CUSTOMER SEGMENTATION

Table of Contents

Introduction

In today's competitive travel industry, understanding customer behavior is essential for business growth. This report analyzes a dataset of 2,000 customers to segment them based on demographic and behavioral attributes such as age, gender, and income. I used Python for data processing, exploratory data analysis, and visualization. Additionally, machine learning techniques, specifically clustering algorithms like K-Means++ and Agglomerative Clustering, were applied to identify distinct customer segments. These methods allowed us to uncover patterns and relationships within the data, which informed the development of targeted marketing strategies tailored to each customer group.

Combined elbow and silhouette methods to determine the optimal number of clusters and used K-means++ and Agglomerative clustering techniques to profile the customers. By categorizing customers into meaningful groups, we can create personalized marketing campaigns, ultimately improving business performance and customer satisfaction.

Exploratory data analysis

Descriptive statistics

First, conducted a descriptive analysis. In Table 1, we can observe that customers' ages range from 20 to 76 years, with an average age of 40. Additionally, 75% of the total customers are under the age of 48.

In terms of income, the minimum income is $35,832, while the maximum is $309,364, with an average income of $137,516.20. Furthermore, 75% of the customers have incomes below $171,232.5.

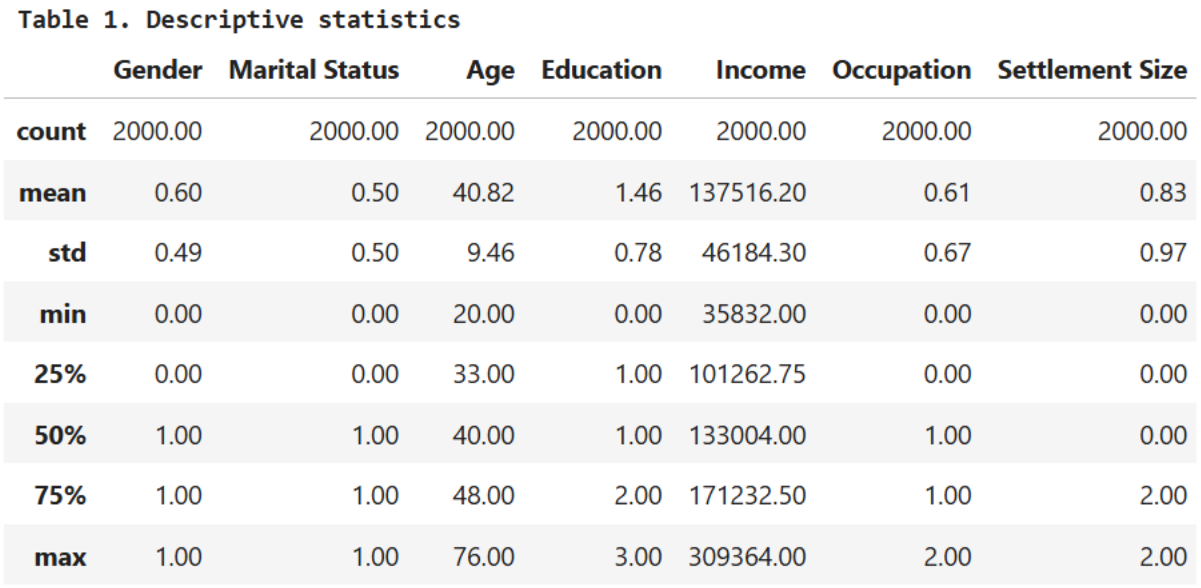


Table 1. Descriptive statistics

	Gender	Marital Status	Age	Education	Income	Occupation	Settlement Size
count	2000.00	2000.00	2000.00	2000.00	2000.00	2000.00	2000.00
mean	0.60	0.50	40.82	1.46	137516.20	0.61	0.83
std	0.49	0.50	9.46	0.78	46184.30	0.67	0.97
min	0.00	0.00	20.00	0.00	35832.00	0.00	0.00
25%	0.00	0.00	33.00	1.00	101262.75	0.00	0.00
50%	1.00	1.00	40.00	1.00	133004.00	1.00	0.00
75%	1.00	1.00	48.00	2.00	171232.50	1.00	2.00
max	1.00	1.00	76.00	3.00	309364.00	2.00	2.00

