Spotify Dashboard

Sahar Khanlari

February 2025

1. Introduction

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique that helps in visualizing high-dimensional datasets. This report outlines the design process, rationale, and prototype development of the **Spotify Dashboard**. The dashboard provides insights into Spotify's audio features by clustering songs based on their characteristics. Additionally, we review related work and present the insights discovered through our analysis.

2. Design Process and Rationale

2.1 Data Collection and Preparation

The dataset comprises two CSV files: `high_popularity_spotify_data.csv` and `low_popularity_spoti`
These datasets were combined, and duplicate tracks were removed to ensure data integrity.
We selected key audio features, including *energy, tempo, danceability, loudness, liveness, valence, speechiness, instrumentalness, mode, key, duration (ms), and acousticness*, for PCA.

2.2 Principal Component Analysis (PCA)

PCA was performed to reduce the dimensionality of the dataset to two principal components, facilitating visualization. The data was standardized before applying PCA to ensure that each feature contributed equally to the analysis. The PCA results were added to the original dataset as `pca_x` and `pca_y`.

2.3 Clustering

We employed the K-Means clustering algorithm to identify patterns within the data. The optimal number of clusters was determined using the Elbow Method, which suggested four distinct clusters. The clustering results were visualized on a scatterplot, with each cluster represented by a different color.

3. Prototype Development

The dashboard was developed using Python for data preprocessing, clustering, and dimensionality reduction, and JavaScript for creating interactive visualizations in the browser.

Python libraries such as `pandas` were used for data manipulation, `scikit-learn` for implementing PCA and K-means clustering, and `matplotlib` for basic plotting during analysis. For front-end visualization, JavaScript was employed with the support of the `D3.js` library, enabling dynamic and responsive charts.

The prototype features:

- **Scatterplot:** Displays clusters based on PCA components, with each cluster represented by a different color from a vibrant color palette. Clicking on a data point filters the other visualizations to display information relevant to the selected cluster.
- **Bar Chart:** Shows the average values of selected audio features (e.g., energy, danceability, loudness) for each cluster. The bar chart dynamically updates when a cluster is selected.
- **Date Interval Selector:** A slider-based tool that allows users to filter songs by their release date. The selector updates dynamically based on the selected cluster or filters.
- **Top 10 Genres, Artists, and Tracks:** Provides insights into the most popular genres, artists, and tracks within the dataset. Rankings are based on popularity metrics. Users can filter by genre or artist, which updates other visualizations accordingly.
- **Dataset Overview:** A summary panel displaying key statistics such as total number of songs, average energy, danceability, valence, and loudness. This helps users grasp the dataset's general characteristics.

The dashboard incorporates interactive features such as real-time data filtering, hover effects, and responsive design adjustments based on screen size.

4. Discovered Insights

Several key insights were uncovered through our analysis:

- **Cluster Characteristics:** Different clusters exhibited distinct audio feature profiles. One cluster showed high energy and loudness (upbeat, danceable tracks), while another had high acousticness and instrumentalness (mellow, instrumental songs).

- **Genre and Artist Trends:** The Top 10 Genres and Artists highlighted popular genres like K-pop and artists such as Bruno Mars and Rosé.
- **Temporal Patterns:** Most songs were released in recent years, with exceptions in rock music.
- **Dataset Structure:** The dataset, built from playlists, includes duplicate songs and shows bias toward recent releases.
- **Feature Correlations:** Energy and loudness show a strong positive correlation, while acousticness correlates negatively with both. Danceability correlates moderately with valence, indicating that danceable songs often have a happier tone.

5. Conclusion

The Spotify Dashboard effectively visualizes complex audio data, enabling the discovery of meaningful patterns and trends. By integrating PCA and clustering, it provides valuable insights into song characteristics, genre popularity, and temporal trends. This approach could be extended to enhance music recommendation systems and user experiences in streaming platforms.