# Mid-Semester Dataset and Documentation Update

Group 3: Jeff Fang, Shoi Rathi, Havyarimana Charles, Weiting Yang, Yuktha Sureshkumar

**Dataset Content Description:**

The dataset described here focuses on analyzing music trends over time by integrating information from Spotify, the *Billboard HOT 100 Songs Spotify Data (1946-2022)* from Kaggle, and historical events in the United States. This dataset aims to analyze music trends over time by incorporating data about music tracks and historical events that may have influenced genre evolution, popularity, and listener preferences. It includes a combination of quantitative music metrics, rankings, and significant historical events, to help provide a comprehensive overview of how music and societal changes interact with each other.

The music data fields in the dataset include track-specific information such as the song titles, artist names, album names, and genres. Each song also includes its rank on the Billboard HOT 100, showing its historical popularity. The inclusion of Billboard rankings in the dataset is very important to our project, as it provides a reliable measure of a song's commercial success and cultural impact over time. Billboard rankings are widely recognized as a benchmark for music popularity, reflecting the tastes and preferences of the broader public. By tracking a song's position on the *Billboard HOT 100*, the dataset can capture shifts in music consumption, listener behavior, and trends, helping us discover what was popular in different eras. By incorporating Billboard rankings, the dataset can quantify shifts in music's popularity, making it possible to analyze the correlation between musical characteristics and historical events. For example, you can observe how certain genres rise or fall in popularity during specific decades and how these changes align with the broader cultural and political landscape. This allows for more concrete comparisons between the music's features and the historical context, revealing patterns and trends that might otherwise be overlooked in a purely qualitative analysis.

The dataset also incorporates a variety of audio characteristics from Spotify, which provide a deeper layer of analysis for understanding music trends and how they have evolved over time. Each of these characteristics offers specific insights into a song's musical features, helping to paint a clearer picture of how music's mood, style, and listener appeal have shifted alongside societal changes.

The tempo of a song, measured in beats per minute, indicates the song's speed and rhythm. This feature is especially useful for understanding the energetic nature of different musical eras. For instance, faster tempos often correspond with dance-centric genres like disco in the 1970s or electronic dance music (EDM) in the 2010s, while slower tempos might be more prevalent in genres like soul or ballads, which often dominate during more introspective or reflective periods.

The energy attribute measures a track's intensity and liveliness, helping to classify songs along a spectrum from relaxed to highly vigorous. High-energy music, which often includes rock, hip-hop, and punk, can reflect periods of political unrest or social upheaval, as these genres frequently channel strong emotions or calls for action. Conversely, low-energy tracks, like those

found in acoustic genres, might be more prominent during times when softer, more soothing music is preferred, such as during economic downturns when listeners may seek more calming music as a coping mechanism.

The danceability of a track measures how suitable it is for dancing, factoring in rhythm stability, beat strength, and overall tempo. Danceable tracks often correlate with periods of cultural exuberance, such as the rise of disco in the 1970s, hip-hop's dance evolution in the 1990s, and the mainstream surge of pop in the 2000s and 2010s. This attribute allows for analysis of how music catered to social gatherings and entertainment trends during different timeframes.

Acousticness quantifies the likelihood that a track is acoustic, offering insights into shifts between more natural, raw sounds and digitally produced music. High levels of acousticness, common in folk, classical, and some indie genres, often rise in popularity during movements that emphasize authenticity, minimalism, or a return to simpler roots. Low acousticness, on the other hand, corresponds with the rise of heavily produced, synthesized sounds seen in genres like electronic music, indicating cultural trends favoring innovation and modernity.

Valence measures the positivity conveyed by a track, offering insights into the overall mood of music during different eras. High valence tracks are generally perceived as cheerful, happy, and uplifting, while low valence tracks tend to be sad, moody, or serious. This feature can be particularly useful in analyzing how music responds to historical contexts. For example, during periods of social optimism or economic growth, high-valence music (e.g., pop, dance) might become more prevalent, while low-valence music (e.g., blues, grunge) might gain traction during times of political turmoil, economic recessions, or societal discontent.

By including release dates, the dataset allows for temporal analysis across different decades, helping to identify trends within these audio characteristics over time. This chronological context is essential for mapping out how music characteristics align with or diverge from historical events and cultural shifts, providing insights into how societal contexts have influenced the sound, mood, and style of popular music.

