IE6600: Data Analysis and Visualization Project Report

Project 3: Netflix

Madison Rodriguez, Adam Luk, Ameya Patil, Leela Maunika Talluri

Due Date: 6/23/2024

1. Introduction

1.1. Objective

The primary objective of this project is to conduct a comprehensive analysis of Netflix data to uncover trends, patterns, and insights that can inform strategic decisions. As a leading global streaming platform, Netflix possesses extensive data covering various aspects such as its content library and financial performance. By analyzing this data, we aim to answer key questions regarding the trends in the release and addition of movies and shows over the years, variations in the duration and number of seasons of shows, distribution patterns of content ratings and countries of origin, and the actors most frequently cast in Netflix productions each year. The insights derived from this analysis will help in understanding Netflix's growth trajectory, content strategy, and user engagement patterns.

1.2. Dataset Description

For this project, we have used multiple datasets related to Netflix, which include:

- **NFLX Stock Data :** Historical stock price and trading volume data for Netflix.
- **Netflix Titles:** Details of movies and shows available on Netflix, including release dates, duration, rating, country, and cast information.
- **Movies on Streaming Platforms:** Data on movies released on Netflix and other platforms with their release years and age categories.

2. Data Acquisition and Preparation

2.1. Initial Inspection

After reading in the datasets, we performed an initial inspection to understand their structure and contents. For each dataset, we utilized the **describe()** function to generate summary statistics for numerical columns, and **describe(include='objects')** to include object-type columns, giving us an overview of the categorical data. Additionally, we used the **info()** function to assess the data types and check for missing values.

The initial inspection of the NFLX Stock Data(df_stocks), which includes columns for Open, High, Low, Close, Adjusted Close, and Volume, reveals significant insights about Netflix's market activity. Each column has 5116 entries, indicating a complete dataset without missing values. The average stock price is approximately $117.82, with an average trading volume of about 16.42 million shares. The data shows high volatility, with standard deviations around $167.87 for stock prices and 19.10 million shares for trading volume. The stock prices range from a minimum close price of $0.37 to a maximum of $691.69, while trading volumes vary from 285,600 to 323.41 million shares. These statistics highlight Netflix's

substantial stock price growth and significant variability in trading activity over time, reflecting its dynamic and evolving market presence.

During the initial inspection of the Netflix Titles(df_titles) dataset, which encompasses various columns such as show_id, type, title, director, cast, country, date_added, rating, duration, listed_in, and description, several key observations emerged. The dataset contains a total of 8807 entries, of which most columns had complete data. However, columns like director, cast, and country show some missing values, which will be evaluated during the data cleaning process. The Netflix Titles dataset had all show_id and title being unique across the dataset, which means no duplicated entries. The dataset predominantly consists of 6131 "Movie" entries, outnumbering "TV Show" entries .The dataset reported 748 unique country entries, with the United States appearing most frequently (2818 times). However, since there are only 195 recognized countries, we needed to investigate why this discrepancy occurred. It was found that some entries listed multiple countries in their names. To address this, we undertook a process to split these entries and extract individual countries correctly. The dataset includes 17 distinct ratings, with "TV-MA" being the most prevalent (3207 entries), indicating a substantial presence of mature content. Additionally, it encompasses 220 unique durations, with "1 Season" being the most common (1793 times), and offers a wide array of 514 unique genre categories, encompassing popular genres such as "Dramas, International Movies". These insights highlight Netflix's commitment to offering diverse and engaging content to its global audience.

The Movies on Streaming Platform’s (df_platforms) dataset offers comprehensive insights into streaming services and content attributes. In the categorical data, which includes 9515 entries, each title is unique, and the only missing values occur in the Age column and Rotten Tomatoes. Age ratings show diversity, with categories like "18+" being most common. The numerical data further reveals key statistics about platform availability and content years. The dataset spans productions from 1914 to 2021, with an average production year around 2007. It indicates that Netflix offers content about 39% of the time, Hulu 11%, Prime Video 43%, and Disney+ nearly 10%. These figures provide valuable insights into the distribution of content across major streaming platforms. These initial observations are crucial for understanding the content landscape and platform preferences, informing strategic decisions in content acquisition and audience engagement strategies within the streaming services industry.

