Exploring and Classifying genres in Music

Aditya Kanthale, Amanuel Odicho, Vivek Patel

1 Introduction

Music has played a prominent role in shaping our culture and society, and thus has risen to become among the most popular ways to appreciate unique and fascinating pieces of art. The art form has rapidly grown from being sold through cassette tapes and CD's to being abundantly available through several music streaming services. In 2021 alone, the market of global music streaming was valued at 29.45 billion US dollars, with an increasing growth rate every year following the rise of smart devices and expansion of the availability of the internet. (Grand View Research EDIT THIS). It goes without saying that this is quite a lucrative sector of the economy. Companies such as Spotify and Apple Music, the two biggest streaming platforms, are in tight competition for the market share. Therefore, they have to focus on gaining and also retaining their customers, lovers and listeners of music.

The primary focus of our research project is to explore new and innovative ways to expand the general palette of music that is currently trending. This will benefit both artists that are currently struggling to grow their audience, and also the customers to come across new music in possibly different genres. We will do this by exploring similarities between two songs in different genres. This can be further developed into taking into account a history of songs listened to instead of just one song, and using that information to predict what songs the customer might listen to from a different genre. Along with this, we aim to build a model that can predict if a song will become popular based on the technical metrics of the song. Lastly, we will do some exploratory data analysis which can also be displayed to the artist in a dashboard analytics fashion, similar to YouTube analytics after you upload a video. This can give insights to the artist on areas to improve to get a greater reach. Our hope is that this will diversify the music industry more than pop and hip-hop, and also let artists be able to understand how to improve through real time analytics.

2 Data Extraction

Existing analyses on music data almost exclusively uses Spotify data due to their rich API collections. For our project, we decided to collect the data from multiple sources. It was one of our assumption that this would complement the Spotify data to give us better results. We validate this assumption later in the analysis.

2.1 RateYourMusic Data

RateYourMusic is a social platform where its users review their music. It's a large community with more than 0.5 million users, and an average of 332 reviews per album. It also curates the lists of top albums' every year/per genre based on these reviews.

With our project goal to increase the genre pallette of a listener, RYM community helps us cover a breadth of genres compared to Spotify, which mainly focus on the listener's history and is more skewed towards pop genres. Also, the data for ratings, reviews and number of reviews will provide us as a metric for our song recommendation.

We customized our scraper over an open-source project rymscraper. We collected the top 20 albums of every year starting from 1960. Here are the some *labels* from RYM:

```
['rym rating', 'ratings', 'reviews']
```

2.2 Spotify Data

Spotify is the largest music streaming platform with around 500 million users and largest song library. The Spotify Web API provide data about music artists, albums, and tracks, directly from the Spotify Data Catalogue. For retrieving the data we used spotipy, a Python library wrapper over Spotify's APIs. Here are some features that we extracted:

```
Audio features:
['danceability', 'energy', 'key', 'loudness', 'mode',
'speechiness', 'acousticness', 'instrumentalness', 'liveness',
'valence', 'tempo', 'type', 'uri', 'duration_ms', 'time_signature']

Song metadata:
['id', 'name', 'album', 'date', 'genres', 'track_href', 'analysis_url',
'popularity']
```

2.3 Exploratory Data Analysis

After fully processing the data, we can finally do some exploratory analysis. We were interested to figure out if there was any bias in our data, so we aggregated by genre. Since every song has a genre attribute, we can easily take advantage of python library functions to help us group by genre, and then count the number of songs in each genre. The bar plot that was produced from this helped us visualize the skewed distribution of genres in our dataset from RYM. There are several options on how to fix this, one such solution is to balance the percentage of the dataset each genre takes. This will minimize the data for bias towards any particular genre. 1

