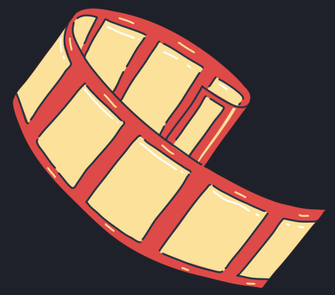
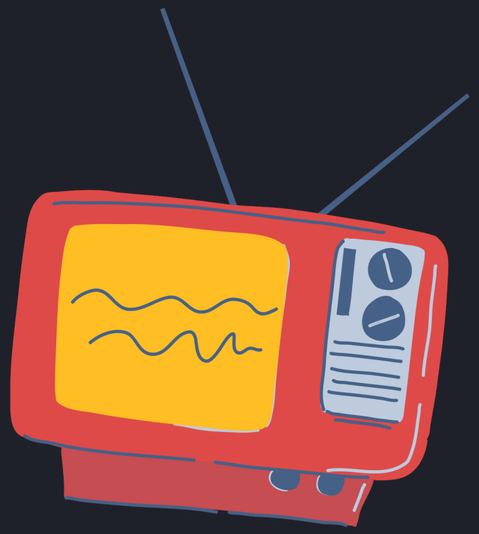
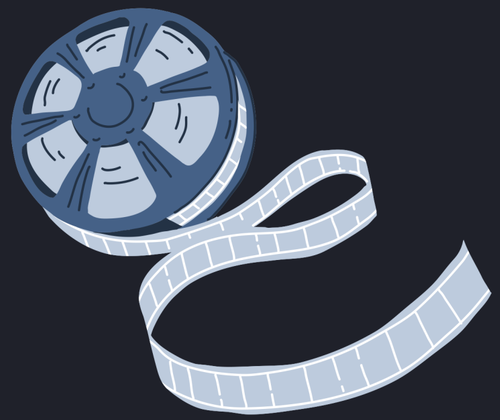
IMDB DATA ANALYSIS PROJECT








INTRODUCTION

In our IMDb data analysis project, we set sail into the expansive sea of data encapsulated within the IMDb dataset. This reservoir of information comprises diverse tables, offering intricate details about movies, directors, actors, genres, and ratings. Employing SQL as our compass, we navigate through this wealth of data to address a spectrum of advanced business questions. From discerning trends in genre preferences to evaluating the impact of directors and actors, our project aims to provide actionable insights for strategic decision-making.

DATASET

- **Director Mapping :** This table Associates directors with movies they have directed, aiding exploration of directorial impact.
- **Genre :** This table Classifies movies into different genres, enabling genre-based analysis of performance metrics.
- **Name :** This table Contains information about individuals involved in movie production, such as actors, directors, and producers.
- **Ratings :** This table Provides ratings and reviews for movies, allowing evaluation of critical and audience reception for each film.
- **Role Mapping :** This table Maps actors/actresses to their roles in movies, facilitating analysis of individual performance.

