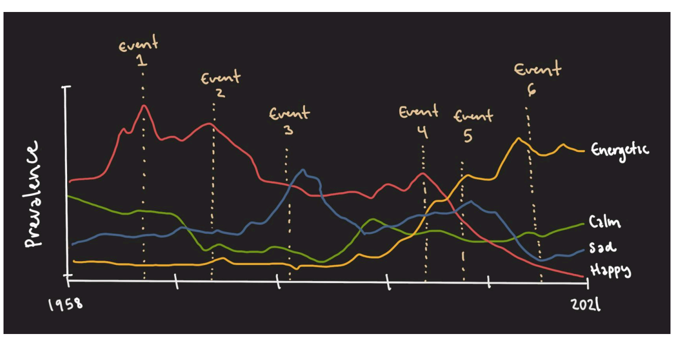
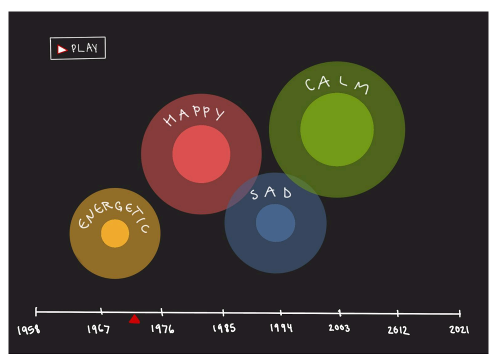
1. Introduction

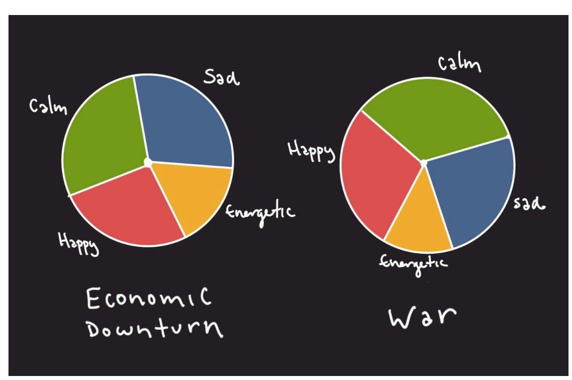
- **Purpose**: This project explores the relationship between historical events and music preferences, aiming to reveal how external events like elections, pandemics, or social movements may impact the types of music people choose to listen to. By analyzing changes in music preferences—particularly the prevalence of different emotional categories (e.g., sad, happy, calm, energetic)—we can gain insights into collective emotional trends during various time periods. These visualizations provide a historical perspective on how music may reflect or influence public sentiment and emotional trends over time.
- **Audience**: This project is intended for individuals interested in the intersections of music, psychology, and history, including researchers, students, and music industry professionals. It is also relevant to anyone interested in understanding how music trends may reveal broader emotional patterns in response to social or political events.

