Customer Segmentation using Clustering Models: A Comparison of Agglomerative and K-Means Algorithms

by

REHA HAZIR

A Master's Thesis
Submitted to the Applied Sciences Department
GISMA Business School
In Fulfillment of the Requirements
For a Master's Degree
March 2023

Abstract

The notion of comprehending customers' needs and preferences from an individualized standpoint lies at the core of effective marketing strategies. Businesses that proficiently cater to the specific needs and preferences of their customers are more likely to foster a devoted customer base and augment their revenue. As such, companies invest significant resources in identifying and analyzing customer behavior and preferences. One of the ways in which companies can gain insights into customer behavior is by leveraging machine learning techniques. These techniques facilitate the analysis of large volumes of customer data and enable the identification of patterns and trends that would be arduous to discern using traditional methods. Therefore, the aim of this report is to investigate the use of clustering, an unsupervised machine learning technique, in analyzing the preferences of customers in a grocery store setting. The focus will be on hybrid clustering essentials, which involves the segmentation of customers based on demographic factors such as age, gender, income level, education level, and behavioral factors such as purchasing behavior and frequency of response to the promotions and the campaigns. To accomplish this objective, various clustering algorithms such as partition-based and hierarchical clustering algorithms will be utilized and will be compared according to the results obtained. Each algorithm has its strengths and weaknesses, and a comparative analysis of the results obtained from each algorithm will be performed to identify the most suitable clustering method for the dataset.

Table Of Contents

1. Introduction

The technological infrastructure of today has various impacts on professional work life, including marketing departments of companies. They are now incorporating technological tools that can work in sync with traditional strategic methods and support decision-making mechanisms. Analyzing data and deriving insights about products or customers is one of the important ones. The widespread use of digital devices and the increasing use of social media have significantly increased data production in recent years, and companies have access to large datasets in the digital space. Brands like Coca Cola, Amazon, Nike, which can be referred to as industry giants, are actively working on customer segmentation methods to provide better products and advertising experiences. Evaluating these large data sets for customer and product evaluations is invaluable for every company, regardless of the industry. As customer segmentation is an increasingly popular technique used by companies of all sizes to gain a better understanding of their customers and to tailor their products, services, and marketing efforts to meet the specific needs and preferences of different customer segments: it involves dividing a company's customer base into distinct groups based on common characteristics such as demographics, psychographics, and behavioral patterns. Through the data collected from clients, any association can identify patterns and trends that may not be immediately apparent, but, by discovering the preferences and behaviors of each customer segment, it can pay off itself with effective marketing campaigns that resonate with their target audience.

In this thesis, I will be focusing on a customer segmentation project via clustering methods to segment customers based on a specific product. By segmenting customers by their demographic characteristics and gaining insight about their preferences and behaviors, organizations can initiate targeted marketing campaigns and strengthen customer satisfaction. Through this project, my task is to provision a practical of how clustering methods can be used to segment customers and robust marketing strategies. This thesis aims to contribute to the field of data science literature by making it vocational for organizations to implement cluster methods to segment their customers based on their demographic scale and behavioral preferences.

2. Literature Review

Customer Segmentation allows firms to identify and better understand their customers, which in turn can lead to more effective marketing campaigns and better customer service (Schiffman & Kanuk, 2010)[1]. For example, a company may discover that a particular segment of customers is more likely to purchase its products during a specific time of year or that they prefer a particular type of marketing message. Or a company may use different marketing channels and messages to target different customer segments, based on their preferred communication methods and interests. By understanding the needs and wants of different customer segments, companies can create customized products and services that better meet those needs, and in turn, build more loyal and satisfied customers (McDonald, 2012)[2]. This can improve customer satisfaction and loyalty, leading to increased sales and revenue. But, in today's business landscape, all these examples are only possible as businesses become more data-driven. Therefore, it is important to keep in mind that customer segmentation techniques have also evolved to keep up with the times. Traditional segmentation techniques were often based on assumptions about customer needs and preferences, resulting in a one-size-fits-all approach to marketing. Unlike the above examples, the one-size-fits-all approach assumes that all the clients of an organization have a similar intention and connection with the company. Because the product usage or the purpose of the product is the same, the intention of demands to buy that product were also seen as the same. It is more like prescribing the same medication or treatment to all patients with a particular condition, without considering individual differences such as age, gender, medical history, or lifestyle. However, according to Han and Kamber (2006)[3], with the rise of data-driven analysis, instead of having the mainstream of universally applicable for each client approach, businesses can have access to more customer data, which allows for more accurate and personalized segmentation. Data-driven customer segmentation also presents another advantage on traditional approaches which is neediness for high-quality data. The traditional approach to customer segmentation was often limited by the availability and quality of data, resulting in a less accurate and less actionable segmentation. However, with the increased availability and sophistication of data analytics tools, businesses can make more precise and effective customer segmentation, which can improve customer engagement and drive business growth.

On top of that, some research has consistently shown that data-driven segmentation is a highly effective technique for improving business performances. In a study published in the Journal of Business Research (Verhoef et al., 2009)[4], it was found that companies who implemented customer segmentation saw significant improvements in customer satisfaction, retention, and loyalty. The study also found that companies who used segmentation to tailor their marketing efforts were more likely to generate higher revenues and profits compared to those who used a more generalized approach. Data-driven customer segmentation techniques have established trust with concrete results in today's world, instead of traditional approaches, and continue to exist as an indispensable department for companies. The fact that new technological developments have filled the areas where traditional approaches have been insufficient has been seen as an important development, and data science has become one of the important areas where it studies in the literature.

2.1 Segmentation Advantages & Benefits

2.1.1 Personalized Marketing

Personalized marketing is a powerful advantage of customer segmentation, as it can significantly improve the effectiveness of marketing campaigns, increase customer engagement and loyalty, and ultimately drive business growth. It refers to the practice of tailoring marketing messages and offers to individual customers based on their characteristics, behaviors, and preferences. According to a study by Epsilon, personalized emails have an average open rate of 29.9%, compared to 18.3% for non-personalized emails (Chen & Barnes, 2007)[5]. This indicates that personalized marketing can significantly improve the effectiveness of marketing campaigns and lead to higher engagement and conversion rates. Additionally, personalized marketing has been shown to increase customer loyalty and retention. A study by Infosys found that 86% of customers are more likely to make a purchase from a company that offers personalized experiences, and 73% of customers said that they would be more likely to repeat business with a company that provides personalized shopping experiences (Verhoef, Neslin, & Vroomen, 2007)[6]. By using customer segmentation to tailor marketing messages and offers to individual customers, companies can create a more personalized and engaging customer

experience, which can lead to stronger customer relationships and increased customer lifetime value.

2.1.2 Better Customer Experience

By understanding the needs and preferences of different customer groups, companies can tailor their products, services, and marketing strategies to meet the specific needs of each segment, resulting in a more personalized and satisfying customer experience. According to a study by McKinsey (2018), companies that provide a personalized customer experience can achieve revenue increases of 5% to 10% and cost reductions of 10% to 20%. This is because personalized experiences lead to higher customer loyalty and satisfaction, which in turn drives repeat business and referrals[7]. Therefore, individualized marketing achieved through client segmentation has the potential to contribute to increased revenue, lower expenses, higher levels of customer loyalty and satisfaction, and an overall more appealing customer experience.

2.1.3 Efficient Resource Allocation & Profitability

The benefits of customer segmentation, which typically include higher profitability and more efficient resource allocation, are often emphasized. Businesses are able to tailor their marketing efforts to the specific requirements and preferences of each customer segment if they divide their client base into discrete segments based on qualities, habits, or requirements that are comparable to those of the segment. This enables firms to more effectively manage their resources by concentrating on the portions of their market that generate the highest profits and avoiding spending money on the segments of their market that generate the lowest profits. This, in turn, has the potential to boost profitability by optimizing income from each sector while simultaneously reducing expenses related to marketing.

2.2 Segmentation Disadvantages & Limitations

2.2.1 Oversimplification Issue

Oversimplification is a common problem in customer segmentation because it occurs when the segmentation model fails to capture the full complexity and diversity of customer behavior. In other words, the segmentation approach may group customers into broad or vague categories that do not accurately reflect their unique needs, preferences, or behaviors. This can result in segments that are too general or not meaningful in terms of actionable insights. For example, imagine a segmentation model that divides customers into only two segments: high-income and low-income. While this may seem like a simple and straightforward approach, it oversimplifies the complexity of customer behavior and may not provide enough information to tailor marketing strategies effectively. Customers within each segment may have vastly different needs, behaviors, and preferences that are not captured by this oversimplified model. Lemon and Verhoef (2016)[8], argue that customer segmentation can be oversimplified and fail to capture the full complexity of customer behavior. They suggest that marketers should use a more comprehensive approach that takes into account the entire customer journey, including pre-purchase, purchase, and post-purchase stages. By doing so, marketers can gain a more nuanced understanding of customer behavior and create more effective marketing strategies that resonate with individual customers. For example, two customers with similar income levels may have very different shopping habits, priorities, and preferences. Without accounting for these differences, the segmentation model may not provide enough detail to create targeted marketing strategies that resonate with individual customers.

As a result, oversimplification in customer segmentation can lead to ineffective or irrelevant marketing strategies that fail to engage customers and drive business growth. To avoid this problem, it's important to use a nuanced and comprehensive approach to segmentation that takes into account a variety of variables and criteria that accurately capture the diversity and complexity of customer behavior.

2.2.2 Data Quality and Collection

The quality of data is a critical factor in customer segmentation because segmentation models rely on accurate, complete, and up-to-date data to identify meaningful patterns and insights. If the data is incomplete, inaccurate, or biased, it can lead to segmentation results that are also incomplete, inaccurate, or biased. Being inaccurate, poor data quality results in missing or incomplete data points that can affect the validity and reliability of segmentation models. Han

and Kamber (2011)[9], highlighted that missing data can pose a significant challenge to creating effective and useful segmentation models. They noted that when some attributes are missing, the clustering results may not be reliable or meaningful, which can limit the ability to draw actionable insights from the segmentation analysis. To illustrate, they provided an example of how missing data can impact customer segmentation, emphasizing the importance of addressing missing data to ensure the accuracy and usefulness of the segmentation models. It means that if data professionals are interested in segmenting customers by their purchase behavior and some customers have missing purchase records, it may be difficult to accurately cluster them with other customers who have complete records. Another issue with data quality is the potential for bias. Biases can be introduced through a variety of factors, such as sampling methods or data collection techniques. If the data used to create segmentation models is biased, it can lead to segmentation results that are not representative of the overall population, leading to ineffective marketing strategies. Similarly, if the data collection techniques used to gather data for the segmentation analysis are biased, it can skew the results in favor of certain segments and limit the ability to draw insights that accurately represent the population as a whole. This can result in missed opportunities, as well as increased costs associated with ineffective marketing campaigns. To mitigate the risks associated with bias, it is essential to use rigorous sampling methods and data collection techniques that minimize the potential for bias to be introduced into the data.

As a result, data quality is a critical limitation for customer segmentation, and companies should prioritize data quality and integrity throughout the segmentation process to ensure that the resulting segments accurately reflect customer behavior and preferences. This can involve regular data cleaning and validation, ensuring that data sources are reliable and up-to-date, and accounting for potential biases in the data.

2.2.3 Accountability

Accountability is a concern when it comes to customer segmentation because it involves the use of customer data, which can be sensitive and personal. Organizations that use customer data for segmentation purposes are responsible for ensuring that they comply with relevant data privacy laws and regulations, such as GDPR, CCPA, or other local data protection laws. If customer data is misused or mishandled, it can lead to legal and reputational consequences,

including fines, lawsuits, and loss of customer trust. Therefore, organizations must be transparent about their data usage policies and provide customers with options to control how their data is collected, processed, and used. Also, there is a risk of bias in customer segmentation models, which can perpetuate existing inequalities and discrimination. For example, if a company's segmentation model is based on demographic or socioeconomic factors, it may result in discriminatory practices or exclusion of certain groups of customers. Therefore, organizations must ensure that their segmentation models are fair, transparent, and ethical, and they must be accountable for any negative impacts that may result from their use.

