Visualization of Netflix IMDb Scores Dataset in Streamlit

Aditya Chaudhary, Susil Raj Neupane
Student, Department of Computer Science and Engineering
Kathmandu University, Dhulikhel Nepal
chaudariaditya@gmail.com, susilrajneupane@gmai.com

***Abstract*—This project explores the visualization of IMDb scores and votes for Netflix content using Python and Jupyter Notebook, with a focus on interactive presentation through the Streamlit library. Leveraging data retrieved from Kaggle[1], the Python programming language is employed to analyze and visualize the IMDb scores and corresponding votes for Netflix shows and movies, and pandas library is used for the data wrangling whereas the seaborn and plotly are used to give pictorial representation of data. Jupyter Notebook facilitates an organized and interactive development environment, while Streamlit is utilized to create a web-based interface for a dynamic and user-friendly demonstration. Through this innovative approach, the project aims to provide a visually engaging platform for users to explore and understand the distribution and trends of IMDb scores and votes within the Netflix content catalog.**

***Impact Statement* — Our "Netflix IMDb Scores and Votes" app, powered by Streamlit, Python, and key libraries, transforms analytics and user engagement. It simplifies interpreting IMDb score and vote distributions for Netflix content, aiding data-driven decisions. The user-friendly interface encourages personalized exploration, and automated processing enhances efficiency. This scalable solution positions us competitively in the dynamic streaming industry, facilitating meaningful insights and an enriched user experience.**

***Index Terms*—IMDb, Streamlit, Pyplot, Seaborn.**

I. INTRODUCTION

The advent of digital streaming services has revolutionized the way we consume media. Among these services, Netflix stands out as one of the most popular, boasting a vast array of movies and TV shows from various countries. This project aims to perform a data analysis and visualization based on the IMDb scores and votes of Netflix content. The focus is on understanding the relationship between scores, votes, and various other factors in the Netflix dataset. Data visualization will play a crucial role in this analysis, as it allows us to understand complex data sets by representing them in a graphical format. Tools like Python's libraries Matplotlib and Seaborn, along with Plotly and Streamlit, can be used to create interactive web dashboard.

The primary goal of this project is to conduct a comprehensive analysis of IMDb ratings to gain insights into audience preferences for movies and TV shows on Netflix. In addition to this overarching objective, specific goals include identifying and highlighting top-rated content to assist users in discovering the most well-received movies and shows. The project also aims to investigate the dynamic nature of IMDb ratings and votes over time, offering valuable insights into the evolving popularity of Netflix content. A key aspect of achieving these objectives involves leveraging Streamlit to develop an interactive platform. This platform is designed to empower users, enabling them to effortlessly explore and engage with the visualized data, fostering a more immersive and informed viewing experience on Netflix.

II. METHODOLOGY

A. Data Cleaning and Preprocessing:

a. *Data Retrieval and Preprocessing*: The initial phase involves getting IMDb scores and vote data for Netflix content. Python, with the Pandas library, is employed to efficient manipulate and clean the data, ensuring it is ready for analysis. This step is crucial for accurate and meaningful visualizations.

```
df.head()
```

	index	id	title	type	description	release_year	age_certification	runtime	imdb_id	imdb_score	imdb_votes
0	0	tm84618	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	1976	R	113	tt0075314	8.3	795222.0
1	1	tm127384	Monty Python and the Holy Grail	MOVIE	King Arthur, accompanied by his squire, recrui...	1975	PG	91	tt0071853	8.2	530877.0
2	2	tm70993	Life of Brian	MOVIE	Brian Cohen is an average young Jewish man, bu...	1979	R	94	tt0079470	8.0	392419.0
3	3	tm190788	The Exorcist	MOVIE	12-year-old Regan MacNeil begins to adapt an e...	1973	R	133	tt0070047	8.1	391942.0
4	4	ts22164	Monty Python's Flying Circus	SHOW	A British sketch comedy series with the shows ...	1969	TV-14	30	tt0063929	8.8	72895.0

