Team Member Name: Sistata Bagale

Group: 30

Project Title: Customer Segmentation Analysis using Clustering technique

Table of Contents

Introduction

In today's world, businesses seeking to improve customer satisfaction and profitability must understand their customers behaviors. Customer segmentation, which groups customers based on shared traits like buying habits and preferences, is a key strategy for tailoring marketing efforts, refining products, and fostering stronger customer connections.

Clustering techniques, part of unsupervised machine learning, are essential for analyzing customer segmentation. Unlike supervised learning, which depends on labeled data, clustering algorithms organize data points according to their natural similarities without needing predefined categories. This approach enables businesses to reveal hidden patterns within their customer information.

K-means and spectral clustering are two popular clustering techniques, each effective in different contexts. K-means is a simple and efficient method that divides data into a predetermined number of clusters by minimizing variance within those clusters. It works particularly well for spherical clusters and is computationally efficient for larger datasets. In contrast, spectral clustering utilizes a graph-based approach, allowing it to capture complex relationships and non-linear patterns, making it ideal for clusters with irregular shapes or when dealing with high-dimensional data.

This project will apply both K-means and spectral clustering to a market dataset to perform an in-depth customer segmentation analysis and assess the performance of these methods.

Problem Statement

Grasping customer segments is essential for businesses looking to refine their marketing strategies and elevate the customer experience. Conventional marketing approaches frequently use a one-size-fits-all model, which can fail to connect with varied customer groups. The modern customer craves personalization. Studies show that over two-thirds of consumers expect a one-on-one experience, even at scale [1]. Customer segmentation is crucial for understanding the audiences on a deeper level. By identifying distinct customer groups, the businesses can uncover the motivations behind their behavior, tailor strategies to meet their specific needs, and foster long-term loyalty [1]. This project aims to explore: How can clustering techniques effectively segment customers to boost targeted marketing efforts and enhance overall business performance?

Objectives

This project aims to:

1. Build the clustering models to segment the customers.
2. Tune the hyperparameters and compare the evaluation metrics.

Research Questions

- How do the performance and effectiveness of K-means compare to spectral clustering in this context?
- What distinct customer segments can be identified using better clustering models from 1?

- What business insights can be derived from the identified customer segments?

Methodology

Data

The data for this project was taken from Kaggle. The dataset contains information about the People, Place, Products and Promotion. The dataset has 2240 observations and 29 variables as:

ID: Customer's unique identifier

Year_Birth: Customer's birth year

Education: Customer's education level

Marital_Status: Customer's marital status

Income: Customer's yearly household income

Kidhome: Number of children in customer's household

Teenhome: Number of teenagers in customer's household

Dt_Customer: Date of customer's enrollment with the company

Recency: Number of days since customer's last purchase

Complain: 1 if the customer complained in the last 2 years, 0 otherwise

MntWines: Amount spent on wine in last 2 years

MntFruits: Amount spent on fruits in last 2 years

MntMeatProducts: Amount spent on meat in last 2 years

MntFishProducts: Amount spent on fish in last 2 years

MntSweetProducts: Amount spent on sweets in last 2 years

MntGoldProds: Amount spent on gold in last 2 years

NumDealsPurchases: Number of purchases made with a discount

AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise

AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise

AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise

AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise

AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise

Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

NumWebPurchases: Number of purchases made through the company’s website

NumCatalogPurchases: Number of purchases made using a catalogue

NumStorePurchases: Number of purchases made directly in stores

NumWebVisitsMonth: Number of visits to company’s website in the last month

Data Preparation

Data Cleaning, and feature engineering

For this clustering analysis, several data preparation steps were undertaken to ensure the dataset was clean and ready for unsupervised learning. The key tasks involved handling missing values, feature engineering, transforming categorical variables to reduce multicollinearity, and addressing the presence of outliers in numerical features.

The dataset contained missing values in the *income* feature. To address this, I employed mean imputation, which is suitable for continuous variables with a small number of missing values. This approach helps to maintain the overall distribution of the income variable and prevents the clustering algorithm from being negatively impacted by missing information.

