**Business Case: Netflix - Data Exploration and Visualisation**

Data Set provided -https://d2beiqkhq929f0.cloudfront.net/public\_assets/assets/000/000/940/original/netflix.csv

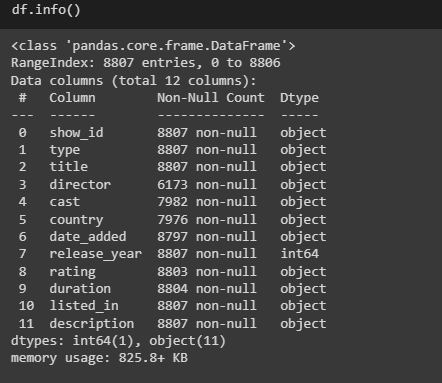
Google Collab Link - https://colab.research.google.com/drive/10WkY7MqQpeRZTP9LV2I3F3HazgLGYKg1?usp=sharing

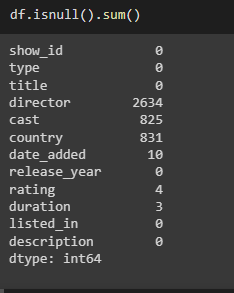
The dataset provided to you consists of a list of all the TV shows/movies available on Netflix:

**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:** TV Rating of the movie/show  
**Duration:** Total Duration - in minutes or number of seasons  
**Listed\_in:** Genre  
**Description:** The summary description

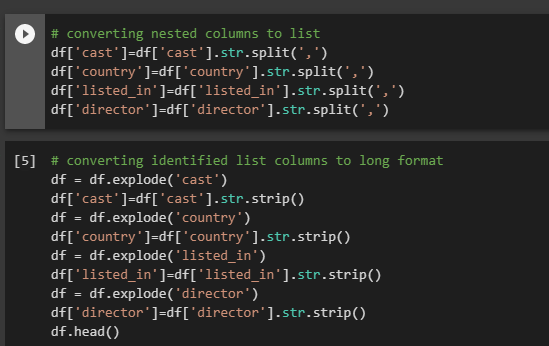
1. **Cleaning Data**

A total of 8807 rows are provided in the dataset with a few of the columns having null values. Below is the screenshot





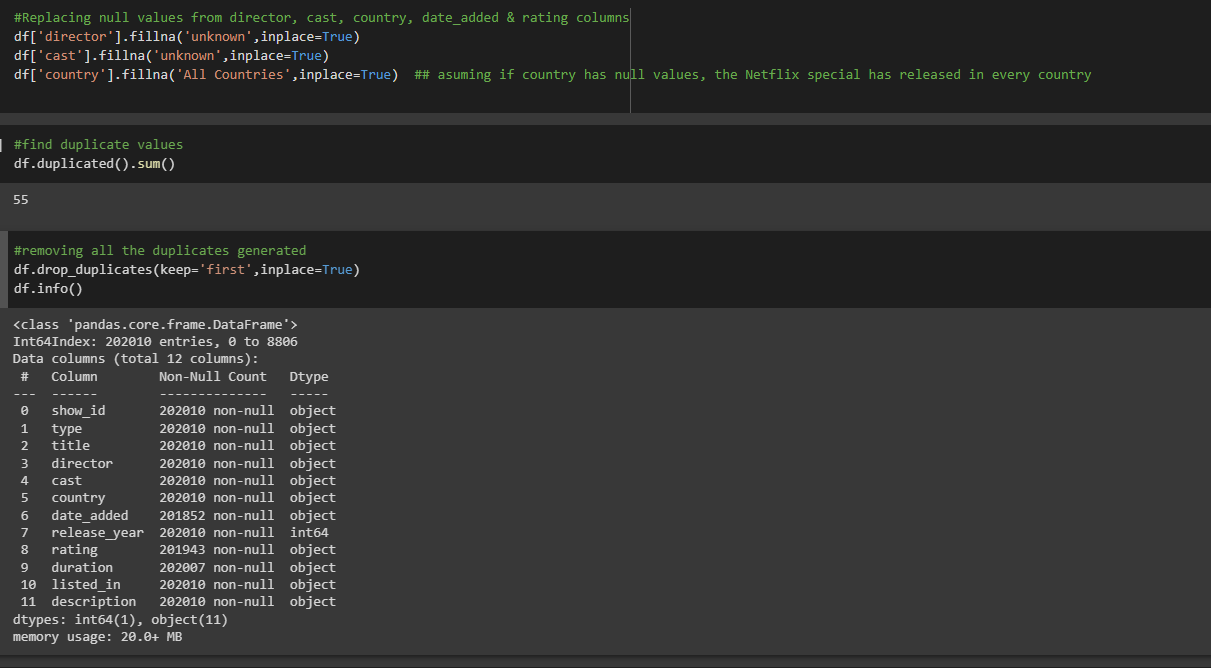
There also has been the observation that columns like director, cast, country and listed\_in have comma-separated values. In order to deal with the above-mentioned issue, I converted the columns into list format where the dividing parameter is ‘,’ for a particular cell and then use explode function for assigning individual values from the list to individual cells in the columnar format.



