Data Insights on Netflix

CS418: Introduction to Data Science

Professor Sourav Medya

April 30th, 2024

Team Black Hawks

Venkata Rohith Kumar, Rithwik Vamshi,

Asritha, Sanjna, Bhargavram

Table of Content

1. Introduction

Netflix is one of the biggest online streaming platforms around the world offering a diverse range of movies and TV shows. With around 8000 plus offerings, the platform caters to all age groups. Over the last few years, Netflix has been offering original content based on real life events and is involved in investing in different content types from around the world to expand its digital library. The project's main objective is to learn and summarize the historical trends seen to understand the customer's preferences to better advise on the strategic decision-making. This will help Netflix offer tailored content packages for audiences from different countries.

Data Source

Main dataset: https://www.kaggle.com/datasets/shivamb/netflix-shows?resource=download

External dataset: https://www.kaggle.com/datasets/ashishgup/netflix-rotten-tomatoes-metacritic-imdb/

Problem Statement

Through our analysis and visualizations, we aim to answer the following research questions:

1. What are the popular content types and how have they evolved over the years?
2. What are some of the current strategies that are working best for countries around the world?
3. How can we predict which new titles will be a hit?
4. Who should be Netflix partnering with for newer and hit content?

Overview of the Dataset

Our dataset contains 12 fields providing information on different titles, actors, cast, countries, directors, ratings, durations, descriptions and so on. We prepared our data by merging two datasets obtained from Kaggle's Netflix titles and additional ratings from reputable sources like Rotten Tomatoes, and IMDb. This merging process was essential to combine valuable insights from both datasets, resulting in a more comprehensive dataset for analysis. The accessibility of the data was convenient as it was pre-collected and readily available, eliminating the need for extensive data collection efforts. Integration was facilitated by matching titles from both datasets, ensuring that all relevant information was included in the combined dataset. The table below provides a detailed overview on the different columns.

Fields/Columns	Description
show_id	Show Identifier
type	Type of content (Movies/TV Shows)
title	Title of the content shown
director	Director of the content shown
cast	The cast members present in the content
country	The country producing the content
date_added	The date when a content was added to Netflix
release_year	The release year of a content
rating	The rating given to a content (PG-13, TV-14 etc.)
duration	Total duration of the content (minutes/hours/seasons)
listed_in	The genre or category this content is listed in
description	Description of what type of content (Thriller/Hidden gem etc)

Analytics tools

Python: Python libraries such as Pandas, NumPy, Matplotlib, Plotly etc. were employed for preprocessing and cleaning of the data. Upon cleaning, Python was used to analyze and visualize the data and conduct statistical analysis.

2. Data Preprocessing and Cleaning

Before diving into the analysis, we observed a few inconsistencies/errors with the data set. To get an accurate analysis, we performed preprocessing and cleaning to ensure the data was free of errors.

Replacing NULL values:

Our dataset had the following null values:

- **Director:** The director field had around 2634 null values comprising around 30% of the dataset. These values were replaced by **'Unknown'.**
- **Cast:** The cast field had around 825 null values comprising around 10% of the dataset. These values were replaced by '**Unknown'.**
- **Country:** The country field had around 831 null values comprising approximately 10% of the dataset. These values were replaced by **'Unknown'.**

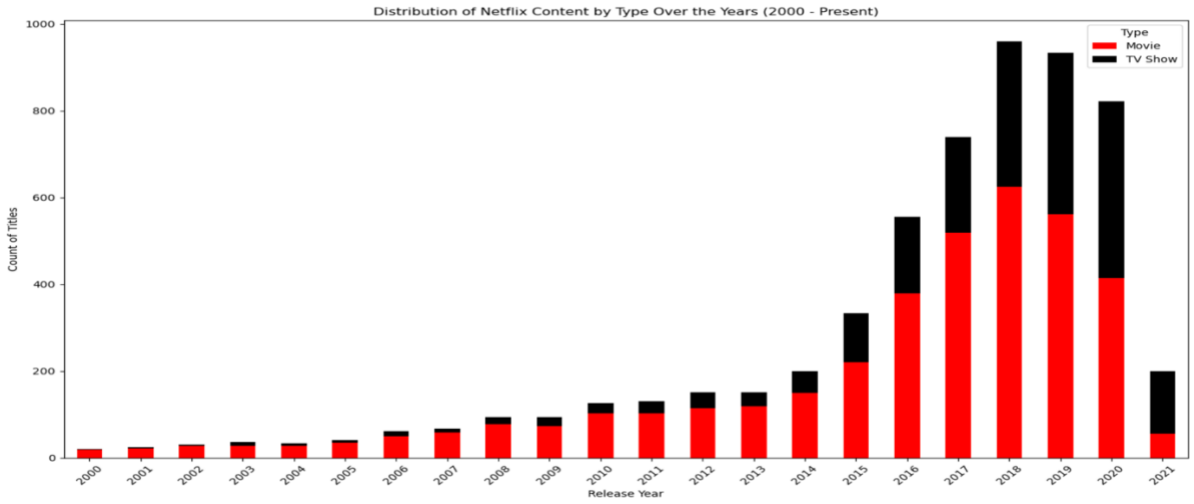
Extracting Data:

Using the date value function, we extracted the duration from the dataset. In addition to this, we extracted the main countries and genres from the list of countries and genres making it easier for the analysis and the visualizations. Further, the date format was converted to a standard date format. The missing values were dropped instead of replacing them. Though the missing values accounted for a smaller portion of the dataset, replacing it with **'Unknown'** caused a few errors with the time value function. Hence, the missing values were dropped.

Lastly, our dataset consisted of only categorical variables. Hence, to perform statistical analysis and visualizations, we incorporated an external dataset that contained quantitative variables to assist with our findings. The dataset consisted of additional columns such as score (IMDb score, Hidden gem score), awards received, awards nominated for, viewer engagement etc.

