LYRICALITYCS - PROCESS BOOK

OVERVIEW

DATA & RESEARCH QUESTION: For this project, we use three Kaggle datasets to analyze the relationship between musical success, song genre, and lyrical content: *Spotify Tracks Dataset* (containing 89,741 Spotify songs, including features such as tempo, danceability, energy, and associated genres, *960K Spotify Songs With Lyrics Data* (containing lyrics for 955,320 Spotify songs) and *Top Hits Spotify from 2000-2019* (listing 2,000 top-charting songs over two decades). We unified these three datasets into one and performed emotion analysis on English songs to explore the relationship between song lyrics' sentiments, a song's musical characteristics, genre, and its popularity.

DATA PROCESSING: To obtain a single unified dataset, we were careful when first choosing them to ensure they were all were based on Spotify data. The *Spotify Tracks Dataset* and *960K Spotify Songs With Lyrics Data* were matched directly using track IDs. *Top Hits Spotify from 2000-2019* did not include track IDs, so we matched songs by carefully cleaning and aligning artist names, song titles, and track features. After merging, our final dataset consisted of **80,274 songs with matching lyrics**, including **1,868 top hits**, and covering **114 unique genres**, without any missing or null values.

For emotional analysis, we focused on **songs with English lyrics**, labeled using Google's langdetect library. We classified song emotions using HuggingFace's SamLowe/roberta-base-go_emotions model, which predicts **28 fine-grained emotion categories** (such as admiration, amusement, anger, joy, sadness, among others). We excluded the "neutral" category from our main analysis to focus on more emotionally expressive content. To check the model's validity, we manually evaluated a sample of 100 songs and found an **81% human agreement rate**, confirming the model's generally reliable performance for our purposes.

VISUALIZATION CHOICES & IMPLEMENTATIONS: We implemented 5 total visualizations, all with some degree of interactivity. The visualizations are the following:

1. A 3D visualization of all the songs in the dataset, colored by emotion. A vector was created using all the song features (using embeddings for the lyrics), and LDA was used to reduce them to 3 dimensions. This presents an overview of the emotion space, showing that songs with different lyrical emotions sometimes have distinct musical characteristics.
2. A more grounded comparison of emotions' relation to song features via a radar plot showing the normalized average value for each song intrinsic feature, per emotion.

3. A word-cloud capturing the relationship between the language themes and trends in a songs' lyrics and the emotion it is associated with, proving an interesting exploration in its own right as well as serving as validation of our emotion classification.
4. A violin plot with an overlayed swarm-plot, used to compare the distribution of song characteristics for song which were top hits against all other songs.
5. A stacked bar-plot that captures emotion and top hit distribution across genres, seeking to understand how the former vary as a function of the song genre.

Details on each visualization, as well as their evolution from ideation to final result are presented in the following pages.

DATA STORY: At the start, we set out to explore what the impact of song lyrics' sentiments was. We have found that **there is a relationship between at least some song musical characteristics and the emotion they express in their lyrics**. This makes sense, as the artists likely try to make use of these features to better convey the emotion that underlies their song. However, emotions are very complex: for example, you can feel blissfully happy, and you can feel a bittersweet kind of happiness. As such, the same base emotion can be expressed in many different ways, with each song featuring different nuances and variants. On top of that, emotions are subjective! Thus, **it makes sense we were able to find a relationship between emotion and song characteristics, but that song emotions by themselves don't capture all the variance in song features - because sometimes different song features are used to convey different nuances of that emotion!**

Then, we turned to look at songs' popularity and song genres. We found that **song popularity goes beyond the data we analyzed here, likely more dependent on real world popularity than song's musical characteristics**. As for song genres? They allowed us to see how the entire ecosystem comes together: **how some genres are used to portray certain emotions; how other genres are more flexible with regards the song emotion, showing a fairly uniform distribution across them; and how genre is better able to predict a song's commercial success than its intrinsic characteristics or the emotions it portrays**.

