# Netflix Data: Cleaning, Analysis and Visualization

## About the Dataset:

Netflix is a popular streaming service that offers a vast catalog of movies, TV shows, and original content. This dataset is a cleaned version of the original version, which can be found here. The data consists of content added to Netflix from 2008 to 2021. The oldest content is as old as 192,5 and the newest as 2021. This dataset will be cleaned and visualized with Python code in a Jupyter notebook. The purpose of this dataset is to test my data cleaning and visualization skills. The cleaned data can be found below
and the Tableau dashboard can be found here.

## Data Cleaning (We are going to):

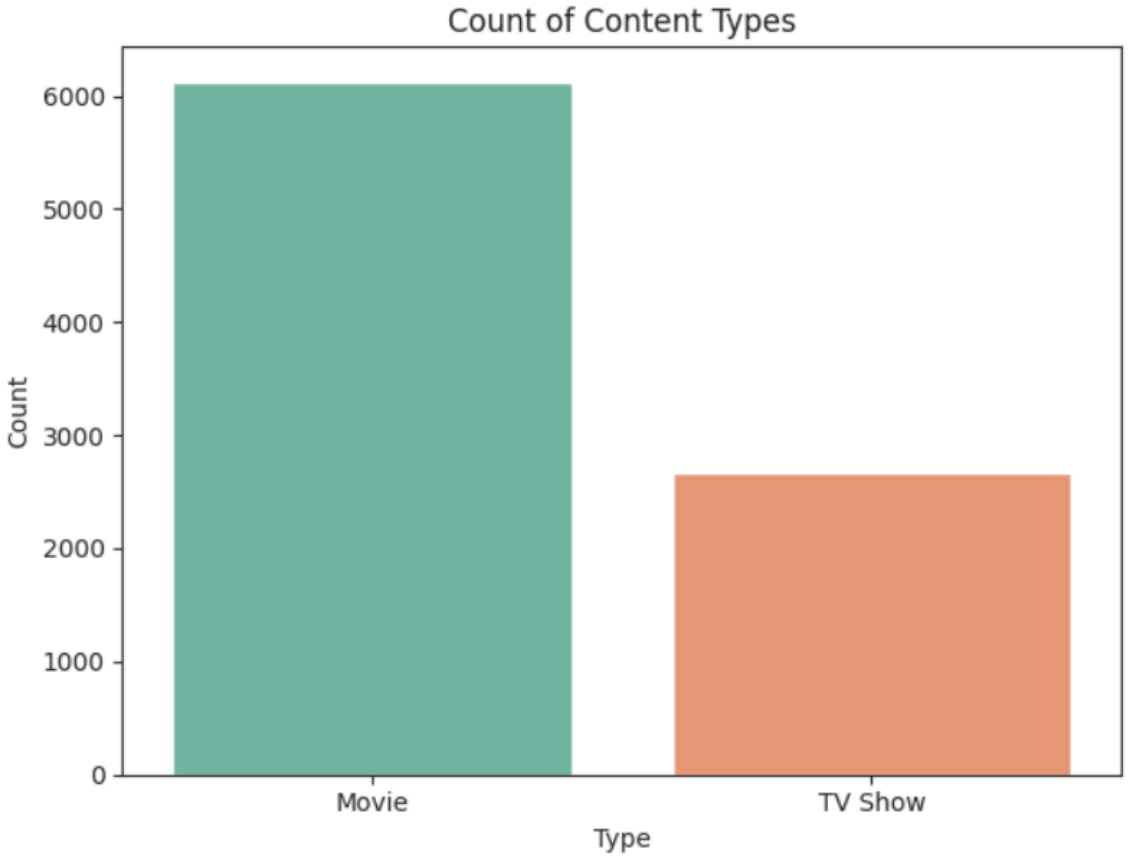
1. Treat the Nulls
2. Treat the duplicates
3. Populate missing rows
4. Drop unneeded columns
5. Split column

## NetflixData:

Cleaning, Analysis, and Visualization (BeginnerMLProject) This project involves loading, cleaning, analyzing, and visualizing data from a Netflix dataset. We'll use Python libraries like Pandas, Matplotlib, and Seaborn to work through the project. The goal is to explore the dataset, derive insights, and prepare for potential machine learning tasks

