MANI CHOKKARA
DATA ANALYST

Netflix Movies and TV Shows Data Analysis using SQL

1. Project Title & Problem Statement

Title: Netflix Movies and TV Shows Data Analysis using SQL

Problem Statement:
The aim of this project is to analyze Netflix's catalog of movies and TV shows using SQL to uncover meaningful patterns and insights that support business decision-making. The dataset contains detailed information about content type, ratings, release year, duration, countries, and categories. The key focus is to understand how Netflix structures its content library, identify viewer-oriented patterns, and answer critical business questions such as:

- What is the distribution between movies and TV shows on the platform?
- Which ratings are most common for different content types?
- How has Netflix's content evolved over the years and across different countries?
- What categories, durations, and keywords appear most frequently in the catalog?

2. Dataset Collection

The data for this project is sourced from the Kaggle dataset:

Dataset Link: Dataset

3. Data Understanding & Documentation

Data Understanding

1. Column Summary

Basic details about each field in the Netflix dataset:

- **show_id** – Unique identifier
- **type** – Movie or TV Show
- **title** – Name of the content
- **director** – Director(s) of the title
- **cast** – List of actors
- **country** – Country of production
- **date_added** – When content was added to Netflix
- **release_year** – Original release year
- **rating** – Maturity rating (e.g., TV-MA, PG-13)
- **duration** – Movie length (minutes) or TV show seasons

2. Data Dictionary

- **type:** Categorical variable for content classification.
- **rating:** Indicates audience suitability.
- **duration:** Requires separation into numeric duration + unit.
- **director, cast, country:** Multi-value text fields with missing values.
- **date_added:** To be converted into proper date format.

3. Early Observations

- Movies dominate the dataset compared to TV shows.
- **TV-MA** is the most frequent rating.
- Significant missing values in *director*, *cast*, and *country*.
- Most content appears after **2015**, aligning with Netflix's global expansion.
- USA and India are leading content-producing countries.

4. Missing Values & Outliers

- **High missing**: director
- **Medium missing**: cast, country
- **Low missing**: date_added
- Outliers include very old release years and unusual durations.

5. KPI Requirements

- Movies vs TV Shows count
- Rating distribution
- Top countries producing Netflix content
- Titles added by year/month
- Most common actors/directors
- Duration analysis (shortest/longest titles)

6. Assumptions

- Missing values treated as "Unknown".
- The first country listed is considered primary.
- Duration always contains a numeric value and unit.
- Ratings assumed accurate.

4. Data Cleaning & Preprocessing

Tools: Python, Excel Power Query

Common preprocessing steps:
✔ Remove duplicates
✔ Treat missing values
✔ Convert data types
✔ Standardize date-time
✔ Create useful columns (feature engineering)
✔ Encode categories if needed

5. Insights & Recommendations

1. Content Distribution

Analysis revealed that movies constitute a larger portion of the Netflix catalog compared to TV shows. This indicates Netflix's strategic focus on maintaining an extensive movie library to cater to users preferring shorter, standalone content formats.

2. Dominant Ratings

The most frequently occurring ratings across both movies and TV shows were TV-MA, TV-14, and R. This trend reflects a platform leaning towards mature and adult-oriented content, suggesting that Netflix targets a predominantly adult audience demographic.

3. Country-wise Contribution

Country analysis indicated that the United States produces the highest number of titles available on Netflix.
Other strong contributors include India, the United Kingdom, and Canada, showing Netflix's global content acquisition strategy with a significant focus on English-language markets.

4. Trends in Indian Content Production

A year-wise breakdown of Indian content showed notable increases in recent years, particularly after 2018.
This highlights Netflix's growing investment in regional and local Indian productions, aligning with its strategy to expand in emerging markets.

5. Duration-Based Insights

By converting and analyzing the duration field:

- The platform hosts movies with widely varying lengths.
- Several TV shows have more than 5 seasons, indicating strong viewer engagement and long-running franchises.

6. Genre and Keyword Behaviors

Genre analysis showed heavy representation in:

- Documentaries
- International content
- Drama
- Comedy

Keyword classification (based on terms like kill and violence) revealed a sizable portion of titles with intense or action-driven narratives.

7. Actor and Director Analysis

Actor-based insights highlighted recurring appearances of prominent figures, such as those in the Indian film industry.
Similarly, director-based analysis (e.g., identifying works of Rajiv Chilaka) allowed mapping of creator-specific contributions.

8. Data Quality and Structure

The dataset contained several NULL or missing values across key columns. After cleaning, the dataset became consistent and valid for analytical purposes. This highlights the importance of preprocessing before deriving insights.