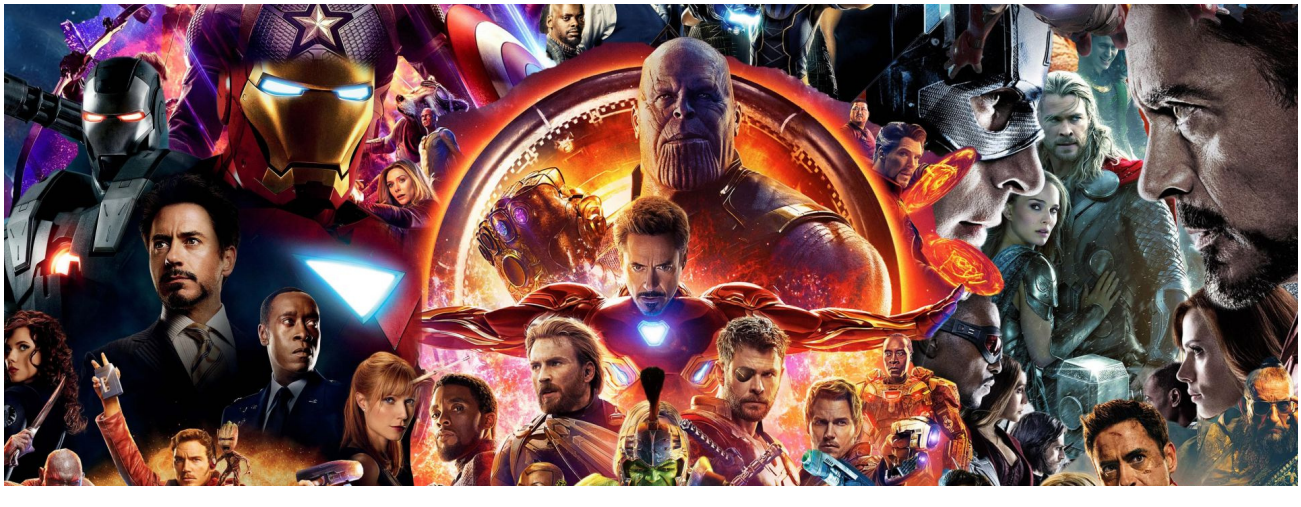
RSVP Movies CASE STUDY

Data science is the art of turning data into action.

Overview

This assignment aims to give an idea of applying Advance SQL in a real business scenario. In this assignment apart from applying techniques of SQL for Data Analysis I also developed a basic understanding of quantitative analysis and understand how data is used to draw meaningful insights that can help the production company to start their new project.



Business Understanding

Data analysis is the process of turning data into insights

RSVP Movies is an Indian film production company that 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 its every move analytically based on data and has approached you for help with this new project. You have been provided with the data on the movies that have been released in the past three years. You have to analyze the data set and draw meaningful insights that can help them start their new project.

Business Objective

Data analysis is the process of exploring, cleaning, transforming, and modeling data to extract useful information that supports decision-making

As a data analyst and an SQL expert. You have to use SQL to analyze the given data and give recommendations to RSVP Movies based on the insights.

