Assignment_1 Statistical Machine Learning

Manikandan Ravichandran 19200314

26/02/2020

Loading Data Set

We will be removing the Categorical Data (Non Numercial Data) to perform K-Means to perform a efficient. But prior to clustering we would be required to scale the data into a standardized data.

```
spotify1 <- read.csv('data_spotify_songs.csv') # Loading Dataset from the Working Directory
spotify=spotify1[,-1:-3] # Removing Categorical Values to perform Clustering using K means
#summary(spotify)
```

Scaling the Dataset:

We perform the Scaling of the Data set ,as Standardization transforms the relative weight of each variable equal by converting each variable to a relative distance.

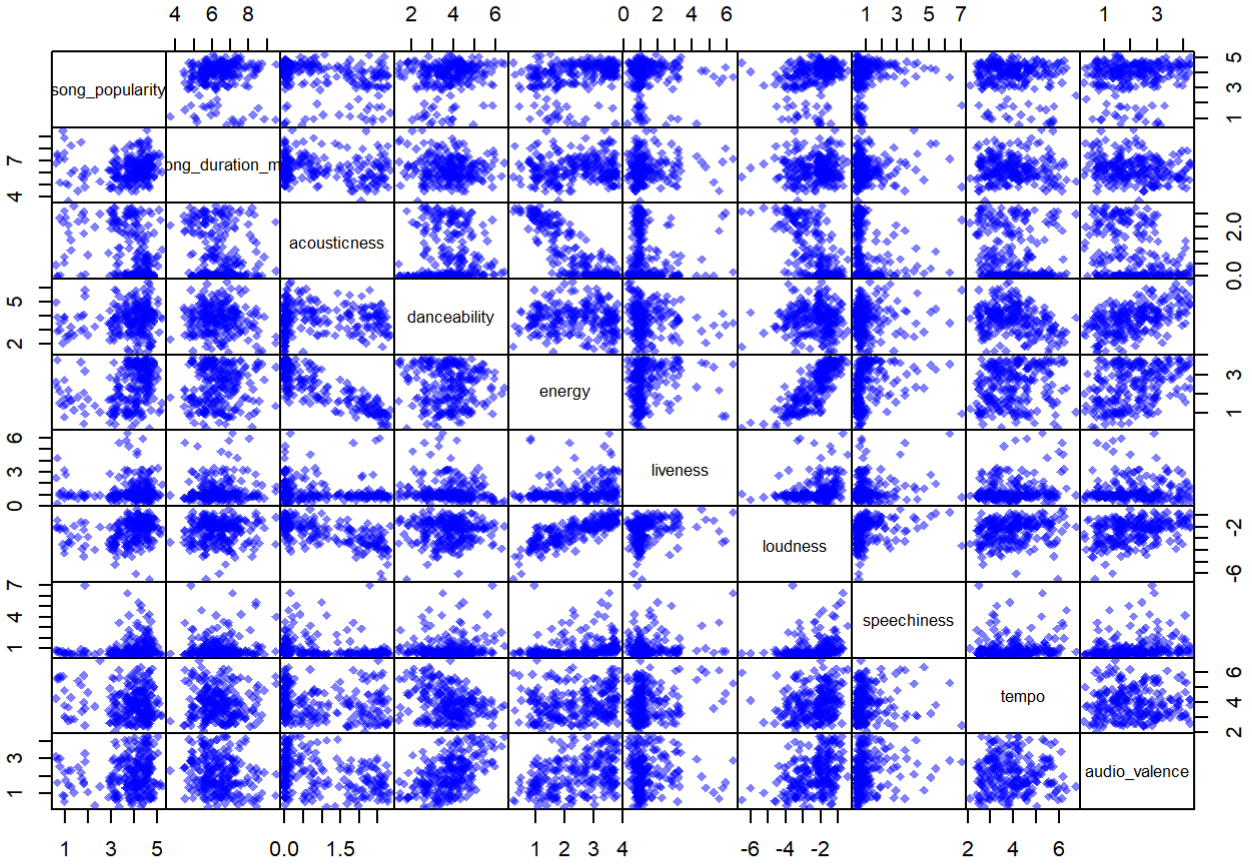
The Scaling process involves either by subraction the columns values by its own mean or by dividing its own value with its Standard Deviation.

