Assignment 2

The multi_kmeans function records the local optima found in multiple kmeans analyses. Study the function below, run the Hartigan's data and the iris data. Review the output. Answer the questions where indicated in the program below:

```
##reducing the number of starts here to 2, try several print functions to understand what the function does at each step ###
multi_kmeans=function(data,k,nstarts=2,alg=c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen")){

for (i in 1:nstarts) {

        kout=kmeans(data,k,iter.max=10000,algorithm=alg)
        sse=round(sum(kout$withinss),digits=5)
```

###Question1 What is kout and sse? ####

```
        if (i == 1) {
        freqmat=rbind(c(sse,1))
        bestcluster=kout$cluster
        bestcenters=kout$centers
        bestsize=kout$size } else {
```

#####Question 2 What is being recorded here where i==1? freqmat=? bestcluster=? Bestcenter=? bestsize=? and below, freqmat[,1]=??? #####

```
        if (sse < min(freqmat[,1])) {
                bestcluster=kout$cluster
                bestcenters=kout$centers
                bestsize = kout$size
        }
        foundit=0
```

####Question 3 – what is dim(freqmat)[1]? freqmat[j,1] = ? freqmat[j,2]=? what happens with break?

```
        for (j in 1:(dim(freqmat)[1])){
                if (freqmat[j,1]==sse) {
                        freqmat[j,2]=freqmat[j,2]+1
                        foundit=1
```

```
                        break
                        }
                }
```

###Question 4 – describe what the next section of code does, including the variables, the for loop. Describe each line of code.

```
                if(foundit==0){
                        freqmat=rbind(freqmat,c(sse,1))
                        torder=order(freqmat[,1])
                        freqmat2=freqmat
                        for (j in 1:(dim(freqmat)[1])){
                                freqmat[j,]=freqmat2[torder[j],]
                                }
                }
        }

}
multi_kmeans=list(ssefreq=freqmat,bestcluster=bestcluster,bestcenters=bestcenters,bestsize=
bestsize)
}

#reading in Hartigan's data

energy=c(11,8,13,12,6,4,5,5)
protein=c(29,30,21,27,31,29,36,37)
calcium=c(1,1,1,1,2,1,1,1)
data=cbind(energy,protein,calcium)

set.seed(99999)
try1=multi_kmeans(data,3) #defaults to Hartigan-Wong algorithm - Fewer Local Optima
try1
try2=multi_kmeans(data,3,alg=c("MacQueen")) #running MacQueen's algorithm - more Local
Optima
try2

data(iris)  ####Read in iris data and look at results with a larger data set#####
 data=iris[,1:4]
 set.seed(1234)
```

```
try1=multi_kmeans(data,3) #defaults to Hartigan-Wong algorithm - Fewer Local Optima
try2=multi_kmeans(data,3,alg=c("MacQueen")) #running MacQueen's algorithm - more Local Optima

try1
try2
```

####Question 5 – Describe the resulting output in try1 and try2. What is the difference in try1 and 2? What is stored in ssefreq[,1] and ssefreq[,2]?

####Question 6 - bump up nstarts to 1000, and then describe the output you get after try1 and try2. Is there any difference when you adjust nstarts?

```
plot(data, col = try1$bestcluster)
points(try1$centers, col = 1:3, pch = 8)
```

Read the pdf file in Blackboard on Ward's method.

The data used for the next portion of the assignment is the Doubs river fish communities data from the PhD thesis of Verneaux (1973). See this website for a description:
http://www.davidzeleny.net/anadat-r/doku.php/en:data:doubs

Question 6 - see below - include your code and output and answers to any questions

Read in the data from 'http://www.davidzeleny.net/anadat-r/data-download/DoubsSpe.csv'####
Look at the first 6 lines of the data input to verify it read correctly###
###Run descriptive statistics using apply to find the min, median, mean, sd, max, and save your results as a data frame, rounding to 1 decimal###
###Print out your data frame and look at the values of your statistics#####
###Based on the values of your descriptive statistics, would you recommend standardizing the data???###

set.seed (99999)

Question 7 – Compute a dissimilarity matrix using the Euclidean distance, then perform single linkage agglomerative clustering. Plot a dendogram of your result. Include your code and results. Based on this result, how would you describe the data set in terms of number of clusters? Do you notice any chaining? Chaining is when a pair is linked to a third object, which is in turn linked to another and so forth.

Question 8 – Using your dissimilarity matrix from Q7 perform complete linkage agglomerative clustering. Plot a dendogram of your result. Include your code and results. Based on this result, how would you describe the data set in terms of number of clusters? Given the data consists of sites along a river (with the numbers following the stream), does this result place closely similar sites in the same groups? Why would 2 perfectly valid clustering methods produce such different results?

Question 9 – repeat Q8 using the average or UPGMA and ward d2 method. Show your code and results. Which of the 4 methods do you feel is the best method and why?

Question 10 – Compute Agglomerative Hierarchical clustering using AGNES. ?agnes in R. The advantage of agnes is it prints the agglomerative coefficient which measures the clustering structure of the data for measures: ward, single, complete, and average. Print your ac values for all 4 methods. Which method has the highest clustering structure? Plot your tree.

Question 11 - Using cutree, look at agnes's wards method with 2 and with 4 clusters. Print out the cluster membership numbers for each of the 2 and 4 solutions (use table). Do both solutions have good distributions, meaning not many clusters with few observations? Run the aggregate code below. Comparing the medians across both the 2 and 4 group solutions, what are your observations, in general?

```
aggregate(data,list(groups.2),median)
aggregate(data,list(group.4),median)
```