Intro To Database System Lab

CSC217(FALL-2024/MWE-22732)

Report :

Netflix Data Analysis

Instructor: Muhammad Irtiza

Prepared by:

Aakanksha Pardeep (20221-32743)

Alisha Nasir (20221-33353)

Saima Batool (20221-33102)

Date of submission: 19th December 2024

Contents

Problem Statement:

The world of online streaming services is growing rapidly, and with platforms like Netflix offering an extensive collection of movies, TV shows, documentaries, and more, the amount of data generated is immense. For content managers, marketers, and decision-makers, being able to efficiently analyze this data to gain insights into user preferences, content trends, and platform performance is crucial. This study seeks to address key business questions surrounding Netflix's catalog using SQL-based data analysis to derive insights from available content data. The following problem statement explores how the SQL queries below can provide valuable business intelligence:

1. **Content Classification and Trends**:
   How does the distribution of content between movies and TV shows compare on Netflix? What are the most common ratings across different content types? This will help in understanding user preferences and categorizing content based on viewership patterns.
2. **Yearly and Genre-Based Trends**:
   What trends can be observed in the release of movies and TV shows, particularly by country and genre? Analyzing the number of releases, genres, and content from different countries like India will enable better content planning and regional strategy alignment.
3. **Top Content and Creators**:
   How can content, such as the longest movie or the top-rated TV shows, be identified? This will provide content managers with insights into user interests, enabling the platform to showcase the most engaging content. Additionally, finding out which actors have appeared the most in movies produced in specific countries will provide insights into popular talent.
4. **Content Categorization**:
   What impact do keywords like "violence" and "kill" in content descriptions have on content categorization? By classifying content as 'Good' or 'Bad', managers can evaluate the content based on user feedback and potentially take action on sensitive topics. Moreover, using triggers can automate categorization, making the system more responsive to data changes.
5. **Content Availability and Regional Focus**:
   Which countries have the most content available on Netflix, and how can the platform focus more on localized content? By understanding the distribution of content by country, Netflix can refine its localization strategies and content acquisition.

By addressing these questions, the analysis will generate insights such as:

- Which countries contribute the most to the platform’s content library.
- The most common ratings for movies and TV shows, helping refine content acquisition and recommendation strategies.
- A classification of content based on genres and regions, optimizing content marketing and recommendations.
- A deeper understanding of content types, with a focus on documentaries and other specific genres.

In conclusion, this study aims to leverage SQL queries to provide strategic insights into Netflix’s global content portfolio, contributing to better decision-making processes for content managers, marketers, and business analysts.

Abstract:

The rise of online streaming platforms, particularly Netflix, has transformed the entertainment industry by providing vast libraries of content to users worldwide. With millions of movies, TV shows, documentaries, and other types of media, managing and extracting meaningful insights from this massive amount of data has become a significant challenge for content managers, marketers, and analysts. This study explores the use of SQL-based data analysis to derive actionable insights from Netflix's content dataset, aiming to answer critical business questions that can optimize platform operations and content strategies.

The analysis focuses on 15 specific business problems, such as comparing the number of movies versus TV shows, identifying the most common ratings for different content types, and finding the longest movie or the top actors in specific regions. It further delves into the categorization of content based on keywords such as "violence" or "kill" to identify content trends, while also providing insights into the average content release rates by country. Additionally, the study evaluates the availability of content across different countries and how it influences regional strategy, with a focus on top countries contributing the most content.

By utilizing SQL queries, the study examines how Netflix's data can be analyzed to identify patterns in content consumption, evaluate the effectiveness of content strategies, and provide better recommendations to users. Furthermore, the study integrates automated categorization through triggers, enabling real-time updates on content classification.

