CUSTOMER SEGMENTATION USING

CLUSTER ANALYSIS

A REPORT PRESENTED BY
K.A.L.S. Kulasooriya (S/17/404)
K.M.S.S.B. Kulasekara (S/17/403)

to the Department of Statistics and Computer Science of the
FACULTY OF SCIENCE
AND
to the ZONE 24x7, Sri Jayewardenepura Kotte

in partial fulfillment of the requirement
for the award of the degree of

BACHELOR OF SCIENCE HONOURS IN DATA SCIENCE

of the

UNIVERSITY OF PERADENIYA
SRI LANKA
2023

DECLARATION

I hereby declare that the Project Summary Report entitled **("Customer Segmentation with cluster analysis"**) is an authentic record of our own work as a requirement of the four-months project under the course of '**Independent study in Data Science (DSC3263)**' during the period from 01/12/2022 to 25/03/2023 for the award of the degree of B.Sc. Honors Study in Data Science from Department of Statistics and Computer Science Faculty of Science University of Peradeniya, under the guidance of Dr. Sachith Abeysundara (Head of the department) , Prof. Roshan D. Yapa and Ms. B. R. Pavithra M. Basnayake.

.

(Signature of student)
K.M.S.S.B Kulasekara (S/17/403)

(Signature of student)
K.A.L.S Kulasooriya (S/17/404)

Date: _____

Certified by:

1. **Supervisor (Name):…… Date: ……..**

   (Signature):……..

2. **Head of the Department (Name): …… Date : ……..**

   (Signature):……..

Department Stamp:

ABSTRACT

The project on customer segmentation using cluster analysis aims to segment customers based on their similarities in terms of demographic, geographic, and behavioral attributes. The study involves the use of various clustering techniques, including k-means clustering, K-Mode clustering, Gaussian Mixture model clustering, Agglomerative clustering, and K-prototype clustering to identify distinct customer segments. The project analyzes a large Amazon sales dataset of customer transaction and demographic data, which is preprocessed and transformed to extract relevant features. The project's outcomes include the identification of customer segments and their characteristics, such as age, income, location, buying behavior, and preferences. The project's results provide valuable insights to marketing teams in developing targeted marketing campaigns, improving customer experience, and increasing customer satisfaction and loyalty. The project's methodology and findings demonstrate the effectiveness of cluster analysis as a powerful tool for customer segmentation and marketing strategy development.

Keywords: - Customer, Customer segmentation, Cluster analysis, RFM, K mean, K mode, Agglomerative, Gaussian mixture model, Elbow method, Python, Machine learning

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to our project supervisors **Prof. Roshan D. Yapa & Ms. B. R. Pavithra M. Basnayake** for their invaluable guidance and support throughout the duration of this project. Their patience and expertise have been instrumental in helping us navigate the challenges and complexities of this project.

We would also like to thank our external supervisor **Mr. Prasan Ratnayake** at Zone 24*7 for his valuable insights and contributions to this project. His expertise and collaborative spirit have been essential in ensuring the success of this project.

Finally, we would like to express our appreciation to our family and friends for their unwavering support and encouragement during this project. Their love and understanding have been a constant source of motivation and inspiration.

Table of Contents

List of Figure

List of Tables

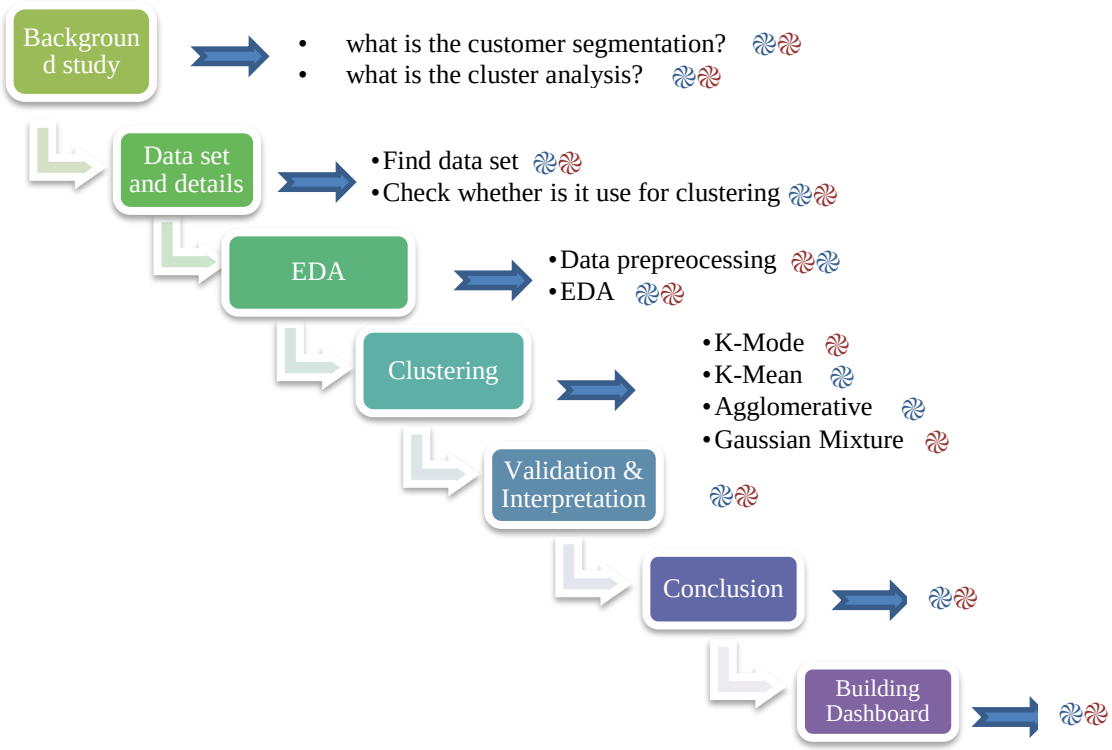
.

CHAPTER 01

1.1. **Error! Bookmark not defined.**We are engaging in a customer segmentation project collaboration with the zone 24*7 company. Zone24x7 specializes in offering end-to-end technology consulting and engineering services encompassing both hardware and software. The services portfolio includes Enterprise Software Applications, Big Data & Data Science Engineering, Embedded Systems Engineering, Remote Monitoring & IoT, Machine Learning, Cognitive Vision, Robotics, and Innovation Services with Technology Proof of Concept Development. Zone24x7 promotes a culture of customer-centric innovation, continuous learning, professionalism, caring for oneself and others, and integrity to create a diverse, inclusive, and thriving workplace for all its associates.

1.2. **Error! Bookmark not defined.**

Lasantha
Sumedha



CHAPTER 02

2.1. **Error! Bookmark not defined.**

Customer segmentation is a crucial aspect of any business's marketing strategy. It involves dividing a large customer base into smaller groups based on shared

characteristics such as age, gender, location, purchase behavior, and preferences. By doing so, businesses can tailor their marketing efforts to each segment's specific needs, resulting in a more targeted and effective approach to customer engagement.

One of the most popular and effective methods of customer segmentation is the use of cluster analysis. Cluster analysis is a statistical technique that allows businesses to group customers based on their similarities and differences across a variety of variables. This method can be used to identify distinct customer groups, understand their behaviors and preferences, and develop targeted marketing strategies to meet their needs.

The process of customer segmentation using cluster analysis typically involves four main steps: data collection, data preprocessing, clustering, and interpretation. Data collection involves gathering relevant customer data, such as demographic information, purchase history, and online behavior. Data preprocessing involves cleaning and preparing the data for analysis, such as removing outliers and missing values.

The clustering step involves applying a clustering algorithm to the preprocessed data to group customers based on their similarities and differences. There are several clustering algorithms available, such as k-means clustering, hierarchical clustering, and fuzzy clustering. The choice of algorithm will depend on the nature of the data and the objectives of the analysis.

Finally, the interpretation step involves analyzing the results of the clustering to understand the characteristics of each customer segment and develop targeted marketing strategies to meet their needs. This could involve developing personalized promotions, improving product offerings, or enhancing customer service based on the specific needs and preferences of each segment.

Overall, customer segmentation using cluster analysis is an essential tool for businesses looking to improve their marketing efforts and better engage with their customers. By identifying distinct customer groups and tailoring their strategies to meet their specific needs, businesses can improve customer loyalty, increase sales, and drive overall business success.

In this project, we have used cluster analysis to segment Amazon's customers based on sales data from four USA regions (Northeast, South, Midwest, and West).

This paper is organized as follows: chapter 2 introduces the problem along with a brief description of the Customer segmentation and cluster analysis and how is it related to Sri Lanka and other countries. Chapter 3 describes the Methodology section. Results and Discussion are presented in Section 4.

2.1.1 Customer Segmentation and Machine Learning:

Using machine learning algorithms to find new segments is one method of client segmentation. Machine learning consumer segmentation enables advanced algorithms to reveal insights and groups that marketers may struggle to obtain on their own. Marketers who create a feedback loop between their segmentation model and campaign results will see their customer groups improve over time. In these circumstances, the machine learning model will be able to not only fine-tune its segment definitions but also determine whether one segment outperforms the others, thus maximizing marketing performance.

2.1.2 Problem Statement

Customers are of different character and have different needs. Therefore, it is not optimal to have the same strategy and marketing for every customer. For a company, it is important to segment their customers and identify the differences between the customer segments to easier meet the customer needs and take care of those customers who are of high importance for the company.

2.2. Literature Review

Customer segmentation is a widely studied topic in marketing research. Clustering algorithms have been increasingly used for customer segmentation as they provide an objective and data-driven approach. In this literature review, we will examine the applications of four popular clustering algorithms: K-means, K-modes, Gaussian mixture models, and agglomerative clustering.

K-means is a well-established and widely used clustering algorithm that partitions data into a predefined number of clusters based on the Euclidean distance between data points. It has been used for customer segmentation in several studies, including a study by (Kilari, 2022), which applied K-means to segment customers based on their demographic and purchase history data, and identified five distinct customer segments. Another study by (Makara, 2021),K-means to segment customers based on their online shopping behavior, and identified three distinct customer segments.

K-modes is a variation of K-means that is designed for categorical data. It uses a dissimilarity measure based on the number of mismatches between categories, and can handle nominal, ordinal, and binary data. It has been applied to customer segmentation in several studies, including a study by (Aprilliant, The k-modes as Clustering Algorithm for Categorical Data Type, 2023),which used K-modes to segment customers based on their churn status, and identified four distinct customer segments.

Gaussian mixture models (GMM) are probabilistic models that assume data is generated from a mixture of Gaussian distributions. GMM can capture complex patterns and underlying structures in data and can handle mixed data types. It has been applied to customer segmentation in several studies, including a study by (Salamzadeh, Ebrahimi, Soleimani, & Fekete-Farkas, 2022),which applied GMM to segment customers based on their transactional behavior and identified three distinct customer segments.

Agglomerative clustering is a hierarchical clustering method that iteratively merges the closest pairs of clusters until all points belong to a single cluster. It can handle various distance measures and can generate a dendrogram to visualize the hierarchy of clusters. It has been applied to customer segmentation in several studies, including a study by (Nazari, Kang, Asharif, Sung, & Ogawa, 2015), which used agglomerative clustering to segment customers based on their online shopping behavior, and identified four distinct customer segments.

In conclusion, clustering algorithms such as K-means, K-modes, GMM, and agglomerative clustering have been successfully applied to customer segmentation in various studies. The choice of clustering algorithm depends on the nature of the data and the research objectives. These algorithms provide a data-driven approach to customer segmentation, which can lead to more effective marketing strategies and improved business performance

2.2.1. Related to Sri Lanka

Customer segmentation using cluster analysis can be applied to any country, including Sri Lanka. Sri Lanka is a country in South Asia that has a diverse population with different demographic, geographic, and behavioral characteristics. Sri Lanka's E-commerce system

differs from that of other countries due to the nation's different, geographical, social, and economic qualities. Thus, conducting customer segmentation in Sri Lanka can help businesses to better understand their customers' needs and preferences, which can lead to more effective marketing strategies and improved customer satisfaction.

Nowadays, the analysis of customer behavior is necessary for active organizations in the field of E-commerce which deal with many customers with different characteristics. In summary, customer segmentation using cluster analysis can be a valuable tool for businesses operating in Sri Lanka, as it can help them to better understand their customers and create more effective marketing strategies.

2.2.2. Related to other countries

Customer segmentation using cluster analysis is a technique used to group customers based on similar characteristics or behaviors. It can be applied to customers in any country, not just Sri Lanka.

In fact, customer segmentation using cluster analysis is commonly used in marketing and business strategy across many different countries. By segmenting customers into groups with similar needs, interests, and behaviors, businesses can tailor their marketing strategies and product offerings to better meet the needs of each segment, ultimately leading to increased customer satisfaction and loyalty.

Some examples of how customer segmentation using cluster analysis might differ across countries could include differences in consumer preferences, cultural norms, and economic factors. For instance, customers in different countries may have different preferences for product features or styles or may prioritize different values when making purchasing decisions. Additionally, economic factors such as income levels and market competition can also influence customer behavior and segmentation strategies.

Overall, while customer segmentation using cluster analysis is a valuable tool for businesses in any country, the specific characteristics and behaviors of customers may vary depending on the country and cultural context.

CHAPTER 03

3.1 Introduction to the Dataset

The dataset used for this research is the Amazon Sales E-commerce dataset.

- Source - Kaggle
- Link - https://www.kaggle.com/datasets/earthfromtop/amazon-sales-fy202021
- Name - Amazon Sales FY2020-21
- This data set includes details about amazon sales from 2020 to the 2021 year in the four US regions which are the South, Midwest, west, and northeast.
- The dataset contains 286,000 records with variables such as Item_Id', 'SKU', 'Quantity_Ordered', 'Price', 'Value', 'Discount_Amount', 'Total', 'Category', 'Payment_Method', 'By_St', 'Customer_Id', 'Year', 'Month', 'Ref_Number', 'Name_Prefix', 'First_Name', 'Middle_Initial', 'Last_Name', 'Gender', 'Age', 'Full_Name', 'Email', 'Signed_Date', 'Phone_Number', 'Place_Name', 'County', 'City', 'State', 'Zip_Code', 'Region', 'User_Name', and 'Discount_Percent'.
- 'Dataset contains 35 variables including 23 categorical variables and 12 numerical variables which contain 7 integer variables and 5 decimal variables which are in Table 1.

Table 1: Dataset Summary

S/N	Attribute Name	Attribute Type
1	Order Id	Categorical
2	Order Date	Categorical
3	Status	Categorical
4	Item Id	Numerical (Integer)
5	SKU	Categorical
6	Quantity Ordered	Numerical (Integer)
7	Price	Numerical (Float)
8	Value	Numerical (Float)
9	Discount Amount	Numerical (Float)
10	Total	Numerical (Integer)
11	Category	Categorical

12	Payment Method	Categorical
13	By St	Categorical
14	Customer Id	Numerical (Integer)
15	Year	Numerical (Integer)
16	Month	Categorical
17	Ref Number	Numerical (Integer)
18	Name Prefix	Categorical
19	First Name	Categorical
20	Middle Initial	Categorical
21	Last Name	Categorical
22	Gender	Categorical
23	Age	Numerical (Integer)
24	Full Name	Categorical
25	Email	Categorial
26	Signed Date	Categorical
27	Phone Number	Categorical
28	Place Name	Categorical
29	County	Categorical
30	City	Categorical
31	State	Categorical
32	Zip Code	Numerical (Integer)
33	Region	Categorical
34	Username	Categorical
35	Discount Percent	Numerical (Float)

3.2 RFM

RFM stands for Recency, Frequency and Monetary.

