SDSC3002 Project Report

Group: Minor in Data Science

Project Title: Credit Card Customer Segmentation

Table of Contents

Section 1 - Introduction

To promote the growth of a business, it is crucial to formulate an effective marketing strategy to cater to different customers' needs. This project takes credit card services as a model. The goal is to investigate what kind of credit card services should be offered to serve different groups of customers based on the results of customer clustering so that the company's resources can be allocated efficiently and provide suitable services for different groups of card holders such as sending appropriate advertisements to them.

Section 2 - Description of dataset

We made use of the "Credit Card Dataset for Clustering" dataset provided by Arjun Bhasin on Kaggle. It is composed of 8950 records and 18 attributes about customers' purchase and payment habits over the last half year [1], including:

Code	Discerption	Remarks
CUSTID	Identification of Credit Card holder (Categorical)	
BALANCE	Balance amount left in their account to make purchases	
BALANCE_FREQUENCY	How frequently the Balance is updated, score between 0 and 1	1 = frequently updated 0 = otherwise
PURCHASES	Amount of purchases made from account	
ONEOFF_PURCHASES	Maximum purchase amount done in one-go	
INSTALLMENTS_PURCHASES	Amount of purchase done in instalment	
CASH_ADVANCE	Cash in advance given by the user	
PURCHASES_FREQUENCY	How frequently the Purchases are being made, score between 0 and 1	1 = frequently purchased 0 = otherwise
ONEOFF_PURCHASES_FREQUENCY	How frequently Purchases are happening in one-go	1 = frequently purchased 0 = otherwise
PURCHASES_INSTALLMENTS_FREQUENCY	How frequently purchases in instalments are being done	1 = frequently done 0 = otherwise
CASH_ADVANCE_FREQUENCY	How frequently the cash in advance being paid	
CASH_ADVANCE_TRX	Number of Transactions made with "Cash in Advance"	
PURCHASES_TRX	Number of purchase transactions made	
CREDIT_LIMIT	Limit of Credit Card for user	
PAYMENTS	Amount of Payment done by user	
MINIMUM_PAYMENTS	Minimum amount of payments made by user	
PRC_FULL_PAYMENT	Percent of full payment paid by user	

There were 313 missing values in the MINIMUM_PAYMENTS column and 1 missing value in the CREDIT_LIMIT column, which needed to be handled beforehand. After clustering, we observed which features dominated how customers were aggregated. Based on how the customers behaved in terms of the 18 features, we could design appropriate marketing strategies to fulfil them.

Section 3 - Observation of dataset

A heatmap (Figure 1) was built to show the correlation between attributes of the data set according to Pearson's correlation (Figure 2). Lighter colour indicated two attributes were highly correlated and darker colour suggested the opposite.

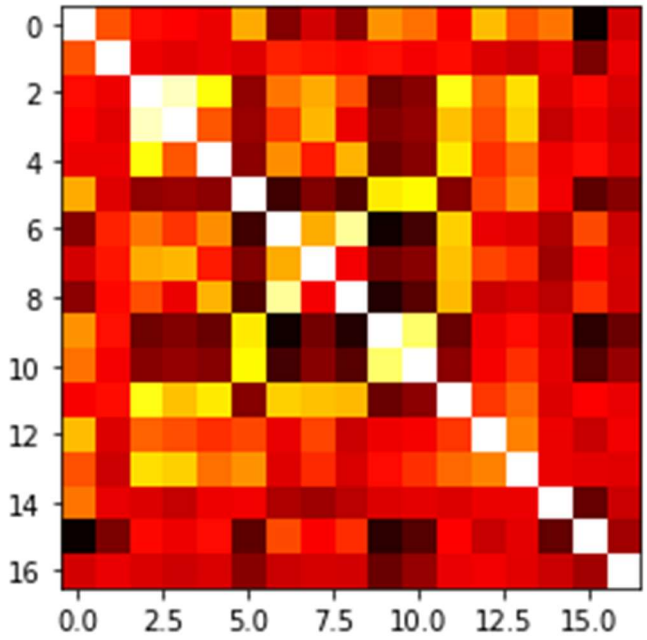


Figure 1: Heatmap based on Pearson's correlation matrix

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Figure 2: Pearson's correlation

Histogram was built for various attributes, including Balance Frequency, Purchases Frequency, One-go Purchases Frequency, Purchases Frequency in Instalment, "Cash in Advance" Paying Frequency, Number of Transactions made with "Cash in Advance", Number of purchase transactions made, Tenure of credit card service for users (Figure 3).

