prime video

Data analysis using sql

-By Mohammad Arshad

mdarshad16

Contents

mdarshad16

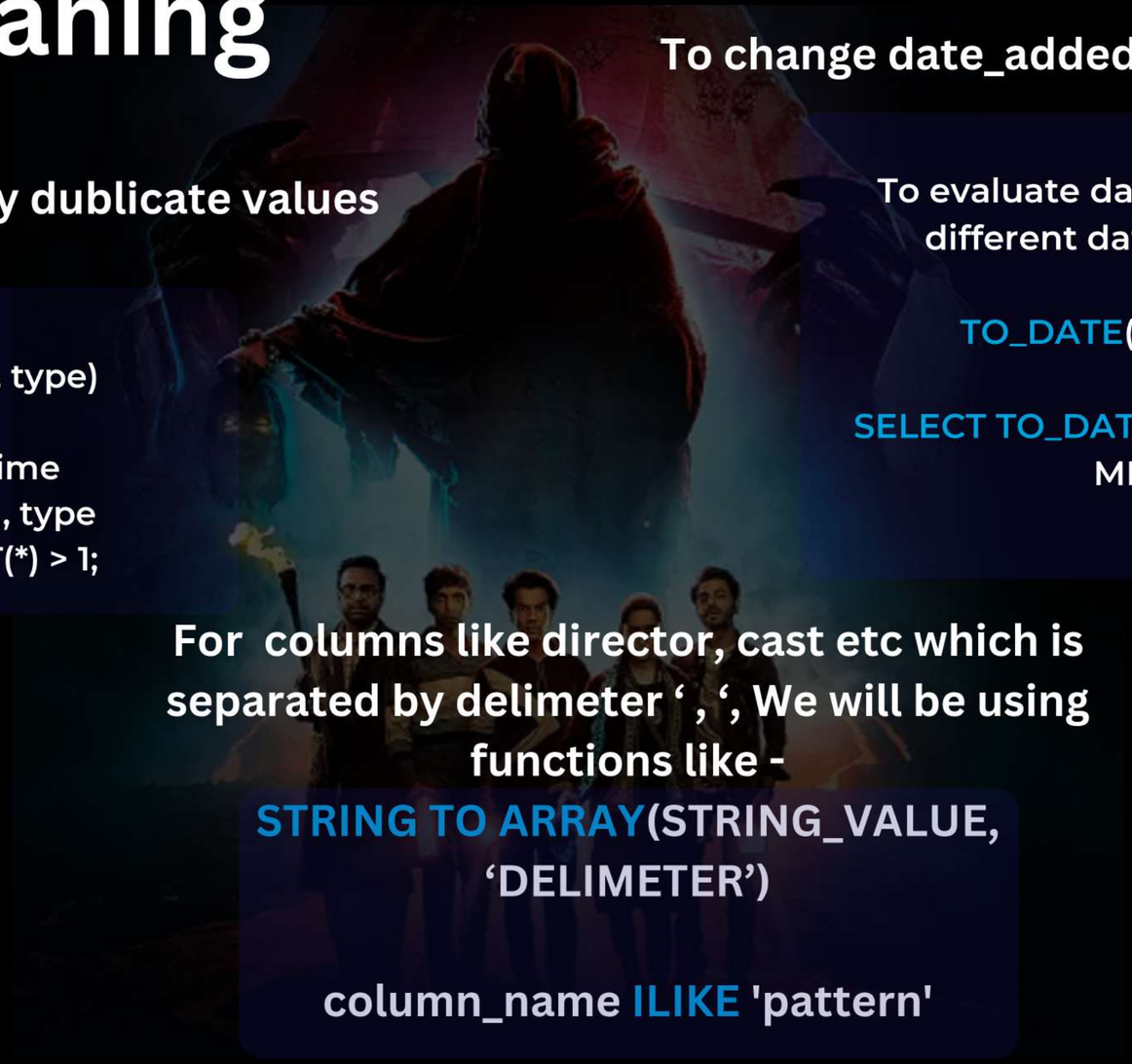
Data Structure

I downloaded this dataset from Kaggle that includes a table named amazon_prime. This table had 12 columns, which contained information such as titles, types (movies or TV shows), release years, ratings, comments, etc. The dataset had a substantial amount of data, allowing me to perform meaningful analysis and gain insights into various aspects of Amazon Prime content. It provided a great opportunity to explore trends and patterns in the types of content available on the platform.