Here we follow the 2nd method, As we will be evaluating the Standard Deviation and then divide all the values in observation with respective Standard Deviation.

```
#scaleddataset<-scale(spotify, scale = TRUE)
spotifysd <- apply(spotify,2, sd)  # We will compute the Standard Deviation for the numeric v
alues i the data set
scaleddataset <- sweep(spotify, 2, spotifysd, "/") # Here we divide the values with its own S
D as a process of scaling
```

A Quick Look how th varibale are appeared

```
pairs(scaleddataset,gap =0,pch =18,col =adjustcolor(4,0.5))
```



We could visualise that the genres in the dataset seem to have considerable amount of overlapping from which we wouldnt achieve the exact result of categorising the cluster variables. This leads us to follow the K-Means clustering methods and validate it using **Internal Validation** and **External Validation**

K -Means Clustering:

Inorder to find the exact or range of Clusering value that can be used in the process can be narroed or analysed by Clustering Validation.

We will perform the validation using the following Types of Validation:

Internal Validation

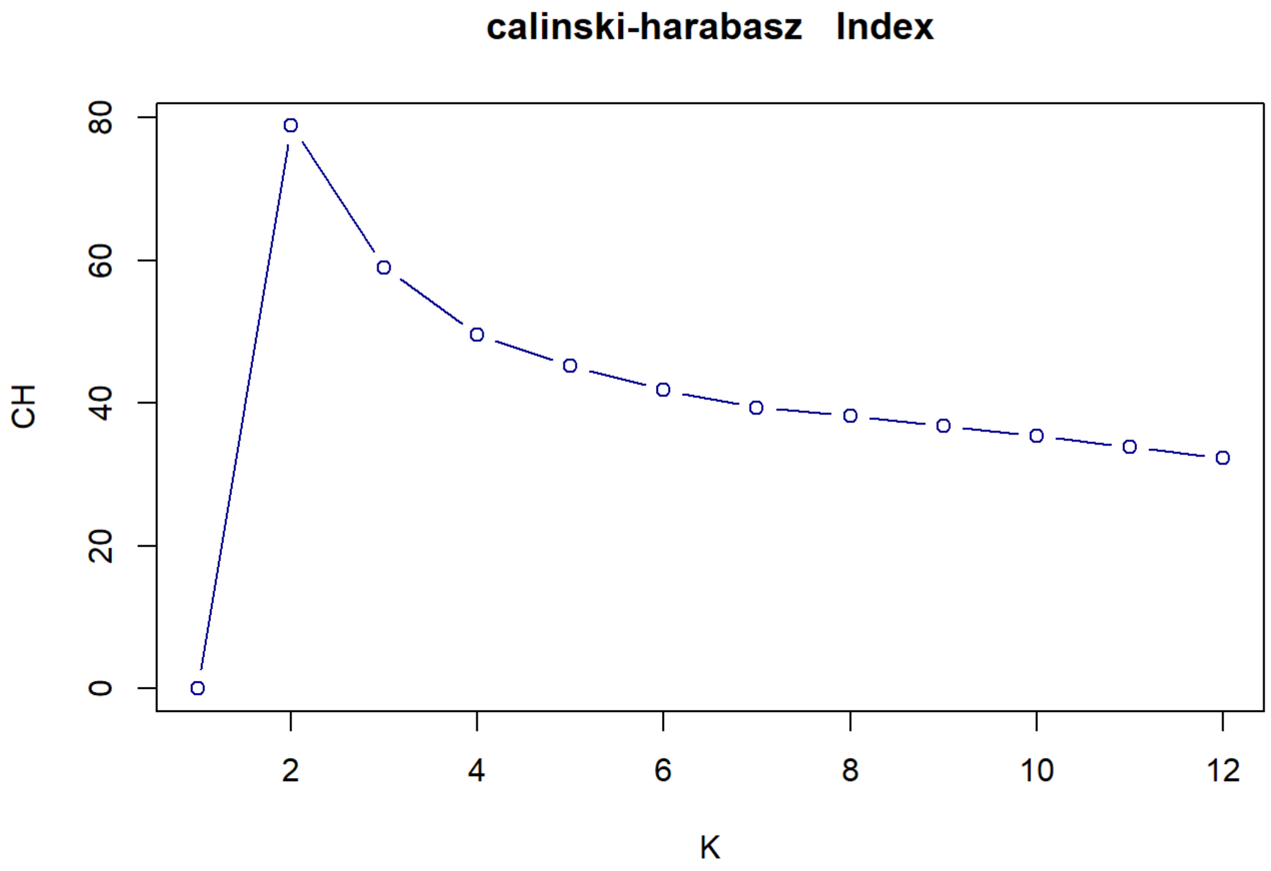
- **Calinski-Harabasz** index
- **Silhouette**

