**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.**

**Data Description:**

**The file EastWestAirlinescontains information on passengers who belong to an airline’s frequent flier program. For each passenger the data include information on their mileage history and on different ways they accrued or spent miles in the last year. The goal is to try to identify clusters of passengers that have similar characteristics for the purpose of targeting different segments for different types of mileage offers**

**ID --Unique ID**

**Balance--Number of miles eligible for award travel**

**Qual\_mile--Number of miles counted as qualifying for Topflight status**

**cc1\_miles -- Number of miles earned with freq. flyer credit card in the past 12 months:**

**cc2\_miles -- Number of miles earned with Rewards credit card in the past 12 months:**

**cc3\_miles -- Number of miles earned with Small Business credit card in the past 12 months:**

**1 = under 5,000**

**2 = 5,000 - 10,000**

**3 = 10,001 - 25,000**

**4 = 25,001 - 50,000**

**5 = over 50,000**

**Bonus\_miles--Number of miles earned from non-flight bonus transactions in the past 12 months**

**Bonus\_trans--Number of non-flight bonus transactions in the past 12 months**

**Flight\_miles\_12mo--Number of flight miles in the past 12 months**

**Flight\_trans\_12--Number of flight transactions in the past 12 months**

**Days\_since\_enrolled--Number of days since enrolled in flier program**

**Award--whether that person had award flight (free flight) or not**

**#Hierarchical clustering**

**#using readx1 package for reading excel file**

install.packages("readxl")

library(readxl)

data <- read\_xlsx(file.choose(),2)# Read xlsx file

View(data)

head(data)

**#display the structure of the data**

str(data)

**# here we can see that variable cc1\_miles,cc2\_miles,cc3\_miles and Award? are categorical form**

**#1 = under 5,000**

**#2 = 5,000 - 10,000**

**#3 = 10,001 - 25,000**

**#4 = 25,001 - 50,000**

**#5 = over 50,000**

**# we have to take average of cc1\_miles,cc2\_miles,cc3\_miles above respective range**

data$cc1\_miles = ifelse(data$cc1\_miles==1,3500,

ifelse(data$cc1\_miles==2,8000,

ifelse(data$cc1\_miles==3,20000,

ifelse(data$cc1\_miles==4,38000,

ifelse(data$cc1\_miles==5,65000,0)))))

data$cc2\_miles = ifelse(data$cc2\_miles==1,3500,

ifelse(data$cc2\_miles==2,8000,

ifelse(data$cc2\_miles==3,20000,

ifelse(data$cc2\_miles==4,38000,

ifelse(data$cc2\_miles==5,65000,0)))))

data$cc3\_miles = ifelse(data$cc3\_miles==1,3500,

ifelse(data$cc3\_miles==2,8000,

ifelse(data$cc3\_miles==3,20000,

ifelse(data$cc3\_miles==4,38000,

ifelse(data$cc3\_miles==5,65000,0)))))

View(data)

Head(data)

ID#` Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles Bonus\_trans

*<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

1 1 28143 0 0 3500 3500 174 1

2 2 19244 0 0 3500 3500 215 2

3 3 41354 0 0 3500 3500 4123 4

4 4 14776 0 0 3500 3500 500 1

5 5 97752 0 0 3500 3500 43300 26

6 6 16420 0 0 3500 3500 0 0

# ... with 4 more variables: Flight\_miles\_12mo *<dbl>*, Flight\_trans\_12 *<dbl>*,

# Days\_since\_enroll *<dbl>*, `Award?` *<dbl>*

**#Normalize the data**

**#here i am taking 2 to 11 columns**

normalized\_Airline\_data <- scale(data[2:11]) #excluding the ID and Award column before normalizing

View(normalized\_Airline\_data)

Head(normalized\_Airline\_data)

Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles Bonus\_trans

[1,] -0.4510844 -0.1862754 -0.6475233 -0.08692286 -0.05703711 -0.7026984 -1.1039265

[2,] -0.5393894 -0.1862754 -0.6475233 -0.08692286 -0.05703711 -0.7010007 -0.9998011

[3,] -0.3199912 -0.1862754 -0.6475233 -0.08692286 -0.05703711 -0.5391853 -0.7915505

