**IMDB DATASET ANALYSIS**

CONTEXT:

**OBJECTIVE:** To examine data from IMDB movie dataset and emphasize vital SQL concepts to perform data-driven operations using SQL query, supporting in decision-making, research and extract valuable information to generate meaningful insights to identify anomalies with deeper understanding of concepts.

**SCOPE:** SQL query analysis in the IMDB dataset is vast, enclosing details from basic data retrieval to complex data analysis performing queries to fetch movie details and advanced tasks like identifying patterns in ratings, performance of actors/ directors and generating detailed reports or insights for decision-making.

**AUDIENCE:** IMDB movie dataset serves diverse audience from casual viewers to professionals in movie industry like,

* Movie Enthusiasts
* Film-Making Professionals
* Scholars Researching Film History
* Tech Enthusiasts
* Marketing Professionals
* Critics and Journalists

ANALYSIS:

**DATA COLLECTION:** IMDB dataset contains movie data categorized into six tables based on the type of records with primary key and foreign key relating the records of different tables. Tables details like,

* Movie-Title,Year,Duration,Country,Income,Language,Productioncompany
* Genre- Genres associated with each movie.
* Director-Mapping- Maps movies to their Directors.
* Role Mapping: Maps actors/actresses to movies.
* Names- Information about people.
* Ratings: Contains ratings information for movies, including the average rating, total votes, and median rating.

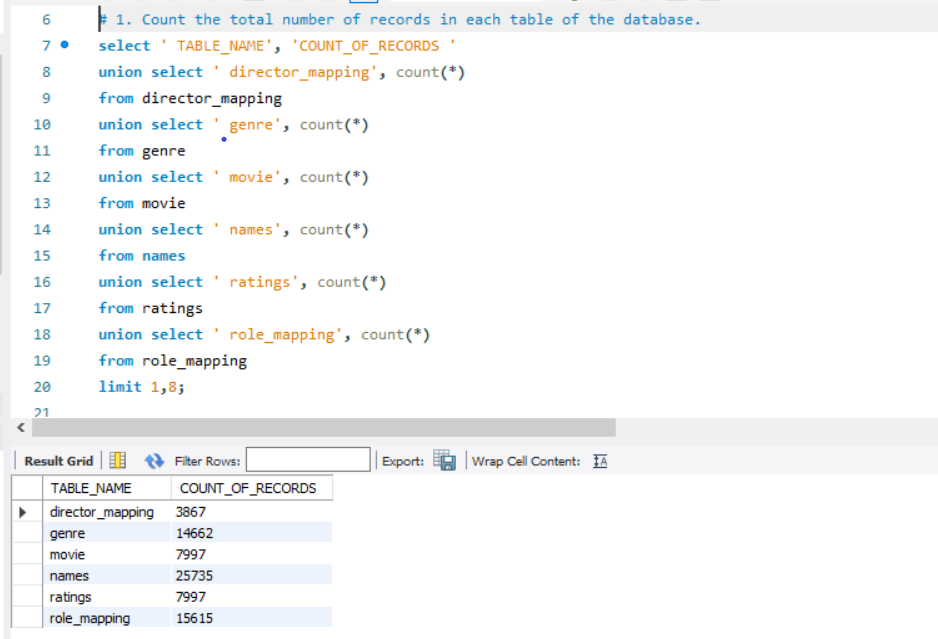
**QUERIES PERFORMED:**

1. **QUERY- Count the total number of records in each table of the dataset.**

In dataset IMDB, using the ‘COUNT()’ function to count the number of records in a table in dataset. The COUNT() function is often used with a ‘SELECT’ statement. Total number of records,

* Director-mapping-3867,
* genre-14662,
* movie-7997,
* names-25735,
* ratings-7997,
* role-mapping-15615.

**OUTPUT:**

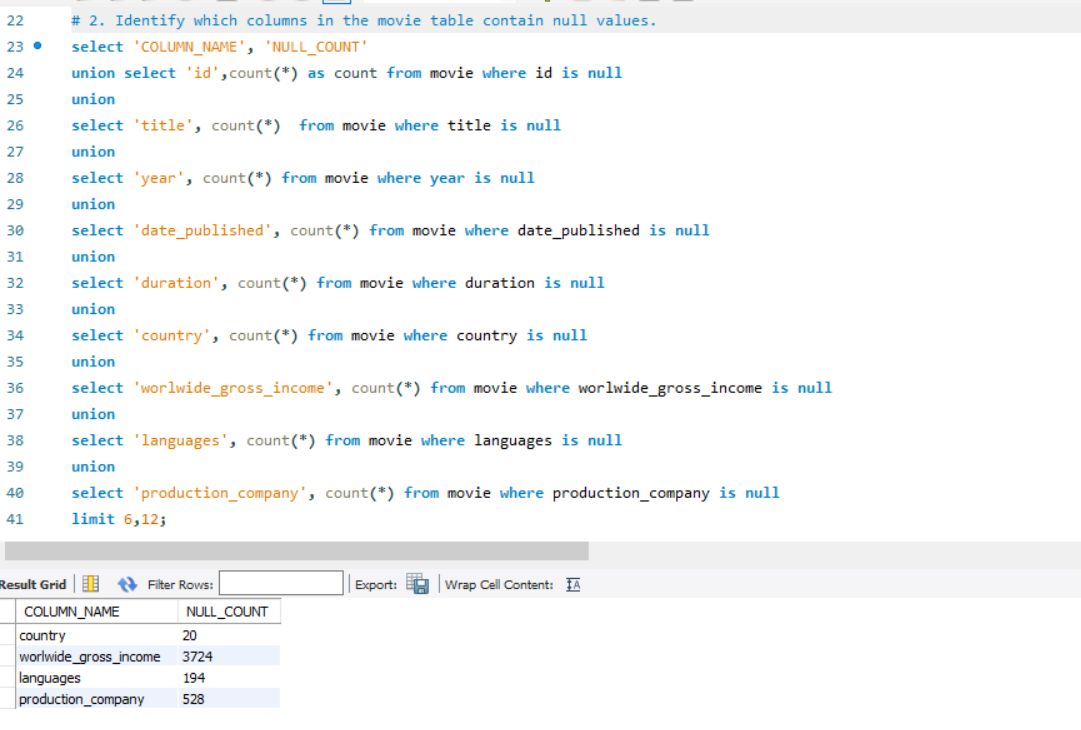
****

1. **QUERY- Identify which columns in the movie table contain null values.**

In IMDB dataset, to check for NULL values in a column is by using the ‘IS NULL’ condition in SQL. Columns containing null values are,

* Country,
* Worldwide-gross-income,
* Languages,
* Production-company.

