ARISTOTLE
UNIVERSITY
OF THESSALONIKI

MSc LOGISTICS AND SUPPLY CHAIN MANAGEMENT
SCHOOL OF ECONOMICS

THE ART OF CUSTOMER SEGMENTATION USING MACHINE LEARNING CLUSTERING METHOD

Masters Dissertation

Author: Christos Galanis

Supervisor: Dr. Athanasios Tsadiras

Thessaloniki, Greece, March/2023

Christos Galanis

BSc Economics Science & Business Administration

Masters Dissertation

A thesis submitted in partial fulfillment of the requirements of the degree of Logistics & Supply Chain Management

Supervisor Professor
Dr. Athanasios Tsadiras

Approved by the three-member examination committee on dd/mm/yyyy

Name/Surname 1	Name/Surname 2	Name/Surname 3
....	....	....

Christos Galanis

....

Abstract

Purpose of the study: The research was conducted to discover the current status of machine learning clustering algorithms used mainly for customer segmentation. Furthermore, an application of a traditional and well-known clustering model i.e. K-means on a dataset took place.

Methodology: In order the objective of the study to be achieved a thoroughly Systematic Literature Review (SLR) approach was followed. This method provides a systematic way of identifying articles based on criteria set by the author and eventually it is completed by determining the remaining articles. These articles are used in the research. Furthermore, a dataset selected from Kaggle was analyzed. This analysis was done using the python programming language in the jupyter notebook environment.

Findings: Using the SLR method the past and present literature remained were presented. Moreover, using the k-means algorithm in the dataset 5 groups were created. The appropriate number to be taken as the optimal was supported by the elbow method. A variety of figures were used to draw insights after the implementation of the algorithm for example the distribution of the segments and the groups behavior based on specific features.

Acknowledgements

First of all, I would like to thank my supervisor Dr. Athanasios Tsadiras. Thank you for the excellent academic advice, support and for the guidance throughout this project. Also, a big thank you to my master's program professors since they taught me about the Logistics & Supply Chain Management field and for their advices as well. Most importantly, I am grateful for my family's unconditional love and support throughout my academic journey.

Contents

Abstract iv

Acknowledgements v

Contents vi

List of Figures viii

List of Tables viii

Chapter 1: Introduction 9

1.1 Basic Notion 9

1.2 Background Literature 12

1.3 Purpose of the study 16

1.4 Methodology 16

1.5 Outline of the Thesis 18

Chapter 2: Systematic Literature Review Methodology 20

2.1 Introduction 20

2.2 Justification of Systematic Literature Review 21

2.3 Steps of Systematic Literature Review 22

2.3.1 Formulate Research Scope 22

2.3.2 Search Strategy Implementation 24

2.3.3 Data Extraction and Analysis 26

2.4 Descriptive analysis and reporting 27

2.4.1 Overview 27

2.4.2 Year of publication 27

2.4.3 Geographical Presentation of Articles 28

Chapter 3: Comprehension of the articles and Discussion 29

3.1 Introduction 29

3.2 Unsupervised machine learning clustering method description 29

3.3 Categories of clustering methods 30

3.3.1 Partitional clustering 30

3.3.2 Hierarchical clustering 31

3.4 Similarity - Proximity measure description 32

3.5 Introduction of K-means 33

3.6 Metaheuristic algorithms 34

3.7 Silhouette and Elbow methods 35

3.7.1 Silhouette method 35
3.7.2 Elbow method 35
3.7.2.1 Finding the K value when the line is smooth 36
3.8 Other approaches for finding the optimal K value 37
3.9 Multiple clustering algorithms 38
3.9.1 Introduction of K-means++ 38
3.9.2 Introduction of K -means|| 38
3.9.3 Automatic c-K-means algorithm 38
3.9.4 K-medoids application 39
3.9.5 A modification of spectral clustering algorithm 40
3.9.6 K-prototypes machine learning approach 40
3.9.7 Gaussian model 40
Chapter 4: Analysis of a dataset 42
4.1 Customer Segmentation 42
4.2 Processing the dataset 43
4.2.1 Data Cleaning 45
4.2.2 Feature Engineering 47
4.2.3 Managing the categorical variables 50
4.2.4 Correlation among the attributes 58
4.3 Standardization 59
4.4 Dimensionality Reduction 60
4.5 Clustering 62
4.6 Visualization and evaluation of the model 64
Chapter 5: Conclusions and future applications 69
5.1 Recap 69
5.2 Fulfillment of Research Objectives 69
5.3 Research Questions answered 70
5.4 Limitations based on Methodology 70
References 72

List of Figures

Figure 1.1.: Steps of the thesis 18

Figure 2.2: Chapter 2 break down 20

Figure 2.3: Broader terms... 23

Figure 2.4: Year of Selected Articles Distribution... 28

Figure 2.5: Countries of published articles 28

Figure 3.6: A curve with a visible elbow point. b A curve with an ambiguous elbow point (Shi, et al., 2021) 36

Figure 4.7: Barplots and Boxplots for the corresponding features.................................. 53

Figure 4.8: Heatmap correlation of the attributes... 59

Figure 4.9: A visualization on the reduced space... 61

Figure 4.10: Illustration of elbow method.. 63

Figure 4.11: Capturing the data on the reduced space 65

Figure 4.12: Distribution of the clusters on the whole data 66

Figure 4.13: Illustration of the clusters regarding their income and spendings 67

Figure 4.14: Distribution of the number of accepted promotions per cluster 68

List of Tables

Table 2.1: Articles remained by searching in each Database.. 24

Table 2.2: Criteria for inclusion 24

Table 2.3: Literature Review Evaluation based on criteria 25

Chapter 1: Introduction

1.1 Basic Notion

It is undeniable today that technology is rising. More and more people are using smart devices which makes their life easier for example they can communicate with their friends across distances instantly with nearly a push of a button and almost with very low cost. Another instance is that one can search for a preferable place to visit and by using GPS or some other app can go to their destination with very accurate instructions. However, smart devices and technology is not only used for personal purposes. Nowadays, it is essential for a company to use technology in order to run its operations with the purpose of optimization. The rise of industrial 4.0 technologies, digital era, 5G technologies and other cutting-edge systems have as a result the enormous increase of data (Yudhistyra, et al., 2020). In addition, there are technologies such as Information and Communication Technologies (ICT), RFID, Internet of Things (IoT), Enterprise Resource Planning (ERP) and more (Arunachalam, et al., 2018) which are playing a vital role in increasing the volume of data in tremendously rhythms.

As Arunachalam, et al., (2018) mentioned the notion of using sophisticated technologies across organizations and specifically in supply chains have brought to the surface the term of the new era called big data. The data is everywhere from the individual using its device to private enterprises, organizations, governments and public sector. In general, it is everywhere in the world in the 21st century since everything has to do with technology. There are a lot of applications where big data can be leveraged and be useful for human beings for instance predicting the outcome of a basketball game by analyzing enormous data. Here it is important to make a reference to the term "logistics & supply chain management" which is highly well-known. Big data is highly used in this sector as mentioned by Yudhistyra, et al., (2020). Everything in the world is connected to each other making the world a global supply chain. Arunachalam, et al., (2018) referred that managers can take operation and strategic decisions either mid-term or long-term based on data driven results from big data analytics and not relying mostly on their intuitions, beliefs, experiences and feelings.

The previous notion was made in order to understand the idea that supply chain has been made a global phenomenon across the whole planet. We mainly use supply chain term in order to describe the departments and operations that are connected each another inside a company or outside. The latter is known when an organization has connections that exceeds its environment making them to have a global supply chain network. From the previous description, one can easily understand that supply chains are everywhere in the world. With the technology eruption nowadays large datasets. Data accuracy and insights are characteristics that flows across supply chains in supplier networks (highlighted by Columbus, 2015, as mentioned by Yudhistyra, et al., 2020).

In a world of data, companies try to find and unveil value of the data. However, difficulties are always there when they try to extract the hidden value. As Yesudas et al., (2014) referred with the help and support of advanced analytics companies try to capture, store and analyze large volume of data across their supply chains (Arunachalam, et al., 2018). With the evolution of ICT technologies information flows across supply chain networks and organizations can have the opportunity and ability to monitor, capture and analyze the data so as to make decisions (referred by Chae and Olson (2013) cited by Arunachalam, et al., (2018)). It is fascinating that today's 90% of data is being produced approximately in the last two years as stated by Fawcett and Waller (2014) cited by Arunachalam, et al., (2018). Big data analytics tools and methods can help companies to improve. But in order to use these tools organizations need to comprehend the idea of the advanced usage of analytics methods with the purpose of extracting values of their decisions (quoted by Tien (2015) referred by Arunachalam, et al., (2018)).

Logistics and Supply Chain Management

It is important a reference to be made to the term "Logistics & Supply Chain Management". Over the years supply chain and logistics have been used a lot. However, there are no solid definitions about these two terms since they have a background which is relevant with transportations, information services, economics and more. The first time that the term "Supply Chain Management" came to the surface was in a paper almost 40 years ago as discussed by Oliver * Weber (1982) cited by Swanson, et al., (2017). As Yudhistyra, et al., (2020) cited Dupuit (1952) and Pienaar (2009) highlighted that logistics management is a mit of military sciences and methods that include maintenance, procurements, facilities, personnel, transport and more. On the other hand, supply chain is the broader view. The terms are specified bellow according to Rushton et al., (2017) as mentioned by Yudhistyra, et al., (2020):

Logistics is mainly material management and distribution of these materials. Also, supply of raw materials, components, goods and more are included with the aim to deliver the products to the final client.

The view of Supply Chain is wider. It consists of suppliers, logistics and customers. Specifically, in order to be made clear logistics are part of a supply chain and they support goods and information to flow across a supply chain meaning from the suppliers to the end customers. Instances of the previous definition can include warehouses, inventory and transportation management, information systems and many more. The purpose of a supply chain is to create value for their customers and in order this to be achieved the flow of goods and services should be flow with the most optimized manner (mentioned by Aamer (2018); Aamer and Sawhney (2004); Chopra and Meindl (2013); Sahara et al., (2019); Yani et al., (2019) cited by Aamer, et al., (2021)).

Big Data

What is big data? As pointed out by Arunachalam, et al., (2018) the Oxford English Dictionary (OED) (2016) described that it is "facts and statistics collected together for reference or analysis" or "the quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signal and recorded on magnetic, optical, or mechanical recording media.". So, the first definition refers to the capture of numbers and facts for analysis with the use of mathematical and statistical methods and tools without the support of a computer (Arunachalam, et al., 2018). The second term, points out the importance of the computer with the aim to perform analysis and to store and transfer data.

In 1997 NASA scientists were doing research by analyzing large amount of data and they faced challenges in the computer systems (Aryal, et al., 2020). Again, the Oxford English Dictionary described "Big Data" as enormous large data sets which if analyzed by computers they can reveal trends, patterns and associations (Arunachalam, et al., 2018).

Big Data Analytics can be used for a wide applications inside a supply chain area including purchasing, production, transportation, sales, customers and many more (Seyedan & Mafakheri, n.d.).

Machine Learning

What is Machine Learning? It is the building and learning algorithms which can learn from pattern recognition and computational learning theory and can make predictions on a dataset. It is a field of computer science. It is necessary to be referred that it is better to use these processes by creation of a model aiming to make decisions based on data-driven insights rather than choices that weights on static program instructions (Simon, et al., 2015). Tom Mitchell from Carnegie Mellon University stated that "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E." as cited by Simon, et al., (2015, p. 1).

Machine learning can be defined into two types of tasks (Simon, et al., 2015):

Supervised machine learning: This means that the program has learned and subsequently “trained” on a pre-defined set of data set specifically it is called “training examples”. After that, having capture and analyze the patterns and trends of the specific dataset then can leverage this ability aiming to make predictions and conclusions when given a new data.

Unsupervised machine learning: The program learns and unveils patterns and relationships from a dataset without any pre-defined set of “training examples”.

Clustering algorithms

First of all, we need to make a reference that clustering is an unsupervised machine learning algorithm. This means that groups data into clusters according to similarities and patterns of data points characteristics and features found (defined by Jain, 2010; Abualigah, (2019) as cited Ezugwu, et al., (2022)). Over the years, a lot of clustering algorithms have been created and implemented in order to solve clustering problems (Zhou et al., 2019; Abualigah et al., (2018) as mentioned by Ezugwu, et al., (2022)). As Tan (2018) have highlighted clustering techniques can be splitted into two categories hierarchical and partitional (Ezugwu, et al., 2022). However, when dealing with real world problems having real datasets there is no prior number of naturally occurring groups or clusters in the data objects (Liu et al., (2011) highlighted by Ezugwu, et al., (2022)). Hence, there is brought to the surface a term of automatic data clustering algorithms which Ezugwu, et al., (2022) have referred to their paper. This is no more no less than clustering methods try to automatically determine the number of clusters in a data set without any prior information of the dataset features and attributes (Ezugwu, (2020a) specified by Ezugwu, et al., (2022).

Customer segmentation

Separating a large group of customers into smaller groups is called customer segmentation. This group of people are devised based on a variety of characteristics based on their social, behavioral and consumption elements and attributes. This is significantly important since capturing and identifying a variety of consumption behaviors and preferences of consumers can result in creating and offering personalized products and services for each and every customer groups (Li, et al., 2021).

1.2 Background Literature

As far as now, basic terms like logistics and supply chain management, big data analytics, machine learning, clustering techniques and customer segmentation have been discussed and explained so as to provide the reader a view of the main ideas that will be used on this thesis. First of all, this research was written in order to bring to the surface clustering algorithms that can be used for solving supply chain management problems and specifically for customer segmentation. In order to achieve the previous, it is vital to explore what the past and present literature refers so as to adjust the research objects and questions. This can be done by implementing a Systematic Literature Review method (SLR). The SLR approach is thoroughly discussed in Chapter 2. After implementing the SLR methodology the main idea of the articles remained will be described in this stage since it is vital for the reader to understand what the literature is about when implementing these screening criteria by the SLR approach.

First of all, Giri & Chen, (2022) highlighted the importance of demand forecasting in the fashion and apparel retail sector using deep learning. As they pointed out, in comparison to other industries fashion apparel retail companies have a variety of challenges and problems when it comes to predicting future demand that need to address. This is mainly because fashion products have a short life cycle, lack of historical data, have periodic seasonal trends and patterns and high uncertainty on the market. Giri & Chen, (2022) proposed a paper where they used European fashion retailer data and they created a forecasting system that combined image product features attributes of clothes with its sales data aiming to predict future demand values. Subsequently, the model has been tested in test items and its results are promising. By clustering based on product's image similarity and its sales profile then clusters of products were created. Finally, the proposed model was able to find its best match in these clusters of products and predict the weekly sales of new fashion apparel (Giri & Chen, 2022).

Another article found by the SLR is the journal paper written by Lee & Mangalaraj, (2022) where it described the big data analytics in supply chain management sector. As they pointed out with the evolution in machine learning and computing infrastructure big data analytics are more than ever significant in the supply chain. This paper made a systematic literature review of studies existed in big data analytics and aimed to identify any research gap in the literature.

One more scholarly journal article made by Hu, et al. (2021) tried to unveil patterns or to seek spatially linked customer behaviors in insurance. Spatial analysis can be ranged from univariate descriptive statistics to complex multivariate analysis. Data taken from an insurance company in Ireland was used in order to see if the publicly spatially linked demographic census data are helpful in modeling customers' behavior for example stopping payment of customer that have premium contracts. The company have a significant benefit from this since it can capture and prevent such incidents. The model used spatial clustering with census data aiming to unveil spatial behavior of Irish customers' life. The contribution of this paper was that it captured the spatial characteristics of Irish consumers from the reliable data taken by the government in the life insurance sector (Hu, et al., 2021).

There are traditional clustering algorithms for example K-means and more that depend on knowing the prior number of clusters. On the other hand, as it was mentioned previously on this stage, in real life problems with real data sets one does not know the number of clusters. Here come the automatic clustering algorithms where sophisticated automatic clustering techniques can determine the optimal number of clusters in data objects. This notion is described and presented in the paper of E. Ezugwu1, et al. (2021) where they made a systematic taxonomical overview and bibliometric analysis of progress and trends in automatic clustering algorithms from the early attempts in the 1990s until 2021.

When it comes to defining the exact number of clusters there are methods commonly used such as the Elbow method (Shi, et al., 2021). With this algorithm the discriminant of the cluster numbers is depending on the manual identification of the elbow points on the curve that is visualized. However, when the plotted curve is fairly smooth then it is hard for the analysts to identify clearly the elbow point and subsequently the optimal number of clusters. Thus, Shi, et al., (2021) wrote a paper where they proposed a discriminant method to yield a statistical metric so as to estimate the optimal number of clusters. As Shi, et al., (2021) highlighted the results found from this research showed that their proposed method is better than the Silhouette method which is widely used.

In the paper of Raj & Vidyaathulasiraman, (2021) it is presented the application of K-Medoid principal for clustering e-Learners. They used both the Elbow and Silhouette methods for determining the value of K clusters. Finally, they proved that Silhouette method is greater since it best predicts the value of K groups.

A very vital aspect of every company is to calculate and monitor their customer churn metrics in order to increase their profits. Subsequently, Sharma, et al., (2021) developed spectral clustering. In clustering the similarity measure plays a significant role for projecting churn. The linear Euclidean distance was replaced by the non-linear S-distance (Sd) and the Sd is concluded from the context of S-divergence. Furthermore, in their work they have applied three (3) existing algorithms meaning k-means, density-based spatial clustering with noise and the SC in 15 databases. The results showed that the proposed algorithm was better in terms of its Jaccard index, f-score, recall, precision and accuracy.
After that, they tried to test the significance of the clustering results by the Wilcoxon's signed-rank test, Wilcoxon's rank-sum test, and sign tests. The outcomes were positive for the proposed algorithm most in the case of clusters of arbitrary shape (Sharma, et al., 2021).

Furthermore, in the scholarly journal article named “Know Your Clients’ Behaviors: A Cluster Analysis of Financial Transactions” a modified behavioral finance recency, frequency, monetary model is applied for quantifying investor behaviors. More specifically, in Canada, financial advisors and dealers are required to collect and maintain information regarding their client also known as “know your client” (KYC). Examples of this could be age of risk tolerance, for investor accounts. A dataset of 50.000 accounts and 23.000 clients Thompson, et al., (2021) applied unsupervised machine learning clustering algorithms with purpose of finding clues of investors that have similar behavior. The results were interesting since they found that information such as gender, residence region, and marital status does imply client behavior rather than eight variables for trade and transaction frequency and volume which provides more information. In order to predict and understand better investor behaviors financial advisors and investors could be supported to utilize more advanced metrics by the outcome of this paper.

