A PROJECT REPORT

on

“Netflix Data Analysis”

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

Tools and Technique Laboratory

BY

Divita Topno	2105455
Satvik Agarwal	21052869
Hardik Chauhan	21052904
Zoyah Afsheen Sayeed	21052990
Abhinav Aakash	21053242

UNDER THE GUIDANCE OF
Prof. Deependra Singh

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024
March 2024

A PROJECT REPORT

on

“Netflix Data Analysis”

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

Tools and Techniques Laboratory

BY

Divita Topno	2105455
Satvik Agarwal	21052869
Hardik Chauhan	21052904
Zoyah Afsheen Sayeed	21052990
Abhinav Aakash	21053242

UNDER THE GUIDANCE OF
GUIDE NAME
Prof. Deependra Singh

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAE, ODISHA -751024
March, 2024

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certified that the project entitled

"Netflix Data Analysis"

submitted by

Divita Topno	2105455
Satvik Agarwal	21052869
Hardik Chauhan	21052904
Zoyah Afsheen Sayeed	21052990
Abhinav Aakash	21053242

is a record of Bonafede work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2024, under our guidance.

Date: 25/03/2024

(Prof. Deependra Singh)
Project Guide

Acknowledgements

We are profoundly grateful to **Prof. Deependra Singh** of **Affiliation** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. His critical reviews and valuable suggestions at different stages of the project were of immense help.

Divita Topno
Satvik Agarwal
Hardik Chauhan
Zoyah Afsheen Sayeed
Abhinav Aakash

ABSTRACT

The project "Netflix Data Analysis" aims to provide a comprehensive analysis of the viewing habits and preferences of Netflix users. Leveraging a large dataset of Netflix's content library and user viewing history, the project employs advanced data analysis techniques to uncover patterns and trends.

Key areas of exploration include genre popularity, viewing time patterns, and regional preferences. The project also investigates the correlation between a show's attributes (such as genre, cast, and director) and its viewership numbers.

The insights derived from this analysis can assist Netflix in making data-driven decisions regarding content acquisition, production, and recommendation. Furthermore, it can enhance the personalization of user experience, thereby increasing viewer engagement and satisfaction.

The project underscores the power of data analysis in transforming raw data into actionable insights, demonstrating its potential in the context of the rapidly evolving digital streaming industry.

Contents

Chapter 1

Introduction

The digital revolution has transformed the entertainment industry, with streaming platforms like Netflix leading the charge. As one of the world’s largest on-demand streaming services, Netflix has amassed a wealth of data from its over 200 million subscribers. This report aims to delve into the depths of this data, leveraging the power of Python libraries to extract meaningful insights.

Python, a versatile and powerful programming language, offers a plethora of libraries designed for data analysis. Libraries such as Pandas for data manipulation, NumPy for numerical computations, Matplotlib for data visualization, will be instrumental in our analysis.

The primary objective of this analysis is to understand the viewing habits and preferences of Netflix subscribers. This includes identifying popular genres, understanding peak viewing times, and determining factors that influence a show’s success. Additionally, we aim to predict future trends and provide recommendations for content creation and acquisition.

The data for this analysis is sourced from Netflix’s extensive database, which includes information on user demographics, viewing history, and content details. It’s important to note that all data used in this analysis is anonymized to ensure user privacy.

This report is structured as follows: First, we will conduct an exploratory data analysis to understand the structure and characteristics of the data. Next, we will preprocess the data, handling missing values and outliers, and transforming variables as needed. We will then proceed with our in-depth analysis, using statistical methods and machine learning algorithms to answer our research questions. Finally, we will present our findings and discuss their implications.

Through this report, we hope to demonstrate the potential of data analysis in driving strategic decision-making in the entertainment industry. By harnessing the power of Python and its libraries, we can transform raw data into actionable insights, providing a competitive edge in the rapidly evolving streaming landscape. This analysis is not just about understanding the past and present, but also about predicting the future, enabling Netflix to continue delivering content that resonates with its global audience.

Chapter 2

Basic Concepts

This section provides an overview of the fundamental concepts and techniques used in this project. The project utilizes several Python libraries, including NumPy, Pandas, Matplotlib, and Seaborn, to analyse and visualize data from Netflix.

