What makes a song popular?

Visualizing the Spotify Dataset

S670: Exploratory Data Analysis
Final Project Report
By: Shreya Paul, Siddartha Rao & Vishal Singh

I. Executive Summary:

Through this project, we have tried to visualize and study the Spotify dataset. Our research is focused on predicting the attributes of a song which predicts its popularity as well as finding if duration of songs impacts their popularity in any way. The dataset used for this project was provided by Spotify on Kaggle and consisted of 232,000 entries. It had data on title, singer, popularity, loudness, duration, danceability, speechiness, tempo and so on for more than 20 genres of music.

For the purpose of this project, we decided to consider only six genres viz. country, pop, electronic, hip-hop, jazz and rock. On initial investigation using box plots, we found that the dataset consisted of outliers. We removed the outliers and post treatment, our dataset consisted of 4385 entries.

On visualizing our final dataset, we decided to use a linear model to predict popularity. We decided to use popularity as our dependent variable and loudness, duration, speechiness, danceability, tempo, liveness and valence as independent variables. Our final model consisted of an interaction of duration with loudness, liveness, tempo and danceability and had an adjusted R squared of 0.61. Our main findings were-

- Pop music in general is more popular than other genres of music
- Popularity of songs decreases with increase in duration when loudness, tempo and speechiness are high
- Interaction of duration with loudness, tempo, liveness and danceability gives us a better model than no interaction.

Later in the project, we also used our model to predict the popularity of a very popular, medium popular and not very popular songs and found that our model predicted values very close to the actual values of popularity for the chosen songs.

This project is just a starting point of the bigger research that we plan to conduct. In future, we are interested in looking at other variables which predict popularity and interaction of genre with different attributes of the song. By doing this, we want to create a model that can predict popularity better than our current model.

II. Data

The data was provided by Spotify and was made available on Kaggle
The dataset had 232,000 rows and 18 columns, consisting of 27 different Genres.
For this project we took a subset of the dataset and filtered it for six genres – Pop, Electronic, Rock, Hip-Hop, Country, Jazz
After the removal of outliers our final dataset contains 4385 rows.
These are the different variables/features in our dataset

Popularity: This is the dependent variable(Y). It is a measure of how popular the song is, it ranges from 0-100.
Loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Values typically range in between -60 and 0 db
Duration: The duration of the track in milli seconds.
Speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values typically range in between 0 and 1
Danceability: 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.
Tempo: 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 and derives directly from the average beat duration
Liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live
Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive, while tracks with low valence sound more negative

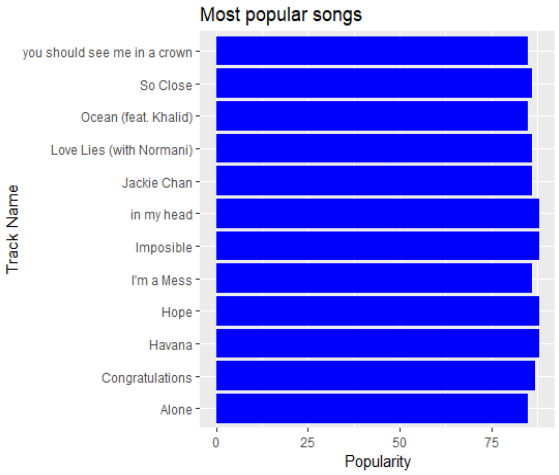
III. Research Questions

1. **What features determine the popularity of a song in a genre?**
   We have different features for all the songs, and it might be possible that not all features contribute to determine popularity of a song in different genres, in our first research question, we look at how different features effect songs of different genre to determine popularity
2. **How duration of songs affects their popularity based on other attributes?**
   We look at the feature 'duration' to see how it effects in determining the popularity of a song, we also look at how 'duration' interacts with different categorical features.

Ultimately, we want to build a model to predict popularity of songs so that anyone interested in making new songs can use our model to check how popular it will be! Using our model, they'll also be able to select the duration of their song based on other attributes to maximize popularity.

IV. Analysis

We began by looking at a few of the basic graphs for our data to get a better sense of our dependent variable 'Popularity' since it is a very subjective term. We look at the most popular and least popular songs from out dataset.