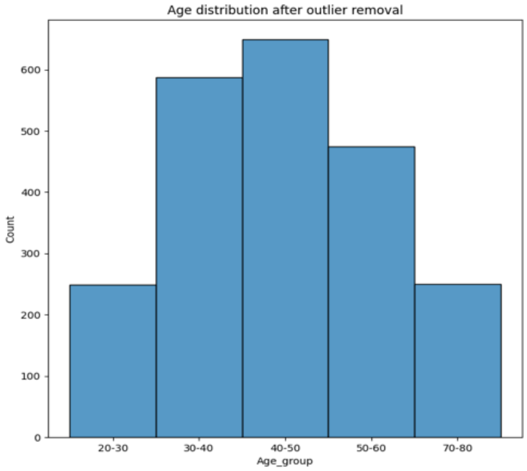
In this dataset, *Marital_Status* and *Education* were only categorical variables. Categories under these variables could be combined to one. For example, “alone”, “YOLO”, “absurd” were combined into one level “single”. This consolidation reduces the number of unique categories, helping to avoid overfitting in the clustering process and improving the interpretability of the resulting clusters. Likewise, “basic”, “2n cycle" were combined as "undergrad","masters", "phd" were combined to "postgrad". This would help in reducing the multicollinearity which can lead to unreliable and distorted clustering results.

I created new column *Age* from customer's *year_birth* to explore the age distribution of the customer. From the *Kidhome* and *Teenhome,* a new column, k*ids,* was created to get the sense of spending behavior of a person based on the number of children they have. Likewise, total number of family members was extracted from the marital status and the number of kids. I created another feature *'Total_Spent'* by summing the amount that a customer spent on a product. The redundant columns were dropped off the dataset.

Outliers can distort clustering results, particularly in distance-based algorithms like K-means. To mitigate this, I employed visual inspection (boxplots and histograms) and statistical methods (z-scores) to identify outliers. Outliers were present in the age and income columns. For age, I decided to remove values above 80 to avoid extreme influence. Likewise, outliers in the income column were removed to ensure a more accurate clustering.

Exploratory Data Analysis

I performed exploratory data analysis to get some insights on customers. The distribution of age of customer suggests that a significant portion of the customer base falls within the age range of 40-50 years.



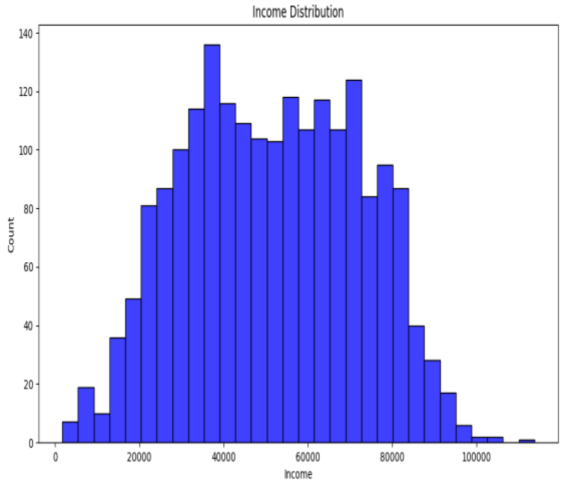


Figure 1: Distribution of age

Figure 2: Income Distribution of Customer

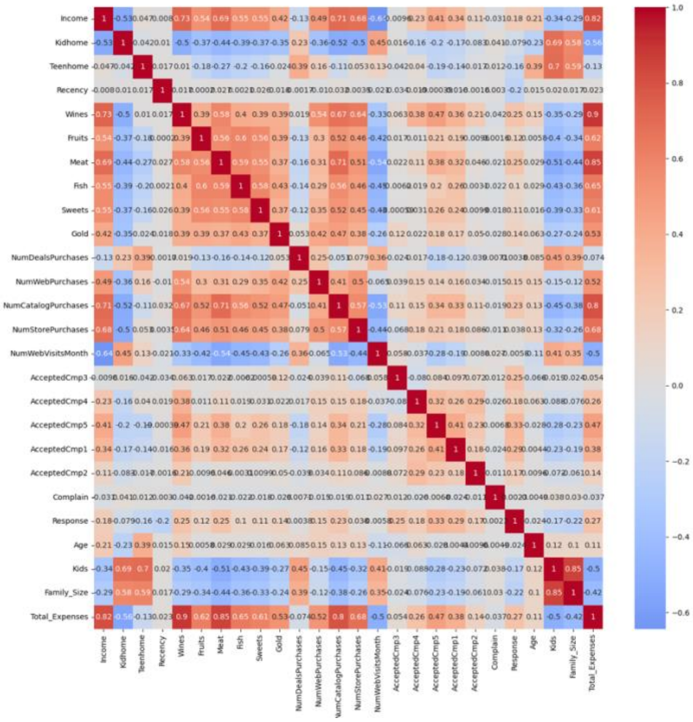


Figure 3: Correlation among the features

Correlation plot among the features shows that some of the features show high correlation with the values >0.8.

