**Clustering**

**Q). 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? CHAR--Has member earned miles with airline freq. flyer credit card in the past 12 months (1=Yes/0=No)?

cc2\_miles? CHAR--Has member earned miles with Rewards credit card in the past 12 months (1=Yes/0=No)?

cc3\_miles? --Has member earned miles with Small Business credit card in the past 12 months (1=Yes/0=No)?

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

**Ans**:

> library(readxl)

> Airlines <- read\_excel("EastWestAirlines.xlsx",

+ sheet = "data")

1. **Hierarchical Clustering**

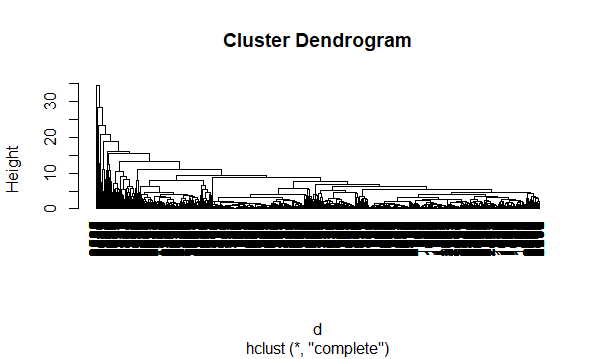
> normalized\_data<-scale(Airlines[,2:7])

> normalized\_data<-scale(Airlines[,2:11])

> d <- dist(normalized\_data, method = "euclidean") # distance matrix

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

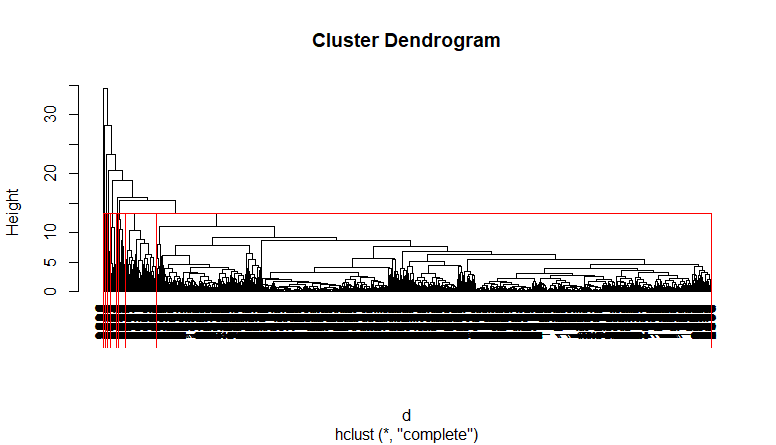
> plot(fit, hang=-1)



**Given DataFrame contains 3999 observations, which is huge no. of records for hierarchical Clustering. But as per problem let us start grouping the Clusters.**

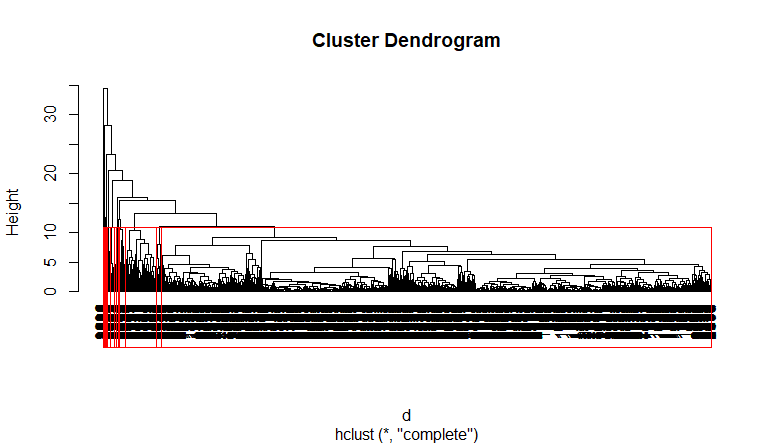
**Start with k= 10 clusters.**

> rect.hclust(fit, k=10, border="red")



**K= 15 clusters**

> rect.hclust(fit, k=15, border="red")



**So From above two plots of Hierarchical clustering with k=10 and k=15, for larger datasets, are Insignificant.**

So let us move to The K means clustering.

1. **K means clustering:**

> fit <- kmeans(normalized\_data,3)

> str(fit)

List of 9

$ cluster : int [1:3999] 3 3 3 3 1 3 1 3 2 1 ...

$ centers : num [1:3, 1:10] 0.42301 1.21837 -0.30666 -0.00728 0.84164 ...

