**Advanced SQL – Reinforcement Project**

**OBJECTIVE**

The objective of these SQL queries is to efficiently retrieve, aggregate, and analyze data from relational databases to derive meaningful insights. By using various SQL operations such as filtering, grouping, joining, and aggregation, the goal is to identify trends, summarize key metrics, and support data-driven decision-making. These queries enable the extraction of relevant information, comparison across different categories, and identification of patterns to inform business strategies and performance evaluations. Ultimately, SQL empowers analysts to work with large datasets effectively and deliver actionable insights for improved decision-making.

**INTRODUCTION**

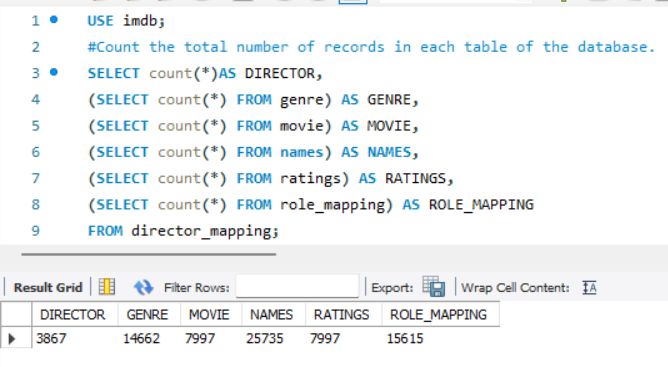
SQL queries play a crucial role in data analysis by enabling efficient retrieval, manipulation, and aggregation of data from relational databases. These queries allow analysts to filter, group, and summarize large datasets, uncovering trends and insights essential for decision-making. By utilizing functions like joins, aggregations, and subqueries, SQL empowers users to analyze complex data structures and support business strategies with accurate, actionable results. Ultimately, SQL queries are indispensable for transforming raw data into meaningful insights for better operational and strategic planning.

**Queries**

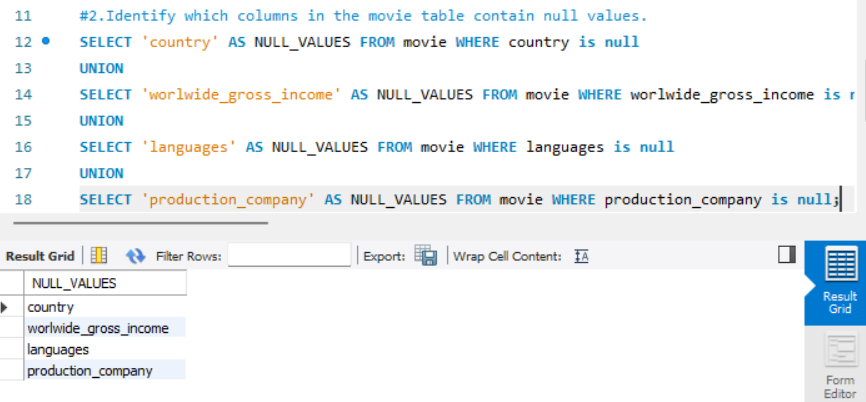
**1.Count the total number of records in each table of the database.**

It retrieves the number of entries in director\_mapping, genre, movie, names, ratings, and role\_mapping tables. The result shows that the names table has the highest count (25,735 records), while the director table has the lowest (3,867 records), giving an overview of the

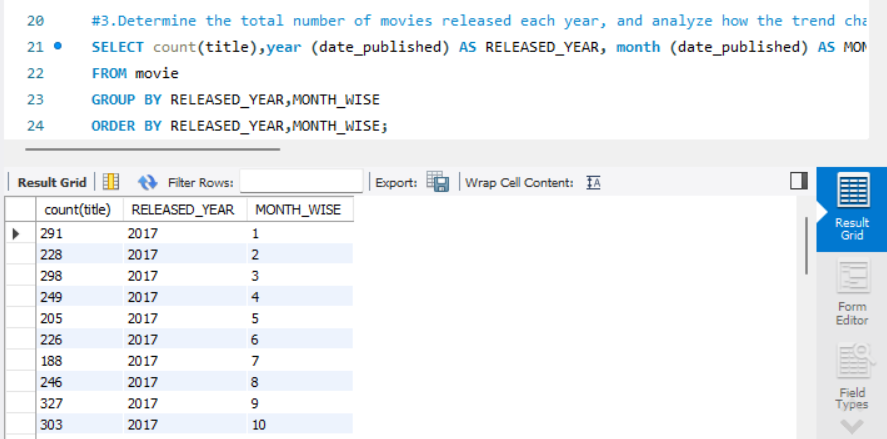
dataset distribution.



