CUSTOMER SEGMENTATION AND PROFILING USING CLUSTERING TECHNIQUE

A Data Analysis and Visualization Project

Presented To:

Dr. A. Shobanadevi (Associate Professor, DSBS, SRM IST)
Vaskar Deka (Associate Professor, IT, Guwahati University)

Presented By:

Pijush Pathak
RA2011027010152
pp5512@gami.com

Abstract

As the grocery market industry evolves and matures, retailers are increasingly motivated to seek data and strategies that can help them segment and understand their customers in a concise yet informative manner. While many grocery operators perceive state-mandated traceability as a necessary requirement, it presents a valuable opportunity for internal customer analysis. Traditional segmentation analysis often focuses on demographics or RFM (Recency-Frequency-Monetary) segmentation. However, these methods alone may not provide deep insights into a customer's purchasing behaviour. This report aims to segment customers using grocery-specific data (such as product preferences ,purchase history , accepted promotional campaigns) and machine learning techniques (such as K-Means and Agglomerative Hierarchical Clustering). The goal is to uncover novel approaches to explore a grocery store's consumer base. The findings reveal the presence of approximately four or five customer clusters, each exhibiting unique purchasing traits that define them. While the results are meaningful, this report could further benefit from exploring additional clustering algorithms, comparing results across different grocery stores within the same region, or investigating segmentations in other geographic markets.

Contents

1 Introduction

1.1 The Business Problem

Collecting, manipulating, and analyzing data is an essential part of any retail company's operations. Data is generated from various sources such as shipments, tickets, employee logs, and digital interactions, providing valuable insights into how a company operates. Access to extensive data allows for a clearer understanding of the business and enables the discovery of previously unseen details that can drive innovation.

However, working with large and complex real-world data poses challenges. While performance metrics and interactive dashboards provide superficial information such as sales figures and top products, they often lack the depth required for advanced data mining and analysis. To optimize commercial practices, companies are increasingly motivated to explore phenomena and data that require more in-depth investigation.

Effective utilization of data science and data mining practices allows companies to delve deeper into their operational strategies, leading to improved optimization. This paper focuses on addressing the question of predicting customer preferences and behaviours, aiming to understand the defining traits of customers, and forecast their purchasing preferences.

If a grocery store wants to expand into new markets, understanding their customer base becomes crucial. It involves not only knowing what products customers prefer but also their purchasing patterns, frequency, and lifetime value.

By integrating machine learning techniques and conventional business understanding, the study explores customer segmentation to uncover purchasing patterns and behaviours of the customers.

1.2 Acquisition of Data

Finding readied, usable data for analysis in a business context is a rarity. As such, it is imperative to collect as much data as possible, but also in a format that meets a wide variety of financial, ethical, and computational considerations.

The acquisition of data in a grocery retail store involves the collection and gathering of various types of information related to the store's operations, customers, products, and transactions. Grocery stores employ multiple methods and technologies to acquire relevant data for analysis and decision-making purposes. Here are some common ways in which grocery retail stores acquire data:

1. **Point of Sale (POS) Systems**: The primary source of data acquisition in grocery stores is through POS systems. These systems record transactional data, including items purchased, prices, quantities, and timestamps. POS systems capture detailed information about sales, inventory levels, and customer interactions at checkout.

2. **Loyalty Programs:** Many grocery retailers offer loyalty programs where customers sign up and provide their personal information in exchange for discounts, rewards, or personalized offers. Loyalty programs enable stores to gather valuable customer data such as demographics, shopping preferences, purchasing habits, and transaction history.
3. **Customer Surveys and Feedback**: Grocery stores often conduct surveys or collect feedback from customers to gather insights into their shopping experiences, preferences, and satisfaction levels. This data helps retailers understand customer needs, identify areas for improvement, and tailor their offerings accordingly.
4. **Website and Mobile Apps**: Online grocery shopping platforms and mobile apps allow retailers to acquire data about customer behaviour, preferences, and browsing patterns. Data on product views, searches, and purchases provide insights into customer interests and enable personalized recommendations.
5. **Supply Chain and Inventory Systems**: Data related to inventory levels, product availability, and supply chain operations are collected through integrated systems. This helps retailers optimize inventory management, track product movements, and identify potential bottlenecks or inefficiencies.
6. **Sensor Technologies**: Some grocery stores use sensors, such as RFID (Radio-Frequency Identification), to track product movement within the store. These sensors provide real-time data on stock levels, shelf replenishment needs, and store layout optimization.
7. **Social Media and Online Reviews**: Monitoring social media platforms and online review sites allows grocery retailers to capture customer feedback, sentiment analysis, and brand perception. This data helps in understanding customer opinions, addressing concerns, and managing reputation.
8. **Market Research and External Data Sources**: Grocery retailers may also acquire data from external sources such as market research firms, industry reports, government statistics, and demographic data. This information provides insights into market trends, competitor analysis, and consumer behaviour outside of the store's own data.

By aggregating and analyzing data from these various sources, grocery retail stores can gain valuable insights into customer preferences, optimize operations, improve inventory management, personalize marketing efforts, and make informed business decisions to enhance the overall shopping experience.

After taking the above processes and considerations into account, it was possible to collect the relevant data in a single query using the software’s SQL editor. The data was then outputted into a CSV file (with around 2240 rows) for easy viewing, importing, and analysis.

1.3 Scope of Analysis

In general, the methods used to gather the data for this project can easily be extended into other relevant contexts/analyses. While there is clear value in using the same data to investigate purchasing patterns or to build an item based collaborative filtering recommender system, neither of these is the focus for this paper. The scope of the paper is limited to the following four intertwined goals:

1. To cluster customers based on common purchasing behaviours for future operations/marketing projects

2. To incorporate best mathematical, visual, programming, and business practices into a thoughtful analysis that is understood across a variety of contexts and disciplines

3. To investigate how similar data and algorithms could be used in future data mining projects

4. To create an understanding and inspiration of how data science can be used to solve real-world problems

Before delving into the details of the project and its implications, the next chapter discusses what customer segmentation analysis is and the reasons for its importance.

2 Customer Segmentation Analysis

2.1 Brief Introduction

For a retailer, understanding the components of their consumer base is key to maximizing their potential in a market; the retailer that attracts the most customers will acquire the most market share. In fact, the high costs of gaining a new customer or getting back an old customer force retailers to seriously consider how to allocate resources to optimize not just volume of customers, but the retention of them as well . Additionally, it is a common understanding in the retail industry that the Pareto Principle—more likely than not—applies to the company: 80% of profits come from 20% of the customers . One crucial reason why this principal hold is because retail businesses thrive on repeat purchases . Therefore, a net change of one customer can significantly impact a business' profit in the long run. Therefore, it is generally in the best interest of the retailer to devote efforts to retaining customers by understanding them on as deep of a level, as necessary.

However, examining the intricate, rich relationships between a retailer and their consumer base involves understanding how different components of the base behave. Namely, how different segments of customers act similarly or differently from other segments . One method of approaching customer understanding is through the lens of customer segmentation. In short, customer segmentation analysis is the process of grouping customers in such a way that customers within one group are like each other but different from customers in other groups. In general, there are two paths of segmentation: a priori and post hoc. A priori analysis involves creating segments or groups of customers based on predetermined criteria or knowledge before examining the actual customer data. The segments are predefined, and customers are then assigned to these segments based on their characteristics. The focus in this approach is on the segments themselves, rather than individual customer data.

On the other hand, post hoc analysis uses the customer data itself to form segments. This approach has become more prominent with advancements in data collection and reliability. Modern retailers and data scientists often prefer post hoc analysis because it leverages the available data to identify meaningful customer segments. This method will be the focus of the paper, as it aligns with the advancements in technology and data collection that have made it a valuable segmentation technique in the retail industry.

While the goal of customer segmentation analysis has been consistent among retailers for many years, approaches in the past relied on much weaker analytical techniques than available today. It is nonsensical to blame companies in the past who failed to utilize their data properly; the technology and data infrastructure simply were not ubiquitous or cheap enough to allow for companies to collect massive amounts of data as they do today. Yet, many companies still found rudimentary methods to attempt to understand their customers, the most traditional involving purely demographic analysis . Demographic analysis is segmenting customers solely on demographic features, such as age, sex, race, or income. It is built upon the assumption that retail behaviour is defined by the demographics of the surrounding neighbourhood of a store's consumer base. Furthermore, demographic analysis also thrived because it became a quick, cheap, and easy model to predict how new customers would interact.

Instead of attempting to divide customers based on their demographics, retailers began segmenting their customers based on their purchasing patterns, mostly using a technique known as the Recency-Frequency Monetary (RFM) method . A standard implementation of the RFM model is cheap and simple: once each of the components are defined in a way that makes them easy to collect, it is a relatively menial task for a retailer to visualize the results, which makes interpretation easy as well.

To perform customer segmentation analysis at a high level, retailers have begun to incorporate aspects of machine learning into the analysis of their customers. More specifically, retailers are utilizing unsupervised machine learning tools such as clustering and dimensionality reduction to approach analysis in ways that cannot be matched without machine learning. Instead of focusing on only a few features or customers at a time, it is possible to write programs and implement algorithms that can consider several more features or several more instances than traditional spreadsheets can hold or process. Because of this massive potential, retailers across all industries are attempting to leverage clustering algorithms such as K-Means or hierarchical clustering to segment their customers more accurately and quickly. The faster and better retailers can cluster their customers, the quicker they can market to them and thus acquire market share.

2.2 Challenges of Performing Analysis

Performing analysis on a dataset from a grocery retail store can come with several challenges. Some common challenges include:

1. **Data Quality**: Ensuring the accuracy, completeness, and consistency of the data can be a challenge. Issues such as missing data, data entry errors, inconsistencies in naming conventions, and data duplication can affect the reliability and validity of the analysis.
2. **Data Integration**: Grocery retail stores often have multiple sources of data, such as point of sale systems, inventory management systems, customer loyalty programs, and online platforms. Integrating data from these disparate sources and aligning them for analysis can be complex and time-consuming.
3. **Large Data Volume**: Grocery retail generates vast amounts of data, including transactional data, customer data, inventory data, and more. Analyzing such large volumes of data requires robust computing resources and efficient data processing techniques.
4. **Data Privacy and Security**: Grocery retail data often contains sensitive information such as customer names, addresses, payment details, and purchasing habits. Ensuring data privacy and implementing robust security measures to protect the data from breaches or unauthorized access is crucial.
5. **Complex Data Relationships**: Data in the grocery retail industry can have intricate relationships. For example, analyzing customer purchasing behavior may require linking transactional data with customer demographics, loyalty

program data, and promotional campaign data. Managing and understanding these complex data relationships is a challenge in analysis.

6. **Data Analysis Skills and Expertise**: Extracting meaningful insights from the grocery retail dataset requires skilled data analysts or data scientists who possess both domain knowledge and expertise in data analysis techniques. Finding and retaining qualified professionals can be a challenge for organizations.
7. **Evolving Data Landscape**: The grocery retail industry is constantly evolving, with new technologies, channels, and customer preferences emerging. Analyzing the dataset needs to consider these changes, and techniques applied today may need to be adjusted in the future to adapt to evolving data landscapes.

Overcoming these challenges requires a combination of data management strategies, advanced analytics tools, data governance practices, and a skilled analytical team. With proper planning, data preparation, and analysis techniques, grocery retail stores can leverage their dataset to gain valuable insights for informed decision-making and improved business outcomes.

3 Clustering Using Machine Learning Methods

While many applications of machine learning, such as regression and classification, focus on predicting the outcome or value of an instance, these applications do not attempt to understand similarities between instances, just the relationship between instances and their respective outputs. Thus, when it comes to searching for algorithms or methods that look for similarities between features of instances, the focus must turn from supervised machine learning to unsupervised machine learning.

In technical terms, clustering is an unsupervised machine learning technique that groups instances into clusters based on the similarities between instances. These just states that clustering is one way of viewing or evaluating data by looking at the natural groupings or segments that separate instances in the data. However, it is difficult to appreciate clustering without first fully understanding what it means for instances to be considered similar.

