# DATA ANALYSIS ON IMDB DATASET

**MySQL** 

# **OESON GLOBAL**

Submitted by:
Monika Saxena
Meera Radhish

**Supriya Meshram** 

**Supervised by: Aritri Debnath** 

# **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.

You are a data analyst and an SQL expert. You have to use SQL to analyse the given data and give recommendations to RSVP Movies based on the insights. For your convenience, the entire analytics process has been divided into four segments, where each segment leads to significant insights from different combinations of tables. The questions in each segment with business objectives are written in the script given below. You have to write the solution code below every question and submit the same SQL script file with the solution in the 'Submission' segment.

# **MYSQL-PROJECT**

```
USE imdb;
/* Now that you have imported the data sets, let's explore some of the
tables.
To begin with, it is beneficial to know the shape of the tables and
whether any column has null values.
 Further in this segment, you will take a look at 'movies' and 'genre'
tables.*/
-- Segment 1:
QUES-1
-- Q1. Find the total number of rows in each table of the schema?
-- Type your code below:
Sol 1:
SELECT Count(*) FROM movie;
-- No. of rows: 7997
SELECT Count(*) FROM genre;
-- No. of rows: 14662
SELECT Count(*) FROM director_mapping;
-- No. of rows: 3867
SELECT Count(*) FROM names;
-- No. of rows: 25735
SELECT Count(*) FROM ratings;
-- No. of rows: 7997
SELECT Count(*) FROM role_mapping;
-- No. of rows: 15615
QUES 2
-- Q2. Which columns in the movie table have null values?
-- Type your code below:
Sol 2:
SELECT Sum(CASE
      WHEN id IS NULL THEN 1
      ELSE 0
```

```
END) AS id_null,
Sum(CASE
   WHEN title IS NULL THEN 1
  ELSE 0
  END) AS title_null,
Sum(CASE
   WHEN year IS NULL THEN 1
  ELSE 0
 END) AS year_null,
Sum(CASE
   WHEN date_published IS NULL THEN 1
  ELSE 0
  END) AS date_published_null,
Sum(CASE
   WHEN duration IS NULL THEN 1
  ELSE 0
 END) AS duration_null,
Sum(CASE
   WHEN country IS NULL THEN 1
  ELSE 0
  END) AS country_null,
Sum(CASE
   WHEN worlwide_gross_income IS NULL THEN 1
  ELSE 0
  END) AS worlwide_gross_income_null,
Sum(CASE
   WHEN languages IS NULL THEN 1
  ELSE 0
 END) AS languages_null,
Sum(CASE
   WHEN production_company IS NULL THEN 1
```

ELSE 0

END) AS production\_company\_null

FROM movie;

# Found null in below given columns (count mentioned)

- country- 20
- worlwide\_gross\_income 3724
- languages 194
- production\_company 528

## QUES 3

- -- Now as you can see four columns of the movie table has null values. Let's look at the at the movies released each year.
- -- Q3. Find the total number of movies released each year? How does the trend look month wise? (Output expected)
- /\* Output format for the first part:

+   Year	+	+ number_of_movies		
2017   2018   2019	+     	2134	I	

Output format for the second part of the question:

+			+	
į	month_num	I	number_of_movies	
+	 1		134	
	2		231	
	•			
+			+ */	

-- Type your code below:

# Sol3:

# The total number of movies released each year

SELECT year,

Count(title) AS NUMBER\_OF\_MOVIES

FROM movie

GROUP BY year;

	year	NUMBER_OF_MOVIES
•	2017	3052
	2018	2944
	2019	2001

# - 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
	7	493
	8	678
	9	809
	10	801
	11	625
	12	438

-- March has highest and December has least no. of films released.

