Project Report For CS661: BIG DATA VISUAL ANALYTICS
2023-2024 Semester II

Project Title - AUDIO AURA

Team members: 1. Indraneel Rajeevan (231110403), 2. Akash Shivaji Varude (231110006), 3. Darshan Jain (231110009), 4. Nitish Kumar (231110033), 5. Pankaj Siddharth Nandeshwar (231110034), 6. Pranjal Maroti Nandeshwar (231110035), 7. Shaurya Agarwal (231110046), 8. Deepen Shrestha (231110061)

Member emails: {indraneel23@iitk.ac.in, akashshiv23@iitk.ac.in, darshanj23@iitk.ac.in, nitishk23@iitk.ac.in, spankaj23@iitk.ac.in, pranjalmn23@iitk.ac.in, shauryaa23@iitk.ac.in, deepens23@iitk.ac.in}

IIT Kanpur

1. Introduction

With *Audio Aura*, the convergence of sight and sound opens doors to untapped opportunities, empowering music label companies to expand their influence and solidify their position in diverse cultural landscapes. *Audio Aura* isn't just about listening to music; it's about seeing it come to life in a multitude of ways. Through genre and artist selection visualizations, users can delve into the vast world of music, exploring their favorite genres and discovering new artists effortlessly. The fusion of genres takes this exploration to the next level, offering a dynamic experience where boundaries blur and creativity thrives.

Album genre analysis provides invaluable insights into the evolving landscape of music, allowing music label companies to tailor their strategies to meet the ever-changing demands of their audience. Similarly, analyzing individual songs by genre sheds light on trends and patterns, enabling labels to anticipate shifts in consumer preferences and adapt accordingly. The analysis for explicitness ensures that content is curated responsibly, aligning with the values and expectations of diverse audiences worldwide.

Moreover, *Audio Aura* doesn't just stop at surface-level analysis; it delves deep into the nuances of music, providing insights and trend analysis based on songs' properties. Whether it's tempo, key, or mood, every element contributes to a richer understanding of the music landscape, empowering labels to make informed decisions that resonate with their target demographic.

Furthermore, the music scale analysis offers a unique perspective on the harmonic structure of songs, enhancing the appreciation of musical composition and fostering a deeper connection between artists and their audience. Lastly, with KNN music recommendation, users are guided on a personalized journey through the vast realm of music, ensuring that every listening experience is tailored to individual preferences and tastes. In essence, *Audio Aura* isn't just a tool for visualization; it's a catalyst for innovation and transformation in the music industry, revolutionizing the way we experience and interact with music.

2. Tasks

2.1 Genre and Artist selection

To analyze music data, we begin with the broadest of filters related to songs i.e. genre. The goal here is to extract insights into the top genres and artists within a specific region. To provide music producers with valuable knowledge about the preferences of audiences in different countries and identify successful artists within those genres.
Specifically, the tasks included:

1. Identifying the top genres in a given country based on streams/popularity.

2. Determining the top artists within each of the identified genres.

2.2 Genre Fusion

Exploring the evolving music landscape, we've observed a rising trend of genre fusion worldwide. This involves blending diverse musical styles without regard for geographical boundaries. By analyzing data on source and target genres, as well as metrics like popularity and average streams, we've uncovered intriguing combinations.

Examples include Europe's fusion of Latin rhythms with electronic beats, the incorporation of hip-hop elements into traditional Asian music, and the emergence of rap-trap fusion in North America. These innovative blends have garnered widespread appeal, highlighting the diminishing distinctions between genres and interconnecting cultures.

Our findings underscore how the modern world embraces diversity and musical innovation. Through data visualization, we explored how genre fusion continues to push boundaries and redefine the global music scene, fueled by diversity and creativity.

2.3 Albums Genres Analysis

To compare average ratings of albums across genres for identifying trends in audience preferences and genre popularity. This analysis helps music analysis companies understand which genres are more positively received by listeners, as well as those that may need improvement or targeted marketing efforts. So, Companies can focus their resources on genres that are more likely to attract a larger audience and drive engagement.

To provide summary statistics for album wise average ratings for quick overview of ratings distribution within genres.

Analysis of relationship between the average rating and the number of ratings for albums in each genre. It can be helpful to know the genres with lower average ratings but a high number of ratings or it can also reveal whether genres with higher average ratings also got larger number of ratings or not.

Investigating the number of reviews for albums across genres. A music production company may want to know whether there are any genres which are dominating the other genres significantly in terms of audience engagement. Hence it is essential to know comparison between reviews of top genres or selected genres

To know what type of songs are there in the album. What kind of content the songs have. Genre wise analysis of this can help to know the music analysis companies about the diversity of emotional content within genres

To know common themes and topics associated with albums in each genre. This can help music companies to identify the key themes and topics that resonate with audiences within each genre. Which can provide insights for content creation and marketing strategies

To be able to do all above mentioned tasks for one or more selected genres. Also in specified period(start date and end date).

2.4 Songs with Genre Analysis

Analyzing the top 100 songs annually provides insight into the most popular songs each year, reflecting trends and preferences in music consumption over the years. It indicates which artists have a larger share of the streaming market and offers insights into listener preferences for explicit content. Furthermore, it highlights trends in song length, revealing potential shifts in audience preferences.

The analysis of the all-time top 5000 songs by genre provides valuable insights into the music landscape. It helps identify trends in song production style, such as the use of acoustic instrumentation versus electronic elements and speech-like characteristics. This is particularly evident in the comparison of acousticness and speechiness among songs. Furthermore, exploring how different song features interact and contribute to a song's popularity within the genre offers a deeper understanding of listener preferences. For instance, the correlation between tempo, valence, and the popularity of songs provides insights into the emotional content and rhythm that resonate with listeners. Overall, this analysis sheds light on popular songs' characteristics and reveals the diverse musical tastes across different genres.

2.5 Analysis for Explicitness

In the dynamic landscape of the music industry, the incorporation of explicit content in songs has been a subject of both artistic expression and commercial strategy. However, the decision to include explicitness warrants a nuanced understanding of its alignment with societal trends and its reception across diverse regions and demographics. In this task, aim is to address key questions surrounding the viability of explicit content in music production using visualization. By understanding societal trends, regional variations, and audience preferences, decision-makers can make informed choices regarding content creation, distribution, and marketing strategies, ultimately optimizing both artistic expression and commercial success in an ever-evolving industry landscape.

2.6 Insights and Trend Analysis using songs' properties

Even some songs of best musicians, best vocalists and best genres fail to perform well in the market. One of the reasons is hidden in the mix of song's properties and future market trends for music. Visually understanding the interrelation of different song's properties is very important to generate a music that has higher chances of success in the market and hence can generate better business for music producers.

The objective is to design and develop "Mix & Master" an innovative visualization tool, poised to revolutionize the music industry's understanding of trends and patterns. Harnessing the power of interactive visualizations using Heatmap and Whiscus plots, Mix & Match aims to provide music producers with intuitive access to intricate insights derived from song properties and streaming data. It will uncover the insights that can play a crucial role in the success of future songs in a particular country.

2.7. Music Scale Analysis

In the ever-evolving music industry, one of the challenges is to create songs that captivate audiences and stand out from the competition. Music directors and producers must navigate a complex array of musical elements to produce tracks that not only appeal to listeners but also maximize commercial success. Understanding the relationship between musical scales and specific song properties can be a crucial factor in this process, offering insights into which scales are best suited for different musical contexts.

2.8 KNN Music Recommendation

As the domain of music analysis advances, the capability to quantify and predict music track attributes through computational approaches is becoming increasingly essential. Such methodologies not only facilitate a deeper understanding and classification of music but also improve the user experience on digital music platforms via personalized recommendations and the creation of automated playlists. Essential elements like danceability, energy, valence, and tempo are critical in the algorithmic evaluation of music tracks, offering a quantifiable framework to assess their dynamic and emotional aspects. Specifically, danceability indicates a track's suitability for dancing, energy reflects its vigor, valence denotes the conveyed emotional positivity, and tempo specifies the track's speed. For predictive analytics, machine learning techniques such as the K-Nearest Neighbors (KNN) are utilized. KNN classifies music tracks by comparing them to other tracks with known properties, using these key factors. Through such analysis, KNN effectively predicts a track's genre or mood, thereby forming the backbone of advanced music recommendation systems. This introduction to the key elements of music analysis and the implementation of KNN underscores the synergy between music theory and contemporary computational methods, underscoring their importance in today's digital music era.

2.9 Big Data handling

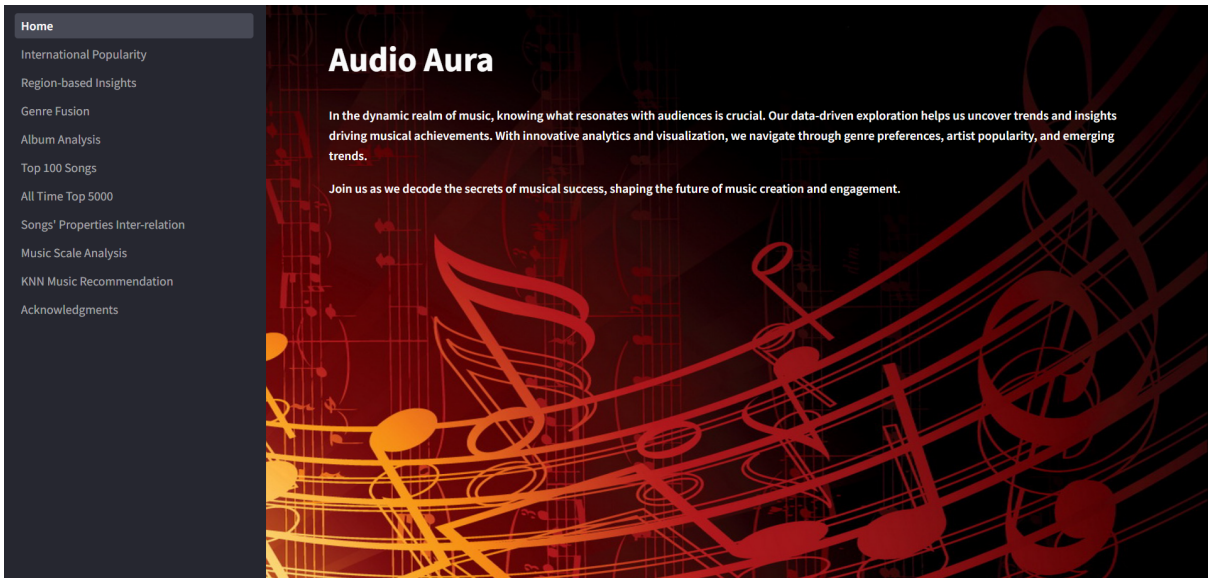
In addition to analyzing music data to extract insights, another critical task in this project was to address the challenge of handling large datasets efficiently. Given the substantial size of the datasets, loading them quickly and optimizing the memory usage is really necessary.

3. Proposed Solution

To bring our vision into action, we used a comprehensive tech stack that combines robust data processing capabilities with a user-friendly front-end interface. Python[1] served as the core language for data manipulation and analysis, with Pandas[2] enabling efficient data processing and Numpy[3] for numerical operations and NPZ compression. We structured our codebase in a package-oriented manner, ensuring modularity and reusability for future enhancements. This approach adhered to the team's requirements, as the majority decided to proceed with Python and Streamlit[4].

The user interface was built using Streamlit, chosen for its simplicity and flexibility in creating interactive data applications. We focused on making the interface responsive, allowing users to interact with visualizations smoothly on various devices. To enhance the user experience, we employed a consistent color scheme and designed the layout to be intuitive for individuals with some background in the music domain, along with general audience. This approach ensured that users could quickly grasp the relationships from our dataset without extensive technical training.

To support collaboration and version control, we utilized GitHub, which allowed the team to work concurrently and track changes efficiently. VS Code was largely helpful as a development environment, providing powerful debugging and code editing tools. For visualization testing and deployment, we used Chromium-based browsers, ensuring compatibility across platforms. By combining these technologies, we built our Visualization and Analysis system that is not only functional and efficient but also visually appealing and accessible to a broader audience.



3.1. Genre and Artist selection

3.1.1 Dataset

The dataset used here is **Spotify Weekly Top 200 Songs Streaming Data**[5] from kaggle. It contains the weekly chart data from December 2016.

3.1.2 Choice of visualization

To address the first task outlined above, we propose a **Sunburst Graph** for showing the best and the worst genres in a region.

