SQL Case Study - Streaming Behavior Analysis

By: Luke Bailey

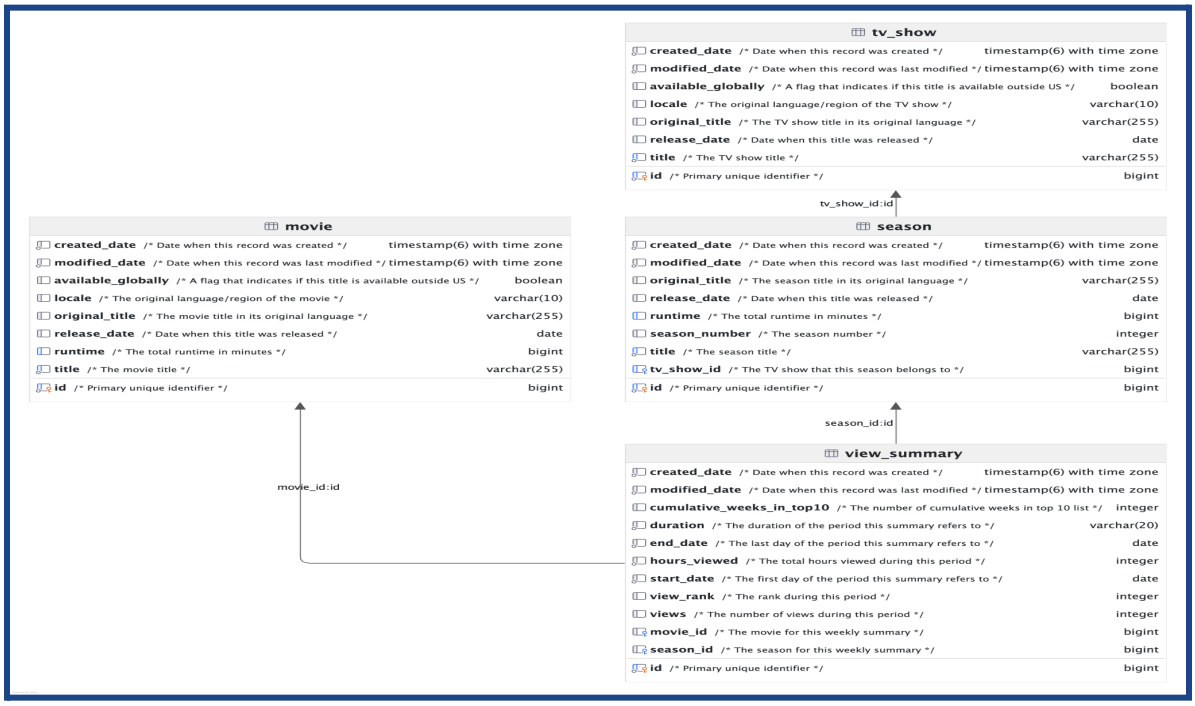
Introduction & About the Dataset

The purpose of this project is to develop, practice, and demonstrate **SQL** query skills including Joins, Sub-Queries, Window Functions, and more. I'll be writing SQL scripts for the publicly available Netflix database, posted to github by user Luis Rocha. I'll be using a local **MySQL** server along with **MySQL Workbench** to create the database and run my queries. The queries will be written based on a list of business questions that I believe will lead to significant insights. The project will conclude with a summary of these insights.

Project Procedure:

1. Familiarize myself with the database
2. Create a list of business questions
3. Query the database to answer business questions
4. Document and record findings
5. Draw useful insights

The Netflix DB is a relational database that's designed to imitate a streaming platform. It contains comprehensive entries about films, tv-shows, and their attributes. Additionally it contains some user data like "hours watched" which can be useful for analysis and business insights. The data model can be seen below.



Business Questions

Based on my understanding of the relational database, I've compiled a list of questions that I intend to answer by writing SQL scripts.

