Billboard Hot Songs Analysis with Spotify

Yuwei (Johnny) Meng

15 Mar 2024

Source Code on GitHub: https://github.com/BullDF/JSC370_project

Introduction

Music, as a human culture, takes a large part in people's life for long. In the current era where technology thrives, countless digital musical softwares emerge, allowing people to listen to music at anytime, anywhere in the world. With easy access to music, factors related to the famousness and popularity of music are of interest to researchers and musicians so that more and more captivating songs can be created. Hence, through data exploration, data visualization, and statistical analysis, this project aims to explore the factors that make certain musical tracks popular.

As an entry point to the project, a dataset on Kaggle (see reference) is downloaded and exploited. This dataset contains the "Hot 100" Billboard songs every week starting from 1958. The dataset contains 330087 rows and 7 columns that include information about the date of the song on Billboard, the name, the artist(s), and some about the rank on Billboard. In this project, only the name and the artist(s) columns are used.

Among the leading musical softwares, Spotify is a world-changing one. Founded in 2006 in Sweden, Spotify gradually attracted more and more users and built up a massive repository of worldwide music. Of this reason, a major portion of the data utilized in this data analysis project is obtained from Spotify through the Spotify web API (see reference) and deemed credible and reliable. This portion of data contains the name of tracks, the artist(s), the popularity, and in particular, some audio features (see appendix for full list) of tracks.

Methodology

The Billboard hot songs dataset from Kaggle is used as a reference. Starting with this list of 330087 tracks, to prevent overloading the Spotify API and keep the data at a manageable size, I randomly sampled 5000 tracks for the remaining analysis. Then I used the Spotify web API via the `Spotipy` Python library (see reference) to access these songs on Spotify and extracted the audio features and popularity for these 5000 tracks. Given that some of these tracks are unavailable on Spotify, I obtained a dataset of 3823 tracks with their corresponding audio features. For analysis, I split the data into three datasets: (1) The tracks and their audio features, (2) The name of the artists, and (3) The matching of track ids and artist ids (for joining the tracks and artists datasets).

Upon obtaining the datasets, I noticed that most audio features are in percentage. For convenience, I rescaled these audio features by multiplying them by 100. I also converted the duration of tracks in milliseconds into the duration type by first dividing by 1000 for conversion into seconds and then calling the `seconds_to_period` function in the `lubridate` package. For text analysis, I employed the `unnest_tokens` function in the `tidytext` package on the name of the tracks for tokenization. A tokenized version with stopwords removed is also considered.

To discover the underlying patterns in the data, multiple types of graphs were created for exploratory purposes. First, I created several bar charts on the tokenized names of the tracks and the mode of the tracks (i.e. major vs. minor), and counted the number of tracks for each artist. As an easy visualization, wordclouds are created for viewing the most frequent words that appear in the names of the tracks. Subsequently, I plotted histograms and scatterplots to visualize the distribution and association between the features extracted using the Spotify API. Lastly, I conducted some regression analysis and attempted to fit statistical models to better understand the underlying trends.

Results

The following charts were created to explore the data. First, I plotted a bar chart to count the number of tracks in major keys vs. in minor keys:

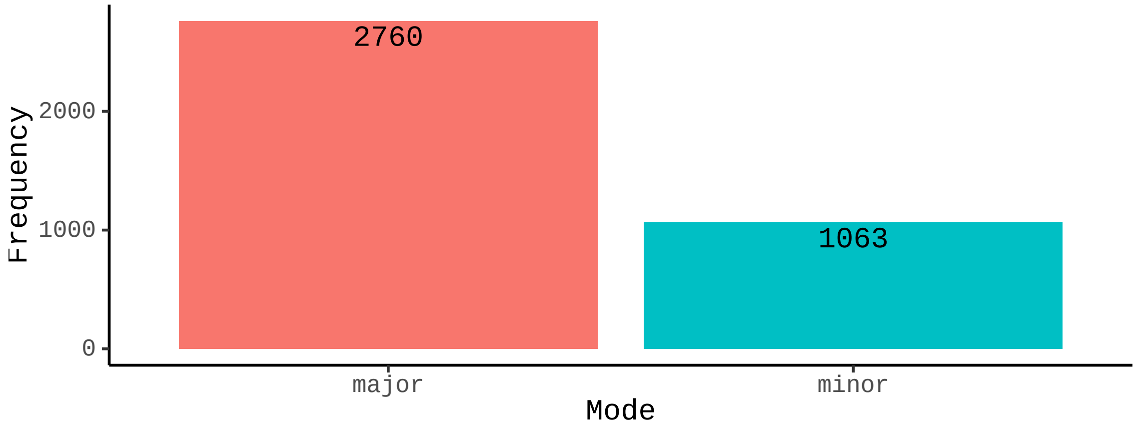

Figure 1: Counts of Major vs. Minor Tracks

In *Figure 1*, we can see that among the 3823 available tracks on Spotify, 2760 of them are in major keys while the remaining 1063 of them are in minor keys. These numbers show that tracks in major keys are favored and more likely to be on the Billboard hot songs. Particularly, this result manifests that people are generally happy as major keys tend to sound joyful and lively while minor keys are more on the gloomy side.

As we tokenized the names of the tracks and removed stopwords, we can also look at the 20 most frequent words that appear in the names of the tracks in the form of a wordcloud:

The wordcloud shows that the word *love* is the most frequent word since it appears in the largest font. This is not surprising because romance has always been an important theme in modern music. The words *remaster*, *version*, and *remastered* appear in similar frequency. These words indicate that sometimes the original version of a track is not as popular. A *remastered version* might be favored by more listeners.

Another question of interest is whether certain artists appear more frequently on the Billboard than others. *Figure 2* below investigates this question.

