**Topic-Clustering**

**DRAFT 2 - GROUP 1**

**Objectives:** develop a customer segmentation to define marketing strategy from credit card data

**INTRODUCTION**

What

Given the dataset This Project aims to develop a customer segmentation to define marketing strategy. The sample Dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months

Why:

Customer Segmentation is a classical problem , using clustering we will be able to make classify customers and see the optimal types of customers based on credit card information like purchases and credit history. This can help the marketing team to make informed decisions on their target audience.

How:

Using Multiple Clustering Techniques including K-means and Hierarchical Clustering

**Literature Review:**

**Normalization** is a technique for organizing data in a database. It is important that a database is normalized to minimize redundancy (duplicate data) and to ensure only related data is stored in each table. It also prevents any issues stemming from database modifications such as insertions, deletions, and updates

We normalized the data to minimize redundancy (duplicate data) and to ensure only related data is stored in each table. Our data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data, such as k-nearest neighbors, It also prevents any issues stemming from database modifications such as insertions, deletions, and updates.

**K-Means Clustering**: K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

There are three popular methods to find optimal clusters:

* Elbow Method
* Silhouette Method
* GAP Statistics

**Elbow**:In the Elbow method, it is plotted a line chart using within-groups the sum of squares of the model which is applied the Kmeans algorithm with the range of different k values. Because a line plot seems like an arm, this method is named Elbow Method. SSE is decreasing as increasing k value because the distance between clusters decreases. Our goal is to choose k value corresponding to the "Elbow point" in the chart. This point is also the minimum SSE value for our model.

**Silhouette**:In the Silhouette method, it is plotted a line chart like an Elbow Method using the range of different k values. While the Elbow Method tries to calculate the error between the clusters, the Silhouette Method tries to estimate the average distance between clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. therefore, our goal is to decide k value corresponding to a higher average distance between clusters.

**The Gap statistic**: Gap statistic is a standard method for determining the number of clusters in a set of data. The Gap statistic standardizes the graph of log(Wk), where Wk is the within-cluster dispersion, by comparing it

to its expectation under an appropriate null reference distribution of the data. We suggest to use Wk instead of log(Wk), and to compare it to the expectation of Wk under a null reference distribution. In fact, whenever a number fulfills the original Gap statistic inequality, this number also fulfills the inequality of a Gap statistic using Wk, but not vice versa.

**DESCRIPTION OF DATA SET**

This case requires to develop a customer segmentation to define marketing strategy. The  
sample Dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables.

Following is the Data Dictionary for Credit Card dataset :-

**CUST**ID : Identification of Credit Card holder (Categorical)

**BALANCE** : Balance amount left in their account to make purchases   
**BALANCEFREQUENCY** : How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)  
**PURCHASES** : Amount of purchases made from account  
**ONEOFF**PURCHASES : Maximum purchase amount done in one-go  
**INSTALLMENTSPURCHASES** : Amount of purchase done in installment  
**CASH**ADVANCE : Cash in advance given by the user  
**PURCHASESFREQUENCY** : How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)  
**ONEOFFPURCHASESFREQUENCY** : How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)  
**PURCHASESINSTALLMENTSFREQUENCY** : How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)  
**CASHADVANCEFREQUENCY** : How frequently the cash in advance being paid  
**CASHADVANCETRX** : Number of Transactions made with "Cash in Advanced"  
**PURCHASES**TRX : Number of purchase transactions made  
**CREDITLIMIT** : Limit of Credit Card for user  
**PAYMENTS** : Amount of Payment done by user  
**MINIMUM\_PAYMENTS** : Minimum amount of payments made by user  
**PRCFULLPAYMENT** : Percent of full payment paid by user  
**TENURE** : Tenure of credit card service for user

**Method Description:**

1. Import Libraries
2. **Import Data**
3. Data Processing –Describe the data find and remove null values from columns using **na.omit()** method and, Remove columns which are not needed like custID.
4. Perform EDA and Feature Engineering, Scaling the features for proper results using scale() function and visualized the dataset using Histogram and Correlation matrix.
5. KMeans Clustering :Even if the cluster number is not optimal for the dataset, the algorithm splits into the given number of groups. It can be got different results even if the same code runs, by changing the starting point of the model. Therefore randomization is the essential point of the Kmeans algorithm. We can start four centers for our model.
6. Determining Optimal Clusters: There is a problem that must be solved because the distribution of the size between clusters seems unbalanced. There may be several reasons for this, we could start with control the number of clusters whether the cluster number is true. There are different methods to find the optimal cluster number in the Kmeans algorithm.

