BUSA3020

Individual Assignment
Customer Segmentation

Student Name: Minh Nguyet (Caitlyn) Ngo

Student ID: 47677813

Word count: 1126 words

Table of Contents

1. Introduction:

This report is to perform customer segmentation analysis for a supermarket chain using clustering techniques in Python based on the provided dataset. This dataset has basic information of 4000 customers including gender, marital status, education, income, etc., collected through loyalty cards that they use at checkout. At the end of this report, actionable insights for management and marketing strategies will be provided regarded on analysis to improve customer satisfaction.

2. Exploratory Data Analysis:

a. Overview:

	Gender	Marital Status	Education	Settlement Size	Occupation	Income	Age
count	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000	4000.000000
mean	0.489500	0.510500	1.708250	1.090000	1.200750	134353.792250	39.946250
std	0.499952	0.499952	1.024155	0.869246	0.526326	48533.567076	10.269724
min	0.000000	0.000000	0.000000	0.000000	0.000000	35832.000000	18.000000
25%	0.000000	0.000000	1.000000	0.000000	1.000000	97815.250000	32.000000
50%	0.000000	1.000000	1.000000	1.000000	1.000000	122607.000000	38.000000
75%	1.000000	1.000000	3.000000	2.000000	2.000000	165920.250000	47.000000
max	1.000000	1.000000	3.000000	2.000000	2.000000	309364.000000	76.000000

The table above shows the descriptive statistics of this dataset, which contains the basic information of 4000 customers. The mean section shows that the customers' average age is nearly 40 years old, and their average income is at around 134,353. While the min, max, and 25%, 50%, and 75% columns show the distribution of the data, which enables to visualise and understand the dataset better.

b. Correlation:

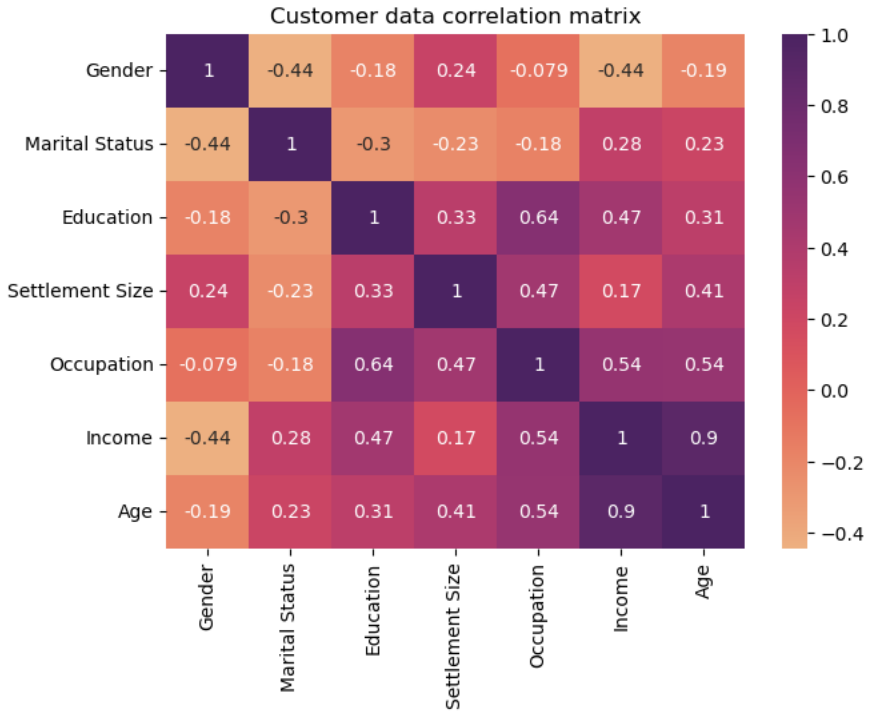

Figure 1. Correlation heatmap

The picture of heatmap shows that if the colour is more purple, the these correspond features will have more strong *positive* relationship (positive correlation) with each other. On the other hand, if the colour becomes lighter, the relationship will be weaker, and if it falls to less than zero, than it becomes *negative* relationship (negative correlation). For instance, "Age" and "Income" have a very strong positive correlation, while "Income" and "Gender" have a negative correlation.

c. Data visualisation:

