Assignment1

SIVASHANKAR-19203161

13/02/2020

Introduction

The dataset is about audio features for a collection of songs extracted from the music streaming platform Spotify. We already know that the songs are classified as Rock, Pop and Acoustic. We are going to perform clustering analysis in this dataset to find a set of clusters that might group these songs in a different way.

Loading the dataset

Before loading the dataset, we will load the libraries which are required for the analysis.

```
library(e1071)
```

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

```
library("cluster")
```

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

The dataset is loaded and it is scaled for better analysis. The columns which are removed for scaling are genre, song_name and artist. These are non-numeric values and it can't be scaled. Moreover they are not useful for analysis. Hence these values are removed from the analysis.

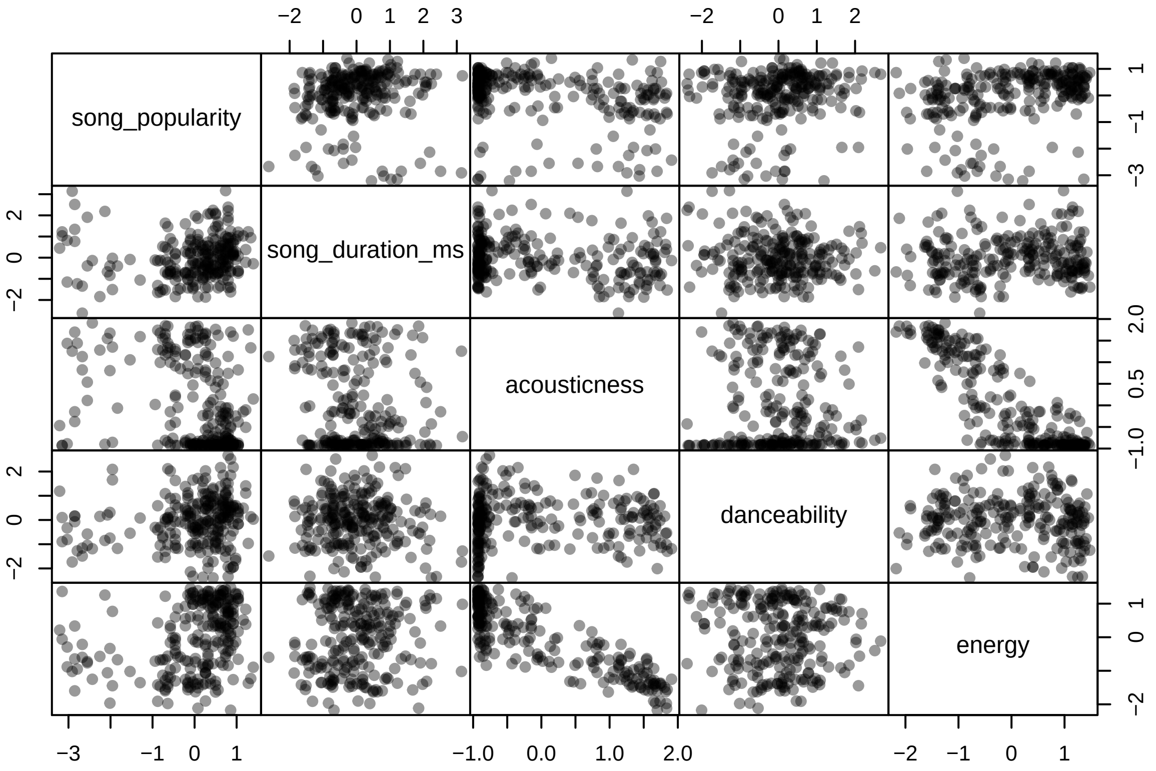
```
load("data_spotify_songs.rda")
data = data.frame(scale(spotify[-c(1,2,3)]))
head(data,3)
```

```
##   song_popularity song_duration_ms acousticness danceability     energy
## 2       0.1401785       -0.2532293   -0.8999343   -0.1660636  0.9464688
## 3       0.7384401        0.1620791   -0.9061221    1.1348938 -0.5875902
## 4       0.6187878       -0.2532293   -0.8531627   -0.7731771  1.4066864
##     liveness   loudness speechiness       tempo audio_valence
## 2 -0.4678201  0.1023366  -0.2700317 -0.51063349    -0.4668657
## 3  0.7433486 -0.3961496   0.1958749  0.09939666    -0.6700066
## 4 -0.5172555  0.6176613   0.6364261  0.05233017    -1.2264361
```

visualizing the data

By doing visualization, we might get some idea about the number of clusters present in the dataset.

```
pairs(data[,c(1,2,3,4,5)], gap = 0, pch = 19, col = adjustcolor(1, 0.4))
```



We are only visualizing a set of columns only, because if we visualize all the columns, then the plots will be hard to read. Here we can see that the possible number of clusters can be 2 or 3. This is evident if we look at the plot song_popularity vs danceability and song_popularity vs acousticness.

