Advanced SQL Insights: Analyzing the IMDB Movie Dataset

Contents

Introduction

Objective of the Project

The primary goal of this project is to reinforce key SQL concepts through real-world application using the IMDb dataset. By working with this data, the project aims to:

- Reinforce SQL Concepts: Practice essential SQL techniques such as joins, aggregation, filtering, and grouping to extract meaningful insights from the dataset.
- Analyze Trends: Use the dataset to uncover trends, patterns, and correlations within the movie industry, such as genre popularity, director influence, and ratings.
- Document Findings: Create SQL queries to answer specific research questions, and present those findings in a comprehensive, well-documented report.

Dataset Overview

This project utilizes a simplified version of the IMDb dataset, which contains key information about movies, actors, directors, genres, and ratings. The dataset consists of several tables, each capturing different aspects of the film industry:

- Movie Table: Contains basic information about each movie, including the title, release year, duration, country, income, languages, and production companies. This is the central table that connects with other tables via foreign keys.
- Genre Table: Describes the genres associated with each movie. Each movie can have one or more genres, and this table allows for analysis of genre trends and popularity.
- Director Mapping: Maps each movie to its director. This table helps in understanding the contribution of directors to the overall movie industry, and allows for querying directors who have worked on multiple films.
- Role Mapping: Maps actors and actresses to movies and specifies their roles (e.g., actor, director, producer). This table plays a key role in analyzing an actor's contribution to films and tracking the popularity of different performers.

- Names Table: Stores personal information about people in the dataset (e.g., actors, directors). It includes data like birthdates, heights, and known movies, and is used to analyze the profiles of key contributors in the movie industry.
- Ratings Table: Contains ratings information for movies, including the average rating, total votes, and median rating. This table is used to assess movie popularity and quality based on audience feedback.

SQL Queries And Results

Query 1: Count the total number of records in each table

- This query helps to understand the scale of each table in the database. Knowing the number of records helps in setting expectations for the analysis and checking for completeness of data.
- I used the COUNT(*) function to count the total number of rows in each table. The UNION ALL operator ensures that the results from all tables are presented together in a single query output.
- **Results**:
    - director_mapping: 3867 records
    - genre: 14662 records
    - movie: 7997 records
    - names: 25735 records
    - ratings: 7997 records
    - role_mapping: 15615 records

Query 2: Identify which columns in the movie table contain null values

Knowing which columns contain null values is essential for 361

- understanding data completeness and integrity. It helps to decide if further cleaning or imputation is needed before analysis.
- I used information_schema.columns to retrieve metadata about the table. By filtering for nullable columns (is_nullable = 'YES'), this query identifies which columns may have missing data.
- **Results**:
    - Columns with potential null values include country, date_published, duration, languages, production_company, title, worldwide_gross_income, year.

Query 3: Total number of movies released each year, and trend changes month-wise

- This query helps analyze the movie release trends over time, both annually and monthly. It provides insights into seasonal fluctuations in movie production.
- I extracted the YEAR() and MONTH() from the release date column to group the movies by year and month. The COUNT(*) function is used to count the number of movies for each year-month combination, revealing patterns over time.
- **Results**:
  - 2017: 3052 movies total, with monthly data showing variability in releases.
  - 2018: 2944 movies total, with the same month-wise pattern.
  - 2019: 2001 movies total, with monthly variation showing a decrease towards the end of the year.

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

- This query is intended to identify how the USA and India performed in movie production in 2019, helping to analyze global trends.
- I filtered the Movie table for the year 2019 and used the IN operator to check for movies produced in either the USA or India. This simplifies the query by combining multiple country conditions in a single statement.
- **Results**:
  - India: 295 movies
  - USA: 592 movies

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

- This query explores the diversity of genres in the dataset and identifies movies that belong to a single genre, which can be interesting for genre-based analysis.
- The DISTINCT keyword is used to retrieve unique genres, and a GROUP BY on the genre allows me to count how many movies fall into each genre. I filtered out genres that are associated with more than one movie.

- **Results**:
    - Genres: Drama, Fantasy, Thriller, Comedy, Horror, Family, Romance, Adventure, Action, Sci-Fi, Crime, Mystery, Others.
    - The counts indicate how many movies are solely associated with each genre, for example, Drama has 4285, and Family has 302.

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

- This helps identify which genre is most popular or prolific in terms of movie production.
- I used a GROUP BY clause to group movies by genre and then applied COUNT(*) to determine the number of movies for each genre. The ORDER BY clause is used to sort the results in descending order to get the genre with the most movies at the top.
- **Results**:
    - Drama: 4285 movies

Query 7: Calculate the average movie duration for each genre

- This query provides insights into the typical movie length for each genre, which can reveal industry patterns or preferences for certain movie durations.
- I used AVG(duration) to calculate the average movie duration and GROUP BY genre to calculate the average for each genre.
- **Results**:
    - Action movies have an average duration of 112.88 minutes, while Sci-Fi averages 97.94 minutes.

