Exploratory Data Analysis of Netflix Dataset

By: Shivangi

**Abstract:** This study presents an Exploratory data analysis (EDA) of Netflix dataset to understand content, trends, and distribution on the platform.By various data visualisation techniques, we analyze categorical and numerical features, handling missing values and derive meaningful insights.

**Introduction:** The Netflix dataset provides an overview of the content available in the dataset, including information on titles, directors, cast, release\_year, rating and many more. As Netflix is large and global entertaining library, It needs to understand the distribution and trends for both consumer and creator.

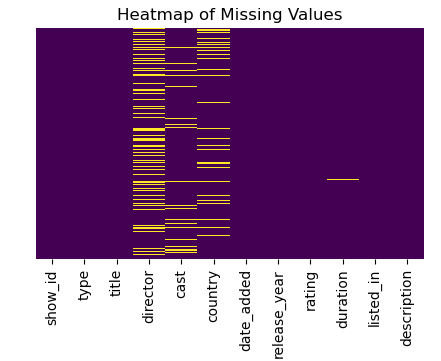
**Data Dictionary:** The dataset has 12 columns and 8807 rows each representing a unique Netflix title.

| **Serial No.** | **Field Name** | **Data Type** | **Possible Values** | **Purpose** |
| --- | --- | --- | --- | --- |
| 1 | show\_id | Categorical | Unique ID for each show | Identifies each record uniquely |
| 2 | type | Categorical | "Movie", "TV Show" | Distinguishes between movies and TV shows |
| 3 | title | Categorical | Various string values | Title of the movie or TV show |
| 4 | director | Categorical | Names of directors, NaN for missing values | Provides the director's name |
| 5 | cast | Categorical | Names of cast members, NaN for missing values | Lists the cast members of the movie or TV show |
| 6 | country | Categorical | Country names, NaN for missing values | Indicates the country where the content was produced |
| 7 | date\_added | Datetime | Dates in the format "Month Day, Year" | Shows the date when the content was added to Netflix |
| 8 | release\_year | Numerical | Year of release (e.g., 2020) | Indicates the year the movie or TV show was released |
| 9 | rating | Categorical | Various ratings like "PG-13", "TV-MA", NaN | Shows the content rating |
| 10 | duration | Categorical | Duration for movies in minutes, or number of seasons | Indicates the duration of movies or the number of seasons for TV shows |
| 11 | listed\_in | Categorical | Genres or categories (e.g., "Documentaries", "Comedies") | Categorizes the content by genre or type |
| 12 | description | Categorical | Various string values | Provides a brief description of the content |

**Problem Statement:** To analyze Netflix's content library to identify trends in content distribution, popular genres, and key contributors. This analysis will provide insights that can help in taking decisions in content creation for producers and consumers.

