**NETFLIX DATA ANALYSIS**

**INTRODUCTION**

Netflix has become one of the world’s leading entertainment platforms, offering a vast collection of movies and TV shows to millions of subscribers across the globe. Its library spans multiple genres, ratings, and production origins, catering to diverse audience preferences. As the streaming industry becomes increasingly competitive, analyzing Netflix's content catalog provides valuable insights into its strategy, growth patterns, and areas for improvement.

Understanding how Netflix structures and curates its library is crucial for evaluating its position in the market. For instance, identifying the most common genres or countries contributing to its content can reveal its target demographics. Similarly, exploring the frequency of TV shows versus movies provides insights into Netflix's focus on episodic versus long-form storytelling.

This report leverages data from a publicly available Netflix dataset sourced from Kaggle. The dataset contains essential metadata, such as the type of content (Movies or TV Shows), genres, release years, countries of origin, directors, cast members, ratings, and more. By addressing 15 specific questions, this analysis aims to uncover trends, patterns, and unique characteristics of Netflix's offerings.

The insights derived from this report can aid in understanding content diversity, identifying popular creators and actors, and evaluating Netflix’s efforts to cater to global audiences. Additionally, it provides a foundation for Netflix to enhance its data-driven content recommendations and further optimize its library for user engagement. The questions analyzed here represent common challenges and strategic decision points that can guide future content curation, licensing, and marketing initiatives.

**OBJECTIVES OF NETFLIX DATA ANALYSIS**

The ultimate objective of this Netflix data analysis is to **enhance Netflix's content strategy** by using data-driven insights. This will allow the platform to:

* Optimize content offerings across regions and genres.
* Better understand audience preferences and engagement patterns.
* Make more informed decisions about which genres, countries, or types of content to invest in or expand.

The analysis will provide actionable insights that will support **data-backed decision-making** in content acquisition, production, and regional strategies.

**PROBLEM STATEMENT**

With the growing competition in the streaming industry, platforms like Netflix need to continually evaluate and optimize their content libraries to meet the evolving preferences of their global audience. Understanding the distribution of content types, popular genres, countries of origin, and audience ratings is crucial for improving user satisfaction and retention. Additionally, analyzing the performance of creators, actors, and long-running TV shows can provide insights into viewer demand and highlight areas for further investment. This analysis aims to address key questions related to Netflix's content catalog, providing a data-driven approach to evaluate its current offerings, identify trends, and uncover opportunities to enhance its competitive edge in the market.

### ****NETFLIX DATASET DESCRIPTION****

Dataset Link : <https://www.kaggle.com/datasets/shivamb/netflix-shows?resource=download>

The Netflix dataset sourced from Kaggle contains information about movies and TV shows available on the Netflix platform. This dataset is useful for analyzing content trends, patterns, and strategies.

Below is a description of each column in the dataset:

1. **show\_id**
   * **Type:** String
   * **Description:** A unique identifier for each show in the dataset.
2. **type**
   * **Type:** String
   * **Description:** Specifies whether the content is a "Movie" or a "TV Show".
3. **title**
   * **Type:** String
   * **Description:** The title of the movie or TV show.
4. **director**
   * **Type:** String
   * **Description:** The director(s) of the content. This field may be blank for certain entries, especially TV shows or group-produced works.
5. **cast**
   * **Type:** String
   * **Description:** The list of main actors in the content.
6. **country**
   * **Type:** String
   * **Description:** The country where the movie or TV show was produced. This field may have missing values.
7. **date\_added**
   * **Type:** Date
   * **Description:** The date when the content was added to Netflix. This is useful for tracking content additions over time.
8. **release\_year**
   * **Type:** Integer
   * **Description:** The year the content was originally released.
9. **rating**
   * **Type:** String
   * **Description:** The rating assigned to the content (e.g., PG-13, TV-MA), indicating its suitability for different audience groups.
10. **duration**
    * **Type:** String
    * **Description:** Specifies the runtime of a movie (in minutes) or the number of seasons for a TV show.
11. **listed\_in**
    * **Type:** String
    * **Description:** The genres or categories the content belongs to (e.g., Comedy, Drama, Documentary).
12. **description**
    * **Type:** String
    * **Description:** A brief summary or synopsis of the content.

### ****POTENTIAL USES OF THE DATASET****

1. Analyzing the ratio of movies to TV shows on Netflix.
2. Examining genre diversity and content distribution by country.
3. Understanding trends in content addition (e.g., by year or region).
4. Identifying popular ratings and their frequency in the catalog.
5. Evaluating contributions of specific directors, actors, or genres.

**COMPREHENSIVE LIST OF NETFLIX BUSINESS PROBLEMS**

Below is the consolidated list of business problems, organized into **sections** for clarity and logical grouping.

