# Netflix Titles Data Summary

![](data:None;base64,)

## About the dataset

This report summarizes a dataset of 8807 Netflix titles, including both movies and TV shows. The dataset contains 12 columns, providing information such as the unique show ID (show\_id), the type of content (type), the title (title), director (director), cast (cast), country of origin (country), date added to Netflix (date\_added), release year (release\_year), content rating (rating), duration (duration), genre categories (listed\_in), and a brief description (description).

Based on the first five sample data entries, the dataset includes a variety of content, ranging from documentaries like "Dick Johnson Is Dead" to international TV shows such as "Blood & Water" and "Kota Factory". The release years of these initial entries are relatively recent, with most titles released in 2020 or 2021. The content ratings vary, including PG-13 and TV-MA, suggesting a diverse audience range. The duration column reflects the difference between movies (e.g., "90 min") and TV shows (e.g., "2 Seasons"). The average release year across the entire dataset is 2014.18, with a standard deviation of 8.82, indicating a range of release years from 1925 to 2021.

## Relevant Inquiries

### Q1.What is the yearly trend of content releases (movies vs. TV shows) on Netflix? Show the number of titles released each year by content type.

![](data:None;base64,)

#### Content Release Overview

* **Overall Trend**: Both movie and TV show content releases on Netflix have shown a **significant upward trend** over the years, particularly accelerating from the early 2000s and peaking around 2019-2020.
* **Historical Context**: Content releases were **minimal in earlier decades**, with only a few titles per year for both movies and TV shows, starting as early as 1925.

#### Movie Release Trends

* **Growth Trajectory**: Movie releases experienced a **gradual increase** from the mid-20th century, followed by a **sharp acceleration** from approximately 2010 onwards.
* **Peak Releases**: The number of movie releases reached its **highest point** around **2019-2020**, with a maximum of **767 titles** released in a single year.
* **Recent Decline**: Following the peak, there appears to be a slight **decline in movie releases** in the most recent year (2021).

#### TV Show Release Trends

* **Growth Trajectory**: TV show releases also demonstrated a **steady increase**, though their growth rate was generally **slower** than that of movies for most of the period. A more noticeable increase in TV show releases began around 2015.
* **Peak Releases**: TV show releases peaked around the same period as movies, with a maximum of **436 titles** released in a single year.
* **Volume Comparison**: The **volume of TV show releases consistently remained lower** than that of movie releases throughout the observed period.

#### Conclusion and Insights

* **Dominance of Movies**: Movies have consistently been released in **higher volumes** than TV shows on Netflix, especially during the period of rapid content expansion.
* **Accelerated Growth Post-2010**: The period from **2010 to 2020** marked a significant boom in content production for Netflix, with both movies and TV shows experiencing their most substantial growth during this decade.
* **Maturity and Potential Saturation**: The slight decline in releases observed after the peak around 2019-2020 for both content types might indicate a **maturation of content output** or a shift in content strategy.

### Q2.What is the distribution of content duration for movies and TV shows? Show the frequency of different movie lengths and TV show season counts.

![](data:None;base64,)

![](data:None;base64,)

#### Movie Duration Distribution

* **Range of Durations**: Movie durations on Netflix range significantly, from a **minimum of 3 minutes to a maximum of 312 minutes**. The average movie duration is approximately **110.40 minutes**.
* **Most Frequent Durations**: The distribution shows that movies around **90-100 minutes** are the most frequent. Specifically, the highest frequency is observed for movies around **93-97 minutes**, with **93 minutes** having the highest count of **152 movies**, followed closely by **94 minutes** with **147 movies**, and **95 minutes** with **145 movies**.
* **Distribution Pattern**: The frequency of movies generally **decreases as duration increases**, indicating a preference for shorter to medium-length films.

#### TV Show Season Count Distribution

* **Range of Seasons**: TV shows on Netflix vary from **1 season to 17 seasons**. The average number of seasons for a TV show is approximately **8.20 seasons**.
* **Dominant Season Count**: A significant majority of TV shows have **1 season**, accounting for **1793 titles**. This is by far the most frequent season count.
* **Decreasing Frequency with More Seasons**: The frequency of TV shows **drastically decreases as the number of seasons increases**. For instance, while 1-season shows are dominant, 2-season shows are significantly fewer (425 titles), and shows with 10 or more seasons are very rare (e.g., 17-season shows only have 1 title).