- Recency, frequency, monetary value (RFM) is a marketing analysis tool used to identify a firm's best clients based on the nature of their spending habits.
- An RFM analysis evaluates clients and customers by scoring them in three categories: how recently they've made a purchase, how often they buy, and the size of their purchases.
- RFM analysis helps firms reasonably predict which customers are likely to purchase their products again, how much revenue comes from new (vs. repeat) clients, and how to turn occasional buyers into habitual ones.

The RFM model is based on three quantitative factors:

1. **Recency**: How recently a customer has made a purchase
2. **Frequency**: How often a customer makes a purchase
3. **Monetary value**: How much money a customer spends on purchases

3.3 Clustering Techniques

3.3.1 K-Mean Clustering Algorithm

The most well-known unsupervised partitioning clustering approach is K-Means Clustering. This clustering approach, commonly known as the centroid-based technique, divides data into non-hierarchical categories. The dataset is separated into a collection of k groups in this sort of partitioning, where K is the number of pre-defined groups or clusters. When compared to another cluster centroid, the cluster center is built so that the distance between data points in one cluster is as low as possible.

According to the Figure 1 these are the steps of K mean algorithm.

Step 1: Select the number K to determine the number of clusters.

Step 2: At random, select K locations or centroids. (It's possible that it's not the same as the incoming dataset.)

Step 3: Form the preset K clusters by assigning each data point to the centroid that is closest to it.

Step 4: Calculate the variance and move the centroid of each cluster.

Step 5: Reverse the previous three steps, reassigning each data point to the cluster's new closest centroid.

Step-6: Go to step-4 if there is a reassignment; otherwise, go to FINISH.

Step 7: The model is now complete.

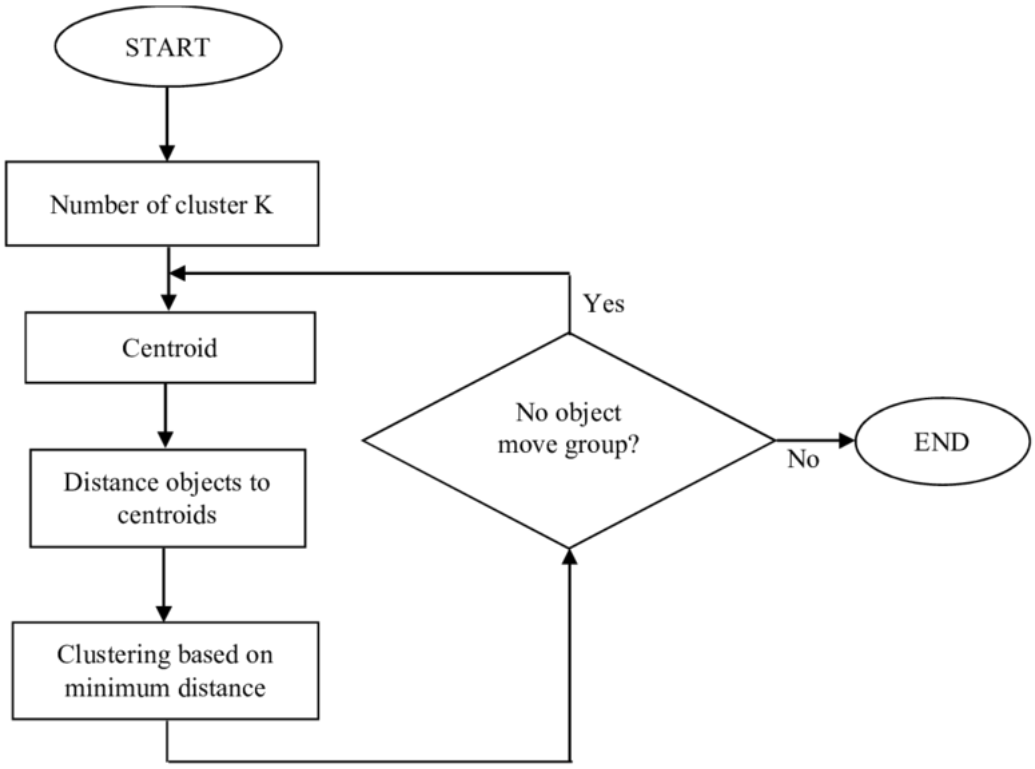


Figure 1: Flow of K mean algorithm

3.3.2 Agglomerative Clustering

Agglomerative clustering requires a distance metric to determine the distance between observations. The choice of distance metric depends on the data and the research question. In retail data, we might use Euclidean distance, Manhattan distance, or cosine similarity, among others. The choice of distance measures is a critical step in clustering. It defines how the similarity of two elements (x, y) is calculated and it will influence the shape of the clusters.

Euclidean distance:

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Manhattan distance:

$$d_{man}\,(x,y) = \sum_{i=1}^{n} |(x_i - y_i)|$$

Where x and y are two vectors of length n.

The agglomerative clustering algorithm works by first assigning each observation to its own cluster. It then iteratively merges the two closest clusters into a new cluster until all observations belong to a single cluster. There are different methods for determining which two clusters to merge at each iteration, such as single linkage, complete linkage, average linkage, and ward method. The choice of linkage method also depends on the data and the research question.

Once the clustering is complete, we need to determine the optimal number of clusters. One approach is to use the elbow method, which involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and selecting the number of clusters at the "elbow" of the curve, where the reduction in WCSS starts to level off. The flow of agglomerative clustering algorithm illustrated in Figure 2.

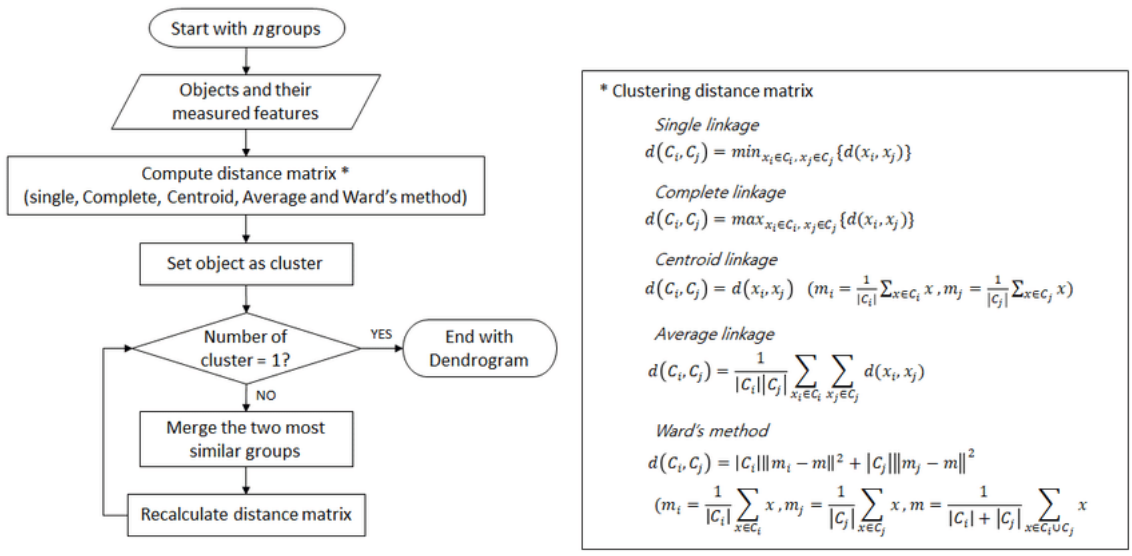


Figure 2:Flow of agglomerative clustering

3.3.3 Gaussian Mixture Model Algorithm

A Gaussian mixture model (GMM) is a probabilistic model for data clustering that considers all data points to be derived from a mixture of a finite Gaussian distribution with unknown parameters. The commonly used clustering approaches, such as k-means are distance-based. While GMM is based on a probability model rather than an objective function of distance measures. GMM assumes that the dataset follows a mixture model of probability distributions so that each cluster is represented by a parametric probability

density and the entire cluster structure can be modeled by a finite mixture. Mathematically, GMM is defined as a parametric probability density function that can be represented as a weighted sum of *k* Gaussian components. Each component is characterized by a simple parametric form. The GMM can be written as,

$$p_M(x) = \sum_{i=1}^{k} \propto_i \; p(x|\mu_i, \Sigma_i)$$

where $p\ (x|\mu_i, \Sigma i)$ is known as the *j*-th components of the mixture with μ_i and Σ_i, and $\propto_i$is called the mixture coefficient and must satisfy $0 \leq \propto_i \leq 1$ together with $\sum_i^k \propto_i = 1$ ‘

In GMM, the probability density function is Gaussian distribution defined as follows

$$p(x) = \frac{1}{(2\pi)^{d/2}\ |\Sigma|^{1/2}}\ e^{-1/2(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

Here are the steps involved in the GMM clustering technique:

1. Initialize the number of clusters: The first step is to decide the number of clusters you want to create. You can either choose this number manually or use a statistical method such as the Bayesian Information Criterion (BIC) to determine the optimal number of clusters.
2. Initialize the parameters: Next, you need to initialize the parameters of the Gaussian distributions. These parameters include the mean, covariance, and weight of each Gaussian component.
3. Expectation step: In this step, you assign each data point to a cluster based on the probability of it belonging to each of the Gaussian distributions. This is done using Bayes' theorem.
4. Maximization step: After the data points have been assigned to clusters, the parameters of each Gaussian distribution are updated using the maximum likelihood estimation (MLE) technique.
5. Repeat the steps 3 and 4 until convergence: Steps 3 and 4 are repeated until convergence is achieved, which is typically determined when the change in the log-likelihood function falls below a certain threshold.

6. Predict the clusters: Once the GMM model has converged, you can use it to predict the cluster membership of new data points based on their probabilities of belonging to each Gaussian component.

Finally, we need to interpret the clusters. This involves examining the characteristics of each cluster, such as the mean values of the variables, to identify patterns and insights. We may also want to compare the clusters to external variables, such as customer demographics or purchase behavior, to gain further insights.

3.3.4 K-Mode Clustering Algorithm

k-Modes clustering was first introduced by Huang (1998) as a clustering method developed from the k-Means method. Therefore, efficient k-Modes is as efficient as k-Means that is used for categorical data. Modifications made to the k-Means method are:

a) The distance between two data points X and Y is the number of observations in X and Y whose values are different (dissimilarity measure), formally formulated as follows:

$$d_1(X,Y) = \sum_{i=1}^{n} \delta(x_i, y_i)$$

where

$$\delta(x_i, y_i) = \begin{cases} 0\,; x_i = y_i \\ 1\,; x_i \neq y_i \end{cases}$$

with x_i is the value of the i-th observation of the X data, y_i is the value of the i-th observation of the Y data, and n is the number of observation.

b) Change the means to modes

c) Use frequency to search for modes

Table 2:Example table of K mode algorithm

ID	Category_1	Category_2	Category_3	Category_4
1	0	1	0	0
2	0	0	1	0
3	0	1	0	1
4	0	0	1	0
5	0	1	0	0

Centroid	0	1	0	0

For example, with the following four dummy data, the centroid point is built from the mode of each column. The mode is the data value that occurs the most. The formation of the centroid is done by looking for the mode of each column.

The following are the steps for k-Modes-based clustering (Huang 2008):

1. Select the k initial mode
2. Allocate the observation to the closest cluster based on a simple dissimilarity measure. Update each cluster mode after each allocation.
3. After all the observations have been allocated to a cluster, check the dissimilarity value of each observation against the mode. If an observation turns out that the closest mode is in another cluster, move the observation to the appropriate cluster and update the mode of both clusters
4. Repeat step 3 until none of the observation change to another cluster

How to choose the optimal number of clusters?

For K Modes, plot cost for a range of K values. Cost is the sum of all the dissimilarities between the clusters. Select the optimal K where you observe an elbow-like bend with a lesser cost value.

3.4 Experimental Setup

3.4.1 Environment Setup

- Jupyter Notebook
    - Jupyter Notebook is a server-client program that allows you to edit and run notebook documents, code, and data using a web browser. The Jupyter Notebook App can be operated locally on a PC with no internet connection (as described in this article) or remotely on a server with internet access. Users can build and organize processes in data science, scientific computing, computational journalism, and machine learning using the versatile interface.
- Google colab

- A cloud-based development environment offered by Google called Colab enables to execute Python code in an interface similar to a Jupyter notebook. It offers free access to a virtual machine that has strong hardware tools like CPUs, GPUs, and TPUs. Because of this, users can carry out complex computations and machine learning tasks without the hassle of building up a local environment or needing specialized hardware. Furthermore, Colab makes it simple to collaborate on projects by enabling numerous users to access the same notebook at once.

- Visual Code Studio

  - Visual Studio Code (VS Code) is a free and open-source source code editor developed by Microsoft. It is a lightweight and cross-platform application that is available on Windows, macOS, and Linux. VS Code is designed to be highly customizable and extensible, with support for a wide range of programming languages, frameworks, and tools. It includes features such as debugging, syntax highlighting, intelligent code completion, and Git integration, making it a popular choice among developers for their code editing needs. Additionally, VS Code has a large and active community that contributes to its ecosystem by creating and maintaining extensions, themes, and other useful tools.

3.4.1 Programming Language

- Python

  - Python is a high-level, general-purpose programming language that is used for a wide range of applications, including web development, scientific computing, data analysis, artificial intelligence, and more. It was first released in 1991 by Guido van Rossum and has since become one of the most popular programming languages in the world.

- Cascading Style Sheets

  - Cascading Style Sheets a stylesheet language used to define the layout, formatting, and appearance of web pages. With CSS, developers can separate the presentation of a web page from its content, which helps to make the code more organized and easier to maintain.

3.4.2 Packages and Libraries

- NumPy
    - NumPy is a package that contains multidimensional array objects and tools for manipulating them. NumPy is a Python library that allows us to perform mathematical and logical operations on arrays. NumPy is widely used in combination with SciPy and Matplotlib (Scientific Python) (plotting library). This combination is frequently used as a substitute for MATLAB, a prominent technical computing platform. The Python counterpart to MATLAB, on the other hand, is today regarded as a more contemporary and comprehensive programming language.
- Pandas
    - Pandas is a Python toolkit for data science, data analysis, and machine learning that is open-source. It is based on NumPy, a multi-dimensional arrays-supporting library. Pandas, being one of the most widely used data manipulation tools, works well with a variety of other Python data science modules.
- Matplotlib
    - For 2D array charts, Matplotlib is a superb Python visualization library. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the entire SciPy stack. The ability to show vast volumes of data in simple images is one of the most essential advantages of visualization. Line, bar, scatter, histogram, and more graphs are available in Matplotlib.
- Seaborn
    - Seaborn is a matplotlib-based open-source Python library. It's used for exploratory data analysis and data visualization. Seaborn makes using data frames and the Pandas library a breeze. The graphs that are generated may also be readily changed
- Plotly
    - Plotly is both open-source and commercial data visualization library that allows users to create interactive charts and graphs in Python, R, and

JavaScript. It provides a wide range of graph types, including line charts, scatter plots, bar charts, heatmaps, and more.