2.2. Data Cleaning

During data cleaning, we standardized all date formats across the four datasets and addressed missing values as follows. Both df_stocks and df_users had no null value. In df_platforms, we encountered null values in the ‘Age’ and ‘Rotten Tomatoes’ columns, which we filled using the mode due to their categorical nature. For df_titles, we filled null values in 'rating', 'duration', and 'country' with mode values. For the ‘date_added’ missing values, we computed the year_diff, based on the difference between date_added and release_year, and determined the average to be 4.7 based on non-null values within the dataset. Subsequently, we applied this average to fill in missing values in the date_added column, ensuring consistency and completeness across the dataset. This process ensured comprehensive data standardization and completeness across all datasets.

3. Exploratory Data Analysis (EDA)

3.1. Overview

In this analysis, we look into various datasets to uncover insights about Netflix and other major streaming platforms. We begin by examining the historical stock prices of Netflix from 2002 to 2022, highlighting significant trends and events that have influenced its market performance. Next, we explore the correlation between Netflix's average annual closing price and the number of movies added each year, providing a perspective on how content volume impacts stock value. Additionally, we analyze the monthly patterns of movie additions to understand Netflix's content release strategy over time. Our geographical analysis through choropleth maps showcases the origins of movies and TV shows on Netflix, revealing the countries with the highest production volumes. We also investigate the yearly distribution of content releases and their ratings to discern trends in Netflix's programming choices. Furthermore, we compare the total count of movies released on different streaming platforms and examine the distribution of age ratings across these platforms. These visualizations collectively offer a comprehensive view of the streaming landscape and Netflix's evolution within it.

3.2. Visualizations

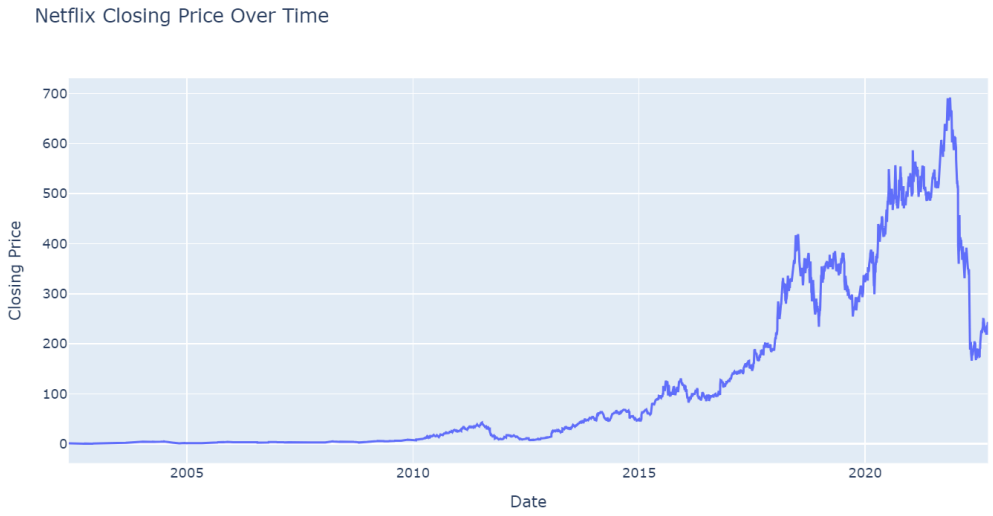

Figure 1 : Netflix Closing Price 2002-2022

From the stock dataset, we've analyzed NFLX closing price from the years 2002 to 2022 and it matches that of the public stock market price. From here we can see interesting trends specifically in 2020 and 2022 where we see a massive dip in price. The dip matches that with the news where in 2020, the COVID pandemic had sprung leading to massive stock sell offs and in 2022 where the stock tumbled nearly 50% due to quarters of subscriber declines.