Journey bengins

```
CREATE TABLE amazon_prime (
show_id VARCHAR(6),
type VARCHAR(10),
title VARCHAR(350),
director VARCHAR(400),
casts VARCHAR(1000),
country VARCHAR(200),
date_added VARCHAR(50),
release_year INT,
rating VARCHAR(15),
duration VARCHAR(15),
listed_in VARCHAR(100),
description VARCHAR(1000)
)
```

Data cleaning



Checking out for any dublicate values

```
SELECT
CONCAT(title, type)
FROM
amazon_prime
GROUP BY title , type
HAVING COUNT(*) > 1;
```

To change date_added to correct datatype

To evaluate date, we will be using different date functions like

```
TO_DATE(string, format)

SELECT TO_DATE('2023-10-14', 'YYYY-MM-DD');
```

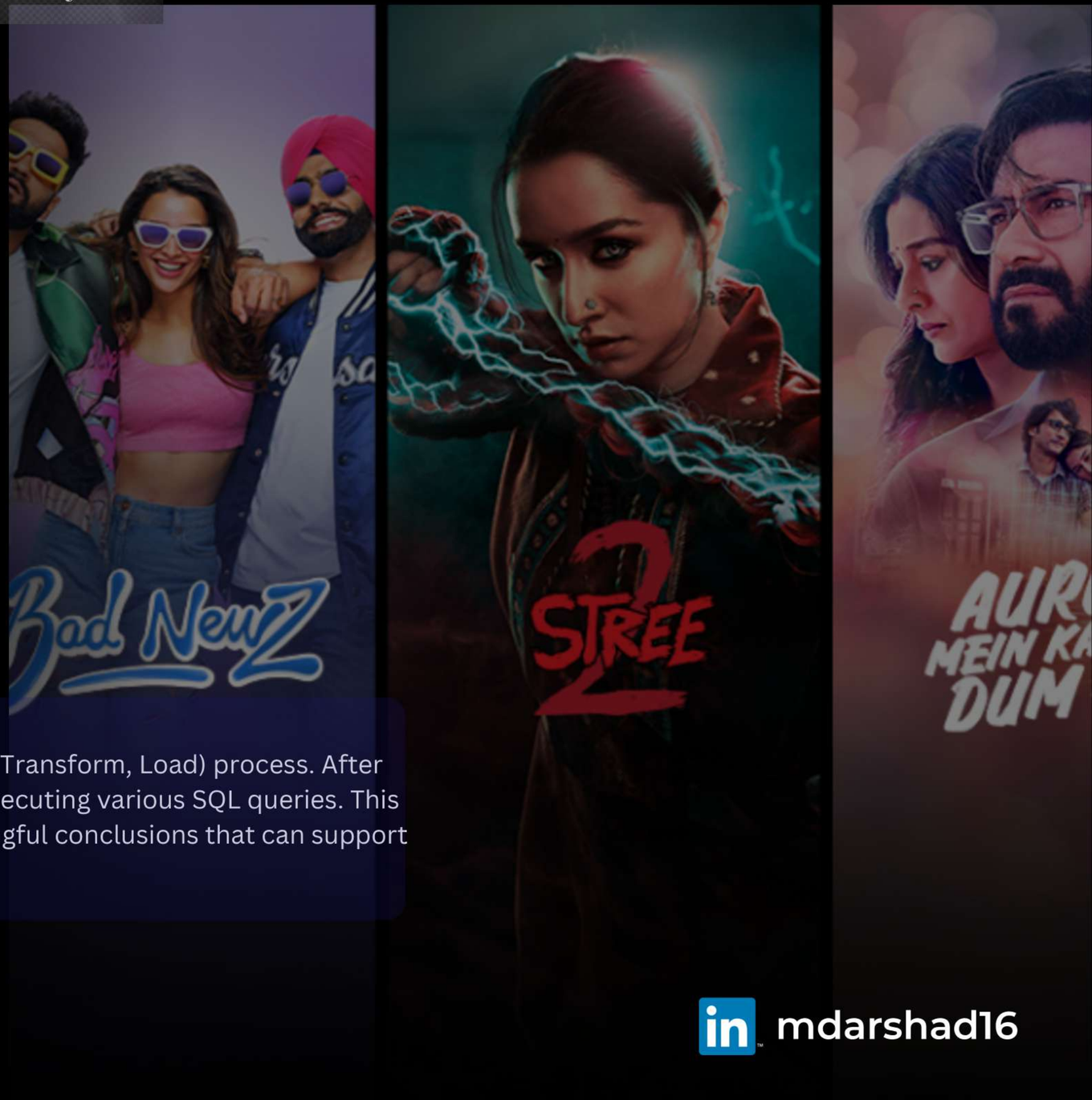
For columns like director, cast etc which is separated by delimeter ‘ , ‘, We will be using functions like -

```
STRING TO ARRAY(STRING_VALUE, ‘DELIMETER’)

column_name ILIKE 'pattern'
```

Data Analysis

We have reached the most exciting part of the process: analyzing the data to extract valuable insights. This analysis will help answer key questions from our stakeholders, providing them with the information they need to make informed decisions. By uncovering trends, patterns, and anomalies in the data, we can offer actionable recommendations that will guide strategic planning and improve overall outcomes. This step is crucial for ensuring that our stakeholders are well-equipped to navigate their challenges and capitalize on opportunities.