/\*The highest number of movies is produced in the month of March. So, now that you have understood the month-wise trend of movies, let's take a look at the other details in the movies table. We know USA and India produces huge number of movies each year. Lets find the number of movies produced by USA or India for the last year.\*/

- -- Q4. How many movies were produced in the USA or India in the year 2019??
- -- Type your code below:

#### **Sol 4:**

```
SELECT
COUNT(DISTINCT id) AS number_of_movies,
year

FROM
movie
WHERE
(UPPER(country) LIKE '%USA%'
OR UPPER(country) LIKE '%India%')
AND year = 2019
GROUP BY
year;

number_of_movies year

1059 2019
```

-- Number of movies produced by USA or India for the year 2019 is "1059".

## Ques 5.

/\* USA and India produced more than a thousand movies(you know the
exact number!) in the year 2019.
Exploring table Genre would be fun!!
Let's find out the different genres in the dataset.\*/

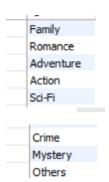
# -- Q5. Find the unique list of the genres present in the data set?

-- Type your code below:

## **Sol 5:**

# SELECT DISTINCT genre FROM genre,





/\* So, RSVP Movies plans to make a movie of one of these genres.
Now, wouldn't you want to know which genre had the highest number of
movies produced in the last year?
Combining both the movie and genres table can give more interesting
insights. \*/

- -- Q6. Which genre had the highest number of movies produced overall?
- -- Type your code below:

#### **Sol 6:**

```
SELECT genre,
Count(mov.id) AS number_of_movies
FROM movie AS mov
INNER JOIN genre AS gen
where gen.movie_id = mov.id
GROUP BY genre
ORDER BY number_of_movies DESC limit 1;
```

genre number\_of\_movies

Drama 4285

--Drama genre had the highest movies produced overall i.e, 4285.

#### Ques 7

 $/\ast$  So, based on the insight that you just drew, RSVP Movies should focus on the 'Drama' genre. But wait, it is too early to decide. A movie can belong to two or more

genres.

So, let's find out the count of movies that belong to only one genre.\*/

## -- Q7. How many movies belong to only one genre?

-- Type your code below:

## **Sol 7**:

```
SELECT genre_count,

Count(movie_id) movie_count

FROM (SELECT movie_id, Count(genre) genre_count

FROM genre

GROUP BY movie_id

ORDER BY genre_count DESC) genre_counts

WHERE genre_count = 1

GROUP BY genre_count;
```

-- 3289 movies have exactly one genre.

#### Ques 8

```
/* There are more than three thousand movies which has only one genre
associated with them.
So, this figure appears significant.
Now, let's find out the possible duration of RSVP Movies' next
project.*/
-- Q8.What is the average duration of movies in each genre?
-- (Note: The same movie can belong to multiple genres.)
/* Output format:
+----+
         | avg_duration |
+-----
| thriller |
                           105
                                             +----+ */
-- Type your code below:
```

#### **Sol 8:**

```
SELECT genre,

Round(Avg(duration),2) AS avg_duration

FROM movie as mov

INNER JOIN genre as gen
```

ON gen.movie\_id = mov.id

# GROUP BY genre

# ORDER BY avg\_duration DESC;

	genre	avg_duration
١	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

-- Duration of Action movies is highest with duration of 112.88 mins whereas Horror movies have least with duration 92.72 mins.

## Ques 9:

/\* Now you know, movies of genre 'Drama' (produced highest in number in 2019) has the average duration of 106.77 mins. Lets find where the movies of genre 'thriller' on the basis of number of movies.\*/

-- Q9.What is the rank of the 'thriller' genre of movies among all the genres in terms of number of movies produced?