One article remained by the SLR approach was written by Cikovic, (2020) where she highlighted the importance of customer profiling for the small and medium-sized enterprises (SMEs). As a result, SME companies will better comprehend their audiences and markets and they can overcome problems and maximize their profits by better marketing strategy. Her conference paper focused on identifying the profile types for a Croatian-based trade and dealer organization. The purpose of the cluster profiling is to get a clear understanding and a picture of who the customer in each cluster belongs. The dataset consisted of customers demographics, psychographics, socioeconomics, product and style preferences and also marketing channel preferences. The research used the Gaussian mixture model to cluster clients. The results obtained by the model application deduced 3 clusters of customers (Cikovic, 2020).

As it was mentioned previously, the K-means algorithm is named one of the ten classic algorithms when it comes to data mining (Zhang, et al., 2018). However, using the earlier described model its initial selection of centers is vulnerable to outliers. The study of Zhang, et al., (2018) proposed the covering K-means algorithm (c-K-means). It calculates both accurate clustering outcomes and is self-adaptively. As a result, it offers a number of clusters based on data features. The paper consists of two phases where in the first executes the covering algorithm which self-organizes and identifies the number of clusters K. The one big advantage of the previous method is that the number of clusters and the intimal centers required not to be prespecified. The second phase started by applying the Lloyd iteration based on the results found on the first phase. As Zhang, et al., (2018) pointed out the significantly advantage that the c-K-means algorithm has since it can solve the problem of large-scale data clustering effectively. They also referred that the accuracy and efficiency that this algorithm shows when experimented on real data surpass the models existed under both sequential and parallel conditions.

1.3 Purpose of the study

As it was mentioned before, technology is rising with tremendously rhythms. Furthermore, the data produced by these systems and technologies is increasing with high volume. The hefty majority of data has been produced in the last two years as described previously. Managers and scholars have raised their interest to this sector. However, regarding past and present literature found by the SLR approach there are limits when it comes to clustering algorithms for solving supply chain management problems and specifically for customer segmentation. Therefore, this research has investigated the current status of machine learning clustering algorithms used mainly for customer segmentation.

The aim of this research was to apply a Systematic Literature Review approach with the purpose of understanding the past and present work done. Subsequently, by acquiring a comprehensive notion of the background literature found by the SLR methodology this thesis is one of the earliest to identify and present an understandable analysis of the machine learning clustering algorithms. Within this framework the research also applied a clustering method in a dataset with the purpose of addressing with accuracy a customer segmentation problem. The results found can be leveraged by individuals, scholars, managers and more. The insights of this research can raise the interest for future academic research.

There were two objectives presented in this thesis and they are illustrated below:

Objective 1: Implementation of SLR approach on Machine Learning Clustering Techniques in a Supply Chain Management problem which is the well-known Customer Segmentation

Objective 2: Application of a clustering algorithm on a real dataset

So, two questions were formulated as shown below:

Research Question 1: What are the insights gained from using a clustering algorithm to segment customers on the dataset?
Research Question 2: Who can use the findings of this research?

1.4 Methodology

This thesis used a variety of published articles relevant to the objectives stated previously with the purpose of obtaining the highest number of articles which satisfied the requirements. This has as a result to provide validated and understandable literature for the reader. The SLR approach was utilized in order to offer a systematic methodology of obtaining literature and in addition a machine learning clustering algorithm was used in a dataset aiming to provide to the reader a thoroughly and comprehensible view of this research.

Literature review is an essential element of academic research. In general, SLR methodology is used to search, locate, assess and synthesize the appropriate articles found from databases aiming to achieve the objectives and the research questions stated. This systematic way of finding literature provides credibility and allows the replication of the literature review (Rother, (2007), Boland, et al., (2017)). In order to foster the literature knowledge further one needs to know the present work done. By searching, analyzing the prior work done in relevant literature one can understand the breadth and depth of the current work and recognize if any gaps there are, unveil inconsistencies, weaknesses and evaluate the current literature. (Xiao & Watson, 2017). Furthermore, answering objectives and questions set up by the researcher is one more advantage of this method.

Literature reviews should be reliable and valid aiming to present the methodology behind the remaining articles used for research (Xiao & Watson, 2017). In this thesis, the researcher utilized the SLR approach so as to obtain up-to-date articles and illustrate the relevant machine learning clustering applications used for supply chain management problems and specifically for customer segmentation. From a total of 217 articles that were initially found by searching after screening by specific factors 11 of them remained for thoroughly analysis and presentation of their context. The previously mentioned procedure is comprehensively described in Chapter 2. Furthermore, a descriptive analysis was conducted aiming to provide to the reader an understanding of the articles in terms of the year of publication and geographic distribution of the articles retained. Continuously, the thesis presented the context of the remaining articles. This aims to explore the current literature and illustrate it to the reader in the most comprehensible way. Exploring and presenting the main findings of the current literature found by the SLR method is vital since it will bring into the surface any scientific literature gap. Furthermore, a machine learning clustering algorithm is used on a dataset in order the reader to get a point of view of how customer segmentation can be done. This dataset is from a well-known machine learning website "Kaggle". The specific dataset was selected since it contains data and values that are easy to be found by a company that may aspire to implement this algorithm to their customers.

The SLR method is thoroughly described in Chapter 2 while the clustering algorithm is being applied on the dataset in Chapter 4 providing all the necessary information to data collection, statistics and finally implementation with the python programming language in jupyter notebook. Below is illustrated an overview of the steps followed for the development of this Thesis.

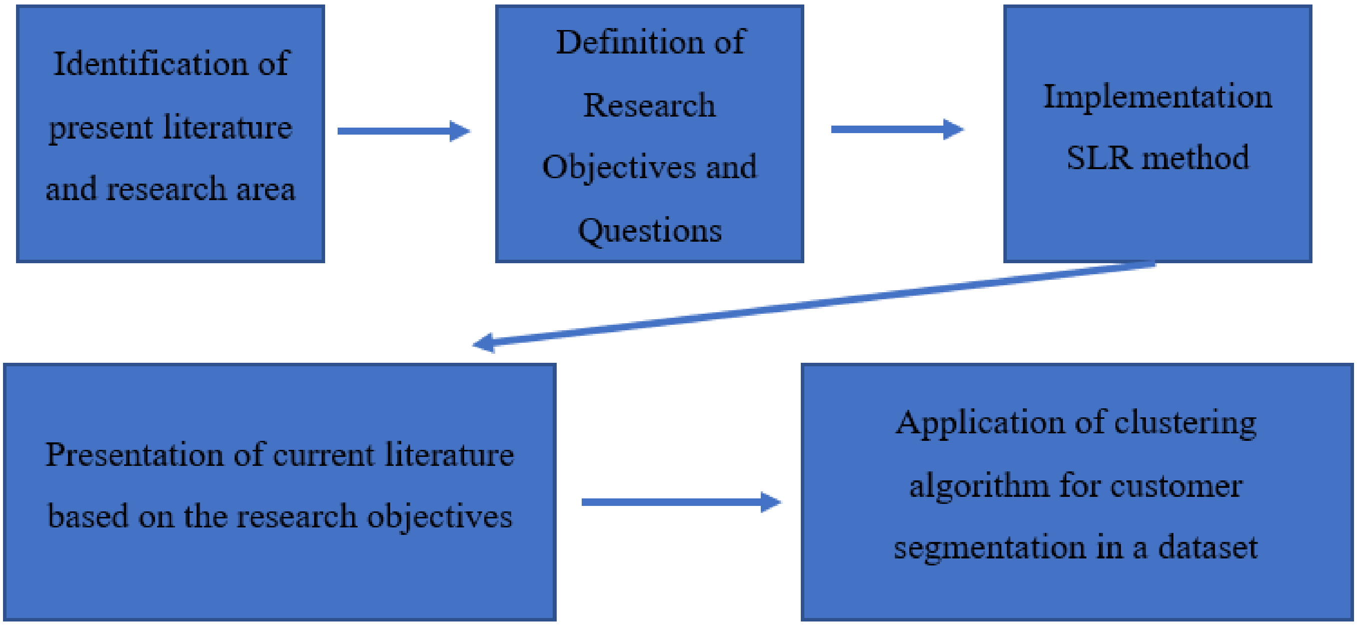
Identification of present literature and research area
Definition of Research Objectives and Questions
Implementation SLR method
Presentation of current literature based on the research objectives
Application of clustering algorithm for customer segmentation in a dataset

Figure 1.1.: Steps of the thesis

1.5 Outline of the Thesis

- The 1st Chapter makes an introduction to the background literature, the current research area, the purpose of this study, defines the research objectives and questions, the methodology followed and highlights the significancy of this thesis.

- Continuously, in Chapter 2 the Systematic Literature Review approach was being followed and presented thoroughly. This ensures the credibility and validity of the articles used in this research. Then follows a descriptive analysis of the literature remained aiming to illustrate a quantitative overview of the articles by providing charts and graphs. This makes it easier to the reader to understand the distribution of age and country publication of each of the articles used in this research.

- Chapter 3 presents the relevant information derived from the selected articles. In other words, a thoroughly discussion of the findings relevant to the objectives and questions was conducted.

- The Chapter 4 was being written in order a clustering algorithm to be applied on a dataset with the support of the python programming language.

- Finally, Chapter 5 concludes the main findings derived from the previous chapters. Moreover, it makes a reference on who can this research aims to help.

Chapter 2: Systematic Literature Review Methodology

A key part of academic research is literature review. In Chapter 2 a thorough Systematic Literature Review (SLR) is conducted in the field of Machine Learning for solving clustering Supply Chain Management (SCM) problem in order to achieve the research objective. In other words, based on the guidelines of the research article “Guidance on Conducting a Systematic Literature Review by Xiao & Watson (2017), a step-by-step approach methodology is presented with rational and justifiable information behind each action in order the reader to get a clear view of the whole SLR methodology. Specifically, by reviewing the relevant literature one can understand the present work done and identify any gaps occurred. After the completion of SLR method a descriptive analysis was done aiming to describe and present in breadth and depth the essential information regarding the main remaining articles. An overview of Chapter 2 is shown below:

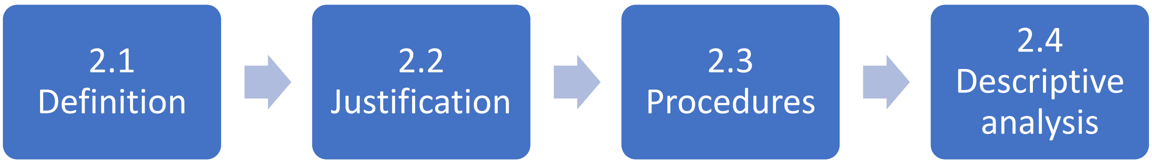

Figure 2.2: Chapter 2 break down

First of all, the definition and the justification of the SLR method is necessary to be introduced since the reader needs to know why an SLR approach is vital for a research thesis. After that, the processes are provided with the necessary details in every step so as to allow one to examine and validate the results found. Last but not least, it is important a reference to be made to the descriptive analysis as the final step.

2.1 Introduction

Most newcomer researchers think that having summaries of papers is enough for a literature review (Webster & T. Watson, 2002). Hart (1998) described the definition of literature review (cited by Levy & J. Ellis ,2006, p. 2) as “the use of ideas in the literature to justify the particular approach to the topic, the selection of methods, and demonstration that this research contributes something new". Furthermore, Levy & J. Ellis (2006) (cited from J. Shaw, 1995) pointed that in order to make a review one needs to be able to explain "how one piece of research builds another".

So, a literature review is more than a collection of papers or research manuscripts. Webster & T. Watson (2002) mentioned that literature review is a process which allows areas of improvement it increases theory development. In other words, having the information of the amount of research exists in a subject area, the researcher is able to detect what the gaps are and unveil uncovered areas for potential research. In this way, a gap that may be brought to the surface between past and present literature and will be closed by future work. So, there is one technique which is the most mainstream approach. This follows a non-systematic procedure of gathering, filtering and assessing literature. Also, this technique is biased on researchers' beliefs and experiences. Therefore, there is increased risk of prejudice in a given research topic.

On the other hand, a more systematic approach is the SLR which uses specific thoroughly methodology steps for locating, gathering, assessing and writing from the available articles. Also, this procedure helps to answer scientific questions and achieve certain objectives stated by the researcher (Jesson, et al., 2011).

2.2 Justification of Systematic Literature Review

The two main literature reviews are described below and the most suitable one is selected with rationale evidence.

As Rother (2007) referred a scientific literature review is no more no less than a search in a database in order to gain information and achieve the objective stated from the academic. She separated the concept of reviewing the articles into two categories: Systematic and narrative review of the literature. In terms of cost and time a better option is narrative view and that it is why it is the most common of these two reviews referred previously (Xiao & Watson, 2017).

When conducting a narrative review, one is not concerned about listing the databases, methodological approaches or the evaluation criteria during the research (Rother, 2007). They are focused on gathering the relevant information in order to achieve their objective (Kastner, et al., 2012). That is why there is lack of impartiality since the usage of narrative review can be subjected to the researcher's point of view, beliefs and experiences (Noordzij, et al., 2011). Rother (2007) pointed that narrative review can support readers with up-to-date information regarding a specific theme or topic. This can be leveraged by education staff. However, the lack of answering to specific quantitative research questions and most importantly the methodological approach that is not followed is making the review referred not suitable for reproduction of data.

In contrast, SLR approach tries to answer specific research questions using methodology which is thoroughly explained step-by-step. The articles found by this review technique are considered original work since they have been found by methodological procedures (Rother, 2007). There is one main disadvantage while conducting an SLR approach. That is, the articles remaining will be subjected to the researchers' actions. In other words, missing relevant literature can happen since it is inevitable to search in every database or one cannot clearly understand the assessment criteria of the articles found.

The most suitable approach of the two literature reviews referred previously is SLR which was used in order to conduct this thesis. Boland, et al., (2017) refers to SLR as a process which can ensure impartiality, replicability and increased quality of literature since it decreases prejudices by selecting solid methodological steps.

2.3 Steps of Systematic Literature Review

This section presents the processes of the SLR methodology in order to make the reader have a clear understanding over the results found allowing them for evaluating the articles and plausible for future research replication and potential enrichment.
The steps of the SLR approach used in this thesis are taken by Boland, et al., (2017) and Xiao & Watson, (2017) which are more or less the same and are the following 6 processes: Planning the Review (1) – which refers to the formulation of the problem e.g. research scope, Developing Search Strategy (2) – which selection criteria such as databases, keywords are selected, Assessment quality (3) – which permits the final assessment of the articles remaining through reviewing abstracts and full – text, Data Extraction (4) – relevant data is extracted aiming to gather the needed information for Analysis and Synthesis (5) and finally the Writing up of the Analysis (6).

2.3.1 Formulate Research Scope

First of all, it is vital to make a reference to the definition of the research scope in order the reader to get familiar with the beginning of the process. A narrative literature review was being conducted so as the researcher to get a clear view of the subject and research theme.

In order to address the research question and to achieve the objectives formulated in chapter 1 there was a need for tow broader terms to be referred. The one was machine learning clustering techniques and the other was customer segmentation. So, the research scope followed specific research keywords and phrases in order to capture the broader terms.

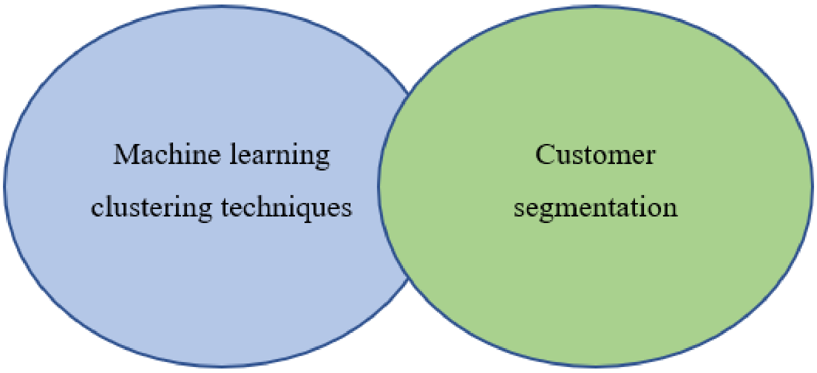

Figure 2.3: Broader terms

Selection of Keywords and declaration of Search string

It is important to select keywords that are going to capture the information needed regarding the objective and research questions. There are two important keywords that need to be discussed and these are the following: “machine learning clustering techniques” which captures the area of clustering methods and “customer segmentation” which again is vital for addressing both the objectives and the research questions. The researcher used the Boolean operators (“OR”, “AND”) to search for articles in databases in the most efficient way. The final strings are shown below:

String 1: "machine learning clustering" OR "machine learning cluster analysis"
String 2: "customer segmentation" OR "customer" OR "market segmentation"

After the initialization of string 1 and string 2 a search was completed in two databases by combining the previous strings in one as shown below:

Search 1 + 2: ("machine learning clustering" OR "machine learning cluster analysis") AND ("customer segmentation" OR "customer" OR "market segmentation")

2.3.2 Search Strategy Implementation

This is the point where the magic happens since by applying this procedure one can obtain the main articles. Searched in three databases and used research strings relevant to the objectives resulted in the final inclusion of the articles found. The next step is to "filter" the articles found by relevant selection criteria of the author's preference and experience.

Data Sources Selection

It is important a reference to be made to the data sources which the researcher used. The first is ProQuest which provides access to dissertations, theses, eBooks, archives and more. Approximately 125 billion digital pages are estimated to be provided by ProQuest and the search is being done through 8 databases. The other data source used in the research is ScienceDirect. It is a website which offers access to more than 18 million pieces of content from more than 4.000 academic journals and 30.000 e-books of the Dutch publisher Elsevier. And finally, Scopus which is the is the largest citation and abstract database of peer-reviewed literature: books, scientific journals and conference proceedings. From the search, a total of 227 articles were remained and then were filtered by the selection criteria and the assessment quality later on. Below the three databases with their results are presented.

Table 2.1: Articles remained by searching in each Database

DATABASE	Remained articles
ProQuest	156
ScienceDirect (Elsevier)	62
Scopus	9

Criteria for articles selection

At this point, criteria for selecting relevant articles are described. From the search procedure a total of 218 articles were found and continuously throughout the process of filtering based on criteria we have the following table.

Table 2.2: Criteria for inclusion

No	Criteria	Description	Reason
1	Full -Text / Open access	161	In order to screen them thoroughly
2	Peer reviewed	70	Articles be validated

3	Last 5 years & in English	60	Up-to-date articles and worldwide English written

From the inclusion criteria a total of 60 articles taken from 3 Databases were found. After that, an assessment of the quality of the remaining articles was conducted.

