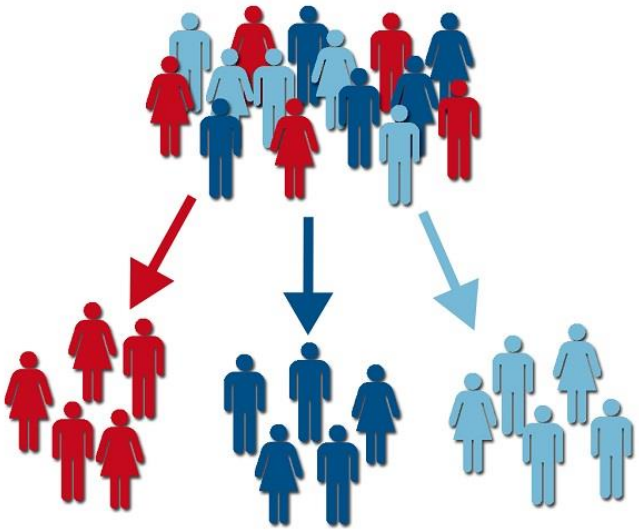
Customer Segmentation Analysis for Developing Tailored Marketing Strategies in a Supermarket Chain



- **Student Name:** Umut Demirhan
- **Student Number:** 46739106
- **Lecturer's Name:** George Milunovich

- **Unit Code:** BUSA8001
- **Unit Name:** Applied Predictive Analytics
- **Assignment Title:** Programming Task 3 - Report
- **Due Date:** 19/May/2023
- **Date of Submission:** 18/May/2023.
- **Word Length (excluding the list of references):** 995 words.

Contents

Introduction

In the fiercely competitive world of retail, understanding the nuances of customer behaviour is a game-changer. It enables businesses to create targeted strategies that resonate with distinct customer groups, leading to enhanced customer satisfaction and improved business performance.

This report aims to uncover insights from our supermarket chain's customer data collected through loyalty cards. Utilizing advanced analytics and machine learning techniques, I will employ customer segmentation to group customers based on common characteristics. These segments will help us understand our customer base and enable personalized marketing. Then, we'll explore our customer data, segment it, interpret the results, and conclude with tailored marketing strategies for each segment.

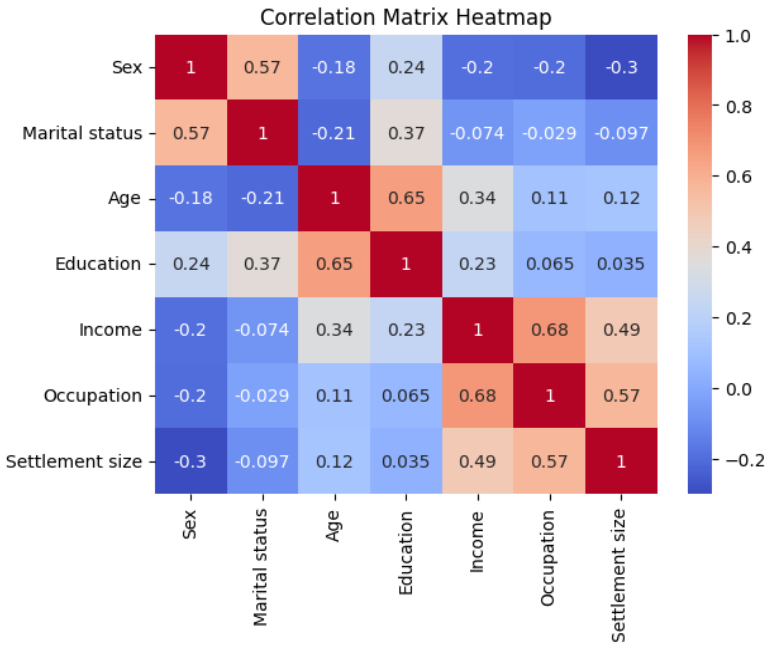
Exploratory Data Analysis

Our dataset, acquired from the supermarket loyalty cards, contains information about 2,000 customers. The dataset includes eight columns, encompassing the ID, Sex, Marital status, Age, Education, Income, Occupation, and Settlement size of each customer.