With this operation performed the total no. of rows increased to 202065 from 8807.

The only operation remaining was to replace the null values in the cast, director and country columns with something meaningful.

I have replaced all the null values in the cast and director columns with ‘**unknown**’ and all the null values in the country column with ‘**All countries**’. And dropped the rows with duplicated values keeping only 1st row.

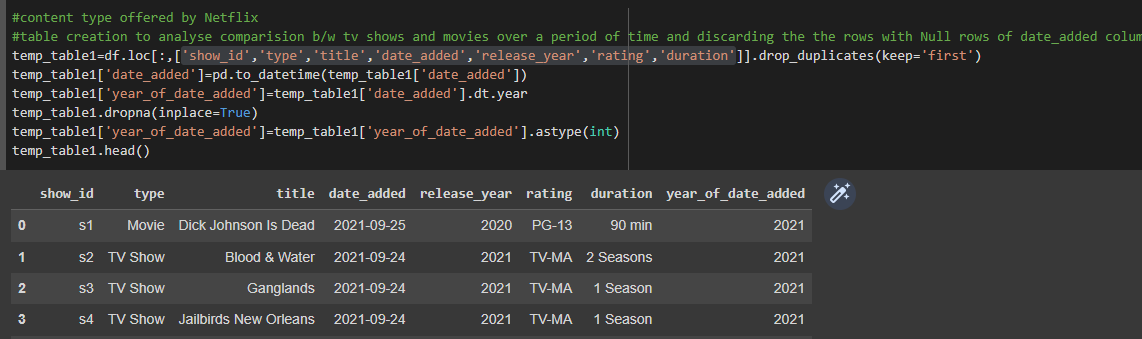


Now the dataframe df will be the table block which will be used for analysis and visualization.

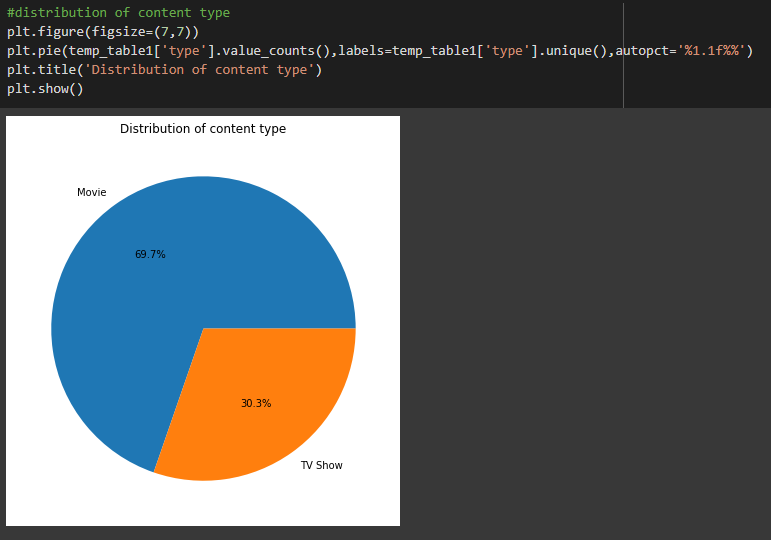
1. **Distribution of Content-Type and rating**

To perform an analysis of the content type and rating we only need a few columns from df table these are: show\_id', 'type', 'title', 'date\_added', 'release\_year', 'rating' and 'duration'.

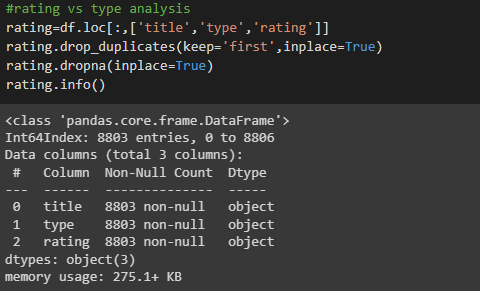
We also need to extract the year form the ‘date\_added’ column to perform the analysis. And drop all the duplicates so that all the rows are unique.