[4,] -0.5837255 -0.1862754 -0.6475233 -0.08692286 -0.05703711 -0.6892000 -1.1039265

[5,] 0.2396479 -0.1862754 1.2445046 -0.08692286 -0.05703711 1.0829857 1.4992070

[6,] -0.5674121 -0.1862754 -0.6475233 -0.08692286 -0.05703711 -0.7099031 -1.2080518

Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll

[1,] -0.3285622 -0.3621226 1.395280

[2,] -0.3285622 -0.3621226 1.379784

[3,] -0.3285622 -0.3621226 1.411744

[4,] -0.3285622 -0.3621226 1.372037

[5,] 1.1547876 0.6924037 1.363805

[6,] -0.3285622 -0.3621226 1.367195

**#create distance matrix using Euclidean distance**

distance<-dist(normalized\_Airline\_data,method = "euclidean")

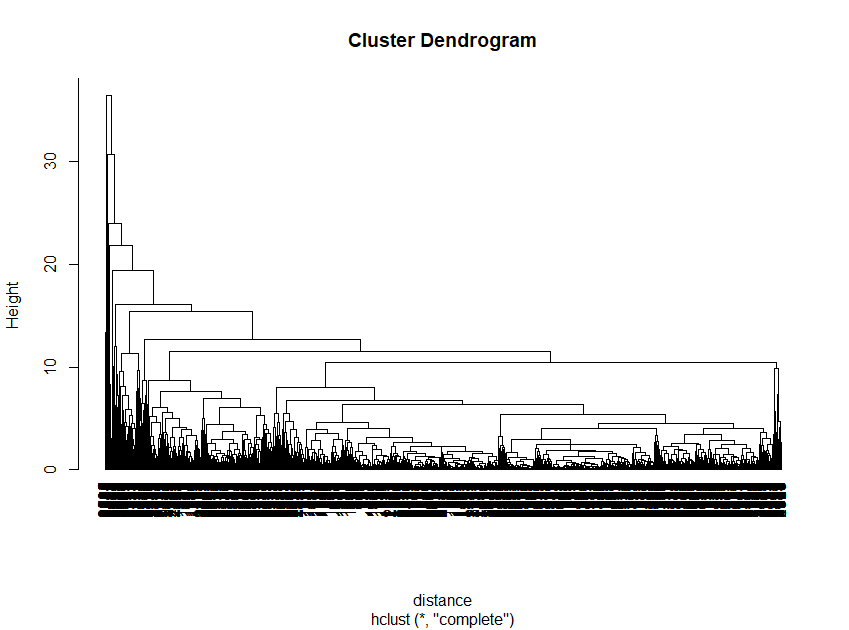
**# create dendrogram using complete linkage**

fit\_Airline\_data1 <- hclust(distance, method="complete")

fit\_Airline\_data1

**#plot the dendrogram**

plot(fit\_Airline\_data1, hang=-1)

**# using complete, it is very difficult to interpret, so we would try different**

**#linkages for better dendrogram**

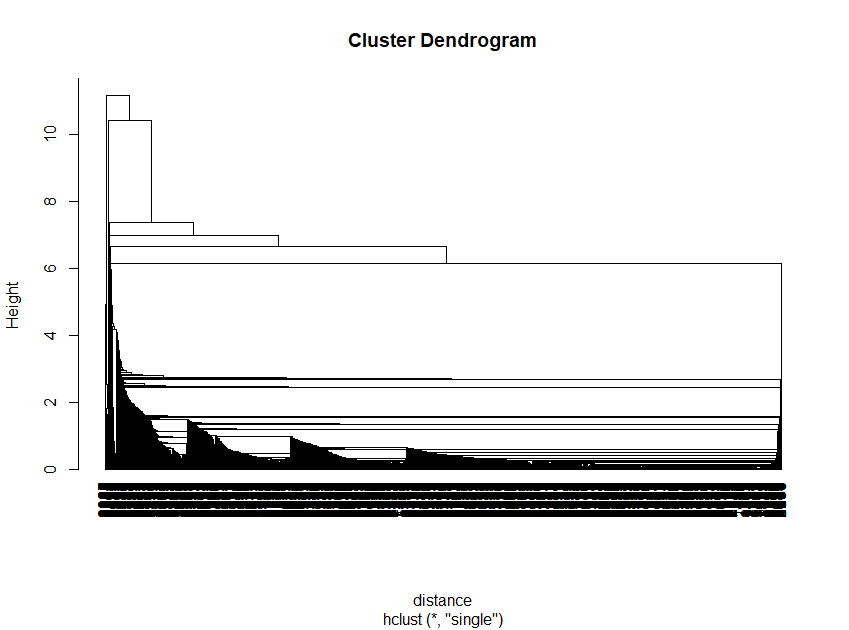
**# next i am using single linkages**

fit\_Airline\_data2 <- hclust(distance, method="single")

