**PREDICTION OF CREDIT CARD SEGMENTATION**

**ARJUN PRAKASH**

**18-11-2019**

1. INTRODUCTION 3-4

* 1. Problem Statement
  2. Data Set

1. METHODOLOGY 5-7
   1. Preprocessing
      1. Detection Of Outliers
      2. Identify the outlier in dataset from IQR method
2. Deriving KPI's 8-9
   1. Monthly average purchase and Cash Advance Amount
   2. Purchased by Type (one-off, installments)
   3. Limit usage (balance to credit limit ratio)
   4. Payments to minimum payments ratio in percentage

# Insights on new KPI 9-11

# Data Standardizing 12

# Principal Components Analysis 13

# MODELING 13-15

* 1. Model Building
     1. K Means Clustering
     2. 5 Cluster
     3. 6 Clusters
     4. 7 Clusters
  2. Checking Performance for k means

1. CONCLUSION 16

8.1 Conclusion

1. APPENDIX A 17-33

Python Code

R CODE

1. **Introduction**
   1. Problem Statement

This case requires trainees 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.

* 1. Data Set

The goal is to build the models which will predict a customer segmentation to define

Marketing strategy. Below is the sample of the data set that we are using to predict the Credit Card Segmentation:

Table 1.1: Credit Card Segmentation Sample Data (Columns: 1-8)

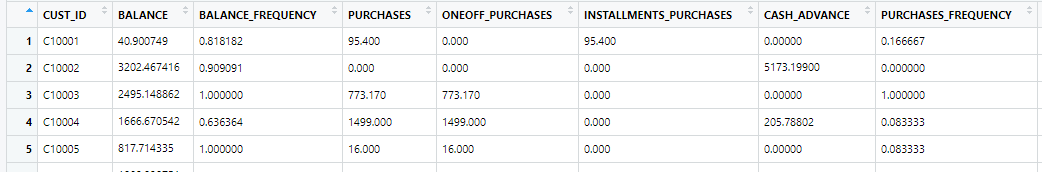


Table 1.2: Credit Card Segmentation Sample Data (Columns: 9-15)

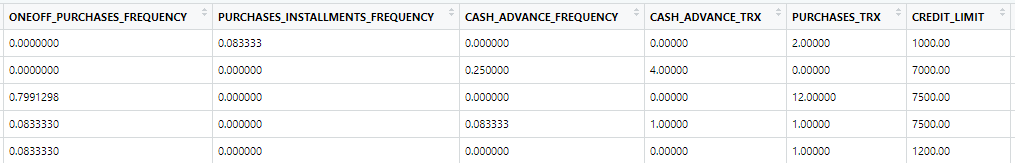
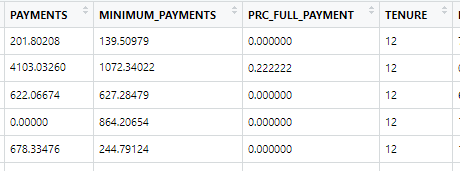


Table 1.3: Credit Card Segmentation Sample Data (Columns: 16-19)



In the below Table we have the following 13 variables, using which we have to predict the Credit Card Segmentation:

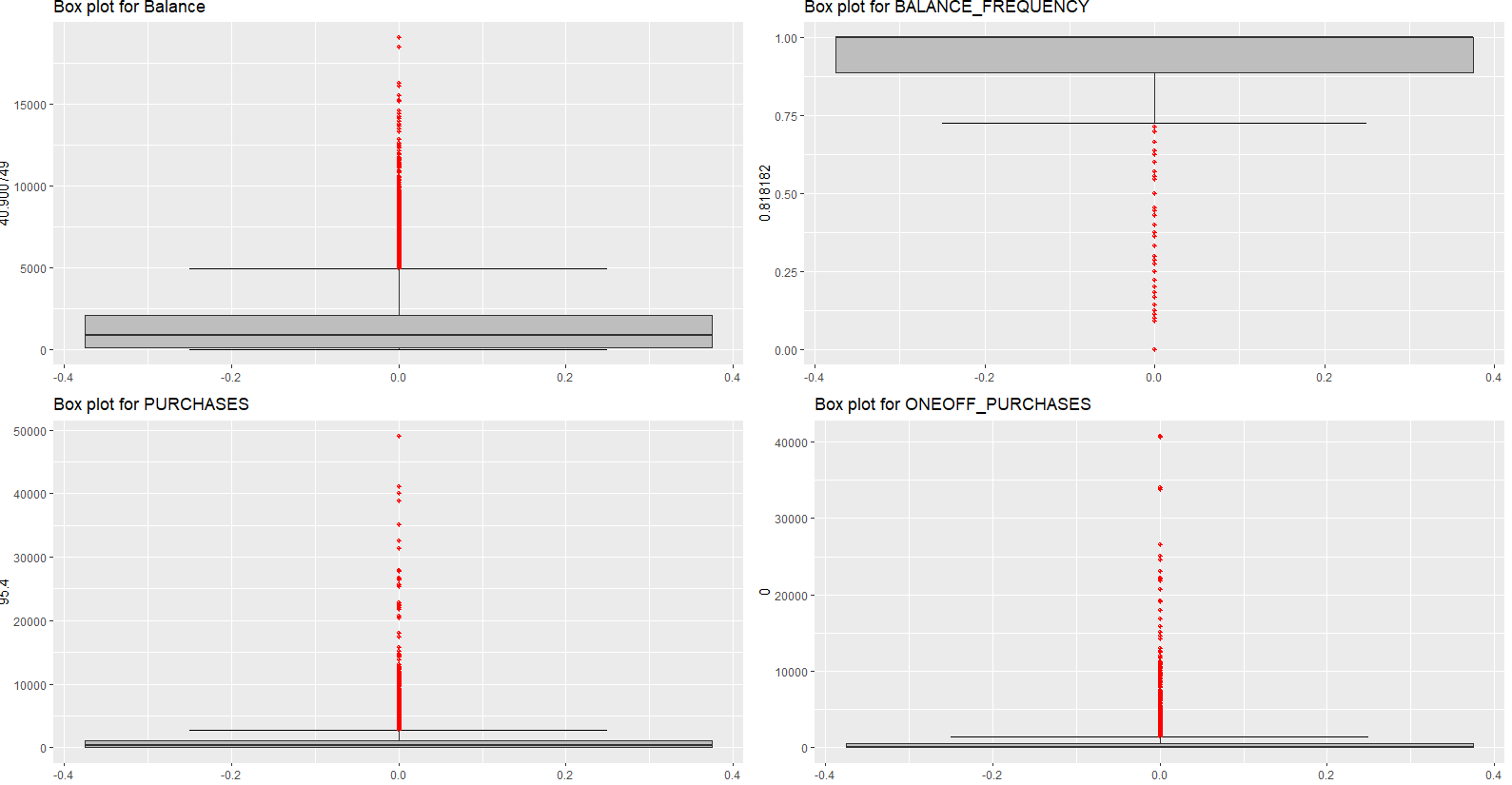
|  |  |
| --- | --- |
| **Sl.No** | **Variables** |
| 1 | CUST\_ID |
| 2 | BALANCE |
| 3 | BALANCE\_FREQUENCY |
| 4 | PURCHASES |
| 5 | ONEOFF\_PURCHASES |
| 6 | INSTALLMENTS\_PURCHASES |
| 7 | CASH\_ADVANCE |
| 8 | PURCHASES\_ FREQUENCY |
| 9 | ONEOFF\_PURCHASES\_FREQUENCY |
| 10 | PURCHASES\_INSTALLMENTS\_FREQUENCY |
| 11 | CASH\_ADVANCE\_ FREQUENCY |
| 12 | AVERAGE\_PURCHASE\_TRX |
| 13 | CASH\_ADVANCE\_TRX |
| 14 | PURCHASES\_TRX |
| 15 | CREDIT\_LIMIT |
| 16 | PAYMENTS |
| 17 | MINIMUM\_PAYMENTS |
| 18 | PRC\_FULL\_PAYMENT |
| 19 | TENURE |