Fig: Original sample of the dataset.

```
df.head(5)
```

index	id	title	type	release_year	age_certification	runtime	imdb_id	imdb_score	imdb_votes
0	tm84618	Taxi Driver	MOVIE	1976	R	113	tt0075314	8.3	795222.0
1	tm127384	Monty Python and the Holy Grail	MOVIE	1975	PG	91	tt0071853	8.2	530877.0
2	tm70993	Life of Brian	MOVIE	1979	R	94	tt0079470	8.0	392419.0
3	tm190788	The Exorcist	MOVIE	1973	R	133	tt0070047	8.1	391942.0
4	ts22164	Monty Python's Flying Circus	SHOW	1969	TV-14	30	tt0063929	8.8	72895.0

Fig: Dataset after preprocessing

b. *Initial Data Examination*: Commencing our analysis, we performed an initial dataset overview to understand its structure, datatypes and identify any missing values. This step was crucial for assessing data quality and guiding subsequent cleaning procedures.

III. VISUALIZATION

During the exploratory data analysis (EDA) phase, Matplotlib and Seaborn emerge as indispensable tools, significantly influencing the depth and quality of our analytical insights. These

powerful libraries not only play a key role but also take center stage in transforming raw data into meaningful visualizations. By harnessing the capabilities of Matplotlib and Seaborn, we are able to craft a diverse range of visual representations, including histograms, scatter plots, and box plots. These visualizations serve as dynamic windows into the intricate details of the IMDb scores and votes dataset, offering a nuanced and comprehensive understanding of the distribution patterns and relationships within the data. Matplotlib's versatility, coupled with Seaborn's high-level interface, empowers our exploratory efforts, allowing us to uncover hidden trends, outliers, and correlations that may influence audience preferences on Netflix. As a result, this phase not only lays the foundation for subsequent analyses but also ensures that our data-driven decisions are grounded in a thorough and visually enriched exploration of the dataset.

In addition, we have the sidebar where we have different filter options. We have multiselect options to choose either movies or shows, along with a double-slided range slider to select a range of years. This feature then adjusts the charts and plots on our dashboard.

A. Category distribution of the MOVIE and SHOW

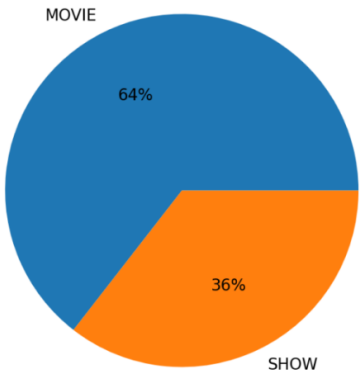

Fig: Data Overview

The chart shows the overall overview of the dataset which can be filtered thorough MOVIE and SHOW, with movies constituting 64% and shows representing 36%.

B. IMDb Scores Distribution

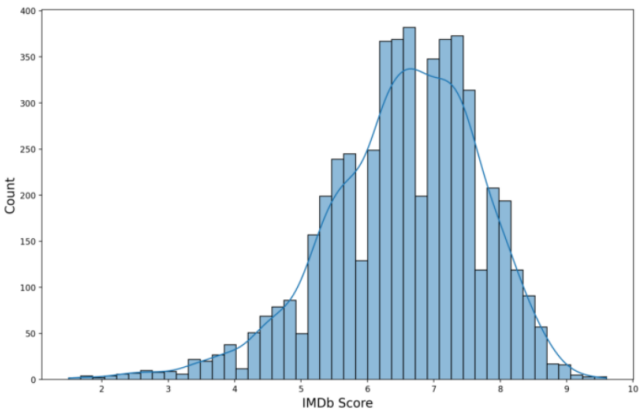

Fig: IMDb Scores Distribution

The histogram displays how many movies and shows have different IMDb scores. The graph indicates that most of them have a score around 6.6. There are only a few with scores below 4 or above 9.

C. Number of releases over years

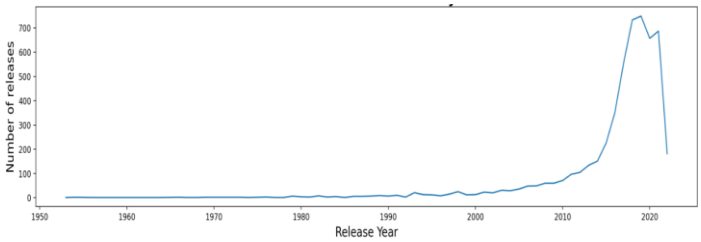

Fig: Number of releases over years

The line chart tracks how many movies and shows Netflix had each year. There aren't many old movies before 1990, probably because they cost a lot to stream, and not many people watch them. After 2010, there's a lot more content, but it drops after 2019 because COVID-19 stopped filming.