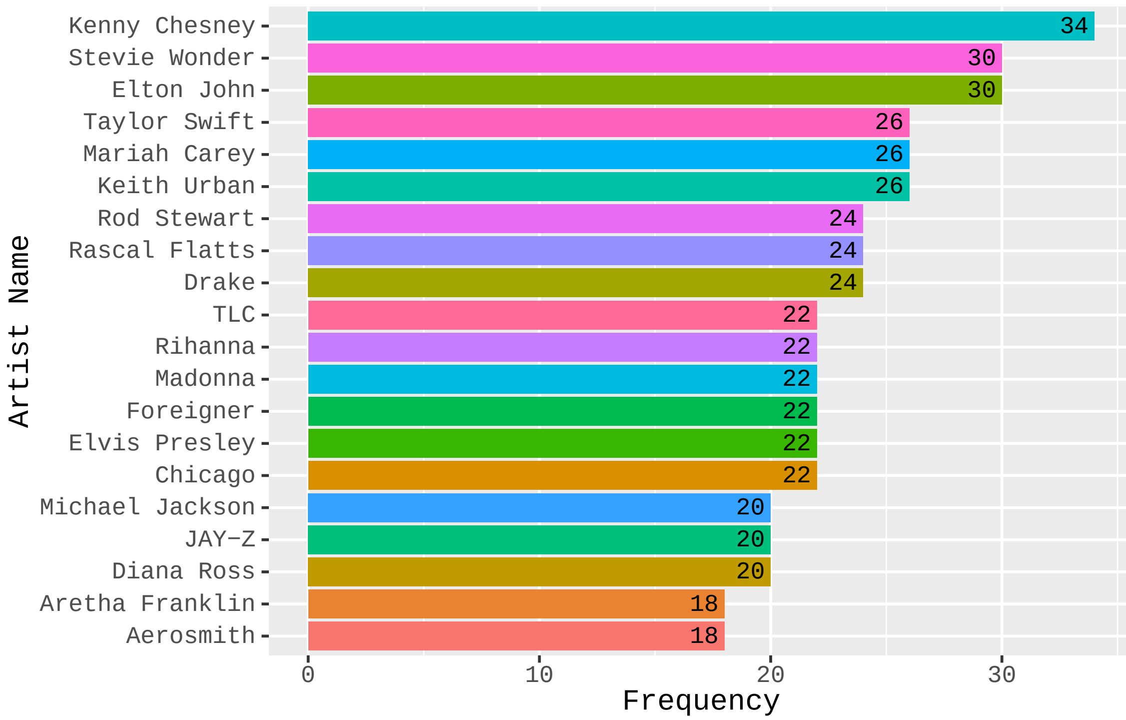

Figure 2: Counts of Artists on Billboard

From *Figure 2*, we can see that *Kenny Chesney* leads the leaderboard, followed by *Stevie Wonder* and *Elton John*. Comparing the numbers, we can conclude that certain artists do appear on Billboard more frequently than others.

Now, to take advantage of the Spotify API, we can look at the association between the measured popularity and various features. Before that, let's examine the distribution of popularity in *Figure 3* below.

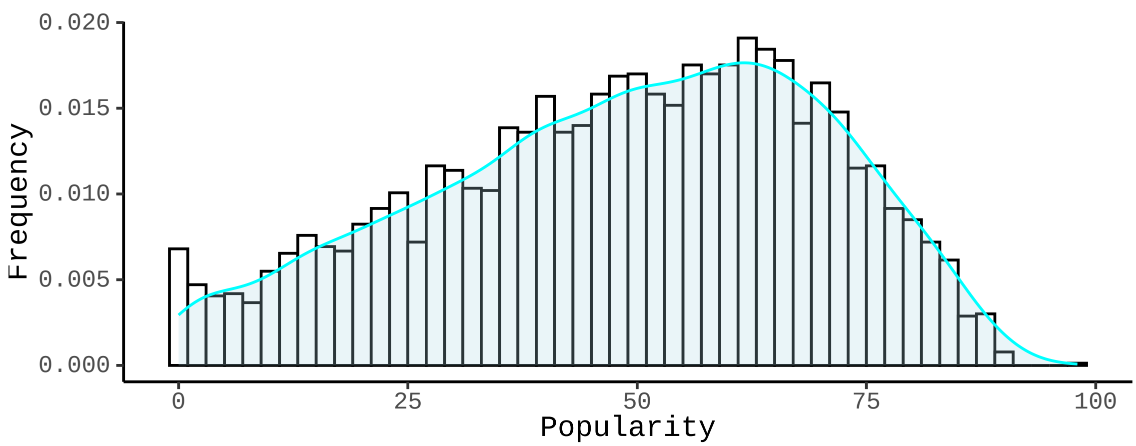

Figure 3: Distribution of Popularity of Billboard Songs with Density

In *Figure 3*, we notice that popularity is unimodal and symmetric with the mean centered at roughly 60 percent. We can also plot the histograms for the year variable and the audio features:

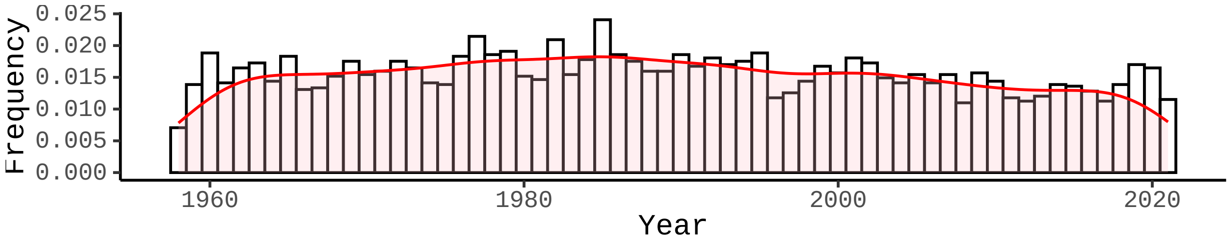

Figure 4: Distribution of Tracks over Time

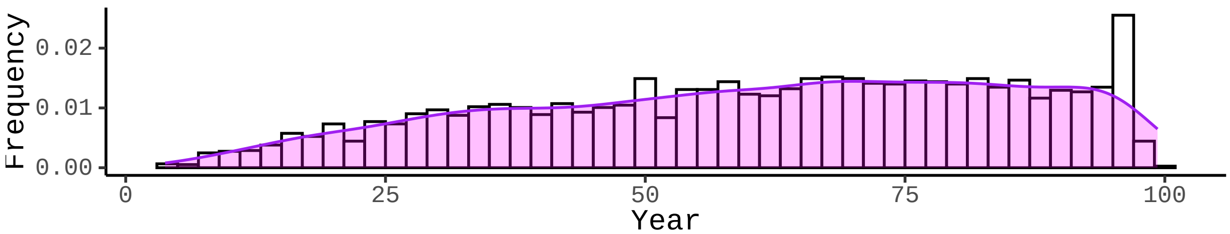

Figure 5: Distribution of Valence with Density

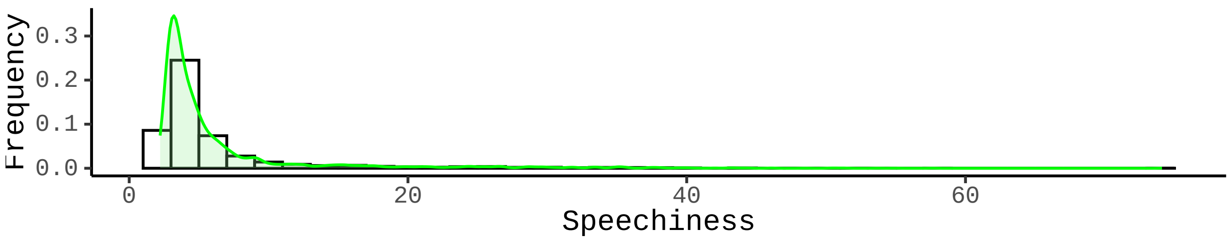

Figure 6: Distribution of Speechiness with Density

From *Figure 4*, the number of tracks that comes from different years is distributed almost uniformly across the timespan of Billboard. Assuming Billboard publishes a similar number of tracks each year, we can summarize that this sample is a representative of all Billboard tracks. For the audio features, I concluded two types of distributions. In the sample, *valence*, *danceability*, *energy*, and *loudness* exhibit a unimodal, symmetric distribution, for which *Figure 5* is a paradigm of this type. The other type is similar to *Figure 6* above, for which the distribution is strongly skewed. This category applies to *speechiness*, *acousticness*, *instrumentalness*, and *liveness*. Since Billboard songs are mostly studio-recordings and sung, it is not surprising that the features *instrumentalness* and *liveness* have a mode around 0 and are skewed to the right. Given the skewed distribution, some transformations might be necessary when fitting the model in subsequent sections.

With a well comprehension of the features, we can start exploring the association between variables. Below is a scatterplot of popularity vs. year.