External Validation

- **Rand and adjusted Rand** index

Let us consider the **Calinski-Harabasz (CH) index** initialy to narrow down the K value and reciprocate and validate in all other validations too.

```r
Kmax <- 12 # seting  maximum K value
wss <- bss <- rep(NA, Kmax) # initializing the vectors used to find CH
for ( Kloop in 1:Kmax ) {
# run kmeans for each value of k
fit <- kmeans(scaleddataset, centers = Kloop, nstart = 20)
wss[Kloop] <- fit$tot.withinss #  total within sum of squares
bss[Kloop] <- fit$betweenss
}
# Deriving CH index
N <- nrow(scaleddataset)
ch <- ( bss/(1:Kmax - 1) ) / ( wss/(N - 1:Kmax) )
ch[1] <- 0  #intialise the initial value to be 0
plot(1:Kmax, ch, type = "b",
     ylab = "CH", xlab = "K",
     main="calinski-harabasz   Index",col="darkblue")
```

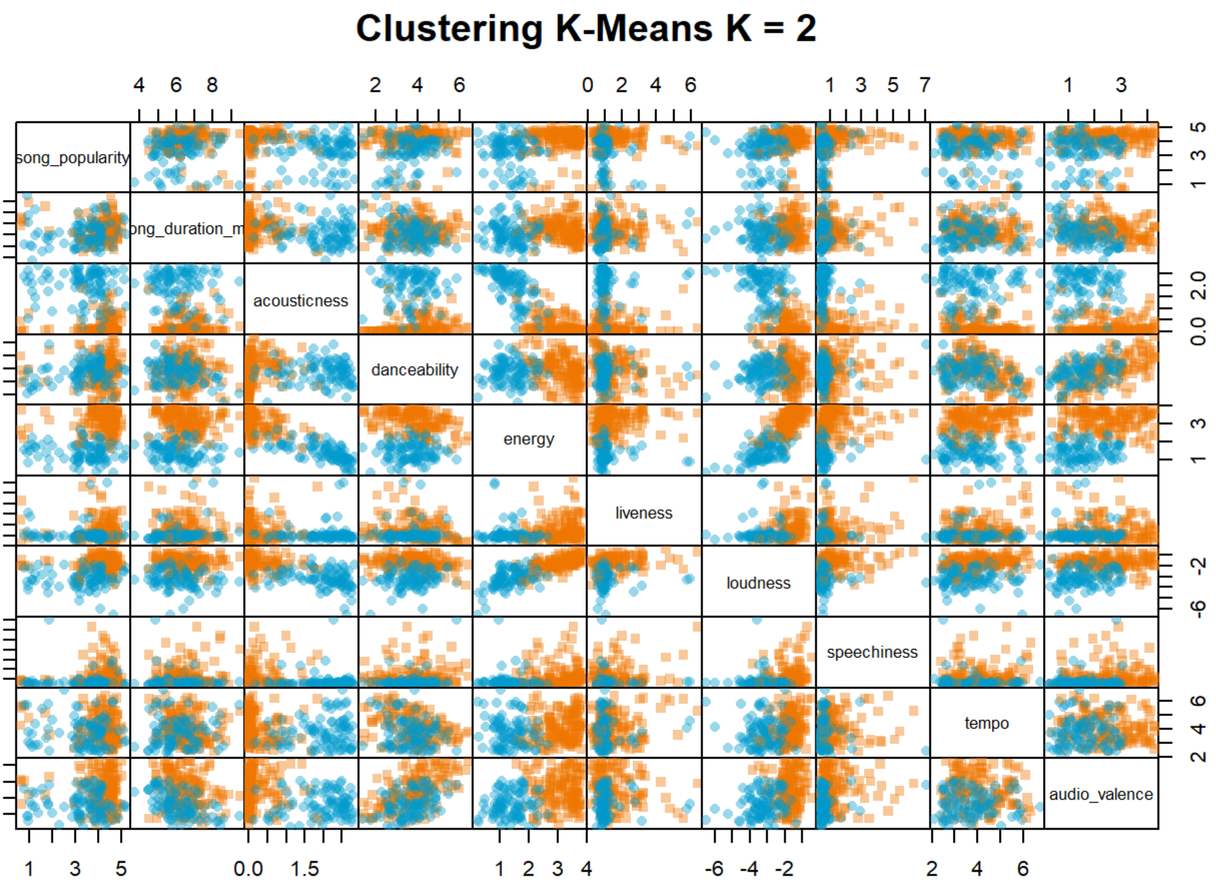


Here from the CH Plot above we can observe that the CH value of k=2 Clusters has the highest CH index and following k=3 has the next highest CH index and the difference of the CH index between k=2 and k=3 seem to be higher than