3.1 Similarity Measures

The success of a clustering algorithm rests upon the ability to choose the proper similarity measure before engaging in clustering. Choosing the best similarity measure, however, depends on an acute awareness of what similarity is and how it can be defined mathematically. Here are some commonly used similarity measures:

1. **Euclidean Distance**: It is the most widely used distance metric and measures the straight-line distance between two data points in Euclidean space. For two data points, A = (a1, a2, ..., an) and B = (b1, b2, ..., bn), the Euclidean distance is calculated as:

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

2. **Manhattan Distance**: Also known as city block distance or L1 distance, it measures the sum of absolute differences between the coordinates of two data points. For two data points A = (a1, a2, ..., an) and B = (b1, b2, ..., bn), the Manhattan distance is calculated as:

$$d(A, B) = |a1 - b1| + |a2 - b2| + ... + |an - bn|$$

3. **Cosine Similarity**: It measures the cosine of the angle between two vectors and is commonly used for text or document clustering.
For two vectors A = (a1, a2, ..., an) and B = (b1, b2, ..., bn), the cosine similarity is calculated as:

$$sim(A, B) = (A . B) / (||A|| * ||B||)$$

where ,

(A . B) denotes the dot product of A and B, and
||A|| and ||B|| denote Euclidean norms of A and B

4. **Pearson Correlation Coefficient**: It measures the linear correlation between two variables and is suitable for clustering techniques that involve continuous or interval data.
   It ranges from -1 (strong negative correlation) to 1 (strong positive correlation).
   For two variables x and y, the Pearson correlation coefficient is calculated as:
   where n is the number of data points.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

5. **Jaccard Similarity**: It is a measure of similarity between two sets and is commonly used in clustering techniques that deal with categorical data or binary attributes. For two sets A and B, the Jaccard similarity is calculated as:

$$\text{sim}(A, B) = |A \cap B| / |A \cup B|$$

These similarity measures can be used in various clustering algorithms such as k- v means, hierarchical clustering, and DBSCAN, among others. The choice of similarity measure depends on the nature of the data and the requirements of the clustering task.

3.2 Centroid-based: K-Means

K-means clustering is widely used in the field of cluster analysis and customer segmentation. K-means is an algorithm designed to group a set of items into K subgroup or clusters. The algorithm is dependent on a manually set value for K. The K centroids are initialized to random observations in the dataset. K-means is then tasked with iteratively moving these centroids to minimize the cluster variance using two steps:

- For each centroid c identify the subset of items that are closer to c than any other centroid using some similarity measure.
- Calculate a new centroid each cluster after every iteration which is equal to the mean vector of all the vectors in the cluster.

This two-step process is repeated until convergence is reached.
The standard implementation of K-means uses Euclidian distance measure described in the section above to find the subset of items that corresponds to each cluster. This is done by calculating mean squared error, which in this case is equivalent with the Euclidian distance, of each item's feature vector with the K centroid and choosing the closest result. However, other distance measures can be used instead of Euclidian distance. Aggarwal et al. claim that for high dimensional data, the choice of distance measure used in clustering is vital for its success.

3.3 Hierarchical based: Agglomerative

Hierarchical clustering is a type of clustering that involves establishing a hierarchy in terms of how similar two clusters are. In agglomerative clustering, each datum starts as its own individual cluster and proceeds to join with the most similar cluster. With numeric data and working with Euclidean Space, the Euclidean Distance Formula is the most common distance/dissimilarity metric. However, since the number of clusters updates each iteration of the algorithm, it is important to keep track of the distance between each cluster to every other cluster via a distance matrix. Once there is only one cluster remaining, the algorithm converges, and the results are commonly visualized with a dendrogram.

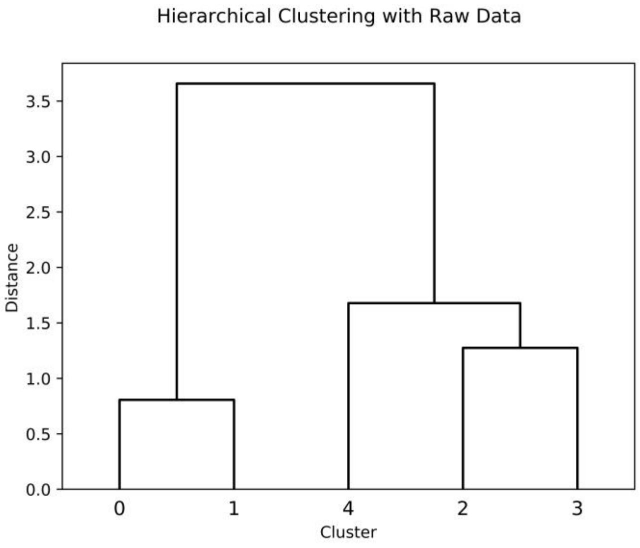

Fig: Dendrogram of Clustered Random Raw Data

However, agglomerative clustering also has limitations. As the algorithm proceeds, the computational complexity increases, making it less efficient for large datasets. The choice of linkage criterion can significantly impact the resulting clusters, and different criteria may yield different outcomes. Furthermore, the method suffers from the "chaining effect," where early merging decisions can propagate errors and affect subsequent merges.

In conclusion, hierarchical-based agglomerative clustering is a versatile technique for data grouping and exploration. It offers flexibility in determining cluster structure and facilitates insights into data patterns at various levels. While it has certain limitations, careful consideration of the linkage criterion and the chaining effect can help mitigate potential drawbacks, making agglomerative clustering a valuable tool in data analysis and exploration.

4 Preparing the Data

The digressions of clustering and customer segmentation analysis were important, but it is now time to think back to the previously stated business problem and the associated data. Although several variables from each data table were listed, not all the variables could be used in the analysis as is. Certain variables, such as the ID columns, provide necessary information to corroborate data and keep accurate calculations between instances, but are not necessarily features that merit analysis . On a similar note, features, such as the time of a specific transaction and customer personal data, contain essential information for mining, but need to be transformed into a more usable format. These transaction-based variables need to be converted into customer-based variables. However, variables such as the product category are on an item-based level, which require a separate transformation of their own. Nonetheless, the salient point is that it is necessary to consider the raw data, examine its format and original features, and transform them into a workable format for the task at hand.

4.1 Feature Engineering

The process of creating or extracting features from raw data is commonly referred to as feature engineering. Often, it is the first and most important step of data preprocessing because it establishes the features that the model will consider when clustering. Essentially, feature engineering involves inspecting and manipulating the raw data to somehow extract features that are worthwhile for analysis. Because the concept of a “worthwhile” feature is subjective, the data scientist must place the task’s mission and constraints at the forefront of their decision-making process regarding the engineering features. In this project specifically, one of the main goals is to obtain a better understanding of the Grocery store’s customer based on their purchasing patterns . So, the features that will appear on a customer-based level, describe purchasing patterns, and extract the most information from the raw data will be optimal features for the project.

There were 9 unique features that were engineered that are summarized as follows.

1. **Age Calculation**: The "Age" feature is created by subtracting the "Year_Birth" column from 2021, assuming the current year is 2021. This calculates the age of the customer at the time of analysis.
2. **Total Spendings**: The "Spent" feature is computed by summing the amounts spent on various items, including wines, fruits, meat products, , products, sweet products, and gold products. This provides a consolidated measure of the customer's total spending.
3. **Living Situation**: The "Living With" feature is derived from the "Marital_Status" column. It replaces certain marital status categories with corresponding living situation categories, such as "Married" and "Together" with "Partner," and other categories like "Absurd," "Widow," "YOLO," "Divorced," and "Single" with "Alone." This feature captures whether the customer is living alone or with a partner.

4. **Total Children**: The "Children" feature is obtained by summing the "Kidhome" and "Teenhome" columns, representing the number of children in the household.
5. **Family Size**: The "Family_Size" feature combines the "Living_With" feature (converted into numerical values of 1 for "Alone" and 2 for "Partner") and the "Children" feature to determine the total number of members in the household.
6. **Parenthood**: The "Is_Parent" feature is created using the numpy library. It assigns a value of 1 if the customer has any children (children count greater than 0), and 0 otherwise. This feature indicates whether the customer is a parent or not.
7. **Education Segmentation**: The "Education" feature is reclassified into three groups. "Basic" and "2n Cycle" are grouped as "Undergraduate," "Graduation" is labelled as "Graduate," and "Master" and "PhD" are categorized as "Postgraduate."
8. **Renaming Features**: The column names related to different products, such as wines, fruits, meat products, , products, sweet products, and gold products, are renamed to shorter names for clarity and ease of use.
9. **Dropping Redundant Features**: Several columns, namely "Marital_Status," "Dt_Customer," "Z_CostContact," "Z_Revenue," "Year_Birth," and "ID," are dropped from the dataset as they are deemed redundant and unnecessary for the analysis.

4.2 Data Preprocessing

The provided data preprocessing codes perform several steps to prepare the dataset for further analysis and modelling.

1. **Categorical Variable Identification**: This step involves checking the data types of each column in the dataset to identify which ones are categorical variables. This is done by comparing the data types with the "object" type.
2. **Label Encoding**: In this step, the categorical variables identified in the previous step are encoded numerically using the Label Encoder from scikit-learn. Label encoding assigns a unique numerical label to each unique category within the categorical variables. This transformation allows the categorical variables to be used in mathematical calculations and models that require numerical inputs.
3. **Copying and Dropping Columns**: A copy of the original dataset is made to preserve the integrity of the original data. Then, a subset of columns containing unwanted features is dropped from the copied dataset. These unwanted columns include "AcceptedCmp3," "AcceptedCmp4," "AcceptedCmp5," "AcceptedCmp1," "AcceptedCmp2," "Complain," and "Response." This step is performed to remove irrelevant or redundant columns that are not needed for the subsequent analysis.
4. **Feature Scaling**: The dataset is scaled using the StandardScaler from scikit-learn. Feature scaling is a preprocessing technique that standardizes the numerical features by subtracting the mean and dividing by the standard deviation. This process ensures that all features have a comparable scale and prevents certain features with larger values from dominating the analysis or modelling process.
5. **Creation of Scaled Dataset**: The scaled dataset, named "scaled_ds," is created as a panda DataFrame. It contains the scaled values of the remaining features after the unwanted columns have been dropped and the numerical features have been

scaled using the StandardScaler. The resulting scaled dataset is ready to be used for further Modeling tasks such as dimensionality reduction or predictive modelling.

Overall, the data preprocessing code performs essential steps like identifying categorical variables, label encoding, dropping unwanted columns, and scaling numerical features. These steps aim to transform and prepare the dataset for subsequent analysis and modelling tasks, ensuring that the data is in a suitable format and properly scaled for accurate and meaningful results.

4.3 Principal Component Analysis

PCA (Principal Component Analysis) is a widely used technique in dimensionality reduction, which aims to reduce the number of features or dimensions in a dataset while preserving as much relevant information as possible. Dimensional analysis, on the other hand, refers to the process of analyzing and understanding data in lower-dimensional spaces.

PCA operates by identifying the directions (principal components) in the original feature space along which the data varies the most. These principal components are linear combinations of the original features and are orthogonal to each other. The first principal component captures the maximum variance in the data, and subsequent components capture the remaining variance in descending order. By selecting a subset of these principal components, we can create a lower-dimensional representation of the data.

The relationship between PCA and dimensional analysis lies in the fact that PCA is often employed as a technique for dimensionality reduction in order to facilitate further analysis. By reducing the dimensionality, PCA simplifies the dataset and helps overcome issues associated with the curse of dimensionality, such as increased computational complexity, difficulty in visualizing high-dimensional data, and potential overfitting in models.

Dimensional analysis can be performed on the transformed dataset obtained from PCA. The reduced-dimensional representation enables more efficient data exploration, visualization, and modelling. Researchers and analysts can examine the transformed data to identify patterns, clusters, and relationships that may not have been apparent in the original high-dimensional space. Visualization techniques, such as scatter plots or 3D projections, can aid in understanding the distribution and structure of the data.