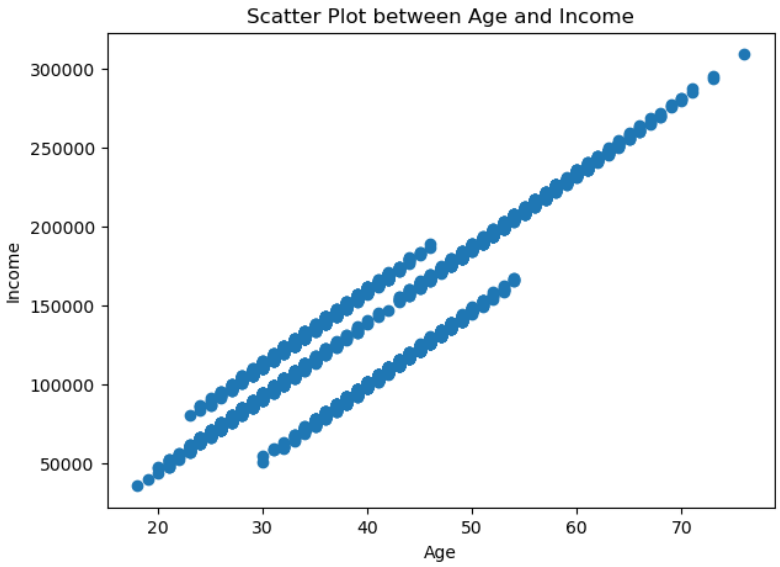

Figure 2. Scatter plot between Age and Income

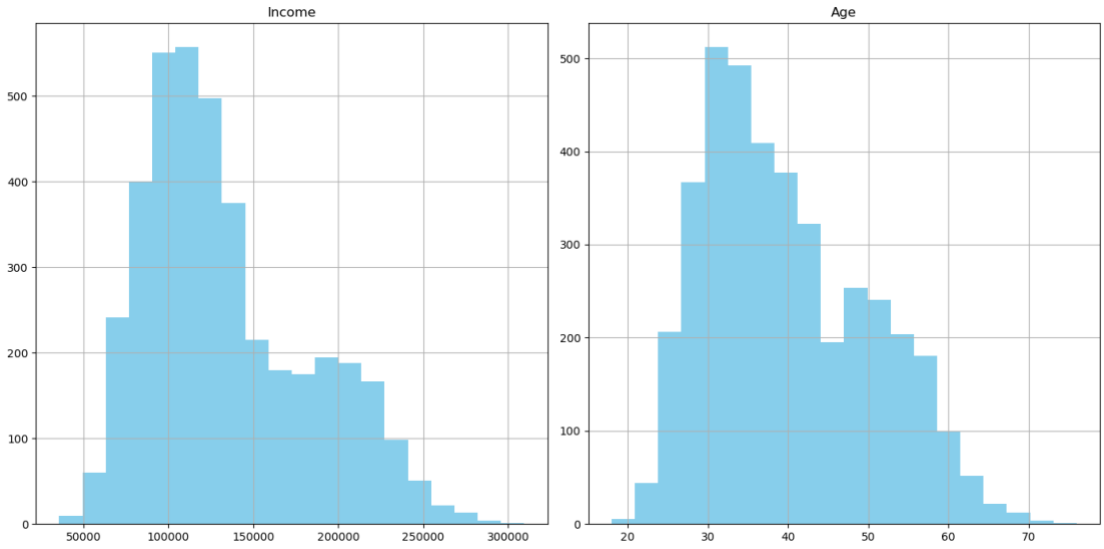

Figure 3. Income & Age distribution charts

These two plots (figure 3) show the relationship between "Age" and "Income" clearer. It is obvious that although there are some exceptions, most of the elder customers have higher income than younger customers in this dataset. Moreover, figure 2 shows that age and income have positive relationship when it shows the higher the age, the higher the income.

