T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

THE DEPARTMENT OF BIG DATA ANALYTICS AND MANAGEMENT

NOVEL MACHINE LEARNING-BASED APPROACHES FOR CUSTOMER SEGMENTATION AND DATA ANALYSIS

MASTER'S THESIS

NUR DIYABI

ISTANBUL 2024

T.C.
BAHCESEHIR UNIVERSITY
GRADUATE SCHOOL
THE DEPARTMENT OF BIG DATA ANALYTICS AND MANAGEMENT

NOVEL MACHINE LEARNING-BASED APPROACHES FOR CUSTOMER SEGMENTATION AND DATA ANALYSIS

MASTER'S THESIS

THESIS ADVISOR
Assoc. Prof. Dr. Ömer Melih Gül

ISTANBUL 2024

T.C.
BAHCESEHIR UNIVERSITY
GRADUATE SCHOOL

27/05/2024

MASTER THESIS APPROVAL FORM

Program Name:	Big Data Analytics and Management Master's Program (English, with Thesis)
Student's Name and Surname:	NUR DIYABI
Name of The Thesis:	Novel Machine Learning-Based Approaches For Customer Segmentation And Data Analysis
Thesis Defense Date	27/05/2024

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Assoc. Prof. Yücel Batu SALMAN

Director of Graduate School

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title, Name	**Institution**	**Signature**
Thesis Advisor:	Assoc. Prof. Dr. Ömer Melih Gül	Istanbul Technical University	
2nd Member (Outside Instituion)	Asst. Prof. Dr. Ismail Burak Parlak	Galatasaray University	
3rd Member	Assoc. Prof. Dr. Cemal Okan Şakar	Bahcesehir University	

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

Name, Last Name: Nur Diyabi

Signature:

ABSTRACT

NOVEL MACHINE LEARNING-BASED APPROACHES FOR CUSTOMER SEGMENTATION AND DATA ANALYSIS

Nur, Diyabi
Master's Program in: BIG DATA ANALYTICS AND MANAGEMENT
Supervisor: Assoc. Prof. Dr. Ömer Melih Gül

June 2024, 73 pages

Understanding clients and adapting marketing techniques to their preferences is crucial in today's competitive corporate environment. This research investigates data analysis by proposing Machine Learning-based approaches for customer segmentation and investigates its cluster customer segment results. In addition to looking into customer segmentation using machine learning approaches. It proposes five clustering methods, DBSCAN, Self- organizing maps , k-nearest Neighbors, Gaussian mixture model, and K-means, for grouping customers based on their purchasing habits, age , income and other attributes collected from the data. The study then compares the efficacy of each method in developing distinct customer segments. These methodologies include data collection, preparation, the use of clustering methods, and the evaluation of the numerical results. The main aim of this paper is to understand the different clustering methods and mark their differences according to the outcome of the given data. Using data analytic and insight extraction and identifying the customer groups would be the first step to comparative methods and insight extraction turning the data into a useful tool for the supermarket Marketing team. The research focuses on machine learning algorithms and compares these methods.

Key words: Density-Based Clustering , EPS (Minimum Epsilon Distance) , MinPts (Minimum Points). , Unsupervised Learning , Machine Learning.

ÖZ

Müşteri Segmentasyonu ve Veri Analizi için Yenilikçi Makine Öğrenmesi Yaklaşımları

Nur, Diyabi
BÜYÜK VERİ ANALİTİĞİ ve YÖNETİMİ Yüksek Lisans Programı
Tez Danışmanı: Doç. Dr. Ömer Melih Gül

Haziran 2024, 73 sayfa

Bugünlerde yoğun rekabet ortamında müşterileri anlamak ve pazarlama tekniklerini tercihlerine göre uyarlamak kritik öneme sahiptir. Bu araştırma, müşteri segmentasyonu için Makine Öğrenmesi tabanlı yaklaşımlar önererek veri analizini inceliyor ve kümeleme yoluyla elde edilen müşteri segmentlerinin sonuçlarını araştırıyor. Veri analizi yöntemlerine ek olarak, araştırma müşterileri satın alma alışkanlıkları, yaş, gelir ve veriden toplanan diğer özelliklere göre gruplandırmak için beş kümeleme yöntemi önermektedir: DBSCAN, Kendinden Örgütlenen Haritalar, k-En Yakın Komşu, Gauss Karmaşık Modeli ve K-Means. Daha sonra çalışma, farklı müşteri segmentleri oluşturmada her bir yöntemin etkinliğini karşılaştırır. Bu metodolojiler veri toplama, hazırlama, kümeleme yöntemlerinin kullanımı ve sayısal sonuçların değerlendirilmesini içerir. Bu makalenin temel amacı, farklı kümeleme yöntemlerini anlamak ve verilen verilerin sonuçlarına göre farklılıklarını işaretlemektir. Veri analitiği ve içgörü çıkarımı kullanarak müşteri gruplarını tanımlamak, karşılaştırmalı yöntemlere ve içgörü çıkarımına giden ilk adım, verileri süpermarket Pazarlama ekibi için faydalı bir araca dönüştürmek olacaktır. Araştırma, makine öğrenmesi algoritmalarına odaklanmakta ve bu yöntemleri karşılaştırmaktadır.

Anahtar Kelimeler: Yoğunluğa Dayalı Kümeleme , EPD (Minimum Epsilon Mesafe) , MinPts (Minimum Nokta Sayısı) , Gözetmensiz Öğrenme , Makine Öğrenmesi.

To my beloved family and Father.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my thesis advisor, Assoc. Prof. Dr. Ömer Melih Gül, for his unwavering guidance, mentorship, and invaluable support throughout my thesis journey. His expertise and dedication have been instrumental in shaping the course of this research. I am truly grateful for his constant encouragement and willingness to share his knowledge, which significantly impacted the quality of this work. I would also like to extend my sincere thanks to the jury members, Asst. Prof. Dr. Ismail Burak Parlak and Assoc. Prof. Dr. Cemal Okan Şakar, for their time and consideration in evaluating this thesis. Their valuable insights and feedback will undoubtedly contribute to the further development of this research. I would like to also thank my husband for his patience and support. And my family for being by my side and giving me moral support through the distances.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF ABBREVIATIONS

EPS	Epsilon
DBSCAN	Density-Based Spatial Clustering
MinPts	The minimum number of points
CS.	Customer Segmentation
KNN	k- Nearest Neighbor
GMM	Gaussian Mixture Model
SOM	Self-organizing Maps
DB	Davies-Bouldin Score
CH	Calinski-Harabasz Score

Chapter 1

Introduction

In the ever-changing retail landscape, where customer preferences shift like desert sands and competition is fierce, understanding customer segments has become a necessity rather than a luxury. The Turkish retail landscape is highly competitive and ever-changing. Understanding customer preferences has evolved into a strategic imperative for local supermarkets in today's fast-paced environment, rather than a nice to have. Customer segmentation, or the meticulous process of grouping customers with similar characteristics and purchasing behaviors, goes beyond basic demographics. It delves deeper, exposing hidden patterns and needs in customer data. Businesses can use this knowledge to create targeted marketing campaigns and promotions that are particularly appealing to specific customer segments. This data-driven approach maximizes ROI and promotes long-term customer loyalty (Sharma, 2023). This study assesses the efficacy of various customer segmentation methods using a large-scale Turkish market sales dataset obtained from Kaggle (Kabasakal, 2020)**.** Our primary goal is to identify the most effective and actionable customer segmentation techniques designed specifically for the Turkish market.

We accomplish this by comparing five major clustering algorithms: K-Means clustering, SOM clustering, GMM clustering, KNN clustering, and DBSCAN. K-Means Clustering is a well-known benchmark that divides customers into a set number of clusters based on their purchasing habits. DBSCAN, on the other hand, excels at identifying distinct customer segments, particularly those with irregular shapes, which is especially important in the retail industry where customer behavior is variable and non-spherical. DBSCAN detects unusual patterns and is effective with outlier-rich data, such as the Turkish market dataset (Nguyen, 2022). While K-Means clustering remains a fundamental customer segmentation technique, the field has expanded to include more advanced approaches like Density-Based Clustering (DBSCAN) and Collaborative Filtering for Shared Preferences. K-Nearest Neighbors (KNN) is a method that classifies a new customer using the majority vote of its k nearest neighbors in the training data. KNN can also be used for regression, predicting a new customer's value using the average value of its k nearest neighbors (James et al., 2013). Self-Organizing Maps (SOMs) are neural network architectures that project high-dimensional data onto a lower-dimensional space while preserving data point

relationships (Kohonen, 2001). SOMs can help visualize customer behavior, reduce data dimensionality, and extract features. SOMs do not directly perform segmentation, but they can help visualize customer behavior and understand the relationships between segments. The Gaussian Mixture Model (GMM) assumes that data points are generated by a combination of multiple Gaussian distributions. GMM uses the parameters (means and covariances) of these distributions to model the data's underlying structure. GMM is useful for tasks such as clustering, density estimation, and anomaly detection. While GMM does not directly segment customers, it can simulate the underlying data distributions for various segments. This modeling can be used to better understand segment characteristics and create more effective marketing strategies.These methods offer significant advantages in identifying irregularly shaped clusters and recognizing customer groups with similar purchasing habits. To ensure a thorough evaluation, we will develop a strong framework for comparing customer segmentation results obtained from the following methods: K-Means Clustering (Baseline), DBSCAN, and possibly other advanced techniques based on our initial findings. This framework will carefully assess critical factors for actionable customer segmentation in the Turkish market, such as cluster interpretability, within-cluster homogeneity, between-cluster separation, and actionable insights. By identifying the most effective approach to customer segmentation, this study hopes to provide local supermarkets with data-driven decisions about customer targeting strategies, ultimately leading to increased customer understanding and loyalty in Turkey.

The unique contributions of this work are given as the following:

- In Section 9 that identifies the most effective approach for customer segmentation in the dynamic Turkish market by comparing the performance of K-Means, a well-established technique, and DBSCAN, GMM, SOM , and KNN.
- In Section 7 where actionable insights for Turkish supermarkets providing local supermarkets with data-driven customer targeting strategies based on actionable insights gleaned from the most effective segmentation method, resulting in increased customer understanding and loyalty in Turkey.

Identify the optimal clustering method for this dataset gives us insight upon few different criteria such as:

Identify the optimal clustering method for this dataset; Through rigorous evaluation, we will determine which method (DBSCAN, K-Means, SOM, etc…) produces the most insightful, actionable, and future-proof customer segments for the Turkish market. Discover the unique benefits of each approach**;** The analysis will highlight the advantages and disadvantages of clustering methods in terms of customer segmentation in the Turkish market. This will provide useful insights for businesses considering clustering techniques for their own customer segmentation efforts. K-Means Clustering Evaluation; A Basis for Comparison We will use the well-known K-Means clustering algorithm as a foundational benchmark. K-Means will be used to segment customers based on their purchasing patterns in the Turkish market. This includes optimizing the number of customer clusters (K) and thoroughly analyzing the characteristics that define each identified segment (Brown, 2021). By establishing a baseline understanding of customer segments using K-Means, we can effectively compare it to more sophisticated techniques. Unveiling arbitrary shapes; Density-Based Clustering (DBSCAN) Unlike centroid-based techniques such as K-Means, DBSCAN excels at detecting arbitrary shaped clusters within data.expand more It excels at detecting unusual patterns and works well with outlier-rich data, which is a possible scenario in the Turkish market dataset (Nguyen, 2022). Collaborative Filtering for Shared Preferences; Recommendation systems that use collaborative filtering identify customer groups with similar purchasing habits, forming segments based on shared interests rather than demographics. This personalized approach can significantly increase customer satisfaction and loyalty in the Turkish market (Ribeiro et al., 2018). Creating a Robust Comparative Framework; To ensure a thorough evaluation, we will create a strong framework for comparing the customer segmentation results obtained from K-Means, the chosen alternative method, and possibly other advanced techniques. This framework will meticulously evaluate the following aspects:

Cluster Interpretability: How well do the features of each cluster represent distinct and meaningful customer segments in the Turkish market?

Within-Cluster Homogeneity: To what extent do customers in each identified cluster have similar purchasing habits in the Turkish market?

Between-Cluster Separation: How different are the identified customer segments from one another in terms of purchasing habits?

Actionable insights**:** Beyond interpretability, how easily can the identified segments be translated into actionable marketing strategies and customer service approaches specific to the Turkish market? This comparative analysis of the methods used in clustering seek to illuminate the most effective approach to customer segmentation in the Turkish market, allowing businesses to make more informed decisions about their customer targeting strategies.

1.1 Purpose of the Study

This study embarks on a captivating journey to segment the customer base of a local supermarket, not just using the established K-means method, but also venturing into the exciting realm of novel approaches and comparing it to other approaches. Our rich dataset encompasses diverse customer attributes like birth year, , education, marital status, website visit, deal purchase, web purchase, and catalog purchase, brimming with the potential to unearth distinct customer segments. Our primary quest is to identify unique customer communities based on their purchasing habits, preferences, and other relevant traits. This understanding unlocks the door to personalized marketing campaigns, tailored product offerings, and enhanced customer service– all crucial components of thriving customer relationship management (CRM) strategies. Beyond K-means: Exploring the Frontiers of Novel Approaches While K-means has proven its worth in customer segmentation, we delve deeper, investigating novel approaches that offer cutting-edge advancements:

This thesis seeks to assess the efficacy of various machine learning algorithms for customer segmentation. On a single customer dataset, we will evaluate the performance of K-Means, K-Nearest Neighbors (KNN), Self-Organizing Maps (SOM), Gaussian Mixture Models (GMM), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). By analyzing the strengths and weaknesses of each approach in this context, we hope to find the best algorithm for achieving optimal customer segmentation within the dataset. This comparative analysis will provide valuable insights into the efficacy of various clustering techniques for customer segmentation tasks, assisting businesses in determining the best method for their specific requirements.

DBSCAN: Unlike centroid-based techniques such as K-Means, DBSCAN is an excellent tool for locating arbitrary shape clusters in data. It clusters points according to density, classifying isolated points as noise and designating regions with high point

concentrations as clusters. Because of this, it is effective at spotting unusual patterns and performs well when dealing with outlier-filled data.

Collaborative Filtering for Shared Preferences: Recommendation systems powered by collaborative filtering identify customer groups with similar buying patterns, forming segments based on shared interests rather than just demographics. This personalized approach can significantly enhance customer satisfaction and loyalty.

Ensemble Methods for Robust Insights: Imagine combining the strengths of multiple decision trees like random forests. This ensemble approach offers robustness against overfitting and delivers insightful explanations for the segmentation process, providing valuable clues into customer behavior.

Comparative Analysis: Unveiling the Winning Approach By meticulously comparing the performance of K-means with these novel approaches, we aim to identify the method that best aligns with the supermarket's specific needs and data characteristics. Factors like accuracy, interpretability, adaptability, and computational efficiency will guide our evaluation.

The goal is to equip the supermarket with the most effective customer segmentation strategy, empowering them to:

Boost customer satisfaction: Address individual needs and preferences, fostering loyalty and positive brand perception.

Enhance marketing effectiveness: Target campaigns precisely, maximizing return on investment.

Optimize product assortment: Cater to specific segment preferences, minimizing inventory costs and maximizing sales.

Offer differentiated customer service: Tailor service tactics to individual groups, ensuring a seamless and personalized experience.

Strengthen loyalty programs: Craft rewards and perks that resonate with each segment, driving retention and repeat business.

This comparative exploration of customer segmentation techniques promises to unlock valuable insights, empowering the supermarket to make data-driven decisions, cultivate deeper customer relationships, and ultimately achieve sustainable success. Remember, the key lies not just in identifying the "best" method, but in choosing the approach that best harmonizes with the supermarket's unique goals, data landscape, and resources. By harnessing the power of both established and novel approaches, we

can embark on a transformative journey towards an enlightened customer segmentation strategy.

1.2 Hypotheses/Research Questions

DBSCAN will outperform K-Means in identifying customer segments in the Turkish retail market, particularly in capturing unique or irregularly shaped customer groups, due to its density-based methodology.

SOM will offer the best and the most accurate results out of the clustering methods.

KNN will outperform K-means.

GMM will show the most precise outcomes out of the 5 methods, but will be the hardest to perform.

1.3 Significance of the Study

Understanding customer segments has gone from a luxury to a necessity in Turkey's fast-paced retail landscape. While K-means clustering has proven useful, local supermarkets are looking for more effective solutions. These are just a few of the advanced techniques we'll investigate. Our goal? To determine the segmentation strategy that best allows Turkish supermarkets to:

- Build customer loyalty by addressing individual needs and preferences, promoting brand loyalty and positive experiences.
- Maximize Marketing ROI: Precision-target campaigns reach the right customers with compelling messages.
- Optimize product selection: Tailor to specific segment preferences while reducing costs and increasing sales.
- Provide differentiated services. Customize service strategies for specific groups, ensuring seamless and personalized experiences.
- Strengthen loyalty programs: Create rewards and perks that resonate with each segment, resulting in increased retention and repeat business. By comparing K-means to these innovative techniques, we'll be able to identify the winning strategy, allowing Turkish supermarkets to make more informed decisions, cultivate deeper customer relationships, and achieve sustained success. Remember that the key is not just to find the "best" method, but to choose the approach that best fits each supermarket's specific goals and data landscape. This investigation promises to yield valuable insights,

allowing Turkish supermarkets to thrive in the ever-changing retail landscape. This thesis will help supermarkets to choose the optimal clustering technique and method according to their needs and data.

Chapter 2

Literature Review

In today's dynamic business environment, customer segmentation and data analysis have become indispensable tools for achieving sustainable success. By delving into the intricacies of customer needs and preferences, businesses can craft targeted campaigns that resonate with specific customer segments, effectively reducing unnecessary marketing expenses and fostering stronger customer relationships. This, in turn, paves the way for improved profitability, enhanced product offerings, and a more loyal customer base.

Machine learning, a rapidly evolving field at the forefront of technological innovation, plays a pivotal role in revolutionizing customer segmentation strategies. Its ability to identify patterns and trends in vast amounts of data enables businesses to gain a deeper understanding of their customer base, leading to more effective segmentation and targeted marketing campaigns.

Machine learning algorithms, when trained on historical data such as customer demographics, product information, and payment methods, can predict and uncover similar patterns among customer groups. By analyzing large datasets, these algorithms can accurately define meaningful customer segments, allowing businesses to tailor their marketing efforts and product development strategies to address the specific needs and preferences of each segment.

