Name: Tirupati Venkata Sri Sai Rama Raju Penmatsa

Colab link: https://colab.research.google.com/drive/1-umSbvPZZ5t7sV02sq8xm0orBoJGG6zp?usp=sharing

Goal : To perform EDA on a given dataset and to obtain observations from the data.

Dataset : Netflix Dataset (source: https://www.kaggle.com/shivamb/netflix-shows)

Data Preprocessing and cleaning:

When the data was first loaded and checked for fields with null values, we have received a list as the below table, from which we can see that director, country, cast, date_added, rating and duration fields had null values. With director having the most number of null fields.

Labels	No of NA's
director	2634
country	831
cast	825
date_added	10
rating	4
duration	3
show_id	0
type	0
title	0
release_year	0
genres	0
description	0

Some small data cleaning tasks were performed to create a better dataset for the analysis

These cleaning tasks are as described:

1) When going through the fields with “duration” which has fields as null, it was observed that the values got interchanged with “rating”. So all I had to do was replicate the values with the corresponding “rating” field.
   Below table represents the data that has the mismatched values.

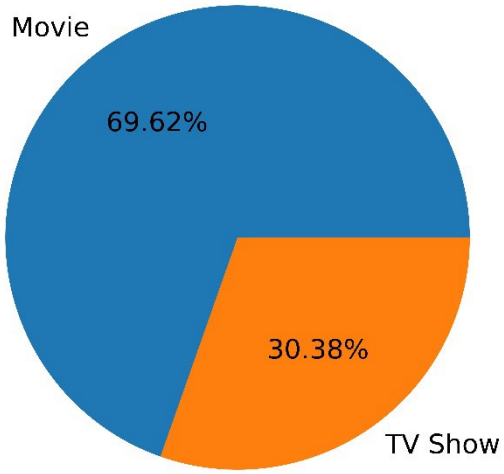
id	title	release_year	rating	duration
8798	Louis C.K.: Live at the Comedy Store	2015	66 min	null
8799	Louis C.K. 2017	2017	74 min	null
8800	Louis C.K.: Hilarious	2010	84 min	null

2) We also had 4 of the movies/tv shows with rating as “null”, which has been changed to “NR” which is one of the ratings that are present in ratings field
3) The null values in “year_added” field was filled with “-1” so as to remove the data while visualizing the charts, also the “year_added” column had data in a different format which can be used to visualize thus data processing was done to convert the fields into datetime and year was extracted from that.
4) In order to get the analysis on fields like “genres” and “cast” which has multiple actors or genres in same row which are separated by “,”, created a new function to count each occurrence after splitting them and give count of each occurrence. This created a list of each occurrence with the count but introduced a new problem of duplicates because of the extra whitespaces.
   So adding a line with .str.strip() to the column fixed the issue of duplicates.

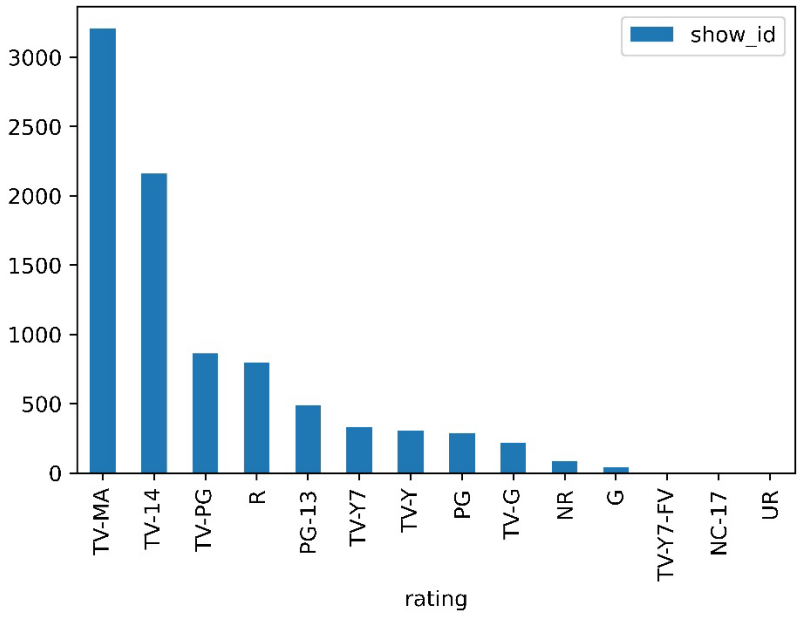
Observations:

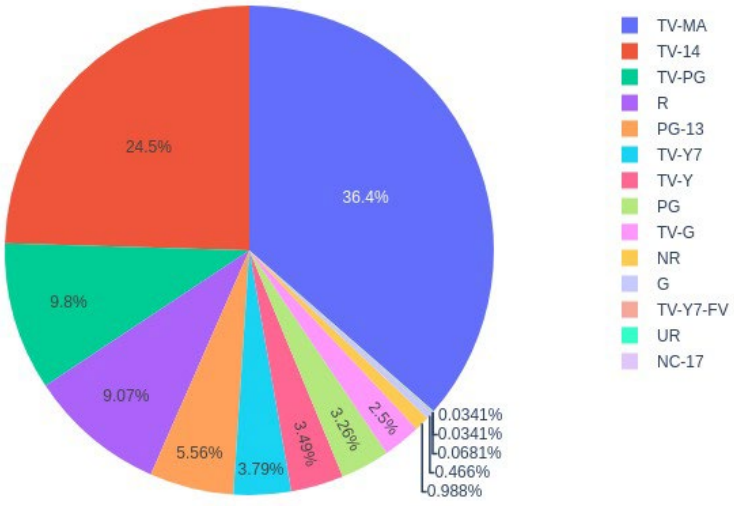
After the above preprocessing of the data, we found out below observations from the chosen dataset:

1) First thing that I had checked was the distribution of the content and percentages of **“Movies vs TV Shows”.** From the below figure we can see that the distribution of movie is 69.62% and TV show as 30.38%



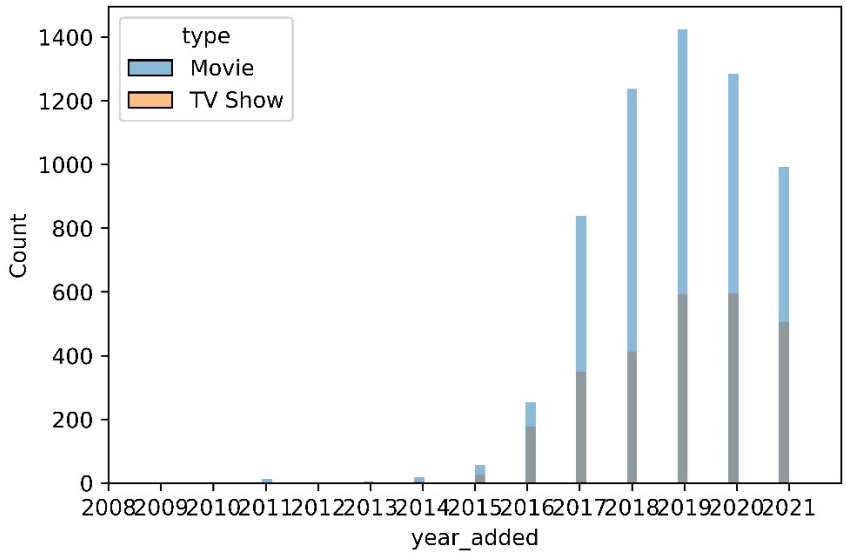
2) Next I found out how the content is distributed in Netflix based on “rating” and got the observation that the content with the most common rating is “TV-MA” with around 3200 Movies and Shows , followed by “TV-14” with 2200 Movies and Shows.