3.1.3 Rationale

Our decision to utilize a sunburst graph for visualizing the hierarchical relationship between countries, genres, and artists was driven by several key factors:

1. **Hierarchical Data Structure:** The nature of our dataset lent itself well to a hierarchical representation, with countries serving as the parent nodes, genres as the first-level child nodes, and artists as the leaf nodes. The sunburst graph is particularly well-suited for visualizing hierarchical data, allowing for intuitive exploration of nested categories.

2. **Interactive Exploration:** The interactive nature of the sunburst graph enables users to dynamically explore the data at multiple levels of granularity. By interacting with the graph, music producers can drill down from the country level to specific genres and artists of interest, facilitating deeper insights into audience preferences and market trends.

3. **Compact Visualization:** Despite the hierarchical structure of the data, the sunburst graph provides a compact and visually appealing representation that effectively communicates complex relationships. This is especially beneficial for presenting large volumes of data in a comprehensible format without overwhelming the user.

3.2 Genre Fusion

Bubble plots visually represent data points as bubbles or circles, with each bubble denoting a data point. They effectively convey relationships between three variables: the source genre, the target genre, and the popularity, representing a third variable.

3.2.1 Dataset

For this particular visualization, we referred to the dataset named "MGD+: An Enhanced Music Genre Dataset with Success-based Networks", which had data related to trending songs, genres, etc. From this dataset, we selected the "genre network" data for visualization. We obtained this dataset from Zenodo[6].

3.2.2 Attributes

From the dataset, we used 5 attributes, namely, source genre, target genre, popularity, average streams, and weight. Using the 3 attributes— popularity, average streams, and weight, we calculated and normalized the popularity of each genre fusion that is possible in that particular country. Using these attributes, we plotted a bubble plot. Following are some features for choosing a bubble plot over other plots for visualization:

1. **Three-Variable Visualization:** Bubble plots excel in visualizing three variables simultaneously, enhancing data comprehension compared to traditional scatter plots.

2. **Size Encoding:** Bubble size encodes quantitative information about the third variable (popularity), allowing easy comparison of relative magnitudes or frequencies across data points. Incorporating bubble size provides richer data representation, particularly useful for datasets with numerous data points or wide-ranging variable values.

3. **Engaging Presentation:** Bubbles make the plot visually engaging, attracting attention to key areas or patterns within the data, and making it accessible to diverse audiences.

4. **Pattern Recognition:** Bubble plots facilitate quick pattern recognition, enabling viewers to identify trends, clusters, and relationships between variables through visual comparison of bubble sizes and positions.

5. **Customization:** Bubble plots offer customization options such as adjusting bubble sizes, colors, and labels, enhancing clarity and interpretation for effective communication of insights from the data.

3.2.3 Insights

We visualized the fusion of different genres that are trending in that particular country in a particular year using bubble plots so as to visually show complex information in simple and more interactive way. By hovering over the desired bubbles, we get to know the normalized popularity of two fusioned genres. Following are some of the insights that we get from the visualization:

1. **Emerging Trends:** Identify evolving genre fusion patterns in different countries over time to anticipate future music preferences.

2. **Market Potential:** Determine countries where genre fusion is growing in popularity, guiding investment decisions for market expansion.

3. **Collaborations:** Identify countries with thriving genre fusion scenes for potential collaborations with local artists and producers.

4. **Content Localization:** Tailor music releases and promotional materials to align with preferences of audiences in different countries.

By leveraging insights, the music label company can develop a focused strategy to maximize profits and establish its presence in diverse global markets.

3.3 Albums Genres Analysis

3.3.1. Dataset

Dataset Name: Top 5000 Albums of All Time - Spotify features[7]
Source: Kaggle
Description:

This dataset is of the top 5000 songs of all time, organized by genre. Each entry includes details such as the artist name, release date, genre(s), and album name. It also provides insights into the songs' reception, with average ratings, number of ratings, and number of reviews indicating their popularity.

For a deeper analysis, the dataset includes several audio features that describe the songs' characteristics. These features include acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, and valence. Acousticness measures the acoustic nature of the song, while danceability indicates how suitable the song is for dancing. Energy quantifies the intensity of the song, and instrumentalness represents the amount of instrumental content. Liveness captures the presence of a live audience, and loudness measures the overall volume. Speechiness indicates the presence of spoken words, and tempo specifies the speed of the song. Valence describes the positiveness conveyed by the song.

Additionally, the dataset includes the duration of each song in milliseconds and the time signature, which denotes the number of beats in each bar of music. This rich collection of features enables detailed analysis and comparison of songs across different genres, providing valuable insights into the evolution of music preferences and styles over time.

3.3.2. Solutions

1. **Average Rating Comparison Across Genres - Box Plot:** We have used box plot visualization for comparing average ratings across genres which provides detailed view of the distribution of ratings within each genre. Each box represents the interquartile range (IQR) of ratings, with the median indicated by the line inside the box. The whiskers extend to the minimum and maximum ratings, excluding outliers, which are plotted as individual points beyond the whiskers. This visualization allows for a quick comparison of the central tendency and spread of ratings among different genres. By identifying genres with higher median ratings or narrower IQRs, music analysis companies can pinpoint genres that are consistently well-received by listeners.

2. **Summary Statistics for Average Ratings - Table:** This table provides a quick overview of the ratings distribution within genres, including statistical measures such as mean, median, minimum, maximum, and quartiles. This table allows music companies to understand the distribution of ratings within each genre.

3. **Relationship Between Average Rating and Number of Ratings - Scatter Plot:** We have used scatter plot visualization to illustrate the relationship between the average rating and the number of ratings for albums in each genre. Each point on the scatter plot represents an album, with the x-axis indicating the number of

ratings and the y-axis indicating the average rating. By examining the distribution of points, music analysis companies can identify genres with lower average ratings but a high number of ratings, as well as genres with higher average ratings and larger numbers of ratings. This analysis can help companies understand the dynamics of audience engagement and genre popularity.

4. **Number of Reviews by Genre - Pie Chart:** The pie chart visualization displays the number of reviews for albums across genres, highlighting genres that are dominating others significantly in terms of audience engagement. This visualization provides a clear comparison of audience engagement between genres. Companies can identify genres that are attracting a larger audience.

5. **Genre and Sentiment Analysis - Stacked Bar Chart:** The stacked bar chart visualizations showcase the distribution of sentiment (positive, negative, neutral) in album descriptions within each genre. Each bar in the chart represents a genre, and the height of the bar is segmented into three parts, each representing the proportion of album descriptions with positive, negative, and neutral sentiment, respectively. By stacking the sentiment categories on top of each other within each genre's bar, we can easily see the overall sentiment composition for each genre and compare them visually. This format is useful for analysis because it provides a clear and intuitive way to understand the relative proportions of each sentiment category within each genre. Also, the stacked bar chart helps us identify trends and patterns in sentiment across genres. For example, we can quickly see if certain genres tend to have more positive or negative sentiment compared to others. This information can be valuable for music analysis companies.

6. **Theme and Topic Analysis - Word Clouds:** The word cloud visualization was used to showcase the most common words or themes found in the descriptions of albums within the selected genres. This visualization is a creative and intuitive way to summarize and display textual data and it also provides quick overview of most prominent terms used in the album descriptions.

7. **Filter and Date Range Selection Functionality:** We have implemented a filter and date range selection functionality in our analysis tool, which allows users to refine the dataset based on selected genres and a specified date range. Users can choose one or more genres from a list of available genres, ranging from the top 3 to 20 genres. And they can also specify a start and end date to filter the dataset based on the release date of the albums.

   This functionality provides users with the flexibility to focus on specific genres or time periods of interest for their analysis. By narrowing down the dataset, users can analyze the average ratings, number of ratings, and other metrics for albums released within a particular genre and time frame. This capability enhances the overall usability of the analysis tool by enabling users to customize their analysis based on their specific requirements and interests.

3.4 Songs with Genre Analysis

3.4.1 Making plots for the Top 100 Songs Annually(2013-2023)

1. **Top 100 Songs by Streams in Selected Year:** Created a bar chart that shows the top 100 songs in the selected year, ranked by the number of streams. The x-axis represents the song names, and the y-axis represents the number of streams.

2. **Distribution of Streams per Artist:** Made a bar chart showing each artist's total number of streams in the selected year. The x-axis represents the artists, and the y-axis represents the total number of streams.

3. **Average Song Duration per Year:** The line chart shows the average duration of songs for each year from 2013 to 2023. The x-axis represents the year and the y-axis represents the average song duration.

4. **Song Duration vs Streams for Top 100 Songs in Selected Year:** This scatter plot shows the relationship between song duration and number of streams for the top 100 songs in the selected year. The x-axis represents the song duration and the y-axis represents the number of streams.

3.4.2 Making plots for the All-time Top-5000 Songs by Genre

1. **Popularity vs Danceability:** Created a scatter plot showing the relationship between the popularity and danceability of songs in the selected genre. The color of the points represents the energy of the songs.

2. **Tempo vs Valence:** Another scatter plot shows the relationship between the tempo and valence of songs in the selected genre. The size of the points represents the popularity of the songs, and the color of the points represents the liveness of the songs.

3. **Energy vs Loudness:** Made a bar chart showing the relationship between the energy and loudness of songs in the selected genre. The color of the bars represents the danceability of the songs.

4. **Acousticness vs Speechiness:** Formed a scatter plot showing the relationship between the acousticness and speechiness of songs in the selected genre. The color of the points represents the valence of the songs.

5. **Feature Comparison:** When a pair of features is chosen, the bar chart contrasts these selected attributes. The hue of the bars is indicative of the songs' popularity.

3.5 Analysis for Explicitness

As we have seen till now our team has analyzed several factors related to songs, genres, people's favourites, regions top songs, genres and artists and many more factors. With these analysis it would definitely be great to have a successful song production. And so among these factors there is one more factor that we thought to be considered that is if the song is explicit it will be encouraged or be seen as out of culture and will be complete flop song that is why we come to analyze such factor.

3.5.1 Explicitness Over the past 70 years

The approach that is used was to firstly detect whether the explicit songs are really trendy or not. So dataset over the top "global" songs for the time period from 1950s to 2020s was analyzed and observed carefully. The data consists of more than 28000 hit songs over the past 70 years. Since the dataset was pretty big, some pre-processing and cleaning was done to make dataset less redundant and more readable. The type of visualization used for this task was time lapse graph with animation to clearly show the direction of trend over the years.
Songs were sorted over the years and the attribute "absence" was extracted from the song properties. And the data was plotted against percentage of obscenity in the content of songs year wise.

3.5.2 Country wise insights for Explicit songs.

From the above visualization it was seen that this trend has gone up over the years. So now moving on we thought to analyze this factor of "explicit" over the regions and country. So the data was extracted for all the top hit songs for years from year 2017 to 2022 country-wise with more than 126k songs combined from this [6] dataset. The year "2017" was chosen as starting point because we saw a steep incline from 2017 in last graph.

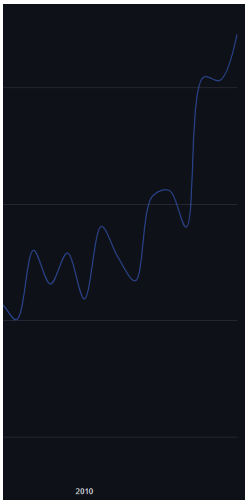


Figure 1: The straight incline is of 2017

Now the visualization chosen was for this is Deck chart (global map api of mapbox)

and then the data was selected country-wise sorted based on popularity but since the popularity was taken by the review factors it was not satisfactory attribute, So rather than that we chose number of total streams. It was more reliable factor to look after, so data was sorted by number of streams. Now the data from top popular songs, its attribute "explicitness" was extracted and plotted on chart country-wise.

The visualization over chart was Scatter-layer Plot, the coordinates for country was passed implicitly and data was mapped accordingly. A slider was also provided to select the top-k percentage of all hit songs. So the corresponding map will tell the percentage of explicit songs in top-k percent songs for that country. On hovering over the map we can see the country name and its value. The palette used was shades of green from lightest to darkest. This particular color palette was chosen because it was easily distinguishable for several countries with similar values from the overall background theme of our project and this gradient color give some sense of similarity between neighbouring points. This visualization tells about the culture about people region-wise and it was further observed that explicit songs is different region-wise which will be later explained in results section.

3.5.3 Trend in recent years

. Now that the geological aspects were seen we moved to check the current pattern of explicit songs in recent years country-wise. So a chart was picked to demonstrate this visualization such that it can show the global average and the trend for the picked country. Input is provided from same above dataset to select countries. The global average is depicted with red-dotted line and for all the other countries it will show subsequent patterns. The chart was points connected over the years and the value of y axis was percentage of explicit songs in all hit songs of that country. The chart was chosen as line chart because it can easily detect the patterns of inclination and declination.

