Exploratory Analysis on Customer Segmentation Data

Multivariate Statistical Analysis Spring 2023

Group Member: Yang Letian, Wang Ruofan, Wu Yue

Abstract

In this post, we used data from kaggle related to credit card customers. We first carry out sufficient data preprocessing, including data cleaning, data description and brief statistics description. After a series of variable adjustment and principal component analysis, we divided users into four categories, and each group showed unique consumption habits and credit behavior. We also selected data related to credit limit according to distance correlation, and classified customers by KMeans algorithm according to these data, and obtained customer groups with low credit, medium credit and high credit. We also perform regression analysis on the credit limit, and give a classification-based random forest regression method that performs better than the ordinary random forest regression. Finally, we also discuss the mainstream clustering methods and the reasons for choosing KMeans in this dataset.

Keywords: PCA , KMeans , Regression , Clustering , Dimension Reduction , Feature selection

Contents

1 Data Preprocessing

1.1 Data Source and Background

The dataset used in this project was obtained from Kaggle, a popular platform for discovering and sharing datasets.

Customer segmentation is a crucial aspect of marketing that involves classifying customers into distinct groups based on shared characteristics such as behavior, preferences, and purchasing habits. This allows companies to develop targeted marketing strategies that are tailored to the needs of each group.

The dataset contains 18 attributes and originally consisted of 8950 samples. Each attribute provides valuable information about the customers, which can be used to perform customer segmentation analysis. The meaning of the attributes is described below:

Attribute	Information	Data type
CUST_ID	Credit card holder ID	object
BALANCE	Monthly average balance (based on daily balance averages)	float64
BALANCE_FREQUENCY	Ratio of last 12 months with balance. (1: Frequently updated, 0: Not frequently updated)	float64
PURCHASES	Total purchase amount spent during last 12 months	float64
ONEOFF_PURCHASES	Total amount of one-off purchases	float64
INSTALLMENTS_PURCHASES	Total amount of installment purchases	float64
CASH_ADVANCE	Total cash-advance amount	float64
PURCHASES_FREQUENCY	Frequency of purchases (Percent of months with at least one purchase). (1: Frequently purchased, 0: Not frequently purchased)	float64
ONEOFF_PURCHASES _FREQUENCY	Frequency of one-off-purchases. (1: Frequently purchased, 0: Not frequently purchased)	float64
PURCHASES_INSTALLMENTS _FREQUENCY	Frequency of installment purchases. (1: Frequently purchased, 0: Not frequently purchased)	float64
CASHADVANCE_FREQUENCY	Cash-Advance frequency	float64
CASH_ADVANCE_TRX	Average amount per cash-advance transaction	int64
PURCHASES_TRX	Average amount per purchase transaction	int64
CREDIT_LIMIT	Credit limit	float64
PAYMENTS	Total payments (Due amount paid by the customer to decrease their statement balance) in the period	float64
MINIMUM_PAYMENTS	Total minimum payments due in the period	float64
PRC_FULL_PAYMEN	Percentage of months with full payment of the due statement balance	float64
TENURE	Number of months as a customer	int64

Overall, this dataset provides a valuable resource for performing customer segmentation analysis and developing targeted marketing strategies based on customer characteristics.

1.2 Data Cleaning

In the data cleaning phase, we performed two main tasks: handling missing values and removing unusual values.

- Handling missing values

  We first removed the missing values from the dataset. The attribute "CREDIT_LIMIT" had one missing value and "MINIMUM_PAYMENTS" had 313 missing values. We removed these missing values from the dataset. Additionally, we removed any duplicate values from the dataset. After this process, we were left with 8636 samples.

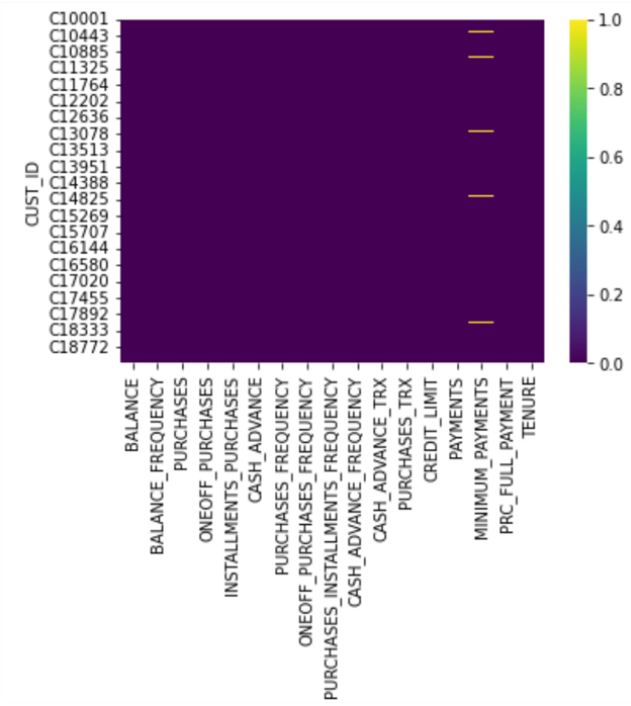


Figure 1: Missing Value Plot

- Removing unusual values

  We observed that the relationship between "PURCHASES", "ONE-OFF_PURCHASES", and "INSTALLMENTS_PURCHASES" was violated in 19 samples. Specifically, some samples had installments purchases that were greater than the total purchases, which is not possible. These values were treated as erroneous and were removed from the dataset. After removing these samples, we were left with 8617 samples.

CUST_ID	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES
C10533	400.41	0.00	489.39
C10772	880.19	0.00	927.45
C10940	3393.25	3364.59	77.66
C11506	130.24	0.00	152.24
C14099	550.62	0.00	583.95
C14810	0.00	0.00	20.00
C15378	205.06	0.00	607.76
C15508	47.69	82.41	0.00
C15897	0.00	0.00	66.95
C16009	65.60	0.00	112.60
C16010	486.27	580.20	0.00
C16133	279.76	0.00	578.55
C16714	468.96	0.00	498.96
C16899	465.50	0.00	513.00
C16978	426.25	0.00	653.55
C16991	339.11	611.65	12.41
C17045	5629.41	0.00	6229.41
C18403	356.77	45.65	333.34
C19075	510.00	0.00	780.00

Figure 2: Samples with unusual purchases

Furthermore, we observed that the attribute ”CASH_ADVANCE_FREQUENCY” should have values between 0 and 1. However, we found eight samples with values that exceeded 1. These values were considered unusual and were removed from the dataset. After this process, we were left with 8609 samples.

