**Analysis Report: Creating Cohort of Songs**

**Executive Summary**

This report presents an analysis of a dataset comprising 1,610 songs by The Rolling Stones, aiming to identify key characteristics and recommend strategies for creating song cohorts. The analysis covers data cleaning, exploratory data analysis (EDA), and clustering. Key findings include: the album "Sticky Fingers (Remastered)" is the most popular, high energy and loudness are strongly correlated, and the optimal number of clusters for the dataset is either 3 (based on the Elbow Method) or 5 (based on the Silhouette Score). The primary goal is to leverage these insights to build effective song cohorts.

**Data Overview**

The dataset consists of 1,610 songs by The Rolling Stones. It contains 18 initial columns, which were processed to 20 columns after data cleaning. The cleaning process involved:

* Dropping an "Unnamed: 0" column.
* Renaming columns like release\_date and track\_number.
* Changing the release\_date data type to datetime and extracting the month and year.
* The dataset was confirmed to have no null or missing values.

**Key Audio Features Analyzed**

The analysis focused on a set of nine key audio features:

* **Danceability:** A measure of how suitable a track is for dancing.
* **Energy:** A perceptual measure of intensity and activity.
* **Loudness:** The overall volume of the track in decibels (dB).
* **Speechiness:** The presence of spoken words in a track.
* **Acousticness:** A confidence measure of whether a track is acoustic.
* **Instrumentalness:** The likelihood that a track contains no vocals.
* **Liveness:** The presence of an audience in the recording.
* **Valence:** A measure of musical positivity (happy vs. sad).
* **Tempo:** The speed or pace of a track in beats per minute (BPM).

**Key Findings**

* **Top Popular Albums:** "Sticky Fingers (Remastered)" is the most popular album with an average popularity score of 53.3. "Some Girls" follows with an average popularity of 48.1.
* **Popular Song Distribution:** Popular songs are not confined to a single decade, but the 1960s-1980s had a higher frequency of high-popularity outliers.
* **Highly Correlated Features:** Energy and Loudness have a strong positive correlation (+0.70). Danceability and Acousticness have a strong negative correlation (-0.52).
* **Popularity Predictors:** Popular music tends to be energetic, loud, and danceable. Songs with high speechiness or acousticness are generally less popular.

**Audio Feature Distribution Insights**

* **Danceability:** Most songs have a moderate danceability score, with the peak around 0.5.
* **Energy & Loudness:** Most songs are high-energy (peaking near 1.0) and loud (peaking around -5 dB).
* **Acousticness & Instrumentalness:** Most tracks are vocal-based and non-acoustic, with these features peaking near 0.
* **Liveness:** The bimodal distribution of liveness (peaks at 0.2 and near 1.0) indicates a mix of both studio and live recordings.
* **Tempo:** The tempo distribution is roughly normal, with a peak around 120 BPM, consistent with many pop and rock songs.

**Correlation Analysis**

A correlation matrix revealed several significant relationships between audio features:

* **Positive Correlations:**
  + **Energy and Loudness (+0.70):** Louder songs are highly likely to be more energetic.
  + **Danceability and Valence (+0.55):** Happier songs tend to be more danceable.
  + **Energy and Valence (+0.51):** Energetic songs are often perceived as more positive.
* **Negative Correlations:**
  + **Danceability and Acousticness (-0.52):** Songs with more acoustic qualities are less danceable.
  + **Energy and Acousticness (-0.36):** Acoustic songs are typically less energetic.
* **Weak Correlations:** Tempo and Instrumentalness show minimal correlation with most other features, suggesting they are relatively independent attributes.

**Popularity Patterns**

* There is **no strong linear correlation** between a song's release year and its popularity.
* High popularity is more common in the earlier decades (1960s-1980s).
* Popularity is positively correlated with danceability, energy, and loudness.
* Popularity is negatively correlated with speechiness and acousticness.

**Clustering Analysis Results**

The analysis used two methods to determine the optimal number of clusters:

* **Elbow Method:** The curve shows a pronounced "elbow" at **K=3**, indicating that 3 clusters provide a good balance between explained variance and model complexity.
* **Silhouette Score:** When clustering on danceability and popularity, the highest silhouette score is found at **K=5**, suggesting that 5 is the optimal number of clusters for this specific combination of features.

**Recommendations**

Based on the analysis, here are recommendations for creating song cohorts:

1. **Prioritize Feature Selection for Cohorts:** Features with strong correlations, like **danceability, energy, and loudness**, should be used as the foundation for defining cohorts.
2. **Define a "Classic Hits" Cohort:** Popularity as a key filter. Identify songs with popularity scores above 50, particularly from the 1960s-1980s, to create a cohort of timeless, high-popularity tracks.
3. **Create Genre-Based Cohorts:** Use the clustering results to create cohorts that represent different musical styles. For example:
   * A **"Dance & Upbeat" cohort** would include songs from the high-danceability, high-energy clusters.
   * A **"Acoustic & Low-Energy" cohort** would be defined by low danceability and low energy/loudness.
4. **Consider a Multi-Cluster Approach:** While the Elbow Method suggests 3 clusters, the Silhouette Score for danceability and popularity suggests 5. Use the 5-cluster model for a more granular segmentation, such as separating high-danceability songs into "medium-energy dance" and "high-energy dance" cohorts.