Visual Analysis 2024-2025

Project Approval Draft

Marco Natale (1929854) – Sahar Khanlari (2107563)

January, 13

Spotify Music Dataset

Dataset Specification: The Spotify Music Dataset consists of two CSV files categorizing songs by popularity: one for high popularity songs and another for low popularity songs. These files will be merged into a unified dataset. Each song includes attributes such as tempo, danceability, energy, loudness, valence, genre, release year, and artist. The merged dataset provides an enriched perspective for analyzing and comparing songs based on their popularity.

https://www.kaggle.com/datasets/solomonameh/spotify-music-dataset?resource=download&select=low_popularity_spotify_data.csv

General Idea:

- **Analytics Part:** The project will use dimensionality reduction techniques to condense high-dimensional song features into a 2D space. This analysis will help identify clusters of songs based on similarities in their musical attributes, with an additional focus on comparing high and low popularity songs.
- **Visual Part:** The project will feature two coordinated and interactive visualizations: (1) a scatterplot representing songs in the reduced-dimensional space, with points color-coded by their popularity category (high or low), and (2) a bar chart showing aggregated feature values (e.g., average danceability, loudness) for selected clusters or popularity groups. Users can interact with the scatterplot to filter and update the bar chart dynamically.

Intended User: The project targets music enthusiasts, and industry professionals (e.g., DJs, playlist curators) who want to explore and analyze music trends, identify clusters of similar songs, or discover patterns in song popularity.

Used Analytics:

1. **Dimensionality Reduction:** PCA or t-SNE will reduce the dataset's numerical features to two dimensions.
2. **Clustering:** Songs in the reduced-dimensional space will be grouped based on proximity, representing similarity in musical features. Popularity will be overlaid as an additional dimension.
3. **Interactive Computation:** When users select clusters or popularity groups, the system will compute additional analytics, such as the cluster centroid's attributes, feature distributions, or the similarity score between clusters.

Relation to Visual Analytics Cycle:

1. **Data Preparation:** The dataset is cleaned, normalized, and prepared for dimensionality reduction. Numerical features like tempo, energy, and valence are standardized.
2. **Dimensionality Reduction:** High-dimensional features are projected onto a low-dimensional space for visualization.
3. **Interactive Exploration:** Users interact with the scatterplot to explore song clusters, trigger additional computations, and refine their analysis.
4. **Feedback Loop:** Insights from user interactions guide further analysis, such as comparing clusters with specific popularity levels or adjusting the dimensionality reduction parameters.
5. **Knowledge Generation:** The coordinated visualizations help users identify trends, relationships, and outliers in the data, supporting informed decisions or discoveries.

Mockup