Project - 6 DS with R
E-commerce - To segment the customers with RFM calculations

```
###  Ingesting Data  ####
e_data <- read.csv(file.choose())
View(e_data)
data1 <- e_data
str(data1)
set.seed(10)




#### Loading necessary libraries  #####
library(dplyr)
library(lubridate)
library(tidyr)




#### Data cleaning #####
#### Allvalues less than 0 are converted to NA and then dropped ####

data1 <- data1 %>%
  mutate(Quantity = replace(Quantity, Quantity<=0, NA),
      UnitPrice = replace(UnitPrice, UnitPrice<=0, NA))

data1 <- data1 %>%
  drop_na()


#### converting character variable to factor #####
#### converting Date from 'char' to 'Date'   ####

data1 <- data1 %>%
  mutate(InvoiceNo=as.factor(InvoiceNo), StockCode=as.factor(StockCode),
      InvoiceDate=parse_date_time(InvoiceDate,  orders = c("dmy")),
CustomerID=as.factor(CustomerID),
      Country=as.factor(Country))

data1 <- data1 %>%
  mutate(total = Quantity*UnitPrice)

head(data1)
```

```
####  RFM calculations #########

df_RFM <- data1 %>%
  group_by(CustomerID) %>%
  summarise(recency= (parse_date_time("1-Jan-2018",  orders = c("dmy"))-
max(InvoiceDate)),
        frequency=n_distinct(InvoiceNo), monetary= sum(total)/n_distinct(InvoiceNo))


df_RFM <- as.data.frame(df_RFM)
df_RFM$recency <- as.integer(df_RFM$recency)
str(df_RFM)
head(df_RFM)
data_RFM <- df_RFM[,-1]

> head(df_RFM)
  CustomerID recency frequency   monetary
1    12346    350       1 77183.6000
2    12347    27      7   615.7143
3    12348    100      4   449.3100
4    12349    43       1 1757.5500
5    12350    335      1   334.4000
6    12352    61       8   313.2550
>




###### k-means clustering  #######

scale_RFM = scale(data_RFM)
dim(scale_RFM)
cluster_up <- kmeans(scale_RFM, 3, iter.max = 10)
str(cluster_up)

cluster_df <- cbind(scale_RFM, ClusterNum = cluster_up$cluster)
dim(cluster_df)
head(cluster_df)
str(cluster_df)
library(mclust)
fit <- Mclust(cluster_df)
plot(fit)
```
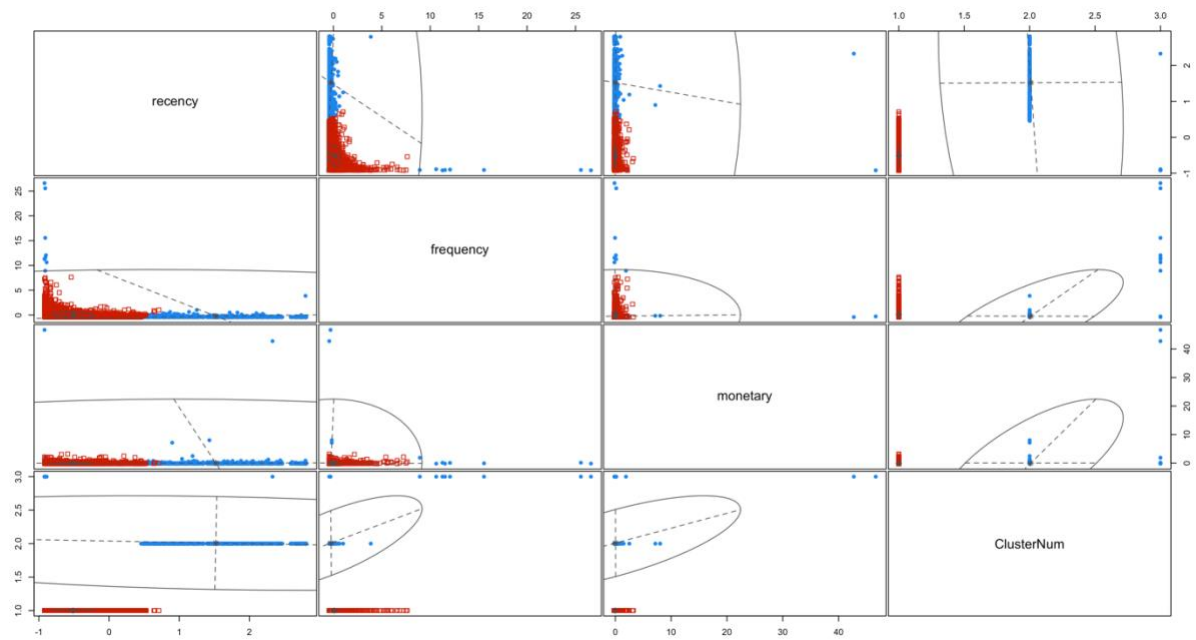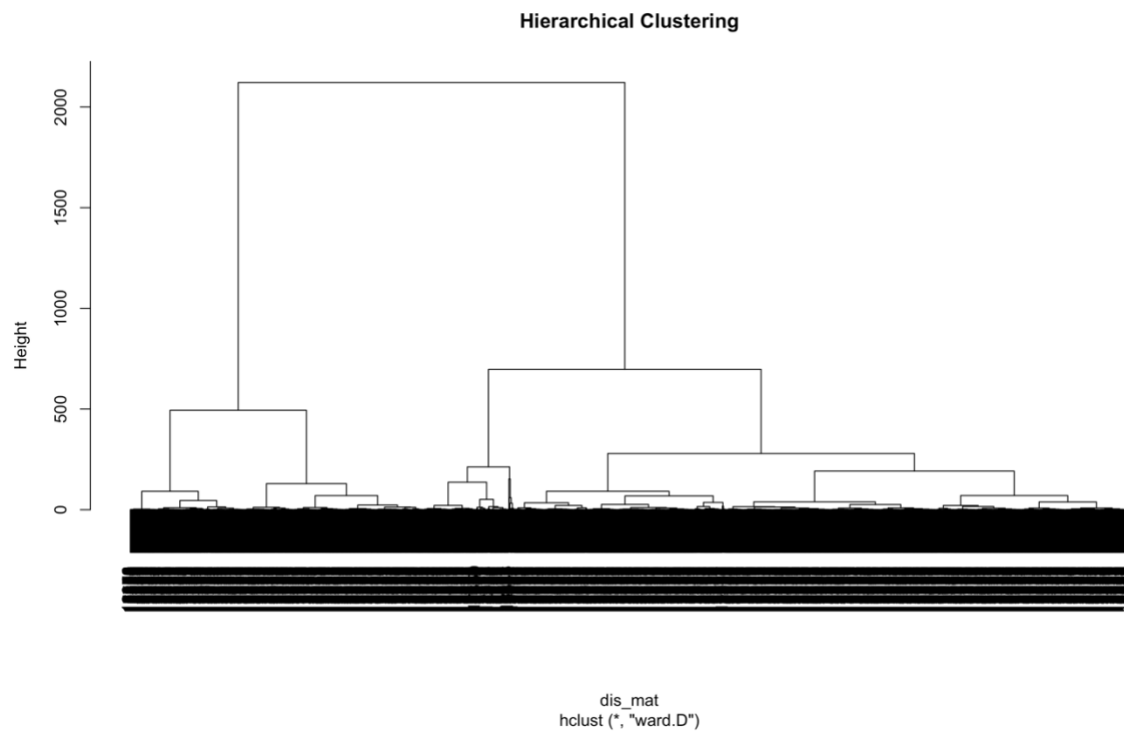
Hierarchical clustering