SQL Queries

Without data, you're just another person with an opinion

```
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.*/
```

```sql
-- Segment 1:


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

SELECT table_name, table_rows
FROM INFORMATION_SCHEMA.TABLES
WHERE TABLE_SCHEMA = 'imdb';

/** OBSERVATIONS **/
-- Table_Name       |  Total number of rows
-- director_mapping | 3867
-- genre            | 14662
-- movie            | 7392
-- names            | 27467
-- ratings          | 8230
-- role_mapping     | 14315




-- 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_NULLS,
SUM(CASE WHEN title IS NULL THEN 1 ELSE 0 END) AS TITLE_NULLS,
SUM(CASE WHEN year IS NULL THEN 1 ELSE 0 END) AS YEAR_NULLS,
SUM(CASE WHEN date_published IS NULL THEN 1 ELSE 0 END) AS DATE_PUBLISHED_NULLS,
SUM(CASE WHEN duration IS NULL THEN 1 ELSE 0 END) AS DURATION_NULLS,
SUM(CASE WHEN country IS NULL THEN 1 ELSE 0 END) AS COUNTRY_NULLS,
SUM(CASE WHEN worlwide_gross_income IS NULL THEN 1 ELSE 0 END) AS WORLDWIDE_GROSS_INCOME_NULLS,
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_NULLS

FROM MOVIE;

/** OBSERVATIONS **/
-- worlwide_gross_income, production_company, languages, country have null values

--  Columns             | null_count
-- worlwide_gross_income |  3724
-- production_company    |  528
-- languages             |  194
-- country               |  20

-- Now as we 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)

-- Type your code below:
```

```sql
SELECT Year, count(id) AS number_of_movies
FROM movie
GROUP BY year
ORDER BY year;

/** OBSERVATIONS **/
--  year  | number_of_movies
-- 2017   |  3052
-- 2018   |  2944
-- 2019   |  2001


SELECT MONTH(date_published) AS month_num, COUNT(id) AS number_of_movies
FROM movie
GROUP BY month_num
ORDER BY month_num ;

/** OBSERVATIONS **/
-- The Highest number of movies is produced in the month of March = 824


/*The highest number of movies is produced in the month of March.
So, now we 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:

SELECT year, COUNT(id) AS number_of_movies
FROM movie
WHERE (country LIKE '%USA%' OR country LIKE '%India%')
AND year = 2019;

/** OBSERVATIONS **/
-- 1059 movies were produced in the UAS and India in the year 2019


/* USA and India produced more than a thousand movies(we 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;

/** OBSERVATIONS **/
-- Drama, Fantasy, Thriller, Comedy, Horror, Family, Romance, Adventure, Action, Sci-Fi, Crime, Mystery, Others.
-- There are 13 unique genre in the dataset.


/* So, RSVP Movies plans to make a movie of one of these genres.

Now, wouldn’t we 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:
```

```sql
SELECT genre, COUNT(movie_id) AS number_of_movies
FROM genre g
INNER JOIN movie m
ON g.movie_id = m.id
GROUP BY genre
ORDER BY number_of_movies desc LIMIT 1;

/** OBSERVATIONS **/
-- Drama(4285) has highest number of movies produced in overall.


/* So, based on the insight that we 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:

WITH count_genre AS
(
SELECT movie_id, COUNT(genre) AS number_of_movies
FROM genre
GROUP BY movie_id
HAVING Number_of_movies = 1
)

SELECT COUNT(movie_id) AS number_of_movies
FROM count_genre;

/** OBSERVATIONS **/
-- 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.)

-- Type your code below:


SELECT genre, ROUND(AVG(duration),2) AS avg_duration
FROM genre AS g
INNER JOIN movie AS m
ON g.movie_id = m.id
GROUP BY genre
ORDER BY AVG(duration) DESC;

/** OBSERVATIONS **/
-- movies of genre 'Drama' (produced highest in number in 2019) has the average duration of 106.77 mins.


/* Now we 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.*/
```

```sql
-- 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)

-- Type your code below:

WITH genre_rank 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_rank
WHERE genre = 'thriller';

/** OBSERVATIONS **/
-- Thriller is in top 3 among all genre in terms of number of movies.


/*Thriller movies is in top 3 among all genres in terms of number of movies

 In the previous segment, we analysed the movies and genres tables.
 In this segment, we will analyse the ratings table as well.
To start with lets get the min and max values of different columns in the table*/

 -- Q10.  Find the minimum and maximum values in  each column of the ratings table except the movie_id column?

 -- Type your code below:


 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?

-- Type your code below:
-- It's ok if RANK() or DENSE_RANK() is used too


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

/** OBSERVATIONS **/
-- Fan and Android Kunjappan Version 5.25 both have an average rating of 9.6.

/* So, now that we 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.
    Type your code below:
```

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


SELECT median_rating, COUNT(movie_id) AS movie_count
FROM ratings
GROUP BY median_rating
ORDER BY median_rating;

/** OBSERVATIONS **/
-- Movies with a median rating of 7 is highest in number i.e.2257.

/* 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)??

