**Git Url :**

<https://github.com/ShilpaAChavan/Datascience-With-R/blob/master/Assignment/Clustering/ClusteringAssignment.R>

**Code :**

######################################CrimeAnalysis###############################

**#Problem Statement:**

**#Perform Clustering for the crime data and identify the number of clusters formed and draw #inferences.**

##################################################################################

#Using Hierarchial clustering

crime\_data <- read.csv(file.choose())

#crime\_data<-read.csv("D:/Shilpa/Datascience/Assignments/ClusteringAssignment/crime\_data.csv")

head(crime\_data)

crimedata <- crime\_data[2:5]

head(crimedata)

#Normalize the data

norm\_crimedata <- scale(crimedata)

head(norm\_crimedata)

d <- dist(norm\_crimedata,method="euclidean")

#calculating distance between clusters using complete linkage.

fit <- hclust(d,method="complete")

#Plot dendogram

plot(fit)

groups <- cutree(fit,k=3)

rect.hclust(fit,k=3,border="blue")

crimedatafinal <- cbind(crime\_data,groups)

crimedatafinal

aggregate(crimedatafinal[,2:6],by=list(crimedatafinal$groups),FUN = mean)

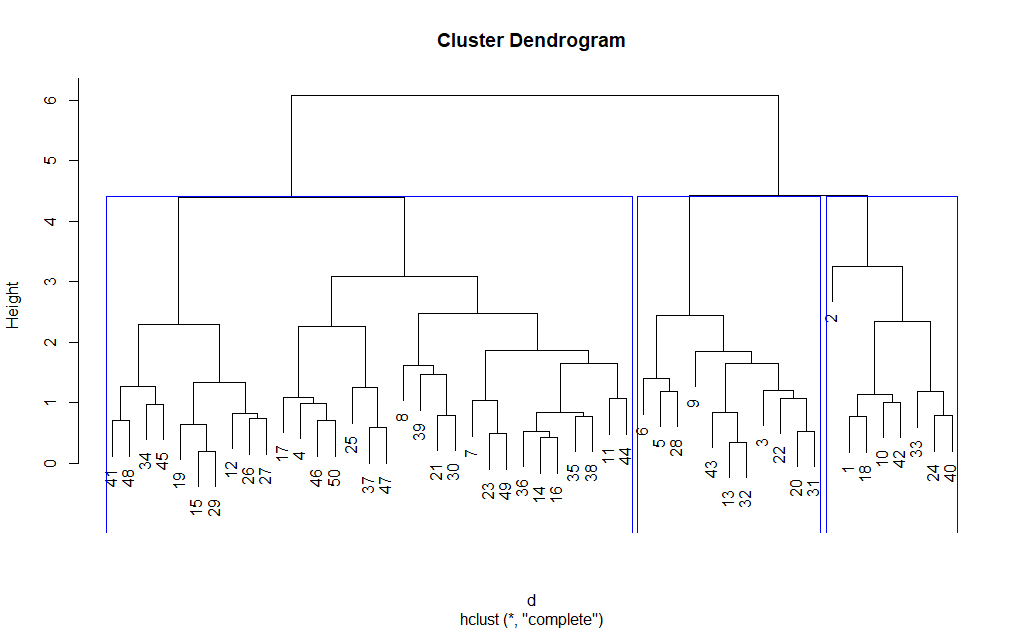
#Group.1 Murder Assault UrbanPop Rape groups

#1 14.087500 252.7500 53.50000 24.53750 1

#2 11.054545 264.0909 79.09091 32.61818 2

#3 5.003226 116.4839 63.83871 16.33871 3

**Dendogram :**



#**Interpretation**: Group 2 Countries have higher rate of crime and Group 3 have

# comparatively lower crime rate.

########################## With Kmeans method ##########################

install.packages("plyr")

library(plyr)