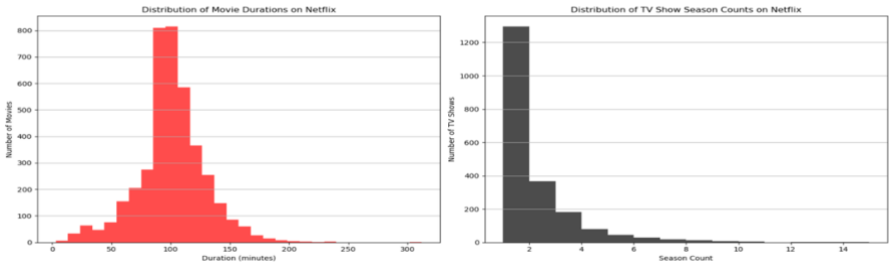
3. Visualizations

3.1 Visualization 1 - Contributed by Rohith.



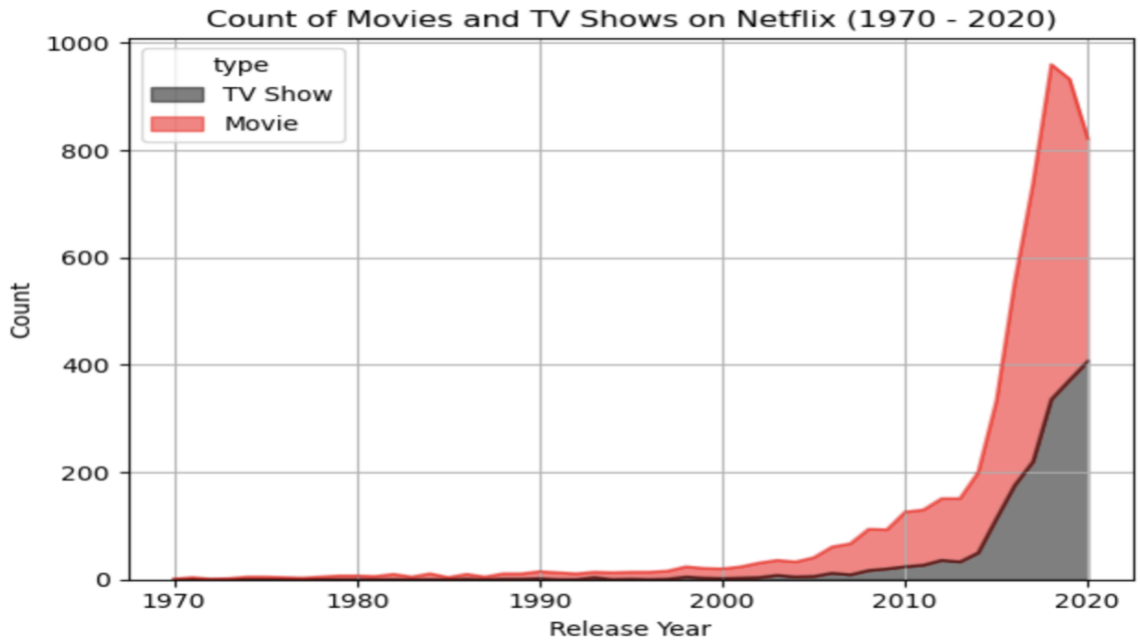
This visualization highlighted the growth in the number of Movies and TV Shows on Netflix over time. The increasing trend in both categories, especially the notable surge in movies around the 2010s, suggests Netflix's expanding library and possibly a strategic shift towards offering a more diverse set of content. This shift might be in response to evolving viewer preferences and the competitive streaming landscape.

3.2 Visualization 3 - Contributed by Rohith.



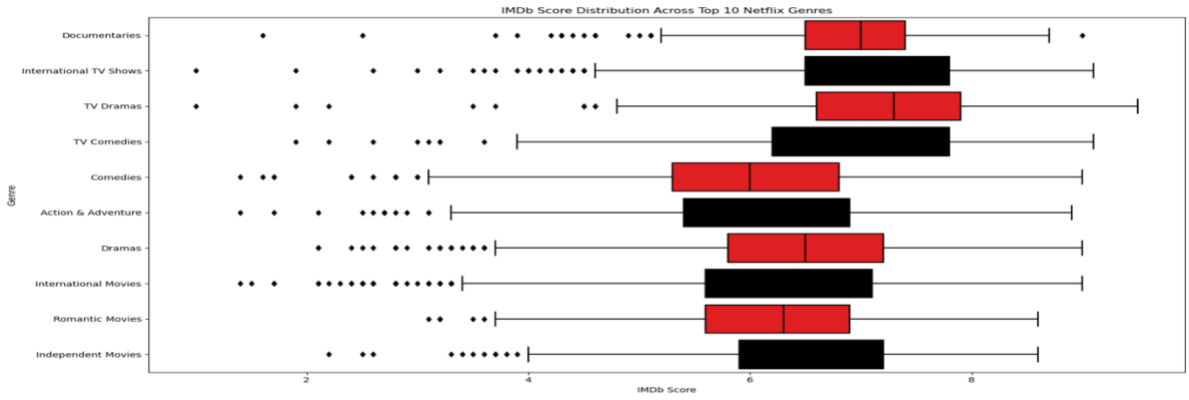
Placing these histograms side by side offers a comparative view of content longevity and investment on Netflix. While movies provide a one-off entertainment experience, TV shows require a more extended commitment from viewers and creators alike. The significant number of movies and the preponderance of TV shows with fewer seasons underscore Netflix's dual strategy: maintaining a broad movie library to cater to diverse tastes and experimenting with TV series to cultivate dedicated viewership over multiple seasons. Netflix’s content portfolio, as suggested by these distributions, might be strategically designed to balance between offering a wide variety of one-time cinematic experiences and developing serialized content that has the potential to build a loyal viewer base over time. Overall, these visualizations shed light on the structure of Netflix's content offerings, suggesting strategic choices in content creation and acquisition to meet diverse viewer preferences and enhance the platform's appeal.

