**Project 02:- To perform data cleaning and analysis for Netflix using SQL.  
  
Introduction :-** The objective of this project is to perform data cleaning and analysis on the Netflix dataset using SQL. Netflix, a leading global streaming platform, possesses a vast amount of data related to its shows. However, to derive valuable insights for business decision-making, this dataset requires thorough cleaning and analysis.

Data cleaning involves handling missing values, inconsistencies, and ensuring data integrity. By addressing these issues, we can ensure the accuracy and reliability of the dataset, enabling us to draw meaningful conclusions and make informed decisions.

**Query for MySQL :**

**Segment 1:**

CREATE database netflix\_dataset;

CREATE TABLE netflix\_dataset (

show\_id INT PRIMARY KEY,

type VARCHAR(20),

title VARCHAR(255),

director VARCHAR(255),

country VARCHAR(255),

date\_added varchar(10),

release\_year INT,

rating VARCHAR(10),

duration VARCHAR(20),

listed\_in VARCHAR(255)

);

DESCRIBE netflix\_dataset;

SELECT COUNT(\*) FROM netflix\_dataset;

SELECT \* FROM netflix\_dataset WHERE rating IS NULL;

**Segment 2:**

SELECT type, COUNT(\*) AS count

FROM netflix\_dataset

GROUP BY type;

SELECT country, COUNT(\*) AS count FROM netflix\_dataset GROUP BY country ORDER BY count DESC LIMIT 10;

SELECT YEAR(date\_added) AS year, COUNT(\*) AS count FROM netflix\_dataset GROUP BY year ORDER BY year;

SELECT YEAR(release\_year) AS year, AVG(duration) AS average\_duration FROM netflix\_dataset GROUP BY year ORDER BY year;

SELECT director, COUNT(\*) AS count FROM netflix\_dataset GROUP BY director ORDER BY count DESC LIMIT 10;

**Segment 3:**

SELECT DISTINCT type

FROM netflix\_dataset;

SELECT type,

COUNT(CASE WHEN type = 'Movie' THEN 1 END) AS movie\_count,

COUNT(CASE WHEN type = 'TV Show' THEN 1 END) AS tvshow\_count,

COUNT(\*) AS total\_count,

ROUND(COUNT(CASE WHEN type = 'Movie' THEN 1 END) \* 100.0 / COUNT(\*), 2) AS movie\_percentage,

ROUND(COUNT(CASE WHEN type = 'TV Show' THEN 1 END) \* 100.0 / COUNT(\*), 2) AS tvshow\_percentage

FROM netflix\_dataset

GROUP BY type;

SELECT type, COUNT(\*) AS count

FROM netflix\_dataset

GROUP BY type

ORDER BY count DESC

LIMIT 5;

SELECT type, SUM(duration) AS total\_duration

FROM netflix\_dataset

GROUP BY type

ORDER BY total\_duration DESC;

**Segment 4:**

SELECT EXTRACT(MONTH FROM date\_added) AS month, EXTRACT(YEAR FROM date\_added) AS year, COUNT(\*) AS count

FROM netflix\_dataset

GROUP BY month, year

ORDER BY year, month;

SELECT CASE WHEN EXTRACT(MONTH FROM date\_added) IN (12, 1, 2) THEN 'Winter'

WHEN EXTRACT(MONTH FROM date\_added) IN (3, 4, 5) THEN 'Spring'

WHEN EXTRACT(MONTH FROM date\_added) IN (6, 7, 8) THEN 'Summer'

WHEN EXTRACT(MONTH FROM date\_added) IN (9, 10, 11) THEN 'Fall'

ELSE 'Unknown'

END AS season, COUNT(\*) AS count

FROM netflix\_dataset

GROUP BY season;

SELECT EXTRACT(MONTH FROM date\_added) AS month, EXTRACT(YEAR FROM date\_added) AS year, COUNT(\*) AS count

FROM netflix\_dataset

GROUP BY month, year

ORDER BY count DESC

LIMIT 1;

**Segment 5:**

SELECT type, rating, COUNT(\*) AS count

FROM netflix\_dataset

GROUP BY type, rating

ORDER BY type, rating;

SELECT rating, AVG(duration) AS avg\_duration

FROM netflix\_dataset

GROUP BY rating

ORDER BY rating;

**Segment 6:**

SELECT t1.type, t2.type, COUNT(\*) AS count

FROM netflix\_dataset t1

JOIN netflix\_dataset t2 ON t1.show\_id != t2.show\_id

WHERE t1.type < t2.type

GROUP BY t1.type, t2.type

ORDER BY count DESC

LIMIT 5;

SELECT type, AVG(duration) AS avg\_duration

FROM netflix\_dataset

GROUP BY type

ORDER BY avg\_duration DESC;

**Segment 7:**

SELECT DISTINCT country

FROM netflix\_dataset;

SELECT country, type, COUNT(\*) AS count

FROM netflix\_dataset

GROUP BY country, type;

SELECT country, AVG(duration) AS avg\_duration

FROM netflix\_dataset

GROUP BY country

ORDER BY avg\_duration DESC;

**Segment 8**: Recommendations for Content Strategy

Based on the analysis, Netflix should focus on producing a diverse range of content that caters to various genres and categories. This will help capture a wider audience and keep viewers engaged.

