Audio Features and Song Popularity

5/2/2023

Data collection and cleaning

```r
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
library(dbplyr)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
df <- read.csv("spotify_dataset.csv")
```

```r
df <- na.omit(df)
df <- df[!duplicated(df), ]
```

```r
dim(df)
```

```
## [1] 41099    20
```

```r
colnames(df)
```

```
##  [1] "track"            "artist"           "uri"              "danceability"    
##  [5] "energy"           "key"              "loudness"         "mode"            
##  [9] "speechiness"      "acousticness"     "instrumentalness" "liveness"        
## [13] "valence"          "tempo"            "duration_ms"      "time_signature"  
## [17] "chorus_hit"       "sections"         "popularity"       "decade"
```

```r
head(df)
```

```
##                   track           artist                                  uri
## 1 Jealous Kind Of Fella    Garland Green spotify:track:1dtKN6wwlolkM8XZy2y9C1
## 2         Initials B.B. Serge Gainsbourg spotify:track:5hjsmSnUefdUqzsDogisiX
## 3          Melody Twist      Lord Melody spotify:track:6uk8tI6pwxxdVTNlNOJeJh
## 4         Mi Bomba Sonó       Celia Cruz spotify:track:7aNjMJ05FvUXACPWZ7yJmv
## 5           Uravu Solla      P. Susheela spotify:track:1rQ0clvgkzWr001POOPJWx
## 6             Beat n. 3  Ennio Morricone spotify:track:32VBSoD2vcoIOiPEvAfFXU
##   danceability energy key loudness mode speechiness acousticness
## 1        0.417  0.620   3   -7.727    1      0.0403        0.490
## 2        0.498  0.505   3  -12.475    1      0.0337        0.018
## 3        0.657  0.649   5  -13.392    1      0.0380        0.846
## 4        0.590  0.545   7  -12.058    0      0.1040        0.706
## 5        0.515  0.765  11   -3.515    0      0.1240        0.857
## 6        0.697  0.673   0  -10.573    1      0.0266        0.714
##   instrumentalness liveness valence   tempo duration_ms time_signature
## 1         0.00e+00   0.0779   0.845 185.655      173533              3
## 2         1.07e-01   0.1760   0.797 101.801      213613              4
## 3         4.42e-06   0.1190   0.908 115.940      223960              4
## 4         2.46e-02   0.0610   0.967 105.592      157907              4
## 5         8.72e-04   0.2130   0.906 114.617      245600              4
## 6         9.19e-01   0.1220   0.778 112.117      167667              4
##   chorus_hit sections popularity decade
## 1   32.94975        9          1    60s
## 2   48.82510       10          0    60s
## 3   37.22663       12          0    60s
## 4   24.75484        8          0    60s
## 5   21.79874       14          0    60s
## 6   65.48604        7          0    60s
```

Data Analysis

Correlation table and Plot

Get the correlations between features and popularity, and sort them decreasingly. From the table below, we observed a strong negative correlation between a song's instrumentalness(probability of being instrumental), acousticness(probability of being acoustics), and duration (in milliseconds) and its popularity. Conversely, we observed a strong positive correlation between a song's danceability, loudness, valence, and energy and its popularity.

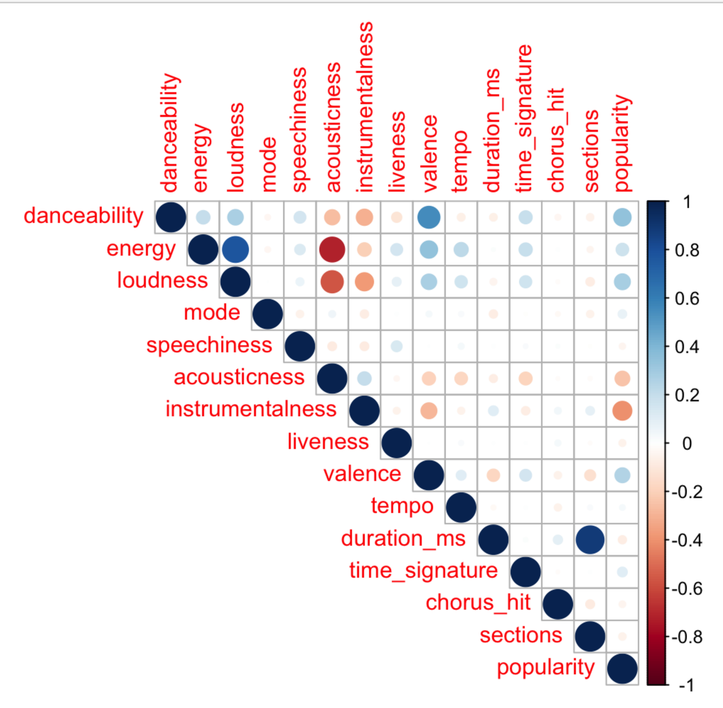
```r
correlations <- cor(df[, c("danceability", "energy", "loudness", "mode", "speechiness",
                           "acousticness", "instrumentalness", "liveness", "valence", "tempo",
                           "duration_ms", "time_signature", "chorus_hit", "sections", "popularity")])

cor_df <- sort(correlations["popularity", ]) %>% as.data.frame()
colnames(cor_df) <- c("correlation")
cor_df
```

```
##                  correlation
## instrumentalness -0.40756039
## acousticness     -0.24595220
## duration_ms      -0.07380557
## sections         -0.05999298
## liveness         -0.05148418
## chorus_hit       -0.04641644
## speechiness      -0.04093589
## tempo             0.03258179
## mode              0.07963298
## time_signature    0.10494133
## energy            0.17711715
## valence           0.25111742
## loudness          0.28597211
## danceability      0.34601993
## popularity        1.00000000
```