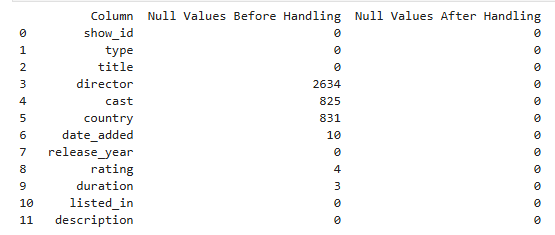
**Data Cleaning and Preprocessing:**

**Handling Missing Values:** Certain columns such as director, cast, country, and date\_added contained missing values.



**To Handle that I have some strategy:**

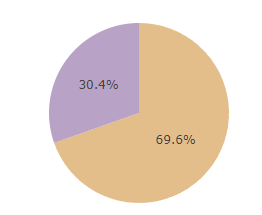
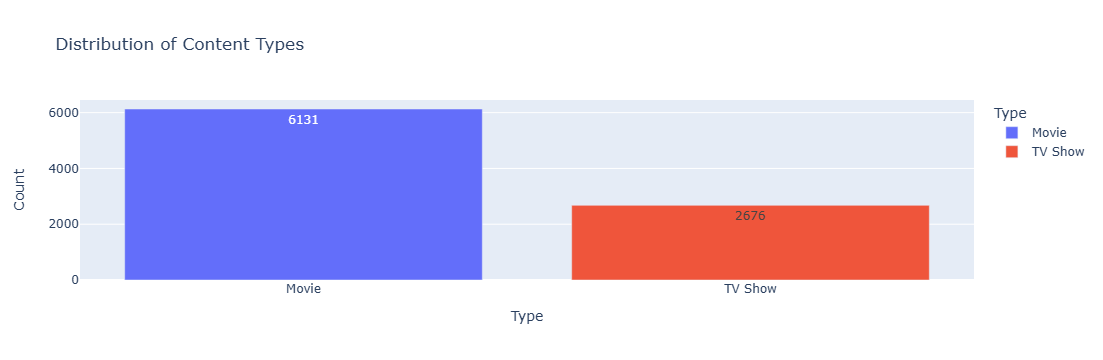
* **Director, Cast, and Country**: Missing values were filled with "Unknown".
* **Rating**: Missing values were filled with "Not Rated".
* **Duration**: For missing duration values, we categorized content as either "Movie" or "TV Show" and filled missing values accordingly (e.g., "0 min" for movies and "0 Seasons" for TV shows).

****

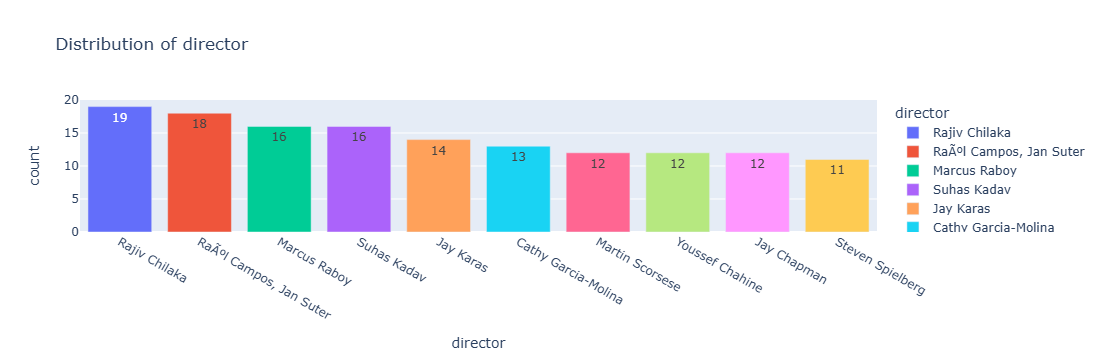
**Exploratory Data Analysis:**

**Univariate analysis:**

**Type Distribution**: A bar chart visualized the distribution between movies and TV shows. On Netflix no. of *movies (69.6%)* are more as compare to TV shows .

** **

**Director Analysis**: We analyzed the top 10 directors by the number of titles they directed on Netflix, excluding unknown values. *Rajiv Chaka* has released more number of movies, he is in top 10 directors.



**Rating distribution**: The distribution of ratings was analyzed to understand Netflix's target audience.*TV-MA (Mature Audience)* has got more rating and that will be in Top-10 ratings. It reflects Netflix’s content strategy, which likely includes targeting adult viewers who prefer more mature shows.

 **TV-MA**: Mature Audience - Content is specifically designed to be viewed by adults and may not be suitable for children under 17.