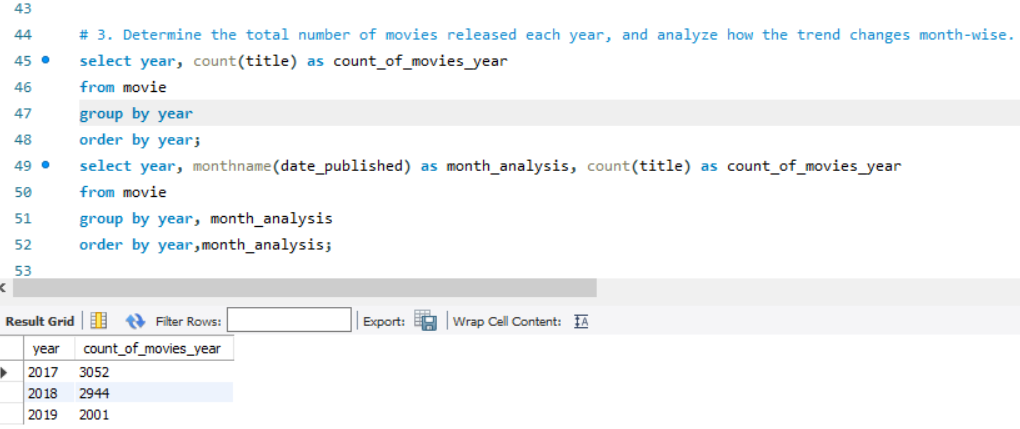
**OUTPUT:**

****

1. **QUERY- Determine the total number of movies released each year, and analyse how the trend changes month-wise.**

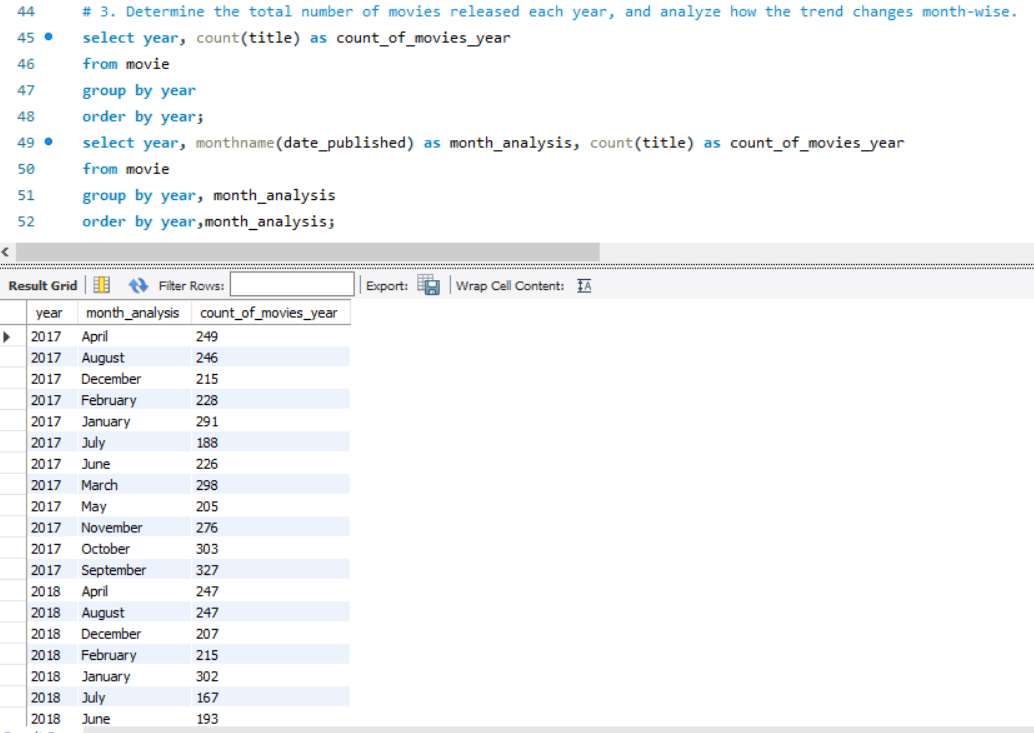
In IMDB dataset, to find the total number of movies released each year, group the data by extracting year from the year column and then count the number of movies for each year. Year 2019 count significantly reduced, 2017 ranks highest among 2018 and 2019.

**OUTPUT:**

****

The movie release trend changes month-wise within each year, you need to group the data by the month extracting from date-published column in movie table. In

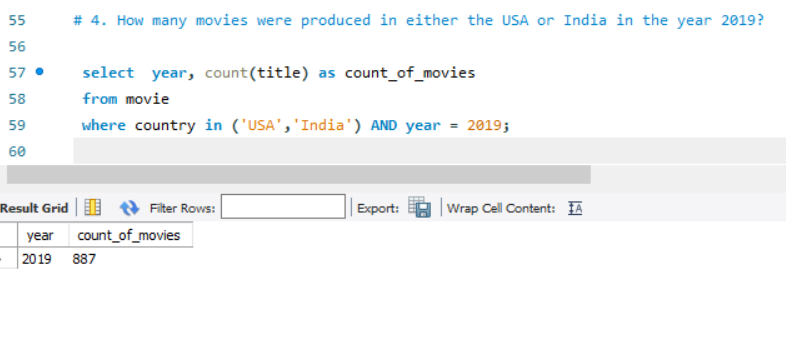
* Year 2017, September- high count 327 and July-low count of 188,
* Year 2018, January- high count 302 and July-low count of 167 movies,
* Year 2019, September- high count 327,July-low count of 188 movies.

****

1. **QUERY- How many movies were produced in either the USA or India in the year 2019?**

To calculate the number of movies produced in either the USA or India in the year 2019 IMDB dataset contains information about movies, such as country and release year.

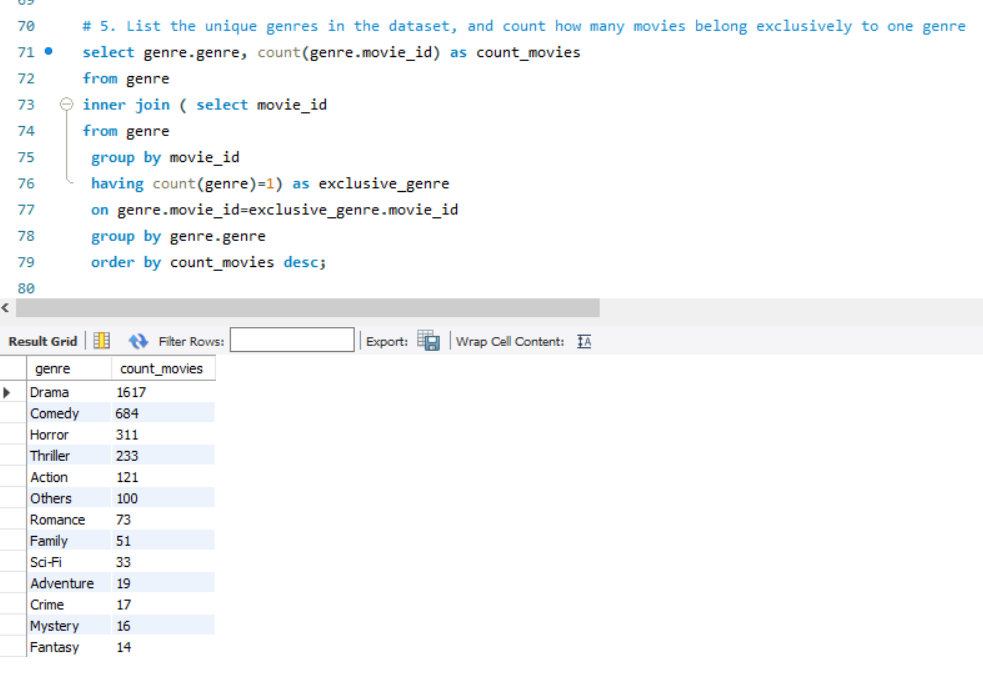
**OUTPUT:**

****

1. **QUERY- List the unique genres in the dataset, and count how many movies belong exclusively to one genre.**

‘SELECT DISTINCT’ will return all unique genre entries from the genre table. Each genre will appear only once in the result, list of unique genres and the count of movies that belong exclusively to one genre found with movie-id count belonging to one genre.

