NETFLIX

SQL - Exploratory Data Analysis

Problem Statement

The goal of this exploratory data analysis is to investigate the patterns, trends, and insights within the Netflix catalog of TV shows and movies. By investigating key factors such as genre, release year, runtime and cast details, we aim to cover correlation between content available and the key trends.

1. **Understand Content Catalog:** Analyze the distribution of Movies vs. TV Shows, genres, and countries to evaluate the diversity of Netflix's offerings.
2. **Identify Content Trends:** Examine release patterns over time to discover key trends in Netflix's content production and acquisition.
3. **Target Audience Analysis:** Assess content ratings to determine the primary audience Netflix serves and identify potential gaps in age group or content type.
4. **Analyze Global Reach:** Investigate which countries produce the most content for Netflix and explore regional opportunities for growth.
5. **Popular Genres and Directors:** Identify the most successful genres, directors, and content categories to inform future production and licensing decisions.

Dataset Information

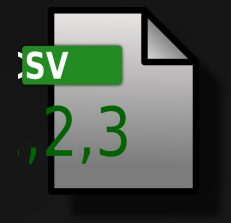
Dataset source: Netflix shows (movies and TV shows)Kaggle

Total records: 8807

Key Attributes:

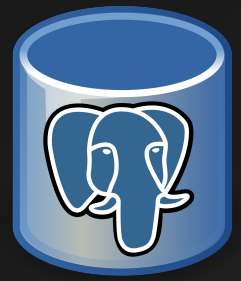
1. show_id : unique identifier for rows(movies and TV shows)
2. type :namely, Movie or TV Show
3. title : name of the TV show or Movie
4. director : director name (blank in few cases)
5. cast : cast (blank in few cases)
6. country : country the show produced in (blank in few cases)
7. date_added : date the show is added on Netflix (YYYY-MM-DD)
8. release_year : year of release
9. rating : rating of the show from 18 categories (blank in few cases)
10. duration : namely, in minutes, or season
11. listed_in : genre
12. description : description text about the show

Data Load



CSV File

Extract and Load



Postgres Table

Preliminary Analysis

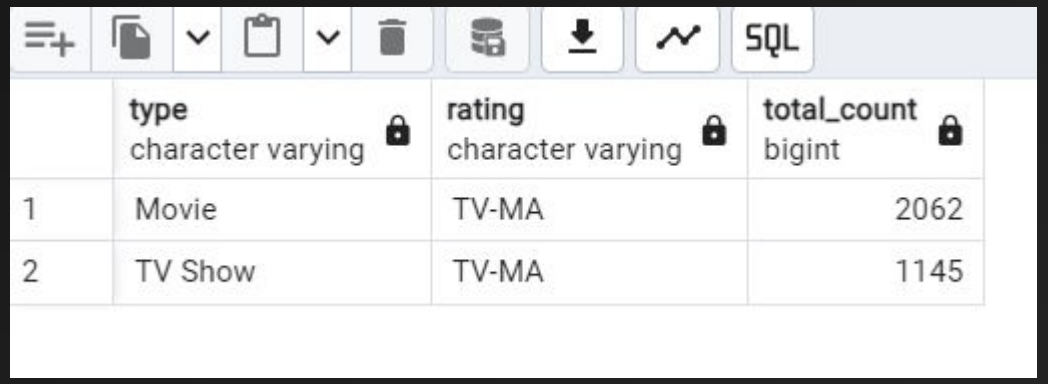
PgAdmin
(PostgreSQL)

SQL functions covered in EDA

- AVG
- CASE
- CAST
- COUNT
- EXTRACT
- GROUP BY
- LIMIT
- STRING_TO_ARRAY
- UNNEST
- ORDER BY
- ROUND
- SUBSTRING
- WHERE
- DISTINCT
- MIN
- ILIKE
- SPLIT_PART
- RANK()

UC#1 Most common rating for Movies and TV shows

```
select  type as show_type,
   rating,
   total_count
from (
select  type,
   rating,
   count(show_id) as total_count,
   rank() over(partition by type order by count(*) desc ) as ranking
from netflixshows a
group by 1,2
order by 3 desc
) as t1
where ranking = 1
;
```



	type character varying	rating character varying	total_count bigint
1	Movie	TV-MA	2062
2	TV Show	TV-MA	1145

UC#2 Avg runtime of TV-shows and movies

```
SELECT type,
        CAST(ROUND(AVG(CAST(split_part(duration, ' ', 1) AS INTEGER)), 2) AS TEXT) ||
       CASE WHEN type = 'Movie' THEN ' minutes'
        ELSE ' Seasons' END average_duration
FROM netflixshows
WHERE duration IS NOT NULL
GROUP BY type;
```

