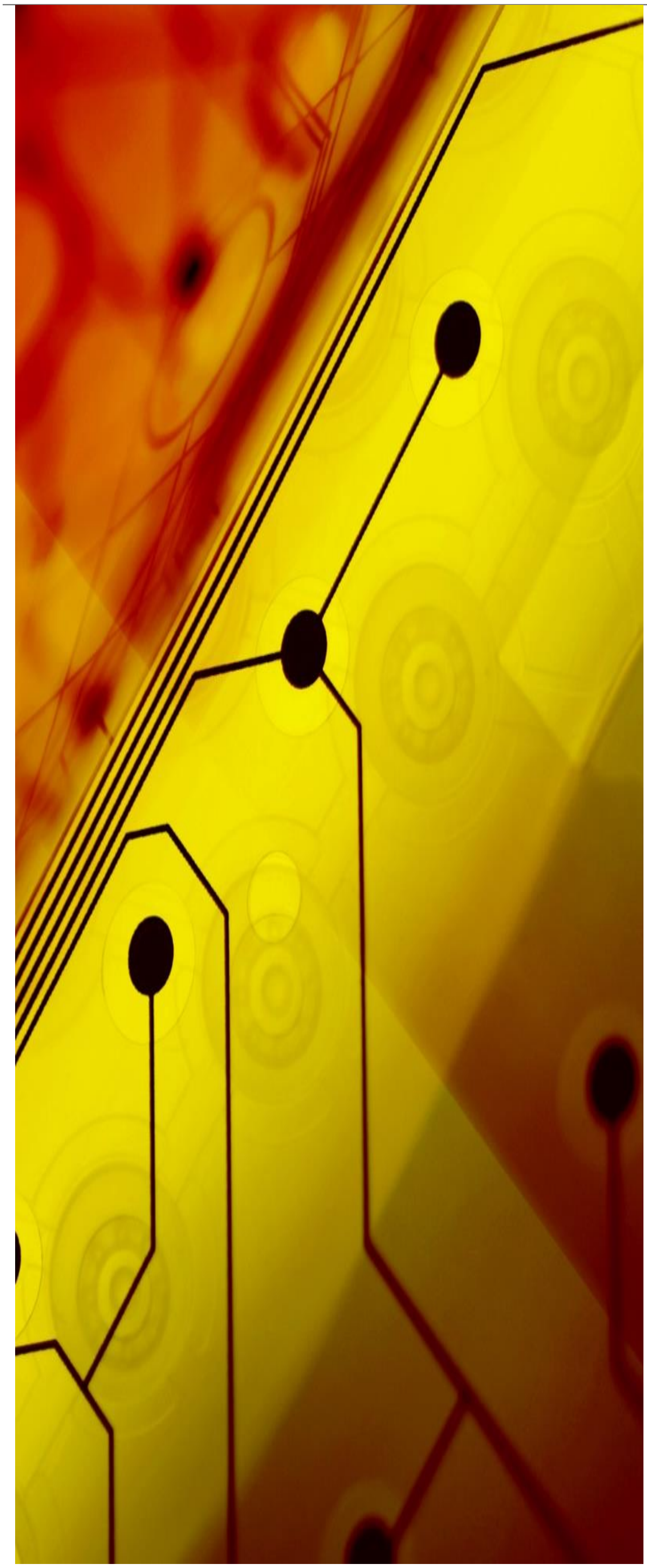


Netflix, Visualization and EDA

AKANKSHA PORWAL

MSC - DATA SCIENCE

SEM-I

202118017

Instructor– Mr. Nishith Kotak

11-12-2021

• Problem Definition:

- ❖ Netflix is one of the largest providers of online streaming services. It collects a huge amount of data because it has a very large subscriber base.

- ❖ Right now, we are living in **“an age of Big data”** trillions of rows of data are being generated every day.

- ❖ With a company valuation of over $164 billion, Netflix has surpassed Disney as the most valued media company in the world.

- ❖ As the company have surpassed 150 million subscribers it is important for that they should increase the customer retention rate, recommend users based on what they prefer more be it be specific genres or age-group and for this the data stored should be analysed.

• Problem explanation:

❖ We can analyse a lot of data and models from Netflix because this platform has consistently focused on changing business needs by shifting its business model from on-demand DVD movie rental and now focusing a lot about the production of their original shows.

❖ Also visualizing this data helps us in curating data into a form that is easily understandable and also helps in highlighting a specific portion. Plain graphs are too boring for anyone to notice and even fail to keep the reader engaged.

❖ Hence, I have created some good visualizations using matplotlib, seaborn and other visualization libraries of python.

- **The project consists of: -**

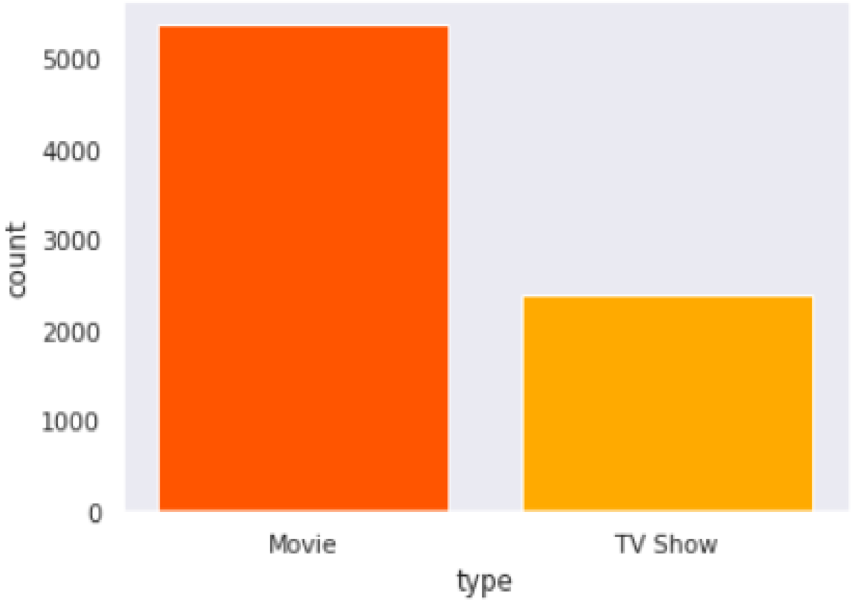
I. Importing libraries
II. Downloading file from Kaggle to environment
III. Pre- Processing
IV. Basic information from the dataset
V. Exploratory Data Analysis