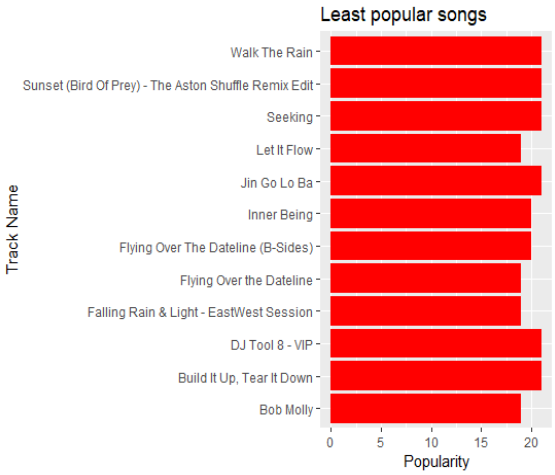
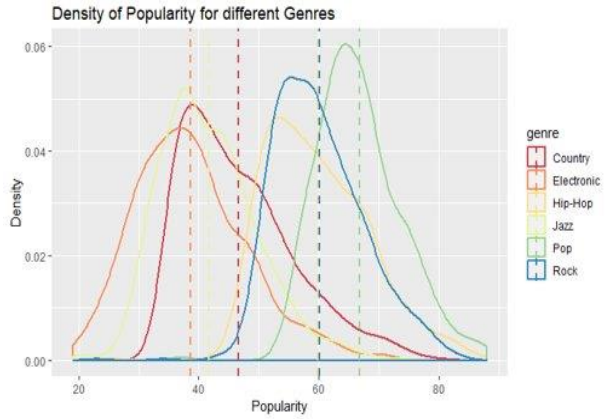


Figure 1 (a) Most popular songs *(b)Least popular songs*

We then visualize the popularity variable of our dataset by making density and box plots for popularity and segregating by genres. The plots below show the spread of popularity of songs based on genre. The plots also show the median value of popularity for songs of each genre.



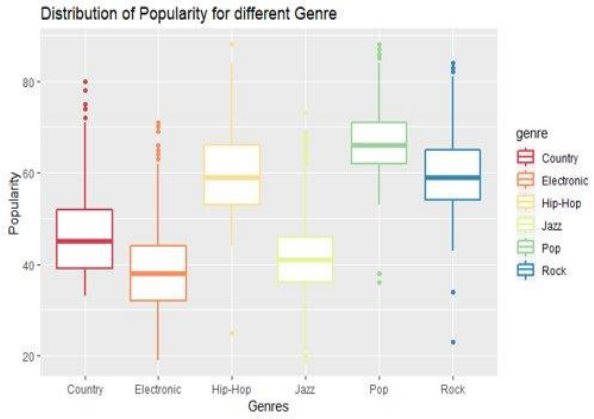


Figure 2 (a) Frequency distribution of popularity by genre *(b)Boxplot of popularity of different songs by genre*

From the density plot and the box plot, we notice in general Pop is most popular genre of music with Pop songs having a median value of popularity of around 65. Furthermore, we notice that the median value of popularity for songs of genres Country, Electronic and Jazz lies between 35 and 45 while for Hip-Hop and Rock, it is around 60. From this information, we deduce that on average Country, Electronic and Jazz genres are not as popular as Hip-Hop and Rock and Pop is the most popular genre of them all. We also notice from the density plot, that the distribution of Electronic, Jazz, Pop and Rock is normal while for Hip-Hop and Country, the distribution is slightly skewed towards the right. This might be due to the existing outliers, which we failed to remove after multiple attempts.

Research Question 1: What features determine the popularity of a song in a genre?
We attempt to look at different features and how they affect the popularity of a song

We first attempt to visualize how value of popularity changed with valence for different genres and the plots are shown below.