- Sckit-Learn
    - It is a free Python machine learning software, sometimes known as sklearn. It is meant to interact with the Python numerical and scientific libraries NumPy and SciPy, and features support vector machines, random forests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering algorithms.
- Scipy
    - SciPy is a Python library used for scientific and technical computing. It is built on top of the NumPy library and provides additional functionality for optimization, signal processing, linear algebra, statistics, and more. SciPy is often used in scientific research, engineering, and data analysis
- Datetime
    - In Python, the datetime module is a built-in module that provides classes for working with dates and times. The datetime class is the most used class in the module, and it represents a date and time in a specific format.
- Streamlit
    - Streamlit is a Python library that allows users to create web applications with interactive data visualization and analysis capabilities using Python scripts. It is designed to simplify the process of building and deploying data science applications and machine learning models.
- Yellowbrick
    - Yellowbrick is a suite of visual analysis and diagnostic tools designed to facilitate machine learning with scikit-learn.

3.5 Methodology

Downloaded data set were in the csv file format. As the first step Dataset was uploaded to the Google collab and Jupyter notebook. Then to identify the formats and data types of each variable first few rows were extracted from the uploaded data set. After looking at the head of the dataset summary statistics were obtained. When going through the

Summary statistics it was identified that before clustering some sort of preprocessing was needed. Null value, Unique value check was carried out first. Dimensions of the dataset and structure of the data set were checked next. value and total columns showed wrong calculation values hence calculated correctly and stored again. Date time columns were shown as numerical columns and therefore transformed into the correct date time format. Then Conducted an EDA analysis for all the variables one by one and plotted the graphs for each variable to get an idea about the variables and is there any relationships or patterns between variables.

After getting the idea about variables and patterns, the dataset was divided into two parts such as completed orders and canceled orders then moved into the clustering techniques. First considered numerical variables in the data set. Using numerical variables, then calculated the RFM columns and removed percentile outliers from the RFM table, hereafter used three clustering techniques for that RFM table such as Kmean, Agglomerative clustering, and Gaussian Mixture Clustering.

In K mean clustering, The Elbow method graph is used to do Clustering using k-means for a range of k clusters (let's say 1 to 20) and found the optimal number of clusters. Using optimal k value then built the model and did the further cluster visualizations and interpretations.

The agglomerative clustering technique used the elbow graph (Same as the previous method) to identify the optimal k. Next, used the ward method and linkage methods to draw the dendrograms. Again, built the agglomerative model to identify the clusters and how they are scattered.

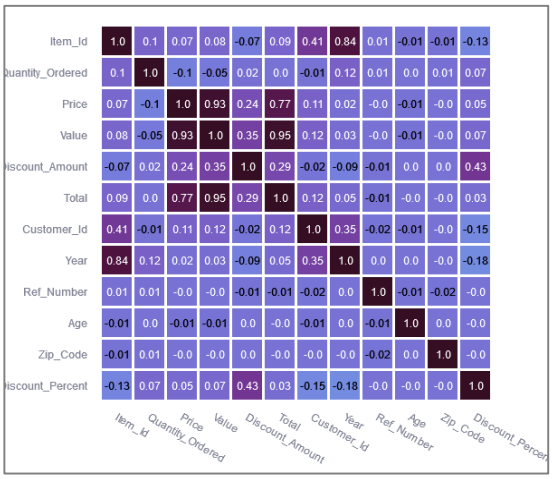
In Gaussian Mixture Model (GMM) clustering algorithm, an Optimal number of clusters was defined by using the lowest Bayesian information criterion (BIC) score. then Built the GMM and did further cluster visualizations and interpretations.

In those, all three methods finally used RFM snake plots to interpret the final clusters.

After that considered categorical variables in the data set. Here used the K-mode clustering algorithm for selected categorical variables and then applied the cost function to determine the optimal number of clusters. Then used Bar graphs with each categorical variable to get an idea about each cluster.

CHAPTER 04

4.1. Results of the explanatory data analysisError! Bookmark not defined.4.1.1 Correlation Matrix



According to the Figure 3 there is a highly positive correlation between Total and Value Variables. Between Customer_Id and Discount percent has a weakly negative correlation.

Figure 3: Correlation Matrix

4.1.2 Order Date

According to the Figure 4 most of the orders are placed in Dec 2020.Total sales of 2021 are higher than the total sales of 2020.

Figure 4: Total sales are from Jan 2020 to Dec 2021

4.1.3 Order

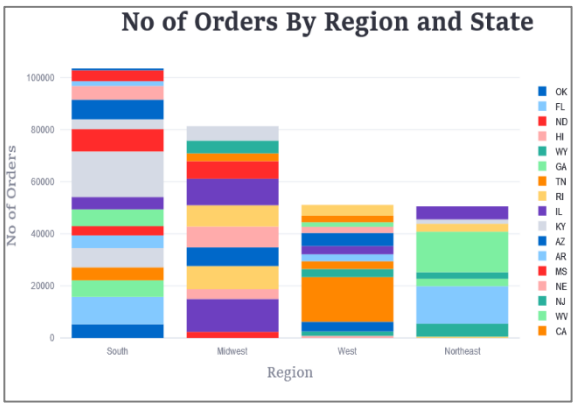


Figure 6: No of orders by state and region

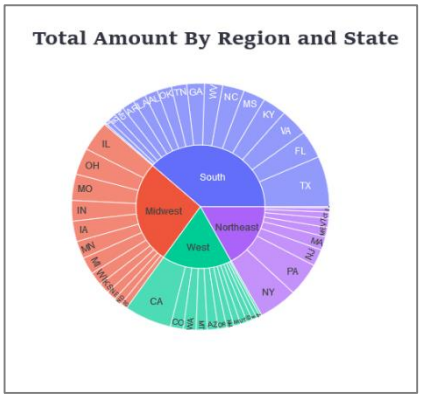


Figure 5: Total amount by state and region

According to the Figure 5 most of the orders are from southern region and least of the orders are from northeast region. And within southern region most of the orders are from state of Texas. As Figure 6 describes highest total amount of sales are from southern region.

4.1.4 Category

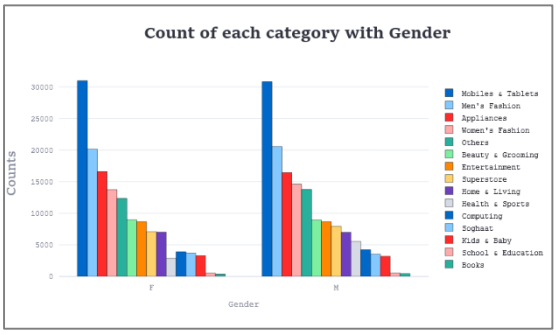


Figure 8: Count of each category by gender

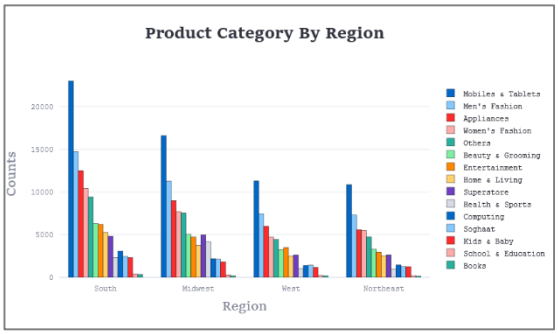


Figure 7:Count of each category by region

According to the Figure7 Mobiles & Tablets is the demanded category in both genders. Males have been ordered more Health & Sports equipment's than the females. As Figure 8 illustrates Mobiles & tablets is the highest demanded category in each region.

4.1.5 Completed Orders

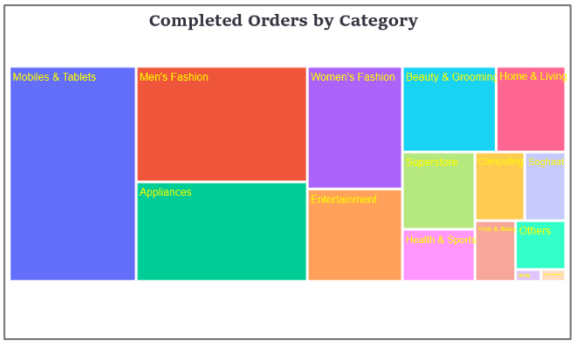


Figure 9 Completed orders by category

Figure 10: Completed orders by region

According to the Figure 9 Mobiles & tablets are the most completed orders then the Men's fashion and then the Appliances. Figure 10 confirms that the fact of most of the orders are from southern region. Equal number of orders completed in west and northeast regions.

4.1.6 Canceled orders

Figure 11 Canceled orders by region

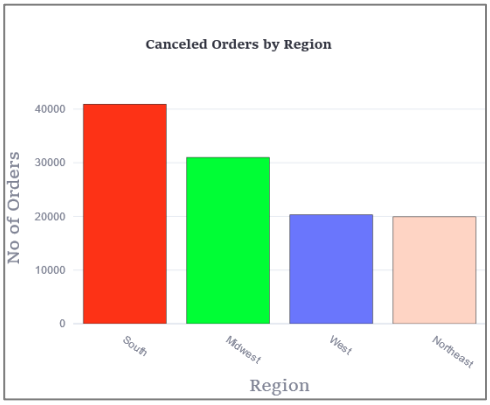


Figure 12 Canceled orders by category

According to the Figure 11 Books are the least canceled orders. The region with the largest number of orders is south, with over 35000 canceled orders.

4.2. Results of the Clustering

After analysis of data to cluster customers using numerical variables, we first classified the customers with RFM features. In order to measure similarities between observations and form clusters they use a distance metric. So, features with high ranges will have a bigger influence on the clustering. Therefore, standardization is required before building a clustering model. Finally, we use K means, Agglomerative and Gaussian mixture model clustering techniques. Then to cluster customers using categorical variables we use K mode clustering technique.

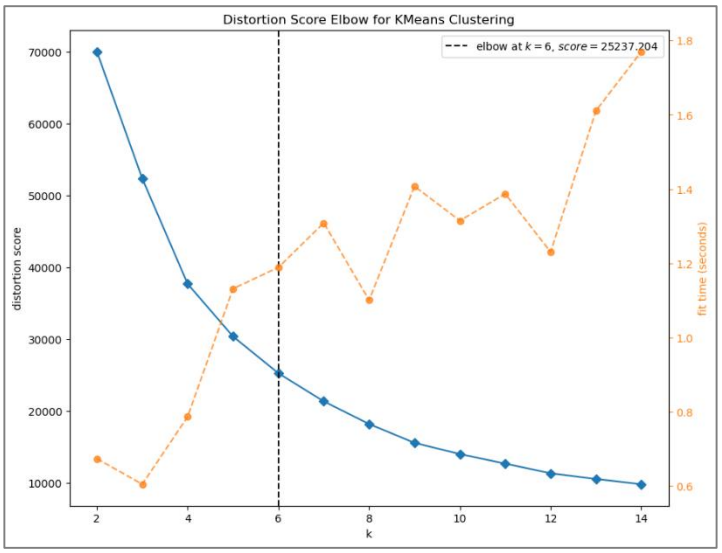
4.2.1 Results of the order completed data

We first applied these clustering techniques to order completed customers.

4.2.1.1 K Mean Clustering Technique

we Identified the optimal number of clusters as 6 using the below graph.

4.2.1.1.1 Elbow Graph



Elbow appear at K=6.

Therefore, optimal number of clusters is 6.

.

Figure 13: Elbow graph of K means algorithm

4.2.1.1.2 Model building

After that built a K mean model using optimal k as 6.

```
kmeans_scaled = KMeans(6)
kmeans_scaled.fit(x_scaled)
```

```
KMeans
KMeans(n_clusters=6)
```

Figure 14: K mean model

4.2.1.1.3 Cluster visualization

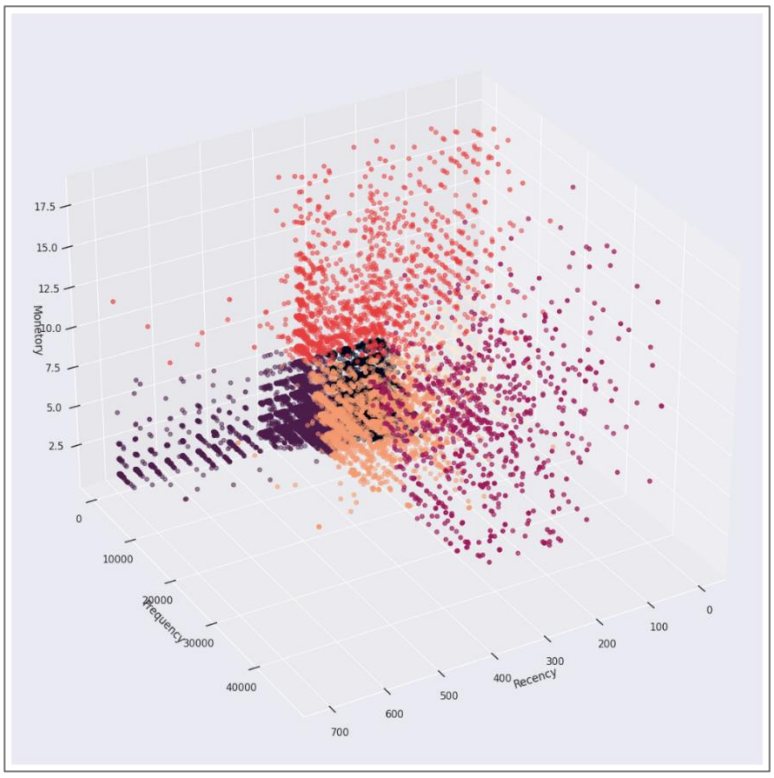


Figure 15: 3D scatterplot of K mean clusters

After the cluster visualization, obtained the summary statistics of each cluster.

4.2.1.1.4 Summary Statistics of all clusters

	Recency			Frequency			Monetory			
	mean	min	max	mean	min	max	mean	min	max	count
cluster										
0	308.397913	34	544	3.518265	1	8	11356.688198	2740.0	26833.240	3066
1	367.126494	302	698	1.621107	1	7	1105.692306	0.0	9999.800	13550
2	256.809211	3	362	7.558114	1	18	29717.786814	17047.2	47438.727	912
3	73.395636	0	161	2.176204	1	7	1649.935927	0.0	20954.800	2429
4	247.069677	3	698	8.987742	6	18	3400.682876	161.7	18782.688	1550
5	237.210827	153	298	1.612374	1	5	1052.144878	0.0	10169.900	8405

Figure 16: Summary statistics of k mean algorithm

Acoording to Figure 16 highest recency, frequency and monetary mean values are in cluster 1,cluster 4 and cluster 2 respectively.