I utilized PostgreSQL to load the data and perform the ETL (Extract, Transform, Load) process. After preparing the data, I am proceeding to extract valuable insights by executing various SQL queries. This approach allows me to analyze the data effectively and derive meaningful conclusions that can support decision-making.

Query: count the number of movies vs tv shows

Purpose: It informs content strategy and investment decisions

```
SELECT
COUNT(show_id), type
FROM
amazon_prime
GROUP BY type
ORDER BY COUNT(show_id) DESC;
```

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

Find the most common ratings for TV shows and movies

Purpose : Identifying the most common ratings for TV shows and movies helps businesses understand viewer preferences and trends. This insight can guide content creation and marketing strategies, improving audience engagement and satisfaction. Analyzing ratings also reveals which genres resonate most with viewers, informing future programming decisions.

```
select type, rating
from (
select type, rating,
count(rating),
rank() over (partition by
type order by count(rating)
desc) as rank
from amazon_prime
group by type, rating
) as t1
where rank = 1;
```

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

list all movies list in a specific year (e.g 2020)

Purpose: Listing all movies from a specific year helps analyze industry trends, such as genre popularity and audience preferences. This information supports informed recommendations, identifies patterns, and aids in forecasting future trends for strategic decision-making in content acquisition and production.

```
SELECT
    *
FROM
    amazon_prime
WHERE
    release_year = 2020
AND
type = 'Movie';
```

First 6

show_id character varying (6)	type character varying (10)	title character varying (350)	director character varying (400)
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson
s17	Movie	Europe's Most Dangerous Man: Otto Skorzeny in Spain	Pedro de Echave García, Pablo Azorín Williams
s79	Movie	Tughlaq Durbar	Delhiprasad Deenadayalan
s85	Movie	Omo Ghetto: the Saga	JJC Skillz, Funke Akindele
s104	Movie	Shadow Parties	Yemi Amodu
s120	Movie	Here and There	JP Habac

Identify the top 5 countries with the most content on Amazon Prime.

Purpose: Understanding which countries have the most content on Amazon Prime allows businesses to tailor content acquisition strategies, optimize marketing efforts, and identify opportunities for market expansion, ultimately enhancing viewer engagement and competitive positioning.

```
SELECT

TRIM(UNNEST(STRING_TO_ARRAY
(country, ','))) AS new_country,
    COUNT(show_id) AS contents
FROM
    amazon_prime
GROUP BY 1
ORDER BY contents DESC
LIMIT 5;
```

new_country text	contents bigint
United States	3690
India	1046
United Kingdom	806
Canada	445
France	393

Identify the longest movie

Purpose: Finding the longest movie helps businesses analyze audience preferences for film length, informing content acquisition and production decisions. This insight can enhance marketing strategies and improve viewer engagement.

```
SELECT
    *
FROM
    amazon_prime
WHERE
    type = 'Movie'
        AND duration = (SELECT
            MAX(duration)
        FROM
            amazon_prime);
```

First 6

show_id character varying (6)	type character varying (10)	title character varying (350)	director character varying (400)	casts character varying (1000)
s52	Movie	InuYasha the Movie 2: The Castle Beyond the Looking Glass	Toshiya Shinohara	Kappei Yamaguchi, Satsuki Yuk
s53	Movie	InuYasha the Movie 3: Swords of an Honorable Ruler	Toshiya Shinohara	Kappei Yamaguchi, Satsuki Yuk
s120	Movie	Here and There	JP Habac	Janine Gutierrez, JC Santos, Vi
s338	Movie	Good Luck Chuck	Mark Helfrich	Dane Cook, Jessica Alba, Dan F
s345	Movie	My Girl 2	Howard Zieff	Anna Chlumsky, Austin O'Brien,
s427	Movie	Cousins	Ainsley Gardiner, Briar Grace-Smith	Rachel House, Briar Grace-Smit

List all movies added in the last 5 years

Analyzing movies added in the last 5 years helps businesses identify trends in content creation, viewer preferences, and market demands, guiding strategic decisions for acquisitions and marketing efforts to enhance audience engagement.

```
select *
from
   amazon_prime
where
  to_date(date_added, 'month dd,
year') >=
current_date - interval
'5 years';
```

show_id character varying (6)	type character varying (10)	title character varying (350)
s1	Movie	Dick Johnson Is Dead
s2	TV Show	Blood & Water
s3	TV Show	Ganglands
s4	TV Show	Jailbirds New Orleans
s5	TV Show	Kota Factory

Find all movies and TV shows by director Rajiv Chilaka

Purpose: Identifying works by Rajiv Chilaka helps businesses understand his influence on content and viewer preferences, guiding targeted marketing strategies and potential collaborations for future projects.

```
select
director,
    count(title)
from amazon_prime
   where director
   ilike '%Rajiv Chilaka%'
   group by director;
```

director character varying (400)	count bigint
Rajiv Chilaka	19
Rajiv Chilaka, Anirban Majumder, Alka Amarkant Dub...	1
Rajiv Chilaka, Binayak Das	1
Rajiv Chilaka, Owll Mina	1

List all TV shows with more than 5 seasons

Purpose: Identifying long-running TV shows helps businesses gauge viewer engagement and content success, guiding acquisition and marketing strategies to enhance audience retention.