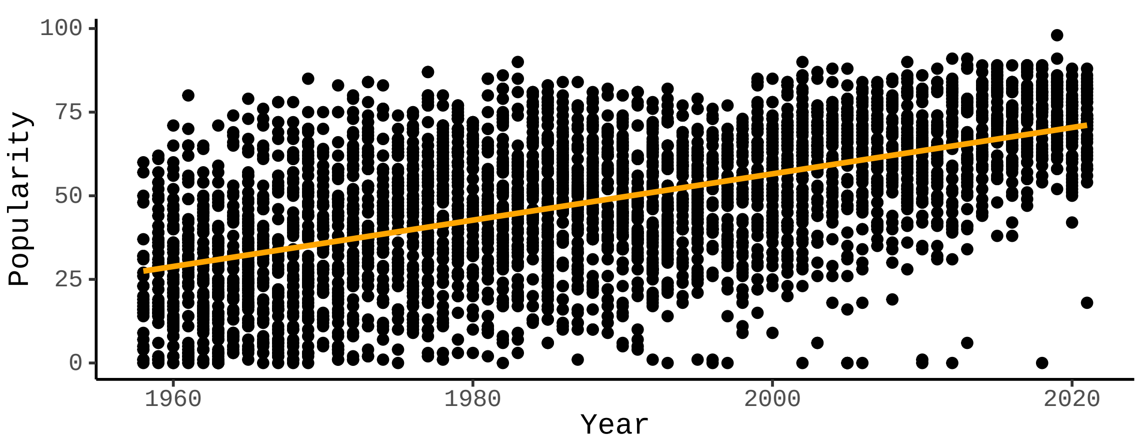

Figure 7: Popularity vs. Time

From the general pattern of the points as well as the fitted regression line, we discover a moderately strong, positive correlation between the popularity and the year of the track on Billboard. In plain words, newer songs are measured by Spotify as more popular than older songs. This is unsurprising if we consider the algorithm Spotify used to compute the popularity. According to the Spotify documentation, the popularity is calculated based on the number of plays and how recent the plays are. This also makes intuitive sense because songs were on Billboard before might not be as popular nowadays. However, this correlation still has practical meaning because the Billboard information and the popularity of tracks come from two different sources (i.e. Billboard vs. Spotify) and so a validation between them is useful. The following summary of a linear model also confirms this finding:

	Estimate	Std. Error	t-value	$\Pr(>\lvert t \rvert)$
(Intercept)	-1328.476	31.634	-42.000	<2e-16
year	0.693	0.016	43.538	<2e-16

R^2	Adjusted R^2	F-statistic	p-value
0.3316	0.3314	1896 on 1 and 3821 DF	<2.2e-16

From the table, the coefficient of the year predictor is 0.693, meaning that an elapse of 1 year is predicted to increase the popularity by 0.693. This confirms the positive correlation. Further, an adjusted R^2 value of 0.3314 implies that 33% of the variation in popularity can be explained by this model.

Besides popularity, we can also explore how some of the audio features changed over time. *Figure 8* and *Figure 9* below show the scatterplots of *energy vs. year* and *danceability vs. year*, respectively.

