Netflix Data Analysis Project

Introduction:

The Netflix dataset contains comprehensive information about movies and TV shows available on the streaming platform. It includes various attributes such as show ID, type (movie or TV show), title, director, cast, country of production, the date it was added to Netflix, release year, rating, duration, genres, and descriptions. This project involves performing SQL operations on this dataset to derive meaningful insights, including content distribution and trends across different categories.

SQL Queries and Results

```
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)
);
```

```
select * from netflix;
```

15 Business Problems

Q1. Count the number of Movies vs TV Shows

select distinct(type), count(*) from netflix
group by type;

	type character varying (10)	count bigint
1	Movie	6131
2	TV Show	2676

Q2. 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;

	type character varying (10)	most_frequent_rating character varying (15)
1	Movie	TV-MA
2	TV Show	TV-MA

Q3. List all movies released in a specific year (e.g., 2020)

select * from netflix
where release_year = 2020;

	g (55)	release_year integer	rating character varying (15)	duration character varying (15)	listed_in character varying (250)
1	2021	2020	PG-13	90 min	Documentaries
2	2021	2020	TV-MA	67 min	Documentaries, Internatic
3	2021	2020	TV-MA	2 Seasons	Crime TV Shows, Spanish
4	2021	2020	TV-MA	3 Seasons	British TV Shows, Internat
5	2021	2020	TV-Y7	1 Season	Kids' TV

Q4. Find the Top 5 Countries with the Most Content on Netflix

select
        unnest(string_to_array(country, ',')) as new_country,
        count(show_id) as total_content
        from netflix
group by 1
order by 2 desc
limit 5;

                                        --OR--

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;

	new_country text	total_content bigint	
1	United States	3211	
2	India	1008	
3	United Kingdo...	628	
4	United States	479	
5	Canada	271	

Q5. Identify the longest movie?

SELECT *FROM netflix
WHERE type = 'Movie'
ORDER BY SPLIT_PART(duration, ' ', 1)::INT DESC;

		date_added character varying (55) 🔒	release_year integer 🔒	rating character varying (15) 🔒	duration character varying (15) 🔒	listed_in characte
1		August 15, 2016	2015	66 min	[null]	Movies
2		September 16, 2016	2010	84 min	[null]	Movies
3		April 4, 2017	2017	74 min	[null]	Movies
4		December 28, 2018	2018	TV-MA	312 min	Dramas
5		June 15, 2021	2021	TV-G	273 min	Docume
6		May 21, 2020	1973	TV-14	253 min	Comedi
7		May 21, 2020	1979	TV-14	237 min	Comedi

Q6. Find content added in the last 5 years

SELECT *
FROM netflix
WHERE TO_DATE(date_added, 'Month DD, YYYY') >= CURRENT_DATE - INTERVAL '5 years';

		country character varying (550)		date_added character varying (55) 🔒	release_year integer 🔒
1		United States		September 25, 2021	2020
2		South Africa		September 24, 2021	2021
3		[null]		September 24, 2021	2021
4		[null]		September 24, 2021	2021
5		India		September 24, 2021	2021

Q7. Find all the movies/TV shows by director 'Rajiv Chilaka'

select * from netflix   -- valid if name is not with someone else with comman .
where director ilike '%Rajiv Chilaka'

                              – OR –

SELECT *FROM (
  SELECT
    *,
    UNNEST(STRING_TO_ARRAY(director, ',')) AS director_name
  FROM netflix
)
WHERE director_name = 'Rajiv Chilaka';

		title character varying (250)		director character varying (550)		casts character va
1		Chhota Bheem - Neeli Pahaadi		Rajiv Chilaka		Vatsal Dube
2		Chhota Bheem & Ganesh		Rajiv Chilaka		Vatsal Dube
3		Chhota Bheem & Krishna: Mayanagari		Rajiv Chilaka		Vatsal Dube
4		Chhota Bheem & Krishna: Pataliputra- City of the Dead		Rajiv Chilaka		Vatsal Dube
5		Chhota Bheem And The Broken Amulet		Rajiv Chilaka		Vatsal Dube
6		Chhota Bheem And The Crown of Valhalla		Rajiv Chilaka		Vatsal Dube

Q8. List All TV Shows with More Than 5 Seasons

select *
from netflix
where
    type= 'TV Show'
    and
    split_part(duration, ' ', 1)::int >5;

	_year		rating character varying (15)		duration character varying (15)		listed_in character varying (250)
1	2021		TV-14		9 Seasons		British TV Shows, Reality TV
2	2021		TV-PG		6 Seasons		Reality TV
3	2021		TV-Y		6 Seasons		Kids' TV
4	1994		TV-PG		9 Seasons		Kids' TV, TV Comedies
5	2021		TV-14		6 Seasons		Crime TV Shows, TV Comedies, TV Dramas
6	2021		TV-MA		7 Seasons		TV Comedies, TV Dramas

NOTE :: is used to convert string to numeric/int.