 **TV-14**: Parents Strongly Cautioned - Content may not be suitable for children under 14.

 **TV-PG**: Parental Guidance Suggested - Some material may not be suitable for children.

 **R**: Restricted - Restricted to viewers over 17, unless accompanied by an adult.

 **PG-13**: Parents Strongly Cautioned - Some material may be inappropriate for children under 13.

 **TV-Y7**: Directed to Older Children - Content suitable for children age 7 and older.

 **TV-Y**: All Children - Content suitable for all children.

 **PG**: Parental Guidance Suggested - Some material may not be suitable for children.

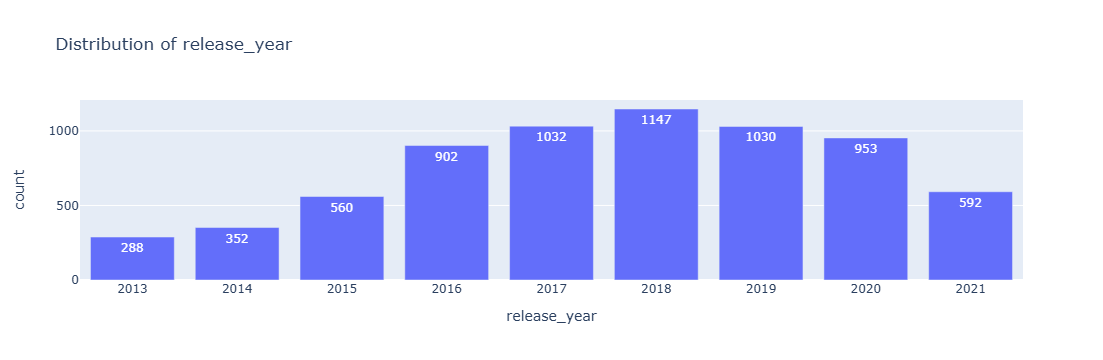
 **TV-G**: General Audience - Content suitable for all ages.

 **G**: General Audience - Content suitable for all ages.



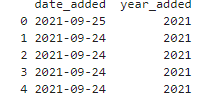
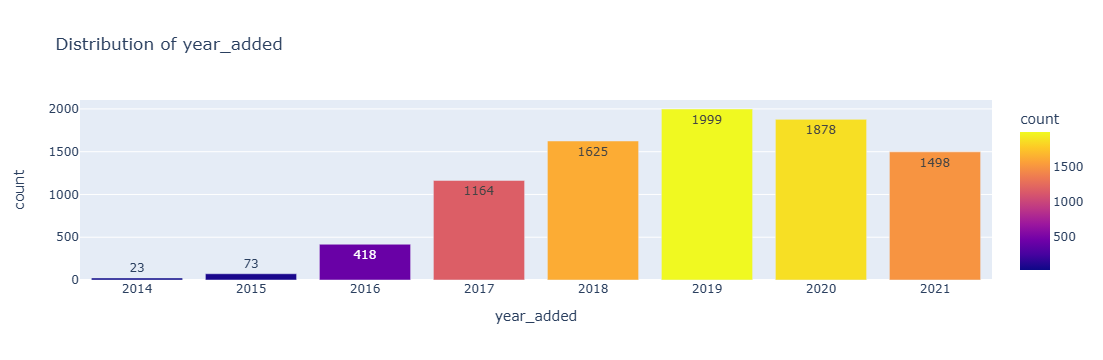
**Numerical Features Analysis**

**Release Year Distribution:** That’s the column when content will released.A bar graph was created to visualize the release year distribution. In 2018 the number of movies and TV shows released most whereas after that it will continuously decreases.

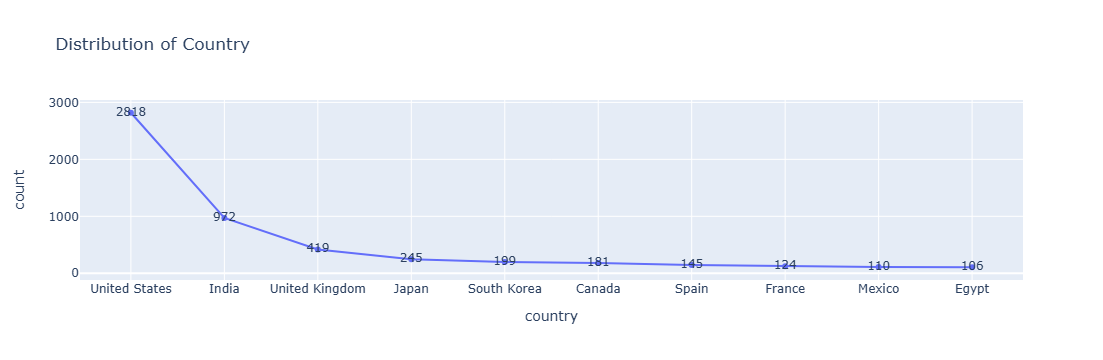


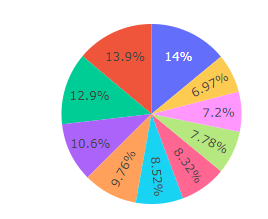
**Date added Distribution**: That’s the column when content has added on Netflix by year. A bar graph has created to visualize the date added distribution. In 2019 more no. of movies and TV shows added on Netflix.