fit\_Airline\_data2

**#plot the dendrogram**

plot(fit\_Airline\_data2, hang=-1)



**# second dendrogram using single linkage is also very difficult to interpret**

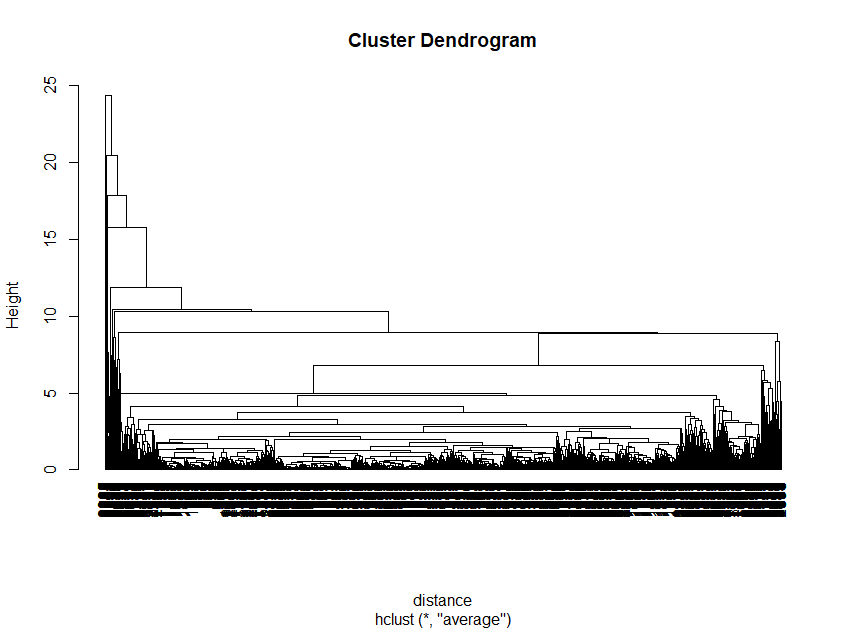
**# using average linkages**

fit\_Airline\_data3 <- hclust(distance, method="average")

fit\_Airline\_data3

#**plot the dendrogram**

plot(fit\_Airline\_data3, hang=-1)



**#using average linkage we got complex dendrogram**

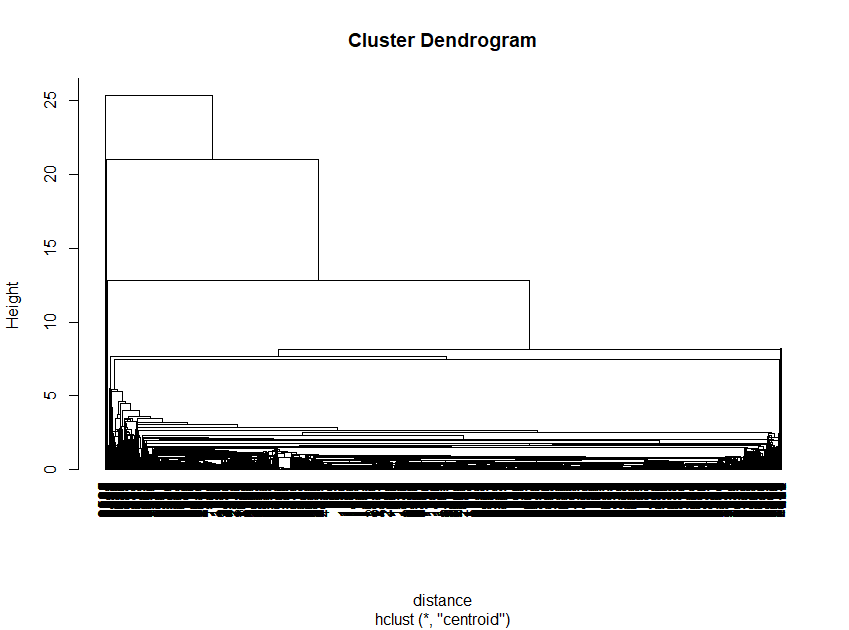
**#using centroid linkage**

fit\_Airline\_data4 <- hclust(distance, method="centroid")