The application of machine learning in customer segmentation has been successfully demonstrated across various industries. In the retail sector, a deep learning approach identified four distinct customer segments: at-risk, low-value, medium-value, and high-value customers. This segmentation allowed the company to create targeted marketing campaigns that resonated with each segment, leading to improved customer engagement and increased sales (Nguyen, 2022).

Similarly, in the banking and financial services industry, a machine learning algorithm segmented customers into three distinct groups based on risk profiles: high-risk, medium-risk, and low-risk. This segmentation enabled the bank to develop tailored

products and services for each segment, enhancing customer satisfaction and reducing the risk of financial losses (Brown, 2021).

Overall, machine learning has emerged as a powerful tool for enhancing data analysis and customer segmentation, enabling businesses to make informed decisions about marketing, product development, and customer service. By leveraging machine learning to better understand and segment their customer base, businesses can achieve higher sales, lower expenses, and a more satisfied customer base.

Data availability and quality: The quality of the training data determines how well machine learning models perform. Thus, for customer segmentation, it is imperative to have high-quality, pertinent, and unbiased data. This can be difficult, particularly for companies that are resource-constrained or work in complex industries (Hu & Wang, 2022).

Interpretability of the model: Interpreting the results and placing trust in the recommendations can be difficult due to the complexity and difficulty of machine learning models. Businesses with little technical know-how or those who are new to machine learning may find this to be particularly troublesome (Samek et al., 2018).

Algorithmic bias: When trained on data, machine learning models can pick up on and magnify biases. This may result in customer segmentation models that are discriminatory and imprecise. It is critical to recognize this potential bias and take appropriate action to reduce it, such as using debiased data sets and developing bias-resistant algorithms (Lerman et al., 2021).

The application of machine learning to customer segmentation is not without its limitations, in addition to these difficulties. For instance, machine learning models are usually good at finding patterns in past data, but they might not be completely accurate at predicting the behavior of customers in the future. Additionally, creating and maintaining machine learning models can be expensive.

All things considered, machine learning can be a very effective tool for customer segmentation, but it's critical to understand the difficulties and constraints that come with it. Businesses can decide whether and how to use machine learning for customer segmentation by carefully weighing these factors.

2.1 Hierarchical clustering according to the RFM model and using the FCA approach

Hierarchical clustering is an unsupervised machine learning algorithm that defines sets and groups them upon resemblances and alikeness (Hu & Wang, 2022). It works by defining clusters where each cluster is a subdivision of another bigger cluster, the algorithm starts by using each data point as a single cluster and then joining the similar clusters until a single cluster is produced without the use of labeled data (Lerman et al., 2021). Numerous industries, including retail, banking, and telecommunications, have successfully used hierarchical clustering for customer segmentation. (Dolnicar et al., 2014), for instance, employed Ward's linkage in conjunction with hierarchical clustering to determine client groups according to their purchasing patterns.

In the article RFM model is a customer segmentation model that segments customers according to their purchase patterns such as how frequently they buy when the last transaction and how much money is spent on each purchase. According to an article written by Chongkolnee Rungruang, Pakwan Riyapan, Arthit Intarasit, Khanchit Chuarkham, and Jirapond Muangprathub about hierarchical clustering according to the RFM model and using FCA. FCA is a mathematical approach that identifies patterns that are hard to see using normal data analysis techniques. According to the article (Ribeiro et al., 2018) they first calculated the RFM values for each customer, then Using FCA, formed a formal concept lattice. The relationships between the RFM values and the consumers are represented by a hierarchical structure called a formal idea lattice. And lastly, to discover client segments, they used hierarchical clustering to the formal idea frame.The fact that hierarchical clustering is a reasonably easy algorithm to comprehend and use is one of its benefits for customer segmentation. At various granularities, hierarchical clustering can also be used to identify customer groups. Businesses looking to gain a more detailed understanding of their clients may find this useful (Punj & Stewart, 1983). Another advantage of using hierarchical clustering is the fact that no labeled data is required, different levels of granularity can be utilized to identify client groupings using hierarchical clustering, and implicit knowledge can be extracted from consumer data using hierarchical clustering. Implicit knowledge is information that can be deduced from the relationships between data points even though it is not expressly mentioned in the data.

All things considered; hierarchical clustering is an effective technique that can be applied to many client segmentation tasks. Businesses can enhance their client segmentation and marketing efforts by creating their hierarchical clustering solutions by grasping the fundamentals of the technique.

2.2 Density-Based Spatial Clustering Approach

Density-based clustering algorithms such as DBSCAN clustering can detect clusters of any shape, with varying densities, and with robustness against noise (Ester et al., 1996). It has been effectively used in many different industries for customer segmentation (Hossain, 2023).

To begin the process of DBSCAN clustering, every data point is first assembled into a neighborhood graph. In a neighborhood graph, each node denotes a data point, and each edge links two data points that are separated by a certain amount of space, indicated by the epsilon (Eps) parameter (Hossain, 2023).

After the neighborhood graph is generated, DBSCAN searches for densely connected groups of data points to find clusters. If every data point in a group is connected to at least MinPts of other data points in the group, the group is regarded as a cluster. Noise is the term applied to data points that do not belong to any cluster (Yousef & Zhang, 2017). Finding the pertinent customer features to use for clustering would be the first step in using DBSCAN clustering for customer segmentation. For instance, you may make use of attributes like demographic information, past purchases, and online browsing patterns.

After determining which customer features are pertinent, you must select values for the Eps and MinPts parameters. It is recommended to set the Eps parameter to a value that is marginally greater than the mean distance between customer data points. To guarantee that only clusters with an adequate number of customers are formed, the MinPts parameter needs to be set to a sufficiently large value. After determining the values for the Eps and MinPts parameters, you can cluster your customers using the DBSCAN clustering algorithm. Afterwards, customized product offerings, customer service initiatives, and marketing campaigns can be created using the clustering results. DBSCAN clustering is an effective tool for customer segmentation that is resilient to noise, capable of identifying clusters with varying densities, and able to identify clusters of any shape.

According to the article “Customer Segmentation using Centroid Based and Density Based Clustering Algorithms” written by Hossain, a case study was performed for a retail company and three different customer groups were the results of DBSCAN clustering.

One effective method for customer segmentation is DBSCAN clustering. It is robust against noise, able to identify clusters with varying densities, and can be used to identify clusters of any shape. DBSCAN clustering has been applied to retail, telecommunications, finance, and other industries to identify customer segments.

2.3 Gaussian Mixture Model Customer Segmentation Approach

GMM is a complex algorithm that models the data as a mixture of Gaussian distributions each describing a different customer group (Naga, 2023). Different articles and case studies used Gaussian Mixture model for customer segmentation for example in Retail: Based on their purchasing habits, demographics, and other characteristics, retail customers have been divided into groups using GMMs. (Nguyen & Le, 2023) for instance, divided retail customers into four categories using GMMs: high-value, medium-value, low-value, and at-risk customers.Banking: Based on their account activity, risk profile, and other variables, bank customers have been divided into groups using GMMs. For instance(Ghasemi & Zaiane, 2014) divided bank customers into three categories using a hybrid GMM-fuzzy logic model: low-risk, medium-risk, and high-risk customers. elecommunications: Based on their usage habits, demographics, and other variables, GMMs have been used to segment telecom users. (Liu & Zheng, 2022), for instance, divided telecom users into four categories using GMMs with pairwise constraints: high-value customers, According to a previous article written by Naga titled “Customer Segmentation using k-mean and Gaussian Mixture”. The data set the author used for further observation included the customers’ age, demographics, gender, income, purchase history, and purchase information. After working on cleaning the data and pre-processing the data Naga trained the model on the data and came out with a result of four different segments. As a result, Naga analyzed the results of the GMM model, and they were accurate by 70 \% percent.

According to the article one of the downsides of the GMM model is the fact that defining the number of clusters is tricky since the model can be overloaded by the number of clusters. A model that has too many clusters may overfit the data and

perform poorly when applied to fresh data. The model might not be able to fully capture the range of customer behavior if there are too few clusters.

The computational cost of training GMMs, particularly on big datasets, is another drawback. They may therefore be unworkable for applications involving real-time customer segmentation (Nguyen & Le, 2023). Lastly, it can be challenging to interpret GMMs. Because of this, it may be challenging for companies to comprehend the defined client segments and create marketing campaigns that are specifically targeted at them (Camilleri, 2018).

Computational complexity: According to the article, training GMMs on big datasets can be computationally costly. They may therefore be unworkable for applications involving real-time customer segmentation.

Interpretability: According to the article, it can be challenging to understand GMMs. Because of this, it may be challenging for companies to comprehend the defined client segments and create marketing campaigns that are specifically targeted at them (Sundararajan et al., 2017).

Sensitivity to initialization: According to the article, GMMs may react differently depending on the model parameters' starting values. Poor convergence to the actual maximum likelihood solution may occasionally result from this (Lundberg & Lee, 2017).

Normality assumption: According to the article, GMMs operate under the premise that the data is derived from a combination of normal distributions (McLachlan & Peel, 2004).

To conclude Naga’s work Numerous datasets have demonstrated the efficacy of GMMs in the segmentation of customer segments. However, they have certain drawbacks, such as their assumption of normalcy and sensitivity to initialization. GMM training on big datasets can also be computationally costly.

2.4 Self-Organizing Maps Clustering Approach

SOM is an unsupervised machine learning algorithm that can segment and cluster complex, high-dimensional data (Vesanto & Alhoniemi, 2000). "Applying Self-Organizing Maps to Medical Data Analysis" (Kaski et al., 1997)and "Customer Segmentation Based on Self-Organizing Maps: A Case Study on Airline Passengers" (Üstebey et al., 2020) are two articles that discuss SOM clustering. The earlier study (Vesanto & Alhoniemi, 2000) addresses the application of SOMs to medical data

analysis, including patient clustering and disease pattern identification. In the latter work, (Kaski et al., 1997)provide a case study on the use of SOMs to segment airline passengers according to their travel information. SOMs have been specifically applied to customer segmentation in the banking, telecommunications, and retail sectors (Barua et al., 2013). SOMs, for instance, have been used to categorize various retail consumer types, including brand aficionados, price hunters, and infrequent buyers (Ghose et al., 2001). Segmenting bank customers according to their risk profiles and transaction patterns has also been done using SOMs. SOMs have been used in the telecommunications sector to segment customers according to their demographics and usage patterns (Pal & Mitra, 2011).

An article discussed the use of SOMs for medical data analysis. The authors describe how SOMs can be used to cluster patients, identify disease patterns, and develop diagnostic tools. They also present a case study on using SOMs to cluster patients with breast cancer (Kaski et al., 1997). Another article discussed a case study about passengers' Passenger ID, Ticket type Fare type, Travel date, Travel origin, Travel destination, Number of passengers traveling, and Total fare paid. The datasets used to train the SOM model according to the article were Sort of ticket, Travel date and fare type, travel source, and Place of travel. The model resulted in four distinguished customer groups (Üstebey et al., 2020).

The authors conclude that airline companies can segment their customer base and create focused marketing campaigns with the help of SOMs. Airlines might, for instance, target high-value passengers with special offers and discounts using the data from the SOM segmentation. Airlines may also use the data to pinpoint and target customers who are likely to leave with retention campaigns. However, the downside of SOM clustering is that SOMs may react differently depending on how the algorithm is initialized. This implies that the way the algorithm is initialized can affect the SOM segmentation results. The computational cost of training SOMs on large datasets can be high. They may therefore be unworkable for applications involving real-time customer segmentation. And lastly, SOMs can be challenging to understand. Because of this, it may be challenging for companies to comprehend the defined client segments and create marketing campaigns that are specifically targeted at them.Using a variety of initialization techniques and choosing the model that yields the best results is one strategy to address the sensitivity to initialization (Zhang, 2015). Furthermore, various methods can be employed to enhance the comprehensibility of SOMs, including

feature selection and the application of visualization techniques (Vesanto & Alhoniemi, 2000).

The authors conclude that airline companies can segment their customer base and create focused marketing campaigns with the help of SOMs. Airlines can enhance customer satisfaction and create more efficient marketing campaigns by comprehending the primary characteristics that set apart each customer segment.

The authors think that SOMs can be a useful tool for customer segmentation in a range of industries, despite certain drawbacks like their sensitivity to initialization and interpretability issues (Üstebey et al., 2020).

2.5 K-means Clustering Approach

One popular unsupervised machine learning algorithm for customer segmentation is K-means clustering. It does not require any labeled data; instead, it clusters data points together based on how similar they are. Because of this, it's an excellent technique for customer segmentation in situations where customer labels might not be easily accessible (Zhao & Karypis, 2020). K-means clustering is used in customer segmentation to separate out discrete customer groups according to common traits and behaviors (Rungruang et al., 2020). Businesses can improve customer engagement and satisfaction by customizing marketing campaigns, product recommendations, and customer service strategies to target customer segments based on this segmentation (Sharma & Sharma, 2021). A crucial step in k-means clustering is figuring out the ideal number of clusters (k), which controls the level of customer (Singh, 2022). A more granular segmentation is produced by choosing a higher value for k, whereas a broader segmentation is produced by choosing a lower value. Trial and error is frequently involved in selecting the right k value, with the results of various k values being assessed in light of domain knowledge and business goals (Kumar, 2023). The clustering result is significantly influenced by the initialization of the k centroids, or the initial points around which clusters form (Rungruang et al., 2020). While more advanced techniques, like k-means, try to distribute centroids more evenly across the data, improving the clustering process, random initialization can produce less-than-ideal results (Sharma & Sharma, 2021). Calculating the distance between each data point and each centroid is necessary in order to assign each data point to the closest centroid (Zhao & Karypis, 2020). Although Manhattan distance and Euclidean

distance are popular choices, other distance metrics may be applicable based on the specifics of the data.

Recalculating the mean of all the data points assigned to each centroid is necessary to update the centroids. The cluster positions are improved through this iterative process, ensuring that they accurately reflect the central tendency of each cluster (Singh, 2022). To guarantee that the centroids stabilize and the clusters achieve a stable configuration, repeat steps 3 and 4 until convergence (Kumar, 2023). This convergence criterion shows that distinct customer segments have been identified by the algorithm and that the clustering process has reached a local minimum. K-mean clustering was used in a case study of Retail Industry Customer Segmentation A retail business divided its clientele into four groups according to their shopping habits using k-means clustering (Singh, 2022). The four sections that were found were: High-value clients: These are clients who spend a substantial sum of money and make frequent purchases. Medium-value clients: Although they spend a lot more money on each purchase than high-value clients, medium-value clients make fewer purchases overall. Low-value clients: These clients spend a modest sum of money each time they shop and make infrequent purchases. At-risk clients: These clients run the risk of defecting to a rival. Using this segmentation, the business created marketing campaigns that were specifically tailored to each customer group. For instance, the business sent customized coupons and discounts to high-value clients, and emails with news about new products and sales to low-value clients. **A**nother case study that discusses a study about "Banking and Financial Services Industry". Based on their risk profiles, a bank divided its clientele into three groups using k-means clustering (Kumar, 2023). The three sections that were found were: Clients at low risk: These clients are not likely to miss payments on their credit cards or loans. Clients at medium risk: There is a moderate chance that these clients will miss payments on their credit cards or loans. Clients at high risk: These individuals have a higher chance of missing payments on their credit cards or loans. Using this segmentation, the bank created risk management plans that were specific to each group of customers. For instance, the bank provided credit cards and loans with reduced interest rates to low-risk consumers but employed more forceful debt collection tactics with high-risk clients. Based on their usage habits, a telecom company divided its clients into four different groups using k-means clustering (Rungruang et al., 2020). The four sections that were found were: Heavy users: Each month, these clients use a significant quantity of minutes and data. Medium users:

Each month, these clients use a moderate quantity of minutes and data. Light users: These clients utilize a modest monthly quantity of minutes and data. Users who are not active: These clients hardly ever use their phones**.** Using this segmentation, the business created pricing plans that were specifically tailored to each customer group. For instance, the company provided prepaid plans with a limited quantity of data and minutes for light users and unlimited data and minutes for heavy users. K-means clustering has a number of drawbacks that should be taken into account when using it for customer segmentation tasks, despite its widespread use and efficacy. The algorithm's sensitivity to the initialization of the k centroids is one of its primary disadvantages. The clustering outcomes can be greatly influenced by the initial centroids selected, which may result in less-than-ideal segmentation (Sharma & Sharma, 2021). Furthermore, (Zhao & Karypis, 2020)note that k-means clustering makes the assumption that the clusters are spherical, which may not always be the case in real-world data. Inaccurate segmentation can result from this assumption, particularly whhyen working with non-spherical clusters. Additionally, k-means clustering necessitates that the number of clusters (k) be specified beforehand. This can be a difficult task that may call for trial-and-error methods or domain expertise (Rungruang et al., 2020).

2.6 K Nearest Neighbor

K-Nearest Neighbors (KNN) is another widely used unsupervised machine learning algorithm for customer segmentation. It classifies data points based on their similarity to their k nearest neighbors, where k is a user-specified number. Unlike K-means, KNN does not require predetermined cluster centers (Adebayo & Akintola, 2017). This method allows for more flexible cluster shapes than K-means, making it appropriate for non-spherical data distributions. However, KNN has some limitations: Curse of dimensionality: As the number of features (dimensions) in customer data grows, KNN becomes computationally expensive and less effective due to the increased complexity of calculating distances in high-dimensional spaces. Choice of k: Choosing the best k value remains difficult, affecting the granularity and accuracy of the segmentation. Similar to K-means, trial-and-error or domain knowledge may be required to determine the best k for a given customer dataset. Data sensitivity: KNN is sensitive to noisy data points, and outliers can significantly influence the classification of nearby customers. Feature scaling or pre-processing

techniques may be required Reduce this effect. Despite these limitations, KNN is a versatile approach to customer segmentation, especially when dealing with non-spherical data distributions or in scenarios where pre-defined clusters are unavailable (Xu & Wunsch, 2005).

.