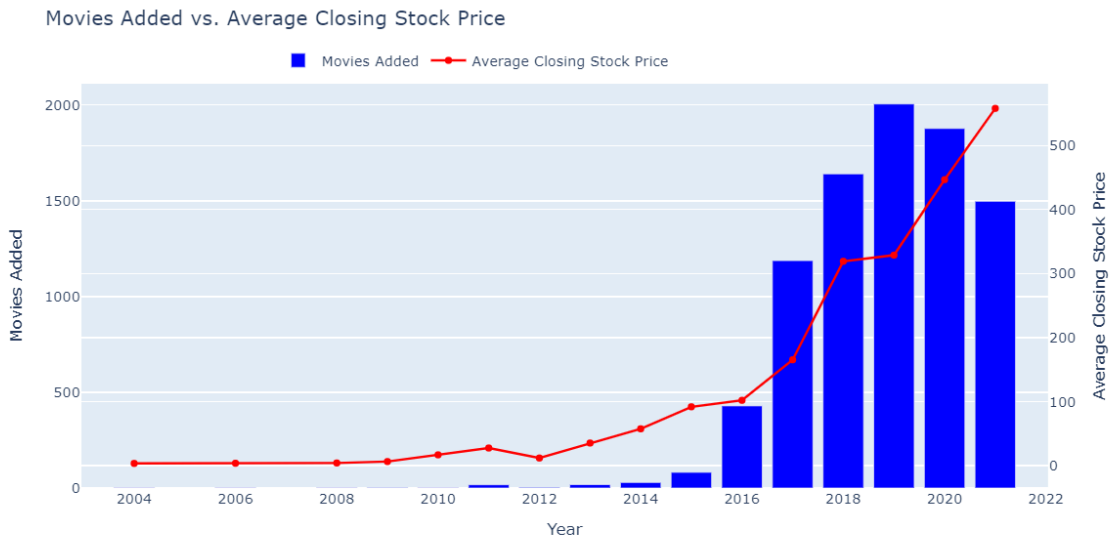

Figure 2: Netflix Average Annual Closing Price vs # of Movies Added

In addition, we've collaborated the Netflix titles dataset to correlate the number of movies added with the increase in average yearly stock price and it seems that there is a positive trend between the two. Unfortunately, there was not enough data to show the connection between the 2022 massive dip.