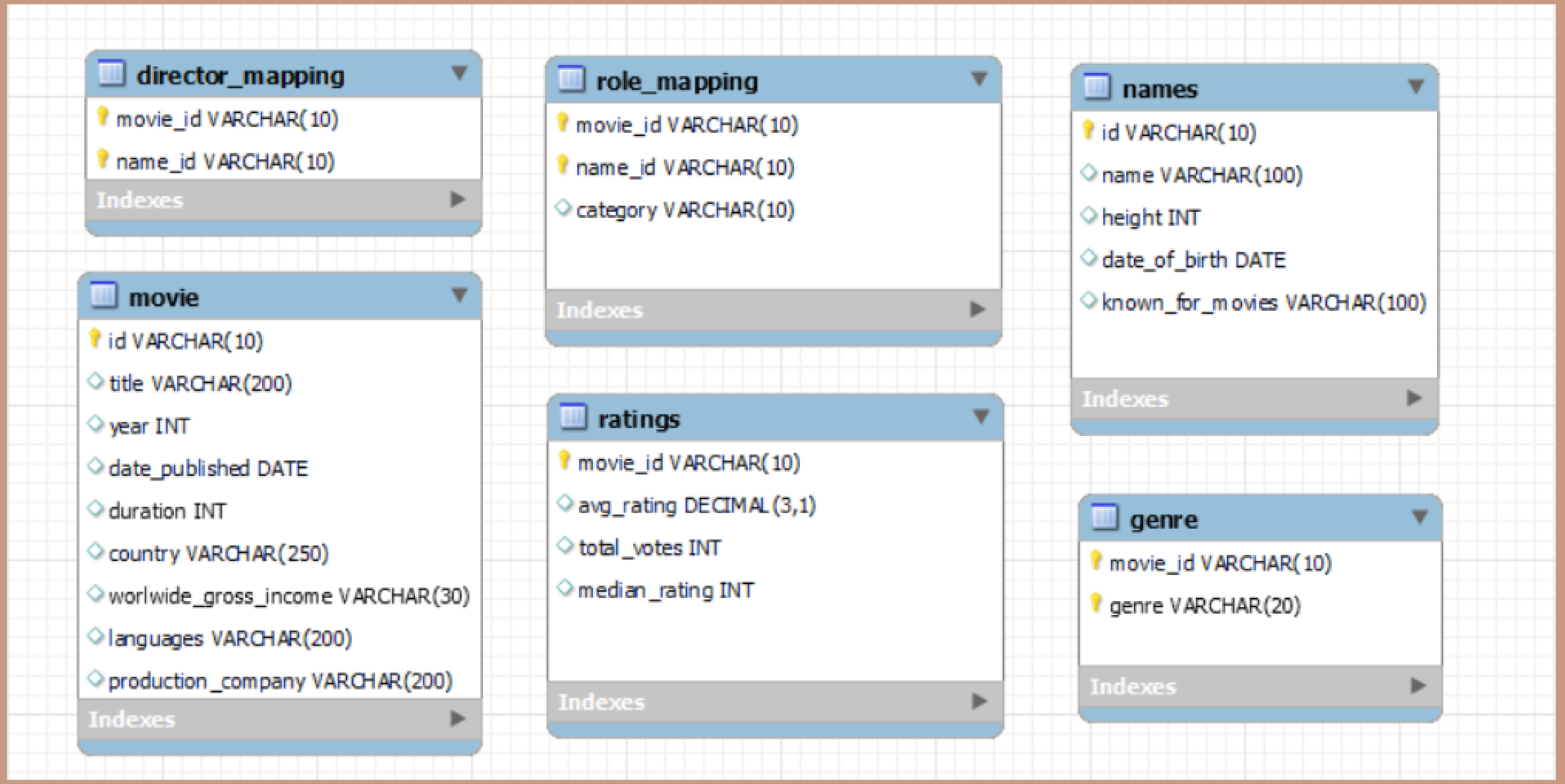
PROJECT GOALS

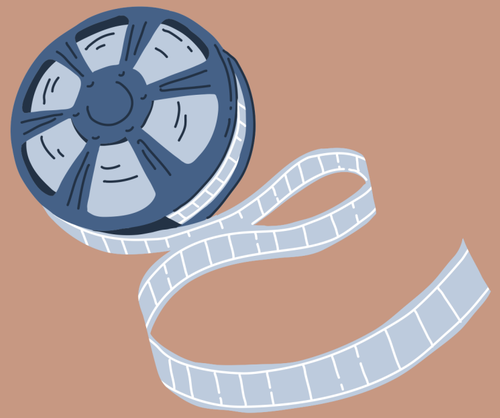
- **Addressing Complex Business Questions:** Utilize the IMDb dataset's breadth to dissect intricate inquiries such as the impact of genres, directors, and actors on movie success.

- **Uncovering Hidden Insights:** Employ SQL to dig deep into the dataset, unveiling subtle yet crucial patterns and correlations that might otherwise remain unnoticed.

- **Storytelling with Data:** Craft compelling narratives that seamlessly weave together the discovered insights, enabling stakeholders to grasp and act upon the implications effectively.

Entity Relationship Diagram





QUESTION 1

Find the total number of rows in each table of the schema?

```
select table_name, table_rows
from information_schema.tables
where table_schema='imdb';
```

QUERY

TABLE_NAME	TABLE_ROWS
director_mapping	3867
genre	14662
movie	8859
names	29523
ratings	7927
role_mapping	16437

RESULT

QUESTION 2

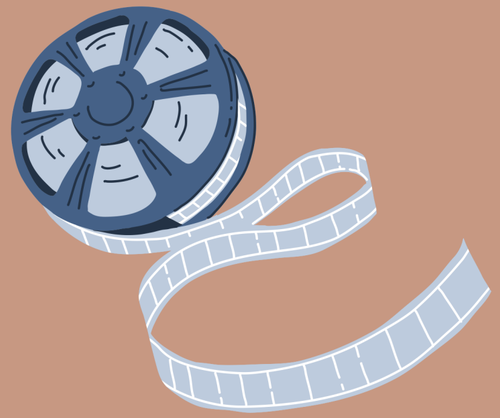
Which columns in the movie table have null values?

```sql
SELECT COLUMN_NAME
FROM information_schema.COLUMNS
WHERE TABLE_SCHEMA = 'imdb'
  AND TABLE_NAME = 'movie'
  AND IS_NULLABLE = 'YES';
```

COLUMN_NAME
title
year
date_published
duration
country
worlwide_gross_income
languages
production_company

QUERY

RESULT



QUESTION 3

Find the total number of movies released each year?

```
-- part 1
select year, count(title) as number_of_movies
from movie
group by 1 ;
```