4.2.1.1.5 Predicted clusters by category, Region and Gender

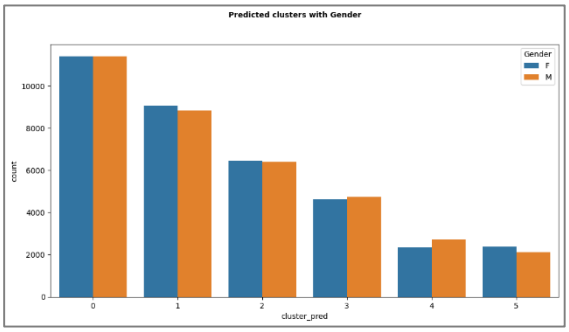


Figure 18: Predicted clusters by gender

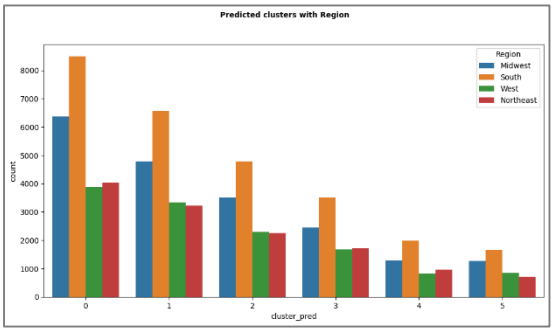


Figure 17:Predicted clusters by region

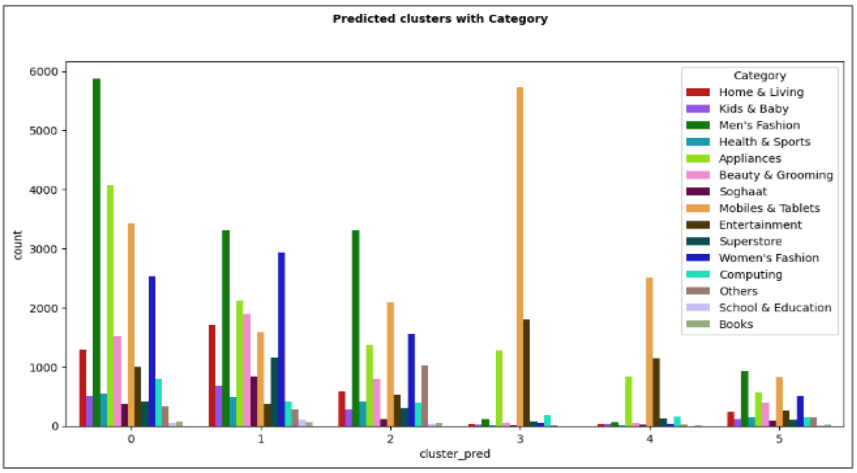


Figure 19:Predicted clusters by category

Table 3 contains the summary of predicted clusters by gender (in Figure 17) , region (in Figure 18) and category (in Figure19).

Cluster Number	Category		Region		Gender	
	Highest	Lowest	Highest	Lowest	Highest	Lowest
0	Men's Fashion	Books	South	Northeast	Female	Male
1	Men's Fashion	Books	South	Northeast	Female	Male
2	Men's Fashion	School & Education	South	West	Male	Female
3	Mobile & Tablets	Books	South	West	Male	Female
4	Men's Fashion	Books	South	Northeast	Female	Male
5	Mobile & Tablets	Books	South	West	Male	Female

Table 3:Summary of predicted clusters in K means

4.2.1.2 Agglomerative Clustering Technique

4.2.1.2.1 Elbow graph

we Identified the optimal number of clusters as 6 using the same elbow graph used in the kmean clustering technique.

4.2.1.2.2 Dendrogram for agglomerative model

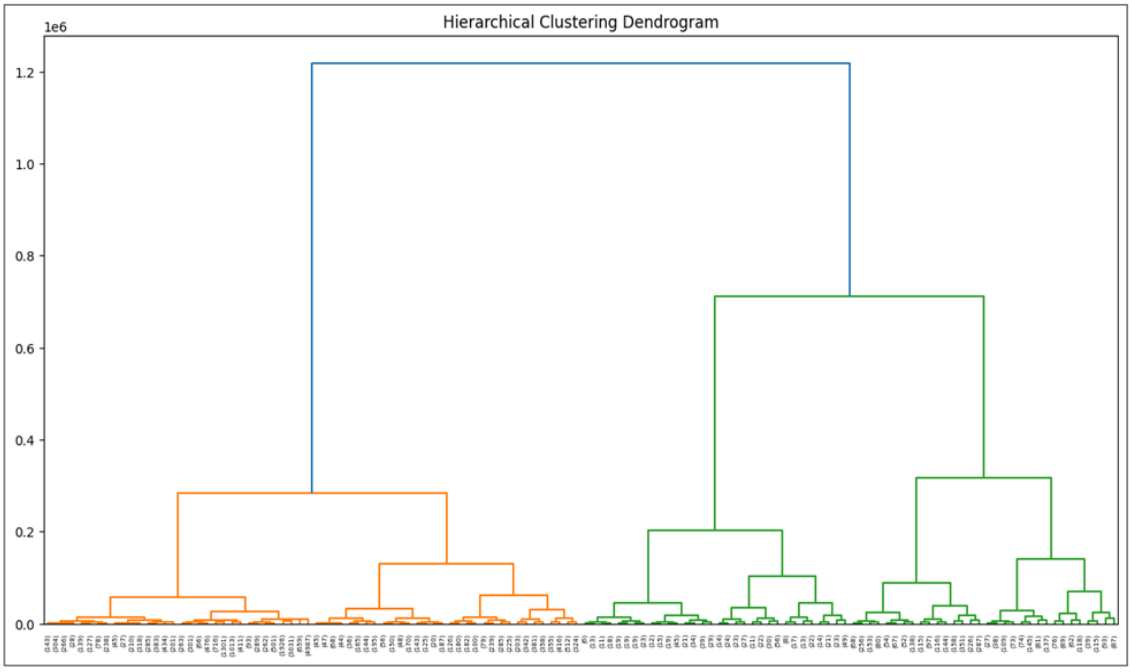


Figure 20: Dendogram of agglomerative clustering (ward method)

4.2.1.2.3 Model building

After that built an Agglomerative model using optimal k as 6.

```
from sklearn.cluster import KMeans,AgglomerativeClustering,DBSCAN
# apply agglomerative algorithm
agglo_model = AgglomerativeClustering(linkage="ward",n_clusters=6)
agglomerative_clusters = agglo_model.fit_predict(x_scaled)
agglomerative_clusters
```

```
array([5, 4, 2, ..., 5, 1, 0])
```

Figure 21:Agglomerative model building

4.2.1.2.4 Cluster visualization

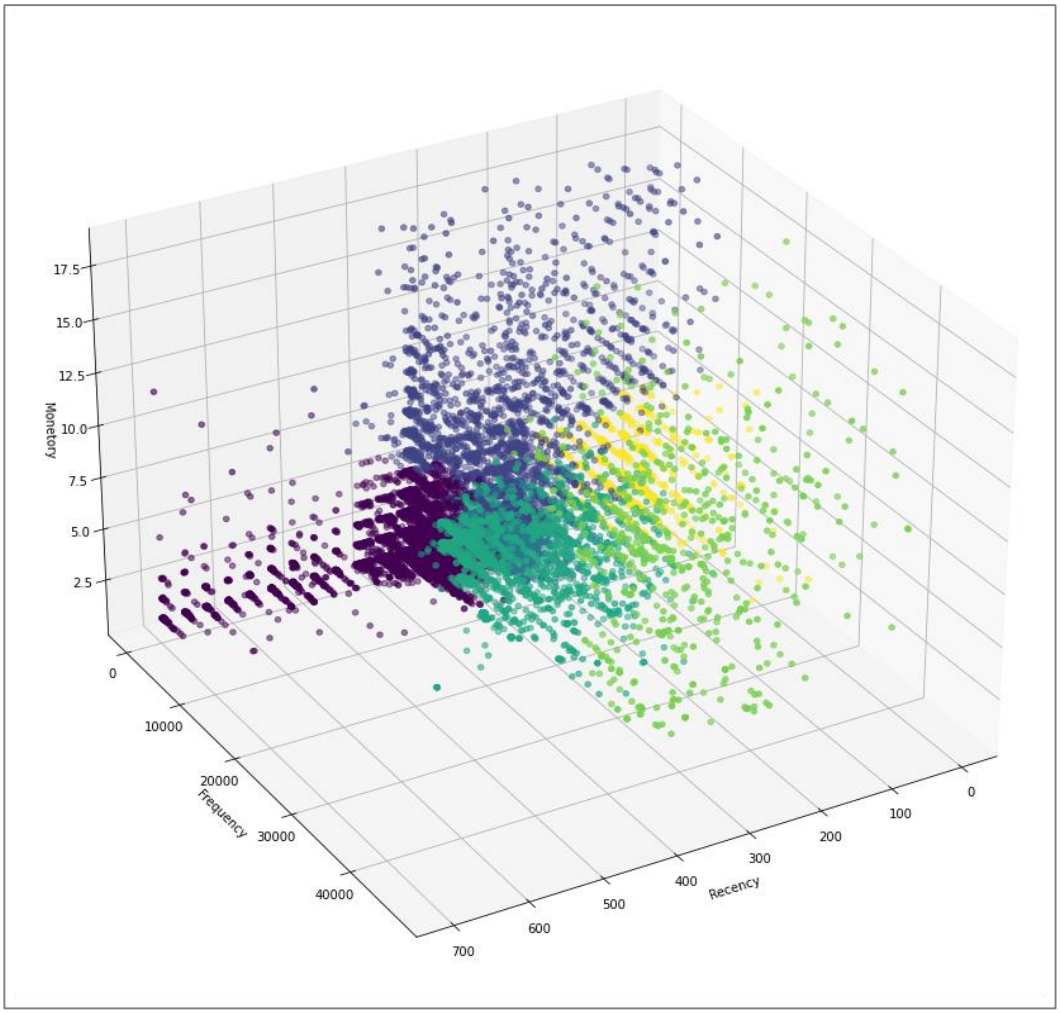


Figure 22:3D scatterplot of agglomerative clustering

4.2.1.2.5 Summary statistics of all clusters

```
rfm_df_AggClusters.groupby('agglomerative_clusters').agg({
    'Recency' : ['mean','min','max'],
    'Frequency' : ['mean','min','max'],
    'Monetory' : ['mean','min','max','count']
})
```

	Recency			Frequency			Monetory			
	mean	min	max	mean	min	max	mean	min	max	count
agglomerative_clusters										
0	367.502119	277	698	1.665205	1	12	1299.065498	0.00	12464.000	13686
1	238.478574	3	577	7.172207	3	18	3182.902019	0.00	16343.750	2497
2	235.788981	114	329	1.463785	1	3	1324.509992	0.00	12214.600	8767
3	319.134131	125	484	3.996155	1	10	14316.670268	6228.83	32333.500	2341
4	240.907131	3	353	8.361526	1	18	33038.839035	16995.70	47438.727	603
5	61.782953	0	160	2.222993	1	11	2106.019815	0.00	27599.070	2018

Figure 23: Summary statistics of agglomerative clustering

Acoording to Figure 23 highest recency, frequency and monetary mean values are in cluster 0,cluster 4 and cluster 4 respectively. 0^{th} cluster contains highest number of customers compare to other 5 clusters.

4.2.1.2.6 Predicted clusters by category, region and gender

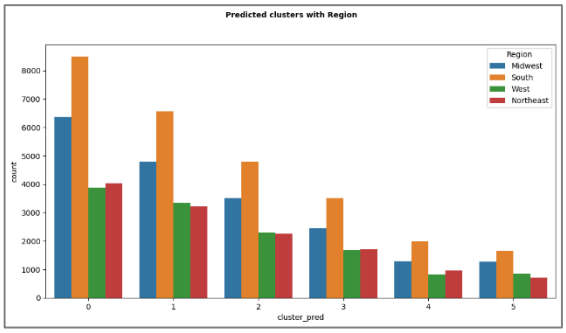


Figure 25:Predicted clusters by region in Agglomerative clustering

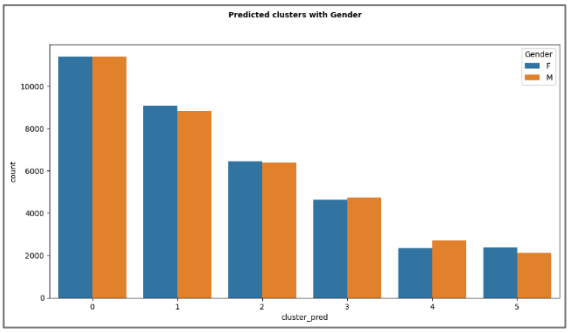


Figure 26: Predicted clusters by gender in Agglomerative clustering

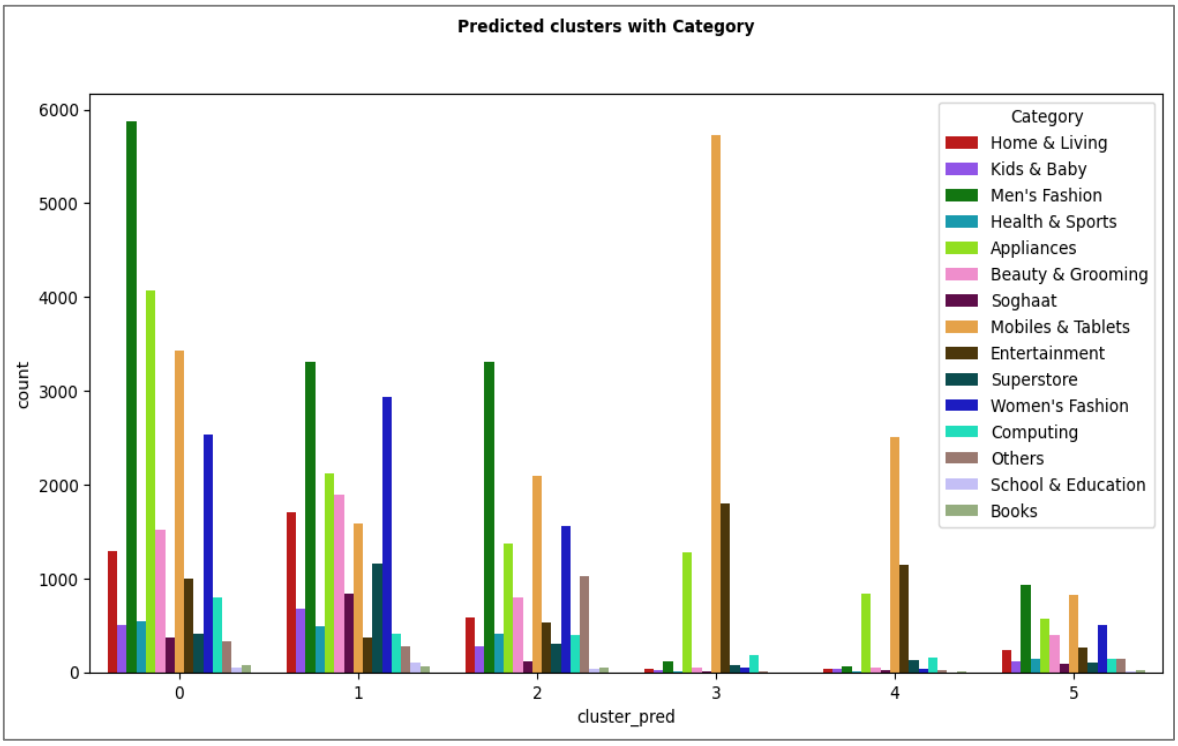


Figure 24:Predicted clusters by category in Agglomerative clustering

Table 4 contains the summary of predicted clusters by gender (in Figure 24) , region (in Figure 25) and category (in Figure 26).