#Elbow chart

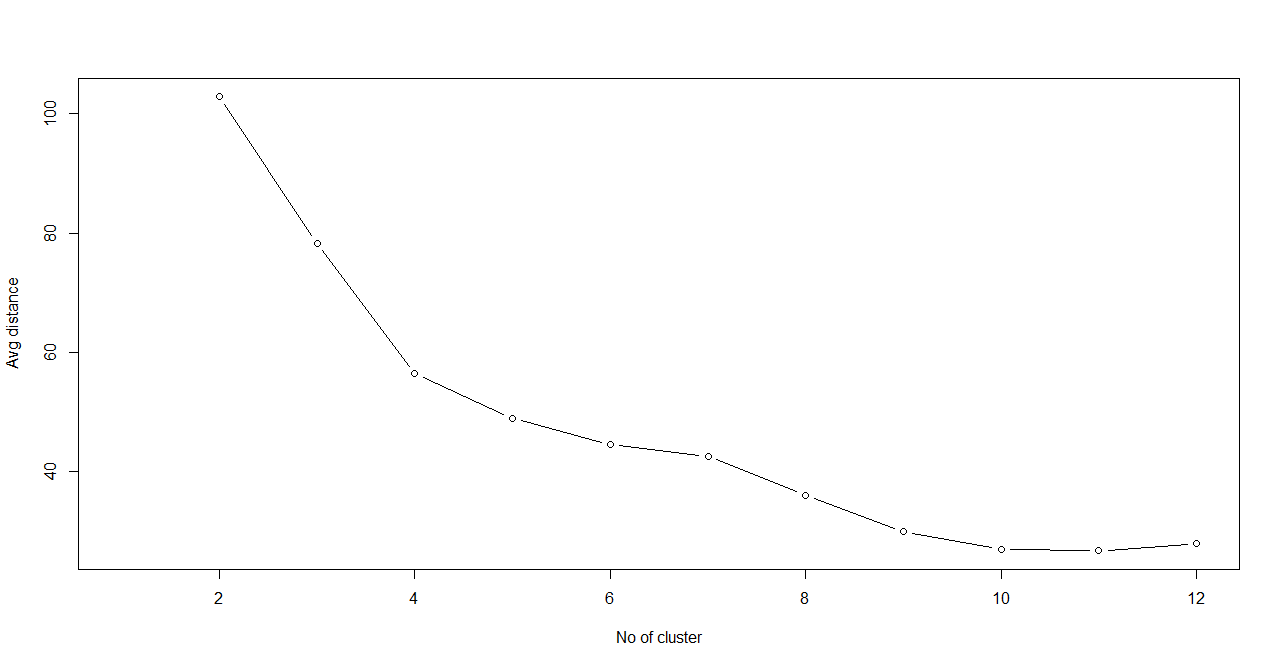
wss <-c()

for(i in 2:12) wss[i] <- sum(kmeans(norm\_crimedata,centers = i)$withinss)

plot(1:12,wss,type="b",xlab="No of cluster",ylab="Avg distance")

#the distortion rate becomes constant is the optimal value.Here k=4.

**Elbow Chart :**



# k=4 optimal kvalue

crimeDataKmeans <- kmeans(norm\_crimedata,4)

str(crimeDataKmeans)

summary(crimeDataKmeans)

crimeDataKmeans$centers

crimeData <- cbind(crime\_data,crimeDataKmeans$cluster)

crimeData

colnames(crimeData)

crimeDataKmeans$centers

aggregate(crimeData[,2:6],by=list(crimeDataKmeans$cluster),FUN = mean)