#### Conclusion and Insights

* **Content Length Disparity**: There is a clear distinction in content length distribution between movies and TV shows on Netflix. Movies tend to cluster around a typical feature film length (around 90-120 minutes), while TV shows are heavily skewed towards single-season productions.
* **Preference for Shorter Formats**: For movies, there's a general trend towards shorter durations, with frequencies declining for longer films. For TV shows, the overwhelming prevalence of single-season series suggests a strong preference for limited series or shows that do not get renewed beyond their initial run.
* **Implications for Content Strategy**: The data indicates that Netflix's content library is rich in standard-length movies and a large volume of single-season TV shows, potentially reflecting a strategy to cater to viewers who prefer shorter, more digestible content or to test new series concepts without long-term commitments.

### Q3.Which country contributes the most content to Netflix, broken down by content type (movie vs. TV show)?

![](data:None;base64,)

#### Top Movie Producing Country

* **Country**: The **United States** is the top country for movie content on Netflix.
* **Content Count**: It contributes **2752** movies to the platform.

#### Top TV Show Producing Country

* **Country**: The **United States** is also the top country for TV show content on Netflix.
* **Content Count**: It contributes **938** TV shows to the platform.

#### Conclusion and Insights

* **Dominant Content Producer**: The **United States** is the leading country in contributing content to Netflix for both **movies (2752 titles)** and **TV shows (938 titles)** .
* **Content Type Distribution**: While the United States leads in both categories, its contribution of movies is significantly higher than its contribution of TV shows, with **nearly three times more movies** than TV shows .
* **Visualization Confirmation**: The analysis is supported by the visualization which explicitly labels "Movies (United States)" and "TV Shows (United States)" as the categories being displayed, indicating their top status .

### Q4.What are the top 10 most frequent genres on Netflix? Show the number of titles belonging to each genre.

![](data:None;base64,)

#### Genre Frequency

* **International Movies**: This genre is the **most frequent**, appearing in **2,752** titles.
* **Dramas**: Following International Movies, Dramas are the second most frequent genre with **2,427** titles.
* **Comedies**: This genre ranks third, with **1,674** titles.
* **International TV Shows**: This genre is fourth, with **1,351** titles.
* **Documentaries**: Documentaries are the fifth most frequent, with **869** titles.
* **Action & Adventure**: This genre accounts for **859** titles.
* **TV Dramas**: This genre has **763** titles.
* **Independent Movies**: This genre is present in **756** titles.
* **Children & Family Movies**: This genre has **641** titles.
* **Romantic Movies**: This genre is the tenth most frequent, with **616** titles.

#### Conclusion and Insights

* **Dominance of International Content**: The top two most frequent genres, **International Movies** and **International TV Shows**, highlight a significant presence of non-domestic content on Netflix, indicating a strong global content strategy or audience preference.
* **Popularity of Core Genres**: **Dramas** and **Comedies** remain highly popular, securing the second and third positions respectively, which suggests a consistent demand for these fundamental storytelling categories.
* **Diverse Content Offering**: The top 10 list showcases a broad range of content, from **Documentaries** and **Action & Adventure** to **Children & Family Movies** and **Romantic Movies**, indicating Netflix's diverse catalog caters to various audience interests.
* **Significant Drop-off in Frequency**: There is a notable decrease in the number of titles from the top-ranking genres (e.g., International Movies with 2752) to the lower-ranking genres within the top 10 (e.g., Romantic Movies with 616), suggesting a concentration of titles within the most popular categories.

### Q5.What is the distribution of content ratings on Netflix? Show the percentage of titles for each rating category.

![](data:None;base64,)

#### Percentage Distribution of Content Ratings

* **TV-MA**: Represents the largest share of content, accounting for **36.44%** of titles.
* **TV-14**: Is the second most common rating, making up **24.55%** of titles.
* **TV-PG**: Constitutes **9.81%** of the content.
* **R**: Accounts for **9.08%** of titles.
* **PG-13**: Represents **5.57%** of the content.
* **TV-Y7**: Makes up **3.80%** of titles.
* **TV-Y**: Accounts for **3.49%** of titles.
* **PG**: Represents **3.26%** of the content.
* **TV-G**: Constitutes **2.50%** of titles.
* **NR**: Accounts for **0.91%** of titles.
* **G**: Represents **0.47%** of the content.
* **TV-Y7-FV**: Makes up **0.07%** of titles.
* **NC-17**: Accounts for **0.03%** of titles.
* **UR**: Represents **0.03%** of the content.

