**Project Objectives and Data**

The main objective is to use exploratory data analysis (EDA) and cluster analysis to create cohorts of songs. These cohorts should be based on various features, with each group containing songs of a similar type. The analysis will help in understanding what factors contribute to the formation of these song cohorts.

The dataset for this project comes from Spotify's API and contains information about all of The Rolling Stones' albums available on the platform. Each song has a unique Spotify ID. The dataset includes variables such as:

* **name**: The song's title.
* **album**: The album's name.
* **release\_date**: The day, month, and year the album was released.
* **acousticness**: A value from 0.0 to 1.0 indicating how acoustic a track is, with 1.0 being highly acoustic.
* **danceability**: A measure from 0.0 to 1.0 of how suitable a track is for dancing.
* **energy**: A measure from 0.0 to 1.0 of intensity and activity, where higher values correspond to faster, louder, and noisier tracks.
* **instrumentalness**: A value from 0.0 to 1.0 predicting if a track contains no vocals.
* **liveness**: A measure of the presence of an audience, with values above 0.8 strongly suggesting a live performance.
* **loudness**: The overall loudness in decibels (dB), typically ranging from -60 to 0 dB.
* **speechiness**: A measure of the presence of spoken words, with higher values indicating more speech-like content.
* **tempo**: The estimated tempo in beats per minute (BPM).
* **valence**: A measure from 0.0 to 1.0 describing the musical positivity, where high values are happy and low values are sad.
* **popularity**: The song's popularity, from 0 to 100.
* **duration\_ms**: The track's duration in milliseconds.

**Steps to Perform the Analysis**

Based on the problem statement, here's a breakdown of the required steps:

**1. Initial Data Inspection and Cleaning 🧹**

* **Examine the data** for duplicates, missing values, irrelevant entries, or outliers.
* **Identify and correct any erroneous entries**.
* **Refine the data** for further processing based on your findings.

**2. Exploratory Data Analysis (EDA) and Feature Engineering 📈**

* **Use visualizations** to determine which two albums should be recommended based on the number of popular songs they contain.
* **Perform EDA** to explore the various features of the songs and identify patterns.
* **Analyze the relationship** between a song's popularity and other factors, and see how this correlation has changed over time.
* **Consider dimensionality reduction techniques**. This is often used to simplify models and visualize high-dimensional data. You would need to provide insights on the significance of these techniques and your observations.

**3. Cluster Analysis 🧠**

* **Identify the right number of clusters** for the data. This could involve methods like the elbow method or silhouette analysis.
* **Apply appropriate clustering algorithms** to group the songs. K-means, DBSCAN, or hierarchical clustering are common choices.
* **Define and interpret each cluster** based on the features that characterize it. For example, a cluster might be defined by high

danceability and energy values, suggesting it contains up-tempo, party-like songs.