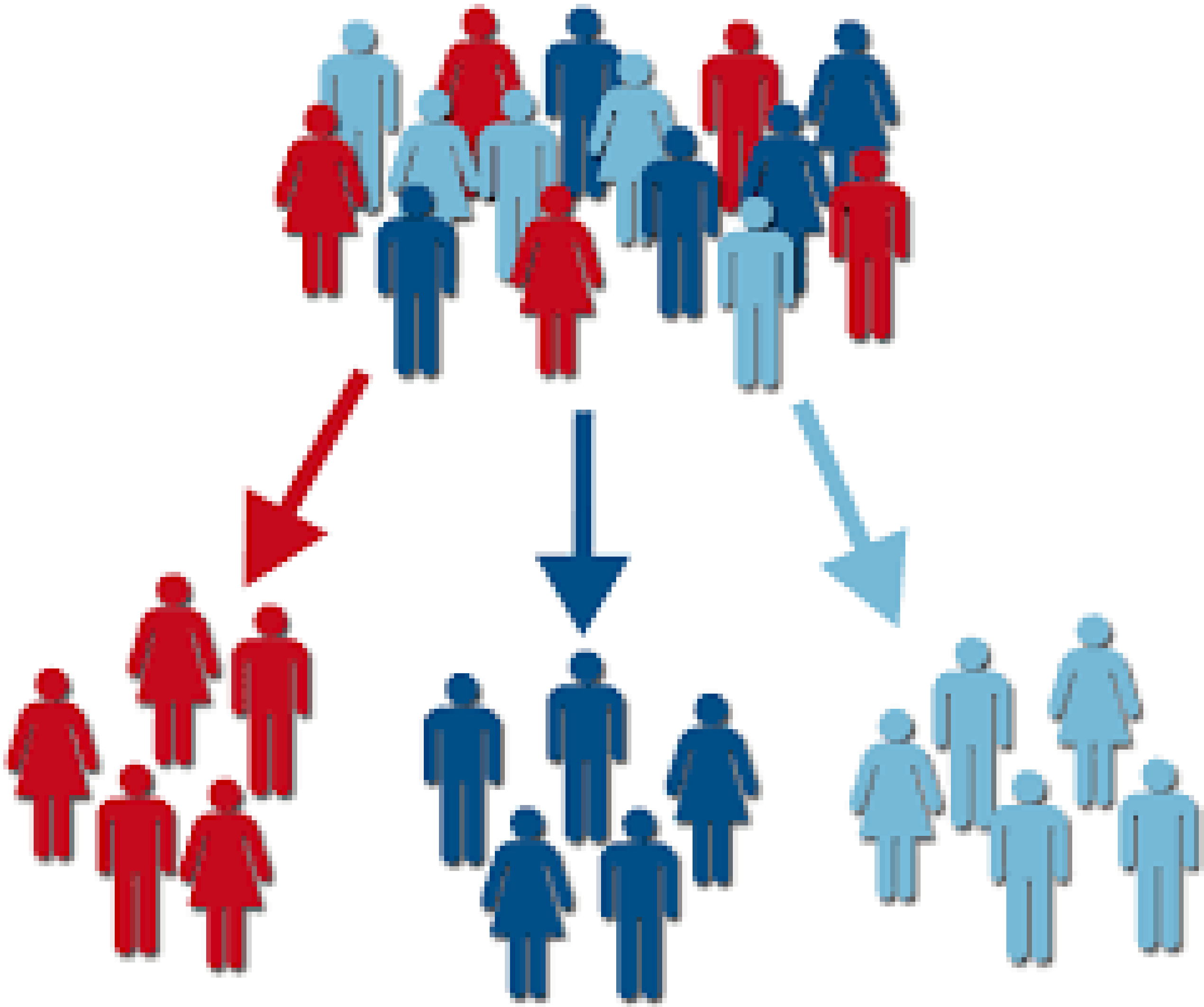
MACQUARIE
University
SYDNEY·AUSTRALIA

CUSTOMER SEGMENTATION ANALYSIS



Author: Jasmine Huynh

Table of Contents

Part 1: Introduction

Customer segmentation is vital for businesses as it allows them to customize their marketing efforts and deliver personalized experiences to different customer groups. For a large supermarket chain, understanding the diverse characteristics and preferences of its customer base is essential for optimizing marketing campaigns, improving customer satisfaction, and driving revenue growth.

In this report, we focus on customer segmentation analysis for the aforementioned supermarket chain. We utilize a dataset containing information on 2000 customers, collected through loyalty cards used at checkout. This dataset provides valuable insights into customer demographics, including age, gender, annual income, and more.

Our main objective is to identify distinct customer segments within the dataset using clustering techniques, such as k-Means and Hierarchical Tree clustering. By grouping customers based on shared characteristics, we aim to uncover meaningful patterns and gain deeper insights into the various customer profiles present in the supermarket chain's customer base.

By the end of this report, the management team will have a comprehensive understanding of the identified customer segments. Armed with this knowledge, they can make informed decisions regarding marketing techniques and strategies tailored to the specific needs and preferences of each customer segment. Ultimately, the goal is to enhance customer engagement, improve customer satisfaction, and drive business growth for the supermarket chain.

Part 2: Exploratory Data Analysis

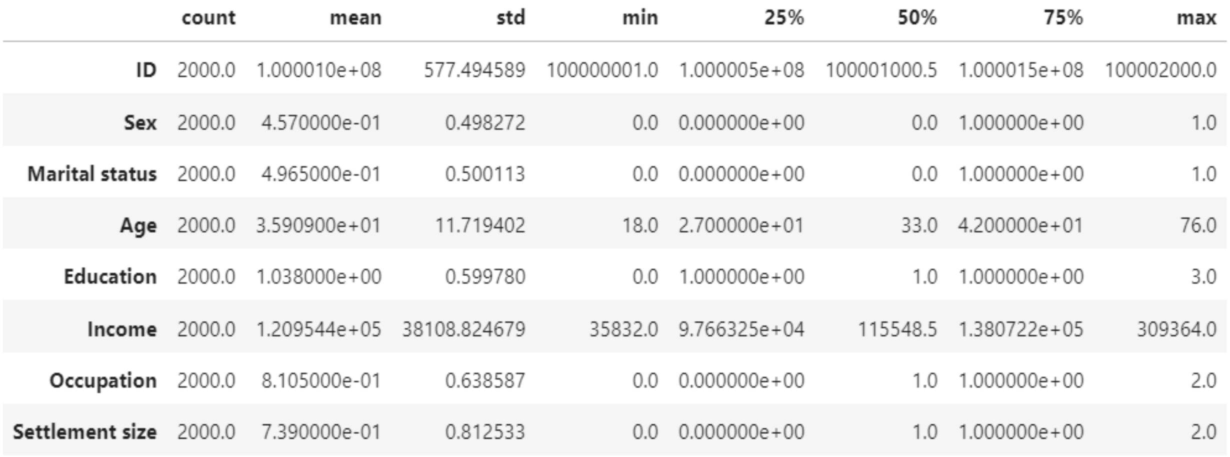
Overall, there are 2000 rows of unique customer data with 8 variables which are ID, Sex, Marital status, Age, Income, Education, Occupation and Settlement size in the data set. They are all numerical variables without any missing values that need to be handled.

