CREDIT CARD SEGMENTATION

Customer segmentation to define marketing strategy

APRIL 9, 2020

CHINMAY KUMAR PRUSTY
EDWISOR

Table of Contents

1. INTRODUCTION

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

Objectives:

1. Advanced data preparation. Build an ‘enriched’ customer profile by deriving ‘intelligent’ KPI’s such as monthly average purchase and cash advance amount, purchases by type (one-off, instalments), average amount per purchase and cash advance transaction, limit usage (balance to credit limit ratio), payments to minimum payments ratio etc.

2. Advanced reporting. Use the derived KPI’s to gain insight on the customer profiles.

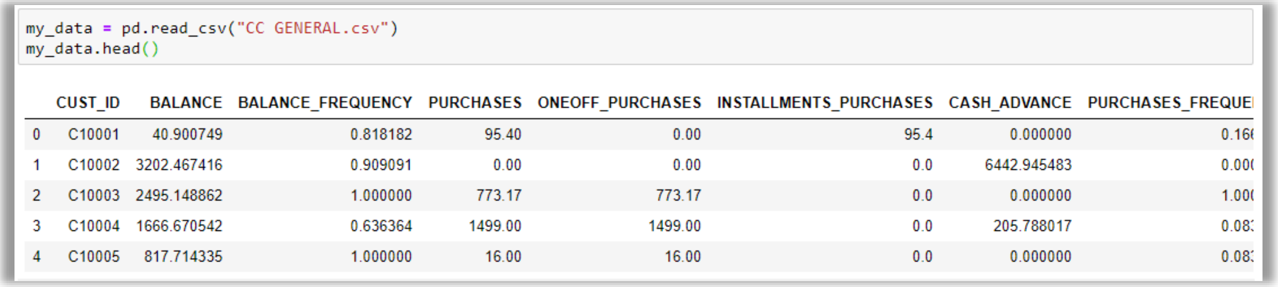
3. Clustering. Apply a data reduction technique factor analysis for variable reduction technique and a clustering algorithm to reveal the behavioural segments of credit card holders.

1.1 Problem Statement:

In this challenge, we need to segment the 9000 customers based on their credit card transaction history of the last 6 months. This falls under unsupervised clustering problem, which means we don’t have any target class here.

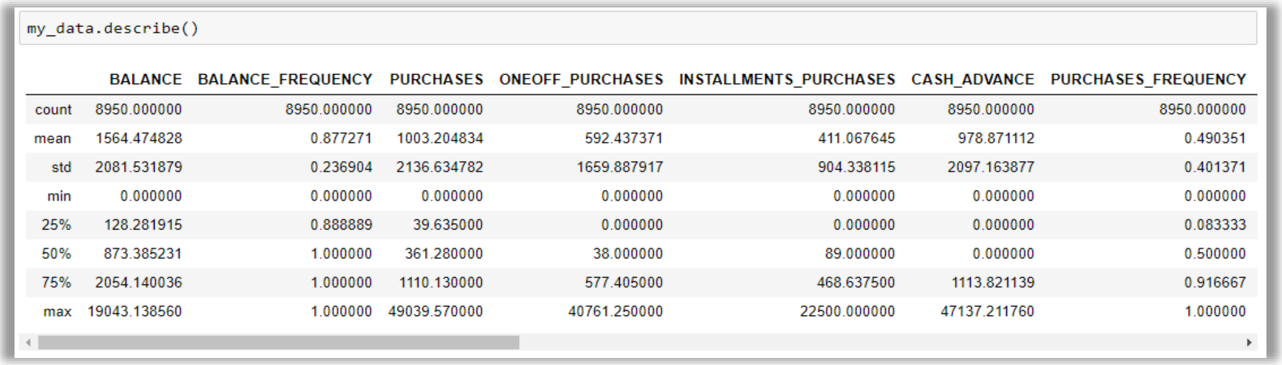
1.2 Dataset Information:

We are provided a dataset containing 8950 rows and 18 columns. These data describe the transaction details of credit card holders.



```
my_data = pd.read_csv("CC GENERAL.csv")
my_data.head()
```

	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUE
0	C10001	40.900749	0.818182	95.40	0.00	95.4	0.000000	0.16
1	C10002	3202.467416	0.909091	0.00	0.00	0.0	6442.945483	0.00
2	C10003	2495.148862	1.000000	773.17	773.17	0.0	0.000000	1.00
3	C10004	1666.670542	0.636364	1499.00	1499.00	0.0	205.788017	0.08
4	C10005	817.714335	1.000000	16.00	16.00	0.0	0.000000	0.08



```
my_data.describe()
```

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY
count	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000
mean	1564.474828	0.877271	1003.204834	592.437371	411.067645	978.871112	0.490351
std	2081.531879	0.236904	2136.634782	1659.887917	904.338115	2097.163877	0.401371
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	128.281915	0.888889	39.635000	0.000000	0.000000	0.000000	0.083333
50%	873.385231	1.000000	361.280000	38.000000	89.000000	0.000000	0.500000
75%	2054.140036	1.000000	1110.130000	577.405000	468.637500	1113.821139	0.916667
max	19043.138560	1.000000	49039.570000	40761.250000	22500.000000	47137.211760	1.000000

1.3 DATA DICTIONARY:

1. CUST_ID: Identification of Credit Card holder (Categorical)

2. BALANCE: Monthly average Balance amount left in their account to make purchases

3. BALANCE_FREQUENCY: How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)

4. PURCHASES: Amount of purchases made from account during last 12 months

5. ONEOFF_PURCHASES: Maximum purchase amount done in one-go

6. INSTALLMENTS_PURCHASES: Amount of purchase done in installment

7. CASH_ADVANCE: Cash in advance given by the user

8. PURCHASES_FREQUENCY: How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)

9. ONEOFF_PURCHASES_FREQUENCY: How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)

10. PURCHASES_INSTALLMENTS_FREQUENCY: How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)

11. CASH_ADVANCE_FREQUENCY: How frequently the cash in advance being paid

12. CASH_ADVANCE_TRX: Average amount per cash advance transaction

13. PURCHASES_TRX: Average amount per purchase transaction

14. CREDIT_LIMIT: Limit of Credit Card for user

15. PAYMENTS: Amount of Payment done by user. Total payments(due amount paid by the customer to decrease their statement balance) in the period.