The findings of this research aim to enhance Netflix's data-driven decision-making, particularly in the areas of content acquisition, marketing, user experience, and regional content strategies. Through actionable insights, the study supports the development of more personalized and effective content recommendations, better understanding of content preferences, and efficient allocation of resources in a competitive streaming market.

https://www.kaggle.com/datasets/shivamb/netflix-shows?resource=download

Queries:

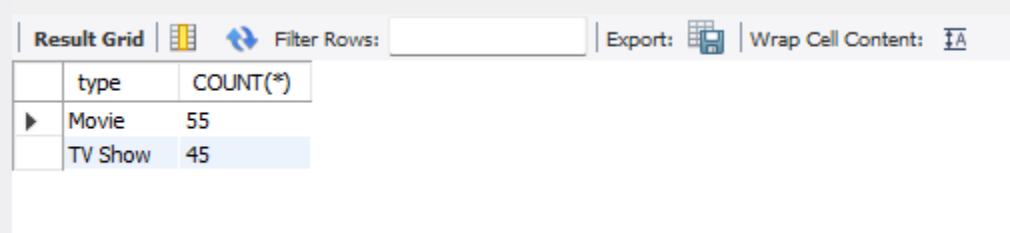
1. Problem: Count the Number of Movies vs TV Shows

Problem: Netflix has various types of content—movies, TV shows, and possibly other content types. The problem is to determine the exact count of each content type (e.g., Movies vs. TV Shows).

Solution: The SQL query groups the data by type (which categorizes each entry as a Movie or TV Show), then uses the COUNT function to return the number of movies and TV shows.

```
-- 1. Count the number of Movies vs TV Shows

SELECT
    type,
    COUNT(*)
FROM netflix
GROUP BY 1;
```

Output:



type	COUNT(*)
Movie	55
TV Show	45

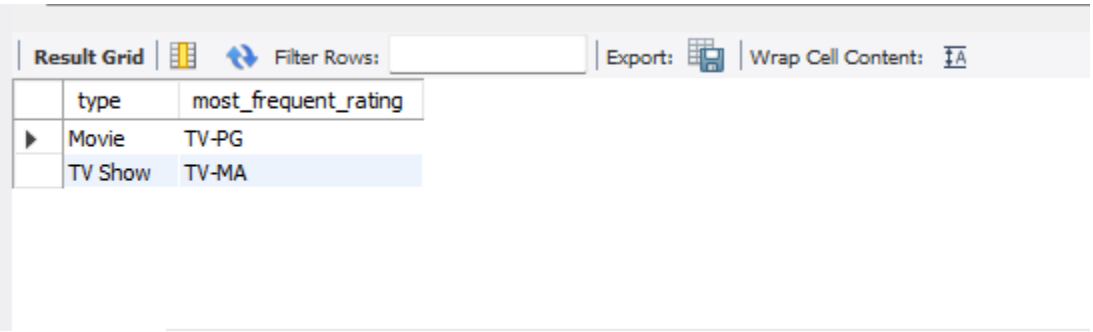
2. Problem: Find the Most Common Rating for Movies and TV Shows

Problem: There is a need to determine the most common (frequent) rating for both Movies and TV Shows. This is useful for understanding the general audience rating distribution for each content type.

Solution: The query counts the number of occurrences of each rating by type (Movies and TV Shows) and returns the rating with the maximum count for each content type.

```sql
-- 2. Find the most common rating for movies and TV shows
SELECT
    rc.type,rc.rating AS most_frequent_rating
FROM (SELECT type,rating,
        COUNT(*) AS rating_count
    FROM netflix
    GROUP BY type, rating
) rc
JOIN (SELECT type,
        MAX(rating_count) AS max_rating_count
    FROM (SELECT type,rating,
            COUNT(*) AS rating_count
        FROM netflix
        GROUP BY type, rating
    ) subquery
    GROUP BY type
) max_rc ON rc.type = max_rc.type
WHERE rc.rating_count = max_rc.max_rating_count
ORDER BY rc.type;
```

Output:



type	most_frequent_rating
Movie	TV-PG
TV Show	TV-MA

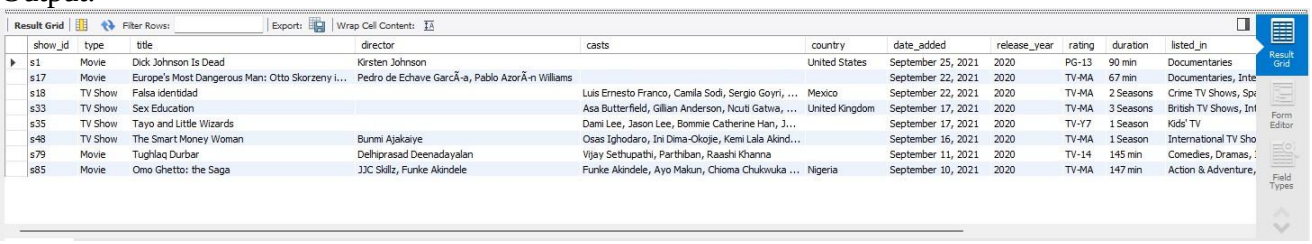
3. Problem: List All Movies Released in a Specific Year (e.g., 2020)

Problem: You want to filter out the movies that were released in a specific year (e.g., 2020). This is useful for understanding Netflix's content release strategy in a particular year.