To address accountability issues in customer segmentation, organizations can implement data governance policies, establish data protection and privacy practices, and adopt ethical guidelines for data usage. They can also provide customers with clear and concise explanations of how their data is used for segmentation purposes, and give them control over their data through opt-in or opt-out mechanisms. By taking these steps, organizations can ensure that they are accountable for their use of customer data and can build trust with their customers.

2.3 Segmentation Approaches

2.3.1 Demographic Segmentation

By segmenting customers into distinct groups based on demographic factors, businesses can obtain a better understanding of the preferences and behaviors of each group. This information can then be utilized to develop targeted marketing campaigns that are more likely to resonate with each customer segment. This approach can involve the creation of personalized promotions, customized product offerings, and tailored advertising content that caters to the unique needs and preferences of each segment.

Demographic segmentation is generally the process of determining the characteristics of a person based on their demographic variables. These variables include age, gender, income, education level, whether they are parents, their occupation, and marital status. These variables are easier to measure than psychographic or behavioral variables, and can provide more accurate long-term insights compared to other models. These data can be valuable, especially

for brands in the healthcare sector and luxury product sales, as they are difficult to change or change slowly over time, making them useful for a long time if the dataset is kept up-to-date. Of course, these statements are only accurate if the datasets remain current.

The development of demographic segmentation dates back to the mid-20th century. During the period when television commercials began to trend, businesses realized the advantages of creating specific target customer groups rather than relying on one-size-fits-all approaches, as it allowed advertisers to reach large audiences with targeted messages based on their viewing habits and demographic characteristics. They discovered that customers did not have similar preferences and needs, but rather, certain groups had specific needs.

The idea that demographic characteristics could be used as a marketing strategy actually began after World War II, when a neo-liberal structure began to dominate the world. Companies used statistical analysis tools to segment customers and measure their behaviors. Throughout time, demographic information has evolved from simple data sets to more complex structures that offer more precise insights for businesses. It is now possible to obtain information on a person's behavior, lifestyle, and habits based on their geographic location. Demographic data sets can also be used to understand market trends. For instance, information about what products high-consumption individuals purchase can provide valuable insights depending on the time period.

2.3.2 Behavioral Segmentation

Behavioral segmentation is similar to demographic segmentation in that it involves dividing customers into specific target groups based on certain characteristics. However, different variables can be used in behavioral segmentation. For example, it may include information on a customer's brand loyalty, purchase frequency, and the benefits they derive from a product. Behavioral segmentation assumes that consumers with similar behaviors have similar needs and preferences, and therefore, respond similarly to marketing messages and product offerings. It is an important approach in understanding buying behavior.
Also, the feature pool is very large in behavioral segmentation. The development of various digital channels such as social media, Point-of-Sales etc has significantly affected the behavioral segmentation position, which enabled the marketers to reach their customers

through multiple channels. Thus, as these data become more complex, the demand for marketers and analysts who are experts in their fields has increased in the right direction. For example, a technical study published in the International Journal of Computer Science & Information Technology conducted customer segmentation behaviorally through a point of sales dataset, and the accuracy and validation of the results were discussed and assessed by marketing experts.[29]

Today, the most researched method in the field of behavioral segmentation is RFM. Due to the fact that it enables businesses to divide their consumer base depending on their transactional behavior, RFM is a widely utilized strategy in behavioral segmentation. Because RFM can give accurate and fast recency, frequency and monetary information about a customer. Along with these customer metrics, it tries to have an idea about the consumer's orientation by determining how the customer responds to the campaigns, how often they spend and how much they spend.[30]

Behavioral segmentation has a relatively shorter history compared to demographic segmentation. It emerged in the late 20th century as businesses began to recognize that consumer behavior and preferences could be used as a basis for market segmentation. The idea of using consumer behavior as a basis for segmentation gained popularity in the 1960s and 1970s, as market research became more sophisticated and businesses sought to gain a better understanding of their customers. This led to the development of new techniques for measuring consumer behavior, such as surveys, focus groups, and observational studies.

Today, behavioral segmentation is a widely used approach to market segmentation, particularly in industries such as retail, e-commerce, and consumer goods. With the rise of big data and machine learning, businesses are able to analyze vast amounts of data on consumer behavior and use it to create highly personalized marketing campaigns and product offerings.

Behavioral Segmentation	**Demographic Segmentation**
Based on behavior of customers	Based on demographic characteristics

better indicator about customer’s actual action	better indicator for general customer”s recognition
Contastanly changing over time	More sustainable over time
applied any product or services	Not applied for all products or services

Figure 1: Differences between Behavioral and Demographic Segmentation

2.3.3 Hybrid Segmentation

Hybrid segmentation can be used to combine multiple data sets for the same purpose when conducting segmentation analysis. For the continuation of this report, a hybrid segmentation analysis will be attempted by combining behavioral and demographic data sets, in order to create an infrastructure where behavioral measurements can be clustered alongside demographic information to gain a more comprehensive understanding of the customer.

A supervised model can be used to estimate the target group for customers who have been grouped through demographic analysis before being included in these tactics in order to enhance the segmentation experience and improve the efficacy of targeted marketing strategies. This can be done in order to improve the effectiveness of targeted marketing strategies. However, an important question is “how distinctly the customers in the cluster differ from other customers”. Before measuring the distinction between the clusters with technical indicators, it is sometimes insufficient to rely solely on demographic data, which is why it is beneficial to include data produced through the customer's interactions with the company's products or services. This will not only provide more features, but also provide a more diverse data set, resulting in more successful segmentation with various insights. It means, customer groups identified through demographic information are established to better understand the customer. However, when the customer is examined through their behavioral movements in the same group as other different behavioral individuals, it can provide a more intense and different perspective, resulting in a stronger outcome. Therefore this study will use a dataset

that combines demographic and behavioral data sets to more accurately define customer groups and evaluate target customer groups with more detailed and diverse arguments.

2.4 Clustering Techniques

2.4.1 Partition-based Clustering

Clustering algorithms generate clusters having similarity between data objects based on characteristics belonging to the same cluster. Due to its ease of implementation and its ability to process high-dimensional datasets with relatively small memory requirements, partition-based clustering is commonly used in projects such as image segmentation, handwriting recognition, customer segmentation, and anomaly detection.[10] The algorithm aims to use a partition based approach which splits a dataset into k clusters, where each observation belongs to the cluster with the nearest mean (centroid).The k-means algorithm works by first randomly selecting k initial centroids, and then iteratively assigning each observation to the nearest centroid and recomputing the centroids based on the new assignments. The process continues until the centroids no longer move or a maximum number of iterations is reached. Although the operations involved in partition-based clustering are easily understandable, there are two significant disadvantages that arise when implementing this technique: the challenge of determining the appropriate number of clusters and selecting initial centroids, and the ability to handle different types of data and frailty on spherical clusters.[11]

During the model implementation stage, it is assumed that every column responsible for carrying out the operations consists of numerical data. Therefore, categorical features need to be preprocessed and encoded before being fed into the model. This presents a disadvantage as analyzing datasets with text data, which are difficult to encode, can be challenging. Another important issue is the determination of the number of clusters during the application of the model. In unsupervised models, since there is no target column, it may not be possible to know or deduce the number of clusters that the relevant dataset needs to be divided into with domain knowledge. This becomes a problem because the value of the number of clusters directly affects all calculations performed, and an incorrect number of clusters may not accurately reflect the results.

The other well-known disadvantages of partition-based clustering are more reliable while identifying spherical clusters with equal variances compared to other clustering methods. According to a study published in the "Journal of Computer Applications" in 2008, all compared partition-based clustering algorithms were found to be more successful in detecting spherical-shaped clusters in small and medium-sized databases.[12] Although it may appear unsuitable for real-world scenarios, partition-based clustering is still widely used in industries such as data mining due to its simplicity, scalability, and effectiveness. And
The most suitable algorithm in which these three advantages are strengthened is the K-means algorithm. For instance, one of the other partition-based algorithms called CLARA (Clustering Large Applications) creates multiple sub-samples from the original dataset to speed up the computation. However, this can lead to a loss of information and potentially affect the accuracy of the clustering results. Another example might be K-Medoid which is more costly than K-Means algorithm because of its complexity.[13] But, also K-means holds some disadvantages. For example, if the data structure may not have a linear structure, clusters may have unequal variances, and the distortion score may be high which means that K-means clustering might yield subpar results if the data contains outliers or noise. Therefore, an attempt will be made to address this issue by balancing the relevant data using Feature Engineering and encoding techniques within the K-Means algorithm. These details will be demonstrated in detail under the Feature Engineering and Label Encoding sections.

2.4.2 Hierarchical Clustering

Hierarchical clustering is a clustering algorithm used to group similar data points together based on their distance or similarity. The similarity measure is typically based on a distance metric such as Euclidean distance or Manhattan distance. It starts with each data point as a separate cluster and then recursively merges clusters until only one cluster is left. Agglomerative clustering is a type of hierarchical clustering, where the algorithm starts by considering each point as its own cluster and iteratively merges the two closest clusters until only one cluster remains. This approach is also known as a bottom-up approach. In agglomerative clustering, the distance between two clusters is typically defined as the distance between the closest pair of points, one from each cluster. There are different ways to measure

the distance between clusters, such as single linkage, complete linkage, and average linkage which will be discussed under the model implementation section.

The process of agglomerative clustering results in a dendrogram, which is a tree-like diagram that illustrates the sequence of merges between clusters. This dendrogram can be cut at different levels to obtain different numbers of clusters. The choice of the number of clusters to select depends on the problem and can be informed by domain knowledge or by visual inspection of the dendrogram. The choice of linkage parameter has a crucial role to find suitable dendrogram on the study. The dendrogram, for instance, will display long vertical branches with shorter horizontal branches that indicate the distance between clusters if the complete linkage approach is applied. The dendrogram will display a succession of small vertical branches with larger horizontal branches that represent the distance between clusters if the single linkage approach is applied. There might occur false interpretations in the results.

There is also one more clustering technique under Hierarchical clustering called as divisive clustering. Unlike agglomerative clustering, it starts with all data points in a single cluster and then splits the clusters iteratively until a stopping criterion is met. It is a top-down approach that also results in a dendrogram, but the height of the branch represents the degree of dissimilarity within a cluster. In terms of computational complexity, agglomerative clustering is also more efficient than divisive clustering because it requires fewer distance calculations.[14] However, divisive clustering may be more effective when the data set is very large or when there is a large variation in cluster sizes.

Additionally, divisive clustering may be preferred in cases where the desired number of clusters is known or when the data has a tree-like structure that lends itself to a top-down approach. Ultimately, the selection of a clustering method should be guided by the specific goals and requirements of the analysis, as well as the characteristics of the data set itself.

2.5 Evaluation Metrics

2.5.1 Silhouette Score

The Silhouette score is a statistical metric that measures how similar an object is to its group and compares this value to other clusters. The values range from -1 to 1. A positive Silhouette score indicates that the observation matches well with its own cluster and poorly with neighboring clusters. Conversely, a low Silhouette score indicates that the corresponding observation matches poorly with its own cluster and better with neighboring clusters.

Peter J. Rousseeuw introduced the Silhouette score in 1987, and it was named after the silhouettes of the clusters. Rousseeuw stated that the Silhouette score is a more informative measure of cluster quality than other metrics, as it can measure both the cohesion and separation of the clusters.[15] However, the Silhouette calculations used today are slightly different from Rousseeuw's original method. This is due to the emergence of different datasets as a result of the development of today's technologies.

