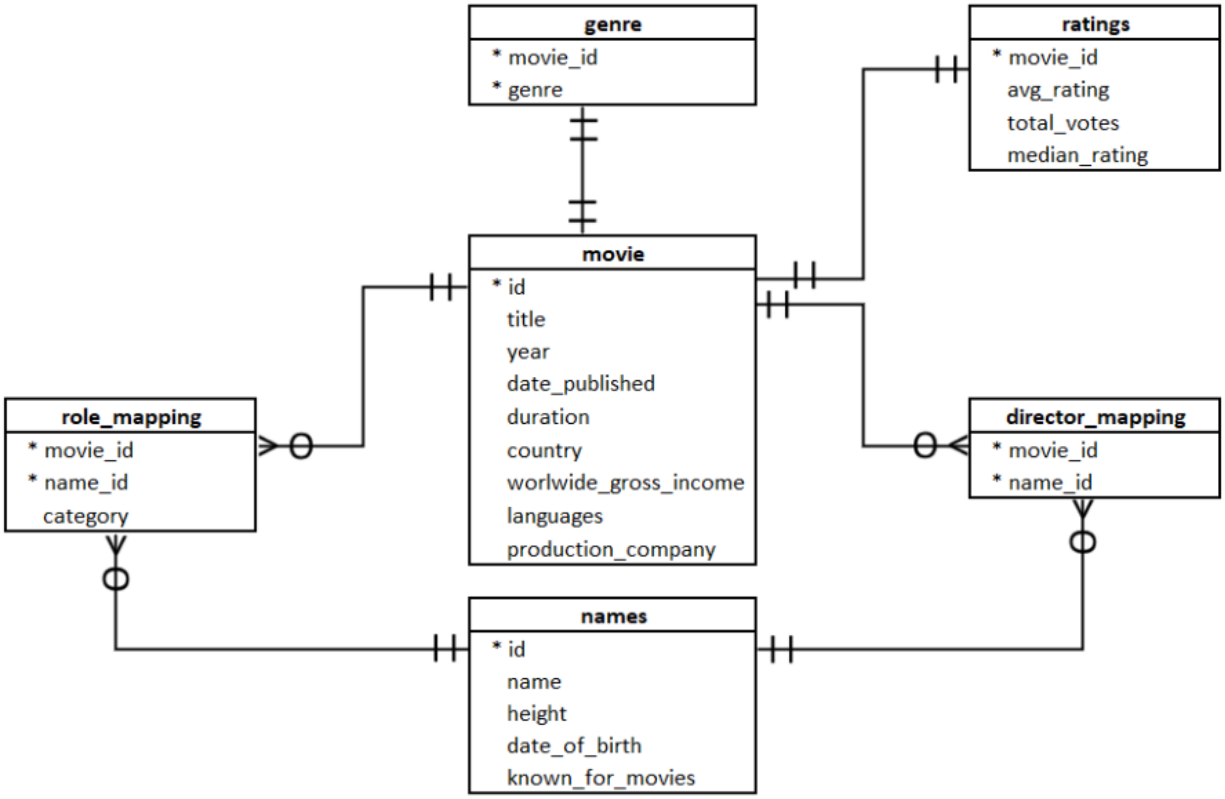
RSVP Movies - Case Study

Problem Introduction

RSVP Movies is an Indian film production company which has produced many super-hit movies. They have usually released movies for the Indian audience but for their next project, they are planning to release a movie for the global audience in 2022.
The production company wants to plan their every move analytically based on data and have approached you for help with this new project. You have been provided with the data of the movies that have been released in the past three years. You have to analyse the data set and draw meaningful insights that can help them start their new project.

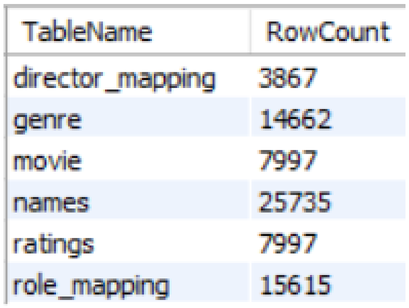


We have focused on 4 major segments:

1. ***Understanding the Data***
2. ***Box Office Performance Analysis***
3. ***Genre Insights***
4. ***Audience Rating Assessment***

1. Count number of rows for each column?

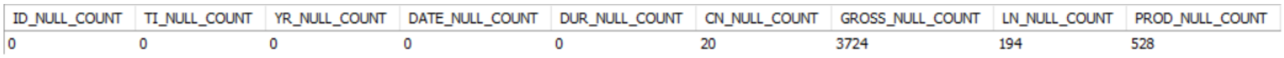
```
SELECT 'director_mapping' AS TableName, COUNT(*) AS RowCount FROM director_mapping
UNION ALL
SELECT 'genre' AS TableName, COUNT(*) AS RowCount FROM genre
UNION ALL
SELECT 'movie' AS TableName, COUNT(*) AS RowCount FROM movie
UNION ALL
SELECT 'names' AS TableName, COUNT(*) AS RowCount FROM names
UNION ALL
SELECT 'ratings' AS TableName, COUNT(*) AS RowCount FROM ratings
UNION ALL
SELECT 'role_mapping' AS TableName, COUNT(*) AS RowCount FROM role_mapping;
```



