Data Visualisation for Netflix Movie & TV Show Dataset

Garima Bansal , Konika Mandal, Nikunj Pansari
Department of Mathematics
Indian Institute of Technology, Patna
garima_2211mc05@iitp.ac.in , konika_2211mc17@iitp.ac.in , nikunj_2211mc21@iitp.ac.in

***Abstract*—Data Visualisation for the Netflix movie & TV show dataset provides a good glimpse and its real-time applications for the recommendation of the appropriate movie or TV show according to the individual's interest. Analysis using Pie Chart, Count plot, Scatter plot & bar plot defines a graphical relationship between the ratings of different movies and TV shows for different years. Additionally, WordCloud is visualized to depict the text data (words), defining the frequency or importance of each of the words. Ultimately, all these visualizations would be helpful for the inference of an effective movie or TV show.**

***Index Terms*—Data Visualisation , Netflix Movie Dataset , movie recommender system.**

I. Introduction

Netflix Movie and TV Shows dataset consists of a variety of movies & TV shows with different attributes namely title, director, cast, country, release date, rating, and duration which also contains some parameters as the metadata helps understand the data efficiently. Real-time applications for this dataset could be beneficial in building a movie or show recommender system based on some metrics like ratings, duration, release year, etc. Choosing a movie or TV show to watch can be difficult at times, and usually when a similar category of movie or shows are available. Thus, there comes the filter of rating and duration, which is useful most of the time. The main objective behind the data visualization is to analyze and explore the content released by Netflix thereby inferring useful insights.

Section II highlights the background and related work for the Netflix dataset and the corresponding metadata description useful for data visualization. Section III illustrates the visualization in the form of pie charts, bar plots, count plots, and scatter plots defining the relationship between important attributes like different ratings and the corresponding movie or TV show. In addition to this, WordCloud describes the important textual data points. Section IV states the analysis of these visualizations providing a good glimpse of the appropriate rating corresponding to a movie. Finally, section V concludes with a summary of the movie & TV show ratings.

II. Background & Related Work

Researchers have defined the exploratory and sentiment analysis of Netflix data using systematic & insightful usage of Python libraries and the corresponding utility tools. Analysis of these data would be highly beneficial in effective decisions making for organizations as well as individuals. Further, it would be helpful in inferring interactive solutions and methods useful for data exploration [3]. Another work depicts the data analysis of the Netflix dataset highlighting different genres of movies, produced by different directors, using python for the visualization. This is carried out by extracting the useful attributes or features from the Netflix dataset [2]. The Netflix dataset consists of various attributes like show_id, type, title, director, cast, country, date_added,release_year, rating, duration, listed_in, and description. Now, metadata used in our data visualization includes rating, type of show & release_year [1].

```
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 movie/show produced
Date_added: date it was added on Netflix
Release_year: release year of movie/show
Rating: TV rating of the movie/show
Duration: in minutes or no. of seasons
Listed_in: genre
Description: the summary description
```

There are different types of ratings that can be defined:

```
TV-MA: mature audience(unfit,under 17 age)
PG-13: parents cautioned(unfit,under 13 age)
PG: parental guidance(unfit, aged 8 or above)
TV-14: parental guidance(unfit under 14 age)
TV-PG: parental guidance(unfit, younger age)
TV-Y: programs aimed at age 2-6
TV-Y7: most appropriate for age 7 & up
R: under 17 requires accompanying parent
TV-G: all ages not necessarily children show
G: appropriate for people of all ages
NC-17: no one 17 and under admitted
NR: has not been submitted for a rating
TV-Y7-fv: program with fantasy violence
UR: uncut version of film that was submitted
```

III. Methodologies

We performed data gathering & data cleaning for the exploration and visualization of the Netflix data.