QUERY

year	number_of_movies
2017	3052
2018	2944
2019	2001

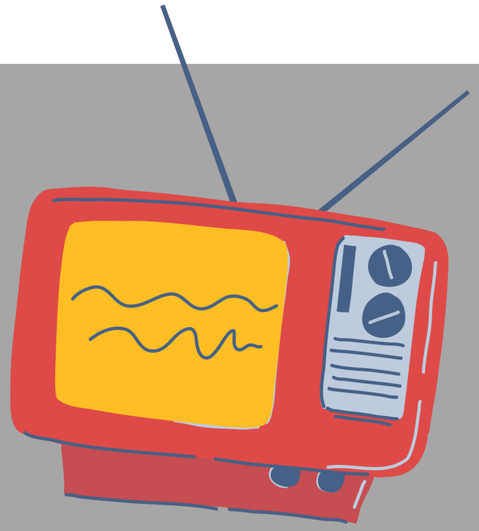
RESULT

QUESTION 4

How many movies were produced in the USA or India in the year 2019?

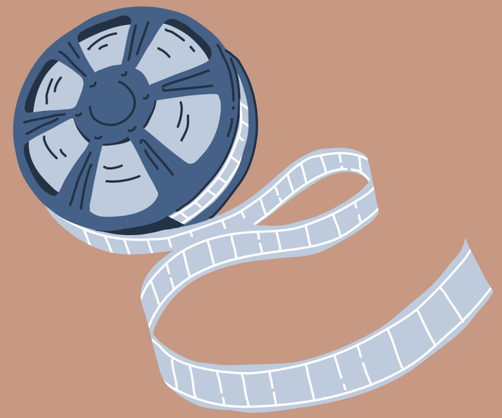
QUERY

```
select count(distinct id) as no_of_movies , year
from movie
where (country like '%USA%' or country like '%INDIA%') and year = 2019;
```



RESULT

	no_of_movies	year
▶	1059	2019



QUESTION 5

Find the unique list of the genres present in the data set?

```
SELECT DISTINCT genre
FROM genre;
```

QUERY

	genre
▶	Drama
	Fantasy
	Thriller
	Comedy
	Horror
	Family
	Romance
	Adventure

RESULT

QUESTION 6

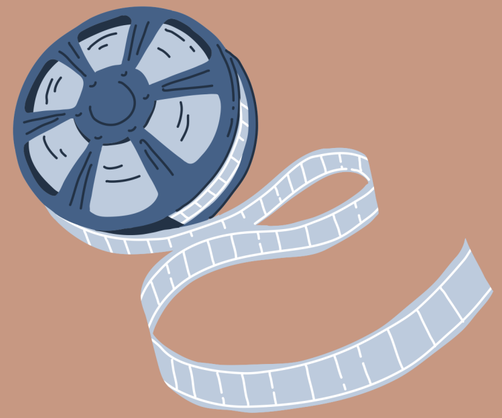
Which genre had the highest number of movies produced overall?

```sql
select genre, count(id) as no_of_movies
from movie m
join genre g
on m.id = g.movie_id
group by 1
order by no_of_movies desc
limit 1;
```

QUERY

	genre	no_of_movies
▶	Drama	4285

RESULT



QUESTION 7

How many movies belong to only one genre?

```
select count(id) as num_movies_with_one_genre
from movie
where id  in (select movie_id from genre
              group by movie_id
              having count(*) = 1 );
```

	num_movies_with_one_genre
▶	3289

QUERY

RESULT

QUESTION 8

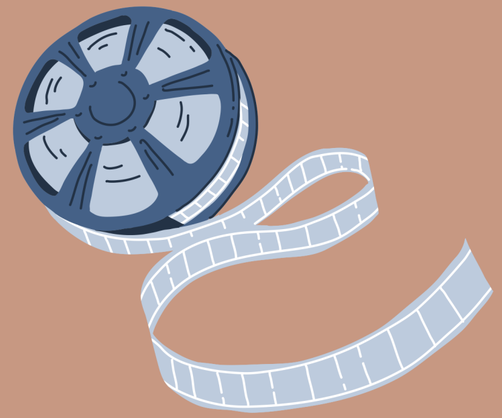
What is the average duration of movies in each genre?

```
SELECT genre, round(AVG(duration),2) AS average_duration
FROM movie
inner join genre
on genre.movie_id = movie.id
WHERE duration IS NOT NULL
GROUP BY genre
order by average_duration desc;
```

QUERY

genre	average_duration
Action	112.88
Romance	109.53
Crime	107.05
Drama	106.77
Fantasy	105.14
Comedy	102.62
Adventure	101.87
Mystery	101.80
Thriller	101.58
Family	100.97
Others	100.16
Sci-Fi	97.94
Horror	92.72

RESULT



QUESTION 9

What is the rank of the ‘thriller’ genre of movies among all the genres in terms of number of movies produced?

```sql
select *, rank() over (order by num_movies desc) as genre_rank
from (select genre, count(*) as num_movies
      from movie m
      join genre g
      on m.id = g.movie_id
      group by 1
      ) wp;
```