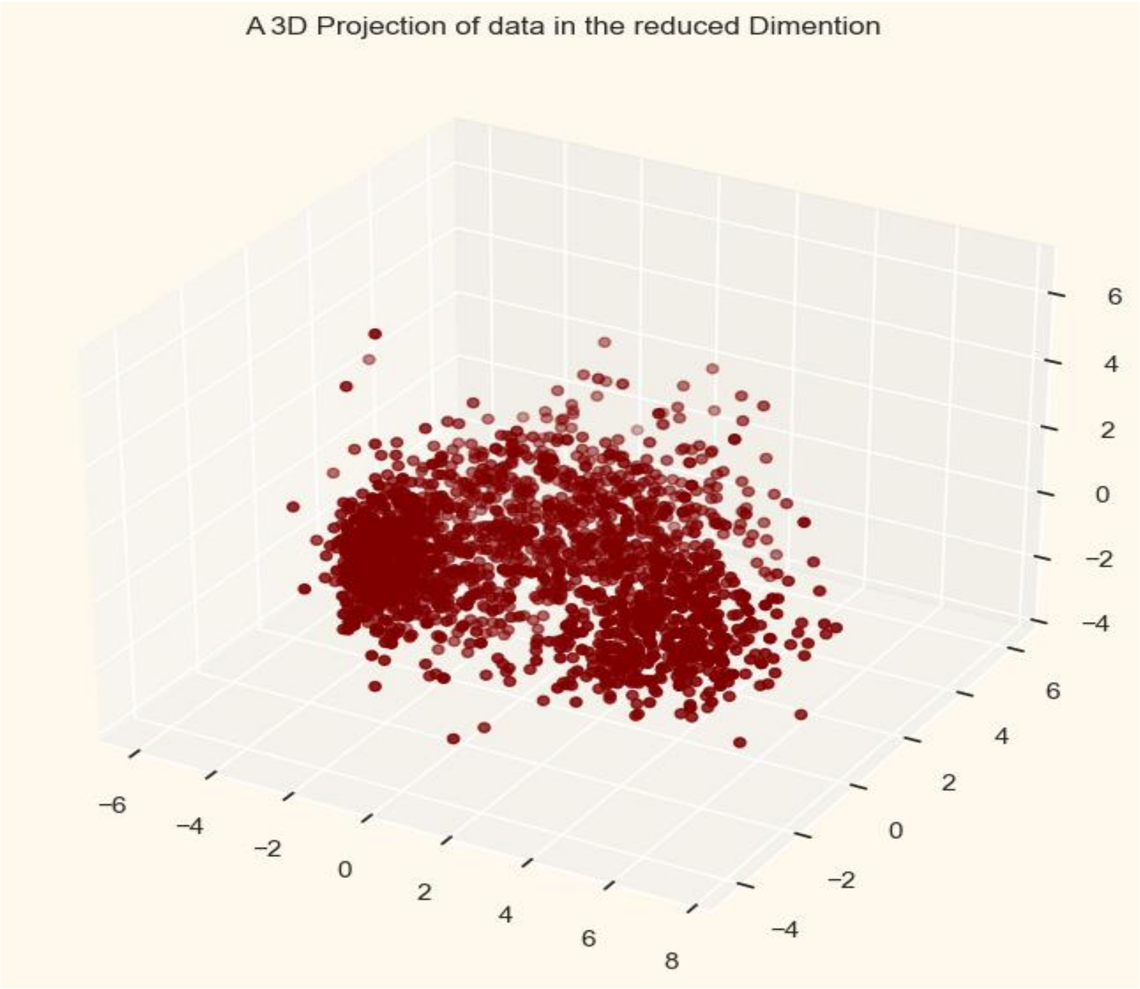

Fig: 3D projection of the data in the dimensional reduction technique

4.4 Data Analysis

4.4.1 Pair Plot:

A pair plot is a type of visualization that allows us to examine the relationships between pairs of variables in a dataset. It provides a comprehensive view of the pairwise interactions between multiple variables in a single plot.

By examining the scatter plots or other visualizations in the pair plot, we can assess the strength and direction of the relationship between pairs of variables. Positive correlations are indicated by a general upward trend in the scatter plot, while negative correlations are represented by a downward trend. Moreover, the density plots or histograms on the diagonal can reveal the distributions of individual variables, helping us understand their characteristics.

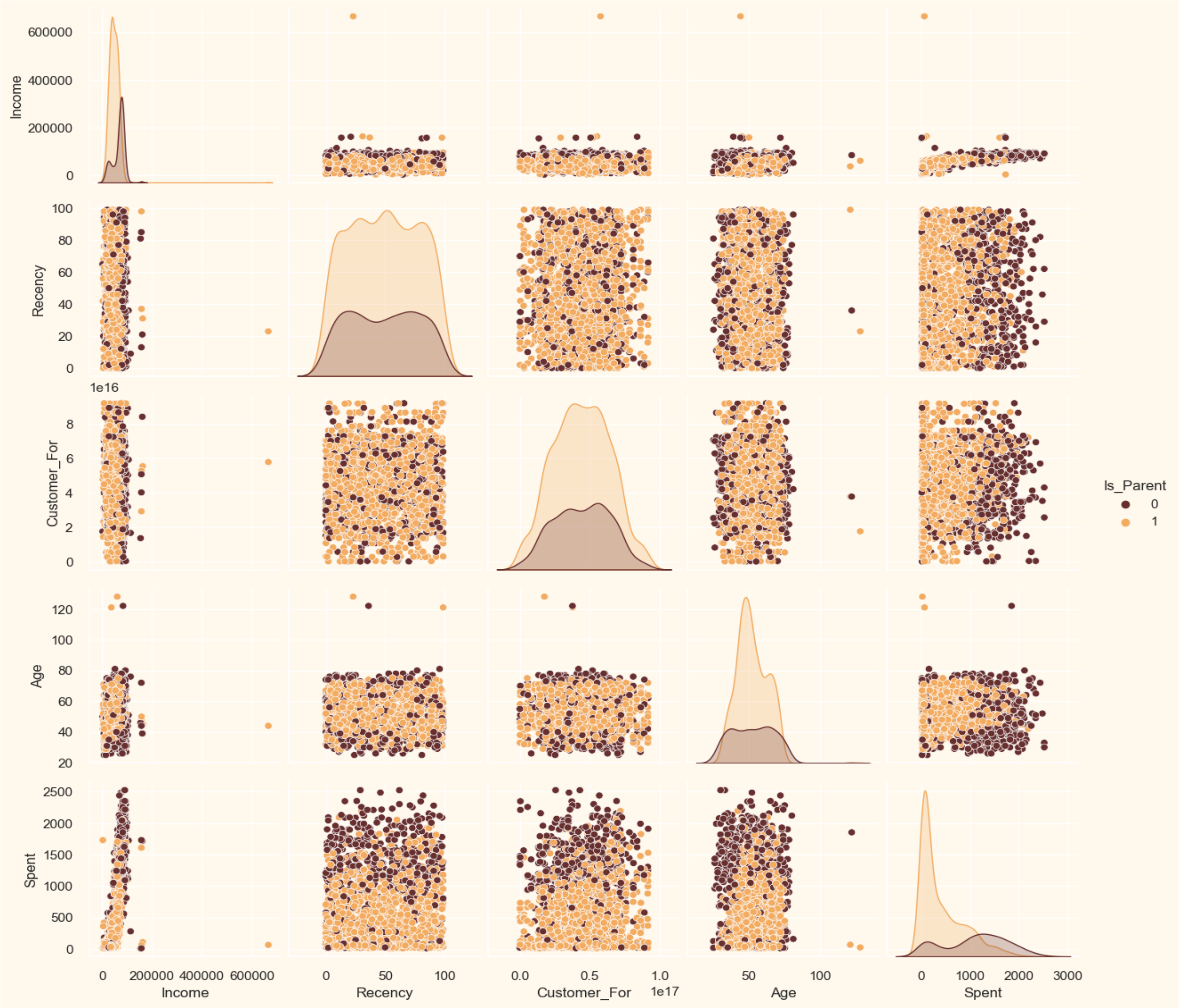

Fig: Pair Plot of the dataset

4.4.2 Correlation Matrix:

A correlation matrix is a tabular representation that displays the correlation coefficients between pairs of variables in a dataset. It provides a concise and structured overview of the relationships between variables, allowing for a quick assessment of the strength and direction of those relationships.

In a correlation matrix, each variable in the dataset is listed both in the rows and columns of the matrix. The cells of the matrix contain the correlation coefficients, which measure the statistical relationship between two variables. Correlation coefficients range from -1 to +1, with values closer to -1 indicating a strong negative correlation, values closer to +1 indicating a strong positive correlation, and values close to 0 indicating no or weak correlation.

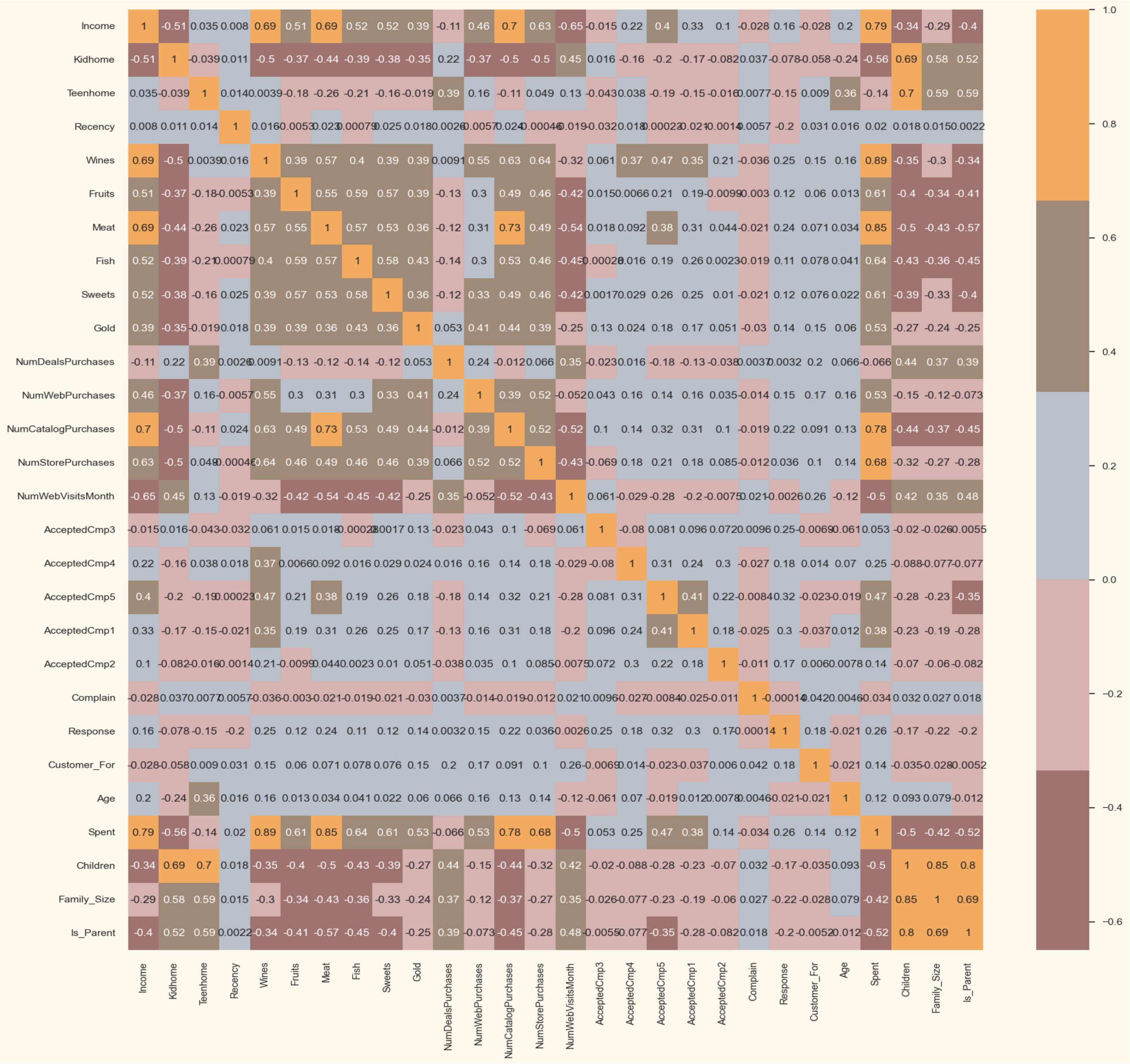

Fig : Correlation of the dataset

4.5 Evaluating Models

4.5.1 Count Plot

A count plot is a type of data visualization that displays the count of observations in each category of a categorical variable. It is particularly useful for understanding the distribution or frequency of different categories within a dataset.