-- (Hint: Use the Rank function)

genre genre_rank	1		movie_count	
drama		2312	+	

2

**Sol 9:** 

SELECT genre,

Round(Avg(duration),2) AS avg\_duration

```
FROM
       movie as mov
INNER JOIN genre as gen
     gen.movie_id = mov.id
GROUP BY genre
ORDER BY avg_duration DESC;
-- Duration of Action movies is highest with duration of 112.88 mins whereas Horror movies
have least with duration 92.72 mins.
WITH genre summary AS
 SELECT
  genre,
      Count(movie_id)
                               AS movie_count,
      Rank() OVER(ORDER BY Count(movie_id) DESC) AS genre_rank
 FROM
        genre
 GROUP BY genre
 )
SELECT *
FROM genre_summary
WHERE genre = "THRILLER";
  genre movie_count genre_rank
 Thriller 1484
-- Thriller genre has 3rd rank with 1484 movies.
-- Segment 2:
Ques 10
-- Q10. Find the minimum and maximum values in each column of the ratings table except the
movie id column?
/* Output format:
+----
----+
```

177

# Sol:

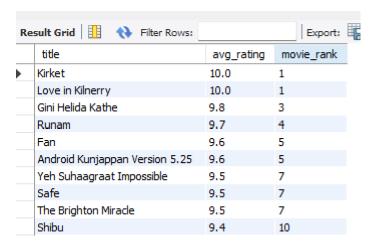
select title, avg\_rating,

rank() over(order by avg\_rating desc)as movie\_rank

from ratings

inner join movie on movie.id =ratings.movie\_id
limit 10;

## **Output:**

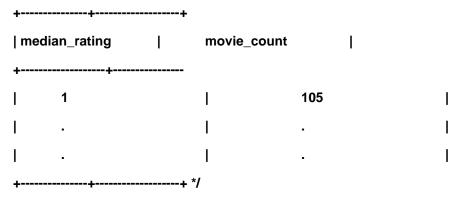


Movie FAN has rating 9.6

# Ques 12

Q12: Summarise the ratings table based on the movie counts by median ratings.

## /\* Output format:

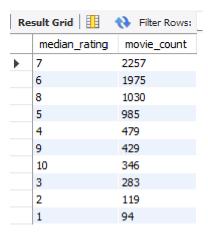


Sol

select median\_rating, count(movie\_id) as movie\_count

from ratings
group by median\_rating
order by movie\_count desc;

# **Output:**

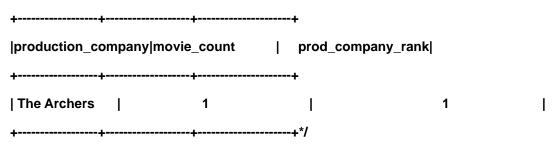


Movies with a median rating of 7 is highest in number.

#### Ques 13

Q13 : Which production house has produced the most number of hit movies (average rating > 8)??

# /\* Output format:



#### Sol:

select production\_company, count(movie\_id) as movie\_count,
rank() over(order by count(movie\_id) desc) as company\_rank
from ratings

inner join movie on movie.id=ratings.movie\_id
where (avg\_rating > 8 )and (production\_company is not null)

## group by production\_company;

## **Output:**

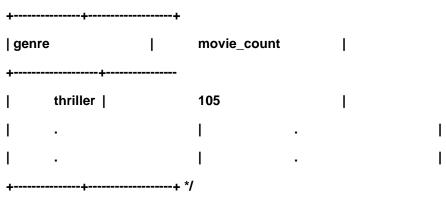


Dream Warrior Pictures & National Theater Live are top the production houses producing hit number of movies.

# Ques 14

Q14: How many movies released in each genre during March 2017 in the USA had more than 1,000 votes?

# /\* Output format:



Sol:

use imdb;

select genre, count(movie.id) as movie\_count

## from movie

inner join genre on genre.movie\_id =movie.id

inner join ratings on ratings.movie\_id=movie.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;

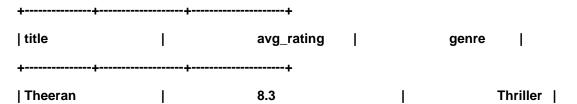
# Output:

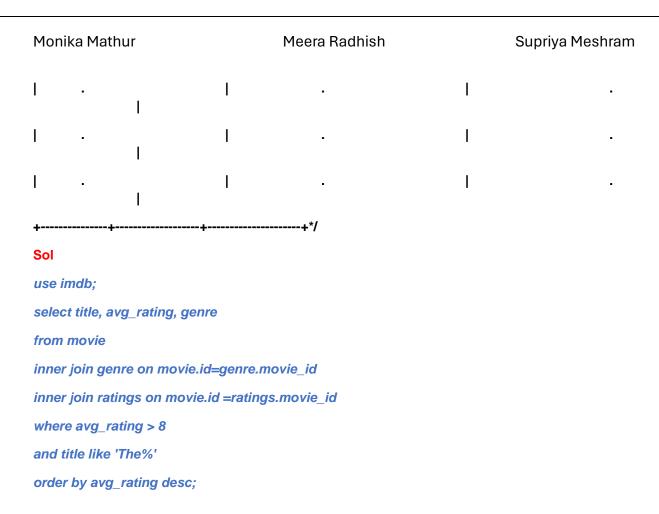


#### **Ques 15**

Q15: Find movies of each genre that start with the word 'The' and which have an average rating > 8?

# /\* Output format:





# Output:

	title	avg_rating	genre
▶	The Brighton Mirade	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

There are 8 movies starting with 'the' in genereal.

Q16: Of the movies released between 1 April 2018 and 1 April 2019, how many were given a median rating of 8?

## Sol:

```
select median_rating,
```

count(\*) as movie\_count

from movie as mov

join

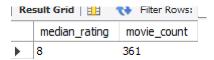
ratings as rat on rat.movie\_id = mov.id

where median\_rating = 8

and date\_published between '2018-04-01' and '2019-04-01'

group by median\_rating;

## **Output:**



There were 361 movies released between 1 April 2018 &1 April 2019

#### **Ques 17**

Q17: Do German movies get more votes than Italian movies?

# Sol:

select country, sum(total\_votes) as total\_number\_votes

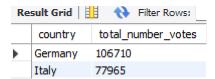
from movie

inner join ratings on movie.id=ratings.movie\_id

where country ='Germany' or country ='Italy'

group by country;

# **Output:**



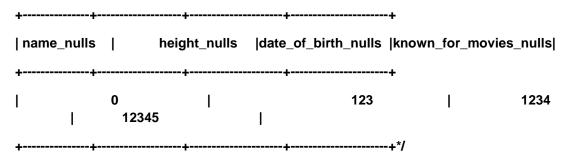
Monika Mathur Meera Radhish Supriya Meshram

Yes, German movies have more votes than Italian movies

#### **Ques 18**

Q18: Which columns in the names table have null values??

/\*Hint: You can find null values for individual columns or follow below output format



#### Sol:

#### select

count(\*)-count(name) as name\_nulls,

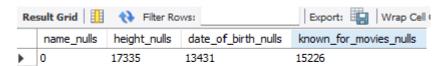
count(\*)-count(height) as height\_nulls,

count(\*)-count(date\_of\_birth) as date\_of\_birth\_nulls,

count(\*)-count(known\_for\_movies) as known\_for\_movies\_nulls

from names;

# **Output:**



There are no Null value in the column 'name'.

#### Ques 19

Q19: 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 > 8.)

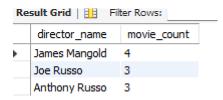
# /\* Output format:

```
+----+
| director_name | movie_count |
+-----
|James Mangold |
                      +----+ */
Sol:
WITH top_rated_genres AS
SELECT
 genre,
 COUNT(m.id) AS movie_count,
RANK () OVER (ORDER BY COUNT(m.id) DESC) AS genre_rank
FROM
 genre AS g
   LEFT JOIN
 movie AS m
 ON g.movie_id = m.id
 INNER JOIN
ratings AS r
 ON m.id=r.movie_id
WHERE avg_rating>8
GROUP BY genre
)
SELECT
n.name as director_name,
COUNT(m.id) AS movie_count
FROM
```

```
names AS n
INNER JOIN
director_mapping AS d
 ON n.id=d.name_id
 INNER JOIN
    movie AS m
 ON d.movie_id = m.id
  INNER JOIN
      ratings AS r
  ON m.id=r.movie id
  INNER JOIN
  genre AS g
  ON g.movie_id = m.id
WHERE g.genre IN (SELECT DISTINCT genre FROM top_rated_genres WHERE genre_rank<=3)
AND avg_rating>8
GROUP BY name
ORDER BY movie_count DESC
```

