Project Report
Spotify Song Reviews and Artist Measures

Nikole Molina, Sarah Phillips, Lizzy Li, Davy Counsell

1. Introduction

Listening to music is a part of many people's daily lifestyle. As of 2022, Spotify alone had 182 million users utilizing its platform.[1] With so many listeners, reviews of songs often influence the popularity of the artist and their music. To analyze this relationship, we utilized data from Spotify 2000 song reviews combined with data of each artist's popularity and Spotify followers collected through the Spotify API.

In this report, we chose song popularity as our target variable and we want to know which variables have a significant impact on predicting song popularity.

2. Data

This data consists of two primary sources of data: Kaggle data on Spotify song reviews, and artists' popularity and their Spotify followers from the Spotify API.

2.1 Kaggle Dataset

The first data source we used in this project was an existing dataset from Kaggle of Spotify song reviews.[2] It is a collection of 1994 songs and their ratings from the top 2000 Spotify songs between 1956-2019. We utilized the first 148 rows of the dataset to perform in-depth analysis on and determine what variables are useful in predicting song popularity.

2.2 Spotify API

To identify and collect data specifically on the song artists, we utilized the Spotify API. This involved running a for loop that collected the artist_id for each of the artists in our dataset. We then used the artist_id to run two other for loops that collected the artist popularity and number of Spotify followers each artist had. Then we combine each of the columns into dataframes that would later be merged with our Kaggle dataset. The API code is included in the R script "Spotify_API.R".

2.3 Data Reviews After Integration

We integrated the Kaggle dataset and the data extracted from the Spotify API using the column "Artist". In order to do this, we first created a dataframe that contained columns with each artist name and their specific artist_id (artist). Then after the artist popularity was extracted using the API we created a dataframe that included artist_id and artist popularity (popularity2). Similarly, after the number of Spotify followers was extracted we created a dataframe that included artist_id and number of followers (followers). We then merged artist, popularity2, and followers, by artist_id, into our final dataframe. In order to perform analysis we renamed 3 of the columns, dropped two columns we did not need, and reordered the columns to increase readability.The final, cleaned data frame consisted of 148 rows with 18 variables. The description of each

[1] • Spotify users - subscribers in 2021 | Statista.
[2] Spotify Kaggle Dataset

variable can be seen in *Table 1.* The integration code is also included in the R script "Spotify_API.R".

Table 1 Data dictionary

Column	Type	Description
Index	integer	Unique ID for each song
Artist	text	Name of the artist
Artist.Popularity	integer	How popular the artist is on a scale of 0-100
Spotify.Followers	integer	Number of followers each artist has on Spotify
Title	text	Name of the song
Top.Genre	text	Genre of the song
Year	integer	Year the song was released
Beats.Per.Minute (BPM)	integer	The tempo of the song in beats per minute
Energy	integer	How energetic the song is on a scale of 0-100
Danceability	integer	How easy the song is to move to on a scale of 0-100
Loudness	integer	How loud the song is in decibels, closer to 0 is louder
Liveness	integer	Likeliness of the song being a live recording on a scale of 0-100
Valence	integer	How positive the song is on a scale of 0-100
Length..Duration	integer	How long the song is in seconds
Acousticness	integer	How acoustic teh song is on a scale of 0-100
Speechiness	integer	How much spoken word the song contains
Song.Popularity	integer	How popular the song is on a scale of 0-100

3. Analysis

3.1 Artist Popularity in Relation to Songs

We created a dplyr summary to calculate the total number of songs from each artist, the artists' popularity rating, as well as the minimum, maximum, and average song popularity. Table 2 displays a sample of 10 artists and their calculations.

Table 2 Artist Popularity Summary Table

Artist	**Songs**	**Artist Popularity**	**Minimum Song Popularity**	**Maximum Song Popularity**	**Average Song Popularity**
3 Doors Down	1	74	77	77	77
Adele	2	90	63	73	68
Alanis Morissette	1	72	57	57	57
Backstreet Boys	1	79	77	77	77
Coldplay	4	92	63	84	75
Foo Fighters	4	83	65	76	72
Norah Jones	2	75	71	74	73
Saybia	2	48	51	55	53
U2	2	81	40	57	49
Youssou N'Dour	1	56	59	59	59

Based on the large amount of data collected, we decided to display an example of 10 artists. The data above displays the artists' popularity compared to their combined songs' popularity. The table indicates that 3 Doors Down and Backstreet Boys have the highest average song popularity and not the highest artist popularity, but only 1 song in the data. Whereas Coldplay has the second highest average of 75 with the highest artist popularity of 92, and has 4 songs to complete that average. Our further analysis on the relationship between artists' popularity and song popularity are displayed in the following sections.

