IE425
SPRING 2023

HOMEWORK 3
Hierarchical Clustering & K-Means

FATMANUR YAMAN - 2019402204
MURAT TUTAR - 2020402264

Question a)

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

- **Preprocessing: THE CODE**

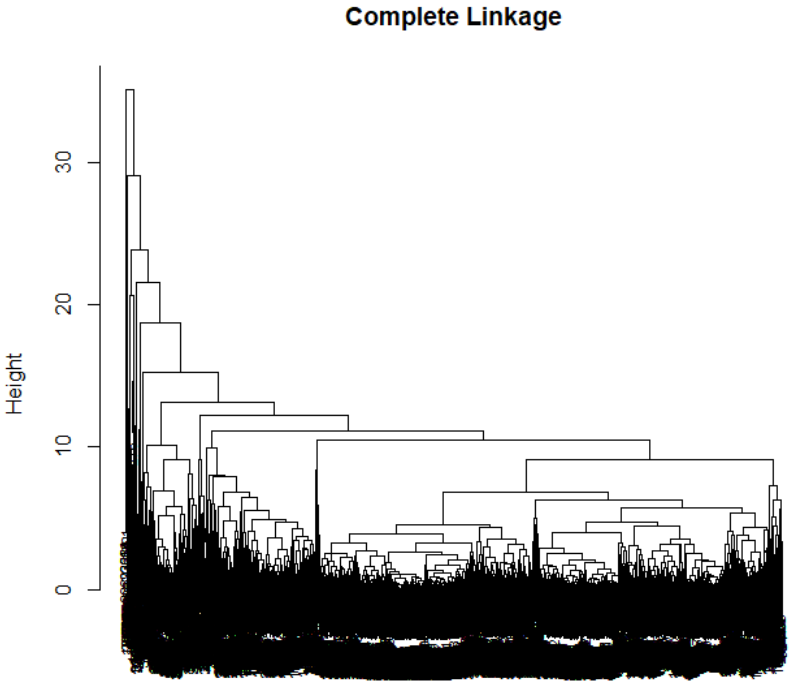
First of all, some preprocessing should be done to convert the categorical attributes (number of miles earned) to the numerical values. I think 1, 2, 3, 4, and 5 can also be used, but they are not reflecting the values they should be. Thus, firstly, these values should be changed. Then, scaling should be done because the values range in different values.

```
raw_data <- read.xlsx("C:\\Users\\fatma\\Desktop\\IE425_HW3\\EastWestAirlines.xlsx", sheet = 3)
raw_data
sum(is.na(raw_data))
str(raw_data)

#Some Preprocessing In The Data
#Convert the categorical attribute to the numerical value.
raw_data$cc1_miles <- ifelse(raw_data$cc1_miles == 1, 2500,
                      ifelse(raw_data$cc1_miles == 2, 7500,
                             ifelse(raw_data$cc1_miles == 3, 17500,
                                    ifelse(raw_data$cc1_miles == 4, 37500,
                                           ifelse(raw_data$cc1_miles == 5, 50000, raw_data$cc1_miles)))))
raw_data$cc2_miles <- ifelse(raw_data$cc2_miles == 1, 2500,
                      ifelse(raw_data$cc2_miles == 2, 7500,
                             ifelse(raw_data$cc2_miles == 3, 17500,
                                    ifelse(raw_data$cc2_miles == 4, 37500,
                                           ifelse(raw_data$cc2_miles == 5, 50000, raw_data$cc2_miles)))))
raw_data$cc3_miles <- ifelse(raw_data$cc3_miles == 1, 2500,
                      ifelse(raw_data$cc3_miles == 2, 7500,
                             ifelse(raw_data$cc3_miles == 3, 17500,
                                    ifelse(raw_data$cc3_miles == 4, 37500,
                                           ifelse(raw_data$cc3_miles == 5, 50000, raw_data$cc3_miles)))))


str(raw_data)
#Scaling
train_scaled=scale(raw_data)
```

