KALINGA UNIVERSITY

RAIPUR (C.G.)

A RESEARCH PAPER

ON

"Netflix Movies and TV Shows — Exploratory Data Analysis (EDA) and Visualization Using Python"

Submitted in Partial fulfillment of the Requirements for the Degree of
Bachelor of Engineering in Computer Science & Engineering

By

VIJAY KUMAR SAHU
(Enrolment no:- 202BTCS291914)

YOGESH KUMAR SAHU
(Enrolment no:- 202BTCS940227)

Under the Guidance of,

Mrs. SWATI TIWARI

Assistant Professor
Dept. of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
KALINGA UNIVERSITY
RAIPUR (C.G.)

LIST OF CONTENTS

1. ABSTRACTION:-

The Netflix dataset analysis focuses on identifying trends in consumer habits and preferences regarding the streaming service. It examines the viewing history of subscribers, the types of content they watch, the time spent watching, and the geographic location of the viewers. The data is collected from a variety of sources, including surveys, reviews, and marketing campaigns. By analyzing the data, researchers can gain insight into the behavior of Netflix subscribers and the effectiveness of the service. The analysis can also help inform decisions on content creation, marketing strategies, and pricing models. The results of the analysis can be used to improve the customer experience and increase customer loyalty.

2. MOTIVATION:-

Netflix is the largest online movie and TV show streaming service on the planet. Its service is widely available in many countries including but not limited to the United States, India, South Korea, Japan, and many more. The service was first introduced as a DVD rental service on the Internet and later, the founder and CEO of the company Reed Hastings transitioned to a

revolutionary way of delivering movies and TV shows through its website allowing many users to directly stream their favorite contents on their Internet enabled devices including desktop computers, laptops, tablet PCs, mobile phones, and many more. With its a whole new approach of delivering shows and movies, the sales of Netflix went up exponentially. Since then, the platform created its own recommender system to understand what types of movies and TV shows the users would like to watch, what kind of style of cinematography they liked the most, and how they consume their favorite TV shows. With such analysis, the company released 'The House of Cards', which became a huge success in the history of streaming service providers. With the power of data analysis, more users were attracted to the platform, and many users tend to spend most of their time watching shows and movies on Netflix. With this approach, we would like explore the dataset to understand the trend of movies and TV shows on Netflix

3. INTRODUCTION :-

N**etflix, Inc.** is an American technology and media services provider and production company headquartered in Los Gatos, California. Netflix was founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California. The company's primary business is its subscription-based streaming service, which offers online streaming of a library of films and television series, including those produced in-house.

Netflix is a popular entertainment service used by people around the world. This EDA will explore the Netflix dataset through visualizations and graphs using python libraries, matplotlib, and seaborn.

We used TV Shows and Movies listed on the Netflix dataset from Kaggle. The dataset consists of TV Shows and Movies available on Netflix as of 2019. The dataset is collected from Flixable, which third-party Netflix search engine.

4. LITERATURE REVIEW:-

We first wanted to get an overview of the dataset that we were dealing with. First we loaded up tidyverse for a simple data anlaysis purpose. We got the dataset from Kaggle and we are going to utilize data that the Kaggle website provides to understand the trend of movies and TV shows released on the platform. This dataset consists .From the code, we could see the column names that the CSV file contains. We will utilize the following columns to understand what movies and TV shows were released in specific year, what genres they were, date when they were released and the rating the audience gave and so on.

Import Libraries

Importing the libraries we need.