2.1 Summary statistics and distribution of each individual variables

According to summary statistic table (graph ii.1.1) and the distribution graph of each individual variables (graph ii.1.2):

Variables	The distribution
Sex	Based on graph ii.1.2 - chart 1, it suggests a slightly higher proportion of males compared to females.
Marital status	Based on graph ii.1.2 - chart 2, it suggests a fairly balanced distribution between single and non-single customers.
Age	- The age analysis reveals that the average age of the customers is approximately 35.909, with a standard deviation of 11.719402 (graph ii.1.1 - row 5). The minimum recorded age is 18, while the maximum age is 76 (graph ii.1.1 - row 4). - Examining the histogram of age (graph ii.1.2 - chart 6), we observe a right-skewed distribution, indicating that the majority of customers are concentrated in the young and middle-age range, which spans from 27 to 42. - For better segmentation and analysis, we can categorize the age into three groups: 1. Young age: Customers aged 18 to 30. 2. Middle-age: Customers aged 31 to 60. 3. Old age: Customers aged 61 to 76. By dividing the age range into these groups, we can gain insights into the distribution and characteristics of different customer segments.
Education	Based on graph ii.1.2 - chart 3, there is a majority of customers with high school education, followed by a significant number with university degrees, and a smaller proportion in the "Other/Unknown" category, while there is a small representation in the graduate school qualifications.
Income	- The mean income is $120,954.419, with a standard deviation of $38,108.824679. The minimum income is $35,832, and the maximum income is $309,364 (graph ii.1.1 - row 6). - By examining the histogram of income (graph ii.1.2 - chart 7), we can observe a relatively normal distribution, indicating a wide range of income levels among the customers. - To facilitate segmentation and analysis, we can classify the customers into the following groups: 1. High income: Customers with incomes greater than $120,954.419, indicating above-average income levels.

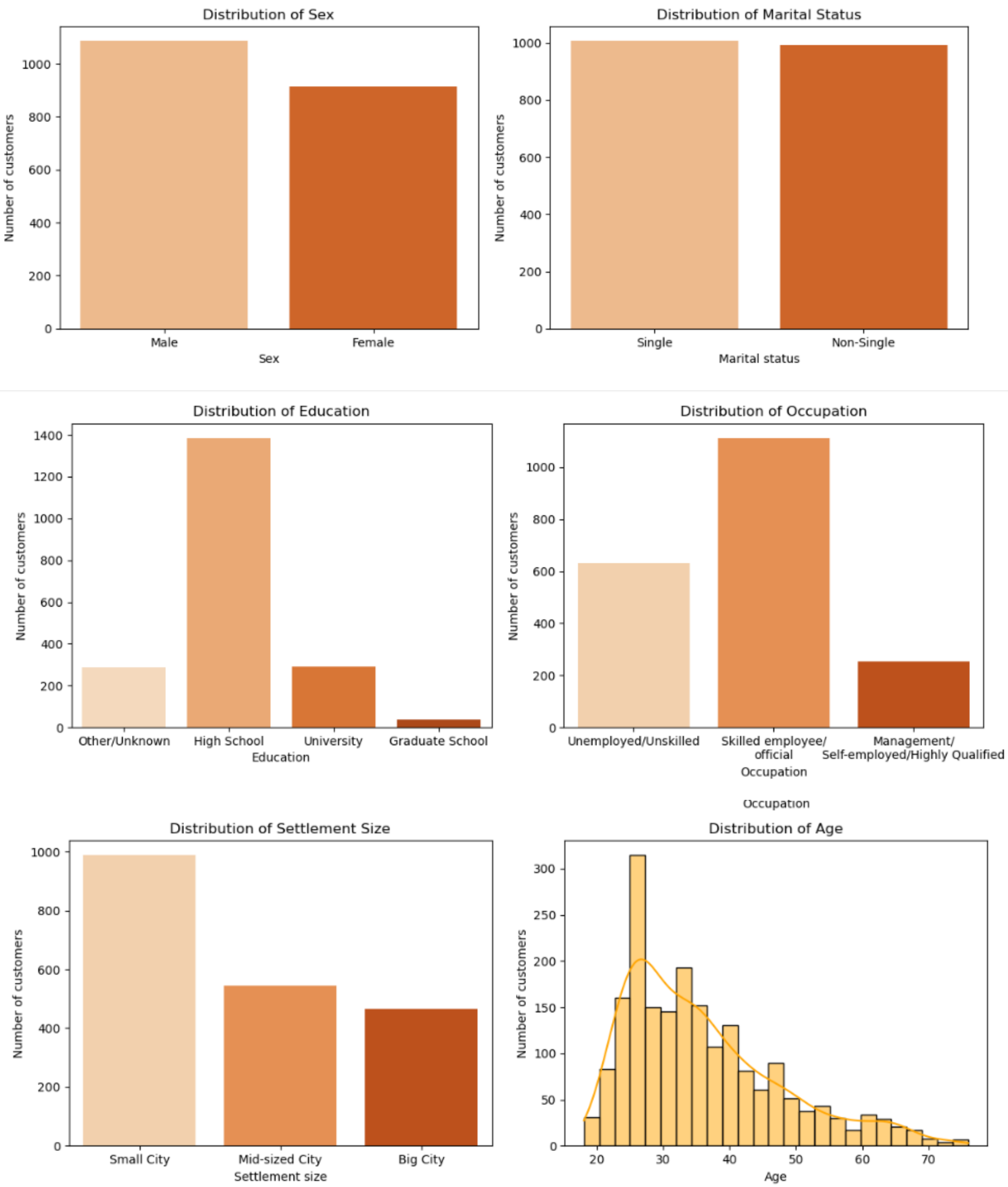
	2. Average income: Customers with incomes around the mean income of $120,954.419, representing a moderate income level. 3. Low income: Customers with incomes lower than $120,954.419, indicating below-average income levels. These income groupings will provide a framework for further analysis and insights into the customer base.
Occupation	Based on graph ii.1.2 - chart 4, a significant number of customers are skilled employees or officials, followed by a notable presence of unemployed or unskilled individuals and the least number is individuals in management, self-employment, or highly qualified positions.
Settlement size	Based on graph ii.1.2 - chart 5, the distribution of customers based on city size reveals that the lowest number of customers reside in small cities, followed by a slightly higher number in mid-size cities, and the highest number in big cities.