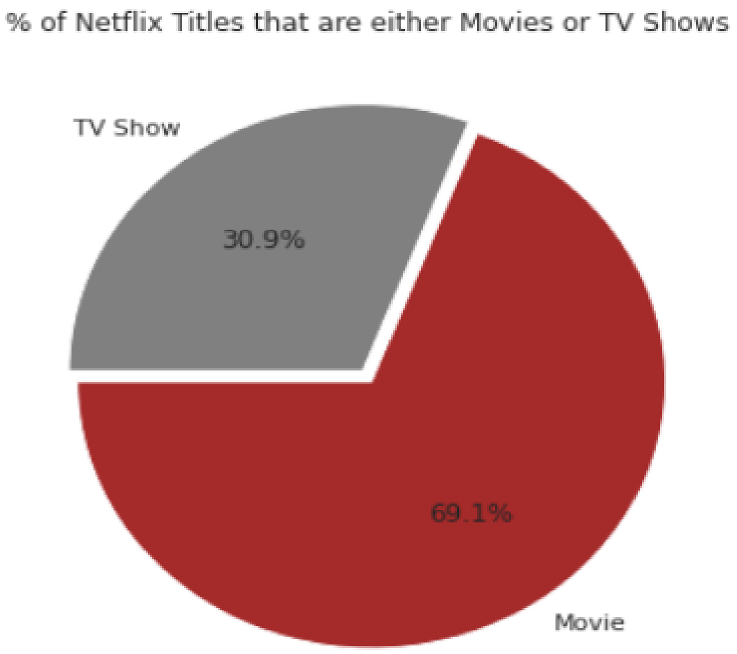
- **Charts/maps Used for EDA:**
  1. Count plot by Seaborn
  2. Heatmap by seaborn
  3. Pie chart by matplotlib
  4. Funnel plot by plotly. Express
  5. Kde plot by seaborn
  6. Bar plot by seaborn

•Result Analysis performed:-

Inference 1:

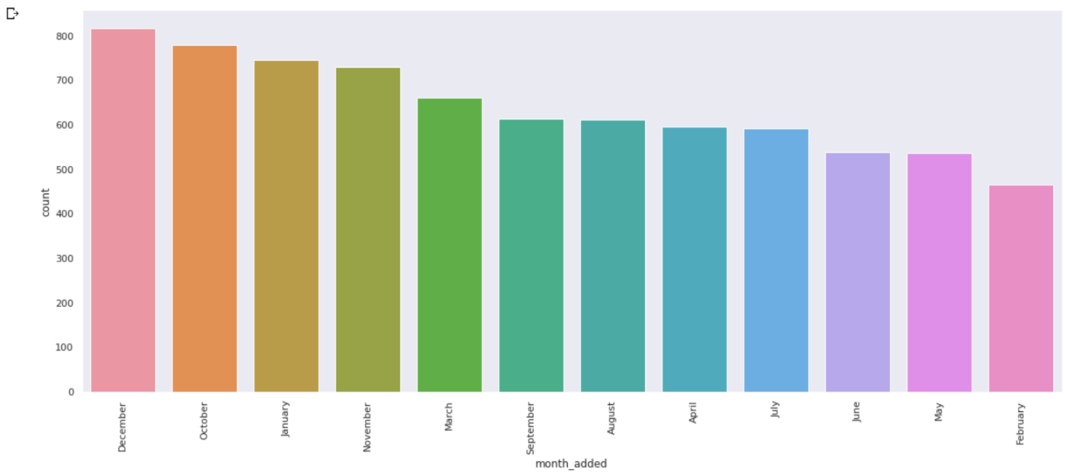
We found out that there are more movies on Netflix than T. V.
Shows.Here, 30.9% of Netflix titles are TV Shows and 6 9.1% are Movies.





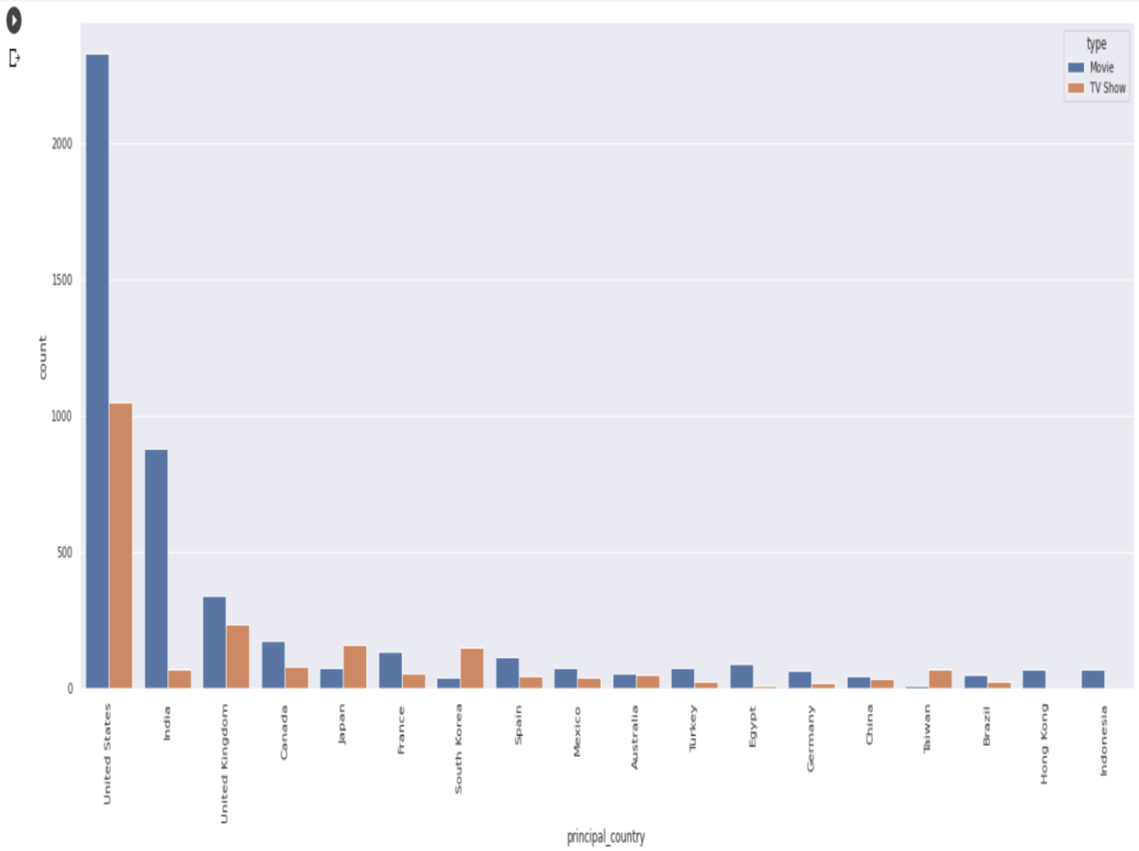
Inference 2:

Most of the directors prefer to release their Movies & Tv Shows in December. Since December is the Month of Vacations.

