**Exploratory Data Analysis on Netflix Dataset of Movies and Shows**

**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**

Country where the movie / show was produced

**date\_added**

Date it was added on Netflix

**release\_year**

Actual Release year of the move / show

**rating**

TV Rating of the movie / show

**duration**

Total Duration - in minutes or number of seasons

**listed\_in**

category of the content like- documentary,International TV Shows, TV Dramas, TV Mysteries etc

**description**

a short description about what the show or movie is all about

**Importing the required libraries**

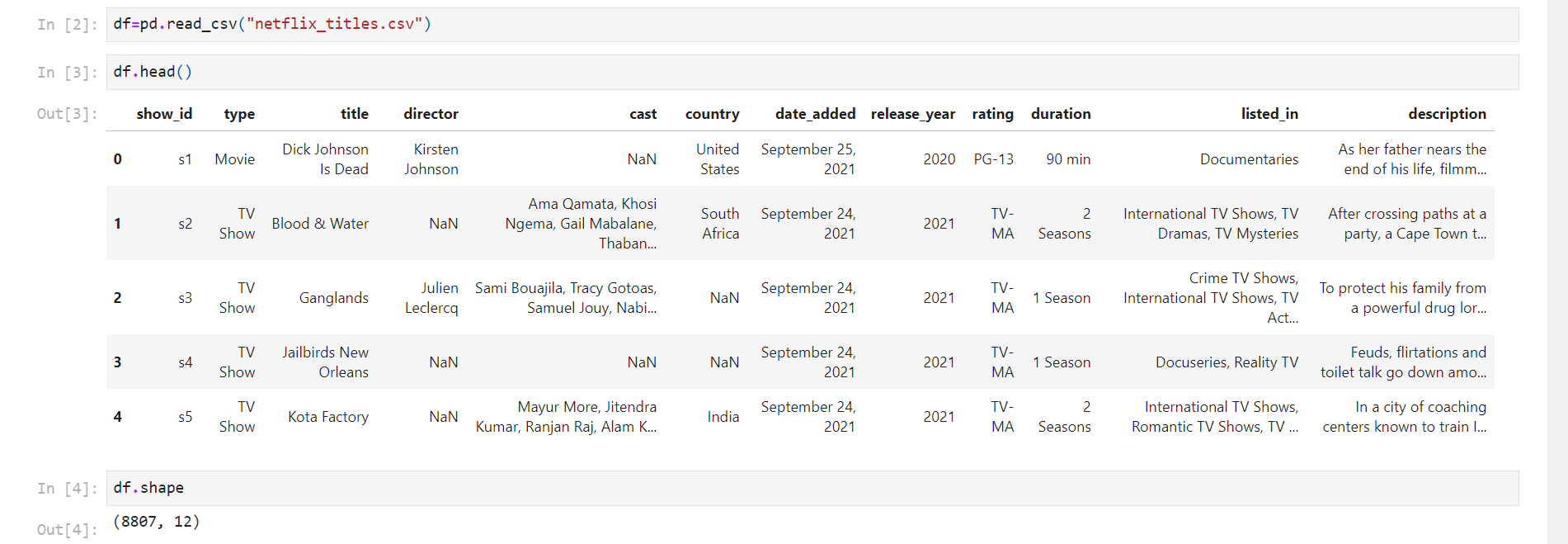
**import** pandas **as** pd

**import** numpy **as** np

**import** matplotlib.pyplot **as** plt

**import** seaborn **as** sns

**A quick look at the DataFrame**

****

**Using describe method on numeric column(s) to get statistical insights about the data**

*#describe method gives the various statistical values of all the numeric column in the dataframe*

newdf2**.**describe()

|  | **release\_year** |
| --- | --- |
| **count** | 5332.000000 |
| **mean** | 2012.742123 |
| **std** | 9.625831 |
| **min** | 1942.000000 |
| **25%** | 2011.000000 |
| **50%** | 2016.000000 |
| **75%** | 2018.000000 |
| **max** | 2021.000000 |

