Wordcount: 957

1. Introduction

The travel agency faces challenges in implementing targeted marketing and engaging its diverse customer base. It has collected a large database with different customer information but lacks clear insights into its customer segments. This report aims to identify distinct customer segments within a dataset of 2,000 customers, enhancing the agency's marketing strategies and improving customer satisfaction. To achieve this, clustering techniques will be employed in Python to analyse key customer segments, then recommendations will be provided accordingly.

2. Exploratory Data Analysis

The dataset has 2000 records with no missing values. Numerical features are *Age* and *Income*, while ordinal categorical features are *Gender*, *Marital Status*, *Education*, *Occupation*, and *Settlement Size*.

	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
50%	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

Fig 1: Basic summary statistic

From Table 1, average customer’s age is 40.8, while average income is 137,515.2 USD/year. I could also find the maximum income (309,364 USD/year) and minimum income (35,832 USD/year).

Two histograms will be used to describe data distribution of the numerical columns. From histogram of age, most customers are between 30 and 50 years old.

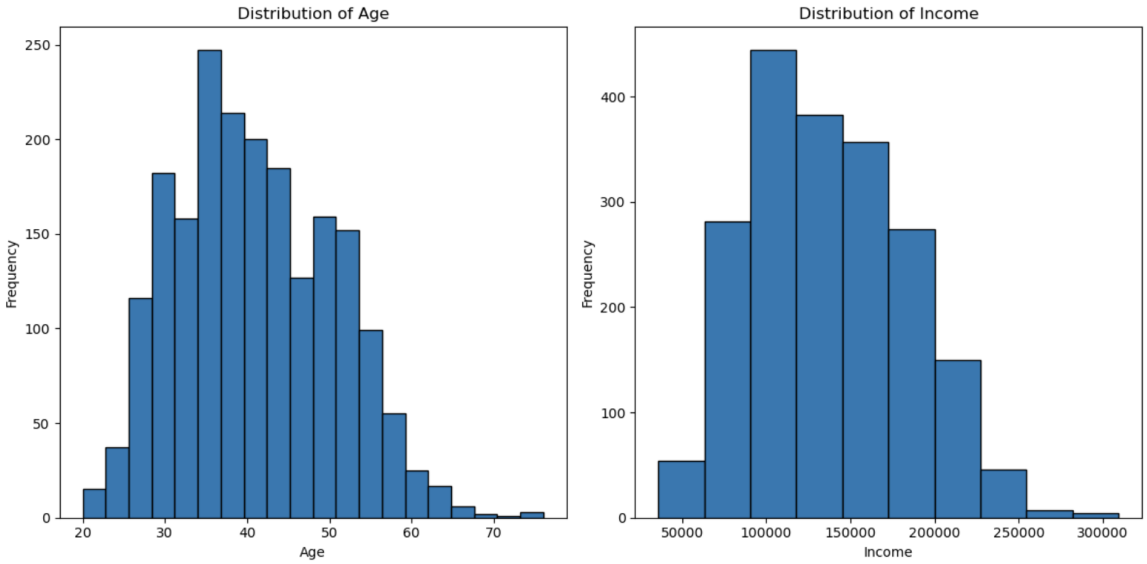


Fig 2: Age and Income distribution of 2000 customers

The histogram of income is right-skewed. Most customers earn between 50,000-150,000 USD/year. The number of customers earning higher incomes decreases significantly as income range increases.