Initially, The most significant criticism of the Silhouette score was that it does not provide accurate results for clusters with uneven sizes.[16] As a result, modified Silhouette scores were introduced, such as the Weighted Silhouette Score and the Adjusted Silhouette Score. Also, the traditional silhouette score calculations used the Euclidean distance metric. However, it was later discovered that other known metrics, such as Manhattan distance, may be more consistent with the nature of the clustered dataset.[17] Similarly, Kamber and Pei (2011) discuss that the use of cosine similarity calculations is more appropriate than Euclidean distance, especially for text datasets.[18]

According to traditional calculations, it was assumed that the generated clusters have roughly spherical shapes. However, to solve clustering problems that contain irregular cluster shapes, the Density-Based Silhouette Score has been proposed, which takes into account the density of points within a cluster. Lastly, the subject that contributes to the development of the silhouette score is its visualization. From the first day it was announced to this day, the silhouette score uses its own name heatmaps and plots today.

2.5.2 Calinski-Harabasz Index

The Calinski-Harabasz Index, which is often referred to as the Variance Ratio Criterion, is a criterion for measuring the quality of clustering that is utilized in the process of determining how successful a clustering method is. It is calculated using the ratio of the variation across clusters to the variance within clusters for a given clustering solution. That a successful clustering solution should have clusters that are well-separated from each other (that is, have a high between-cluster variance) and closely clustered within themselves (that is, have a low intra-cluster variance) is the concept that lies behind the Calinski-Harabasz Index. A high CHI value signifies that there is a better separation between the clusters since it indicates that there is a high variance between the clusters in comparison to the variance that exists within the clusters.

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1}$$

Figure 2: Calinski-Harabasz Index Formula

where $tr(B_k)$ represents the trace between the group dispersion matrix and $tr(W_k)$ represents the trace of the within-cluster dispersion matrix defined by:

$$W_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T$$

$$B_k = \sum_{q=1}^{k} \sum_{z \epsilon C_q} (x - c_q)(x - c_q)^T$$

E : set of data

s : CHI value

n_E: total number of observations

k : the total number of clusters

c_q : center of the cluster

c_E: center of E

n_q : number of points in the cluster q

Figure 3: Group Dispersion and Within-Cluster Dispersion Formulas

There is also another measure called the “Davies-Bouldin Index”. The Calinski-Harabasz Index (CHI) and the Davies-Bouldin Index (DBI) are both measures of clustering quality that can be used to evaluate the effectiveness of a clustering algorithm. However, they differ in terms of the type of information they use to evaluate the clustering solution. For instance, the CHI is calculated by dividing the between-cluster variance by the within-cluster variance. The underlying assumption is that a good clustering solution should include clusters that are well-separated from one another (i.e., have a high between-cluster variance) and closely grouped within themselves (i.e., have a low intra-cluster variance).[19] The CHI evaluates the ratio of these two variances to determine how effectively the clusters are separated. The DBI, on the other hand, is calculated using the average similarity of each cluster to its most similar cluster as well as the average size of each cluster. The idea is that a successful clustering solution should contain clusters that are both internally compact and well-separated from one another. The DBI computes the ratio of average dissimilarity across clusters to maximum intra-cluster dissimilarity, weighted by cluster size. Therefore, a low DBI score suggests that the clusters are well-separated and compact internally.

3. Methodology

There are various ways to segment customers, such as needs-based, psychographic, and value-based approaches. However, for the purposes of this report and the available dataset, It will be focused on exploring the details of demographic and behavioral segmentation methods.

Customer segmentation cannot be a one-time exercise that is done and then used indefinitely. Customer preferences and behaviors can change over time, and therefore, customer segmentation should be continuously reviewed and updated to reflect any changes in the market or customer base. However, demographic segmentation, as opposed to psychographic and behavioral segmentation, is relatively stable over time and may not have significant changes that can deeply affect its analysis. Therefore, demographic approaches can be used to support long-term supervised models. Companies can evaluate their customers by comparing them using demographic metrics to determine their market strategies and develop models accordingly. Therefore, the use of demographic approaches is essential for building a data pipeline that companies can use in the long term.

In this study, quantitative research will be conducted with a grocery setting dataset for analysis of demographic information, product interactions, response to promotions, and location of customers over the course of a year. The study aims to compare the results of agglomerative and k-means algorithms that were chosen for customer segmentation, to determine which algorithm plays a more prominent role in hybrid segmented analysis. The roadmap for this study includes data preparation, data preprocessing, model implementation and analyzing the results.The first main duty on this quantitative research is to apply a complete four-stage data pre-processing procedure to transform the data into a suitable form for the model. During this process, important parameters for the data pipeline such as number of clusters (k) and the components of PCA(Principal Component Analysis) will be reviewed to ensure the success of the analysis. Finally, the resulting segments and the technical abilities of the two models will be compared.

3.2 Data Preparation

First step is to include the columns that have been determined based on the results of my data analysis in a clear and concise manner. During this process, if there is a possibility that two columns have a correlation based on domain knowledge, I prefer to either create a new column or transform an existing column to reach the necessary conclusion.

When working with data, it's important to make sure that the variables you're using are well-defined and meaningful. In the case of the ‘Year_Birth’ column, the values represented the year in which the customer was born. While this information is useful, it can also be a bit cumbersome to work with. For example, if you have customers born in the years 1950, 1960, and 1970, the values in the column will be 1950, 1960, and 1970 respectively. These numbers aren't standardized in any way, which can make it difficult to compare them directly. To address this issue, the column was renamed to ‘Age’. This makes the values in the column smaller and more standardized. For example, if you have customers who are 70, 60, and 50 years old, the values in the column will be 70, 60, and 50 respectively. This standardization makes it easier to compare the values and analyze them in a meaningful way. In addition to the benefits of standardization, renaming the column to ‘Age’ also makes it easier to understand the data. Age is a concept that most people are familiar with, and it's more intuitive than birth year. When analyzing the outputs of a model or presenting the data to others, using age as a variable is likely to be more accessible and easier to understand.

Another important aspect of the dataset is the ability to create new features based on the existing ones. For instance, when examining what products our customers purchase and how they make their payments, a new feature may be needed to see their overall purchasing power. In this case, a new column called ‘TotalSpent’ can be created to display the total amount spent by each customer. This new column can then be placed below the existing columns, and all the observations can be sorted accordingly. By creating this new column, a more complex information schema can be simplified, and all observations can be assigned different numerical values with different weights by simply adding them up. This can provide a better understanding of our customers' purchasing power and enable more informed decision-making.

Finally, it should be noted that some columns in the dataset may contain unnecessary information and could be removed. One of the potential issues that may arise from having too many uncorrelated data columns is the problem of overfitting, which could result in the accuracy of the resulting model being compromised. Another reason for removing certain

columns is that the necessary information may have already been extracted from some columns and a new column has been created as a result. Therefore, the old column should be deleted to avoid any potential issues with high correlation among columns that could affect the accuracy of the model. As a result, the columns ‘Marital_Status’, ‘Dt_Customer’, ‘Z_CostContact’, ‘Z_Revenue’, ‘Year_Birth’, and ‘ID’ should be removed. It should be noted that the primary objective of our report was to measure how customers respond to promotions, and the information contained in these columns does not seem to be relevant to the outcome of our project. Of course, it is not always necessary to delete entire columns, and in some cases, if the information within the dataset is of little support to the research's effectiveness, it should not be completely removed but rather simplified. The column that indicates marital status in the dataset has a classification feature, but it has been unnecessarily detailed for the purpose of the report. While the analysis process of the project only requires whether the customer lives alone or with someone, the column contains very detailed classifications such as ‘Married’, ‘Together’, ‘Absurd’, ‘Widow’, ‘YOLO’, ‘Divorced’, and ‘Single’. This column can be simplified by grouping them under two headings, ‘Partner’ and ‘Alone’. The same issue is present in the Education feature as well. Some classification values are meaningless, so I simplify them to provide a clearer form and remove columns from the dataset that do not have the potential to add value or have been modified.

After the columns in our dataset have been organized and simplified for better understanding and to collect all necessary information properly, the preprocessing stage can now be entered. Here, technical issues that may arise in our dataset can be addressed and worked towards being resolved. It is critical in this stage to ensure that the dataset is suitable for analysis and that any issues that may affect the accuracy or reliability of the results are identified and resolved.

3.3 Data Preprocessing

3.3.1 Outliers

In general, all datasets have a clean set of data. However, there were anomalies within some columns that could affect the performance of the model. For example, there are 3 observations above the age column of 120 years old. When these observations were examined in detail, it was found that one of them also showed an anomaly in the income column. Besides being

entered incorrectly, what makes them anomalies is that they are very far away from other data points in terms of distance. The fourth closest data point to the ages of 120, 121, and 128 is 81 years old. Therefore, this could create a problem for the model accuracy.

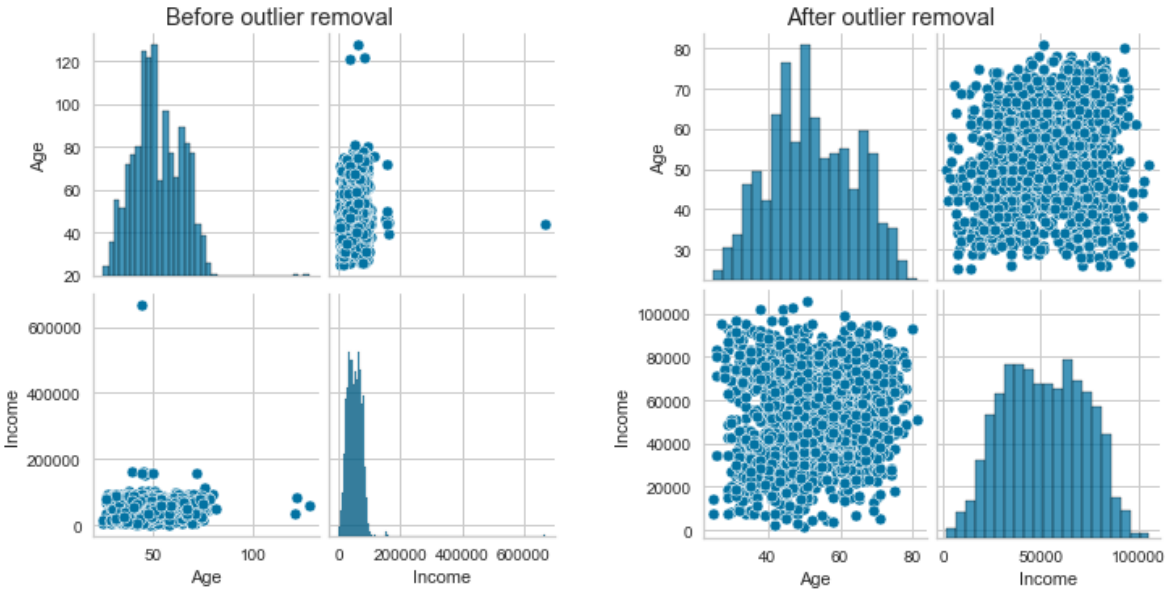


Figure 4: Before and After Removing Outliers

I set the limit to 100 for age and the threshold value to 110000 for income. According to the results above, all observations were noticeably spread out over the table after setting these limits.

3.3.2 Label Encoding

In the data preprocessing stage, I transformed the categorical values of the data into numerical columns by using the encoding technique and assigning them a certain value. The main reason for doing this is that every data type that goes into PCA must be numerical. Therefore, the transformation was necessary to ensure that the data was suitable for PCA.

Label Encoding is used to convert categorical variables into numerical variables. This technique involves assigning each unique category within a categorical variable with a numerical label, usually starting from 0 up to the number of distinct categories minus 1.

In our “Education” feature, I have three different values: Undergraduate, Graduate, and Postgraduate. In addition to converting this column from an object column to a numerical column, there are also weights associated with each value. As the level of education obtained follows the order of Postgraduate > Graduate > Undergraduate. However, the encoding technique assigns values based on the order in the column obtained. For instance, the value of the first observation is ‘Graduate’, which takes the value of ‘0’. Nevertheless, if we were to sort them according to their weights, the ‘Graduate’ value should have taken the value of ‘1’. When requesting the mapping of values in a standard LabelEncoder() application, the obtained result is as follows..

```
LE = LabelEncoder()
LE.fit(df['Education'])
mapping = dict(zip(LE.classes_, LE.transform(LE.classes_)))
mapping
Output:
{'Graduate': 0, 'Postgraduate': 1, 'Undergraduate': 2}
```

Figure 5: LabelEncoder Codeline

According to the order of the ‘Education’ column, the second value that comes after the ‘Graduate’ value is ‘Postgraduate’. Therefore it takes the value of ‘1’, while the “Undergraduate” value takes the value of ‘3’. Because it is the last one in the order. It is known that LabelEncoder assigns values from the top of the column downwards, so a custom encoding will be created.