fit\_Airline\_data4

**#plot the dendrogram**

plot(fit\_Airline\_data4, hang=-1)



**#using centriod linkage we got complex dendrogram**

**#Ward's minimum variance method: It minimizes the total within-cluster variance.**

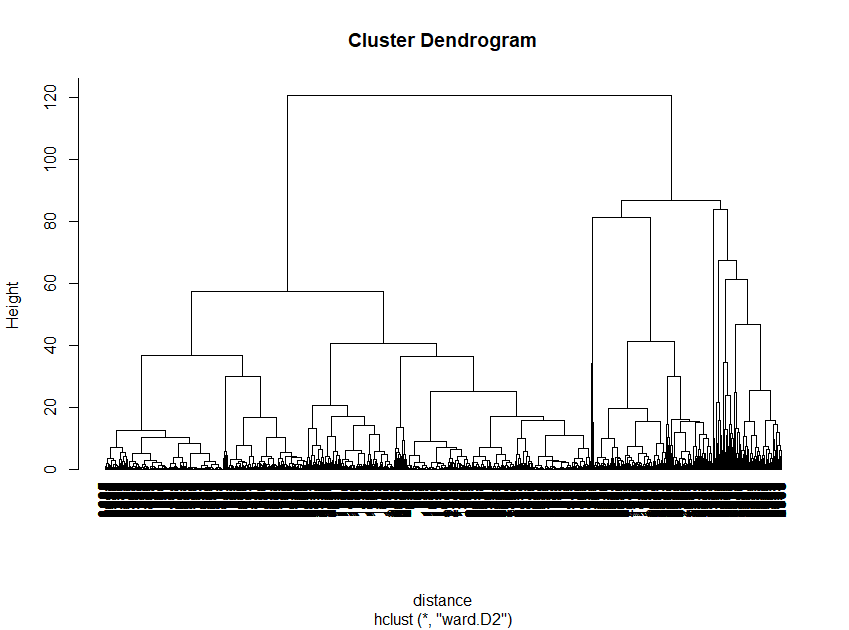
**#At each step the pair of clusters with minimum between-cluster distance are merged**.

fit\_Airline\_data5 <- hclust(distance, method="ward.D2")

fit\_Airline\_data5

**#plot the dendrogram**

plot(fit\_Airline\_data5, hang=-1,)



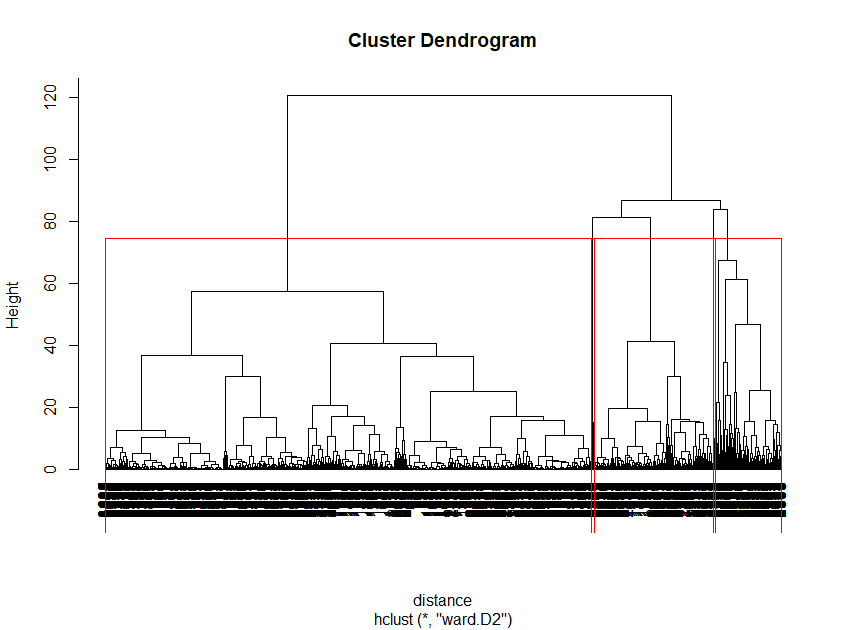
**#using ward.D2 linkage we got better dendrogram**

