CUSTOMER SEGMENTATION REPORT

Tony Lu

Contents

1. Introduction

This report aims to segment the customer sample population provided in accompanying dataset by dividing customers into groups based on common characteristics through the use of k-means and agglomerative clustering methods. Customer segmentation allows for targeted marketing techniques according to cluster characteristics.

2. Exploratory Data Analysis

An initial investigation is performed on the overall data to identify their characteristics and discover any patterns that may be present.

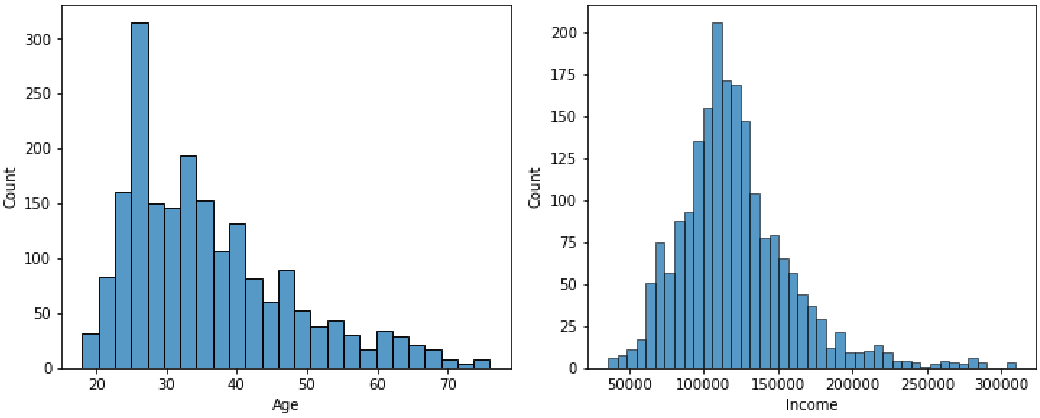

Figure 2.1 – Numerical Variables

From the histograms in Figure 2.1, it can be determined that the distribution of both numerical variables age and income are skewed right, therefore comparisons between the central tendency of distributions will be made with the medians of both variables. The median age is 33 and the median income is $115,548.50 in the data.

Age and income have dissimilar scales; with age ranging between 18 and 76 and income ranging between $35,832.00 and $309,364.00.

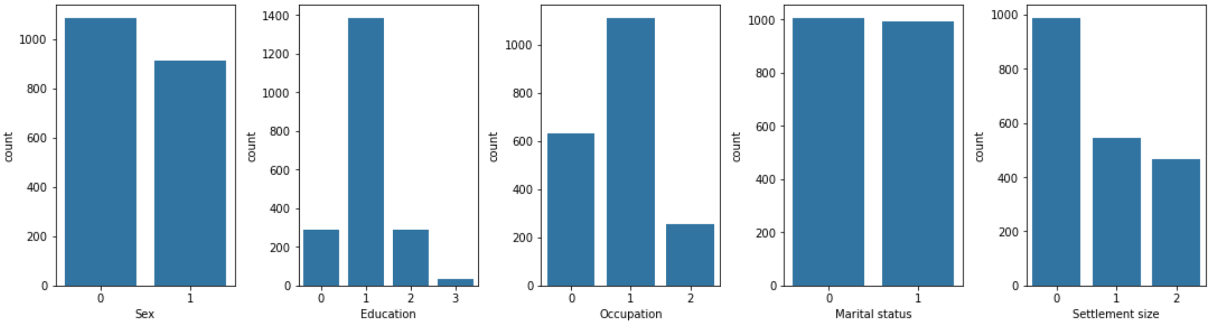

Figure 2.2 – Categorical Variables

From the count charts Figure 2.2, the distributions of the categorical variables can be observed (*see Appendix 1 for legend*).