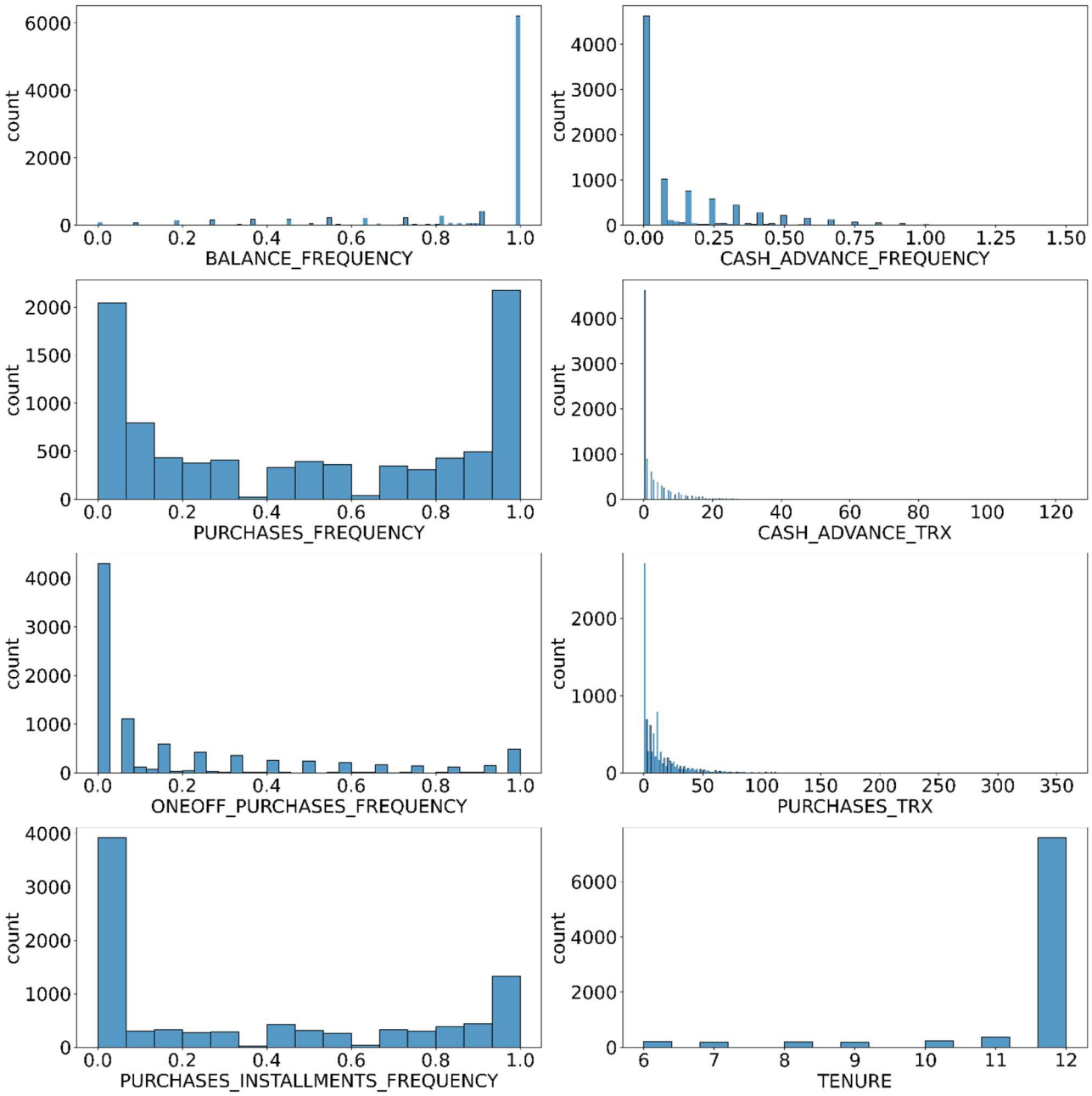


Figure 3: Histogram of various attribute

Section 4 - Description of data mining techniques with benefits

Some pre-processing of the data and overview of the data were done before clustering. Missing values were filled with the mean of the existing values in their own column. Since customers' ID were categorical data, that column was dropped before clustering.

Hierarchical Clustering

Two clustering techniques were demonstrated in this project, hierarchical clustering and K-mean clustering. Hierarchical clustering was adopted as the baseline method as it was a deterministic approach to guarantee reproducible results. For each linkage criterion, “Single”, “Complete”, “Average” and “Ward”, a dendrogram was generated.

Hierarchical Clustering Procedures [2]

1. Consider each node as a single cluster
2. Calculate the distance of one cluster to other clusters
3. Merge clusters based on linkage criterion to form a single cluster
4. Recalculate linkage distance between clusters’ centroids and merge clusters
5. Merge until one cluster left

Hierarchical Clustering Linkage Criterion

There is no gold-standard definition for inter-cluster distance. Typically, there are four linkage criteria to measure the distance, single linkage, complete linkage, average linkage and Ward linkage.

Single linkage (Figure 4) and complete linkage (Figure 5) search for the minimum [2] and maximum [2] distance between points in two clusters respectively.

$$d(C_i, C_j) = \min_{a \in C_i, b \in C_j} d(a, b)$$

Figure 4: Single-linkage algorithm

$$d(C_i, C_j) = \max_{a \in C_i, b \in C_j} d(a, b)$$

Figure 5: Complete-linkage algorithm

Average linkage (Figure 6), as the name suggests, looks for the average distance [2] between pairs of observations in two clusters.

In terms of these three methods, the greater the distance, the better the result. In contrast, Ward's method (Figure 7) approaches the problem by summing up the intra-variance of two clusters [2]. The formula below does not take the average but it poses no impact on the result. The value as low as possible is preferred such that the two clusters are not likely to be considered as one cluster.

$$d(C_i, C_j) = \sum_{a \in C_i, b \in C_j} \frac{d(a, b)}{|C_i||C_j|}$$

Figure 6: Average-linkage algorithm

$$d(C_i, C_j) = \sum_{a \in C_i \cup C_j} \left\| a - \overline{C_\iota \cup C_J} \right\|$$

Figure 7: Ward's method

K-mean clustering

On top of hierarchical clustering, we further conducted k-mean clustering to compare their performance. The main benefit of k-mean clustering over hierarchical clustering was that it was much less computationally expensive as it did not need to store a $n \times n$ distance matrix to construct the dendrogram. The procedures of k-mean clustering are as follows [3]:

1. Randomly pick k points as the centroids
2. Compute the distance between each centroid and other points to assign the points to their nearest centroid
3. Compute the mean value for every cluster and set it to be the centroid
4. Repeat the process until no centroid shifted to new position

Since k-mean clustering highly depends on the choice of centroids at the beginning, it might generate different results in every iteration, unlike hierarchical clustering.

K-mean++

Instead of randomly picking k points of initial centroids, K-mean++ selects initial cluster centroids for k-mean clustering in a "smarter" way to speed up convergence. The procedures are as follows [4]:

1. Randomly pick a point x as a centroid
2. Compute the distance between each point and its centroid
3. Find the point whose distance from its centroid is the largest to be the next centroid (Figure 8)

$$c = argmax\left[\frac{d^2}{d_1^2 + d_2^2 + \cdots + d_n^2}\right]$$

Figure 8: Calculation for the next centroid

4. Repeat step 2 and 3 until the desired k centroids are found

The idea is to find the point farthest away from the existing centroids to be the next centroid in order to avoid picking two random points close to each other to be centroids.

Section 5 – Evaluation

Two quantitative techniques were adopted to evaluate the performance of the clustering, elbow method and calculation of silhouette score.

Silhouette Score

Silhouette score is a metric to measure the effectiveness of a clustering technique. Its value ranges from -1 to 1. The higher the score, the farther apart the clusters are. If the points are assigned to a wrong cluster such that the intra-cluster distance is even greater than the inter-cluster distance, the silhouette score will become negative. Zero means there is no clear distinction between clusters [5]. We aimed for the k value that produced the highest score in k-mean clustering.