With this graph we can easily deduce that if the trend is decreasing and is below global average there is no need to publish your explicit song in that country and waste money on marketing. Similarly we can seen that if country trend for explicit songs is increasing and is also above global average it will not matter that much.

3.6 Mix & Master: Songs' Properties Inter relation

Mix and Master provides the crucial and beautiful insights related to songs' properties that are hidden in the raw dataset and also provide analysis of number of streams using wiscus plot. These insights will help music producers to analyse the trends of popular songs or non-popular songs based on song properties for the selected country and in the selected year.
It will also give the idea about the future of music industry in a particular country by analysing the average number of streams over the years using the whisker's box plot.
We can filter the data based on popularity of songs to compare the variance of properties of popular songs and nun-popular songs. Stream filter will help to filter based on number of streams, it will give insights of the number of streams of popular songs and that of nom-popular songs.
Whiscus plot will give insights about the comparison of average number of streams over the years and we can also compare average number of streams for different country.

3.6.1 Heatmap between Songs' Properties

Heatmap provides insights between relationship of diffeent songs' properties. We can also filter data based on popularity scores and also based on number of streams, that allow us to analyse the trends in popular songs vs non popular songs. This analysis will help music producers by giving them idea about how well they must mix the properties in order to make the songs popular. It will also tell them for what not to try based on the insights of non popular songs.
The Heatmap gives answers to questions like::

1. Should we add/remove loudness in acoustic songs?
2. Should we increase/decrease valence in songs with danceability?
3. What duration of acoustic songs or songs with danceability is preferred by audience?
4. Should we consider instrumentalness in loud songs or not?

3.6.2 Whisker's Plot (Box plot) for number of Streams

This tells the music producers about the average, median and minimum numer of streams of songs in a particular country for a particular year.
The data can be filtered based of popularity as that we can get the idea of what is the stream statistics of popular songs and num popular songs respectively.
It also has a filter based on number of streams that will give idea of what is the persantage of songs that has views in that selected range and about its average and median number of streams.

The following questions can be answered easily based on the whisker's plot:

1. What is the overall average, median, maximum and minimum number of streams of songs in the selected country for the selected year?
2. Number of streams analysis of Popular songs.

3. Number of stream analysis of non popular songs.
4. Should we launch our music in that particular country based on number of stream analysis
5. Is the average number of streams increasing year by year?

3.6.3 Dataset Used

I have used a subset of MGD+ dataset [6].

This dataset is well structured into different sections, and into different regions(countries). Data for each year is in different .csv file making it easier to load the dataset in real time within no time, hence enhancing the user experience.

NOTE: I have got some meaning insights and analysis with the help of "Mix & Master" that I have discussed in details in the results section (**4.7**). Also there is a case study on Bolivia. Please do refer them to get the practical significance.

3.7. Music Scale Analysis

Our project aims to provide insights into the factors that contribute to a song's success in today's competitive music industry. As part of this broader effort, a specialized utility for Music Scale Analysis has been developed, designed to investigate how various musical scales relate to key song properties, such as valence, energy, speechiness, acousticness, instrumentalness, danceability, and liveness. By examining these relationships, our team seeks to offer practical guidance to music professionals, enabling them to create music that resonates with audiences and drives commercial success.

To achieve this, a large dataset [5] was processed, initially weighing 796 MB. The original data contained extensive information on various song attributes, but not all were relevant to this analysis. Through careful data reduction, unnecessary attributes were removed, retaining the most significant ones, and the dataset was compressed into a manageable 13.4 MB file. The final dataset consists of song properties stored as `float32` values ranging from 0 to 1, while the song key is represented as `int8`, indicating one of the 12 possible notes. This streamlined dataset serves as the foundation for Music Scale Analysis.

This utility offers several key functionalities, including a Heatmap to visualize the relationships between a chosen song property and all music scales, a Box Plot to compare a specific scale's distribution of a given property, and a Pie Chart to examine the relationship between a specific scale and multiple song properties. Additionally, the dashboard allows users to customize music notation (Sharp or Flat) and provides a reset button to start over. By focusing on these features, this utility provides a versatile and valuable tool for exploring the connections between music scales and song properties, offering actionable insights for music directors, composers, and producers.

3.7.1. Dataset

The dataset for this utility was derived from raw data sourced from Kaggle. This initial dataset contained a comprehensive set of attributes describing various aspects of songs, but it was unwieldy due to its size (796 MB in CSV format). To make the data manageable for analysis and visualization apropos of this task, a multi-step process to shrink and optimize the dataset for the task at hand has been carried out.

First, unnecessary attributes were identified and removed, focusing on the ones relevant to music scale analysis. The primary goal was to retain information that would help in exploring the relationships between musical scales and key song properties, such as valence, energy, speechiness, acousticness, instrumentalness, danceability, and liveness. Although this initial reduction cut down the dataset's size, it was still several hundred megabytes.

Next, data compression was targeted and the storage format was optimized. By converting the data from `CSV` to `NPZ` (a compressed format offered by `Numpy`), the dataset size was reduced to 13.4 MB, a significant improvement in terms of manageability and efficiency. This compression was achieved by converting the song properties to `float32`,

given their values ranged between 0 and 1, and representing the song key as `int8`, reflecting one of the 12 musical notes (C, C#, D, ...). The dataset was now lightweight enough for efficient processing and analysis while retaining the essential information.

The final dataset retained the key song properties and musical scale attributes required for this utility. The final optimized dataset became the foundation for this analysis, enabling us to visualize and derive insights into the relationships between musical scales and song properties, ultimately contributing to the broader Music Analysis project.

3.7.2. Attributes

The dataset for this utility comprises a set of attributes that describe essential song properties and the derived music scales. The scales were derived from two existing attributes: "key" and "mode". The "key" attribute represents one of 12 integers, indicating the musical note (C, C#, D, etc.), while the "mode" attribute is a `boolean` indicating whether the song is in a major scale (True) or minor scale (False). By deriving the scales from these two attributes, the dataset's size was further reduced, with "key" stored as `int8` and "mode" as `bool`, providing an efficient representation of musical scales.

Valence:
Represents the musical positiveness conveyed by a track, with higher values indicating more positive or happy music.

Energy:
Measures the intensity and activity in a song, with energetic tracks feeling fast, loud, and noisy.

Speechiness:
Detects the presence of spoken words in a track, with higher values indicating more speech-like content.

Acousticness:
Indicates how acoustic a song is, with 1.0 representing an entirely acoustic track.

Instrumentalness: Reflects the likelihood of a track containing no vocals, with values above 0.5 suggesting an instrumental focus.

Danceability: Describes a track's suitability for dancing, considering tempo, rhythm stability, and beat strength.

Liveness: Represents the probability of live audience presence during recording, with higher values indicating more live performances.

By deriving the music scales from existing attributes and optimizing the dataset's structure, a significant size reduction was achieved while preserving the essential information needed for analysis. This approach allowed the development of this utility to focus on key song properties and their relationships with musical scales, providing a solid foundation for this utility.

3.7.3. Use Cases

This dashboard has been designed to offer a variety of use cases that provide valuable insights into the relationships between musical scales and key song properties. By enabling music professionals to visualize these connections, the system contributes to informed decision-making in the music composition process, ultimately enhancing song quality and commercial success. Here is a detailed summary of the primary use cases provided by this dashboard:

3.7.3.1. Heatmap for Overall Trends

This use case allows users to visualize the relationships between a chosen song property and all music scales. The heatmap provides a comprehensive overview of how different scales influence specific attributes, such as valence, energy, or danceability. By using a color-coded matrix, users can quickly identify trends and patterns, guiding them toward the scales that best align with the desired song properties. This feature is particularly useful for understanding broader correlations and selecting scales with the desired characteristics.

3.7.3.2. Box Plot for Scale-Specific Analysis

The box plot use case helps users visualize the distribution of a chosen song property within a specific music scale. By plotting the data as a box plot, the utility shows the median, quartiles, and potential outliers, allowing users to assess the range and variability of a song property for each scale. This feature provides more detailed insights into how scales compare against each other for specific attributes, enabling music directors and composers to make data-driven decisions when choosing a scale for a song.

3.7.3.3. Pie Chart for Property Contribution

This use case allows users to visualize the relationship between a chosen music scale and any subset of the song properties. The pie chart offers a quick overview of how a specific scale contributes to various attributes, highlighting the proportionate influence of each property. By examining the pie chart, users can determine which song properties are most associated with a particular scale, providing a broader context for music composition decisions.

3.7.3.4. Customizable Music Notation

The dashboard also supports customization of music notation, allowing users to switch between sharp and flat notations. This flexibility caters to different user preferences and ensures the analysis is accessible to a wider audience. The ability to toggle between notations makes the utility user-friendly, especially for those with varying backgrounds in music theory.

3.7.3.5. Reset Button for Fresh Start

This use case provides a feature to reset the Music Scale Analysis system's state, enabling users to start fresh and clear any complex configurations. By allowing a reset, the utility

ensures that users can easily return to a baseline state, reducing the risk of confusion and facilitating experimentation. This feature is useful for both new and experienced users, providing a safety net during analysis.

Overall, these use cases demonstrate the versatility of this dashboard, offering a range of tools to explore and analyze the relationships between music scales and song properties. This comprehensive approach empowers music professionals to make informed choices in their creative processes, enhancing the quality and appeal of their compositions.

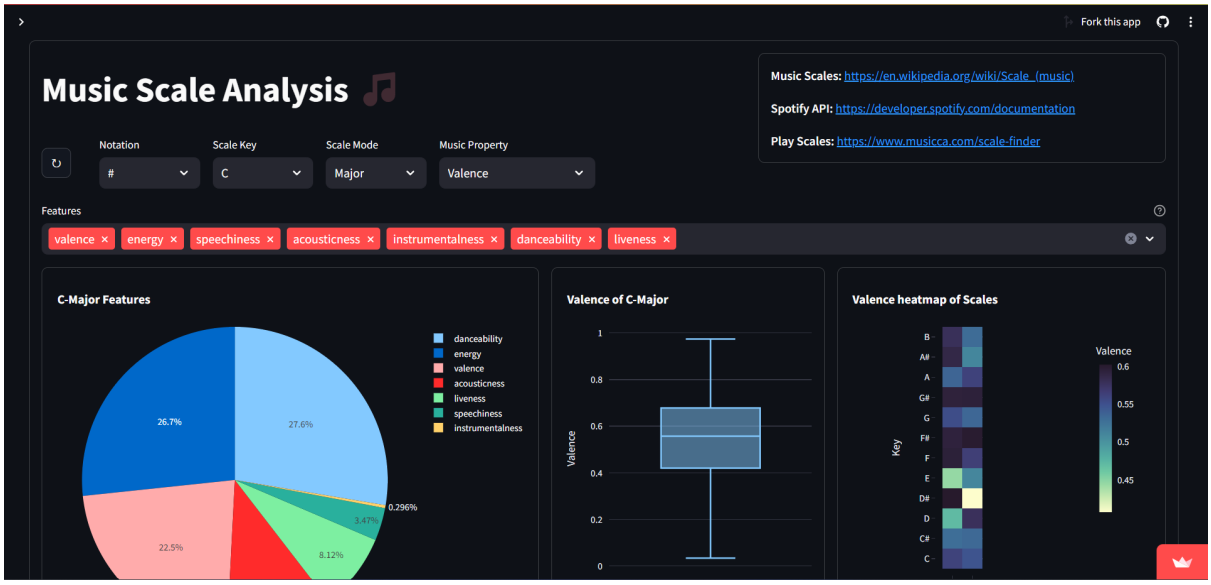


Figure 2: Dashboard for Music Scale Analysis

3.8 KNN based Music Recommendation

3.8.1 Music mainly depends on the below given 4 factor:

1. **Danceability:** Measures how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A higher score in danceability typically indicates that a song is easier to dance to. This attribute is often quantified using algorithms that assess the presence of a consistent beat and suitable tempo for dancing.

2. **Energy:** reflects the intensity and activity of a track. It is a measure that typically considers dynamic range, perceived loudness, timbre, onset rate, and general entropy. High-energy tracks feel fast, loud, and noisy (like hard rock or heavy metal), whereas low-energy tracks might be more acoustic, softer, and mellow (like soft jazz or classical music).

3. **Valence** measures the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g., happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g., sad, depressed, angry). This factor is often determined by the modes, keys, lyrics, and harmonic progressions within the music.

4. **Tempo** refers to the speed at which a piece of music is played, and it's measured in beats per minute (BPM). Tempo is a crucial element in determining the mood and genre of a track; for instance, a slower tempo might be typical of a blues or ballad, whereas a faster tempo might be indicative of genres like disco or techno.

