I. Introduction

This report identifies customer segments to optimize marketing strategies for a large travel agency. The dataset comprises information on age, gender, income, marital status, education, occupation, and settlement size from 2,000 customers. The methodology includes exploratory data analysis to understand and visualize dataset, followed by clustering techniques using the Elbow Method and Silhouette Plots to determine the optimal number of clusters. K-means++ and Agglomerative Clustering segment customers, with tailored recommendations provided for each segment to enhance engagement.

II. Exploratory Data Analysis

Summary statistics

	Gender	Marital Status	Age	Education	Income	Occupation	Settlement Size
count	2000.0000	2000.0000	2000.0000	2000.0000	2000.0000	2000.0000	2000.000
mean	0.6045	0.5005	40.8235	1.4565	137516.1965	0.6125	0.834
median	1.0000	1.0000	40.0000	1.0000	133004.0000	1.0000	0.000
min	0.0000	0.0000	20.0000	0.0000	35832.0000	0.0000	0.000
max	1.0000	1.0000	76.0000	3.0000	309364.0000	2.0000	2.000

Figure 1: Summary statistics of seven columns in the dataset

The dataset includes 2,000 entries with seven variables covering demographic and socio-economic details. Gender and Marital Status are nominal categorical variables, while Education, Occupation, and Settlement Size are ordinal. Age and Income, the numerical variables, have average values of 40.8 years and $137,516, both exceeding their medians.

Distribution of Age and Income

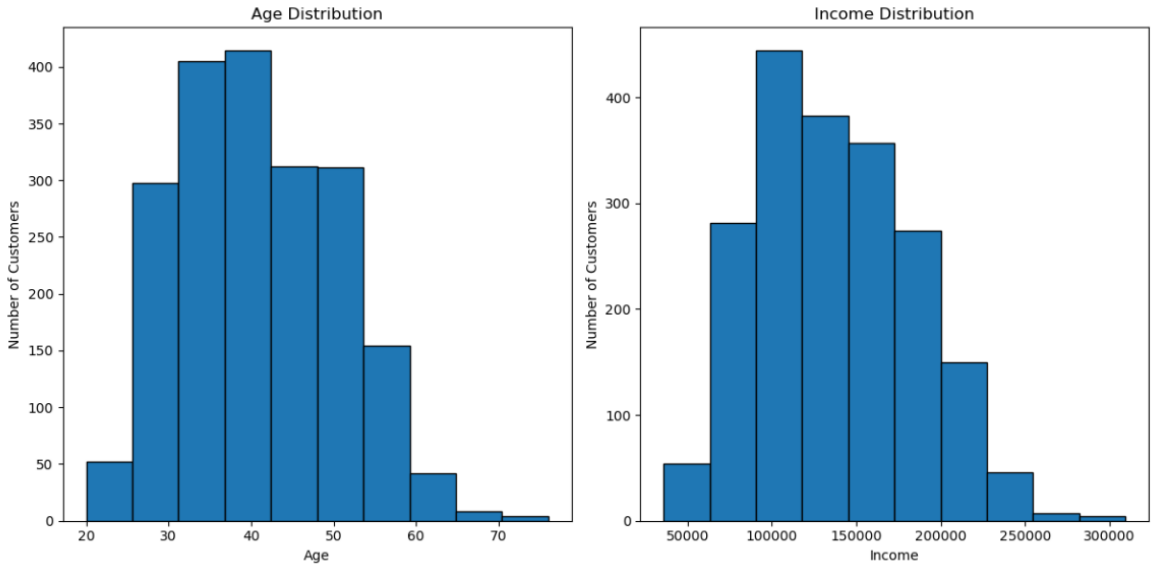

Figure 2: Distribution of Age and Income

Figure 2 shows that the age distribution is symmetrical, peaking around 40 years, with fewer younger (<30) and older (>60) individuals. Income distribution is right-skewed, showing most customers earning $70,000 to $150,000, with fewer high-income earners.

Distribution of categorical variables

Figure 3: Distribution of categorical variables

Figure 3 illustrates that the dataset consists predominantly of females (60.4%), with marital status evenly distributed. Education levels are primarily represented by high school graduates (36.8%) and university graduates (37.9%). Almost half of the customers are unemployed or unskilled (49.8%), and a majority live in small cities (56.1%).