Solution: A simple WHERE clause filters the release_year for the year in question (2020).

```
33    -- 3. List all movies released in a specific year (e.g., 2020)
34
35    SELECT *
36    FROM netflix
37    WHERE release_year = 2020;
38
```

Output:



show_id	type	title	director	casts	country	date_added	release_year	rating	duration	listed_in
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson		United States	September 25, 2021	2020	PG-13	90 min	Documentaries
s17	Movie	Europe's Most Dangerous Man: Otto Skorzeny i...	Pedro de Echave GarcÃ-a, Pablo AzorÃ-n Williams			September 22, 2021	2020	TV-MA	67 min	Documentaries, Inte
s18	TV Show	Falsa identidad		Luis Ernesto Franco, Camila Sodi, Sergio Goyri, ...	Mexico	September 22, 2021	2020	TV-MA	2 Seasons	Crime TV Shows, Spa
s33	TV Show	Sex Education		Asa Butterfield, Gillian Anderson, Ncuti Gatwa, ...	United Kingdom	September 17, 2021	2020	TV-MA	3 Seasons	British TV Shows, Int
s35	TV Show	Tayo and Little Wizards		Dami Lee, Jason Lee, Bommie Catherine Han, J...		September 17, 2021	2020	TV-Y7	1 Season	Kids' TV
s48	TV Show	The Smart Money Woman	Bunmi Ajakaiye	Osas Ighodaro, Ini Dima-Okojie, Kemi Lala Akind...		September 16, 2021	2020	TV-MA	1 Season	International TV Sho
s79	Movie	Tughlaq Durbar	Delhiprasad Deenadayalan	Vijay Sethupathi, Parthiban, Raashi Khanna		September 11, 2021	2020	TV-14	145 min	Comedies, Dramas,
s85	Movie	Omo Ghetto: the Saga	JJC Skillz, Funke Akindele	Funke Akindele, Ayo Makun, Chioma Chukwuka ...	Nigeria	September 10, 2021	2020	TV-MA	147 min	Action & Adventure,

4. Problem: Find the Top 5 Countries with the Most Content on Netflix

Problem: This query helps in identifying which countries contribute the most content to Netflix's global library, which is crucial for understanding content distribution.

Solution: The procedure splits the countries from the comma-separated list and counts the occurrences of each country. The result is sorted in descending order, showing the top 5 countries with the most content.

```sql
-- 4. Find the top 5 countries with the most content on Netflix
DELIMITER $$
CREATE PROCEDURE GetTopCountries()
BEGIN
-- Simple query to count occurrences of each country from the comma-separated list in the 'country' column
    SELECT country,
        COUNT(*) AS total_content FROM (
        SELECT
            TRIM(SUBSTRING_INDEX(SUBSTRING_INDEX(country, ',', n.n), ',', -1)) AS country
        FROM netflix
        JOIN (SELECT 1 AS n UNION ALL SELECT 2 UNION ALL SELECT 3 UNION ALL SELECT 4 UNION ALL
            SELECT 5 UNION ALL SELECT 6 UNION ALL SELECT 7 UNION ALL SELECT 8 UNION ALL
             SELECT 9 UNION ALL SELECT 10) n
        ON LENGTH(country) - LENGTH(REPLACE(country, ',', '')) >= n.n - 1
        WHERE country IS NOT NULL
    ) AS countries
    GROUP BY country
    ORDER BY total_content DESC
    LIMIT 5;
END$$
DELIMITER ;

CALL GetTopCountries();
```

Output:

United States	25
Japan	14
United Kingdom	8
India	7

Result 26

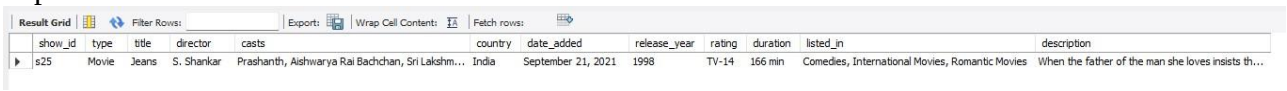
5. Problem: Identify the Longest Movie

Problem: To help users find the longest movie on Netflix, this query determines the movie with the maximum duration.