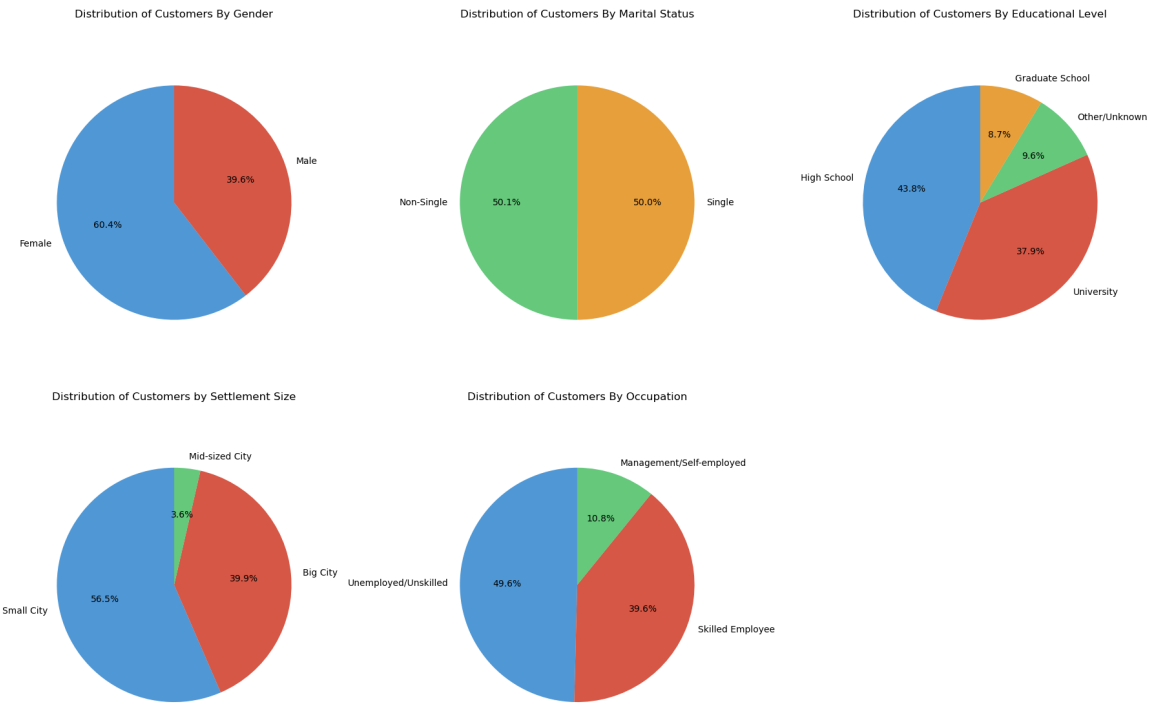


Fig 3: Proportion of Customers by Gender, Marital Status, Educational Level, Settlement Size, and Occupation

From the pie charts above, the majority of customers are female, while less than 40% are males. There is almost a balance between single and non-single customers.

Most customers either completed high school or attended university, while very few have graduate-level education. A majority of customers live in smaller cities, followed by those in big cities, while only a few reside in mid-sized cities.

3. Customer segmentation

Firstly, numerical features like age and income will be standardised to ensure consistency in clustering analysis. After using the Elbow Method and three appropriately chosen Silhouette Plots with the number of clusters equal to 2,3,4 respectively, I found that the optimal of customer segments (clusters) is 2 (as can be seen in the plot). Optimal number will be the point where adding more cluster stops making big improvement.

