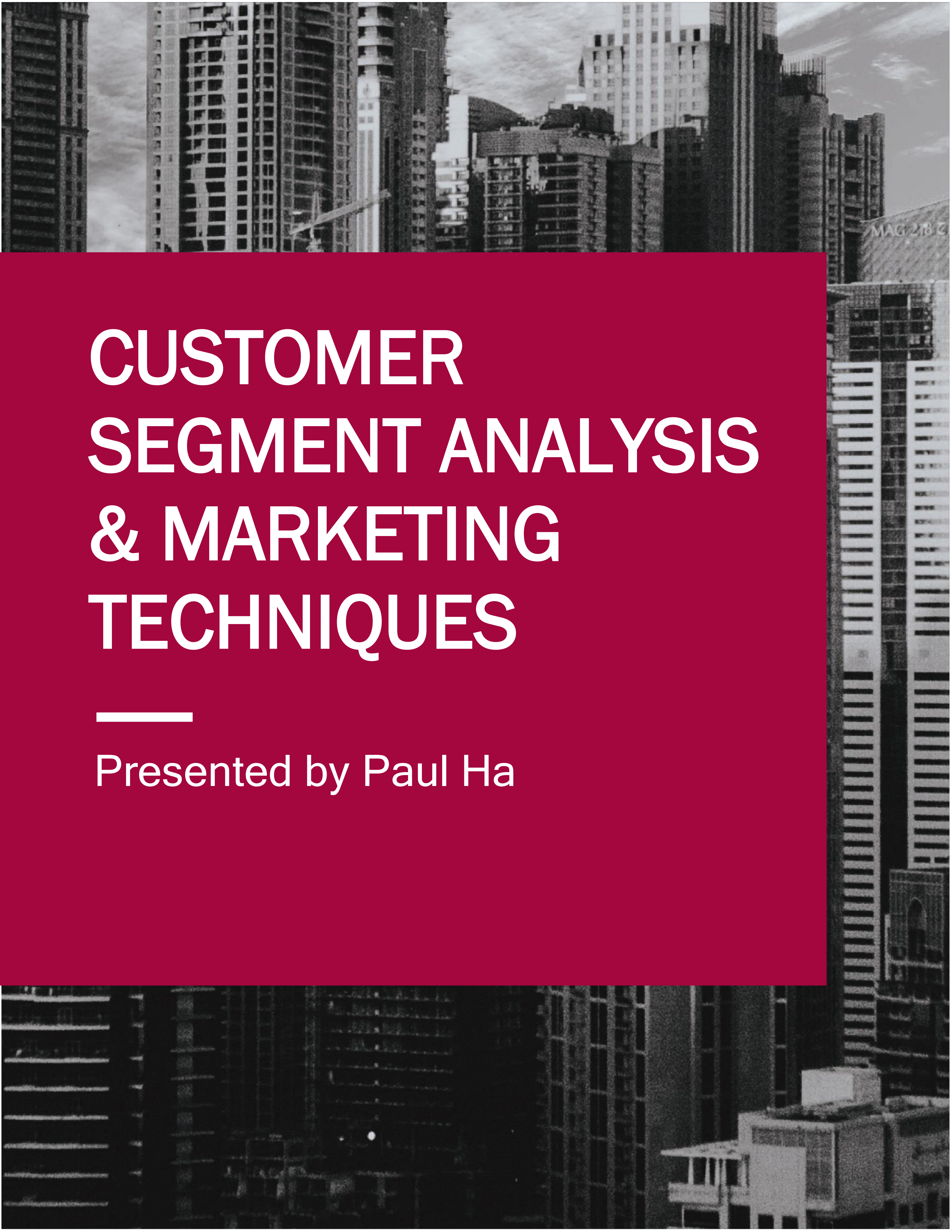


CUSTOMER SEGMENT ANALYSIS & MARKETING TECHNIQUES

Presented by Paul Ha

Table of Contents

Part 1: Introduction

Customer segmentation plays a crucial role for businesses by enabling them to tailor marketing efforts and deliver personalized experiences to various customer groups. For large supermarket chains, understanding the diverse preferences and characteristics of their customers is essential to optimize marketing campaigns, enhance customer satisfaction, and drive revenue growth.

This report focuses on conducting a customer segmentation analysis for a supermarket chain. The analysis utilizes a dataset containing information from 4,000 customers, collected via loyalty cards at the checkout. This dataset provides insights into various customer demographics, such as age, gender, and income.

The primary objective is to identify distinct customer segments within this dataset using clustering techniques, including k-Means and Hierarchical Tree clustering.

By grouping customers with similar attributes, the report aims to uncover valuable patterns and provide a better understanding of the different customer profiles within the supermarket's customer base.

By the end of the analysis, the supermarket's management team will have a detailed overview of the identified customer segments. This knowledge will enable them to make data-driven marketing decisions and tailor their strategies to meet the unique needs of each segment. The ultimate goal is to boost customer engagement, enhance satisfaction, and foster business growth.

