DATA MINING ANALYSIS OF ECONOMIC INDICATORS OF COUNTRIES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ERDEM GÜNGÖR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

AUGUST 2020

Approval of the thesis:

DATA MINING ANALYSIS OF ECONOMIC INDICATORS OF COUNTRIES

submitted by **ERDEM GÜNGÖR** in partial fulfillment of the requirements for the degree of **Master of Science** i**n Statistics, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar

Dean, Graduate School of **Natural and Applied Sciences**

Head of Department, Statistics

Doç. Dr. Ceylan Talu Yozgatlıgil

Statistics, METU

Examining Committee Members:

Prof. Dr. İnci Batmaz

Statistics, METU

Doç. Dr. Ceylan Talu Yozgatlıgil

Statistics, METU

Doç. Dr. Könül Bayramoğlu Kavlak

Industrial Emgineering, Bogazici

Date: 17.08.2020

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Erdem Güngör

Signature :

ABSTRACT

DATA MINING ANALYSIS OF ECONOMIC INDICATORS OF COUNTRIES

Güngör, Erdem
Master of Science, Statistics
Supervisor: Doç. Dr. Ceylan Talu Yozgatlıgil

August 2020, 240 pages

Data Mining is becoming a famous analysis day by day to reveal the hidden information within big data. In the study, we use data mining techniques on the economic indicators of the countries. The four data mining techniques are to be implemented on the dataset. Making homogenous groups of the countries whose economic characteristics are similar are obtained by the Clustering Algorithm. After the clustering algorithm is performed, we pass to Association Rule Data Mining to investigate the most exported products by Switzerland to the other countries. With the clustering and association rule mining, we complete the first stage of the data mining that is so-called as unsupervised learning. In the second stage, we build up both classification and regression models with panel data based on the new variables that are obtained by the Principal Component Analysis. The main aim of the second stage is to determine the most important economic predictor variables that have an effect on the grouping of the countries and have an effect on the main economic indicators such as Gross Domestic Product (GDP), Gross National Product (GNP), etc.

Keywords: Data Mining, Clustering, Association, Classification, Panel data analysis

ÖZ

ÜLKELERİN EKONOMİK GÖSTERGELERİNİN VERİ MADENCİLİĞİ ANALİZİ

Güngör, Erdem
Yüksek Lisans, İstatistik
Tez Yöneticisi: Doç. Dr. Ceylan Talu Yozgatlıgil

Ağustos 2020, 240 sayfa

Veri Madenciliği, verinin içinde saklı gizli bilgiyi ortaya çıkarmak için kullanılan ve günden güne popüler olan bir yöntemdir. Bu çalışmada, veri madenciliği yöntemleri, ülkelerin ekononomik indikatörleri üzerinde uygulanmaktadır. Dört veri madenciliği tekniği veri seti üzerinde uygulanmaktadır. Ekonomik açıdan birbirine benzeyen ülke gruplarının oluşturulması Kümeleme Algoritması ile elde edilmektedir. Kümeleme algoritmasından sonra, İsviçre tarafından dünyaya en fazla ihraç edilen ürünleri soruşturmak için İlişki Kuralları Veri Madenciliği kullanılır. Kümeleme ve İlişki Kuralı Veri Madenciliğiyle, veri madenciliğinin ilk kısmı olan ve denetimsiz öğrenme olarak adlandırılan kısım tamamlanmış olmaktadır. İkinci kısımda, veri boyutunu indirgemek amacıyla kullanılan Temel Bileşenler Analizi ile elde ettiğimiz yeni değişkenlere dayalı sınıflandırma ve panel veri ile regresyon modelleri inşaa ederiz. İkinci kısmın temel amacı, ülkelerin gruplandırılması ve Gayrı Safi Yurtiçi Hasıla ve Gayri Safi Milli Hasıla gibi ana ekonomik göstergeler üzerinde etkisi olan en önemli ekonomik tahmin edici değişkenleri belirlemektir.

Anahtar Kelimeler: Veri Madenciği, Kümeleme, İlişik, Sınflandırma, Panel Veri Analizi

Dedication

ACKNOWLEDGMENTS

I wish to express my deepest gratitude to my supervisor Assoc. Prof. Dr. Ceylan Talu Yozgatlıgil for her guidance, advice, criticism, encouragements and insight throughout the research.

I would like to thank Prof. Dr. Özlem İlk Dağ for her suggestions and comments.

I would like to thank my family for their excellent encouragements and their wonderful suggestions.

I would also like to thank my friend Alp Ceyhan for his help.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

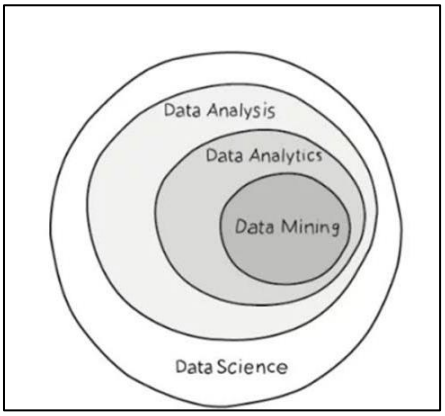
FIGURES

CHAPTER 1

INTRODUCTION

1.1 Data Mining

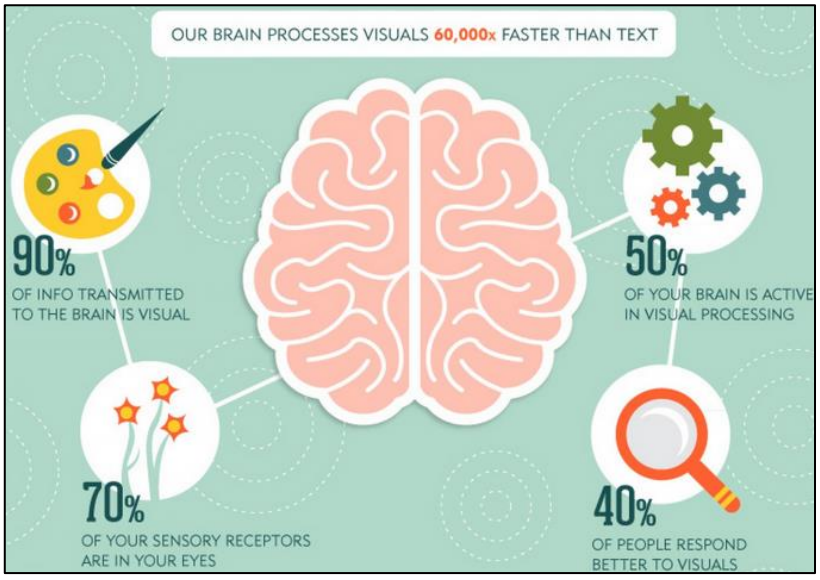
Data Mining is an improving concept but actually, this is the name of the process of getting useful information from data. If we implement the analysis on the large set of data, it simply means that we are mining valuable information that is not seen by using traditional analysis. Therefore, we need to improve existing methods to figure out how to deal with big data. The term that is called data mining comes from the fact that we are handling a large body of information and getting the information hidden in big datasets. The main difference of data mining from other data analysis terms such as data science, statistical data analysis, statistics, and machine learning is the way in discovering knowledge. When we look at the below figure, we better understand the place of data mining among other sciences.



(What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, Big Data and Predictive Analytics?)

Figure 1.1 Data Mining Place among Other Sciences

1.1.1 What is Data Mining?

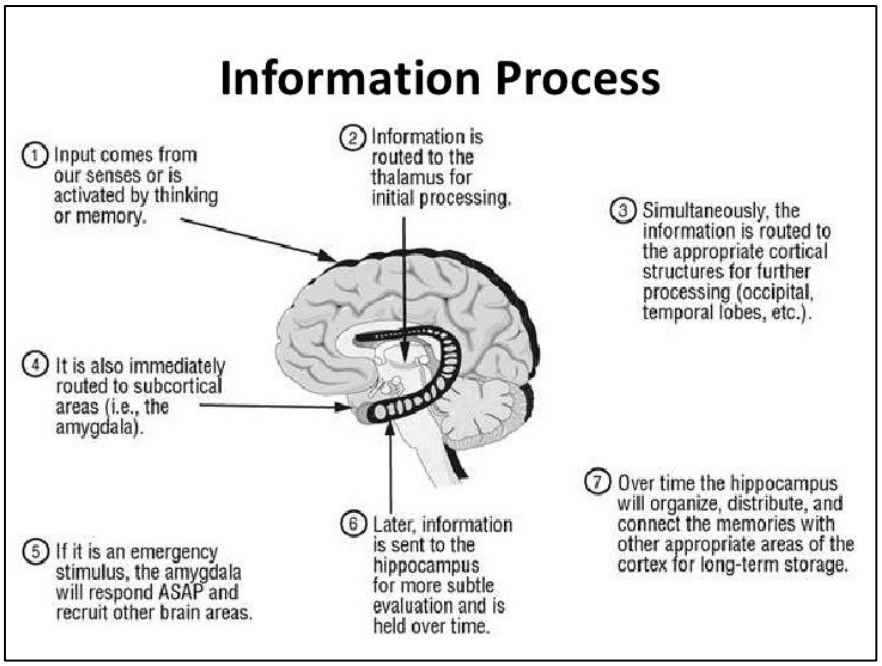


(SocialMediaToday, Advanseez)

Figure 1.2 Information Processing Rates in Brain

Are you aware that each day, you are processing the large body of information in seconds even in milliseconds by your eyes? Your eyes are the biggest source of the information transmitted to the brain by the neural networks. 90% of the information coming from many areas is caught up with the eyes and the brain processes this information. The human brain can reach the remarkable feat of processing an image seen for just 13 milliseconds. This lightning speed invalidates the previous record speed of 100 milliseconds reported by previous studies. The transmission of the information is performed by the cells that are called neurons in the brain. Neurons work together and transmit information through the pathways. These pathways are called neural networks. They are also called information roads. After making inferences about the incoming signals, gain understanding from the signals is the output of the process. Since information is transmitted throughout the neural networks, the processing speed of the incoming signals depends on how effectively neural networks are organized to make the information reached the point of processing in the brain.It means that if neural networks know the pathways of the

information more compared with the previous transmission of the same and related information, this contributes significantly to the increment of the processing speed. This is an indication of how much the repetition of information is important.



(Wicaksana, 2011)

Figure 1.3 Brain Parts

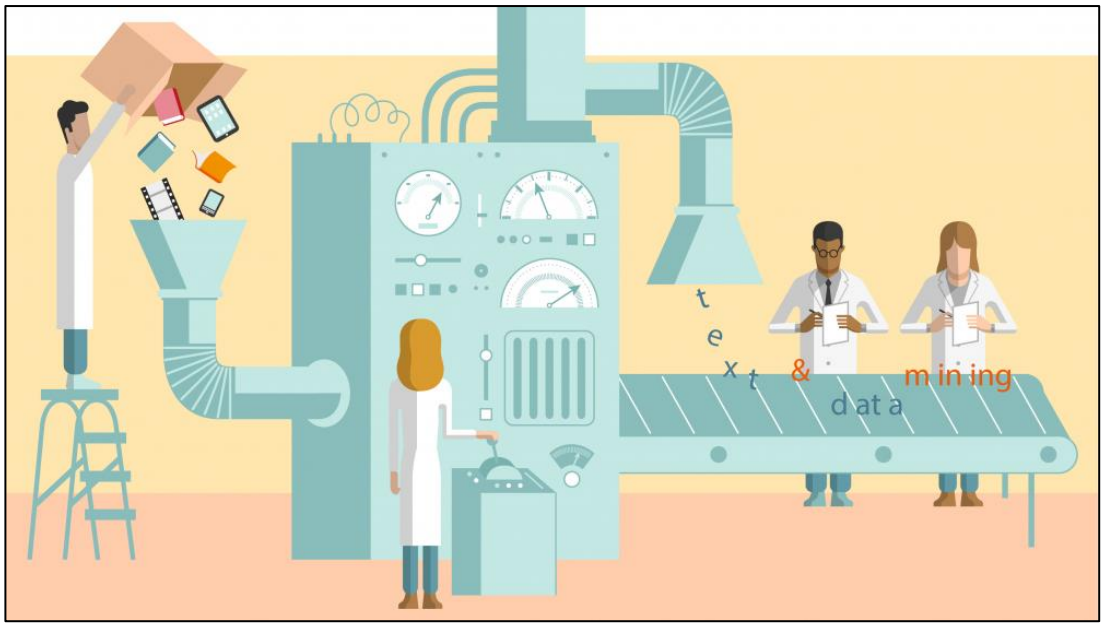
Information processing starts with input from the sensory organs, which transform physical stimuli such as touch, heat, sound waves, or photons of light into electrochemical signals. The sensory information is repeatedly transformed by the algorithms of the brain in both bottom-up and top-down processing. Once information is processed to a degree, an attention filter decides how important the signal is and which cognitive processes it should be made available to. In order for the brain to process information, it must first be stored. There are multiple types of memory, including sensory, working, and long-term. First, information is encoded.

There are types of encoding specific to each type of sensory stimuli. For example, verbal input can be encoded structurally, referring to what the printed word looks like, phonemically, referring to what the word sounds like, or semantically,

referring to what the word means. Once information is stored, it must be maintained. Some animal studies suggest that working memory, which stores information for roughly 20 seconds, is maintained by an electrical signal looping through a particular series of neurons for a short period of time. There are numerous models of how the knowledge is organized in the brain, some based on the way human subjects retrieve memories, others based on computer science, and others based on neurophysiology.

The semantic network model states that there are nodes representing concepts and that the nodes are linked based on their relatedness. For example, in a semantic network, "chair" might be linked to "table," which can be linked to "wooden," and so forth. The connectionist model states that a piece of knowledge is represented merely by a pattern of neuronal activation rather than by meaning. There is not yet a universally accepted knowledge organization model, because each has strengths and weaknesses. As it is clearly understood from the brain process and its steps, each day we are exposed to take lots of information coming from a variety of resources and turn the signals into valuable information and we can use this information and can store them to utilize them later on.

At any moment, our brains process information according to their regions and separate information from the others to react to and respond to what we have gotten. For example, our brains do not mix signals attributed to different sensory sources. Neural Networks directly mines information and makes them reach suitable places of parts of the brain and the related cortexes produce information and replies. This is the simplest and the most complicated example of mining the data and the illustration of turning information into valuable knowledge. Therefore, we actually are mining the data each second even each millisecond to produce information and reply. All of these give us a definition of data mining.



(vizyonergenç, 2019)

Figure 1.4 Data Mining

Data Mining (DM) is the art of extracting valuable and useful information that is hidden within the large body of information and dataset. DM is becoming popular nowadays. By collecting and investigating data, we are able to discover and find some patterns that are one of the outputs of data mining. These patterns can be more frequently seen as dummy variables, names, sequential data, or non-sequential data, etc. By transmitting those patterns to the visual area, we can produce more clear results. As we mentioned in the previous pages, the visualization of the data is the most important way to see the invisible information that is hidden in the large body of information. The amount the data or the volume of the data plays a key role in finding the desired results. Since we are dealing with big data in DM, we can produce lots of patterns, correlations, dependencies, and models with the data mining techniques. By visualization methods, we can gain structural learning more easily and see the relationships better. In the next pages, we are going to provide more visual representations in order for the reader to better understand and make inferences on the findings.

1.1.2 What are Data Mining (DM) Steps?

The more information we have, the more detailed research we need to perform, and the more steps we need to complete to find the desired results. DM provides us with useful and hidden information. We cannot talk about whether or not additional knowledge is included in the data. Thus, bringing some facts to light requires steps on the way to reach the aims. DM is mainly composed of 7 different stages.

Data Integration: In the beginning, we need to collect data coming from a variety of resources. It can include economic data, health data, transformation data, climate data, space data, etc. As it is clearly understood from the name of this step, we are making use of a variety of resources of the information coming from all sciences. Data Mining begins with a collection of big data because we need to mine some useful knowledge among large information. A Combination of the datasets and building up big data is a starting point for us to take a look at the big picture of the data.

Data Selection: After collection of the data is done, we need to select the data by our aims. If we handle all data, we can end up with a dead end. The selection of the relevant data helps us figure out the most important answers. According to our aims, the selection procedure must be performed for better understanding big data and must cover the most important and critical elements of the data.

Data Cleaning: Due to the fact that we are interested in big data, it is most probably to have incompatible data that includes inconsistent data, noisy data, etc. After selecting the relevant information, we pass to the stage of cleaning the selected data. Owing to natural reasons, collected data sets can be in different forms. In order to strike a balance of combined datasets, it is requested to have the same forms of the data.

Data Transformation: In order to prepare data for mining, we need to make some transformations. It is because of the reasons of the existence of having different features of datasets and having the obligatory to read data on software that we are going to use. Non-identification of the data on software like R-STUDIO, ORANGE DATA MINING, etc. can lead us to come across complex situations. One of the most valuable extractions of the relevant information is to interpret the numbers. As long as the transformations are made well, we need to presuppose possible outcomes and possible interpretations of the outputs. Therefore, in the stage of transformation, not only do we need to take the data transformation into account but also we have got to make inferences of likely outcomes or likely interpretations of the results.

Data Mining: After transformation, we are ready to mine data and get useful and hidden patterns. With the implementation of the data mining techniques, the discovery of shape and patterns start. By implementing data mining techniques, we visualize a lot of shapes and plots to explore data. The plots and shapes provide insight into the relevant data sets. Their explanatory and exploratory powers help us see the invisible information in the assessments. Thus, we need to draw special attention to graphs to gain an understanding on the related data sets. The process of evaluation gives place to the analysis of the relevant data sets and after making an interpretation of the processed shapes and graphs, we can get benefit from usual statistical data analysis and other analyses.

Pattern Evaluation and Knowledge Presentation: This step involves visualization, transformation, removing redundant patterns, etc. from the patterns we generated. The irrelevant patterns that are not informative are removed from analysis in order to get focused on the relevant plots.

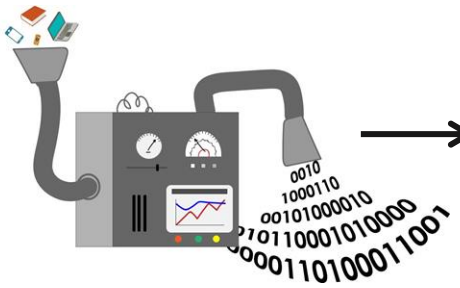
Decision and Use of Discovered Knowledge: This step helps the user to make use of the knowledge that is obtained from the assessments of the analyses and interpretations of data mining. However, the results of mining data do not have to have meaningful comments as we expect, because the information or output must

give the new information. Comments are made not only by available knowledge but also the outputs of data mining.

In Summary;

Collection of data coming from a variety of sources forms big data

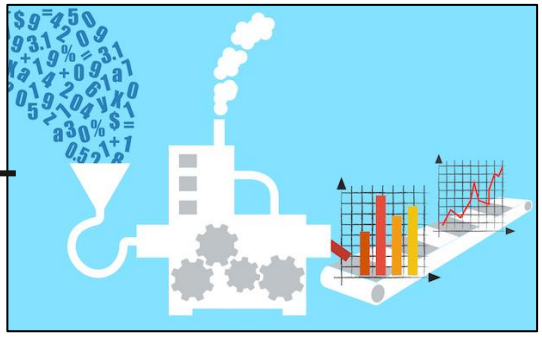
(Sinnreich & Romzek, 2018)



With data mining techniques, we are producing some patterns

(McKell, 2018)

By means of patterns, we are discovering knowledge and figuring out the hidden information



(*3 pictures banded together by ERDEM GÜNGÖR*)

Figure 1.5 Data Mining Steps in Summary

The main focus of the study is the exhibition of implementation of the data mining techniques on the data sets that are obtained by data.worldbank.org and tradema.org bank. Apart from the statistical data analysis; we focus on discovering the hidden information in the data sets. This study can be an example to show how to carry out the data mining techniques on the data sets.

1.2 Motivation of the study

In our study, we are aimed at performing data mining techniques on data sets that we collect and combine. This data is mostly made up of economic indicators that determine and give the clue for the economic situation of a country.

The first aim is to investigate the economic situation of the countries by implementing data mining techniques. We also assess the similarities and differences in economies of the countries.

The second aim is to try to bring hidden and useful information to light with the help of data mining techniques. Since the economy is a wide topic, the reasons and activities that make countries go one more step ahead by comparison with the other countries are showed.

The third aim is to draw a comparison between data mining techniques and give the best one according to the accuracy rates of each method. Due to the existence of many data mining techniques, we choose the best one that gives more logical results. Of course, we do not come across meaningful results. However, they can be informative. Just as making comments on the outputs, we make interpretations of the utility and accuracy of data mining techniques.

The fourth aim is to build up economic models by using data mining techniques. In the next chapters, it is mentioned that regression is one of the data mining techniques. Through data mining, we are going to depict graphical models that are used for predicting target variable that is the same as a dependent variable in the regression. However, for our study, we perform panel data analysis because our data sets contain information about the year from 2000 to 2015 for each individual.

The main purpose of the study is to find suitable explanations for our aims. In the following chapters, we can enhance the aims, but for now, it is good to know the aims that we ordered above.

1.3 Research Questions of the Study

In the study, together with the determined aims, we are looking for some research questions to find answers. They are parallel to our aims but those questions are the main points and targets to be investigated.

They are ordered as following;

- **Can we obtain more information about the countries' past, current, and future economic situations by using data mining procedures?**

It is a very well known fact that countries' economic background provides us with their current economic situation that affects the future economic situation of countries. Most of the time, though statistical data analysis such as logistic regression analysis, multivariate regression analysis, simple linear regression analysis, and generalized linear regression analysis, we can reach suitable solutions and process of the economy and get the view about for the current situations. With the data mining techniques, we are going to delve into data sets to find more information on the way to make better predictions and descriptions of the economy-related information that exist in datasets.

- **What are the most important economic indicators that enable a country to improve?**

Throughout years, economy experts have struggled with making inferences about the economic situation of a country by making use of some indicators that give clues about the economic situation. The most important indicators to produce outcomes and make comments on the economic situation of a country are known as Gross State Product, Employment Rate, Unemployment Rate, Inflation Rate, Imports and Exports of Goods and Services, Total Debt Services, etc. As we bring an explanation to data, the random variables that have explanatory power to explain the economic situation of a country are to be obtained and separated as the most significant variables.

- **Do the data mining techniques/tasks help us to see the invisible information/models hidden in the large body of databases that are made up of economic indicators?**

As we clearly state under the description of the data mining terminology, it is a tool that gives us a chance to discover the hidden information. Thus, the main focus of the study is intensified around the data mining to bring information that is not visible by traditional techniques into the light. Moreover, data mining and data science are closely aligned and connected. Data mining gives us results with the greatest and the most beneficial ways. Employing data mining, we go to outcomes rapidly and effectively.

- **Can the process of digging through data to discover hidden connections and predict future economic trends contribute to our understanding of statistical data analysis?**

Until the terminology of data mining comes into existence, we are doing a similar thing with statistical data analysis. The main thing that separates traditional statistical data analysis from data mining is coming from the fact that data mining handles big data and the purpose of the data mining vary by analysis goals. Data mining is separated from statistical data analysis by the ways of analyzing the data. It is certain that in data mining we have aims but we are trying to reach this aim by looking at datasets from a general point of view utilizing the different techniques. On the way to get results, we are also aimed at figuring out different results that are not seen in traditional statistical data analysis. Giving results out of obtained ones by different methods, the data mining describes the findings in different formats and it responds to the aims and research questions in different ways.

1.4 Literature Overview

To be able to give the previous study on data mining neatly, it is given in table format.

Table 1.1 Literature Overview

LITARATURE OVERVIEW - DATA MINING		
No.	**Writers & Authors**	**Study Sumamry**
1	Schroeder Jr. (1998)	Neural Networks utilized for the problems of business
2	Exner(1998)	Data Mining Steps and Knowledge Discovery Process
3	Lingras & Yao (1998)	Investigation of basic ideas of data mining employing the theory of rough set
4	Mcclean et al. (2000)	Usage of domain knowledge, e.g., integrity constraints or a concept hierarchy, to re-design the database and assign sets to which missing or unacceptable outlying information may belong.
5	Skarmeta et al.(2000)	Usage of a semi-supervised agglomerative hierarchical clustering (ssAHC) algorithm to text categorization which is made up of allocating text documents to predefined categories.
6	Zanasi(1998)	Examination of the data mining application to competitive intelligence analysis
7	Oguchi et al.(1999)	A large-scale Parallel Computer Cluster is set by connecting 100 personal computers by means of a general-purpose ATM network. Applications to parallel data mining are examined and explained.Parallel Data Mining will be built up.
8	Oh et al.(1999)	It is aimed at advancing the performance and features of our 2-D related protein databases, with particular concentration on the tools which are utilized for database mining by means of a systematic analysis of knowledge known as post-planned analysis.
9	King et al.(2000)	Presentation of a novel data-mining approach to estimate protein functional class from sequence.
10	Duan et al.(2000)	Investigation of conclusion that the interaction between aromatic and backbone amide groups is of general significance to protein structure because of its strength and prevelance occurrence.
11	Sticht(2002)	Application of data mining on Structural Biology
12	Pistilli & Arnold(2010)	Application of data mining to develop student access
13	Wade(2006)	Attempt to bring together the data reduction techniques by virtue of data mining and knowledge discovery process to get the information which can be important for correct implementation and examination of monitoring

Table 1.1 (continued)

		operation from large datasets that are found in plant operations.
14	Zhang & Block(2009)	Application of data mining techniques approach to select the most important variables and dispose of insignificant variables from the database.
15	Benoît(2002)	Data Mining and Knowledge Discovery Process.
16	Bath(2004)	Application of data mining techniques in exploring knowledge in the field of health and medicine.
17	Schumaker et al.(2010)	Application of data mining techniques in exploring information to improve sport players' success rate and increase the capability of having important knowledge obtained by virtue of the sport acitivites, managers,coaches and scouts.
18	Nicholson (2003)	Application of data mining to design classification on web pages if or not the web pages contain scholarly contexts .Predictive models that are based on four techniques to carry out an estimation the classification of the web page are displayed.They are logistic regression,non parametric disciriminant analysis,classification trees and neural networks.
19	Chen(2004)	Web Mining to figure out significant statistical patterns that display the dependencies between texts. Examination of the web contexts and pages.
20	Gluck(2001)	Application of data mining visualization technique that is called seriation to carry out an explanatory data analysis.The visualizations acquired by seriation technique are employed to explain environment risk assessments.
21	Wang & Yang(2007)	Suggestion of the new segmentation algorithm by mining Web data by means of search engines. As Chinese is an ideographic character-based language, the words in the texts are not seperated by white spaces. Therefore, Indexing of Chinese documents is not possible without a suitable segmentation algorithm. The new segmentation algorithm will be employed to mine web data.
22	Barry et al.(2007)	Improvement of new and novel way for analysis Web pages.This mining procedure deals with the similarities and differences between queries in search engines and detects the seqeunce of query transformations.They are illustrated as graphically by enabling readers to realize and analyze the things easily and provide richest understanding for the searchers.
23	Thelwall et al.(2010)	Application of data mining for sentiment analysis in one of social networks called Myspace.Researchers were trying to figure out whether or not comments are including either positive or negative emotions.
24	Rokach et al. (2011)	Investigation of evaluation of the researches by bringing together three important data mining tenchniques which are logistic regression,decision trees and artificial neural networks.
25	Giudici (2001)	Study of one of the data mining techniques that is so-called Bayesian Netwok Structure
26	Skica et al.(2015)	Implementation of data mining techniques to identify the relationships between the economy and general government sector size.Detection of the most significant variables that are utilized to ascertain the relation between econıomy and the size of general government sector.

Table 1.1 (continued)

27	Costantini et al.(2001)	Examination of non-linear Principal Component Analysis that is way of integrating all variables under more than 1 PCs with optimal-scaling
28	Vaduvescu et al. (2009)	Investigation of old photographic plates and CCD image archives in order to bring hidden and unexplored information into the open.By virtue of this study, researchers attempt to build up the orbits of Near Earth Asteroids(NEAs)
29	Weber et al.(2009)	Study on data mining and machine learning methods for an analysis of complex systems in computational biology in order to construct mathematical modelings and prediction of gene-expression patterns.
30	Gieger et al.(2004)	Investigation on interpretation of the genes expression with help of text mining which is the other area to mine the text data.
31	Glaser et al. (2004)	Data mining Implementation on the risk levels of diseases that are attributed to food.
32	Gould (2004)	Study on side-effects of the drugs that are unlikely to be realized during clinical trials.Thus, scientists have gathered a bounty of reports of adverse effects of drugs so as to get to the bottom of problems with respect to drug usage.
33	Unwin (2004)	Study on the detection of common problems and their potential solutions in science and engineering applications.
34	Charaniya(2009)	Investigation of the transcroptome data so as to brighten the efficiency characteristic by taking comparions of the transcriptomes of several NS0 cell lines with a wide range of antibody productivity into consideration.
35	Qin (2000)	Survey about the Konwledge Discovery Process and examination of its concepts and applications to extract information from vast amount of databases.Research handles the process of how Knowledge Discocery Process relates to knowledge management.
36	Strickland & Willard(2000)	Examination of current immigration system
37	Žalik(2005)	Knowledge about data mining and its usage to bring the informartion to light by employing the data mining tools.Data Mining Application on GSM interface at stationary telelphone stations and explore navigation patterns through the web pages in order to provide and build up fast accessibility to information.
38	Eis et al. (2000)	Data Mining on The Kinease Sequence Database(KSD).This database forms an accumulation of protein kinase sequences grouped into families by homollogy of their catalytic domains.KSD enables the researchers to figure out and investigate family and protein names and provides researches with statistical tools for analysis.
39	Burghaus et al. (2004)	Data Mining Implementation of the Complex Polymer Process
40	Mazzatorta et al. (2002)	Scaling Procedure while mining data so as to extract informartion toxicity value in fish.
41	Du et al. (2002)	Data Mining Appplcation on alkenes for searching an accurate quantitative relationship between the molecular structure and retention indices of gas chromatography and remarks on finding a new variable that contributes signifificantly to prediction accuracy of regression models.

Table 1.1 (continued)

42	Marx et al. (2003)	Data Mining and Supervised Machine Learning Implementation on the clusters of the genes obtained by Nationall Cancer Institute
43	Adams & Schubert(2004)	Data Mining approach to polymer science and disputes on possible relationships between quantitative structure and property
44	Teckentrup et al.(2004)	Implementation of Kohonen Neural Networks on drug discovery process in order to make an analysis of combinatorial libraries for the similarity and diversity and to select descriptors for structure- activity relationships.
45	Helma et al.(2004)	Discovering the utility of data mining and machine learning algorithms for the induction of mutagenicity structure-activity relationships (SARs) from noncongeneric data sets
46	Divsalar et al. (2012)	Data Mining Application in building up models for bankruptcy predicition by virtue of gene expression programming and multi-expression programming.
47	Ohrenberg et al. (2005)	Data-mining Applications and Evolutionary Optimization to improve the efficiency of high-throughout experimentation (HTE) to discover new materials, drugs, or catalysts.
48	Hori et al. (2002)	Apriori Algorithm Implementation on the database watchdog task.
49	Shankar & Winer (2006)	Data Mining Applications in looking for customer relationships
50	Mena (1996)	Survey about accounting for competitive intelligence by applying data mining and data warehouses to artificial intelligence.
51	Köksal et al. (2010)	İmalat Sektöründe Kalite İyileştirmede Veri Madenciliği Tekniklerinin Kullanimi

1.5 Dissertation Flow

In order for the reader to keep track of the study flow, let us summarize the general concepts of the study.

Chapter 1 explains the introduction to the study and the general definition of Data Mining. We introduce data mining in this chapter by making a connection with the information process that occurs in the human brain so that readers can better understand the logic behind the terminology of data mining. In chapter 1, we are aimed at giving a variety of studies that are implemented with relevant to data mining.

Chapter 2 brings an explanation to the dataset that is used for data mining. All variables are given and are categorized according to our aims. The response

variables that are called as target variables in data mining are to be introduced. Moreover, the explanatory variables are determined. By means of classification of all variables, the illustrations and interpretations of the results are expected to be simplified to some extent.

Chapter 2 keeps going on an explanation of data mining techniques that are utilized in the following chapters.

Chapter 3 starts with the application of the first data mining techniques in data. The first data mining technique is called **Clustering.** We are implementing clustering types on the most important economic indicators to design clusters of the countries. Data points within the designed cluster have high similarity between the other data points within the same cluster and have a high difference between the other data points that are placed on different clusters. We perform all cluster types on selected economic indicators of a country to distinguish the countries from each other. Then, we put one of them into the study to obtain the clusters of countries to see which countries belong to which clusters. In the study, only one clustering algorithm that performs the best is illustrated. This simply shows the comparison between the economic levels of the countries provides insight into a better understanding of economical differences between the countries, therefore we can gain a view on the countries that have a dynamic economic activity.

Chapter 3 continues with the second data mining technique which is called **Association Rule**. Association Rule is mostly used data mining technique in market basket analysis to show how frequently an item set places in a transaction of a good. Therefore, we are obligated to turn the market basket analysis into the form that gives us important results according to data on hand. Under the association rule, we talk about apriori algorithm technique to figure out the most frequently together exported products of Switzerland to the world. This rule provides us with relationships and interesting associations among large sets of data.

Chapter 4 begins with building up models. The first model comes from probabilistic graphical models. Thus, chapter 4 deals with the most important data

mining technique that gives us graphical models that are called **Classification.** The classification technique is the best known technique of setting up models based on discrete and continuous random variables. This technique is categorized under the predictive tasks of data mining. Moreover, Classification is a data mining procedure that assigns items in a collection to target categories or classes. The aim of classification is to accurately predict the target class for each case in the data. In literature, classification techniques vary by purposes such as Decision Trees, Random Forest, Bayesian Network Structures, Support Vector Machines, etc.

Chapter 4 goes ahead with handling the fourth data mining technique that is called **Regression.** We know from the experience that regression analysis can be implemented for the sake of designing predictive models. There are plenty of regression types in literature. According to data and our aim, we benefit from the other regression types such as logistic regression, Poisson Regression, Simple Linear Regression, Generalized Linear Regression. In the study, we also take the longitudinal into account to derive the marginal, transitional, and random-effects model.

Chapter 5 is the last part of the study. In this part, we summarize what we have done throughout the study. Because we bring results and conclusion at the end of each part of the analysis, it is not needed to put all the results on chapter 5.

END OF DATA MINING - INTRODUCTION

End of Chapter 1

"Data Mining is the invisible part of visible information."

(GÜNGÖR, 2020)

CHAPTER 2

EXPLORATION OF THE DATA

2.1 Introduction to Data

The data is made up of economic indicators, health, and transportation of the 80 countries that are Best Countries 2019 **(2019 U.S. News & World Report LP).** Overall, 69 random variables (economic indicators, health, military, social, migration, and transportation data) including year effect are taken into consideration to analyze the data sufficiently. Since most of our data related to the economy dominate the other data types that we mentioned above, we make inferences on the economic situations of the countries. Therefore, it is clearly understood from the research questions and aims, the implementation of data mining on whether or not having knowledge on the economic variables gives information about economic situation of a country is the main focus on the way to better understand past, current, future economic activities of a country. We use the tasks of data mining that are predictive and descriptive so as to fulfill our aims and figure out answers to the research questions. Before going with data analysis, the most important response variables are listed below in order for readers to better understand the factors' (or attributes in data mining and explanatory variables in statistics) effects in the process of data analysis.

Datasets, as we stated at the beginning of this session, are composed of 80 Countries. We collect this data for each country throughout 15 successive years from 2000-2015 by the website of **https://data.worldbank.org**. In our datasets, we deal with missing cases by utilizing missing data analysis with the help of packages that are available in R Software. For example, the missing cases are quite more in North Korean. About 90% of data information for North Korean is missing. Instead of imputing the missing cases for this country, we remove this county from the

analysis, because if we implement some ways of handling missing cases, it cannot represent the reality of observations. However, we get the use of the variables of this country as much as possible in order to mine data. One of the main focuses of mining data is to search for the factors behind any links that are placed between economic variables. The data set is composed of annual data.

2.2 Indicator Variables

2.2.1 Response Variables (Target Variables, Dependent Variable)

Our data is made up of 69 variables. The number of data points is approximately 100.000. Among the 69 indicator variables, for the modeling part in chapter 4, we focus on the specific random variables that are dependent. The dataset collected from hundreds of different areas of life allows us to determine important patterns and capture the hidden information and variables to explain the variability on the response variables. The target variables that are determined are ordered below,

- GNI (GNP), Atlas method (current US$)
- GNI (GNP) per capita, Atlas method (current US$)
- GNI (GNP), PPP (current international $)
- GNI (GNP) per capita, PPP (current international $)
- GNI (GNP) per capita growth (annual %)
- GNI (GNP) (current US$)
- GNI (GNP) growth (annual %)
- GDP (current US$)
- GDP growth (annual %)
- GDP per capita (current US$)
- GDP per capita growth (annual %)
- GDP, PPP (current international $)
- GDP per capita, PPP (current international $)

- It is stated in the literature resources (LaComb, 2019) that there is no substantive difference between Gross National Income and Gross National Product. GNI is simply the current way of stating the concept known as Gross National Product. These two terms are designating the same concept that is the total output of new final goods and services produced by country's residents and the productive factors they own, regardless of where in the world that production takes place.
- **The "I" in GNI and "P" in GDP are** changeable because, in national income accounting, all productions provide an equal amount of income for someone and vice versa. Moreover, we could not reach the variable of GNI per capita (current $) to make the analysis effective. Instead of utilizing this variable, we use the variable of GNI per capita with Atlas Method (current $). Even if there is no significant difference between these two terms, they are slightly different concepts. In this study, we are going to use the Gross National Income that is the same concept as the Gross National Product. The only difference is summarized at the below table;

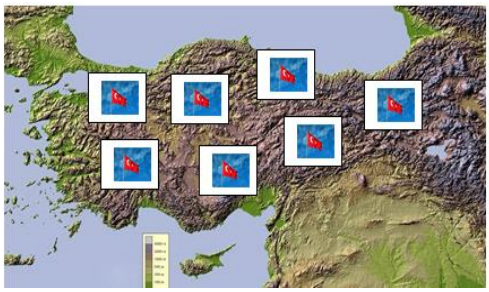
Table 2.1 Differences between GNP vs GNI

Differences	GNP(Gross National Product)	GNI(Gross National Income)
Definition	Gross national product includes the earnings from all assets owned by residents. It even includes earnings that don't flow back into the country. It omits the earnings of all foreigners living in the country, even if they spend it within the country. GNP only reports how much is earned by the country's citizens and businesses, no matter where it is spent in the world.	GNI measures income earned, including income from investments, that flows back into the country.
Calculation	**GNP** = GDP – NFIA = GNP = GDP + [(income earned on all foreign assets – income earned by foreigners in the country)].	**GNI** = GNP + [(income spent by foreigners within the country) – (foreign income not remitted by citizens)].

As it is seen from the table, there is no significant difference between these two terms. The only difference is attributed to the condition of whether or not the foreign incomes happened within the border of the country. Therefore, we do not need to make so many efforts to distinguish these terms. During the analysis of data, we are making comments on the GNI.

Now, let us look at the main differences between the two main variables that are GNI (GNP, Gross National Income/Gross National Product) and GDP (Gross State Product).

a-) <u>GNP</u>



b-) <u>GDP</u>

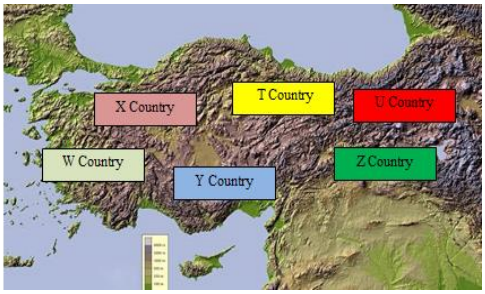


Table 2.2 Comparison of GNP vs GDP

Differences	GNP(GNI,Gross National Product)	GDP(Gross Domestic/State Product)
Definition	Total income earned by the nation's factors of productions, regardless of where they are located,meaning GNP is the market value of goods and services produced by all citizens of a country both domestically and abroad.	Total income earned by domestically-located factors of production, regardless of nationality.
Meaning	The worth of goods and services produced by the county's citizens irrespective of the geographical location is known as Gross National Product (GNP)	The worth of goods and services produced within the geographical limits of the county is known as Gross Domestic Product (GDP).
Indicator	Gross National Product represents how the citizens are subsidizing to the country's economy.Measurement of production by nationals.	Measurement of the overall health and size of the country's economy.Measurement of domestic production.
Study Area	Study of how residents in a country are contributing to the economy	Study of Domestic economy of a country
What prodution?	Production of products by the enterprises sustained by the residents of the country.	Production of products within the country's borderline.
What is taken into account?	Goods and services produced by citizens living outside the country	Goods and services produced by foreigners within that country
What is not calculated?	Goods and services produced by foreigners within that country	Goods and services produced by citizens outside the country

Table 2.2 (continued)

Productivity Measure	Productivity is measured on an international scale.	Productivity is measured on a local scale.
Calculation	GNP can be computed by adding consumption, government spending, capital spending by businesses, and net exports (exports minus imports) and net income by domestic residents and businesses from overseas investments. This figure is subtracted from the net income earned by foreign residents (NFIA) and businesses from domestic investment.	**Consumption(C):** The price value of the consumption of goods and services obtained and consumed by the country's households/residents. This forms the largest part of GDP
		Government Spending (G): All consumption, investment, and payments made by the government for utilization currently.
		Capital Expenditure by Businesses/Investment(I): Spending on purchases of fixed belongings and unsold stock by non-public businesses
		Net Exports(X): Represents the country's balance of trade (BOT), where a positive number bumps up the GDP as country exports more than it imports, and vice versa(exports-imports)
	GNP = GDP – NFIA = GNP = GDP + [(income earned on all foreign assets – income earned by foreigners in the country)].	**GDP =** Consumption+ Investment + Government Spending + Net export. = C+I+G+X

Table 2.3 Summary of GDP, GNI and GNP

Income Earned by:	GDP	GNI	GNP
Residents in Country	C+I+G+X	C+I+G+X	C+I+G+X
Foreigners in Country	Includes	Includes If Spent in Country	Excludes All
Residents Out of Country	Excludes	Includes If sent Back	Includes All
Foreigners Out of Country	Excludes	Excludes	Excludes

2.2.2 Exploratory Variables (Predictor Variables, Independent Variables)

In DM terminology, the exploratory variables are known as attributes, and response variables are known as target variables. During the data mining analysis, we call those variables either predictor, independent, exploratory, or attributes. Apart from 13 variables that are ordered under the head of the response variable category, we have 59 variables, as well in which we state them as exploratory variables. The independent variables include different datasets that are in different forms such as rate, countable, and numeric. In order to keep away from the different scales of the variables, in the next sections, we have got to require the standardization of the response and predictor variables.

2.3 Market Exchange Rates, Purchasing Power Parity and Big Mac Index

In the following chapter, we set up clusters of the countries by countries' economic values. The main purpose of performing the clustering analysis separates the countries from each other by response variables that are stated in the previous part (2.2.1.).

When drawing comparisons between the countries that utilize different currencies, it is needed to turn values, Gross State Product, into a common currency. This can be performed in two ways.

2.3.1 Market Exchange Rates (1$ = 1.10€)

A market exchange rate is the value of one nation's currency versus the currency of another nation or economic zone. This is one way to compare the economic activities of the countries with each other. For example, how many Turkish Liras does it take to buy one euro? As of 18.01.2020, the exchange rate is 6.52, meaning it takes 6.56 Turkish Lıra to buy 1 euro. Before moving on to details, let us give two important definitions under the exchange rates.

Appreciation: the increase in the value of one currency compared to another (e.g. the euro appreciates against the dollar when its value increases from 1.06 dollars per euro to 1.07 dollars per euro).

Depreciation: The loss of value of one currency compared to another (in the previous example, the euro appreciates while the dollar depreciates; more dollar units are required to purchase a single euro)

However, there are some difficulties in terms of making comparisons between the countries when the exchange rate is used.

The first one is attributable to rapid changes in the exchange rates. Rapid changes in exchange rates cannot seem to be meaningful to make comments on the current

economic situation of the countries and give rise to alterations of the variable in interest such as GDP, meaning artificial changes can cause unreal interpretations of the economic activities. For example, a one-month depreciation of the US$ by 8% against the Japanese Yen would reduce the dollar value of the Japanese economy by 8%. It is understood from this example that the fluctuations in the rates are more to do with changes in the exchange rate than changes in the underlying state of the Japanese economy. Thus, drawing comparison by depending on the fluctuations in the exchange rates cannot give us valuable information about the economic situations of the countries.

The second challenge is coming from the determination criteria of market exchange rates. Market Exchange Rates are determined by the demand and supply of currencies, which indicate changes in imports and exports of traded goods and services. However, not all countries make trades of their goods and services at the same rate, so currency values are not determined in a coherent way.

2.3.2 Purchasing Power Parities (PPPs)

The purchasing power is one of the alternative ways to deal with making comments on the countries' economic welfare. PPP is a term that is introduced in 1954 in order to explain the quantity of currency needed to purchase a given unit of a good, a common basket of goods and services. It simply works on the equation of the purchasing power of two currencies by taking the cost of living and inflation differences into account. PPP provides us with valuable comparisons of the economic productivity and standards of living between countries. It does it regardless of depending on the exchange rates of the countries instead does it throughout a basket of goods approach.

Calculation Purchasing Power Parity

$$\boldsymbol{S = \frac{P_1}{P_2}, where}$$

$$S = Exchange\ Rate\ of\ cuurency\ 1\ to\ currency\ 2$$

$$P_1 = Cost\ of\ good\ X\ in\ currency\ 1$$

$$P_2 = Cost\ of\ good\ X\ in\ currency\ 2$$

To understand BMI better, we want to explain it with an example.

Suppose that China has a higher GDP per capita (US$20) than the USA (US$16). This is simply an indication that the average Chinese person makes $4 more than the average American. However, this does not necessarily refer to that the Chinese are wealthier. Again suppose that one gallon of orange juice costs $8 in China, and $2 in the USA, i.e. $8 buys a good in China that can be purchased in the USA for $2.1 gallon of orange juice is taken as a reference good in this example. Simply one gallon of orange juice can be bought in China, versus 4 gallons in the USA, with an equivalent amount of money. We can calculate a PPP index for China as compared to the USA equal to ¼. According to orange juice prices, Americans have stronger purchasing power or can buy more value with their money.

The USA has a PPP-adjusted GDP of $16, which has not changed since it is the reference currency. China's GDP is only $10 when adjusted for PPP. This is calculated by multiplying China's unadjusted GDP by the PPP index. In reality, a much wider range of goods that include much more than just orange juice is taken to calculate the PPP index, so that it accurately reflects the average cost of living.

Country	Per Capita GDP (Nominal)	Per Capita GDP (PPP)
United States	47,100	47,400
Germany	40,500	35,900
United Kingdom	36,200	35,100
Japan	42,500	34,200
Mexico	8,900	13,800
Brazil	10,100	10,900
China	4,300	7,400
India	1,200	3,400

(MARIADOSS)

Figure 2.1. Per Capita GDP (Nominal) versus Per Capita GDP (PPP) of Select Countries (2010), Currency is 1$

2.3.3 Big Mac Index (BMI)

Big Mac Index is developed as one of the comparison criteria used for performing cross-check comparisons of the economy of the countries on the basis of purchasing power parities. The main market basket in BMI is made up of the prices of McDonald's Big Mac as an assessment. Simply a burger is replaced in the baskets of goods and it is known as the Big Mac PPP or Burgonomics. Day by day, BMI is becoming more popular in terms of assessing the economic activities of the countries. Some websites like statistica.com are utilizing BMI to keep track of local purchasing power parities.

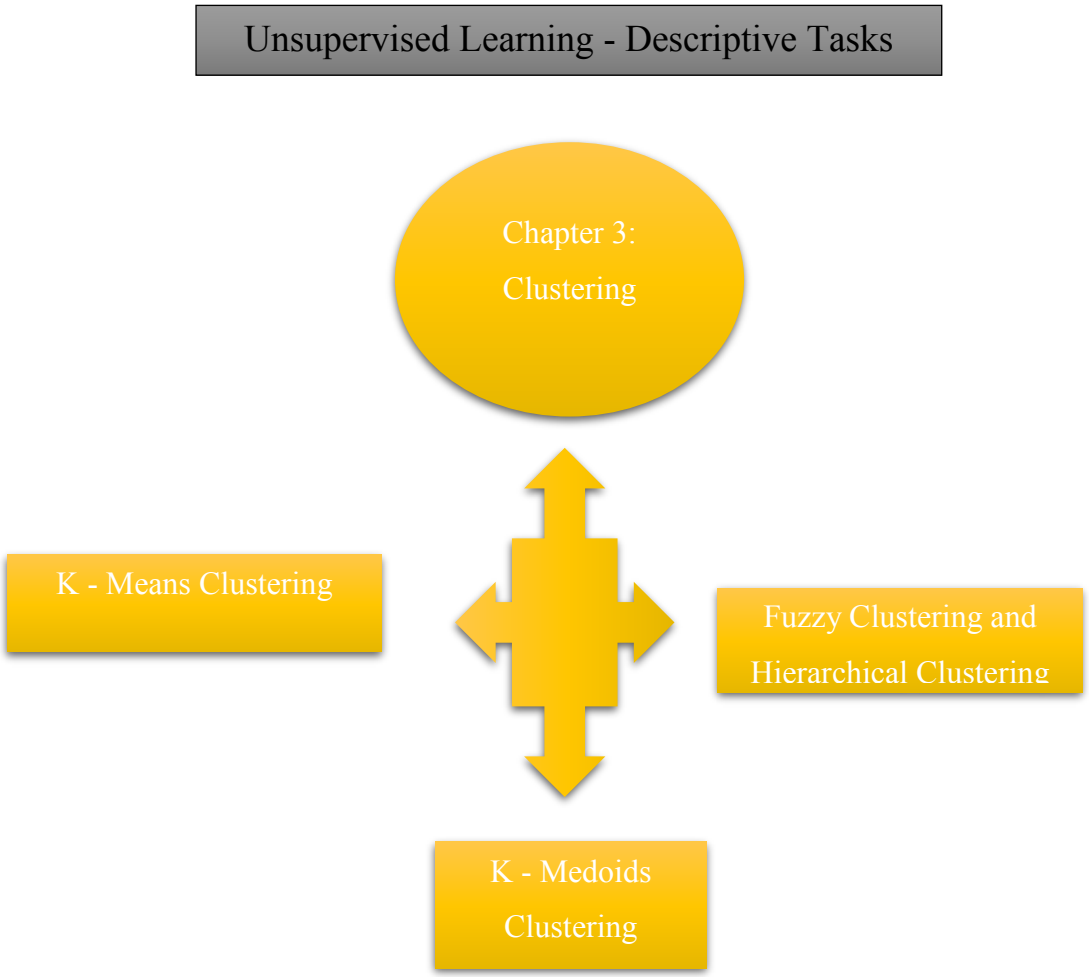
To understand BMI better, we want to explain it with an example.

In January 2019, the Economist summarized that the British pound was undervalued by %24 against the U.S. dollar on the basis of BMI. That is a Big Mac costs $5.58 in the U.S. and 3.19 pounds in the U.K. That discrepancy implies;

$$\textbf{\textit{the exchange rate by BMI}}(\textbf{\textit{PPP}}) = \frac{\mathbf{3.19}}{\mathbf{5.58}} = \mathbf{0.57\%}$$

However, when we look at the market exchange rates at that time, it is 0.78%. As it is clearly stated from this example, there is a huge difference between rates of comparisons. Although there is plenty of criticism about the usage of BMI as a comparison way due to the fact that McDonalds is not found in every country, according to an economist, it is a fairly accurate real-world indicator of PPP since the price of a Big Mac must take local costs of raw materials, labor, taxes and business premises into consideration.

2.4 Overview of Data Mining Techniques to be used likely



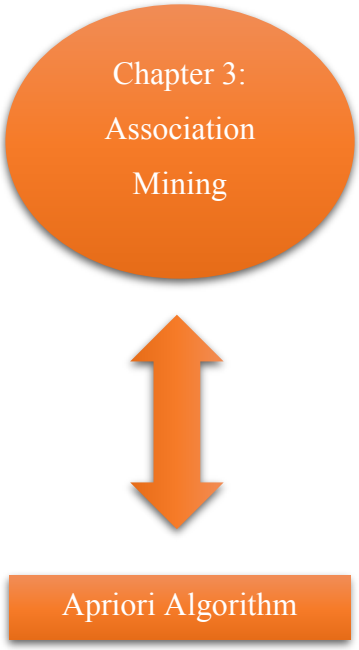


Figure 2.2. Unsupervised Learning Data Mining Techniques
that are used in the study

Data mining techniques can be divided into two groups in accordance with the purpose of mining (Note: In this part of the study, we give the definitions of well-known data mining techniques for the sake of gaining a little information about for which methods are applied to mine the data)

2.4.1 Unsupervised Learning

Unsupervised Learning is the most important way of discovering knowledge. In this learning algorithm, we are just trying to come out with invisible information by using all variables in the dataset. There is no categorization of the variables to perform data mining, meaning there is no either target variable (response variable) or explanatory variable (predictor variable). Its aim is to uncover unknown patterns in the dataset.

2.4.1.1 Clustering (CL)

Clustering is one way to implement an unsupervised learning algorithm in the dataset to discover similar characteristics and to build up separate clusters by finding the nearest data points. It splits data into groups by means of mathematical distance formulations. As it is stated that the main aim of clustering is to build up clusters in which the data points in each cluster are close to each other. The clustering algorithm does not take the analysis of individual data points into consideration that makes it impossible to make inferences on the single data point.

There are plenty of methods under clustering. K-means, K-medoids, Fuzzy clustering, Hierarchical clustering, etc. are the most used ones to discover the similar groups.

2.4.1.2 Association Rule Mining (ASM)

Another unsupervised learning algorithm is called Association Mining. Its aim is to identify sets of items that frequently are sold/bought together in the dataset. The dataset used for ASM is not made up of quantitative; on the contrary, the dataset is composed of qualitative texts or codes representing the items or kinds of stuff. The main usage area of ASM is market basket analysis to detect the most frequently sold items at the same time. By means of this mining type, owners of markets can regulate the places of items, and can effectively improve the relationships with the customers by meeting the requirements of customers in time. Giving precedence to items that are sold together can provide with sales increases in the market.

There are plenty of algorithms to enforce the market basket analysis. The most used one is called Apriori Algorithm. It is the first known algorithm for answering a lot of questions about which items are sold together. There are also other algorithms such as AprioriTid, Apriori hybrid, and Tertius algorithms.

