IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

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

Salim Umar Ladan,

Stephen Ladu Sebit Amosa

Eseto Lomumba Patric

Salim Isah

Maryam Abubakar Yazid

Mr. Pawan Kumar, Assistant professor
Department of Computer Science and Information Technology
Kalinga University, Village – Kotni, Near Mantralaya, Naya Raipur (C.G.), India-492101

Abstract

The Movies 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 area of the audience. The data is gathered from various sources, including surveys, reviews, and marketing campaigns. Through careful data analysis, researchers can uncover valuable insights on the behavior of Movies subscribers and the effectiveness of the service The analysis can provide valuable guidance for making decisions about content creation. marketing strategies, and pricing models.The results of the analysis can be used to improve thecustomer experience and increase customer loyalty.

Keywords— Python, Data Science, Data Analysis, Pandas, Data Visualization, Matplotlib, Seaborn.

I. INTRODUCTION

Exploratory data analysis (EDA) [17] is a valuable approach for analyzing data sets to summarize their main characteristics, often using visual methods. While a statistical model can be utilized, the primary focus of EDA is to uncover insights beyond formal modeling or hypothesis testing. This approach, promoted by John Tukey, encourages statisticians to explore data and potentially form hypotheses that could lead to new data collection and experiments. It's important to note that EDA is different from initial data analysis (IDA), which is more focused on checking assumptions for model fitting, hypothesis testing, handling missing values, and making necessary variable transformations. EDA encompasses IDA and aims to maximize the analyst's insight into the underlying structure of a data set while providing all the specific items that an analyst would want to extract, such as a well-fitting, parsimonious model, and a list of outliers. [17]

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 weare 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 seethe 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*

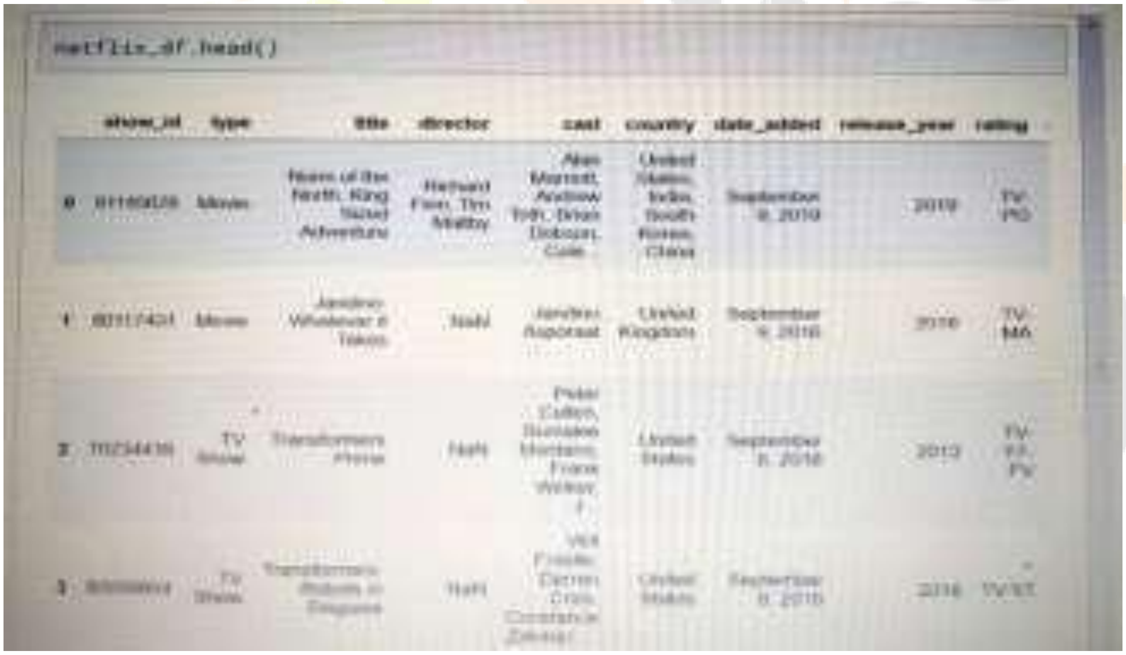
To import library functionality into your code, use the keyword import. Using the import keyword at the top of your code file, you can import certain library functions or the entire library at once.