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

First 6

character varying (6)	type character varying (10)	title character varying (350)	director character varying (400)	casts character varying (1000)
s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Hollywood
s56	TV Show	Nailed It	[null]	Nicole Byer, Jacques Torres
s66	TV Show	Numberblocks	[null]	Beth Chalmers, David Holt, Marcel McCalla, Teresa Gallagher
s68	TV Show	Saved by the Bell	[null]	Mark-Paul Gosselaar, Tiffani Thiessen, Mario Lopez, Lark Voorhies
s83	TV Show	Lucifer	[null]	Tom Ellis, Lauren German, Kevin Alejandro, D.B. Woodside, Lesley-
s275	TV Show	Grace and Frankie	[null]	Jane Fonda, Lily Tomlin, Martin Sheen, Sam Waterston, June Diane

count the number of contents item in each genre

Purpose: Finding the longest movie helps businesses analyze audience preferences for film length, informing content acquisition and production decisions. This insight can enhance marketing strategies and improve viewer engagement.

```
SELECT
    COUNT(show_id),

TRIM(UNNEST(STRING_TO_ARR
AY(listed_in, ','))) AS genre
FROM
    amazon_prime
GROUP BY genre
```

1351	International TV Shows
219	Sports Movies
102	LGBTQ Movies
69	Teen TV Shows
375	Music & Musicals
16	TV Shows

Find each year and the average number of content released by Amazon prime India, returning the top 5 years with the highest average release content

Purpose: Analyzing the average annual content released by Amazon prime India helps businesses understand trends in content production and viewer engagement, guiding strategic decisions on acquisitions and marketing efforts to align with audience expectations and optimize performance in the Indian market.

release_year integer	country character varying (200)	show_count bigint
2017	India	101
2018	India	94
2019	India	87
2020	India	75
2016	India	73

```
SELECT release_year,
    country,
    show_count,
    AVG(show_count) OVER () AS
average_show_count
FROM (
  SELECT release_year,
      country,
       COUNT(show_id) AS
show_count
  FROM amazon_prime
  WHERE country = 'India'
  GROUP BY release_year,
country
) AS yearly_counts
ORDER BY show_count DESC
LIMIT 5;
```

list all the movies that are documentaries

Identifying documentary films helps businesses understand viewer interest in real-life stories and factual content, guiding content acquisition and marketing strategies to cater to audiences seeking informative and engaging programming.

```
select *
 from
amazon_prime
  where  listed_in
ilike
'%Documentaries
%';
```

show_id character varying (6)	type character varying (10)	title character varying (350)	director character varying (400)
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson
s17	Movie	Europe's Most Dangerous Man: Otto Skorzeny in Spain	Pedro de Echave García, Pa
s46	Movie	My Heroes Were Cowboys	Tyler Greco
s69	Movie	Schumacher	Hanns-Bruno Kammertöns,
s89	Movie	Blood Brothers: Malcolm X & Muhammad Ali	Marcus Clarke
s92	Movie	The Women and the Murderer	Mona Achache, Patricia To

find all contents without a director

Purpose: Identifying content without a director helps businesses pinpoint gaps in their catalog, informing acquisition strategies to enhance the quality and credibility of their offerings while ensuring a more comprehensive viewing experience for audiences.

```
SELECT
    *
FROM
    amazon_prime
WHERE
    director IS NULL
```

show_id character varying (6)	type character varying (10)	title character varying (350)	director character varyi
s2	TV Show	Blood & Water	[null]
s4	TV Show	Jailbirds New Orleans	[null]
s5	TV Show	Kota Factory	[null]
s11	TV Show	Vendetta: Truth, Lies and The Mafia	[null]
s15	TV Show	Crime Stories: India Detectives	[null]

How many movies has actor Salman Khan appeared in over the last 10 years?