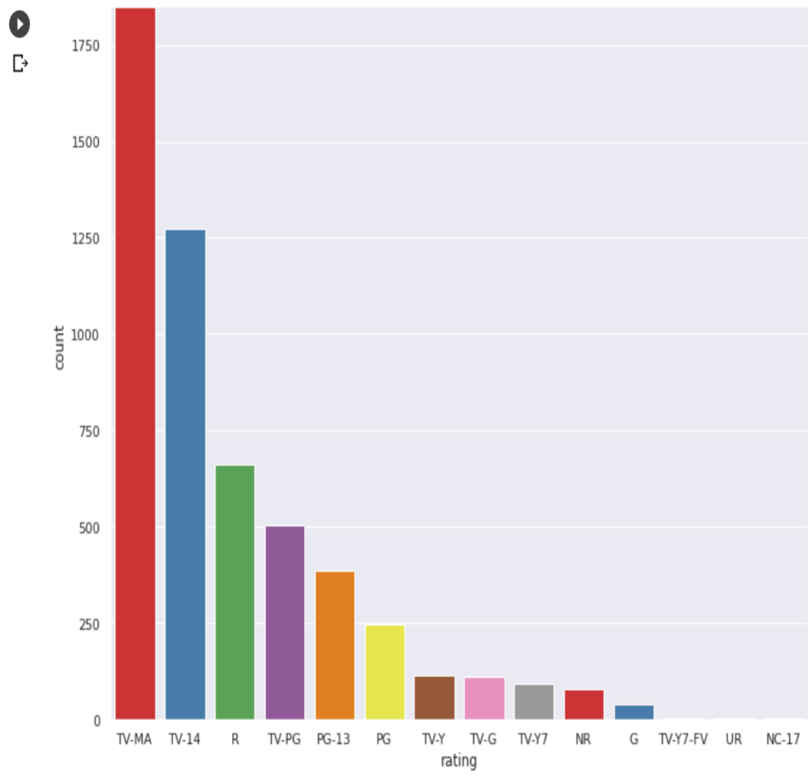


Inference 3:

United States provides the Highest number of Movies & Tv Shows, then at 2nd place India provides the Highest number of Movies.



Inference 4:
Movie Rating Analysis: -

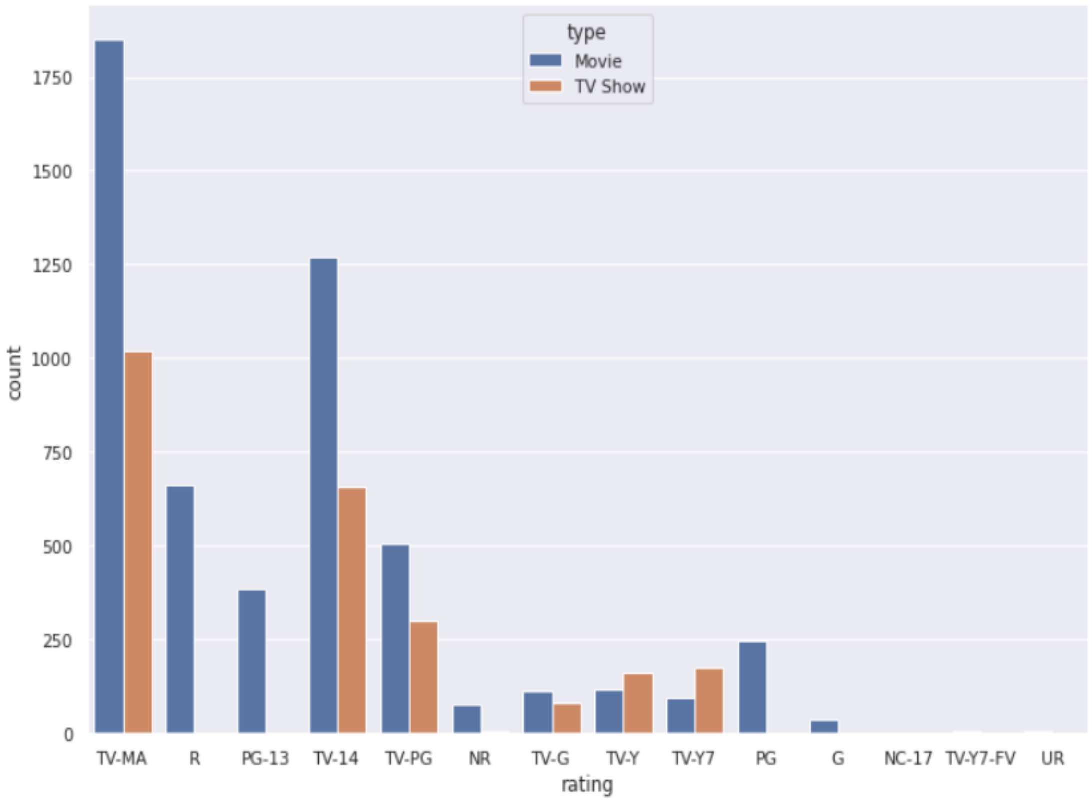


- TV-MA:This program is specifically designed to be viewed by adults and therefore may be unsuitable for children under 17.
- TV-14:This program contains some material that many parents would find unsuitable for children under 14 years of age.
- TV-PG:This program contains material that parents may find unsuitable for younger children.
- R:Under 17 requires accompanying parent or adult guardian,Parents are urged to learn more about

the film before taking their young children with them.