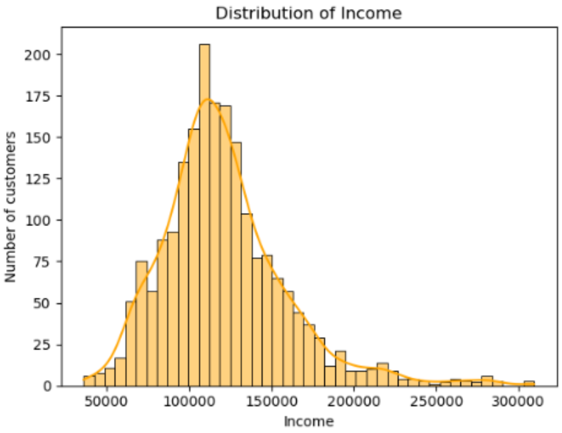


	count	mean	std	min	25%	50%	75%	max
ID	2000.0	1.000010e+08	577.494589	100000001.0	1.000005e+08	100001000.5	1.000015e+08	100002000.0
Sex	2000.0	4.570000e-01	0.498272	0.0	0.000000e+00	0.0	1.000000e+00	1.0
Marital status	2000.0	4.965000e-01	0.500113	0.0	0.000000e+00	0.0	1.000000e+00	1.0
Age	2000.0	3.590900e+01	11.719402	18.0	2.700000e+01	33.0	4.200000e+01	76.0
Education	2000.0	1.038000e+00	0.599780	0.0	1.000000e+00	1.0	1.000000e+00	3.0
Income	2000.0	1.209544e+05	38108.824679	35832.0	9.766325e+04	115548.5	1.380722e+05	309364.0
Occupation	2000.0	8.105000e-01	0.638587	0.0	0.000000e+00	1.0	1.000000e+00	2.0
Settlement size	2000.0	7.390000e-01	0.812533	0.0	0.000000e+00	1.0	1.000000e+00	2.0

graph ii.1.1: *summary statistics of all variables in data set*

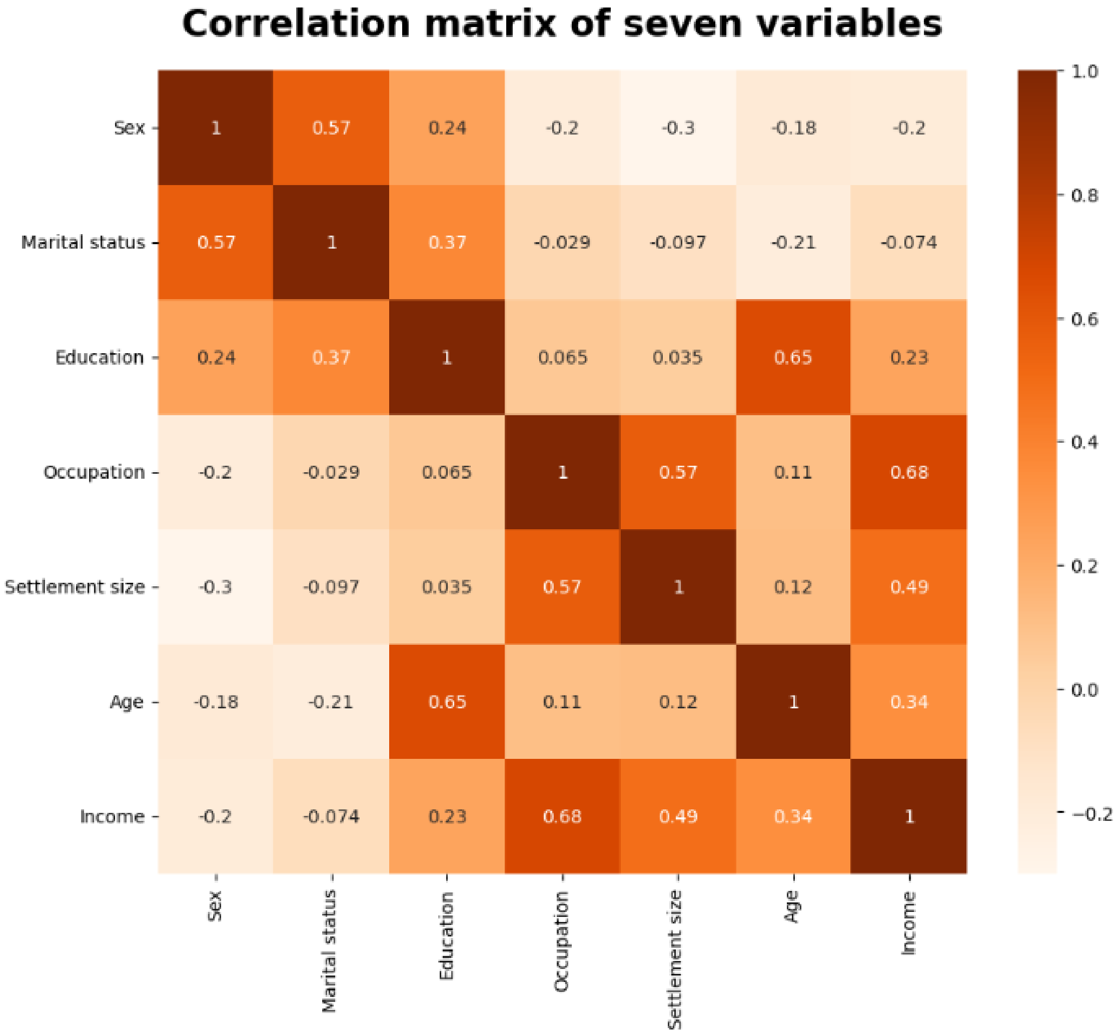
Bar charts and Histogram to display the distribution of each variable