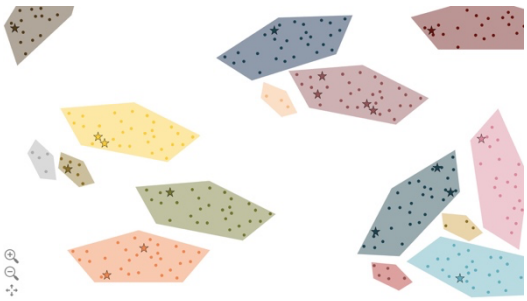
TEAM MEMBERS CONTRIBUTIONS

All members contributed equally to the project, and worked on most things together. Nevertheless, we outline what each member's main focuses were:

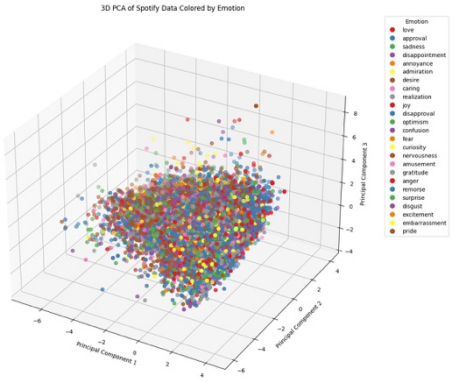
- **Beatriz** (Figure 1, Figure 2, Figure 4, website, all milestones)
- **Leonardo** (Figure 1, Figure 2, Figure 3, Figure 4, all milestones)
- **Kyuhee** (Figure 1, Figure 5, all milestones)

VISUALIZATIONS' EVOLUTION

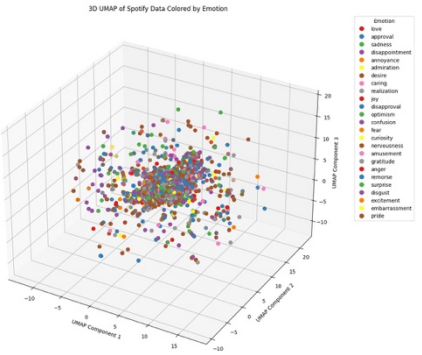
VISUALIZATION 1 – OVERVIEW USING SONG EMBEDDINGS



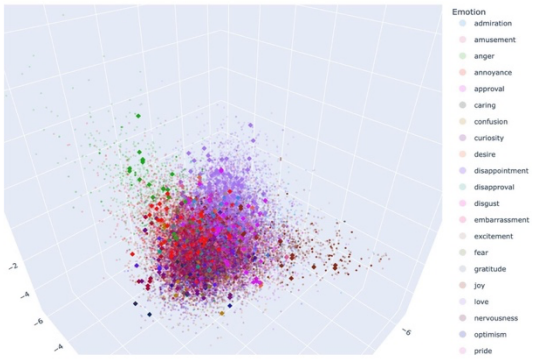
ORIGINAL IDEA & MOTIVATION: From the start, with this visualization, we wanted to provide an intuitive and engaging entry point into the several different dimensions we are exploring (song's intrinsic features, top hit quality, and emotion). By making these dimensions visible and navigable but not overwhelming, the user is primed to dive deeper into the data to obtain the full, detailed picture of the relationship between these features. To do this, we wanted to project these high-dimensional features onto a 2D or 3D space to observe the natural groups and clusters that emerge.



ITERATIONS: We began by applying PCA to several combinations of data (from just the song numeric features to adding genre categories and embeddings of lyrics) to capture the song's profile as best as possible. However, this projection did not reveal clear and distinct clusters when colored by song emotion. We experimented with different emotion groupings (e.g., only separating by positive, negative or neutral emotions) as well as alternative dimensionality reduction techniques like UMAP to better capture non-linear relationships. Still, the clusters remained overlapping or poorly defined.



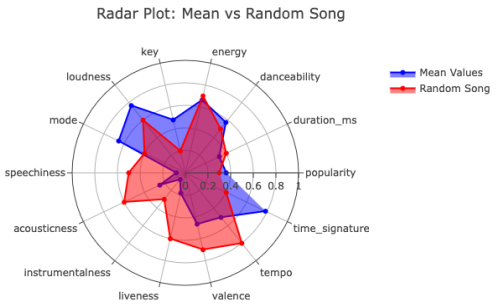
Finally, we employed LDA, which allowed us to preserve some of the emotional separation while still projecting the full song data onto lower dimensional spaces. We can observe some separation between emotions, but, at the center, there remains an overlapping region of emotions. This shows that while it's impossible to fully explain song features using song lyrics' emotions, lyrics' emotions do explain some of the variability we can observe in song features, motivating us to dive deeper to understand what those differences are. Finally, we add some visual separation between normal songs and top hits.