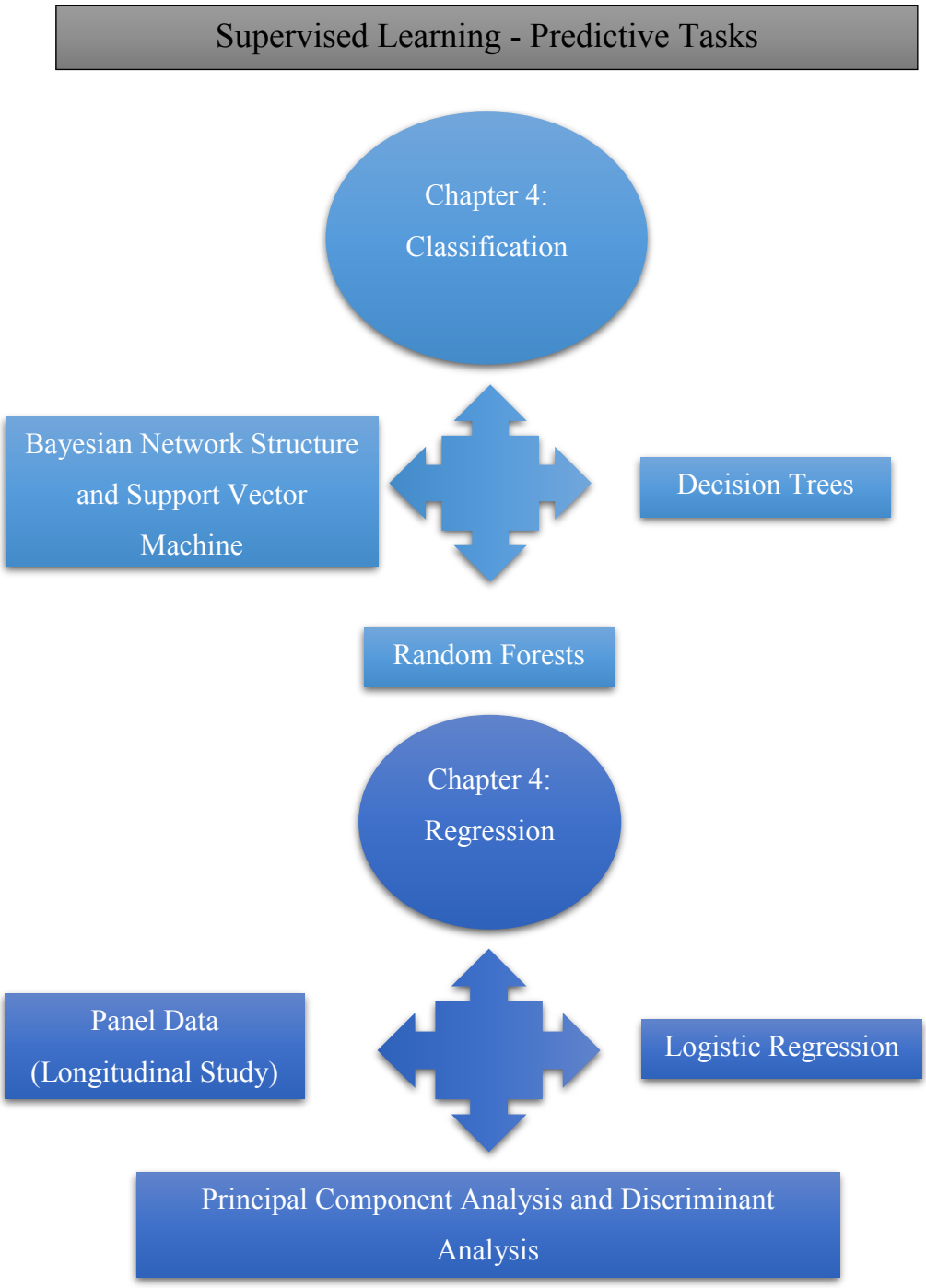


Figure 2.3. Supervised Learning Data Mining Tecniques used in the study

2.4.2 Supervised Learning

Supervised Learning is another data mining learning type that uncovers patterns and relationships between variables by using a labeled training dataset, meaning the dataset contains a known target variable (dependent or response variable). By building up statistical relationships between the response and independent variables, we produce statistical models and probabilistic graphical models to explain the variation in the response variable taking all possible independent variables into consideration. Its aim is to turn the dataset into real and actionable insights. Thus, when we have a target variable (response variable) to be predicted on the basis of a given set of explanatory variables, we need to concentrate on the data mining way that is so-called supervised learning.

2.4.2.1 Classification (CLA)

The first data mining technique under a supervised learning algorithm is known as classification. Classification is a DM function that assigns the variables in the dataset to target categories or classes. The main aim of classification is to accurately predict the target class for each sample in the dataset. For instance, a classification model, known as probabilistic graphical models can be designed to identify loan applicants as low, medium, or high credit risks. As an example, someone takes a look at weather conditions to take a walk. Here taking a walk becomes our target variable, meaning that is whether or not a person takes walk (0-person takes walk; 1-person does not walk). By taking the weather or other variables such as health of a person at the time of walk into consideration, the decision on taking a walk can be made.

There are many methods under the classification such as Decision Trees, Random Forests, Bayesian Network Structures, Neural Networks, and Support Vector Machines.

Now, let us summarize the supervised learning and unsupervised learning algorithm with a simple table;

Table 2.4 Supervised Learning & Unsupervised Learning

Key Differences	Supervised machine learning technique	Unsupervised machine learning technique
Logic	Input and output variables are given and models are formed by the known information.	Only input data is used to bring the hidden information into the open.Output data is the hidden information in the data sets.
Used Data	Algorithms are trained using data that is so called train data to derive models.	Algorithms are employed for the data sets that keep information hidden inside itself.
Algorithms	Decision Tree, Random Forest, Naive Bayesian Structure, Support Vector Machine Models are used for classification modeling. All regression modelings are utilized to derive models.	K-means, K-medoids, Fuzzy Clustering and Hierarchical Clustering Techniques are used for grouping the data points.Association Rule is utilized for determination of mostly sold or bought items together in market basket analysis.
Accuracy of Results	More reliable and accurate results	Less reliable and accurate results

2.4.2.2 Regression (RG)

The second DM technique under a supervised learning algorithm is known as regression. Regression is a way of statistical measurement utilized in every science that tries to identify and determine the strength of the relationships between the explanatory variables (independent variable) and the response variable (target variable, dependent variable). By means of regression techniques, we are aimed at building up some useful models that have the power of explaining the variation in the response variable. The main purpose of this technique is to help the readers as well as professionals to bring the existence of relationships between variables into the open and design models to be used for predicting future values of the response variable. For example, regression helps predict the future values of Gross Domestic Product (GDP) by taking the economic variables such as unemployment rate, inflation rate, employment rate etc. into consideration.

There are plenty of regression types such as logistic regression, generalized regression models, simple linear regression, passion regression, panel data, etc.

2.5 Treatment of Missing Cases

The dataset, as mentioned in the part of Introduction to data, is made up of 69 variables and it contains data relevant to the economy, health, sociology, and military for 80 countries that are ordered “Best Countries in 2019”. For each country, we have the data of 15 successive years. Therefore, with the simple multiplication, the dataset consists of almost 15*80*72 = 86400 observations. Of course, it is inevitable that we have missing values but this does not cause any problem on the way to mine the data. During the analysis, we are not worried about the existence of the missing cases. Because of the fact that we have the data of 15 successive years for each country, if we have missing cases in 2005, for example, we can put the values of either 2006 or 2004 into 2005. This kind of treatment of the missing case is just a simple example and is known as “Last Observed Variable

(LOV)". Notwithstanding the LOV, there are also other ways to treat missing cases such as "the Complete Case Method" that deletes all rows that include missing cases. Due to the reason of the loss of information attributable to this method, most of the time, readers or analyzers prefer keeping away from using this method. In software like R, there is a package called "mice" that enables the treatment of missing cases effectively. By comparison with the complete case method, we are confident that there is no serious information loss. Now, let us look at the types of missing cases. The dataset that is going to be used for clustering, classification, and regression includes some missing cases. Thus, before moving on to the analysis, we perform missing data analysis to make imputation. The types of missing cases are missing at random (MAR), Missing Completely at Random (MCAR), and Missing Not at Random (Non-ignorable/Non-response) (MNAR). We need to understand the missingness mechanism in our dataset. When we consider South Korea, since many observations are missing, we can be suspicious about the MNAR type missingness mechanism. Some other countries contain missing cases for different countries that display the existence of the type of MCAR because the missing values seem to be caused by the non-extraction for some years. Although there are some consecutive years that include missing cases, missingness seems to be attributable to not getting information for some years. In order to deal with missing cases and perform the data mining in an accurate way, MCAR is treated with the techniques of last Observed Variable and mean imputation. The reason why we use the mean imputation technique is because of the fact that we have fewer missing cases for some countries. If we had more missing cases and treated them with mean imputation, it would have led to a biased situation. Moreover, the countries that include a serious number of missing cases are removed from the dataset because there is no data and information about the variables.

2.5.1 Missing at Random (MAR)

Missing at random implies that the inclination for a data point to be missing is not relevant to the missing data, but it is related to some of the observed variables. There is a systematic association between the propensity of missing values and the observed data, but not the missing data. All of these definitions mean that missingness is random.

Only younger people have missing values for IQ. In that case, the probability of missing data on IQ is related to age. The assumption that the mechanism is MAR can not unfortunately be verified because it is not testable that the probability of missing data on a variable is only a function of other measured variables.

Table 2.5 MAR example

Complete Data		Incomplete Data	
25	120	25	MAR
26	125	26	
27	114	27	
28	112	28	
29	110	29	
45	90	45	90
48	100	48	100
52	92	52	92
65	85	65	85
75	111	75	111
90	112	90	112
age	IQ score	age	IQ score

2.5.2 Missing Completely at Random (MCAR)

Missing completely at random (MCAR) is the only missing data mechanism that can be confirmed by statistical tests. Missing data are MCAR when the probability of missing data on a variable is unrelated to other measured variables and is

unrelated to the variable with missing values itself. That means the missingness on the variable is completely unsystematic

Table 2.6 MCAR example

Complete Data		Incomplete Data	
25	120	25	120
26	125	26	125
27	114	27	MCAR
28	112	28	
29	110	29	
45	90	45	90
48	100	48	100
52	92	52	MCAR
65	85	65	85
75	111	75	MCAR
90	112	90	112
age	IQ score	age	IQ score

2.5.3 Missing Not at Random (Non-ignorable/Non-response) (MNAR)

Data are missing not at random (MNAR) when the missing values on a variable/attribute are related to the values of that variable (age) itself. For example, when data include missing cases on IQ level of persons and only the people with low IQ values have missing observations for IQ variable.

Table 2.7 MNAR example

Complete Data		Incomplete Data	
25	120	25	120
26	125	26	125
27	114	27	114
28	112	28	112

Table 2.7 (continued)

29	110	29	110
45	90	45	MNAR
48	100	48	100
52	92	52	MNAR
65	85	65	MNAR
75	111	75	111
90	112	90	112
age	IQ score	age	IQ score

Since the ways of handling missing data cases are out of the scope, here we are just giving the names of these ways in a table shown below;

Table 2.8 Treatment Techniques for Missing Cases

Deletion Methods	Listwise Deletion
	Pairwise Deletion
Imputation Methods	Mean Imputation
	Regression Imputation
	Matching Methods
	Last Observed Carried Forward
	Multilayer Perceptron
	Expectation Maximization (EM)
	Markov Chain Monte Carlo (MCMC)

2.6 The Research Topic Overview

In this research study, we are mainly focusing on the economic data in to bring the hidden information behind the economic growth into the open. Besides the economic data, it is expected that the concentration and thinking on the data of health, sociology, and military provide some undiscovered knowledge. As a whole, we keep track of important clues about;

- The Factors (Health, Sociology, Military, and others) Behind Economic Growth,
- The Most Important Variables That Play a Direct Role in The Economic Growth,
- The Role of All Economic Indicators in Determination of the Countries That Have High Economic Level.
- The Role of Imports and Exports in Economy,
- The Most Frequently Exported Items from a Country to Other Countries,
- The Statistical and Graphical Economic Models, Health Models, Sociology Models, and Military Models.

END OF DATA MINING – EXPLORATION OF THE DATA

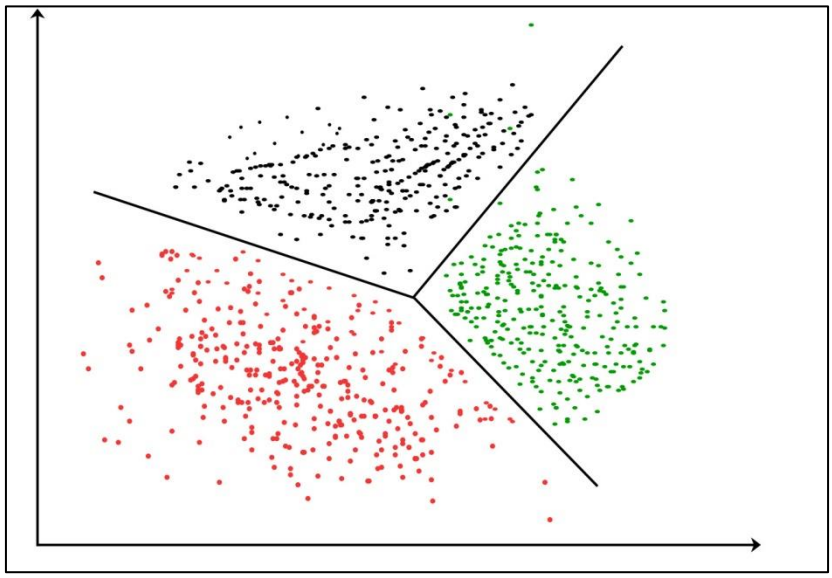
End of Chapter 2

CHAPTER 3

APPLICATIONS OF UNSUPERVISED DATA MINING ON DATASET

3.1 Clustering

Clustering is the first and most used one of the data mining techniques to describe data under unsupervised learning. The reason why clustering is utilized for the sake of describing the data comes from the fact that its aim is to make inferences on the data by separating variables from each other and search for the variables that are closest to each other. Clustering is the most very well-known technique to distinguish the data points from each other by the similarity indexes.



(Chatterjee, 2019)

Figure 3.1. Clustering of the data points

Over the last decades, the increasing interest in the knowledge discovery process paves the way for the new methods to better understand the data to discover the hidden knowledge. What fascinates the researches in figuring out the new information comes from the interest in finding ways of how grouping similar items

are performed. It is completely the determination or process of organizing data points in such a way that objects within a group must be similar to each other and different from other groups, namely the greater similarity within a cluster, the greater difference between groups. In order to implement the clustering algorithm, we apply for some mathematical distance formulations. Using the well-known distance algorithms, users are trying to search for the best one to diminish the damage that is attributed to the wrong selection of distance algorithms. Thus, the clustering algorithm has types according to the procedures or steps to obtain the best clusters.

3.1.1 Data Preparation – Data Pre-Processing

In order to perform cluster analysis and research in R, it is needed to perform some regulations on the data.

1. **Rows that are individuals and columns that are attributes must be in the form of numeric format.**

We perform a clustering analysis on the chosen variables that are stated in the previous section. Those variables are numeric.

2. **Any missing case must be treated or removed from the data.**

In our datasets, there is no available dataset for the countries of North Korean, Qatar, Myanmar and Iraq. For the countries of Qatar, Myanmar, and Iraq, we fill the missing cases by the method of last observed variable. Because there is no even data for North Korean, we remove this country from the cluster analysis.

3. **Data must be treated in a way that clustering analysis can be implemented suitably and can provide well-designed visualizations.**

The related dataset is made up of information that is coming from successive years from 2000 to 2015. Each country has a dataset of 15 years. During the experiments, we have realized that we cannot get comprehensive and suitable graphs and results

in case we take the year effect into consideration. Therefore, we calculated an average of 15 years' datasets for each country.

4. Standardization of the datasets

The most important step in data preparation for clustering is the standardization of the datasets to provide comparability of the variables with each other. Data points consist of variables whose means are equal to zero and whose standard deviations are eqal to one.

5. Treatment of class of the dataset

The last step before passing to cluster analysis in R is to make the format of data into a matrix class. Our dataset includes a factor variable that shows the level of income of countries by the variable of GNI. We add this variable to analysis in order to capture the information about whether or not the grouping of countries by this variable provides insight into the contributions to clustering.

Table 3.1 Standardized Average Numeric Dataset-1

	GNI by Atlas Method	GNI per capita by Atlas Method	GNI by PPP	GNI per capita by PPP	GNI	GDP	GDP per capita	GDP by PPP	GDP per capita by PPP
	GNI-A	GNI-A-PC	GNI-PPP	GNI-PPP-PC	GNI	GDP	GDP-PC	GDP-PPP	GDP-PPP-PC
MAR	-0.3299	-0.8296	-0.3593	-0.9218	-0.3327	-0.3361	-0.8185	-0.3624	-0.9014
PER	-0.3112	-0.7517	-0.3331	-0.7856	-0.3128	-0.3129	-0.7319	-0.3305	-0.7533
AGO	-0.3395	-0.8223	-0.3942	-0.9619	-0.3401	-0.3407	-0.7901	-0.3930	-0.9161
ARE	-0.2356	1.0417	-0.2322	2.6218	-0.2349	-0.2390	0.9765	-0.2379	2.3839
ARG	-0.1928	-0.5291	-0.1421	-0.3933	-0.1886	-0.1844	-0.5079	-0.1332	-0.3775
AUS	0.1201	1.2183	-0.0733	0.6732	0.1320	0.1514	1.2428	-0.0628	0.6583
AZE	-0.3545	-0.7636	-0.4008	-0.6538	-0.3564	-0.3592	-0.7359	-0.4018	-0.6113
BEL	-0.1355	1.1661	-0.2507	0.7721	-0.1377	-0.1425	1.0342	-0.2574	0.6605
……	……	……	……	……	……	……	……	……	……
PPP = Power Purchasing Rate									

The above table is our dataset-1 which is composed of standardized numeric data. For example, say that GNI-A, which stands for the GNI by Atlas Method, for the

country of AUS is 0.1201. This number is the average of 15 years standardized numeric data of the country of AUS under this variable.

Table 3.2 Unstandardized Average Percentage Dataset-2

	GNI per capita growth(% annually)	**GNI growth(% annually)**	**GDP growth(% annually)**	**GDP per capita growth(% annually)**
	GNI.PC.G	**GNI.G**	**GDP.G**	**GDP.PC.G**
MAR	3.1576	4.4680	4.4681	3.1577
PER	4.0314	5.0454	5.1708	4.1560
AGO	5.1421	8.9298	7.1922	3.4795
ARE	-2.9945	4.2749	4.7758	-2.2917
ARG	1.7657	2.8366	2.7216	1.6518
AUS	1.5903	3.0595	3.0185	1.5509
AZE	10.6752	11.9845	10.8039	9.5052
BEL	1.0395	1.6560	1.5742	0.9576
……	……	……	……	……

The above table is our dataset-2 which is made up of ustandardized percentage data. The reason why we do not need to perform standardization for this dataset is that there are no scale differences between the variables. All of them are percentages. For example, the variable GNI.PC.G which stands for GNI per capital growth is 1.59 for the country of AUS. Note that again this is the average of the percentages of 15 successive years of the variables/attributes.

Note that: Although we used many clustering algorithms, here only the results of the k-medoids algorithm are given because, in the classification part of the study, the response variables attributed to k-medoids give the highest accuracy value. That indicates that groups of the countries are determined better by means of the k-medoids algorithm as compared to the other clustering methods.

3.1.2 K-medoids Algorithm on chosen variables

K-Medoids Clustering (Partitioning around Medoid (Representative Points of each cluster))

A medoid can be defined as a point in the cluster whose distances with all the other points in the cluster are minimum. Each cluster is represented by one of the data points in the clusters and they are called cluster medoids. The terminology of the medoid points to an object included in a cluster in which the average distance between it and all the other members of the clusters is minimal. Medoid indicates the most centrally located point in the subgroup.

Those points correspond to arithmetic means that are calculated in the k-means algorithm. The most important difference between the k-means and the k-medoids algorithm is attributed to the fact that the k-means algorithm works with the means calculated at each iteration of the algorithm. However, it is known that the mean term is mostly affected by the outliers by comparison with the other descriptive measurements such as median, quartiles, etc. On the contrary, the k-medoids algorithm uses randomly selected data points as cluster centers and because there is no mean calculation during the iteration of the k-medoids, it is understood that k-medoids algorithm is more robust to outlier as compared to the k-means algorithm. This is the basic disadvantage of the k-means algorithm and one can choose this algorithm to carry out clustering to avoid the negative effects of outliers because the usage of the k-means algorithm while keeping outliers in dataset lay the ground for clustering the outliers as different groups, and this situation can mislead readers about the interpretations of the data points that are classified in a distinct cluster.

K-medoids Algorithm Steps;

1. Searching for “k” representative objects or medoids among the data points among the dataset.
2. Assignment of each observation to the nearest medoid and setting up clusters by means of mathematical distance formulations.
3. Determination of the objective function which is the summary of minimum dissimilarities of all data points to their nearest medoid.

4. Calculation of the arithmetic means is not implemented to determine new cluster centers.
5. Repetition of 2 and 3 steps to acquire clusters whose objective function is smaller among all objective functions obtained by utilizing different medoids.

Note: *Euclidean and Manhattan distance are mostly used measures to calculate the dissimilarity between the data points. However, in the literature, it is highly recommended to use the latter because Manhattan distance is more robust to the outliers as compared to the Euclidean Distance. (Penn State Science, 2018).Although Manhattan distance is preferred to Euclidean Distance, because of the problems that we came across during the analysis in case of using manhattan distance, we use Euclidean Distance Measurement.*

As a result, in k-medoids clustering;

- It works better in case data include outlier points.
- Centroids are known as medoids also called representative data points.
- Manhattan distance is recommended to be used.

K-Medoids Clustering based on the optimal number of clusters

First of all, let us define the methods to choose the optimal number of clusters, and then we perform clustering analysis on the scaled numeric dataset-1 and percentage dataset-2.

Elbow Method (EM): The **EM** is an experiential method of making comments and validation of consistency within-cluster analysis developed to help figure out how many clusters to be chosen for the sake of separating objects from each other. As compared to the other methods, the findings of EM are less reliable and can give rise to uncertainty in the number of clusters. The main function that is used for calculating the "elbow" point in EM is directly calculated by **within sum of squares** of the clusters. The Elbow method looks at the total WSS as a function of the number of clusters. One should select several clusters so that adding another cluster to all processes does not improve much better than the total WSS.

The algorithm of the Elbow Method works simply as following;

- Calculation clustering algorithm (k-means clustering or another clustering techniques) for the different values of k.
- For each k, the total within-cluster sum of square (WSS) is computed.
- Draw the plot of wss by the different number of clusters.
- Determination of bend point, which is called “knee” or “elbow”.
- Bend point is thought of as the best number of clusters.

Average Silhouette Method (ASM): The ASM approach gauges the quality of a clustering. Its function determines how well each data point lies within its own cluster. A high average silhouette method calculates the average silhouette of objects for the different number of clusters k. The optimal number of clusters is the one that maximizes the average silhouette over a range of possible values k.

The algorithm of the Average Silhouette Method works simply as following;

- Calculation clustering algorithm (k-means clustering or another clustering techniques) for the different values of k.
- For each k, we compute the average silhouette of the observations. (avh.sil).
- Draw the plot of avg.sil by the different number of clusters.
- The location of the maximum points is thought of as suitable number of clusters.

Gap Statistics Method (GSM): The gap statistic draws comparisons of the total within-cluster variation for different values of k with their expected value under determined a null reference distribution of data. The desired optimal number of clusters is a value that gives rise to maximation of the gap statistics, meaning the largest gap statistics provide us with the best number of clusters.

The algorithm of the Gap Statistic Method works simply as following;

- Clustering the observed data, for the different number of clusters

$$k = 1, \dots \dots \dots . k_{max}$$

And calculate the corresponding total within-cluster variation W_k.

- Produce B reference data sets with a random uniform distribution. Clustering of these reference datasets with different number of clusters.

$$k = 1, \dots \dots \dots . k_{max}$$

And calculate the corresponding total within-cluster variation W_{kb}.

- Compute the estimated **gap statistic** as the deviation of the observed W_k value from its expected value W_{kb} under the null hypothesis:

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} \log(W_{kb}*) - \log(W_k)$$

Compute the standard deviation of the statistics.

- Lastly, select the number of clusters as the smallest value of k such that the gap statistic is within one standard deviation of the gap at

$$k + 1: Gap(k) \geq Gap(k1) - s_{k+1}$$

Note: Using B = 500 is suggested to get quite precise results because according to the literature overview, the "gap plot" stays unchanged after another run.

R package – NbClust – Method: During the literature overview, we run across a package that is called **NbClust.**

This package provides 30 indices to determine the optimal number of clusters and suggests the best clustering scheme from the different outputs acquired by assessing all combinations of the number of clusters, distance measures, and clustering methods.

Note: The indexes that come up after implementing the NbClust Method to be shown during the assessment of the best number of clusters. The optimal number of clusters that are obtained by means of NbClust are used for hiearchical clustering. The results of hierarchical clustering are shown in Appendix A.

Table 3.3 Summary of Optimal Number of Clusters for dataset-1&2 by K-Medoids Clustering

Methods	Optimal Number of Clusters for dataset-1	Optimal Number of Clusters for dataset-2
Elbow Method	**3,5**	**3,5**
Average Silhouette Method	**3**	**2**
Gap Statistic Method	**2**	**1**

Now, let us look at the visualization of clusters and silhouette plots under the k medoids algorithm for the dataset-1 when k = 5 that forms the best clustering scheme.

For dataset-1

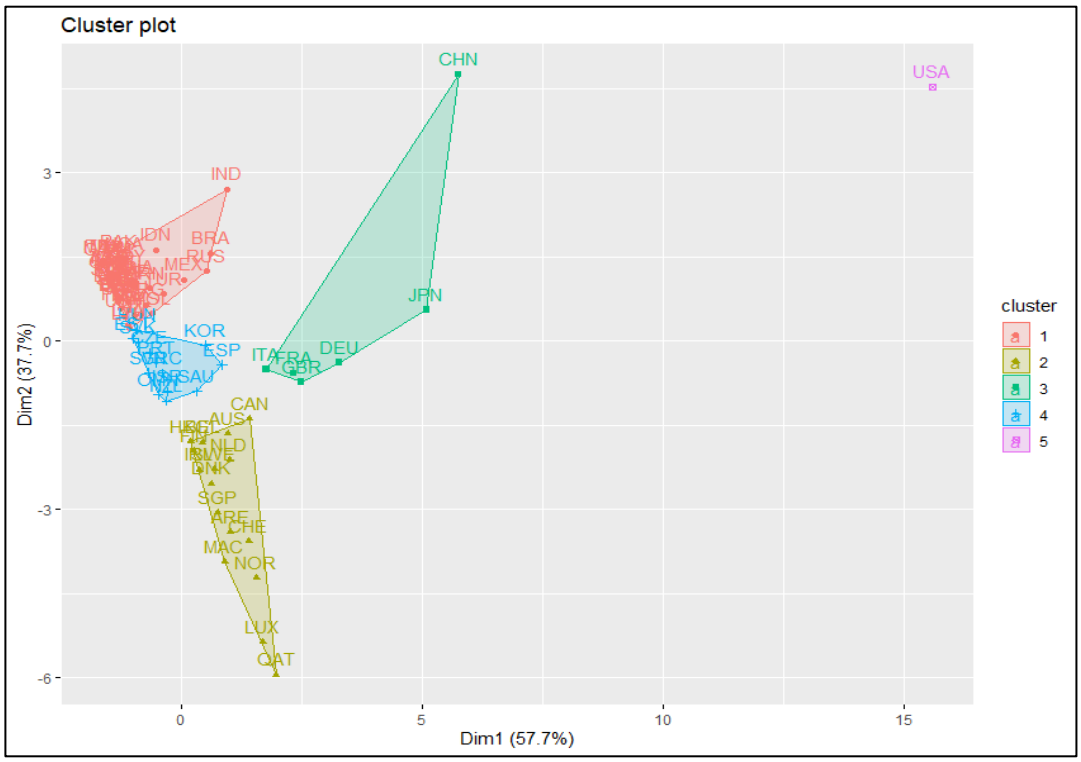


Figure 3.2 K-Medoids for k=5 for the scaled-numeric dataset-1

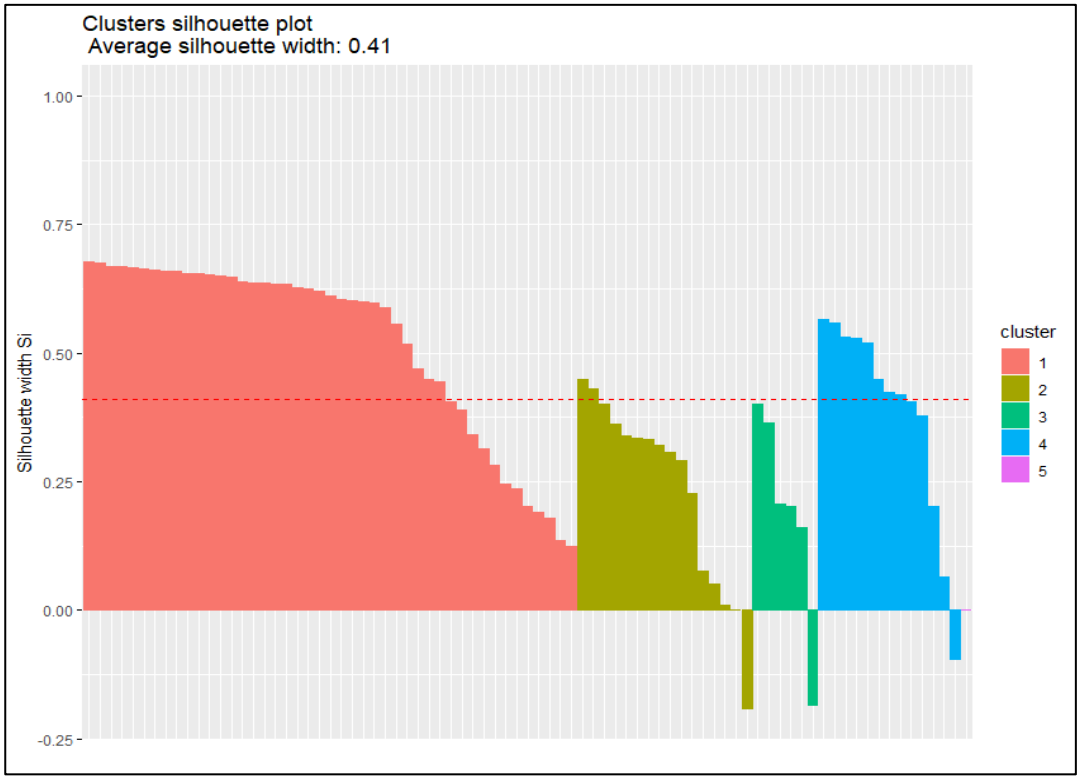


Figure 3.3 Silhouette Plot -k-medoids - k=5, dataset-1

To sum up, we can suggest the k-medoids algorithm for the dataset-1 with the optimal number of clusters that are k=3, 4, 5 respectively. Whilst the graphics that are drawn by k=3 and k=4 give similar results with one another, the cluster plot presented when k=5 provides us with a logical clustering, as well. In order to decide on the best number of clusters, we can get benefit from the silhouette plots and we can view the number of observations whose silhouette scores are negative under the clusters designed for k=3, 4, 5, respectively.

Assessment of Silhouette Plot resulted from the clustering of dataset-1

Silhouette plots provide us the information about the goodness of the clustering mechanism. For the dataset-1, k=3, 4, and k=5 can be preferable. All of these clusters designed for k=3, 4, and 5 can be used for separating the countries from each other. When k=3 and k=5, we have less number of countries whose silhouette numbers are smaller than 0. However, we do not have more observations under the

clustering when k=4. We have just one more observation that is clustered wrongly as compared to clusters built up when k=3 and 5. Therefore, for me, I can take k=5, because we know that CHN and JPN behave similar economic patters. Moreover, by using the k-medoids algorithm we can know for sure that the outliers' effects are minimized. Overall, k=3, k=4, and k=5 can be chosen and preferred. We are of the opinion that the optimal number of clusters (k) can be taken as 5 to draw meaningful results.

When k=5; the countries that are wrongly clustered based on k-medoids are;

	cluster	neighbor	sil_width
CAN	2	4	-0.0003795182
HKG	2	4	-0.1912994583
ITA	3	4	-0.1842597820
HUN	4	1	-0.0952302599

Therefore, for the dataset-1 that is made up of numeric values, we obtain a better an d desirable number of groups of the countries when k is chosen as 5. Now, let us lo ok at the visualization of clusters and silhouette plots under the k medoids algorith m for the dataset-2 when k = 3 that forms the best clustering scheme.

For dataset-2

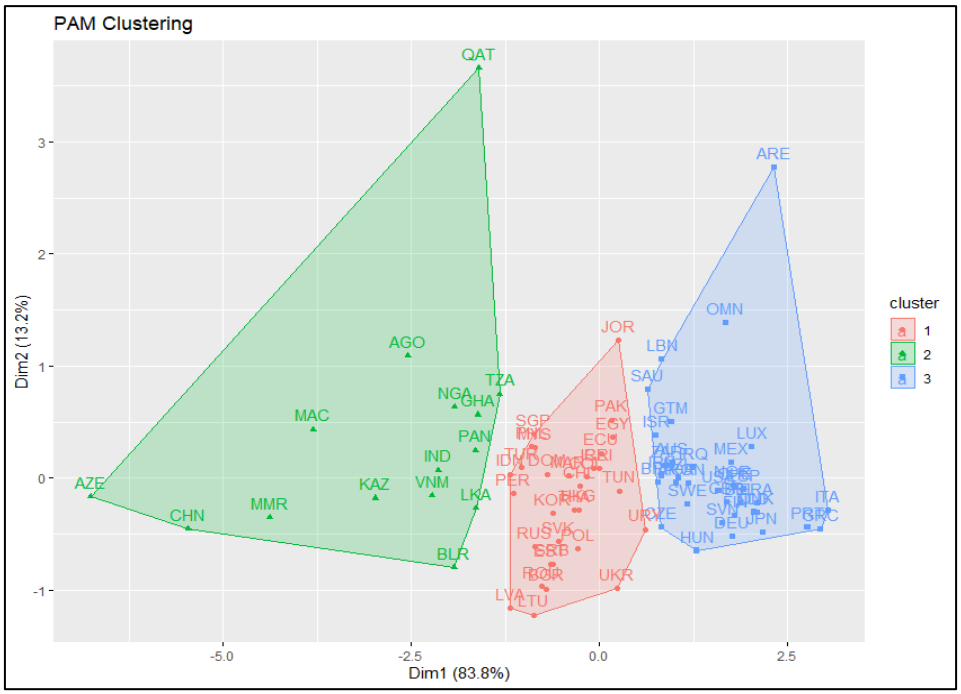


Figure 3.4 K-Medoids for k=3 for the percentage dataset-2

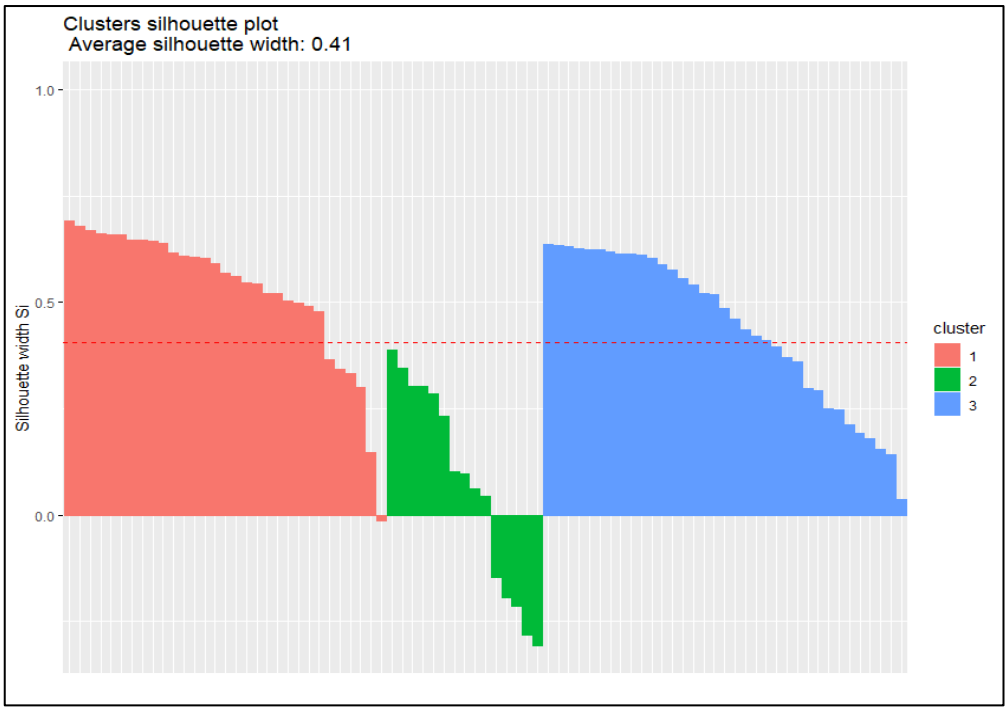


Figure 3.5 Silhouette Plot -k-medoids - k=3, dataset-2

In conclusion, we can see that when k=3, 4, and 5, we can get good separation of countries by courtesy of k-medoids method. In order to choose one among these numbers of clusters, we looked at the assessments of silhouette plots for k=3, 4, and 5 and we decide to take k=3 as the best and optimal number of clusters.

Assessment of Silhouette Plots resulted from clustering of dataset-2

When k=3; the countries that are wrongly clustered based on k-medoids are;

	cluster	neighbor	sil_width
URY	1	3	-0.01332717
GHA	2	1	-0.14572045
PAN	2	1	-0.19405255
BLR	2	1	-0.21424203
TZA	2	1	-0.28301216
LKA	2	1	-0.30711631

When k=3, we can see that 5 of the countries placed on the cluster 2 seem to be clustered wrongly and it is seen also that those countries' neighbor cluster is 1, meaning cluster 1 may be more suitable for these countries. Totally, 6 countries are separated wrongly because each country's sil_width is smaller than 0. Moreover,

the number of countries that are wrongly clustered is less than the case when k=4 and k=5. Thus, it is suitable to choose k=3 as the best and optimal number of clusters for the dataset-2. In the following conclusion part, we show the groups of the countries in the case when k=3,4 and 5 for the dataset-1 and when k=3 and k=4 for the dataset-2.

3.1.3 Conclusion and Discussion

K-Medoids Algorithm: **The conclusions are made on the basis of chosen countries.** Possible Clusters coming from k-medoids algorithm are summarized at the following table;

Table 3.4 K-Medoids Clusters for dataset-1&2

Countries	Scaled-Numeric Dataset-1			Percentage Dataset-2	
	k=3	k=4	k=5	k=3	k=4
MAR	1	1	1	1	1
PER	1	1	1	1	1
AGO	1	1	1	2	2
ARE	2	2	2	3	3
ARG	1	1	1	3	3
AUS	2	2	2	3	3
AZE	1	1	1	2	2
BEL	2	2	2	3	4
BGR	1	1	1	1	1
BLR	1	1	1	2	2
BRA	1	1	1	3	3
CAN	2	2	2	3	3
CHE	2	2	2	3	4
CHL	1	1	1	1	1
CHN	1	1	3	2	2
COL	1	1	1	1	1
CRI	1	1	1	1	1
CZE	1	3	4	3	3
DEU	2	2	3	3	4
DNK	2	2	2	3	4
DOM	1	1	1	1	1
ECU	1	1	1	1	1
EGY	1	1	1	1	3
ESP	2	3	4	3	4
EST	1	3	4	1	1
FIN	2	2	2	3	4
FRA	2	2	3	3	4
GBR	2	2	3	3	4
GHA	1	1	1	2	2
GRC	1	3	4	3	4

Table 3.4 (continued)

GTM	1	1	1	3	3
HKG	2	2	2	1	1
HUN	1	3	4	3	4
IDN	1	1	1	1	1
IND	1	1	1	2	2
IRL	2	2	2	1	1
IRN	1	1	1	3	3
IRQ	1	1	1	3	3
ISR	2	3	4	3	3
ITA	2	2	3	3	4
JOR	1	1	1	1	3
JPN	2	2	3	3	4
KAZ	1	1	1	2	2
KOR	1	3	4	1	1
LBN	1	1	1	3	3
LKA	1	1	1	2	2
LTU	1	1	1	1	1
LUX	2	2	2	3	4
LVA	1	1	1	1	1
MAC	2	2	2	2	2
MEX	1	1	1	3	4
MMR	1	1	1	2	2
MYS	1	1	1	1	1
NGA	1	1	1	2	2
NLD	2	2	2	3	4
NOR	2	2	2	3	4
NZL	2	3	4	3	3
OMN	1	3	4	3	3
PAK	1	1	1	1	3
PAN	1	1	1	2	2
PHL	1	1	1	1	1
POL	1	1	1	1	1
PRT	1	3	4	3	4
QAT	2	2	2	2	2
ROU	1	1	1	1	1
RUS	1	1	1	1	1
SAU	2	3	4	3	3
SGP	2	2	2	1	1
SRB	1	1	1	1	1
SVK	1	3	4	1	1
SVN	1	3	4	3	4
SWE	2	2	2	3	3
THA	1	1	1	1	1
TUN	1	1	1	1	3
TUR	1	1	1	1	1
TZA	1	1	1	2	2
UKR	1	1	1	1	1
URY	1	1	1	1	3
USA	3	4	5	3	4
VNM	1	1	1	2	2
ZAF	1	1	1	3	3

Note that:

People that want to take a look at groups of countries deeply must pay attention to COLOR not numbers because in the part of programming, for example, when k=4 USA is clustered in C3, however, when k=5 USA is clustered in C2. Groups of the countries are represented with the same color.

K-medoids algorithm results that are coming from the investigation of scaled-numeric dataset-1 propose us to divide the scaled-numeric dataset-1 into **3, 4 or 5 clusters**. The number of clusters of dataset-1 is achieved by giving a reasonable and logical number and by different methods and available packages in R that are improved to find the optimal number of clusters.

It is clear to say that most of the countries are grouped together even if k=3, 4, and 5. When k is increased from 3 to 4 then 5, we can see that some countries such as CZE, ESP, and ISR that are grouped in the same cluster are again being clustered in the same group. This may indicate that the closeness and association between these countries when k=2 is also much greater than that the other countries. Moreover, for the other countries such as HUN, GRC and EST are clustered in 1. When k=3, they are all together moving on to clusters 3 and 4 when k=4 and 5, respectively. Those are the main findings coming from the k-medoids algorithm. We put here the most logical number of clusters. We can arrange the clusters by those numbers for scaled-numeric dataset-1.

The other countries such as TUR, RUS, and BRA placed in the same group are again grouped in the same circle even if there is an increment from k=3 to k=4 and 5. This is a sign that the countries that are grouped together throughout the number of clusters that are 3, 4, and 5 are considered as having similar economic patterns. The USA comes to the front as a country separated always as a single-point cluster. The reason why the USA is not clustered with the other countries is because the related values of USA are excessive and behaving as outliers. Since we know from the fact that the k-medoids algorithm is not affected by excessive values, the USA accounts for its own single-point cluster.

K-medoids algorithm results that are coming from the investigation of percentage dataset-2 propose us to divide the percentage dataset-2 into either **3 or 4 clusters**. The number of clusters of dataset-2 is achieved by giving a reasonable and logical number and by different methods and available packages in R that are improved to find the optimal number of clusters.

As it is seen clearly from the result of clustering dataset-2, when k=3 and 4, almost quarter number of the countries are changing their clusters and passing on to the new clusters indicating the neediness of possessing one more cluster. Thus, k=3 and k=4 can be a preferable number of the clusters to make inferences on the groups of the countries showing distinct economic growth patterns. TUR and RUS are seen in the same group even if k goes from 3 to 4 just the same as the USA and JPN are displaying the same movement by existing in the same clusters when k is moving from 3 to 4. Because we have less number of the countries that are wrongly clustered when k=3, the optimal and best number of the groups can be taken as k=3.

The results of k-means, fuzzy, and hierarchical clustering algorithm are added to Appendix A in order for the reader to observe the clusters under different techniques.

END OF DATA MINING-CLUSTERING

3.2 Association Rule Mining (ASM)

Association Rule Mining (ASM) is one of the ways to deal with figuring out the events that happen frequently together existing in datasets. The main aim is to find the objects that occur together frequently by comparison with the other groups of items. The data sets are made up of transactions that form a group of items in market basket analysis.

For example, people who buy milk are more likely to buy eggs. It can be explained in such a way that if someone buys milk, then he/she can also buy eggs. It is important to note that there is just a probability of buying eggs. There is no absolute knowledge on the certainty of the events. In the following sections, we give a clear example of this rule by showing also how the apriori algorithm is working.

AR data mining is mostly used in the market basket analysis. Market Basket Analysis is one of the key techniques that is utilized to come plenty of relations into existence. It enables us to identify and detect relationships between the items or products that people buy together frequently. The best and known algorithm is called “Apriori Algorithm (AA)”.

3.2.1 Apriori Algorithm (AA)

Apriori Algorithm is the most used method and classical algorithm for ones who are interested in AR data mining. It is employed for bringing frequent item sets and relevant association rules into the open. It maps out to operate on a database including a lot of transactions coming from items or products that are bought by customers in a market. The thinking on the way to discover itemsets that are sold and bought together becomes a very important strategy for market owners to make a research on what items are taken interest together during shopping of customer at a given interval and to increase the profits and to earn more money.

Therefore, this way of mining data is a crucial and effective step for Market Basket Analysis.

This algorithm helps customers reach items and purchase their items with an easer way and also this algorithm helps owners of the market make a planned layout of their items. This leads us to increment in the sales of markets. AA is used for many purposes to bring the associations between items into the open. It is employed in the field of healthcare to detect adverse drug reactions by taking all patients records into account. The rules originating from the analysis of healthcare provide researchers with information about what combinations of medications and patient characteristics bring about specific diseases. In order to better understand this algorithm step by step, let us look at an example ***that is taken from Professor Anita Wasilewska Lecture Notes.***

Table 3.5 Transaction IDs and Products/Items

TID	Item IDS
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Before performing AA algorithm, first of all, we need to specify and give some threshold values to the algorithm. These threshold values are determined at the beginning to make an accord on the number of association rules. This can be changed by the aims of the study. It can show changes by the analysts' point of view.Minimum Support and Minimum Confidence are determined as **equal to 2** and **70%, respectively in the example by Wasilewska.** Instead of support, one can give support rates that are the percentage of frequency of each item in the list of transactions.

Step 1: Finding of candidate and frequent item sets.

Table 3.6 Candidate Itemsets and Frequent Itemsets with 1 item

Candidate Items (C1)			**Frequent Items (L1)**		
>>Scan TID for count of each candidate	Itemset	Sup.Count	>>Compare candidate support count with minimum support count	Itemset	Sup.Count
	{I1}	6		{I1}	6
	{I2}	7		{I2}	7
	{I3}	6		{I3}	6
	{I4}	2		{I4}	2
	{I5}	2		{I5}	2

The set of frequent itemsets1, called L1, is made up of the candidate itemsets (C1) that are satisfying the rule of being greater than the threshold value of the support that is 2.

Step 2: Finding of candidate and frequent item sets with 2 items

Table 3.7 Candidate Itemsets and Frequent Itemsets with 2 items

Generation of C2		**Candidate Items (C2)**			**Frequent Items (L2)**		
>>Generation of candidate itemset 2 by taking the union of L1 and L1	Itemset	>>Scan TID for count of each candidate	Itemset	Sup.Count	>>Compare candidate support count with minimum support count	Itemset	Sup.Count
	{I1, I2}		{I1, I2}	4		{I1, I2}	4
	{I1, I3}		{I1, I3}	4		{I1, I3}	4
	{I1, I4}		{I1, I4}	1		{I1, I5}	2
	{I1, I5}		{I1, I5}	2		{I2, I3}	4
	{I2, I3}		{I2, I3}	4		{I2, I4}	2
	{I2, I4}		{I2, I4}	2		{I2, I5}	2
	{I2, I5}		{I2, I5}	2		Support of other item sets are smaller than 2. They are eliminated.	
	{I3, I4}		{I3, I4}	0			
	{I3, I5}		{I3, I5}	1			
	{I4, I5}		{I4, I5}	0			

Step 3: Finding of candidate and frequent item sets with 3 items

$$\boldsymbol{C3 = L2 \cup L2}$$
$$= \{\{\boldsymbol{I1, I2, I3}\}, \{\boldsymbol{I1, I2, I5}\}, \{\boldsymbol{I1, I3, I5}\}, \{\boldsymbol{I2, I3, I4}\}, \{\boldsymbol{I2, I3, I5}\}, \{\boldsymbol{I2, I4, I5}\}\}$$

In this step, in order to find the C3 and L3, we can simply take a look at whether or not all subset of the item sets with 3 items are included in L2. C3 is made up of the item sets with 3 items in which all subsets of itemsets with 3 items are included in the L2. Thus, C3 is made up of $\{\boldsymbol{I1, I2, I3}\}$ ***and*** $\{\boldsymbol{I1, I2, I5}\}$ **because all subsets of these itemsets are available in the L2. The others are removed from C3.**

Table 3.8 Candidate Itemsets and Frequent Itemsets with 3 items

Candidate Items (C3)			**Frequent Items (L3)**		
>>Scan TID for count of each candidate	Itemset	Sup.Count	>>Compare candidate support count with minimum support count	Itemset	Sup.Count
	{I1, I2, I3}	2		{I1, I2, I3}	2
	{I1, I2, I5}	2		{I1, I2, I5}	2

Step 4: Finding of candidate and frequent item sets with 4 items

The next candidate itemsets include 4 items. From the union of L3 and L3, we obtain the item set which is made up of $\{\boldsymbol{I1, I2, I3, I5}\}$**.**

When we look at the support of this item set, it is equal to 1, then, we could not take this as frequent itemset. Thus, the item sets that are going to be used are made up of L1, L2, and L3.

Step 5: Building up Association Rules

Up to now, we have only used support count as a threshold value to decide on which itemsets to be taken into consideration to produce rules. After forming itemsets, we set up Association Rules based on all item sets that are placed on L1, L2, and L3. In this stage, we are going to use the minimum confidence threshold value as criteria for the assessment of the possibility of the rule.

For example, from L3,

$$\boldsymbol{R1: I1\ and\ I2 \rightarrow I5}$$

$$Confidence = \frac{Support\ Count\ \{I1, I2, I5\}}{Support\ Count\ \{I1, I2\}} = \frac{2}{4} = 50\%$$

At the beginning of the example, we determined 70% as threshold confidence value. The confidence of 1st rule is smaller than 70%. Thus, we reject the above rule 1.

$$\boldsymbol{R2: I1\ and\ I5 \rightarrow I2}$$

$$Confidence = \frac{Support\ Count\ \{I1, I2, I5\}}{Support\ Count\ \{I1, I5\}} = \frac{2}{2} = 100\%$$

The confidence of 2nd rule is greater than 70%. Thus, we accept the above rule 2.

$$\boldsymbol{R3: I2\ and\ I5 \rightarrow I1}$$

$$Confidence = \frac{Support\ Count\ \{I1, I2, I5\}}{Support\ Count\ \{I2, I5\}} = \frac{2}{2} = 100\%$$

The confidence of 3rd rule is greater than 70%. Thus, we accept the above rule 2.

R2 and R3 are coming from L3. The algorithm assesses all the L1, L2 and L3 to derive association rules. So for example, what does R2 mean?

Say that I1 is bread, I2 is yogurt and I5 is egg. This rule says that

- If people buy bread and egg together, then, they also have tendecy to buy yogurt with confidence 100%.

3.2.2 Data Preparation – Data Pre-Processing

Datasets that are going to be utilized are composed of exports of Switzerland to 99 Countries throughout the years 2001, 2009, and 2018. The reason why Switzerland is chosen as a country to be investigated is attributable to the research that is performed by U.S.News in 2019. *(U.S.News, 2019)*. The selection procedure

of the best country is performed under 9 distinct categories and then the countries are scored. The indexes that are taken into account to choose the best one are Adventure, Citizenship, Cultural Influence, Entrepreneurship, Heritage, Movers, Open for Business, Power, and Quality of Life. We take all the countries that have trade policies with Switzerland. Switzerland's major exports are machinery and equipment, chemical-pharmaceautical products, watches and textiles, and apparel. Raw materials, food, vegetable oils, and fuel account for almost one-quarter of total imports and they are transported by rail, truck, and barge. In order to apply the association rule mining to databases, we need to have appropriate forms of datasets that are suitable for mining. AR deals with the datasets that are composed of transactions. For example, a person goes to market in Thursdays only once and buys milk, bread, and cigarette. This makes one transaction in a basket and this transaction record contains 3 goods. If the same person goes to market twice on the same day, this forms a second transaction that includes the same or different items. This is the most important aspect of designing datasets for association rule mining. Therefore, if this situation can be thought in an excel table, rows are corresponding to the transactions and each transaction can be called as times that customer goes to the shopping center and buys something. Therefore, the same customer can lead to two different transactions on the condition that times of shopping are different. What we mean by saying times is that time can be a time of day, a day of the week, a week of the month, a month of the year.

Table 3.9 Typical Transactions

Customers	**Transaction**	**Items Sold by Market**
Customer 1	TDI-1-time(1)	A,B,C,D
Customer 1	TDI-2-time(2)	A,B,D,E
Customer 2	TDI-3-time(1)	D,C,A,B
Customer 3	TDI-4-time(2)	A,B,C,U

Table 3.9 (continued)

Customer 4	TDI-5-time(2)	A,D,C,E

From Table 3.9, there are five transactions. It can be seen that TD-1 and TD-2 are formed by the same customer in different times. In time (2), Customer-1, Customer-3, and Customer-4 are shopping together and form different transactions. Therefore, transactions only are designed depending on different times and different customers. This condition is sometimes known as "on the same visit" instead of "at the same time". Databases that we are going to use are made up of transactions of exports of Switzerland to 99 countries throughout the years of 2001, 2009 and 2018. For this study, 99 countries that are importing product from Switzerland account for the transactional data, and we build up association rules based on the transactions of exported products of three years (2001, 2009 and 2018). Thus, in our first dataset by which time is the year of 2018, we have 99 transactions attributed to 99 different countries' imports from Switzerland. A quick appearance of the data sets can be viewed in Appendix B: Association Rule.

3.2.3 Frequency Table of Products/Product Groups on given Data Set based on Exports of Switzerland to 99 Countries.

Analysis of frequency table is mostly used method to analyze the association between transactions. By means of assessing the frequency table analysis, we can gain insight into the most frequently exported products prior to passing to the Apriori Algorithm.

Exported Products of the Year of 2018

First of all, let us look at the frequency of items in each Mostly Bought/Exported Items/Products Group (MBIG) for the year of 2018. Suppose that we want to examine the frequencies/supports rate, of products that are either exported lonely or together with other products. For that reason, say that we want to see exported groups of products or products whose support rates are greater than

0.1, meaning products or product groups must be in the list of transactions at least 10 out of the recorded 99 transactions in MBIGs.

Table 3.10 Support Rates/Counts by MBIGs when min.support = 0.1 & Year 2018

Minimum Support = 0.1			
No	**items/products/product groups**	**support**	**count**
1	{MBIG1=3002,MBIG2=3004}	0.1212	12
2	{MBIG1=3004,MBIG4=9102}	0.1010	10
3	{MBIG1=3004,MBIG3=9102}	0.1212	12
4	{MBIG1=3004,MBIG2=3002}	0.3333	33
5	{MBIG1=3004}	0.5354	53
6	{MBIG2=3002}	0.3838	38
7	{MBIG2=3004}	0.2323	23
8	{MBIG3=9102}	0.2222	22
9	{MBIG4=9102}	0.2121	21
10	{MBIG1=3002}	0.1515	15
11	{MBIG1=7108}	0.1414	14
12	{MBIG3=3002}	0.1212	12
13	{MBIG3=3004}	0.1111	11
14	{MBIG10=9021}	0.1111	11
15	{MBIG7=9999}	0.1010	10
16	{MBIG4=9101}	0.1010	10
17	{MBIG5=9102}	0.1010	10

Item names are ordered below together with definitions;

3002: Human blood; animal blood prepared for therapeutic, prophylactic or diagnostic uses; antisera and other blood fractions and immunological products, whether or not modified or obtained by means of biotechnological processes; vaccines, toxins, cultures of micro-organisms (excluding yeasts) and similar product

3004: Medicaments consisting of mixed or unmixed products for therapeutic or prophylactic uses, put up in measured doses "incl. those in the form of transdermal administration" or in forms or packings for retail sale (excluding goods of heading 3002, 3005 or 3006

9102: Wrist-watches, pocket-watches and other watches, incl. stop-watches (excluding of precious metal or of metal clad with precious metal)

7108: Gold, incl. gold plated with platinum, unwrought or not further worked than semi-manufactured or in powder form

9021: Orthopaedic appliances, incl. crutches, surgical belts and trusses; splints and other fracture appliances; artificial parts of the body; hearing aids and other appliances which are worn or carried, or implanted in the body, to compensate for a defect or disability

9101: Wrist-watches, pocket-watches and other watches, incl. stop-watches, with case of precious metal or of metal clad with precious metal (excluding with backs made of steel)

9999: Commodities not elsewhere specified

Upon the min. support is equal to 0.1 that means that count number of the products in each MBIG is greater than at least 10, 17 products/product groups comes to the front. For example, the product 3002 of MBIG1 and 3004 of MBIG2 are both found together in MBIG1 and MBIG2 in 12 records of transactions. This situation is also valid for rows numbered 2, 3 and 4. This can make us think that 3004 and 3002 products can be exported together. The frequency of the products is summarized below. It can be seen that the products that 3004, 3002, 9102, 7108, 9021 and 9101 are most frequently exported items by the countries.