3.2 Evaluate Artist.Popularity, Energy, and BPM

We are looking to determine which variables out of artist popularity, energy, and BPM is the best predictor for a song's overall popularity. In order to determine this, we used chi-squared hypothesis testing and correlation tests to determine whether these variables were independent of the song's popularity. For each chi-squared test, the null hypothesis is that the variable is independent of the target variable. If they are not independent, that means the variable is likely a predictive variable of the song's popularity. For each correlation test, the null hypothesis will be that there is no correlation between the variable and the target variable.

For artist popularity, the chi-squared test provided a significantly low p-value, below .001. This is below the threshold of .05, which means that we reject the null hypothesis of independence, and that there is likely a relationship between artist popularity and a song's popularity. The correlation test we ran confirmed this, as the p-value on this test is also less than .001, meaning we reject the null hypothesis that there is no correlation between the variables.

For energy, the chi-squared test provided a p-value of .0875, which is greater than our threshold of 0.05. Due to this, we fail to reject the null hypothesis of independence. Looking at the correlation test, we see a p-value of .1945, so we also fail to reject this null hypothesis of no correlation.

For BPM, the chi-squared test provided a p-value of .3629. This is significantly greater than the threshold of .05, meaning we fail to reject the null hypothesis of independence. The correlation test presents a p-value of .3103. This means we fail to reject the null hypothesis of no correlation between the variables.

3.3 Evaluate Artist.Popularity, Valence, and Length..Duration

We created a binary column that labeled each artist as either popular or not popular to get a better understanding of the relation between artists' popularity and song popularity. We defined that artists' popularity scores less or equal to 50 as not popular and larger than 50 as popular. Based on our criteria, 16 song artists are labeled as not popular and 132 song artists are labeled as popular. Table 3 displays a dplyr summary table of song popularity by artist popularity.

Table 3

ArtistEvaluation	Total Songs	Min	Avg	Max
Not popular	16	36	46.8	59
Popular	132	16	59.7	84

We created a ValenceEvaluation binary column that labeled valence as either positive or not positive to get a better understanding of the relation between valence and song popularity. We defined the valence score less or equal to 50 as not positive and larger than 50 as positive. Based on our criteria, 82 song valences are labeled as not positive and 66 song valences are labeled as positive. Table 4 displays a dplyr summary table of song popularity by valence.

Table 4

ValenceEvaluation	Total Songs	Min	Avg	Max

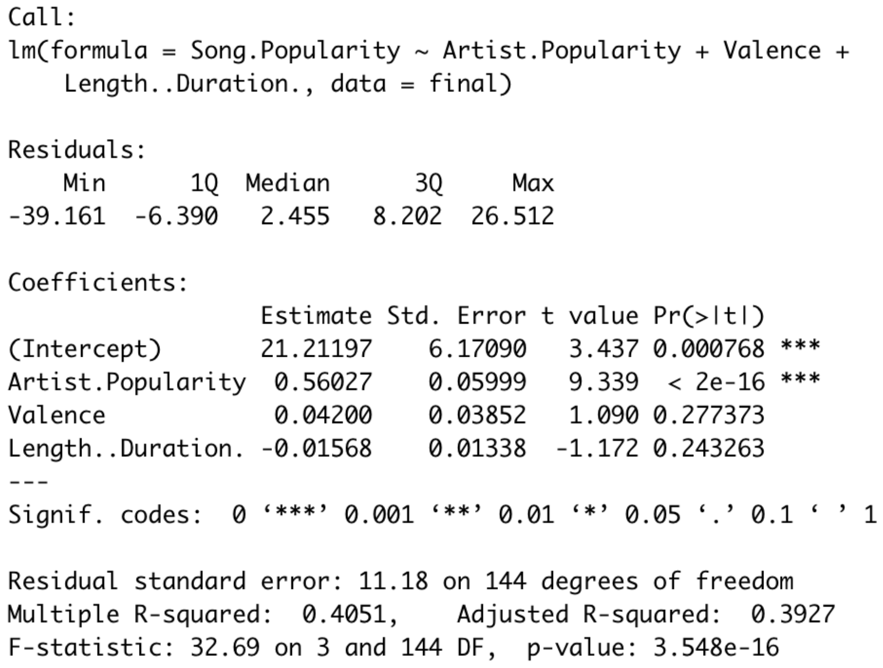
Not positive	82	19	57.2	84
Positive	66	16	59.6	82

From our data dictionary, we knew that song lengths were now scored from 0-100, so we used the qwraps2 package to create a summary table for length. Table 5 shows the summary of the length of our 148 songs.