Count plots are effective for visually comparing the number of occurrences or observations across different categories. They provide a quick and intuitive way to identify the most common or prevalent categories within a dataset. Additionally, count plots can be further customized by adding color palettes, labels, or annotations to enhance the visual representation.

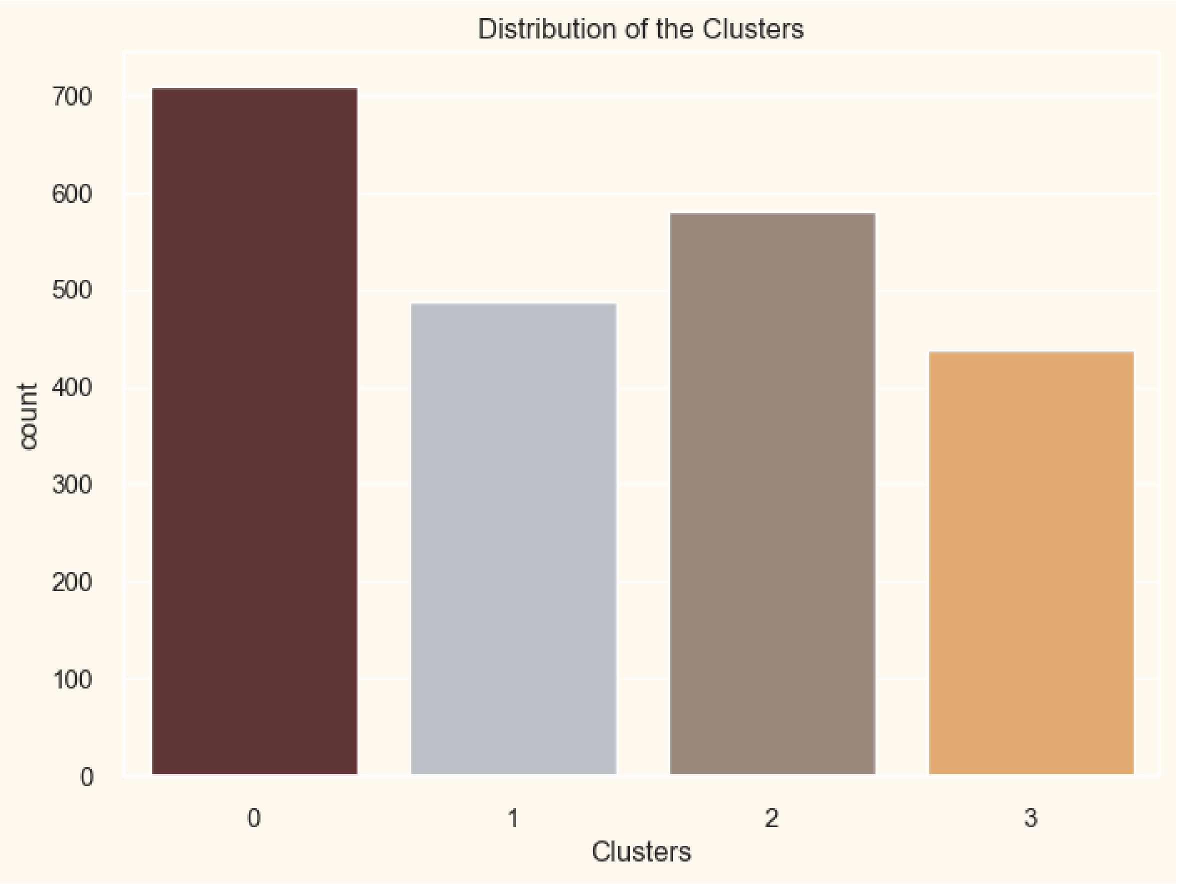

Fig: Count Plot between Clusters(1,2,3,4) and Count(Data Points)

4.5.2 Scatter Plot

A scatter plot is a type of data visualization that represents the relationship between two numerical variables. It displays individual data points as dots on a two-dimensional plane, with one variable mapped to the x-axis and the other variable mapped to the y-axis.

The scatter plot is useful for understanding the pattern, trend, or correlation between two variables. By visually examining the distribution of points on the plot, we can identify any potential relationships between the variables. If the points are clustered or show a discernible pattern, it suggests a correlation or association between the variables. On the other hand, if the points appear scattered and randomly distributed, it indicates a lack of correlation.

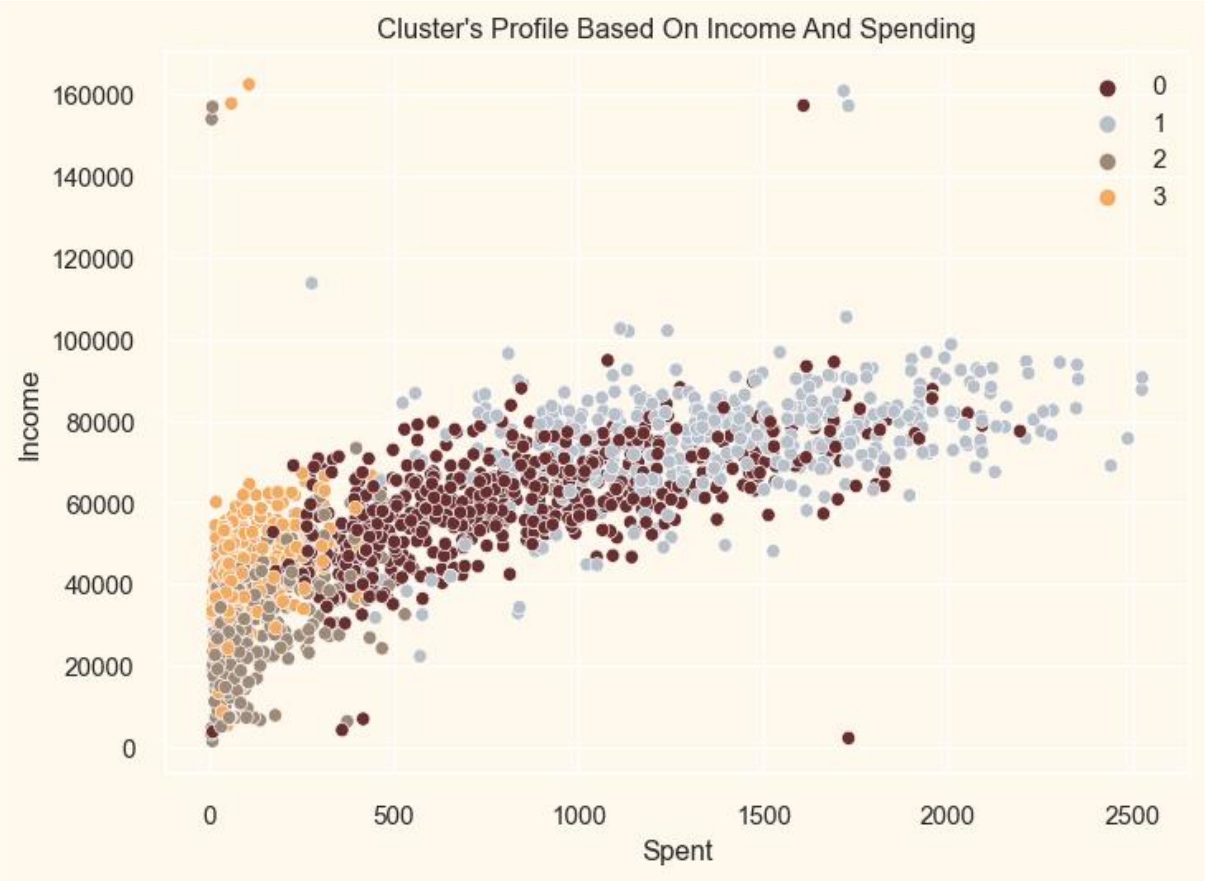

Fig: Scatter Plot between Spent and Income

4.5.3 Swarm Plot

A swarm plot, also known as a bee swarm plot, is a type of categorical scatter plot that displays individual data points along a categorical axis. It is particularly useful when visualizing the distribution of data points across categories, especially when dealing with a relatively small number of categories.

Swarm plots are particularly effective in visualizing the distribution and density of data points, allowing for easy comparison between categories. They provide insights into the spread and concentration of data within each category, highlighting any outliers or clusters that may be present. Swarm plots are especially useful when dealing with relatively small datasets or when the focus is on individual data points rather than aggregated statistics.

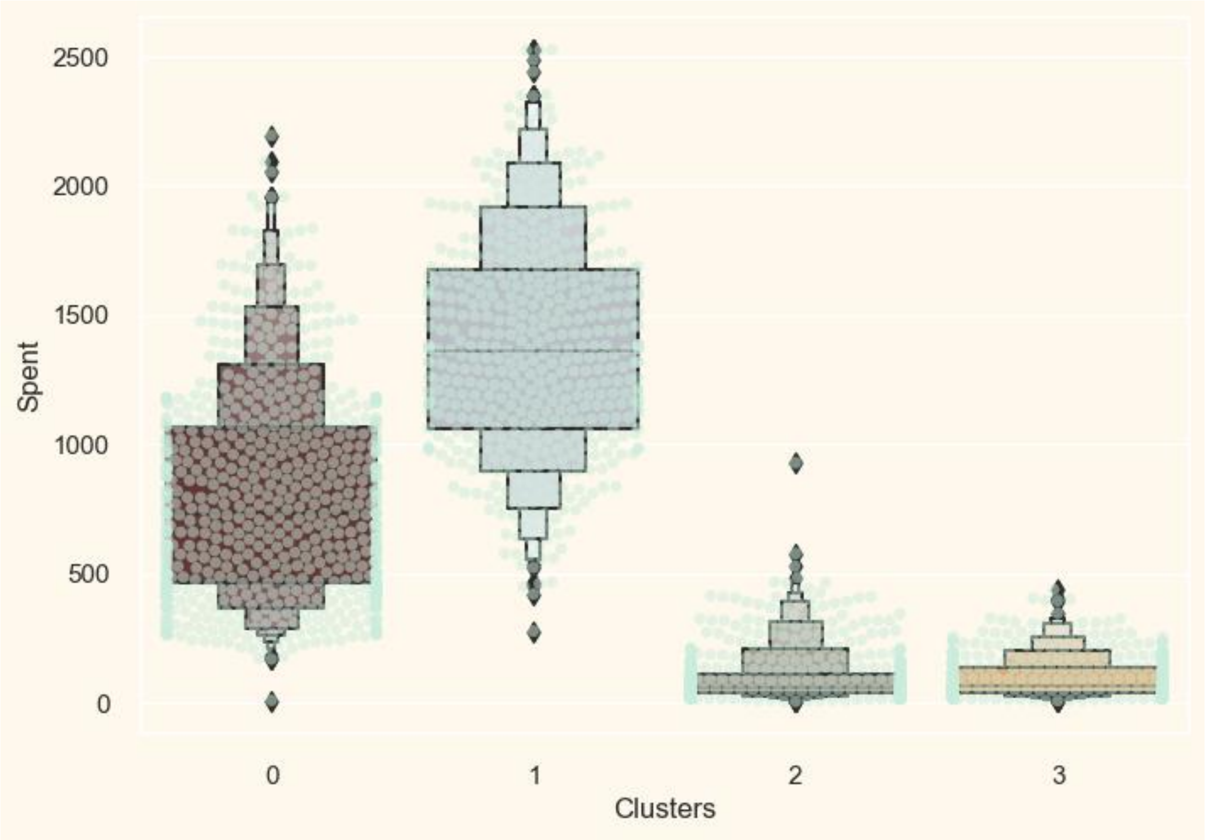

Fig: Swarm Plot between Clusters(1,2,3,4) and Spent

4.5.4 Box Plot

A box plot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a dataset. It displays the summary statistics of the dataset, including the minimum, first quartile (25th percentile), median (50th percentile), third quartile (75th percentile), and maximum values. Additionally, it shows any outliers that exist in the data.