Solution: This query filters for type = 'Movie' and orders the data by the movie's duration, descending, to get the longest movie.

```sql
-- 5. Identify the longest movie
SELECT *
FROM netflix
WHERE type = 'Movie'
ORDER BY CAST(SUBSTRING_INDEX(duration, ' ', 1) AS UNSIGNED) DESC
LIMIT 1;
```

Output



show_id	type	title	director	casts	country	date_added	release_year	rating	duration	listed_in	description
s25	Movie	Jeans	S. Shankar	Prashanth, Aishwarya Rai Bachchan, Sri Lakshm…	India	September 21, 2021	1998	TV-14	166 min	Comedies, International Movies, Romantic Movies	When the father of the man she loves insists th…

6. Problem: List All TV Shows with More than 5 Seasons

Problem: This query identifies TV shows that have a substantial number of seasons (more than 5), which could be an indication of popularity or content longevity.

Solution: By filtering based on the duration of TV shows, this query finds those with more than 5 seasons.

```sql
-- 6. List all TV shows with more than 5 seasons
SELECT *
FROM netflix
WHERE type = 'TV Show'
  AND CAST(SUBSTRING(duration, 1, LENGTH(duration) - 7) AS UNSIGNED) > 5;
```

Output:

show_id	type	title	director	casts	country	date_added	release_year	rating	duration	listed_in	description
s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Holl…	United Kingdom	September 24, 2021	2021	TV-14	9 Seasons	British TV Shows, Reality TV	A talented batch of amat
s56	TV Show	Nailed It		Nicole Byer, Jacques Torres	United States	September 15, 2021	2021	TV-PG	6 Seasons	Reality TV	Home bakers with a terri
s66	TV Show	Numberblocks		Beth Chalmers, David Holt, Marcel Mcalla, Fer…	United Kingdom	September 15, 2021	2021	TV-Y	6 Seasons	Kids' TV	In a place called Numberl
s68	TV Show	Saved by the Bell		Mark-Paul Gosselaar, Tiffani Thiessen, Mario Lo…	United States	September 15, 2021	1994	TV-PG	9 Seasons	Kids' TV, TV Comedies	From middle school to coll
s83	TV Show	Lucifer		Tom Ellis, Lauren German, Kevin Alejandro, D.B.…	United States	September 10, 2021	2021	TV-14	6 Seasons	Crime TV Shows, TV Comedies, TV Dramas	Bored with being the Lord

7. Problem: Count the Number of Content Items in Each Genre

Problem: This query helps categorize and count the number of content items in each genre, providing insights into content diversity.

Solution: The query extracts individual genres from the listed_in column, which holds comma-separated genre values, and counts occurrences for each.

```sql
-- 7 . Count the number of content items in each genre
SELECT genre, COUNT(*) AS total_content
FROM (
    SELECT TRIM(SUBSTRING_INDEX(SUBSTRING_INDEX(listed_in, ',', 1), ',', -1)) AS genre FROM netflix
    WHERE listed_in IS NOT NULL
    UNION ALL
    SELECT TRIM(SUBSTRING_INDEX(SUBSTRING_INDEX(listed_in, ',', 2), ',', -1)) AS genre FROM netflix
    WHERE listed_in IS NOT NULL
    UNION ALL
    SELECT TRIM(SUBSTRING_INDEX(SUBSTRING_INDEX(listed_in, ',', 3), ',', -1)) AS genre FROM netflix
    WHERE listed_in IS NOT NULL
    UNION ALL
    SELECT TRIM(SUBSTRING_INDEX(SUBSTRING_INDEX(listed_in, ',', 4), ',', -1)) AS genre FROM netflix
    WHERE listed_in IS NOT NULL
    UNION ALL
    SELECT TRIM(SUBSTRING_INDEX(SUBSTRING_INDEX(listed_in, ',', 5), ',', -1)) AS genre FROM netflix
    WHERE listed_in IS NOT NULL
) AS genres
GROUP BY genre
ORDER BY total_content DESC;
```