2.1 NumPy

NumPy, or Numerical Python, 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. It is fundamental for scientific computing with Python as it contains among other things a powerful N-dimensional array object, sophisticated (broadcasting) functions, tools for integrating C/C++ and Fortran code, useful linear algebra, Fourier transform, and random number capabilities.

2.2 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. It provides data structures for efficiently storing large datasets and tools for data wrangling and analysis. In this project, it is used for data loading and preprocessing. The library allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.

2.3 Matplotlib

Matplotlib is a plotting library for the Python programming language and it has a numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. Matplotlib is also a popular library for creating static, animated, and interactive visualizations in Python. It can be used in Python scripts, the Python and IPython shell, web application servers, and more. It is used in this project for creating a bar chart for day-level analysis of content added to Netflix.

2.4 Seaborn

Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. Seaborn helps you explore and understand your data. Its plotting functions operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. It is a dataset-oriented and declarative library, which works well with matplotlib.

2.5 Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. In this project, the ‘netflix_titles.csv’ file is loaded using Pandas. The date_added column is converted to datetime format, and new columns for day_added, year_added, and month_added are created. The data is then checked for null values.

2.6 Data Visualization

Data visualization is the graphical representation of information and data. It uses statistical graphics, plots, information graphics, and other tools to communicate information clearly and efficiently. In this project, a bar chart is created to perform analysis of content added to Netflix.

Chapter 3

Problem Statement

The purpose of this report is to conduct a comprehensive and detailed analysis of the Netflix dataset. Netflix, as one of the world’s leading entertainment services with over 200 million paid memberships in more than 190 countries, provides a rich dataset for analysis. The dataset includes a wealth of information about movies, TV shows, documentaries, and more, spanning several years and numerous countries.

The primary focus of this analysis is to understand the release patterns of movies and TV shows on the platform. Release patterns can provide valuable insights into Netflix’s content strategy, including when new content is typically released and how this may vary by month or day of the week. This analysis could potentially reveal whether Netflix tends to release more content during certain times of the year, such as during holiday seasons or summer months when viewership may be higher.

In addition to release patterns, the analysis will also delve into the genres of the released content. Netflix offers a wide variety of genres, from action and adventure to comedies, dramas, documentaries, and more. By analysing the genres of the released content, we can gain insights into the types of content that Netflix prioritizes. This could potentially reveal trends in viewer preferences, as well as Netflix’s strategy in catering to a diverse global audience.

The analysis will involve several steps, including data cleaning and preprocessing, feature extraction, and data visualization. Data cleaning and preprocessing will ensure the integrity of the analysis by removing any null or NaN values and formatting the date to a specific format conducive to analysis. Feature extraction will involve extracting the release dates of each movie or TV show, as well as the genre information. Finally, data visualization using Matplotlib will provide a visual representation of the release patterns and genre distribution, making the results of the analysis clear and easy to understand.

The ultimate goal of this report is to provide a comprehensive overview of Netflix’s content strategy as revealed through its release patterns and genre distribution. The insights gained from this analysis could potentially inform future content strategies, both for Netflix and for other players in the entertainment industry. It is hoped that this report will serve as a valuable resource for anyone interested in understanding the dynamics of content release strategies in the streaming entertainment industry.

3.1 Project Planning

The project commenced with the crucial step of loading the Netflix dataset onto Google Colab. Google Colab, a cloud-based data analysis tool, was chosen for its robust data processing capabilities and its user-friendly interface that allows for seamless interaction with the data. This platform not only provides powerful computational resources but also supports collaborative work, making it an ideal choice for this project.

Once the dataset was successfully loaded onto Google Colab, the next critical phase was data cleaning. This is a fundamental step in any data analysis project as the quality of data directly impacts the outcomes of the analysis. In this project, data cleaning involved meticulously going through the dataset and removing all NaN or null values. This process ensured that the dataset was free of any gaps or inconsistencies that could potentially skew the results of the analysis.

