The Data Behind our Daily Listening

Cade Crandall, Sayuri Garud, Katy Jendrzey, Colin Siles
CSCI 403 Final Project
May 5th, 2021
Github Repo

I. Introduction

About 200 years ago, Henry Wadsworth Longfellow, an American poet, quoted “Music is the universal language of mankind” [1]. Besides the debate of whether music is indeed a universal language, we can all agree that music plays a powerful role in connecting people. As of June 2019, 68% of adults ranging from 18-34 years old reported listening to music everyday [2].

The four of us are a part of this 68%, and we all primarily use Spotify to participate in our daily listening. We only make up a tiny fraction of Spotify users, though — Spotify holds the position of being the most popular audio streaming service with 356 million users (over 44% of those holding a paid monthly subscription) across the world [3]. Although Spotify’s services include other audio entertainment such as podcasts, our primary focus is music, their biggest customer pull.

With such a large amount of users, Spotify is collecting data from an extremely diverse group of listeners. While we couldn’t find individual user data, Spotify (along with many other music databases) keeps records of popular music along with song data such as regional popularity, song attributes, genre, etc. Through Kaggle, we found many large datasets created from Spotify’s usage across different populations. These datasets helped us explore the patterns over time that made our daily listening so unique.

Kaggle offered us a lot of unique data, and we ended up storing more datasets in the database than we actually used [4, 5]. Of these data sets, we used 4: a dataset that tracked Spotify’s top 200 songs over a recent 3 year period, a dataset of Spotify’s popular songs and their song attributes, a dataset of popular songs specific to different regions, and a dataset of Billboard’s top 100 songs tracked over a 20 year period. While we wanted to stick mainly to Spotify, other music popularity charts such as Billboard provided good background information about trends in certain areas such as genre of music.

Some of our data was also pulled from the Spotify Web API [6]. Neither of the Kaggle datasets list a license, but Spotify has a more thorough terms of service that prohibits certain use cases for legal purposes, the primary one being that the API cannot be used for commercial purposes without permission. Our use of the API does not violate the terms of service, as far as we are aware.

We loaded a few different tables from across our datasets. All of the tables were loaded into the schema “colinsiles.” A description of the primary tables we used are included below.

- `world_rankings`**:** Lists the top 200 songs (by track name and artist), and their ranking in fifty-four regions around the world, in each week of 2017
- `billboard_hot_100`: Lists the songs (and their rankings) that appeared on the Billboard Hot 100 (by track name and artist, as well as other miscellaneous information about the song) for each week from July 1999 to July 2019. Also includes information about how long the track had been on the Billboard Hot 100, and the highest ranking it reached.

- `spotify_top_weekly`**:** Lists the top 200 songs (by track name, artist, and featured artists, if applicable), along with the number of streams they received that week, in the US, from December 2016 to July 2019.
- `song_attributes`**:** Lists various quantitative attributes of a collection of songs (identified by the track name, artist, and album). This table includes data from the Kaggle dataset, as well as additional data that was pulled from the Spotify Web API to make it more complete. There are thirteen unique attributes, many of which are scaled between 0 and 1, and are self explanatory. "Loudness" values are reported in dBFS, with all values being negative, but more negative values indicating lower volumes. Popularity is a metric from Spotify from 1-100, with a higher number indicating a song was more popular in July 2019. Finally, valence is scaled from 0 to 1, and indicates the positivity of the track, with higher values indicating a more positive track.

II. Visualization

We used Python, primarily Jupyterhub, to translate our findings and different trends related to our datasets into visuals. We used a Pandas Dataframe to store the dataset and then proceeded to use Matplotlib and Plotly Express to plot our data.

We analysed the data from the `song_attributes` table which consists of different attributes of songs from July 2019 (table is described above). The figure below shows the general trends for songs from 2019. For example, most songs fell under 0-0.25 range of acousticness, instrumentality, and speechiness. The danceability graph is approximately normal depicting that the songs ranged from 0.25-1. The number of songs that were explicit in 2019 were significantly lower than those that were not explicit which might not be the case anymore.