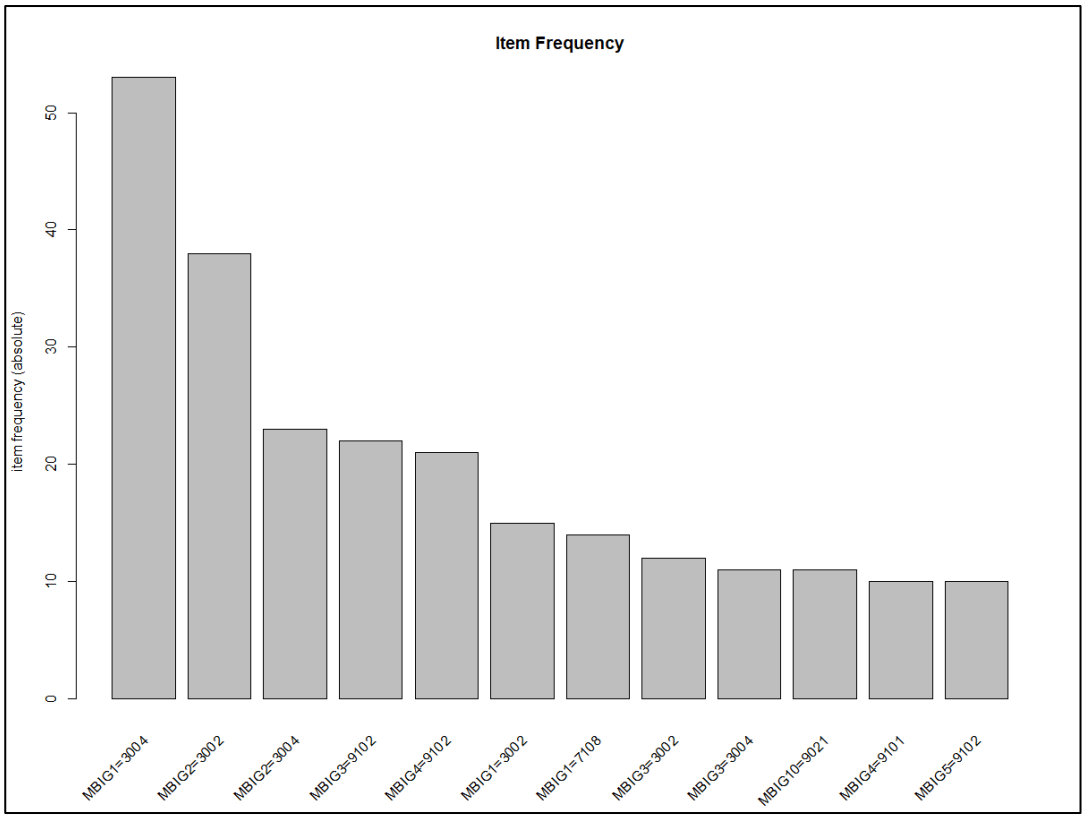


Figure 3.6 Frequency of Top 12 Items by MBIG & min.support rate = 0.1 & Year 2018

Now, in order to get more rules, we go to decrement in the threshold value for Minumum Support. When we decrease Minumum support from 0.1 to 0.075, we get more rules because of the number of rules whose support values are greater than 0.075.

Exported Products of the Year of 2009

After carrying out a frequency analysis of mostly bought/exported products by Switzerland for the year of 2018, let us move on to the frequency investigation of products that are exported for the year of 2009. The below frequency table is designed by the same minimum support as we did to set up a frequency table of the exported products of the year of 2018.

Table 3.11 Support Rates/Counts by MBIGs when min.support = 0.1 & Year 2009

Minimum Support = 0.1			
No	**items/products/product groups**	**support**	**count**
1	{MBIG1=3004,MBIG5=9102}	0.1111	11
2	{MBIG1=3004,MBIG3=3002}	0.1111	11
3	{MBIG1=3004,MBIG2=3002}	0.3333	33
4	{MBIG1=3004}	0.6364	63
5	{MBIG2=3002}	0.3636	36
6	{MBIG2=3004}	0.1717	17
7	{MBIG3=3002}	0.1515	15
8	{MBIG5=9102}	0.1414	14
9	{MBIG3=9102}	0.1212	12
10	{MBIG2=9102}	0.1212	12
11	{MBIG4=9999}	0.1111	11
12	{MBIG4=3002}	0.1010	10

Item names are ordered below together with definitions;

3002: Human blood; animal blood prepared for therapeutic, prophylactic or diagnostic uses; antisera and other blood fractions and immunological products, whether or not modified or obtained by means of biotechnological processes; vaccines, toxins, cultures of micro-organisms (excluding yeasts) and similar product

3004: Medicaments consisting of mixed or unmixed products for therapeutic or prophylactic uses, put up in measured doses "incl. those in the form of transdermal administration" or in forms or packings for retail sale (excluding goods of heading 3002, 3005 or 3006

9102: Wrist-watches, pocket-watches and other watches, incl. stop-watches (excluding of precious metal or of metal clad with precious metal)

9999: Commodities not elsewhere specified

By comparison with the frequencies of the products that are coded 3004 and 3002 in the year of 2018, we can see that by the year of 2009, the numbers of countries

that export these products from Switzerland are greater than that of 2018. For example, 4th row is representing that 3004 coded products are bought from 63 countries over the recorded 99 countries, and 3002 coded products are purchased from 36 countries over the recorded 99 countries. Moreover, the products that are coded 3002 and 9102 are less preferable by the countries in the year of 2009 because the numbers of the countries that export 3002 and 9102 coded products in MBIG2 and MBIG3 respectively are less than that of the countries in the year of 2018. This may be an indication to some extent that the 3002 coded product is exported with the product 3004 later on by the year of 2018. Furthermore, we want to home in on the second row of the above table that displays the purchasing frequency of 3004 and 3002 coded products. Here we see that 3004 from MBIG1 and 3002 from MBIG3 are seen 11 transactions of the union of these groups. However, when we look at the table of the year of 2018, it can be clearly seen that 3002 coded products move from MBIG3 to MBIG2 showing that 3002 coded products are anymore becoming mostly exported products after 3004. Moreover, 9012 coded products are becoming more popular by the year 2018 and entering into groups of products sold together. As a result, upon minimum support is equal to 0.1, 3002 and 9102 coded products are entering the list of mostly exported products by Switzerland to 99 countries.

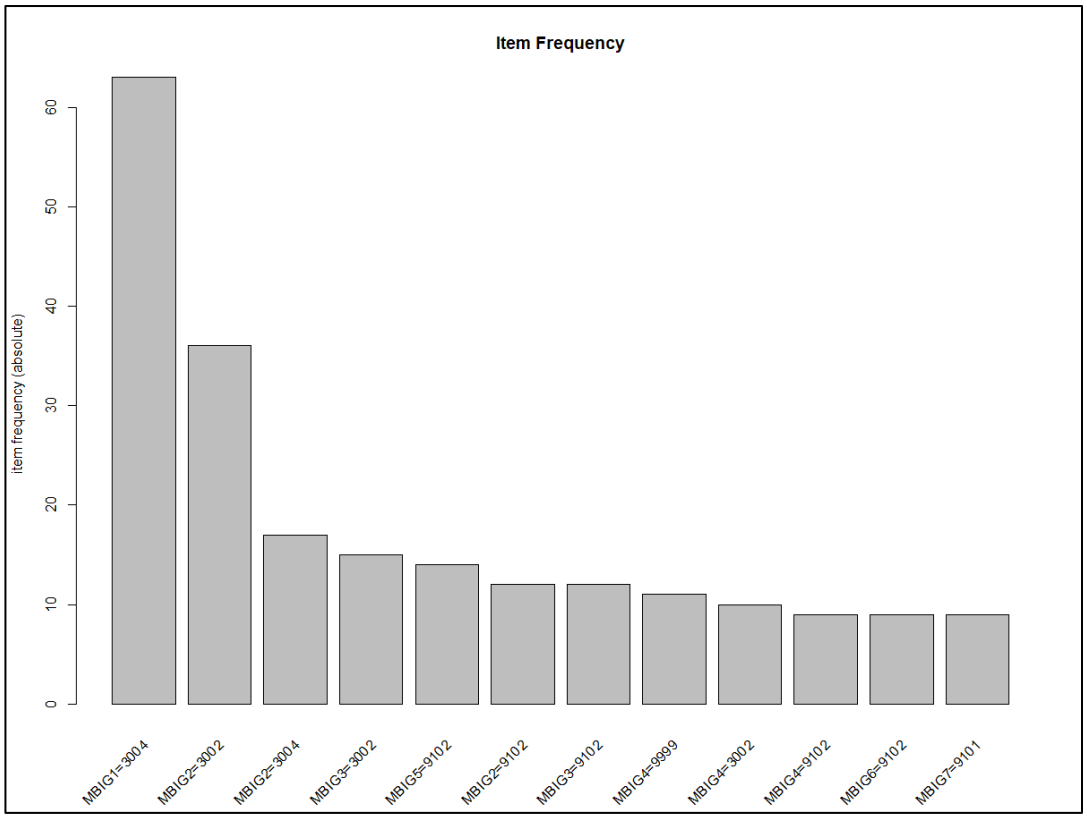


Figure 3.7 Frequency of Top 12 Items by MBIG & min.support rate = 0.1 & Year 2009

The frequency of the products is summarized in the above figure. It can be seen that the products coded 3004, 3002, 9102, and 9101 are the most frequently exported items by the countries. By the year of 2018, the products coded 7108 and 9021 are entering the list of most frequently exported products together.

Now, in order to get more rules, we go to decrement in the threshold value for Minumum Support. When we decrease Minumum support from 0.1 to 0.075, we get more rules because of the number of the rules whose support values are greater than 0.075.

Exported Products of the Year of 2001

After carrying out a frequency analysis of mostly bought/exported products by Switzerland for the year of 2018 and 2009, let us move on to the frequency investigation of products that are exported for the year of 2001. The below

frequency table is designed by the same minimum support as we did to set up a frequency table of the exported products of the year of 2018 and 2009.

Table 3.12 Support Rates/Counts by MBIGs when min.support = 0.1 & Year 2001

Minimum Support = 0.1			
No	**items/products/product groups**	**support**	**count**
1	{MBIG1=3004,MBIG3=9102}	0.1111	11
2	{MBIG1=3004}	0.5657	56
3	{MBIG2=3004}	0.1616	16
4	{MBIG3=9102}	0.1616	16
5	{MBIG2=9102}	0.1414	14
6	{MBIG6=9102}	0.1010	10

Item names are ordered below together with definitions;

3002: Human blood; animal blood prepared for therapeutic, prophylactic or diagnostic uses; antisera and other blood fractions and immunological products, whether or not modified or obtained by means of biotechnological processes; vaccines, toxins, cultures of micro-organisms (excluding yeasts) and similar product

3004: Medicaments consisting of mixed or unmixed products for therapeutic or prophylactic uses, put up in measured doses "incl. those in the form of transdermal administration" or in forms or packings for retail sale (excluding goods of heading 3002, 3005 or 3006

9102: Wrist-watches, pocket-watches and other watches, incl. stop-watches (excluding of precious metal or of metal clad with precious metal)

The frequency statistics of the products based on the year of 2001 can be seen from the above table. When the minimum support is determined as 0.1, we want to see products whose number of appearance in the list must be equal to or greater than 10 out of the recorded 99 transactions. The products that we know from the exported products list of the year of 2009 and 2018 can be seen in the list of the year of 2001. However, only one product that is coded as 3002 is not available in the table

when minimum support is equal to 0.1. Throughout the years from 2001 to 2018, we can make an inference such that the product coded 3002 variable is taken interest by the coming of year of 2009. Moreover, the product coded 7108 variable raises awareness of the countries with the coming of year of 2018. 3004 coded products are protecting its place in the list of the most frequently exported products and do not lose its popularity during the 18-year period. Therefore, we can say that 3002 coded products are becoming prevalent in the list with the year of 2009, and 7108 coded products catch attention with the year of 2018.

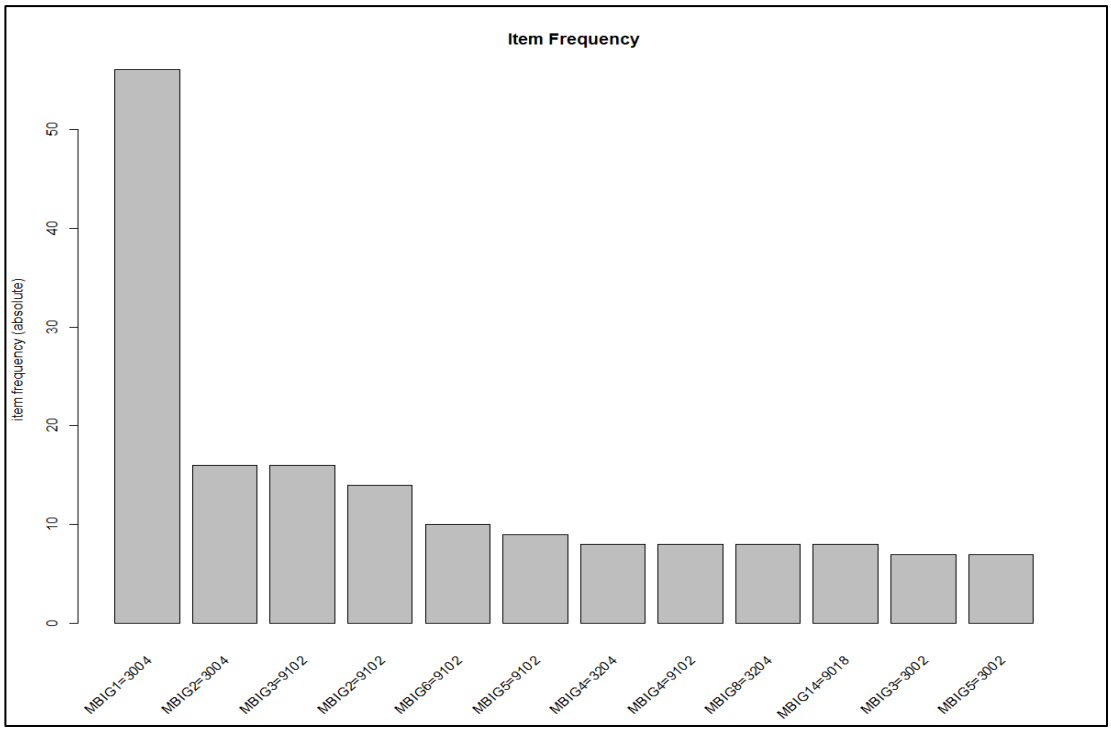


Figure 3.8 Frequency of Top 12 Items by MBIG & min.support rate = 0.075 & Year 2001

The above figure proves what we made inference previously because we can see that the product 3002 is seen at the right end side of the figure indicating that its frequency in the table of most frequently exported products is less as compared to the products 3004 and 9102. Moreover, in this figure one product catches our attention that is coded 3204 whose number of appearances is greater than that of 3002. However, the improvement in the trade volume of product coded 3002 overtakes the trade volume of 3204 as time passed.

Now, in order to get more rules, we go to decrement in the threshold value for Minumum Support. When we decrease Minumum support from 0.1 to 0.075, we get more rules because of the number of the rules whose support values are greater than 0.075.

To sum up, throughout 18-year period, the 3004 and 9102 coded products are the most frequently exported products together. We can summarize the most important findings as following;

- 2001 >> **3004**(Medicaments consisting of mixed or unmixed products for therapeutic or prophylactic uses, put up in measured doses "incl. those in the form of transdermal administration" or in forms or packings for retail sale (excluding goods of heading 3002, 3005 or 3006)) and **9102**(Wrist-watches, pocket-watches and other watches, incl. stop-watches (excluding of precious metal or of metal clad with precious metal) are the main products.
- 2009 >> **3004**(Medicaments consisting of mixed or unmixed products for therapeutic or prophylactic uses, put up in measured doses "incl. those in the form of transdermal administration" or in forms or packings for retail sale (excluding goods of heading 3002, 3005 or 3006)) and **9102**(Wrist-watches, pocket-watches and other watches, incl. stop-watches (excluding of precious metal or of metal clad with precious metal) are the main products. **3002**(Human blood; animal blood prepared for therapeutic, prophylactic or diagnostic uses; antisera and other blood fractions and immunological products, whether or not modified or obtained by means of biotechnological processes; vaccines, toxins, cultures of micro-organisms (excluding yeasts) and similar product) is the new product.
- 2018 >> **3004**(Medicaments consisting of mixed or unmixed products for therapeutic or prophylactic uses, put up in measured doses "incl. those in the form of transdermal administration" or in forms or packings for retail sale (excluding goods of heading 3002, 3005 or 3006)), **9102**(Wrist-watches, pocket-watches and other watches, incl. stop-watches (excluding

of precious metal or of metal clad with precious metal) are the main products and **3002**(Human blood; animal blood prepared for therapeutic, prophylactic or diagnostic uses; antisera and other blood fractions and immunological products, whether or not modified or obtained by means of biotechnological processes; vaccines, toxins, cultures of micro-organisms (excluding yeasts) and similar product)are main products.7108 is the main products. **7108(**Gold, incl. gold plated with platinum, unwrought or not further worked than semi-manufactured or in powder form) is the new product.

3.2.4 Apriori Algorithm (AA) on given Data Set based on Exports of Switzerland to 99 Countries

After assessing the most frequently exported products from Switzerland to 99 countries for the years of 2001, 2009 and 2018, we move on to struggling for finding the association rules. From the analysis of frequency table, the expected association rules must include at least one of the products that are ordered and shown in the above frequency tables obtained by the minimum support 0.1 and 0.075, respectively.

The most important step in the usage of the AA Algorithm is to determine parameter values that are provided into algorithms during the analysis. As we stated in the explanation part of AA, we are obligatory to determine 2 important parameters that are “Minimum Support (MS)” and “Minimum Confidence (MC)” into consideration.

In order to get away from the problems that are possible to be encountered during figuring out rules, we are required to take two important points into account. Firstly, the big MS values limit us to finding more rules and can cause us to ignore many of the rules that can be significant. The small MS values can provide us with more specific and meaningless rules that lead to insignificant results. Therefore,

while determining parameter values, it can be useful and beneficial to take the track of the results of the frequency table into consideration.

1st Apriori Algorithm for Exported Products of the Year of 2018

The first parameter that is used in the AA is given below and let us find the rules.

Table 3.13 Parameter Space-1 & Year 2018

Parameter Space	Parameter Values
Minimum Support	0.1
Minimum Confidence	0.50
Minimum Length of the Rule	2

Now, let us find the association rules based on parameters shown in Table 3.13.

Table 3.14 AR based on Parameter Space-1 & Year 2018

Number of Rules	lhs		rhs	support	confidence	lift	count
1	{MBIG1=3002}	=>	{MBIG2=3004}	0,12	0,80	3,44	12
2	{MBIG2=3004}	=>	{MBIG1=3002}	0,12	0,52	3,44	12
3	{MBIG2=3002}	=>	{MBIG1=3004}	0,33	0,87	1,62	33
4	{MBIG1=3004}	=>	{MBIG2=3002}	0,33	0,62	1,62	33
5	{MBIG3=9102}	=>	{MBIG1=3004}	0,12	0,55	1,02	12

The first association rules are shown in the above table. According to parameter space, only 5 rules are suggested by the AA. The table is designed by ordering the lift value from largest to smallest one. The reason why we order rules by lift is due to the fact that lift value is the most valuable and certain criteria to choose the best rules. Item names are ordered in the previous part.

- The first rule says;

{MBIG1=3002}	=>	{MBIG2=3004}

- If a country imports the product coded 3002 from Switzerland, then the same country is also buying the product coded 3004 on a given interval.

- The 5th rule says;

{MBIG3=9102}	=>	{MBIG1=3004}

- If a country imports the product coded 9102 from Switzerland, then the same country is also buying the product coded 3004 on a given interval.

The same interpretations can also be made for the other rules.

Now, in order to see the association rules in Table 3.14 together, we can get benefit from the following graph that is drawn with confidence on y-axis and support on x-axis. The darker red a data point has, the highest lift value an association rule has.

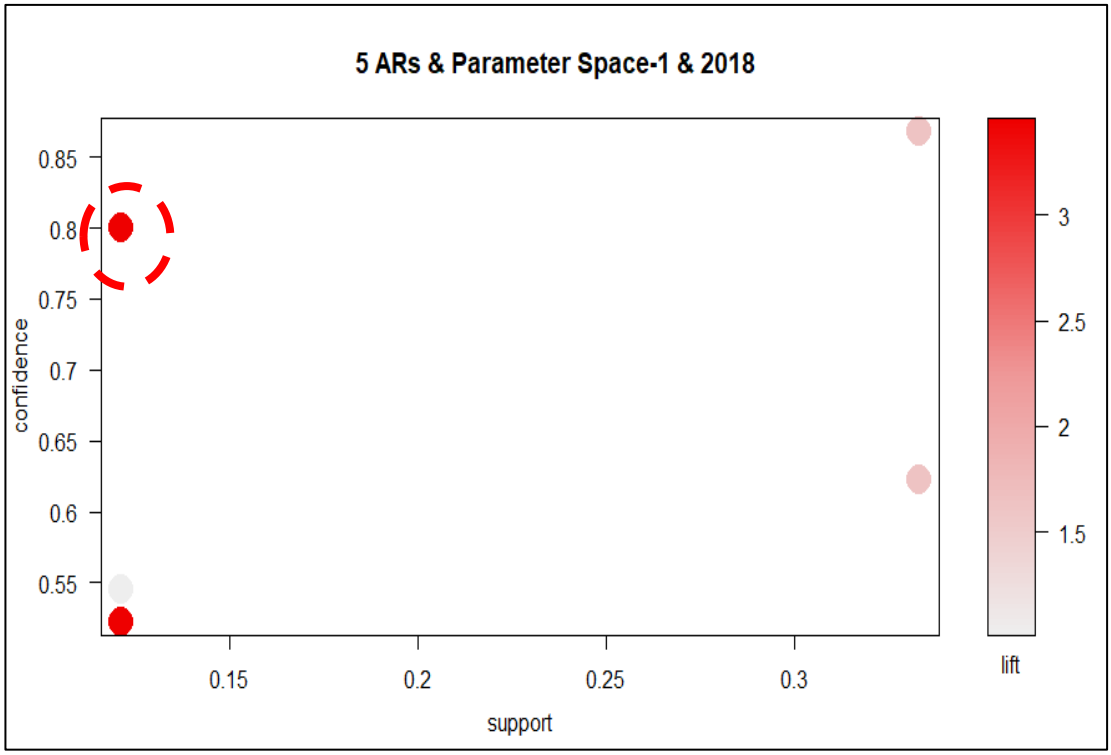


Figure 3.9 5 ARs & Parameter Space-1 & Year 2018 & Scatter Plot

The above figure can help discover the best rules at first glance. The selection of the best rules is performed by determining the biggest lift value. We can see that the biggest lift values are being represented with a darker red color. Thus, we need to pay attention to the left part of the figure. From the left part, if the degree of red is the same across the rules, the second assessment must be made based on confidence. Then, we can see that the point shown within the red dashed circle is the most powerful rule among 5 rules. This rule corresponds to the first rule in the above table. Another visualization of the rules can be illustrated using a matrix plot. The following matric plot represents the rules based on the lift values. The matrix plot rules are divided into two groups. The left side of the rules is called Antecedent (LHS) and the right side of the rules is called Consequent (RHS).

Table 3.15 Antecedent and Consequent for 5 Rules

Itemsets in Antecedent (LHS)				
1	2	3	4	5
{MBIG1=3002}	{MBIG2=3004}	{MBIG2=3002}	{MBIG1=3004}	{MBIG3=9102}
Itemsets in Consequent (RHS)				
1	2	3	4	
{MBIG1=3004}	{MBIG2=3002}	{MBIG2=3004}	{MBIG1=3002}	

Now, let us visualize the rule in the different graphical representations.

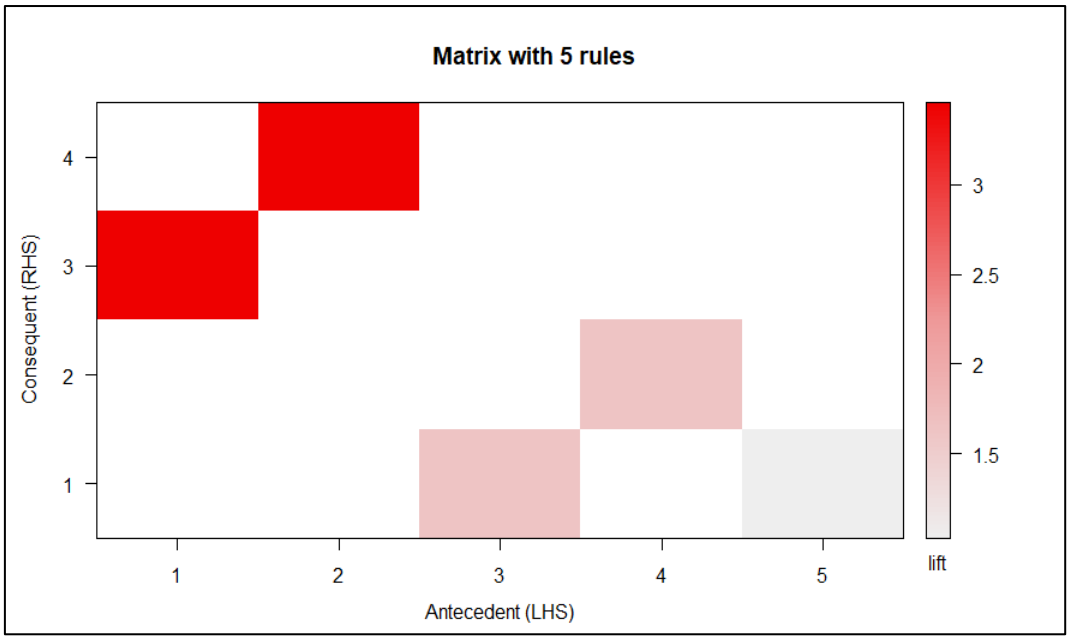


Figure 3.10 ARs & Parameter Space -1 & Year 2018 & Matrix Plot

Reading the above table can also help us understand the best rules based on the biggest lift value. **Table 3.15** is constructed with the matrix plot. For example, in the matrix plot, the x-axis represents the left-hand side of the rules and the y-axis displays the right-hand side of the rules. From the matrix plot, when the antecedent is equal to 5, then the consequent becomes 1. This forms a rule and the lift value of this rule is understood by the degree of color. If the color is darker red, it means that lift value is high. The lift value for the rule of 5>>1 is the smallest one by comparison with all the other 4 rules.

2nd Apriori Algorithm for Exported Products of the Year of 2018

The second parameter that is used in the AA is given below and let us find the rules.

Table 3.16 Parameter Space-2 & Year 2018

Parameter Space	Parameter Values
Minumum Support	0.075
Minumum Confidence	0.50
Minumum Length of the Rule	2

Now, let us find the association rules based on parameters shown in Table 3.16.

Table 3.17 ARs based on Parameter Space-2 & Year 2018

Number of Rules	lhs		rhs	support	confidence	lift	count
1	{MBIG1=3002}	=>	{MBIG2=3004}	0.12	0.80	3.44	12
2	{MBIG2=3004}	=>	{MBIG1=3002}	0.12	0.52	3.44	12
3	{MBIG1=3004,MBIG3=9102}	=>	{MBIG2=3002}	0.09	0.75	1.95	9
4	{MBIG2=3002,MBIG3=9102}	=>	{MBIG1=3004}	0.09	1.00	1.87	9
5	{MBIG2=3002}	=>	{MBIG1=3004}	0.33	0.87	1.62	33
6	{MBIG1=3004}	=>	{MBIG2=3002}	0.33	0.62	1.62	33
7	{MBIG3=9102}	=>	{MBIG1=3004}	0.12	0.55	1.02	12

By the parameters given in Table 3.16, we reach to seven rules in Table 3.17. Interpretations can be made in the same manner as we write in the previous part. The table is designed by ordering the lift value from the largest to the smallest one. Item names are ordered in the previous part.

- The 3rd rule says;

{MBIG1=3004,MBIG3=9102}	=>	{MBIG2=3002}

- If a country imports the products coded 3004 and 9102 together, respectively, then the same country has the potency to buy the product coded 3002. Be careful that this rule does not imply that the reverse of this explanation is true. If it were true, we would have detected the reversed version of this rule in the table.

- The 7th rule says;

{MBIG3=9102}	=>	{MBIG1=3004}

If a country imports the product coded 9102 from Switzerland, then the same country is also buying the product coded 3004 on a given interval.

Now, let us visualize the rule in the different graphical representations.

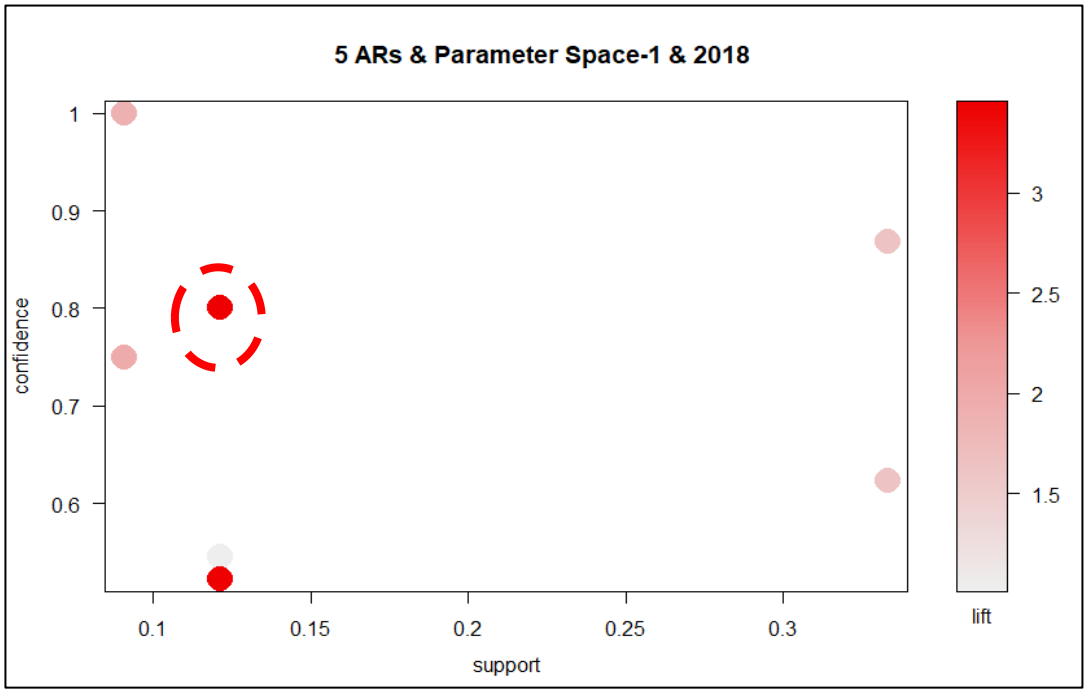


Figure 3.11 ARs & Parameter Space-2 & Year 2018 & Scatter Plot

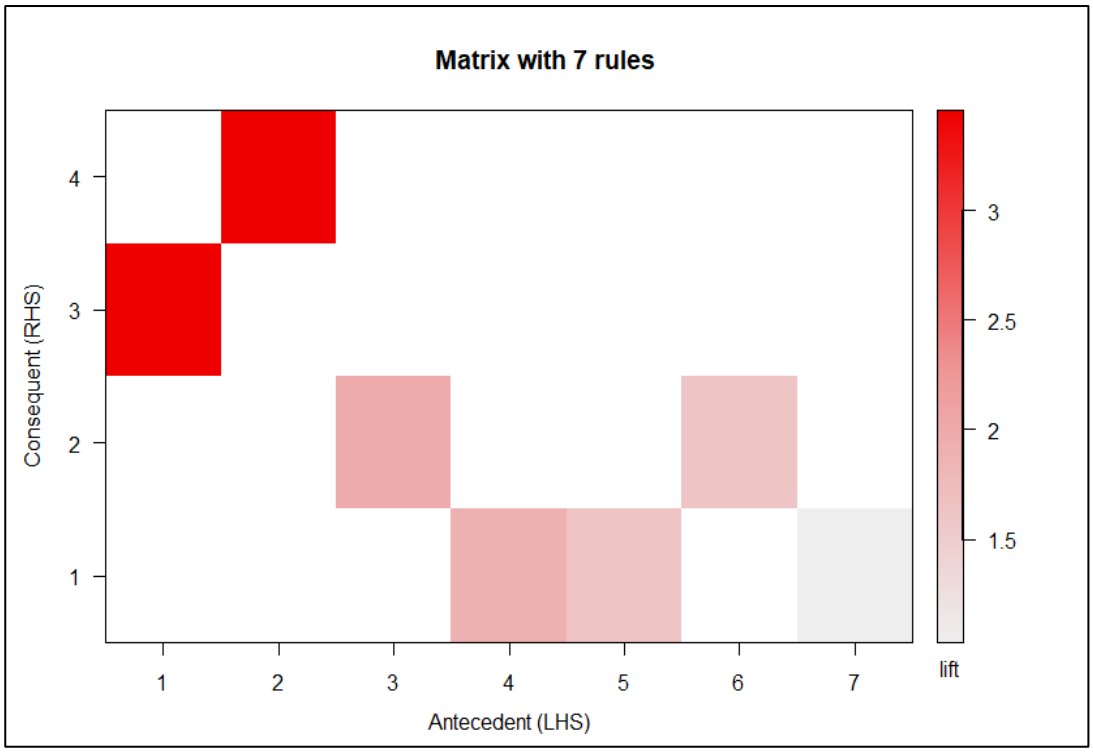


Figure 3.12 ARs & Parameter Space -2 & Year 2018 & Matrix Plot

- 1st figure represents all the rules. It is clear that two of the red points are darker than the others, meaning their lift values are greater than that of others. One of the rules that have higher lift values has also higher confidence value. Thus, the most powerful and beneficial rule is shown within the red dashed circle.
- 2nd figure displays all the rules. We have 7 distinct Antecedents and 4 different Consequents. 1st and 2nd antecedents can be paid attention because their corresponding lift values are greater than those of the others. This happens always if you order the rules by the lift value.

1st Apriori Algorithm for Exported Products of the Year of 2009

The first parameter that is used in the AA is given below and let us find the rules.

Table 3.18 Parameter Space-1 & Year 2009

Parameter Space	Parameter Values
Minumum Support	0.1
Minumum Confidence	0.50
Minumum Length of the Rule	2

Now, let us find the association rules based on parameters shown in Table 3.18.

Table 3.19 ARs based on Parameter Space-1 & Year 2009

Rules	lhs		rhs	support	confidence	lift	count
1	{MBIG1=3004}	=>	{MBIG2=3002}	0.33	0.52	1.44	33
2	{MBIG2=3002}	=>	{MBIG1=3004}	0.33	0.92	1.44	33
3	{MBIG5=9102}	=>	{MBIG1=3004}	0.11	0.79	1.23	11
4	{MBIG3=3002}	=>	{MBIG1=3004}	0.11	0.73	1.15	11

The first association rules are shown in the above table. According to the parameter space, only 4 rules are suggested by the AA. The table is designed by ordering the lift value from the largest to the smallest one. The reason why we order rules by lift is due to the fact that lift value is the most valuable and certain criteria to choose the best rules. Item names are ordered in the previous part.

- The first rule says;

{MBIG1=3004}	=>	{MBIG2=3002}

- If a country imports the product coded 3004 from Switzerland, then the same country is also buying the product coded 3002 on a given interval.
- The 3rd rule says;

{MBIG5=9102}	=>	{MBIG1=3004}

- If a country imports the product coded 9102 from Switzerland, then the same country is also buying the product coded 3004 on a given interval.

The same interpretations can also be made for the other rules.

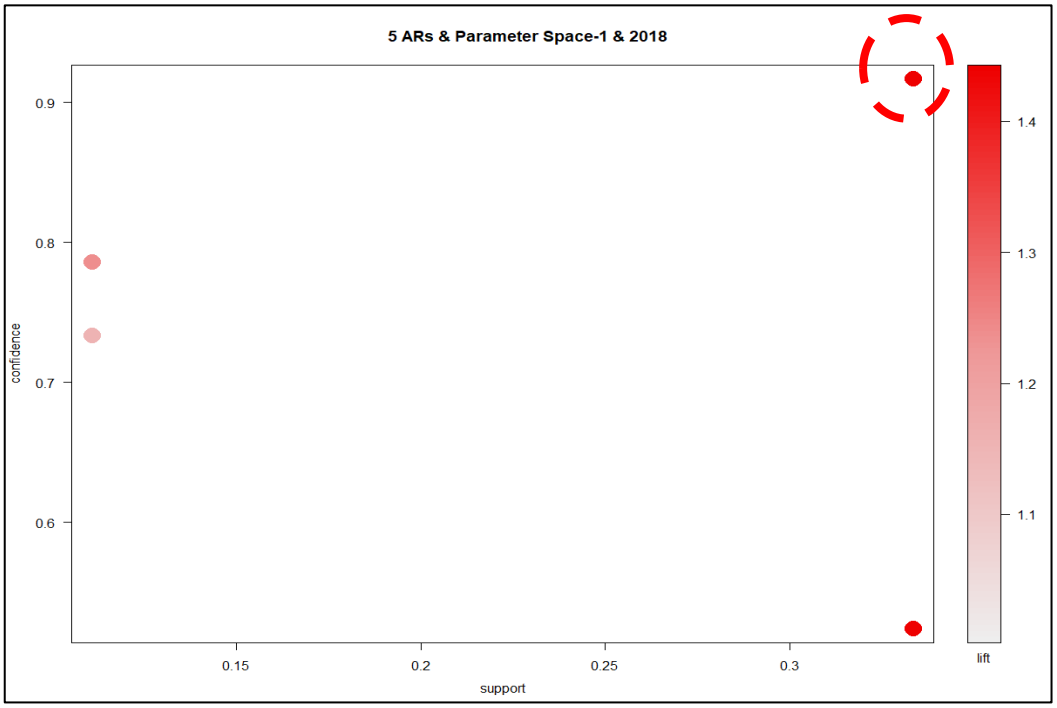


Figure 3.13 ARs & Parameter Space-1 & Year 2009 & Scatter Plot

Table 3.20 Antecedent and Consequent for 4 Rules

Itemsets in Antecedent (LHS)			
1	2	3	4
{MBIG1=3004}	{MBIG2=3002}	{MBIG5=9102}	{MBIG3=3002}
Itemsets in Consequent (RHS)			
1	2		
{MBIG1=3004}	{MBIG2=3002}		

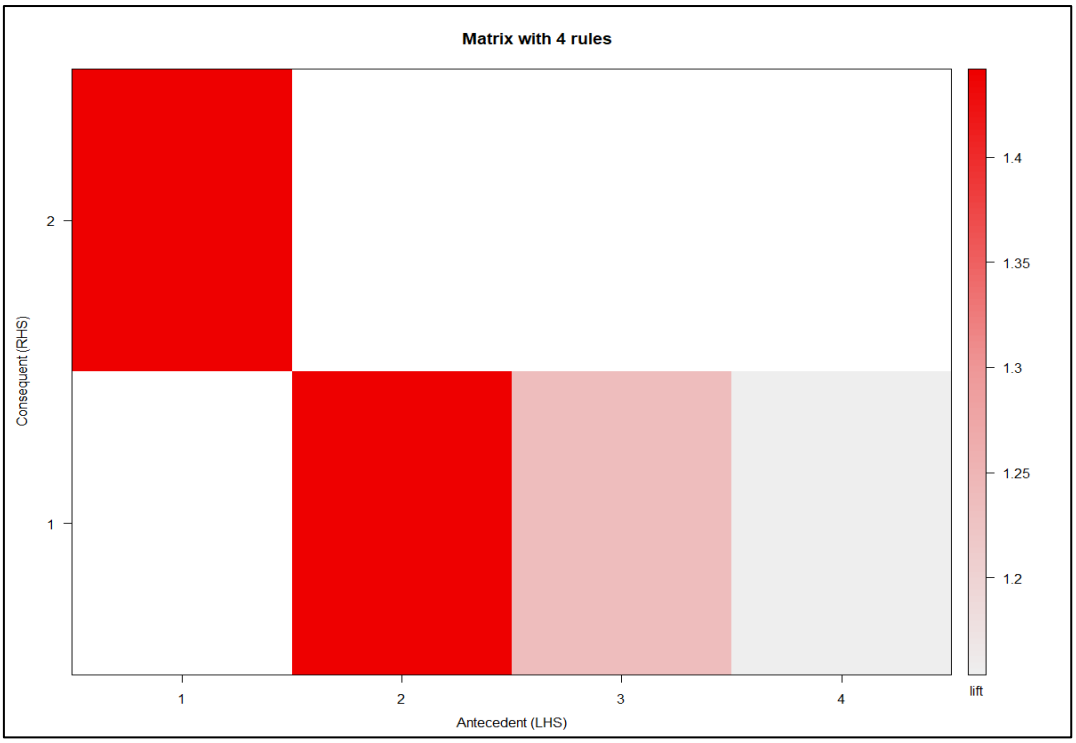


Figure 3.14 ARs & Parameter Space -1 & Year 2009 & Matrix Plot

The above two graphical representations are summarizing all the rules obtained by using parameter space-1.

- 1st figure shows 4 rules and the lift value of each rule is deciphered by the degree of red color. The most powerful rule is displayed by the red data point existing within the red dashed line because its confidence and lift values are greater than the values of the others.
- 2nd figure offers us to see the rule on the matrix surface. X-axis includes the products which exist at the left-hand side of ARs. From the table of "Antecedent and Consequent for 4 Rules", the related products in the association rules are summarized by the antecedents that show products placed on the left-hand side of the rules and the consequents that display products placed on the right-hand side of the rules. The matrix graph helps us understand which rules' lift values are greater, which rules are preferable as compared to others. In this plot, we cannot see the products' names.

By comparison with the rules that are coming from the investigation of the association rules that are obtained by the parameter space-1 & Year 2018, the products that are involved in the rules are similar to each other. For example, the products coded 3004 and 3002 are more associated with each other. Another group of products associated with each other includes 9102 and 3004.

2nd Apriori Algorithm for Exported Products of the Year of 2009

The first parameter that is used in the AA is given below and let us find the rules.

Table 3.21 Parameter Space-2 & Year 2009

Parameter Space	Parameter Values
Minumum Support	0.075
Minumum Confidence	0.50
Minumum Length of the Rule	2

Now, let us find the association rules based on parameters shown in Table 3.21.

Table 3.22 ARs based on Parameter Space-2 & Year 2009

Rules	lhs		rhs	support	confidence	lift	count
1	{MBIG1=3004}	=>	{MBIG2=3002}	0.33	0.52	1.44	33
2	{MBIG2=3002}	=>	{MBIG1=3004}	0.33	0.92	1.44	33
3	{MBIG7=9101}	=>	{MBIG1=3004}	0.08	0.89	1.40	8
4	{MBIG4=3002}	=>	{MBIG1=3004}	0.08	0.80	1.26	8
5	{MBIG5=9102}	=>	{MBIG1=3004}	0.11	0.79	1.23	11
6	{MBIG3=3002}	=>	{MBIG1=3004}	0.11	0.73	1.15	11
7	{MBIG4=9999}	=>	{MBIG1=3004}	0.08	0.73	1.14	8
8	{MBIG2=9102}	=>	{MBIG1=3004}	0.08	0.67	1.05	8

Table 3.22 (continued)

9	{MBIG3=9102}	=>	{MBIG1=3004}	0.08	0.67	1.05	8

The first association rules are shown in the above table. According to parameter space, only 7 rules are suggested by the AA. The table is designed by ordering the lift values from largest to smallest one. The reason why we order rules by lift is due to the fact that lift value is the most valuable and certain criteria to choose the best rules. Item names are ordered in the previous part.

- The first rule says;

{MBIG1=3004}	=>	{MBIG2=3002}

- If a country imports the product coded 3004 from Switzerland, then the same country is also buying the product coded 3002 on a given interval.

- The 3rd rule says;

{MBIG7=9101}	=>	{MBIG1=3004}

- If a country imports the product coded 9101 from Switzerland, then the same country is also buying the product coded 3004 on a given interval.

The graphical representations of the above rules are shown below;

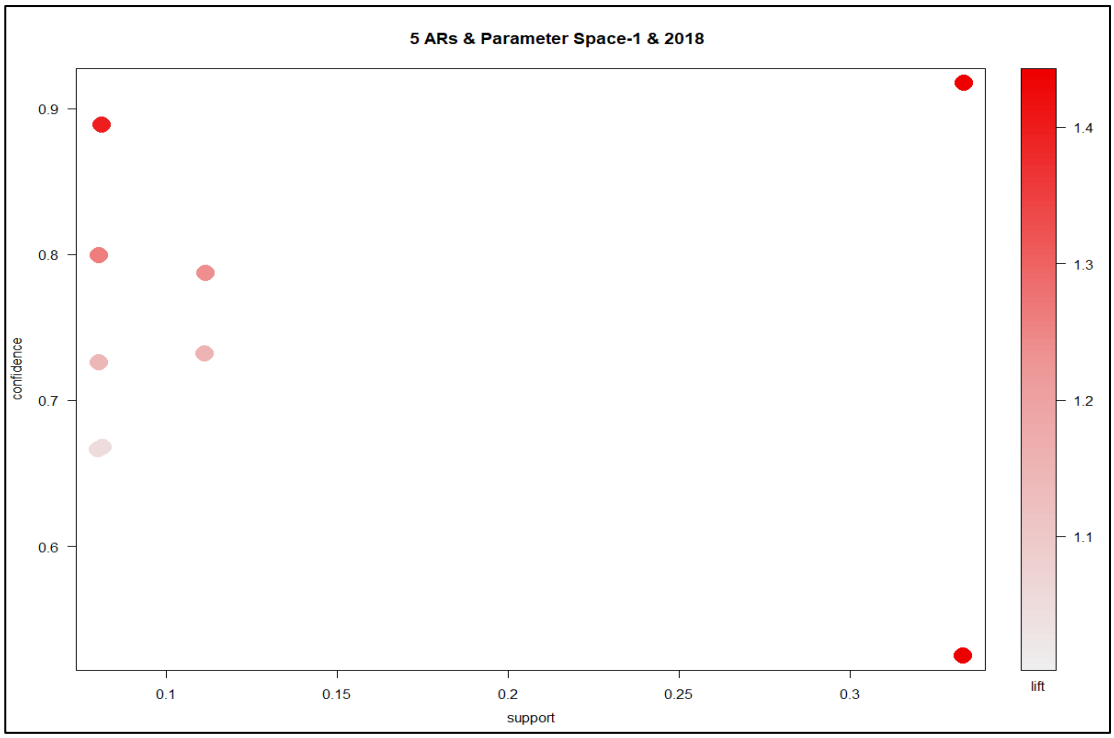


Figure 3.15 ARs & Parameter Space-2 & Year 2009 & Jittered Scatter Plot

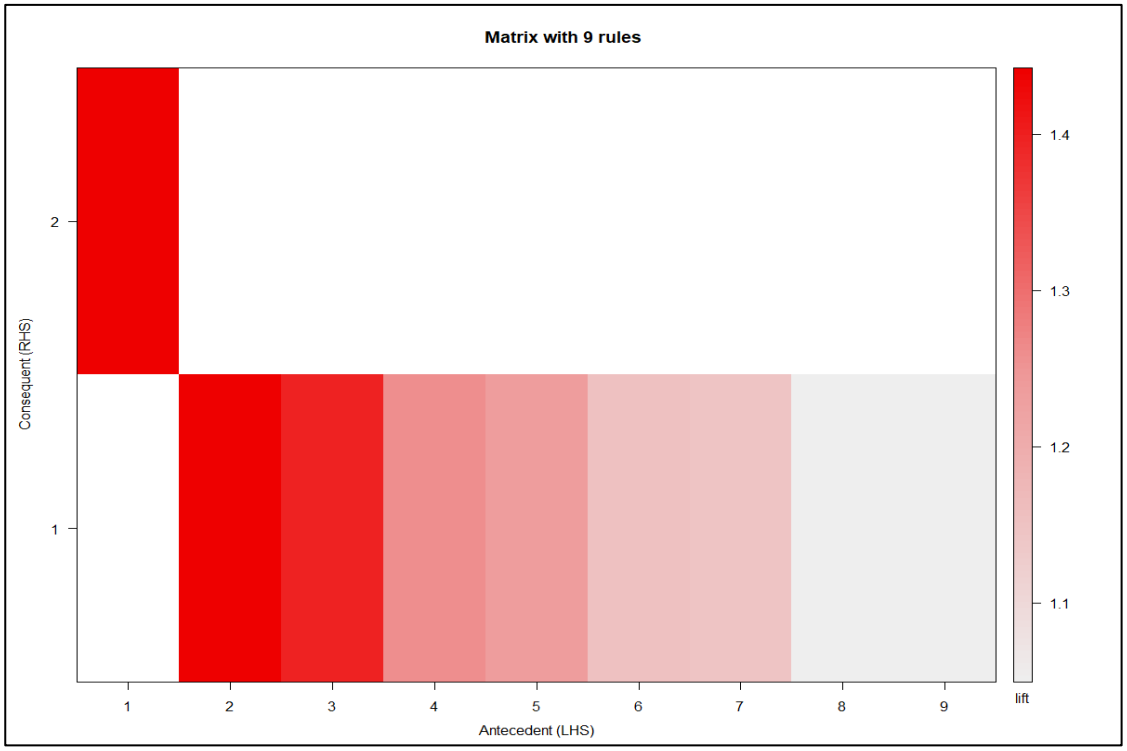


Figure 3.16 ARs & Parameter Space -2 & Year 2009 & Matrix Plot

1st Apriori Algorithm for Exported Products of the Year of 2001

The first parameter that is used in the AA is given below and let us find the rules

Table 3.23 Parameter Space-1 & Year 2001

Parameter Space	Parameter Values
Minumum Support	0.05
Minumum Confidence	0.50
Minumum Length of the Rule	2

Now, let us find the association rules based on parameters shown in Table 3.23.

Table 3.24 ARs based on Parameter Space-1 & Year 2001

Rules	lhs		rhs	support	confidence	lift	count
1	{MBIG14=9021}	=>	{MBIG1=3004}	0.06	0.86	1.52	6
2	{MBIG5=3002}	=>	{MBIG1=3004}	0.06	0.86	1.52	6
3	{MBIG3=3002}	=>	{MBIG1=3004}	0.06	0.86	1.52	6
4	{MBIG2=3808}	=>	{MBIG1=3004}	0.05	0.83	1.47	5
5	{MBIG4=9102}	=>	{MBIG1=3004}	0.06	0.75	1.33	6
6	{MBIG4=3204}	=>	{MBIG1=3004}	0.06	0.75	1.33	6
7	{MBIG9=9999}	=>	{MBIG1=3004}	0.05	0.71	1.26	5
8	{MBIG5=3808}	=>	{MBIG1=3004}	0.05	0.71	1.26	5
9	{MBIG3=9102}	=>	{MBIG1=3004}	0.11	0.69	1.22	11
10	{MBIG2=9102}	=>	{MBIG1=3004}	0.09	0.64	1.14	9
11	{MBIG14=9018}	=>	{MBIG1=3004}	0.05	0.63	1.10	5
12	{MBIG6=9102}	=>	{MBIG1=3004}	0.06	0.60	1.06	6

The first association rules are shown in the above table. According to parameter space, only 12 rules are suggested by the AA. The table is designed by ordering the lift values from largest to smallest one. The reason why we order rules by lift is due to the fact that lift value is the most valuable and certain criteria to choose the best rules. Item names are ordered in the previous part.

- The first rule says;

{MBIG14=9102}	=>	{MBIG1=3004}

- If a country imports the product coded 9102 from Switzerland, then the same country is also buying the product coded 3004 on a given interval.

We can realize from the above results that although MBIG shows the increase in the most bought item group, we can come across the same products that are associated with each other. This is an indication of the power of the algorithm that finds the products whose degree of preference is different from others.

- The 3rd rule says;

{MBIG3=3002}	=>	{MBIG1=3004}

- If a country imports the product coded 3002 from Switzerland, then the same country is also buying the product coded 3004 on a given interval.

2nd Apriori Algorithm for Exported Products of the Year of 2001

The second parameter that is used in the AA is given below and let us find the rules

Table 3.25 Parameter Space-1 & Year 2001

Parameter Space	Parameter Values
Minumum Support	0.075
Minumum Confidence	0.50
Minumum Length of the Rule	2

Now, let us find the association rules based on parameters shown in Table 3.25.

Table 3.26 ARs based on Parameter Space-2 & Year 2001

Rules	lhs		rhs	support	confidence	lift	count
1	{MBIG3=9102}	=>	{MBIG1=3004}	0,11	0,69	1,22	11
2	{MBIG2=9102}	=>	{MBIG1=3004}	0,09	0,64	1,14	9

The products coded 9102 and 3004 are most frequently sold/bought products together in the year of 2001. It can be seen that under the given parameters, these products are sold together and their degree of preference is higher than the degree of preference of buying some other products together. The description of the rules can be defined as we did in the previous parts. Item names are ordered in the previous part.

3.2.5 Conclusion and Discussion

- **Association Rules attributed to 1st and 2nd Apriori Algorithms for Exported Products for the years of 2018 and 2009.**

Now, let us look at and draw comparisons between the rules that are obtained by means of the first parameter spaces for the data sets of the year of 2009 and the year of 2018 transactions.

Table 3.27 ARs by 1st Parameter Space for the years of 2009 & 2018

2009-1st Parameter Space					2018-1st Parameter Space				
Rules	lhs		rhs	lift	Rules	lhs		rhs	lift
1	{MBIG1=3004}	=>	{MBIG2=3002}	1.4	1	{MBIG1=3002}	=>	{MBIG2=3004}	3.44
2	{MBIG2=3002}	=>	{MBIG1=3004}	1.4	2	{MBIG2=3004}	=>	{MBIG1=3002}	3.44
3	{MBIG5=9102}	=>	{MBIG1=3004}	1.2	3	{MBIG2=3002}	=>	{MBIG1=3004}	1.62

Table 3.27 (continued)

4	{MBIG3=3002}	=>	{MBIG1=3004}	1.2	4	{MBIG1=3004}	=>	{MBIG2=3002}	1.62
NO RULE					5	{MBIG3=9102}	=>	{MBIG1=3004}	1,02

Table 3.27 displays that the products coded 3004, 3002, and 9102 are most frequently sold/bought products together. For 9 year periods from 2009 to 2018, there is clear evidence that the interaction and association between the products coded 3004 and 3002 had increased. Furthermore, the preference to buy 3002 coded product seems to increase in the year of 2018 because of the fact that MBIG1 includes this product. The product coded 9102 comes to 3rd place of mostly sold/bought product list in the year of 2018 (5th row and MBIG3=9102). The association rules developing out of the analysis of the transactions of the year 2009 include also 9102 coded product based rules. However, the place of this product in the list of the most sold/bought product in the year of 2009 and 2018 is different from each other and this product in the year of 2009 is less preferred by the year of 2009, but it is more preferable to be bought together with the products coded 3002 & 3004. This can give us information about the trade policy of the countries that are importing from Switzerland. When we look at the lift values, we can reach higher lift values that indicate the power of the rule. For example, the lift value of the same rules increased from 1.4 to 3.44 from 2009 to 2018. This situation strengthens the place of 3002 coded products in the most bought itemsets with 3004.