It should be noted that encoding may present certain limitations in cases where there are no weight differences between data points. While label encoding allows for the conversion of data into a machine-readable format, it assigns a unique numerical label, starting from ‘0’, to each

class of data. This can result in priority issues during the training of datasets, as a label with a higher numerical value may be perceived to have a greater priority than a label with a lower value.

Let's assume that there is a column in the dataset that contains the countries of customers, such as France, Germany, USA, etc. According to label encoding, the first value in the column (France) would be assigned a numerical label of ‘0’, followed by Germany with ‘1’, and USA with ‘2’.

When there are no intrinsic weight differences between categorical values, encoding techniques such as label encoding can potentially distort the data when used in machine learning models. The problem stems from the fact that numerical labels assigned to each category may inadvertently imply a ranking or priority among them. For example, since 2 > 1 mathematically, a model may falsely infer that a category assigned a higher numerical label is more important, even if there is no inherent ordering among the categories.

3.3.3 Standardization

The other important step in the data preprocessing stage is to standardize the dataset. The main idea of standardization is to center the data at a mean of ‘0’ and calculate the standard deviation as ‘1’. This step ensures that all variables contribute equally to the PCA (Principal Component Analysis) and helps to avoid the dominance of certain variables.

The main purpose of PCA (Principal Component Analysis) is to find new uncorrelated data sets within our dataset and to describe all variances from the widest perspective. If the data has different scales and ranges, larger numbers can dominate the statistical analysis. As a result, incorrect and biased results can occur.

To avoid this, the numeric columns will be scaled to put them all in the same range. For example, the column named “Education” column which is encoded and it has integer values between ‘1’ and ‘3’. However, at the same time, the “TotalSpent” column contains values between ‘5’ and ‘2525’. If this dataset did not be scaled before putting it into PCA (Principal Component Analysis), the “TotalSpent” column will carry more variance and have a greater

impact on the PCA (Principal Component Analysis) components than the “Education” column. As a result, the PCA components may vary more with respect to the “TotalSpent” column, leading to unwanted effects on the results. Therefore, the scales of the variables in the dataset must be standardized. For example, while the “TotalSpent” column contains values between ‘5’ and ‘2525’, the “Education” column only has integer values between ‘1’ and ‘3’. Thus, the “TotalSpent” column can be transformed into a scale between ‘0’ and ‘1’. This way, both columns will be on the same scale and the PCA (Principal Component Analysis) results will be more consistent. The calculation of the standardization score, also known as the z-score, is a value obtained by subtracting the mean of a variable from each data point and dividing it by the standard deviation. The equation for this can be seen in the following figure.

$$z = \frac{X - \mu_p}{S_p}$$

X: variable ,
μ: population mean,
s: sample standard deviation,
n: instance (rows),
p: features (columns),

Figure 6: Standardization equation

Assuming that the data values in the “TotalSpent” column is represented as $x_1, x_2, \ldots, x_n$, then x_1= 1617. By taking the sum of all the values in the column and dividing it by the total number of values (denoted as N) $\mu_p = \frac{\Sigma x}{N}$ obtained the average value of the “TotalSpent” column, μ_{25}, which is approximately equal to 607.075. Then, by subtracting the average value μ_{25} from x_1, got a result of 1009.92. Finally, by dividing this result by the standard deviation,

$\sqrt{\frac{\Sigma(X-\bar{x})^2}{n-1}}$, z score will be equal to 1.679323. This process scales the "TotalSpent" column, which originally contained values between 5 and 2525, to a range between approximately -1 and 3.18, preserving the structure of the data while standardizing it.

3.3.4 Dimensionality Reduction

The term "dimensionality" refers to the number of features or variables within a dataset. A dataset containing a significantly higher number of variables than observations is known as a "high dimensional dataset". Analyzing such datasets presents challenges due to the "*curse of dimensionality*", a phenomenon arising from an excessive number of variables relative to the sample size, leading to sparse data and numerous possible feature combinations, making it difficult to identify significant patterns. Moreover, such datasets can produce computational complexity, significant storage requirements, and the potential for misinterpreting results. It addresses these challenges by transforming high-dimensional data into a lower-dimensional representation that retains most of the relevant information, while reducing redundancy and noise. This can lead to more efficient processing, easier visualization, and improved predictive performance in downstream modeling tasks.

The use of PCA in this project was generally designed according to the advantages and disadvantages of the model. First of all, generally in every type of project dimensionality reduction reduces time and the storage required. This provides compact working flow and streamline accessibility to results. Also there is a linear relationship between two or more independent variables, which can cause problems in the analysis. But with implementation of Dimensionality reduction, interpretation of the parameters of the clustering techniques will be improved because of multicollinearity and provide more reliable results[20]. The "Wines" and "TotalSpent" columns within the dataset are two columns that feed each other during the feature engineering process. They have the highest correlation of '0.9' in the correlation graph (Appendix B). By applying dimensionality reduction, this issue is addressed and resolved.

According to Bellman's book, Adaptive Control Processes: A Guided Tour, there is a term called "curse of dimensionality" which refers to the difficulties and limitations encountered when analyzing and interpreting high-dimensional data.[21] As the number of dimensions in the

data increases, the volume of the data space grows exponentially, leading to a sparsity problem and making it difficult to obtain meaningful insights from the data. To solve this problem Bellman mentions that one way to reduce the curse of dimensionality is to make certain that the number of independent variables required to describe the problem is as small as possible.[22] Which it produces that the answer could be dimensionality reduction. Similarly, although there are '29' different columns in this dataset, only '10' of them have a high correlation issue. In fact, the result analysis will be carried out only on two features. Therefore, after deleting some of the columns in the dataset, the remaining ones will be used to apply dimensionality reduction, resulting in the highest quality variances. After these procedures, it is important to decide on the dimension of the data. Because the optimal number of components is the minimum number required to capture most of the variation in the data for the effective analysis. Finding the optimal number of principal components, or "n_components", there is a method called "Scree plot" which will be implemented to find the perfect number of the PCA components.

a.Scree plot

Although scree plots were originally used in factor analysis(Cattell, 1966), they have since been adopted for use in principal component analysis (PCA) as well. In fact, scree plots are now one of the most commonly used methods for determining the number of principal components to retain in a PCA. It is a graphical representation of the eigenvalues associated with each principal component in a principal component analysis (PCA). The scree plot displays the eigenvalues of each principal component in descending order, with the number of principal components plotted on the x-axis and the corresponding eigenvalues plotted on the y-axis.

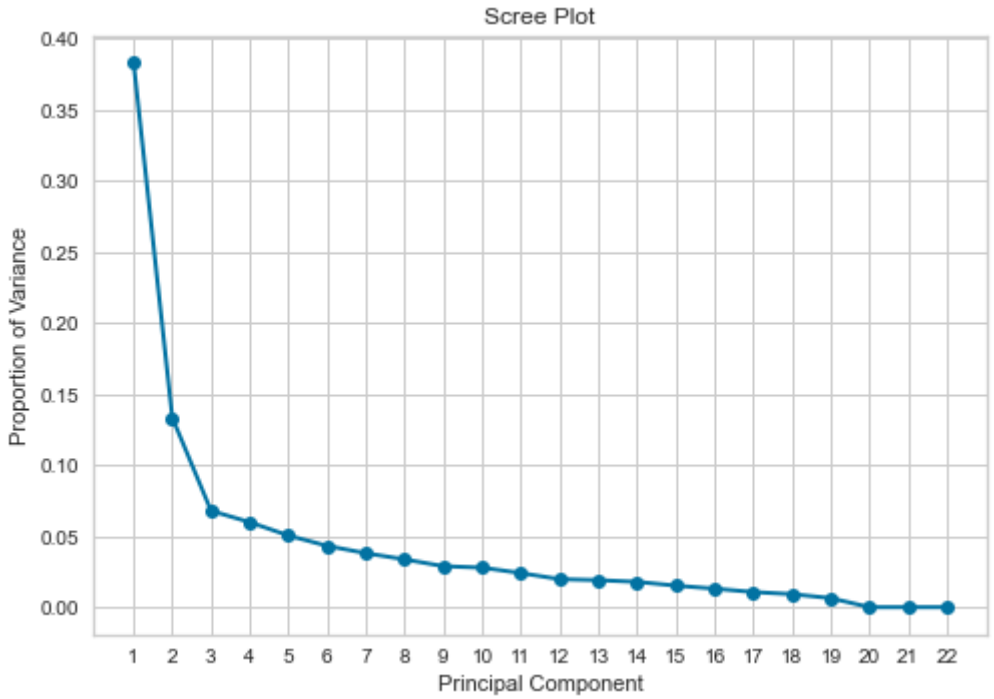


Figure 7: Scree Plot

The number of principal components to retain is often chosen at the point where the steep drop in eigenvalues levels off, which is known as the “elbow point”. The rationale behind this is that the first few principal components capture the majority of the variation in the data, while the later ones capture only a small amount of additional variation. Therefore, according to Figure.11 the best suitable :n_components’ indicate the number of three. It gave enough information by itself to use PCA with the ‘3’ component. Because “elbow point” is obvious, not in smooth or uneven shapes. Also Cumulative Explained Variance plot might be used. But in such a high dimensional dataset, this graph for more to solve the high collinearity issue.[23]

Depending on the analysis's context, changes in silhouette scores between the PCA and before and after can signify different things. In some circumstances, it can mean that PCA has enhanced clustering performance by reducing the dimensionality of the data while maintaining the clustering structure. This could be viewed as a successful outcome.

However, in other circumstances, the variations in silhouette scores might be a sign that PCA corrupted the data's clustering structure, leading to subpar clustering performance. This could be viewed as a poor result. Because of this, it's crucial to carefully examine the variations in silhouette scores before and after PCA and interpret them in light of the particular analysis.

3.4 Model Implementation

3.4.1 K-means Clustering

The most crucial step in K-means clustering is choosing the number of clusters with "n_clusters" parameter. There are numerous ways that can be used for this, but other parameters must be chosen before using them. Therefore, a loop employing the "GridSearchCV" technique will be executed to determine the ideal K-means algorithm settings. The crucial detail in this case is that the code cell was executed without the "n_clusters" argument being one of the inputs. The fact that the silhouette score will be used as the primary score because it is an unsupervised model which has no target feature to pick best parameters. Thereby, the parameters with the highest silhouette score will be chosen as the outcome.

```
param_grid = {
    'init': ['k-means++', 'random'],
    'n_init': ['auto', 1, 5, 10],
    'max_iter': [i for i in range(1, 20, 5)],
    'algorithm':['lloyd', 'elkan', 'auto']
}
output:
Best parameters:  {'algorithm': 'lloyd', 'init': 'k-means++', 'max_iter':
```

```
1, 'n_init': 'auto'}
```

Figure 8: Parameters for “GridSearchCV”

The selected parameters of the K-means algorithm for “GridSearchCv” are the initial positions of the centroids (init), the maximum number of iterations (max_iter), the algorithm to use for computing (algorithm), the number of times that the algorithm will be run (n_init).

“init” is used to determine the initial positions of the centroids. Initialization methods have three basic perspectives. First of them, which is used as “k-means++”, initializes the centroids by first selecting one centroid at random from the dataset. After that, it selects subsequent centroids from the remaining data points using a probability that is proportional to the squared distance from the current centroid that is the closest to the initial selection. This approach not only improves the quality of the clustering result but also helps to ensure that the centroids are kept sufficiently apart from one another. Second of them, which is used as “random”, it simply randomly selects “k” data points as the initial centroids. Because of k-means++ initialization is an improved method of selecting initial centroids that addresses the issue of poor initialization, random initialization would not be the best parameter for this report.