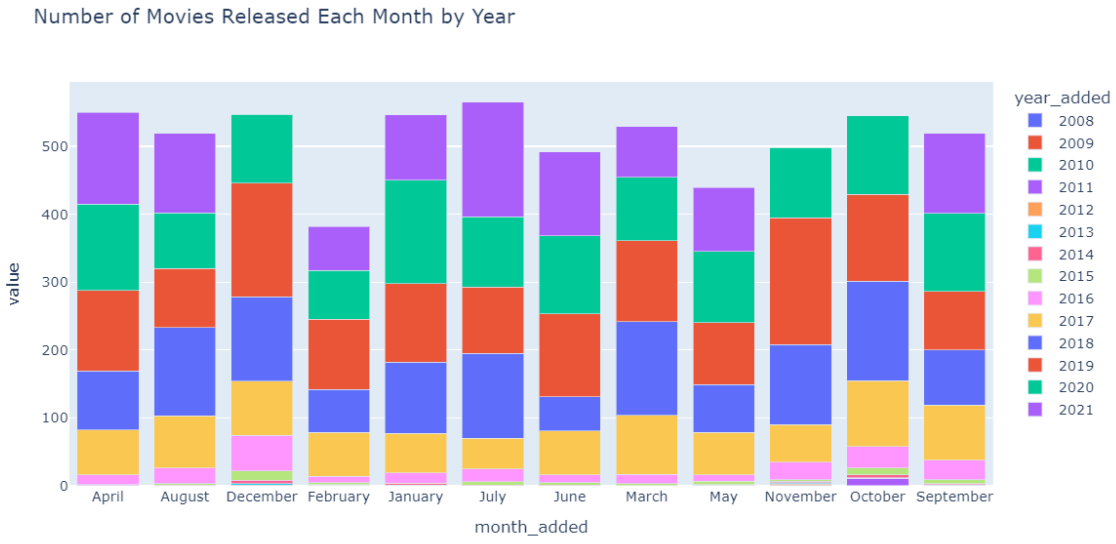

Figure 3: Number of Movies Added for Each Year/Month

Additionally, we wanted to see how Netflix approached adding movies in each discrete month. We analyzed that February was the lowest number of movies added, while July seemed to have the most total of movies added. Interestingly, you can see the periods where more movies were

added in a specific month compared to the rest for certain years as if there is A/B testing and experimentation to see when it is the best period to add movies (i.e December 2009).

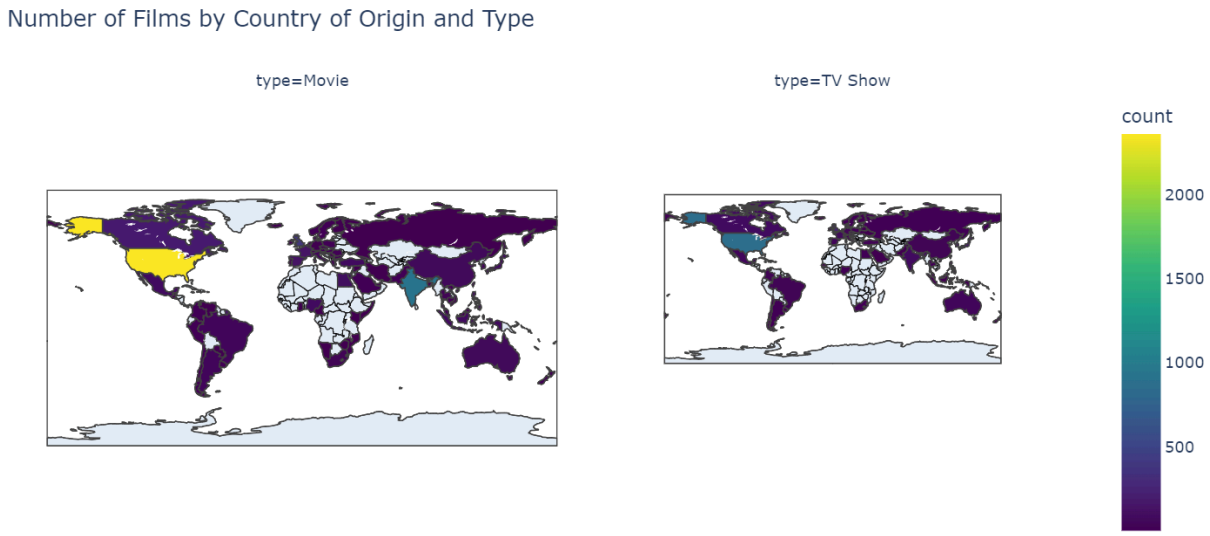

Figure 4: Movie and TV show Origin Country of Release Count

Lastly from the Netflix title dataset, we wanted to see the countries of origins where there is a significant release of films. So we created two choropleth maps to analyze where movies and tv shows had the amost abundant releases. For movies the United States and India have the most production of movies, while the United States has the most production of television shows.