Following the data cleaning, the date was formatted to a specific format that would be more conducive to analysis. This step was crucial as it standardized the date format across the dataset, thereby facilitating easier manipulation and analysis of the data. The process involved converting the date into a uniform format and then extracting the release dates of each movie or TV show. This extraction provided valuable data points for the subsequent analysis.

The final step in the project planning was the planning of data visualization. Given the large size of the Netflix dataset and the complexity of the data, visualizing the data was identified as a key strategy to better understand the data and draw meaningful insights. For this purpose, Matplotlib, a popular data visualization library in Python, was chosen. The plan was to use Matplotlib to create a variety of graphs that would show release patterns on a day-wise, month-wise, and year-wise basis. These visualizations would not only aid in the analysis of the data but also in the presentation of the findings in a manner that is easy to understand and interpret.

In summary, the project planning phase laid a strong foundation for the project by setting up the necessary tools and processes. It ensured that the dataset was clean, standardized, and ready for analysis. It also planned for effective data visualization to support the analysis and presentation of the findings. This meticulous planning set the stage for a successful data analysis project.

3.2 Project Analysis

The project analysis began with data exploration. The dataset, sourced from Kaggle, initially contained the following features: ‘show_id’, ‘type’, ‘title’, ‘director’, ‘cast’, ‘country’, ‘date_added’, ‘release_year’, ‘rating’, ‘duration’, ‘listed_in’, and ‘description’. These features provided a comprehensive view of each movie or TV show, including its type (movie or TV show), title, director, cast, country of origin, date added to Netflix, release year, rating, duration, genres (listed_in), and a brief description.

To enhance the dataset and facilitate more detailed analysis, new features were added: ‘day_added’, ‘year_added’, and ‘month_added’. These features were derived from the ‘date_added’ feature and provided more granular information about when each movie or TV show was added to Netflix. This enriched dataset paved the way for a more in-depth analysis of release patterns.

The data analysis was conducted using two powerful Python libraries: Matplotlib and Seaborn. Matplotlib, a versatile plotting library, was used to create a variety of graphs to visualize the release patterns. Seaborn, a statistical data visualization library based on Matplotlib, was used to create more attractive and informative statistical graphics.

The analysis focused on understanding when most movies/TV shows are released on Netflix and the genres of the released content. This was achieved by plotting graphs showing the number of contents released on the platform against the month of the year when the content was released. These visualizations provided a clear picture of the release patterns and helped identify any trends or patterns.

In summary, the project analysis phase involved a thorough exploration of the dataset, the addition of new features to enrich the dataset, and the use of data visualization techniques to analyse and understand the release patterns and genres of content on Netflix. The insights gained from this analysis are expected to provide a deeper understanding of Netflix’s content strategy and viewer preferences.

3.3 System Design

The system design for this project was primarily concerned with the design of the data analysis system.

3.3.1 Design Constraints

The main constraint encountered during the design of the system was the handling of missing or null values in the dataset. These values had to be carefully handled to ensure they did not skew the results of the analysis. The solution to this constraint was to remove these NaN or null values from the dataset. This process ensured that the dataset was clean and ready for analysis.

Another design constraint was the need to add new features or attributes to the data for the analysis. This was addressed by adding new features such as ‘day_added’, ‘year_added’, and ‘month_added’ to the dataset. These features provided more granular information about when each movie or TV show was added to Netflix, thereby enriching the dataset and facilitating a more detailed analysis.

3.3.2 System Architecture

The system architecture for this project consisted of several components. The data was first loaded and cleaned in Google Colab. Then, the cleaned data was analysed using various Python libraries, including Pandas for data manipulation, Matplotlib for data visualization, and Seaborn for creating more attractive and informative statistical graphics. The results of the analysis were then interpreted and reported.

This system design allowed for a streamlined and efficient analysis of the Netflix dataset, providing valuable insights into release patterns and content genres. The use of powerful Python libraries like Pandas, Matplotlib, and Seaborn ensured that the system was capable of handling the complex tasks of data cleaning, manipulation, analysis, and visualization. The system was designed with flexibility in mind, allowing for adjustments and additions to be made as needed during the analysis process.

Chapter 4

Implementation

1. **Data Collection**: The first step in our analysis is data collection. We gathered data from Kaggle.