16. MINIMUM_PAYMENTS: Minimum amount of payments made by user

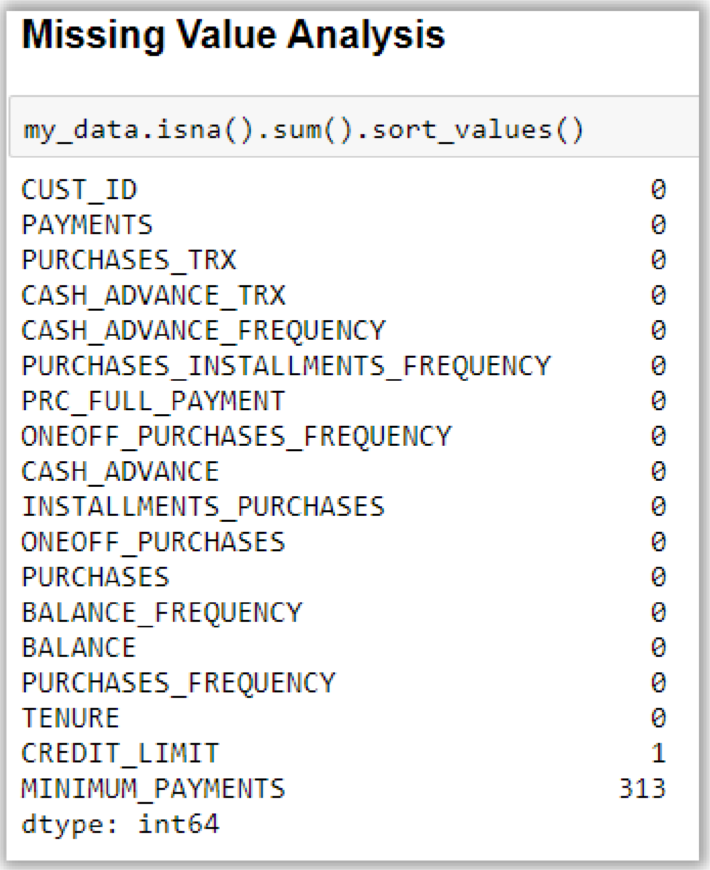
17. PRC_FULL_PAYMENT: Percent of full payment paid by user

18. TENURE: Tenure of credit card service for user

2. DATA PREPROCESSING AND EDA

2.1 Missing Value Analysis:

Here we will be looking for missing values in each column of the dataset. And if there are any missing values, we will impute them with mean, median or KNN method.



Missing Value Analysis

```
my_data.isna().sum().sort_values()
```

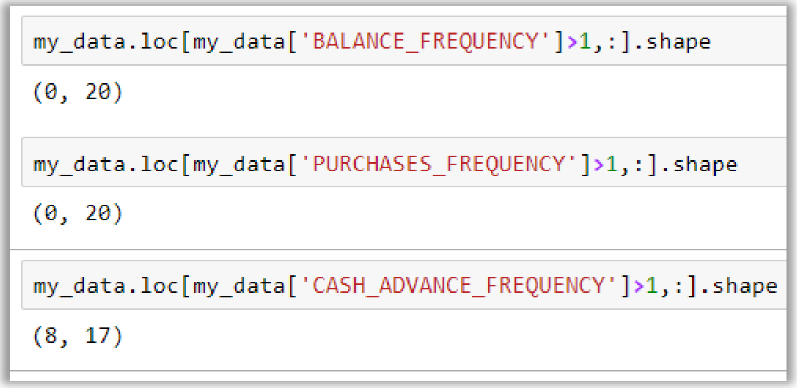
```
CUST_ID                             0
PAYMENTS                            0
PURCHASES_TRX                       0
CASH_ADVANCE_TRX                    0
CASH_ADVANCE_FREQUENCY              0
PURCHASES_INSTALLMENTS_FREQUENCY    0
PRC_FULL_PAYMENT                    0
ONEOFF_PURCHASES_FREQUENCY          0
CASH_ADVANCE                        0
INSTALLMENTS_PURCHASES              0
ONEOFF_PURCHASES                    0
PURCHASES                           0
BALANCE_FREQUENCY                   0
BALANCE                             0
PURCHASES_FREQUENCY                 0
TENURE                              0
CREDIT_LIMIT                        1
MINIMUM_PAYMENTS                  313
dtype: int64
```

CREDIT_LIMIT and MINIMUM_PAYMENTS were having missing values as shown in the figure above.

In this case, I have used median method to impute missing values since these two columns were skewed, and the presence of outliers would have a greater impact on the mean method.

2.2 Cleaning Noise:

We know that the value of frequency columns lies between 0 – 1. So, wherever the value is greater than 1 we will drop them.



```
my_data.loc[my_data['BALANCE_FREQUENCY']>1,:].shape
```

```
(0, 20)
```

```
my_data.loc[my_data['PURCHASES_FREQUENCY']>1,:].shape
```

```
(0, 20)
```

```
my_data.loc[my_data['CASH_ADVANCE_FREQUENCY']>1,:].shape
```

```
(8, 17)
```

From the above analysis, I found that CASH_ADVANCE_FREQUENCY is having 8 rows where the value is greater than 1. So, I dropped these rows.

The shape of the data after cleaning is: (8942, 17)

2.3 Generating New KPI's:

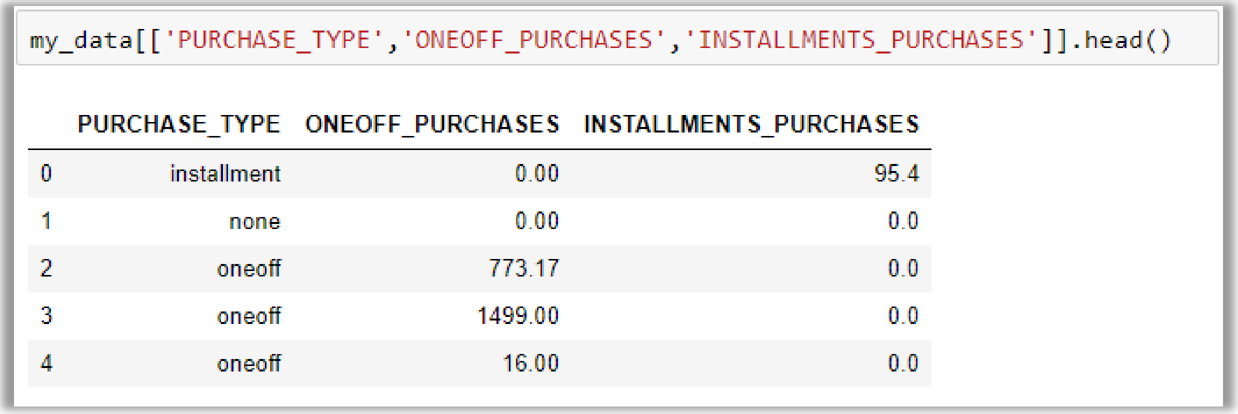
1. Monthly average purchase and monthly cash advance amount

This is calculating by dividing the values of monthly average purchase and monthly cash advance with the value of tenure.

