**Sinan Demirhan 2016402330**

**IE 425**

**Homework 3 (due May 21 23:59pm)**

**East-West Airlines is trying to learn more about its customers. Key issues are their flying patterns, earning and use of frequent flyer rewards, and use of the airline credit card. The task is to identify customer segments via clustering. The file EastWestAirlines.xls contains information on 4000 passengers who belong to an ariline’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 charactersitics for the purpose of targeting different segments for different types of mileage offers.**

**a. Apply hierarchical clustering with Euclidean distance and complete linkage. How many clusters appear to be appropriate?**

**library(class)**

**library(CRAN)**

**library(cluster)**

**library(factoextra)**

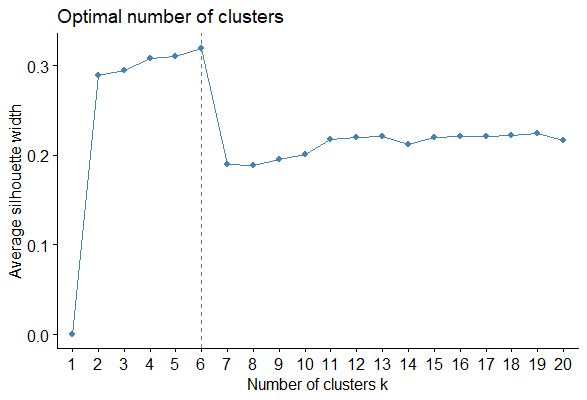
**data=data[,2:12]**

**data\_scaled<-scale(data)**

**d <- dist(data\_scaled, method = "euclidean")**

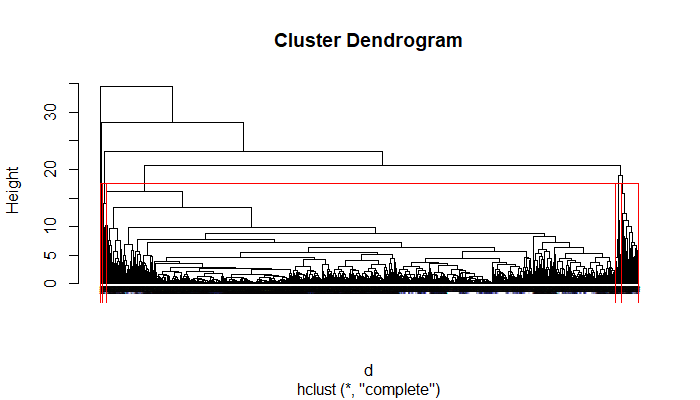
**hc1 <- hclust(d, method = "complete")**

**fviz\_nbclust(data\_scaled,FUN=hcut,method="silhouette",k.max = 20, nboot = 100**)



**plot(hc1, cex = 0.2, hang = -1)**

**rect.hclust(hc1,k=6)**



**clust<-cutree(hc1, k =6)**

To select 6 cluster is appropriate for our data set.

Table cluster

cluster size ave.sil.width

1 1 3782 0.56

2 2 43 0.45

3 3 127 0.16

4 4 28 0.30

5 5 15 0.56

6 6 4 0.42

Table cluster shows the cluster sizes and average silhouette values.

**b. Compare the cluster centroids to characterize the different clusters and try to give each cluster a label.**

**group= sapply(1:6, function(i, dat, clusters){ind = (clusters == i)**

**colMeans(dat[ind,])**

**}, data\_scaled, clust)**

group

[,1] [,2] [,3] [,4] [,5] [,6]

Balance -0.076399952 -0.04688380 0.59397240 7.27191389 0.63963927 0.57948685

Qual\_miles -0.008716739 -0.15621606 0.24604733 0.30904865 -0.08442237 0.26223984

cc1\_miles -0.018835354 -0.66814372 0.41426275 1.09804087 1.02195660 0.31990627

cc2\_miles -0.098229603 9.03712348 -0.09822960 -0.09822960 -0.09822960 -0.09822960

cc3\_miles -0.058695896 -0.06275873 -0.06275873 -0.06275873 15.64434291 -0.06275873