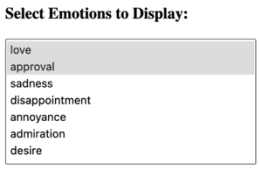
FINAL VERSION: Satisfied with our preliminary results, we implement a visualization for the website using D3.js. We add a slight rotating animation, distinct music icons for top hits, and several modes of user interactivity: from navigating the plot by dragging the mouse and zooming in and out, to hovering over points to obtain information about them, to being able to select which emotions to display at any given time (the image we present contains all emotions).

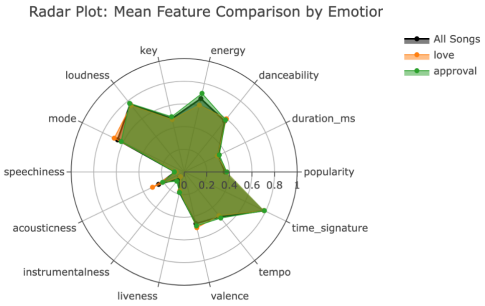
VISUALIZATION 2 - RADAR PLOT OF SONG FEATURES ACROSS EMOTIONS

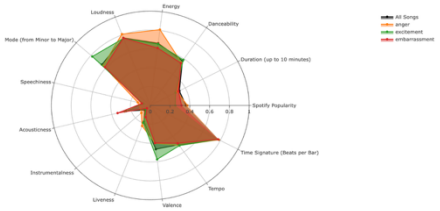
ORIGINAL IDEA & MOTIVATION: This visualization's goal was to explore the relationship between the emotional content of song lyrics and their intrinsic musical features (such as tempo, energy, valence, etc.). By mapping these emotional averages across song attributes in a radar plot, we aim to reveal patterns or correlations that might be less obvious in traditional charts.



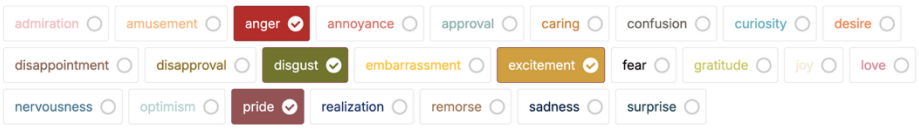
ITERATIONS: We encountered significant challenges due to the wide range of values across different song features. This variability made it difficult to compare features directly on the radar plot, as some axes dominated the scale while others were compressed. To address this, we normalized all the intrinsic song features to a common scale, ensuring that each feature contributed equally to the visualization and improved comparability. Another challenge was the overcrowding caused by displaying all emotions simultaneously. The overlapping lines and labels made the plot cluttered and hard to interpret. To solve this, we implemented interactivity that allows users to select which emotions to display on the radar plot. This user-driven filtering reduced visual noise and enabled more focused exploration of specific emotional profiles.

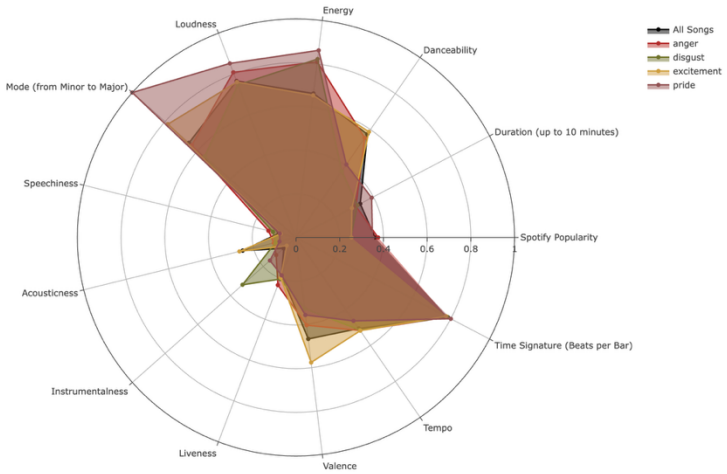






FINAL VERSION: Our final iteration is seamlessly integrated with the website, and the checkboxes are consistently color-coded by emotion (in an emotion-color assignment that is maintained across the entire website), making it easy to understand and interact with. Functionally, the final radar plot offers a clean and intuitive visualization of emotional profiles across key song features. The interactive tooltips and legend improve usability. This plot effectively showcases that certain emotions tend to be associated with particular ranges of energy, acousticness, valence, among others. This corroborates our findings from Visualization 1 that song emotions are correlated with at least some non-lyrical song features!