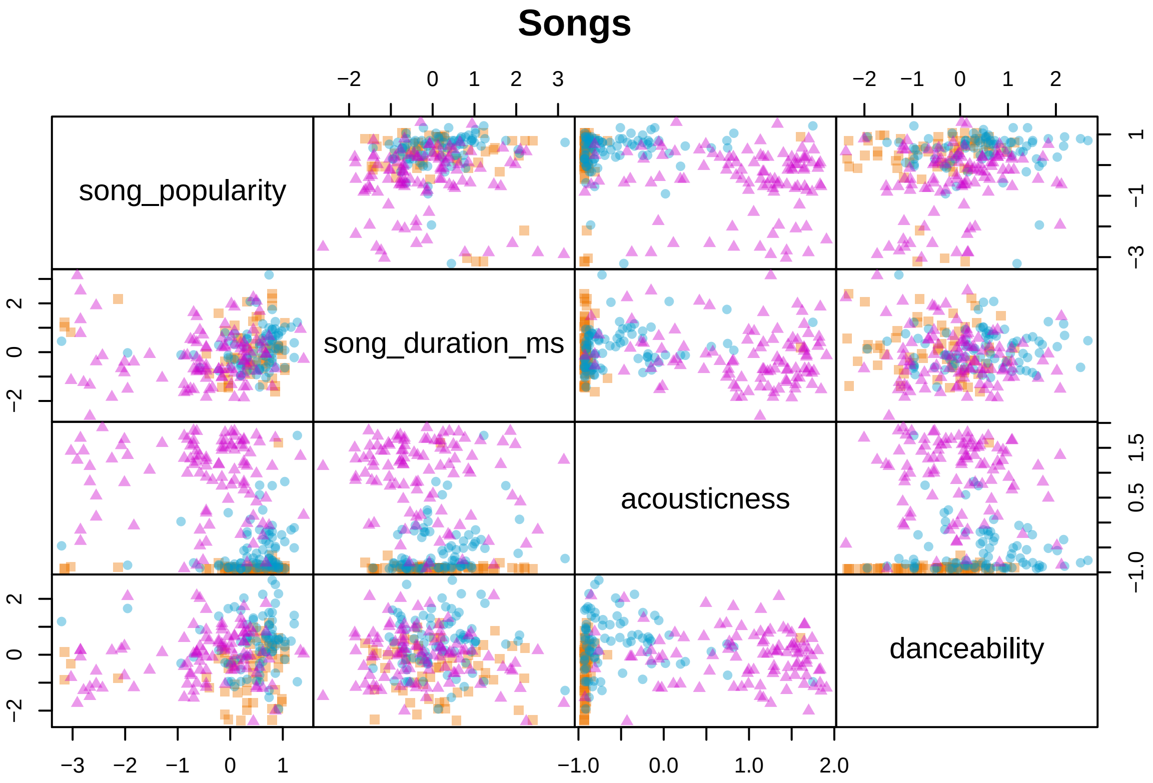
Model training

Let us create 2 models fit2 and fit3. These models represents 2 and 3 clusters respectively.