To determine an optimal number of clusters, we plotted silhouettes for every value of k, which can be calculated by the following steps [5]:

1. Compute the average distance a_i of point i to other points in the same cluster
2. Compute the average distance b_i of point i to other points in the cluster nearest to its own cluster
3. Compute silhouette coefficient shown in Figure 9

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

Figure 9: Calculation for silhouette coefficient [5]

Ideally, b_i tends to infinity thus $a_i \ll b_i$ (the intra-cluster distance $\ll$ inter-cluster distance), so that the clusters are distinguishably separable, i.e., $s_i = 1$.

Elbow Method

The Elbow method targets a suitable number of clusters so that the clusters are well separated without over-fitting [6]. When the number of clusters equals the total number of points, the sum of squared distance between each point and its cluster's centroid becomes zero. Despite zero error, clustering points in this way is overfitting and does not achieve the objective of clustering. Grouping all points into one cluster gives meaningless results either. Rather than aiming for zero squared error, the k-mean cluster looks for the k value at the point when further increasing the number of clusters only makes little decrease in the error. In an SSE-k graph, SSE should drop sharply at the beginning but shortly become stable and steadily decrease. The k value at that turning point is the target. In practice, it can be determined by finding the largest orthogonal distance between the secant line of the curve to the curve based on the fact that SSE always dropped as k increased (Figure 10).

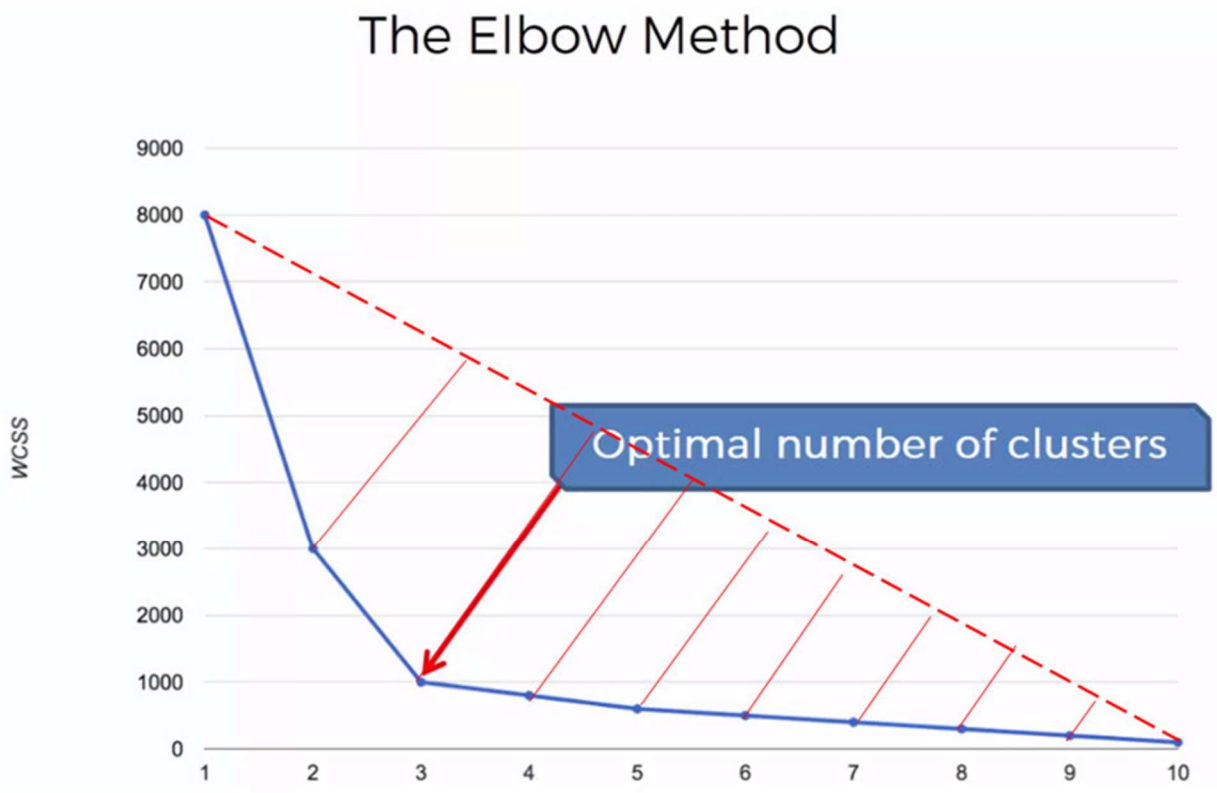


Figure 10: Brief explanation about Elbow method [7]

PCA

PCA converts a set of observations of possibly correlated variables into a set of linearly uncorrelated variables called principal components such that projecting the data on the space formed by these principal components could retain most of the information [8]. The more information to be retained, the principal components need to be preserved. This project reduced a 17-dimensional data into 2 in order to visualize the result of clustering on a 2D graph. Details of procedures can be found in figure 11.

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

$$x_j^i = \frac{x_j^i - \bar{x}_j}{\sigma_j} \forall j$$

$$C = \frac{1}{m}\sum_i^m (x_i)(x_i)^T, \qquad C \in \mathbb{R}^{n \times n}$$

$$u^T C = \lambda u$$

$$U = \left\{ \begin{bmatrix} u_1 \end{bmatrix}, \begin{bmatrix} u_2 \end{bmatrix}, \dots, \begin{bmatrix} u_n \end{bmatrix} \right\}, \qquad u_i \in \mathbb{R}^n$$

$$x_{new}^i = \begin{bmatrix} u_1^T x^i \\ u_2^T x^i \\ \dots \\ u_k^T x^i \end{bmatrix}, \qquad x_{new}^i \in \mathbb{R}^k$$

Figure 11: Brief explanation on calculating PCA [8]

Section 6 - Results and Discussion

Hierarchical clustering

The performance of hierarchical clustering was examined with silhouette scores. Since computation of silhouette score involved inter-cluster distance, Ward's method exhibited the highest silhouette score at $k = 2$ and $k = 3$.

