Netflix

SQL Data Analysis

Business Report

C Jyoshitha

9-25-2024

Executive Summary

This report offers a comprehensive look into the Netflix dataset, shedding light on the **content distribution** across different types, such as movies and TV shows, and examining the diversity in geographical representation. The analysis shows that a **majority of Netflix's content** comprises movies, and a significant portion of this content is targeted at **mature audiences**. This suggests that Netflix is catering to a diverse yet predominantly adult demographic, likely driven by demand for mature themes and complex narratives.

A significant trend highlighted in the data is the **geographical dominance of U.S.-based productions**. The platform has historically focused on American content, which reflects both its origin and its largest market. However, there is also evidence of a **growing presence of Indian titles**, indicating Netflix's strategic focus on expanding its influence in non-Western markets, particularly in countries with large, content-hungry populations like India. This shift could be a response to Netflix's efforts to become more global and to appeal to diverse audience tastes.

Additionally, the dataset provides insight into **key contributors** behind Netflix's content, such as directors, actors, and production teams. Analysis of the **cast and crew** reveals trends in actor popularity and the types of projects frequently led by particular directors. Furthermore, the categorization of content through **descriptions and genres** helps identify viewer preferences and potential areas for future growth in Netflix's content strategy. Understanding these patterns is vital for predicting how Netflix might tailor its content to meet evolving viewer demands in the future.
"Data is a reflection of the world we live in—by interpreting it wisely, we discover more than just trends; we unveil stories that shape tomorrow."

Table of Contents

1. Introduction

1.1 Origin of the Dataset

The dataset used for this analysis, titled **netflix_titles.csv**, is sourced from **Kaggle**, a renowned platform for data science projects and competitions. The dataset contains detailed information about shows and movies available on Netflix, including essential attributes such as:

- **Show ID**: A unique identifier for each title.
- **Type**: Whether the content is a movie or TV show.
- **Title**: The name of the content.
- **Director**: The director(s) responsible for the content.
- **Cast**: The actors featured in the title.
- **Country**: The primary country where the content was produced.
- **Date Added**: When the content was made available on Netflix.
- **Release Year**: The year the content was originally released.
- **Rating**: The content's rating (e.g., TV-MA, PG-13).
- **Duration**: The length of the movie or number of seasons for TV shows.
- **Listed Genres**: The genres the content is classified under.
- **Description**: A brief synopsis of the content.

The dataset serves as a comprehensive source of information, providing a clear picture of Netflix's vast library, the creators behind it, and its global reach.

1.2 Essentials of the Netflix Dataset

The Netflix dataset is pivotal for analyzing the platform's **content landscape**, offering insights into the **types of content** available and their distribution across different **regions, genres, and ratings**. By leveraging the detailed attributes in this dataset, the analysis seeks to uncover patterns in:

- **Viewer preferences**: Identifying popular types of content and ratings favored by different demographics.
- **Geographical representation**: Understanding which countries dominate Netflix's library and identifying emerging markets.
- **Genre and content trends**: Assessing how Netflix curates its content across various genres and how these have evolved over time.

This analysis is particularly useful for content strategy, enabling Netflix to tailor its offerings according to **regional trends** and shifting viewer interests. Additionally, the dataset provides a foundation for examining **director participation, actor prominence**, and content categorization based on themes like violence or family-friendly viewing, helping to understand how **content diversity** aligns with audience expectations.

2. Data Analysis

2.1 Data Loading

The dataset was initially loaded into a Pandas DataFrame and written into an SQLite database, database_netflix.db, facilitating efficient data querying and manipulation.

2.2 SQL Queries and Insights

2.2.1 Content Type Distribution

Query: Count of content types (Movies and TV Shows)

Sql <code>

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

Results:

- Movies: 5880
- TV Shows: 2585

2.2.2 Most Frequent Ratings by Content Type

Query: Find the most frequent rating for each content type

Sql <code>

```
WITH RatingCounts AS (
SELECT type, rating, COUNT(*) AS rating_count
FROM netflix_titles
GROUP BY type, rating
),
RankedRatings AS (
SELECT type, rating, rating_count,
RANK() OVER (PARTITION BY type ORDER BY rating_count DESC) AS rank
FROM RatingCounts
)
SELECT type, rating AS most_frequent_rating FROM RankedRatings WHERE rank = 1;
```

Results:

- Movies: Most frequent rating is "TV-MA".
- TV Shows: Most frequent rating is "TV-MA".