I have extracted the year from the date\_added column to create a new feature year\_added by which we get some more insights**.**

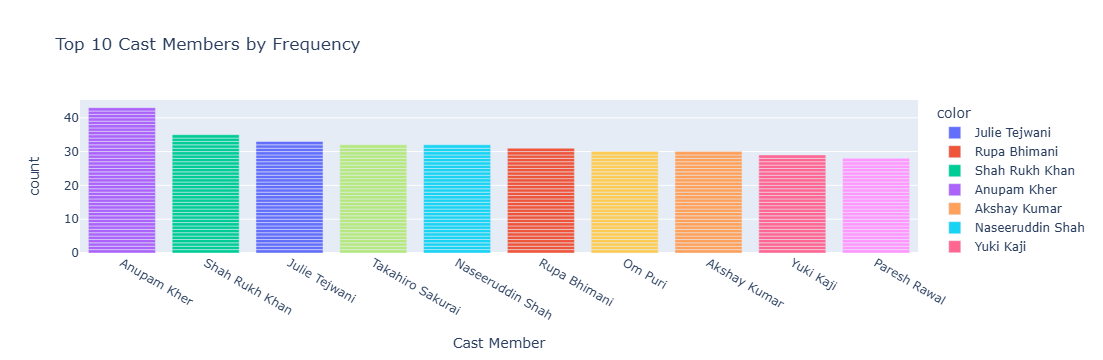
**Distribution of Country:** Top 10 countries where Netflix is using mostly. Netflix has more user from United State and India.



**Listed Distribution:** This will helping us to understand which types of content are leading the charge on Netflix. 

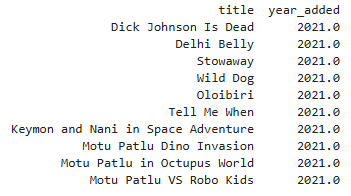
Dramas, International Movies has more charged whereas Documentaries and International movies has lesser charged it means user’s are not interested in Documentaries.

**Cast Distribution:** Top10 Cast member , this will helping us to understand who is the best cast. Anupam kher and Sahrukh khan are top 2 best performer.

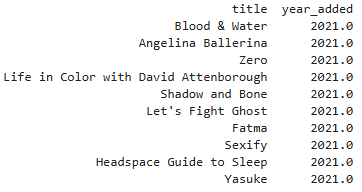


**Bivariate Analysis:**

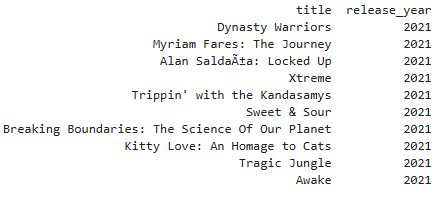
**Recent movies distribution by date added:** Top 10 recent movies which will added on Netflix. Dick Johnson is Dead that’s the most recent added movie on Netflix.



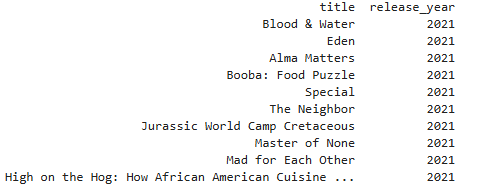
**Recent TV shows distribution by date added:** Top 10 recent TV shows which will added on Netflix. Blood and water that’s the most recent added movie on Netflix.