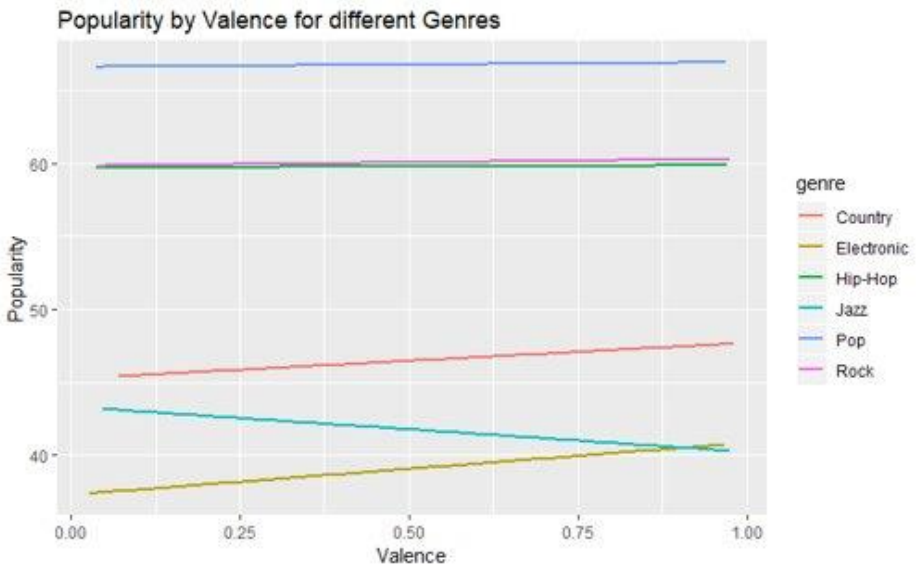


Figure 3 Popularity by Valence for different Genres

We notice from the above plot, that there is no change in popularity with increase in valence for Pop, Rock and Hip-Hop genres. However, with increase in valence, the popularity increases for Electronic and Country genres. Moreover, we notice that for Jazz music, popularity decreases with increase in valence.

We next look at how popularity changed with acousticness for different genres.

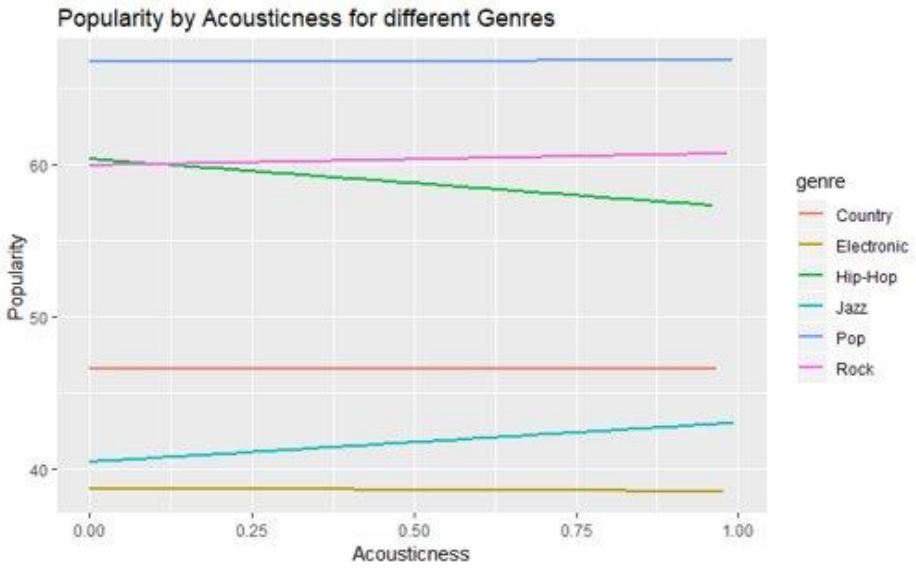


Figure 4 Popularity by Acousticness for different Genres

From the above plot, we see that the value of popularity remains the same with increase in acousticness for Pop, Electronic and Country genres. However, popularity increases with increase in acousticness for Rock and Jazz genres while it decreases for songs of Hip-Hop genre.

Lastly, we look at changes in popularity for different genres with change in speechiness.

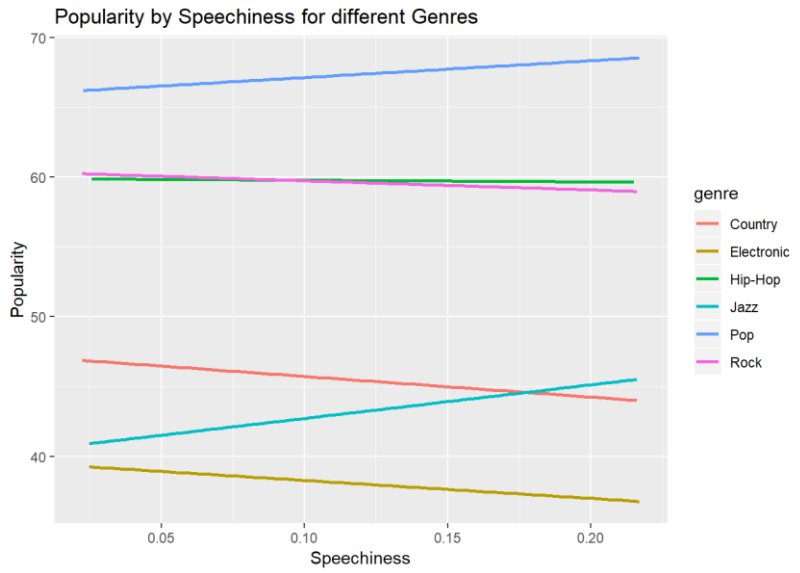


Figure 5 Popularity by Speechiness for different Genres

We see that for genres Jazz and Pop, the popularity increased with an increase in Speechiness whereas it decreased with an increase in speechiness for Country and Electronics.