Further, we utilize corrplot to generate an upper correlation matrix. We see that danceability, energy, and loudness are sharing common underlying features that may contribute to popularity.

```r
corrplot(correlations, method = "circle", "upper")
```



Person's test on correlation

Hypothesis:

H_0 : There is no correlation between the variables.

H_1 : There is a correlation between the variables.

Given that the p-values from our correlation tests are all significantly less than 0.05, we can confidently reject the null hypothesis of no correlation on the significance level of 0.05, and conclude that there are indeed statistically significant relationships between danceability, loudness, and valence with popularity in the Spotify dataset

```r
cor.test(df$danceability, df$popularity, method = "pearson")
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  df$danceability and df$popularity
## t = 74.765, df = 41097, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3374809 0.3545020
## sample estimates:
##       cor 
## 0.3460199
```

```r
cor.test(df$loudness, df$popularity, method = "pearson")
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  df$loudness and df$popularity
## t = 60.5, df = 41097, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2770702 0.2948249
## sample estimates:
##       cor 
## 0.2859721
```

```r
cor.test(df$valence, df$popularity, method = "pearson")
```

```
## 
##  Pearson's product-moment correlation
## 
## data:  df$valence and df$popularity
## t = 52.593, df = 41097, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2420371 0.2601538
## sample estimates:
##       cor 
## 0.2511174
```

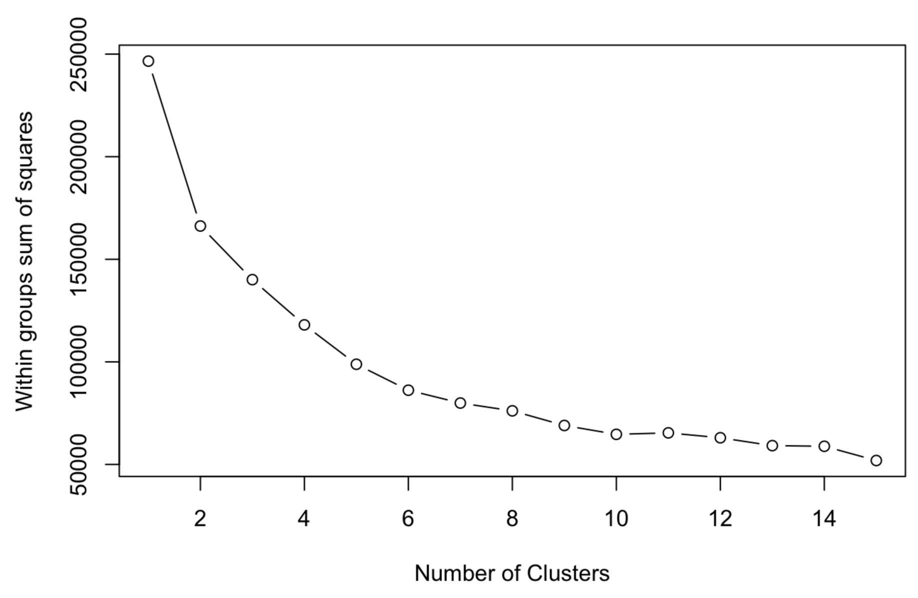
K-means clustering on features

From the matrix, we retrieve six features that are most correlated to the popularity as a data frame, and we apply k-means clustering to see it there is any underlying pattern. We choose K as 9 from the elbow method by the elbow method.

```r
features <- df[, c("track","danceability", "energy", "loudness", "popularity", "instrumentalness", "acousticnes
s")]

scaled_features <- scale(features[,-1])

wss <- (nrow(scaled_features)-1)*sum(apply(scaled_features,2,var))
for (k in 1:15) {
    wss[k] <- sum(kmeans(scaled_features, centers=k)$withinss)
}
plot(1:15, wss, type="b", xlab="Number of Clusters",
    ylab="Within groups sum of squares")
```

