****

**KAVYA KALYAN**

**Introduction**

In recent years, streaming services have transformed the landscape of entertainment consumption, with Netflix emerging as a leading platform in this digital revolution. As of 2023, Netflix boasts a vast library of films and television series that spans diverse genres, languages, and cultures, catering to a global audience of millions. The ability to analyze this extensive content database provides valuable insights into viewing trends, content diversity, and audience preferences.

This project aims to conduct a comprehensive analysis of Netflix's content using SQL, focusing on several key objectives. First, we will explore the distribution of content types, distinguishing between movies and TV shows, to understand their relative prevalence. Next, we will identify the most common ratings assigned to these content types, shedding light on audience perceptions and content quality.

Further, we will analyze the content based on release years, countries of origin, and durations, revealing trends in production and geographic diversity. Lastly, we will examine genre preferences and the contributions of various directors, offering insights into the creative influences that shape Netflix's offerings.

By leveraging SQL for data analysis, this project will not only enhance our understanding of Netflix's content landscape but also highlight the potential of data-driven approaches in the entertainment industry.

**Objectives**

* Analyze the distribution of content types (movies vs TV shows).
* Identify the most common ratings for movies and TV shows.
* List and analyze content based on release years, countries, and durations.
* Analyze the genre preferences and director contributions.

**About The Dataset**

Netflix stands as a leading force in the realm of media and video streaming. With a staggering array of over 8000 movies and TV shows accessible on their platforms as of mid-2021, their global subscriber count exceeds over 200 million. This tabulated dataset comprehensively catalogs all offerings on Netflix, including vital details such as cast, directors, ratings, release year, duration and more.

**Dataset Overview**

The Netflix titles dataset is a comprehensive compilation of movies and TV shows available on Netflix, covering various aspects such as the title type, director, cast, country of production, release year, rating, duration, genres(listed\_in), and a brief description. This dataset is instrumental for analyzing trends in Netflix content, understanding genre popularity, and examining the distribution of content across different regions and time periods. The key attributes of the column are given below.

* **Show\_id:** Unique ID for every Movie / TV show.
* **Type:** Identifier - A Movie or TV Show.
* **Title:** Title of the Movie / TV show.
* **Director:** Director of the Movie.
* **Cast:** Actors involved in the movie/show.
* **Country:** The country where the movie/show was produced.
* **Date\_added:** Date it was added on Netflix.
* **Release\_year:** Actual Release year of the movie/show.
* **Rating:** Maturity Rating of the movie/show.
* **Duration:** Total Duration - in minutes or number of seasons.
* **Listed\_in:** Genre of contents.
* **Description:** The summary description.

**Methodology**

**1) Data Collection:**

The data for this project was collected from Kaggle, a reputable platform known for hosting a wide variety of datasets across different domains. The Netflix dataset is a comprehensive compilation of movies and TV shows available on Netflix, covering various access such as the title type, director, cast country of production, release year, rating, duration, genres (listed in), and a brief description. This dataset is instrumental for analyzing trends in Netflix content, understanding genre popularity, and examining content distribution across different regions and periods.

**2) Data Cleaning and Preprocessing:**

The process conducts the data cleaning tasks to address any missing values, duplicate entries, and outliers ensuring the dataset’s integrity and readiness for analysis. This step is crucial for maintaining accuracy in subsequent analysis.

**3) Exploratory Data Analysis (EDA):**

Exploratory Data Analysis (EDA) of the Netflix dataset involves a systematic examination of the content library to uncover patterns, trends, and anomalies within the data. Initially, we analyze the distribution of content types—movies and TV shows—to understand their prevalence. Following this, we assess the ratings assigned to various titles, providing insight into audience reception and content quality. We also explore the release years and countries of origin, which helps identify trends in global content production and availability. Additionally, genre analysis reveals viewer preferences, while an examination of directors highlights influential figures in Netflix's catalog. Overall, EDA serves as a foundational step in understanding the dataset's characteristics and guiding further analyses.

**4) Data Visualizations:**

Data visualization helps to communicate trends and patterns in a dataset effectively.

**Data Analysis:**

**Count the Number of Movies vs TV Shows:**

SELECT type, COUNT(\*) as total FROM netflix\_titles GROUP BY type;

|  |  |
| --- | --- |
| Type | Count |
| Movie | 6131 |
| TV Show | 2676 |

The dataset contains 6131 movies and 2676 TV shows, indicating a higher prevalence of movies than TV shows.