2. Purchase by type (one-off, installments)

Here we have 4 categories:

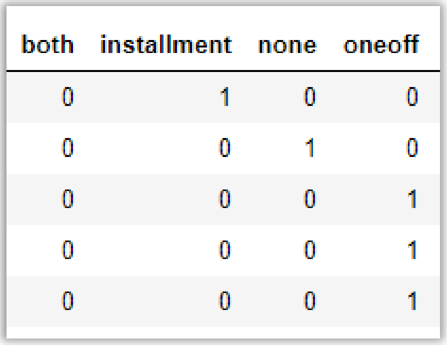
- **None:** one-off purchases = 0 and installment purchases = 0
- **One-Off:** one-off purchases > 0 and installment purchases = 0
- **Installment:** one-off purchases = 0 and installment purchases > 0
- **Both:** one-off purchases > 0 and installment purchases > 0



```
my_data[['PURCHASE_TYPE','ONEOFF_PURCHASES','INSTALLMENTS_PURCHASES']].head()
```

	PURCHASE_TYPE	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES
0	installment	0.00	95.4
1	none	0.00	0.0
2	oneoff	773.17	0.0
3	oneoff	1499.00	0.0
4	oneoff	16.00	0.0

Since there are 4 categories of PURCHASE_TYPE, I have created 4 dummy columns of each purchase type.



both	installment	none	oneoff
0	1	0	0
0	0	1	0
0	0	0	1
0	0	0	1
0	0	0	1

3. Limit usage (balance to credit limit ratio)

A **credit limit** is the maximum amount that you can spend with a **credit card**. Having high **limits** lets you spend more and can be good for your **credit** scores. And Balance is the amount left in their account to make purchases. So, having more limit usage means the user is more likely to have good credit score.

4. Payment to minimum payments ratio

The **minimum payment** is the lowest **amount** of money that you are required to **pay** on your **credit card** statement each month

Payment is the amount paid by the user.

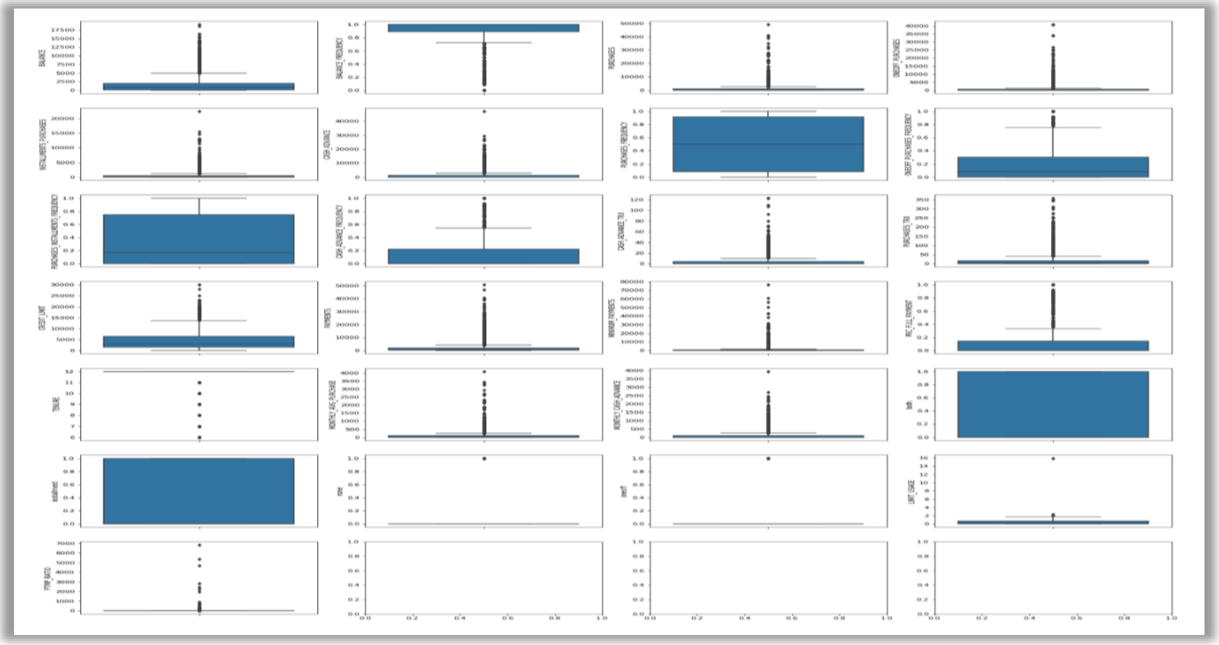
Hence if the ratio is having high value, it means the customer is clearing his debts timely.

2.4 Outlier Analysis

What is an Outlier?

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error. An outlier can cause serious problem while training the predictive model.

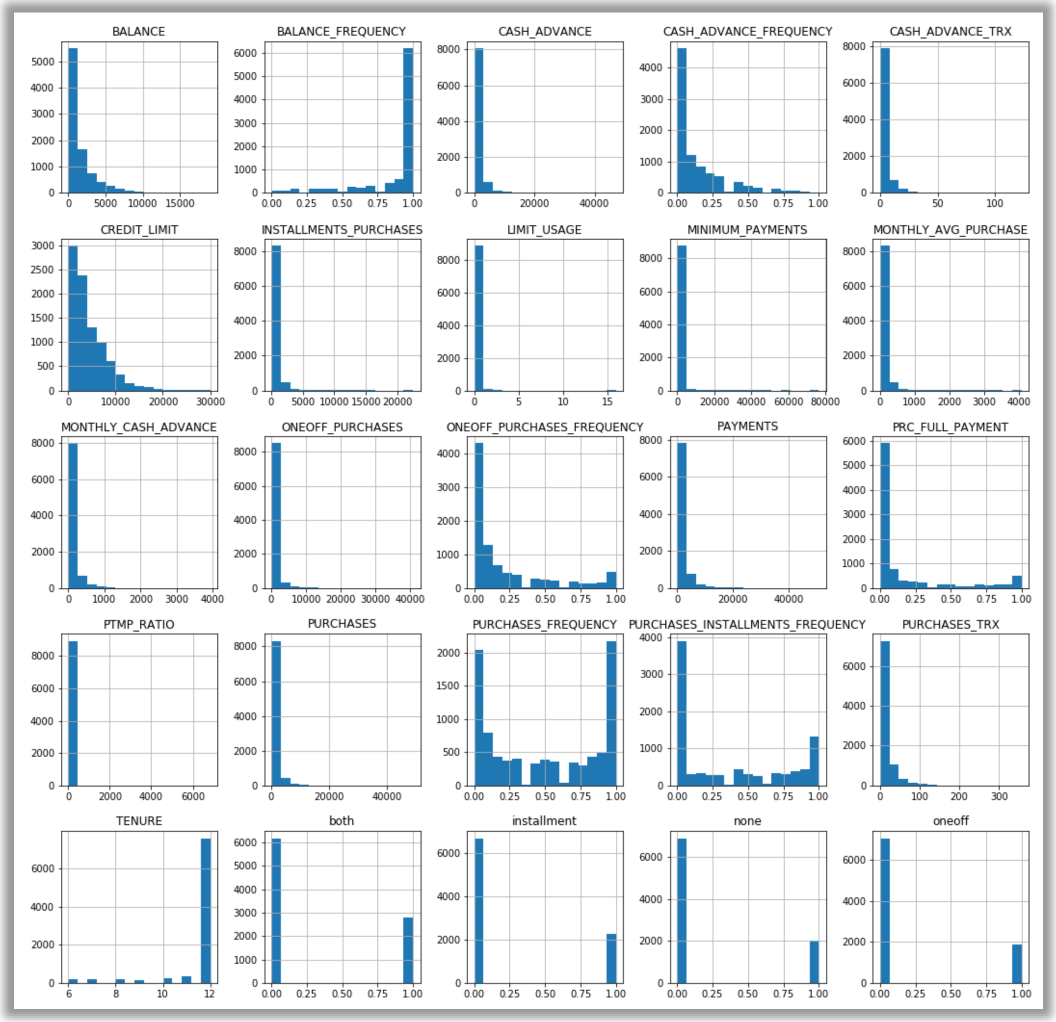
I have used Boxplot to visualize outliers.