newdf2**.**dtypes

show\_id object

type object

title object

director object

cast object

country object

date\_added object

release\_year int64

rating object

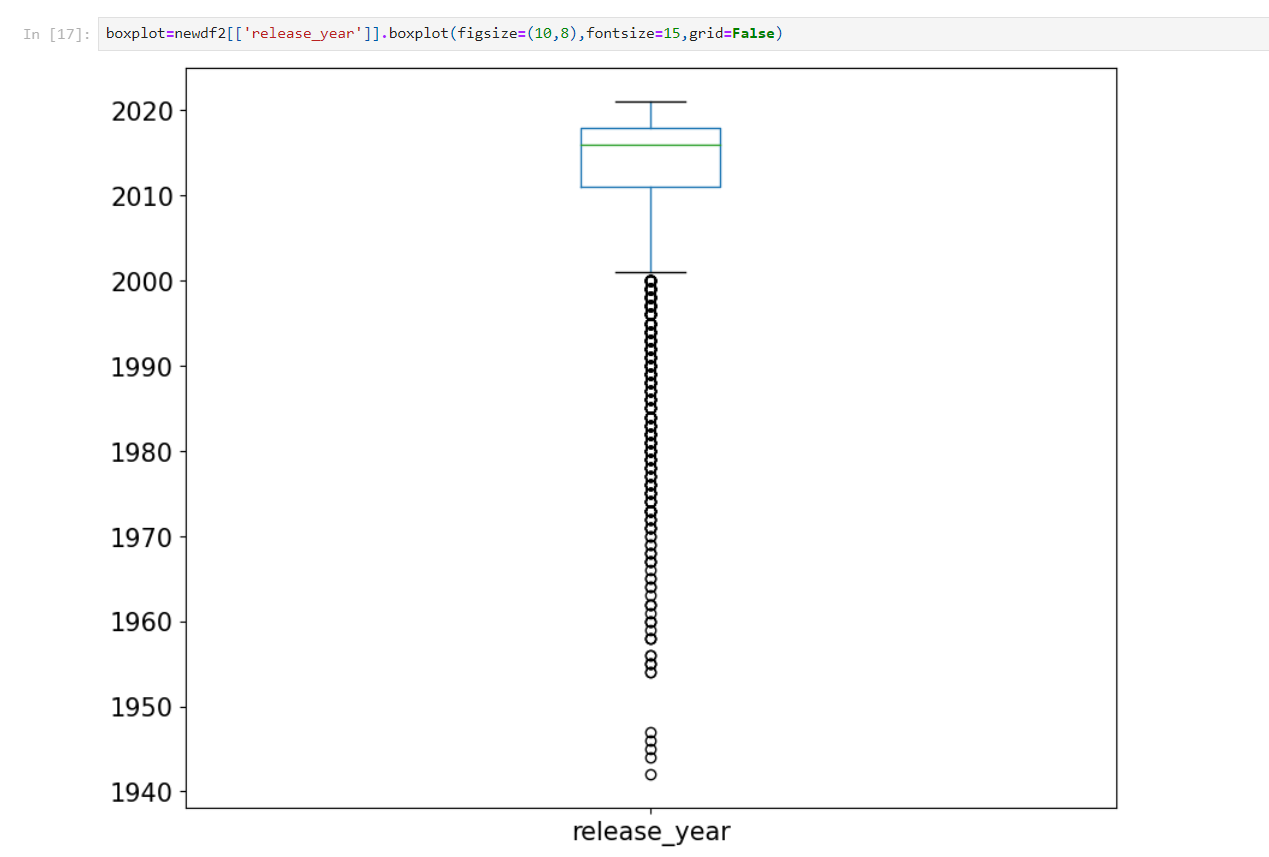
duration object

listed\_in object

description object

dtype: object

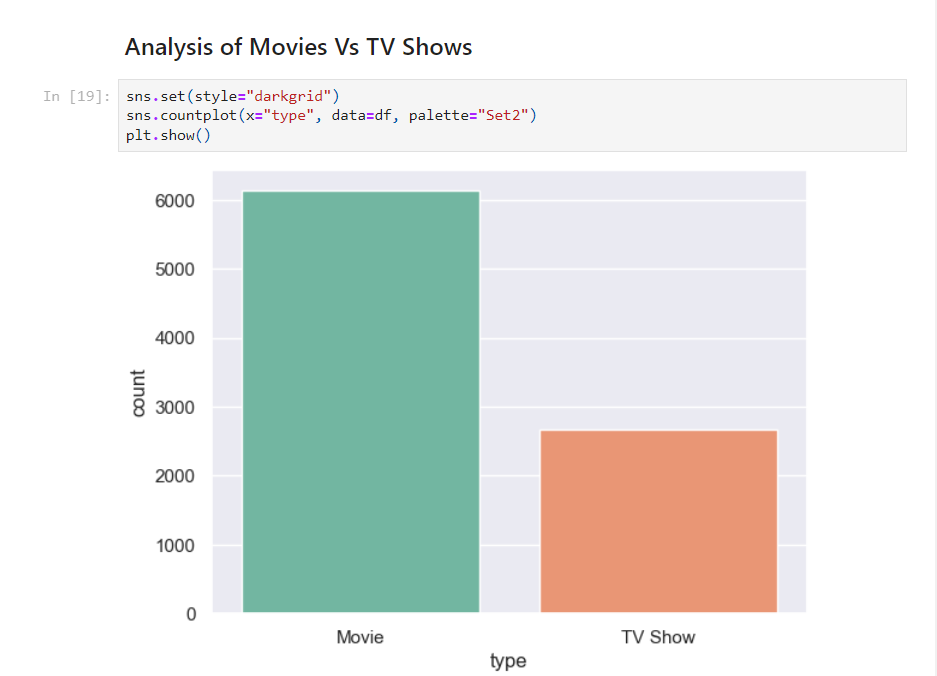
**Data Visualizations and drawn conclusions**