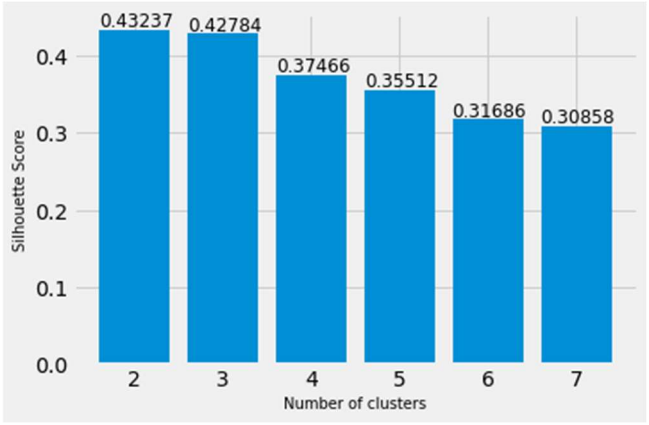


Figure 12: Silhouette scores of Ward's methods from $k = [2,8]$

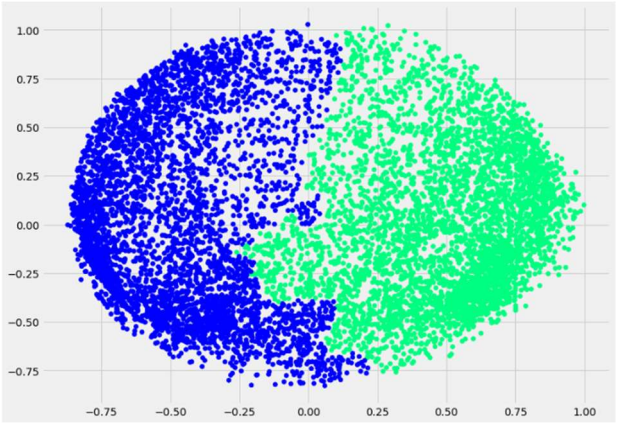


Figure 13: Hierarchical Clustering (Ward's methods) with highest Silhouette scores

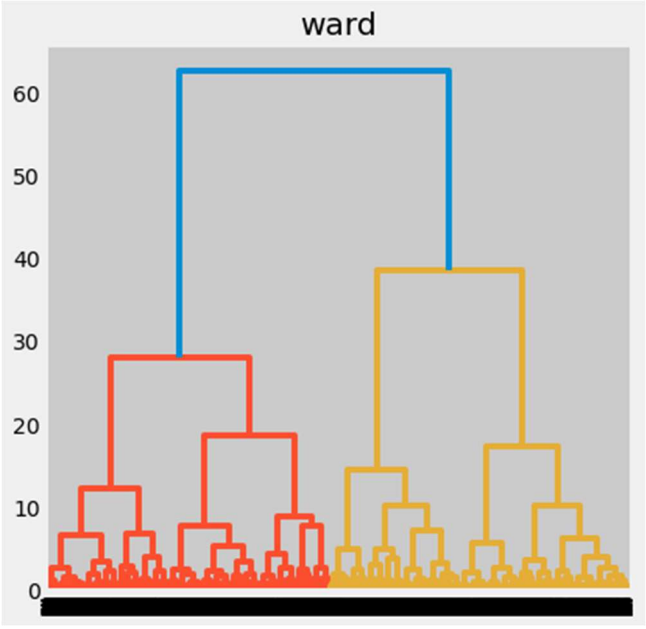


Figure 14: Dendrogram of Ward's method

The single linkage seriously mis-assigned the data to wrong clusters which lead to a negative silhouette score. Even though the score was positive at $k = 2$, it was still much lower than other methods' scores.

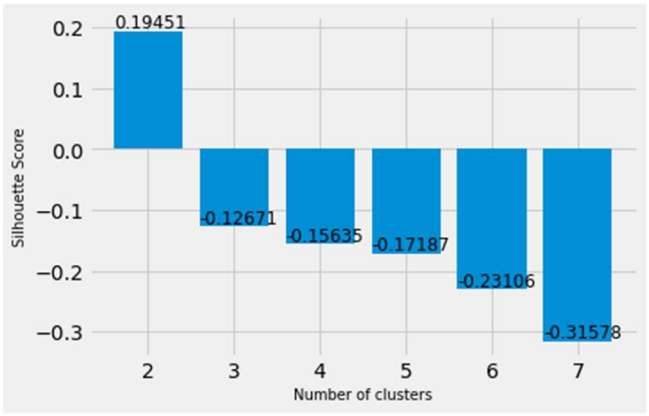


Figure 15: Silhouette scores of Single Linkage from $k = [2,8]$

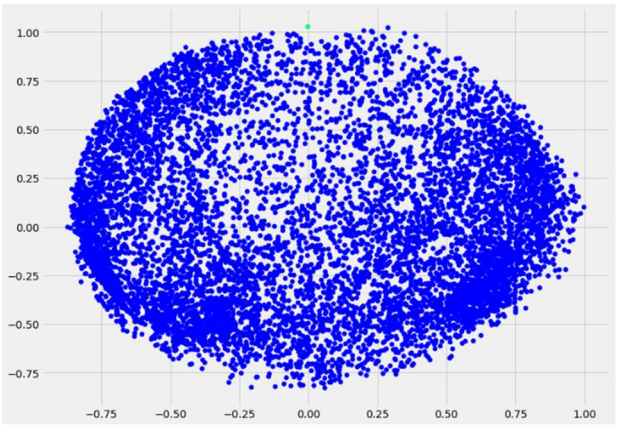


Figure 16: Hierarchical Clustering (Single Linkage) with highest Silhouette scores

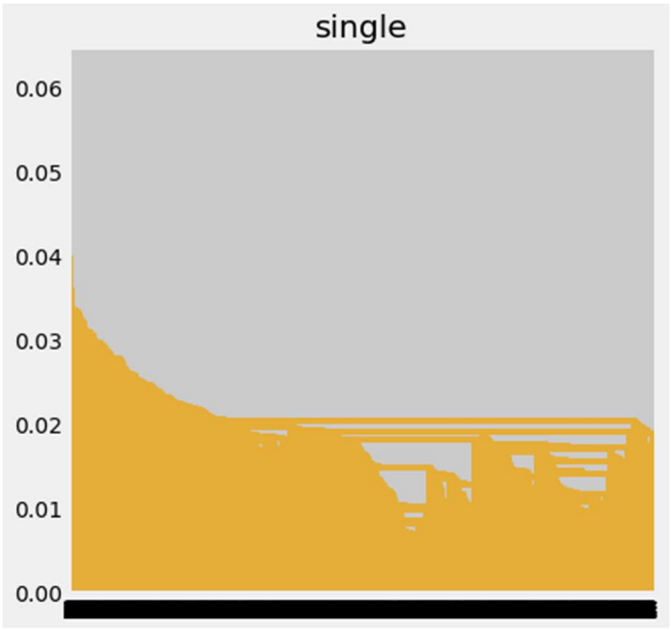


Figure 17: Dendrogram of Single Linkage

The complete linkage also had a lower score than Ward's method.

