Kun Yan

Dataset: For my project, I have used the "Top 1000 Songs" dataset sourced from Spotify playlists, which provides musical features for songs spanning decades from the 1920s to the 2010s. I found it from https://www.kaggle.com/datasets/srishreya/top-1000-songs-from-1920-to-2020 (Kaggle). The dataset includes attributes such as danceability, acousticness, energy, valence, and tempo. Each row in the dataset represents a song, and the features provide quantitative insights into its characteristics. This data allows us to analyze trends in music over time and identify patterns within different eras.

Project Idea and Implementation: I implemented two main analyses in my program: Principal Component Analysis (PCA) and K-means clustering. PCA was used to reduce the dimensionality of the dataset from five features (danceability, acousticness, energy, valence, and tempo) to two principal components. This technique highlights the most influential factors that explain the variance in the data, enabling a simpler representation for further analysis. Following PCA, I implemented k-means clustering to group songs based on their feature similarities. The algorithm grouped the songs into two distinct clusters: one characterized by high-energy, danceable tracks with lower acousticness and the other featuring low-energy, acoustic-focused songs with more balanced attributes. Additionally, I calculated feature averages for each cluster and across decades to observe how musical attributes like danceability and energy evolved over time.

Result: The analysis groups the dataset into two distinct clusters based on musical features. Cluster 0 includes 402 songs characterized by high energy, low acousticness, and fast tempos, representing modern, upbeat music styles. Cluster 1, with 597 songs, features tracks with high acousticness and lower energy, indicative of acoustic or slower-paced genres. These clusters offer valuable insights into the diversity of music trends and can be used for genre-specific recommendations or trend analysis.

The clustering and decade-wise analysis revealed significant insights into music trends. Songs in the first cluster were associated with high danceability and energy, often representing modern music genres like pop, EDM, and rock. These songs typically exhibited lower acousticness, reflecting the dominance of electronic and digital production in recent decades. Songs in the second cluster were more acoustic-focused with lower energy, often aligning with ballads, folk, or classical music. These tracks featured higher acousticness and balanced valence, reflecting a broader emotional range. Decade-wise, danceability, energy, and tempo increased significantly from the 1980s onward, driven by the rise of dance and electronic music, while acousticness decreased, highlighting the shift from traditional instrumentation to digital production techniques.