- **Data Gathering**: Collection of the data from an open-source platform.
- **Modules & Imported Libraries**: numpy, pandas, pandas_profiling, matplotlib, seaborn, worldcloud & counter from collections, are some of the python packages utilised.
- **Data Cleaning**: Utilisation of python libraries for getting rid of redundant data, or minimizing the presence of anomalies prior to the deployment of the data for analysis.Data Cleansing is the process of detecting and changing raw data by identifying incomplete, wrong, repeated, or irrelevant parts of the data.

TABLE I
Before data cleaning

Show id	director	release year	rating
s5542	Louis C.K.	2017	74 min
s5795	Louis C.K.	2010	84 min
s5814	Louis C.K.	2015	66 min

It can be seen from (Table I), in the column 'rating' with values 74 min, 84 min, and 66 min are the irrelevant data which is replaced as 'NR'(No rating) since the correct 'rating' of those particular movies or shows is not known (Table II).

TABLE II
After data cleaning

Show id	director	release year	rating
s5542	Louis C.K.	2017	NR
s5795	Louis C.K.	2010	NR
s5814	Louis C.K.	2015	NR

For the data visualization of the Netflix dataset, data analysis is illustrated as:

- **Correlation Heat Map**: It illustrates the dependency factor within the attributes. The features such as type & release_year have a medium shade of blue color which indicates that there is a variation among these attributes. Additionally, the release_year directly doesn't imply the type of the movie or the show. Overall, the HeatMap represents the degree of relationship between -1 and 1 (Table III). The dark blue color depicts a high correlation, while the light blue color shows less correlation between features namely release_year, type & rating in the dataset (fig 1).

TABLE III
Correlation matrix of heatmap

Attributes	release year	type	rating
release year	1.000	0.166	0.136
type	0.166	1.000	0.342
rating	0.136	0.342	1.000

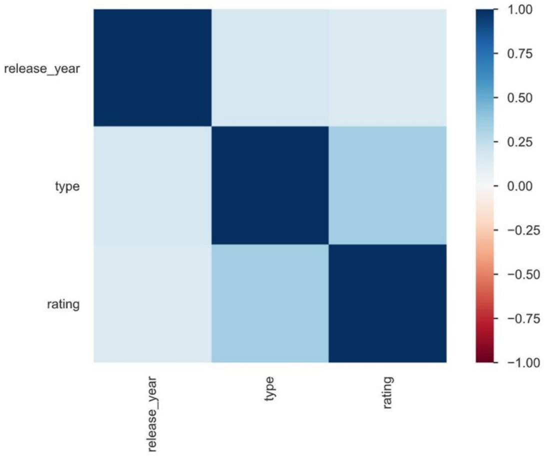

Fig. 1. HeatMap : Correlation Matrix

- **Count plot**: The x-axis shows type of rating, while the y-axis represents the count of those ratings. Using seaborn.countplot() (fig 2).

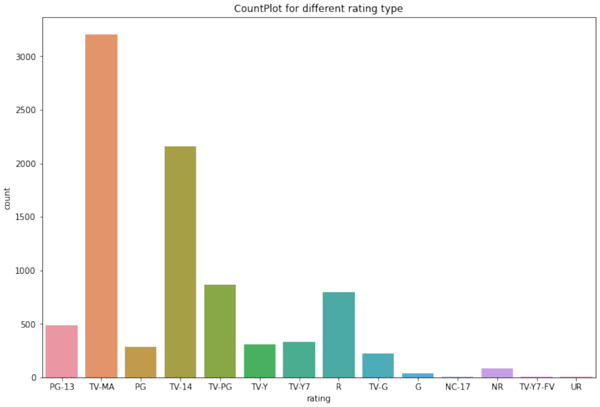

Fig. 2. Count Plot for different rating type

- **Scatter plot**: Plotted the data points on horizontal & vertical axis in an attempt to show how much rating is affected by type i.e. tv shows and movies. Also, the scatter plot shows the comparison between the two types using seaborn.scatterplot() (fig 3).