From the above visualization it is clear that most of the variables are having outliers.

But then, so many outliers are there and we don’t want to lose information. So, I will keep the outliers for now.

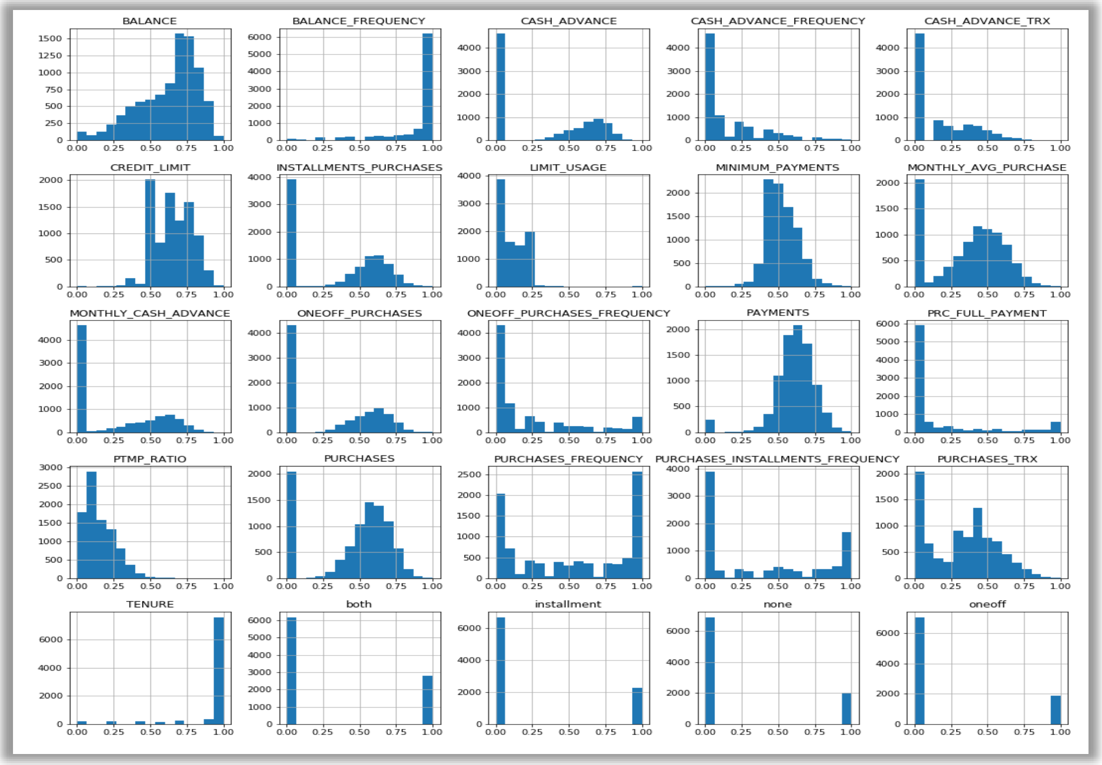
2.5 Checking the distribution of data (using histogram):



From the above visualization, I can see that most of the data is positively skewed.

So I have applied log(x+1) transformation to remove the outlier effect.

2.6 Distribution after Log Transformation:



2.5 Feature Scaling (Normalization):

What is Normalization?

It is the process of reducing unwanted variables either within or between the variables. It brings all of the variables into proportion with one another. The values range between 0 to 1.

It is sensitive to outliers, that’s why I have first treated outliers before performing this step.

Formula for normalization:

Values(new) = (Value – min_Value)/(Max_value – min_Value)

2.6 Correlation Analysis

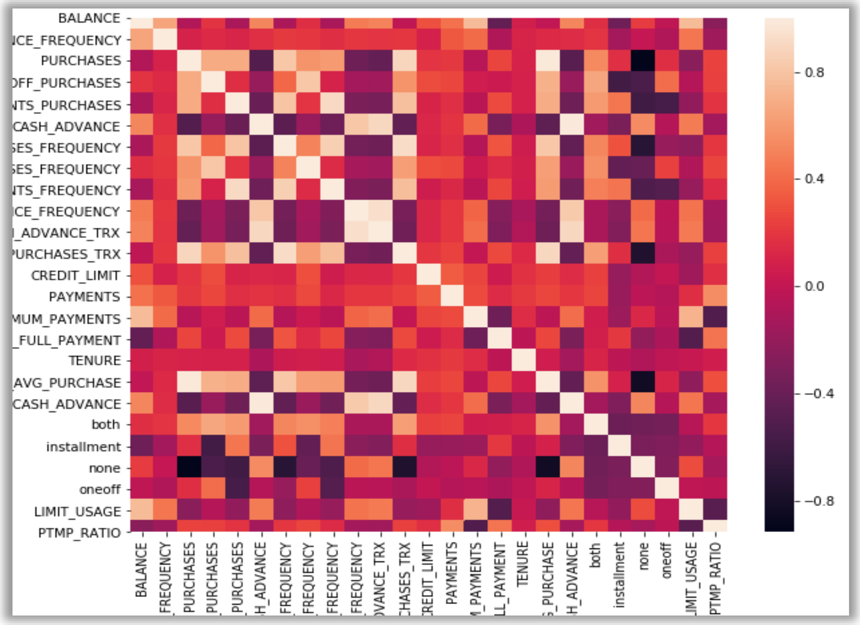
Correlation is a **statistical** technique that can show whether and how strongly pairs of variables are related.

Types of Correlation

- Positive Correlation – when the value of one variable increases with respect to another.
- Negative Correlation – when the value of one variable decreases with respect to another.
- No Correlation – when there is no linear dependence or no relation between the two variables.

What is multicollinearity?

Multicollinearity is a state of very high intercorrelations or inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable.



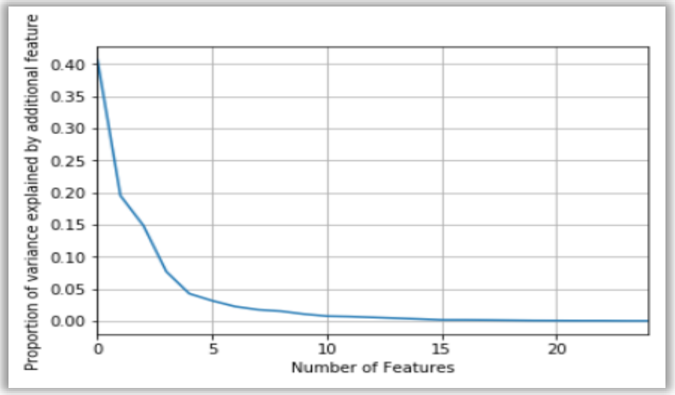
Many variables are highly correlated to each other, making it possible to use PCA for reducing the dimension.