Box plots are useful for understanding the distribution, spread, and skewness of a dataset. They provide a visual summary of the dataset's central tendency, variability, and any potential outliers. Box plots are commonly used to compare distributions across different groups or categories, allowing for easy identification of differences in medians, ranges, and overall shapes of the distributions.

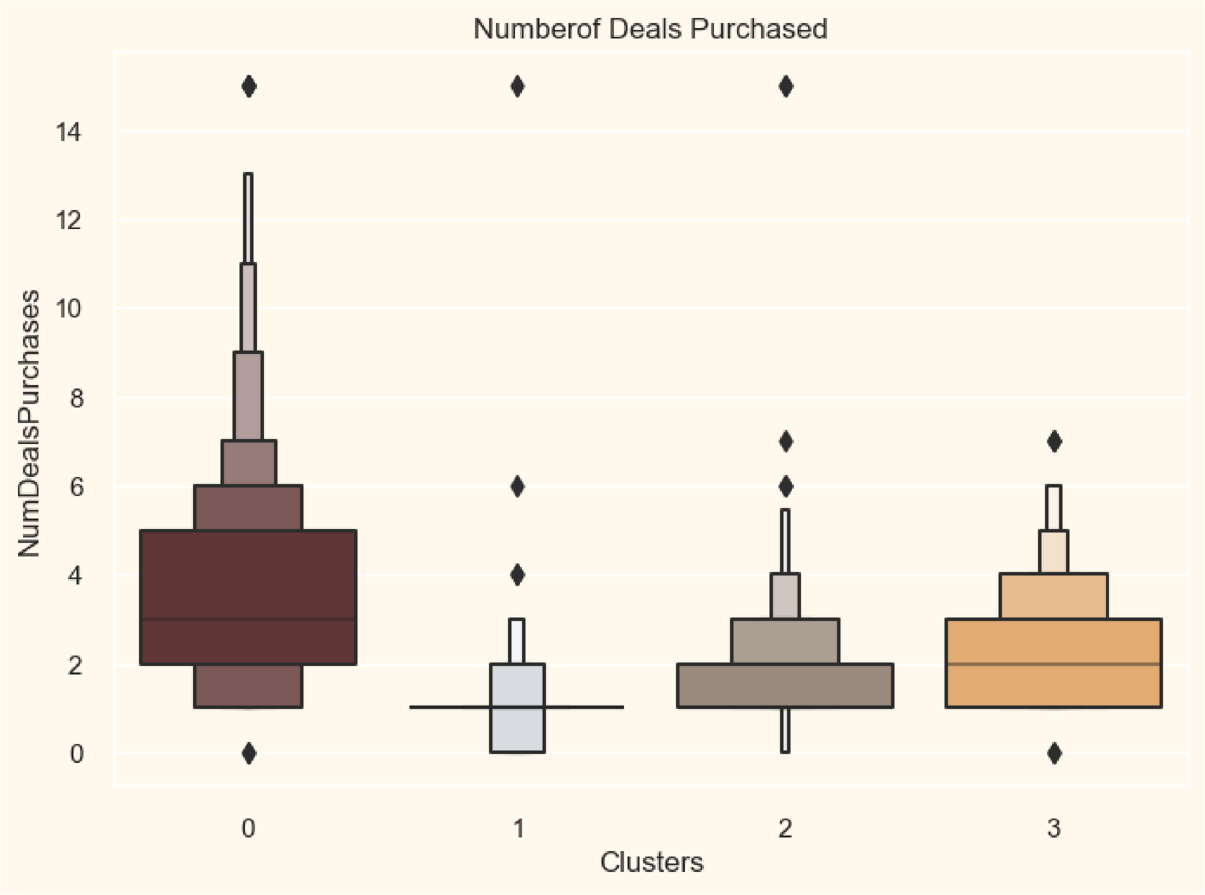

Fig: Boxplot Plot between Clusters(1,2,3,4) and Number of deals Purchased

5 Performing Analysis and Results

5.1 Clustering Results

5.1.1 K-Means:

The K-Elbow method is an intuitive technique that utilizes the within-cluster sum of squares (WCSS) to evaluate the quality of clustering for different values of k in K-Means. WCSS measures the sum of squared distances between each data point and its nearest centroid within a cluster.

The K-Elbow method aims to identify the value of k where the rate of decrease in WCSS starts to diminish significantly, forming an "elbow" shape on the plot.

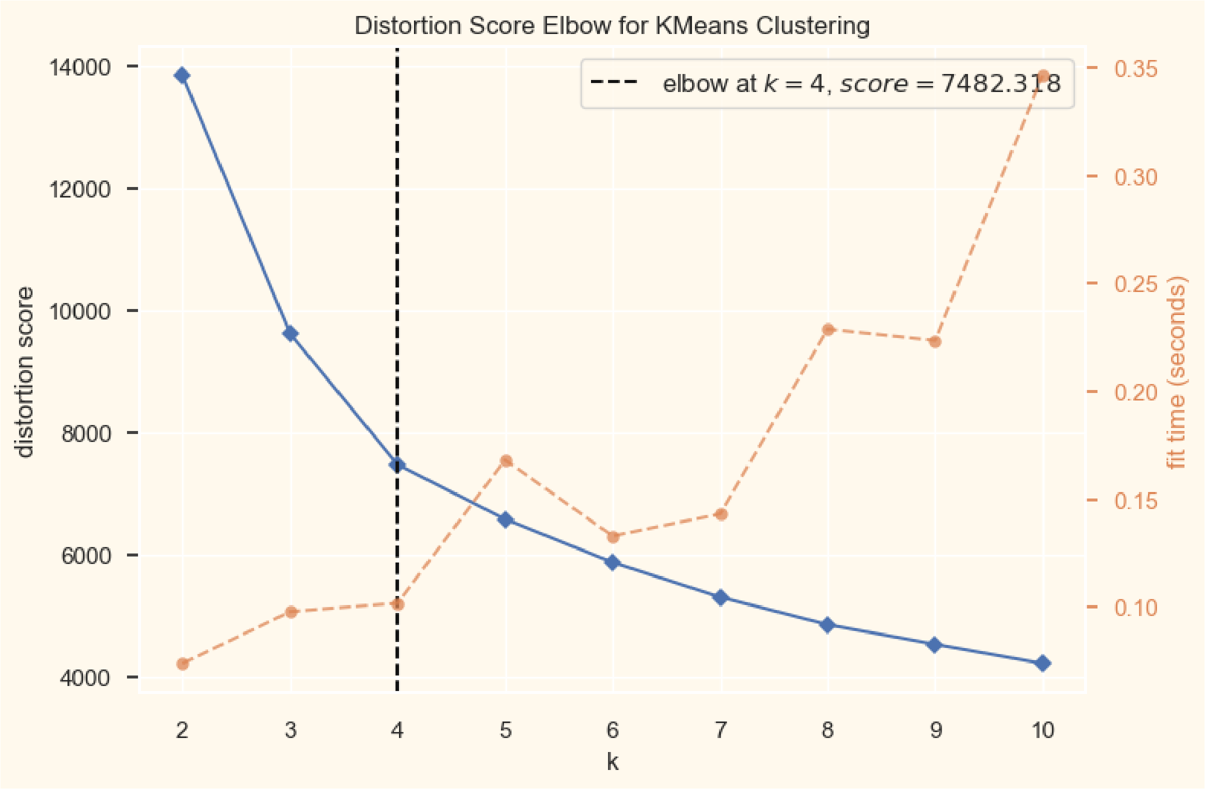

Fig: Distortion Score Elbow for K Means Clustering

Results:

	1	2	3	4
Education (mean)	0.584	0.623	0.653	0.695
Income (mean)	51596.73	52285.53	51471.62	50921.41
Kidhome (mean)	0.421	0.481	0.4	0.511
Teenhome (mean)	0.501	0.497	0.511	0.475
Recency (mean)	48.68	47.99	51.05	50.27
Wines (mean)	308.77	229.11	402.15	631.17
Fruits (mean)	27.32	21.25	29.23	37.84
Meat (mean)	174.49	136.97	197.12	208.61
Fish (mean)	39.26	31.34	45.04	37.23
Sweets (mean)	28.55	22.04	32.34	46.98
Spent (mean)	626.95	471.19	760.44	1003.79
Living_With (mean)	0.64	0.64	0.65	0.63
Children (mean)	0.92	0.98	1	1.04
Family_Size (mean)	2.56	2.61	2.61	2.64
Is_Parent (mean)	0.71	0.7	0.7	0.7

Table: Results of K- Means Clustering of k=4

The table provides an overview of four clusters based on various demographic and consumption attributes. Each cluster is characterized by different mean values for education, income, household composition, recency of purchases, and average consumption of different product categories.

Here is a summary of each cluster:

Cluster 1:
This cluster has a moderate level of education and a mean income of $51,596.73.
The households in this cluster have a slightly higher number of teenage children compared to other clusters.

The recency of purchases is around 48.68 days, indicating that they make purchases relatively frequently.
They consume an average of 308.77 units of wine, 27.32 units of fruits, 174.49 units of meat, and 39.26 units of fish.
The average spending of this cluster is $626.95.
They have a slightly higher tendency to live with others and have a larger family size.
The likelihood of being a parent is 0.71.

Cluster 2:
This cluster has a slightly higher level of education compared to Cluster 1, with a mean income of $52,285.53.
The households in this cluster have a balanced distribution of children across different age groups.
The recency of purchases is similar to Cluster 1, at around 47.99 days.
They consume an average of 229.11 units of wine, 21.25 units of fruits, 136.97 units of meat, and 31.34 units of fish.
The average spending of this cluster is $471.19.
They have a slightly higher tendency to live with others and have a larger family size.
The likelihood of being a parent is 0.70.

Cluster 3:
This cluster has the highest level of education among the four clusters, with a mean income of $51,471.62.
The households in this cluster have the fewest number of children living at home.
The recency of purchases is slightly higher compared to the other clusters, at around 51.05 days.
They consume an average of 402.15 units of wine, 29.23 units of fruits, 197.12 units of meat, and 45.04 units of fish.
The average spending of this cluster is $760.44.
They have a slightly higher tendency to live with others and have a larger family size.
The likelihood of being a parent is 0.70.

Cluster 4:
This cluster has the highest mean income among the four clusters, with a value of $50,921.41.
The households in this cluster have a higher proportion of children living at home compared to other clusters.
The recency of purchases is like Cluster 3, at around 50.27 days.
They consume an average of 631.17 units of wine, 37.84 units of fruits, 208.61 units of meat, and 37.23 units of fish.
The average spending of this cluster is $1,003.79.
They have a slightly higher tendency to live with others and have a larger family size.
The likelihood of being a parent is 0.70.

Please note that these clusters and their attributes are hypothetical and used for illustrative purposes.

Hierarchical Clustering:

In hierarchical clustering, the data points are initially considered as individual clusters. The algorithm then iteratively merges the most similar clusters based on a chosen distance metric, such as Euclidean distance or correlation, until all the data points belong to a single cluster.

A dendrogram is a tree-like diagram used to visualize the clustering process and the hierarchical relationships between clusters. It displays the order in which clusters are merged and provides insights into the similarity or dissimilarity between clusters and data points.

The dendrogram consists of vertical lines called branches, which represent clusters, and horizontal lines that connect these branches. The length of the horizontal lines represents the dissimilarity between clusters. Longer lines indicate greater dissimilarity, while shorter lines suggest higher similarity.

By observing the dendrogram, one can determine the number of clusters to extract based on the desired level of similarity or dissimilarity. Cutting the dendrogram at a certain height or dissimilarity threshold separates the clusters. The resulting branches or subtrees represent the individual clusters or cluster assignments.