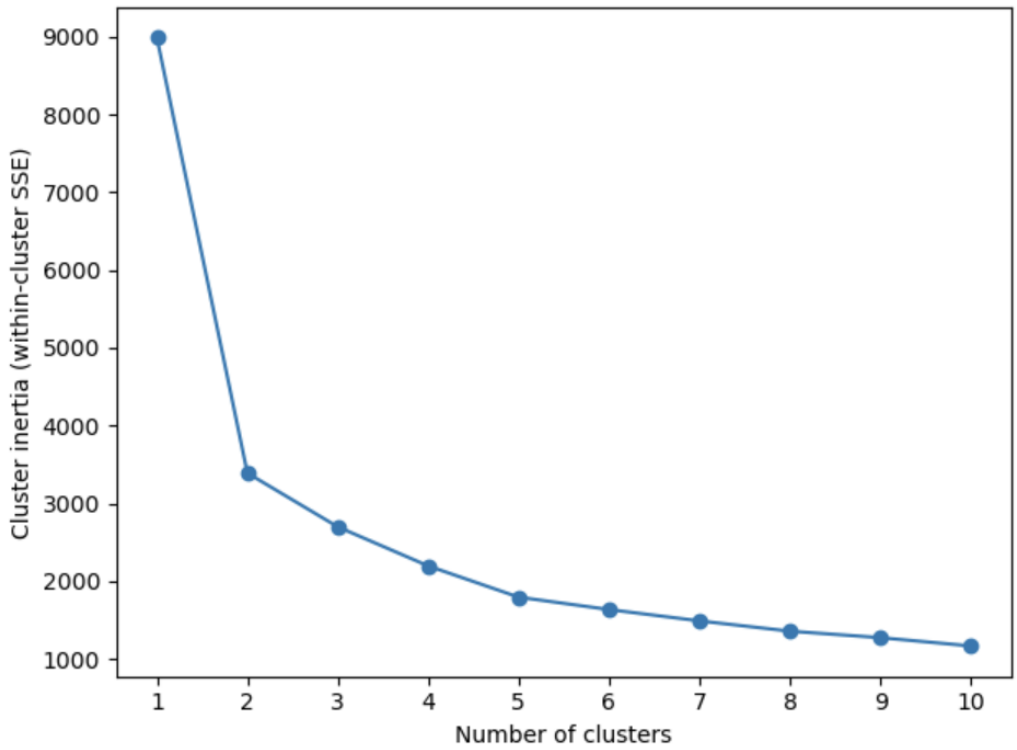
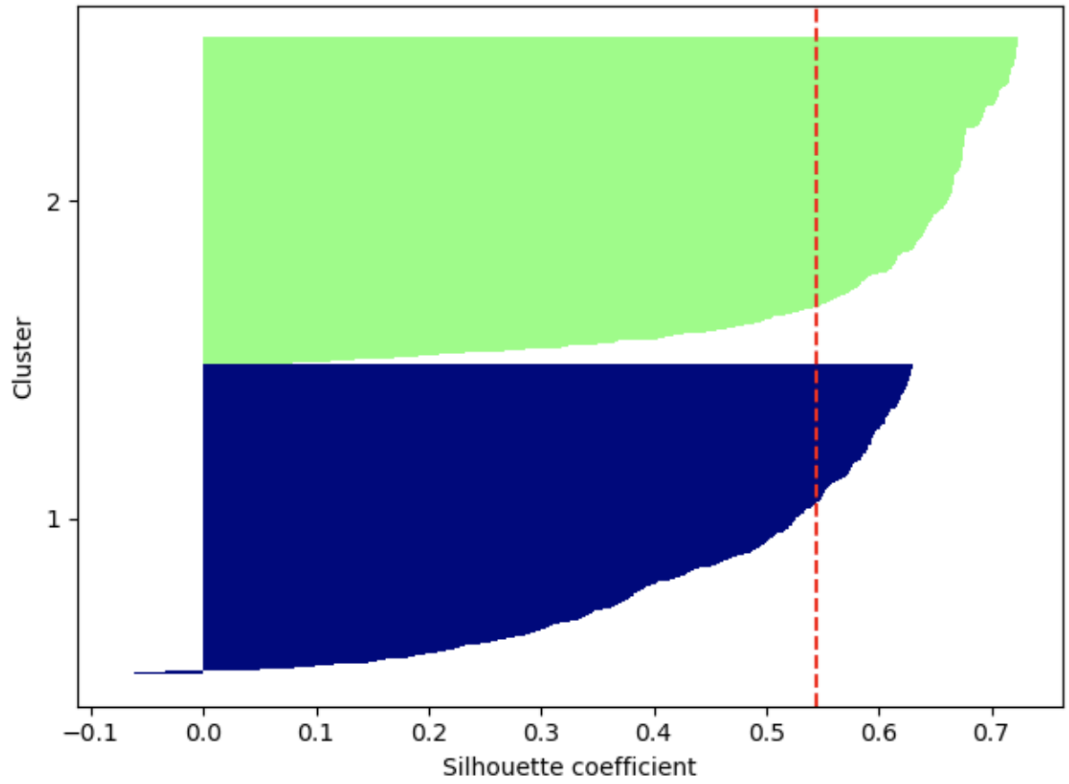
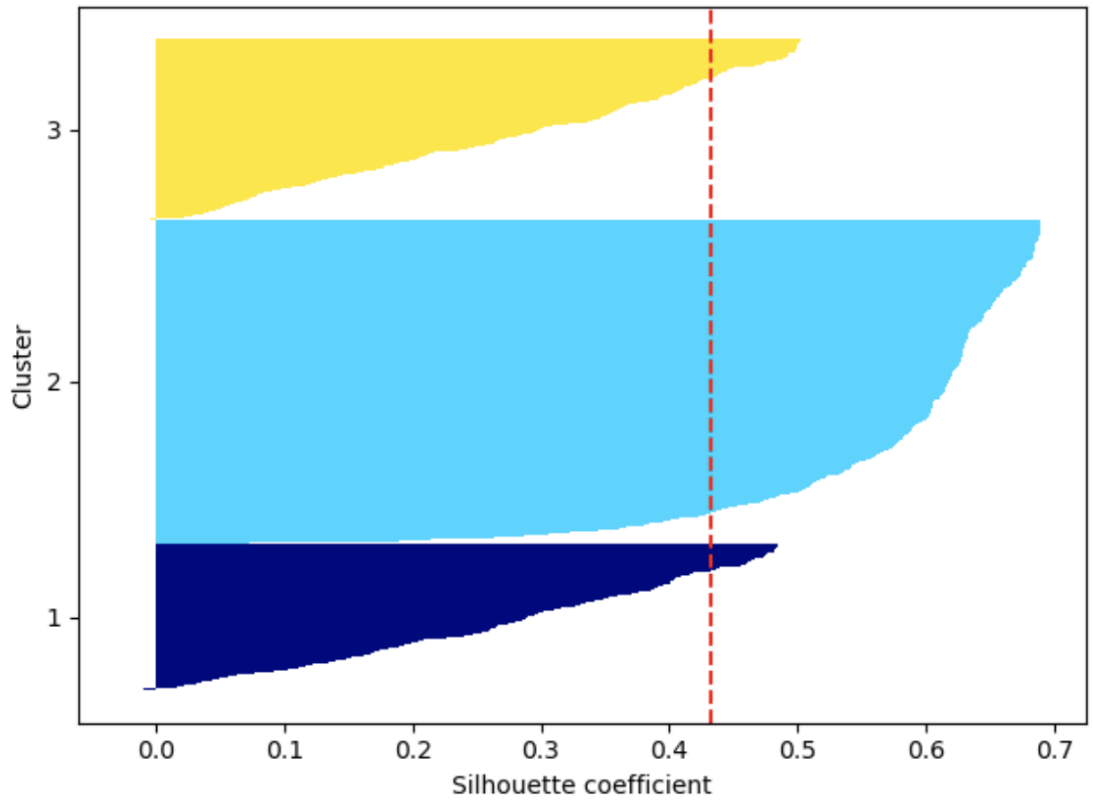