3.8.2 K-Nearest Neighbors (KNN)

KNN is a simple, versatile, and widely used machine learning algorithm that can be used for both classification and regression tasks, including music prediction based on attributes like danceability, energy, valence, and tempo. KNN works by finding the 'k' nearest data points in feature space (based on a chosen distance metric like Euclidean distance) to the new point, and makes predictions based on the majority label of these neighbors for classification, or the average for regression. In the context of music analysis, KNN can predict the genre, mood, or other characteristics of a music track by comparing it to a dataset of labeled tracks. Using these factors and the KNN algorithm, you can build models that predict various attributes of music tracks, enhancing applications such as music recommendation systems or automatic playlist generation.

3.8.3 Using SQL query

And if user don't want to go with our recommendation system. They can perform an SQL query in our database and get the required song. It is implemented with the help of pandasql.

3.9 Big Data Handling

To tackle the challenges associated with handling big data in our project, we devised a comprehensive solution aimed at optimizing dataset loading speed and memory usage while maintaining data integrity and accuracy. The proposed solution encompassed the following strategies:

1. **Column selection and filtering:** Keeping relevant columns for analysis and discarding unnecessary data to reduce dataset size and improve loading speed. This involved conducting a thorough review of dataset requirements and selecting only the columns essential for the analysis tasks at hand.

2. **Data type conversion:** Converting data types to more memory-efficient alternatives, such as float32 or int32, where appropriate. By downsizing data types without compromising precision, we aimed to significantly reduce memory usage and enhance processing efficiency.

3. **Dataset splitting:** Datasets were split into subsets to be loaded individually only for the pages where they are required. The aim is again to minimize the data required for a page.

4. Results

4.1 Genre/Artist Selection

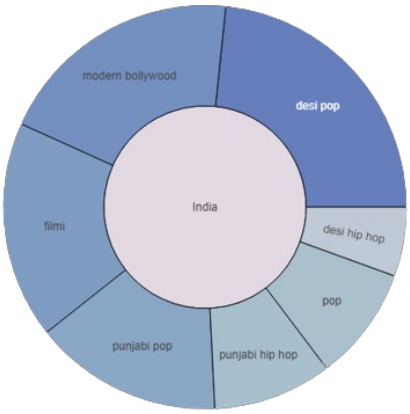


Figure 3: Top 7 genres in India.

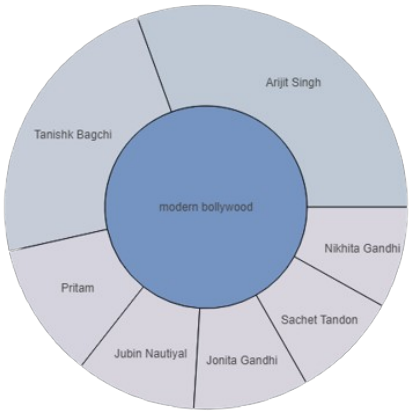


Figure 4: Top 7 artists in Modern Bollywood genre in India.

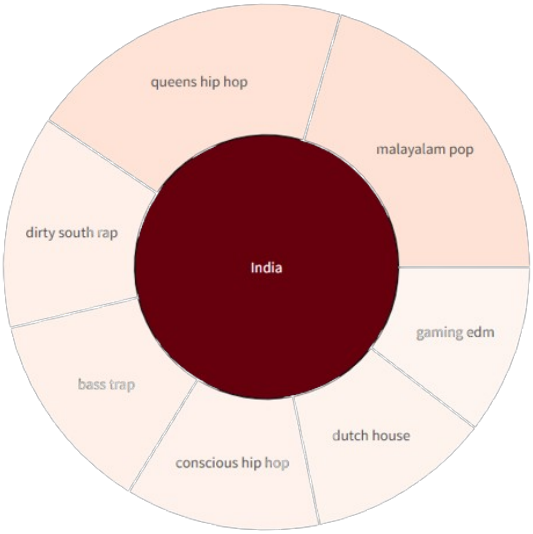


Figure 5: Bottom 7 genres in India.

4.2 Genre Fusion Analysis

Figure 6: Top 20 Most Popular Genre Fusion in Switzerland in 2022

4.3 Album Analysis

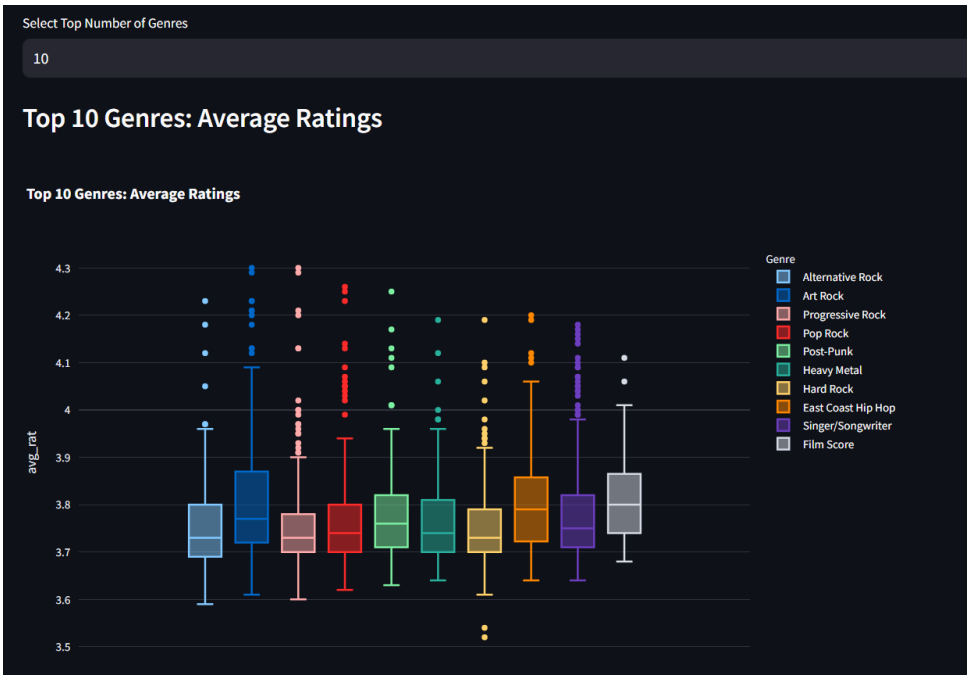


Figure 7: Top 10 Genres by Average Ratings

Figure shows the rating wise top 10 genres. Points above and below the box plots are outliers. Each of the point represents album. While color of point represents a genre the album belongs to

Summary Statistics

Genre	count	mean	std	min	25%	50%	75%	max
Alternative Rock	153	3.7575	0.1026	3.59	3.69	3.73	3.8	4.23
Art Rock	113	3.82	0.1558	3.61	3.72	3.77	3.87	4.3
East Coast Hip Hop	115	3.8098	0.1135	3.64	3.725	3.79	3.855	4.2
Film Score	112	3.81	0.0859	3.68	3.74	3.8	3.8625	4.11
Hard Rock	159	3.7572	0.0994	3.52	3.7	3.73	3.79	4.19
Heavy Metal	140	3.7619	0.0948	3.64	3.7	3.74	3.81	4.19
Pop Rock	142	3.7792	0.1261	3.62	3.7	3.74	3.8	4.26
Post-Punk	114	3.7864	0.1146	3.63	3.71	3.76	3.82	4.25
Progressive Rock	281	3.7578	0.0983	3.6	3.7	3.73	3.78	4.3
Singer/Songwriter	338	3.7817	0.1061	3.64	3.71	3.75	3.82	4.18

Figure 8: Summary stats of Top 10 Genres

By looking at the statistics we can see singer/songwriter genre performs exceptionally well in top 5000 albums. Followed by progressive rock genre. This summary can help music companies to take informed decisions by merely looking at these statistical values

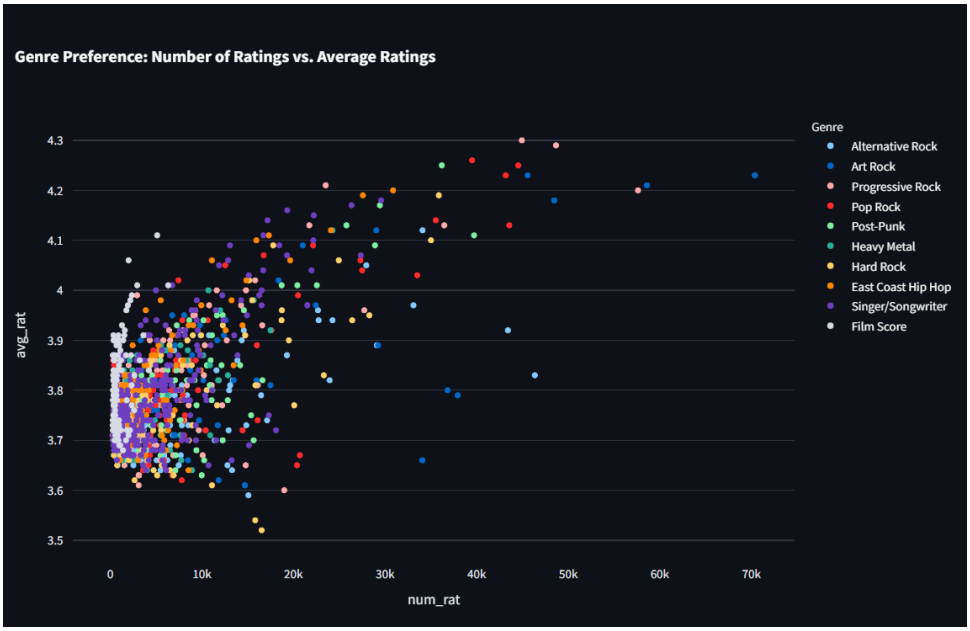


Figure 9: Average ratings vs number of ratings

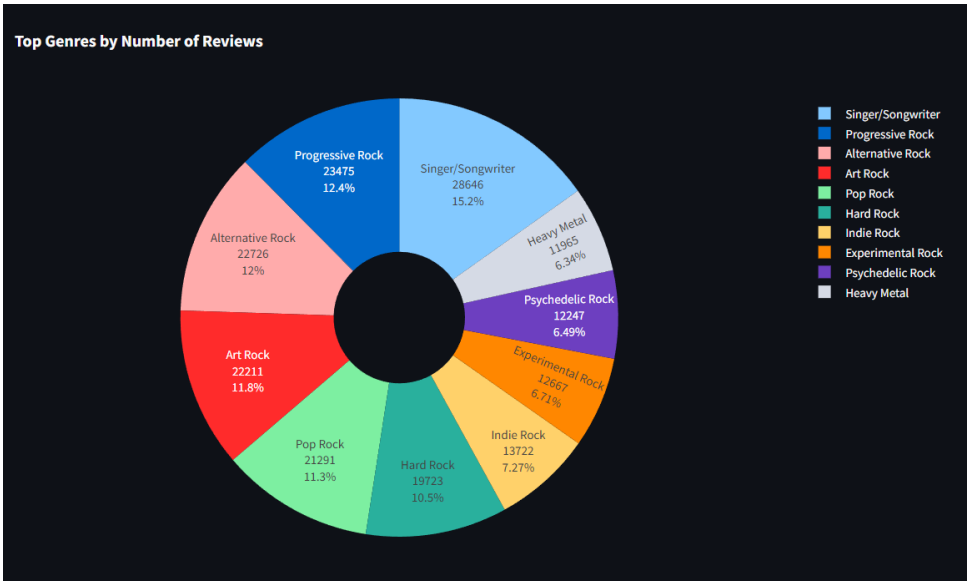


Figure 10: Number of reviews by genre

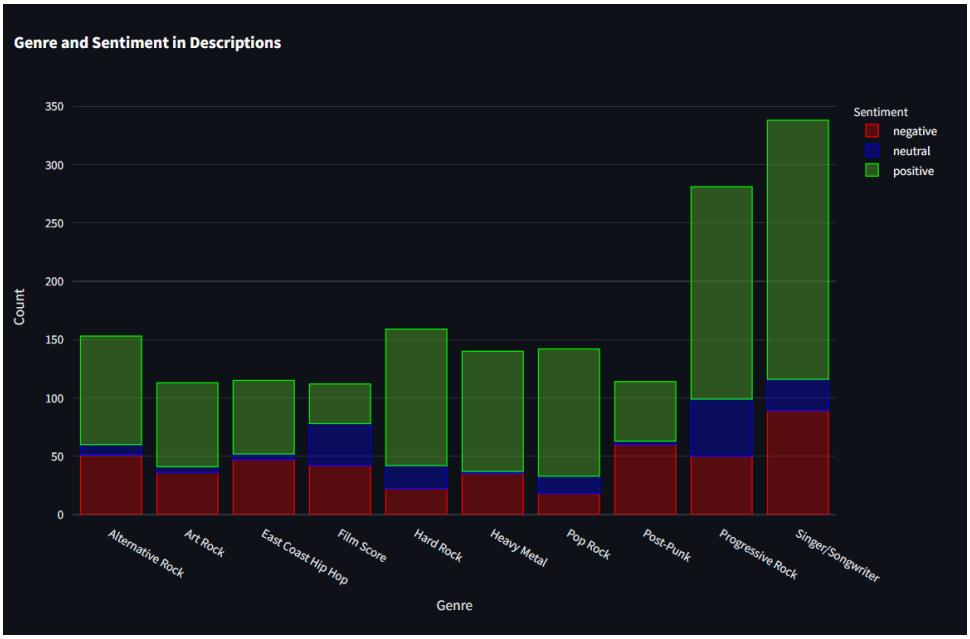


Figure 11: Genre wise sentiment analysis

Each album has an attribute description which roughly describes what types of songs it contains. We have done sentiment analysis on that description to get idea of sentiments. Red part in stacked bar chart represents negative sentiment, green means positive sentiment and blue means neutral. Companies can decide by selecting the genre and viewing at this chart what types of songs to produce. Also, to know the subjects of sentiments, below represented word cloud will help to get rough idea about the topics of songs particular album contains.