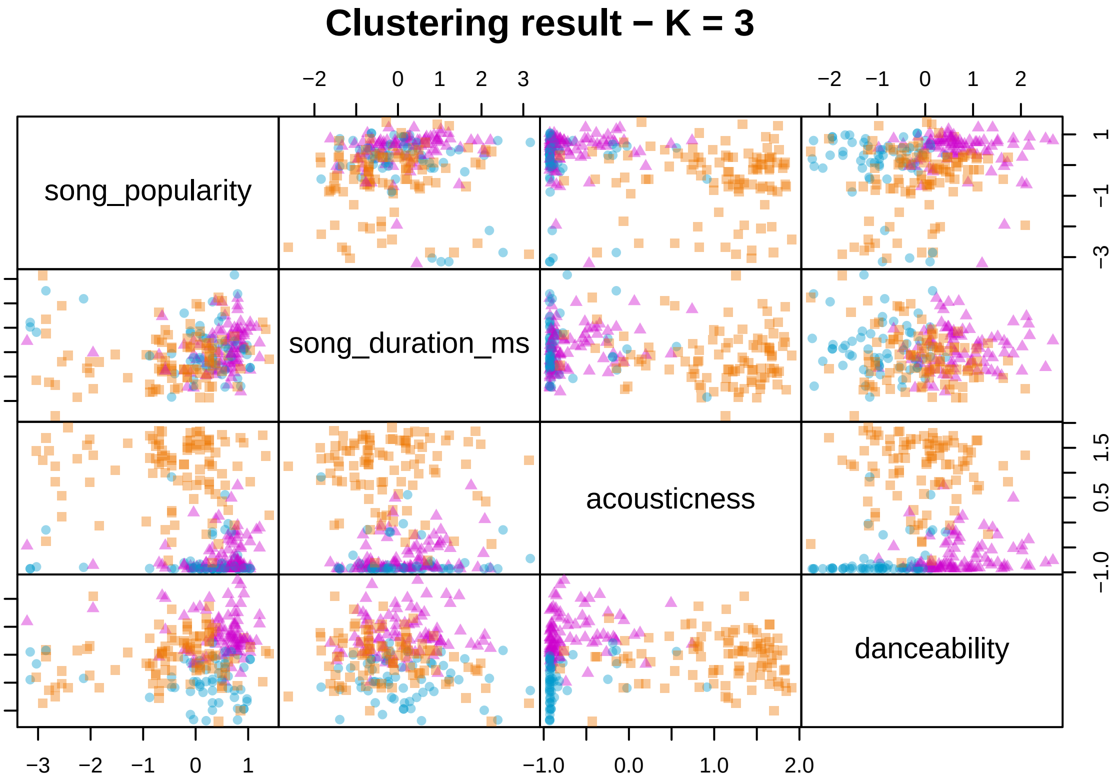
```r
fit2 = kmeans(data,centers = 2,nstart = 30)
fit3 = kmeans(data,centers = 3,nstart = 30)
```

In the below code, the comparison is shown between the cluster types given in dataset and clusters found using K means algorithm.

```r
par(mfrow=c(2,1))

symb <- c(15, 16, 17)
col <- c("darkorange2", "deepskyblue3", "magenta3")

# plot with symbol and color corresponding to the genre

pairs(data[c(1,2,3,4)], gap = 0, pch = symb[spotify$genre],
col = adjustcolor(col[spotify$genre], 0.4),
main = "Songs")
```