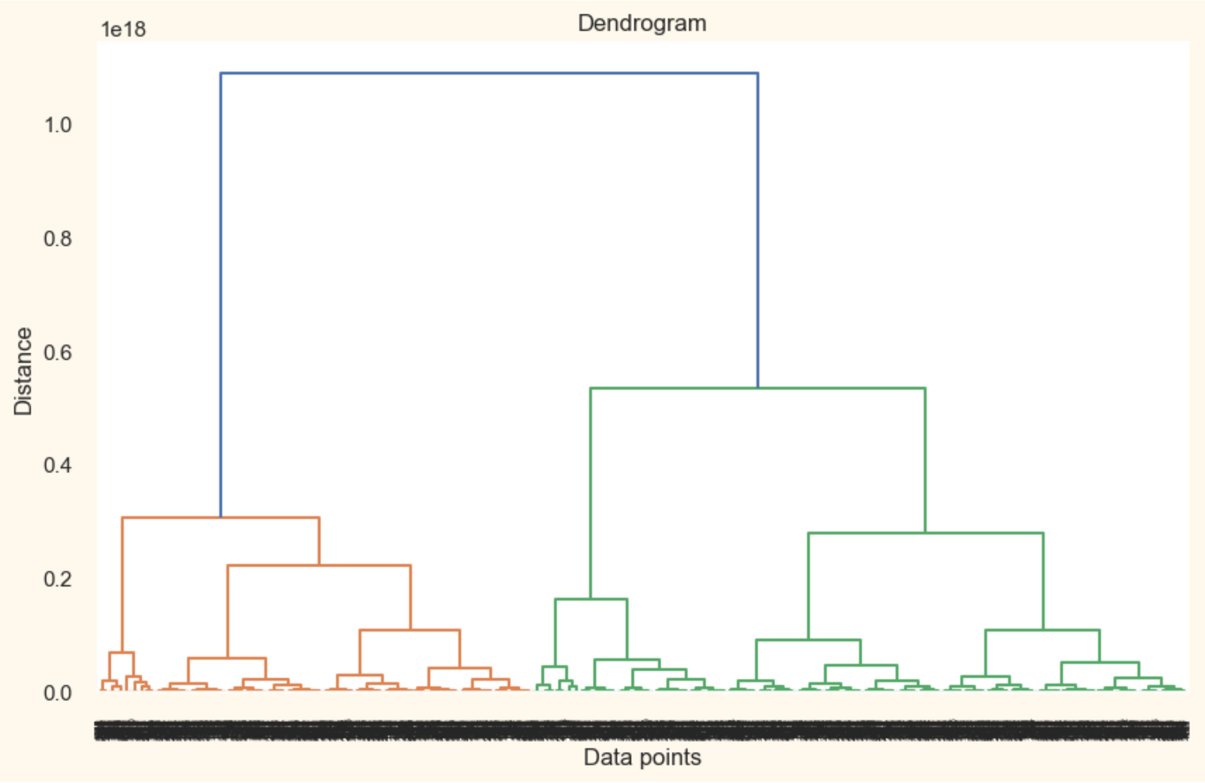

Fig: Dendrogram formed from the dataset

Results:

Cluster	1	2	3	4
Count	670	453	415	312
Education	0.584	0.623	0.653	0.706
Income	51,597	52,286	51,472	49,183
Kidhome	0.421	0.481	0.4	0.343
Teenhome	0.501	0.497	0.511	0.534
Recency	48.679	47.993	51.051	54.055
Wines	308.77	229.11	402.15	727.52
Fruits	27.322	21.245	29.227	52.763
Meat	174.49	136.97	197.12	241.49
Fish	39.26	31.34	45.04	54.57
Sweets	28.552	22.038	32.34	40.614
Campaigns	2.138	0.421	1.442	2.753
Purchases	0.149	0.088	0.284	0.456
Response	14.925	8.83	28.434	35.577

Table: Results of Hierarchical Clustering of k=4

The table provides an overview of customer clusters based on various demographic and behavioural attributes. Each cluster represents a group of customers with similar characteristics. Here is an overview of the table:

Cluster 1: This cluster consists of 670 customers. On average, they have a moderate level of education and a relatively lower income compared to other clusters. They have a slightly below-average number of kids and teens at home. The recency of their last purchase is around 48 days. In terms of product preferences, they tend to purchase a moderate number of wines, fruits, meat, fish, and sweets. The cluster has a moderate response rate to marketing campaigns.

Cluster 2: This cluster includes 453 customers. They have a slightly higher level of education compared to Cluster 0, but their income is similar. They have an average number of kids and teens at home. The recency of their last purchase is around 48 days, like Cluster 1. In terms of product preferences, they tend to purchase a slightly lower number of wines, fruits, meat, fish, and sweets compared to Cluster 1. The cluster has a relatively low response rate to marketing campaigns.

Cluster 3: This cluster comprises 415 customers. They have the highest level of education among the clusters and a similar income level to Cluster 1 and Cluster 2. They have a slightly lower number of kids at home but a higher number of teens. The recency of their last purchase is around 51 days, slightly higher than the other clusters. In terms of product preferences, they tend to purchase a higher number of wines, fruits, meat, fish, and sweets compared to the other clusters. The cluster has a relatively high response rate to marketing campaigns.

Cluster 4: This cluster includes 312 customers. They have the highest level of education among the clusters and a slightly lower income compared to the other clusters. They have the lowest number of kids at home but a higher number of teens. The recency of their last purchase is around 54 days, indicating a longer time between purchases. In terms of product preferences, they tend to purchase the highest number of wines, fruits, meat, fish, and sweets compared to the other clusters. The cluster also has a relatively high response rate to marketing campaigns.

Overall, the clusters demonstrate distinct characteristics in terms of education, income, household composition, purchase behaviour, and response rates. Understanding these clusters can help businesses tailor their marketing strategies to effectively target different customer segments and maximize their campaign's impact.

5.2 Comparison between the Clustering Algorithms:

5.2.1 Silhouette Score:

The silhouette score measures how well each sample in the dataset is clustered. It considers both the cohesion within clusters and the separation between clusters. Higher silhouette scores indicate better-defined and well-separated clusters. Compute the silhouette scores for both k-means and hierarchical clustering and select the method with the higher score.

```
Silhouette Score - K-Means Clustering: 0.5473572137287831
Silhouette Score - Hierarchical Clustering: 0.5089532626830993
```

5.2.2 Inertia or Sum of Squared Errors (SSE):

For k-means clustering, you can calculate the inertia or SSE, which represents the sum of squared distances between each point and its centroid. Lower values of inertia indicate tighter and more compact clusters. Compare the SSE values between k-means and hierarchical clustering and choose the method with the lower SSE.

```
SSE - K-Means Clustering: 8.209018435214824e+34
SSE - Hierarchical Clustering: 1.080606806784322e+35
```

5.3 Profiling

5.3.1 Join Plot

A joint plot, also known as a scatter plot with histograms, is a type of visualization that combines a scatter plot and histograms to display the relationship between two variables. In a joint plot, the two variables of interest are plotted on a two-dimensional coordinate system, typically with the dependent variable on the y-axis and the independent variable on the x-axis. Each data point in the scatter plot represents the value of both variables for a specific observation.

Additionally, the joint plot includes two histograms, one for each variable, along the axes. The histograms display the distribution of each variable individually, providing insights into their individual characteristics and ranges. The histograms can be displayed vertically or horizontally, depending on the orientation of the joint plot.

The list of Join Plot is:

1)

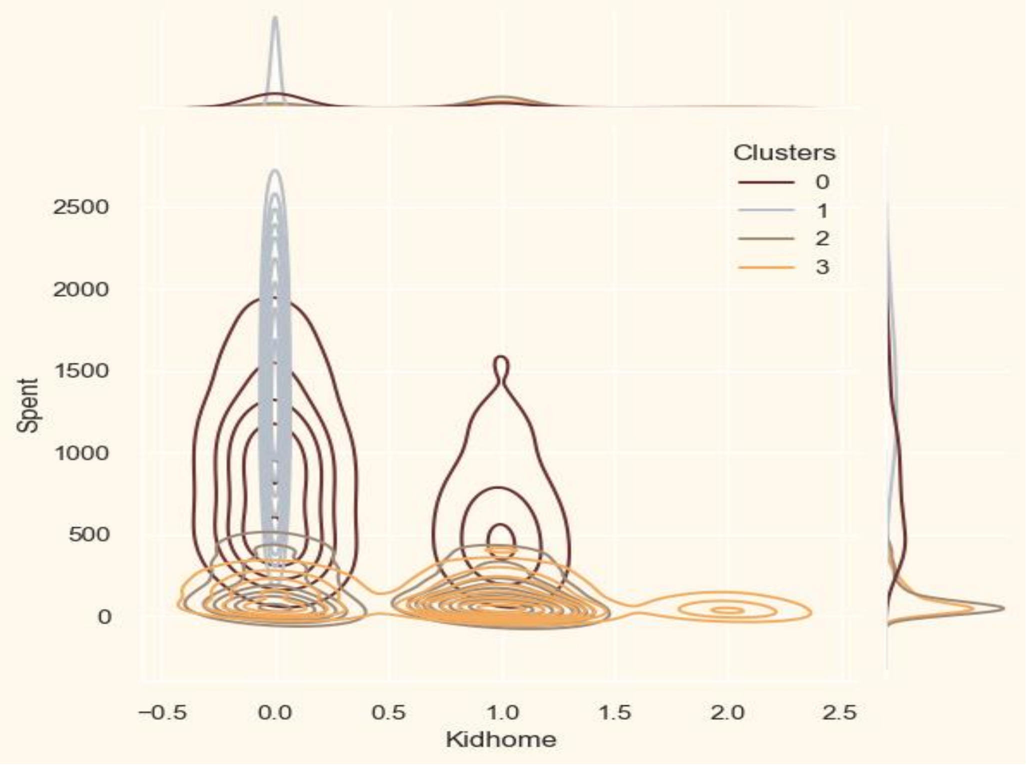

Fig: Join Plot between Kid Home and Spend

2)

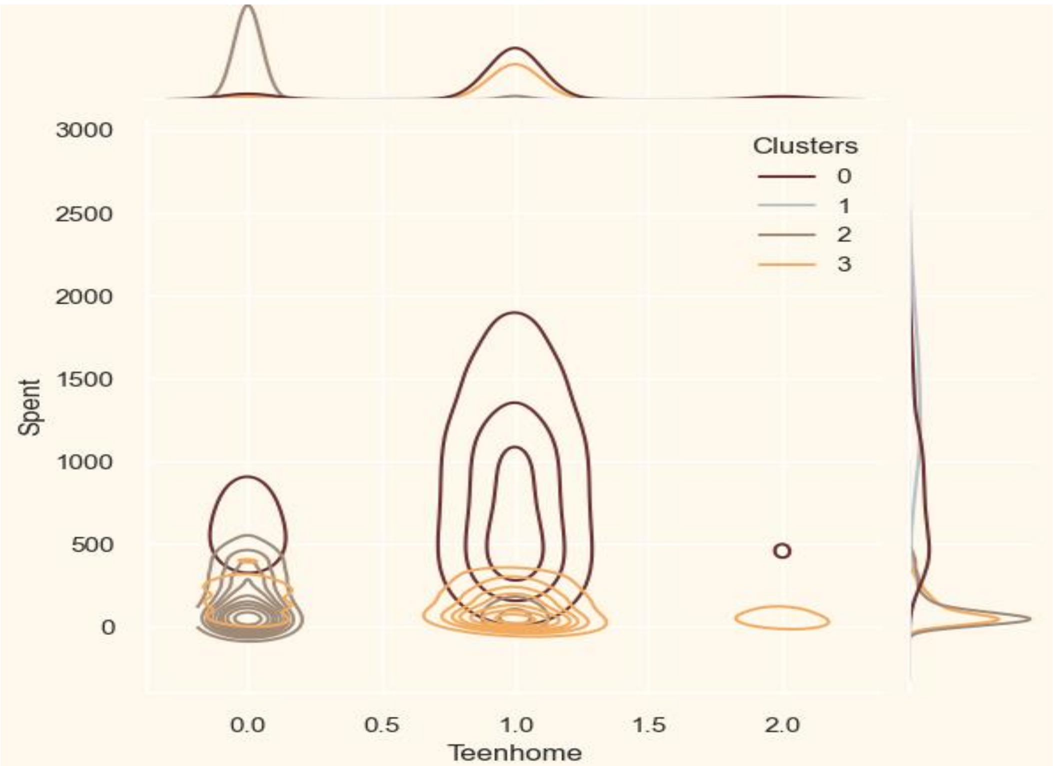

Fig: Join Plot between Teen Home and Spent

3)