3. *Loading the Dataset*

Sure, you can use the following code to load a CSV file using the Pandas library:

python import pandas as pd # Load the CSV file data = pd.read_csv. Got it, the dataset is named movies_df.. *Movies_df = pd.read_csv("Movies_titles.csv")*

To check the first 5 data, you can use the head() function in Python pandas like this



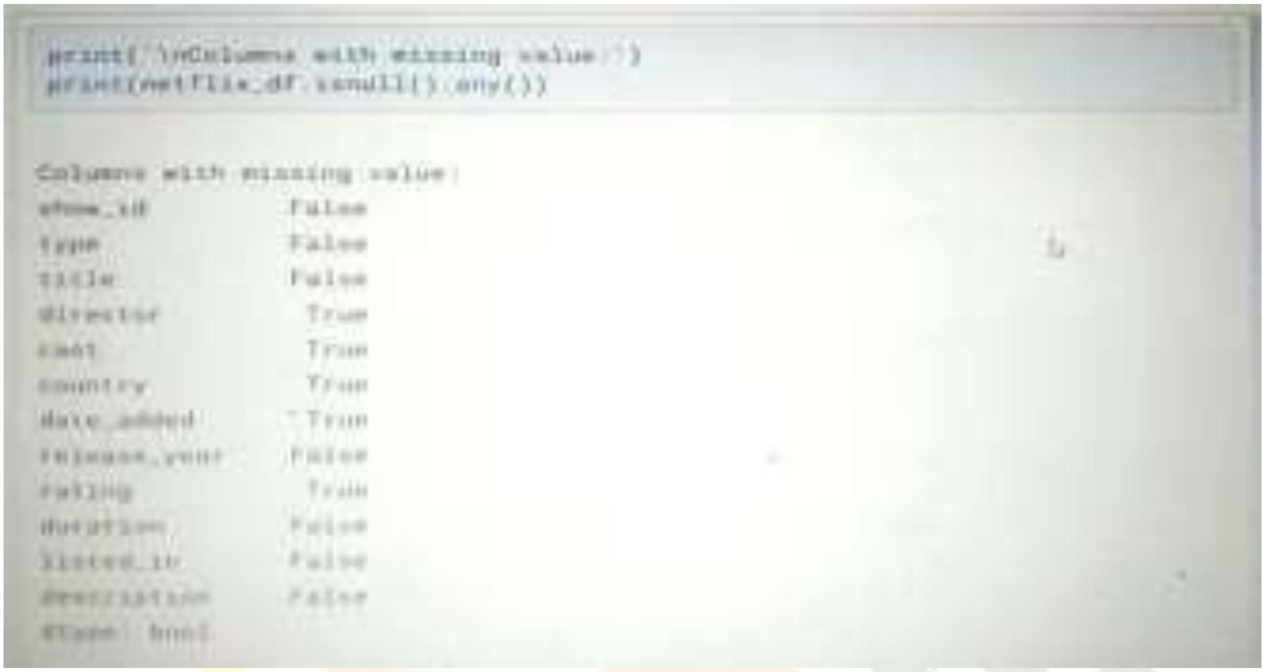
```
netflix_df.head()
```

	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 analysis above shows that, the data set contains more than 6234 titles with total of 12 descriptions. Furthermore, it appears that a particular movie/TV show have data frame without proper rating.