QUERY

genre	num_movies	genre_rank
Drama	4285	1
Comedy	2412	2
Thriller	1484	3
Action	1289	4
Horror	1208	5
Romance	906	6
Crime	813	7
Adventure	591	8
Mystery	555	9
Sci-Fi	375	10
Fantasy	342	11
Family	302	12
Others	100	13

RESULT

QUESTION 10

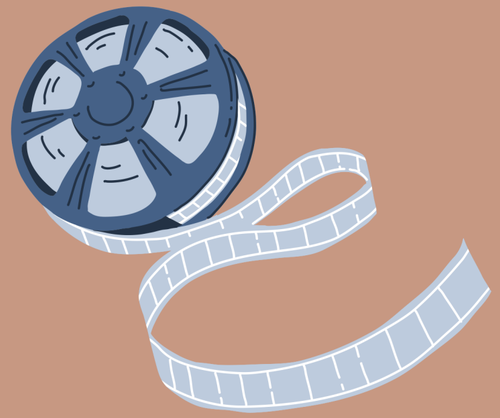
Find the minimum and maximum values in each column of the ratings table except the movie_id column

```
SELECT MIN(avg_rating),MAX(avg_rating),MIN(total_votes),MAX(total_votes),MIN(median_rating),MAX(median_rating)
FROM ratings;
```

QUERY

	MIN(avg_rating)	MAX(avg_rating)	MIN(total_votes)	MAX(total_votes)	MIN(median_rating)	MAX(median_rating)
▶	1.0	10.0	100	725138	1	10

RESULT



QUESTION 11

Which are the top 10 movies based on average rating?

```
select title, avg_rating,
dense_rank() over (order by avg_rating desc) as movie_rank
from movie m
join ratings r
on m.id = r.movie_id
limit 10;
```

QUERY

	title	avg_rating	movie_rank
▶	Kirket	10.0	1
	Love in Kilnerry	10.0	1
	Gini Helida Kathe	9.8	2
	Runam	9.7	3
	Fan	9.6	4
	Android Kunjappan Version 5.25	9.6	4
	Yeh Suhaagraat Impossible	9.5	5
	Safe	9.5	5
	The Brighton Miracle	9.5	5
	Shibu	9.4	6

RESULT

QUESTION 12

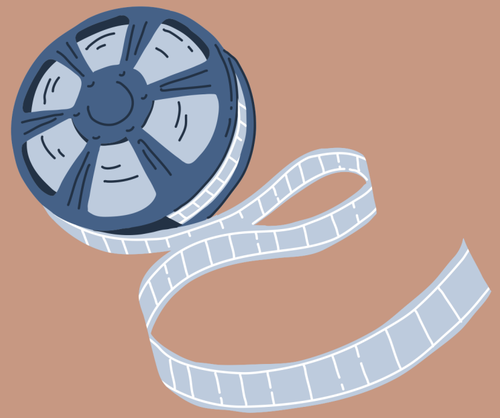
Summarise the ratings table based on the movie counts by median ratings.

```sql
select median_rating, count(movie_id) as movie_count
from ratings
group by 1
order by movie_count desc;
```

QUERY

median_rating	movie_count
7	2257
6	1975
8	1030
5	985
4	479
9	429
10	346
3	283
2	119
1	94

RESULT



QUESTION 13

Which production house has produced the most number of hit movies (average rating > 8)?

```sql
with cte as (
select production_company,
       count(movie_id) as movie_count
from movie m
join ratings r
on m.id = r.movie_id
where avg_rating > 8 and production_company is not null
group by 1
order by 2 desc)
select *, dense_rank() over (order by movie_count desc ) as prod_company_rank
from cte;
```

QUERY

production_company	movie_count	prod_company_rank
Dream Warrior Pictures	3	1
National Theatre Live	3	1
Lietuvos Kinostudija	2	2
Swadharm Entertainment	2	2
Panorama Studios	2	2
Marvel Studios	2	2
Central Base Productions	2	2
Painted Creek Productions	2	2
National Theatre	2	2
Colour Yellow Productions	2	2

RESULT

QUESTION 14

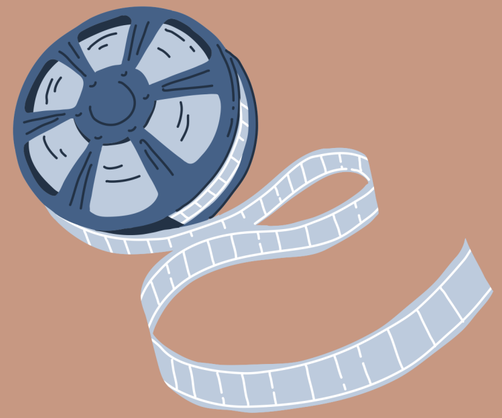
How many movies released in each genre during March 2017 in the USA had more than 1,000 votes?

```sql
SELECT genre,
       COUNT(*) AS movie_count
FROM movie m
join genre g ON m.id = g.movie_id
JOIN ratings r ON m.id = r.movie_id
WHERE m.country = 'USA'
  AND MONTH(m.date_published) = 3
  AND YEAR(m.date_published) = 2017
  AND r.total_votes > 1000
GROUP BY genre;
```

	genre	movie_count
▶	Action	4
	Comedy	8
	Crime	5
	Drama	16
	Fantasy	2
	Mystery	2
	Romance	3
	Sci-Fi	4
	Thriller	4
	Horror	5
	Family	1