between any other cluster numbers. Which means that all the other clusters such that k=3 and above has lower variances between them , Narrowing our predictions to evolve arounf k=2 and k=3.

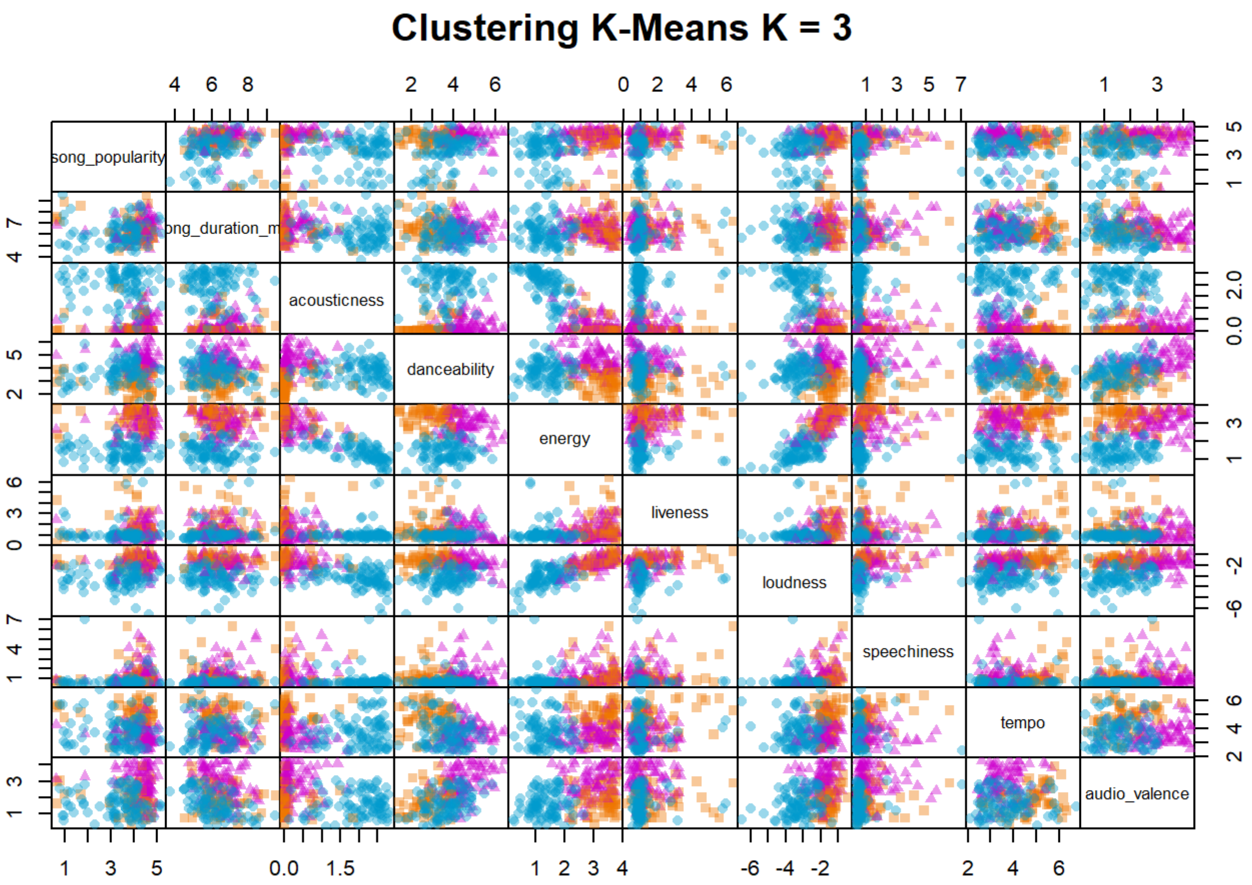
Now Let us perform the K-Means for K=2 and 3 which are most prominent changes in CH value as rest of the K clusters doesnt visualise much of impact on the Data set

```r
kmeanfit2 <- kmeans(scaleddataset, centers = 2, nstart = 50) # K-Mean clustering for 2 Cluster
kmeanfit3 <- kmeans(scaleddataset, centers = 3, nstart = 50) # K-Mean clustering for 3 Cluster

symb <- c(15, 16, 17)
col <- c("darkorange2", "deepskyblue3", "magenta3")
# Let us Visualise the Clustering Plot for K-Means K=2,3.
# K=2
pairs(scaleddataset, gap = 0, pch = symb[kmeanfit2$cluster],
      col = adjustcolor(col[kmeanfit2$cluster], 0.4),
      main = "Clustering K-Means K = 2")
```



```r
# K=3
pairs(scaleddataset, gap = 0, pch = symb[kmeanfit3$cluster],
      col = adjustcolor(col[kmeanfit3$cluster], 0.4),
      main = "Clustering K-Means K = 3")
```