- **Hierarchical Clustering: THE CODE**

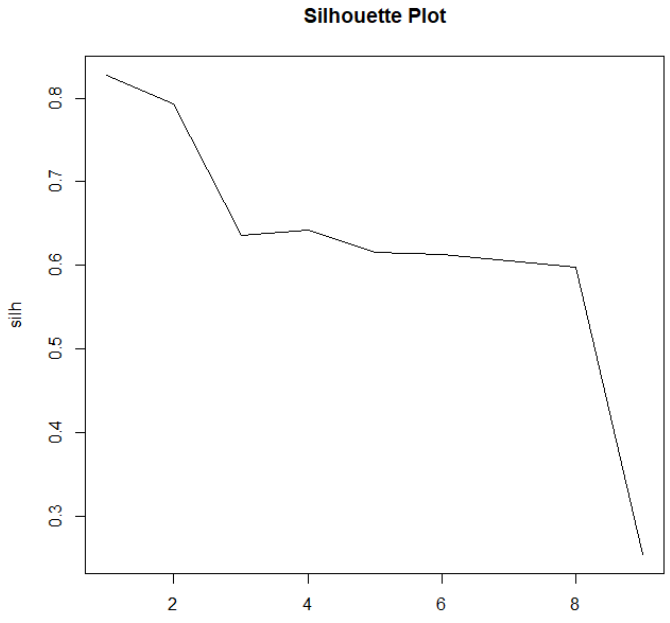
```
#1. Hierarchical Clustering
hc_complete=hclust(dist(train_scaled), method="complete")
#Plot the tree.
plot(hc_complete,main="Complete Linkage", xlab="", cex=.7)

#Finding the Best NUmber of Clusters With Silhouette Index
silh=c()
for (k in 2:10){
  kume=cutree(hc_complete,k=k)
  X_sil=silhouette(kume, dist(train_scaled))
  silh[k-1]=mean(X_sil[,3])
}
data.frame(k=2:10,silh)
plot(silh, main = "Silhouette Plot", type="l")
#Choose the number with maximum Silhouette value.
best_k <- which.max(silh)+1
cat("The best number of clusters is: " , best_k)
#Cut the tree from the index of the maximum Silhouette value.
clusters <- cutree(hc_complete, k = best_k)
#See the distribution of classes.
table(clusters)
```

- **Hierarchical Clustering: THE OUTPUT**



k	silh
2	0.8271479
3	0.7933231
4	0.6357979
5	0.6417347
6	0.6150863
7	0.6125680
8	0.6059403
9	0.5982916
10	0.2543520

The k vs. silhouette value Table



- **Hierarchical Clustering: THE COMMENT**

A hierarchical clustering with Euclidean distance and complete linkage was applied to the dataset. Then, to determine the number of clusters, the silhouette index was used. **The maximum Silhouette value was obtained in the 2nd index. Thus, the appropriate number of clusters is 2.**

In addition to this, when we look at the number of instances in each class, it was found that there are only 2 values in Class 2. These values can be outliers. Actually, we removed them and tried again to build a hierarchical clustering tree, but the number of instances in Class 2 increased to 11. It is such a small increase. Thus, there should be something wrong with the hierarchical clustering in this dataset.

Question b)

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

- **Cluster Centroids: THE CODE**

```
#b. Finding the centroids of the clusters.
clusters <- cutree(hc_complete, k = best_k)
centroids <- aggregate(train_scaled, by = list(clusters), FUN = mean)
centroids
```

- **Cluster Centroids: THE OUTPUT**

```
  Group.1            ID       Balance    Qual_miles     cc1_miles     cc2_miles     cc3_miles
1       1 -0.0003466613 -0.0001642088 -0.0003556443 -0.0002179852  4.542317e-05  2.957415e-05
2       2  0.6928025826  0.3281712151  0.7107550699  0.4356434039 -9.077821e-02 -5.910394e-02

  Bonus_miles  Bonus_trans Flight_miles_12mo Flight_trans_12 Days_since_enroll         Award
-0.001226132 -0.003329203       -0.00929074    -0.006282624      0.0003674573 -0.0006523668
 2.450425764  6.653411228       18.56754317    12.555824369     -0.7343634811  1.3037551185
```

- **Cluster Centroids: THE COMMENT**

As can be seen from the output table above, the center of the “Award”, “Bonus_miles” , and “Flight_trans_12” (number of flights made in the previous 12 months) indicate that the Class 2 makes more flights than Class 1. People in Class 2 fly often, collect bonuses more, and get awards more. **Thus, Class 2 can be seen as “Business&Premium Class” who flies regularly, and Class 1 can be seen as “Economy Class” who chooses planes as transportation if there is no other choice.** However, as I said above, the number of instances in Class 2 is too low to make a conclusion.