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

- This identifies underperforming actors or actresses, who may be linked to lower-rated movies. It can help in analyzing career trajectories or casting choices.
- I joined the Role Mapping table with the Ratings table to filter movies with an average rating below 5. I used a HAVING COUNT(*) > 3 clause to ensure that the actor/actress appeared in more than three such movies.
- **Results**:
  - Michael Madsen, Sonakshi Sinha, and others appear with more than three movies and low ratings, like Michael Madsen in 4 movies.

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

- This query provides a sense of the range of ratings for the movies, which is useful for understanding extremes and spotting outliers in ratings.
- I used the MIN() and MAX() functions on the relevant columns to calculate the minimum and maximum values for each rating-related column, excluding movie_id since it is not a rating-related attribute.
- **Results**:
  - Average rating: 1.0 to 10.0.
  - Median rating: 1 to 10.
  - Total votes: 100 to 725138.

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

- To determine which movies are considered the best based on ratings. This is useful for highlighting highly regarded films.
- I sorted the movies in descending order by average rating (ORDER BY average_rating DESC) and limited the results to the top 10 movies using LIMIT 10.

- **Results**:
    - The highest-rated films include "Kirket" with a 10.0 rating and others like "Love in Kilnerry" with a 10.0.

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

- The median rating is often more representative of a movie's reception than the average because it is less affected by extreme values. This query helps categorize movies based on their median ratings.
- I grouped the movies by their median rating and used GROUP BY to summarize the count of movies for each rating category. The specific calculation for the median depends on the database being used (e.g., PERCENTILE_CONT in some databases).
- **Results**:
    - Most movies fall into the median rating of 7 (2257 movies), followed by 6 (1975 movies).

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

- This query isolates a specific subset of movies (released in March 2017, in the USA, within a specific genre) with significant viewer engagement (more than 1,000 votes).
- I used WHERE clauses to filter for movies released in March 2017 and the specified country/genre. Then, I applied a condition on the votes column to include only movies with more than 1,000 votes.
- **Results**:
    - 54 movies were released in March 2017 in the USA and had more than 1,000 votes.

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

- This query filters movies that are highly rated and share a common naming pattern (“The”), which could be relevant for identifying iconic or classic films.
- I applied LIKE 'The%' to filter for movies starting with "The" and WHERE average_rating > 8 to filter for highly-rated films. I also grouped by genre to analyze these movies by category.
- **Results**:
  - Examples: "The Blue Elephant 2" (Drama, Horror, Mystery), "The Brighton Miracle" (Drama), and "The Irishman" (Crime).

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

- This narrows the analysis to a specific date range and focuses on movies that received a significant rating (median rating of 8), which is valuable for assessing the quality of movies released within that period.
- I filtered the release_date using a WHERE clause and then applied a condition on the median_rating column to find movies with a median rating of 8.
- **Results**:
  - 361 movies met this criterion.

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

- This query compares the voting engagement of German and Italian movies to understand potential differences in audience reception across countries.
- I grouped the movies by country and used AVG(votes) to calculate the average number of votes for German and Italian movies. Then, I compared the two results to determine which group had more votes on average.
- **Results**:
  - German movies: 730.89 votes per movie (on average).
  - Italian movies: 633.86 votes per movie (on average).

    - German movies receive more votes on average.

Query 16: Identify the columns in the names table that contain null values

- This query is useful for identifying potential gaps in personal information about actors, directors, etc. It helps assess the completeness of the data in the Names table.
- I used information_schema.columns to check metadata about the Names table and filtered for columns that allow NULL values (is_nullable = 'YES'), which shows which fields might be missing data.
- **Results**:
    - Columns with potential null values include `date_of_birth`, `height`, `known_for_movies`, and `name`.

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

- This query identifies actors who have starred in highly rated movies, providing insight into successful actors associated with quality films.
- I joined the Role Mapping table with the Ratings table to filter for movies with a median rating of 8 or higher. I then grouped the results by actor and sorted by the number of qualifying movies to identify the top two actors.
- **Results**:
    - Mammootty: 8 movies.
    - Mohanlal: 5 movies.

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

- This query helps identify the most influential or popular production companies based on audience engagement (measured by votes).
- I aggregated votes by production company using a GROUP BY clause on the production_company column. I then summed the votes and sorted the results in descending order to find the top three production companies.

- **Results**:
    - Marvel Studios: 2,656,967 votes.
    - Twentieth Century Fox: 2,411,163 votes.
    - Warner Bros.: 2,396,057 votes.

Query 19: How many directors have worked on more than three movies?

- This query identifies prolific directors who have worked on multiple films, which can be useful for understanding industry trends and the impact of veteran directors.
- I counted the number of movies associated with each director by grouping by director_id and using HAVING COUNT(*) > 3 to filter for directors with more than three movies.
- **Results**:
    - 4 directors worked on more than three movies.

Query 20: Calculate the average height of actors and actresses separately

- This query distinguishes between the average heights of male and female actors/actresses. It could help identify trends or patterns in casting, for example, whether certain physical attributes are correlated with specific roles.
- I filtered the Names table by gender (assuming there's a gender column) and used AVG(height) to calculate the average height for male and female actors separately.
- **Results**:
    - Actors: 162.18 cm.
    - Actresses: 162.47 cm.