QUERY

RESULT



QUESTION 15

Find movies of each genre that start with the word ‘The’ and which have an average rating > 8?

```
select title, avg_rating, genre
from movie m
join genre g
on m.id =  g.movie_id
join ratings r
on m.id = r.movie_id
where title like 'The%' and avg_rating > 8
order by 2 desc;
```

	title	avg_rating	genre
▶	The Brighton Miracle	9.5	Drama
	The Colour of Darkness	9.1	Drama
	The Blue Elephant 2	8.8	Drama
	The Blue Elephant 2	8.8	Horror
	The Blue Elephant 2	8.8	Mystery
	The Irishman	8.7	Crime
	The Irishman	8.7	Drama
	The Mystery of Godliness: The Sequel	8.5	Drama
	The Gambinos	8.4	Crime
	The Gambinos	8.4	Drama
	Theeran Adhigaaram Ondru	8.3	Action
	Theeran Adhigaaram Ondru	8.3	Crime
	Theeran Adhigaaram Ondru	8.3	Thriller
	The King and I	8.2	Drama
	The King and I	8.2	Romance

QUERY

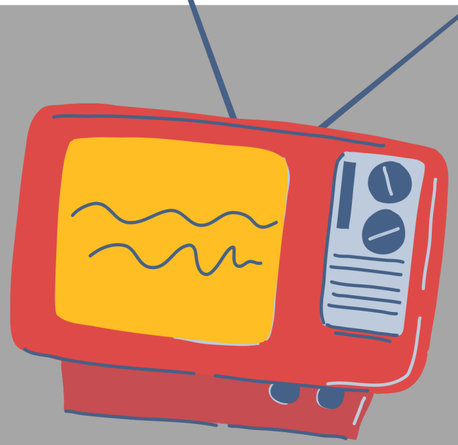
RESULT

QUESTION 16

The movies released between 1 April 2018 and 1 April 2019, how many were given a median rating of 8?

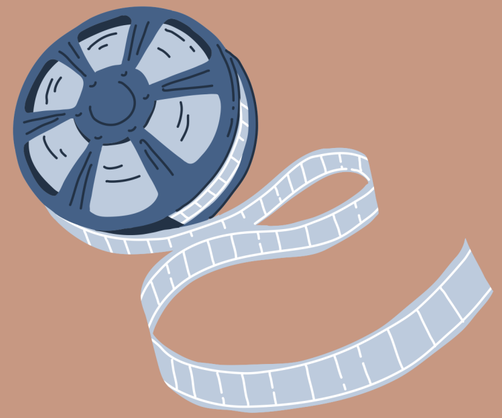
```sql
select median_rating, count(*) as movie_count
from movie m
join ratings r
on m.id = r.movie_id
where date_published between '2018-04-01' and '2019-04-01' and median_rating = 8
group by 1;
```

QUERY



RESULT

	median_rating	movie_count
▶	8	361



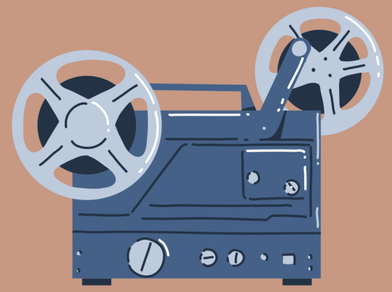
QUESTION 17

Do German movies get more votes than Italian movies?

```sql
SELECT 'Germany' AS country,SUM(total_votes) as Votes
FROM movie m
JOIN ratings r ON m.id = r.movie_id
WHERE m.country like "%Germany%"
UNION ALL
SELECT 'Italy',SUM(total_votes) as Votes
FROM movie m
JOIN ratings r ON m.id = r.movie_id
WHERE m.country like "%Italy%";
```

QUERY

	country	Votes
▶	Germany	2026223
	Italy	703024



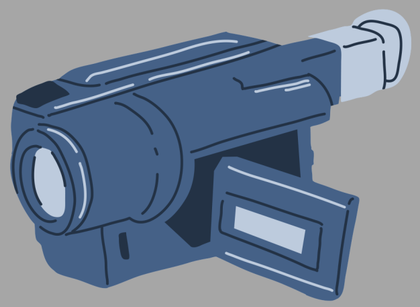
RESULT

QUESTION 18

Which columns in the names table have null values?