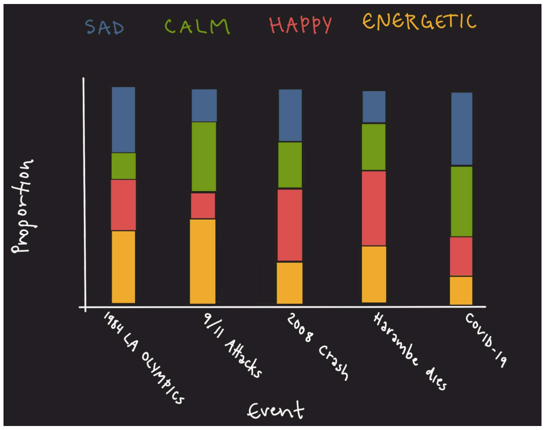
2. Project Overview

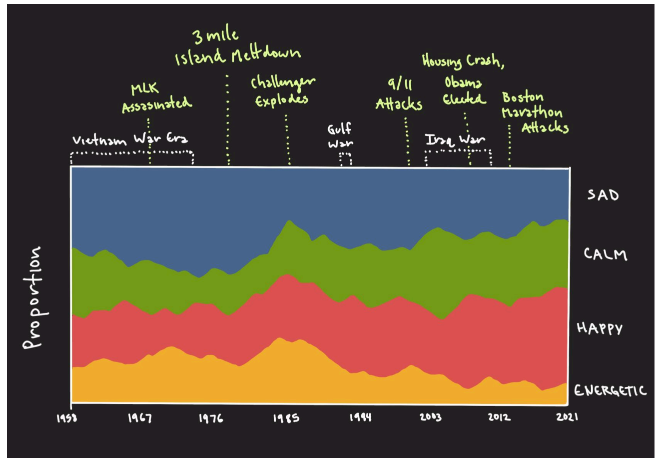
- **Objectives**: The primary goal of this project is to investigate how music and history are correlated. We wish to reveal patterns that teach us something about the role music plays in the public sphere, and how emotions in music reflect the current state of history.
- **Data Sources**: The data is sourced from 3 places. Our first data source is a kaggle dataset titled [278K Emotion Labeled Spotify Songs](). This data contains a large volume of Spotify songs, along with various associated traits found from spotify. This is our main dataset. The second source is another kaggle dataset titled [Billboard "The Hot 100" Songs](). This dataset contains the top 100 songs in the United States every week from 1958-2021, along with their previous position on the top 100, and weeks on the top 100. Unfortunately, these 2 datasets do not have a direct mapping, so we will need to use the spotify API to bridge the two sets. In later stages of the project, we intend to hand-curate a dataset of historical events during the time period being visualized.
- **Initial Design Ideas:** These are some sketches we created for the project proposal:











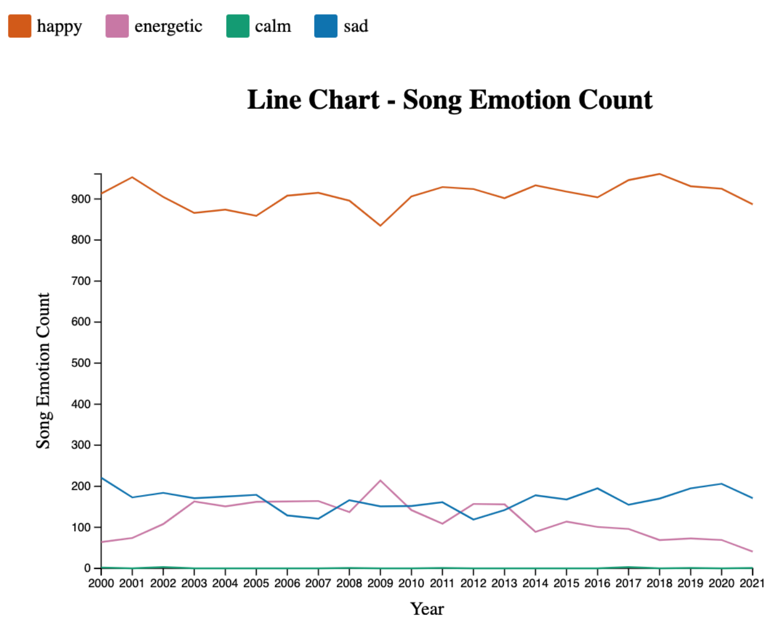
3. Data Preparation

- **Data Collection**: Getting the initial datasets was a very straightforward process. We used the raw CSVs to import Top 100 and Emotion Labeled Songs into pandas dataframes in google colab. We then needed to find a way to map the data from one set into the other, and fill in missing values. The Top 100 dataset included the Song Title, Artist, and other information about the song's rank. The Emotion Labeled Songs dataset included the Spotify URI, the emotion labels, and other audio feature data. In order to join these two datasets, we need to use the Spotify API to search for a song using Title and Artist, get the URI from that search, and join it to the Emotion Labeled Songs on the URI.
- **Data Challenges and Manipulation:** The Top 100 dataset provided the top 100 songs for every week spanning 1958-2021. If we had performed the process described above for every song in the set, taking into consideration the Spotify rate limits, pulling the data would have taken approximately 700 hours. 29 days of data collection was untenable for the scope of the semester, so we decided to pivot. Our first consideration was the time span. Although there are many interesting events that occurred pre-2000, the majority of the class (our primary audience) only has living memory from around 2000 on. Due to this and our time considerations, we decided to limit the scope of our years to 2000-2021. The second consideration was the granularity of the time steps. We found that when a song was in the top 100 one week, there were good odds that it was also present in the other weeks of the month. This inspired us to aggregate our data by month, and only keep unique song values for each month. This cut down the number of individual queries needed by a factor of ~4. We also reasoned that it would be cleaner to represent data with months on the X axis, since representing every week 2000-2021 would have required 1,092 ticks on the X axis, while representing every month would only require 252.
- **Data Cleaning**: We performed a left join in order to use all songs in the Top 100, however, for many of the songs in the Top 100 there was no corresponding song in the Emotion Labeled Songs set. To impute this data we used the machine learning model provided by the authors of the Emotion Labeled Songs dataset that they used to produce the label. This required using the song's Spotify URI to retrieve the necessary audio features, feed them into the model, and save the labels to our dataset. There were a handful of songs that do not exist on Spotify, so we manually added those songs by listening to them and deciding for ourselves which label they deserved.