***graph ii.1.2:** the distribution of each individual variables in the data set*

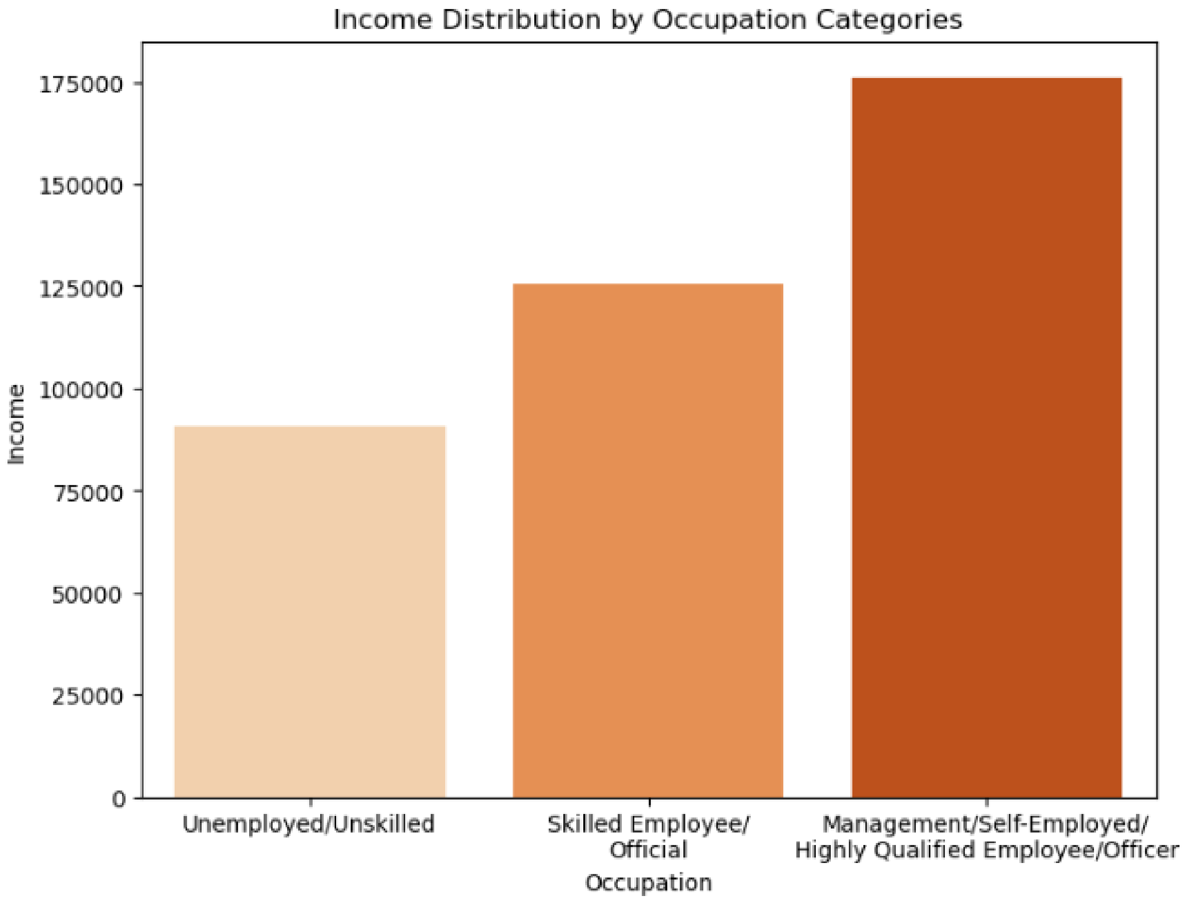
2.2 Relationship between variables



graph ii.2.1: *the correlation matrix of seven variables*

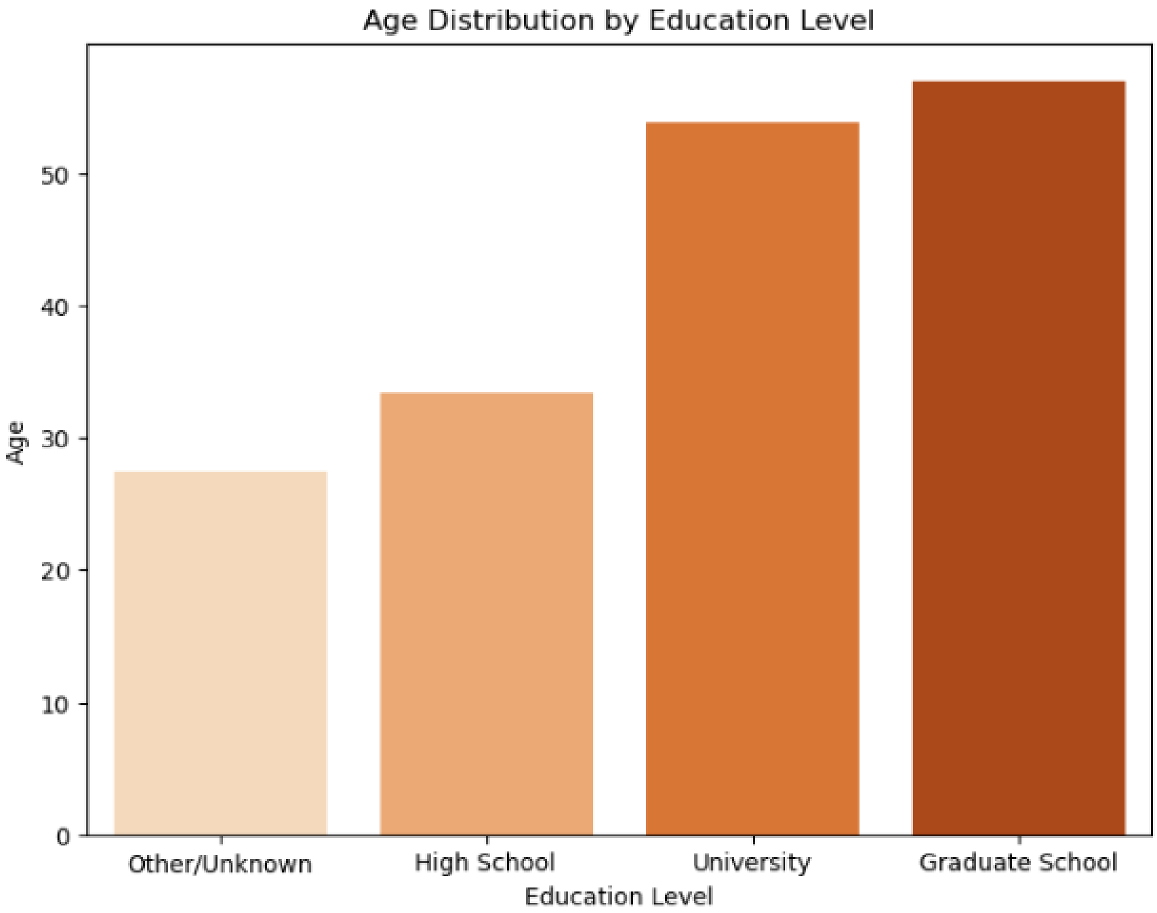
Overall, firstly, the correlation coefficients suggest that there is minimal linear relationship between Marital Status and Occupation (-0.029), Marital Status and Settlement Size (-0.097), Marital Status and Income (-0.074), Education and Occupation (0.065), and Education and Settlement Size (0.035) (graph ii.2.1). These values indicate that being single or non-single does not significantly impact the differences in Income, Occupation, or Settlement Size. Similarly, the level of education does not significantly influence the differences in Occupation and Settlement Size.

Secondly, there is quite a strong positive relationship between Income and Occupation (0.68), Age and Education (0.65) (graph ii.2.1), indicating that these variables are positively correlated.



graph ii.2.2: *customer income based on different occupation categories*

To be more specific, the graph above reveals a clear trend between occupation and income. As the level of occupation increases, particularly in the category of management/highly qualified staff, customers tend to have higher incomes.



