To Perform EDA on Netflix Movies and TV shows

Vikas Reddy Bodireddy(HG8118), Vikas rayala(VV6339)

1.Introduction:

Netflix, Inc., a prominent media services provider and production company headquartered in Los Gatos, California, was established in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California. At its core, the company offers a subscription-based streaming service, making a vast library of films and television series accessible to a global audience.

Netflix's streaming service is a household name, widely embraced by individuals from all corners of the world, serving as a key player in the entertainment industry. This exploratory data analysis (EDA) delves into a Netflix dataset, using R libraries such as ggplot and plotly to create informative visualizations and graphs. The dataset comprises information on TV shows and movies available on Netflix as of 2021, with the dataset sourced from Flixable, a reputable third-party Netflix search engine.

The EDA's primary objective is to ensure data cleanliness and organization, establishing a robust foundation for subsequent analysis. By maintaining well-structured and tidy data, Netflix can drive meaningful analyses that have the potential to enhance viewership, boost sales, and refine customer recommendations, further solidifying its position as a leader in the entertainment streaming industry.

2.Data Description:

This is Netflix Movies and TV Shows dataset taken from kaggle, which originated from a third party search engine platform called flixable. The dataset contains information about more than 8000 movies and TV shows by mid-2021.

Using this dataset we can check if there is more movies or Tv shows in the Site, to check which countries view more on the website, check which who are the top directors on the channel, what are the ratings for different type of content.

This dataset has 8807 observations with 12 variables with 11 variables Type, Title, Director, Cast, Country, Date added, Release year, Rating, Dutration, Listed in and Description useful for EDA and 12th variable Show Id used as indexing number for rows, While Date and Release dates are Character variables we have later converted them to Date variables for analysis. Whose description is given in below *Table i*.

Table i

Variables	**Type**	**Description**
Show_Id	numeric	Uniques Id for every Movie/ TV show
Type	categorical	Identifier Used To know If it is a Movie or TV show.
Title	Nominal	Title of Movie/TV Show
Director	Nominal	Director of Movie/ TV Show
Cast	Nominal	Names of cast in the Show/ Movie
Country	Nominal	Where movie produced
Date added	Date Variable	Date added to netflix
Release year	Date Variable	Actual Release of show
rating	Categorical	Types of ratings

duration	Numeric	Duration of the Show/ Movie
Listed in	Categorical	Genre(Documentaries, International shows, etc.)
Description	Nominal	Summary Description of the Show/Movie

The dataset has many missing values, Which we have tried to eliminate and produce efficient Results. But it turned out there are many missing values in Directors names which can reduce the dataset size, So we have used the dataset as it is and remove these values or add dummy values wherever needed.

The Summary statistics of the dataset are shown in fig 2.1.

3. Methods:

3.1 Data Cleaning and Modification:

During the data exploration and cleaning process, it came to our attention that three variables had a significant number of missing values. Specifically, the "Director" variable had 2634 missing values, the "Cast" variable had 825 missing values, and the "Country" variable had 831 missing values. While we considered removing rows with missing values to tidy up the dataset, we decided against removing more than 100 rows due to the missing values in the "Cast" and "Country" variables. However, we recognized that removing 2634 rows due to missing director information would substantially impact the dataset's results. As a result, we opted to retain rows with missing values in these variables for further exploration.

Our primary focus is on uncovering key insights that are essential for understanding analytics and how they can enhance our business. To achieve this, we conduct various analyses to assess the data's relevance and its potential to drive business improvements. Which are well understood from below Graphical analysis.

3.2 Graphical Analysis and Exploration:

3.2.1 Movies vs TV Shows:

The Graphical analysis to check the ratio of Movies versus Tv Shows show that there is a strong domination of movies on Tv Shows(nearly 70% of content streamed are movies while the rest of 30% TV Shows). As shown in *fig 3.2.1-1*.

The removal of rows with missing values resulted in a skewed dataset, with 97% movies and only 3% TV shows. as shown in *fig 3.2.1-2*.

This happened because the rows with missing data, especially regarding director names, cast, and production country, predominantly contained TV shows. This outcome doesn't align with our analysis goals and is undesirable. The skewed data could lead to misleading results, and we may need to address this issue to ensure a more accurate representation of the content distribution.

3.2.2 Content over Years:

When we examined the content trends over the years using the release year and content type, we made an intriguing discovery. Between 1925 and 1975, movies and TV shows were released in roughly equal proportions. However, from 1976 onwards, there was a significant increase in content production for both movies and TV shows. Notably, movies outpaced TV shows, with double the number of movies produced compared to TV shows during this period. The peak years for both movies and TV series releases were 2013 and 2018. As shown in *fig 3.2.2-1* .

