CS 3300 Project 2 Report: Evolution of Acousticness in Popular and Unpopular Songs on Spotify (2010-2022)

Nuray Ozden (nyo3), Richard Pignatiello (rap347), Richard Kim (rk625), Yuki Suwabe (ys462)

Dataset Description

The two original datasets we chose for our visualization relate to songs on Spotify, a music subscription streaming service. The first dataset is called 'Spotify Top Hit Playlist (2010-2022)' from Kaggle, and it contains information on all of the 100 songs on Sptofy's 'Top Hits' Playlist every year between 2010 and 2023. There are 23 attributes in this dataset, but we chose this specific subset for our visualization: *acousticness*, *danceability*, *energy*, *instrumentalness*, *liveness*, *loudness*, *speechiness*, *tempo*, and *valence* for attributes with 'Float' values, and *artist_popularity*, *duration_ms*, *key*, *mode*, *track_popularity*, *year*, and *time_signature* for attributes with 'Int' values.

The other dataset we used for our visualization is 'Spotify unpopular songs' from Kaggle, which contains audio characteristics on the 10,000 most unplayed songs on Spotify as of 2022. This dataset contains 17 attributes, and the subset we chose for our visualization is the same as the 'Spotify Top Hit Playlist (2010-2022)' for the 'Float' attributes, but only *key*, *mode*, *popularity*, and *duration_ms* for the 'Int' attributes since this dataset did not contain the others.

Lastly, we decided to add a third dataset called 'Genre of Artist', which has 700 different artists and their genre(s) for an extra visual representation. It was a second file in the 'Spotify unpopular songs' dataset from Kaggle and has attributes *artist_name*, *artist_id*, and *genre*. However, some reformatting had to occur for the *genre* to be joined properly by the artist. Since the *artist_id* attribute does not exist in the other two datasets, it had to be joined on *artist_name*. Additionally, some of the artists had multiple genres; for example, in 'Genre of Artist', the artist 'Wild Powwers' has the genres of ['cascadia psych', 'seattle indie'] in an array format. We needed to ensure it would not come in as a string, but an array, so the necessary transformations occurred as the data was being loaded.

The idea behind using both of these datasets, which contain the 'best' and 'worst' songs on Spotify, is to allow the user to compare and contrast the attributes between the two groups while allowing for filtering on other attributes. While there is a discrepancy in the range of years, we chose to tackle this by comparing across different years for the 'best' songs, and just 2022 for the 'worst' songs, on each feature. To accomplish this goal, our group created multiple visualizations that sit on top of one another so that the user may choose what attribute(s) to compare the averages of the 'best' and 'worst' songs on, while also looking at how the comparison changed over time.

Visual Design Rationale

Our visualizations contain two types: the default visual, which is a scatter plot, and the specific category visualizations, which are each line graphs.

The scatterplot is the general overview of the average feature values for different musical attributes across different years and the average unpopular score. The x-axis has the different features we chose to analyze, and the y-axis has the average feature score of those 100 songs of that year, where each of the green dots represents one year and the black dot represents the unpopular songs. Arranging the visual in this manner allows the user to get a holistic view of a few different observations; how the different features have changed over time (how the dots change color), the magnitude of how much the features have changed over time (how wide the dots spread), how the popular songs compare to the unpopular songs in each feature (comparing to the black dot with a black line), and how these comparisons differ across features (how does the average unpopularity score differ in danceability vs. energy?) Thus, many different analyses can co-occur in this visual. Additionally, you can hover over the spots to get the specific values, since it may be difficult as you go further right. The marks in this visual are the dots and small green horizontal lines indicating the unpopularity dot, and the channels include the color and position of the dots on the x and y-axes. In order to clearly articulate the difference between the popular group and the unpopular group, the color scales over the years were made to scale over a fairly similar color, although distinguishable. Although this allowed us to clearly contrast the difference between the two groups it did make it harder for the user to be able to distinguish between each year.

The line plots allow the user to take a closer look at the selected audio feature by choosing from a drop-down menu below the visual. The y-axis represents that feature's average score, and the x-axis is the year. Each line plot's green line takes the 100 songs that were Top Hits that year, obtains the average score on that specified feature, and displays how that changed between 2010 and 2022. The black line is the average across the 100 unpopular songs in 2022, so it is a straight horizontal line. This line plot compares the 'best' and 'worst' songs based on the chosen feature and shows how the 'best' score has changed between 2010 and 2022. Additionally, you can hover over the black line to see the specific value of the unpopular metric for that category as well as hover over each year on the green line to see the specific metric of the category from that year. The marks are the lines, and the channels are the colors (black vs. green) and the position on the x and y-axis. This visualization allows the user to see both the difference between the unpopular score and the popular scores per year as well as the trend for specific features over time. However, this visualization made it difficult for viewers to see individual points every year and offers no way to compare the unpopular and popular score per year since there is only an average unpopular score over all the years, not per individual year.

Overview of Design

Many design decisions went into our two groups of visualizations. We first made our scatterplot visualization to compare how each unpopular metric did compared to their popular counterparts. Because of the tradeoffs mentioned above we decided to include a hover function to allow users to see the metric of the specific years to give them more information. Through making this visualization we noticed a general trend in some of the colors, which indicates the year, which suggested how each category changed over the years based on trend. This prompted us to explore further how each category evolved over the course of 13 years and how those changes compared to the general metric of the unpopular songs.

As for choosing the scatterplot and the line plots, the scatterplot made the most sense because it allowed multiple features to sit next to one another horizontally, for comparison, while being able to look at the feature scores and how they changed over time with a dot per year. This made much more visual sense than a bar graph, given there are thirteen years between 2010 and 2022 as well as the unpopular score; it would have been too overcrowded and very difficult to extract any meaning from. The line plot seemed like the obvious choice since time was on the x-axis, and we wanted the user to see how the score changed over the years. Additionally, having the drop-down menu to switch between the scatter plot to the different line plots is a user-friendly, familiar, and fun way to lay each visual on top of one another and allow the user flexibility in what they want to see, as opposed to just putting each one next to each other.

The Story

Music is one of the largest forms of media consumption today, especially for college-aged individuals such as us. We hope that our visuals convey a story about the music in Spotify, the #1 music streaming platform in the world, and uncover insights on what makes a song ‘good’ and what makes a song ‘bad’. There are some surprising findings that our visuals display; for example, the unpopular songs all have higher average acousticness and higher average instrumentalness than the popular songs across all years from 2010 to 2022. This visualization allows us to take a closer look at the music we listen to and understand the patterns and trends, across features and years, that may not seem obvious.

Contributions

- Nuray wrote the report on the visualization developed while coordinating with her group members to accumulate all the necessary information. She spent around ~4 hours on her work.
- Richard Kim worked on developing and designing the visualization. He spent around ~5 hours on this project.

- Yuki also worked on the development of the visualization. Specifically the main graph and the design of the visualization. She spent around ~7 hours on the project.
- Richard Pignatiello worked on brainstorming ways to visualize the data and data ingestion/clean up. He spent ~4 hours on this project.

Citations

JosephineLSY. (2022). Spotify Top Hit Playlist 2010-2022 [Data set]. Kaggle. https://www.kaggle.com/datasets/josephinelsy/spotify-top-hit-playlist-2010-2022

EstienneGGX. (Year). Spotify Unpopular Songs [Data set]. Kaggle. https://www.kaggle.com/datasets/estienneggx/spotify-unpopular-songs/data?select=z_genre_of_artists.csv