**2.Identify which columns in the movie table contain null values.**

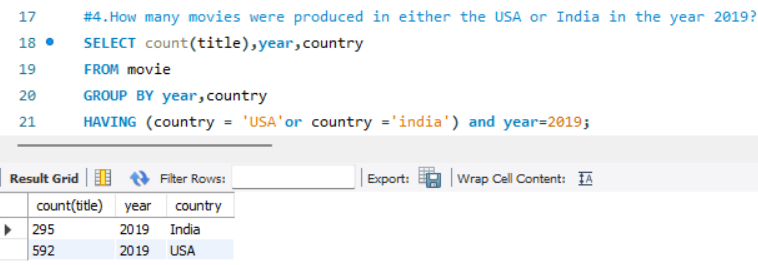
It checks for missing values in country, worldwide\_gross\_income, languages, and production\_company using the UNION operator. The result confirms that all these columns contain NULL values.

**3.Determine the total number of movies released each year, and analyze how the trend changes month-wise.**

The result shows monthly movie release counts for the year 2017, with March (298 movies) and September (327 movies) having the highest releases. The data is grouped by year and month and sorted accordingly. 

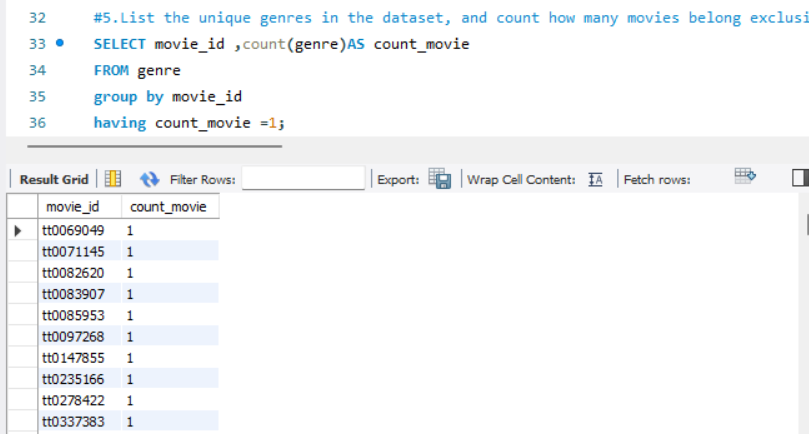
**4.How many movies were produced in either the USA or India in the year 2019?**

The result shows that 295 movies were produced in India, while 592 movies were produced in the USA. The query groups data by year and country, filtering only 2019 and the specified countries.



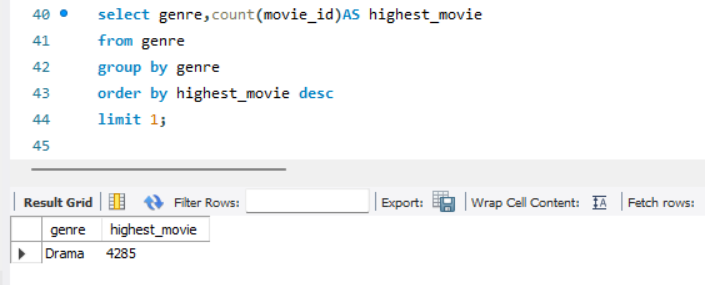
**5.List the unique genres in the dataset, and count how many movies belong exclusively to one genre.**

It counts the number of genres associated with each movie\_id, groups them, and filters results where the count is exactly one. The result displays a list of movies with a single unique genre.



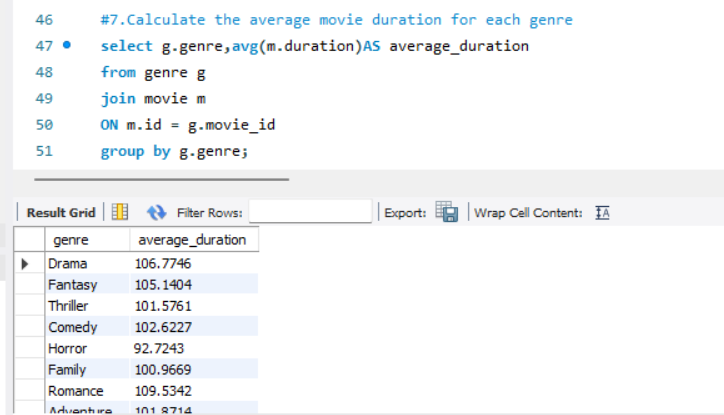
**6. Which genre has the highest total number of movies produced?**

This SQL query identifies the genre with the highest number of movies in the dataset. It uses the COUNT(movie\_id) function to count the movies in each genre, then groups the results by genre. The query orders the genres by the movie count in descending order and limits the result to only the genre with the highest number of movies, which in this case is "Drama" with 4,285 movies.