III. Customer Segmentation

1. Standardize numeric variables

The Age and Income columns are standardized to ensure they are on the same scale for accurate customer segmentation analysis (Figure 4)

	Gender	Marital Status	Education	Occupation	Settlement Size	Age_scaled	Income_scaled
0	1	1	2	1	2	-0.192892	-0.150483
1	0	0	1	0	0	-1.250703	-1.238852
2	1	0	0	0	0	-0.616016	-0.659462
3	0	1	2	1	0	1.605387	1.656471
4	1	1	2	1	2	0.441795	0.446623

Figure 4: First five rows of the dataset after standardizing age and income columns

2. Optimal number of customer segments

Elbow Method

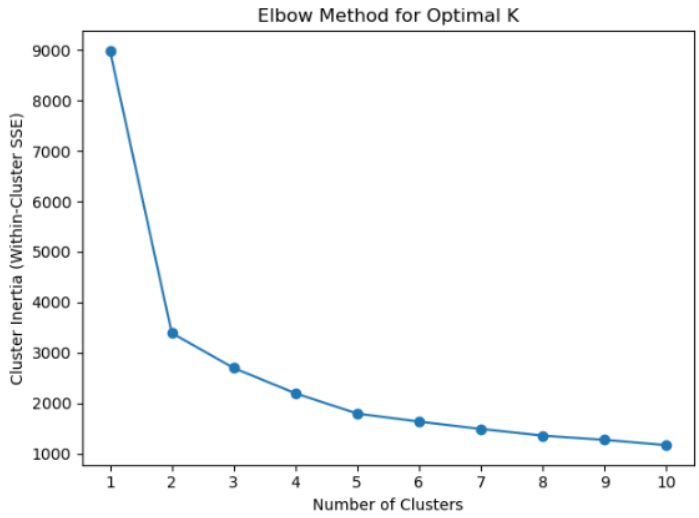

Figure 5: Elbow method for choosing optimal number of clusters

The Elbow method identified two as the optimal number of clusters in Figure 5, where adding more clusters shows minimal performance improvement.

Silhouette Plots

Silhouette plots were used to assess clustering quality, revealing the highest average silhouette score at two clusters, at 0.54 (Figure 6). This reinforces the choice of two clusters as optimal.

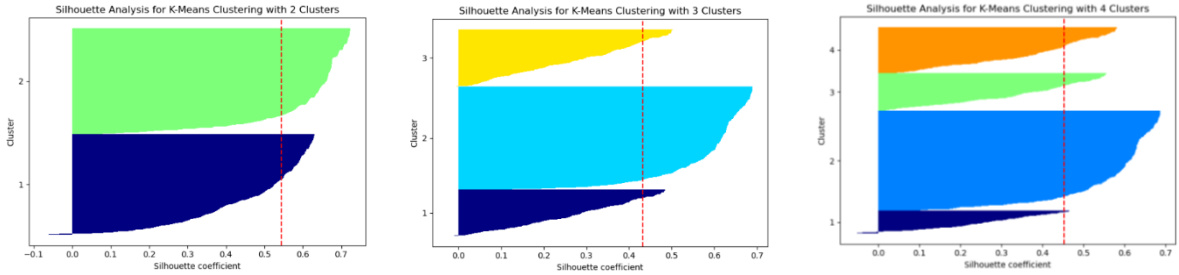

Figure 6: Silhouette Analysis for K-Means Clustering with 2, 3, and 4 Clusters

3. Customer segments

K-means++ and agglomerative clustering techniques were employed to identify two customer segments.

K-means++

This technique identified two clusters as below:

	KMeans++	Gender	Marital Status	Education	Occupation	Settlement Size	Age_scaled	Income_scaled	Number of Customer
0	Cluster 1	0.855533	0.991803	2.102459	1.214139	1.635246	0.777997	0.778479	976
1	Cluster 2	0.365234	0.032227	0.840820	0.039062	0.070312	-0.741528	-0.741987	1024

Figure 7: K-Means++ result

	KMeans++	Gender	Marital Status	Education	Occupation	Settlement Size	Average Age	Average Income	Number of Customer
0	Cluster 1	Female	Non-Single	University	Skilled Employee/Official	Big City	48.18	173469.68	976
1	Cluster 2	Male	Single	High School	Unemployed/Unskilled	Small City	33.81	103248.03	1024