Then, let us look at and draw comparisons between the rules that are obtained by means of the second parameter spaces for the data sets of the years of 2009 and the year of 2018 transactions.

Table 3.28 ARs by 2nd Parameter Space for the years of 2009 & 2018

2009-2nd Parameter Space					2018-2nd Parameter Space				
Rules	lhs		rhs	lift	Rules	lhs		rhs	lift
1	{MBIG1=3004}	=>	{MBIG2=3002}	1.44	1	{MBIG1=3002}	=>	{MBIG2=3004}	3.44
2	{MBIG2=3002}	=>	{MBIG1=3004}	1.44	2	{MBIG2=3004}	=>	{MBIG1=3002}	3.44
3	{MBIG7=9101}	=>	{MBIG1=3004}	1.40	3	{MBIG1=3004,MBIG3=9102}	=>	{MBIG2=3002}	1.95
4	{MBIG4=3002}	=>	{MBIG1=3004}	1.26	4	{MBIG2=3002,MBIG3=9102}	=>	{MBIG1=3004}	1.87
5	{MBIG5=9102}	=>	{MBIG1=3004}	1.23	5	{MBIG2=3002}	=>	{MBIG1=3004}	1.62
6	{MBIG3=3002}	=>	{MBIG1=3004}	1.15	6	{MBIG1=3004}	=>	{MBIG2=3002}	1.62
7	{MBIG4=9999}	=>	{MBIG1=3004}	1.14	7	{MBIG3=9102}	=>	{MBIG1=3004}	1.02
8	{MBIG2=9102}	=>	{MBIG1=3004}	1.05	NO RULE				
9	{MBIG3=9102}	=>	{MBIG1=3004}	1.05					

As for the above Table 3.28 that shows the rules by the 2nd parameter spaces, we can get a better understanding of which products are more associated with each other because we are decreasing support threshold value to bring more rule into the open. The rules that are based on the data of the year of 2009 are mostly including the products as same as what we have found by the 1st parameter space. Differently, we come upon two more products coded 9101 and 9999 that have the potency to be associated with the products coded 3004. The 9101 coded products come into the list of mostly bought item groups of some countries in the 7th order (3rd row and MBIG7). We expect this product to be also associated with the product 9102 because they are under the same category but the results inform us about no association between these products. After 9 years period, when the threshold value of support is decreased, the rules that are composed of 3 products/items start to come into existence. These rules contain the products coded 3004, 3002 and 9102 as illustrated in the 3rd and the 4th rules in the table of "ARs

by 2nd Parameter Space for the years of 2009 & 2018" under the head of 2018-2nd parameter space. We want to draw attention to these rules. As you see, the 3rd and 4th rules say that if countries import both the products coded 3002 and 9102, they also buy 3004 and if countries import both the products coded 3004 and 9102, they also buy 3002. However, this rule does not tell that if countries import both the products coded 3002 and 3004, they also buy 9102. This indicates to some extent that the power of association between the products coded 3004 and 3002 is greater than both the power of association between the products coded 3004 and 9102 or between the products coded 3002 and 9102.

To sum up, the AR rules based on parameter space-1&2 for the years of 2009 and 2018 say that the power of associations is greater between the products coded 3004 and 3002. Moreover, the product coded 9102 seems to be frequently bought and sold together with the product coded 3004 and 3002.

END OF DATA MINING-ASSOCIATION RULE MINING

End of Chapter 3

CHAPTER 4

APPLICATIONS OF SUPERVISED DATA MINING ON DATASET

4.1 Principal Component Analysis (PCA)

Large data sets including multiple samples and variables are accumulated every day by computers and researchers in various fields, such as economy, bio-medical, marketing, etc. Discovering information from a large body of information requires certain techniques for performing analysis on those data sets containing multiple variables. For this reason, Multivariate Analysis (MVAN) is used for analyzing a data set including more than one variable. One of the main difficulties existing in MVAN is the problem of visualizing data that has many variables. In data sets with many variables, it is inevitable that some variables are correlated to each other. The existence of the correlation in the data sets gives rise to redundancy in the data. When we come across a situation like this, it is logical to put the variables that are correlated to each other under a single variable. By means of this way, we can simplify the problem by replacing a group of correlated variables with a single new variable.

Principal Component Analysis is a statistical method utilized for reaching to this methodology. PCA forms a new set of variables that are so-called Principal Components or Dimensions. Each PC comes from a linear combination of the original variables. The obtained PCs are not anymore correlated to each other and it is said that PCs are orthogonal to each other. Therefore, after implementing PCs, redundant information is removed from the data sets. There are a couple of distinct Principal Component Methods. The usage of PCs changes by the types of variables. Apart from PCA, the other methods such as Correspondence Analysis, Multiple Correspondence Analysis are dealing with the data sets with categorical variables. Because of the fact that our data sets are made up of continuous variables, PCA is taken into consideration to build up the new variables.

In section 3.1, under the unsupervised learning system, we dealt with cluster analysis on significant variables of such as GDP, GNI, etc. that have the potency to describe the economic situation of a country to bring the hidden information about which countries are economically more close to each other into the open. By the way, the variables that were utilized in the cluster analysis are called as response variables under the supervised learning system. The procedure is done by using the R packages that are so-called "FactoMiner" and "factoextra" that enable us to obtain PCs directly without using any other function. By means of using this package, the variables that are correlated to each other are grouped together and we made inferences on the clusters of the countries. As for the supervised learning system, for this time, we use the same R package to group the predictor variables that are coming from the data set. In the following sections, data sets are to be illustrated and explained by the tables that are put on both here and Appendix C.

4.1.1 PCA elements

In PCA, as we said above, we make use of the R packages called "FactoMiner" to obtain numeric solutions and principal components (dimensions) and "factoextra" to draw and extract graphical representations of the Principal Components. The function that we are going to use is called "PCA" coming from the package "FactoMiner". After carrying out this function, we get results that are ordered below table and we call them PCA elements. PCA elements contain all the information about PCs, the attributes/predictors/explanatory variables that are at the columns of the data sets, and individuals that are at the rows of the data sets. During the analysis of PCA, we make interpretations of the below outputs and plots.

PCA Elements and Results

In R, we apply for the "FactoMiner" package to reduce the dimension of the data set. After implementing PCA, we can extract below elements coming from the PCA under the "FactoMiner" package.

Table 4.1 Principal Component Analysis Results – Package "FactoMiner"

No	name	description
1	$eig	eigenvalues
2	$var	results for the variables
3	varcoord	coord. for the variables
4	varcor	correlations variables - dimensions
5	varcos2	cos2 for the variables
6	varcontrib	contributions of the variables
7	$ind	results for the individuals
8	indcoord	coord. for the individuals
9	indcos2	cos2 for the individuals
10	indcontrib	contributions of the individuals
11	$call	summary statistics
12	$call$centre	mean of the variables
13	$call$ecart.type	standard error of the variables
14	$call$row.w	weights for the individuals
15	$call$col.w	weights for the variables

4.1.2 Data Standardization

Data Standardization is the most important part to deliver the suitable PCs because the data sets contain variables with different measurements such as kilometers, kilowatts, kilograms, etc. Without making standardization of the variables, PC is negatively affected by this situation. The main aim is to enable the variables to be comparable to each other. Most of the time, variables are treated to have standard deviation that is equal to 1 and mean that is equal to 0.

4.1.3 Principal Component Analysis on Data Sets

The study data sets are composed of 69 distinct variables/attributes. As it is stated, chapter 2 is dedicated to explaining the indicator variables in which 13 variables out of 69 variables are accepted as response variables and the rest of 56 variables are taken as predictor variables. For the supervised learning algorithm part, 56 continuous predictor and 1-time variables are utilized to design classification and regression models. As we said at the beginning, we remove some countries from the data set owing to possessing a multitude of missing values. In order to build up models under the supervised learning algorithm, Principal Component Analysis is performed for both response and predictor variables. Before moving on to PCA, missing values existing in response variables side are treated with imputation method that is so-called Last Observed Variable. Missing cases existing in predictor variables side are treated with both Last Observed Variable for the countries that contain missing cases for certain years and mean imputation technique for the countries that include missing cases for a 16-years period. The reason for not removing the missing cases for a variable throughout the 16 years is because these countries have information about the other variables. Thus, we did not lose information, and also the number of countries that include missing cases throughout 16 years was at most 3. The imputePCA function under the package of missMDA is employed to fill the missing values. The main reason why we use the mean imputation to fill missing cases existing in predictor variables side is that we have missing cases in each country for the same variables and some countries do not even have data. Therefore, instead of getting rid of those countries including missing cases, we use the mean imputation technique to treat missing cases. We go on with the mean imputation on PCA because the package that we use in PCA provides us with the best results with mean imputation. Therefore, we decide to keep going with treating missing cases with the mean imputation technique.

4.1.3.1 Variables used in PCA (Response Variables and Predictor Variables)

The basic aim of the supervised learning algorithm is to deliver models from the training data. In order to derive models, we need to have a response and predictor variables. However, because of the high dimensionality problem, we reduce the dimension of the data sets into fewer numbers of new variables that are so-called Principal Components. After that by means of the new variables, the models are going to be set up based on the PCs. After reducing the dimension of the data sets, PCs are our main interest variables and the comments and inferences are made according to these variables. Therefore, the training and test data sets are made up of PCs. After determining the training data, the predictions based on the test data enable us to assess the accuracy of the models that originate from Classification and Regression Models.

In the visualization part of PCs, the names of the attributes/variables in the columns cover plenty of places in the graphics. In order to avoid any confusion by showing the variables at the same time and graph, the new names for the variables are going to be employed. The new shorted variable names are shown in Appendix C.

In the PC results, we are also getting benefits from the new names in that some of the variable names are so long. They can be found in Appendix C.

4.1.3.2 PCs of Response Variables and Predictor Variables

After renaming the variable for the sake of easiness, we can pass to figuring out PCs. Before moving on to finding PCs, it is important to note that the package "FactoMiner" helps us standardize the response and predictor variables by means of PCA function that provides us with standardized variables with a single code that is written as **scale=TRUE existing in the functional form.**

PCs for Response Variables

Step 1- Correlation Circle

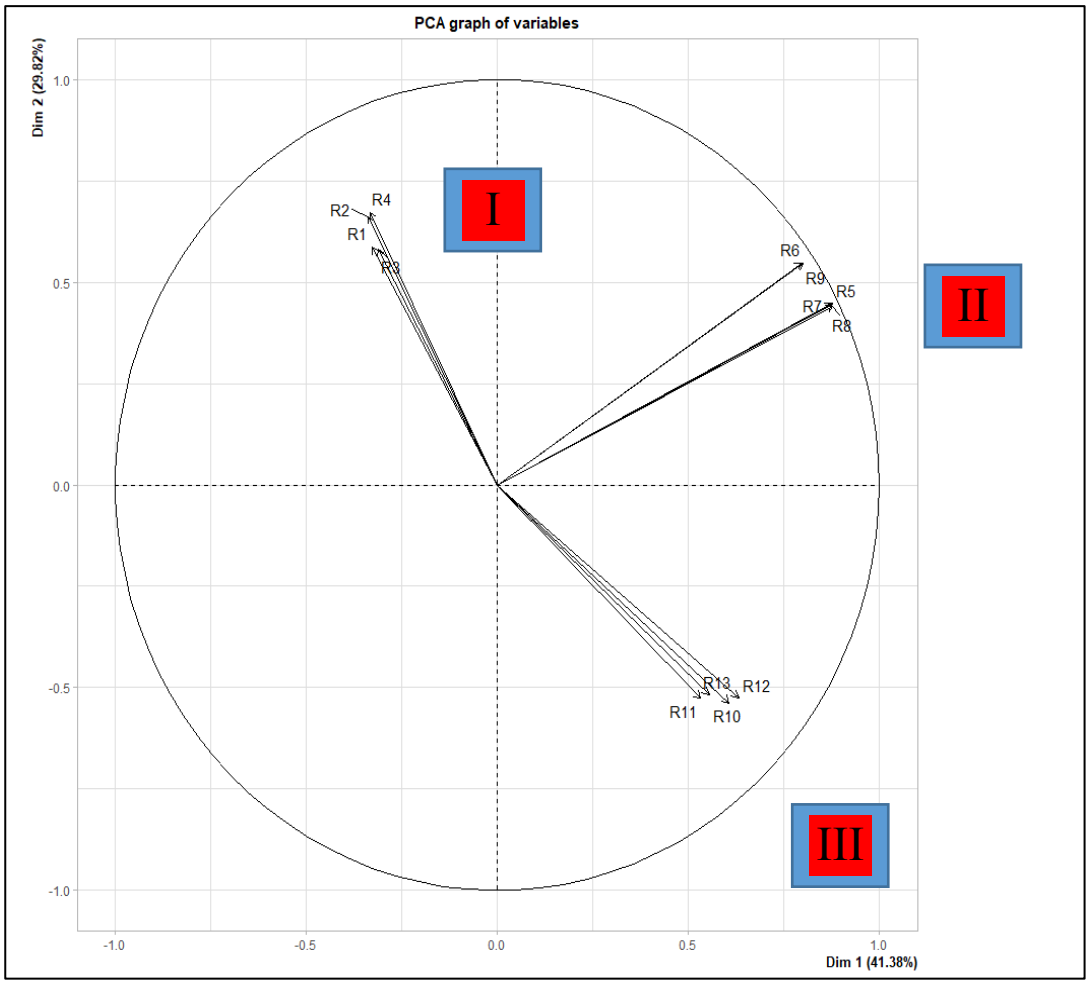


Figure 4.1 Correlation Circle - Response Variables

The correlation circle enables us to see the correlation between the variables/attributes and principal components that are called dimensions in the above graph. The correlation between variables and PCs are known as coordinates of the variables on the PC. It can be seen that the variables R1, R2, R3, and R4 are on the same side of the graph and found together. This indicates that those variables are correlated with each other. R5, R6, R7, R8, and R9 variables are seen together and R10, R11, R12, and R13 are also found together. Those variables that are both found together and look at the same direction are said to be correlated to

each other. Furthermore, we can make an inference that the variables included in I and III are said to be negatively correlated to each other because their directions are opposite to each other.

As for the correlation of those groups of the variables (I, II and III) with the PCs, we can see that 1st group of the variables are more correlated to PC2 (dimension2) because its coordinates take higher values of coordinates on dimension 2 by comparison with the coordinates on dimension 1. For the second group of the variables, they are taking higher coordinate values on dimension 1 as compared to the coordinates on dimension 2, indicating that the 2nd group of the variables are more correlated to PC1. Lastly, the 3rd group of the variables has the same amount of the correlation with dimension 1(PC1) and dimension 2(PC2).

With regards to the amount of variance retained by each PC, while PC1 takes the responsibility of explaining almost 41% of the total variance attributed to the response variables and PC2 has the capability of explaining almost 29% of the total variance attributed to the response variables. Totally, over 70% of the total variance is explained by the two PCs. In order to get the desired level of the results, the threshold value must be at least 70%, meaning the total variance attributed to PCs must exceed the level of 70%. Thus, for this reason, we take one more PC into account. Now, let us look at the eigenvalues that are variances maintained by each PC to make a decision on how many PCs can be added to the analysis.

Note that we put a limit on the number of dimensions on the algorithm to give us at most 5 PCs.

Step 1I- Coordinates (Correlations) of the variables on PCs (Dimensions) and Eigenvalues (Variances retained by each PCs)

Table 4.2 Coordinates (Correlations) of the variables on PCs

Groups	Response Variables	Principal Components				
		PC1	PC2	PC3	PC4	PC5
		Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
I	R1	-0.33	0.59	0.64	0.31	-0.15
	R2	-0.34	0.66	0.55	0.34	0.12
	R3	-0.31	0.58	0.64	-0.34	-0.1
	R4	-0.33	0.67	0.54	-0.29	0.2
II	R5	0.88	0.45	-0.11	0.01	0.03
	R6	0.8	0.55	-0.14	-0.01	-0.07
	R7	0.88	0.45	-0.11	0.01	0.03
	R8	0.88	0.44	-0.11	0.01	0.03
	R9	0.8	0.55	-0.14	-0.01	-0.07
III	R10	0.61	-0.54	0.52	0.04	0.23
	R11	0.53	-0.53	0.61	-0.02	-0.22
	R12	0.63	-0.52	0.5	0.02	0.25
	R13	0.55	-0.52	0.6	-0.04	-0.22

We can see from the above table that variables existing in I and III are also correlated with the PC3. That may indicate that PC3 is also thought of as a significant new variable. In order to know for sure, let us look at the eigenvalues (variances) retained by PC3.

Table 4.3 Eigenvalues (Variances)

Principal Components	Eigenvalue	Percentage of Variance	Cumulative Percentage of Variance
comp1	5.38	41.38	41.38
comp2	3.88	29.82	71.20
comp3	2.75	21.13	92.33
comp4	0.41	3.19	95.52
comp5	0.32	2.44	97.96
comp6	0.13	1.02	98.98
comp7	0.11	0.84	99.82

Table 4.3 (continued)

comp8	**0.02**	**0.14**	**99.96**
comp9	**0.00**	**0.02**	**99.98**
comp10	**0.00**	**0.01**	**99.99**
comp11	**0.00**	**0.01**	**100.00**
comp12	**0.00**	**0.00**	**100.00**
comp13	**0.00**	**0.00**	**100.00**

From the above table, it can be said that Dimension 3 (PC3) is contributing significantly to overall variation with an amount of 2.75 that corresponds to 21.13% increases in the total amount of variation.

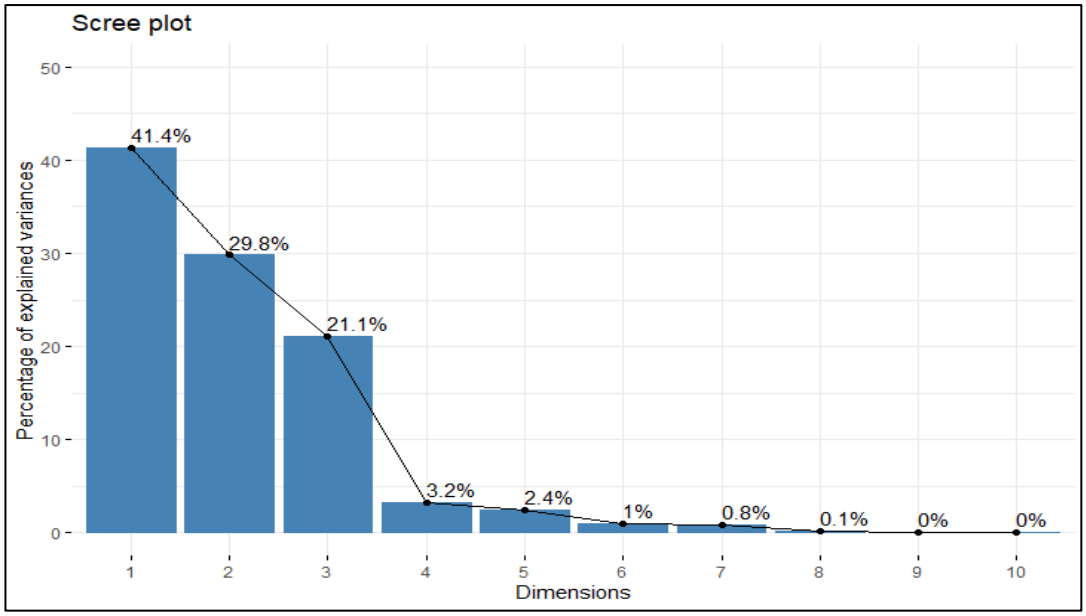


Figure 4.2 Percentage of explained variance retained by each PCs.

Scree plot enables us to make the last decision on how many PCs can be accepted as the new variables. Up to now, our findings indicate that 3 PCs (PC1, PC2, and PC3) must be used to make inferences on the available variables.

Therefore, instead of using 13 variables in the models, we can get benefit from 3 PCs and use them as the new response variables. The following table is the new response data set that is going to be used for building up classification and regression tree(CART) models.

Table 4.4 New Response Variable - 3 PCs

Countries_Years	coord.Dim.1 =PC1	coord.Dim.2 = PC2	coord.Dim.3 = PC3
CHE00	0.379	-1.046	1.502
CHE01	0.92	-2.297	0.079
CHE02	1.034	-2.485	-0.048
CHE03	0.946	-1.944	0.952
CHE04	1.114	-1.974	1.388
CHE05	1.110	-1.671	2.112
CHE06	1.383	-2.025	2.139
CHE07	1.897	-2.986	1.392
CHE08	2.366	-3.802	0.795
CHE09	2.034	-2.821	2.099
CHE10	1.919	-2.19	3.283
CHE11	2.926	-4.034	1.769
CHE12	2.716	-3.462	2.543
CHE13	2.842	-3.529	2.760
CHE14	2.895	-3.608	2.726
CHE15	2.792	-3.374	2.999
JPN00	4.025	1.413	-0.128
•••	•••	•••	•••

Now, let us figure out Principal Components for the original predictor variables.

PCs for Predictor Variables

Step 1- Correlation Circle

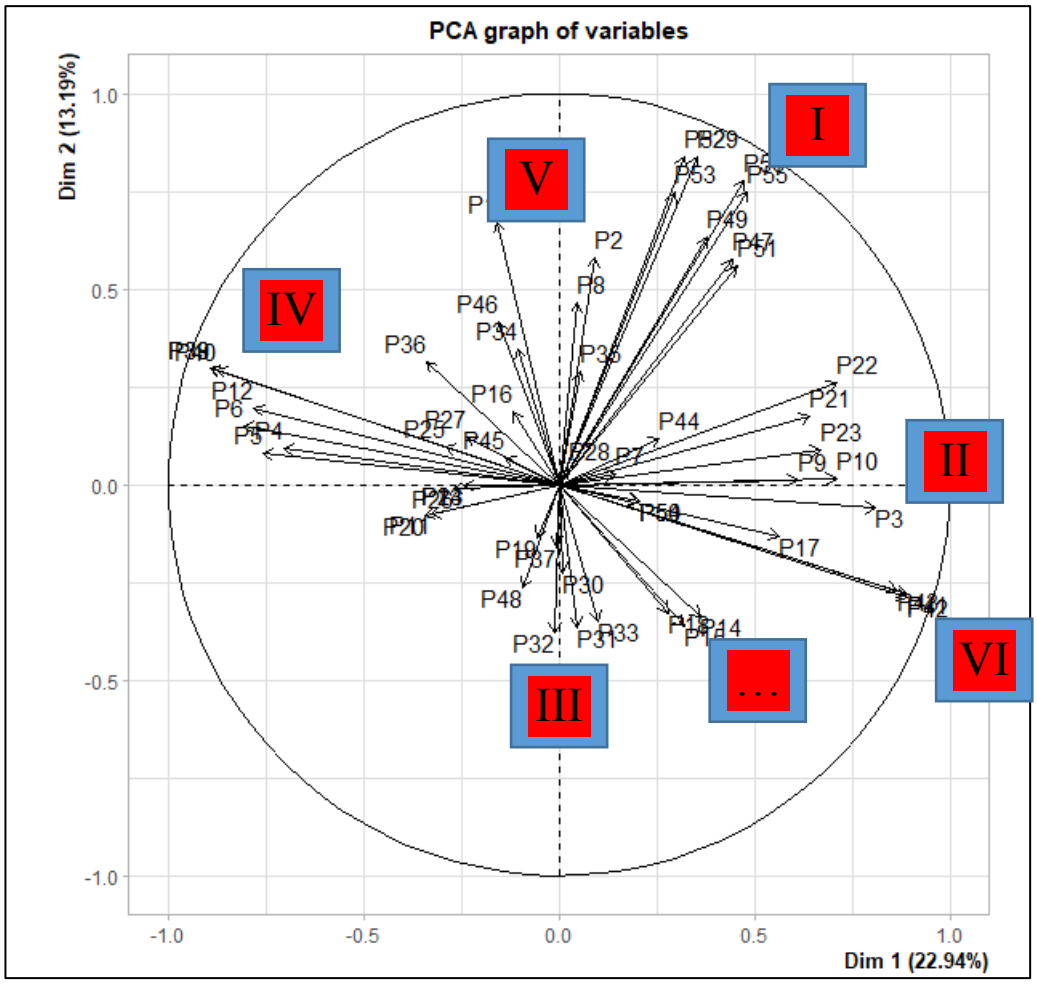


Figure 4.3 Correlation Circle - Predictor Variables

As we stated in the assessment of the PCs for the response variables, the correlation circle provides us with the information about which variables are more correlated to each other and which variables are not correlated to each other and gives us a rough idea about the possible number of dimensions (PCs). For those predictor variables, it seems that predictor variables have the potency to be represented more than at least 5 PCs.

As regards to the dimensions (PCs), the PC1 explains 22% of the total variation attributed to predictor variables included in the PC1.In order for the reader to better

understand which variables are contributing to PC1, we can simply take a look at the variables that are laying throughout the x-axis. For example, the variables placed on group II are the variables that combine under dimension 1. Their linear combinations provide us with some part of the new variable called PC1.

As for dimension 2 (PC2), it is less capable of explaining the variation by comparison with the first one. It explains 13% of the total variation by courtesy of the variables existing in dimension 2 such as P46, P34, and P2 placed under V group. PCs make good work to detect the variables that are correlated to each other because from the "correlation circle scatter plot", we can see that most of the variables are grouped together, indicating that the variables have the potency to be linearly associated with each other. Before moving on to the modeling part, the treatment of the existence of the correlation in the predictor variables is crucial to not come across a multicollinearity problem during the analysis. Moreover, most of the variables are away from the origin. This is an indication that those variables are well represented on the factor map.

Note that we put a limit on the number of dimensions on the algorithm to give us at most 20 PCs

Step 1I- Coordinates (Correlations) of the variables on PCs (Dimensions) and Eigenvalues (Variances retained by each PCs)

Table 4.5 Coordinates of the variables on PCs (Ps)

Original Predictor Variables (OP)	Principal Components												
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Dim. 5	Dim. 6	Dim. 7	Dim. 8	Dim. 9	Dim.10	Dim.11	Dim.12	Dim.13
OP1	-0.16	0.67	-0.03	-0.004	-0.24	0.4	-0.07	-0.04	-0.21	-0.1	0.05	0.19	0.04
OP2	0.09	0.58	-0.21	0.28	-0.09	-0.08	-0.46	-0.01	0.14	-0.1	0.08	0.23	0.1
OP3	0.81	-0.06	0.18	-0.16	-0.18	-0.02	0.02	-0.19	-0.03	-0.01	0.01	-0.21	0.14
OP4	-0.7	0.09	-0.03	0.09	0.21	-0.17	0.07	0.24	-0.17	0.15	0.06	-0.11	-0.13
OP5	-0.76	0.08	-0.14	0.06	0.23	-0.23	-0.16	0.24	0.05	0.15	0.11	-0.07	0.06
OP6	-0.8	0.15	-0.11	0.01	0.31	0	0.02	0.13	-0.09	0.1	0.03	0.09	-0.21
OP7	0.14	0.03	-0.07	0.33	-0.43	-0.04	0.16	-0.28	0.33	0.09	-0.03	-0.1	0.27
OP8	0.05	0.47	-0.22	0.27	-0.09	-0.18	-0.46	-0.03	0.26	-0.12	0.13	0.25	0.21

Table 4.5 (continued)

OP9	**0.61**	**0.01**	**0.3**	**0.42**	**0.05**	**-0.21**	**-0.05**	**0**	**-0.14**	**0.02**	**-0.06**	**0.03**	**-0.28**
OP10	**0.71**	**0.02**	**0.21**	**0.14**	**0.12**	**-0.21**	**-0.14**	**-0.13**	**-0.13**	**0.01**	**-0.07**	**0.03**	**-0.31**
OP11	**-0.33**	**-0.08**	**-0.18**	**0.43**	**0.59**	**0.21**	**-0.11**	**-0.32**	**-0.13**	**-0.13**	**-0.02**	**-0.12**	**0.14**
OP12	**-0.78**	**0.2**	**0.08**	**-0.17**	**0.06**	**0.03**	**0.12**	**-0.12**	**0.13**	**-0.13**	**-0.15**	**0.05**	**-0.13**
OP13	**-0.24**	**0**	**0.24**	**0.76**	**-0.18**	**0.04**	**0.18**	**0.13**	**-0.24**	**0.25**	**0.23**	**-0.03**	**0**
OP14	**0.36**	**-0.34**	**0.45**	**-0.06**	**0.35**	**0.4**	**-0.04**	**0.38**	**0.04**	**-0.05**	**0.15**	**0.06**	**0.04**
OP15	**0.32**	**-0.36**	**0.35**	**-0.18**	**0.33**	**0.48**	**0**	**0.37**	**0.13**	**-0.08**	**0.13**	**0.07**	**0.1**
OP16	**-0.12**	**0.19**	**0.3**	**0.25**	**-0.14**	**0.4**	**0.08**	**-0.09**	**-0.07**	**0.15**	**-0.03**	**0.2**	**-0.01**
OP17	**0.56**	**-0.13**	**0.2**	**-0.1**	**0.05**	**-0.01**	**-0.15**	**0.11**	**-0.13**	**0.22**	**-0.33**	**0.02**	**0.34**
OP18	**0.28**	**-0.33**	**0.4**	**-0.04**	**0.31**	**0.48**	**0.04**	**0.4**	**0.13**	**-0.06**	**0.13**	**0.11**	**0.1**
OP19	**-0.05**	**-0.14**	**0.1**	**0.27**	**-0.02**	**-0.23**	**-0.1**	**0.2**	**-0.1**	**0.59**	**-0.38**	**0.07**	**0.31**
OP20	**-0.34**	**-0.08**	**-0.19**	**0.41**	**0.59**	**0.2**	**-0.12**	**-0.29**	**-0.11**	**-0.14**	**0.01**	**-0.13**	**0.13**
OP21	**0.64**	**0.17**	**0.01**	**-0.37**	**0.02**	**0.18**	**0.06**	**-0.31**	**0.03**	**0.28**	**0.21**	**-0.02**	**0.03**
OP22	**0.71**	**0.26**	**0**	**-0.29**	**0.11**	**0.13**	**-0.02**	**-0.18**	**0.02**	**0.29**	**0.19**	**0.07**	**-0.09**
OP23	**0.67**	**0.09**	**0.14**	**-0.36**	**0.05**	**0.22**	**-0.05**	**-0.26**	**-0.02**	**0.33**	**0.19**	**0.03**	**-0.03**
OP24	**-0.24**	**0**	**0.24**	**0.76**	**-0.18**	**0.04**	**0.18**	**0.13**	**-0.24**	**0.25**	**0.23**	**-0.03**	**0**
OP25	**-0.29**	**0.1**	**0.16**	**0.28**	**-0.09**	**0.26**	**0.26**	**0.01**	**0.61**	**0.06**	**-0.18**	**-0.16**	**-0.06**
OP26	**-0.27**	**-0.01**	**-0.28**	**0.07**	**-0.08**	**-0.24**	**-0.32**	**0.23**	**0.24**	**-0.07**	**0.21**	**0.05**	**0.3**
OP27	**-0.23**	**0.12**	**0.09**	**0.31**	**-0.15**	**0.33**	**0.28**	**-0.03**	**0.6**	**0.03**	**-0.17**	**-0.12**	**-0.06**
OP28	**0.02**	**0.04**	**0.12**	**0.27**	**-0.33**	**0.16**	**0.26**	**-0.05**	**0**	**-0.28**	**0.36**	**-0.05**	**0.05**
OP29	**0.35**	**0.84**	**-0.22**	**0.08**	**0.02**	**0.15**	**0.02**	**0.07**	**-0.04**	**0**	**-0.02**	**0**	**-0.09**
OP30	**0.01**	**-0.23**	**0.23**	**0.29**	**-0.06**	**0.1**	**0.14**	**-0.07**	**0.11**	**0.13**	**-0.21**	**0.43**	**-0.11**
OP31	**0.05**	**-0.37**	**-0.82**	**0.07**	**0**	**0.18**	**-0.03**	**0.05**	**0.07**	**0.19**	**0.05**	**0.16**	**-0.11**
OP32	**-0.01**	**-0.38**	**-0.77**	**0.14**	**-0.11**	**0.14**	**0.06**	**0.1**	**0**	**0.21**	**0.09**	**0**	**-0.07**
OP33	**0.1**	**-0.35**	**-0.78**	**0.06**	**0.04**	**0.21**	**-0.04**	**0.04**	**0.08**	**0.16**	**0.02**	**0.23**	**-0.13**
OP34	**-0.11**	**0.35**	**0.72**	**0.09**	**0.19**	**-0.22**	**-0.25**	**-0.12**	**0.16**	**0.16**	**0.09**	**-0.02**	**-0.07**
OP35	**0.05**	**0.29**	**0.43**	**-0.1**	**0.34**	**-0.11**	**-0.36**	**-0.21**	**0.31**	**0.22**	**-0.03**	**0.13**	**-0.08**
OP36	**-0.34**	**0.31**	**0.67**	**0.12**	**-0.05**	**-0.23**	**-0.04**	**0**	**-0.03**	**-0.03**	**0.2**	**-0.2**	**0.05**
OP37	**-0.01**	**-0.16**	**-0.34**	**-0.26**	**-0.05**	**0.04**	**0.03**	**-0.04**	**0.01**	**0.33**	**0.19**	**-0.49**	**0.08**
OP38	**-0.89**	**0.3**	**0.08**	**-0.23**	**0.02**	**0.08**	**0.06**	**-0.04**	**-0.02**	**0.08**	**0**	**0.07**	**0.01**
OP39	**-0.89**	**0.3**	**0.07**	**-0.18**	**0.03**	**0.07**	**0.07**	**-0.02**	**-0.05**	**0.1**	**0**	**0.06**	**0.02**
OP40	**-0.87**	**0.29**	**0.08**	**-0.27**	**-0.01**	**0.08**	**0.06**	**-0.05**	**-0.01**	**0.05**	**-0.02**	**0.07**	**0.02**
OP41	**0.89**	**-0.28**	**-0.04**	**0.25**	**-0.01**	**-0.08**	**-0.06**	**0.04**	**0.02**	**-0.12**	**-0.02**	**-0.01**	**-0.02**
OP42	**0.89**	**-0.29**	**-0.06**	**0.19**	**-0.03**	**-0.07**	**-0.06**	**0.02**	**0.04**	**-0.12**	**-0.01**	**-0.04**	**-0.02**
OP43	**0.87**	**-0.27**	**-0.04**	**0.31**	**0.01**	**-0.09**	**-0.06**	**0.06**	**0**	**-0.1**	**-0.01**	**0**	**-0.03**
OP44	**0.26**	**0.12**	**0.24**	**-0.37**	**-0.56**	**-0.15**	**0.21**	**0.31**	**-0.1**	**-0.03**	**-0.13**	**0.15**	**-0.1**
OP45	**-0.14**	**0.07**	**0.02**	**-0.25**	**-0.32**	**-0.3**	**-0.21**	**0.35**	**0.22**	**0.07**	**0.35**	**-0.01**	**-0.07**
OP46	**-0.15**	**0.42**	**-0.07**	**-0.18**	**-0.17**	**0.2**	**0.13**	**0.04**	**-0.33**	**-0.18**	**-0.17**	**0.08**	**0.36**
OP47	**0.44**	**0.58**	**-0.12**	**0.22**	**0.15**	**-0.21**	**0.12**	**0.21**	**-0.05**	**-0.02**	**-0.11**	**-0.08**	**-0.01**
OP48	**-0.09**	**-0.26**	**0.27**	**0.03**	**-0.19**	**0.26**	**-0.03**	**-0.25**	**-0.13**	**0.13**	**0.24**	**0.19**	**0.18**

Table 4.5 (continued)

OP49	0.38	0.63	-0.1	0	0.15	0.17	-0.04	0.25	0.09	0.07	-0.13	-0.22	0
OP50	0.21	-0.04	-0.07	-0.06	0.38	-0.51	0.62	-0.08	0.1	-0.01	0.16	0.25	0.16
OP51	0.45	0.56	-0.13	-0.03	0.33	-0.1	0.28	0.19	0.13	0.06	-0.04	-0.08	0.08
OP52	0.32	0.84	-0.2	0.14	-0.03	0.11	0.04	0.03	-0.05	0	0.05	0.01	-0.07
OP53	0.3	0.75	-0.18	0.1	0.02	0.04	0.12	-0.04	0	-0.03	0.13	0.06	-0.08
OP54	0.21	-0.04	-0.06	-0.06	0.38	-0.51	0.62	-0.08	0.1	-0.01	0.16	0.25	0.16
OP55	0.48	0.75	-0.12	0.02	0.09	0.17	0.09	0.08	-0.06	0.07	0.01	-0.01	0.08
OP56	0.47	0.78	-0.17	0.01	0.13	0.1	0.12	0.13	-0.04	0.03	-0.03	-0.04	0.04

The coordinates of the variables are ordered in the above table. Coordinates are called either as the coefficient of the variables or correlation of the variables with the dimensions (PCs).

Table 4.6 Eigenvalues (Variances)

Principal Components	Eigenvalue	Percentage of Variance	Cumulative Percentage of Variance
comp1	12.85	22.94	22.94
comp2	7.38	13.19	36.13
comp3	4.83	8.62	44.75
comp4	3.70	6.61	51.37
comp5	2.90	5.18	56.55
comp6	2.81	5.02	61.57
comp7	2.23	3.99	65.56
comp8	1.83	3.26	68.82
comp9	1.73	3.09	71.90
comp10	1.51	2.70	74.60
comp11	1.31	2.34	76.95
comp12	1.20	2.15	79.10
comp13	1.11	1.98	81.07
comp14	1.05	1.88	82.95
comp15	0.96	1.72	84.67
comp16	0.86	1.54	86.22
comp17	0.82	1.46	87.68
comp18	0.77	1.38	89.06
comp19	0.65	1.16	90.22
comp20	0.62	1.11	91.33
comp21	0.60	1.07	92.40
comp22	0.54	0.97	93.37
comp23	0.48	0.86	94.23
comp24	0.40	0.72	94.95

Table 4.6 (continued)

comp25	**0.35**	**0.63**	**95.58**
comp26	**0.32**	**0.57**	**96.15**
comp27	**0.27**	**0.49**	**96.64**
comp28	**0.27**	**0.49**	**97.13**
comp29	**0.22**	**0.40**	**97.53**
comp30	**0.19**	**0.34**	**97.86**
comp31	**0.17**	**0.30**	**98.16**
comp32	**0.16**	**0.28**	**98.45**
comp33	**0.13**	**0.23**	**98.67**
comp34	**0.10**	**0.18**	**98.86**
comp35	**0.10**	**0.17**	**99.03**
comp36	**0.09**	**0.16**	**99.19**
comp37	**0.08**	**0.15**	**99.34**
comp38	**0.07**	**0.12**	**99.46**
comp39	**0.06**	**0.11**	**99.57**
comp40	**0.06**	**0.10**	**99.67**
comp41	**0.05**	**0.08**	**99.75**
comp42	**0.04**	**0.06**	**99.82**
comp43	**0.03**	**0.06**	**99.88**
comp44	**0.02**	**0.04**	**99.92**
comp45	**0.01**	**0.03**	**99.94**
comp46	**0.01**	**0.02**	**99.97**
comp47	**0.01**	**0.01**	**99.98**
comp48	**0.01**	**0.01**	**99.99**
comp49	**0.00**	**0.01**	**100.00**
comp50	**0.00**	**0.00**	**100.00**
comp51	**0.00**	**0.00**	**100.00**
comp52	**0.00**	**0.00**	**100.00**
comp53	**0.00**	**0.00**	**100.00**
comp54	**0.00**	**0.00**	**100.00**
comp55	**0.00**	**0.00**	**100.00**
comp56	**0.00**	**0.00**	**100.00**

In the PCA for the response variable, we have determined the threshold value as 90%, meaning we take the principal components whose total variance is exceeding 90%. As for PCA for the predictor variables, we determined the threshold value as 80% to avoid taking more PCs into consideration. When we look at the above table of eigenvalues corresponding to PCs, 81% of the total variation is explained by the 13 PCs. Therefore, our new predictor variables are made up of those PCs. Now, let us look at the scree plot of the eigenvalues.

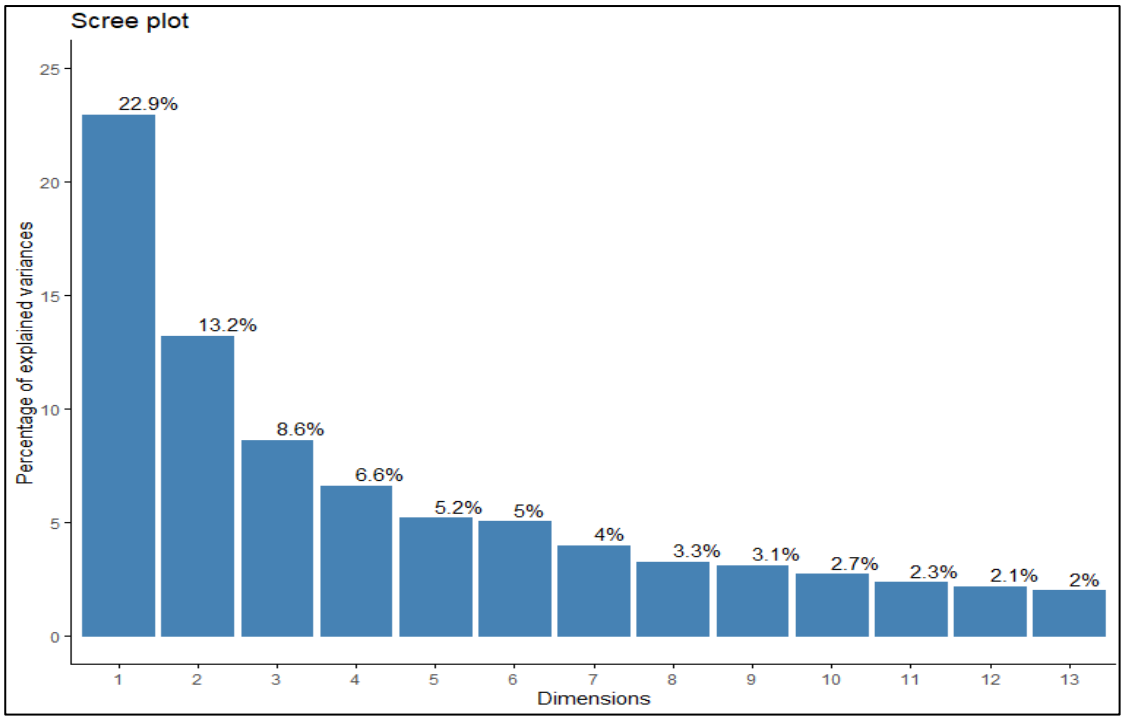


Figure 4.4 Percentage of explained variance retained by each PCs.

It is clear from the above scree plot; we are also taking the PCs that are contributing less to the explained variation into account. The choice of how many PCs to be included in the analysis is left to the analysts. It is suggested in the literature that the PCs whose eigenvalues are greater than 1 must be paid attention according to the rule that is called as "eigenvalues-greater-than-one rule". *(Kiaser, 1960)*

Therefore, instead of using 56 variables in the models, we can get benefit from 13 PCs and use them as the new predictor variables. The following table is the new predictor data set that is going to be used for building up classification and regression tree (CART) models. Those PCs are called as features in tree-based models. The above table is the new predictor variables that are used instead of the original ones.

Table 4.7 New Predictor Variables – 13 PCs

Countries_ Years	coord.Dim.1 = PC1	coord.Dim.2 = PC2	coord.Dim.3 = PC3	coord.Dim.4 =PC4	coord.Dim.5 = PC5	…	coord.Dim.13 =PC13
CHE00	2.897	0.079	1.667	-1.467	0.637	…	-0.12
CHE01	2.740	0.059	1.995	-1.099	0.158	…	-0.348
CHE02	2.826	0.03	1.804	-1.264	-0.006	…	-0.399
CHE03	2.987	0.069	1.441	-1.392	0.034	…	-0.365
CHE04	3.249	0.022	1.324	-1.383	0.462	…	-0.063
CHE05	3.481	0.104	1.297	-1.301	0.907	…	0.326
CHE06	3.647	0.604	1.521	-1.129	0.883	…	0.206
CHE07	3.674	0.678	1.653	-1.184	0.835	…	0.184
CHE08	3.653	0.536	1.758	-1.064	0.804	…	0.356
CHE09	4.016	0.645	1.447	-1.67	0.429	…	0.229
CHE10	4.092	0.629	1.293	-1.273	1.069	…	0.829
CHE11	4.135	1.024	1.572	-1.158	0.431	…	0.4
CHE12	4.454	0.933	1.511	-1.578	0.673	…	0.523
CHE13	4.400	0.48	1.525	-1.56	0.926	…	0.76
CHE14	4.240	0.847	1.428	-1.461	0.14	…	0.175
CHE15	4.736	1.609	1.255	-1.605	0.753	…	0.521
JPN00	4.238	3.463	0.606	-0.487	-1.246	…	-1.087
…	…	…	…	…	…	…	…

So, in the end, what can we understand from this analysis? In order to understand what exactly PC does, we can take a look at the mathematical background of this methodology. Each Principal Component is nothing but the linear combination of the original variables. For example, in Table 4.5**,** we get PCs for the original predictor values. Dimension one that is also so-called Principal Component 1 is made up of a linear combination of the 56 original variables. In order to see the formula for PC1, it is given in detail below;

$$PC1 = -0.16OP1 + 0.09OP2 + 0.81OP3 - 0.70OP4 - 0.76OP5 + \cdots \ldots . . + 0.47OP56$$

$OPs\ are\ corresponding\ to\ original\ variables.$

$The\ coefficients\ of\ OPs\ in\ the\ above\ linear\ equations\ are\ also$

$known\ as\ either\ coordinates\ or\ correlations$

$with\ respect\ to\ Principal\ Component\ 1.$

Moreover, the best property of PCs is that PCs are orthogonal to each other, meaning they are independent of each other.

Apart from PCs, there are also some other methods to reduce the dimension of the data set. Because of not having any categorical variable in the data sets, we apply for Principal Component Analysis to go to the decrement in the dimension of the predictor variables. In the case that we have categorical variables in the data set, we need to go with some other techniques such as Correspondence Analysis. Principal Component Analysis is so useful in the situations that we derive statistical models because the variables that are used in the modeling part can be correlated to each other. This situation can cause the modeling to give nonmeaningful results. To remove the correlation that is so-called multicollinearity between the predictor variables, someone can apply for dimension reduction techniques, or someone can throw some of the variables out of the models to avert the correlation problem.

END OF DATA MINING-PRINCIPAL COMPONENT ANALYSIS

4.2 Classification

From now on, we pay attention to the modeling of the variables that are coming from PCA. In order to set off modeling, first of all, we begin to derive tree-based models, and then the models based on the regression is going to be handled in 4.3.

Classification is one of the modeling data mining techniques under supervised learning. The main aim of the classification is to find out probabilistic and graphical models to explain the variation in the response variable. Classification is based on the logic of prediction of a certain outcome/response based on given input/predictor variables. For the sake of prediction of the outcome, the algorithms used in classification use training data sets to design the models. The algorithm struggles for discovering the relationships between the predictor variables/attributes/explanatory/independent variables and outcomes/responses/dependent variables. The models are set up by means of different algorithms.

In the clustering analysis, the clusters were newly formed. However, most of the time, groups of interest items/objects are already known. The need to assign new objects to the related groups or clusters is crucial to maintain the categorization. The terms of classification bring into the open to accomplish this task. Classification is a way of organizing and categorizing the variables into distinct clusters where the data is already labeled. The process of correctly classifying the variables into different groups and deriving models are needed to be dealt with the following steps;

I. Model Construction is performed based on training data sets.
II. Model Evaluation is carried out based on test data sets.
III. Model Accuracy and Model Comparisons.

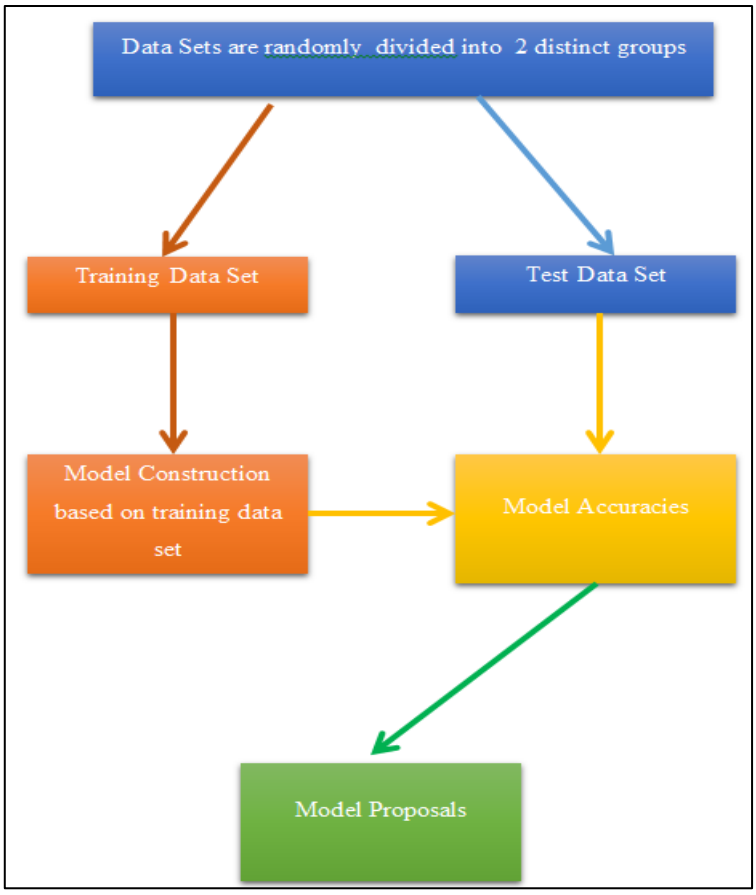


Figure 4.5 Modeling Steps

The most used, important and well-known classification techniques are ordered below;

- **Decision Trees (DT)**
- **Random Forest (RF)**
- **Bayesian Classification (BC)**
- **Support Vector Machines (SVM)**
- Artificial Neural Network (ANN)
- K-Nearest Neighbor (KNN)
- Regression Trees (RT)

Note that: In our study, we mainly focus on Decision Trees, Random Forests, BC, and SVM to produce some useful tree-based models.

4.2.1 Determination of the Target Variable by using the new response variables originating from Principal Component Analysis.

In chapter 3, we dealt with the clustering of the countries by their economical levels. For the sake of bringing the hidden information into the open, percentage and numeric data sets were used for detecting the clusters of the countries. Our data is in the shape of longitudinal data and in order for the reader and ourselves to make good interpretations and inferences on the groups of the countries and build up well-designed visualization of the clusters of the countries. The year effect was removed and for each country, the averages of the economic values of 16 years periods were taken into consideration. After that, we performed cluster analysis and we suggested some useful clusters developing out of four different cluster techniques. Before CART modeling, we had determined the new response and new predictor variables. Classification models are mostly designed to predict categorical variables. However, there is a problem. Our new response variables that are coming from PCs are made up of numeric values and the year effect is included. We thought that instead of using the new response variables as the target variables in Classification models, we can take the clusters of the countries into account and turn them into new response variables. However, the clusters that we have found in section 3.1 were not formed by taking the year effect into consideration. Therefore,

- The new categorical response variables are to be formed by means of the new response variables attributable to Principal Components in Classification Models. New categorical response variables are designed by means of clustering technique without going into detail as we did in section 3.1.
- In section 4.3, we implement panel data analysis. The new response numeric variables that are Principal Components are taken as the new dependent variables. Therefore, the modelings that use both categorical new

response variables and numeric new response variables are constructed in our study.

- For the evaluation of the performance of classification models, the overall classification accuracy and the Kappa coefficient are used.
  (Soleymani et al,2015,s.6)

4.2.1.1 New Categorical Response Variables coming from the new response variables that are Principal Components for 13 original response variables

- According to the results of the ***k-means*** algorithm, for the two data sets that are numeric and percentage, k=4 can be suggested as the number of groups/categories for the groups among the different values of k.
- According to the results of the ***k-medoids*** algorithm, for the two data sets that are numeric and percentage, k=4 can be suggested as the number of groups/categories for the groups among the different values of k.
- According to the results of the ***fuzzy (fanny***) algorithm, for the two data sets that are numeric and percentage, k=4 can be suggested as the number of groups/categories for the groups among the different values of k.
- According to the results of the ***hiearchical*** algorithm, for the two data sets that are numeric and percentage, k=4 can be suggested as the number of groups/categories for the groups and ward. D2 method is preferred to design clusters among the different values of k.

Someone asks the reason of why we take these values as indexes in order to make clusters. We can take those as indexes because even if we did not think the year effect in section 3.1, the suggested number of the clusters and methods provide us with similar results and give us prior knowledge about the possible numbers for the k.

4.2.2 Decision Tree Models (DTM)

In the previous part, we said that we need to find the new categorical variables that are attributed to the new continuous response variables that are coming from the Principal Component Analysis. After we carry out the implementation of clustering techniques for the variables of the new response variables based on the indexes that are determined as the best ones in section 3.1, the following data set with the new categorical variables are obtained.

Table 4.8 New data set to be used for CART Models & dimension 1296 x 17

Countries_ Years	R1	R2	R3	R4	P1=PC1	P2=PC2	…	P13=PC13
CHE00	GroupC	GroupA	GroupA	GroupA	2.897	0.079	…	-0.120
CHE01	GroupC	GroupA	GroupA	GroupA	2.740	0.059	…	-0.348
CHE02	GroupC	GroupA	GroupA	GroupA	2.826	0.030	…	-0.399
CHE03	GroupC	GroupA	GroupA	GroupA	2.987	0.069	…	-0.365
CHE04	GroupC	GroupA	GroupA	GroupA	3.249	0.022	…	-0.063
CHE05	GroupC	GroupA	GroupA	GroupA	3.481	0.104	…	0.326
CHE06	GroupC	GroupA	GroupA	GroupA	3.647	0.604	…	0.206
CHE07	GroupC	GroupA	GroupA	GroupA	3.674	0.678	…	0.184
CHE08	GroupC	GroupA	GroupA	GroupB	3.653	0.536	…	0.356
CHE09	GroupC	GroupA	GroupA	GroupA	4.016	0.645	…	0.229
CHE10	GroupC	GroupA	GroupA	GroupA	4.092	0.629	…	0.829
CHE11	GroupC	GroupA	GroupA	GroupA	4.135	1.024	…	0.400
CHE12	GroupC	GroupA	GroupA	GroupA	4.454	0.933	…	0.523
CHE13	GroupC	GroupA	GroupA	GroupA	4.400	0.480	…	0.760
CHE14	GroupC	GroupA	GroupA	GroupA	4.240	0.847	…	0.175
CHE15	GroupC	GroupA	GroupA	GroupA	4.736	1.609	…	0.521
JPN00	GroupC	GroupA	GroupA	GroupA	4.238	3.463	…	-1.087
…	…	…	…	…	…	…	…	…

- The groups under R1, R2, R3, and R4 are coming from k-means, k-medoids, fuzzy, and hierarchical clustering algorithms based on the best indexes that we have found in clustering part of the study, respectively. Under the decision tree algorithm, there are two important ways to determine the most important predictor variables that have an important

effect on the response variable. They are **Information Gain Index and Gini Indexes**.

a.) Information Gain

The first measurement to figure out the best attribute that has the potency to explain the variation in the response variables is called as Information Gain. As it is understood from the name, the main aim is to find the predictor that contains much valuable information for the response variable (target value). The best attribute is also known as the root node in DT. This provides us with the remaining entropy when we split the datasets along with the feature values. Then, we subtract this value from the originally calculated entropy of the datasets to observe how much splitting of this attribute reduces the original entropy that gives the information gain of a feature. The formulation of information gain measurement is as following;

Formulation:

$$\boldsymbol{Information\ Gain\ (Feature\ )} \\ \boldsymbol{= Entropy\ (\ Dataset = Parent\ ) - Entropy\ (\ Feature)}$$

The feature with the largest information gain should be used as the root node to start to build the decision tree. Before moving on to what Gini index does to determine which features to be chosen as the most important ones, we need to talk about the definition of the entropy.