Table 4:Summary of predicted clusters in Agglomerative clusters

Cluster Number	Category		Region		Gender	
	Highest	Lowest	Highest	Lowest	Highest	Lowest
0	Men's Fashion	School & Education	South	West	Equal	Equal
1	Men's Fashion	Books	South	Northeast	Male	Female
2	Men's Fashion	School & Education	South	Northeast	Male	Female
3	Mobile & Tablets	Books	South	West	Female	Male
4	Mobile & Tablets	School & Education	South	West	Female	Male
5	Men's Fashion	School & Education	South	Northeast	Male	Female

4.2.1.3 Gaussian Mixture Model (GMM) Clustering Technique

4.2.1.3.1 Gaussian distribution

Before applying the model, we checked about is data followed a Gaussian distribution or not. Here used the Shapiro-Wilk test to determine the distribution.

```
# Extract the RFM values into a numpy array
rfm_values = rfm_df[['Recency', 'Frequency', 'Monetory']].values

# Perform Shapiro-Wilk test on the RFM values
stat, p = shapiro(rfm_values)

# Print the results
print('Statistic=%.3f, p=%.3f' % (stat, p))
if stat < 1.96:
    print('The RFM data is likely Gaussian.')
else:
    print('The RFM data is likely not Gaussian.')

Statistic=0.308, p=0.000
The RFM data is likely Gaussian.
```

Figure 27: Shapiro Wilik test

By the above results, concluded that the data follows a Gaussian distribution at 5% significant level.

4.2.1.3.2 BIC score graph

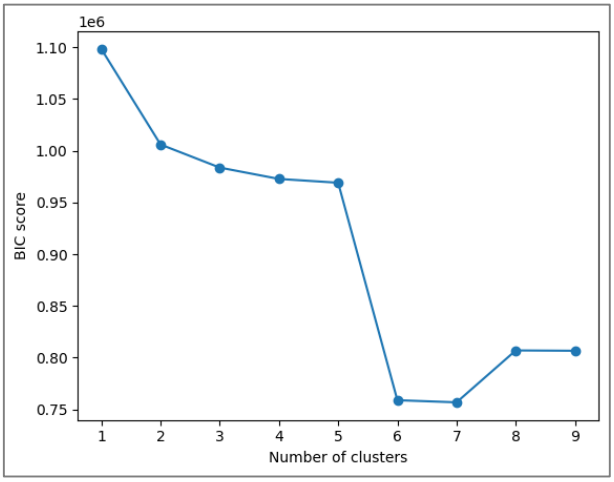


Figure 28: BIC score

Then, We Identified the optimal number of clusters as 7 using the BIC Score graph.

4.2.1.3.3 Model building

```
# Fit the Gaussian mixture model
gmm = GaussianMixture(n_components=7, random_state=42)
gmm.fit(rfm_df_GauClusters)

# Predict the clusters
clusters = gmm.predict(rfm_df_GauClusters)
```

Figure 29: Gussian Mixture model building

After that built a GMM algorithm using optimal k as 7.

4.2.1.3.4 Cluster visualization

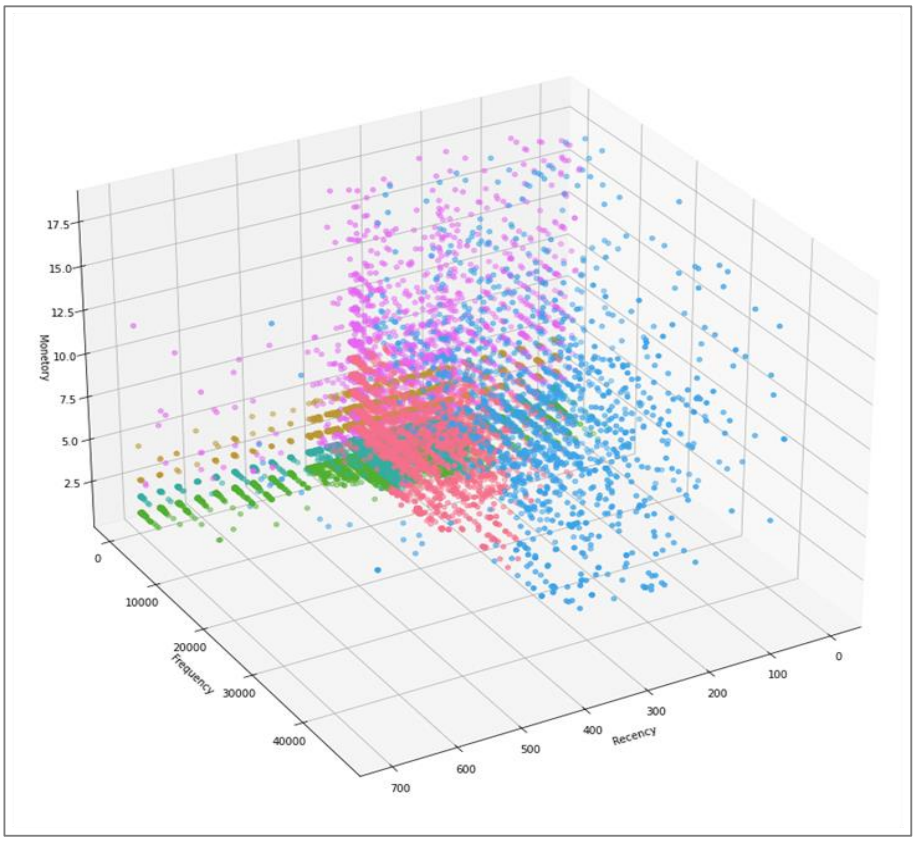


Figure 30: 3D scatterplot of GMM clusters

4.2.1.3.5 Summary statistics of all clusters

```
rfm_df_GauClusters.groupby('GauCluster').agg({
    'Recency' : ['mean','min','max'],
    'Frequency' : ['mean','min','max'],
    'Monetory' : ['mean','min','max','count']
})
```

	Recency			Frequency			Monetory			
	mean	min	max	mean	min	max	mean	min	max	count
GauCluster										
0	243.843034	93	417	3.470018	1	6	9345.612343	2548.97	19664.400	1134
1	274.508731	0	698	3.505239	3	5	504.481434	0.00	1313.600	2577
2	299.938318	2	698	1.000000	1	1	1015.967353	0.00	9059.800	14980
3	292.013573	2	698	2.000000	2	2	1612.547106	0.00	12511.800	5452
4	224.294535	3	577	7.694819	1	18	22733.960836	4865.10	47438.727	1409
5	227.480854	3	698	7.242310	3	18	1853.995300	161.70	6063.380	1593

Figure 31: Summary statistics of GMM

4.2.1.3.6 Predicted clusters by category, gender, region

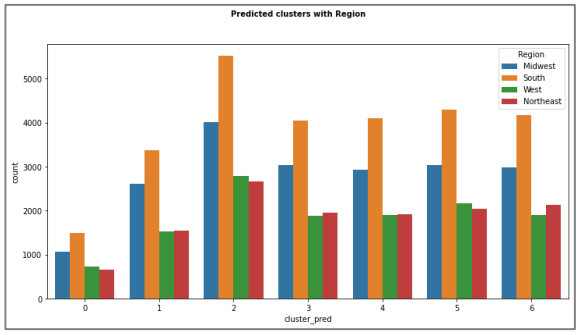


Figure 34:Predicted clusters by region in GMM

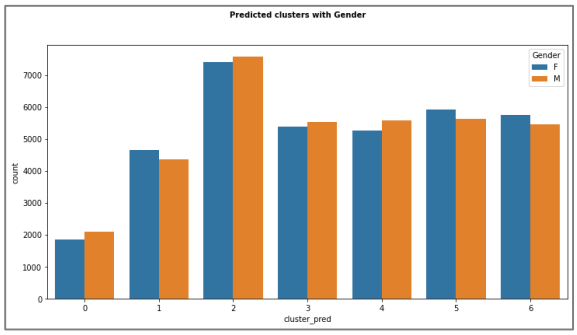


Figure 33: Predicted clusters by gender in GMM

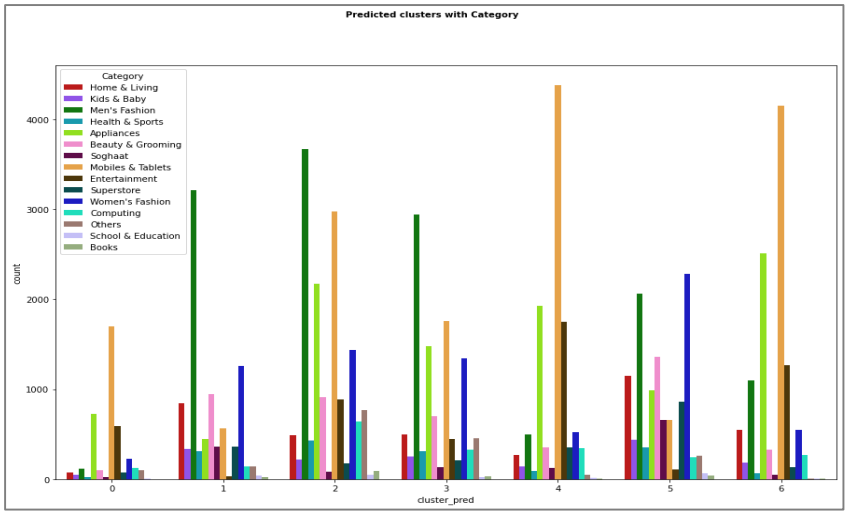


Figure 32: Predicted clusters by category in GMM

Table 5 contains the summary of predicted clusters by gender (in Figure 32) , region (in Figure 33) and category (in Figure 34)

Table 5:Summary of predicted clusters in GMM

Cluster Number	Category		Region		Gender	
	Highest	Lowest	Highest	Lowest	Highest	Lowest
0	Mobile & Tablets	School & Education	South	Northeast	Female	Male
1	Men's Fashion	Books	South	West	Male	Female
2	Men's Fashion	School & Education	South	Northeast	Female	Male
3	Men's Fashion	School & Education	South	West	Female	Male
4	Mobile & Tablets	Books	South	West	Female	Male
5	Women's Fashion	Books	South	Northeast	Male	Female
6	Mobile & Tablets	Books	South	West	Male	Female

4.2.1.4 RFM based analysis

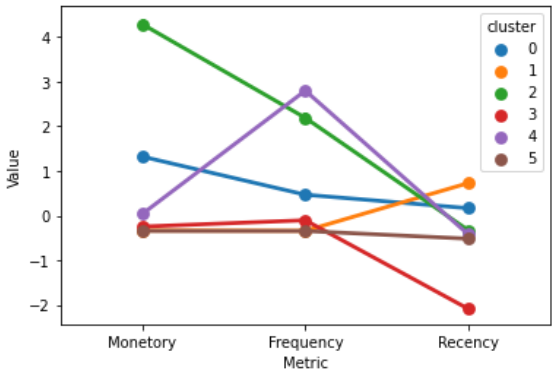


Figure 36: Snakeplot of Kmean clusters

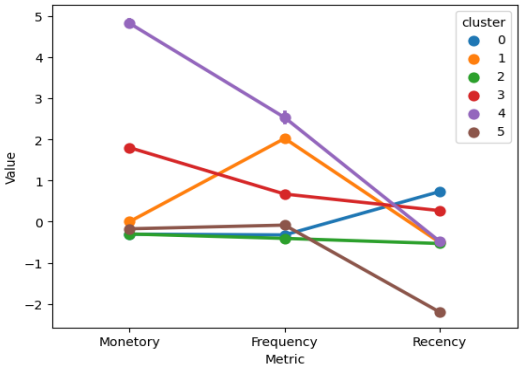


Figure 35 : Snakeplot of agglomerative clusters

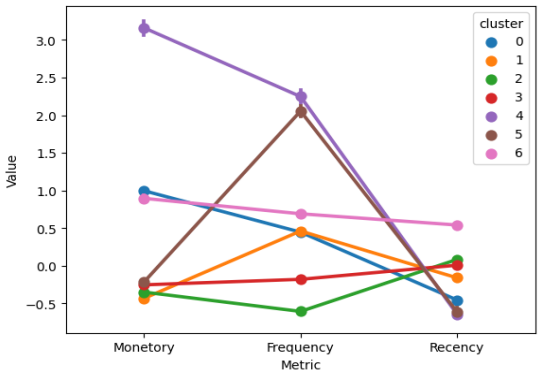


Figure 37: Snakeplot of GMM clusters

Table 6 contains the summary of above snakeplots.

Table 6: RFM Interpretation on completed data

Type of Customer	**K-mean**		**Agglomerative**		**GMM**	
	Cluster number	*Total as a percentage*	*Cluster number*	*Total as a percentage*	*Cluster number*	*Total as a percentage*
Loyal Customers	5,2	11.15%	3,4	9.84%	0,4	8.47%
New customers	3,4	15.40%	1,5	15.09%	1,5	13.94%
Lost/Churned	1	45.32%	0	45.75%	2	50.08%
At Risk	0	28.16%	2	29.30%	3,6	27.47%

4.2.1.5 K Mode Clustering Technique

4.2.1.5.1 Encoding

First created a categorical table and applied Label encoding because Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.

Category	Payment_Method	By_St	Gender	State	Is_Discount	Region	age_bin
Men's Fashion	cod	Net	F	OK	False	South	30-45
Men's Fashion	cod	Net	F	OK	False	South	30-45
Appliances	Easypay	Net	M	FL	True	South	19-30
Appliances	Easypay	Net	M	FL	True	South	19-30
Home & Living	Easypay	Net	M	FL	True	South	19-30

```
from sklearn import preprocessing
le = preprocessing.LabelEncoder()
cat_df = cat_df.apply(le.fit_transform)
cat_df.head()
```

	Category	Payment_Method	By_St	Gender	State	Is_Discount	Region	age_bin
2	8	6	0	0	36	0	2	2
3	8	6	0	0	36	0	2	2
24	0	0	0	1	9	1	2	1
25	0	0	0	1	9	1	2	1
26	6	0	0	1	9	1	2	1

Figure 38: Label encoding

4.2.1.5.2 Cost function elbow graph

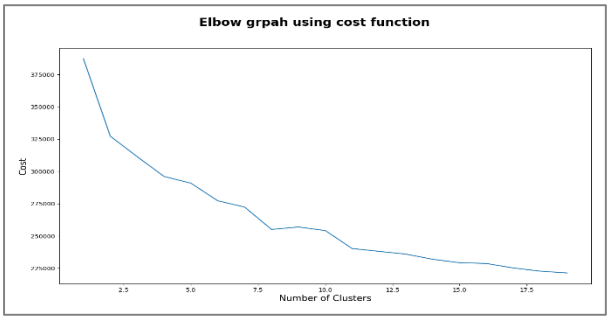


Figure 39: Cost function elbow graph in K mode algorithm

Then We Identified the optimal number of clusters as 10 using the Cost Function Elbow graph.