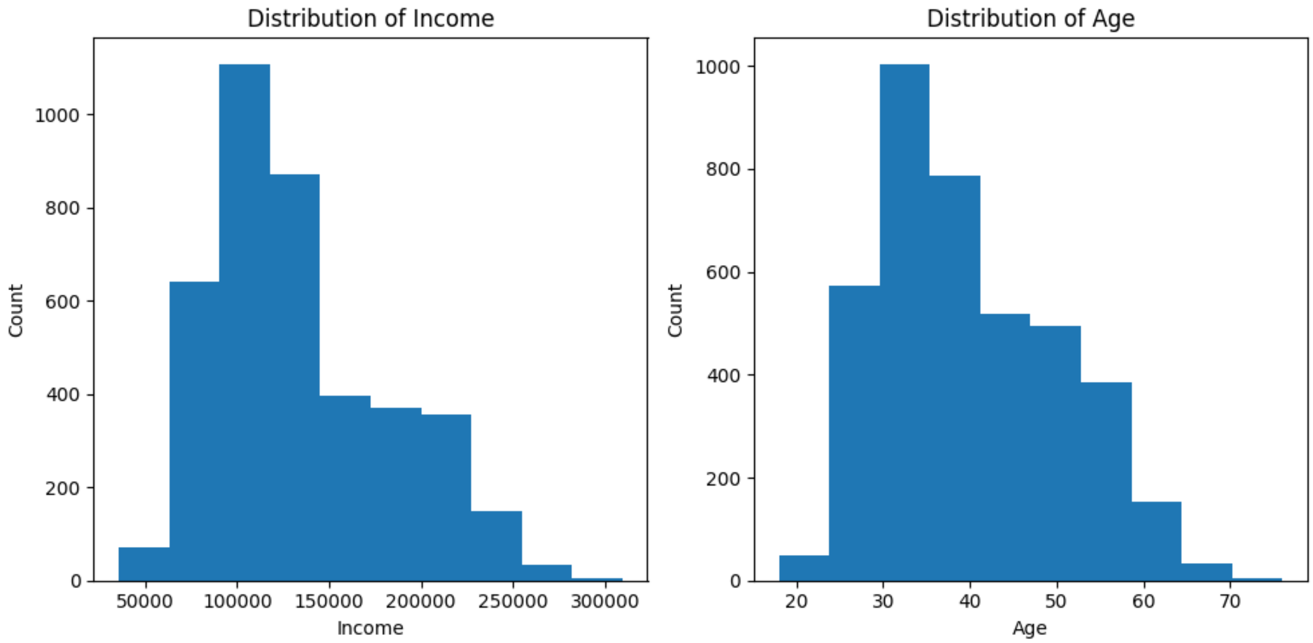
Part 2: Exploratory Data Analysis

Overall, the dataset contains 4000 rows of unique customer data with 7 variables Gender, Marital Status, Education, Settlement Size, Occupation, Income, Age. They are all numerical variables without any missing values that need to be handled.

2.1 Summary statistics and distribution of each individual variables

Gender	51.05% are male, 48.95% are female
Marital Status	51.05% are single, 48.95% are non-single
Education	The majority of customers have a high school education as their highest level of qualification (1609), followed by those from graduate school (1270), and university (707). Customers with "Unknown" education has lowest representation (414).
Settlement Size	The distribution of customers based on city size reveals that the highest number of customers reside in big cities, followed by small cities, and the lowest number in mid-sized cities.
Occupation	94% of customers are skilled employees, or at a management level, and the least number is unemployed individuals.

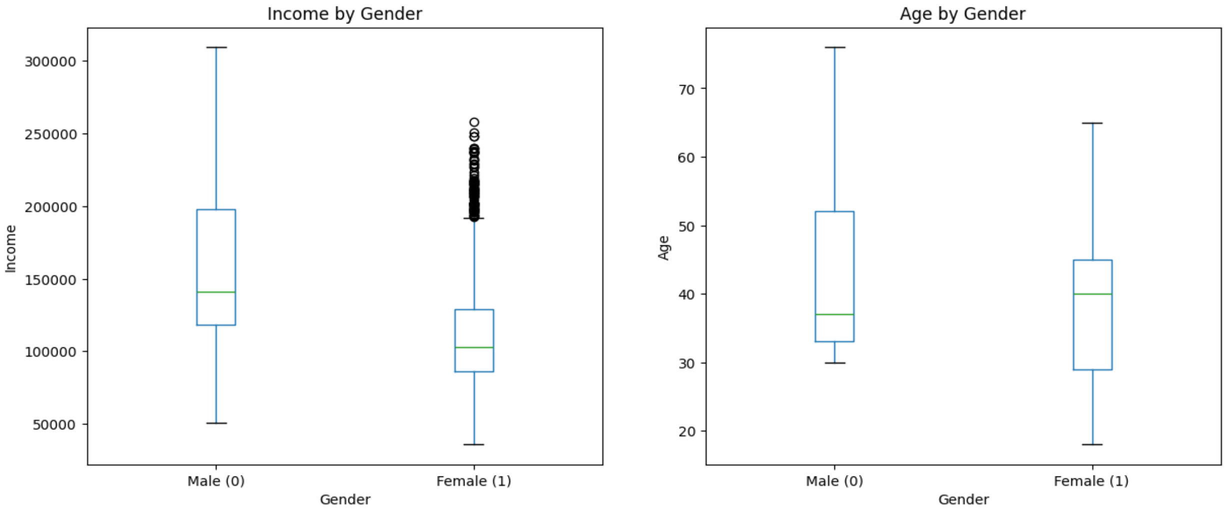
Age	The average age is 39.94, with a standard deviation of 10.26. The youngest customers are 18 years old, while eldest customers are 75 years old. Looking at the graph 2.1.1, the distribution seems to be normal, bell-shaped, a little bit right skewed, with the majority of the distribution is in range 25 to 55 years old.
Income	The mean income is $134,353.79, with a standard deviation of $48533.57. The maximum income is $309,364 and the minimum income is $35,832. By examining the histogram of income, income values are normally distributed, indicating a wide range of income levels among the customers.



graph 2.1.1: *distribution of income and age variables*

The graph 2.1.2 shows that **male customers generally earn higher incomes** than female customers. Male customers appear to be more financially stable, with a larger proportion falling into higher income brackets. In contrast, **female customers tend to have lower incomes on average**, although there is a small group of high-income females who stand out from the majority.

