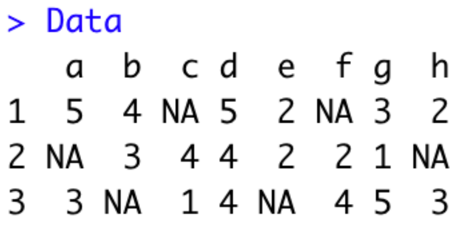
Name : Manasa Challa
UB Person ID : 50356441

REPORT HOMEWORK – 3

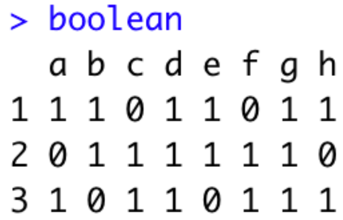
Question1:

Created the table :

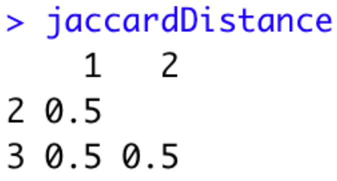


```
> Data
   a  b  c d  e  f g  h
1  5  4 NA 5  2 NA 3  2
2 NA  3  4 4  2  2 1 NA
3  3 NA  1 4 NA  4 5  3
```

a.) Treating the utility matrix as Boolean, compute the Jaccard distance between each pair of users.

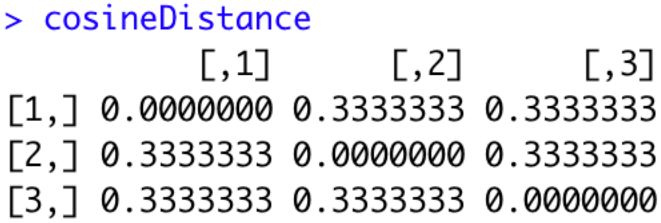


```
> boolean
  a b c d e f g h
1 1 1 0 1 1 0 1 1
2 0 1 1 1 1 1 1 0
3 1 0 1 1 0 1 1 1
```



```
> jaccardDistance
    1   2
2 0.5    
3 0.5 0.5
```

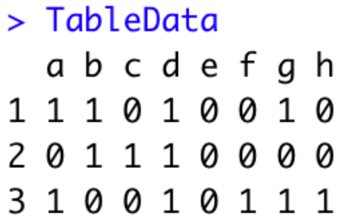
b.) Repeat Part A, but use the cosine distance.



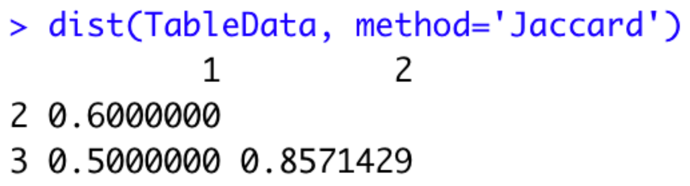
```
> cosineDistance
          [,1]      [,2]      [,3]
[1,] 0.0000000 0.3333333 0.3333333
[2,] 0.3333333 0.0000000 0.3333333
[3,] 0.3333333 0.3333333 0.0000000
```

c.) Treat ratings of 3, 4, and 5 as 1, and ratings 1, 2, and blank as zero. Compute the Jaccard distance between each pair of users.

After changing 3, 4, and 5 as 1, and 1, 2, and blank as zero, our table looks like:

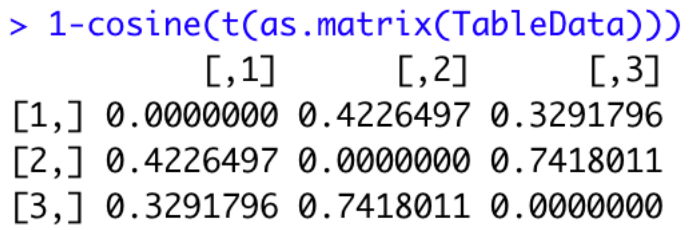


```
> TableData
  a b c d e f g h
1 1 1 0 1 0 0 1 0
2 0 1 1 1 0 0 0 0
3 1 0 0 1 0 1 1 1
```



