```
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:




-- Q1. Find the total number of rows in each table of the schema?
-- Type your
code below:

SELECT COUNT(*) FROM director_mapping;

SELECT COUNT(*) FROM GENRE ;

SELECT
COUNT(*) FROM  MOVIE;

SELECT COUNT(*) FROM  NAMES;

SELECT COUNT(*) FROM  RATINGS;

SELECT
COUNT(*) FROM  ROLE_MAPPING;



-- Q2. Which columns in the movie table have null values?
--
Type your code below:

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 title_NULL_COUNT,
       Sum(CASE
             WHEN year IS NULL
THEN 1
             ELSE 0
           END) AS year_NULL_COUNT,
       Sum(CASE

WHEN date_published IS NULL THEN 1
             ELSE 0
           END) AS
date_published_NULL_COUNT,
       Sum(CASE
             WHEN duration IS NULL THEN 1

  ELSE 0
           END) AS duration_NULL_COUNT,
       Sum(CASE
             WHEN country IS
NULL THEN 1
             ELSE 0
           END) AS country_NULL_COUNT,
       Sum(CASE

    WHEN worlwide_gross_income IS NULL THEN 1
```

```
            ELSE 0
          END) AS
worlwide_gross_income_NULL_COUNT,
       Sum(CASE
             WHEN languages IS NULL THEN 1

          ELSE 0
           END) AS languages_NULL_COUNT,
       Sum(CASE
             WHEN
production_company IS NULL THEN 1
             ELSE 0
           END) AS
production_company_NULL_COUNT
FROM   movie;



-- 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:

SELECT year,
       Count(title) AS NUMBER_OF_MOVIES
FROM   movie
GROUP  BY year;


SELECT Month(date_published) AS MONTH_NUM,
       Count(*)              AS
NUMBER_OF_MOVIES
FROM   movie
GROUP  BY month_num
ORDER  BY month_num;



/*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:

-- Pattern matching using LIKE operator for country
column
SELECT Count(DISTINCT id) AS number_of_movies, year
FROM   movie
WHERE  ( country LIKE
'%INDIA%'
          OR country LIKE '%USA%' )
       AND year = 2019;

-- 1059 movies were
produced in the USA or India in the year 2019


/* 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:

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:


SELECT     genre,

   Count(m.id) AS number_of_movies
FROM       movie       AS m
INNER JOIN genre       AS
g
where      g.movie_id = m.id
GROUP BY   genre
ORDER BY   number_of_movies DESC limit 1
;



/* 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:


-- Using genre table to find
movies which belong to only one genre
-- Grouping rows based on movie id and finding the
distinct number of genre each movie belongs to
-- Using the result of CTE, we find the count of
movies which belong to only one genre
WITH movies_with_one_genre
     AS (SELECT movie_id
```

```
   FROM   genre
          GROUP  BY movie_id
          HAVING Count(DISTINCT genre) = 1)
SELECT
Count(*) AS movies_with_one_genre
FROM   movies_with_one_genre;

-- 3289 movies belong to only
one genre


/* 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:

--
Finding the average duration of movies by grouping the genres that movies belong to
SELECT
genre,
           Round(Avg(duration),2) AS avg_duration
FROM       movie                  AS
m
INNER JOIN genre                  AS g
ON      g.movie_id = m.id
GROUP BY   genre
ORDER BY
avg_duration DESC;

-- Action genre has the highest duration of 112.88 seconds followed by
romance and crime genres.


/* 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:

--
CTE : Finds the rank of each genre based on the number of movies in each genre
-- Select query
displays the genre rank and the number of movies belonging to Thriller genre
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" ;

-- Thriller has
rank=3 and movie count of 1484


/*Thriller movies is in top 3 among all genres in terms of
number of movies
 In the previous segment, you analysed the movies and genres tables.
 In this
segment, you will analyse the ratings table as well.
To start with lets get the min and max
values of different columns in the table*/




-- Segment 2:


-- 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                |               177
2000                           |               0              |        8
|
+---------------+------------------+---------------------+----------------------+-----------
------+-----------------+*/
-- Type your code below:

-- Using MIN and MAX functions for the
query
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;


/* So, the minimum and maximum values in each column of
the ratings table are in the expected range.