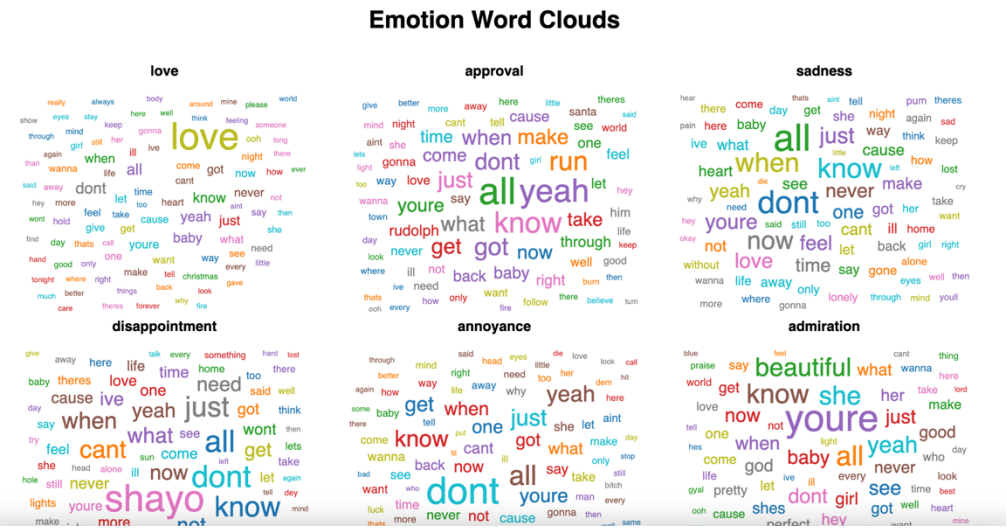




VISUALIZATION 3 - WORD CLOUD OF MOST COMMON LYRIC WORDS PER EMOTION

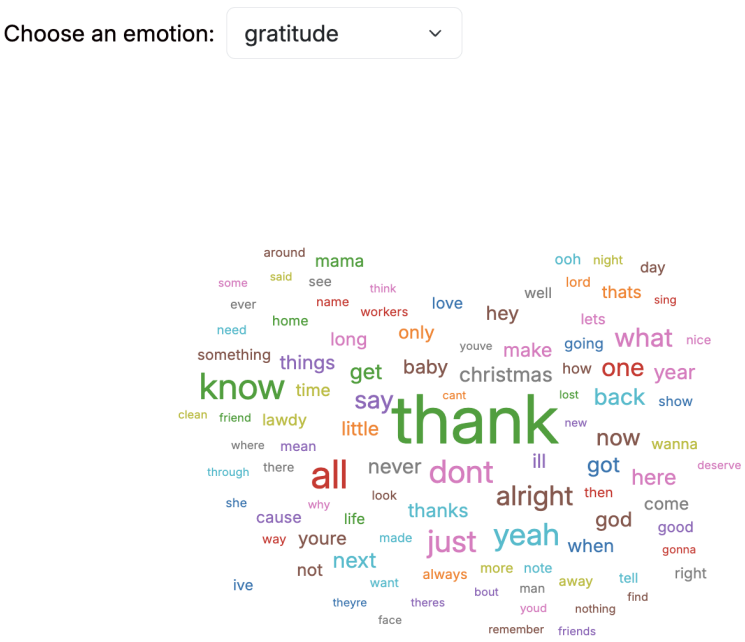
ORIGINAL IDEA & MOTIVATION: The idea behind this visualization is to highlight the most frequently used words in song lyrics associated with specific emotions. By generating word clouds for each emotion category, the goal is to provide an intuitive and visually appealing way to understand common lyrical themes and vocabulary tied to different emotional states. This approach helps in quickly grasping how language use varies across emotional contexts in music, and it also serves as a publicly accessible validation of our emotion labeling.

