Customer Segmentation Analysis – Travel Agency Dataset

1. Introduction

This report aims to overcome our current limitation of lacking targeted marketing through a comprehensive customer segmentation analysis. It begins with an exploratory data analysis (EDA) to identify key patterns and trends. Then, K-means++ and agglomerative clustering techniques are applied to segment the customers. Finally, tailored marketing strategies are recommended to ensure we effectively target each customer group. Throughout this report, we will use the dataset of 2,000 customers with their seven attributes, including age, income, gender, education level, settlement size, occupation, and marital status.

2. Exploratory Data Analysis

In this section, EDA will be conducted by first summarising the key statistics for the seven variables, as shown in the results below.

Figure 1. Summary of key statistics

	Gender	Marital Status	Age	Education	Income	Occupation	Settlement Size
count	2000.00	2000.0	2000.00	2000.00	2000.00	2000.00	2000.00
mean	0.60	0.5	40.82	1.46	137516.20	0.61	0.83
std	0.49	0.5	9.46	0.78	46184.30	0.67	0.97
min	0.00	0.0	20.00	0.00	35832.00	0.00	0.00
25%	0.00	0.0	33.00	1.00	101262.75	0.00	0.00
50%	1.00	1.0	40.00	1.00	133004.00	1.00	0.00
75%	1.00	1.0	48.00	2.00	171232.50	1.00	2.00
max	1.00	1.0	76.00	3.00	309364.00	2.00	2.00

On average, our customers are 41 years old, earning an annual income of $137,516. The table shows the significant variance in income distribution (from $35,832 to $309,364),

indicating diverse economic profiles across our customer base. Similarly, the broad age range of 20 to 76 also highlights our customers' demographic diversity.