-- Type your code below:

SELECT production_company, COUNT(id) AS movie_count,
DENSE_RANK() OVER (ORDER BY COUNT(id) DESC) AS prod_company_rank
FROM movie AS m
INNER JOIN ratings AS r
ON m.id = r.movie_id
WHERE avg_rating > 8 AND production_company IS NOT NULL
GROUP BY production_company
ORDER BY movie_count DESC;

/** OBSERVATIONS **/
-- Dream Warrior Pictures or National Theatre Live or both
-- It's ok if RANK() or DENSE_RANK() is used too


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

-- Type your code below:

SELECT g.genre, COUNT(g.movie_id) AS movie_count
FROM genre g
INNER JOIN ratings rt
ON g.movie_id = rt.movie_id
INNER JOIN movie m
ON rt.movie_id = m.id
WHERE m.country LIKE '%USA%' AND rt.total_votes > 1000 AND MONTH(date_published) = 3
AND YEAR = 2017
GROUP BY g.genre
ORDER BY movie_count DESC;

/** OBSERVATIONS **/
-- Drama has the maximum number of movies in the March, 2017 and followed by Action.


-- Q15. Find movies of each genre that start with the word ‘The’ and which have an average rating > 8?

-- Type your code below:

SELECT title, avg_rating, genre
FROM genre AS g
INNER JOIN ratings AS r
```

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


-- We should also try our 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:

SELECT median_rating, count(movie_id) AS movie_count
FROM movie AS m
INNER JOIN ratings AS r
ON m.id = r.movie_id
WHERE median_rating= 8 AND date_published BETWEEN '2018-04-1' AND '2019-04-1'
ORDER BY avg_rating DESC;

/** OBSERVATIONS **/
-- 361 movies have released between 1 April 2018 and 1 April 2019.



-- Q17. Do German movies get more votes than Italian movies?
-- Type your code below:

SELECT country, sum(total_votes) AS total_votes
FROM movie AS mv
INNER JOIN ratings AS ra
ON mv.id = ra.movie_id
WHERE UPPER(country) LIKE 'GERMANY' or country LIKE 'ITALY'
GROUP BY country;


/** OBSERVATIONS **/
-- Yes German movies get more votes than Italian movies.

/* Now that we 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.*/

-- Q18. Which columns in the names table have null values??

-- Type your code below:

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;


/** OBSERVATIONS **/
-- No null values in name column.
-- height_nulls, date_of_birth_nulls, known_for_movies_nulls columns have null values.


/* 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?
```

```sql
-- Type your code below:

WITH top_three_genres AS

(SELECT genre, COUNT(mv.id) AS movie_count
FROM movie AS mv
INNER JOIN genre AS gen
ON gen.movie_id = mv.id
INNER JOIN ratings AS rat
ON rat.movie_id = mv.id
WHERE avg_rating > 8
GROUP BY genre
ORDER BY count(mv.id) DESC
LIMIT 3)

SELECT na.name AS director_name, COUNT(dir.movie_id) AS movie_count
FROM director_mapping AS dir
INNER JOIN genre AS gen USING (movie_id)
INNER JOIN names AS na ON na.id = dir.name_id
INNER JOIN ratings AS ra USING (movie_id)
INNER JOIN top_three_genres USING (genre)
WHERE avg_rating > 8
GROUP BY na.name
ORDER BY movie_count DESC LIMIT 3;


/** OBSERVATIONS **/
-- James Mangold is the rank 1 director with 4 movie_count followed by Soubin Shahir and  Joe Russo.


/* 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?

-- Type your code below:

SELECT n.name AS actor_name, COUNT(rt.movie_id) AS movie_count
FROM names n
INNER JOIN role_mapping rm
ON n.id = rm.name_id
INNER JOIN ratings rt
ON rm.movie_id = rt.movie_id
WHERE median_rating >= 8
AND category = 'actor'
GROUP BY n.name
ORDER BY movie_count DESC LIMIT 2;

/** OBSERVATIONS **/
-- Mammootty have median_rating more than 8. Mohanlal's median rating is 5.

/* 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?

-- Type your code below:

WITH top_prod_companies AS
(
SELECT production_company,SUM(total_votes) AS vote_count,
DENSE_RANK() OVER(ORDER by SUM(total_votes) DESC) AS prod_comp_rank
```

```
DENSE_RANK() OVER(ORDER by SUM(total_votes) DESC) AS prod_comp_rank
FROM movie m
INNER JOIN ratings rt
ON m.id = rt.movie_id
GROUP BY production_company
)
SELECT * FROM top_prod_companies
WHERE prod_comp_rank <= 3;


/** OBSERVATIONS **/
-- Marvel Studios(2656967), Twentieth Century Fox(2411163), Warner Bros.(2396057).
-- 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.)


SELECT * FROM
(
SELECT name AS actor_name, SUM(total_votes) AS total_votes,
COUNT(m.id) AS movie_count, ROUND(SUM(avg_rating * total_votes) / SUM(total_votes), 2) AS actor_avg_rating,
RANK() OVER(ORDER by SUM(avg_rating * total_votes) / SUM(total_votes) DESC) AS actor_rank
FROM movie m
INNER JOIN
ratings rt
ON m.id = rt.movie_id
INNER JOIN
role_mapping rm
ON rt.movie_id = rm.movie_id
INNER JOIN
names n
ON rm.name_id = n.id
WHERE
category = 'actor' AND country LIKE '%India%'
GROUP BY name
HAVING COUNT(m.id) >= 5
)
A WHERE actor_rank = 1;


/** OBSERVATIONS **/
-- Top actor is Vijay Sethupathi with averagse rating of 8.42.



-- 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.)

-- Type your code below:

SELECT * FROM
(
SELECT name AS actress_name,
SUM(total_votes) AS total_votes, COUNT(m.id) AS movie_count,
ROUND((SUM(avg_rating * total_votes) / SUM(total_votes)), 2) AS actress_avg_rating,
RANK() OVER(ORDER by (SUM(avg_rating * total_votes) / SUM(total_votes)) DESC, SUM(total_votes) DESC) AS actress_rank
FROM movie m
INNER JOIN ratings rt
ON m.id = rt.movie_id
INNER JOIN role_mapping rm
ON m.id = rm.movie_id
```

```
ON m.id = rm.movie_id
INNER JOIN names n
ON rm.name_id = n.id
WHERE category = 'actress'
AND country LIKE '%India%'
AND languages LIKE '%Hindi%'
GROUP BY name
HAVING COUNT(m.id) >= 3 LIMIT 5
)
A WHERE actress_rank <= 5;
```

```
/** OBSERVATIONS **/
-- 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:

SELECT title,
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'
WHEN avg_rating < 5 THEN 'Flop movies'
END AS avg_rating_category
FROM movie m
INNER JOIN genre g
ON m.id = g.movie_id
INNER JOIN ratings rt
ON m.id = rt.movie_id
WHERE genre = 'Thriller';
```