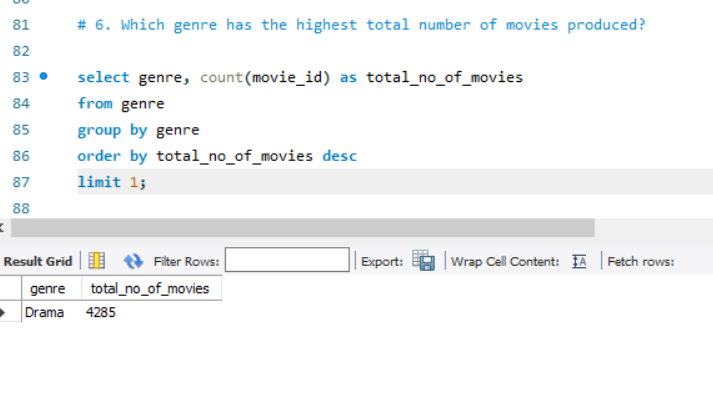
**OUTPUT:**

****

1. **QUERY- Which genre has the highest total number of movies produced?**

In IMDB dataset, select genre and count() function counts movies belong to each genre, ‘Group by’ the result set by the genre field for each unique genre. Order by highest and limit first row indicates ‘Drama genre’ with 4285 movies.

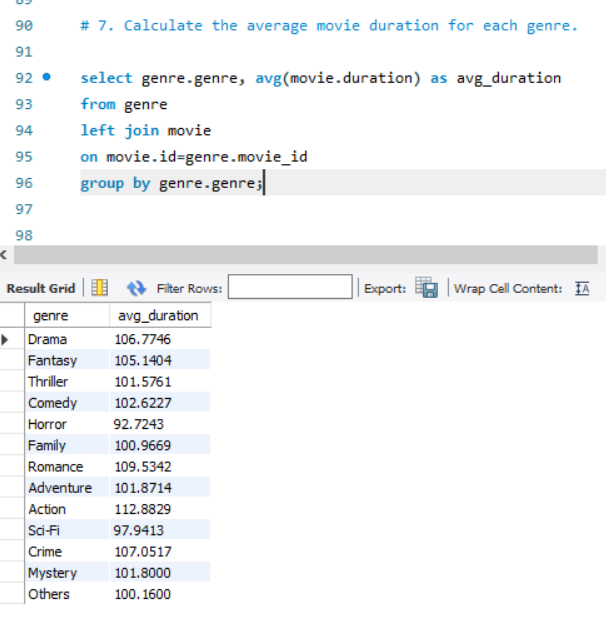
**OUTPUT:**



1. **QUERY- Calculate the average movie duration for each genre.**

In IMDB dataset, calculate the average duration for each genre, and calculate the average duration of movies in each genre, group by genre using joins combining movies and genre tables.

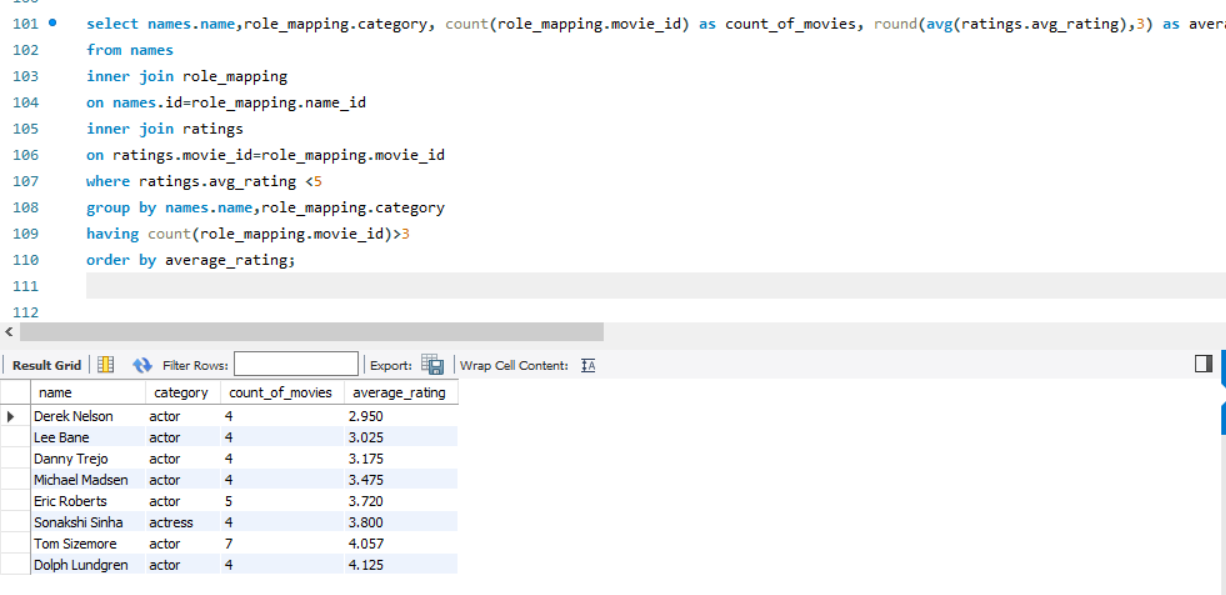
**OUTPUT:**

****

1. **QUERY- Identify actors or actresses who have appeared in more than three movies with an average rating below 5.**

In IMDB dataset, Join the names, role-mapping, ratings tables to connect movies, category, and their ratings. Filter movies with a rating below 5. Group the results by actor and names. Ensure the actor has appeared in more than three such movies.

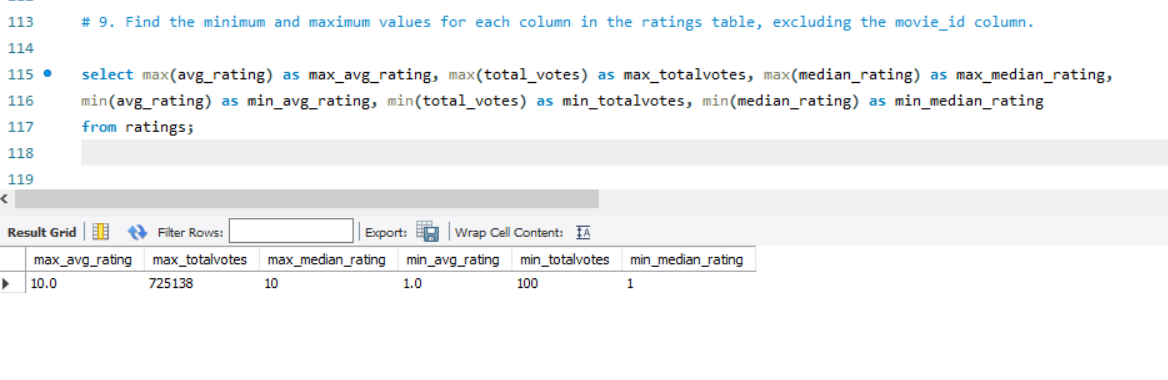
**OUTPUT:**

****

1. **QUERY- Find the minimum and maximum values for each column in the ratings table, excluding the movie-id column.**

MIN() computes the minimum value in the rating table, represents the lowest rating in the ratings table. MAX() computes the maximum value in the rating table, represents the highest rating in the ratings table.

