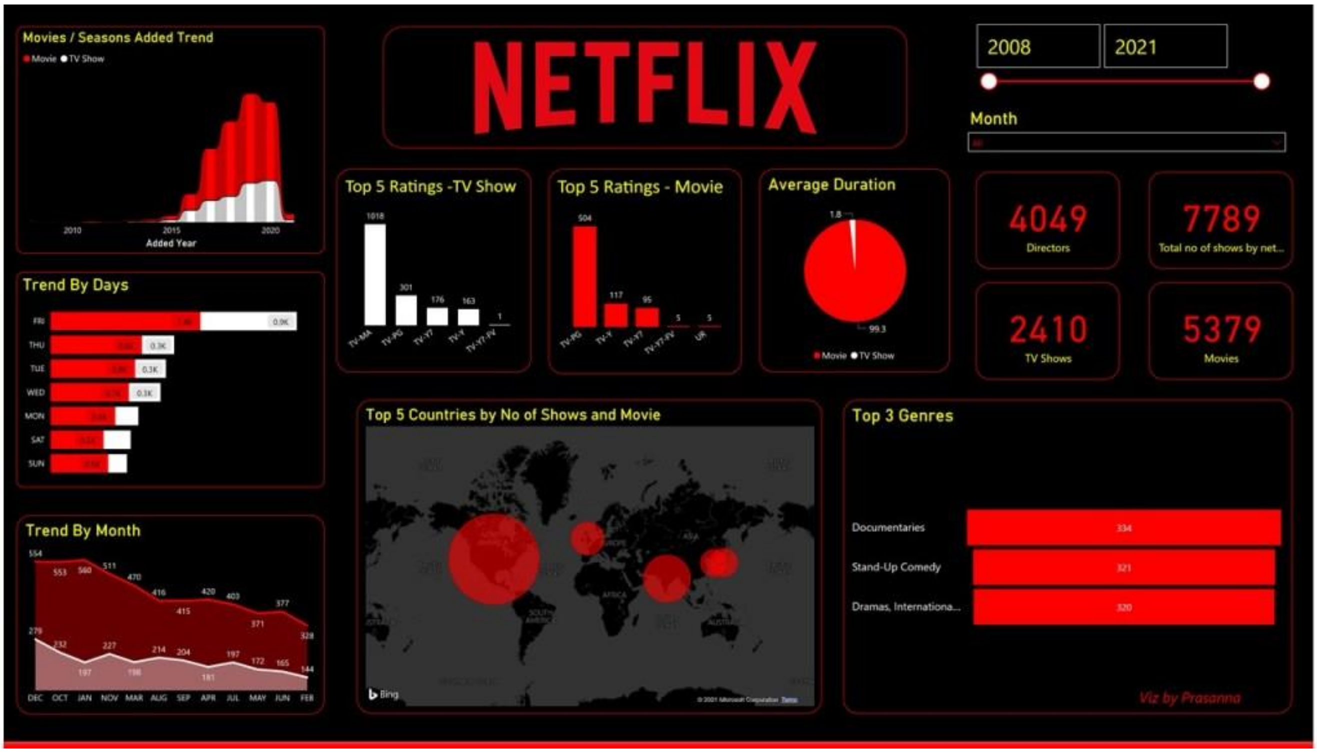
Netflix Data: Cleaning, Analysis, and Visualization

Tarun C
8th sem, , ECE
NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY,
BANGALORE-560064



Abstract

With the rapid rise of streaming platforms, Netflix has emerged as a global leader in online content distribution. This project focuses on analyzing Netflix's catalog of TV shows and movies through data cleaning, exploration, and visualization using Python. The dataset spans content added from 2008 to 2021 and includes information such as type, genre, director, country, release year, and rating.

The analysis began with cleaning the dataset by handling missing values, removing duplicates, and converting data types. A variety of visualizations were then used to uncover patterns in content distribution, genre frequency, release trends over time, rating distributions, and country-wise contributions.

Key insights include the dominance of movies over TV shows in Netflix's catalog, the growing volume of content added between 2017 and 2020, and the prominence of genres like Dramas and Comedies. The project also identified top directors, most active countries, and rating preferences, culminating in a comprehensive view of Netflix's global content strategy.