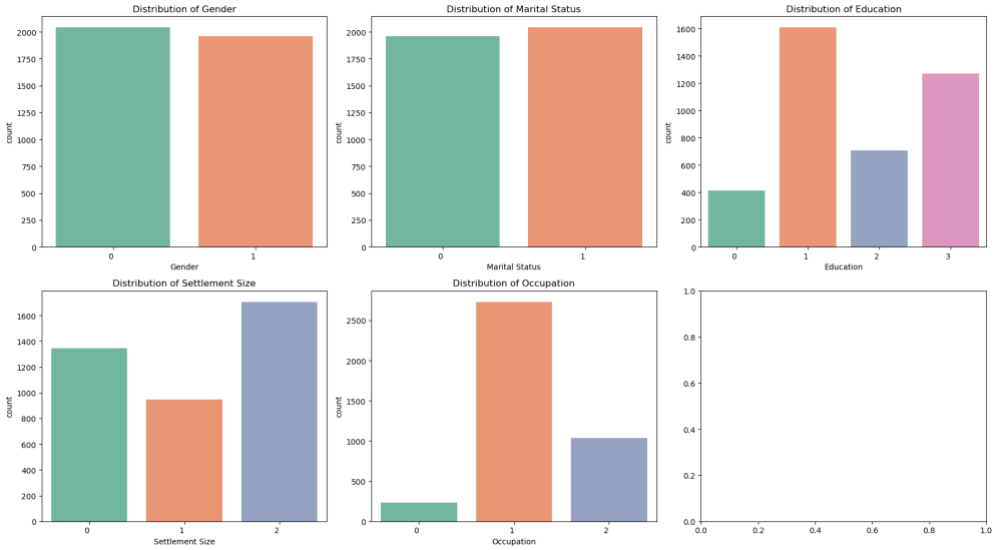

Figure 4. Distribution of other features

These charts show that this dataset has the number of males and females are quite identical, there is no noticeable different among marital status. Meanwhile, most of the customers have finished high school, they mostly live in the big city, and most of them are skilled employee.

3. Find the optimal clusters.

In this report, to find the optimal clusters, Elbow Method and Silhouette score are used.

Elbow method is to find the range of optimal values, and Silhouette will be used to determine which value is the best to use.

a. Elbow Method:

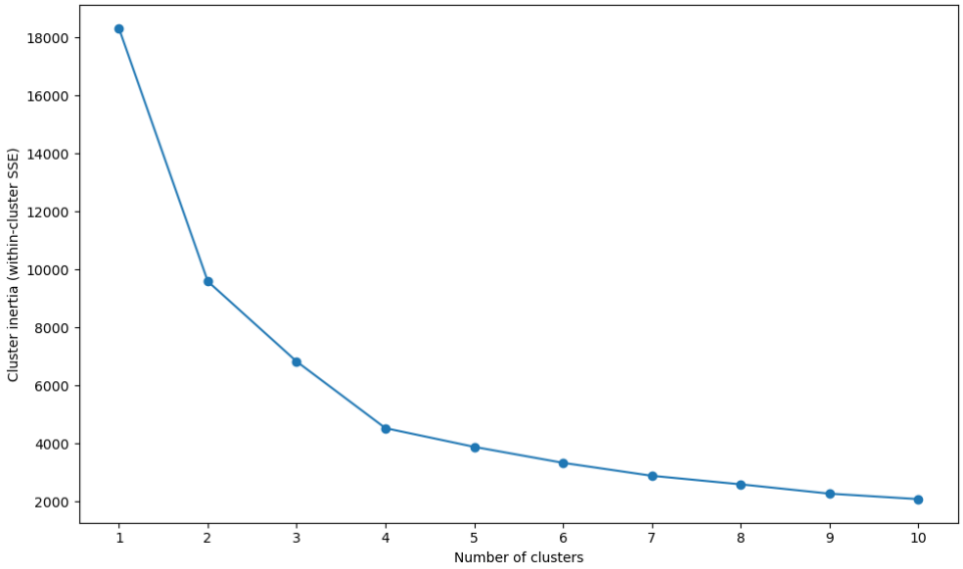

Figure 5. Elbow method

When running the code shown above, the elbow happens at around 2 to 4 clusters, when the cluster inertia (within-cluster SSE) drops dramatically, and then declines slightly after that. Therefore, the number of clusters from 2 to 4 will be used to analysed further to determine the optimal number of clusters by using Silhouette method.

b. Silhouette method:

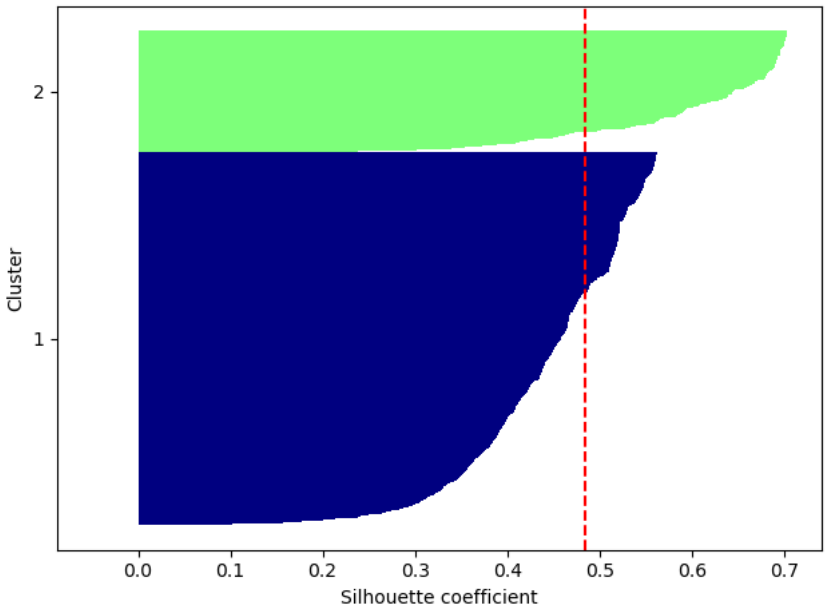

Figure 6. Silhouette method with n_clusters = 2

With number of clusters equals 2, Silhouette score is 0.48, and the shape of the first cluster does not look good, therefore, this cannot be the best number of clusters.

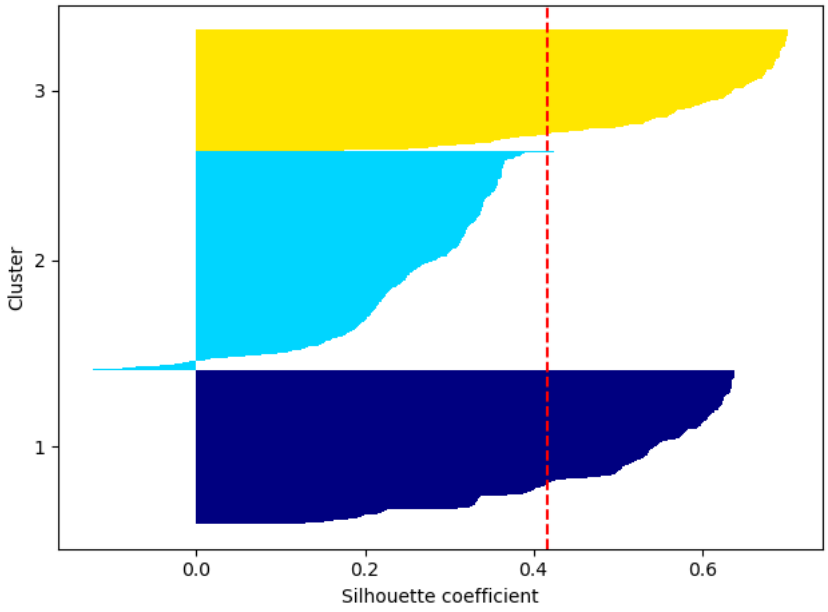

Figure 7. Silhouette method with n_clusters = 3

With number of clusters equals 3, Silhouette score is 0.42, and the shape of the second cluster does not look good, therefore, it cannot be the best number of clusters.