- **Feature Selection**: The main features that are important to us are Emotion, Date (by month) and Rank. Emotion and Date are self-explanatory, but we chose to also incorporate Rank so that we could see how the popularity of the songs weighted their relative standing in the Top 100.
- **Exploratory Data Analysis (EDA)**: Our initial findings show that the top category is overwhelmingly ‘Happy’ for all months 2000-2021. There are very few ‘Calm’ songs, usually 0 for most months. When we graphed the average rank of each emotion we found that the average rank of each label Happy, Energetic, and Sad are quite comparable. ‘Calm’ usually sits at 0, however when there is a ‘Calm’ song in the month, it usually ranks fairly low. We have not created our historical events labels yet, but we noticed that there were interesting dips in the popularity of ‘Happy’ songs after significant events such as 9/11/2001 and the 2008 economic crash.

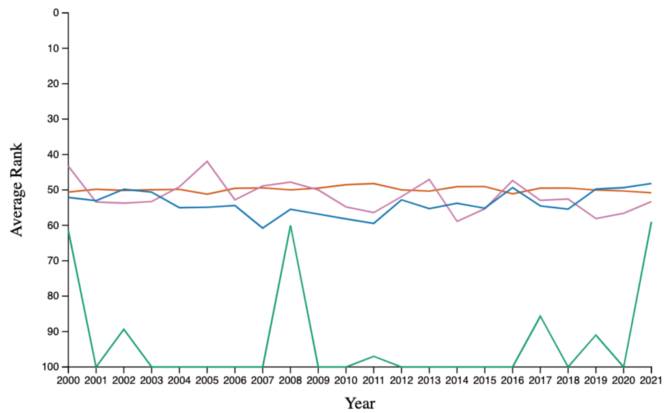
4. Visualization Design @ Project Milestone

- **Chart Selection**: For the first charts to implement, we chose a line chart of emotion count by year, a line chart of average rank per emotion by year, an animated area chart showing emotion per year, and a stacked bar chart of emotion proportion of songs per year.



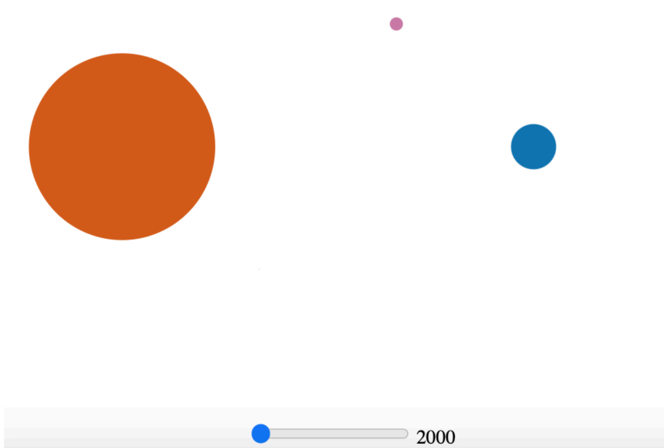
This chart shows the number of songs belonging to each emotion category aggregated by year. We thought that a line chart was an appropriate choice for this visualization because it shows the trends over time. Clearly, most of the Top 100 songs are ‘Happy’.