**1. Content Type and Distribution**

1. Count the number of Movies vs TV Shows.
2. Identify underrepresented genres on Netflix.
3. Count the number of content items in each genre.
4. List all TV shows with more than 5 seasons.
5. Evaluate the lifespan of TV shows (average number of seasons).

**2. Ratings and Content Categorization**

1. Find the most common rating for movies and TV shows.
2. Analyze the distribution of family-friendly content vs mature content.
3. Categorize content based on the presence of keywords ('kill' and 'violence') in the description field. Label such content as 'Bad' and the rest as 'Good.' Count how many items fall into each category

**3. Content Trends and Additions**

1. List all movies released in a specific year (e.g., 2020).
2. Find content added in the last 5 years.
3. Analyse the days of release of movies.

**4. Regional Focus and Representation**

1. Find the top 5 countries with the most content on Netflix.
2. Analyze country-wise contributions to Netflix’s library.
3. Find the average release year for content produced in a specific country

**5. Talent and Collaboration Analysis**

1. List all movies/TV shows by director 'Rajiv Chilaka.'
2. Find how many movies actor 'Salman Khan' appeared in the last 10 years.
3. Find the top 10 actors who have appeared in the highest number of movies in their career.
4. Identify the most frequent directors.
5. Analyze recurring actor collaborations.

**6. Duration and Performance**

1. Identify the longest movie and TV show duration.
2. Highlight long-running TV shows with exceptional durations.

**7. Missing Metadata and Documentaries**

1. Find all content without a director.
2. List all movies that are documentaries.

### METHODOLOGY USING SQL

The methodology for analyzing Netflix's dataset using SQL involves several key steps that include data cleaning, querying, and analysis. Below is a step-by-step breakdown of the process:

#### **1. Data Inspection and Cleaning:**

* **Data Inspection:** First, the dataset was reviewed to understand the structure of key columns such as show\_id, type, title, director, cast, country, date\_added, release\_year, rating, duration, listed\_in, and description.
* **Missing Data Handling:**
  + Missing or NULL values were handled by excluding rows where necessary (e.g., for analyses based on release\_year or country).
  + Missing cast or director fields were either excluded from certain analyses or replaced with "NULL".
  + Inconsistent formats in duration (e.g., movies vs. TV shows) were addressed using string manipulation to parse the data properly.
* **Data Type Conversion:**
  + The date\_added field was converted to the DATE format for temporal analysis.
  + The release\_year field was cast as an integer for numerical analysis.

**INSIGHTS AND FINDINGS**

CREATE TABLE netflix

(

show\_id varchar(6),

type varchar(10),

title varchar(150),

director varchar(225),

casts varchar(1000),

country varchar(200),

date\_added varchar(50),

release\_year int,

rating varchar(10),

duration varchar(15),

listed\_in varchar(100),

description varchar(250)

)

SELECT \* FROM netflix

**--BUSINESS PROBLEMS—**

**--1.Count the number of Movies vs TV Shows.**

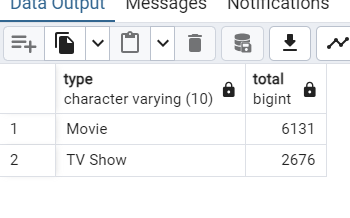
SELECT

type,

COUNT(\*) as total

from netflix

GROUP BY type



**--2.Identify underrepresented genres on Netflix**.

WITH GenreCounts AS (

SELECT

TRIM(genre) AS genre

FROM

netflix,

UNNEST(string\_to\_array(listed\_in, ',')) AS genre

)

SELECT

genre,

COUNT(\*) AS genre\_count

FROM

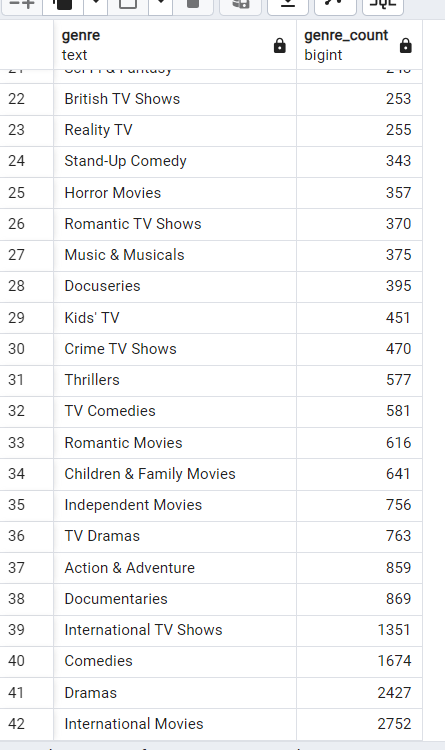
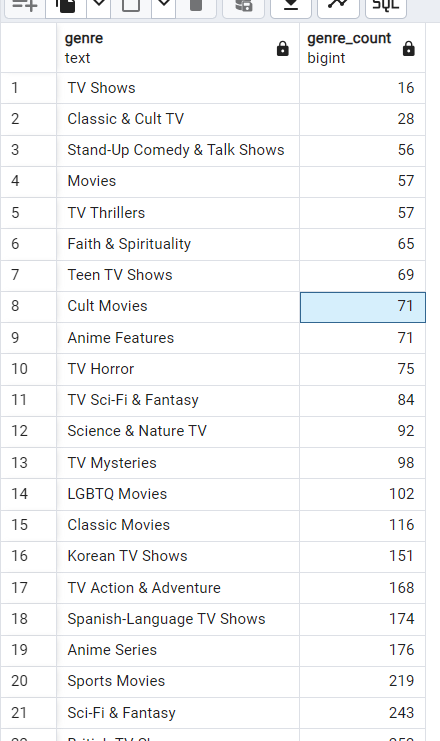
GenreCounts