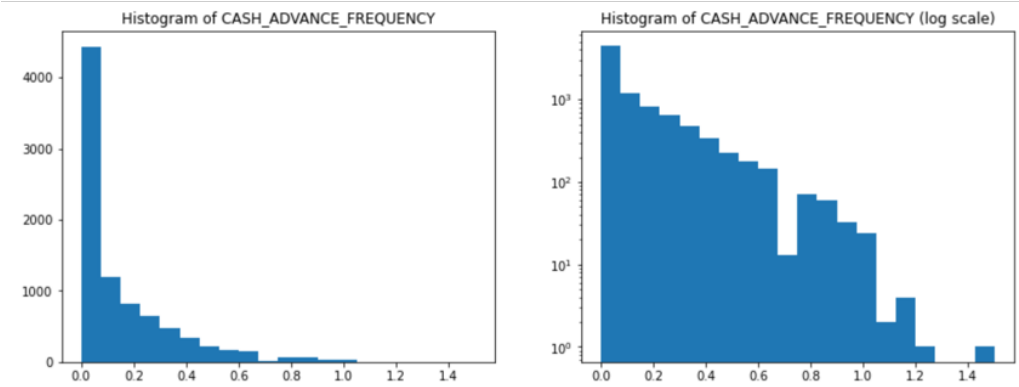
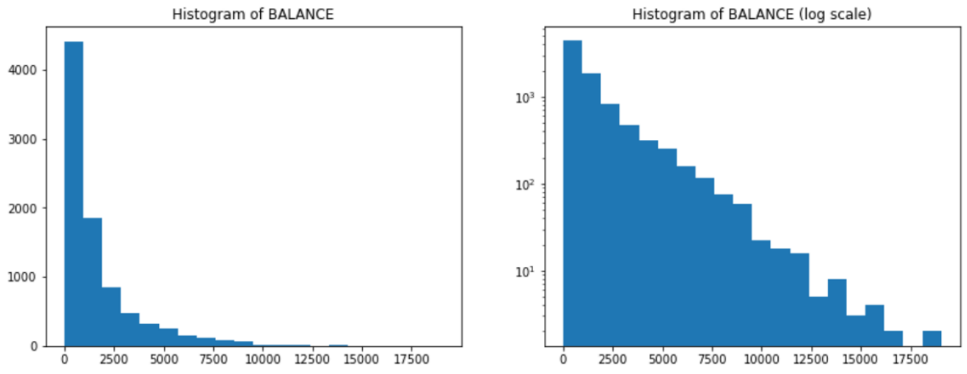


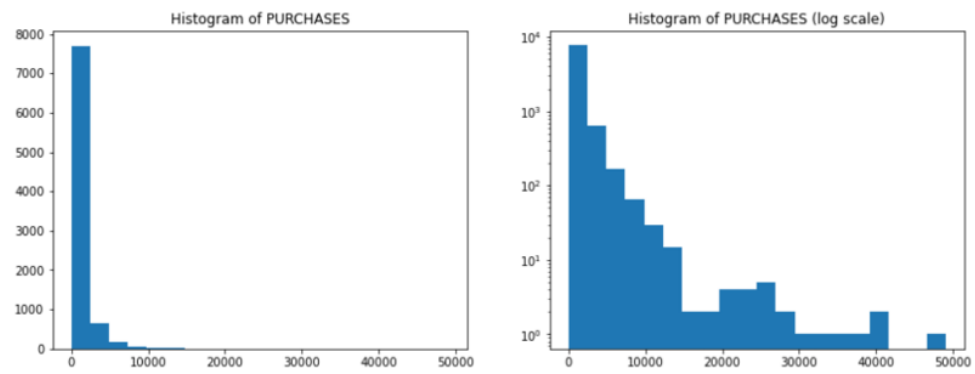
Figure 3: Histogram of Original CASH_ADVANCE_FREQUENCY

1.3 Data Description

In presenting the data, we used two types of y-axis scales: the original values and a log-scaled y-axis to better visualize the distribution of the data.

- Right-skewed data





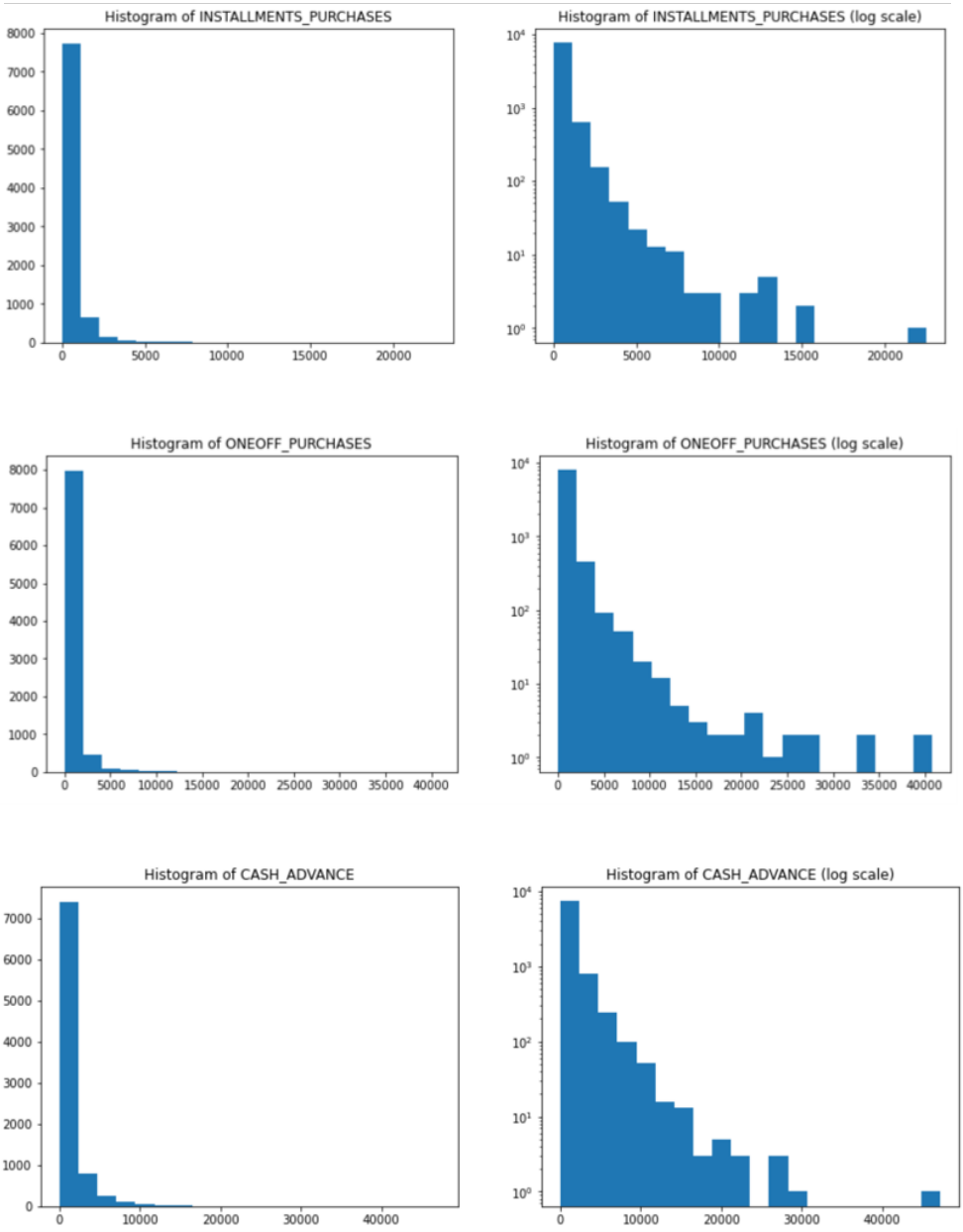


Figure 4: Histograms of Five Variables

As shown in the above graphs, attributes related to money such as "BALANCE", "PURCHASES", "INSTALLMENTS_PURCHASES", "ONEOFF_PURCHASES", and "CASH_ADVANCE" exhibit a clear right-skewed distribution. In subsequent analyses such as linear regression, we transformed these attributes using a log transformation.