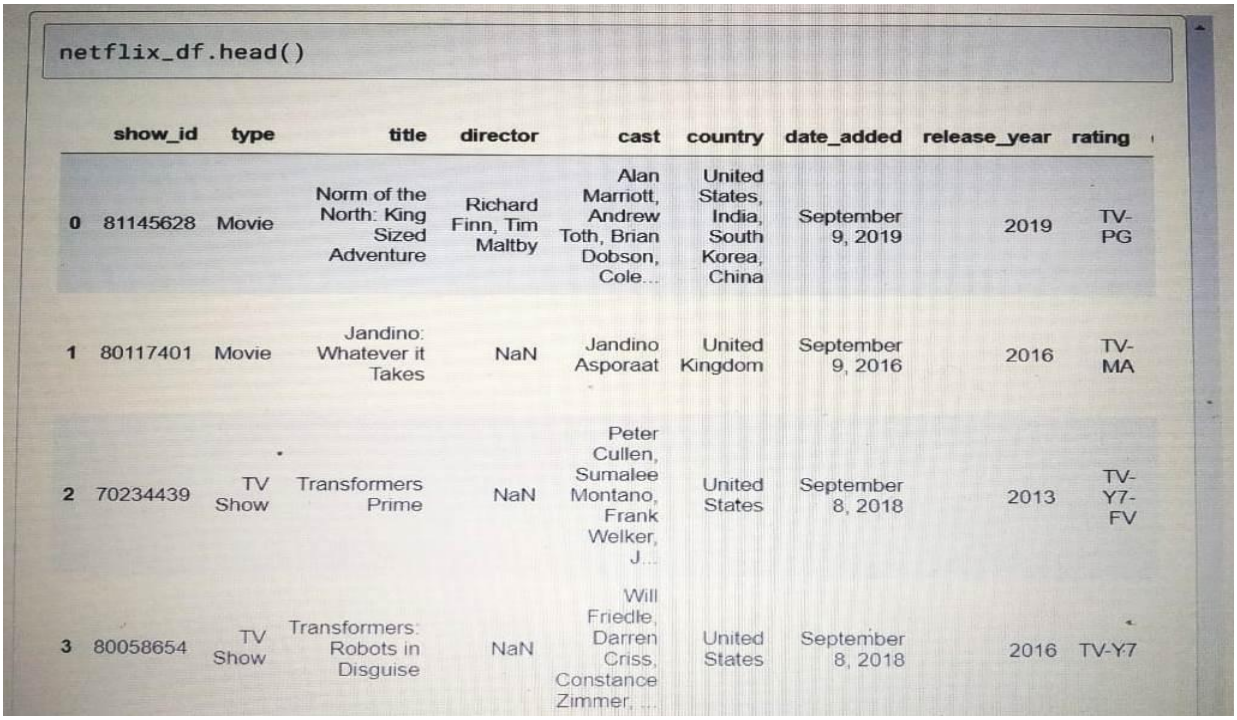
```
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
```

Loading The Dataset

Using Pandas Library, we'll load the CSV file. Named it with netflix_df for the dataset.

```
netflix_df = pd.read_csv("netflix_titles.csv")
```

Let's check the first 5 data.



```
netflix_df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating
0	81145628	Movie	Norm of the North: King Sized Adventure	Richard Finn, Tim Maltby	Alan Marriott, Andrew Toth, Brian Dobson, Cole...	United States, India, South Korea, China	September 9, 2019	2019	TV-PG
1	80117401	Movie	Jandino: Whatever it Takes	NaN	Jandino Asporaat	United Kingdom	September 9, 2016	2016	TV-MA
2	70234439	TV Show	Transformers Prime	NaN	Peter Cullen, Sumalee Montano, Frank Welker, J...	United States	September 8, 2018	2013	TV-Y7-FV
3	80058654	TV Show	Transformers: Robots in Disguise	NaN	Will Friedle, Darren Criss, Constance Zimmer, ...	United States	September 8, 2018	2016	TV-Y7

The dataset contains over 6234 titles, 12 descriptions. After a quick view of the data frames, it looks like a typical movie/TVshows data frame without ratings. We can also see that there are NaN values in some columns.

5. PROBLEM IDENTIFICATION:–

Data Profiling & Cleaning

Data Cleaning means the process of identifying incorrect, incomplete, inaccurate, irrelevant, or missing pieces of data and then modifying, replacing, or deleting them as needed. Data Cleansing is considered as the basic element of Data Science.

```
print('\nColumns with missing value:')
print(netflix_df.isnull().any())

Columns with missing value:
show_id          False
type             False
title            False
director          True
cast              True
country           True
date_added        True
release_year     False
rating            True
duration         False
listed_in        False
description      False
dtype: bool
```

From the info, we know that there are 6,234 entries and 12 columns to work with for this EDA. There are a few columns that contain null values, “director,” “cast,” “country,” “date_added,” “rating.”