Evaluation of remained articles

In order to ensure that the articles retained are satisfying the criteria formulated an assessment is usually conducted. This increases credibility and validity of the research since from the 60 articles they will remain only these that are more relevant to the thesis research. Therefore, a three-step process was followed by evaluating each of the articles by its title, abstract and full-text aiming to compare its context to this research.

Screening by tittles

Firstly, as mentioned previously, reading the tittles and evaluating each article took place in this stage. The retained articles are linked with the thesis objectives and they can more or less answer the research questions stated in Chapter 1. Hence, from the 60 articles found only 19 are relevant to our research goals

Screening by abstract

At this point, reading the abstract of the remining 19 articles was followed. Finally, 12 articles were retained.

Screening Full-Text

Having remained 12 articles after screening them by title and abstract it was time to read their full content aiming to evaluate if they are linked to the research objectives. In the end, 11 articles found to be relevant to the research.

It is important a reference to be made to the assessment criteria. In other words, in order the reader to gain a comprehensively understanding of the literature assessment procedure the following table described the reasons behind each exclusion article.

Table 2.3: Literature Review Evaluation based on criteria

Processes	Description
Screening by title	Firstly, a total of 60 articles found before going to the next step which was to read their title. After that, 19 articles were found to be relevant to the research objectives whereas the others had

	a focus on different research areas such as 3PLs, infrastructure and more. Examples, of articles excluding from the research are the following: “Machine learning methods to improve the operations of 3PL logistics” (Tufano, et al., 2020) and “Clustering algorithm-based network planning for advanced metering infrastructure in smart grid” (Gallardo, et al., 2021).
Screening by Abstract	After the first step, the articles remained were screened by their abstracts and 7 of them were excluded. In other words, their research focus was irrelevant to this thesis objectives such as industrial context, twitter real time trends and more. Some of the examples are the following: “A survey of clustering algorithms for an industrial context” (Benabdellah, et al., 2019) and “Real-Time Twitter Trend Analysis Using Big Data Analytics and Machine Learning Techniques” (Rodrigues, et al., 2021).
Screening by Full-Text	After remaining with 12 articles, it was time to read their full context so as to gain a clear view of objectives, methodology and results. An article mainly focused on employee point of interest was excluded which was the following: “PREDICTIVE ANALYTICS IN EMPLOYEE CHURN: A SYSTEMATIC LITERATURE REVIEW” (Ekawati & University, 2019). **Hence, a total of 11 articles were remained for the thesis.**

2.3.3 Data Extraction and Analysis

According to Boland, et al., (2017) there are 4 key steps involved in this stage which are the following:

- Understanding the data in order to be extracted
- Creation of data tables
- Completion of data tables
- Reporting and presenting the extracted data

Based on the above steps a thoroughly descriptive analysis was conducted and is presented in the next pages.

Interpretation of findings and writing up procedure

One very important stage of the SLR is the summarization and description of the articles found. In other words, an elaboration of each of the remaining articles took place in Chapter 3 with the goal the reader to understand the current literature and the research done so far based on the objectives and questions set in Chapter 1.

2.4 Descriptive analysis and reporting

2.4.1 Overview

A descriptive analysis of the articles remained took place in order to gain some information about them. Specifically, the articles are classified based on their year of publication and geographic region. The goal of this descriptive analysis is to present to the reader solid information about the articles remained since these articles will be explained in Chapter 3 in more detail. This has as a result to increase the validity of the articles.

2.4.2 Year of publication

In total, 11 articles have remained through the SLR approach. All of them, were published within the last five (5) years based on the criteria selected in Chapter 2. This results in obtaining findings that are recently written. As it is observed from the figure below the hefty proportion of articles were published in 2021 (approx. 55%).

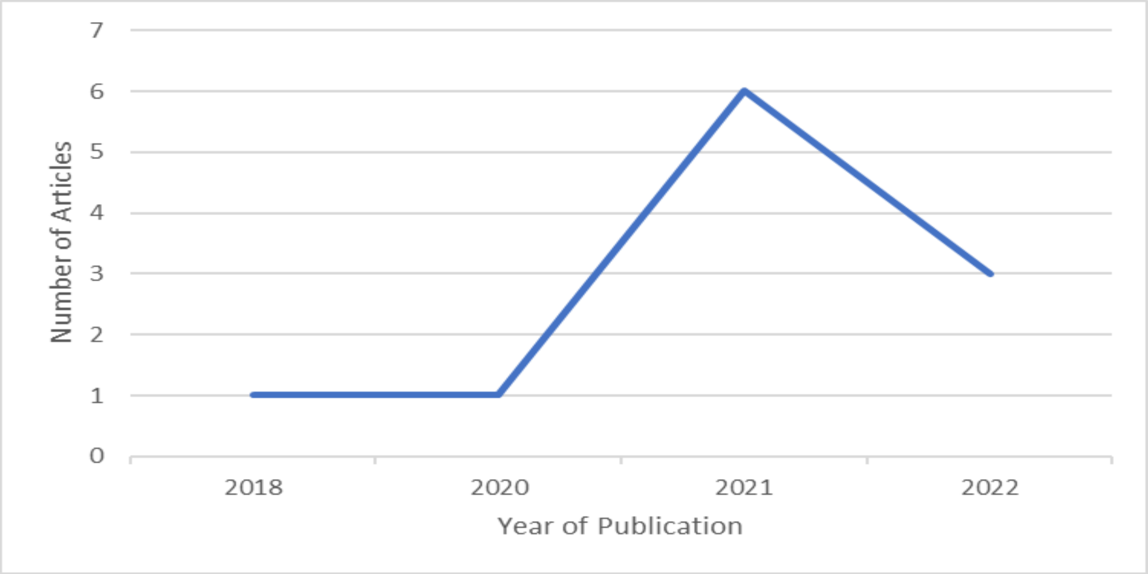

Figure 2.4: Year of Selected Articles Distribution

2.4.3 Geographical Presentation of Articles

The vast majority of articles was from Switzerland in approximately 36%. The geographical focus of 27% of the articles was from United Kingdom. Below, on Figure it is illustrated the distribution of the remaining articles of this thesis.

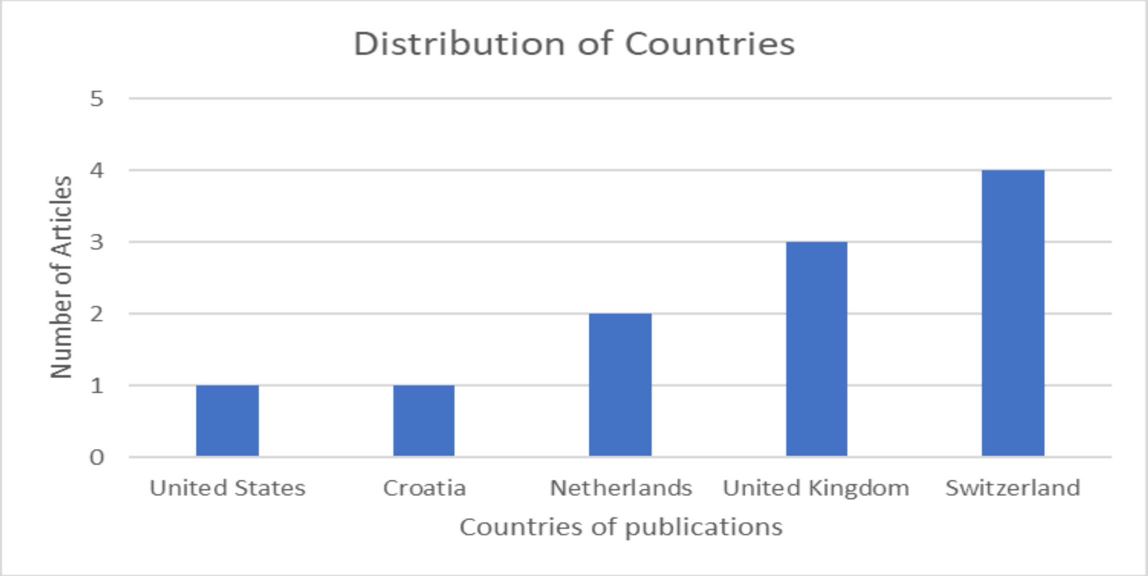

Figure 2.5: Countries of published articles

Chapter 3: Comprehension of the articles and Discussion

3.1 Introduction

In this stage, the researcher presented a thoroughly analysis of the clustering algorithms found in the articles remained. In other words, an introduction to clustering and its categories was made. Clustering methods were described in order to give the reader an idea of the literature found by the SLR approach. It is important to be mentioned that the presentation of the findings of each article will be mainly inside the context of the research objectives stated in Chapter 1.

It is undoubtedly the tremendously increase of data due to the development of big data systems and technologies e.g., internet of things, mobile applications, e-commerce and more. Nowadays, organizations can handle more input data than ever before since they have machines where they can handle this data. As Zhang, et al., (2018) stated that researchers face challenges regarding big data and in order to mitigate them they want to use appropriate tools. Aiming to extract the most important information from big data is required to utilize machine learning approaches according to Zhang, et al., (2018). One of the most widely used machine learning algorithm is clustering which is applied in many sectors like statistics, bioinformatics and more. Researchers have found a variety of clustering algorithms with each own of them with their own style and optimization methods.

3.2 Unsupervised machine learning clustering method description

Machine learning clustering belongs to unsupervised machine learning methods targeting at finding the optimal groups of unlabeled objects based on their similarity of calculated intrinsic features (Landers & Duperrouzel, (2019) as cited by E. Ezugwu1, et al., (2021)). Also, Giri & Chen (2022) referred to the term "clustering" in their research and its goal is by using distance metric to group items that have similar characteristics. In other words, when there is a dataset with a set of data vectors $X = \{x1, x2, \ldots, x_n\}$, then the algorithm tries to group them based on their similarity to the same cluster.

The most important aspects of this algorithm are the specification of the optimal number of clusters and secondly the assignment of all data groups to clusters correctly. Below is presented a formula such as the assignment of N objects to K clusters to be implemented:

$$S(N,K) = \frac{1}{K!}\sum_{I=0}^{K}(-1)^{K-i}\binom{K}{i}i^N$$

The S (N, K) is perceived as the Stirling number of the second kind as it was recognized by E. Ezugwu1, et al., (2021). Many times, the number of clusters is unknown and below is illustrated the search space size for detecting the correct number of groups:

$$B(N) = \sum_{K=1}^{N} S(N, K)$$

In agreement with E. Ezugwu1, et al., (2021) the B(N) is called Bell number. Furthermore, Falkenauer, (1998) recognized that the clustering application of determining the optimal number of groups is NP-hard (nondeterministic polynomial time) when K>3. Cowgill, et al., (1999) stated when there is the case of moderate-sized problems, the clustering process could require computationally challenging (E. Ezugwu1, et al., 2021).

It is important to be mentioned that the data which is used for input are unlabeled. The output resulted from the clusters can be utilized for another machine learning objective. Russell, et al., (2010) highlighted as quoted by Giri & Chen (2022) that there is a variety of clustering methods for instance K-means, hierarchical, and probabilistic clustering.

3.3 Categories of clustering methods

Clustering algorithms can be partitional clustering, hierarchy clustering, density-based clustering, model-based clustering, graph theory-based clustering, grid-based clustering (Madhulatha, (2012) as reproduced by Zhang, et al., (2018)).

In this stage, some clustering algorithms are being described thoroughly: partitional and hierarchical clustering (Ramadas & Abraham, (2019) as mentioned by E. Ezugwu1, et al., (2021)).

3.3.1 Partitional clustering

Partitional clustering methods: “decompose the datasets into a set of disjoint clusters based on specific optimization criteria” (E. Ezugwu1, et al., 2021, p. 3) for example prototype-based algorithms, graph-based (hierarchical agglomerative clustering) methods, density-based algorithms and hybrid algorithms. In addition to the above approaches, we can encounter some of the traditional clustering methods such as K-means, fuzzy c-means and simulated annealing (E. Ezugwu1, et al., 2021). Partitional clustering can be exploited in two methods: fuzzy and hard. Hard clustering refers to patterns which are binary. Hence, each pattern is included to only one cluster.

Furthermore, the above cluster method splits a dataset into predetermined number of groups (KR, (2008) cited by E. Ezugwu1, et al., (2021). Continuously, as it is referred by E. Ezugwu1, et al., (2021) a dataset X is partitioned into J non-overlapping clusters $B = \{b_1, b_1, \dots, b_K\}$. In addition, the three conditions bellow should be all satisfied:

$$B \neq \emptyset, i = 1{,}2, \dots, K$$

$$U_{i=1}^{K} b_i = X$$

$$B_i \cap B_i = \emptyset, i, j = 1, \dots, K \; and \; i \neq j$$

The most well-known algorithm of hard clustering is k-means method which requires a prespecified number of clusters E. Ezugwu1, et al., (2021). Its goal is to minimize the sum-of-squared-error criterion (Hartigan & Wong, (1979) and Jaccard, (1901) pointed out by E. Ezugwu1, et al., (2021)). Jain, (2010) presented those 50 years after the creation of k-means and it is still used because it is simple and it requires low computational complexity (E. Ezugwu1, et al., 2021). However, since it requires a predefined number of clusters it faces challenges when applied in real-world clustering problems. Therefore, upgrades of this approach have been done so as to detect clusters automatically. Some of these new features are mentioned in E. Ezugwu1, et al., (2021) research such as X-means (Pelleg & Moore, 2000) and G-means algorithms (Hamerly & Elkan, 2003).

On contrast, fuzzy clustering assumes that there is a non-binary relationship between patterns and clusters. Therefore, there are different degrees for the patterns that are assigned to each cluster (E. Ezugwu1, et al., 2021). That is, the dataset is specified in “fuzzy sets” where each pattern may be included to more than one group simultaneously with a specific degree of membership $u_j \in [0; 1]$. The two conditions below should be satisfied from the membership value of the ith pattern within the jth group (E. Ezugwu1, et al., 2021).:

$$\sum_{j=1}^{K} u_j(X_i) = 1, i = 1, \dots, N,$$

$$\sum_{i=1}^{N} u_j(X_i) < N, j = 1, \dots, K.$$

According to Bezdek, (1981) as reproduced by E. Ezugwu1, et al., (2021) there is an extension of fuzzy k-means algorithm which is called fuzzy c-means method.

3.3.2 Hierarchical clustering

Hierarchical clustering methods: these are iterative-based clustering processes that create results identical to a hierarchical tree or dendrogram which illustrates a sequence of nested grouping of the objects included in a dataset (Chang, et al., (2010) and Changa & Yeung, (2007) as highlighted by E. Ezugwu1, et al., (2021)). In other words, as E. Ezugwu1, et al., (2021) referred the process starts by the creation of N successive clustering levels and the next clustering is based on the solution found at the previous level. This means that there is no need for priori knowledge regarding the number of clusters. It is pointed out by E. Ezugwu1, et al., (2021) that the clusters results are static since the objects in each cluster cannot be moved to another. In addition, two characteristics of them are the arbitrary decision making and time complexity. Jain, (2010) presented the two categories of hierarchical clustering Agglomerative and divisive methods. From them, the most popular are the single-link and complete-link algorithms respectively (E. Ezugwu1, et al., 2021).

3.4 Similarity - Proximity measure description

One of the most important characteristics when applying clustering algorithms is the similarity measure. Selecting the correct proximity measure is vital since memberships are specified for every object in a given dataset X. Ezugwu1, et al., (2021) referred that the proximity measure could be "either a distance (dissimilarity) or a similarity between a pair of objects, between an object and a prototype, or between a pair of prototypes". The output clusters formed rely on the selected proximity measure (Maulik & Saha, (2010) as mentioned by Ezugwu1, et al., (2021)).

Two widely used proximity measures are presented below:
The Minkowski metric [130], or Lp-norm, is a dissimilarity measure specified as:

$$d_p(x, y) = (\sum_{i=1}^{D} |x_i - y_i|^p)^{1/p}$$

In the case of x and y there are D-dimensional data vectors. A reference needs to be made that when p = 2, the Minskowski is converted to the Euclidean distance or L_2-norm indicated as d_e (x, y). Furthermore, when p = 1, then the Minkowski metric becomes the well-known Manhattan distance or L_1-norm, and lastly when p → ∞ then there is the Chebyshev distance or L_∞-norm. The last case is calculated as follows:

d_∞ (x, y) = $max_{1 \le i \le D} |x_i - y_i|$

A second option of measuring the similarity between two vectors is the cosine of the angle between them and it is calculated as below:

$$\cos(x, y) = \frac{x^T y}{\|x\| \|y\|}$$

In the case of $\|.\|$ it is perceived as L_2-norm. Das, et al., (2009) pointed out that the relationship between the cosine similarity and Euclidean distance can convert the cosine similarity into a dissimilarity measure and it can be computed as below:

$$d_{\cos(x,y)} = 1 - \cos(x, y) = \frac{1}{2} d_e^2(x, y)$$

Bottou & Bengio, (1995) and Sharma & Seal, (2021) as cited by Sharma, et al., (2021) presented some clustering algorithms using linear distance for example Pearson correlation, Euclidean, Manhattan, Kendall correlation, Eisen cosine correlation, Bit-Vector, Spearman correlation, Hamming, the Jaccard Index and the Dice Index. Despite the fact that, in the above approaches there is a lot of literature done, when it comes to non-linearity in clustering the research done is not the same Sharma, et al., (2021). The latter targets to find more accurate boundary between two groups.

3.5 Introduction of K-means

The most widely used machine learning method is k-means where the easiness in its implementation and comprehensive are advantages and the challenging selection of the k value and initial centers are its drawbacks (Bottou & Bengio, (1994) as presented by Zhang, et al., (2018)). The latter have a significantly impact on the final results. Proposals for improving initial K-means centers has been made and these improvements helped the performance of the Lloyd iterations regarding convergence and quality (Ostrovsky, et al., (2012), Arthur & Vassilvitskii, (2007) and Bahmani, et al., (2012) as mentioned by Zhang, et al., (2018)).

However, K-means requires to predefine the number of clusters and this is a challenge when working with real world problem data. Zhang, et al., (2018) presented in their paper an algorithm called C- K - means clustering algorithm where it does not require to set a number of clusters but it automatically calculates them based on the data characteristics and features. Also, it is independent of the initial centers. Some of the advantages that this method has according to Zhang, et al., (2018) is its accuracy, faster and efficiency in high dimensional data. Furthermore, it is easy to implement and have good scalability.

