# Statistical Machine Learning-Assignment 1

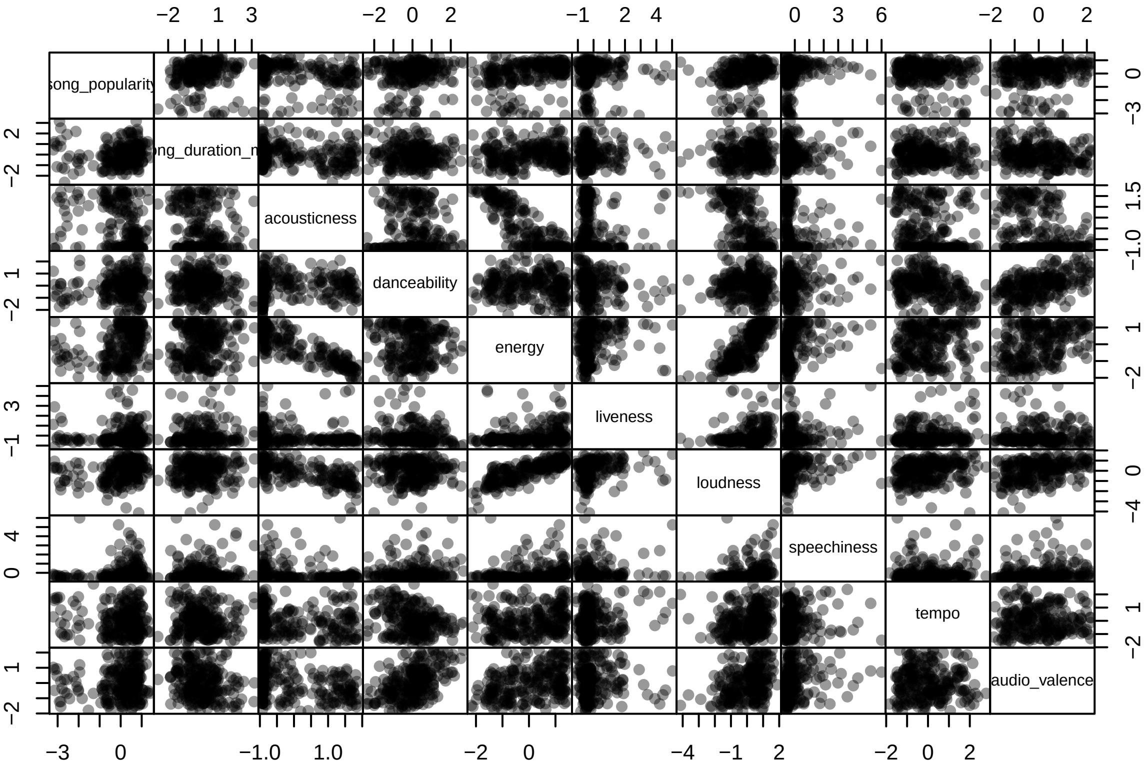
Nisarga G-19203753

25/02/2020

Task: Complete the cluster analysis of the Spotify audio features data using k-means by peforming internal and external validation.

Load data set: Load the data and drop only those columns which are not numeric for the pair plot to work.Using 'scale' standardize the data so that it has mean 0 and variance 1.

```r
data <- read.csv("data_spotify_songs.csv")
X <- data[,-c(1,2,3)]
X <- scale(X)
pairs(X, gap = 0, pch = 19, col = adjustcolor(1, 0.4))
```



We can observe that there is no clear distinction between the observations to determine the number of clusters even though the data is grouped into two clusters.

Clustering Validation: There are 2 types of validation - internal and external. Internal validation is useful to select the number of clusters and check whether appropriate k is used. This can be done in 2 ways- Calinski-Harabasz index and Silhouette.