Table 5

	Final (N=148)
Min length	144
Max length	639
Mean length sd	258.22 ± 70.07

Since there are no categorical independent variables in our dataset, we used a linear regression model to examine the impact of Artist.Popularity, valence, and Length..Duration on the outcome of our target variable.



```
Call:
lm(formula = Song.Popularity ~ Artist.Popularity + Valence + 
    Length..Duration., data = final)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.161  -6.390   2.455   8.202  26.512 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       21.21197    6.17090   3.437 0.000768 ***
Artist.Popularity  0.56027    0.05999   9.339  < 2e-16 ***
Valence            0.04200    0.03852   1.090 0.277373    
Length..Duration. -0.01568    0.01338  -1.172 0.243263    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.18 on 144 degrees of freedom
Multiple R-squared:  0.4051,	Adjusted R-squared:  0.3927 
F-statistic: 32.69 on 3 and 144 DF,  p-value: 3.548e-16
```

Fig 1.

Figure 1 coefficients show that when Artist.Popularity, valence, and length are 0, song popularity is 21.21197. Increasing Artist. Popularity by 1 will increase song popularity by 0.56027.

Increasing songs' valence by 1 will increase the song popularity by 0.04200. Increasing songs' duration by 1 will decrease song popularity by -0.01568. Then we checked the linearity of each independent variable.

Figure 1 shows that there is no linear relationship between song popularity and songs' valence. Figure 2 shows that there is a linear relationship between song popularity and artists' popularity. The figure also indicates a positive relationship that song popularity tends to increase with the increase of artist popularity.

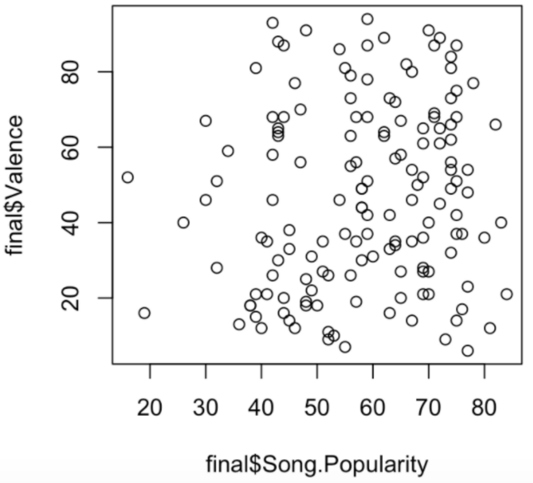


Figure 1

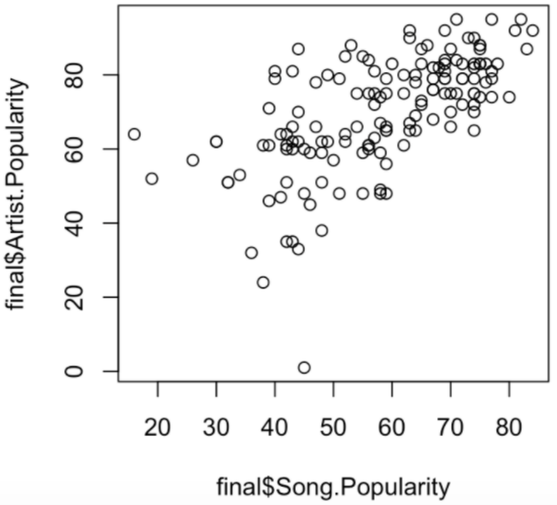


Figure 2

Figure 3 shows that there is a linear relationship between song popularity and the length of the songs. The figure also illustrates that song popularity tends to increase when the length of the song decreases.