Data Preprocessing

Standardization

I applied feature scaling to ensure that all variables contributed equally to the clustering process. Features were standardized using z-score normalization. This step was critical to ensure that features with larger numeric ranges did not disproportionately influence the distance metric.

Dimensionality Reduction (Principal Component Analysis)

I performed correlation test amongst the features (figure 3) and it was observed that many features were highly correlated. Due to this, there might be the redundancy in the data information and such features might reflect similar patterns in the data. Therefore, I decided to perform Principal Component Analysis which performs the linear transformation on the data, projecting the original features to new sets of uncorrelated components i.e., principal components.

Table 1: Variance Explained by Principal Components

	Principal Components	Variance Explained (%)		Principal Components	Variance Explained (%)
0	PC1	37.75	10	PC11	2.47
1	PC2	12.77	11	PC12	2.11
2	PC3	6.79	12	PC13	2.03
3	PC4	6.21	13	PC14	1.84
4	PC5	5.09	14	PC15	1.72
5	PC6	4.79	15	PC16	1.27
6	PC7	4.38	16	PC17	1.12
7	PC8	3.18	17	PC18	0.69
8	PC9	2.95	18	PC19	0.00
9	PC10	2.79	19	PC20	0.00

Analyzing the explained variance, it was observed that the first three principal components account for 57.31% (table 1) of the total variance. To capture 90% of the variance, approximately 10 principal components would be required (figure 4), indicating a complex dataset with multiple interrelated factors. Despite this, a comparison of clustering performance with 3 and 10 PCs revealed that 3 PCs yielded better results. This suggests that adding more principal components might introduce noise or irrelevant information, potentially hindering the clustering process. I decided to move forward with 3 PCs.

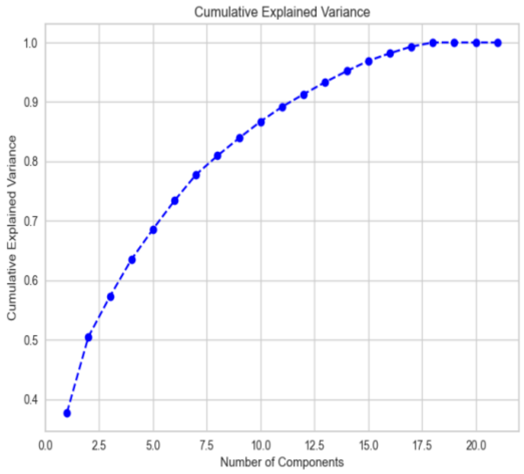


Figure 4: Cumulative Explained variance by PCs

To understand the underlying patterns captured by the principal components, I analyzed the loadings of PCA (appendix). PC1 is primarily driven by income and product purchases, suggesting a general spending behavior. PC2 reflects demographic and family-related factors, such as age, family size, and education. PC3 seems to capture lifestyle and personal characteristics, including education and marital status.

K-means Clustering

K-means clustering divides a dataset into 'k' distinct clusters. The algorithm begins by randomly selecting 'k' data points as initial cluster centroids then iteratively performs two steps:

1) Assignment: Each data point is assigned to the nearest centroid based on a distance metric.

2) Update: The centroids are recalculated as the mean of the assigned data points. This process continues until convergence, where the cluster assignments stabilize [3].

Elbow method:

To determine the optimal number of clusters, elbow method was used. By plotting the Within-Cluster Sum of Squares against the number of clusters, we can visually identify the "elbow" point.