Purpose: Assessing Salman Khan's recent film appearances helps businesses evaluate his popularity and inform casting and marketing strategies to attract audiences and maximize box office success.

```
select title, casts,
date_added
from amazon_prime
where casts Ilike
'%Salman khan%'
and
release_year >
extract(YEAR from
current_date)-10
```

title character varying (350)	casts character varying (1000)
Prem Ratan Dhan Payo	Salman Khan, Sonam Kapoor, Anu
Paharganj	Lorena Franco, Bijesh Jayarajan, N

top 10 actors who have appeared in highest number of movies produced in India

Purpose: Identifying the top actors by film appearances helps businesses understand talent popularity and market trends, guiding casting decisions and marketing strategies to enhance audience engagement and boost box office performance.
4o mini

total_content bigint	actors text
40	Anupam Kher
34	Shah Rukh Khan
31	Naseeruddin Shah
29	Akshay Kumar
29	Om Puri

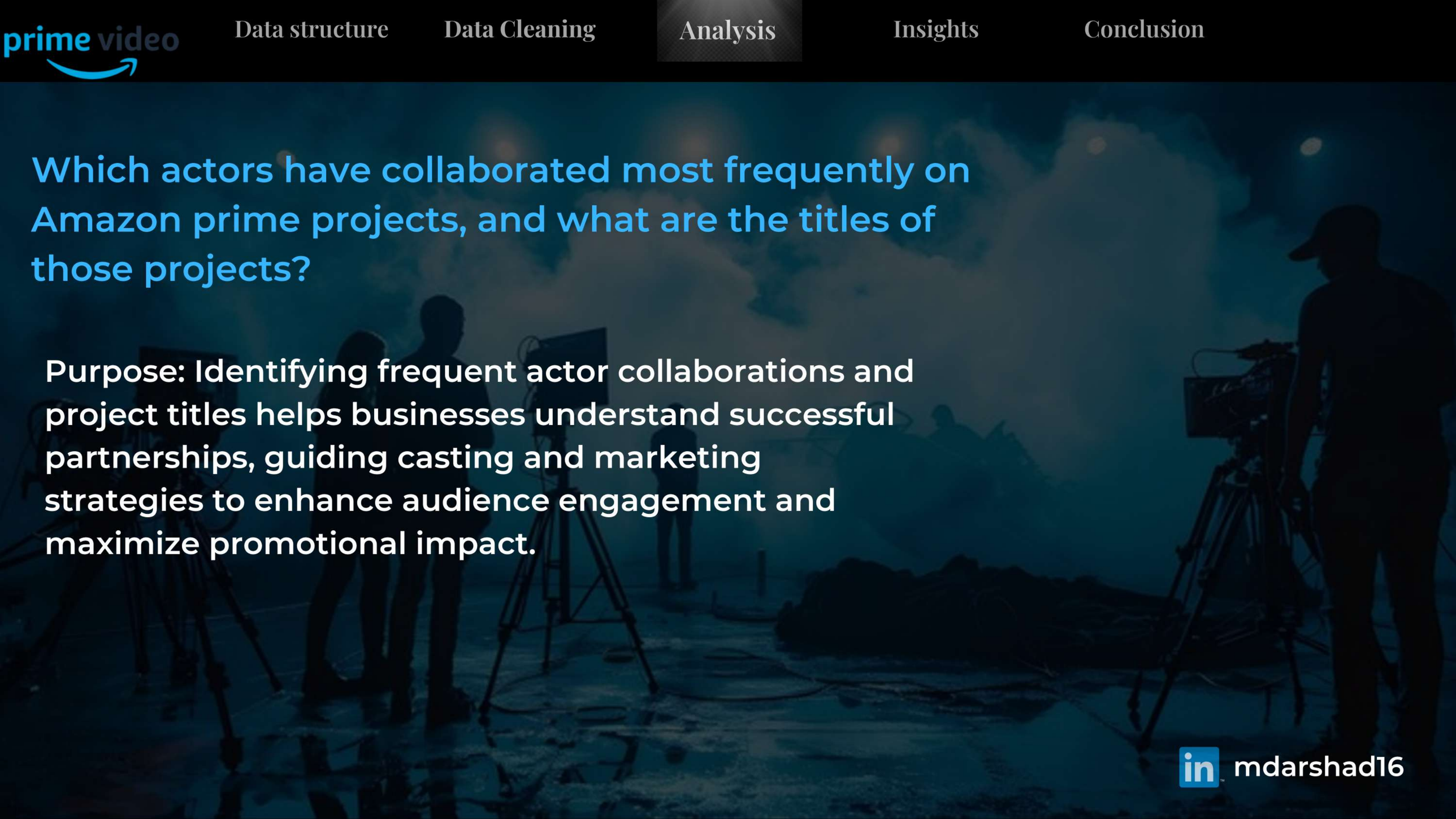
```
select count(show_id)
as total_content,
trim(unnest(string_to
_array(casts, ',')))  as
actors
from amazon_prime
where country Ilike
'%India%'
group by actors
order by
total_content desc
limit 5;
```

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.

