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

Vijay Kumar Sahu[1], Bigit Krishna Goswami[2], Moses Benjamin[3], Aviral Upadhyay[4],
[5]Mr. Omprakash Dewangan Assistant professor
[1,2,3,4,5]Department of Computer Science and Information Technology
[1,2,3,4,5]Kalinga University, Village – Kotni, Near Mantralaya, Naya Raipur (C.G.), India-492101

[1]Viijjaayy2000@gmail.com, [2]bigitkrishna.goshwam@kalingauniversity.ac.in,
[3]mosesjosephbenjamintutu@gmail.com, [4]aviralupadhyay433@gmail.com,
[5]Omprakash.dewangan@kalingauniversity.ac.in

DOI:10.48047/ecb/2023.12.si4.704

***Abstract*— 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.**

***Index Terms*— Python, Data Science, Data Analysis, Pandas, Data Visualization, Matplotlib, Seaborn.**

I. INTRODUCTION

Netflix, 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 is a third-party Netflix search engine.

II. LITERATURE REVIEW

1. *Review Stage*

We first wanted to get an overview of the dataset that we were dealing with. First, we loaded up tidy verse for a simple data analysis 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 of. From the code, we could see the column names that the CSV file [1] contains. We will utilize the following columns to understand what movies and TV shows were released in a specific year, what genres they were, date when they were released and the rating the audience gave and so on.

2. *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
```

3. *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.

Fig 1. Glimpse of first 5 entries of dataset

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

III. PROBLEM IDENTIFICATION

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.

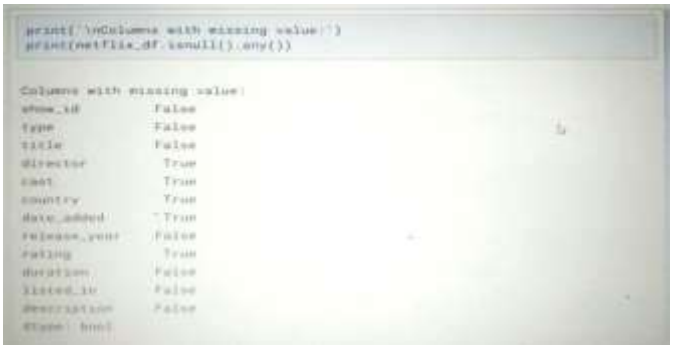

Fig 2: Columns with missing value

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".

Fig 3: Count of 'null' values observed in dataset

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 [2]. Italicize symbols (T might refer to temperature, but T is the unit tesla). Refer to "(1)," not "Eq. (1)" or "equation (1)," except at the beginning of a sentence: "Equation (1) is ... ."

IV. 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("CountryUnavailable", 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 [3]. 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.

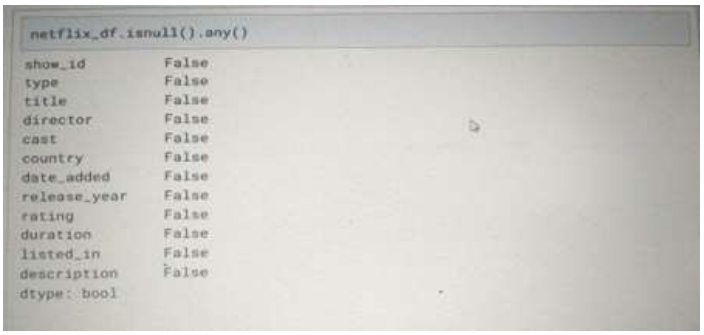

Fig 4: Segregated data verification

V. 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 arrays.

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 representation of data.

VI. 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 [4].

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

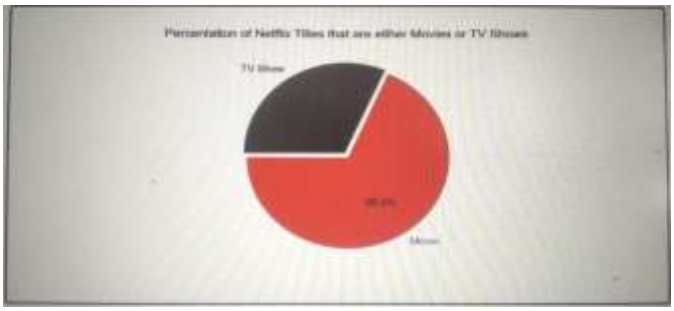

Fig 5: Percentile of Netflix Titles that are either Movies or TV Shows

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 [5], 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()
```

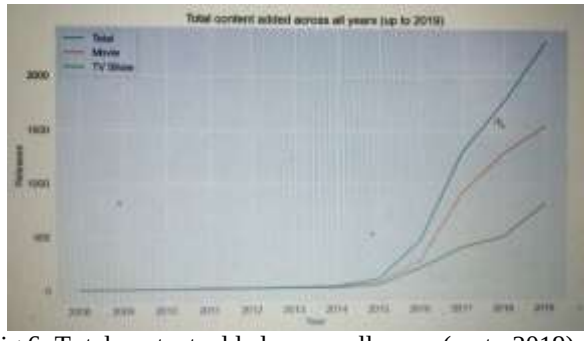

Fig 6: Total content added across all years (up to 2019)