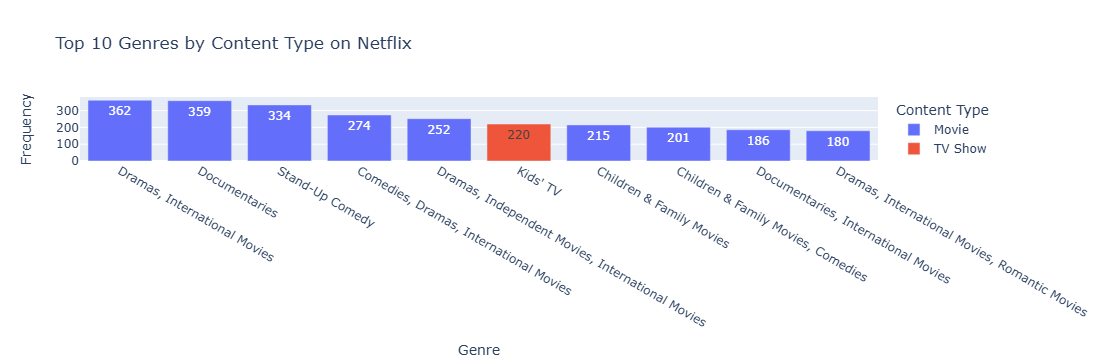
**Recent movies distribution by release year:** Top 10 recent movies released in market. Dynasty warrior that’s the most recent released movie in market.



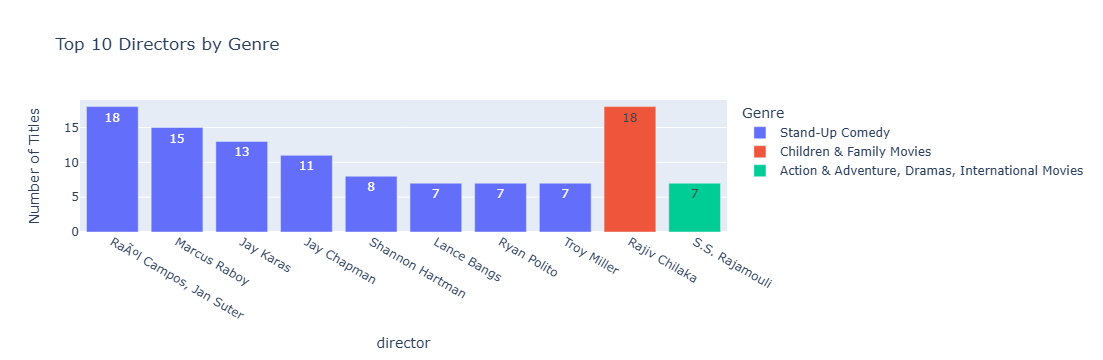
**Recent TV shows distribution by release year:** Top 10 recent TV shows released in market. Blood and water that will be released as that movie comes in market but after that it will be changed second one released movie was Eden.



**Genres by Content Type on Netflix:** This provide a comparison of genre popularity between movies and TV shows on Netflix, making it easy to identify which genres are most common within each content type.

****

**Director by Genre on Netflix:** This provide a comparison of genre popularity by directors on Netflix, making it easy to identify which genres are most released by directors.

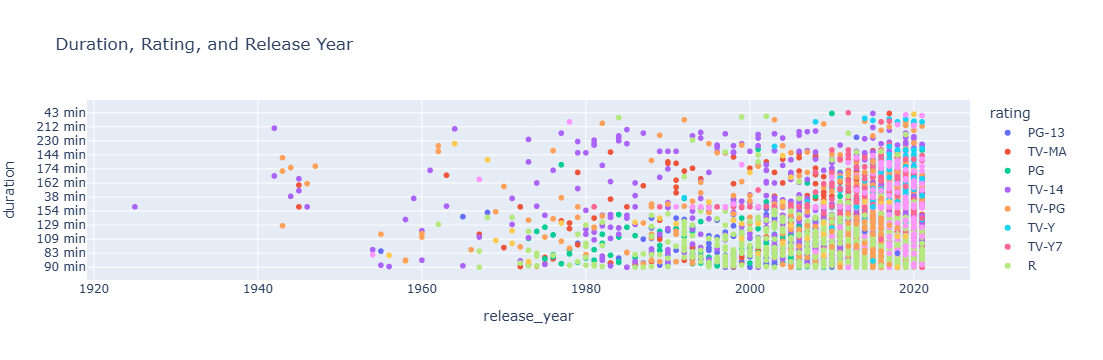


**Multivariate Analysis:**

**Duration, Rating and Release year:** The scatter plot reveals a general trend in content duration over the years. There has been a gradual increase in the average duration of content since the year 2000.