Figure 8: Interpretation of K-Means++ result

Agglomerative clustering

This technique identified two clusters as below:

	Agglomerative	Gender	Marital Status	Education	Occupation	Settlement Size	Age_scaled	Income_scaled	Number of Customer
0	Cluster 1	0.774721	0.856406	1.925193	1.043852	1.384351	0.661907	0.662055	1163
1	Cluster 2	0.367981	0.005974	0.805257	0.013142	0.069295	-0.919710	-0.919916	837

Figure 9: Agglomerative clustering result

	Agglomerative	Gender	Marital Status	Education	Occupation	Settlement Size	Average Age	Average Income	Number of Customer
0	Cluster 1	Female	Non-Single	University	Skilled Employee/Official	Mid-sized City	47.08	168092.73	1163
1	Cluster 2	Male	Single	High School	Unemployed/Unskilled	Small City	32.13	95030.53	837

Figure 10: Interpretation of Agglomerative clustering result

Cluster interpretation of two techniques

	K-Means++	**Agglomerative Clustering**
Cluster 1	Middle-aged, non-single females in their late 40s, employed as skilled employees or officials, with university degrees, average annual incomes of $173,469.68, living in big cities.	Middle-aged, non-single females in their late 40s, employed as skilled employees or officials, with university degrees, earning an average annual income of $168,092.73, living in mid-sized cities.
Cluster 2	Young, single males in their early 30s, unemployed or unskilled, with high school education, earning an average annual income of $103,248.03, residing in small cities.	Young, single males in their early 30s, unemployed or unskilled, with high school education, earning an average annual income of $95,030.53, residing in small cities.

Comparison of two techniques

The customer segments identified by both techniques exhibit significant similarities, indicating overlap across clusters, with income levels, city sizes, and customer distribution variations.

	K-Means++	**Agglomerative Clustering**
Similarities in Cluster 1	middle-aged, non-single females in their late 40s, with university degrees and skilled employment	
Similarities in Cluster 2	young, single males in their early 30s, with high school education and unemployed or unskilled, residing in small cities	
Difference in Cluster 1		
Located in	Big cities	Mid-sized cities
Average annual income	$173,469.68	$168,092.73
Number of customers	976	1163
Difference in Cluster 2		
Average annual income	$103,248.03	$95,030.53
Number of customers	1024	837

IV. Recommendations

Cluster 1: Middle-aged, non-single females in their late 40s, employed as skilled employees or officials, with university degrees, average annual incomes of $173,469.68, and living in big cities.

No	**Marketing Strategy**	**Marketing technique**
1	Offer luxury and unique travel experiences	Develop premium travel packages featuring luxury resorts, exclusive tours, and culturally enriching experiences highlighting sophistication.
2	Promote wellness-focused travel options	Create packages centered on relaxation, incorporating spa retreats and yoga vacations that resonate with their commitment to health and well-being.
3	Develop family-friendly travel packages	Design packages that include kid-friendly activities, accommodations, and multi-generational experiences.

Cluster 2: Young, single males in their early 30s, unemployed or unskilled, with high school education, earning an average annual income of $103,248.03, and residing in small cities.

No	Marketing Strategy	Marketing technique
1	Provide budget-friendly travel packages	Design packages that combine services such as transportation and accommodation, using special deals to reduce financial barriers.
2	Provide adventure travel packages	Feature affordable destinations that offer adventure experiences such as hiking and cave exploration.
3	Leverage social media and influencer marketing	Run campaigns collaborating with travel bloggers, creating content showcasing affordable and adventurous experiences tailored to young travelers.

V. Conclusion

This report provides a detailed analysis of customer segmentation, including exploratory data analysis, segmentation techniques, and personalized recommendations. The Elbow methods and Silhouette plots indicate that the dataset can be divided into two segments. K-means++ and agglomerative clustering identify Cluster 1 as middle-aged, non-single females in their late 40s, earning an average annual income of $173,469.68 and living in big cities. Cluster 2 consists of young, single males in their early 30s, with an average annual income of $103,248.03, residing in small cities. Recommendations focus on luxury, wellness, and family travel for Cluster 1, and budget-friendly, adventure options for Cluster 2.

Word count: 938