**#next step is to cut the dedrogram**

**#here i am going to cut dendrogram into five cluster**

Airline\_groups <- cutree(fit\_Airline\_data5, k=5)# cut tree into 5 clusters

rect.hclust(fit\_Airline\_data5, k=5, border="red")



**#convert groups information into a matrix for better understanding**

Airline\_cluster<-as.matrix(Airline\_groups)

View(Airline\_cluster)

head(Airline\_cluster)

[,1]

[1,] 1

[2,] 1

[3,] 1

[4,] 1

[5,] 2

[6,] 1

**#create dataframe to combine membership and original data**

final\_data <- data.frame(data, Airline\_cluster)

View(final\_data)

head(final\_data)

ID. Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles Bonus\_trans

1 1 28143 0 0 3500 3500 174 1

2 2 19244 0 0 3500 3500 215 2

3 3 41354 0 0 3500 3500 4123 4

4 4 14776 0 0 3500 3500 500 1

5 5 97752 0 0 3500 3500 43300 26

6 6 16420 0 0 3500 3500 0 0

Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll Award. Airline\_cluster

1 0 0 7000 0 1

2 0 0 6968 0 1

3 0 0 7034 0 1

4 0 0 6952 0 1

5 2077 4 6935 1 2

6 0 0 6942 0 1

**#here i am going to change the position of the column Airline\_cluster in to first**

final\_data1 <- final\_data[,c(ncol(final\_data),1:(ncol(final\_data)-1))]

View(final\_data1)

head(final\_data1)

Airline\_cluster ID. Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles

1 1 1 28143 0 0 3500 3500 174

2 1 2 19244 0 0 3500 3500 215

3 1 3 41354 0 0 3500 3500 4123

4 1 4 14776 0 0 3500 3500 500

5 2 5 97752 0 0 3500 3500 43300

6 1 6 16420 0 0 3500 3500 0

Bonus\_trans Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll Award.

1 1 0 0 7000 0

2 2 0 0 6968 0

3 4 0 0 7034 0

4 1 0 0 6952 0

5 26 2077 4 6935 1

6 0 0 0 6942 0

**#display first 10 ID of each clusters**

|  |
| --- |
| **cluster1 <-subset(final\_data1,Airline\_cluster==1)**  **> cluster1$ID.[1:10]**  **[1] 1 2 3 4 6 7 8 10 11 13**  **> cluster2 <-subset(final\_data1,Airline\_cluster==2)**  **> cluster2$ID.[1:10]**  **[1] 5 9 23 36 43 44 46 51 60 64**  **> cluster3 <-subset(final\_data1,Airline\_cluster==3)**  **> cluster3$ID.[1:10]**  **[1] 12 16 17 21 33 42 53 57 68 73**  **> cluster4 <-subset(final\_data1,Airline\_cluster==4)**  **> cluster4$ID.[1:10]**  **[1] 109 389 861 915 1047 1257 1313 1895 1934 1963**  **> cluster5 <-subset(final\_data1,Airline\_cluster==5)**  **> cluster5$ID.**  **[1] 202 386 951 1283 1808 1935 2064 2183 2296 3000 3021 3225 3390 3634 3802** |
|  |
| |  | | --- | | > | |

#**display the average of the Number of flight miles in the past 12 months of each cluster**

tapply(final\_data1$Flight\_miles\_12mo,Airline\_cluster,mean)

1 2 3 4 5

135.6844 3189.3753 270.7493 506.6667 692.6667

**#here we can see that second cluster has the highest average of the Number of flight miles in the past 12 months**

**#Award--whether that person had award flight (free flight) or not**

tapply(final\_data1$Award.,Airline\_cluster,sum)

1 2 3 4 5

763 257 447 8 6

**#here we can see that first cluster passanger has got more awards**