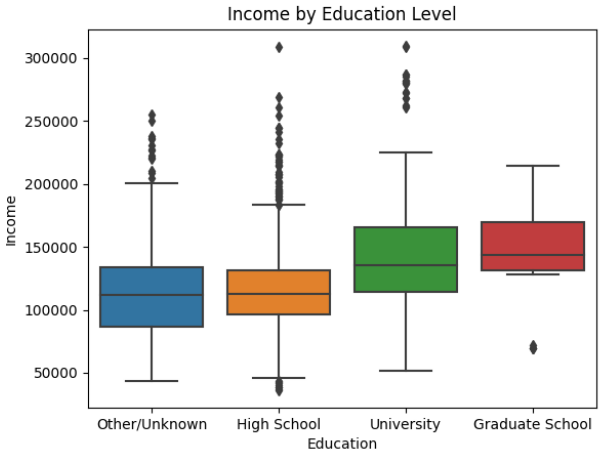
Upon analysing the summary statistics of the data, the following key observations emerge:

- Gender is slightly skewed towards males (54.3%).
- Marital status is almost evenly split between singles and non-singles.
- Customers are predominantly young to middle-aged, ranging from 18 to 76 with a median age of 33 years.
- Most customers have a high school education (69%).
- Average annual income is approximately $120,954, indicating a relatively affluent customer base.
- Most customers are skilled workers, officials, or hold management positions.
- Customers are fairly distributed across small, mid-sized, and large cities.

In terms of correlations, Income and Occupation are correlated, as are Age and Education. For example, the heatmap below suggests higher-income customers tend to have more skilled occupations, and older customers tend to have higher levels of education.



I used visuals such as bar charts, histograms, and boxplots during the initial data exploration. These showed that income varies a lot among high school graduates. Interestingly, those with graduate degrees rarely earn more than $200,000 a year, with some earning just above $50,000.



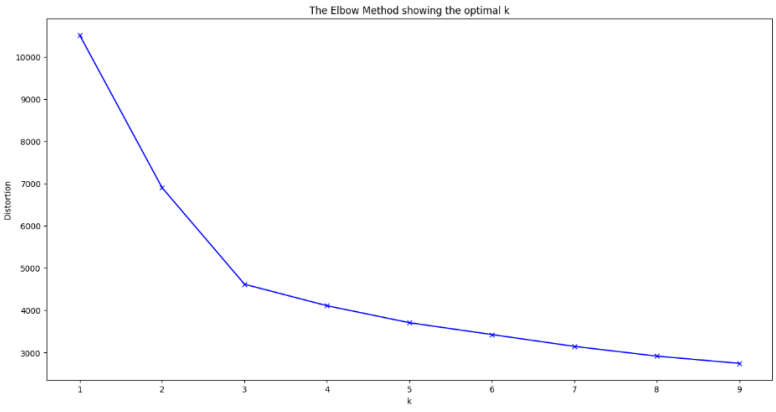
This dataset has further potential for exploration, as detailed in the Jupyter file for additional explanatory analysis. However, the primary objective of this report is to focus on customer segmentation.

Customer Segmentation

To understand our customers better, I employed clustering analysis, a powerful unsupervised machine learning technique. This method groups customers into unique segments based on their shared characteristics.

I used two clustering techniques: K-Means and Hierarchical Clustering, to identify customer segments. K-Means efficiently finds groupings, while Hierarchical Clustering visually displays the clustering process and doesn't need a predefined cluster number. For accurate results, I standardized the continuous data and kept categorical data unchanged before applying these methods.

I employed the Elbow Method and Silhouette Analysis for both K-Means and Hierarchical Clustering to determine the optimal number of clusters. These techniques provide a heuristic to identify the point where adding more clusters does not significantly improve the intra-cluster homogeneity.