Purpose: Categorizing content based on specific keywords helps businesses assess the nature of their offerings, informing content moderation and marketing strategies. This analysis can guide decisions on promoting or restricting certain types of content to align with audience preferences and brand values.

count bigint	category text
8465	Good
342	Bad

```
with cte_table as(
select * ,
case when description
Ilike '%kill%'
or description Ilike
'%violence%' then 'Bad'
else 'Good'
end category
from amazon_prime)
select count(*),
category
from cte_table
group by category;
```



Which actors have collaborated most frequently on Amazon prime projects, and what are the titles of those projects?

Purpose: Identifying frequent actor collaborations and project titles helps businesses understand successful partnerships, guiding casting and marketing strategies to enhance audience engagement and maximize promotional impact.

```
with actor_list as (
select title, trim(unnest(string_to_array(casts,
',')))  as actor
from amazon_prime),
actor_pairs as
(select
a1.actor as actor1,
a2.actor as actor2,
count(*) as collaboration_count,
   STRING_AGG(DISTINCT a1.title, ', ') AS
project_titles
from actor_list a1
join actor_list a2 on a1.title = a2.title and
a1.actor < a2.actor
group by a1.actor, a2.actor
)
select actor1, actor2, collaboration_count,
project_titles
from actor_pairs
ORDER BY
   collaboration_count DESC;
```

actor1 text	actor2 text	collaboration_count bigint	project_titles text
Julie Tejwani	Rupa Bhimani	31	Antariksha Ke Rakhwale, Chhota Bheem
Julie Tejwani	Rajesh Kava	24	Chhota Bheem, Chhota Bheem - Neeli Pa
Rajesh Kava	Rupa Bhimani	22	Chhota Bheem, Chhota Bheem - Neeli Pa
Jigna Bhardwaj	Julie Tejwani	21	Chhota Bheem, Chhota Bheem - Neeli Pa
Jigna Bhardwaj	Rajesh Kava	20	Chhota Bheem, Chhota Bheem - Neeli Pa

Which genres tend to have the highest average ratings, and how do these ratings vary by country?

Purpose: Analyzing which genres have the highest average ratings and how these ratings vary by country helps identify audience preferences and trends. This insight can inform content acquisition strategies, marketing efforts, and production decisions to better cater to viewer interests in specific regions.

mdarshad16

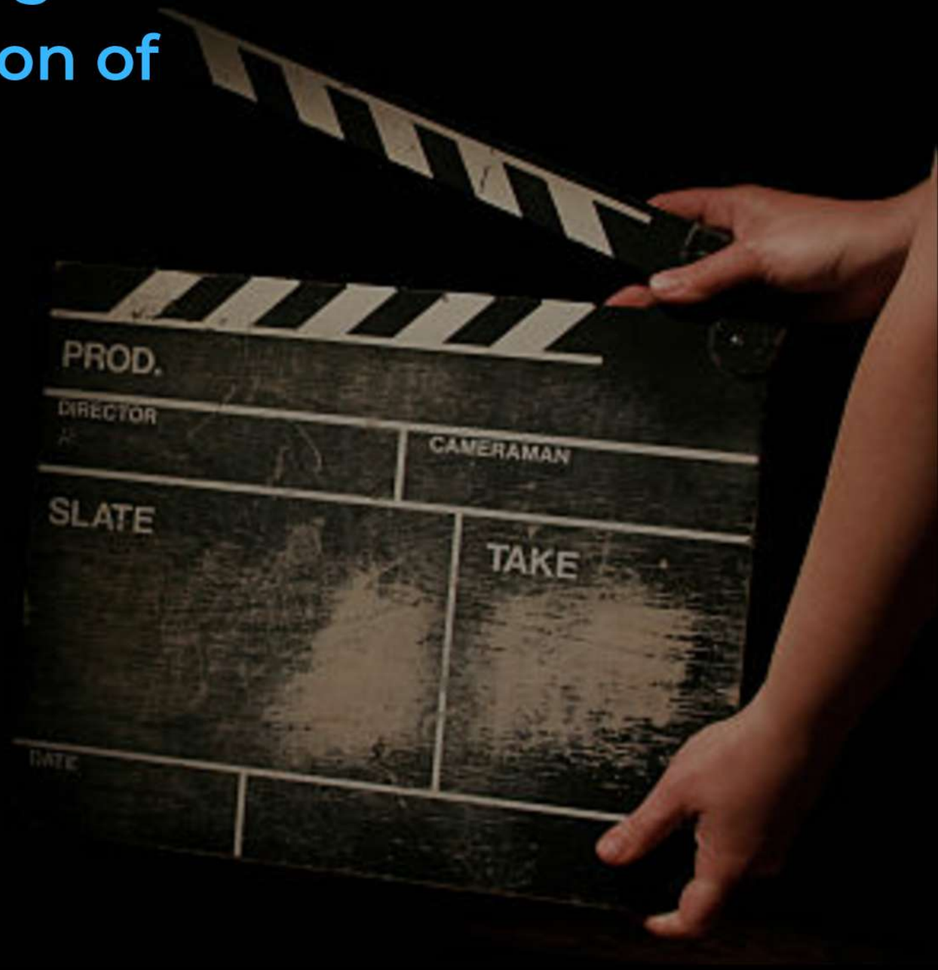
```sql
WITH genre_country AS (
  SELECT
    TRIM(UNNEST(STRING_TO_ARRAY(listed_in, ','))) AS genre,
    TRIM(UNNEST(STRING_TO_ARRAY(country, ','))) AS country,
    CASE
      WHEN rating = 'G' THEN 1
      WHEN rating = 'PG' THEN 2
      WHEN rating = 'PG-13' THEN 3
      WHEN rating = 'R' THEN 4
      WHEN rating = 'NC-17' THEN 5
      ELSE NULL
    END AS rating_value
  FROM amazon_prime
),
avg_rating AS (
  SELECT
    genre,
    country,
    AVG(rating_value) AS avg_rating
  FROM genre_country
  WHERE rating_value IS NOT NULL
  GROUP BY genre, country
)

SELECT
  country,
  genre,
  avg_rating
FROM avg_rating
ORDER BY avg_rating DESC;
```