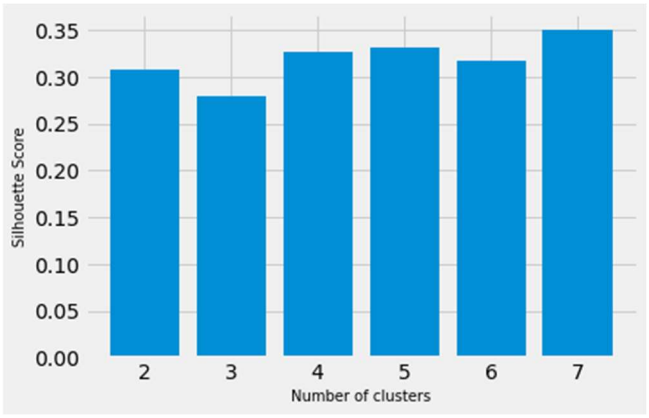


Figure 18: Silhouette scores of Complete Linkage from $k = [2,8]$

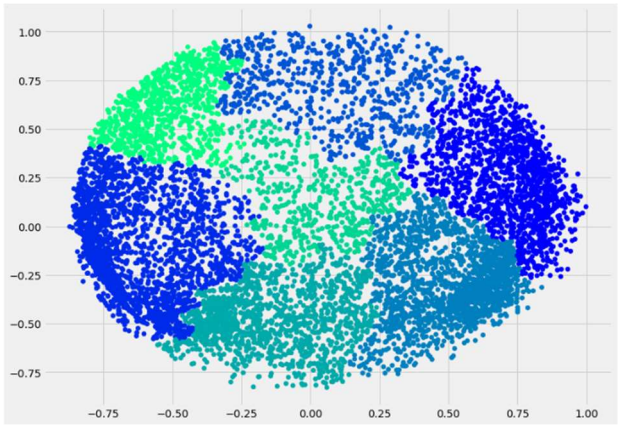


Figure 19: Hierarchical Clustering (Complete Linkage) with highest Silhouette scores

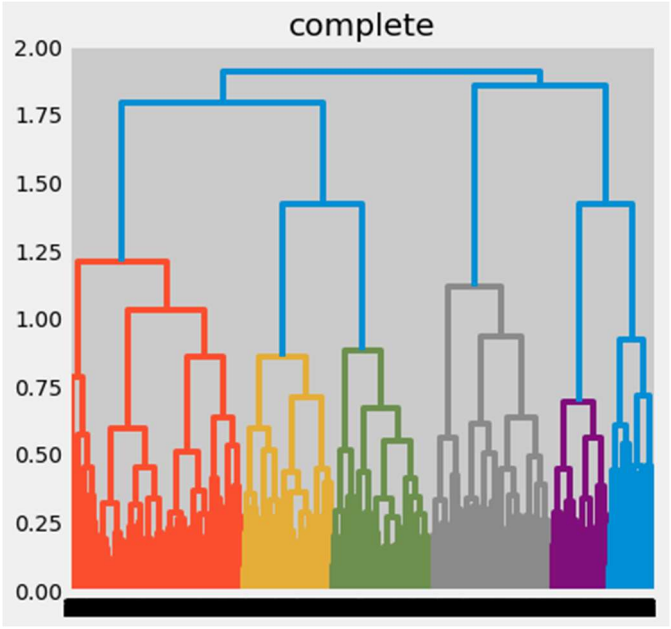


Figure 20: Dendrogram of Complete Linkage

Although the average linkage had comparable score with the Ward's method, the Ward's method still had a slightly better performance than the average linkage.

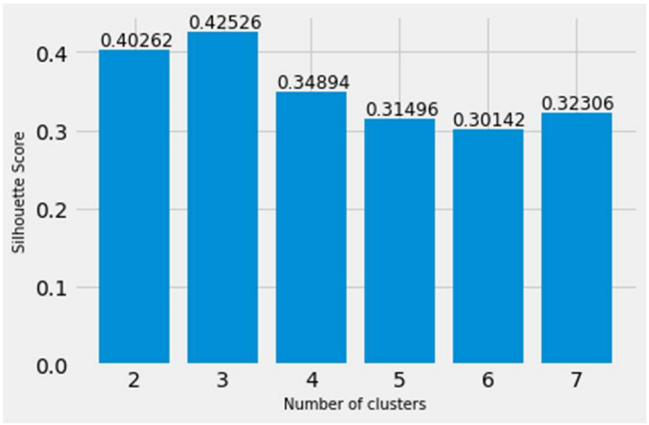


Figure 21: Silhouette scores of Average Linkage from $k = [2,8]$

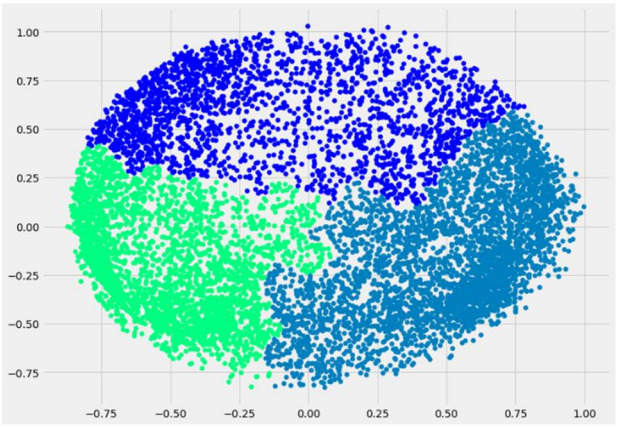


Figure 22: Hierarchical Clustering (Average Linkage) with highest Silhouette scores