```
show_id          0
type             0
title            0
director      1969
cast           570
country        476
date_added      11
release_year     0
rating          10
duration         0
listed_in        0
description      0
dtype: int64
```

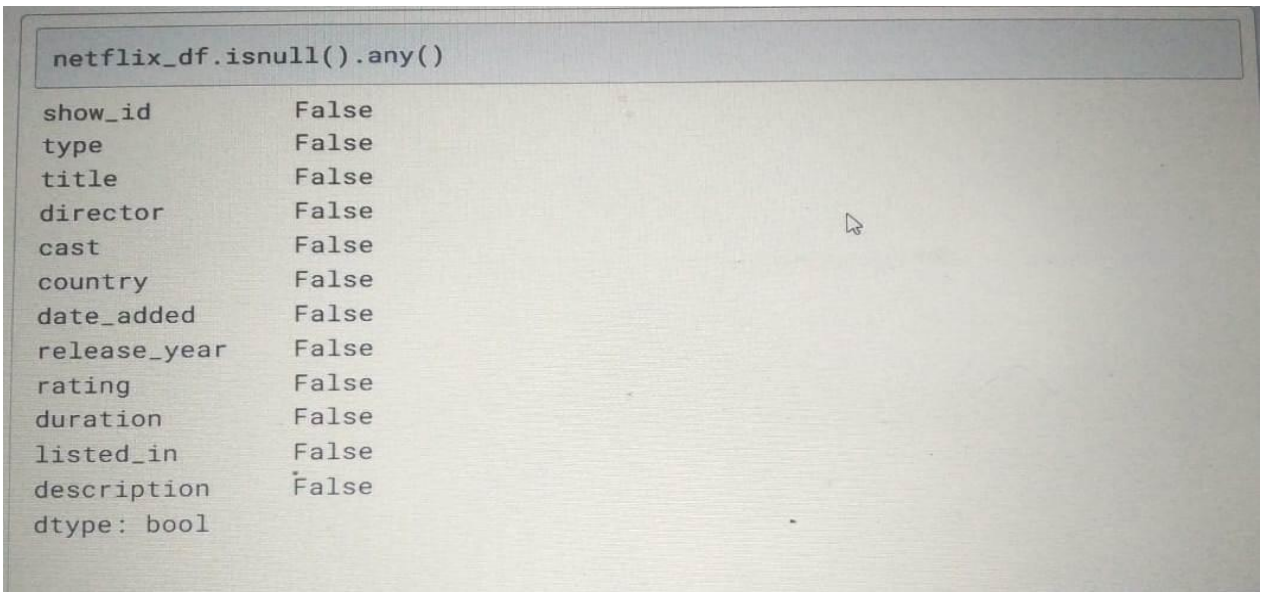
There are a total of 3,036 null values across the entire dataset with 1,969 missing points under “director” 570 under “cast,” 476 under “country,” 11 under “date_added,” and 10 under “rating.” We will have to handle all null data points before we can dive into EDA and modeling.

6. METHODOLOGY:-

Imputation is a treatment method for missing value by filling it in using certain techniques. Can use mean, mode, or use predictive modeling. In this module, we will discuss the use of the fillna function from Pandas for this imputation. Drop rows containing missing values. Can use the drone a function from Pandas.

```
netflix_df.director.fillna("No Director", inplace=True)
netflix_df.cast.fillna("No Cast", inplace=True)
netflix_df.country.fillna("Country Unavailable", inplace=True)
netflix_df.dropna(subset=["date_added", "rating"], inplace=True)
```

The easiest way to get rid of them would be to delete the rows with the missing data for missing values. However, this wouldn't be beneficial to our EDA since it is a loss of information. Since "director," "cast," and "country" contain the majority of null values, we chose to treat each missing value is unavailable. The other two label "date_added" and "rating" contain an insignificant portion of the data, so it drops from the dataset. Finally, we can see that there are no more missing values in the data frame.



7. TECHNOLOGIES USED:-

- PANDAS- Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

- NUMPY- NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these array

- MATPLOTLIB- **Matplotlib** is an amazing visualization **library** in **Python** for 2D plots of arrays. **Matplotlib** is a multi-platform data visualization **library**

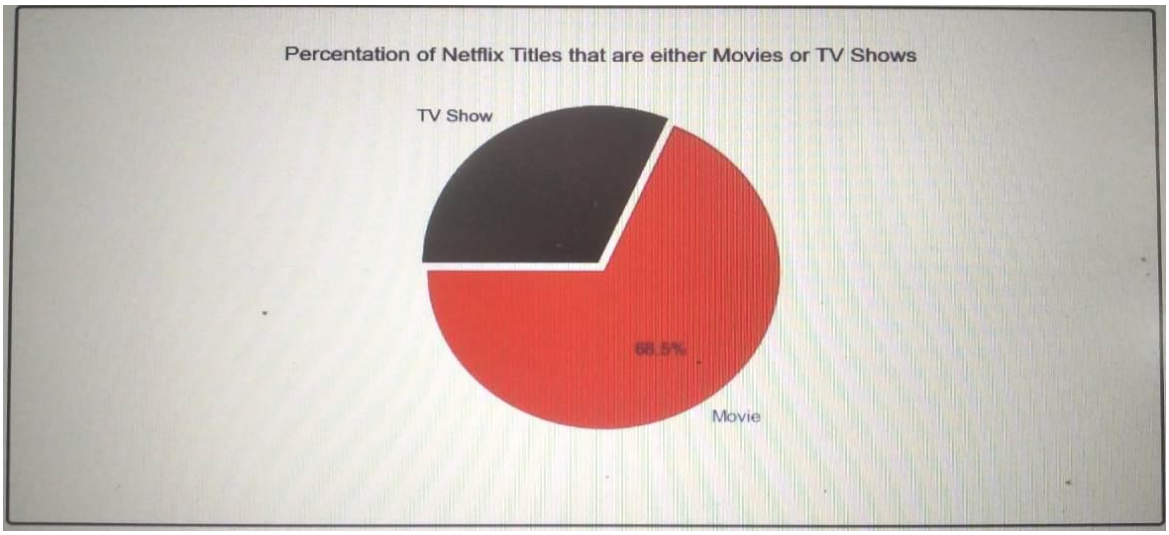
- SEABORN- **Seaborn** is a Python data visualization **library** based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical

8. EXPLORATORY ANALYSIS AND VISUALIZATION:-

1) Netflix Content By Type:-

Analysis entire Netflix dataset consisting of both movies and shows. Let's compare the total number of movies and shows in this dataset to know which one is the majority.