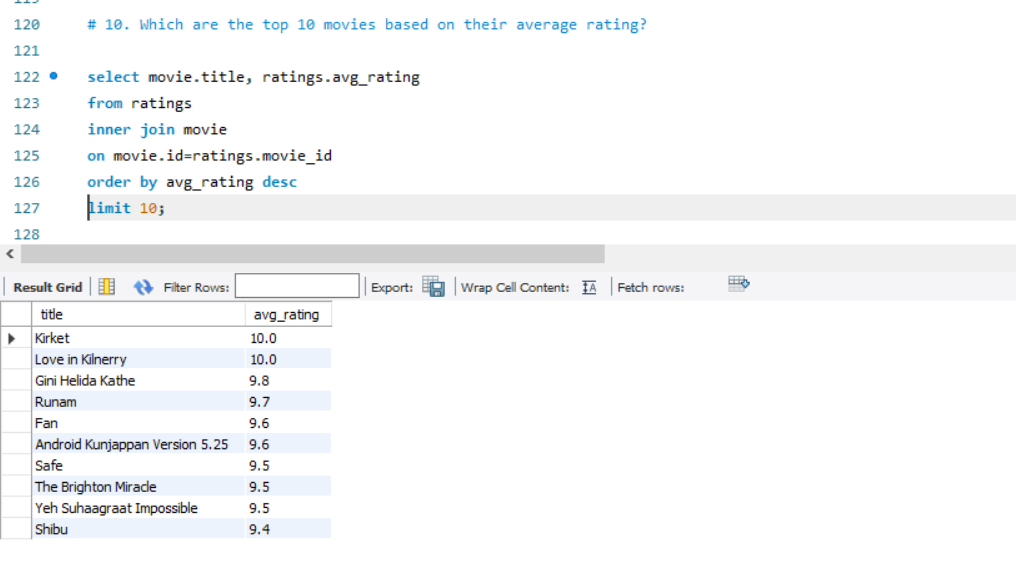
**OUTPUT:**

****

1. **QUERY- Which are the top 10 movies based on their average rating?**

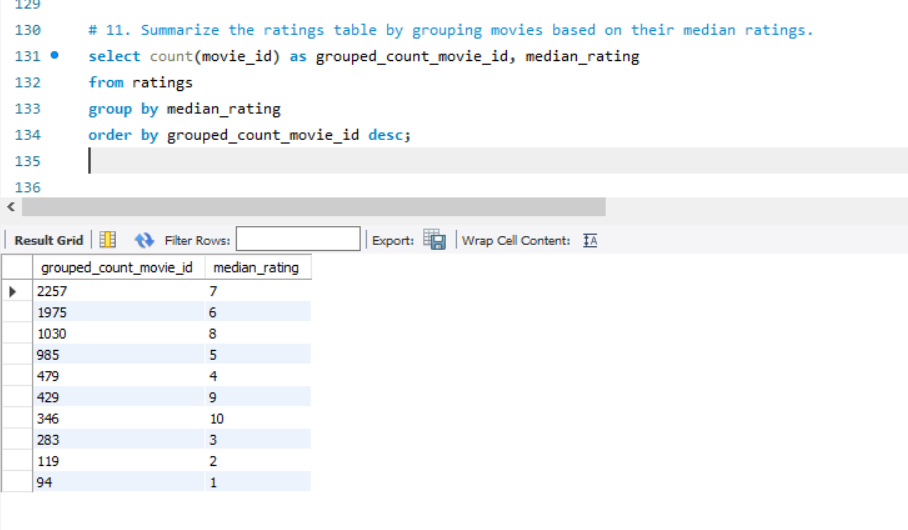
In IMDB dataset, calculate the average rating for each movie and sort the result in descending order of the average rating and Limit the result to the top 10 movies.

**OUTPUT:**

****

1. **QUERY- Summarize the ratings table by grouping movies based on their median ratings.**

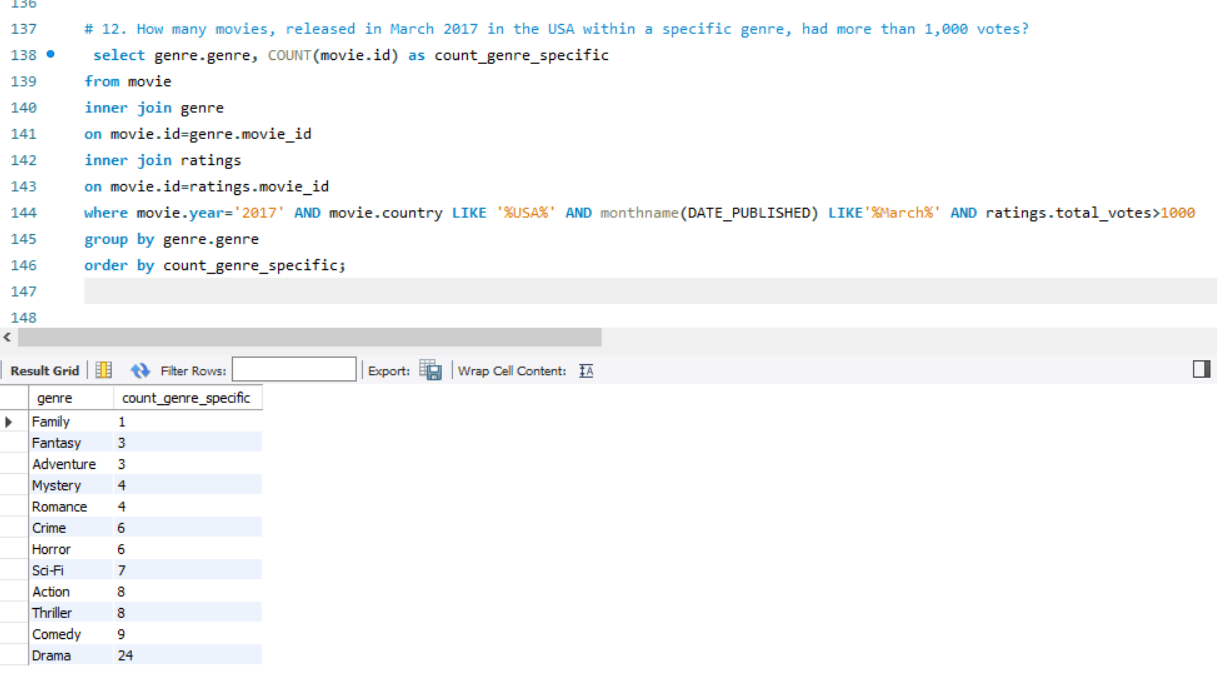
In IMDB dataset, To summarize the ratings table by grouping movies based on their median ratings, calculate the median for each movie’s ratings by grouping movie-id.