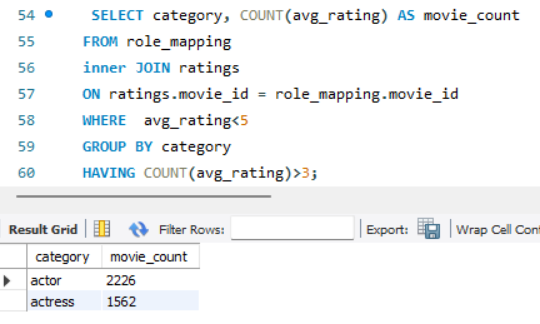
**7. Calculate the average movie duration for each genre.**

It uses the AVG(m.duration) function to find the average duration of movies in each genre, joining the genre and movie tables based on the movie's ID. The results are grouped by genre, providing the average movie duration for each genre in the dataset.



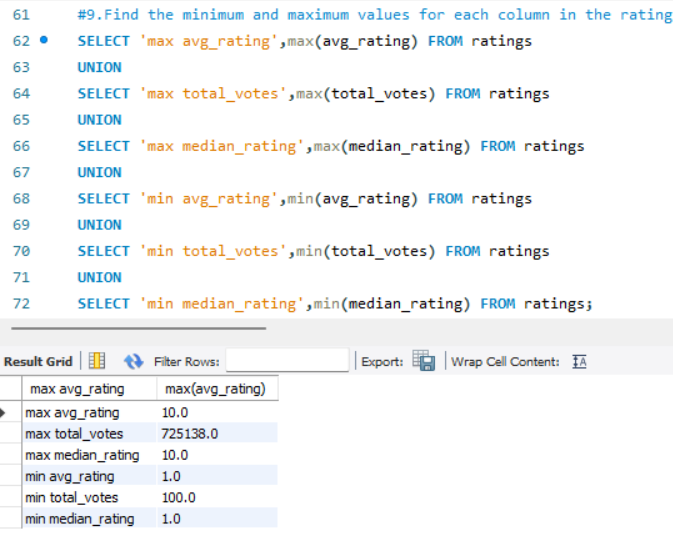
**8. Identify actors or actresses who have appeared in more than three movies with an average rating below 5.**

.The query joins the role\_mapping and ratings tables, filters the results by movies with an average rating less than 5, and groups the results by category. It uses HAVING to filter out categories with fewer than three such movies, displaying the count of movies per category.



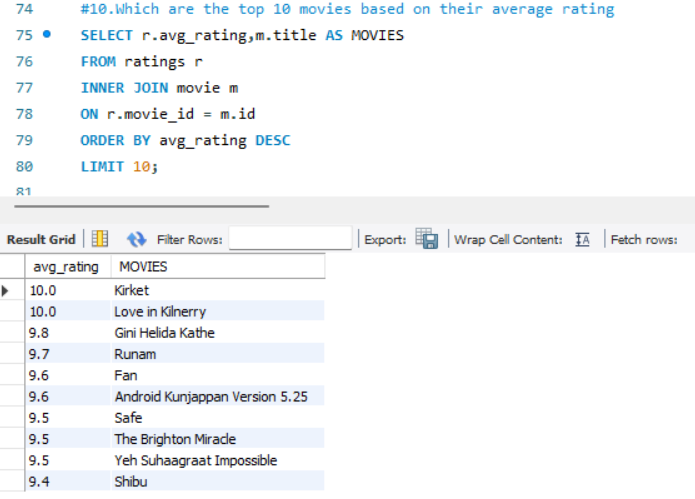
**9. Find the minimum and maximum values for each column in the ratings table, excluding the movie\_id column**.

Ratings table: avg\_rating, total\_votes, and median\_rating. It uses the MAX and MIN functions to find the highest and lowest values for each of these columns, combining results using UNION. The output provides insight into the range of ratings and votes in the dataset.



