Chelsea Alford
Marlon Do Couto
Yujia Huo
Richard Locarni

Spotify Predictive Model:

Data Mining for the Next Big Hit

Business Understanding

In a world with billions of songs written by billions of different artists across time and around the world, what makes a song popular? What perfect cadences and synchronized instrumentation combines to create the next top hit? Everyone in the music industry is working to answer this question, because there is significant profit in making and producing popular music. In the last decade, the top earning musician, Dr. Dre, made $950 million dollars (Greenburg) . So, with so much money in the music industry to be made, how can producers and artists come closer to making the next big hit?

As a team of music-lovers, we set out to answer this question. Our team decided to look into data from Spotify, a popular music streaming app. With 248 million monthly active listeners (Silva), Spotify's data promises a plethora of information about what people are listening to. Through the data, we want to identify why Spotify users are listening to certain songs over others. If we could accurately identify the musical attributes that make up a top hit, then we could build a model to predict the popularity of songs. This model would be useful to everyone in the music industry, from singers, musicians, djs, and producers, and it would be extremely valuable. Given the almost universal market that the music industry has, making something that is popular among most listeners promises high returns. Therefore, we set out to mine Spotify's data and eventually build a model that could predict the popularity of any given song, to help music writers, artists, and producers create more popular and successful music.

Data Understanding

We found our Spotify data on Kaggle, a data science company owned by Google that is home to countless datasets published by various data scientists. The dataset includes five different tables which segment the data in different ways. There is one main table with all the data, one with data segmented by artist, one segmented by year, and two tables with data segmented by genre. We used different tables within the database depending on what we were analyzing in the data, but for the most part we decided to focus on the genre datasets because this information proved the most useful (more on this in the data preparation section).

The data has information on Spotify's songs based on 18 different musical variables: duration, key, mode, time signature, acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, valence, tempo, track id, uri, track reference, audio analysis, and type **(Figure 1)**. We did not use any of the last 5 variables (track id, uri, track reference, audio analysis, and type) because these variables all had to do with how each song is stored in Spotify's database and therefore irrelevant to the analysis which we hope to achieve. In our analysis, we focused on the variables in the data which contribute to the musical quality of a particular song like energy and acousticness. Many of these variables represent scales in which a song receives a score based on a certain musical quality. For example, acousticness measures the level of acoustic sound quality in a particular song.

Though some of these variables like danceability or instrumentalness may seem like subjective qualities, we perceive little to no bias in the Spotify data. These variables are determined based on a scientific analysis of music and though Spotify does not disclose the exact method used to

assign different songs with values for these variables, the explanation provided on their website (link in works cited) provides us with confidence that the data is unbiased.

Data Preparation

Before starting our analysis, we performed a series of data preparation techniques to ensure that the data we used was devoid of missing or incomplete records, improperly formatted or structured data, and inconsistent values. First, we examined the missing records for both data sets that we used ("spotify.csv "and "data_w_genres.csv"). Neither data set had missing or incomplete records. We also checked if there were any improperly formatted or structured data points by summarizing all the variables in the data sets. We observed that the format of observations in the release date variable were inconsistent. Some of the formats were "Year-Month-Date" while others contained just "Year". Finally, we checked for any abnormal values.

To determine the dataset for our final regression model, we first filtered all records in 2019 from "spotify.csv". However, after looking into variables such as duration, danceability, energy, liveness, and loudness, we did not spot any general correlations between song popularity and these variables. Therefore, we decided against using the dataset segmented by year and instead chose to use 27,621 observations from "data_w_genres.csv". When we looked into the genres variable, we noticed that there were too many different categories to perform a successful analysis. With 2,663 genres, we would not be able to glean any useful conclusions due to the over-specific segmentation. As shown in **Figure 11**, we segmented the data into four basic categories: Pop, Rock, Rap, Jazz. Since we assume that different styles of music must have different criteria for becoming popular, looking into different genres would be more meaningful for our analysis.

Modeling

With the overall goal of creating a model to predict the popularity of songs within various genres based on certain musical characteristics, we started modeling with a simple approach. As previously mentioned, we divided our analysis into four main genre categories: Jazz, Pop, Rap and Rock. This division was essential in our modeling, allowing us to find better correlations between variables and improve our analysis.

In our first model **(Figure 7)**, we analyzed the following variables: danceability, energy, liveness, and loudness. We found an R-squared of 0.1908 for the Rock model, 0.2243 for Jazz, 0.2353 for Pop, and 0.2774 for Rap. To strengthen our model, we decided to explore which other variables might correlate with popularity.

We analyzed which variables had the strongest correlation within each genre by using a correlation matrix **(Figure 5)** including every variable in the dataset. Since each genre differs in style and sound from one another, we assumed the most significant variables would be different across various genres. Looking particularly at the popularity response variable, we noticed a wide range of correlations across different genres. In Jazz, energy had the strongest positive correlation with popularity at 0.31 while acousticness had the strongest negative correlation at -0.5. The same is true of Pop, in which the strongest positive correlation with popularity is energy (0.32) while acousticness had the strongest negative correlation (-0.42). This pattern is also observed in Rock songs. We observed from the correlation matrices that some of the variables not previously explored could have an impact in the response variable.

To further improve our model, we decided to include all other variables contained in the Spotify data, adding acousticness, instrumentalness, speechiness, tempo, and valence. The inclusion of the entire dataset significantly improved our models for three genres: Rock, Jazz, and Pop. The R-squared for the new complete model is 0.2906 for Rock, 0.387 for Jazz, 0.3565 for Pop. For rap, the new model R-squared is at 0.2617, showing 0.0157 less correlation than the previous model with fewer independent variables. Therefore, we continued our observations with the original rap model **(Figure 8)**.

To further improve the models, we also worked with transformations. Our first step was to filter out songs with zero popularity in the Spotify data. This served two purposes: to account only for songs that had streamings and to allow us to apply the Box-Cox transformation method. We found λ values of 1.15, 0.83, 1.39, and 1.64 for Rock, Jazz, Pop, and Rap, respectively **(Figure 9)**. We also tried performing interactions for all four models, but none of the interactions performed were statistically significant.

Our findings within the models can be seen in **Figure 10** and are summarized as follows:

- Rock: The final R-squared is 0.2275. Because the λ for rock was so close to 1, we decided to keep the model untransformed. Additionally, we eliminated duration since the variable does not have a significant impact on the popularity value. AIC for the model without duration is also lower at 31,500 versus 31,502 for the previous model with the variable.

- Jazz: Because the λ value for the jazz genre was close to 1, we also decided not to transform this model. After eliminating independent variables speechiness, tempo, and instrumentalness for their high p-values, the final R-squared is 0.38. However, AIC for the first complete model is lower at 8,211 against 8,215 for the second model, which did not

include speechiness, tempo, and instrumentalness. Therefore, we decided to keep the first model with an R-squared of 0.387, since these three variables appear to have some explanatory value to the model.