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

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

-- 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 UNBOUNDED PRECEDING) AS moving_avg_duration
FROM movie m
INNER JOIN genre g
ON m.id = g.movie_id
GROUP BY genre
ORDER BY genre;
```

```
/** OBSERVATIONS **/
--   The average is constantly aboce 100 for any two consecutive genres.
-- 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.)

  -- Top 3 Genres based on most number of movies

  WITH top_genres AS
  (
  SELECT genre, COUNT(m.id) AS movie_count,
  RANK() OVER( ORDER BY COUNT(m.id) DESC) AS genre_rank
  FROM movie m
  INNER JOIN genre g
  ON g.movie_id = m.id
  INNER JOIN ratings r
  ON r.movie_id = m.id
  WHERE avg_rating > 8
  GROUP BY genre LIMIT 3
  ),
  movie_summary AS
  (
  SELECT genre, YEAR, title AS movie_name, CAST(REPLACE(REPLACE(IFNULL(worlwide_gross_income, 0), 'INR', ''), '$', '') AS DECIMAL(20))
  AS worlwide_gross_income,


-- Converting worldwide_gross_income datatype from 'varchar' to decimal

DENSE_RANK() OVER(PARTITION BY YEAR
ORDER BY CAST(REPLACE(REPLACE(IFNULL(worlwide_gross_income, 0), 'INR', ''), '$', '') AS DECIMAL(20)) DESC) AS movie_rank
FROM movie m
INNER JOIN genre g
ON m.id = g.movie_id
WHERE genre IN (SELECT genre FROM top_genres)
GROUP BY movie_name
)

SELECT * FROM movie_summary
WHERE movie_rank <= 5
ORDER BY YEAR;


-- 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?