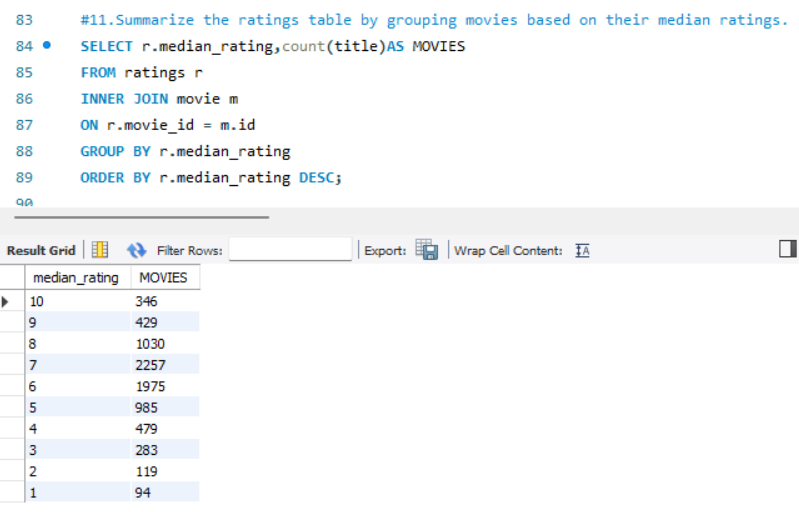
**10. Which are the top 10 movies based on their average rating?**

It joins the ratings and movie tables using the movie\_id column, sorts the results in descending order of avg\_rating, and limits the output to 10 records. The result grid displays movies like *Kirket* and *Love in Kilnerry* with perfect 10.0 ratings, followed by other highly-rated films.



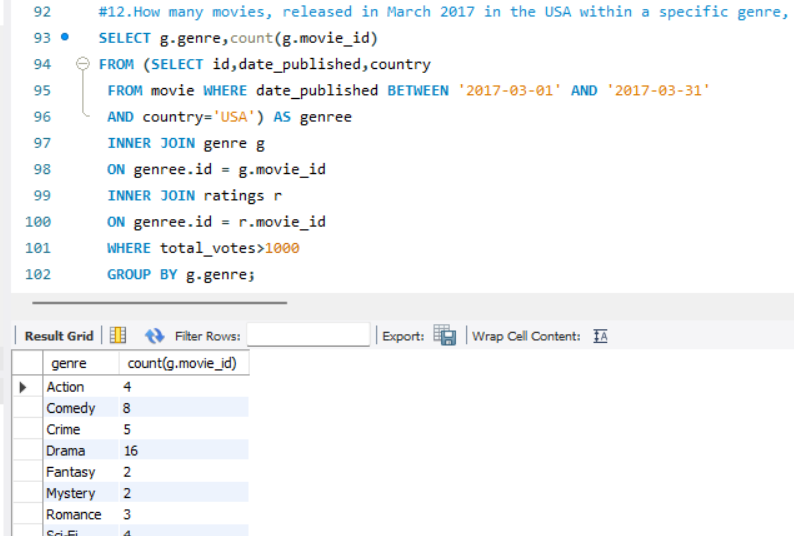
**11. Summarize the ratings table by grouping movies based on their median ratings.**

It joins the movie table, counts the number of movies for each median rating, and orders the results in descending order. The result grid shows that the highest number of movies have a median rating of 7 (2,257 movies), followed by ratings of 6 and 8. This analysis helps in understanding the distribution of movie ratings.



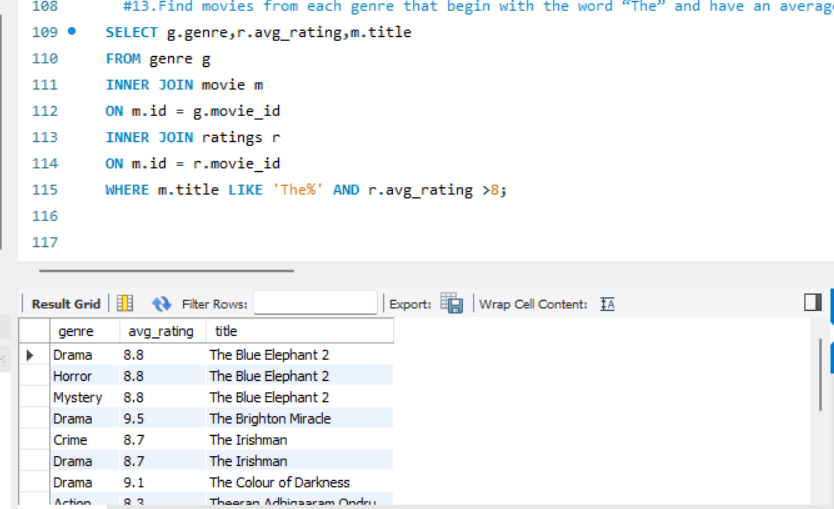
**12. How many movies, released in March 2017 in the USA within a specific genre, had more than 1,000 votes?**