3.3 Visualization 8 - Contributed by Rohith.



The area graph displays the count of movies and TV shows on Netflix from 1970 to 2020, a hypothesis that can be assumed from this graph might been that Netflix has significantly increased its content library over the years, particularly in the last decade. This graph can be used to understand how the company's content strategy has evolved in response to changing viewer preferences and the competitive last decade streaming media. The sharp increase in movies, especially, may reflect an emphasis on providing a diverse range of films to attract a more audience

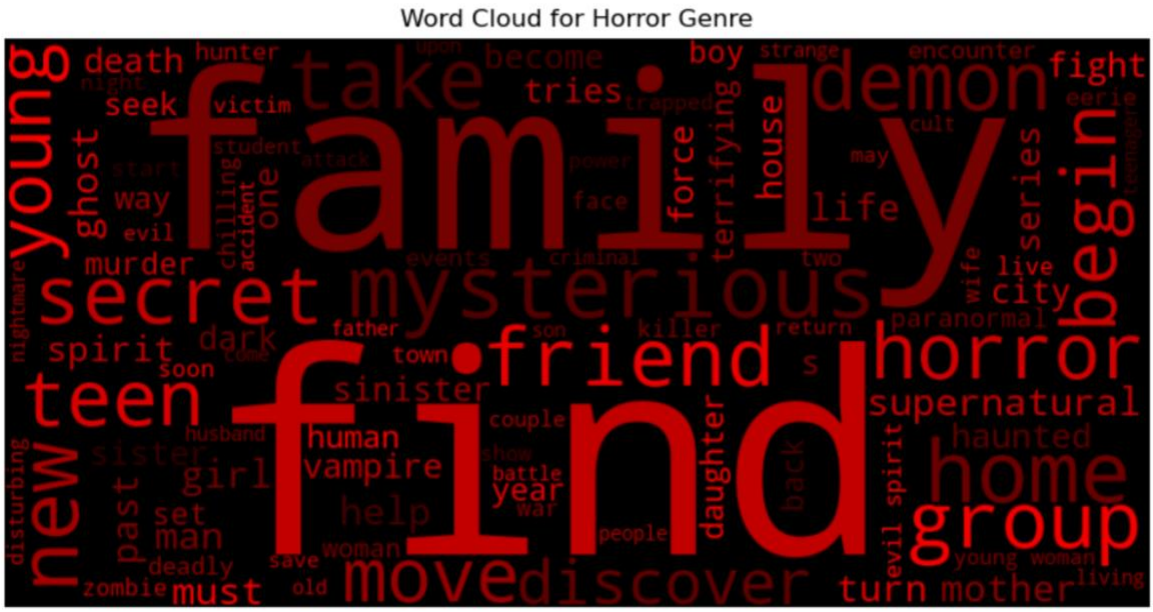
3.4 Visualization 2 - Contributed by Rithwik.



The box plot shows IMDb score distributions across various Netflix genres, the hypothesis could be exploring the quality of content as seen by viewers across different genres. The intent of this analysis is to identify which genres consistently deliver content that is highly rated, as well as to understand the range

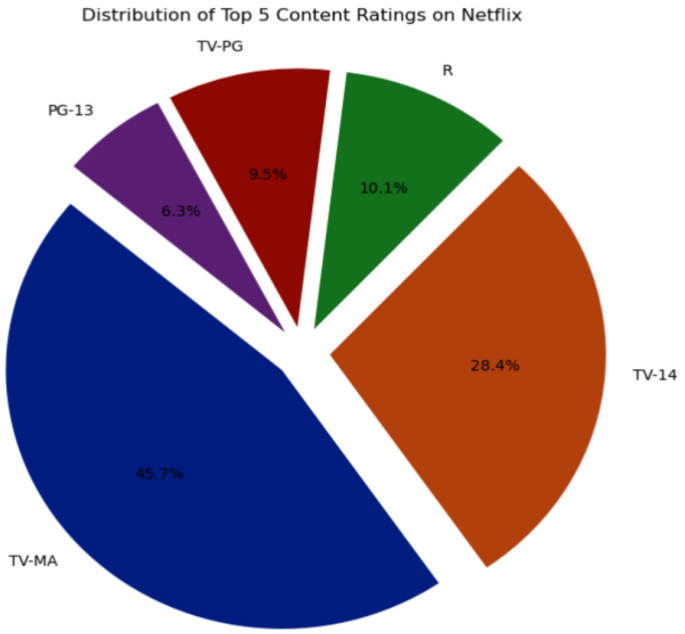
and distribution of viewer opinions within each genre. Such insights can be useful for Netflix's decisions regarding which genres to invest in and what type of content to prioritize to maintain high viewer satisfaction and engagement. The dots outside the boxes seen are the outliers. They represent IMDb scores that fall outside the typical range (IQR), either unusually high or low compared to the bulk of scores in that genre.

3.5 Visualization 4 - Contributed by Rithwik.



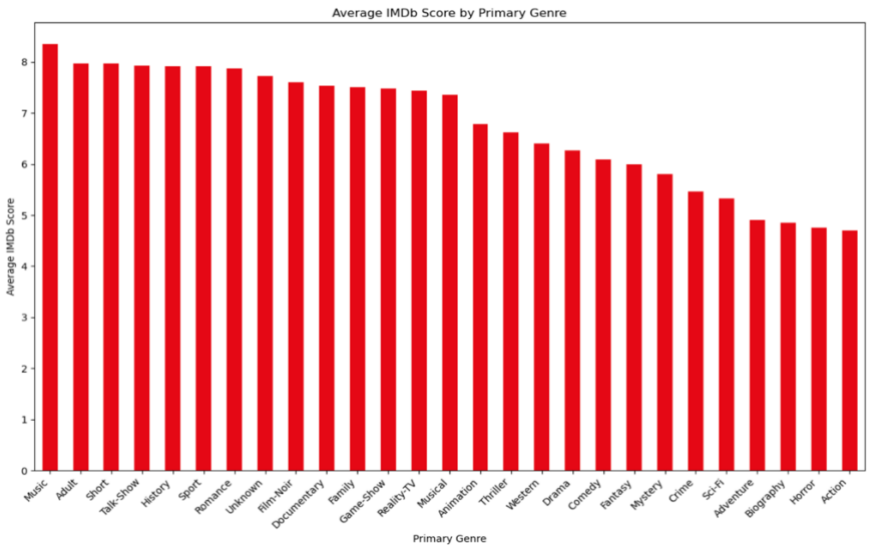
This is a word cloud for the horror genre, which is a visual representation of text data where the size of each word indicates its frequency or importance within a certain context—in this case, likely keywords associated with horror movies or TV shows on a platform such as Netflix. Prominent words like "horror," "mysterious," "supernatural," and "family," among others, suggest common themes and elements that audiences might expect from horror content. The hypothesis behind creating such a word cloud could be to visually analyze and present the most frequent elements or themes that define the horror genre on Netflix. This can help content creators and marketers to understand popular tropes within the genre and tailor their productions and promotions to match audience expectations and trends.

3.6 Visualization 9 - Contributed by Rithwik.



The pie chart depicting the distribution of titles among different content rating categories offers insights into the prevalence of ratings on Netflix. The visualization indicates a preference for mature content, as evidenced by the prominence of TV-MA-rated titles. This observation suggests that the content selection strategy tailored to cater to mature audiences on the platform.

