🎵 Spotify SQL Data Analysis Project Report

Author: Prince Kashav
Database: SPOTIFY_db
Tools Used: PostgreSQL

---

1. Project Overview

The purpose of this SQL project is to analyze a Spotify dataset containing information about songs, artists, albums, and engagement metrics (streams, views, likes, comments).
The project focuses on **data exploration, cleaning, and deriving business insights** from the streaming data using SQL.

This analysis helps understand **which artists, tracks, and albums perform the best**, and how various musical attributes (like energy, danceability, or liveness) influence performance across platforms like Spotify and YouTube.

---

2. Database Design

Table: spotify

Column	Data Type	Description
artist	VARCHAR(255)	Name of the artist
track	VARCHAR(255)	Track title
album	VARCHAR(255)	Album name
album_type	VARCHAR(255)	Type of album (e.g., single, album)
danceability	FLOAT	Measure of how danceable the track is
energy	FLOAT	Energy level of the track
loudness	FLOAT	Overall loudness
speechiness	FLOAT	Amount of spoken words in the track
acousticness	FLOAT	Measure of how acoustic the track is
instrumentalness	FLOAT	Probability that the track is instrumental
liveness	FLOAT	Likelihood the track was performed live
valence	FLOAT	Musical positivity of the track
tempo	FLOAT	Beats per minute (BPM)
duration_min	FLOAT	Duration in minutes

Column	Data Type	Description
title	VARCHAR(255)	Video title
channel	VARCHAR(255)	YouTube channel name
views	FLOAT	Total number of views
likes	BIGINT	Number of likes
comments	BIGINT	Number of comments
licensed	BOOLEAN	Whether the track is licensed
official_video	BOOLEAN	Indicates if it's an official video
stream	BIGINT	Number of Spotify streams
energy_liveness	FLOAT	Combined metric for analysis
most_playedon	VARCHAR(255)	Platform with most streams (Spotify / YouTube)

---

3. Data Cleaning

To ensure data accuracy and reliability:

- Checked for missing or invalid durations:
- SELECT * FROM spotify WHERE duration_min = 0;
- DELETE FROM spotify WHERE duration_min = 0;
- Verified total rows before and after cleaning.
- Ensured all metrics are numeric and valid.

---

4. Exploratory Data Analysis (EDA)

Analysis	Description	Query Reference
Total rows	Checked total number of records	Q1
Unique artists and albums	Counted distinct entries	Q2 – Q4
Duration analysis	Found longest and shortest songs	Q5 – Q6
Data cleanup	Removed invalid durations	Q7 – Q8

---

5. Business Insights

5.1 High-Level Analysis

- **Top streamed tracks**: Identified tracks with more than **1 billion streams**.
- **Top albums by streams**: Ranked albums based on **total cumulative streams**.
- **Licensed content performance**: Calculated total comments for all licensed tracks.
- **Album type breakdown**: Found number of tracks released as *singles*.

5.2 Artist-Level Insights

- Counted number of unique tracks per artist.
- Identified **most productive artists** and those with **most viewed or streamed songs**.
- Used **window functions** to get **top 3 most viewed tracks per artist**.

---

6. Analytical Insights

6.1 Musical Attributes

- Computed **average danceability per album**.
- Ranked **top 10 tracks by energy** levels.
- Found **energy difference (max–min)** per album to analyze variation.

6.2 Platform Comparison

- Compared **YouTube vs Spotify** performance:
- HAVING SUM(CASE WHEN most_playedon = 'Youtube' THEN stream END) >
- SUM(CASE WHEN most_playedon = 'Spotify' THEN stream END);
- Highlighted tracks performing better on YouTube.

6.3 Advanced Analytical Queries

- Found tracks where **liveness > average liveness**, indicating potential live performances.
- Calculated **energy-to-liveness ratio** for identifying energetic live tracks.
- Calculated **cumulative likes** ordered by views using window functions.

---

7. Key Insights and Findings

✅ The dataset provides a holistic view of artist and track performance across platforms.
✅ High energy and danceability correlate with higher views and streams.
✅ Licensed and official videos receive significantly higher engagement.
✅ YouTube often outperforms Spotify in terms of views for visual content.
✅ Certain albums show wide energy variation, indicating stylistic diversity.
✅ Cumulative engagement metrics help identify audience interaction trends.

---

8. Conclusion

This project demonstrates the power of **SQL for data cleaning, exploration, and analytics**. By applying aggregation, window functions, and conditional logic, we derived valuable business insights from Spotify's music data.

The results can help:

- **Artists** identify which tracks resonate most.
- **Labels** allocate promotional resources.
- **Streaming platforms** improve recommendation systems.

---

9. Future Improvements

- Integrate additional metadata like release year or genre.
- Visualize query outputs using Power BI or Tableau.
- Automate updates using scheduled database scripts.

---

10. Files Included in GitHub Repository

File Name	Description
Spotify_SQL_Project.sql	Contains all SQL queries and setup commands
Spotify_SQL_Project_Report.docx	This analytical report with explanations and insights