```
plt.figure(figsize=(12,6))
plt.title("Percentation of Netflix Titles that are either Movies or TV Shows")
g = plt.pie(netflix_df.type.value_counts(),
explode=(0.025,0.025), labels=netflix_df.type.value_counts().index,
colors=['red','black'],autopct='%1.1f%%', startangle=180)
plt.show()
```

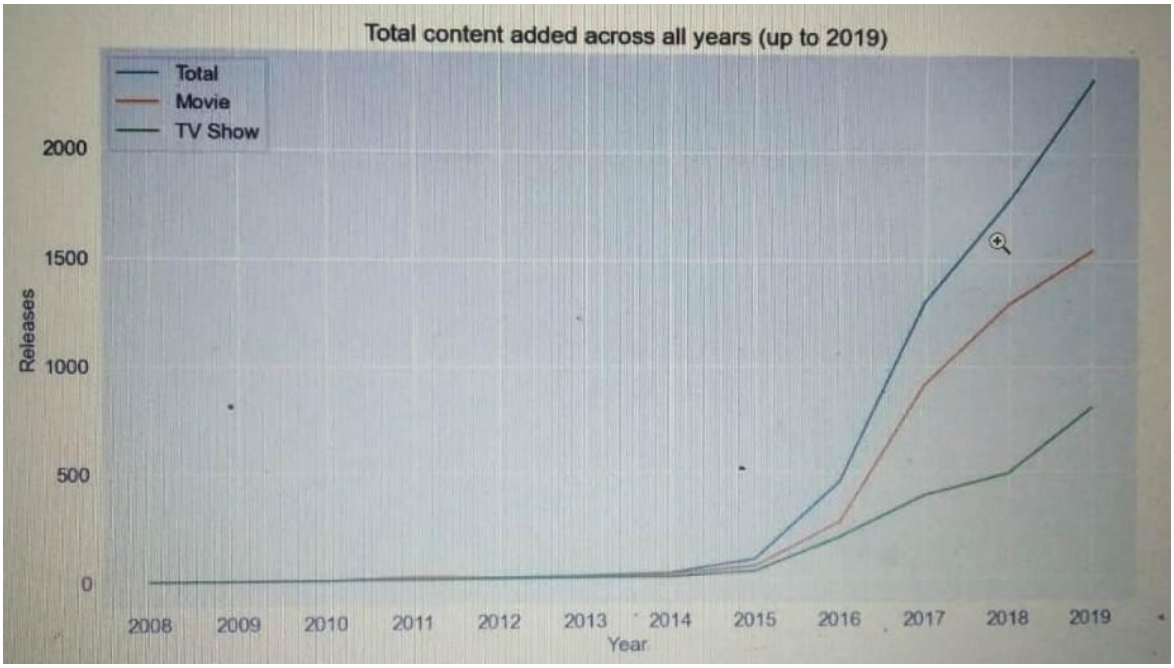


So there are about 4,000++ movies and almost 2,000 TV shows, with movies being the majority. There are far more movie titles (68,5%) that TV shows titles (31,5%) in terms of title.

2) Amount of Content as a Function of Time:-

Next, we will explore the amount of content Netflix has added throughout the previous years. Since we are interested in when Netflix added the title onto their platform, we will add a "year_added" column to show the date from the "date_added" columns.