****

the above boxplot is an unconventional one because we usually dont have boxplot for a column which tells the release year of the content, the mean is 2012,which means that

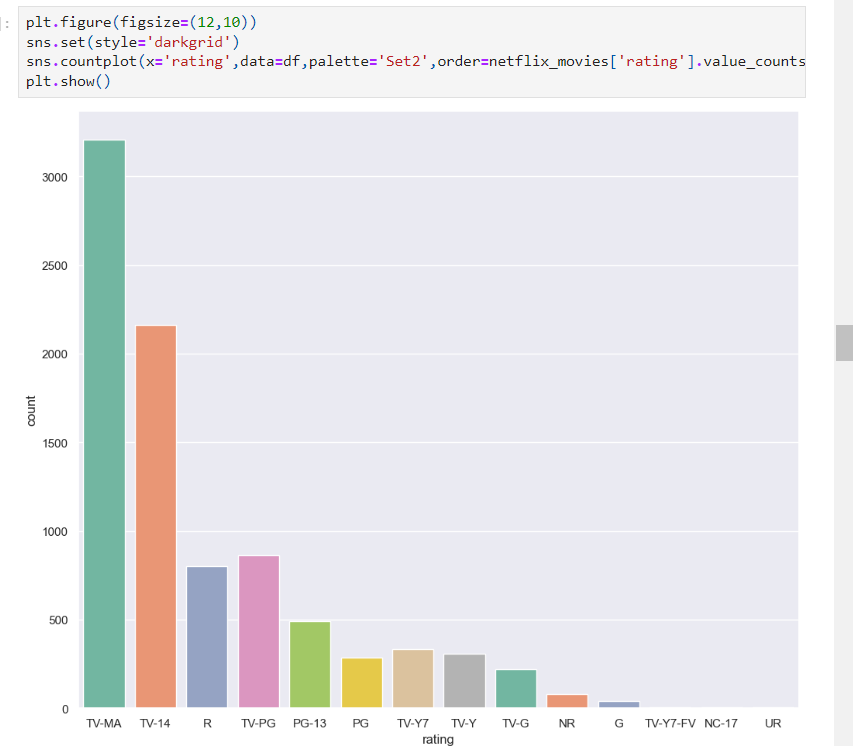
***Insight1:***

### MOST OF THE CONTENT ON NETFLIX ARE RELATIVELY MODERN CONTENT WHICH WERE MADE/CREATED IN 2010s.



***Insight2:***

### Clearly there are more movies on netflix than TV Shows.



***Insight3:***

###### The rating certificate for the movies shows that to which audience it can be shown appropriately

###### Clearly netflix contains more than 3000 movies of rating type TV-MA which is a rating for matured audience

###### The next most frequent set of movies on netflix are of TV-14 type which may not be shown to children below 14 years of age

###### 3rd most frequent category is R which states that the movie contains some scenes which is not suitable for children below 17 years.

###### 

***Insight4:***

***From the heatmap***

### If the latest year 2019 is considered, January and December were the months when comparatively much less content was released.Therefore, these months may be a good choice for the success of a new release!

