What Makes Popular Music?

Marc Stern

1. Description

1.1 Basic Information

There are two major metrics in 2021 that determine popular music in the United States music industry: Billboard charts, and streaming popularity.

With the Billboard charts, there is one major chart called the Hot 100, which is supposed to represent the most popular 100 songs in America for a given week. However, there is a long-standing notion in the music industry that big-money interests, such as wealthy artists and record labels, often influence the Hot 100 chart, misrepresenting what are truly the most popular songs of that week. This hypothesis of chart manipulation has been written about recently in Digital Music News, and dates back even further to a 1996 New York Times article, showing that the topic has persisted through the various changes in the music industry, pre and post digital.

With streaming, there is more data than ever that is recorded for each song. This includes meta-data, such as song names, artists names, and genres. There is also numerical data that analyzes the musical content of each song, such as fundamental musical metrics like tempo and time signature, as well as more interpretive metrics such as energy, danceability, and valence.

1.2 Project Objectives

This project seeks to answer two major questions:

1) Is the Billboard Hot 100 chart representative of the most popular music?
2) What makes music popular?

1.3 Description of the Data Set

I have downloaded three data sets for this project to ensure that all of my bases are covered for drawing my conclusions.

The first data set is a data set compiled from the Spotify API to include the musical elements of over 600,000 songs released between 1922 and 2021. This data set will be crucial in providing context for the songs in the Billboard Hot 100 by including popular songs that do not live on the Billboard chart. One key metric from Spotify's API is the popularity rating.

The other two data sets are specific to the Billboard Hot 100, including data from 1958 through 2020. One data set is taken from Billboard's website itself. Another is compiled from the Spotify API to include the musical elements of the songs from the Billboard chart. This will be my main data set to do genre and other musical element trends.

The Spotify API musical metrics that will be used for analysis below, with descriptions of the metric taken from the API documentation.

- **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** – 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.
- **Energy** – 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.
- **Explicit** – Whether or not the song has explicit content.
- **Genres** – A list of the genres used to classify the album.
- **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.
- **Popularity** – The popularity of the track. The value will be between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are.
- **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.

2. Exploratory Data Analysis

2.1 Data Preparation

The process of data preparation begins with loading the three data sets into separate data frames for individual analyses. Knowing that much of my analyses would need to separate the data by units of time, the first operation is to set the indices of the Billboard Hot 100 and Spotify data sets to datetime objects using the week ID for Billboard, and release date for Spotify. I also dropped unnecessary categories like Spotify url from the dataframes. For the Billboard data set with Spotify API metrics, I dropped rows where the music metric data was null, as it would not provide anything for analysis. Later on, I created different data frames using the time series object indices to select different time ranges for data analyses. This includes selecting only the Billboard Hot 100 data from 2017-2020.

2.2 Data Analysis using Descriptive Statistics

One of the primary questions I ask is: how accurate does the Billboard Hot 100 chart represent popular music? Since Spotify has a lot of songs, most of which are not popular, I needed to understand the range of popularity I was working with and how it compared to Billboard. One problem I encountered was that Billboard's data dates back to 1958, and while Spotify has songs that were released as early as 1922, there is more to consider here. First of all, Billboard represents popular music in the week at a particular time, while Spotify's popularity metric is ever changing, and is as updated as recently as this data set was captured from the API, in April 2021. It is safe to assume that older music is no longer as popular, despite it once upon a time having charted on the Billboard Hot 100 chart. Additionally, Spotify was founded in 2008, making it less than 15 years old. Music older than 15 years was not new when it was released on Spotify, lowering the likelihood of that music to be popular on the service. To solve this issue of relevance being a primary factor in Spotify's popularity algorithm, I decided to use only Billboard data from 2017-2020, as mentioned in section 2.1, so that the comparisons would be fairly recent, and more accurate. Spotify, the leading streaming service, also has become most popular, dominating the music industry listening market, as early as 2015, but more so by 2017, as seen in this separate dashboard I created in Tableau from Recording Industry Association of America data. A final note: songs can remain on the Billboard charts for as long as they are deemed among the most popular 100 songs, and old music can re-enter the charts, as seen with many christmas songs. Therefore, I did not see it fit to only include songs from the most recent year as that could exclude older songs that have lived on the Billboard charts for more than a year.