```
fig, ax = plt.subplots(figsize=(13, 7))
sns.lineplot(data=netflix_year_df, x='year', y='date_added')
sns.lineplot(data=movies_year_df, x='year', y='date_added')
sns.lineplot(data=shows_year_df, x='year', y='date_added')
ax.set_xticks(np.arange(2008, 2020, 1))
plt.title("Total content added across all years (up to 2019)")
plt.legend(['Total','Movie','TV Show'])
plt.ylabel("Releases")plt.xlabel("Year")
plt.show()
```



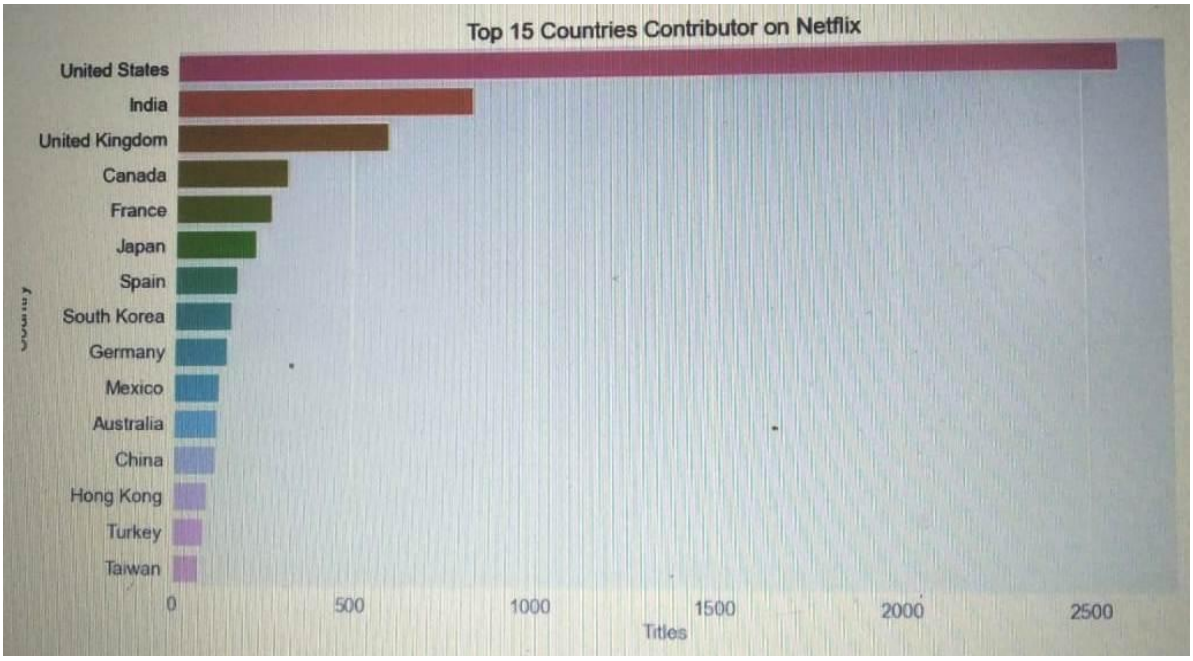
Based on the timeline above, we can conclude that the popular streaming platform started gaining traction after 2013. Since then, the amount of content added has been increasing significantly. The growth in the number of movies on Netflix is much higher than that on TV shows. About 1,300 new movies were added in both 2018 and 2019. Besides, we can know that Netflix has increasingly focused on movies rather than TV shows in recent years

3) Countries by the Amount of the Produces Content:-

Next is exploring the countries by the amount of the produces content of Netflix. We need to separate all countries within a film before analyzing it, then removing titles with no

countries available.