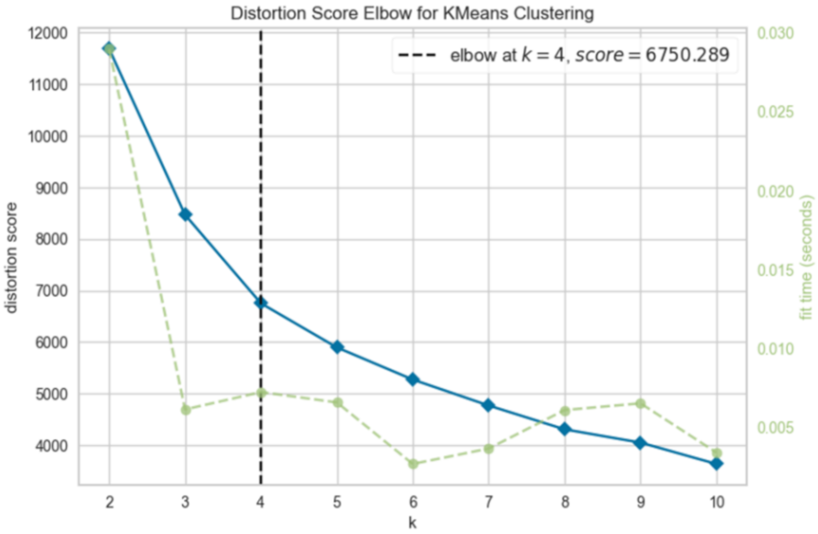


Figure 5: Optimal number of clusters by Elbow method

The elbow method, based on the distortion score, indicates that 4 is the optimal number of clusters. This is because the distortion score decreases significantly up to 4 clusters, and adding more clusters doesn't lead to a substantial reduction (figure 5). However, to gain further insights, I decided to explore clustering with 2, 3, and 4 clusters and compare their performance using various evaluation metrics.

2D visualization of k-means with two, three and four are shown in the figure below. This shows that the data points are separated into 2 clusters without any visual overlapping. However, the cluster 1 seems more dispersed while cluster 0 is more compact. K-means with 3 and 4 clusters gave distinct clusters. However, there is visible overlapping of datapoints between the clusters.

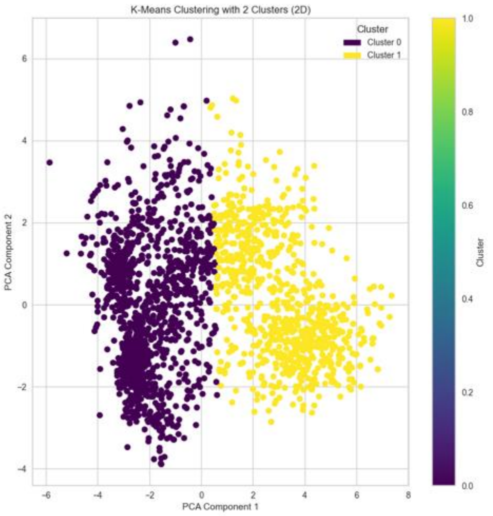


Figure 6: K-means with 2 clusters

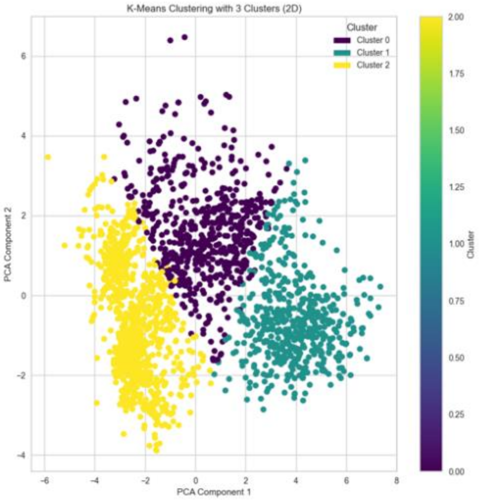


Figure 7: K-means with 3 clusters

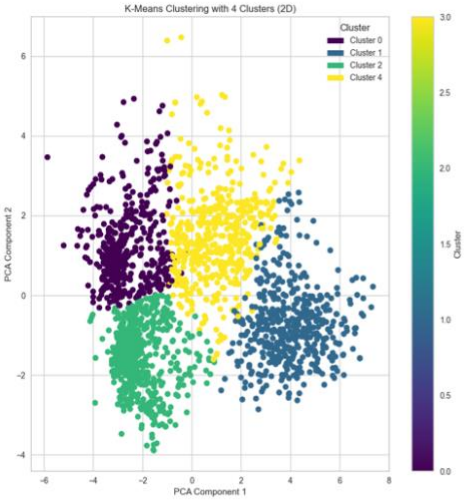


Figure 8: K-means with 4 clusters

Spectral Clustering

Spectral clustering is a clustering technique that leverages the graph-based representation of data. It uses the eigenvalues and eigenvectors of the similarity matrix to map data points to a lower-dimensional space, where clustering becomes more straightforward [5]. Steps while performing spectral clustering are:

- Graph Construction:

  Spectral clustering involves constructing a similarity graph based on the distance between data points. A similarity matrix is created, and the Laplacian matrix is computed from it.
- Eigenvalue Decomposition**:**

  Perform eigenvalue decomposition of the Laplacian matrix to find the eigenvectors corresponding to the smallest eigenvalues. These eigenvectors will be used to form a new feature space for clustering.

For spectral clustering also, I compared the three values of 'n' (number of clusters) i.e. 2, 3, and 4. The 2D visualization of clusters using spectral clustering algorithm are shown in figures 8, 9 and 10. From the visualization of plot, spectral clustering performed like k-means with 2 clusters (as we can see the similar patterns in the datapoints). However, it produces different clustering pattern for 3 and 4 clusters. The clusters seem sparser with n=3 and 4 and there are more overlapping data points.