# 1. Content Type Distribution

### 1.1 Count of Content Types




**This picture displays** the count of Movies and TV Shows available on Netflix.

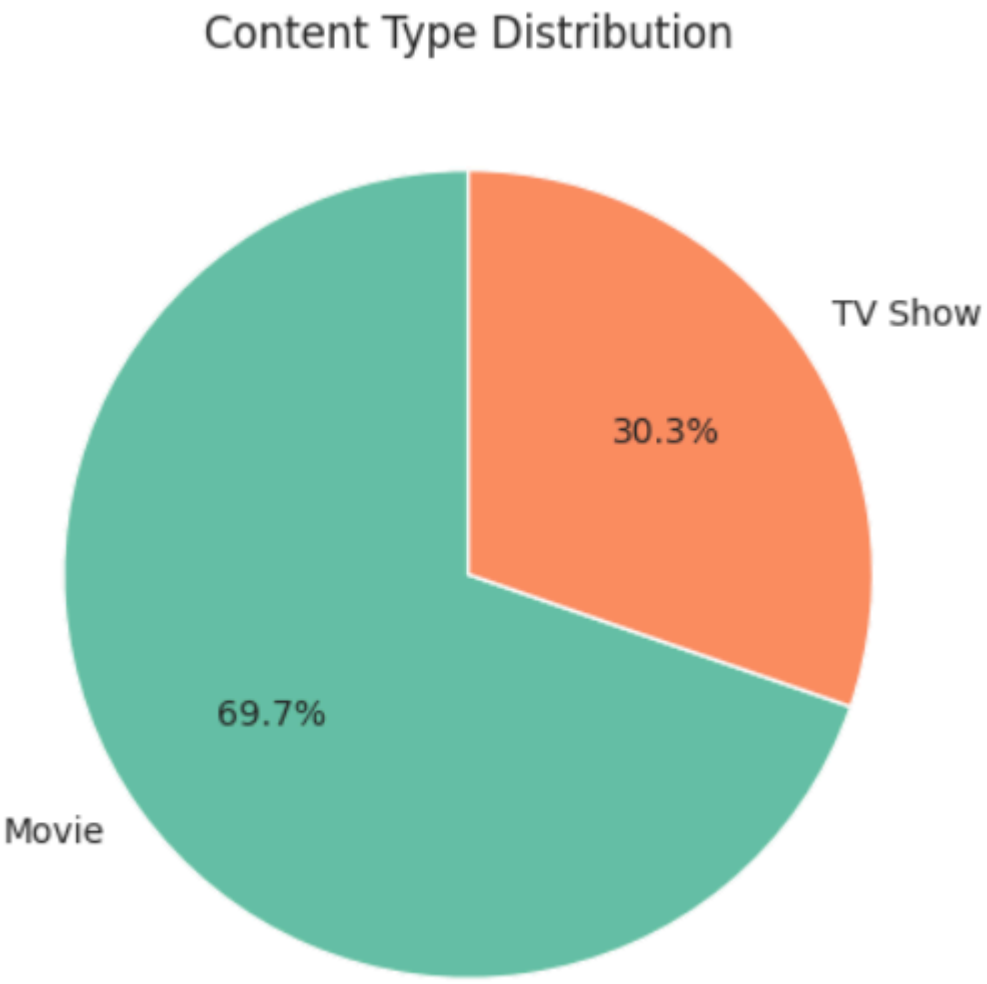
**Observations:**

- There are over 6000 movies, while the number of TV shows is just under 3000.
- The gap between the two types is significant, nearly 2:1.

**Insight:**

The platform prioritizes a diverse movie selection, though TV shows still form a substantial part of the offering. This reflects a strategy to cater to different viewer preferences, with more emphasis on one-off content (movies).

**1.2 Content Type Distribution**

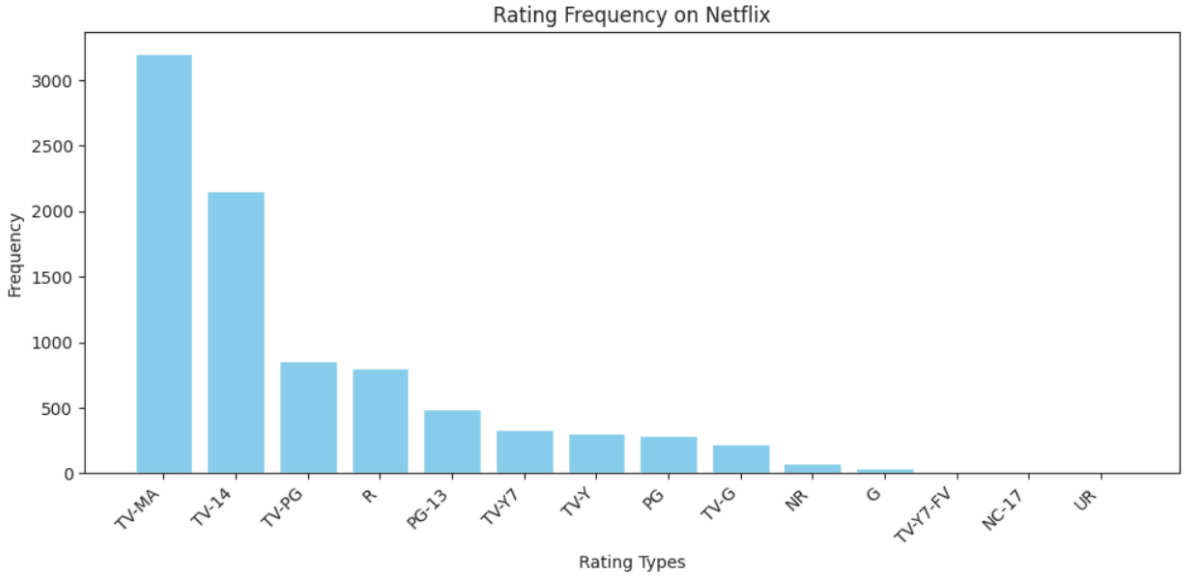



**This picture displays** the overall proportion of Movies and TV Shows available on Netflix.

**Observations:**
- Movies dominate the content library, making up approximately 69.7%.
- TV Shows account for the remaining 30.3%.

**Insight:**
Netflix offers more movies than TV shows, suggesting its platform still maintains a movie-heavy catalog, possibly due to licensing flexibility or production volume.