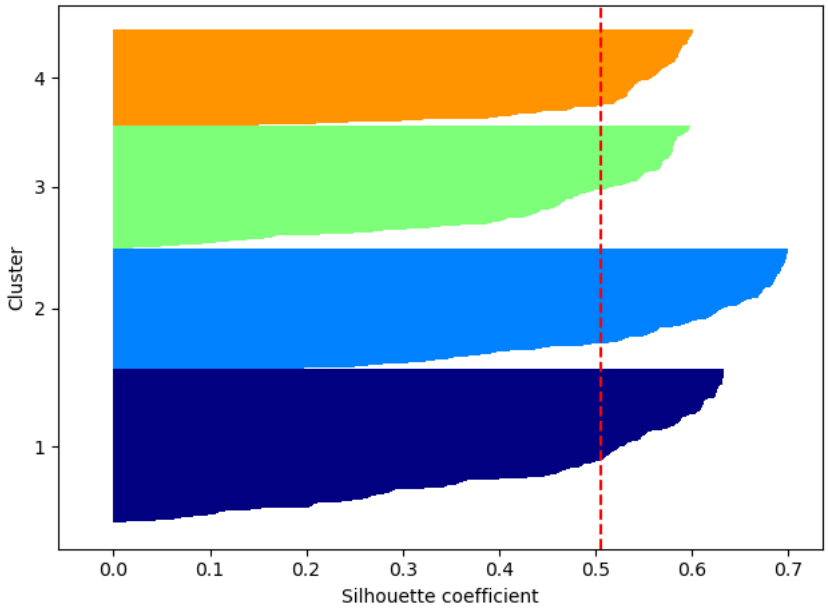

Figure 8. Silhouette method with n_clusters = 4

With number of clusters equals 4, Silhouette score is 0.50, the shape of all clusters looks quite even, and this is the highest Silhouette score among these three mentioned figures. Therefore, with this dataset, 4 is the best number of clusters for customer segmentation.

4. Agglomerative Clustering.

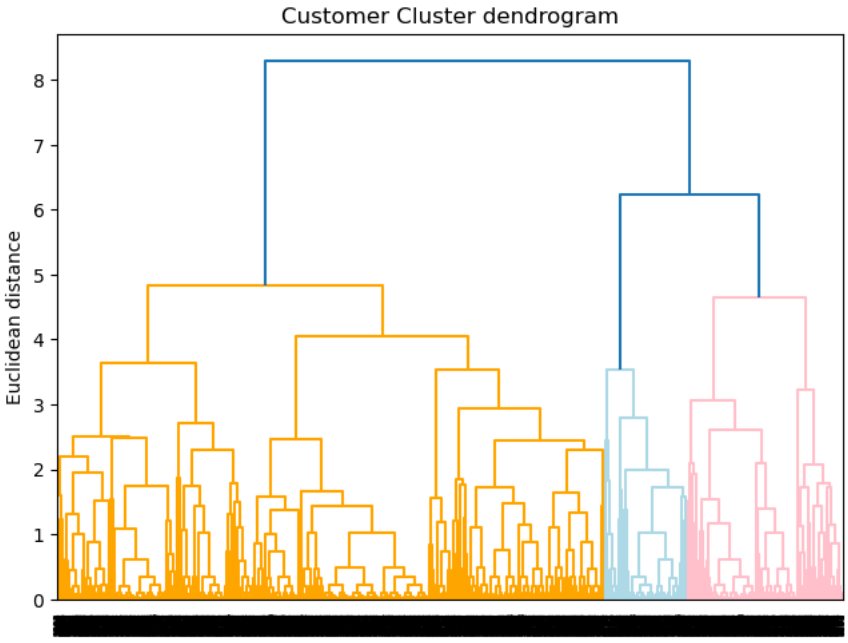

Figure 9

4 clusters to be chosen based on this graph.

cluster	Gender	Marital Status	Education	Settlement Size	Occupation	Income	Age	customer counts
0	0.082305	0.995885	2.987654	1.868313	1.888889	240571.674897	61.460905	243
1	0.425711	0.366702	2.503688	1.387777	1.657534	185768.228662	50.452055	949
2	0.570911	0.336753	1.650573	0.943271	1.036210	94218.529270	30.774291	1657
3	0.510860	0.776716	0.865334	0.891399	0.915725	127317.202433	39.946134	1151

Figure 10. Table presenting customer counts

There are 4 clusters of customers by using this method.

The first is a cluster of senior-aged, non-single males with high education level and high income, living in the big city.

The second is a cluster of high-aged, single individuals with high education level and high income, living in the mid-sized city.

The third is a cluster of individuals (mostly females) in 30s, mostly single, have lower income, mostly living in the mid-sized city.

The fourth is a cluster of individuals nearly in the middle-aged, mostly non-single, have moderate income, mostly living in the mid-sized city.

5. Customer segmentation using KMeans++ method & Recommendations.

a. Customer segmentation.

	Gender	Marital Status	Education	Settlement Size	Occupation	Income	Age
0	0.265712	0.827367	0.945107	0.019093	0.833731	0.016599	-0.262518
1	0.309426	0.440574	3.000000	1.813525	1.889344	1.472173	1.374023
2	0.631420	0.132931	2.276939	1.026183	1.163142	-0.930166	-1.042276
3	0.897933	0.568475	0.589147	1.998708	0.976744	-0.689989	0.030902