GROUP BY

genre

ORDER BY

genre\_count ASC;



**--3.Count the number of content items in each genre.**

WITH GenreCounts AS (

SELECT

TRIM(genre) AS genre

FROM

netflix,

UNNEST(string\_to\_array(listed\_in, ',')) AS genre

)

SELECT

genre,

COUNT(\*) AS genre\_count

FROM

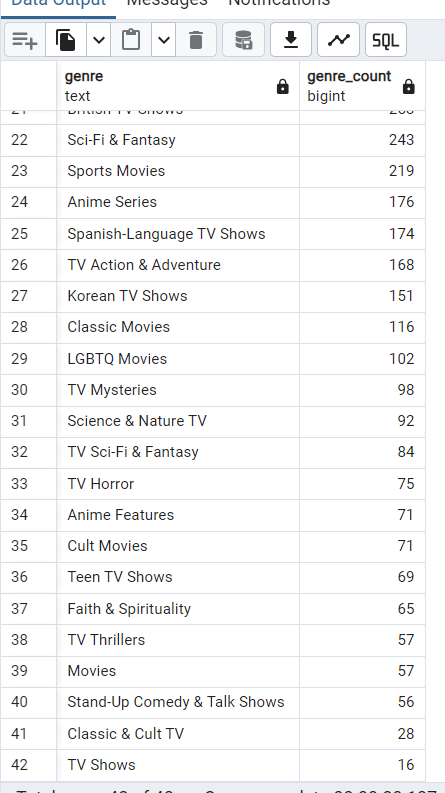
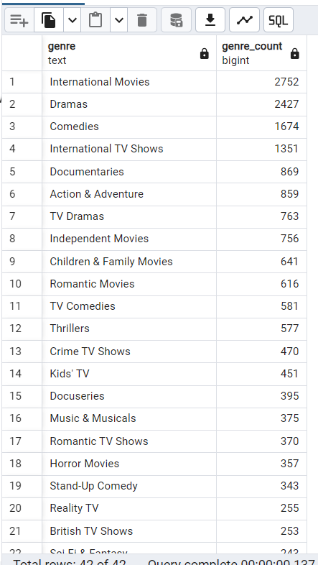
GenreCounts

GROUP BY

genre

ORDER BY

genre\_count DESC;



**--4.List all TV shows with more than 5 seasons**

SELECT

title

FROM

netflix

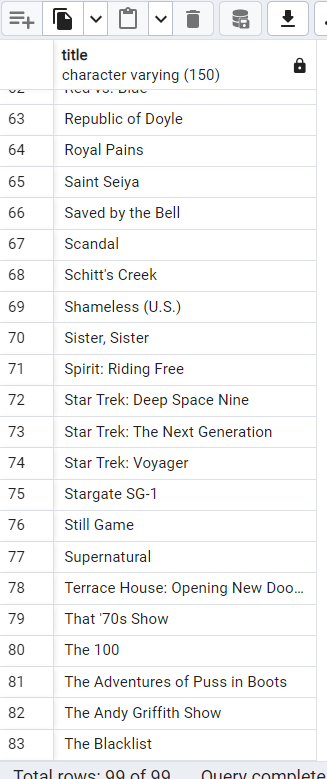
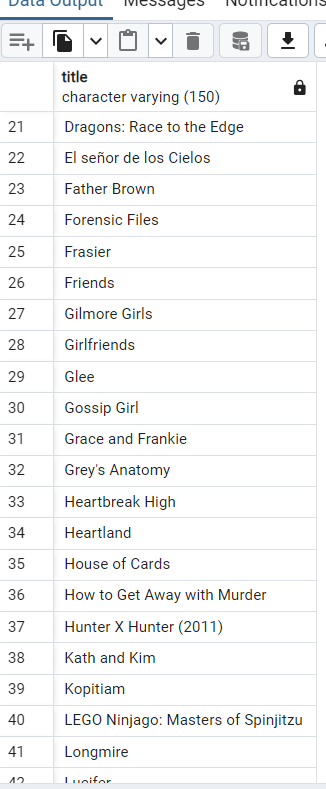
WHERE

type = 'TV Show'