- Distribution of frequency-type data

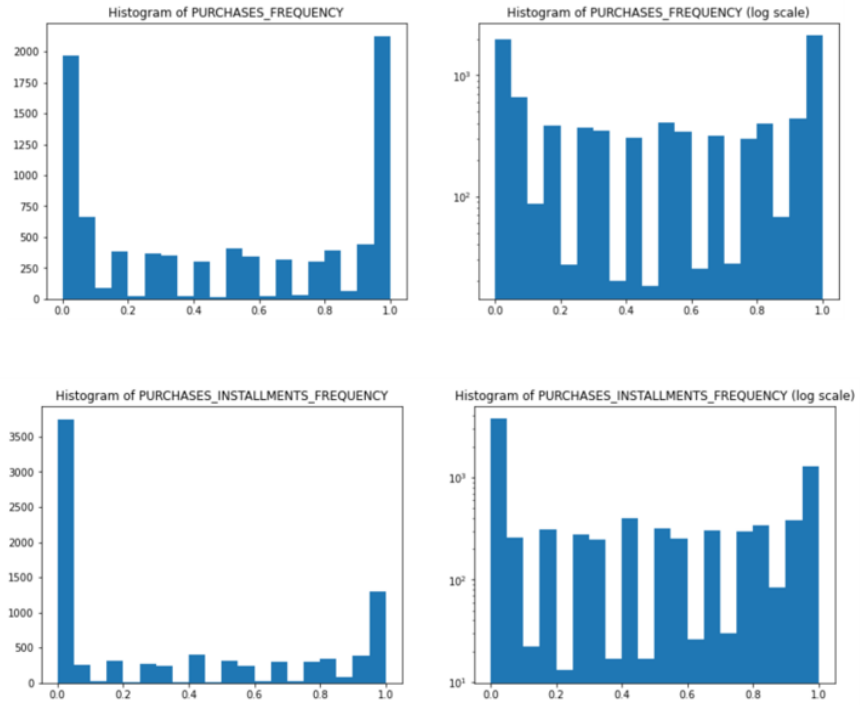


Figure 5: Histograms of PURCHASES_FREQUENCY and PURCHASES_INSTALLMENTS_FREQUENCY

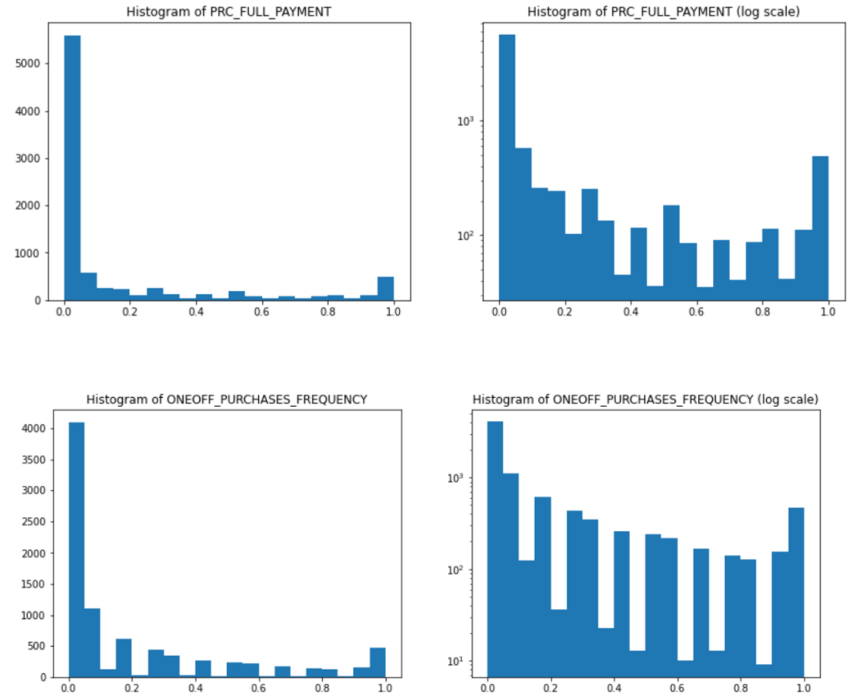


Figure 6: Histograms of PRC_FULL_PAYMENT and ONEOFF_PURCHASES_FREQUENCY

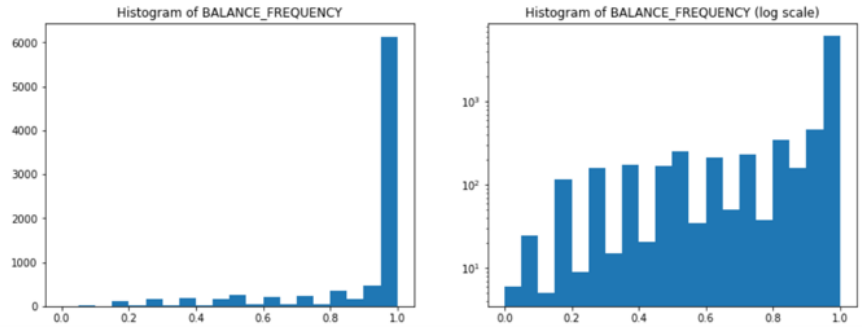


Figure 7: Histogram of BALANCE_FREQUENCY

Through observation, we found that "PURCHASES_FREQUENCY" and "PURCHASES_INSTALLMENTS_FREQ exhibit a high-low-high pattern, while "PRC_FULL_PAYMENT" and "ONEOFF_PURCHASES_FREQUENCY" exhibit a high-low-high pattern with a small cluster of samples at the high end (close to 1). "BALANCE_FREQUENCY" exhibits a clear left-skewed distribution.

- Categorical variables

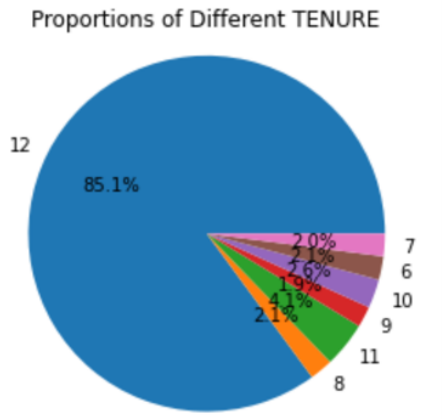


Figure 8: Pie Chart of TENURE

As shown in the graph, the "TENURE" variable is an ordered categorical variable and is the only categorical variable in our dataset. The majority of customers have a tenure of 12 months, accounting for 85.1% of the total sample.

1.4 Brief statistics description

In this dataset, the average balance across all customers is approximately $1,564.47, ranging from a minimum of $0 to a maximum of $19,043.14. On average, customers make purchases worth $1,003.20, with a minimum of $0 and a maximum of $49,039.57. The average credit limit is $4,494.45, ranging from a minimum of $50 to a maximum of $30,000.

Customers make an average payment of $1,733.14, with a minimum of $0 and a maximum of $50,721.48. A total of 86.32% of customers have a minimum payment amount, with an average minimum payment of $864.21. Additionally, 15.37% of customers pay their balance in full each month.

The average tenure of customers, i.e., how long they have been a customer, is 11.52 months. Overall, these statistics provide a valuable insight into the financial behavior of customers in this dataset.

In summary, our dataset contains various types of attributes with different distributions. By visualizing the data in different ways, we can gain insights into the characteristics of the data and identify potential issues that may need to be addressed indata analysis.

2 Customer Segmentation

2.1 Initial Clustering

We used our original dataset for the first clustering. We used Silhouette Method and the Elbow Method to find the optimal number of clusters.