Q9. Count the Number of Content Items in Each Genre

select
    unnest(string_to_array(listed_in, ',')) as genre,
    count(show_id) as total_content
    from netflix

group by 1;

	genre text		total_content bigint	
1	International TV Shows		774	
2	Stand-Up Comedy		9	
3	Spanish-Language TV Shows		172	
4	Romantic Movies		613	
5	Anime Features		21	
6	TV Horror		11	
7	Stand-Up Comedy & Talk Shows		34	
8	Kids' TV		63	
9	Docuseries		221	
10	Children & Family Movies		605	
11	Movies		57	

Q10. Find each year and the average numbers of content released in India on netflix.

```
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;
```

	country character varying (550)	release_year integer	total_release bigint	avg_release numeric
1	India	2017	101	10.39
2	India	2018	94	9.67
3	India	2019	87	8.95
4	India	2020	75	7.72
5	India	2016	73	7.51

Q11. List All Movies that are Documentaries

```
select * from netflix
where listed_in like '%Documentaries%'
```

	ease_year eger	rating character varying (15)	duration character varying (15)	listed_in character varying (250)
1	2020	PG-13	90 min	Documentaries
2	2020	TV-MA	67 min	Documentaries, International Movies
3	2021	PG	23 min	Documentaries
4	2021	TV-14	113 min	Documentaries, International Movies, Sports Movies
5	2021	PG-13	96 min	Documentaries, Sports Movies
6	2021	TV-14	92 min	Documentaries, International Movies

Q12. Find All Content Without a Director

select * from netflix
where director is null;

	title character varying (250)	director character varying (550)	ca ch
1	Blood & Water	[null]	A
2	Jailbirds New Orleans	[null]	[n
3	Kota Factory	[null]	N
4	Vendetta: Truth, Lies and The Mafia	[null]	[n
5	Crime Stories: India Detectives	[null]	[n

Q13. 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;

	director character varying (550)	casts character varying (1050)	c c
1	Sooraj R. Barjatya	Salman Khan, Sonam Kapoor, Anupam Kher, Neil Nitin Mukesh, Sanjay Mishra, Swara Bhaskar, Deepak Dobriyal, Arman Kohli, Manoj Joshi, Samaira Rao, Karuna Pandey, Sucheta Kh...	
2	Rakesh Ranjan Kumar	Lorena Franco, Bijesh Jayarajan, Neet Chowdhary, Rajeev Gaursingh, Karran Jeet, Rajesh Sharma, Gautam Kurup, Poonam Mathur, Salman Khan, Manish Devkush	

Q14. 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(show_id)
from netflix
group by 1
order by 2 desc
limit 10;

	actor (text)	count (bigint)
1	Anupam Kher	39
2	Rupa Bhimani	31
3	Takahiro Sakurai	30
4	Julie Tejwani	28
5	Om Puri	27
6	Shah Rukh Khan	26
7	Rajesh Kava	26
8	Yuki Kaji	25
9	Andrea Libman	25
10	Paresh Rawal	25

Q15.Categorize Content Based on the Presence of 'Kill' and 'Violence' Keywords using cet.

```
with new_table as
(select *,
        case when
                        description ilike '%kill%' or
                        description ilike '%violence%'
        then 'Bad_Content'
        else 'Good_Content'
        end as category
from netflix
)
SELECT
   category,
   COUNT(*) AS content_count
        from new_table
        group by category;
```

```
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;
```

	category text	content_count bigint
1	Good	8465
2	Bad	342

Conclusion:

The analysis of the Netflix dataset through SQL provided insights into the types of content available on the platform and their characteristics, including their distribution across genres, countries, and years. The SQL queries helped in understanding content trends, viewing patterns, and provided valuable data points for further content strategy on the platform. These insights can help inform Netflix's strategic decisions and better align offerings with audience preferences.