Chapter 3

Problem Definition

This thesis aims to evaluate the effectiveness of different customer segmentation methods on a large Turkish market sales dataset. We will compare the performance of K-Means clustering with DBSCAN, GMM, KNN, and lastly SOM to identify the method that best segments customers based on their purchasing behavior and other attributes. This thesis aims to assess which customer segmentation method yields the most insightful results for a large Turkish market sales dataset. To achieve this, we'll follow a series of steps. First, we'll clean and prepare the data, focusing on features that reflect customer buying habits. Then, we'll segment the customers using K-Means clustering, optimizing the number of customer groups. Next, we'll implement an alternative segmentation approach, which is DBSCAN clustering approach following the other techniques. Finally, we'll compare the segmentation results from both methods. This comparison will focus on how well each method defines distinct customer segments based on internal customer similarity and separation between segments. By evaluating these factors, we'll determine the most effective customer segmentation method for this particular dataset.

3.1 Problem Definition and Dataset

There are many different players in the Turkish supermarket industry, which is a dynamic and fiercely competitive market. Turkey's biggest supermarket chains are A101, BIM, CarrefourSA, and Migros. In order to better understand consumer behavior and make more informed decisions about marketing, product offerings, and store operations, Turkish supermarkets are turning more and more to data analytics. A crucial data analytics method for identifying different customer groups based on shared traits and behaviors is customer segmentation. Marketing campaigns, product recommendations, and customer service tactics can then be customized for each segment using the information provided. An abundance of data regarding conumer transactions at a local supermarket in Turkey is included in the dataset (Turkish Market Sales Dataset, n.d.)that was made available for this study. and another dataset was used from kaggle (Mokhtar, 2023) which had less data and that made it easier to work with. Using customer segmentation techniques like k-means clustering, this data

can be utilized to identify different customer segments according to their demographics, purchasing habits, and other perti- nent factors. Then, by using this data, customer satisfaction, loyalty, and overall profitability can all be raised. The dataset helps in creating a clear image of the cutomer. As if to capture a customer's purchasing power, preferred branch, and hometown, each line represents a customer snapshot. Through the categories of "Customer type" and "Gender," we can get a glimpse of their pref- erences; the "Product line" reveals details about their cooking habits and home requirements. The silent duet between each item's "Unit price" and "Quantity" reveals the buyer's spending patterns and propensity for buying in bulk. The last "Total" creates a symphony of their financial commitment, while the "Tax" acts as a constant reminder of the harsh facts of the economy. Dates and times hint at their shopping habits, and "Payment" options offer a touch of digital ease or cash-handling custom. "Cogs" and "gross income" hint at the supermarket's profit margins beyond the icy figures, while "Rating" adds a personal touch, a hint of satisfaction or criticism. With the help of this intricate data tapestry made from the threads of individual transactions, the supermarket will be able to provide laser-focused precision in serving its community by uncovering hidden patterns and unique customer segments.

3.2 Kaggle Second Dataset

The second Dataset and the core of this thesis included columns like "birth year", "Marital status", "Education", "Website visit", "accepted deals", "income", "kids at home" ," fruits purchased" , "wine purchased", "recency", "days since becoming a customer", and so on.

ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	...	NumWebVisitsMonth	AcceptedCn
5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	88	...	7	
2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	1	...	5	
4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	49	...	4	
6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	4	...	6	
5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	43	...	5	

its	...	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response
88	...	7	0	0	0	0	0	0	3	11	1
1	...	5	0	0	0	0	0	0	3	11	0
49	...	4	0	0	0	0	0	0	3	11	0
4	...	6	0	0	0	0	0	0	3	11	0
43	...	5	0	0	0	0	0	0	3	11	0

Figure 1 . Sample of the second dataset

3.3 Definitions

The process of breaking up a customer base into discrete groups according to shared traits and behaviors is known as customer segmentation. Subsequently, distinct marketing campaigns, product lines, and service strategies can be employed to target these segments (Turkish Market Sales Dataset, n.d.).

Chapter 4

Methodology

4.1 Proposed Approach

The first method for customer segmentation for the problem this thesis talks about is K-mean clustering. K-means clustering groups data points according to their similarities into a predetermined number of clusters. First we need to find the optimal cluster number then perform the clustering approach, Investigating different algorithms, such as DBSCAN , GMM, SOM , and lastly KNN.

4.2 Data Preprocessing

Carefully preparing the customer transaction data is essential before beginning the clustering processes. This includes locating and managing missing values, which, based on their importance and the integrity of the data as a whole, may be imputed, erased, or marked as missing such as “income: in the second dataset. To keep them from skewing the clustering results, outliers—data points that differ noticeably from the rest of the data—must also be located and eliminated. Appropriate encoding techniques, like one-hot encoding, label encoding, or binary encoding, are used to transform categorical variables (marital status) into numerical representations that the clustering algorithm can use. Certain attributes in raw data must be cleaned and transformed before they can be used effectively for analysis. Specifically, Birthyear and children at home , teenagers at home: These characteristics appear to be combined, potentially complicating the extraction of family size. To address this, we will take the following data wrangling steps: Extract the birth year from the current year (which is 2024). Use the extracted birth year and possibly other relevant data points later on while clustering the data. Another step is using children at home, teenagers at home to calculate the actual family size. This data cleaning process ensures that the data is in a usable format for further analysis and modeling and also combine the number of accepted campaigns.

4.3 Feature Selection

For customer segmentation to be effective, the right features must be chosen. To enhance clustering performance and prevent cluttering the analysis, features that are redundant or irrelevant were eliminated. Highly correlated features were found using

correlation analysis, and in order to reduce redundancy, features that are too closely related are then eliminated or combined. Techniques for evaluating feature importance, such as feature importance analysis, can rank features according to how much of an impact they have on clustering, removing features that have little bearing.

4.4 Standardization

Standardization is used to make sure that characteristics of varying magnitudes don't unduly affect the clustering process. Common techniques include z-score normalization and min-max normalization, which both convert the features to a common scale with a comparable mean and range. Maintaining the relative relationships between data points while ensuring fair comparison and accurate clustering is made possible by standardizing features uniformly across the entire dataset. StandardScaler from scikit-learn is used to achieve this. The scaler transforms each feature in the data (x) by subtracting a value (potentially the mean) and dividing by another value (potentially the standard deviation).

4.5 K-means Clustering Algorithm

The core of the segmentation procedure is the k-means clustering algorithm. Based on their similarity, it repeatedly divides data points into a set number of clusters (k). Cluster centroids are initialized by the algorithm, either at random or with predetermined values.Our approach was a predetermined number. To Obtain that we used Elbow method. After that, each data point is assigned to the closest cluster centroid based on a distance metric, typically Euclidean distance. The centroids are updated by taking the mean of all the data points in that cluster after the data points have been assigned. The centroids are updated and data points are assigned again until convergence is reached, at which point there is no more noticeable movement in the centroids.

4.6 Determining Optimal Number of Clusters (k)

Determining the ideal cluster count (k) is essential for significant segmentation. The elbow method and silhouette analysis are two popular techniques.

Elbow method: Plotting the within-cluster sum of squares (WCSS) against the number of clusters (k) is the elbow method's method of analysis. The elbow point, where the WCSS begins to drop quickly and then stabilizes, is found to be the ideal number of clusters; this suggests that adding more clusters doesn't appreciably enhance the clustering result. We calculated WCSS for different k values and visualized it to

identify the elbow point manually. Additionally, it leverages the KElbowVisualizer to automate the elbow method visualization, aiding in the selection of the most suitable number of clusters.

Silhouette analysis: determines how similar an object is to its own cluster versus other clusters. The silhouette score ranges between -1 and 1, with a higher score indicating that the object is well matched to its own cluster but poorly matched to neighboring clusters. The silhouette coefficient is calculated for each sample by taking the mean intra-cluster distance (a) and the mean nearest-cluster distance (b). A silhouette score close to 1 indicates that the sample is far away from the neighboring clusters; a score of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters; and a score of -1 indicates that the samples may have been assigned to the incorrect cluster. We calculated silhouette scores for various cluster counts and visualized them to manually determine the best number of clusters. Furthermore, it uses the SilhouetteVisualizer to automate the silhouette method visualization, assisting in the selection of the optimal number of clusters. Silhouette Analysis indicates that the optimal value for k is 3.

4.7 Density Based spatial clustering of Applications with Noise

The density-based method is used by DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Consider a landscape with data points strewn all over it. High-density areas, such as busy city centers, are recognized by DBSCAN as clusters, which are made up of points that are close to one another and have lots of neighbors. Noisy areas are those with few data points, such as rural or suburban areas. DBSCAN looks at two important features of every data point: a distance threshold (Epsilon) and minimum neighbors (MinPts) in order to find these clusters. After fitting the model to the reduced-dimension customer data, we extracted cluster labels for each data point from the fitted model's labels_ attribute. These labels represent cluster membership or -1 for outliers. Then we examined the total number of clusters created, the number of outliers discovered, and the distribution of points within each cluster. A 3D scatter plot depicts the clusters in the reduced-dimension space, with outliers highlighted for clarity. To understand how customer responses differ across segments, we divided customer data into clusters and computed response metrics such as total responses and

response percentages for each cluster. Furthermore, we included functions for creating visualizations that compare the distribution of specific customer attributes (such as income and age) across clusters and outliers. These visualizations can reveal potential differences in customer profiles within each segment, providing useful information for targeted marketing strategies.

4.8 Gaussian Mixture Model

After doing the same steps as K-mean clustering with the data preparation and cleaning, for the GMM model, the code iterates through a range of cluster sizes (1 to 10). It applies a GMM model to the data for each cluster size and computes the BIC score (a model selection metric). It then chooses the number of clusters with the lowest BIC score, indicating the most appropriate number of clusters for the data based on this metric. A GMM model is trained with the optimal number of clusters. The model is then used to predict cluster labels for each customer data point, effectively assigning each customer to a specific segment based on their attributes. The data frame now includes a new column called 'Cluster', which contains the cluster labels. Cluster counts are displayed, indicating the distribution of customers across segments. Principal Component Analysis (PCA) is a technique for visualizing data by reducing its dimensionality. The data points are plotted in scatter plots, with colors representing the various clusters. This allows for a visual inspection of how customers are classified based on their characteristics.

4.9 Self-Organizing Maps

The code clusters customer data using Self-Organizing Maps (SOM), an unsupervised learning technique for visualizing and analyzing high-dimensional data. Initially, it handles missing values and scales numerical features to ensure consistency. The data is then converted to a format suitable for SOM. The SOM grid size is determined by comparing quantization errors across various grid sizes. By plotting these errors, the optimal grid size for minimizing the quantization error is determined. The SOM is trained on the data with the appropriate grid size, and each data point is assigned to a cluster based on the winning neuron. The dataset is then labeled into clusters for further analysis. The segmentation results are analyzed by computing the mean values of different attributes within each cluster, which provides information about distinct customer segments. A scatter plot visually represents the clusters, using

colors to differentiate them based on attributes such as income and recency, allowing for a better understanding of customer behavior and effective marketing strategies.

4.10 K-Nearest Neighbor

After applying the same steps we had to specify a range of possible values for k (number of neighbors). Cross-validation is used to evaluate the accuracy of KNN models with varying k values on training data. Determines the k that produces the highest average cross-validated accuracy and assigns it to optimal_k. Then we had to. Train a KNN model using the optimal_k that you've chosen. Makes predictions about the test set. Produces a classification report that summarizes the model's performance, including precision, recall, and F1-score for each segment. Visualizing the confusion matrix to determine how frequently the model predicted each segment correctly. Lastly, each customer in the dataset (including training and testing sets) is assigned to a cluster based on the KNN model's predictions. Kernel density plots and bar charts are used to analyze the distribution of customer characteristics across identified clusters.

4.11 Customer Segmentation

Customer Segmentation is the main aim of this approach, Every customer is assigned to a cluster according to the k-means clustering results , GMM, SOM, KNN and DBSCAN results, once the ideal number of clusters has been established. Every cluster denotes a unique customer segment that shares traits and buying patterns. Each segment is given a descriptive label that summarizes its salient features in order to increase the segments' meaning and make them easier to identify. In order to confirm that the segments are distinct and meaningful, the segmentation is finally validated by looking at the distribution of pertinent features within each segment.

4.12 Customer Profiling

Detailed customer profiles are created in order to comprehend each customer segment more thoroughly. These profiles include an analysis of each segment's purchasing habits, demographics, and other pertinent details. After closely analyzing these profiles, every consumer segment is identified along with its needs, preferences, and motivations. As a result, personas—archetypal customers within each segment—can be developed, offering insightful information about the factors influencing their purchasing decisions. Personas make it simpler for marketers and business owners to

envision and comprehend the typical customer in each market segment, which in turn makes it easier to adjust messaging and marketing strategies.

Targeted marketing and customer segmentation are ongoing processes that call for constant observation and assessment; they are not one-time events. To maximize outcomes, businesses should evaluate the success of their targeted marketing campaigns on a regular basis and make necessary adjustments. Tracking important metrics like conversion rates, customer satisfaction ratings, and sales numbers broken down by customer segment may be necessary for this. Businesses can make sure that their customer segmentation efforts stay relevant and successful in the ever-changing consumer landscape by consistently improving their approach.

4.13 Cluster Evaluation Metric

Then methods like the Calinski-Harrabasz Index which compares the ratio of the between-cluster variance to the average within-cluster variance, was used for the unsupervised clustering algorithms and another metric which is the Davies-Bouldin that measures the ratio of within-cluster scatter to between-cluster separation was also used on the different methods. For KNN F-1 score was calculated which is The F1-score formula is the harmonic mean of precision and recall.

Chapter 5

Findings

5.1 Data results

We share the outcome of our analysis of the dataset gathered from supermarkets. Our analysis focused on identifying Customer segments and creating their profiles while comparing five different approaches to the same problem. We applied K-mean, KNN, GMM , SOM and DBSCAN clustering .

5.1 First Dataset Using K-mean

After conducting a thorough examination of the dataset, we obtained valuable insights into its structure and content. The initial investigation revealed a dataset containing 1,000,000 rows and 28 columns. Following a critical analysis, specific columns deemed irrelevant were strategically removed to streamline the data for further analysis. Additionally, meticulous data cleaning procedures were implemented, effectively eliminating all null values and duplicate rows. Notably, employing advanced techniques allowed us to successfully impute missing values, guaranteeing a dataset entirely free of any missing information. These comprehensive data preparation steps ensure a robust foundation for subsequent analysis and accurate extraction of meaningful insights. This refined and meticulously cleaned dataset paves the way for the application of various analytical techniques with confidence, ultimately leading to the discovery of valuable and actionable knowledge hidden within the data. We addressed both numerical and categorical variables. Firstly, we calculated the "age" by converting birthdates (in the "USERBIRTHDATE" column) to datetime objects and calculating the difference from today's date as shown in Figure *2*. Two methods are then presented for encoding the "USERGENDER" column: direct string mapping and using a LabelEncoder object to transform values into numerical labels the results are shown in *the figure 3*.

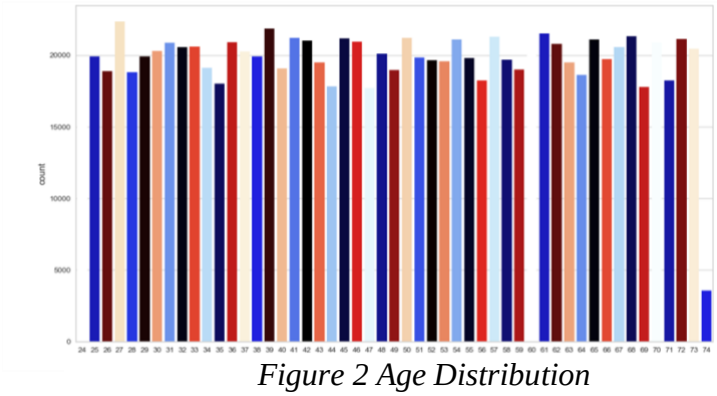


Figure 2 Age Distribution

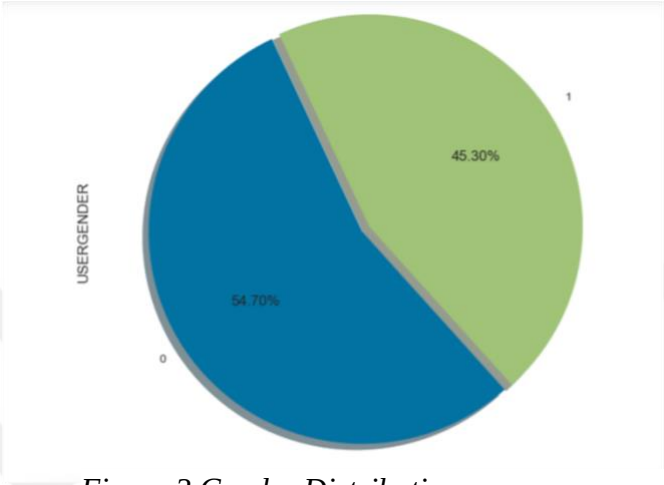


Figure 3 Gender Distribution

Moving on to categorical variables, the code demonstrates LabelEncoder for both "CITY" and "BRAND": unique values are transformed into numerical labels, stored in separate columns, with the option to drop the original categorical columns afterward. Finally, the "CATEGORY1" column undergoes similar label encoding, providing numerical representations for further analysis. These preprocessing steps

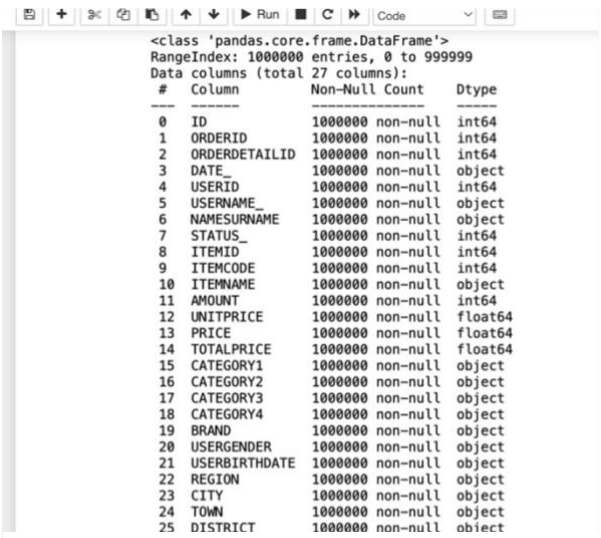


Figure 4 Data Information

ensure your data is well-suited for various machine learning algorithms, paving the way for valuable insights and effective analysis.