```
filtered_countries =
netflix_df.set_index('title').country.str.split(', ',
expand=True).stack().reset_index(level=1, drop=True);
filtered_countries = filtered_countries[filtered_countries != 'Country
Unavailable']
plt.figure(figsize=(13,7))
g = sns.countplot(y = filtered_countries,
order=filtered_countries.value_counts().index[:15])
plt.title('Top 15 Countries Contributor on Netflix')
plt.xlabel('Titles')
plt.ylabel('Country')
plt.show()
```

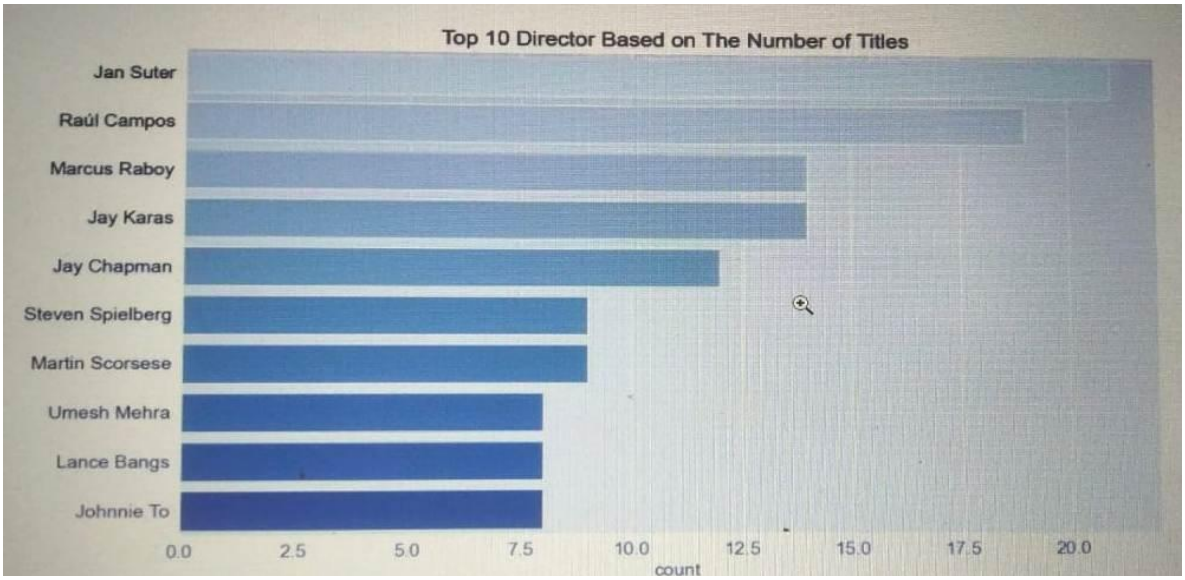


From the images above, we can see the top 15 countries contributor to Netflix. The country by the amount of the produces content is the United States.

4) Top Directors on Netflix:-

To know the most popular director, we can visualize it.

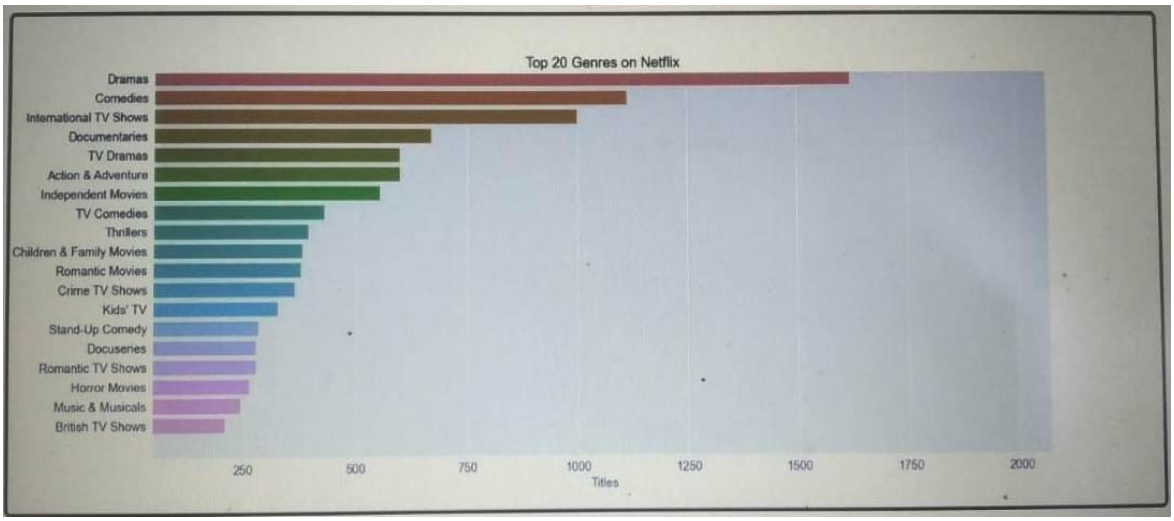
```
filtered_directors = netflix_df[netflix_df.director != 'No
Director'].set_index('title').director.str.split(', ',
expand=True).stack().reset_index(level=1, drop=True)
plt.figure(figsize=(13,7))
plt.title('Top 10 Director Based on The Number of Titles')
sns.countplot(y = filtered_directors,
order=filtered_directors.value_counts().index[:10],
palette='Blues')
plt.show()
```



The most popular director on Netflix, with the most titles, is mainly international.

5) Top Genres on Netflix:-

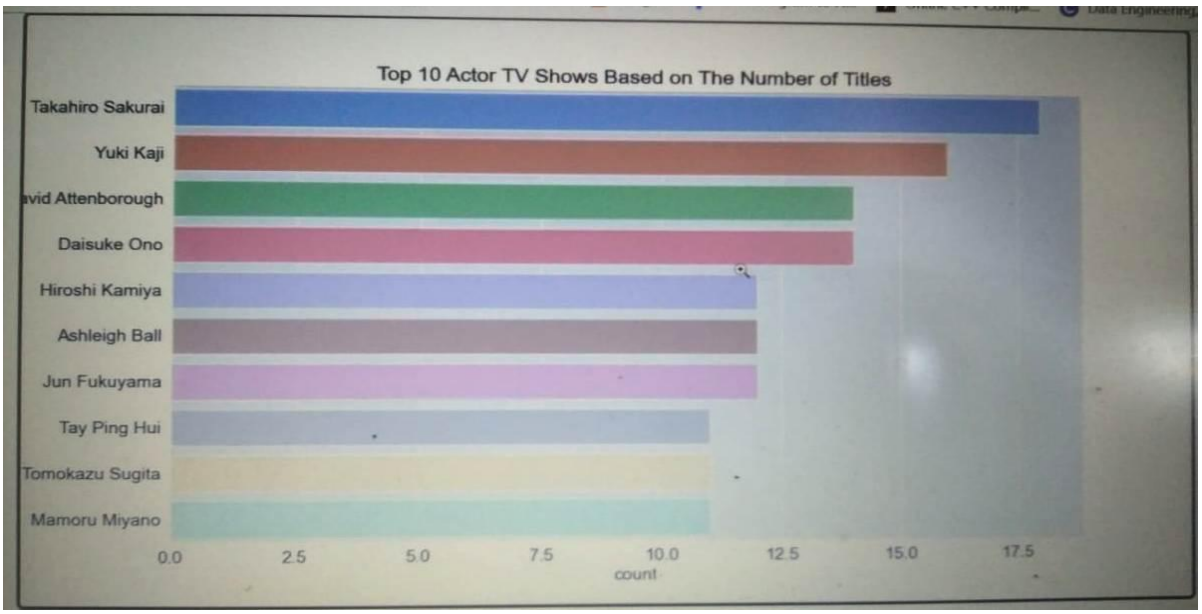
```
filtered_genres =
netflix_df.set_index('title').listed_in.str.split(', ',
expand=True).stack().reset_index(level=1, drop=True);
plt.figure(figsize=(10,10))
g = sns.countplot(y = filtered_genres,
order=filtered_genres.value_counts().index[:20])
plt.title('Top 20 Genres on Netflix')
plt.xlabel('Titles')
plt.ylabel('Genres')
plt.show()
```



From the graph, we know that International Movies take the first place, followed by dramas and comedies.

6) Top Actor for TV Show on Netflix based on the number of titles:-

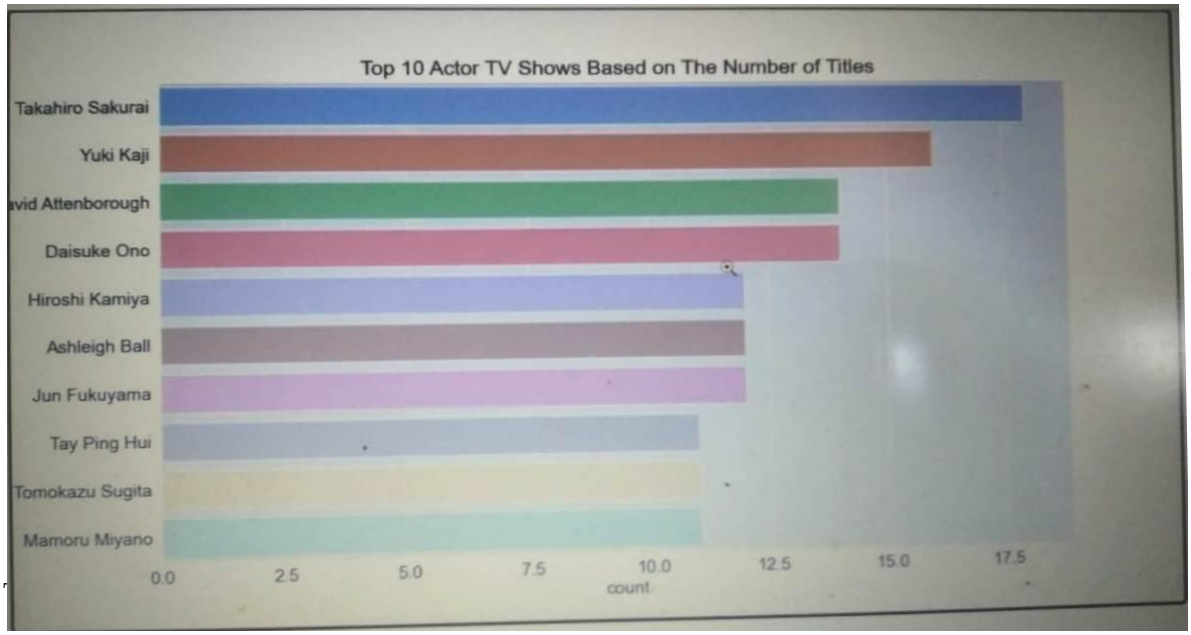
```
filtered_cast_shows =
netflix_shows_df[netflix_shows_df.cast != 'No
Cast'].set_index('title').cast.str.split(', ',
expand=True).stack().reset_index(level=1, drop=True)
plt.figure(figsize=(13,7))
plt.title('Top 10 Actor TV Shows Based on The Number of Titles')
sns.countplot(y = filtered_cast_shows,
order=filtered_cast_shows.value_counts().index[:10], palette='pastel')
plt.show()
```



The top actor on Netflix TV Show, based on the number of titles, is Takahiro Sakurai.

7) Top Actor for Movie on Netflix based on the number of titles:-