II. PROBLEM IDENTIFICATION

Absolutely, data cleaning is indeed a crucial process that involves identifying and addressing data which is not accurately, completely, relevant, missing through modification or erasing the data completely. (medium.com)

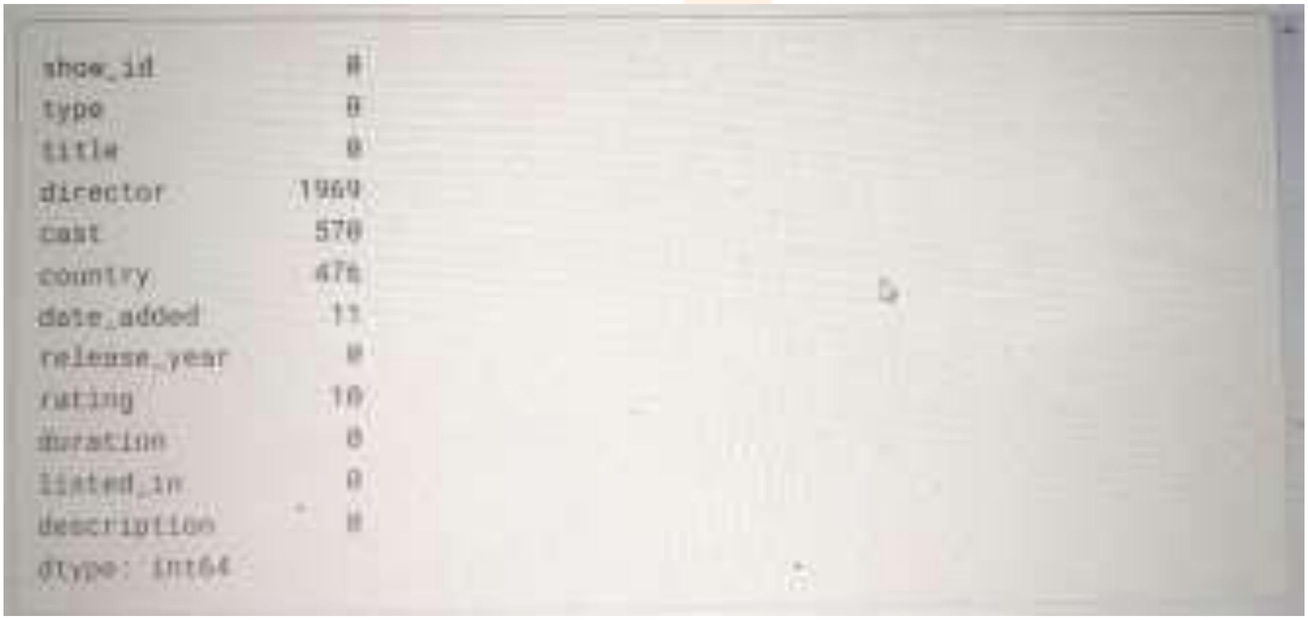


```
print('\nColumns with missing value:')
print(netflix_df.isnull().any())
```

```
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
```

Fig 2: Columns with missing value

The above picture shows that, there are 6234 input and 12 columns to analyse. It's worth noting that there are severanull values, including "director," "cast," "country," "date_added," and "rating." [16]



```
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
```

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

The analysis indicate that a total 3,036 null values and 1,969 points missing within "director" 570 points missing within "cast" 476 points missing within "country" 11 under "date added" lastly,10 points missing within "rating".