“max-iter” is a parameter that specifies the maximum number of iterations that the algorithm will run before terminating, even if convergence has not been reached. The key point of this parameter is that the algorithm will finish, at which point it will return the best clustering result it has discovered up to this point. However, it is essential to keep in mind that an excessively high value for the max_iterand parameter will result in an overfitting of the data. Therefore, the loop range will be ‘1’ to ‘20’ preventing the overfitting issue. One of the other parameters is known as the "algorithm" parameter that is used to determine the centroids of the clusters. There are several popular variations of the K-means algorithm, including the “Lloyd” and “Elkan” options, in addition to the default settings. It is possible for the speed of the K-means clustering as well as the amount of memory that it uses to be affected by the algorithm that is used, and other algorithms may be more suitable for various kinds of datasets. By calculating the mean of the points in each cluster, Lloyd's algorithm updates the cluster centroids. The

Elkan algorithm, on the other hand, establishes a lower constraint on the separation between a point and a centroid by means of a triangle inequality. The algorithm is made more efficient by using this lower bound to stop conducting pointless distance calculations during the algorithm's assignment phase. Although the Elkan approach may not converge for some datasets, it is typically faster than Lloyd's algorithm, especially for low-dimensional data.[24]

Lastly, if we discuss the n_init parameter, it controls the number of centroids that are randomly initialized before running the k-means algorithm. Therefore it very much depends on the value of init, especially when 'n_init' = 'auto', the number of runs depends on the value of init: '10' if using 'init' = 'random', '1' if using 'init' = 'k-means++'.[25] The likelihood of discovering the global minimum of inertia increases with increasing "n_init" value, but the method will run more slowly.

Number of Clusters (k)

The most crucial step in conducting a good analysis is figuring out how many clusters will be used to separate the consumers after the primary parameters have been established. In order to accomplish this, the well-known elbow technique will be used first, and the data obtained in various scenarios (with various numbers of clusters) will then be examined based on this strategy in order to get the most suitable outcome.

a. Elbow Method:

Including the calculation of Euclidean distance here is important because we are generally calculating the homogeneity of the data, and this is the underlying main concept behind the elbow method. It is used to measure the distance between two data points. The number of dimensions we consider depends on the dataset we are working with. For example, in this report, we have a three-dimensional dataset.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Figure 9: 3-Dimensional Euclidean Distance Formula

While considering the report's dataset, which comprises 2020 observations in a three-dimensional space. First random observation in our dataset be denoted by $(x_1, y_1, z_1) = (5.03, -0.174, 2.67)$ and the values for the second random observation are $(x_2, y_2, z_2) = (-2.88, 0.0168, -1.88)$. To calculate the Euclidean distance, we need to find the differences between the x, y, and z coordinates of the two points. For example, the difference between the x-coordinates is $(x_2 - x_1) = -7.92$, the difference between the y-coordinates is $(y_2 - y_1) = 0.190$, and the difference between the z-coordinates is $(z_2 - z_1) = -4.55$. Next, we square each of these differences (SSD), and add them together: $d = \sqrt{(-7.920812)^2 + (0.190866)^2 + (-4.552321)^2}$. Finally, we take the square root of the sum to get the Euclidean distance, $\sqrt{83.49}$, which is approximately 9.13.

Of course, these results include the results of two random observations. The important point here is that when calculating WCSS, one of the observations should contain the data of the centroid of the relevant cluster. Thus, by measuring the distance of each data point to the center of the cluster it belongs to, the homogeneity of each cluster is measured. Therefore, a lower WCSS score indicates a more homogeneous clustering. So, in order to compute the WCSS score, it is necessary to have an already clustered dataset. However, assuming that a dataset in which we don't know how many clusters it will be allocated, is already clustered, may seem odd. Therefore, the below-mentioned approach will be followed: The relevant clustering algorithm (K-Means) for '1' to '20' different clusterings will be iterated and the results of WCSS score for each clustering will be added to a list. Then, we will display this data in a table. Therefore each observations' distance will be calculated in '20' different scenarios.

As a first step, the K-Means algorithm assigning each data point will be runned to its closest centroid. By iterating K-Means algorithm for '1' to '20' clusterings, I had the centroids for each cluster by taking the mean of all the data points assigned to that cluster for '20' different

situations. To get those calculations basically the attribute of the KMeans clustering model used in Python which is “kmeans.cluster_centers_”.

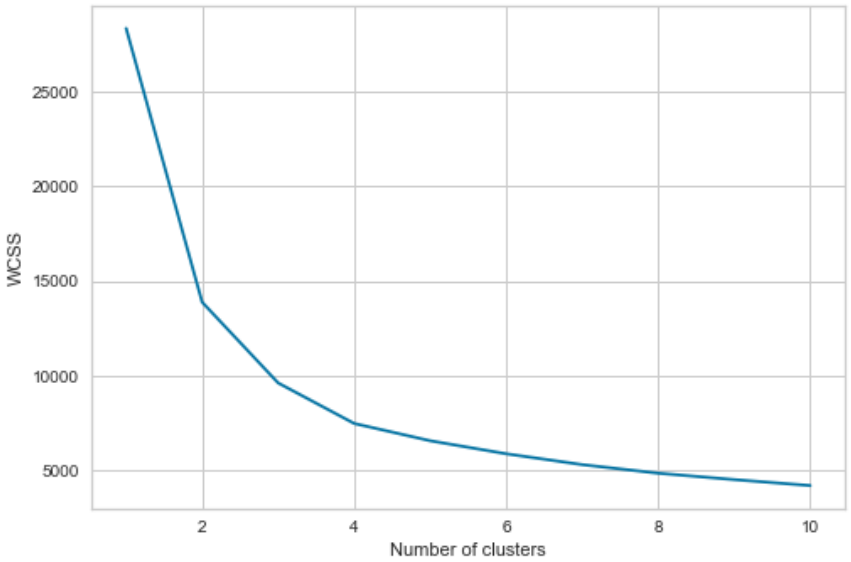


Figure.10: WCSS score for K-Means Clustering

Looking at the above table, we can see the rough version of the elbow graph. The values start to plateau after (k) = ‘2’, but they reach their lowest scores after (k) equals to ‘4’. After that, all the obtained WCSS scores carry almost the same level of homogeneity, so the most valuable clustering value that can be taken is between ‘2’ to ‘4’. Like the above statement explained, the elbow technique is a heuristic for figuring out what amount of clusters are suitable for a particular dataset. The task was to determine the k-value at which the rate of SSD loss begins to level off, generating a “elbow” in the plot. Normally, the number of clusters (k) is increased, which generally results in a decrease in SSDs since a larger number of centers can be used to cluster the data. However, in our dataset as (k) is further increased, the rate of decrease in SSDs begins to level off, eventually leading to the production of a curvy shape in the Elbow plot. As known, the elbow plot's curvature is frequently the outcome of an imbalance between the effectiveness of the clustering and the quantity of clusters applied. The outcome of this trade-off might have emerged because of the selected clustering algorithm.

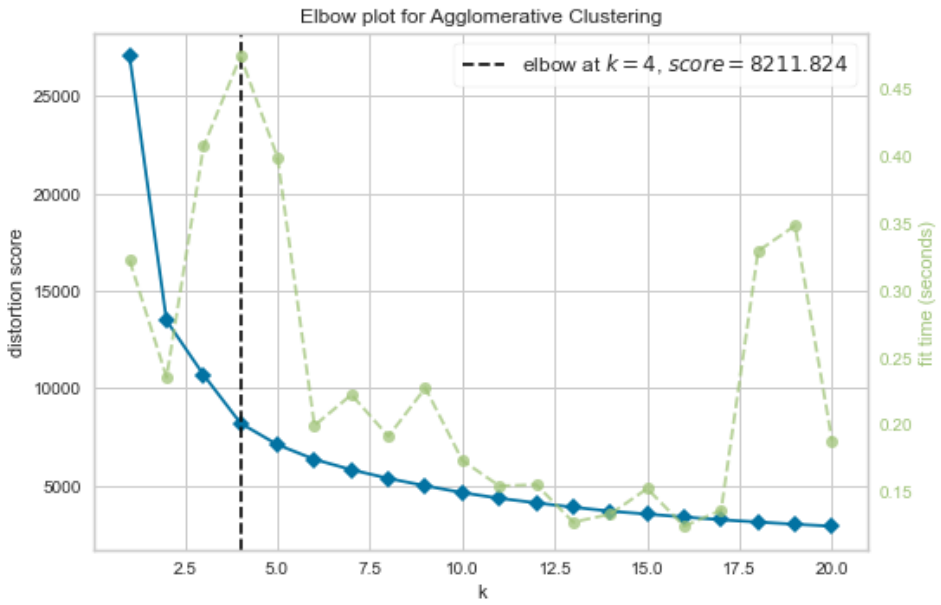


Figure 11: Elbow Method for Agglomerative Clustering

Therefore, Elbow Visualizer is implemented to see if the curvy shape is smooth or bumpy. It also applied for Agglomerative Clustering and the distortion score is about ‘1000’ points higher and the plot’s curvature has a slightly more smoother shape. But still it was hard to detect a suitable number for models.It means that the general K-Means and Agglomerative models has the best potential while the number of cluster is between ‘2’ and ‘4’ and looking for this graphs it is hard to say which one would be the most suitable one. For this reason, analytical inferences will be applied into 3 different scenarios which contain the situations of (k) equals ‘2’, ‘3’ and ‘4’.

b. Analytical Inference

In this section, since the graph obtained from the Elbow method alone is not convincing enough, scenarios should be examined one by one to understand the preferred number of clusters according to the model's movement methods.This way, scenarios with ‘2’, ‘3’, and ‘4’ cluster numbers in the relevant model iteration will be analyzed in detail, and the most accurate (k) variable will be sought. During the analysis, explanations will be made about the statistical

structure of the clusters and the insights they can provide. While doing this, the suitability of the graphical analysis for demographic segmentation will be discussed.

In the very first scenario, when (k) equals ‘2’, it should be noted that our silhouette score is currently at its highest level, indicating a scenario in which two clusters are notably distinct from one another. In such a circumstance, the primary indicator that must be scrutinized is the balance of observation quantities between clusters. Out of a total of ‘2020’ observations, ‘1229’ observations are present in “Cluster 0” while ‘791’ observations are present in “Cluster 1”. Consequently, it can be inferred that the clusters do not possess an equal amount of observations. Furthermore, by observing the income balance of each customer based on total expenses, it becomes apparent that there exists a clear-cut distinction between the clusters.

Figure 12: Cluster’s Profile Based on Income and Total Spending (n_clusters=2)

The average income of customers in “Cluster 0” is around ‘39000’, while their average expenses are only ‘200’. On the other hand, customers in “Cluster 1” have an average income of approximately twice that, around ‘72000’, but their average expenses is about ‘1250’. These findings are consistent with logical expectations when looking at the other features such as “Wines”, “Meat”, “Fish”. It can be said that the resulting clusters from clustering are well-separated. However, the imbalance in the number of observations between clusters is due to the fact that there are not enough observations, especially for “Cluster 1”, in the dataset. This imbalance is not related to the clustering process itself and does not have any effect on the clustering model. But the main problem arises when examining the impact of the “Education”, “Children”, and “NumDealPurchases” columns on the clustering model. It can be seen that the model only separates customers based on their income and spending, regardless of their education level, number of children, or shopping behavior related to campaigns. Therefore, it equalizes these demographic variations and separates customers only around the correlation between “Income” and “TotalSpent” features. Since the dataset cannot achieve demographic segmentation, having two clusters does not add value to the analysis. Therefore, a different value should be preferred for clustering models.