It joins the movie, genre, and ratings tables, groups results by genre, and orders them accordingly. The result grid shows that *Drama* had the highest count (16 movies), followed by *Comedy* (8) and *Crime* (5). This analysis helps in understanding genre popularity during that period.



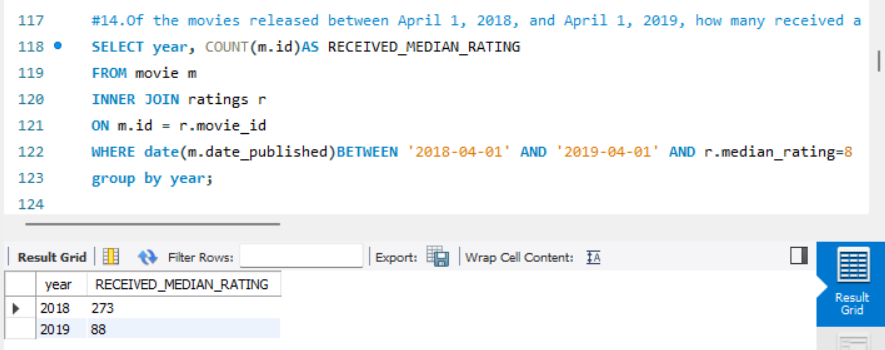
**13. Find movies from each genre that begin with the word “The” and have an average rating greater than 8.**

It joins the genre, movie, and ratings tables to filter relevant movies. The result grid includes highly-rated films like *The Blue Elephant 2* (8.8), *The Brighton Miracle* (9.5), and *The Irishman* (8.7). This analysis helps in identifying top-rated movies with a common title pattern across genres.



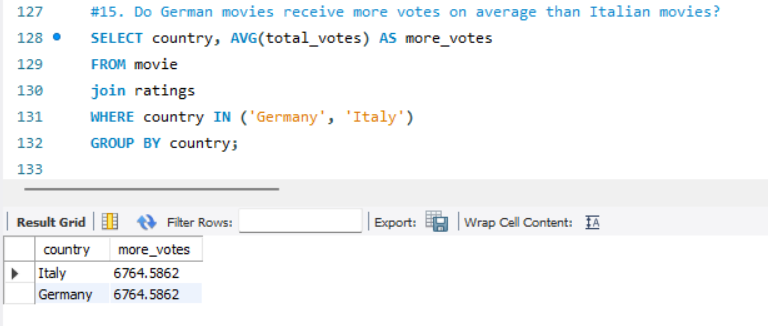
**14. Of the movies released between April 1, 2018, and April 1, 2019, how many received a median rating of 8?**

It joins the movie and ratings tables and filters results based on the specified date range and rating criteria. The result grid shows that 273 movies met this condition in 2018, while 88 movies did in 2019. This analysis helps track highly-rated movie trends over the given period.



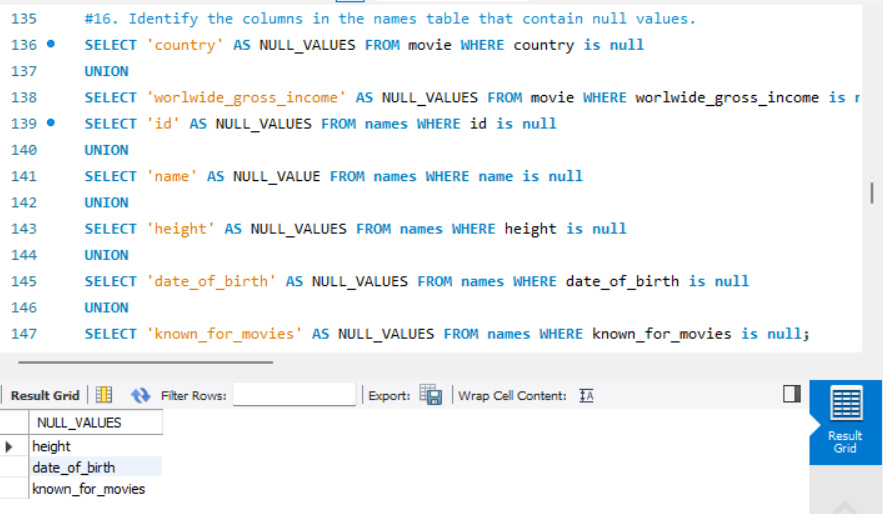
**15. Do German movies receive more votes on average than Italian movies?**