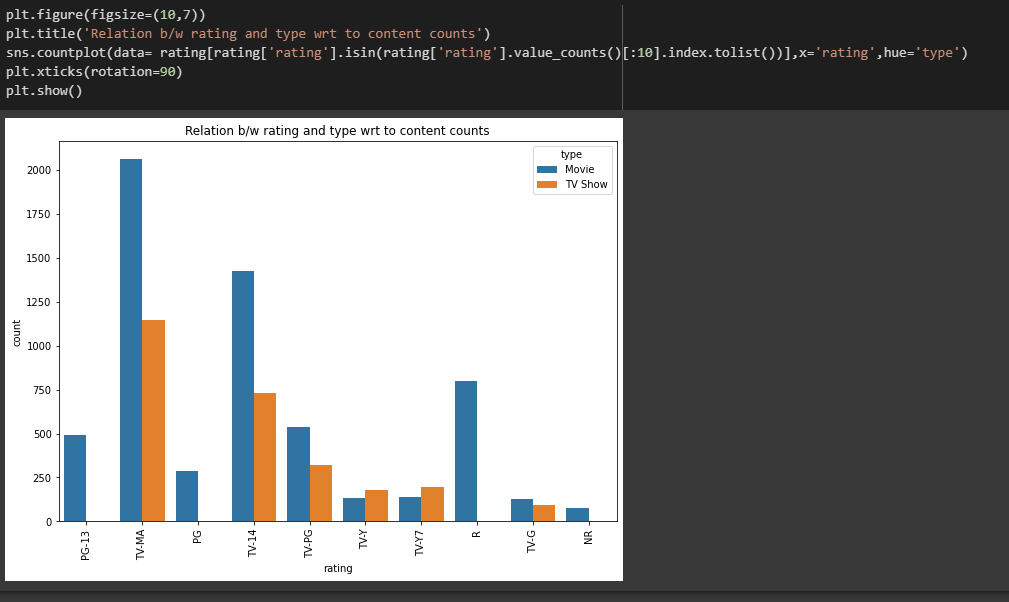
We have learnt from Content-Type distribution that of all the titles released by Netflix approx. 31% are TV show Type and 69% are movie type content



For rating type analysis we only need 3 columns from df table, which are 'title', 'type' & 'rating'. And drop all the duplicate and null values if any.



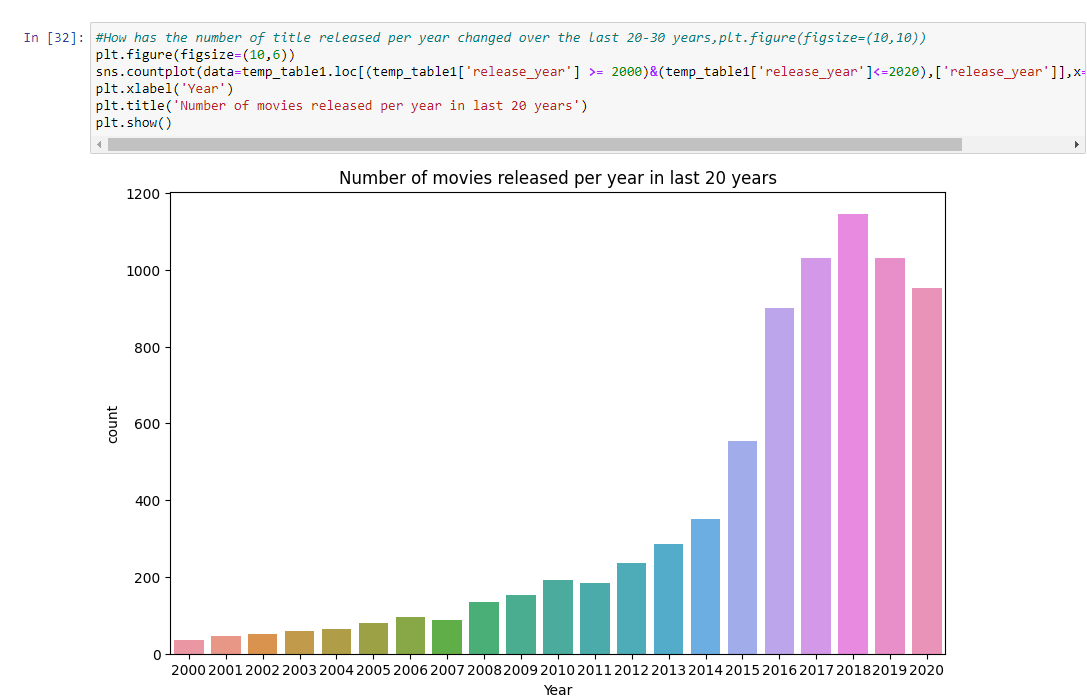
We have learnt that movies and TV shows of TV-MA rating type have the highest no. of content followed by TV-14 and TV-PG. There is another shocking observation that there are no TV Shows on Netflix with content types PG-13, PG and R but there is a good number of Movies present.