In the paper of Raj & Vidyaathulasiraman, (2021) are described cases where they used clustering methods for selecting optimal number of clusters. Some of the cases are presented below:

- Syakur, et al., (2017) used the K-Means clustering algorithm with the support of Elbow method in their paper.
- The same strategy was integrated by Humaira & Rasyidah, (2020) in their conference paper.

- Hmednaa, et al., (2019) found the favorite learning preferences of the learners utilized the K-Means clustering algorithm.

Joshi & Jain, (2018) applied the K-Means and K-Medoids clustering methods with the aim to build a model for projecting the students' academic performance.

K-means approach is presented below in compliance to Zhang, et al., (2018):
First of all comes the randomly selection of k data points from the dataset. Then the distance is calculated between each data point in the dataset and each center in the initial centers. In this stage, an assignment of data points is given to clusters and the result is that every data point knowing which center is closest. Then, repeatedly follows the process of updating the center of the clusters and each data point be assigned to the nearest center of the cluster. This creates a set of cluster centers and it is continuously generated until the new set of cluster centers is less or equal to the set of former cluster centers. This is known as local search and is called Lloyd iteration.

3.6 Metaheuristic algorithms

There are some disadvantages when using simple clustering algorithms and this is mainly appeared when the number of features, instances and dimensionality of the data objects rises. In this case, a hefty percentage of the approaches are trapped into a local optima (Abualigah, et al., (2016) and Abualigah, et al., (2016) as presented by E. Ezugwu1, et al., (2021)). Their effectiveness is relied heavily on the first result (E. Ezugwu1, et al., 2021).

In order to mitigate these drawbacks, there have been developed some nature-inspired metaheuristic algorithms as mentioned by E. Ezugwu1, et al., (2021). For example, particle swarm optimization (PSO) (Eberhart & Kennedy, 1995), the firefly algorithm (FA) (Agbaje, et al., 2019), differential evolution (DE) (Storn & Price , 1997), genetic algorithm (GA) (Goldberg & Holland , 1988), artificial bee colony (ABC) (KARABOGA , 2005) and many more. It is rational to think that their names are inspired from the contexts of the nature occurring phenomena that is why they are perceived as nature-inspired metaheuristic algorithms. Furthermore, according to Kuo, et al., (2014) as highlighted by E. Ezugwu1, et al., (2021) there approaches that try to minimize the similarity within each cluster and simultaneously find the max of the dissimilarity between them. That is, they are referred as optimization problems and they fell into the new era of clustering paradigm.

3.7 Silhouette and Elbow methods

In the field of Data analytics, when clustering it is a challenge to predict the cluster number.
When using traditional K-means or K-medoids clustering algorithm it is vital to determine the correct value of K. Raj & Vidyaathulasiraman, (2021) presented in their paper two classical ways of predicting the cluster number: Elbow or Silhouette method. In addition, these two methods were presented Masud, et al., (2018) as referred by Shi, et al., (2021).

First of all, lets distinguish the two approaches described previously.

3.7.1 Silhouette method

The research of Giri & Chen (2022) used the silhouette score (Aranganayagi & Thangavel, 2007) in order to select the optimum k. Furthermore, Shi, et al., (2021, p. 3) mentioned the Silhouette method and it mainly utilizes the average distance between one data point and others in the same cluster and the average distance among different clusters to score the clustering result" (Rodriguez, et al., (2019), Rousseeuw, (1987)). Specifically, "A is the mean intra-cluster distance and B represents the mean nearest-cluster distance" Shi, et al., (2021, p. 3).

In other words, this approach computes the cohesion which is the similarity of an object to its cluster and then the separation which is the comparison to their clusters. According to Arbelaitz, et al., (2013) as cited by Shi, et al., (2021) the Silhouette method is better to use since it can evaluate the best number of clusters in most distinct cases. The values range from -1 to 1 and if a value is close to -1 this means that the item is allocated to the wrong cluster whereas if the value is close to 1 indicates the sample is better clustered.
The Equation (1) can be seen below:

$$Silhouette\ Coefficient = \frac{b-a}{\max\,(a,b)}\ (1)$$

3.7.2 Elbow method

One traditional method of selecting the number of clusters is through the Elbow method (Ketchen & Shook, (1996) as highlighted by Shi, et al., (2021)). Its notion starts by setting K equals to 2 as the initial best K value and then increasing K by step 1 until it reaches the maximum predetermined estimated potential correct cluster number. Then it finds clearly the optimal cluster number K that is consistent with the plateau. Shi, et al., (2021) described the Elbow method where one can obtain the optimal K value when there is an elbow point. This means that before taking the value K, the cost reduces to the cost peak value and then after passing the point of K, it increases to the cost peak value without any change.

Another explanation of the Elbow method was made by Raj & Vidyaathulasiraman, (2021). The value of clusters can range from 1 to n. First step is to determine the Within-Cluster Sum of Square (WCSS) for number of K. In other words, the goal is to compute for every value of K the number of squared distances between each point and the centroid (Raj & Vidyaathulasiraman, 2021). When WCSS is plotted and K equals to 1 then WCSS value is the highest. As the clusters increase the WCSS value decreases. If it is not a smooth line then the correct value of K corresponding to the number of clusters will be the point which forms an elbow point (Raj & Vidyaathulasiraman, 2021).

However, the selection of the correct K is depending on the humans. Furthermore, when it is the curve of the plot fairly smooth then it is difficult even for the experts to recognize the Elbow point.

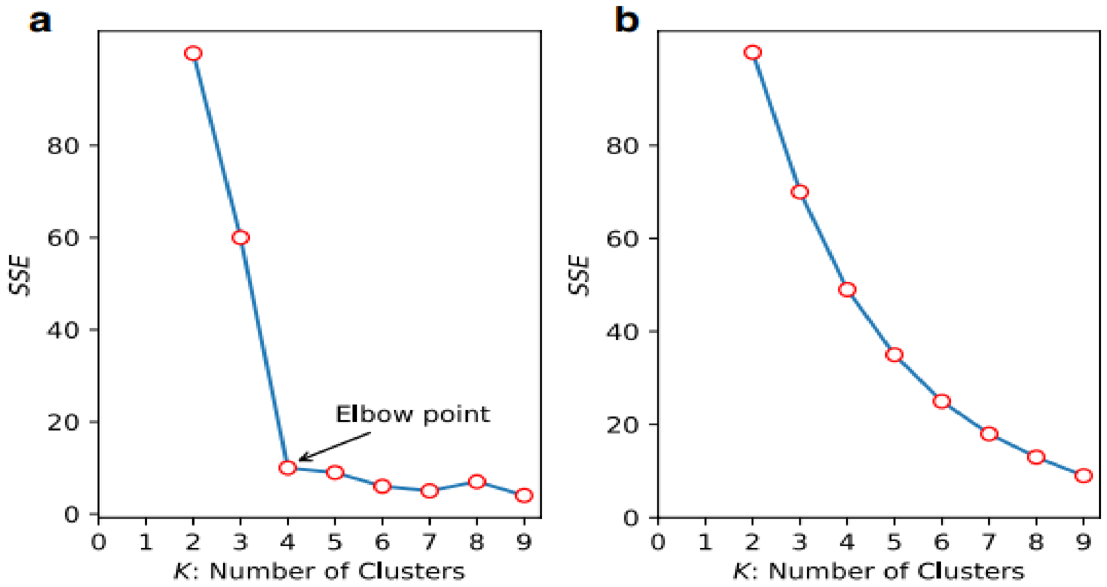

Figure 3.6: A curve with a visible elbow point. b A curve with an ambiguous elbow point (Shi, et al., 2021)

3.7.2.1 *Finding the K value when the line is smooth*

As it was mentioned above, sometimes when the curve is smooth even empirical analysts cannot identify absolute the number of K. In some instances, there is enough prior information regarding the dataset so as to obtain the appropriate number of clusters. On the other hand, in general there is not sufficient data for selecting the optimal K that it can be used for clustering algorithms (Shi, et al., 2021). In the latter case, Shi, et al., (2021, p. 2) pointed out that the optimal cluster number can be found “if you specify a potential estimated range of values for an unknown number of clusters”.

Shi, et al., (2021) proposed a new approach so as to calculate a metric to recognize the Elbow point when this is not clearly visible but smooth. First of all, the average degree of distortion taken by the Elbow approach is normalized to the range of 0 to 10. Then, this output is exploited so as to determine the cosine of intersection angles between elbow points. Continuously, the result from the calculation of the intersection cosine angles and the arccosine theorem are utilized aiming to evaluate the intersection angles between elbow points. Its result is an index which is perceived as the optimal cluster number (Shi, et al., 2021).

3.8 Other approaches for finding the optimal K value

At this point, it was useful to refer a variety of methods that can be exploited with the aim to find the optimal number of clusters. Tibshirani, et al., (2002) and Yuan & Yang, (2019) as cited by Shi, et al., (2021) highlighted the steps for using the gap static method as follows: it gets the result from K-means, then it makes a comparison between the output found previously with the change in intra-cluster dispersion and finally it gains the correct number of clusters. Based on the K-means or expectation maximum (EM) the v-fold cross-validation method can be used for calculate the correct number of clusters (BURMAN, (1989) and Yu, et al., (2018) as specified by Shi, et al., (2021)). Another well-known method is the Hierarchical agglomerative clustering which runs the K-means N times and from its output a dendrogram can be used for selecting the optimal cluster number (Liu, et al., (2017) and Dash, et al., (2003) as reproduced by Shi, et al., (2021)).

In addition, according to Posada & Buckley, (2004) as mentioned by Shi, et al., (2021) there are some methods that can calculate the optimal cluster number based on information criteria. That is, the Bayesian information criterion (BIC) or Akaike information criterion (AIC) where they are utilized in the X-means machine learning clustering algorithm with the purpose of finding the appropriate number of clusters in the analyzed dataset (Ding, et al., (2017) as recognized by Shi, et al., (2021)). Sugar & James, (2011) as pointed out by Shi, et al., (2021) made clear that the rate distortion theory can also be utilized for obtaining the optimal cluster number. Furthermore, Shi, et al., (2021) have referred to the article of Xu & Wunsch, (2005) where they highlighted the cross-validation method which gives a number of clusters providing cluster stability. This produces similar clusters from the original dataset. Nainggolan, et al., (2019) mentioned that this method is “stable for input randomization as pointed out by Shi, et al., (2021).

3.9 Multiple clustering algorithms

3.9.1 Introduction of K-means++

In K-means algorithm the solution found is local and the selection of K clusters can significantly impact the results of the clustering method. Since, we want to find the optimal global solution then it needs to repeatedly select initial centers and obtain the final values by changing these initial centers. In order to overcome this disadvantage of K-means researchers as highlighted by Zhang, et al., (2018) have found the so-called K -means++.

This algorithm was initially proposed by Arthur & Vassilvitskii, (2007) as cited by Zhang, et al., (2018). The goal of this method is to determine the initial centers one by one in a way that the generation of the current cluster center is based on all of the previously acquired cluster centers. It is important to make a reference to the initialization idea behind the two algorithms referred meaning K-means and K -means++. The first method tries to select decentralized initial centers where the second selects clustering centers while gives prioritization on the data points away from the centers selected before. Although, K -means++ have advantages the initialization process has inherent sequential execution properties. This is a disadvantage of this algorithm as it makes it not scalable. In other words, the k centers value must cross the dataset k times and the generation of the current cluster center is based on the previous found centers. This creates limitations when working with large-scale datasets.

3.9.2 Introduction of K -means||

Aiming to address the above disadvantages of the referred algorithms the researchers proposed a new initialization method called K -means|| (Bahmani, et al., (2012) as cited by Zhang, et al., (2018)). The goal of this approach relies on the sampling strategy during each traverse and suggests an overlapping factor. When sample points are crossed in a nonuniform manner each time and the sample procedure is repeatedly for X iterations, then the X is the clustering cost of the clusters selected. The acquisition of the centers of N sample points can be done by repeated sampling. As Zhang, et al., (2018, p. 4) pointed out the “The number of intermediate centers is larger than k and much smaller than the original data size”. Continuously, the process follows the assignment of weights to center points in the set of center points. Then, the center points of these weights are clustered. The final k centers are found. The final stage requires that the k points obtained operate as initial centers for the Lloyd iteration (Bahmani, et al., (2012) as specified by Zhang, et al., (2018)).

3.9.3 Automatic c-K-means algorithm

In the real-world data clustering methods can face challenges when selecting the optimal value of clusters. Traditional clustering approaches for instance K-Medoids, K-Means, and Chameleon rely heavily on having a prior knowledge of the number of clusters. However, in real world problems these methods may face difficulties when they do have an unknown number of clusters. The paper of E. Ezugwu1, et al., (2021) described a systematic taxonomical overview and bibliometric analysis regarding the progress and the trends in metaheuristic clustering methods.

Zhang, et al., (2018) proposed the c-K-means algorithm where it can automatically predict the number of clusters based on data features. It consists of two stages: the initialization of the covering algorithm (CA) and the Lloyd iteration of the K-means. With the order that were referred initially CA starts by identifying the k value based on the similarities in the data without having predefined the number of clusters nor the manual selection of the initial centers. The so called "blind" feature is the k value where is not predefined. The results found from phase one are used in stage two where Lloyd iteration is performed. It is important to refer that this algorithm includes the advantages of both CA and K-means since it has shown good scalability on the results extracted by the experiments and it can be exploited so as to solve large-scale clustering problems. The results found according to Zhang, et al., (2018) of the c-K-means approach outperforms in terms of efficiency and accuracy the methods existing under both sequential and parallel conditions.

3.9.4 K-medoids application

Hu, et al., (2021) took data from an insurance company and made research to question if the publicly spatially linked demographic census data can be utilized with efficiency in order to model consumers' lapse behavior specifically in life insurance policies for example stopping payment of premiums. They made a variety of useful groups with clustering spatial units based on people's characteristics lived there taken by the census data. A multi-phase model was proposed using k-medoids clustering method. In order to be made clear, their census data consisted of small area and each of them had various features such as age distribution, marital status and so on. The utilized K-medoids method in their paper made groups from the census data. The optimal K needed to be found so an elbow plot was employed. From a range of 1 to 20 clusters the K took values and the algorithms run multiple times. Then, the total within-cluster sum of distance was calculated for each value of K. After that, a plot was created with the sum of distances against the number of clusters K. A question may arise in this stage and it is as follows: what is the optimal K after the plot? A bend i.e., elbow found in the plot can be exploited as the appropriate number of clusters (Hu, et al., 2021). Continuously, when the clusters have been identified it starts the analysis of each cluster so as to be profiled and described.

In addition, another application of K-medoid clustering algorithm was done by Raj & Vidyaathulasiraman, (2021) when they tried to group e-learners into clusters based on their style and learning preferences. They used both Elbow and Silhouette method. However, the latter was proved to predict better the cluster value.

3.9.5 A modification of spectral clustering algorithm

According to Sharma, et al., (2021) almost all the traditional unsupervised machine learning methods have been used in the industrial field mainly in churn projection aiming to seek and identify the behavior of customers. These approaches depend heavily on "*data/features, similarity/distance measure, objective functions, initial cluster centers and the clustering algorithm itself*" (Sharma, et al., 2021, p. 1). Sharma, et al., (2021) proposed a modified spectral clustering (SC) approach by replacing the conventional linear Euclidian distance with the non-linear S-distance which is taken from the S-divergence (SD).

3.9.6 K-prototypes machine learning approach

Thompson, et al., (2021) with the usage of recency, frequency and monetary (RFM) model quantified investor behaviors from a financial investment dealer dataset (Lumsden, et al., (2008) as cited by Thompson, et al., (2021)). They applied the k-prototypes machine learning clustering algorithm by Huang, (1997). The K-prototypes approach exploited in their paper have similarities with the K-means method. Apart from this, K-prototypes integrated processes to include data which is categorical. After that, they developed a high-dimensional visualization by predicting it into lower-dimensional space. To make this happen, they utilize the t-distributed stochastic neighbor embeddings (t-SNE) (Laurens van der Maaten & Hinton, (2008) as mentioned by Thompson, et al., (2021)). This approach gathers instances of data that have similarities. In addition to this, a term called "perplexity" was highlighted by Laurens van der Maaten & Hinton, (2008) as cited by Thompson, et al., (2021) and it is a measurement of the algorithm for the effectiveness of the nearest neighbors which in general range between 5 to 50. For selecting the optimal perplexity value there is no traditional method. However, Laurens van der Maaten & Hinton, (2008) pointed out that larger datasets need larger perplexity value as presented by Thompson, et al., (2021).

3.9.7 Gaussian model

Nowadays, comprehending customers is vital for companies. Therefore, customer profiling is a strategic move that can be exploited from enterprises. Cikovic, (2020) utilized the Gaussian mixture model (GMM) to cluster customers' data taken from a Croatian-based trade and dealer company. This approach was first introduced by Pearson (1894) as cited by Cikovic, (2020). In addition, its evolution is due to EM algorithm (Expectation Maximization) (Dempster et al. in 1978 as presented by Cikovic, (2020)). Gaussian mixture model which is used for machine learning clustering objectives has the following advantages: It is a probabilistic approach of finding clusters of instances meaning that the probability of each instance to belong to each cluster is generated. Therefore, classifications are created by assigning instances to the most likely cluster. Mixture modeling is known to be flexible.

Chapter 4: Analysis of a dataset

4.1 Customer Segmentation

In this notebook, an unsupervised machine learning algorithm is used in order to divide the dataset into smaller groups with similar characteristics. More specifically, certain attributes such as purchase history, demographics, spending habits and more can be exploited so as to find groups with similar behavior.

Clustering algorithms are a common tool for performing unsupervised customer segmentation. These algorithms work by identifying patterns in the data and grouping similar observations together. For example, a clustering algorithm could make some groups of a dataset with customers based on the products they bought, how often they make purchases, or how much they spend. Also, they might include some demographics attributes such as their country, city and many more.

Some common clustering algorithms used for customer segmentation include K-means, Hierarchical clustering and DBSCAN.

The findings and the results of this analysis can be used by individuals and organizations either in the private or public sector. In the private sector, managers can understand better their customers and be more effective in their decisions about how to market and serve different groups of customers. On the other hand, in the public sector such as universities or individuals can be helped from this analysis for further research.

About the dataset

The dataset was taken from https://www.kaggle.com/ which is a website for machine learning. It includes datasets that can be used for machine learning applications.