Figure 1

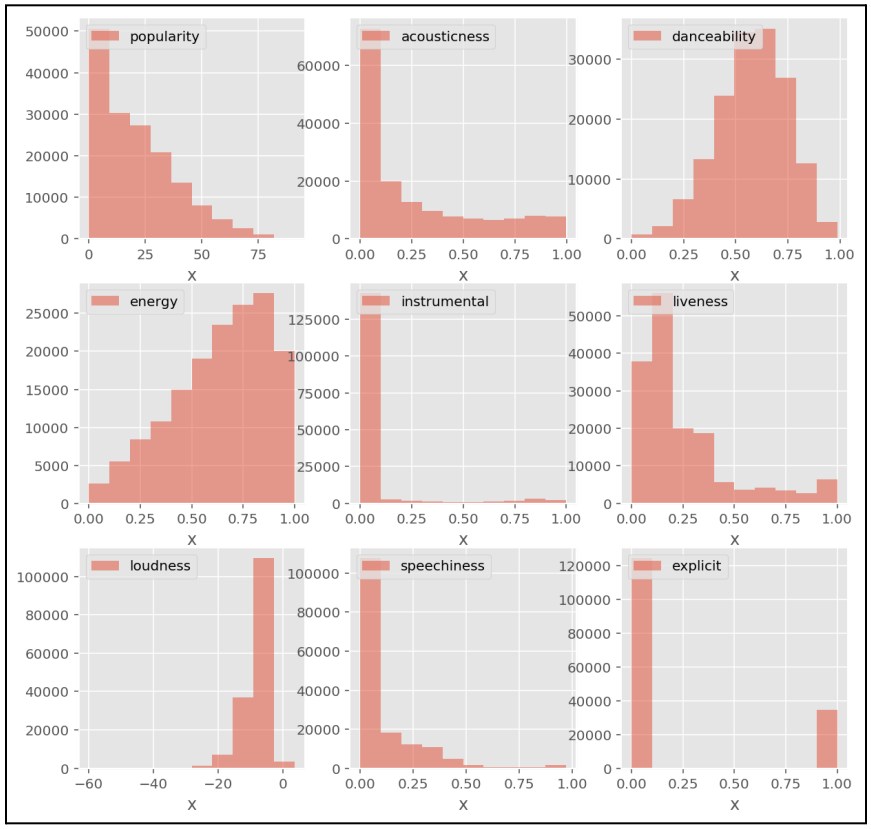
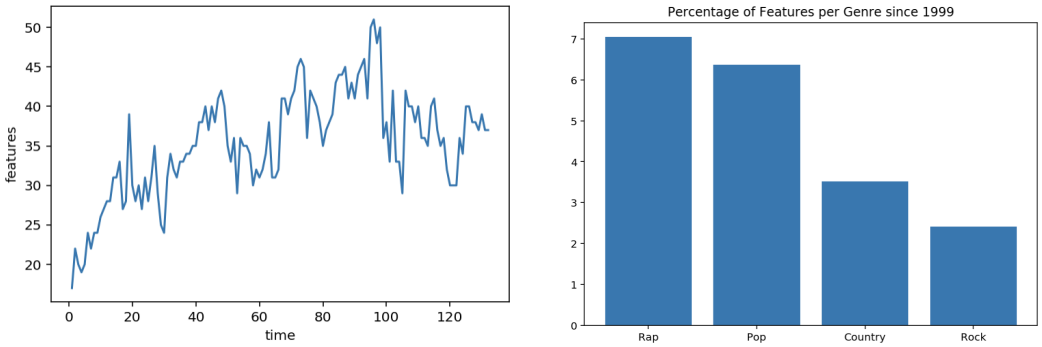


Figure 1. This figure displays the different overlying trends of the attributes of the songs in 2019

We have also noticed a heavy trend, just in the past few years of listening to music, that rap music has overtaken much of what we listen to. This is fitting, because Hip Hop is now the most popular music genre, surpassing Rock music. We analyzed the amount of features in the `spotify_top_200` table , and there seems to have been an upward trend from the beginning of 2017 to 2019. Coincidentally, with the rising popularity of rap music, we found that the percentage of top rap songs in the `billboard_hot_100` dataset that contained featured artists were higher than any other genre. This percentage correlated with the rise of rap music, and rise of features in the most popular songs.