**Analyzing IMDB ratings to get top rated movies on Netflix**

###### for this, first,we will download the dataset which contains imdb ratings of movies and then we will correlate both the tables to find ratings of movies/shows on Netflix.

imdb\_ratings**=**pd**.**read\_csv('IMDb ratings.csv',usecols**=**['weighted\_average\_vote'])

imdb\_titles**=**pd**.**read\_csv('IMDb movies.csv', usecols**=**['title','year','genre'])

ratings **=** pd**.**DataFrame({'Title':imdb\_titles**.**title,

'Release Year':imdb\_titles**.**year,

'Rating': imdb\_ratings**.**weighted\_average\_vote,

'Genre':imdb\_titles**.**genre})

ratings**.**drop\_duplicates(subset**=**['Title','Release Year','Rating'], inplace**=True**)

ratings**.**shape

###### (85852,4)

###### Performing inner join on the ratings dataset and netflix dataset to get the content that has both ratings on IMDB and are available on Netflix.

ratings**.**dropna()

joint\_data**=**ratings**.**merge(newdf2,left\_on**=**'Title',right\_on**=**'title',how**=**'inner')

joint\_data**=**joint\_data**.**sort\_values(by**=**'Rating', ascending**=False**)

###### *Insight5:*

#### The top 10 rated movies on Netflix are:

###### 

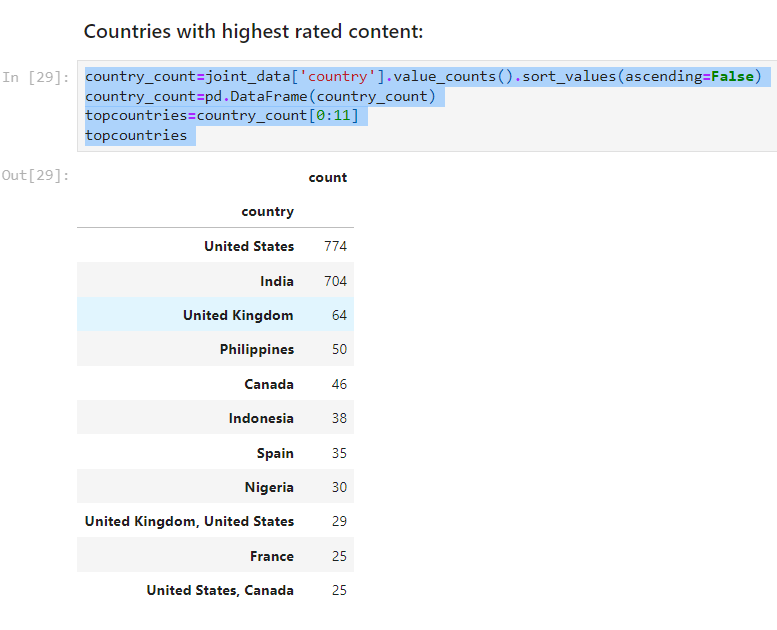
#### Countries with highest rated content:

country\_count**=**joint\_data['country']**.**value\_counts()**.**sort\_values(ascending**=False**)

country\_count**=**pd**.**DataFrame(country\_count)

topcountries**=**country\_count[0:11]

topcountries

****

**import** plotly.express **as** px

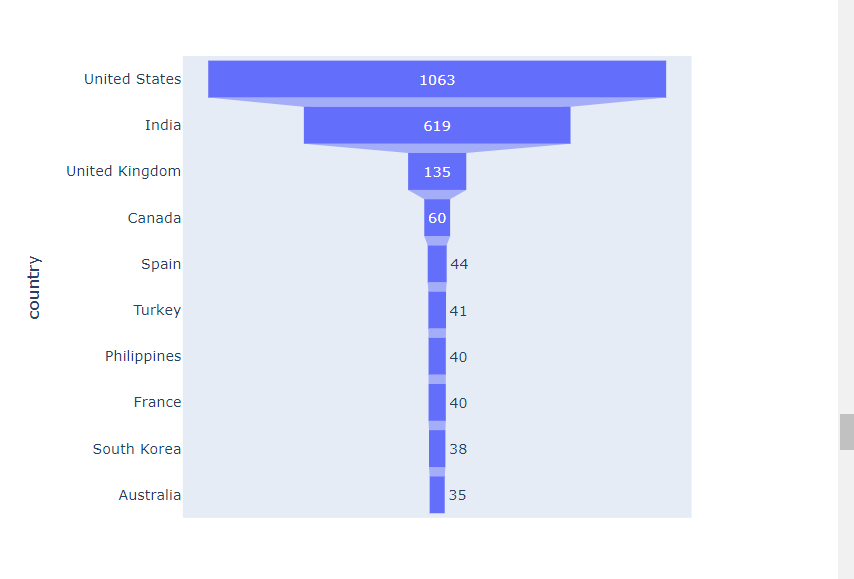
data **=** dict(

number**=**[1063,619,135,60,44,41,40,40,38,35],

country**=**["United States", "India", "United Kingdom", "Canada", "Spain",'Turkey','Philippines','France','South Korea','Australia'])

fig **=** px**.**funnel(data, x**=**'number', y**=**'country')

fig**.**show()