It can be observed that there are more male customers than female in the dataset. Single customers marginally outnumber non-single customers. The majority of the customers in the dataset live in small cities, followed by mid-sized cities and large cities respectively. The majority of the customers in the dataset have completed high school, a smaller proportion having completed university level, and a very small proportion having completed graduate school. However, there sizeable proportion of customers whose education level is unknown. The majority of the customers in the dataset are skilled employees, followed by unemployed/unskilled and management/self-employed.

3. Clustering Segmentation

This analysis will be applying the K-means and Agglomerative clustering methods. These methods are based on Euclidian distance and will perform poorly with discrete values under the categorical variables. Therefore the variables *Age* and *Income* will be used to form clusters.

Number of Clusters: To determine the optimal number of clusters, the elbow method is employed.

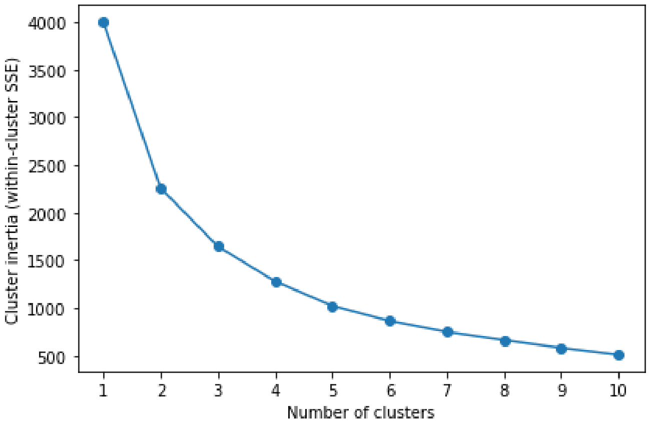

Figure 3.1 – Cluster Inertia

From Figure 3.1, it can be estimated the optimal number of clusters to be 3, as change in cluster inertia drops significantly after 3 clusters.

Scaling: As both clustering methods are reliant on Euclidian distance, the dissimilarity in the scales of the variables (age and income) give them differing weightings. Therefore, the variables are standardized to equal variance, so they are weighted equally with respect to each other.

Clustering and Analysis:

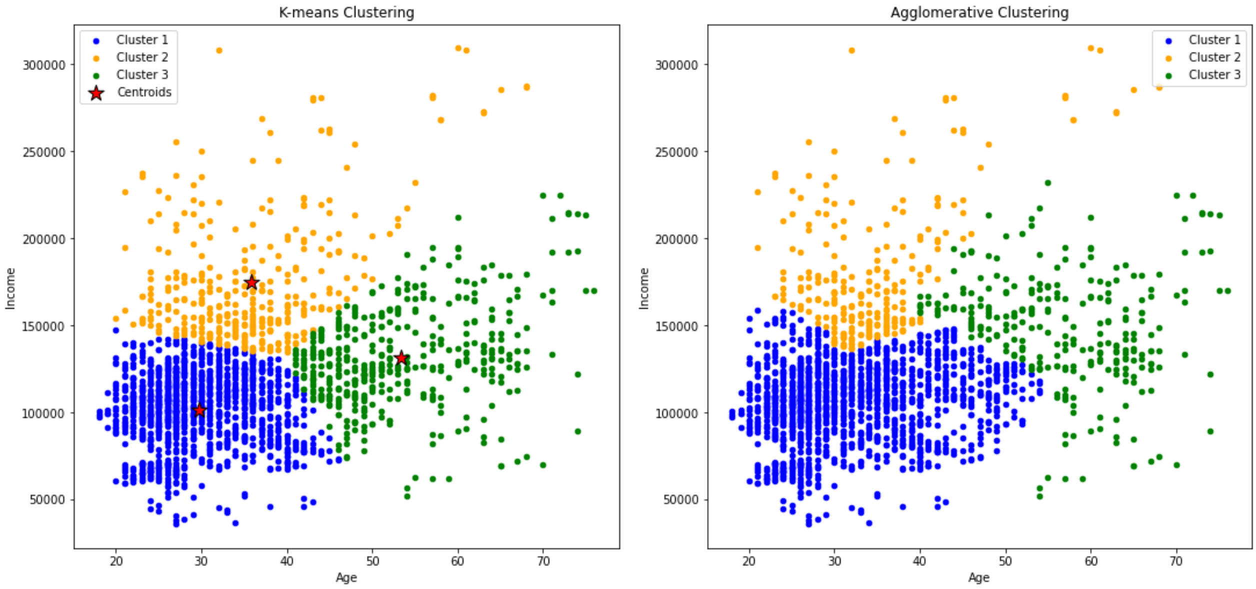

Figure 3.2 – Cluster Plots

Figure 3.2 provides a visual representation of the 3 clusters for both methods employed. Relative cluster size is consistent across methods. Cluster 1 is the largest cluster (K-means = 1212, Agglomerative = 1468), cluster 2 in the middle (K-means = 434, Agglomerative = 276) and cluster 3 the smallest (K-means = 354, Agglomerative = 256).

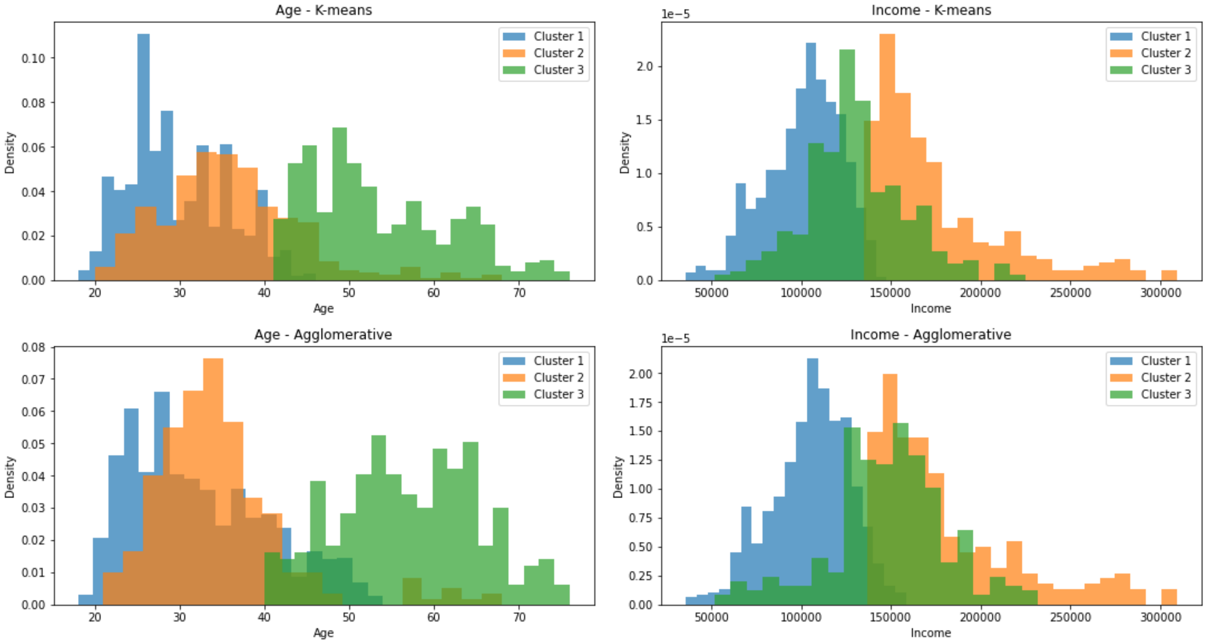

Figure 3.3 – Numerical Variable Distributions

From Figure 3.3, the varying distribution of each cluster can be observed.

Regarding clusters formed by k-means, Cluster 1 occupies the lower age and income ranges, with median age of 28 and median income of $104,292.00. Cluster 2 occupies a relatively higher age range than Cluster 1, and the highest income range, with median age of 35 and median income of $162,483.00. Cluster 3 occupies the highest age range and a middling income range, with a median age of 51 and median income of $127,788.50.