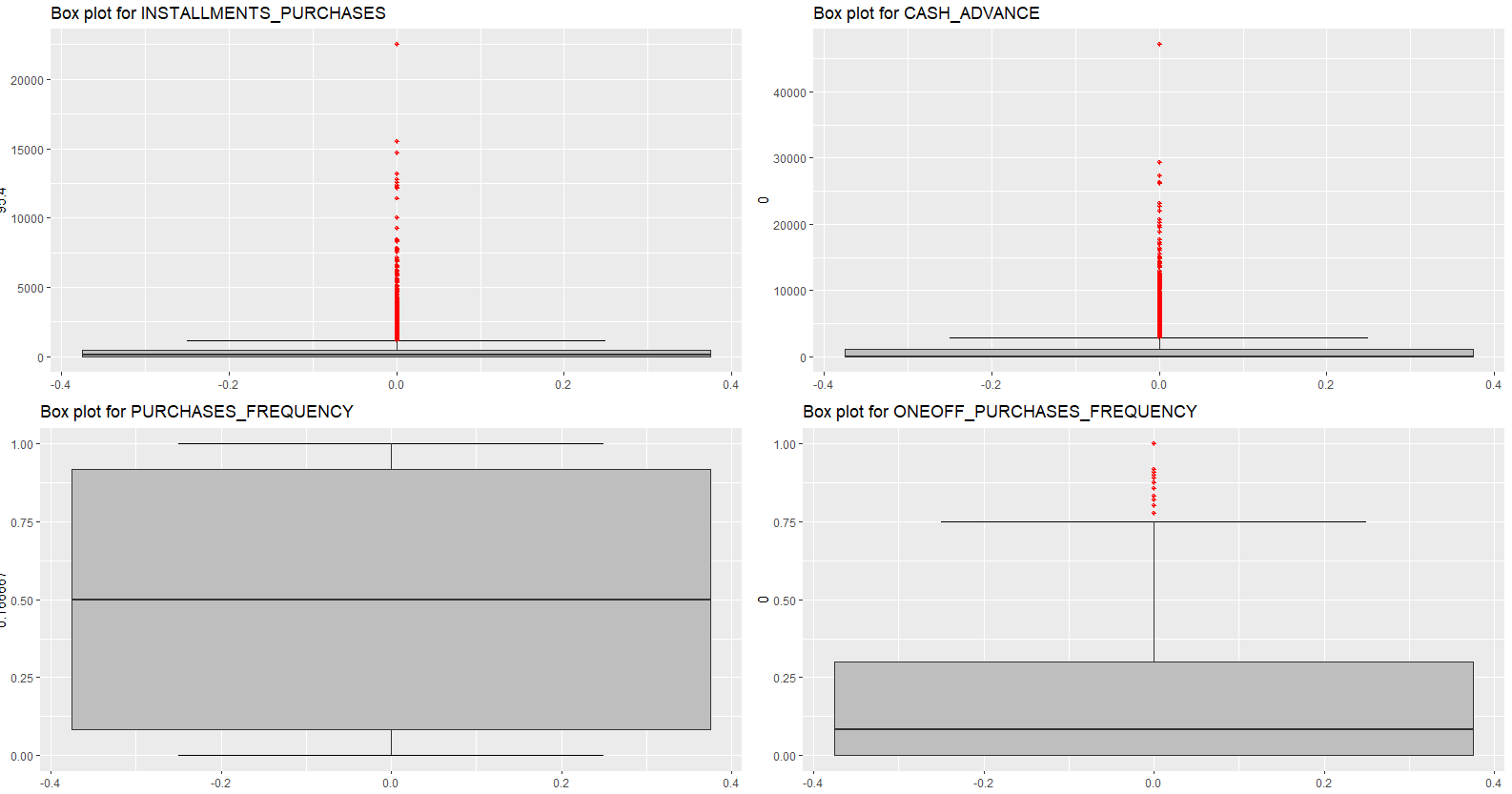
Table 1.3: Credit Card Segmentation

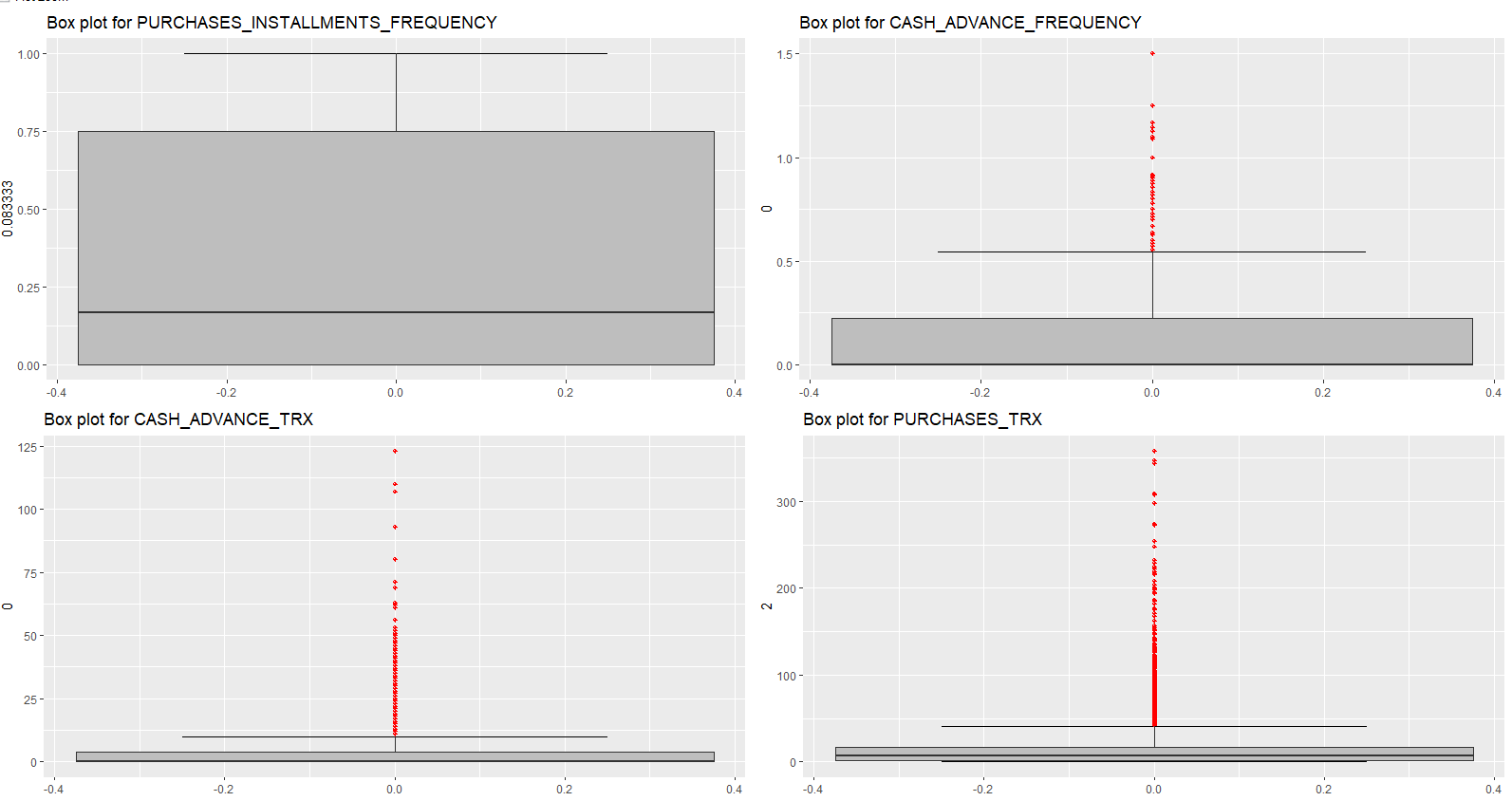
1. **Methodology**
   1. Pre-Processing

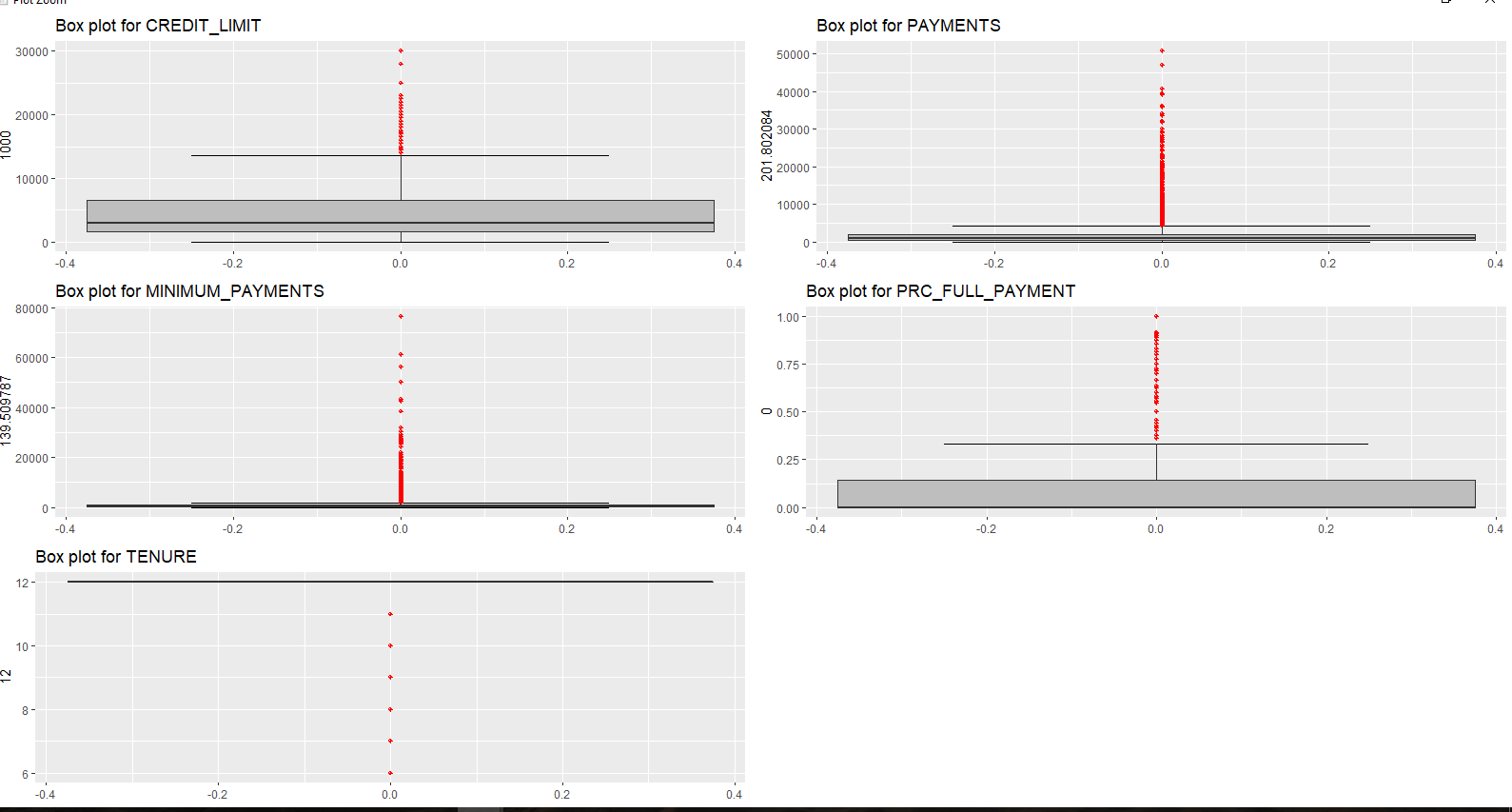
2.2.1 Detection of Outliers

Outliers:An outlier is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the datasetOutliers are detected using boxplots. Below figure illustrates the boxplots.









2.2.2 Identify the outlier in dataset from IQR method

Outliers can be identified using the IQR method, wherein the Inter Quartile Range (IQR) is calculated and the minimum and maximum values are calculated for the variables. Any value ranging outside the minimum and maximum value are discarded.

**3 Deriving KPI's**

Key Performance Indicators

A Key Performance Indicator (KPI) is a measurable value that demonstrates how effectively a company is achieving key business objectives. Organizations use KPIs to evaluate their success at reaching targets.