What is Dimensionality Reduction?

Dimensionality reduction is the process of **reducing** the number of random variables under consideration, by obtaining a set of principal variables.

I have used PCA to reduce the dimensions.

PCA (Principal Component Analysis) is a projection-based method which transforms the data by projecting it onto a set of orthogonal axes.



Since 2 variables are explaining most of the data we will take 2 PCA.

3. Modelling:

We will be using K Means Clustering and Hierarchical Clustering algorithms for this dataset.

What is Clustering??

Clustering is basically a technique that groups similar data points such that the points in the same group are more similar to each other than the points in the other groups. The group of similar data points is called a Cluster.

What is K Means Clustering?

K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into **k clusters** in which each observation belongs to the **cluster** with the nearest **mean** (**cluster** centers or **cluster** centroid), serving as a prototype of the **cluster**.

What is Hierarchical Clustering?

Hierarchical clustering is one of the popular and easy to understand clustering technique. This clustering technique is divided into two types:

1. Agglomerative
2. Divisive

I have used the Agglomerative technique**.**
Agglomerative Hierarchical clustering Technique: In this technique, initially each data point is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed.

The basic algorithm of Agglomerative is straight forward.

- Compute the proximity matrix
- Let each data point be a cluster
- Repeat: Merge the two closest clusters and update the proximity matrix
- Until only a single cluster remains

3.1 Evaluation Metrics:

I have covered two metrics in this report:

- Elbow method
- Silhouette analysis

3.1.1 Elbow Method

Elbow method gives us an idea on what a good *k* number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. We pick *k* at the spot where SSE starts to flatten out and forming an elbow.
Sometimes it's still hard to figure out a good number of clusters to use because the curve is monotonically decreasing and may not show any elbow or has an obvious point where the curve starts flattening out.

3.1.2 Silhouette Analysis

Silhouette analysis can be used to determine the degree of separation between clusters. For each sample:

- Compute the average distance from all data points in the same cluster (ai).
- Compute the average distance from all data points in the closest cluster (bi).
- Compute the coefficient:

$$\frac{b^i - a^i}{max(a^i, b^i)}$$

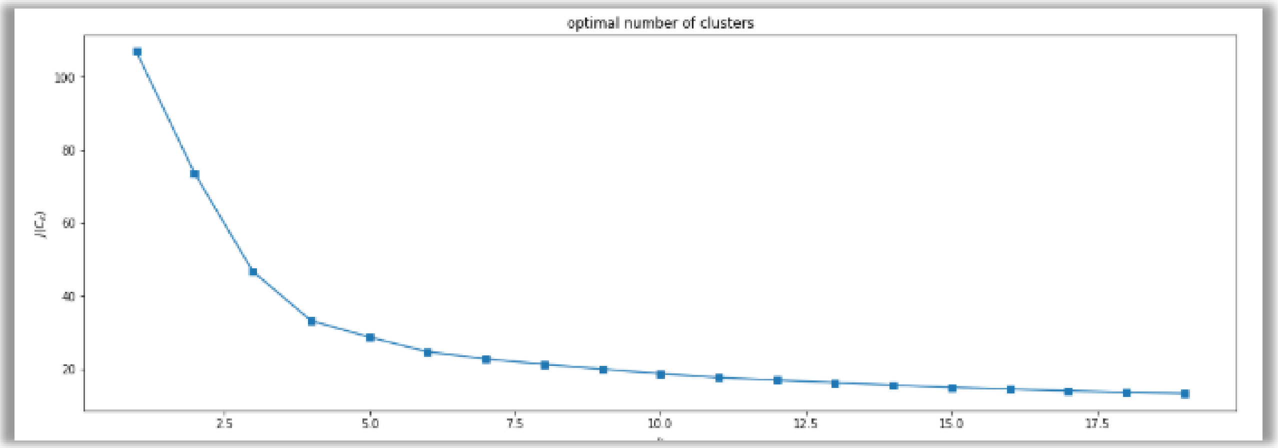
The coefficient can take values in the interval [-1, 1].

- If it is 0 –> the sample is very close to the neighboring clusters.
- If it is 1 –> the sample is far away from the neighboring clusters.
- If it is -1 –> the sample is assigned to the wrong clusters.

Therefore, we want the coefficients to be as big as possible and close to 1 to have good clusters.

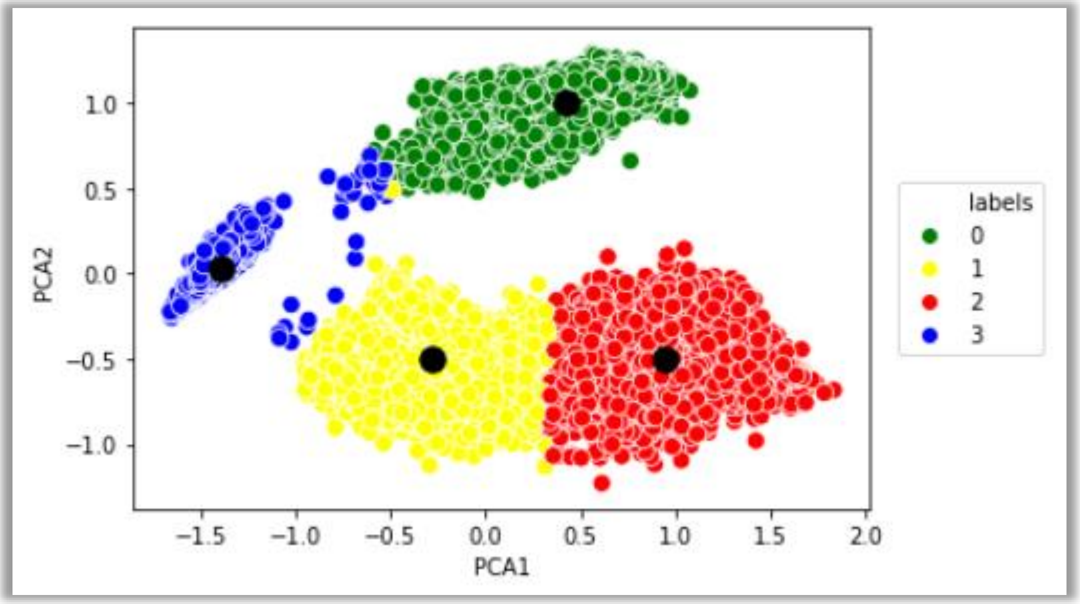
3.2 K Means Clustering

3.2.1 Identifying the optimum number of clusters



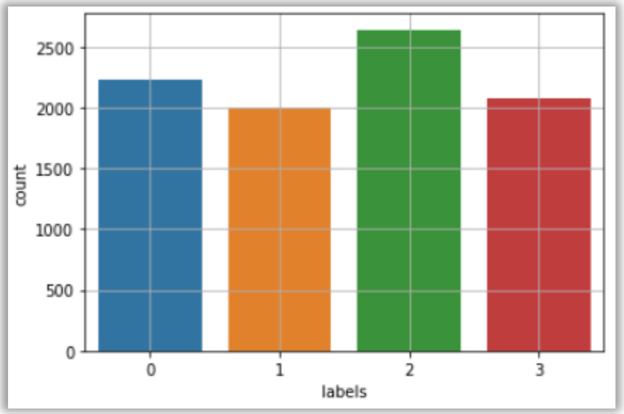
An elbow is observed at 4th point, so I will consider 4 as the optimal no. of clusters for our model.

