Project Proposal

Project Overview

Music streaming platforms rely heavily on data-driven insights to understand listener preferences, track performance, and relationships between musical attributes. With the growth of large-scale music datasets, there is increasing opportunity to analyse how measurable audio features relate to engagement outcomes such as popularity and musical characteristics (like valence and energy).

This project explores a large Spotify track dataset to investigate whether quantifiable audio features and genre classifications can act as indicators of popularity and musical characteristics. In addition, the project explores how these insights can support the foundations of a simple recommendation system.

The analysis combines exploratory data analysis, statistical investigation, and introductory machine learning techniques to derive interpretable insights from music streaming data.

Dataset Description

The dataset used in this project is the Spotify tracks dataset sourced from Kaggle. The original dataset contains **114,000 entries,** and **21 columns**. The tracks span across **125 distinct genres**, with each track associated with a range of metadata and audio features.

1. **Track_id** - Unique identifier assigned to each track by Spotify.
2. **Track_name** - The name of the song as listed on Spotify.
3. **Artists** - A string containing the name(s) of the artist(s) associated with the track. Multiple artists may be listed for collaborative tracks.
4. **Album_name** - The name of the album on which the track appears.
5. **Popularity** - A numerical score ranging from 0 to 100 indicating the track’s popularity on Spotify, based on streaming activity and engagement.
6. **Duration_ms** - The length of the track measured in milliseconds.
7. **Explicit** - A binary indicator specifying whether the track contains explicit content.
8. **Danceability** - A measure of how suitable a track is for dancing, based on tempo, rhythm stability, and beat strength.
9. **Energy -** A measure representing the intensity and activity of a track, with higher values indicating louder and more dynamic music.

10. **Key -** The musical key in which the track is composed, represented as an integer using standard pitch class notation (e.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on). A value of −1 indicates that no key was detected for the track
11. **Loudness -** The overall loudness of a track measured in decibels (dB).
12. **Mode -** The musical modality of the track, where 1 represents major and 0 represents minor.
13. **Speechiness -** A measure of the presence of spoken words within a track.
14. **Acousticness -** A confidence measure of whether a track is acoustic in nature.
15. **Instrumentalness -** A measure predicting whether a track contains no vocals.
16. **Liveness -** A measure indicating the likelihood that the track was recorded during a live performance.
17. **Valence -** A measure describing the musical positiveness conveyed by a track, with higher values representing happier or more upbeat music.
18. **Tempo -** The estimated speed of the track measured in beats per minute (BPM).
19. **Time_signature -** The estimated time signature of the track, representing the number of beats per bar.
20. **Track_genre** - The primary genre assigned to the track by Spotify.

Project Aim and Objectives

Aim

The primary aim of this project is to analyse whether audio features and genre classifications can be used to explain or predict track popularity and musical characteristics, and to explore how these features can support basic music recommendation logic.

Objectives

- To clean and prepare a large-scale music dataset for reliable analysis
- To explore relationships between genre and engagement metrics such as popularity
- To examine whether specific audio features act as indicators of other musical attributes (e.g. danceability)
- To assess correlations between emotional tone (valence), tempo, loudness, and danceability
- To experiment with regression and machine learning techniques for modelling musical features
- To lay groundwork for a simple recommendation approach based on audio similarity

Research Questions

Based on initial exploration of the dataset, the following research questions guide the project:

Research Questions

1. Is musical genre an indicator of track popularity on Spotify?
    - **Hypothesis:** Tracks belonging to certain genres have significantly higher average popularity scores than others.
2. Do certain audio features (e.g. loudness, energy, tempo) show strong relationships with musical characteristics (e.g. valence and danceability)?
    - **Hypothesis**: Tempo is positively correlated with danceability.
3. Do collaborative tracks (tracks with multiple artists) differ in popularity or behaviour compared to solo tracks?
    - **Hypothesis:** Tracks involving multiple artists demonstrate different popularity patterns compared to solo-artist tracks.

NOTE:

For the sake of clarity and consistency within this analysis, some of the variables derived from the dataset are grouped into two categories: **audio features** and **musical characteristics**.

Audio features - Variables that describe the technical and signal-level properties of a track. Things that are less subjective. In this project, they are the following variables:

- Energy
- Loudness
- Tempo
- Speechiness
- Instrumentalness

Musical characteristics - Variables that describe the perceptual qualities of a track, particularly in relation to emotional tone and movement. In this project, they are the following variables:

- Danceability
- Valence

- Acousticness

Tools and Technologies

The project uses the following tools and technologies:

- **Python** for data analysis and modelling
- **Pandas & NumPy** for data cleaning, manipulation, and feature engineering
- **Matplotlib & Seaborn** for visualisation (histograms, box plots, bar charts, heatmaps)
- **Scikit-learn** for regression modelling and evaluation
- **TensorFlow / Keras** for exploratory neural network modelling and embeddings
- **Jupyter Notebooks** for reproducible analysis and documentation
- **Tableau/ Power BI** for dashboarding
- **Streamlit** for dashboarding and presentation

Expected Deliverables

By the end of the project, the following deliverables are expected:

- A fully cleaned dataset
- Exploratory visualisations illustrating relationships between genres, audio features, and popularity
- An exploratory machine learning model demonstrating recommendation potential.
- A concise analytical dashboard showcasing key insights
- A recommendation tool based on users input.
- A presentation documenting key findings, limitations, and conclusions