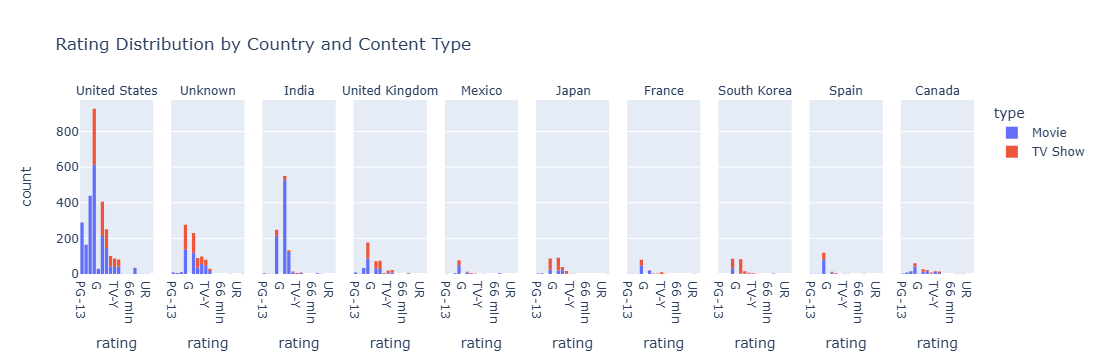
TV-MA rating tends to have longer durations, particularly in the last decade, indicating a shift towards more mature content. TV-PG content generally has shorter durations compared to TV-MA content, suggesting that content targeting younger audiences.

In 2020 the rating and duration will gradually increase it may be due to covid panedemic.

****

**Analysis of Rating Distribution by Country and Content Type :** The histogram visualizes the distribution of content ratings across various countries, categorized by content type (Movies or TV Shows).

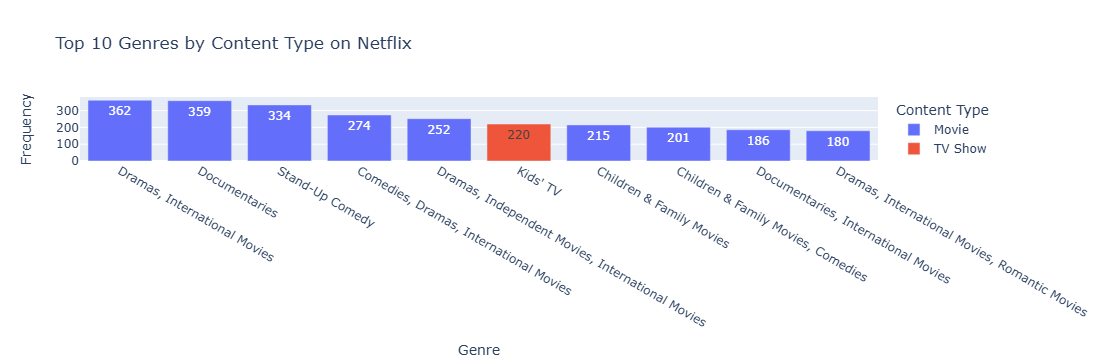
India shows content rated as 'TV-14' and 'TV-MA', indicating that content is suitable for teenagers and mature audiences, producing more TV-MA content might be strategic for markets like India and the United States, where this rating is popular.

