Netflix: Data Cleaning, Analysis and Visualization

1. Count the number of Movies vs TV Shows

```
SELECT type, COUNT(*) as Total_count

    FROM Netflix
    GROUP BY type
```

Output:

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

2. Find the most common rating for movies and TV Shows

```
SELECT type, rating FROM

(

SELECT type, rating, count(*),

RANK() OVER(PARTITION BY type ORDER BY count(*) DESC) as ranking

FROM Netflix

GROUP BY 1,2

) AS t1

where ranking=1
```

Output:

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

3. List all movies released in a specific year (eg., 2020)

```
SELECT * FROM Netflix

WHERE type= 'Movie' and release_year = 2020
```

Output: File attached

4. Find the top 5 countries with the most content on Netflix

```
SELECT
   TRIM(UNNEST(STRING_TO_ARRAY(country, ','))) AS new_country,
   COUNT(show_id) AS Total_content
FROM Netflix
GROUP BY TRIM(UNNEST(STRING_TO_ARRAY(country, ',')))
ORDER BY 2 DESC
LIMIT 5
```

Output:

	new_country text	total_content bigint
1	United States	3690
2	India	1046
3	United Kingdom	806
4	Canada	445
5	France	393

5. Identity the longest movie or TV Show duration

```
SELECT *
FROM Netflix
WHERE type = 'Movie' AND
duration = (select max(duration) from Netflix)
```

Output: File attached.

6. Find the content added in the last 5 years

```
SELECT * FROM Netflix
WHERE TO_DATE(date_added, 'month DD, yyyy') >= CURRENT_DATE - INTERVAL '5 YEARS'
```

Output: File attached.

7. Find all the movies/TV Shows by director 'Rajiv Chilaka'

```
SELECT * FROM Netflix
WHERE director ILIKE '%rajiv chilaka%'
```

Output: File attached

8. List all TV Shows with more than 5 seasons

```
SELECT * FROM Netflix
WHERE type = 'TV Show' AND
SPLIT_PART(duration, ' ',1)::NUMERIC > 5
```

Output: File attached

9. Count the number of content items in each genre

```
SELECT
  TRIM(UNNEST(STRING_TO_ARRAY(listed_in, ','))) as genre,
  COUNT(show_id)
FROM Netflix
GROUP BY 1
```

Output: File attached

10.1 find the average release year for content produced in a specific country

```
SELECT
    TRIM(UNNEST(STRING_TO_ARRAY(country, ','))) as country,
            AVG(release_year)
            FROM Netflix
            GROUP BY 1
```

Output: File attached

10.2 Find each year and the average numbers of content release in India on Netflix.

Return top 5 year with highest average content release.

```
SELECT EXTRACT(YEAR FROM TO_DATE(date_added, 'Month DD, YYYY')) as date,
COUNT(*),
```

```
ROUND(
COUNT(*)::numeric/(SELECT count(*) FROM Netflix WHERE country ilike '%india%')::numeric * 100
  ,2) as avg_content
FROM Netflix
WHERE country ilike '%india%'
GROUP BY 1
```

Output:

	date numeric	count bigint	avg_content numeric
1	2018	349	33.37
2	2016	13	1.24
3	2019	218	20.84
4	2021	105	10.04
5	2020	199	19.02
6	2017	162	15.49

11. List all movies that are documentaries

```
SELECT * FROM Netflix
WHERE type = 'Movie' AND
listed_in ILIKE '%documentaries%'
```

Output: File attached.

12. Find all content without a director

```
SELECT * FROM Netflix WHERE director is null
```

Output: File attached.

13. List the movies in which the actor 'Salman khan' appeared in last 10 years

```
SELECT * FROM Netflix where casts ilike '%salman khan%'
            and
   release_year >= EXTRACT(YEAR FROM CURRENT_DATE) - 10
```

Output: File attached.

14. Find the top 10 actors who have appeared in the highest no.of movies produced

```
SELECT
UNNEST(STRING_TO_ARRAY(casts, ',')) as actors,
count(*) as total_movies
from Netflix
where country ilike '%india%'
group by 1
order by 2 desc
limit 10
```

Output:

	actors text	total_movies bigint
1	Anupam Kher	36
2	Om Puri	26
3	Boman Irani	25
4	Paresh Rawal	25
5	Shah Rukh Khan	25
6	Akshay Kumar	23
7	Naseeruddin Sh...	20
8	Amitabh Bachch...	20
9	Kareena Kapoor	20
10	Asrani	17

15. Categorize the content based on the presence of the keywords ' kill' and ' violence' in the description field. Label content containing these keywords as 'Bad' and all other content as 'Good'. Count how many items fall into each category.

```
WITH new_table
AS
(
SELECT *,
CASE WHEN description ilike '%kill%' or
        description ilike '%violence%' THEN 'Bad_Content'
            ELSE 'Good_Content'
            END category
```

```
FROM Netflix

)

SELECT

   category,

        count(*) as total_content

        FROM new_table

        GROUP BY 1
```

Output:

	category text	total_content bigint
1	Bad_Content	342
2	Good_Content	8465