Title: Assignment #2: (*Analysis of PCA*)

Purpose: *This assignment is to study the PCA technique when it is applied on Kmeans and Hierarchical clustering. To determine if doing PCA actually helps us in improving the results.*

Dataset(s): *The dataset used the Wine Quality white csv dataset present in the UCI library.*

Link to the dataset: [Wine_Quality_White]

Approach:

- *First, the dataset is loaded, checked if there are any NA's and omitted if any. (This dataset had none)*
- *The dataset had the "quality" classification to be ranged from 3-9 and since there was a suggestion to move it to (0-2), (3-5), (6-8), (9-10) – I classified the dataset to 3 sets (since there wasn't any (0-2)). So the "quality" classification was moved to be 3 clusters.*
- *Performed the normal Kmeans clustering on the dataset after scaling and removing the last column out. Took the K value to be 3 and checked the results.*

```
             1    2    3
Excellent    0    3    2
Good       967 1278 1008
Not Good   839  346  455
```

- *Applied the PCA on the dataset and tried to find the variance and plotted the PCA transformation along the variance to determine which value would be an ideal component number. I took the components to be five, since I felt it sort of balanced between 4-6.*
- *Next I applied K-means clustering to the dataset transformed under PCA, and found that the betweenSS/TotalSS actually increased from 43% to 54%.*
- *This is when the data is scaled.*
- *When the data is not scaled, and used as such the PCA and the clustering values differ drastically.*

```
table(winequality_white$quality,wine_kmeans$cluster)

             1    2    3
Excellent    3    2    0
Good      1358 1329  566
Not Good   616  467  557
```

```
table(winequality_white$quality,km_wine_pca$cluster)

             1    2    3
Excellent    0    2    3
Good       566 1328 1359
Not Good   559  466  615
```

- *The between and totss values actually increased from 73% to 80%.*
- *Using just one Principal Component provided the clustering results to be more accurate than the one in the table mentioned.*

Summary:

PCA on this dataset did not help much. It didn't significantly help in improving the dataset before scaling. PCA is extremely sensitive to scaling. I am not sure how it is sensitive to scaling, but the same dataset without scaling showed significantly improved results for the clustering than the dataset which was scaled. KMeans also is sensitive to scaling.

Appendix :

This includes the R code that I have done to obtain the analysis for the PCA:

```
winequality_white <- read_delim("~/Downloads/Education/Spring 17/R for Data
Scientists/R_CSV/winequality-white.csv", ";", escape_double = FALSE, trim_ws = TRUE)
#remove the last item
winequality_white$quality[winequality_white$quality>=9] <- "Excellent"
winequality_white$quality[winequality_white$quality==8] <- "Good"
winequality_white$quality[winequality_white$quality==7] <- "Good"
winequality_white$quality[winequality_white$quality==6] <- "Good"
winequality_white$quality[winequality_white$quality==5] <- "Not Good"
winequality_white$quality[winequality_white$quality==4] <- "Not Good"
winequality_white$quality[winequality_white$quality==3] <- "Not Good"
winequality_white$quality[winequality_white$quality==2] <- "Bad"
winequality_white$quality[winequality_white$quality==1] <- "Bad"
winequality_white$quality[winequality_white$quality==0] <- "Bad"

#change the scale and remove the last item
new_wine <- winequality_white
new_wine$quality = NULL
#new_wine <- scale(new_wine) –This line when uncommented gives slightly bad results due
to sensitivity for scaling.

#run the normal kmeans - since there was no bad wine, i used only three samples.
wine_kmeans <- kmeans(new_wine,3)
table(winequality_white$quality,wine_kmeans$cluster)
wine_kmeans$betweenss/wine_kmeans$totss
#now apply PCA to the components
wine_pca <- prcomp(new_wine)
biplot(wine_pca,cex=c(1/3,1/2), scale=0)
```

```
wine_pca.var =wine_pca$sdev ^2
pve=wine_pca.var/sum(wine_pca.var)
plot(cumsum (pve ), xlab=" Principal Component ", ylab ="
   Cumulative Proportion of Variance Explained ", ylim=c(0,1) ,
   type='b')

#after principal component and variance - 4-6 anything should do good - i am taking 5
km_wine_pca = kmeans(wine_pca$x[,1:1],3) #1 PC is also sufficient for classification, since that provided a betwenss/totss value to be 80.9% compared to 2PC's of 74%
table(winequality_white$quality,km_wine_pca$cluster)
km_wine_pca$betweenss/km_wine_pca$totss
```

For H Clustering, the results seems to be improving as well using PCA, but preventing the data to be scaled was useful in both the scenarios.

```
winequality_white <- read_delim("~/Downloads/Education/Spring 17/R for Data Scientists/R_CSV/winequality-white.csv", ";", escape_double = FALSE, trim_ws = TRUE)
#remove the last item
winequality_white$quality[winequality_white$quality>=9] <- "Excellent"
winequality_white$quality[winequality_white$quality==8] <- "Good"
winequality_white$quality[winequality_white$quality==7] <- "Good"
winequality_white$quality[winequality_white$quality==6] <- "Good"
winequality_white$quality[winequality_white$quality==5] <- "Not Good"
winequality_white$quality[winequality_white$quality==4] <- "Not Good"
winequality_white$quality[winequality_white$quality==3] <- "Not Good"
winequality_white$quality[winequality_white$quality==2] <- "Bad"
winequality_white$quality[winequality_white$quality==1] <- "Bad"
winequality_white$quality[winequality_white$quality==0] <- "Bad"

#change the scale and remove the last item
winequality_white <- read_delim("~/Downloads/Education/Spring 17/R for Data Scientists/R_CSV/winequality-white.csv", ";", escape_double = FALSE, trim_ws = TRUE)
new_wine <- winequality_white
new_wine$quality = NULL
new_wine <- scale(new_wine)

#now we do the Hierarchical clustering
new_wine.complete =  hclust(dist(new_wine),method="complete")

#plot for comparison to know where to cut
par(mfrow=c(1,1))
plot(new_wine.complete,main="Complete Linkage", xlab="", sub="", cex=.9)
```

```
table(cutree(new_wine.complete,3),winequality_white$quality)

#now use the same PCA obtained for the previous method and see if it improves the results
pca_wine.complete =  hclust(dist(wine_pca$x[,1:1]),method="complete")
par(mfrow=c(1,1))
plot(pca_wine.complete,main="Complete Linkage", xlab="", sub="", cex=.9)

table(cutree(pca_wine.complete,3),winequality_white$quality)
```