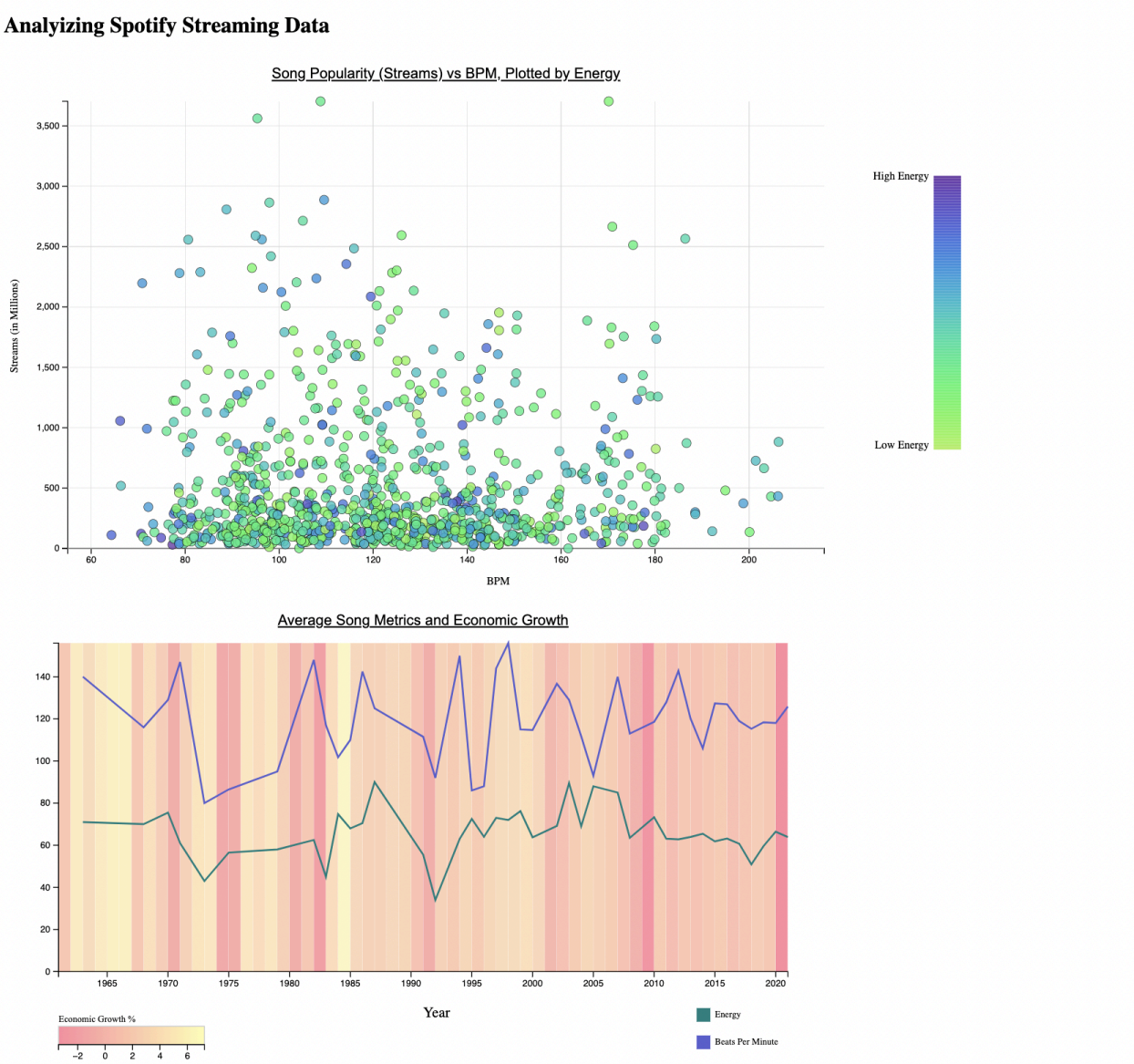
CS 3300 Project 1: Spotify Data Analysis

Suhani Patel, Rhea Verma, Aneesha Kodati

Visualizations



These two visualizations analyze Spotify streaming data, specifically focusing on song popularity and how it correlates with beats per minute (BPM), energy, and economic growth over time. These visualizations aim to uncover trends between song metrics and larger societal or musical patterns.

Data and Data Cleaning

We got all of our data from Kaggle. The first dataset we used was Spotify Most Streamed Songs. For visualization 1, we cleaned this dataset to remove fields we did not need (specifically removing everything except for streams, bpm, and energy_% as well as removing any NAs for released_year value rows). The streams field represents the number of streams of that song on Spotify, the energy_% field represents the amount of perceived energy the song has, and the bpm represents the beats per minute for each song.

The second dataset we used was USA GDP Growth. From this dataset we extracted the year and the growth, and discarded other columns such as GDP per capita and total GDP since these generally always increase due to inflation and did not really give us any relative information on the economic status of the US that year. Since the growth was in a string, we stripped "%" from it, and converted it to a float from 0-100.

The x-axis of this visualization is the year, so we needed to group by year and aggregate the Spotify metrics in order to find the average metrics for beats per minute (BPM) and liveliness per year. We then did a right join on the GDP growth and the Spotify metrics since there we wanted GDP data for every year to fill the heatmap but there weren't all the years in the spotify data.

All of this cleaning was done in a Python file using Pandas with the path data/cleaner.py.