Output:

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

genre	total_content
International Movies	73
Kids' TV	38
Thrillers	34
Children & Family Movies	33
International TV Shows	32
Reality TV	32
TV Dramas	30
Action & Adventure	28
TV Comedies	26
Dramas	24
Comedies	24
Documentaries	15
Anime Features	12
TV Action & Adventure	10
TV Sci-Fi & Fantasy	10
Docuseries	9
TV Mysteries	9
Crime TV Shows	7
Sports Movies	7
Music & Musicals	7
Romantic Movies	6
British TV Shows	5
Sci-Fi & Fantasy	5
Romantic TV Shows	5
Horror Movies	4
Korean TV Shows	4
TV Thrillers	3
Anime Series	2
Independent Movies	2

Result 29

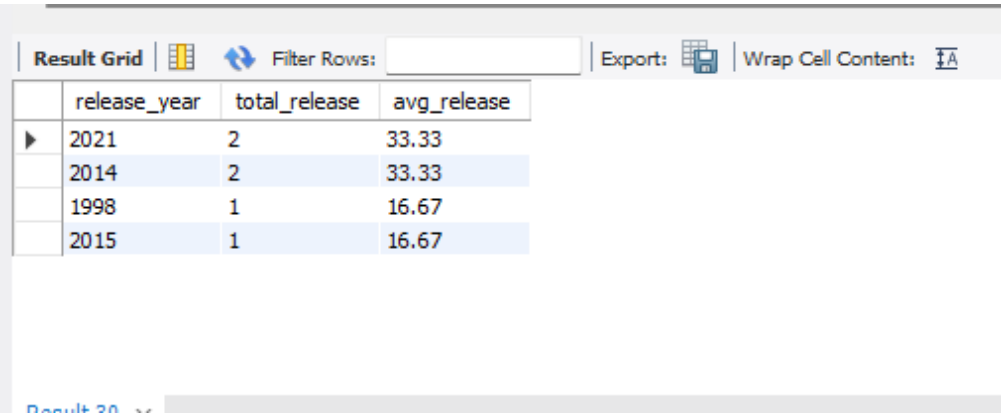
8. Problem: Find Each Year and the Average Number of Content Released by India on Netflix

Problem: You need to understand the trend of Netflix content released by India over the years, including the average number of releases.

Solution: This query calculates the yearly release count by India, compares it to the total releases, and returns the top 5 years with the highest average release rates.

```
-- 8 . Find each year and the average numbers of content release by India on netflix.
-- return top 5 year with highest avg content release !
SELECT
    release_year,
    COUNT(show_id) AS total_release,
    ROUND(
        COUNT(show_id) / total_releases_per_year.total_releases * 100, 2) AS avg_release
FROM netflix
JOIN (
    SELECT
        COUNT(show_id) AS total_releases
    FROM netflix
    WHERE country = 'India'
) AS total_releases_per_year
WHERE country = 'India'
GROUP BY release_year
ORDER BY avg_release DESC
LIMIT 5;
```

Output:



release_year	total_release	avg_release
2021	2	33.33
2014	2	33.33
1998	1	16.67
2015	1	16.67

9. Problem: List All Movies that are Documentaries

Problem: This query filters all the movies categorized as documentaries, providing insights into non-fictional content on Netflix.

Solution: Using the LIKE operator, it identifies content with 'Documentaries' in the listed_in field.

```
-- 9. List all movies that are documentaries
SELECT *
FROM netflix
WHERE listed_in LIKE '%Documentaries%';
```

