5/21/2024

BUSA3020: Advanced Analytics Techniques

Assessment 2 Report

aditya agarwal
46184821

Introduction

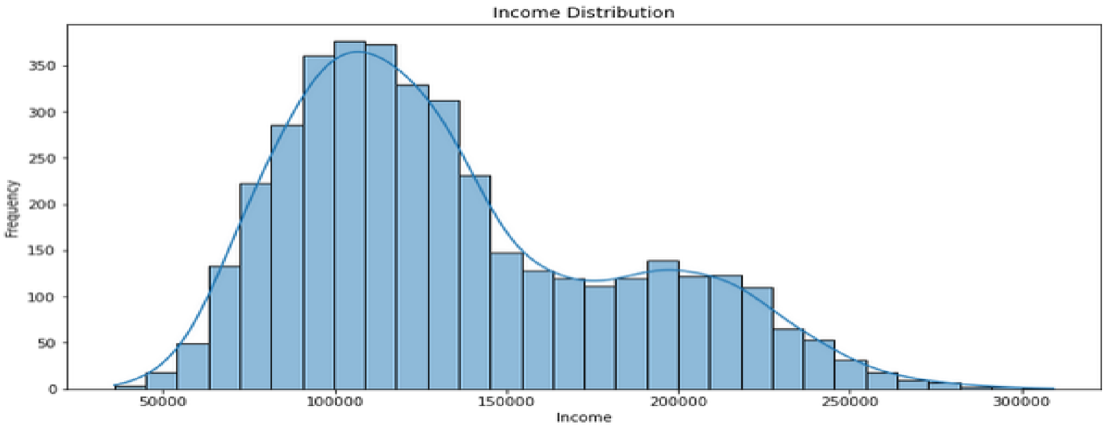
A large supermarket chain seeks to improve its marketing effectiveness by segmenting its customer base. The objective is to identify distinct customer segments using demographic and purchasing data collected through loyalty cards. This segmentation will allow the supermarket chain to create targeted marketing campaigns, optimize product offerings, and enhance the overall customer experience.

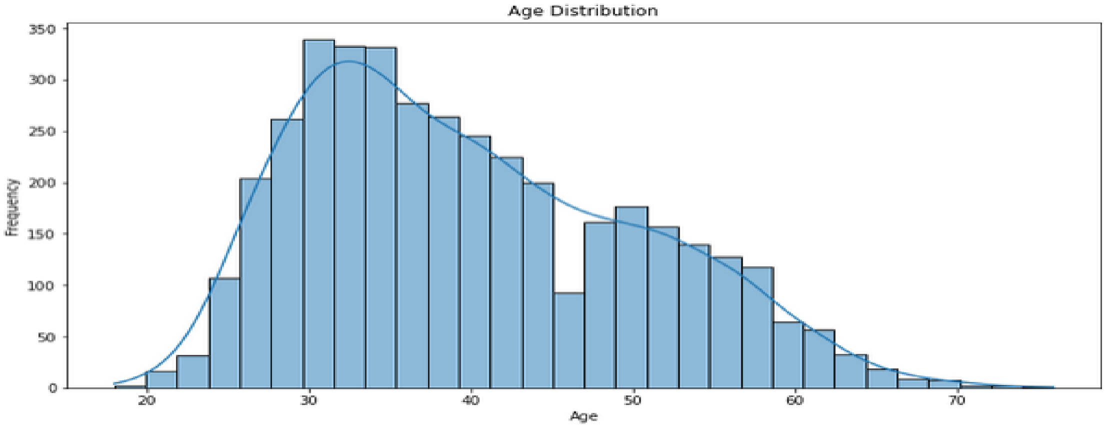
Approach used and Explanation of the dataset:

To address this challenge, a customer segmentation analysis is conducted using a dataset comprising 4,000 customers. The dataset includes variables such as age, gender, annual income, education, marital status, settlement size, and occupation. The analysis proceeded through several key steps:

1. **Data Exploration and Preprocessing:** The dataset was examined to understand its structure, handle missing values, map categorical variables, and visualize data distributions, providing an overview of customer demographics and purchasing behaviours.
2. **Feature Scaling and Encoding:** Numerical features were scaled, and categorical variables were encoded to ensure equal contribution to the clustering process.
3. **Optimal Number of Clusters Identification:** The Elbow Method and Silhouette Analysis identified four as the optimal number of clusters for K-means++ and Agglomerative Clustering.
4. **Cluster Formation:** Clusters were created using K-means++ and Agglomerative Clustering, grouping customers based on similarities.
5. **Cluster Profiling and Interpretation:** Clusters were profiled by calculating average feature values and identifying common attributes, revealing differences in age, income, education, occupation, and settlement size.
6. **Visualization of Clusters:** Principal Component Analysis (PCA) was used to visualize clusters in 2D, aiding in the interpretation and presentation of the results.