# **Output:**

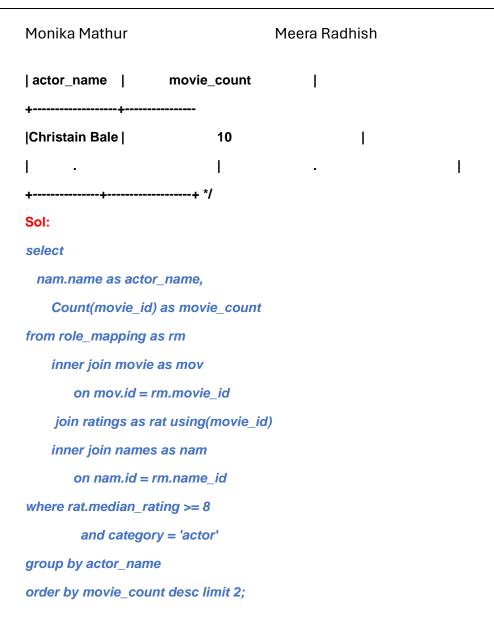
LIMIT 3;



Top 3 Directors are James Mangold, Joe Russo & Anthony Russo

Q20: Who are the top two actors whose movies have a median rating >= 8?

/\* Output format:



Supriya Meshram

# Output:

		-
	actor_name	movie_count
۰	Mammootty	8
	Mohanlal	5

Top two actors are Mammootty and Mohanlal

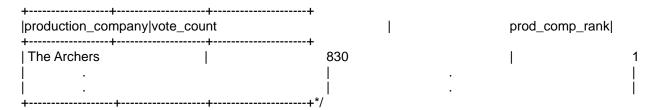
/\* Have you find your favourite actor 'Mohanlal' in the list. If no, please check your code again.

**RSVP Movies plans to partner with other global production houses.** 

Let's find out the top three production houses in the world.\*/

# Q21. Which are the top three production houses based on the number of votes received by their movies?

#### /\* Output format:



Type your code below:

#### Sol:

```
SELECT

m.production_company,

SUM(r.total_votes) AS vote_count,

RANK() OVER (ORDER BY SUM(r.total_votes) DESC) AS prod_comp_rank

FROM

imdb.movie m

JOIN

imdb.ratings r ON m.id = r.movie_id

GROUP BY

production_company

ORDER BY

vote_count DESC

LIMIT 3;
```

Re	esult Grid 🔠 🙌 Filter	Rows:	Export:	Wrap Cell Content:	<u> ‡ A</u>
	production_company	vote_count	prod_comp_rank		
•	Marvel Studios	2656967	1		
	Twentieth Century Fox	2411163	2		
	Warner Bros.	2396057	3		

/\*Yes Marvel Studios rules the movie world.

So, these are the top three production houses based on the number of votes received by the movies they have produced.

Since RSVP Movies is based out of Mumbai, India also wants to woo its local audience.

RSVP Movies also wants to hire a few Indian actors for its upcoming project to give a regional feel.

Let's find who these actors could be\*/

#### Ques 22

Q22. Rank actors with movies released in India based on their average ratings. Which actor is at the top of the list?