country text	genre text	avg_rating numeric
Spain	International Movies	4.50
Belgium	Independent Movies	4.25
Colombia	[null]	4.00
Australia	Dramas	4.00
Bulgaria	Dramas	4.00
New Zealand	Dramas	4.00
Argentina	International Movies	4.00
Indonesia	[null]	4.00
Czech Republic	Thrillers	4.00

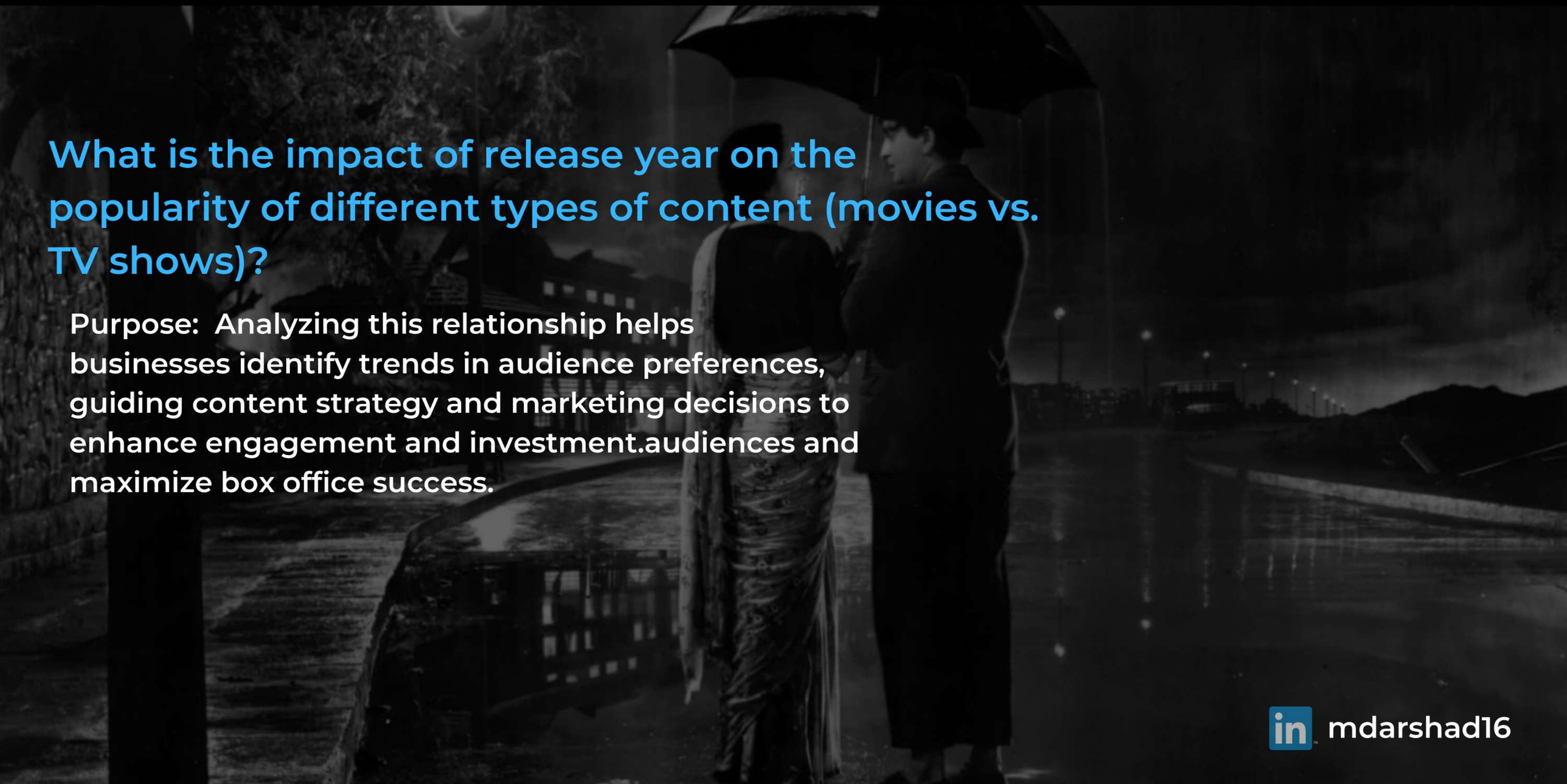
Which directors have shown versatility by working across multiple genres, and what is the distribution of their work?

