Spotify Analysis

Katie Herlihy, Jack Cronin, David Barnes,
Jamar Manning and Brendan Chua

9/29/2020

Introduction

We are executives of a record label, and are hoping to run some analytics to help us gain some insight into what makes popular music popular, so that we can better identify key features that make up a successful song and use that information in the future.

Our Data

Our data was retrieved from Kaggle.com, and it is best described as the top tracks from 1920-2020, with approximately 10,000 songs selected per genre

```
head(SpotifyFeatures)
```

```
## # A tibble: 6 x 18
##   genre artist_name track_name track_id popularity acousticness danceability
##   <chr> <chr>       <chr>      <chr>         <dbl>        <dbl>        <dbl>
## 1 Movie Henri Salv… C'est bea… 0BRjO6g…          0        0.611        0.389
## 2 Movie Martin & l… Perdu d'a… 0BjC1Nf…          1        0.246        0.59
## 3 Movie Joseph Wil… Don't Let… 0CoSDzo…          3        0.952        0.663
## 4 Movie Henri Salv… Dis-moi M… 0Gc6TVm…          0        0.703        0.24
## 5 Movie Fabien Nat… Ouverture  0IuslXp…          4        0.95         0.331
## 6 Movie Henri Salv… Le petit … 0Mf1jKa…          0        0.749        0.578
## # … with 11 more variables: duration_ms <dbl>, energy <dbl>,
## #   instrumentalness <dbl>, key <chr>, liveness <dbl>, loudness <dbl>,
## #   mode <chr>, speechiness <dbl>, tempo <dbl>, time_signature <chr>,
## #   valence <dbl>
```

Additional variables assigned to each song:

(as defined by Spotify)

Additional variables assigned to each song:

(as defined by Spotify)

danceability, loudness, speechiness, popularity, mode, key,
acousticness, liveness, instrumentalness, valence

Data Errors

Upon further inspection, we decided to clean up our data so any models we run will have accurate implications. This meant that we excluded the *Comedy* genre, because comedy specials are not the concern of our label executives and it may skew our findings. We are also converting the duration from ms to seconds. We also detected a spelling error that resulted in duplicate columns, so corrected that as well.

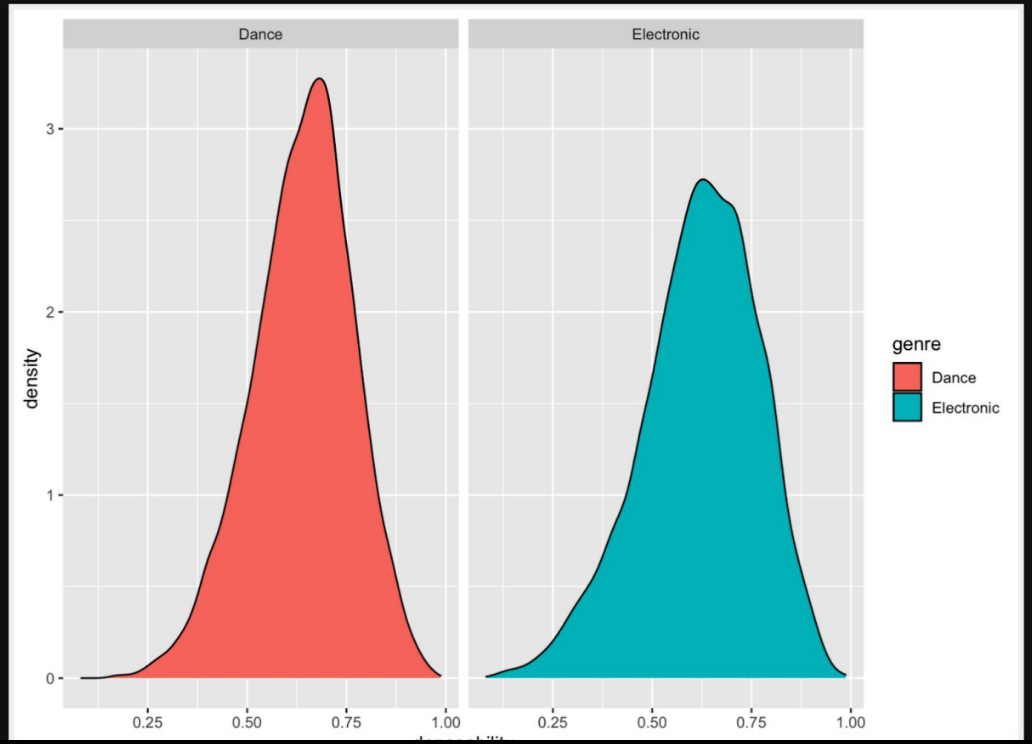
```
SpotifyFeatures <- SpotifyFeatures %>%
  filter(genre !="Comedy") %>%
  filter(genre != "Children's Music")

SpotifyFeatures <- SpotifyFeatures %>%
  mutate(duration_ms= duration_ms/1000) %>%
  rename(duration_s = duration_ms)
```

Problem #1: *Music Festival*

Our label is sponsoring a music festival and we are trying to decide which genre to invite to perform. As an executive producer, we believe that *danceability* is a huge factor in attendee satisfaction.

