Credit Card EDA PCA and Models

```
library(dbscan)
library(fpc)
library(factoextra)
library(rattle.data)
library(plyr)
library(dplyr)
library(sqldf)
library(lubridate)
library(data.table)
library(ggplot2)
library(anytime)
library(reshape2)
library(tidyr)
library(evaluate)
library(shape)
library(cluster)


memory.limit()
memory.limit(size=90000)

library(readr)
cc_data <- read_csv("C:/Users/daveh/Desktop/CC_Data.csv",
    col_types = cols(`Agency Number` = col_number(),
        `Posted Date` = col_date(format = "%m/%d/%Y"),
        `Transaction Date` = col_date(format = "%m/%d/%Y")))

colnames(cc_data)[11] <- "Merchant.Category"

cc_data <- as.data.table(cc_data)

View(cc_data)

colnames(cc_data)
```

Dates Manipulation

```
Mo_Date <- anydate(c(cc_data$`Year-Month`))
cc_months <- as.numeric(month(Mo_Date))
cc_days <- as.numeric(days(Mo_Date))
cc_weeks <- as.numeric(weeks(Mo_Date))
```

Monthly purchase frequency by merchant category or number of same merchant monthly transactions

```
ff1 <-setDT(cc_data)[, .("Transaction.Frequency" = .N),
                     by = .(Merchant.Category, Agency.Name, cc_months,
Transaction.Date)][order(Merchant.Category, decreasing = FALSE)]
ff1



join(cc_data, ff1,type = "inner")%>%
```

```
  select(-Agency.Number, -Cardholder.First.Initial, -Cardholder.Last.Name, -
Description, -Amount, -Vendor, -Posted.Date, -Transaction.Date)
```

Feature 1: Monthly purchase frequency ratio by merchant category (number of same merchant monthly transactions)

```
f1 <- ff1%>%group_by(Merchant.Category, Agency.Name) %>%
  mutate(Monthly.Purchase.Frequency.Ratio = Transaction.Frequency
/mean(Transaction.Frequency))%>%
  select(-Transaction.Frequency)
f1
```

Feature 2: Daily transacion ratio by merchanct category

```
f2 <- ff1%>% mutate(Daily.Purchase.Frequency.Ratio
=Transaction.Frequency/sum(Transaction.Frequency)/365,
Daily.Purchase.Frequency = Transaction.Frequency/365)%>%
  select(-cc_months, -Transaction.Frequency,-Daily.Purchase.Frequency)

f2
```

Feature 3: Weekly transaction frequency ratio by merchant category

```
f3 <-setDT(cc_data)[, .(Weekly.Transaction.Frequency.Ratio = .N/52),
                    by = .( Merchant.Category, Agency.Name,
Transaction.Date)][order(Agency.Name, decreasing = FALSE)]
```

Mean, max, min transaction frequency by agency name and merchant category

```
ff4<- cc_data %>% group_by(Agency.Name, Merchant.Category, `Year-Month`) %>%
       summarise(Mean.Trans.Amount = mean(Amount), Max.Trans.Amount =
max(Amount),
       Min.Trans.Amount = min(Amount), Trans.Freq =n())

library(plyr)
join(cc_data, ff4,type = "inner")%>%
  select(-Agency.Number, -Cardholder.First.Initial, -Cardholder.Last.Name, -
Description, -Amount, -Vendor, -Posted.Date, -Transaction.Date)
```

Feature 4: Max ratio by agency name

```
f4 <- ff4%>%group_by(Agency.Name, Merchant.Category) %>%
  mutate(Max.Amount.Ratio = Mean.Trans.Amount/Max.Trans.Amount)%>%
  select(-Trans.Freq, -Mean.Trans.Amount, -Max.Trans.Amount, -
Min.Trans.Amount)
```

Feature 4 dataframe rehaped to wide columns and each row reps an agency name for each a merchant category

```
f4.dcast <- dcast(f4,  Agency.Name + Merchant.Category ~`Year-Month`,
value.var = 'Max.Amount.Ratio')
f4.dcast[is.na(f4.dcast)]<- "0"
f4.dcast<- as.matrix(f4.dcast)
f4.dcast <- as.data.frame(f4.dcast)
```

Feature 6: Annual mean transaction count ratio by agency

```
f6 <- ff4%>%group_by(Agency.Name, Merchant.Category) %>%
  mutate(Mean.Trans.Count.Ratio = Trans.Freq /mean(Trans.Freq)) %>%
```

```
  select( -Mean.Trans.Amount, -Max.Trans.Amount, -Min.Trans.Amount, -
Trans.Freq)
```

Feature 7: Min ratio by agency name

```
f7 <- ff4%>%group_by(Agency.Name, Merchant.Category) %>%
  mutate(Min.Amount.Ratio = Mean.Trans.Amount/Min.Trans.Amount) %>%
  select(-Trans.Freq, -Mean.Trans.Amount, -Max.Trans.Amount, -
Min.Trans.Amount)
```

PCA analysis: Dimension reduction

```
dff <- cbind(cc_data[,11],f1[,5],Daily.Purchase.Frequency.Ratio = f2[,4],
f3[,4], f4[,4],f6[,4], f7[,4], ff4[,3:5 ])

dff2 <- dff[,2:10]


## 75% of the sample size
smp_size <- floor(0.001 * nrow(dff2))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(dff2)), size = smp_size)

train <- dff2[train_ind, ]
test <- dff2[-train_ind, ]

dim(train)
dim(test)

cc.data.pca <- prcomp(train,
                 center = TRUE,
                 scale. = TRUE)
print(cc.data.pca)

summary(cc.data.pca)
```

Meanshift

```
library(MeanShift)
library(ggfortify)

set.seed(123)

# mean and sd
cc.data.sd<-scale(cc.data.pca$x[,1:3])
cc.data <-t(cc.data.sd)
dim(cc.data)


# MeanShift clustering (Option 1)
CC.Clusters<- msClustering( cc.data, h=1.0 )
print(CC.Clusters$labels)
```