General Idea

Our project was inspired by the absence of an integrated platform capable of exploring the multifaceted evolution of music over time. While extensive datasets for **Spotify tracks** exist, with comprehensive information about the songs, it is not accompanied by a tool that allow us to visualize these information.

Our objective is to develop an interactive dashboard that identifies meaningful trends in musical characteristics and popularity. By consolidating Spotify's audio features with external sources from Kaggle and other websites.

Datasets

The primary dataset was sourced from Kaggle (Spotify Tracks Dataset) and contains $89,741$ songs, each associated with specific genres and various audio attributes. To ensure relevance, we do not utilize the entire corpus; rather, we filter these tracks against a secondary dataset.

The second dataset, also from Kaggle (The Hot 100 Songs), represents the music industry standard record chart in the United States, published weekly by Billboard magazine. These rankings are calculated based on sales, radio airplay, and online streaming metrics.

To further enrich our data, we performed webscraping on the WhoSampled database. This process allowed us to integrate the year of publication and quantitative metadata regarding samples, covers, and remixes (both incoming and outgoing). This added layer provides a deeper understanding of a song's cultural influence and historical context.

The resulting master dataset consists of $2,655$ unique observations across the following 15 attributes:
track_id, artists, track_name, popularity, duration_ms, danceability, energy, speechiness, acousticness, instrumentalness, loudness, liveness, valence, track_genre, year, out_connections, and in_connections.
With $2,655$ rows and 16 columns, the dataset yields an **AS index** of approximately $42,000$ data points ($2,655 \times 16 = 42,480$).

In addition to this master file, we maintain a longitudinal dataset of $46,000$ rows from the Billboard Hot 100 (filtered for Spotify compatibility). This secondary structure enables the creation of temporal evolution charts, specifically focusing on genre-ranking shifts over time. This dual-dataset approach provides a robust framework for evaluating the evolution of musical tastes and industry trends across several decades.

Intended Users

The target user is a music analyst or professional curator (e.g. a Spotify playlist editor or a radio music director). This user needs to move beyond simple "top charts" to understand the structural similarities between songs, identify emerging genre trends over time, and curate playlists based on specific audio profiles (e.g. sad but high-energy songs).

Dimensionality Reduction

Our project will include a **dimensionality reduction** technique so that the user can visualize the high-dimensional audio features of the dataset on a **2D scatterplot**.

The analysis of the data done by dimensionality reduction, particularly with non-linear techniques, requires significant computation. Therefore, it is not feasible to run the full algorithm in

real-time while the user interacts with the application. For this reason, we decided to perform a preprocessing phase in which we run the algorithm and store the result in the dataset, adding two new columns (e.g. $tsne_x$ and $tsne_y$) for the projected coordinates.

For our project, we will use **t-SNE** (t-Distributed Stochastic Neighbor Embedding). We chose t-SNE over other methods like PCA or MDS because it is particularly effective at preserving local structure, meaning it excels at keeping similar songs close together in the low-dimensional space. This makes it ideal for our goal of revealing distinct "clusters" of music (e.g. separating acoustic ballads from high-energy pop tracks) that might be overlapped or obscured in a linear projection.

Possibly, we will run the t-SNE algorithm with different perplexity values during the preprocessing stage and let the user toggle between these pre-calculated views. This would allow the user to see the data structure at different scales, balancing between local details and global geometry, without needing to wait for a live recalculation.

Visualization

Our project will contain the following visualizations:

- **Scatter plot (t-SNE Projection)**: representing the Dimensionality Reduction results. This view will map the high-dimensional audio features (danceability, energy, acousticness, etc.) onto a 2D plane, clustering songs that "sound similar" near each other.
- **Bubble plot**: used to visualize the relation between the attributes of the dataset, allowing to select the axis's attributes.
- **Box plots**: used to plot the statistical distribution of specific audio attributes.
- **Parallel coordinates**: used to visualize the audio profile of selected tracks at once. This will feature vertical axes for numerical Spotify features (*Danceability, Energy, Speechiness, Acousticness, Instrumentalness, Liveness, Valence*), allowing users to see the shape of a song or a cluster of songs.
- **Ranking bump chart**: used to visualize the temporal evolution of the **Top 10 genres**. Since standard line charts can be messy for rankings, we will use a bump chart where the X-axis represents time (Years) and the Y-axis represents the Rank (1st to 10th). Each genre is represented by a colored "flow" line that moves up or down between ranks as time progresses. This allows the user to trace macro-trends in music history, such as the gradual rise of Latin music or the stability of Pop, without the noise of individual song fluctuations.

User Interaction and Analytic

All the visualizations (or surely most of them) in the project will be related to each other: if the user interacts with one part, the others will change accordingly. More specifically:

- The **t-SNE scatterplot** will have a brushing feature (lasso or rectangular selector) to select clusters of similar-sounding movies/songs. When a cluster is selected here, the Parallel Coordinates will immediately update to show the specific audio profile of that cluster (e.g. revealing that this specific group of songs is defined by High Energy but Low Valence).
- The **bubble plot** will also have a brushing feature, allowing the user to select specific moods. This selection will filter the list of songs shown in the other views.

- The **parallel coordinates** chart will have a brushing feature on each vertical axis. This allows the user to perform complex queries visually, such as "Show me only songs with High Danceability AND Low Acousticness." The Scatterplot will highlight the songs that meet these criteria.
- **Triggered Analytic** (centroid calculation): When a cluster of songs is selected in the t-SNE scatterplot (via brushing), the system will trigger a real-time computation to calculate the "Centroid" (average audio profile) of that specific selection. The Parallel Coordinates chart will then overlay this calculated "average song" as a bold line on top of the individual tracks. This allows the user to instantly see the mathematical "fingerprint" of the selected cluster.
- The **bump chart** will serve as a context filter. Hovering over a specific genre line (e.g. 'Rock') will highlight all 'Rock' songs in the t-SNE scatterplot and Parallel Coordinates, allowing the user to drill down from the macro-trend of the genre to the specific audio features that define it.

Mockup of the user interface (DRAFT)