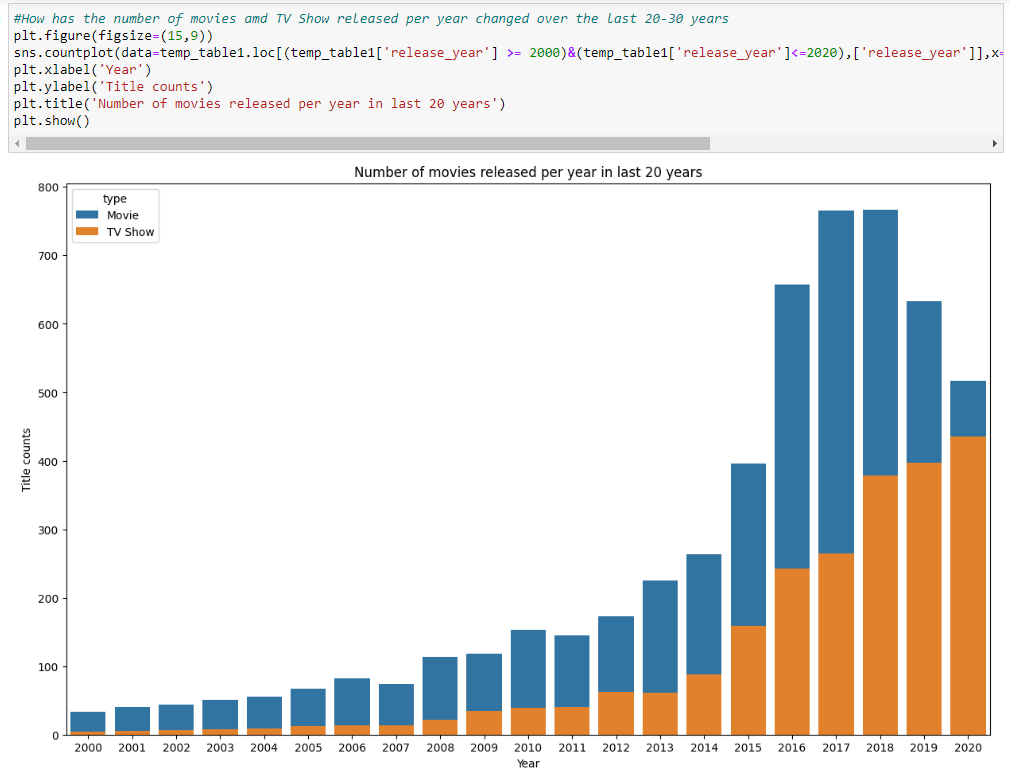
1. **How has the number of titles released per year changed over the last 20 years?**

To analyse this question we are using the same columns which I have mentioned in the content type, which are show\_id', 'type', 'title', 'date\_added', 'release\_year', 'rating' and 'duration'.

For movies we got the following results

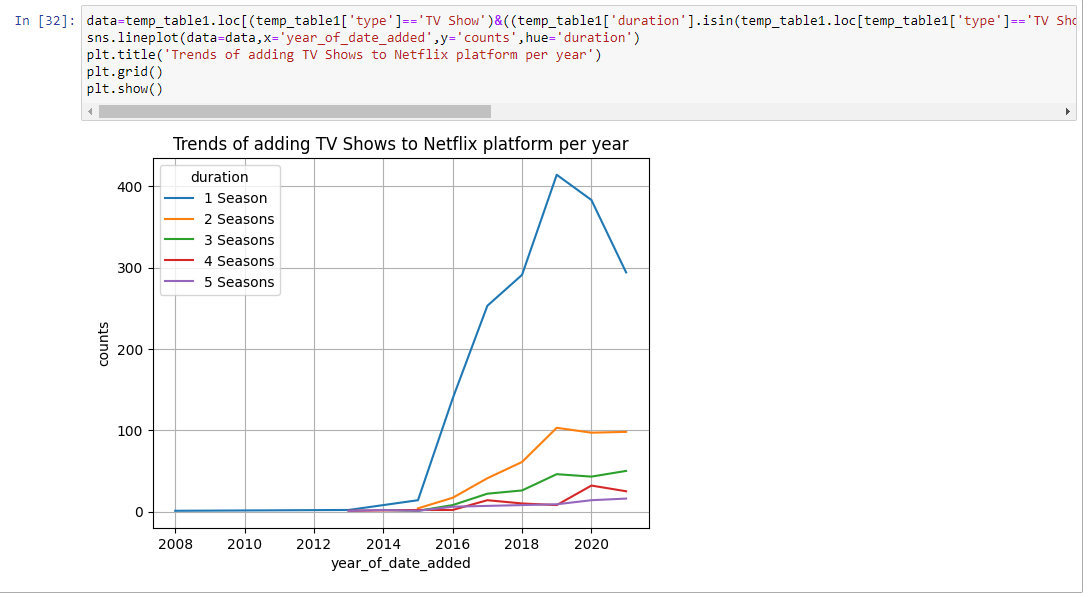


We observe that maximum no of content of movie type was released in the year 2018 and then we see slight decline. But when combine TV show type content over the release year we observe a new trend picking up