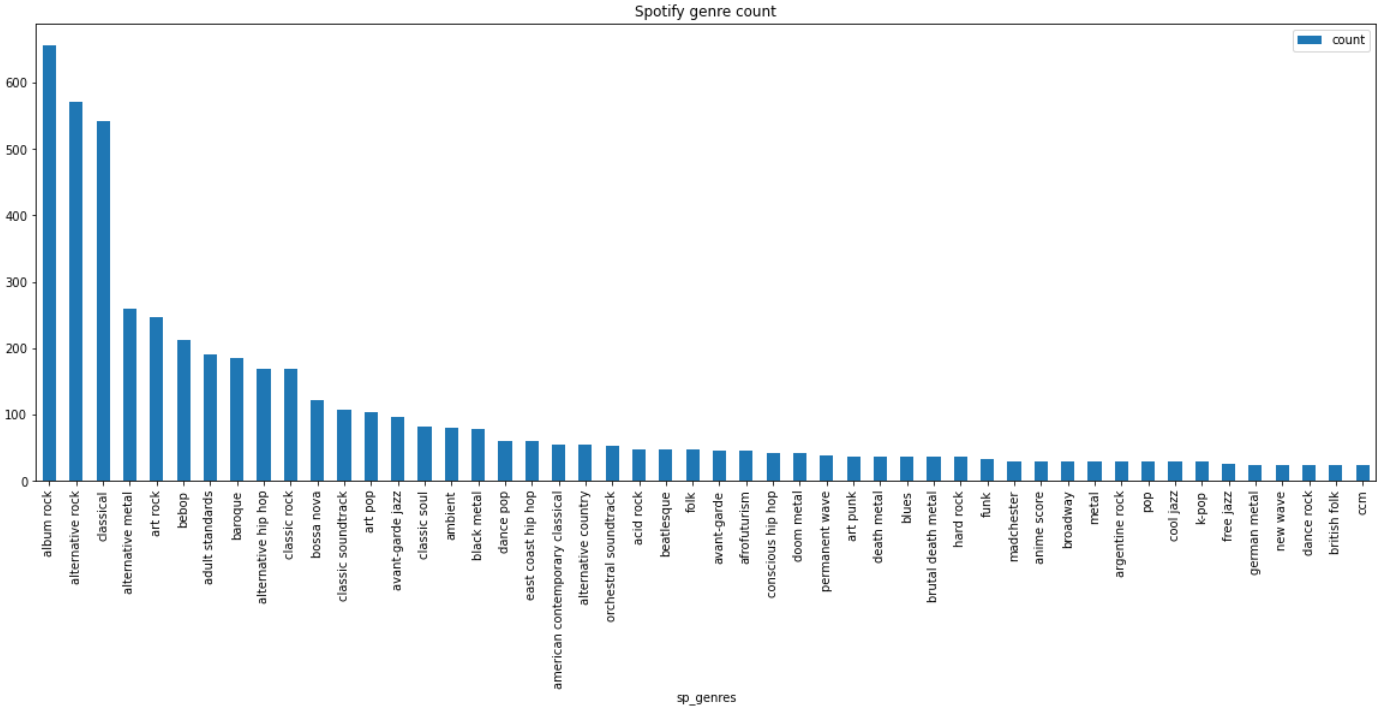

Figure 1: Spotify Genre popularity

To give us an idea on how to approach the popularity prediction model we started looking into the popularity metric provided by the Spotify dataset. This is a metric from 0 - 100 and measures the popularity rating of a song, presumably by number of listeners and some other factors.

Nonetheless, we took a similar approach to the first EDA and we grouped by genre, but we found the mean of all songs' popularity attribute added

in each genre. This visualization allowed us to grasp which genre had the highest "hit" rate, meaning the number of songs produced in a genre vs the number of songs that got very high popularity. If a genre had to produce 1000 songs to get 1 song to be popular, that's a worse ranking than a genre that produces a popular song for every 100 songs produced. This is all relative and this is displayed neatly using the bar plot again. Knowing this information, it can help us fine tune parameters later on with modeling algorithms. 2