Left Figure. Number of songs with featured artists (out of 200)
Right Figure. Percentage of top songs with featured artists, by genre

When looking at the data for top songs that incorporate features, we noticed a slight trend between late 2016 and 2019. Our data was taken weekly since the last weekly top 200 release in 2016. This graph plots the number of songs in the top 200 per week that contained features, in comparison to 200 total songs. We saw a plateau in October 2018 with over 25% of the songs in Spotify's top 200 containing a featured artist.

By looking at all songs in the Billboard Hot 100 since 1999, we can see that the percentage of songs with features per genre is historically high in rap. As witnessed in our culture today, rap and hip hop have overtaken charts and become the highest selling and streamed genre, and the spike in songs that feature outside artists have supported this genre's influence on our daily listening. This also shows that rap, as a genre, is becoming much more streamlined and efficient in the sense of collaboration – many rap artists are beginning to recognize that they cannot make top selling music alone, and by working with other artists, more creativity can be used towards higher quality (and best-selling) music.

Analysis was also done to investigate how the attributes of popular songs have changed over the past twenty years. The mean value of the attribute for songs on the Billboard Hot 100 (weighted by number of weeks the song was on the Hot 100), for each of the 13 attributes was computed for each year. The results were plotted to identify interesting patterns. A collection of some of these plots are included in **Figure 2**, below. The rest of the graphs that showed interesting patterns are included in the Appendix.