## 2. Rating Frequency on Netflix




**This picture displays** the frequency of all content ratings found on Netflix.

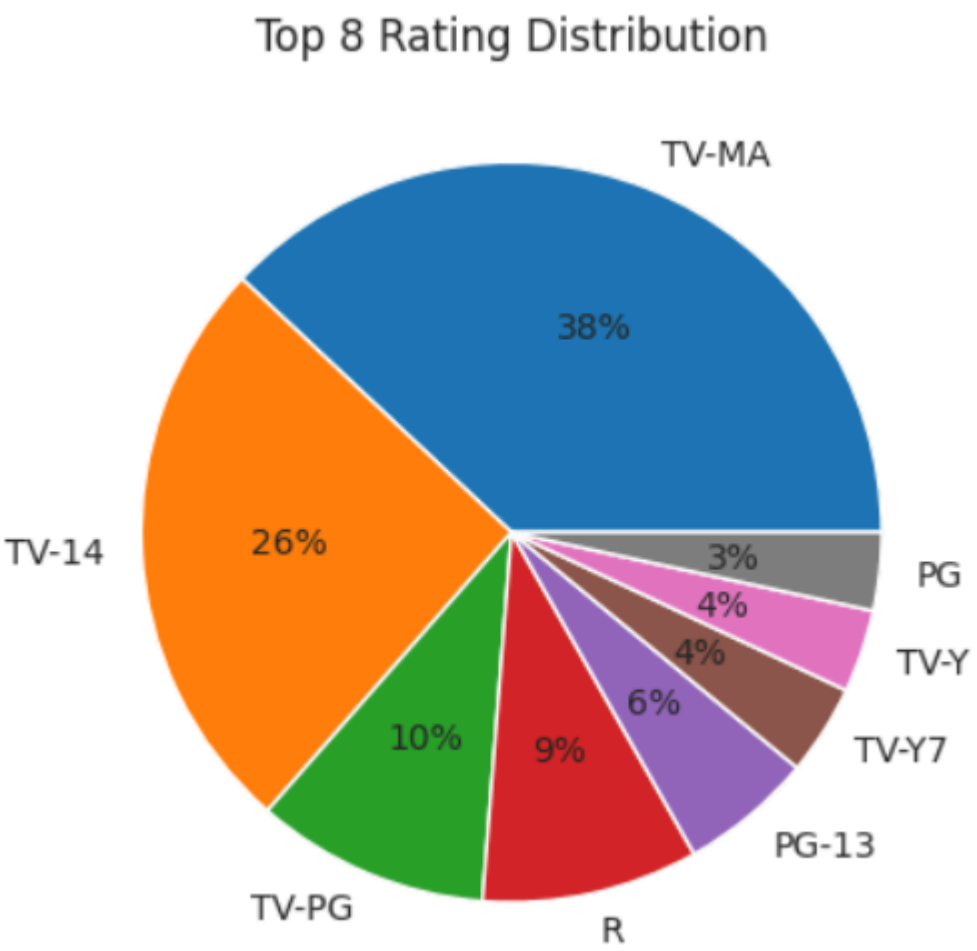
**Observations:**

- TV-MA and TV-14 are the most frequent, with over 3000 and 2100 instances respectively.
- Ratings like TV-PG and R also have noticeable representation, each with around 800–900 entries.
- Ratings such as G, UR, NC-17, and TV-Y7-FV appear extremely rarely.

**Insight:**

Most Netflix content skews toward mature and teenage audiences, while content rated for general or young audiences is comparatively rare, reflecting targeted demographics.