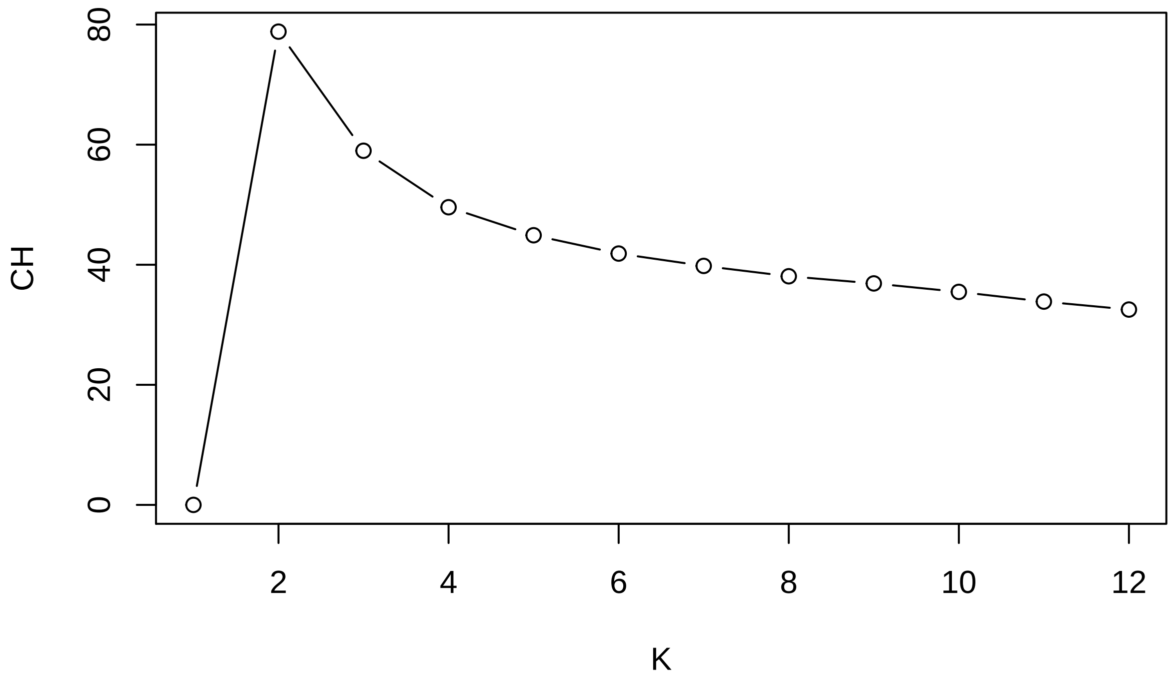
Calinski-Harabasz index: We run K-means on the spotify songs data for different values of K. Storing within sum of squares and the between sum of squares, with the output we can calculate and plot the Calinski-Harabasz index.

```
K <- 12 # set K max
wss <- bss <- rep(NA, K) # initialize empty vector

for ( k in 1:K ) {
 # run kmeans for each value of k
 fit <- kmeans(X, centers = k, nstart = 50)
 wss[k] <- fit$tot.withinss # store total within sum of squares
 bss[k] <- fit$betweenss
}

# compute calinski-harabasz index
N <- nrow(X)
ch <- ( bss/(1:K - 1) ) / ( wss/(N - 1:K) )
ch[1] <- 0

plot(1:K, ch, type = "b", ylab = "CH", xlab = "K")
```



From the plot we can infer that CH index is maximum i.e 80 for k=2. By calculating using table function we can see that 96(acoustic) observations fall into one cluster whereas 143(pop and rock) observation fall into cluster 2.

```
fit<-kmeans(X,centers =2)
table(fit$cluster)
```

```
##
##   1   2
## 143  96
```

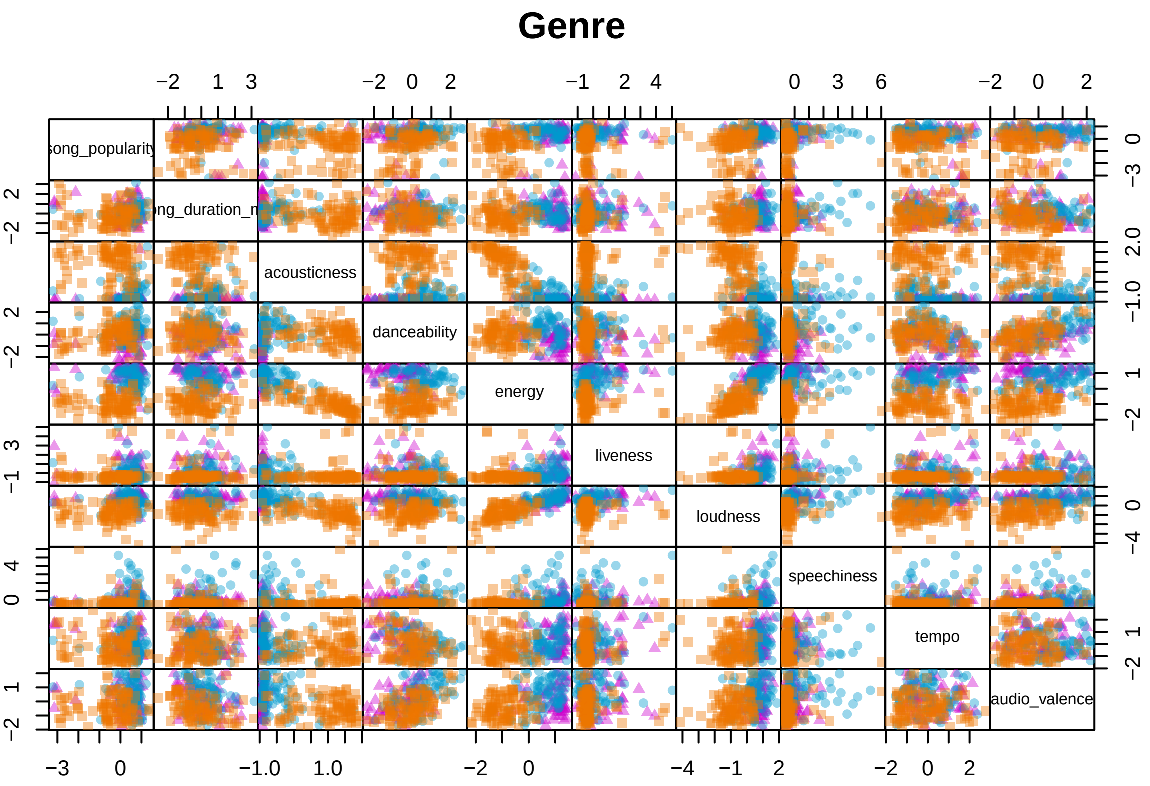
```
table(data$genre)
```

```
##
## acoustic      pop     rock
##      100       80       59
```

