Advanced SQL – Reinforcement Project – IMDB Dataset

The dataset provided is a simplified version of the IMDb database, structured to capture essential information about movies, their genres, actors, directors, ratings, and more. This database consists of several tables that contain various details such as:

1. **Movie**: Contains basic information about each movie, including title, release year, duration, country, income, languages, and production companies.
2. **Genre**: Describes the genres associated with each movie.
3. **Director Mapping**: Maps movies to their directors.
4. **Role Mapping**: Maps actors/actresses to movies and specifies the role category (e.g., actor, director, producer).
5. **Names**: Stores information about people (actors, directors, etc.), including their birthdates, heights, and known movies.
6. **Ratings**: Contains ratings information for movies, including the average rating, total votes, and median rating.

Project Objective

The primary objective of this project is to:

- Reinforce key SQL concepts such as joins, aggregation, filtering, and grouping.
- Analyze and extract meaningful insights from a real-world movie dataset.
- Document, present, and improve communication skills by creating a detailed report and a presentation of findings.

Queries to be Performed

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

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

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

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

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

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

7. Calculate the average movie duration for each genre.

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

9. Find the minimum and maximum values for each column in the ratings table, excluding the movie_id column.

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

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

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

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

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

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

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

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

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?
20. Calculate the average height of actors and actresses separately.
21. List the 10 oldest movies in the dataset along with their title, country, and director.
22. List the top 5 movies with the highest total votes, along with their genres.
23. Identify the movie with the longest duration, along with its genre and production company.
24. Determine the total number of votes for each movie released in 2018.
25. What is the most common language in which movies were produced?

Deliverables of the Project

The deliverables for this project should include the following:

1. **SQL Code**: The full set of SQL queries used to answer the questions above. Each query should be properly commented to explain its logic.
2. **Screenshots**: Screenshots of the results of each query. This should include the SQL query and the corresponding output, clearly labelled.
3. **Documentation**: A comprehensive report that includes:
    - An introduction to the dataset and its structure.
    - A description of each query and the rationale behind it.
    - The answers to the questions derived from the queries.
4. **Presentation**: A presentation with 10 slides that summarizes the project. The presentation should include:
    - A brief introduction to the dataset.
    - Key findings from the analysis.
    - Conclusion and insights based on the queries performed.

Evaluation Rubric

Criteria	Points
SQL Code (correctness, clarity and efficiency of SQL Code)	40
Screenshots	20
Documentation	20
Presentation	20
Total	**100**