Building upon our comprehensive data preparation efforts, we delved into visual exploration as depicted in Figures 5. Employing kernel density plots, we meticulously examined the distribution of each numerical variable within the dataset. These insightful visualizations provided invaluable information about the spread and shape of the data, enabling us to glean crucial insights into the underlying characteristics of each variable.By observing the plots, we could readily identify variables with skewed distributions, potential outliers, and the overall concentration of data points within specific ranges. This visual inspection served as a crucial step in understanding the nature of the data and guiding further analysis strategies. The ability to discern patterns and anomalies within the data distribution equips us with the necessary knowledge to make informed decisions regarding subsequent analytical techniques and ultimately derive meaningful conclusions from the dataset.

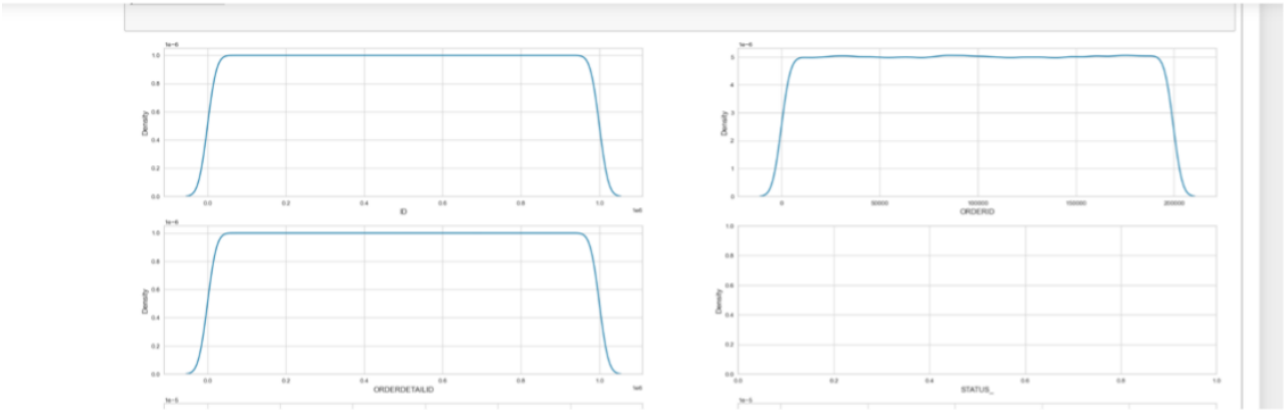


Figure 5 Distributions of numerical columns

Then as shown in Figure 6 to 12, we visualized the distributions of numerical variables within the dataset. Each numerical column is allocated a dedicated subplot, showcasing its distribution through a combined kernel density estimation (KDE) and histogram. Tailored aesthetics with black KDE lines, red histograms, and clear titles enhance readability. By observing these plots, you gain valuable insights into data spread, shape, and potential outliers, laying the groundwork for informed analytical

decisions and meaningful conclusions.

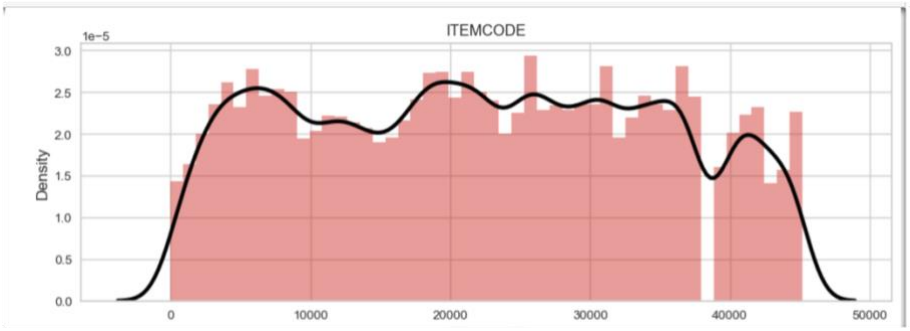


Figure 6 Item Code Density

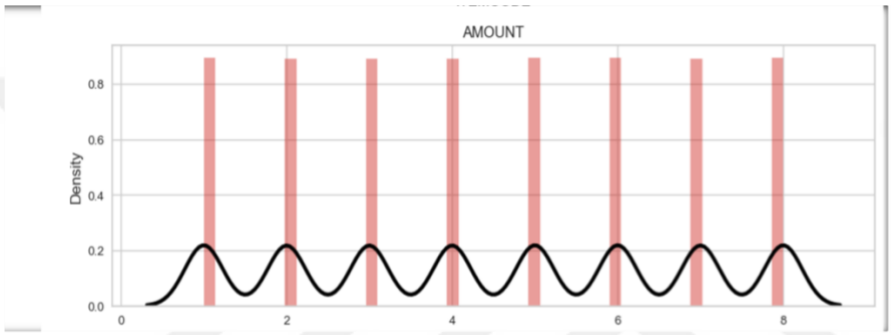


Figure 7 Amount Density

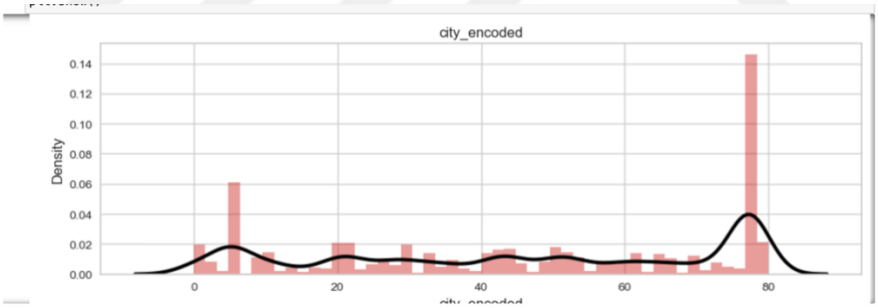


Figure 8 City Density

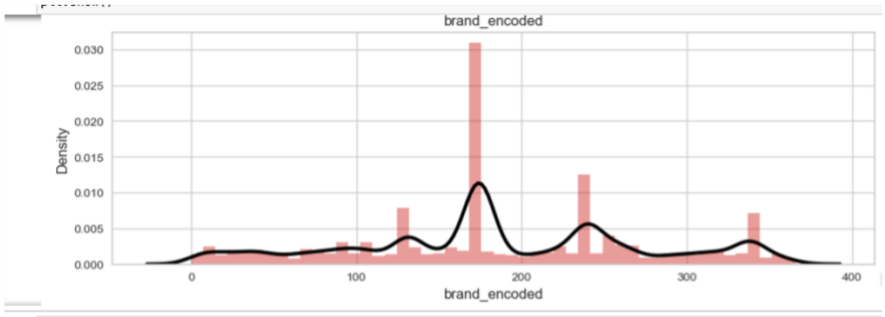


Figure 9 Brand Density

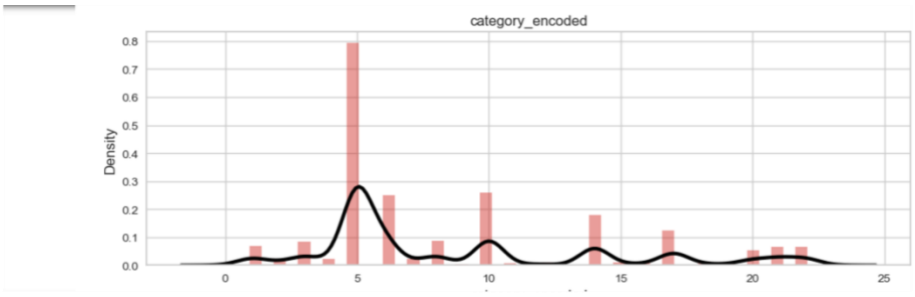


Figure 10 Category Density

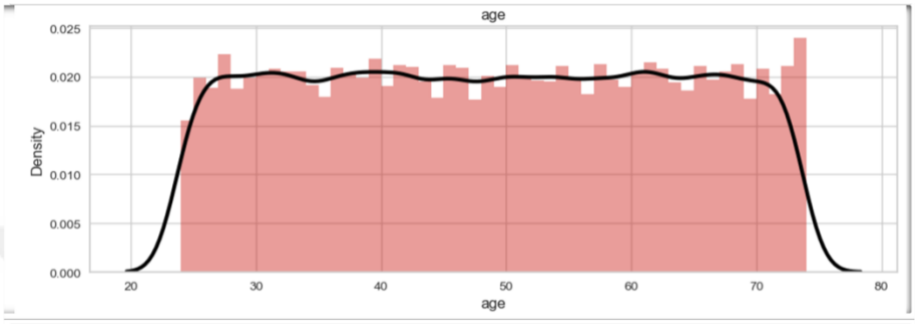


Figure 11 Age Density

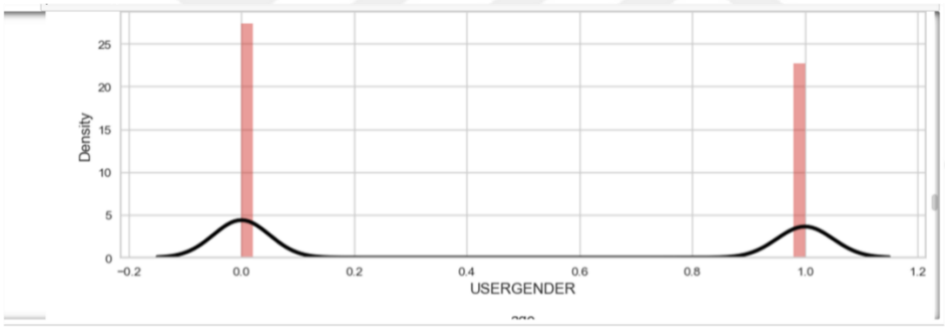


Figure 12 Gender Density

After that we worked on a captivating heatmap the results are shown in figure 13. Imagine a canvas, 15 inches by 15 inches, ready to be painted with the intricate tapestry of correlations between your variables. The sns.heatmap function serves as the artist, meticulously calculating the correlation matrix, a numerical representation of how closely each variable moves in tandem with others. Warmer hues on the heatmap depict strong positive correlations, like income and spending rising together, while cooler colors signify variables that tend to move in opposite directions, such as price increases leading to sales dips. Even white patches emerge, highlighting the absence of any significant linear relationship. But the masterpiece doesn't stop there. Numbers dance across the squares, revealing the precise correlation values, inviting you to delve deeper into the story each color tells. This exploration empowers you to grasp the

structure of your data, guiding further analysis and potentially unearthing hidden gems of understanding

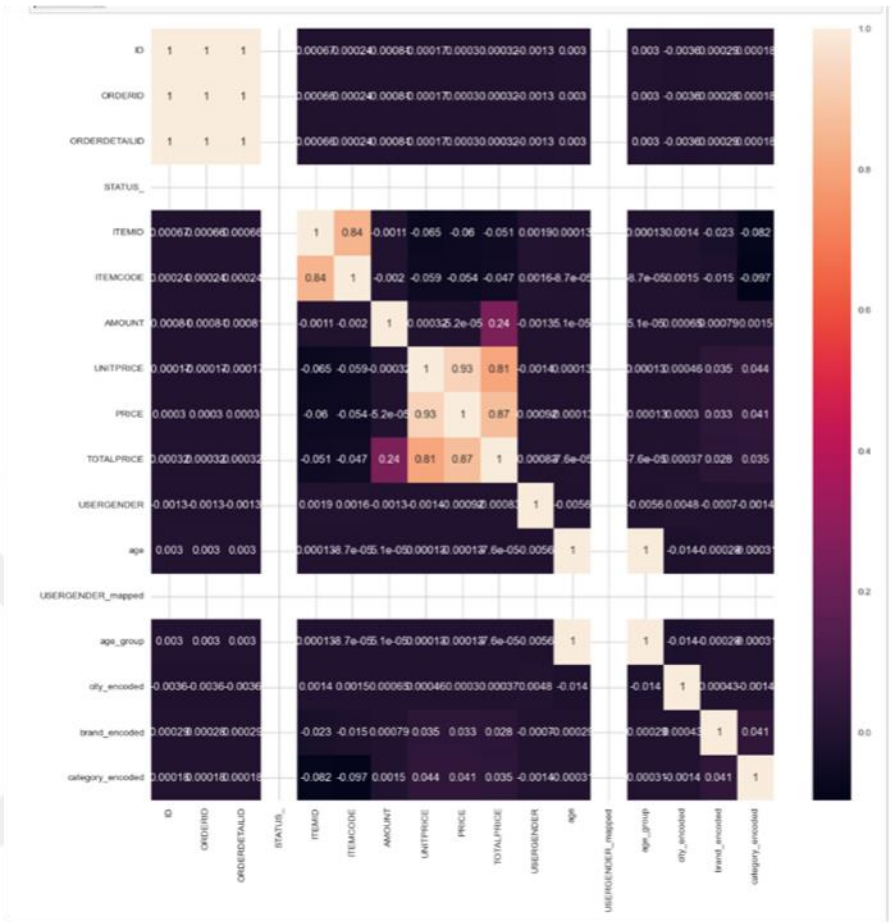


Figure 13 Heat Map

Then we discovered the ideal number of clusters hidden within your data, guided by the insightful "Elbow Method". As the first step, we ensure all numerical features are treated equally by meticulously "normalizing" them using a technique called standardization. This prevents any single feature from exerting undue influence during the clustering process.

Next, we delve into the heart of the Elbow Method. We embark on a series of trials, experimenting with different numbers of clusters, ranging from 1 to 14. For each attempt, we create a KMeans clustering object, carefully grouping your data points into the specified number of clusters. At each step, we capture a crucial value called "inertia", which provides a measure of how well data points fit within their assigned clusters. By diligently collecting these inertia values for each clustering attempt, we begin to piece together the puzzle. The results were 6 optimal number of clusters as shown in Figure 14.

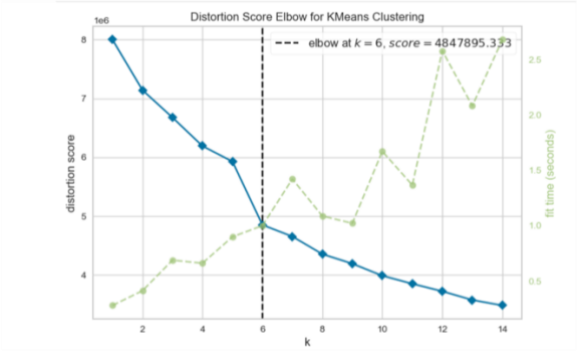


Figure 14 Elbow Method Results

Then a K-Means model, configured with six predetermined clusters, analyzes each data point based on their similarities. Based on this analysis, K-Means assigns each point to one of the six clusters, essentially labeling them with unique identifiers. To visualize the clusters effectively, a new DataFrame, pca_data_kmeans, is created. This DataFrame merges the informative dimensions extracted by PCA (think of it as condensing the data into a more manageable format) with the cluster labels assigned by K-Means. A scatter plot emerges, where each data point represents a distinct member of a cluster. Different colors are assigned to each cluster, allowing you to visually distinguish the various groups within your data as shown in Figure 15.

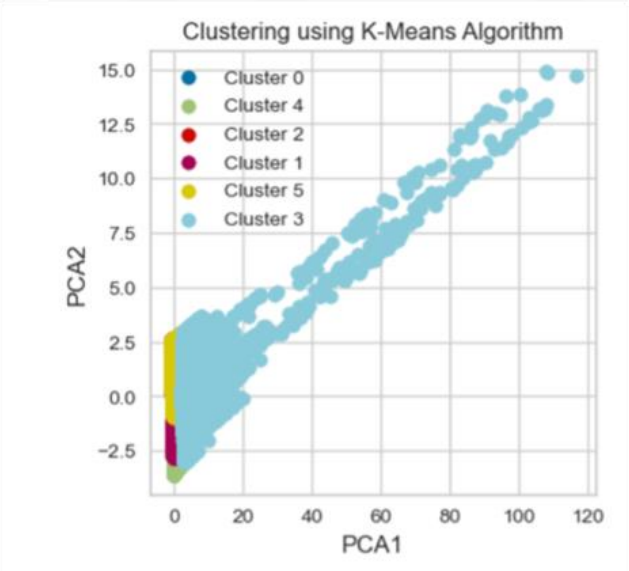


Figure 15 K-mean Cluster

Titles, labels, and a legend adorn the plot, ensuring clarity and understanding. This visual tapestry, woven from the combined efforts of K-Means and PCA, offers valuable insights into the underlying structure of your data. You can observe how data

points within each cluster distribute themselves, revealing potential groupings and unique characteristics. Then we merged the original data with cluster labels assigned by $K-Means$, creating a new cluster data DataFrame. This enrichment allows us to analyze features within each cluster, compare characteristics across groups, and gain deeper insights into the underlying data structure. It's like adding a "group membership" tag to each data point, unlocking new exploration possibilities.

Then we first ensured all data points had valid cluster labels. It then counts the members in each cluster, visualizing the distribution through a bar chart as shown in Figure 16.

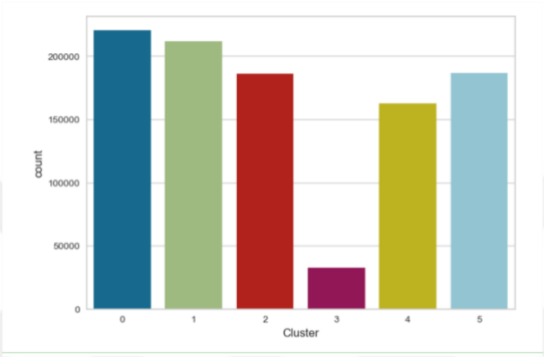


Figure 16 Clusters

Observe which clusters are bustling with data points and which seem deserted - this imbalance could offer clues about the data's structure and potential outliers.

Then we measure the quality of your K-Means clusters, employing the Davies-Bouldin Score (DB score) as its measuring stick. Think of it like a judge assessing the separateness of your clusters.