The age distribution shows that **male customers are generally older**, with the youngest male customers in the dataset around **30 years old** and the oldest around **75 years old**. On the other hand, **female customers cover a wider age range**, starting from their **early 20s** to around **60 years old**.



graph 2.1.2: *income by gender and age by gender*

<u>Marital Status By Gender</u>

- *The number of single females is almost double that of single males. In contrast, for non-single individuals, the number of females is roughly half that of males.*

<u>Education By Gender</u>

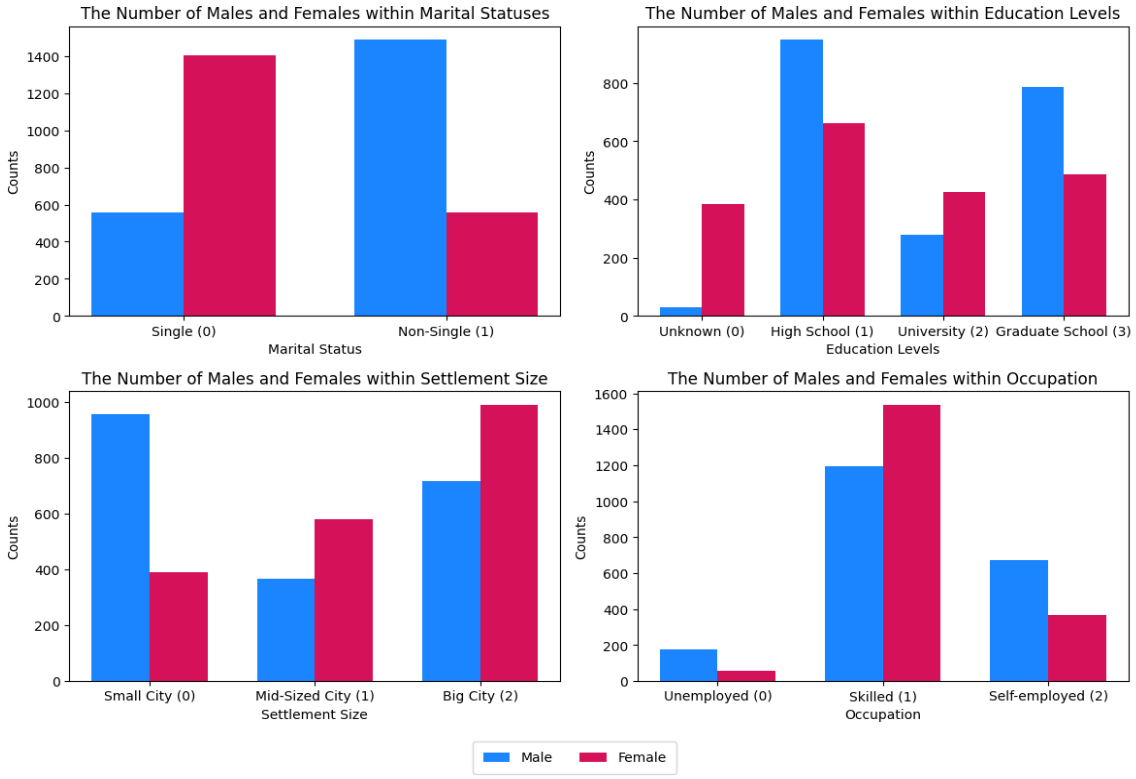
- *Among whose education is unknown, females significantly outnumber males (approximately 400 and 15 respectively).*
- *The education level distribution shows that among university level, females are slightly more numerous than males.*
- *At the high school and graduate school level, males outnumber females (around 1000 and 700 respectively at high school, and around 800 and 500 respectively).*

<u>Settlement Size By Gender</u>

- *In small cities, the number of males is roughly 2.5 times higher than that of females, nearly 950 compared with 400 respectively.*
- *In mid-sized and big cities, the number of females are slightly larger than that of males.*