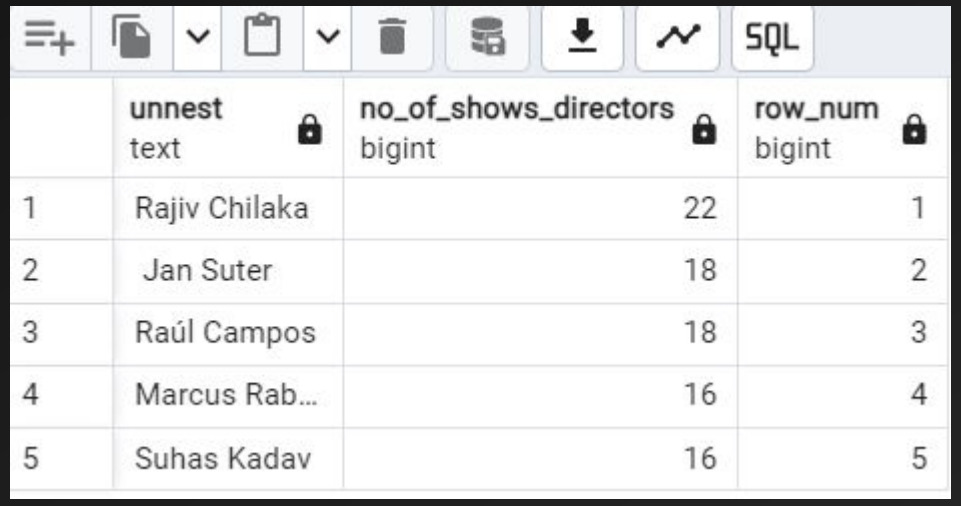
Data Output | Messages | Notifications

SQL

	type character varying	average_duration text
1	Movie	99.58 minutes
2	TV Show	1.76 Seasons

UC#3 Find top 5 Directors with most movies and TV shows

```
select *
from (
select  unnest(string_to_array(director, ',')),
        count(show_id) as no_of_shows_directors,
        row_number() over( order by count(show_id)  desc) as row_num
from netflixshows
group by 1 ) as t1
where row_num <=5
;
```



	unnest text	no_of_shows_directors bigint	row_num bigint
1	Rajiv Chilaka	22	1
2	Jan Suter	18	2
3	Raúl Campos	18	3
4	Marcus Rab…	16	4
5	Suhas Kadav	16	5

UC#4 Top 3 directors who worked as actors

```
--Top 3 director who worked as actor and director in most movies
WITH DIR AS
(
SELECT
      DISTINCT director
FROM netflixshows
WHERE director<>''
and type='Movie'
)
, SUBSET AS
(
SELECT
      d.director,
      COUNT(1) AS count_of_appearance
FROM netflixshows A
JOIN DIR D
ON ((A.cast LIKE '%'|| D.director||',%' )OR (A.cast LIKE '%, '||D.director))
where a.type='Movie'
GROUP BY d.director)
, RN AS
(
SELECT
      director,
      count_of_appearance, rank() over (order by count_of_appearance DESC) RNK
FROM SUBSET
)
SELECT *
FROM RN
WHERE RNK<=3
ORDER BY RNK;
```

Data Output | Messages | Notifications

	director character varying	count_of_appearance bigint	rnk bigint
1	James Franco	19	1
2	Aamir Khan	16	2
3	Tinnu Anand	16	2

UC#5 Top 3 most popular genres based on country using a dense rank to break ties

	country text	genre text	total_shows bigint	top_genre_ranking bigint
1	AFGHANISTAN	international movies	1	1
2	AFGHANISTAN	documentaries	1	1
3	ALBANIA	international movies	1	1
4	ALBANIA	dramas	1	1
5	ALGERIA	dramas	3	1
6	ALGERIA	international movies	3	1
7	ALGERIA	classic movies	1	2
8	ALGERIA	independent movies	1	2
9	ANGOLA	international movies	1	1
10	ANGOLA	action & adventure	1	1
11	ARGENTINA	international movies	58	1
12	ARGENTINA	dramas	35	2
13	ARGENTINA	spanish-language tv shows	18	3
14	ARMENIA	international movies	1	1
15	ARMENIA	documentaries	1	1
16	AUSTRALIA	dramas	38	1
17	AUSTRALIA	international tv shows	31	2
18	AUSTRALIA	international movies	30	3