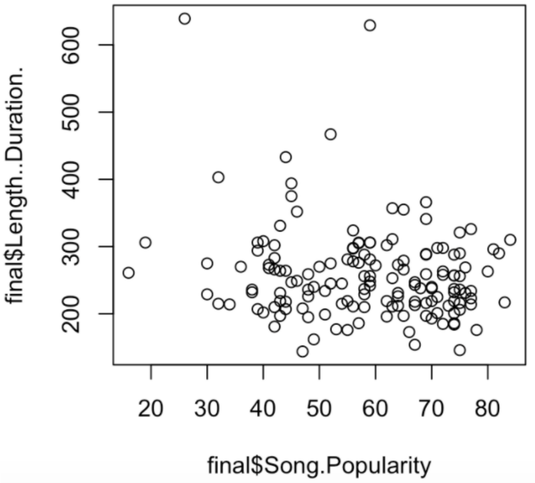
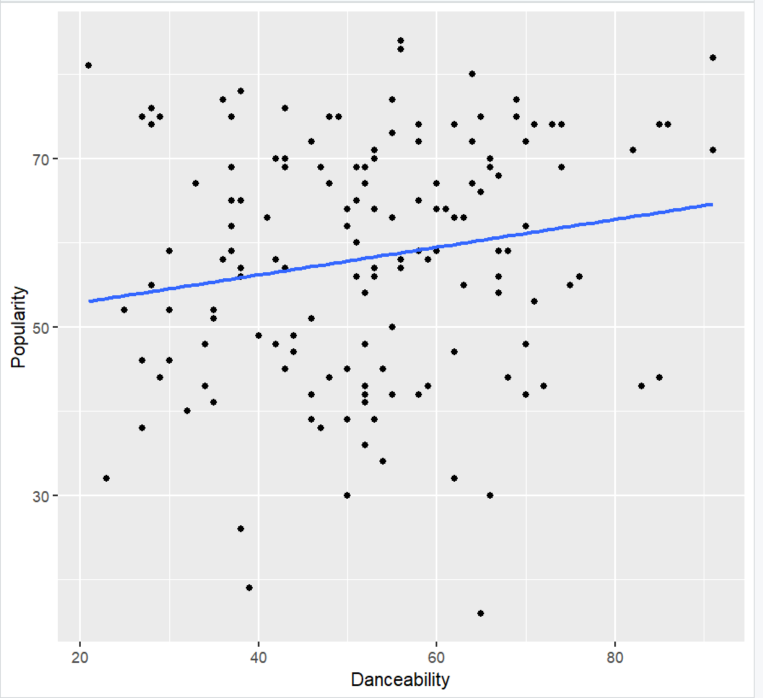
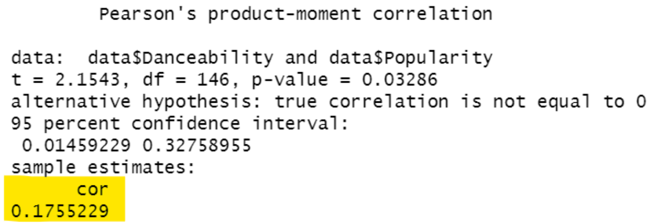


Figure 3

3.4 Evaluate Danceability, Acousticness, and Song Popularity

What impact does danceability and acousticness have on a songs popularity? In order to determine the relationship that each variable has on our target variable, popularity, we created scatter plot visualizations. The figure below graphs danceability against popularity for each song in our dataset. The plot appears to be mostly random, however, when inserting a line of best fit we can conclude that there is a weak, positive relationship between danceability and popularity. This means that the higher danceability scores are associated with higher popularity scores. The correlation coefficient when comparing danceability and popularity is 0.1755229, which supports our conclusion that danceability and popularity have a weak, positive relationship.