3.2.3 Content by Country:

When examining the countries that have contributed the most content to the streaming platform, the United States clearly stands out as the leading producer of both movies and TV shows. It's followed by the United Kingdom, which is notable for its contribution to TV shows, and India, which takes the lead in the production of movies. However, when we consider the available content on the platform, India surpasses the United Kingdom in terms of the quantity of content.

As always, the United States maintains its top position. Which can be seen from fig *3.2.3-1 to 3.2.3-3*.

3.2.4 Successful Directors:

While conducting our analysis to identify directors who have left a significant mark on the platform's content, we found that Director Rajiv Chilka has set a remarkable record by directing over 18 shows and movies, making a historic impact. He is closely followed by directors Raul Compos and Jan Suter, who have each directed nearly 16 shows. Refer *fig 3.2.4-1* .

3.2.5 Other Key aspects:

Next, we delved into several key aspects of the platform's content. We examined the year with the highest number of releases, observed different ratings for content types, analyzed the most viewed genres, determined the average screen time for movies, and investigated the most commonly occurring words in the descriptions of movies and shows. .

Our analysis revealed a notable increase in the rate of content releases, both for movies and TV shows, starting from 2010. The period from 2016 to 2019 witnessed the highest number of movie releases, while TV shows maintained a consistent growth trajectory, eventually surpassing movies in 2021. *Refer fig 3.2.5-1* .

In terms of content ratings, we found that TV-MA is the most prevalent, followed by TV-14, collectively making up a third of the platform's content. This indicates a significant portion of the content is not suitable for children and teenagers under the age of 17. Refer *fig 3.2.5-2* .

Regarding genres and themes, the platform is primarily dominated by international shows, followed by dramas and comedies.

Our time series analysis demonstrated a decrease in the duration of movies, particularly since digital platforms started streaming content. Notably, there was a significant increase in movie duration between 1960 and 1965. *Refer fig 3.2.5-3* .

Finally, our investigation into frequently used words in movie and show descriptions revealed the prominence of words like "life," "women," "documentary," "love," "teen," "friend," "series," "family," and "world." as shown in *fig 3.2.5-4* .

These findings provide valuable insights into the content landscape and trends on the platform.

4. Results:

Based on the analysis and visualizations presented in our Netflix project report, the key high-level takeaways can be summarized as follows:

- Approximately 70% of the content on Netflix consists of movies, with TV shows making up the remaining 30%.
- The annual release of titles on Netflix has seen a consistent increase over time, with 2018 marking the year with the highest number of movie releases.
- The United States is the primary producer of Netflix content, contributing over 50% of the titles.

India and the UK also play significant roles in content production.

- TV-MA, TV-14, and R-rated content are prevalent in the Netflix catalog, indicating a substantial presence of adult-oriented content.
- The average duration of movies on Netflix has remained relatively stable, typically ranging from 90 to 120 minutes.
- Netflix's content library is rich in genres, with dramas, comedies, action, and international movies and shows being among the most common.
- Descriptions of movies often feature words like "life," "family," and "friends," while TV shows mention "series," "stories," and "world."
- There are a few directors who have directed 10 or more movies or shows for Netflix, signifying their significant contributions to the platform.

These insights provide actionable information for optimizing content selection, release planning, audience targeting, and enhancing the overall viewer experience on the platform, ultimately contributing to business growth and success.

5. Conclusion:

In conclusion, this exploratory analysis of the Netflix content library revealed key insights into the platform's focus on movies, rapid growth in production, distribution across countries, and prevalence of adult content and certain genres. The results quantify Netflix's priorities and can inform future decisions around content and business strategy. Further analysis on additional variables would provide further valuable perspectives.

6. Appendix:

GitHUB ((R-Code) :

[Vikas Github]

[Vikas Rayala github]