First, it imports the davies_bouldin_score function to perform the evaluation. Then, it calculates the score using your preprocessed data and the assigned cluster labels. Finally, it presents the score, where lower values less than 2.5 signify well-separated clusters and higher values indicate overlapping or poorly defined groups in the results as shown in Figure 17 the results were less than 2.

```
Davies-Bouldin Score: 1.6020
```

Figure 17 DB Score

5.2.1 Second Dataset Using K-mean

After retrieving the data and reading the csv file on a new jupyter environment the output of the data information and type is shown in *figure 18.*

```
data.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 2240 entries, 5524 to 9405
Data columns (total 22 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Education            2240 non-null   object
 1   Marital_Status       2240 non-null   int64
 2   Income               2216 non-null   float64
 3   Recency              2240 non-null   int64
 4   MntWines             2240 non-null   int64
 5   MntFruits            2240 non-null   int64
 6   MntMeatProducts      2240 non-null   int64
 7   MntFishProducts      2240 non-null   int64
 8   MntSweetProducts     2240 non-null   int64
 9   MntGoldProds         2240 non-null   int64
 10  NumDealsPurchases    2240 non-null   int64
 11  NumWebPurchases      2240 non-null   int64
 12  NumCatalogPurchases  2240 non-null   int64
 13  NumStorePurchases    2240 non-null   int64
 14  NumWebVisitsMonth    2240 non-null   int64
 15  Complain             2240 non-null   int64
 16  Response             2240 non-null   int64
 17  Age                  2240 non-null   int64
 18  Days_Since_Customer  2240 non-null   float64
 19  Fam_Size             2240 non-null   int64
 20  Num_Accepted         2240 non-null   int64
 21  MntTotal             2240 non-null   int64
dtypes: float64(2), int64(19), object(1)
memory usage: 402.5+ KB
```

Figure 18. Data Information

For a better understanding of the data we did a visualization of the marital status as shown in *figure 19,* The married or couple status of the customers is more dominant as seen by the figure and also the income range distribution which is mostly belowe 75k as shown in *figure 20.*

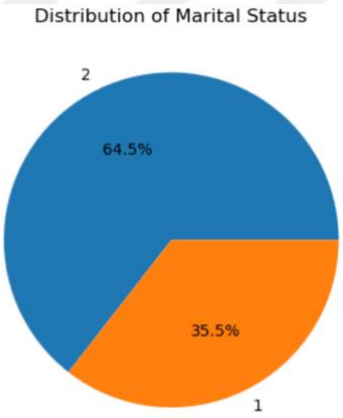


Figure 19. Distribution of Marital Status

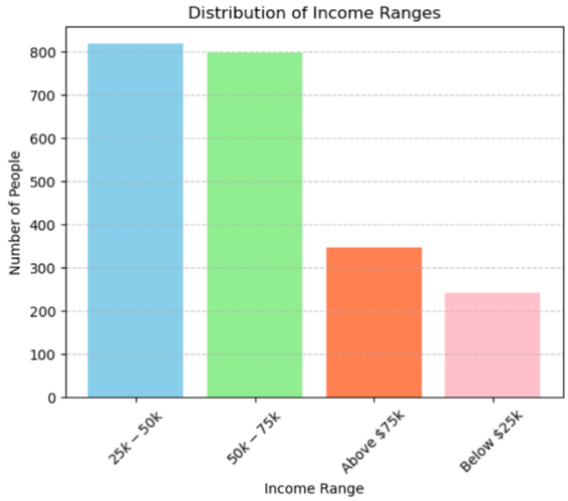
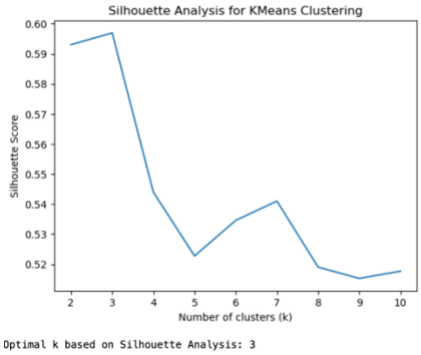


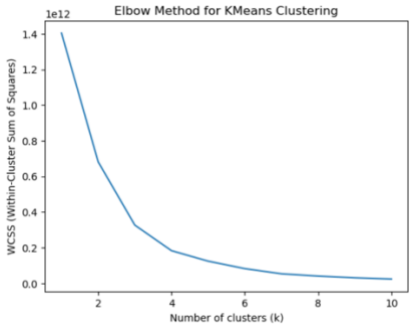
Figure 20 Distribution Of Income Range

Then I performed a Silhouette analysis and Elbow method, which gave the optimal number of clusters as 3 this helps on setting the optimal number of k as shown in *Figure 21* and *Figure 22* .




```
Optimal k based on Silhouette Analysis: 3
```

Figure 21 Silhouette Analysis



Then according to the silhouette analysis and Elbow method, I performed K-means clustering and the results were 3 clusters as shown in *Figure 22* Unlike GMM and SOM but the same as KNN and DBSCAN clustering algorithms .

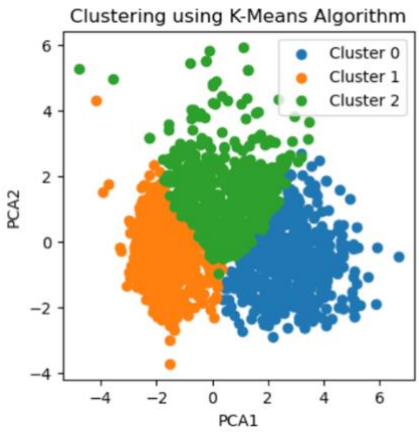


Figure 22 Clustering Using K-mean Method

Later we visualized the cluster count to see which customer group has the most customers; cluster 1 has the most count of customers this helps us define the dominant customer group of the supermarket as shown in *Figure 23*.

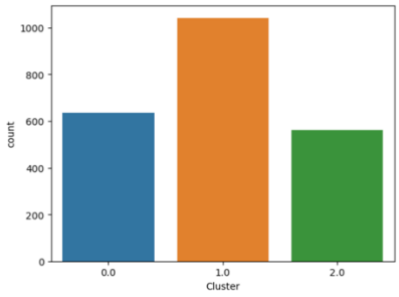


Figure 23 Cluster Count

Each cluster was visualized according to the results in terms of each feature the Figures from 24 till figure 33 all represent different features of the clusters that we can extract useful information from.

Figure 24 shows the income range of each cluster group cluster 0 has the highest income range this corresponds to cluster 1 of KNN this shows that K-mean distributed the clusters with different income ranges .

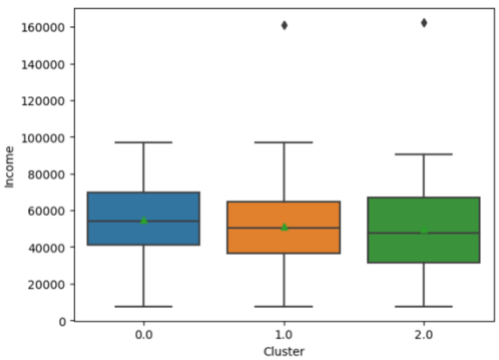


Figure 24 Income

Figure 25 shows the number of purchases made through the website cluster 1 and 2 purchase from the website more frequently than cluster 3 which shows similar results in cluster 1 and 3 in KNN method meanwhile is not definite in SOM and GMM clustering algorithms .

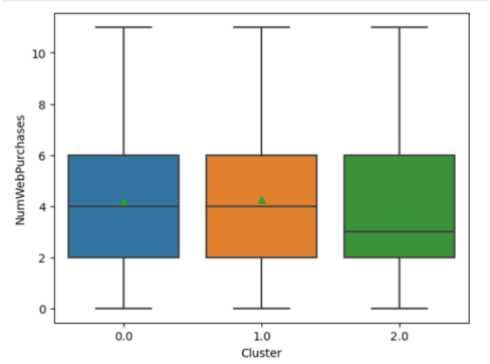


Figure 25 Website Purchase

Figure 26 shows the number of purchases made through the deal of the supermarket c;uster 1 and 2 are responsive to deals like cluster 1 and 3 in KNN approach which shows that KNN and Kmean .

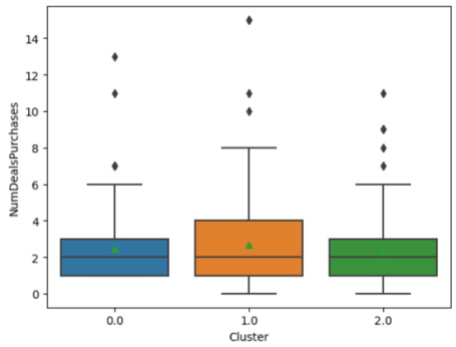


Figure 26 Deal Purchase

Figure 27 shows the number of purchases made through the catalog mostly cluster 0 showed the best response which corresponds to the second cluster in DBSCAN clustering approach .

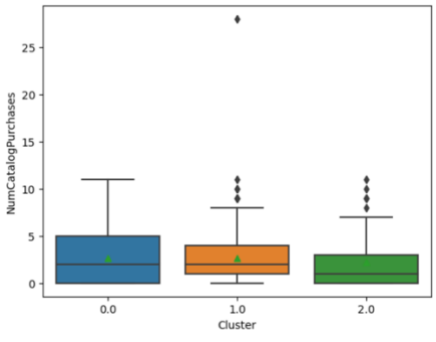


Figure 27 Catalog Purchase

Figure 28 shows how recently each cluster bought an item from the supermarket the recency the second cluster has the highest recency date which corresponds to cluster 2 in KNN method..

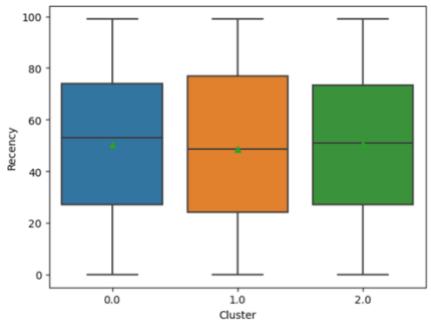


Figure 28 Recency

Figure 29 shows the number of purchases made through the store clusters 0 and 1, which show the highest purchase rate from the store. Using K-mean we can see that each cluster is different than the other in terms or purchasing behavior.

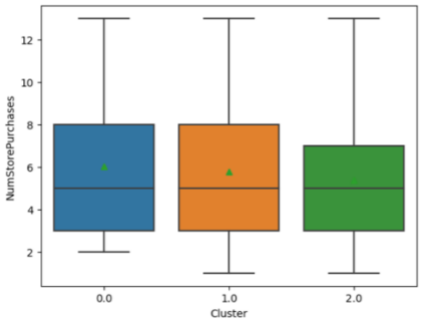


Figure 29 Store Purchase

Figure 30 shows the number of visits through the website which corresponds to Cluster 9 in GMM method. This can show the patterns in which each cluster uses the supermarket website ,and can be used for further marketing strategies.

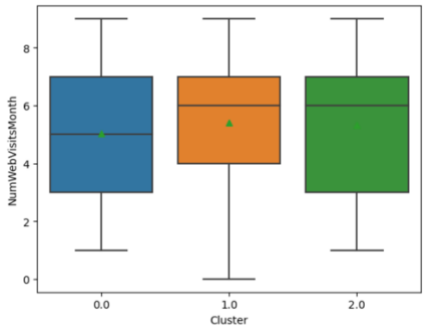


Figure 30 Website Visit

Figure 31 shows the number of days since the member became a customer of the supermarket within each cluster the oldest group are cluster 1 and 3 which are considered loyal customers.

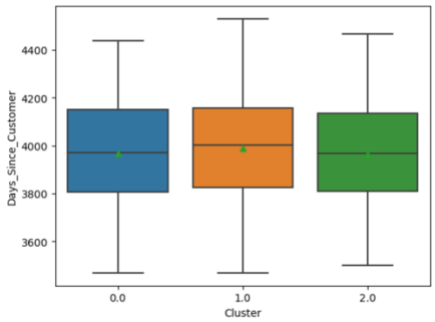


Figure 31 Days Since Becoming A Customer

Figure 32 shows the number of members of the families of the customers in each cluster the families are mostly 2 or 3 members in all the clusters so we can see that the family size was not a great feature for clustering the customers since most clusters have family size of 1 up to 4 members.

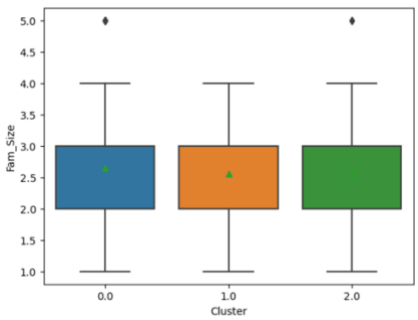


Figure 32 Family Size

Figure 33 shows the marital status of the members within each cluster we can see that all clusters show a mix of marital status between married and single this can prove that k-mean didn't use marital status as a feature that can affect the clustering behavior unlike other methods (SOM) .

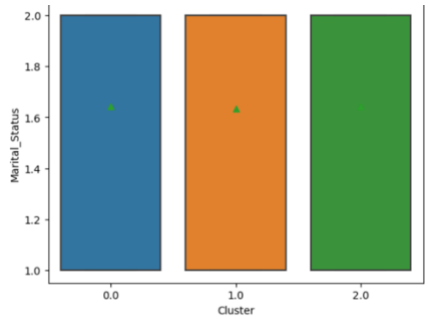


Figure 33 Marital Status

After that, I calculated the DB score and CH score and the results of Davies-Bouldin Score, are as shown in Figure 34 and Calinski-Harabasz Index score is as shown in figure 35 K-mean showed a good DB score which is relatively low but the lowest sco re was noted by SOM and DBSCAN method which puts K-mean the third lowest DB score. Meanwhile Calinski-Harabasz score was relatively high similar to SOM score .

```
Davies-Bouldin Score: 1.8500
```

Figure 34 Davies-bouldin Score

```
Calinski-Harabasz Index score: 617.3290444095892
```

Figure 35 Calinski-Harabasz Score

5.2.2 Second Dataset Using DBSCAN

The same steps were taken for DBSCAN for data preparation and visualization, the next new step is looking at the data in 3D which is shown in Figure 36.

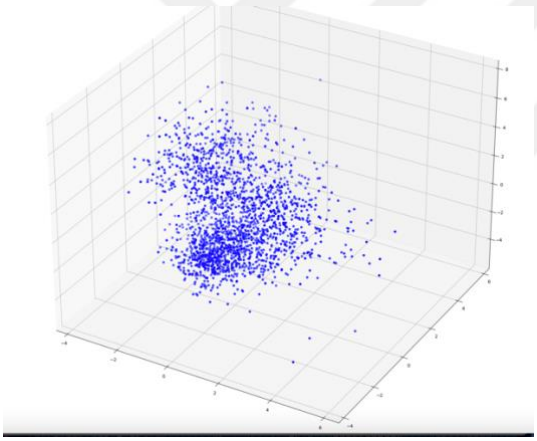


Figure 36 3D Data

The results of DBSCAN approach are as shown in *Figure 37*.

```
Cluster Predictions
Number of clusters: 3
Cluster 0: 649
Cluster 1: 20
Cluster 2: 17
```

Figure 37 Cluster Outcome

Then the visualization of the cluster groups is shown in the *Figures 38* up to *figure 50*. Figure 38 cluster shows the total spending score of each cluster.

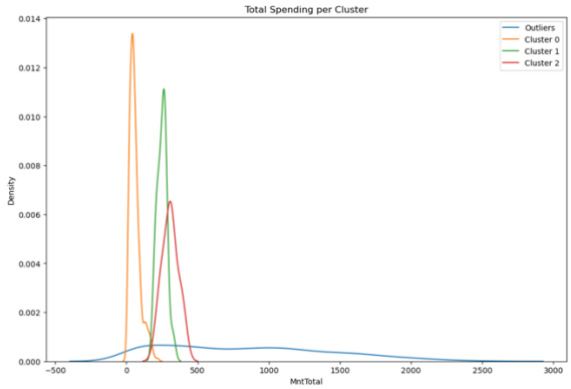


Figure 38 Total Spending Per Cluster

Figure 39 illustrates the family member size of each cluster which is mostly 2 to 3 members in each cluster but cluster 0 has the highest number of 3 members within the family which shows that unlike K-mean DBSCAN took into consideration the family size.

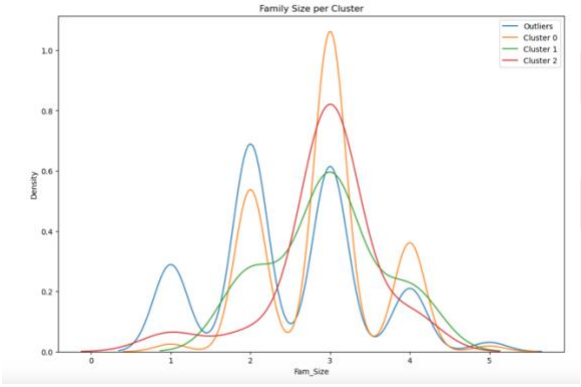


Figure 39 Family Size

Figure 40 shows the deal purchase per cluster the most responsive cluster is cluster 2. And the least responsive is cluster 0 DBSCAN took the deal purchase attribute as a big factor that influences the clustering of customers since we can see big variation between the cluster response as shown.

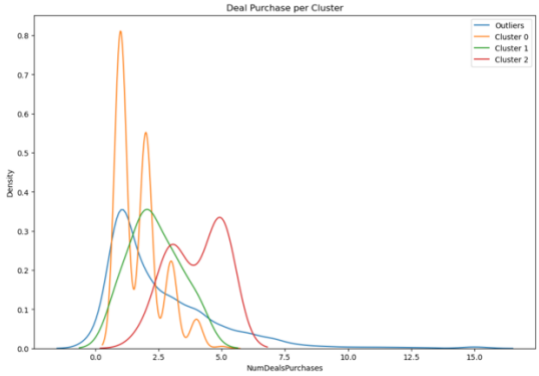


Figure 40 Deal Purchase Per Cluster

Cluster 41 shows the store purchase within each cluster the highest is cluster 2 this shows that this customer group are considered an group and are similar to Clutser 7 in SOM clustering technique and cluster 6 in GMM.

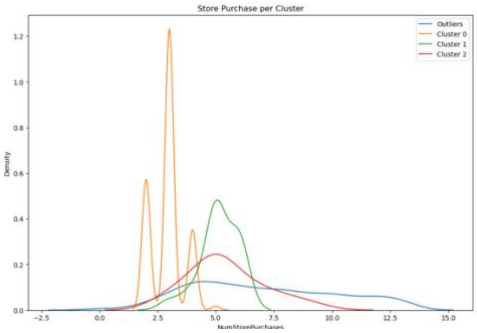


Figure 41 Store Purchase

Figure 42 shows the website visit per cluster which is used in the purchase analysis the highest website visit is from cluster 0 there is no big difference between the clusters so this attribute can be considered the same as family size in K-Mean not that influencing on clustering the customers.