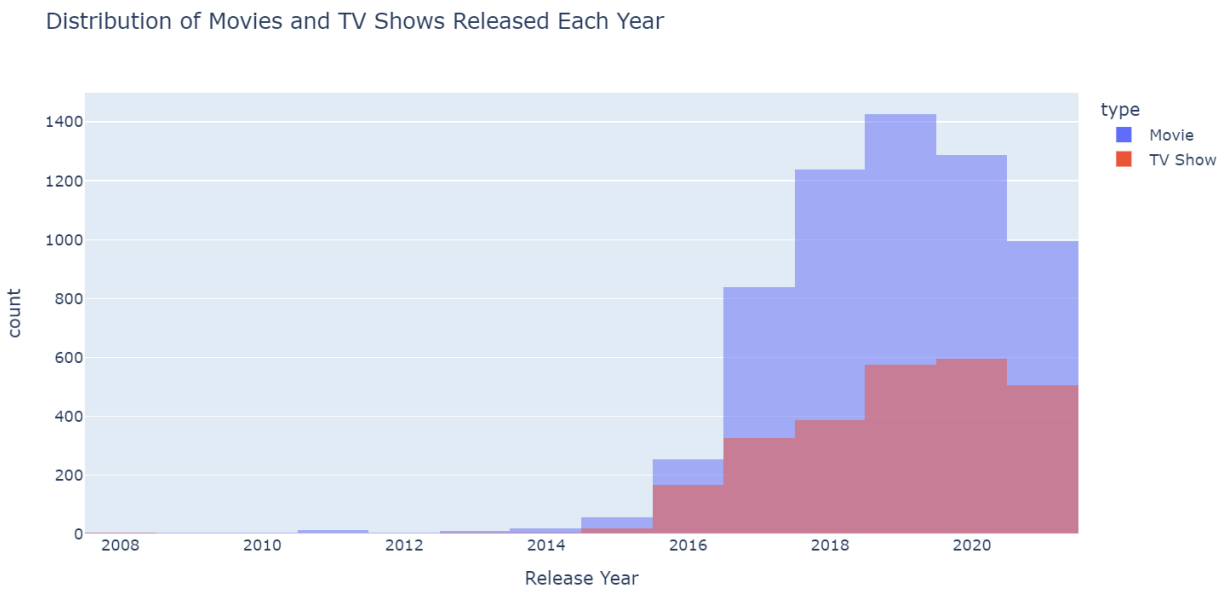

Figure 5: Distribution of Movies and TV Shows Released on Netflix Each Year

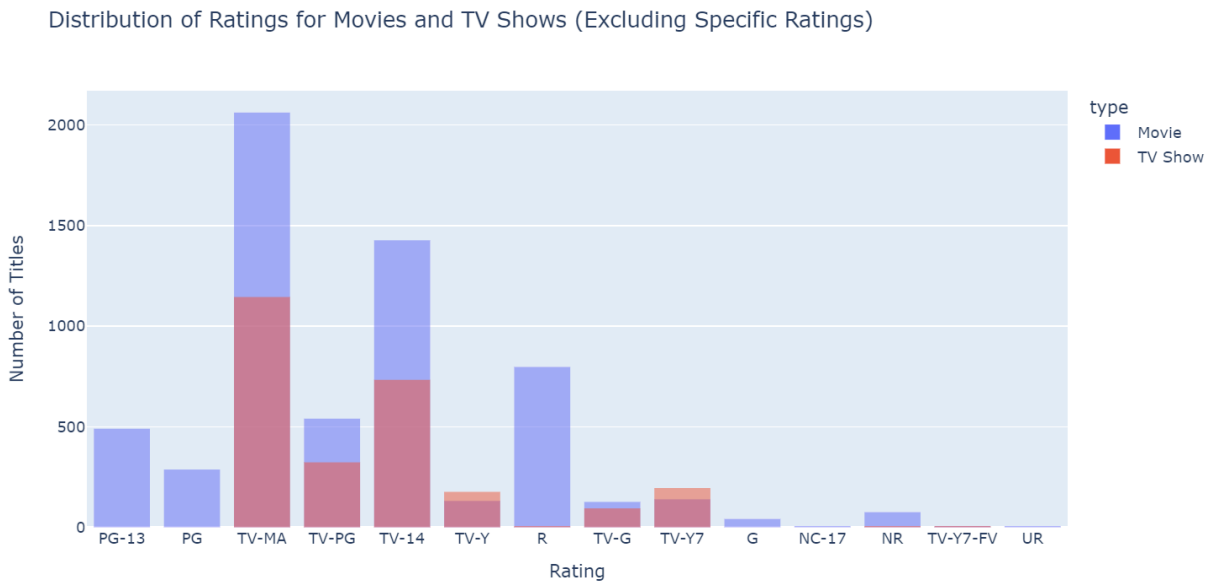

Figure 6: Distribution of Movies and TV Shows Ratings on Netflix

Next we also found that 2019 has the most significant amount of releases of film and there is a decrease from there on, it is unsure if that 2020 pandemic contributed to this. Additionally, we analyzed the distribution of ratings and found TV-MA was the most even and abundant distribution for both shows and movies. While movies claimed the most pg-13 and rated R ratings, most likely because it is more difficult to air TV shows with these ratings.