When the value of (k) is ‘3’, “Cluster 0” and “Cluster 1” do not have significantly different values compared to the previous clustering structure. However, this time the model discovers an intermediate class, “Cluster 2”, in the graph that focuses on income-based spending. The average income of this class is around ‘60,000’ while their spending is approximately ‘815’, which is the midpoint of “Cluster 0” and “Cluster 1” values. When the observation quantities of the clusters are examined, “Cluster 0” has fewer observations, ‘948’ compared to the previous clustering structure but is almost equal to the total observation quantities of “Cluster 1”, ‘535’ and “Cluster 2”, ‘615’.

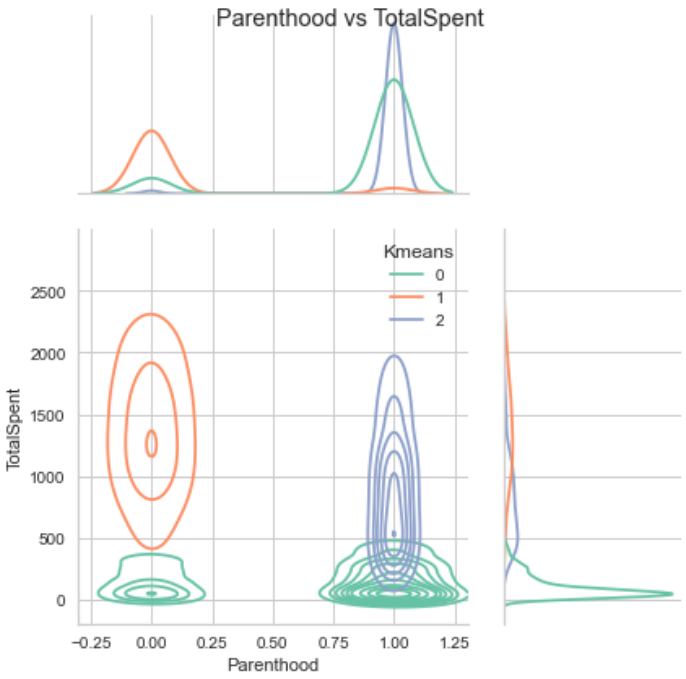


Figure 13: Cluster's Profile Based on Parenthood and Total Spending (n_clusters=3)

The crucial part in this segmentation is that the model has taken into account other demographic characteristics and each cluster has its own unique identities. For instance, in "Cluster 1", where the intensity of spending values are between '1000' and '1500', none of the customers are parents. Or, only a few customers in "Cluster 1" and "Cluster 2" have an undergraduate education level, while the majority have either graduate or postgraduate education levels. Choosing three clusters instead of two for the required number of clusters by the model can result with a much more varied outcome.

When (k) equals '4', all clusters contain a similar amount of observations. "Cluster 0" and "Cluster 3" represent the group with low income and low spending, while "Cluster 1" and "Cluster 2" represent the group with high income and high spending. Since the silhouette score is at its lowest, the distance between data points within clusters is very close. In this scenario, the dataset is not inadequate in terms of observations, but rather inadequate in terms of features. Since it cannot find different variations and weights of data points, the clustering structure created for (k) = '2' seems like a divided version of this scenario. In fact, when

looking at other demographic features, the group with low income and spending has been divided into two, and their demographic features are almost identical. The only clear distinction is that both groups with low and high income spend slightly more compared to the other group with similar demographic features.

In conclusion, the question of how many clusters the models should have can be answered by analyzing the individual scenarios for each value of (k) to gain a better understanding of the clustering structure. Therefore, for this specific demographic customer segmentation project, it was determined that the most appropriate choice is (k) = ‘3’.

3.4.2 Agglomerative Clustering

In terms of deciding the “n_clusters” parameter, the agglomerative method follows the same procedures as the K-means model. The difference is that, while the parameters for k-means clustering are determined using “GridSearchCV” technique, the agglomerative method requires a modified code to be iterated. This is because, in scenarios where the value of (k) is set to ‘3’, both models can yield a complex result that meets the desired analyses in the Analytical Inference part which under 3.4.1 section. Therefore, for agglomerative clustering, it is best to compare the silhouette scores obtained for scenarios with more than three clusters and find the best parameters accordingly. In other words, since the model's highest silhouette score is known to be achieved when (k) is equal to ‘2’, the number ‘3’ will serve as a threshold, and scenarios with higher values will be iterated.

While SSD was important to calculate euclidean distance for k-means algorithm, linkage criterion is the second most important step for an agglomerative model. Because the agglomerative model’s concept is to combine the closest pairs of clusters until there is only one cluster left. To make this combination, the process requires a way to measure the distance or similarity between clusters, which is where linkage comes in. It defines how the distance between two clusters is computed based on the distances between their individual data points. There are different types of linkage criteria such as “ward”, “average”, “complete”. While ward linkage measures the distance between clusters based on the increase in the sum of squared distances (SSD) when the two clusters are merged, complete linkage measures the distance between clusters based on the maximum distance between any two points in the two clusters. The two linking criteria' treatment of cluster size and shape is where they diverge most.

Complete linkage typically results in clusters that are smaller and more spherical, whereas Ward linkage typically results in clusters that are more balanced and evenly proportioned.

The other important parameter is “distance_threshold” which provides flexibility to the modely on deciding on the number of clusters. Because, it controls the granularity of the resulting clusters. It specifies the maximum distance between the two nearest neighbors, beyond which clusters are not merged. Therefore ir provides a flexible way to decide on the number of clusters based on the desired level of granularity in the clustering solution. In the loop results, all the silhouette scores where distance_threshold fails to be the most suitable parameter, while(k) = 3. This is because the clustering algorithm identified the optimal number of clusters in the data, and the distance_threshold parameter was not able to capture this structure.

As a result, an agglomerative model with a "ward" linkage method and 3 clusters is obtained. When the cluster assignments of the two models are examined, only a small portion of 8.96% of the '2022' observations do not match. All the other data points are in the same clusters regardless of what model implemented, which shows that both agglomerative and k-means algorithms follow a similar path.

3.5 Clusters

First, regardless of the value of the silhouette score, it is presumed that the scenario with three clusters in both the k-means and agglomerative algorithms is the most effective for customer segmentation on analytical scales. In addition to the analysis, the “Total_Promos” column, which contains information about five distinct campaigns, will be included and examined. In this section, a comparison will be made between agglomerative clustering and k-means clustering algorithms, and the current results will be discussed and by analyzing the “NumDealsPurchased” column, which displays the amount of shopping conducted based on the characteristics of customer segments, customers will be determined how they respond to campaigns.

3.5.1 Low-buyers

Considering the predominant characteristics of the consumers within this cluster, it is appropriate to refer to them as “LowBuyer”. In terms of spending, they are representative of the economic category between ‘0’ and ‘500’. The age group is relatively unimportant. They span a broad range of ages and are predominantly composed of parents or children. This is the group that participates in campaigns the least. They have returned in no more than two of five campaigns. Nobody has returned three times. However, in terms of the number of purchases made in response to the campaign, they rank second out of the three categories. From this, we can conclude that despite being the group with the lowest return to campaigns, their post-return purchasing is substantial.

3.5.2 Mid-buyers

Considering the predominant characteristics of the consumers in this cluster, it is appropriate to refer to them as “MidBuyer”. In terms of spending, they are closer to the top economic group than LowBuyers, and the disparity between them is wider. Their minimum age requirement is approximately 40 years old. There are children in this customer's residence, and this individual is generally a teenager. The majority have at least a bachelor's degree. Customers have responded to four out of five campaigns. They are the most prolific purchasers throughout the campaign in the dataset.

3.5.3 High-buyers

Given the predominant characteristics of this cluster's customers, it is appropriate to refer to them as “HighBuyers”. This demographic represents the highest spending class. The majority of the total spending data falls between ‘1000’ and ‘2000’ on average. Similar to the “LowBuyer” category, this group spans a broad age range. Only a tiny minority of households have children. None of this group's members are parents. Comparable to the “MidBuyer” category, the majority of their education level consists of at least a bachelor's degree. Similar to the middle segment group, this cluster contains consumers who responded to four of five campaigns. However, this demographic is the least likely to buy during campaigns.

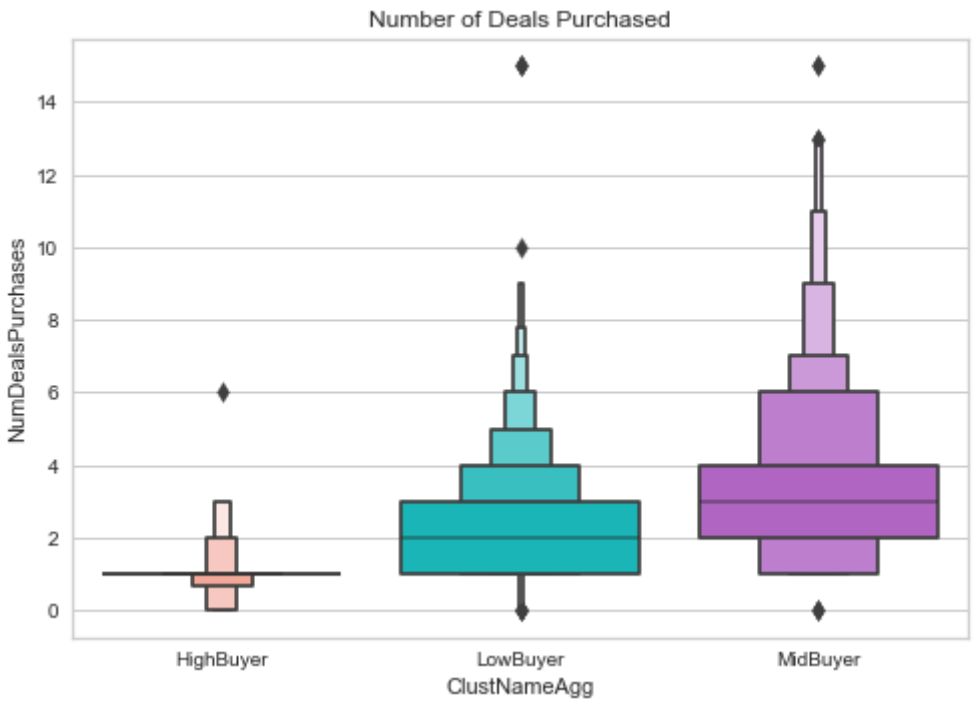


Figure 14: Promo Deals by Clusters

4. Result Interpretations and Findings

The first point to be compared is speed. When the two models are compared according to the relevant dataset, it has been determined that the k-means clustering algorithm gives faster results than agglomerative. The reason for this lies in two main factors: simpler computational structure and fewer calculations. Whereas the K-means algorithm is an iterative clustering algorithm that begins with an initial set of centroids and assigns each data point to its nearest centroid, Agglomerative clustering is a hierarchical clustering algorithm that builds a tree of clusters which begins by treating each data point as a single cluster and then iteratively merges the closest pair of clusters until all data points belong to a single cluster. The agglomerative model is pushed into the background by this form of computational structural diversity. This is due to the fact that computing the distance between all pairs of data points can be computationally expensive, especially for large datasets. The number of iterations necessary for convergence in the K Means Clustering method is typically quite low, particularly in situations

when the number of clusters under consideration is low.In the speed tests conducted within this project, while k-means method generally gave results between ‘0.05’ seconds and ‘0.12’ seconds, the time taken by agglomerative algorithm was between ‘0.170’ and ‘0.5’ seconds.