```
filtered_cast_movie =
netflix_movies_df[netflix_movies_df.cast != 'No
Cast'].set_index('title').cast.str.split(', ',
expand=True).stack().reset_index(level=1, drop=True)
plt.figure(figsize=(13,7))
plt.title('Top 10 Actor Movies Based on The Number of Titles')
sns.countplot(y = filtered_cast_movie,
order=filtered_cast_movie.value_counts().index[:10], palette='pastel')
plt.show()
```

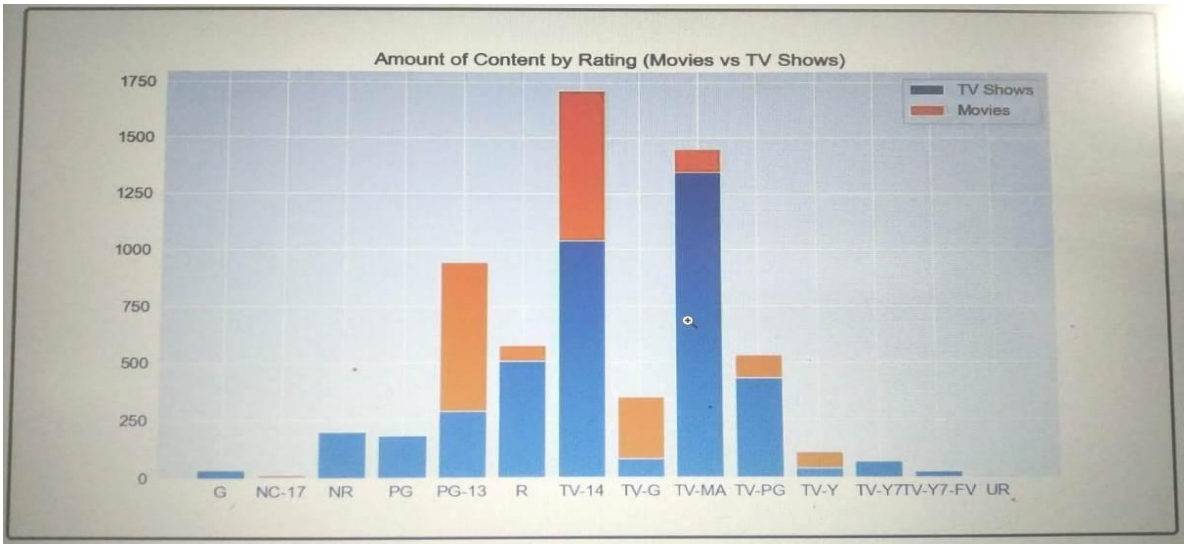


8) Amount of Content By Rating:-