2. **Data Cleaning**: Once we have collected the data, we clean it by removing duplicates, handling missing values, and correcting inconsistent data entries.

3. **Data Exploration**: We then explore the data to understand the patterns and trends. This involves generating descriptive statistics and visualizations.

4.1 Methodology

1. **Objective**: The objective of our analysis is the cornerstone of our project. It’s crucial to define it clearly and precisely. For instance, our objective is to understand user behaviour on Netflix. This involves analysing viewing patterns, preferences, and habits of users. We might also aim to predict future trends based on historical data. This could help Netflix in making strategic decisions about content creation and acquisition.

2. **Users**: The intended users of our analysis are the stakeholders who will benefit from the insights we generate. These could be internal teams within Netflix such as the content team, which could use the analysis to understand what type of content resonates with the audience, or the marketing team, which could use the insights to tailor their campaigns. The analysis could also be used to enhance the viewing experience of Netflix users by providing them with personalized content recommendations.

3. **Data Sources**: The data sources for our analysis would primarily be Kaggle. We might also consider other relevant data sources such as social media platforms for sentiment analysis or third-party databases for additional movie metadata.

4. **Data Analysis Techniques**: Our data analysis techniques would involve a combination of statistical methods. Initially, we would perform data cleaning to handle missing or inconsistent data. This would be followed by exploratory data analysis where we generate descriptive statistics and visualizations to understand the patterns and trends in the data. These models could help us predict user behaviour, recommend movies, or even forecast future trends.

5. **Outcomes**: The expected outcomes of our analysis would depend on our objectives. Our objective is to understand user behaviour, and to predict future trends.

4.2 Result Analysis

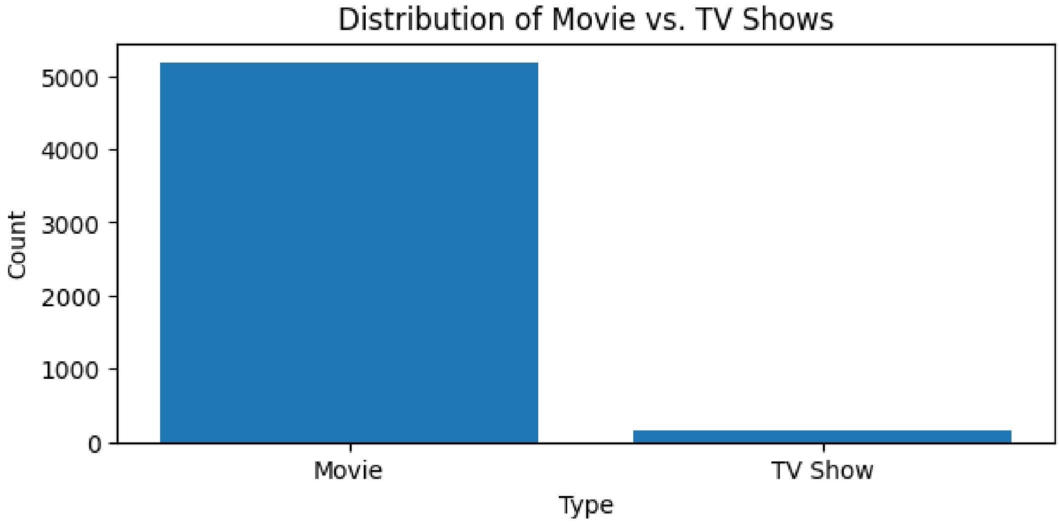

Figure 4.2.1 Type of content v/s Number of releases

Netflix offers a wide variety of content to cater to the diverse tastes of its global audience. The platform primarily features two types of content: Movies and TV Shows. As per the data provided, Netflix has released a substantial number of movies, totalling 5185. This reflects the platform's extensive movie collection, offering a rich selection of genres, languages, and themes. On the other hand, the number of TV Shows is relatively fewer, with 147 releases. Despite being fewer in number, these TV Shows provide audiences with longer narratives and character development, often leading to dedicated fan bases. The disparity in numbers could also be attributed to the fact that producing a TV Show involves a more extended commitment both in terms of production and viewer engagement compared to movies.