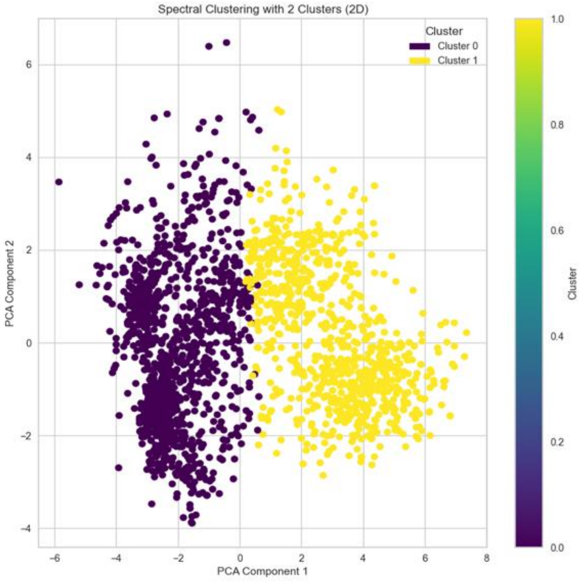


Figure 9: Spectral Clustering with 2 clusters

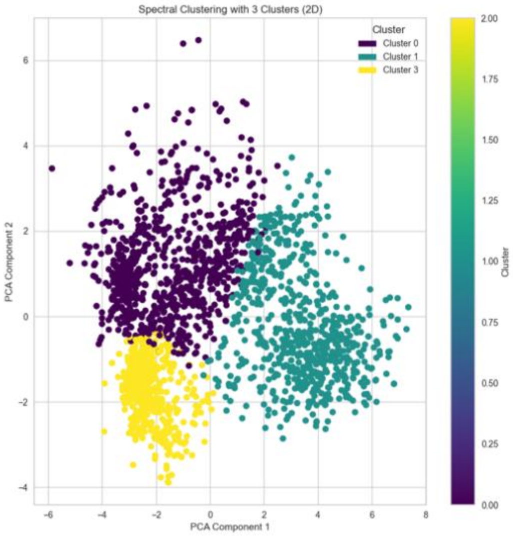


Figure 10: Spectral Clustering with 3 clusters

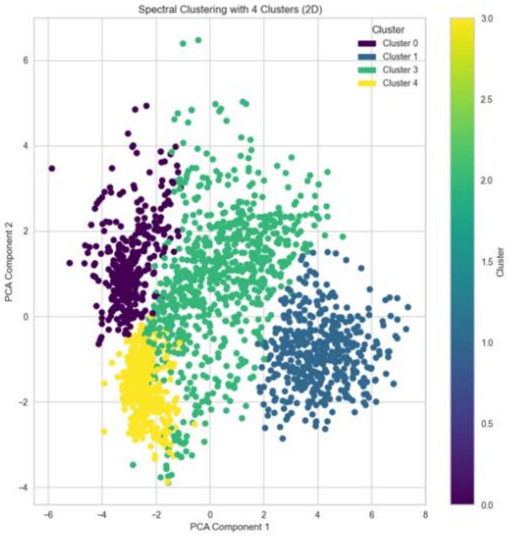


Figure 11: Spectral Clustering with 4 clusters

However, we can come to a more robust conclusion about which model performed better, by comparing the performance of these two algorithms using evaluation metrices. The quality of the clustering was further assessed using Silhouette score and Dunn's index.

Evaluation

Silhouette Score:

$$s = \frac{b - a}{\max(a, b)}$$

where:

"a" is the mean distance between a sample and all other points in the same cluster.

"b" is the mean distance between a sample and all other points in the next nearest cluster.

This metric measures how similar a data point is to its own cluster compared to other clusters. The score ranges from -1 to 1, where a higher score indicates better-defined clusters [6].

Dunn's Index:

The Dunn's Index is calculated by taking the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance. Higher the DI, more compact the cluster is and well separated from other clusters [6].