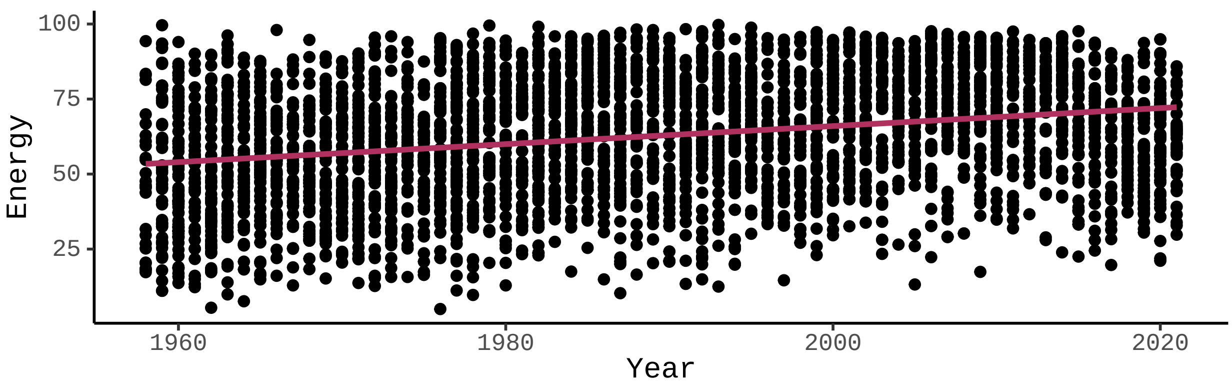

Figure 8: Energy vs. Time

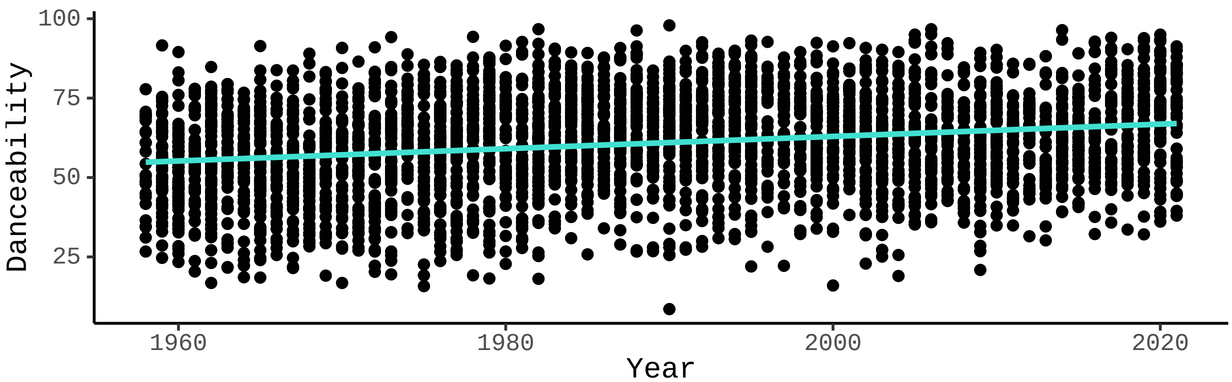

Figure 9: Danceability vs. Time

The plots reveal that the correlations between energy and time, and between danceability and time are both positive. Even though the correlations are weak, there is a general pattern that people were writing more and more energetic and danceable songs.

Summary

In the above analysis, we explored multiple factors and discovered patterns that are shared among Billboard songs. For example, about 3/4 of the tracks in my random sample are in major keys, while the other small portion are in minor keys. For the names, most artists like to name their songs with the word *love*, which reflects that romance is a main theme in modern music. Remastered versions of tracks are also popular among Billboard songs. Among all the artists whose songs were nominated on Billboard, Kenny Chesney has the most songs on the board, followed by Stevie Wonder and Elton John.

Furthermore, we also discovered that the variable of interest in this data exploration, *popularity*, is distributed approximately normally among the range of possible values. This characteristic satisfies the assumption for a regression analysis. The number of songs in each year in this sample exhibits a uniform distribution, which indicates that the sample is a representative of all the Billboard tracks. Among the audio features, *speechiness*, *acousticness*, *instrumentalness*, and *liveness* are skewed in distribution, while *valence*, *danceability*, *energy*, and *loudness* mostly follow a Gaussian distribution. As we formed a linear regression model on popularity vs. year, we noticed that popularity is positively correlated with time, which means that the later the songs on Billboard, the more popular they currently are. Lastly, energy and danceability are both positively correlated with time. This means that the later the tracks were on Billboard, the higher their energy and danceability.

The current exploration has some limitations. For instance, the above analysis only involves a restricted amount of data. If possible, more data should be pulled using the Spotify API for a more comprehensive exploration and analysis. Additionally, only a simple linear regression model was fit in the above analysis, which does not have much expressive power to model a complicated correlation. Hence, more complicated models should be built to better capture the intrinsic patterns within the data.

References

- Billboard "The Hot 100" Songs: https://www.kaggle.com/datasets/dhruvildave/billboard-the-hot-100-songs
- Spotify Web API Documentation: https://developer.spotify.com/documentation/web-api
- Spotipy Python Library Documentation: https://spotipy.readthedocs.io/en/2.22.1/

Appendix

Full List of Spotify Audio Features

Feature	Description
acousticness	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	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.
duration_ms	The duration of the track in milliseconds.

Feature	Description
energy	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. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
instrumentalness	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
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. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.
mode	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
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 above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may 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 other non-speech-like tracks.

Feature	Description
valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a 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).