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 [6]. 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 [7][8].

```
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 below, we can see the top 15 countries contributor to Netflix. The country with the most amount of the produces content is the United States [9].

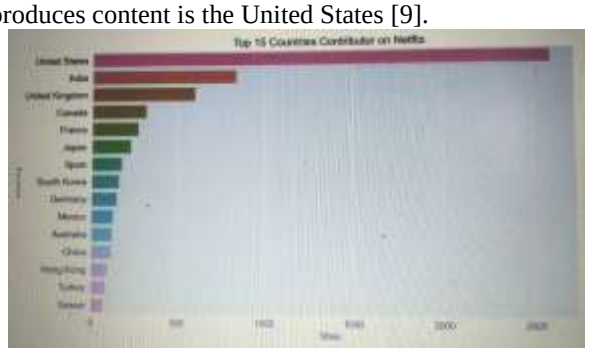

Fig 7: Countries Contributor on Netflix

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.

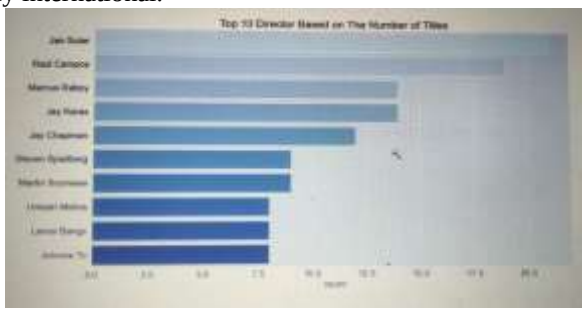

Fig 8: Top 10 Director Based on The Number of Titles

5. *Top Genres on Netflix*

```
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()
```

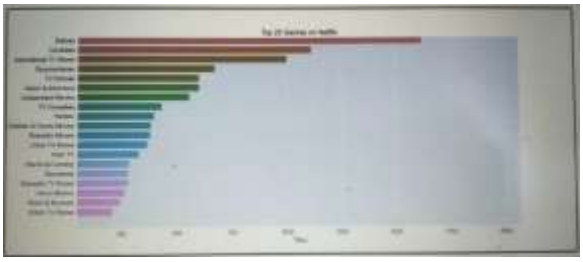

Fig 9: Total content added across all years (up to 2019)

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 Takshiro Sakurai [10].

7. *Top Actor for Movie on Netflix based on the number of titles*

```
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()
```

The top actor on Netflix Movies, based on the number of titles, is Anupam Kher.

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()
```

Fig 11: Top 10 Actor in TV Shows based on the Number of Titles (Left) & Amount of Content by Rating (Movies vs TV Shows) (Right)

VII. CONCLUSION

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:

1. The most content type on Netflix is movies.

2. The popular streaming platform started gaining fraction after 2014. Since then, the amount of content added has been increasing significantly.

3. The country by the amount of the produces content is the United States.

4. The most popular director on Netflix, with the most titles observed for Jan Suter.

5. International Movies is a genre that is mostly in Netflix.

6. The most popular actor on Netflix TV Shows based on the number of titles is Takahiro Sakurai.

7. The most popular actor on Netflix movie, based on the number of titles, is Anupam Kher.

REFERENCES

[1] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 1, no. 1, pp. 24-45, Jan.-March 2004, doi: 10.1109/TCBB.2004.2.

[2] E. Handschin, F. C. Schweppe, J. Kohlas and A. Fiechter, "Bad data analysis for power system state estimation," in IEEE Transactions on Power Apparatus and Systems, vol. 94, no. 2, pp. 329-337, March 1975, doi: 10.1109/T-PAS.1975.31858.

[3] V. Gowri, B. Harish, F. Ahmed and M. Srinath, "Netflix Stock Price Movements Insights from Data Mining," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-4, doi: 10.1109/MysuruCon55714.2022.9972547.

[4] A. Batch and N. Elmqvist, "The Interactive Visualization Gap in Initial Exploratory Data Analysis," in IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 278-287, Jan. 2018, doi: 10.1109/TVCG.2017.2743990.

[5] J. C. Roberts, "State of the Art: Coordinated & Multiple Views in Exploratory Visualization," Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007), Zurich, Switzerland, 2007, pp. 61-71, doi: 10.1109/CMV.2007.20.

[6] A. Meyer-Baese, A. Wismueller and O. Lange, "Comparison of two exploratory data analysis methods for fMRI: unsupervised clustering versus independent component analysis," in IEEE Transactions on Information Technology in Biomedicine, vol. 8, no. 3, pp. 387-398, Sept. 2004, doi: 10.1109/TITB.2004.834406.

[7] C. M. Choy, M. K. Co, M. J. Fogel, C. D. Garrioch, C. K. Leung and E. Martchenko, "Natural Sciences Meet Social Sciences: Census Data Analytics for Detecting Home Language Shifts," 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM), Seoul, Korea (South), 2021, pp. 1-8, doi: 10.1109/IMCOM51814.2021.9377412.

[8] T. Zhang and C. . -C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," in IEEE Transactions on Speech and Audio Processing, vol. 9, no. 4, pp. 441-457, May 2001, doi: 10.1109/89.917689.

[9] Yao Wang, Zhu Liu and Jin-Cheng Huang, "Multimedia content analysis-using both audio and visual clues," in IEEE Signal Processing Magazine, vol. 17, no. 6, pp. 12-36, Nov. 2000, doi: 10.1109/79.888862.

[10] Z. Lv, H. Song, P. Basanta-Val, A. Steed and M. Jo, "Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics," in IEEE Transactions on Industrial Informatics, vol. 13, no. 4, pp. 1891-1899, Aug. 2017, doi: 10.1109/TII.2017.2650204.