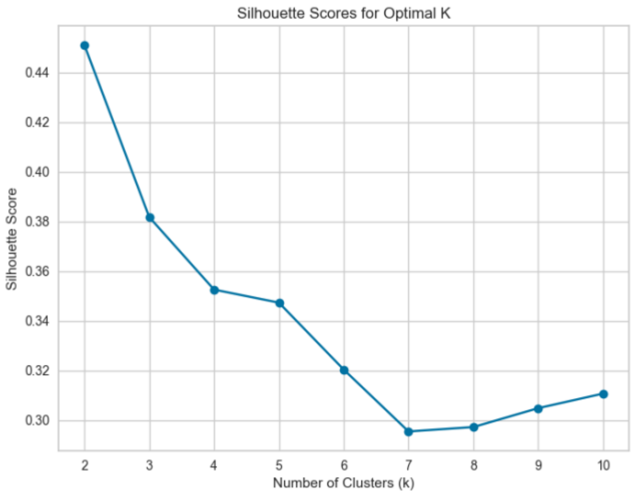


Figure 12: Silhouette Score for optimal k

Table 2: Silhouette Score comparison

Number of Clusters	K-means Clustering	Spectral Clustering
2	0.451	0.449
3	0.382	0.345
4	0.353	0.314

Table 3: Dunn's Index value

Number of Clusters	K-means Clustering	Spectral Clustering
2	0.0078	0.017
3	0.0013	0.011
4	0.0011	0.009

From the evaluation of clustering method, it was observed that k-means outperforms spectral clustering in terms of the Silhouette Score, particularly when k = 3 and k = 4. The silhouette scores suggest that the optimal number of clusters is likely 2 or 3, with k = 2 being a relatively better choice in terms of cluster separation for both algorithms. However, spectral clustering generally has a higher Dunn's Index, suggesting that it may produce better-defined clusters (more separated and less internally spread) compared to K-means.

Based on the evaluation metrices, k=2 is likely the best number of clusters for this dataset.

<u>Results and Discussion</u>

I analyzed the customer's income versus spending plots. However, the clusters did not appear well-separated (figure 13). This might be because of the overlapping clusters and k-mean failed to separate the data well. Due to this, I explored hierarchical agglomerative clustering which handles overlapping clusters better than K-Means. Upon evaluating the performance of agglomerative clustering using evaluation metrices, it performed close to k-means. The Silhouette Score for agglomerative clustering was 0.360 using 3 clusters which is comparable to k-mean's 0.382.

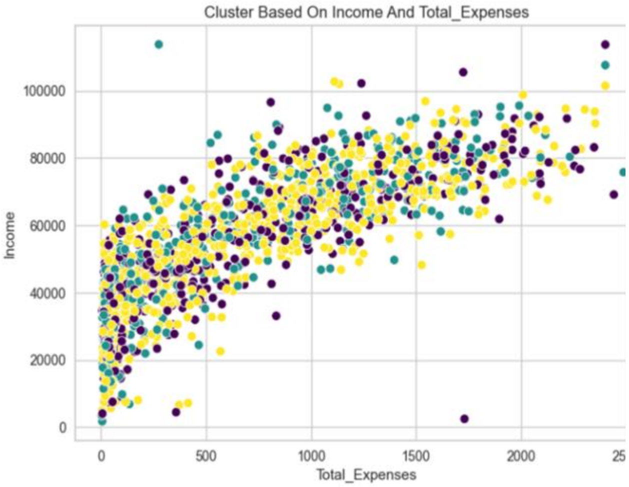


Figure 13: Cluster based on Income vs Spending (k-means)

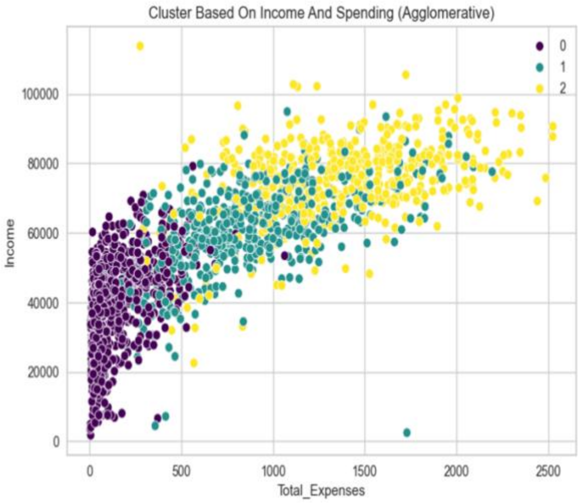


Figure 14: Cluster based on Income vs Spending (Agglomerative)

From the figure 14, we could say that hierarchical agglomerative clustering formed the clustering pattern that would make more sense than the k-means. Income and spending both increases from cluster 0 to 2. Cluster 0 shows the group of people with low income and low expenses while cluster 2 shows high income and high expenditure.

To gain insights into customer behavior, I examined their interactions with promotions and deals. The analysis revealed that customers in Cluster 0 were more receptive to promotions and actively purchased deals. In contrast, customers in Cluster 2 exhibited lower engagement with both promotions and deals.

