SOC - Security Analyst Interview Exercise Documentation of results

By Mariana de San Juan

Understanding the assignment

From two files of datasets in CSV format is necessary to answer the following questions by SQL Queries.

- According to the provided dataset, how many 2018 films were categorized as a Comedy?
- According to the provided dataset, how many 2018 films got a score of 8.0 or higher? (Note that this will require joining the two datasets together)
- What was the best film of 2018?
- Do audiences prefer longer films, or shorter films? You may choose to simply outline your methodology to approach this problem.

The plan

For this assignment, we are going to use HeidiSQL as the database management system (DBMS) to analyze datasets and obtain results through SQL queries.

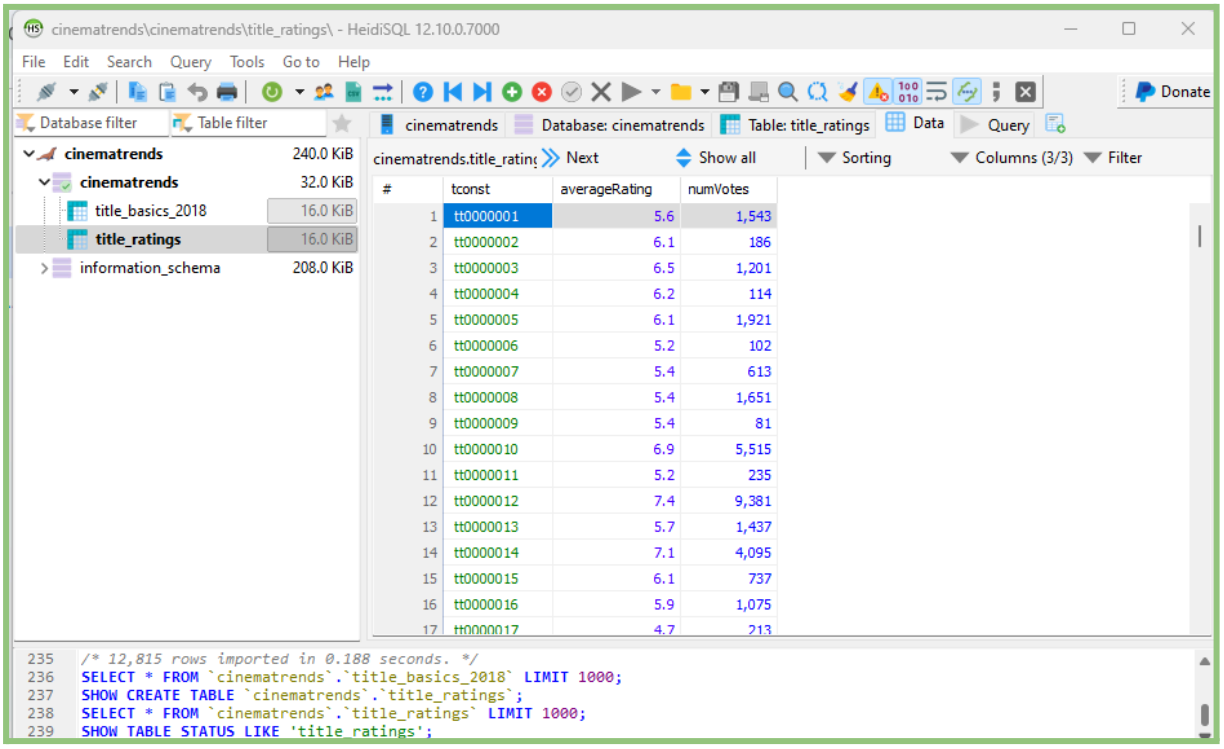
Hands-on step by step

The database model was created under the name 'cinematrends' to proceed with the creation of tables with the following column and data type:

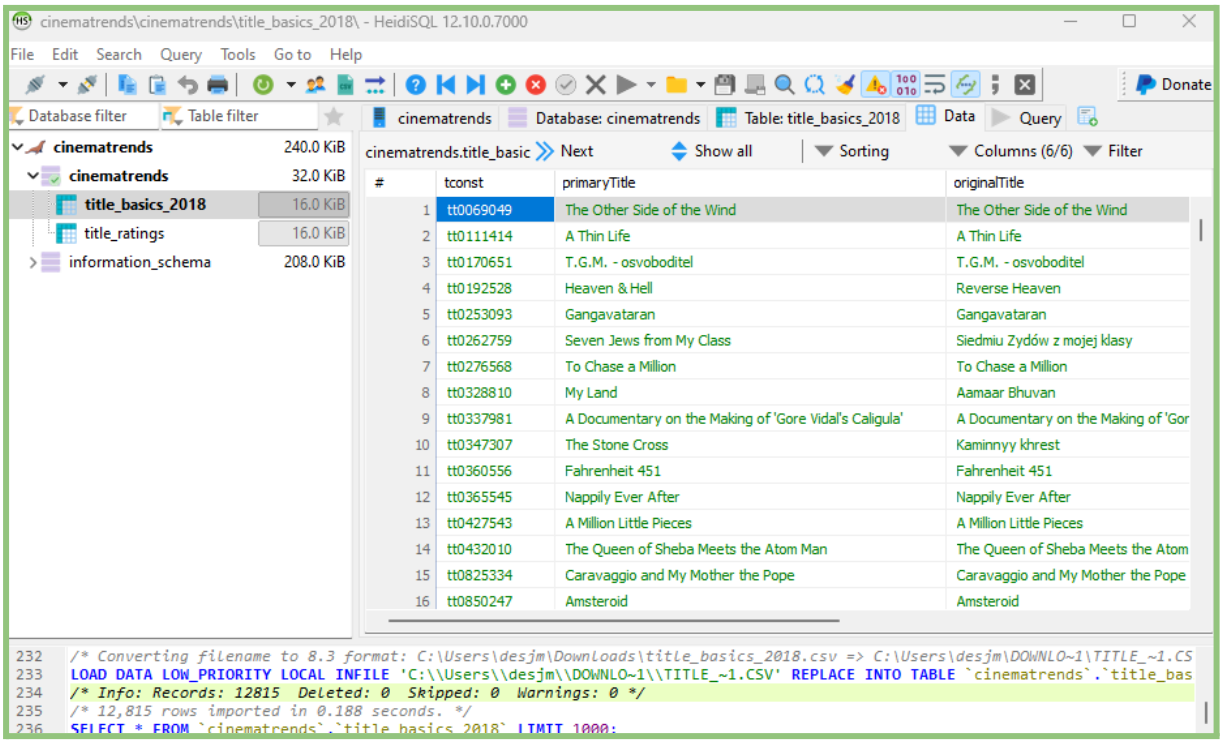
title_basics_2018.csv		
Field	datatype	description
tconst	string	alphanumeric unique identifier of the title
primaryTitle	string	the more popular title / the title used by the filmmakers on promotional materials at the point of release
originalTitle	string	original title, in the original language
year	YYYY	represents the release year of a title
runtimeMinutes	int	primary runtime of the title, in minutes
genres	string array	includes up to three genres associated with the title

title_ratings.csv		
Field	datatype	description
tconst	string	alphanumeric unique identifier of the title
averageRating	float	weighted average of all the individual user ratings
numVotes	int	number of votes the title has received

The next step is importing the datasets from the title_ratings.csv and title_basics_2018.csv files in order to populate the database.



Screenshot 1. Evidence of dataset loaded into the tittle_ratings table.