TableName	RowCount
director_mapping	3867
genre	14662
movie	7997
names	25735
ratings	7997
role_mapping	15615

2. Which columns in the movie table are having null values?

```
SELECT
SUM(CASE WHEN id IS NULL THEN 1 ELSE 0 END) AS ID_NULL_COUNT,
SUM(CASE WHEN title IS NULL THEN 1 ELSE 0 END) AS TI_NULL_COUNT,
SUM(CASE WHEN year IS NULL THEN 1 ELSE 0 END) AS YR_NULL_COUNT,
SUM(CASE WHEN date_published IS NULL THEN 1 ELSE 0 END) AS DATE_NULL_COUNT,
SUM(CASE WHEN duration IS NULL THEN 1 ELSE 0 END) AS DUR_NULL_COUNT,
SUM(CASE WHEN country IS NULL THEN 1 ELSE 0 END) AS CN_NULL_COUNT,
SUM(CASE WHEN worlwide_gross_income IS NULL THEN 1 ELSE 0 END) AS GROSS_NULL_COUNT,
SUM(CASE WHEN languages IS NULL THEN 1 ELSE 0 END) AS LN_NULL_COUNT,
SUM(CASE WHEN production_company IS NULL THEN 1 ELSE 0 END) AS PROD_NULL_COUNT
FROM movie;
```

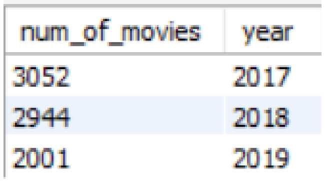


ID_NULL_COUNT	TI_NULL_COUNT	YR_NULL_COUNT	DATE_NULL_COUNT	DUR_NULL_COUNT	CN_NULL_COUNT	GROSS_NULL_COUNT	LN_NULL_COUNT	PROD_NULL_COUNT
0	0	0	0	0	20	3724	194	528

3. Find total number of movies released each year ? How does the trend look month wise?

```
SELECT count(id) as num_of_movies, year
FROM movie
GROUP BY year;

--For monthly count
```

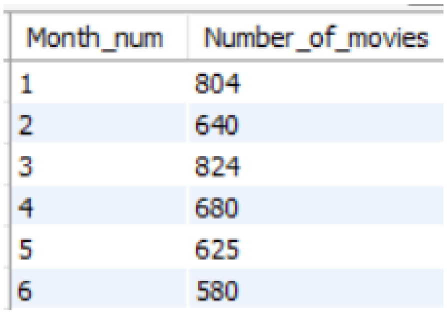
```
SELECT Month(date_published) as month_num, count(id) as num_of_movies
FROM movie
GROUP BY month_num
ORDER BY num_of_movies;
```



num_of_movies	year
3052	2017
2944	2018
2001	2019

4. Number of movies released each month?

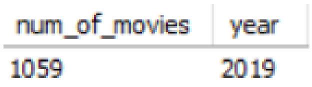
```
SELECT Month(date_published) as Month_num, count(*) as Number_of_movies
FROM movie
GROUP BY Month_num
ORDER BY Month_num;
```



Month_num	Number_of_movies
1	804
2	640
3	824
4	680
5	625
6	580

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

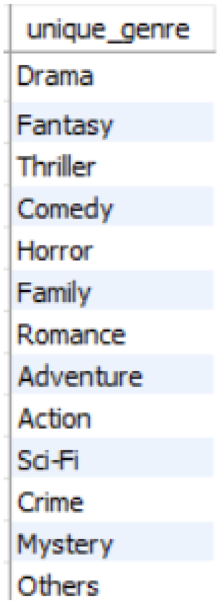
```
SELECT COUNT(DISTINCT(id)) as num_of_movies, year
FROM movie
WHERE country like '%India%' or country like'%USA%'
GROUP BY year
HAVING year = 2019;
```



num_of_movies	year
1059	2019

6. Find unique list of genre present in the dataset?

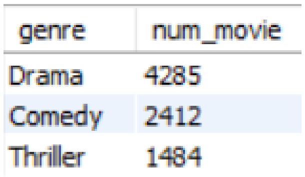
```
SELECT distinct genre as unique_genre
FROM genre;
```



unique_genre
Drama
Fantasy
Thriller
Comedy
Horror
Family
Romance
Adventure
Action
Sci-Fi
Crime
Mystery
Others