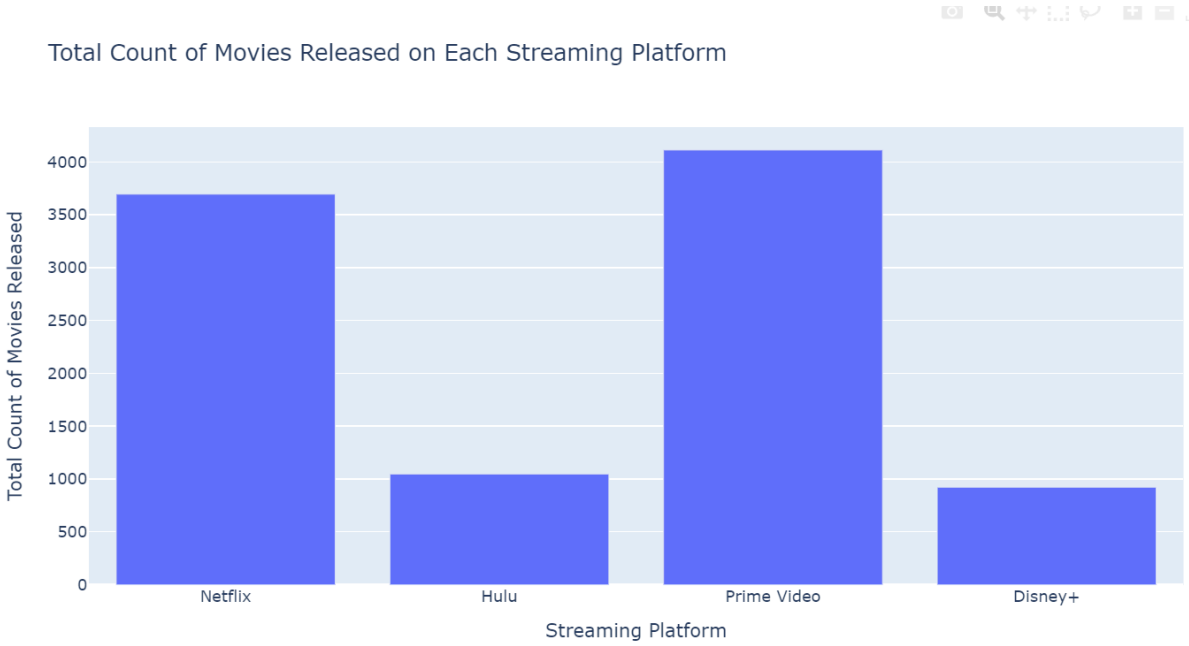

Figure 7: Total Count of Movies Released on Streaming Platforms

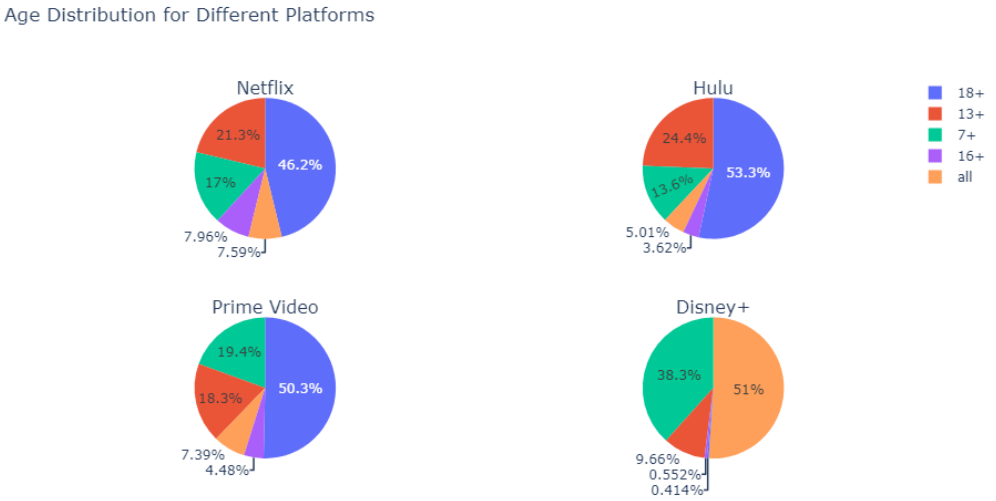

Figure 8: Distribution of Age Ratings for Different Platforms

Next, we analyzed a dataset of total movies released on each streaming platform. Surprisingly Prime Video had the highest considering it was founded way after Netflix and Hulu. Disney+ had the lowest amount, not surprising because it was founded in 2019. Lastly we were curious on the age distribution of the movies for each platform, every platform but Disney+ had a majority of films for 18+ while Disney+ had nearly 51% for all and 38.3# for 7+, not surprising because it is catered towards younger viewers.