E. IMDb Scores vs Runtime

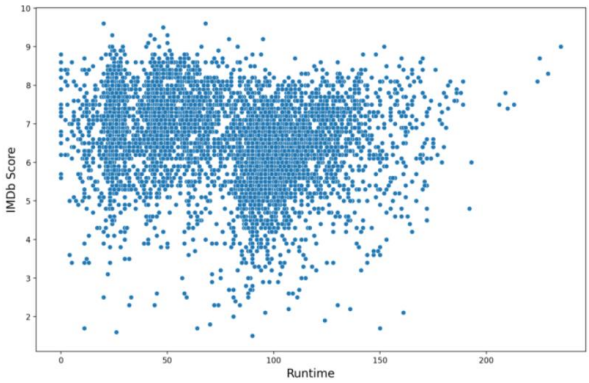

Fig: IMDb scores vs runtime

This scatter plot compares IMDb scores to movie runtimes. Each point represents a film, showcasing whether there's a connection between a movie's duration and its IMDb rating. This visual analysis offers a brief yet insightful snapshot of the relationship between these two crucial elements in film evaluation.

G. IMDb Score based on Age Certification

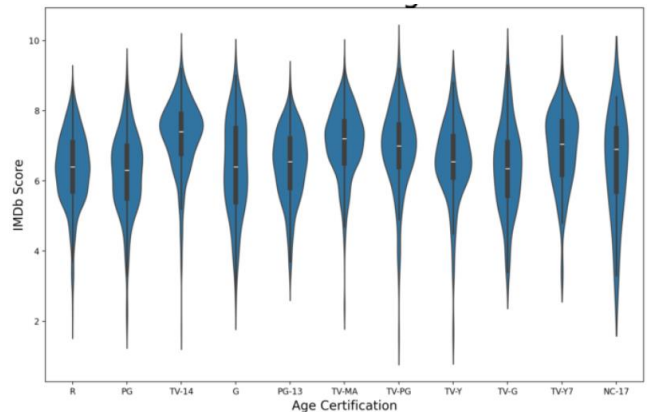

Fig: IMDb Scores based on Age Certification

The given violin plot shows the distribution of IMDb score of multiple category of age-certifications. For each category it shows the minimum, 1st quartile, median, 3rd quartile and maximum IMDb score. For instance: PG category median IMDb score is around 6.3.

G. Top IMDb Scores by Title

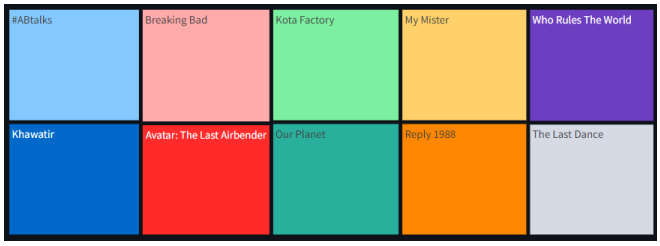

Fig: Top IMDb scores by Title

The given treemap shows that the top ten rated in IMDb on which “ABtalks” holds on first position whereas the “The Last Dance” on the tenth position.

IV CONCLUSION

In conclusion, our project has effectively developed and deployed a Streamlit dashboard visualizing Netflix IMDb scores. The dashboard showcases key insights such as top-rated movies and shows, age certifications, runtime, and more. Leveraging the Streamlit framework, we've created an interactive and user-friendly platform for exploring Netflix content. The successful deployment on the Streamlit community cloud[2] ensures accessibility and availability to a wider audience, contributing to an enriched viewing experience.

REFERENCES

[1] *Netflix IMDB scores*. (2023, December 3). Kaggle. https://www.kaggle.com/datasets/thedevastator/netflix-imdb-scores

[2] *Streamlit Community Cloud • Streamlit*. (n.d.). https://streamlit.io/cloud