**OUTPUT:**

1. **QUERY- How many movies, released in March 2017 in the USA within a specific genre, had more than 1,000 votes?**

To find how many movies were released in March 2017 in the USA within a specific genre and having more than 1,000 votes, filters based on, release date being in March 2017, country being the USA, genre of the movie and movie with more than 1,000 votes.

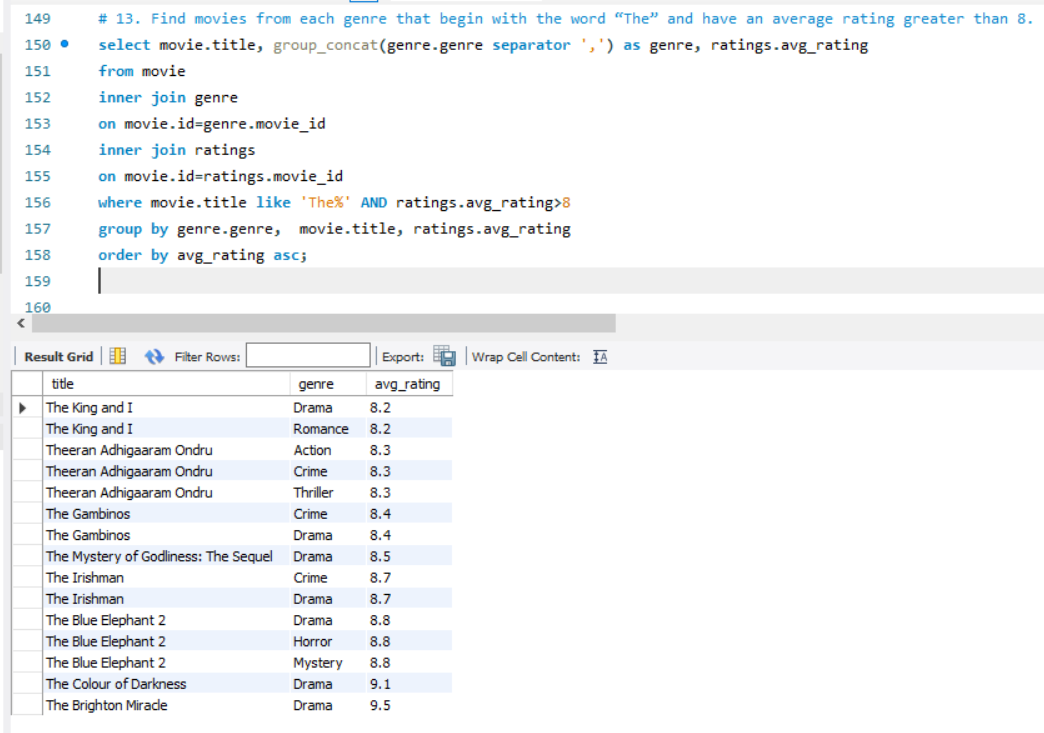
**OUTPUT:**

****

1. **QUERY- Find movies from each genre that begin with the word “The” and have an average rating greater than 8.**

In IMDB dataset, Filter movies whose titles begin with the word "The" using ‘LIKE’ operator. Calculate the average rating for each movie. Group the results by genre and filter movies with an average rating greater than 8.

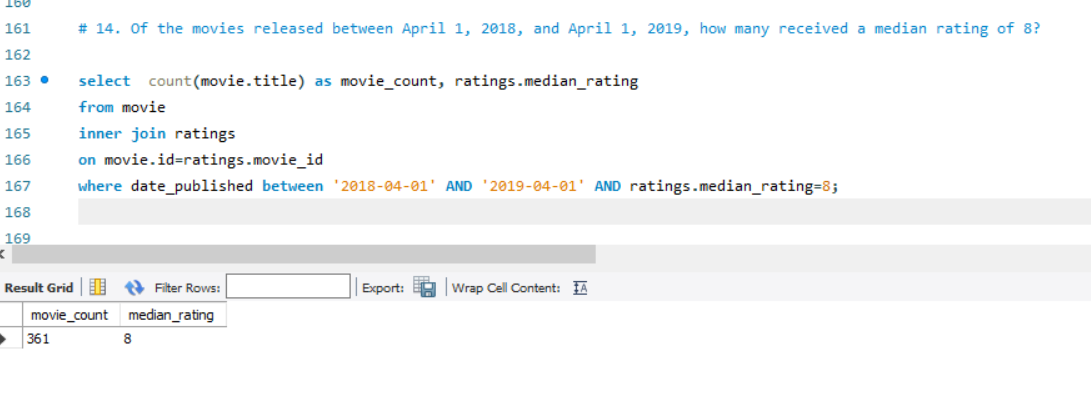
**OUTPUT:**

****

1. **QUERY- Of the movies released between April 1, 2018, and April 1, 2019, how many received a median rating of 8?**

In IMDB dataset, to find the number of movies released between April 1, 2018, and April 1, 2019, that received a median rating of 8, Filter movies released within April 1, 2018, and April 1, 2019. Filter movies that have a median rating of 8. Count the number of movies.

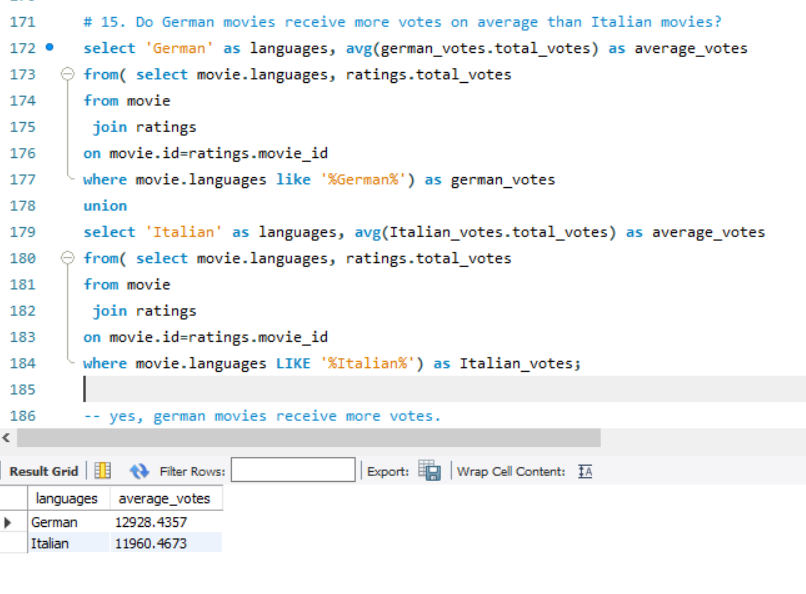
**OUTPUT:**

****

1. **QUERY- Do German movies receive more votes on average than Italian movies?**

In IMDB dataset, to determine if German movies receive more votes on average than Italian movies, Filter movies by country ‘Germany and Italy’ and Group the movies by country. Calculate the average number of votes for each country movies and Compare the average number of votes between the two countries.

