**Customer Segmentation: RFM Model & K-means Clustering Using R**

#Read the data from the working directory, create your own working directly to read the dataset.

setwd("C:/Users/Deep/Desktop/data")

data1 <- read.csv ("C:/Users/Deep/Desktop/data/

customer\_seg.csv",header=TRUE,sep=",")

## Performing Data exploration

#perform exploratory data analysis to know about the data

# display top 6 rows of dataset to see how data look like

head (data1)

# display bottom 6 rows

tail(data1)

# describe the structure of data

str(data1)

#display the column name of the data

names(data1)

#display the summary or descriptive statistics of the data

summary(data1$Amount)

#Let’s check the missing values present in the data

sum(is.na(data1))

#Unique number of Invoice

length(unique(data1$Invoice\_No))

#Unique customer\_id

length(unique(data1$Customer\_ID))

#installing dplyr package

install.packages("dplyr")

library(dplyr)

#Displaying date format for invoice\_date

data2 <- data1 %>%

mutate(Invoice\_Date=as.Date(Invoice\_Date, '%m/%d/%Y'))

NOW <- as.Date("2011-12-12", "%Y-%m-%d")

#Structure of data2 after changing the date format

str(data2)

**#Building RFM Model**

#Calculating Recency,Frequency and Monetary table

R\_table <- aggregate(Invoice\_Date ~ Customer\_ID, data2, FUN=max)

R\_table$R <-as.numeric(NOW - R\_table$Invoice\_Date)

F\_table <- aggregate(Invoice\_Date ~ Customer\_ID, data2, FUN=length)

M\_table <- aggregate(Amount ~ Customer\_ID, data2, FUN=sum)

#Merging the datasets, remove the unnecessary column and rename the columns

RFM\_data <- merge(R\_table,F\_table,by.x="Customer\_ID", by.y="Customer\_ID")

RFM\_data <- merge(RFM\_data,M\_table, by.x="Customer\_ID", by.y="Customer\_ID")

names(RFM\_data) <- c("Customer\_ID","Invoice\_Date", "Recency", "Frequency", "Monetary")

#Display top 6 observations from RFM\_data

head(RFM\_data)

#RFM scoring

#Rsegment 1 is very recent while Rsegment 5 is least recent

RFM\_data$Rsegment <- findInterval(RFM\_data$Recency, quantile(RFM\_data$Recency, c(0.0, 0.25, 0.50, 0.75, 1.0)))

#Fsegment 1 is least frequent while Fsegment 5 is most frequent

RFM\_data$Fsegment <- findInterval(RFM\_data$Frequency, quantile(RFM\_data$Frequency, c(0.0, 0.25, 0.50, 0.75, 1.0)))

#Msegment 1 is lowest sales while Msegment 5 is highest sales

RFM\_data$Msegment <- findInterval(RFM\_data$Monetary, quantile(RFM\_data$Monetary, c(0.0, 0.25, 0.50, 0.75, 1.0)))

#concatenating the RFM score into a single column

RFM\_data$Con <- paste(RFM\_data$Rsegment,RFM\_data$Fsegment,RFM\_data$Msegment)

#Total of RFM score

RFM\_data$Total\_RFM\_Score <- c(RFM\_data$Rsegment + RFM\_data$Fsegment+RFM\_data$Msegment)

#Display top 20 observations from RFM\_data after scoring

head(RFM\_data,20)

#Display the structure of RFM\_data

str(RFM\_data)

#to keep only selected variables from RFM data

clus\_df<-RFM\_data[,c(3,4,5)]

#Display top 6 observations from clus\_df

head(clus\_df)

#setting seed

set.seed(123)

#Applying K-means

km <-kmeans(clus\_df,centers =5,nstart = 30)

#Describing structure of km

str(km)

#Printing km

print(km)

#Computing centers

km$centers

#Assigning clusters to the each data point

km$cluster

#Computing ‘Within Sum of Squares’

km$withinss

#Computing ‘Total of the Within Sum of Squares’

km$tot.withinss

#Computing ‘Between Sum of Squares’

km$betweenss

#Computing ‘Total Sum of Squares’

km$totss

#installing factoextra package

install.packages("factoextra")

#clustering algorithms and visualization

library(factoextra)

fviz\_cluster(km, data = clus\_df)

#Combining clus\_df data and km$clusters by cbind

finalclus<-cbind(clus\_df,km$cluster)

#Displaying top 20 observations from finalclus

head(finalclus,20)

#Keeping customer id from RFM\_data

cus\_id <-RFM\_data[(1)]

#Combining cus\_id and finalclus by cbind

finalclusid<-cbind(cus\_id,finalclus)

#Displaying top 20 observations from finalclusid

head(finalclusid,20)