Fig 4: Elbow Method

```
silhouette_avg: 0.54
```




```
silhouette_avg: 0.43
```



```
silhouette_avg: 0.45
```

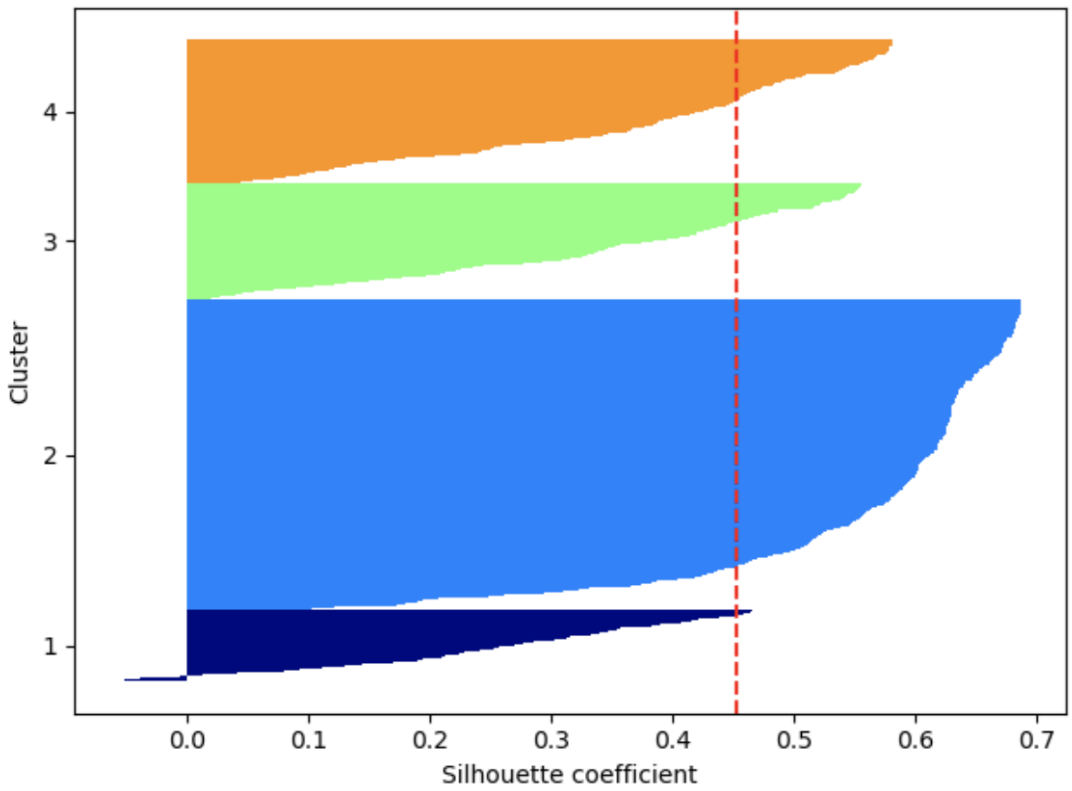


Fig 5: Average Silhouette coefficient when n = 2 is the highest compared to n=3 and n=4, so the optimal number of clusters is 2.

After that, I used KMeans++ and Agglomerative clustering to group customers into different segments based on their similarity. The results, with 2 segments, show average values for each feature. I converted back the scaled values of age and income, so real values will be presented.