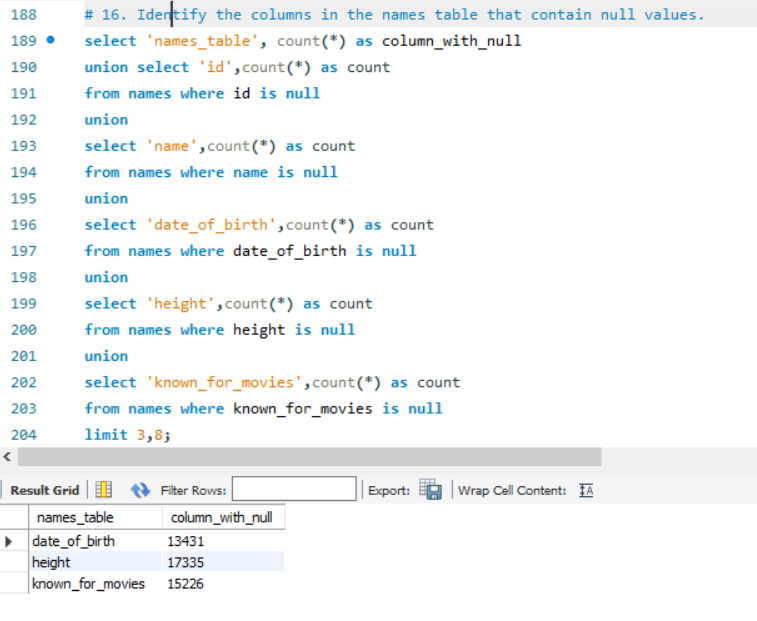
**OUTPUT:**

****

1. **QUERY- Identify the columns in the names table that contain null values.**

In IMDB dataset, To identify the columns in the names table that contain null values, use the COUNT() function along with a UNIQUE statement to count null values in each column.

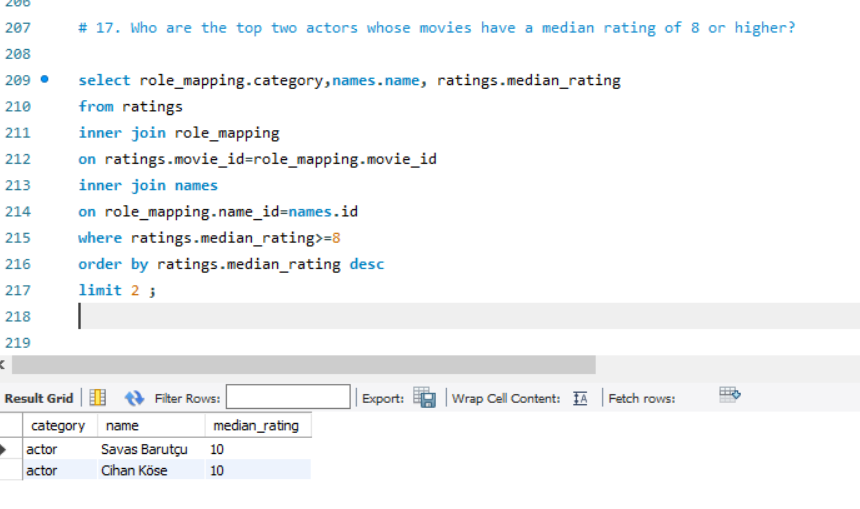
**OUTPUT:**

****

1. **QUERY- Who are the top two actors whose movies have a median rating of 8 or higher?**

In IMDB dataset, to identify the top two actors whose movies have a median rating of 8 or higher, Identify movies with a median rating of 8 or higher, Determine the actors by combining ratings and names, role-mapping tables. Rank the actors based on the number of movies, limit the top two actors.

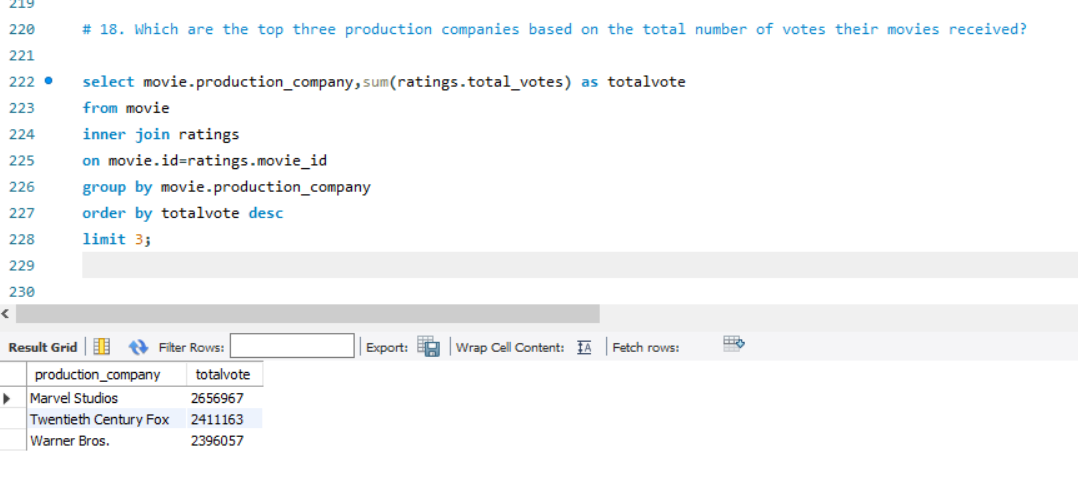
**OUTPUT:**

****

1. **QUERY- Which are the top three production companies based on the total number of votes their movies received?**

In IMDB DATASET, to find the top three production companies based on the total number of votes their movies received, Identify the movies and their associated production companies, Calculate the total number of votes each movie received, Group by production company, Order the results by the total votes in descending order. Select the top three production companies.

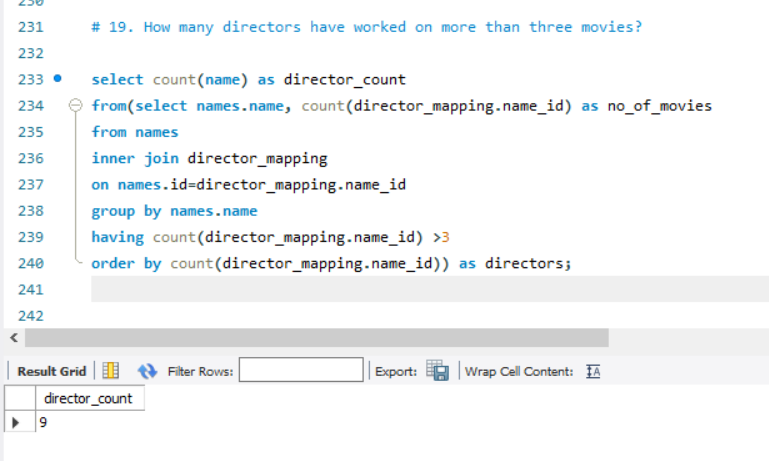
**OUTPUT:**

****

1. **QUERY- How many directors have worked on more than three movies?**

In IMDB dataset, to find out how many directors have worked on more than three movies, create relations between movies and directors using joins. count the number of movies each director has worked on. Filter directors worked on more than three movies. Count the number of directors.