Research Question 2: How duration of songs affects their popularity based on other attributes?

We then decided to explore how popularity changed with change in duration segregated by genre.

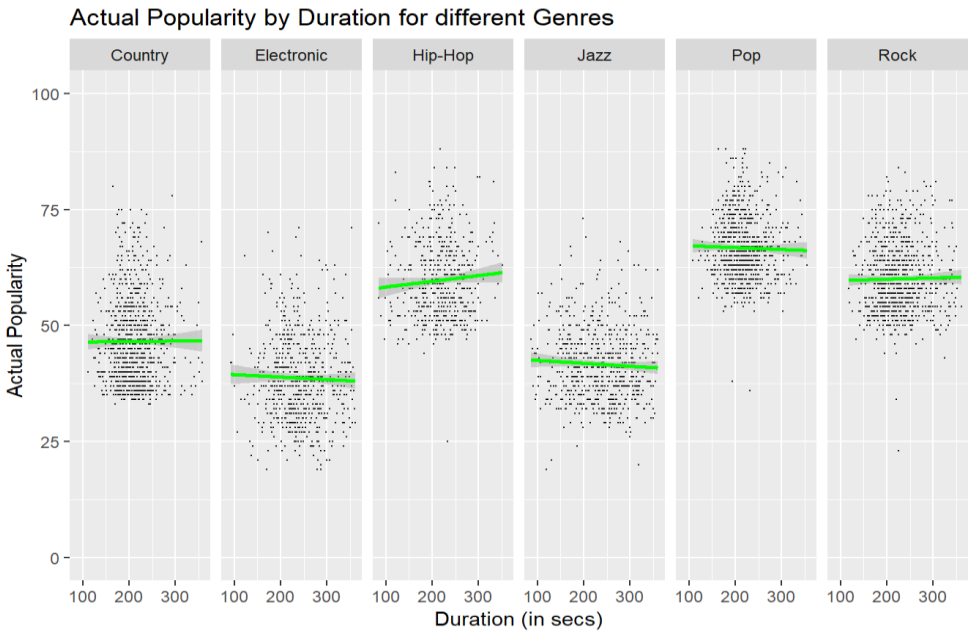


Figure 6 Actual Popularity by Duration for different Genres

On analyzing the above plot, we notice that popularity increases with increase in duration for songs of Country, Hip-Hop and Rock genres. On the other hand, popularity decreases by a very small amount with increase in duration for songs of Electronic, Jazz and Pop genres.

Next, we investigated the interaction of duration with other attributes of the song such as loudness, liveness, danceability, speechiness, tempo valence and acousticness. However, after

using anova for the models with interaction and the models without interaction, we noticed that the models with interaction terms for loudness, tempo, liveness and danceability were the better models. For the rest of the attributes, we determined that the models without interaction terms were the better models.