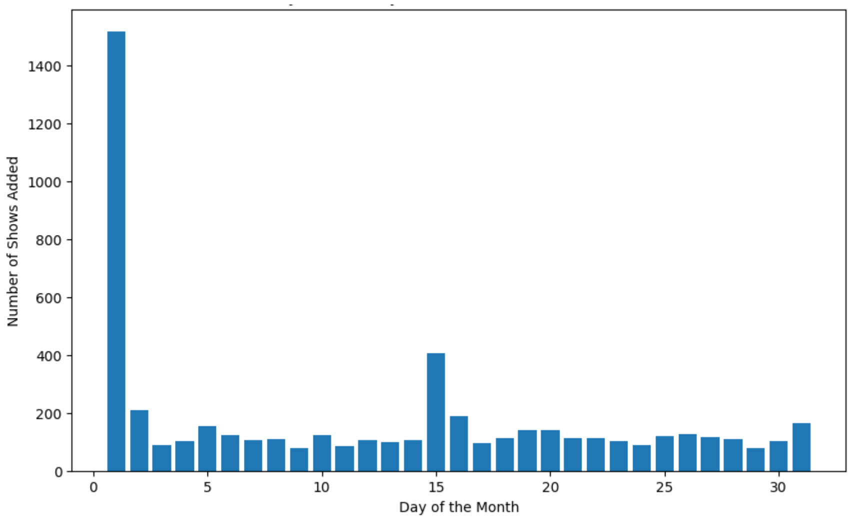

Figure 4.2.2 Relationship between Number of shows added v/s Day of its release

The data provided gives us an interesting insight into the distribution of content releases on Netflix based on the day of the month. It appears that the 1st day of the month sees the highest number of releases, with a total of 1519. This could be due to the start of a new month being a strategic time for introducing new content to viewers.

The 15th of the month also sees a significant number of releases, totaling 408. This mid-month surge could be a strategy to keep the audience engaged throughout the month. The rest of the days have a more evenly distributed number of releases, ranging from 77 to 208, ensuring a steady flow of new content for viewers. Towards the end of the month, there is a slight increase in the number of releases, with the 31st day having 165 releases. This could be a strategy to end the month on a high note and keep viewers excited for the upcoming month. Overall, the distribution of content releases throughout the month seems to be a strategic decision by Netflix to keep its audience engaged and looking forward to new content. It’s fascinating to see how data can provide insights into such strategies.

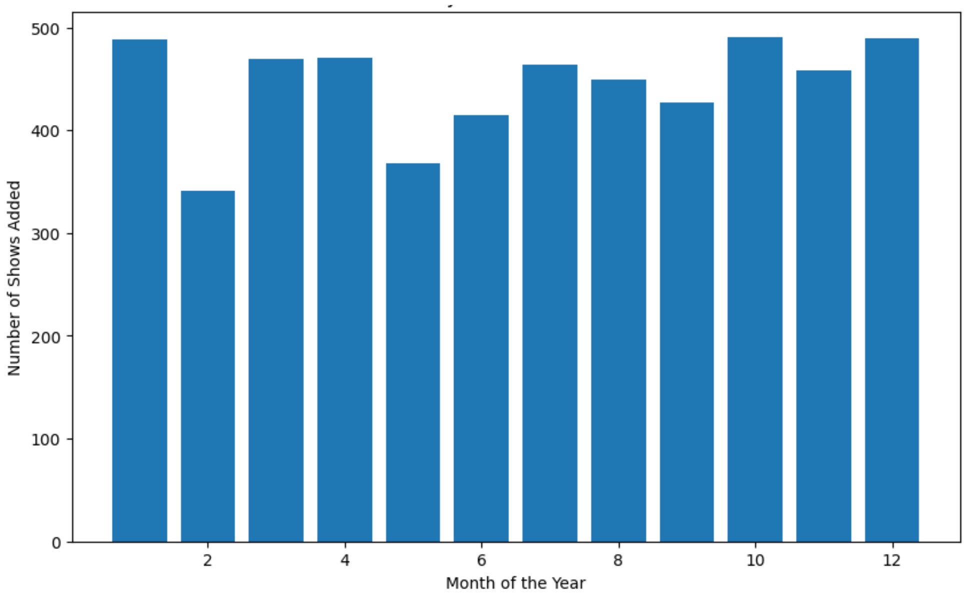

Figure 4.2.3 Relationship between Number of shows added v/s month of its release

The data provided gives us an interesting insight into the distribution of content releases on Netflix based on the month of the year. The months with the highest number of releases are January, October, and December, with 489, 491, and 490 releases respectively. This could be due to strategic planning by Netflix to start and end the year with a large number of new releases, and also to cater to the increased viewership during the holiday season in October and December.