For this analysis, I used the https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis which contains information regarding customers from a groceries company's database. The different attributes are illustrated below:

People

- **ID:** Customer's unique identifier
- **Year_Birth:** Customer's birth year
- **Education:** Customer's education level
- **Marital_Status:** Customer's marital status
- **Income:** Customer's yearly household income
- **Kidhome:** Number of children in customer's household
- **Teenhome:** Number of teenagers in customer's household
- **Dt_Customer:** Date of customer's enrollment with the company
- **Recency:** Number of days since customer's last purchase
- **Complain:** 1 if the customer complained in the last 2 years, 0 otherwise

Products

- **MntWines:** Amount spent on wine in last 2 years
- **MntFruits:** Amount spent on fruits in last 2 years
- **MntMeatProducts:** Amount spent on meat in last 2 years

- **MntFishProducts:** Amount spent on fish in last 2 years
- **MntSweetProducts:** Amount spent on sweets in last 2 years
- **MntGoldProds:** Amount spent on gold in last 2 years

Promotion

- **NumDealsPurchases:** Number of purchases made with a discount
- **AcceptedCmp1:** 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- **AcceptedCmp2:** 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- **AcceptedCmp3:** 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- **AcceptedCmp4:** 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- **AcceptedCmp5:** 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- **Response:** 1 if customer accepted the offer in the last campaign, 0 otherwise

Place

- **NumWebPurchases:** Number of purchases made through the company's website
- **NumCatalogPurchases:** Number of purchases made using a catalogue
- **NumStorePurchases:** Number of purchases made directly in stores
- **NumWebVisitsMonth:** Number of visits to company's website in the last month

Goal

The overall target is to divide the customers into similar groups.

Acknowledgements

The dataset for this project is provided by Dr. Omar Romero-Hernandez. Furthermore, the notebooks below were used as a reference for some code, ideas, and explanations in this analysis:

- Customer Segmentation: Clustering - https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering - created by KARNIKA KAPOOR
- Customer Segmentation using Clustering - https://www.kaggle.com/code/johnybhiduri/customer-segmentation-using-clustering - created by JAINENDRA BHIDURI

4.2 Processing the dataset

```
import pandas as pd

#load the dataset
data = pd.read_csv(r'C:\Users\galan\OneDrive\Έγγραφα\Masterthesis_L&SCM\marketing_campaign (2).csv', sep='\t')

#printing the top 5 rows of the dataset
data.head()
```

```
     ID  Year_Birth   Education Marital_Status   Income  Kidhome  Teenhome  \
0  5524        1957  Graduation         Single  58138.0        0         0
```

```
1  2174        1954  Graduation        Single  46344.0        1         1
2  4141        1965  Graduation      Together  71613.0        0         0
3  6182        1984  Graduation      Together  26646.0        1         0
4  5324        1981         PhD       Married  58293.0        1         0

  Dt_Customer  Recency  MntWines  ...  NumWebVisitsMonth  AcceptedCmp3  \
0  04-09-2012       58       635  ...                  7             0   
1  08-03-2014       38        11  ...                  5             0   
2  21-08-2013       26       426  ...                  4             0   
3  10-02-2014       26        11  ...                  6             0   
4  19-01-2014       94       173  ...                  5             0   

   AcceptedCmp4  AcceptedCmp5  AcceptedCmp1  AcceptedCmp2  Complain  \
0             0             0             0             0         0   
1             0             0             0             0         0   
2             0             0             0             0         0   
3             0             0             0             0         0   
4             0             0             0             0         0   

   Z_CostContact  Z_Revenue  Response  
0              3         11         1  
1              3         11         0  
2              3         11         0  
3              3         11         0  
4              3         11         0  

[5 rows x 29 columns]
```

```
#checking if there are any missing values and the data types of each attribute

data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   ID                   2240 non-null   int64  
 1   Year_Birth           2240 non-null   int64  
 2   Education            2240 non-null   object 
 3   Marital_Status       2240 non-null   object 
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64  
 6   Teenhome             2240 non-null   int64  
 7   Dt_Customer          2240 non-null   object 
 8   Recency              2240 non-null   int64  
 9   MntWines             2240 non-null   int64  
 10  MntFruits            2240 non-null   int64  
 11  MntMeatProducts      2240 non-null   int64  
 12  MntFishProducts      2240 non-null   int64  
 13  MntSweetProducts     2240 non-null   int64  
 14  MntGoldProds         2240 non-null   int64  
 15  NumDealsPurchases    2240 non-null   int64  
 16  NumWebPurchases      2240 non-null   int64  
 17  NumCatalogPurchases  2240 non-null   int64  
 18  NumStorePurchases    2240 non-null   int64  
 19  NumWebVisitsMonth    2240 non-null   int64  
 20  AcceptedCmp3         2240 non-null   int64  
 21  AcceptedCmp4         2240 non-null   int64  
 22  AcceptedCmp5         2240 non-null   int64  
 23  AcceptedCmp1         2240 non-null   int64  
 24  AcceptedCmp2         2240 non-null   int64  
 25  Complain             2240 non-null   int64  
 26  Z_CostContact        2240 non-null   int64  
 27  Z_Revenue            2240 non-null   int64  
 28  Response             2240 non-null   int64  
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
```

```
#checking if there are any null values in each attribute
data.isnull().sum()
```

```
ID                      0
Year_Birth              0
Education               0
Marital_Status          0
Income                 24
Kidhome                 0
Teenhome                0
Dt_Customer             0
Recency                 0
MntWines                0
MntFruits               0
MntMeatProducts         0
MntFishProducts         0
MntSweetProducts        0
MntGoldProds            0
NumDealsPurchases       0
NumWebPurchases         0
NumCatalogPurchases     0
NumStorePurchases       0
NumWebVisitsMonth       0
AcceptedCmp3            0
AcceptedCmp4            0
AcceptedCmp5            0
AcceptedCmp1            0
AcceptedCmp2            0
Complain                0
Z_CostContact           0
Z_Revenue               0
Response                0
dtype: int64
```

4.2.1 Data Cleaning

From the above results we can see that there are 24 missing values in Income attribute

So, we are going to drop the rows that have missing values.

```
data.dropna(axis=0, inplace = True)
```

```
data
```

```
         ID  Year_Birth   Education Marital_Status   Income  Kidhome  \
0      5524        1957  Graduation         Single  58138.0        0
1      2174        1954  Graduation         Single  46344.0        1
2      4141        1965  Graduation       Together  71613.0        0
3      6182        1984  Graduation       Together  26646.0        1
4      5324        1981         PhD        Married  58293.0        1
...     ...         ...         ...            ...      ...      ...
2235  10870        1967  Graduation        Married  61223.0        0
2236   4001        1946         PhD       Together  64014.0        2
2237   7270        1981  Graduation       Divorced  56981.0        0
2238   8235        1956      Master       Together  69245.0        0
2239   9405        1954         PhD        Married  52869.0        1

      Teenhome Dt_Customer  Recency  MntWines  ...  NumWebVisitsMonth  \
0            0  04-09-2012       58       635  ...                  7
1            1  08-03-2014       38        11  ...                  5
2            0  21-08-2013       26       426  ...                  4
3            0  10-02-2014       26        11  ...                  6
4            0  19-01-2014       94       173  ...                  5
...        ...         ...      ...       ...  ...                ...
2235         1  13-06-2013       46       709  ...                  5
2236         1  10-06-2014       56       406  ...                  7
2237         0  25-01-2014       91       908  ...                  6
2238         1  24-01-2014        8       428  ...                  3
2239         1  15-10-2012       40        84  ...                  7

      AcceptedCmp3  AcceptedCmp4  AcceptedCmp5  AcceptedCmp1  AcceptedCmp2  \
0                0             0             0             0             0
1                0             0             0             0             0
2                0             0             0             0             0
3                0             0             0             0             0
4                0             0             0             0             0
...            ...           ...           ...           ...           ...
2235             0             0             0             0             0
2236             0             0             0             1             0
2237             0             1             0             0             0
2238             0             0             0             0             0
2239             0             0             0             0             0

      Complain  Z_CostContact  Z_Revenue  Response
0            0              3         11         1
1            0              3         11         0
2            0              3         11         0
3            0              3         11         0
4            0              3         11         0
...        ...            ...        ...       ...
2235         0              3         11         0
2236         0              3         11         0
2237         0              3         11         0
2238         0              3         11         0
```

```
2239         0              3         11          1

[2216 rows x 29 columns]
```

4.2.2 Feature Engineering

We make a new attribute which indicates the number of days a customer is registered to the system. In order to do this, the Dt_Customer attribute should be converted into date datatype.

```
data['Dt_Customer'] = pd.to_datetime(data['Dt_Customer'])

data['Customer_days'] = data[['Dt_Customer']].max() - data[['Dt_Customer']]

data['Customer_days'] = pd.to_numeric(data['Customer_days'])

data[['Customer_days']]
```

```
          Customer_days
0     83894400000000000
1     10800000000000000
2     40780800000000000
3      5616000000000000
4     27734400000000000
...                 ...
2235  46742400000000000
2236   5270400000000000
2237  27216000000000000
2238  27302400000000000
2239  67564800000000000

[2216 rows x 1 columns]
```

Also, we can find the age of each customer

```
data["Age"] = 2022 - data['Year_Birth']
```

There are some attributes which indicate the number of purchased products. From all of these we can create a new attribute called total_num_purchased which illustrates the total amount of products bought.

```
data['total_num_purchased'] = data['MntWines'] + data['MntFruits'] + data['MntMeat
Products'] + data['MntFishProducts'] + data['MntSweetProducts'] + data['MntGoldPro
ds']

data[['total_num_purchased']]
```

```
      total_num_purchased
0                    1617
1                      27
2                     776
3                      53
4                     422
...                   ...
2235                 1341
```

```
2236                  444
2237                 1241
2238                  843
2239                  172

[2216 rows x 1 columns]
```

We can create a new attribute so as to have the number of children of each customer by adding the kidhome and teenhome columns

```
data['children'] = data['Kidhome'] + data['Teenhome']

data['children']
```

```
0       0
1       2
2       0
3       1
4       1
       ..
2235    1
2236    3
2237    0
2238    1
2239    2
Name: children, Length: 2216, dtype: int64
```

Now, let's create a new feature called family_members so as to have the number of the family size

```
#Check the distinct values of Marital_Status attribute

data['Marital_Status'].value_counts()
```

```
Married     857
Together    573
Single      471
Divorced    232
Widow        76
Alone         3
YOLO          2
Absurd        2
Name: Marital_Status, dtype: int64
```

```
data['family_members'] = data["Marital_Status"].replace({"Single": 1,"Divorced": 1
,"Widow": 1,"Alone": 1,"Absurd": 1,"YOLO": 1, "Married":2, "Together":2})+ data['c
hildren']

data['family_members']
```

```
0       1
1       3
2       2
3       3
4       3
```

```
       ..
2235    3
2236    5
2237    1
2238    3
2239    4
Name: family_members, Length: 2216, dtype: int64
```

Dropping some non-important features. The goal in a clustering problem is to group similar observations together based on the values of the attributes. In our case, characteristics like the ID, year of birth, Dt_Customer, Z_CostContact, Z_Revenue do not provide any information about the similarity of the observations. This means that they can be removed. Moreover, including attributes like the previous can also cause the model to overfit.

```
data.drop(['ID', 'Year_Birth', 'Dt_Customer', 'Z_CostContact', 'Z_Revenue'], axis=
1, inplace = True)
```

```
data.head()
```

```
    Education Marital_Status   Income  Kidhome  Teenhome  Recency  MntWines  \
0  Graduation         Single  58138.0        0         0       58       635   
1  Graduation         Single  46344.0        1         1       38        11   
2  Graduation       Together  71613.0        0         0       26       426   
3  Graduation       Together  26646.0        1         0       26        11   
4         PhD        Married  58293.0        1         0       94       173   

   MntFruits  MntMeatProducts  MntFishProducts  ...  AcceptedCmp5  \
0         88              546              172  ...             0   
1          1                6                2  ...             0   
2         49              127              111  ...             0   
3          4               20               10  ...             0   
4         43              118               46  ...             0   

   AcceptedCmp1  AcceptedCmp2  Complain  Response      Customer_days  Age  \
0             0             0         0         1  83894400000000000   65   
1             0             0         0         0  10800000000000000   68   
2             0             0         0         0  40780800000000000   57   
3             0             0         0         0   5616000000000000   38   
4             0             0         0         0  27734400000000000   41   

   total_num_purchased  children  family_members  
0                 1617         0               1  
1                   27         2               3  
2                  776         0               2  
3                   53         1               3  
4                  422         1               3  

[5 rows x 29 columns]
```

Let's take a look at some statistics of the dataset

```
data.describe()
```

```
              Income      Kidhome     Teenhome      Recency     MntWines  \
count    2216.000000  2216.000000  2216.000000  2216.000000  2216.000000
mean    52247.251354     0.441787     0.505415    49.012635   305.091606
std     25173.076661     0.536896     0.544181    28.948352   337.327920
min      1730.000000     0.000000     0.000000     0.000000     0.000000
25%     35303.000000     0.000000     0.000000    24.000000    24.000000
50%     51381.500000     0.000000     0.000000    49.000000   174.500000
75%     68522.000000     1.000000     1.000000    74.000000   505.000000
max    666666.000000     2.000000     2.000000    99.000000  1493.000000

         MntFruits  MntMeatProducts  MntFishProducts  MntSweetProducts  \
count  2216.000000      2216.000000      2216.000000       2216.000000
mean     26.356047       166.995939        37.637635         27.028881
std      39.793917       224.283273        54.752082         41.072046
min       0.000000         0.000000         0.000000          0.000000
25%       2.000000        16.000000         3.000000          1.000000
50%       8.000000        68.000000        12.000000          8.000000
75%      33.000000       232.250000        50.000000         33.000000
max     199.000000      1725.000000       259.000000        262.000000

       MntGoldProds  ...  AcceptedCmp5  AcceptedCmp1  AcceptedCmp2  \
count   2216.000000  ...   2216.000000   2216.000000   2216.000000
mean      43.965253  ...      0.073105      0.064079      0.013538
std       51.815414  ...      0.260367      0.244950      0.115588
min        0.000000  ...      0.000000      0.000000      0.000000
25%        9.000000  ...      0.000000      0.000000      0.000000
50%       24.500000  ...      0.000000      0.000000      0.000000
75%       56.000000  ...      0.000000      0.000000      0.000000
max      321.000000  ...      1.000000      1.000000      1.000000

          Complain     Response  Customer_days          Age  \
count  2216.000000  2216.000000   2.216000e+03  2216.000000
mean      0.009477     0.150271   4.423735e+16    53.179603
std       0.096907     0.357417   2.008532e+16    11.985554
min       0.000000     0.000000   0.000000e+00    26.000000
25%       0.000000     0.000000   2.937600e+16    45.000000
50%       0.000000     0.000000   4.432320e+16    52.000000
75%       0.000000     0.000000   5.927040e+16    63.000000
max       1.000000     1.000000   9.184320e+16   129.000000

       total_num_purchased     children  family_members
count          2216.000000  2216.000000     2216.000000
mean            607.075361     0.947202        2.592509
std             602.900476     0.749062        0.905722
min               5.000000     0.000000        1.000000
25%              69.000000     0.000000        2.000000
50%             396.500000     1.000000        3.000000
75%            1048.000000     1.000000        3.000000
max            2525.000000     3.000000        5.000000

[8 rows x 27 columns]
```

4.2.3 Managing the categorical variables

```
#keeping the attributes names in a list

data_col = data.columns
data_col = list(data_col)
data_col
```

```
['Education',
 'Marital_Status',
 'Income',
 'Kidhome',
 'Teenhome',
 'Recency',
 'MntWines',
 'MntFruits',
 'MntMeatProducts',
 'MntFishProducts',
 'MntSweetProducts',
 'MntGoldProds',
 'NumDealsPurchases',
 'NumWebPurchases',
 'NumCatalogPurchases',
 'NumStorePurchases',
 'NumWebVisitsMonth',
 'AcceptedCmp3',
 'AcceptedCmp4',
 'AcceptedCmp5',
 'AcceptedCmp1',
 'AcceptedCmp2',
 'Complain',
 'Response',
 'Customer_days',
 'Age',
 'total_num_purchased',
 'children',
 'family_members']
```

In order to check the distribution of categorical and numerical attributes, barplots and boxplots created respectively below.

```
import matplotlib.pyplot as plt

fig = plt.figure(figsize=(18, 36))

# Increase the height and width of the subplots
plt.subplots_adjust(wspace=0.3, hspace=0.4)

i = 0
num_cols = []
object_cols = []
# Loop through the columns of the dataframe
for column in data_col:
    # Check the data type of the column
    if data[column].dtype in ['int64', 'float64']:
        ## Create box plots for the numerical input variables
        num_cols.append(column)
        i += 1
        plt.subplot(8,4, i)
        plt.boxplot(data[column])
        plt.xlabel(column)
    else:
        # Create bar plots for the categorical input variables
        object_cols.append(column)
        i += 1
        plt.subplot(8,4, i)
        plt.bar(data[column].value_counts().index, data[column].value_counts().val
ues)
        plt.xticks(rotation=45)
        plt.xlabel(column)
plt.show()
```

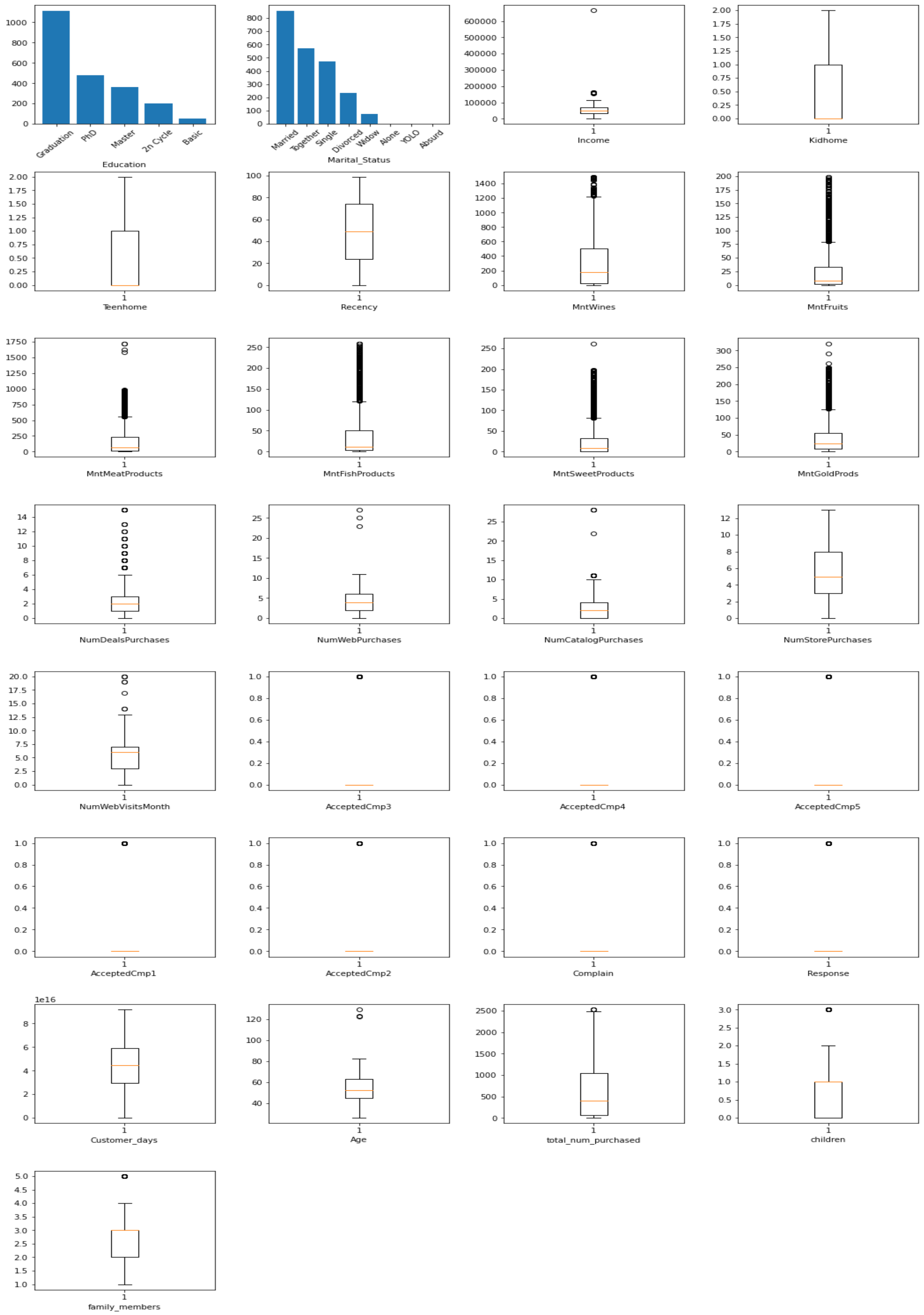

Figure 4.7: Barplots and Boxplots for the corresponding features

From the above figure we can draw some information:

Categorical features:

- Most of the customers have graduated and also regarding the marital status most of the customers are married.