### 3. Top 8 Rating Distribution




**This picture displays** the proportion of different content ratings in the top 8 categories on Netflix.

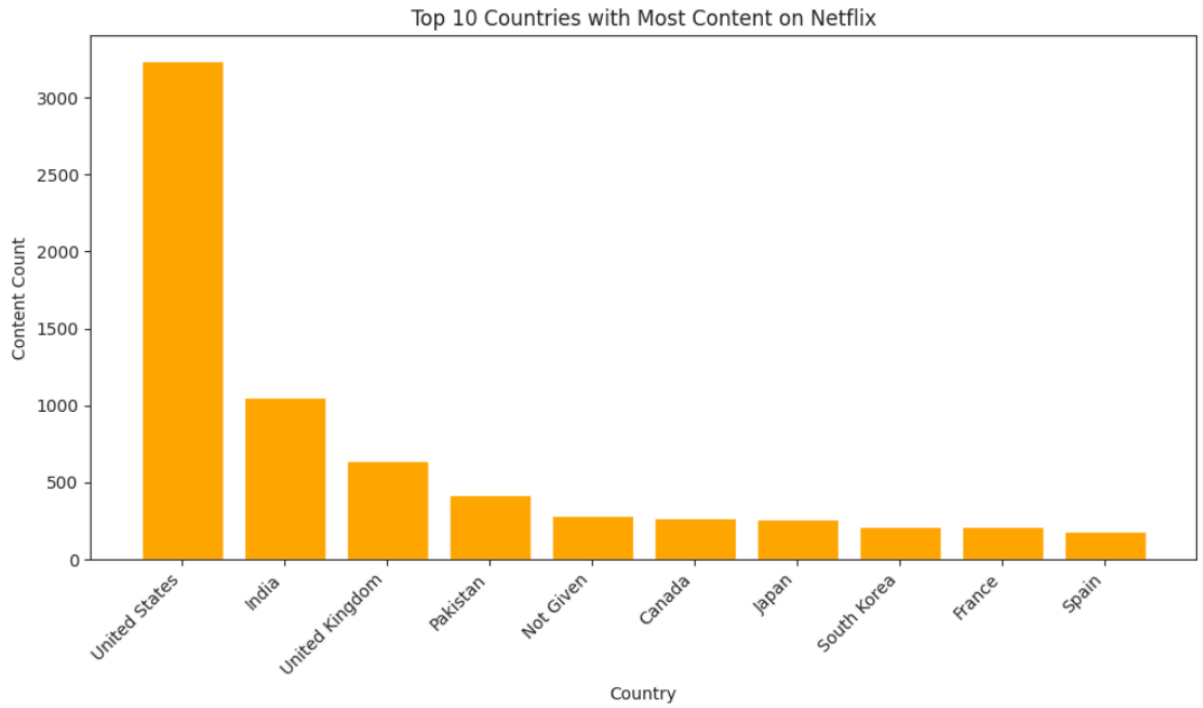
**Observations:**

- TV-MA dominates with 38%, followed by TV-14 at 26%.
- TV-PG and R ratings account for 10% and 9% respectively.
- The remaining ratings (PG-13, TV-Y7, TV-Y, PG) each make up a small percentage (3–6%).

**Insight:**

A significant portion of Netflix content is intended for mature audiences (TV-MA and TV-14), indicating a focus on adult or teen viewers. Content appropriate for younger children makes up a relatively smaller share.

## 4. Top 10 Countries with Most Content on Netflix




**This picture displays** the distribution of Netflix content count among the top 10 contributing countries.

**Observations:**

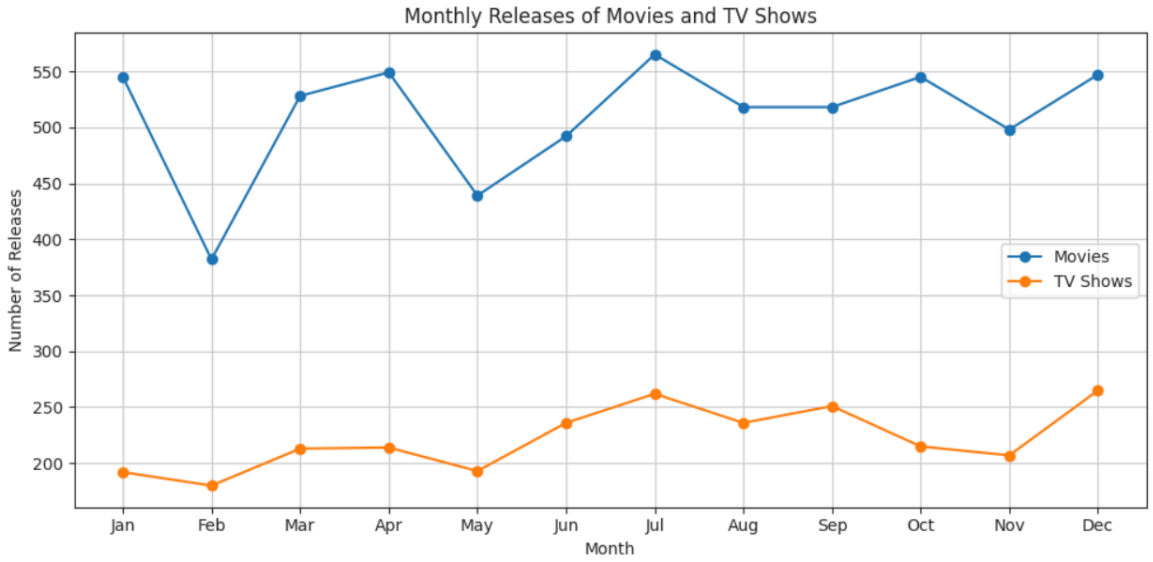
- The United States leads by a wide margin, contributing over 3200 pieces of content.
- India ranks second with just above 1000, followed by the United Kingdom and Pakistan.
- Other countries such as Canada, Japan, South Korea, France, and Spain contribute relatively fewer titles.
- A small portion of the data is labeled "Not Given," indicating missing or unspecified country information.

**Insight:**

Netflix's content library is heavily dominated by the United States, indicating a strong presence of U.S.-based productions. This suggests regional biases in available titles and may reflect content licensing or production trends.

# Monthly and Yearly Trends

## 5.1 Monthly Releases of Movies and TV Shows




This picture displays the number of movie and TV show releases per month throughout a calendar year.

**Observations:**

- **Movies are released in higher volumes** every month compared to TV shows.

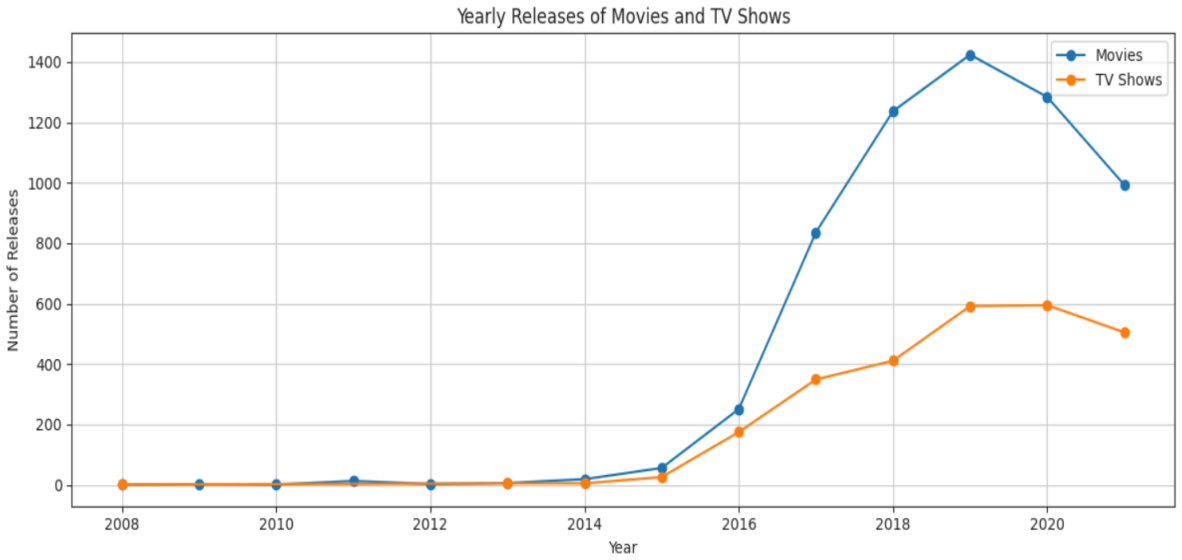
- There are noticeable spikes in movie releases during **January, April, July, and December**, possibly due to seasonal trends or holiday releases.