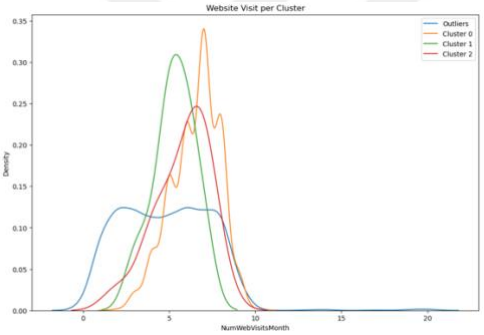


Figure 42 Website Visit Per Cluster

Figure 43 shows the catalog purchase within each cluster. Cluster 2 has the highest catalog purchase rate the values differ between the clusters so DBSCAN took into consideration this attribute like SOM and GMM methods but unlike KNN method .

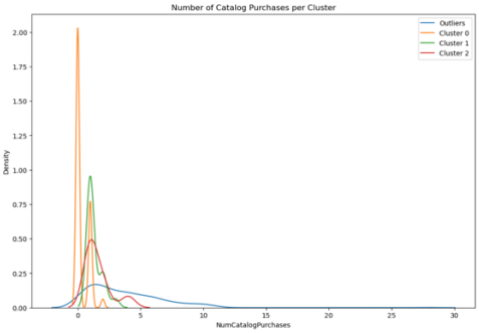


Figure 43 Catalog Purchase

Figure 44 shows the website purchase per cluster , cluster 2 shows the best outcome for the purchases made through a website they are considered the active customers

through the website and the high spenders . We can see that DBSCAN focus son the purchase behavior and patterns to define the clusters.

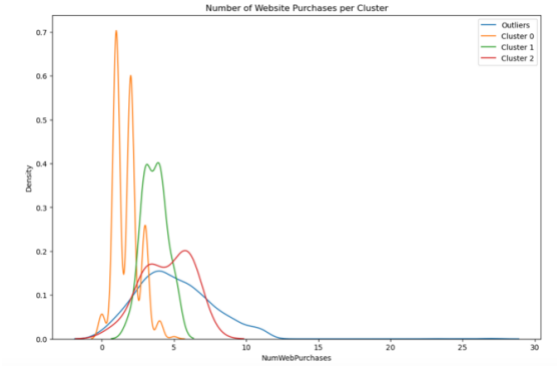


Figure 44 Website Purchase

Figure 45 shows the age group each cluster mostly belongs to, cluster 2 is mostly middle-aged, but cluster 1 is mostly in retirement age most customers are relatively close in age.

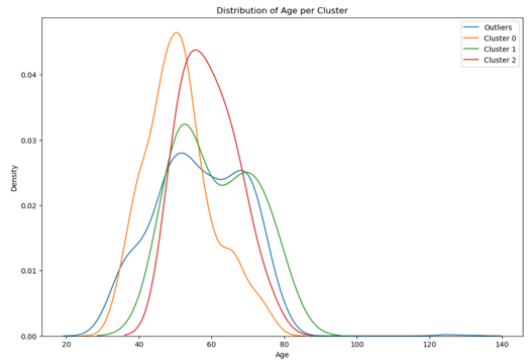


Figure 45 Age

Figure 46 shows the number of accepted campaigns that each cluster responded to, the highest response came from cluster 2 which is also shown from their spending patterns and deal, website, and store purchases behavior proven by the previous results .

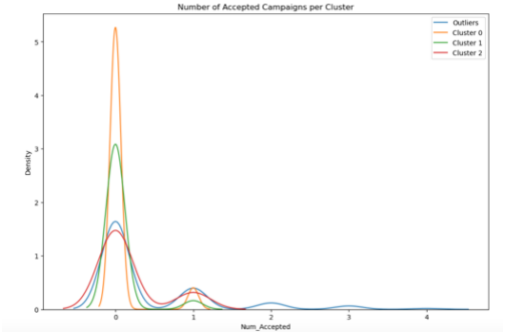


Figure 46 Number of Accepted Campaigns

Figure 47 shows the income range of each cluster. Cluster 0 shows the least income range out of the clusters .

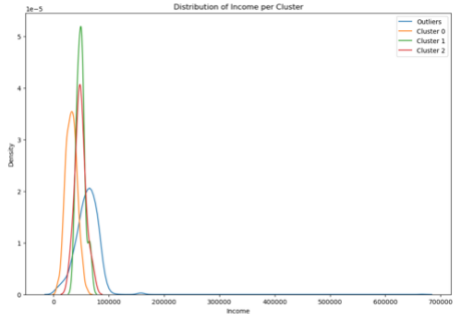


Figure 47 Income Per Cluster

Figure 48 shows how recent the members of each cluster bout an item from the supermarket and the most recent is from cluster 1 meanwhile the furthest is from cluster 0.

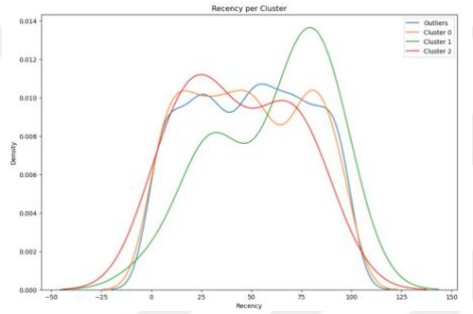


Figure 48 Recency

Figure 49 shows the number of days since becoming a member and customer of the supermarket, cluster 2 shows the most furthest date but there is no big difference between the customers which shows that DBSCAN focuses on the purchasing behavior while slightly showing influence by other attributes like days since beoming a customer.

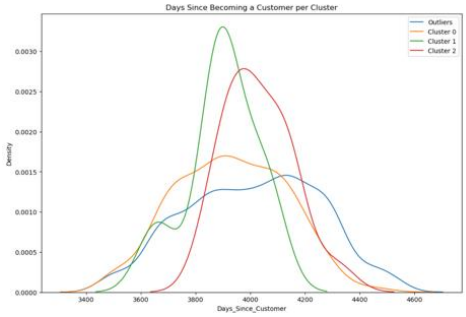


Figure 49 Days Since Becoming a Customer

Figure 50 shows the education level each cluster has the highest number of, cluster 1 is mostly PhD and master holders, meanwhile, cluster 0 is mostly graduates and PhD holders and cluster, lastly, cluster 2 is mostly graduates and master graduates.

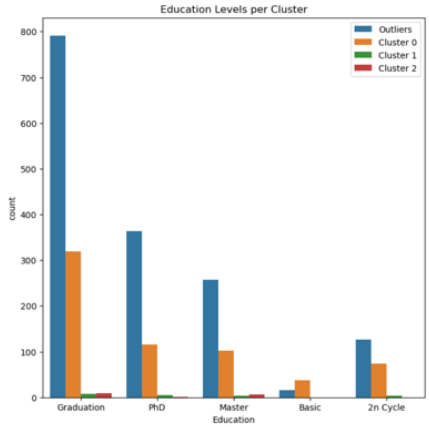


Figure 50 Education Level

Then I calculated Calinski-Harabasz and Davies-Bouldin score as shown in *Figure 51* and *Figure 52* DBSCAN shows the second lowest score of Davies-Bouldin and a relatively better score than the other methods for CH score .

```
Calinski-Harabasz Index score: 302.33947778652623
```

Figure 51 Calinski-harabasz Score

```
Davies-Bouldin score: 1.255466045205454
```

Figure 52 Davies-Bouldin Score

5.2.3 Second Dataset Using GMM

After doing the same steps as DBSCAN and k-mean, I had to find the optimal number of clusters using BIC as shown in *Figure 53*.

```
Optimal number of clusters: 10
```

Figure 53 Optimal Number of clusters

Then I had to perform the GMM approach of the clustering analysis problem, after performing the method using the optimal number of clusters, I visualized the output as shown in *Figure 54*.

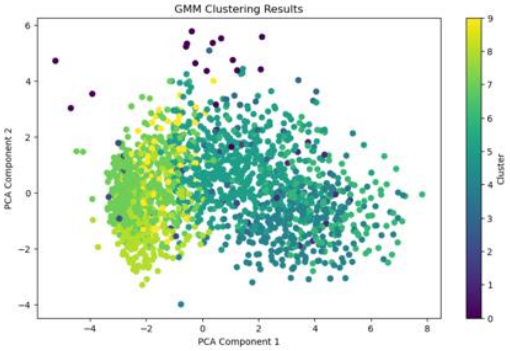


Figure 54 GMM Clustering Results

From *Figure 55* to *Figure 68* are the results of each cluster analysis from the GMM clustering method according to the features.

Figure 55 shows the various spending scores of the 10 clusters. Cluster 6 shows the most spending rate out of the other clusters which makes the customers of that cluster group the most spenders in the supermarket and in the active group the values of the spending behavior vary largely between the customers in the analysis.

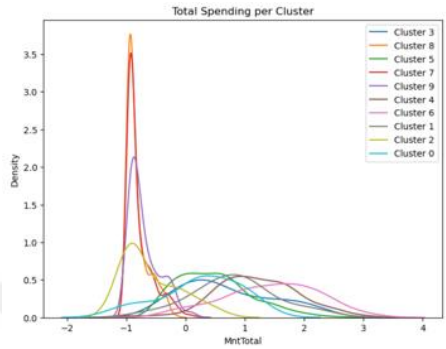


Figure 55 Total Spending Per Cluster

Figure 56 shows the family size of the 10 clusters, the outcomes vary from 2 to 3 family members. Cluster 0 had 4 members which represented the customer group that have a big family meanwhile the other clusters have a moderate family size such as cluster 7 averaging with 2 family members .

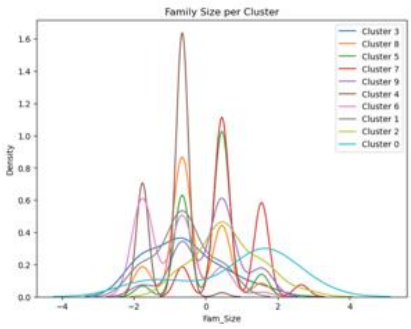


Figure 56 Family Size

Figure 57 notes down the deal purchase that each cluster responded to leaving cluster 0 to be the most cluster that bought from deals. The biggest response was shown in cluster 0 leaving these customers to be the accepters of deals of that supermarket. GMM is one of the few methods that grouped customers who didn't show any response to deals as a group by themselves.

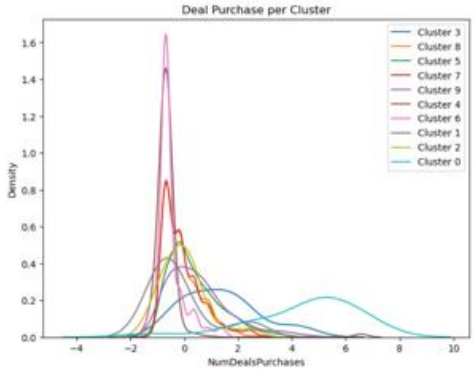


Figure 57 Deal Purchase

Figure 58 shows the store purchase of each cluster showing cluster 1 as the cluster that bought from the store. Meanwhile, cluster 7 were the least customers who bought from the store, this helps in analyzing the spending patterns of the customers. We can also see that there is a customer group with zero store purchases.

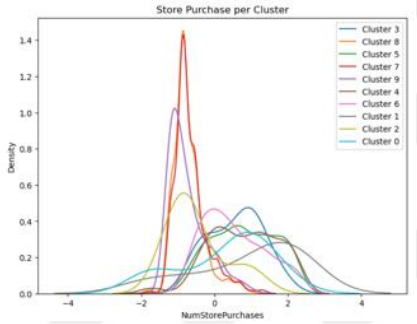


Figure 58 Store Purchase Per Cluster

Figure 59 shows the website visit per cluster, the highest website visit was from cluster 0. And the lowest was from cluster 4. Cluster zero also previously showed the most deal purchase.

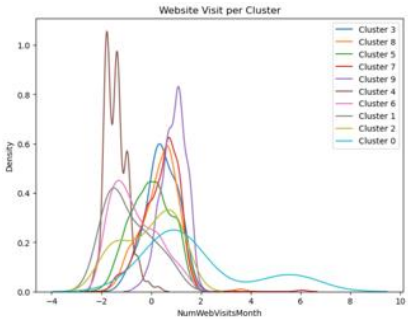


Figure 59 Website Visit Per Cluster

Figure 60 shows the catalog purchase per cluster which showed a good response from mostly all clusters but especially cluster 3 and cluster 6 since they are not similar to the results of store purchase we can see that the values of catalog purchases are not treated only as spending scores in this method.

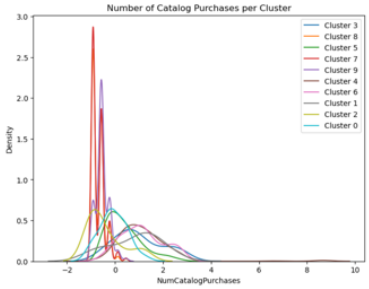


Figure 60 Catalog Purchase Per Cluster

Figure 61 shows the website purchase by each cluster putting cluster 1 at the top. The results of the website, catalog and deal purchase can show the GMM method used in these purchases pattern to define the clusters. These values reflect strongly on the customer divisions.

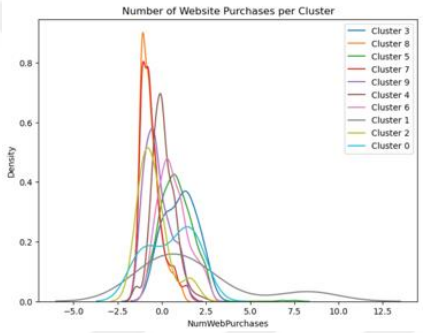


Figure 61 Website Purchase Per Cluster

Figure 62 shows the age group each cluster belongs to the oldest group is cluster 2 meanwhile the youngest are cluster 9 and 7 but most customers of the supermarket are middle aged customers.

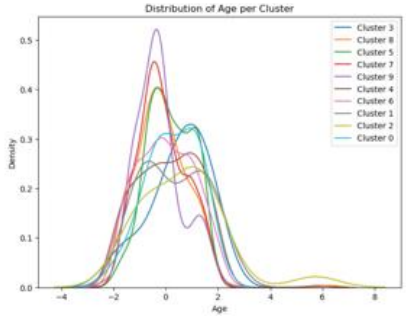


Figure 62 Distribution of Age Per Cluster

Figure 63 shows the number of accepted campaigns by each cluster outing cluster 6 with the highest acceptance rate.

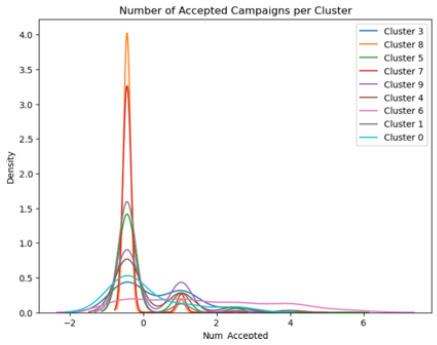


Figure 63 Accepted Campaigns per Cluster

Figure 64 shows the income distribution of each cluster leaving cluster 2 with the highest income range leaving these customers to be the richest out of the other customer groups.

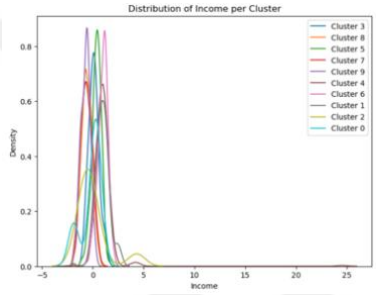


Figure 64 Income Distribution

Figure 65 shows the recency in which each cluster recently bought an item from the store, this helps in analyzing the spending scores and purchase rate the most recent group and with the highest number of customers was cluster 4.

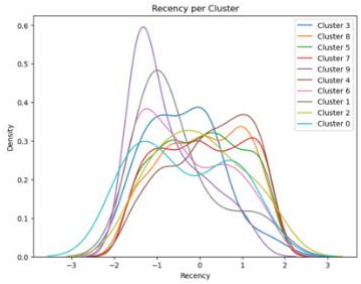


Figure 65 Recency

Figure 66 shows how many days ago each customer became a customer of that store cluster 0 show the furthest date making them the oldest customer group of that supermarket.

Figure 66 Days Since Becoming a Customer

Figure 67 shows the education level of each cluster most of the clusters are graduates but cluster 7 shows a high number of customers who have a PhD degree all of the clusters have of each education level so we can note that education level is not a huge factor in clustering the customers but the information can be helpful in analyzing each cluster.

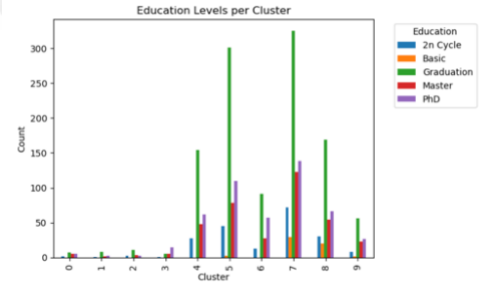


Figure 67 Education Levels per Cluster

After analyzing the results, I applied the two cluster analysis methods, the Davies-Bouldin score and the Calinski-Harabasz score as shown in *Figure 68* and *Figure 69*. The DB score is considered relatively high compared to the scores of the other methods.

```
Calinski-Harabasz Index score: 184.92106309530934
```

Figure 68 Calinski-Harabasz Score

```
Davies-Bouldin score for GMM: 2.2875329080832243
```

Figure 69 Davies-Bouldin Score

5.2.4 Second Dataset Using SOM

After applying the same methods and techniques as K-mean and DBSCAN for data preprocessing and cleaning I had to find the optimal grid size as shown in *figure 70* using quantization error.