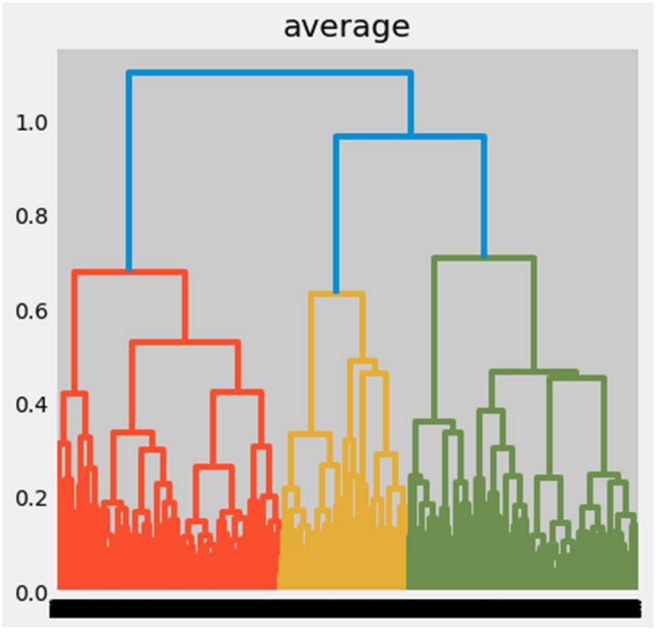


Figure 23: Dendrogram of average linkage

According to the Silhouette scores, Ward's method was chosen, and $k = 3$ was chosen over $k = 2$ for the hierarchical clustering (Figure 24), in order to design a more precise strategy as there were only trivial differences in their scores.

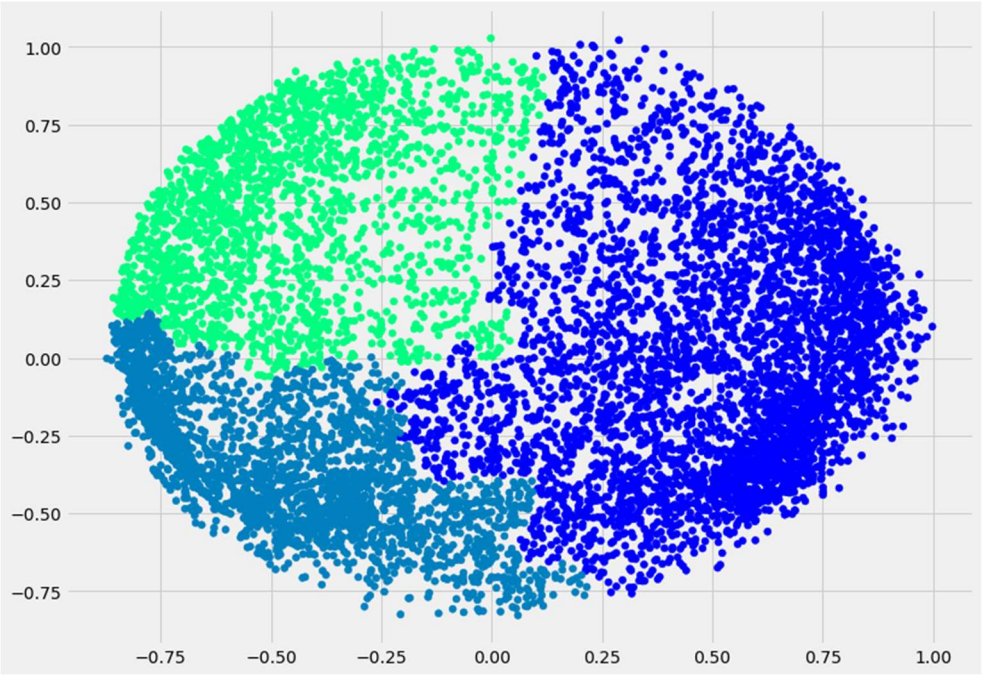


Figure 24: Hierarchical Clustering (Ward's Method) with 3-Clusters

K-mean clustering

Since the data were densely packed and there was no outlier, k-mean and k-mean++ produced almost identical results. In terms of convergence speed, k-mean++ (around 5.9s) was indeed higher than k-mean (6.4s) by about 0.5s. k-mean++ was adopted eventually to ensure the stability of the results. Regarding the elbow method analysis, the turning point of the curve in the SSE-k graph below was at around $k = 3$ and $k = 4$ which produced the largest inter-cluster distance.

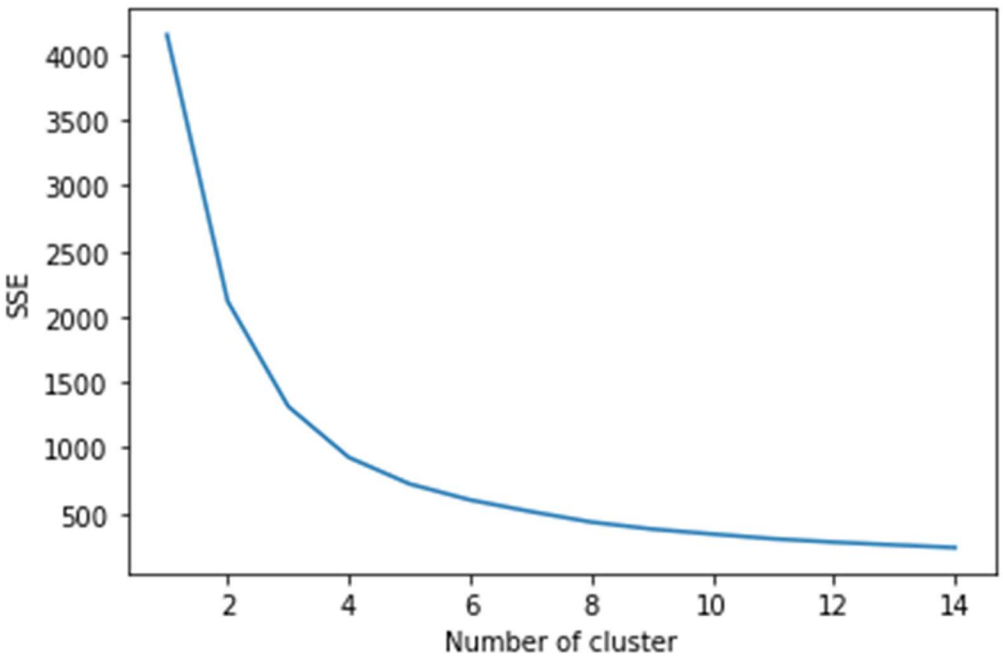


Figure 25: K-mean++ SSE

Based on the result of silhouette score, $k = 2$ and $k = 3$ were suggested, and based on the result from the elbow method, $k = 3$ and $k = 4$ were suggested. Therefore, 3 was indeed the optimal k value in addition to consideration of marketing strategy analysis (Figure 26).