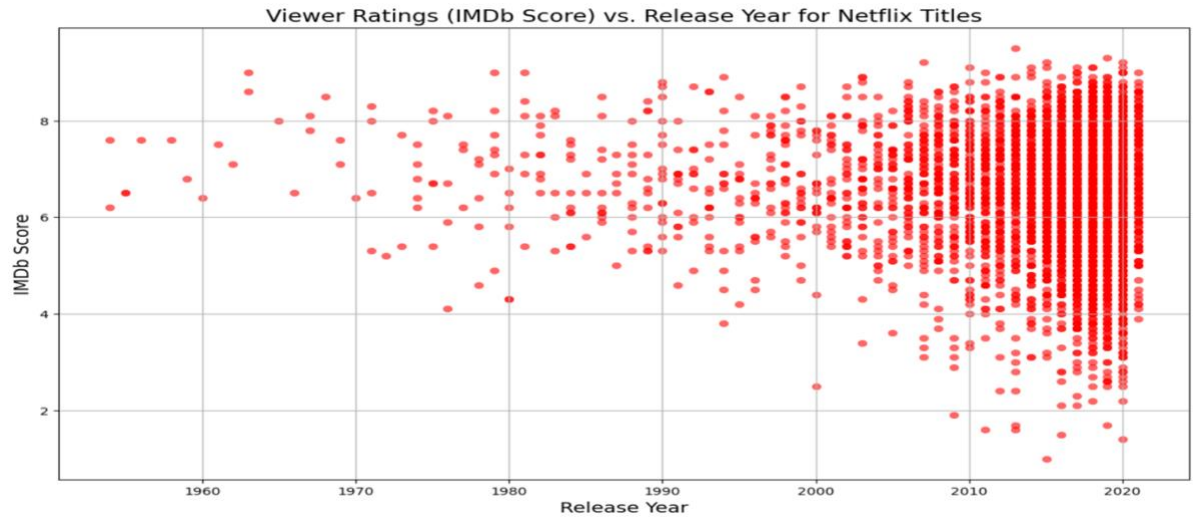
3.7 Visualization 5 - Contributed by Asritha.



Average IMDb Score by Primary Genre Hypothesis: The success of new titles can be partially predicted by their genre, with certain genres consistently performing better in terms of viewer ratings. Explanation: This

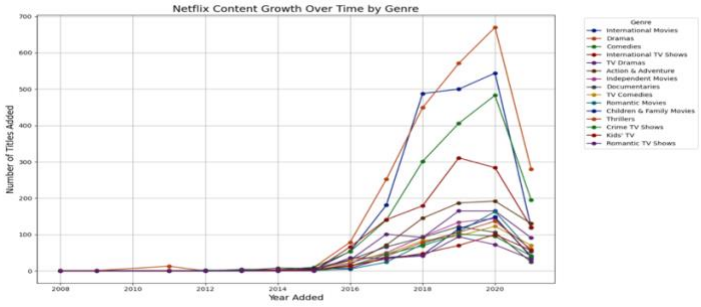
bar chart displays the average IMDb score (used as a proxy for title success) for various primary genres. Higher average scores suggest that titles within these genres tend to be received more favorably by audiences, potentially indicating a safer investment for new productions. Why It's Interesting: By understanding which genres are more likely to yield successful titles, Netflix can make informed decisions about which types of new content to develop or acquire. This insight can guide strategic planning and resource allocation to maximize audience engagement and satisfaction. Responsible Member: Casey, focusing on predicting the success of new titles, would find this analysis invaluable. It provides a data-driven basis for forecasting the potential reception of titles based on their genre, aiding in the strategic selection of content.

3.8 Visualization 6 - Contributed by Sanjna.



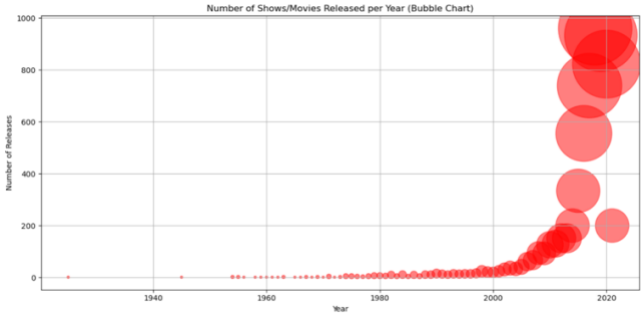
The scatter plot of viewer ratings (IMDb Score) against the release year for Netflix titles offers a way to assess the perceived quality of content over time. This visualization can uncover insights such as: Content Quality Over Time: By observing the distribution of ratings across different release years, you can evaluate whether the quality of content, as rated by viewers, has been consistent, improved, or declined over time. High-Quality Content Identification: Identifying titles with high ratings across different eras can help pinpoint standout content, which can be further analyzed to understand the characteristics of successful titles.

3.10 Visualization 10 - Contributed by Sanjna.



This visualization provides insights into how the variety and volume of content in different genres have evolved on Netflix over the years. By tracking the number of titles added each year across various genres, you can identify trends in content acquisition and production, such as: Content Strategy Insights: Identifying which genres have seen significant growth can indicate Netflix's strategic focus and possibly reflect changing viewer preferences. For example, a surge in documentaries might suggest a growing interest in educational or real-life content. This visualization directly addresses the problem statement regarding how the popularity and evolution of different content types have changed over time, providing a clear visual narrative of content trends on Netflix.

