# **Netflix Movies and TV Shows Data Analysis using SQL**

****

## **Overview**

This project involves a comprehensive analysis of Netflix's movies and TV shows data using SQL. The goal is to extract valuable insights and answer various business questions based on the dataset. The following README provides a detailed account of the project's objectives, business problems, solutions, findings, and conclusions.

## **Objectives**

* Analyze the distribution of content types (movies vs TV shows).
* Identify the most common ratings for movies and TV shows.
* List and analyze content based on release years, countries, and durations.
* Explore and categorize content based on specific criteria and keywords.

## **Dataset**

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

* Dataset Link: [Movies Dataset](https://www.kaggle.com/datasets/shivamb/netflix-shows?resource=download)

## 

## 

## **Schema**

DROP TABLE IF EXISTS netflix;

CREATE TABLE netflix

(

show\_id VARCHAR(5),

type VARCHAR(10),

title VARCHAR(250),

director VARCHAR(550),

casts VARCHAR(1050),

country VARCHAR(550),

date\_added VARCHAR(55),

release\_year INT,

rating VARCHAR(15),

duration VARCHAR(15),

listed\_in VARCHAR(250),

description VARCHAR(550)

);

## **Business Problems and Solutions**

### **1. Count the Number of Movies vs TV Shows**

SELECT

type,

COUNT(\*)

FROM netflix

GROUP BY 1;

Objective: Determine the distribution of content types on Netflix.

### **2. Find the Most Common Rating for Movies and TV Shows**

WITH RatingCounts AS (

SELECT

type,

rating,

COUNT(\*) AS rating\_count

FROM netflix

GROUP BY type, rating

),

RankedRatings AS (

SELECT

type,

rating,

rating\_count,

RANK() OVER (PARTITION BY type ORDER BY rating\_count DESC) AS rank

FROM RatingCounts

)

SELECT

type,

rating AS most\_frequent\_rating

FROM RankedRatings

WHERE rank = 1;

Objective: Identify the most frequently occurring rating for each type of content.

### **3. List All Movies Released in a Specific Year (e.g., 2020)**

SELECT \*

FROM netflix

WHERE release\_year = 2020;

Objective: Retrieve all movies released in a specific year.

### **4. Find the Top 5 Countries with the Most Content on Netflix**

SELECT \*

FROM

(

SELECT

UNNEST(STRING\_TO\_ARRAY(country, ',')) AS country,

COUNT(\*) AS total\_content

FROM netflix

GROUP BY 1

) AS t1

WHERE country IS NOT NULL

ORDER BY total\_content DESC

LIMIT 5;

Objective: Identify the top 5 countries with the highest number of content items.

### **5. Identify the Longest Movie**

SELECT \*

FROM netflix

WHERE type = 'Movie'

ORDER BY SPLIT\_PART(duration, ' ', 1)::INT DESC;

Objective: Find the movie with the longest duration.

### **6. Find Content Added in the Last 5 Years**

SELECT \*

FROM netflix

WHERE TO\_DATE(date\_added, 'Month DD, YYYY') >= CURRENT\_DATE - INTERVAL '5 years';

Objective: Retrieve content added to Netflix in the last 5 years.

### **7. Find All Movies/TV Shows by Director 'Rajiv Chilaka'**

SELECT \*

FROM (

SELECT

\*,

UNNEST(STRING\_TO\_ARRAY(director, ',')) AS director\_name

FROM netflix

) AS t

WHERE director\_name = 'Rajiv Chilaka';

Objective: List all content directed by 'Rajiv Chilaka'.

### **8. List All TV Shows with More Than 5 Seasons**

SELECT \*

FROM netflix

WHERE type = 'TV Show'

AND SPLIT\_PART(duration, ' ', 1)::INT > 5;

Objective: Identify TV shows with more than 5 seasons.

### **9. Count the Number of Content Items in Each Genre**

SELECT

UNNEST(STRING\_TO\_ARRAY(listed\_in, ',')) AS genre,

COUNT(\*) AS total\_content

FROM netflix

GROUP BY 1;

Objective: Count the number of content items in each genre.

### **10.Find each year and the average numbers of content release in India on netflix.**

return top 5 year with highest avg content release!

SELECT

country,

release\_year,

COUNT(show\_id) AS total\_release,

ROUND(

COUNT(show\_id)::numeric /

(SELECT COUNT(show\_id) FROM netflix WHERE country = 'India')::numeric \* 100, 2

) AS avg\_release

FROM netflix

WHERE country = 'India'

GROUP BY country, release\_year

ORDER BY avg\_release DESC

LIMIT 5;

Objective: Calculate and rank years by the average number of content releases by India.

### **11. List All Movies that are Documentaries**

SELECT \*

FROM netflix

WHERE listed\_in LIKE '%Documentaries';

Objective: Retrieve all movies classified as documentaries.

### **12. Find All Content Without a Director**

SELECT \*

FROM netflix

WHERE director IS NULL;

Objective: List content that does not have a director.

### **13. Find How Many Movies Actor 'Salman Khan' Appeared in the Last 10 Years**

SELECT \*

FROM netflix

WHERE casts LIKE '%Salman Khan%'

AND release\_year > EXTRACT(YEAR FROM CURRENT\_DATE) - 10;

Objective: Count the number of movies featuring 'Salman Khan' in the last 10 years.

### **14. Find the Top 10 Actors Who Have Appeared in the Highest Number of Movies Produced in India**

SELECT

UNNEST(STRING\_TO\_ARRAY(casts, ',')) AS actor,

COUNT(\*)

FROM netflix

WHERE country = 'India'

GROUP BY actor

ORDER BY COUNT(\*) DESC

LIMIT 10;

Objective: Identify the top 10 actors with the most appearances in Indian-produced movies.

### **15. Categorize Content Based on the Presence of 'Kill' and 'Violence' Keywords**

SELECT

category,

COUNT(\*) AS content\_count

FROM (

SELECT

CASE

WHEN description ILIKE '%kill%' OR description ILIKE '%violence%' THEN 'Bad'

ELSE 'Good'

END AS category

FROM netflix

) AS categorized\_content

GROUP BY category;

Objective: Categorize content as 'Bad' if it contains 'kill' or 'violence' and 'Good' otherwise. Count the number of items in each category.

## **Findings and Conclusion**

* Content Distribution: The dataset contains a diverse range of movies and TV shows with varying ratings and genres.
* Common Ratings: Insights into the most common ratings provide an understanding of the content's target audience.
* Geographical Insights: The top countries and the average content releases by India highlight regional content distribution.
* Content Categorization: Categorizing content based on specific keywords helps in understanding the nature of content available on Netflix.

This analysis provides a comprehensive view of Netflix's content and can help inform content strategy and decision-making.

## **Author - Vipin Kumar**

This project is part of my portfolio, showcasing the SQL skills essential for data analyst roles. If you have any questions, feedback, or would like to collaborate, feel free to get in touch!

### **Stay Updated and Join the Community**

For more content on SQL, data analysis, and other data-related topics, make sure to follow me on social media and join our community:

* YouTube:<https://youtube.com/@vipinsiracademy1671?si=dNraa9Dlu9z7-iwL>
* Instagram: <https://www.instagram.com/_vipin_kumar16/#>
* LinkedIn: [www.linkedin.com/in/ vipin-kumar-0a2638250](http://www.linkedin.com/in/%20vipin-kumar-0a2638250)
* Git Hub :<https://github.com/vipin-kumar16/Capstone-EDA-Project>

Thank you for your support, and I look forward to connecting with you!