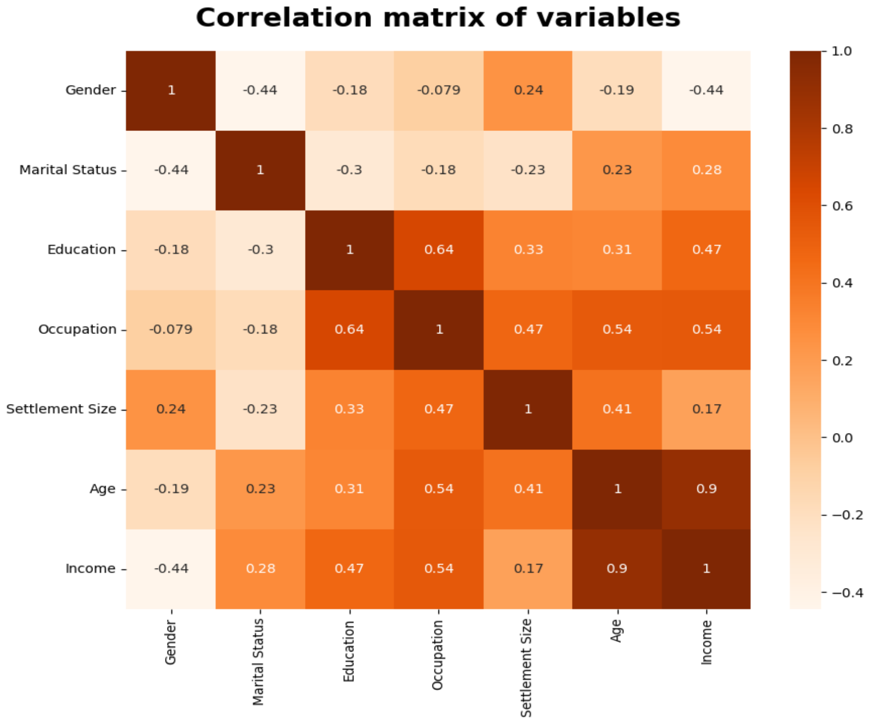
<u>Occupation By Gender</u>

- *Among skilled workers, females are slightly more numerous than males.*
- *Among unemployed and self-employed individuals, there are more males than females.*
- *Among skilled workers, females are slightly more numerous than males.*
- *Among unemployed and self-employed individuals, there are more males than females.*



graph 2.1.3: *gender statistics*

2.2 Relationships between variables



graph 2.2.1: *correlation matrix*

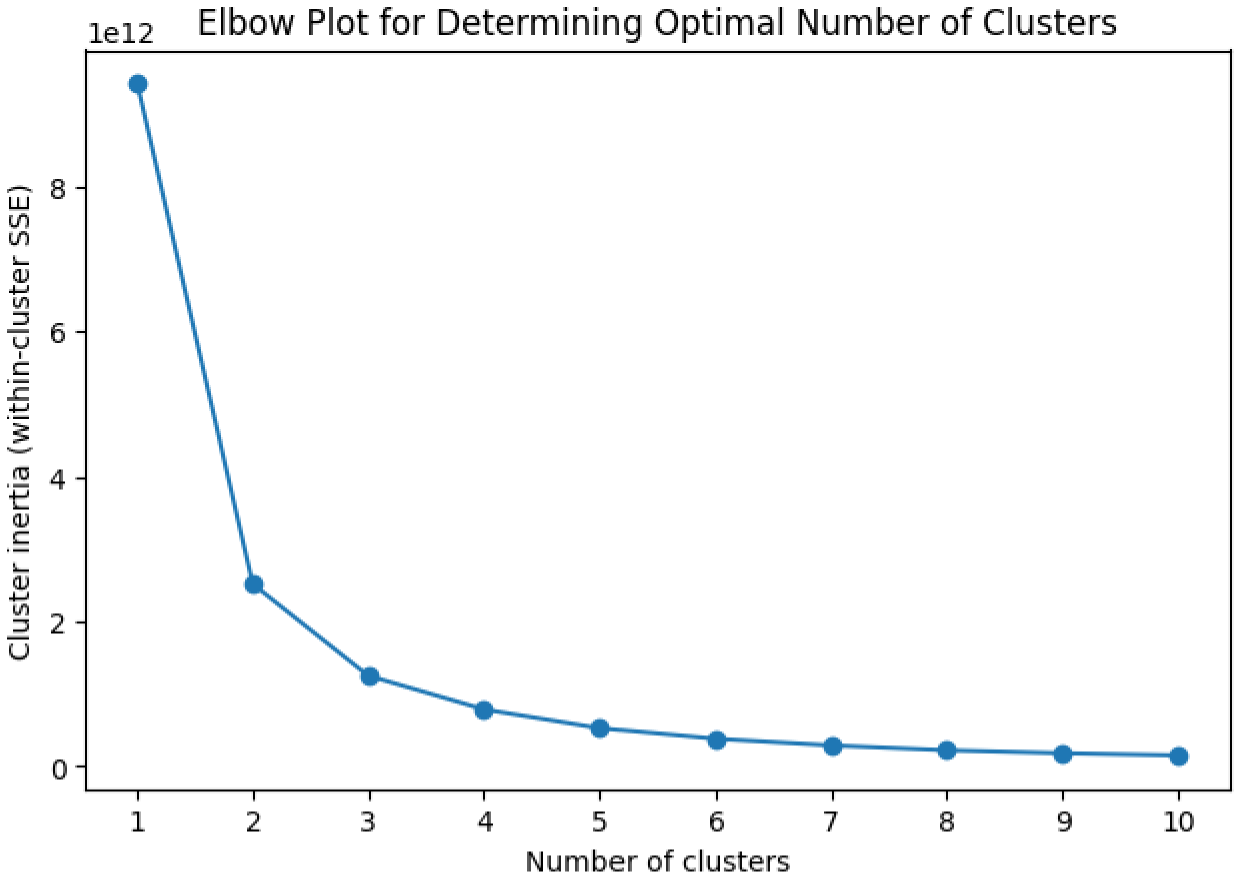
Age and income are highly correlated (0.9), followed by occupation and education (0.64).

- It is understandable that the higher the level of education is, the higher level of the occupation is.
- Income rises when customers get older. Probably, they have more experience than younger people, so they are paid higher.

Part 3: Clustering

To conduct customer segmentation, we apply k-means and agglomerative clustering algorithms. The elbow method is used to identify the optimal number of clusters for the k-means approach.