Bonus\_miles -0.065151687 -0.10165261 1.15913274 1.70872967 3.17929372 2.00776234

Bonus\_trans -0.091361246 0.61777389 1.89931065 1.00089286 1.71439935 6.00262786

Flight\_miles\_12mo -0.135315754 0.08753845 3.41819524 0.63157499 0.03328853 13.92645087

Flight\_trans\_12 -0.146379668 0.22031926 3.63386776 1.11609729 0.05968793 12.62173226

Days\_since\_enroll -0.022402820 -0.07245492 0.40250550 1.31581426 0.23984262 -0.92890286

Award. -0.041409194 0.05177741 0.91246469 1.00795819 0.33748514 1.30375512

We can differ clusters 4,5 and 6 from others,because their means in balance, cc3\_miles ,fligth miles and bonus trans are highly different than others.So they can be thought as different class.

**c. To check the stability of the clusters, remove a random 5% of the data (by taking a random sample of 95% of the records, namely 200 records), and repeat the analysis. Does the same picture emerge? Use 425 as the seed.**

**set.seed(425)**

**train=sample(1:3999,200)**

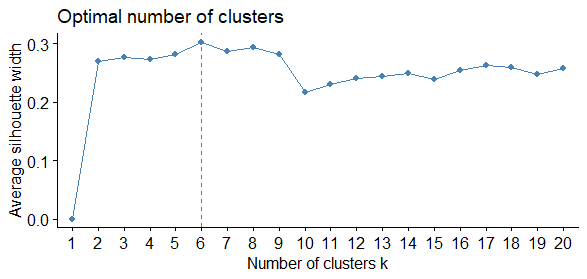
**random\_data=data[train,]**

**scaled\_rand<-scale(random\_data)**

**rand\_dist <- dist(scaled\_rand, method = "euclidean")**

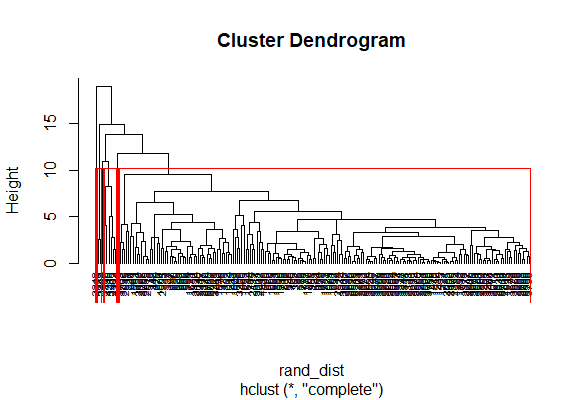
**hc2 <- hclust(rand\_dist, method = "complete" )**

**fviz\_nbclust(scaled\_rand,FUN=hcut,method="silhouette",k.max = 20, nboot = 100)**



**plot(hc2, cex = 0.6, hang = -1)**

**rect.hclust(hc2,k=6)**



There is no significant difference between the whole data and the 5% randomed data

**d. Use k-means algorithm with the number of clusters you found in part (a). Does the same picture emerge?**

**kmeans6<-kmeans(data\_scaled,centers =6,nstart = 120 )**

**table(kmeans6$cluster)** #the clusters in kmeans algorithm

1 2 3 4 5 6

146 15 43 2075 904 816

**table(clust)** #the clusters in hierarhical algorithm

clust

1 2 3 4 5 6

3782 43 127 28 15 4

The values are very different.So we can not say that they are similar.

**e. Which clusters would you target for offers, and what type of offers would you target to customers in that cluster?**

I would target the fourth cluster because their balance are high.Because their balance are high I would make them some offers in business class and I would also make some offers to the other passengers in the economic class because their balance are low.