Figure 12: Word cloud of genres' descriptions

4.4 Top Songs Analysis

4.4.1 Plots for the Top 100 Songs Annually(2013-2023)

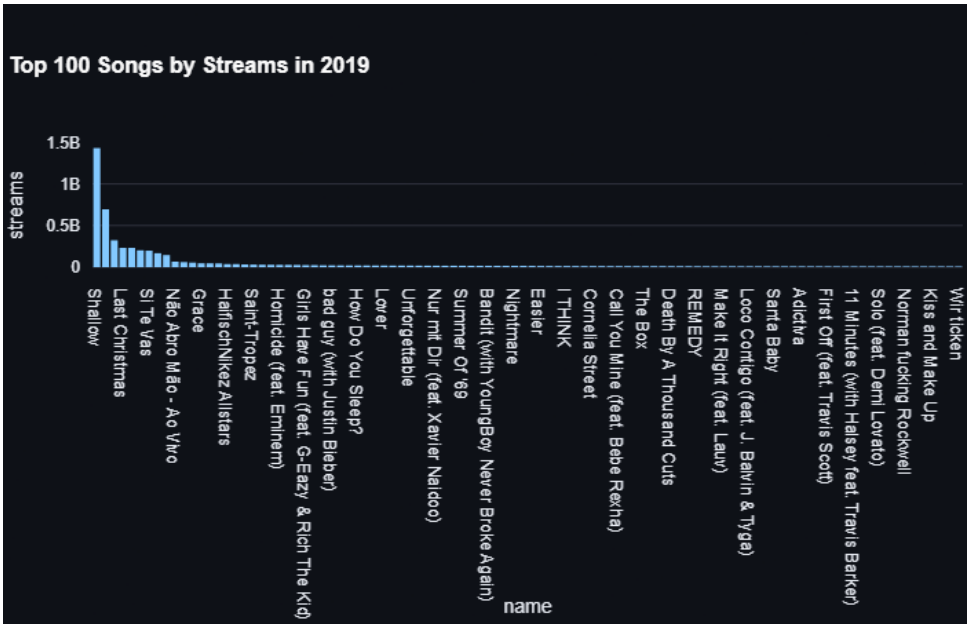


Figure 13: Top 100 Songs by Streams in Selected Year

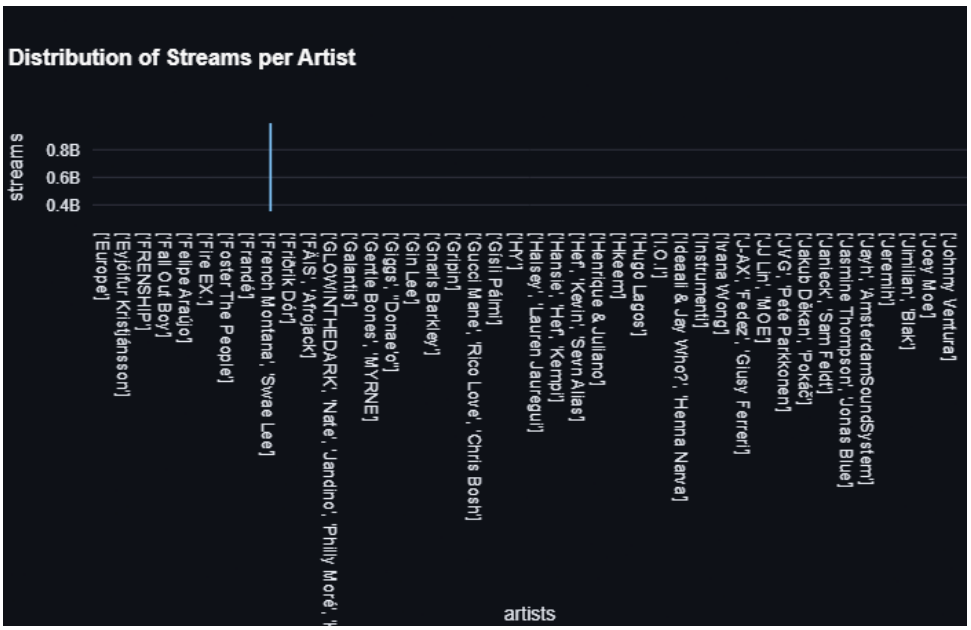


Figure 14: Distribution of Streams per Artist

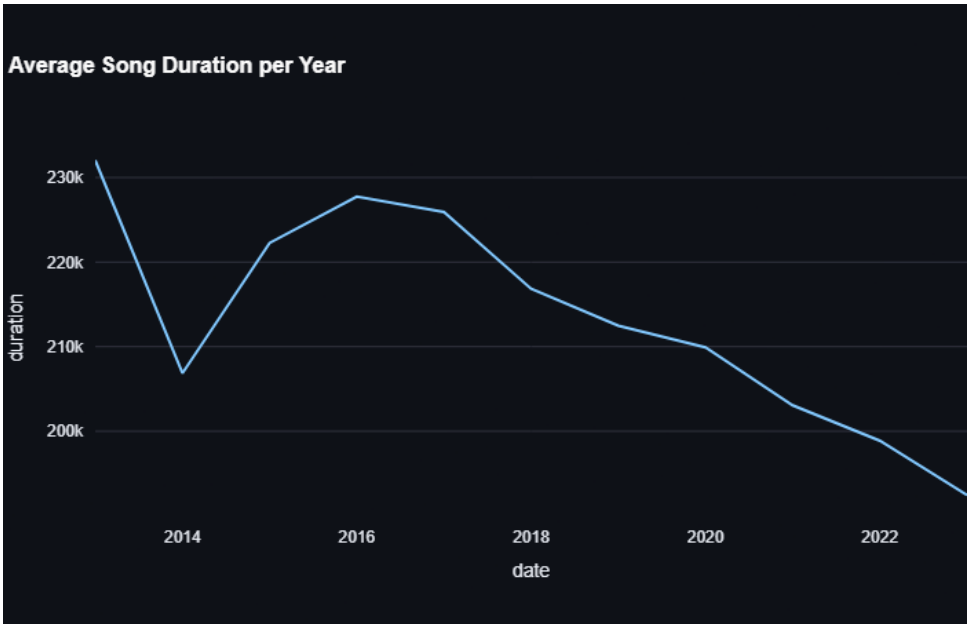


Figure 15: Average Song Duration per Year

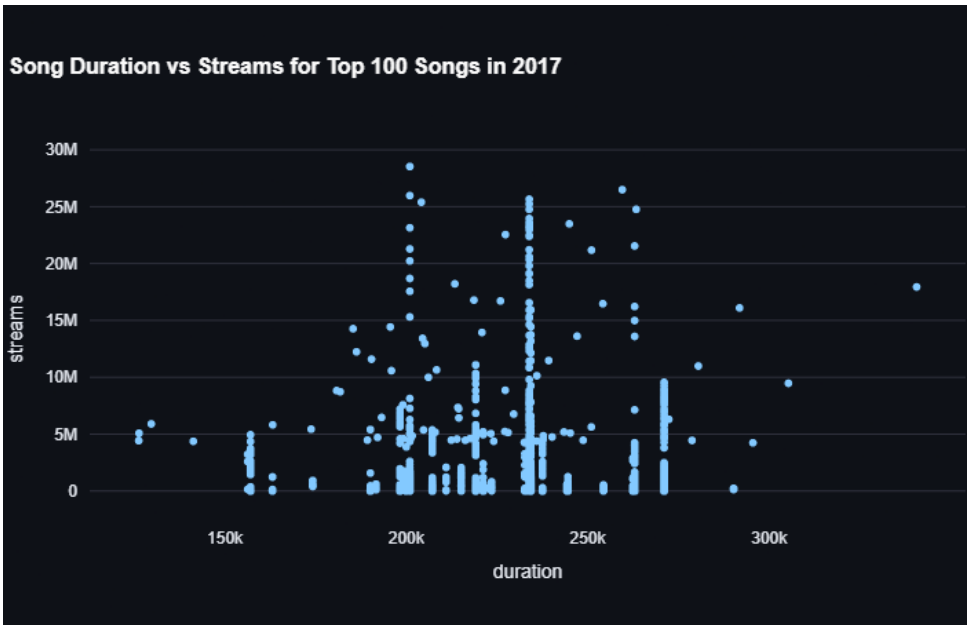


Figure 16: Song Duration vs Streams for Top 100 Songs in Selected Year

4.4.2 Plots for the All-time Top-5000 Songs by Genre

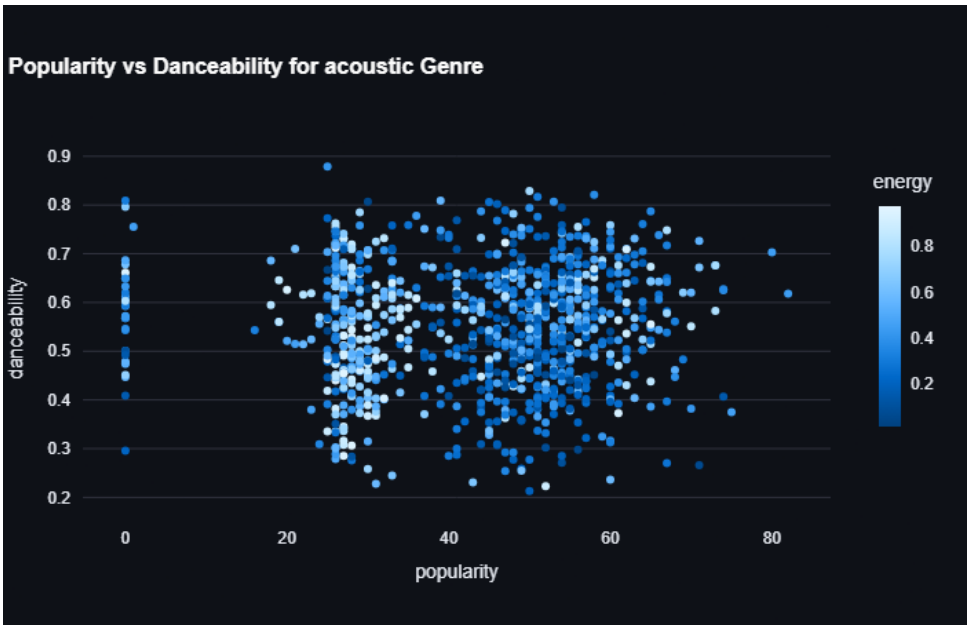


Figure 17: Popularity vs Danceability

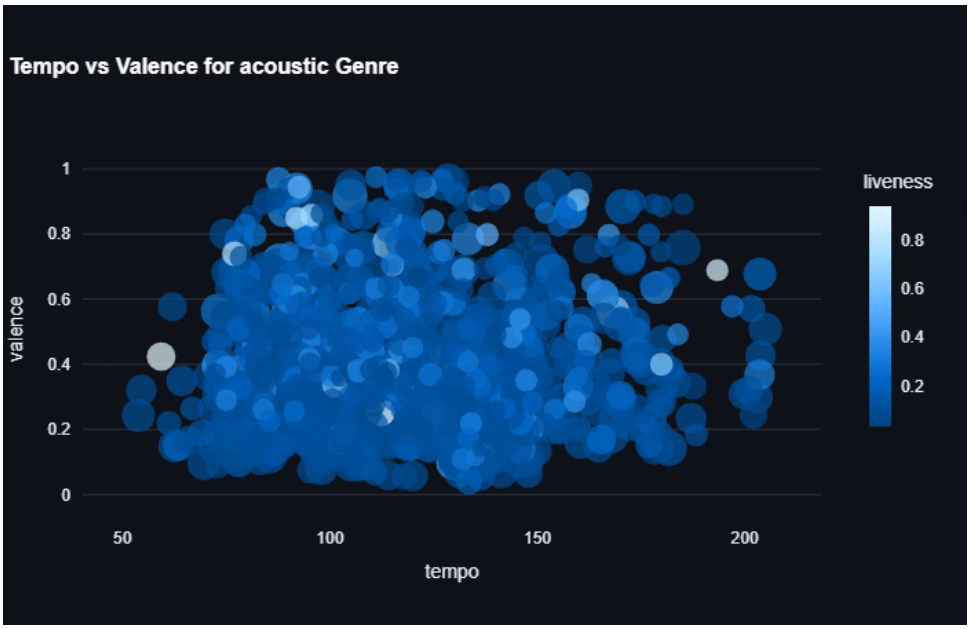


Figure 18: Tempo vs Valence

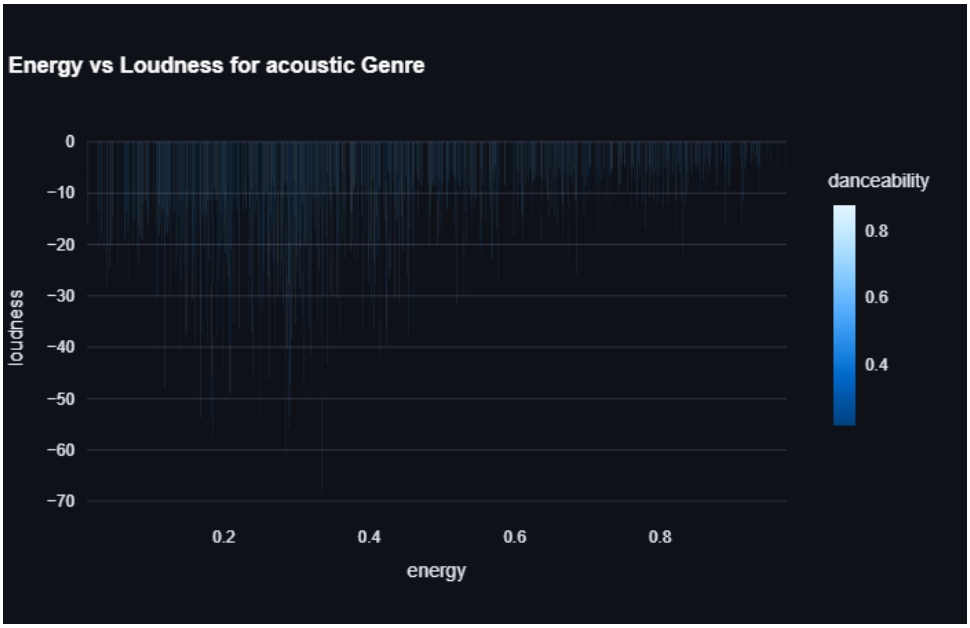


Figure 19: Energy vs Loudness

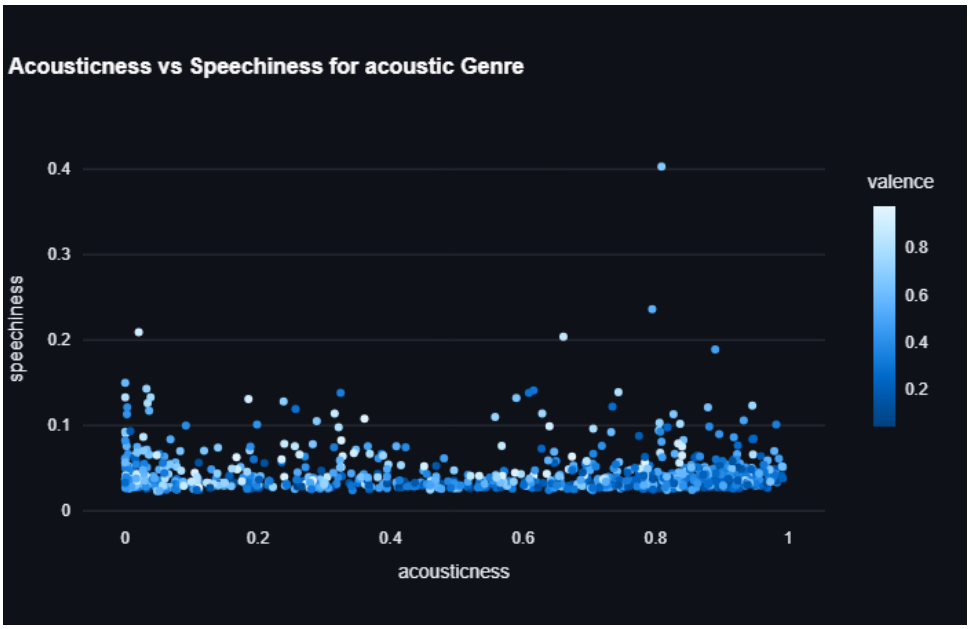


Figure 20: Acousticness vs Speechiness

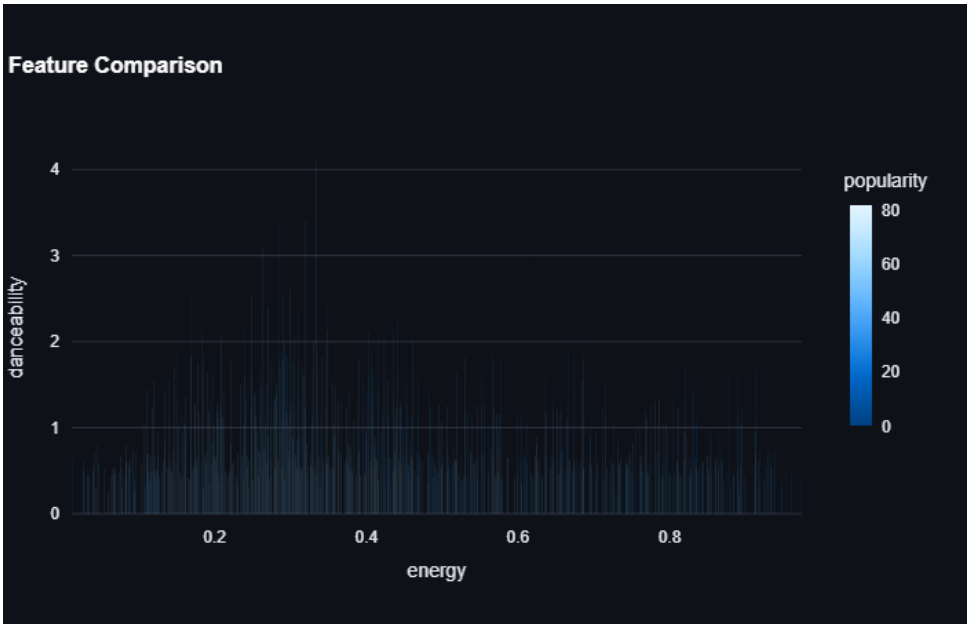


Figure 21: Feature Comparison

4.5 Resulting analysis of Explicitness

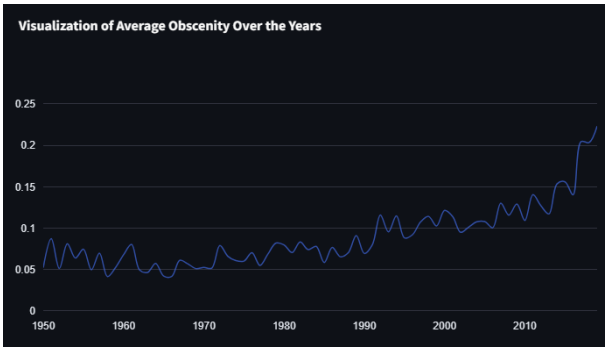


Figure 22: Visualization of Average Obscenity over the years

The above figure shows clearly of how the trend for explicit songs has increased over the last 70 years. And also it can be seen that the line takes steep incline growth after 2015.