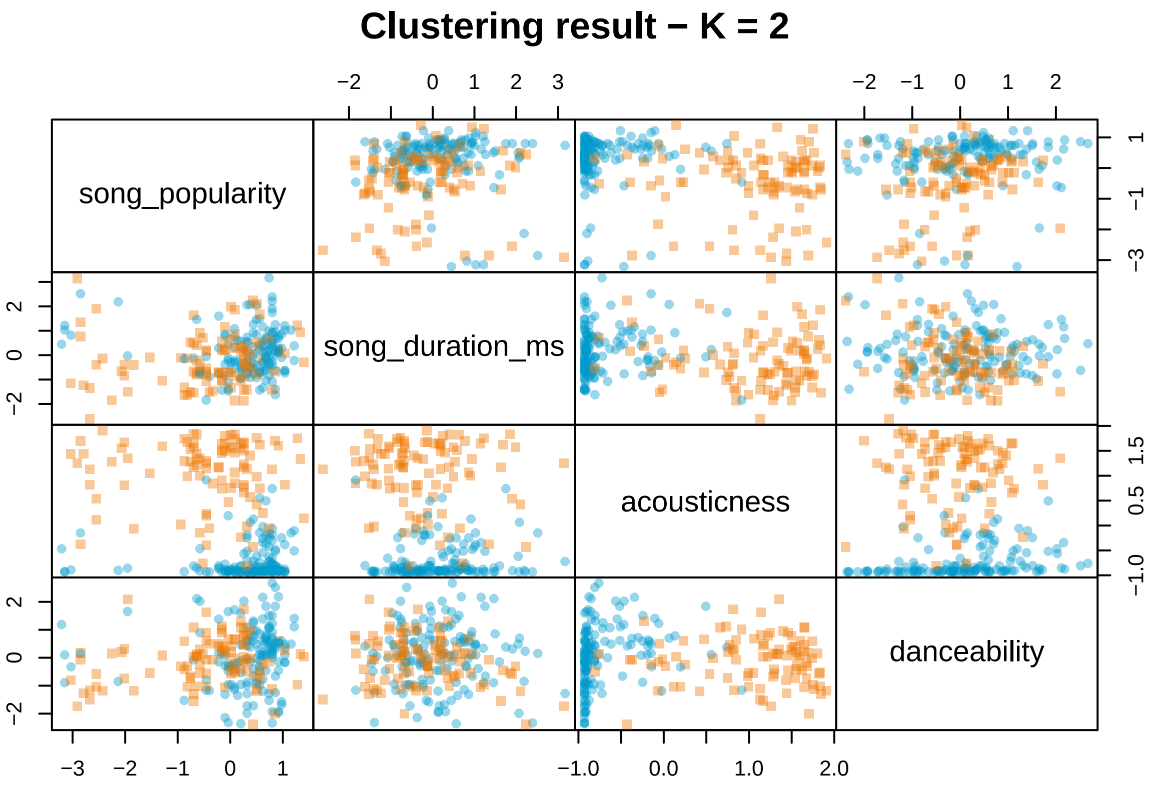


```
pairs(data[c(1,2,3,4)], gap = 0, pch = symb[fit3$cluster],
col = adjustcolor(col[fit3$cluster], 0.4),
main = "Clustering result - K = 3")
```



Most of them are clustered perfectly by the K-means algorithm for k=3. The colours might be different but it classified the clusters in the right groups. Let us visualize the model where k= 2.

```
# plot with symbol and color corresponding to the species
pairs(data[c(1,2,3,4)], gap = 0, pch = symb[fit2$cluster],
col = adjustcolor(col[fit2$cluster], 0.4),
main = "Clustering result - K = 2")
```



CLUSTER VALIDATION

Now, we have to validate the model to confirm that whether k=3 or 2.

We have 2 types of validation. Internal validation and External validation. Internal validation is performed to choose the best possible value of K. External validation is performed to measure the performance of the model.