Figure 2

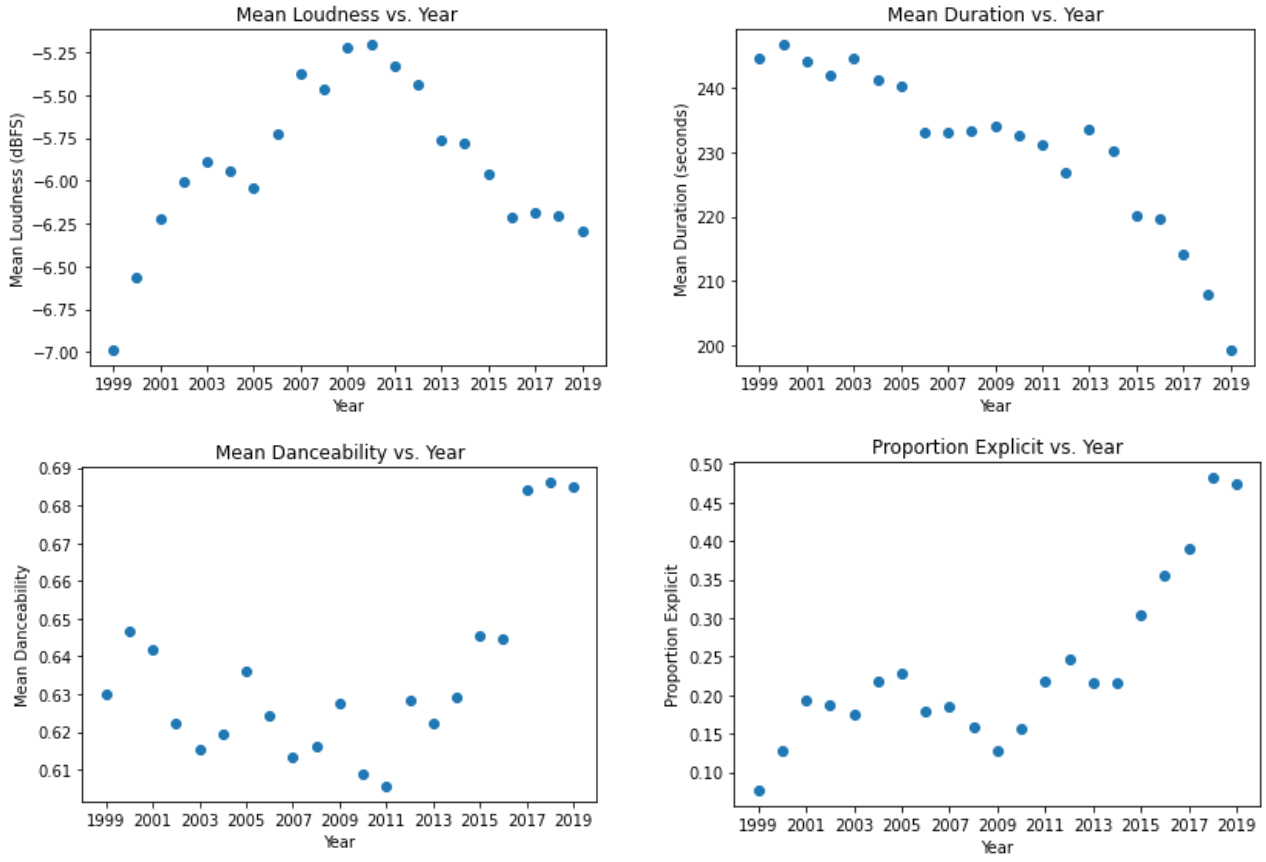


Figure 2 shows the mean value of various attributes of songs on the Billboard Hot 100 since 1999

Some attributes, including instrumentalness, did not show any meaningful patterns, but many attributes did. Danceability and acousticness reached their lowest values in 2010, and have been increasing since. Conversely, energy, loudness, and tempo all reached a peak in 2010, and have been decreasing since. This result for loudness matches known patterns of music becoming louder in the 2000s. Additionally, valence and duration have consistently fallen since 1999, with the duration of the average Billboard Hot 100 song falling from four minutes, to about three minutes over that time period. Finally, the proportion of explicit songs has increased by nearly five times since 1999, from 10% to nearly 50%. It is possible that this is due to the rising popularity of Rap music that we noted earlier.

Additionally, an analysis was performed to determine how many songs reach the Billboard Hot 100, and how long they remain there. The results are shown in **Figure 3,** below.