3.9 Visualization 7 - Contributed by Bhargavram.



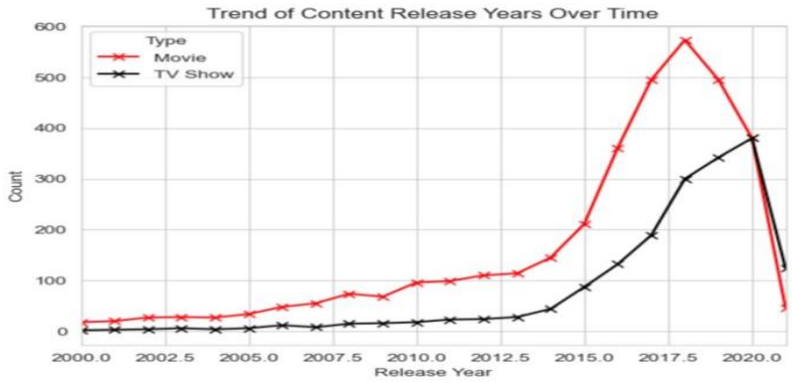
The bubble chart visualizes the number of shows or movies released per year, with each bubble's size proportional to the release count. This type of visualization is quite effective for showing the volume of output each year relative to others, as the differences are immediately apparent through bubble size. The trend shows a relatively low and stable number of releases from the 1950s up to around the late 1990s. Starting in the late 1990s, there's a noticeable upward trend in the number of releases. There is a particularly large increase in releases in recent years, with the largest bubbles appearing towards the right side of the chart, indicating a possible boom in the industry or a change in the way content is produced and released (such as the rise of streaming services). The size of the bubbles is carefully scaled so that they represent the data accurately without overwhelming the chart. Data Source Reliability: Considering the reliability and completeness of the data source, as the sharp increase in recent years may reflect changes in data collection methods or market disruptions.

4. Descriptive Analytics

1. What are the popular content types and how have they evolved over the years?

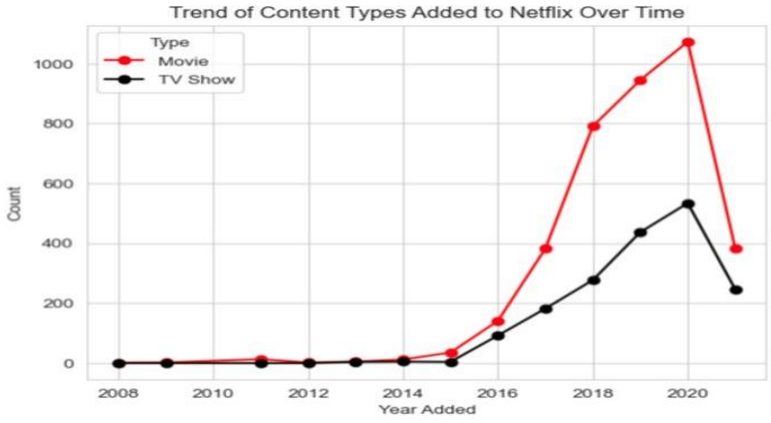
To find the popular content types and how they have evolved over time, we used time series analysis to examine the changing popularity using the date_added and the release_year fields. Tracking the trend over the years will help us examine how the content's release years relate to the date added onto Netflix. The main objective here is to find how the content types have varied over the years. Apart from determining if the audience prefers watching TV shows or movies, we also want to understand if the year added influences the viewer engagement. By analyzing this, we can obtain valuable insights on past data trends on Netflix content types.

The below two visualizations displays information on the past trends for Netflix over the years:



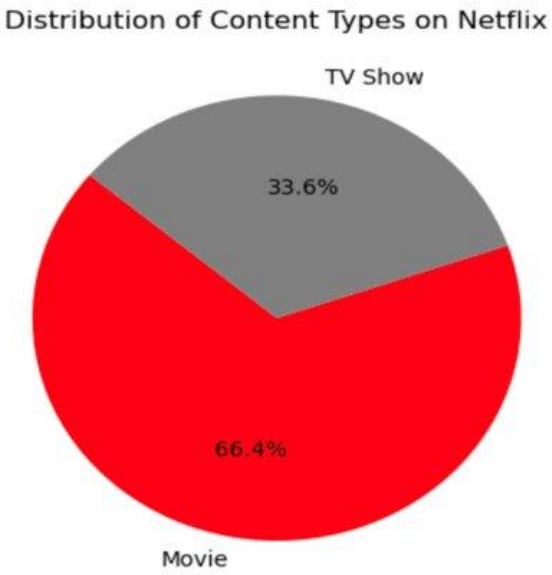
Trend of Content Release Years on Netflix Over Time:

A gradual increase since the year 2000 in the content release years can be seen in the above graph. However, in recent years there is a steep increase indicating that Netflix is continuously adding newer content keeping it fresh for the audiences and up to date.



Trend of Content Types Added to Netflix Over Time:

The line plot above displays the number of movies and TV shows added to the platform year on year. We can observe an upward trend in both the content types indicating there is an increase in the content added over the years however, there is significantly a greater number of movies compared to the TV shows added to the digital library.



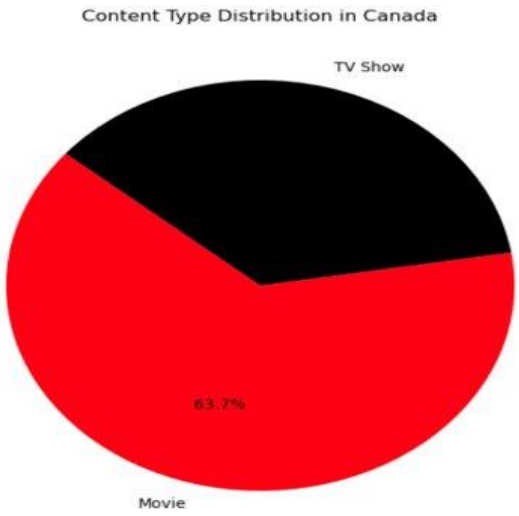
The pie chart above represents the distribution of content types (Movies and TV shows) on Netflix. We can see a high percentage of distribution in movies compared to TV shows.

Key findings:

Based on the above visualizations, we can infer that movies are more popular compared to the TV shows though there is a significant increase in the number of TV shows added to Netflix over the years. Netflix data shows an increase in addition to recent releases indicating that there is a demand for newer and fresh content among the audience. Furthermore, Netflix should focus on producing and investing in movies while also recognising the growth in the demand for TV shows to increase customer base.

2. What are some of the current strategies that are working best for countries around the world?

To understand the current strategies that are working best, we need to perform a few different analyzes such as content distribution analysis, Genre popularity, rating distributions and language preference analysis. We start by analyzing the 'Country Availability' column. We take Canada as an example to perform the different analyses.