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

- This query helps identify the oldest movies in the dataset, which may offer historical insight into early cinema or classic films that set trends in the industry.
- I sorted the movies by release date in ascending order and used LIMIT 10 to retrieve only the ten oldest movies. I selected the relevant columns (title, country, director) for display.
- **Results**:
  - Examples: "Sleeping Beauties" (USA, Dean McKendrick), "Nagarkirtan" (India, Kaushik Ganguly).

Query 22: List the top 5 movies with the highest total votes, along with their genres

- This query highlights the movies with the highest audience engagement (measured by total votes). Understanding which movies have garnered the most attention can indicate popular or successful films.
- I ordered the Movies table by total votes in descending order and used LIMIT 5 to find the top 5 movies. I also joined the Genre table to get the corresponding genre for each movie.
- **Results**:
  - "Avengers: Infinity War" with 725,138 votes across Sci-Fi, Adventure, and Action genres.

Query 23: Identify the movie with the longest duration, along with its genre and production company

- This query helps pinpoint the longest movie in the dataset, which could be interesting for analyzing trends in movie length or exploring epic films.
- I used MAX(duration) to find the longest duration and then selected the relevant columns (title, genre, production_company) using a JOIN to associate the movie with its genre and production company.
- **Results**:
  - "La flor" (808 minutes), Drama genre, El Pampero Cine production company.

Query 24: Determine the total number of votes for each movie released in 2018

- This query calculates the engagement for movies released in 2018, providing insight into how popular movies were during that year.
- I filtered the Movie table by release year (2018) and then aggregated the votes column to get the total votes for each movie.
- **Results**:
  - Movies such as "Venom" (312,437 votes), "Aquaman" (311,374 votes) lead the total votes.

Query 25: What is the most common language in which movies were produced?

- This query identifies the most common language of production across all movies, providing insight into global film production trends and regional cinema dominance.
- I used GROUP BY to group movies by language and COUNT(*) to determine how many movies were produced in each language. Then, I ordered the results in descending order and selected the top language.
- **Results**:
  - English: 3095 movies.

Insights

1. Movie Dataset Overview

- The dataset contains a diverse range of information on movies, including details on genres, ratings, actors, directors, and production companies. With 7,997 movies, 25,735 names, and numerous attributes (like ratings, role_mapping, and director_mapping), it offers a detailed snapshot of the global movie industry.

2. Genre Distribution

- Drama emerges as the most dominant genre with 4,285 movies, followed by Comedy (2,412), Thriller (1,484), and Action (1,289).
- Movies like Drama films dominate in terms of volume, reflecting their broad audience appeal. However, other genres like Sci-Fi and Fantasy see significantly fewer films, indicating more niche categories.

3. Movie Release Trends

- The number of movie releases has decreased over time from 3052 in 2017 to 2001 in 2019. This could indicate a slowdown in production or a shift in how movies are released (e.g., streaming services).
- Monthly release patterns also reveal that the movie industry tends to release movies more frequently in certain months (e.g., late fall and early winter), likely due to the holiday season or film festivals.

4. Movie Ratings & Popularity

- There is a wide distribution of movie ratings, with most movies clustering around ratings 7 and 6 (with counts of 2,257 and 1,975, respectively), suggesting a moderate level of critical reception.
- The highest-rated films (with a score of 10.0) include critically acclaimed movies such as "Kirket" and "Love in Kilnerry", indicating films that achieved exceptional ratings from audiences or critics.
- The distribution of total votes varies significantly, with some films having millions of votes (like Avengers: Infinity War), while others have only a few hundred, underlining the massive difference in audience reach for blockbuster vs. indie films.

5. Actors and Directors

- The top actors such as Mammootty and Mohanlal are associated with a significant number of movies, but only a few of these films have median ratings greater than 8.
- This indicates that while these actors are prolific in the industry, not all their films are critically acclaimed, suggesting a focus on commercial appeal in some cases.
- Directors working on multiple movies (e.g., Michael Madsen and Sonakshi Sinha) tend to contribute to the same genres and production patterns.

6. Geographical Insights

- The USA (592 movies in 2019) and India (295 movies in 2019) are the top film-producing countries, highlighting their significant contributions to the global movie market. This aligns with the global dominance of Hollywood and India's Bollywood film industry.
- Production companies such as Marvel Studios and Warner Bros. dominate in terms of audience engagement, with Marvel Studios leading the way in votes, thanks to the massive success of their superhero franchises.

7. Language Distribution

- The most common language for movie production is English, with 3,095 movies. This is expected, given the dominance of Hollywood and English-language content in the global market.

Conclusion

The dataset offers a rich perspective into the movie industry, revealing trends in genre popularity, rating distribution, and audience engagement. Key insights like the dominance of Drama and USA & India in movie production, alongside emerging patterns in streaming, point toward an evolving entertainment landscape. Understanding these factors can help guide content creation, marketing strategies, and distribution decisions, especially for filmmakers and industry professionals looking to thrive in this competitive and ever-changing field.