This implies there are no outliers in the table.
```

```
Now, let’s find out the top 10 movies based on average rating.*/

-- Q11. Which are the top
10 movies based on average rating?
/* Output
format:
+---------------+-------------------+---------------------+
|
title                        |                avg_rating        |                movie_rank
|
+---------------+-------------------+---------------------+
| Fan                        |                9.6                        |
|
|        .                          |                .                          |
|        .                          |                .                          |
|        .                          |                .                          |
|
+---------------+-------------------+---------------------+*/
-- Type your code below:
--
It's ok if RANK() or DENSE_RANK() is used too

-- Finding the rank of each movie based on it's
average rating
-- Displaying the top 10 movies using LIMIT clause
SELECT     title,

avg_rating,
           Rank() OVER(ORDER BY avg_rating DESC) AS movie_rank
FROM       ratings
                           AS r
INNER JOIN movie                                  AS m
ON
  m.id = r.movie_id limit 10;

WITH MOVIE_RANK AS
(
SELECT     title,
           avg_rating,

        ROW_NUMBER() OVER(ORDER BY avg_rating DESC) AS movie_rank
FROM       ratings
                   AS r
INNER JOIN movie                                  AS m
ON         m.id =
r.movie_id
)
SELECT * FROM MOVIE_RANK
WHERE movie_rank<=10;

-- Top 3 movies have average
rating >= 9.8


/* Do you find you favourite movie FAN in the top 10 movies with an average
rating of 9.6? If not, please check your code again!!
So, now that you know the top 10 movies,
do you think character actors and filler actors can be from these movies?
Summarising the
ratings table based on the movie counts by median rating can give an excellent insight.*/

--
Q12. Summarise the ratings table based on the movie counts by median ratings.
/* Output
format:

+---------------+-------------------+
|
median_rating        |        movie_count                |
+-------------------+----------------
|        1                          |                105                          |
|        .                          |
.                        |
|        .                          |                .                          |
+---------------+-------------------+ */
```

```
-- Type your code below:
-- Order
by is good to have

-- Finding the number of movies vased on median rating and sorting based on
movie count.
SELECT median_rating,
       Count(movie_id) AS movie_count
FROM   ratings
GROUP
BY median_rating
ORDER  BY movie_count DESC;


/* Movies with a median rating of 7 is highest
in number.
Now, let's find out the production house with which RSVP Movies can partner for its
next project.*/

-- Q13. Which production house has produced the most number of hit movies
(average rating > 8)??
/* Output
format:
+-----------------+-----------------+--------------------+
|production_company|movi
e_count
|        prod_company_rank|
+-----------------+-----------------+--------------------+
| The
Archers           |               1                   |                          1
|
+-----------------+-----------------+--------------------+*/
-- Type your code below:

--
CTE: Finding the rank of production company based on movie count with average rating > 8
using RANK function.
-- Querying the CTE to find the production company with rank=1
WITH
production_company_hit_movie_summary
     AS (SELECT production_company,

Count(movie_id)                     AS MOVIE_COUNT,
               Rank()

OVER(
                   ORDER BY Count(movie_id) DESC ) AS PROD_COMPANY_RANK
         FROM
ratings AS R
               INNER JOIN movie AS M
                       ON M.id =
R.movie_id
         WHERE  avg_rating > 8
                AND production_company IS NOT
NULL
         GROUP  BY production_company)
SELECT *
FROM
production_company_hit_movie_summary
WHERE  prod_company_rank = 1;

-- Dream Warrior Pictures
and National Theatre Live production houses has produced the most number of hit movies (average
rating > 8)
-- They have rank=1 and movie count =3

-- It's ok if RANK() or DENSE_RANK() is
used too
-- Answer can be Dream Warrior Pictures or National Theatre Live or both


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

SELECT
genre,
       Count(M.id) AS MOVIE_COUNT
FROM   movie AS M
       INNER JOIN genre AS G

       ON G.movie_id = M.id
       INNER JOIN ratings AS 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;

-- 24 Drama movies
were released during March 2017 in the USA and had more than 1,000 votes
-- Top 3 genres are
drama, comedy and action during March 2017 in the USA and had more than 1,000 votes




-- Lets
try to analyse with a unique problem statement.