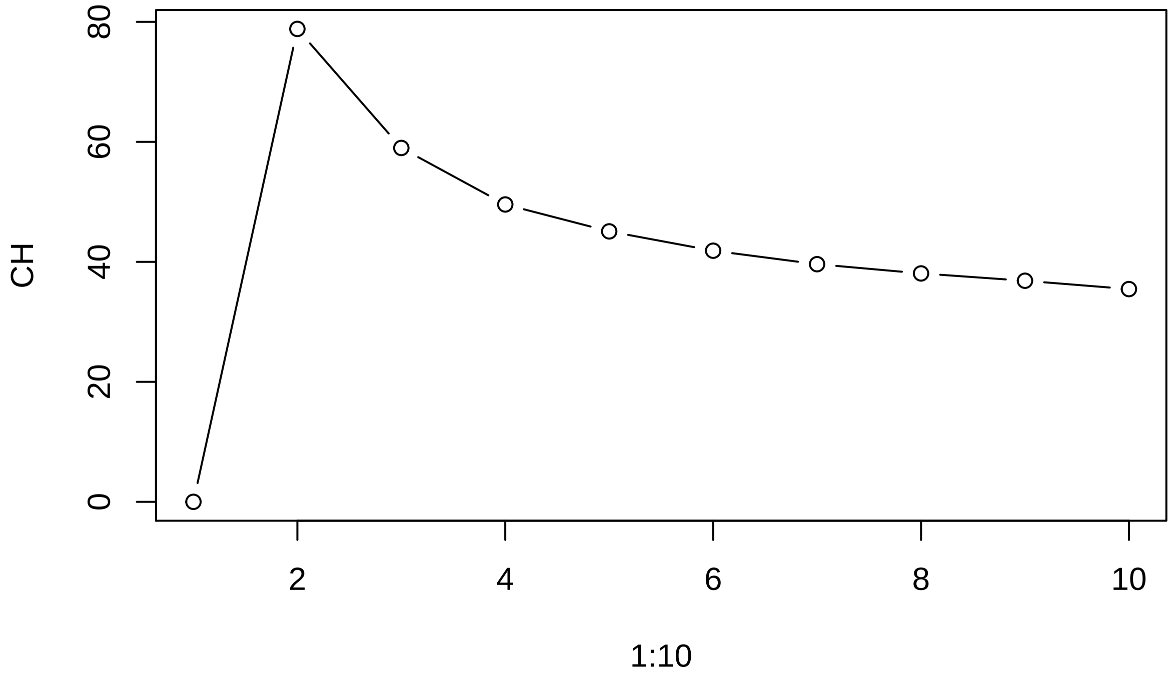
INTERNAL VALIDATION

In internal validation, we look at Calinski-Harabasz index and Silhouette methods.

Calinski-Harabasz index

We will run K-means algorithm for k value starting from 1 to 10. We will choose the K value whose corresponding CH value is maximum.

```
nc = 10

N = nrow(data)
Wss = rep(0,nc)
Bss = rep(0,nc)
for(k in 1:nc){
  fit <- kmeans(data, centers = k, nstart = 30)

  Wss[k] = fit$tot.withinss
  Bss[k] = fit$betweenss

}

CH = ( Bss/(1:nc - 1)) / (Wss/(N -  1:nc))
CH[1] = 0

plot(1:10,CH,type = "b")
```



Calinski-Harabasz index tells us that k=2 is the right choice. We will now look at the other validation method to confirm this statement.

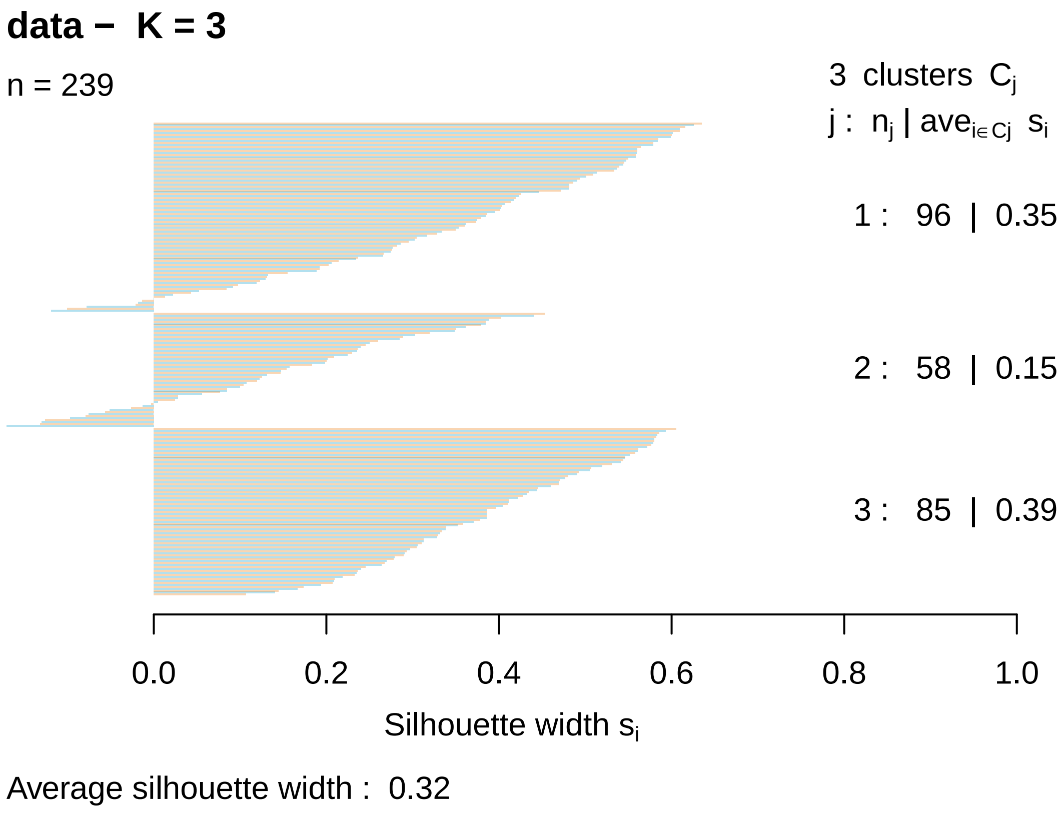
Silhouette

Now we test the models for which k=2 and 3 respectively. The model which has a maximum average silhouette width is chosen to be the right choice.