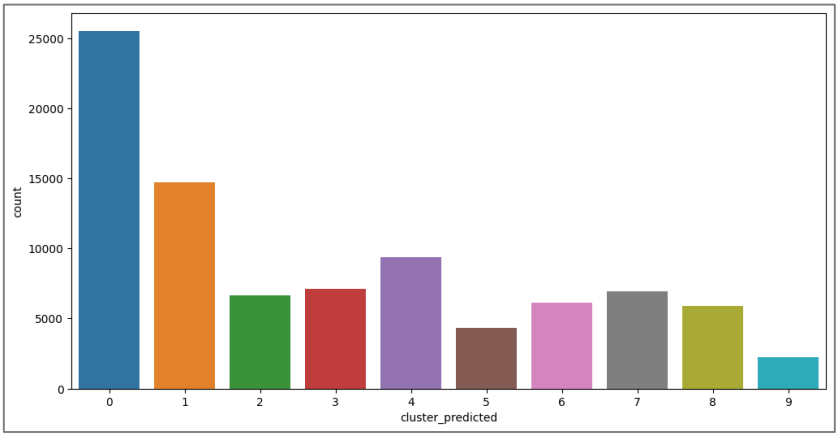
4.2.1.5.3 Model building

```
## Choosing K=10
km_cao = KModes(n_clusters=10, init = "Cao", n_init = 1, verbose=1)
fitClusters_cao = km_cao.fit_predict(cat_df)

Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run 1, iteration: 1/100, moves: 16636, cost: 254195.0
```

Figure 40: K mode model building

4.2.1.5.4 Cluster visualization



0	25538
1	14739
2	6650
3	7086
4	9375
5	4328
6	6157
7	6948
8	5892
9	2255

Figure 41: Cluster counts of K mode algorithm

4.2.1.5.5 Predicted clusters by category, gender and region

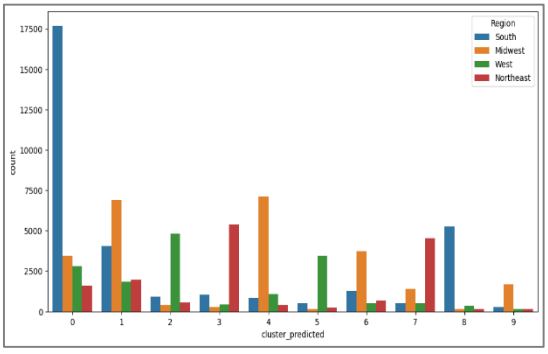


Figure 43: Predicted clusters by region in K mode clusters

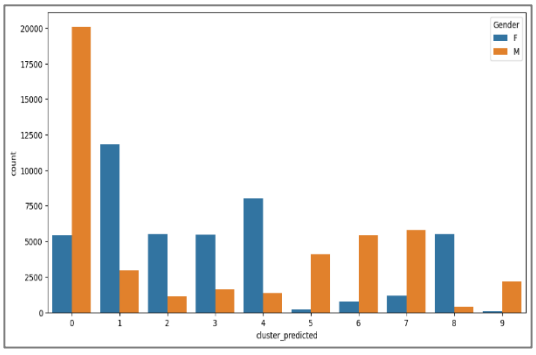


Figure 42: Predicted clusters by gender K mode clusters

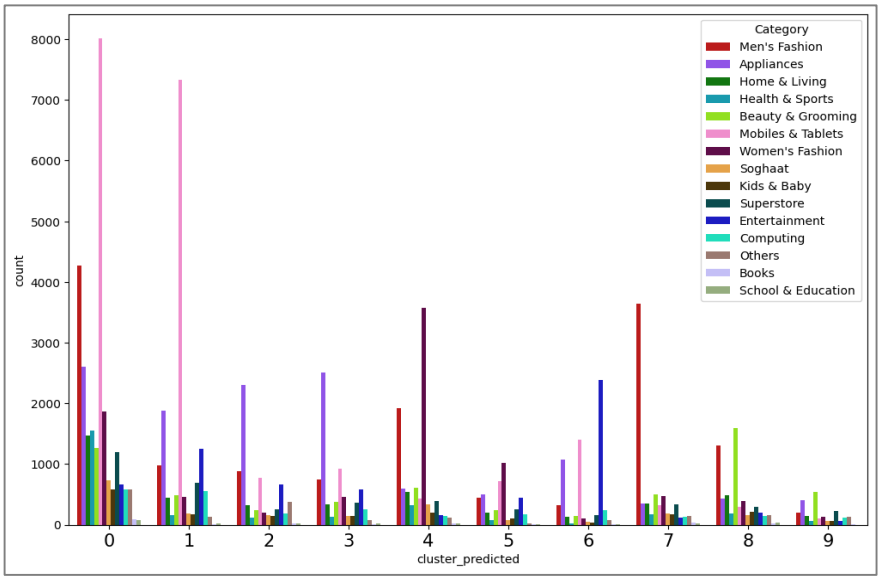


Figure 44: Predicted clusters by category K mode clusters

Table 7 contains the summary of predicted clusters by gender (in Figure 42) , region (in Figure 43) and category (in Figure 44).

Table 7:Summary of predicted clusters in K mode clusters

Cluster	Details
0	category- highest in men's fashion Gender- highest male Region- highest south region
1	category- highest in mobile & tablets Gender- highest female Region- highest Midwest region
2	category- highest in men's fashion Gender- highest female Region- highest West region
3	category- highest in Appliances Gender- highest female Region- highest North East region
4	category- highest in Women's fashion Gender- highest female Region- highest Midwest region
5	category- highest in Mobile and tablets Gender- highest male Region- highest Midwest region

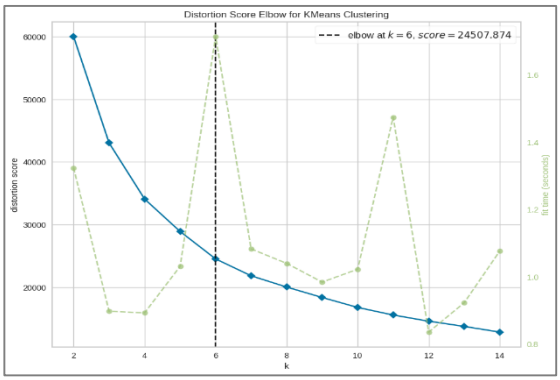
6	category- highest in Appliances Gender- highest male Region- highest North East region
7	category- highest in Men's Fashion Gender- highest female Region- highest South region
8	category- highest in Mobile and tablets Gender- highest male Region- highest West region
9	category- highest in Mobile and tablets Gender- highest female Region- highest South region

4.2.2 Results of the order canceled data

Now we applied these techniques in to order canceled data.

4.2.2.1 K Mean Clustering Technique

4.2.2.1.1 Elbow Graph



Elbow appear at K = 6. Therefore, optimal number of clusters is 6.

Figure 45: LBow graph of kmean Cancelled data

4.2.2.1.2 Model building

After that built a Kmean model using optimal k as 6.

```
kmeans_scaled = KMeans(6)
kmeans_scaled.fit(x_scaled)
```

```
KMeans(n_clusters=6)
```

Figure 46:kmean Model building for Cancelled data

4.2.2.1.3 Cluster visualization

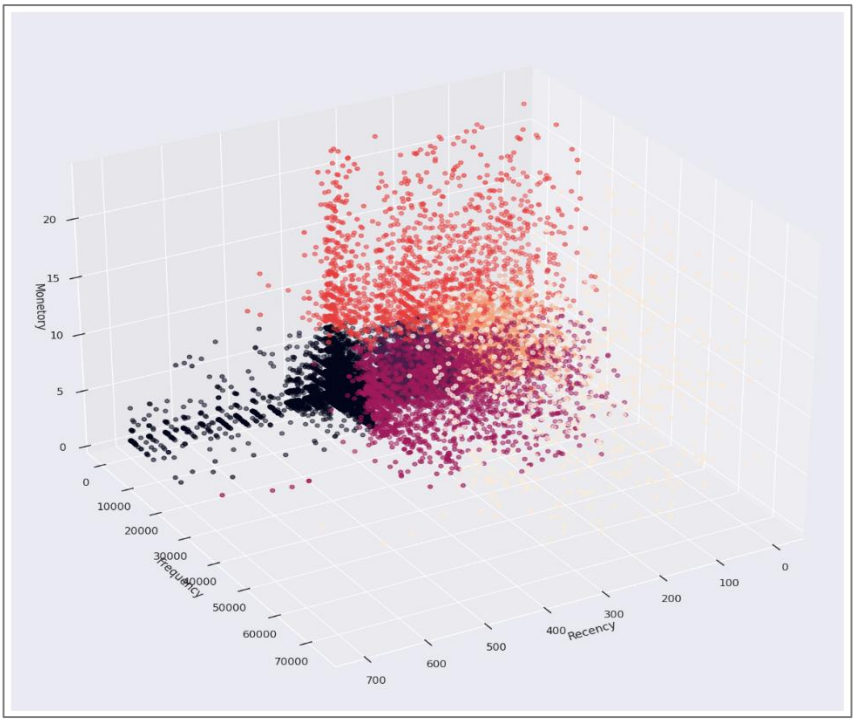


Figure 47: 3D Scatteplot of Kmean(Cancelled data)

4.2.2.1.4 Summary statistics of all the clusters

	Recency			Frequency			Monetory			
	mean	min	max	mean	min	max	mean	min	max	count
cluster										
0	360.847789	290	699	1.872760	1	9	2717.802283	0.00	23790.000	10492
1	224.884352	147	294	1.811667	1	6	2113.380827	0.00	13799.800	10800
2	248.330489	0	698	4.149079	1	11	20514.733167	7739.96	45900.000	2227
3	223.750314	0	547	11.557862	7	23	6606.432210	0.00	31718.347	1590
4	86.015174	0	175	2.357562	1	8	3589.191763	0.00	23873.700	4086
5	172.066667	0	638	9.991667	1	23	47704.544007	26133.08	73336.176	720

Figure 48: Summary statistics of all the clusters in kmean(cancelled)

4.2.2.2.5 Predicted clusters by region, gender and category

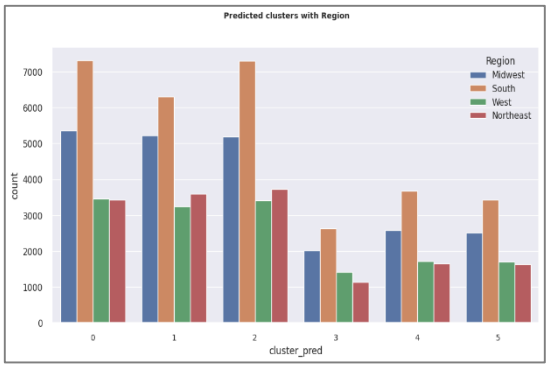


Figure 50: Predicted clusters by region in Kmean clusters(Cancelled data)

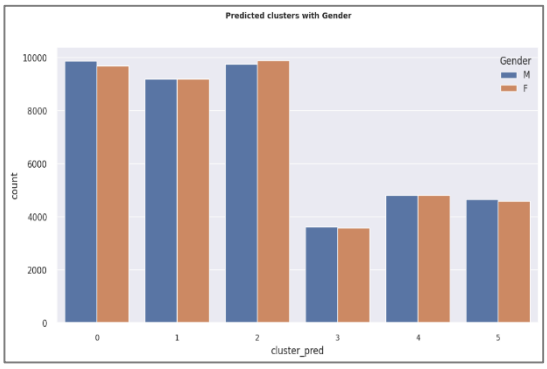


Figure 49: Predicted clusters by gender in Kmean clusters(Cancelled data)

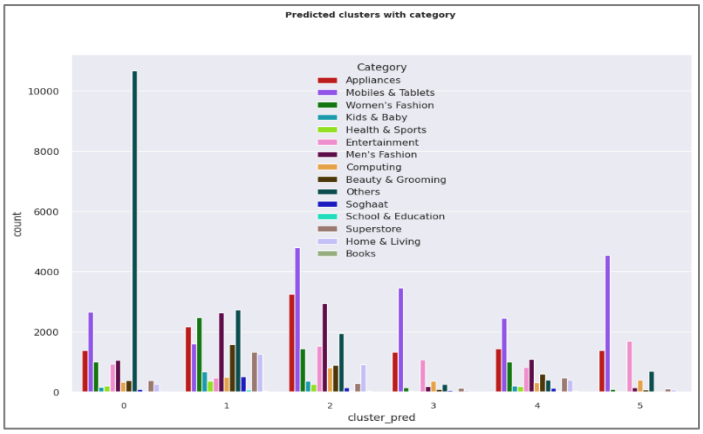


Figure 51: Predicted clusters by category in Kmean clusters(Cancelled data)

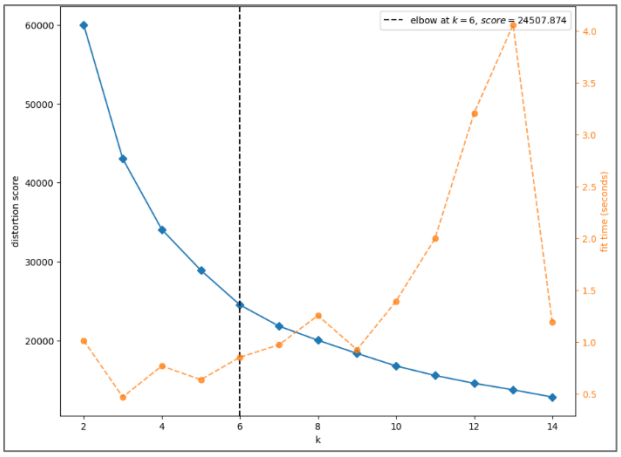
Table 8 contains the summary of predicted clusters by gender (in Figure 49) , region (in Figure 50) and category (in Figure 51).

Table 8: Summary of predicted clusters in K means (canceled data)

Cluster Number	Category		Region		Gender	
	Highest	Lowest	Highest	Lowest	Highest	Lowest
0	Others	School & Education	South	Northeast	Male	Female
1	Others	Books	South	West	Equal	Equal
2	Mobile & Tablets	School & Education	South	West	Female	Male
3	Mobile & Tablets	School & Education	South	Northeast	Male	Female
4	Mobile & Tablets	Books	South	Northeast	Equal	Equal
5	Mobile & Tablets	School & Education	South	Northeast	Male	Female

4.2.2.2 Agglomerative Clustering Technique

4.2.2.2.1 Elbow graph



Elbow appear at K=6.

Therefore, optimal number of clusters is 6.