Purpose: Analyzing directors' versatility across genres helps identify creative talents and influences in the industry, guiding strategic decisions for content development, partnerships, and marketing campaigns. This insight can enhance programming diversity and attract broader audiences.



```
select trim(director), genre,
count(*)as genre_count
from (
select
TRIM(UNNEST(STRING_TO_ARRAY
(director, ','))) AS director,
TRIM(UNNEST(STRING_TO_ARRAY
(listed_in, ','))) AS genre
from amazon_prime)
as subquery
where director is not null and
genre is not null
group by director, genre
having count(*) > 1
order by genre_count desc;
```

director text	genre text	genre_count bigint
Rajiv Chilaka	Children & Family Movies	22
Raúl Campos	Stand-Up Comedy	18
Suhas Kadav	Children & Family Movies	16
Marcus Raboy	Stand-Up Comedy	15
Jay Karas	Stand-Up Comedy	14
Jay Chapman	Stand-Up Comedy	11
Shannon Hartman	Stand-Up Comedy	9
Don Michael Paul	Action & Adventure	9
Hakan Algül	Comedies	8
Hanung Bramantyo	Dramas	8



What is the impact of release year on the popularity of different types of content (movies vs. TV shows)?

Purpose: Analyzing this relationship helps businesses identify trends in audience preferences, guiding content strategy and marketing decisions to enhance engagement and investment.audiences and maximize box office success.

```
select type,
release_year, count(title)
as Total_release
from amazon_prime
where type = 'Movie'
group by type,
release_year
order by count(title) asc;
```

From top

type character varying (10)	release_year integer	total_release bigint
Movie	1947	1
Movie	1966	1
Movie	1959	1
Movie	1963	1
Movie	1961	1

type character varying (10)	release_year integer	total_release bigint
TV Show	1925	1
TV Show	1946	1
TV Show	1991	1
TV Show	1945	1
TV Show	1974	1

```
select type,
release_year, count(title)
as Total_release
from amazon_prime
where type = 'TV Show'
group by type,
release_year
order by count(title) asc;
```

From bottom

Movie	2020	517
Movie	2019	633
Movie	2016	658
Movie	2017	767
Movie	2018	767

TV Show	2017	265
TV Show	2021	315
TV Show	2018	380
TV Show	2019	397
TV Show	2020	436

Insights

- **Content Distribution:** A comparison of the number of movies versus TV shows reveals viewer preferences for specific content types, guiding production decisions.
- **Rating Trends:** Identifying the most common ratings for TV shows and movies helps understand audience expectations and influences content development.
- **Yearly Releases:** Analyzing movies released in specific years highlights trends over time and can inform marketing strategies for new releases.
- **Geographical Insights:** The top countries with the most content on Amazon prime provide insights into regional content preferences, aiding localization strategies.
- **Longest Movies:** Identifying the longest movies may indicate trends toward epic storytelling and audience interest in immersive content.
- **Recent Additions**: A list of movies added in the last five years helps assess how content libraries evolve and adapt to viewer demands.
- **Director Contributions:** Finding content by specific directors (like Rajiv Chilaka) showcases individual contributions and can highlight potential for future collaborations.
- **Genre Versatility**: Directors working across multiple genres indicate creative flexibility and can help in curating diverse content offerings to attract varied audiences.

Conclusion

In conclusion, the insights derived from the analysis of various queries reveal significant patterns in content consumption and production. Understanding the distribution of movies versus TV shows, rating trends, and geographical preferences equips businesses with the knowledge needed to make informed strategic decisions. Additionally, identifying key contributors such as directors and actors highlights opportunities for future collaborations and content diversification. By leveraging these insights, organizations can enhance viewer engagement, optimize content libraries, and better align their offerings with audience preferences, ultimately driving growth and success in a competitive market.

mdarshad16

mdarshad16