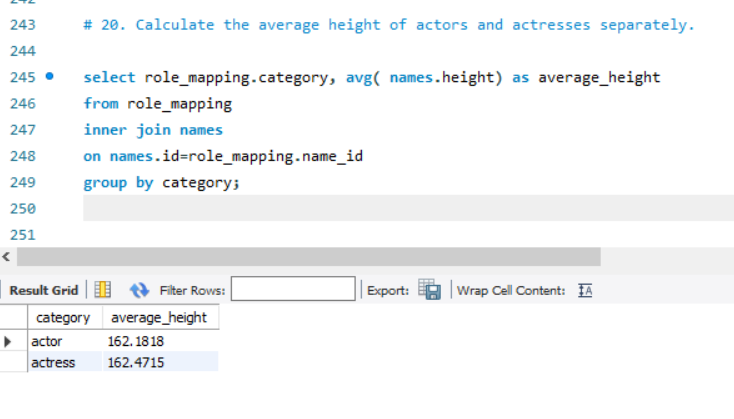
**OUTPUT:**

****

1. **QUERY- Calculate the average height of actors and actresses separately.**

In IMDB dataset, To calculate the average height of actors and actresses separately, Identify the category, Group the actors by category ‘actors and actresses’, Calculate the average height of actors and actresses.

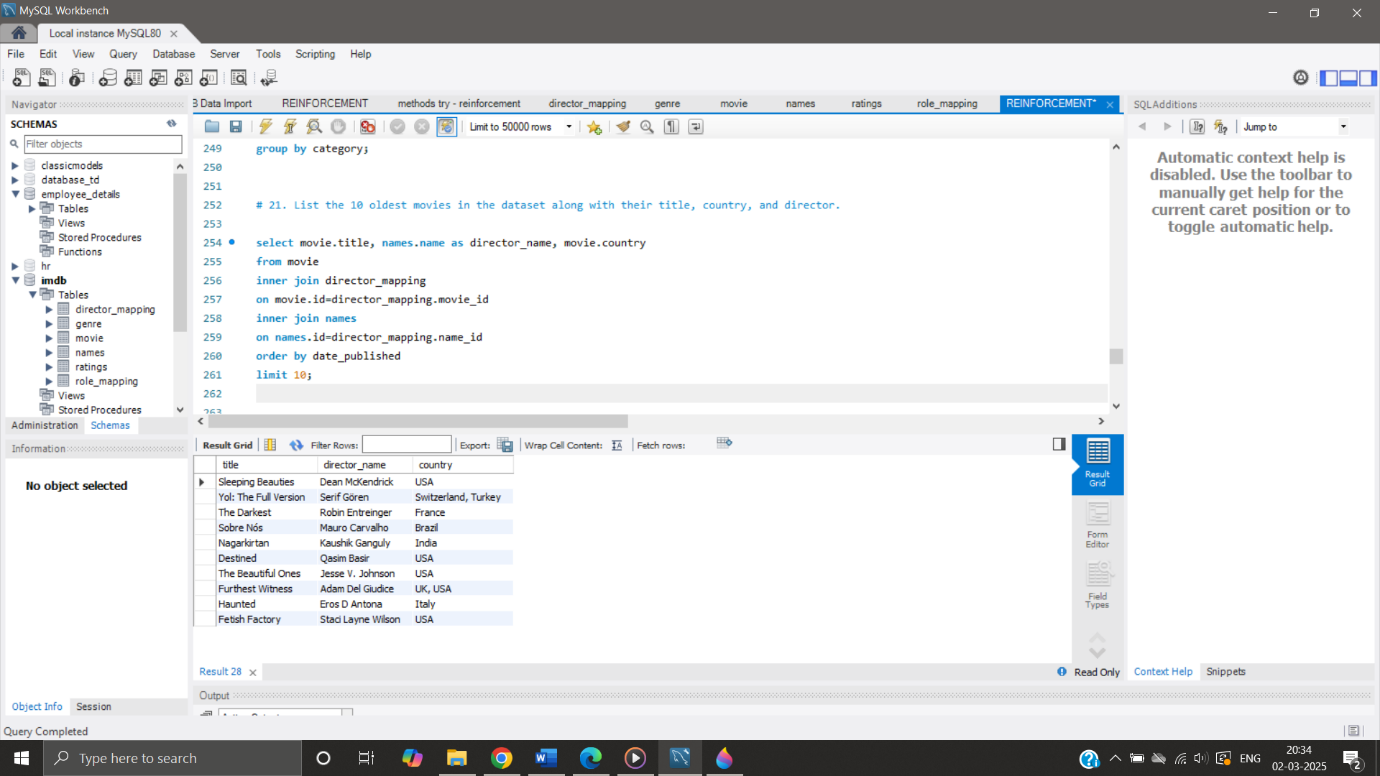
**OUTPUT:**

****

1. **QUERY- List the 10 oldest movies in the dataset along with their title, country, and director.**

In dataset, to list the 10 oldest movies in the dataset along with their title, country, and director, Sort the movies by their release year in ascending order to get the earliest, Select columns: title, country, and director, Limit the results to the 10 oldest movies.

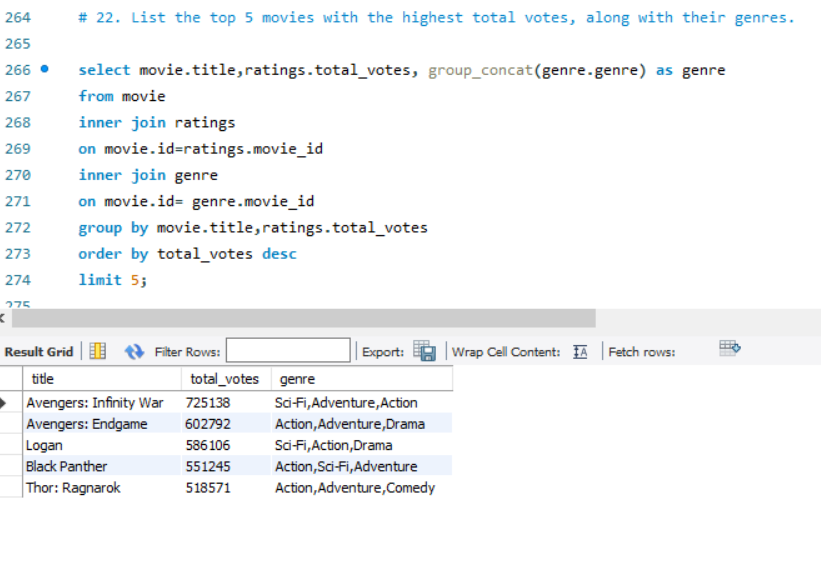
**OUTPUT:**

****

1. **QUERY- List the top 5 movies with the highest total votes, along with their genres.**

In dataset, to list the top five movies with the highest total votes along with their genres, Sum the total votes for each movie, using Joins combine the movies with their genres. Sort by the total votes in descending order. Limit the result to the top 5 movies.