Split Criteria - Entropy: Entropy is used to gauge the impurity or randomness of a dataset.

$$\boldsymbol{Entropy\ (x) = -\sum (P(x = k) * \log_2 P(x = k))}$$

Where P (x = k) is the probability that a target feature takes a specific value, k.

- **2 classes: Max entropy is 1.**
- **4 classes: Max entropy is 2.**
- **8 classes: Max entropy is 3.**

- **16 classes: Max entropy is 4**

b.) Gini Index

Gini Index or Gini impurity gauges the degree of probability of a particular variable being wrongly classified when it is randomly chosen. If all elements pertain to a single class, then it can be called pure. The value of the Gini index varies between 0 and 1, where 0 represents that all elements are randomly distributed across various classes. A Gini score of 0.5 displays equally distributed elements into some classes.

Formulation: Split Criteria – Gini Index

$$\boldsymbol{Gini = 1 - \sum_{i=1}^{n} p_i^2}$$

Where p_i is the probability of an object being classified to a particular class. While setting up DT, we prefer choosing the attribute/feature with the smallest Gini index as the root node.

Note that: We built up models based on R1, R3, and R4 response variable and we saw that none of the models obtained by using R1, R3, and R4 provide us with the desired accuracy levels. This indicated that k-means, fuzzy, and hierarchical clustering algorithms are less capable of designing the well- designed clusters of the countries as compared to the k-medoids algorithm. Thus, we are just adding the models to the study based on R2 to the analysis. The other models can be reached in Appendix D.

4.2.2.1 Modeling Based on R2 categorical response variable

We set up models based on the second categorical response variables (R2) that are made up of multi classes and coming from the k-medoids clustering algorithm. Moreover, we are going to use new predictor variables that are found by courtesy of the Principal Components. We know for sure that the predictor variables are independent and orthogonal to each other.

- **Modeling by Gini Index**

Step I: Building up Models – DTM7

The 7th and 8th models are to be set up with the gini index.

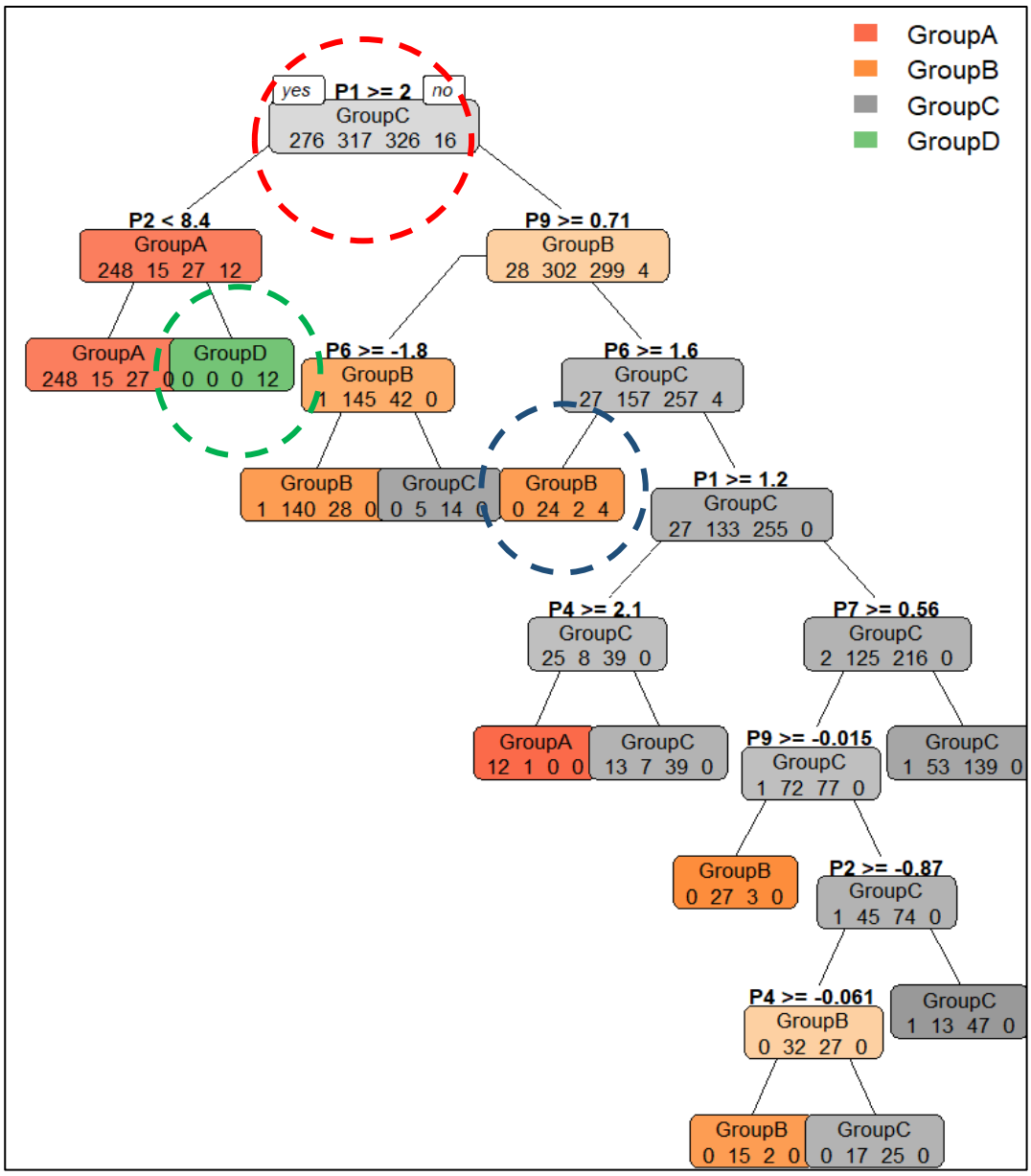


Figure 4.6 Decision Tree Model-7(DTM7) & R2 & Gini &cp=0.01

Let us make inferences on the decision tree model 7 (DTM7);

- According to the results based on Gini index, the most important variable is the same as that of the results of information gain that is P1. There is no change in the rule scores.
- After splitting of the root node is performed on the basis of the rule of P1>=2, the next splittings are carried out by means of the variables that are P2 and P9 that are the same features as the splittings that are formed in the models that come out by using the R1 categorical response variable. On the side of the P2 separation part, when P2<8.4, then most of the countries in group D are separated from the tree and become a terminal node/leaf node shown in green dashed circle. By looking at the numbers that are placed on each node under the group name, we can see that at the beginning we have 16 items/countries of group D shown in the red dashed circle in the figure. Then, 12 items of group D shown in the green dashed circle are separated successfully out of 16 items. The rest of group D items/countries are grouped as B that is shown with the blue dashed line in the figure. P1 and P2 variable seem to be enough in order to separate the group D from the other groups by Figure 4.6. The same interpretation can be also made for the other features and classes (group A, group B and group C).
- The classification of the group A carried out by the predictor variables is also completed earlier than that of group B and group C because the bottom leaf nodes are composed of those groups and more branching is required for discrimination of those groups from the other groups. We need to pay attention to the fact that classifications of the groups by using predictor variables are not accurately grouped. We discuss this issue by giving a confusion matrix in order to have an idea about the predicted power of the features.

Now, let us find the best Complexity Parameter (CP) value and derive the models by the best CP value again.

Table 4.9 Complexity Parameters attributed to DTM7

CP	nsplit	rel error	xerror	xstd
0.3678	0	1.0000	1.0279	0.0236
0.1642	1	0.6322	0.6847	0.0250
0.0361	2	0.4680	0.4746	0.0232
0.0197	3	0.4319	0.4434	0.0228
0.0148	4	0.4122	0.4450	0.0228
0.0107	9	0.3383	0.4466	0.0228
0.0100	11	0.3169	0.4466	0.0228

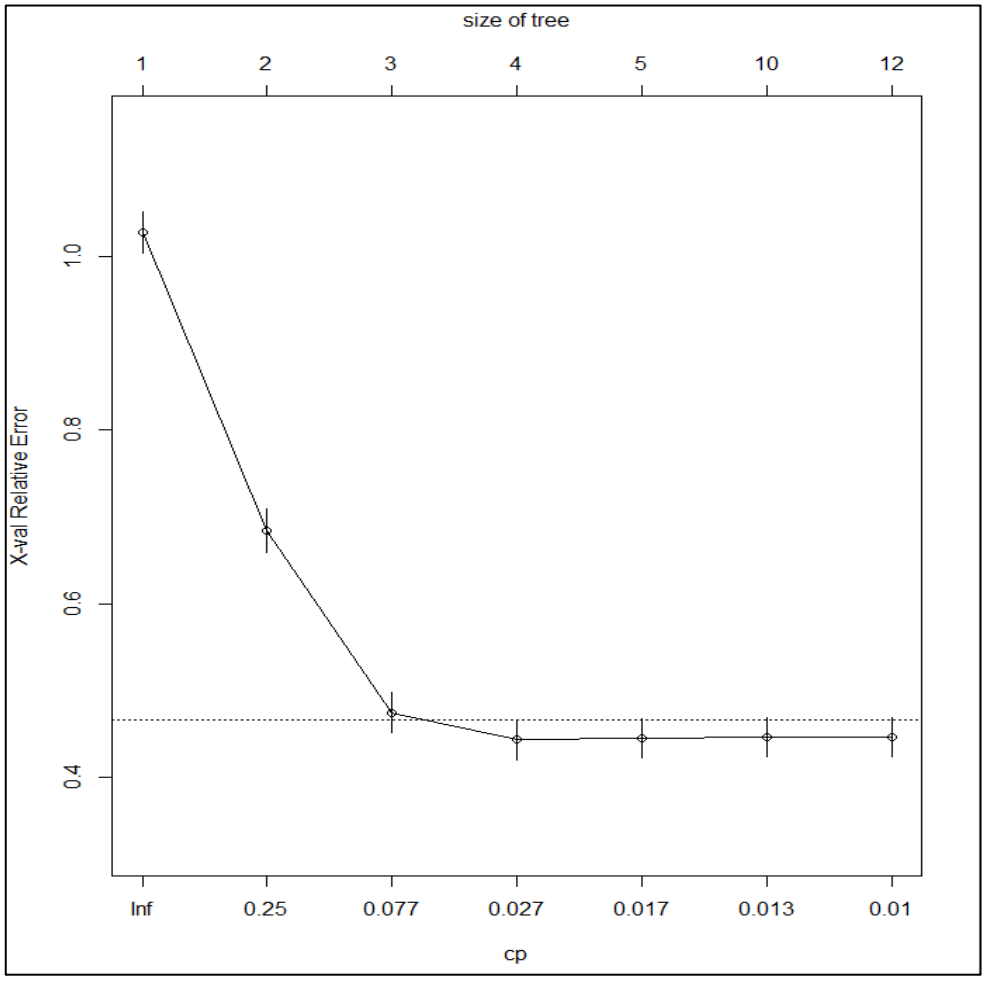


Figure 4.7 Complexity Parameters attributed to DTM7

From the above figure and table, it can be understood that the improvements are decreased after the tree is split either 3 or 4 times because there occurs the minimum improvement to decrease the relative error when the value for cp is

decreased. In order to sidestep crowded trees that have more branching, more splits, we can prune the trees when the improvement of the trees is less in case of splitting. Moreover, we can determine cp value for this modeling as 0.0148 which is so close to the default value of 0.01 and the cp value which is found by using the information gain. Someone can choose the 0.077 value as the best value for cp by looking at the figure. However, by taking the overfitting problem into account, we decide to take the cp value as 0.0148. Then, the model of the 8^{th} decision tree is achieved by the new indexes. We saw that the 8^{th} model modified by the new cp values provided us with small accuracy and kappa values as compared to the 7^{th} model (DTM7). The results of the accuracy and kappa values are explained below.

- Accuracy values for DTM7 and DTM8 are 0.7729 and 0.7424, respectively and they are above 70%. This means that two of the models can be preferred. Someone can choose DTM7 as the best because of the higher accuracy rate. However, for someone, DTM7 can not be preferred due to the fact that this model can lead to an overfitting problem that is caused by over branching. By accuracy rates, DTM7 is chosen by taking the possibility of the existence of an overfitting problem into consideration.
- Kappa values for DTM7 and DTM8 are 0.6628 and 0.6154, respectively. These values fall inside the "moderate" part of the level of agreement indicating that these modes can be used for further predictions.
- As for sensitivity and specificity interpretations, sensitivity and specificity scores of group A and group D are higher than that of groups B and C. This is beucase the most of the elements existing in classes A and D are classified more early as compared to others. This can provide these groups with better separation from the others. Balanced accuracy results also indicate that the countries of groups A and D are easily obtained with the highest accuracy rates by comparison with the other countries that are placed as groups B and C.

Furthermore, all of the sensitivity, specificity, and balanced accuracy rates that are originating from DTM7 and DTM8 are close to that of DTM5 and DTM6 that are obtained by using information gain index.

- As a result, when R2 is used as a response variable, the best decision model is DTM7 that is obtained by Gini index and complexity parameter that is equal to 0.01.
- The following graph summarizes the accuracy and kappa values across the models DTM5 - DTM6 that are achieved by Information Gain Index and DTM7 - DTM8 which are obtained by Gini Index, respectively.

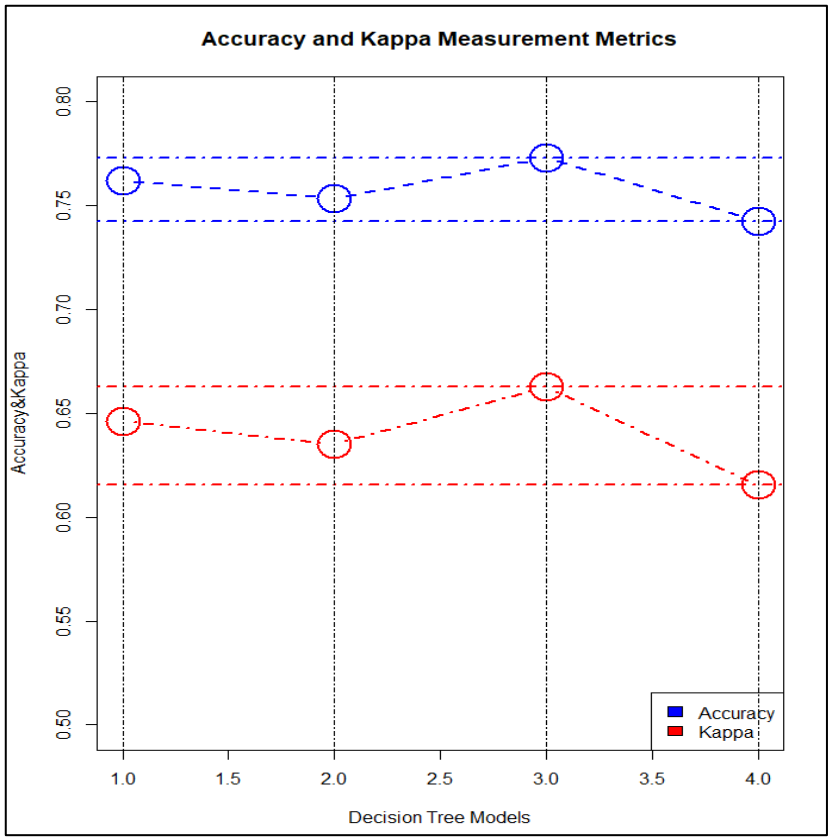


Figure 4.8 Summary of Accuracy & Kappa Values for decision models

According to the latest findings and the figure above, it is clearly understood that the Gini Index with the default value of cp which is 0.01 gives the best accuracy

and kappa values. Therefore, the Gini Index is preferred to the Information Gain index to build up decision tree modelings.

4.2.3 Random Forest (RF) Models

After building up the decision tree modes, we are moving on to designing the random forest models (RFM) that are a collection of the decision tree models which are set up by means of using the different features. This modeling is called as ensembling models. The most important benefit of these models can enable us to assess plenty of decision trees that are built up by randomly selected features. Therefore, randomization on the features leads us to think of resampling methods.

The most often used resampling method is known as bootstrapping. Bootstrap Aggregation is a simple and very powerful ensembling method. It is called as Bagging. As for what ensemble method means, it is a technique that sets up a combination between the predictions from multiple machine learning algorithms to provide more accurate predictions than any individual model. The main aim of this method is to reduce the high variance that is attributed to overfitting models. We know that decision trees have the problem of overestimating the target variable that is known as the response variable. Because overestimation of the target variable can lead to overfitting problem in decision models, the random forest comes into play at this point to remove the overfitting problem and reduce the high variance.

Random Forests (RF) are an improvement over bagged decision trees. Combining predictions from multiple decision tree models in ensembles works always better on the condition that the predictions developing out of sub-models are uncorrelated or at best weakly correlated. RF changes the algorithm in such a way that sub-trees are learned so that the resulting predictions from all of the sub-trees have less correlation. In classification and regression tree models, upon selecting a split point, the learning algorithm is persuaded to look into all features and all features' values to make a selection of the most optimal split point that is the most important variable for the target variable. RF changes those steps by designing as much

decision trees as possible so that the learning algorithm is just made work for a randomly selected sample of features.

The number of features that are used at each split point (m) must be specified as a parametric value to the algorithm.

In literature,

- For classification, the default value for m is $m = \sqrt{p}$.
- For regression, the default value for m is $m = {}^{p}/_{3}$ *(Wikipedia, n.d.)*

Where m is randomly selected features that can be looked over at a split point and p is the number of input variables (the number of all features in data set). For example, if a data set has 25 features totally, then

- m=5

Lastly, the estimated performance of random forest is measured by a metric that is called Out-Of-Bag (OOB). The less OOB the model has, the more confidence we get by the related models.

Out-of-bag error: After creating the classifiers (S trees), for each$(x_{i,}y_i)$ in the original training set i.e. T, select all $T_k$which does not include$(x_{i,}y_i)$. This subset, pay attention, is a set of boostrap datasets which does not contain a particular record from the original dataset. This set is called out-of-bag examples.

4.2.3.1 Modeling Based on R2 categorical response variable

We set up models based on the second categorical response variable (R2) that is made up of multi classes and originating from the k-medoids clustering algorithm. Moreover, we are going to use new predictor variables that are found by courtesy of the Principal Components. We know for sure that the predictor variables are independent and orthogonal to each other.

Step I: Building up Models – RF3 and Confusion Matrix by the Bootstraping Methodology (default values: m=sqrt (p=13) and ntree = 500)

Table 4.10 RFM3 & mtry =3 & ntree =500 & importance =FALSE

Call: Random Forest Model -3					
randomForest(formula = R2 ~ ., data = datakmtrain)					
Type of random forest: classification					
Number of trees: 500					
No. of variables tried at each split: 3					
OOB estimate of error rate: 20.32%					
Confusion matrix:					
Classes	**GroupA**	**GroupB**	**GroupC**	**GroupD**	**class error**
GroupA	258	4	14	0	0.0652
GroupB	12	228	79	0	0.2853
GroupC	12	68	251	0	0.2417
GroupD	2	0	0	12	0.1429

The random forest is obtained by the default values in which 500 decision trees and 3 features, which are variables randomly sampled as candidates at each split, out of 13 features have been used to make predictions. Moreover, the importance of the variable is not assessed here. They will be given at the end of the classification modeling part. Then, the above results can be summarized and explained as follows;

- Ensembling model is made up of 500 decision trees, namely 500 Decision Tree Models (DTM).
- The number of candidate features used in splitting is 3 out of 13 that is close to sqrt (p=13). These parameters cannot give us insight into the most important variables. They can only provide us with the idea of how many features to be taken into consideration to reach the desired results.
- Out-Of-Bag Error is 20.32%. This is different than the validation scores that are accuracy rate and kappa value. From the above results, therefore, we can say that the rows not used in building up models are roughly predicted 80.5% of (1-20.32%). 80.5% is the rate that the groups of the countries that

are not used to derive the models which are obtained by the bootstrapping samples of training data set are predicted well. As compared to the validation score, the OOB score is calculated on data that was not necessarily used in deriving the classification model. While the validation/accuracy score is calculated by using the ensemble of decision trees, the OOB score is computed employing only a subset of DTs not containing the OOB sample (unseen rows) in their bootstrap training dataset.

- As for the class errors, the most of group B countries are misclassified because its class error is higher among the others. The group A countries seem to be best classified because of having the smallest class error.
- The diagonal elements of the confusion matrix say that most of countries' groups are accurately classified in the training data set. Therefore, the models that are designed by the default values work good.

The class errors and out of bag error can be seen from the following figure;

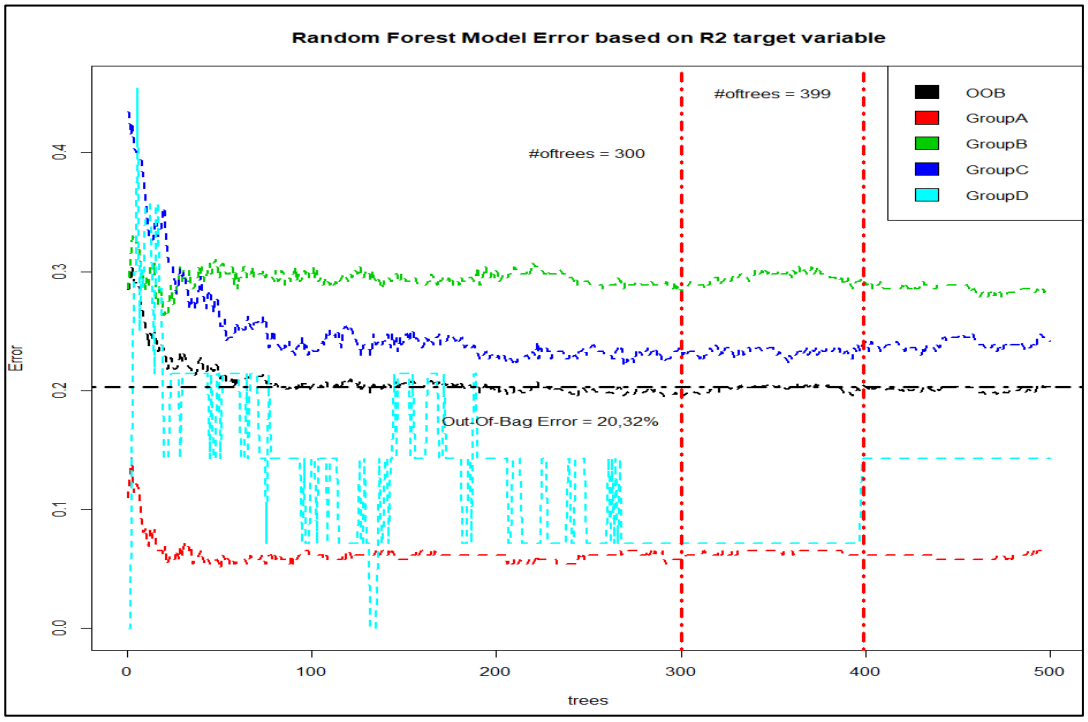


Figure 4.9 RFM3 - OOB & Class Errors & Default Values (ntree =500 and mtry = 3)

- The figure enables us to see easily which countries are accurately classified or predicted. The black dashed line shows Out-Of-Bag Error. It is clear from the figure that the misclassifications of countries grouped A and D are less than that of countries grouped as B and C. As for the number of the trees, it can be seen that the error rates start stabilizing between 300 and 399 trees shown vertical red dashed lines, and after 400 trees, error rates slightly are increasing for all classes. Thus, we take 300 trees as a suitable number of trees in the following RFM4.

Step II: Prediction by using test data

After building up our random forest model, we can assess the accuracy of the model by using validation scores. The test data is made up of 25% of the data set.

- The accuracy rate is 0.7781 that is above 70% and the kappa value is 0.6727 that indicates that the model is at a moderate level. Random forest model (RFM3) works well on the test data by these evaluation metrics. From the confusion matrix that contains sensitivity and balanced accuracy rates, it is seen that the group A countries are mostly predicted accurately. Secondly, group B countries are rightly predicted after group A. By comparison with the Out-Of-Bag Error rate, the accuracy rate is very close to 1- (Out-Of-Bag Error). The compromise between these two measures strengthens the accuracy of model results. Note that while OOB is a good measurement for small data sets and does not require two data sets that are train and test, accuracy rates cannot perform well on the small data sets because we have large data sets. We see that the most accurate classifications are performed for the countries grouped A and D. However, when the model is implemented on the test data, we can see that the most accurate classifications are performed for the countries grouped as A and B. This may be because the number of the countries grouped as D in test data is less than that of the training data. As a result, it is understood that when a new data comes into play, the group B countries are classified more accurately than the countries grouped as D. In order to make a decision on

the number of features to be used in each split, we can get benefit from the following table.

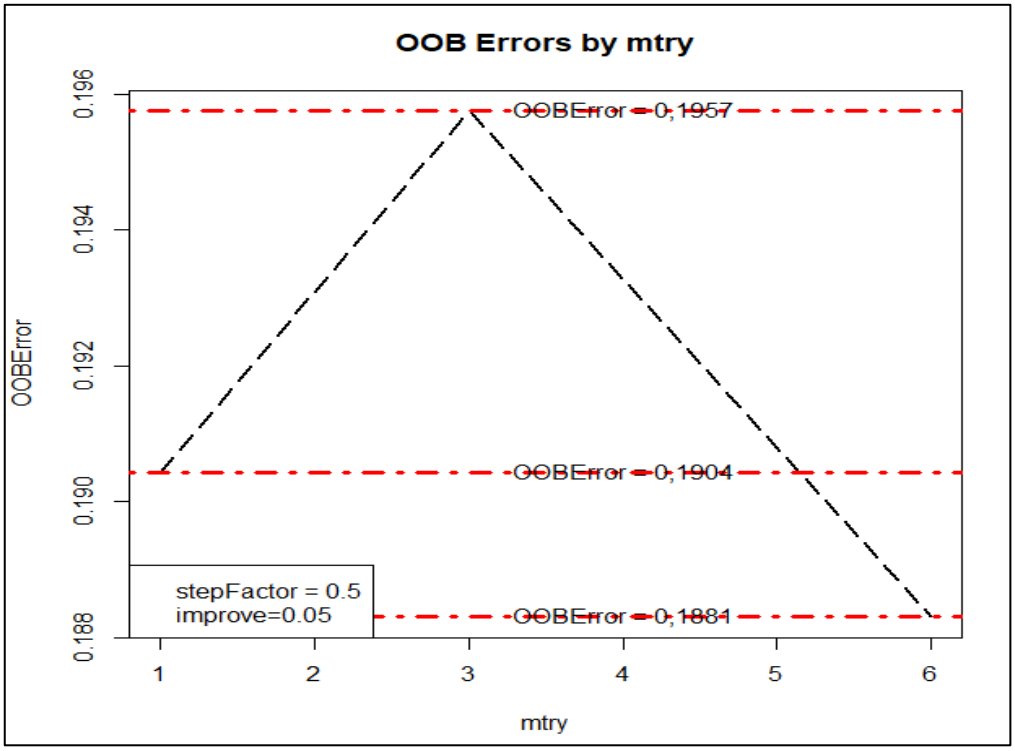


Figure 4.10 OOB Errors by mtry which is the *m* value

According to the figure, mtry, which is the *m* value, must be equal to 6 because of OOB Error. Moreover, in step I, we see that the number of trees to catch the stability on OOB Error is provided slowly after 300 trees are built up. Therefore, the usage of default value that is 500 trees can damage to our analysis. Thus, we select 300 decision trees to set up another random forest model (RFM4). Then, we also drew the conclusions about the new random forest model (RFM4) based on 300 trees and the new m value that is 6. We saw that accuracy and kappa values for RFM4 are slightly smaller than that of RFM3. Thus, we selected RFM3 as the random forest model.

Comparison with the best decision tree model DTM7

- The validation scores of RFM3 and RFM4 are close to the accuracy rate of the best and useful decision tree model that is coded as DTM7.

- Because we get benefits from more information by means of RFM3, this model is preferred to DTM7.
- RFM3 removes the overfitting problem so those models are preferred to DTM7.
- RFM is providing the balance between the bias and variance, therefore, it is suggested to go with these models, and making inferences based on these models enables us to know for sure that we strike a balance between the evaluation metrics.

4.2.4 Bayesian Network Classification (Naive Bayesian Classification (NBC)) Models

The third classification model is designed employing the Naïve Bayesian Classification Method. The Naïve Bayesian Classification (NBC) is based on Bayes Theorem. A naïve Bayesian model is easy to design. Iteration cannot be used for producing a bunch of predictions that make it especially useful for large data sets. Despite the simplicity of the NBC, it often performs surprisingly well to make predictions by comparison with the Decision Tree Models and Random Forest models. During the analysis, it is tried to be proven whether or not this is an actually true explanation.

Algorithm: Bayes Theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c), P(x), and\ P(x|c)$. Naïve Bayes classification assumes that the effect of the value of a predictor (x) on a given class(c) is independent of the values of the other predictors. This assumption is called **class conditional independence.**

Assumption: Predictors must be independent.

Then, the following two formulations summarize the way behind NBC;

$$\boldsymbol{P(Class_j|x) = \frac{P(x|Class_j) \times P(Class_j)}{P(x)}}$$

$\boldsymbol{P(Class_j|x)}$: *the posterior probability of Class j given a predictor x.*

$\boldsymbol{P(x|Class_j)}$: *the likelihood; the probability of the predictor given a Class j and calculated from the training data set.*

$\boldsymbol{P(Class_j)}$: *the prior probability of Class j: It is what we know about the class distribution before we consider the predictors.*

$\boldsymbol{P(x)}$: *the evidence. In practice there is interest only in the numerator and denominator is effectively constant.*

- Applying the independence assumption of the predictors.

$$\boldsymbol{P(x|Class_j) = P(x_1|Class_j) \times P(x_2|Class_j) \times \ldots \times P(x_k|Class_j)}$$

- Substituting the independence assumption, then we can obtain the Posterior Probability of Class j given a new instance of predictors.

$$\boldsymbol{P(Class_j|x') = P(x_1'|Class_j) \times P(x_2'|Class_j) \times \ldots \times P(x_k'|Class_j) \times P(Class_j)}$$

We can ignore the normalizing constant because it is used to make the total probability equal to 1.

4.2.4.1 Assumption Check

Before we move on to building up Naïve Bayesian Classification Models (NBCM), we need to check the independence assumption of the predictor variables. It is known that by means of the principal component analysis, we determine 13 Principal Component instead of using 56 predictor variables. One of the main advantages of utilizing PCs in the place of original variables is that PCs are orthogonal to each other, meaning they are independent of each other. Thus, we are

very lucky to have PCs. The usage of the PCs removes the multicollinearity that gives rise to the problem of correlated predictors in the analysis.

One of the additional issues in NBCM is that since Naïve Bayes utilizes the feature probabilities conditioned on each class, we come across a serious problem when new data has a feature value that never occurs for one or more levels of a response class. The results are $P(x_i|C_k) = 0$ for this individual feature. This zero ripples through the entire multiplication of all features and causes posterior probability to be zero for that class.

A solution to this problem involves using the **Laplace Smoother.** The Laplace Smoother adds a small number to each of the counts like 1 in the frequencies for each feature, which provides that each feature has a nonzero probability of occurring for each class. Typically, a value of one to two for the Laplace smoother is sufficient and the most used ones. Thus, in the case that we have a conditional probability that is equal to 0, we can apply for this way in solving the above mentioned problem.

The following figure shows the pairwise correlation of the predictor variables by Pearson correlation measurement;

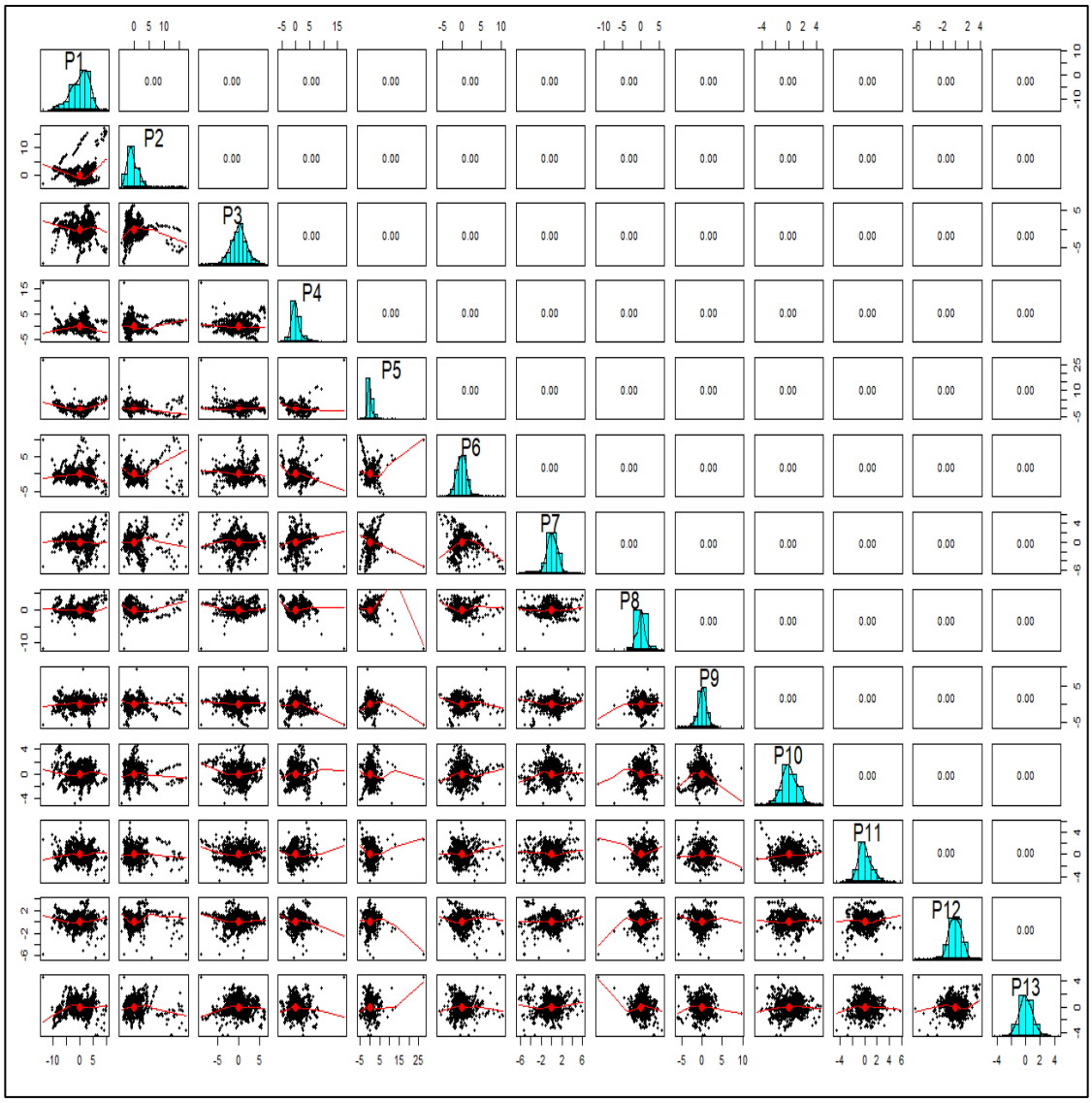


Figure 4.11 Independence between variables

From Figure 4.11, it is clearly seen that there is no linear pattern between the predictor variables. This fact is also understood from the correlation values between the predictor variables. They are equal to almost 0. This means that we satisfy the assumption of independence of the predictor variables.

After the independence assumption is satisfied, we know for sure that the predictor variables are independent and orthogonal to each other. There are two important packages to build up NBCM. One of them is called “naiveBayes” and another is called “e1071”. “naiveBayes” function assumes that the predictor variables are

normally distributed. In Table 4.11, it is seen that none of the predictor variables except P10 satisfy the normality assumption because we reject the null hypothesis that is

$$H_0 : Variables\ are\ normally\ distributed$$

$$H_1 : Variables\ are\ not\ normally\ distributed$$

We are going to use "naiveBayes" in order to get benefit from the kernel density trick. However, we will also add the models that are designed as if the normality assumption is satisfied in order to draw a comparison between the findings.

Table 4.11 Shapiro-Wilk Normality Test for predictor variables

Predictors	p_values
P1	2,25E-09
P2	1,72E-34
P3	1,01E-05
P4	4,79E-22
P5	5,46E-35
P6	2,41E-22
P7	2,45E-18
P8	4,64E-15
P9	2,26E-15
P10	1,75E+00
P11	5,50E-06
P12	2,07E-09
P13	2,63E-02

From the table, we can see that the normality assumption is not satisfied. The model that assumes that predictor variables are coming from a normal distribution is not as accurate as the model that is built up by kernel trick. Thus, it can be said that the normality assumption must be satisfied to derive useful models under NBCM. The accuracy results are shared in the following sections. Thus, for Bayesian Structure Models, the best model is NBCM4 that is set up under the kernel trick.

4.2.4.2 Modeling Based on R2 categorical response variable

We set up models based on the second categorical response variable (R2) that is made up of multi classes and originating from the k-medoids clustering algorithm. Moreover, we use new predictor variables that are found by courtesy of the Principal Components. We know for sure that the predictor variables are independent and orthogonal to each other.

Step I: Building up Models – NBCM3 as if normality assumption of predictor variables is satisfied

Table 4.12 NBCM3 & Under the Normality

Naive Bayes Classification Model-3 under normality					
Call:					
naive_bayes.formula(formula = R2 ~ ., data = datakmtrain)					
Laplace smoothing: 0 **(I)**					
A priori probabilities: **(II)**					
	GroupA	GroupB	GroupC	GroupD	
	0,2936	0,3393	0,3521	0,0148	
::: (P1=PC1) (Gaussian) **(III)**					
P1	GroupA	GroupB	GroupC	GroupD	P1
mean	3,339	-1,973	-1,234	7,480	mean
sd	1,104	3,232	2,738	2,416	sd
::: P2, P3 ……..P13 (Gaussian) **(IV)**					
P…	GroupA	GroupB	GroupC	GroupD	P…
mean	…	…	…	…	mean
sd	…	…	…	…	sd

- **I** is the value of Laplace smoothing value**.** Again, if we take Laplace smoothing value as 1, there is no difference between the results when we do not take Laplace smoothing into account. Thus, there is no need to adjust for the model to get better outcomes. **II** is the prior probabilities. Our response variable consists of 4 groups/classes. It is clear that the group C countries are at the majority. Then, group B countries come after. **III** displays the mean and standard deviation of P1 variables when the target variables is filtered as group A, group B, group C, and

group D. **IV** provides us with the same results as **III**, but we do not show them here because the numbers are just mean and standard deviations. Now, let us look at the boxplot and density plot of some important features (not all of them) by taking groups/classes into consideration to make assessments on the feature effects on the response variable. Note that we only show the top 3 features (PCs) that behave differently for the distinct groups/classes.

- **Step II: Visualization of some of the important features (Principal Components)**

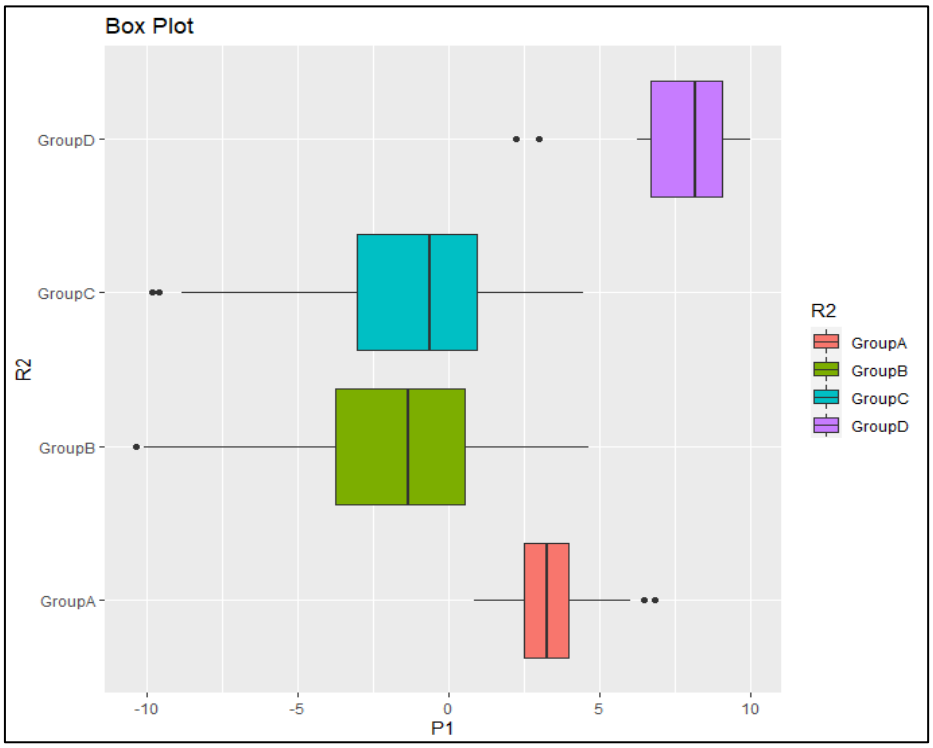


Figure 4.12 Box plot & P1 vs. R2

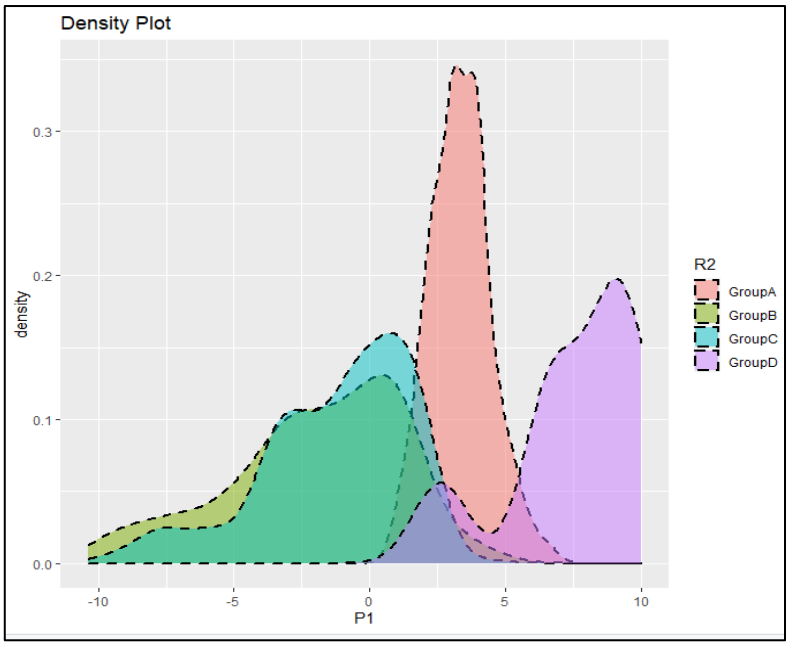


Figure 4.13 Density Plot & P1 vs. R2 & Conditional Probability

From the box and density plots that are drawn by P1 vs. R2, we can see that the P1 feature seems to behave slightly differently for the distinct groups. This is the indication that ***maybe*** the P1 variable is one of the important variables to be used in the classification in order to make predictions of the group of countries.

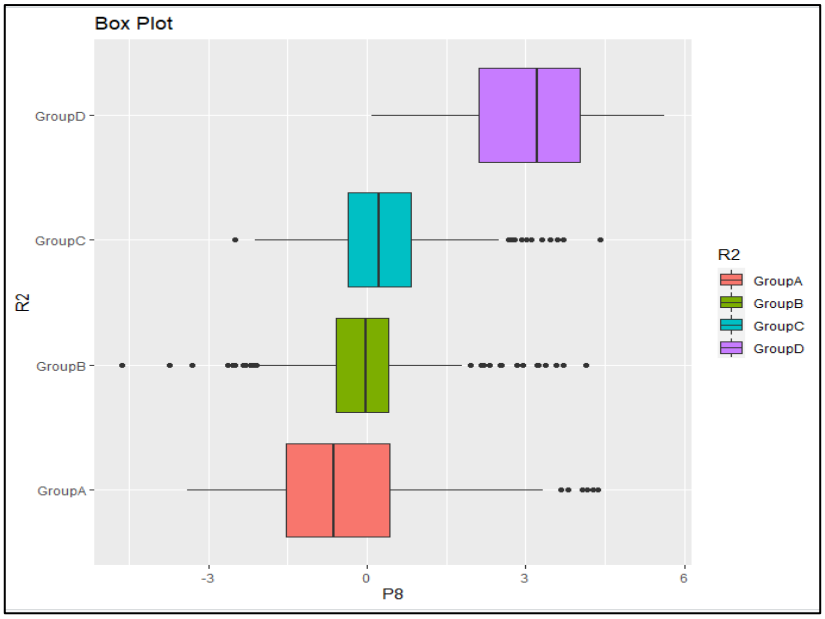


Figure 4.14 Box plot & P8 vs. R2

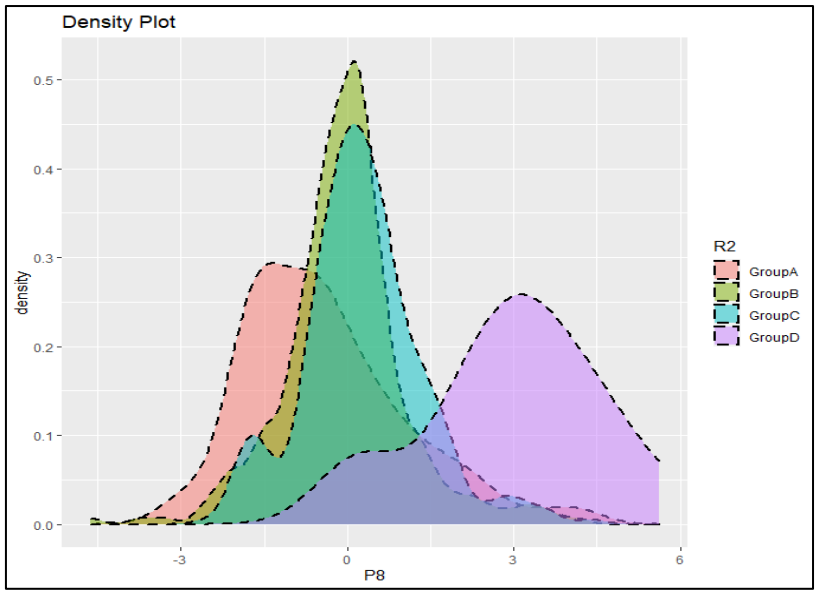


Figure 4.15 Density Plot & P8 vs. R2 & Conditional Probability

From the box and density plots that are drawn by P8 vs. R2, we can see that the P8 feature seems to behave slighlty differently for the distinct groups. This is the indication that ***maybe*** the P8 variable is one of the important variables to be used in the classification to predict the countries' classes.

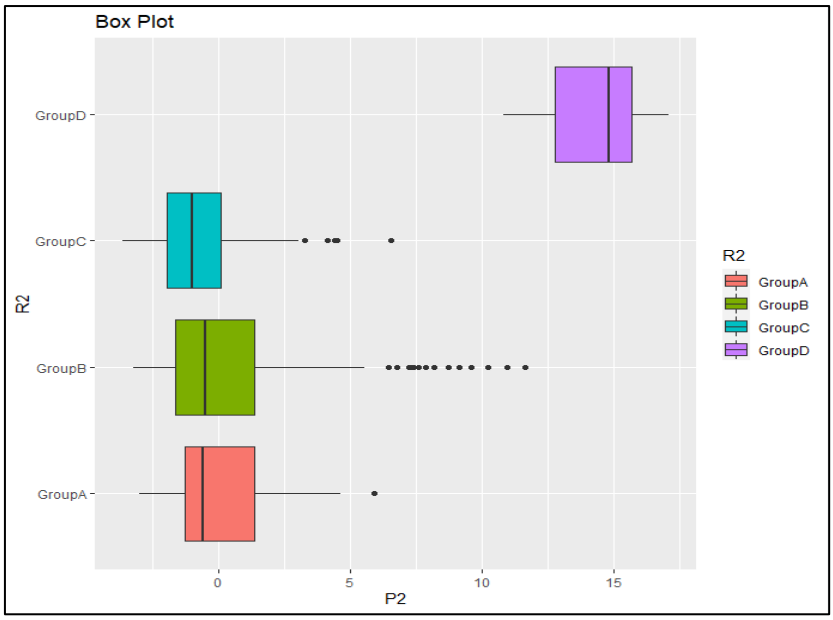


Figure 4.16 Box plot & P2 vs. R2

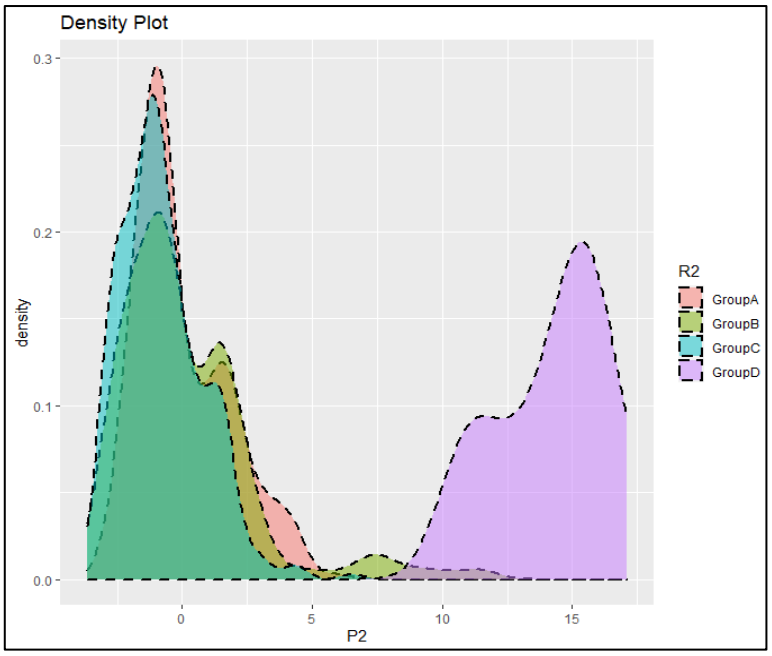


Figure 4.17 Density Plot & P2 vs. R2 & Conditional Probability

From the box and density plots that are drawn by P2 vs. R2, we can see that the P2 feature seems to behave highly differently for the distinct groups. This feature is accepted as the best distinguishing attribute for group D countries. However, from the plot, we can not make assessments such that this variable has an effect on the separation of the groups apart from group D countries. This is the indication that ***maybe*** the P2 variable is one of the important variables to be used in the classification to mainly separate the countries grouped as D from the other countries. If this feature has the capability of making separation of the certain grouped countries, it is highly possible to insert that feature into the model.

Step III: Assessment of NBCM3 designed via normality by making predictions on the test data and Assessment of NBCM4 designed via Kernel Trick making predictions on the test data

- Accuracy values for NBCM3 and NBCM4 are 0.6657 and 0.7022, respectively and the accuracy of NBCM4 is above 70%. This means that NBCM4 can be preferred to NBCM3. We see that when we perform the Naïve Bayesian Classification Model as if predictor variables are coming

from a normal distribution, we obtain small accuracy rates. This situation indicates that although the normality assumption of the predictor variables is not satisfied, performing NBCM based on the normality assumption can lead to meaningless outcomes. As for NBCM4 that is designed by kernel function, we reach a better accuracy rate.

- Kappa values for NBCM3 and NBCM4 are 0.502 and 0.5604, respectively. The kappa values of the two models fall inside the "moderate" part of the level of agreement, indicating that two models can be used for further predictions. However, as for which model to be employed to predict the country classes, the NBCM4 model must be selected because of the high accuracy rate.
- As for sensitivity and specificity interpretations, sensitivity and specificity scores of group A and group D are higher than that of groups B and C. This is because most of the elements/countries existing in classes A and D are classified better as compared to others. Two models can provide these groups to better separation from the others. Balanced accuracy results also indicate that the countries of groups A and D are easily obtained with the highest accuracy rates by comparison with the other countries that are placed in groups B and C.
- NBCM4, which is chosen as the best one to make predictions of the classes based on the R2 target variable, predicts the group A and group D countries better by comparison with the others. When we look at the models that are attributed to Decision Tree and Random Forest to predict the target variable R2 coming from the k-medoids algorithm, the results of NBCM4 show similarity with the decision tree models about the classification of group A and group D countries. As regards to the similarity between the models of RFM and NBCM4, these models only enable group A countries to separate from the other countries with a good accuracy rate and predict those countries in a good measurement. These outcomes display that the models coming from DTMs and NBCMs are producing similar results by comparison with the RFMs. Although DTMs and NBCMs match each other

in the findings, RFMs decompose from those models with slightly different results and higher accuracy rates. As a result, up to now, RFMs provide us with the best models.

- The most important variables that are used for NBCM can be found by means of the "rminer" package. The most important variables (Principal Components) ranged from high importance to low importance are ordered P6, P9, P2, P5, P4, P7, P1, P13, P8, P12, P11, P3 ,and P10.

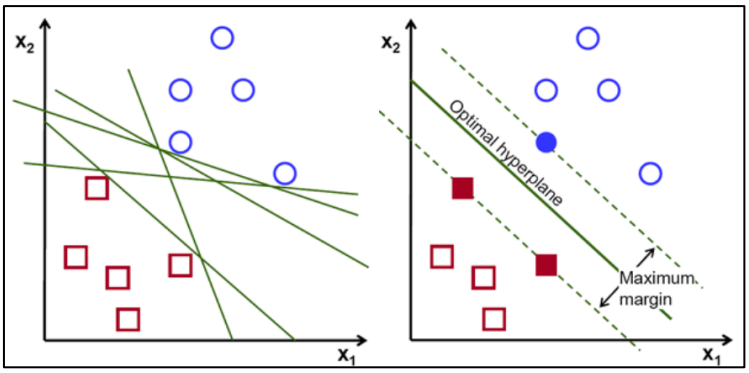
4.2.5 Support Vector Machine Models (SVMM)

Up to now, we built up decision tree models, random forest models, and naïve Bayesian classification models. The best and useful models had been obtained by means of the random forest models because of high accuracy and kappa values. Lastly, in this section, we try to set up models based on support vector machines (SVM). As we know what SVM is doing, it is aimed at finding the best hyperplanes by means of support vectors coming from the distinct groups in order to make the best separation between the groups/classes. A support vector machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm provides an optimal hyperplane that categorizes new examples. It does this by minimizing the margin between the data points near the hyperplane.

Large Margin Intuition

In logistic regression, we take the output of the linear function and squash the value within the range of [0,1] using the **sigmoid function**. If the squashed value is greater than a threshold value (0.5) we assign it to a label 1, else we assign it to a label 0. **In SVM**, we take the output of the linear function and if that outcome is greater than 1, we identify it with one class and if the output is -1, we identify is with another class. Since the threshold values are changed to 1 and -1 in SVM, we obtain this reinforcement range of values ([-1, 1]) which acts as margin.