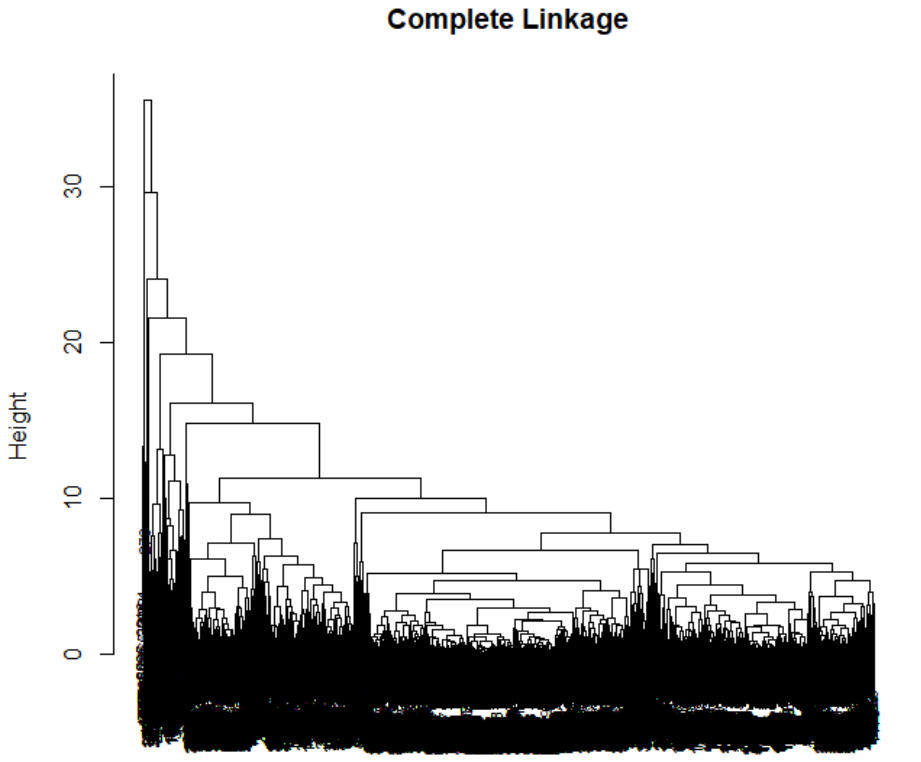
Question c)

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.

- **Removed Data: THE CODE**

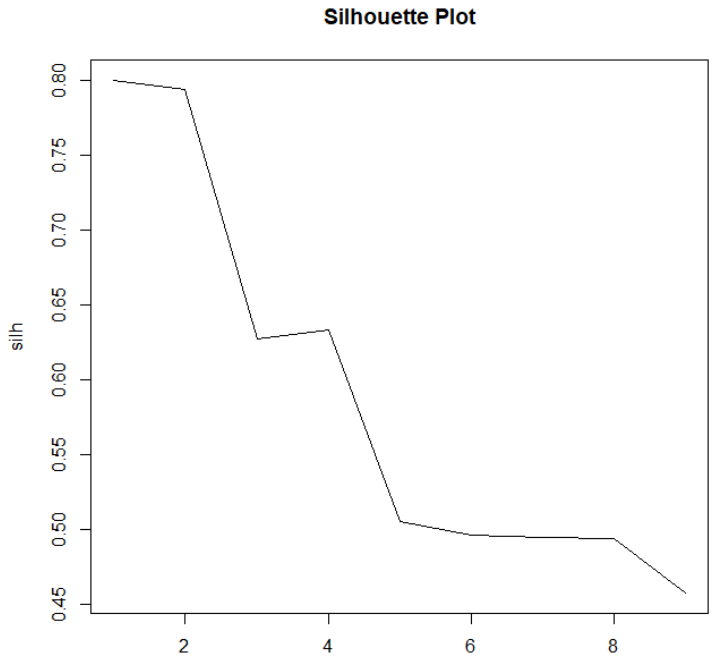
```r
#C. Remove Data Randomly
set.seed(425)
index_to_remove <- sample(1:nrow(raw_data),200)
train_removed <- raw_data[-index_to_remove,]
nrow(train_removed)

#Scaling
train_scaled <- scale(train_removed)
#Hierarchical Clustering
hc_complete=hclust(dist(train_scaled), method="complete")
plot(hc_complete,main="Complete Linkage", xlab="", cex=.7)

#Finding the Best NUmber of Clusters With Silhouette Index
silh=c()
for (k in 2:10){
  kume=cutree(hc_complete,k=k)
  x_sil=silhouette(kume, dist(train_scaled))
  silh[k-1]=mean(x_sil[,3])
}
data.frame(k=2:10,silh)
plot(silh, main = "Silhouette Plot", type="l")
#Find the Best Number of Clusters.
best_k <- which.max(silh)+1
cat("The best number of clusters is: " , best_k)
clusters <- cutree(hc_complete, k = best_k)
table(clusters)
#Find the center of the centroids.
clusters <- cutree(hc_complete, k = best_k)
centroids <- aggregate(data_removed, by = list(clusters), FUN = mean)
centroids
```

- **Removed Data: THE OUTPUT**



k	silh
2	0.8000186
3	0.7938714
4	0.6274477
5	0.6331417
6	0.5054293
7	0.4961736
8	0.4946986
9	0.4939958
10	0.4577711

The k vs. Silhouette Value Table



Group.1	ID	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles
1	1	-0.0003466613	-0.0001642088	-0.0003556443	-0.0002179852	4.542317e-05	2.957415e-05
2	2	0.6928025826	0.3281712151	0.7107550699	0.4356434039	-9.077821e-02	-5.910394e-02

Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award
-0.001226132	-0.003329203	-0.00929074	-0.006282624	0.0003674573	-0.0006523668
2.450425764	6.653411228	18.56754317	12.555824369	-0.7343634811	1.3037551185

- **Removed Data: THE COMMENT**

The 200 rows of data were removed from the original dataset by selecting randomly. But, the situation is nearly the same as the non-removed dataset. **The appropriate number of classes are 2 as before, and also the centers of centroids are very close to the previous results.** Class 2 is “Business & Premium Class” and Class 1 is the people who fly rarely.