7. Which genre has the highest number of movies?

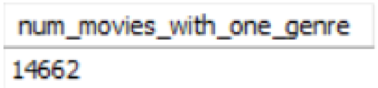
```
SELECT g.genre , count(m.id) as num_movie
FROM movie m
INNER JOIN genre g on m.id = g.movie_id
GROUP BY g.genre
ORDER BY num_movie DESC
LIMIT 1;
```



genre	num_movie
Drama	4285
Comedy	2412
Thriller	1484

8. How many movies belong to only one genre?

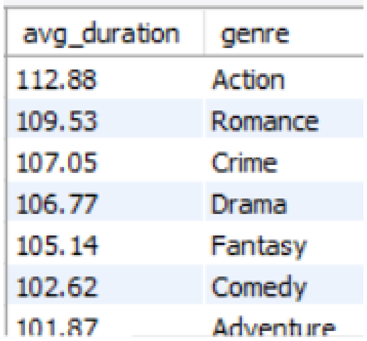
```
WITH GenreCount AS (
SELECT m.id , count(g.genre) as num_genre
FROM movie m
INNER JOIN genre g on m.id = g.movie_id
GROUP BY g.genre,m.id)

SELECT Count(*) AS num_movies_with_one_genre
FROM GenreCount
WHERE num_genre = 1;
```



num_movies_with_one_genre
14662

9. What is the average duration of movies in each genre?

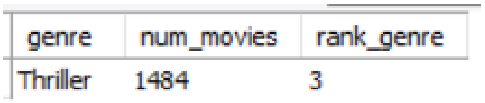
```
SELECT round(avg(m.duration), 2) as avg_duration, g.genre
from movie m
INNER JOIN genre g on g.movie_id = m.id
GROUP BY g.genre
ORDER BY avg_duration DESC;
```



avg_duration	genre
112.88	Action
109.53	Romance
107.05	Crime
106.77	Drama
105.14	Fantasy
102.62	Comedy
101.87	Adventure

10. What is the rank of 'Thriller' genre in nterms of number of movies produced?

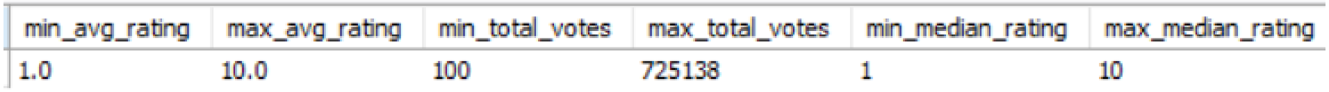
```
WITH genre_summary AS(
SELECT genre,
COUNT(movie_id) as num_movies,
rank() over (order by COUNT(movie_id) DESC) as rank_genre
from genre
group by genre
)

SELECT * FROM genre_summary WHERE genre = 'thriller';
```



genre	num_movies	rank_genre
Thriller	1484	3

11.Exploring Ratings — Min and Max Values

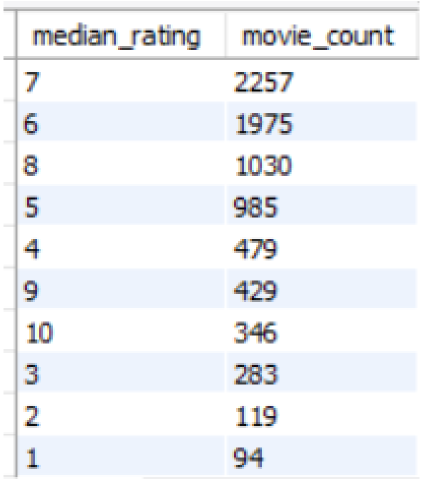
```
SELECT
MIN(avg_rating) AS min_avg_rating, MAX(avg_rating) AS max_avg_rating,
MIN(total_votes) AS min_total_votes, MAX(total_votes) AS max_total_votes,
MIN(median_rating) AS min_median_rating, MAX(median_rating) AS max_median_rating
FROM ratings;
```



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

12. Summarize the ratings table based on movie counts by median ratings?

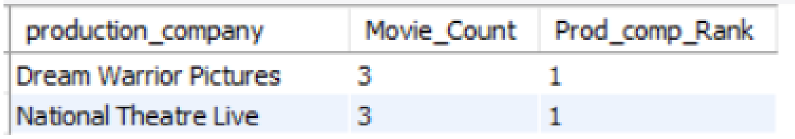
```
SELECT median_rating,
COUNT(movie_id) as movie_count
FROM ratings
GROUP BY median_rating
ORDER BY movie_count DESC;
```



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

13. Which production house has produced most hit movies (average rating > 8)?

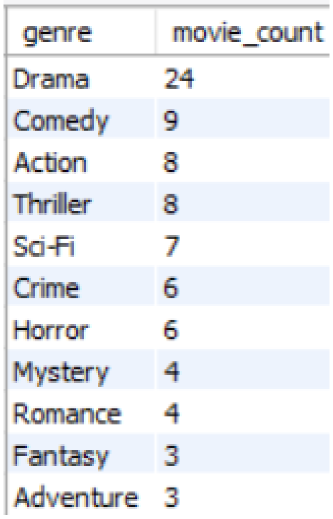
```
WITH MovieRatings AS (
SELECT
m.production_company,
count(m.id) as Movie_Count,
RANK() OVER (ORDER BY count(m.id) DESC) AS Prod_comp_Rank
FROM
movie m
INNER JOIN ratings r ON r.movie_id = m.id
WHERE r.avg_rating > 8
AND m.production_company IS NOT NULL
GROUP BY m.production_company
)
SELECT * FROM MovieRatings
WHERE Prod_comp_Rank = 1;
```



production_company	Movie_Count	Prod_comp_Rank
Dream Warrior Pictures	3	1
National Theatre Live	3	1

