Joshua Lin, Jolene Okamoto, Son Nguyen, Ji Yoon Rhee

22 March 2022

Project: Identifying trends/patterns in genres and songs on Spotify

Phase 2: Data Acquisition and Preparation

Before we specify the data source(s) we plan to use for this project, we as a group have decided to modify our objectives. Before, we wanted to identify any trends/patterns in Global Top 200 songs and make predictions using the model based on Top 200 songs. Here, we modified our objectives as follows:

1. Identify any trends or patterns in genres and songs on Spotify that make them more well-known and popular.
2. If such a trend/pattern exists, predict which types of songs will be more popular using the identified trends.

Data Source

The data source we have chosen is named “Spotify Multi-Genre Playlists Data”. The data set is split into 7 different genres: alternative, blues, hiphop, indie, metal, pop, and rock. For each genre, there is data on song popularity, danceability, energy, key, loudness, and mode.

https://www.kaggle.com/siropo/spotify-multigenre-playlists-data

Related Work

We found two sources that performed similar work. Both of the links below use similar Spotify datasets to predict song popularity based on different song attributes, such as danceability and energy. While they are very similar to the work we want to do, they are not

exactly the same because the datasets used in the links below provide more song information compared to the dataset we plan to use. For example, the examples have information on songs' time signatures, tempo, instrumentalness, etc., which is information we do not have.

https://towardsdatascience.com/predicting-popularity-on-spotify-when-data-needs-culture-more-than-culture-needs-data-2ed3661f75f1

https://github.com/MattD82/Predicting-Spotify-Song-Popularity/blob/master/README.md

Additional Resources

We plan to use the following libraries:

- *Pandas*
- *Matplotlib.pyplot*
- *Seaborn*
- *Statsmodels.formula.api*
- *Sklearn.linear_model*
- *Numpy*
- *Scipy*

EDA Preparation Steps

Since it is the early stages of EDA, we inspected the data set, identified its dimensions for any null or invalid values, and replaced or removed the identified data accordingly. We also checked the types of the columns and made sure all the columns had the proper data type. Lastly, we checked for and removed duplicate rows. Since we want to determine the most popular genre and then possibly make a predictive model based on the genre determined from the previous

step, we ran the same iterative process of looking for and replacing any invalid/null values for each genre. In total, we had datasets for seven different genres. All seven datasets had no null-values and the data looked good. There was also no significant number of "0" values for the columns of numerical data (e.g., danceability, energy). There were no duplicate rows nor data type conflict.

To first narrow down which genre out of the seven was the most popular, we calculated the mean popularity score for each genre and compared them to find the genre with the highest average popularity score. We used the range and median values from the set of descriptive statistics for each genre to determine whether there were any outliers within the dataset.

We have determined that out of the seven genres, pop music has the highest average score for popularity and therefore is the most popular genre. From this point on, we visualized each criterion (e.g., danceability, energy, valence) based on the popularity score as scatter plots, as this type of graph is mostly used to visually aid in observing correlation between two variables.

For danceability, energy, valence, loudness, liveness, tempo, and acousticness, there was no visible correlation with the popularity score. All the correlation coefficients were below 0.2, which confirms that there was no strong correlation. The only visible pattern we were to somewhat able to observe is that time signatures and duration of songs in milliseconds had pretty constant patterns to it. Many pop songs had the time signature of 4/4 and the duration in milliseconds that ranges from 100000 to mid-300000 ms. Danceability and loudness had the highest correlation coefficients at 0.13 and 0.12, respectively. Even though the correlations were very low, we think that the reason for these attributes having higher correlations is because of the

increasing popularity and use of the TikTok app, songs that are considered “danceable” attract a larger audience and therefore have an increased number of plays.