Machine Learning

Course-End Project Problem Statement

Course-End Project: Creating Cohorts of Songs

Problem Scenario: The customer always looks forward to specialized treatment, whether shopping over an e-commerce website or watching Netflix. They want what they might like to see. To keep the customers engaged, it is also crucial for companies to always present the most relevant information. Spotify is a Swedish audio streaming and media service provider. The company has over 456 million active monthly users, including over 195 million paying subscribers, as of September 2022. The company intends to create cohorts of different songs that will aid in the recommendation of songs to users based on various relevant features. Each cohort would contain similar types of songs.

Problem Objective:

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

Note: Download ***Data Dictionary – Creating cohorts of songs.xlsx*** from the course resource section in the LMS.

Data Description:
This dataset contains data from Spotify's API about all albums for the Rolling Stones listed on Spotify. It is important to note that all songs have unique IDs.

Variable	Description
name	It is the name of the song.
album	It is the name of the album.
release_date	It is the day, month, and year the album was released.
track number	It is the order the song appears on the album.
id	It is the Spotify id for the song.
uri	It is the Spotify URI for the song.
acousticness	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0

	represents high confidence the track is acoustic.
danceability	It describes how suitable a track is for dancing based on a combination of musical elements, including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is the least danceable, and 1.0 is the most danceable.
energy	It is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
instrumentalness	It predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
liveness	It detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides a strong likelihood that the track is live.
loudness	The overall loudness of a track in decibels (dB) and loudness values are averaged across the entire track and

	are useful for comparing the relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 dB.
speechiness	It detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g., talk show, audiobook, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
tempo	The overall estimated tempo of a track is in beats per minute (BPM), and in musical terminology, the tempo is the speed or pace of a given piece and derives directly from the average beat duration.
valence	A measure from 0.0 to 1.0 describes the musical positiveness conveyed by a track, and tracks with high valence sound more positive (e.g., happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g., sad, depressed, angry).
popularity	It is the popularity of the song from 0 to 100.
duration_ms	It is the duration of the track in milliseconds.

Steps to Perform:

- Initial data inspection and data cleaning:

    - Check whether the data has duplicates, missing values, irrelevant (erroneous entries) values, or outliers.

- Depending on your findings, clean the data for further processing.

- Perform Exploratory Data Analysis and Feature Engineering:

    - Use appropriate visualizations to find out which two albums should be recommended to anyone based on the number of popular songs in an album.
    - Perform exploratory data analysis to dive deeper into different features of songs and identify the pattern.
    - Discover how a song's popularity relates to various factors and how this has changed over time.
    - Comment on the importance of dimensionality reduction techniques, share your ideas and explain your observations.

- Perform Cluster Analysis:
    - Identify the right number of clusters
    - Use appropriate clustering algorithm
    - Define each cluster based on the features