K means Clustering Documentation

Business Problem: 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.

About the data:

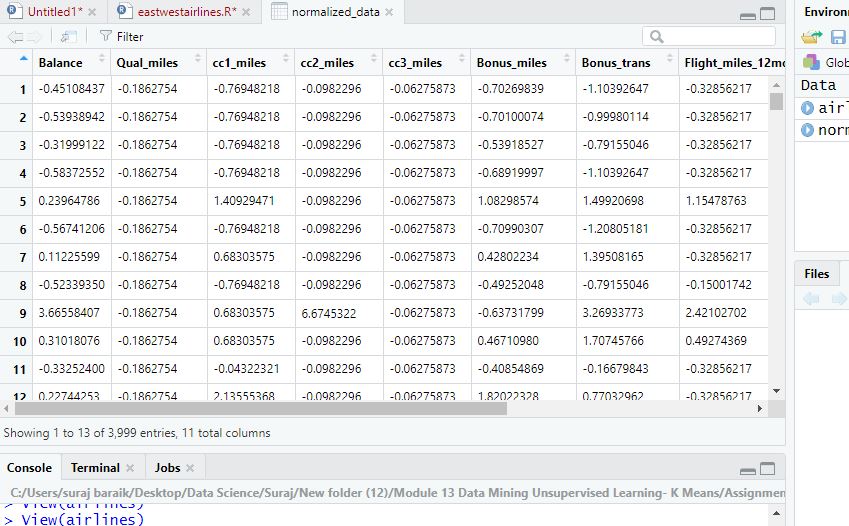
The data file is of .xlsx format. To access the file of .xlsx format , we have to call the library of “openxlsx”

library(openxlsx)

The data file contain 12 columns and total of 3999 entries.

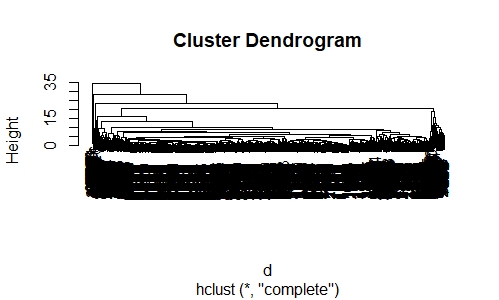
Normalizing continuous columns to bring them under same scale

normalized\_data<-scale(airlines[,2:12])

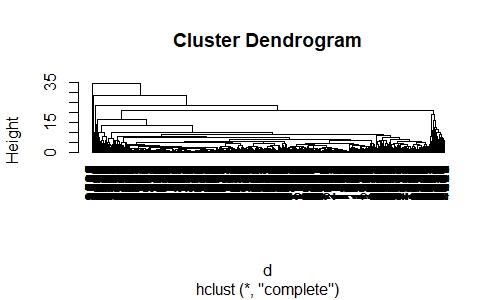


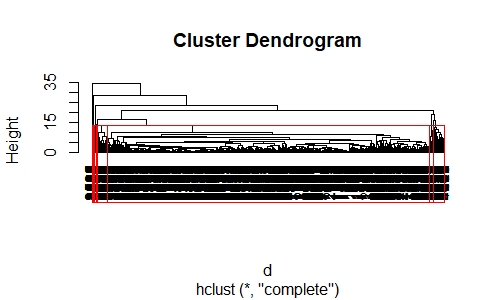
Perform Euclidean distance matrix

display dendrogram



Display Dendogram (plot(fit, hang=-1))





membership<-as.matrix(groups) # groups or cluster numbers

View(final)

aggregate(airlines[,-1],by=list(final$membership),mean)

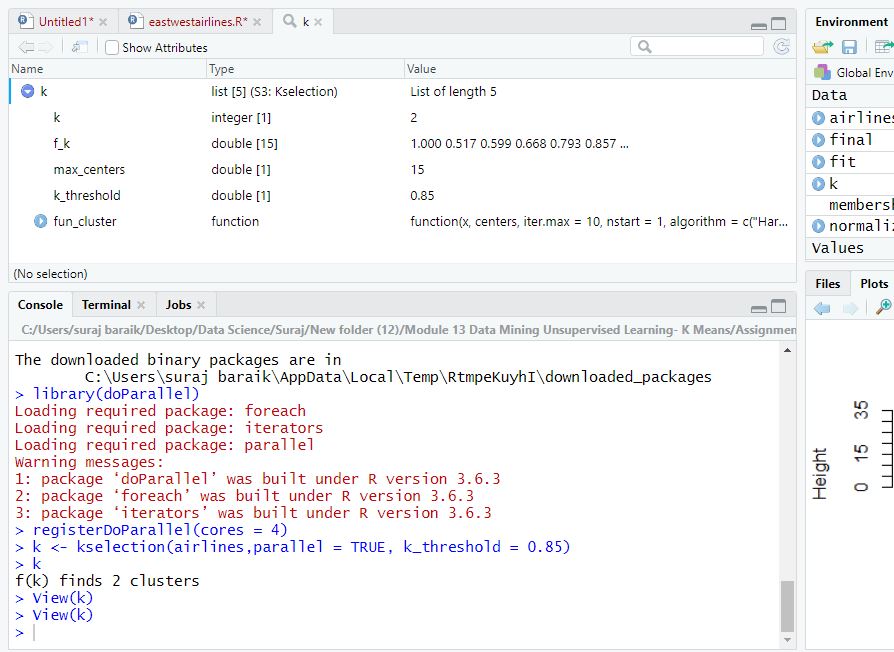
# as dataset containing large no of observations we are not able to decide no.of clusters so we will go for kmeans clustering