A support vector machine (SVM) is the last supervised learning algorithm that we are going to deal with in our study. It is one of the modes like DT, RF, and BC that uses classification algorithms. The algorithm of this classification is similar to the algorithm that is performed by logistic regression. Both of them try to capture the line that separates the points from each other. Thus, the main aim of SVM is to figure out a hyperplane (1-dimensional, 2-dimensional, etc.) that distinctly classifies the data points in N-dimensional space (N =the number of features /attributes



(Gandhi, 2018)

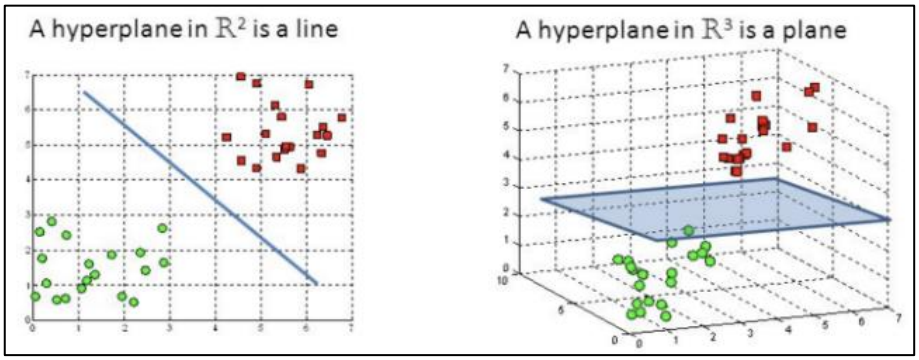
Figure 4.18 Possible hyperplanes and Optimal hyperplane

From the above figure, possible hyperplanes are drawn for the sake of finding the best one that separates the points from each other. In the second figure, an optimal hyperplane is shown and it is found by means of the points that are placed on the edge of each cluster of the points. The distances of those points to each other is called "**maximum margin".** In order to separate data points from each other, there are many possible hyperplanes. Our objective is to find a plane that has the maximum margin that is the maximum distance between the data points of the classes. Thus, maximizing this distance is the main task so that incoming data points can be correctly classified with more accurate rates.

Support Vector Machine Termilonogies: Hyperplanes and Support Vectors

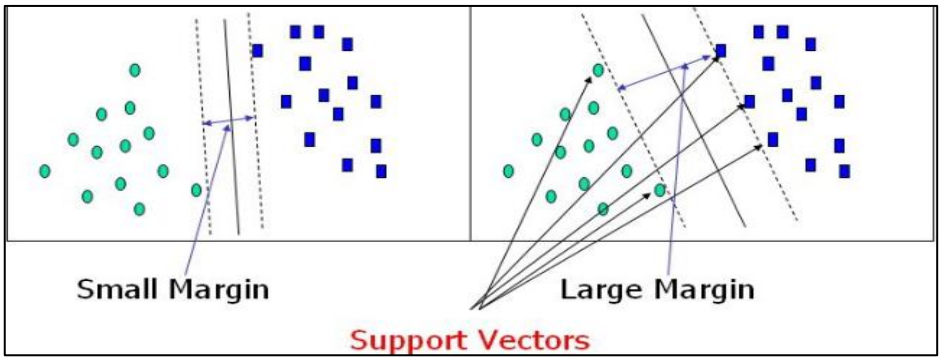
Hyperplanes are also known as decision boundaries which help classify the data points. In the beginning, the data points that are lying down at the opposite sides of

the hyperplane can be caused by distinct classes. The dimension of the hyperplanes depends on the number of attributes/features. For example, if the number of features is 2(age and height), then the hyperplane is just made up of a single line. Moreover, if the number of the features is 3(age, weight, and height), then the hyperplane is composed of a two-dimensional plane. As the number of features increases, it becomes so difficult to find a hyperplane.



(Gandhi, 2018)

Figure 4.19 Hyperplanes



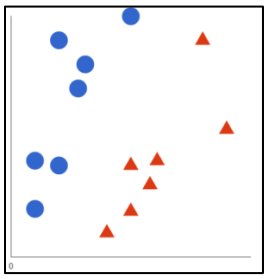
(Gandhi, 2018)

Figure 4.20 Margin and Support Vectors

Support Vectors are the data points that are closed to hyperplane and have an impact on the placement of the hyperplane. By courtesy of these support vectors, the maximization of the margin of the classifier is obtained. Those points are the main elements of the classification that help us build up support vector machines.

The basics of SVM and the working principle of SVM are demonstrated with the following example.

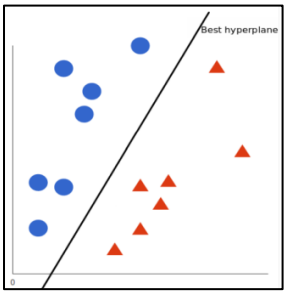
Let us imagine we have two features that are x and y. The training data about these features is summarized below;



(Stecanella, 2017)

Figure 4.21 Training Data & x and y features

An SVM takes these data points and by means of the support vectors, we can draw a hyperplane that is 1-dimensional which is a single line.

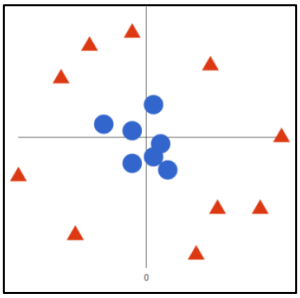


(Stecanella, 2017)

Figure 4.22 One dimensional Hyperplane & A simple line

For SVM, the hyperplane is chosen so that the distance to the nearest elements is the largest. The above example is illustrated in the case that the data points are easily separated from each other with a single line. What if we have a pattern as

following.



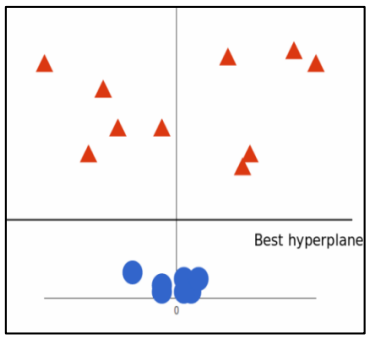
(Stecanella, 2017)

Figure 4.23 Two features not seperable with a single line

From the above figure, it is clear that we cannot separate the data points from each other via a hyperplane that is single line. However, the vectors/data points are very segregated from each other. It looks as though they are easily separated from each other. For the sake of overcoming this problem, a new variable or dimension is produced. Until now we had only 2 dimensions, x and y. We form a new dimension called z and we put a condition onto this new dimension so that it can be calculated by means of the formulation given as follows;

$$x^2 + y^2 = \mathrm{z}$$

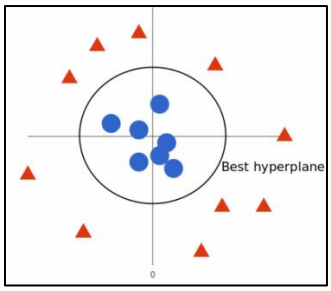
This is the equation for a circle.



(Stecanella, 2017)

Figure 4.24 3-dimensional space

Note that since we are in three dimensions now, the hyperplane is a plane parallel to the x-axis at a certain z.



(Stecanella, 2017)

Figure 4.25 Best hyperplane

At the last step, we see that our decision boundary is a circle that separates data points from each other in a perfect way.

In the last example, we find a way to make a separation of the data points that are non-linear from each other by mapping those vectors into higher dimensions. However, most of the time performing this is computationally expensive and for the more features, it can lead to some confusion. SVM can get by with the dot products of the features. This indicates that we can avoid from calculations that take serious time. Now;

- Think that the new space that is wanted

$$x^2 + y^2 = \mathrm{z}$$

- Find out the dot product of the features.

$$a * b = x_a * x_b + y_a * y_b + z_a * z_b$$

$$a * b = x_a * x_b + y_a * y_b + \left(x^2{}_a + y^2{}_a\right) * \left(x^2{}_b + y^2{}_b\right)$$

- Order SVM to perform its thing, but using the new dot product, we call this a **kernel trick**.

The kernel trick that enables us to avoid a lot of expensive calculations is the terms that we are sometimes going to come across in the SVM. The kernel trick is not a term that is at the scope of SVM. It can be only used to reach decision boundaries.

Cost Function and Gradient Updates

Whilst performing a support vector machine analysis, the main objective is to separate the data points from each other as much as possible. This is implemented by increasing the margin between the data points. The loss function that helps to maximize the margin between the data points is called as hinge loss. An R package that is used for classification based on the support vector machine is "e1071". Before moving on to the analysis of SVM, let us look at some terms that are important for better understanding the results of classification and shown in the following table;

Table 4.13 Support Vector Machines Elements in svm function of "e1071" R package

kernel
the kernel used in training and predicting. You might consider changing some of the following parameters, depending on the kernel type.
linear:
$u'v$
polynomial:
$(\gamma u'v+coef0)degree$
radial basis:
default = $e(-\gamma\|u-v\|2)$; γ=gamma
sigmoid:
$\tanh(\gamma u'v+coef0)$ **; γ=gamma**
degree
parameter needed for kernel of type polynomial **(default: 3)**
gamma
parameter needed for all kernels except linear **(default: 1/(data dimension))**
coef0

Table 4.13 (continued)

parameter needed for kernels of type polynomial and sigmoid **(default: 0)**
cost
cost of constraints violation **(default: 1)**---it is the 'C'-constant of the regularization term in the Lagrange formulation.
If cost is too high, it will mean high penalty for non-seperable points.What may happen in case cost is too high is that model stores too many support vectors and that causes overfitting problem whereas if the cost value is too small, we may end up with underfitting and we may have poor modeling whose results is not accurate, at all.
The cost parameter penalizes large residuals. So a larger cost will result in a more flexible model with fewer misclassifications. In effect the cost parameter allows you to adjust the bias/variance trade-off. The greater the cost parameter, the more variance in the model and the less bias.
epsilon
Epsilon in the insensitive-loss function **(default: 0.1**).When epsilon is increasing, it is understood that we should have actually more supported vectors.
Traditional $\epsilon\epsilon$-SVR works with the epsilon-insensitive hinge loss. The value of ϵ defines a margin of tolerance where no penalty is given to errors.
Remember the support vectors are the instances across the margin, i.e. the samples being penalized in which slack variables are non-zero.
The larger ϵ is, the larger errors you admit in your solution. By contrast, if $\epsilon \rightarrow 0+$, every error is penalized: you end with many (tending to the total number of instances) support vectors to sustain that.

4.2.5.1 Modeling Based on R2 categorical response variable

We set up models based on the second categorical response variables (R2) that are made up of multi classes and originating from the k-medoids clustering algorithm. Moreover, we are going to use new predictor variables that are found by courtesy of the Principal Components. We know for sure that the predictor variables are independent and orthogonal to each other.

Building up Models – SVMM4 after adjusments of the parameters (Tuning of the default parameters)

For the SVMM3, we used the default value for the cost value that is 1.

In order to find the best SVMM to estimate R2 categorical response variable, we need to form a parameter space that is composed of the epsilons that are ranged from 0 to 1 and costs that are ranged from 2^1 to 2^9 in order to figure out misclassification error. Then, we have 11 different epsilon parameters and 9 distinct cost parameters. Totally, we assess 99 models. Then, we find the best model.

Table 4.14 Tuning the Paremeters of SVMM3

best.tune(method = svm, train.x = R2 ~ ., data = datakmtrain, ranges = list(epsilon = seq(0, 1, 0.1), cost = 2^(1:9)))			
- best parameters:			
epsilon		cost	
0		16	
best performance: 0.1606383			
- Detailed performance results:			
epsilon	cost	error	dispersion
0.0	2	0.1851064	0.0423161
0.1	2	0.1851064	0.0423161
0.2	2	0.1851064	0.0423161
0.3	2	0.1851064	0.0423161
0.4	2	0.1851064	0.0423161
…	…	…	…
0.0	16	0.1606383	0.04356077

From the above table, the best parameters for epsilon and cost are 0 and 16, respectively. Upon these parameters are taken into account, the misclassification error becomes minimum.

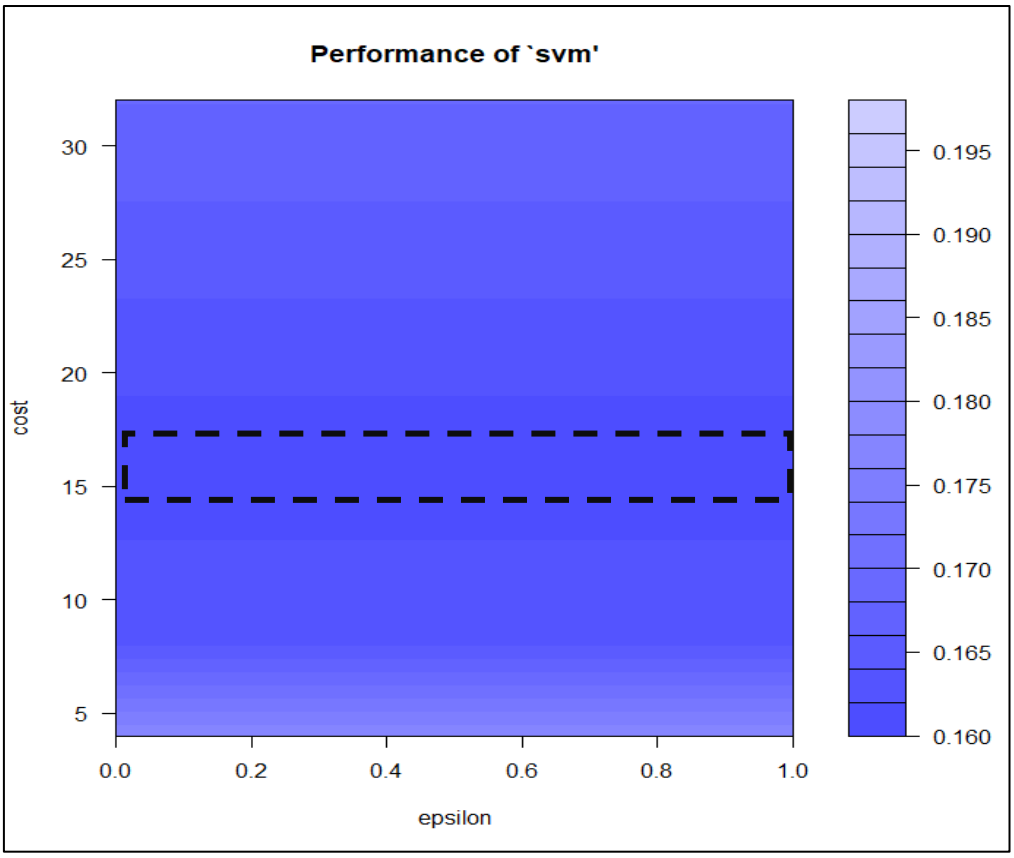


Figure 4.26 Performance of SVMM

The black dashed rectangular displays the area of the minimum misclassification error part. Therefore, the best SVMM is designed under the parameters that epsilon is equal 0 and the cost is equal to 16.

Table 4.15 SVMM4 under the new parameters

Call:
best.tune(method = svm, train.x = R2 ~ ., data = datakmtrain, ranges = list(epsilon = seq(0, 1, 0.1), cost = 2^(1:9)))
Parameters:
SVM-Type: C-classification **(I)**
SVM-Kernel: radial **(II)**

Table 4.15 (continued)

cost: 16 **(III)**
Number of Support Vectors: 493 **(IV)**
(97 188 197 11) **(IV)**
Number of Classes: 4
Levels:
GroupA GroupB GroupC GroupD

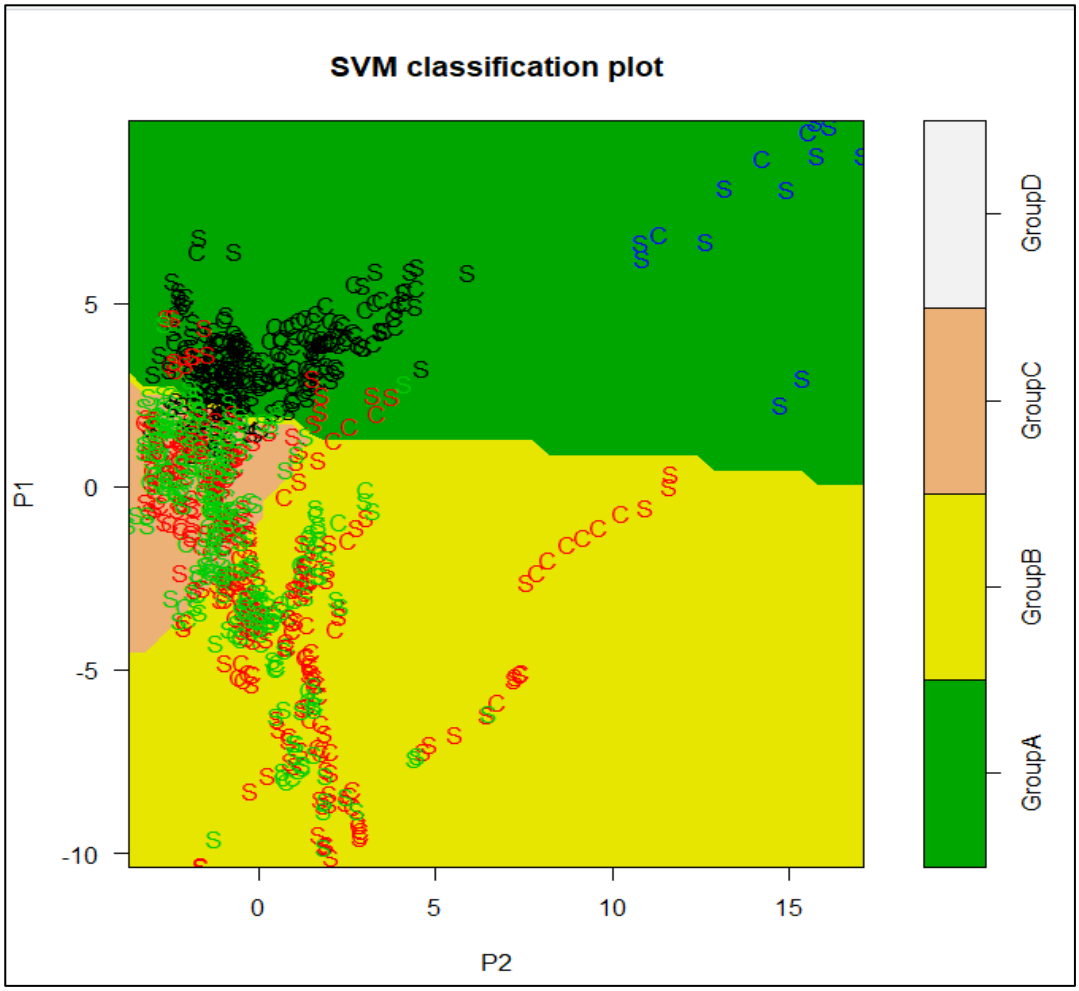


Figure 4.27 SVMM4 Classification Plot & P1~P2 & Based on new parameters

- **I** indicates that because the target variable is made up of groups, this model is described by classification models. **II** indicates the best Kernel function that is used for better classification of the countries' groups. Tuning procedure determines the radial kernel function as the best one by comparison with the other kernel functions that linear, polynomial a

sigmoid. **III** is accepted as 16. **IV** is the total number of support vectors that are coming from all groups. 97 data points out of support vectors are attributed to groupA, 188 is developing out of group B, 197 is originating from group C and 11 is coming from group D countries. When we compare the number of support vectors of SVMM3, the less number is obtained. This removes the danger of overfitting, low bias, and high variance. The trade-off between the bias and variance seems to be completed in a good way providing better modeling.

```
          Reference
Prediction GroupA GroupB GroupC GroupD
    GroupA    270      3      4      0
    GroupB      2    286     19      0
    GroupC      4     30    308      0
    GroupD      0      0      0     14

Overall Statistics

               Accuracy : 0.934
                 95% CI : (0.9162, 0.9491)
    No Information Rate : 0.3521
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9022
```

Figure 4.28 Confusion Matrix by train data

When we look at the above figure, it can be understood that most of the countries are predicted accurately. Accuracy and kappa rates are high enough to say that this model can be better by comparison with the SVMM3. However, to suggest this model, it is required to investigate the behavior of this model on the test data set. In the next step, we evaluate the models based on the test data.

Step III: Model Evaluations

- Accuracy values for SVMM3 and SVMM4 are 0.7584 and 0.8230, respectively and the accuracy rates of the two models are above 75%. This means that those models are significantly useful for making prediction of the countries' group. We see that when we perform the Support Vector Machine Model by the default values of parameters, we obtain smaller accuracy rates. This situation indicates that although the default values are giving satisfying results, performing SVMM based on the new parameters

can contribute significantly to meaningful outcomes. As for SVMM4 that is designed by kernel function called radial and new parameters that we have defined by means of tuning the parameters of SVMM, we reach a better accuracy rate.

- Kappa values for SVMM3 and SVMM4 are 0.6449 and 0.7389, respectively. The kappa values of two models fall inside "substantial" part of the level of agreement indicating that two models can be confidently used for further predictions. However, as for which model to be employed to predict the country classes and to obtain good predictions, the SVMM4 model must be selected because of high accuracy and kappa rates.
- As for sensitivity interpretations of SVMM4, the sensitivity scores of group A and group D are higher than that of groups B and C. This is because most of the elements existing in classes A and D are classified better as compared to others. Two models can provide these groups with better separation from the others. Balanced accuracy of SVMM4 results also indicates that the countries of groups A and B are easily obtained with the highest accuracy rates by comparison with the other countries that belong to groups C and D.
- In the Naïve Bayesian Modeling part, we stated that the best models to predict the R2 response variable coming from k medoids algorithm is developing out of Random Forest Models. After analyzing Support Vector Machine Models, we realized that the accuracy rate of SVMM4 is the highest value that we had ever found. However, these results do not mean that RFM is not anymore valid for making predictions. Note that we are only finding useful models. Therefore, up to now, among all modelings that are set up to predict the R2, RFMs and SVMMs can be preferred to the other ones originating from decision tree and naïve Bayesian classification models. The most important variables ranged from high importance to low importance are ordered P1, P9, P4, P7, P5, P6, P11, P2, P10, P12, P13, P8, and P3.

4.2.6 Conclusion and Discussion

4.2.6.1 The best and useful models Based on R2 categorical response variable

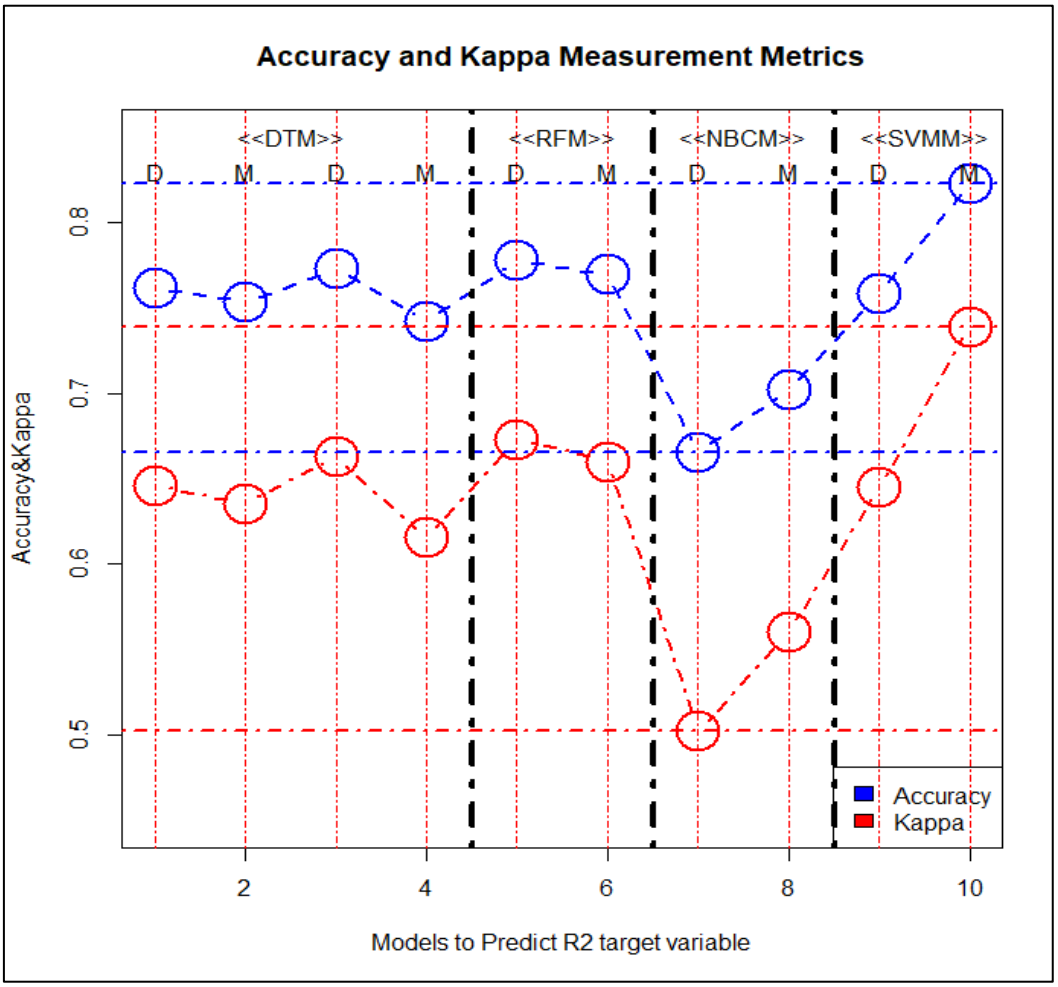


Figure 4.29 Models to Predict R2 target variable (D=default and M=modified)

The best and useful models to predict R2 response variables are designed by Support Vector Machine and Random Forest Models.

Now, in order to see the accuracy, kappa values and computational time of the above mentioned and suggested models, we can take a look at below table,

Table 4.16 Accuracy & Kappa Values and Computational Time in R Software

Useful Classification Models	Accuracy	Kappa	Computational Time in R Software
DTM7(Decision Tree Models)	0.7729	0.6628	2.3 seconds
RFM3 (Random Forest Models)	0.7781	0.6727	3.6 seconds
NBCM4 (Naive Bayesian Classification Models)	0.7022	0.5604	1.9 seconds
SVMM4 (Support Vector Machine Models)	0.8230	0.7389	2.1 seconds

When we look at the accuracy and kappa values for each model, the highest accuracy and kappa values are obtained by support vector machine models. This makes those models effective in designing classification models to estimate the groups of the countries in this study. Therefore, it is suggested that Support Vector Machine models be used for the identification and estimation of the class of the objects based on predetermined predictor variables. As regards computational time, R software spends 2.4 seconds to derive the above models on the average. Support Vector Machine Models come out earlier than the other models. The reason why R spends a little bit much time give random forest models is that those models are made up of plenty of decision tree models. R firstly derives the decision tree model many times and then it takes the average estimated value of each decision tree. Therefore, this process takes a little bit much time as compared to other classification models. Now, let us turn our attention to the most important original predictor variables that have an effect on the classification of the countries. For that purpose, we are aimed at taking the above 4 useful models' results into consideration to analyze which original predictor variables to be included in assessments of the economic situation of a country.

Table 4.17 Important PCs (P) & Original Predictor variables (OP) correlated with PCs

The predictor variables (Ps) that are used in CART models are obtained by Principal Component Analysis on 56 original predictor variables(OP). Therefore, the Ps are corresponding to PCs below and; the response variables (Rs) are the new categorical response variables that we have obtained by means of performing the k-means, k-medois, fuzzy and hierarchical clustering algorithms on the new numeric response variables coming from Principal Component Analysis of 13 original dependent variables.				
Categorical Response Variable	**Best and Useful Model**	**Top 5 Important Variables(PCs)**	**Union of Important Variables (PCs=Ps) for All Models**	**Top 5 Original Important Predictor Variables that are mostly correlated(+/-) to union of important variables(PCs)**
R2	**DTM7**	**PC1,PC9,PC2,PC6, PC7**	**PC1 & PC2 & PC4 & PC5 & PC6 & PC7 & PC8 & PC9 & PC10**	**PC1:OP41,OP42,OP43,(-OP39),(-OP38) PC2:OP29,OP52,OP56,OP55,OP53 PC4:OP13,OP24,OP11,OP9,OP20 PC5:OP11,OP20,(-OP44),(-OP7),OP54 PC6:(-OP50),(-OP54),OP15,OP18,OP14 PC7:OP54,OP50,(-OP8),(-OP2),(-OP35) PC8:OP18,OP14,OP15,OP45,(-OP11) PC9:OP25,OP27,OP7,(-OP46),OP35 PC10:OP19,OP37,OP23,OP22,OP21**
	RFM4 (instead of RFM3)	**PC1,PC9,PC2,PC6, PC7**		
	NBCM4	**PC6,PC9,PC2,PC5, PC4**		
	SVMM4	**PC1,PC9,PC4,PC7, PC5**		

The above table can be read as follows (***The corresponding names of the orginal variables are available Table 5.12 under Appendix C.);***

When we use R2 new categorical response variable coming from 3 Principal Components that are obtained by using 13 original response variables, we have 4 different best models coming from different terminologies. The best and useful models are DTM7, RFM4, NBCM4, and SVMM4. Among those models, SVMM4 has the highest accuracy rate. As for the top 5 important variables, this column shows the most significant new predictor variables used in the models. Then, in

order to observe the common important variables utilized in the model, we can take a look at the union of important variables (PCs=Ps) for all models. Lastly, for the end column of the table, Top 5 Original Important Predictor Variables that are mostly correlated with the union of important variables (PCs) are put to indicate which variables can have an impact on the classification of the countries. For ones that want to see the original predictor variables, the table in Appendix C is used to arrive at the decision about which original predictor variables are the most important to accurately classify the countries by their economic indicators. Note that in this table, the original predictor variables are displayed with the letter **OP. According to Table 4.16, we find the most important new predictor variables that are called Principal Components. In order to get an understanding of which original predictor variables play a role in making the Principal Components important, we can take a look at the following notes.**

$\boldsymbol{R2}$: Categorical response variable coming from k_medoids clustering algorithm

$\boldsymbol{PC_{WS\_VE}}$: PC1 is attributable mostly to **wage and salaried workers and vulnerable employment**-related variables.

$\boldsymbol{PC_{T\_IS}}$: PC2 is attributable mostly to **trade and industry**-related variables.

$\boldsymbol{PC_{IS\_I}}$: PC4 is attributable mostly to **industry and inflation**-related variables.

$\boldsymbol{PC_{I}}$: PC5 is attributable mostly to **inflation** related-variables.

$\boldsymbol{PC_{IN(-)}}$: PC6 is attributable mostly to **foreign direct investment**-related variables. PC6 is negatively correlated with the investment.

$\boldsymbol{PC_{IN(+)}}$: PC7 is attributable mostly to **foreign direct investment**-related variables. PC7 is positively correlated with the investment.

$\boldsymbol{PC_{T}}$: PC8 is attributable mostly to **trade**-related variables.

PC_{IS_M}: PC9 is attributable mostly to **industry and manufacturing**-related variables.

$PC_{TI(+)}$: PC10 is attributable mostly to **net barter trade index**-related variables. PC10 is positively correlated with the net barter trade index.

4.3 Regression Models – Panel (Longitudinal) Data Analysis

In the last section of the study, we are going to build up models based on the regression analysis. Because the data set includes the year information, the best way to design models is carried out by means of Panel (Longitudinal) Data Analysis (PDA). With the help of PDA, the marginal, transitional, and random effects models are designed under the "gee" R package.

4.3.1 Determination of the Target Variable by using the original response variables.

In section 4.1, we performed Principal Component Analysis to reduce the dimension of the data. For that reason, PCA was implemented on 13 predetermined and original response variables and 56 predetermined and original predictor variables. Then, the results showed that 3 PCs explain at least 90% of the variation included in the 13 predetermined and original response variables and 13 PCs explain the at least 80% of the variation included in the 56 predetermined and original predictor variables.

Then, we come to the last analysis that is called panel data analysis. In this part, if 3 new response variables satisfied the normality assumption that is the main assumption of the models, we would have used the 3 PCs as new responses and 13 PCs as new predictor variables. However, although all the transformations are performed so that the new response variables can be normally distributed, we could not satisfy the assumption of normality. Therefore, instead of using the new response variables that are obtained by Principal Component Analysis, we have

decided to use the original response variables. Among 13 predetermined and original response variables, 4 out of 13 are selected to build up models. If we used PCs as response variables, we would have gotten benefit from all response variables, but we fail to satisfy the normality assumption. Then, the data set the for response variables that are going to be employed is displayed below;

Table 4.18 Response Variables for PDA (PPP=Power Purchasing Rate)

COUNTRY	GDP (current BillionUS$)=OR1	GDP, PPP (current international BillionUS$)=OR2	GNI (current BillionUS$)=OR3	GNI, PPP (current international BillionUS$)=OR4
CHE	272.06	256.87	290.15	273.95
CHE	278.63	265.90	289.70	276.47
CHE	301.42	274.77	309.44	282.08
CHE	352.91	276.78	376.10	294.96
CHE	394.16	289.45	418.64	307.43
CHE	408.69	301.74	442.98	327.06
CHE	430.92	337.57	463.07	362.76
CHE	479.91	375.64	482.83	377.92
CHE	554.36	402.15	518.15	375.88
CHE	541.51	400.89	551.81	408.52
CHE	583.78	415.25	618.07	439.64
CHE	699.58	444.55	707.65	449.68
CHE	668.04	462.61	682.99	472.97
CHE	688.50	486.24	703.03	496.50
CHE	709.18	506.89	712.23	509.08
CHE	679.83	529.57	695.95	542.12
JPN	4,887.52	3,404.30	4,957.34	3,452.93
JPN	4,303.54	3,493.11	4,369.13	3,546.35
...	...	...	...	...

Table 4.19 Predictor Variables for PDA

COUNTRY	PC1	PC2	PC3	…..	PC14(time)
CHE	8.395	0.006	2.777	…	0
CHE	7.510	0.003	3.982	…	1
CHE	7.986	0.001	3.256	…	2
CHE	8.924	0.005	2.076	…	3
CHE	10.559	0.000	1.753	…	4
CHE	12.117	0.011	1.683	…	5
CHE	13.301	0.365	2.313	…	6
CHE	13.497	0.460	2.733	…	7
CHE	13.346	0.288	3.090	…	8
CHE	16.127	0.416	2.093	…	9
CHE	16.742	0.396	1.672	…	10
CHE	17.096	1.048	2.470	…	11
CHE	19.835	0.870	2.283	…	12
CHE	19.364	0.230	2.326	…	13
CHE	17.974	0.717	2.039	…	14
CHE	22.430	2.590	1.575	…	15
JPN	17.958	11.991	0.367	…	0
JPN	14.980	8.900	0.021	…	1
…	…	…	…	…	…

Table 4.17 includes the response variables that we are going to use in the Panel Data Analysis. Each one is renamed as OR1, OR2, OR3, and OR4 respectively.

Table 4.18 contains the predictor variables. The predictor variables are composed of;

- ✓ 13 new predictor variables that are used also in classification modeling.
- ✓ PC14 is a time variable that is going to be used to bring the time effect into the open whether or not the variation attributed to time can explain the variation in the response variable.
- ✓ YR is added to draw effective visualizations.

Note that: For each response variables OR1, OR2, OR3, and OR4, we are going to build up models. Totally, 4 different marginals, 4 different transitionals, and 4 different random effect models are obtained and the useful models for each

response variable are selected. We take the model built up based on the only OR2 into consideration because according to model evaluation metrics (MEM), we obtained the best results with the models that are designed under the OR2 response variable. The MEM of other models can be reached in Appendix E.

Main Aim of PDA: The main aim is to find the effect of predictor variables on the response variable by exploring data and suggesting some models.

Research Questions of PDA

- What is the reaction or behaviour of the OR1, OR2, OR3, and OR4 as a change occurs in the predictor variables?

Statistically, we are looking for the answer to the question stated below

"Are the predictor variables in relation to response variables ?"

- Can the change or variation in the values of OR1, OR2, OR3, and OR4 be explained by depending on the predictor variables?

Statistically, we are looking for the answer to the question stated below

Do the predictor variables explain the variability in the response variable?

- Can the alteration in the values of OR1, OR2, OR3, and OR4 be explained by the predictor variables change arising from the difference between COUNTRIES and within each COUNTRY over the years?

Statistically, we are looking for the answer to the question stated below

Do the variability stemming from between individuals, within the individuals and year effect contribute to the variation in response variables?

PDA is performed under 3 sections as follows;

- ✓ **Exploratory Longitudinal Data Analysis**

In this section, by means of visualizations, we gain insight into the most probable predictor variables that are in relation to the response variables. This

lays the groundwork for better understanding of which predictor variables to be added to the modeling.

✓ **Estimation of the Variance-Covariance Structure and Normality Check**

In this section, we estimate **Σ and we check if or not** the response variables satisfy the normality assumption. If the normality assumption is not satisfied, the transformations on the response variables are carried out.

✓ **Models**

Lastly, after the estimation of Variance- Covariance structure and normality check, we move on to designing the models.

Note that: we had built up models taking all the response variables into consideration. After carrying out model evaluation indexes, we see that the best model was obtained in the case that **OR2** was used. Then, we focused on the models developing out of the usage of the OR2 response variable. In the end, we observed that we decide to suggest the **Random Intercept Effect Model** when model evaluation metrics are assessed. Thus, we exhibit the results of the related model.

4.3.2 Exploratory Longitudinal Data Analysis (ELDA)

Let us look at the first of the appearance of the data set that is called "datam" in R with the help of head and summary functions.

```
> head(datam)
  ID Countries_Years COUNTRY      OR1      OR2      OR3      OR4       PC1          PC2      PC3      PC4          PC5
1  1          CHE00     CHE 272.0555 256.8696 290.1456 273.9501  8.394843 0.0062378783 2.777271 2.153486 4.058603e-01
2  1          CHE01     CHE 278.6313 265.8996 289.7025 276.4651  7.509872 0.0034749532 3.981743 1.207282 2.506470e-02
3  1          CHE02     CHE 301.4168 274.7684 309.4414 282.0836  7.985690 0.0009077887 3.255901 1.598603 3.855547e-05
4  1          CHE03     CHE 352.9148 276.7782 376.1023 294.9636  8.923848 0.0047713488 2.075891 1.938140 1.186400e-03
5  1          CHE04     CHE 394.1637 289.4518 418.6411 307.4264 10.558517 0.0004731277 1.753181 1.913481 2.133157e-01
6  1          CHE05     CHE 408.6894 301.7406 442.9822 327.0594 12.117279 0.0108665775 1.682767 1.691314 8.232454e-01
        PC6         PC7      PC8          PC9      PC10     PC11      PC12        PC13 PC14 PC15 PC16 PC17 PC18   YR
1 0.5669751 0.180560963 3.446633 0.0299856575 0.4283947 2.657028 2.0357683 0.014496504    0    3    1    1    1 2000
2 0.1916557 0.002719753 3.202035 0.1112326981 0.2111805 1.336825 1.8979757 0.121447864    1    3    1    1    1 2001
3 0.2354826 0.088561634 2.934716 0.0014994415 0.1893192 1.327604 1.8889191 0.159014786    2    3    1    1    1 2002
4 0.1387430 0.211962680 2.794044 0.1317610857 0.3808722 1.453317 1.8020994 0.133424199    3    3    1    1    1 2003
5 0.6730847 0.166930157 2.975337 0.0006768677 0.3299300 2.051921 0.8631935 0.003998952    4    3    1    1    1 2004
6 1.2023601 1.912423139 3.431894 0.1893373309 0.4586189 2.712331 0.3250983 0.106097539    5    3    1    1    1 2005
```

Figure 4.30 Main Data Long Form

PC15, PC16, PC17, and PC18 variables are the categorical response variables that are corresponding to R1, R2, R3, and R4 in CART models. We put those variables into analysis in order to derive graphical representations.

```
> summary(datam)
       ID      Countries_Years      COUNTRY              OR1                OR2                OR3          
 Min.   : 1   Length:1296        Length:1296        Min.   :    4.983   Min.   :   13.18   Min.   :    4.836  
 1st Qu.:21   Class :character   Class :character   1st Qu.:   48.001   1st Qu.:  107.37   1st Qu.:   45.208  
 Median :41   Mode  :character   Mode  :character   Median :  166.287   Median :  279.87   Median :  162.564  
 Mean   :41                                         Mean   :  671.471   Mean   :  919.34   Mean   :  672.296  
 3rd Qu.:61                                         3rd Qu.:  443.832   3rd Qu.:  790.15   3rd Qu.:  447.344  
 Max.   :81                                         Max.   :18219.298   Max.   :19820.98   Max.   :18704.317  

      OR4                PC1                PC2                PC3                PC4                PC5          
 Min.   :   12.67   Min.   :  0.000   Min.   :  0.0000   Min.   : 0.0000   Min.   :  0.0000   Min.   :  0.000  
 1st Qu.:  102.59   1st Qu.:  1.590   1st Qu.:  0.6253   1st Qu.: 0.2816   1st Qu.:  0.1601   1st Qu.:  0.082  
 Median :  274.81   Median :  6.253   Median :  1.9546   Median : 1.6637   Median :  0.9193   Median :  0.466  
 Mean   :  916.37   Mean   : 12.849   Mean   :  7.3838   Mean   : 4.8300   Mean   :  3.7040   Mean   :  2.902  
 3rd Qu.:  786.44   3rd Qu.: 14.758   3rd Qu.:  4.5586   3rd Qu.: 5.1181   3rd Qu.:  3.0531   3rd Qu.:  1.598  
 Max.   :19745.79   Max.   :195.812   Max.   :292.0117   Max.   :81.0605   Max.   :301.6101   Max.   :754.919  

      PC6                PC7                PC8                PC9                PC10               PC11          
 Min.   :  0.0000   Min.   : 0.0000   Min.   :  0.00000   Min.   : 0.00000   Min.   : 0.0000   Min.   : 0.00001  
 1st Qu.:  0.1909   1st Qu.: 0.1098   1st Qu.:  0.08791   1st Qu.: 0.08868   1st Qu.: 0.1156   1st Qu.: 0.10332  
 Median :  0.7815   Median : 0.5067   Median :  0.45552   Median : 0.46929   Median : 0.5985   Median : 0.45524  
 Mean   :  2.8113   Mean   : 2.2321   Mean   :  1.82509   Mean   : 1.72954   Mean   : 1.5119   Mean   : 1.31316  
 3rd Qu.:  2.2367   3rd Qu.: 1.6312   3rd Qu.:  2.04843   3rd Qu.: 1.69161   3rd Qu.: 1.9978   3rd Qu.: 1.54388  
 Max.   :111.7740   Max.   :39.2530   Max.   :137.03662   Max.   :92.25473   Max.   :22.5020   Max.   :32.15763  
```

Figure 4.31 Summary of the data (not all of them illustrated)

In Figure 4.30, our main data is displayed. You can see that OR1, OR2, OR3, and OR4 are our response variables and the others are predictor variables. Figure 4.31 takes a glimpse of the summary of the data. We can see some summary statistics. Panel data is composed of the variables of individuals that are observed throughout 16 years (2000 to 2015).

```
> table(datam$ID)

 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16
28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54
16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16
55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16
```

Figure 4.32 Each Country is followed up 16 times from 200 to 2015

From the above Figure 4.32, we can see that each country is followed up 16 times from the years of 2000 to 2015. This means that our data is fully constrained (balanced data). As for the issue of whether or not we have missing cases in our data, as it is seen at the output below that none of the variables include missing values (NA).

```
> summary(is.na(datam))
     ID          Countries_Years  COUNTRY          OR1             OR2             OR3
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:1296      FALSE:1296      FALSE:1296      FALSE:1296      FALSE:1296      FALSE:1296
    OR4             PC1             PC2             PC3             PC4             PC5
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:1296      FALSE:1296      FALSE:1296      FALSE:1296      FALSE:1296      FALSE:1296
    PC6             PC7             PC8             PC9             PC10            PC11
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:1296      FALSE:1296      FALSE:1296      FALSE:1296      FALSE:1296      FALSE:1296
    PC12            PC13            PC14            PC15            PC16            PC17
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:1296      FALSE:1296      FALSE:1296      FALSE:1296      FALSE:1296      FALSE:1296
    PC18            YR
 Mode :logical   Mode :logical
 FALSE:1296      FALSE:1296
```

Figure 4.33 Missing Cases Investigation

Because we treated missing cases at previous parts, from the above output, we have no missing cases.

Reminder: Missing cases are treated with mostly "LAST OBSERVED VARIABLE (LOV)" Technique for the countries that have a loss of information only for certain years. We have 78 variables when we collect the data. Since 9 of those predictor variables contain a multitude of missing cases; they are removed from the analysis. The rest of 56 predictors and 13 response variables are including a small amout of missing cases only for some years. Then, in order to treat those missing cases, LOV and mean imputation technique was utilized. That means, 69 variables are treated with LOV and mean imputation techniques. The rest 3 predictor variables out of 56 predictor variables are important and include more missing cases by comparison with the 53 predictor variables. They are treated with mostly mean imputation technique in the Principal Component Analysis. The treatment of missing cases existing in 13 predetermined and original response variables are performed before making clustering analysis.

Our data is **person-period data** namely, it is in **the long-form**, because of the structure of "**One row for each subject at each time measurement."**

Because we have 14 numeric new predictor variables coming from Principal Component Analysis, first of all, let us look at the correlation plot of the variables. If we draw all of the variables in the scatter plot matrix, visualization may not give us valuable information about the possible relationships between the

predictor and response variables. The visualization contains the variables that are shown in the following correlation plot.

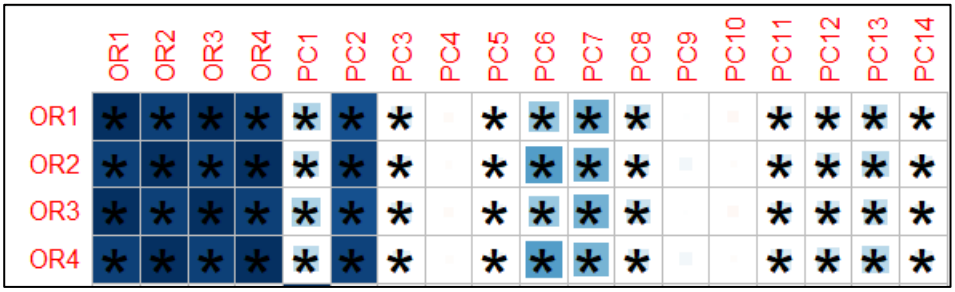


Figure 4.34 Correlation Plot

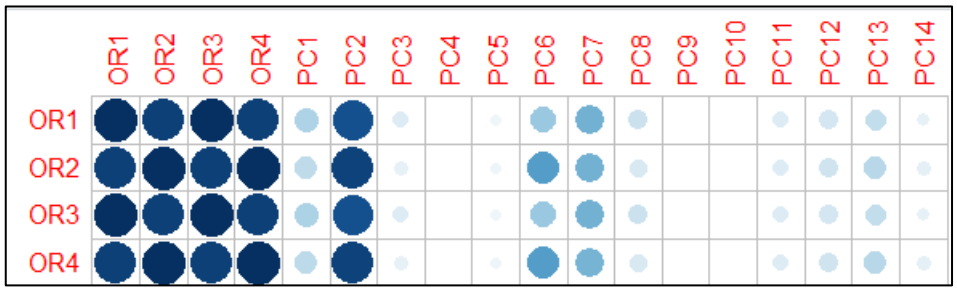


Figure 4.35 Correlation Plot-2

The correlation plot tells us that the response variables are significantly correlated with the variables that are shown with stars. Therefore, the response variables are correlated with the variables that are PC1, PC2, PC3, PC5, PC6, PC7, PC8, PC11, PC12, PC13, PC14(time). The response variables seem to be uncorrelated with the variables that are PC4, PC9, and PC10. For visualization, PC1, PC2, PC6, PC7, and PC13 are drawn because of a high correlation with the response variable. The high correlation is displayed with the blueness and size of the circle. In order to better understand which variables are the most significant to explain the variation in the response variables, the modeling part provides us with the most certain information.

❖ **Assessment the association of the variables with Exploratory Plots (Spaghetti, Trellis, Histograms, Mosaic Plots etc.)**

Scatter PLOTS

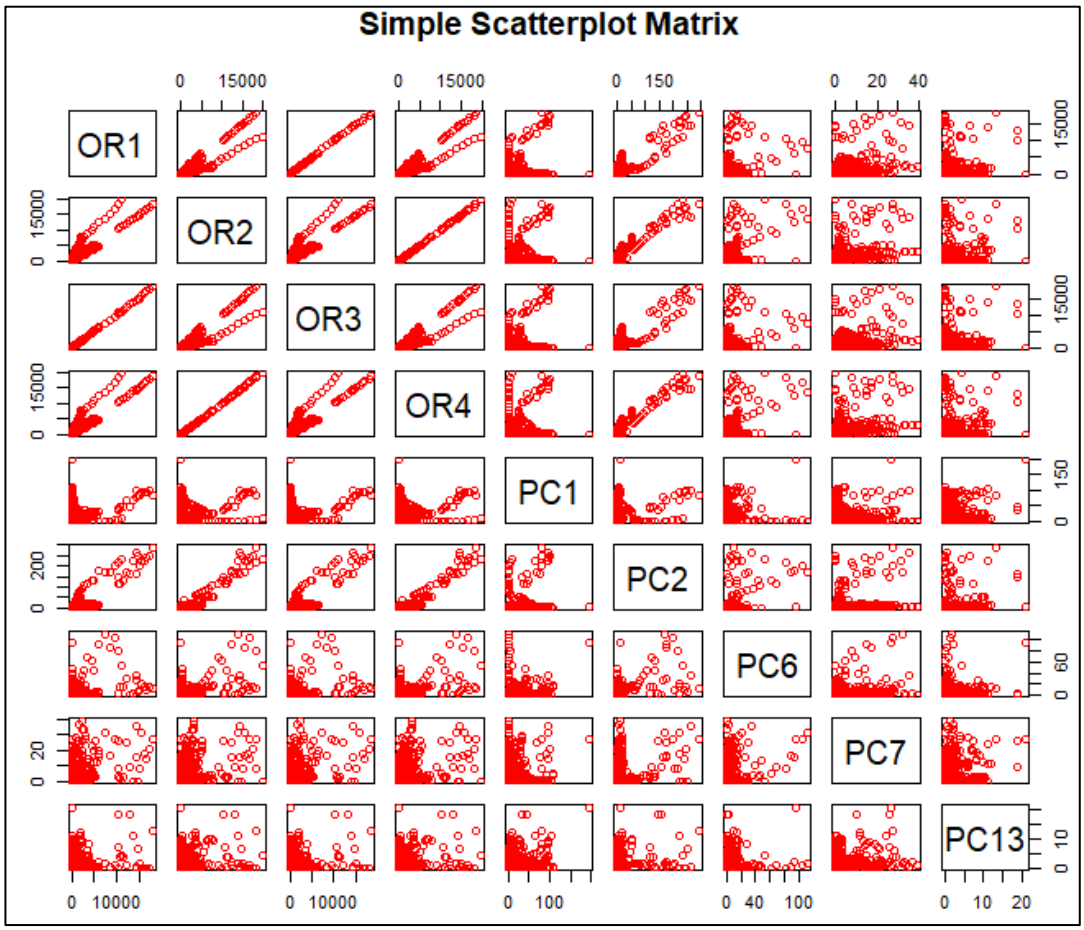


Figure 4.36 Scatter Plots & OR1, OR2, OR3 and OR4 vs. PC1, PC2, PC6, PC7 and PC13

The above pairwise scatter plot illustrates which variables are in relation with predictor variables. We can see that while the relationship between the predictor variables of PC1 and PC2 and OR1, OR2, OR3 and OR4 seems to be more clear, we cannot bring the same interpretation for the other variables. Moreover, we can see that the response variables are highly correlated with each other. We can learn this fact from the correlation plot 1&2 and a simple scatterplot matrix. Again, it is seen that the predictor variables that are drawn the above scatterplot matrix are uncorrelated with each other and there should not be in relation to one another because they are orthogonal to each other.

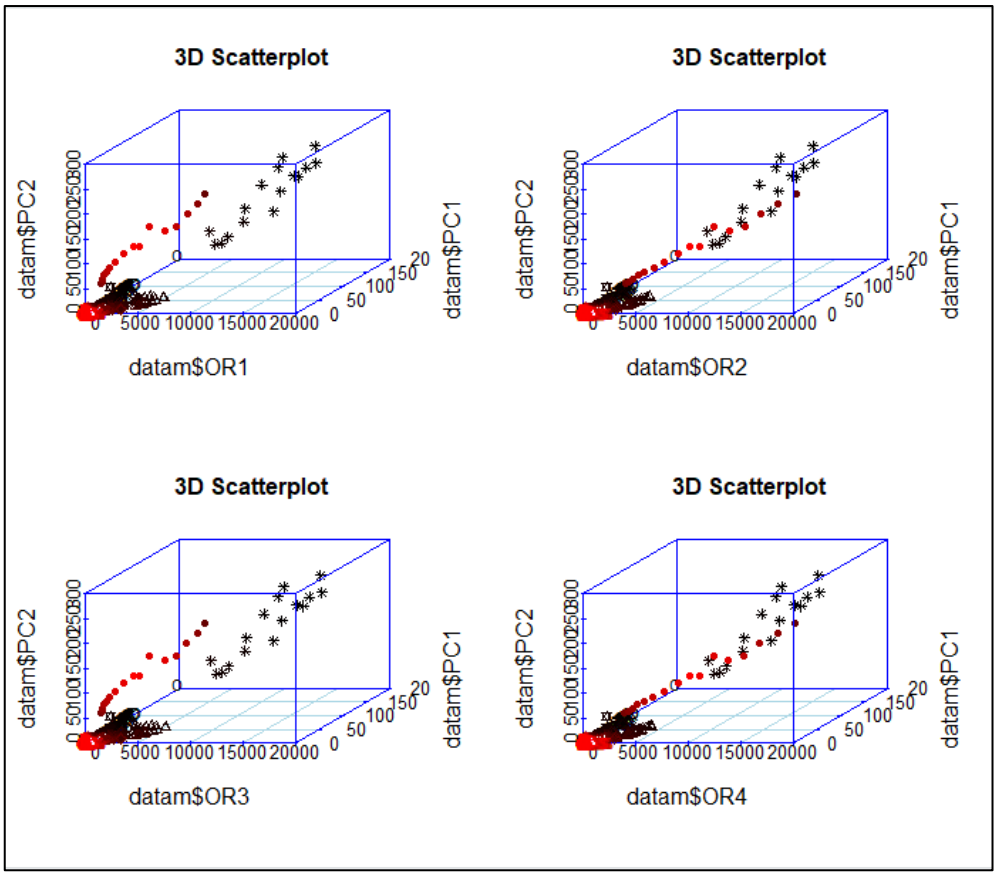


Figure 4.37 3d Scatter Plots & OR1, OR2, OR3 and OR4 vs. PC1 and PC2 & Point Shapes represent each country

Interpretation: 3D Scatter Plots are drawn by the response variables with the first two predictor variables. From the above plots, it can be seen that there is a positive relationship between the response and predictor variables. Because the shapes of the points represent the countries, across 4 plots, the same countries are almost exhibiting the same behavior in which as PC1 and PC2 increase, the response variables are also increasing. This indicates that the variables **PC1 and PC2** can be significant variables to explain the variation in the response variables.