Output:

show_id	type	title	director	casts	country	date_added	release_year	rating	duration	listed_in
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson		United States	September 25, 2021	2020	PG-13	90 min	Documentaries
s17	Movie	Europe's Most Dangerous Man: Otto Skorzeny i...	Pedr			September 22, 2021	2020	TV-MA	67 min	Documentaries, Internatio
s46	Movie	My Heroes Were Cowboys	Tyler Greco			September 16, 2021	2021	PG	23 min	Documentaries
s69	Movie	Schumacher	Hanns-Bruno KammertÃ¶ns, Vanessa NÃ¶cker, ...	Michael Schumacher		September 15, 2021	2021	TV-14	113 min	Documentaries, Internatio
s89	Movie	Blood Brothers: Malcolm X & Muhammad Ali	Marcus Clarke	Malcolm X, Muhammad Ali		September 9, 2021	2021	PG-13	96 min	Documentaries, Sports Mov
s92	Movie	The Women and the Murderer	Mona Achache, Patricia Tourancheau		France	September 9, 2021	2021	TV-14	92 min	Documentaries, Internatio
s97	Movie	If I Leave Here Tomorrow: A Film About Lynyrd ...	Stephen Kijak	Ronnie Van Zandt, Gary Rossington, Allen Collin...	United States	September 7, 2021	2018	TV-MA	97 min	Documentaries, Music & Mu

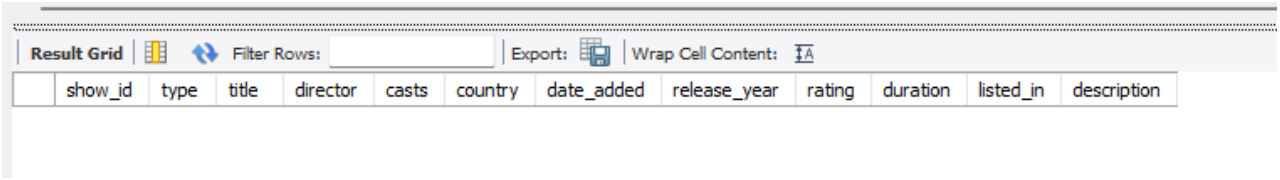
10. Problem: Find All Content Without a Director

Problem: This query identifies all content entries (movies/TV shows) that do not have an associated director, helping to identify gaps in content metadata.

Solution: A straightforward query checks for NULL in the director field.

```
-- 10. Find all content without a director
SELECT *
FROM netflix
WHERE director IS NULL;
```

Output



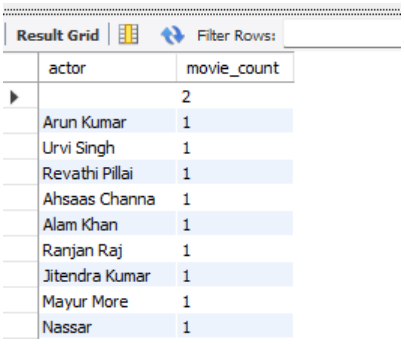
11. Problem: Find the Top 10 Actors Who Have Appeared in the Highest Number of Movies Produced in India

Problem: This query identifies the top 10 actors who have appeared the most in movies produced in India, which can help Netflix assess the popularity of certain actors in Indian productions.

Solution: It extracts actor names from the casts column and counts their appearances, filtering for content produced in India.

```sql
-- 11. Find the top 10 actors who have appeared in the highest number of movies produced in India.

SELECT actor, COUNT(*) AS movie_count
FROM (
    SELECT TRIM(SUBSTRING_INDEX(SUBSTRING_INDEX(casts, ',', n.n), ',', -1)) AS actor
    FROM netflix
    JOIN (SELECT 1 AS n UNION SELECT 2 UNION SELECT 3 UNION SELECT 4 UNION SELECT 5 UNION
                SELECT 6 UNION SELECT 7 UNION SELECT 8 UNION SELECT 9 UNION SELECT 10) n
    ON LENGTH(casts) - LENGTH(REPLACE(casts, ',', '')) >= n.n - 1
    WHERE country = 'India' AND casts IS NOT NULL
) AS actors
GROUP BY actor
ORDER BY movie_count DESC
LIMIT 10;
```

Output:



12. Problem: Categorize Content Based on Keywords in Description (e.g., 'kill', 'violence')

Problem: This query categorizes content as 'Good' or 'Bad' based on the presence of certain keywords in the description, such as "kill" or "violence", and counts the number of items in each category.