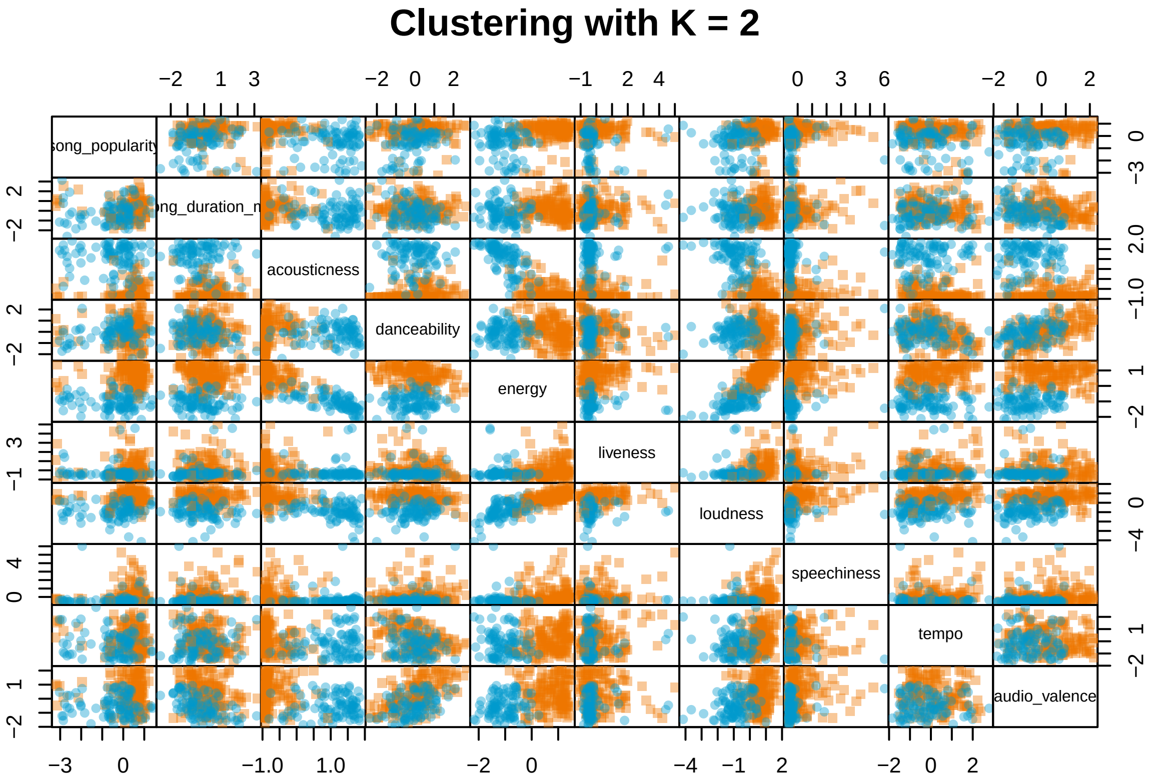
```
tabl <-table(fit$cluster,data$genre)
tabl
```