3.2.2 Model 1(with 4 clusters):

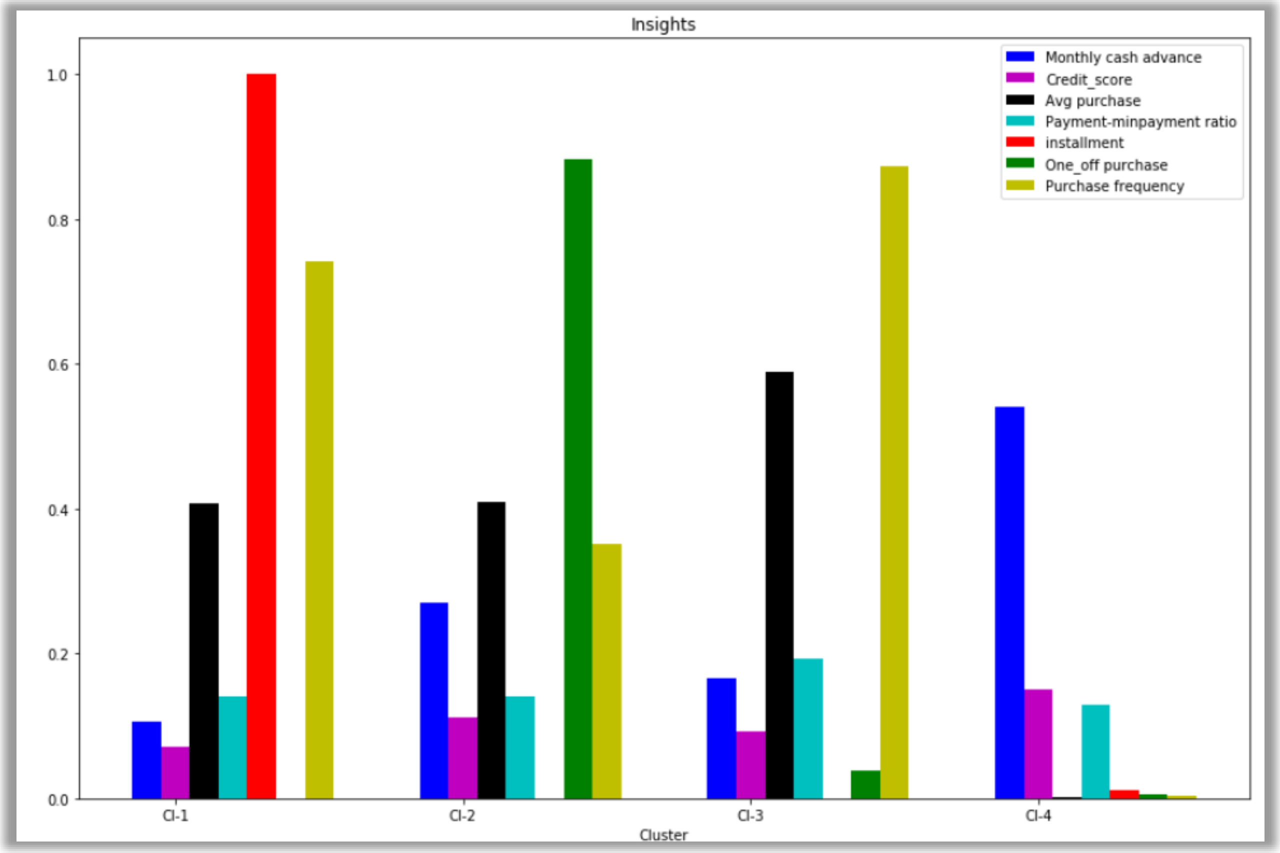


I can see that the model is able to distinguish between clusters. So now let’s count the no. of customers in each cluster and analyze them.