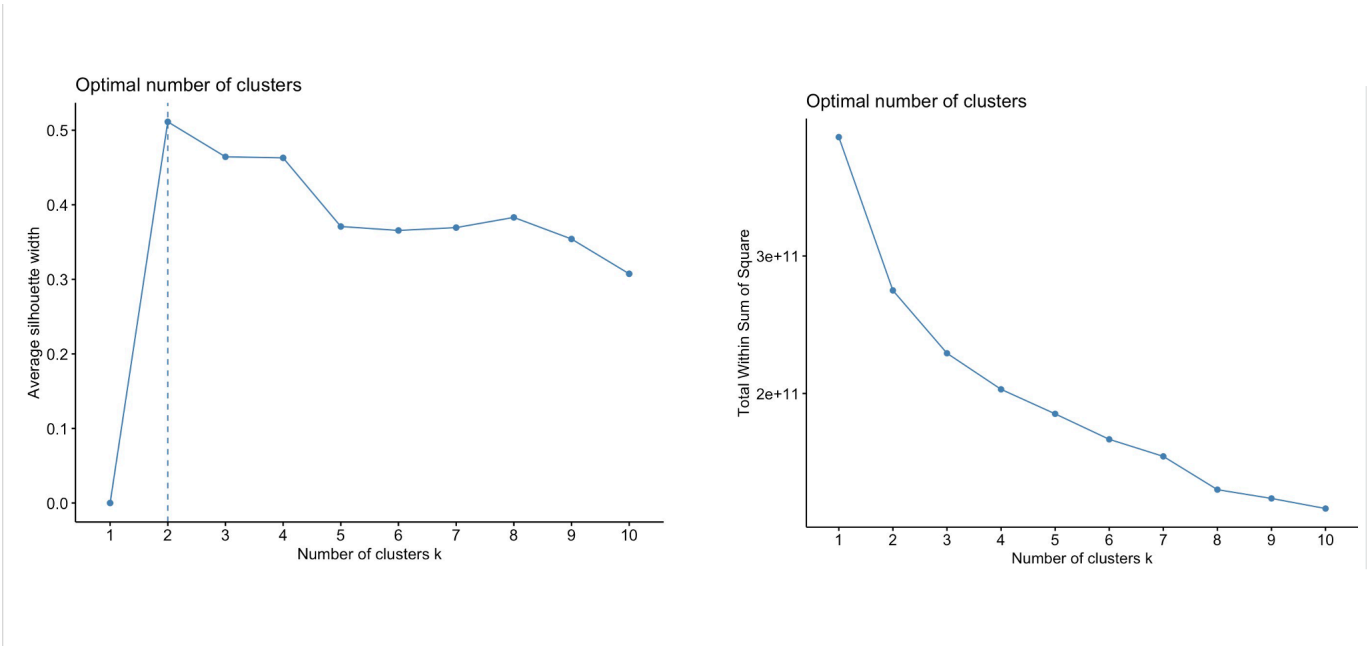


Figure 9: Results of Silhouette Method and the Elbow Method

From the results, we can see that the silhouette analysis is very determinable whereas it is more difficult to understand the plot generated when using the elbow method. It seems that two clusters are the optimal number.

We generated the cluster plot and realized that there are many outliers affecting the result but not giving much information, so we decided to remove the outliers by z-score method and remove them when the absolute value of their z-score exceeds 3. This has made our plot more readable.

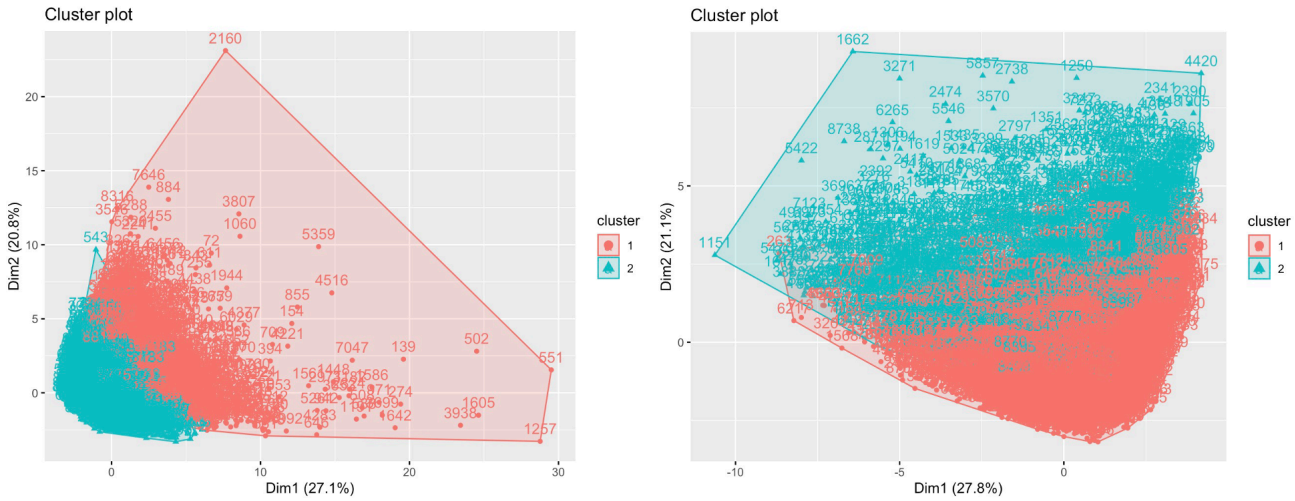


Figure 10: Results of Clustering before and after Outlier Removal

Then, we observed the mean values of the variables per cluster below, it is not difficult to determine that there are major differences in the clusters generated. It could take some time to understand the distinctiveness of each cluster.

	variable	cluster	BALANCE_…[1]	BALAN…[2]	PURCH…[3]	ONEOF…[4]	INSTA…[5]	CASH_…[6]	PURCH…[7]	ONEOF…[8]	PURCH…[9]	CASH_…×
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	1	846.	0.894	562.	283.	280.	458.	0.468	0.144	0.359	0.100
2	1	2	2304.	0.933	1221.	779.	442.	1186.	0.550	0.307	0.378	0.140

Figure 11: Values by Clusters

Generally, it seems that credit card holders in cluster 1 tend to spend more money more frequently. This group could be considered as the Big Spenders group. Their Credit Limits are higher than that of the Credit limits of the other group. The amount of PURCHASES made from the accounts in Cluster 1 are also much higher.

Also, the difference in frequency of ONEOFF PURCHASES is greater than the difference when viewing standard PURCHASES, meaning the BIG SPENDER Group appropriately earn their name as they will more frequently and spontaneously make purchases. This leads one to believe that perhaps these accounts are the accounts of high earners.