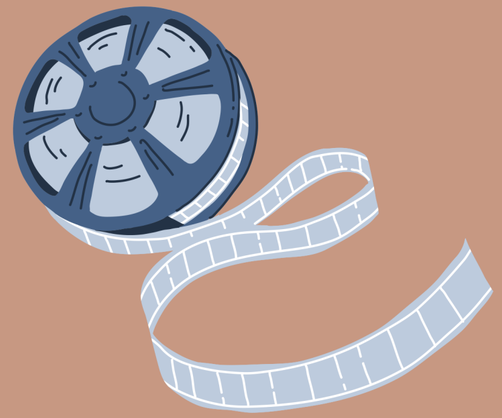
```sql
select
sum(case when name is null then 1 else 0 end) as name_nulls,
sum(case when height is null then 1 else 0 end) as height_nulls,
sum(case when date_of_birth is null then 1 else 0 end) as date_of_birth_nulls,
sum(case when known_for_movies is null then 1 else 0 end) as known_for_movies_nulls
from names;
```

QUERY



RESULT

	name_nulls	height_nulls	date_of_birth_nulls	known_for_movies_nulls
▶	0	17335	13431	15226



QUESTION 19

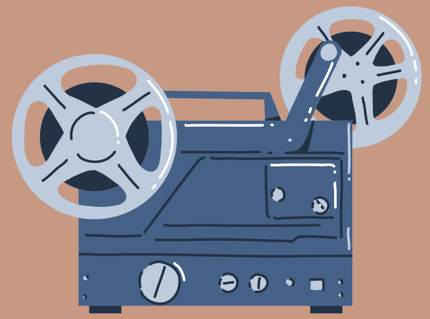
Who are the top three directors in the top three genres whose movies have an average rating > 8?

QUERY

```
with top_3_genre as
( select  genre,
   count(m.id) as movie_count,
   rank() over (order by count(m.id) desc ) as genre_rank
   from movie m
   inner join genre g
   on m.id = g.movie_id
   inner join ratings r
   on m.id = r.movie_id
   where avg_rating > 8
   group by 1
   limit 3 )
select n.name as director_name,
       count(d.movie_id) as movie_count
from director_mapping as d
inner join genre g
using (movie_id)
inner join names as n
on n.id = d.name_id
inner join top_3_genre
using (genre)
inner join ratings
using  (movie_id)
where  avg_rating > 8
group by name
order by  movie_count desc limit 3 ;
```

RESULT

director_name	movie_count
James Mangold	4
Anthony Russo	3
Soubin Shahir	3



QUESTION 20

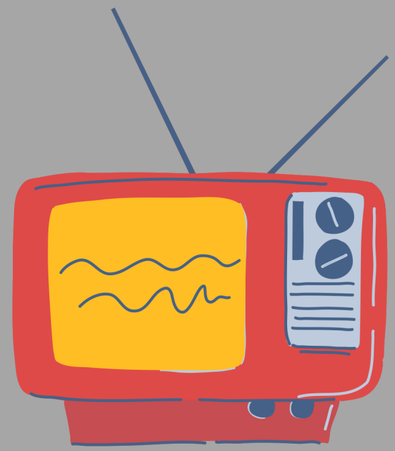
Who are the top two actors whose movies have a median rating >= 8?

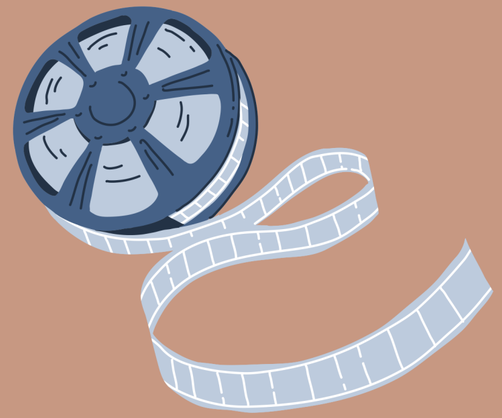
QUERY

```sql
select n.name as actor_name,
    count(movie_id) as movie_count
from role_mapping rm
inner join names n
on rm.name_id = n.id
inner join genre g
using (movie_id)
inner join ratings r
using (movie_id)
inner join movie m
on rm.movie_id = m.id
where category = 'actor' and r.median_rating >= 8
group by name
order by movie_count desc
limit 2;
```

RESULT

actor_name	movie_count
Mammootty	16
Mohanlal	13

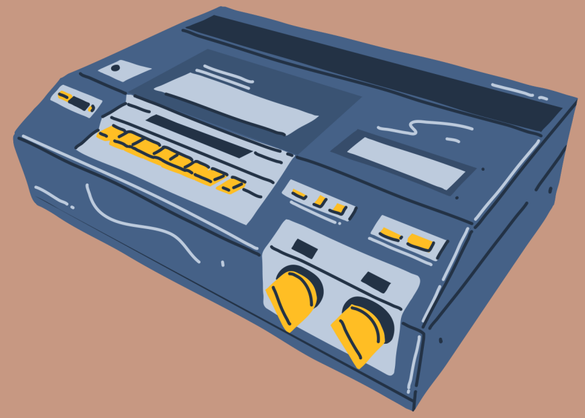




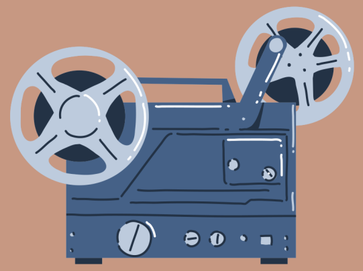
QUESTION 21

Which are the top three production houses based on the number of votes received by their movies?

```
select production_company,
sum(total_votes) as vote_count,
rank() over (order by sum(total_votes) desc) as prod_comp_rank
from movie m
inner join ratings r
on m.id = r.movie_id
group by 1
limit 3;
```