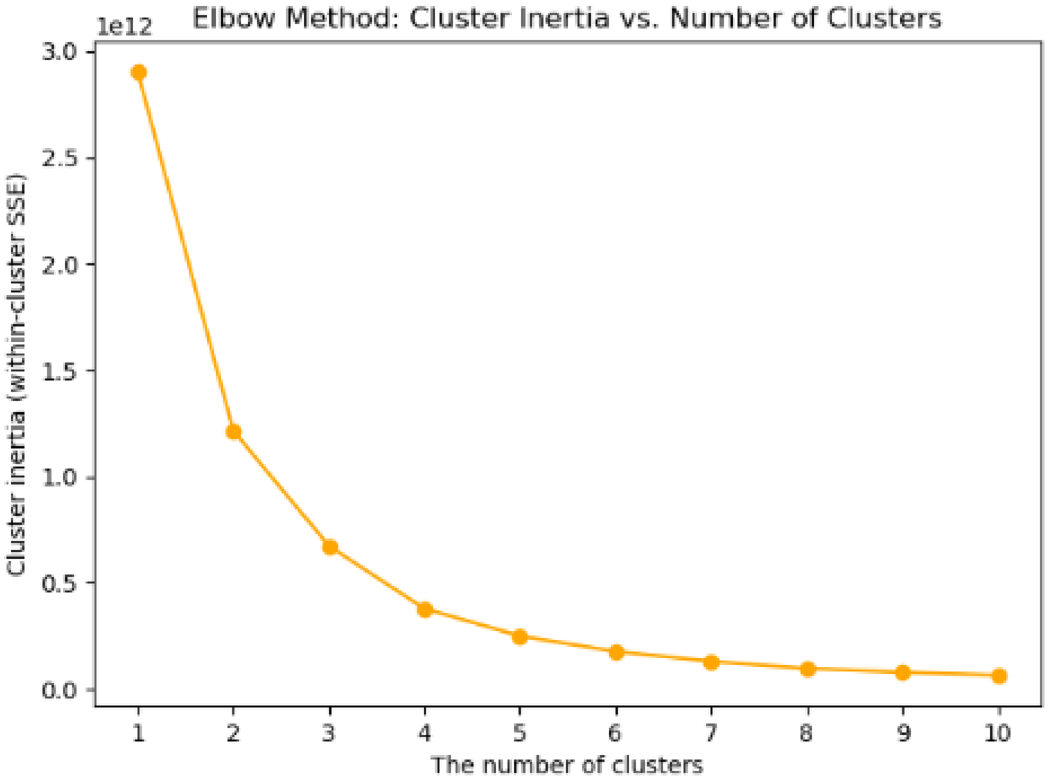
graph ii.2.3: age distribution by education level

The graph above clearly shows a positive relationship between education level and age. Customers with higher education, especially those with graduate school qualifications, tend to be older.

Part 3: Customer segmentation

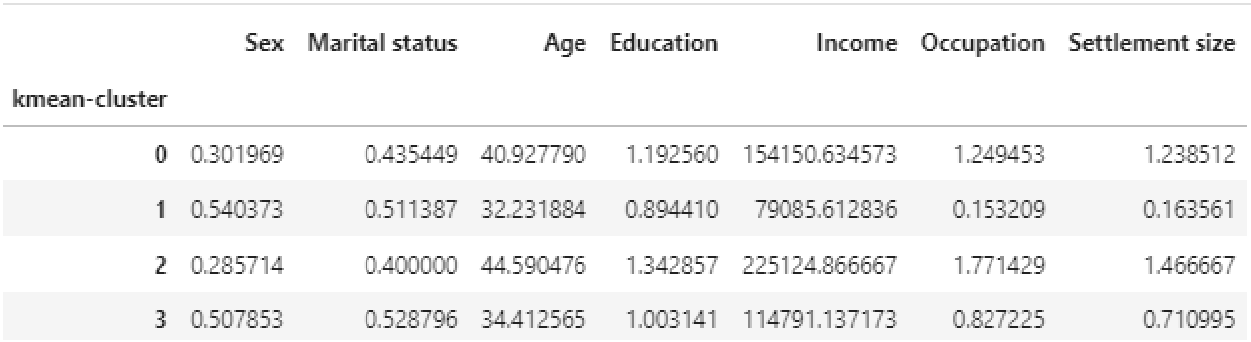
To perform customer segmentation, we utilize k-means clustering and agglomerative clustering algorithms provided by scikit-learn. The elbow method is employed to determine the optimal number of clusters in the k-means approach.

3.1 K-means cluster:



graph iii.1.1: elbow method: cluster inertia vs. number of clusters

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 4. Therefore, we proceeded to compute the mean values of the variables for each customer group within the clusters 0, 1, 2, and 3.



	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
kmean-cluster							
0	0.301969	0.435449	40.927790	1.192560	154150.634573	1.249453	1.238512
1	0.540373	0.511387	32.231884	0.894410	79085.612836	0.153209	0.163561
2	0.285714	0.400000	44.590476	1.342857	225124.866667	1.771429	1.466667
3	0.507853	0.528796	34.412565	1.003141	114791.137173	0.827225	0.710995

graph iii.1.2: *the mean values of the variables for each customer group within the k-mean clusters 0, 1, 2, and 3*

The profile of 4 group of customer based on the graph above are:

Group 1: The "Skilled Middle-age Single Male" segment comprises middle-aged males with high school education, who are skilled employees. They reside in mid-sized cities and have an average income.

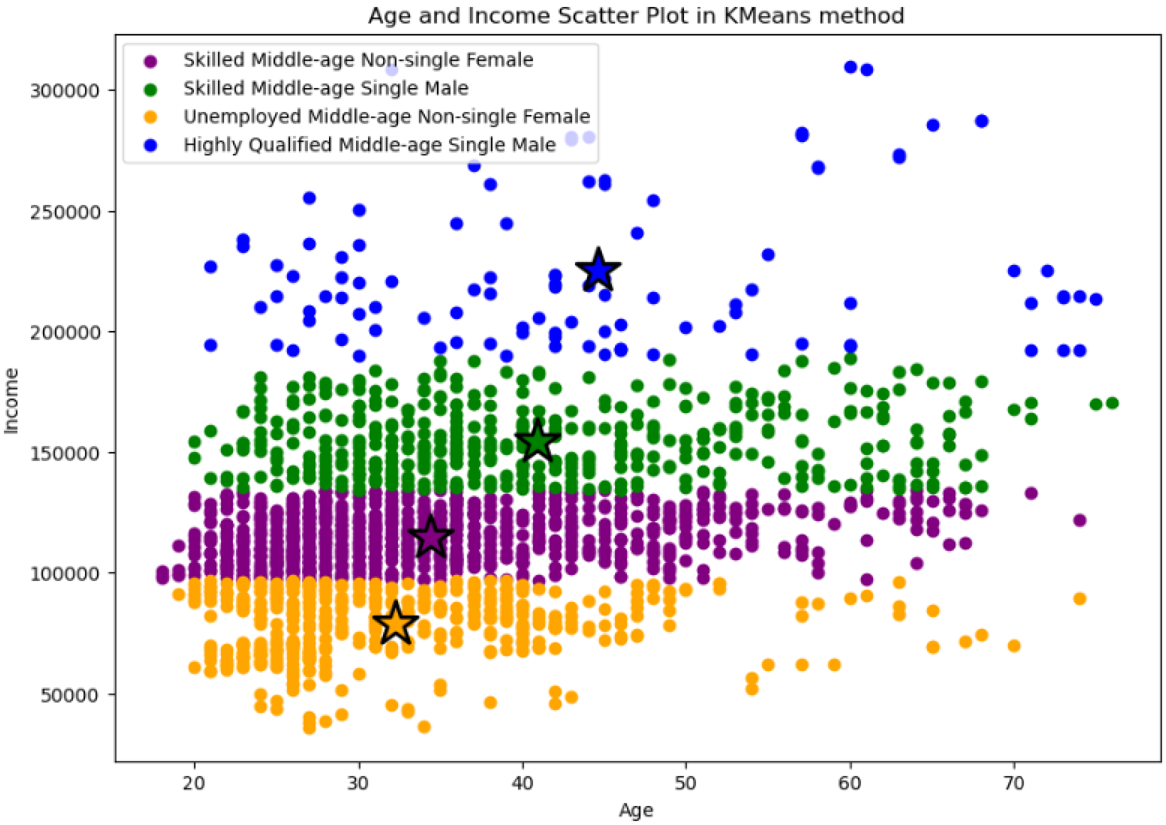
Group 2: The "Unemployed Middle-age Non-single Female" segment consists of middle-aged females with high school education who are unemployed. They live in small cities and have low income.