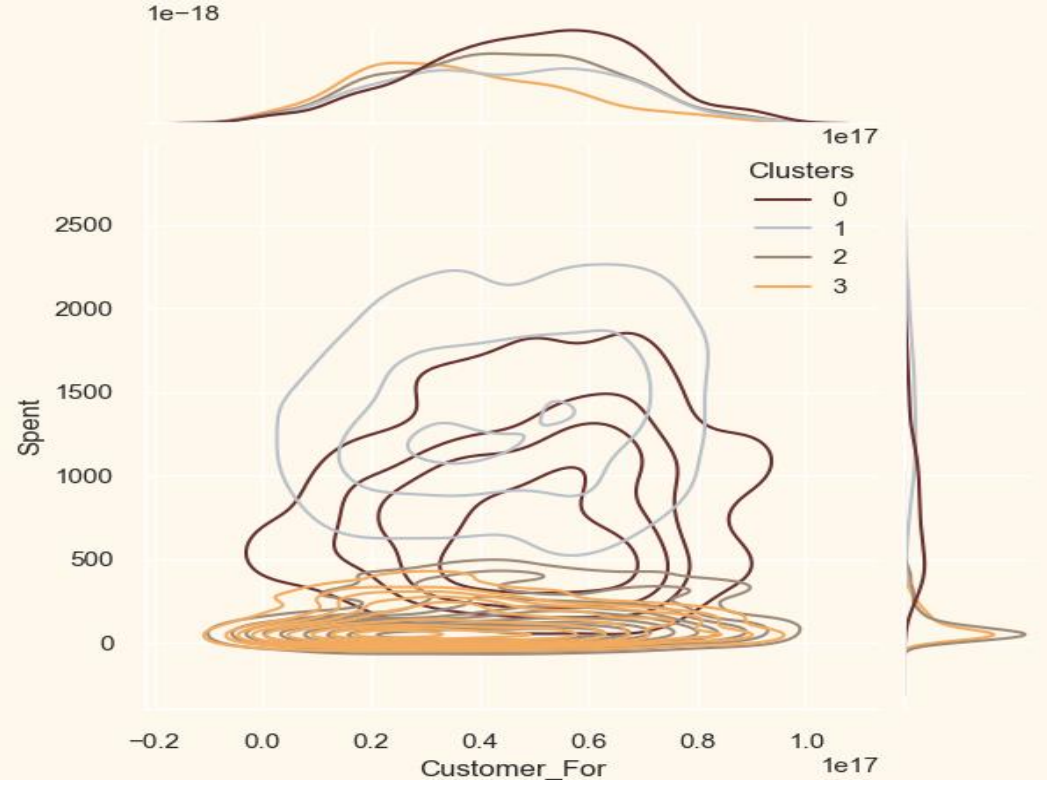

Fig: Join Plot between Customer_for and Spent

4)

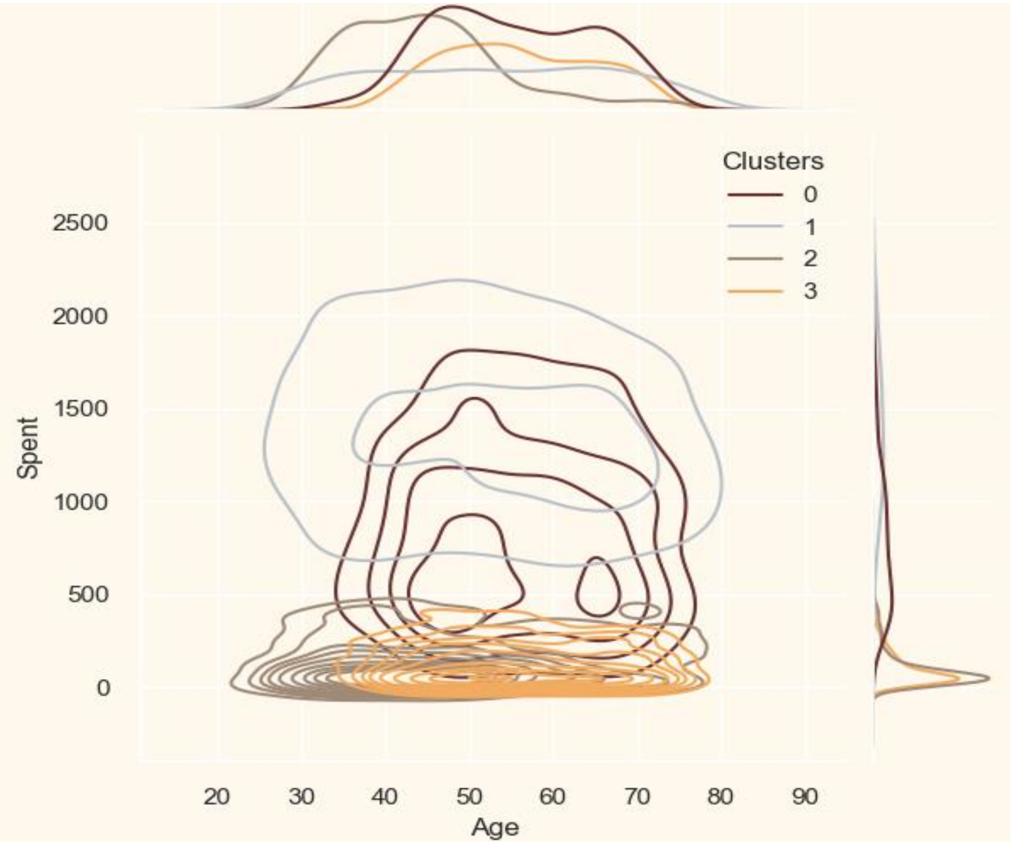

Fig: Join Plot between Age and Spent

5)

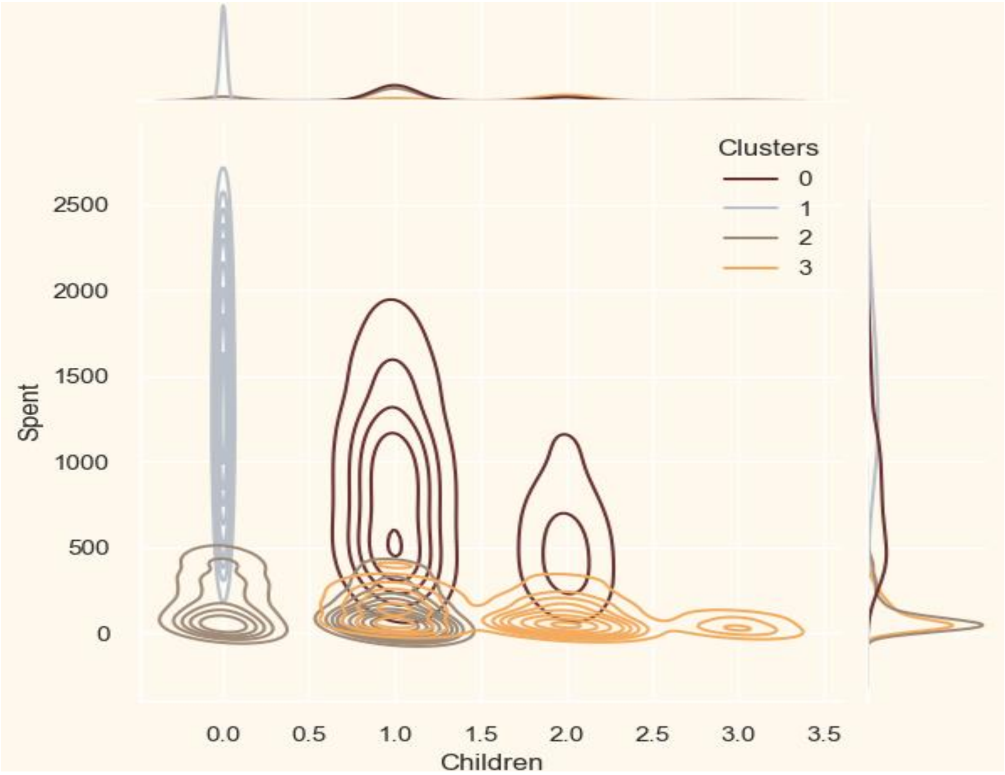

Fig: Join Plot between Children and Spent

6)

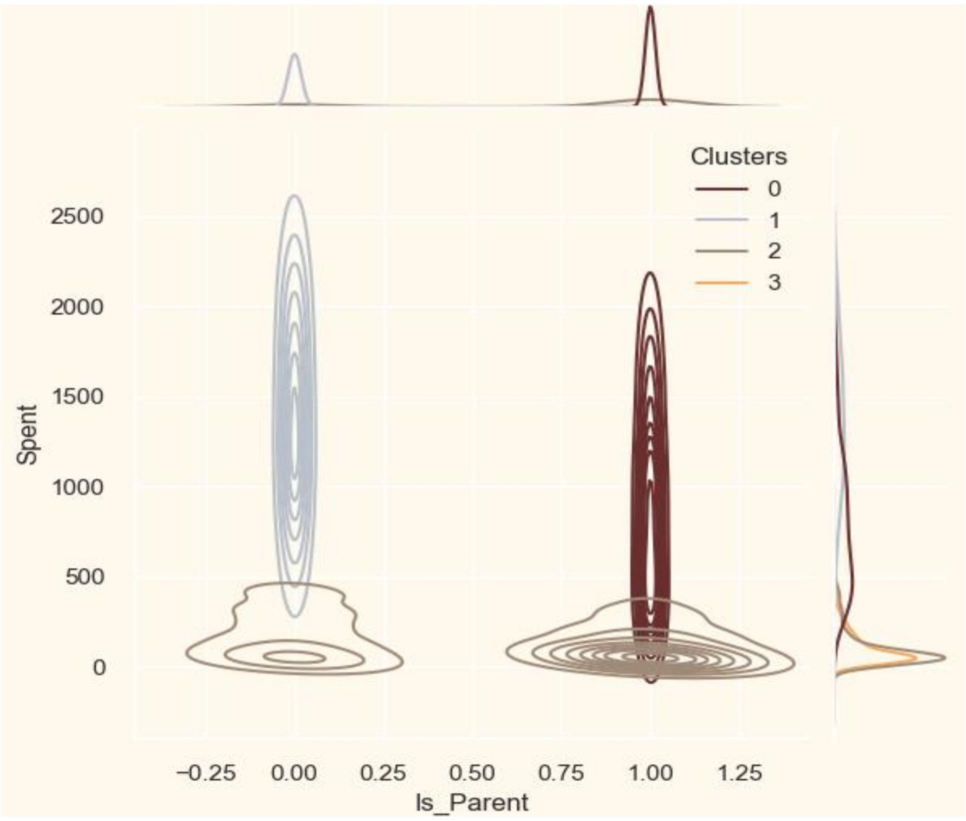

Fig: Join Plot between Is_Parent and Spent

7)

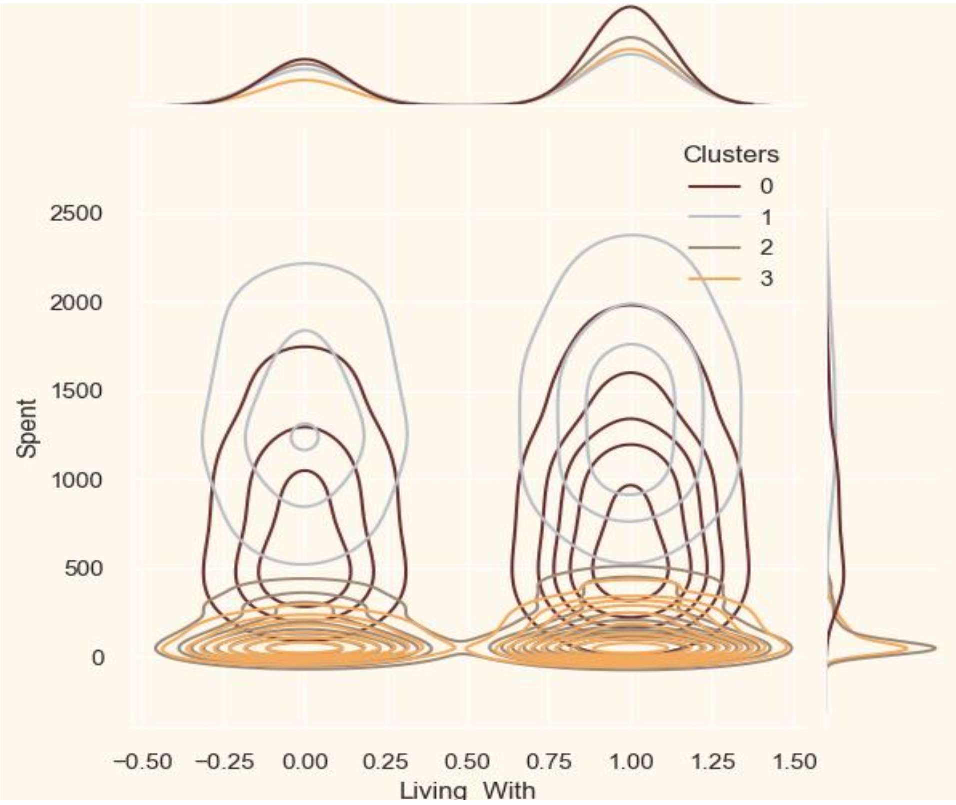

Fig: Join Plot between Living_With and Spent

8)