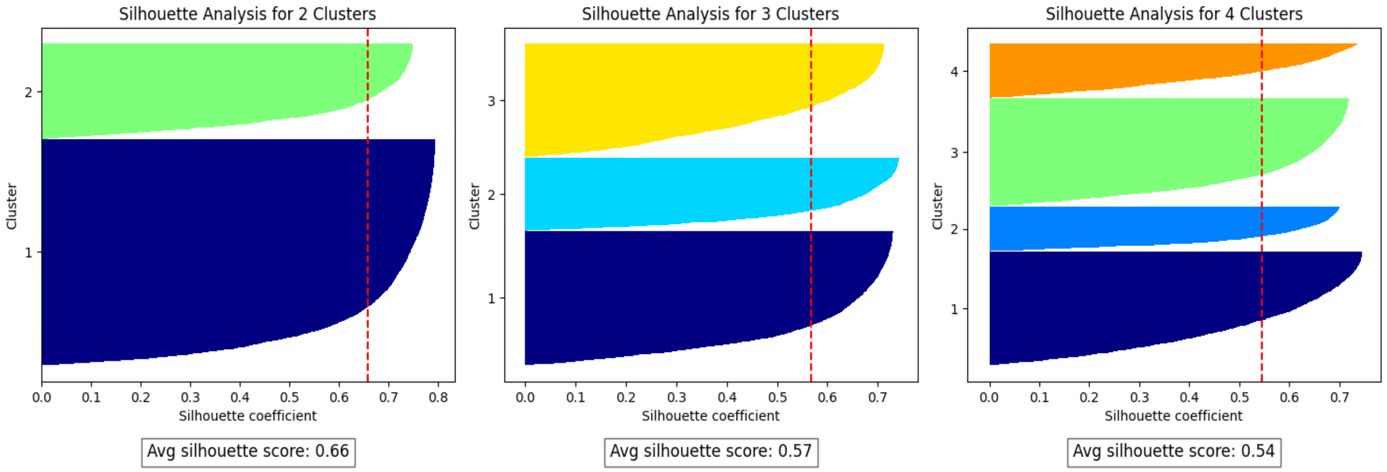
3.1 K-Means++ clustering:



***graph 3.1.1:** elbow method*

Based on the application of the k-means algorithm, we **conducted the elbow method analysis to determine the optimal number of clusters**. The elbow plot revealed that the ideal number of clusters is 2, 3, or 4.

To further validate the results, we used Silhouette Scores, a metric that evaluates the quality of clustering. In simple terms, the Silhouette Score measures how well each customer fits within its assigned group (or cluster) and how distinct the groups are from one another. A higher score indicates that the clusters are more clearly defined and meaningful.

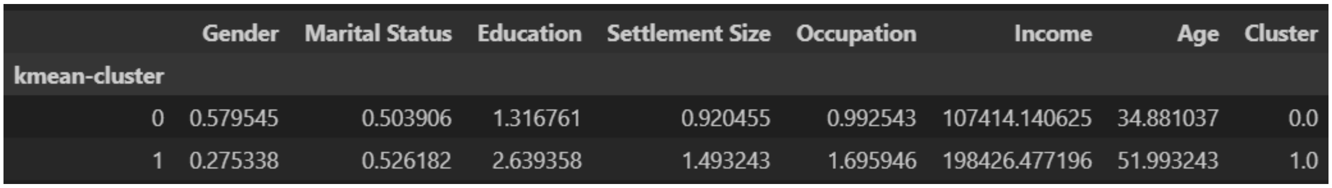


graph 3.1.2: *Silhouette scores*

For this analysis:

- 2 clusters achieved the highest Silhouette Score of 0.66, indicating strong and distinct groupings.
- 3 clusters had a lower score of 0.57, suggesting slightly less clear separation.
- 4 clusters had the lowest score of 0.54, meaning the clusters were less distinct and cohesive.

These results confirm that **2 clusters** provide the most effective segmentation for this dataset.