Figure 52: Elbow graph for Agglomerative Clustering (Cancelled data)

4.2.2.2.2 Dendrogram for agglomerative model

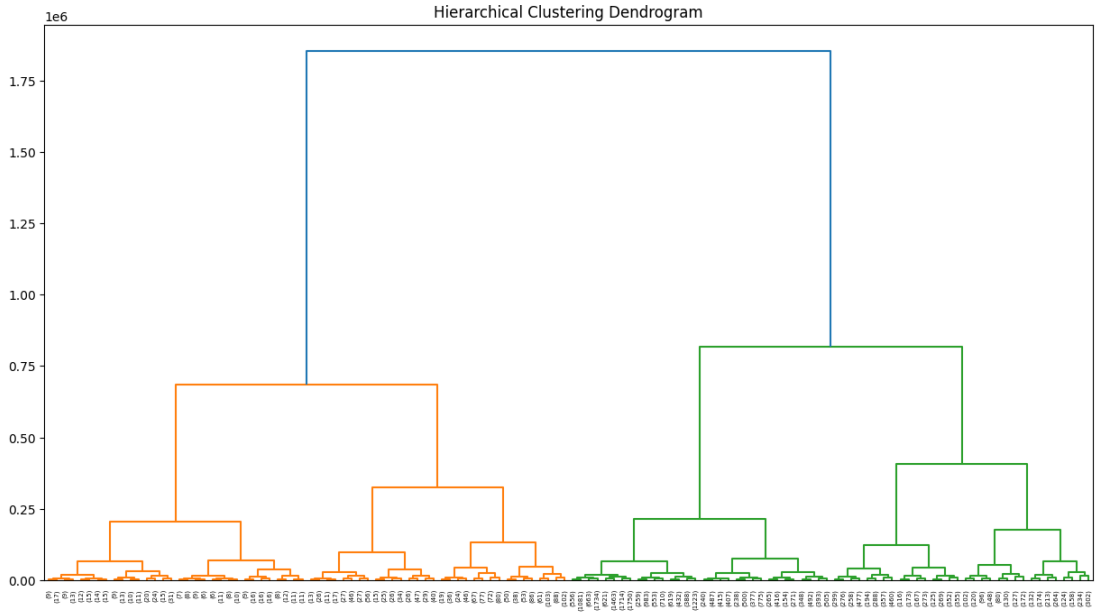


Figure 53: Dendrogram for agglomerative model (cancelled data)

4.2.2.2.3 Model building

After that built an Agglomerative model using optimal k as 6.

```
from sklearn.cluster import KMeans,AgglomerativeClustering,DBSCAN
# apply agglomerative algorithm
agglo_model = AgglomerativeClustering(linkage="ward",n_clusters=6)
agglomerative_clusters = agglo_model.fit_predict(x_scaled)
agglomerative_clusters
```

```
array([4, 1, 4, ..., 5, 4, 5])
```

Figure 54: Model bulding of agglomerative model (cancelled data)

4.2.2.2.4 Cluster visualization

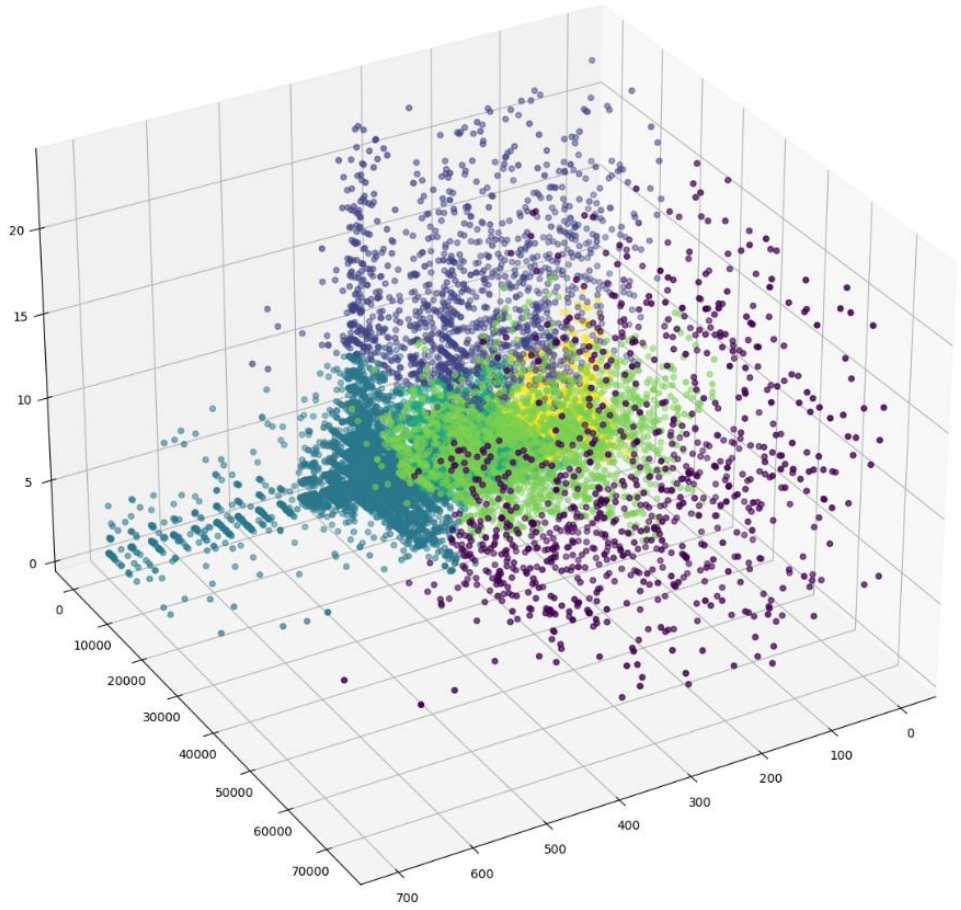


Figure 55: 3D Scatteplot of Kmean(Completed data)

4.2.2.2.5 Summary statistics of all clusters

```
rfm_df_AggClusters.groupby('agglomerative_clusters').agg({
    'Recency' : ['mean','min','max'],
    'Frequency' : ['mean','min','max'],
    'Monetory' : ['mean','min','max','count']
})
```

	Recency			Frequency			Monetory			
	mean	min	max	mean	min	max	mean	min	max	count
agglomerative_clusters										
0	197.721186	0	638	8.152580	1	23	44697.052050	19989.535	73336.176	911
1	197.953147	0	547	11.431469	6	23	5714.448259	0.000	34974.870	1430
2	361.641130	263	699	2.034059	1	10	3455.645714	0.000	32226.800	10834
3	222.536964	66	320	1.774165	1	6	2058.168857	0.000	13799.800	11349
4	191.223137	0	455	4.811857	1	16	16893.447139	3589.800	40590.770	2429
5	76.852465	0	156	2.034099	1	9	1965.381205	0.000	13265.720	2962

Figure 56: Summary statistics of Agglomerative Clustering(Cancelled data)

4.2.2.2.6 Predicted clusters by region, gender and category

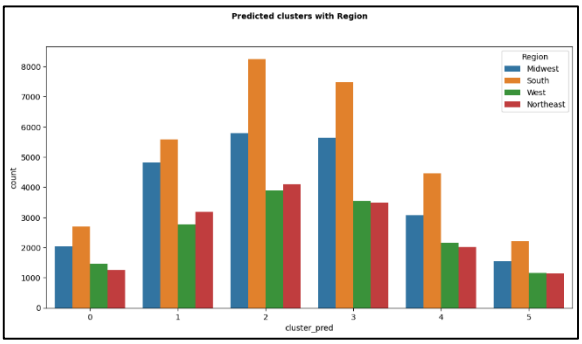


Figure 57:Predicted clusters by region in Agglomerative clusters (Cancelled data)

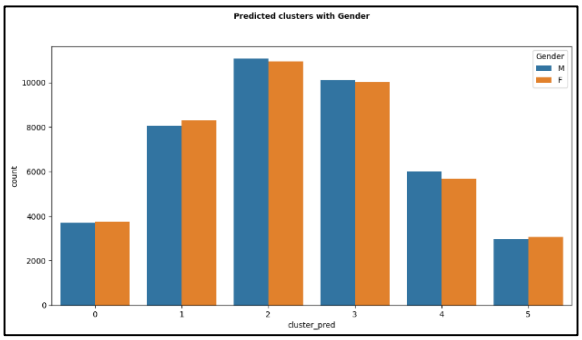


Figure 58: Predicted clusters by gender in Agglomerative clusters (Cancelled data)

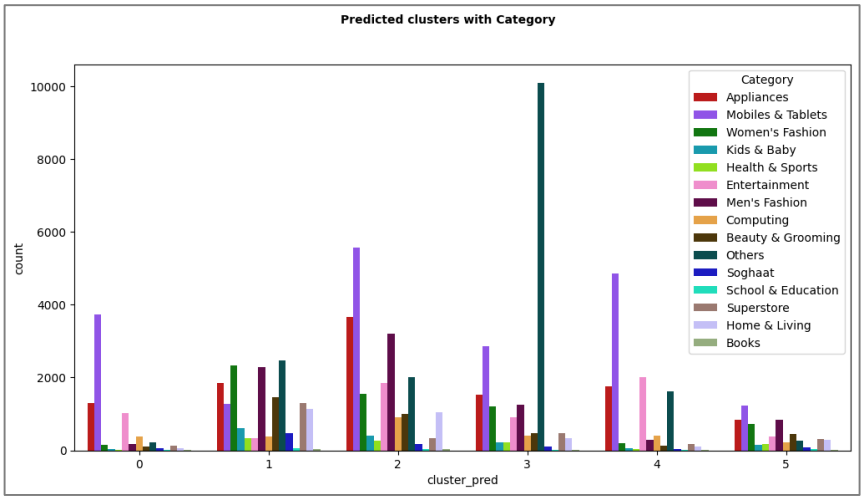


Figure 59: Predicted clusters by category in Agglomerative clusters(Cancelled data)

Table 9 contains the summary of predicted clusters by region (in Figure 57) , gender (in Figure 58) and category (in Figure59).

Table 9: Summary of predicted clusters in Agglomerative Clustering (Cancelled)

Cluster Number	Category		Region		Gender	
	Highest	Lowest	Highest	Lowest	Highest	Lowest
0	Mobile & Tablets	Books	South	Northeast	Female	Male
1	Others	Books	South	West	Female	Male
2	Mobile & Tablets	Books	South	West	Male	Female
3	Others	School & Education	South	Northeast	Male	Feale
4	Mobile & Tablets	Books	South	Northeast	Male	Female
5	Mobile & Tablets	Books	South	Northeast	Female	Male

4.2.2.3 Gaussian Mixture Model (GMM) Clustering Technique

4.2.2.3.1 Gaussian distribution

Before applying the model, we checked about is data followed a Gaussian distribution or not. Here used the Shapiro-Wilk test to determine the distribution.

```
# Extract the RFM values into a numpy array
rfm_values = rfm_df[['Recency', 'Frequency', 'Monetory']].values

# Perform Shapiro-Wilk test on the RFM values
stat, p = shapiro(rfm_values)

# Print the results
print('Statistic=%.3f, p=%.3f' % (stat, p))
if stat < 1.96:
    print('The RFM data is likely Gaussian.')
else:
    print('The RFM data is likely not Gaussian.')

Statistic=0.332, p=0.000
The RFM data is likely Gaussian.
```

Figure 60: Shapiro Wilk Test (Cancelled data)

The above results concluded that the data follows a Gaussian distribution at a 5% significant level. Then, We Identified the optimal number of clusters as 9 using the BIC Score graph.

4.2.2.3.2 BIC score graph

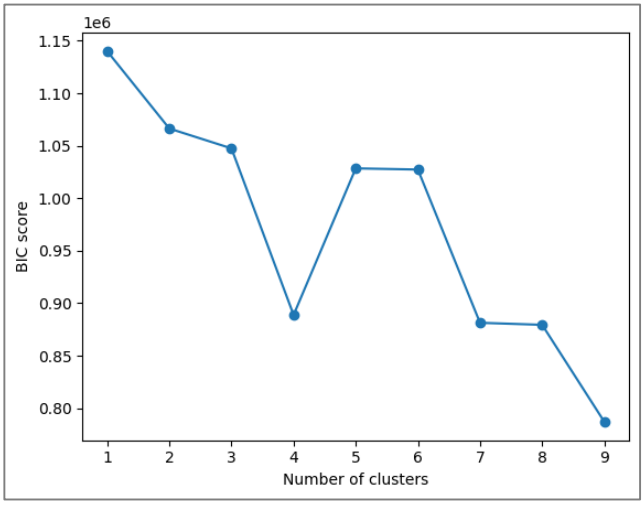


Figure 61: BIC graph for GMM (Cancelled data)

By this BIC graph, obtained optimal k as 9.

4.2.2.3.3 Model building

```
# Fit the Gaussian mixture model
gmm = GaussianMixture(n_components=9, random_state=42)
gmm.fit(rfm_df_GauClusters)

# Predict the clusters
clusters = gmm.predict(rfm_df_GauClusters)
```

Figure 62: Model building for GMM(Cancelled data)

4.2.2.3.4 Cluster visualization

Again, Clusters visualized by Using 3D scatter plot.

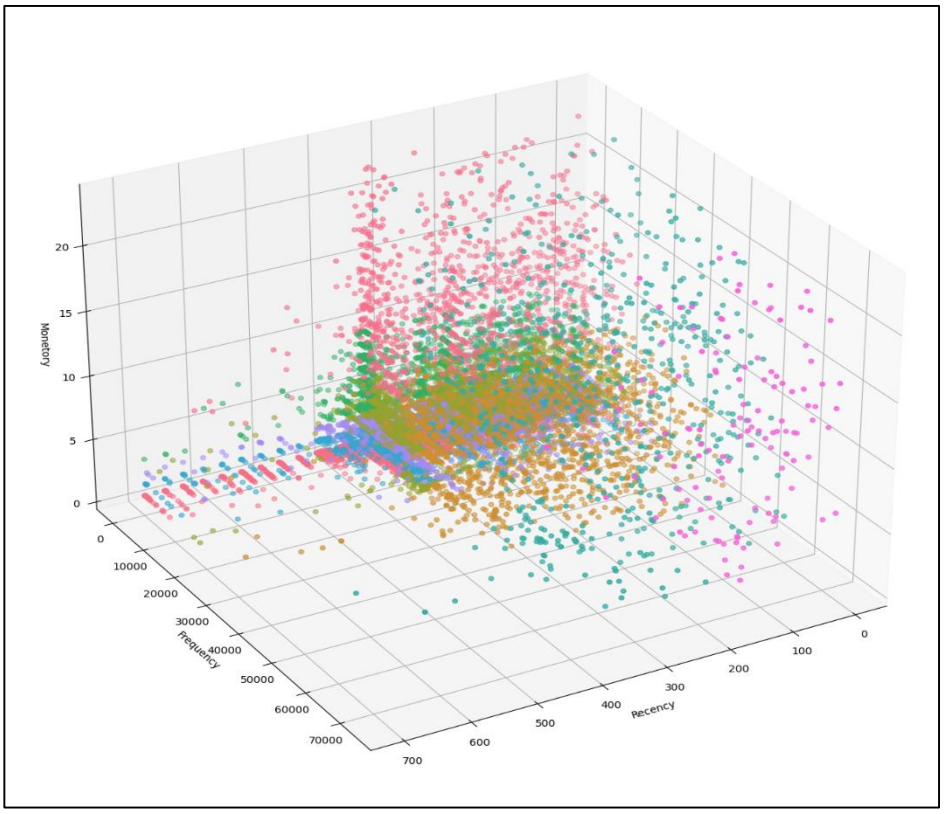


Figure 63: 3D Scatterplot of GMM (Cancelled data)

4.2.2.3.5 Summary statistics of all clusters

	Recency			Frequency			Monetory			
	mean	min	max	mean	min	max	mean	min	max	count
GauCluster										
0	265.202272	0	699	1.000000	1	1	1797.316399	0.00	12999.500	13556
1	218.419901	0	698	5.053634	1	11	25688.530428	13496.20	46679.600	1417
2	245.799045	0	698	4.263484	1	7	9550.338396	2840.00	19400.000	2095
3	244.767865	0	699	5.444740	4	10	1335.055303	0.00	3659.760	2253
4	197.953287	0	638	11.456747	1	23	41308.187642	10720.02	73152.600	578
5	258.307587	0	699	2.000000	2	2	3283.341889	0.00	21315.200	5865
6	252.925938	0	699	3.000000	3	3	4971.981899	0.00	29496.600	2957
7	116.741935	0	419	12.500000	3	23	64691.108766	51762.50	73336.176	124
8	220.695327	0	639	12.121495	1	23	6453.508248	0.00	18752.940	1070

Figure 64: Summary statistics of GMM Clustering(Cancelled data)

4.2.2.3.6 Predicted clusters by category

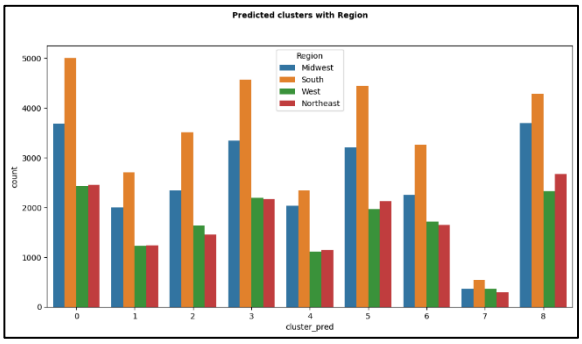


Figure 66: Predicted clusters by region in GMM clusters (Cancelled data)