Pie chart

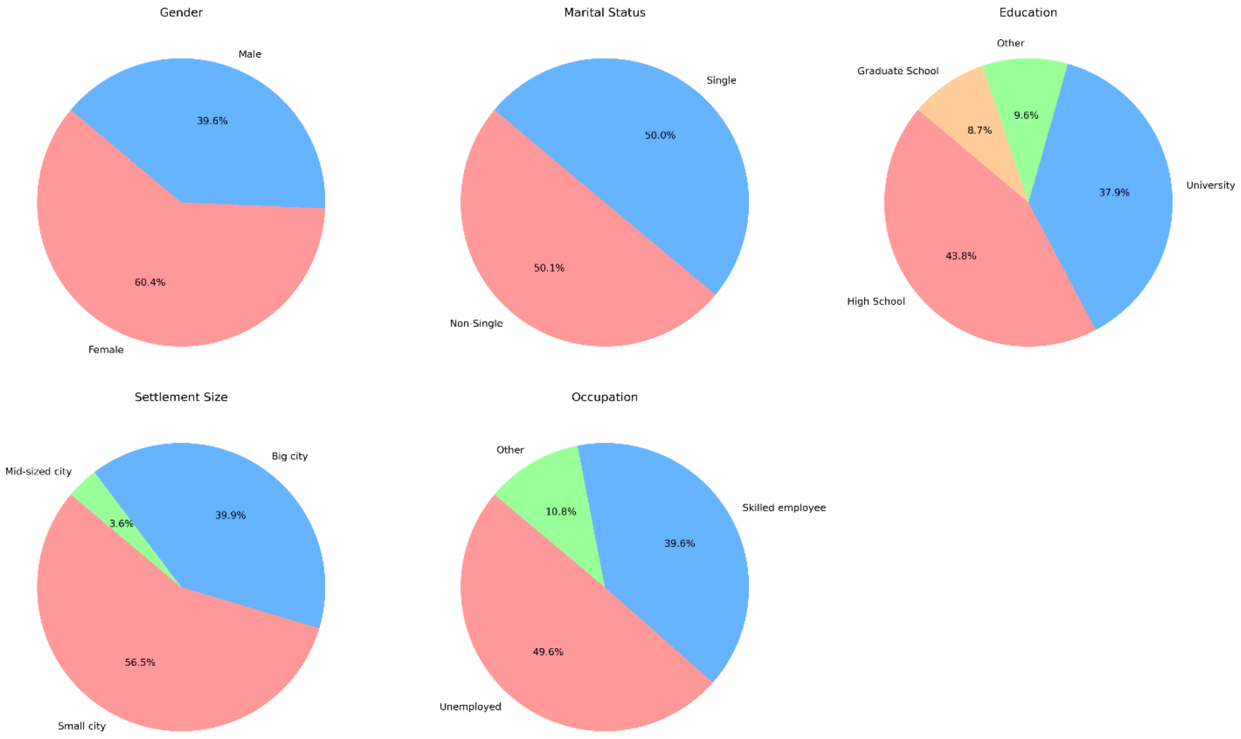
In Graph 1, illustrated the proportion of different demographic categories. 60.4% of customers are female. Half are Single, and the other half are Non-Single (divorced, separated, married, or widowed).

Regarding education, 43.8% have a high school education, 37.9% attended university, and 8.7% completed graduate school, while 9.6% are in the other/unknown category.

In terms of location, most customers live in small cities (56.5%), followed by big cities (39.9%), with only 3.6% in mid-sized cities.

For occupation, 39.6% are skilled employees or officials, 49.6% are unemployed or unskilled, and the rest work in management, are self-employed, or are highly qualified employees.