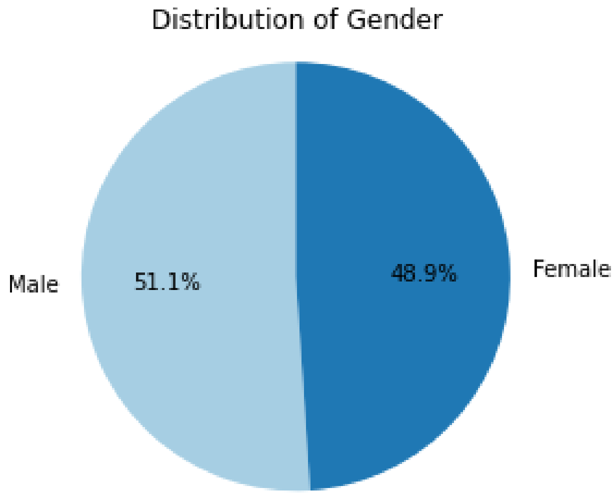
Exploratory Data Analysis



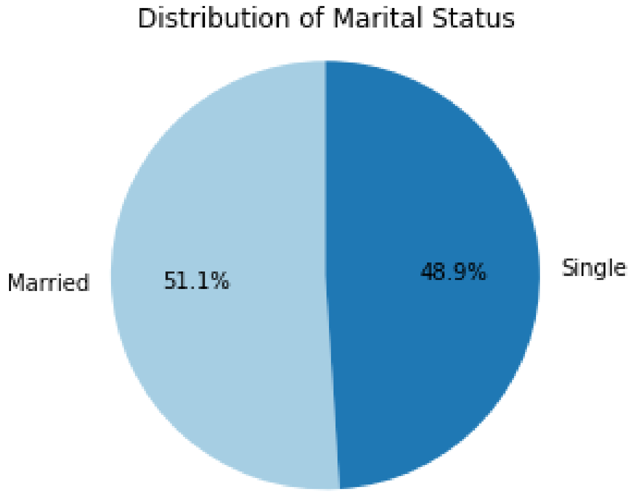


The graphs illustrate the distribution of income and age among our supermarket customers, offering valuable insights for strategic marketing initiatives. The Income Distribution graph indicates that a significant majority of the customers fall within the income range of $50,000 to $150,000, with the highest concentration centred around $100,000. This suggests that the customer base predominantly comprises middle-income earners. Additionally, there is a noticeable, albeit smaller, segment of high-income customers earning above $200,000, highlighting a diverse economic profile within our clientele.

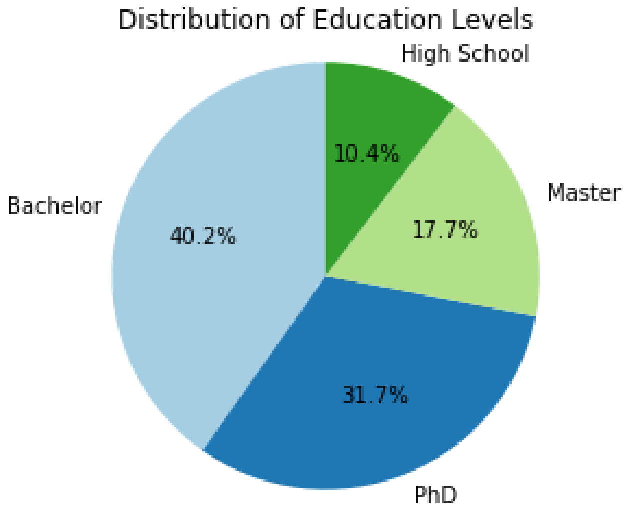
Similarly, the Age Distribution graph reveals that most of the customers are aged between 25 and 45 years, with a pronounced peak around the age of 30. This demographic trend suggests that the customer base is largely composed of younger adults. The frequency of customers declines significantly in the age groups above 50, indicating a smaller proportion of older customers.



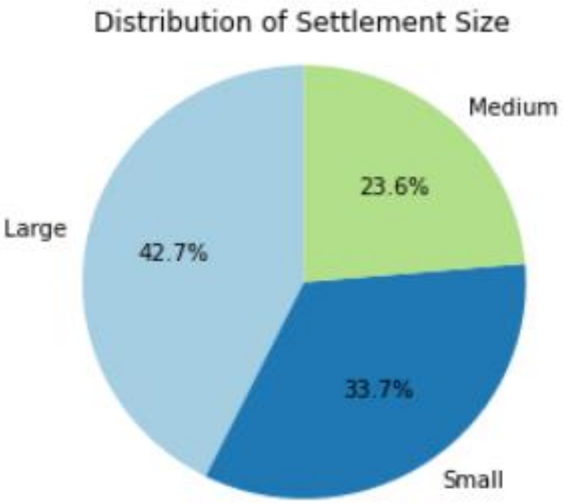
The first pie chart, **Distribution of Gender**, shows a nearly equal split between male and female customers, with males constituting 51.1% and females making up 48.9% of the total customer base.



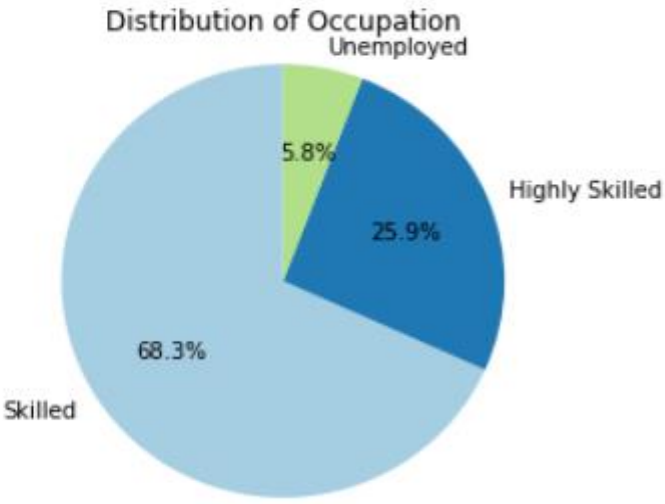
The second pie chart, **Distribution of Marital Status**, reveals a similar near-equal distribution, with 51.1% of customers being married and 48.9% being single. This information indicates that the customer base is almost evenly divided between married and single individuals.



The third pie chart, **Distribution of Education Levels**, provides insights into the educational background of our customers. The majority hold a bachelor's degree (40.2%), followed by those with a PhD (31.7%). Customers with a master's degree account for 17.7%, while those with only a High School education represent 10.4% of the customer base. This data indicates that a significant portion of the customers are highly educated.