ITERATIONS: The first version of the word cloud was created using D3.js. It partly displayed all emotions simultaneously in a static grid, with words sized by frequency.



One major challenge was the long loading time when generating word clouds directly from the full CSV data, which contained a large volume of lyrics and words. This made the visualization too slow and inefficient to use. To optimize performance, we preprocessed the data by filtering out stop-words and retaining only the top 100 most frequently used words per emotion. We then cached this reduced dataset by saving as a JSON file, allowing the word cloud to load much faster and provide a smoother user experience.

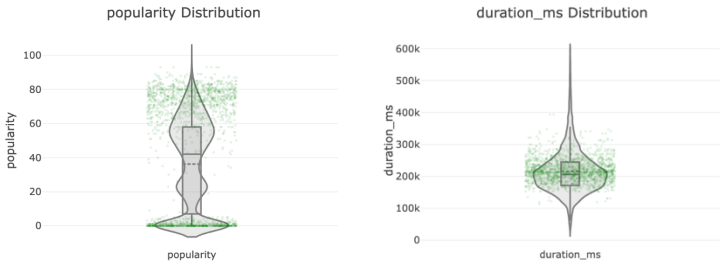
FINAL VERSION: The final word cloud visualization features an interactive design where users can select the emotion of interest, updating the word cloud accordingly. Words are sized proportionally to their frequency (after stop-word removal), providing a clear and engaging representation of common lyric choices per emotion. The clean layout and interactive controls enhance the user's ability to explore and compare lyrical themes linked to emotional content – and verifies that the inferred emotions song are relevant and reliable!



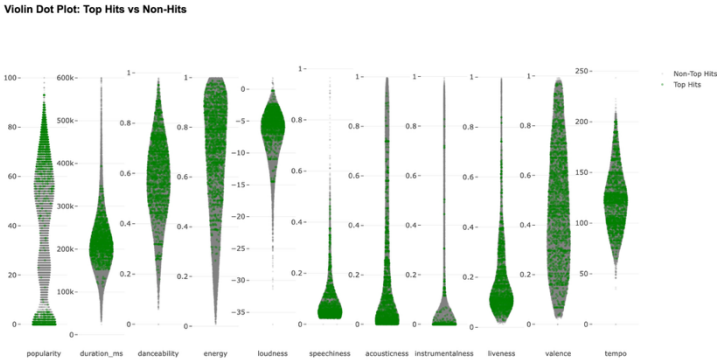
VISUALIZATION 4 - TOP HIT SONGS FEATURES' DISTRIBUTION VS. NON-TOP HITS

ORIGINAL IDEA & MOTIVATION: The goal of this visualization was to compare the distribution of intrinsic song features (such as tempo, energy, loudness, etc.) between top hit songs and non-top hits. By visually contrasting these groups, the aim was to identify distinguishing characteristics that may contribute to a song's commercial success. A swarm-plot was chosen for its ability to show individual data points while illustrating overall distribution, making it easier to detect patterns and outliers within each category.

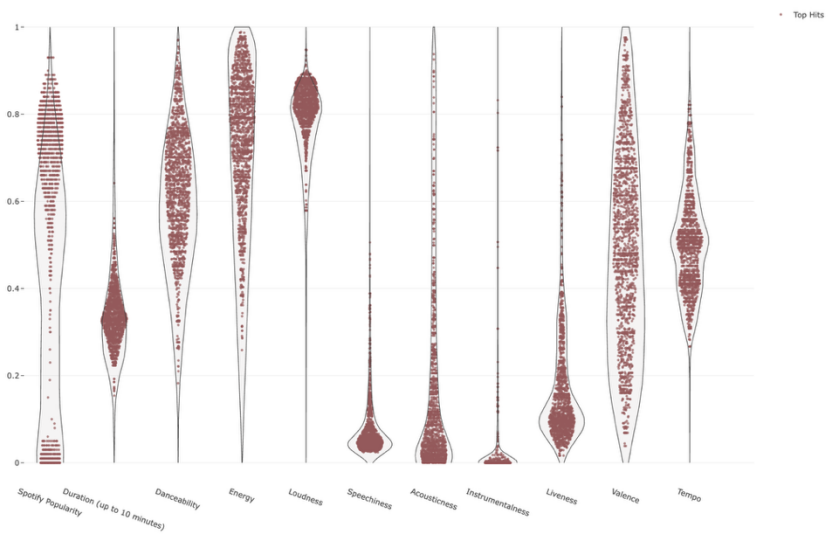
ITERATIONS: The first implementation used D3.js to create a basic swarm-plot comparing feature distributions for the two groups. While the plot effectively displayed data points, the visualization became cluttered with overlapping points, especially for features with many songs.