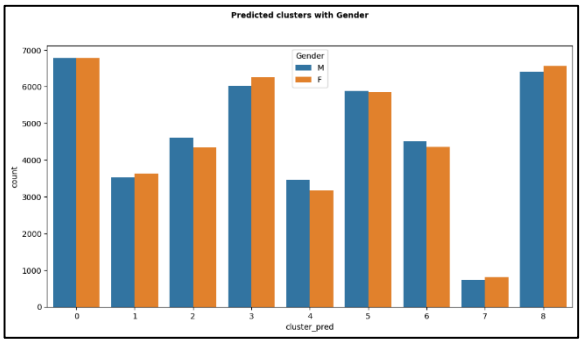


Figure 65: Predicted clusters by gender in GMM clusters (Cancelled data)

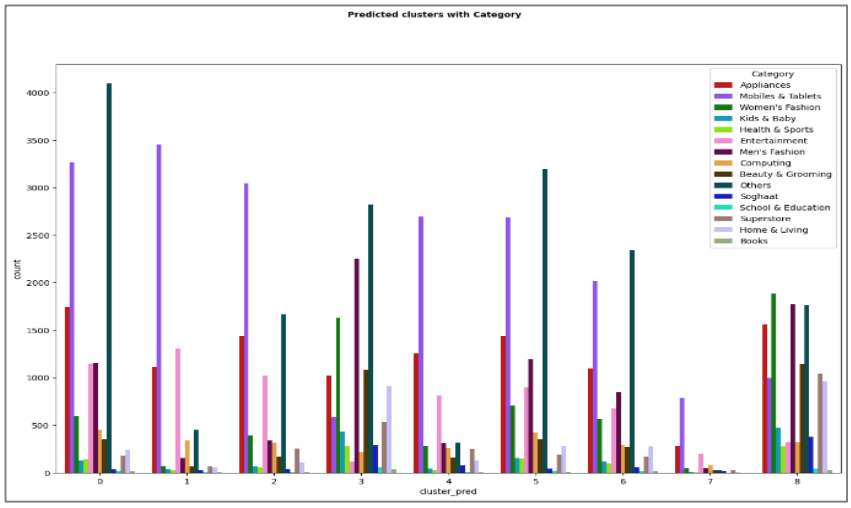


Figure 67: Predicted clusters by category in GMM clusters (Cancelled data)

Table 10 contains the summary of predicted clusters by gender (in Figure 65) , region (in Figure 66) and category (in Figure 67).

Table 10: Summary of predicted clusters in GMM (canceled data)

Cluster Number	Category		Region		Gender	
	Highest	**Lowest**	**Highest**	**Lowest**	**Highest**	**Lowest**
0	Others	Books	South	West	Equal	Equal
1	Mobile & Tablets	Books	South	West	Female	Male
2	Mobile & Tablets	Books	South	Northeast	Male	Feale
3	Others	Books	South	Northeast	Female	Male
4	Mobile & Tablets	Books	South	West	Male	Female
5	Others	Books	South	West	Male	Female
6	Others	Books	South	Northeast	Male	Female
7	Mobile & Tablets	Entertainment	South	Northeast	Female	Male
8	Women's Fashion	Books	South	West	Female	Male

4.2.2.4 RFM analysis

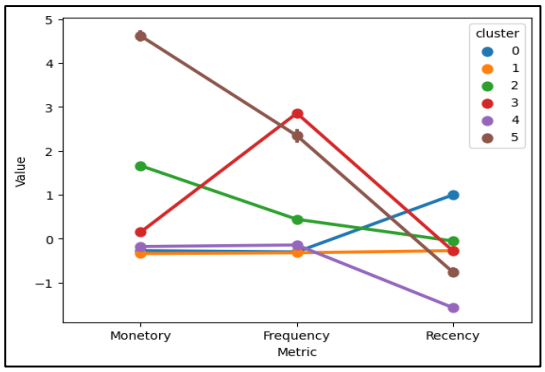


Figure 68:Snakeplot of K mean clusters (canceled data)

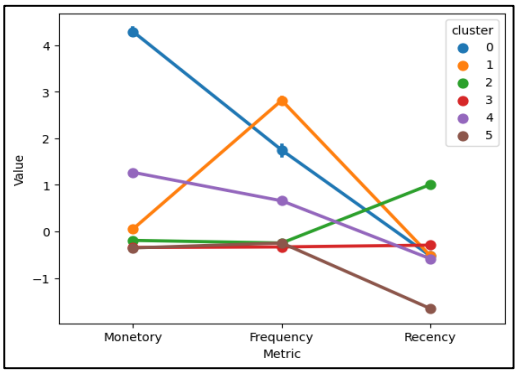


Figure 69:Snakeplot of agglomerative clusters (cancel data)

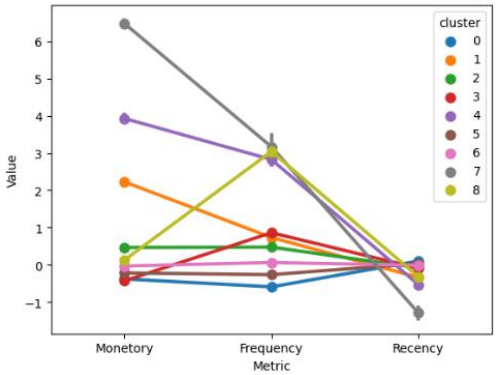


Figure 70:Snakeplot of Kmean clusters

Table 11 contains summary of above snakeplots.

Table 11:RFM interpretation (canceled data)

Type of Customer	**K-mean**		**Agglomerative**		**GMM**	
	Cluster number	*Total as a percentage*	*Cluster number*	*Total as a percentage*	*Cluster number*	*Total as a percentage*
Loyal Customers	5,4	16.06%	4,0	11.16%	1,2,4,7	14.08%
New customers	3	5.31%	1	4.78%	3,8	11.10%
Lost/Churned	0,1	71.17%	2,3	74.15%	0,5	64.92%
At Risk	2	7.44%	5	9.90%	6	9.88%

Error! Bookmark not defined.

4.3. Discussion

As shown in Figure 4, total sales for 2021 are greater than for 2020, indicating that the Covid-19 pandemic has an impact on overall sales. This project successfully identified four customer segments based on their purchasing behavior. These segments can be used by the Amazon sales company to develop targeted marketing strategies and improve customer satisfaction.

According to Table 6, risk customer percentages of K mean, agglomerative and GMM clusters are nearly equivalent which is 28.3% from total order completed customers. Also, the new customer's percentage of all 3 techniques are nearly equivalent which is 14.81% from total order completed customers. In k mean and agglomerative clustering techniques percentage of lost/churned customers from order completed customers are nearly equivalent which is around 45%, but in GMM it is around 50%. The percentage of loyal customers from total order completed customers are differing each other in all 3 clustering techniques.

In K mode clustering we conclude that comparing to other payment methods all cluster shows highest frequency in cash on delivery method. Also, in Figure 43 illustrates that the 0th 7th and 9th clusters contains more customers from southern region. by using k mode algorithm, we cannot get a clear picture of customers that we are interested.

In K means, agglomerative and GMM clustering techniques for order canceled data illustrates that only the at-risk customer segment is nearly equivalent in all 3 techniques. As an average percentage it is around 8.5%.

4.3.1 Model Validation

The Silhouette score in the clustering algorithm is between -1 and 1. This score represents how well the data point has been clustered, and scores above 0 are seen as good, while negative points mean our algorithm has put that data point in the wrong cluster.

4.3.1.1. Completed dataset

```
from sklearn.cluster import KMeans, AgglomerativeClustering
from sklearn.mixture import GaussianMixture
from sklearn.metrics import silhouette_score, adjusted_rand_score, normalized_mutual_info_score

silhouette_kmeans = silhouette_score(x_scaled, labels_kmeans)
silhouette_agglo = silhouette_score(x_scaled, labels_agglo)
silhouette_gmm = silhouette_score(x_scaled, labels_gmm)
```

```
print("KMeans - Silhouette Score:", silhouette_kmeans)
print("Agglomerative Clustering - Silhouette Score:", silhouette_agglo)
print("GMM - Silhouette Score:", silhouette_gmm)
```

```
KMeans - Silhouette Score: 0.3933125871986843
Agglomerative Clustering - Silhouette Score: 0.3683009299934663
GMM - Silhouette Score: 0.11901770415761323
```

Figure 71: Silhouette scores of all 3 methods (Completed data)

All the three clustering techniques silhouette score is greater than 0. Therefore, we can conclude that all the three methods are useful for clustering this canceled dataset based on RFM values.

4.3.1.2. Cancelled dataset

```
from sklearn.cluster import KMeans, AgglomerativeClustering
from sklearn.mixture import GaussianMixture
from sklearn.metrics import silhouette_score, adjusted_rand_score, normalized_mutual_info_score

silhouette_kmeans = silhouette_score(x_scaled, labels_kmeans)
silhouette_agglo = silhouette_score(x_scaled, labels_agglo)
silhouette_gmm = silhouette_score(x_scaled, labels_gmm)
```

```
print("KMeans - Silhouette Score:", silhouette_kmeans)
print("Agglomerative Clustering - Silhouette Score:", silhouette_agglo)
print("GMM - Silhouette Score:", silhouette_gmm)
```

```
KMeans - Silhouette Score: 0.4109093040322848
Agglomerative Clustering - Silhouette Score: 0.37730673789777047
GMM - Silhouette Score: 0.06306936227121537
```

Figure 72: : Silhouette scores of all 3 methods (Cancelled data)

All the three clustering techniques silhouette score is greater than 0. Therefore, we can conclude that all the three methods are useful for clustering this canceled dataset based on RFM values.

4.3.2 Actions to take for the retaining the customers

Table 12: Actions for retaining customers

Customer Segment	Actionable Insight
Loyal Customers Spend good money with us often & are responsive to promotions	• Upsell higher-value products • Ask for reviews • Engage them • offer membership/loyalty programs • recommend other products to them
New Customers Bought more recently but haven't spent much	• Provide onboard supporting • Give them early success • Start building relationships
At Risk Customers Spent big money and purchased often but haven't purchased for a long time ago	• Send personalized emails to reconnect • offer more discounts • Win them back with newer products • Reconnect with them
Lost Customers Lowest Recency, Frequency & Monetary scores	• revive interest in reach out campaign • Recreate brand value • offer other relevant products and special discounts

CHAPTER 05

5.1 Limitations and Challenges

There are several limitations and challenges associated with customer segmentation in the retail sector. Some of the key challenges include:

1. **Data quality and availability**: One of the main challenges faced in customer segmentation with cluster analysis is the availability and quality of data. It is essential to have accurate and comprehensive data on customer characteristics and behavior in order to effectively segment them. However, data may be incomplete, outdated, or inaccurate, which can hinder the accuracy of the cluster analysis.
2. **Identifying relevant variables**: Another challenge is identifying the most relevant variables to include in the cluster analysis. There may be a large number of variables available, but not all of them may be relevant or useful for customer segmentation. It is important to carefully select variables that are meaningful and relevant to the customer segments being targeted.
3. **Ensuring data privacy**: Finally, it is important to ensure that customer data is handled and used in a way that respects privacy and complies with relevant regulations and laws. This can be a challenge when working with large amounts of data and using advanced analytical techniques like cluster analysis.

5.2 Future work

- Derive a model which is suitable for both numerical and categorical variables such as k-prototype algorithm.
- Compare the results of different clustering algorithms: To better understand the strengthens and weakness of each algorithm.
- Incorporate psychographic and Behavioral data :- psychographic data such as social class and lifestyle ,behavioral data such as loyalty status and purchase occasion can be helpful in customer segmentation.
- Upscaling the project with feature selection engineering techniques such as Principle component analysis|(PCA), Linear Discriminant Analysis(LDA) or Future important ranking to identify the more relevant features for clustering.

APPENDIX

Here is the notebook links which were used to implement all above outputs.

- EDA
  https://colab.research.google.com/drive/1K5nW9HO9k1lcStd5kZ36-x86-JeUfB0A?usp=share_link
- Completed data
  https://colab.research.google.com/drive/1I1yDmXSdhRMS_C0jNO0q5lDbkv32Y7Za#scrollTo=xJB58QAa9TI8
- Cancelled Data
  https://colab.research.google.com/drive/1gkxR00uPuTSpfbgAz6u2Hj35ZsgXCwWm#scrollTo=xJB58QAa9TI8

REFERENCES

Aprilliant, A. (2023, January). The k-modes as Clustering Algorithm for Categorical Data Type. *The k-modes as Clustering Algorithm for Categorical Data Type*. Retrieved from https://medium.com/geekculture/the-k-modes-as-clustering-algorithm-for-categorical-data-type-bcde8f95efd7

Aprilliant, A. (2023, January). The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical). *The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical)*. Retrieved from https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb

Ds_mt. (2022, March). Using K-Means to detect changes in a retail store | Towards Data Science | Towards Data Science. *Using K-Means to detect changes in a retail store | Towards Data Science | Towards Data Science*. Retrieved from https://towardsdatascience.com/using-k-means-to-detect-changes-in-a-retail-store-96af1476dd9f

Gauravduttakiit. (2020, August). Clustering using K-means + Hierarchical + PCA. *Clustering using K-means + Hierarchical + PCA*. Retrieved from https://www.kaggle.com/code/gauravduttakiit/clustering-using-k-means-hierarchical-pca

Kilari, H. (2022, April). Customer Segmentation using K-Means Clustering. *IJERT*. doi:10.17577/IJERTV11IS030152

Makara, I. (2021). A Clustering Approach to Market Segmentation Using Integrated Business Data. *A Clustering Approach to Market Segmentation Using Integrated Business Data*. Retrieved from http://erepository.uonbi.ac.ke/handle/11295/160755

Murphy, C. (2022, November). What Is Recency, Frequency, Monetary Value (RFM) in Marketing? *What Is Recency, Frequency, Monetary Value (RFM) in Marketing?* Retrieved from https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp#:~:text=An%20RFM%20analysis%20evaluates%20clients,each%20of%20the%20three%20categories

Nazari, Z., Kang, D., Asharif, M. R., Sung, Y.-W., & Ogawa, S. (2015, November). A new hierarchical clustering algorithm. *2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*. doi:10.1109/iciibms.2015.7439517

Salamzadeh, A., Ebrahimi, P., Soleimani, M., & Fekete-Farkas, M. (2022, September). Grocery Apps and Consumer Purchase Behavior: Application of Gaussian Mixture Model and Multi-Layer Perceptron Algorithm. *Journal of risk and financial management, 15,* 424. doi:10.3390/jrfm15100424

Singh, A. (2022, June). Build Better and Accurate Clusters with Gaussian Mixture Models. *Build Better and Accurate Clusters with Gaussian Mixture Models*. Retrieved from https://www.analyticsvidhya.com/blog/2019/10/gaussian-mixture-models-clustering/

Vipulgandhi. (2019, September). Gaussian Mixture Models Clustering - Explained. *Gaussian Mixture Models Clustering - Explained*. Retrieved from https://www.kaggle.com/code/vipulgandhi/gaussian-mixture-models-clustering-explained