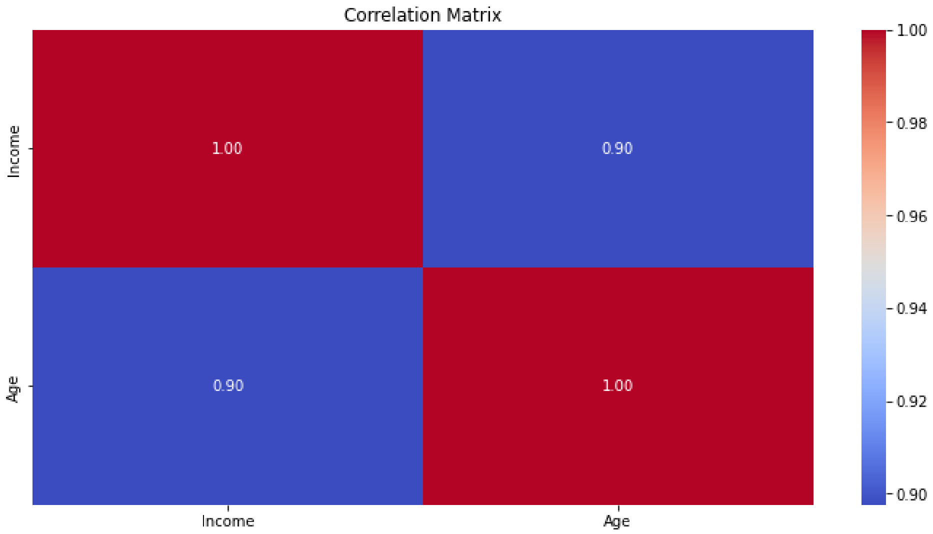


The first pie chart, **Distribution of Settlement Size**, shows that the largest segment of the customers, 42.7%, reside in large settlements. This is followed by 33.7% of customers living in small settlements and 23.6% in medium-sized settlements.



The second pie chart, **Distribution of Occupation**, reveals that a significant majority of our customers, 68.3%, are skilled workers. Highly skilled professionals make up 25.9% of the customer base, while only 5.8% are unemployed. This occupational breakdown highlights that the customer base predominantly consists of working professionals.

Overall, these distributions and charts help us understand the demographic and socio-economic profile of our customer base, enabling us to design more effective and targeted marketing strategies that cater to the diverse needs of our customers.

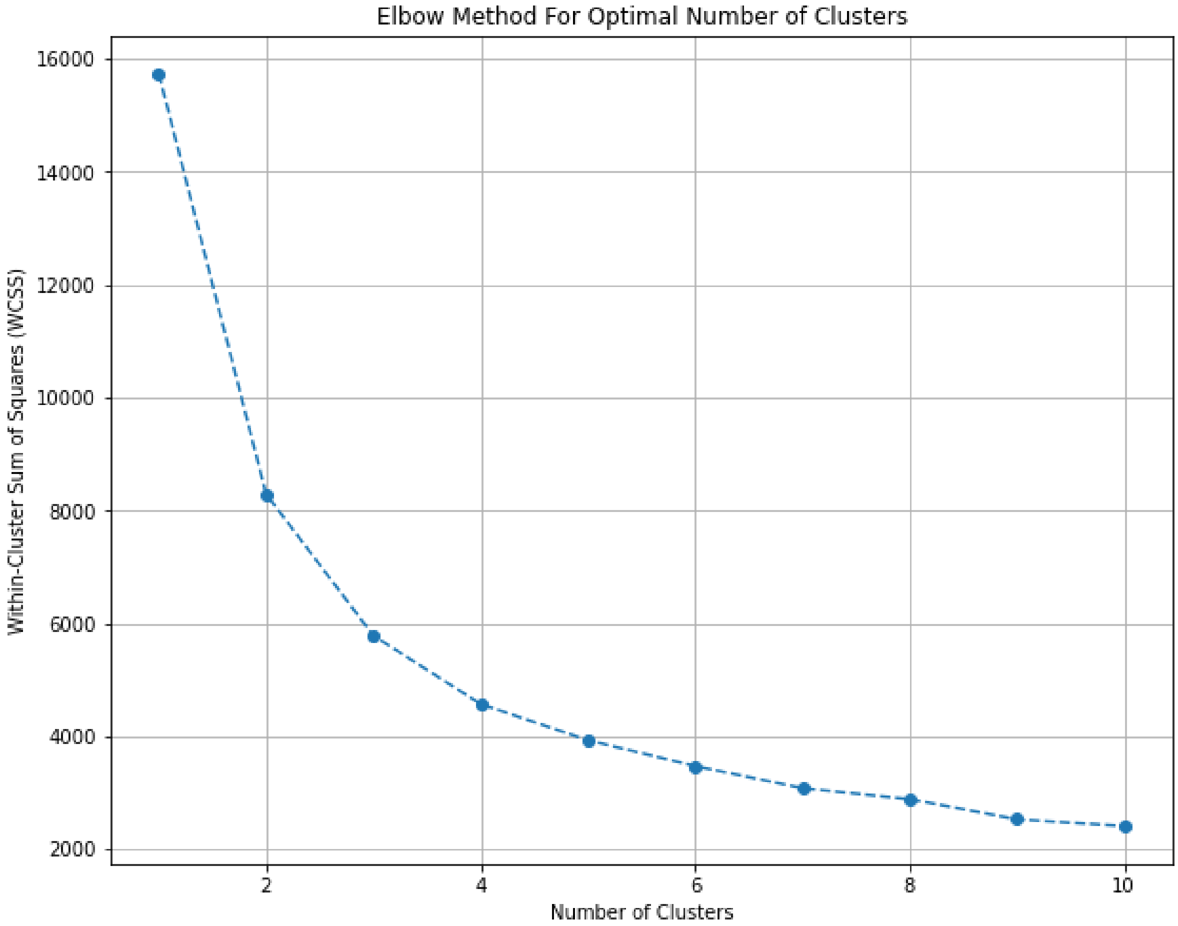


In this correlation matrix, we observe a strong positive correlation of 0.90 between income and age.

This high correlation coefficient suggests that, generally, as age increases, income also tends to increase among our customers. This relationship is important for understanding the economic behaviour of different age groups within our customer base.

Customer Segmentation

- **The Elbow method** used to find optimal number of customer segments/clusters.