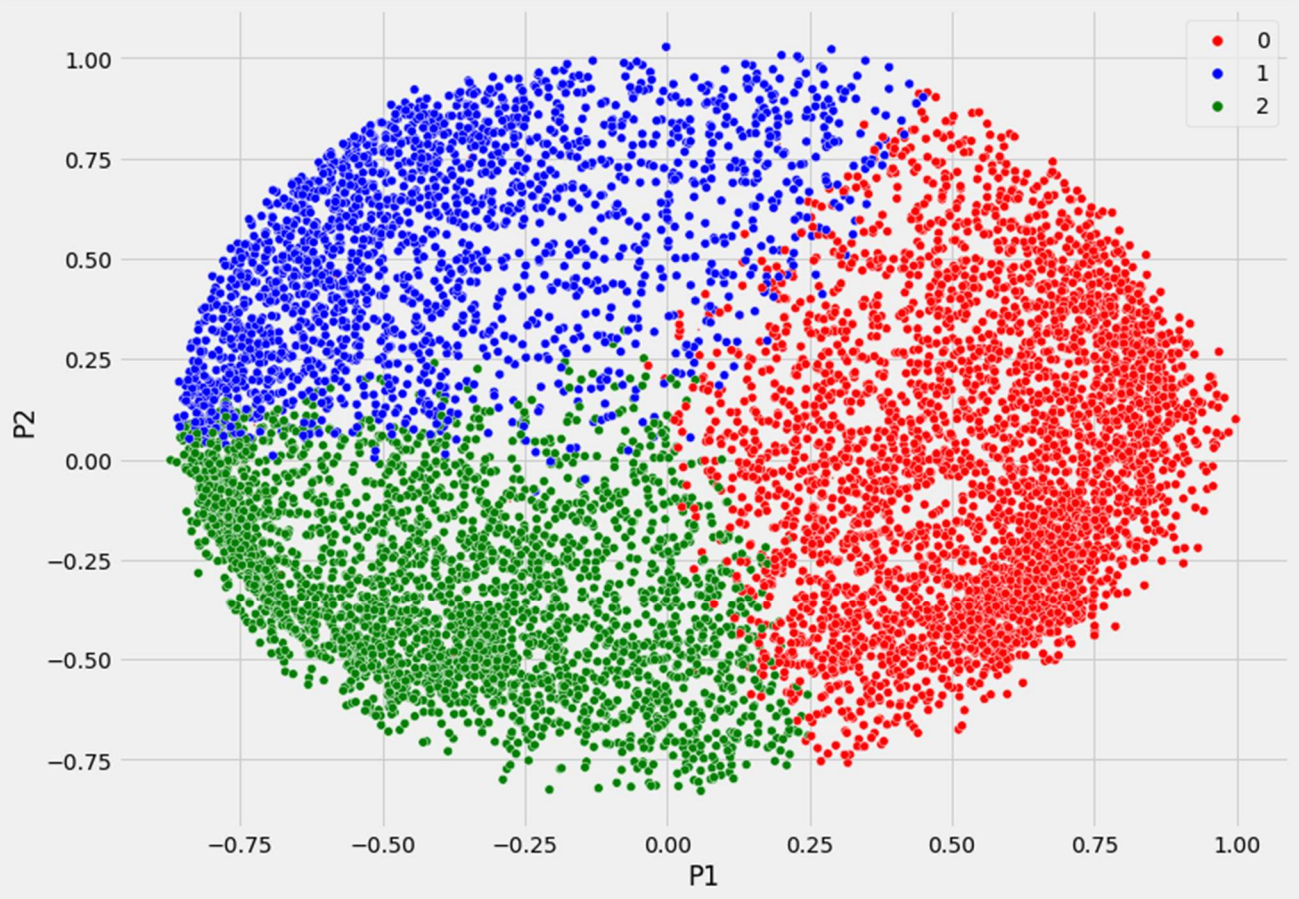
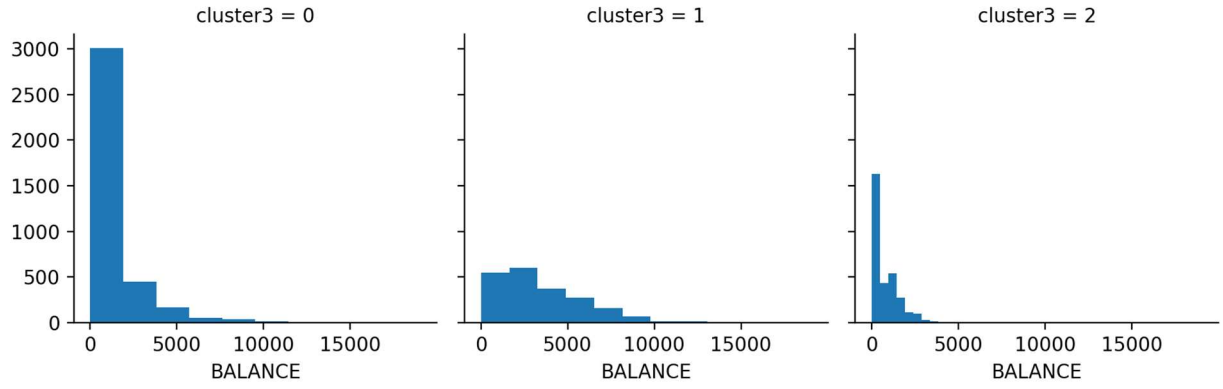


Figure 26: K-mean++ with 3- Clusters

Section 7 – Conclusion

In conclusion, we can divide the customers into 3 clusters with the k-mean method. Characteristics of every customer cluster are illustrated with histograms (Figure 27).