- Rap: After the transformation, the new R-squared for this model is 0.2749. We observed that after eliminating independent variables with high p-values such as speechiness, we did not lose much explanatory power of the model (R-squared lowered to 0.2745). The AIC for both models was similar at 22,593 and 22,592. We recommend the use of the second transformed model without speechiness.

- Pop: Because the λ value is again within the range of 1, we decided not to transform the model. The final R-squared is 0.3565. No variable was eliminated in the process due to a high p-value.

Evaluation

When starting our Spotify data analysis, we expressed reservations about the predictive accuracy of any potential model. These reservations were based on our knowledge of the music industry and the absence of a particular "formula" to produce a popular song. However, after stratifying our data by genre and performing power transformations, we developed models with greater predictive accuracy than expected.

Our team developed four models in this analysis, one each for Rock, Rap, Jazz, and Pop. Our model's predictive power ranged from a low R-squared of .2275 for Rock to a high R-squared of .387 for Jazz. Considering the somewhat random nature of popular music, we believe that even a

model with an R-Squared of .2275 would be helpful to industry professionals. With this model, artists and producers gain insight into what drives success in a given musical genre. Although our model cannot guarantee the success of any particular song, it predicts a song's highest probability of success.

According to our regression models, the three variables that most impact popularity within Jazz music are: valence, danceability and liveness. Danceability has a positive correlation with popularity for jazz music, while valence and liveness show a negative correlation. Though Jazz music is most often performed and listened to live, the model shows that the higher the liveness quality of a jazz song, the less popular it is likely to be. This may be because the sound quality of live music is worse than studio records on average and most people on Spotify are looking for a quality soundtrack. Therefore, a record producer for Jazz should follow this trend, seeking a higher quality sound than live jazz as well as balancing the blues vibe and danceability of Jazz music.

Similar to Jazz, danceability and valence also are the most impactful variables on the popularity of Rock music. People who love Rock expect to hear less valence and more danceability from a track. As opposed to Jazz, Rock fans prefer fewer spoken words in a soundtrack. A probable explanation for this is that Rock fans may not care about the content of the lyrics; instead, they appreciate the emotion conveyed from the music more. Therefore, a music producer should focus on the emotion delivered by Rock music.

Valence and danceability also have a significant impact on the popularity of Pop music **(Figure 2)**. Additionally, Pop fans prefer a more energetic track. Record producers should focus on producing an energetic track, whether through rhythmic beats or exciting lyrics, to make their Pop song a big hit.

When producing a Rap song, valence and danceability are the two factors that should be considered most **(Figure 2)**. Similarly to the other three genres, Rap fans favor music with less valence and more danceability. Moreover, Rap fans prefer a track containing less instrumentation, making instrumentalness another impactful variable on the popularity of Rap songs. Therefore, Rap producers should pay attention to lyrical quality and avoid heavy instrumentation in their songs.

Admittedly, there are some pitfalls in our model. First, the model in its current state cannot reflect the ever-changing dynamics of consumer taste in the music industry. It would be beneficial to include a machine learning element in the model to update consumer preferences in real time. Furthermore, developing a "catch-all" type of model for music production may dissuade record companies from signing otherwise popular, unconventional artists that do not fit the standard definition of popular music. Since the model would likely not predict popularity for such unconventional artists, it would miss this particular opportunity for a significant increase in revenues.

Additionally, if a large amount of popular music follows a model, consumers may get tired of the model's musical style. Only catering to market preferences might undermine the overall quality of music and betray music producers' professional intuition. The diversity of the music in the industry would be limited, thus compromising innovation of music across genres. This development would render the original model useless and without a dataset reflective of new consumer tastes, it may be difficult to construct a new model.

Deployment

Businesses can implement our model in a variety of ways. Major music companies, such as Universal Music Group and Sony Music Entertainment, can use this model to decide which artists to sign. For example, if a company is deciding between two rap artists, it can pick the artist whose sound best fits our model to have the greatest chance of success. The selection of artists may be more beneficial to smaller record companies, such as Carpark Records and Third Man Records, as these firms have limited resources to acquire talent. Furthermore, these smaller companies tend to deal with artists that are not already popular and therefore cannot rely on reputation to increase popularity. Therefore, our selected variables are of great importance to small record firms.

Our model is also beneficial at micro-levels of music production. For example, an artist can change elements of a specific song during recording to better fit our model. As a result, the song will have a greater chance of becoming popular than if it was unchanged.

Return on investment for companies that implement this model is hard to calculate without a large future dataset. Due to the unpredictability of the music industry and our model's relatively low R-squared value, it is unlikely that all firms will see immediate improvement in revenues directly after implementation. However, a firm using our model consistently for multiple years will likely realize greater returns than firms not using our model. Once the model has been used, we can collect data to quantify its success. Then determining a return on investment figure and required rate of return would be somewhat simple and we would adjust the price of licensing our model accordingly.

Return on investment will also be impacted based on the scale with which our models are implemented into a business. It would be possible for a company to use our model for artists with a large number of songs to determine which variables are most important to that particular artist. This would be helpful, as fans of a certain artist tend to have similar tastes, allowing this model to cater new songs directly to this artist's fanbase.

Appendix

Figure 1: Variables Description

Key	Value Type	Value Description
duration_ms	Int	The duration of the track in milliseconds.
Key	Int	The estimated overall key of the track. Integers map to pitches using standard Pitch Class Notation.
Mode	Int	Mode indicates the modality of a track (major or minor), the type of scale in which its melodic content is derived. Major represented by 1, minor represented by 0.
acousticness	float	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
danceability	float	Danceability describes how suitable a track is for dancing based on a combination of tempo, rhythm stability, beat strength and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
Energy	float	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.

instrumentalness	float	Predicts whether a track contains no vocals. The closer to 1.0, the greater likelihood the track contains no vocals.
Liveness	float	Detects the presence of an audience in the recording. Higher liveness values represent and increased probability the track was performed live.
Loudness	float	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness. Values typically range between -60 and 0 dB.
speechiness	float	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording, the closer to 1.0. Values between 0.33 and 0.66 describe tracks that contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and non-speech-like tracks.
Valence	float	A measure from 0.0 to 1.0 describing musical positiveness conveyed by the track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Tempo	float	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology tempo is the speed or pace of a given piece.