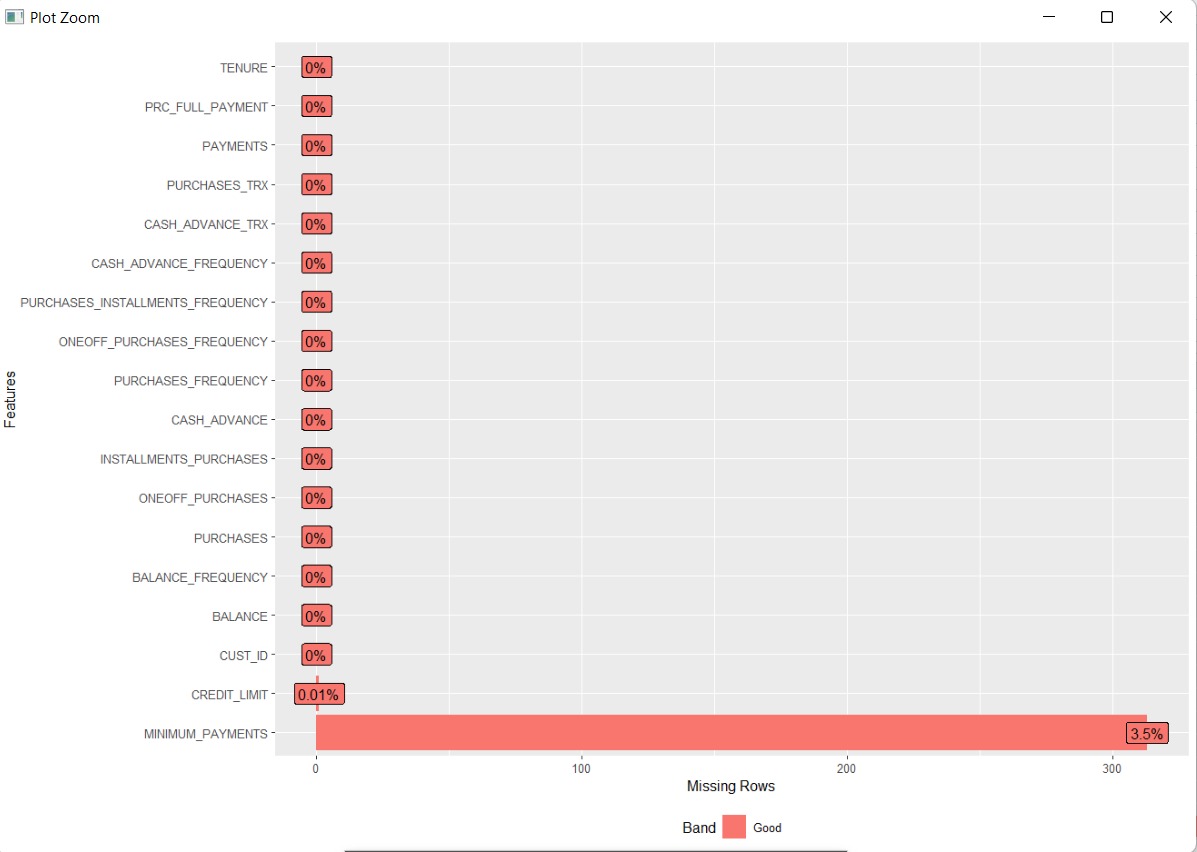
6a)Elbow Method

6b) Silhouette Method

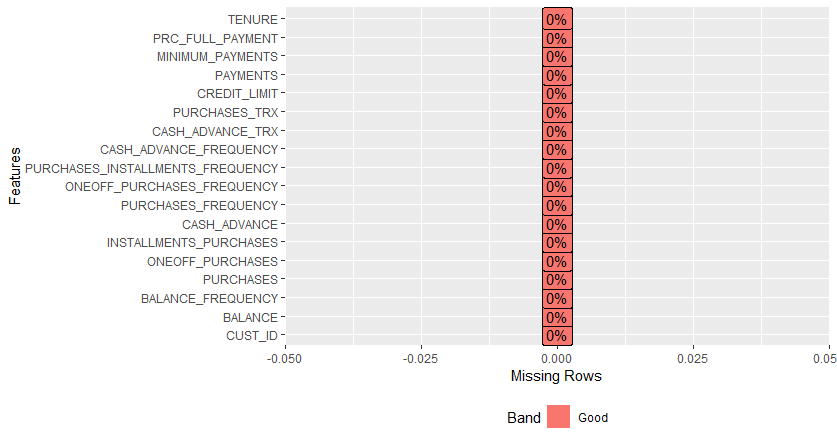
6c) GAP Statistics