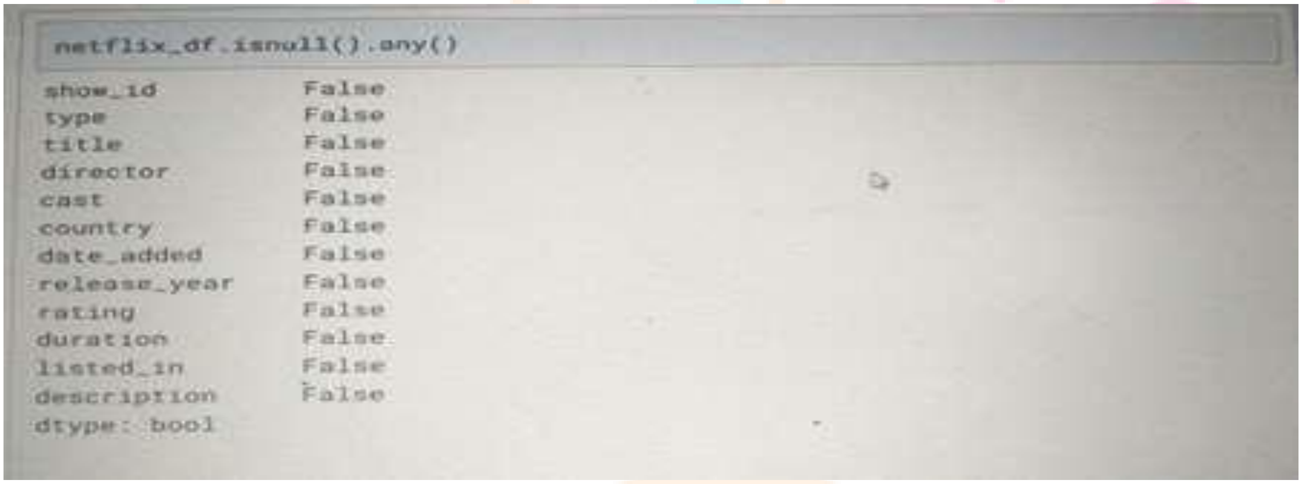
III. METHODOLOGY

Imputation is indeed a treatment method for handling missing values by filling them in using various techniques. This module will focus on the use of pandas fillna function for imputation and dropping cells which carries the missing values. Lastly we will use the pandas dropna function for this purpose. (www.midium.com)

Movies_df.director.fillna("No Director", inplace=True) Movies_df.cast.fillna("No Cast", inplace=True)
Movies_df.country.fillna("CountryUnavailable", inplace=True)
Movies_df.dropna(subset=["date_added","rating"],inplace= True)

It's true that the simplex method to handle missing values is by erasing the entire rows with the missing data, but this approach can lead to a loss of valuable information for our EDA. Since "director," "cast," and "country" carry the highest null value, one can choose to consider each missing value separately as “unavailable” [16]

Fig 4: Segregated data verification



```
netflix_df.isnull().any()

show_id           False
type              False
title             False
director          False
cast              False
country           False
date_added        False
release_year      False
rating            False
duration          False
listed_in         False
description       False
dtype: bool
```

IV. TECHNOLOGIES USED

PANDAS- Pandas is an essential Python library designed for efficient data manipulation and analysis. It provides essential data structures and operations for manipulating numerical tables and time series. [17]

NUMPY- NumPy is a Python library that enhances the language with support for large, multi-dimensional arrays and matrices. It also includes a wide range of high-level mathematical functions designed to operate on these arrays. [11]

MATPLOTLIB- Matplotlib is an incredible data visualization library in Python specifically designed for 2D plots of arrays. It is a versatile data visualization library that can be used across multiple platforms. [19]

SEABORN- Seaborn is a Python data visualization library that is built on top of matplotlib. It offers a high-level interface for creating visually appealing and informative statistical representations of data.
[19]