-- Note: The actor should have acted in at least five Indian movies.

act

-- (Hint: You should use the weighted average based on votes. If the ratings clash, then the total number of votes should act as the tie breaker.)

```
/* Output format:
+----
-----+
----+
     Yoqi Babu |
                                    3455
                                                      11
                                                      +----+
----+*/
Type your code below:
Sol:
 SELECT n. name AS actor name,
       SUM(r.total votes) AS sum total votes,
  COUNT(r.movie id) AS movie count,
       ROUND(Sum(avg_rating * total_votes) / SUM(total_votes), 2) AS actor_avg_rating,
  DENSE RANK() OVER (
                                     ORDER BY
ROUND(SUM(avg_rating * r.total_votes) / SUM(r.total_votes), 2) DESC,
SUM(r.total_votes) desc
                                     ) AS actor_rank
FROM 'names' n
  JOIN
  role_mapping rm ON n.id = rm.name_id
  JOIN
         ON r.movie_id = rm.movie_id
  ratings r
  JOIN
                  ON
  movie m
                              m.id = r.movie id
WHERE m.country LIKE '%India%' AND
      rm.category LIKE '%Actor%'
GROUP BY actor name
HAVING movie count >=5
LIMIT 1;
Result Grid Filter Rows:
                           Export: Wrap Cell Content: IA
          sum_total_votes movie_count actor_avg_rating actor_rank
  actor_name
```

8.42

Vijay Sethupathi is in top actor list

Vijay Sethupathi 23114

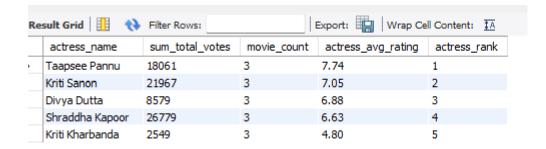
5

Q23. Find out the top five actresses in Hindi movies released in India based on their average ratings?

- -- Note: The actresses should have acted in at least three Indian movies.
- -- (Hint: You should use the weighted average based on votes. If the ratings clash, then the total number of votes should act as the tie breaker.)

act

```
/* Output format:
+----
-----+
            total votes | movie count
| actress name |
+-----+----
    Tabu
                              3455
                                            11
                   +----
----+*/
Type your code below:
Sol:
SELECT n. `name` AS actress name,
       SUM(r.total votes) AS sum total votes,
     COUNT(r.movie id) AS movie count,
       ROUND(SUM(avg rating * total votes) /
SUM(total votes), 2) AS actress avg rating,
     DENSE RANK() OVER (
ROUND(SUM(avg rating * r.total votes) / SUM(r.total votes), 2)
desc,
                     SUM(r.total votes) DESC
                               ) AS actress rank
FROM `names` n
    role mapping rm ON n.id = rm.name id
    JOIN
    ratings r ON r.movie id = rm.movie id
    JOIN
                              m.id = r.movie id
    movie m
                   ON
WHERE m.languages LIKE '%Hindi%' AND
      rm.category LIKE '%Actress%' AND
    country='india'
GROUP BY actress name
HAVING movie count >=3
LIMIT 5;
```



/\* Taapsee Pannu tops with average rating 7.74.
Now let us divide all the thriller movies in the following categories
and find out their numbers.\*/

#### Ques 24

Q24. Select thriller movies as per avg rating and classify them in the following category:

```
Rating > 8: Superhit movies
Rating between 7 and 8: Hit movies
Rating between 5 and 7: One-time-watch movies
Rating < 5: Flop movies
```

-- Type your code below:

#### Sol:

```
SELECT
 m.id,
  m.title,
  AVG(r.avg_rating) AS avg_rating,
  CASE
    WHEN AVG(r.avg_rating) > 8 THEN 'Superhit movies'
    WHEN AVG(r.avg_rating) BETWEEN 7 AND 8 THEN 'Hit movies'
    WHEN AVG(r.avg_rating) BETWEEN 5 AND 7 THEN 'One-time-watch movies'
    ELSE 'Flop movies'
  END AS movie_category
FROM
  imdb.movie m
JOIN
  imdb.ratings r ON m.id = r.movie_id
  imdb.genre g ON m.id = g.movie_id
WHERE
  g.genre = 'Thriller'
GROUP BY
  m.id, m.title
ORDER BY
  avg_rating DESC;
```



/\* Until now, you have analysed various tables of the data set. Now, you will perform some tasks that will give you a broader understanding of the data in this segment.\*/

#### Ques 25

Q25. What is the genre-wise running total and moving average of the average movie duration? -- (Note: You need to show the output table in the question.)

