Spotify Popularity Predictors

by Hayden Bean & Maya Barlow
Denison University, Professor Zhe Wang

Abstract

This project aims to answer the question of what makes modern songs popular, and what factors impact the overall popularity of songs on the streaming platform Spotify.

```
 acousticness        artists          danceability     duration_ms       energy           explicit
 Min.   :0.000002   Length:4370        Min.   :0.0000   Min.   : 14629   Min.   :0.000281   Min.   :0.0000
 1st Qu.:0.030225   Class :character   1st Qu.:0.4880   1st Qu.:182151   1st Qu.:0.445000   1st Qu.:0.0000
 Median :0.165000   Mode  :character   Median :0.6070   Median :218942   Median :0.615000   Median :0.0000
 Mean   :0.295990                      Mean   :0.5972   Mean   :222116   Mean   :0.595110   Mean   :0.2584
 3rd Qu.:0.525000                      3rd Qu.:0.7180   3rd Qu.:252350   3rd Qu.:0.773000   3rd Qu.:1.0000
 Max.   :0.996000                      Max.   :0.9770   Max.   :886293   Max.   :0.998000   Max.   :1.0000
      id              instrumentalness
 Length:4370        Min.   :0.000000
 Class :character   1st Qu.:0.000000
 Mode  :character   Median :0.000039
                    Mean   :0.135390
                    3rd Qu.:0.071875
                    Max.   :0.964000
      key            liveness          loudness          mode            name          popularity
 Min.   : 0.000   Min.   :0.0193   Min.   :-33.966   Min.   :0.0000   Length:4370        Min.   : 0.0
 1st Qu.: 2.000   1st Qu.:0.0969   1st Qu.:-10.270   1st Qu.:0.0000   Class :character   1st Qu.:40.0
 Median : 5.000   Median :0.1250   Median : -7.168   Median :1.0000   Mode  :character   Median :57.0
 Mean   : 5.233   Mean   :0.1861   Mean   : -8.229   Mean   :0.6689                      Mean   :52.5
 3rd Qu.: 8.000   3rd Qu.:0.2280   3rd Qu.: -5.298   3rd Qu.:1.0000                      3rd Qu.:69.0
 Max.   :11.000   Max.   :0.9920   Max.   :  0.674   Max.   :1.0000                      Max.   :92.0
   release_date       speechiness        tempo           valence           year
 Length:4370        Min.   :0.0000   Min.   :  0.00   Min.   :0.0000   Min.   :1924
 Class :character   1st Qu.:0.0312   1st Qu.: 96.12   1st Qu.:0.2810   1st Qu.:1989
 Mode  :character   Median :0.0459   Median :119.05   Median :0.4650   Median :2011
                    Mean   :0.0759   Mean   :119.81   Mean   :0.4759   Mean   :2001
                    3rd Qu.:0.0903   3rd Qu.:139.95   3rd Qu.:0.6600   3rd Qu.:2017
                    Max.   :0.9800   Max.   :217.93   Max.   :0.9800   Max.   :2020
```

Introduction

This data set deals with a number of variables that impact the popularity of songs. The variable for popularity was created by means of an algorithm, mainly based off of how many plays a song has and how recent those plays were. All variables are measured on a scale from zero to one. The variables we will focus on for this research are: energy, tempo, danceability, acousticness, liveness, valence, loudness, and instrumentalness.

Methodology

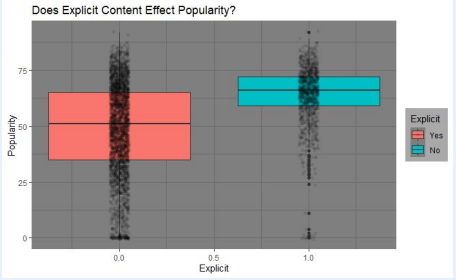
We first used the streams variable for each month to narrow the datasets down and select songs with a large number of streams. This ensures a relatively popular grouping of songs and it also helps filter out any outliers. Then, using the popularity variable plotted many linear regressions against various variables, we tested to see which variables were statistically significant.

```
Residuals:
     Min       1Q   Median       3Q      Max
-72.789 -10.215   1.073  11.461  78.292

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.974e+01  2.826e+00  24.678  < 2e-16 ***
energy          -1.349e+01  2.206e+00  -6.115 1.05e-09 ***
danceability     3.385e+01  1.762e+00  19.208  < 2e-16 ***
acousticness    -1.028e+01  1.211e+00  -8.486  < 2e-16 ***
instrumentalness -6.668e+00  1.446e+00  -4.612 4.10e-06 ***
liveness        -5.086e+00  1.716e+00  -2.965  0.00304 **
tempo            2.281e-02  8.650e-03   2.637  0.00839 **
loudness         1.971e+00  1.022e-01  19.282  < 2e-16 ***
valence         -2.117e+01  1.244e+00 -17.021  < 2e-16 ***
key              1.297e-03  6.832e-02   0.019  0.98486
duration_ms     -6.738e-06  3.817e-06  -1.766  0.07755 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.4 on 4359 degrees of freedom
Multiple R-squared:  0.338,     Adjusted R-squared:  0.3364
F-statistic: 222.5 on 10 and 4359 DF,  p-value: < 2.2e-16
```

The above visual is a summary of a linear regression with all the variables. Each variable with one or more stars next to them indicates that there is a possible correlation between them. By repeating this process with each individual variable, we can determine which one has the strongest correlation with song popularity.

Results

By conducting multiple linear regressions as well as two sample t-tests, we discovered that the variable “explicit” has the strongest correlation with song popularity.




```
        Welch Two Sample t-test

data:  spotdata2000$popularity and
spotdata2000$explicit
t = 171.47, df = 4373.1, p-value < 2.2e-16
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 51.64639 52.84102
sample estimates:
 mean of x  mean of y
52.5020595  0.2583524
```

The t-test demonstrates the strength of the relationship between these two variables. The p-value is very low.

This means that the difference between the two group averages did not happen by chance. The above plot shows that songs are more popular when they contain explicit content. The blue box, or songs with explicit content, has a higher average popularity score than the red box, or songs without explicit content.

Conclusion

From this data we can conclude that there is a positive correlation between the content of a song and it’s overall popularity. This does not mean that if a song has explicit lyrics it causes it to be more popular. There is no cause and effect relationship between the variables. Instead, we can conclude that when a song is explicit, we see an increase in the popularity variable.

Correlation ≠ Causation

Acknowledgements

https://www.kaggle.com/yamaerenay/spotify-data set-19212020-160k-tracks?select=data.csv
https://spotifycharts.com/regional
https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/