Numerical features:

- Income: one customer seems to have income around 600.000.
- Age: two customers are more than 120 years old.

The two above observations of the numerical features can lead us to delete these extreme values.

Moreover, it is important to observe that there are a lot of extreme values to attributes relevant to the number of purchases of products like number of fruits, number of wines etc. If we take into account that these variables store the total number of purchases a customer makes to each type of product for 2 years, it seems logic to have some non-average customers that have bought a lot of products.

```
#excluding some extreme values

data = data[(data["Age"]<98)]
data = data[(data["Income"]<600000)]

data
```

```
       Education Marital_Status   Income  Kidhome  Teenhome  Recency  \
0     Graduation         Single  58138.0        0         0       58   
1     Graduation         Single  46344.0        1         1       38   
2     Graduation       Together  71613.0        0         0       26   
3     Graduation       Together  26646.0        1         0       26   
4            PhD        Married  58293.0        1         0       94   
...          ...            ...      ...      ...       ...      ...   
2235  Graduation        Married  61223.0        0         1       46   
2236         PhD       Together  64014.0        2         1       56   
2237  Graduation       Divorced  56981.0        0         0       91   
2238      Master       Together  69245.0        0         1        8   
2239         PhD        Married  52869.0        1         1       40   

      MntWines  MntFruits  MntMeatProducts  MntFishProducts  ...  \
0          635         88              546              172  ...   
1           11          1                6                2  ...   
2          426         49              127              111  ...   
3           11          4               20               10  ...   
4          173         43              118               46  ...   
...        ...        ...              ...              ...  ...   
2235       709         43              182               42  ...   
2236       406          0               30                0  ...   
2237       908         48              217               32  ...   
2238       428         30              214               80  ...   
2239        84          3               61                2  ...   

      AcceptedCmp5  AcceptedCmp1  AcceptedCmp2  Complain  Response  \
0                0             0             0         0         1   
1                0             0             0         0         0   
```

```
2                0             0             0         0         0
3                0             0             0         0         0
4                0             0             0         0         0
...            ...           ...           ...       ...       ...
2235             0             0             0         0         0
2236             0             1             0         0         0
2237             0             0             0         0         0
2238             0             0             0         0         0
2239             0             0             0         0         1

          Customer_days  Age  total_num_purchased  children  family_members
0     83894400000000000   65                 1617         0               1
1     10800000000000000   68                   27         2               3
2     40780800000000000   57                  776         0               2
3      5616000000000000   38                   53         1               3
4     27734400000000000   41                  422         1               3
...                 ...  ...                  ...       ...             ...
2235  46742400000000000   55                 1341         1               3
2236   5270400000000000   76                  444         3               5
2237  27216000000000000   41                 1241         0               1
2238  27302400000000000   66                  843         1               3
2239  67564800000000000   68                  172         2               4

[2212 rows x 29 columns]
```

We create a new list that has first the numerical features and then the categorical in order to use it for the next stage which is the one-hot encoder

```
final_data = num_cols + object_cols
final_data
```

```
['Income',
 'Kidhome',
 'Teenhome',
 'Recency',
 'MntWines',
 'MntFruits',
 'MntMeatProducts',
 'MntFishProducts',
 'MntSweetProducts',
 'MntGoldProds',
 'NumDealsPurchases',
 'NumWebPurchases',
 'NumCatalogPurchases',
 'NumStorePurchases',
 'NumWebVisitsMonth',
 'AcceptedCmp3',
 'AcceptedCmp4',
 'AcceptedCmp5',
 'AcceptedCmp1',
 'AcceptedCmp2',
 'Complain',
 'Response',
 'Customer_days',
 'Age',
 'total_num_purchased',
 'children',
 'family_members',
 'Education',
 'Marital_Status']
```

Machines do not understand categorical features. Therefore, we need to convert these features into numerical ones. For this reason, we use the OneHotEncoder from the sklearn library which converts the categorical variables into a numerical form. The result is a binary vector with dimensionality equal to the number of categories (if drop_first parameter is set to True then the first dimension meaning the first category is not used. This is done for dimensionality reduction purposes). All the values in the vector are zero except for the index which corresponds to the category every time.

```
from sklearn.preprocessing import OneHotEncoder

X = data[object_cols].values

# OneHotEncoder() object
data_ohe = OneHotEncoder()

encoded_data = data_ohe.fit_transform(X).toarray()

encoded_data = pd.get_dummies(data[final_data], drop_first=True)

encoded_data
```

```
C:\Users\galan\anaconda3\lib\site-packages\scipy\__init__.py:138: UserWarning: A N
umPy version >=1.16.5 and <1.23.0 is required for this version of SciPy (detected
version 1.23.4)
  warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion} is requir
ed for this version of "
```

```
       Income  Kidhome  Teenhome  Recency  MntWines  MntFruits  \
0     58138.0        0         0       58       635         88
1     46344.0        1         1       38        11          1
2     71613.0        0         0       26       426         49
3     26646.0        1         0       26        11          4
4     58293.0        1         0       94       173         43
...       ...      ...       ...      ...       ...        ...
2235  61223.0        0         1       46       709         43
2236  64014.0        2         1       56       406          0
2237  56981.0        0         0       91       908         48
2238  69245.0        0         1        8       428         30
2239  52869.0        1         1       40        84          3

      MntMeatProducts  MntFishProducts  MntSweetProducts  MntGoldProds  ...  \
0                 546              172                88            88  ...
1                   6                2                 1             6  ...
2                 127              111                21            42  ...
3                  20               10                 3             5  ...
4                 118               46                27            15  ...
...               ...              ...               ...           ...  ...
2235              182               42               118           247  ...
```

```
2236               30                 0                 0             8  ...
2237              217                32                12            24  ...
2238              214                80                30            61  ...
2239               61                 2                 1            21  ...

      Education_Graduation  Education_Master  Education_PhD  \
0                        1                 0              0   
1                        1                 0              0   
2                        1                 0              0   
3                        1                 0              0   
4                        0                 0              1   
...                    ...               ...            ...   
2235                     1                 0              0   
2236                     0                 0              1   
2237                     1                 0              0   
2238                     0                 1              0   
2239                     0                 0              1   

      Marital_Status_Alone  Marital_Status_Divorced  Marital_Status_Married  \
0                        0                        0                       0   
1                        0                        0                       0   
2                        0                        0                       0   
3                        0                        0                       0   
4                        0                        0                       1   
...                    ...                      ...                     ...   
2235                     0                        0                       1   
2236                     0                        0                       0   
2237                     0                        1                       0   
2238                     0                        0                       0   
2239                     0                        0                       1   

      Marital_Status_Single  Marital_Status_Together  Marital_Status_Widow  \
0                         1                        0                     0   
1                         1                        0                     0   
2                         0                        1                     0   
3                         0                        1                     0   
4                         0                        0                     0   
...                     ...                      ...                   ...   
2235                      0                        0                     0   
2236                      0                        1                     0   
2237                      0                        0                     0   
2238                      0                        1                     0   
2239                      0                        0                     0   

      Marital_Status_YOLO  
0                       0  
1                       0  
2                       0  
3                       0  
4                       0  
...                   ...  
2235                    0  
2236                    0  
```

```
2237                    0
2238                    0
2239                    0

[2212 rows x 38 columns]
```

4.2.4 Correlation among the attributes

This is an important process since it presents how strong is the correlation between the features. We can compute the Pearson coefficient which produces the pairwise correlations of the dataset attributes.

```
correlations = encoded_data.corr()

# Compute the absolute values of the correlations
abs_correlations = correlations.abs()

#Creating a heatmap with seaborn library

import seaborn as sns


f, ax = plt.subplots(figsize=(60, 60))

# Generating a custom diverging colormap
cmap = sns.color_palette("Blues", as_cmap=True)

# Drawing the heatmap
sns.heatmap(correlations, cmap="YlGnBu", vmax=1, vmin=-1, center=0,
            annot=True, fmt='.2f', square=True, linewidths=.5, cbar_kws={"shrink":
.5})

# Show the plot
plt.show()
```

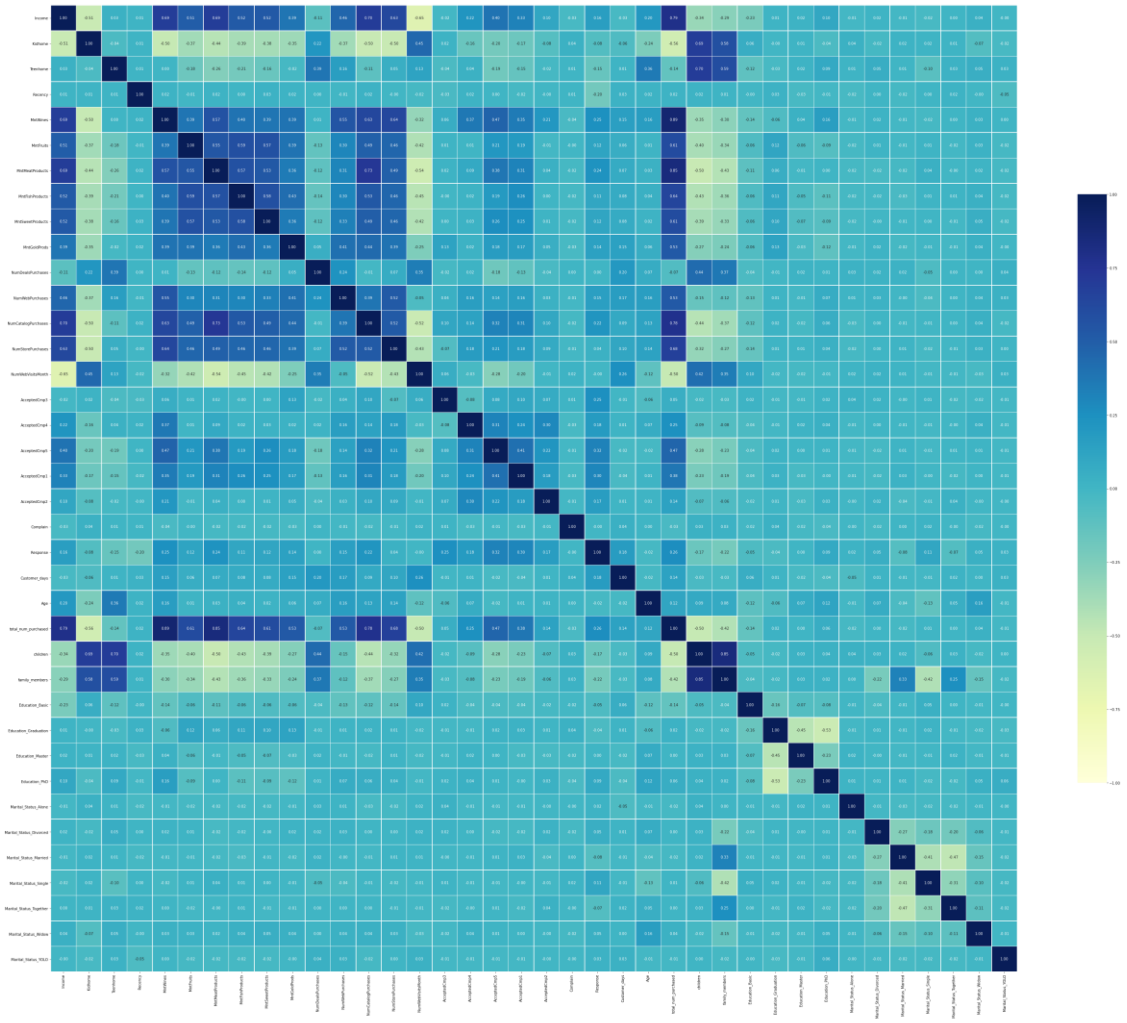

Figure 4.8: Heatmap correlation of the attributes

4.3 Standardization

Now, we should standardize the data since we are going to use k-means clustering algorithm. This is very important because this method uses Euclidean distance measure to calculate the similarity between data observations. Furthermore, this measure is sensitive to the scale of the variables, therefore if there are any variables with a larger scale, they will have a greater impact on the clustering results. Using standardization on the dataset will transform the variables so as to have a mean of 0 and a standard deviation of 1. This means that the variables have similar scale and it allows for a fairer comparison of their influence on the clustering results.

```
from sklearn.preprocessing import StandardScaler

# Create a StandardScaler object
scaler = StandardScaler()

# Standardize the data
stand_data = scaler.fit_transform(encoded_data)

stand_data
```

```
array([[ 0.28710487, -0.82275354, -0.92969866, ..., -0.58988012,
         -0.18862801, -0.03008284],
        [-0.26088203,  1.04002111,  0.90809708, ..., -0.58988012,
         -0.18862801, -0.03008284],
        [ 0.9131964 , -0.82275354, -0.92969866, ...,  1.69525969,
         -0.18862801, -0.03008284],
        ...,
        [ 0.23334696, -0.82275354, -0.92969866, ..., -0.58988012,
         -0.18862801, -0.03008284],
        [ 0.80317156, -0.82275354,  0.90809708, ...,  1.69525969,
         -0.18862801, -0.03008284],
        [ 0.04229031,  1.04002111,  0.90809708, ..., -0.58988012,
         -0.18862801, -0.03008284]])
```

4.4 Dimensionality Reduction

We will create a new target space with 3 dimensions by applying PCA technique on the standardized data and then use this new 3d space into k-means clustering technique.

This is a very important process since PCA (Principal Component Analysis) is a method used for dimensionality reduction. Using PCA with k-means clustering algorithm is quite useful because it can help to reduce the number of features in the dataset. This results in making the clustering procedure more computationally efficient and can also help to avoid overfitting. Furthermore, PCA is a technique which helps to identify the underlying structure of the dataset by identifying the principal components which can explain the most variation in the dataset. The principal components found can be utilized as inputs to the k-means algorithm. Having the most important parts of the data can help to improve the quality of the clustering results.

```
#Creating a new target space with 3 dimension

# Import PCA from scikit-learn
from sklearn.decomposition import PCA

pca = PCA(n_components=3)

PCA_data = pd.DataFrame(pca.fit_transform(stand_data), columns=(["dim1","dim2", "d
im3"]))

print("Dimensionality:", PCA_data.shape)
```

```
Dimensionality: (2212, 3)
```

```
PCA_data
```

```
          dim1      dim2      dim3
0     4.692403 -0.943147 -0.389921
1    -2.853770 -0.315184 -0.419588
2     2.037152 -0.533357 -1.407003
3    -2.703264 -1.535852 -0.593795
4    -0.646819  0.399667 -0.123425
...        ...       ...       ...
2207  2.359628  1.740971 -1.994264
2208 -2.302000  4.513061  1.515251
```

```
2209  2.386060 -1.182756  0.782209
2210  1.499541  1.640682 -0.710429
2211 -2.389719  2.080685  1.610928

[2212 rows x 3 columns]
```

```
#A 3D visualization of the data in the 3D Reduced Space
x =PCA_data["dim1"]
y =PCA_data["dim2"]
z =PCA_data["dim3"]

#Plotting the figure
fig = plt.figure(figsize=(12,10))
ax = fig.add_subplot(111, projection="3d")
ax.scatter(x,y,z, c="blue", marker="o" )
ax.set_title("A 3D visualization of the data in the 3D Reduced Space")
plt.show()
```

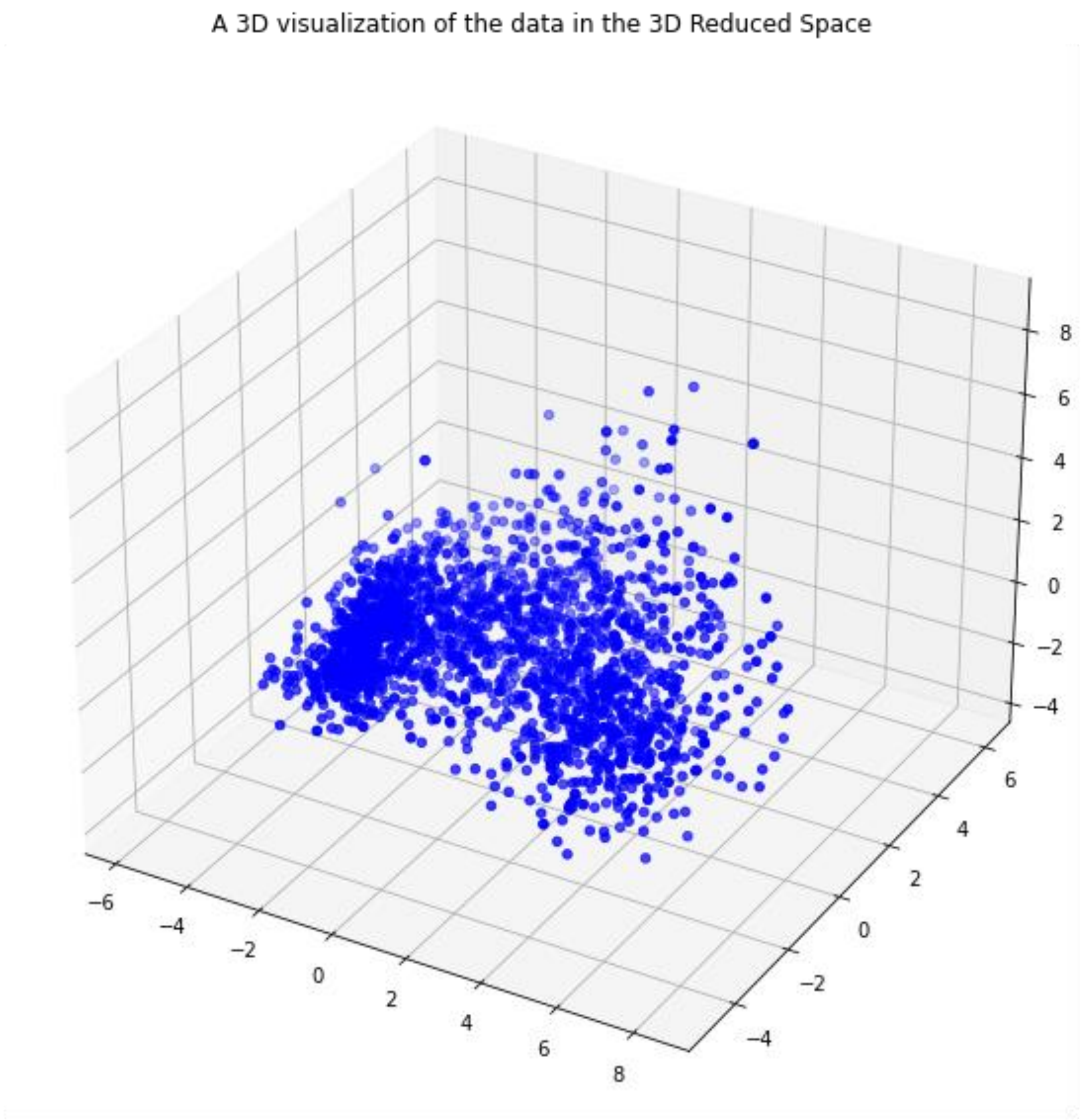

Figure 4.9: A visualization on the reduced space

4.5 Clustering

In this step we will use the K-means clustering method on the reduced target space with only 3 dimensions. This is a widely used technique in pattern recognition, data mining, and machine learning because it can unveil the structure and patterns in large and complex datasets.