-- 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
|
|        .                        |                .                          |
|        .                        |                .                          |
|        .                        |                .                          |
|
+---------------+-------------------+---------------------+*/
-- Type your code below:

--
Query to find:
-- 1. Number of movies of each genre that start with the word ‘The’ (LIKE
operator is used for pattern matching)
-- 2. Which have an average rating > 8?
-- Grouping
by title to fetch distinct movie titles as movie belog to more than one
genre

SELECT
title,
avg_rating,
genre
FROM movie m
INNER JOIN ratings r
ON m.id =
r.movie_id
```

```
INNER JOIN genre g
ON g.movie_id = m.id
WHERE title LIKE'The%' AND
avg_rating>8
ORDER BY avg_rating DESC;

-- There are 8 movies which begin with
"The" in their title.
-- The Brighton Miracle has highest average rating of 9.5.
--
All the movies belong to the top 3 genres.




-- You should also try your hand at median
rating and check whether the ‘median rating’ column gives any significant insights.
-- Q16.
Of the movies released between 1 April 2018 and 1 April 2019, how many were given a median
rating of 8?
-- Type your code below:

-- BETWEEN operator is used to find the movies released
between 1 April 2018 and 1 April 2019
SELECT median_rating, Count(*) AS movie_count
FROM
movie AS M
       INNER JOIN ratings AS 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;

-- 361 movies have released between 1 April 2018 and 1 April 2019 with a median
rating of 8



-- Q17. Do German movies get more votes than Italian movies?
-- Hint: Here you
have to find the total number of votes for both German and Italian movies.
-- Type your code
below:

SELECT country, sum(total_votes) AS votes_count
FROM movie as m
INNER JOIN ratings as
r
ON r.movie_id=m.id
WHERE country = 'germany' OR country = 'italy'
GROUP BY country;


-- By
observation, German movies received the highest number of votes when queried against language
and country columns.

-- Answer is Yes


/* Now that you have analysed the movies, genres and
ratings tables, let us now analyse another table, the names table.
Let’s begin by searching
for null values in the tables.*/




-- Segment 3:


-- 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                |               123
|
+---------------+-------------------+---------------------+----------------------+*/
-- Type
your code below:
-- NULL counts for individual columns of names table
SELECT Count(*) AS
name_nulls
FROM   names
WHERE  NAME IS NULL;

SELECT Count(*) AS height_nulls
FROM
names
WHERE  height IS NULL;

SELECT Count(*) AS date_of_birth_nulls
FROM   names
WHERE
date_of_birth IS NULL;

SELECT Count(*) AS known_for_movies_nulls
FROM   names
WHERE
known_for_movies IS NULL;

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

-- Height, date_of_birth,
known_for_movies columns contain NULLS

/* There are no Null value in the column 'name'.
The
director is the most important person in a movie crew.
Let’s find out the top three
directors in the top three genres who can be hired by RSVP Movies.*/

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

-- CTE: Computes the top 3 genres using average rating > 8 condition and
highest movie counts
```

```
-- Using the top genres derived from the CTE, the directors are found
whose movies have an average rating > 8 and are sorted based on number of movies made.

WITH top_3_genres AS
(
           SELECT     genre,
                      Count(m.id)
                 AS movie_count ,
                      Rank() OVER(ORDER BY Count(m.id) DESC)
AS genre_rank
           FROM       movie                                     AS m

INNER JOIN genre                                    AS g
           ON         g.movie_id = m.id

          INNER JOIN ratings AS r
           ON         r.movie_id = m.id
           WHERE
avg_rating > 8
           GROUP BY   genre limit 3 )
SELECT     n.NAME            AS
director_name ,
           Count(d.movie_id) AS movie_count
FROM       director_mapping  AS
d
INNER JOIN genre G
using     (movie_id)
INNER JOIN names AS n
ON         n.id =
d.name_id
INNER JOIN top_3_genres
using     (genre)
INNER JOIN ratings
using
(movie_id)
WHERE      avg_rating > 8
GROUP BY   NAME
ORDER BY   movie_count DESC limit 3
;

-- James Mangold , Joe Russo and Anthony Russo are top three directors in the top three
genres whose movies have an average rating > 8



/* James Mangold can be hired as the
director for RSVP's next project. Do you remeber his movies, 'Logan' and 'The Wolverine'.