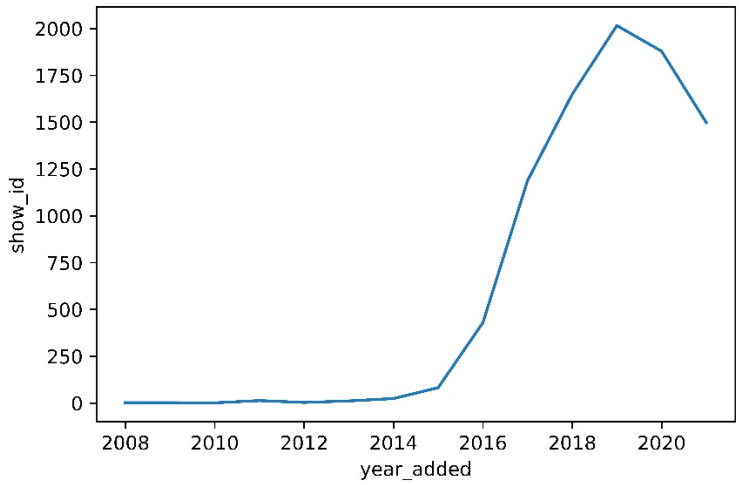


3) Created a word cloud using Seaborn and wordcloud library on “genres” column to find out the most common genres and type, and TV Shows , International Movies, Dramas were some of the most frequent genres and types.

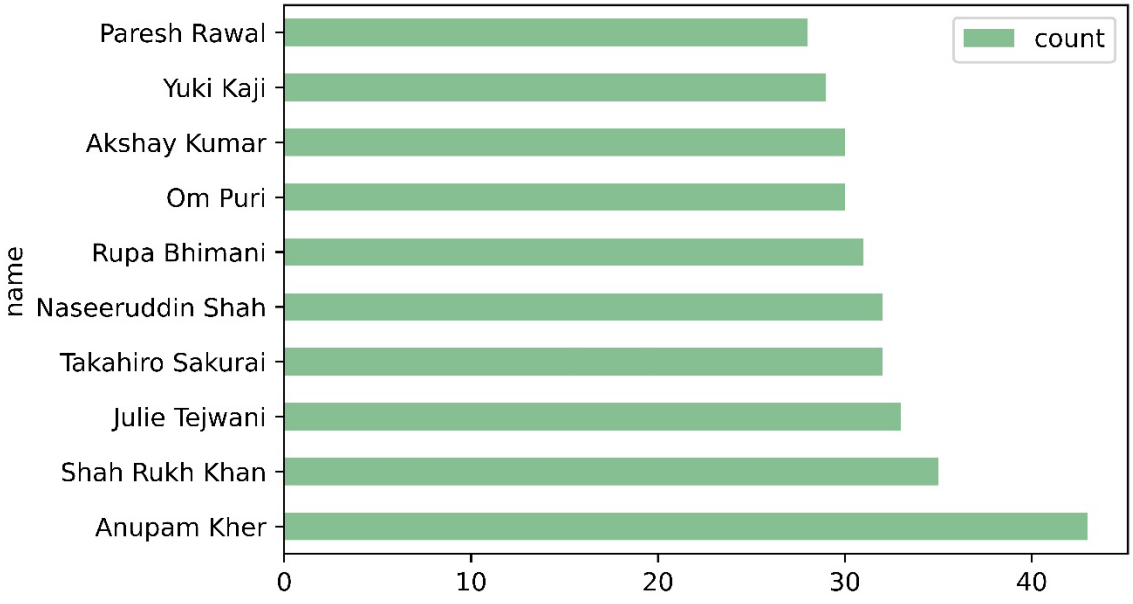
4) Next using Seaborn histplot, analyzed the how many movies and tv shows each were added on yearly basis from 2008 to 2021. Largest year for “Netflix” in terms of content that has been added to the library. In terms of “Movies” 2019 was their biggest addition with almost 1450 movies being added and in terms of “TV shows” 2020 was the biggest year but just slightly more than 2019 with 620 shows as compared to 590 in the previous year.