Figure 2: Danceability Across Genres

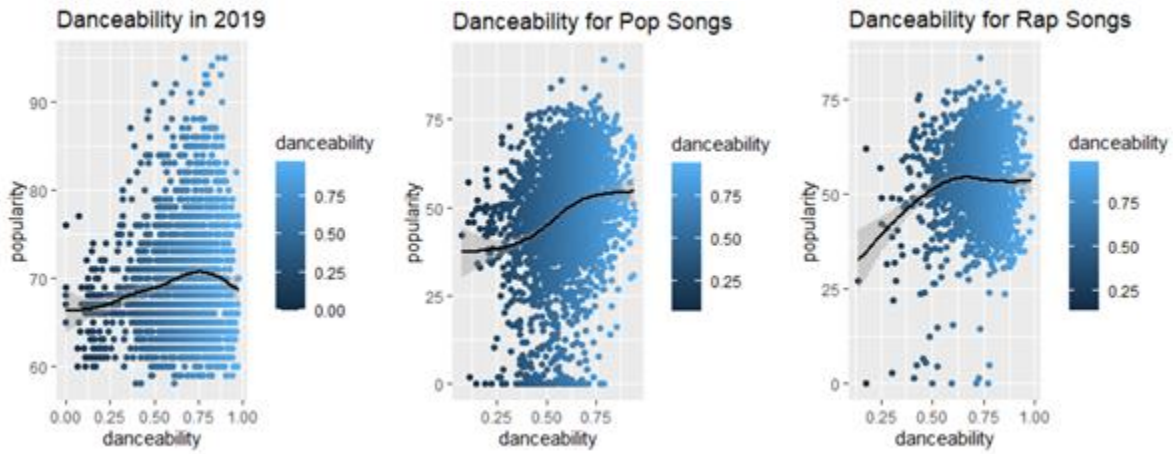


Figure 3: Popularity and Duration

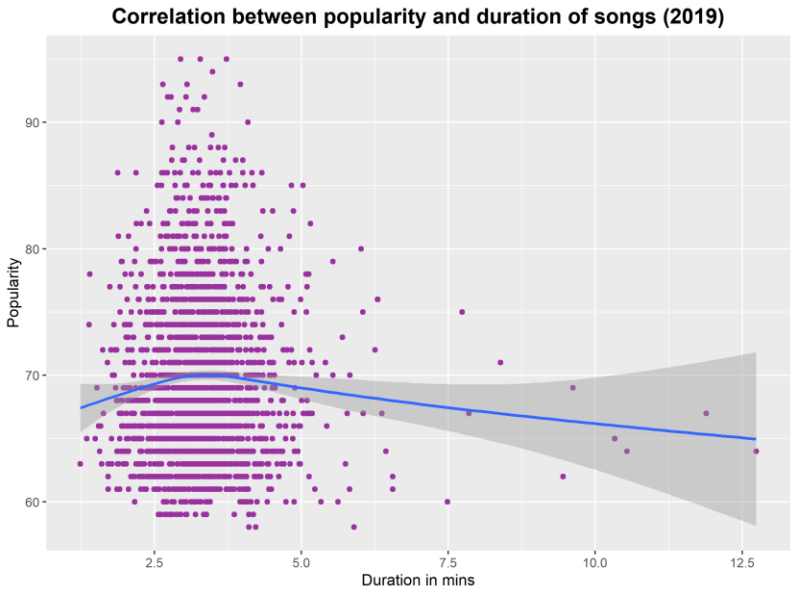
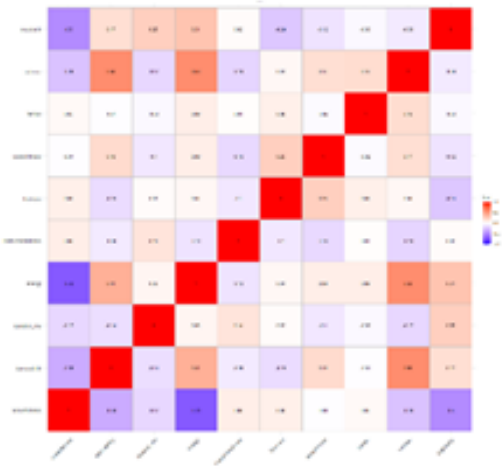


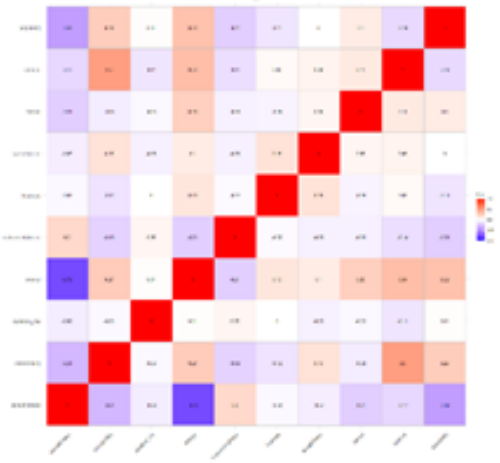
Figure 4: Correlation with Popularity

	Variable	Corr_to_popularity
1	explicit	0.101986982
2	danceability	0.094068757
3	speechiness	0.035280525
4	loudness	0.019618968
5	tempo	0.002087824
6	valence	0.001678888
7	key	-0.011470126
8	instrumentalness	-0.020991792
9	energy	-0.026314457
10	mode	-0.031425554
11	liveness	-0.034425273
12	acousticness	-0.041603899
13	duration_ms	-0.122831592
14	duration_min	-0.122831592
15	year	-0.208615705