> *Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.*

From this graph we see that the distribution of danceability between Electronic and Dance music is very similar, with a little bit of a wider distribution within the Electronic genre.



Next, we wanted to see how the average danceability between these two genres compared.

We see that the average danceability is very similar between these two categories- *but are they significantly different?*

```
DanceDanceabilityMean
```

```
## [1] 0.6381911
```

```
ElectronicGenreMean
```

```
## [1] 0.6195422
```

Testing Approach

We are going to run a two-sided hypothesis test. We are wondering *if the danceability between Dance and Electronic music is significantly different*, so we can determine which genre will be a better fit for our event.

$$d = difference$$
$$h_0 : d = 0$$
$$h_1 : d \neq 0$$

Running our T-Test

```
##
##  Welch Two Sample t-test
##
## data:  DanceGenre$danceability and ElectronicGenre$danceability
## t = 9.1792, df = 17987, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01466672 0.02263114
## sample estimates:
## mean of x mean of y
## 0.6381911 0.6195422
```

Test Results

Our test shows that the danceability between Dance and Electronic music are significantly different, by looking at both the confidence interval and the p-value. While we were not expecting the danceability of the two categories to be significantly different because the averages were so similar, we see now that the difference of means is indeed significant and therefore we will choose to feature Dance music in our festival.

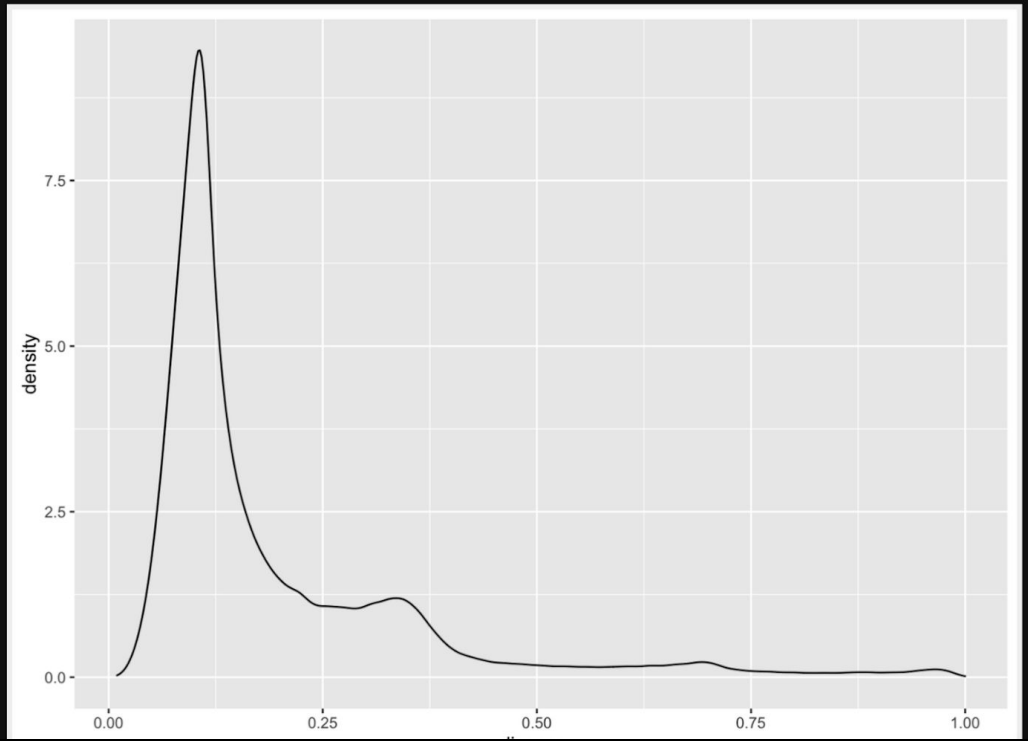
Problem #2: *Introducing more live song versions*

Many of our artists have expressed interest in releasing live versions of their most popular songs, and we are wondering if having a higher liveness leads to above average popularity. If this is true, we will consider more live versions of songs in order to increase total revenues.

Liveness detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

‘Popularity’ is determined by the popularity of the song lately (default country = US). Ranked on a scale 1-100.

We first ran a density plot for liveness, and found it very skewed to the right. Because of this, we decided to use the median as the defining point for ‘high’ or ‘low’ liveness as opposed to the mean.