### Analysis of duration of movies:

netflix\_movies2**=**newdf2[newdf2['type']**==**'Movie']

netflix\_movies2['duration']**=**netflix\_movies2['duration']**.**str**.**replace(' min','')

netflix\_movies2['duration']**=**netflix\_movies2['duration']**.**astype(str)**.**astype(int)

netflix\_movies2['duration']

7 125

9 104

12 127

24 166

27 103

...

8801 96

8802 158

8804 88

8805 88

8806 111

Name: duration, Length: 5185, dtype: int32

### Analysis on Genres:

g**=**{k: v **for** k, v **in** sorted(g**.**items(), key**=lambda** item: item[1], reverse**=** **True**)}

fig, ax **=** plt**.**subplots()

fig **=** plt**.**figure(figsize **=** (10, 10))

x**=**list(g**.**keys())

y**=**list(g**.**values())

ax**.**vlines(x, ymin**=**0, ymax**=**y, color**=**'green')

ax**.**plot(x,y, "o", color**=**'maroon')

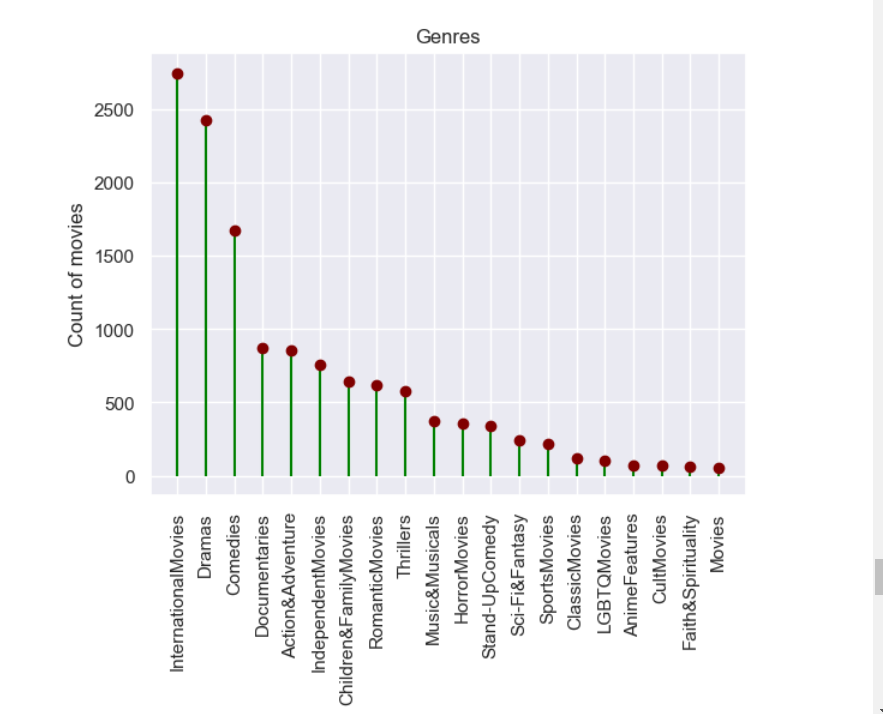
ax**.**set\_xticklabels(x, rotation **=** 90)

ax**.**set\_ylabel("Count of movies")

*# set a title*

ax**.**set\_title("Genres");

plt**.**show()



***Insight6:***

Therefore, it is clear that international movies, dramas and comedies are the top three genres that have the highest amount of content on Netflix.

**Analysis of directors and number of movies they get to make in a country**

netflix\_fr**=**df[df['country']**==**'France']

nannef**=**netflix\_fr**.**dropna()

**import** plotly.express **as** px

fig **=** px**.**treemap(nannef, path**=**['country','director'],

color**=**'director', hover\_data**=**['director','title'],color\_continuous\_scale**=**'Purples')

fig**.**show()

****

***Insight7:***

##### *Above analysis show that France is a good example where every director is given chance to make good number of films and not a single director has too many movies also, not a single director has too few movies.*