#Group.1 Murder Assault UrbanPop Rape crimeDataKmeans$cluster

#1 13.93750 243.62500 53.75000 21.41250 1

#2 5.65625 138.87500 73.87500 18.78125 2

#3 3.60000 78.53846 52.07692 12.17692 3

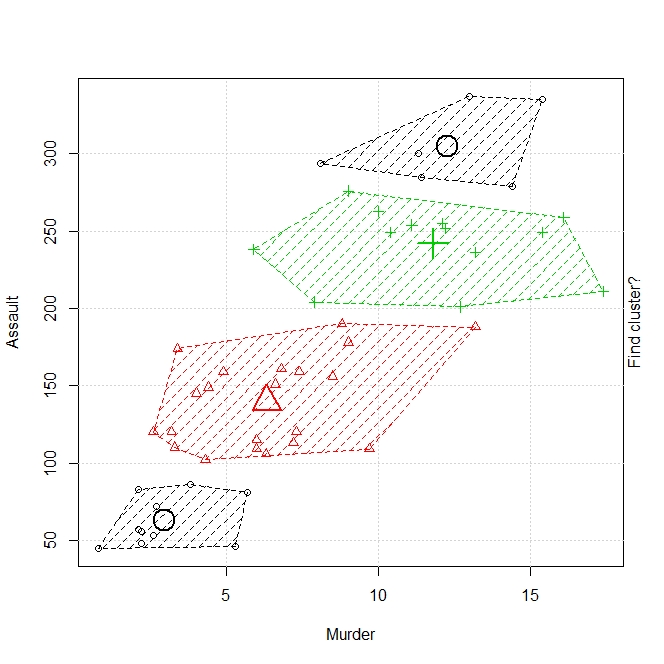
#4 10.81538 257.38462 76.00000 33.19231 4

install.packages("animation")

library(animation)

windows()

km <- kmeans.ani(crimedata,4)



#**Interpretation** : Group 4 countries have higher crime rate with higher assaults as well

#there urban population is dense followed by Group 1 countries, then group 2 and

# group3.

**Git URL:**

<https://github.com/ShilpaAChavan/Datascience-With-R/blob/master/Assignment/Clustering/ClusterAssignment_AirlinesEx.R>

############################################################################

**#Problem Statement:**

**#Perform clustering (Both hierarchical and K means clustering) for the airlines data to obtain optimum number of clusters.**

**#Draw the inferences from the clusters obtained.**

############################################################################

library(readr)

library(readxl)

#eastWestAirlines <- read\_excel("D:/Shilpa/Datascience/Assignments/ClusteringAssignment/EastWestAirlines.xlsx",sheet = "data")

eastWestAirlines<-read\_excel(file.choose(),2)

View(eastWestAirlines)

head(eastWestAirlines)

normEastWestAirlines <- scale(eastWestAirlines[,2:12])

distEastWestAirlines <- dist(normEastWestAirlines,method = "euclidean")

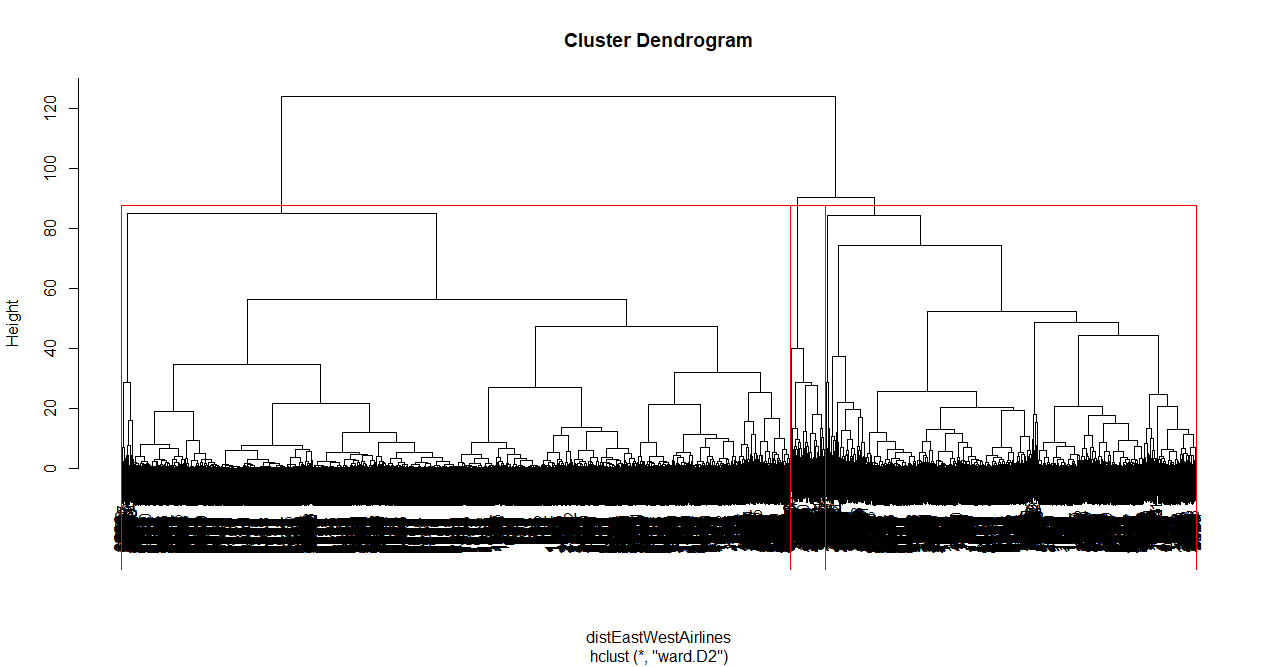
#Calculate distance between clusters using complete linkage