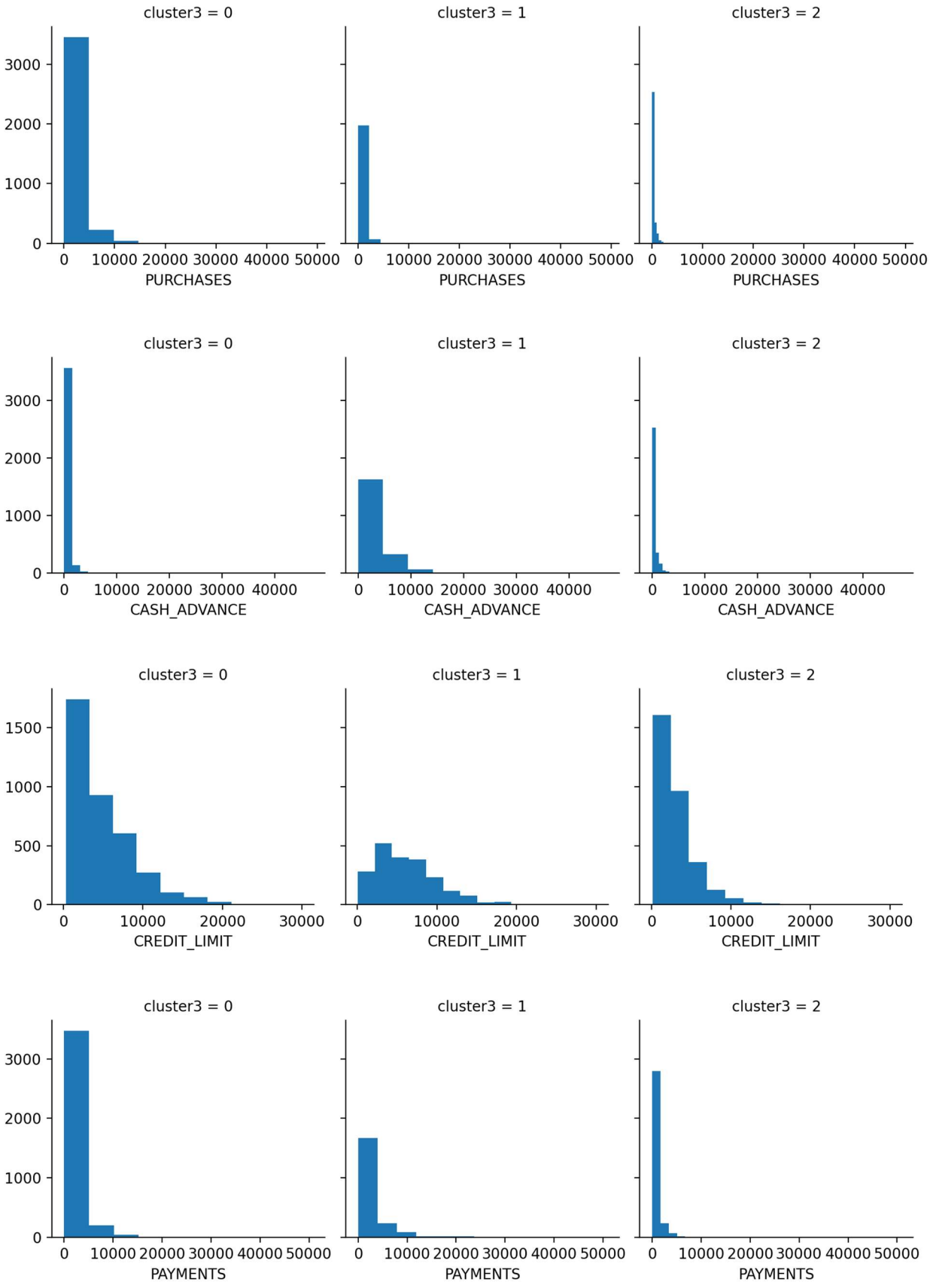


Figure 27: Histograms about the statistics of the clusters

The first group of customers has low balance in their account. They have low purchase frequency and the lowest credit card limit. Therefore, these customers are less active and not the main target customers. Credit card companies should assign fewer resources to these customers.

The second group of customers has the highest purchase frequency, medium account balance and low cash advance, which means they usually make purchases with credit cards instead of cash loans. They are the main target of the credit card company, so more resources should be allocated to them. For example, companies may encourage them to apply for more credit cards and offer better cash back in order to maintain the customer loyalty.

As for the last group of customers, they have the highest balance, highest cash advance and highest credit limit. It is likely that these customers are using their credit card as a loan, so credit card companies can assign resources related to loans to them. For example, they can offer a higher credit limit or lower loan interest rate to them, or even encourage them to apply for personal loan products. Histograms about the statistics of the clusters and A summary table are shown below (Figure 28).

	Cluster 1	Cluster 2	Cluster 3
Account Balance	Low	Medium	High
Purchase frequency	Low	High	Medium
Credit limit	Low	Medium	High
Cash advance	Medium	Low	High

Figure28: A summary of the 3 clusters.

Section 8 – References

[1] F. Murtagh, "A Survey of Recent Advances in Hierarchical Clustering Algorithms," *The Computer Journal, Volume 26, Issue 4,* p. 354–359, 11 1983.

[2] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed, "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity," *Middle-East Journal of Scientific Research 12 (7),* pp. 959-963, 7 12 2012.

[3] savyakhosla, "ML | K-means++ Algorithm," GeeksforGeeks, 13 7 2021. [Online]. Available: https://www.geeksforgeeks.org/ml-k-means-algorithm/. [Accessed 20 4 2022].

[4] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020,* pp. 747-748, 20 11 2020.

[5] Purnima Bholowalia, Arvind Kumar, "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN," *International Journal of Computer Applications (0975 – 8887),* pp. 17-24, 9 11 2014.

[6] A. Perera, "Finding the optimal number of clusters for K-Means through Elbow method using a mathematical approach compared to graphical approach," Linkedin, 2 10 2017. [Online]. Available:

https://www.linkedin.com/pulse/finding-optimal-number-clusters-k-means-through-elbow-asanka-perera. [Accessed 20 4 2022].

[7] NCSS, LLC, "Principal Components Analysis," NCSS Statistical Software, [Online]. Available: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Principal_Components_Analysis.pdf. [Accessed 20 4 2022].

[8] A. Bhasin, "Credit Card Dataset for Clustering," Kaggle, 2 5 2018. [Online]. Available: https://www.kaggle.com/datasets/arjunbhasin2013/ccdata. [Accessed 20 4 2022].

Section 9 – Work Allocation

Item	Person in charge		
	Chak Hei LAM	Chun Wai LEUNG	Ming Chi WONG
Code	X	X	X
Presentation Slides	X	X	X
Report	X	X	X

Appendix

MinorInDataScience-code.zip

CC_GENERAL.csv

3002project.ipynb

README.md