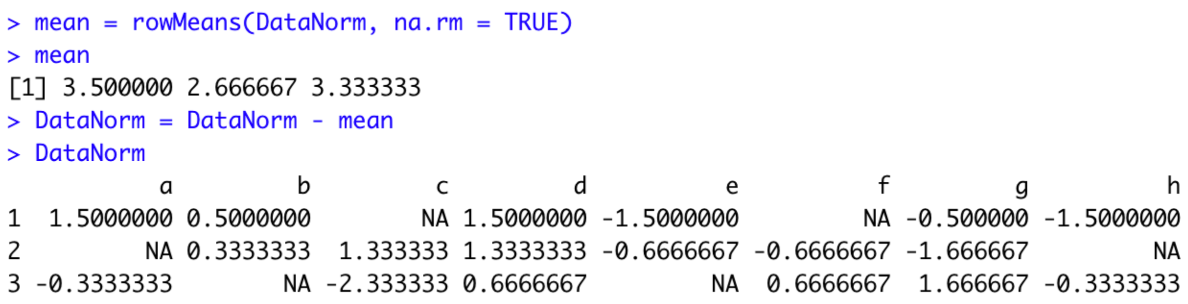
```
> dist(TableData, method='Jaccard')
          1         2
2 0.6000000          
3 0.5000000 0.8571429
```

d.) Repeat Part C, but use the cosine distance.



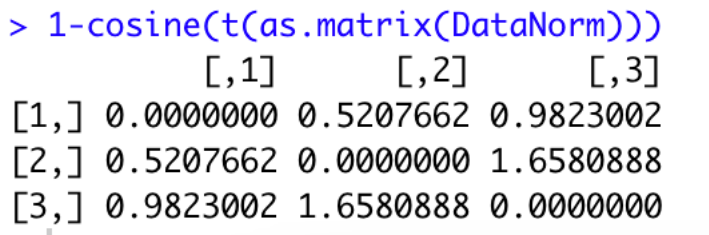
```
> 1-cosine(t(as.matrix(TableData)))
          [,1]      [,2]      [,3]
[1,] 0.0000000 0.4226497 0.3291796
[2,] 0.4226497 0.0000000 0.7418011
[3,] 0.3291796 0.7418011 0.0000000
```

e.) Normalize the matrix by subtracting from each nonblank entry the average value for its user.



```
> mean = rowMeans(DataNorm, na.rm = TRUE)
> mean
[1] 3.500000 2.666667 3.333333
> DataNorm = DataNorm - mean
> DataNorm
           a         b         c         d          e          f         g          h
1  1.5000000 0.5000000        NA 1.5000000 -1.5000000         NA -0.500000 -1.5000000
2         NA 0.3333333  1.333333 1.3333333 -0.6666667 -0.6666667 -1.666667         NA
3 -0.3333333        NA -2.333333 0.6666667         NA  0.6666667  1.666667 -0.3333333
```

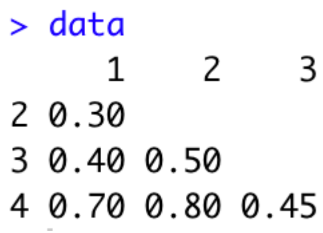
f.) Using the normalized matrix from Part E, compute the cosine distance between each pair of users.



```
> 1-cosine(t(as.matrix(DataNorm)))
          [,1]      [,2]      [,3]
[1,] 0.0000000 0.5207662 0.9823002
[2,] 0.5207662 0.0000000 1.6580888
[3,] 0.9823002 1.6580888 0.0000000
```

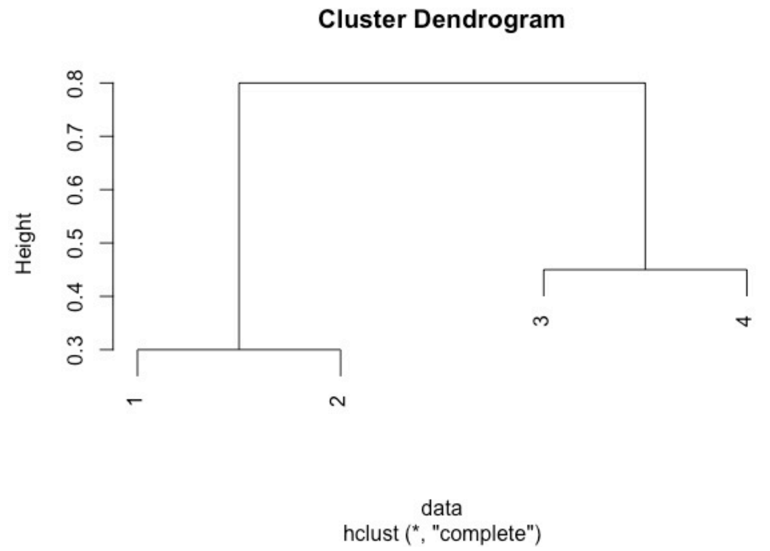
Question 2:

a.) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage.