2.2.3 Content Released in 2020

Query: Retrieve content released in 2020

Sql <code>

```
SELECT * FROM netflix_titles WHERE release_year = 2020;
```

Results: Details of all titles released in 2020 for trend analysis.

2.2.4 Country Representation

Query: Count of titles by country

Sql <code>

```
SELECT country, COUNT(*) AS total_content FROM netflix_titles WHERE country IS NOT NULL GROUP BY country ORDER BY total_content DESC LIMIT 5;
```

Results:

- United States: Highest number of titles.
- India: Notable presence with a growing catalog.

2.2.5 Duration of Movies

Query: Retrieve Movies sorted by duration

Sql <code>

```
SELECT * FROM netflix_titles WHERE type = 'Movie' ORDER BY CAST(substr(duration, 1, instr(duration, ' ') - 1) AS INTEGER) DESC;
```

Results: Insights into movie lengths and trends.

2.2.6 Recent Content (Last 5 Years)

Query: Content added in the last 5 years

Sql <code>

```
SELECT * FROM netflix_titles WHERE date_added >= DATE('now', '-5 years');
```

Results: Analysis of recent viewer interests.

2.3 Detailed Insights

2.3.1 Director Analysis

Query: Count of titles directed by 'Rajiv Chilaka'

Sql <code>

```
SELECT * FROM netflix_titles WHERE director IS NULL;
```

Results: Titles without specified directors.

2.3.2 Actor Popularity in Indian Movies

Query: Top actors in Indian movies

Python<code>

```
top_actors_count = top_actors.value_counts().head(10);
```

Results: Top actors in Indian movies for audience preference analysis.

2.3.3 Content Categorization by Description

Query: Categorizing content based on descriptions

Sql <code>

```
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_titles
) AS categorized_content GROUP BY category, type ORDER BY type;
```

Results:

Category	Type	Content Count
Bad	Movie	251
Good	Movie	5880
Bad	TV Show	91
Good	TV Show	2585

3. Conclusion

The analysis of the Netflix dataset has revealed key patterns in the distribution of content across the platform. One of the standout findings is that the majority of Netflix's content is made up of **movies**, with a significant percentage of the library sourced from the **United States**. However, **Indian titles** have shown notable growth, reflecting Netflix's increasing focus on expanding its global content reach. This trend indicates a shift towards greater **geographical diversity**, offering viewers a wider range of choices across different cultural backgrounds and languages.

Another important observation is that the **content rating system** leans heavily toward mature audiences, with the majority of titles categorized under ratings like **TV-MA** and **R**. However, Netflix has also curated a substantial amount of content with a **"Good" rating**, which suggests that the platform prioritizes content quality and curates titles that align with **viewer preferences**. The analysis of content descriptions further reinforces Netflix's strategy of offering a broad mix of genres to cater to diverse tastes, from family-oriented films to action-packed blockbusters.

Projections

Future analyses could delve deeper into **viewer ratings** and **reviews** to provide a more granular understanding of how specific titles perform after release. Additionally, examining the impact of **release timing**—whether certain times of the year (e.g., holidays) see higher engagement with specific genres—could offer valuable insights into **viewership trends**. This would help Netflix optimize its release schedules and further refine its content curation strategy to ensure high engagement levels across different regions and demographics.

4. References

- **Kaggle:** Netflix Dataset - https://www.kaggle.com/datasets/shubhendra/netflix-shows-dataset
- **Project**: https://colab.research.google.com/drive/1scoU_AjQ2V8c1uCAkQudPC-RelTyEQCE?usp=sharing
- **Github**: https://github.com/Jyoshitha04/SQL
- **LinkdIn**: https://www.linkedin.com/in/c-jyoshitha