Graph 1: Demographic distribution



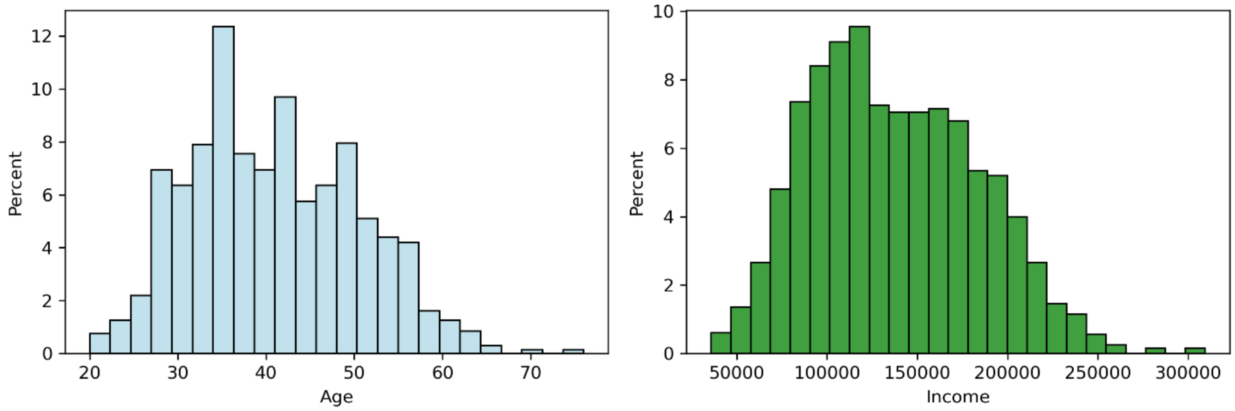
Histogram

In Graph 2, displayed the percentage distribution of customers' age and income.

For age, the majority of customers are between 30 and 50 years old, with the most frequent age group being in the mid-40s, where approximately 12% of the population is concentrated. Very few customers are over the age of 70.

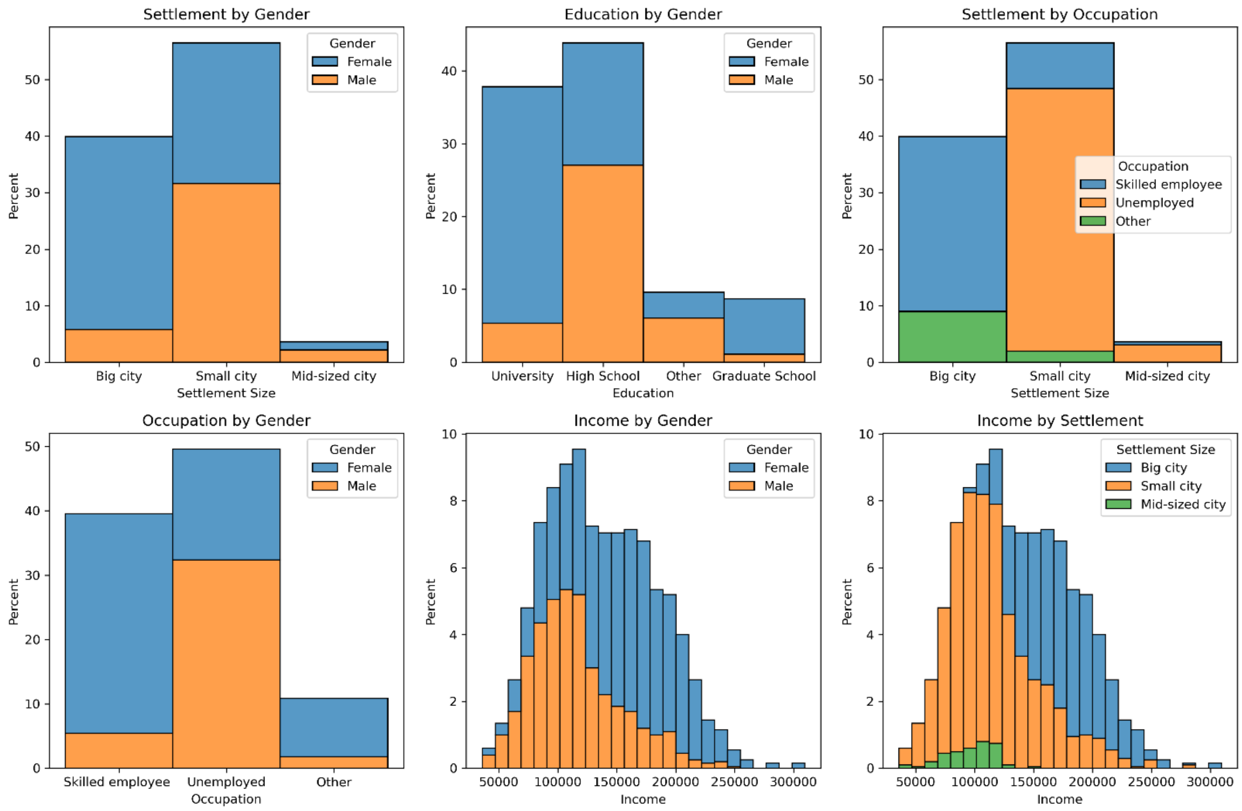
For income, which ranges from $50,000 to $300,000, the majority of customers earn between $50,000 and $150,000. The most common income group falls between $100,000 and $150,000, accounting for around 9-10% of the population. A smaller proportion of customers earn above $150,000.