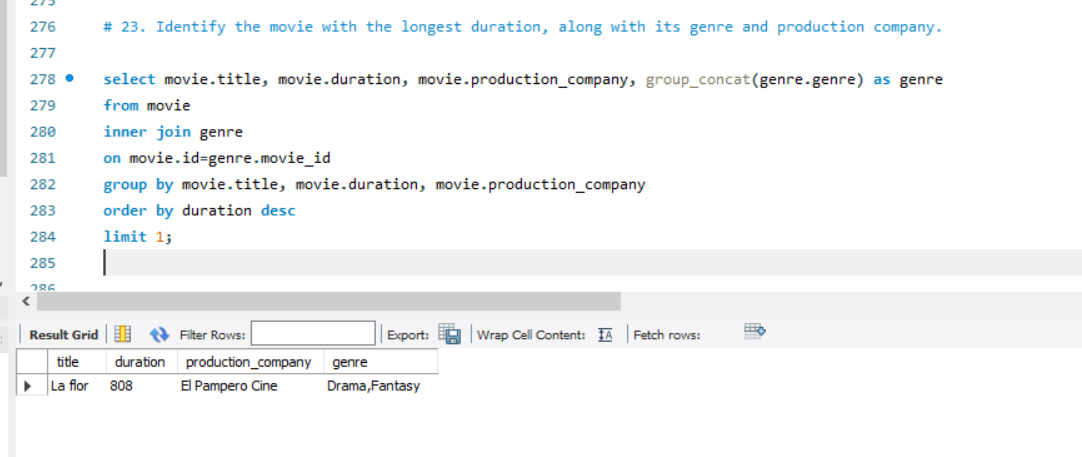
**OUTPUT:**

****

1. **QUERY- Identify the movie with the longest duration, along with its genre and production company.**

In IMDB dataset, To identify the movie with the longest duration along with its genre and production company Identify the movie with the maximum duration. Join the movies with their genre and production company. Display information: movie title, genre, production company, and duration.

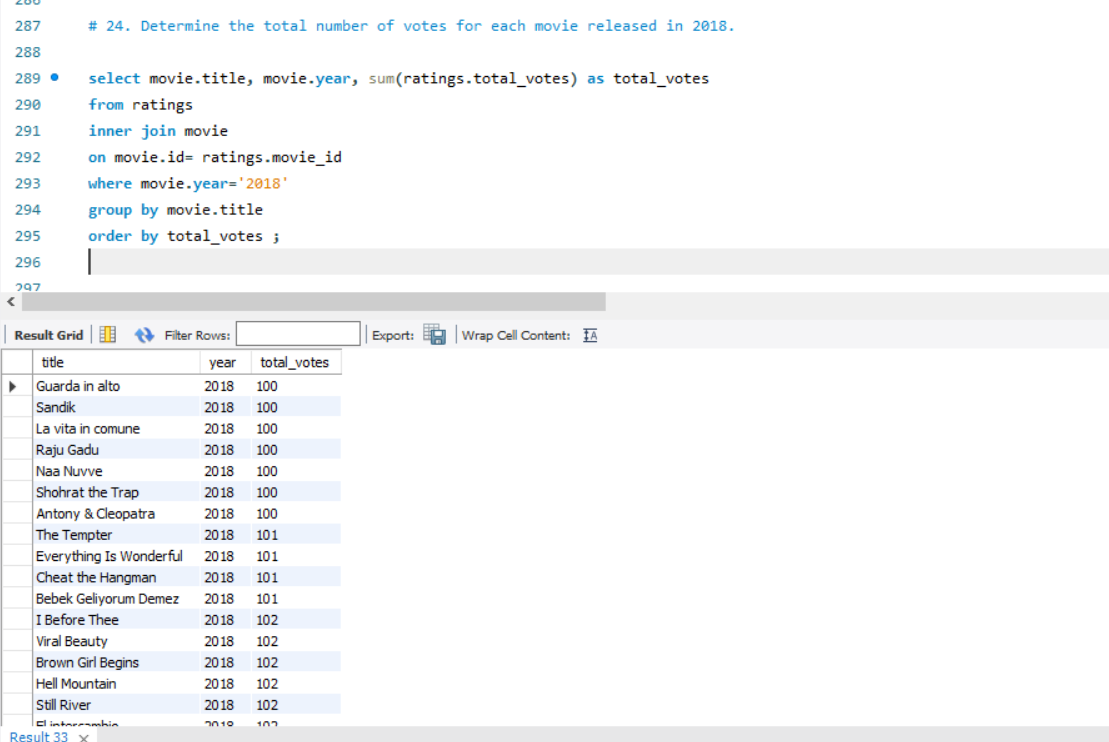
**OUTPUT:**

****

1. **QUERY- Determine the total number of votes for each movie released in 2018.**

In IMDB dataset, To determine the total number of votes for each movie released in 2018, Filter movies released in 2018, Sum the number of votes for each movie, Group the results by movie to get the total votes.

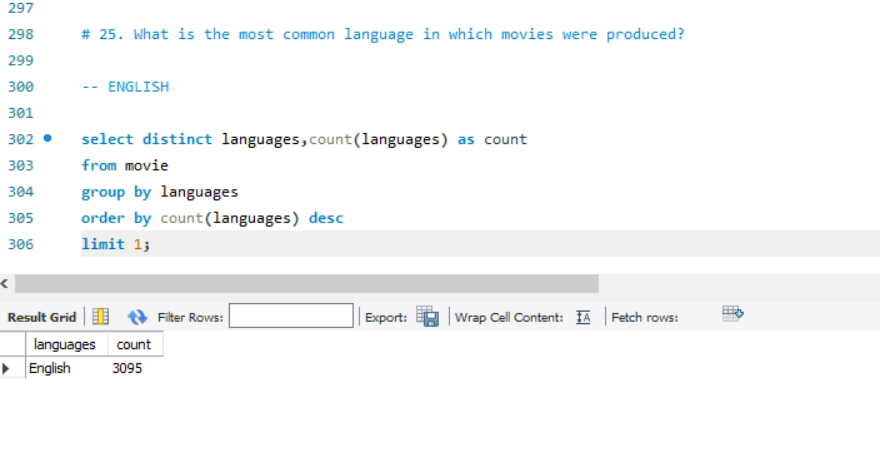
**OUTPUT:**

****

1. **QUERY- What is the most common language in which movies were produced?**

In IMDB dataset, determine the most common language in which movies were produced, Identify the language for each movie in the dataset, Count the occurrences of each language, Sort the results by the count of occurrences. Select the language with the highest count.

**OUTPUT:**

****

**CONCLUSION:**

The IMDB movie dataset analysis provides valuable insights on year-month analysis, viewer’s interests based on ratings, genre most preferred and movie performance along with preferred actor-actress and language, which can be used to inform future business strategies and decision-making by leveraging these findings, company improves customer satisfaction, increase enthusiasts/ viewers and optimize operations.

**FUTURE RECOMMENDATIONS:**

* Target Marketing Campaigns to engage viewer’s
* Optimization of new Ideas and Genre
* Website Enhancement to enhance Interaction
* Customer Retention Strategies.

THANK YOU

REPORTED BY,

VEENA SRI PM,

DATA ANALYTICS AND DATA SCIENCE,

RP-33.