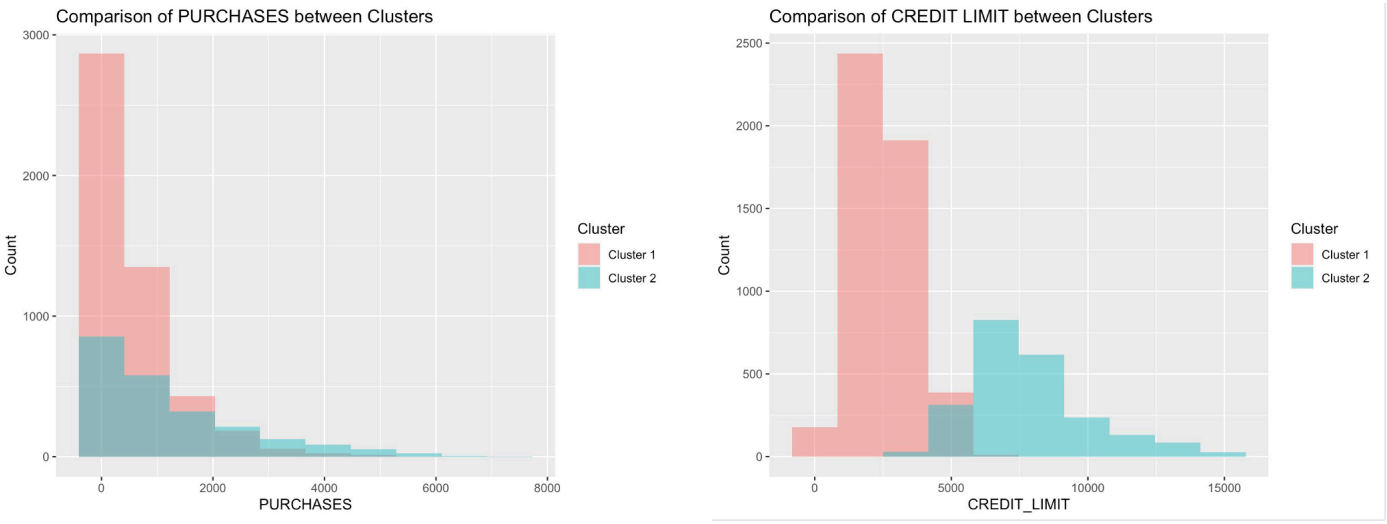


Figure 12: Comparison of Purchases and credit Limit by Clusters

2.2 Principal Component Analysis

The results above lead us to further think whether the type of purchases should be the leading factor in this case. We add four dummy variables in the dataset for installments, one-off, doing none, and doing both. Then we did a principal component analysis on this dataset.

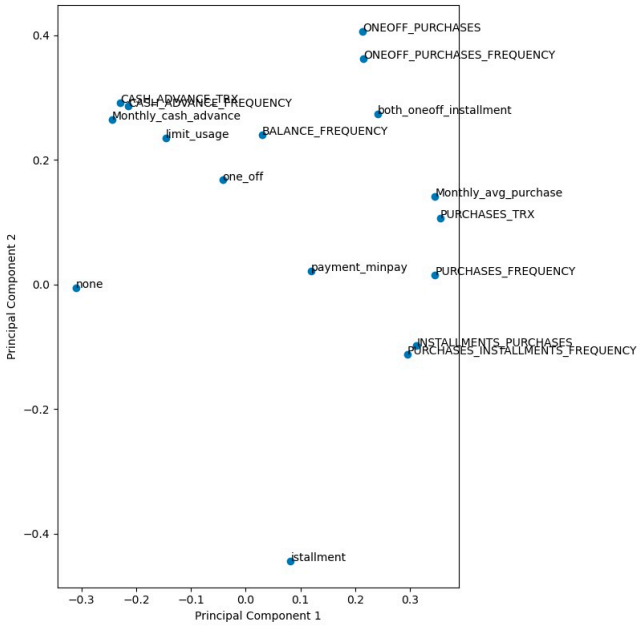


Figure 13: Visualization of the first two PCs

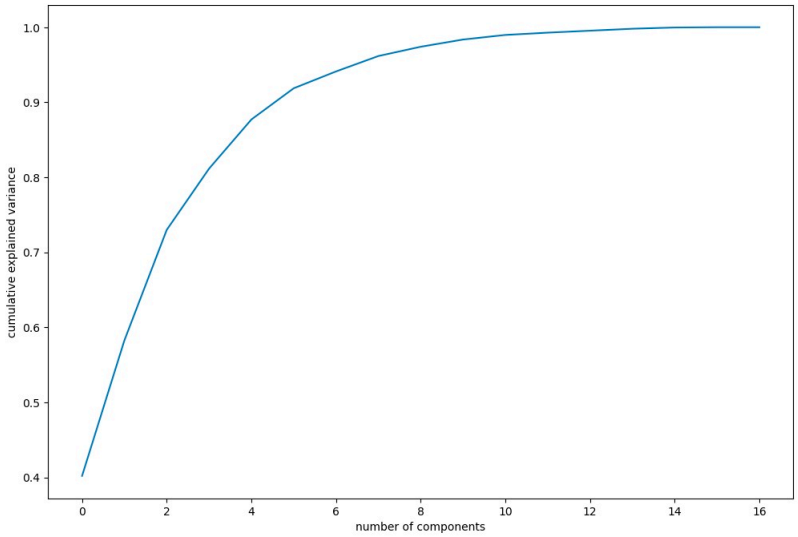


Figure 14: Plot of cumulative explained variance

From the plot, we can see that the total of the five first components can explain over 90 percent of the variance. So, we chose the first five PCs to represent our data.

2.3 Adjusted Clustering

On this dataset, we again used Silhouette Method and the Elbow Method to find the optimal number of clusters. From these two analyses, we found 3 and 4 seem to be the optimal number for our clustering, so we go with 4 for further K-means analysis.