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

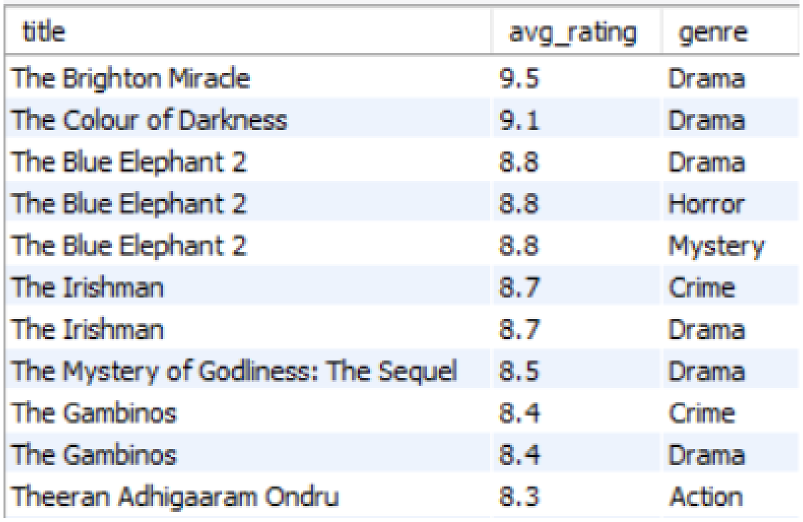
```
SELECT genre, count(m.id) as movie_count
FROM movie m
INNER JOIN genre g on g.movie_id = m.id
INNER JOIN ratings r on r.movie_id = m.id
where year = 2017 and month(date_published) = 3
and country like '%USA%'
and total_votes > 1000
GROUP BY genre
ORDER BY movie_count DESC;
```



genre	movie_count
Drama	24
Comedy	9
Action	8
Thriller	8
Sci-Fi	7
Crime	6
Horror	6
Mystery	4
Romance	4
Fantasy	3
Adventure	3

15. Find movies of each genre starting with 'The' and having average rating > 8?

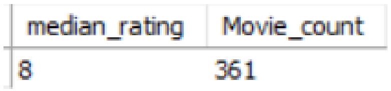
```
SELECT title, avg_rating, genre
FROM movie m
INNER JOIN genre g on g.movie_id = m.id
INNER JOIN ratings r on r.movie_id = m.id
WHERE avg_rating > 8
AND title like 'The%'
ORDER BY avg_rating DESC;
```



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

16. You should also try your hand at median rating and check whether the 'median rating' column gives any significant insights of the movies released between 1 April 2018 and 1 April 2019, how many were given a median rating of 8?

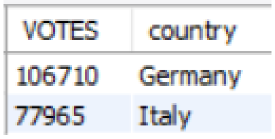
```
SELECT median_rating, COUNT(*) as Movie_count
FROM movie m
INNER JOIN ratings r on r.movie_id = m.id
WHERE median_rating = 8
AND date_published between '2018-04-01' and '2019-04-01'
GROUP BY median_rating;
```



median_rating	Movie_count
8	361

17. Do German movies got more votes than Italian movies?

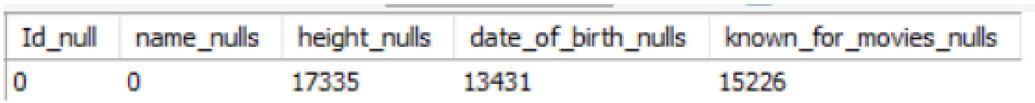
```
SELECT Sum(total_votes) AS VOTES, country
FROM movie m
INNER JOIN ratings r on m.id = r.movie_id
WHERE country in ('Germany', 'Italy')
GROUP BY country;
```



VOTES	country
106710	Germany
77965	Italy

18. Which columns in the name table have null values?

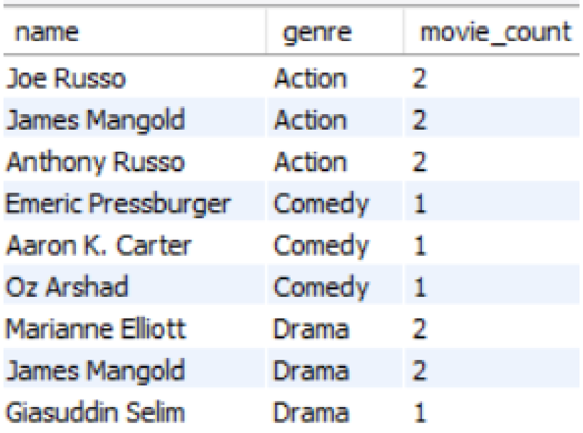
```
SELECT
SUM(CASE WHEN id IS NULL THEN 1 ELSE 0 END) AS Id_null,
SUM(CASE WHEN name IS NULL THEN 1 ELSE 0 END ) AS name_nulls,
SUM(CASE WHEN height IS NULL THEN 1 ELSE 0 END ) AS height_nulls,
SUM(CASE WHEN date_of_birth IS NULL THEN 1 ELSE 0 END ) AS date_of_birth_nulls,
SUM(CASE WHEN known_for_movies IS NULL THEN 1 ELSE 0 END ) AS known_for_movies_nulls
FROM names;
```