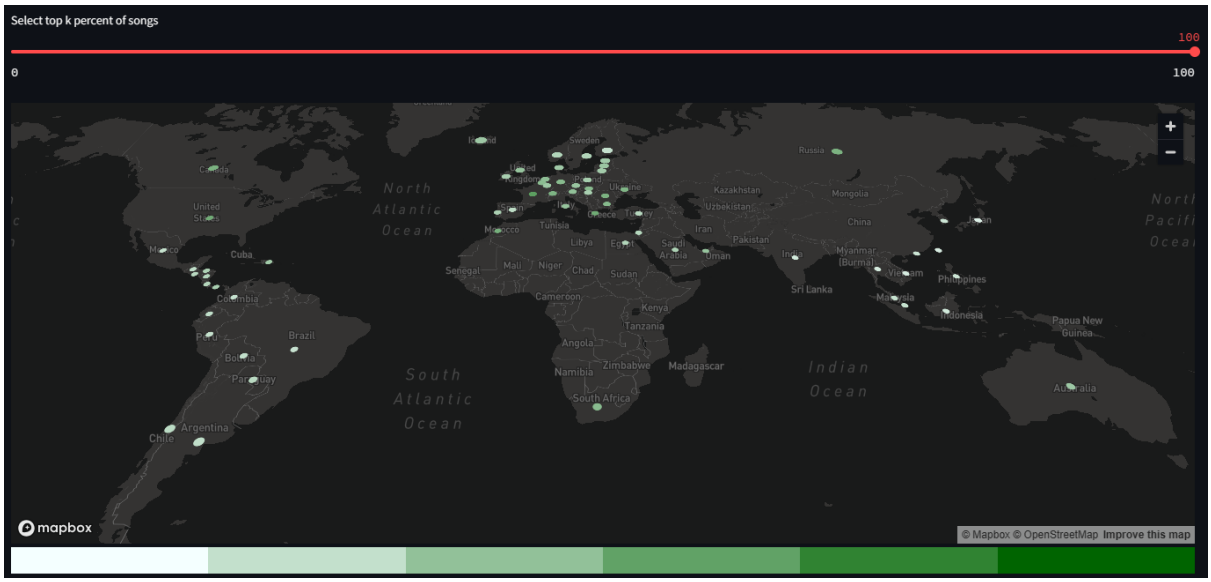


Figure 23: Map visualization of percentage of explicit songs within most popular songs

The above world map shows some good insights. Note that the global average for percentage of global songs is observed 42.2%. And if we see the data for all popular songs India and other southern Asian countries and southern American countries have very low numbers (8-14%) for such songs, we can say that songs tastes in these countries does not align much with such songs. And similarly we can see that other western countries like US, European countries have more than global average. These all countries have numbers in between 49-62%.

Finally in figure below we can see the red - dotted line as global average for percentage of explicit songs. It can be quite evident for some countries like France, USA they have constant trend in high numbers. And countries like India shown in dark blue colour is very far below from these countries. And also some middle eastern countries are shown great increase in their taste for explicit songs. The light pink line shows corresponding values for Peru which has doubled its values in last 3 years.