Question d)

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

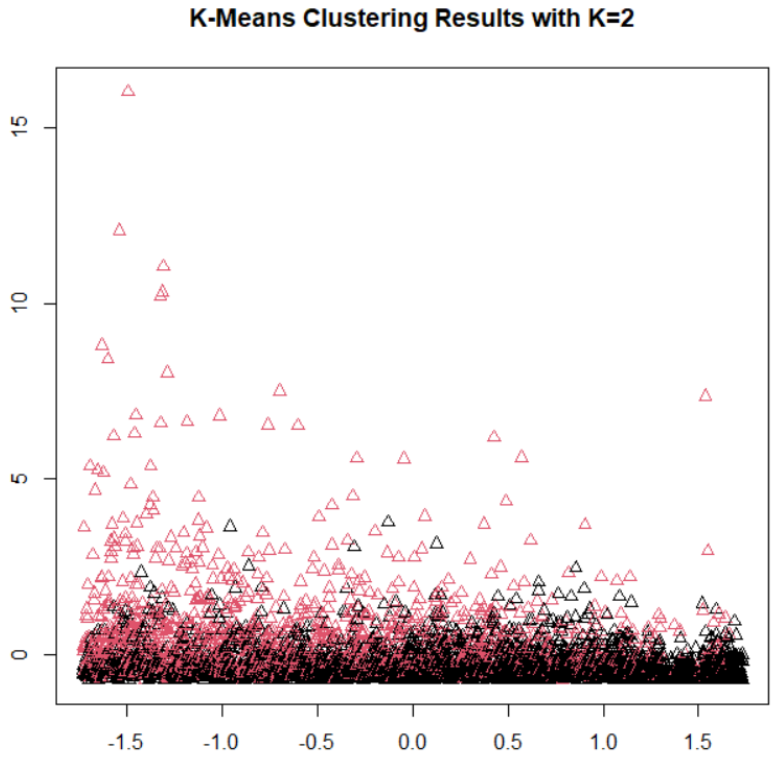
- **KMEANS: THE CODE**

```
#d. KMeans
km <- kmeans(x=train_scaled, 2, nstart=20)
km
km$cluster
km$tot.withinss
km$size
plot(train_scaled, col=(km$cluster), main="K-Means Clustering Results with K=2", xlab="", ylab="", pch=2, cex=1)
#Find The Centroids of The Classes
centroids <- aggregate(train_scaled, by = list(km$cluster), FUN = mean)
centroids
```

- **KMEANS: THE OUTPUT**

```
> km$size
[1] 2666 1133
```

The Sizes Of The Classes




```
  Group.1         ID    Balance  Qual_miles  cc1_miles   cc2_miles   cc3_miles
1       1  0.2184608 -0.2840014 -0.07748897 -0.5059966  0.01924069 -0.05156798
2       2 -0.5140482  0.6682680  0.18233504  1.1906329 -0.04527421  0.12134177

Bonus_miles Bonus_trans Flight_miles_12mo Flight_trans_12 Days_since_enroll      Award
 -0.4586389  -0.3932173        -0.1723195      -0.1944674       -0.2226379 -0.2932744
  1.0791979   0.9252580         0.4054756       0.4575905        0.5238771  0.6900878
```

- **KMEANS: THE COMMENT**

It can be seen from the class centroid table above, the K-Means algorithm also suggests to divide the classes as it was divided previously (Class 2: regular fliers, Class 1: sometime-users). **The people in Class 2 fly more, spend more, use awards miles more, and earn more bonuses.** But, it diverges at one point. **It provides more balanced classes according to the number of instances in each class.** There are 2666 instances in Class 1 and 1133 instances in Class 2.

Question e)

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

Class 1: These people should be encoıarged to fly more often with discounts and promotions. **I would prefer to target these people for offers, because Class2 will fly in any case and spend their money.** And also, why these people do not prefer flying should be examined to suggest powerful solutions. In addition to these, gaining a new loyal customer will be the best part of targeting these people.

Class 2: These are the people who already fly often and visit a lot of places. These people love feeling precious and luxurious things. Offering new experiences to these people like having a shower in the plane or tasting meals from different cuisines. These people will fly in any case, the important point is being the brand that these people always prefer by providing them indispensable experiences on the air.