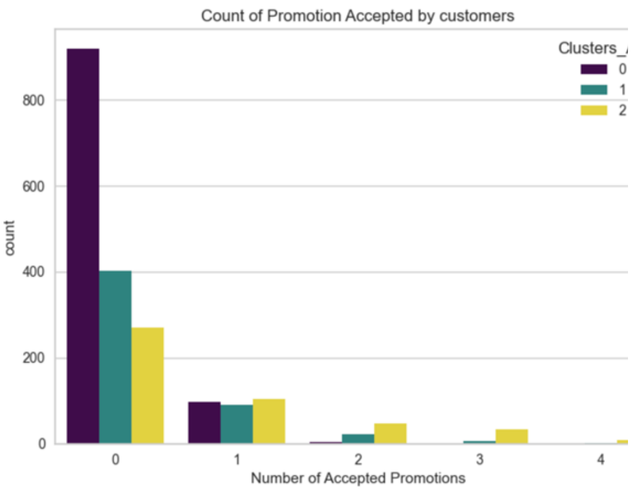


Figure 15: Promotion Accepted by Customers

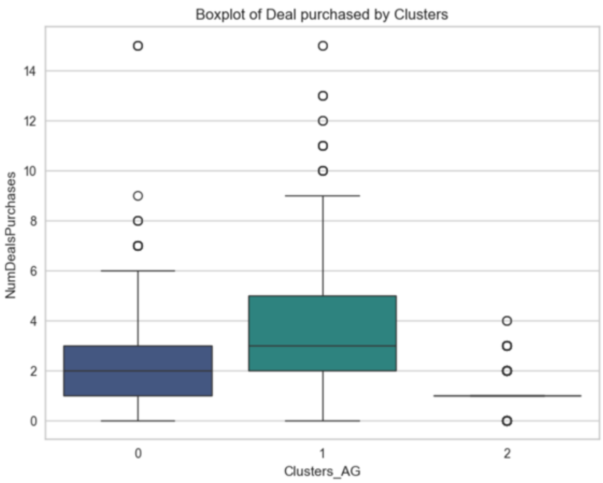


Figure 16: Deal purchased by Cluster

Examining family-oriented features, I observed distinct patterns across the clusters (figure 18-20 in appendix). Clusters 0 and 1 tend to have families of 2-3 members, while Cluster 2 typically has smaller families of 1-2 members. In terms of children, Cluster 0 and 1 have 1-2 kids, while the majority of people in Cluster 2 are childless. Additionally, the age distribution varies across clusters. Cluster 0 represents a younger demographic with less age diversity, Cluster 2 corresponds to an older demographic with a wider age range, and Cluster 1 appears to be a transitional group between the two.

Conclusion

This project utilized clustering techniques to segment customers based on their demographic and behavioral characteristics. Three distinct customer segments were identified:

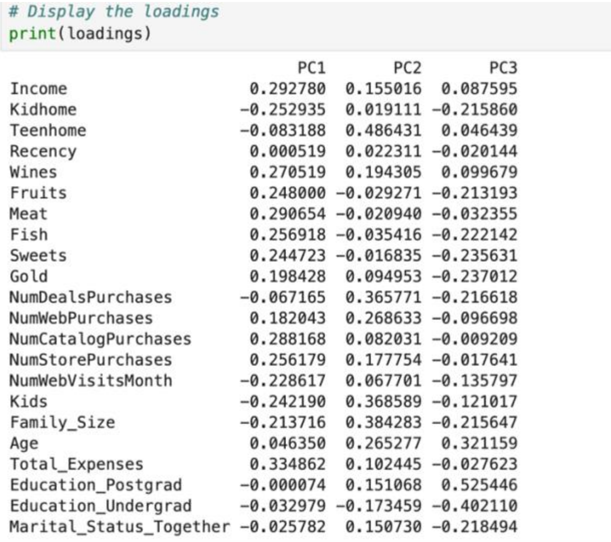
- **Segment 1:** Low-income, young families. Marketing strategies should emphasize affordability and family-oriented offerings.
- **Segment 2:** Mid-income, transitional demographic. Marketing efforts should focus on mid-range products and personalized recommendations.
- **Segment 3:** High-income, older demographic. Marketing strategies should emphasize premium products, exclusive experiences, and long-term value.