On the other hand, February has the fewest releases with 341, which might be due to it being the shortest month of the year. The rest of the months have a more evenly distributed number of releases, ranging from 368 to 471, ensuring a steady flow of new content for viewers throughout the year.

Overall, the distribution of content releases throughout the year seems to be a strategic decision by Netflix to keep its audience engaged and looking forward to new content. It’s fascinating to see how data can provide insights into such strategies.

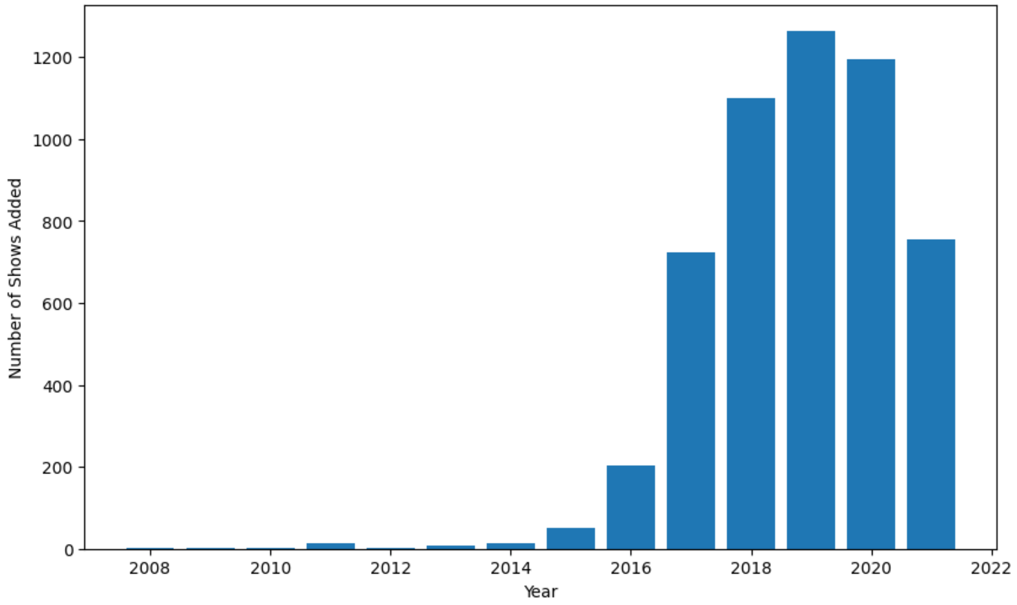

Figure 4.2.4 Relationship between Number of contents added v/s Year of release

The data provided offers a fascinating look at the growth of Netflix’s content library over the years. In 2008, the platform had only one release, but this number has grown exponentially over time. The growth was relatively slow until 2015, with only 50 releases. However, in 2016, the number of releases increased dramatically to 202, indicating a significant expansion in Netflix’s content production and acquisition. The year 2017 saw an even more substantial increase, with 724 releases, more than tripling the previous year’s count. This trend continued in 2018 and 2019, with 1101 and 1265 releases respectively, reflecting Netflix’s aggressive strategy to broaden its content library. In 2020, the number of releases slightly decreased to 1194, possibly due to the global pandemic’s impact on content production. Despite this, Netflix managed to release a significant amount of content. In 2021, the platform released 755 pieces of content, indicating a continued commitment to providing a diverse range of viewing options for its audience. Overall, the data shows a clear upward trend in the number of releases over the years, highlighting Netflix’s growth as a global entertainment platform.