..- attr(\*, "dimnames")=List of 2

.. ..$ : chr [1:3] "1" "2" "3"

.. ..$ : chr [1:10] "Balance" "Qual\_miles" "cc1\_miles" "cc2\_miles" ...

$ totss : num 39980

$ withinss : num [1:3] 11397 5396 10758

$ tot.withinss: num 27551

$ betweenss : num 12429

$ size : int [1:3] 1340 163 2496

$ iter : int 3

$ ifault : int 0

- attr(\*, "class")= chr "kmeans"

**Need to append the formed cluster to the final dataframe.**

> final <- data.frame(Airlines,fit$cluster)

> head(final)

ID. Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles Bonus\_trans Flight\_miles\_12mo

1 1 28143 0 1 1 1 174 1 0

2 2 19244 0 1 1 1 215 2 0

3 3 41354 0 1 1 1 4123 4 0

4 4 14776 0 1 1 1 500 1 0

5 5 97752 0 4 1 1 43300 26 2077

6 6 16420 0 1 1 1 0 0 0

Flight\_trans\_12 Days\_since\_enroll Award. fit.cluster

1 0 7000 0 3

2 0 6968 0 3

3 0 7034 0 3

4 0 6952 0 3

5 4 6935 1 1

6 0 6942 0 3

**Finalising the dataset by shifting the last column to the first no.**

> final2<-final[,c(ncol(final),1:(ncol(final)-1))]

> head(final2)

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

1 3 1 28143 0 1 1 1 174 1

2 3 2 19244 0 1 1 1 215 2

3 3 3 41354 0 1 1 1 4123 4

4 3 4 14776 0 1 1 1 500 1

5 1 5 97752 0 4 1 1 43300 26

6 3 6 16420 0 1 1 1 0 0

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

1 0 0 7000 0

2 0 0 6968 0

3 0 0 7034 0

4 0 0 6952 0

5 2077 4 6935 1

6 0 0 6942 0

> aggregate(Airlines[,2:11],by= list(fit$cluster),FUN=mean)

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

1 1 116230.13 138.4851 3.697761 1.001493 1.032090 39158.463 18.774627

2 2 196382.98 795.2638 2.171779 1.036810 1.030675 31958.589 27.404908

3 3 42697.48 104.6138 1.172676 1.020032 1.000401 4359.234 6.719151

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

1 319.5276 0.9559701 4836.710

2 5402.3681 15.9141104 4719.601

3 212.7444 0.6482372 3693.763

**Visual Representation:**

> install.packages("kselection")

> library(kselection)

> install.packages("doParallel")

> library(doParallel)

> registerDoParallel(cores = 3)

> k <- kselection(normalized\_data,parallel = F, k\_threshold = 0.95, max\_centers = 14)

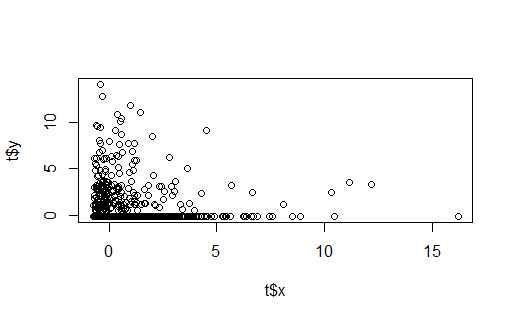
> k

f(k) finds 7 clusters

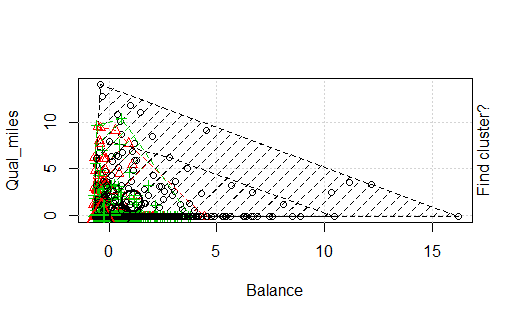
> library(animation)

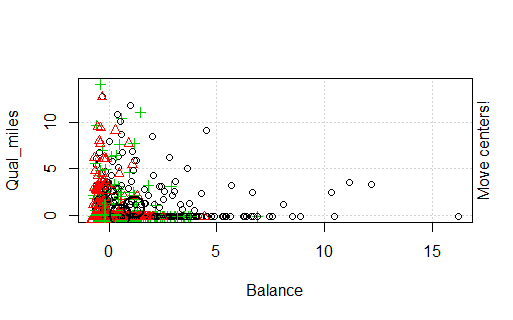
> t<-xy.coords(normalized\_data)

> plot(t)



> km<-kmeans.ani(normalized\_data,4)





**Conclusion : For a large no of records (dataset ) K means clustering is significant over Hierarchical clustering.**