Trellis PLOTS

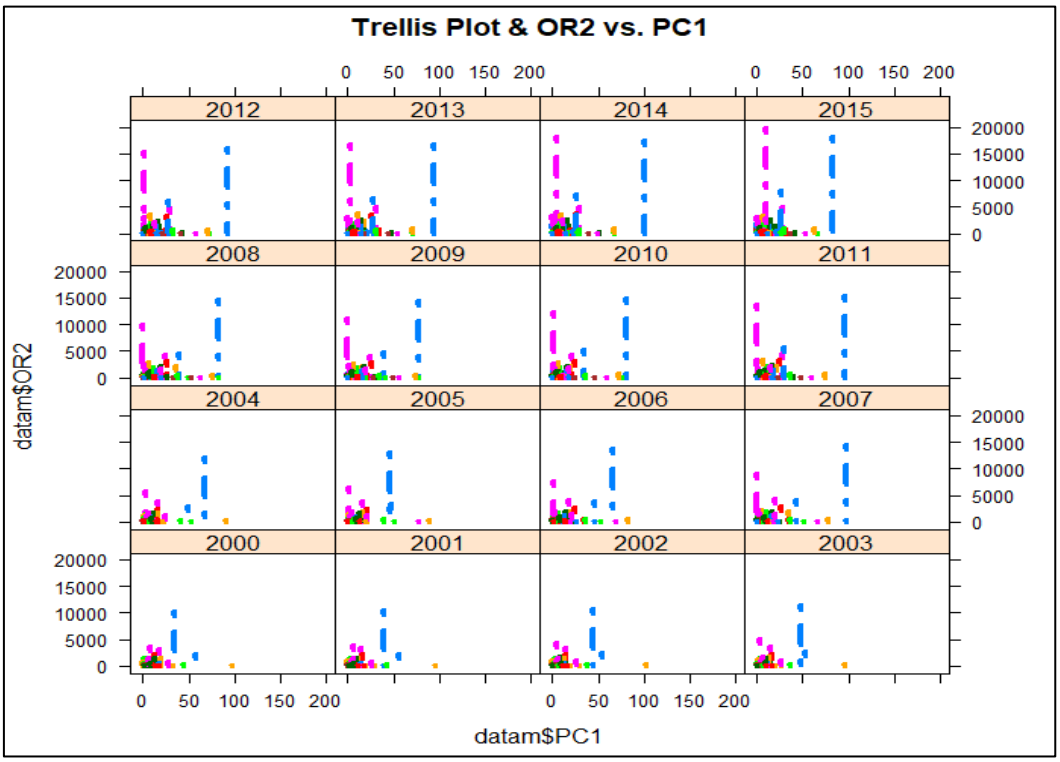


Figure 4.38 Trellis Plot & OR2 vs. PC1 & 81 Countries shown with lines coloured differently

Interpretation: From the above trellis plot that is drawn by response variable of OR2 with the first predictor variable (PC1), on the condition that a change occurs in the years, we can see that the response variables seem to be increasing especially for some countries as an increment in PC1 occurs and time passes. For example, the country showed the colors blue and pink represents the above defined behavior. This behavior is also observed when PC2 is taken into consideration, as well. Hence, we can infer that the effect of **year, PC1, and PC2** are significant to bring an explanation to the variation in the response variables.

Spaghetti PLOTS

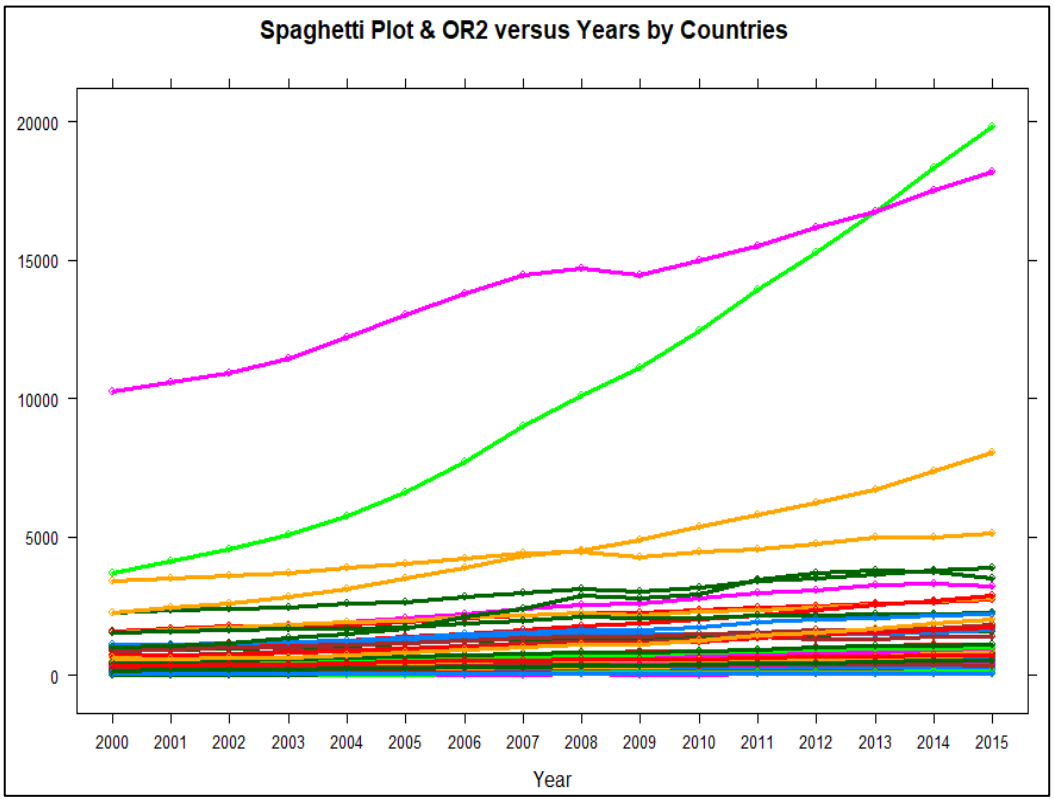


Figure 4.39 Spaghetti Plot & OR2 vs. Year by Countries

Interpretation-1: The above spaghetti plot gives us information about the behavior trend of OR2 response variables for all countries shown distinct colored lines as time passed. We had drawn the other scatter plots based on OR1, OR3, and OR4 and we see some similarities between the scatter plots. The reason behind this situation is attributed to the type of response variables. Those variables are similar to each other. The response variables OR1 and OR3 are calculated by usual methods whereas OR2 and OR4 are computed by taking the power purchasing rate into consideration. However, the change in the response variables across the years is similar to each other without thinking of the type of response variable.

Interpretation-2: As years passed, an increment happens in the values of the responses especially for some countries. Each line corresponds to a single country. However, the values of response variables for most of the countries do not seem to change across the years.

Interpretation-3: In terms of the modeling approach, this may indicate that the variation attributed to the year effect of countries cannot contribute to the variation in the response variable. This may also mean that random time effect model that is built up based on country-based year effect can not be suggestable and cannot be a useful one. In order to analyze this situation, we modeled a random effect model taking the country-based year effect into account and it is understood from the model results that the variation attributed to country-based year effect does not seem to be important to explain additional variation in the response variable. This fact is also observed from the spaghetti plot which gives us information about the change in the values of response variable across the countries as time passed. From the spaghetti plots, it can only be seen that the response variables of 3 or 4 countries are increasing as years passed. However, we know that the OR2 values of countries show important differences. There is a big variation in the OR2. This can be considered as a sign that the individual country effect on explaining the variation in the response variable can be important. In order to observe this assumption, we designed random intercept models. The findings proved us right and we see that the variation attributed to individual effect seems to be enough high to say that the countries' effect on explaining the variation in the response variable is significant.

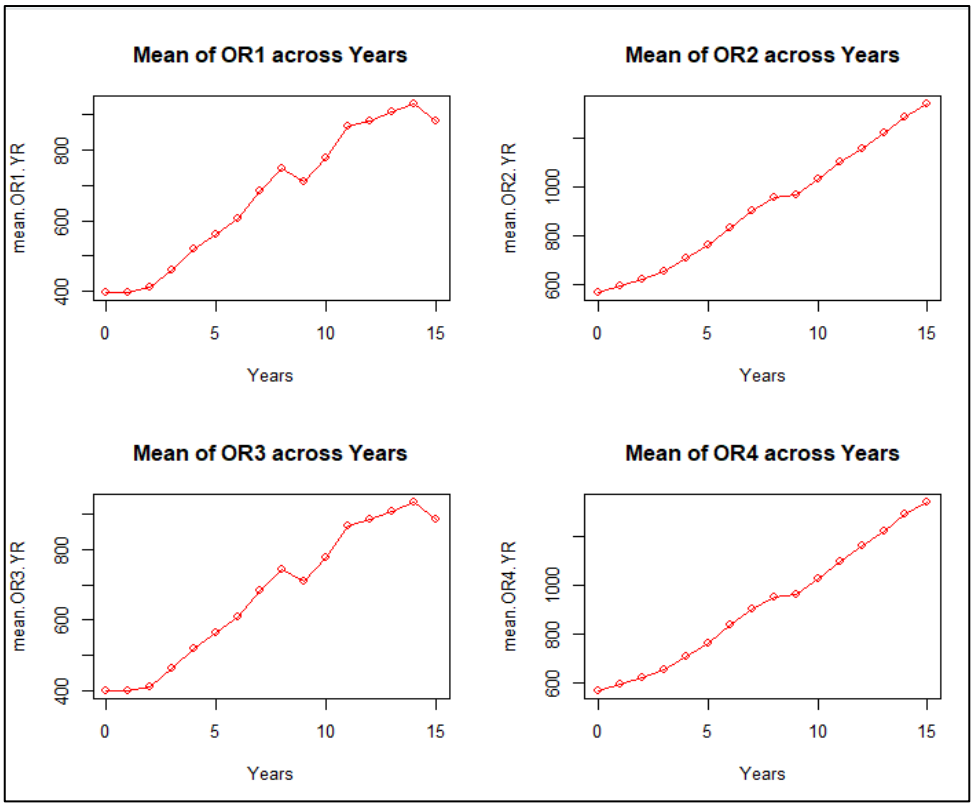


Figure 4.40 Scatter Plots & Mean of ORs vs. Years

Interpretation: We draw the above scatter plots in order to observe the effects of the year on the response variables. We can see that as years passed the mean value of the response variables is increasing. These plots strongly imply that the variation attributed to the year can contribute to explaining the variation in the response variable. It is highly possible to see the variable of the year as an important variable. According to these plots, in the modeling part, we built up marginal and transition models. Morever, inferences that are made on the basis of the above plots were proved, meaning year effect came to the front as one of the most important variables to explain the variation in the response variable. However, because we preferred to put the random effect model, the results about the marginal and transition models were not given here.

Loess Smoothing Strategy

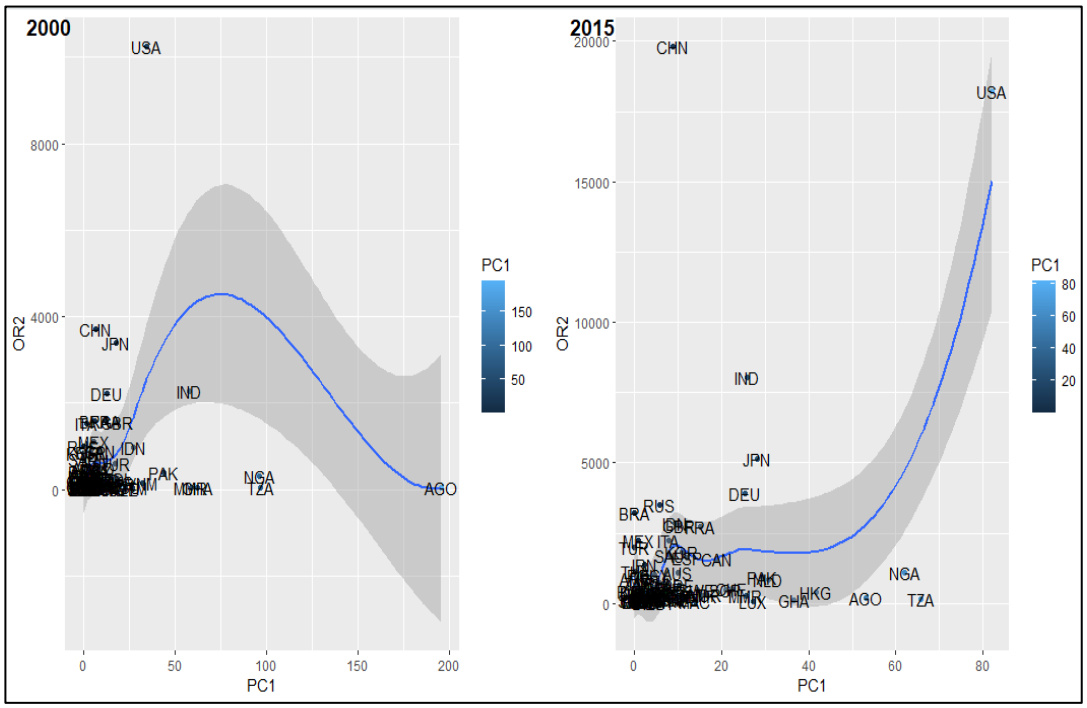


Figure 4.41 Loess Smoothing & OR2 vs. PC1 & Change from 2000 to 2015

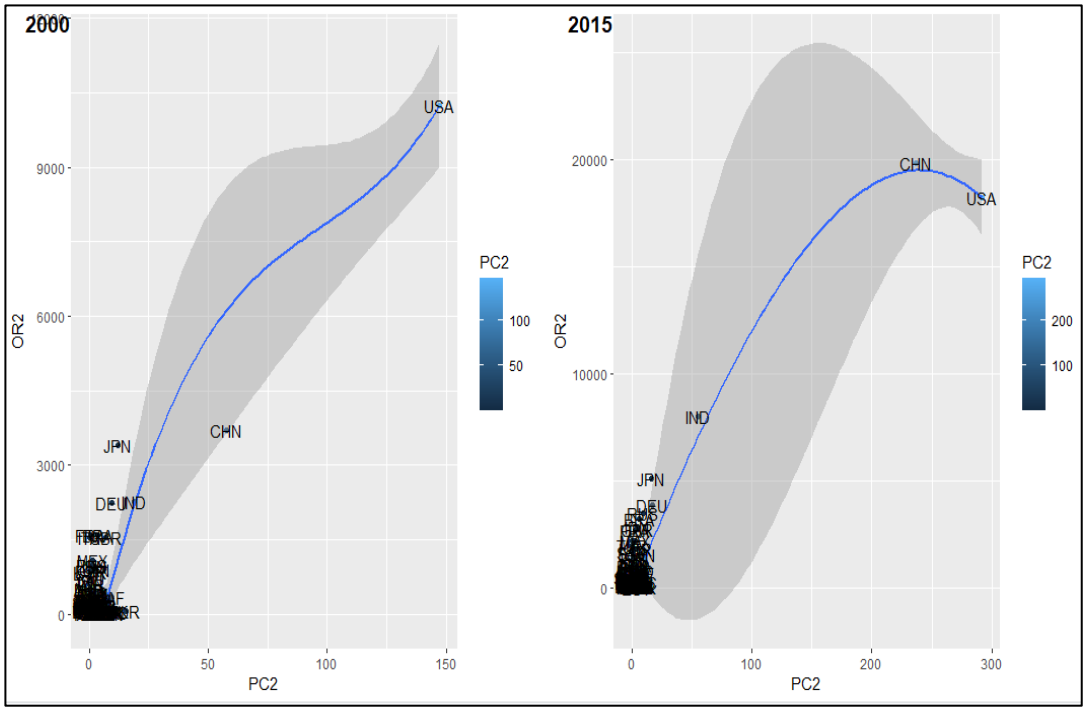


Figure 4.42 Loess Smoothing & OR2 vs. PC2 & Change from 2000 to 2015

Interpretation: When we implement the Loess Smoothing method to fit the model that is made up of OR2 and PC1 variables for the years 2000 and 2015, we can see that by 2015, the linearity seems to be more powerful than that of 2000. This indicates that the linearity between PC1 and OR2 variables increased as time passed. The linearity seems to be caused by countries that have excessive values. Moreover, the existence of non- linearity relationship in Figure 4.41 dominates the linearity. The excessive points are also seen in Figure 4.4 that are called as outliers. As for the loess smoothing plots that are composed of OR2 and PC2 variable for the years 2000 and 2015, the linearity is seen at both of the plots. However, the linearity does not turn into a non-linear relationship as shown in Figure 4.41. Thus, these methods tell that there are relationships between the variables PC1&PC2 and OR2.

4.3.3 Estimation of the Variance-Covariance Structure and Normality Check

Response Variable: OR2

a-) Estimation of the Variance – Covariance Structure for OR2

Let us estimate what the possible sigma Σ.could be.

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.97	0.95	0.94	0.92	0.91	0.90	0.89	0.88
2001	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.97	0.96	0.95	0.93	0.92	0.91	0.90	0.89
2002	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.98	0.96	0.95	0.94	0.93	0.92	0.91	0.90
2003	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.97	0.96	0.95	0.94	0.93	0.92	0.91
2004	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.97	0.96	0.95	0.94	0.93	0.92
2005	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.97	0.96	0.96	0.95	0.94	0.93
2006	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.97	0.97	0.96	0.95	0.95
2007	0.98	0.98	0.98	0.99	0.99	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.98	0.97	0.97	0.96
2008	0.97	0.97	0.98	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.98	0.97
2009	0.95	0.96	0.96	0.97	0.98	0.98	0.99	0.99	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.98
2010	0.94	0.95	0.95	0.96	0.97	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
2011	0.92	0.93	0.94	0.95	0.96	0.96	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99
2012	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00
2013	0.90	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00
2014	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.97	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00
2015	0.88	0.89	0.90	0.91	0.92	0.93	0.95	0.96	0.97	0.98	0.99	0.99	1.00	1.00	1.00	1.00

Figure 4.43 Correlation Matrix of OR2 based on years & Pearson Method

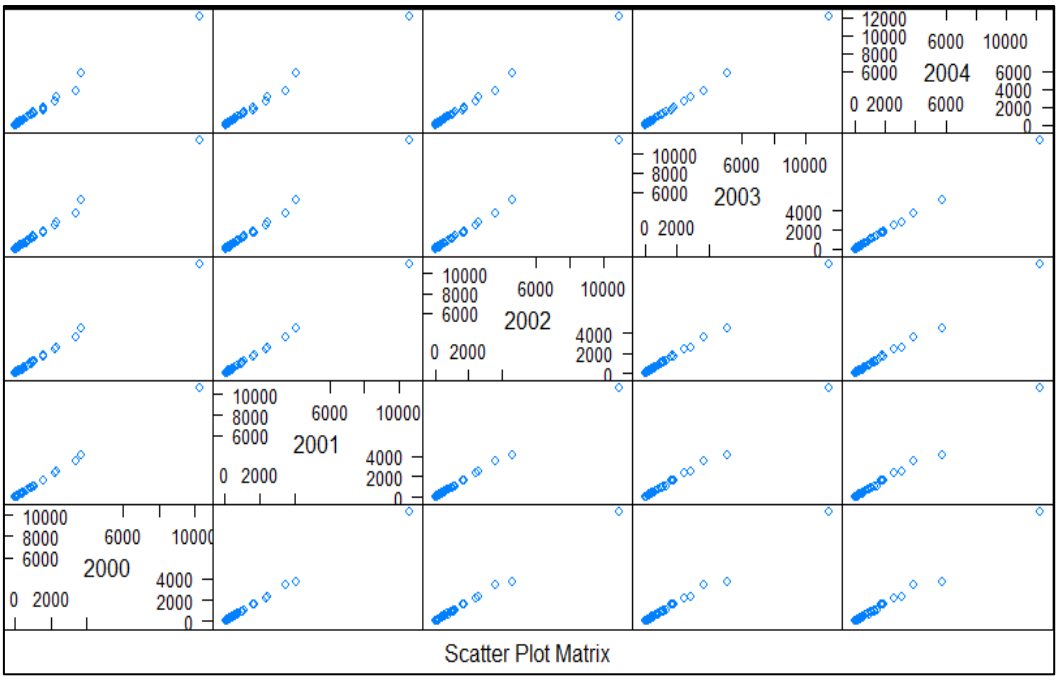


Figure 4.44 Scatter Plot & Correlation Plot of OR2 based on years

Interpretation: With the above correlation matrix and splom plot, we can see that off-diagonal correlation values are very close to each other. This variance-covariance structure is very similar **to "compound symmetry" or "exchangeable" structure**. Therefore, for the modeling part, we are assuming an "exchangeable" variance-covariance structure to derive marginal models.

b-) Normality Check of OR2

Table 4.20 Normality Assumption Check of OR2 & Before and After Transformation-***Shapiro Wilk Normality Test***

Years	Before Transformation	After transformation (log)
	p_valuebefore	p_valueafter
yr2000	1.9E-10	0.227
yr2001	2.2E-10	0.211
yr2002	2.5E-10	0.186
yr2003	2.6E-10	0.150
yr2004	2.7E-10	0.151
yr2005	2.7E-10	0.135
yr2006	3.2E-10	0.118

Table 4.19 (continued)

yr2007	3.5E-10	0.100
yr2008	4.1E-10	0.126
yr2009	3.5E-10	0.189
yr2010	3.3E-10	0.168
yr2011	3.2E-10	0.118
yr2012	2.5E-10	0.120
yr2013	2.1E-10	0.119
yr2014	1.8E-10	0.113
yr2015	1.5E-10	0.133

$$H_0: Observations\ are\ \ normally\ distributed$$

$$H_1: Observations\ are\ not\ normally\ distributed$$

$$\alpha = 0.01\ (significance\ level)$$

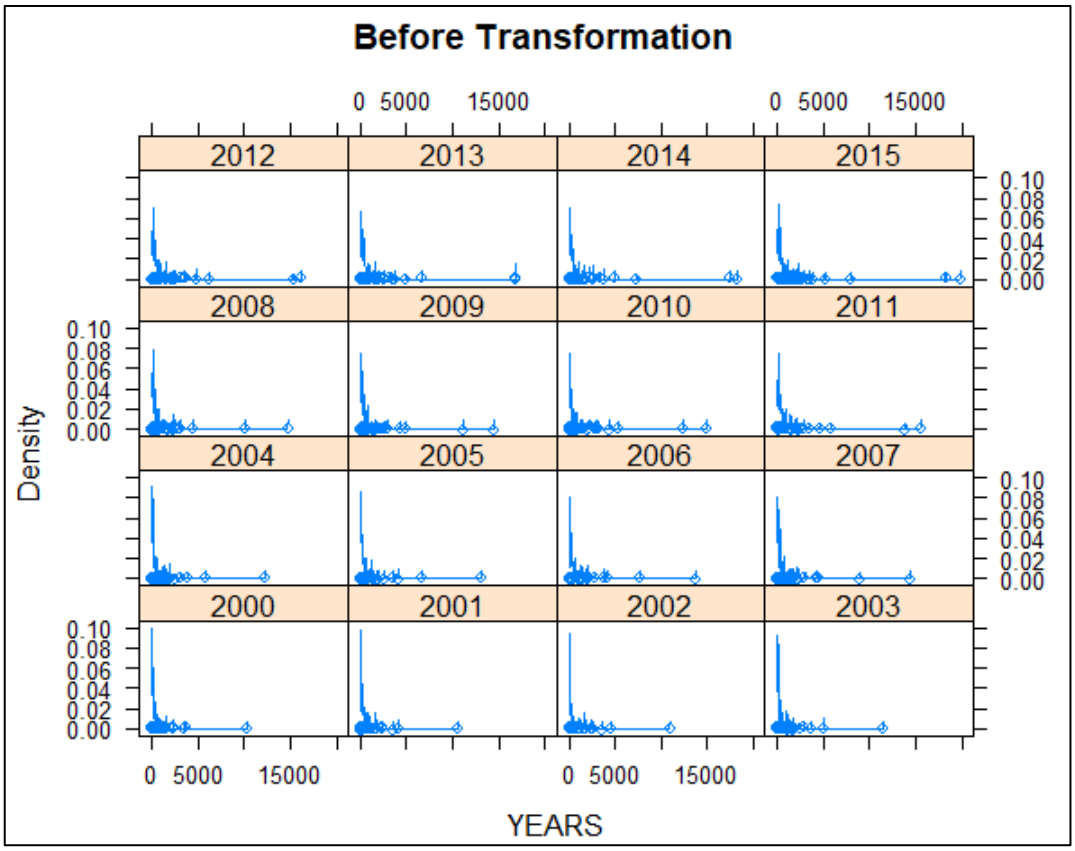


Figure 4.45 Density Plot for time based OR2 before transformation

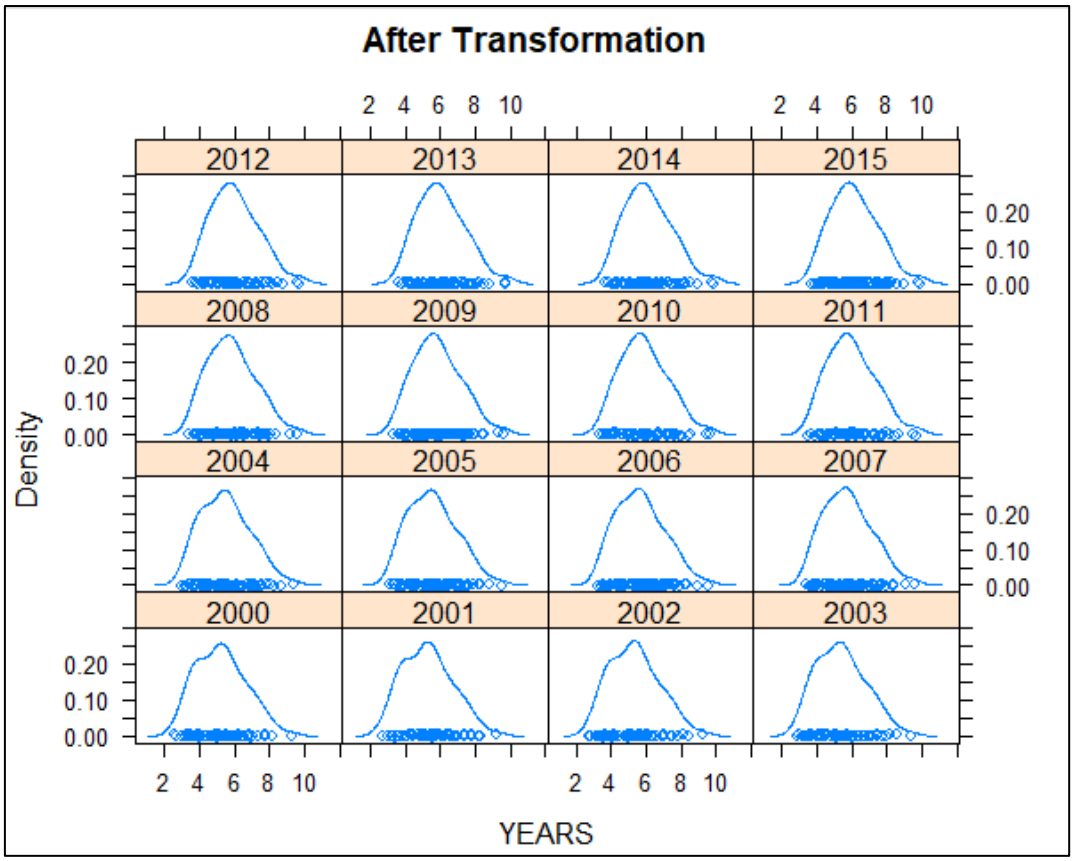


Figure 4.46 Density Plot for time based OR2 after transformation

As regards the normality checking of the response variable OR2, we can see that all of the p_values are smaller than (alpha =0.01), which means that we reject the normality of responses. Then, we need to perform a transformation. We did implement a ***"log transformation"*** to satisfy the normality. After putting this transformation into force, we satisfied the normality in that the p_values are greater than the alpha=0.01. Thus, for the modeling part, OR2 is taken as log (OR2). Moreover, from the density plots, while the density of original response variables does not seem to be normally distributed, after the transformation; we obtain density plots in the shape of normal distribution. For the other response variables OR1, OR3, and OR4, the same transformations were implemented and we satisfied the normality assumption.

4.3.4 Models

After estimating the variance-covariance structure, we come to the end. In this part, we design models based on four response variables that are OR1, OR2, OR3, and OR4. For the modeling part, we take the logarithm of the response variables in order to satisfy the normality assumption as shown in the previous part. For each response variable, 3 models (marginal, transitional, and random effect models) are obtained. Totally, 12 models are to be displayed for 4 response variables. In order to gauge the accuracy of the models, we divide the data into 2 groups that are called "train data set" and "test data set". By means of train data set, the models are brought into the open. Then, by virtue of the test data set, we measure the models' accuracy. For that reason, 86 % of the main data set is employed as a train data set and 14 % of the main data set is utilized as a test data set. By these measurements, the data set belonging to 70 Countries pertains to the train data set and the data set belonging to 11 Countries is dedicated to the test data set.

Note-1: Model Evaluation Metrics are to be displayed at the Conclusion & Discussion Part. For the modeling part, the significance level is accepted as 0.05.

Null and Alternative Hypothesis for models

$$H_0: Related\ Predictor\ variable\ is\ not\ significant$$

$$H_1: Related\ Predictor\ variable\ is\ significant$$

$\alpha = 0.05\ (significance\ level)$ And $Z\alpha_{/2} = 1.96\ (Z - tabulated\ value)$

- ✓ Estimated Variance-Covariance Structure is "**Compound Symmetry(exchangeable)"**
- ✓ Normality assuomption is satisfied when log transformation is applied and significance level is taken as 0.01

4.3.4.1 Models for OR2 response variable

In order to draw models based on the response variable OR2, we took marginal models, transitional models, and random effect models into consideration. According to the results, all the models mentioned above provide us with significant results. The marginal model is giving us the same predictor variables as Random Effect Models and Transition Models. However, it can not give us information about the individual effect of the countries on the response variable. As for transition models, by comparison with the marginal models, these models additionally provide us with the information about effect of one another variable that is called past information on the variation of the response variable. However, again as a marginal model, it can not enable us to see the individual effect/country effect/random intercept effect on the response variable. As regards random effect models, we have the same important variables as marginal and transition models. In this model, we also know whether or not the individual intercept model exists. By the results, the variation attributed to individual country effect seems to be high. Therefore, we are going to prefer to put random effect model into study.

Random Effects Model

Reduced Random Effects Model based on Subject Variation

We build up models based on the change in the values of the subjects in order to explain the variability in the response variable. The main aim is to focus on whether or not the variability attributed to the difference between the countries/individuals/subjects accounts for some part of the variation in the response variable.

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method
['lmerModLmerTest']
Formula: OR2 ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC9 + PC1
0 +      PC11 + PC13 + PC14 + (1 | ID)
   Data: datamtrain

REML criterion at convergence: -1253

Scaled residuals:
   Min     1Q Median     3Q    Max
-5.282 -0.460  0.025  0.549  3.927
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 ID       (Intercept) 1.9873   1.410
 Residual             0.0103   0.101
Number of obs: 1120, groups:  ID, 70

Fixed effects:
             Estimate Std. Error         df t value Pr(>|t|)
(Intercept)  5.52e+00   1.69e-01  6.95e+01   32.69  < 2e-16 ***
PC1         -1.14e-02   7.38e-04  1.05e+03  -15.41  < 2e-16 ***
PC2          2.00e-03   3.95e-04  1.04e+03    5.07  4.6e-07 ***
PC3         -2.27e-03   8.82e-04  1.04e+03   -2.57     0.01 *
PC4         -8.99e-03   1.36e-03  1.04e+03   -6.59  6.9e-11 ***
PC5          6.36e-03   1.50e-03  1.04e+03    4.23  2.6e-05 ***
PC6          5.77e-03   8.84e-04  1.04e+03    6.52  1.1e-10 ***
PC7         -4.94e-03   1.16e-03  1.04e+03   -4.27  2.2e-05 ***
PC9         -4.09e-03   9.36e-04  1.04e+03   -4.37  1.4e-05 ***
PC10        -1.10e-02   2.57e-03  1.04e+03   -4.29  1.9e-05 ***
PC11         6.99e-03   1.73e-03  1.04e+03    4.03  6.0e-05 ***
PC13        -1.39e-02   2.37e-03  1.04e+03   -5.85  6.5e-09 ***
PC14         6.14e-02   6.81e-04  1.04e+03   90.11  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*** and * Variables are significant ones in which we also obtained some of the important ones such as PC2, PC6, PC14 in reduced marginal and transition models that only include important variables. The most important part of this model is to determine whether or not the Random Effect (Country/Individual/Subject Effect) contributes significantly to the explanation in the change or variation of the response variable. Random Effect part of the model is seen from the red area. We can see that the variability seems to be important because the variance attributable to random intercept seems to be at the normal level and can have an effect on explaining the variation in the response variable. Therefore, adding a random effect to the model coming from the countries' individual effect may work well. This random effect adds an additional intercept effect (b_0) to the model. As regards the effects of variables, the coefficients of PC2, PC5, PC6, PC11, and PC14 variables are positive, meaning that they have a positive effect on the response variable. Upon there is an increment in these variables, the response variable grows as much as the number of coefficients of the above mentioned variables.

4.3.5 Conclusion and Discussion

Table 4.21 Model Evaluation Metrics for in case OR2 is used as response variable

Model Evaluation Metrics	OR2 response variable							
	Marginal Models		Transition Models		Random Effects Models based on subject variation		Random Effects Models based on time variation	
	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)
Root Mean Square Error (RMSE)	1.330	1.320	1.16	1.16	1.33	1.34	1.28	1.28
Mean Square Error (MSE)	1.780	1.730	1.34	1.34	1.78	1.78	1.63	1.64
Quasi-Likelihood Criteria (QIC)	830	830	1096	1096	NAN	NAN	NAN	NAN
Sum of Absolute Error (SAE)	213	210	179	179	213	214	202	202
Bias (bias)	-0.568	-0.588	1.09	1.09	-0.568	-0.567	-0.715	-0.725
Sum of Squared Errors (SSE)	313	305	221	220	313	314	288	289
Relative Squared Error (RSE)	1.480	1.450	1.14	1.14	1.48	1.49	1.36	1.37
Mean Absolute Error (MAE)	1.210	1.200	1.09	1.09	1.21	1.21	1.15	1.15
Mean Absolute Percent Error (MAPE)	0.260	0.257	0.218	0.218	0.26	0.26	0.252	0.252
Sum of Squared Log Error (SSLE)	8.370	8.160	8.29	8.22	8.37	8.38	7.79	7.81
Mean Squared Log Error (MSLE)	0.048	0.046	0.0503	0.0498	0.0476	0.0476	0.0443	0.0444
Percent Bias (percent_bias)	-0.162	-0.165	0.218	0.218	-0.162	-0.162	-0.187	-0.189
Relative Absolute Error (RAE)	1.290	1.270	1.16	1.16	1.29	1.29	1.22	1.22

Table 4.21 (continued)

Root Mean Squared Log Error (RMSLE)	**0.218**	**0.215**	**0.224**	**0.223**	**0.218**	**0.218**	**0.21**	**0.211**
Root Relative Squared Error (RRSE)	**1.220**	**1.200**	**1.07**	**1.07**	**1.22**	**1.22**	**1.17**	**1.17**

- The table summarizes the model evaluation metrics (MEM). We put 15 MEMs into force to assess which models are good at prediction. The MEMs are obtained by means of test data. At the beginning of the modeling, we had said that the data is divided into two groups. One of them is called "train data" that is made up of 86% of the data and the other is called "test data" that is composed of 14% of the data.
- According to results, the reduced models for each type of model can be chosen as useful models because of model parsimony and slight decrements/increments in MEM values. The model parsimony means that the model that has less and important variables should be preferred. Therefore, in order to build up models based on the OR2 response variable, **the reduced marginal, reduced transition and reduced random effect model that is set up on the basis of the subject effect can be taken as useful models. For someone who is looking for the effect of past information can get benefit from the reduced transition model because all of the transition models include past information as an important variable. Moreover, for someone who is interested in the variation attributed to the individual effect on the response variables can go on with the random effect intercept model. Marginal models are also preferred to make predictions for the new data sets.**
- For the purpose of making inferences on the most important predictor variables, we are going to choose a useful model for the OR2 response variable. After showing the most important variables that are coming from Principal Component Analysis, we are going to delve into the original predictor variables (OP) that are mostly correlated to the important predictor variables. Thus, for the response variable OR2, we are taking the

Reduced Random Effects Intercept Model into consideration to make comments on the predictor variables. The following table summarizes the original predictor variables that are correlated to the important PCs.

Table 4.22 Important PCs and Important Original Predictor Variables

Response Variables	Selected Useful Model	Important Predictor Variables (PCs)	Top 5 Original Important Variables that are mostly correlated(+/-) to important variables(PCs)
OR2	Random Effect Intercept Model (Individual Effect)	PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC9, PC10, PC11, PC13, PC14	PC1:OP41,OP42,OP43,(-OP39),(-OP38) PC2:OP29,OP52,OP56,OP55,OP53 PC3:(-OP31),(-OP33),(-OP32),OP34,OP36 PC4:OP13,OP24,OP11,OP9,OP20 PC5:OP11,OP20,(-OP44),(-OP7),OP54 PC6:(-OP50),(-OP54),OP15,OP18,OP14 PC7:OP54,OP50,(-OP8),(-OP2),(-OP35) PC9:OP25,OP27,OP7,(-OP46),OP35 PC10:OP19,OP37,OP23,OP22,OP21 PC11:(-OP19),OP28,OP45,(-OP17),OP48 PC13:OP46,OP17,OP19,(-OP10),OP26

The above table can be read as following (***The corresponding names of the original variables are available Table 5.12 under Appendix C.);***

For example;

1- Upon our response variable is OR2, our useful model is the Random Effect Intercept Model.

2- Then, when we take a look at the important predictor variables that are coming from Principal Component Analysis, they are PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC9, PC10, PC11 and PC13, and time-variable named PC14. Those are our significant predictor variables in order to build up the models.

3- So as to see which original predictor variables are more correlated to PCs, Table 4.5 is reviewed.

4- As an illustration, we see the PC1 is mostly correlated to OP41, OP42, OP43, (-OP39), and (-OP38). (**Sign (-) means that PC is negatively correlated to the original variable).**

5- OP41 = Wage and salaried workers, total (% of total employment) (modeled ILO estimate) from Table 5.12. The correlation between PC1 and OP41 is 0.89. This is also coefficient in PC as following;

(I): PC1 = 0.89*OP41 + 0.89*OP42+0.87*OP43……-0.89*OP38

6- Therefore, if PC1 is important, OP41 is also said to be slightly more significant than OP43 because PC1 is more related to OP41.

7- As a result, when our response variable is OR2= GDP, PPP (current international BillionUS$), and log transformation are performed on this response variable, the new dependent variables become log (GDP). In the Reduced Random Effect Model for this transformed target variable, we know that the coefficient of PC1 is negative, meaning as PC1 increases, there happens decrement in the log of the response variable. Therefore, in order to observe the effect of OP1 on the log (GDP, PPP), we need to keep the other original predictor variables shown in **(I)** constant and when there is an increment in the OP1 value, this gives rise to increment in the PC1. Then if PC1 increases, the log (GDP) decreases.

As a result, the model that we obtained can be written mathematically as below.

Table 4.23 Names of Cofficients

	Estimate	**Names of Coefficients**
(Intercept)	5.52e+00	β_0
PC1	-1.14e-02	β_1
PC2	2.00e-03	β_2
PC3	-2.27e-03	β_3
PC4	-8.99e-03	β_4
PC5	6.36e-03	β_5

Table 4.23 (continued)

PC6	5.77e-03	β_6
PC7	-4.94e-03	β_7
PC9	-4.09e-03	β_8
PC10	-1.10e-02	β_9
PC11	6.99e-03	β_{10}
PC13	-1.39e-02	β_{11}
PC14 = time	6.14e-02	β_{12}

$$OR2_{it} = \beta_0 + b_{0i} + \beta_1 * PC_1 + \beta_2 * PC_2 + \cdots .. \beta_{12} * PC_{14} + \varepsilon_{it}$$

$$b_{0i}$$

$$: Invidual\ intercept\ effect\ (Individual\ Country\ effect)$$

$$PC\ : Predictor\ Variables\ (Principal\ Components)$$

$$PC_{14} : Time\ variable$$

$$\beta : Coefficietns\ in\ the\ front\ of\ PCs.$$

$$\varepsilon_{it} : Error\ term$$

After the modeling part, we understood that PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC9, PC10, PC11, PC13, and PC14(time) are the significant variables to explain the variation in the response variable. Now, in order to better understand how the original predictor variables play a role in Principal Components and which original variables make their related Principal Components important, we can rewrite the above modeling by the degree of the relation of the original variables with Principal Components. Let us write all model again;

$$OR2_{it} = \beta_0 + b_{0i} + \beta_1 * PC_{WS\_VE} + \beta_2 * PC_{T\_IS} + \beta_3 * PC_{UNE} + \beta_4 * PC_{IS\_I} + \beta_5 * PC_I + \beta_6 * PC_{IN(-)} + \beta_7 * PC_{IN(+)} + \beta_8 * PC_{IS\_M} + \beta_9 * PC_{TI(+)} + \beta_{10} * PC_{TI(-)} + \beta_{11} * PC_{IF} + \beta_{12} * time + \varepsilon_{it}$$

$OR2_{it}$: **log of** GDP, PPP (current international BillionUS$)

PC_{WS_VE}: PC1 is attributable mostly to **wage and salaried workers and vulnerable employement**-related variables.

PC_{T_IS}: PC2 is attributable mostly to **trade and industry**-related variables.

PC_{UNE}: PC3 is attributable mostly to **unemployement**-related variables.

PC_{IS_I}: PC4 is attributable mostly to **industry and inflation**-related variables.

PC_{I}: PC5 is attributable mostly to **inflation**-related variables.

$PC_{IN(-)}$: PC6 is attributable mostly to **foreign direct investment**-related variables. PC6 is negatively correlated with the investment.

$PC_{IN(+)}$: PC7 is attributable mostly to **foreign direct investment**-related variables. PC7 is positively correlated with the investment.

PC_{IS_M}: PC9 is attributable mostly to **industry and manufacturing**-related variables.

$PC_{TI(+)}$: PC10 is attributable mostly to **net barter trade index**-related variables. PC10 is positively correlated with the net barter trade index.

$PC_{TI(-)}$: PC11 is attributable mostly to **net barter trade index**-related variables. PC11 is negatively correlated with the net barter trade index.

PC_{IF}: PC13 is attributable mostly to **individual life**-related variables

END OF DATA MINING-REGRESSION MODELING

End of Chapter 4

CHAPTER 5

CONCLUSION

This study is aimed at showing how to implement well-known data mining techniques on the data sets that are made up of economic indicators, to discover the hidden information about the possible economic relationships between the countries and to determine most frequently sold/exported products by Switzerland to the other countries.

As it is stated at the beginning of the study, the main considerations were simply to compare the countries' economic levels with each other so that we can understand which countries show similarities and differences with each other. Moreover, the other main focus area was to build up models based on the response variables that are really important economic indicators of a country such as Gross Domestic Product (GDP), Gross National Income (GNI), etc. by using the well-known data mining techniques.

This study is crucial for investigating the economical situation of the countries. Raising awareness of the similarities and differences between the countries can lay the groundwork for making the future studies focus on gaining insight into the possible solutions to handle the differences. In terms of educational purposes, this study can be a guideline to display how to effectively implement the data mining techniques such as clustering, association rule, classification, and regression models on the data sets.

Specifically, the study sought answers to the research questions by means of data mining techniques. In the end, we obtained some useful results and answers to our research questions. Chapter 1 provided readers the main aim of the study, the research questions that we had struggled for finding answers throughout the study, and the data mining techniques that we had used for making analysis of the data sets. The data mining is a technique that struggles for discovering the invisible

information of visible information. Although the determination of the research questions seems to be meaningless in data mining terminology, the reason why we put those questions on the study is because of our expectations about the results. We know that a supervised learning system of data mining is aimed at drawing models by setting up relationships between the variables. Since we also take advantage of the supervised learning algorithm, the pre-determined research questions were answered under this learning algorithm. In chapter 2, we started to talk about the description of the data. In order for the reader to better understand and not cause any confusion, the variables in the study were simply introduced. In the beginning, we had missing cases. In order to fill these missing values, the general treatments for missing cases were suggested and the missing cases in the data were treated with Last Observed Variable and mean imputation method.

In chapter 3, Data Mining comes into force and under the unsupervised learning algorithm, the two data mining techniques were considered. We performed the analysis of unsupervised learning by means of the data mining techniques that are clustering and association rule. We implemented the first analysis that is clustering on the data sets in order to make clusters of the countries. In order to get to the bottom of the best clusters of the countries, the four clustering methods were employed and one of the results was shared in section 3.1 and the other results were put into Appendix. The second analysis that is Association Rule was implemented on the 3 different high volumed data sets that are made up of the transactions. These transactions are composed of exported products of Switzerland to the other countries for the years of 2001, 2009 and 2018. The country of Switzerland was selected according to the best countries of 2019 published in U.S. News magazine. According to the analyses, we reach significant results in chapter 3. The first findings were obtained by the clustering algorithm. Among the clustering techniques, the k medoids algorithm made successfully the separation of the countries. Not affected by the anomaly points in the data sets, the k medoids algorithm provided us the best representative clusters of the countries. The possible number of clusters were taken either 3, 4, or 5. For each number of clusters, the one

cluster included only one country that is the USA. This can lead us to think that the USA is economically an improved/improving country among the others. Because the k medoids algorithm is not affected by the outliers, the USA comes as the only country that is contained in the only one cluster. In order to observe which countries are economically at a similar level with the USA, the other algorithms such as k means might be used. With the analysis of data sets, we both had the opportunity for discovering groups of the countries based on economic indicators and drawn comparisons between the different clustering techniques. Those were already stated as our main aims at the beginning of the study. After we performed clustering analysis and determined clusters of the countries, we passed to the next data mining technique that is called as association rule. This technique enabled us to gain insight into which products were most frequently exported/sold by Switzerland to the other countries. One of the most famous algorithms, which is called "Apriori", was utilized and the results showed that the products coded 3004 and 3002 were exported together by Switzerland, meaning if a country bought the product either 3004 or 3002, that country also imported the product coded either 3004 or 3002. This was one of the results of this technique. The reason why we selected this way of analysis was to show how a country performs so-called market basket strategy to analyze trade policy. In terms of educational purposes, the implementation of this algorithm on the exported products of a country was the first one in the literature. As for what the future studies can be, because of the fact that Apriori algorithm takes frequency into consideration, the analysts cannot know for sure the price of the products and we cannot have the information on the contribution of the exported products to the economy of the country. Therefore, a different data set can be collected so that the trading volume of the product can be based even if its frequency is less compared to others. The product whose trading volume is bigger than any other product can contribute to the economy of a country more than that of others, meaning instead of dealing with most frequently exported products,future studies can concentrate on the products whose prices are greater than that of others.

As for chapter 4, we passed to the second stage of data mining which is called supervised learning. At the beginning, Principal Component Analysis (PCA) was carried out to reduce the dimension of the data sets and combine some variables that are correlated to each other. After performing PCA, the new response and the new predictor variables were achieved, and by means of the PCA, we got rid of the problem of multicollinearity that is a serious problem during the modeling analysis. The PCA provided us 4 new response variables attributed to predetermined 13 original response variables and also it provided us 13 new predictor variables coming from 56 original predictor variables. The selection of the number of Principal Components (PC) was performed by predetermined threshold values for the total variance. Thus, the number of PCs can vary from analysis to analysis. After the new variables, which are Principal Components, are found, the modeling part of the study was set off.

In chapter 4, under supervised learning, the two data mining techniques were thought to build up both classification and panel data models. Those techniques were used for drawing models in order to predict the response variables by using predictor variables. For the classification modeling part, the 4 methods that are Decision Tree Modeling, Random Forest Modeling, Naïve Bayesian Classification Modeling, and Support Vector Machine Modeling were implemented and we reached the important modeling results. The response variables of the classification modeling were obtained by means of using the new response variables, namely, in the classification modeling, the response variables should have been categorical most of the time, and however, we had new numeric response variables. Thus, we performed a clustering algorithm by using the best parameters that come out in chapter 3 on the new numeric response variables and finally, we achieved 4 different categorical response variables. Each one was coming from different clustering techniques. Then, the models were built up for those new categorical response variables. By means of using the categorical response variable of R2 that comes from the k medoids clustering algorithm, we obtained the model with the highest accuracy rate. This indicated that k-medoids algorithm provided with the

best groups of the countries. The 4 models that are Decision Tree, Random Forest, Naïve Bayesian Classification and Support Vector Machine Models were achieved and we observed that when the categorical variables come from the groups formed by k medoids clustering algorithm, we had the highest accuracy rate by means of Support Vector Machine Models (SVMM). Therefore, it was suggested that SVMM be taken into consideration because it gave us the best score of accuracy rate. As for the panel data modeling part, we tried to get the benefit from the new response and predictor variables. However, while we did not come across any problem in the new predictor variables like multicollinearity problem, the new response variables did not satisfy the main assumption of panel data analysis that the response variables must come from the normal distribution. Although we tried every method such as transformation of the response variables to satisfy the normality assumption, we could not get the desired results. Then, we went back to the original data set and we selected the 4 important response variables that are considered as the most important economic indicator variables which inform us about the economic situation of a country. After the selection of the response variables, Panel Data Analysis was carried out by using new predictor variables and we achieved some useful models. One of these models that was designed under the response variable OR2 (GDP, PPP) was selected as a useful one by assessing the model evaluation metrics. The results of Panel Data Modeling showed that the best model was Random Effect Model that takes the intercept effect into consideration. This model says that each country has an individual effect on the variance in the response variable.

The results of the study suggest that the concept of data mining can effectively be implemented on the economical data sets. This study is therefore a pioneering attempt to adapt the concept of data mining to economical indicators of the countries. In this study, we used well-known data mining techniques and we obtained significant results. By means of famous data mining techniques, researchers can play more with the data and can turn it into a shape in which the different data analysis ways might be utilized to arrive at different results. Thus, for

further studies, different data analysis ways can be employed to obtain more different results from what we found in this study.

Thank you.