install.packages("kselection")

library(kselection)

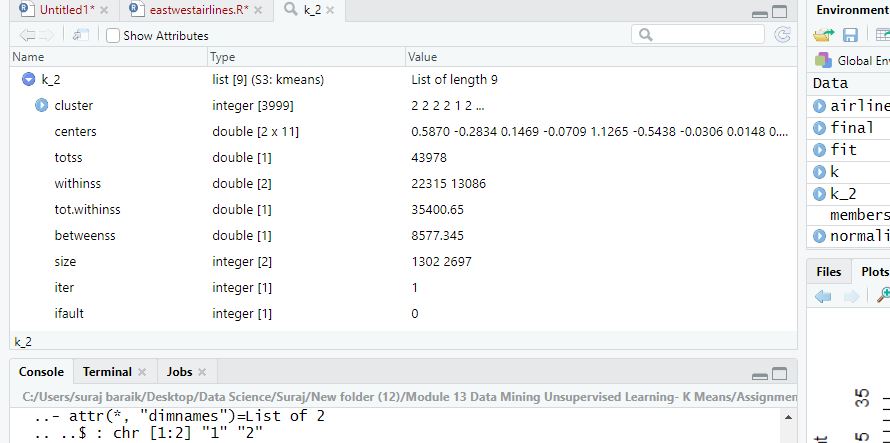
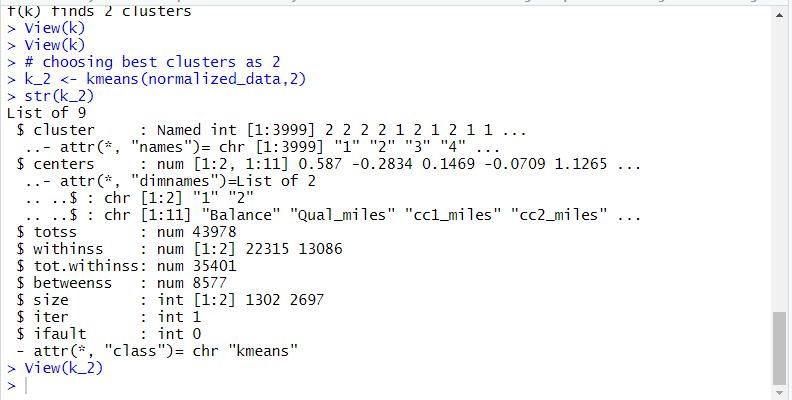
install.packages("doParallel")

k <- kselection(airlines,parallel = TRUE, k\_threshold = 0.85)

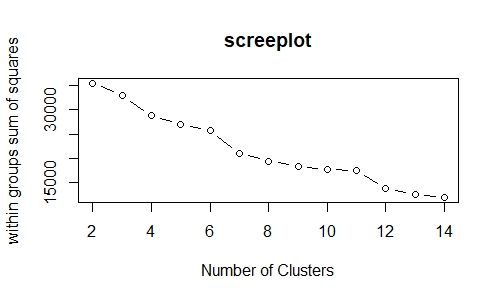


#from k selection it shows that 2 clusters is optimum number, choosing best clusters as 2.

k\_2 <- kmeans(normalized\_data,2)