Regularly monitoring and analyzing viewer feedback, ratings, and preferences can inform content decisions and help identify emerging trends or genres that are gaining popularity.

Expanding into new markets and exploring partnerships with international production companies can provide opportunities for growth and audience expansion.

Investing in original content creation, including movies and TV shows, can differentiate Netflix from competitors and foster viewer loyalty.

**Objective :**

The objective of the project is to perform data cleaning and analysis on the Netflix dataset to gain valuable insights for business decision-making. The project aims to achieve the following objectives:

1. Data Cleaning: Ensure the dataset is accurate, consistent, and free from errors or missing values. By cleaning the data, we can enhance the quality of the analysis and avoid misleading results.
2. Content Analysis: Analyze the distribution of content types (movies vs. TV shows), identify top countries with the highest number of productions, investigate the trend of content additions over the years, analyze the relationship between content duration and release year, and identify directors with the most content. These analyses provide an understanding of the content offerings on Netflix and help identify trends and patterns.
3. Genre and Category Analysis: Determine the unique genres and categories present in the dataset, calculate the percentage of movies and TV shows in each genre, identify the most popular genres/categories based on the number of productions, and calculate the cumulative sum of content duration within each genre. This analysis helps in understanding viewers' preferences and can aid in content curation and production decisions.
4. Release Date Analysis: Determine the distribution of content releases by month and year, analyze seasonal patterns in content releases, and identify months and years with the highest number of releases. This analysis can inform content scheduling and marketing strategies.
5. Rating Analysis: Investigate the distribution of ratings across different genres and analyze the relationship between ratings and content duration. This analysis can help understand viewer sentiment and preferences.
6. Co-occurrence Analysis: Identify the most common pairs of genres/categories that occur together in content and analyze the relationship between genres/categories and content duration. This analysis can provide insights into content combinations that are well-received by viewers.
7. International Expansion Analysis: Identify the countries where Netflix has expanded its content offerings, analyze the distribution of content types in different countries, and investigate the relationship between content duration and country of production. This analysis can inform international content strategies and localization efforts.
8. Recommendations for Content Strategy: Based on the analysis, provide recommendations for the types of content Netflix should focus on producing. Identify potential areas for expansion and growth based on the analysis of the dataset. These recommendations aim to support strategic decision-making and content planning to enhance viewer satisfaction and grow the platform's user base.

Overall, the project's goal is to leverage SQL expertise to clean and analyze the Netflix dataset, derive meaningful insights, and make data-driven recommendations to support Netflix's content strategy and business success.

Netflix, a leading global streaming platform, possesses a dataset containing information about its shows. However, the dataset requires cleaning and analysis to derive valuable insights for business decision-making. As a data analyst with SQL expertise, your objective is to perform data cleaning and analysis on the Netflix dataset to help the company gain insights into their content offerings.

Segment 1: Database - Tables, Columns, Relationships

* Identify the tables in the dataset and their respective columns.
* Determine the number of rows in each table within the schema.
* Identify and handle any missing values in the dataset.

Segment 2: Content Analysis

* Analyse the distribution of content types (movies vs. TV shows) in the dataset.
* Determine the top 10 countries with the highest number of productions on Netflix.
* Investigate the trend of content additions over the years.
* Analyse the relationship between content duration and release year.
* Identify the directors with the most content on Netflix.

Segment 3: Genre and Category Analysis

* Determine the unique genres and categories present in the dataset.
* Calculate the percentage of movies and TV shows in each genre.
* Identify the most popular genres/categories based on the number of productions.
* Calculate the cumulative sum of content duration within each genre.

Segment 4: Release Date Analysis

* Determine the distribution of content releases by month and year.
* Analyse the seasonal patterns in content releases.
* Identify the months and years with the highest number of releases.

Segment 5: Rating Analysis

* Investigate the distribution of ratings across different genres.
* Analyse the relationship between ratings and content duration.

Segment 6: Co-occurrence Analysis

* Identify the most common pairs of genres/categories that occur together in content.
* Analyse the relationship between genres/categories and content duration.

Segment 7: International Expansion Analysis

* Identify the countries where Netflix has expanded its content offerings.
* Analyse the distribution of content types in different countries.
* Investigate the relationship between content duration and country of production.

Segment 8: Recommendations for Content Strategy

* Based on the analysis, provide recommendations for the types of content Netflix should focus on producing.
* Identify potential areas for expansion and growth based on the analysis of the dataset.

Evaluation pointers:

- The tasks are correctly identified and executed.

- The solution output matches the expected output.

- The query is optimised and syntactically correct.

- Proper aliases are used

- If required any, appropriate comments are written.

- The code is written concisely with appropriate indentations.