Line Chart - Average Rank Per Year

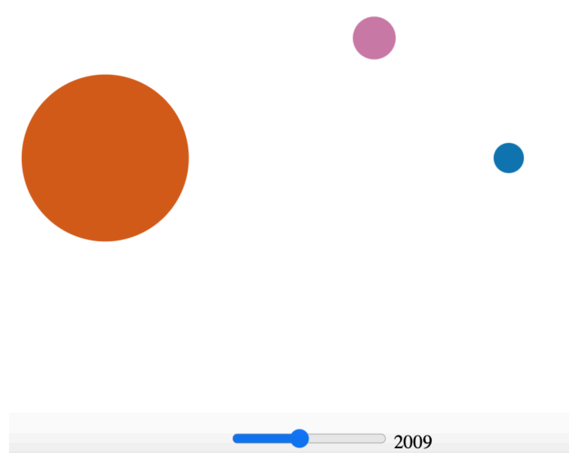


Our second visualization shows the average rank for each emotion label over time. This allows us to compare how popular the songs of each emotion were, not just how many were in the Top 100. We can see that ‘Happy’ songs have a stable average ranking around 50%, meaning there is a fairly even distribution of ‘Happy’ songs throughout the top songs. ‘Calm’, on the other hand, rarely makes the top 100, and when it does, it is usually a lower ranking song.

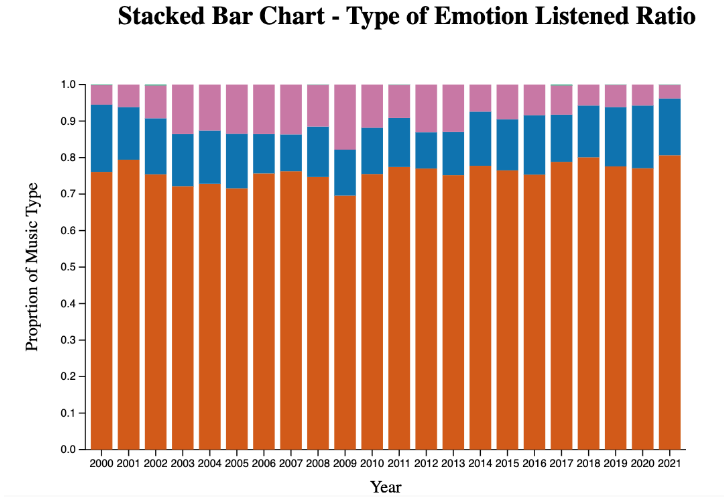
Circle Chart - Song Emotion



Circle Chart - Song Emotion



Our third chart is an interactive bubble graph showing the number of songs of each category by year. As you drag the slider along, the bubbles change in size to represent the prevalence of each label. Unfortunately, ‘Calm’ is so small that it appears nonexistent. We may consider doing log count or some other transformation in order to get a better visualization in the future.



Our final visualization is a stacked bar chart showing the proportion of each emotion in the top songs per year. It is a different way of visualizing the trends shown in the line graph. Again, 'Calm' is so small it barely appears.

- **Color & Style Choices**: We chose to use a colorblind friendly palette that corresponded to each emotion using our associations of color and mood. We chose warm colors for the positive, upbeat emotions, and cool colors for low energy, melancholic moods. Currently, our font size is a bit too small and we need to figure out a way to increase the size without making the axes too crowded.
- **Design Evolution:** We pivoted from using months on the x axis to years, since the number of ticks for each month would have been very hard to see. In the future, we may use the data by month as inflection points on the line chart, but just label the years so that the axis labels are not too crowded. We are still aiming to implement the designs from our mock-ups, and there will be some UI changes in the future.

5. Evaluation

So far our visualizations are showing that emotional listening trends do differ over time, but the relative ranking of each emotion stays relatively stable. 'Happy' is always on top, 'Sad' and 'Energetic' are fighting for second place, and 'Calm' is barely in the running. Some of the more interesting aspects of our visualization will come once we have labels on the charts showing when notable historical events happened. We have already noticed that the popularity of 'Happy' songs trend down after events like 9/11 and the 2008 recession. Something we could do to improve our visualizations would be to use transformations to un-skew our data, increase font size, and allow interactivity so that the user can highlight certain things to learn more.