* 1. Monthly average purchase and Cash Advance Amount

Monthly Average Purchase

0 7.950000

1 0.000000

2 64.430833

3 124.916667

Cash Advance Amount

0 0.000000

1 536.91212

2 0.000000

3 17.149001

4 0.000000

3.2 Purchased by Type (one-off, installments)

4 types of purchase behavior has been seen

1. People not done both type of purchases (one-off, installments)

2. People done both type of purchases (one-off, installments)

3. People done only one-off payment type of purchases

4. People done only installments payment type of purchases

* 1. Limit usage (balance to credit limit ratio)

Limit usage

0 0.040901

1 0.457495

2 0.332687

3 0.222223

4 0.681429

* 1. Payments to Minimum Payments ratio in Percentage

Minimum Payments ratio

0 1.446508

1 3.826241

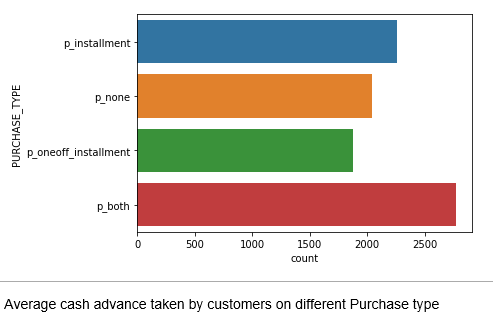
2 0.991682

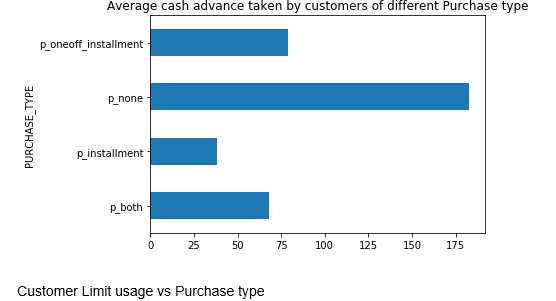
