Project title : `spotify_songs`

`Domain` : Music / Streaming Analytics

Project Goal: Analyze song performance, compare songs by genre, language, duration, and explicit content, and extract valuable insights to support decision-making in the music industry.

Project Description :

Data sourced from **Kaggle**, containing information about songs including:

- **Song Attributes:** `[song_title]`, `[artist]`, `[album]`, `[genre]`, `[language]`, `[duration]`
- **Performance Metrics:** `[popularity]`, `[stream]`
- **Classifications:** `[explicit_content]`, `[popularity_level]`, `[streams_level]`, `[duration_minute]`, `[date_group]`
- **Production Data:** `[composer]`, `[producer]`

Process :

- **Questions**

1. Which songs have the highest popularity scores?
2. Which songs have the highest number of streams?
3. How many songs of each genre fall into high, medium, and low popularity levels?
4. What is the distribution of songs by popularity levels per genre, and what is the average streams per genre?
5. Are explicit songs more popular than non-explicit songs?
6. Which years had the highest number of new song releases?
7. Do newer songs get more streams than older or middle-period songs?
8. Which languages are most commonly used in songs?
9. Which languages have songs with the highest average popularity?
10. How does song duration (short, normal, long) affect average popularity?
11. How does song duration (short, normal, long) affect average streams?
12. How are songs distributed across duration levels and popularity levels?
13. How are songs distributed across duration levels and streams levels?
14. How many explicit vs non-explicit songs exist in each streams level (Low, Medium, High)?
15. How many explicit vs non-explicit songs exist in each date group (new, middle, old)?

16. How are song durations distributed across release periods (new, middle, old)?
17. Does the combination of song duration and explicit content affect popularity?
18. How do streams vary across different genres and release periods?
19. How are explicit vs non-explicit songs distributed across popularity levels?
20. How do average streams vary by language across different popularity levels?
21. Do songs from major labels (Universal, Sony, Warner, Def Jam) get more streams than indie labels

- # Data Cleaning
  - No duplication
  - Deal with outliers
    - There is outliers in [`duration] and its count 349` it will be deleted
  - Deal with nulls
    - Filling nulls in [language] by mode
    - Filling nulls in [duration] by mean
    - Delete [collaboration] Because it contains many nulls = 35000
  - Feature Engineering
    - Create columns like
      - popularity_level
      - duration_minute
      - streams_level
      - date_group
    - the columns that created will Facilitate analysis and comparison across different groups

- # DATA EXPLORATION

  - **Univariate Analysis:**
    Studied each column separately. For numeric features (`popularity`, `stream`), we used **histograms** and **boxplots** to check distributions and outliers. For categorical features (`genre`, `language`, `duration_minute`, `explicit_content`, `date_group`, `popularity_level`, `streams_level`), we used **bar charts** and **frequency tables** to understand counts and proportions.

- **Bivariate Analysis:**
  Examined relationships between two variables, such as **Streams vs Popularity**, **Genre vs Popularity Level**, or **Explicit Content vs Streams**, using **scatterplots**, **boxplots**, and **cross-tabulations**.
- **Multivariate Analysis:**
  Analyzed combined effects of multiple factors (e.g., **Genre + Duration + Explicit Content**) to see their impact on streams and popularity.
- **Target Analysis:**
  Compared distributions of all features across different **performance levels** (`popularity_level`, `streams_level`) to identify which attributes are associated with higher-performing songs.

- ## Extracting Insights

  After exploring the Spotify dataset, several key insights were identified that:

  - **Highlight the main factors associated with higher song performance**, such as high popularity, high streams, explicit content, certain genres, and song duration.
  - **Compare different groups** (New vs Middle vs Old releases, Short vs Normal vs Long songs, Explicit vs Non-Explicit) to observe their impact on popularity and streaming numbers.
  - **Study the interaction of multiple factors together** (e.g., Genre + Duration + Explicit Content) and their combined relation to streams and popularity.
  - **Identify general patterns that differentiate top-performing songs from less popular ones**, including trends across languages, streaming levels, and release periods.

- ## Visualization :

  Data visualization was applied to better understand the dataset

  • **Numeric features:** Histograms and boxplots to study distributions and detect outliers.

  • **Categorical features:** Bar charts to show frequency counts.

  • **Relationships:**

Scatterplots, heatmaps, and crosstabs to explore variable relationships.

• **Target analysis:** Compared patients with and without heart disease to identify key influencing factors.

• Tools

Excel

Python (pandas,numpy,matplotlib)

Sql server