AND CAST(SPLIT\_PART(duration, ' ', 1) AS INTEGER) > 5

ORDER BY

title;

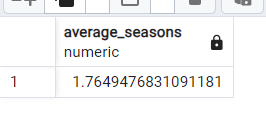
**--5.Evaluate the lifespan of TV shows (average number of seasons).**

SELECT

AVG(CAST(SUBSTRING(duration FROM 1 FOR POSITION(' ' IN duration)-1) AS INTEGER)) AS average\_seasons

FROM netflix

WHERE type = 'TV Show';



**--6.Find the most common rating for movies and TV shows.**

SELECT

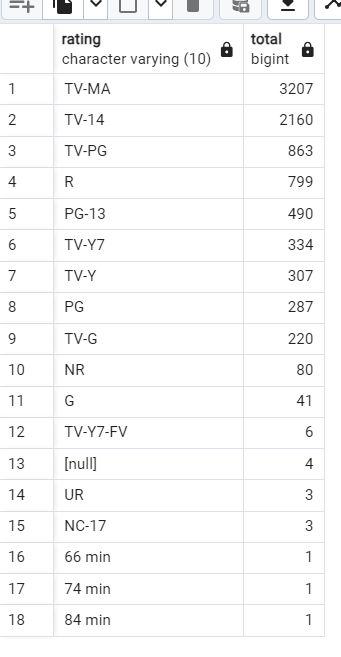
rating,

COUNT(\*) as total

FROM netflix

GROUP BY rating

ORDER BY total DESC



**--7.Analyze the distribution of family-friendly content vs mature content.**

SELECT

CASE

WHEN rating IN ('G', 'PG', 'TV-Y', 'TV-Y7', 'TV-G', 'TV-PG') THEN 'Family-Friendly'

WHEN rating IN ('PG-13', 'R', 'NC-17', 'TV-14', 'TV-MA') THEN 'Mature'

ELSE 'Other'

END AS content\_type,

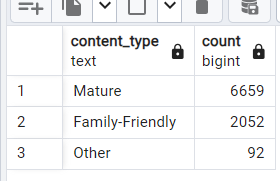
COUNT(\*) AS count

FROM netflix

WHERE rating IS NOT NULL

GROUP BY content\_type

ORDER BY count DESC;



**--8.Categorize content based on the presence of keywords ('kill' and 'violence') in the description field. Label such content as 'Bad' and the rest as 'Good.' Count how many items fall into each category.**

SELECT

CASE

WHEN LOWER(description) LIKE '%kill%' OR LOWER(description) LIKE '%violence%' THEN 'Bad'

ELSE 'Good'

END AS content\_category,

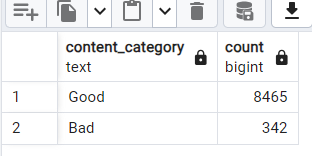
COUNT(\*) AS count

FROM netflix

WHERE description IS NOT NULL

GROUP BY content\_category

ORDER BY count DESC;



**--9.List all movies released in a specific year (e.g., 2020).**

SELECT title FROM netflix

WHERE release\_year='2020'