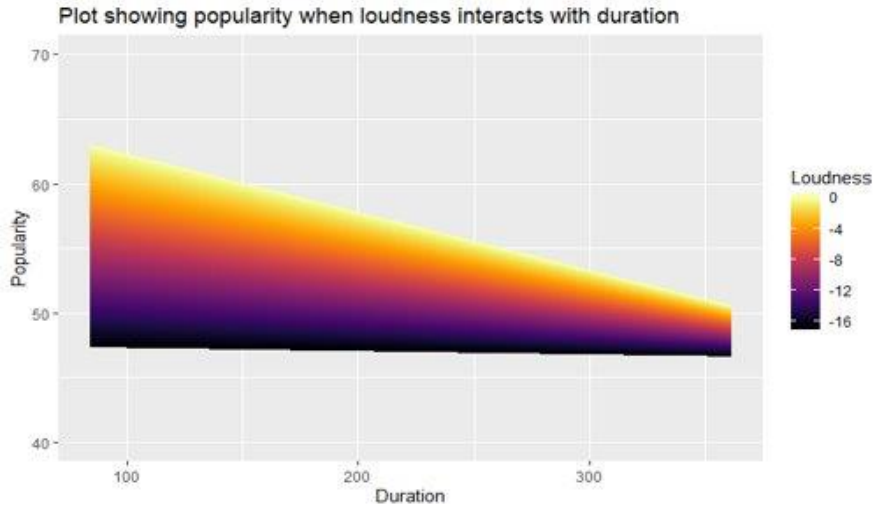


Figure 7 Popularity by duration interacting with loudness

Our comparative analysis between models using anova showed that the model containing interaction between duration and loudness is better than the model containing no interaction term. The above plot shows that popularity of songs with extremely high levels of loudness decreases with increase in duration while popularity of songs with extremely low levels of loudness remains the same with increase in duration.

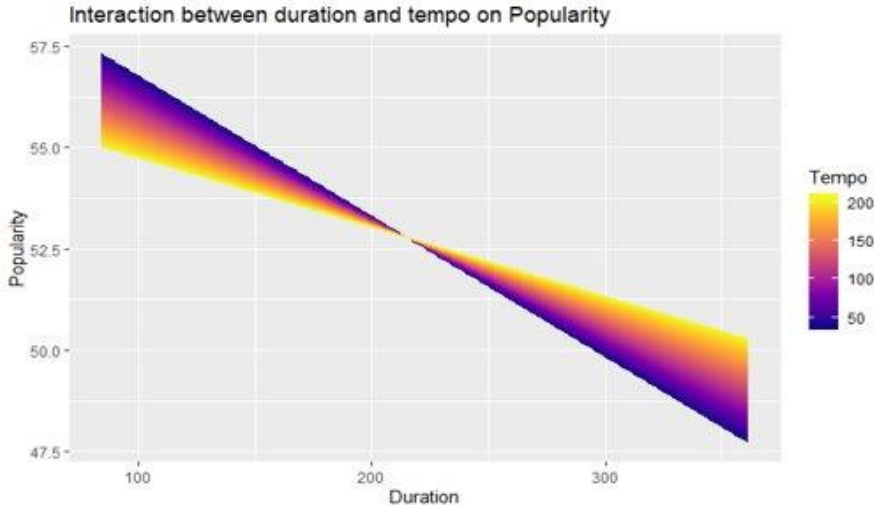


Figure 8 Popularity by duration interacting with tempo

Our analysis also found that the model containing interaction of duration and tempo is better than a model containing no such interaction. And, as we can see from the above plot, as duration increases, the popularity of songs having extremely low levels of tempo decreases more rapidly than popularity of songs with extremely high levels of tempo, both for songs of duration above and below 200 seconds.

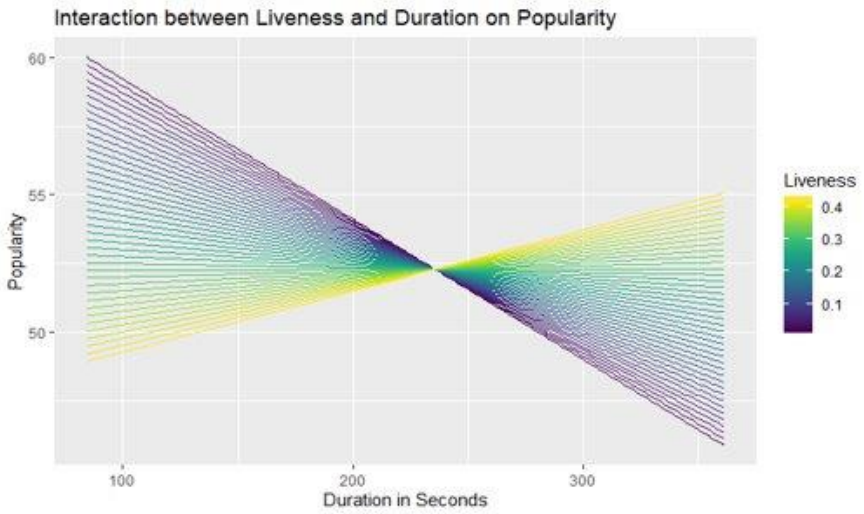


Figure 9 Popularity by duration interacting with liveness

We also found that a model containing the interaction between duration and liveness is better than a model with no interaction. As we can see from the plot above, the popularity of songs with low values of liveness decreases with increase in duration whereas the popularity of songs with high values of liveness increases with increase in duration.