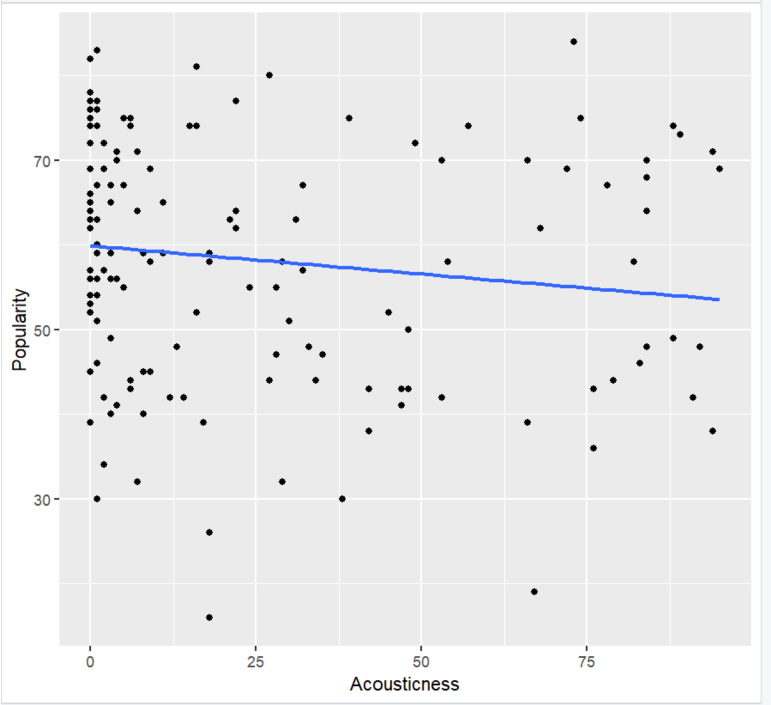


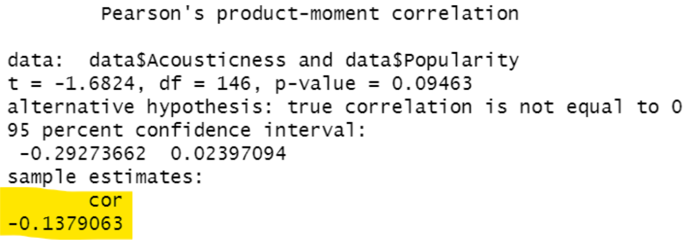


```
        Pearson's product-moment correlation

data:  data$Danceability and data$Popularity
t = 2.1543, df = 146, p-value = 0.03286
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01459229 0.32758955
sample estimates:
      cor
0.1755229
```

We also examined the relationship between acousticness and popularity. It is visualized with the scatter plot below. This distribution is comparable to the danceability vs. popularity scatter plot because there is no obvious pattern within the data points. The dataset also appears to contain many songs that have an acousticness score of 0, causing more of the data points to be on the left side of the scatter plot. After inserting a line of best fit, we determined that acousticness and popularity have a weak, negative relationship. This means that as a song's acousticness score increases, the songs popularity tends to decrease. The correlation coefficient aligns with this conclusion because it is -0.1379063.



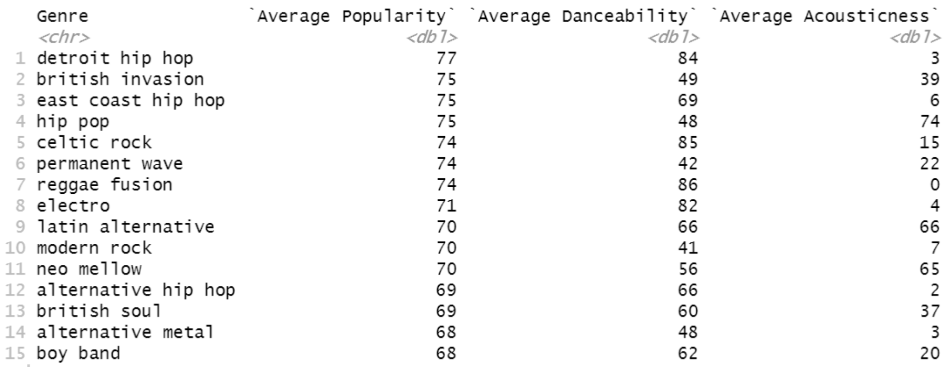


```
	Pearson's product-moment correlation

data:  data$Acousticness and data$Popularity
t = -1.6824, df = 146, p-value = 0.09463
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.29273662  0.02397094
sample estimates:
       cor
-0.1379063
```