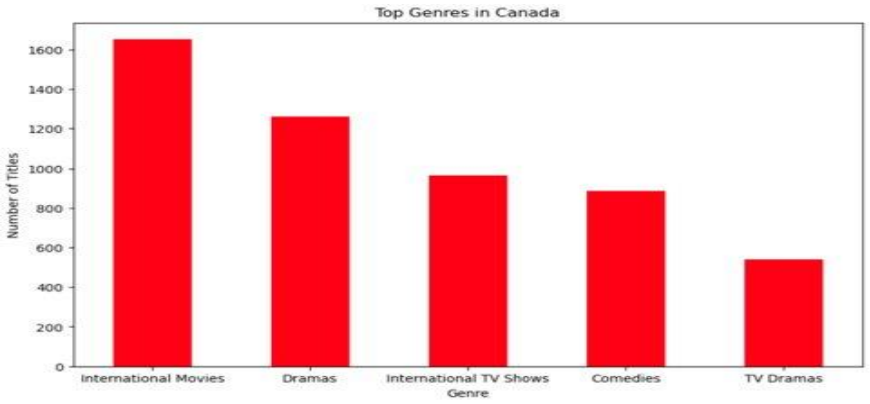


Content distribution in Canada:

Movies: 3,041 titles

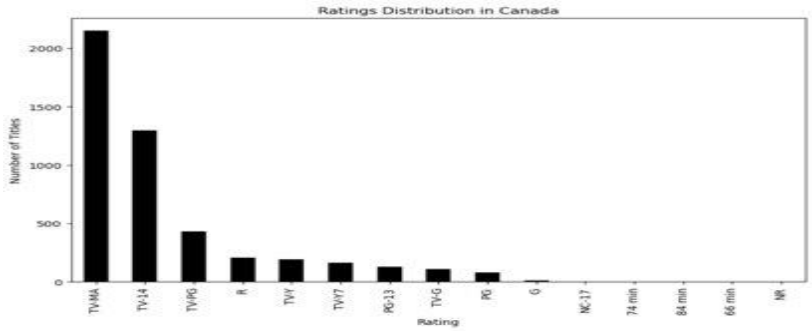
TV Shows: 1,734 titles

The above pie chart displays a high distribution of movies over TV shows in Netflix available in Canada. This could be due to the viewer preferences like the content distribution as a whole.

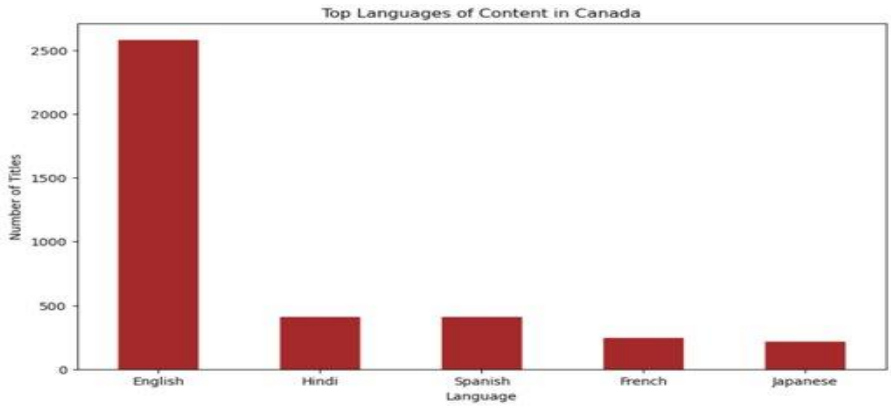


International Movies: 1,653 titles, Dramas: 1,263 titles, International TV Shows: 966 titles, Comedies: 885 titles, TV Dramas: 541 titles

The top genres in Canada are international movies, Dramas and International TV shows. As observed previously, this could be due to the global appeal or the cultural difference making it attractive for audiences in Canada to prefer these genres.



A significant portion of content is rated as TV-MA for Canada indicating that the content is more inclined towards a mature audience followed by TV-14 making it more appealing to general audience and older teenagers.



English is the primary language for Canada aligning closely with that of the United States and the U.K. Further, there is a good portion of content preferred by the audience in Hindi, Spanish, Japanese and French. This could be due to Canada's audience preferring multicultural content around the world and a portion of Canada's population speaking these languages.

Key findings:

Canada's audience prefers more movies compared to TV shows; hence Netflix should focus on producing and investing more in genres of international movies and dramas. There should also be a mix of content for adults and teenagers made available. Lastly, focus on content in English and French for the general audience and content in Hindi for the South Asian audience should be done.

5. Predictive Analytics

3. How can we predict which new titles will be a hit?

To predict which new titles will be a hit, we need to use machine learning models that can analyze past data and patterns for correlation. We will consider 'IMDb score' as our key attribute to measure our hits. Other potential success indicators such as viewership numbers, votes, awards won and so on through 'Feature Selection' will be used. we focus on predicting the potential success of new titles on Netflix. By using the IMDb score as a proxy for hit status, we leverage exploratory data analysis (EDA) to identify key features that signal a title's success. Our model then performs regression analysis on preprocessed data to predict which new shows and movies are likely to capture audiences and gain critical acclaim.

We started with our baseline model- Linear Regression. We later learnt that it wasn't the best choice for predicting IMDb scores because it assumes a straightforward relationship between features, doesn't handle unusual data well, and can't easily manage complex patterns or categories without a lot of extra work. On the other hand, RandomForestRegressor is often more effective because it can automatically manage these complexities, handle both numerical and categorical data without prior conversion, and provides insights on which features most influence the IMDb scores.

The model's accuracy is quantitatively assessed using metrics like Mean Squared Error (MSE) and R-squared, ensuring that our predictions are both precise and reliable.

R- square	**0.8687244053500376**
Mean Squared Error (MSE)	**0.177694919925121**

Key findings:

The model's performance is quite strong, which is promising for using it to predict whether new titles will be hits based on the defined IMDb score threshold. We can determine the hit status of a title if the predicted IMDb score is above or below a certain limit.