```
order = netflix_df.rating.unique()
count_movies = netflix_movies_df.groupby('rating')
['title'].count().reset_index()
count_shows = netflix_shows_df.groupby('rating')
['title'].count().reset_index()

count_shows = count_shows.append
([{"rating" : "NC-17", "title" : 0},{"rating" : "PG-13", "title" : 0},{"rating" :
"UR", "title" : 0}], ignore_index=True)
count_shows.sort_values(by="rating", ascending=True)

plt.figure(figsize=(13,7))
plt.title('Amount of Content by Rating (Movies vs TV Shows)')
plt.bar(count_movies.rating, count_movies.title)
plt.bar(count_movies.rating, count_shows.title, bottom=count_movies.title)
plt.legend(['TV Shows', 'Movies'])
plt.show()
```



9. CONCLUSIONS-

We have drawn many interesting inferences from the dataset Netflix titles; here's a summary of the few of them: We have drawn many interesting inferences from the dataset Netflix titles; here's a summary of the few of them:

- The most content type on Netflix is movies,
- The popular streaming platform started gaining traction after 2014. Since then, the amount of content added has been increasing significantly.
- The country by the amount of the produces content is the United States,
- The most popular director on Netflix , with the most titles, is Jan Suter.
- International Movies is a genre that is mostly in Netflix
- The most popular actor on Netflix TV Shows based on the number of titles is Takahiro Sakurai,
- The most popular actor on Netflix movie, based on the number of titles, is Anupam Kher.

10. FUTURE WORK:-

Given more time we would work to develop or improve on several ideas. Further work could also have been done with applying deep learning to this problem. A few viable methodologies might be to use techniques for unsupervised learning from the numerical data and maybe used in other neural networks to perform classification.

Team Membership and Attestation of Work

VIJAY KUAMR SAHU and YOGESH KUMAR SAHU agree to participate and work on this project with the understanding of the project.

Video Presentation Link

11. REFERENCES:-

[1]. You can download the data and python code document via GitHub: https://github.com/dwiknrd/medium-code/tree/master/netflix-eda.

[2]. https://www.kaggle.com/datasets/shivamb/netflix-shows/versions/2?resource=download

[3]. https://medium.com/analytics-vidhya/netflix-movies-and-tvshows-exploratory-data-analysis-eda-and-visualization-using-python-80753fcfcf7