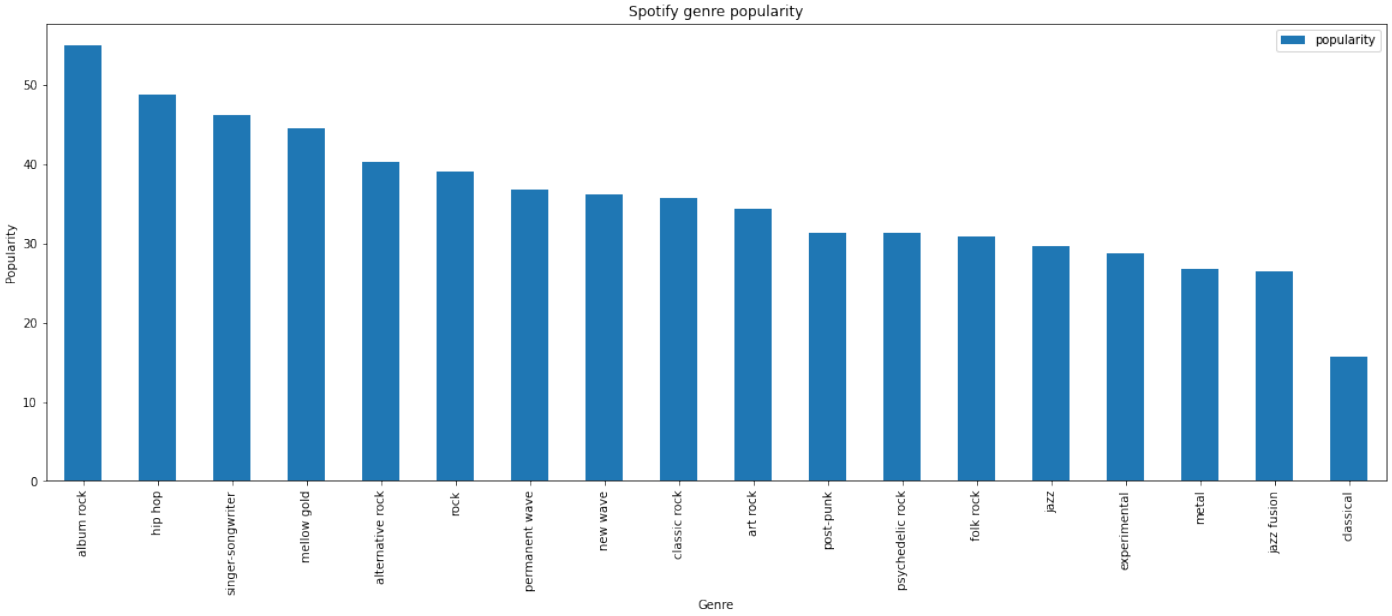

Figure 2: Spotify Genre popularity

While we are at this stage, we also thought of analyzing the actual RYM ratings for each song and finding the average rating within each genre. We can cross reference the top genres here with the top genres from the previous EDA and find some sort of relationship between what people think of a song vs. how popular it is. Surprisingly, we were not able to find any positive correlation or any indication where one causes the other, in fact it seemed random. Genres that had high ratings from users were not always popular genres, and popular genres were not always highly ranked, and vice versa. Here is a quick look into the bar plot produced for Spotify genre vs RYM rating. 3