It calculates the average number of votes (total\_votes) for movies from Germany and Italy by grouping results based on the country. The output shows that both countries have the same average votes, 6764.5862, indicating no difference in voting trends between them.



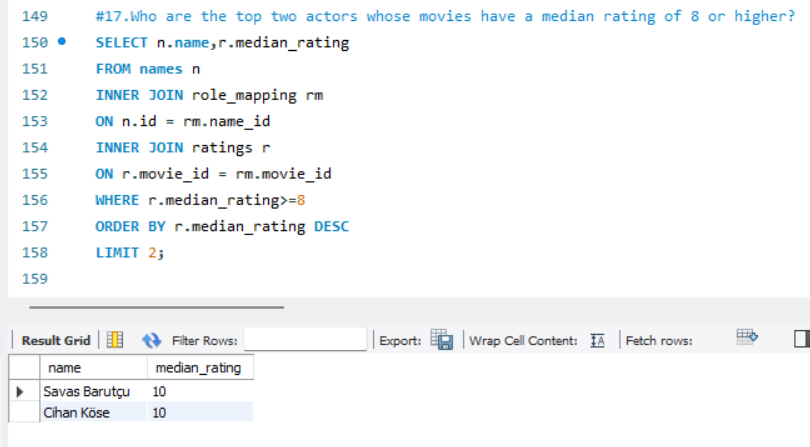
**16. Identify the columns in the names table that contain null values.**

It multiple columns, including height, date\_of\_birth, and known\_for\_movies, using UNION statements to list only those with missing data. The result shows that these three columns have NULL values.

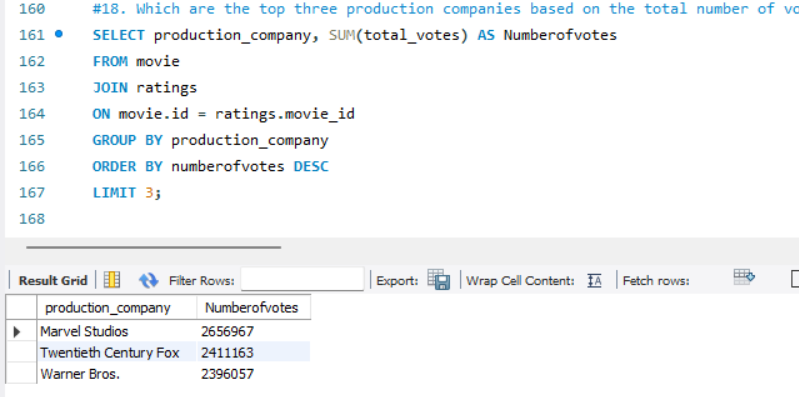


**17. Who are the top two actors whose movies have a median rating of 8 or higher?**

By joining the names, role\_mapping, and ratings tables, it filters movies with a median rating above 8 and orders them in descending order. The top two actors found are Savas Barutçu and Cihan Köse, both having a perfect median rating of 10.