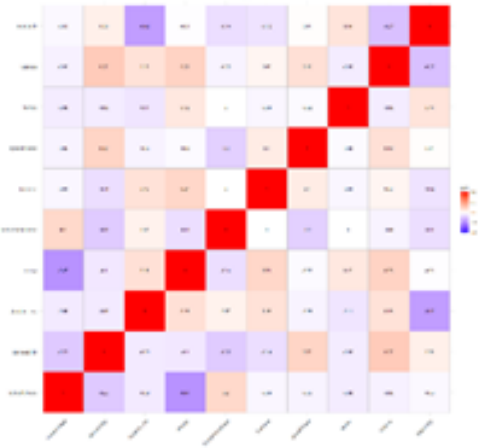
Figure 5: Correlation Matrices by Genre



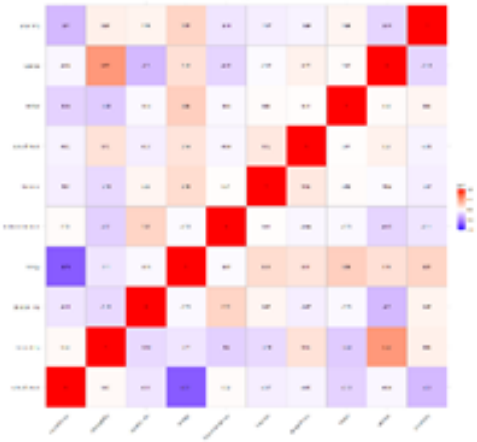
Correlation Matrix for Jazz Genre



Correlation Matrix for Pop Genre

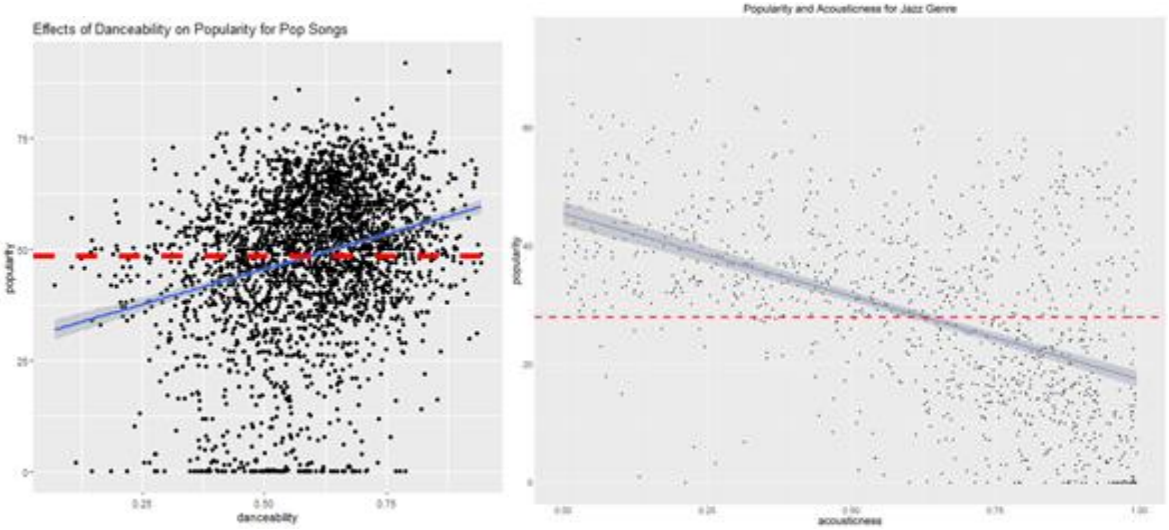


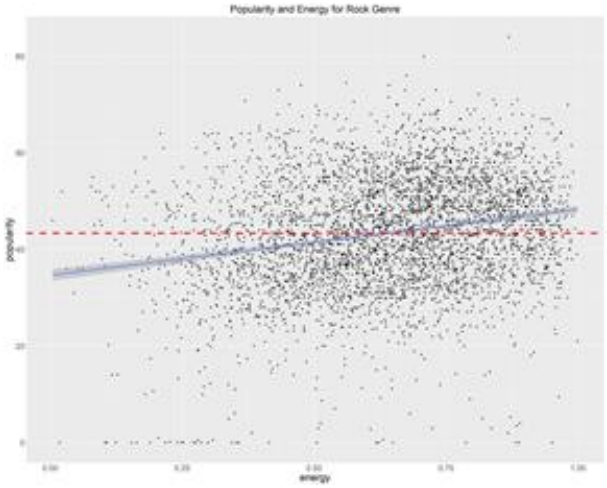
Correlation Matrix for Rap Genre



Correlation Matrix for Rock Genre

Figure 6: Simple Linear Models by Genre




Popularity and Energy for Rock Genre
popularity
energy

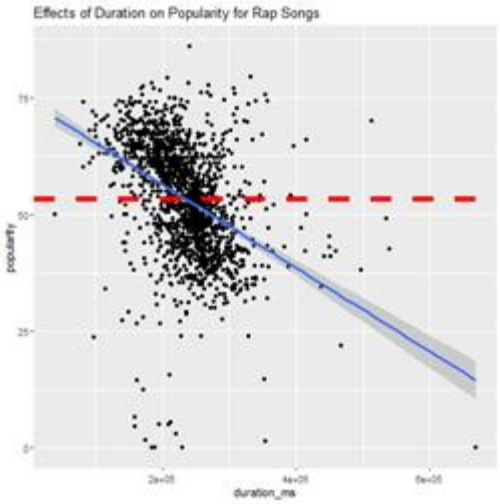

Effects of Duration on Popularity for Rap Songs
popularity
duration_ms

Figure 7: Initial Models by Genre

```
Call:
lm(formula = popularity ~ acousticness + danceability + duration_ms +
    energy + instrumentalness + liveness + loudness + speechiness +
    tempo + valence + popularity, data = rock_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-54.256  -5.753  -0.288   6.093  56.711

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.006e+01  2.270e+00  26.461  < 2e-16 ***
acousticness     -9.860e+00  9.144e-01 -10.782  < 2e-16 ***
danceability      2.424e+01  1.611e+00  15.048  < 2e-16 ***
duration_ms       3.593e-06  2.414e-06   1.488  0.13673
energy           -9.293e+00  1.738e+00  -5.346 9.47e-08 ***
instrumentalness -2.424e+00  8.945e-01  -2.709  0.00677 **
liveness         -4.145e+00  1.403e+00  -2.955  0.00314 **
loudness          1.395e+00  7.232e-02  19.295  < 2e-16 ***
speechiness      -2.315e+01  3.000e+00  -7.716 1.49e-14 ***
tempo             2.622e-02  8.040e-03   3.261  0.00112 **
valence          -1.939e+01  1.056e+00 -18.351  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.877 on 4186 degrees of freedom
Multiple R-squared:  0.2906,    Adjusted R-squared:  0.2889
F-statistic: 171.5 on 10 and 4186 DF,  p-value: < 2.2e-16
```

First Rock model with all variables included

```
Call:
lm(formula = popularity ~ acousticness + danceability + duration_ms +
    energy + instrumentalness + liveness + speechiness + tempo +
    valence + popularity, data = only_jazz_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-38.633  -8.708  -0.946   7.369  41.726

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.994e+01  4.173e+00   7.174 1.39e-12 ***
acousticness     -1.913e+01  2.183e+00  -8.764  < 2e-16 ***
danceability      2.553e+01  4.292e+00   5.947 3.72e-09 ***
duration_ms       2.382e-05  4.258e-06   5.595 2.83e-08 ***
energy            1.655e+01  3.460e+00   4.783 1.97e-06 ***
instrumentalness -1.531e+00  1.320e+00  -1.160   0.2461
liveness         -2.341e+01  3.513e+00  -6.664 4.34e-11 ***
speechiness      -1.744e+01  7.575e+00  -2.302   0.0215 *
tempo             4.050e-02  2.075e-02   1.952   0.0512 .
valence          -2.963e+01  2.867e+00 -10.334  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.33 on 1033 degrees of freedom
Multiple R-squared:  0.387,     Adjusted R-squared:  0.3817
F-statistic: 72.46 on 9 and 1033 DF,  p-value: < 2.2e-16
```