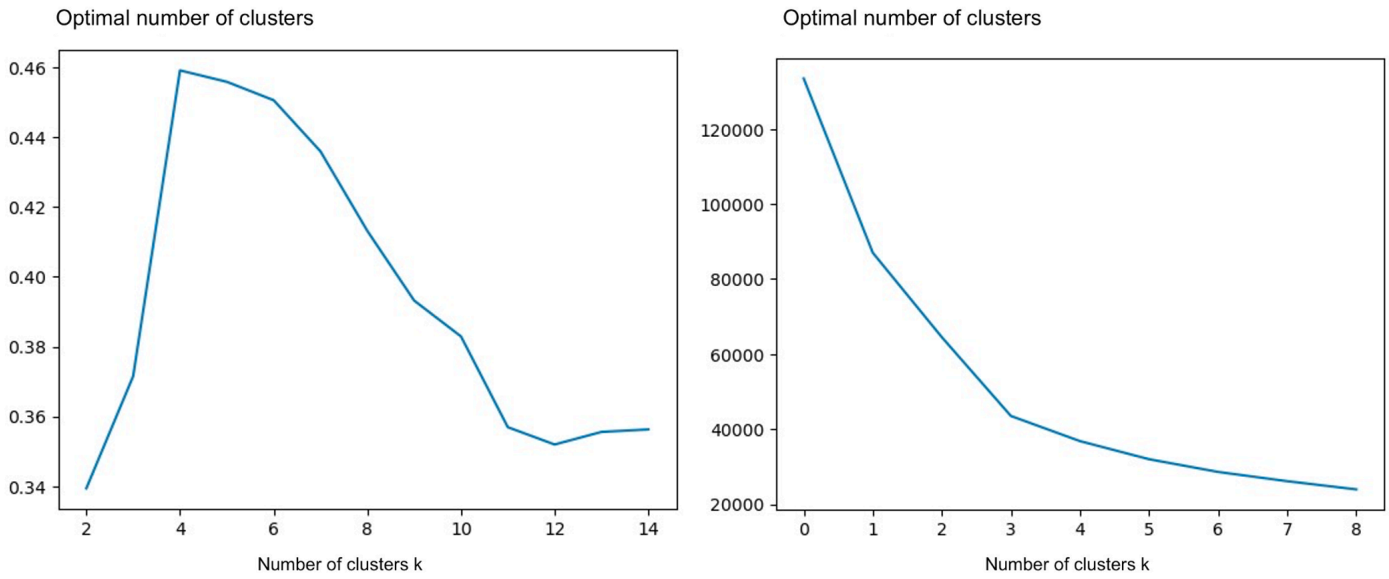


Figure 15: Results of Silhouette Method and the Elbow Method

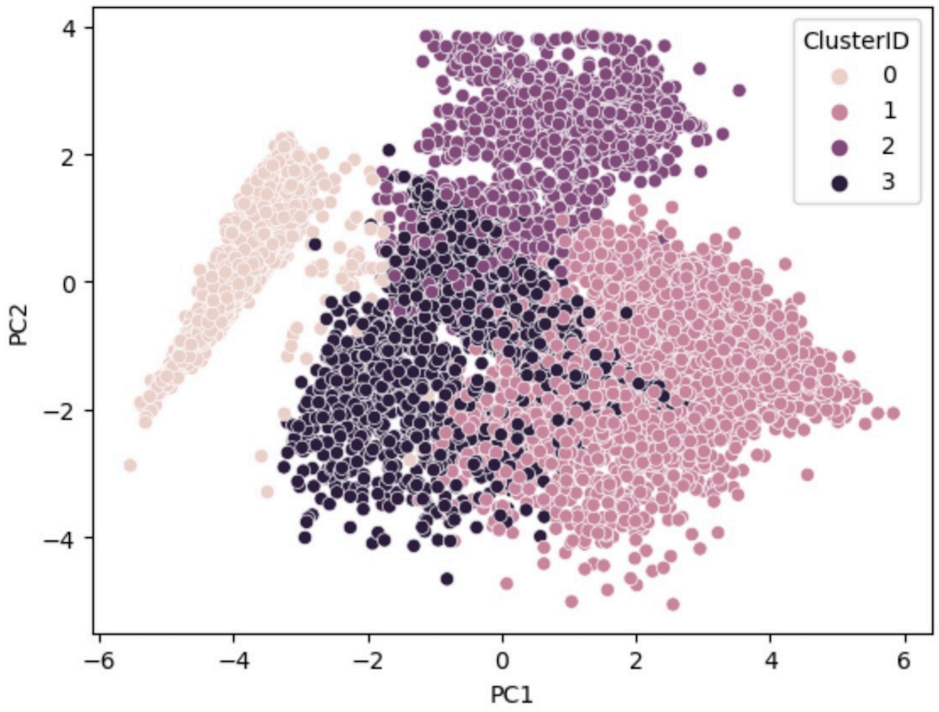


Figure 16: Adjusted Clustering Result

We can see that even though there are still many overlapping points, the total separation seems to be better than the previous one, so we take a closer look at the statistics of the customers in each cluster.

	0	BALANCE	PURCHASES	PURCHASES_FREQUENCY	PURCHASES_TRX	Monthly_avg_purchase	Monthly_cash_advance	limit_usage	CASH_ADVANCE_TRX	payment_minpay	both_oneoff_installment	istallment	one_off	none	CREDIT_LIMIT
0	0	2177.572394	1.859344	0.003758	0.045933	0.159337	186.298043	0.576217	6.552632	9.927979	0.002392	0.017225	0.003349	0.977033	4055.582137
1	1	1811.244331	2280.612488	0.802265	33.135292	193.759754	67.644533	0.354616	2.808125	7.264788	1.000000	0.000000	0.000000	0.000000	5748.836753
2	2	798.267082	543.583589	0.703035	12.051144	47.560400	33.474821	0.264156	1.018843	13.404629	0.002243	0.997757	0.000000	0.000000	3338.238396
3	3	1429.021056	787.353671	0.321170	7.118997	69.758276	77.843485	0.378727	2.864995	5.561421	0.003735	0.000000	0.996265	0.000000	4512.905630

Figure 17: Variables' values by Clusters

The clustering result by this method is highly interpretable. We noticed that Cluster 3 customers are doing maximum ONEOFF transactions and has the least payment ratio among all the clusters. Cluster 0 is the group of customers who have the highest Monthly cash advance and doing both installment as well as ONEOFF purchases, have comparatively good credit scores but have poor average purchase scores. Cluster 1 customers have maximum Average Purchase and good Monthly cash advance, but this cluster doesn't do installment or ONEOFF purchases. Cluster 2 is doing maximum installments, has a maximum payment to minimum payment ratio, and doesn't do one-off purchases.

3 Credit Limit Analysis

In this data, credit limit is very valuable, and it is very related to the credit rating of customers. Next, we conduct research around this data.

In this part, we used the distance correlation method to find the columns most correlated with the credit limit, and then used these data as feature values to classify customers using K-Means method. The optimal number of customer categories is determined by silhouette score and elbow method. Finally, according to the obtained classification, the data of customers under different categories were studied, and the customers were divided into low credit customers, medium credit customers and high credit customers, and the relevant characteristics and business suggestions were given.

Finally, we tried to use credit limit as the response variable and used the distance correlation method to find other relevant columns to fit it. We used a random forest to fit the model, and then we optimized it: we combined the previous classification, ran a random forest regression on each class, and adjusted the hyperparameters, which showed that we could improve the fit of the model very well.

3.1 Data Preprocessing

We found that some data in the whole data had too large Balance, so we eliminated them. At the same time, it is found that the cash-advance-frequency of some data is greater than 1, which is not practical. We also remove it.

3.2 Feature Selection

We use distance correlation to measure the relationship between variables and credit limit.

The distance correlation between two variables X and Y is defined as the normalized distance covariance between them:

$$dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X)dVar(Y)}}$$

where dCov(X,Y) denotes the distance covariance between X and Y, and dVar(X) and dVar(Y) denote the distance variances of X and Y, respectively.

The distance covariance between X and Y is defined as:

$$dCov(X,Y) = \sqrt{dCov(X,X)dCov(Y,Y)}$$

where dCov(X,X) and dCov(Y,Y) are the distance covariances of X and Y with themselves, respectively.

The distance covariance between X and itself is defined as:

$$dCov(X,X) = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}b_{ij}}$$

where n is the number of observations, and a_{ij} andb_{ij} are the distance matrices of X and Y, respectively.

The distance between two observations i and j in X is defined as:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

where p is the number of features in X, and x_{ik} and x_{jk} are the feature values of the i-th and j-th observations in X, respectively.

The distance matrix a_{ij} is then defined as:

$$a_{ij} = d_{ij} - \bar{d}i\cdot - \bar{d}\cdot j + \bar{d}..$$

where $\bar{d}i\cdot$ is the mean distance of observation i to all other observations in X, $\bar{d}\cdot j$ is the mean distance of observation j to all other observations in X, and $\bar{d}..$ is the mean distance between all pairs of observations in X.

The distance covariance between X and Y can be calculated in a similar way:

$$dCov(X,Y) = \sqrt{\frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}b_{ij}}$$

where b_{ij} is the distance matrix of Y, defined in the same way as a_{ij}.

Once the distance covariance and variances are calculated, the distance correlation coefficient can be obtained by dividing the distance covariance by the product of the distance standard deviations:

$$dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X)dVar(Y)}}$$

The advantages of distance correlation are as follows:

- Detection of non-linear correlation: distance correlation method can detect non-linear correlations, which cannot be captured by traditional correlation methods such as Pearson correlation coefficient. This allows distance correlation method to provide a more comprehensive description of the correlation between two vectors.

- Insensitivity to scale and transformation: distance correlation method is insensitive to the scale and transformation of the vectors, as it only focuses on the distance relationship between the vectors. This allows distance correlation method to handle vectors with different scales and units, and is not affected by linear transformations.

- No assumption of data distribution: distance correlation method does not require the assumption of data distribution, as it is a distance-based method that only considers the distance relationship between the vectors. This allows distance correlation method to be applied to various types of data, including continuous, discrete, ordered, and unordered data.