In the below, let us check for the model k=3.

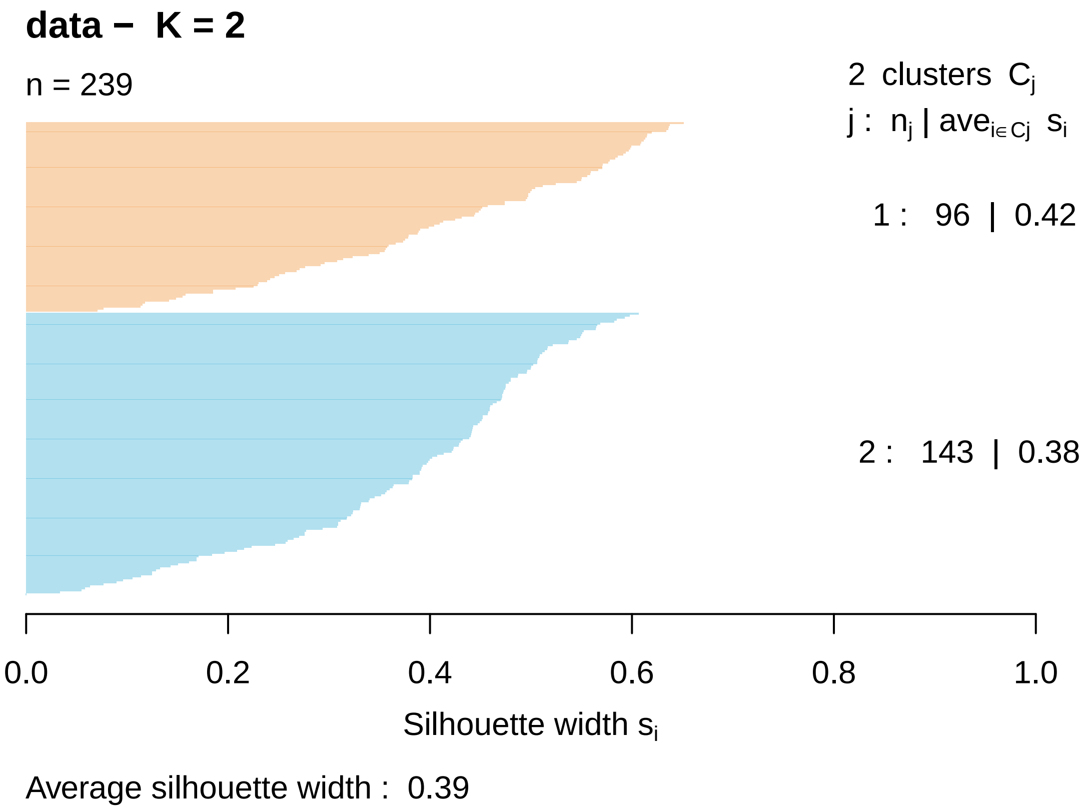
```
d3 <-dist(data,method ="euclidean")^2

sil3 <-silhouette(fit3$cluster, d3)
```

```
col <-c("darkorange2","deepskyblue3")
plot(sil3,col =adjustcolor(col,0.3),main ="data -  K = 3")
```



For k=3, we got the value 0.32. If the average silhouette width for the model k=2 is large, then that model is considered.

```
d2 <-dist(data,method ="euclidean")^2

sil2 <-silhouette(fit2$cluster, d2)

col <-c("darkorange2","deepskyblue3")
plot(sil2,col =adjustcolor(col,0.3),main ="data -  K = 2")
```



The average silhouette width of this model is greater than the above model(k=3). Moreover, there are some negative values present for the model k=3. Hence the model for k=2 is chosen for further analysis.

External Validation

In external validation, we will look at Rand index and Adjusted Rand index values. Adjusted Rand Index values are preferred for analysis.

```
tab = table(fit2$cluster,spotify$genre)
classAgreement(tab)
```

```
## $diag
## [1] 0.3179916
##
## $kappa
## [1] 0.02648874
##
## $rand
## [1] 0.7342569
##
## $crand
## [1] 0.4741481
```

The classification rate is poor as the diag value is around 0.3

The performance of the model is around 0.47 (adjusted Rand Index) and the Rand index is 0.73.