Despite variances in distribution and range in clusters formed by agglomerative clustering, the pattern holds across the methods.

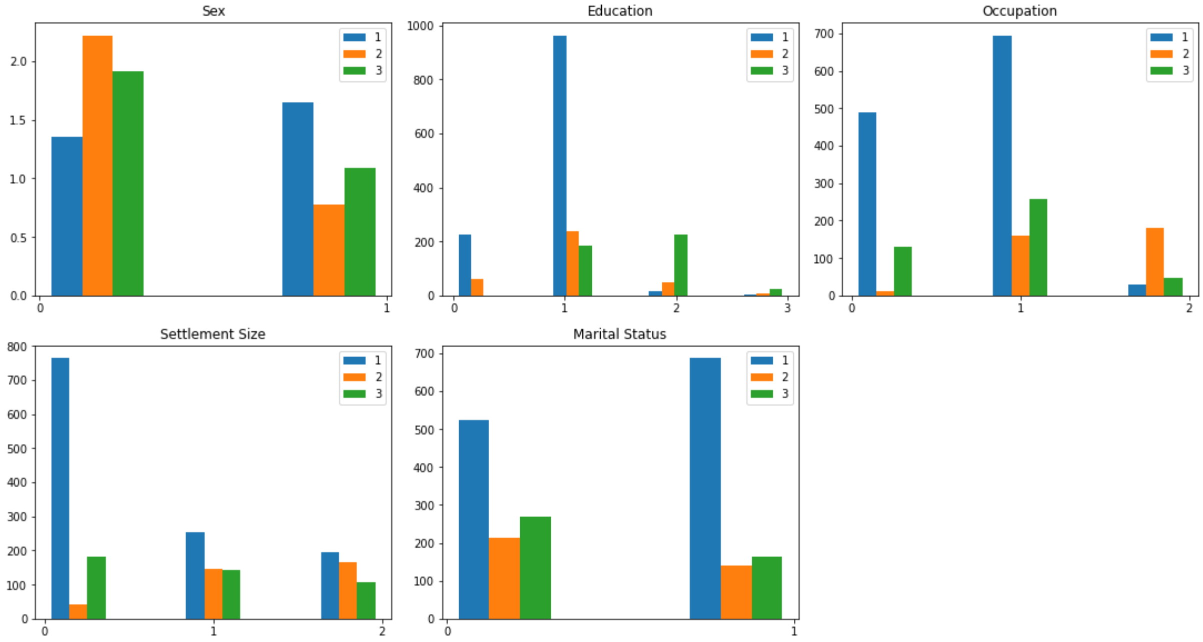

Figure 3.4 – Categorical Distributions (K-means)