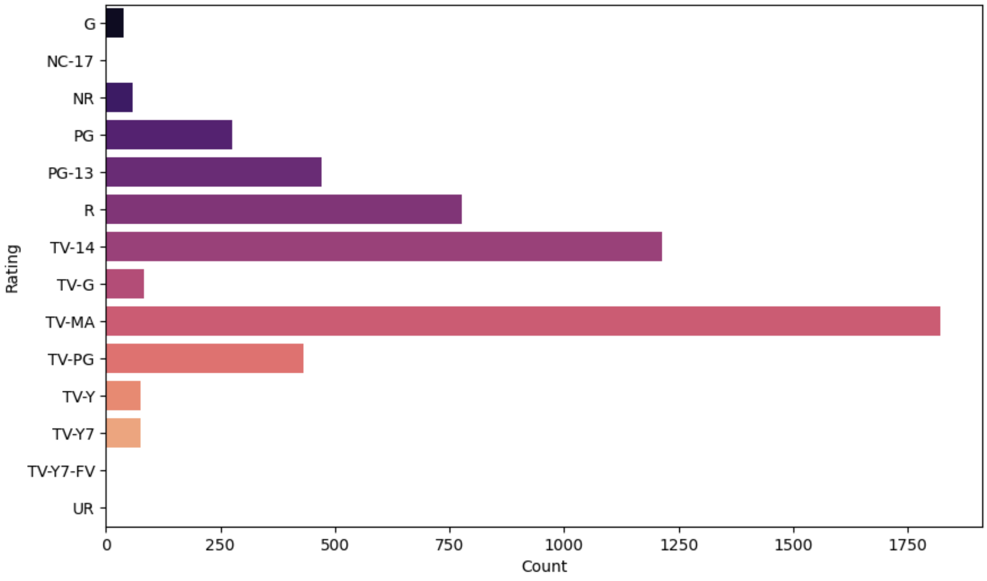

Figure 4.2.5 Genre of the content v/s Number of content

The data provided gives us an interesting insight into the distribution of content releases on Netflix based on the content rating and the month of release. Content with a rating of `TV-MA` has the highest number of releases, with both the year and month of addition totalling 3205. This suggests that Netflix has a significant amount of content aimed at mature audiences. Following `TV-MA`, content with a `TV-14` rating has the second-highest number of releases, totalling 2157 for both the year and month of addition. This indicates a substantial amount of content suitable for viewers aged 14 and above. For content rated `R`, which is restricted to viewers over 17 unless accompanied by an adult, there are 799 releases. Content rated `PG-13` and `TV-PG`, which suggest parental guidance, have 490 and 861 releases respectively. Content with a `G` rating, which is suitable for all ages, has 41 releases. This shows that there is a smaller amount of content specifically aimed at younger audiences. Interestingly, there are also a few releases that are listed with their duration as the rating, such as `66 min`, `74 min`, and `84 min`, each with one release. Overall, the data shows a wide range of content ratings in Netflix's releases, catering to a diverse audience with different viewing preferences.