RESULT



QUERY

	production_company	vote_count	prod_comp_rank
▶	Marvel Studios	2656967	1
	Twentieth Century Fox	2411163	2
	Warner Bros.	2396057	3

QUESTION 22

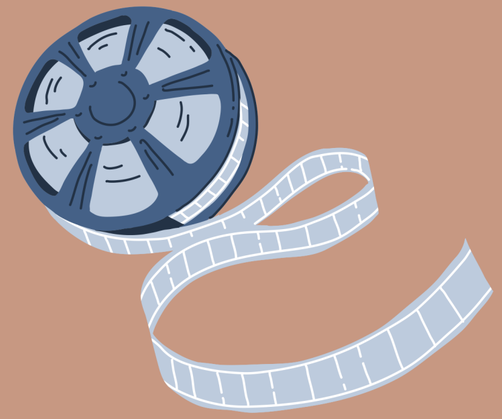
Rank actors with movies released in India based on their average ratings. Which actor is at the top of the list?

QUERY

```sql
with actor_summary as(
    select
        n.name as actor_name,
        sum(total_votes) as total_votes,
        count(r.movie_id) as movie_count,
        round(sum(avg_rating * total_votes) / sum(total_votes), 2) as actor_avg_rating
    from  movie as m
        inner join ratings as r on m.id = r.movie_id
        inner join role_mapping as rm on m.id = rm.movie_id
        inner join names as n on rm.name_id = n.id
    where
        category = 'ACTOR'
        and country = 'india'
    group by  1
    having movie_count >= 5
)
select
    *,
    rank() over (order by actor_avg_rating desc) as actor_rank
from actor_summary
limit 5;
```

RESULT

actor_name	total_votes	movie_count	actor_avg_rating	actor_rank
Vijay Sethupathi	23114	5	8.42	1
Fahadh Faasil	13557	5	7.99	2
Yogi Babu	8500	11	7.83	3
Joju George	3926	5	7.58	4
Ammy Virk	2504	6	7.55	5



QUESTION 23

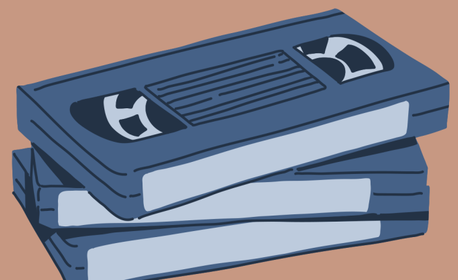
Find out the top five actresses in Hindi movies released in India based on their average ratings?

```
with actresses_summary as
(
  select n.name as actresses_name,
  sum(total_votes) as total_votes,
  count(r.movie_id) as movie_count,
  round(sum(avg_rating * total_votes) / sum(total_votes), 2) as actress_avg_rating
from movie as m
  inner join ratings as r on m.id = r.movie_id
  inner join role_mapping as rm on m.id = rm.movie_id
  inner join names as n on rm.name_id = n.id
  where category = 'Actress' and country = "INDIA" and languages like '%Hindi%'
  group by 1
  having movie_count >=3
  )
  select *,
  rank() over (order by actress_avg_rating desc) as actress_rank
  from actresses_summary;
```

QUERY

RESULT

actresses_name	total_votes	movie_count	actress_avg_rating	actress_rank
Taapsee Pannu	18061	3	7.74	1
Kriti Sanon	21967	3	7.05	2
Divya Dutta	8579	3	6.88	3
Shraddha Kapoor	26779	3	6.63	4
Kriti Kharbanda	2549	3	4.80	5
Sonakshi Sinha	4025	4	4.18	6

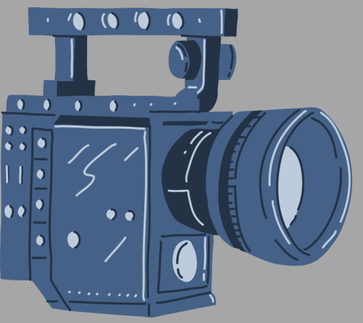


QUESTION 24

Select thriller movies as per avg rating and classify them in the following category:

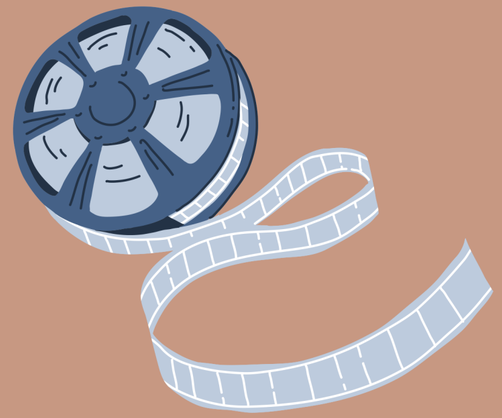
QUERY

```sql
with thriller_movies as
( select distinct title,
    avg_rating
    from movie m
    inner join ratings r on m.id = r.movie_id
    inner join genre g on m.id = g.movie_id
    where genre like 'Thriller'
)
  select *,
  case
     when avg_rating > 8 then 'Superhit movies'
     when avg_rating between 7 and 8 then 'Hit movies'
     when avg_rating between 5 and 7 then 'One-time-watch movies'
     else 'Flop movies'
     end as avg_rating_category
from thriller_movies;
```



RESULT

title	avg_rating	avg_rating_category
Der müde Tod	7.7	Hit movies
Fahrenheit 451	4.9	Flop movies
Pet Sematary	5.8	One-time-watch movies
Dukun	6.9	One-time-watch movies
Back Roads	7.0	Hit movies
Countdown	5.4	One-time-watch movies
Staged Killer	3.3	Flop movies
Vellaipookal	7.3	Hit movies
Uriyadi 2	7.3	Hit movies
Incitement	7.5	Hit movies
Rakshasudu	8.4	Superhit movies
Trois jours et ...	6.6	One-time-watch movies
Killer in Law	5.1	One-time-watch movies
Kalki	7.3	Hit movies



QUESTION 25

What is the genre-wise running total and moving average of the average movie duration?

```sql
SELECT genre,
        round(avg(duration),2) as avg_duration,
        sum(round(avg(duration),2)) over(order by genre rows unbounded preceding) as running_total_duration,
        avg(round(avg(duration),2)) over(order by genre rows 5 preceding) as moving_avg_duration
from movie as m
inner join genre as g
on m.id= g.movie_id
group by genre
order by genre;
```

QUERY

RESULT

genre	avg_duration	running_total_duration	moving_avg_duration
Action	112.88	112.88	112.880000
Adventure	101.87	214.75	107.375000
Comedy	102.62	317.37	105.790000
Crime	107.05	424.42	106.105000
Drama	106.77	531.19	106.238000
Family	100.97	632.16	105.360000
Fantasy	105.14	737.30	104.070000
Horror	92.72	830.02	102.545000
Mystery	101.80	931.82	102.408333
Others	100.16	1031.98	101.260000
Romance	109.53	1141.51	101.720000
Sci-Fi	97.94	1239.45	101.215000
Thriller	101.58	1341.03	100.621667

QUESTION 26

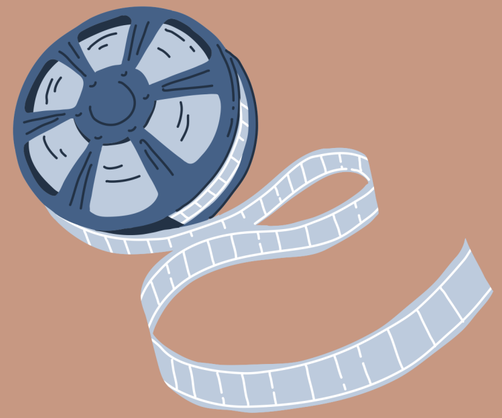
Which are the top two production houses that have produced the highest number of hits (median rating >= 8) among multilingual movies?

```sql
with production_company_summary as
( select production_company,
   count(*) as movie_count
   from movie m
   inner join ratings r on m.id = r.movie_id
   where median_rating >= 8 and production_company is not null
   and position(',' in languages ) > 0
   group by 1
)
 select *,
        rank() over( order by movie_count desc) as prod_comp_rank
 from production_company_summary
 limit 2;
```

QUERY

	production_company	movie_count	prod_comp_rank
▶	Star Cinema	7	1
	Twentieth Century Fox	4	2

RESULT



QUESTION 27

What is the genre-wise running total and moving average of the average movie duration?

```sql
with actresses_summary as
( select n.name as actress_name,
         sum(total_votes) as total_votes,
         count(r.movie_id) as movie_count,
         round(sum(avg_rating * total_votes) /sum(total_votes)) as actress_avg_rating
     from movie m
     inner join ratings r on m.id = r.movie_id
     inner join genre g using(movie_id)
     inner join role_mapping rm using(movie_id)
     inner join names n on rm.name_id = n.id
     where category = 'actress' and genre = 'Drama' and avg_rating > 8
     group by 1
)
 select * ,
     rank() over (order by movie_count desc ) as actress_rank
from actresses_summary
limit 3;
```

QUERY

RESULT

actress_name	total_votes	movie_count	actress_avg_rating	actress_rank
Parvathy Thiruvothu	4974	2	8	1
Susan Brown	656	2	9	1
Amanda Lawrence	656	2	9	1

THE END

RESOURCE PAGE