Solution: The query uses a CASE statement to categorize content, and the LIKE operator to search for specific keywords in the description field.

```
/*
Question 12:
Categorize the content based on the presence of the keywords 'kill' and 'violence' in
the description field. Label content containing these keywords as 'Bad' and all other
content as 'Good'. Count how many items fall into each category.
*/
SELECT category,type,COUNT(*) AS content_count
FROM (SELECT *,CASE WHEN description LIKE '%kill%' OR description LIKE '%violence%' THEN 'Bad'ELSE 'Good'END AS category
    FROM netflix
) AS categorized_content
GROUP BY category, type
ORDER BY content_count DESC;
```

Output:

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

category	type	content_count
Good	Movie	52
Good	TV Show	44
Bad	Movie	3
Bad	TV Show	1

 Problem: Content on Netflix needs to be automatically categorized based on certain keywords in the description (such as "kill" or "violence") for content moderation or filtering purposes.

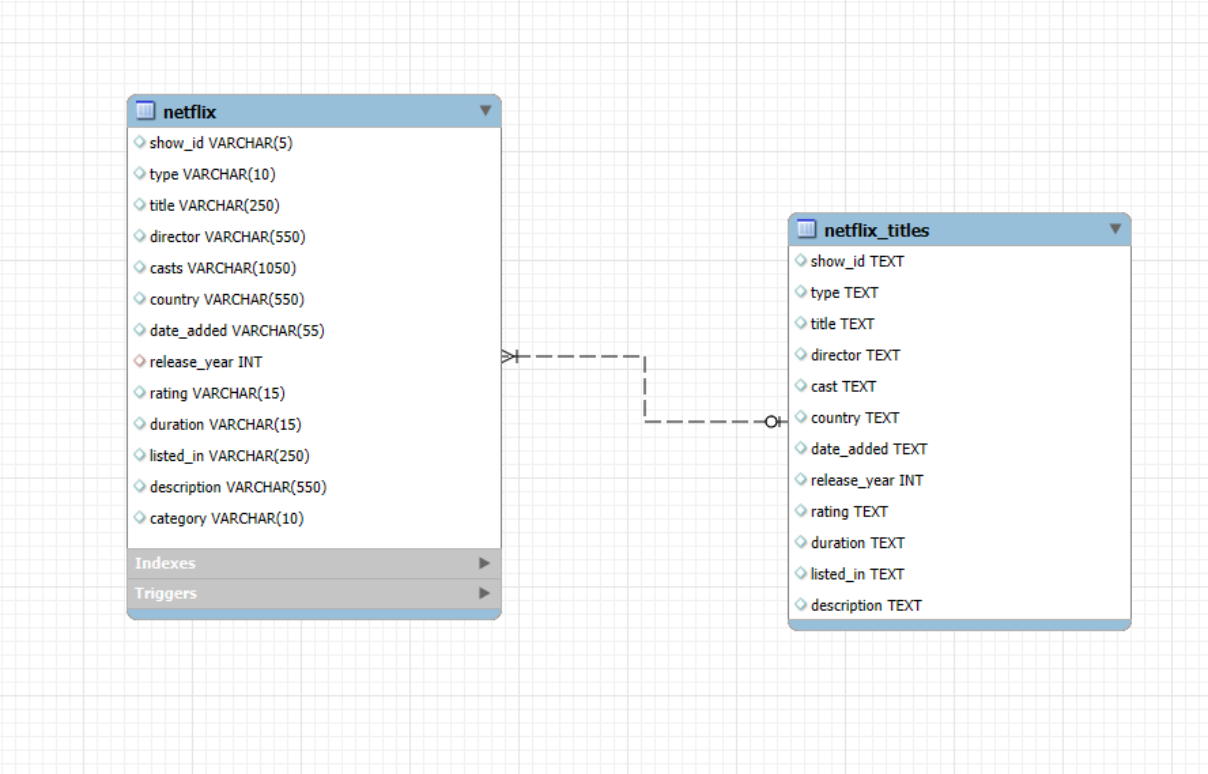
 Solution: A **MySQL trigger** is implemented to automatically assign a category ("Bad" or "Good") to each new row inserted into the netflix table based on the content of the description field. This ensures that content is categorized in real-time without manual intervention.

```
-- Triggers
-- Ensure the category column exists
ALTER TABLE netflix
ADD COLUMN category VARCHAR(10);

-- Trigger to update category when a new row is inserted
DELIMITER $$

CREATE TRIGGER update_category_before_insert
BEFORE INSERT ON netflix
FOR EACH ROW
BEGIN
    -- Check if description contains 'kill' or 'violence'
    IF NEW.description LIKE '%kill%' OR NEW.description LIKE '%violence%' THEN
        -- Set category to 'Bad' if keywords are found
        SET NEW.category = 'Bad';
    ELSE
        -- Otherwise, set category to 'Good'
        SET NEW.category = 'Good';
    END IF;
END $$

DELIMITER ;
```