First Jazz model with all variables included

```
Call:
lm(formula = popularity ~ acousticness + danceability + duration_ms +
    energy + instrumentalness + liveness + speechiness + tempo +
    valence + popularity, data = only_pop_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-59.504  -8.215   0.534   9.189  42.230

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.808e+01  2.785e+00  13.674  < 2e-16 ***
acousticness     -1.228e+01  1.407e+00  -8.729  < 2e-16 ***
danceability      4.288e+01  2.249e+00  19.072  < 2e-16 ***
duration_ms      -1.438e-05  4.215e-06  -3.412 0.000654 ***
energy            1.460e+01  2.064e+00   7.070 1.91e-12 ***
instrumentalness -1.240e+01  1.507e+00  -8.228 2.80e-16 ***
liveness         -1.141e+01  2.213e+00  -5.157 2.67e-07 ***
speechiness      -1.462e+01  3.888e+00  -3.759 0.000174 ***
tempo             5.834e-02  1.195e-02   4.883 1.10e-06 ***
valence          -3.547e+01  1.403e+00 -25.288  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.58 on 3021 degrees of freedom
Multiple R-squared:  0.3565,    Adjusted R-squared:  0.3546
F-statistic:   186 on 9 and 3021 DF,  p-value: < 2.2e-16
```

First Pop model with all variables included

```
Call:
lm(formula = popularity ~ duration_ms + energy + liveness + loudness,
    data = only_rap_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-63.222  -6.019   0.500   6.362  38.374

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.973e+01  2.356e+00  42.321  < 2e-16 ***
duration_ms -7.898e-05  4.601e-06 -17.164  < 2e-16 ***
energy      -2.016e+01  2.500e+00  -8.063  1.4e-15 ***
liveness    -5.749e+00  2.258e+00  -2.546    0.011 *
loudness     1.888e+00  1.255e-01  15.038  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.04 on 1684 degrees of freedom
Multiple R-squared:  0.2773,    Adjusted R-squared:  0.2756
F-statistic: 161.5 on 4 and 1684 DF,  p-value: < 2.2e-16
```

First Rap model with all variables included

Figure 8: Secondary Models by Genre

```
Call:
lm(formula = popularity ~ danceability + duration_ms + energy +
    liveness + loudness, data = rock_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-53.978  -6.081   0.065   6.354  62.073

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.859e+01  1.683e+00  34.819  < 2e-16 ***
danceability  7.726e+00  1.258e+00   6.140 9.00e-10 ***
duration_ms   2.392e-05  2.414e-06   9.909  < 2e-16 ***
energy       -1.185e+01  1.286e+00  -9.221  < 2e-16 ***
liveness     -6.905e+00  1.401e+00  -4.928 8.64e-07 ***
loudness      1.810e+00  6.782e-02  26.684  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.989 on 4164 degrees of freedom
Multiple R-squared:  0.2078,    Adjusted R-squared:  0.2069
F-statistic: 218.5 on 5 and 4164 DF,  p-value: < 2.2e-16


Call:
lm(formula = popularity ~ acousticness + danceability + duration_ms +
    energy + instrumentalness + liveness + loudness + speechiness +
    tempo + valence + popularity, data = rock_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-42.524  -5.791  -0.406   5.910  57.120

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.047e+01  2.197e+00  27.527  < 2e-16 ***
acousticness     -8.261e+00  8.861e-01  -9.323  < 2e-16 ***
danceability      2.361e+01  1.547e+00  15.266  < 2e-16 ***
duration_ms       7.028e-06  2.458e-06   2.859  0.00427 **
energy           -1.056e+01  1.677e+00  -6.296 3.37e-10 ***
instrumentalness -2.136e+00  8.623e-01  -2.477  0.01329 *
liveness         -3.965e+00  1.343e+00  -2.953  0.00317 **
loudness          1.469e+00  6.972e-02  21.070  < 2e-16 ***
speechiness      -1.590e+01  3.012e+00  -5.279 1.37e-07 ***
tempo             2.192e-02  7.724e-03   2.837  0.00457 **
valence          -1.866e+01  1.017e+00 -18.351  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.435 on 4159 degrees of freedom
Multiple R-squared:  0.2941,    Adjusted R-squared:  0.2925
F-statistic: 173.3 on 10 and 4159 DF,  p-value: < 2.2e-16
```

Models for Rock after filtering for popularity greater than zero: on top the model with preliminary variables; in the bottom, the model with all variables included.

```
Call:
lm(formula = popularity ~ danceability + duration_ms + energy +
    liveness + loudness, data = only_jazz_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-30.431  -9.041  -0.989   7.922  45.745

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.519e+00  4.367e+00   1.264    0.207
danceability  1.052e+01  4.181e+00   2.517    0.012 *
duration_ms   3.172e-05  4.735e-06   6.699 3.51e-11 ***
energy        2.500e+01  3.600e+00   6.945 6.83e-12 ***
liveness     -2.942e+01  3.834e+00  -7.673 3.99e-14 ***
loudness     -2.645e-01  1.647e-01  -1.606    0.108
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.52 on 999 degrees of freedom
Multiple R-squared:  0.1893,    Adjusted R-squared:  0.1852
F-statistic: 46.65 on 5 and 999 DF,  p-value: < 2.2e-16


Call:
lm(formula = popularity ~ acousticness + danceability + duration_ms +
    energy + instrumentalness + liveness + speechiness + tempo +
    valence + popularity, data = only_jazz_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-33.498  -8.472  -0.906   7.457  40.372

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.914e+01  4.257e+00   6.846 1.33e-11 ***
acousticness     -1.763e+01  2.193e+00  -8.041 2.52e-15 ***
danceability      2.725e+01  4.362e+00   6.247 6.20e-10 ***
duration_ms       1.888e-05  4.294e-06   4.396 1.22e-05 ***
energy            1.619e+01  3.475e+00   4.659 3.61e-06 ***
instrumentalness -1.199e+00  1.321e+00  -0.908   0.3643
liveness         -2.118e+01  3.640e+00  -5.820 7.93e-09 ***
speechiness      -1.313e+01  9.866e+00  -1.331   0.1834
tempo             4.214e-02  2.128e-02   1.981   0.0479 *
valence          -2.956e+01  2.872e+00 -10.294  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.11 on 995 degrees of freedom
Multiple R-squared:  0.3512,    Adjusted R-squared:  0.3454
F-statistic: 59.85 on 9 and 995 DF,  p-value: < 2.2e-16
```

Models for Jazz after filtering for popularity greater than zero: on top the model with preliminary variables; in the bottom, the model with all variables included.