-- Type your code below: **Sol**:

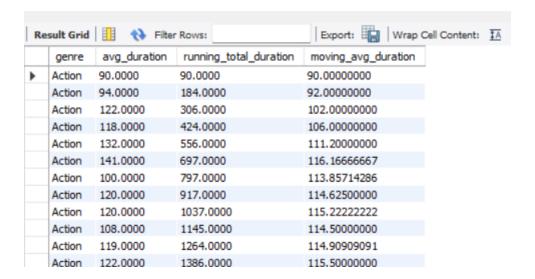
running\_total\_duration,

#### **SELECT**

g.genre,

AVG(m.duration) AS avg\_duration, SUM(AVG(m.duration)) OVER (PARTITION BY g.genre ORDER BY m.id) AS

```
AVG(AVG(m.duration)) OVER (PARTITION BY g.genre ORDER BY m.id) AS moving_avg_duration FROM imdb.genre g
JOIN imdb.movie m ON g.movie_id = m.id
GROUP BY g.genre, m.id
ORDER BY g.genre, m.id;
```



- -- Round is good to have and not a must have; Same thing applies to sorting
- -- Let us find top 5 movies of each year with top 3 genres.

Q26. Which are the five highest-grossing movies of each year that belong to the top three genres?

-- (Note: The top 3 genres would have the most number of movies.)

/* Output format:			+		+	
	 	+ year			movie_name	
		+	-+			
comedy			2017		indian	
				•	I	
				•	I	
					I	

```
+----
----+*/
-- Type your code below:
-- Top 3 Genres based on most number of movies
Sol:
WITH RankedMovies AS (
 SELECT
   m.id,
   m.title as movie_name,
   m.year,
   m.worlwide_gross_income,
   g.genre,
   RANK() OVER (PARTITION BY m.year, g.genre ORDER BY m.worlwide_gross_income
DESC) AS movie_rank
 FROM
   imdb.movie m
 JOIN
   imdb.genre g ON m.id = g.movie_id
SELECT
 genre,
 year,
 movie name,
 worlwide_gross_income,
 movie_rank
FROM
 RankedMovies
WHERE
 movie rank <= 5
ORDER BY
 genre, year, movie_rank;
```

Re	Result Grid   Filter Rows: Export: Wrap Cell Content: IA						
	genre	year	movie_name	worlwide_gross_income	movie_rank		
•	Action	2017	Winner	INR 250000000	1		
	Action	2017	Beyond Skyline	\$ 992181	2		
	Action	2017	Zashchitniki	\$ 9765483	3		
	Action	2017	V.I.P.	\$ 9710283	4		
	Action	2017	Overdrive	\$ 9650552	5		
	Action	2018	The Villain	INR 1300000000	1		
	Action	2018	Simmba	\$ 9865268	2		
	Action	2018	Sin-gwa ham-kke: In-gwa yeon	\$ 97962238	3		
	Action	2018	Traffik	\$ 9515914	4		
	Action	2018	Ying	\$ 91708374	5		
_	1. 5.4	2010					

-- Finally, let's find out the names of the top two production houses that have produced the highest number of hits among multilingual movies.