Figure 2. Customer Age Distribution

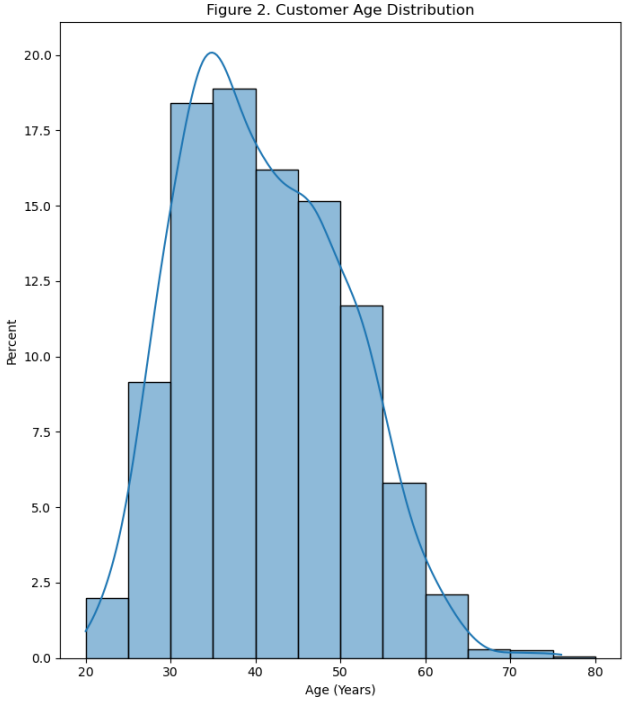


Figure 3. Customer Income Distribution

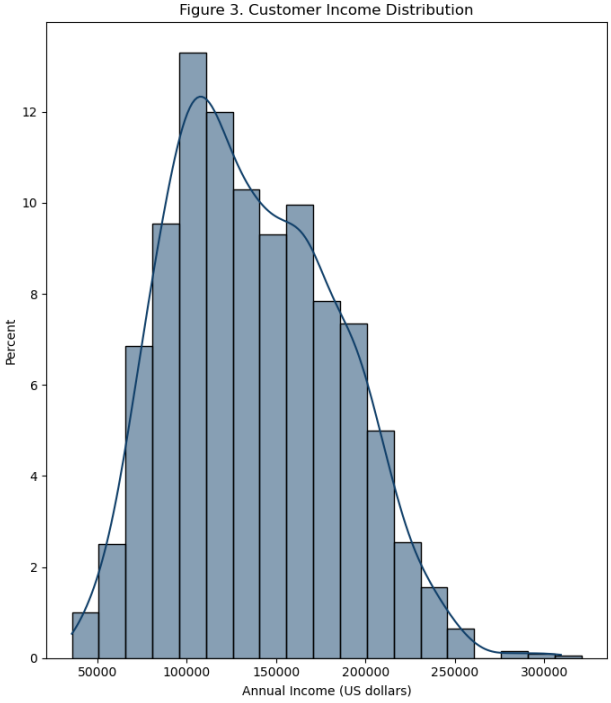
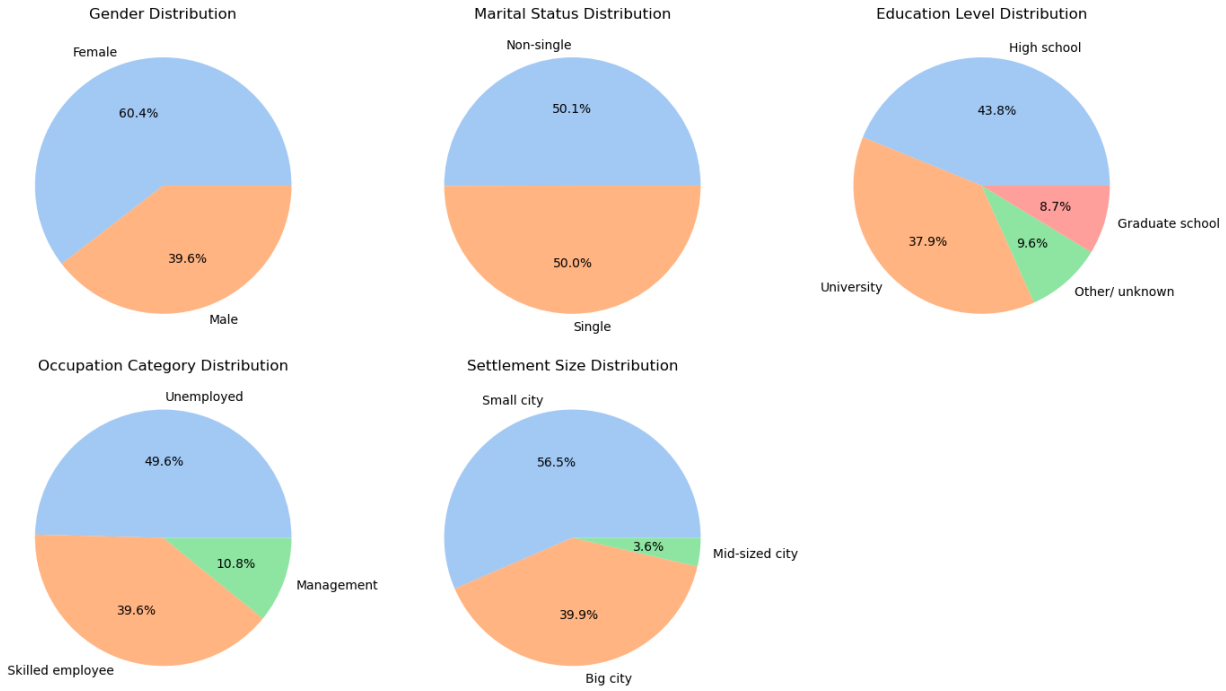


Figure 2 shows that most of our customers are middle-aged (30 to 50), with only a small number of older customers over 65. Additionally, most customers fall within the middle-income range of $80,000 to $170,000, while a smaller portion earns more than $200,000.

Figure 4. Categorical/ Ordinal variables distribution



Moving to categorised attributes, most customers are women (60%), with 43.8% having completed high school, and nearly half (49.6%) being unemployed. The majority live in smaller cities (49.6%), and they are evenly split between single and non-single individuals.