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

```
par(mfrow =c(2,1))
symb <-c(15,16,17)
col <-c("darkorange2","deepskyblue3","magenta3")
pairs(X,gap =0,pch =symb[data$genre],col =adjustcolor(col[data$genre],0.4),main ="Genre")
```



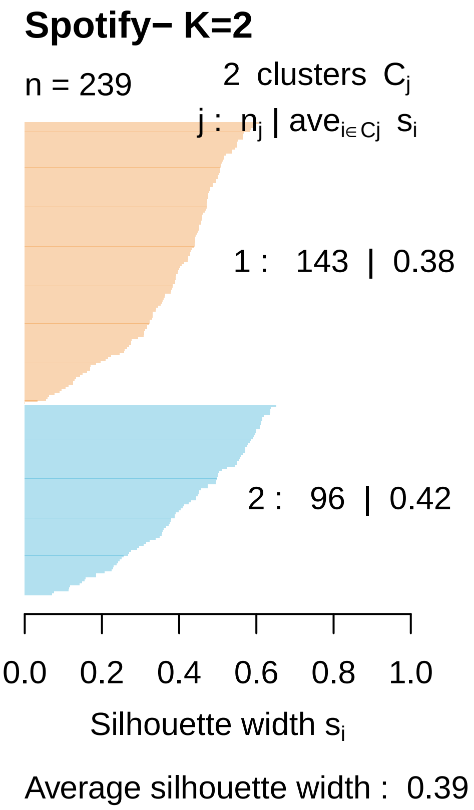
```
pairs(X,gap =0,pch =symb[fit$cluster],col =adjustcolor(col[fit$cluster],0.4),main ="Clustering with K =
```

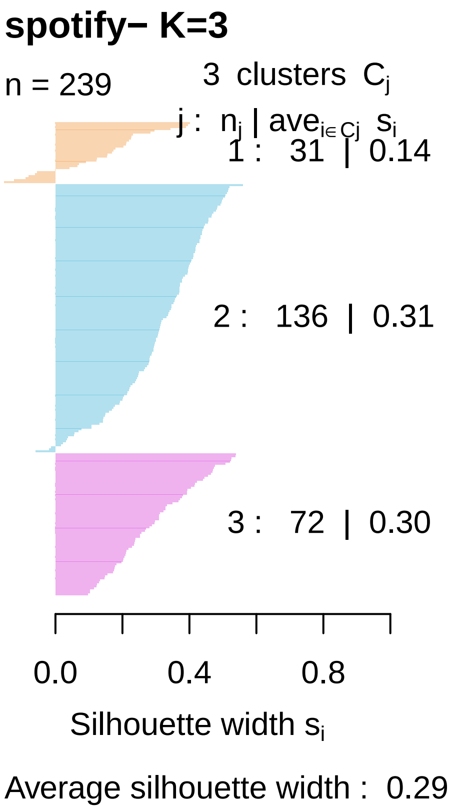


By comparing the two plots above,as the colour order do not necessarily match we can infer that blue in the cluster plot is same as purple in the genre plot.Also, yellow in the cluster is same as blue in the genre plot.

Silhouette: The silhouette is a measure of how close each point in one cluster is to points in the neighboring clusters.Observation with high silhouette are close to their own cluster members while those with low silhouette are close to members of neighbouring cluster.The cluster package has a function called silhouette which computes the silhouette for each observation. The function takes two arguments: a clustering of the data and a distance matrix.

```
fit3 <-kmeans(X,centers =3)
library(cluster)
d <- dist(X, method = "euclidean")^2
sil2 <-silhouette(fit$cluster, d)
sil3 <-silhouette(fit3$cluster, d)
# produce the two silhouette plots
col <- c("darkorange2", "deepskyblue3", "magenta3")
par( mfrow = c(1,2) )
plot(sil2, col = adjustcolor(col[1:2], 0.3),main="Spotify- K=2")
plot(sil3, col = adjustcolor(col, 0.3),main="spotify- K=3")
```





The plot with higher average silhouette value will be much likely to be true than the one with the lower value.Here we can see that the silhouette plot with k=2 has higher value than the one with k=3.Hence, we can confirm that the number of clusters is 2.

External Validation:we compare the clustering to an external reference clustering, the one with the kmeans versus the genre-wise partition.

Rand Index:Need to quantify the agreement between two clustering partitions using classAgreement function.

```r
library(e1071)
classAgreement(tab= tabl)
```

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

Rand index is 0.73 which is almost 1 indicates that there is an agreement that there are only 2 clusters in the spotify data set.