1. What are the 10 most watched movies and tv-shows of all time?
2. How does viewership differ between globally available content and US-only content?
3. Which movies and shows have spent the most total weeks in the top 10?
4. How does user viewership between movies and tv shows compare?
5. Which movies/tv shows had the highest first week viewership?
6. What is the average number of views for tv shows that have 3+ seasons?
7. What period of time yields the most releases for movies and tv shows?
8. How many movies/ tv shows are available in the US versus globally?
9. What movie/ tv shows have the highest viewership by year?
10. Which movie/ tv show had the highest drop in viewership after the first week?

Queries and Techniques

To answer the questions laid out in the previous section, I wrote a series of SQL queries that aimed to achieve:

- Simplicity in their construction
- Efficiency in their execution time
- Readability and organization in their responses

Question	Scripts + Queries	Used
What are the 10 most watched movies and tv-shows of all time?	ten_most_watched.sql • 1.1: Top 10 Movies by hours viewed • 1.2: Top 10 TV Shows by hours viewed • 1.3: Top 10 Movies/TV Shows by hours viewed	SUM, AS, JOIN, GROUP BY, ORDER BY, LIMIT
How does viewership differ between globally available content and US-only content?	global_v_us.sql • 2.1: Global versus US Total Hours for Movie • 2.2: Global versus US Total Hours for TV Shows	SUM, AS, JOIN, GROUP BY, ORDER BY, LIMIT, CASE WHEN, THEN, ELSE
Which movies and shows have spent the	top_ten_weeks.sql • 3.1: Movies and TV Shows by total weeks in top 10	SUM, AS, JOIN, GROUP BY, ORDER BY, LIMIT, MAX

most total weeks in the top 10?		
How does user viewership between movies and tv shows compare?	compare.sql ● 4.1: Total Hours by Type ● 4.2: Average Hours per Title ● 4.3: Average Weekly Hours Viewed	SUM, AS, JOIN, GROUP BY, ORDER BY, LIMIT, AVG, Sub Queries
Which movies/tv shows had the highest first week viewership?	week_one.sql ● 5.1: First week views for movies and tv shows	JOIN, ROW_NUMBER(), PARTITION BY, WITH, OVER(), WITH, AS
What is the average number of views for tv shows that have 3+ seasons?	long_show.sql ● 6.1: Average views for tv shows >= 3 seasons ● 6.2: Average views for tv shows < 3 seasons	WITH, HAVING, JOIN, AVG(), COUNT()
What period of time yields the most releases for movies and tv shows?	popular_times.sql ● 7.1: Releases by Year ● 7.2: Releases by Month	COUNT(), YEAR(), MONTHNAME(), UNION ALL, IS NOT NULL
How many movies/ tv shows are available in the US versus globally?	availability.sql ● 8.1: Titles US versus Global	COUNT(), CASE, WHEN, THEN, ELSE, UNION ALL, GROUP BY
What movie/ tv shows have the highest viewership by year/month?	views_by_time.sql ● 9.1: Highest Viewership by Year ● 9.2: Highest Viewership by Month	JOIN, ROW_NUMBER(), PARTITION BY, WITH, OVER(), WITH, AS, YEAR(), MONTHNAME()
Which movie/ tv show had the highest drop in viewership after the first week?	highest_drop.sql ● 10.1: Movie Drop ● 10.2: TV Drop ● 10.3: Movie and TV Drop	WITH(), ROW_NUMBER(), JOIN, ROUND(), UNION ALL, Arithmetic

Each of the queries were run, and the results were organized into an Excel dashboard for visualization.

Insights and Takeaways

Throughout my exploratory data analysis process, I noted various key insights that have been summarized below.

- The top movie is Leave the World Behind with 687 Million hours viewed.
- The top tv show is Bridgerton with 2.7 Billion hours viewed.
- TV shows consistently outperform movies in watch hours, likely due to TV shows having more watchable hours.
- Movies are more US centric, while TV shows have a larger international audience
- The Menendez Brothers dropped 69.8% of viewers after the first week.
- Most releases occurred between 2021 and 2023, likely because of Netflix's aggressive approach post pandemic