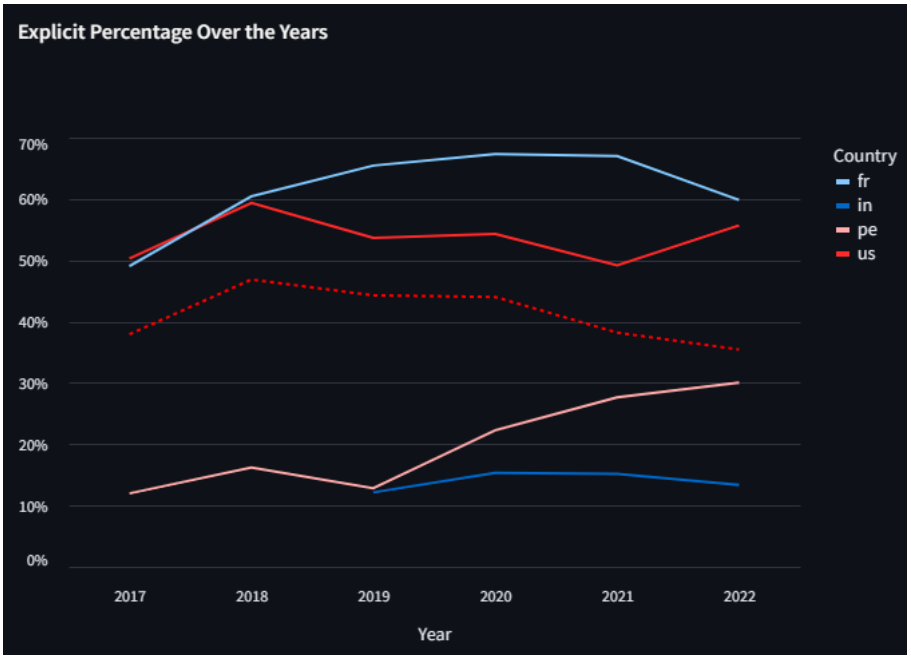


Figure 24: Explicit content over the years

4.6 Interface of Mix and Master

I have build a visually attractive, user friendly, scalable and fast UI for "Mix and Master". User can select country and year for which he/she wants to do the analysis.
Only the dataset for the selected country and year will be loaded at a time, making it super responsive.
For this analysis, Heatmap and Whiscus box plot are most suitable visualizations. User can also filter data based on popularity scores and number of streams.
Here are the snapshots:

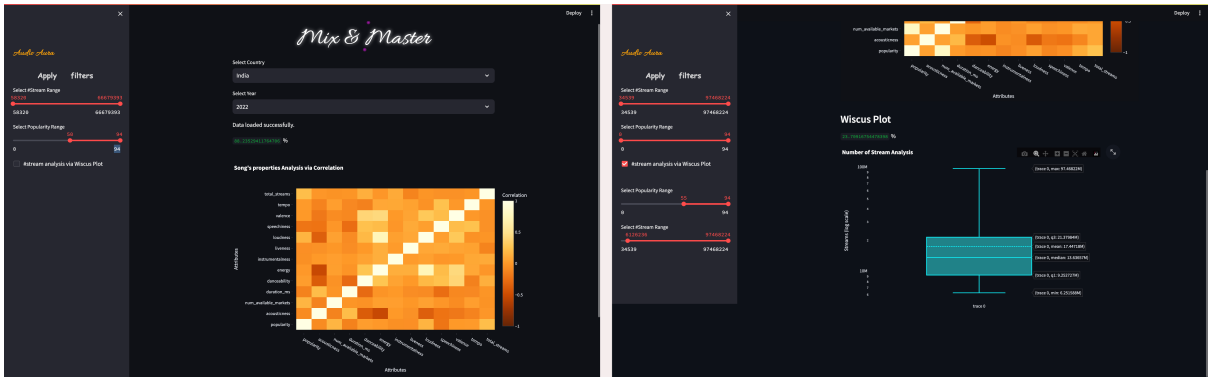


Figure 25: Interface of Mix & Master: Heatmap and Whiscus Plot

4.6.1 Generic Insights via Mix and Master

These are some beautiful analysis that I have done using MIX & MASTER:

- Acoustic songs has a highly negative correlation with loudness, making an acoustic song that is loud might fail badly in the market.
- Songs with danceability has positive correlation with valence and that makes sense as we want positive environment with songs with danceability.
- Track length of acoustic songs has a positive correlation, audience prefer longer duration acoustic songs.
- Track length of songs with danceability has a large negative correlation, that implies that audience prefer shorter duration songs with danceability.
- Songs with loudness has very high energy, are high in danceability and valence.
- If songs are popular, they get more number of streams.
- Over the years, the average number of streams is increasing, for example in India, showing growing demand of music in India and a positive sign of good market for music producers.

4.6.2 Analysis by Heatmap:

textbfSample analysis: In the following heatmap, we can see the following patterns:

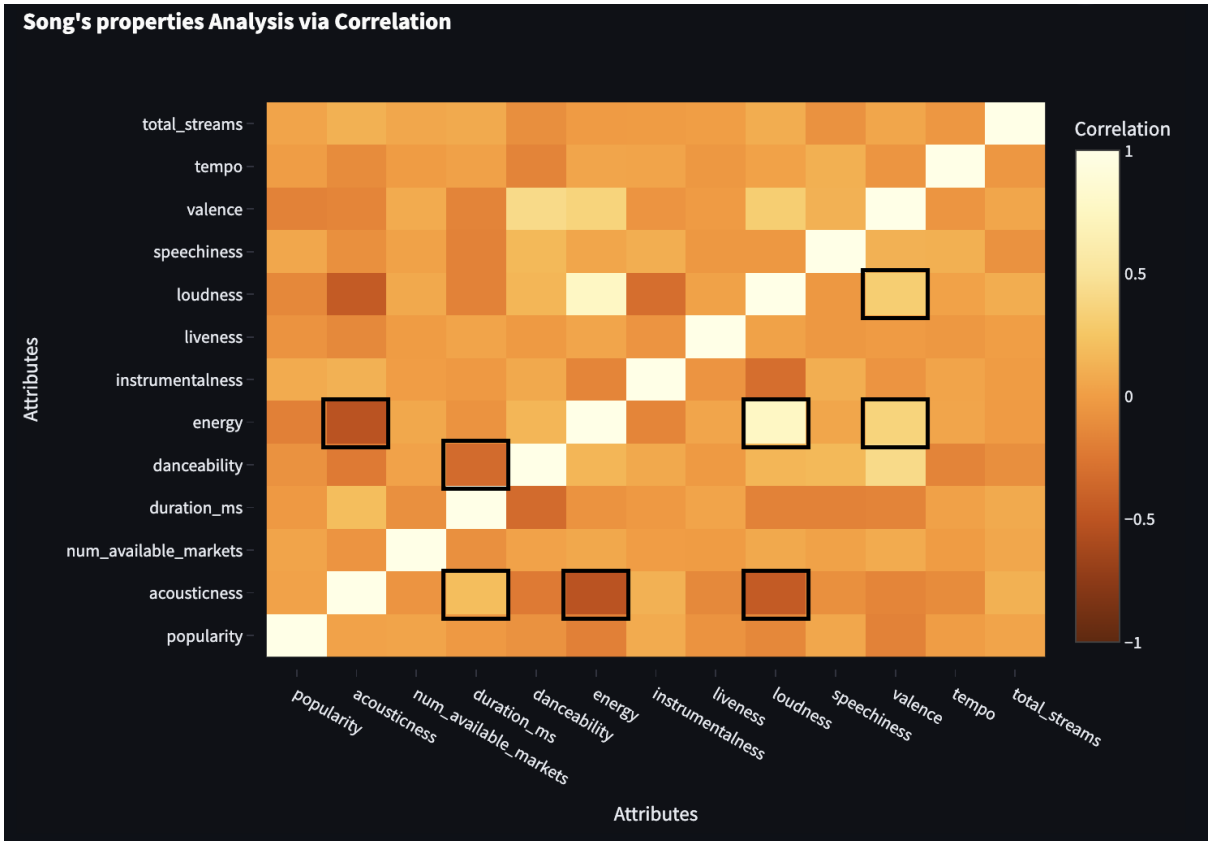


Figure 26: Heatmap of Popular songs of India in 2019

- Songs with energy are loud.
- Songs with danceability are shorter in duration.
- Songs with energy are not acoustic in nature.
- Songs with acousticness are longer in duration.
- Songs with danceability have valence in them.
- Songs with energy have valence in them

NOTE: These patters were observed in most of the countries.

In order in make a new song popular in a particular country, here India, we need to maintain this correlation, else our song may get flop in the market,leading to losses.

4.6.3 Analysis by Whisker's Plot:

In the following plots, we can get the following insights:

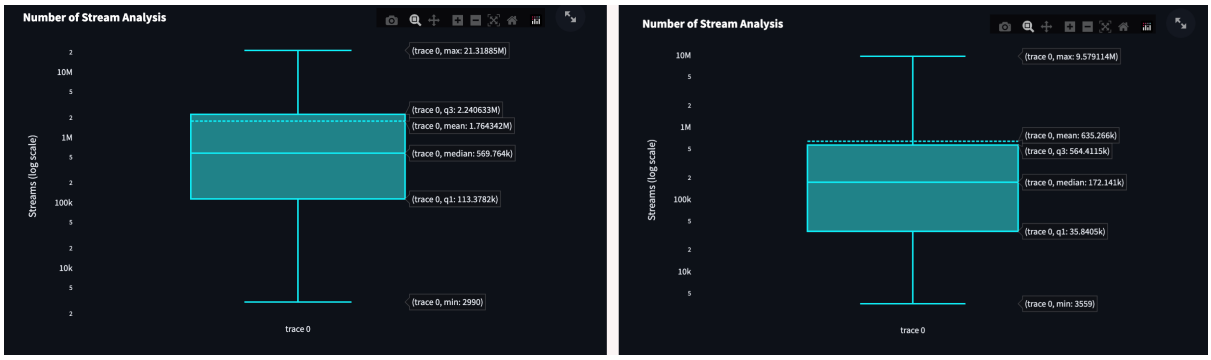


Figure 27: India's Popular songs 2019 vs India's Popular songs 2019

- Songs with popularity more than 60, have an average streams of 1.7 millions with a median streams of around 570,000.
- Songs which were not that much popular, popularity below 30, have a median streams of 172,000.

In this way we can get the insights like in which country even if the song does not become that popular, how many streams it can still get, like in India-2019, it was 172,000, which is not that bad. Hence we can conclude that India is a good market for Music Industry.

4.6.4 Stream analysis over the years:

This gives the idea of the future demand of music in a particular country.
Suppose, we want to figure out if the music industry is going up or down in the future.

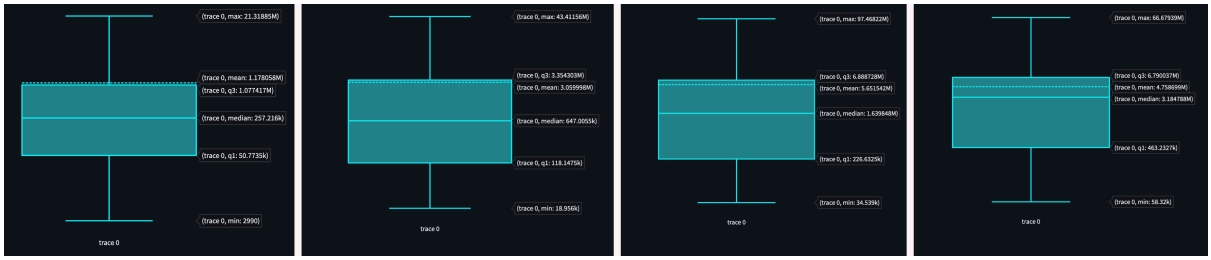


Figure 28: Stream analysis of India in 2019-2022

For India, we can clearly see the increasing pattern in median of number of streams of songs:
2019: 257,000
2020: 647,000
2021: 1,630,000
2022: 3,184,000

Clearly, India is a hot market for Music Industry as the number of streams is on an upsurge year by year.
If, for a country, this pattern does not follow, then that country might not be a good market for music industry, and launching songs in that country may not be a good idea in terms of businesses.

4.6.5 Bolivia's Case Study:

In 2019, Songs with instrumentalness, had a large negative correlation with popularity, implies that those songs were flop in the market.
But, when we filter the data towards more popular songs, we see that songs with instrumentalness were not that bad.

This can be seen in the following heatmap:

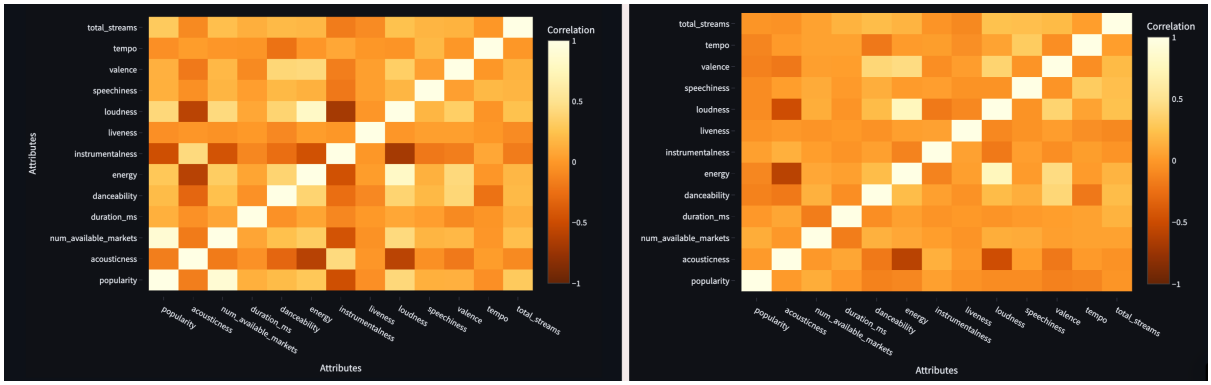


Figure 29: Bolivia: 2019: Left: Non-popular Right: Popular

Here, we observed that when the instrumentalness songs were flop in the market, there was a negative correlation with loudness, danceability and energy. Generally people love songs with danceability, instrumentalness and loudness which was not in this case, hence this might be the reason that these songs didn't preformed well.
But if we see in the heatmap of more popular songs, it can be observed that these properties were balanced well i.e., large negative correlation was not observed in popular songs.