We used these parameters to create a “HighLiveness” subset from our main data:

```
HighLiveness <- SpotifyFeatures %>%
  filter(liveness>0.128)
HighLiveness
```

```
## # A tibble: 104,539 x 18
##    genre artist_name track_name track_id popularity acousticness danceability
##    <chr> <chr>       <chr>      <chr>         <dbl>        <dbl>        <dbl>
##  1 Movie Henri Salv… C'est bea… 0BRjO6g…          0        0.611        0.389
##  2 Movie Martin & l… Perdu d'a… 0BjC1Nf…          1        0.246        0.59
##  3 Movie Fabien Nat… Ouverture  0IuslXp…          4        0.95         0.331
##  4 Movie Le Club de… Les bisou… 0VSqZ3K…         10        0.319        0.598
##  5 Movie Leopold St… Symphony … 0XKgego…          0        0.921        0.191
##  6 Movie Richard M.… Keys of L… 0pXwl2C…          0        0.97         0.4
##  7 Movie Michel Roux Les avent… 0uWUjxM…          0        0.548        0.588
##  8 Movie Jean Claud… Diane      0vS7Zid…          0        0.7          0.625
##  9 Movie Bernard Mi… Ultra Man… 0x8xSao…          3        0.488        0.744
## 10 Movie Henri Salv… Veunise    113pHPs…          1        0.381        0.451
## # … with 104,529 more rows, and 11 more variables: duration_s <dbl>,
## #   energy <dbl>, instrumentalness <dbl>, key <chr>, liveness <dbl>,
## #   loudness <dbl>, mode <chr>, speechiness <dbl>, tempo <dbl>,
## #   time_signature <chr>, valence <dbl>
```

Testing Approach

$$h_0 : Average\ Popularity\ of\ High\ Liveness <= 41.3$$
$$h_1 : Average\ Popularity\ of\ High\ Liveness > 41.3$$

Running our T-Test

```
##
##  One Sample t-test
##
## data:  HighLiveness$popularity
## t = -14.083, df = 104538, p-value = 1
## alternative hypothesis: true mean is greater than 43.13
## 95 percent confidence interval:
##  42.2736     Inf
## sample estimates:
## mean of x
##  42.36317
```

Test Results

After running this test, we fail to reject the null hypothesis; the average popularity for songs with ‘high liveness’ is not significantly above average.

Problem #3: *Factors of Popularity*

As a music label, we see that times are changing and it's important to stay on top of current trends. We are very interested in identifying the factors that influence the popularity of songs. We believe that popularity is the most valuable KPI, and it is the most directly related to revenue.

We first ran a single variable regression analysis of genre on popularity.

We expected the genre in which songs lie to be pretty closely correlated with whether a song was popular or not, and our findings support this

```
## genreHip-Hop          49.1206     0.8962  54.808  < 2e-16 ***
## genreIndie            45.3990     0.8961  50.664  < 2e-16 ***
## genreJazz             31.5219     0.8961  35.175  < 2e-16 ***
## genreMovie             2.8716     0.8973   3.200  0.00137 **
## genreOpera             4.0331     0.8969   4.497 6.91e-06 ***
## genrePop              57.2881     0.8962  63.925  < 2e-16 ***
## genreR&B              43.0062     0.8964  47.976  < 2e-16 ***
## genreRap              51.2313     0.8963  57.161  < 2e-16 ***
## genreReggae           26.2868     0.8966  29.320  < 2e-16 ***
## genreReggaeton        28.4404     0.8965  31.725  < 2e-16 ***
## genreRock             50.3169     0.8962  56.142  < 2e-16 ***
## genreSka              19.3098     0.8965  21.539  < 2e-16 ***
## genreSoul             37.7253     0.8964  42.088  < 2e-16 ***
## genreSoundtrack       24.6523     0.8960  27.513  < 2e-16 ***
## genreWorld            26.2216     0.8963  29.254  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.715 on 217616 degrees of freedom
## Multiple R-squared:  0.6798, Adjusted R-squared:  0.6797
## F-statistic: 1.925e+04 on 24 and 217616 DF,  p-value: < 2.2e-16
```

As people who want to really gain a deeper understanding of the music, we want to look into more than just genre as a predictor of popularity; we want to know what factors of the music explain the other 30% of variance in popularity.

We decide to run a multivariable regression model, using acousticness, loudness and danceability as independent variables against popularity.