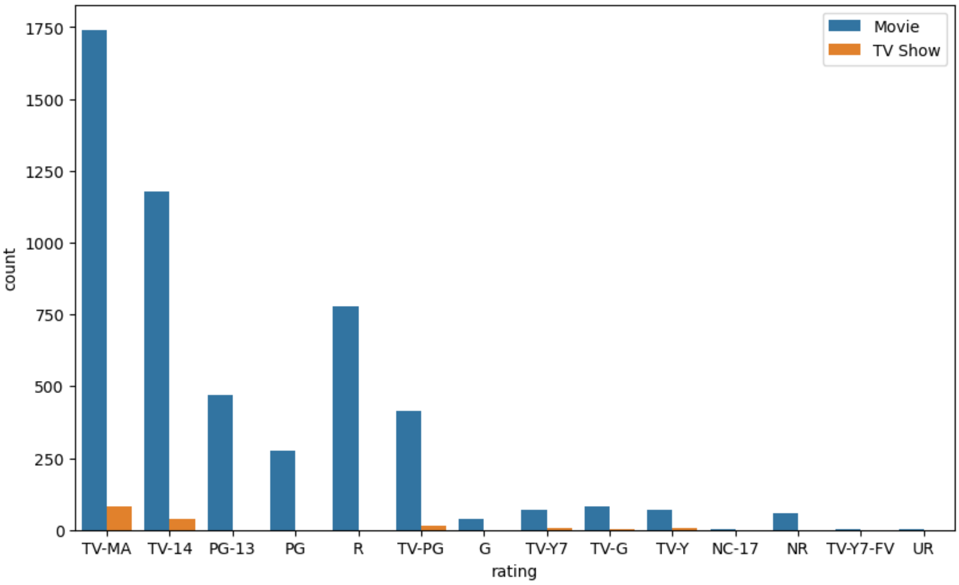

Figure 4.2.6 Relationship between type of content v/s Total count of content

The data provided gives us an interesting insight into the distribution of content across different ratings and their release timings. It appears that the content rated as 'TV-MA' has the highest number of releases, with 1822 releases in a particular year and month. This is followed by 'TV-14' with 1214 releases, and 'R' with 778 releases in the same period.
On the other hand, content with ratings 'NC-17', 'TV-Y7-FV', and 'UR' have the least number of releases, with just 2, 3, and 3 releases respectively. This could be indicative of the type of audience that the platform is targeting, with a clear preference for more mature content ('TV-MA', 'TV-14', 'R').
It's important to note that this data is specific to the year and month of release. A different time frame might present a different distribution. This highlights the dynamic nature of content distribution based on ratings over time.

Chapter 5

Conclusion and Future Scope

5.1 Conclusion

The comprehensive study conducted on Netflix’s data analytics primarily revolved around the analysis of user watch specifics, the quantity and genre of content, and their respective ratings. The research was meticulously carried out, with a keen focus on the aforementioned parameters, to gain a deeper understanding of the content strategy employed by Netflix.

The key findings of the study were quite revealing. It was observed that between the years 2018 and 2020, a majority of the shows and TV series were released on the platform. This indicates a strategic move by Netflix to consistently add fresh content to their platform throughout the year, thereby ensuring that the viewers always have something new to look forward to. Interestingly, it was also noted that the first three weeks of every month saw the majority of these releases. This could potentially be a strategic decision to engage viewers early in the month, possibly aligning with common billing cycles.

A significant observation from the study was that from 2008 to 2020, there were considerably more movies released than TV shows. This could be indicative of the viewer’s preference for movies over TV shows during this period, or perhaps a strategic decision by Netflix based on other factors such as production costs or licensing agreements.

In terms of content ratings, the study found that TV-MA and TV-14 rated content were released more frequently than others. This could suggest that such content is more popular among Netflix’s target audience, or it could reflect the nature of the content that is being produced and made available for streaming.

The study made use of several analytical tools and programming languages, including Matplotlib, NumPy, Pandas, Python, and Seaborn. These tools were instrumental in processing and analysing the data, and in generating meaningful insights.

However, like any research, this study was not without its challenges. Handling corrupt data, such as NAN values, null values, or duplicated data, posed a significant challenge. These issues had to be addressed to ensure the accuracy and reliability of the study’s findings.

5.2 Future scope

Looking ahead, there are several exciting avenues for future research in this area. One potential area of focus could be a more granular analysis of user behaviour. By studying viewing patterns and preferences in more detail, it would be possible to further enhance the content recommendation algorithms, thereby providing a more personalized and enjoyable viewing experience for users.

Another promising direction for future research could be to include more recent data in the analysis. This would provide a more current understanding of trends on the platform, and could potentially reveal new insights about the evolving preferences and behaviours of Netflix’s user base.

The challenges encountered in this study, particularly in handling corrupt data, present an opportunity for improvement in future research. Developing more sophisticated data cleaning techniques could help overcome these challenges, and would contribute to the robustness and reliability of future studies.

Finally, an interesting area for future research could be to explore the impact of release patterns and ratings on user engagement and satisfaction. Understanding how these factors influence viewer behaviour could provide valuable insights for content strategy and decision-making. This, in turn, could contribute to an improved user experience and potentially higher viewer retention for Netflix. Ultimately, the goal of such research would be to contribute to the ongoing efforts to understand and enhance the viewer experience on Netflix.