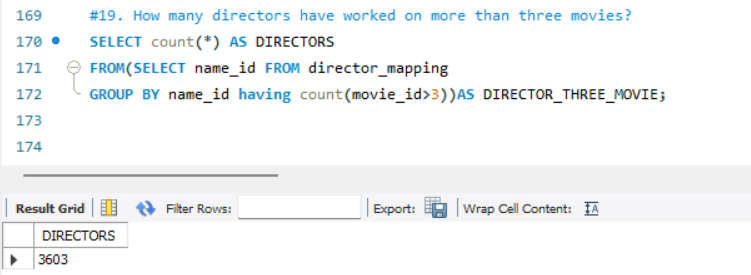


**18. Which are the top three production companies based on the total number of votes their movies received?**



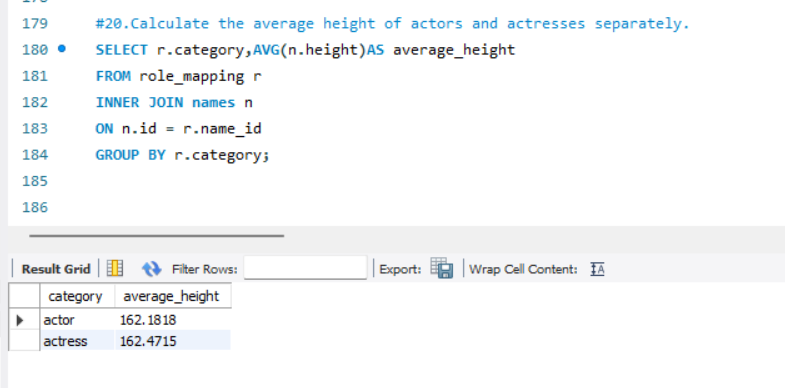
**19. How many directors have worked on more than three movies?**

It joins the movie and ratings tables, groups by production\_company, and sums up the total\_votes. The results show that Marvel Studios leads with 2,656,967 votes, followed by Twentieth Century Fox with 2,411,163 votes, and Warner Bros. with 2,396,057 votes.

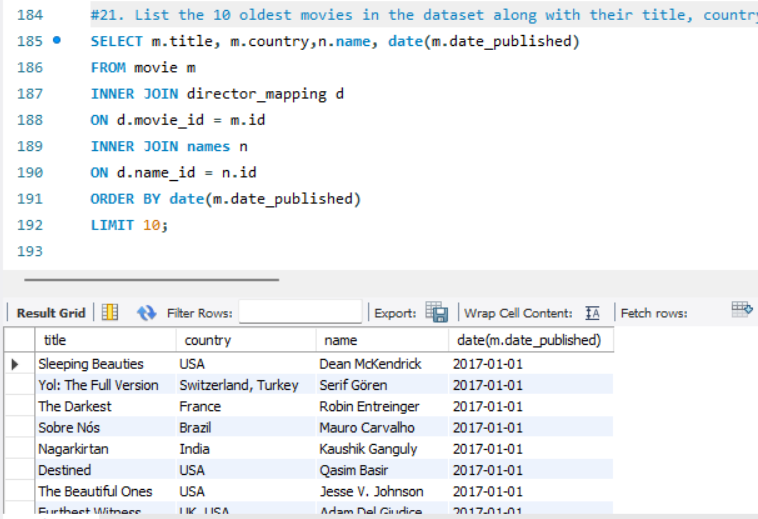


**20. Calculate the average height of actors and actresses separately.**

It joins the role\_mapping and names tables based on name\_id and groups the results by category. The results indicate that the average height of actors is 162.18 cm, while the average height of actresses is 162.47 cm.

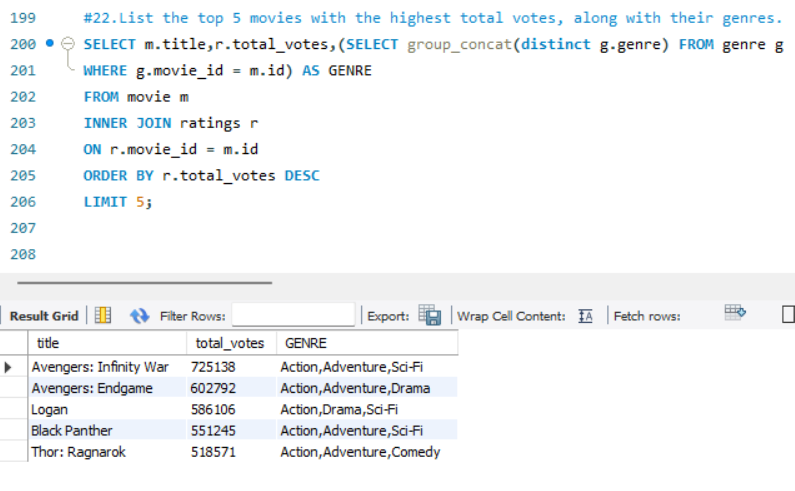
21. List the 10 oldest movies in the dataset along with their title, country, and director.

It joins the movie, director\_mapping, and names tables and orders the results by date\_published in ascending order. The output lists films such as "Sleeping Beauties" (USA) and "Yol: The Full Version" (Switzerland, Turkey), all published on January 1, 2017.