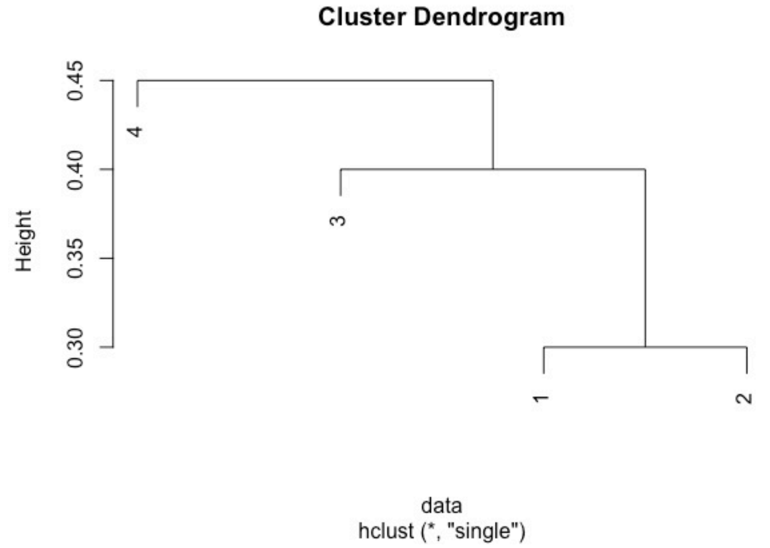
```
> data
     1    2    3
2 0.30          
3 0.40 0.50     
4 0.70 0.80 0.45
```

Name : Manasa Challa
UB Person ID : 50356441



At the height of 0.3 and 0.45, it can be observed that (1,2) and (3,4) form 2 clusters respectively. A bigger cluster is seen at height a of 0.8.

b.) Repeat (a), this time using simple linkage clustering.

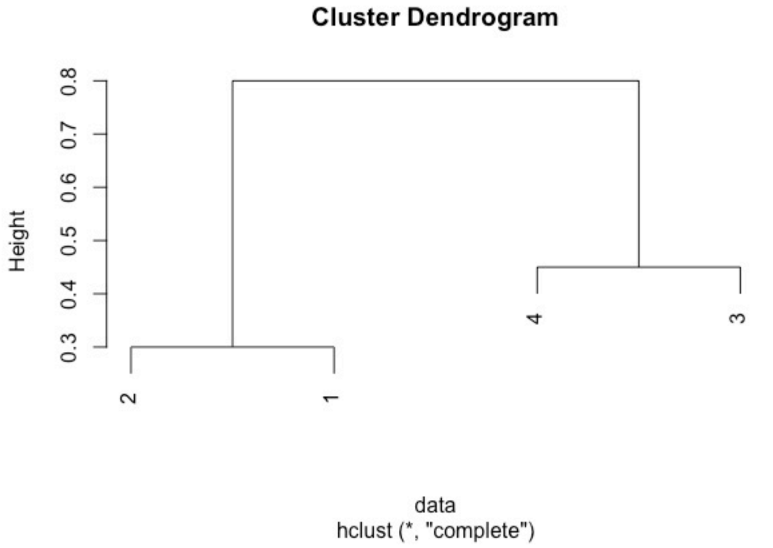


At the height of 0.3 it can be observed that ((1,2),3) and 4 form 3 clusters.

c.) Suppose that we cut the dendrogram obtained in (a) such that two clusters result. Which observations are in each cluster ?
#Ans : We obtain clusters (1,2) and (3,4).

d.) Suppose that we cut the dendrogram obtained in (b) such that two clusters result. Which observations are in each cluster ?
#Ans : We obtain clusters ((1,2),3) and (4).

e.) Draw a dendrogram that is equivalent to the dendrogram in (a), for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.

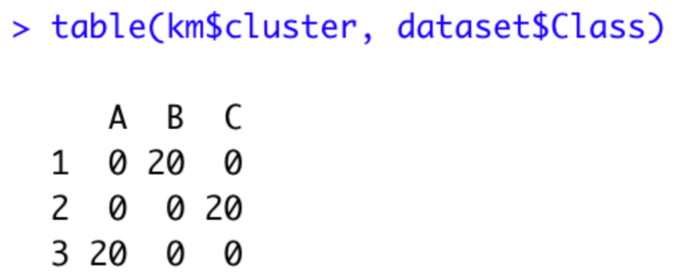


Question 3:

a.) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total) and 50 variables.

20 observations were generated in each of the three classes with mean as 5,15,25 respectively and with standard deviation as 2 in each of them.