Now,
let’s find out the top two actors.*/

-- Q20. Who are the top two actors whose movies have a
median rating >= 8?
/* Output format:

+---------------+-------------------+
|
actor_name        |        movie_count                |
+-------------------+----------------
|Christain
Bale        |                10                         |
|        .                        |                .                          |
+---------------+-------------------+ */
-- Type your code
below:

SELECT N.name          AS actor_name,
       Count(movie_id) AS movie_count
FROM
role_mapping AS RM
       INNER JOIN movie AS M
               ON M.id = RM.movie_id

INNER JOIN ratings AS R USING(movie_id)
       INNER JOIN names AS N
               ON N.id =
```

```
RM.name_id
WHERE  R.median_rating >= 8
AND category = 'ACTOR'
GROUP  BY actor_name
ORDER  BY
movie_count DESC
LIMIT  2;

-- Top 2 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:
+------------------+-------------------+---------------------+
|production_company|vot
e_count                        |               prod_comp_rank|
+------------------+-------------------+---------------------+
|
The Archers                |               830                        |               1
|        .                                |                .                        |
|        .                                |                .                        |
|
+-------------------+-------------------+---------------------+*/
-- Type your code
below:


-- Approach 1: Using select statement
SELECT     production_company,

Sum(total_votes)                            AS vote_count,
           Rank() OVER(ORDER BY
Sum(total_votes) DESC) AS prod_comp_rank
FROM       movie
AS m
INNER JOIN ratings                                      AS r
ON         r.movie_id =
m.id
GROUP BY   production_company limit 3;

-- Approach 2: using CTEs
WITH ranking AS(
SELECT
production_company, sum(total_votes) AS vote_count,
        RANK() OVER(ORDER BY SUM(total_votes)
DESC) AS prod_comp_rank
FROM movie AS m
        INNER JOIN ratings AS r ON r.movie_id=m.id
GROUP BY
production_company)
SELECT production_company, vote_count, prod_comp_rank
FROM ranking
WHERE
prod_comp_rank<4;

-- Top three production houses based on the number of votes received by
their movies are Marvel Studios, Twentieth Century Fox and Warner Bros.



/*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.*/

-- 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                  |
|
+---------------+------------------+---------------------+---------------------+----------
------+
|        Yogi Babu        |                        3455        |               11
|                 .               |                         .               |
 .                 |           .                            |                .                |
|                 .               |                         .               |               .
|
|                 .               |                         .               |               .
|
+---------------+------------------+---------------------+---------------------+----------
------+*/
-- Type your code below:

WITH actor_summary AS (SELECT N.NAME

           AS actor_name,
                total_votes,

   Count(R.movie_id)
                AS movie_count,

               Round(Sum(avg_rating * total_votes) / Sum(total_votes), 2)
                AS
actor_avg_rating
         FROM   movie AS M
                INNER JOIN ratings AS R

            ON M.id = R.movie_id
                INNER JOIN role_mapping AS RM

       ON M.id = RM.movie_id
                INNER JOIN names AS N
                        ON
RM.name_id = N.id
         WHERE  category = 'ACTOR'
                AND country =
"india"
         GROUP  BY NAME
         HAVING movie_count >= 5)
SELECT
*,
Rank()OVER(ORDER BY actor_avg_rating DESC) AS actor_rank
FROM   actor_summary;


-- Top
actor is Vijay Sethupathi followed by Fahadh Faasil and Yogi Babu.

-- Top actor is Vijay
Sethupathi

-- 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                |
|actress_rank
|
+---------------+-------------------+---------------------+----------------------+----------
------+
|        Tabu                |                           3455        |                11
|                .                |                         .                |
.                  |          .                          |                .               |
|                .                |                         .                |             .
|
|                .                |                         .                |             .