Figure 3

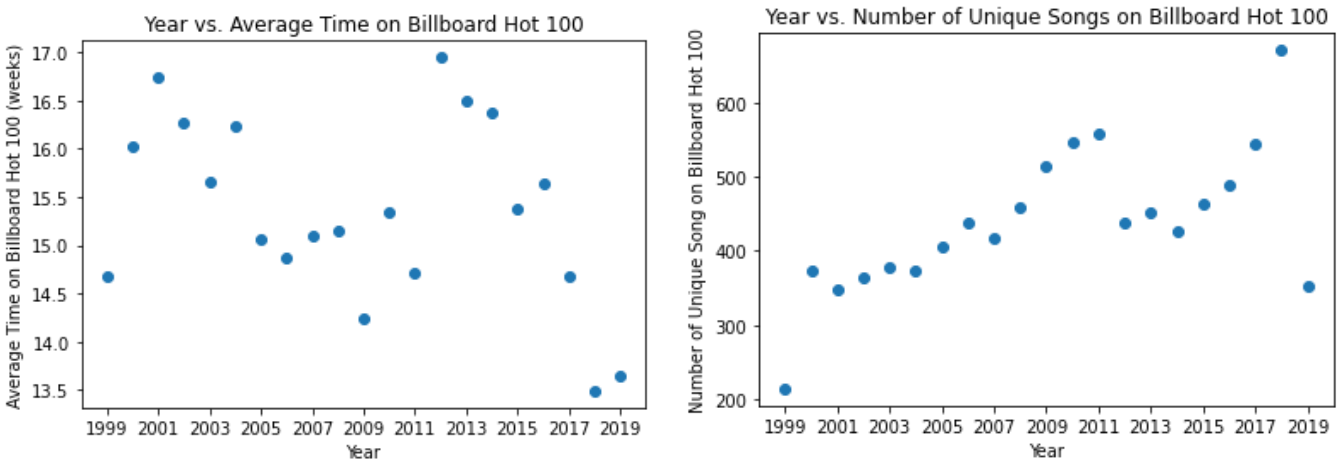
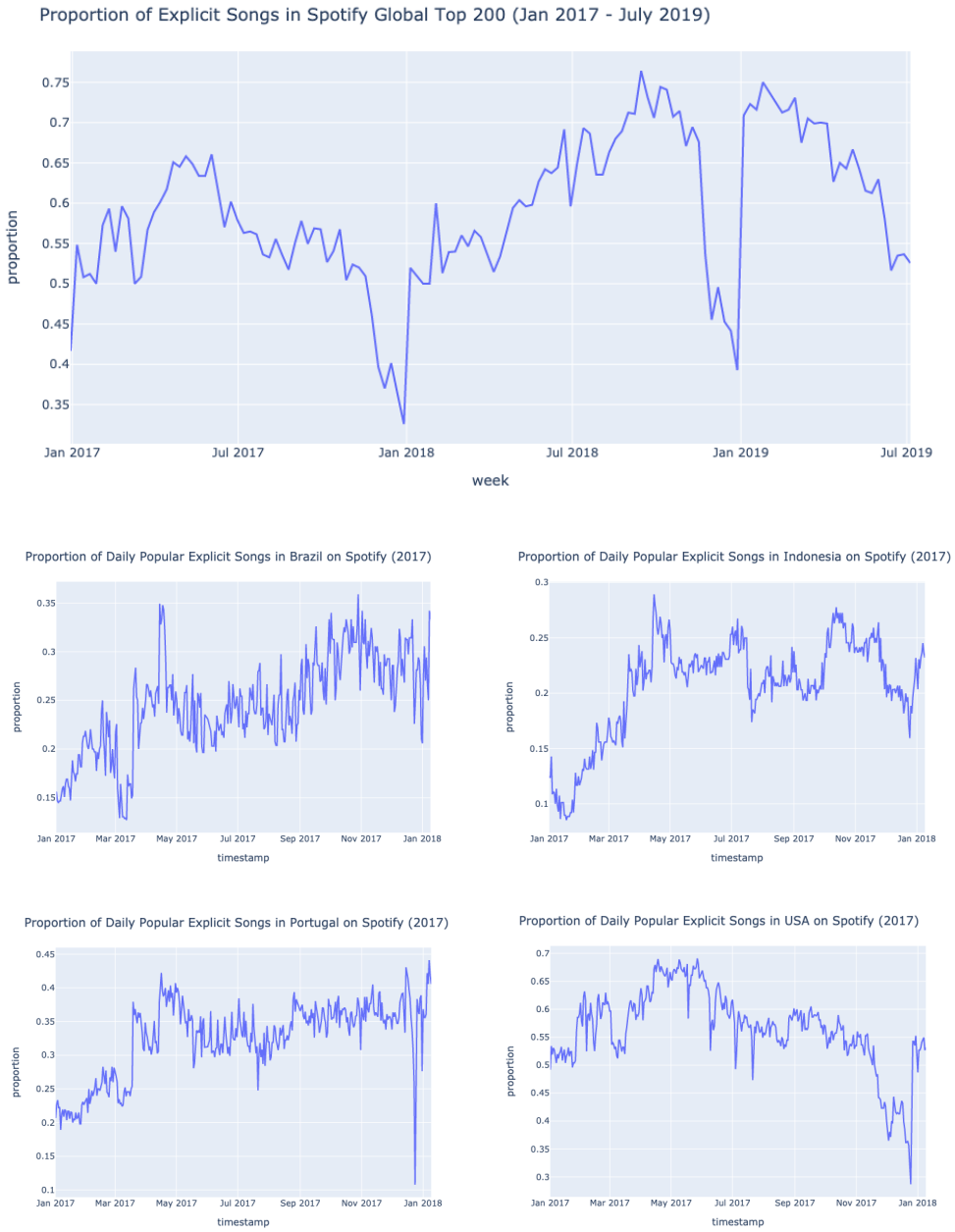