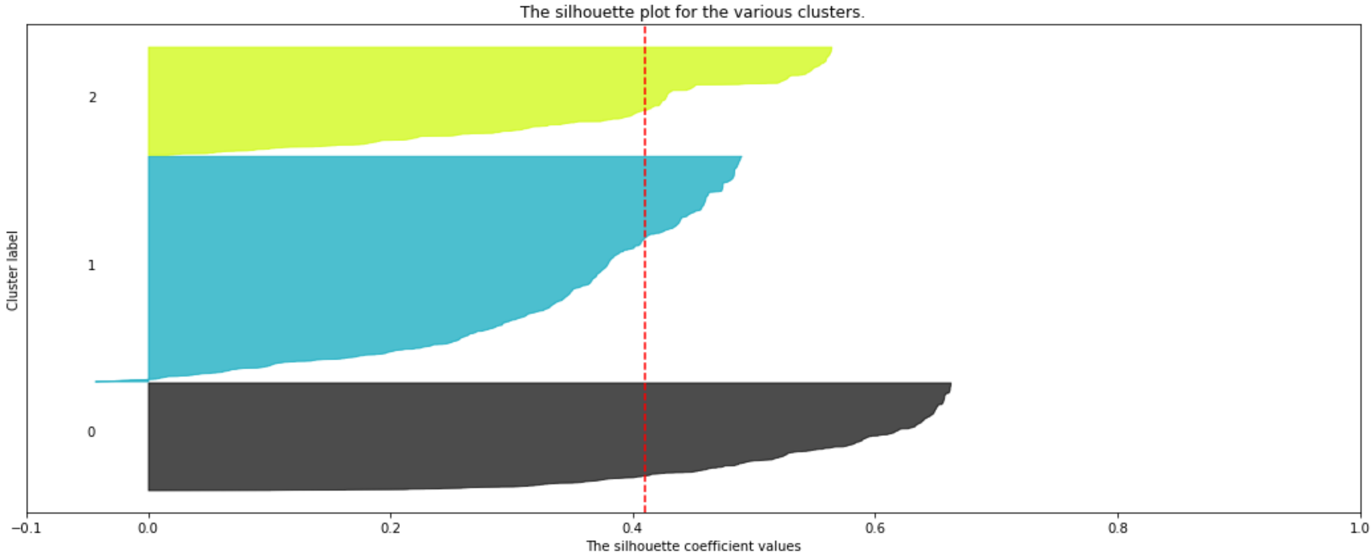
The plot illustrates the Elbow Method, which is used to determine the optimal number of clusters for k-means clustering.

The x-axis represents the number of clusters, while the y-axis shows the within-cluster sum of squares (WCSS), a measure of the variance within each cluster.

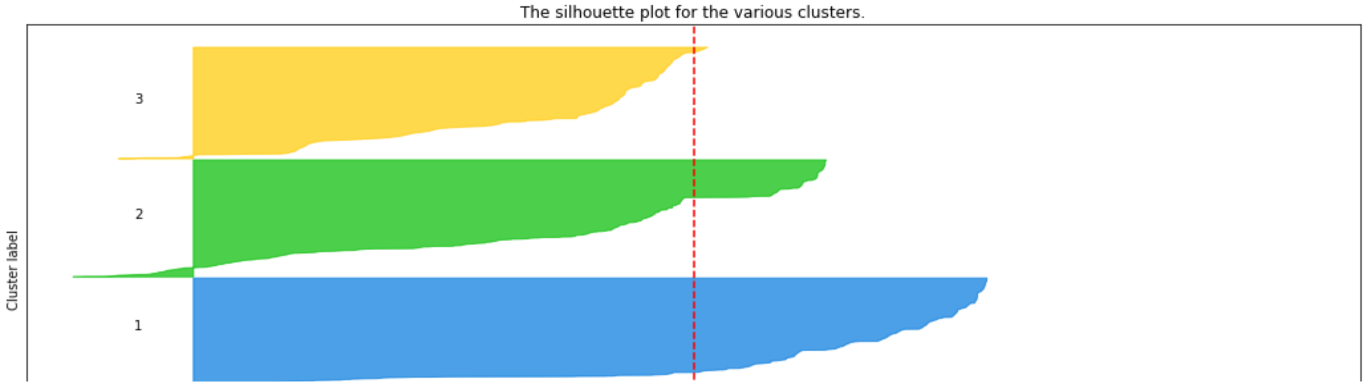
As the number of clusters increases, the WCSS decreases, indicating that the clusters are becoming more compact and better defined. Initially, there is a steep decline in WCSS, which starts to level off as the number of clusters increases. The "elbow" point, where the rate of decrease sharply slows, is typically considered the optimal number of clusters. In this plot, the elbow appears to be at around 3 to 4 clusters. This suggests that using 3 or 4 clusters would likely provide a good balance between compactness and simplicity, capturing the underlying structure of the data without overcomplicating the model.

- **Silhouette plots** used to find optimal number of customer segments/clusters.