**Find the Most Common Rating for Movies and TV Shows:**

-- Find the most common rating for Movie

SELECT rating, COUNT(\*) AS rating\_count

FROM netflix\_titles

WHERE type = 'Movie'

GROUP BY rating ORDER BY rating\_count DESC

LIMIT 1;

-- Find the most common rating for TV Shows

SELECT rating, COUNT(\*) AS rating\_count

FROM netflix\_titles

WHERE type = 'TV Show'

GROUP BY rating

ORDER BY rating\_count DESC

LIMIT 1;

|  |  |
| --- | --- |
| Type | Count |
| Movie | TV-MA |
| TV Show | TV-MA |

The most common rating for both movies and TV shows in your data is TV-MA.

**Find out the top 5 countries with the most content on Netflix:**

SELECT TRIM(SUBSTRING\_INDEX(SUBSTRING\_INDEX(country, ',', n.n), ',', -1)) AS new\_country, COUNT(show\_id) AS total\_count

FROM netflix\_titles

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 CHAR\_LENGTH(country) - CHAR\_LENGTH(REPLACE(country, ',', '')) >= n.n - 1

GROUP BY new\_country

ORDER BY total\_count DESC

LIMIT 5;

|  |  |
| --- | --- |
| Country | Total |
| United States | 3649 |
| India | 972 |
| United Kingdom | 419 |
| Japan | 245 |
| South Korea | 199 |

From the given table, we can see the total content available on Netflix for various countries. The United States has the highest content total at 3649 titles, followed by India with 972 titles. The United Kingdom has 419 titles, while Japan has 245. It appears that the United States significantly leads in the number of available titles, indicating its dominance in content production and availability on the platform. Additionally, the presence of multiple entries for the United States suggests that the data may need clarification or consolidation to accurately reflect the total.

**Most popular Genres on Netflix:**

SELECT TRIM(SUBSTRING\_INDEX(SUBSTRING\_INDEX(listed\_in, ',', n.n), ',', -1)) AS genre, COUNT(\*) AS total\_count

FROM netflix\_titles

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 CHAR\_LENGTH(listed\_in) - CHAR\_LENGTH(REPLACE(listed\_in, ',', '')) >= n.n - 1

GROUP BY genre

ORDER BY total\_count DESC

LIMIT 5;

|  |  |
| --- | --- |
| Genre | Count |
| International Movies | 2752 |
| Dramas | 2427 |
| Comedies | 1674 |
| International TV Shows | 1351 |
| Documentaries | 869 |

The data shows a strong preference for international content, with "International Movies" (2752) and "International TV Shows" (1351) leading in the count, suggesting high demand for globally diverse programming. Dramas (2427) and Comedies (1674) follow closely, indicating these genres are also widely favoured. Documentaries, though lower in count (869), cater to a niche audience, reflecting a specialized but engaged viewer base. Overall, the genre distribution highlights a broad selection catering to varied interests, emphasizing international and dramatic content.

**Which Year had the most releases?**

SELECT release\_year, COUNT(\*) as Count from netflix\_titles GROUP BY release\_year ORDER BY Count DESC LIMIT 5;

|  |  |
| --- | --- |
| Year | Count |
| 2018 | 1147 |
| 2017 | 1032 |
| 2019 | 1030 |
| 2020 | 953 |
| 2016 | 902 |

The year with the highest number of releases was 2018, which featured a total of 1147 titles. This significant figure indicates a peak in content production during that year, suggesting that Netflix was actively expanding its library. In comparison, 2017 and 2019 followed closely with 1032 and 1030 releases, respectively, while 2020 and 2016 had fewer releases at 953 and 902. This trend may reflect the evolving strategy of Netflix in content acquisition and production during those years.

**To identify the Longest Movie:**

SELECT title, duration

FROM netflix\_titles

WHERE type = 'Movie'

ORDER BY CAST(SUBSTRING\_INDEX(duration, ' ', 1) AS UNSIGNED) DESC

LIMIT 1;

The longest movie is 'InuYasha the Movie 2: The Castle Beyond the Looking Glass' with a duration of 99 min minutes.

**Find the most prolific directors:**

SELECT director, COUNT(\*) AS title\_count

FROM netflix\_titles

WHERE director IS NOT NULL AND director != 'Unspecified'

GROUP BY director

ORDER BY title\_count DESC

LIMIT 5;

|  |  |
| --- | --- |
| Director | Count |
| Rajiv Chilaka | 22 |
| Raúl Campos | 18 |
| Marcus Raboy | 16 |
| Suhas Kadav | 16 |
| Jay Karas | 15 |

The data shows Rajiv Chilaka as the most prolific director with 22 titles, suggesting strong demand for his work, likely in animation or children’s programming. He is followed by Raúl Campos (18), Marcus Raboy (16), Suhas Kadav (16), and Jay Karas (15), all of whom have directed a similar number of titles, indicating steady productivity. These directors may work in popular genres with recurrent content, such as comedy, family, or documentary programming. Their high output reflects both demand and consistency, highlighting their significance and sustained relevance in the industry.