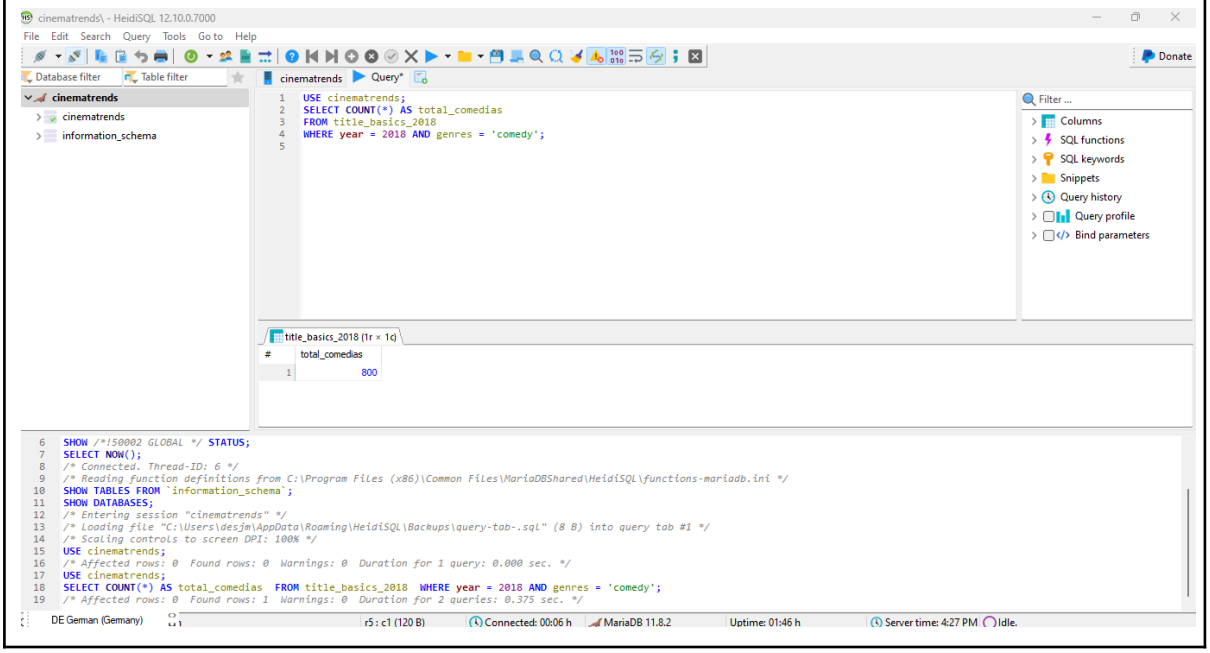


Screenshot 2. Evidence of dataset loaded into the title_basics_2018 table.

Results

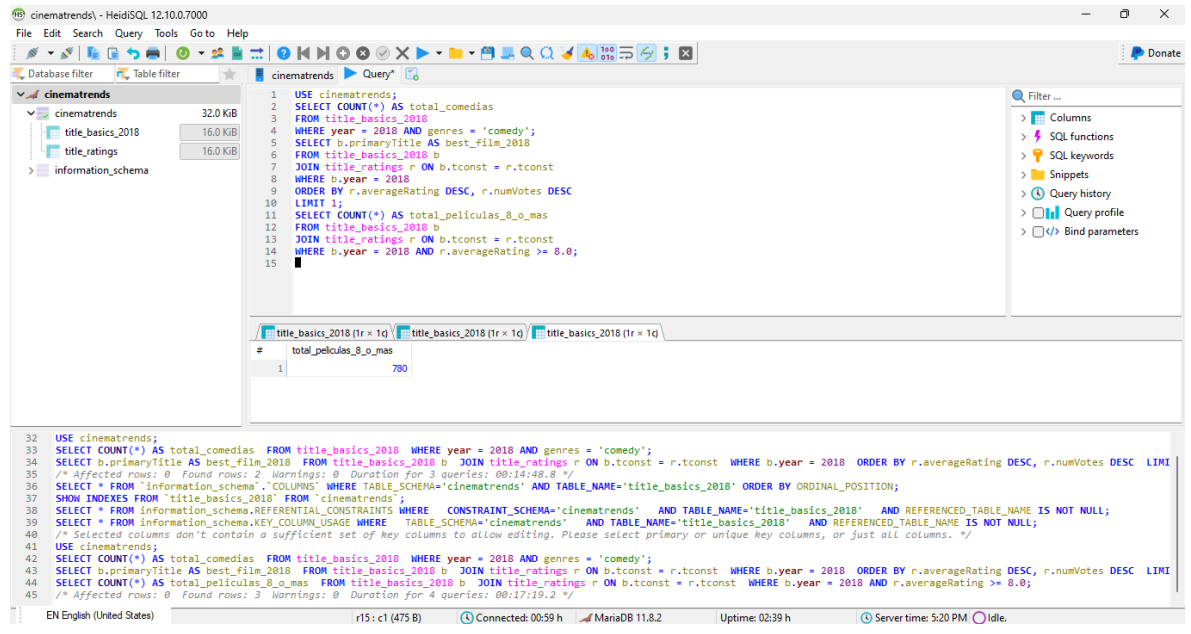
The following step is opening a Query tab and writing the transactions to retrieve the necessary information to answer the questions of this assignment. Now that the datasets have been successfully loaded into our database, we will proceed with crafting SQL queries to address each of our research questions.