Figure 3 shows the number of unique songs that reached the Billboard Hot 100each year between 1999 and 2019 (left), and the average number of weeks that a song that reached the Billboard Hot 100 remained there.

Although these plots show less clear patterns than the song attribute analysis, they still show interesting patterns. Generally, it seems that more unique songs are reaching the Billboard Hot 200, nearly doubling from 2000 to 2018. Average number of weeks on the Billboard Hot 100 seems more noisy, but there has been a consistent drop from seventeen weeks to fourteen weeks between 2011 and 2018.

The proportion of popular songs that are explicit was particularly interesting to our group. Over the time range 2017 to 2019, we can see that the proportion of explicit songs appears to be a bit cyclical throughout the year, despite the fact that explicit songs are becoming more popular overall. In Figure 4, it is clear that the last weeks of the year have the lowest proportion of popular explicit songs, likely due to users listening to Christmas music, while the rest of the year enjoys a much higher proportion of explicit songs.

Figure 4



Not every country reported in the Spotify dataset mirrors the global trend of decreased popularity in explicit songs in late December. Out of the above selected countries, this trend is most prominent in Western Europe, North America, and Japan (no other East Asian countries have data). In general, the proportion of popular explicit songs is highest in these countries. Brazil, for example, experiences a sharp decline in the prevalence of explicit popular music in late February - Early March 2017, corresponding with the national holiday Carnaval. Carnaval is held the Friday afternoon before Ash Wednesday and celebrates the beginning of Lent, a Catholic holiday. There are other countries that mirror a weaker version of this pattern, including Argentina, Brazil, and France, all of which have decent sized Catholic populations.

III. Challenges

We had a bit of trouble also setting up Jupyterhub. One of our group members has yet to take a class where they would learn more about graphical design using Jupyterhub, so there was a bit of a learning curve between project partners with the amount of data and graphing tools used.

Another challenge was loading the data into the database. Although many tables required a relatively simple copy command, some tables required additional steps to make them usable for our analysis. In particular, the `billboard_hot_100` table had inconsistent ways of indicating null values, with some columns using the string "NA," and other columns using empty strings. To resolve this issue, these problematic columns were loaded in the database as `TEXT`, marked explicitly as null based on how that column indicated null values, and then converted to the appropriate type. Additionally, the "explicit" column of the `song_attributes` table was initially loaded as a `BOOLEAN` column, but was transformed to store integers, making it easier to find proportions by using the aggregate function `AVG`, rather than more complicated alternatives.

There were also challenges with pulling data from Spotify. In particular, Spotify has multiple versions of the same song, often with the same song attributes. We resolved this problem by pulling the first version Spotify listed. Other datasets also didn't list track names and artists in the same way Spotify expected them to be queried, so punctuation was stripped from track and artists names before querying the Spotify API. This still resulted in a small number of songs that weren't found in the Spotify database. Finally, Spotify applied rate limits to pulling data through their API, which required workarounds to obtain a complete dataset.

Bibliography

[1] J. Gottlieb, “New Harvard study says music is universal language,” *Harvard Gazette*, 25-Nov-2019. [Online]. Available: https://news.harvard.edu/gazette/story/2019/11/new-harvard-study-establishes-music-is-universal/.

[2] Published by Statista Research Department and J. 8, “Music listening habits in the U.S. by age 2019,” *Statista*, 08-Jan-2021. [Online]. Available: https://www.statista.com/statistics/749666/music-listening-habits-age-usa/.