- TV show releases remain relatively steady with slight peaks in **June, July, September, and December**.

**Insight:**

Movie release cycles are more seasonal and strategic, while TV show releases follow a steadier distribution. December and July are hot months for both formats, likely aligned with school holidays and viewer demand.

### 5.2 Yearly Releases of Movies and TV Shows




This picture displays the yearly trend in the number of movie and TV show releases from 2008 to 2021.

**Observations:**

● Both movies and TV shows saw **rapid growth between 2015 and 2019**, with movies peaking in 2019 and TV shows in 2019–2020.

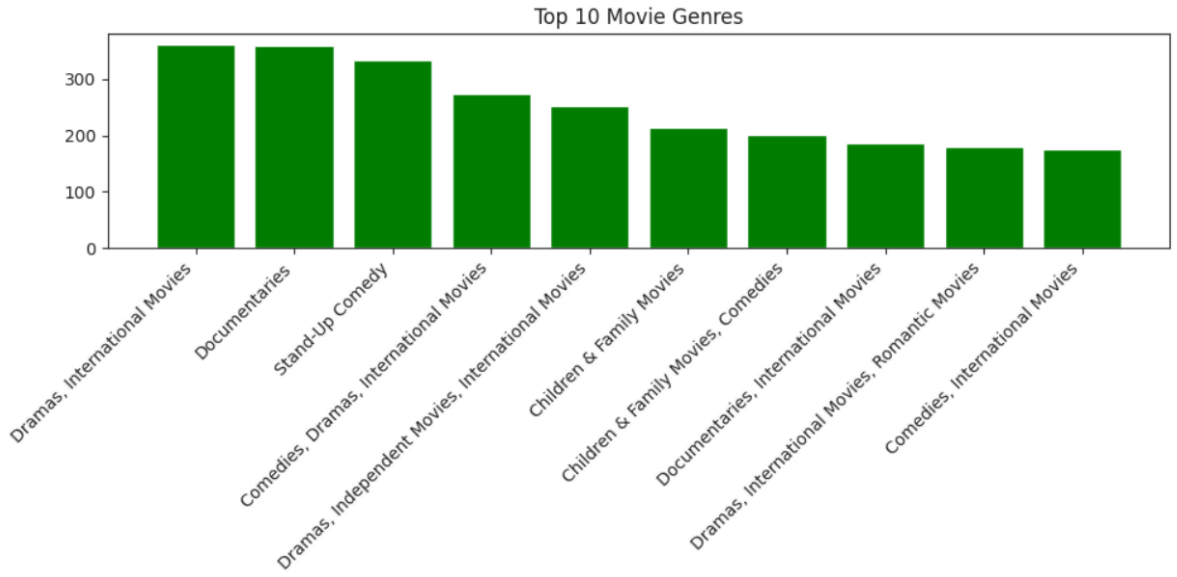
● After 2019, both categories experienced a **decline**, likely due to the **COVID-19 pandemic's impact on production**.

● Prior to 2015, growth was minimal, especially for TV shows.

**Insight:**

The streaming era (post-2015) drove a boom in content production, especially for movies. The pandemic introduced a temporary setback, but the data reflects how the content industry scaled rapidly in recent years.

**6. Distribution of Top 10 Movie Genres**




This picture displays the distribution of the top 10 movie genres based on the number of releases.

**Observations:**

● The most frequent genre is **“Dramas, International Movies”**, followed closely by **Documentaries** and **Stand-Up Comedy**.

● Many genres are combinations, such as **“Comedies, Dramas, International Movies”** and **“Children & Family Movies, Comedies”**, indicating genre blending is common.