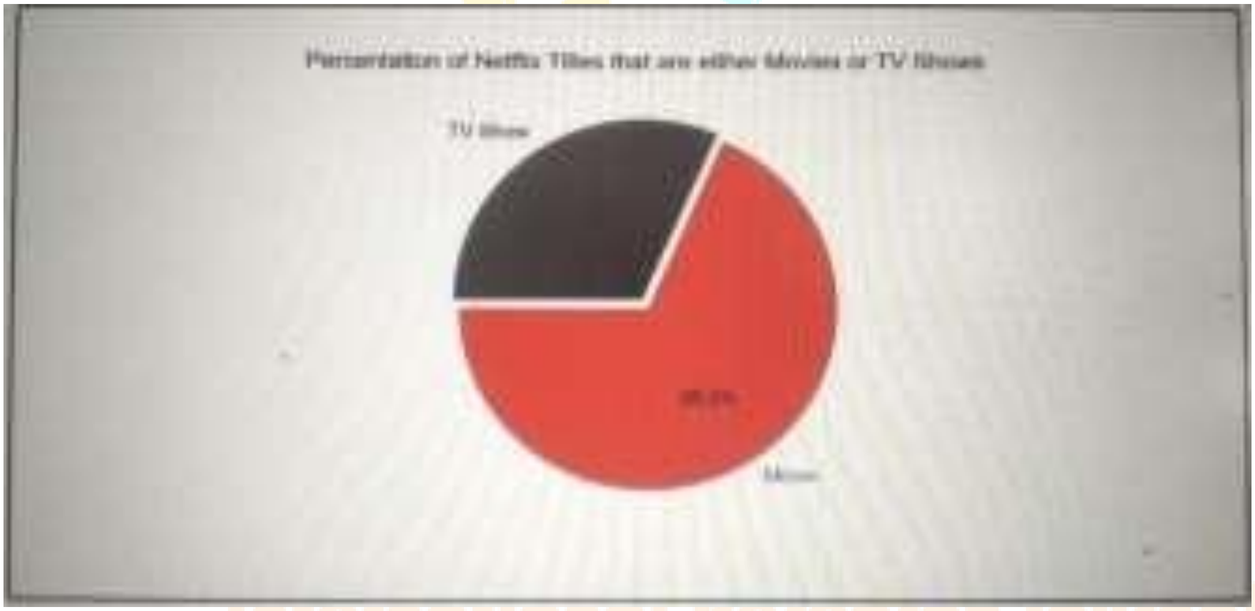
Exploratory Data Analysis

1. ***Movies content by Type***

Let's analyze the entire Movies dataset, which includes both movies and shows, to compare the total number of each and determine the majority. [16].

```
plt.figure(figsize=(12,6))
plt.title(“Percentation of Movies Titles that are either Moviesor TV Shows”)
g = plt.pie(Movies_df.type.value_counts(), explode=(0.025,0.025),
labels=Movies_df.type.value_counts().index, colors=[‘red’,’black’],autopct=’%1.1f%%’, startangle=180)
plt.show. [16]
```

There are over 4,000 movies and almost 2,000 TV shows, making movies the majority. Movie titles make up 68.5% of the total, while TV show titles account for 31.5%. [16]



tile of Movies Titles that are either Movies orTV Shows

2. **Content Volume Over Time**

Next, we will examine the quantity of content added to Movies over the past years. As we are focused on the addition dates of titles to their platform, we will create a "year_added" column to extract the dates from the "date_added" column. [16]

```
fig, ax = plt.subplots(figsize=(13, 7)) sns.lineplot(data=Movies_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)”) [16] plt.legend([‘Total’,’Netflix’,’TV Show’])
plt.ylabel(“Releases”)plt.xlabel(“Year”)
plt.show.
```

