Creating Cohorts of Songs

Problem Scenario:

The customer always looks forward to specialized treatment, whether shopping on an e-commerce website or watching Netflix. The customer desires content that aligns with their preferences. To maintain customer engagement, companies must consistently provide the most relevant information.

Starting with Spotify, a Swedish audio streaming and media service provider, boasting over 456 million active monthly users (including more than 195 million paid subscribers as of September 2022), the company aims to create cohorts of different songs to enhance song recommendations. These cohorts will be based on various relevant features, ensuring that each group contains similar types of songs.

Problem Objective:

As a data scientist, you should perform exploratory data analysis and cluster analysis to create cohorts of songs. The goal is to better understand the various factors that create a cohort of songs.

Data Description:

The dataset comprises information from Spotify's API regarding all albums by the Rolling Stones available on Spotify. It's crucial to highlight that each song possesses a unique ID.

Question 1: Initial Data Inspection and Cleaning

The dataset contains 1,610 entries with no missing or duplicate values. All data types are appropriate. The release_date column can be converted to datetime for time-series analysis if needed.

Question 2: Data Refinement

Outliers:

- 'track_number' goes up to 47 (likely multi-disc albums).
- 'duration_ms' has a max over 900,000 (likely long live tracks).

Recommendations:

- Consider filtering long tracks if only analyzing typical studio songs.

- Convert release_date to datetime if temporal patterns are relevant.

Question 3a: Most Popular Albums

Top albums by count of songs with popularity >= 50:

1. Sticky Fingers (Remastered)

2. Exile On Main Street (2010 Re-Mastered)

3. Let It Bleed

These albums are strong candidates for recommendation based on song popularity.

Question 3b-3c: Feature Correlation Analysis

Most features show weak correlation with popularity. However:

- Danceability, acousticness, and loudness show slight positive correlation.

- Liveness, energy, and instrumentalness show slight negative correlation.

Strong internal correlations:

- Energy and loudness (0.70)

- Valence and danceability (0.55)

- Energy and liveness (0.51)

Question 3d: Dimensionality Reduction with PCA

PCA was used to reduce the feature space to two dimensions. The resulting plot showed discernible groupings, suggesting that clustering is appropriate.

Question 4a: Optimal Number of Clusters

Silhouette score analysis showed the best clustering structure at 2 clusters.

Question 4b-4c: Cluster Descriptions

Cluster 0:

- Higher popularity

- More danceable, acoustic, and positive
- Lower energy and speechiness
- Shorter duration

Cluster 1:

- Lower popularity
- Higher energy, liveness, and speechiness
- Longer and louder tracks

Interpretation:

- Cluster 0 likely includes upbeat studio tracks.
- Cluster 1 likely includes live performances or experimental content.