3 0.000000

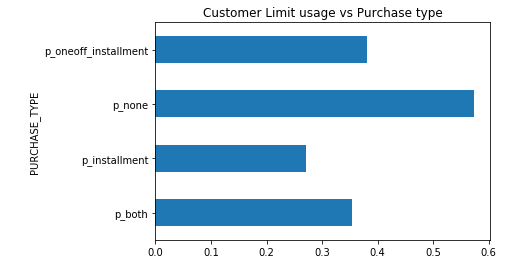
4 2.771075

**4 Insights on new KPI**

Plot graph on Purchase types







Heat Map – Show correlation between variable

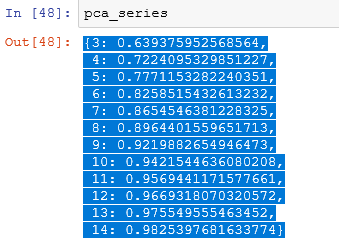


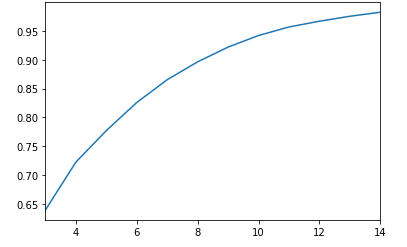
**5 Data Standardizing**

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization

**6 Performance Component Analysis (PCA)**

Principal component analysis (PCA) is a technique that is useful for the compression and classification of data. The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables, smaller than the original set of variables, that nonetheless retains most of the sample's information

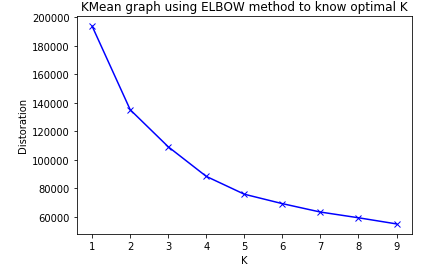




**7 Model Building**

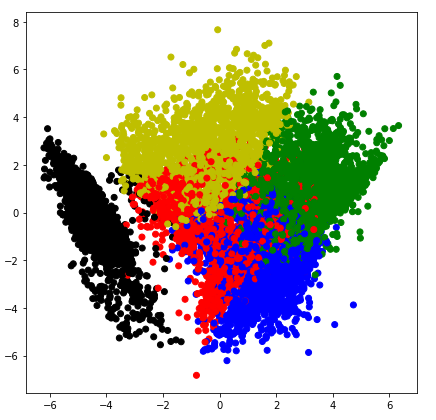
7.1.1 K Means Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. ... In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible



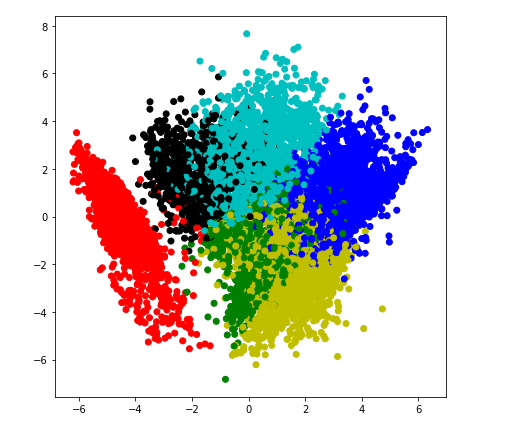
From the Elbow method we can take optimal K value as 5 since there are no drastic changes after 5.So we consider 5 Cluster initially and start building the model.

**7.1.2 5 Cluster**



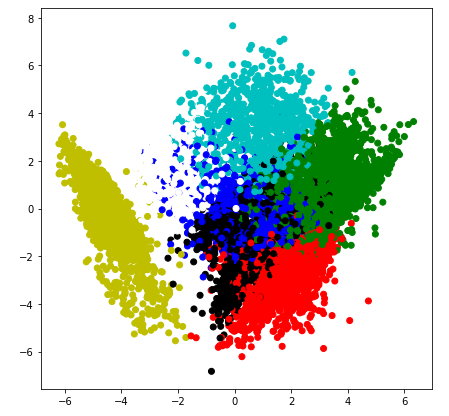
The Cluster in Black seems to be well segregated compred to other clusters.

**7.1.2 6 Cluster**



The cluster in red seems to be more segregated from other clusters.

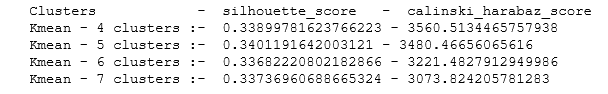
**7.1.4 7 Cluster**

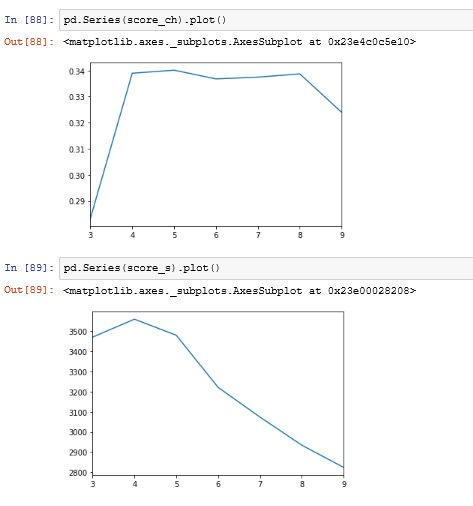


The cluster in yellow is more segregated than the others.

**7.2 Checking performance metrics for Kmeans**

We have used the calinski harabaz , silhouette methods used to validate the performance of K means Clusters and below are the scores





8 CONCLUSION

Marketing Strategy Suggested:

1. Group with both type of purchases

* They are potential target customers who are paying dues and doing purchases and maintaining comparatively good credit score ) -- we can increase credit limit or can lower down interest rate -- Can be given premium card /loyality cards to increase transactions

2 Group not done both type of purchases

* They have poor credit score and taking only cash on advance. We can target them by providing less interest rate on purchase transaction

1. Group done both type of purchases

* This group is has minimum paying ratio and using card for just oneoff transactions (may be for utility bills only). This group seems to be risky group.

4. Group done only installments payment type of purchases

* This group is performing best among all as cutomers are maintaining good credit score and paying dues on time. -- Giving rewards point will make them perform more purchases.

**9 Appendix A**

**PYTHON CODE:-**



**R CODE:-**

#Clear the environment

rm(list = ls())

rm(library)

#Load the Libraries

library(DataCombine)

library(ggplot2)

library(gridExtra)

library(dplyr)

library(magrittr)

library(corrplot)

library(corrgram)

library(psych)

library(GPArotation)

library(paran)

library(NbClust)

#Set Working directory

setwd("C:/Users/rnp/Desktop/Arjun/Data Science/Project3")

#Load the Credit card dataset

credit\_df<- read.csv('credit-card-data.csv',header = T)

head(credit\_df)

summary(credit\_df)

View(credit\_df)

dim(credit\_df)

#Distribution of data attributes

hist\_plot1 = ggplot(credit\_df, aes(BALANCE))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of Balance")+theme(text = element\_text(size = 10))

hist\_plot2 = ggplot(credit\_df, aes(BALANCE\_FREQUENCY))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of BALANCE\_FREQUENCY")+theme(text = element\_text(size = 10))

hist\_plot3 = ggplot(credit\_df, aes(PURCHASES))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of PURCHASES")+theme(text = element\_text(size = 10))

hist\_plot4 = ggplot(credit\_df, aes(ONEOFF\_PURCHASES))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of ONEOFF\_PURCHASES")+theme(text = element\_text(size = 10))

hist\_plot5 = ggplot(credit\_df, aes(INSTALLMENTS\_PURCHASES))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of INSTALLMENTS\_PURCHASES")+theme(text = element\_text(size = 10))

hist\_plot6 = ggplot(credit\_df, aes(CASH\_ADVANCE))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of CASH\_ADVANCE")+theme(text = element\_text(size = 10))

hist\_plot7 = ggplot(credit\_df, aes(PURCHASES\_FREQUENCY))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of PURCHASES\_FREQUENCY")+theme(text = element\_text(size = 10))

hist\_plot8 = ggplot(credit\_df, aes(ONEOFF\_PURCHASES\_FREQUENCY))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of ONEOFF\_PURCHASES\_FREQUENCY")+theme(text = element\_text(size = 10))