With the above explanation, I first looked at the unique songs from Billboard between 2017-2020, and compared them to the entire Spotify data set. Of the 1,559 songs on Billboard, only 59.1% were found in the Spotify data set. Then, of these 922 songs in common, only 33.4% were found in the 1,559 most popular songs on Spotify. This shows that at least in 2021, only 33.4% of the songs on the Billboard Hot 100 chart were still among the most popular music. It does not definitely say that Billboard is not accurately representing popular music, but

the remaining 66.6% did not meet the test of time with regard to popularity. It's also worth pointing out that Spotify had an average popularity rating of 25.69, which was far lower than Billboard's 72.39 (for 2017-2020), and even farther from Spotify's top songs at 79.47. Another pointer to Billboard not being aligned with Spotify in terms of popularity.

2.3 Other Techniques

Genre is a particularly messy data metric as there are technically no limits as to what a genre can be called, and the relationship between songs and genres is one to n, as one song can have multiple genre tags. In order to solve this problem, I looped through the data, in one case counting the genre tags, and in a second case adding the song ID to lists within a dictionary, for more music metric analysis. However, in both cases I used the most popular seven broad genre categories: Pop, Hip-Hop (also matched on "Rap"), R&B (also matched on "Soul"), Rock, Country, Dance (also matched on "House"), and Latin (also matched on "Reggaeton"). This accounts for the majority of the 804 songs with genre data, while the remaining minority is accounted for with Other, and NA (for empty strings).

The first instance of genre data is organized by decade, from 1950's to 2020's. It is worth noting that the 1950's and 2020's do not include a full decade worth of data, however I still include them because they are aggregated by the mean, so the scale (from 0-100%) is equal to those with full decades of data.

The second instance of organizing the data by genre, I created a separate data frame for each genre category, dropping unnecessary columns of data like song name and genre, as this data is used for comparing music metrics by genre. I then add a new column for each data frame with the desired genre category, and concatenated these nine data frames into one. With this data frame, I created a new dataframe to keep only the numerical data for comparison. With this data frame, I grouped the rows by the genre, and aggregated the values with the mean.

3. Data Visualization

3.1 Design of Data Visualization

The first question, asking about the accuracy of Billboard's Hot 100 chart, uses visuals to aid the descriptive statistical analyses described in section 2.2. Here, I used basic pie chart to compare portions of the whole when looking at the amount of songs from Billboard found in Spotify, and furthermore in the top 1,559 songs on Spotify. I also used a bar chart to show the popularity differences between Billboard, Spotify, and the top Spotify songs.

The second question regarding what metrics are common in popular music are much more involved. I used two separate categories of visualizations, one comparing metrics between Billboard Hot 100 songs by genre, some of which over time, others using the 2017-2020 data. In order to chart genre distribution over time interactively with Plotly, I had to take a wide data

frame and make it long using the melt function. Then I multiply a list of the year range by nine in order to accommodate for nine categories of data being melted into a single value column.

The second category is comparing Billboard to Spotify, grouping Billboard into categories by year bins (2017-2020 and pre 2017), and grouping Spotify by the top 1,559 songs and the rest of the data. These visualizations look at the metrics of acoustincess, danceability, energy, explicit, mode, and valence using radar charts to easily compare multiple metrics across categories.