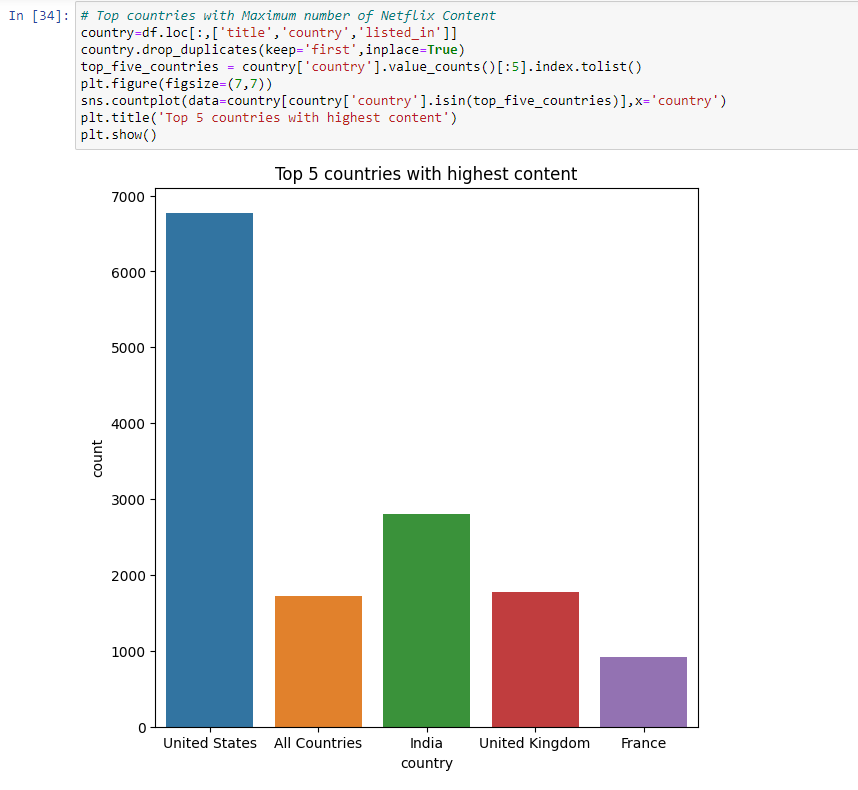
**Here we observe that TV Show type content is getting popular, though we its not large as Movie type content but the time spent by the users in Netflix for TV Show type content are usually more than that of the time spent by the users in Netflix for Movie Type content**

Netflix is also investing lot of time and effort for adding TV show type content, below are the trends



Here we see there is decline in Season 1 Duration type TV Show because few of the TV Shows may have become more popular and Netflix reinvested on same title for new seasons.

1. **Top countries with the Maximum number of Netflix Content**

For analysing we only took 3 columns i.e 'title', ’type’, 'country' and 'listed\_in'. We also dropped duplicate rows. As discussed in the Data Cleaning section, it is assumed that ‘Country’ column with Null values is replace with “All Countries”, which signifies this is common content and is available in al the countries where Netflix is available. 

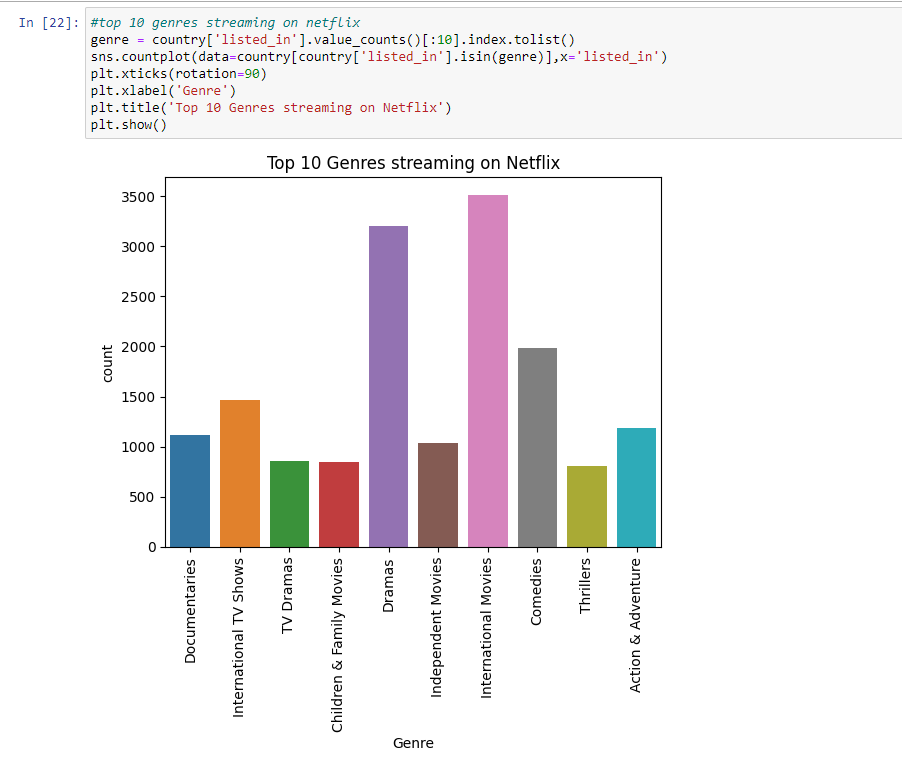
Here we observe that United states, India and United Kingdom are the top 3 countries where Netflix has max number of contents available.