Id_null	name_nulls	height_nulls	date_of_birth_nulls	known_for_movies_nulls
0	0	17335	13431	15226

19. Who are the top three directors in the top three genres whose movies have an average rating 8? (Hint: The top three genres would have the most number of movies with an average rating > 😎

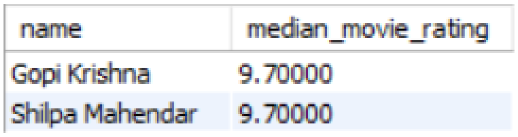
```
WITH RatedMovies AS (
-- Filter movies with average rating >= 8
SELECT r.movie_id, r.avg_rating
FROM ratings r
WHERE r.avg_rating >= 8
),
GenreRankings AS (
-- Get top 3 genres by the number of highly rated movies
SELECT g.genre, COUNT(g.movie_id) AS movie_count,
ROW_NUMBER() OVER (ORDER BY COUNT(g.movie_id) DESC) AS genre_rank
FROM RatedMovies rm
JOIN genre g ON rm.movie_id = g.movie_id
GROUP BY g.genre
ORDER BY movie_count DESC
LIMIT 3 -- Select the top 3 genres
),
DirectorRankings AS (
-- For each genre, rank the directors by the number of highly rated movies
SELECT dm.name_id, g.genre, COUNT(dm.movie_id) AS movie_count,
ROW_NUMBER() OVER (PARTITION BY g.genre ORDER BY COUNT(dm.movie_id) DESC) AS
director_rank
FROM RatedMovies rm
JOIN genre g ON rm.movie_id = g.movie_id
JOIN director_mapping dm ON rm.movie_id = dm.movie_id
WHERE g.genre IN (SELECT genre FROM GenreRankings) -- Filter to top 3 genres
GROUP BY dm.name_id, g.genre
)

SELECT n.name, dr.genre, dr.movie_count
FROM DirectorRankings dr
JOIN names n ON dr.name_id = n.id
WHERE dr.director_rank <= 3;
```



name	genre	movie_count
Joe Russo	Action	2
James Mangold	Action	2
Anthony Russo	Action	2
Emeric Pressburger	Comedy	1
Aaron K. Carter	Comedy	1
Oz Arshad	Comedy	1
Marianne Elliott	Drama	2
James Mangold	Drama	2
Giasuddin Selim	Drama	1

20. Who are the top two actors whose movies have a median range > =8?

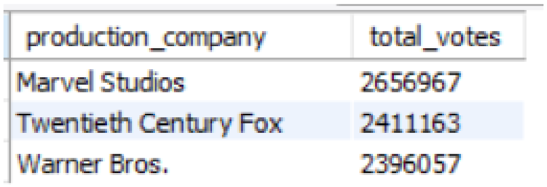
```
WITH ActorMovies AS (
SELECT rm.name_id, rm.movie_id, r.avg_rating
FROM role_mapping rm
JOIN ratings r ON rm.movie_id = r.movie_id
WHERE rm.category = 'Actor' -- Assuming 'Actor' is the category for actors
),
ActorMedianRatings AS (
SELECT name_id, COUNT(movie_id) AS movie_count,
avg(avg_rating) AS median_movie_rating
FROM ActorMovies
GROUP BY name_id
HAVING median_movie_rating >= 8
)
SELECT n.name, am.median_movie_rating
FROM ActorMedianRatings am
JOIN names n ON am.name_id = n.id
ORDER BY am.median_movie_rating DESC
LIMIT 2;
```



name	median_movie_rating
Gopi Krishna	9.70000
Shilpa Mahendar	9.70000

21. Which are the top 3 production houses based on the number of votes received by their movies?

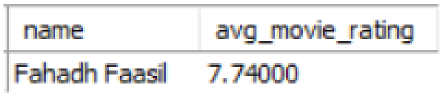
```
WITH ProductionVotes AS (
SELECT m.production_company, SUM(r.total_votes) AS total_votes
FROM movie m
JOIN ratings r ON m.id = r.movie_id
GROUP BY m.production_company
)
SELECT production_company, total_votes
FROM ProductionVotes
ORDER BY total_votes DESC
LIMIT 3;
```



production_company	total_votes
Marvel Studios	2656967
Twentieth Century Fox	2411163
Warner Bros.	2396057

22. Rank actors with movies released in India based on their average rating. Which actor is at the top of the list? Note: The actor should have acted at least in 5 Indian movies

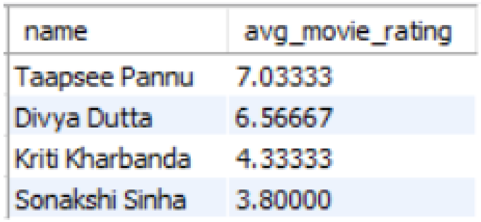
```
WITH ActorMovies AS (
SELECT rm.name_id, r.avg_rating, rm.movie_id
FROM role_mapping rm
JOIN ratings r ON rm.movie_id = r.movie_id
JOIN movie m ON m.id = rm.movie_id
WHERE m.country = 'India' -- Filter Indian movies
),
ActorRating AS (
SELECT name_id, AVG(avg_rating) AS avg_movie_rating, COUNT(movie_id) AS movie_count
FROM ActorMovies
GROUP BY name_id
HAVING COUNT(movie_id) >= 5
)
SELECT n.name, ar.avg_movie_rating
FROM ActorRating ar
JOIN names n ON ar.name_id = n.id
ORDER BY ar.avg_movie_rating DESC
LIMIT 1; -- This will return the top actor
```