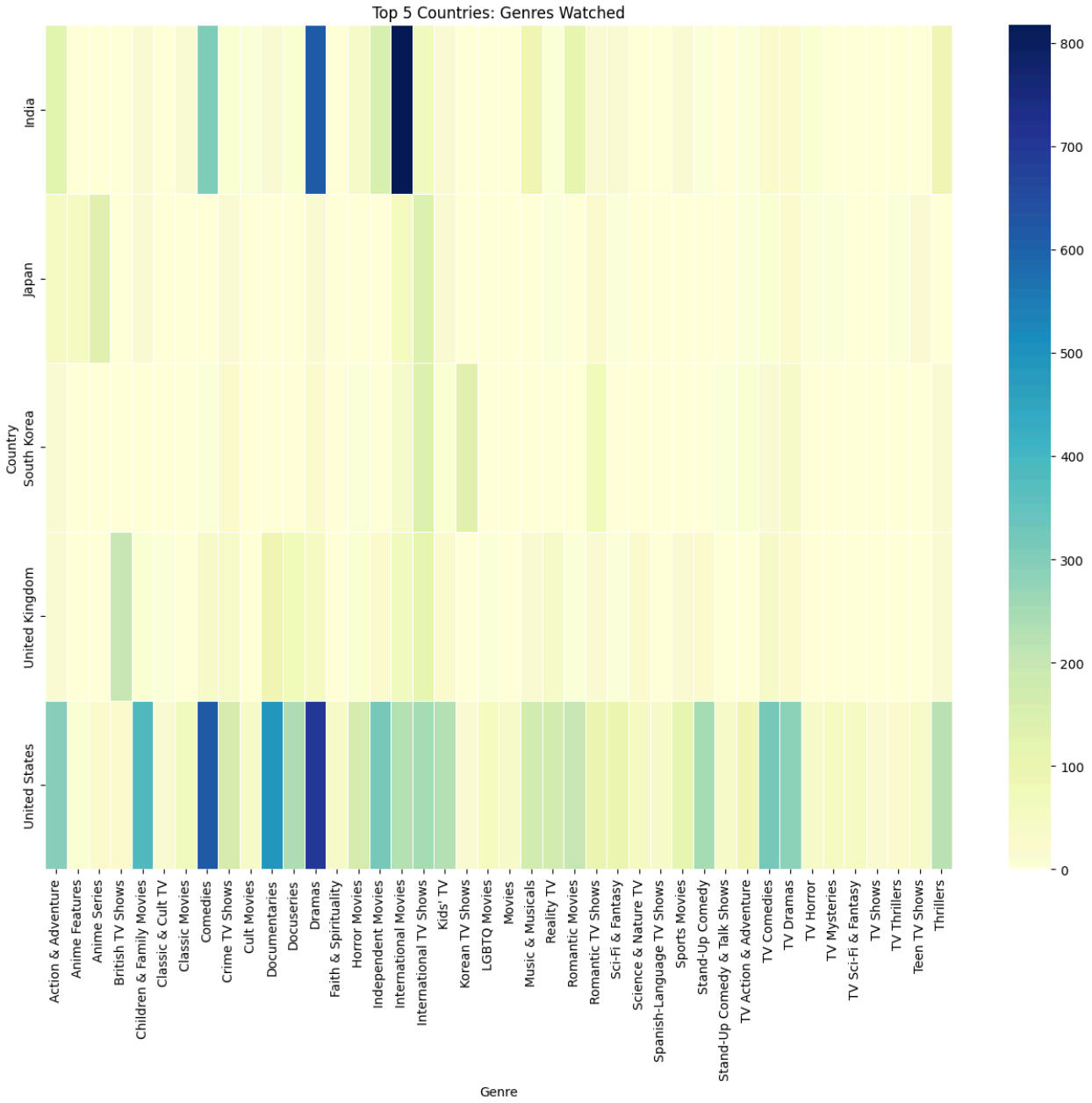

Figure 9: Top 5 Countries Vs Genre

Dramas are the most popular genres in both India and the United States when it comes to international films. In India, audiences prefer International films. Similarly, in the United States, dramas are consistently among the most popular genres. Dramas captivate audiences by providing profound insights into the human condition and prompting meaningful reflections, whether they are about historical events, current issues, or personal struggles.

4. Conclusion

4.1. Summary of Findings

This analysis provides valuable insights into the strategies and trends of Netflix and other major streaming platforms. We observed that Netflix's stock prices are influenced by significant global events, with notable dips during the COVID-19 pandemic and in response to declining subscriber numbers. There is a positive correlation between the number of movies added to Netflix and its average annual stock price, suggesting that increasing content volume can enhance market value. Monthly patterns in movie additions indicate strategic experimentation with release timings. Geographically, the United States and India are major contributors to Netflix's movie catalog, while the United States dominates TV show production. The peak year for Netflix releases was 2019, with a decline following the onset of the pandemic. Rating distributions reveal a preference for TV-MA content, while movies dominate the PG-13 and R categories. Prime Video leads in total movie releases, highlighting its aggressive content expansion despite its later market entry. Disney+ focuses on family-friendly content, catering primarily to younger audiences.