airlineClust <- hclust(distEastWestAirlines,method = "ward.D2")

plot(airlineClust)

# 3 clusters are formed at height 85.

**Dendogram:**



groups <- cutree(airlineClust,k=3)

table(groups)

rect.hclust(airlineClust,k=3,border="red")

airlinesFinal <- cbind(eastWestAirlines,groups)

colnames(airlinesFinal)

aggregate(eastWestAirlines[,2:12],FUN = mean,by=list(airlinesFinal$groups))

#Group.1 Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles Bonus\_trans

# 1 46718.86 9.274407 1.242266 1.023303 1.000000 5037.793 7.091201

# 2 116314.45 363.839130 3.498551 1.000000 1.035507 37150.357 18.066667

# 3 134880.89 393.323077 2.430769 1.000000 1.000000 36582.169 29.338462

#Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll Award?

# 221.1671 0.7002812 3772.786 0.1880273

# 377.0000 1.1500000 4696.888 0.6630435

# 5915.5231 16.6384615 4599.608 0.7538462

**#Interpretation**: It seems the clusters are classified using balance, bonus miles and

days since enrolled.

Cluster 1(Non frequent flyer: relatively recent passengers) :2489 total in the group the passengers are relatively new with low balance and bonus miles.

Cluster 3(Frequent flyer: old passengers): 130 total in this group. These are old passengers with high balance and bonus miles, also with more number of flight miles in the past 12 months

Cluster 2: passengers can be said as intermediate flyers.

#################### Using K-Means#########################################

airlineKmeans <- kmeans(normEastWestAirlines,5)

str(airlineKmeans)

summary(airlineKmeans)