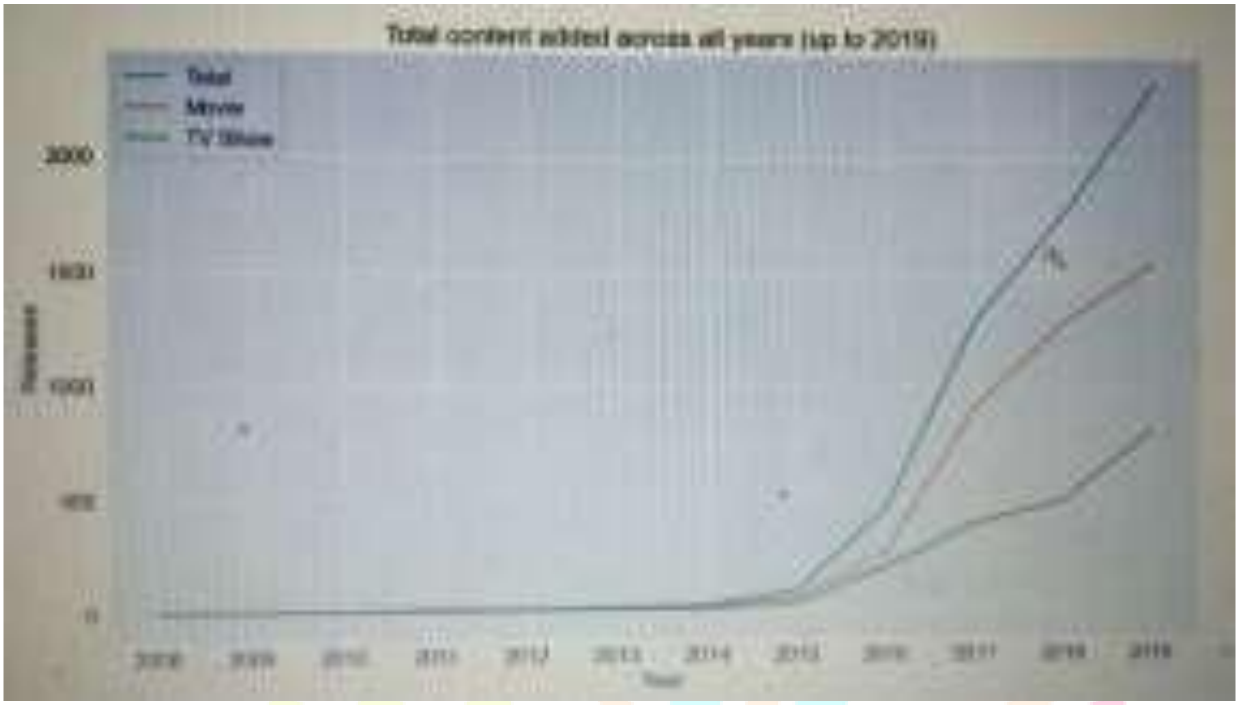


Fig 6: Total added content

Looking at the picture above it indicates that, the most known streaming platform gained its popularity after 2013. After that, the level of added content increased.[16]

3. *Content Production by Country*

Next, we'll explore the countries producing content for Netflix. We'll separate all countries within a film before analysis and remove titles with no available country information. [16]

```
filtered_countries = Movies_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 Movies’) plt.xlabel(‘Titles’)
plt.ylabel(‘Country’) plt.show. [16]
```

From the images below, it's clear that the United States is the top contributor to Netflix's content production. [16]

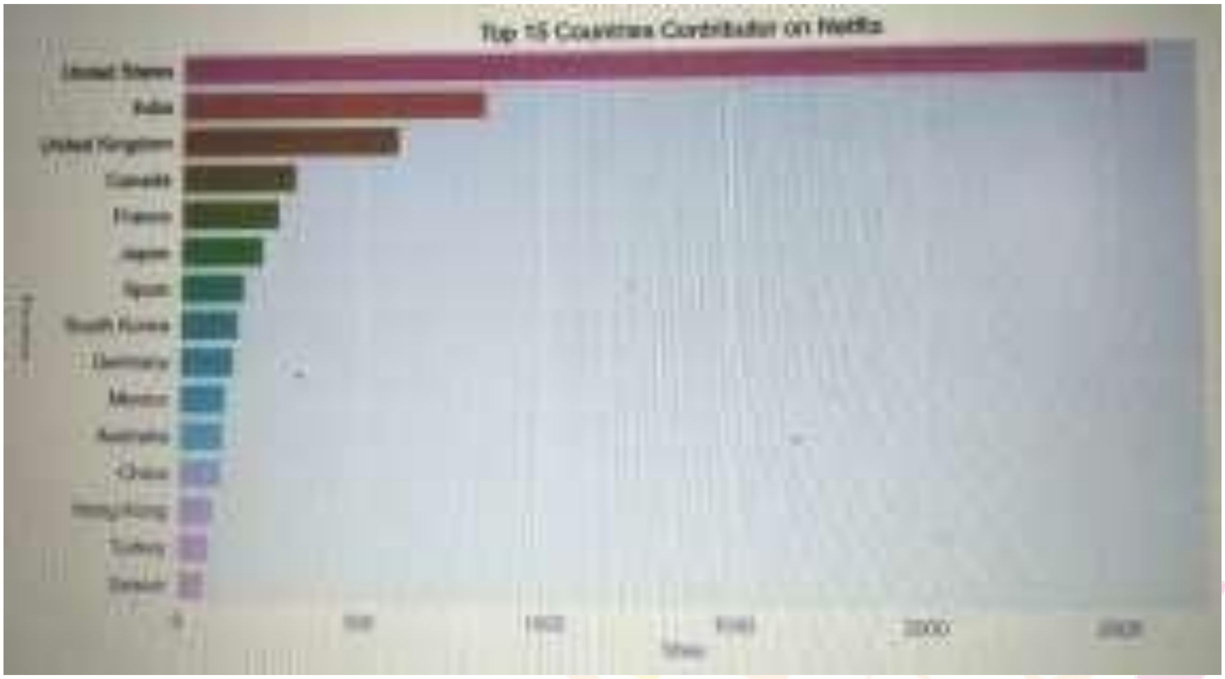


Fig 7: Countries Contributor on Movies

4. ***Top Directors on Movies***

To identify the most popular director, we can visualize the data. [16]