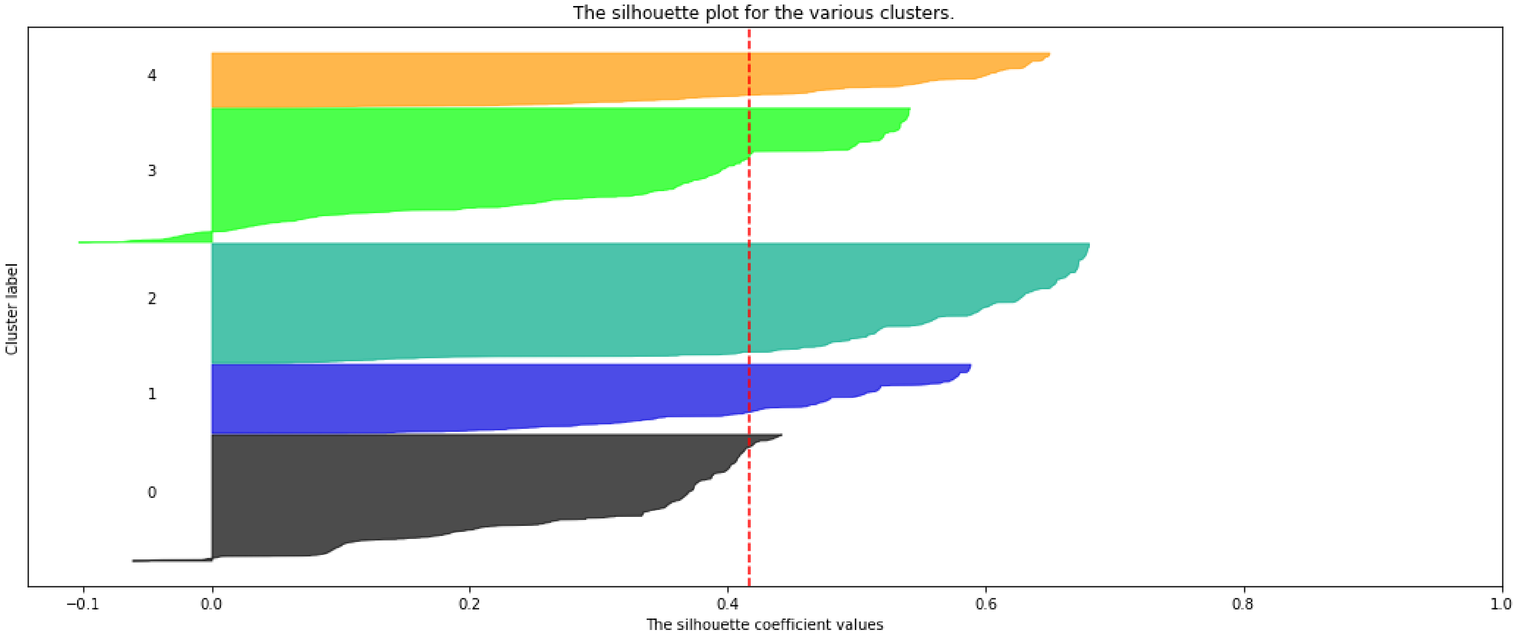
```
For n_clusters = 3 The average silhouette_score is : 0.41053927869916423
```



```
For n_clusters = 4 The average silhouette_score is : 0.42964047875343225
```



```
For n_clusters = 5 The average silhouette_score is : 0.4165513868453014
```



```
Best cluster number: 4 with a silhouette score of 0.42964047875343225
```

The silhouette plots visually represent the quality of clustering for different numbers of clusters (3, 4, and 5) using the silhouette score as a metric. The silhouette score measures how similar a point is to its own cluster compared to other clusters, with values ranging from -1 to 1. A higher silhouette score indicates better-defined clusters.

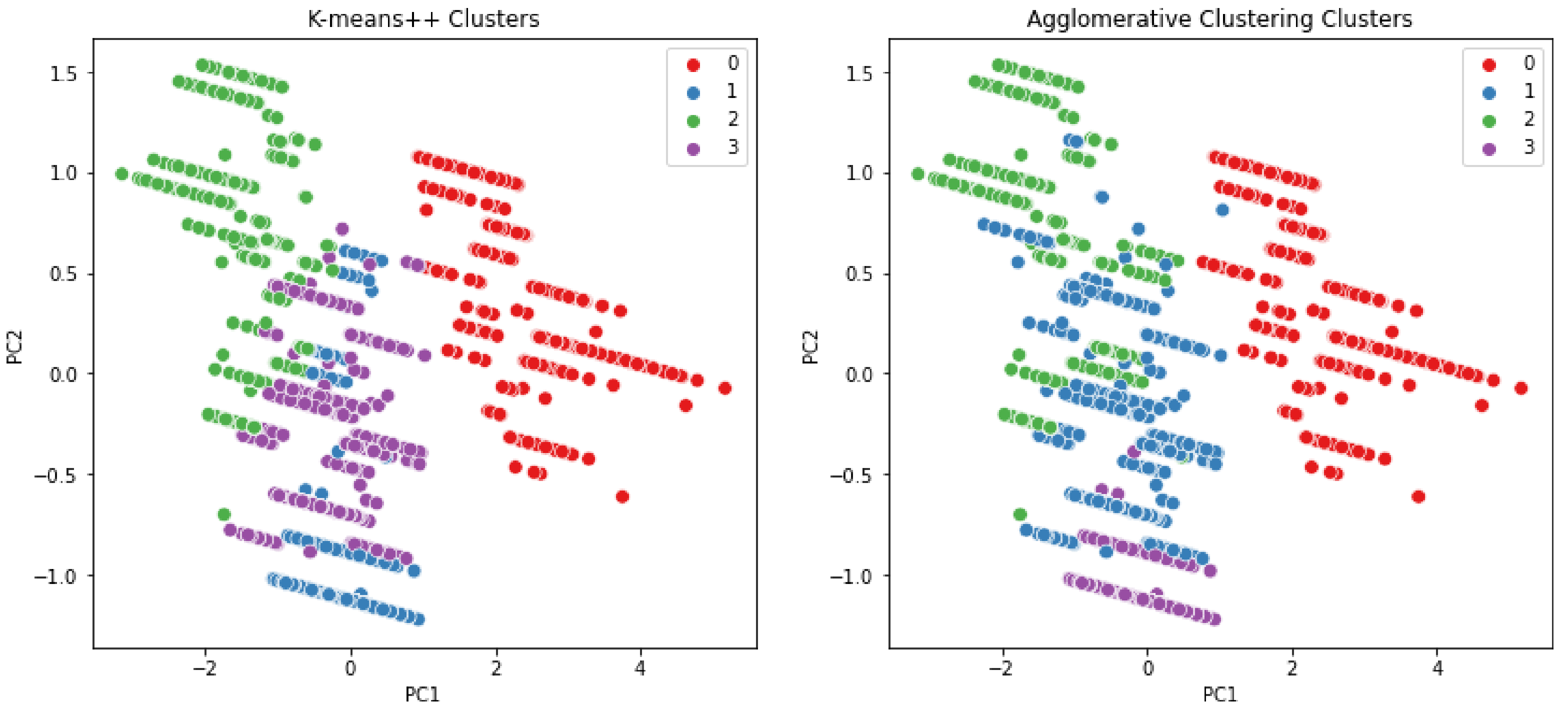
In the first plot, with **3 clusters**, the average silhouette score is 0.4105. Each cluster is shown in a different colour, and the width of each plot indicates the silhouette coefficient for points within that cluster.