Graph 2. Age and Income Distribution



In Graph 3, we can observe that most customers live in big or small cities. Big city residents are mostly female, while small city residents have an even gender balance. The majority of those with university or graduate school education are women. Regarding occupations, most small city residents are unemployed, while big city residents are predominantly skilled employees or in other occupations, with most skilled employees being women. Additionally, the histograms show that women and big city residents tend to have higher incomes than their counterparts.

Graph 3. Demographic distributions by features

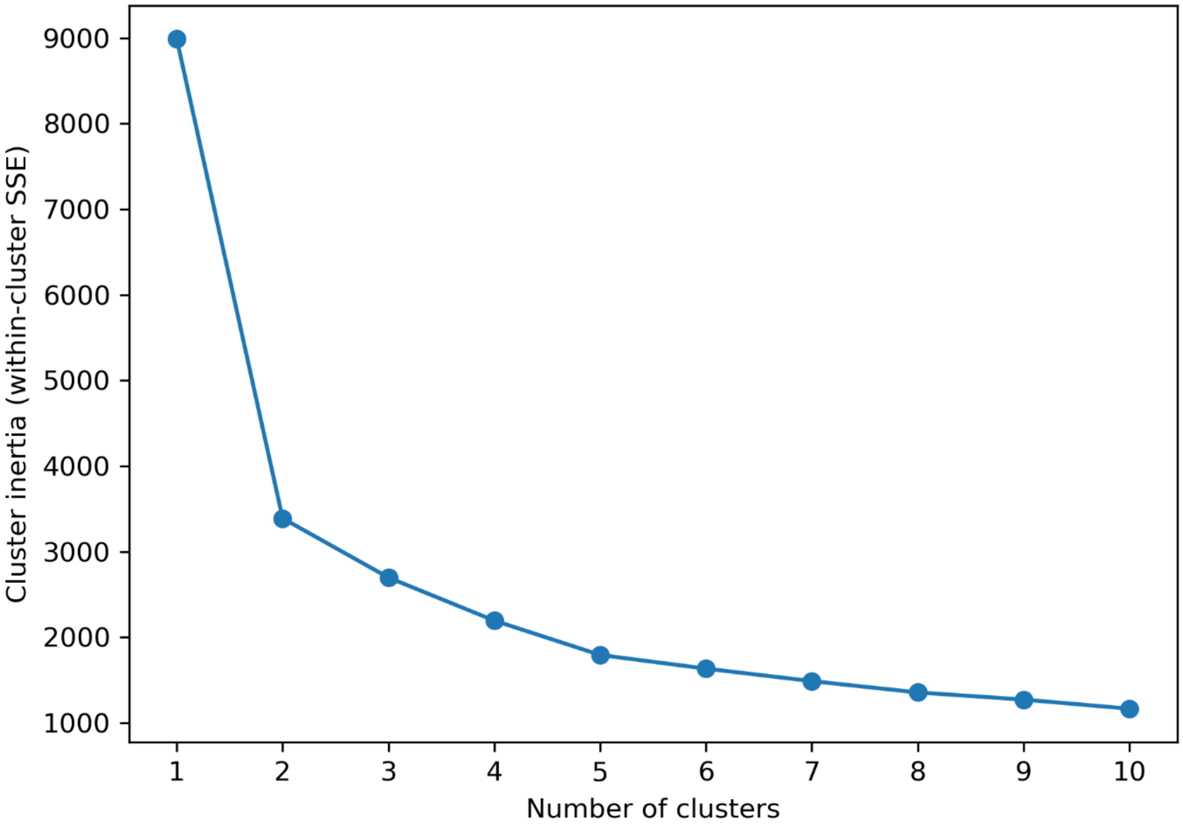


Customer segmentation

Defining number of clusters by elbow method

After standardization, used the Elbow Method to find the optimal number of clusters. In our case, the elbow is clearly visible at 3 clusters (as shown in Graph 4), indicating that 3 is the optimal number of clusters for this dataset.