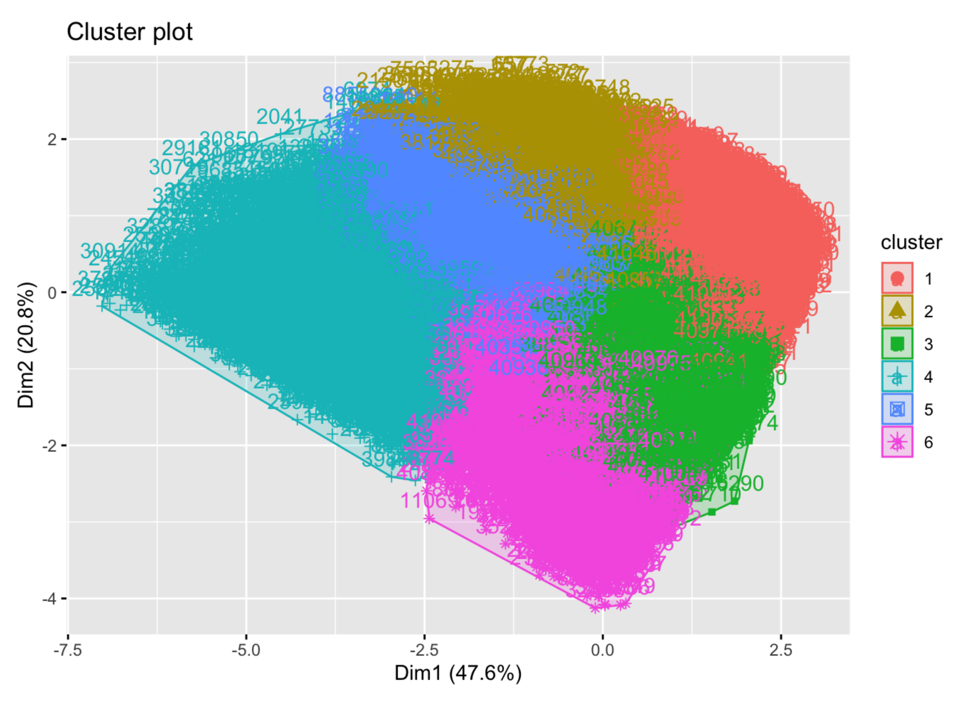



```r
# K-means clustering
set.seed(123)
km.res <- kmeans(scaled_features, 6, nstart = 20)
```

```r
km.res$size
```

```
## [1] 13340  6833  7485  3556  6386  3499
```

```r
fviz_cluster(km.res, data = scaled_features)
```



In cluster 2, we observe a interesting pattern that the acousticness can contribute to one's popularity if it is slow and quiet, which lead us to think that in the music industry, acoustic music can be popular if they are more traditional, solely and slowly performed in one instrument.

And we can also come up with an obvious conclusion that if one has danceability, energy, and loudness, one is more likely to be popular.

```r
cluster_summary <- aggregate(scaled_features, by=list(cluster=km.res$cluster), FUN=mean)
cluster_summary
```

```
##   cluster danceability     energy   loudness popularity instrumentalness
## 1       1    0.5992647  0.5720563  0.5603926  1.0000608       -0.4546250
## 2       2   -0.1359211 -0.5934044 -0.2218034  1.0000608       -0.4535804
## 3       3   -0.1204800  0.7777388  0.5473921 -0.9996476       -0.3833085
## 4       4   -1.1068991 -1.5530381 -1.9291713 -0.9644822        2.0284206
## 5       5   -0.1638928 -0.9101914 -0.5859280 -0.9999148       -0.3687304
## 6       6   -0.3374957  0.5536473  0.1556401 -0.8221519        2.0505074
##   acousticness
## 1   -0.6794388
## 2    0.5909972
## 3   -0.6374355
## 4    1.4536389
## 5    1.0190206
## 6   -0.5372883
```

Conclusion

In this analysis, we delve into the intricacies of Spotify music data, focusing on their distinct features. By employing statistical techniques such as correlation tables and plots, Pearson's correlation test, and K-means clustering, we unravel the impact of these features on the popularity of the tracks. Our findings offer a comprehensive perspective on how specific characteristics can influence the listeners' preferences, thereby affecting a song's overall popularity.

In Cluster 2, an intriguing pattern emerges suggesting that a song's acousticness can boost its popularity if it is slow and tranquil. This observation prompts us to infer that within the music industry, acoustic music can gain popularity if it resonates with traditional styles, focuses on a single instrument, and maintains a slow tempo.