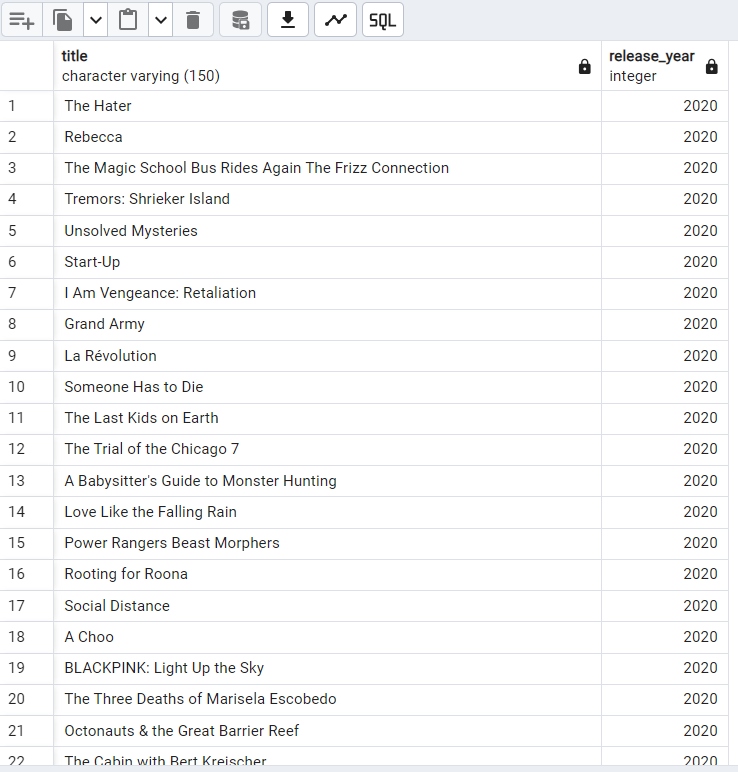
****

**--10.Find content added in the last 5 years.**

SELECT title,release\_year FROM netflix

WHERE release\_year BETWEEN 2020 AND 2024

ORDER BY release\_year ASC;



**--11. Analyse the days of release of movies.**

SELECT

TO\_CHAR(TO\_DATE(date\_added, 'Month DD,YYYY'), 'Day') AS day\_of\_week,

COUNT(\*) AS release\_count

FROM

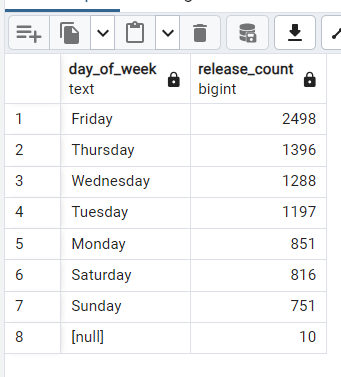
netflix

GROUP BY

day\_of\_week

ORDER BY

release\_count DESC;



**--12.Find the top 5 countries with the most content on Netflix.**

SELECT

country,

COUNT(\*) as Number\_of\_releases

FROM

(

SELECT unnest(string\_to\_array(country, ', ')) AS country

FROM netflix

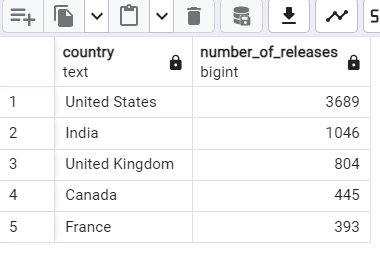
WHERE country IS NOT NULL

) AS country\_list

GROUP BY country

ORDER BY Number\_of\_releases DESC

LIMIT 5;



**--13.Analyze country-wise contributions to Netflix’s library**.

SELECT

country,

COUNT(\*) as Number\_of\_releases

FROM

(

SELECT unnest(string\_to\_array(country, ', ')) AS country

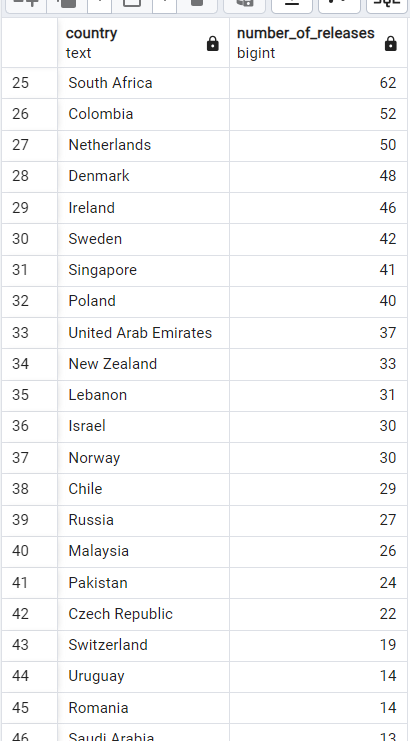
FROM netflix

WHERE country IS NOT NULL

) AS country\_list

GROUP BY country

ORDER BY Number\_of\_releases DESC



**--14. Find the average release year for content produced in a specific country.**

SELECT

country,

ROUND(AVG(CAST(release\_year AS INTEGER)),0) AS average\_release\_year

FROM

(

SELECT unnest(string\_to\_array(country, ', ')) AS country, release\_year

FROM netflix

WHERE country IS NOT NULL AND release\_year IS NOT NULL

) AS country\_list

GROUP BY country

ORDER BY average\_release\_year DESC



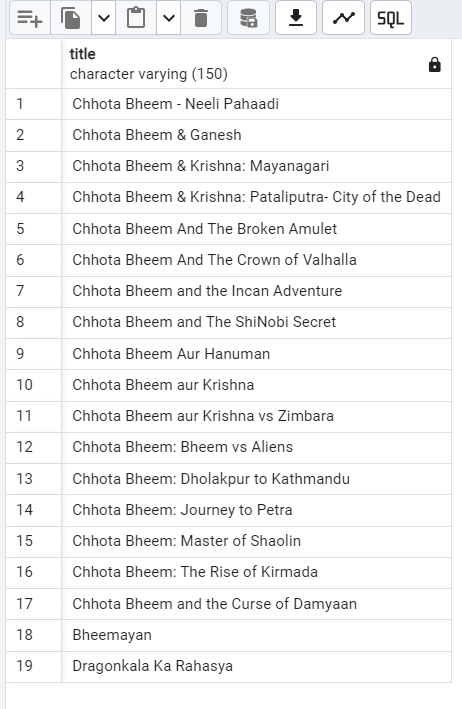
**--15.List all movies/TV shows by director 'Rajiv Chilaka.'**