3.2 Implementation of Data Visualization

A. Spotify vs Billboard

The first visualization one that consists of two pie chart, the first comparing the total number of Billboard Hot 100 songs from 2017-2020 that exist in the Spotify data set, and the second (on the right) comparing the total number of songs from the 59.1% of the Billboard data set that are in the Spotify dataset, to specifically those in the top 1,559 songs most popular songs in the Spotify dataset. Using color to show the Yes and No categories, with annotations showing the exact percentages.
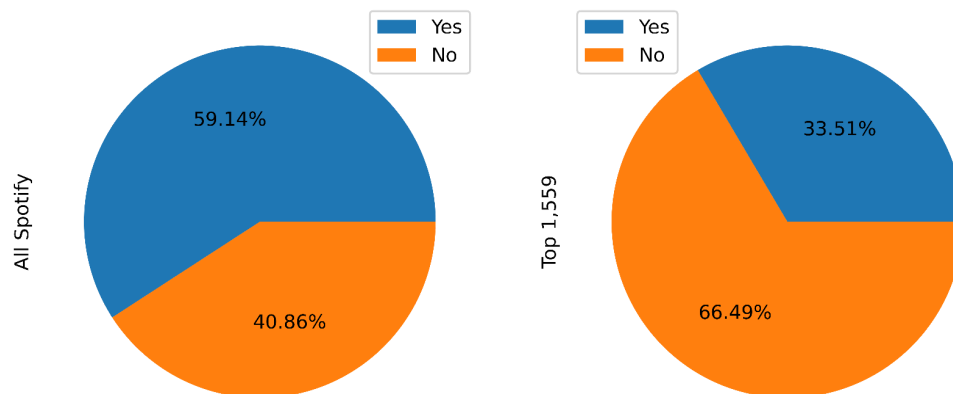
The next chart also looks at a popularity comparison using a bar chart to show the popularity averages for three categories: The whole Spotify data set, Billboard Hot 100 songs from 2017-2020, and the top 1,559 songs on Spotify.

Popularity Comparison
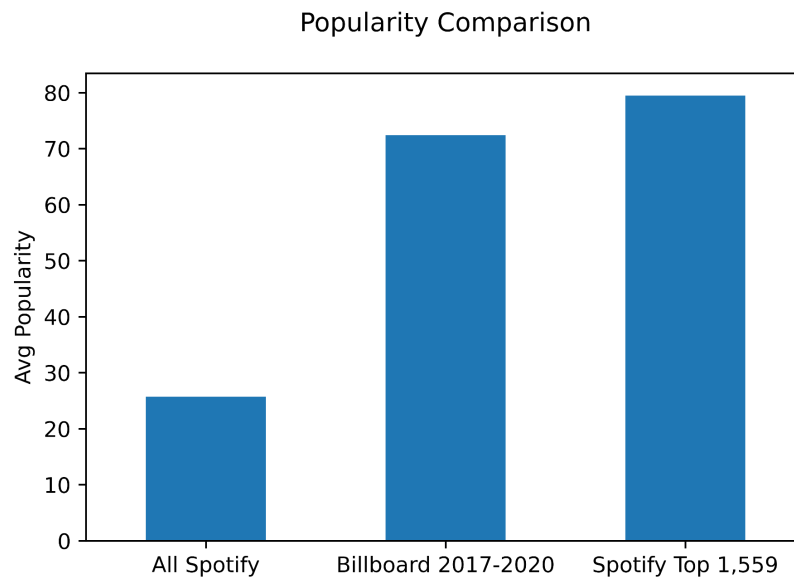
The final comparison between Spotify and Billboard uses a radar chart to compare six musical metrics over four categories of data. Color allows the categories to be easily determined for comparison. This chart in particular shows that "Other Spotify", meaning not in the most popular 1,559 songs, show a much lower rating in energy and danceability than the other three categories, each which belongs to a broader category of popular music.

**Spotify vs. Billboard Hot 100 Music Metric Comparison**
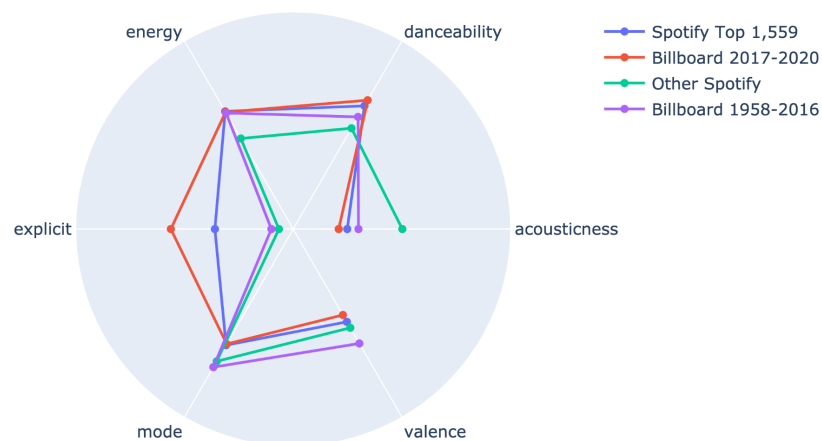
B. Billboard Hot 100 Genres Over Time