dis_mat <- dist(scale_RFM, method = 'euclidean')
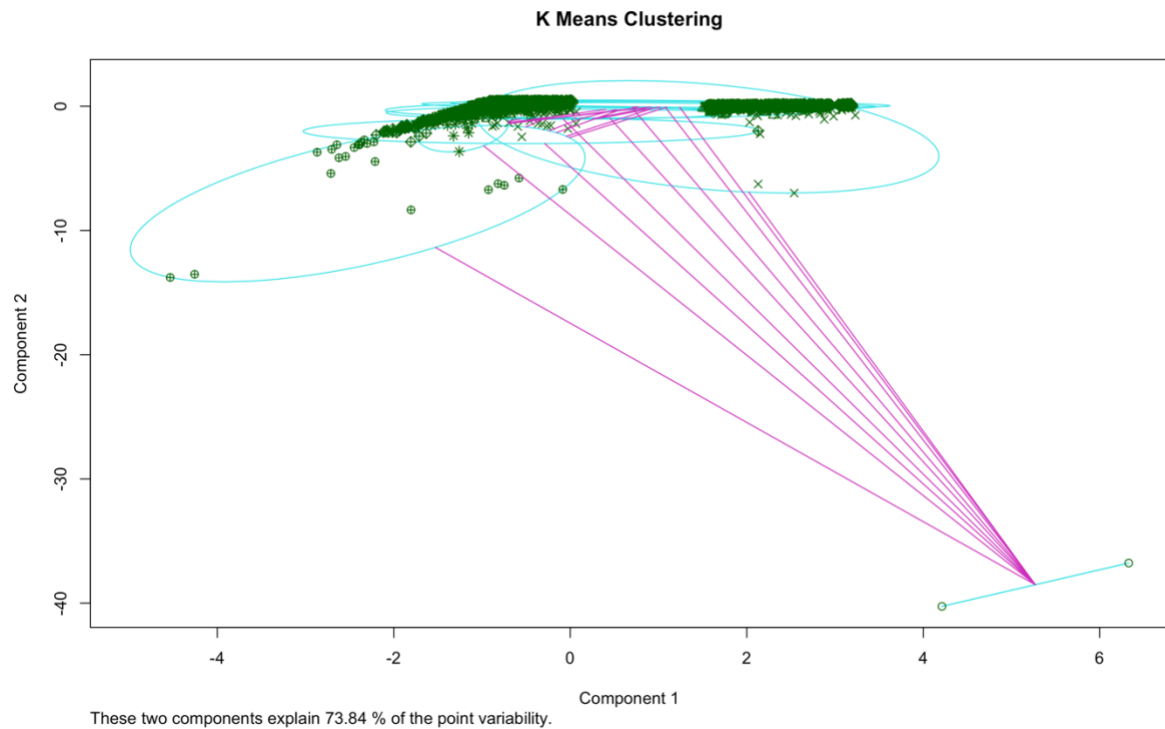hclus <- hclust(dis_mat, method = 'ward.D')

Dendogram

plot(hclus, labels = as.character(df_RFM$CustomerID),
    main = 'Hierarchical Clustering')

Hierarchical Clustering

dis_mat
hclust (*, "ward.D")

###Compare and visualize clusters
# Visualize the clusters
library(cluster)
cluster::clusplot(cluster_df, labels, main = 'K Means Clustering')

K Means Clustering

These two components explain 73.84 % of the point variability.

cluster::clusplot(scale_RFM, labels, main = 'H Clustering')

H Clustering

These two components explain 75.38 % of the point variability.