****

**What I Explore using the Netflix dataset:**

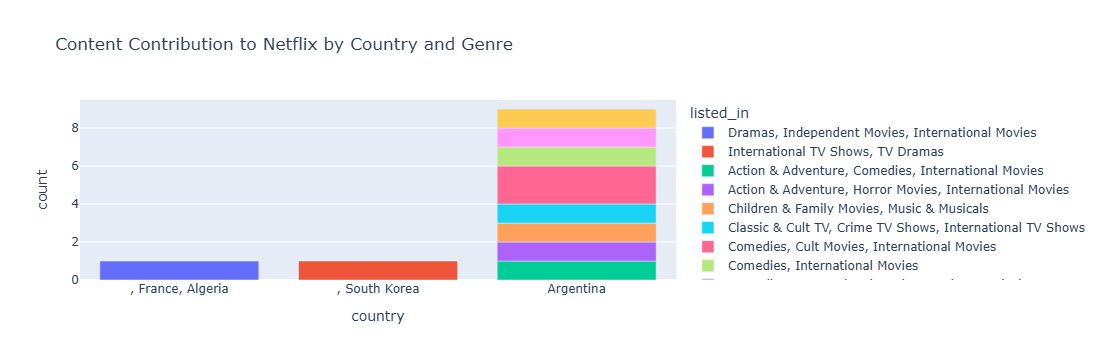
1. What are the most common genres on Netflix, and how do they differ between movies and TV shows?

* The most common genres could include Documentaries, International Movies, and Dramas for movies, while Kid’s TV for TV shows.

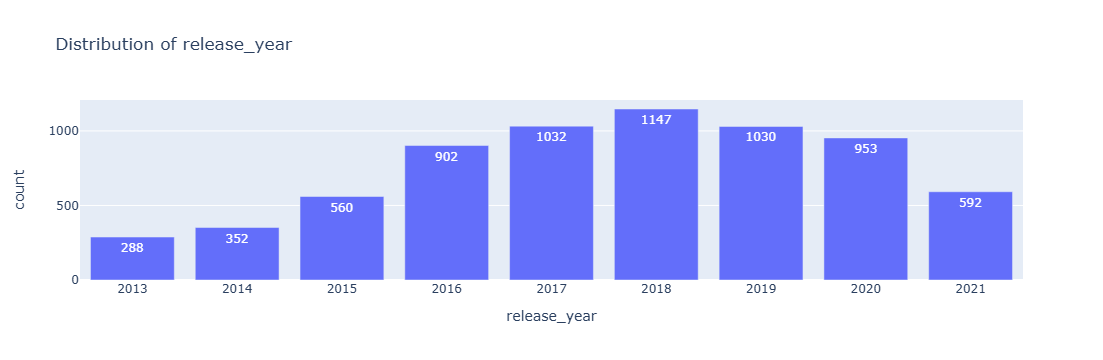
****

1. Which countries contribute the most content to Netflix, and what genres are they most known for?

* France, Algeria: Known for contributing content in genres such as Drama and International Movies.
* **South Korea:** A major contributor to Netflix's especially in genres such as International TV shows, TV Dramas.
* **Argentina**: This country is noted for its content in genres such as Drama, Comedy, and Romance.



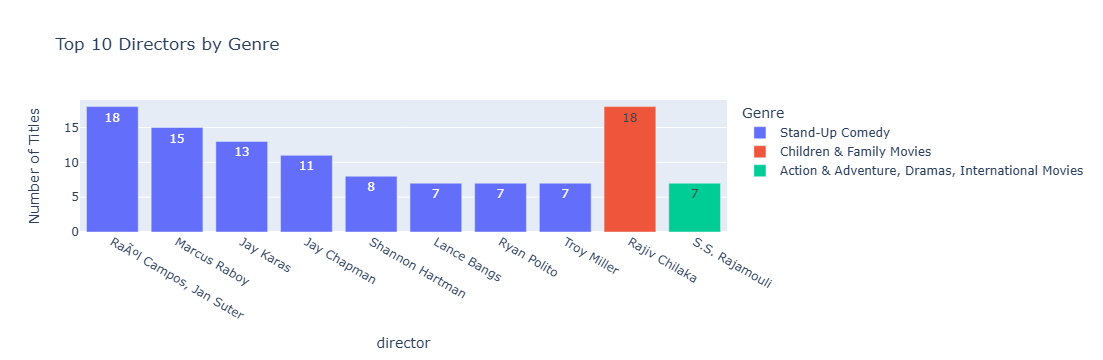
1. What are the trends in content release years on Netflix? Are there any noticeable patterns in content addition over time?