3.2.3 Count of customers in each cluster:

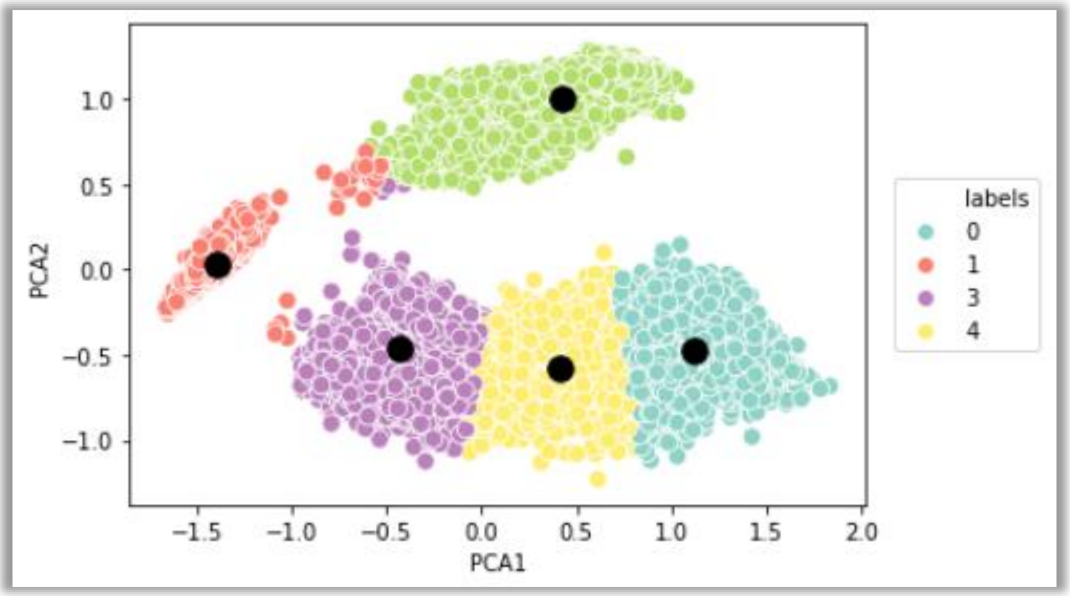


3.2.4 Visualizing the KPI's of each cluster:

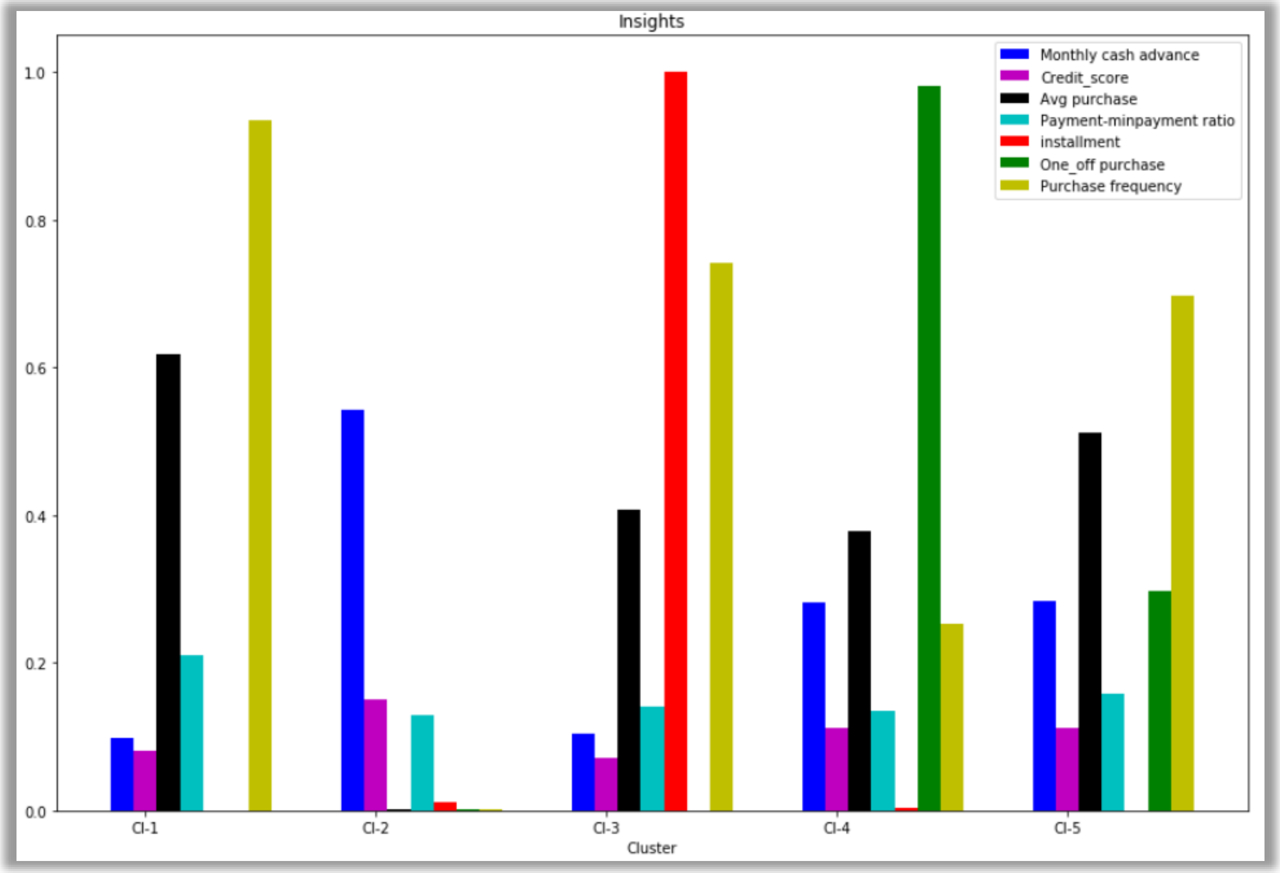


Clearly the clusters are showing distinguishable behaviors among them.

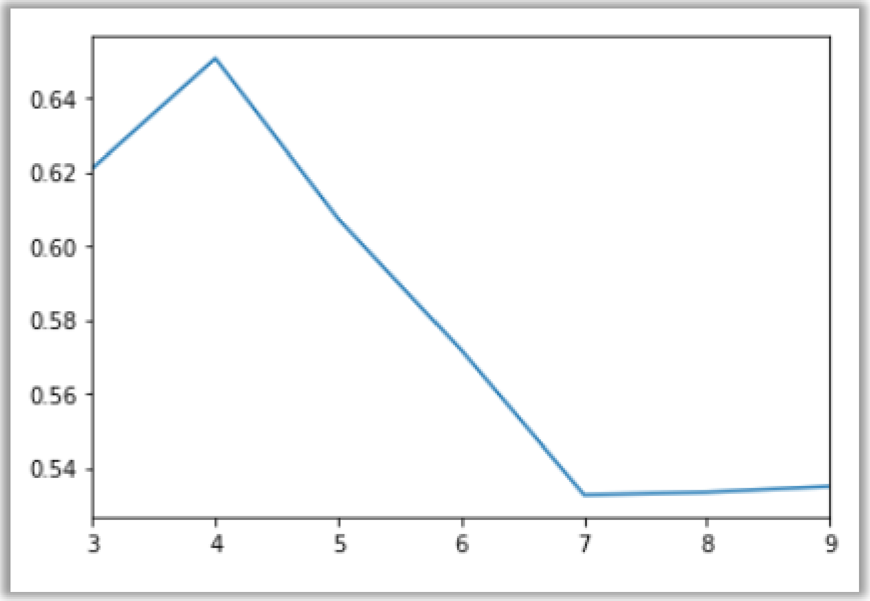
3.3 Model 2 (with 5 clusters):