# Q27. Which are the top two production houses that have produced the highest number of hits (median rating >= 8) among multilingual movies?

```
/* Output format:
+----+
|production company |movie count
  prod_comp rank|
+----+
| The Archers |
                        830
   •
                        1
                       +----+*/
-- Type your code below:
SELECT production company,
       COUNT(id) AS movie_count,
       DENSE RANK() OVER (ORDER BY COUNT(id) DESC) AS
prod comp rank
FROM movie m
INNER JOIN ratings r
ON m.id = r.movie id
WHERE median rating >= 8 AND languages LIKE '%,%' AND
production_company IS NOT NULL
GROUP BY production company
LIMIT 2;
Result Grid | III 💎 Filter Rows: | Export:
```

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

-- Multilingual is the important piece in the above question. It was created using POSITION(',' IN languages)>0 logic -- If there is a comma, that means the movie is of more than one language

#### **QUES 28:**

-Q28. Who are the top 3 actresses based on number of Super Hit movies
(average rating >8) in drama genre?

/\* Output format:
+-----+
| actress\_name | total\_votes | movie\_count | actress
+-----+
| Laura Dern | 1016 | 1

#### **Ques 29:**

Q29. Get the following details for top 9 directors (based on number of movies)
Director id
Name
Number of movies
Average inter movie duration in days
Average movie ratings
Total votes
Min rating
Max rating

#### total movie durations

AVG(r.avg\_rating) AS avg\_rating,

```
Format:
+----
_______
----+
| director id | director name | number of movies |
    avg inter movie days | avg rating
                                | total votes |
min rating | max rating | total duration |
______
                                                5
|nm1777967
              A.L. Vijay
______
____*/
-- Type you code below:
Sol:
SELECT
 dm.name_id AS director_id,
 n.name AS director_name,
 COUNT(DISTINCT dm.movie_id) AS number_of_movies,
 AVG(DATEDIFF(m.date_published, COALESCE((SELECT MAX(m_prev.date_published)
             FROM movie m_prev
             WHERE m_prev.id < m.id
             AND m_prev.id IN (SELECT dm_prev.movie_id
                  FROM director_mapping dm_prev
                  WHERE dm_prev.name_id = dm.name_id
            ), m.date_published))) AS avg_inter_movie_days,
```

```
SUM(r.total_votes) AS total_votes,
  MIN(r.avg_rating) AS min_rating,
  MAX(r.avg_rating) AS max_rating,
  SUM(m.duration) AS total_duration
FROM
  imdb.director_mapping dm
JOIN
  imdb.names n ON dm.name_id = n.id
JOIN
  imdb.movie m ON dm.movie_id = m.id
JOIN
  imdb.ratings r ON m.id = r.movie_id
GROUP BY
  dm.name_id
ORDER BY
  number_of_movies DESC
LIMIT 9;
```

# **Executive Summary**

- There are 13 distinct genres on which RSVP movies can make a movie.
- The highest number of movies released in March while the lowest number of movie releases was in December.
- USA and India are producing huge number of movies each year.
- Most movies produced in Drama genre followed by Comedy and Thriller. Hence focus on these categories will make RSVP successful.
- While producing Action genre has to be high(112.88 min) followed by Romance and Crime genres.
- Production Houses like Dream Warrior Pictures and National Theater Live Pictures can be considered as they have produced most of the hit movies and having average rating greater than 8.
- German movies will be more profitable compared to Italian movies based on the highest votes.
- Highest votes received by Marvel Movies followed by twentieth Century Fox and Warner Bros. Hence these can be considered as world wide release partner.
- James Mangold can be cosidered for next project as he is the top director in top 3 genres with highest superhit movies.
- For casting, the top 2 actors Mammootty n Mohanlal should be considered for the next project as they have record of most superhit movies.
- Based on the superhit movies Parvathi Thiruvorathu should be cosidered for actress
- Top Actor & Actress in Indian Movies are Vijaysethupathi and Tapsipannu

Therefore it would be a great success if RSVP movies produce a Drama film with James Mangold as Director, Dream Warior Pictures or National Theater Live as Production House also considering Marvel Studios for Box Office Success and choosing Mammotty, Mohanlal, Vijay sethupathi, Parvathy Thiruvorathu or Tapseepannu for casting.