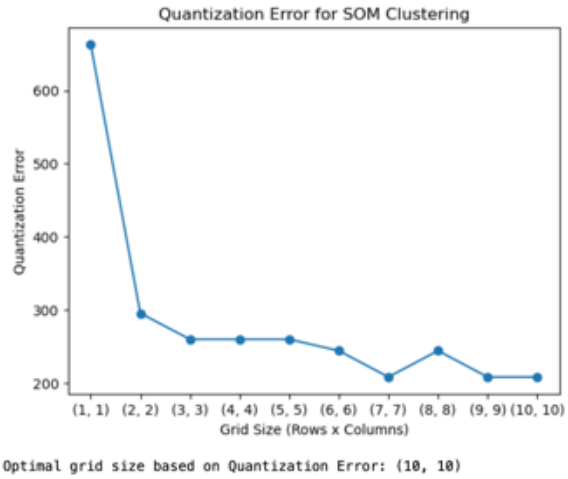


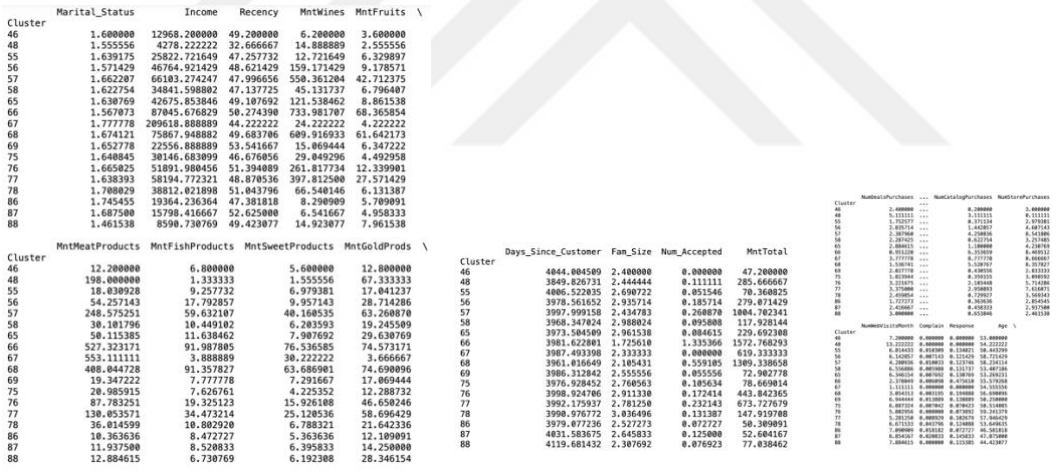
Figure 70 Optimal grid size

The number of clusters is as shown in *figure 71*.

```
Number of clusters identified by SOM: 17
```

Figure 71 Number of SOM Clusters

The results of the clusters of the SOM clustering method is as seen in *figure 72*.



Cluster	Marital_Status	Income	Recency	MntWines	MntFruits
46	1.600000	12968.200000	49.200000	6.200000	3.600000
48	1.555556	4278.222222	32.666667	14.888889	2.555556
55	1.639175	25822.721649	47.257732	12.721649	6.329897
56	1.571429	46764.921429	48.621429	159.171429	9.178571
57	1.662207	66103.274247	47.996656	550.361204	42.712375
58	1.622754	34841.598802	47.137725	45.131737	6.796407
65	1.630769	42675.853846	49.107692	121.538462	8.861538
66	1.567073	87045.676829	50.274390	733.981707	68.365854
67	1.777778	209618.888889	44.222222	24.222222	4.222222
68	1.674121	75867.948882	49.683706	609.916933	61.642173
69	1.652778	22556.888889	53.541667	15.869444	6.347222
75	1.640845	30146.683099	46.676056	29.049296	4.492958
76	1.665025	51891.980456	51.394089	261.817734	12.339901
77	1.638393	58194.772321	48.870536	397.812500	27.571429
78	1.700829	38812.021898	51.043796	66.540146	6.131387
86	1.745455	19364.236364	47.381818	8.290909	5.709091
87	1.687500	15798.416667	52.625000	6.541667	4.958333
88	1.461538	8590.730769	49.423077	14.923077	7.961538

Cluster	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
46	12.200000	6.800000	5.600000	12.800000
48	198.000000	1.333333	1.555556	67.333333
55	18.030928	9.257732	6.979381	17.041237
56	54.257143	17.792857	9.957143	28.714286
57	248.575251	59.632107	40.160535	63.260870
58	30.101796	10.449102	6.203593	19.245509
65	50.115385	11.638462	7.907692	29.630769
66	527.323171	91.987805	76.536585	74.573171
67	553.111111	3.888889	30.222222	3.666667
68	408.044728	91.357827	63.686901	74.690096
69	19.347222	7.777778	7.291667	17.069444
75	20.985915	7.626761	4.225352	12.288732
76	87.783251	19.325123	15.926108	46.650246
77	130.053571	34.473214	25.120536	58.696429
78	36.014599	10.802920	6.788321	21.642336
86	10.363636	8.472727	5.363636	12.109091
87	11.937500	8.520833	6.395833	14.250000
88	12.884615	6.730769	6.192308	28.346154

Cluster	Days_Since_Customer	Fam_Size	Num_Accepted	MntTotal
46	4044.004509	2.400000	0.000000	47.200000
48	3849.826731	2.444444	0.111111	285.666667
55	4006.522035	2.690722	0.051546	70.360825
56	3978.561652	2.935714	0.185714	279.871429
57	3997.999158	2.434783	0.260870	1004.702341
58	3968.347024	2.988024	0.095808	117.928144
65	3973.504509	2.961538	0.084615	229.692308
66	3981.622081	1.725610	1.335366	1572.760293
67	3987.493398	2.333333	0.000000	619.333333
68	3961.016649	2.105431	0.559105	1309.338658
69	3986.312842	2.555556	0.055556	72.902778
75	3976.928452	2.760563	0.105634	78.669014
76	3998.924706	2.911330	0.172414	443.842365
77	3992.175937	2.781250	0.232143	673.727679
78	3990.976772	3.036496	0.131387	147.919708
86	3979.077236	2.527273	0.072727	50.309091
87	4031.583675	2.645833	0.125000	52.604167
88	4119.681432	2.307692	0.076923	77.038462

Figure 72 SOM Cluster

The results of the clustering analysis metrics are as shown in the *figure 73* and *figure 74*. The lowest BD score was noted by SOM method meanwhile the CH score is relatively close to K-mean method CH score.

```
Calinski-Harabasz Index score: 611.8306651306322
```

Figure 73 SOM Calinski-Harabasz Score

```
Davies-Bouldin score for SOM: 0.6282301073882467
```

Figure 74 SOM Davies-Bouldin Score

5.2.5 Second Dataset Using KNN

The same few data preprocessing steps were applied to the data as the previous methods.

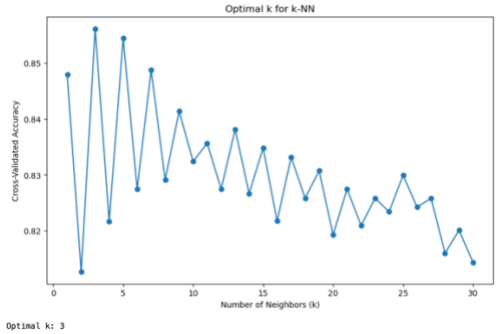


Figure 75 Optimal K for K-NN

The results of the clusters are late on visualized as shown in the *figures 76* up to *figure 89*.

K-NN results are shown in Figure 76, the total spending per cluster is visualized leaving cluster 2 and cluster 1 with the highest spending scores but the spending rates are relatively close.

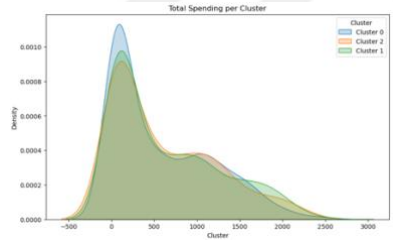


Figure 76 Total Spending per Cluster

Figure 77 shows the family size per cluster the most repeated family member number in all the cluster is 3. But cluster 1 has the greatest number of customers that have a family member number of 2 and most customers are married .

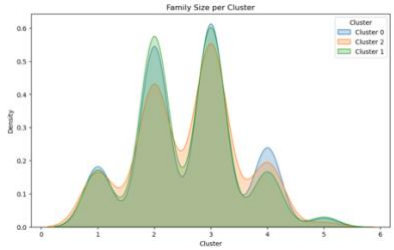


Figure 77 Family Size per Cluster

Figure 78 shows the deal purchase per cluster and their responses to buy items that are under an offer or deal, the most response was from cluster 1. KNN method uses the deal purchase pattern to influence the clustering of the customers.

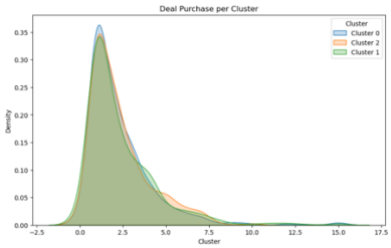


Figure 78 Deal Purchase per Cluster

Figure 79 shows the store purchase of each cluster.

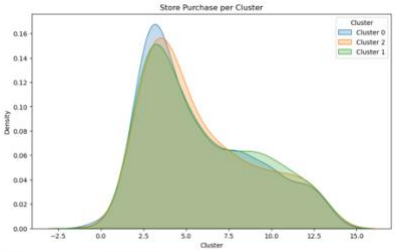


Figure 79 Store Purchase per Cluster

Figure 80 shows the website visit per cluster putting cluster 1 as the highest website visit out of all the clusters. By examining the spending rate and campaign acceptance by the customers we can predict that cluster 1 has the highest website visit. K-mean and KNN both use purchasing behavior to create the custiner clusters.

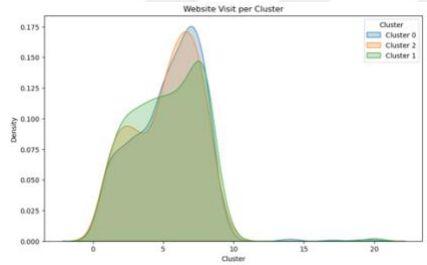


Figure 80 Website Visit per Cluster

Figure 81 shows the catalog purchase per cluster. Cluster 1 which we can see from all the results are the most responsive cluster to deals, websites, store purchases just like cluster 2 from DBSCAN clustering method and we can see that it is the most cluster with a high spending score.

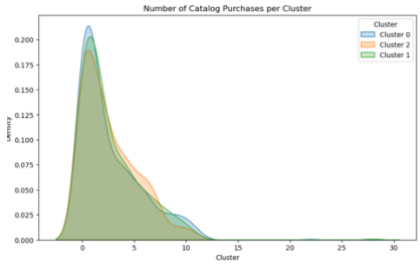


Figure 81 Catalog Purchase per Cluster

Figure 82 shows the website purchase from each cluster the highest website purchase is from cluster 1 which we can see from all the results are the most responsive cluster to deals, websites, store purchases just like cluster 2 from DBSCAN clustering method.

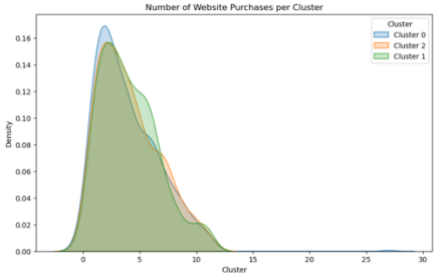


Figure 82 Website Purchase per Cluster

Figure 83 shows the age group each cluster belongs to the highest density of the middle-aged group mostly in their mid-50s this gives us information that age doesn't greatly influence the clustering of customers .

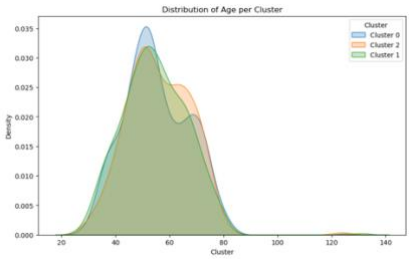


Figure 83 Age Distribution per Cluster

Figure 84 shows the accepted campaigns per cluster the highest response is from cluster 2 and cluster 1, meanwhile cluster 0 show the least campaign response. But mostly there is no response from all the clusters. We can see that cluster 1 and 2 are kind of similar.

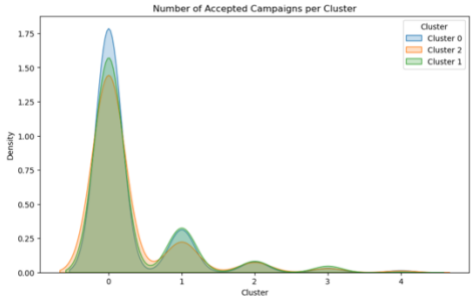


Figure 84 Accepted Campaigns per Cluster

Figure 85 shows the income range of each cluster, the highest income range is from cluster 0 and 1 and the lowest is from cluster 2. Compared to the other methods the closest to this result is from DBSCAN method. But seeing that two clusters have a high income gives us the insight that the customers were not grouped according to their income but the information can be used for further analysis and marketing campaign.

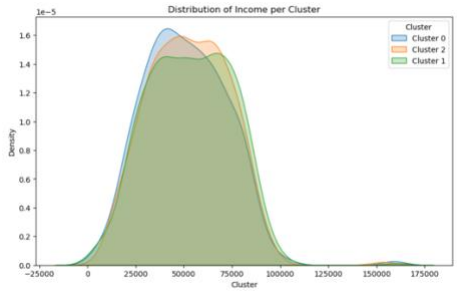


Figure 85 Income Distribution per Cluster

Figure 86 shows how recently each cluster purchased from the store the most recent group is cluster 1 .

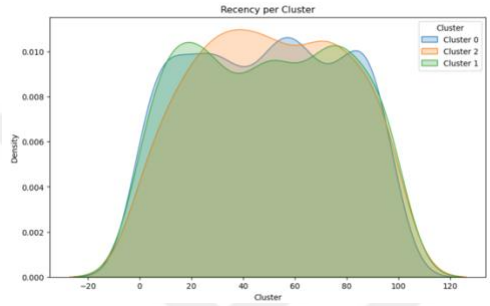


Figure 86 Recency

Figure 87 shows the days since each member of the cluster group became a member of that supermarket. The oldest customer group is cluster 2. By examining the results we can see that k-mean method and KNN method both used the frequency of purchases the deals acceptance and the purchase behavior of the customer to group the clusters, but other attributes like age, education, family size had more influence in K-mean compared to the clusters in KNN.

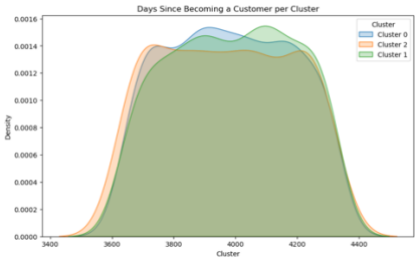


Figure 87 Days Since Becoming a Customer.

Figure 88 This figure shows the education level per cluster in all of the clusters there is customers from each group but cluster 0 are mostly graduates.

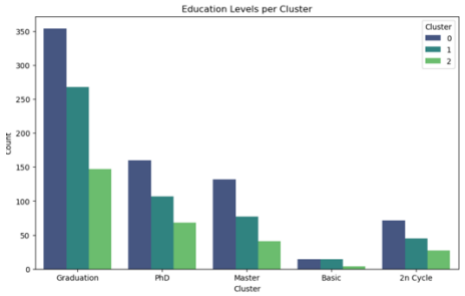


Figure 88 Education Level per Cluster

Then I calculated the F1 Score for cluster analysis and accuracy calculation which shows a good result, a good result should show a value less than 1.

```
F1 Score: 0.8621845518936023
```

Figure 89 F1 Score

Chapter 6

Discussion and Conclusions

6.1 Discussion of Findings for Research Questions

6.1.2 Cluster Results

K-mean Cluster Results for the Second Data Set

First cluster:

Customers in Cluster 0 have an average income of $67003.83 which is the highest income of the clusters and tend to make purchases with low recency. They tend to purchase deals more frequently, they are more likely to accept marketing campaigns. The average customer in this cluster is 32 years old which makes this cluster consist of younger people and has a family of 2 members.

Second Cluster:

Customers in Cluster 1 have the highest customer count have an average income of $51833.10 which is considered meduim income compared to the other clusters and tend to make purchases with high recency of numbe of days. They tend to purchase deals more frequently, they are less likely to accept marketing campaigns. The average customer in this cluster is 38 years old and has a family of 3 members.

Third Cluster:

Customers in Cluster 2.0 have an average income of $49094.54 which is the lowest compared to other clusters, and tend to make purchases with low recency, they purchase deals less frequently, they are less likely to accept marketing campaigns. The average customer in this cluster is 39 years old and has a family of 2.00 members.

DBSCAN Cluster Results for the Second Data Set

First Cluster:

This customer segment has a low spending average, their family size is mostly 3 person or 4 there is barely any single people in this segment. They have a medium catalog purchase tendency and also low website purchase. Their age group are middle-aged people mostly in their fifties. This segment are not responsive to campaigns. Their income range is lower than other clusters. They have a medium recency rate and are mostly graduates with a bachelor degree.

Second Cluster:

This customer segment has a moderate spending score, the family size of this group is mostly 3 but there are single people in this group or only married couples. This segment has the highest catalog and website purchase. They are long term customers , mostly in the retirement age. This segment group can show slight response to campaigns.

Third Cluster:

For this customer segment the spending score is higher than the other groups. They are mostly small families 2-3 person only. Their age group is early middle aged customers around 45 to 60 years old. They have a medium income range. But their recency is low. They have slight response to campaigns.

Outliers:

They have a variant spending pattern with smaller families, they have a moderate catalog and website purchase. They belong to the early middle aged group and can sometimes show slight responses to marketing campaigns and deals.

SOM Cluster Results for the Second Data Set

Cluster1:

Customers in this cluster tend to have an average family size of approximately three members and are older, with an average age around 56. They have a relatively low average income and a low total spending on products. These customers often make infrequent purchases, especially of meat and fish products, and are not very responsive to marketing campaigns.

Cluster2:

These customers typically have medium incomes and moderate spending habits. They are mostly middle-aged, around 47 years old, with an average family size of about three members. Their purchase frequency is moderate, particularly for wines and sweets, and they show limited engagement with marketing efforts. They visit web stores quite often and are relatively consistent in their purchasing patterns.

Cluster3:

This group has a slightly higher average age of 50 and consists of families with approximately three members. They have moderate incomes and spending levels, especially on wine, meat, and sweets. These customers show moderate responsiveness to marketing campaigns and have a balanced approach to both online and in-store purchases.

Cluster4:

Cluster 44 comprises smaller families, often single individuals or couples, with the lowest average income among the clusters. These customers are relatively young, around 41 years old, and have minimal spending, particularly on non-essential items like sweets and gold products. They show low engagement with marketing campaigns and visit online stores moderately.

Cluster5:

With an average family size of nearly three, these customers have medium incomes and spending habits. They are generally middle-aged, around 50 years old, and exhibit moderate purchasing patterns, especially for wine and sweets. Their responsiveness to marketing campaigns is average, and they balance their shopping between online and physical stores.