According to this project, when memory consumption tests were performed, no substantial discrepancies were detected in the names of the two models. The memory utilization of clustering techniques such as agglomerative and k-means can rely on numerous aspects such as the amount of data points, the dimensionality of the data, the distance metric employed, and the implementation of the method. Therefore, it may be difficult to find a clear demonstration or rule of thumb for comparing the memory utilization of various algorithms. However, generally it is acceptable that K-means technique consumes less memory than agglomerative clustering because it involves keeping only the cluster centroids and the cluster assignments of each data point. While in K-means, the memory usage is mostly governed by the number of clusters and the number of data points, agglomerative clustering needs keeping the distance matrix between all pairs of data points and the clustering tree. The problem of storing a distance matrix between each pair is that the size of this matrix rises quadratically with the number of data points[26], which can become prohibitive for huge datasets. It performs memory-intensive tasks for agglomerative clustering. But it might deliver a huge advantage that enables the application of more sophisticated linking criteria, which can capture different types of linkages between clusters. For example, a single linkage criterion joins two clusters based on the least distance between any two points from each cluster, while complete linkage criterion connects two clusters based on the greatest distance between any two points from each cluster. By employing multiple connection criteria, it is feasible to capture different characteristics of the clustering structure and create more meaningful clusters. Another advantage of maintaining the distance matrix enables for the employment of density-based clustering algorithms, which can discover clusters of different forms and sizes. Density-based algorithms such as DBSCAN and OPTICS employ the distance matrix to find regions of high density and segregate them from regions of low density. This method may prove especially useful in the case of datasets that have structures that are either irregular or complex, in which case conventional clustering algorithms may have difficulty locating meaningful clusters.

Finally, I will be using the “Calinski-Harabasz Index” to measure the clustering quality of the data. The “Calinski-Harabasz Index” is a measurement that examines the quality of clustering by determining how well clusters are separated from one another. A higher value of the index

suggests that the clusters are more well defined. When applied to the incumbent data, the Calinski-Harabasz Index for k-means clustering comes in at approximately ‘2022’, whereas the value for agglomerative clustering is approximately ‘1830’. For the specified dataset and number of clusters, it appears that the k-means clustering method was able to produce clusters that were more precisely defined than those produced by the agglomerative clustering algorithm. The “Calinski-Harabasz Index” is not, however, the only factor that should be considered when comparing two methods; instead, the qualities of the data and the issue at hand should be the primary considerations. It is important to note that the choice of clustering algorithm is dependent on the characteristics of the data and the issue at hand. The disparity in the values of the “Calinski-Harabasz Index” between the two algorithms may also imply that the data may be better suited for partitioning-based clustering methods, rather than partitioning-based clustering methods based on hierarchical clustering methods. Therefore, it is important to look at the unmatched data types between the clusters created by the two models. The most notable point in this analysis is that ‘94’ observations clustered as “HighBuyer” in the k-means model were classified as “LowBuyer” in the agglomerative model. There are four identical reasons. Firstly, the nature of the dataset is more close to k-means clustering. For example having a linear shape, being a small dataset, not giving identical data points for hierarchical clustering might be the cause of this reason. The other reason is, different parameters might have caused the problem. For example, in the research agglomerative clustering uses “Ward” linkage which minimizes the variance within each cluster. On the other hand, “Lloyd” algorithms aim to minimize the sum of squared differences between each data point and the centroid of its assigned cluster. It means that while the “Ward” linkage method tends to produce clusters of similar sizes and compact shapes, the Lloyd algorithm tends to produce clusters of different sizes and shapes. Lastly, The 94 observations may have unique combinations of demographic and behavioral characteristics that make them more similar to one group of customers than another or there could be errors or noise in the data that affect the clustering results. But after careful consideration, missing, inaccurate or any unique characteristic regarding the observations couldn't be found.

As a consequence of this, it has been established that the typical structure of our data set is more suited to the k-means algorithm than it is to the agglomerative approach in terms of going through a quick and efficient segmentation process. This conclusion was reached as a consequence of the fact that it has been found that the distinctive structure of our data set. The k-means method may be an operative choice in customer segmentation projects when the

primary factors are dictated by the demographic structure of the consumers because it is known that common demographic datasets are comparable to the dataset that was utilized in this research.

5. Future of Research Capacity

With the emergence of big data, businesses are confronted with an ever-increasing volume of data that can be difficult to analyze and interpret. Clustering techniques are becoming increasingly important in order to make sense of this data and obtain actionable insights. Clustering can aid in the identification of patterns, the grouping of similar clients, and ultimately the improvement of business decisions.

In the future, clustering algorithms will become more sophisticated and capable of managing even larger and more complex datasets through an incremental process. This will most likely involve advancements in machine learning, enabling algorithms to automatically recognize patterns and segment customers without requiring input from individuals. In a survey of more than 160 US-based executives from companies using or planning to use advanced technologies in digital marketing, more than one-third (35.2%) of respondents indicated that improved customer segmentation is one of the most important ways they are experimenting with AI in digital marketing. In addition, more than half of respondents (54%) are presently using AI or machine learning to provide customers with a personalized experience, with another third (39%) planning to do so in the future.[27] Even now, prominent corporations have developed AI-supported customer segmentation products. By leveraging AI-driven customer segmentation, such as IBM Watson's customer segmentation tool, many Salesforce AI-supported tools, Google Cloud Platform, Microsoft Dynamics 365 AI-driven insights, Amazon's automated segmentation system, etc., brands and organizations can gain a competitive edge in the current market and drive revenue growth.

As businesses continue to collect information from a broader variety of sources, such as social media, online behavior, and mobile devices, clustering techniques will need to be modified to effectively manage this information. In the future, in addition to the expansion of data

collection sources, new data collection frameworks may also emerge. Graph databases and graph modeling frameworks may be an appropriate example for this. According to DB-Engines Ranking, the prevalence of graph databases has been steadily increasing for years.[28] This may pave the way for developing new algorithms explicitly designed to work with unstructured data, or for leveraging advances in classification techniques to better comprehend customer sentiment and preferences. Increasing importance of personalized marketing is another trend that will likely influence the future of clustering consumer segmentation. Taking into account the products of the aforementioned companies, personalized marketing strategies on the market also contribute to the expansion of the market for the responsible business. As consumers become more knowledgeable and demand more personalized experiences, businesses will need to rely on clustering to better understand their customers and modify their marketing strategies accordingly.

6. Critical Evaluation

During the process of determining the number of clusters in our dataset and in the analysis stages, it was generally noted that the imbalance in the dataset was insufficient, or that there was not a dataset with a wide variety of features for analysis. Just like the examination of these processes, the inadequacies of the dataset in terms of data points prevented the clusters from being more prominent, and the occurrence of a unique column prevented the analysis from deepening. A dataset with more rows and more observations in terms of data points, containing deeper correlations and insights, could help turn the results observed in this thesis into a much more unbiased and externally influenced study.

Another approach to using data augmentation to balance out the number of observations between the clusters by increasing the amount of data in smaller clusters. This technique is commonly used in computer vision and NLP projects to increase the diversity of the dataset. However, it is important to be aware that adding variations to the data can result in the loss of original data or introduce biases if not implemented carefully. Common approach to data augmentation is generative models which can learn the underlying patterns and relationships between customer attributes and generate new synthetic customer profiles that are similar to the existing ones. However, the implementation of these generative models can be challenging

and require high computational complexity, which may make them difficult to apply in practice.

Finally, one of the most important categories in our dataset is the customer response to campaigns, which is captured in the columns “AcceptedCmp1”, “AcceptedCmp2”, “AcceptedCmp3”, “AcceptedCmp4”, and “AcceptedCmp5”. If we had more detailed strategic information about these campaign plans, we could perform more in-depth analysis and draw more detailed insights from the results. For example, these campaigns might offer price discounts or "buy one, get one free" promotions, or they may be loyalty programs developed by the store to improve its customer portfolio. Not knowing these details limits the ability to draw meaningful insights beyond what the numerical values in the analysis can show.

7. Conclusion

With the increasing use of digital products, the amount of data has also grown at an unstoppable pace. Companies that closely follow new technological changes use the latest technological tools in marketing to maintain a competitive edge. While it is the case, many organizations and businesses aim to maintain the strongest possible connection with their customers and understand their needs and preferences for their product and services. Therefore, one of the most important topics in recent years, personalized marketing strategies have become key elements for different sectors. Companies in banking sector, e-commerce sector, retailing sector, finance sector etc. find attractive the idea of personalized marketing strategies, thus, have taken advantage of the invested in this area by resorting to various software updates to take advantage of the new technological innovations . With the customer segmentation concept finding its place in data science literature, by evaluating segmentation with the support of machine learning algorithms, has become one of the crucial developments. This thesis statement examines a hybrid segmentation approach that consists of the demographic and behavioral analysis of customers by using clustering methods in unsupervised machine learning algorithms, which allows marketing professionals to find patterns and insights with the inference of customer clusters based on the clients demographic and purchasing behavior characteristics. In this particular grocery store setting in the retailing sector, the analysis

revealed that customers with a broad range of spending habits were more responsive to the store's campaigns than those with medium and low spending habits. Furthermore, partition-based and hierarchical clustering models were analyzed, and it was found that KMeans performed better than agglomerative models with the consideration of the provided dataset's limitations. This thesis concludes itself with the aim of supporting other studies in clustering methods in customer segmentation projects and demonstrating to current organizations the potential impact of clustering models on customer segmentation procedures. In the future, new data structures and frameworks will continue to evolve in the literature and the related industries, allowing for even deeper and larger datasets to be analyzed while finding admires the related improvements.

9. References

[1] Schiffman, L. G., Schiffman, L., & Kanuk, L. L. (2010). Consumer Behavior (10th ed.). Pearson

[2] McDonald, M. (2012). Market Segmentation: How to Do It and How to Profit from It, Revised 4th Edition. John Wiley & Sons

[3] Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd ed.). Morgan Kaufmann

[4] Verhoef, P. C., Lemon, K. N., Parasuraman, A. P., & Schlesinger, L. A. (2009). Customer experience creation: Determinants, dynamics and management strategies. Journal of Retailing, 85(1), 31-41. doi: 10.1016/j.jretai.2008.11.001

[5] Chen, Y.-H., & Barnes, S. J. (2007). Initial trust and online buyer behavior. Industrial Management & Data Systems, 107(1), 21-36. doi: 10.1108/02635570710719034

[6] Verhoef, P. C., Neslin, S. A., & Vroomen, B. (2007). Multichannel customer management: Understanding the research-shopper phenomenon. International Journal of Research in Marketing, 24, 129-148.

[7] McKinsey & Company. (2018). The Business Value of Design. [online] Available at: https://www.mckinsey.com/business-functions/mckinsey-design/our-insights/the-business-value-of-design (Accessed: 2 Mar. 2023)

[8] Lemon, K. N., & Verhoef, P. C. (2016). Understanding Customer Experience Throughout the Customer Journey. Journal of Marketing, 80(6). https://doi.org/10.1509/jm.15.0420

[9] Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Elsevier.

[10] Boopathi, G. and Arockiasamy, S. (2016), ‘An Image Compression Approach using Wavelet Transform and Modified Self Organizing Map’, in Thangaraj, R. and Pandian, M. (eds.) Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, Springer, Singapore, pp. 331-341.

[11] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020), 'The k-means algorithm: A comprehensive survey and performance evaluation', Electronics, 9(8), 1295. Available at: https://doi.org/10.3390/electronics9081295 (Accessed: 6 Mar. 2023)

[12] Boomija, M.D. (2008). Comparison of partition based clustering algorithms. Journal of Computer Applications, Prathyusha Institute of Technology and Management.

[13] J, S. S., & Pandya, S. (2016). An Overview of Partitioning Algorithms in Clustering Techniques. International Journal of Computer Applications, 149(10), 21-26.

[14] Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining. Pearson Education.

[15] Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

[16] Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). Clustering Validation: A Concise Overview. IEEE Transactions on Knowledge and Data Engineering, 16(9), 1-16.

[17] Wang, H., Wang, W., & Yang, J. (2014). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Transactions on Intelligent Systems and Technology, 5(1), 1-31.

[18] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

[19] Haitian Wei. (2020, January 2). How to measure clustering performances when there are no ground truth? [Blog post]. Retrieved from https://medium.com/@haataa/how-to-measure-clustering-performances-when-there-are-no-ground-truth-db027e9a871c

[20] Abigyhan, S. (2020, June 14). Importance of Dimensionality Reduction. Analytics Vidhya. https://medium.com/analytics-vidhya/importance-of-dimensionality-reduction-d6a4c7289b92

[21] Bellman, R. (1961). Adaptive Control Processes: A Guided Tour (Princeton University Press).

[22] Bellman, R. E. (1957). Dynamic Programming. Princeton University Press.