#### Visual Representation of Content Rating Distribution

* The bar chart visually confirms the dominance of **TV-MA** and **TV-14** ratings, which are significantly higher than other categories.
* The distribution shows a clear descending trend, with fewer titles in the less common rating categories like G, NC-17, and UR.

#### Conclusion and Insights

* **Dominance of Mature Content**: The distribution clearly indicates that **TV-MA** (36.44%) and **TV-14** (24.55%) are the most prevalent content ratings on Netflix, collectively accounting for over **60%** of all titles. This suggests a strong focus on content suitable for mature audiences or those aged 14 and above.
* **Significant Share of General Audiences/Parental Guidance Content**: Ratings like **TV-PG** (9.81%), **R** (9.08%), and **PG-13** (5.57%) also hold substantial portions, indicating a diverse offering that includes content requiring parental guidance or restricted to certain age groups.
* **Limited Children's and Niche Content**: Content suitable for younger audiences (e.g., TV-Y7, TV-Y, TV-G, G) and very specific adult ratings (NC-17, UR) represent a much smaller percentage of the total library, with each category typically below 5% and some even below 1%. This highlights a comparatively smaller emphasis on content for very young children or highly restricted adult content.

### Q6.Which directors have the most titles on Netflix? Show the number of titles directed by each director.

![](data:None;base64,)

#### Top Directors by Title Count

* **Unknown**: The category for titles with **missing director information** accounts for the highest number of entries, with **2,634 titles**. This indicates a significant portion of the dataset lacks director attribution.
* **Rajiv Chilaka**: This director has the highest number of attributed titles, with **19 titles**.
* **Raúl Campos, Jan Suter**: This directorial duo is responsible for **18 titles**.
* **Marcus Raboy**: This director has contributed **16 titles**.
* **Suhas Kadav**: This director also has **16 titles**.

#### Conclusion and Insights

* **Dominance of Unattributed Titles**: The analysis clearly shows that the largest single category of titles on Netflix, by a significant margin, is those with **unspecified or missing director information**, represented by the 'Unknown' category. This category alone accounts for **2,634 titles**, far exceeding any individual director's contribution.
* **Top Individual Directors**: Excluding the 'Unknown' category, **Rajiv Chilaka** stands out as the director with the most titles (19), followed closely by **Raúl Campos, Jan Suter** (18), and then **Marcus Raboy** and **Suhas Kadav** (both with 16 titles). This highlights a relatively small number of individual directors who have contributed multiple titles to the platform.
* **Skewed Distribution**: The distribution of titles per director is highly skewed, with a vast majority of directors having only one title, while a few (or the 'Unknown' category) account for a much larger share. The average number of titles per director is approximately **1.94**, but this average is heavily influenced by the 'Unknown' category's large count.

### Q7.Which actors appear most frequently together in Netflix content? Visualize the co-occurrence of actors in different titles.

![](data:None;base64,)

* **Top Co-occurring Pair**: The pair **Julie Tejwani & Rupa Bhimani** appeared together most frequently, with a count of **31**.
* **High Co-occurrence Group**: A group of actors including **Julie Tejwani, Rupa Bhimani, Rajesh Kava, Jigna Bhardwaj, Vatsal Dubey, Mousam, and Swapnil** show high rates of co-occurrence, with counts ranging from **14 to 31**.
* **Other Notable Pairs**: Other frequently co-occurring pairs include **John Paul Tremblay & Robb Wells (15 times)**, **Michael Palin & Terry Jones (14 times)**, and various combinations of **Eric Idle, John Cleese, Terry Jones, Michael Palin, Terry Gilliam, and Graham Chapman (12-14 times)**.
* **Minimum Co-occurrence for Top 50**: The least frequent pair among the top 50 co-occurred **12 times**.