- PG-13:Some material may be inappropriate for children under 13. Parents are urged to be cautious. Some material may be inappropriate for pre-teenagers.
- NR or UR:If a film has not been submitted for a rating or is an uncut version of a film that was submitted
- PG:Some material may not be suitable for children,May contain some material parents might not like for their young children.
- TV-Y7:This program is designed for children age 7 and above.
- TV-G:This program is suitable for all ages.
- TV-Y:Programs rated TV-Y are designed to be appropriate for children of all ages. The thematic elements portrayed in programs with this rating are specifically designed for a very young audience, including children ages 2-6.
- TV-Y7-FV:is recommended for ages 7 and older, with the unique advisory that the program contains fantasy violence.
- G:All ages admitted. Nothing that would offend parents for viewing by children.
- NC-17: No One 17 and Under Admitted. Clearly adult. Children are not admitted.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f51a05a3e90>
```



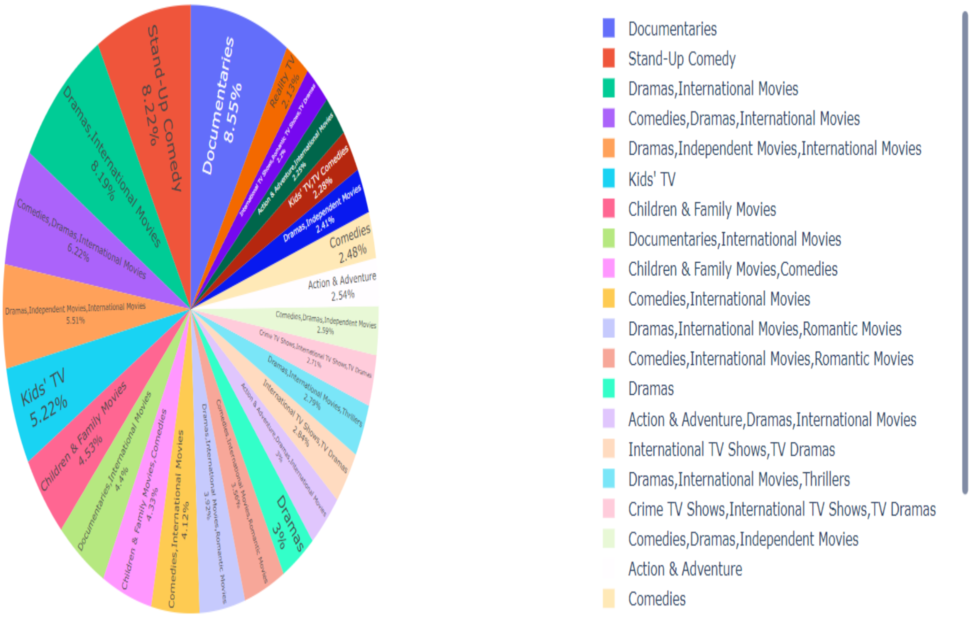
The largest count of movies are made with the 'TV-MA' rating. “TV-MA" is a rating assigned by the TV Parental Guidelines to a television program that was designed for mature audiences only.

Second largest is the 'TV-14' stands for content that may be inappropriate for children younger than 14 years of age.

Third largest is the very popular 'R' rating. An R-rated film is a film that has been assessed as having material which may be unsuitable for children under the age of 17 by the Motion Picture Association of America.

Inference 5:

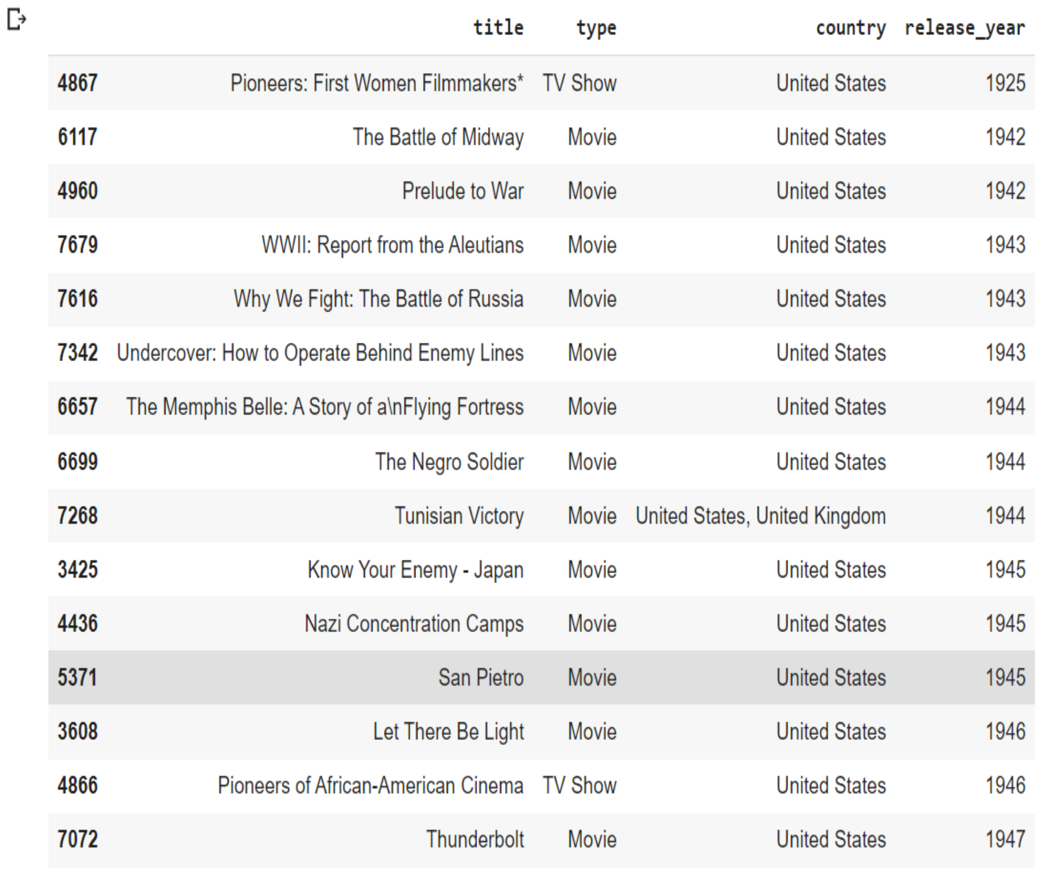
Netflix provides "Documentry" type Movies & TvShows most then in the 2nd place it provides Stand Up Comedy most.



Inference 6:

Oldest movie available on Netflix was: -

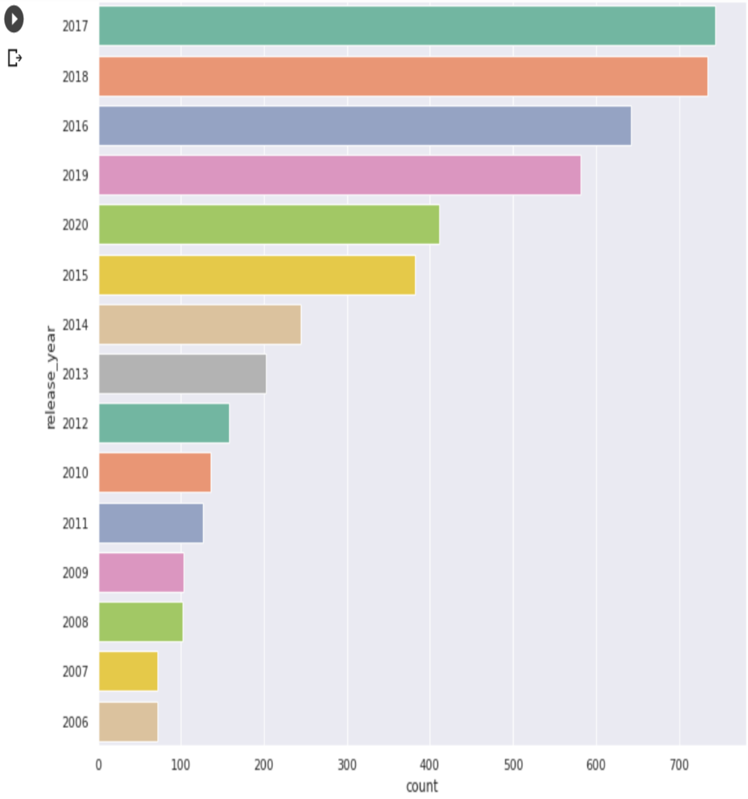