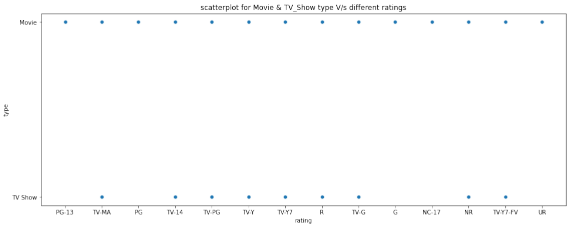

Fig. 3. Scatter plot for different ratings

- **Pie-chart**: Categorical variables such as ratings are used to represent the percentage of a particular data from the whole pie. It shows TV-MA is the highest rated, while UR & NC-17 are the least rated in the dataset (fig 4).

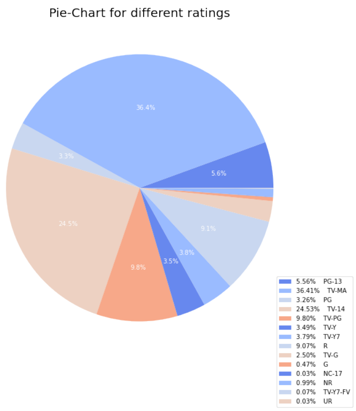

Fig. 4. Pie-Chart for different ratings

- **Line plot**: Calculated the frequencies of movies and tv shows which were released in different years and are available on Netflix by using value_counts().plot(kind="line") from the pandas' library (fig 5,6). Also, plotted value_counts in descending order. Here the x-axis depicts the year, while the y-axis depicts the frequencies of TV shows & movies.

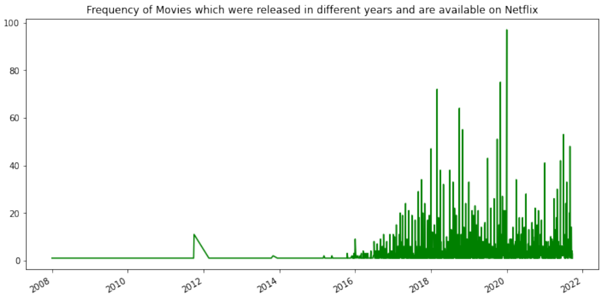

Fig. 5. Frequency of Movie added for different release year

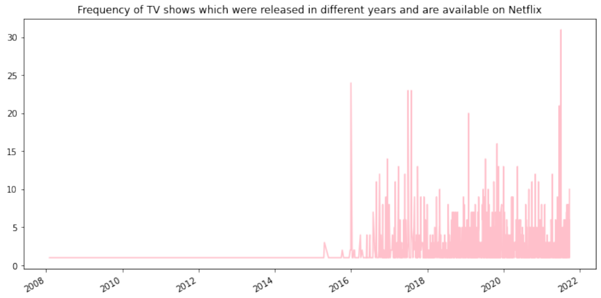

Fig. 6. Frequency of TV show added for different release year

- **WordCloud**: Utilization of wordcould() from word cloud library and matplotlib.subplots() to highlight significant textual data points for maintaining the figure size. Here the size of each word indicates its importance or frequency in the whole dataset (fig 7).

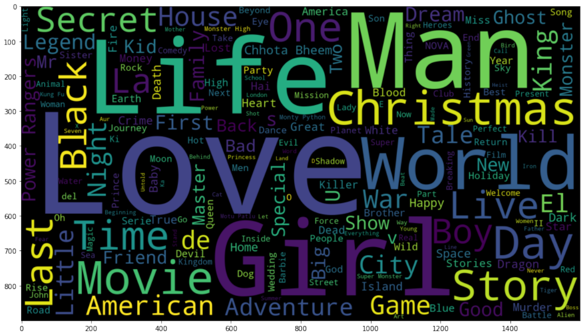

Fig. 7. WordCloud for different words

IV. Analysis & Results

From the data cleaning & data visualization of the Netflix dataset, the following results are obtained:

- The dataset is pretty clean except for the **rating** , **date_added** , **duration** , **director** , **cast** & **country** columns. Next, we imputed the empty values in the data for these attributes.