In the following years, Bolivia's Music Industry kept note of this and songs with instrumentalness got better popularity.

Conclusion: Mix and Master can act as a great tool for Music Producers to analyse the correlation of song's properties over the years for different countries. As we have seen, it can also analyse the future pattern of music market in a particular country. Most importantly, it can give the visual analysis of songs' properties of popular songs, and non-popular songs, comparing which, we can increase the chances of our future releases to be popular in the market.

4.7. Insight from Music Scale Analysis

Imagine a music director (say, Alice) tasked with creating a song that brings high profit. As the romantic genre is highly popular worldwide, it is reasonable that Alice may choose to compose a romantic song. To accomplish this task, Alice can seek assistance from this Music Scale Analysis dashboard to explore the relationships between musical scales and key song properties. The goal for Alice, in such a scenario, would be to identify the scales that tend to produce the desired romantic feel while ensuring an optimal production environment.

Figure 30: Speechiness of G# Minor

For a song to sound more romantic, Alice can focus on achieving a natural tone, relying on high speechiness to avoid excessive artificial effects. By using the Box Plot visualization in this utility, Alice can discover that G# Minor would be the best key for this task, indicating a scale that could provide the desired level of speechiness. The Box Plot summary statistics suggests low values, hinting at a left-skewed distribution. However, this is not entirely surprising since modern songs tend to have low speechiness compared to other song properties, a pattern further supported by the Pie Chart visualizations. Considering all the plots in entirety, G# Minor stands out as the most suitable scale for romantic compositions.

Next, Alice may desire to ensure a comfortable experience for their singer, requiring a scale with low liveness to avoid the complexities of live performances. The Heatmap visualization in this utility indicates that F Minor is the best scale for this task, as it has the lowest liveness. However, to maintain the original sharpness from G# Minor, Alice would decide to transpose the composition from F Minor to F# Minor (on the basis of nearest sharp key), as supported by the Heatmap with almost no change in liveness. This change would maintain the low liveness while adding a subtle difference in tonality.

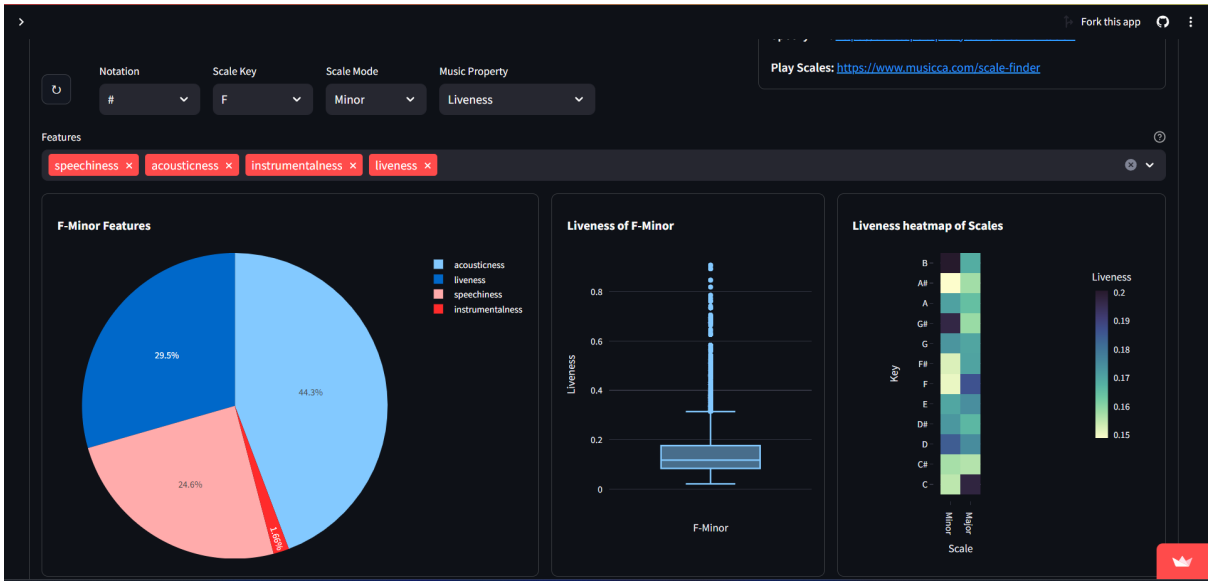


Figure 31: Liveness of F Minor

Interestingly, the song Tum Hi Ho appears to have followed this exact pattern, with the raw composition in F Minor and the film version in F# Minor. This finding can be verified from a music scale finder and an analyzer. As one further explores this Music Scale Analysis utility, it can be noticed that random combinations within the F Minor and F# Minor scales often produces a romantic sound, with F# Minor having a slightly higher note, adding a touch of melancholy to the overall feel from F Minor. These insights suggest that experimenting with F Minor and F# Minor could lead to creating more quality romantic songs, offering valuable guidance for music directors and producers seeking to make popular and commercially successful music.

4.8 KNN Music Recommendation

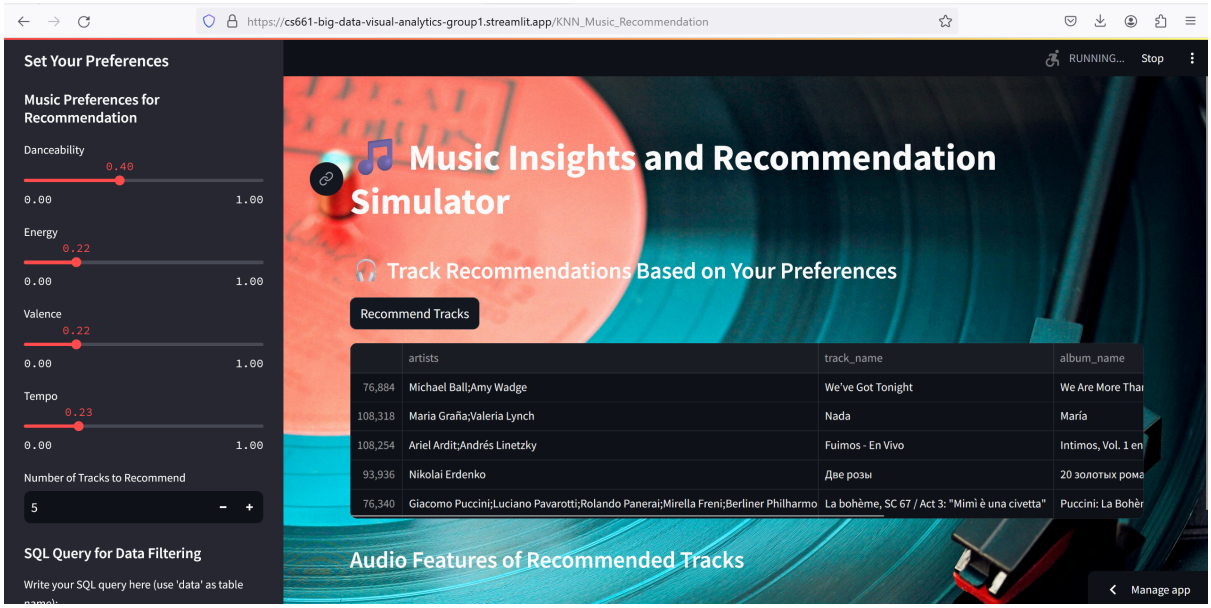


Figure 32: KNN music recommendation

Figure 33: Scatter plot matrix of Recommendation

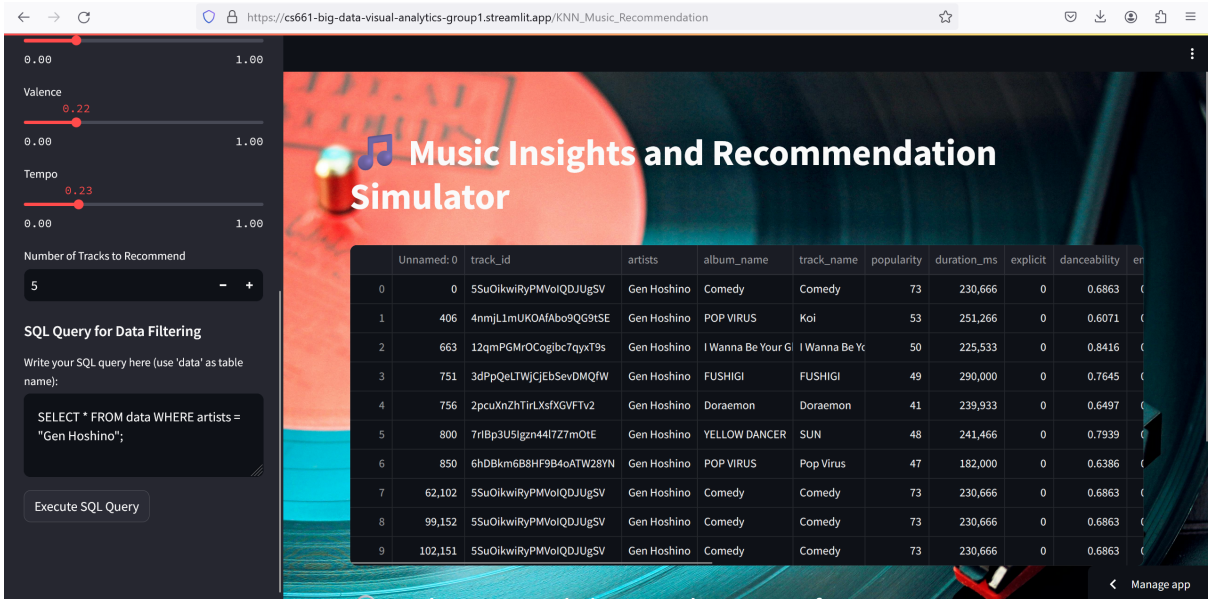


Figure 34: Performing SQL Query in Our Data

4.9 Big Data handling

The implementation of our proposed solution yielded significant improvements in dataset loading speed, memory efficiency, and overall performance. Key results obtained from addressing the challenges of handling big data include:

1. Substantial Reduction in Loading Time
2. Reduction in total memory usage from about 1 GB to sub 100 MB.

5. Conclusion

In conclusion, our project on visualization and analysis of music data has yielded valuable insights and solutions to address the complexities of the modern music industry. Through the integration of advanced data analytics techniques, interactive visualizations, and efficient data processing strategies, we have empowered music producers with the tools and knowledge to make informed decisions and navigate the dynamic landscape of the music market.

From the identification of top genres and artists to the analysis of deep song properties to the optimization of dataset handling for scalability and efficiency, our project has demonstrated the trans-formative potential of data-driven approaches in music production and marketing. By leveraging the wealth of available music data, we have provided actionable insights into audience preferences, genre trends, musical scales, and artist success metrics, enabling stakeholders to capitalize on emerging opportunities and stay ahead of industry trends.

Furthermore, our project underscores the importance of interdisciplinary collaboration and innovation in addressing the multifaceted challenges of the music industry. By bringing together expertise from data science, music production, and visual communication, we have developed a holistic approach to music data analysis that bridges the gap between raw data and strategic decision-making.

6. Link to source code:

1. **Project Repository (GitHub):**
   https://github.com/deepen-stha/CS661-Big-Data-Visual-Analytics
2. **Web Application (deployed):**
   https://cs661-big-data-visual-analytics-group1.streamlit.app/

References

[1] Python. https://www.python.org.

[2] Pandas. https://pandas.pydata.org.

[3] Numpy. https://numpy.org.

[4] Streamlit. https://streamlit.io.

[5] Yejie. Spotify Weekly Top 200 Songs Streaming Data. https://www.kaggle.com/datasets/yelexa/spotify200, 2022.

[6] Danilo B. Seufitelli, Gabriel P. Oliveira, Mariana O. Silva, and Mirella M. Moro. MGD+: An Enhanced Music Genre Dataset with Success-based Networks. https://doi.org/10.5281/zenodo.8086643, 2023.

[7] Lucas Cantu. Top 5000 Albums of All Time - Spotify feature. https://www.kaggle.com/datasets/lucascantu/top-5000-albums-of-all-time-spotify-features, 2022.