Pioneers: First Women Filmmakers, was a TV Show by U.S and released in the year 1925. Also, all of the oldest Movies & TV Shows on Netflix are from United State.

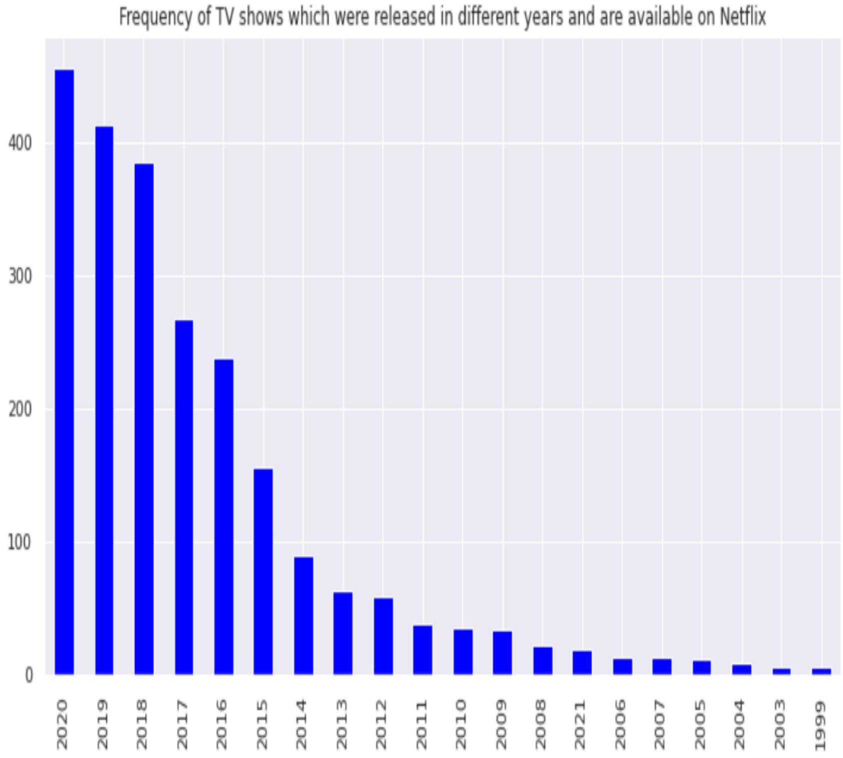


	title	type	country	release_year
4867	Pioneers: First Women Filmmakers*	TV Show	United States	1925
6117	The Battle of Midway	Movie	United States	1942
4960	Prelude to War	Movie	United States	1942
7679	WWII: Report from the Aleutians	Movie	United States	1943
7616	Why We Fight: The Battle of Russia	Movie	United States	1943
7342	Undercover: How to Operate Behind Enemy Lines	Movie	United States	1943
6657	The Memphis Belle: A Story of a\nFlying Fortress	Movie	United States	1944
6699	The Negro Soldier	Movie	United States	1944
7268	Tunisian Victory	Movie	United States, United Kingdom	1944
3425	Know Your Enemy - Japan	Movie	United States	1945
4436	Nazi Concentration Camps	Movie	United States	1945
5371	San Pietro	Movie	United States	1945
3608	Let There Be Light	Movie	United States	1946
4866	Pioneers of African-American Cinema	TV Show	United States	1946
7072	Thunderbolt	Movie	United States	1947

Inference 7:

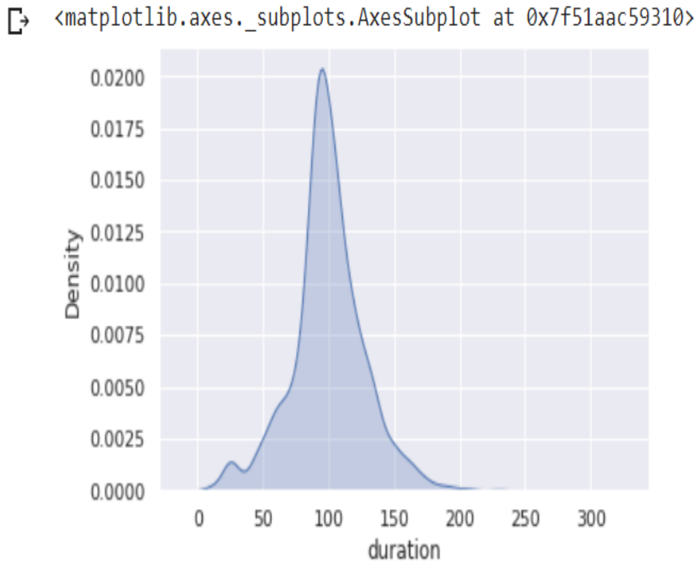
2017 was the year when most of the movies were released and 2020 was the year when most of TV shows were released.


2017
2018
2016
2019
2020
2015
2014
2013
2012
2010
2011
2009
2008
2007
2006
release_year
0
100
200
300
400
500
600
700
count


Frequency of TV shows which were released in different years and are available on Netflix
400
300
200
100
0
2020
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2021
2006
2007
2005
2004
2003
1999

Inference 8:

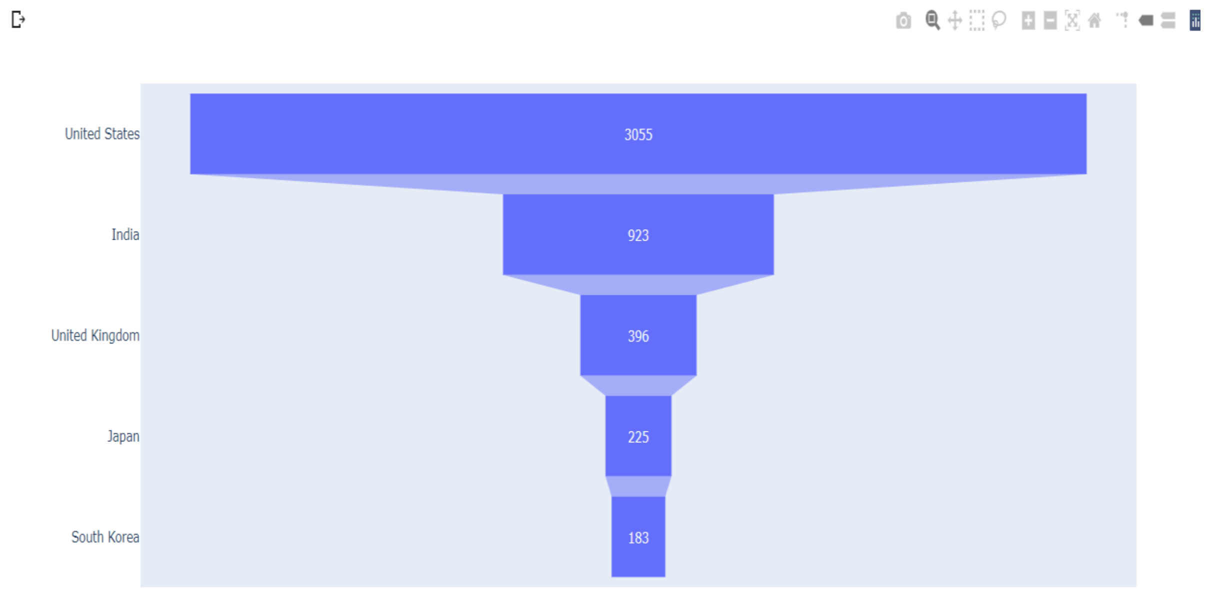
A good number of movies on Netflix are among the duration of 75-120 mins. It is acceptable considering the fact that a fair amount of the audience cannot watch a 3-hour movie in one sitting.



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f51aac59310>
```