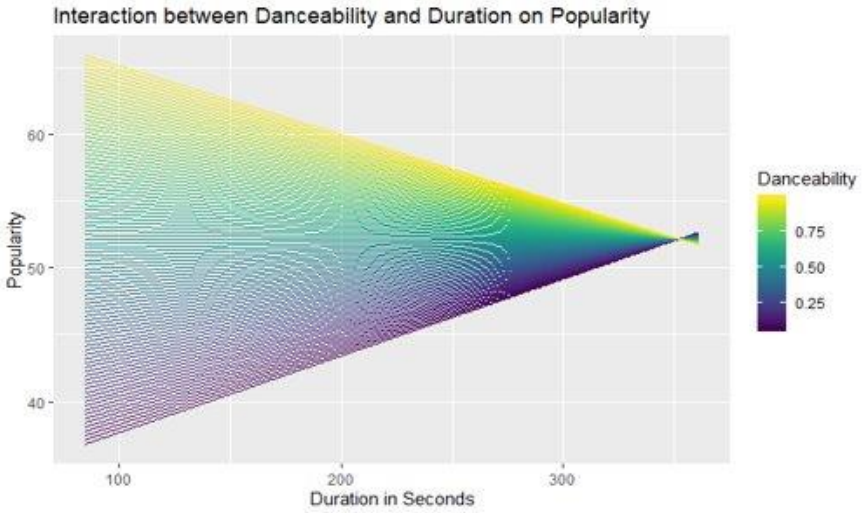


Figure 10 Popularity by duration interacting with danceability

Finally, our study found that a model with an interaction between duration and danceability is better than a model with no such interaction. Also, the above plot shows that with increase in duration, popularity of songs having high levels of danceability decreases whereas for songs having low levels of danceability popularity increases.

Therefore, we can conclude that the interaction of loudness, tempo, liveness and danceability with duration does have an impact on predicting popularity and adds value to our model.

V. Linear Model

Using the features given for the songs, we built a prediction model which would predict the popularity of a song. Since the popularity of a song ranges from 0-100, we fit a Linear Regression model to our data using the features given. We convert speechiness, loudness, tempo, and liveness to categorical variables of different number of classes, we also include the interaction between duration and these categorical variables as these are found to have interaction from our research question 2

Our equation for the fit model can be expressed as

$$Prediction = lm(duration * (loudness + liveness + Tempo + Dancibility) + Valence + Genre + Acousticness + Dancibilty + mode + speechiness)$$

We obtain an R-squared value of 0.61, which proves that our model is decent at predicting the popularity of a song for the given set of features. When we look at the residual plot, we see that the residuals are randomly distributed across the mean line and do not show any trend. This is also proves that our model is a good predictor of the dependent variable.

Figure 11 Residual plot of the linear model

We next look at the plot of predicted values for our model, we see that even if the plot looses some of the gradient in the fitted lines for different genres, we can still see the predicted values are quite close to the actual values shown in *figure 6.* We can also see that the order of the mean predicted values for different genres is same as the actual values.

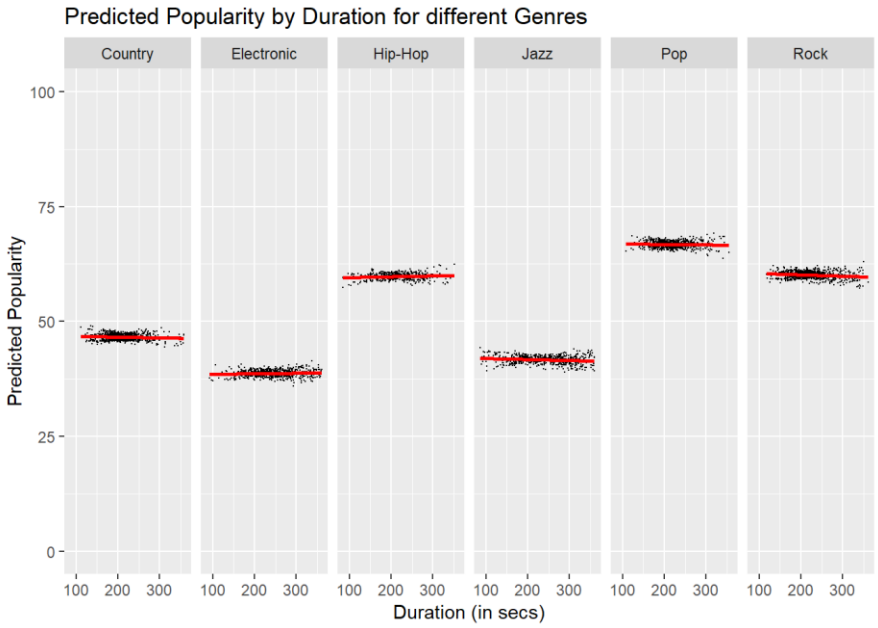


Figure 12 Predicted Popularity by Duration for different Genres

For checking the performance of our prediction model, we finally compare the actual and predicted values for different categorical variables. We plot a tile plot for actual and predicted popularity values for different values of Tempo