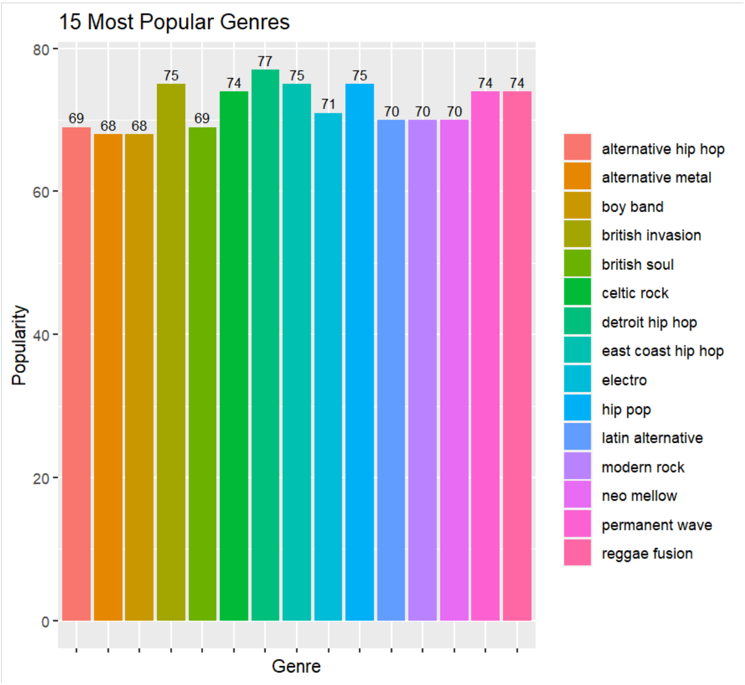
3.5 Top Genres

What genres are most popular? Determining the significance of certain numeric variables in our dataset did provide useful insight, but we also wanted to examine the relationship the genre has with song popularity, danceability, and acousticness. Below is a pivot table that is grouped by the top 15 most popular genres. The table also includes each genre's average popularity, average danceability, and average acousticness scores. Earlier we determined that danceability has a weak, positive relationship with popularity and acousticness has a weak, negative relationship with popularity. Each of these relationships are represented in the pivot table.



```
   Genre                 `Average Popularity` `Average Danceability` `Average Acousticness`
   <chr>                                <dbl>                  <dbl>                  <dbl>
 1 detroit hip hop                         77                     84                      3
 2 british invasion                        75                     49                     39
 3 east coast hip hop                      75                     69                      6
 4 hip pop                                 75                     48                     74
 5 celtic rock                             74                     85                     15
 6 permanent wave                          74                     42                     22
 7 reggae fusion                           74                     86                      0
 8 electro                                 71                     82                      4
 9 latin alternative                       70                     66                     66
10 modern rock                             70                     41                      7
11 neo mellow                              70                     56                     65
12 alternative hip hop                     69                     66                      2
13 british soul                            69                     60                     37
14 alternative metal                       68                     48                      3
15 boy band                                68                     62                     20
```

To visualize the top 15 genres from our dataset we created a barplot using the Genre and Average Popularity columns from the pivot table. This allowed us to conclude that the top genre from our dataset is Detroit Hip Hop. There are also three different types of hip hop genres in the top 15.



4. Conclusion

This project incorporates data from Kaggle on Spotify's top 2000 songs from 1956-2019 with data scraped from Spotify API. Our final dataset only contains Spotify's top songs from 2000-2009 to analyze which features have significant relationship with song popularity. We performed 4 analyses considering whether the artist popularity, song energy, BPM, valence, length, danceability, acousticness had a significant impact with songs popularity through summary statistics, hypothesis testing, linear regression, and visualizations. The first analysis

focuses on the relationship between artist popularity and the popularity of their songs. Through hypothesis testing, we see that artist popularity is correlated and not independent of a song's popularity, as we rejected null hypotheses for both tests. These hypothesis tests show artist popularity as a seemingly good predictive variable for determining a song's popularity. Artist popularity was also tested as a high attributor to song popularity in our linear regression model. Overall, artist popularity had a significant impact on song popularity in our dataset compared to other variables. The second analysis focuses on energy and BPM and how those variables influence the target variable of a song's popularity. Through hypothesis tests, we see that neither energy or BPM could be determined to be correlated or not independent of a song's popularity. Therefore, the hypothesis tests have shown energy and BPM as poor predictive variables for the target variable of a song's popularity. The third analysis focused on the whether valence and song length are song attributors to song popularity. Length had a negative linear relationship with song popularity and there is no relationship between song valence and song popularity. Song popularity will tend to have lower scores when the length of the songs are too long. Therefore, artists should consider song length when making music. Through visualizations, we determined that danceability has a weak, positive relationship with song popularity and acousticness has a weak, negative relationship with song popularity. These relationships were concluded through scatter plots. This was also seen within each genre of music in our dataset. Lastly, through a bar plot, we determined that the hip hop genre is very prevalent in our dataset, with Detroit Hip Hop having the highest average song popularity.
Our data has limitations on having a small dataset of only 148 observations and only includes the songs from 2000-2009. We could get a better result if we include more observations from a larger year range.