hist\_plot9 = ggplot(credit\_df, aes(PURCHASES\_INSTALLMENTS\_FREQUENCY))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of PURCHASES\_INSTALLMENTS\_FREQUENCY")+theme(text = element\_text(size = 10))

hist\_plot10 = ggplot(credit\_df, aes(CASH\_ADVANCE\_FREQUENCY))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of CASH\_ADVANCE\_FREQUENCY")+theme(text = element\_text(size = 10))

hist\_plot11 = ggplot(credit\_df, aes(CASH\_ADVANCE\_TRX))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of CASH\_ADVANCE\_TRX")+theme(text = element\_text(size = 10))

hist\_plot12 = ggplot(credit\_df, aes(PURCHASES\_TRX))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of PURCHASES\_TRX")+theme(text = element\_text(size = 10))

hist\_plot13 = ggplot(credit\_df, aes(CREDIT\_LIMIT))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of CREDIT\_LIMIT")+theme(text = element\_text(size = 10))

hist\_plot14 = ggplot(credit\_df, aes(MINIMUM\_PAYMENTS))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of MINIMUM\_PAYMENTS")+theme(text = element\_text(size = 10))

hist\_plot15 = ggplot(credit\_df, aes(PRC\_FULL\_PAYMENT))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of PRC\_FULL\_PAYMENT")+theme(text = element\_text(size = 10))

hist\_plot16 = ggplot(credit\_df, aes(TENURE))+theme\_bw()+geom\_histogram(fill='blue', bins = 20)+ggtitle("Distribution of TENURE")+theme(text = element\_text(size = 10))

gridExtra::grid.arrange(hist\_plot1,hist\_plot2,hist\_plot3,hist\_plot4,hist\_plot5,hist\_plot6,hist\_plot7,hist\_plot8,hist\_plot9,

hist\_plot10,hist\_plot11,hist\_plot12,hist\_plot13,hist\_plot14,hist\_plot15,hist\_plot16,nrow=4,ncol=4)

# Find Missing values

missing\_value<-data.frame(missing\_value=apply(credit\_df,2,function(x){sum(is.na(x))}))

missing\_value

#impute missing values with mean

credit\_df$CREDIT\_LIMIT[is.na(credit\_df$CREDIT\_LIMIT)] <- mean(credit\_df$CREDIT\_LIMIT, na.rm = TRUE)

credit\_df$MINIMUM\_PAYMENTS[is.na(credit\_df$MINIMUM\_PAYMENTS)] <- mean(credit\_df$MINIMUM\_PAYMENTS, na.rm = TRUE)

# Check for Missing values after imputation

missing\_value<-data.frame(missing\_value=apply(credit\_df,2,function(x){sum(is.na(x))}))

missing\_value

#\*\*\*\*\*\*\*\* Outlier detection \*\*\*\*\*\*\*\*#

box\_plot1 = ggplot(aes\_string(y = credit\_df$BALANCE), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$BALANCE)+ggtitle(paste("Box plot for Balance"))

box\_plot2 = ggplot(aes\_string(y = credit\_df$BALANCE\_FREQUENCY), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$BALANCE\_FREQUENCY)+ggtitle(paste("Box plot for BALANCE\_FREQUENCY"))

box\_plot3 = ggplot(aes\_string(y = credit\_df$PURCHASES), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$PURCHASES)+ggtitle(paste("Box plot for PURCHASES"))

box\_plot4 = ggplot(aes\_string(y = credit\_df$ONEOFF\_PURCHASES), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$ONEOFF\_PURCHASES)+ggtitle(paste("Box plot for ONEOFF\_PURCHASES"))

box\_plot5 = ggplot(aes\_string(y = credit\_df$INSTALLMENTS\_PURCHASES), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$INSTALLMENTS\_PURCHASES)+ggtitle(paste("Box plot for INSTALLMENTS\_PURCHASES"))

box\_plot6 = ggplot(aes\_string(y = credit\_df$CASH\_ADVANCE), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$CASH\_ADVANCE)+ggtitle(paste("Box plot for CASH\_ADVANCE"))

box\_plot7 = ggplot(aes\_string(y = credit\_df$PURCHASES\_FREQUENCY), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$PURCHASES\_FREQUENCY)+ggtitle(paste("Box plot for PURCHASES\_FREQUENCY"))

box\_plot8 = ggplot(aes\_string(y = credit\_df$ONEOFF\_PURCHASES\_FREQUENCY), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$ONEOFF\_PURCHASES\_FREQUENCY)+ggtitle(paste("Box plot for ONEOFF\_PURCHASES\_FREQUENCY"))

box\_plot9 = ggplot(aes\_string(y = credit\_df$PURCHASES\_INSTALLMENTS\_FREQUENCY), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$PURCHASES\_INSTALLMENTS\_FREQUENCY)+ggtitle(paste("Box plot for PURCHASES\_INSTALLMENTS\_FREQUENCY"))

box\_plot10 = ggplot(aes\_string(y = credit\_df$CASH\_ADVANCE\_FREQUENCY), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$CASH\_ADVANCE\_FREQUENCY)+ggtitle(paste("Box plot for CASH\_ADVANCE\_FREQUENCY"))

box\_plot11 = ggplot(aes\_string(y = credit\_df$CASH\_ADVANCE\_TRX), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$CASH\_ADVANCE\_TRX)+ggtitle(paste("Box plot for CASH\_ADVANCE\_TRX"))

box\_plot12 = ggplot(aes\_string(y = credit\_df$PURCHASES\_TRX), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$PURCHASES\_TRX)+ggtitle(paste("Box plot for PURCHASES\_TRX"))

box\_plot13 = ggplot(aes\_string(y = credit\_df$CREDIT\_LIMIT), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$CREDIT\_LIMIT)+ggtitle(paste("Box plot for CREDIT\_LIMIT"))

box\_plot14 = ggplot(aes\_string(y = credit\_df$PAYMENTS), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$PAYMENTS)+ggtitle(paste("Box plot for PAYMENTS"))

box\_plot15 = ggplot(aes\_string(y = credit\_df$MINIMUM\_PAYMENTS), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$MINIMUM\_PAYMENTS)+ggtitle(paste("Box plot for MINIMUM\_PAYMENTS"))