Output:

```
Please enter the details for the prediction:
Enter type (Movie/TV Show): Movie
Enter director: Steven Spielberg
Enter country: United States
Enter genre: Fantasy
Predicted IMDb Score: 6.19
```

6. Prescriptive Analytics

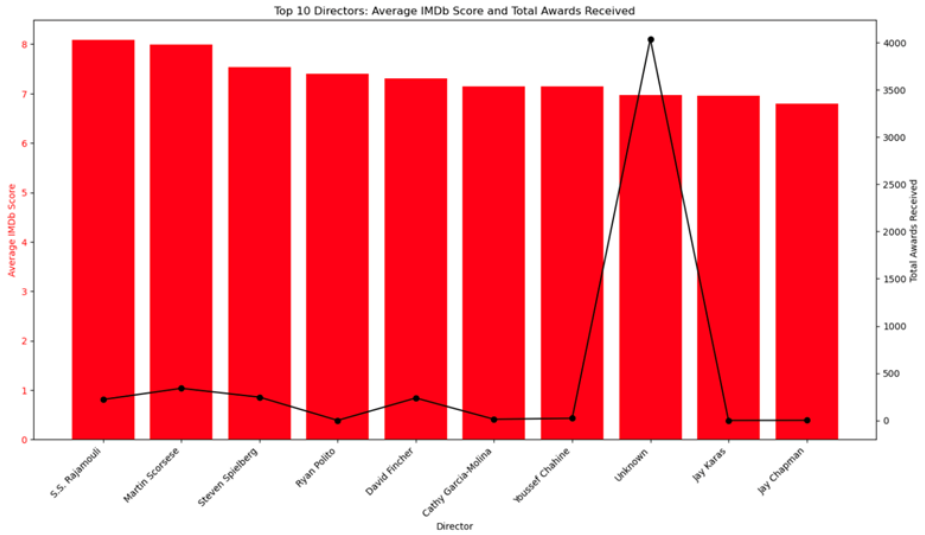
4. Who should be Netflix partnering with for newer and hit content?

For determining which industry figures Netflix should partner with, considering both critical acclaim and popularity, we focused on the highest ratings (IMDb Score), IMDb votes and awards. To visualize this data for the top 10 directors, cast members, writers, and production houses, we plotted each of the three metrics (Average IMDb Score, Total Awards Received, and Total IMDb Votes) on the same chart with normalized values. The normalization method used in the code is Min-Max normalization. This method is particularly used as it brings all values into the range [0, 1] while preserving the relationships among the original data points and allows comparison of these metrics side-by-side. This technique scales the data linearly between a minimum and maximum value, typically 0 and 1. The formula for Min-Max normalization is:

$$x\text{-normalized} = x - \min(x) \,/\, (\max(x) - \min(x))$$

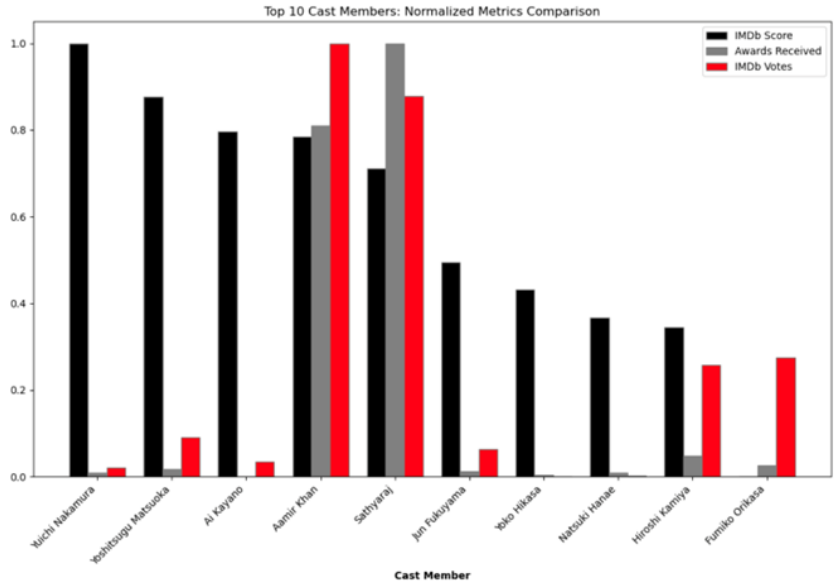
Analysis of directors:

We looked at the average IMDb Score, IMDb votes and total number of awards received for the content they have directed. We only considered directors with a significant number of works to ensure we are making recommendations based on a consistent track record. The below directors have shown their ability to produce content that is critically acclaimed, reach and impact on viewers, and recognized by awarding bodies making them potentially strong partners for developing new content.



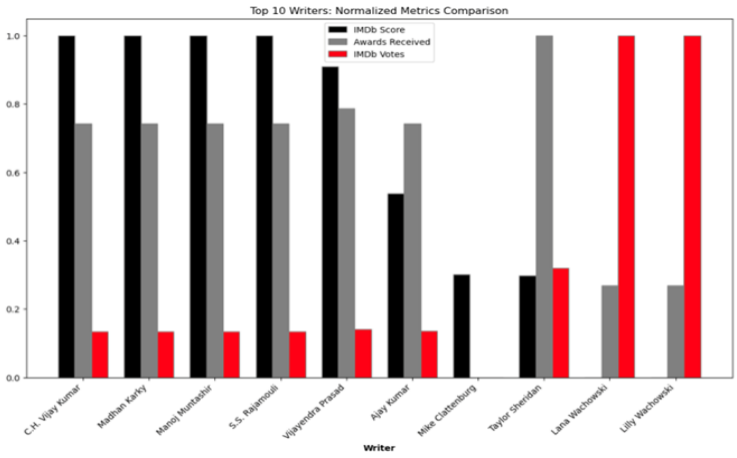
Analysis of cast members:

We looked at each cast member's average IMDb Score for content in which they have appeared, total awards received and nominated for, total number of IMDb votes for their content and number of titles each cast member has appeared. The below cast members not only have high average IMDb Scores, suggesting they choose projects that are well-received, but also have a significant number of IMDb votes, indicating that their films are popular with a broad audience.



