DATA ANALYSIS ON IMDB DATASET

MySQL

OESON GLOBAL

Submitted by:

Monika Saxena

Meera Radhish

Supriya Meshram

Supervised by: Aritri Debnath

Monika Mathur Meera Radhish Supriya Meshram

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.

Monika Mathur Meera Radhish Supriya Meshram

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           |      2134                  |
|      2018           |              .                     |
|      2019           |              .                     |
+---------------+-------------------+


Output format for the second part of the question:
+---------------+-------------------+
|      month_num      |      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.

Ques 4

```
/*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;
```

	genre
▶	Drama
	Fantasy
	Thriller
	Comedy
	Horror

Family
Romance
Adventure
Action
Sci-Fi

Crime
Mystery
Others

Ques 6

```
/* 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

```
/* 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:

+---------------+-------------------+
| genre         |      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
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)


/* Output format:
+---------------+-------------------+---------------------+
| genre                |              movie_count    |
       genre_rank     |
+---------------+-------------------+---------------------+
|drama                 |      2312                   |                    2
+---------------+-------------------+---------------------+*/
-- Type your code below:
```

Sol 9:

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;
-- 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	3

-- 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:

```
+---------------+-------------------+---------------------+------------
----------+-----------------+-----------------+
```

```
| min_avg_rating|     max_avg_rating |       min_total_votes    |
       max_total_votes        |min_median_rating|min_median_rating|
+---------------+-------------------+---------------------+------------
----------+-----------------+-----------------+
|              0              |                     5              |
+---------------+-------------------+---------------------+------------
----------+-----------------+-----------------+*/
-- Type your code below:
```

Sol 10:

```
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;
```

Ques 11

Q11 : Which are the top 10 movies based on average rating?

```
/* Output format:
+---------------+-------------------+---------------------+
| title         |     avg_rating    |     movie_rank      |
+---------------+-------------------+---------------------+
| Fan           |       9.6         |          5          |
|     .         |        .          |          .          |
|     .         |        .          |          .          |
|     .         |        .          |          .          |
+---------------+-------------------+---------------------+*/
```

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:

Result Grid | Filter Rows: | Export:

title	avg_rating	movie_rank
Kirket	10.0	1
Love in Kilnerry	10.0	1
Gini Helida Kathe	9.8	3
Runam	9.7	4
Fan	9.6	5
Android Kunjappan Version 5.25	9.6	5
Yeh Suhaagraat Impossible	9.5	7
Safe	9.5	7
The Brighton Miracle	9.5	7
Shibu	9.4	10

Movie FAN has rating 9.6

Ques 12

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

```
/* Output format:
+---------------+-------------------+
| median_rating |      movie_count  |
+-------------------+----------------
|      1        |        105        |
|      .        |        .          |
|      .        |        .          |
+---------------+-------------------+ */
```

Sol

```
select median_rating, count(movie_id) as movie_count
```

from ratings

group by median_rating

order by movie_count desc;

Output:

Result Grid | Filter Rows:

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

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:

```
+-----------------+-----------------+--------------------+
|production_company|movie_count      |  prod_company_rank|
+-----------------+-----------------+--------------------+
| The Archers     |        1        |          1         |
+-----------------+-----------------+--------------------+*/
```

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:

Result Grid | Filter Rows: | Export:

production_company	movie_count	company_rank
Dream Warrior Pictures	3	1
National Theatre Live	3	1
Lietuvos Kinostudija	2	3
Swadharm Entertainment	2	3
Panorama Studios	2	3
Marvel Studios	2	3
Central Base Productions	2	3
Painted Creek Productions	2	3
National Theatre	2	3
Colour Yellow Productions	2	3
The Archers	1	11
Blaze Film Enterprises	1	11
Bradeway Pictures	1	11
Bert Marcus Productions	1	11
A Studios	1	11

***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:

```
+---------------+-------------------+
| genre         |  movie_count      |
+-------------------+---------------
|     thriller  |   105             |
|     .         |        .          |
|     .         |        .          |
+---------------+-------------------+ */
```

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:

Result Grid | Filter Rows:

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
Family	1

Ques 15

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

/* Output format:

```
+---------------+-------------------+---------------------+
| title         |     avg_rating    |        genre        |
+---------------+-------------------+---------------------+
| Theeran       |       8.3         |      Thriller       |
```