The next two visuals show genre tag distribution by genre over time. The first line chart shows the average % of genre tags per genre over time grouped by decades. The following line chart groups by year from 2000-2021, to show more specific trends this century. Color is used to discern the categories in the charts. A user-defined color map dictionary was created to keep consistency between all the visuals in this section.

Avg % of Genre Tags on Billboard Hot 100 per Decade

Avg % of Genre Tags on Billboard Hot 100 in the 21st Century
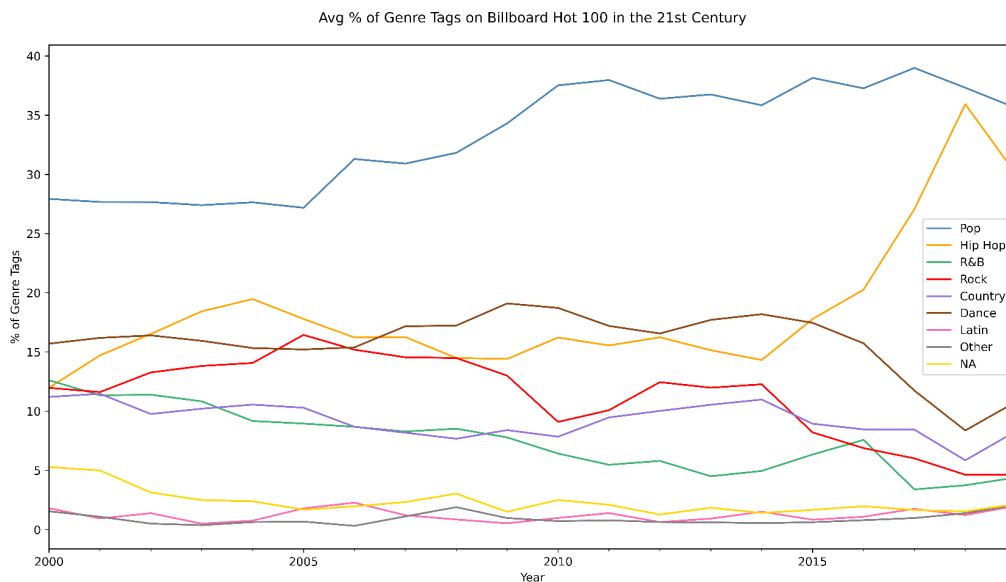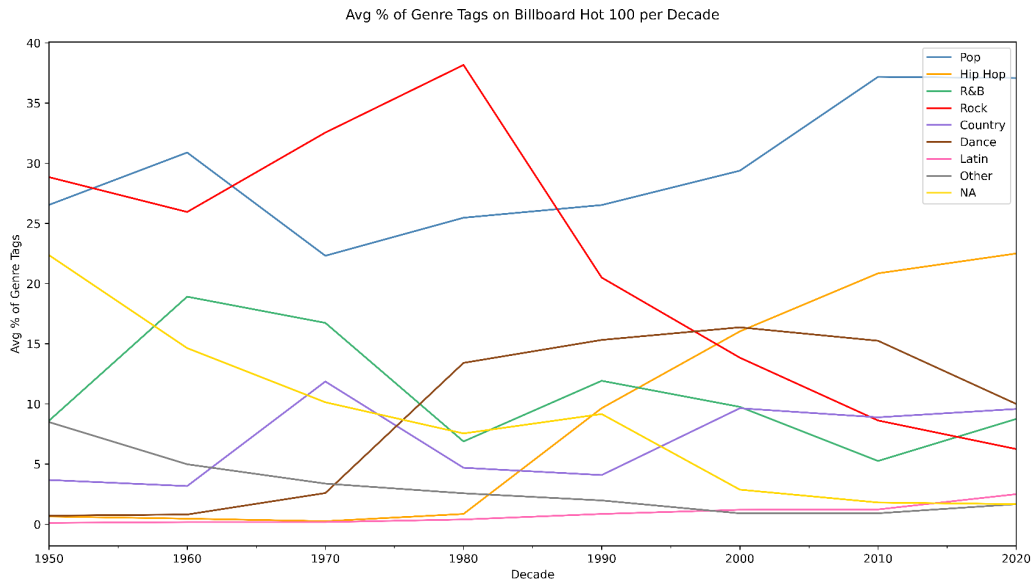
The third visualization in this section uses the same color scheme for clarity, however is an interactive bar chart to display the distribution of genre tags over time for each year between 1958 and 2020. The following visual is not interactive while opened in PDF format.