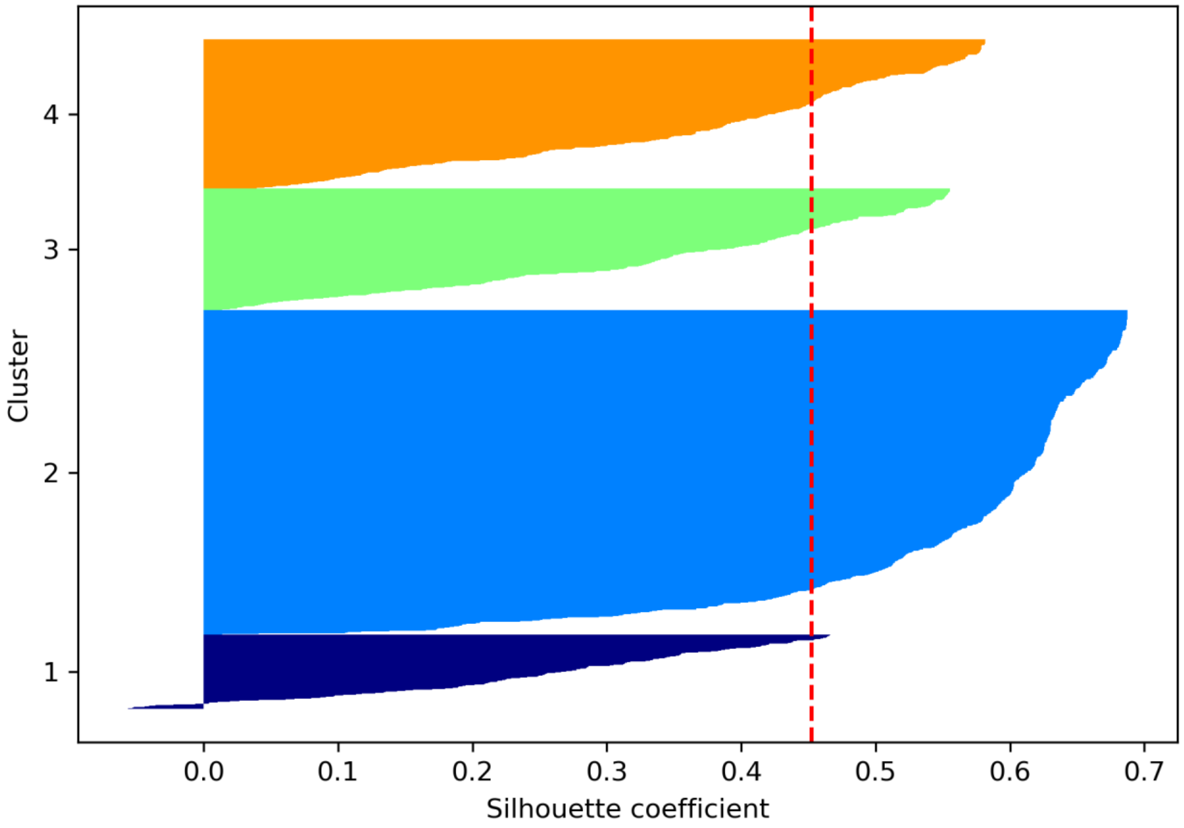
Graph 4. Clusters by Elbow Method



Defining number of clusters using Silhouette plot

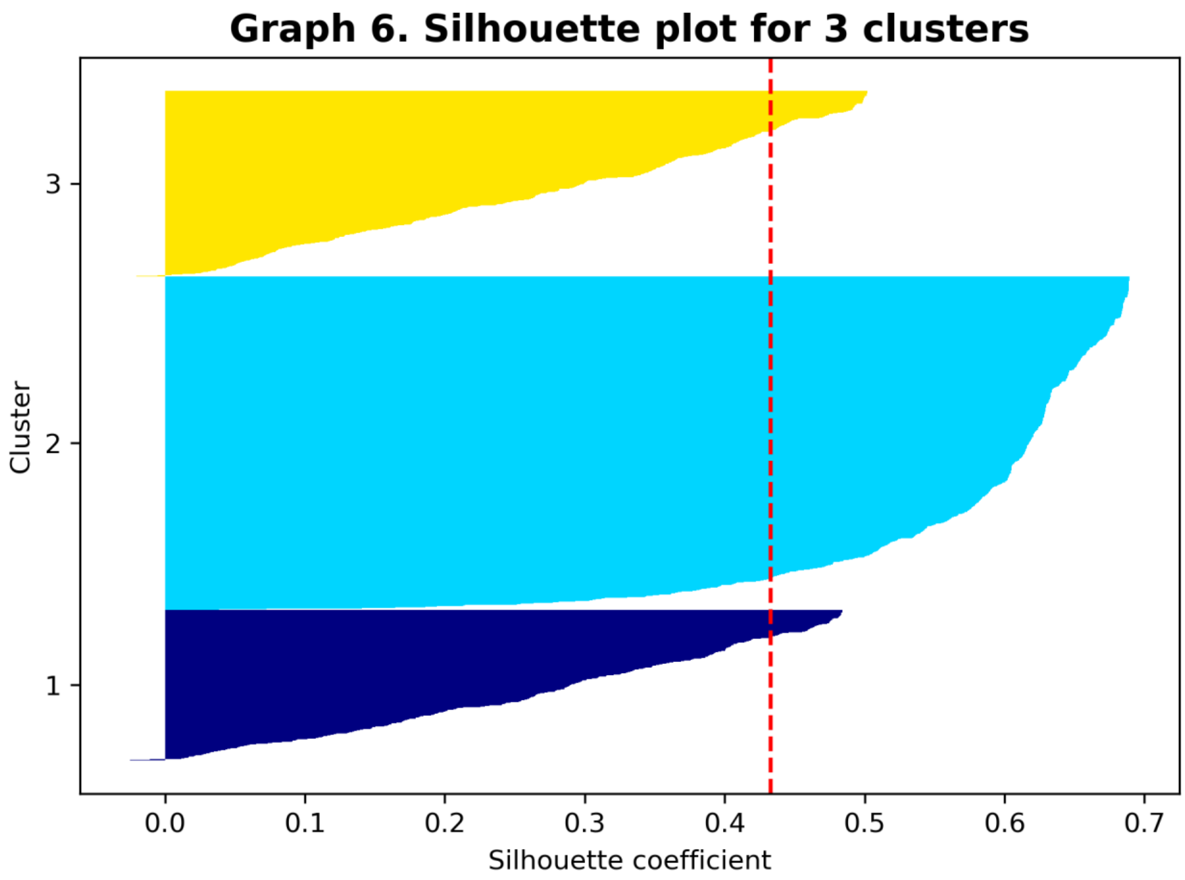
Since the Elbow method indicates that optimal cluster number is 3, we plotted silhouette plot for 4, 3, and 2 clusters.

First, 4 clusters. A silhouette average of 0.45 suggests that the clustering is moderately well-defined (Graph 5).



Graph 5. Silhouette plot for 4 clusters

For 3 clusters, the silhouette average is 0.43 (Graph 6), which is slightly lower than the silhouette average of 0.45 for 4 clusters.



For 2 clusters, the silhouette average is 0.54, which is higher than both the 0.45 for 4 clusters and 0.43 for 3 clusters. The higher silhouette score suggests that the 2-cluster solution provides the best-defined separation among the clusters. Therefore, we will proceed our analysis with 2 clusters.

Estimating clusters using K-means++

Table 3 presents the characteristics of the two clusters identified by the K-means++ algorithm.