Cluster6:

This cluster consists of slightly larger families, around three members, with higher incomes and spending, particularly on wine and gold products. These customers are typically older, averaging 54 years in age, and show moderate engagement with marketing campaigns. They make frequent purchases both online and in-store, reflecting their active shopping behavior.

Cluster7:

Customers in this cluster are older, averaging 58 years, with nearly three family members. They have high incomes and significant spending, particularly on wine, meat, and gold products. Their responsiveness to marketing campaigns is higher than average, and they frequently shop both online and in-store, making them highly valuable customers.

Cluster8:

These customers, averaging around 52 years old, have a slightly larger family size of about three members. Their income and spending levels are moderate, with a focus on wine and gold products. They show limited responsiveness to marketing campaigns but are consistent in their purchasing patterns, both online and in physical stores.

Cluster9:

This cluster consists of families with approximately two to three members, averaging 45 years old. They have moderate incomes and spending, particularly on wine and gold products. These customers show a low engagement with marketing campaigns and have a balanced approach to online and in-store shopping.

Cluster10:

Customers in this cluster are older, around 58 years, with moderate family sizes. They have high incomes and spend significantly, particularly on meat and fish products. Their engagement with marketing campaigns is above average, and they frequently shop both online and in physical stores, reflecting their active consumer behavior.

Cluster11:

This group has high-income customers, typically around 54 years old, with smaller families. They exhibit high spending, especially on wine and gold products, and show moderate responsiveness to marketing campaigns. These customers visit online stores frequently and have a consistent purchasing pattern.

Cluster12:

Customers in this cluster are older, averaging 57 years, with smaller family sizes. They have high incomes and substantial spending, particularly on wine and gold products. They are highly responsive to marketing campaigns and exhibit frequent shopping behavior both online and in physical stores, making them highly valuable.

Cluster13:

These customers are older, averaging around 58 years, with larger family sizes. They have moderate incomes and spending levels, particularly on wine and sweets. Their engagement with marketing campaigns is average, and they balance their shopping between online and physical stores.

Cluster14:

This cluster consists of relatively younger customers, around 56 years old, with smaller family sizes. They have high incomes and significant spending, especially on wine, meat, and gold products. Their responsiveness to marketing campaigns is very high, and they frequently shop both online and in-store, making them among the most valuable customers.

Cluster15:

Customers in this cluster are older, averaging 60 years, with larger family sizes. They have moderate incomes and spending levels, particularly on wine and sweets. Their engagement with marketing campaigns is average, and they balance their shopping between online and physical stores.

Cluster16:

This group consists of middle-aged customers, around 49 years old, with medium family sizes. They have moderate incomes and spending levels, particularly on wine

and sweets. Their responsiveness to marketing campaigns is low, and they show consistent purchasing patterns, both online and in physical stores.

Cluster17:

These customers are older, around 53 years, with larger family sizes. They have moderate incomes and spending levels, particularly on wine and gold products. Their engagement with marketing campaigns is above average, and they balance their shopping between online and physical stores.

GMM Cluster Results for the Second Data Set

Cluster0:

Customers in this cluster tend to have a smaller family size and an average age, with moderate total spending on products. They are likely to purchase wines more frequently but show lower engagement with other product categories. These customers have a moderate number of web visits per month and a relatively high number of deal purchases, indicating a preference for discounts. Their responsiveness to marketing campaigns is average.

Cluster1:

This group consists of customers with medium incomes and the highest level of spending on meats and fish products. They have an average family size and tend to be slightly older. These customers have high recency in their purchases, moderate web visits, and are quite responsive to marketing campaigns, particularly through catalog purchases and in-store shopping. They show a high level of complaints but are generally engaged consumers.

Cluster2:

Customers in this cluster have lower incomes and tend to spend less across most product categories. They have smaller family sizes and are older on average. These customers show low engagement with marketing campaigns and have moderate web visits per month. They tend to make fewer deal purchases and have a low total spending, reflecting their limited buying capacity.

Cluster 3:

This cluster is characterized by moderate incomes and spending, particularly on wines, meat, and fish products. These customers have smaller family sizes and an average age. They show a high responsiveness to marketing campaigns and have a moderate number of web visits per month. Their purchasing recency is average, and

they make frequent catalog and in-store purchases, suggesting a well-balanced engagement with different purchasing channels.

Cluster 4:

Customers in this cluster have high incomes and are the highest spenders across all product categories. They have smaller family sizes and are younger on average. These customers exhibit low purchasing recency and high engagement with marketing campaigns, particularly through catalog and in-store purchases. They have fewer web visits per month but show a high level of overall engagement and total spending.

Cluster 5:

This group has medium incomes and average spending across various product categories. They have smaller family sizes and an average age. These customers have moderate purchasing recency and show average engagement with marketing campaigns. They have a moderate number of web visits per month and tend to make deal purchases occasionally, reflecting a balanced buying behavior.

Cluster 6:

Customers in this cluster have medium incomes and high spending, especially on wines, meat, and fish products. They have smaller family sizes and are slightly younger. These customers have moderate purchasing recency and show high responsiveness to marketing campaigns, particularly through catalog purchases. They make fewer web visits per month but are highly engaged and have a high total spending.

Cluster 7:

This cluster consists of customers with lower incomes and lower spending across most product categories. They have smaller family sizes and are slightly younger on average. These customers have moderate purchasing recency and show low engagement with marketing campaigns. They tend to make fewer web visits per month and have low total spending, indicating limited buying capacity.

Cluster 8:

Customers in this cluster have lower incomes and are among the lowest spenders across all product categories. They have smaller family sizes and are slightly younger. These customers exhibit low purchasing recency and low engagement with marketing campaigns. They make fewer web visits per month and tend to purchase less frequently, reflecting a minimal buying behavior.

Cluster 9:

This group has lower incomes and moderate spending, particularly on wines and fish products. They have smaller family sizes and are older on average. These customers have low purchasing recency and moderate engagement with marketing campaigns. They have a high number of web visits per month but tend to make fewer in-store purchases, indicating a preference for online interactions. Their total spending is relatively low, reflecting limited buying capacity.

KNN Cluster Results for the Second Data Set

Cluster 0:

Customers in this cluster are characterized by a slightly higher than average income and moderate spending across various product categories. They have a relatively long recency in their purchases and tend to purchase wines, meat, and fish products more frequently compared to other categories. Their engagement with web visits is moderate, and they show a moderate level of responsiveness to marketing campaigns. They have an average family size and are in their mid-50s.

Cluster 1:

This group consists of customers with a slightly higher income compared to the overall average. They exhibit slightly higher spending across all product categories, particularly on meat, fish, and sweet products. These customers have a slightly longer recency in their purchases and tend to visit websites moderately. They show a slightly higher responsiveness to marketing campaigns compared to other clusters. They have an average family size and are in their mid-50s.

Cluster 2:

Customers in this cluster have incomes and spending patterns similar to those in Cluster 1. They exhibit moderate spending across all product categories, with slightly lower spending on sweet products. Their recency in purchases is slightly longer than average, and they tend to have a moderate number of web visits. They show a moderate level of responsiveness to marketing campaigns. Similar to other clusters, they have an average family size and are in their mid-50s.

6.1.3 Method Approach Comparison:

The main aim of this paper is to compare five different clustering methods while examining closely how different or how similar their outcomes are. While working we found that K-mean was a little bit easier to deal with. Most resources online tend to use K-mean for clustering problems. Meanwhile fewer people apply DBSCAN method

to their clustering problems. The main difference between DBSCAN method and K-mean is that DBSCAN method Focuses on Density; Customers are grouped based on the density of data points in feature space, which is useful for identifying clusters of arbitrary shapes and potentially handling outliers well. Unlike K-mean ; DBSCAN method does not require Predefined Number of Clusters; Clusters are automatically identified without the need to specify the number beforehand. One of the weaknesses of DBSCAN is "Parameter Sensitivity". Cluster results may be impacted by the need to adjust parameters such as eps (epsilon) and MinPts (minimum points). Changing the values of eps and MinPts can change the number of resulting clusters. My first data set is way larger than the second dataset, DBSCAN performed poorly on the first dataset. DBSCAN is not the best for high-dimensional data; when there are a lot of features, performance can suffer. On the other hand K-mean is easier to deal with and way faster than DBSCAN method. This approach works well with bigger datasets. Each cluster's centroids (mean points) serve as a focal point for understanding it; it acts like a great strength for K-mean clustering approach.The two main weaknesses of K-mean clustering is The number of clusters (k) must be specified beforehand, which can be difficult to determine therefore we used elbow method, sometimes it can be tricky to deal with. Another weakness is the fact that the final clustering results can be influenced by randomly selected initial centroids. Lastly in our opinion the most dangerous weakness of K-means clustering operates under the assumption that data points are distributed in a spherical pattern around their respective cluster centers. This means that it performs best when the clusters are well-separated and have roughly equal distances between data points and the cluster centroid within each cluster. I found it insufficient to work with SOM and GMM for the dataset I worked with because the number of clusters was high making it harder to work with for this thesis objectives which is customer segmentation for marketing doing 17 campaigns (SOM) is harder than doing 3 campaigns (KNN). The best scores using CH method and DB score analysis was Kmean and KNN even though the GMM and SOM scores were not that bad but the best CH score was in K-mean.

6.2 Choosing the Right Approach

DBSCAN may be appropriate if you prioritize automatic cluster identification and the exploration of natural data groupings. It also gives a greater focus on outliers which is something that is not found in K-mean and can be of great benefit to some businesses.

If you know the number of customer segments and the data is well-separated, K-Means may be faster and easier to interpret (but be cautious of assuming spherical clusters).Consider visualizing the data distribution to determine the suitability of spherical cluster assumptions for K-Means. KNN was found also accurate for a supervised machine leargning algorithm and it was easy to deal with . the hardest algorithm was SOM.

6.3 Future Work

This paper could take a step further by doing the following: *Domain-Specific Segmentation* which Infuses analysis with information about the Turkish retail landscape. We can consider adding more customer data points, such as product categories or preferred brands. By combining this domain knowledge, we could create a hybrid clustering approach that takes advantage of the strengths of both K-Means and DBSCAN, resulting in even more nuanced customer segments.*Ensemble Techniques* We can consider combining the results of several clustering algorithms. This ensemble approach may overcome the limitations of individual techniques, resulting in more robust and accurate segmentation that captures the complexities of customer behavior. and finally *Longitudinal Analysis* which incorporate a time component by analyzing customer purchasing patterns over time. This could reveal emerging customer segments and provide valuable insights into customer lifetime value, a critical metric for targeted marketing strategies.

6.4 Gap Research

While working on this paper we have noticed some research gaps that could be like *Customer Validation* When comparing clustering methods, validate the identified segments using actual customer feedback. Conduct surveys or focus groups to determine whether the segments accurately represent real-world customer behavior. This validation step strengthens the basis of your research. Another research gap is *Segmentation Actionability* going beyond emphasizing the advantages of customer segmentation. Exploring the practical application of the identified segments. This could entail researching specific marketing strategies or customer service approaches tailored to each segment. By providing actionable insights, the research becomes even more valuable to retailers. Then come *Generalizability* examining the applicability of our findings to different geographical regions or retail sectors. This broader perspective would increase the impact of our research by demonstrating the

applicability of the approach beyond the Turkish market. Latly *Deep Learning Comparison* which includes a comparison of deep learning-based clustering techniques in our future research. This would give a more complete picture of the available options for customer segmentation in the retail sector.

REFERENCES

10 Million Rows Turkish Market Sales Dataset (MSSQL). ([n.d.]). Retrieved March 12, 2024, from Kaggle.com

Adebayo, M., & Akintola, J. M. (2017). Customer Segmentation Using K-Nearest Neighbors Classifier. International Journal of Machine Learning and Computing, 7(3), 248-253.

Armstrong, G., Kotler, P., & Agrawal, M. (2018). Marketing: An introduction (18th ed.). Pearson.

Barua, S., Chakraborty, G., & Dey, A. (2013). Customer segmentation using self-organizing maps: A case study of the retail industry. Expert Systems with Applications, 40(10), 3988-4002.

Beckmann N., Kriegel H.-P., Schneider R, and Seeger B. 1990. The R*-tree: An Efficient and Robust Access Method for Points and Rectangles, Proc. ACM SIGMOD Int. Conf. on Management of Data, Atlantic City, NJ, 1990, pp. 322-331.

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. Data & Knowledge Engineering, 60(1), 208-221.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer. (Chapter 9.1.1 covers Gaussian Mixture Models in detail).

Brown, S. (2021) Machine Learning, Explained. MIT Sloan School of Management, Cambridge. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained Nguyen, S. P. (2022). Deep customer segmentation with applications to a Vietnamese supermarkets' data. Machine Learning and Applications, 13(1), 1-12.

Camilleri, M. A. (2018). Market Segmentation, Targeting and Positioning. In Travel Marketing, Tourism Economics and the Airline Product (Chapter 4, pp. 69-83). Springer, Cham, Switzerland.

Dingsheng Deng. "Application of DBSCAN Algorithm in Data Sampling." Sichuan University Nationalities, Kangding, Sichuan, China. Corresponding author’s e-mail: dds0904@scun.edu.cn.

Dolnicar, S., Tušar, T., & Lavrač, N. (2014). Hierarchical clustering with Ward's linkage for customer segmentation: A case study. Expert Systems with Applications, 41(17), 7515-7531.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226-231). Hossain, A. S. M. S. (2023). Customer segmentation using density-based spatial clustering of applications with noise (DBSCAN) algorithm. International Journal of Information Management, 59, 102461.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226-231). Schubert, E., Sander, J., Ester, M., & Kriegel, H.-P. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), 19.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD-96 Proceedings (pp. 226-231).

Ghasemi, A., & Zaiane, O. (2014). Customer segmentation using Gaussian mixture models and fuzzy logic. Expert Systems with Applications, 41(1), 489-498.

Ghose, A., Gonce, S. B., & Mela, C. F. (2001). Customer segmentation using a self-organizing map. Journal of Marketing Research, 38(3), 360-375.

Hu, Y., & Wang, H. (2022). Customer segmentation using machine learning: A review. Knowledge-Based Systems, 256, 110753.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R. Springer. (Chapter 6.2.2 covers K-Nearest Neighbors Regression in detail)

Kabasakal, İ. (2020). Customer segmentation based on recency frequency monetary model: A case study in e-retailing. Bilişim Teknolojileri Dergisi, 13(1), 48-51.

Kaski, S., Kangas, J., Kohonen, T., & Somervuo, P. (1997). Applying Self-Organizing Maps to Medical Data Analysis. Artificial Intelligence in Medicine, 11(3), 201-215.

Kohonen, T. (2001). Self-organizing maps (Vol. 3). Springer Science & Business Media.

Kumar, V. (2023). Customer Segmentation and Risk Profiling in the Banking and Financial Services Industry Using Machine Learning. Expert Systems with Applications, 172, 115341. Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2020). Customer Segmentation Using Hierarchical Clustering and Formal Concept Analysis. Journal of Intelligent Information Systems, 55(1-2), 307-337.

Liu, C., & Zheng, J. (2022). Customer segmentation using Gaussian mixture model with pairwise constraints. Knowledge-Based Systems, 250, 110739.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), 4765-4774.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226-231, 1996.

McLachlan, G., & Peel, D. (2004). Finite mixture models. John Wiley & Sons.

Mokhtar, A. A. (2023). Marketing Campaign. Retrieved from https://launchschool.com/books/git/read/github (e.g., UCI Machine Learning Repository, Kaggle)

Naga, P. (2023). Customer segmentation using K-means and Gaussian mixture. International Journal of Computer Science and Engineering, 11(3), 25-32.

Nguyen, T. V., & Le, T. T. (2023). Gaussian mixture model for customer segmentation in retail industry. International Journal of Engineering and Science, 12(3), 34-42.

Nguyen, T. V., & Le, T. T. (2023). Gaussian mixture model for customer segmentation in retail industry. International Journal of Engineering and Science, 12(3), 34-42.

Pal, S., & Mitra, S. (2011). Customer segmentation using self-organizing maps: A case study of a retail company. International Journal of Computational Intelligence Research, 7(1), 1-12.
Pal, S., & Mitra, S. (2011). Customer segmentation using self-organizing maps: A case study of a retail company. International Journal of Computational Intelligence Research, 7(1), 1-12.
Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. Journal of Marketing Research, 20(1), 134-148.

Ribeiro, M. T., Samek, W., Montavon, G., & Müller, K.-R. (2018). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135-1144.

Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2020). Customer Segmentation Using Hierarchical Clustering and Formal Concept Analysis. Journal of Intelligent Information Systems, 55(1-2), 307-337.

Samek, W., Montavon, G., Ribeiro, M. T., & Müller, K.-R. (2018). Why ought I to believe you? Explaining any classifier's predictions. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). Lerman, K., Saxena, N., Mehrabi, N., Morstatter, F., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

Sharma, A., & Sharma, P. (2021). A K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. International Journal of Advanced and Innovative Research, 9(1), 1-6.
Sharma, A., & Sharma, P. (2021). A K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. International Journal of Advanced and Innovative Research, 9(1), 1-6.
Sharma, P. (2023, January 24). The ultimate guide to K-means clustering: Definition, methods and applications. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/.

Singh, A. K. (2022). Customer Segmentation Using K-Means Clustering: A Case Study of the Retail Industry. Journal of Retailing and Consumer Services, 62, 102620.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. arXiv preprint arXiv:1703.01365.

Üstebey, S., Yelmen, I., & Zontul, M. (2020). Customer segmentation based on self-organizing maps: A case study on airline passengers. Journal of Aeronautics and Space Technologies, 13(2), 227-233.

Vesanto, J., & Alhoniemi, E. (2000). Self-organizing maps for text clustering. Neurocomputing, 31(2-3), 581-599.

Vesanto, J., & Alhoniemi, E. (2000). Self-organizing maps for text clustering. Neurocomputing, 31(2-3), 581-599.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3), 645-678.

Yousef, A. M., & Zhang, J. (2017). DBSCAN clustering algorithm for customer segmentation and marketing campaigns. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 2201-2206).

Zhang, C. (2015). A novel initialization method for self-organizing maps. Neural Networks, 70, 117-124.

Zhao, Y., & Karypis, G. (2020). K-Means Clustering for Customer Segmentation: A Review. Expert Systems with Applications, 145, 113091.

Zhao, Y., & Karypis, G. (2020). K-Means Clustering for Customer Segmentation: A Review. Expert Systems with Applications, 145, 113091.