Genre Distribution of the Billboard Hot 100 Chart Over Time
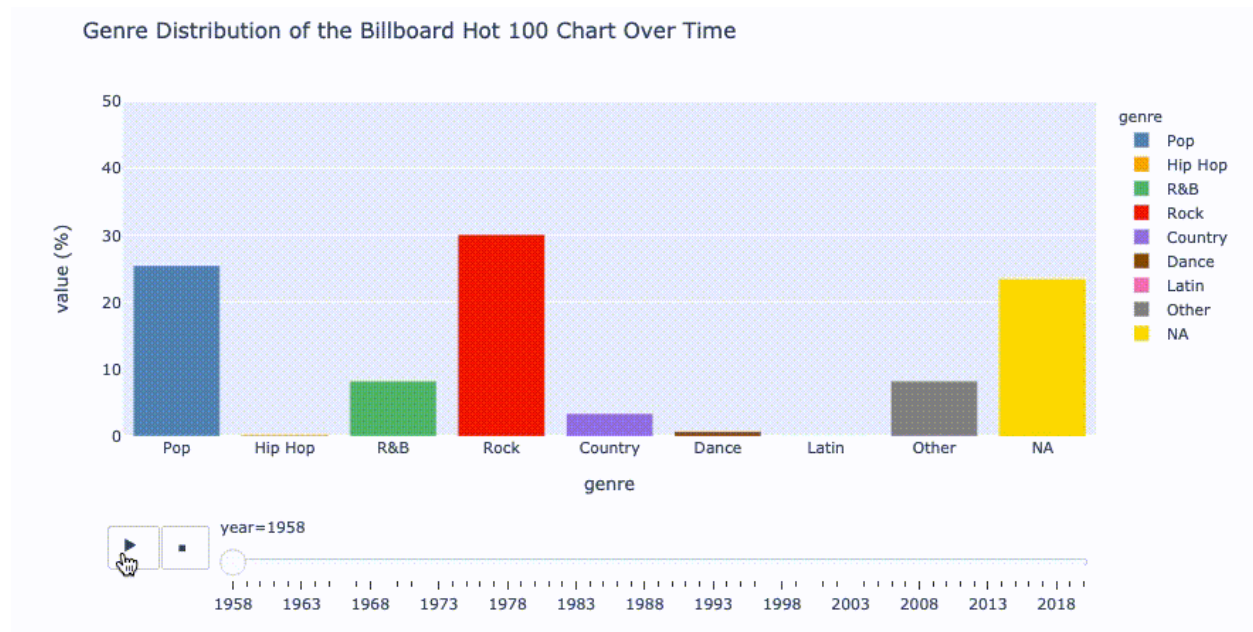
All three charts in this section show a steep rise in Hip Hop since 2015 in the Hot 100 chart, with a steady incline since the 1980's. Meanwhile Rock, which dominated the 1970's and 1980's, and since fallen to an all time low in 2020.