The second plot, with **4 clusters**, has a slightly higher average silhouette score of 0.4296. This indicates better-defined clusters compared to the 3-cluster solution. The silhouette scores for each cluster are more consistent, suggesting that 4 clusters might be a better representation of the data structure.

The third plot, with **5 clusters**, shows an average silhouette score of 0.4165. While this is an improvement over the 3-cluster solution, it is slightly lower than the 4-cluster solution. The 5-cluster plot demonstrates more variation in silhouette scores within clusters, indicating that adding more clusters does not necessarily improve the clustering quality significantly.

Overall, the 4-cluster solution appears to be the most optimal, as it has the highest average silhouette score of 0.4296, suggesting better-defined and more distinct clusters compared to the other configurations. This indicates that partitioning the data into 4 clusters provides a balanced and meaningful segmentation of the customer data.

- **Comparing** the customer segments/clusters of ***kMeans++*** and ***agglomerative*** methods



The scatter plots compare the results of two clustering methods: K-means++ and Agglomerative Clustering, both visualized using Principal Component Analysis (PCA) to reduce the data to two dimensions. Each point represents a customer, and the colours indicate the cluster to which the customer belongs.

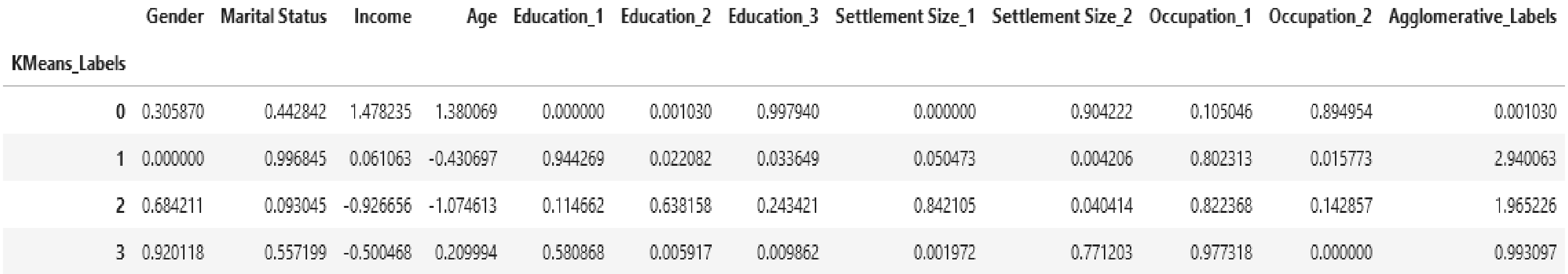
In the **K-means++ Clusters** plot (left), four distinct clusters are clearly separated, indicating that the K-means++ algorithm effectively grouped customers based on their similarities. The clusters are color-coded as red (Cluster 0), green (Cluster 1), blue (Cluster 2), and purple (Cluster 3). The distinct separation of clusters suggests that the algorithm has successfully identified meaningful patterns in the data.

The **Agglomerative Clustering Clusters** plot (right) also shows four clusters, using the same colour scheme. While the clusters are similarly well-separated, there are slight differences in the cluster boundaries and the distribution of points compared to the K-means++ results. This indicates that while both methods are effective, they may identify slightly different groupings based on the data.

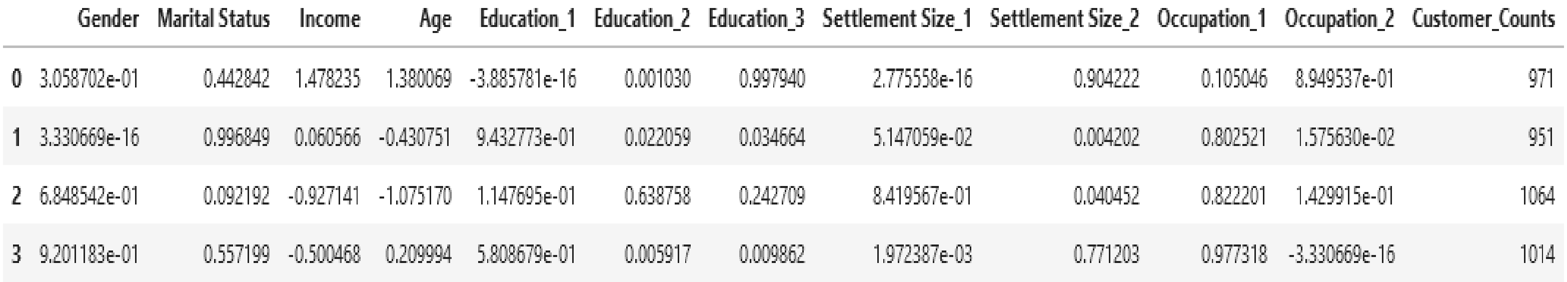
These visualizations confirm that both clustering methods provide valuable insights into customer segmentation.

- *Provide tables presenting the cluster centres (all 7 variables) and customer counts for both clustering techniques.*
- *Interpret each of the identified clusters (for both clustering techniques) in terms of customer attributes, i.e. profile the clients in each cluster.*

#K-means++ Clustering



	Gender	Marital Status	Income	Age	Education_1	Education_2	Education_3	Settlement Size_1	Settlement Size_2	Occupation_1	Occupation_2	Agglomerative_Labels
KMeans_Labels												
0	0.305870	0.442842	1.478235	1.380069	0.000000	0.001030	0.997940	0.000000	0.904222	0.105046	0.894954	0.001030
1	0.000000	0.996845	0.061063	-0.430697	0.944269	0.022082	0.033649	0.050473	0.004206	0.802313	0.015773	2.940063
2	0.684211	0.093045	-0.926656	-1.074613	0.114662	0.638158	0.243421	0.842105	0.040414	0.822368	0.142857	1.965226
3	0.920118	0.557199	-0.500468	0.209994	0.580868	0.005917	0.009862	0.001972	0.771203	0.977318	0.000000	0.993097