Inference 9:

Naturally, United States has the most content that is created on Netflix in the tv series category.

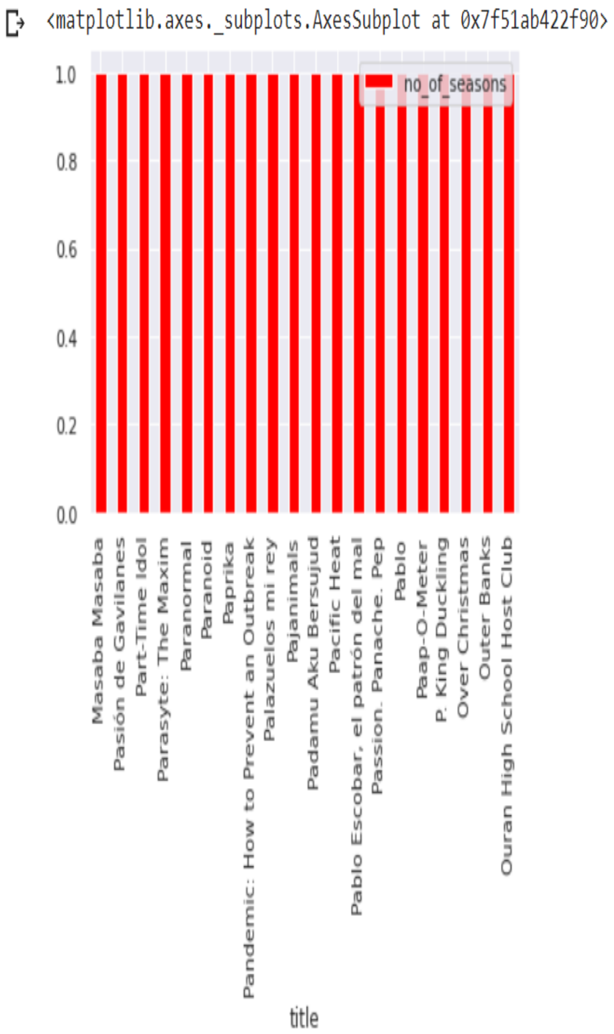


Inference 10:

NCIS, Grey's Anatomy and Supernatural are amongst the tv series that have highest number of seasons.

Inference 11:

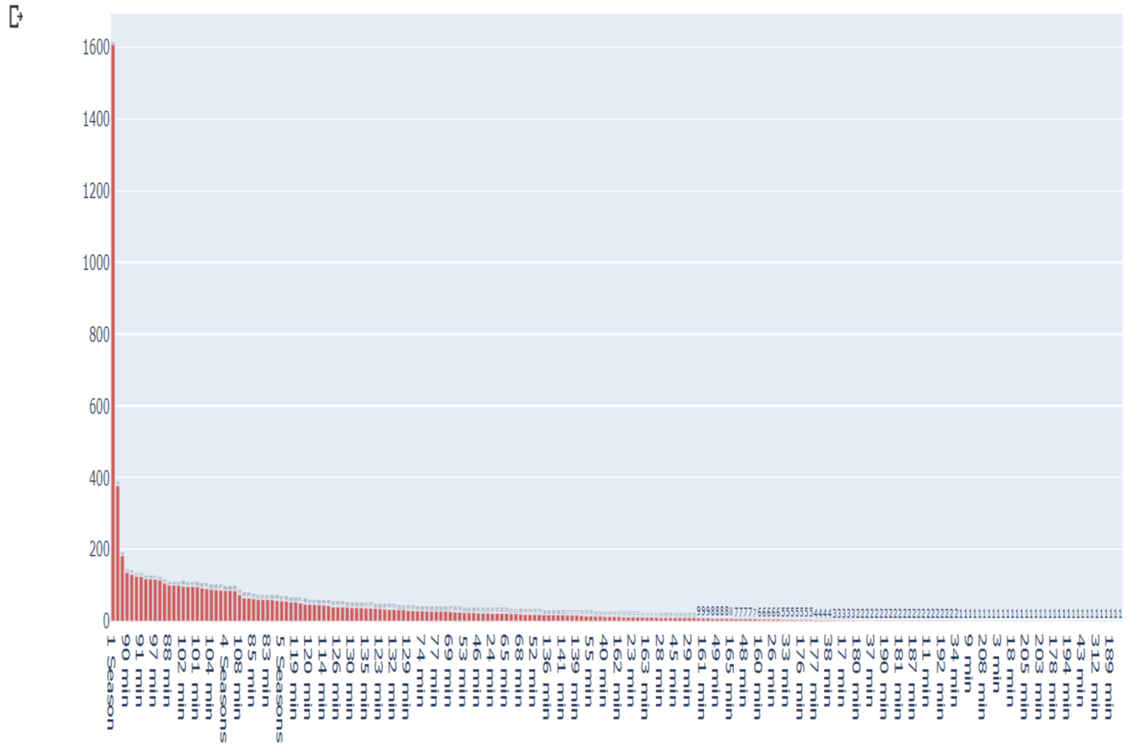
These are some binge-worthy shows that are short and have only one season like Masaba-Masaba, Outer Banks etc.