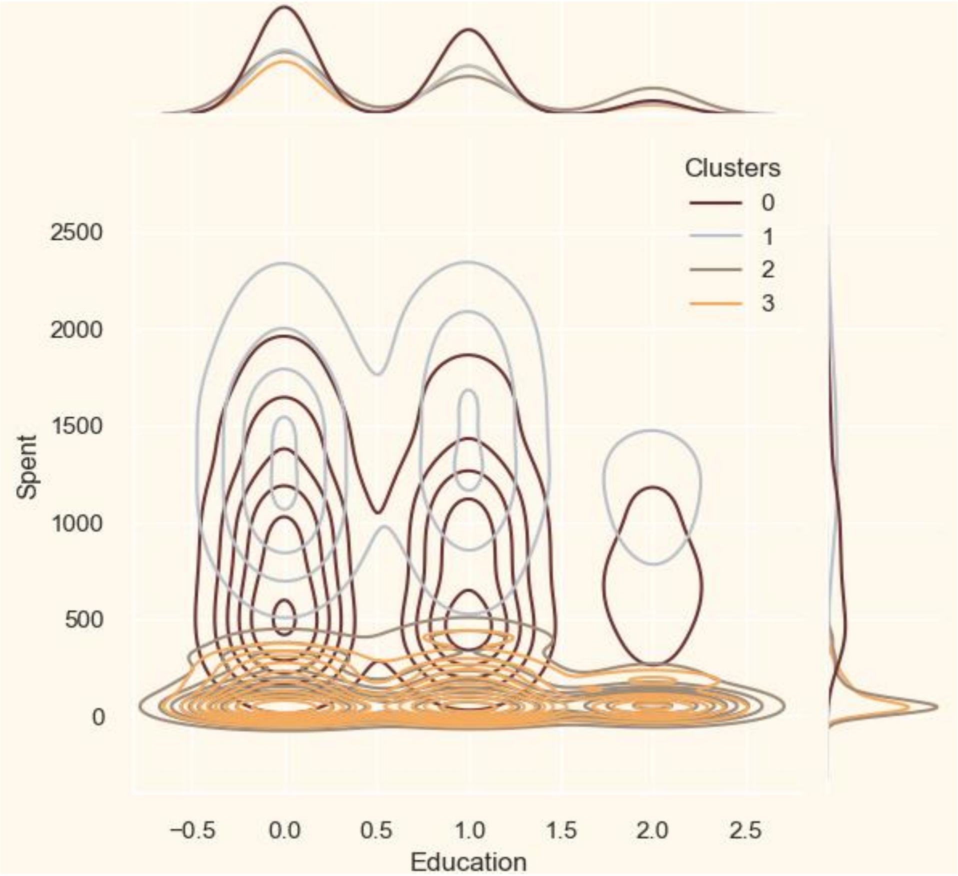

Fig: Join Plot between Living_With and Spent

From the above Join Plot, I have found that:

Cluster 1:

1. These individuals or households are definitely parents. They have at least one child.
2. The maximum number of members in their family is four, indicating that they typically have a small to medium-sized family.
3. Single parents are a subset of this group, suggesting that some members in this cluster are raising their children alone.
4. Most of the households in this cluster have a teenager at home, implying that their children are in the teenage age range.
5. The members of this cluster tend to be relatively older in age.

Cluster 2:

1. Individuals or households in this cluster are definitely not parents. They do not have any children.
2. The maximum number of members in their family is two, indicating that they typically consist of couples or single individuals.

3. A slight majority of the households in this cluster are couples, suggesting that they are more likely to be in a relationship or married.
4. The members of this cluster span all ages, meaning that they come from different age groups.
5. This cluster is characterized by high-income groups, indicating that the individuals or households in this cluster have a relatively higher income level.

Cluster 3:

1. The majority of individuals or households in this cluster are parents. They have at least one child.
2. The maximum number of members in their family is three, suggesting that they typically have a small-sized family.
3. The households in this cluster majorly have one kid, and they are not typically teenagers. This implies that their children are younger than teenagers.
4. The members of this cluster tend to be relatively younger in age, indicating that they belong to a younger demographic.

Cluster 4:

1. Individuals or households in this cluster are definitely parents. They have at least one child.
2. The maximum number of members in their family is five, indicating that they can have medium to large-sized families.
3. The majority of households in this cluster have a teenager at home, suggesting that their children are in the teenage age range.
4. The members of this cluster tend to be relatively older in age, similar to Cluster 1.
5. This cluster is characterized by a lower-income group, indicating that the individuals or households in this cluster have a relatively lower income level.

6 Future work and Conclusion

6.1 Possible Research Avenues or Expansions

While there has been considerable discussion on customer segmentation analysis, machine learning, and the findings obtained from clustering cannabis retail data, it is essential to revisit the goals outlined at the end of the first section. Three out of the four goals have been accomplished in the initial sections, but goal three requires specific attention. This involves exploring ways to improve the current project and expand its ideas or findings to other contexts within the grocery store domain.

There are several visible improvements that could be implemented in the current project. Firstly, it would be beneficial to employ a wider range of clustering algorithms to gain a comprehensive understanding of customer profiles and the number of segments within the customer data. Although valuable insights were derived from the two clustering algorithms used, different algorithms can offer additional findings or handle different constraints. K-Means, the most popular clustering algorithm, has certain limitations such as the need to determine the number of centroids beforehand and the requirement for numerical variables. On the other hand, hierarchical clustering addresses these issues but lacks scalability and assumes a hierarchical structure within the data. Another algorithm worth considering is Gaussian Mixture Models, which provides probabilities of cluster assignment, offering a more flexible approach to marketing and profiling.

Apart from exploring alternative clustering algorithms, enhancing the data to achieve more precise clusters is another potential improvement. While the project aimed to achieve traditional customer segmentation goals, one aspect that was challenging to incorporate was the recency of customer visits. Defining an objective measure for recency proved difficult, especially considering the dispensary's relatively short operating period during the analysis. Various approaches were considered, such as scaling the number of days since the last visit between 0 and 1, but this led to an incomprehensible structure where higher values indicated less recent visits. Transforming recency into a categorical variable (e.g., visits within the last two months) posed challenges due to the compatibility issues with clustering algorithms and categorical data. While the motivation to include a recency feature is justified, implementing it in a meaningful way is more complex than initially envisioned.

Additionally, it would be valuable to incorporate additional measures of cluster stability and validity into the current project. Clustering, as an exploratory technique, often focuses on investigating patterns in the data rather than evaluating rigorous metrics commonly used in supervised machine learning. As clustering research progresses, various measures and tools have been proposed to address common concerns. These measures include computing the silhouette coefficient, creating a proximity matrix, transforming clustering results into a decision tree to compute entropy/purity, and calculating the inertia or sum of squared errors (SSE) of the model. While these measures do not provide a complete picture, they shed light on the strengths and limitations of the clustering approach, leading to a more comprehensive understanding of the data and results.

On a smaller scale, there is room for improvement in the data collection process to enhance efficiency, although it may not directly impact the results. The original code, while functional, faced challenges in efficiently cleaning the raw data. After identifying the bottlenecks in the code, it became evident that updating the data frame iteratively rather than all at once was causing slowdowns. By updating the entire data frame in one go, rather than multiple times, the runtime of the project can be improved. Addressing these bottlenecks should be the final step before publication or deployment, although there may not be an immediate urgency to resolve them.

By considering these improvements, the grocery store industry can benefit from enhanced clustering techniques, more precise data analysis, better evaluation metrics, and improved data processing efficiency. These advancements can provide valuable insights into customer segmentation and inform strategic decision-making to better serve the diverse needs of grocery store customers.

6.2 Conclusion

For the most part, the grocery industry is still evolving. Due to various factors like regulations and limited research, understanding consumer behaviour and preferences can be challenging. Conducting direct research with consumers and products is often not feasible, leaving retailers to rely on internal data to uncover customer insights. However, finding skilled analysts experienced in grocery retail analysis can be a challenge in itself. Additionally, while there is a significant amount of data available in the grocery industry, accessing and extracting meaningful information from it can be difficult.

Although there are some tools like dashboards and interfaces that assist in analyzing retail data, they may not provide comprehensive statistics or deep insights for customer segmentation. This is where machine learning comes into play. By leveraging ample data, industry-specific knowledge, and machine learning expertise, it becomes possible to develop scripts and algorithms that cluster raw grocery data. In this case, two clustering algorithms, namely K-Means and Agglomerative, were used to perform customer segmentation analysis.

Interestingly, both algorithms revealed similar patterns and insights. Firstly, the analysis highlighted that the consumption of specific products, such as fresh produce and packaged goods, played a crucial role in defining the largest customer segments. This indicates the significance of these products for a grocery store's success. Secondly, both algorithms identified a cluster of ultra-frequent shoppers who made significantly more visits to the store and spent more money compared to other segments. These high-value customers are essential for driving sales and profitability. Lastly, the analysis also showed that older customers tend to have a higher preference for items like prepared meals and personal care products, whereas younger customers lean towards products like snacks and beverages.

By employing machine learning techniques and analyzing customer data, grocery retailers can gain valuable insights into consumer preferences and behaviours. This information can be utilized to optimize product offerings, marketing strategies, and overall store operations to better serve their customers' needs.

References

Bhatnagar, Amit; Ghose, S. (2004), ‘A latent class segmentation analysis of e-shoppers’, Journal of Business Research 57, 758–767.

Chen, D., Sain, S. L. & Guo, K. (2012), ‘Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining’, Journal of Database Marketing & Customer Strategy Management 19(3), 197–208.
URL: https://doi.org/10.1057/dbm.2012.17

Cooil, B., Aksoy, L. & Keiningham, T. L. (2008), ‘Approaches to customer segmentation’, Journal of Relationship Marketing 6(3-4), 9–39.

Marcus, C. (1998), ‘A practical yet meaningful approach to customer segmentation approach to customer segmentation’, Journal of Consumer Marketing 15, 494–504.

Mattson, M. P. (2014), ‘Superior pattern processing is the essence of the evolved human brain’, Frontiers in Neuroscience 8, 265.

Morrison, C., Gruenewald, P. J., Freisthler, B., Ponicki, W. R. & Remer, L. G. (2014), ‘The economic geography of medical cannabis dispensaries in california’, International Journal of Drug Policy 25(3), 508 – 515.
URL: http://www.sciencedirect.com/science/article/pii/S0955395913002387

Rajaraman, A. & Ullman, J. D. (2011), Mining of Massive Datasets, Cambridge University Press, New York, NY, USA.

Rogers, S. & Girolami, M. (2016), A First Course in Machine Learning, Second Edition, Chapman & Hall/CRC.

Roux, M. (2018), ‘A comparative study of divisive and agglomerative hierarchical clustering algorithms’, Journal of Classification 35(2), 345–366. URL: https://doi.org/10.1007/s00357-018-9259-9

Shepitsen, A., Gemmell, J., Mobasher, B. & Burke, R. (2008), Personalized recommendation in social tagging systems using hierarchical clustering, in ‘Proceedings of the 2008 ACM

Conference on Recommender Systems’, RecSys ’08, ACM, New York, NY, USA, pp. 259–266.
URL: http://doi.acm.org/10.1145/1454008.1454048

Chat GPT