airlineKmeans$centers

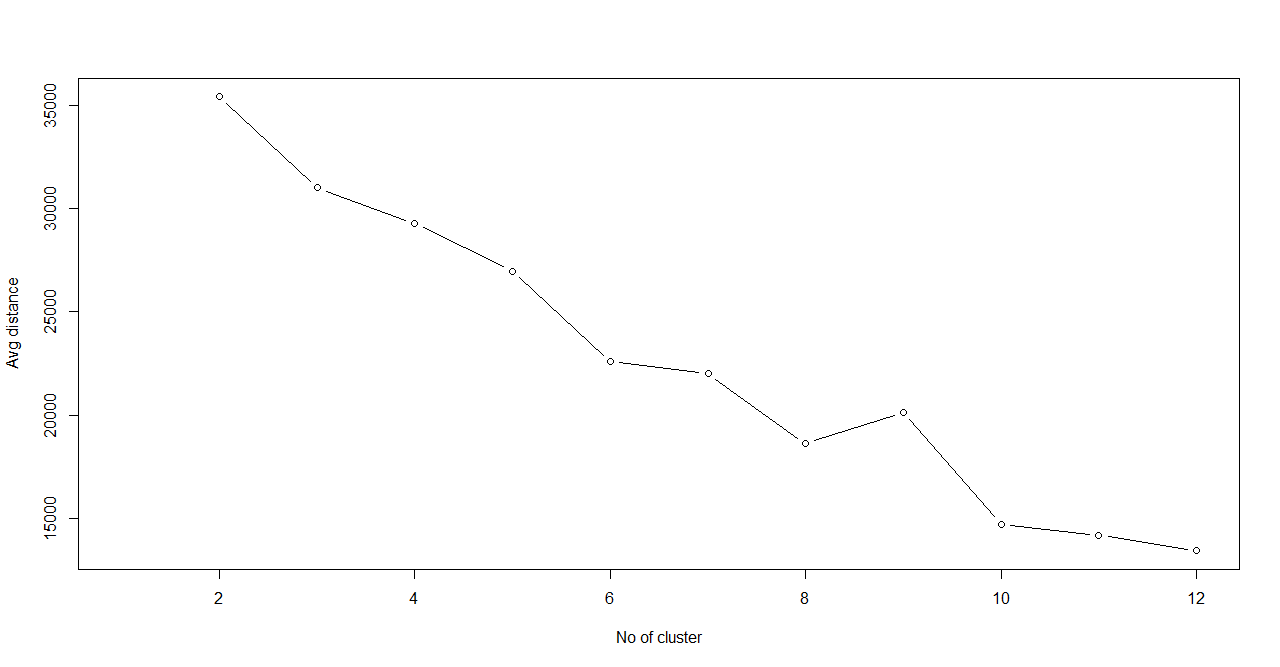
#Elbow chart

wss <-c()

for(i in 2:12) wss[i] <- sum(kmeans(normEastWestAirlines,centers = i)$withinss)

plot(1:12,wss,type="b",xlab="No of cluster",ylab="Avg distance")

#the distortion rate becomes constant is the optimal value.Here k=3.

**ElbowChart**: 

airlineKmeans <- kmeans(normEastWestAirlines,3)

summary(airlineKmeans)

airlineKmeans$centers

eastWestAirlines <- cbind(eastWestAirlines,airlineKmeans$cluster)

colnames(eastWestAirlines)

aggregate(eastWestAirlines[,2:12],FUN = mean,by=list(airlineKmeans$cluster))

#Group.1 Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles Bonus\_trans

# 1 116978.71 156.82049 3.731533 1.002383 1.034154 40417.577 18.960286

# 2 44370.06 96.83333 1.235820 1.019037 1.000389 4831.835 7.005439

# 3 197873.37 780.89157 2.150602 1.036145 1.030120 31562.446 27.066265

#Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll Award?

# 342.4543 1.0119142 4884.592 0.6528991

# 200.6974 0.6262626 3704.443 0.2039627

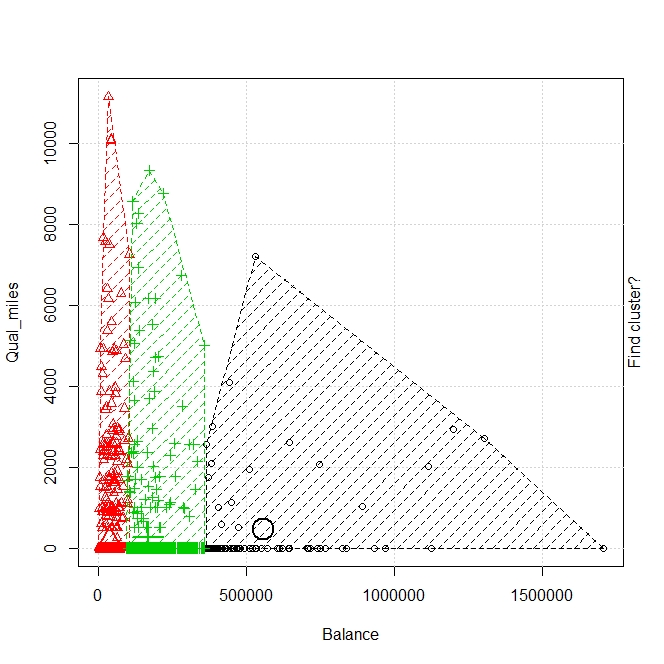
# 5373.6024 15.7048193 4730.018 0.8072289

install.packages("animation")

library(animation)

windows()

km <- kmeans.ani(eastWestAirlines[2:12],3)



**#Interpretation:**

Cluster 2(Non frequent flyer: relatively recent passengers) :2574 total in the group the passengers are relatively new with low balance and bonus miles.

Cluster 3(Frequent flyer: old passengers) : 166 total in this group. These are old passengers with high balance and bonus miles, also with more Number of flight miles in the past 12 months

Cluster 1 (1259 passengers): passengers can be said as intermediate flyers.

**Conclusion** : Recently added passengers are more in number which are non-frequent

flyer, discounts on fare other offerings should be given to improve the number fly.