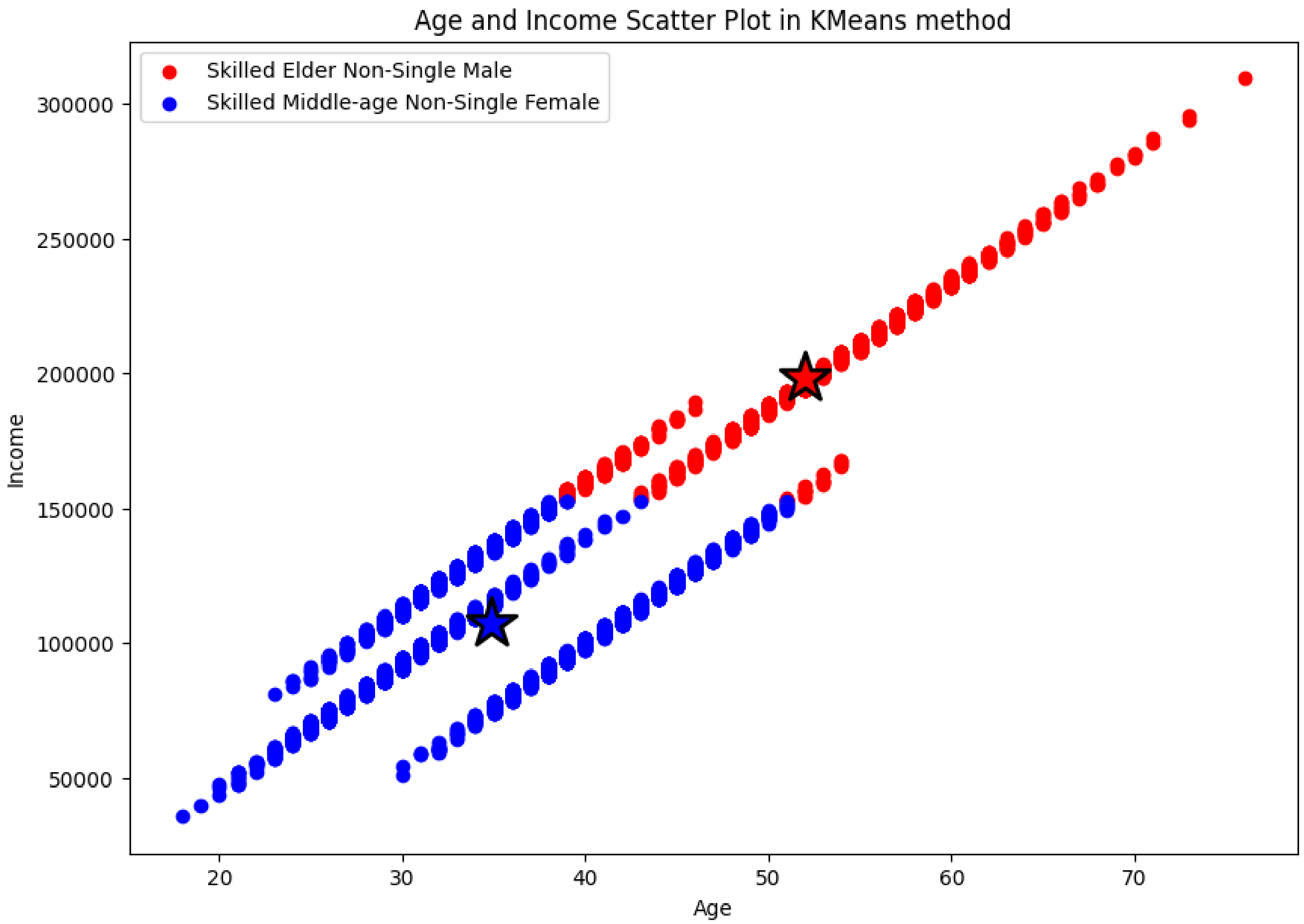
	Gender	Marital Status	Education	Settlement Size	Occupation	Income	Age	Cluster
kmean-cluster								
0	0.579545	0.503906	1.316761	0.920455	0.992543	107414.140625	34.881037	0.0
1	0.275338	0.526182	2.639358	1.493243	1.695946	198426.477196	51.993243	1.0

graph 3.1.3: mean values of the variables for each customer group

The profile of 2 groups of customers based on the graph above are:

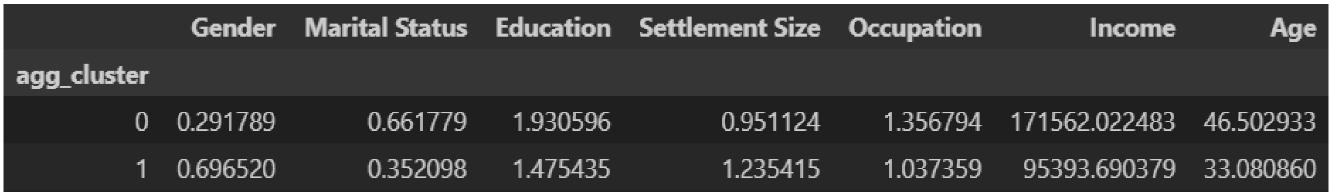
- **Group 1:** The "Skilled Middle-age Non-Single Female" segment comprises middle-aged females with high school education and skilled occupation. They reside in small cities and have a moderate income.
- **Group 2:** The "Skilled Elder Non-Single Male" segment comprises elder males with high graduate education and skilled occupation. They reside in big cities and have a high income.

The scatter plot below provides a clear visual representation of the distribution of income and age across customer groups identified using the KMeans method. It displays all data points along with the centroids of each cluster, enabling us to examine the relationship between age and income within each group.



graph 3.1.4: *age and income scatter plot in k-means method*

3.2 Agglomerative clustering:



	Gender	Marital Status	Education	Settlement Size	Occupation	Income	Age
agg_cluster							
0	0.291789	0.661779	1.930596	0.951124	1.356794	171562.022483	46.502933
1	0.696520	0.352098	1.475435	1.235415	1.037359	95393.690379	33.080860

graph 3.2.1: *mean values of the variables for each customer group*

The profile of 2 groups of customers based on the graph above are:

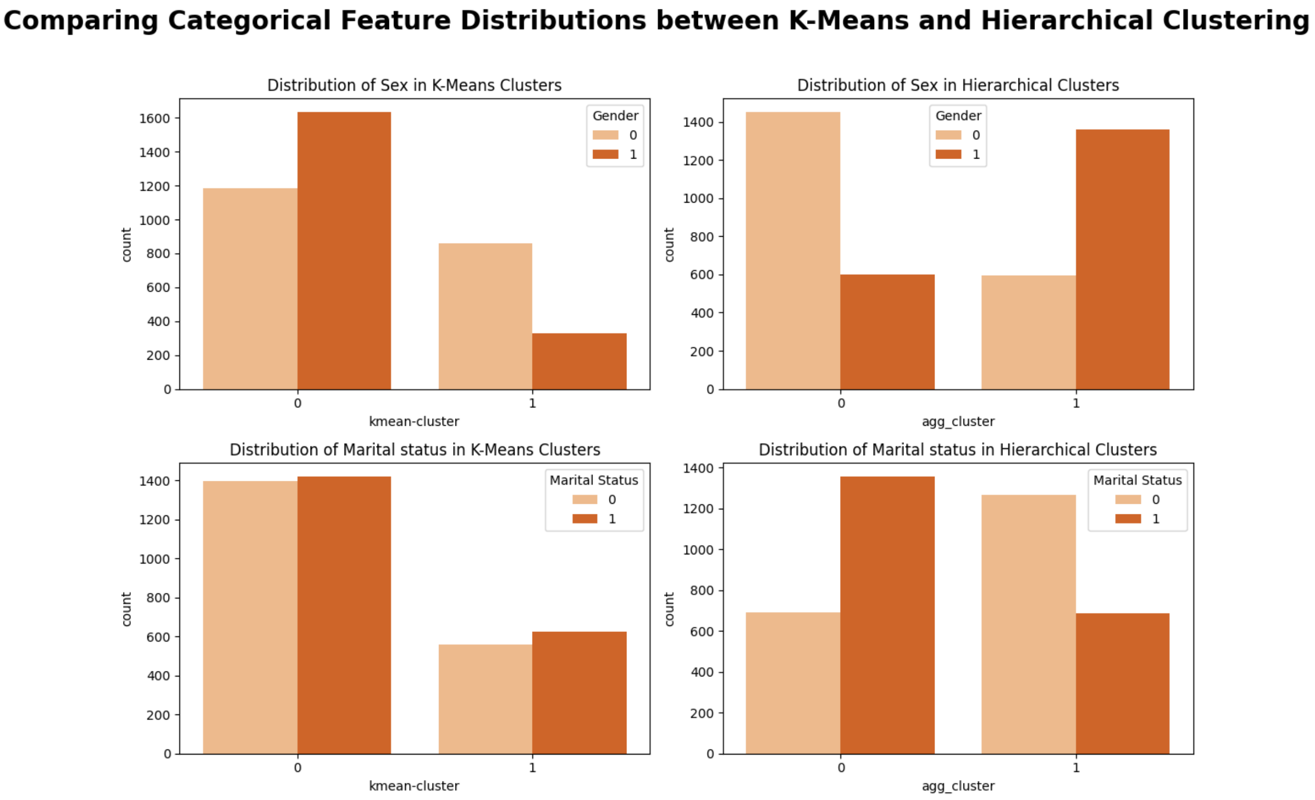
- **Group 1:** The “Skilled Middle-age Non-Single Male” segment comprises middle-aged males with university education and skilled occupation. They reside in mid-sized cities and have a high income.
- **Group 2:** The “Skilled Middle-age Single Female” segment comprises middle-aged females with high school education and skilled occupation. They reside in mid-sized cities and have a moderate income.

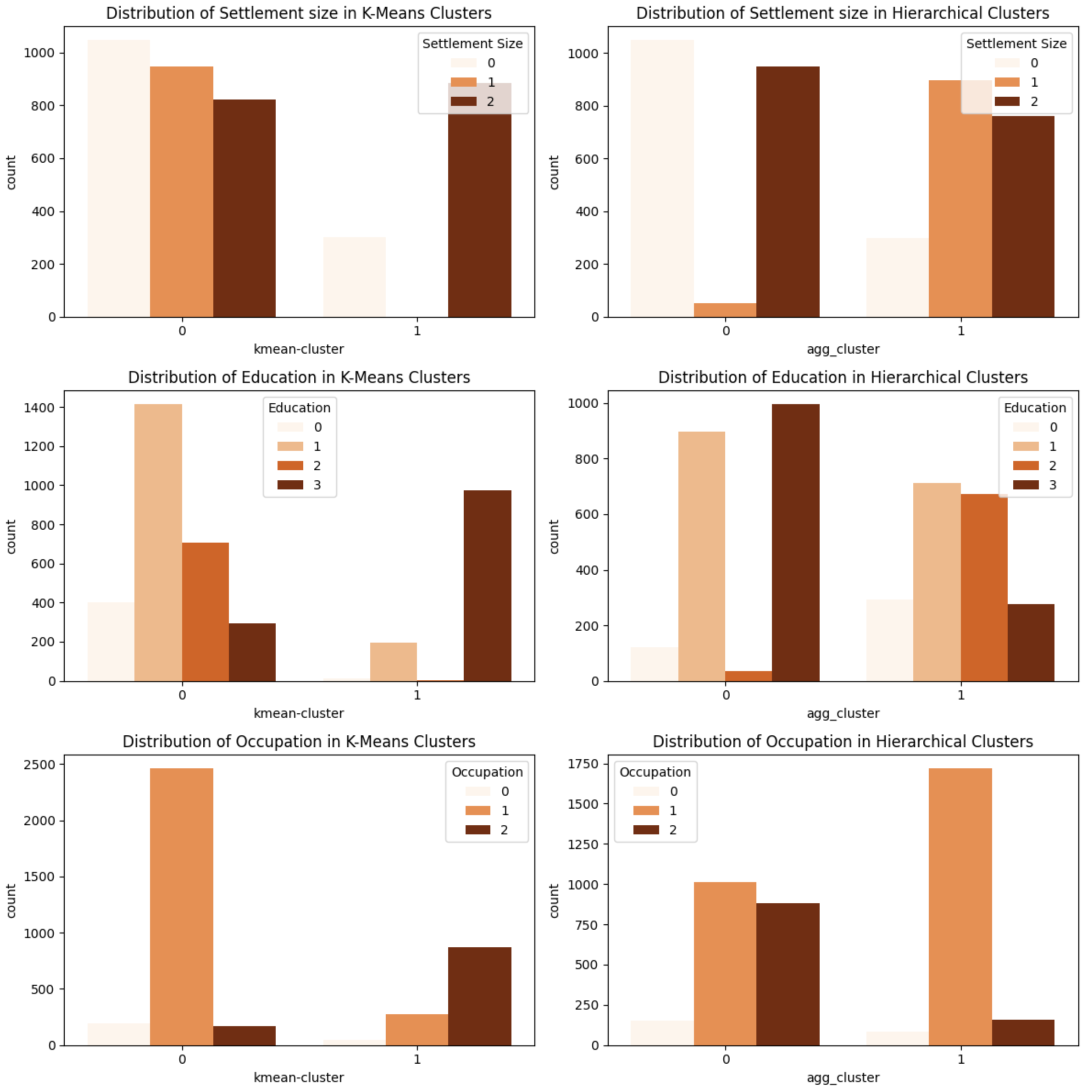
3.3 Comparison between K-Means++ and Agglomerative Clustering:

K-Means++ and Agglomerative Clustering differ significantly in their approach to grouping data, as reflected in the resulting clusters. K-Means++ forms distinct, concentrated clusters, such as Cluster 0, which predominantly includes individuals with lower education and unemployment levels. This method prioritizes clearer separation within the dataset, resulting in clusters that are highly focused on specific feature combinations.

In contrast, Agglomerative Clustering produces more balanced and diverse clusters. For example, Cluster 1 in Agglomerative Clustering includes individuals across a broader range of education levels and settlement sizes. This method captures a more nuanced distribution of features, reflecting greater diversity within each cluster.

While both methods show consistent patterns in categorical features such as gender and marital status, **Agglomerative Clustering better represents the complexity of the dataset by creating clusters that accommodate a mix of feature values. K-Means++, however, is more effective when distinct groupings are needed for targeted analysis**. The choice between these methods depends on whether the focus is on precise segmentation (K-Means++) or a broader representation of feature diversity (Agglomerative Clustering). Both have unique strengths, making them suitable for different segmentation objectives.

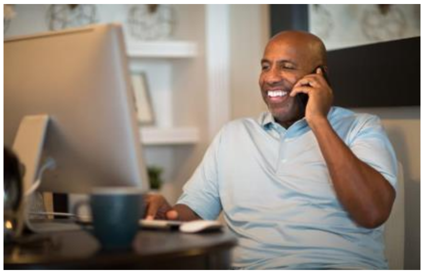




***graph 3.3.1:** comparing categorical feature distributions*

between Kmeans++ and Agglomerative clustering

Part 4: Recommendations

Customer Profile	Suggested Marketing Approach
 Skilled Elder Non-Single Male with high income	• Provide premium food products and gourmet ingredients to cater to their refined preferences and higher income levels. • Organize cooking classes or workshops that highlight gourmet recipes and advanced culinary techniques. • Collaborate with local fitness centers or health clubs to offer exclusive discounts or partnerships that promote healthy living.

 Skilled Middle-age Non-Single Female with moderate income	- Focus on promoting family-oriented products and offer discounts for large purchases of household essentials. - Provide in-store childcare services or play areas to support customers with caregiving responsibilities. - Establish a dedicated section featuring healthy meal ideas and nutritional advice tailored for families.
 Skilled Middle-age Non-Single Male with high income	- Organize events or campaigns focused on quick and convenient food options for busy professionals. - Highlight ready-to-eat meals, pre-packaged foods, and grab-and-go options to suit their time constraints. - Offer loyalty rewards or discounts to regular customers within this demographic.
 Skilled Middle-age Single Female with moderate income	- Organize events or campaigns focused on quick and convenient food options for busy professionals. - Highlight ready-to-eat meals, pre-packaged foods, and grab-and-go options to suit their time constraints. Offer loyalty rewards or discounts to regular customers within this demographic.

For all groups of customers:

- Develop a loyalty program that monitors each customer's purchasing habits, enabling the supermarket to send personalized email campaigns with tailored offers.
- Collaborate with nutrition experts to create a "Fresh" handbook featuring meal recipes, listing specific ingredients, and their original prices available at the supermarket.
- Conduct random online customer surveys to gain insights into shopping preferences and priorities for each customer group.

Part 5: Conclusions

This report highlights customer segmentation analysis conducted for a large supermarket chain using clustering techniques. By analyzing a dataset of 4,000 customers, valuable insights were obtained into demographics, such as sex, marital status, age, income, education, occupation, and settlement size. Through exploratory data analysis, patterns emerged in these variables, revealing strong positive correlations between income and occupation, as well as age and education, while minimal relationships were observed between others.

Using K-means and Hierarchical clustering methods, four distinct customer segments were identified. Both methods provided consistent clustering results, with segments defined by the average values of key variables. Tailored marketing recommendations were proposed for each segment to improve customer engagement and satisfaction through personalized approaches.

In conclusion, leveraging customer segmentation and crafting marketing strategies based on these distinct segments empowers the supermarket chain to boost business growth. By addressing the specific needs and preferences of each group, the company can enhance customer experiences, strengthen satisfaction, and foster long-term loyalty.