- Counted the total number of movies and TV shows from the dataset. It can be observed that count of movies is almost two times the count of TV shows.

- Table IV depicts the count of total number of movies and TV shows from different countries, thus observed that the United States has added the maximum number of Movies and TV shows on Netflix.

TABLE IV
Countrywise counts of movies and TV shows

Country name	**Counts of movies and TV shows**
United States	2818
India	972
unavailable	831
United Kingdom	419
Japan	245
South Korea	199
Canada	181
Spain	145
France	124
Mexico	110

- Also counted the total number of ratings for movies and TV shows. From fig 2, it can be observed that maximum movies and TV shows have got a TV-MA rating.

- Now, compared the rating of the TV show vs the rating of the movie using a scatter plot. And from fig 3, we can observe that no TV show has got PG-13, PG, G, NC-17, and UR ratings whereas movies have got these ratings.

- Fig 4 depicts how much content is available with different types of ratings on Netflix. It is evident from the dataset that 36.41% of the total available content has a TV-MA rating and only 0.03% of the total content has a UR rating.

- Movie and TV Shows tend to have an increasing trend every year. Movie outperform every year than TV shows. From fig 5, it is evident that the Highest number of Movies was added around the year 2020 on Netflix. However, fig 6 interprets that the highest number of TV show was added around the year 2022.

- Plotted the top 10 genres for the content provided in the data and drill deeper into them to get insights enabling us to filter out the genres our users are interested in. It is observed that the users are mostly interested in Dramas, International Movies in movies, and kid's TV in TV shows available on Netflix (fig 8).

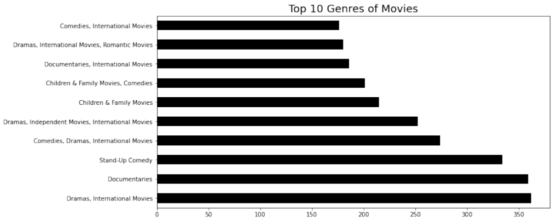

Fig. 8. Top 10 Genre of Movies

- Plotted the top 10 genres for the content provided in the data and drill deeper into them to get insights enabling us to filter out the genres our users are interested in. It is observed that the users are mostly interested in Dramas, International Movies in movies, and kid's TV in TV shows available on Netflix (fig 9).

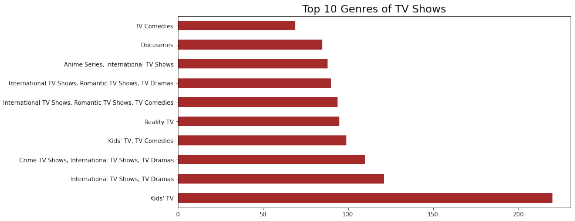

Fig. 9. Top 10 Genre of TV shows

Created a word cloud based on the relevance of the Genres in the dataset. Fig 7 shows the frequency of the word ”Love” is maximum in the dataset.

V. Conclusion

Performed a relatively precise data analysis to determine the genre of a given TV show & movie. Processing and narrowing down the features of the Netflix dataset by identifying the top ten countries with the top ten genres the audience watch and feeding this data to figure out the trend over the years using which was useful in categorizing the data in different age groups. Techniques like expanding the dataset & feature set selection might have the potential to improve upon results in the future.

References

[1] https://www.kaggle.com/datasets/shivamb/netflix-shows.

[2] Vybhav Achar Bhargav, Seongwoo Choi, and David Haddad. Data analysis on netflix datasets. 03 2022.

[3] Karthik Babu Vadloori and Shriya Madhavi Sanghishetty. Exploratory and sentiment analysis of netflix data. *International Journal of Engineering Research & Technology (IJERT)*, 10(9):213–217, 2021.