This algorithm has some advantages such as its simplicity. To apply this method is easy to understand and implement, and it can be applied to a wide range of datasets. Furthermore, k-means is computationally efficient and thus can be used for large datasets.

Additionally, k-means algorithm can be utilized as a preprocessing step for other machine learning algorithms, for example regression and classification. Using clustering and recognizing patterns of the structure of the data can improve the performance of the other machine learning methods.

Some of the applications that k-means can be used are image segmentation, market segmentation, anomaly detection and many more.

Below, we use the elbow technique which is a vital method in k-means clustering since it helps to identify the optimal number of clusters by measuring the within-cluster sum of squares (WCSS) for each value of k and selecting the value of k where the rate of change in WCSS begins to slow down. This results in avoiding overfitting and increasing the performance of the k-means model.

```
from sklearn.cluster import KMeans

inertias = []

# Run k-Means 20 times for different numbers of initial clusters (notice how n_clu
ster changes at each step of the loop)
for i in range(2, 20):
    km = KMeans(n_clusters=i, init='k-means++', n_init=10, max_iter=300, random_st
ate=42)
    km.fit(PCA_data)
    inertias.append(km.inertia_)

# Plot the distortions of the 20 k-Means executions
plt.plot(range(2, 20), inertias, marker='o')
plt.xlabel('Number of clusters')
plt.ylabel('Distortion')
plt.tight_layout()

plt.show()
```

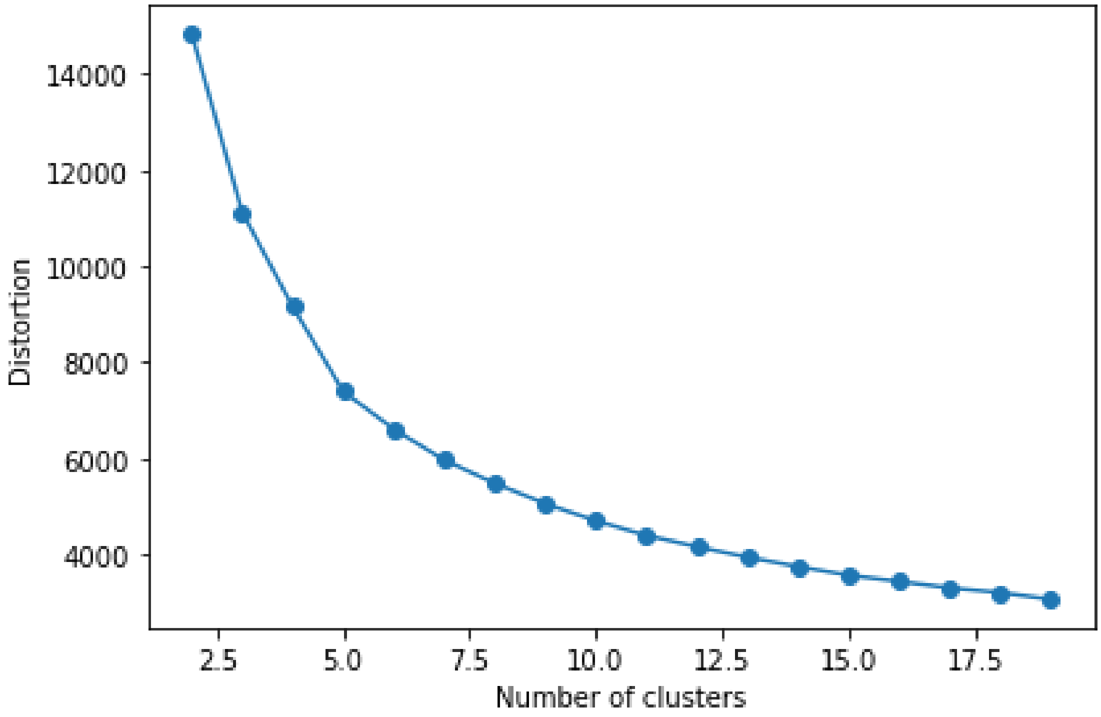

Figure 4.10: Illustration of elbow method

From the above diagram we can see that the optimal number of clusters is 5. This is the point where the 'Elbow' is creating.

```
# Initialize k-Means and its hyper-parameters
km = KMeans(n_clusters=5, init='k-means++', n_init=10, max_iter=300, random_state=
42)

# Execute k-Means on the data points X
data_pred = km.fit_predict(PCA_data)

PCA_data['Clusters'] = data_pred

data['Clusters'] = data_pred
```

Inertia is utilized as a measure of the compactness of the clusters in k-means clustering algorithm. The smaller the values the better the clustering.

This value is important as a quantitative measure of the quality of the clustering, to determine the optimal number of clusters. Also, it is a measure of the similarity between clusters and provides a way to compare the homogeneity of different clusters.

The degree to which the data observations within a cluster are similar to each other is referred as Homogeneity of clusters. Having a high degree of homogeneity within a cluster in a k-means algorithm indicates that the data points within that cluster are close to the centroid of that cluster. Conversely, a low degree of homogeneity within a cluster means that the data points within that cluster are far from the centroid of that cluster.

```
# Print the inertia of our k-Means clustering
print('Distortion (inertia): %.3f' % km.inertia_)
```

```
Distortion (inertia): 7395.752
```

4.6 Visualization and evaluation of the model

In this stage we will visualize the clusters by making a 3D plot. Furthermore, we will use a metric and some figures so as to try and evaluate the performance of the model.

It is important to know that since we used an unsupervised machine learning model such as k-means clustering we do not have a tagged feature to evaluate or score our model meaning we do not know the true class of each point. Therefore, in this step we just try to study the patterns and trends in the clusters created and determine the nature of the clusters' patterns.

All in all, we will use apart from the silhouette score also exploratory data analysis and draw some conclusions.

Below the silhouette score is used. It is a metric utilized to evaluate the quality of clustering in k-means. Specifically, it provides a way to evaluate the quality of clustering which is independent of the number of clusters, and it can be used to determine the optimal number of clusters.

It takes into account both the similarity within clusters and the dissimilarity between clusters which is an important indicator of the cluster's homogeneity.

In general, a silhouette score ranges from -1 to 1. A good clustering will have a silhouette score close to 1 and a poor clustering will have a score close to -1. A silhouette score of around 0 indicates that the sample is on or close to the decision boundary between two clusters, which is not the ideal case.

```
from sklearn.metrics import silhouette_score

score = silhouette_score(PCA_data, km.labels_)

print(score)
```

```
0.47864094417225306
```

From the above result we can observe that the silhouette score is around 0.48 which means that the data point is moderately well-matched to its own cluster. Nevertheless, not as well-matched as it could be. A silhouette scores between 0.4 and 0.5 are considered to be moderate which means that there could be some room for improvement.

In general, a moderate silhouette score could mean that the cluster is not very compact or well-defined, or that the data observations within the cluster are somehow similar to data points in other clusters.

```
#we use a variable pal so as to pass a list of hex colors to be used in the next v
isualization figures
import matplotlib.colors as mcolors

pal = ["#0047ab", "#aa201e", "#7ee23b", "#ffff66","#525659"]
custom_cmap = mcolors.ListedColormap(pal)

#plotting a figure in order to make a 3d projection on the reduced space after app
lying PCA and k-means clustering

fig = plt.figure(figsize=(10,8))
ax = plt.subplot(111, projection='3d', label="bla")
```

```
ax.scatter(x, y, z, s=40, c=PCA_data["Clusters"], marker='o', cmap = custom_cmap)
ax.set_title("Clusters")
plt.show()
```

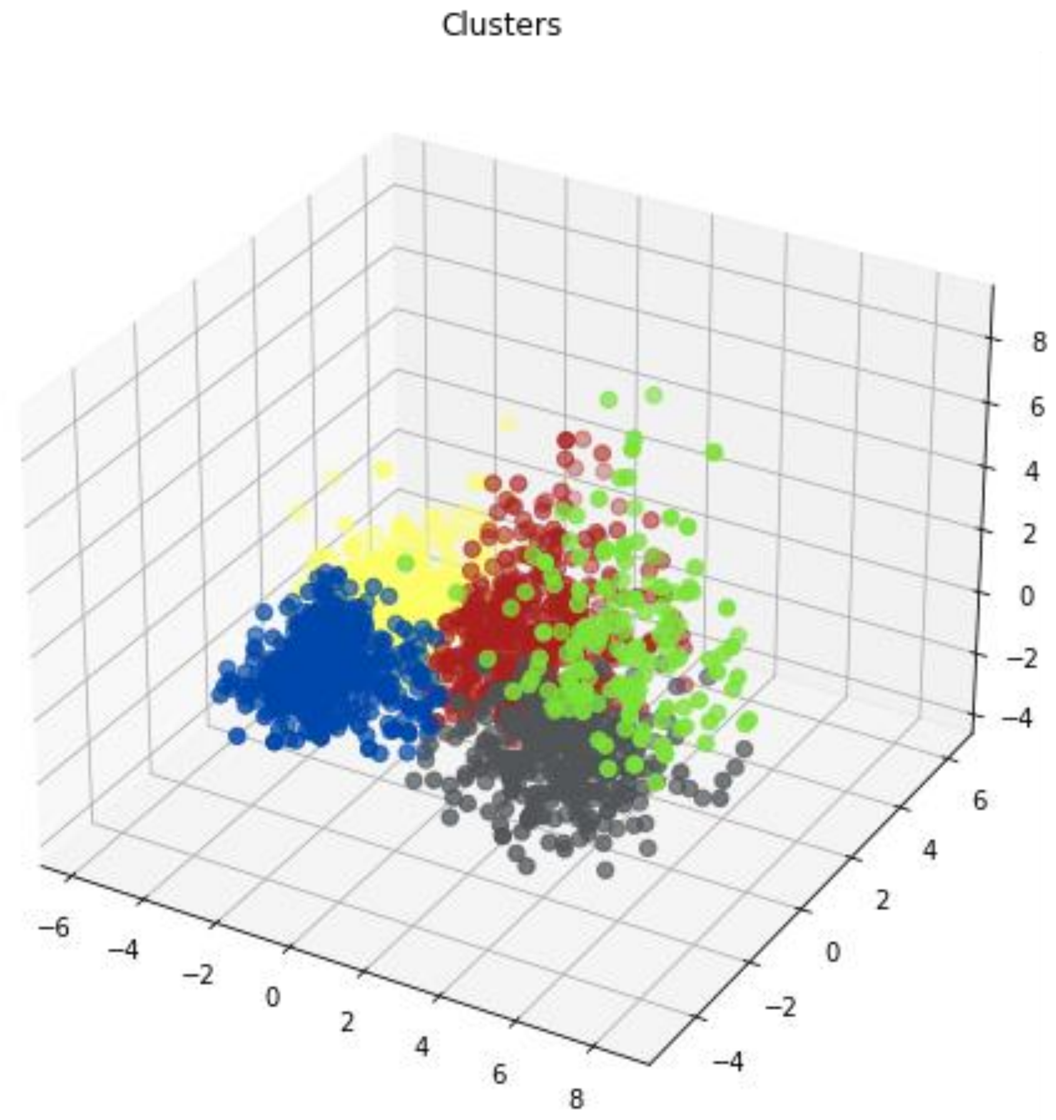

Figure 4.11: Capturing the data on the reduced space

From the above figure it seems roughly that the clusters are more or less capturing good the data in the reduced space.

Next, we plot the distribution of each cluster in the dataset with all the dimensions.

```
#Plotting countplot of clusters
pl = sns.countplot(x=data["Clusters"], palette= pal)
pl.set_title("Distribution of the clusters")
plt.show()
```

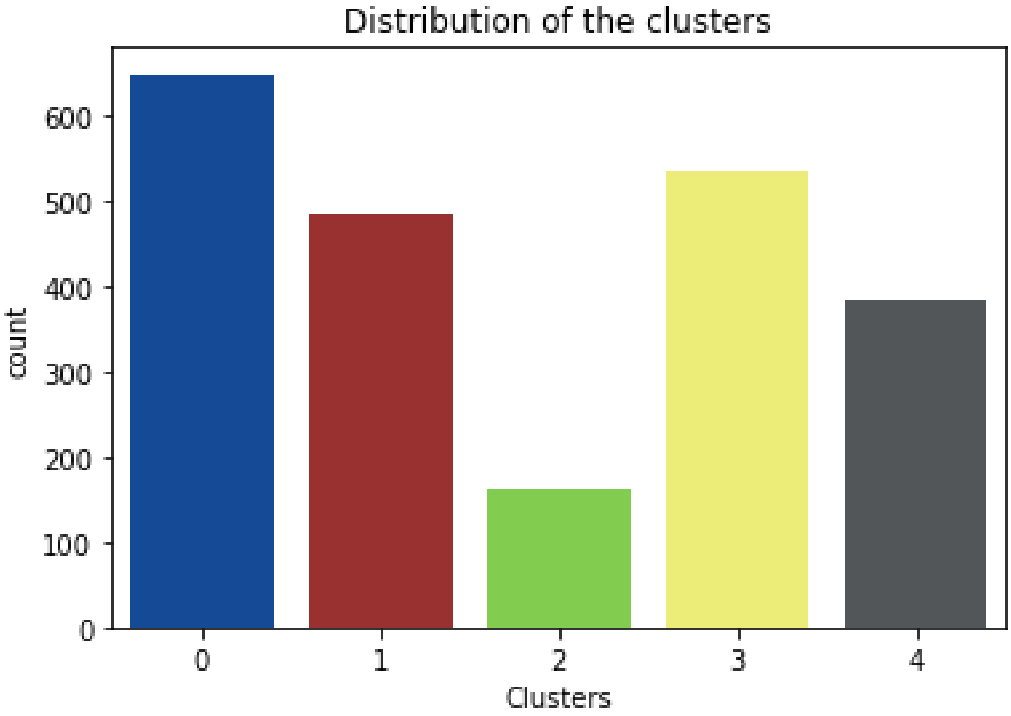

Figure 4.12: Distribution of the clusters on the whole data

We can observe from the above figure that the clusters are fairly distributed except cluster 2. Specifically, roughly 180 customers have been assigned to cluster 2. Whereas cluster 0 has approximately 600 customers assigned to it.