As in the elbow method, the optimal number of clusters is at the "elbow" of the plot, where the WCSS starts to decrease more slowly; Therefore, I settled on three clusters. When interpreting these clusters, I found each representing a unique customer profile. For instance, one cluster may represent young, single individuals with low income, while another may represent older, married, high-income individuals.

Cluster ID	Sex (% Female)	Marital Status (% Not Single)	Avg. Age	Education (Mode)	Avg. Income ($)	Occupation (Mode)	Settlement Size (Mode)
0	25.3	40.4	32	University	137358	Skilled Employees/Officials	Big Cities
1	36.1	47.7	55	Graduate School	154392	Skilled Employees/Officials	Mid-sized to Big Cities
2	63.8	56.9	32	University	97063	Unemployed/Unskilled	Small Cities

The table above shows the outcome of K-Means clustering, presenting 'Sex' and 'Marital status' as percentages for easy comprehension, while 'Age' and 'Income' are displayed as averages. 'Education', 'Occupation', and 'Settlement Size' are represented by their most common values, known as the mode. This analysis has successfully identified three distinct customer segments:

- "Young Professionals in Big Cities" (Cluster 0) consists of non-single, university-educated males in their 30s, with incomes around $137,000.
- "Senior, Educated Individuals" (Cluster 1) are predominantly older males with higher education, earning around $154,000.
- "Young, Lower-Income Singles" (Cluster 2) are primarily single females in their 30s earning about $97,000.

Cluster	Sex (% Female)	Marital Status (% Married)	Avg. Age	Education (Mode)	Avg. Income ($)	Occupation (Mode)	Settlement Size (Mode)
0	39.578454	51.053864	52.604215	university	153663.459	skilled employee	small city
1	62.785637	54.624592	30.899891	high school	97155.6703	unemployed	small city
2	25.688073	41.743119	32.047401	high school	133040.489	skilled employee	big city

The table above the outcome of Hierarchical clustering. Both K-Means and Hierarchical clustering identified three distinct customer groups, but with minor differences. In general, we can see a similar clustering. Cluster 0 of K-Means is almost same with Cluster 2 of Hierarchical clustering expect the education. Similarly, Cluster 1 of K-Means with Cluster 0 of Hierarchical clustering, and Cluster 2 of K-Means with Cluster 1 of Hierarchical Clustering. They differ on small fractions in each feature. the results are consistent. Therefore, I will provide recommendations based on K-Means clustering results.

Evaluating these models using the silhouette score, a measure of cluster cohesion and separation, I found that both clustering methods performed similarly. However, the K-Means algorithm delivered a slightly superior score of 0.4223 compared to Hierarchical Clustering's 0.4107. Thus, K-Means demonstrated a slight edge in this context.

Both results are consistent. Therefore, I will provide recommendations based on K-Means clustering results.

Recommendations

Considering our segmentation analysis, I propose the following strategic marketing approaches for each cluster:

1. Young Urban Professionals: They present an opportunity for targeted marketing of premium products. High earners in urban areas would respond well to promotions for premium products and services, preferably through digital platforms. Trend-based, high-quality campaigns should be the focus.

2. Mature, High-Income Suburbanites: Given their higher income and potential for frequent purchases, campaigns focused on quality and premium service would resonate well. Loyalty programs can be considered for ensuring repeat business.

3. Young, Single, Price-Sensitive: For this price-conscious group, marketing should emphasize value. Promotions like sales, discounts, and partnerships with affordable brands would be effective. Highlighting product durability and flexible payment options could also appeal.

Continuous evaluation of these strategies is vital to meeting evolving customer needs and preferences.

Conclusion

Our analysis segmented customers into three unique groups using exploratory data analysis and clustering techniques. These groups are young urban professionals; mature, high-income suburbanites; young, single, price-sensitive individuals. Understanding these groups allows tailored marketing strategies that cater to each segment's unique needs, enhancing customer satisfaction and loyalty. As a result, we can refine our strategies, improve engagement, and foster business growth. In addition, this data-driven approach ensures that I stay responsive to our customers, making our supermarket chain their preferred choice.