name	avg_movie_rating
Fahadh Faasil	7.74000

23. Find out the Top 5 actresses in Hindi movies released in India based on their average rating? Note: the actress should have acted in atleast 3 indian movies.

```
WITH ActressMovies AS (
SELECT rm.name_id, r.avg_rating, rm.movie_id
FROM role_mapping rm
JOIN ratings r ON rm.movie_id = r.movie_id
JOIN movie m ON m.id = rm.movie_id
WHERE m.country = 'India' AND m.languages = 'Hindi' -- Filter Hindi movies in India
AND rm.category = 'Actress' -- Assuming 'Actress' is the category for actresses
),
ActressRating AS (
SELECT name_id, AVG(avg_rating) AS avg_movie_rating, COUNT(movie_id) AS movie_count
FROM ActressMovies
GROUP BY name_id
HAVING COUNT(movie_id) >= 3
)
SELECT n.name, ar.avg_movie_rating
FROM ActressRating ar
JOIN names n ON ar.name_id = n.id
ORDER BY ar.avg_movie_rating DESC
LIMIT 5;
```



name	avg_movie_rating
Taapsee Pannu	7.03333
Divya Dutta	6.56667
Kriti Kharbanda	4.33333
Sonakshi Sinha	3.80000

24. Select the thriller movies and classify them in the following category:

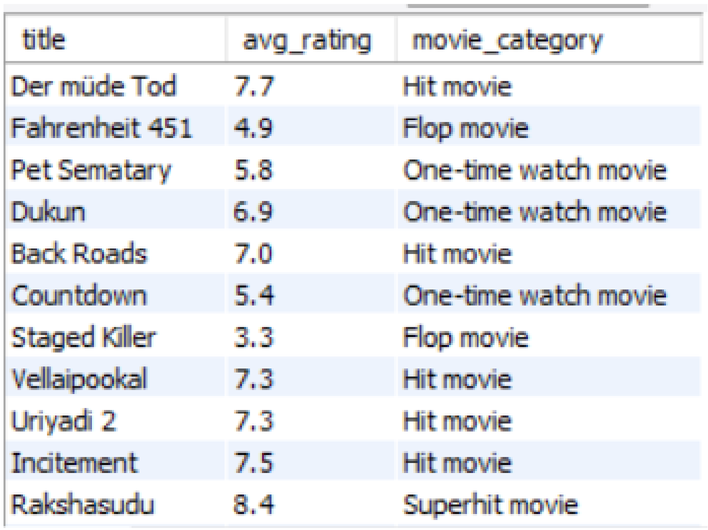
Rating>8: Superhit movies

Rating between 7,8: Hit mmovie

Rating between 5 and 7: One time watch movie

Rating < 5: Flop movie

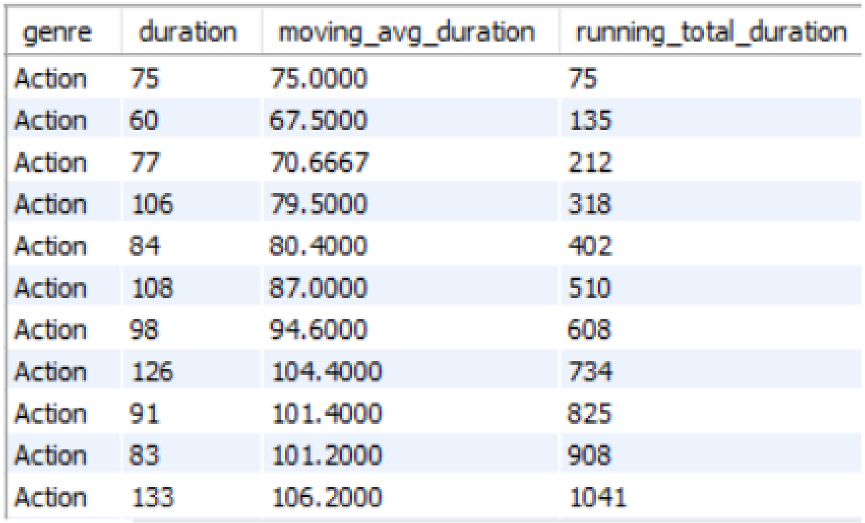
```
SELECT m.title, r.avg_rating,
CASE
WHEN r.avg_rating > 8 THEN 'Superhit movie'
WHEN r.avg_rating BETWEEN 7 AND 8 THEN 'Hit movie'
WHEN r.avg_rating BETWEEN 5 AND 7 THEN 'One-time watch movie'
WHEN r.avg_rating < 5 THEN 'Flop movie'
END AS movie_category
FROM movie m
JOIN ratings r ON m.id = r.movie_id
JOIN genre g ON m.id = g.movie_id
WHERE g.genre = 'Thriller'; -- Assuming 'Thriller' is the genre for thriller movies
```



title	avg_rating	movie_category
Der müde Tod	7.7	Hit movie
Fahrenheit 451	4.9	Flop movie
Pet Sematary	5.8	One-time watch movie
Dukun	6.9	One-time watch movie
Back Roads	7.0	Hit movie
Countdown	5.4	One-time watch movie
Staged Killer	3.3	Flop movie
Vellaipookal	7.3	Hit movie
Uriyadi 2	7.3	Hit movie
Incitement	7.5	Hit movie
Rakshasudu	8.4	Superhit movie