Below, we use the total number of items purchased attribute and the income attribute in order to observe how the clusters behave between these two attributes in 2d space.

```
#Creating a 2d plot
pl = sns.scatterplot(data = data,x=data["total_num_purchased"], y=data["Income"],h
ue=data["Clusters"], palette= pal)
pl.set_title("Clusters based on spendings and income attributes")
plt.legend()
plt.show()
```

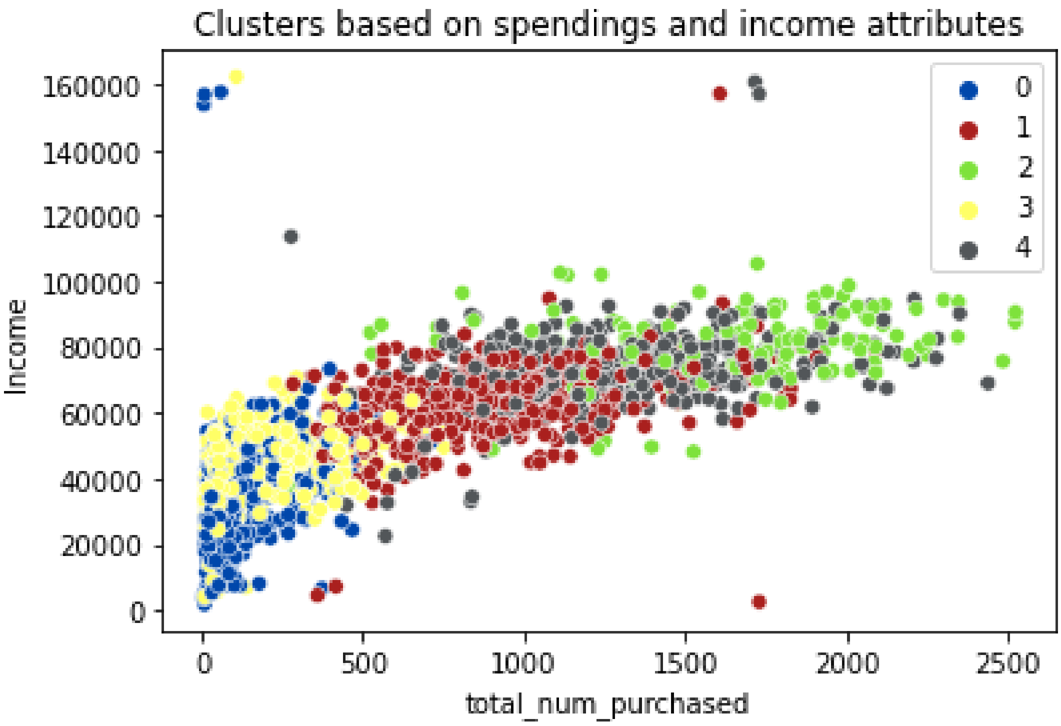

Figure 4.13: Illustration of the clusters regarding their income and spendings

From the above figure we can draw some conclusions for the 5 clusters created regarding the customers' income and spending habits. More specifically, we can see:

- **Group 0:** low spendings and low income
- **Group 1:** low spendings and high income
- **Group 2:** high spendings and high income
- **Group 3:** low spendings and average income
- **Group 4:** average spendings and high income

Last but certainly not least, let's examine how the campaigns did by plotting the sum of campaigns and check which group accepted the most campaigns.

```
#Creating a new feature to get a sum of accepted campaigns
data["Sum_promo"] = data["AcceptedCmp1"]+ data["AcceptedCmp2"]+ data["AcceptedCmp3
"]+ data["AcceptedCmp4"]+ data["AcceptedCmp5"]
#Plotting count of total promotions accepted.
plt.figure()
pl = sns.countplot(x=data["Sum_promo"],hue=data["Clusters"], palette= pal)
pl.set_title("Count of promotions accepted")
pl.set_xlabel("Number Of promotions accepted")
plt.show()
```

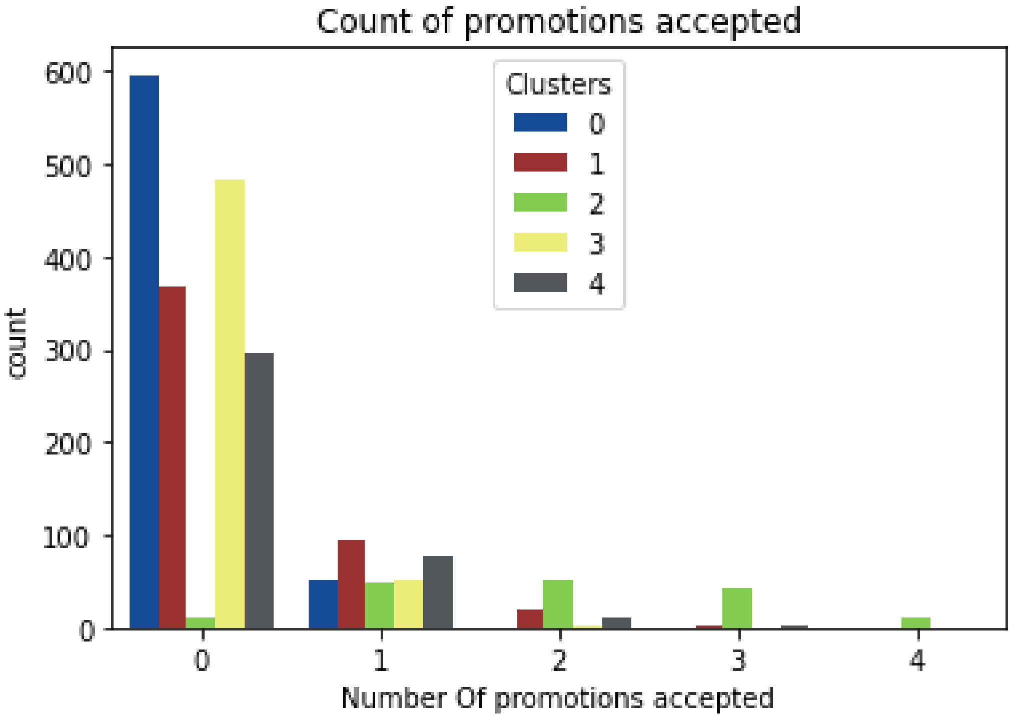

Figure 4.14: Distribution of the number of accepted promotions per cluster

The most important observation we can conclude from the above figure is that a small proportion of each of the 5 groups have accepted a promotion. Most of the groups have not accepted any of the promotions.

Also, we can see that group 2, despite having the lowest number of customers, it has some customers who have accepted 4 times some of the promotions that the company did in the past.

It is vital to conclude that the promotional activities that the company did were not so accurate. It might be needed a better strategy regarding the targeted groups and more well-planned structured campaigns.

All in all, after cleaning the data, creating some attributes, applying standardization, reducing the number of features with PCA and applying k-means clustering algorithm we draw some conclusions based on exploratory analysis.

Chapter 5: Conclusions and future applications

5.1 Recap

First of all, before we proceed to the main findings from this research it is vital to remember what has been done in this Thesis as far. An introduction to basic terms was done in Chapter 1 so as to get the reader familiar with some basic terms encountered in this research. Moreover, the main idea of each of the remaining articles found by the Systematic Literature Review approach was presented. Continuously, the purpose of the study was stated. After that, in Chapter 2 the SLR approach was illustrated comprehensively step by step and visualizations charts were created. In the next chapter, this research tried to grasp the most relevant information of each article in relation to the research objectives. Last but certainly not least, in Chapter 4 a thorough analysis of a dataset was conducted using python in jupyter notebook. This Chapter concludes the information learned from this research and answers the research objectives and questions.

5.2 Fulfillment of Research Objectives

First of all, the two research objectives stated in Chapter 1 are presented below alongside with justification answers:

Objective 1: Implementation of SLR approach on Machine Learning Clustering Techniques in a Supply Chain Management problem which is the well-known Customer Segmentation
Objective 2: Application of a clustering algorithm on a real dataset

1. For the first objective a comprehensive SLR method was taken in order to illustrate the transparent procedure and provide the tools for future replication and research in the upcoming years. On top of that, a descriptive analysis was done with the aim to support the SLR approach.
2. For the second objective, a dataset was analyzed taken from Kaggle website. Python programming language and jupyter notebook were utilized during the analysis. The clustering algorithm used was K-means which is the most well-known clustering method.

5.3 Research Questions answered

The two research questions are illustrated again and using the information retrieved from the research are addressed:

Research Question 1: What are the insights gained from using a clustering algorithm to segment customers on the dataset?

For the clustering of the dataset in this Thesis k-means algorithm was used. The results found were solid important. Specifically, customers were divided into 5 groups of people after obtaining that the optimal number of segments equals to 5. Also, the distribution of the groups was more or less the same except for one segment. Furthermore, an illustrated figure was created regarding customers' income against their spending habits. Lastly, a figure about the groups of people accepting the promotional offers by the company was made. All of the above, are critical to determine not only how the customers are separated to a number of groups but also if one prefers more insights about the groups created, they can dive in more detail using further analysis.

Research Question 2: Who can use the findings of this research?

The results of this analysis can be utilized by a wide range of audiences, including individuals and groups of people either in the academic or the private sector. The results generated meaning findings and insights can provide valuable information for researchers in various fields, as well as be taken into consideration as informative information in decision-making for individuals and organizations in the private sector. An important aspect of this analysis is that it offers versatility meaning it is applicable across diverse users and domains.

5.4 Limitations based on Methodology

During this research some limitations have come to the surface and it is important to present them thoroughly.

- The first limitation concerns the SLR method which was presented in Chapter 2. The author searched in databases for scientific articles. However, this search could be different if time had passed with the enlargement of the research. Moreover, the evaluation criteria were established by the author based on his preferences and experience. This means that there is a probability of leaving relevant literature. For example, a constraint to keep only articles written in English language could leave articles that are published in other languages and they could be relevant to the research goals.

- Another limitation is based on the dataset selected. This means that the analysis for example methods and the algorithm used can provide different results if they are applied on another dataset.

- Time is the most important factor when it comes to limitations. It is obvious that this dissertation has been affected by the element of time from conducting the SLR method to selecting and analyzing the dataset and the wiring procedure of the overall Thesis.

References

Aamer, A., Yani, L. P. E. & Priyatna, A. I. M., 2021. Data Analytics in the Supply Chain Management: Review of Machine Learning Applications in Demand Forecasting. *Operations and Supply Chain Management An International Journal.*

Abualigah, L., Khader, A. & Beta, M. A. A., 2016. *A krill herd algorithm for efficient text documents clustering.* s.l., s.n.

Abualigah, L., Khader, A. & Beta, M. A. A., 2016. *Multi-objectives-based text clustering technique using K-mean algorithm.* s.l., s.n.

Agbaje, M., Ezugwu, A. E.-S. & Els, R., 2019. Automatic Data Clustering Using Hybrid Firefly Particle Swarm Optimization Algorithm.

Aranganayagi, S. & Thangavel, K., 2007. *Clustering categorical data using silhouette coefficient as a relocating measure.* s.l., Proceedings of the International Conference on Computational Intelligence and Multimedia Applications.

Arbelaitz, O. et al., 2013. An extensive comparative study of cluster validity indices.

Arthur, D. & Vassilvitskii, S., 2007 . k-means++: the advantages of careful seeding.

Arunachalam, D., Kumar, N. & Kawalek, J. P., 2018. Understanding big data analytics capabilities in supply chain management: Unravelling the issues, challenges and implications for practice. *Transportation Research Part E.*

Aryal, A., Liao, Y., Nattuthurai, P. & Li, B., 2020. The emerging big data analytics and IoT in supply chain management: a systematic review.

Bahmani, B. et al., 2012 . Scalable k-means++.

Benabdellah, A. C., Benghabrit, A. & Bouhaddou, I., 2019. A survey of clustering algorithms for an industrial context. In: s.l.:s.n.

Bezdek, J. C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms.

Boland, et al., 2017. *Doing a systematic review : a student's guide.* s.l.:s.n.

Bottou, L. & Bengio, Y., 1994. Convergence properties of the K-means algorithms.

Bottou, L. & Bengio, Y., 1995. Convergence Properties of the K-Means Algorithms.

BURMAN, P., 1989. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods.

Changa, H. & Yeung, D.-Y., 2007. Robust path-based spectral clustering.

Chang, D.-X., Zhang, X.-D. & Zheng, C.-W., 2010. A robust dynamic niching genetic algorithm with niche migration for automatic clustering problem.

Cikovic, K. F., 2020. *CUSTOMER PROFILES IN THE ANTIQUES AND COLLECTIBLES INDUSTRY IN CROATIA USING GAUSSIAN MIXTURE MODEL CLUSTERING: AN EMPIRICAL STUDY.* s.l., s.n.

Cowgill, M. C., Harvey, R. J. & Watson, L. T., 1999. A genetic algorithm approach to cluster analysis.

Dash, M., Liu, H., Scheuermanna, P. & Tan, K. L., 2003. *Fast hierarchical clustering and its validation.* s.l.:s.n.

Das, S., Chowdhury, A. & Abraham, A., 2009. *A Bacterial Evolutionary Algorithm for automatic data clustering.* s.l., s.n.

Davies, D. L. & Bouldin, D. W., 1979. A Cluster Separation Measure.

Ding, J., Tarokh, V. & Yang, Y., 2017. Bridging AIC and BIC: A New Criterion for Autoregression.

E. Ezugwu1, A. et al., 2021. Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature.

Eberhart, R. & Kennedy, J., 1995. *A new optimizer using particle swarm theory.* s.l., s.n.

Ekawati, A. D. & University, B. N., 2019. Predictive Analytics in Employee Churn: A Systematic Literature Review.

Ezugwu, A. E. et al., 2022. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence.*

Falkenauer, E., 1998. Genetic algorithms and grouping problems.

fdsafa, n.d. afafaa.

Gallardo, J. L., Ahmed, M. A. & Jara , N., 2021. Clustering algorithm-based network planning for advanced metering infrastructure in smart grid.

Giri, C. & Chen, Y., 2022. Deep Learning for Demand Forecasting in the Fashion and Apparel Retail Industry.

Goldberg, D. E. & Holland , J. H., 1988. Genetic Algorithms and Machine Learning.

Hamerly, G. & Elkan, C., 2003. Learning the k in k-means.

Hartigan, J. A. & Wong, M. A., 1979. Algorithm AS 136: A K-Means Clustering Algorithm.

Hmednaa, B., ElMezouary, A. & Baza, O., 2019. How Does Learners' Prefer to Process Information in MOOCs? A Data-driven Study.

Huang, Z., 1997. Clustering large data sets with mixed numeric and categorical values.

Humaira, H. & Rasyidah, R., 2020. *Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm.* s.l., s.n.

Hu, S., Adrian , O., Sweeney, J. & Ghahramani, M., 2021. A spatial machine learning model for analysing customers’ lapse behaviour in life insurance.

Jaccard, P., 1901. Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines..

Jain, A. K., 2010. Data clustering: 50 years beyond K-means.

Jesson, J., Matheson, L. & Lacey, F. M., 2011. *Doing your literature review: traditional and systematic techniques.* s.l.:s.n.

Joshi, P. & Jain, P., 2018. Prediction of Students Academic Performance Using K-Means and K-Medoids Unsupervised Machine Learning Clustering Technique.

KARABOGA , D., 2005. AN IDEA BASED ON HONEY BEE SWARM FOR NUMERICAL OPTIMIZATION.

Kastner, M. et al., 2012.

Ketchen, D. & Shook, C., 1996. The application of cluster analysis in strategic management research: an analysis and critique.

KR, Z., 2008. An efficient k-means clustering algorthm.

Kuo, R. et al., 2014. Automatic kernel clustering with bee colony optimization algorithm.

Landers, J. R. & Duperrouzel, B., 2019. Machine Learning Approaches to Competing in Fantasy Leagues for the NFL.

Laurens van der Maaten & Hinton, G., 2008. Visualizing Data using t-SNE.

Lee, I. & Mangalaraj, G., 2022. Big Data Analytics in Supply Chain Management: A Systematic Literature Review and Research Directions. *Big Data and Cognitive Computing.*

Levy, Y. & J. Ellis , T., 2006. A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research.

Liu, H., Fen, L., Jian, J. & Chen, L., 2017. Overlapping Community Discovery Algorithm Based on Hierarchical Agglomerative Clustering. *International Journal of Pattern Recognition and Artificial Intelligence.*

Li, Y. et al., 2021. Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. *Applied Soft Computing.*

Lumsden, S.-A., Beldona, S. & Morrison, A. M., 2008. Customer Value in an All-Inclusive Travel Vacation Club: An Application of the RFM Framework.

Madhulatha, T. S., 2012. An Overview on Clustering Methods.

Masud, M. A. et al., 2018. a new approach for identifying the number of clusters and initial cluster centres.

Maulik, U. & Saha, I., 2010 . Automatic Fuzzy Clustering Using Modified Differential Evolution for Image Classification.

Nainggolan, R., Perangin , R. a., Simarmata, E. & Tarigan, A. F., 2019. Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method.

Noordzij, M., Zoccali, C., Dekker, F. W. & Jager, K. J., 2011. Adding Up the Evidence: Systematic Reviews and Meta-Analyses.

Ostrovsky, R., Rabani, Y., Schulman, L. J. & Swamy, C., 2012. The effectiveness of lloyd-type methods for the k-means problem.

Pelleg, D. & Moore, A., 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters.

Posada, D. & Buckley, T. R., 2004. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests.

Raj, S. A. P. & Vidyaathulasiraman, 2021. Determining Optimal Number of K for e-Learning Groups Clustered using K-Medoid. *International Journal of Advanced Computer Science and Applications.*

Ramadas, M. & Abraham, A., 2019. Metaheuristics For Data Clustering And Image Segmentation.

Rodrigues, A. P. et al., 2021. Real-Time Twitter Trend Analysis Using Big Data Analytics and Machine Learning Techniques.

Rodriguez, M. Z. et al., 2019. Clustering algorithms: A comparative approach.

Rother, E. T., 2007. Systematic literature review X narrative review.

Rousseeuw, P. J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.

Seyedan, M. & Mafakheri, F., n.d. Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities.

Sharma, K. K. & Seal, A., 2021. Spectral embedded generalized mean based k-nearest neighbors clustering with S-distance.

Sharma, K. K., Seal, A., Herrera-Viedma, E. & Krejcar, O., 2021. An Enhanced Spectral Clustering Algorithm with S-Distance.

Shi, C. et al., 2021. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm.

Simon, A., Deo, M. S., Venkatesan, S. & Babu, D. R. R., 2015. An Overview of Machine Learning and its Applications. *International Journal of Electrical Sciences & Engineering (IJESE).*

Storn, R. & Price , K., 1997. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces.

Sugar, C. A. & James, G. M., 2011. Finding the Number of Clusters in a Dataset.

Swanson, D., Goel, L., Francisco, K. & Stock, J., 2017. Applying theories from other disciplines to logistics and supply chain management: a systematic literature review. *Transportation Journal.*

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S. & Satoto, B. D., 2017. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster.

Thompson, J. R. J., Feng, L., Reesor, R. M. & Grace , C., 2021. Know Your Clients' Behaviours: A Cluster Analysis of Financial Transactions.

Tibshirani, R., Walther, G. & Hastie, T., 2002. Estimating the number of clusters in a data set via the gap statistic.

Tufano, A., Accorsi, R. & Manzini, R., 2020. *Machine learning methods to improve the operations of 3PL logistics.* s.l., s.n.

Webster, J. & T. Watson, R., 2002. Analyzing the Past to Prepare for the Future: Writing a Literature Review.

Xiao, Y. & Watson, M., 2017. Guidance on Conducting a Systematic Literature Review. *Journal of Planning Education and Research.*

Xu, R. & Wunsch, D., 2005. Survey of clustering algorithms.

Yuan, C. & Yang, H., 2019. Research on K-Value Selection Method of K-Means Clustering Algorithm.

Yudhistyra, W. I., Risal, E. M., Raungratanaamporn, I.-s. & Ratanavaraha, V., 2020. Exploring Big Data Research: A Review of Published Articles from 2010 to 2018 Related to Logistics and Supply Chains.

Yu, S.-S.et al., 2018. Two improved k-means algorithms.

Zhang, Y. et al., 2018. Self-Adaptive K-Means Based on a Covering Algorithm.