Write Up: Analyzing The Rolling Stones Spotify Data

This project aimed to perform an in-depth analysis of The Rolling Stones' Spotify data, covering data preparation, exploratory data analysis (EDA), feature engineering, dimensionality reduction, and cluster analysis to uncover insights into their musical characteristics and popularity trends.

1. Data Preparation and Cleaning

The initial phase involved loading the dataset and conducting thorough data inspection. Key steps included:

- Duplicate Check: Confirmed no duplicate rows were present in the dataset.
- Missing Values: Identified no missing values across any columns.
- Data Types: Converted the release_date column to datetime objects to facilitate time-series analysis.
- Irrelevant Features: Dropped the 'Unnamed: 0' column, which was found to be redundant.

2. Exploratory Data Analysis (EDA) and Feature Engineering

EDA was performed to understand the dataset's structure, identify patterns, and engineer new features. Key insights and tasks included:

- Popular Album Identification (3.a):
    - Defined 'popular songs' as those with a popularity score of 30 or higher (based on the average popularity of 20.79).
    - Identified the top albums by the number of popular songs. The two albums with the highest number of popular songs were 'Exile On Main Street (2010 Re-Mastered)' (18 popular songs) and 'Blue & Lonesome' (12 popular songs).
    - Visualized the top 10 albums by popular song count using a bar plot.
- Correlation Analysis of Audio Features (3.b):
    - Generated a correlation matrix and heatmap of key audio features (energy, danceability, popularity, acousticness, instrumentalness, liveness, loudness, speechiness, tempo, valence, duration_ms).
    - Observed weak correlations between most individual audio features and song popularity, suggesting that popularity is influenced by a complex interplay of factors rather than a single feature.
- Popularity vs. Various Factors (3.c):
    - Popularity Evolution Over Time: Created a release_year column and plotted the average song popularity across release years, revealing trends in The Rolling Stones' popularity over their career.
    - Individual Feature Relationships: Explored the relationship between popularity and acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, and valence through scatter plots and correlation coefficients. These analyses largely confirmed weak linear relationships.

3. Dimensionality Reduction (PCA)

To identify underlying patterns and simplify the high-dimensional audio feature space, Principal Component Analysis (PCA) was applied. (3.d)

- Significance: PCA was used to reduce the number of features while retaining the maximum variance, making the data more manageable for visualization and subsequent analysis, and potentially reducing noise and multicollinearity.
- Application: Audio features were standardized and PCA was applied.
- Explained Variance: The explained variance ratio for each principal component and the cumulative explained variance were analyzed. It was observed that a significant portion of the variance could be explained by a few principal components.
- Component Loadings: The loadings were examined to understand the contribution of original features to each principal component, providing insight into what each component represents (e.g., PC1 captures general 'activity' while PC2 captures 'acousticness' vs. 'danceability' differences).

4. Cluster Analysis

K-Means clustering was performed on the scaled audio features to group songs with similar characteristics.

- Optimal Number of Clusters (4.a):
    - The Elbow Method was used, plotting the Within-Cluster Sum of Squares (WCSS) for various numbers of clusters. An 'elbow' was observed around k=3 or k=4, suggesting these as potential optimal numbers.
- K-Means Clustering and Visualization (4.b):
    - K-Means was performed with k=3 (chosen for better visualization and interpretability based on a previous iteration).
    - Cluster assignments were added to the DataFrame.
    - PCA was used to reduce the data to two dimensions for visualizing the clusters on a scatter plot, showing distinct groupings of songs.
- Cluster Definition (4.c):
    - The mean values of audio features for each cluster were calculated to characterize them:
        - Cluster 0: Acoustic & Mellow: Songs with higher acousticness, lower energy, moderate danceability, and the shortest duration.
        - Cluster 1: Energetic & Danceable: Songs characterized by high danceability, energy, and valence (positivity), with lower acousticness and moderate loudness. These songs tend to be more upbeat.
        - Cluster 2: Loud Live Performances: Songs exhibiting very high energy, high loudness, and high liveness, indicating a strong likelihood of being live recordings. They also have the highest speechiness and longest duration, and are the least danceable.

Conclusion

This analysis provided a comprehensive understanding of The Rolling Stones' music, from identifying their most popular albums to categorizing their songs into distinct clusters based on audio features. While direct correlations between single features and popularity were weak, the clustering revealed inherent groupings that could be used for recommendation systems or further targeted analysis. The dimensionality reduction with PCA helped condense complex audio characteristics into interpretable components.