The primary challenge was managing overplotting and clutter in the swarm-plot, which obscured data density and trends. To address this, we experimented with adding jitter to reduce overlapping points and adjusting marker size and transparency for better visibility. Additionally, the scales of some features varied significantly, making direct comparison difficult. To improve interpretability, normalization or rescaling was implemented to align feature ranges across top hits and non-top hits, ensuring a clear and interpretable comparison.



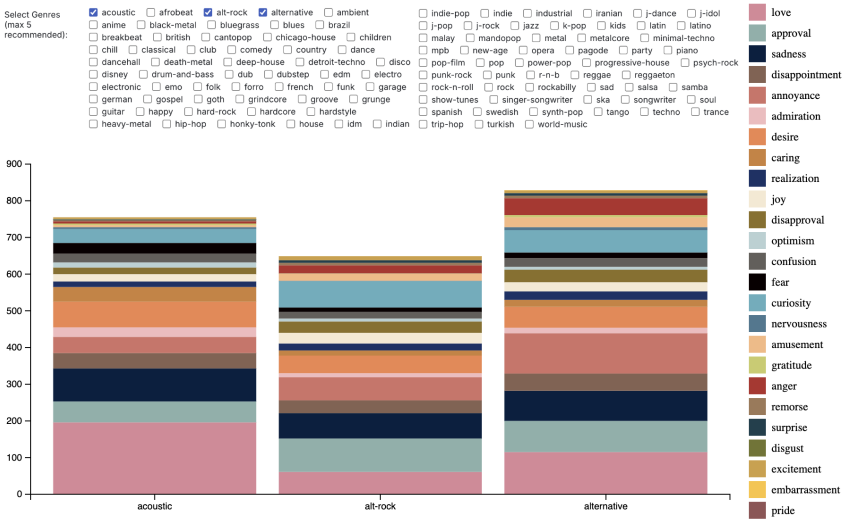
FINAL VERSION: The final visualization combines a refined swarm-plot with semi-transparent points and slight jitter, alongside summary statistics overlays. This balanced approach preserves individual song data while clearly illustrating differences in the features' distributions between top hits and non-top hits. The plot successfully captures the differences between the two groups: while top hits seem to be slightly higher on loudness, danceability and duration when compared to the average song, these differences are not significant, implying that what makes a song become very popular is probably tied to extrinsic song features (e.g., artist popularity, recording agency, marketing budget) rather than intrinsic song features.



VISUALIZATION 5 - RELATING SONG EMOTION, GENRE, AND POPULARITY

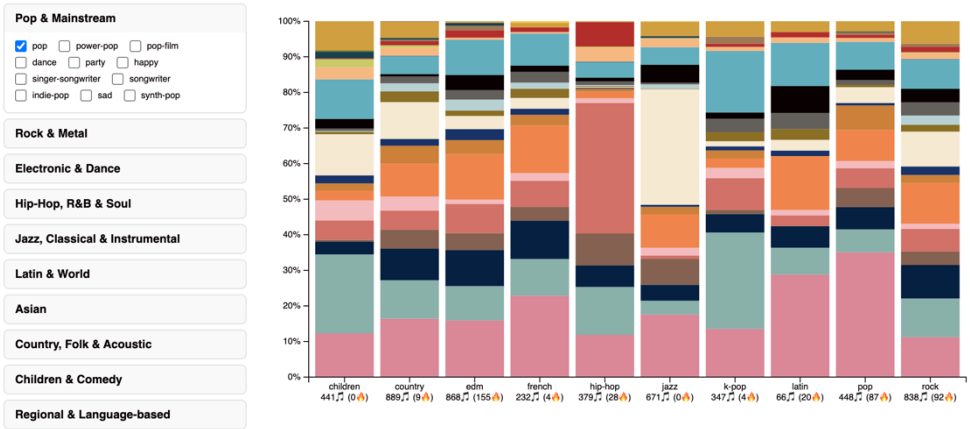
ORIGINAL IDEA & MOTIVATION: The goal of this visualization was to explore whether different music genres are characterized by different emotional expressions in their lyrics. Our hypothesis was that genres like "rock" or "hip-hop" might have distinct emotional patterns compared to others like "jazz" or "children's music." In parallel, we also wanted to investigate the distribution of commercial top hits across genres. Are some genres dominated by certain emotions or much more likely to contain top hits than other genres?