```
|        .                    |              .                    |                    .
                 |
|        .                    |              .                    |                    .
                 |
|        .                    |              .                    |                    .
                 |
+---------------+-------------------+---------------------+*/
```

Sol

```
use imdb;
select title, avg_rating, genre
from movie
inner join genre on movie.id=genre.movie_id
inner join ratings on movie.id =ratings.movie_id
where avg_rating > 8
and title like 'The%'
order by avg_rating desc;
```

Output:

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

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

Ques 16

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:

Result Grid | Filter Rows:

median_rating	movie_count
8	361

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:

Result Grid | Filter Rows:

country	total_number_votes
Germany	106710
Italy	77965

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

+----------+----------+----------+----------+

| name_nulls | height_nulls |date_of_birth_nulls |known_for_movies_nulls|

+----------+----------+----------+----------+

| 0 | 123 | 1234 | 12345 |

+----------+----------+----------+----------+*/

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:

Result Grid | Filter Rows: | Export: | Wrap Cell

name_nulls	height_nulls	date_of_birth_nulls	known_for_movies_nulls
0	17335	13431	15226

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  |        4          |
|      .        |        .          |
|      .        |        .          |
+---------------+-------------------+ */
```

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
LIMIT 3;
```

Output:

Result Grid | Filter Rows:

director_name	movie_count
James Mangold	4
Joe Russo	3
Anthony Russo	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:

+----------+----------+

| actor_name | movie_count |

+----------+----------

|Christain Bale | 10 |

| . | . |

+----------+----------+ */

Sol:

```
select
  nam.name as actor_name,
     Count(movie_id) as movie_count
from role_mapping as rm
     inner join movie as mov
          on mov.id = rm.movie_id
      join ratings as rat using(movie_id)
     inner join names as nam
          on nam.id = rm.name_id
where rat.median_rating >= 8
            and category = 'actor'
group by actor_name
order by movie_count desc limit 2;
```

Output:

	actor_name	movie_count
▸	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.*/

Ques 21

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

/* Output format:

```
+-----------------+-------------------+--------------------+
|production_company|vote_count                              |                    prod_comp_rank|
+-----------------+-------------------+--------------------+
| The Archers                    |                    830                          |                    1
|          .                                          |                    .                                |
|          .                                          |                    .                                |
+-------------------+-------------------+--------------------+*/
```

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;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

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.

-- (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:
+---------------+-------------------+---------------------+------------
----------+-----------------+
| actor_name    |      total_votes          |      movie_count            |     act
+---------------+-------------------+---------------------+------------
----------+-----------------+
|      Yogi 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
    ratings r        ON   r.movie_id =  rm.movie_id
    JOIN
    movie m                    ON                    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:

actor_name	sum_total_votes	movie_count	actor_avg_rating	actor_rank
Vijay Sethupathi	23114	5	8.42	1

Vijay Sethupathi is in top actor list

Ques 23

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.)

```
/* Output format:
+---------------+-------------------+---------------------+------------
----------+-----------------+
| actress_name |      total_votes          |      movie_count            |     act
+---------------+-------------------+---------------------+------------
----------+-----------------+
|      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 (
                                                ORDER BY
ROUND(SUM(avg_rating * r.total_votes) / SUM(r.total_votes), 2)
desc,
                                SUM(r.total_votes) DESC
                                                ) AS actress_rank

FROM `names` n
      JOIN
      role_mapping rm    ON    n.id = rm.name_id
      JOIN
      ratings r          ON    r.movie_id =  rm.movie_id
      JOIN
      movie m              ON             m.id = r.movie_id

WHERE m.languages LIKE '%Hindi%' AND
         rm.category LIKE '%Actress%' AND
      country='india'