4.2. Implications

These findings underscore the dynamic nature of the streaming industry and Netflix's adaptive strategies. Understanding these trends can help stakeholders make informed decisions regarding content production, marketing, and platform development. For consumers, it offers insights into content availability and platform focus.

5.3. Future Work

Future research could delve deeper into the causal relationships between content additions and stock performance, explore the impact of exclusive releases and original productions, and analyze viewer engagement metrics across different platforms and regions. Additionally, examining the effects of emerging competitors and market saturation on streaming platforms would provide a comprehensive understanding of the industry's trajectory.

6. Reference

Bhatia, R. (2021). *Movies on Netflix, Prime Video, Hulu and Disney+*. Kaggle.com. https://www.kaggle.com/datasets/ruchi798/movies-on-netflix-prime-video-hulu-and-disney

Bansal, S. (2021). *Netflix Movies and TV Shows*. Kaggle.com. https://www.kaggle.com/datasets/shivamb/netflix-shows

Aj. (2024). *Netflix Userbase Dataset*. Kaggle.com. https://www.kaggle.com/datasets/arnavsmayan/netflix-userbase-dataset/data

Chauhan, A. (2022). *Netflix | Stock Market Analysis | Founding Years*. Kaggle.com. https://www.kaggle.com/datasets/whenamancodes/netflix-stock-market-analysis-founding-years