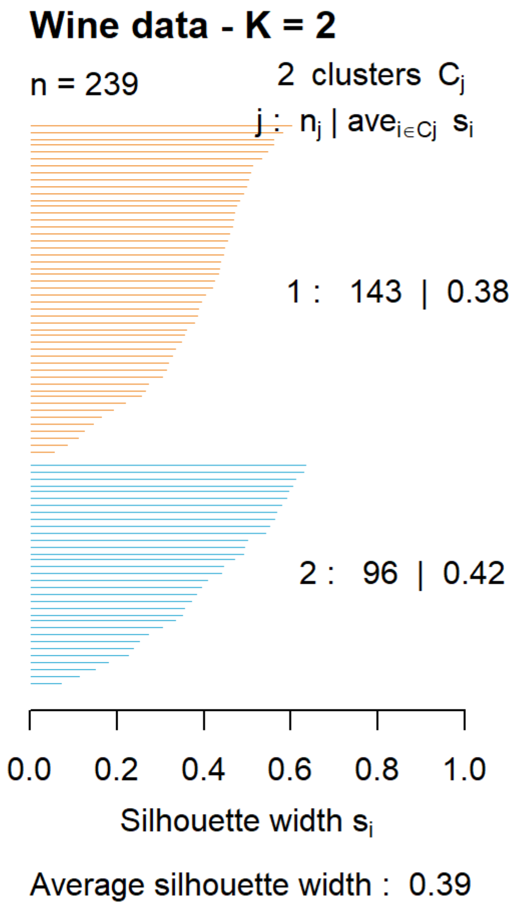
We could see that K=2 has comparably less interactions or overalapping between the clusters than K=3, where in there are ovelapping in both hence we need other validation methods to verify the number of clusters and theor observations.