	Gender	Marital Status	Income	Age	Education_1	Education_2	Education_3	Settlement Size_1	Settlement Size_2	Occupation_1	Occupation_2	Customer_Counts
0	3.058702e-01	0.442842	1.478235	1.380069	-3.885781e-16	0.001030	0.997940	2.775558e-16	0.904222	0.105046	8.949537e-01	971
1	3.330669e-16	0.996849	0.060566	-0.430751	9.432773e-01	0.022059	0.034664	5.147059e-02	0.004202	0.802521	1.575630e-02	951
2	6.848542e-01	0.092192	-0.927141	-1.075170	1.147695e-01	0.638758	0.242709	8.419567e-01	0.040452	0.822201	1.429915e-01	1064
3	9.201183e-01	0.557199	-0.500468	0.209994	5.808679e-01	0.005917	0.009862	1.972387e-03	0.771203	0.977318	-3.330669e-16	1014

Cluster 0:

- Gender: Majority male (30.6%)

- Marital Status: Mixed (44.3% married)

- Income: High income

- Age: Older customers

- Education: Mostly graduate school (99.8%)

- Settlement Size: Predominantly big cities (90.4%)

- Occupation: Majority highly qualified (89.5%)

Profile: High-income, older male professionals, mostly well-educated, living in big cities. These customers are likely to value premium products and services and have significant purchasing power.

Cluster 1:

- Gender: All male
- Marital Status: Mostly married (99.7%)
- Income: Moderate income
- Age: Younger customers
- Education: High school (94.4%)
- Settlement Size: Small and mid-sized cities
- Occupation: Skilled employees (80.2%)

Profile: Younger, married males with moderate income and high school education, living in smaller cities. These customers might be focused on family and household needs.

Cluster 2:

- Gender: Majority female (68.4%)
- Marital Status: Mostly single (9.3%)
- Income: Low income
- Age: Younger customers
- Education: Mixed (graduate school 24.3%)
- Settlement Size: Predominantly small cities
- Occupation: Unemployed/unskilled

Profile: Younger, low-income females, mostly single, living in small cities. These customers may prioritize affordability.

Cluster 3:

- Gender: Predominantly female (92.0%)

- Marital Status: Mixed (55.7% married)

- Income: Low income

- Age: Middle-aged

- Education: High school(58.1%)

- Settlement Size: Predominantly big cities (77.1%)

- Occupation: Skilled employees (97.7%)

Profile: Middle-aged, predominantly female customers with low income and high school education, living in big cities.

K-means++ Clustering Results:

- High-income older males: Cluster 0 could be targeted with premium products.

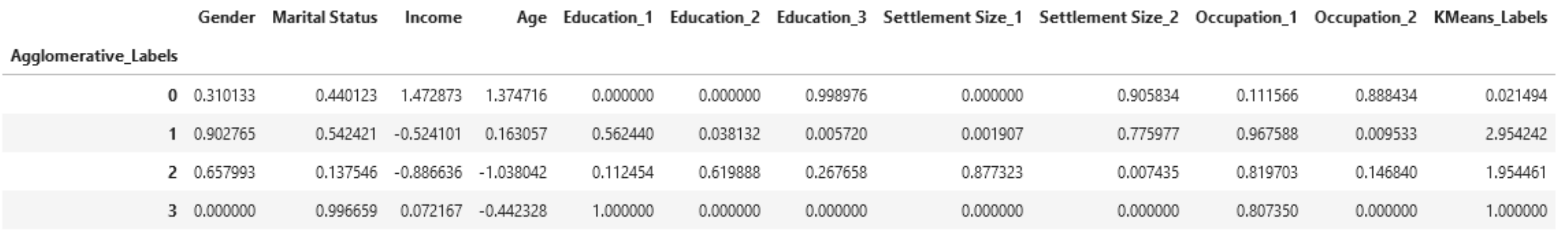
- Married younger males: Cluster 1 might benefit from family-oriented promotions.

- Low-income younger females: Cluster 2 may need budget-friendly options.

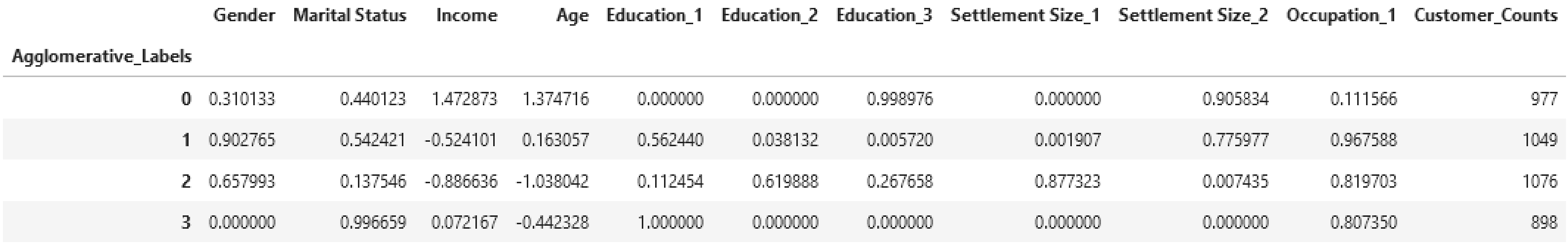
- Middle-income middle-aged females: Cluster 3 might appreciate community engagement initiatives.

- *Provide tables presenting the cluster centres (all 7 variables) and customer counts for both clustering techniques.*
- *Interpret each of the identified clusters (for both clustering techniques) in terms of customer attributes, i.e. profile the clients in each cluster.*

# Agglomerative Clustering



	Gender	Marital Status	Income	Age	Education_1	Education_2	Education_3	Settlement Size_1	Settlement Size_2	Occupation_1	Occupation_2	KMeans_Labels
Agglomerative_Labels												
0	0.310133	0.440123	1.472873	1.374716	0.000000	0.000000	0.998976	0.000000	0.905834	0.111566	0.888434	0.021494
1	0.902765	0.542421	-0.524101	0.163057	0.562440	0.038132	0.005720	0.001907	0.775977	0.967588	0.009533	2.954242
2	0.657993	0.137546	-0.886636	-1.038042	0.112454	0.619888	0.267658	0.877323	0.007435	0.819703	0.146840	1.954461
3	0.000000	0.996659	0.072167	-0.442328	1.000000	0.000000	0.000000	0.000000	0.000000	0.807350	0.000000	1.000000