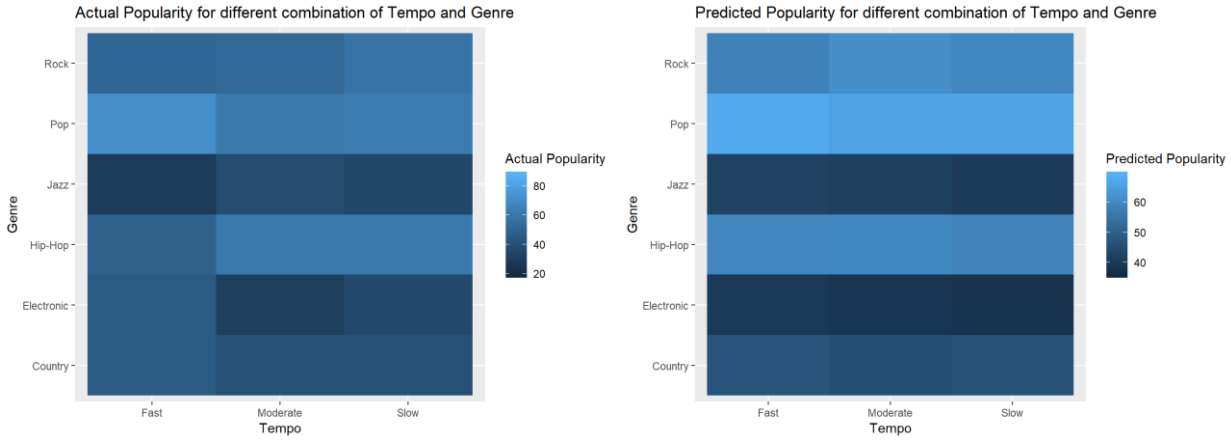


Figure 2 (a) Actual popularity values *(b) Predicted popularity values*

Below is a list of a few songs along with their actual and predicted values

Song Name	Genre	Actual Popularity	Predicted Popularity
Havana	Pop	88	66.60
D'yer Mak'er	Rock	57	60.95
We got Us	Country	43	47.16
Lionheart	Jazz	46	42.48
Rule The World (feat. Ariana Grande)	Hip-Hop	80	59.07

Table 1: Actual popularity and predicted popularity of Songs

VI. Conclusion

After performing analysis, we can conclude the following for the research questions we had intended to answer.

- **Different features have different effect on the popularity of songs**, we saw this by looking at features i.e. valence, acousticness, speechiness and how they had different effects on the popularity of songs of different genre. We could say that certain features characterize a genre and increasing/decreasing that feature for song in that genre will affect popularity of that song. For example, in case of Jazz we saw that popularity of Jazz songs increased with increase in acousticness, this could be because jazz instruments are mostly acoustic and having more of that feature make a song more jazz-y and hence, increase its popularity. We can also see that popularity increase for Jazz songs with decrease in Valence which is the measure of how positive a song is. We could comment that Jazz appeals to a more mature audience who seek a deeper meaning in the lyrics which often are not positive.
- **We could see significant interaction between the features which we inspected** namely duration with Loudness, Tempo, Danceability, and Liveness. The interaction between variables shouldn't be ignored while creating a predictive model using song features. We could look at more interaction in the future to build a better model for this problem.
- Using the given variable and including a few of the interactions **gave decent results on the linear predictive model**, this could be further improved by including more data, including more interactions, building a non-linear model or decision tree based predictive model.