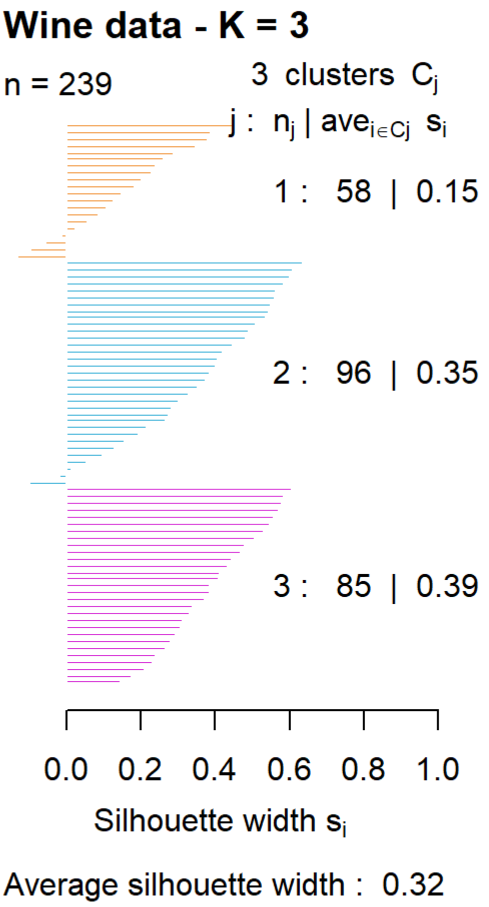
Silhouette :

```
# Silhouette is derived by using the Euclidean Distance and k means which in order uses the s
quared euclidean distance value.
library(cluster)
d <- dist(scaleddataset, method = "euclidean")^2

Silhouette2 <- silhouette(kmeanfit2$cluster, d) #Calculating Silhoutte index for K=2 Clusters
Silhouette3 <- silhouette(kmeanfit3$cluster, d) #Calculating Silhoutte index for K=3 Clusters

par( mfrow = c(1,2) )
plot(Silhouette2, col = adjustcolor(col[1:2], 0.8),
      main = "Wine data - K = 2")
plot(Silhouette3, col = adjustcolor(col[1:3], 0.8),
      main = "Wine data - K = 3")
```





The average silhoutette is larger for the solution with **2 clusters** with an average silhouette of **0.39** than the one with **3** which has a average silhouette of **0.32**.

As we could find some **Negative Silhouette values in the 3 clusters(K=3)**, which is not found in 2 clusters and also the Average values of 3 clusters is lesser than the 2 clusters.

Rand and adjusted Rand index

This method is used to compare the observations in the cluster by a reference partion.

Where in this makes feasable to find which number of clusters have irrelavent (Not a perfect data belonging to the cluster)

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.6.2
```

```
genre <- factor( spotify1$genre, labels = c("rock", "pop", "acoustic")) #factoring the music
 genre and assigning them to variable type

tab <- table(kmeanfit2$cluster, genre)
tab
```

```
##    genre
##     rock pop acoustic
##   1   10  75       58
##   2   90   5        1
```

```
# The function to compute Rand and Adjusted Rand Index

classAgreement(tab)  # Tabluating 2 Cluster values
```

```
## $diag
## [1] 0.06276151
##
## $kappa
## [1] -0.5234626
##
## $rand
## [1] 0.7342569
##
## $crand
## [1] 0.4741481
```

we could observe one cluster holds most of the Rock Genre as the other cluster holds both pop and acoustic.

The Rand value of 0.7342569 is predominently near to to the value 1, which describes that k=2 clusters matches best in this dataset. Wherein The Rand values might change obeservation to oberservation in a dataset.It is better to consider the CRand (Adjusted Rand index) which does'nt change in a higher variation yet remains contant throughout the dataset.

when K=3 clusters is performend here; We could observe that the one cluster has predominantly of Rock genre songs and the rest of the 2 cluster has mixed songs of both pop and acoustics majoritily.

Conclusion:

As we performed our analysis on the Dataset containing the audio features for a collection of songs extracted from the music streaming platform Spotify, can be categorised into 2 clusters by predicting the clusters based on the data observations rather than the original 3 Genres.