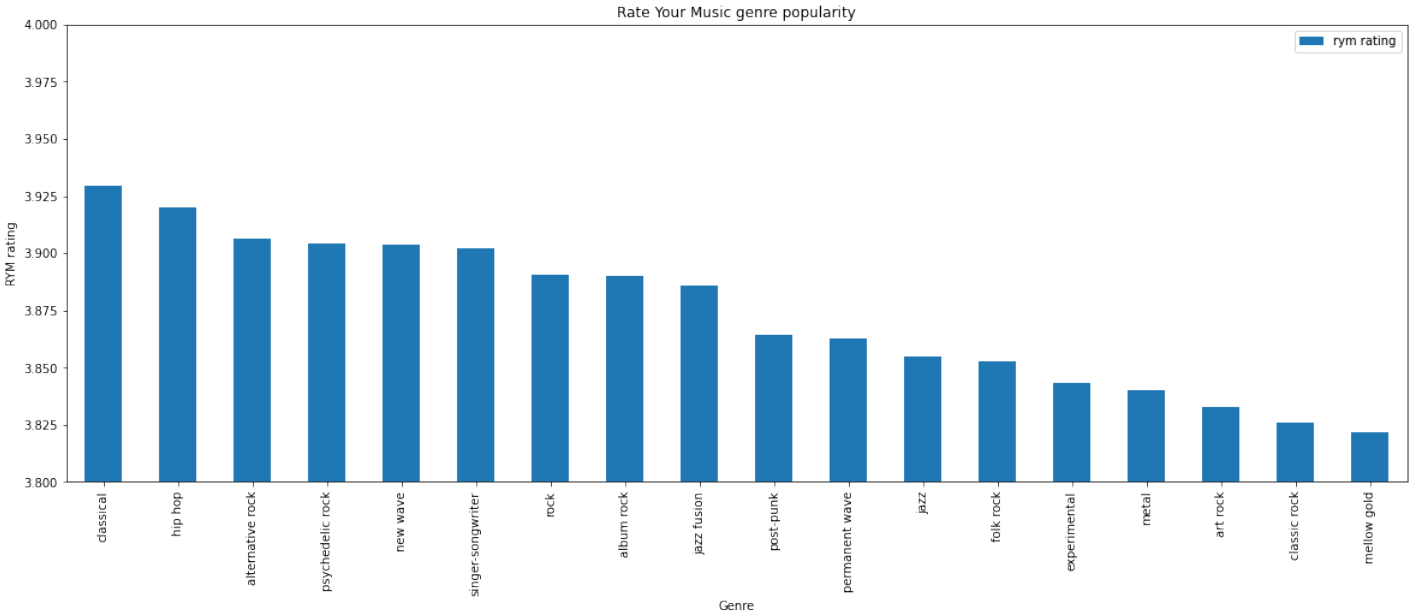

Figure 3: RYM Genre popularity

Lastly, we were able to take advantage of the extremely helpful Python library Seaborn to create a beautiful heatmap of a confusion matrix of all our data and its features. This shows a detailed summary of our data through the permutation of correlation between each and every feature. Instinctively, we thought this would be helpful to select features for more advanced analysis and modeling later on. Unfortunately, as you can see, none of the features that we wanted to investigate such as popularity did not have any significantly negative or positive correlation with any other features. Furthermore, we combined the top features that had the highest correlation with popularity and tried to use that to fine tune our models. We will delve further into it in the next section. 4

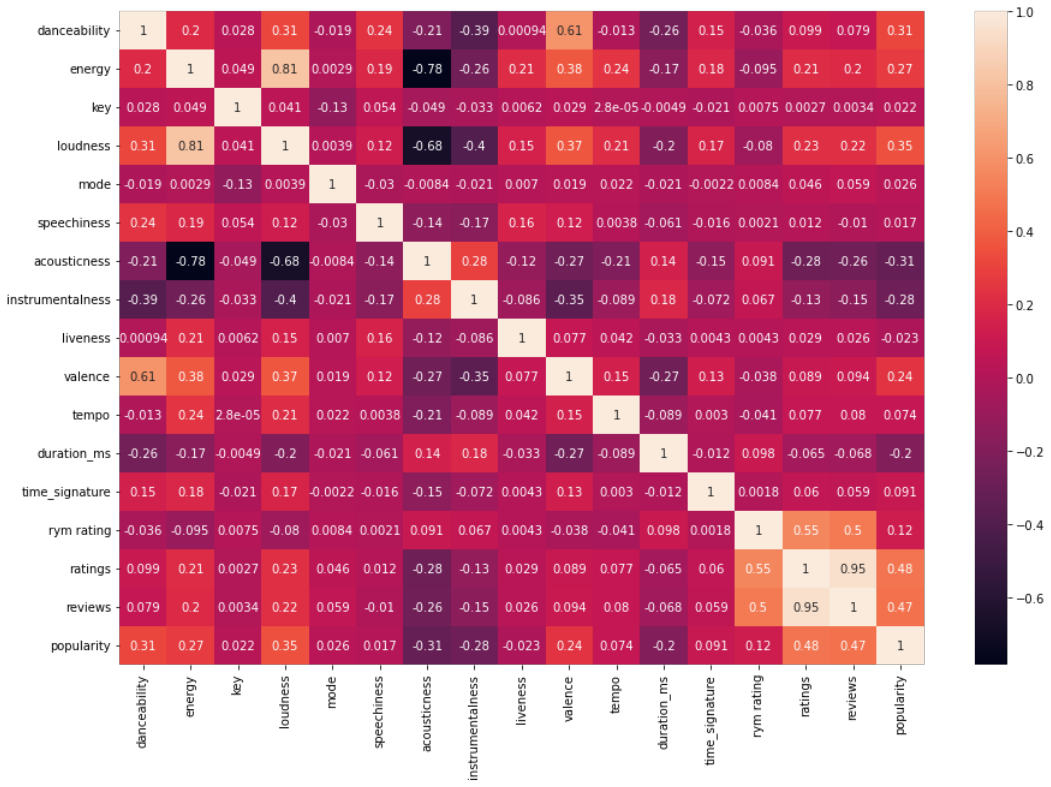

Figure 4: Heatmap correlation matrix

3 Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec ultricies dui a elit egestas rutrum. Duis ac hendrerit felis. Quisque vehicula sodales mattis. Donec eget viverra ex. Suspendisse dignissim nisl nec dignissim ultrices. Sed et bibendum magna. Curabitur egestas venenatis gravida. Suspendisse potenti. Vestibulum ut dictum lectus. Aliquam volutpat dolor lacus, sed facilisis nisl convallis sit amet. Maecenas egestas consectetur felis sed volutpat. Quisque sit amet tincidunt odio. Nullam quam orci, faucibus vitae nunc vitae, sollicitudin commodo orci. Integer molestie luctus leo, vestibulum pretium enim molestie ac. Etiam laoreet vestibulum lobortis.

4 References

: Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec ultricies dui a elit egestas rutrum. Duis ac hendrerit felis. [1] [2]