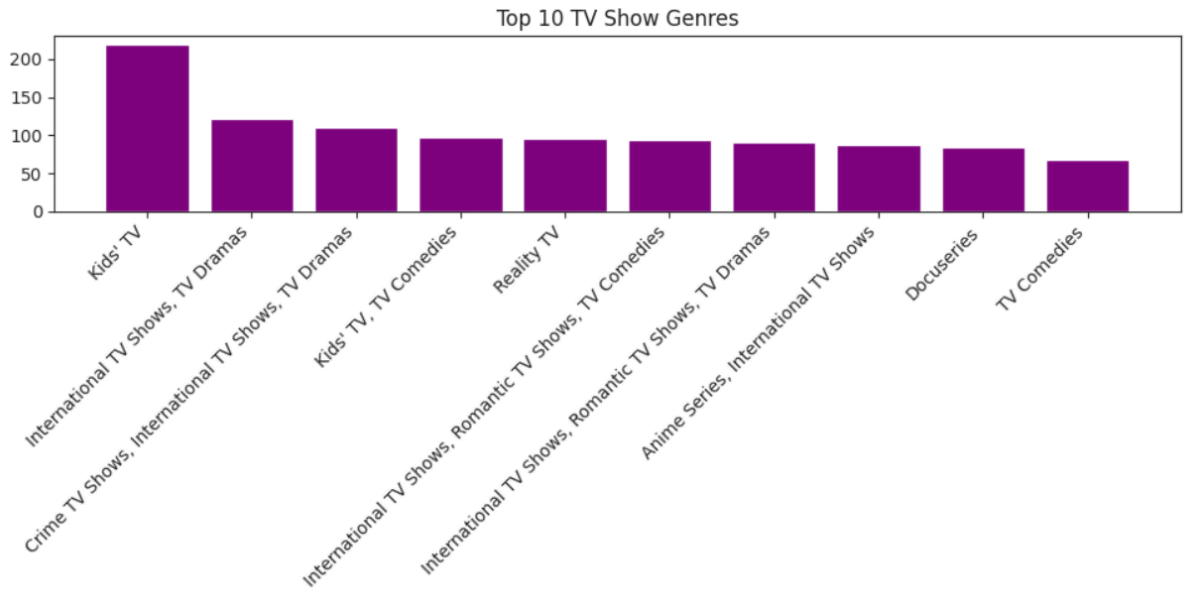
● Genres involving **International Movies** dominate the top 10, appearing in at least 6 out of 10 entries.

● The distribution shows a gradual decline in frequency from the top genre to the tenth.

**Insight:**

International and dramatic storytelling forms the core of movie content, with comedy and children’s content also being significant. Genre hybridity is a key trend in movie production.

**7. Distribution of Top 10 TV Show Genres**




This picture displays the distribution of the top 10 TV show genres based on the number of releases.

**Observations:**

● **Kids' TV** is by far the most widely released TV genre.

● International and regional hybrid genres such as **"International TV Shows, TV Dramas"** are frequent, indicating cross-border content popularity.

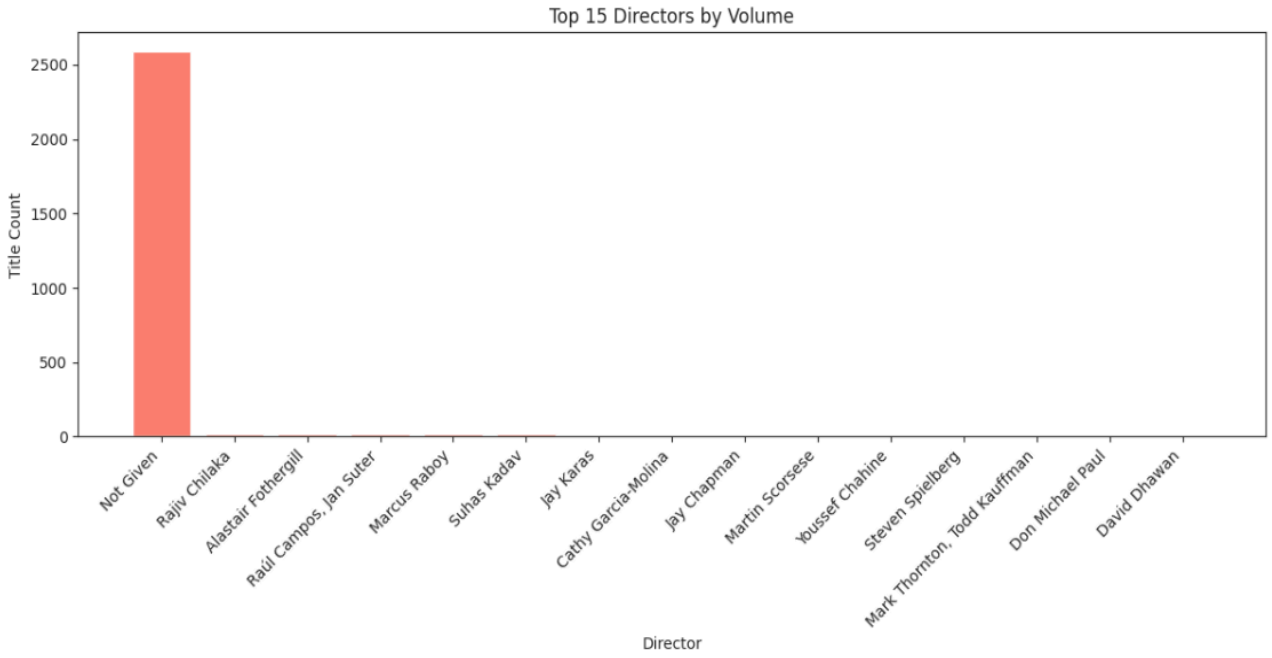
● Genres like **Reality TV**, **Anime**, and **Docuseries** show that TV content is more varied and niche compared to movies.

● The number of releases drops sharply after the top genre.

**Insight:**

TV show production focuses on specific audience segments like children and fans of reality or anime, with a strong lean toward international content and serialized formats.

**8. Top 15 Directors by Volume**




This picture displays the top 15 directors with the highest number of released titles, based on available metadata.

**Observations:**

● An overwhelming number of titles (2600) have **no director listed (“Not Given”)**, dominating the chart.