	Gender	Marital Status	Income	Age	Education_1	Education_2	Education_3	Settlement Size_1	Settlement Size_2	Occupation_1	Customer_Counts
Agglomerative_Labels											
0	0.310133	0.440123	1.472873	1.374716	0.000000	0.000000	0.998976	0.000000	0.905834	0.111566	977
1	0.902765	0.542421	-0.524101	0.163057	0.562440	0.038132	0.005720	0.001907	0.775977	0.967588	1049
2	0.657993	0.137546	-0.886636	-1.038042	0.112454	0.619888	0.267658	0.877323	0.007435	0.819703	1076
3	0.000000	0.996659	0.072167	-0.442328	1.000000	0.000000	0.000000	0.000000	0.000000	0.807350	898

Cluster 0:

- Gender: Majority male (31.0%)

- Marital Status: Mixed (44.0% married)

- Income: High income

- Age: Older customers

- Education: Mostly graduate school (99.9%)

- Settlement Size: Predominantly big cities (90.6%)

- Occupation: Majority highly qualified (88.8%)

Profile: High-income, older male professionals, well-educated, living in big cities. These customers value premium products and services.

Cluster 1:

- Gender: Predominantly female (90.3%)

- Marital Status: Mixed (54.2% married)

- Income: Low income

- Age: Middle-aged

- Education: High school (56.2%)

- Settlement Size: Predominantly big cities (77.6%)

- Occupation: Skilled employees (96.8%)

Profile: Middle-aged, predominantly female customers with low income and high school education, living in big cities.

Cluster 2:

- Gender: Majority female (65.8%)

- Marital Status: Mostly single (13.8%)

- Income: Low income

- Age: Younger customers

- Education: Mixed (graduate school 26.8%)

- Settlement Size: Predominantly small cities (87.7%)

- Occupation: Unemployed/unskilled (81.9%)

Profile: Younger, low-income females, mostly single, living in small cities. These customers prioritize affordability.

Cluster 3:

- Gender: All male

- Marital Status: Mostly married (99.7%)

- Income: Moderate income
- Age: Younger customers
- Education: High school (100%)
- Settlement Size: Small and mid-sized cities
- Occupation: Skilled employees (80.7%)

Profile: Younger, married males with moderate income and high school education, living in smaller cities. These customers might be focused on family and household needs.

Agglomerative Clustering Results:

- High-income older males: Cluster 0 aligns closely with K-means++ Cluster 0.
- Middle-income middle-aged females: Cluster 1 is similar to K-means++ Cluster 3 but with slightly different demographic.
- Low-income younger females: Cluster 2 is consistent with K-means++ Cluster 2.
- Married younger males: Cluster 3 corresponds well to K-means++ Cluster 1.

Recommendations

The K-means++ clustering method identified four distinct customer segments, each with unique characteristics. By understanding these segments, targeted marketing strategies can be created to effectively engage and satisfy each group.

Cluster 0 is characterized by higher income and older age. Customers in this segment are predominantly from larger settlements and are engaged in skilled occupations. Marketing strategies for this group should focus on premium products and services that emphasize quality and exclusivity. Personalized offers such as loyalty programs, premium memberships, and high-end product recommendations can appeal to their higher purchasing power. Additionally, highlighting convenience and luxury in marketing messages can resonate well with this segment.

Cluster 1 comprises mostly married individuals with moderate to lower incomes and younger age. This group has a significant representation of bachelor's degree holders. Marketing efforts should focus on value-for-money products and family-oriented promotions. Offering discounts on essential household items, family bundles, and affordable yet stylish products can attract this segment. Communication should emphasize practicality, savings, and benefits for the whole family, leveraging social media platforms and family-centric events to enhance engagement.

Cluster 2 consists of younger customers with lower incomes and diverse education levels. This segment is likely to be more price-sensitive and tech-savvy. Marketing strategies should include promotional offers, discounts, and seasonal sales to attract budget-conscious consumers. Leveraging digital marketing techniques such as targeted social media ads, influencer partnerships, and email marketing campaigns can effectively reach this tech-oriented group. Additionally, offering instalment payment options or loyalty rewards for frequent purchases can increase their purchasing frequency and brand loyalty.

Cluster 3 is predominantly female, with a mix of educated individuals holding bachelor's and master's degrees. This segment has an average income and age profile. Marketing strategies should focus on products and services that cater to women's interests and needs, such as fashion, beauty, and wellness products. Personalized marketing through email campaigns, tailored product recommendations, and content marketing that highlights trends and lifestyle tips can engage this group. Collaborating with female influencers and bloggers can also enhance brand visibility and credibility among this segment.

By tailoring marketing strategies to the specific characteristics and preferences of each customer segment identified by the K-means++ method, the company can improve customer satisfaction, enhance engagement, and drive sales growth. This targeted approach ensures that the marketing efforts are relevant and resonate with each unique group, ultimately fostering stronger customer relationships and loyalty.

Conclusion

In this report, a comprehensive customer segmentation analysis was conducted for a large supermarket chain using a dataset of 4,000 customers collected through loyalty cards. The analysis involved an initial exploration of the dataset, including checking for missing values, understanding the data structure, and visualizing distributions of various features. Categorical variables were mapped to meaningful labels, and numerical features were standardized to facilitate effective clustering.

Two clustering methods, K-means++ and Agglomerative Clustering, were employed to identify distinct customer segments. The optimal number of clusters was determined using the Elbow Method and Silhouette Analysis, with four clusters emerging as the most suitable solution. These clusters were then visualized using Principal Component Analysis (PCA) to provide a clear understanding of the groupings.

The characteristics of each customer segment were summarized, highlighting key features such as gender, marital status, income, age, education, settlement size, and occupation. Based on these insights, targeted marketing strategies were suggested for each segment, emphasizing personalized and relevant approaches to meet the unique needs and preferences of the identified groups.

Overall, this report demonstrates the application of data-driven techniques to achieve effective customer segmentation, providing actionable insights that can enhance marketing efforts and drive business growth for the supermarket chain.