**Directors Creating Both Horror and Comedy Movies:**

SELECT director

FROM netflix\_titles

WHERE director IS NOT NULL

AND type = 'Movie'

AND (listed\_in LIKE '%Horror%' OR listed\_in LIKE '%Comedy%')

GROUP BY director

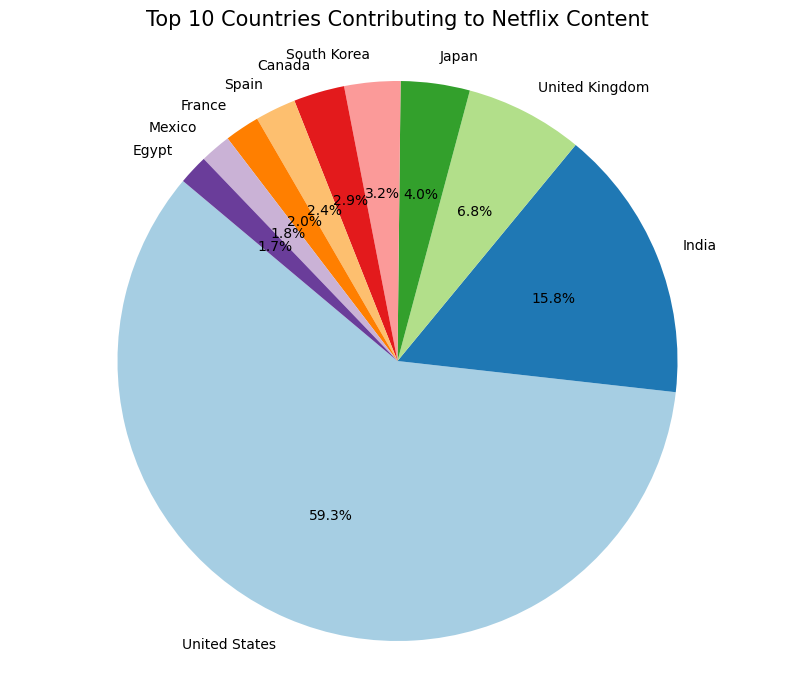
HAVING SUM(listed\_in LIKE '%Horror%') > 0

AND SUM(listed\_in LIKE '%Comedy%') > 0;

Identified 2 directors who have created both horror and comedy movies, with Steve Brill and Oz Rodriguez noted for their substantial contributions to both genres.