Acousticness is the relative metric of the track being acoustic (0-1); Loudness is the relative loudness of the track in the typical range [-60, 0] in decibel (dB)

We chose these factors because they had the highest R^2 for the remaining variables when they were all run individually against popularity:

```
##
## Call:
## lm(formula = popularity ~ acousticness + danceability + loudness,
##     data = SpotifyFeatures)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.03 -10.36   1.71  10.97  56.12
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.168077   0.151477  265.18   <2e-16 ***
## acousticness -6.223624   0.137294  -45.33   <2e-16 ***
## danceability 17.793429   0.202188   88.00   <2e-16 ***
## loudness      0.519987   0.008175   63.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.61 on 217637 degrees of freedom
## Multiple R-squared:  0.1735, Adjusted R-squared:  0.1735
## F-statistic: 1.523e+04 on 3 and 217637 DF,  p-value: < 2.2e-16
```

We get some good insights from running this regression analysis. These 3 factors explain approximately 17% of the data- when we think about how much variance was left over from the genre linear regression, we are pretty happy with these insights.

It seems that generally a unit increase of acousticness results in -6 units of popularity, and a unit increase of danceability results in a 17 unit increase in popularity. A less helpful finding was that a unit increase of loudness leads to only a .52 unit increase of popularity.

How we'll use these results:

We know that while genre is probably the best indicator of popularity, we now know which variables we can run when asked to choose between multiple songs that are within a single genre.

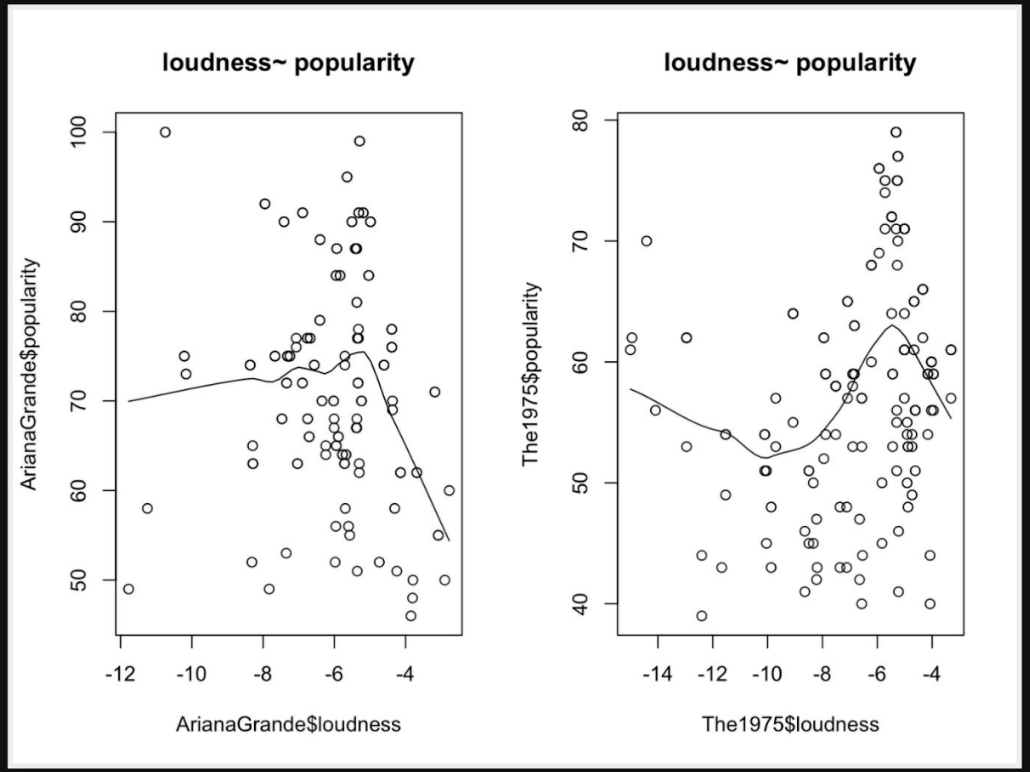
For example, if the ultimate goal for a record launch is popularity, we can use a regression analysis similar to this to determine which songs should be selected as the album's singles.

Concluding Thoughts

While we believe that analytics can be a very strong when trying to reach definite conclusions about a particular dataset, we want to acknowledge that many other factors come into play when considering music.

Music is a very subjective field, and we would be cautious of placing too much confidence in the numbers. Trends can very quickly change, and no analysis can determine if a song is ‘good’ or ‘bad’. For example, earlier we ran a regression model identifying loudness as one of the main predictors in popularity.

Let’s take a look at how loudness plays into popularity for two different artists, both very successful in their respective genres:



While we made conclusions about loudness within the dataset as a whole, these models may be most useful when run within a single genre or for a single artist. Each artist within a genre may have different variables that deem them ‘successful’, however our analysis offers a very generic overview of Spotify’s top tracks within the last 100 years.

Additional Data

We would have liked to see the date of each song's release, so that we could better observe industry trends within genre throughout time.

We also would have liked some data regarding song engagement, specifically number of downloads or radio streams. This would give us more insight as to what contributes to the definition of what is classified as 'popular'.