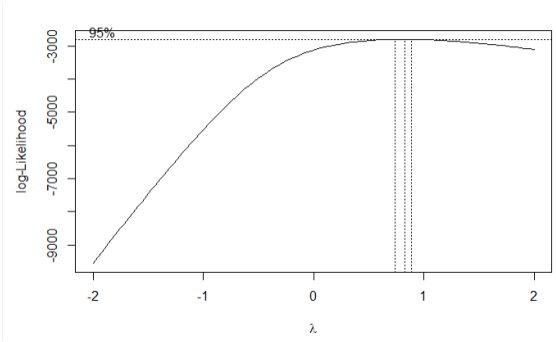
```
Call:
lm(formula = popularity ~ danceability + duration_ms + energy +
    liveness + loudness, data = only_pop_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-59.115  -7.415   1.594   9.975  54.475

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.894e+01  2.496e+00  23.613  < 2e-16 ***
danceability  1.651e+01  2.043e+00   8.081 9.21e-16 ***
duration_ms   1.450e-05  4.564e-06   3.177  0.00150 **
energy       -5.591e+00  2.077e+00  -2.692  0.00713 **
liveness     -1.565e+01  2.353e+00  -6.653 3.40e-11 ***
loudness      2.061e+00  1.124e-01  18.335  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.8 on 3025 degrees of freedom
Multiple R-squared:  0.2353,    Adjusted R-squared:  0.234
F-statistic: 186.1 on 5 and 3025 DF,  p-value: < 2.2e-16


Call:
lm(formula = popularity ~ acousticness + danceability + duration_ms +
    energy + instrumentalness + liveness + speechiness + tempo +
    valence + popularity, data = only_pop_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-59.504  -8.215   0.534   9.189  42.230

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.808e+01  2.785e+00  13.674  < 2e-16 ***
acousticness     -1.228e+01  1.407e+00  -8.729  < 2e-16 ***
danceability      4.288e+01  2.249e+00  19.072  < 2e-16 ***
duration_ms      -1.438e-05  4.215e-06  -3.412 0.000654 ***
energy            1.460e+01  2.064e+00   7.070 1.91e-12 ***
instrumentalness -1.240e+01  1.507e+00  -8.228 2.80e-16 ***
liveness         -1.141e+01  2.213e+00  -5.157 2.67e-07 ***
speechiness      -1.462e+01  3.888e+00  -3.759 0.000174 ***
tempo             5.834e-02  1.195e-02   4.883 1.10e-06 ***
valence          -3.547e+01  1.403e+00 -25.288  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.58 on 3021 degrees of freedom
Multiple R-squared:  0.3565,    Adjusted R-squared:  0.3546
F-statistic:   186 on 9 and 3021 DF,  p-value: < 2.2e-16
```

Models for Pop after filtering for popularity greater than zero: on top the model with preliminary variables; in the bottom, the model with all variables included.

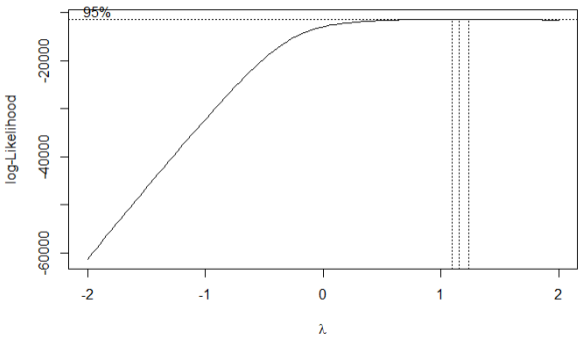
```
Call:
lm(formula = popularity ~ duration_ms + energy + liveness + loudness,
    data = only_rap_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-63.562  -6.037   0.447   6.209  38.259

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.008e+02  2.289e+00  44.042   <2e-16 ***
duration_ms -7.939e-05  4.547e-06 -17.459   <2e-16 ***
energy      -2.198e+01  2.429e+00  -9.049   <2e-16 ***
liveness    -5.418e+00  2.189e+00  -2.475   0.0134 *
loudness     1.849e+00  1.217e-01  15.190   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.718 on 1679 degrees of freedom
Multiple R-squared:  0.282,     Adjusted R-squared:  0.2802
F-statistic: 164.8 on 4 and 1679 DF,  p-value: < 2.2e-16


Call:
lm(formula = popularity ~ acousticness + danceability + duration_ms +
    energy + instrumentalness + liveness + speechiness + tempo +
    valence + popularity, data = only_rap_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-54.113  -5.783   0.120   6.279  37.928

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.398e+01  3.103e+00  20.620  < 2e-16 ***
acousticness      3.797e+00  1.694e+00   2.241  0.02516 *
danceability      1.041e+01  2.338e+00   4.453 9.03e-06 ***
duration_ms      -7.742e-05  4.699e-06 -16.474  < 2e-16 ***
energy            8.847e+00  2.194e+00   4.033 5.75e-05 ***
instrumentalness -1.087e+01  2.328e+00  -4.670 3.25e-06 ***
liveness         -5.812e+00  2.268e+00  -2.563  0.01048 *
speechiness      -7.228e-01  2.362e+00  -0.306  0.75966
tempo             3.649e-02  1.139e-02   3.203  0.00139 **
valence          -1.710e+01  1.527e+00 -11.194  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.888 on 1674 degrees of freedom
Multiple R-squared:  0.2587,    Adjusted R-squared:  0.2548
F-statistic: 64.92 on 9 and 1674 DF,  p-value: < 2.2e-16
```

Models for Rap after filtering for popularity greater than zero: on top the model with preliminary variables; in the bottom, the model with all variables included.

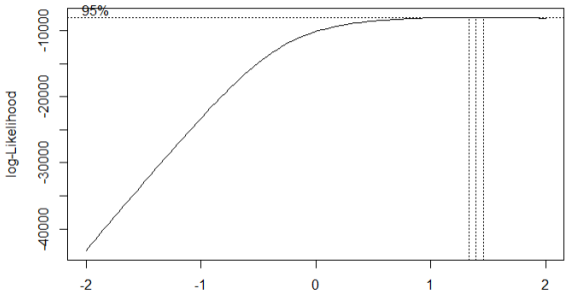
Figure 9: Lambda Values by Genre



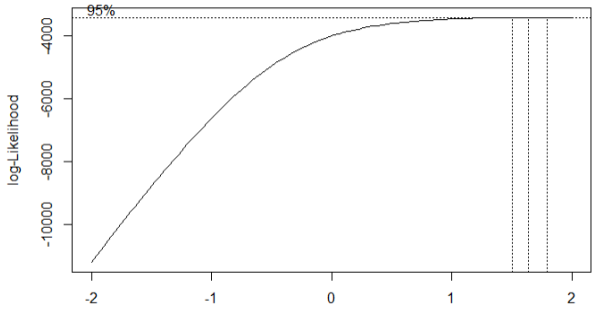
Lambda for Jazz.



Lambda for Rock.



Lambda for Pop.



Lambda for Rap.