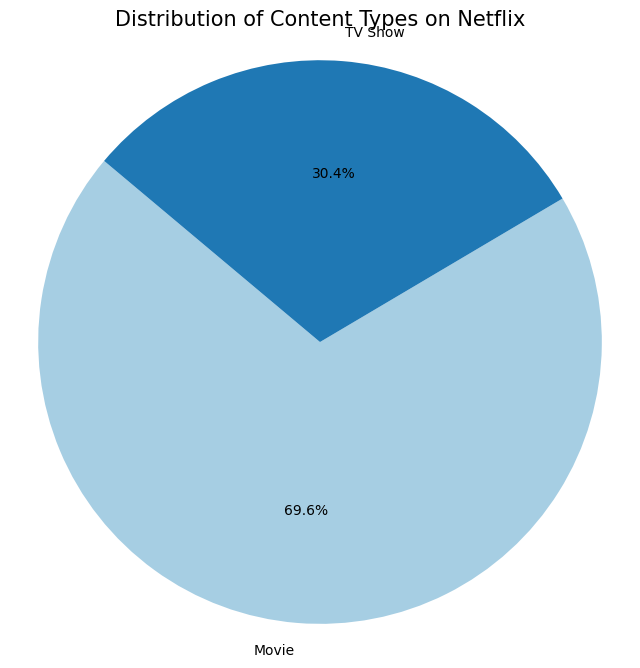
**DATA VISUALISATIONS**

* **Top 10 countries as contributors to Netflix:**

****

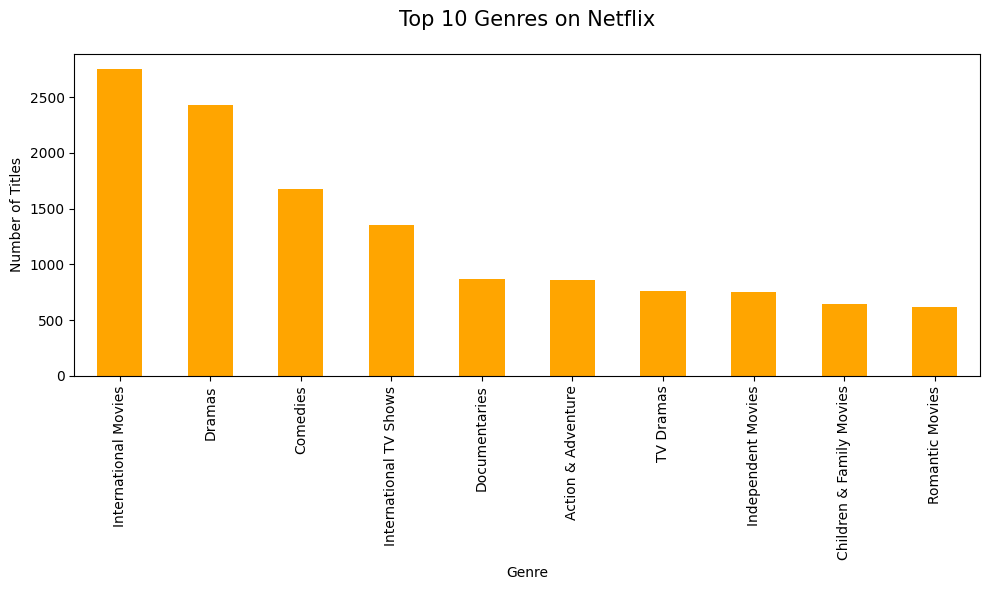
The United States has the highest contributor with 59.3% of videos on Netflix. India surprisingly ranks second with 15.8%, followed by the United Kingdom with 6.8%.

* **The types and their total numbers:**

****

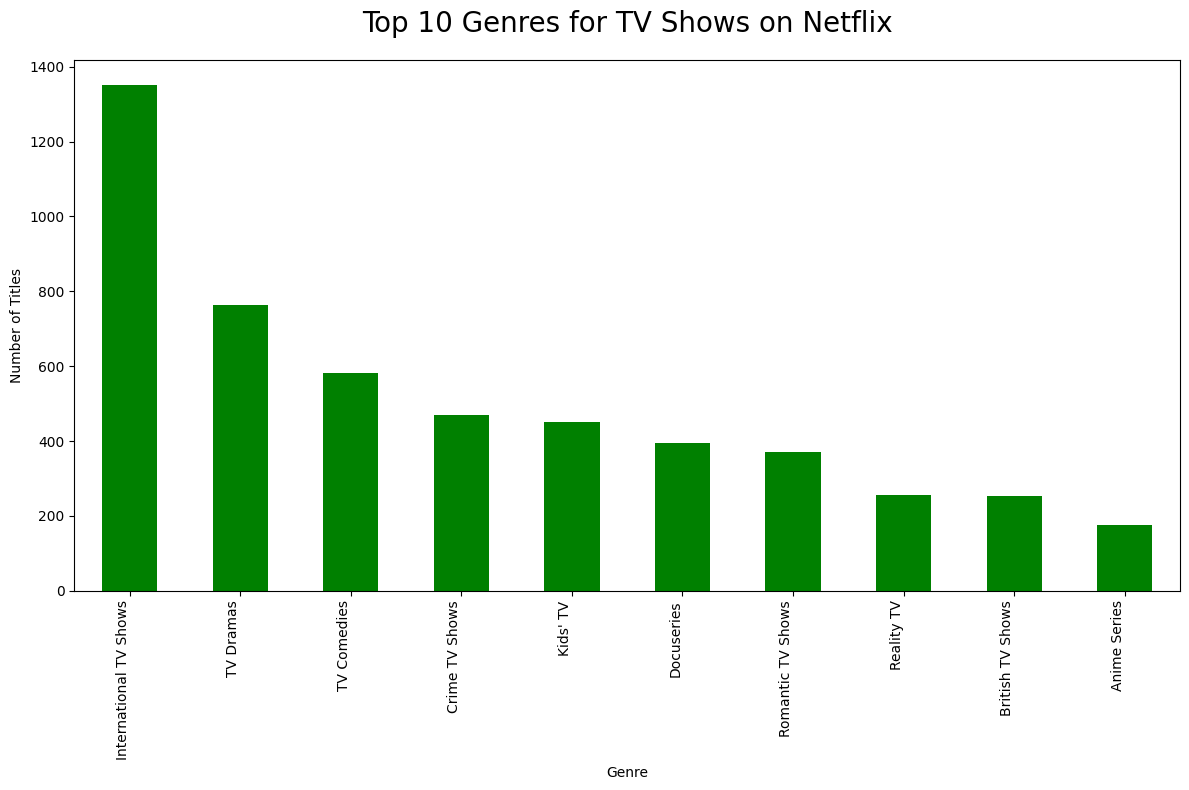
We can see that movies have a viewer rate of 69.6% which is much higher than that of TV shows which is 30.4%. Netflix viewers explore more Movies than TV Shows overall.