According to the provided dataset, how many 2018 films were categorized as a Comedy?



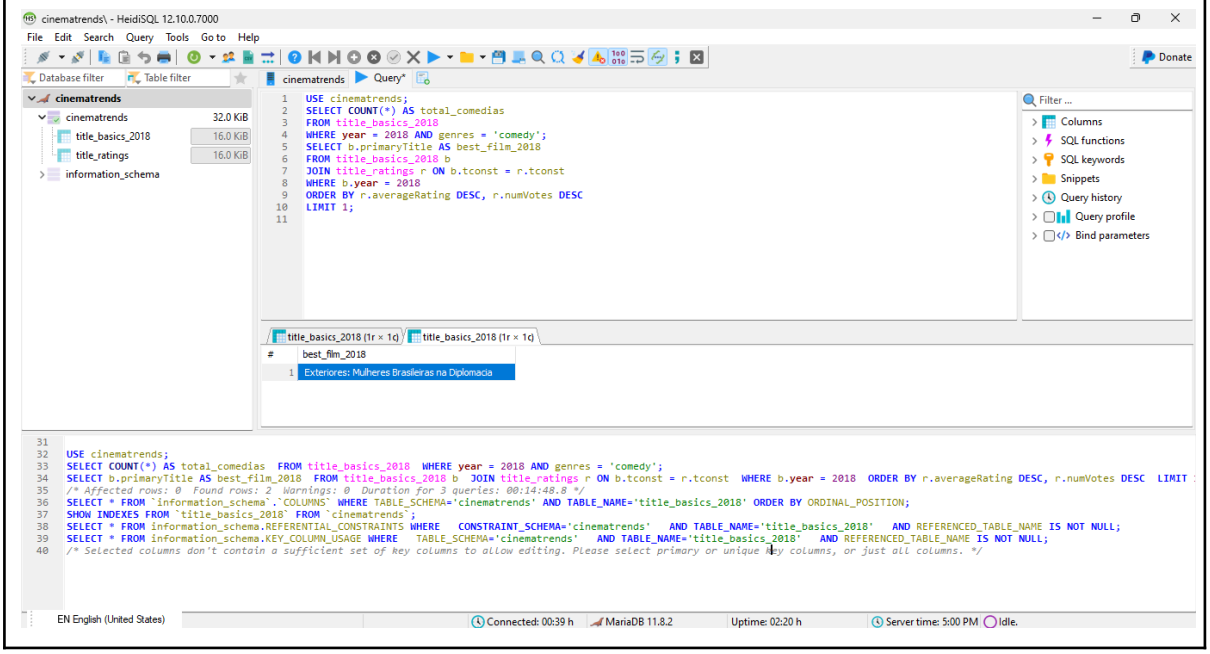
2.According to the results of the Query 800 films from 2018 were categorized as a Comedy.

According to the provided dataset, how many 2018 films got a score of 8.0 or higher? (Note that this will require joining the two datasets together)