3.3.1 Visualizing the KPI's of clusters:



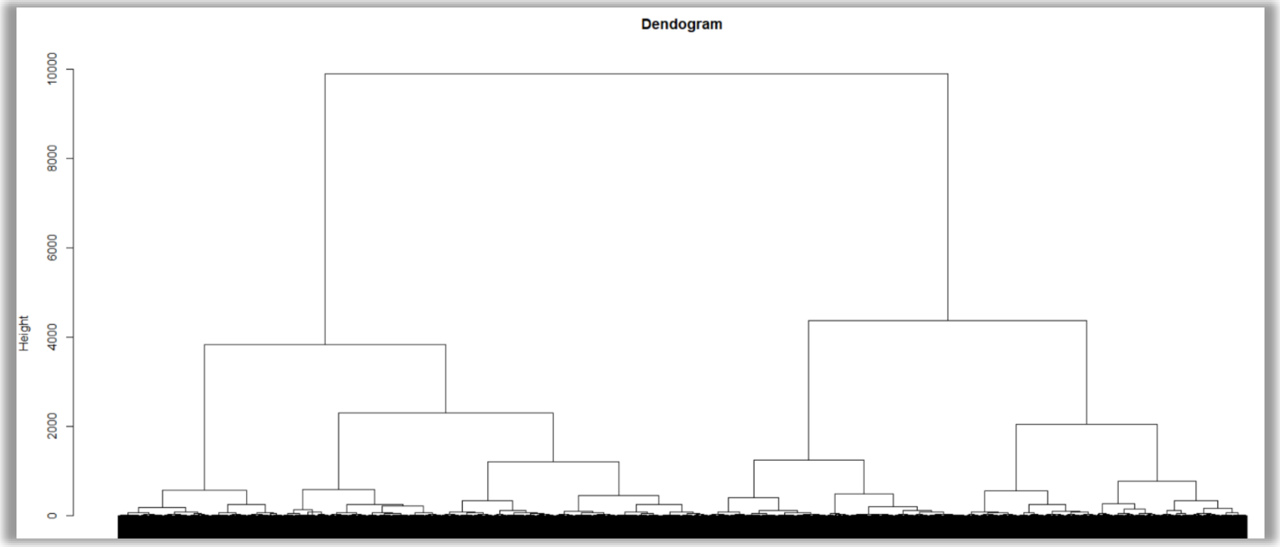
3.3.2 Evaluating K-Means Model



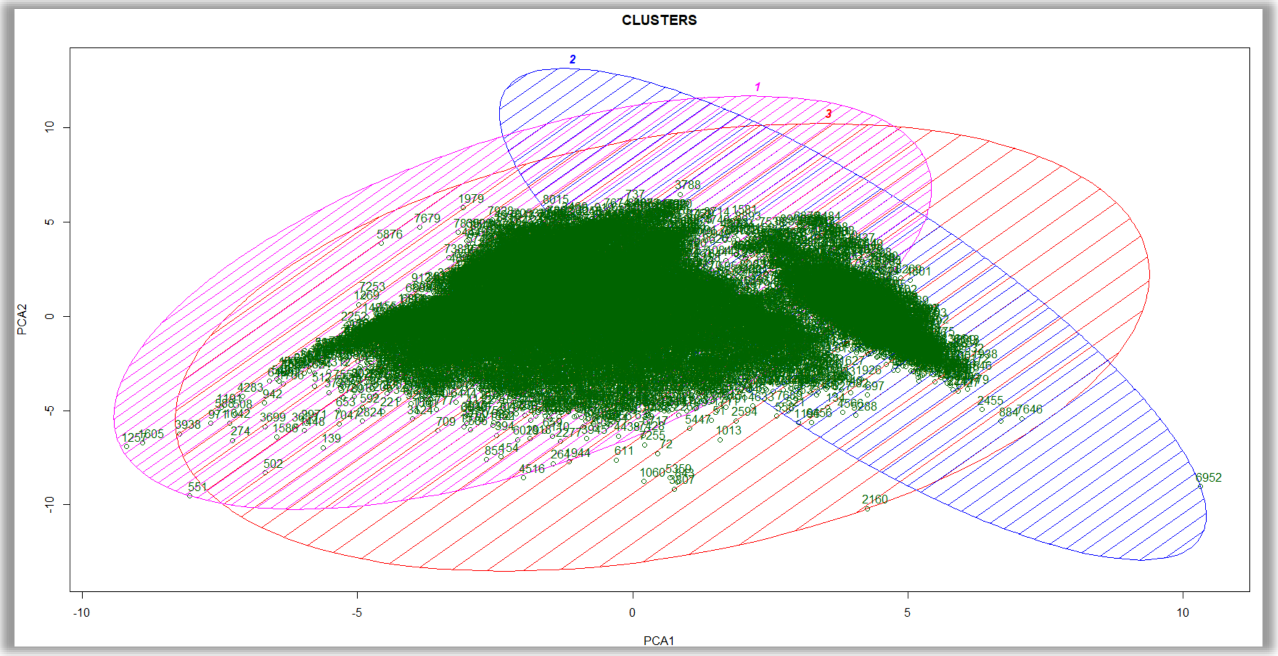
The model is performing well with 4 clusters. So, I will discard the model 2 of K Means.

3.4 Hierarchical Clustering

3.4.1 Dendrogram:



3.4.2 Clusters:

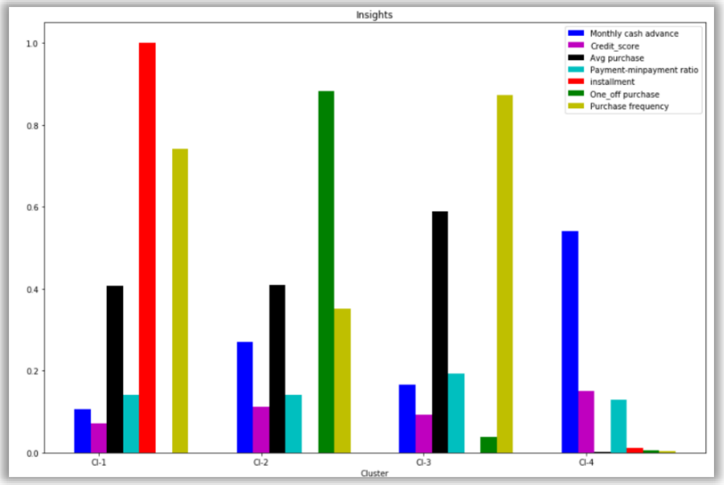


Since the clusters are overlapping a lot, this is not a good model for this data set.

When compared to K-Means, k-means separated the clusters quite well.

3.5 Marketing Strategies:

Lets now again analyze the clusters formed by k-means model 1.



Cluster 1:

- Customers are having very high installment purchases
- Extremely low One-off purchases
- Purchase frequency is also good

Cluster 2:

- Extremely low installment purchases
- Very high one-off purchases
- High Cash Advance

Cluster 3:

- Average purchase amount is highest
- Very low installment purchases
- Low one-off purchases
- Purchase frequency is also very high

Cluster 4:

- One-off and installment purchases, both are very low.
- Very high cash advance
- Average purchase amount is also very low
- Credit score is highest

3.5.1 Possible Strategies for respected clusters:

Cluster 1:

This group is the best among all, as their installment purchases is very high and these customers are maintaining good credit score by paying dues on time. We can give them rewards to make them perform more purchases.

This group of customers have good monthly transactions and good monthly spend. These are the cardholders who are vested in our loyalty program. Bonus points or cash-back incentives encourage more frequent purchases and higher spend. For example, “Earn three times the points for all grocery purchases in the next 30 days,” or “Spend $3,000 or more and get 5% cash back.”

Cluster 2:

These customers are using card for only one-off transactions. This can be risky. So, we can give them offers on installment purchases as Credit card companies tend to make more profit through installment purchases.

Cluster 3:

These are the potential target customers who are paying dues and doing purchases and maintaining comparatively good credit score. Basically, these are the customers who use their credit card on a regular basis.

These are the cardholders who have high monthly spend and high monthly transactions.

So, we must retain these customers by providing premium card/loyalty or can lower down interest rate on purchases to increase transactions or give some offers.

For example, “Use your card to pay for utilities and get two times the points,” or “Get 10,000 bonus miles for all travel purchases made in the next 90 days.”

Cluster 4:

This group is having low credit score and are taking only cash transactions. We can provide them low interest to make them more purchase transactions. These may be the cardholders that have frequent, low-ticket purchases or who may be splitting spend across multiple cards. The goal is to get a higher share of wallet. For example, “Spend $3,000 or more over the next 30 days and get a $25 statement credit.”

References:

- https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
- https://medium.com/@TheDataGyan/day-8-data-transformation-skewness-normalization-and-much-more-4c144d370e55
- https://www.rdocumentation.org/packages/factoextra/versions/1.0.6/topics/fviz_nbclust
- https://thefinancialbrand.com/75905/credit-card-marketing-strategies-bank-credit-union-roi/
- https://stackoverflow.com/questions/37931327/barplot-with-multiple-columns-in-r