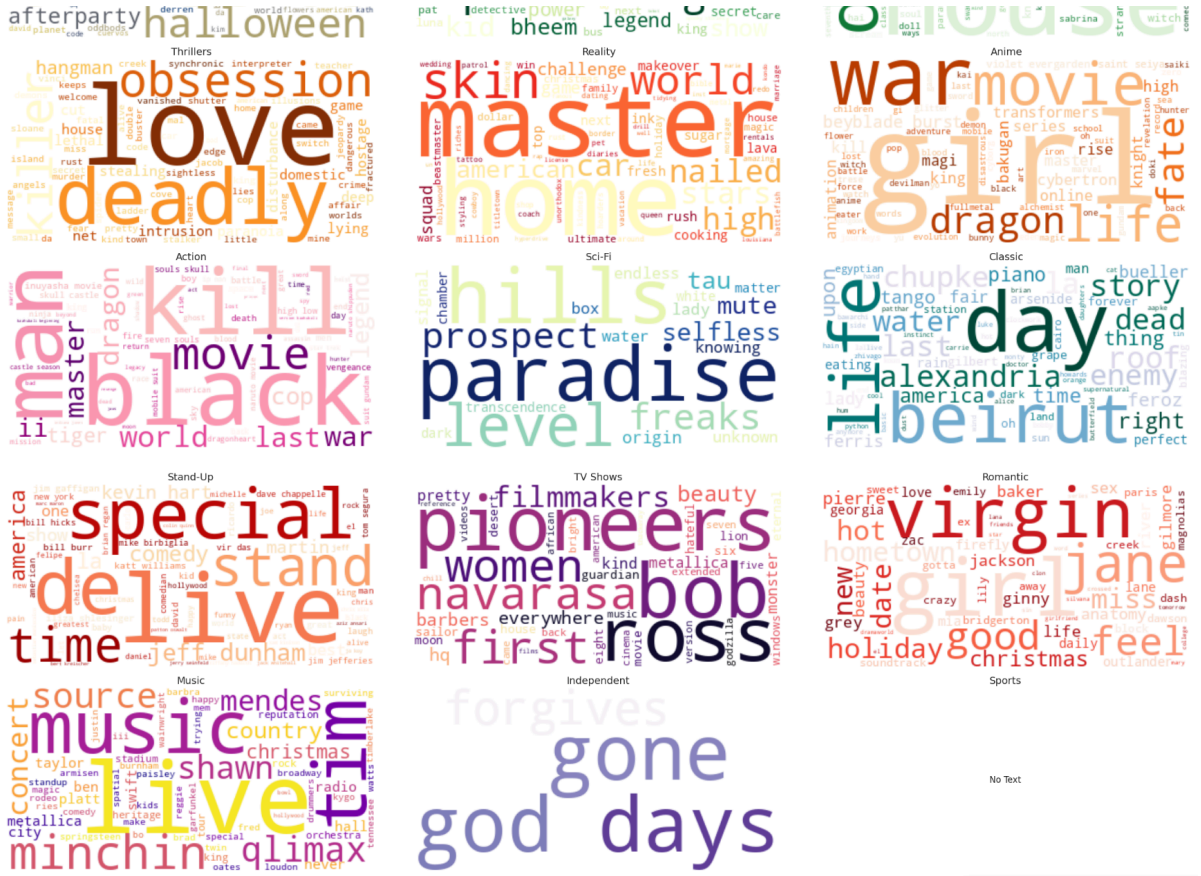
● Directors with the highest documented volumes include **Rajiv Chilaka**, **Alastair Fothergill**, and **Raúl Campos & Jan Suter**.

● Well-known names like **Martin Scorsese** and **Steven Spielberg** appear but with relatively lower volume counts.

**Insight:**

Director metadata is often missing, limiting analytics. However, where available, a mix of prolific children's content creators (e.g., Chilaka) and documentary or mainstream directors populate the top list.

**9. Word Clouds by Genre Type**




This picture displays word clouds showing the most frequent title words across various genres such as Action, Comedy, Anime, Stand-Up, TV Shows, Romantic, Music, Classic, etc.

**Observations:**

● Words like **"love," "deadly," "war," "master," "girl," "live,"** and **"special"** appear prominently in their respective genres, reflecting recurring themes.

● **"Bob Ross"** and **"pioneers"** stand out in the TV Shows genre, suggesting notable figures or documentary titles.