C. Billboard Hot 100 Popularity by Genre

The two bar charts in this section are easy to understand, and use the Billboard Hot 100 data from 2017-2020. The first bar chart shows the breakdown of genre in this grouping by percentage. The second chart shows average popularity rating on Spotify by genre. Both charts exclude the "NA" category, as it is not useful for comparison here. What is interesting in these two charts is that while Pop and Hip Hop make up the majority of Billboard charting songs in this time frame with over 30% each, they both fall in the bottom half of the popularity chart. Latin music on the other hand only accounts for less than 2.5% of the charting songs, yet, the songs that do chart, are of the most popular category.

## Genre Tags for Billboard Hot 100 Songs from 2017-2020
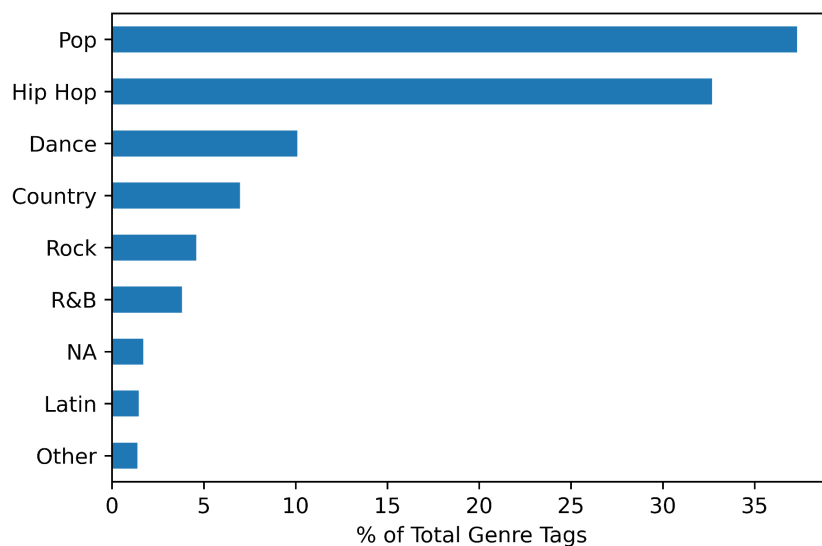
Spotify Average Popularity Ratings for Songs
Charting on Billboard Hot 100 from 2017-2020
by Genre
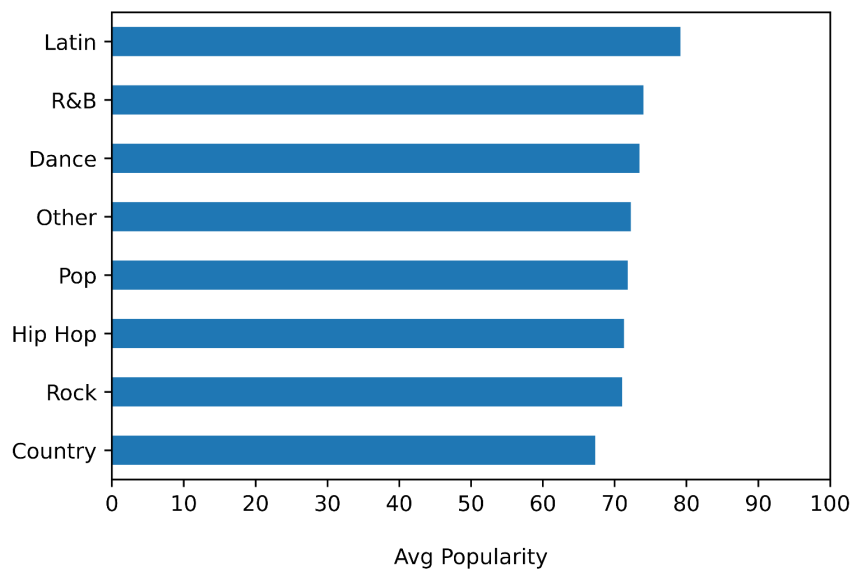
D. **Billboard Hot 100 Music Metrics by Genre**

The final chart is another radar chart that uses the same metrics as the one in section A, however the categories this time are by genre, once again excluding "NA" as a genre category. This is a particularly interesting chart as the colors help identify the categories to easily compare some stark differences between genres.