3. According to the datasets there are 780 films from 2018 with a score of 8.0 or higher.

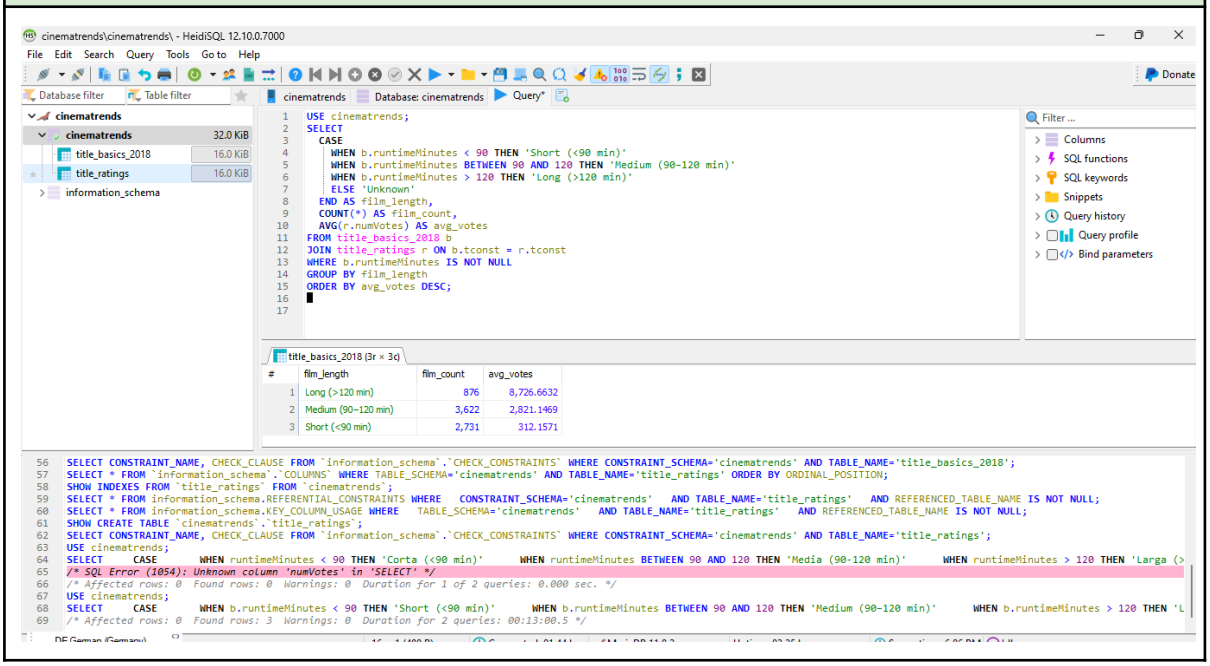
What was the best film of 2018?



4. IMDb crowns the top film by balancing the highest ratings with the greatest number of votes ensuring quality meets popularity. Under this method, the best film of 2018 is: Exteriores: Mulheres Brasileiras na Diplomacia

Do audiences prefer longer films, or shorter films? You may choose to simply outline your methodology to approach this problem.

I wanted to see if people prefer longer or shorter films, so I compared the duration of each movie with the number of audience votes. More votes usually mean more interest. By grouping movies by length and checking the average votes, I got a better idea of what kind of runtime tends to attract the most attention.



5. My analysis suggests that audiences tend to prefer longer films, as these titles show higher engagement through a greater number of votes.

My SQL Script

```
SELECT COUNT(*) AS total_peliculas_8_o_mas
FROM title_basics_2018 b
JOIN title_ratings r ON b.tconst = r.tconst
WHERE b.year = 2018 AND r.averageRating >= 8.0;
SELECT
  CASE
    WHEN b.runtimeMinutes < 90 THEN 'Short (<90 min)'
    WHEN b.runtimeMinutes BETWEEN 90 AND 120 THEN 'Medium (90–120 min)'
    WHEN b.runtimeMinutes > 120 THEN 'Long (>120 min)'
    ELSE 'Unknown'
  END AS film_length,
  COUNT(*) AS film_count,
  AVG(r.numVotes) AS avg_votes
FROM title_basics_2018 b
JOIN title_ratings r ON b.tconst = r.tconst
WHERE b.runtimeMinutes IS NOT NULL
GROUP BY film_length
ORDER BY avg_votes DESC;
```

Conclusion

SQL is a versatile language used to analyze and extract insights from large datasets, tailored to the specific needs of data analysis. With just a few queries, we can uncover meaningful answers. However, the ability of engineers and analysts to solve problems effectively plays a crucial role in minimizing bias and ensuring reliable conclusions. In fact my SQL script has a mix of english and spanish but this is a test, in the workplace I can stick to a single language english or spanish.