[23] Kumar, S., (2020), "How to remove Multicollinearity in dataset using PCA?", Towards Data Science, December 19, 2020. Available at: https://towardsdatascience.com/how-to-remove-multicollinearity-in-dataset-using-pca-4b4561c28d0b (Accessed: 7 Mar. 2023)

[24] Gupta, N. (2015). Using the Triangle Inequality to Accelerate k-Means.

[25] scikit-learn. (n.d.). sklearn.cluster.KMeans. Retrieved April 21, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[26] Wittek, P. (2014). Unsupervised Learning. University of Borås, Sweden. Available online 29 September 2014.

[27] Carolyn Crafts. (2022, December 02). How Brands are Using Artificial Intelligence in Customer Segmentation. FullSurge. https://www.fullsurge.com/blog-old/how-brands-use-artificial-intelligence-in-customer-segmentation-0#:~:text=By%20analyzing%20customer%20data%2C%20AI,personalized%20experiences%20for%20their%20customers

[28] McKenna, B. (2022, November 23). Why graph technology is for all our digital futures. ComputerWeekly.com. https://www.computerweekly.com/blog/Data-Matters/Why-graph-technology-is-for-all-our-digital-futures

[29] Yoseph, F., Malim, N. H. A. H., & AlMalaily, M. (n.d.). New behavioral segmentation methods to understand consumers in retail industry. School of Computer Science, Universiti Sains Malaysia, & Faculty of Engineering, Mansoura University.

[30] DOĞAN, O., AYÇİN, E., & BULUT, Z. A. (2017). Customer segmentation by using RFM model and clustering methods: A case study in retail industry. Journal of Business Research-Turk, 9(2), 76-97.

[31] Patel, A. (n.d.). Customer Personality Analysis. Kaggle. Retrieved on [14 Feb. 2023] from https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

10. Appendix

Appendix A

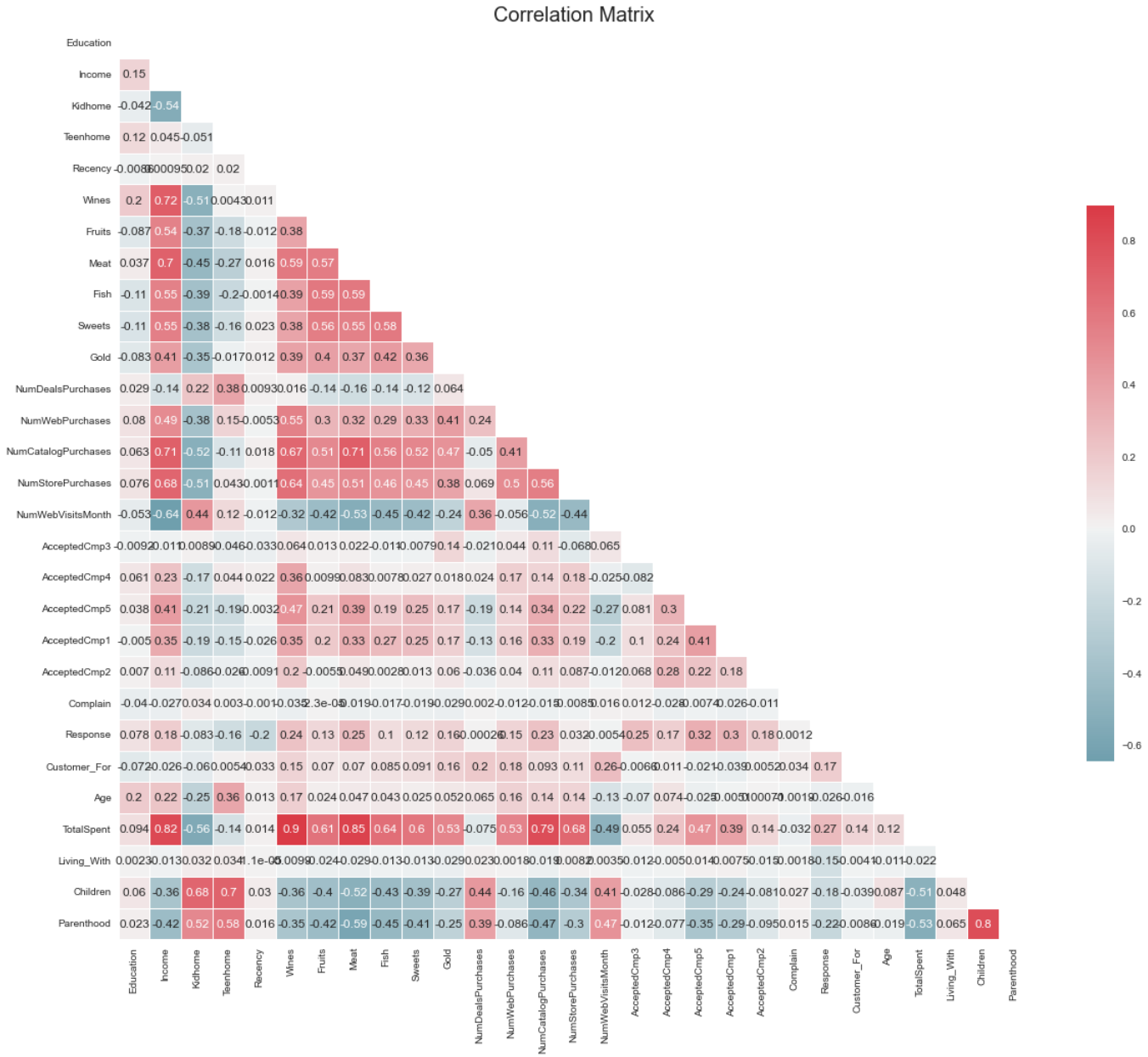
Column Name	Description
ID	Customer's unique identifier
Year_Birth	Customer's birth year
Education	Customer's education level
Marital_Status	Customer's marital status
Income	Customer's yearly household income
Kidhome	Number of children in customer's household
Teenhome	Number of teenagers in customer's 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
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
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
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

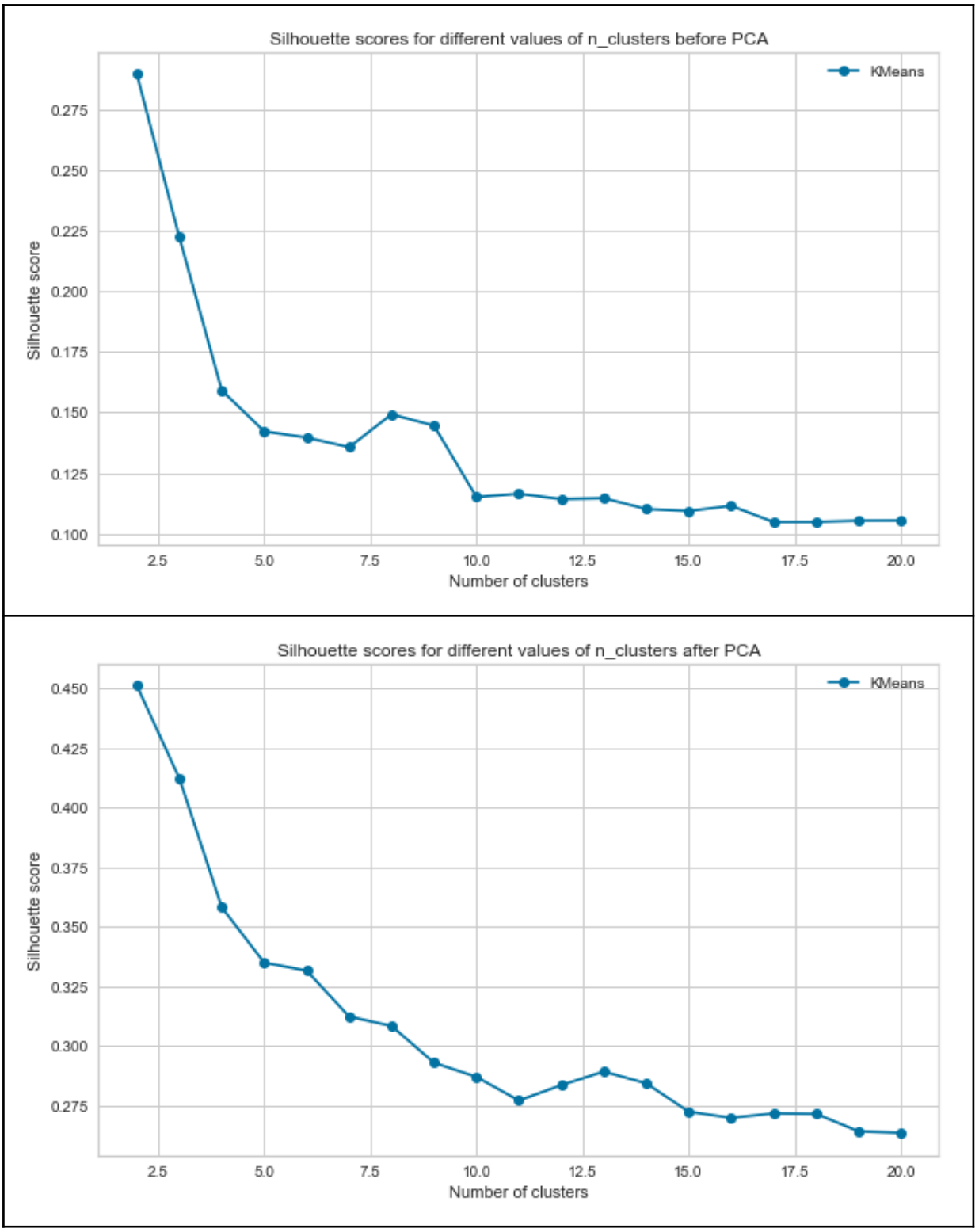
Appendix B

Education	Income	NumDealsPurchases	Age	TotalSpent	Living_With	Children	Parenthood
1	58138.0	3	66	1617	0	0	0
1	46344.0	2	69	27	0	2	1
1	71613.0	1	58	776	1	0	0
1	26646.0	2	39	53	1	1	1
2	58293.0	5	42	422	1	1	1
2	62513.0	2	56	716	1	1	1
1	55635.0	4	52	590	0	1	1
2	33454.0	2	38	169	1	1	1
2	30351.0	1	49	46	1	1	1
2	5648.0	1	73	49	1	2	1
0	7500.0	1	47	61	1	0	0
1	63033.0	1	64	1102	0	0	0
2	59354.0	3	71	310	0	2	1
1	17323.0	1	36	46	1	0	0
2	82800.0	1	77	1315	0	0	0
1	41850.0	3	43	96	1	2	1
1	37760.0	2	77	317	1	0	0
2	76995.0	2	74	1782	1	1	1
0	33812.0	2	38	133	0	1	1
1	37040.0	1	41	316	1	0	0
1	2447.0	15	44	1730	1	1	1
2	58607.0	3	74	972	1	1	1
2	65324.0	3	69	544	1	1	1
1	40689.0	7	72	444	1	1	1
1	18589.0	2	54	75	0	0	0
1	53359.0	4	47	257	1	2	1
1	38360.0	2	34	131	1	1	1

Appendix C


Correlation Matrix
Education
Income
Kidhome
Teenhome
Recency
Wines
Fruits
Meat
Fish
Sweets
Gold
NumDealsPurchases
NumWebPurchases
NumCatalogPurchases
NumStorePurchases
NumWebVisitsMonth
AcceptedCmp3
AcceptedCmp4
AcceptedCmp5
AcceptedCmp1
AcceptedCmp2
Complain
Response
Customer_For
Age
TotalSpent
Living_With
Children
Parenthood
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Appendix D


Silhouette scores for different values of n_clusters before PCA
KMeans
Silhouette score
0.275
0.250
0.225
0.200
0.175
0.150
0.125
0.100
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Number of clusters
Silhouette scores for different values of n_clusters after PCA
KMeans
Silhouette score
0.450
0.425
0.400
0.375
0.350
0.325
0.300
0.275
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Number of clusters