Figure 5. Customer Age Distribution Across Genders

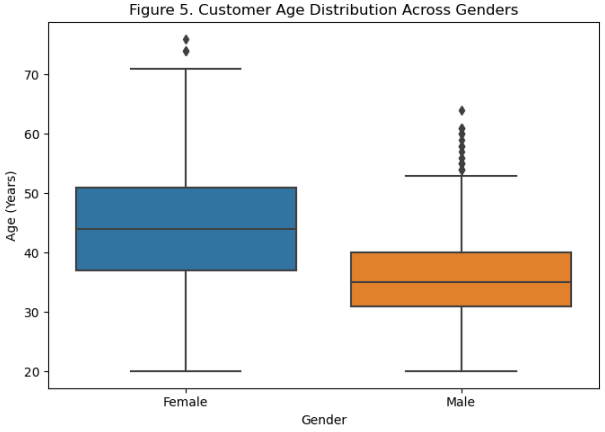
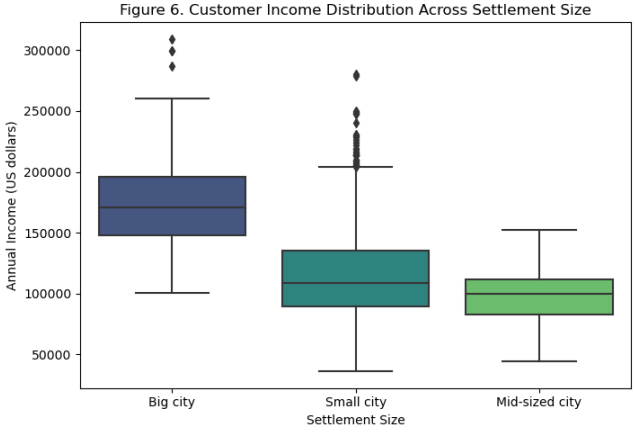
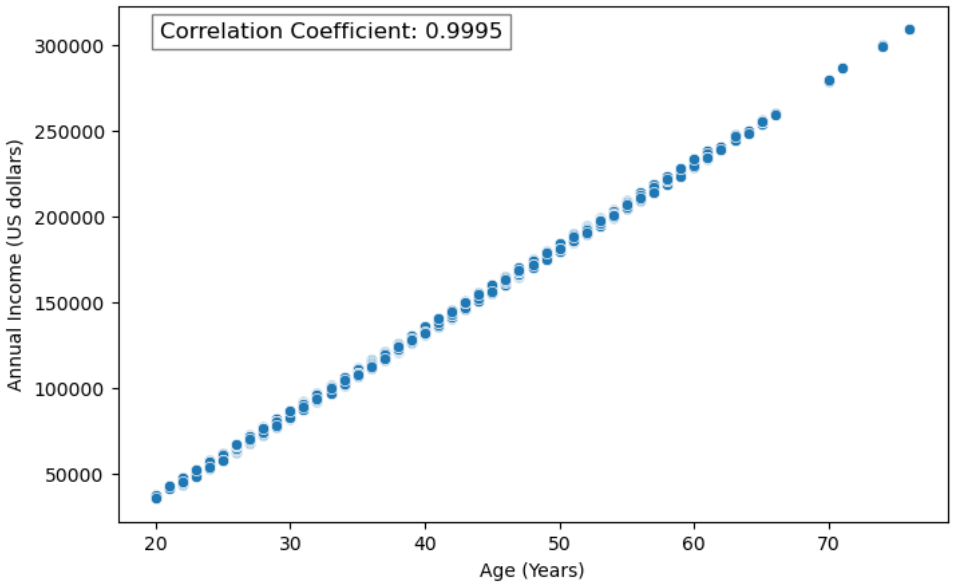


Figure 6. Customer Income Distribution Across Settlement Size



Next, the data also reveals that female customers tend to be older, mostly between 40 and 50, while male customers are generally from 30 to 40. Interestingly, customers in mid-sized cities have the lowest incomes (below $100,000), while those in small cities earn over $100,000 on average. The highest earners are from big cities, with an average income of around $175,000.

Figure 7. Age vs. Income



Finally, Figure 7 highlights a strong positive connection between age and income, suggesting elder customers tend to earn more.