Group 3: The "Highly Qualified Middle-age Single Male" segment includes middle-aged males with high school education, who are highly qualified employees. They reside in mid-sized or big cities and have high income.

Group 4: The "Skilled Middle-age Non-single Female" segment consists of middle-aged females with high school education, who are skilled employees. They reside in mid-sized cities and have high income.

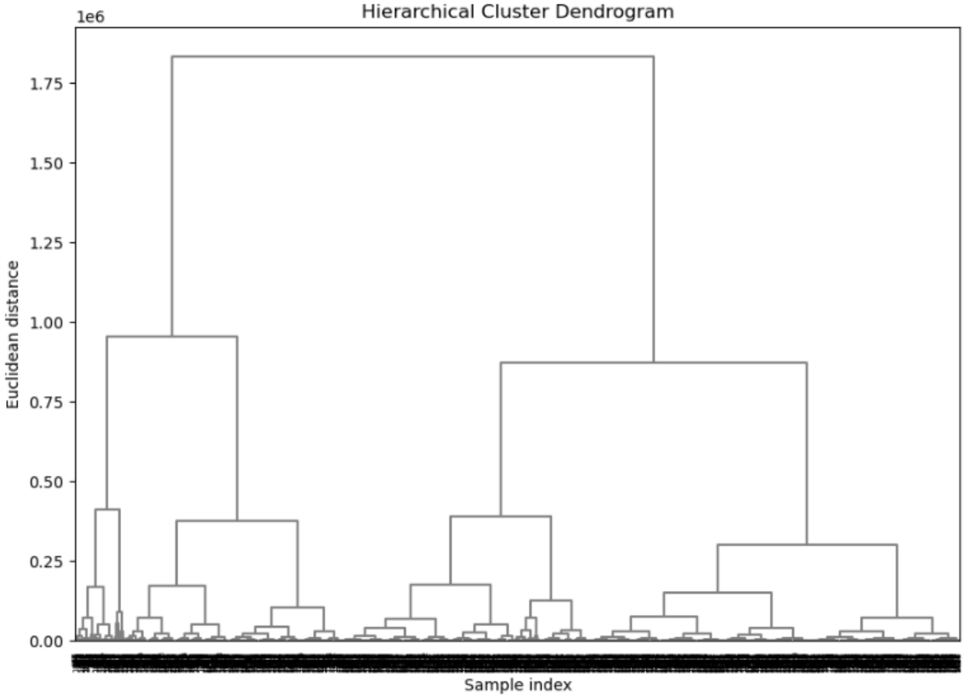
Please note that the age and income levels used in the segmentation are determined based on the summary statistics table created in Part 1.

The following scatter plot provides a comprehensive visual representation of the income and age distribution within customer groups obtained from the KMeans method. It showcases all data points and centroids for each cluster, allowing us to analyze the correlation between age and income within each group:



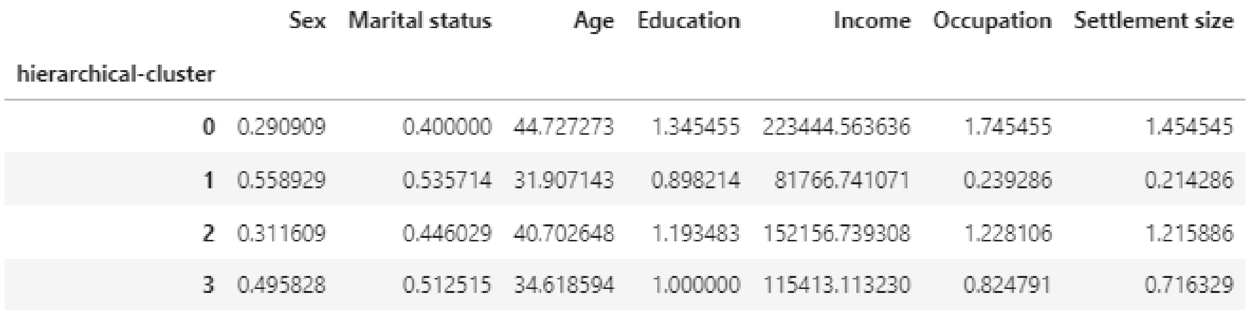
graph iii.1.3: age and income scatter plot in k-means method

3.2 Hierarchical cluster:



graph iii.2.1: hierarchical cluster dendrogram

After applying the agglomerative algorithm, we conducted hierarchical clustering using a distance matrix and generated a dendrogram to identify the optimal number of clusters. The analysis indicated that the ideal number of clusters is 4. Subsequently, we calculated the average values of the variables for each customer group within the clusters labeled as 0, 1, 2, and 3.



	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
hierarchical-cluster							
0	0.290909	0.400000	44.727273	1.345455	223444.563636	1.745455	1.454545
1	0.558929	0.535714	31.907143	0.898214	81766.741071	0.239286	0.214286
2	0.311609	0.446029	40.702648	1.193483	152156.739308	1.228106	1.215886
3	0.495828	0.512515	34.618594	1.000000	115413.113230	0.824791	0.716329

graph iii.2.2: *the mean values of the variables for each customer group within the k-mean clusters 0, 1, 2, and 3*

The profile of 4 group of customer based on the graph above are:

Group 1: The "Highly Qualified Middle-age Single Male" segment consists of middle-aged males with high school education, who are highly qualified employees. They reside in a mid-sized or big city and have high income.

Group 2: The "Unemployed Middle-age Non-single Female" segment comprises middle-aged females with high school education, who are unemployed. They reside in a small city and have low income.