REFERENCES

Aram Sinnreich Associate Professor of Communication Studies, & Barbara Romzek Professor of Public Administration and Policy. (2019, August 27). To serve a free society, social media must evolve beyond data mining. Retrieved from https://theconversation.com/to-serve-a-free-society-social-media-must-evolve-beyond-data-mining-94704

McKell, K. (2018, March 27). Data mining: How it works, why it's important. Retrieved from https://universe.byu.edu/2018/03/27/data-mining-1/

Mariadoss,B.J.(n.d.). Retrieved From Https://Opentext.Wsu.Edu/Cpim/Chapter/5-2-Classifying-World-Economies/

Team, C. (n.d.). MoData. Retrieved from https://www.mo-data.com/what-is-the-difference-between-data-analytics-data-analysis-data-mining-data-science-machine-learning-big-data-and-predictive-analytics/

(n.d.). Retrieved from https://twitter.com/advanseez

Wicaksana, S. (2011, October 23). The way the brain learns best. Retrieved from https://www.slideshare.net/wicaksana/the-way-the-brain-learns-best

Chatterjee, I. (2019, October 22). A Comparative study of Clustering Algorithms. Retrieved from https://medium.com/analytics-vidhya/comparative-study-of-the-clustering-algorithms-54d1ed9ea732

Sinnreich, A., & Romzek, B. (2019, August 27). To serve a free society, social media must evolve beyond data mining. Retrieved from https://theconversation.com/to-serve-a-free-society-social-media-must-evolve-beyond-data-mining-94704

(n.d.). Retrieved from https://vizyonergenc.com/icerik/5-temel-soruda-veri-madenciligi-data-mining-nedir

Cluster. (n.d.). Retrieved May 16, 2020, from https://www.rdocumentation.org/packages/cluster/versions/2.1.0/topics/pam

Cluster Validation Statistics: Must Know Methods. (2018, October 22). Retrieved May 16, 2020, from https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/

Godfrey, K., Kassambara, Romero, J., Kumar, V., Cassiano, M., & G., G. (2018, October 21). Determining the Optimal Number Of Clusters: 3 Must Know Methods. Retrieved May 16, 2020, from

https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

Ggplot2 scatter plots: Quick start guide - R software and data visualization. (n.d.). Retrieved May 16, 2020, from http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization

Kabacoff, R. (n.d.). Scatterplots. Retrieved May 16, 2020, from https://www.statmethods.net/graphs/scatterplot.html

Splom. (n.d.). Retrieved May 16, 2020, from https://plotly.com/r/splom/

Cluster Validation Statistics: Must Know Methods. (2018, October 22). Retrieved May 16, 2020, from https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/

Sign In. (n.d.). Retrieved May 16, 2020, from https://rpubs.com/aephidayatuloh/clustervisual

Khongfak, S. (2019, December 25). The Ultimate Guide to Cluster Analysis in R. Retrieved May 16, 2020, from https://www.datanovia.com/en/blog/cluster-analysis-in-r-practical-guide/

K-means Cluster Analysis. (n.d.). Retrieved May 16, 2020, from https://uc-r.github.io/kmeans_clustering

What is a Dendrogram? How to use Dendrograms. (2020, April 20). Retrieved May 16, 2020, from https://www.displayr.com/what-is-dendrogram/

Im, S. (2017, February 20). Dendrogram. Retrieved May 16, 2020, from https://ldld.samizdat.cc/2017/dendrogram/

{{metadataController.pageTitle}}. (n.d.). Retrieved May 16, 2020, from https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781784390815/9/ch09lvl1sec97/transforming-data-into-transactions

Kadimisetty, A. (2018, May 13). Association Rule Mining in R. Retrieved May 16, 2020, from https://towardsdatascience.com/association-rule-mining-in-r-ddf2d044ae50

Prabhakaran, S. (n.d.). Eval (ez_write_tag ([[728,90],'r_statistics_co-box-3','ezslot_4',109,'0','0']));Association Mining (Market Basket Analysis). Retrieved May 16, 2020, from http://r-statistics.co/Association-Mining-With-R.html

Sign In. (n.d.). Retrieved May 16, 2020, from https://rpubs.com/dnchari/associationRules

Kassambara. (2019, December 25). Top R Color Palettes to Know for Great Data Visualization. Retrieved May 16, 2020, from https://www.datanovia.com/en/blog/top-r-color-palettes-to-know-for-great-data-visualization/

U.S.News. (2019). Best Countries 2019, 1–3. Retrieved from https://www.usnews.com/media/best-countries/overall-rankings-2019.pdf

Rouse, M. (2018, November 02). What is association rules (in data mining)? - Definition from WhatIs.com. Retrieved May 16, 2020, from https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining

Transactional Data. (n.d.). Retrieved May 16, 2020, from https://www.sciencedirect.com/topics/computer-science/transactional-data

Itc. (n.d.). Trade statistics for international business development. Retrieved May 16, 2020, from https://www.trademap.org/Index.aspx

Workman, D. (2020, April 30). Switzerland's Top 10 Exports. Retrieved May 16, 2020, from http://www.worldstopexports.com/switzerlands-top-10-exports/

Itc. (n.d.). Trade statistics for international business development. Retrieved May 16, 2020, from https://www.trademap.org/Product_SelCountry_TS.aspx?nvpm=1%7C757%7C%7C%7C%7CTOTAL%7C%7C%7C4%7C1%7C1%7C2%7C2%7C1%7C1%7C1%7C1

Decision Tree in R with Example. (n.d.). Retrieved May 16, 2020, from https://www.guru99.com/r-decision-trees.html

Machine Learning Decision Tree Classification Algorithm - Javatpoint. (n.d.). Retrieved May 16, 2020, from https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Will. (2016, February 27). Learn by Marketing. Retrieved May 16, 2020, from http://www.learnbymarketing.com/481/decision-tree-flavors-gini-info-gain/

RekhaMolala. (2020, March 23). Entropy, Information gain and Gini Index; the crux of a Decision Tree. Retrieved May 16, 2020, from https://blog.clairvoyantsoft.com/entropy-information-gain-and-gini-index-the-crux-of-a-decision-tree-99d0cdc699f4

Gini Index for Decision Trees. (2020, March 05). Retrieved May 16, 2020, from https://blog.quantinsti.com/gini-index/

Wjohnson. (n.d.). Wjohnson/lbm. Retrieved May 16, 2020, from https://github.com/wjohnson/lbm/blob/master/examples/decision-tree-flavors.R

Random forest. (2020, May 02). Retrieved June 08, 2020, from https://en.wikipedia.org/wiki/Random_forest

Understanding Confusion Matrix in R. (n.d.). Retrieved May 16, 2020, from https://www.datacamp.com/community/tutorials/confusion-matrix-calculation-r

Siddhant, Shuvayan, Chetan66, Puneet_r, & Akash1694. (2015, December 21). How does Complexity Parameter (CP) work in decision tree. Retrieved May 16, 2020, from https://discuss.analyticsvidhya.com/t/how-does-complexity-parameter-cp-work-in-decision-tree/6589

Learn by Marketing. (n.d.). Retrieved May 16, 2020, from http://www.learnbymarketing.com/tutorials/rpart-decision-trees-in-r/

McHugh, M. (2012). Interrater reliability: The kappa statistic. Retrieved May 16, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Yiu, T. (2019, August 14). Understanding Random Forest. Retrieved May 16, 2020, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Brownlee, J. (2019, August 12). Bagging and Random Forest Ensemble Algorithms for Machine Learning. Retrieved May 16, 2020, from https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/

Bhatia, N. (2019, June 27). What is Out of Bag (OOB) score in Random Forest? Retrieved May 16, 2020, from https://towardsdatascience.com/what-is-out-of-bag-oob-score-in-random-forest-a7fa23d710

Hatipoglu, E. (2018, July 13). Machine Learning -Classification - Support Vector Machine- Kernel Trick- Part 10. Retrieved May 16, 2020, from https://medium.com/@ekrem.hatipoglu/machine-learning-classification-support-vector-machine-kernel-trick-part-10-7ab928333158

Stecanella, B. (2020, March 30). An Introduction to Support Vector Machines (SVM). Retrieved May 16, 2020, from https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

FreeCodeCamp.org. (2020, February 10). SVM Machine Learning Algorithm Explained. Retrieved May 16, 2020, from https://www.freecodecamp.org/news/support-vector-machines/

Support Vector Machines in R. (n.d.). Retrieved May 16, 2020, from https://www.datacamp.com/community/tutorials/support-vector-machines-r

Gandhi, R. (2018, July 05). Support Vector Machine - Introduction to Machine Learning Algorithms. Retrieved June 09, 2020, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Stecanella, B. (2020, March 30). An Introduction to Support Vector Machines (SVM). Retrieved June 09, 2020, from https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

E1071. (n.d.). Retrieved May 16, 2020, from https://www.rdocumentation.org/packages/e1071/versions/1.7-3/topics/svm

Kowalczyki, A. (2018, November 14). Support Vector Regression with R. Retrieved May 16, 2020, from https://www.svm-tutorial.com/2014/10/support-vector-regression-r/

Almohamad, T. (2017, February 05). Hi, could anyone tell how the Epsilon-SVR perform the regression in Support Vector Machines (SVM)? Retrieved May 16, 2020, from https://www.researchgate.net/post/Hi_could_anyone_tell_how_the_Epsilon-SVR_perform_the_regression_in_Support_Vector_Machines_SVM

A Step By Step Guide to Implement Naive Bayes in R. (2019, May 22). Retrieved May 16, 2020, from https://www.edureka.co/blog/naive-bayes-in-r/

Understanding Naïve Bayes Classifier Using R. (2018, January 22). Retrieved May 16, 2020, from https://www.r-bloggers.com/understanding-naive-bayes-classifier-using-r/

Sign In. (n.d.). Retrieved May 16, 2020, from https://rpubs.com/riazakhan94/naive_bayes_classifier_e1071

Zhang, Z. (2016, June). Naïve Bayes classification in R. Retrieved May 16, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4930525/

(n.d.). Retrieved May 16, 2020, from https://www.saedsayad.com/naive_bayesian.htm

Scatter Plot Matrices - R Base Graphs. (n.d.). Retrieved May 16, 2020, from http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs

Normality Test in R. (n.d.). Retrieved May 16, 2020, from http://www.sthda.com/english/wiki/normality-test-in-r

Admin. (2019, April 22). How to Conduct an Anderson-Darling Test in R. Retrieved May 16, 2020, from https://www.statology.org/how-to-conduct-an-anderson-darling-test-in-r/

Ggplot2 density plot: Quick start guide - R software and data visualization. (n.d.). Retrieved May 16, 2020, from http://www.sthda.com/english/wiki/ggplot2-density-plot-quick-start-guide-r-software-and-data-visualization

Sign In. (n.d.). Retrieved May 16, 2020, from https://www.rpubs.com/prashant2007/536375

Correlation matrix: A quick start guide to analyze, format and visualize a correlation matrix using R software. (n.d.). Retrieved May 16, 2020, from http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software

(n.d.). Retrieved May 16, 2020, from http://www.math.armstrong.edu/statsonline/5/5.3.2.html

Dieye, I. (1967, August 01). Can we use the "predict" function for a gee model? Retrieved May 16, 2020, from https://stats.stackexchange.com/questions/316198/can-we-use-the-predict-function-for-a-gee-model?noredirect=1

How to Measure the Accuracy of Predictive Models. (n.d.). Retrieved May 16, 2020, from http://www.acheronanalytics.com/acheron-blog/how-to-measure-the-accuracy-of-predictive-models

Accuracy and Errors for Models. (n.d.). Retrieved May 16, 2020, from https://rcompanion.org/handbook/G_14.html

Root Mean Square Error. (n.d.). Retrieved May 16, 2020, from https://www.rforge.net/doc/packages/hydroGOF/rmse.html

MLmetrics. (n.d.). Retrieved May 16, 2020, from https://www.rdocumentation.org/packages/MLmetrics/versions/1.1.1/topics/MSE

(n.d.). Retrieved May 16, 2020, from https://www.ibm.com/support/knowledgecenter/SSLVMB_subs/statistics_casestudies_project_ddita/spss/tutorials/gee_wheeze_fit.html

YouthPrankYouthPrank 2311 silver badge66 bronze badges, & Jacobsgjacobsg 12166 bronze badges. (1967, July 01). R 'newdata' had 7 rows but variables found have 182 rows. Retrieved May 16, 2020, from https://stackoverflow.com/questions/47047204/r-newdata-had-7-rows-but-variables-found-have-182-rows

Parellada, A., & Parellada, A. (1965, June 01). Predict () Function for lmer Mixed Effects Models. Retrieved May 16, 2020, from https://stats.stackexchange.com/questions/174203/predict-function-for-lmer-mixed-effects-models

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., Fox, J. (2020, April 14). Predict.merMod: Predictions from a model at new data values in lme4: Linear Mixed-Effects Models using 'Eigen' and S4. Retrieved May 16, 2020, from https://rdrr.io/cran/lme4/man/predict.merMod.html

ArXiv, E. (2020, April 02). New Measure of Human Brain Processing Speed. Retrieved May 16, 2020, from https://www.technologyreview.com/2009/08/25/210267/new-measure-of-human-brain-processing-speed/

(n.d.). Retrieved May 16, 2020, from http://www.teachnology.com/teachers/methods/info_processing/

Braaten, E. (2020, April 17). Slow Processing Speed and the Brain. Retrieved May 16, 2020, from https://www.understood.org/en/learning-thinking-differences/child-learning-disabilities/information-processing-issues/at-a-glance-4-ways-brain-structure-and-chemistry-may-affect-processing-speed

Wicaksana, S. (2011, October 23). The way the brain learns best. Retrieved May 16, 2020, from https://www.slideshare.net/wicaksana/the-way-the-brain-learns-best

Oppong, T. (2018, October 03). 30 Days to a Smarter Brain (How to Rapidly Improve How You Think). Retrieved May 16, 2020, from

https://medium.com/personal-growth/30-days-to-a-smarter-brain-how-to-rapidly-improve-how-you-think-5a5fca4db3cc

Data Mining Explained. (n.d.). Retrieved May 16, 2020, from https://www.microstrategy.com/us/resources/introductory-guides/data-mining-explained

Scientific articles. (n.d.). Retrieved May 16, 2020, from http://gen.lib.rus.ec/scimag/?q=data+mining

Skica, T., Rodzinka, J., & Mroczek, T. (2015, September 01). Data Mining Approach in Determining the Relationships Between the Economy and the General Government Sector Size. Retrieved May 16, 2020, from https://content.sciendo.com/view/journals/fiqf/11/3/article-p1.xml

Amadeo, K. (2020, May 13). What Does Gross National Product Say About a Country? Retrieved May 16, 2020, from https://www.thebalance.com/what-is-the-gross-national-product-3305847

Adhikari, S. (2019, August 04). 14 Differences between GDP and GNP. Retrieved May 16, 2020, from https://www.publichealthnotes.com/differences-between-gdp-and-gnp/

Chappelow, J. (2020, April 29). Gross Domestic Product – GDP. Retrieved May 16, 2020, from https://www.investopedia.com/terms/g/gdp.asp

Amadeo, K. (2020, April 06). What Gross National Income Says About a Country Retrieved May 16, 2020, from https://www.thebalance.com/gross-national-income-4020738

Purchasing_power_parity. (2020, January 27). Retrieved May 16, 2020, from https://www.economicsonline.co.uk/Global_economics/Purchasing_power_parity.html

Elisabet.furioc. (2017, June 02). The foreign exchange market: Exchange rate systems: BBVA. Retrieved May 16, 2020, from https://www.bbva.com/en/foreign-currency-market-exchange-rate-systems/

Eckland, C. (2018, December 29). Types of Exchange Rates: Fixed, Floating, Spot, Dual etc - Interpretation. Retrieved May 16, 2020, from https://efinancemanagement.com/international-financial-management/types-of-exchange-rates

Chen, J. (2020, March 02). Exchange Rate Definition. Retrieved May 16, 2020, from https://www.investopedia.com/terms/e/exchangerate.asp

Zhao, Y. (2017). *DATA MINING APPLICATIONS WITH R*. Place of publication not identified: ELSEVIER ACADEMIC Press.

Zhao, Y. (2013). *R and data mining: Examples and case studies*. Oxford: Academic.

King, R. S. (2015). *Cluster Analysis and Data Mining: An Introduction*. Mercury Learning.

Mishra, P. (2016). *R data mining projects*. Place of publication not identified: Packt Publishing Limited.

Cichosz, P. (2015). *Data mining algorithms: Explained using R*. Chichester: John Wiley & Sons.

Torgo, L. (2017). *Data mining with R: Learning with case studies*. Boca Raton: CRC Press, Taylor & Francis Group.

Taniar, D. (2008). *Data mining and knowledge discovery technologies*. Hershey: IGI Pub.

Makhabel, B. (2015). *Learning data mining with R develop key skills and techniques with R to create and customize data mining algorithms*. Birmingham: Packt Publ.

Cirillo, A. (2017). *R data mining: Implement data mining techniques through practical use cases and real-world datasets*. Birmingham, England: Packt Publishing.

Kassambara, A. (2017). *Practical guide to Cluster Analysis in R: Unsupervised machine learning*. S.l.: CreateSpace Independent Publishing Platform.

Kassambara, A. (2017). *Practical guide to principal component methods in R*. United States: Stdha.com.

Cichosz, P. (2015). *Data mining algorithms: Explained using R*. Chichester: John Wiley & Sons.

What is the difference between Bagging and Boosting? ⋆ Quantdare. (2017, October 06). Retrieved May 16, 2020, from https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/

Nagpal, A. (2017, October 18). Decision Tree Ensembles- Bagging and Boosting. Retrieved May 16, 2020, from https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9

Machine learning predicts behavior of biological circuits. (2019, October 02). Retrieved May 16, 2020, from https://www.sciencedaily.com/releases/2019/10/191002165235.htm

Data Mining Techniques for Churn Mitigation/Detection. (n.d.). *Developing Churn Models Using Data Mining Techniques and Social Network Analysis,* 41-62. doi:10.4018/978-1-4666-6288-9.ch003

Siguenza-Guzman, L., Saquicela, V., Avila-Ordóñez, E., Vandewalle, J., & Cattrysse, D. (2015). Literature Review of Data Mining Applications in Academic Libraries. *The Journal of Academic Librarianship, 41*(4), 499-510. doi:10.1016/j.acalib.2015.06.007

Literature Review. (2018). *Advances in Data Mining and Database Management Predictive Analysis on Large Data for Actionable Knowledge,* 14-58. doi:10.4018/978-1-5225-5029-7.ch002

Apriori Algorithm - CodeBatch. (n.d.). Retrieved May 16, 2020, from https://www.sites.google.com/site/getallcodesyouwant/data-mining/apriori-algorithm

Mohityadav. (2020, April 04). Apriori Algorithm. Retrieved May 16, 2020, from https://www.geeksforgeeks.org/apriori-algorithm/

Soleymani, A., Pennekamp, F., Petchey, O. L., & Weibel, R. (2015). Developing and Integrating Advanced Movement Features Improves Automated Classification of Ciliate Species. *Plos One, 10*(12). doi:10.1371/journal.pone.0145345

APPENDIXES

Appendix A: Clustering

a-) Clustering Results (k-means, fuzzy and hierarchical clustering)

K-means Algorithm: The conclusions are made on the basis of the chosen countries. (if people want more information about the other countries clusters, they can assess them from written notes and the above tables)

Possible Clusters coming from k-means algorithm are summarized at the following table;

Table 5.1. K-Means Clusters for dataset-1&2

Countries	Scaled-Numeric Dataset-1		Percentage Dataset-2	
	k=4	k=5	k=3	k=4
MAR	1	1	2	3
PER	1	1	2	3
AGO	1	1	2	1
ARE	4	3	3	4
ARG	1	1	3	4
AUS	4	5	3	4
AZE	1	1	1	2
BEL	4	5	3	4
BGR	1	1	2	3
BLR	1	1	2	1
BRA	1	1	3	4
CAN	4	5	3	4
CHE	4	3	3	4
CHL	1	1	2	3
CHN	2	4	1	2
COL	1	1	2	3
CRI	1	1	2	3
CZE	1	1	3	4
DEU	4	5	3	4
DNK	4	5	3	4
DOM	1	1	2	3
ECU	1	1	2	3
EGY	1	1	2	3
ESP	4	5	3	4
EST	1	1	2	3

Table 5.1 (continued)

FIN	4	5	3	4
FRA	4	5	3	4
GBR	4	5	3	4
GHA	1	1	2	1
GRC	1	5	3	4
GTM	1	1	3	4
HKG	4	5	2	3
HUN	1	1	3	4
IDN	1	1	2	3
IND	1	1	2	1
IRL	4	5	2	3
IRN	1	1	3	4
IRQ	1	1	3	4
ISR	1	5	3	4
ITA	4	5	3	4
JOR	1	1	2	3
JPN	2	4	3	4
KAZ	1	1	1	1
KOR	1	5	2	3
LBN	1	1	3	4
LKA	1	1	2	1
LTU	1	1	2	3
LUX	4	3	3	4
LVA	1	1	2	3
MAC	4	3	1	2
MEX	1	1	3	4
MMR	1	1	1	2
MYS	1	1	2	3
NGA	1	1	2	1
NLD	4	5	3	4
NOR	4	3	3	4
NZL	4	5	3	4
OMN	1	5	3	4
PAK	1	1	2	3
PAN	1	1	2	1
PHL	1	1	2	3
POL	1	1	2	3
PRT	1	1	3	4
QAT	4	3	2	1
ROU	1	1	2	3
RUS	1	1	2	3
SAU	4	5	3	4
SGP	4	3	2	3
SRB	1	1	2	3
SVK	1	1	2	3
SVN	1	1	3	4
SWE	4	5	3	4
THA	1	1	2	3

Table 5.1 (continued)

TUN	1	1	2	3
TUR	1	1	2	3
TZA	1	1	2	1
UKR	1	1	2	3
URY	1	1	3	4
USA	3	2	3	4
VNM	1	1	2	1
ZAF	1	1	3	4

Fuzzy Clustering Algorithm: The conclusions are made on the basis of the chosen countries. (if people want more information about the other countries clusters, they can assess them from written notes and the above tables)

Possible Clusters coming from fuzzy clustering algorithm are summarized at the following table;

Table 5.2 Fuzzy Algorithm Clusters for dataset-1&2

Countries	Scaled-Numeric Dataset-1		Percentage Dataset-2	
	k=4	k=5	k=4	k=5
MAR	1	1	1	1
PER	1	1	2	2
AGO	1	1	2	3
ARE	2	2	3	4
ARG	3	3	3	4
AUS	2	2	3	4
AZE	1	1	2	3
BEL	2	2	4	5
BGR	1	3	1	2
BLR	1	3	2	3
BRA	3	4	3	4
CAN	4	5	3	4
CHE	2	2	4	5
CHL	3	3	1	1
CHN	4	4	2	3
COL	1	1	1	1
CRI	1	3	1	1
CZE	3	4	3	4
DEU	4	5	4	5
DNK	2	2	4	5
DOM	1	1	1	2
ECU	1	1	3	1
EGY	1	1	3	4
ESP	4	4	4	5
EST	3	3	1	2
FIN	2	2	4	5
FRA	4	5	4	5

Table 5.2 (continued)

GBR	4	5	4	5
GHA	1	1	2	3
GRC	3	4	4	5
GTM	1	1	3	4
HKG	2	2	1	1
HUN	3	3	4	5
IDN	1	1	2	3
IND	3	4	2	3
IRL	2	2	1	1
IRN	3	3	3	4
IRQ	1	1	3	4
ISR	4	4	3	4
ITA	4	5	4	5
JOR	1	1	3	4
JPN	4	5	4	5
KAZ	3	3	2	3
KOR	3	4	1	2
LBN	1	1	3	4
LKA	1	1	2	3
LTU	3	3	1	2
LUX	2	2	4	5
LVA	3	3	1	2
MAC	2	2	2	3
MEX	3	3	4	5
MMR	1	1	2	3
MYS	3	3	1	2
NGA	1	1	2	3
NLD	2	2	4	5
NOR	2	2	4	5
NZL	4	4	3	4
OMN	3	4	3	4
PAK	1	1	3	4
PAN	3	3	2	3
PHL	1	1	1	2
POL	3	3	1	1
PRT	3	4	4	5
QAT	2	2	2	3
ROU	3	3	1	2
RUS	3	4	1	2
SAU	4	4	3	4
SGP	2	2	1	2
SRB	1	1	1	2
SVK	3	3	1	2
SVN	3	4	4	5
SWE	2	2	3	4
THA	1	3	1	1
TUN	1	1	3	4
TUR	3	3	2	2
TZA	1	1	2	3
UKR	1	1	1	1
URY	3	3	3	4
USA	4	5	4	5

Table 5.2 (continued)

VNM	1	1	2	3
ZAF	1	1	3	4

Hierarchical Clustering Algorithm: **The conclusions are made on the basis of the chosen countries. (if people want more information about the other countries clusters, they can assess them from written notes and the above tables)**

Possible Clusters coming from hierarchical clustering algorithm are summarized at the following table;

Table 5.3 HC Clustering for dataset-1

Countries	Scaled-Numeric Dataset-1									
	Single		Complete		Average		Ward.D		Ward.D2	
	k=4	k=5	k=4	k=5	k=4	k=5	k=4	k=5	k=4	k=5
MAR	1	1	1	1	1	1	1	1	1	1
PER	1	1	1	1	1	1	1	1	1	1
AGO	1	1	1	1	1	1	1	1	1	1
ARE	1	1	2	2	1	2	2	2	2	2
ARG	1	1	1	1	1	1	1	1	1	1
AUS	1	1	1	1	1	1	3	3	3	3
AZE	1	1	1	1	1	1	1	1	1	1
BEL	1	1	1	1	1	1	3	3	3	3
BGR	1	1	1	1	1	1	1	1	1	1
BLR	1	1	1	1	1	1	1	1	1	1
BRA	1	1	1	1	1	1	1	1	1	1
CAN	1	1	1	1	1	1	3	3	3	3
CHE	1	1	2	2	1	2	2	2	2	2
CHL	1	1	1	1	1	1	1	1	1	1
CHN	2	2	3	3	2	3	3	4	4	4
COL	1	1	1	1	1	1	1	1	1	1
CRI	1	1	1	1	1	1	1	1	1	1
CZE	1	1	1	1	1	1	3	3	1	1
DEU	1	1	1	4	1	1	3	4	3	3
DNK	1	1	1	1	1	1	3	3	3	3
DOM	1	1	1	1	1	1	1	1	1	1
ECU	1	1	1	1	1	1	1	1	1	1
EGY	1	1	1	1	1	1	1	1	1	1

Table 5.3 (continued)

ESP	1	1	1	1	1	1	3	3	3	3
EST	1	1	1	1	1	1	1	1	1	1
FIN	1	1	1	1	1	1	3	3	3	3
FRA	1	1	1	4	1	1	3	4	3	3
GBR	1	1	1	4	1	1	3	4	3	3
GHA	1	1	1	1	1	1	1	1	1	1
GRC	1	1	1	1	1	1	3	3	3	3
GTM	1	1	1	1	1	1	1	1	1	1
HKG	1	1	1	1	1	1	3	3	3	3
HUN	1	1	1	1	1	1	1	1	1	1
IDN	1	1	1	1	1	1	1	1	1	1
IND	1	1	1	1	1	1	1	1	1	1
IRL	1	1	1	1	1	1	3	3	3	3
IRN	1	1	1	1	1	1	1	1	1	1
IRQ	1	1	1	1	1	1	1	1	1	1
ISR	1	1	1	1	1	1	3	3	3	3
ITA	1	1	1	4	1	1	3	4	3	3
JOR	1	1	1	1	1	1	1	1	1	1
JPN	1	3	1	4	3	4	3	4	3	3
KAZ	1	1	1	1	1	1	1	1	1	1
KOR	1	1	1	1	1	1	3	3	3	3
LBN	1	1	1	1	1	1	1	1	1	1
LKA	1	1	1	1	1	1	1	1	1	1
LTU	1	1	1	1	1	1	1	1	1	1
LUX	1	1	2	2	1	2	2	2	2	2
LVA	1	1	1	1	1	1	1	1	1	1
MAC	1	1	2	2	1	2	2	2	2	2
MEX	1	1	1	1	1	1	1	1	1	1
MMR	1	1	1	1	1	1	1	1	1	1
MYS	1	1	1	1	1	1	1	1	1	1
NGA	1	1	1	1	1	1	1	1	1	1
NLD	1	1	1	1	1	1	3	3	3	3
NOR	1	1	2	2	1	2	2	2	2	2
NZL	1	1	1	1	1	1	3	3	3	3
OMN	1	1	1	1	1	1	3	3	3	3
PAK	1	1	1	1	1	1	1	1	1	1
PAN	1	1	1	1	1	1	1	1	1	1
PHL	1	1	1	1	1	1	1	1	1	1
POL	1	1	1	1	1	1	1	1	1	1
PRT	1	1	1	1	1	1	3	3	3	3
QAT	3	4	2	2	1	2	2	2	2	2
ROU	1	1	1	1	1	1	1	1	1	1

Table 5.3 (continued)

RUS	1	1	1	1	1	1	1	1	1	1
SAU	1	1	1	1	1	1	3	3	3	3
SGP	1	1	2	2	1	2	2	2	2	2
SRB	1	1	1	1	1	1	1	1	1	1
SVK	1	1	1	1	1	1	1	1	1	1
SVN	1	1	1	1	1	1	3	3	3	3
SWE	1	1	1	1	1	1	3	3	3	3
THA	1	1	1	1	1	1	1	1	1	1
TUN	1	1	1	1	1	1	1	1	1	1
TUR	1	1	1	1	1	1	1	1	1	1
TZA	1	1	1	1	1	1	1	1	1	1
UKR	1	1	1	1	1	1	1	1	1	1
URY	1	1	1	1	1	1	1	1	1	1
USA	4	5	4	5	4	5	4	5	4	5
VNM	1	1	1	1	1	1	1	1	1	1
ZAF	1	1	1	1	1	1	1	1	1	1

.

Table 5.4 HC Clustering for dataset-2

Countries	Percentage Dataset-2									
	Single		Complete		Average		Ward.D		Ward.D2	
	k=4	k=5	k=4	k=5	k=4	k=5	k=4	k=5	k=4	k=5
MAR	1	1	1	1	1	1	1	1	1	1
PER	1	1	1	1	1	1	1	2	1	1
AGO	1	1	2	2	1	1	1	2	2	2
ARE	1	2	3	3	2	2	2	3	3	3
ARG	1	1	3	4	1	1	2	3	3	3
AUS	1	1	3	4	1	1	2	3	3	3
AZE	2	3	4	5	3	3	3	4	4	4
BEL	1	1	3	4	1	1	4	5	3	5
BGR	1	1	1	1	1	1	1	1	1	1
BLR	1	1	1	1	1	1	1	2	2	2
BRA	1	1	3	4	1	1	2	3	3	3
CAN	1	1	3	4	1	1	2	3	3	3
CHE	1	1	3	4	1	1	4	5	3	5
CHL	1	1	1	1	1	1	1	1	1	1
CHN	2	3	4	5	3	3	3	4	4	4
COL	1	1	1	1	1	1	1	1	1	1
CRI	1	1	3	4	1	1	1	1	1	1
CZE	1	1	3	4	1	1	2	3	3	3
DEU	1	1	3	4	1	1	4	5	3	5

Table 5.4 (continued)

DNK	1	1	3	4	1	1	4	5	3	5
DOM	1	1	1	1	1	1	1	1	1	1
ECU	1	1	3	4	1	1	1	1	1	1
EGY	1	1	3	4	1	1	1	1	1	1
ESP	1	1	3	4	1	1	4	5	3	5
EST	1	1	1	1	1	1	1	1	1	1
FIN	1	1	3	4	1	1	4	5	3	5
FRA	1	1	3	4	1	1	4	5	3	5
GBR	1	1	3	4	1	1	4	5	3	5
GHA	1	1	2	2	1	1	1	2	2	2
GRC	1	1	3	4	1	1	4	5	3	5
GTM	1	1	3	4	1	1	2	3	3	3
HKG	1	1	1	1	1	1	1	1	1	1
HUN	1	1	3	4	1	1	4	5	3	5
IDN	1	1	1	1	1	1	1	2	1	1
IND	1	1	1	1	1	1	1	2	2	2
IRL	1	1	3	4	1	1	1	1	1	1
IRN	1	1	3	4	1	1	2	3	3	3
IRQ	1	1	3	4	1	1	4	5	3	5
ISR	1	1	3	4	1	1	2	3	3	3
ITA	1	1	3	4	1	1	4	5	3	5
JOR	1	1	3	4	1	1	2	3	3	3
JPN	1	1	3	4	1	1	4	5	3	5
KAZ	1	1	4	5	1	1	3	4	4	4
KOR	1	1	1	1	1	1	1	1	1	1
LBN	1	1	3	4	1	1	2	3	3	3
LKA	1	1	1	1	1	1	1	2	2	2
LTU	1	1	1	1	1	1	1	1	1	1
LUX	1	1	3	4	1	1	4	5	3	5
LVA	1	1	1	1	1	1	1	1	1	1
MAC	1	1	4	5	1	1	3	4	4	4
MEX	1	1	3	4	1	1	4	5	3	5
MMR	3	4	4	5	3	4	3	4	4	4
MYS	1	1	1	1	1	1	1	2	1	1
NGA	1	1	2	2	1	1	1	2	2	2
NLD	1	1	3	4	1	1	4	5	3	5
NOR	1	1	3	4	1	1	4	5	3	5
NZL	1	1	3	4	1	1	2	3	3	3
OMN	1	1	3	3	1	1	2	3	3	3
PAK	1	1	3	4	1	1	1	1	1	1
PAN	1	1	1	1	1	1	1	2	2	2
PHL	1	1	1	1	1	1	1	2	1	1

Table 5.4 (continued)

POL	1	1	1	1	1	1	1	1	1	1
PRT	1	1	3	4	1	1	4	5	3	5
QAT	4	5	2	2	4	5	2	3	2	2
ROU	1	1	1	1	1	1	1	1	1	1
RUS	1	1	1	1	1	1	1	1	1	1
SAU	1	1	3	4	1	1	2	3	3	3
SGP	1	1	1	1	1	1	1	2	1	1
SRB	1	1	1	1	1	1	1	1	1	1
SVK	1	1	1	1	1	1	1	1	1	1
SVN	1	1	3	4	1	1	4	5	3	5
SWE	1	1	3	4	1	1	2	3	3	3
THA	1	1	1	1	1	1	1	1	1	1
TUN	1	1	3	4	1	1	1	1	1	1
TUR	1	1	1	1	1	1	1	2	1	1
TZA	1	1	2	2	1	1	1	2	2	2
UKR	1	1	3	4	1	1	2	3	3	3
URY	1	1	3	4	1	1	2	3	3	3
USA	1	1	3	4	1	1	4	5	3	5
VNM	1	1	1	1	1	1	1	2	2	2
ZAF	1	1	3	4	1	1	2	3	3	3

Appendix B: Association Rule

a-) Export Products Coded 2 and 4 digits

Table 5.5 Export Products - 2 digits codes

Code-2 Digits	Product label
01	Live animals
02	Meat and edible meat offal
03	Fish and crustaceans, molluscs and other aquatic invertebrates
04	Dairy produce; birds' eggs; natural honey; edible products of animal origin, not elsewhere ...
05	Products of animal origin, not elsewhere specified or included
06	Live trees and other plants; bulbs, roots and the like; cut flowers and ornamental foliage
07	Edible vegetables and certain roots and tubers
08	Edible fruit and nuts; peel of citrus fruit or melons
09	Coffee, tea, maté and spices
10	Cereals
11	Products of the milling industry; malt; starches; inulin; wheat gluten
12	Oil seeds and oleaginous fruits; miscellaneous grains, seeds and fruit; industrial or medicinal ...
13	Lac; gums, resins and other vegetable saps and extracts
14	Vegetable plaiting materials; vegetable products not elsewhere specified or included
15	Animal or vegetable fats and oils and their cleavage products; prepared edible fats; animal ...

Table 5.5 (continued)

16	Preparations of meat, of fish or of crustaceans, molluscs or other aquatic invertebrates
17	Sugars and sugar confectionery
18	Cocoa and cocoa preparations
19	Preparations of cereals, flour, starch or milk; pastrycooks' products
20	Preparations of vegetables, fruit, nuts or other parts of plants
21	Miscellaneous edible preparations
22	Beverages, spirits and vinegar
23	Residues and waste from the food industries; prepared animal fodder
24	Tobacco and manufactured tobacco substitutes
25	Salt; sulphur; earths and stone; plastering materials, lime and cement
26	Ores, slag and ash
27	Mineral fuels, mineral oils and products of their distillation; bituminous substances; mineral **...**
28	Inorganic chemicals; organic or inorganic compounds of precious metals, of rare-earth metals, **...**
29	Organic chemicals
30	Pharmaceutical products
31	Fertilisers
32	Tanning or dyeing extracts; tannins and their derivatives; dyes, pigments and other colouring **...**
33	Essential oils and resinoids; perfumery, cosmetic or toilet preparations
34	Soap, organic surface-active agents, washing preparations, lubricating preparations, artificial **...**

Table 5.5 (continued)

35	Albuminoidal substances; modified starches; glues; enzymes
36	Explosives; pyrotechnic products; matches; pyrophoric alloys; certain combustible preparations
37	Photographic or cinematographic goods
38	Miscellaneous chemical products
39	Plastics and articles thereof
40	Rubber and articles thereof
41	Raw hides and skins (other than furskins) and leather
42	Articles of leather; saddlery and harness; travel goods, handbags and similar containers; articles **...**
43	Furskins and artificial fur; manufactures thereof
44	Wood and articles of wood; wood charcoal
45	Cork and articles of cork
46	Manufactures of straw, of esparto or of other plaiting materials; basketware and wickerwork
47	Pulp of wood or of other fibrous cellulosic material; recovered (waste and scrap) paper or **...**
48	Paper and paperboard; articles of paper pulp, of paper or of paperboard
49	Printed books, newspapers, pictures and other products of the printing industry; manuscripts, **...**
50	Silk
51	Wool, fine or coarse animal hair; horsehair yarn and woven fabric
52	Cotton
53	Other vegetable textile fibres; paper yarn and woven fabrics of paper yarn

Table 5.5 (continued)

54	Man-made filaments; strip and the like of man-made textile materials
55	Man-made staple fibres
56	Wadding, felt and nonwovens; special yarns; twine, cordage, ropes and cables and articles thereof
57	Carpets and other textile floor coverings
58	Special woven fabrics; tufted textile fabrics; lace; tapestries; trimmings; embroidery
59	Impregnated, coated, covered or laminated textile fabrics; textile articles of a kind suitable **...**
60	Knitted or crocheted fabrics
61	Articles of apparel and clothing accessories, knitted or crocheted
62	Articles of apparel and clothing accessories, not knitted or crocheted
63	Other made-up textile articles; sets; worn clothing and worn textile articles; rags
64	Footwear, gaiters and the like; parts of such articles
65	Headgear and parts thereof
66	Umbrellas, sun umbrellas, walking sticks, seat-sticks, whips, riding-crops and parts thereof
67	Prepared feathers and down and articles made of feathers or of down; artificial flowers; articles **...**
68	Articles of stone, plaster, cement, asbestos, mica or similar materials
69	Ceramic products
70	Glass and glassware
71	Gold,natural or cultured pearls, precious or semi-precious stones, precious metals, metals clad **...**
72	Iron and steel

Table 5.5 (continued)

73	Articles of iron or steel
74	Copper and articles thereof
75	Nickel and articles thereof
76	Aluminium and articles thereof
78	Lead and articles thereof
79	Zinc and articles thereof
80	Tin and articles thereof
81	Other base metals; cermets; articles thereof
82	Tools, implements, cutlery, spoons and forks, of base metal; parts thereof of base metal
83	Miscellaneous articles of base metal
84	Machinery, mechanical appliances, nuclear reactors, boilers; parts thereof
85	Electrical machinery and equipment and parts thereof; sound recorders and reproducers, television **...**
86	Railway or tramway locomotives, rolling stock and parts thereof; railway or tramway track fixtures **...**
87	Vehicles other than railway or tramway rolling stock, and parts and accessories thereof
88	Aircraft, spacecraft, and parts thereof
89	Ships, boats and floating structures
90	Optical, photographic, cinematographic, measuring, checking, precision, medical or surgical **...**
91	Clocks and watches and parts thereof
92	Musical instruments; parts and accessories of such articles

Table 5.5 (continued)

93	Arms and ammunition; parts and accessories thereof
94	Furniture; bedding, mattresses, mattress supports, cushions and similar stuffed furnishings; ...
95	Toys, games and sports requisites; parts and accessories thereof
96	Miscellaneous manufactured articles
97	Works of art, collectors' pieces and antiques

Table 5.6. Sub-Product Labels of Exported Products - Pharmaceutical Products

Code-2 digits	**Product label**
30	**Pharmaceutical products**
Code - 4 digits	**Sub-Product label**
3004	Medicaments consisting of mixed or unmixed products for therapeutic or prophylactic uses, put ...
3002	Human blood; animal blood prepared for therapeutic, prophylactic or diagnostic uses; antisera ...
3006	Pharmaceutical preparations and products of subheadings 3006.10.10 to 3006.60.90
3003	Medicaments consisting of two or more constituents mixed together for therapeutic or prophylactic ...
3005	Wadding, gauze, bandages and the like, e.g. dressings, adhesive plasters, poultices, impregnated ...
3001	Dried glands and other organs for organo-therapeutic uses, whether or not powdered; extracts ...

b-) Data sets for Association Rule Mining (2001, 2009 and 2018 export products of Switzerland)

Table 5.7 Exports of Switzerland to 99 Countries in 2001

Exports of Switzerland to the world of countries for the year of 2001	ITEM-1 (sold/bought item-1)	ITEM-2 (sold/bought item-2)	ITEM-3 (sold/bought item-3)	ITEM-4 (sold/bought item-4)
Germany	3002	3004	2941	. . .
United States	3004	7110	9102	. . .
United Kingdom	3004	2933	9102	. . .
China	3004	8479	8477	. . .
France	3004	7113	9102	. . .
India	8477	3004	8448	. . .
Italy	3004	9102	3002	. . .
...	. . .	. . .	. . .	. . .

Table 5.8 Exports of Switzerland to 99 Countries in 2009

Exports of Switzerland to the world of countries for the year of 2009	ITEM-1 (sold/bought item-1)	ITEM-2 (sold/bought item-2)	ITEM-3 (sold/bought item-3)	ITEM-4 (sold/bought item-4)
Germany	3004	3002	2716	. . .
United States	3004	3002	9021	. . .
United Kingdom	3004	3002	7113	. . .

Table 5.8 (continued)

China	9102	3004	8486	. . .
France	3004	3002	2716	. . .
India	3004	7114	9999	. . .
Italy	2716	3002	3004	. . .
...	. . .	. . .	. . .	. . .

Table 5.9 Exports of Switzerland to 99 Countries in 2018

Exports of Switzerland to the world of countries for the year of 2018	ITEM-1 (sold/bought item-1)	ITEM-2 (sold/bought item-2)	ITEM-3 (sold/bought item-3)	ITEM-4 (sold/bought item-4)
Germany	3004	3002	2933	. . .
United States	3004	3002	9021	. . .
United Kingdom	7113	3004	3002	. . .
China	7108	3002	3004	. . .
France	7108	7113	3004	. . .
India	7108	7106	3004	. . .
Italy	3004	7108	2716	. . .
...	. . .	. . .	. . .	. . .

c-) Popular Product Descriptions by the results

Table 5.10 Product Descriptions by Analysis Results

Product Codes	Product Description
3004	Medicaments consisting of mixed or unmixed products for therapeutic or prophylactic uses, put up in measured doses "incl. those in the form of transdermal administration" or in forms or packings for retail sale
3002	Human blood; animal blood prepared for therapeutic, prophylactic or diagnostic uses; antisera and other blood fractions and immunological products, whether or not modified or obtained by means of biotechnological processes; vaccines, toxins, cultures of micro-organisms (excluding yeasts) and similar products
3302	Mixtures of odoriferous substances and mixtures, incl. alcoholic solutions, based on one or more of these substances, of a kind used as raw materials in industry; other preparations based on odoriferous substances, of a kind used for the manufacture of beverages
3808	Insecticides, rodenticides, fungicides, herbicides, anti-sprouting products and plant-growth regulators, disinfectants and similar products, put up for retail sale or as preparations or articles, e.g. sulphur-treated bands, wicks and candles, and fly-papers
7108	Gold, incl. gold plated with platinum, unwrought or not further worked than semi-manufactured or in powder form

Table 5.10 (continued)

9018	Instruments and appliances used in medical, surgical, dental or veterinary sciences, incl. scintigraphic apparatus, other electro-medical apparatus and sight-testing instruments, n.e.s.
9021	Orthopaedic appliances, incl. crutches, surgical belts and trusses; splints and other fracture appliances; artificial parts of the body; hearing aids and other appliances which are worn or carried, or implanted in the body, to compensate for a defect or disability
9102	Wrist-watches, pocket-watches and other watches, incl. stop-watches (excluding of precious metal or of metal clad with precious metal)
9101	Wrist-watches, pocket-watches and other watches, incl. stop-watches, with case of precious metal or of metal clad with precious metal (excluding with backs made of steel)
9108	Watch movements, complete and assembled
9999	Commodities not elsewhere specified

Appendix C: Principal Component Analysis

a-) Renamed Response and Predictor Variables

Table 5.11 Renamed Response Variables

Response Variables	Renamed Response Variables
GDP growth (annual %)	R1
GDP per capita growth (annual %)	R2
GNI growth (annual %)	R3
GNI per capita growth (annual %)	R4
GDP (current BillionUS$)	R5
GDP, PPP (current international BillionUS$)	R6
GNI (current BillionUS$)	R7
GNI, Atlas method (current BillionUS$)	R8
GNI, PPP (current international BillionUS$)	R9
GDP per capita (current ThousandUS$)	R10
GDP per capita, PPP (current international ThousandUS$)	R11
GNI per capita, Atlas method (current ThousandUS$)	R12
GNI per capita, PPP (current international ThousandUS$)	R13

Table 5.12 Renamed Predictor Variables

Predictor Variables	Renamed Predictor Variables
Population	OP1
Surface area (sq. km)	OP2
Life expectancy at birth, total (years)	OP3

Table 5.12 (continued)

Fertility rate, total (births per woman)	**OP4**
Adolescent fertility rate (births per 1,000 women ages 15-19)	**OP5**
Mortality rate, under-5 (per 1,000 live births)	**OP6**
Immunization, measles (% of children ages 12-23 months)	**OP7**
Forest area (sq. km)	**OP8**
Energy use (kg of oil equivalent per capita)	**OP9**
Electric power consumption (kWh per capita)	**OP10**
Inflation, GDP deflator (annual %)	**OP11**
Agriculture, forestry, and fishing, value added (% of GDP)	**OP12**
Industry (including construction), value added (% of GDP)	**OP13**
Exports of goods and services (% of GDP)	**OP14**
Imports of goods and services (% of GDP)	**OP15**
Gross capital formation (% of GDP)	**OP16**
Mobile cellular subscriptions (per 100 people)	**OP17**
Merchandise trade (% of GDP)	**OP18**
Net barter terms of trade index (2000 = 100)	**OP19**
Inflation, consumer prices (annual %)	**OP20**
Domestic credit provided by financial sector (% of GDP)	**OP21**
Domestic credit to private sector (% of GDP)	**OP22**
Domestic credit to private sector by banks (% of GDP)	**OP23**
Industry (including construction), value added (% of GDP)	**OP24**

Table 5.12 (continued)

Industry (including construction), value added (annual % growth)	**OP25**
Intentional homicides (per 100,000 people)	**OP26**
Manufacturing, value added (annual % growth)	**OP27**
Manufacturing, value added (% of GDP)	**OP28**
Scientific and technical journal articles	**OP29**
Transport services (% of service exports, BoP)	**OP30**
Unemployment, total (% of total labor force) (modeled ILO estimate)	**OP31**
Unemployment, female (% of female labor force) (modeled ILO estimate)	**OP32**
Unemployment, male (% of male labor force) (modeled ILO estimate)	**OP33**
Employment to population ratio, 15+, total (%) (modeled ILO estimate)	**OP34**
Employment to population ratio, 15+, female (%) (modeled ILO estimate)	**OP35**
Employment to population ratio, 15+, male (%) (modeled ILO estimate)	**OP36**
Employers, total (% of total employment) (modeled ILO estimate)	**OP37**
Vulnerable employment, total (% of total employment) (modeled ILO estimate)	**OP38**
Vulnerable employment, female (% of female employment) (modeled ILO estimate)	**OP39**
Vulnerable employment, male (% of male employment) (modeled ILO estimate)	**OP40**
Wage and salaried workers, total (% of total employment) (modeled ILO estimate)	**OP41**
Wage and salaried workers, female (% of female employment) (modeled ILO estimate)	**OP42**
Wage and salaried workers, male (% of male employment) (modeled ILO estimate)	**OP43**
Real effective exchange rate index (2010 = 100)	**OP44**
Real interest rate (%)	**OP45**

Table 5.12 (continued)

Personal remittances, received (current US$)	**OP46**
Personal remittances, paid (current US$)	**OP47**
Current account balance (BoP, current US$)	**OP48**
Foreign direct investment, net inflows (BoP, current US$)	**OP49**
Foreign direct investment, net (BoP, current US$)	**OP50**
Foreign direct investment, net outflows (BoP, current US$)	**OP51**
Industry (including construction), value added (constant 2010 US$)	**OP52**
Manufacturing, value added (constant 2010 US$)	**OP53**
Foreign direct investment, net (BoP, current US$)	**OP54**
Exports of goods and services (BoP, current US$)	**OP55**
Imports of goods and services (BoP, current US$)	**OP56**

Appendix D: Classification

a-) Classification Models' Names

Table 5.13 Decision Tree Models

Groups of Countries by k-means -R1 response variable		
dtm	accuracy	kappa
dtm1	0.7611	0.6409
dtm2	0.7006	0.5497
dtm3	0.7803	0.6707
dtm4	0.7548	0.6318
Groups of Countries by k-medoids -R2 response variable		
dtm	accuracy	kappa
dtm5	0.7618	0.646
dtm6	0.7535	0.6353
dtm7	0.7729	0.6628
dtm8	0.7424	0.6154
Groups of Countries by fuzzy -R3 response variable		
dtm	accuracy	kappa
dtm9	0.6355	0.5611
dtm10	0.6647	0.5313
dtm11	0.6795	0.5439
dtm12	0.6647	0.5254
Groups of Countries by hierarchical -R4 response variable		
dtm	accuracy	kappa
dtm13	0.7051	0.565
dtm14	0.6635	0.503
dtm15	0.6955	0.5487
dtm16	0.6635	0.4932

Table 5.14 Random Forest Models

Groups of Countries by k-means -R1 response variable		
rfm	accuracy	kappa
rfm1	0.7978	0.6979
rfm2	0.7865	0.6851
Groups of Countries by k-medoids -R2 response variable		
rfm	accuracy	kappa
rfm3	0.7781	0.6727
rfm4	0.7697	0.6601
Groups of Countries by fuzzy -R3 response variable		
rfm	accuracy	kappa
rfm5	0.75	0.6538
rfm6	0.75	0.6545
Groups of Countries by hierarchical -R4 response variable		
rfm	accuracy	kappa
rfm7	0.7753	0.6657
rfm8	0.7753	0.6651

Table 5.15 Naive Bayesian Classification Models

Groups of Countries by k-means -R1 response variable		
nbcm	accuracy	kappa
nbcm1	0.6517	0.4793
nbcm2	0.7051	0.5606
Groups of Countries by k-medoids -R2 response variable		
nbcm	accuracy	kappa
nbcm3	0.6657	0.502
nbcm4	0.7022	0.5604
Groups of Countries by fuzzy -R3 response variable		
nbcm	accuracy	kappa
nbcm5	0.632	0.497
nbcm6	0.6657	0.541
Groups of Countries by hierarchical -R4 response variable		
nbcm	accuracy	kappa
nbcm7	0.6404	0.4571
nbcm8	0.6854	0.5302

Table 5.16 Support Vector Machine Models

Groups of Countries by k-means -R1 response variable		
svmm	accuracy	kappa
svmm1	0.7809	0.6728
svmm2	0.8118	0.7186
Groups of Countries by k-medoids -R2 response variable		
svmm	accuracy	kappa
svmm3	0.7584	0.6449
svmm4	0.823	0.7389
Groups of Countries by fuzzy -R3 response variable		
svmm	accuracy	kappa
svmm5	0.7219	0.6107
svmm6	0.7725	0.6849
Groups of Countries by hierarchical -R4 response variable		
svmm	accuracy	kappa
svmm7	0.7331	0.6037
svmm8	0.7837	0.6754

Appendix E: Regression

a-) Model Evaluation Metrics for each response variable (OR1, OR2, OR3 and OR4)

Table 5.17 Model Evaluation Metrics in case OR1 is used

Model Evaluation Metrics	OR1 response variable							
	Marginal Models		Transtion Models		Random Effects Models based on subject variation		Random Effects Models based on time variation	
	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)
Root Mean Square Error (RMSE)	1.640	1.630	0.154	0.142	1.64	1.65	1.55	1.57
Mean Square Error (MSE)	2.680	2.660	0.0236	0.0201	2.68	2.72	2.4	2.47
Quasi-Likelihood Criteria (QIC)	809	950	1081	1062	NAN	NAN	NAN	NAN
Sum of Absolute Error (SAE)	247	245	18.7	17.9	247	249	233	234
Bias (bias)	-0.984	-1.030	0.0253	0.0141	-0.984	-0.98	-1.24	-1.24
Sum of Squared Errors (SSE)	472	469	3.9	3.32	472	479	423	434

Table 5.17 (continued)

Relative Squared Error (RSE)	**2.170**	**2.150**	**0.0201**	**0.0171**	**2.17**	**2.2**	**1.94**	**2**
Mean Absolute Error (MAE)	**1.400**	**1.390**	**0.113**	**0.109**	**1.4**	**1.41**	**1.32**	**1.33**
Mean Absolute Percent Error (MAPE)	**0.423**	**0.421**	**0.0296**	**0.0284**	**0.423**	**0.425**	**0.415**	**0.418**
Sum of Squared Log Error (SSLE)	**18.100**	**18.000**	**0.201**	**0.154**	**18.1**	**18.8**	**17**	**17.3**
Mean Squared Log Error (MSLE)	**0.103**	**0.102**	**0.00122**	**0.00093**	**0.103**	**0.107**	**0.0967**	**0.0984**
Percent Bias (percent_bias)	**-0.341**	**-0.353**	**0.00616**	**0.00256**	**-0.341**	**-0.338**	**-0.401**	**-0.404**
Relative Absolute Error (RAE)	**1.560**	**1.540**	**0.128**	**0.123**	**1.56**	**1.57**	**1.47**	**1.48**
Root Mean Squared Log Error (RMSLE)	**0.321**	**0.319**	**0.0349**	**0.0305**	**0.321**	**0.327**	**0.311**	**0.314**
Root Relative Squared Error (RRSE)	**1.470**	**1.470**	**0.142**	**0.131**	**1.47**	**1.48**	**1.39**	**1.41**

Table 5.18 Model Evaluation Metrics in case OR2 is used

Model Evaluation Metrics	OR2 response variable							
	Marginal Models		Transtion Models		Random Effects Models based on subject variation		Random Effects Models based on time variation	
	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)
Root Mean Square Error (RMSE)	1.330	1.320	1.16	1.16	1.33	1.34	1.28	1.28
Mean Square Error (MSE)	1.780	1.730	1.34	1.34	1.78	1.78	1.63	1.64
Quasi-Likelihood Criteria (QIC)	830	830	1096	1096	NAN	NAN	NAN	NAN
Sum of Absolute Error (SAE)	213	210	179	179	213	214	202	202
Bias (bias)	-0.568	-0.588	1.09	1.09	-0.568	-0.567	-0.715	-0.725
Sum of Squared Errors (SSE)	313	305	221	220	313	314	288	289
Relative Squared Error (RSE)	1.480	1.450	1.14	1.14	1.48	1.49	1.36	1.37
Mean Absolute Error (MAE)	1.210	1.200	1.09	1.09	1.21	1.21	1.15	1.15

Table 5.18 (continued)

Mean Absolute Percent Error (MAPE)	**0.260**	**0.257**	**0.218**	**0.218**	**0.26**	**0.26**	**0.252**	**0.252**
Sum of Squared Log Error (SSLE)	**8.370**	**8.160**	**8.29**	**8.22**	**8.37**	**8.38**	**7.79**	**7.81**
Mean Squared Log Error (MSLE)	**0.048**	**0.046**	**0.0503**	**0.0498**	**0.0476**	**0.0476**	**0.0443**	**0.0444**
Percent Bias (percent_bias)	**-0.162**	**-0.165**	**0.218**	**0.218**	**-0.162**	**-0.162**	**-0.187**	**-0.189**
Relative Absolute Error (RAE)	**1.290**	**1.270**	**1.16**	**1.16**	**1.29**	**1.29**	**1.22**	**1.22**
Root Mean Squared Log Error (RMSLE)	**0.218**	**0.215**	**0.224**	**0.223**	**0.218**	**0.218**	**0.21**	**0.211**
Root Relative Squared Error (RRSE)	**1.220**	**1.200**	**1.07**	**1.07**	**1.22**	**1.22**	**1.17**	**1.17**

Table 5.19 Model Evaluation Metrics in case OR3 is used

Model Evaluation Metrics	OR3 response variable							
	Marginal Models		Transtion Models		Random Effects Models based on subject variation		Random Effects Models based on time variation	
	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)
Root Mean Square Error (RMSE)	1.630	1.630	0.147	0.149	1.63	1.64	1.55	1.58
Mean Square Error (MSE)	2.670	2.650	0.0217	0.0222	2.67	2.7	2.41	2.49
Quasi-Likelihood Criteria (QIC)	812	812	1081	1081	NAN	NAN	NAN	NAN
Sum of Absolute Error (SAE)	247	246	18.3	18.4	247	248	234	235
Bias (bias)	-0.988	-1.030	-0.0001	-0.0009	-0.988	-0.987	-1.24	-1.25
Sum of Squared Errors (SSE)	470	467	3.59	3.66	470	475	424	438
Relative Squared Error (RSE)	2.120	2.110	0.0182	0.0186	2.12	2.15	1.92	1.98
Mean Absolute Error (MAE)	1.400	1.400	0.111	0.112	1.4	1.41	1.33	1.33

Table 5.19 (continued)

Mean Absolute Percent Error (MAPE)	**0.427**	**0.426**	**0.03**	**0.0299**	**0.427**	**0.429**	**0.422**	**0.426**
Sum of Squared Log Error (SSLE)	**18.400**	**18.100**	**0.185**	**0.187**	**18.4**	**18.8**	**17.4**	**17.8**
Mean Squared Log Error (MSLE)	**0.105**	**0.103**	**0.00112**	**0.00113**	**0.105**	**0.107**	**0.0987**	**0.101**
Percent Bias (percent_bias)	**-0.344**	**-0.358**	**-0.0011**	**-0.0011**	**-0.344**	**-0.344**	**-0.407**	**-0.411**
Relative Absolute Error (RAE)	**1.550**	**1.540**	**0.124**	**0.125**	**1.55**	**1.56**	**1.46**	**1.47**
Root Mean Squared Log Error (RMSLE)	**0.323**	**0.321**	**0.0334**	**0.0336**	**0.323**	**0.327**	**0.314**	**0.318**
Root Relative Squared Error (RRSE)	**1.460**	**1.450**	**0.135**	**0.136**	**1.46**	**1.47**	**1.38**	**1.41**

Table 5.20 Model Evaluation Metrics in case OR4 is used

Model Evaluation Metrics	OR4 response variable							
	Marginal Models		Transtion Models		Random Effects Models based on subject variation		Random Effects Models based on time variation	
	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)	Model-1 (FM)	Model-2 (RM)
Root Mean Square Error (RMSE)	1.330	1.330	1.14	1.13	1.33	1.34	1.28	1.29
Mean Square Error (MSE)	1.780	1.760	1.29	1.29	1.78	1.8	1.64	1.66
Quasi-Likelihood Criteria (QIC)	835	869	1094	1074	NAN	NAN	NAN	NAN
Sum of Absolute Error (SAE)	215	215	175	175	215	214	203	203
Bias (bias)	-0.574	-0.595	1.06	1.06	-0.574	-0.54	-0.721	-0.719
Sum of Squared Errors (SSE)	313	310	213	212	313	316	289	292
Relative Squared Error (RSE)	1.440	1.430	1.07	1.07	1.44	1.46	1.33	1.35
Mean Absolute Error (MAE)	1.220	1.220	1.06	1.06	1.22	1.21	1.15	1.16
Mean Absolute Percent Error (MAPE)	0.264	0.264	0.214	0.213	0.264	0.263	0.255	0.256

Table 5.20 (continued)

Sum of Squared Log Error (SSLE)	**8.430**	**8.350**	**7.95**	**7.89**	**8.43**	**9.64**	**7.94**	**7.99**
Mean Squared Log Error (MSLE)	**0.048**	**0.047**	**0.0482**	**0.0478**	**0.0479**	**0.0548**	**0.0451**	**0.0454**
Percent Bias (percent_bias)	**-0.165**	**-0.169**	**0.214**	**0.213**	**-0.165**	**-0.155**	**-0.191**	**-0.191**
Relative Absolute Error (RAE)	**1.280**	**1.280**	**1.12**	**1.11**	**1.28**	**1.27**	**1.2**	**1.21**
Root Mean Squared Log Error (RMSLE)	**0.219**	**0.218**	**0.22**	**0.219**	**0.219**	**0.234**	**0.212**	**0.213**
Root Relative Squared Error (RRSE)	**1.200**	**1.200**	**1.03**	**1.03**	**1.2**	**1.21**	**1.16**	**1.16**

END OF STUDY