filtered_directors = Movies_df[Movies_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. (medium.com)

The most popular director on Movies, with the highest number of titles, is predominantly international.

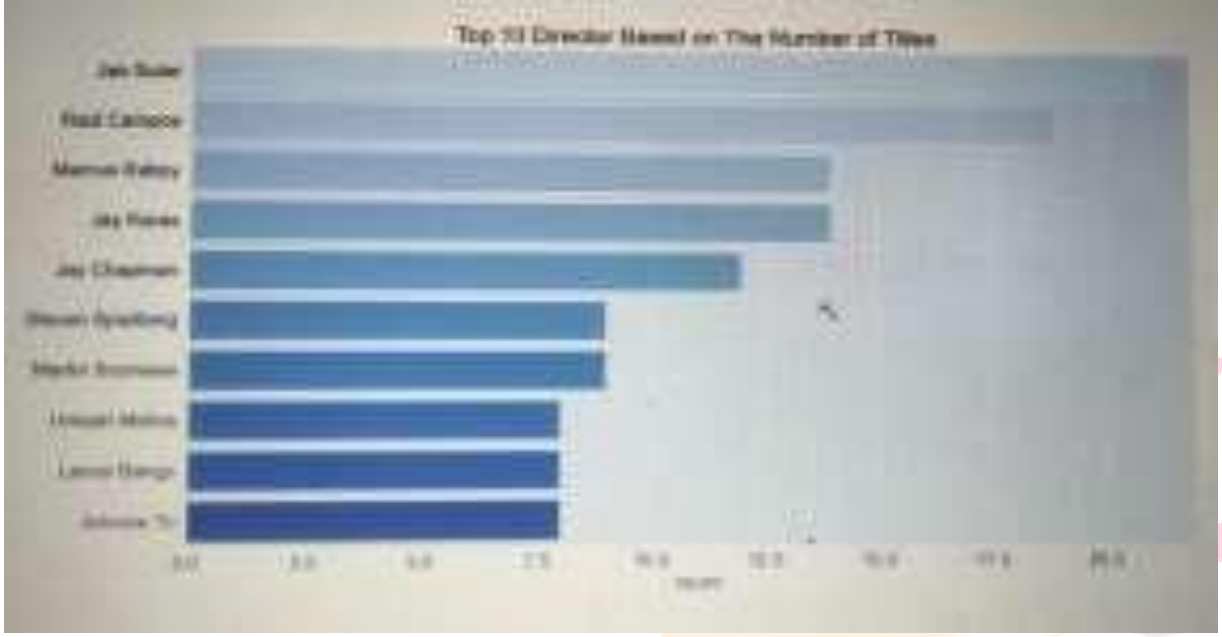


Fig 8: Top 10 Directors Based on Number of Titles

5. ***Top Genres on Movies***

Movies_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 Movies') plt.xlabel('Titles')
plt.ylabel('Genres') plt.show (medium.com).