[3] “Company Info,” Spotify, 28-Apr-2021. [Online]. Available: https://newsroom.spotify.com/company-info/.

[4] “Data on Songs from Billboard 1999-2019,” *Kaggle*, 03-March-2020. [Online]. Available: https://www.kaggle.com/danield2255/data-on-songs-from-billboard-19992019

[5] “Spotify’s Worldwide Daily Song Ranking,” *Kaggle*, 12-Jan-2018. [Online]. Available: https://www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking/metadata

[6] “Web API,” *Spotify*, 2021. [Online]. Available: https://developer.spotify.com/documentation/web-api/

Appendix

Figure 4

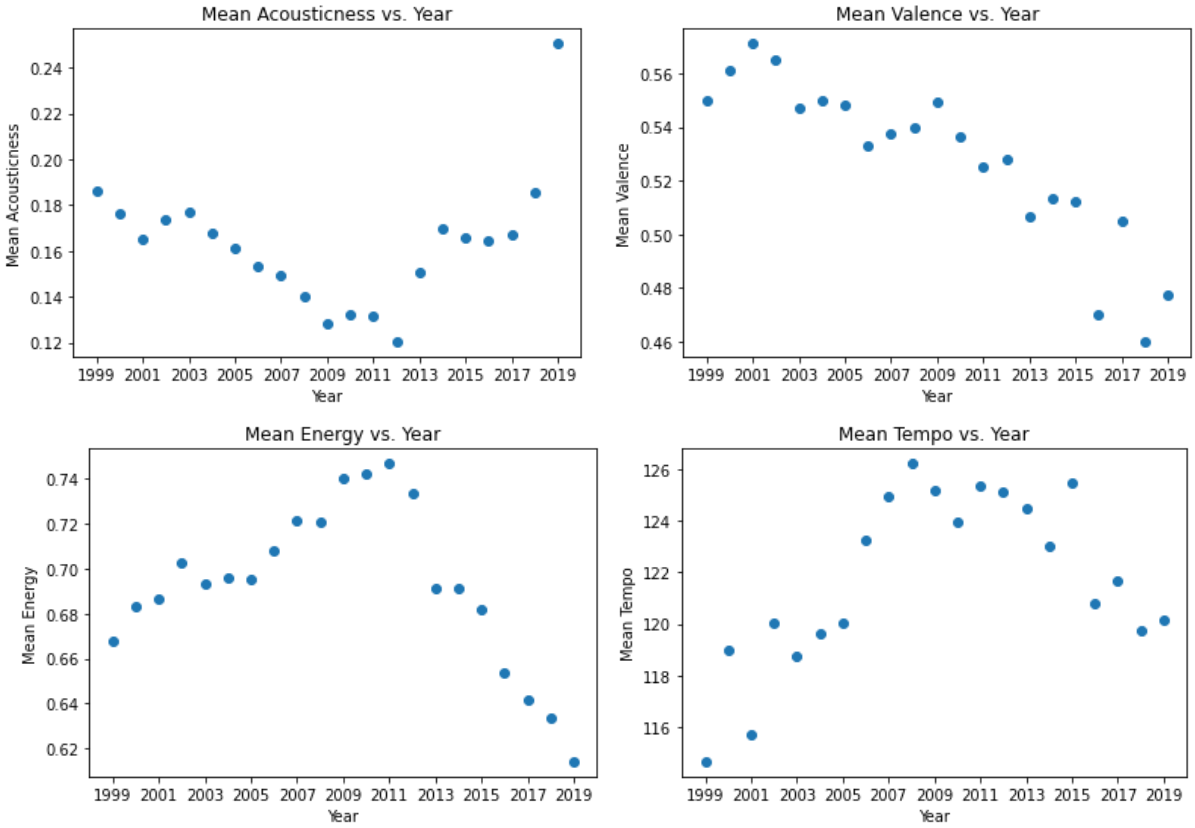


Figure 4 shows additional patterns for song attributes in songs on the Billboard Hot 100 between 1999 and 2019