Figure 10: Final Models by Genre

```
Call:
lm(formula = (popularity^1.64 - 1/1.64) ~ acousticness + danceability +
    duration_ms + energy + instrumentalness + liveness + speechiness +
    tempo + valence, data = only_rap_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-836.69 -128.22   -8.54  127.11  804.06

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.155e+02  6.198e+01  14.770  < 2e-16 ***
acousticness      1.212e+02  3.384e+01   3.580 0.000353 ***
danceability      2.052e+02  4.671e+01   4.393 1.19e-05 ***
duration_ms      -1.614e-03  9.387e-05 -17.192  < 2e-16 ***
energy            1.612e+02  4.381e+01   3.679 0.000242 ***
instrumentalness -2.150e+02  4.651e+01  -4.623 4.07e-06 ***
liveness         -1.300e+02  4.530e+01  -2.869 0.004173 **
speechiness      -4.331e+01  4.718e+01  -0.918 0.358847
tempo             8.996e-01  2.276e-01   3.953 8.03e-05 ***
valence          -3.389e+02  3.051e+01 -11.107  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 197.5 on 1674 degrees of freedom
Multiple R-squared:  0.2749,    Adjusted R-squared:  0.271
F-statistic: 70.51 on 9 and 1674 DF,  p-value: < 2.2e-16


Call:
lm(formula = (popularity^1.64 - 1/1.64) ~ acousticness + danceability +
    duration_ms + energy + instrumentalness + liveness + tempo +
    valence, data = only_rap_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-832.17 -128.43   -8.22  126.93  809.30

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.116e+02  6.184e+01  14.742  < 2e-16 ***
acousticness      1.215e+02  3.384e+01   3.590 0.000340 ***
danceability      1.991e+02  4.623e+01   4.306 1.76e-05 ***
duration_ms      -1.609e-03  9.371e-05 -17.169  < 2e-16 ***
energy            1.640e+02  4.371e+01   3.752 0.000181 ***
instrumentalness -2.078e+02  4.583e+01  -4.533 6.22e-06 ***
liveness         -1.358e+02  4.485e+01  -3.027 0.002506 **
tempo             8.978e-01  2.275e-01   3.946 8.29e-05 ***
valence          -3.423e+02  3.027e+01 -11.308  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 197.5 on 1675 degrees of freedom
Multiple R-squared:  0.2745,    Adjusted R-squared:  0.2711
F-statistic: 79.23 on 8 and 1675 DF,  p-value: < 2.2e-16
```

Final Rap models: on top, model with speechiness; on bottom model without speechiness.

```
Call:
lm(formula = popularity ~ acousticness + danceability + tempo +
    energy + instrumentalness + liveness + speechiness + valence,
    data = rock_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-43.268  -6.588  -0.354   6.644  41.312

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       34.278384   1.666946  20.564  < 2e-16 ***
acousticness      -6.283640   0.908409  -6.917 5.31e-12 ***
danceability      30.106038   1.594379  18.883  < 2e-16 ***
tempo              0.021076   0.008121   2.595  0.00948 **
energy            13.787400   1.264815  10.901  < 2e-16 ***
instrumentalness  -7.227379   0.851047  -8.492  < 2e-16 ***
liveness          -5.922627   1.405094  -4.215 2.55e-05 ***
speechiness      -19.087498   3.160456  -6.039 1.68e-09 ***
valence          -25.108277   0.998856 -25.137  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.924 on 4161 degrees of freedom
Multiple R-squared:  0.2187,    Adjusted R-squared:  0.2172
F-statistic: 145.6 on 8 and 4161 DF,  p-value: < 2.2e-16
```

Final Rock model without duration.

```
Call:
lm(formula = popularity ~ acousticness + danceability + duration_ms +
    energy + liveness + valence, data = only_jazz_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-33.724  -8.369  -1.003   7.645  41.023

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.339e+01  3.604e+00   9.263  < 2e-16 ***
acousticness -1.751e+01  2.183e+00  -8.020 2.94e-15 ***
danceability  2.516e+01  4.267e+00   5.896 5.10e-09 ***
duration_ms   1.848e-05  4.258e-06   4.340 1.57e-05 ***
energy        1.627e+01  3.474e+00   4.684 3.20e-06 ***
liveness     -2.213e+01  3.474e+00  -6.371 2.86e-10 ***
valence      -2.834e+01  2.811e+00 -10.079  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.13 on 998 degrees of freedom
Multiple R-squared:  0.3473,    Adjusted R-squared:  0.3434
F-statistic:  88.5 on 6 and 998 DF,  p-value: < 2.2e-16
```

Final Jazz model without instrumentalness, speechiness, and tempo

```
Call:
lm(formula = popularity ~ acousticness + danceability + duration_ms +
    energy + instrumentalness + liveness + speechiness + tempo +
    valence, data = only_pop_songs)

Residuals:
    Min      1Q  Median      3Q     Max
-59.504  -8.215   0.534   9.189  42.230

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.808e+01  2.785e+00  13.674  < 2e-16 ***
acousticness     -1.228e+01  1.407e+00  -8.729  < 2e-16 ***
danceability      4.288e+01  2.249e+00  19.072  < 2e-16 ***
duration_ms      -1.438e-05  4.215e-06  -3.412 0.000654 ***
energy            1.460e+01  2.064e+00   7.070 1.91e-12 ***
instrumentalness -1.240e+01  1.507e+00  -8.228 2.80e-16 ***
liveness         -1.141e+01  2.213e+00  -5.157 2.67e-07 ***
speechiness      -1.462e+01  3.888e+00  -3.759 0.000174 ***
tempo             5.834e-02  1.195e-02   4.883 1.10e-06 ***
valence          -3.547e+01  1.403e+00 -25.288  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.58 on 3021 degrees of freedom
Multiple R-squared:  0.3565,    Adjusted R-squared:  0.3546
F-statistic:   186 on 9 and 3021 DF,  p-value: < 2.2e-16
```

Final Pop model with all variables included.