3. Customer Segmentation

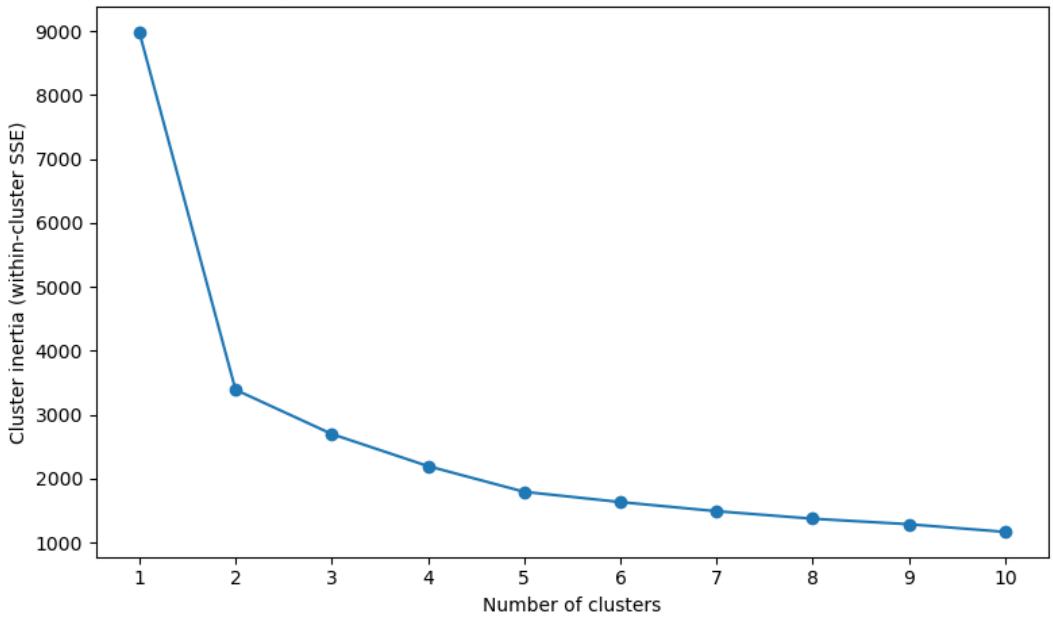
Subsequently, clustering techniques are applied to segment our customers. Since income and age are on different scales, we will first adjust them to be on the same range to ensure the algorithm can fairly group customers based on these attributes.

Figure 8. A snapshot of the data after scaling

	Gender	Marital Status	Age	Education	Income	Occupation	Settlement Size	Age_scaled	Income_scaled
0	1	1	39	2	130568	1	2	-0.192892	-0.150483
1	0	0	29	1	80315	0	0	-1.250703	-1.238852
2	1	0	35	0	107067	0	0	-0.616016	-0.659462
3	0	1	56	2	214000	1	0	1.605387	1.656471
4	1	1	45	2	158138	1	2	0.441795	0.446623

Then, the elbow graph is plotted to determine the ideal number of segments. The results show a sharp drop in inertia (which measures how different the customers are within each group) when dividing them into two groups, with smaller drops for three and four groups.

Figure 9. Elbow Method for Optimal Number of Clusters



Hence, to confirm the best option, Silhouette plots are used to measure how well the groups are separated. The results (Figures 10, 11, and 12) suggest that segmenting the data into two clusters is the best option, as it produces the highest Silhouette score (0.54).

Figure 10. Silhouette Plot for Evaluating the Number of Clusters (K=2)

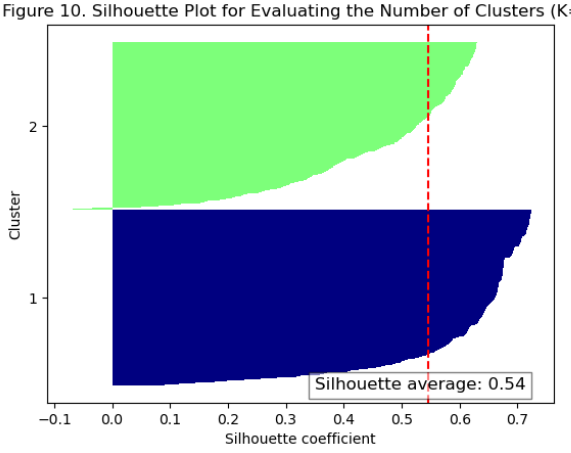


Figure 11. Silhouette Plot for Evaluating the Number of Clusters (K=3)

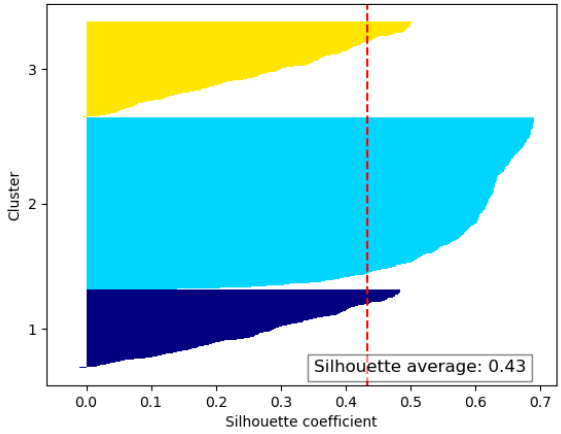
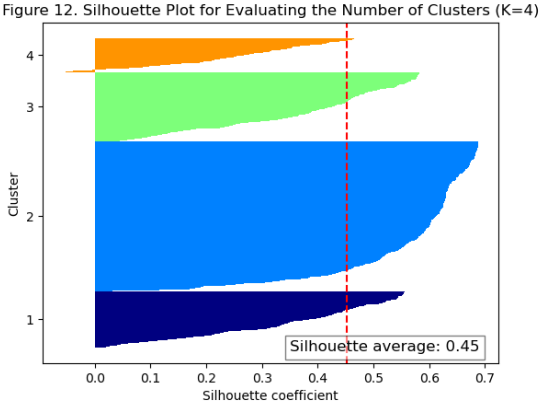