Businesses can, therefore, implement effective marketing strategies that meet the needs and preferences of each customer segment. These actionable recommendations would help organizations in increasing customer engagement, driving revenue growth, effective marketing, and building long-term relationships with their diversified customer base. This segmentation could be a source of informed decision-making and strategic planning for the success of business

References:

1. https://www.qualtrics.com/experience-management/brand/customer-segmentation/
2. https://www.geeksforgeeks.org/k-means-clustering-introduction/
3. https://www.researchgate.net/figure/Accelerated-K-SC-Clustering-Algorithm_fig1_288626888
4. https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/
5. https://www.geeksforgeeks.org/ml-spectral-clustering/
6. https://www.analyticsvidhya.com/blog/2020/10/quick-guide-to-evaluation-metrics-for-supervised-and-unsupervised-machine-learning/
7. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

Appendix:



```
# Display the loadings
print(loadings)
```

```
                             PC1       PC2       PC3
Income                  0.292780  0.155016  0.087595
Kidhome                -0.252935  0.019111 -0.215860
Teenhome               -0.083188  0.486431  0.046439
Recency                 0.000519  0.022311 -0.020144
Wines                   0.270519  0.194305  0.099679
Fruits                  0.248000 -0.029271 -0.213193
Meat                    0.290654 -0.020940 -0.032355
Fish                    0.256918 -0.035416 -0.222142
Sweets                  0.244723 -0.016835 -0.235631
Gold                    0.198428  0.094953 -0.237012
NumDealsPurchases      -0.067165  0.365771 -0.216618
NumWebPurchases         0.182043  0.268633 -0.096698
NumCatalogPurchases     0.288168  0.082031 -0.009209
NumStorePurchases       0.256179  0.177754 -0.017641
NumWebVisitsMonth      -0.228617  0.067701 -0.135797
Kids                   -0.242190  0.368589 -0.121017
Family_Size            -0.213716  0.384283 -0.215647
Age                     0.046350  0.265277  0.321159
Total_Expenses          0.334862  0.102445 -0.027623
Education_Postgrad     -0.000074  0.151068  0.525446
Education_Undergrad    -0.032979 -0.173459 -0.402110
Marital_Status_Together -0.025782  0.150730 -0.218494
```

Figure 17: PCA Loadings

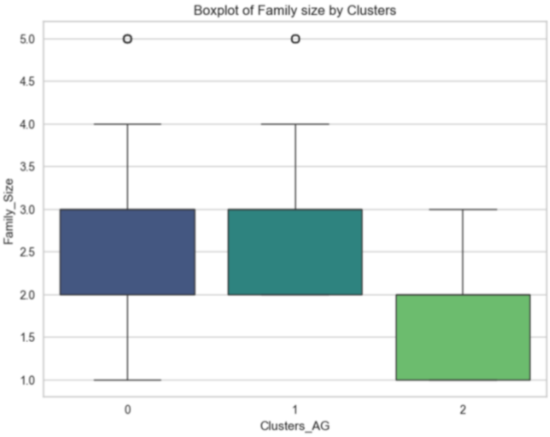


Figure 18: Boxplot of family size by clusters

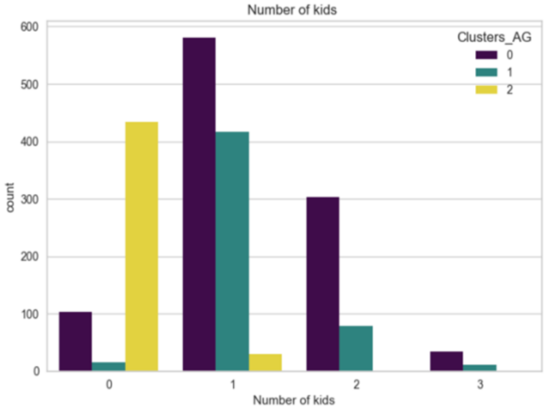


Figure 19: Number of kids by clusters

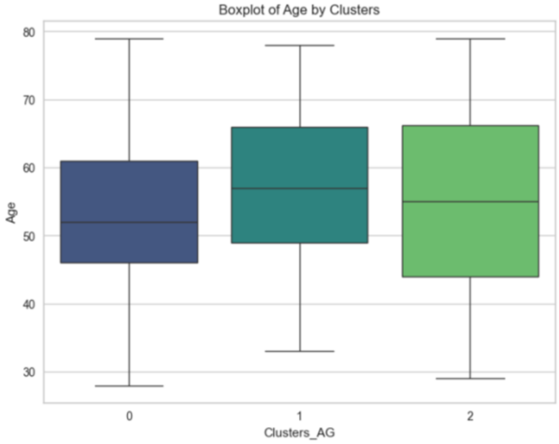


Figure 20: Boxplot of age by clusters