|
+---------------+-------------------+---------------------+----------------------+----------
------+*/
-- Type your code below:
WITH actress_summary AS
(
           SELECT     n.NAME AS
actress_name,
                      total_votes,
                      Count(r.movie_id)
                             AS movie_count,

Round(Sum(avg_rating*total_votes)/Sum(total_votes),2) AS actress_avg_rating
           FROM
  movie                                                 AS m
           INNER JOIN ratings
                                        AS r
           ON         m.id=r.movie_id

INNER JOIN role_mapping AS rm
           ON         m.id = rm.movie_id
           INNER JOIN
names AS n
           ON         rm.name_id = n.id
           WHERE      category = 'ACTRESS'

         AND        country = "INDIA"
           AND        languages LIKE '%HINDI%'

          GROUP BY   NAME
           HAVING     movie_count>=3 )
SELECT   *,
         Rank()
OVER(ORDER BY actress_avg_rating DESC) AS actress_rank
FROM     actress_summary LIMIT 5;


--
Top five actresses in Hindi movies released in India based on their average ratings are Taapsee
Pannu, Kriti Sanon, Divya Dutta, Shraddha Kapoor, Kriti Kharbanda


/* 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.*/


/* 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:

-- Using CASE statements to classify thriller movies as per
avg rating
WITH thriller_movies
     AS (SELECT DISTINCT title,

avg_rating
         FROM   movie AS M
                INNER JOIN ratings AS R

      ON R.movie_id = M.id
                INNER JOIN genre AS G using(movie_id)
          WHERE
 genre LIKE 'THRILLER')
SELECT *,
       CASE
         WHEN avg_rating > 8 THEN 'Superhit
movies'
         WHEN avg_rating BETWEEN 7 AND 8 THEN 'Hit movies'
         WHEN avg_rating
BETWEEN 5 AND 7 THEN 'One-time-watch movies'
         ELSE 'Flop movies'
       END AS
avg_rating_category
FROM   thriller_movies;




/* 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.*/

-- Segment 4:

-- 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_dura
|
+---------------+-------------------+---------------------+----------------------+
|        comdy                 |
                        145                |               106.2          |           128.42
|                .               |                          .                |               .
|
|                .               |                          .                |               .
|                .               |                          .                |               .
|
+---------------+-------------------+---------------------+----------------------+*/
-- Type
your code below:


SELECT genre,
                ROUND(AVG(duration),2) AS avg_duration,

SUM(ROUND(AVG(duration),2)) OVER(ORDER BY genre ROWS UNBOUNDED PRECEDING) AS
running_total_duration,
        AVG(ROUND(AVG(duration),2)) OVER(ORDER BY genre ROWS 10
PRECEDING) AS moving_avg_duration
FROM movie AS m
INNER JOIN genre AS g
ON m.id=
g.movie_id
GROUP BY genre
ORDER BY genre;
```

```
-- 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:
+---------------+-------------------+---------------------+----------------------+-----
------------+
| genre                         |        year                          |        movie_name
|
+---------------+-------------------+---------------------+----------------------+-----------
------+
|        comedy                 |                          2017        |                indian
|                .                 |                          .                 |
      .                   |           .                          |                .
|                .                 |                          .                 |                .
|
|                .                 |                          .                 |                .
|
+---------------+-------------------+---------------------+----------------------+-----------
------+*/
-- Type your code below:

-- Top 3 Genres based on most number of movies


WITH
top3_genre
AS
(
SELECT
genre,
COUNT(movie_id) as movie_count
FROM genre
GROUP BY genre
ORDER BY
movie_count DESC
LIMIT 3
),
top5_movie
AS
(
SELECT
genre,
YEAR,
title as
movie_name,
worlwide_gross_income,
DENSE_RANK() OVER(PARTITION BY year ORDER BY
worlwide_gross_income DESC) AS movie_rank
FROM movie m
INNER JOIN genre g
ON m.id =
g.movie_id
WHERE genre IN(SELECT genre FROM top3_genre)
)
SELECT *
FROM top5_movie
WHERE
movie_rank<=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.
-- 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:
WITH production_company_summary
     AS (SELECT production_company,

Count(*) AS movie_count
         FROM   movie AS m
                inner join ratings AS r

                   ON r.movie_id = m.id
         WHERE  median_rating >= 8

AND production_company IS NOT NULL
                AND Position(',' IN languages) > 0

  GROUP  BY production_company
         ORDER  BY movie_count DESC)
SELECT *,
       Rank()

     over(
           ORDER BY movie_count DESC) AS prod_comp_rank
FROM
production_company_summary
LIMIT 2;

-- Star Cinema and Twentieth Century Fox are the top two
production houses that have produced the highest number of hits among multilingual
movies.