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

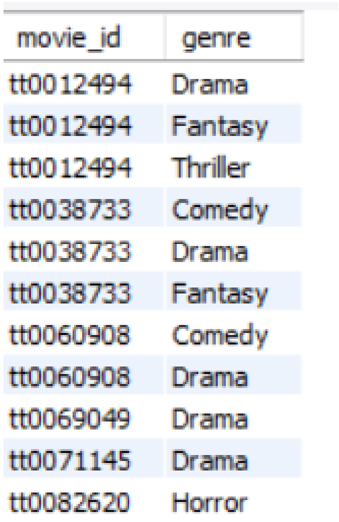
```
WITH GenreDuration AS (
SELECT g.genre, m.duration,
AVG(m.duration) OVER (PARTITION BY g.genre ORDER BY m.date_published ROWS BETWEEN 4
PRECEDING AND CURRENT ROW) AS moving_avg_duration,
SUM(m.duration) OVER (PARTITION BY g.genre ORDER BY m.date_published ROWS BETWEEN
UNBOUNDED PRECEDING AND CURRENT ROW) AS running_total_duration
FROM movie m
JOIN genre g ON m.id = g.movie_id
)
SELECT genre, duration, moving_avg_duration, running_total_duration
FROM GenreDuration;
```



genre	duration	moving_avg_duration	running_total_duration
Action	75	75.0000	75
Action	60	67.5000	135
Action	77	70.6667	212
Action	106	79.5000	318
Action	84	80.4000	402
Action	108	87.0000	510
Action	98	94.6000	608
Action	126	104.4000	734
Action	91	101.4000	825
Action	83	101.2000	908
Action	133	106.2000	1041

26. Which are the 5 highest grossing movies of each year that belongs to top 3 genre?

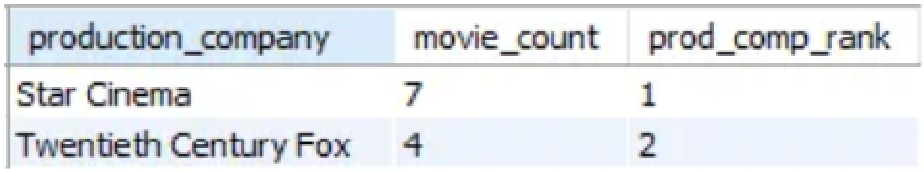
```
WITH TopGenres AS (
SELECT genre, COUNT(movie_id) AS movie_count
FROM genre
GROUP BY genre
ORDER BY movie_count DESC
LIMIT 3 -- Select the top 3 genres
),
YearlyTopMovies AS (
SELECT m.id, m.title, m.year, m.worlwide_gross_income as gross_earnings, g.genre,
ROW_NUMBER() OVER (PARTITION BY m.year, g.genre ORDER BY m.worlwide_gross_income
DESC) AS movie_rank
FROM movie m
JOIN genre g ON m.id = g.movie_id
WHERE g.genre IN (SELECT genre FROM TopGenres)
)
SELECT id, title, year, gross_earnings, genre
FROM YearlyTopMovies
WHERE movie_rank <= 5;
select * from genre;
```



movie_id	genre
tt0012494	Drama
tt0012494	Fantasy
tt0012494	Thriller
tt0038733	Comedy
tt0038733	Drama
tt0038733	Fantasy
tt0060908	Comedy
tt0060908	Drama
tt0069049	Drama
tt0071145	Drama
tt0082620	Horror

27. Which are the top 2 production houses that have produced the highest number of hits(median rating >= 8) among multilingual movies?

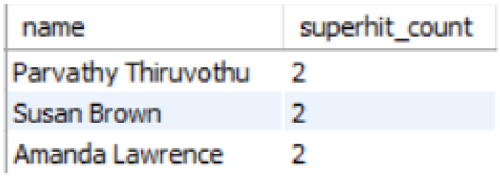
```
WITH HitMovies AS (
SELECT m.id, m.production_company, r.avg_rating
FROM movie m
JOIN ratings r ON m.id = r.movie_id
WHERE r.avg_rating >= 8 #AND m.is_multilingual = 1 -- Assuming there's a
multilingual flag
),
ProductionHouseHits AS (
SELECT production_company, COUNT(id) AS hit_count
FROM HitMovies
GROUP BY production_company
ORDER BY hit_count DESC
)
SELECT production_company, hit_count
FROM ProductionHouseHits
LIMIT 2;
```