#### Actor Co-occurrence Network Visualization

* **Network Representation**: The visualization presents an **Actor Co-occurrence Network** where each actor is represented as a **node**, and a connection (edge) between two nodes indicates that the actors have co-occurred in content.
* **Edge Weight/Thickness**: The thickness or color of the edges is intended to represent the frequency of co-occurrence, highlighting stronger relationships. While the image provided does not clearly show varying edge thickness or color, it does display the co-occurrence count next to some edges (e.g., "15" for John Paul Tremblay & Robb Wells, "13" for Ian James Corlett & Andrea Libman, "12" for Alessandro Juliani & Elyse Maloway, "13" for Yuki Kaji & Takahiro Sakurai).
* **Clustered Collaborations**: The network graph reveals distinct clusters of actors who frequently collaborate. For instance, a large cluster includes **Jigna Bhardwaj, Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Rajesh Kava, Mousam, and Swapnil**, indicating a strong collaborative group.
* **Other Collaborative Groups**: Other noticeable groups include:
* **John Paul Tremblay & Robb Wells**.
* **Eric Peterson, Anna Claire Bartlam, Jamie Watson, and Michela Luci**.
* **Ian James Corlett, Andrea Libman, Alessandro Juliani, Elyse Maloway, Vincent Tong, and Diana Kaarina**.
* **Daisuke Ono, Yuki Kaji, and Takahiro Sakurai**.
* **Eric Idle, Terry Jones, Terry Gilliam, John Cleese, Graham Chapman, and Michael Palin**.

#### Conclusion and Insights

* **Dominant Collaborative Groups**: The analysis clearly identifies several tightly-knit groups of actors who frequently work together. The most prominent group revolves around **Julie Tejwani, Rupa Bhimani, Rajesh Kava, and Jigna Bhardwaj**, suggesting they are key figures in a particular set of productions on Netflix.
* **Strongest Pairings**: The pair **Julie Tejwani & Rupa Bhimani** stands out as the most frequent collaboration, indicating a highly successful or recurring partnership.
* **Network Structure**: The network visualization effectively illustrates these collaborative patterns, showing distinct communities of actors rather than a single, highly interconnected network. This suggests that co-occurrence is often localized within specific production teams or genres.
* **Implications for Content Strategy**: Understanding these frequent co-occurrences can be valuable for content creators and Netflix in identifying successful casting combinations, potential spin-offs, or recurring themes/genres associated with these actor groups.

### Q8.What is the monthly trend of content additions to Netflix? Show the number of titles added each month.

![](data:None;base64,)

#### Data Preparation

* **Data Source**: The analysis began by processing the column from the table (likely ).
* **Date Parsing**: The column was converted into datetime objects, with any unparseable dates being dropped.
* **Monthly Aggregation**: The year and month were extracted from the cleaned column, and the data was then grouped by these year-month combinations.
* **Title Count**: For each year-month period, the number of titles added was counted, resulting in a dataset named . This dataset contains (string) and (numeric) columns.
* **Statistical Summary**: The column in the dataset has a **mean of 80.64**, a **standard deviation of 74.17**, a **minimum of 1.00**, and a **maximum of 257.00** titles added per month.

#### Monthly Title Count Trend

* **Early Period (2008-2015)**: From **2008-01 to 2015-06**, the number of titles added monthly was consistently very low, often **1 to 5 titles**. There was a slight increase around **2012-02 to 2012-06** with **11 titles** added, and a gradual rise starting from **2014-10** (e.g., **13 titles** in **2014-10**, **28 titles** in **2015-10**).
* **Growth Acceleration (2016-2018)**: A significant increase in content additions began in **2016**. For instance, **41 titles** were added in **2016-02**, reaching **95 titles** by **2017-02**. The trend continued upwards, with monthly additions frequently exceeding **100 titles** from **2017-10** onwards (e.g., **123 titles** in **2017-10**, **170 titles** in **2018-06**, and **190 titles** in **2018-10**).
* **Peak and Volatility (2019-2021)**: The period from **2019 to mid-2021** shows the highest number of monthly content additions, but also increased volatility.
* **2019**: Monthly additions were high, with a peak of **191 titles** in **2019-10**.
* **2020**: The highest peak in content additions occurred in **2020-02** with **253 titles**, followed by **242 titles** in **2020-06**. There was a notable dip to **114 titles** in **2020-04**.
* **2021**: The trend continued with high numbers, reaching **257 titles** in **2021-02**, which is the **overall maximum** observed. This was followed by **207 titles** in **2021-04** and **183 titles** in **2021-06**.