Figure 11. Table presenting cluster centers

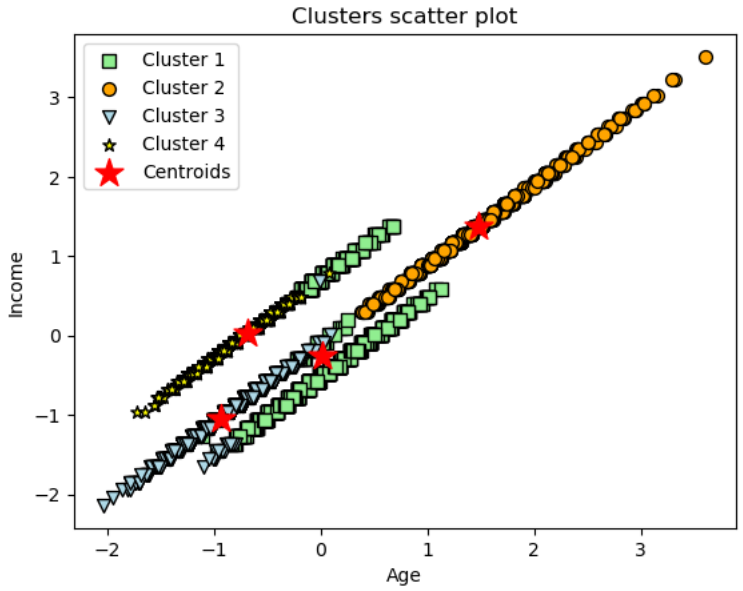

Figure 12. Clusters scatter plot using KMeans++ (standardised data)

cluster	Gender	Marital Status	Education	Settlement Size	Occupation	Income	Age	customer counts
0	0.265712	0.827367	0.945107	0.019093	0.833731	135159.303103	37.250597	1257
1	0.309426	0.440574	3.000000	1.813525	1.889344	205794.655738	54.055328	976
2	0.631420	0.132931	2.276939	1.026183	1.163142	89215.184290	29.243706	993
3	0.897933	0.568475	0.589147	1.998708	0.976744	100870.365633	40.263566	774

Figure 13. Table presenting customer counts

Based on the chart and the table including cluster and customer counts for each cluster above, every cluster has its own customer profiling.

Regarding to the first cluster, which is the largest group of customers, this group can be considered as a cluster of nearly middle-aged, non-single males with moderate to slightly high income, with low education and living in small city.

Turing to the second cluster, this group can be considered as a cluster of older, highly educated males with high income, living in big city.

About the third cluster, this group can be a cluster of younger, university educated individuals with lower income, living in mid-sized city.

The last cluster is mainly a group of middle-aged females with moderate incomes. This group is a cluster of middle-aged females with moderate income, lower education levels, living in big city.

b. Recommendations.

For the first cluster (cluster 0), traditional advertising can be a bright option. With this strategy, company should utilise traditional advertising channels such as local newspapers or community bulletin boards, which are more prevalent in small towns compared to other types of cities. Moreover, for men group, practical solutions can be a good choice. Company should highlight the practical benefits directly and ease of use of products, such as durable and multi-functional items.

For the second cluster, since they have high income, exclusive services can be an outstanding marketing strategy. Company should put effort to offer them exclusive services served only for this group, such as introducing elite membership clubs or loyalty rewards such as luxury travel experiences or special retreats.

For the third cluster, since they are still young and have lower income, affordable and trendy products can be a good strategy. Company should emphasize products' affordability with trendy design to appeal to younger generation. Moreover, developing creative and interactive social media campaigns on platforms such as Instagram and YouTube are the good ways to capture the attention of younger audiences.

For the fourth cluster, with most of the customers are female living in the big city, company can reach them by enhancing urban lifestyle. Company can target them with urban gardening kits, and supermarket delivery services, which fit for their fast-paced lives in the big city.

6. Compare two methods.

With the collective clusters using these two methods, overall, these two methods capture reasonable similar features to group the customers. However, if using Agglomerative, there are some clusters facing overlap issue, since this method gave the information that several customers living in the mid-sized city, while the real data does not.

7. Conclusion.

By exploring data, specific information about how features affect each other has been

provided to cluster customers easier. After that, customer segmentation techniques including K-means++ and Agglomerative were applied to cluster customers. To determine the optimal number of clusters, Elbow and Silhouette methods were used.

With the essential data after applying segmentation techniques, company can have detail information about each group of customers and can launch some marketing strategies targeting to the target audience based on each cluster. These strategies can enable company to enhance customer engagement, increase brand loyalty, hence, improve benefits in the long-term for the company.