SELECT

title

FROM netflix

WHERE director = 'Rajiv Chilaka'



**--16.Find how many movies actor 'Salman Khan' appeared in the last 10 years.**

SELECT

title

FROM

(

SELECT unnest(string\_to\_array(casts, ', ')) AS casts,title

FROM netflix

WHERE country IS NOT NULL AND title IS NOT NULL

) AS cast\_list

WHERE casts = 'Salman Khan'



**--17.Find the top 10 actors who have appeared in the highest number of movies in their**

**career.**

SELECT

casts,

COUNT(\*) as Total\_no\_of\_movies

FROM

(

SELECT unnest(string\_to\_array(casts, ', ')) AS casts,title

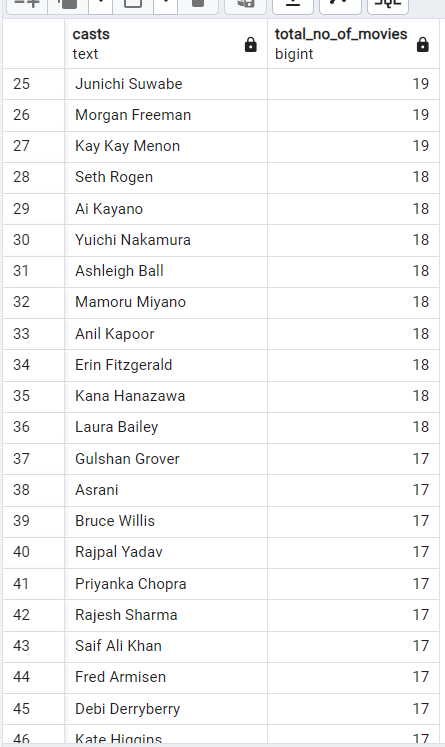
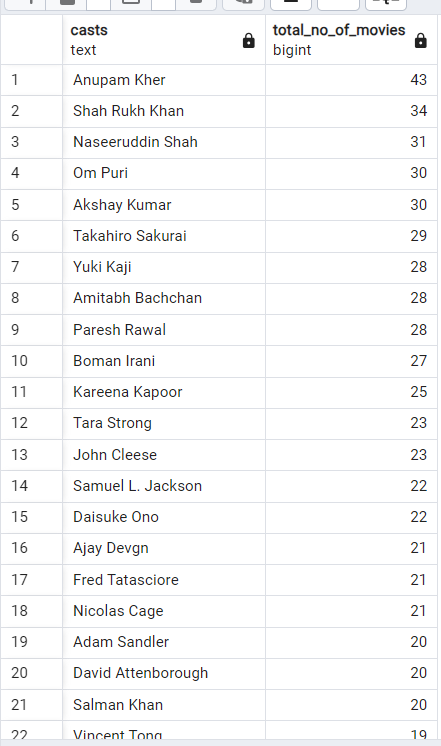
FROM netflix

WHERE country IS NOT NULL AND title IS NOT NULL

) AS cast\_list

GROUP BY casts

ORDER BY Total\_no\_of\_movies DESC;



