Spotify Dashboard

Visual Analytics

Group 12
Marco Natale 1929854
Sahar Khanlari 2107563

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_spotify_data.csv`. 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 Principal Component Analysis (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. The scatterplot is interactive. 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, reflecting the corresponding feature averages. Animations are incorporated for smooth transitions during data updates.
- **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 quick insights into the most popular genres, artists, and tracks within the dataset. The rankings are based on a popularity metric derived from the dataset. Users can filter the data by selecting specific genres or artists, which updates the scatterplot and bar chart accordingly.

- **Dataset Overview:** A summary panel displaying key statistics, including the total number of songs, average energy, danceability, valence, and loudness. This overview helps users quickly grasp the dataset's general characteristics and compare them across different clusters.

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

4 Discovered Insights

Through our analysis, several key insights were uncovered:

- **Cluster Characteristics:** Different clusters exhibited distinct audio feature profiles. For example, one cluster showed high energy and loudness, indicating upbeat, danceable tracks, while another cluster had high acousticness and instrumentalness, suggesting more mellow, instrumental songs.

- **Genre and Artist Trends:** The Top 10 Genres and Artists provided insights into current musical trends, highlighting the popularity of genres like K-pop and artists such as Bruno Mars and ROSÉ.

- **Temporal Patterns:** The Date Interval Selector revealed that the majority of songs in the dataset were released in recent years, with some exceptions, particularly in rock music.

- **Dataset Structure Insights:** We noticed that, because the dataset is built from a collection of playlists, the same song appears multiple times as it is included in different playlists. The dataset also shows a bias towards more recent releases, with significantly more songs from the past few years compared to older tracks.

- **Feature Correlations:** Strong relationships were revealed between certain audio features. For example, energy and loudness have a strong positive correlation, indicating that louder songs tend to be more energetic. In contrast, acousticness shows a strong negative correlation with both energy and loudness, suggesting that acoustic tracks are generally softer and less energetic. Additionally, danceability has a moderate positive correlation with valence, meaning that more 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 techniques, the dashboard provides valuable insights into song characteristics, genre popularity, and temporal trends. This approach can be further extended to enhance music recommendation systems and user experience in streaming platforms.

6 Related Works

In the field of music analytics, dimensionality reduction and clustering techniques have been extensively applied to uncover patterns and enhance music recommendation systems. This section discusses the objectives, datasets, and analytical techniques employed in three related research papers, providing a comparative perspective.

The paper *Analysis of Machine Learning-Based Music Recommendation System Using Spotify Datasets* (Li, 2024) focuses on the application of machine learning techniques, particularly Principal Component Analysis (PCA) and K-means clustering, in music recommendation systems. The study utilizes Spotify datasets to analyze how machine learning can personalize music recommendations by reducing data dimensionality and grouping songs with similar features. PCA helps retain essential information while simplifying the dataset, and K-means clustering identifies consistent musical patterns. The visualization of clusters facilitates a deeper understanding of genre relationships and user preferences.

Classical Music Clustering Based on Acoustic Features (Wang & Haque,

2017) investigates the clustering of 330 classical music pieces from the MusicNet database. Unlike the first study, this research focuses on classical music and employs spectral clustering based on musical note sequences. The authors introduce novel feature extraction methods, such as shingling and chord trajectory matrices, to capture the compositional styles of different eras and composers. The dataset differs significantly from Spotify datasets, emphasizing acoustic features inherent in classical compositions rather than modern streaming metrics.

The third paper, *Enhanced Music Recommendation Systems: A Comparative Study of Content-Based Filtering and K-means Clustering Approaches* (Mukhopadhyay et al., 2024), compares content-based filtering and K-means clustering techniques using an extensive Spotify dataset. The study aims to optimize music recommendation accuracy by evaluating both methodologies. Content-based filtering focuses on audio features like danceability, energy, and loudness, providing highly personalized recommendations. In contrast, K-means clustering groups songs based on shared characteristics, revealing broader patterns within the dataset. The research highlights the superior performance of content-based filtering in recommendation accuracy while acknowledging the robustness of K-means in pattern discovery.

While PCA and K-means clustering are common across the first and third studies, the second paper introduces spectral clustering, demonstrating the adaptability of clustering methods to different music genres and datasets.

6.1 Comparison

Comparing these works to our project, several similarities and differences emerge. Like Li (2024) and Mukhopadhyay et al. (2024), our work employs PCA and K-means clustering to analyze Spotify datasets, focusing on dimensionality reduction and pattern discovery. However, while their studies emphasize music recommendation systems, our project centers on visualizing audio features through a PCA dashboard to derive insights into song characteristics and genre trends. Unlike Wang & Haque (2017), who concentrated on classical music and used spectral clustering, we targeted a broader range of contemporary music from Spotify, leveraging modern streaming data. Despite these differences, the common goal across all studies is to uncover meaningful patterns in musical data, enhancing our understanding of music analytics and recommendation methodologies.