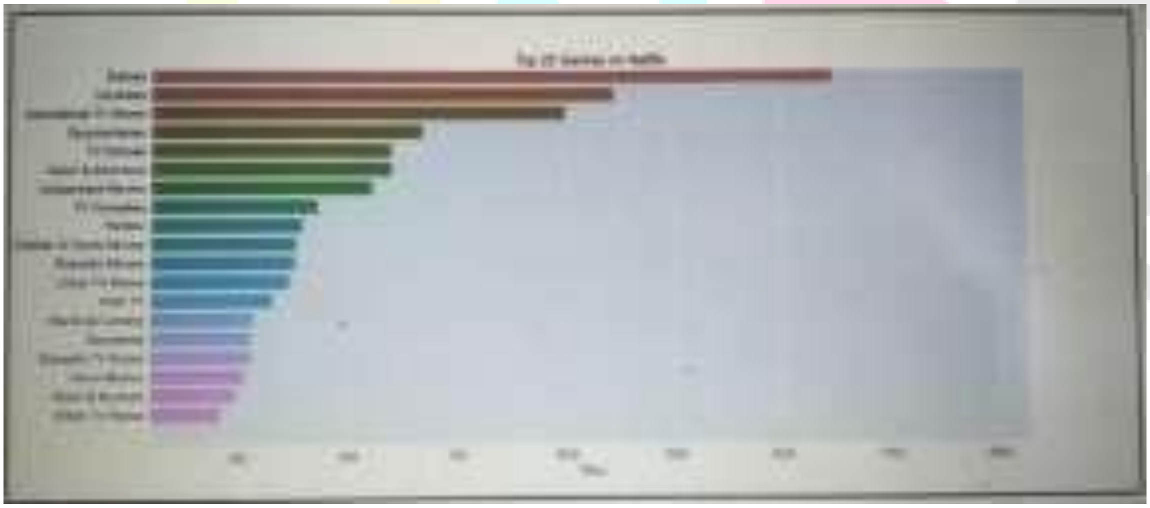


Fig 9: Total content added across all years until 2019

From the graph, it's evident that International content on Netflix takes the top spot, followed by dramas and comedies.

6. ***Top Actor for TV Show on Movies based on thenumber of titles***

```
filtered_cast_shows = Movies_shows_df[Movies_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 ofTitles’)
sns.countplot(y = filtered_cast_shows, order=filtered_cast_shows.value_counts().index[:10], palette=’pastel’)
plt.show (medium.com)
```

The top actor on Movies TV Show, based on the number oftitles, is Takshiro Sakurai [10].

7. ***Top Actor for Netflix on Movies based on thenumber of titles***

```
filtered genres = Movies_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 Movies') plt.xlabel('Titles')
plt.ylabel('Genres') plt.show (medium.com)
```

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

8. ***Amount of Content by Rating***

```
order = Movies_df.rating.unique()
count_Movies = Movies_Movies_df.groupby('rating')['title'].count().reset_index()
count_shows = Movies_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 (medium.com)
```

Fig 11 Top 10 Actors in TV Shows Based on Number of Titles (Left) & Amount of Content by Rating (Movies vs TV Shows) (Right) [20]

V. CONCLUSION

It sounds like you have made some interesting discoveries from your dataset of movie titles. I'd be happy to help you summarize the inferences you've drawn. [20]

REFERENCE:

1. The most content type on Movies is Movies.
2. The popular streaming platform started gaining fractionafter 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 Movies, with the most titles observed for Jan Suter.
5. International Movies is a genre that is mostly in Movies.
6. The most popular actor on Movies TV Shows based on the number of titles is Takahiro Sakurai.
7. The most popular actor on Movies Netflix, 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, "Movies 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 inExploratory 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 IEEETransactions on Industrial Informatics, vol. 13, no. 4, pp. 1891-1899, Aug. 2017, doi: 10.1109/TII.2017.2650204.

[11] www.sist.sathyazbama.ac.in

[12] www.visit.edu.in

[13] www.smujo.id

[14] A review of model and framework for designing mobile learning experience and by Hsu-2015

[15] www.ourgenerationusa.com

[16] www.medium.com

[17] S AISHWARYA-2020-naac.kct.ac.in exploratory Data Analysis on Netflix Data set

[18] www.ece.anits.edu.in

[19] www.docobook.com
[20] www.sist.sathyabama.ac.in
[21] www.medium.com
[22] www.link.springer.com
[23] www.coek.info
[24] www.arxiv.org
[25] www.freepetentsonline.com
[26] www.dx.doi.org
[27] www.asbmur.onlineliberary.widely.comInformation extraction from sound for medical telemonitoring by istrate-2006