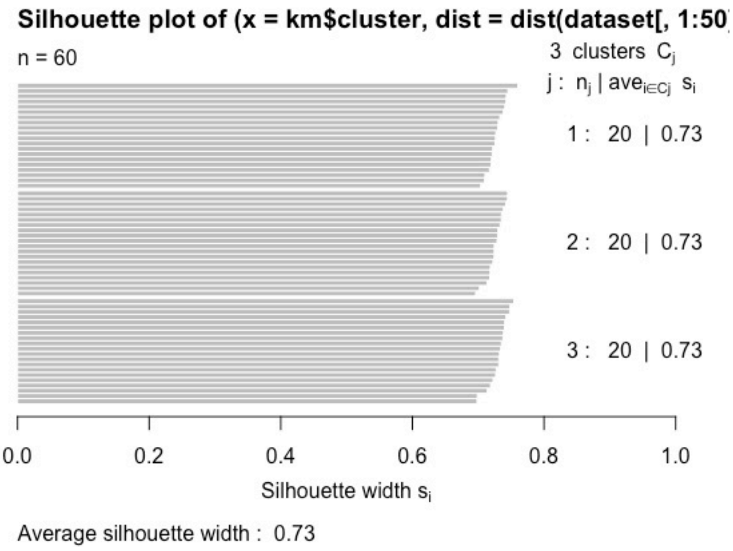
b.) Perform k-means clustering of the observations with K=3. Using the rand index and adjusted rand index, assess how well do the clusters that you obtained in K-means clustering compare to the true labels?



```
> table(km$cluster, dataset$Class)

     A  B  C
  1  0 20  0
  2  0  0 20
  3 20  0  0
```

After calculating the rand index and adjusting it, they turn out to be one that means clustering is same as the original classes.



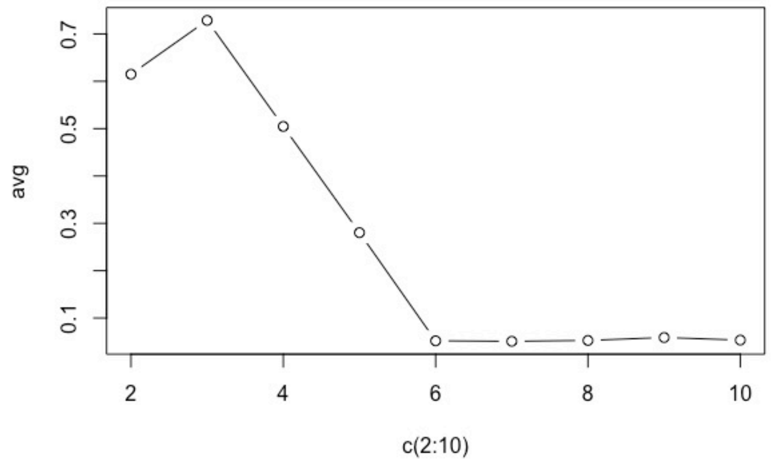
Observations:

1.Average silhouette width = 0.73

2.Signifies that data we have chosen with avg means 5,15,25 are well separated in their clusters

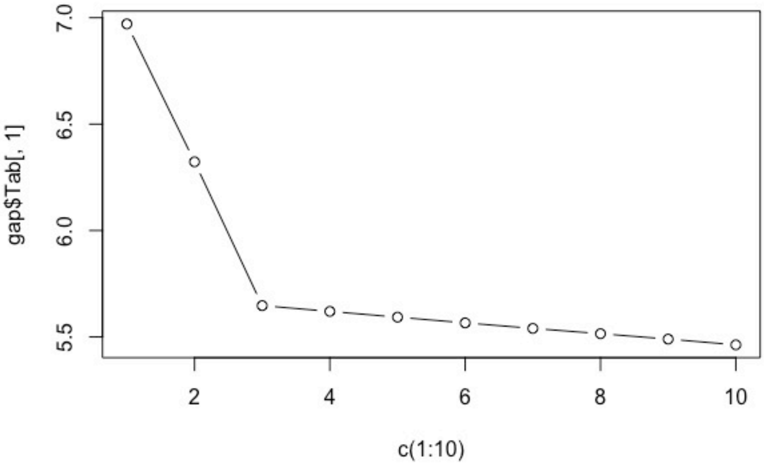
c.) Using silhouette plots, select the optimal number of clusters.

Optimal value of k found to be 3
Largest silhouette distance > 0.7



d.) Using the gap statistics, select the optimal number of clusters.

Name : Manasa Challa
UB Person ID : 50356441



From the figure it can be seen that a elbow formation is found at k=3.