- Ability to handle high-dimensional data: distance correlation method can handle high-dimensional data, as it is a distance-based method that can consider all dimensions of the vectors. This allows distance correlation method to discover correlation patterns in high-dimensional data.

- Statistical significance and interpretability: distance correlation method has statistical significance and interpretability, as it can calculate the significance level and direction of the correlation. This allows distance correlation method to be widelyused in scientific research and data analysis.

The distance correlation results between different features and credit limit are obtained as follows (top 5 most relevant features):

	Balance	Payments 3	Purchases 4	Oneoff-Purchases	Cash-Advance
Correlation	0.4733	0.4183	0.3370	0.3263	0.2963

Table 1: distance correlations between credit limit features.

This is consistent with our cognition: credit card users with higher balance usually have better credit and have higher credit limit; credit card users with higher payment and purchase will also have higher credit limit.

3.3 Customer Clustering

We select the five most relevant columns based on distance correlation: Balance, Purchase, OneOff - Purchase, Payments, Cash - Advance and Credit - Limit are used as user characteristics to carry out the following research.

First, we compute the hopkins statistic to see if the dataset tends to cluster. The Hopkins statistic is a measure of how likely a dataset is to cluster, on a scale of 0 to 1. The closer the Hopkins statistic is to 1, the better fit the dataset is for clustering. When the Hopkins statistic is close to 0, the data set is not suitable for clustering.

The hopkins statistic is computed as: 0.97, which means that this dataset is a good fit for clustering.

Next, we used silhouette score analysis and elbow method to determine the optimal number of clusters.

The Silhouette score ranges from -1 to 1. When the Silhouette score is close to 1, it means that the distance between the sample and other samples in its cluster is small, and the distance between the sample and other clusters is large, so the clustering result is reasonable.

The elbow method is a common method used to determine the number of clusters. In the elbow method, we calculate the sum of squared errors (SSE) of the clustering model for different number of clusters and plot the SSE as the vertical axis and the number of clusters as the horizontal axis. We then look at the chart to find the "elbow" where the SSE reduction starts to slow down, which is the optimal value for choosing the number of clusters.

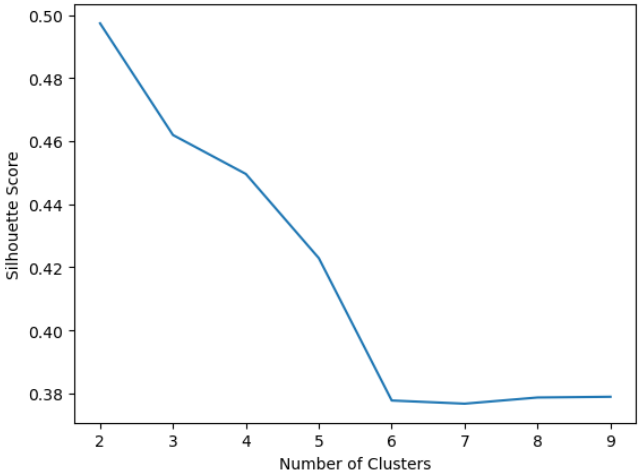


Figure 18: Silhouette Score

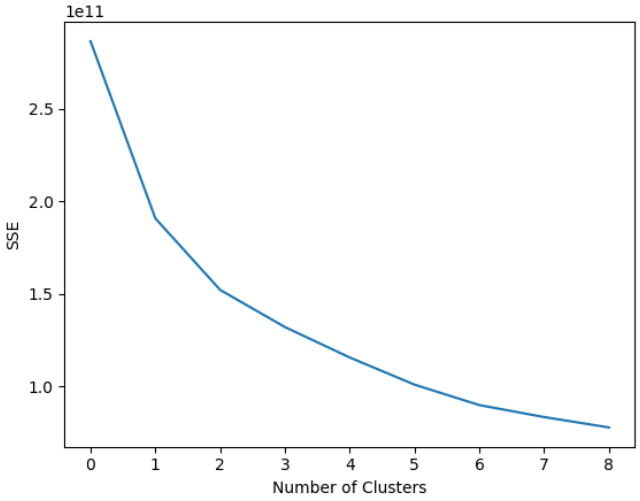


Figure 19: SSE

Figure 20: Best number of clusters

In summary, we choose the number of clusters to be three.

Then , we selected five variables related to Credit-limit and began clustering based on these variables using the KMeans clustering method. We also attempted other clustering methods, which will be discussed later.

KMeans clustering makes the following assumptions:

- Clusters are spherical, equally sized, and isotropic: KMeans assumes that the clusters in the data are spherical in shape and have the same size and variance. This means that the distance between any two points within a cluster should be roughly the same, and the distance between any two points in different clusters should be significantly different.

- Data has continuous features: KMeans assumes that the features in the data are continuous and can be represented as a matrix of real numbers.
- Data is normalized: KMeans assumes that the data is normalized or standardized, meaning that each feature has zero mean and unit variance.
- Number of clusters is known or can be estimated accurately: KMeans requires the number of clusters to be specified before the algorithm is run, or estimated using a suitable method.

Therefore, we standardized the data and began clustering. We first chose the most appropriate K value, which was 3, based on the mean squared error (MSE).

We obtained three clusters. We observed the mean values of the variables for each of the three clusters and found that the differences between the clusters were quite distinct.

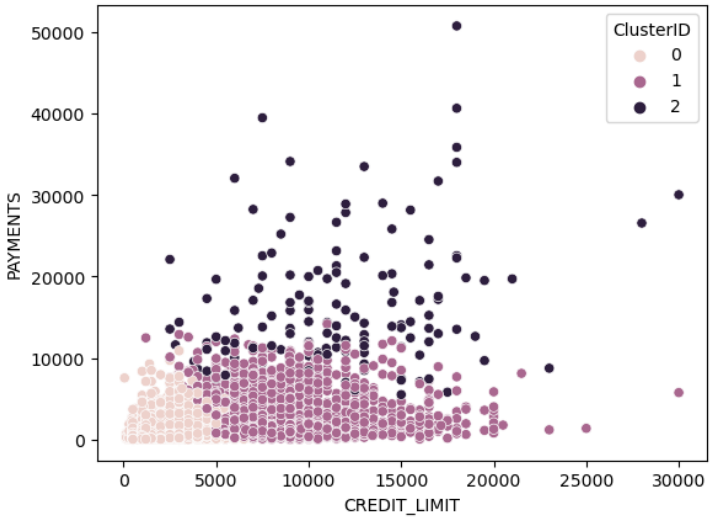


Figure 21: Credit-Limit & Balance

It can be seen from the picture that after using KMeans classification, the effect is good, and the data is significantly divided into three categories

Table 2: Classification result

labels	Balance	Purchases	OneOFF-Purchases	Payments	Credit Limit
0	881.702	599.272	305.348	1005.655	2539.638
1	2890.442	1461.768	905.475	2647.197	8487.17
2	3478.880	9283.352	6688.206	15771.521	11355.298

From the table, we can see that the average credit limit, balance, purchases, payments and one-off purchases of these three types of data are significantly different and show an increasing trend. As far as we know from experience, the larger the credit limit and the larger the balance, the higher the credit degree of the person can be proved. Therefore, the crowd can be divided into low credit group, medium credit group and high credit group through this classification.

3.4 Random Forest Regression

Credit limit data can be used as a standard to measure a user's credit, so it is very valuable to predict the user's credit limit according to the existing data.