Design Rationale

In visualization 1, we created a scatterplot that uses position to map two continuous variables: BPM on the x-axis and streams (in millions) on the y-axis. Each point represents a song, and the color of the point encodes the energy level of the song, using a gradient from green (low energy) to purple (high energy). The decision to use color rather than size or shape to encode energy was made to avoid clutter and ensure easy comparison across songs. Additionally, purple and green are opposite to each other on the color wheel, providing a clear and identifiable contrast. Instead of varying the size of the dots, all circles have a fixed radius of 5 pixels, which allows us to focus attention on color and position without introducing unnecessary complexity or visual noise. To further improve readability, the circles have a stroke (black outline) and opacity set at 0.85, ensuring that points remain distinguishable even in dense areas while giving a subtle visual separation between overlapping points. The axes are kept linear rather than logarithmic to

preserve the natural distribution of both BPM and stream counts. No size differentiation was applied to the circles, focusing the viewer's attention purely on the relationship between energy, BPM, and popularity without introducing an additional visual variable. The design emphasizes clarity by avoiding overlapping elements and ensuring that color is the primary visual channel to convey energy. A scatterplot is the optimal choice for visualizing this dataset because it effectively allows for a clear and intuitive mapping of two continuous variables, BPM and streams, enabling viewers to quickly identify trends or correlations between song tempo and popularity. Additionally, the scatterplot supports multivariate representation by using color to encode energy, allowing for a third dimension of data to be visualized within a two-dimensional space. Scatterplots are also well-suited to large datasets like the Spotify dataset, as each song is represented by an individual point, allowing for detailed visualization without overwhelming the viewer. Scatterplots are optimal for highlighting patterns, clusters, gaps, and outliers in data, making it easy to spot trends.

In visualization 2, we made a line graph of beats per minute and liveliness over the course of years, and made the background a heatmap with yellow corresponding to high GDP growth and red/pink corresponding to either negative growth or low GDP growth. The reason we decided to do it like this is because it was the easiest way to see how changes in GDP corresponded with both of the metrics, since it was directly behind the line graph. We chose pastel yellow and pink because it was not overwhelming the line graphs or making the graph cluttered, the colors were similar enough that the gradient wasn't distracting, and in general we associate red with negatives and yellow with more neutral/positives.

The reason we made it a line graph rather than a scatter plot is that the actual numbers aren't as important as the relative changes year to year to determine if songs got sadder or slower when there was economic distress, and points were not adding that much information but were making it cluttered. We chose a light and dark teal color for the line graphs because they popped out on the yellow/pink background and the colors are also contrasting and looked good together.

We didn't label the y-axis because the units change depending on which metric you're looking at. We also considered adding more metrics but it made the graph more cluttered without adding much additional information. A linear axis was used rather than a logarithmic one because the log scale didn't really make a difference since the data was evenly distributed across the years.

Analysis and Insights

Visualization 1 provides insights into the relationship between song popularity (streams), BPM, and energy. Surprisingly, there is no strong correlation between BPM and the number of streams; popular songs are spread across a wide range of BPM values. This suggests that BPM alone does not drive song popularity. However, there is a slight tendency for more popular songs to have moderate BPMs (around 100-120 BPM). Additionally, the color gradient shows that both high- and low-energy songs can achieve success, indicating that energy levels are not a definitive factor in streaming popularity either. Songs being scattered across different BPM and energy levels reflects the diverse preferences of Spotify listeners and the multifaceted factors that contribute to a song's popularity.

Visualization 2 attempts to compare the economic state of the United States with how happy, upbeat, or fast the music was. There were some years where there were significant jumps, such as 1973 where growth was extremely high and songs got less upbeat and less fast. However, there wasn’t actually any consistent trend. We also wanted to see if the speed and energy of songs got faster or different over the course of years, but it doesn’t seem that there’s much correlation here either. One interesting trend that we did find however is that GDP growth in the United States as a whole slowed down significantly– although this had no impact on music. We also found a slight correlation between energy and bpm since they spiked at roughly the same times.

Team Contributions

Suhani: I spent time cleaning the datasets in our original plan, (Spotify Most Streamed Songs and World Important Events). This took about 1 hour. Then I designed and created visualization 2 using these datasets, which took about 1.5-2 hours to complete. Then, after speaking with a TA and receiving feedback, we realized that visualization 2 needed interactive elements that would not be used in grading to achieve the goal we wanted. I then worked with Aneesha and Rhea to redesign and rethink our approach - we found a new dataset and designed a new visualization. This took about 1 hour. I then wrote the sections on visualization 1 for parts c and d, as well about cleaning the Spotify data for visualization 1. I then presented the visualizations, received feedback, and made changes based on feedback which took about 1.5 hours.

Aneesha: I cleaned data prior to starting visualization 1, and helped plan it out just to get started, which took around 30 minutes. I also found the new dataset for economic growth, which was not in our plan originally. While we tried to incorporate major world events, it was difficult to implement this without an interactive part so I helped find more numerical data that might have to do with the state of people in the country, and cleaned the data and merged it with another dataset. This took around an hour. I worked on creating the heatmap, graphing the line plots, and playing with the colors and design features as well as labels and legends, which took around 2 hours. Finally, I worked on the data cleaning, design rationale, and analysis and insights for this writeup for visualization 2 which took around 1 hour.

Rhea: I found the first dataset Spotify Most Streamed Songs and formulated the research question we would be plotting which took about 30 minutes. Then, I coded the first visualization comparing Song Popularity by BPM and Energy, which took about 3 hours to complete. I also worked with Suhani and Aneesha to rethink our design for visualization 2 given our pivot to a static plot and reformulated our research questions. Finally, I was responsible for combining the two plots into one file and making the theme more cohesive (fonts, title formats, etc.) to give viewers a good experience, and contributed to writing the report for visualization 1, which took about 1 hour.