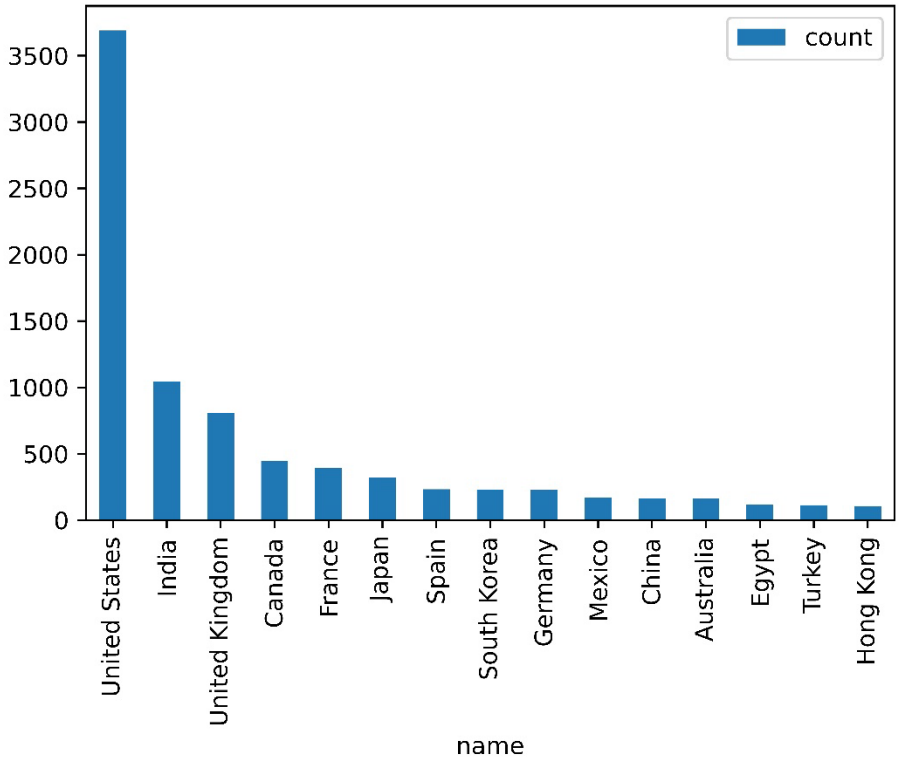
5) Just taking the previous analysis a step further, decided to plot a lineplot using seaborn to find out the overall content addition over the years, to get an overview of rise of the Netflix library over the years. And from the graph it can be inferred that over the years from 2008 to 2014, the content that has been added to the library is very low and almost constant, with 2015 being the year where Netflix according to the graph started investing in the content library, 2016 being the year with almost thrice the amount of content added compared to previous year and 2017 being an exponential growth once again and finally 2019 is the peak year for Netflix in terms of content addition with more than 2000 titles that have been added.



6) Using the previously created function to find the count of each actor, analyzed the top 10 actors with most number of titles with them being on cast. And Anupam Kher has the most number of titles with him being part of the cast with more than 40 movies and next highest is Sharukh Khan with around 35 titles. Below is the visualization of the graph.



7) Next to find out the variety of movies and tv shows based on country, plotted out a graph. From the graph we can see that United States gives the most content with more than 3500 titles, followed by India wiith close to 1100 titles and then from United Kingdom with just shy of 750 titles.



8) Created a country specific subset from the Netflix Dataset and chose the country as “India”, to find out Indian Movies vs Indian TV Shows in the Netflix library. We can infer from the below that there significantly lower number of TV Shows from India compared to Movies.