Here we observe a very small but important detail, that **popularity of TV shows in India hasn’t picked up yet (till 2021), where people still prefer to watch Movie type content over TV Show type content.**

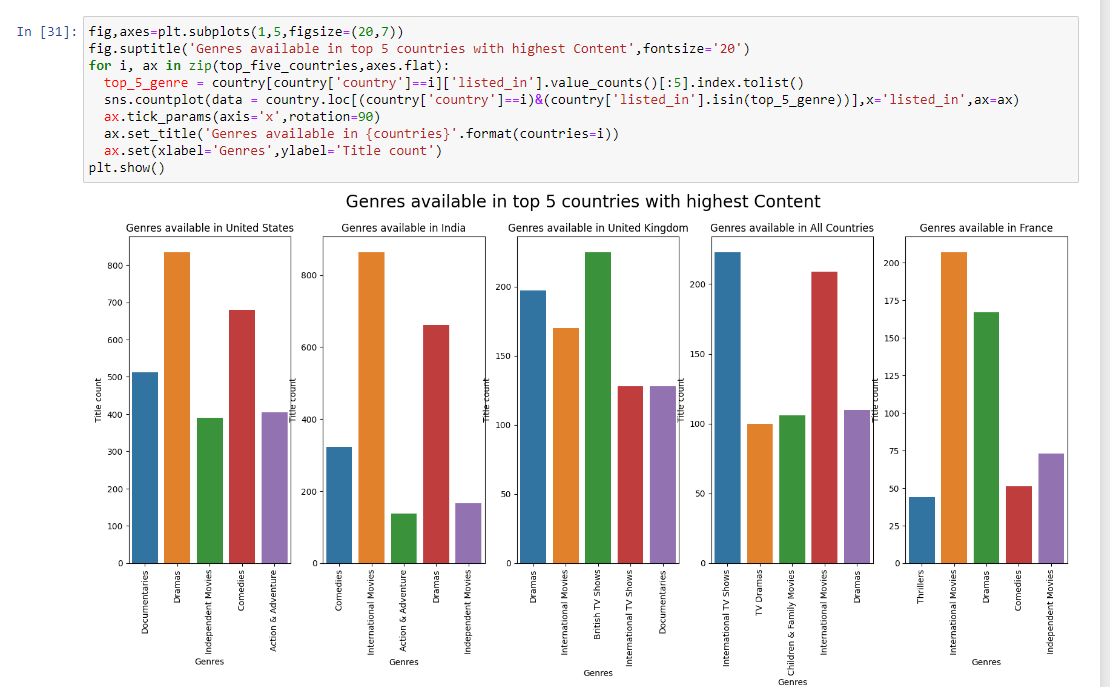
1. **Top 10 genres streaming on Netflix**

To analyse genre distribution, we have considered the same columns as to analyse country-type content distribution. i.e 'title', ’type’, 'country' and 'listed\_in'.



Here we observe that ‘International Movies’ and ‘Dramas’ are the two top Genres which Netflix users like the most, followed by ‘Comedies’ and ‘International TV Shows’.