Group 3: The "Skilled Middle-age Single Male" segment includes middle-aged males with high school education, who are skilled employees. They reside in a mid-sized city and have average income.

Group 4: The "Skilled Middle-age Non-single Female" segment consists of middle-aged females with high school education, who are skilled employees. They reside in a mid-sized city and have average income.

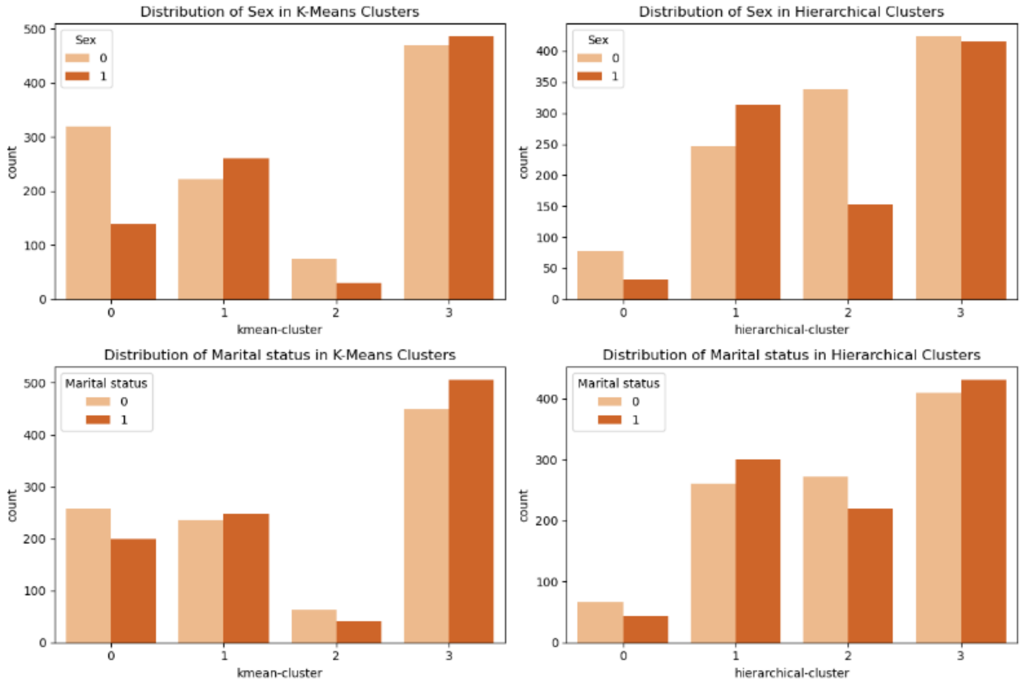
Please note that the age and income levels used in the segmentation are determined based on the summary statistics table created in Part 1.

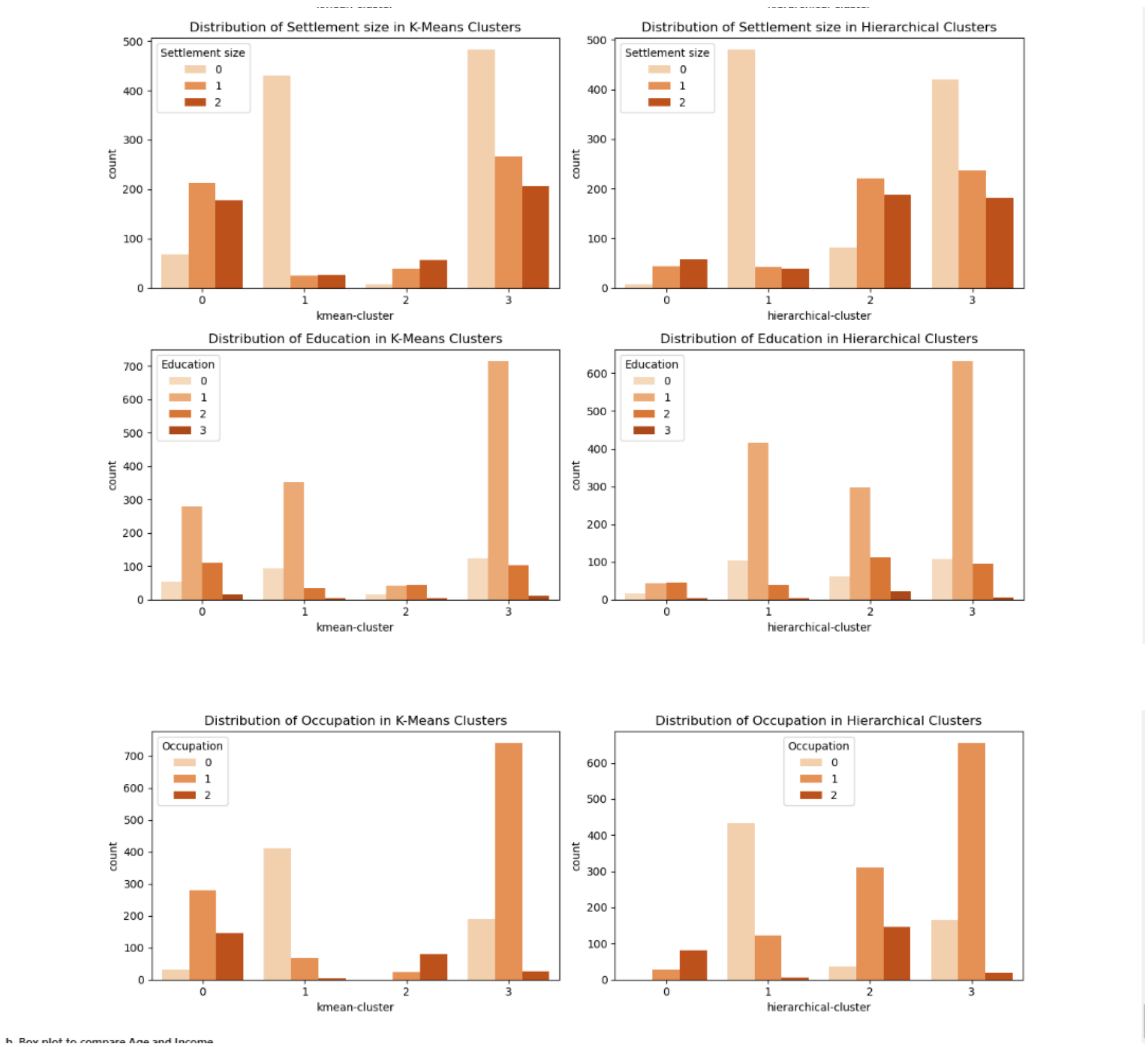
3.3 Comparison between two methods:

Based on the analysis and the provided below graphs, we observe that the two clustering methods have identified similar groups of customers. Specifically, Group 1 from the k-means method corresponds to Group 3 from the hierarchical method, while Group 3 from the k-means method aligns with Group 1 from the hierarchical method. Similarly, Group 2 and 4 from the k-means method are similar to Group 2 and 4 from the hierarchical method, respectively.