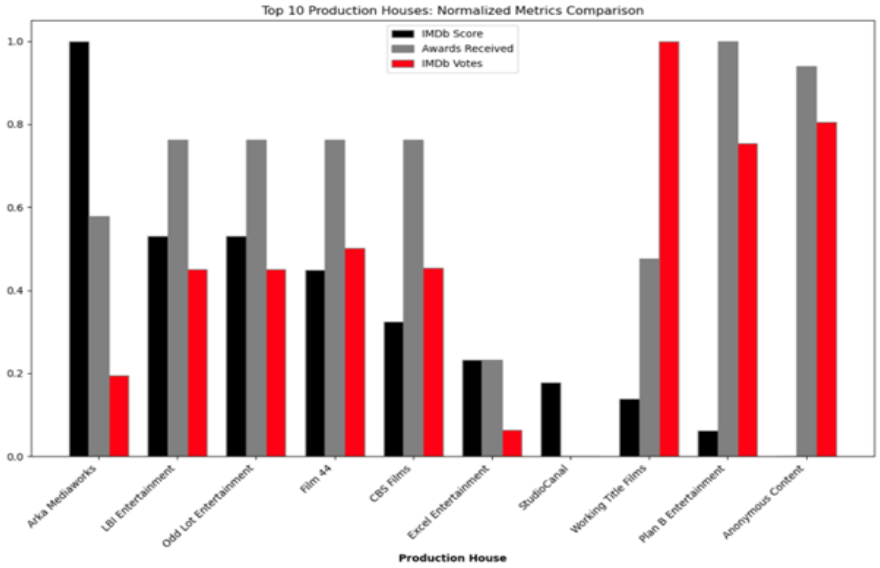
Analysis of writers:

We looked at each writer's Number of titles, Average IMDb, Total number of awards received and nominations and Total number of IMDb votes for the content they have written. The below writers have shown that they can deliver content that resonates with both critics and audiences. Their work has received significant awards and amassed many IMDb votes, indicating their potential to attract viewers.



Analysis of production houses:

We looked at each production house's Number of titles, Average IMDb Score, Total number of awards received and nominations and Total number of IMDb votes of the content they've produced. The below production houses have demonstrated an ability to produce content that resonates with both critics and audiences, and they have a track record of awards recognition. They might be the best to partner with for new content.



Key findings:

Combining these insights, partnerships with industry figures like S.S. Rajamouli, Aamir Khan, Vijayendra Prasad, and production houses like Arka Mediaworks and Working Title Films could potentially lead to creation of content that is well-received, widely watched, critically acclaimed and successful.

7. Additional Work

Apart from the visualizations and models discussed earlier, we have developed a comprehensive dashboard that includes several static visualizations directly addressing our project's problem statements. Additionally, the dashboard features an interactive visualization (multi-linked visualization) where users have the option to select a specific year in which they wants to get more information. Once selected, the top genres with respective average IMDb score are associated with that year are displayed in a dynamic bubble chart. Selecting a genre from this bubble chart triggers a word cloud, which display the most common words used in that genre, offering insights into thematic elements and content styles.

Furthermore, the dashboard is integrated with our predictive model, providing a user-friendly interface where inputs can be submitted to predict the IMDb score and assess the potential success of future movie titles. This feature enhances the dashboard's utility, making it a valuable tool for both analysis and prediction. To facilitate a better understanding and accessibility of our dashboard, we have uploaded a Demo (video walkthrough) to a Google Drive link **(Click here for Demo video)**, which showcases its functionality and various features in action. This UI in a real world scenario is intended to help stakeholders and team members of Netflix effectively utilize the insights from the dashboard for strategic decision-making and content planning.

We used a combination of technologies to achieve this UI :

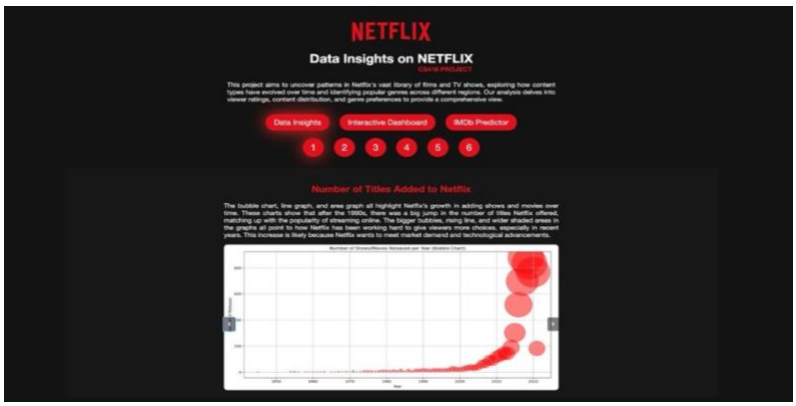
Interactive Dashboard: The interactive dashboard is developed using the Dash library, which is built on top of Plotly. Dash provides a powerful framework for building interactive web-based visualizations using Python. The dashboard integrates these visualizations seamlessly, providing insights into Netflix's content distribution, viewer preferences, and more.

Machine Learning Model: The backend for the IMDb score prediction utilizes Scikit-Learn for implementing the machine learning model. Scikit-Learn offers tools for data preprocessing and model development.

Both these components are linked with your frontend, which is developed using HTML, CSS, and JavaScript. The communication between your frontend and Python backend is managed through a HTTP server, enabling requests from the UI to be processed by Python scripts which then return the results back to the UI.

Dashboard: (Click here for Demo video)

Data Insights



Interactive Dashboard

IMDb Predictor

Note:

GITHUB Repo: https://github.com/Rohith1110/Datascience_Project

Google drive link (UI Demo): https://drive.google.com/file/d/1YMSNR7APX6J6lQ-_i4TXcXnIczLClgLI/view?usp=sharing

8. Results and Conclusion

The project underscores the dynamic nature of content management in the streaming industry and highlights the importance of data-driven decision-making in shaping content strategies. By effectively predicting what new titles are likely to succeed, Netflix can better allocate resources, tailor content for specific markets, and stay ahead in the highly competitive streaming landscape. The findings from this project not only benefit strategic planning but also provide a framework for continuous improvement and adaptation to changing viewer preferences and market dynamics.

Summary of Findings:

Growth of Netflix Library:
Over the years, Netflix has significantly expanded its content library, with a steady increase in the number of movies and TV shows available to its subscribers. This expansion reflects Netflix's strategy to cater to a broad range of viewer preferences and demographics.

Content Composition:
The analysis highlighted that movies constitute the larger portion of Netflix’s catalog, outnumbering TV shows. This might suggest a strategic focus on films, potentially due to their wider appeal or shorter commitment compared to series.

Recent Content Influx:
A significant portion of Netflix's offerings were produced and released in the last twenty years, indicating an emphasis on modern and contemporary content, likely to attract a younger audience base.

Geographical Availability:
The United States boasts the largest selection of Netflix titles, closely followed by Canada. This suggests a strategic focus on North American audiences, who perhaps represent a significant portion of Netflix's market share.

Diversity in Filmmaking:
Netflix's strategy includes a mix of content from both renowned directors and emerging talents, which helps in maintaining a diverse and dynamic catalog appealing to various viewer tastes and preferences.