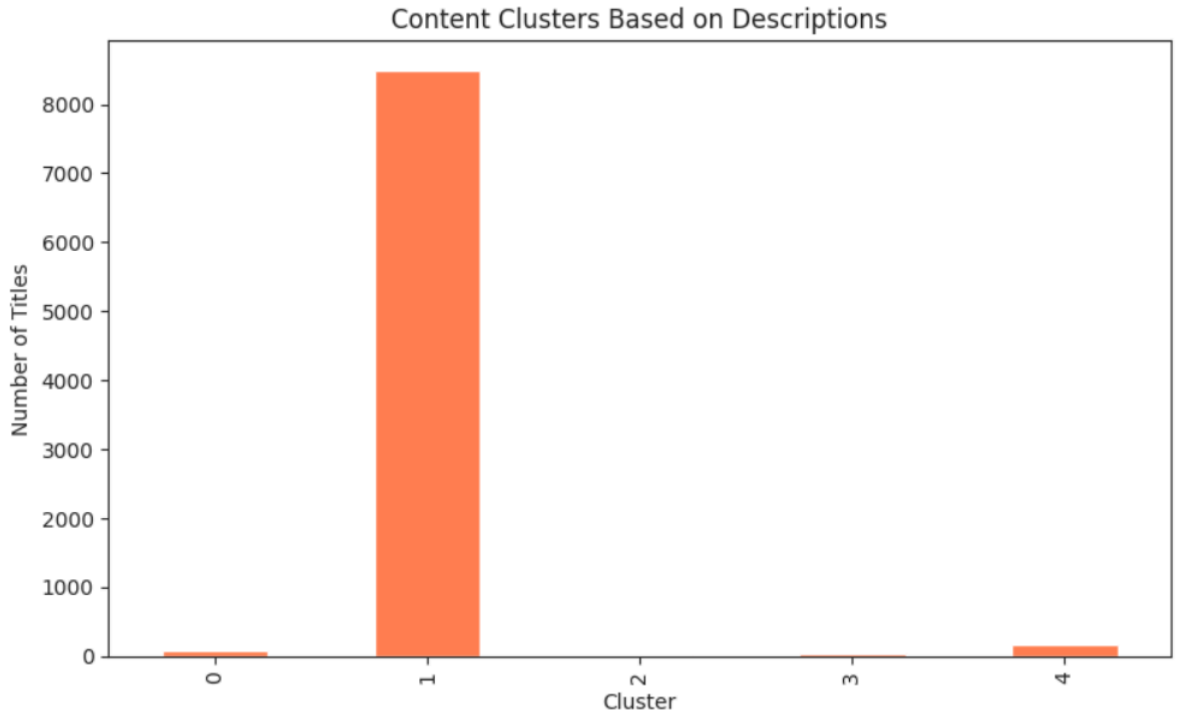
● The Romantic genre is dominated by words like **"virgin," "date," "holiday," "feel,"** and **"jane,"** reflecting relationship and emotional themes.

● Music-related titles include words such as **"live," "music," "mendes," "orchestra,"** and **"country."**

**Insight:**

Title keywords strongly reflect genre expectations, e.g., love and emotion in romantic genres, combat in action/war genres, and performance or artist names in music. These words shape audience perception even before viewing.

**10. Content Clusters Based on Descriptions**




This picture displays the distribution of titles across five content clusters derived from their descriptions using text-based clustering.

**Observations:**

● The vast majority of titles (8500) fall into **Cluster 1**, indicating highly similar description styles or topics.

● Very few titles are spread across the remaining clusters (0, 2, 3, and 4), suggesting limited variation in content description styles.

● Cluster 4 holds a small but notable group, possibly with distinct or niche themes.

**Insight:**

Most content shares a common description pattern, likely generic or similar in structure, indicating a lack of descriptive diversity. This could be due to templated or marketing-driven write-ups, limiting differentiation in metadata.

# Suggested Actions

**1. Improve Metadata Completeness and Quality**

- **Action:** Prioritize filling in missing director, cast, and description fields — particularly for top-performing or recommended content.
- **Why:** Better metadata allows for improved searchability, personalization, and analytics.

**2. Diversify Content Descriptions**

- **Action:** Encourage creative teams to use more descriptive, varied, and engaging language in content summaries.
- **Why:** Over 8500 titles falling into one description cluster shows low variation, which can hurt engagement and recommendation performance.

**3. Content Strategy Optimization**

- **Action:** Invest in high-performing genres like Documentaries and International Dramas, but also explore underrepresented genres like **Sports** or **Classic content**.
- **Why:** There's potential in diversifying offerings to attract niche audiences and boost engagement across more demographics.

**4. Improve Genre Tagging and Search Filters**

- **Action:** Refine or reclassify genre tagging logic to better reflect user-friendly categories and sub-genres.
- **Why:** Many titles appear in overlapping genres (e.g., "Dramas, International Movies"), which may confuse users or skew data insights.

**5. Highlight High-Volume Directors**

- **Action:** Spotlight creators like **Rajiv Chilaka** (Kids' content) or **Alastair Fothergill** (Documentaries) in curated collections.
- **Why:** These individuals contribute significantly and can be promoted as familiar names to viewers.

**6. Release Strategy Planning**

- **Action:** Capitalize on months with higher engagement (e.g., July, December) by releasing flagship titles during these peaks.

- **Why:** Historical monthly release patterns show seasonality — aligning releases with these trends can boost visibility and viewership.

---

## Final Insight

Netflix's strength lies in its global content diversity and volume, but there are critical gaps in metadata quality and description distinctiveness. Addressing these can lead to better recommendation accuracy, user experience, and strategic content planning.

## Conclusion

The analysis reveals key insights into Netflix's content distribution across genres, time, and metadata quality:

- **Content Concentration:** Movie content dominates the platform, both by quantity and release consistency across months and years. However, TV shows, while fewer, show a more stable growth trend from 2016–2020.

- **Genre Diversity:** Top genres for movies are skewed towards **Dramas, Documentaries**, and **Comedies**, while TV shows lean towards **Kids TV**, **TV Dramas**, and **Reality TV**. International content appears heavily in both categories, confirming Netflix's global strategy.

- **Metadata Issues:** A significant portion of data is missing or incomplete, especially with **director information** and repetitive content descriptions. Over 2500 titles lack named directors, and a large majority of content falls into a **single text cluster**, indicating homogeneity in descriptions.

- **Title Word Trends:** Word clouds reflect consistent thematic associations across genres (e.g., "love" in Romantic, "war" in Action), but also reveal some repetitiveness or branding dependency (e.g., "live," "special," "bob ross").

##### THIS IS A WORK DONE BY 'SURENDRAN L'