Output:

```
93  10:29:30  CREATE TRIGGER update_category_before_insert BEFORE INSERT ON netflix FOR EACH ROW BEGIN  -- Check if description contains 'kill' or 'vi...  0 row(s) affected
```

ERD diagram


netflix
show_id VARCHAR(5)
type VARCHAR(10)
title VARCHAR(250)
director VARCHAR(550)
casts VARCHAR(1050)
country VARCHAR(550)
date_added VARCHAR(55)
release_year INT
rating VARCHAR(15)
duration VARCHAR(15)
listed_in VARCHAR(250)
description VARCHAR(550)
category VARCHAR(10)
Indexes
Triggers
netflix_titles
show_id TEXT
type TEXT
title TEXT
director TEXT
cast TEXT
country TEXT
date_added TEXT
release_year INT
rating TEXT
duration TEXT
listed_in TEXT
description TEXT

Conclusion:

The analysis of Netflix's content data provided valuable insights into various aspects of its catalog. By utilizing SQL queries, we were able to extract meaningful information across different dimensions such as content type (Movies vs TV Shows), genre distribution, actor participation, and country-specific content contributions.

Key Findings:

1. **Content Breakdown**: The query to count the number of Movies versus TV Shows helped in understanding the distribution of content types on Netflix. This insight can be useful for content strategy and marketing efforts, as it reveals the emphasis placed on either format.
2. **Popular Ratings**: By identifying the most common ratings for both movies and TV shows, we gain a deeper understanding of how content is perceived in terms of suitability for audiences. This can help Netflix tailor its content offerings to meet user preferences and regulatory standards.
3. **Country-Specific Content**: The analysis of content by country highlighted the regions that contribute the most to Netflix's global content library. This is useful for Netflix's localization strategy, helping to target regions with high demand or expand their catalog in underrepresented areas.
4. **Longest Movies & Documentaries**: Identifying the longest movie and categorizing documentaries further refined content classification. This insight can guide Netflix's marketing strategies by focusing on unique or standout content.
5. **TV Shows with More Seasons**: The query to list TV shows with more than 5 seasons reflects Netflix's investment in long-term content. It offers insight into the platform's approach to fostering engaging series and could inform decisions around renewing or investing in new TV shows.
6. **Actor Contributions**: By identifying the top actors in Indian movies, Netflix can evaluate which stars generate the most viewership and could be used to promote or advertise content, especially for regions like India where Bollywood content is a major draw.
7. **Categorizing Content**: The categorization of content based on keywords like 'kill' and 'violence' in the description enabled the identification of content that may appeal to different audiences. Such data is crucial for content moderation, compliance, and personalization strategies.
8. **Automating Content Categorization**: The implementation of SQL triggers, such as categorizing content automatically based on specific keywords, improves the efficiency of managing the Netflix catalog. It ensures that content is appropriately labeled and can be updated in real-time as new content is added.

Recommendations:

- **Content Strategy**: Continue to analyze and adjust content offerings based on regional popularity, viewer ratings, and genre preferences to cater to diverse global audiences.
- **Focus on Long-term Shows**: Given the success of long-running TV shows, it might be beneficial for Netflix to continue investing in multi-season series to retain subscribers and foster long-term engagement.
- **Actor and Genre Targeting**: By leveraging insights into popular actors and genres, Netflix can refine its content promotion strategies and potentially sign more high-profile actors to boost viewership.
- **Further Automation**: The use of SQL triggers and stored procedures for automated categorization and updates is a step towards streamlining operations. Expanding this approach to other aspects of content management (such as tagging or metadata management) could enhance Netflix's agility in handling its large catalog.

Overall, the analysis demonstrates the power of SQL in extracting actionable business insights from large datasets, and these insights can be directly applied to strategic decision-making processes at Netflix. With continuous refinement and deeper exploration of data, Netflix can further optimize its content offerings and remain a leader in the streaming industry.