GROUP BY actress_name
HAVING movie_count >=3
LIMIT 5;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

actress_name	sum_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

```
/* 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
JOIN
   imdb.genre g ON m.id = g.movie_id
WHERE
   g.genre = 'Thriller'
GROUP BY
   m.id, m.title
ORDER BY
   avg_rating DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

id	title	avg_rating	movie_category
tt10869474	Safe	9.50000	Superhit movies
tt4897596	Digbhayam	9.20000	Superhit movies
tt9390200	Dokyala Shot	9.20000	Superhit movies
tt6271432	Abstruse	9.00000	Superhit movies
tt9900782	Kaithi	8.90000	Superhit movies
tt10975452	Raju Gari Gadhi 3	8.80000	Superhit movies
tt2311530	Lost Angelas	8.80000	Superhit movies
tt5266470	Enigma	8.80000	Superhit movies
tt7286456	Joker	8.80000	Superhit movies
tt8364132	Birbal Trilogy	8.80000	Superhit movies
tt6148156	Vikram Vedha	8.70000	Superhit movies
tt7060344	Ratsasan	8.70000	Superhit movies
tt9378950	Ghost	8.70000	Superhit movies
tt9430780	Bell Bottom	8.70000	Superhit movies

```
/* 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.)

```
/* Output format:
+---------------+-------------------+---------------------+------------
----------+
| genre                |       avg_duration
       |running_total_duration|moving_avg_duration  |
+---------------+-------------------+---------------------+------------
----------+
|       comdy          |                     145            |
106.2     |        128.42              |
|              .              |                     .              |              .
|              .              |                     .              |              .
|              .              |                     .              |              .
+---------------+-------------------+---------------------+------------
----------+*/


-- Type your code below:
```

Sol:

```
SELECT
   g.genre,
   AVG(m.duration) AS avg_duration,
   SUM(AVG(m.duration)) OVER (PARTITION BY g.genre ORDER BY m.id) AS
running_total_duration,
```

```
   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;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

genre	avg_duration	running_total_duration	moving_avg_duration
Action	90.0000	90.0000	90.00000000
Action	94.0000	184.0000	92.00000000
Action	122.0000	306.0000	102.00000000
Action	118.0000	424.0000	106.00000000
Action	132.0000	556.0000	111.20000000
Action	141.0000	697.0000	116.16666667
Action	100.0000	797.0000	113.85714286
Action	120.0000	917.0000	114.62500000
Action	120.0000	1037.0000	115.22222222
Action	108.0000	1145.0000	114.50000000
Action	119.0000	1264.0000	114.90909091
Action	122.0000	1386.0000	115.50000000

```
-- 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.
```

Ques 26

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:
+---------------+-------------------+---------------------+------------
----------+-----------------+
| genre                |       year                  |       movie_name              |
+---------------+-------------------+---------------------+------------
----------+-----------------+
|       comedy         |                      2017    |              indian    |
|              .              |                      .              |              .
|              .              |                      .              |              .
|              .              |                      .              |              .
```

```
+---------------+-------------------+---------------------+------------
----------+-----------------+*/
-- 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;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

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

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

Ques 27

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:
```

Sol:

```
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 | 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                |
```

```
|              .              |                    .              |
|              .              |                    .              |
+---------------+------------------+--------------------+------------
----------+----------------+*/
-- Type your code below:
```

Sol:

```
WITH actress_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),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
              INNER JOIN GENRE AS g
                   ON g.movie_id = m.id
       WHERE lower(category) = 'actress'
               AND avg_rating > 8
               AND lower(genre) = "drama"
       GROUP BY name )
SELECT *,
          Rank() OVER(ORDER BY movie_count DESC) AS
actress_rank
FROM actress_summary LIMIT 3;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

actress_name	total_votes	movie_count	actress_avg_rating	actress_rank
Parvathy Thiruvothu	4974	2	8.25	1
Susan Brown	656	2	8.94	1
Amanda Lawrence	656	2	8.94	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

```
Format:
+---------------+-------------------+---------------------+------------
----------+--------------+--------------+------------+------------+----
------------+
| director_id  |       director_name  |       number_of_movies  |
       avg_inter_movie_days |       avg_rating     | total_votes  |
min_rating     | max_rating | total_duration |
+---------------+-------------------+---------------------+------------
----------+--------------+--------------+------------+------------+----
------------+
|nm1777967            |       A.L. Vijay            |                     5
|       .                     |              .                      |
|       .                     |              .                      |
|       .                     |              .                      |
|       .                     |              .                      |
|       .                     |              .                      |
|       .                     |              .                      |
|       .                     |              .                      |
|       .                     |              .                      |
+---------------+-------------------+---------------------+------------
----------+--------------+--------------+------------+------------+----
------------+

-----------------------------------------------------------------------
---------------------*/
-- 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,
   AVG(r.avg_rating) AS avg_rating,
```

```
    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.