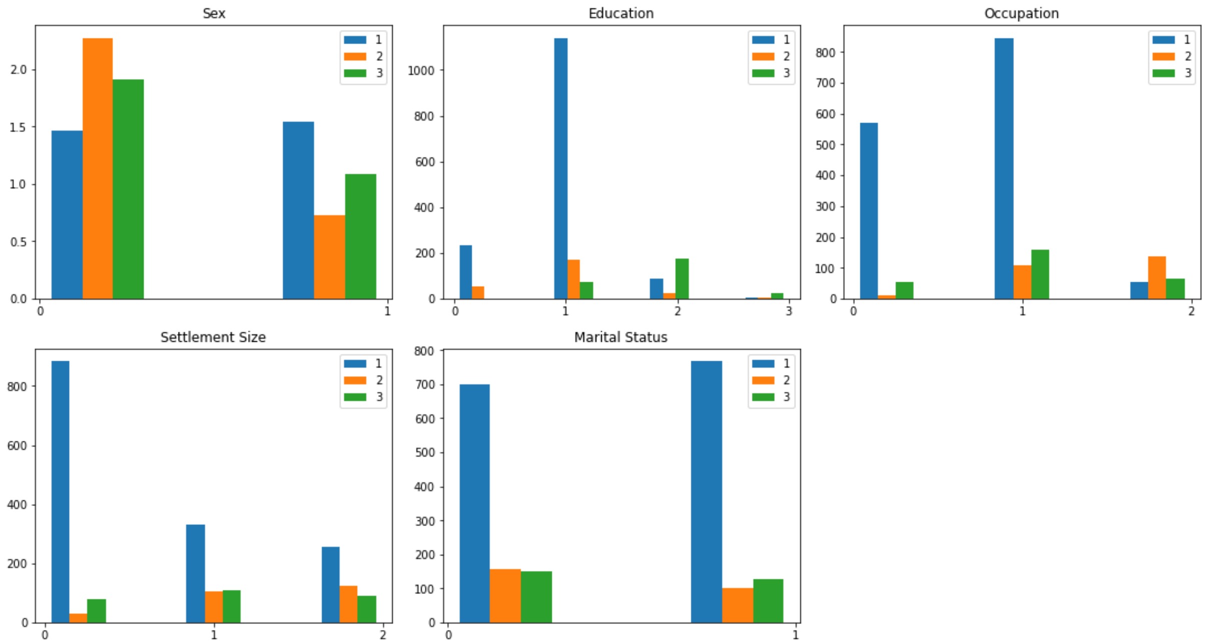

Figure 3.5 – Categorical Distributions (Agglomerative)

From Figure 3.4 and 3.5, the characteristic for each cluster from both methods are observable (*see Appendix 1 for legend*).

The general relative pattern of distribution and proportions for each variable between the clusters remain the same across both methods. Therefore the characteristics of each clusters remains the same.

Cluster	Characteristics
1	Cluster 1 comprises of younger and low-income customers. This is the only cluster that is majority female. The majority of customers in this cluster have a high school level of education (aligning with their younger age) and prefer live in smaller cities. These customers are majority skilled employees but have also the highest rate of unskilled or unemployed. This is also the only cluster with majority non-single marital status. This is the largest cluster by population.
2	Cluster 2 comprises of middle age high income customers. This cluster is majority male and single. Customers in this cluster have similar distribution of education to cluster 1 but prefer living in big cities. The majority of these customers are in managerial positions or self-employed (aligning with their high income).
3	Cluster 3 comprises of older and middle-income customers. This cluster is also majority male and single. This cluster of customers have the highest level of education (aligning with their older age) and prefer live in mid-cities. These customers are majority skilled employees. This is the smallest cluster by population.

4. Recommendation

Cluster	Marketing Recommendation
1	Promote low-cost and discount items through platforms that younger female audiences would engage with (i.e. social media). These items should appeal to married couples, and separated people. These items should not be academic in nature and should appeal to/align with small city lifestyle.
2	Promote expensive items through platforms that middle age high income single men would engage with. These items should not be academic in nature and should appeal to/align with big city lifestyle. These items can also be business related in nature (self-employed customers).
3	Promote general items through platforms that older single men would engage with. These items may be academic in nature and should appeal to/align with mid-city lifestyle.

5. Conclusion

In conclusion, through the use of k-means and agglomerative clustering methods, 3 customers segment have been formed based on common characteristics, allowing for targeted marketing techniques according to such characteristics.

Appendix 1 – Data Legend

Variable	Value	Description
Sex		Biological sex (gender) of a customer.
	0	male
	1	female
Marital status		Marital status of a customer.
	0	single
	1	non-single (divorced / separated / married / widowed)
Education		Level of education of the customer
	0	other / unknown
	1	high school
	2	university
	3	graduate school
Occupation		Occupation of the customer.
	0	unemployed / unskilled
	1	skilled employee / official
	2	management / self-employed / highly qualified employee / officer
Settlement size		The size of the city that the customer lives in.
	0	small city
	1	mid-sized city
	2	big city