7. Reference: [R For DataScience] [Kaggle]

8. Figures:

```
#Summary Stats
summary(netflix_data)
```

```
  show_id             type              title             director
Length:8807        Length:8807        Length:8807        Length:8807
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character


     cast             country           date_added         release_year
Length:8807        Length:8807        Length:8807        Min.   :1925
Class :character   Class :character   Class :character   1st Qu.:2013
Mode  :character   Mode  :character   Mode  :character   Median :2017
                                                         Mean   :2014
                                                         3rd Qu.:2019
                                                         Max.   :2021
   rating            duration          listed_in          description
Length:8807        Length:8807        Length:8807        Length:8807
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

fig 2.1

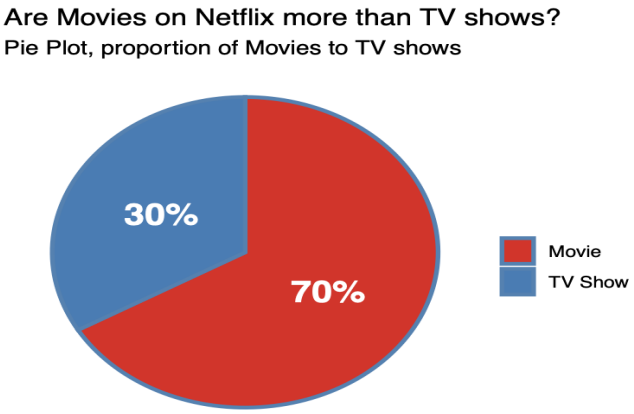

fig 3.2.1-1

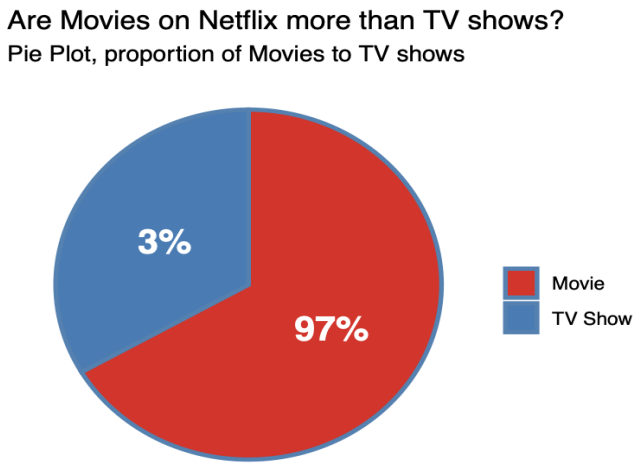

fig 3.2.1-2

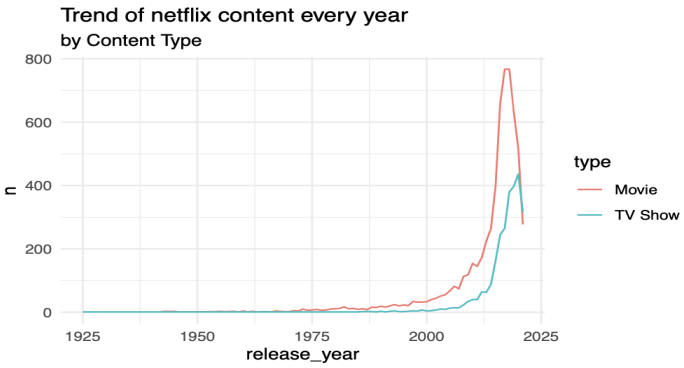

fig 3.2.2-1

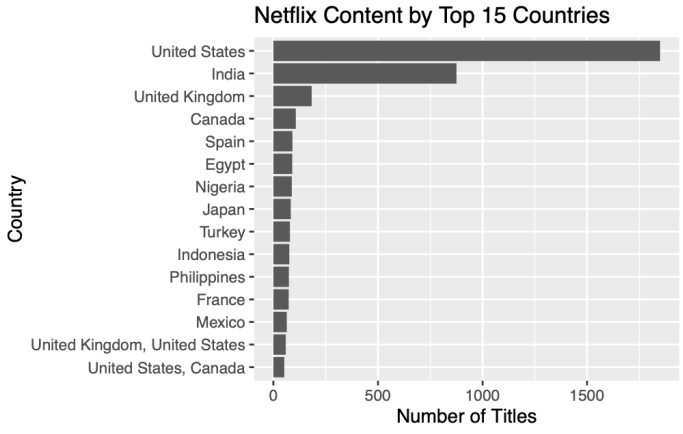

fig 3.2.3-1

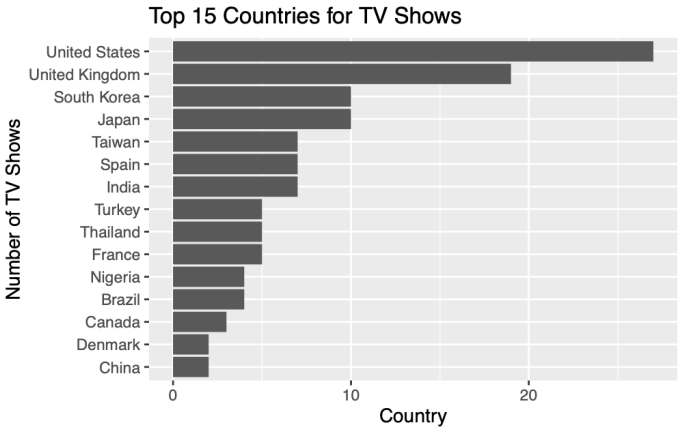

fig 3.2.3-2

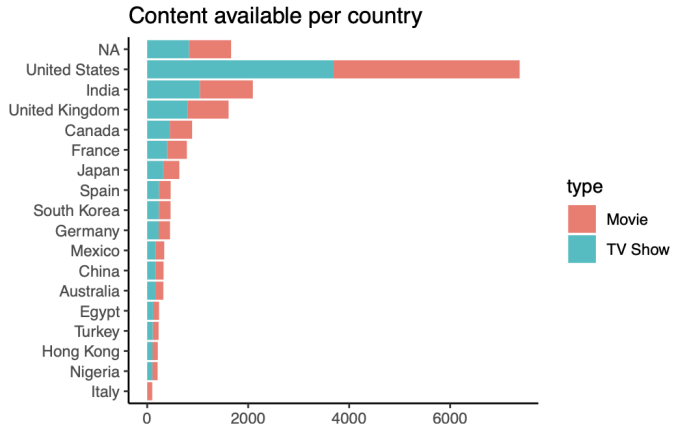

fig 3.2.3-3

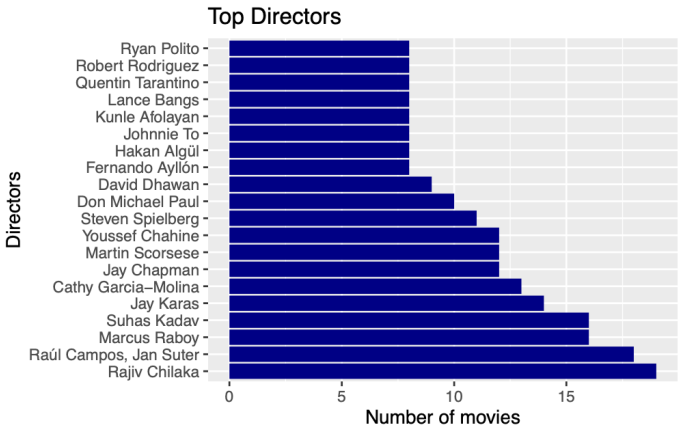

fig 3.2.4-1

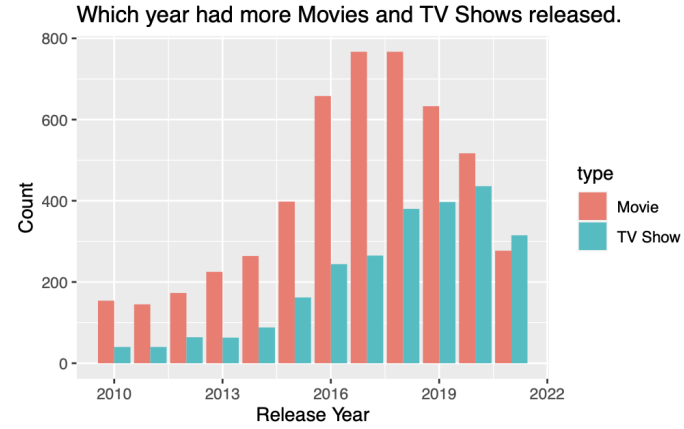

fig 3.2.5-1

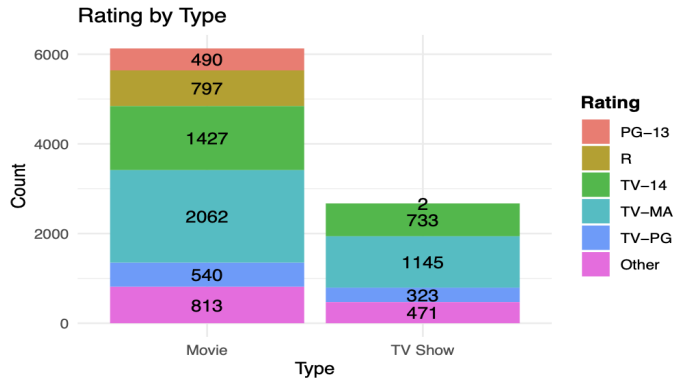

fig 3.2..5-2

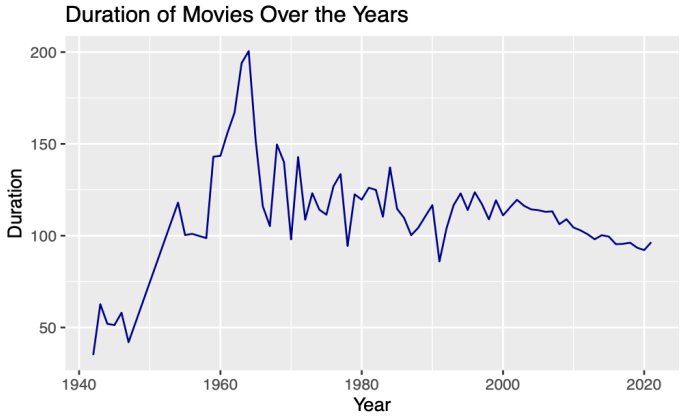

fig 3.2.5-3

fig 3. 2.5-4