box\_plot16 = ggplot(aes\_string(y = credit\_df$PRC\_FULL\_PAYMENT), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$PRC\_FULL\_PAYMENT)+ggtitle(paste("Box plot for PRC\_FULL\_PAYMENT"))

box\_plot17 = ggplot(aes\_string(y = credit\_df$TENURE), data = credit\_df)+stat\_boxplot(geom = "errorbar", width = 0.5) +

geom\_boxplot(outlier.colour="red", fill = "grey" ,outlier.shape=10,outlier.size=1, notch=FALSE) + theme(legend.position="bottom")+labs(y=credit\_df$TENURE)+ggtitle(paste("Box plot for TENURE"))

gridExtra::grid.arrange(box\_plot1,box\_plot2,box\_plot3,box\_plot4,nrow=2,ncol=2)

gridExtra::grid.arrange(box\_plot5,box\_plot6,box\_plot7,box\_plot8,nrow=2,ncol=2)

gridExtra::grid.arrange(box\_plot9,box\_plot10,box\_plot11,box\_plot12,nrow=2,ncol=2)

gridExtra::grid.arrange(box\_plot13,box\_plot14,box\_plot15,box\_plot16,box\_plot17,nrow=3,ncol=2)

#Advance KPI's

#1.monthly average purchase and cash advance amount

credit\_df$MONTHLY\_AVG\_PURCHASE <- credit\_df$PURCHASES/credit\_df$TENURE

credit\_df$CASH\_ADV\_AMOUNT <- credit\_df$CASH\_ADVANCE/credit\_df$TENURE

#2.Purchased by Type (one-off, instalments)

### 4 types of purchase behaviour has been seen

###### 1. People not done both type of purchases (one-off, instalments)

###### 2. People done both type of purchases (one-off, instalments)

###### 3. People done only one-off payment type of purchases

###### 4. People done only installments payment type of purchases

purchase <- function(one\_purchase,instal\_purchase){

if ((one\_purchase == 0) && (instal\_purchase == 0)) {

return ('p\_none')

} else if ((one\_purchase > 0) && (instal\_purchase > 0)) {

return ('p\_both')

} else if ((one\_purchase > 0) && (instal\_purchase == 0)) {

return ('p\_oneoff\_installment')

} else if ((one\_purchase == 0) && (instal\_purchase > 0)) {

return ('p\_installment')

}

}

one\_purchase = credit\_df$ONEOFF\_PURCHASES

instal\_purchase = credit\_df$INSTALLMENTS\_PURCHASES

credit\_df$PURCHASE\_TYPE <- mapply(purchase, one\_purchase,instal\_purchase)

credit\_df$PURCHASE\_TYPE

table(credit\_df$PURCHASE\_TYPE)

#3.Limit usage (balance to credit limit ratio in percentage)

credit\_df$LIMIT\_USAGE= (credit\_df$BALANCE/credit\_df$CREDIT\_LIMIT)\*100

credit\_df$LIMIT\_USAGE

#4.payments to minimum payments ratio in percentage

credit\_df$MIN\_PAYMENT=(credit\_df$PAYMENTS/credit\_df$MINIMUM\_PAYMENTS)\*100

credit\_df$MIN\_PAYMENT

#5 Minimum payment ratio

credit\_df$MIN\_PAYMENTS\_RATIO <- credit\_df$PAYMENTS/credit\_df$MINIMUM\_PAYMENTS

credit\_df$MIN\_PAYMENTS\_RATIO

####Insights on new KPI

#Plot graph on Purchase types for counts

bar\_plot = ggplot(credit\_df, aes(PURCHASE\_TYPE))+theme\_bw()+geom\_bar(fill='blue')+ggtitle("Total counts by Purchase Type")+theme(text = element\_text(size = 10))

plot(bar\_plot)

#Average cash advance taken by customers on different Purchase type

credit\_df %>% group\_by(PURCHASE\_TYPE) %>% summarise(mean\_cash\_adv=mean(CASH\_ADV\_AMOUNT))

# Output

#### A tibble: 4 x 2

#### PURCHASE\_TYPE mean\_cash\_adv

#### <chr> <dbl>

#### 1 p\_both 67.8

#### 2 p\_installment 38.4

#### 3 p\_none 183.

#### 4 p\_oneoff\_installment 79.0

#Plot a graph

credit\_df %>%

group\_by(PURCHASE\_TYPE) %>%

summarise(mean\_cash\_adv=mean(CASH\_ADV\_AMOUNT)) %>%

ggplot(aes(x = PURCHASE\_TYPE, y = mean\_cash\_adv, fill = PURCHASE\_TYPE)) + geom\_bar(stat = "identity") + theme\_classic() +

labs(x = "Purchase Type", y = "Mean cash advance amount", title = paste("Mean cash advance amount summarise by Purchase Type"))

#Customer average Limit usage vs Purchase type

credit\_df %>% group\_by(PURCHASE\_TYPE) %>% summarise(mean\_cash\_adv=mean(LIMIT\_USAGE))

# Output

#### A tibble: 4 x 2

#### PURCHASE\_TYPE mean\_cash\_adv

#### <chr> <dbl>

#### 1 p\_both 35.4

#### 2 p\_installment 27.2

#### 3 p\_none 57.4

#### 4 p\_oneoff\_installment 38.1

#Plot a graph

credit\_df %>%

group\_by(PURCHASE\_TYPE) %>%

summarise(mean\_limit\_usage=mean(LIMIT\_USAGE)) %>%

ggplot(aes(x = PURCHASE\_TYPE, y = mean\_limit\_usage, fill = PURCHASE\_TYPE)) + geom\_bar(stat = "identity") + theme\_classic() +

labs(x = "Purchase Type", y = "Mean Limit usage", title = paste("Mean Limit usage summarise by Purchase Type"))

# Identifying Outliers

mystats <- function(x) {

missing\_val<-sum(is.na(x))

val <- x[!is.na(x)]

mn <- mean(val)

len <- length(val)

std <- sd(val)

min <- min(val)

p1<-quantile(val,0.01)

q1<-quantile(val,0.25)

q2<-quantile(val,0.5)

q3<-quantile(val,0.75)

p90<-quantile(val,0.90)

p95<-quantile(val,0.95)

p99<-quantile(val,0.99)

max <- max(val)