Billboard Hot 100 2017-2020 Music Metrics by Genre
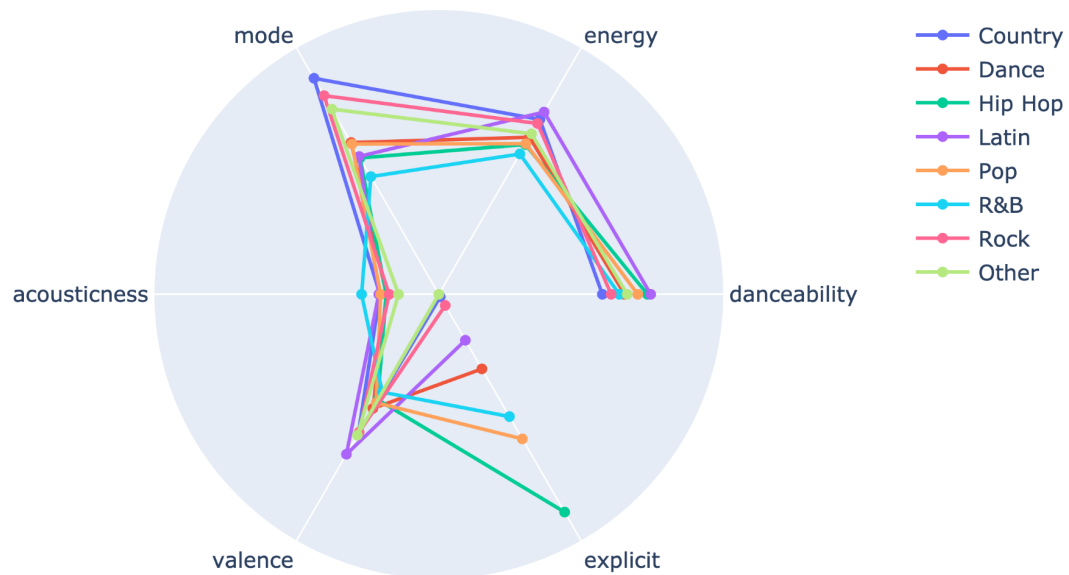
4. Conclusion

In conclusion, I ask again the two main questions of this research: Does Billboard represent the most popular music? What makes music popular?

With regard to the first question, this study can not give a conclusive answer. There are certainly some areas that are worth looking into, but the discrepancy in data between the Billboard data set and the Spotify data set, being that not all of the Billboard songs were in the Spotify data set make it difficult for a direct comparison. It is also important to note again that there are many issues with comparing these two data sets, namely that the Billboard Hot 100 chart represents the most popular music in a given week, while Spotify's popularity rating is not fixed by the week, but instead constantly updating. The data show that it is an area worth

investing further, because it also did not explicitly show that Billboard is accurately representing the most popular music on Spotify, which is the most popular way of music listening in 2021. If I were to draw a specific conclusion, I would argue that there are discrepancies between the Billboard charts and the most popular music on Spotify. To best answer this question, I would need to compare the Billboard songs each week to the updated Spotify popularity ratings for the top 100 songs on Spotify for an extended period of time to determine a stronger conclusion.

The second question, regarding musical metrics shown in popular music does have a stronger answer. It is apparent that Hip Hop and Pop music are the most likely genres to be popular, while if a Latin song makes it to the Billboard Hot 100 chart it is likely to be a monstrous hit. It is also apparent that popular music across genres has a higher rating for energy and danceability than less than popular music. Popular music is less likely to be acoustic, relating to danceability and energy. Interestingly, Hip Hop music, the most common music on the Billboard chart is almost always explicit, showing that explicit lyrics do not hinder the ability for music to become popular, at least among Hip Hop, R&B, and Pop music genres. Explicitness also goes with songs being in a minor key, and conversely against high valence scores. This makes sense as explicit lyrics would not seem to fit with happy sounding songs in major keys. Overall, the data show that high energy music that people can dance to is the foundation of music that charts on Billboard Hot 100, a pinnacle landmark for artist success.