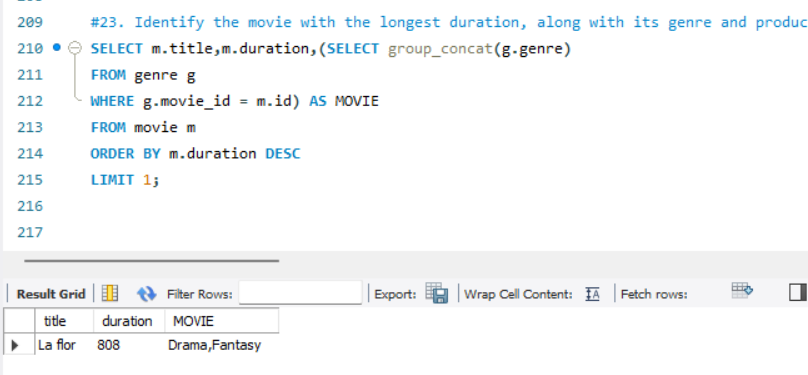


**22. List the top 5 movies with the highest total votes, along with their genres.**

It joins the movie, ratings, and genre tables, using GROUP\_CONCAT to list multiple genres for each movie. The results show popular films like "Avengers: Infinity War" (725,138 votes, Action, Adventure, Sci-Fi) and "Black Panther" (551,245 votes, Action, Adventure, Sci-Fi).

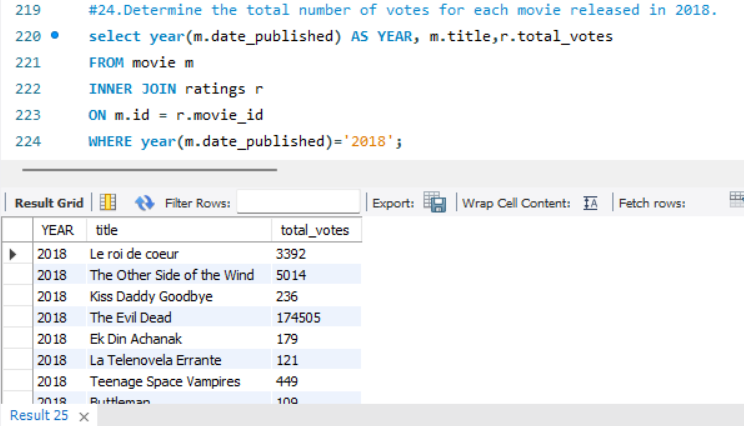
**23. Identify the movie with the longest duration, along with its genre and production company.**

The query joins the movie and genre tables, using GROUP\_CONCAT to list multiple genres. The result shows that "La Flor" has the longest runtime of 808 minutes, categorized under the Drama, Fantasy genres.



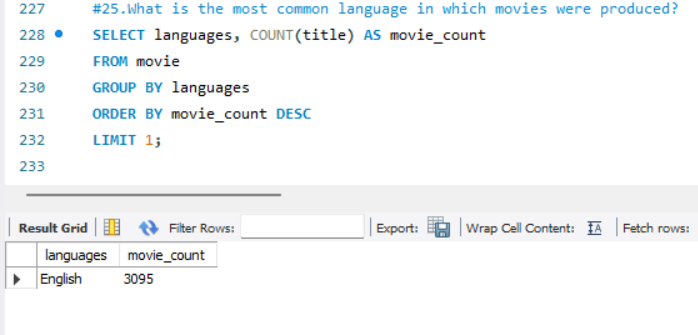
**24. Determine the total number of votes for each movie released in 2018.**

It joins the movie and ratings tables, filtering movies based on the year of publication (2018). The result includes movie titles and their corresponding total votes. Notable entries include "The Evil Dead" with 174,505 votes and "The Other Side of the Wind" with 5,014 votes.



**25. What is the most common language in which movies were produced?**

It counts the number of movies for each language, groups the results, and orders them in descending order. The most common language is English, with 3,095 movies produced.



**CONCLUSION**

The queries analyze various aspects of a movie database. They determine the most common language in movies (English with 3,095 movies) and count how many movies belong exclusively to one genre. The data shows that in 2019, India produced 295 movies, while the USA produced 592 movies. Another query examines the total number of movies released each year and by month. The database also identifies columns containing NULL values, including country, worldwide gross income, languages, and production company. Finally, a query counts the total records in different tables, revealing the names table has the most records (25,735), while the director table has the least (3,867).