Creating Cohorts of Songs
Project Write-Up

Creating Cohorts of Songs Using Spotify Data

1. Introduction

In today's digital world, customers expect personalized experiences while using platforms such as Netflix or Spotify. Spotify, a leading audio streaming platform, aims to recommend songs that match user preferences. One effective way to achieve this is by grouping similar songs into cohorts based on their musical features. These cohorts help improve recommendation systems and enhance user engagement.

2. Objective of the Project

The main objective of this project is to:

- Perform Exploratory Data Analysis (EDA) on Spotify song data

- Identify important musical features influencing song similarity

- Apply cluster analysis to group similar songs into meaningful cohorts

- Understand how different song attributes contribute to clustering

3. Dataset Description -:

The dataset contains information about Rolling Stones songs collected from Spotify's API. Each song is uniquely identified and described using several musical attributes such as energy, danceability, tempo, loudness, and popularity.

Key features include:

- Acousticness

- Danceability

- Energy

- Tempo

- Loudness

- Valence

- Popularity

- Duration

4. Data Inspection and Cleaning -:

Initial data inspection showed that:

- There were no missing values

- No duplicate records were found

- All feature values were within valid ranges

- The dataset was clean and suitable for further analysis after removing irrelevant identifier columns.

5. Exploratory Data Analysis (EDA)

EDA was performed to understand patterns in song features:

- Albums were compared based on average song popularity to identify which   albums are most recommendable.

- Analysis showed that high-energy and danceable songs tend to be more popular.

- Acoustic songs generally had lower popularity compared to energetic tracks.

- Song popularity trends were also observed over time.

- These insights help understand user preferences and listening behavior.

6. Dimensionality Reduction -:

Since the dataset contains many correlated features, Principal Component Analysis (PCA) was used:

- ❖ To reduce dimensionality

- ❖ To remove noise

- ❖ To improve clustering performance

- ❖ To enable easy visualization of song clusters

- ❖ PCA helped summarize the most important information using fewer dimensions.

7. Cluster Analysis -:

To create song cohorts:

- Data was standardized using feature scaling

- The Elbow Method and Silhouette Score were used to determine the optimal number of clusters

- K-Means clustering was applied

- The analysis identified four distinct clusters of songs.

8. Cluster Interpretation :-

Each cluster represents a unique type of song:

- Cluster 1 – High Energy Rock Songs: Loud, energetic, fast-paced tracks

- Cluster 2 – Acoustic & Calm Songs: Soft, emotional, low-energy tracks

- Cluster 3 – Danceable & Popular Songs: High danceability and positive mood

- Cluster 4 – Live Performance Songs: Songs recorded during live performances

9. Conclusion

This project successfully created meaningful cohorts of songs using exploratory data analysis and clustering techniques. The insights gained can help Spotify:

Improve song recommendations

Personalize playlists

Enhance user experience

Overall, clustering songs based on musical features is an effective approach for building intelligent recommendation systems.