Figure 11: Popularity by Genre

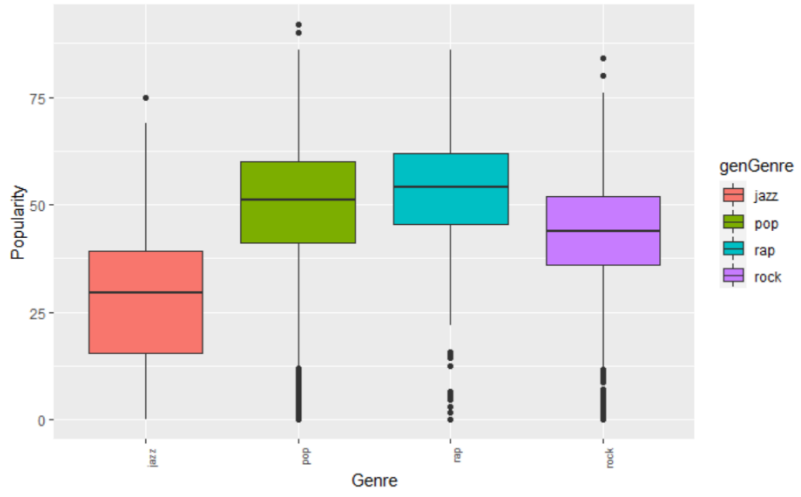


Figure 12: Popularity of Selected Rock Subgenres

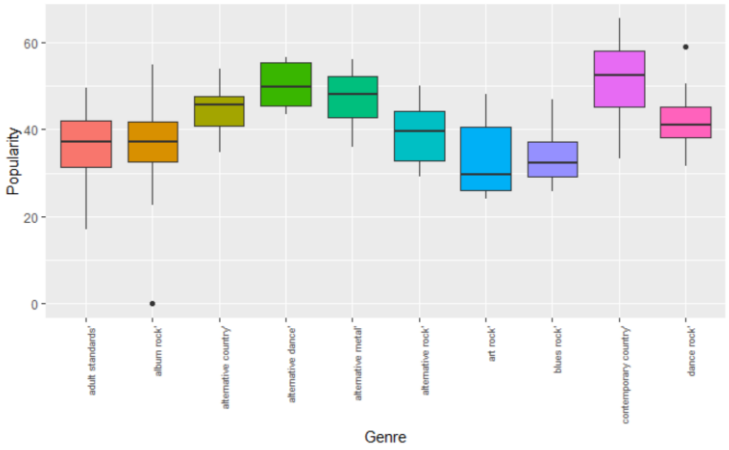


Figure 13: Popularity vs Loudness by Genre

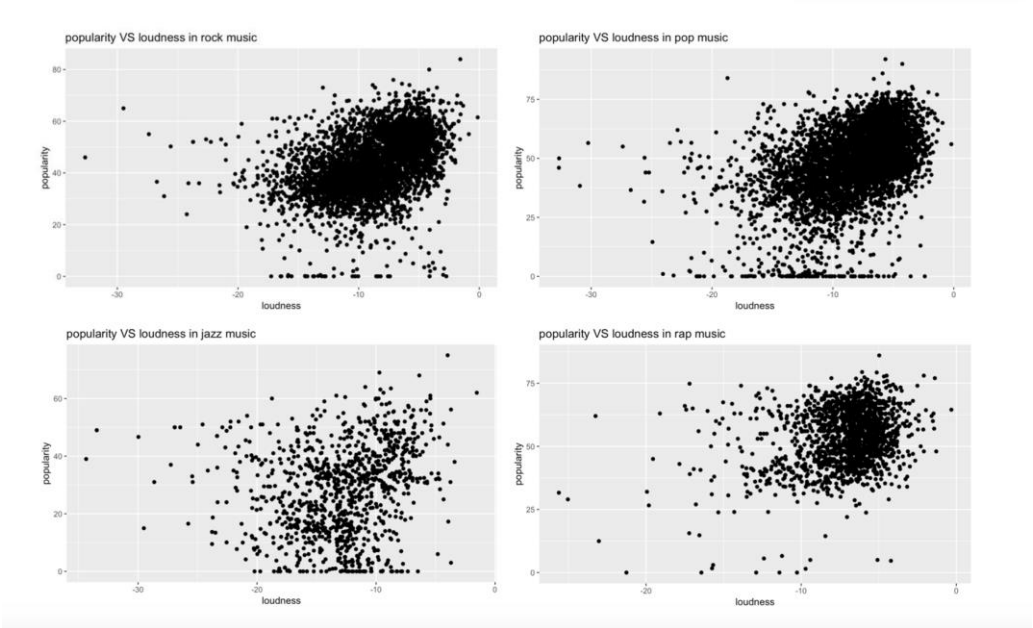
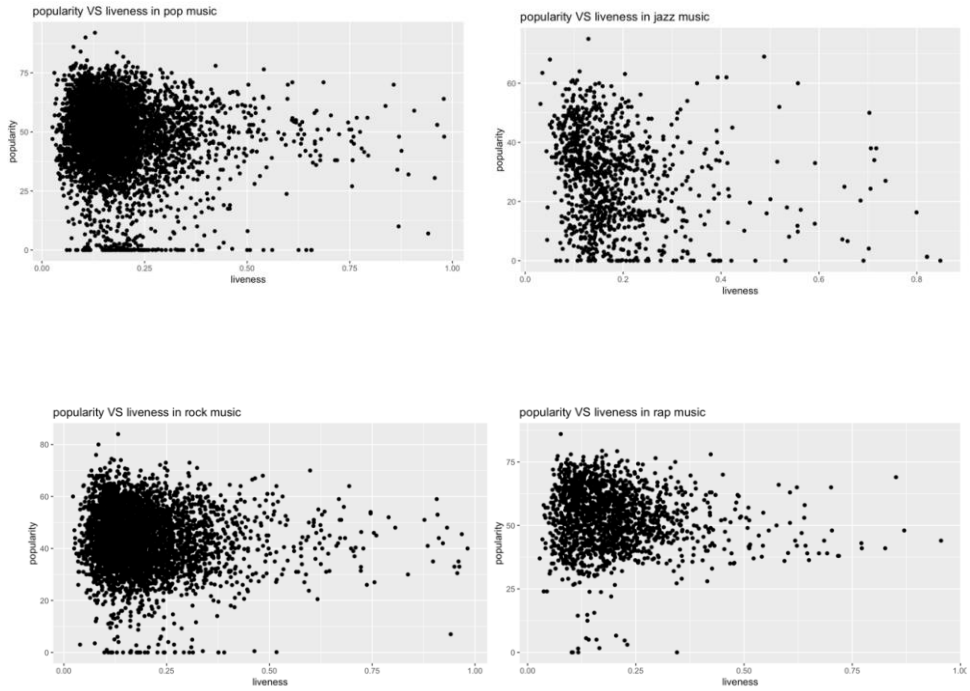


Figure 14: Popularity vs Liveness by Genre



Contributions

Yujia: visualizations, data preparation and evaluation write-up, edit of final paper

Marlon: modeling, modeling write-up, visualizations, edit of final paper

Richard: segmentation/cleaning of data into genres, visualizations, deployment write-up, edit of final paper

Chelsea: visualizations, business understanding and data understanding write-up, edit of final paper

Works Cited

"Get Audio Features for a Track." Spotify for Developers, Spotify, 2020, developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/.

Greenburg, Zack O'Malley. "From Taylor Swift To Dr. Dre: The 10 Top-Earning Musicians Of The Decade." Forbes, Forbes Magazine, 24 Dec. 2019, www.forbes.com/sites/zackomalleygreenburg/2019/12/23/from-beyonc-to-paul-mccartney-the-10-top-earning-musicians-of-the-decade/.

Silva, Matthew De. "Spotify Is Still the King of Music Streaming-for Now." Quartz, Quartz, 28 Oct. 2019, qz.com/1736762/spotify-grows-monthly-active-users-and-turns-profit-shares-jump-15-percent/.