-- 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


-- 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_rank
|
+---------------+-------------------+---------------------+----------------------+----------
------+
|        Laura Dern        |                        1016        |               1
|                .                |                         .               |
 .                   |           .                              |                 .               |
|                .                |                         .               |               .
|
+---------------+-------------------+---------------------+----------------------+----------
------+*/
-- Type your code below:

-- Top 3 actresses based on number of Super Hit movies
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      category = 'ACTRESS'
           AND
avg_rating>8
           AND genre = "Drama"
           GROUP BY   NAME )
SELECT
*,
         Rank() OVER(ORDER BY movie_count DESC) AS actress_rank
FROM     actress_summary
LIMIT 3;

-- Top 3 actresses based on number of Super Hit movies are Parvathy Thiruvothu, Susan
Brown and Amanda Lawrence



/* 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_mo
total_votes  | min_rating        | max_rating | total_duration
|
+---------------+-------------------+---------------------+----------------------+----------
---+--------------+------------+------------+----------------+
|nm1777967                |        A.L.
Vijay                |                          5                  |                177
|
|        .                          |                .                          |
|
|        .                          |                .                          |
|
|        .                          |                .                          |
|
|        .                          |                .                          |
|
|        .                          |                .                          |
|
|        .                          |                .                          |
```

```
|
|        .                        |                .                        |
|
|        .                        |                .                        |
|
+---------------+------------------+--------------------+----------------------+-----------
---+-------------+-----------+-----------+---------------+

------------------------------
-------------------------------------------------------------*/
-- Type you code below:
WITH
next_date_published_summary AS
(
           SELECT     d.name_id,
                      NAME,

                    d.movie_id,
                      duration,

r.avg_rating,
                      total_votes,
                      m.date_published,

                Lead(date_published,1) OVER(partition BY d.name_id ORDER BY
date_published,movie_id ) AS next_date_published
           FROM       director_mapping
                                                                  AS d
           INNER JOIN names
                                                                               AS n

ON         n.id = d.name_id
           INNER JOIN movie AS m
           ON         m.id =
d.movie_id
           INNER JOIN ratings AS r
           ON         r.movie_id = m.id ),
top_director_summary AS
(
       SELECT *,
              Datediff(next_date_published,
date_published) AS date_difference
       FROM   next_date_published_summary )
SELECT   name_id
                      AS director_id,
         NAME                          AS director_name,

        Count(movie_id)               AS number_of_movies,

Round(Avg(date_difference),2) AS avg_inter_movie_days,
         Round(Avg(avg_rating),2)
       AS avg_rating,
         Sum(total_votes)              AS total_votes,

Min(avg_rating)               AS min_rating,
         Max(avg_rating)               AS
max_rating,
         Sum(duration)                 AS total_duration
FROM
top_director_summary
GROUP BY director_id
ORDER BY Count(movie_id) DESC limit 9;
```

EXECUTIVE SUMMARY

In the given IMDB data by RSVP movies following Insights are derived and wecame to the conclusion that:

- Most of the movies were produced in the month of March with an average of about 2300 movies per year.
- Drama was the most popular genre with 4285 number of movies and an avg duration of 106.7746. RSVP movies can focus on this genre for its future films. Action and Thriller genres also hold potential.
- Most of the movies were rated between 6-8 on a median rating scale. Aiming for 8+ on a median rating will increase chances of a superhit movie.
- Dream warrior Pictures and National Theatre Live had produced highest rated films. Star Cinema and Twentieth Century Fox are also good contenders due to high number of multilingual movies as the movie will be for Indian audience primarily.
- Top directors observed from the analysis are James Mangold, Anthony Russo, Joe Russo and Soubin Shahir. With the later 3 tied at #2 spot. RSVP can have its future projects with them.
- Mammooty and Mohanlal are the top actors with highest number of rating.
- Taapsee Pannu can be chosen as actress as she is on the top of the list with average rating 7.74 and maximum votes.
- Marvel Studios(1st) Twentieth Century Fox(2nd) & Warner Bros.(3rd) can be chosen as its Global Partners as number of votes received is maximum of their movies.
- In India, Vijay Sethupati can be focused on as he his popular here.
- Star Cinema and Twentieth Century Fox are the top two production houses that have produced the highest number of hits (median rating >= 8) among.