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

28. Who are the top 3 actresses based on number of superhit movies (avg rating >8) in drama genre?

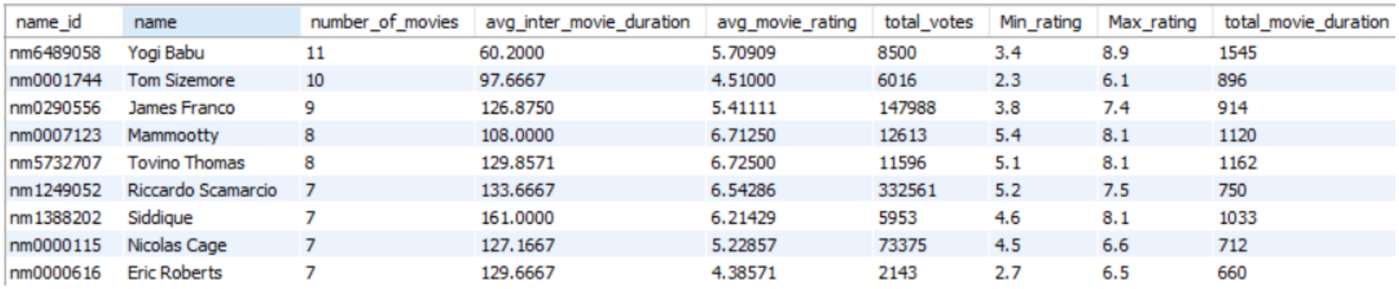
```
WITH SuperhitDramaMovies AS (
SELECT rm.name_id, r.avg_rating, rm.movie_id
FROM role_mapping rm
JOIN ratings r ON rm.movie_id = r.movie_id
JOIN genre g ON rm.movie_id = g.movie_id
WHERE g.genre = 'Drama' AND r.avg_rating > 8 AND rm.category = 'Actress' -- Filter
Drama and Superhit movies
),
ActressSuperhitCount AS (
SELECT name_id, COUNT(movie_id) AS superhit_count
FROM SuperhitDramaMovies
GROUP BY name_id
)
SELECT n.name, asco.superhit_count
FROM ActressSuperhitCount asco
JOIN names n ON asco.name_id = n.id
ORDER BY asco.superhit_count DESC
LIMIT 3;
```



name	superhit_count
Parvathy Thiruvothu	2
Susan Brown	2
Amanda Lawrence	2

29. Get the following details for top 9 directors(based on number of movies): Director_id, name, number_of_movie, Avg inter movie duration in days, avg movie rating, total votes, min rating, max rating, total movie duration

```
WITH DirectorMovies AS (
SELECT dm.name_id, m.id AS movie_id, m.date_published, r.avg_rating, r.total_votes,
m.duration,
LEAD(m.date_published) OVER (PARTITION BY dm.name_id ORDER BY m.date_published) AS
next_movie_date
FROM role_mapping dm
JOIN movie m ON dm.movie_id = m.id
JOIN ratings r ON m.id = r.movie_id
),
DirectorMoviesWithDuration AS (
SELECT name_id, movie_id, date_published, avg_rating, total_votes, duration,
DATEDIFF(next_movie_date, date_published) AS inter_movie_duration
FROM DirectorMovies
),
DirectorStats AS (
SELECT name_id, COUNT(movie_id) AS number_of_movies,
AVG(inter_movie_duration) AS avg_inter_movie_duration,
AVG(avg_rating) AS avg_movie_rating,
MIN(avg_rating) AS Min_rating,
MAX(avg_rating) AS Max_rating,
SUM(total_votes) AS total_votes,
SUM(duration) AS total_movie_duration
FROM DirectorMoviesWithDuration
GROUP BY name_id
)
SELECT ds.name_id, n.name, ds.number_of_movies, ds.avg_inter_movie_duration,
ds.avg_movie_rating,
ds.total_votes, ds.Min_rating, ds.Max_rating, ds.total_movie_duration
FROM DirectorStats ds
JOIN names n ON ds.name_id = n.id
ORDER BY ds.number_of_movies DESC
LIMIT 9;
```



name_id	name	number_of_movies	avg_inter_movie_duration	avg_movie_rating	total_votes	Min_rating	Max_rating	total_movie_duration
nm6489058	Yogi Babu	11	60.2000	5.70909	8500	3.4	8.9	1545
nm0001744	Tom Sizemore	10	97.6667	4.51000	6016	2.3	6.1	896
nm0290556	James Franco	9	126.8750	5.41111	147988	3.8	7.4	914
nm0007123	Mammootty	8	108.0000	6.71250	12613	5.4	8.1	1120
nm5732707	Tovino Thomas	8	129.8571	6.72500	11596	5.1	8.1	1162
nm1249052	Riccardo Scamarcio	7	133.6667	6.54286	332561	5.2	7.5	750
nm1388202	Siddique	7	161.0000	6.21429	5953	4.6	8.1	1033
nm0000115	Nicolas Cage	7	127.1667	5.22857	73375	4.5	6.6	712
nm0000616	Eric Roberts	7	129.6667	4.38571	2143	2.7	6.5	660

Conclusion: This analytical journey equips RSVP Movies with valuable insights. By utilizing data-driven recommendations, they can strategically plan their global

audience-friendly movie project. With a firm grasp of their data, RSVP Movies is poised to create cinematic magic that resonates with audiences worldwide.