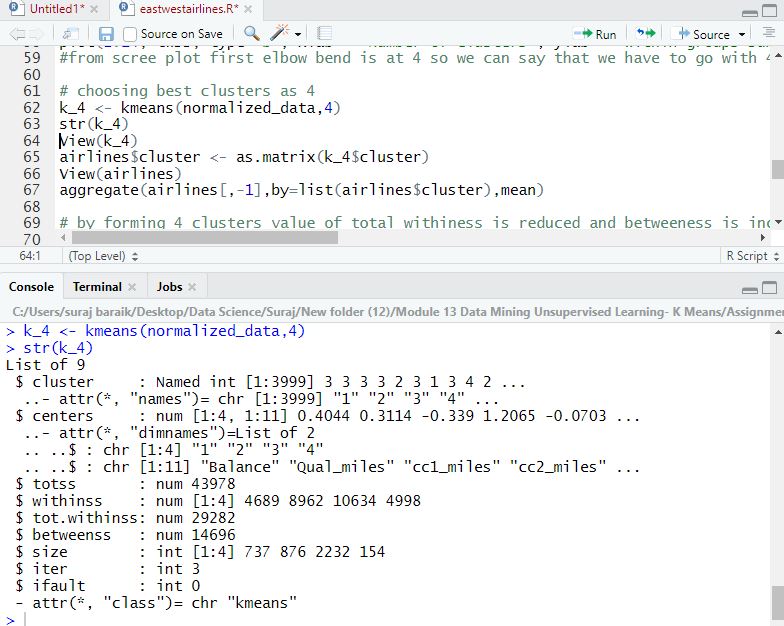
Scree plot for selection of no.of clusters



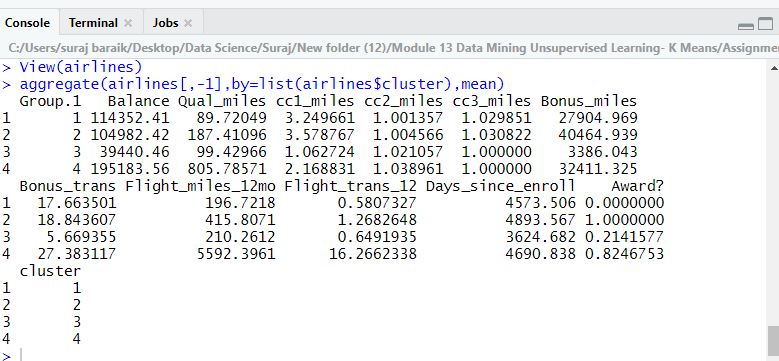
From scree plot first elbow bend is at 4 so we can say that we have to go with 4 no. of clusters, choosing best clusters as 4.

k\_4 <- kmeans(normalized\_data,4)

str(k\_4)



aggregate(airlines[,-1],by=list(airlines$cluster),mean)



# by forming 4 clusters value of total withiness is reduced and betweeness is increased but we will check that once by second elbow turn at 5

# by considering 5 clusters

# choosing best clusters as 4

k\_5 <- kmeans(normalized\_data,4)

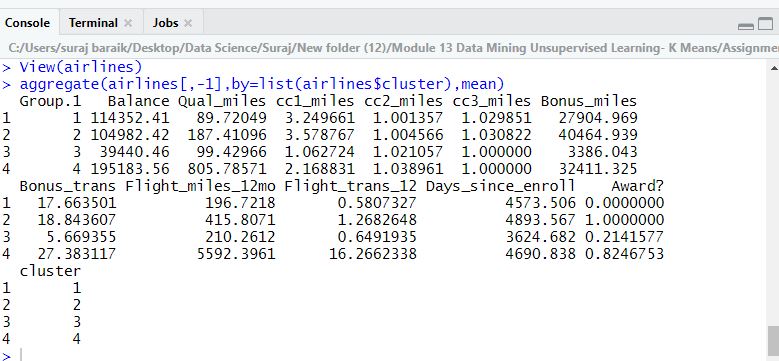
str(k\_5)

View(k\_5)

airlines$cluster <- as.matrix(k\_5$cluster)

View(airlines)

aggregate(airlines[,-1],by=list(airlines$cluster),mean)



# by forming 5 clusters value of total withiness is reduced slightly and betweeness is increased slightly but we will check that once by second elbow turn at 5

# so there is no need to do seperate cluster for slight change so we will finalise with 4 no. of clusters

#Cluster 1 has the largest average values in Qualmiles,Bonus\_miles,Bonus\_trans,Flight\_miles\_12mo,Flight\_trans\_12

#Cluster 1 are infrequent but loyal customers. Mostly contains customers with few miles, but who have been with the airline the longest.

#Customers in cluster1 mostly preferred to travel in last 12 months,Day\_since\_enroll.

#Cluster 2 has the largest average values in Balance,cc1\_miles,Day\_since\_enroll.

#these customers are oldest one and used our service less in last 12 months

#Cluster 3 are frequent user of our airline.as they are not new and uses frequently our service we need to give them promotinal packages.

#Cluster 4 are newest customers of our airline as value of day\_since\_enroll is having less value.