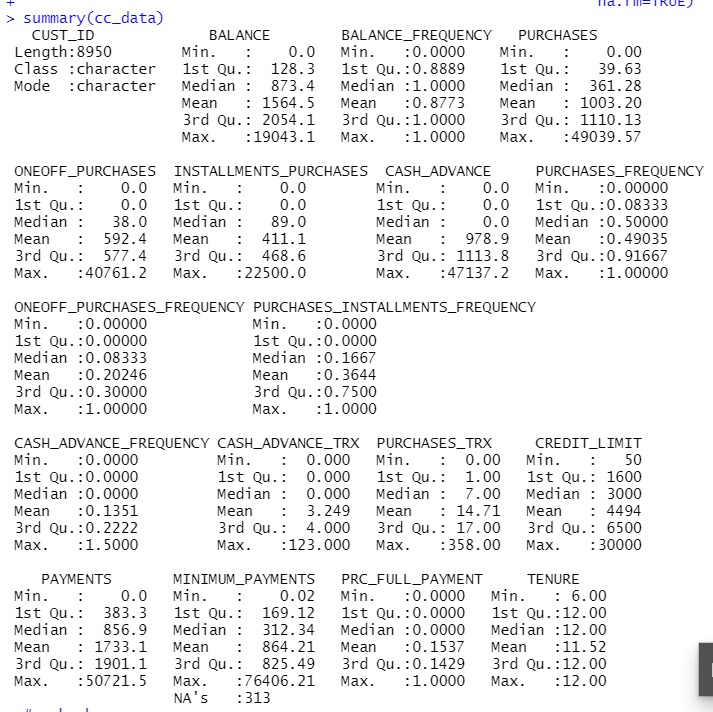
1. Results and clustering
2. Conclusion

Analysis

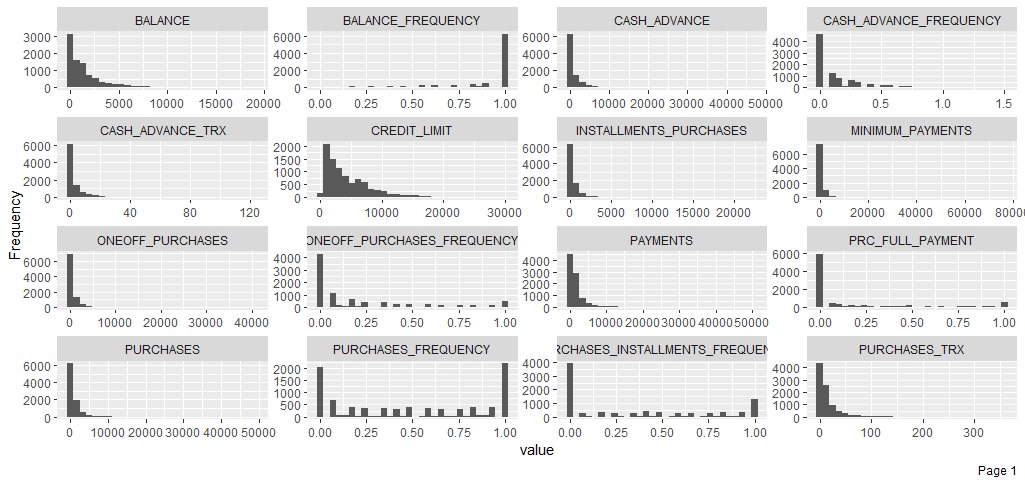
1 Data Cleaning Before:

Data after cleaning:

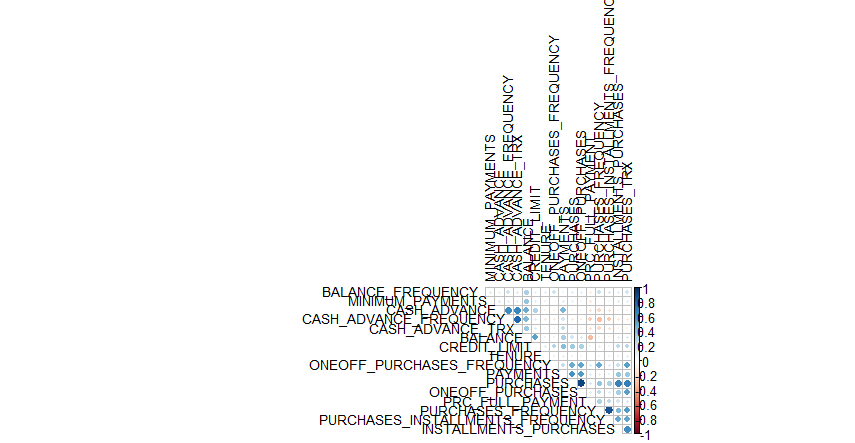


Data Description Summary:

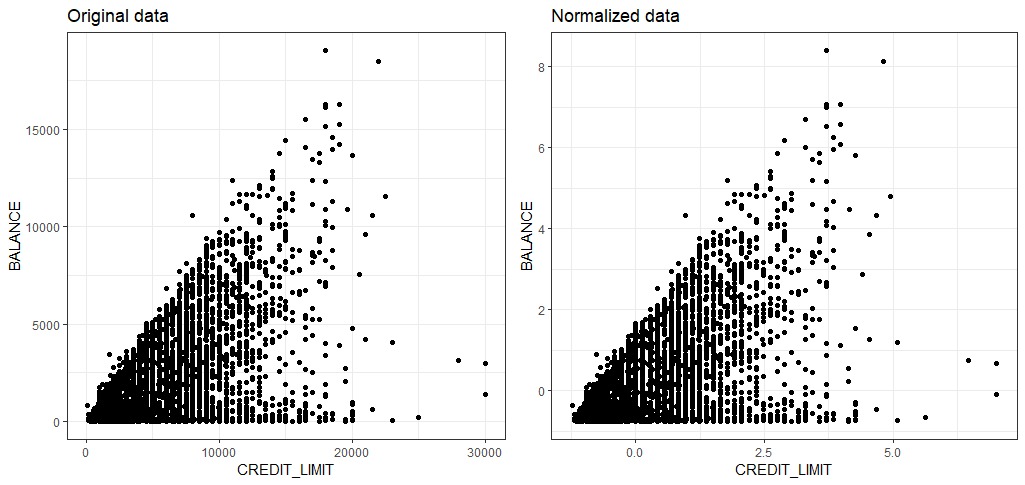
Histogram:

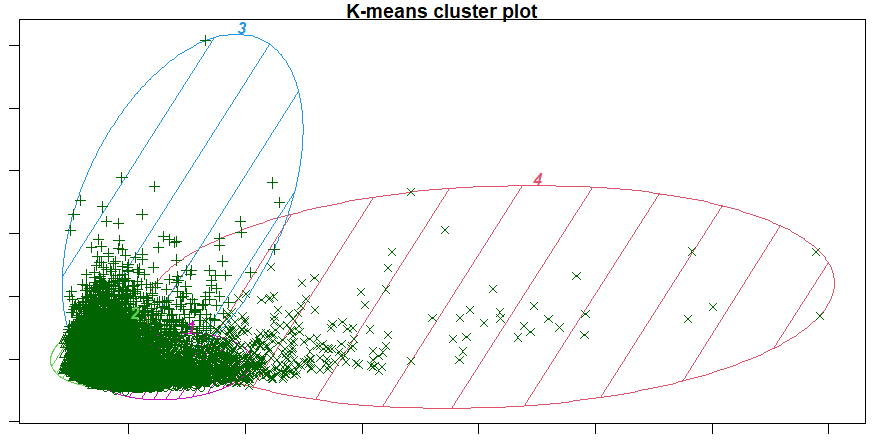


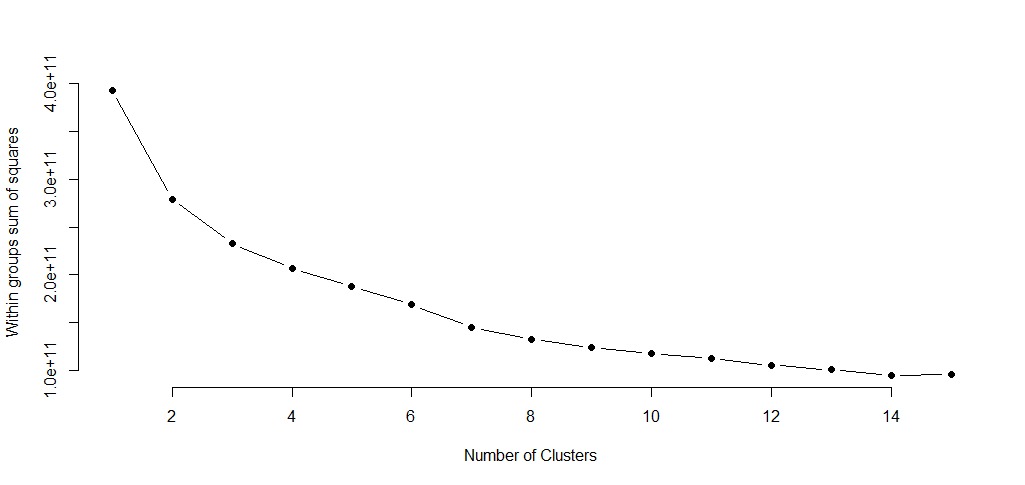
Correlation Matrix:



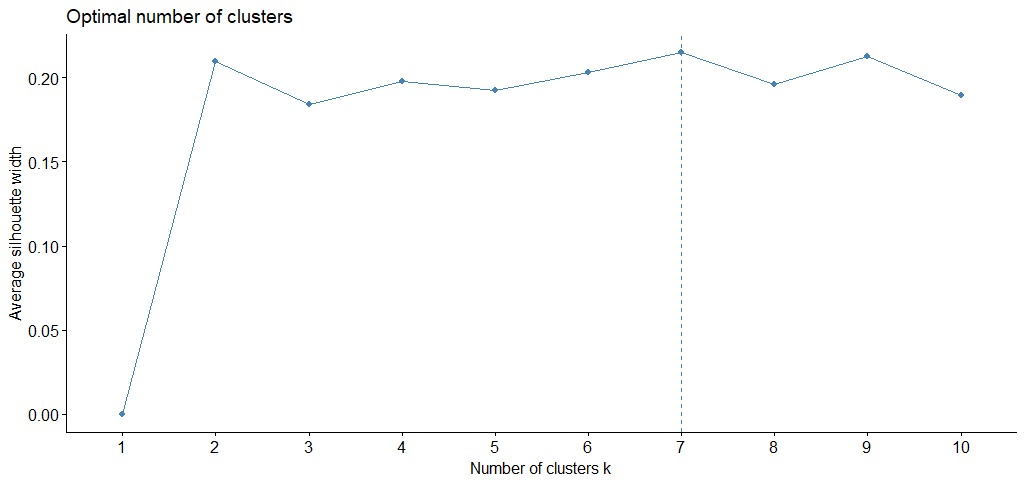
Normalized Data Vs Original Data:



Kmeans using 4 Clusters first:

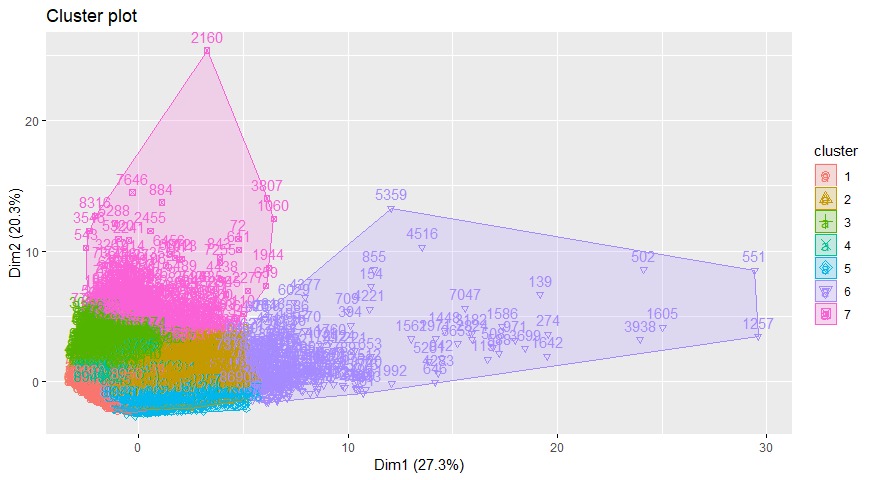
Elbow:

Silhouette:

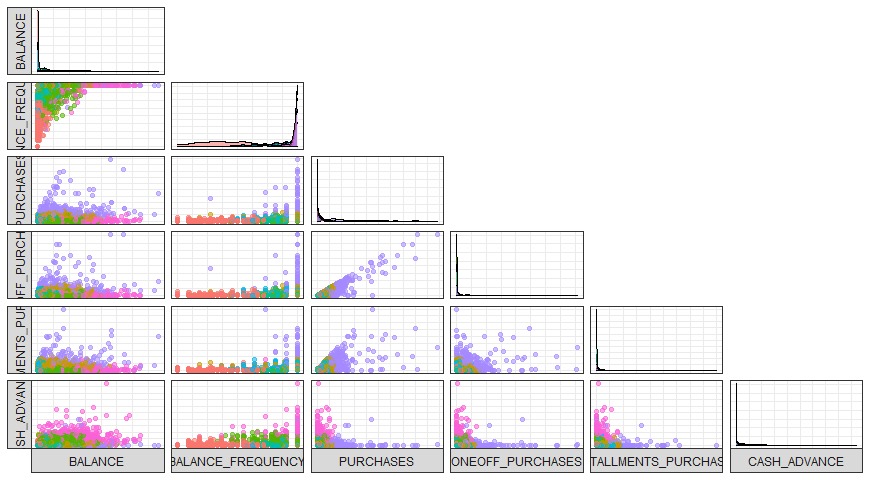


It seems seven clusters are good for our model.

Final Clusters:



Features Of Clusters:



RESULT:

The aim was to segment customers to define a marketing strategy. Below you can see the user groups that we have segmented according to the clusters we have reserved.

**1- Subclass user group(RED):** They are the users with the lowest cash. Credit limits are also low. They don't buy often. They do not perform banking transactions too much and their balances are not updated frequently.

**2- Balanced middle-class user group:** Users who are in the 2nd largest group in this regard when they are considered as one-time purchases, with no significant differences between cash and spending expenses. Installment amounts are above average and it is the second largest group in this regard. The number of purchases is quite high. They are in Group 2 in this field as well. They have an average credit limit.

**3- User group with the smallest expenditures:** They have the third highest balance. They work in advance. They have an average credit limit. Credit card service periods are the lowest.

**4- Small Spending and User Group with the Lowest Credit Limit:** These users are included in the group with the lowest credit limit, but they do not buy too much. They are the largest customer group.

**5- Credit card lovers group:** It is the lowest cash-buying group. They make an average of 80% of their repayments. This group is the third smallest in number and they do not keep their money in the bank.

**6- User group with big expenditures:** It is the second user group with the highest cash. They make expensive purchases and have the highest credit limit. It is the smallest customer group. 7. Users in the group are candidates for entering this group. Loyalty bank applications can be implemented to increase spending.

**7- Frugal user group w/ money:**It is the group with the highest cash. They generally like to trade with cash advance. They pay attention to their balances and expenses at the bank. It is the group that makes the highest payment in terms of reimbursement. Marketing transactions can be carried out in order to ensure that the transactions made in cash are returned to the credit card.