3.1. KMeans ++ clustering

```
Combined Cluster Centers and Customer Counts (KMeans++) with Real Age and Income:
```

	KMeans++	Gender	Marital Status	Education	Occupation	Settlement Size	Real Age	Real Income	Number of Customers
0	1	0.855533	0.991803	2.102459	1.214139	1.635246	48.180118	173469.679044	976
1	2	0.365234	0.032227	0.840820	0.039062	0.070312	33.811723	103248.033451	1024

Fig 6: Result of KMeans++ clustering technique

	KMeans++	Gender	Marital Status	Education	Occupation	Settlement Size	Real Age	Real Income	Number of Customers
0	1	Female	Non-Single	University	Skilled Employee/Official	Big City	48.180118	173469.679044	976
1	2	Male	Single	High School	Unemployed/Unskilled	Small City	33.811723	103248.033451	1024

Fig 7: Interpretation

Cluster 1: This segment comprises 976 customers, largely non-single females, mostly university graduates or above. They hold skilled jobs and predominantly reside in big cities. Their average age is 48.18, with an annual income of 173,469.68 USD (above average).

Cluster 2: This segment includes 1,024 customers, mainly single males with lower education levels, likely high school graduates or less. They tend to be unemployed or in unskilled jobs and predominantly live in small cities. Their average age is 33.81, with an annual income of 103,248.03 USD (which is below average).

3.2. Agglomerative Clustering

```
Combined Cluster Centers and Customer Counts (Agglomerative) with Real Age and Income:
```

	Agglomerative	Gender	Marital Status	Education	Occupation	Settlement Size	Real Age	Real Income	Number of Customers
0	1	0.774721	0.856406	1.925193	1.043852	1.384351	47.082390	168092.726774	1163
1	2	0.367981	0.005974	0.805257	0.013142	0.069295	32.126858	95030.527792	837

Fig 8: Result of Agglomerative clustering technique

	Agglomerative	Gender	Marital Status	Education	Occupation	Settlement Size	Real Age	Real Income	Number of Customers
0	1	Female	Non-Single	University	Skilled Employee/Official	Mid-sized City	47.082390	168092.726774	1163
1	2	Male	Single	High School	Unemployed/Unskilled	Small City	32.126858	95030.527792	837

Fig 9: Interpretation

Cluster 1: This segment has 1,163 customers, largely non-single females with a high education level, likely university graduates or over. They typically hold higher-skilled jobs and reside mainly in mid-sized or big cities. Their average age is 47.08, and average annual income is $168,092.73, above the average.

Cluster 2: This segment includes 837 customers, predominantly single males with lower education levels, likely high school graduates or less. They are often unemployed or in unskilled jobs and primarily live in small cities. Average age is 32.13, with an annual income of $95,030.53, below the average.

3.3. Comparison

The clusters identified by both techniques overlap significantly by identifying 2 customer profiles. Segment 1 consists of older, more affluent, well-educated urban females, mostly non-single, working in higher-skill jobs. Segment 2 includes younger, less affluent males, predominantly single, with lower education levels, unemployed or working in unskilled jobs and living in smaller cities.

Main difference lies in customer distribution. KMeans++ shows a more even spread, while Agglomerative clustering shows a larger difference. Additionally, income level in segment 2 are slightly lower in Agglomerative clustering compared to KMeans++. Settlement size also varied in cluster 1 between both techniques.

4. Recommendations

For the first segment, the company should implement targeted marketing campaigns focused on luxury experiences, such as 5-star accommodations, and wellness retreats that cater to their higher income and desire for high-quality relaxation. Personalized marketing through targeted ads or email campaigns would be effective.

To capitalize on this segment's willingness to spend, the company could promote additional services like premium travel insurance through emails. Utilizing social media platforms such as Facebook, and LinkedIn for visually engaging content will capture their attention, emphasizing luxury family-oriented trips that align with their lifestyles.

With segment 2, the agency should focus on promoting affordable travel packages such as adventure trips, and short getaways. Social media marketing, including Instagram, TikTok and Facebook (which are popular with younger individuals) and influencer collaborations could be highly effective for this segment, given their younger age and digital savviness. This segment responds well to dynamic pricing and time-limited offers, which can be promoted through targeted email and social media ads.

5. Conclusion

To sum up, this report has provided data visualization and performed customer segmentation with machine learning techniques (KMeans++ and Agglomerative). From these insights, this analysis provides the agency with insights that can drive targeted marketing campaigns, improve customer satisfaction, retention and conversion rate with 2 distinct segments, ultimately contributing to its growth.