When we observe top genres consumed by top 5 countries, we observe some like below results

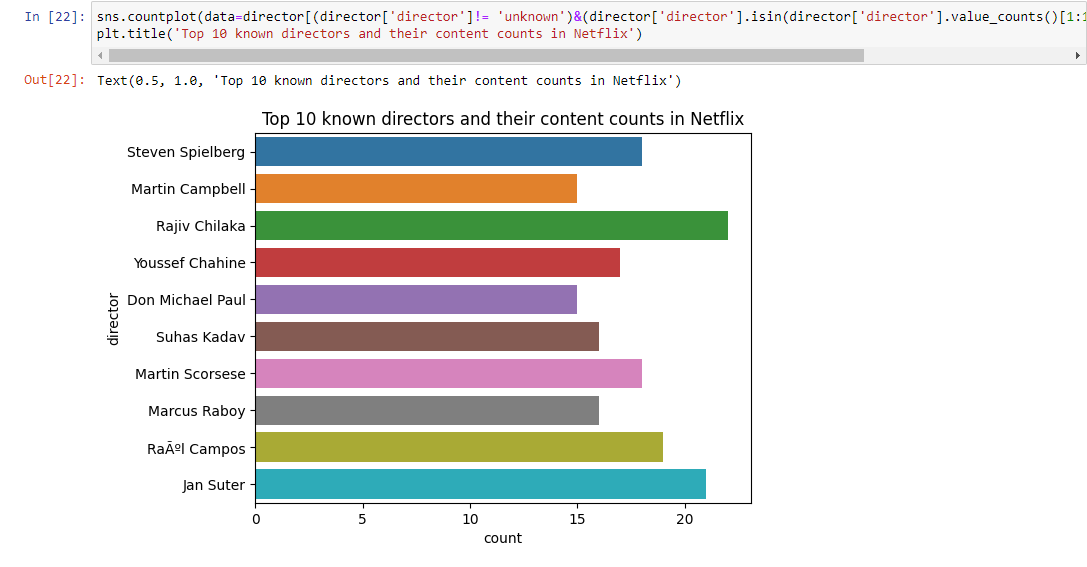


Here we Observe that ‘Dramas’ and ‘Comedies’ are popular Genres in US, ‘International’ and ‘Dramas’ are popular in genres in India, ‘British TV Shows’ and ‘Dramas’ in UK, and ‘International Movies’ and ‘Dramas’ are popular in France.

In General, We have observed that ‘Dramas’ is a popular genre among Netflix Consumers.

1. **Popular director in top countries with the highest content**

For analysing this I have created a temporary table with columns ‘title’, ’director’ and ‘country’ columns, and dropped all duplicates. There are a few directors tagged to a title with Unknown names, which is not considered in the below analysis. We observe that



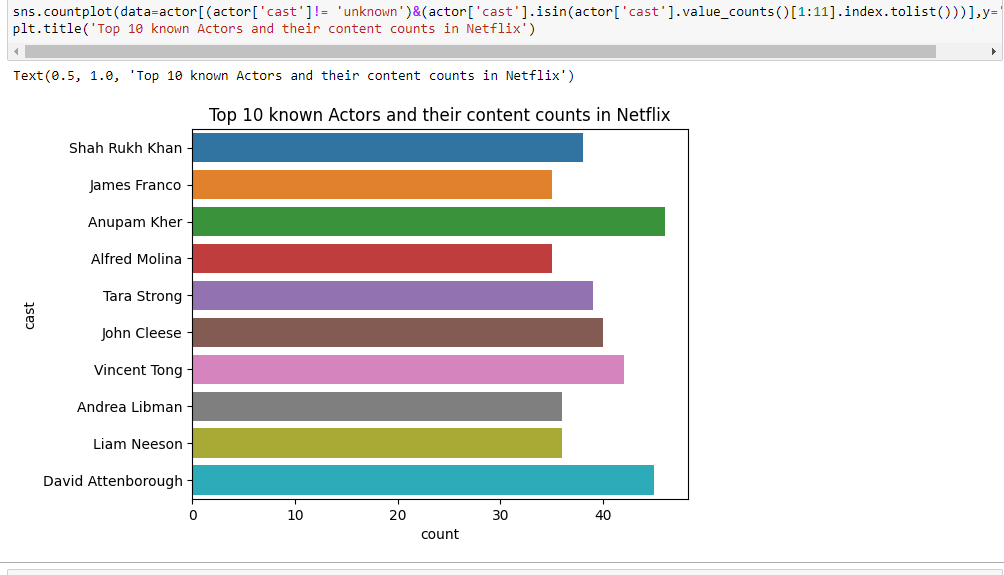
We observe that ‘Rajiv Chilaka’, ‘Jan Suter’ and ‘Steven Spielberg’ are the top 3 directors in creating Netflix Content.

Here are the stats of Top directors were Netflix streams its most content.