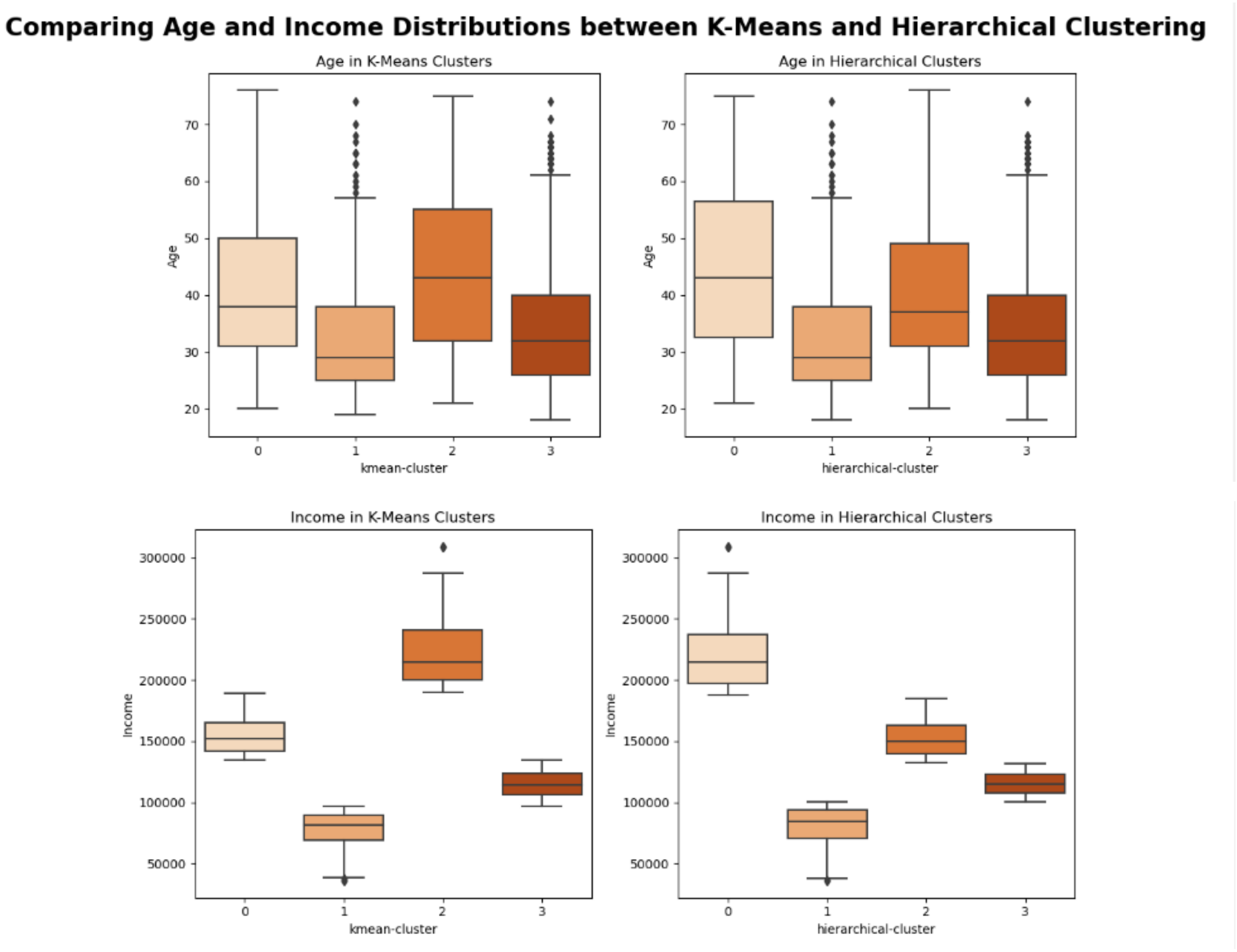
Furthermore, we notice that all the variables in both methods exhibit similar ranges without any significant discrepancies. This suggests that the clustering results are consistent across the different methods, reinforcing the similarity of the identified customer groups.

Comparing Categorical Feature Distributions between K-Means and Hierarchical Clustering





graph iii.3.1: Comparing Categorical Feature Distributions between K-Means and Hierarchical Clustering



graph iii.3.2: Comparing Age and Income Distributions between K-Means and Hierarchical Clustering

Part 4: Recommendation

Customer Profile	Suggestion marketing techniques
 Highly Qualified Middle-age Single Male have high income.	1. Provide gourmet or premium food goods and ingredients that appeal to their refined preferences and higher income levels. 2. Offer cooking lessons or seminars that emphasise gourmet recipes and culinary expertise. 3. Work with local fitness centres or health clubs to provide unique discounts or partnerships to promote a healthy lifestyle.
 Unemployed Middle-age Non-single Femal have low income.	1. Provide budget-friendly meal planning alternatives and low-cost meal packages to assist people save time and money when cooking. 2. Promote in-store or online job ads as well as career development services to help people find work. 3. Establish loyalty programmes that provide unique discounts or prizes on necessary home products.

 Skilled Middle-age Single Male have average income.	1. Plan events or campaigns centred on quick and easy food options for working professionals. 2. To accommodate their time restrictions, highlight convenience foods, pre-prepared meals, and grab-and-go choices. 3. Provide regular buyers in this sector with loyalty benefits or discounts.
 Skilled Middle-age Non-single Female have average income.	1. Highlight family-friendly items and provide discounts on large purchases of home supplies. 2. Provide in-store childcare or play places to accommodate their carer obligations. 3. Create a section with healthy food ideas and nutritional recommendations for their families.

Furthermore, for all groups of customer:
- Use personalised marketing tactics, such as targeted email campaigns or personalised offers based on their purchase patterns.
- Use social media networks to provide recipes, culinary suggestions, and special offers.
- Implement consumer feedback programmes to gather information and enhance the shopping experience overall.

The supermarket may successfully communicate with their target audience and establish long-term consumer loyalty by adapting marketing strategies to the individual demands and preferences of each client group.

Part 5: Conclusion

This report focuses on customer segmentation analysis for a large supermarket chain using clustering techniques. The dataset of 2000 customers provides insights into demographics, and through exploratory data analysis, we identified trends in variables like sex, marital status, age, income, education, occupation, and settlement size. Correlation analysis revealed minimal relationships between certain variables, while strong positive correlations were found between income and occupation, and age and education.

Using K-means and Hierarchical clustering methods, we identified four distinct customer segments. The clustering results were consistent across methods, and the segments were characterized by their mean values in each variable. Recommendations for personalized marketing techniques were provided for each segment, aiming to enhance customer engagement and satisfaction.

In conclusion, by leveraging customer segmentation and tailoring marketing strategies based on the identified segments, the supermarket chain can drive business growth, improve customer experiences, and foster long-term loyalty.