-- Type your code below:

SELECT * FROM
(
SELECT production_company, COUNT(m.id) AS movie_count,
ROW_NUMBER() OVER(ORDER by COUNT(m.id) DESC) AS prod_comp_rank
FROM movie m
INNER JOIN ratings rt
ON m.id = rt.movie_id
WHERE median_rating >= 8
AND production_company IS NOT NULL
AND POSITION(',' IN languages) > 0
GROUP BY production_company
)
a WHERE prod_comp_rank <= 2;


/** OBSERVATIONS **/
-- Star Cinema and Twentieth Century Fox are top two production companies.

-- 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?

-- Type your code below:

SELECT * FROM
(
SELECT name, SUM(total_votes) AS total_votes, COUNT(rm.movie_id) AS movie_count,
AVG(avg_rating) AS avg_rating, ROW_NUMBER() OVER( ORDER by AVG(avg_rating) DESC) AS actress_rank
FROM names n INNER JOIN
role_mapping rm ON n.id = rm.name_id
INNER JOIN ratings rt
ON rm.movie_id = rt.movie_id
INNER JOIN genre g
ON rt.movie_id = g.movie_id
WHERE category = 'actress' AND avg_rating > 8 AND genre = 'Drama'
GROUP BY name
)
a WHERE actress_rank <= 3;


/** OBSERVATIONS **/
-- Sangeetha Bhat, Fatmire Sahiti, Adriana Matoshi are the top 3 actresses based on number of Super Hit movies (average rating >8) in drama genre.


/* 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

-- Type you code below:*/

WITH initial_table AS
(
SELECT dm.name_id AS director_id, n.name AS director_name, m.id, DATEDIFF(LEAD(date_published) OVER(PARTITION BY name ORDER by date_published),
date_published) + 1 AS 'avg_inter_movie_days',
r.avg_rating, r.total_votes, m.duration
FROM movie m
INNER JOIN ratings r
ON r.movie_id = m.id
INNER JOIN director_mapping dm
ON dm.movie_id = m.id
INNER JOIN names n
ON n.id = dm.name_id
)

--  Initial Table has the values for each director and their movies and the inter movie days without aggregate function applied

SELECT director_id,director_name, COUNT(id) AS number_of_movies,
ROUND(AVG(avg_inter_movie_days)) AS avg_inter_movie_days,
AVG(avg_rating) 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 initial_table
GROUP BY director_name
ORDER BY number_of_movies DESC LIMIT 9;


/** OBSERVATIONS **/
-- A.L. Vijay is the Rank 1 Director with maximum 5 number of movies.
```

Summary & Recommendations

Data speaks louder than words

All the below observations are captured with respect to movie data provided for years 2017, 2018 and 2019.

- The highest number of movies is produced in March is 824 movies.
- There are 13 unique genres in the dataset, from which Drama has highest number of movies produced i.e. 4285 and an average duration of 106.77 minutes. So RSVP Movies should focus on the 'Drama' genre.
- Dream Warrior Pictures and National Theatre Live both has produced highest rated films.
- The top actors with highest average median rating are Mammootty with more than 8 and Mohanlal with 5.
- Median Rating of 8+ will increase chances of superhit movie.
- In India, Taapsee Pannu can be chosen as the top actress with average rating 7.74 and top actor is Vijay Sethupathi with average rating 8.42.
- Sangeetha Bhat, Fatima Sahiti, Adriana Matoshi are the top 3 actresses based on the number of Super Hit movies in drama genre.
- Marvel Studios, Twentieth Century Fox and Warner Bros, are the top three production houses based on the number of votes received by the movies they have produced.
- Star Cinema and Twentieth Century are top two production companies that have produced highest number of hits among multilingual movies.
- Star Wars: Episode VIII - The Last Jedi, The Fate of the Furious Despicable Me 3 are the five highest-growing movies of each year that belong to the top three genres.

Prepared by: Rishabh Tiwari

Date: 23/08/2022