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f51ab422f90>
```

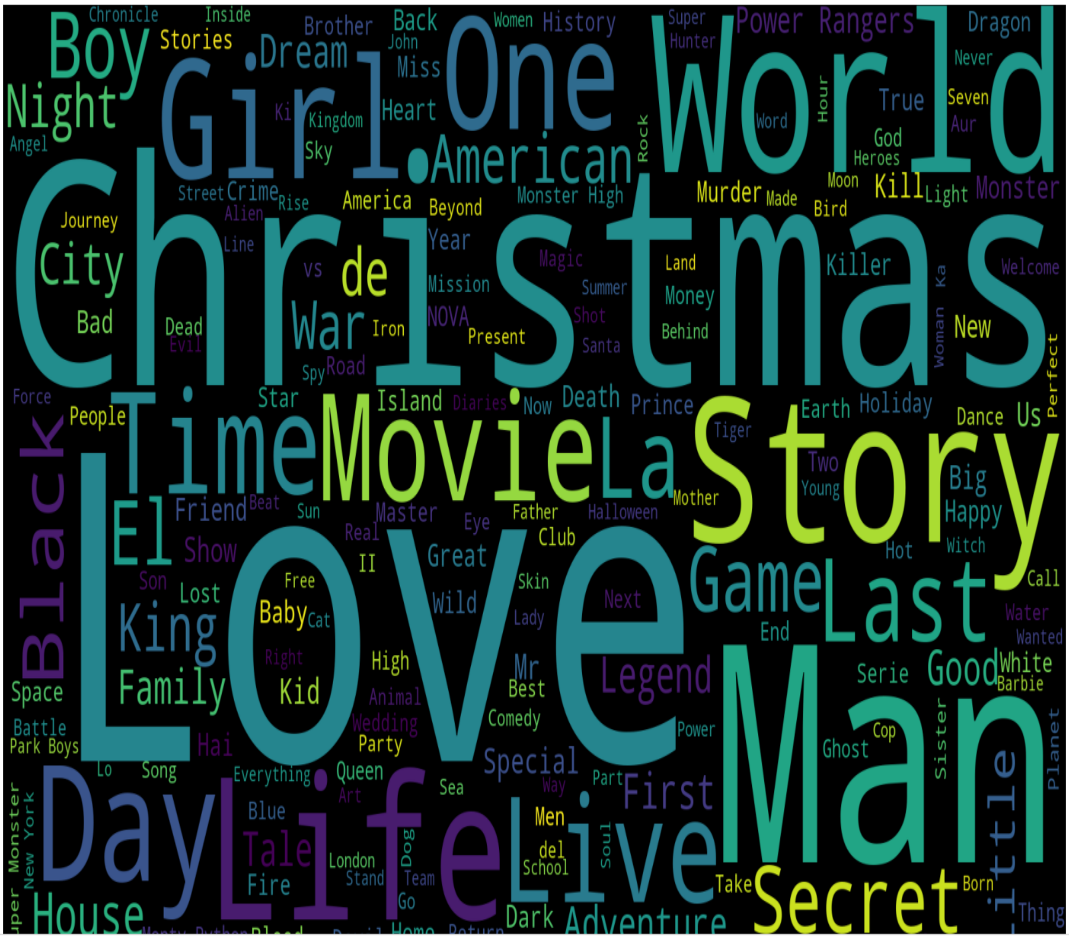
Inference 12:

Having one season is the most preferred duration.



Inference 13:

A visual representation of text that’s based on the freq uency with which a particular word appears



Detailing of steps performed : -

1.Importing the relevant libraries

(I)NumPy- useful in performing operations that are related to linear algebra and for its handling of random numbers, can effectively implement multi-dimensional array objects or reshaping of matrices.

(II)Pandas- python library used for faster data analysis, data cleaning & data pre-processing.

(III)matplotlib- for plot or figure manipulation

(IV)seaborn- is a library mostly used for statistical plotting and it's built on top of matplotlib & provides beautiful styles & colour palettes.

(v)warnings-
lets the user indicate what should happen with differen
t types of warnings

2)Loading the datasets

- ✓ Using pandas read csv method to load our csv file into Netflix data variable. Also, we are checking the first five rows of our data frame with the head command.

3)Getting a short summary of the data

- ✓ .info () method gives us the number of columns, their datatype and count of non-null values

4)Data Cleaning

- ✓ It’s necessary to first deal with the missing values because they are not same as data values. A null value is basically an undefined value of no use.

- ✓ So for that we can check unique value of each column, and them taking the sum of values which are null for all the 12 column’s by **isna().sum()** method and also heatmap which clearly shows the columns which are having null values.

- ✓ **.drop()** method tells python to remove column(s) ,axis=1 tells python that you want to apply function to columns instead of rows.

- ✓ Therefore, we first check unique values for rating and display only nan values. Next, we are only displaying only nan values back in our data frame.

- ✓ We have TV-MA which is the most common rating and hence we can replace all nan values with TV-MA. As we know mode takes only the value with higher number of occurrences so we can replace nan values with it.

- ✓ Now in the fix date_added missing column, I am keeping only not null values back in our data frame thus keeping aside the null values.

- ✓ In the same way country columns mode was USA as Netflix was created in USA and every show is aired on Netflix US only.
- ✓ Now we finally check if our data is cleaned by isna (). sum () and yes 0’s indicate that they are cleaned.
- ✓ Now we are creating partitions for Netflix data as in Netflix shows which will contain only movies data. We can also that check with use of head command.
- ✓ Now in the plot which is seaborn’s countplot we can clearly see that there are more movies on Netflix than TV shows.
- ✓ To better analyse the month and years which are valuable I have created a new column called year and month added. As date_added column was consisting of both year and month we have splitted the values by use of **split()** function.
- ✓ Now, there are specific names or meaning behind each rating, like for example TV-MA rating, are given to program specifically designed to be viewed by adults and therefore may be unsuitable for children under 17 and so on there was a specific age or restriction for each type of rating so that we could know which is target mostly focused by productions. Here can see the usage of **replace**

() method which replaces given value with the new value passed on to the function.

- ✓ Moving forward to country fixups. There are some countries which has multiple values like USA, Brazil or USA, Canada etc so for that I have created a new column principal_country which consists of only first-one so we can check with regions having more productions.

5)Fixing Data Types

- ✓ **.dtypes()** method will provide us the data type of each column. Now we can see that type column should be categorical. So, we are converting it's data type to categorical using pandas **.to_categorical()** method passing in the type column.
- ✓ Similarly for target_ages column and year we have converted to integer using pandas **to_numeric()** function.

6)Data Visualization

- ✓ We have used matplotlib's **figure()** function to create a figure object and we have specified the width and height in pixels using the figure parameter.

- ✓ We have given a title, and we have created a pie chart using function **.pie(),** where we have to pass the dataframe's type column counts and other formatting parameters using **.show()** we can see 30.9% of Netflix titles are TV shows and 69.1% are movies.
- ✓ Next, we are checking which month the directors prefer most to release Movies and TV Shows.
- ✓ Here, also we are using figure function, then **countplot()** function, order is the main argument here where we have added counts of each month.
- ✓ Matplotlib **xticks()** allow us to rotate labels as per our chart.
- ✓ Here we can see most of directors prefer to release their movies and tv shows in the month of December. Since it's the month of vacations.
- ✓ Now to see which state has provided highest number of movies and tv shows again we are using countplot function and here we have specified which column to take, it's counts.
- ✓ So basically, US provides the highest number of movies and Tv shows to which India is the 2nd, similarly for rating, Tv and movies has highest for TV-MA rating.

- ✓ Now with the help of **plotly.express** we are creating a pie chart to see which type of content is most provided by Netflix.
- ✓ We are using genre columns, **value_counts()** for 25 types of content.
- ✓ Next we have specified the data, values, names and labels with the help of **update_traces()** method we have specified textposition inside and info+percent as label.
- ✓ Documentaries constitute 8.55% of the whole content which is the most and second, we have stand-up comedy.
- ✓ Now we are checking oldest movies available on Netflix, first we have sorted the values according to release year using **sort_values()** method of data frame . These are 15 oldest movies , oldest was released in 1925.One thing is to note that almost all from US.
- ✓ Next we have listed total shows by a particular country using pandas **Dataframe()** method which create two dimensional potentially heterogeneous tabular data which does consists of count of each appearing with the .**reset_index()** method for descending counts.
- ✓ Next is one implementation of funnel charts which takes a dictionary y as an input and same

dictionary values on x and y axis. It helps us to analyse better and it's mostly used in different stages of a business process.