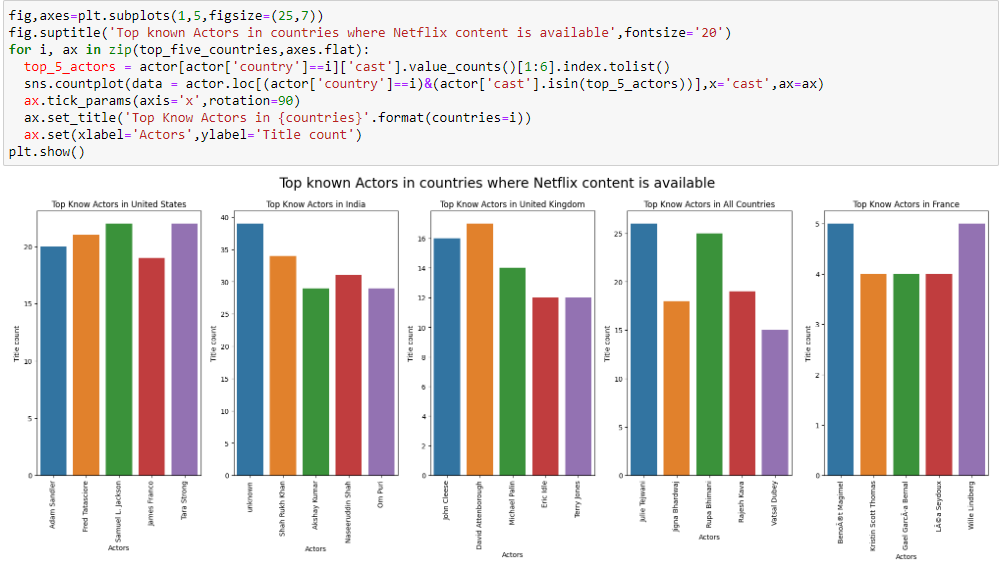


1. **Popular actors in top countries with the highest content**

For analysing this I created a temporary table with columns ‘title’, ’cast’ and ‘country’ columns and dropped all duplicates. There are a few actors tagged to a title with Unknown names, which is not considered in the below analysis. We observe that

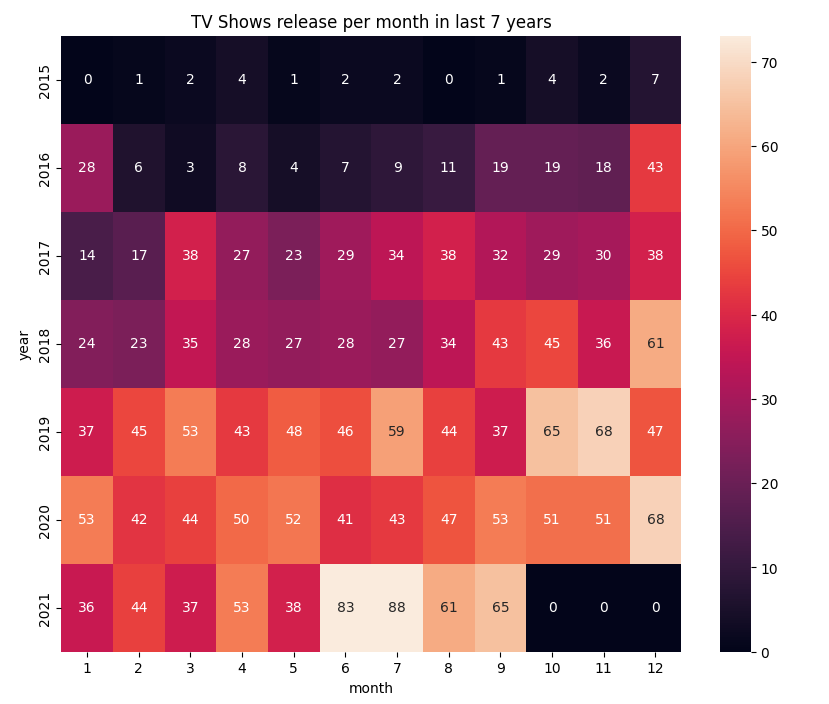


We see that ‘Anupam Kher’, ‘David Attenborough’ and ‘Vincent Tong’ are the top 3 actors whose titles are streaming on Netflix. Here is the country-wise distribution of Netflix cast.



1. **What is the best time to launch a TV show?**

To perform the above-mentioned analysis, column ‘Date\_Added ’ has been converted into datetime type column. And month and year columns are newly created whose origin is from ‘Date\_Added’ column. We have also considered this analysis between 2015 and 2021 and we observe that

We observe the gradual release of content during the months of June, October, November and December from the above heat map.

1. **Recommendations**
2. In order to keep more users engaged in the platform. TV Show type content are the best as it has more screen time than ‘Movie’ type content and the best time to release such type of content are October, November and December
3. Anupam Kher is the most popular actor which users are accepting currently. Netflix should produce titles with Anupam Kher as a cast
4. Rajiv Chilaka Kher is the most popular director which users are accepting currently. Netflix should produce titles with Anupam Kher as a director.
5. Trend for TV Show type content is increasing but the ideal mark is till season 4 after that it is observed that the audience is losing interest.
6. Audiences prefer PG-13, PG and R rating type content in ‘Movies’ type content, not TV Show type content