**--18.Identify the most frequent directors.**

SELECT

director,

COUNT(\*) as Total\_no\_of\_movies

FROM

(

SELECT unnest(string\_to\_array(director, ', ')) AS director,title

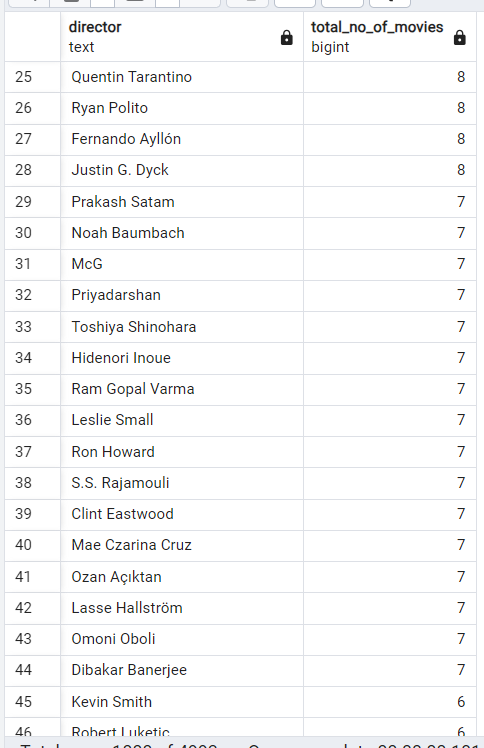
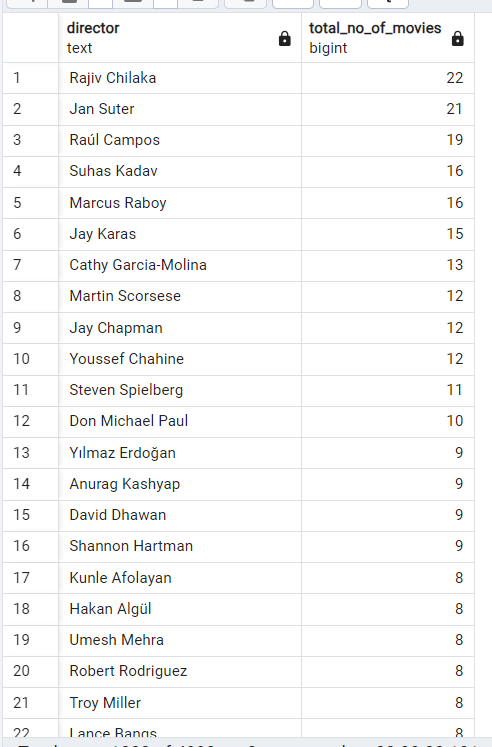
FROM netflix

WHERE director IS NOT NULL AND title IS NOT NULL

) AS director\_list

GROUP BY director

ORDER BY Total\_no\_of\_movies DESC;



**--19.Analyze recurring actor collaborations.**

WITH actor\_pairs AS (

SELECT

unnest(string\_to\_array(casts, ', ')) AS actor,

title

FROM

netflix

WHERE

casts IS NOT NULL AND title IS NOT NULL

),

collaborations AS (

SELECT

a1.actor AS actor\_1,

a2.actor AS actor\_2,

a1.title

FROM

actor\_pairs a1

JOIN

actor\_pairs a2

ON

a1.title = a2.title

WHERE

a1.actor < a2.actor

)

SELECT

actor\_1,

actor\_2,

COUNT(\*) AS collaboration\_count

FROM

collaborations

GROUP BY

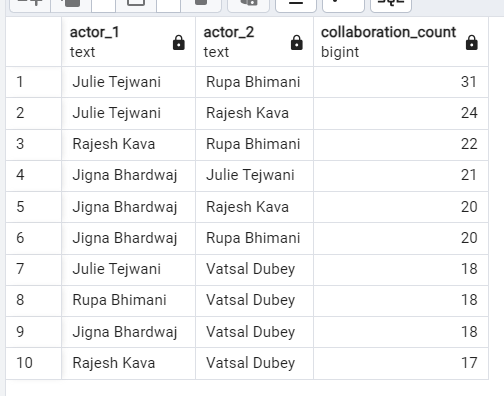
actor\_1,

actor\_2

ORDER BY

collaboration\_count DESC

LIMIT 10;



**--20.Identify the longest movie or TV show duration**

SELECT

title,

duration,

CAST(SUBSTRING(duration FROM '\d+') AS INTEGER) AS movie\_duration

FROM

netflix

WHERE

type = 'Movie' AND

duration IS NOT NULL

ORDER BY

movie\_duration DESC

LIMIT 1;

SELECT

title,

CAST(SPLIT\_PART(duration, ' ', 1) AS INTEGER) as seasons

FROM

netflix

WHERE

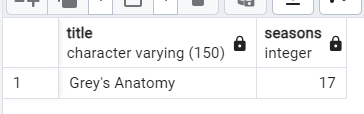
type = 'TV Show'

AND duration IS NOT NULL

ORDER BY

seasons DESC

LIMIT 1;