The historical context data within this dataset plays a crucial role in analyzing music trends by offering insights into how significant cultural, economic, and political events have influenced music over time. Each event in the dataset is not only marked by its date but also accompanied by a detailed description that outlines its broader implications. This helps to identify patterns where major shifts in society may correlate with changes in musical themes, genres, or popularity. For example, technological advancements are included, such as the launch of the first U.S. weather satellite (Tiros I) in 1960 and the invention of the ARPANET in 1969, the precursor to the internet. The dataset allows for exploration of how such technological milestones impacted music production, distribution, and consumption. For instance, the rise of television and radio broadcasting during the postwar years played a major role in the popularization of music, with more homes having access to these technologies by the 1960s, changing the way music reached audiences. By linking these cultural, economic, and political events to music trends, the dataset helps us understand not just what was popular, but why

certain genres or themes gained prominence during specific periods. The dataset's structure enables this kind of detailed analysis, revealing the deeper relationship between music and historical context.

The dataset was compiled from multiple sources, including Spotify data, the Kaggle dataset, and web-scraped historical records. The integration process involved aligning the data by matching time periods, particularly by decades, to enable meaningful analysis of the interplay between music and historical events. In the reconciliation process, irrelevant rows, such as pre-1960 entries, were removed to ensure consistency and focus on relevant time periods, given that Billboard started tracking the HOT 100 in 1958. Additionally, fields that were deemed unnecessary for the analysis, such as Spotify URIs or time signatures, were omitted to retain only attributes that contribute to understanding trends in music characteristics. This careful selection and combination of data sources have ensured that the dataset is both comprehensive and focused, facilitating a detailed exploration of how music and historical contexts intersect over time.

**Data Collection Methodology:**

In the beginning, we started looking for music data from Spotify and Kaggle. In Spotify, music metadata and statistics are collected by Spotify through its API. Various variables such as music tracks, albums, artists, genres, release dates, popularity scores, song characteristics like tempo, energy, and acousticness, and danceability are included. On Kaggle, we found the dataset of Billboard HOT 100 Songs Spotify Data (1946-2022), where it has songs with their rank on Billboard charts, album name, audio features, etc. Additionally, this dataset combines Spotify's real-time metadata and the historical scope of the Kaggle dataset, which allows us to compare contemporary music trends with past decades, facilitating a richer understanding of genre evolution and listener preferences. Our group decided to use this dataset as our foundation and template to create our dataset for this semester-long project. To enhance the analysis of music trends over time, we looked up *"USA History Timeline"* to discover important events or incidents related to music. In this article, it layed out crucial cultural, political, and social events that have historically influenced music. Meanwhile, our group got the opportunity to familiarize ourselves with cultural shifts, political events, economic factors, and social movements happening in the United States that may affect the genre popularity or lyrics of music at that time. For this *"USA History Timeline,"* we decided to apply web scraping learned in this course to contextualize our music dataset.

**Challenges:**

The main challenge in creating the music dataset for this semester-long project was to narrow down our chosen topic and be more specific with our questions in analyzing music trends over time. From the feedback given a few weeks ago, we realized it was too broad and unfit for us to begin analysis. Initially we wanted to study music trends throughout history. This is why we settled on the billboards data spanning from 1946 to 2022. However, to contextualize and

bring meaningful analysis to this project, we decided to create our own dataset to work in tandem with the spotify dataset. We utilized web scraping to collect major cultural events to connect with the later analysis. We will drive insights from segments of history and showcase how music trended during that time.

**Criteria in Feature and Data Points Selection:**

After combining all the data and information collected through Spotify, Kaggle, and *USA History Timeline* and considering the feedback we got, we removed all the rows having music before 1960 because Billboard started tracking the Hot 100 in 1958. We rounded to 1960 so we can better section our analysis into decades. We dropped rows with missing release dates as they would not be useful. We also dropped some columns such as Spotify URI, Song Image, Mode, and Time Signature. The Spotify URLs are not directly useful for our topic analysis; the song images do not contain any attributes that impact the music's characteristics or popularity trends. The Mode was deleted as we already have specific attributes like energy, valence, or danceability to capture the trends changes in music. From our preliminary analysis, we found out that most popular songs follow common time signatures like 4/4, so this column will not affect our later music trends prediction.

**Responsibility and Contributions**

**Charles**: responsible for data cleaning, column adjustment, removal of unnecessary rows and columns from Spotify dataset.
**Jeff** : responsible for compiling information/data historical events, social movements, and political trends that are related to or affected music trends over time.
**Weiting**: responsible for dataset description and organizing the formatting and writing of this midterm project paper
**Shoi**: responsible for web scraping to contextualize the group's music dataset
**Yuktha**: responsible for writing out the data content description

## Reference

America's best history timeline. (n.d.). https://americasbesthistory.com/abhtimeline.html

Dave, D. (2021, November 9). Billboard "The Hot 100" songs. Kaggle. https://www.kaggle.com/datasets/dhruvildave/billboard-the-hot-100-songs

Sharma, T. (2023, June 1). Billboard hot 100 songs Spotify Data (1946-2022). Kaggle. https://www.kaggle.com/datasets/tushar5harma/billboard-hot-100-songs-spotify-data-1946-2022