We analyzed the original data and found that most of the data were non-normal and the linear relationship was not obvious. Therefore, we selected random forest as a non-parametric method for regression analysis, because it was most suitable for the nature of the data.

We constructed the model with the other 17 variables as independent variables and Credit-Limit as dependent variable.

First, we normalize the data to remove the effects of dimensionality, and then we use a random forest for regression. Our random forest model consists of 100 decision trees and then divides the data into training and test sets with a ratio of 8:2. After training on the training set and testing on the test set, the Mean Square Error (MSE) is 0.5272.

After that, we consider combining customer classification with random forest regression to optimize our model. We first fit a random forest model to each of the three categories of customers, resulting in three random forest models.In this way, for new data, we first use the classifier to classify it, and then put it in the random forest model of the predicted class. Finally, we predicted on the test set and got a mean squared error (MSE) of 0.3916。

We can see that this' piecewise prediction 'can greatly improve the accuracy of the model.

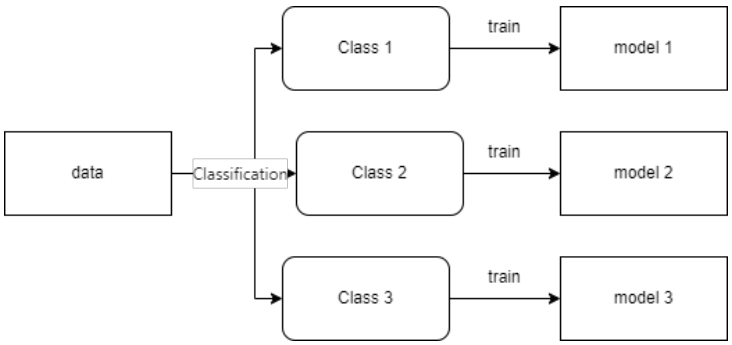


Figure 22: Training flow

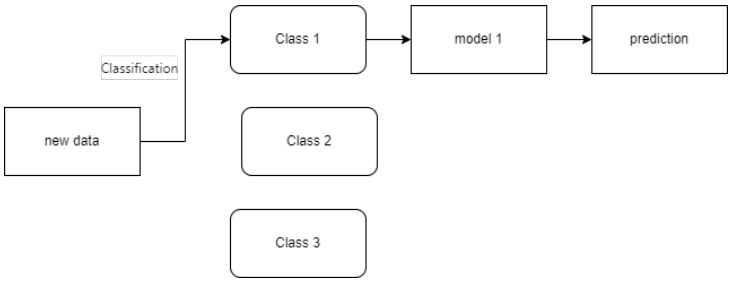


Figure 23: Prediction flow

4 Extension

4.1 Introduction to Fuzzy Clustering

Fuzzy clustering, also known as soft clustering, is a clustering technique that assigns a probability distribution to each data point indicating the likelihood of the point belonging to each cluster. This differs from traditional hard clustering techniques, such as k-means, which assign each data point to a single cluster.

The algorithm for fuzzy clustering can be described as follows:

Algorithm 1 Fuzzy C-Means (FCM)

Require: Data set $\mathbf{X} = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, number of clusters c, fuzzifier parameter m, stopping criterion ϵ

1: Initialize fuzzy membership matrix $\mathbf{U} = [u_{ij}]_{n \times c}$ randomly such that $\sum_{j=1}^{c} u_{ij} = 1$ for all $i = 1, 2, \ldots, n$

2: **repeat**

3: Update cluster centers $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_c]$ as

$$\mathbf{c}_j = \frac{\sum_{i=1}^{n} u_{ij}^m \mathbf{x}_i}{\sum_{i=1}^{n} u_{ij}^m} \quad \text{for} \quad j = 1, 2, \ldots, c \tag{1}$$

4: Update fuzzy membership matrix $\mathbf{U} = [u_{ij}]_{n \times c}$ as

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{||\mathbf{x}_i - \mathbf{c}_j||}{||\mathbf{x}_i - \mathbf{c}_k||} \right)^{\frac{2}{m-1}}} \tag{2}$$

5: **until** $||\mathbf{U}^{(t)} - \mathbf{U}^{(t-1)}|| < \epsilon$

In the above algorithm, $\mathbf{X}$ is the data set, c is the number of clusters, m is the fuzzifier parameter (a value greater than 1 that determines the degree of fuzziness), and ϵ is the stopping criterion (a small value that indicates when the algorithm has converged). The fuzzy membership matrix $\mathbf{U}$ is an $n \times c$ matrix, where u_{ij} represents the degree of membership of data point $\mathbf{x}_i$ to cluster $\mathbf{c}_j$. The cluster centers $\mathbf{C}$ are represented by an $n \times c$ matrix, where each column $\mathbf{c}_j$ represents the center of cluster j.

The algorithm iteratively updates the cluster centers and the fuzzy membership matrix until convergence. The update equations for the cluster centers and fuzzy membership matrix are given by equations (1) and (2), respectively.

Equation (1) shows that the cluster centers are updated as the weighted average of the data, where the degree of membership of each data point determines the weight assigned to it in the average. Specifically, the numerator of the equation represents a weighted sum of all data points, where the weight of each data point is given by its degree of membership to the j-th cluster raised to the power of m. The denominator represents the sum of the weights, which is used to normalize the weighted sum and obtain the updated center of the j-th cluster.

In summary, FCM is a "soft" clustering algorithm that assigns degrees of membership to each data point, allowing it to belong to multiple clusters at the same time. The fuzzy membership matrix and the update equations for the cluster centers enable FCM to handle uncertain and overlapping data.

4.2 Apply to Our Case

The FCM algorithm makes the following assumptions:

- Data should have a clear underlying structure: FCM assumes that the data has a clear underlying structure and can be partitioned into clusters. This means that the data should have some degree of similarity between points within a cluster and dissimilarity between points in different clusters.
- Data should be continuous: FCM assumes that the data is continuous and can be represented as a matrix of real numbers.
- Data should not have missing values: FCM assumes that the data does not have any missing values, as the algorithm cannot handle missing values.
- The number of clusters should be specified: FCM requires the number of clusters to be specified before the algorithm is run.

We observed that k-means clustering has some assumptions about the data, such as assuming that clusters are spherical and that the data is normalized. Our data may not meet these assumptions very well, so we attempted fuzzy clustering instead.

With a setting of K=3 and membership=2, fuzzy clustering partitioned the data into three clusters with sizes of 1855, 1314, and 5286. The silhouette score was calculated and compared to that of k-means clustering, with k-means having a score of 0.5008 and fuzzy clustering having a score of 0.3902. It can be seen that fuzzy clustering is slightly inferior.

However, the advantage of fuzzy clustering is that it has less strict assumptions about the data, which suggests that it has strong potential as an effective clustering tool.

5 Conclusion

5.1 Customer Segmentation

When doing customer segmentation, it initially revealed two distinct groups, with different purchasing behaviours. After variable adjustments and principal component analysis, a more nuanced four-cluster solution was found, with each group exhibiting unique spending habits and credit behaviors.

5.2 Credit Limit Analysis

We use distance correlation to select the data related to credit limit, and classify customers according to them. It is found that customers can be well divided into low, medium and high credit groups, and their characteristics and behavior habits are explained.

We use the random forest model to predict the credit limit value. We find that the prediction after grouping customers produces better results than the ungrouping prediction, which further confirms the effectiveness of customer group segmentation and also gives a credit limit prediction model with good performance.

We also explored fuzzy clustering, which makes less stringent assumptions on the original data. After comparing it with Kmeans, we finally selected the Kmeans results as the basis for the subsequent analysis.