**--21.Highlight long-running TV shows with exceptional durations.**

SELECT

title,

CAST(SPLIT\_PART(duration, ' ', 1) AS INTEGER) as seasons

FROM

netflix

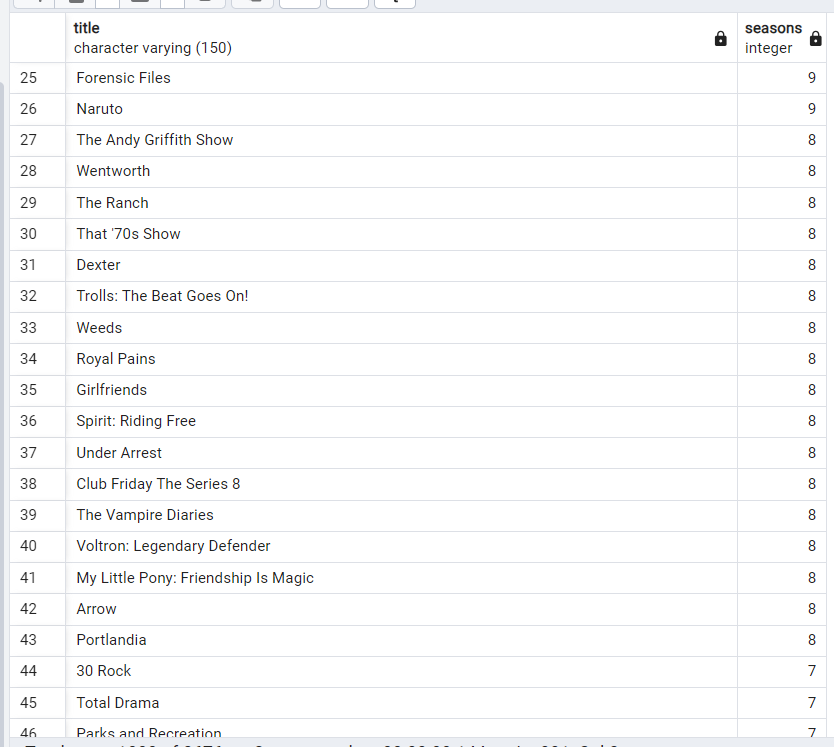
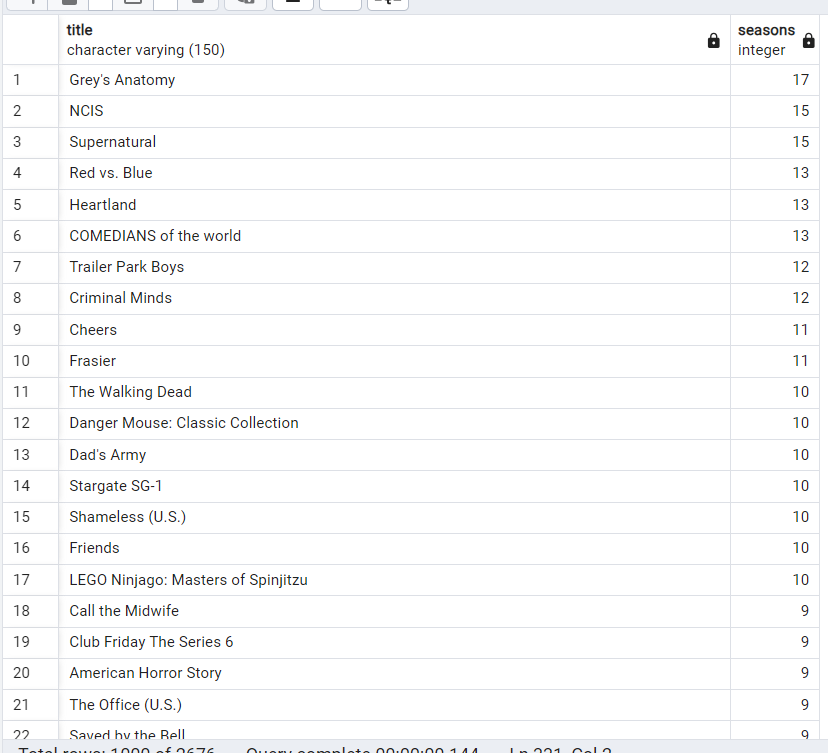
WHERE

type = 'TV Show'

AND duration IS NOT NULL

ORDER BY

seasons DESC;



**--22.Find all content without a director.**

SELECT

title

FROM

netflix

WHERE director IS NULL



**--23.List all movies that are documentaries**

SELECT

title

FROM

netflix

WHERE

type = 'Movie'

AND listed\_in LIKE '%Documentaries%'



**CONCLUSION**

In conclusion, the analysis of Netflix's data reveals key insights into user behaviour, content preferences, and operational trends that are essential for driving business decisions and enhancing user experience. The data analysis highlights patterns in subscriber growth, viewing habits, and content engagement, which can inform future content strategies and marketing efforts.

Key takeaways include the importance of personalized recommendations, the impact of original content on subscriber retention, and the need for adaptive pricing models to maintain growth in competitive markets. Furthermore, understanding the demographic and regional preferences allows Netflix to tailor its offerings to diverse audiences, ensuring continued relevance in a rapidly evolving streaming landscape.

Ultimately, this analysis underscores the significance of leveraging data to optimize user experiences, improve content investments, and make informed strategic decisions that will help Netflix maintain its leadership position in the global entertainment industry.