Cluster 1 consists of 976 customers. It is characterized by a middle-aged, non-single woman, who lives in a big city and has completed university education. She is a skilled employee with an above-average income, earning approximately $173,460 annually.

Cluster 2 includes 1024 customers. This cluster is represented by a single man in his early 30s, with a high school education. He lives in a small city and is unemployed, earning a below-average income of around $103,256 per year.

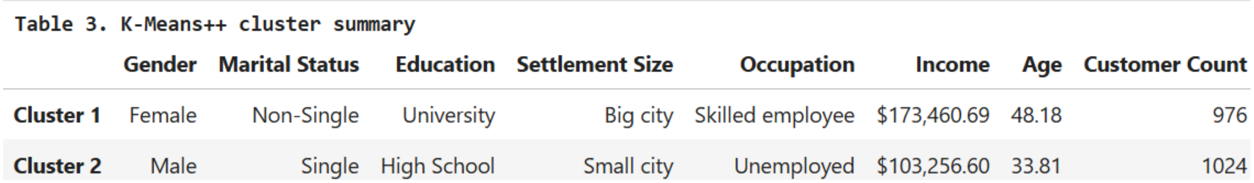


Table 3. K-Means++ cluster summary

	Gender	Marital Status	Education	Settlement Size	Occupation	Income	Age	Customer Count
Cluster 1	Female	Non-Single	University	Big city	Skilled employee	$173,460.69	48.18	976
Cluster 2	Male	Single	High School	Small city	Unemployed	$103,256.60	33.81	1024

Estimating clusters using Agglomerative clustering technique.

Table 4 presents the characteristics of two clusters identified by the K-Means++ algorithm.

Cluster 1 consists of 1163 customers. It is represented by a non-single woman in her late 40s who lives in a big city and with higher education and income (around $168,085).

Cluster 2 includes 837 customers. This cluster is characterized by a single man in his early 30s, who lives in a small city and is unemployed, with a lower education and income (around $95,041).

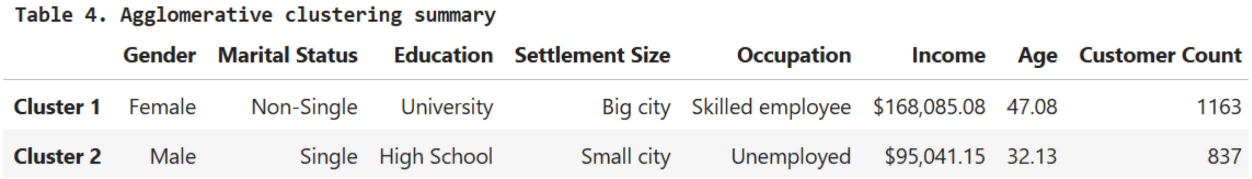


Table 4. Agglomerative clustering summary

	Gender	Marital Status	Education	Settlement Size	Occupation	Income	Age	Customer Count
Cluster 1	Female	Non-Single	University	Big city	Skilled employee	$168,085.08	47.08	1163
Cluster 2	Male	Single	High School	Small city	Unemployed	$95,041.15	32.13	837

For Cluster 1 and Cluster 2, both methods characterize the segment with same profile, with very slight difference in the exact figures of income, age and customer count. Despite small differences, the overall profiles of the clusters from both methods are similar. This consistency supports the segmentation of the customer base into clear groups.

Marketing recommendations

Cluster 1:

- Offer more exclusive and high-end travels and services.
- Offer child friendly travel plans.

Cluster 2:

- Offer affordable packages, discounts, and deals.
- Promote travels and services related to skill development or career growth.
- Referral programs offering discounts or rewards for bringing in new customers can help expand the customer base.
- Location-Specific regional promotions targeting smaller cities.

Conclusion

In this analysis, we began with an exploratory data analysis to examine key statistics from the dataset, including descriptive statistics and visualization. We then standardized numerical variables to ensure they were on the same scale. Using the Elbow Method, we identified 3 as the optimal number of clusters; however, further analysis with silhouette scores suggested that the 2-cluster solution provided the best-defined separation. We proceeded with both K-Means++ and Agglomerative clustering techniques to segment the customers. Both methods produced two clusters with similar profiles. At the end this analysis offered targeted marketing strategies to enhance customer satisfaction and business performance.