**Increase in Content**: there's a sharp rise from 2018 in content addition in recent years, Netflix's expanding global reach and its investment in original content. 

1. Who are the most frequent directors on Netflix, and what type of content do they produce?

Here as we can see that Rajiv Chaka who is top director he has released children and family movies type content

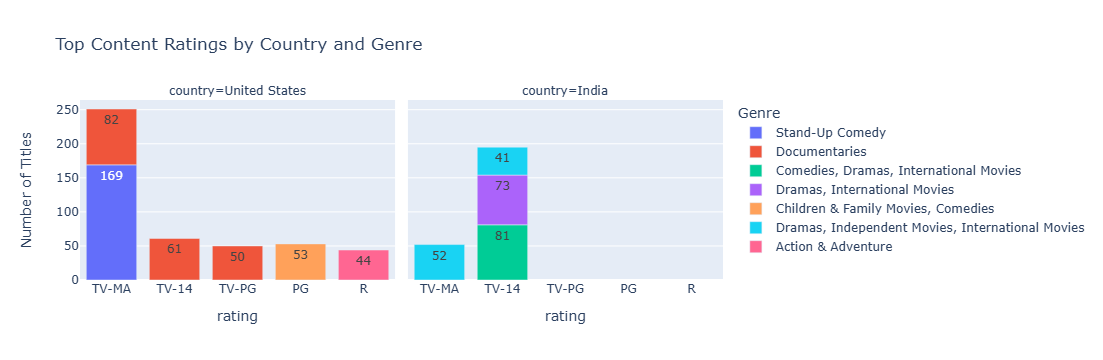
And Standup comedy that’s the most common genre we can observe that every director has worked on that and released that type of content.



1. What are the most common content ratings on Netflix, and how do they correlate with genres or countries?

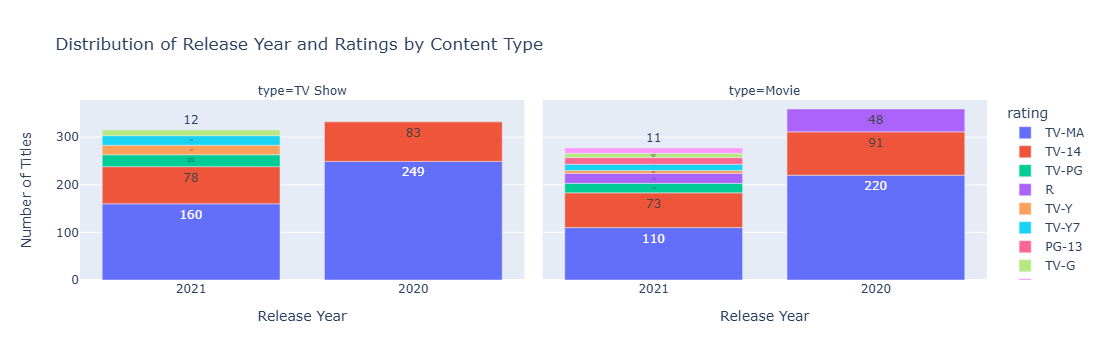
Here, I have selected 2 top performing countries and analysing that what type of content liked by their users and which one has top ratings.

So, as we can see that TV-MA(matured audience) is most rated for Stand-Up comedy which was mostly seen in US. In India the users most rated i.e. TV-14 has Comedy, Drama and International movies.



1. How many TV shows versus movies are available on Netflix, and how do they differ in terms of release years and ratings?

In these analysis we will firstly see the TV show and movie in which the TV-MA is most rated but it has decreases from 2020 to 2021. In fact all ratings are gradually decreases from 2020 to 2021.



1. Find out Top 10 Comedy movies and Drama?



1. Find out Top 2 Movie which has got highest rating:

****

1. Top 2 TV show which has got highest rating:

****