* **To count the top 10 number of Genre:**

****

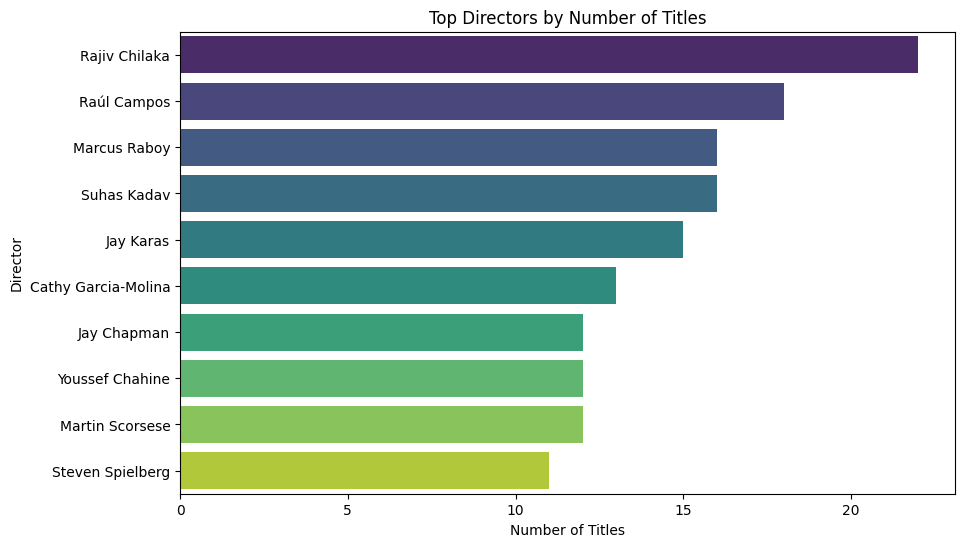
International Movies (2752) have the highest genre rate of movies users explore on Netflix, followed by Dramas (2427) and Comedies (1674).

* **To count the number of each genre of all the TV shows:**

****

International TV Shows have the highest genre rate of TV shows users explore on Netflix, followed by TV Dramas and TV Comedies.

* **To find the top 10 Directors:**



The graph highlights Rajiv Chilaka as the most prolific director, with over 20 titles, followed by Raúl Campos and Marcus Raboy, who also have a significant number of titles. Suhas Kadav and Jay Karas are close behind, indicating that these top five directors are particularly productive. Notable names like Steven Spielberg and Martin Scorsese appear lower on the list, suggesting that while they are well-known, they have fewer titles in this dataset compared to the top directors.

**Conclusion**

This project provides a comprehensive analysis of Netflix's content catalog, offering valuable insights into the platform's composition, genre preferences, audience ratings, and the creative contributors who shape its offerings. The analysis reveals that movies make up the majority of Netflix’s content, with the United States being the primary contributor to the platform's library, followed by India and the United Kingdom. The most common content rating is TV-MA, indicating a strong emphasis on mature content. Genre analysis highlights a clear preference for international content, dramas, and comedies, with international movies and TV shows particularly popular among audiences, reflecting Netflix's focus on global appeal.

Trends in release years show a peak in content production in 2018, indicating an aggressive content expansion strategy during that period. A closer examination of directors shows that Rajiv Chilaka leads in content contribution, followed by Raúl Campos, Marcus Raboy, and others who frequently produce content for the platform. The presence of directors skilled in multiple genres, such as horror and comedy, underscores Netflix’s dedication to diverse, cross-genre offerings.

Through data cleaning, EDA, and visualizations, this project demonstrates how SQL can effectively extract and analyze trends within a vast media library. These insights can guide future strategies for content acquisition, development, and regional focus, ultimately enhancing Netflix’s ability to cater to its diverse and growing global audience. The project illustrates the potential of data-driven approaches in optimizing content libraries and understanding viewer preferences in the streaming industry.