VII. Future Work

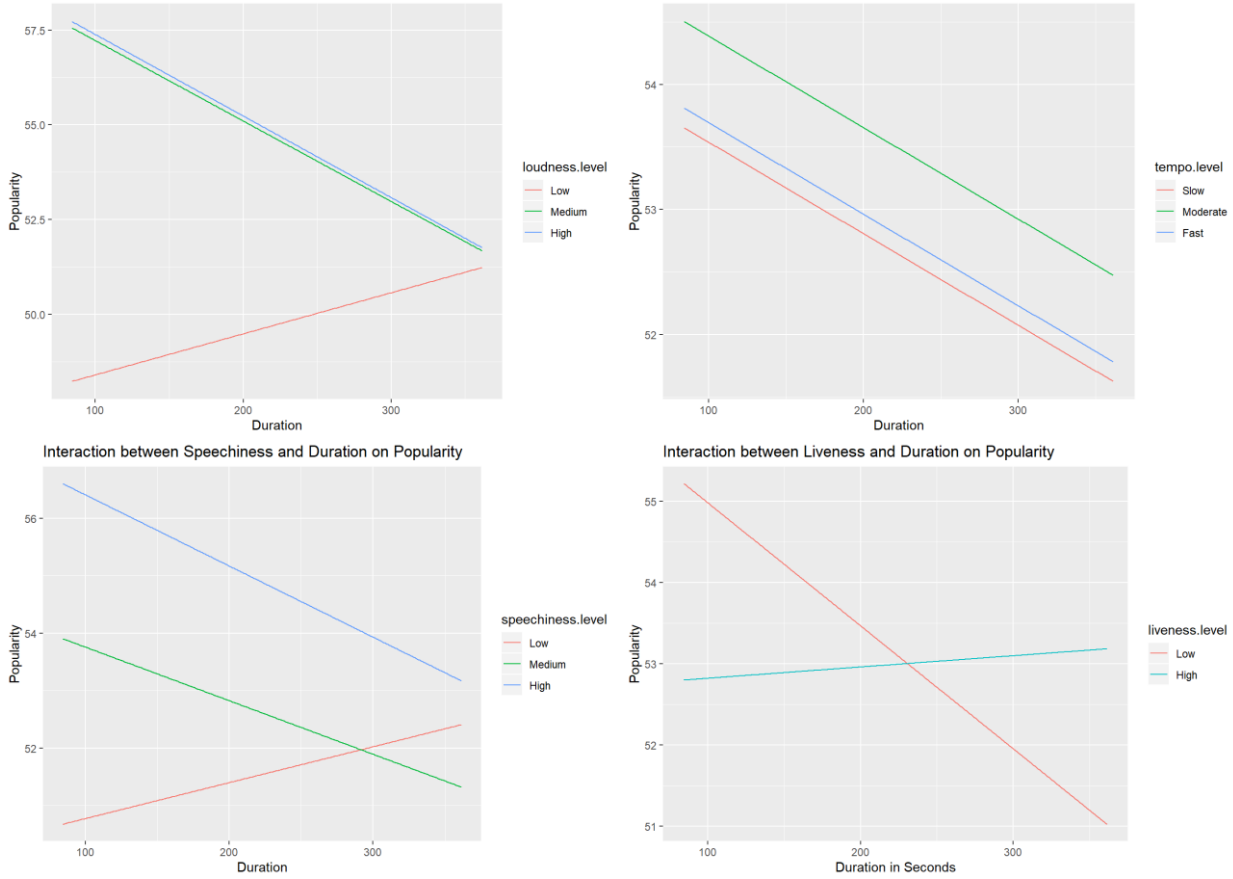
In this project, we only looked at how duration affects the popularity of a song by studying its interaction with other attributes. However, this project did not explore in detail how other attributes of a song determine its popularity. Future work includes exploring the impact of these attributes and their interaction on the popularity of songs. Also, adding as many meaningful predictors to the model as possible. Moreover, future research should also explore if artists, gender of artists and place of origin of songs influence popularity. For example – Are Post Malone's songs in general more popular than Ariana Grande's? Are Spanish songs more popular than Hindi songs? Answering these questions will help build a robust model for predicting popularity.

VIII. References

1. https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/
2. https://rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf
3. https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
4. Data: https://www.kaggle.com/zaheenhamidani/ultimate-spotify-tracks-db

Appendix

We see interaction between duration and variable Loudness, tempo, speechiness, and liveness after converting the latter set of variables to categorical variables as described in notebook – Data Creation. We decided to show the interaction by keeping these variables continuous (Figure 6-10) as we thought they gave a better intuition of the results.



We performed Principal component analysis and plotted the different data points for top 2 principal components (61% of variation explained) by different genre, we couldn't observe any discernible difference in the points for different genres.