#### Conclusion and Insights

* **Overall Growth Trend**: The monthly addition of titles to Netflix shows a clear and **significant upward trend** over time, particularly accelerating from **2016** onwards.
* **Early Slow Growth**: In its early years (2008-2015), Netflix added content at a very slow pace, with monthly additions rarely exceeding single digits.
* **Rapid Expansion**: The period from **2016 to 2021** marks a phase of rapid content expansion, with monthly title additions consistently in the hundreds, indicating a strong focus on growing the content library.
* **Peak Activity**: The highest volumes of content additions were observed in **early 2020** and **early 2021**, with peaks of **253 titles** in **February 2020** and **257 titles** in **February 2021**.
* **Increased Volatility**: While the overall trend is upward, the later years (2019-2021) also show more pronounced fluctuations in monthly additions, suggesting potential seasonality or strategic release schedules.

### Q9.How should we handle the column, which is currently a string (e.g., '100 min'), to extract a numerical value (minutes) for movies, and what approach should be taken for any entries that cannot be converted or are missing?

#### Handling Column for Numerical Extraction

* **New Column Creation**: A new column, , is initialized with **null values (pd.NA)** to store the numerical duration.
* **Type-Specific Processing**: The extraction process is **specifically applied only to entries where the is 'Movie'**. For 'TV Show' entries, the column remains null.
* **Numerical Extraction**: For 'Movie' entries, a **regular expression ()** is used to extract the numerical part from the string (e.g., '90 min' becomes 90).
* **Error Handling and Missing Values**:
* If the extraction fails (e.g., the string is not in the expected 'X min' format or is missing), the function with is used to **assign a null value (NaN)** to for that specific entry.
* The initial null values in for 'TV Show' entries are preserved.
* **Data Type Conversion**: The column is converted to a **nullable integer type ()** to accommodate both integer values and nulls.

#### Overall Approach

* **Dedicated Numerical Column**: Instead of modifying the original column, a **new column ()** is created to store the numerical representation, preserving the original string format.
* **Conditional Extraction**: The numerical extraction is **conditionally applied based on the of content**, ensuring that only movies have a calculated duration in minutes.
* **Robust Error Management**: The approach effectively handles cases where duration strings are malformed or missing by **coercing unparseable values to null**, preventing errors and maintaining data integrity.

### Q10.Identify outliers in the release year data. Use a boxplot to visualize the distribution and highlight any unusual release years.

![](data:None;base64,)

#### Data Overview

* **Number of Records**: The dataset contains **8807** entries for 'Release Year'.
* **Data Type**: The 'Release Year' column is of a **numeric** type, suitable for statistical analysis.

#### Statistical Summary of Release Years

* **Mean Release Year**: The average release year is approximately **2014.18**.
* **Standard Deviation**: The release years have a standard deviation of **8.82**, indicating a relatively narrow spread around the mean.
* **Minimum Release Year**: The earliest recorded release year is **1925.00**.
* **Maximum Release Year**: The latest recorded release year is **2021.00**.

#### Release Year Distribution Visualization

#### Boxplot Analysis

* **Visual Representation**: The boxplot for 'Release Year Distribution' shows a **highly concentrated distribution** of release years.
* **Box and Whiskers**: The box and whiskers are extremely compressed, appearing as a very narrow blue line around the 2020-2021 mark on the y-axis. This indicates that the vast majority of release years are clustered within a very recent period.
* **Outlier Identification**: The boxplot **does not visually display any individual points outside the whiskers**, which would typically represent outliers. This suggests that while there is a wide range from 1925 to 2021, the bulk of the data is so recent that older entries, despite being far from the median, are not explicitly marked as outliers by the boxplot's default calculation, or they are too few to be visible as distinct points given the scale.

#### Conclusion and Insights

* **Concentrated Recent Releases**: The analysis reveals that the majority of content on Netflix has been released in **recent years**, with the mean release year being 2014.18 and the boxplot showing a strong concentration around 2020-2021.
* **Absence of Visible Outliers in Boxplot**: Despite a wide range of release years from 1925 to 2021, the boxplot **does not visually identify any outliers**. This is likely due to the overwhelming number of recent releases, which compresses the interquartile range (IQR) and whiskers, making older, less frequent release years fall within the calculated whisker boundaries or not appear as distinct points due to the visualization's scale.
* **Potential for Hidden Outliers**: While the boxplot doesn't explicitly show them, the significant difference between the minimum release year (1925) and the mean (2014.18) suggests that older titles, though not flagged as statistical outliers by this specific boxplot's rendering, represent a **small proportion of the total content** and are far removed from the central tendency of the data. Further investigation with a different visualization or outlier detection method might be needed to highlight these older entries if they are of specific interest.