- ✓ Next is the year wise analysis of Netflix movies and we can see 2017 was the year when most of the movies were released and in the same way 2020 was the year when most of the Tv shows were released.
- ✓ Next is one implementation of word cloud, it's just a visual representation of text that is based on the frequency with which particular word appears.
- ✓ Here I am using matplotlib's **subplots()** function to first specify height and width in pixels . Also, word cloud library gives functionality for creating wordcloud.
- ✓ In the **wordcloud()** function we have to specify the background,color,width,height and next applying .generate() function on Netflix_data.title column.
- ✓ Next with matplotlib's **imshow()** function we can show our created wordcloud , we have kept axis as off otherwise it would have created a grid over this and have specified the x and y axis.
- ✓ Next is to save particular wordcloud in a file called cast.png using **savefig()** method.

- ✓ Next we have saved the duration of movies in duration variable again by replacing min with null space and then converted str to int for duration.
- ✓ Next with **sns.set()** we have set background style of darkgrid and this is the kde (kernel density plot for duration , this parameter shade=True creates and filled plot.
- ✓ So, a good amount of movies on Netflix are of duration 75-120 min. But then we also know a fact that a fair amount of audience cannot watch a movie in a sitting of 3 hours.
- ✓ Next we can TV Shows with largest TV No. of seasons are NCI'S, Grey's Anatomy and supernatural, we have plotted and sorted top 20 values.
- ✓ For the top duration we are making use of plotly.graph_objects, in the top Duration we are using pandas **Value_counts()** method to count duration then using graph objects we have created a figure which is a Bar plot, using **update_traces()** method we have specified formatting of text.
- ✓ Also we can see that there are some binge worthy shows that only have 1 season.
- ✓ Coming to the top duration we can see having one season is the most preferred duration.

Conclusions: -

It was a wonderful and learning experience for me while working on this project. The joy of working and the thrill involved while tackling the various problems and challenges gave me a feel of the industry. I enjoyed each and every bit of work I had put into this project.

In conclusion, the sole purpose of Netflix or any other company should not be only to make profits. They should care about the users as well, because after all, users use the companies' resources. Those companies that do not care, act responsibly, or are not ethically driven, are less likely to survive in the long run or make profit.

Future Work: -

I can create powerful analytic models using machine learning. These models can process terabytes of data to churn out meaningful information. Judicious use of data analytics is the main reason for a company's success. In fact, big data & analytics are so vital to company's success that you may as well call them an analytics company instead of a media company.

I can make a recommendation system for users using various algorithmic approaches like reinforcement learning, neural networks, causal modelling, probabilistic graphical models, matrix factorization, ensembles, bandits.

The recommendation system will estimate the probability of a user watching a particular title based on the following factors –

- Viewer interactions with Netflix services like viewer ratings, viewing history, etc.
- Information about the categories, year of release, title, genres, and more.
- Other viewers with similar watching preferences and tastes.
- Time duration of a viewer watching a show
- The device on which a viewer is watching.

Learning from the project: -

- I have learnt how to load the dataset of csv format.
- How & why to do data-pre-processing of data, data pre-processing leads to better data sets, that are

cleaner and are more manageable, a must for any business trying to get valuable information from the data it gathers.

- What are the python libraries available for data analysis?

  Some of them are as follows-

NumPy

One of the fundamental packages used for scientific computing with Python is Numpy. Among other things, it contains the following:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions for performing array computations
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra operations, Fourier transforms, and random number capabilities.

Besides this, it can also be used as an efficient multidimensional container of generic data. Arbitrary data types can be defined and integrated with a wide variety of databases.

Pandas

Pandas is a Python package that supports rich data structures and functions for analyzing data and is developed by the PyData Development Team. It is focused on the improvement of Python’s data libraries. Pandas consists of the following things:

- A set of labelled array data structures; the primary of which are Series, DataFrame, and Panel
- Index objects enabling both simple axis indexing and multilevel/hierarchical axis indexing
- An integrated group by engine for aggregating and transforming datasets
- Date range generation and custom date offsets
- Input/output tools that loads and saves data from flat files or PyTables/HDF5 format
- Optimal memory versions of the standard data structures
- Moving window statistics and static and moving window linear/panel regression

Because of these features, Pandas is an ideal tool for systems that need complex data structures or high-performance time series functions such as financial data analysis applications.

Matplotlib

Matplotlib is the single most used Python package for 2D-graphic. It provides both a very quick way to visualize data from Python and publication-quality figures in many formats: line plots, contour plots, scatter plots, or Basemap plot. It comes with a set of default settings, but allows customizing all kinds of properties. However, we can easily create our chart with the defaults of almost every property in Matplotlib.

- Why data analytics is so important for businesses? Using data analysis, we can determine what forms of advertising reach your customers effectively and make an impact that will make customers buy products. Data enables you to understand what methods of advertising your product have the biggest impact on the target audience and at what scale you can adopt such advertising.

Bibliography: -

1. Kaggle: Your Home for Data Science
2. Open Government Data (OGD) Platform India
3. Welcome to Python.org
4. Python Tutorial (w3schools.com)