UC <- mn+2\*std

LC <- mn-2\*std

outlier\_flag<- max>UC | min<LC

return(c(n=len, missing\_val=missing\_val, outlier\_flag=outlier\_flag, mean=mn, stdev=std, min = min, p1=p1,q1=q1,q2=q2,q3=q3,p90=p90,p95=p95,p99=p99,max=max, UC=UC, LC=LC ))

}

colnames(credit\_df)

DS\_Vars <- c(

"BALANCE",

"BALANCE\_FREQUENCY",

"PURCHASES",

"ONEOFF\_PURCHASES",

"INSTALLMENTS\_PURCHASES",

"CASH\_ADVANCE",

"PURCHASES\_FREQUENCY",

"ONEOFF\_PURCHASES\_FREQUENCY",

"PURCHASES\_INSTALLMENTS\_FREQUENCY",

"CASH\_ADVANCE\_FREQUENCY",

"CASH\_ADVANCE\_TRX",

"PURCHASES\_TRX",

"CREDIT\_LIMIT",

"PAYMENTS",

"MINIMUM\_PAYMENTS",

"PRC\_FULL\_PAYMENT",

"TENURE",

"MONTHLY\_AVG\_PURCHASE",

"CASH\_ADV\_AMOUNT",

"LIMIT\_USAGE",

"MIN\_PAYMENT",

"MIN\_PAYMENTS\_RATIO"

)

Outliers<-t(data.frame(apply(credit\_df[DS\_Vars], 2, mystats)))

View(Outliers)

# Outlier Treatment

credit\_df$BALANCE[credit\_df$BALANCE>5727.539]<-5727.539

credit\_df$BALANCE\_FREQUENCY[credit\_df$BALANCE\_FREQUENCY>1.351079]<-1.351079

credit\_df$PURCHASES[credit\_df$PURCHASES>5276.474]<-5276.474

credit\_df$ONEOFF\_PURCHASES[credit\_df$ONEOFF\_PURCHASES>3912.213]<-3912.213

credit\_df$INSTALLMENTS\_PURCHASES[credit\_df$INSTALLMENTS\_PURCHASES>2219.744]<-2219.744

credit\_df$CASH\_ADVANCE[credit\_df$CASH\_ADVANCE>5173.199]<-5173.199

credit\_df$PURCHASES\_FREQUENCY[credit\_df$PURCHASES\_FREQUENCY>1.293092]<-1.293092

credit\_df$ONEOFF\_PURCHASES\_FREQUENCY[credit\_df$ONEOFF\_PURCHASES\_FREQUENCY>0.7991298]<-0.7991298

credit\_df$PURCHASES\_INSTALLMENTS\_FREQUENCY[credit\_df$PURCHASES\_INSTALLMENTS\_FREQUENCY>1.159333]<-1.159333

credit\_df$CASH\_ADVANCE\_FREQUENCY[credit\_df$CASH\_ADVANCE\_FREQUENCY>0.535387]<-0.535387

credit\_df$CASH\_ADVANCE\_TRX[credit\_df$CASH\_ADVANCE\_TRX>16.89812]<-16.89812

credit\_df$PURCHASES\_TRX[credit\_df$PURCHASES\_TRX>64.42513]<-64.42513

credit\_df$CREDIT\_LIMIT[credit\_df$CREDIT\_LIMIT>11771.67]<-11771.67

credit\_df$PAYMENTS[credit\_df$PAYMENTS>7523.271]<-7523.271

credit\_df$MINIMUM\_PAYMENTS[credit\_df$MINIMUM\_PAYMENTS>5525.383]<-5525.383

credit\_df$PRC\_FULL\_PAYMENT[credit\_df$PRC\_FULL\_PAYMENT>0.738713]<-0.738713

credit\_df$TENURE[credit\_df$TENURE>14.19398]<-14.19398

credit\_df$MONTHLY\_AVG\_PURCHASE[credit\_df$MONTHLY\_AVG\_PURCHASE>447.1927] <- 447.1927

credit\_df$CASH\_ADV\_AMOUNT[credit\_df$CASH\_ADV\_AMOUNT>475.2502] <- 475.2502

credit\_df$LIMIT\_USAGE[credit\_df$LIMIT\_USAGE>116.8327] <- 116.8327

credit\_df$MIN\_PAYMENT[credit\_df$MIN\_PAYMENT>24538.99] <- 24538.99

credit\_df$MIN\_PAYMENTS\_RATIO[credit\_df$MIN\_PAYMENTS\_RATIO>245.3899] <- 245.3899

#Check for collinearity using correlation graph

corrgram(credit\_df, order = F, upper.panel=panel.pie, text.panel=panel.txt, main = "Correlation Plot")

#computes the correlation coefficient.

Col\_nums <- credit\_df[DS\_Vars]

comp\_corr<- cor(Col\_nums)

View(comp\_corr)

#Parallel Analysis Scree Plot

parallel <- fa.parallel(comp\_corr, fm = 'minres', fa = 'fa')

scree(comp\_corr, factors=T, pc=T, main="scree plot", hline=NULL, add=FALSE)

#Factor Analysis

FA\_5 <- fa(r=comp\_corr, 5, rotate = "Varimax" , fm="minres")

#View the values

FA\_SORT<-fa.sort(FA\_5)

FA\_SORT$loadings

#Standardizing the data

credit\_scaled <- credit\_df[DS\_Vars]

credit\_scaled <- data.frame(scale(credit\_scaled))

head(credit\_scaled)

#Extract number of cluster to build

#NBclsut\_ext <- NbClust(comp\_corr, min.nc = 5, max.nc = 8, method = "kmeans")

#Bulid the model with K-mean clustering

#With no. of cluster 4,5,6,7

Kmean\_4 <- kmeans(credit\_scaled, 4, nstart = 20)

Kmean\_5 <- kmeans(credit\_scaled, 5, nstart = 20)

Kmean\_6 <- kmeans(credit\_scaled, 6, nstart = 20)

Kmean\_7 <- kmeans(credit\_scaled, 7, nstart = 20)

#summarise the cluster data

Kmean\_4

Kmean\_5

Kmean\_6

Kmean\_7

\