Figure 12. Silhouette Plot for Evaluating the Number of Clusters (K=4)



After that, we proceed to apply both the K-means++ and agglomerative clustering methods to estimate the customer segments. The results of each technique are outlined below.

Figure 13. Customer Segmentation by K-means++

	Avg Age	Avg Income	Gender	Marital Status	Education	Occupation	Settlement Size	No. Customers
Cluster 1	34.0	103257.0	Male	Single	High school	Unemployed / Unskilled	Small city	1024
Cluster 2	48.0	173461.0	Female	Non-single	University	Skilled employee / Official	Big city	976

With K-means++, the first segment consists of 1,024 customers, likely representing early middle-aged, single, unemployed men with high school degrees, middle incomes, and living in small cities. The second segment, with 976 customers, mostly includes non-single, late middle-aged women who are employed as skilled workers, earning high incomes, and living in big cities.

Figure 14. Customer Segmentation by Agglomerative Clustering

	Avg Age	Avg Income	Gender	Marital Status	Education	Occupation	Settlement Size	No. Customers
Cluster 1	47.0	168085.0	Female	Non-single	University	Skilled employee / Official	Mid-sized city	1163
Cluster 2	32.0	95041.0	Male	Single	High school	Unemployed / Unskilled	Small city	837

Similarly, the agglomerative clustering approach yields comparable results with a few differences. One segment contains late middle-aged, high-income, non-single women with university degrees, residing in mid-sized cities, unlike K-means++ where they live in big cities. The other segment is consistent with K-means++ results, featuring early middle-aged, single men with high school diplomas, middle incomes, and living in small cities.

The key distinction between these two methods lies in the distribution of customers across the segments. While K-means++ places fewer customers in the female segment compared to the male group, agglomerative clustering shows the reverse, with 1,163 customers in the female cluster and 837 in the male cluster.

4. Recommendations

Based on the two customer segments identified through K-means++ clustering, specific marketing strategies have been devised to target each group effectively. The female segment, who tends to possess higher income and education levels, is more likely to appreciate premium, personalised travel experiences. To engage this group, the company should collaborate with women's organisations to promote exclusive, high-quality travel packages. Additionally, incorporating testimonials from other women can build trust and enhance appeal to this demographic. Since they are non-single, promoting family or group travel packages as opportunities to create lasting memories with loved ones is particularly beneficial. Moreover, offering customized options and additional perks can elevate the agency's visibility and nurture customer loyalty. Ultimately, capitalising on word-of-mouth marketing from satisfied customers could further expand this segment, as late middle-aged women are often influenced by their social circles.

In contrast, the male segment, comprising single, middle-income, and early middle-aged individuals, is inclined towards adventurous, budget-friendly travel options. Therefore, the focus should be on promoting experiences that resonate with their lifestyle, such as adventure trips, solo travel, or group excursions with friends. To effectively reach this audience, launching a targeted social media campaign with clear, engaging, and relatable messaging is crucial. Additionally, advertising on platforms popular among males, like sports forums, gaming platforms, and hobbyist communities, can effectively capture their attention. Finally, we should also promote special offers or discounts to emphasise affordability and encourage customers to share their travel experiences on social networks. This will then help amplify the campaign’s reach and attract more customers from this segment.

5. Conclusion

In conclusion, this report has successfully addressed our current problem of lacking targeted marketing strategies through the recommendations made after a comprehensive customer segmentation analysis. Through EDA, key customer patterns were uncovered, followed by the application of K-means++ and agglomerative clustering techniques. Two distinct customer segments were identified, leading to the development of tailored marketing strategies for each group. The female segment should be targeted with premium, personalised travel experiences, while the male segment responds better to adventurous, budget-friendly options. These insights will enable more effective targeting and engagement, ultimately improving customer satisfaction and driving revenue growth.