ITERATIONS: Our initial plan was to create a stacked bar chart, where each bar represents a selected genre, and segments within the bar reflect the distribution of lyrical emotions. This would provide a clear comparative view of emotional trends across genres. We also wanted to indicate the proportion of top hits per genre, to explore the link between emotion and popularity. The first version was a basic interactive D3 stacked bar chart where users could pick genres (limited to 5) via checkboxes. Bars represented raw emotion counts, with different colors indicating different emotions. However, this version suffered from poor scalability, minimal interactivity, and unclear legends.



As such, we took several measures to address these visualization challenges. First, the use of raw song counts across genres sometimes resulted in significantly disparate bar heights, which we resolved by switching to percentages, enabling easier more accurate comparisons across genres. Second, genre selection proved overwhelming because users had to navigate a very long flat list of options. To improve this, the interface was enhanced with collapsible grouped checkboxes, such as categories for “Rock & Metal,” making navigation more intuitive. Finally, the emotion legend in the chart was initially unclear and static. This was improved by introducing a hover interaction: users can now hover over emotion colors in the chart to reveal tooltips displaying the emotion name and corresponding count.

FINAL VERSION: The final version is an interactive D3.js stacked bar chart that allows users to select multiple genres, which are then visualized as a stacked bar showing the proportion of lyrical emotions. Hovering over any segment reveals a tooltip with the emotion name and both absolute and percentage song counts. The x-axis label enables an analysis of genre popularity and commercial viability, by adding the total song counts for each genre (🎵) as well as the number of top hits (🔥).



Website & Data Story

Integrating the Visualizations

Our visualizations were carefully designed and ordered to progressively uncover the relationship between emotion, musical features, and commercial success. Visualization 1 began with LDA to explore whether intrinsic song features could naturally separate emotional categories, but the observed overlap justified the need for more targeted, emotion-driven analysis. Visualization 2 responded to this by directly mapping emotions to normalized musical features using an interactive radar plot, allowing users to isolate and investigate specific emotional patterns. To complement this numeric perspective, Visualization 3 introduced word clouds that emphasized the lyrical content behind each emotion, grounding abstract features in real language use. With a foundation in emotion and musicality, Visualization 4 shifted toward commercial relevance, using swarm-plots to examine how feature distributions differ between top-hit and non-hit songs. This paved the way for Visualization 5, which integrated prior insights into a genre-based stacked bar chart, enabling comparative analysis of emotional expression across genres and linking it to popularity metrics. Each visualization built logically on the last, moving from general structure to specific relationships, and from broad exploration to focused insight.

Designing the Website

The website was designed to offer a smooth, linear user experience, with a visually appealing color palette that accommodates over 20 data categories. To avoid overwhelming casual users, detailed data processing explanations are placed on a separate "About" page, while the main webpage focuses on the core data story and research questions. The landing screen immediately orients users to the key questions explored through the visualizations. While the site is fully responsive, it is best viewed on tablets or larger screens due to the space requirements of certain visualizations. A major technical challenge was optimizing performance, as some visualizations are resource intensive. To address this, visualizations begin loading as soon as the page does, allowing them to render while the user reads the landing screen. Additional backend optimizations include preprocessing key data (such as computing emotion word frequencies and normalized values beforehand) so that only final results need to be loaded, ensuring a faster, more seamless experience on all visualizations.