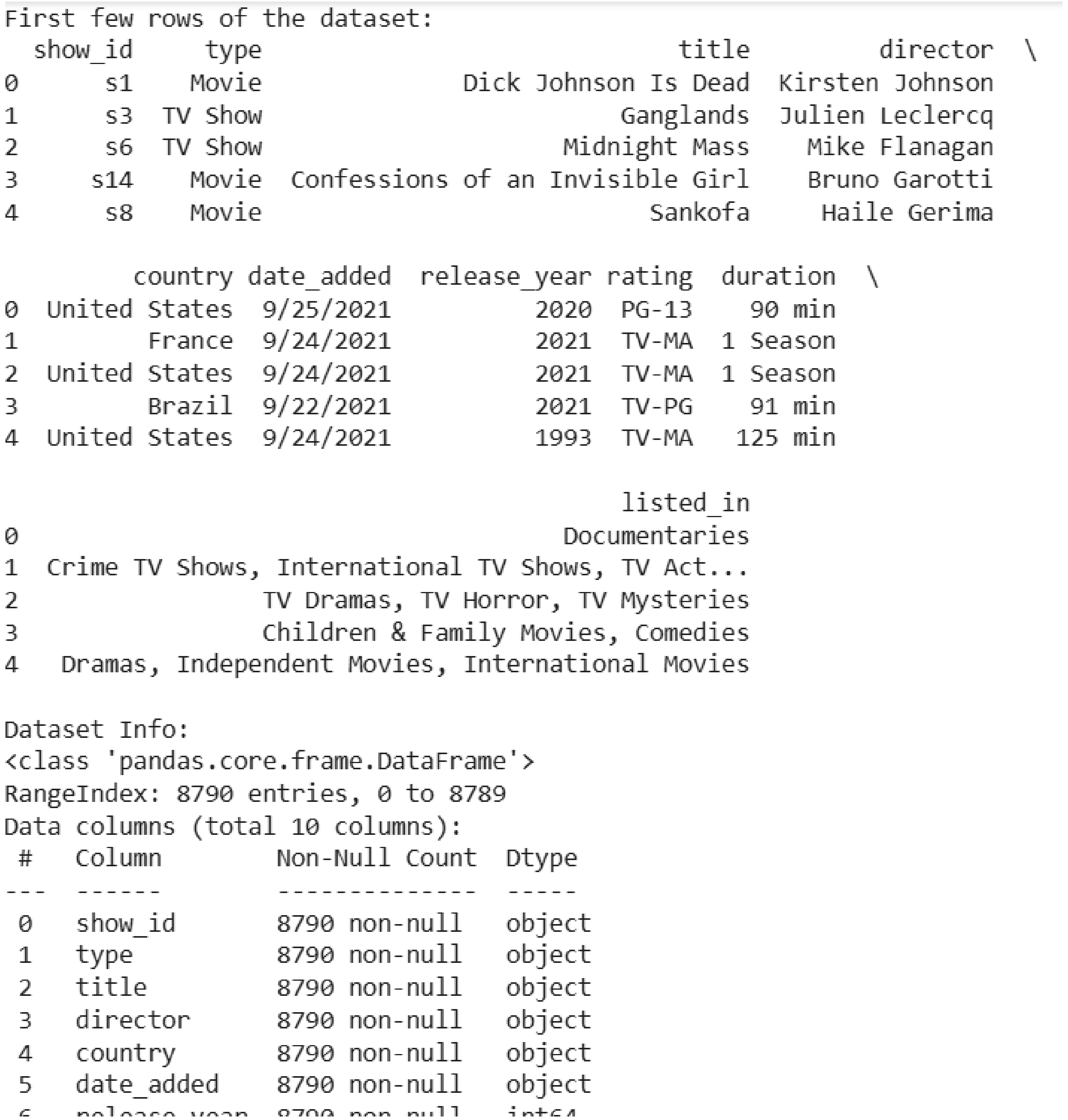
This analysis not only demonstrates fundamental data processing and visualization skills but also sets the foundation for future projects such as recommendation systems and content trend forecasting.

KEYWORDS: Netflix, Data Cleaning, Exploratory Data Analysis, Python, Data Visualization, Streaming Analytics, Genre Analysis, Content Trends, Movie vs TV Show, Rating Distribution, Word Cloud, Time Series Analysis, Country-wise Content, Director Frequency

Data Collection

The dataset and the sources used for this process are listed below :

https://drive.google.com/file/d/1cWcK8cddROe_DSv5zH5Fk70d32tK3ftf/view?usp=sharing



```
First few rows of the dataset:
  show_id     type                           title         director  \
0      s1    Movie            Dick Johnson Is Dead  Kirsten Johnson
1      s3  TV Show                       Ganglands  Julien Leclercq
2      s6  TV Show                   Midnight Mass    Mike Flanagan
3     s14    Movie  Confessions of an Invisible Girl    Bruno Garotti
4      s8    Movie                         Sankofa     Haile Gerima

         country date_added  release_year rating  duration  \
0  United States  9/25/2021          2020  PG-13    90 min
1         France  9/24/2021          2021  TV-MA  1 Season
2  United States  9/24/2021          2021  TV-MA  1 Season
3         Brazil  9/22/2021          2021  TV-PG    91 min
4  United States  9/24/2021          1993  TV-MA   125 min

                                           listed_in
0                                      Documentaries
1  Crime TV Shows, International TV Shows, TV Act...
2                    TV Dramas, TV Horror, TV Mysteries
3                   Children & Family Movies, Comedies
4   Dramas, Independent Movies, International Movies

Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8790 entries, 0 to 8789
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8790 non-null   object
 1   type          8790 non-null   object
 2   title         8790 non-null   object
 3   director      8790 non-null   object
 4   country       8790 non-null   object
 5   date_added    8790 non-null   object
 6   release_year  8790 non-null   int64
```

Figure 1 : *output after loading dataset*

Data Cleaning

Before performing any analysis, the raw Netflix dataset underwent several essential cleaning steps to ensure consistency, accuracy, and usability.

1. Duplicate Removal

- Duplicate rows were identified and removed using drop_duplicates() to ensure each title appears only once.
- Result: Reduced dataset to unique entries, ensuring unbiased visualizations.

2. Handling Missing Values

- Columns such as director and country were considered critical for content insights.
- Rows with missing values in these fields were dropped using dropna().

3. Data Type Conversion

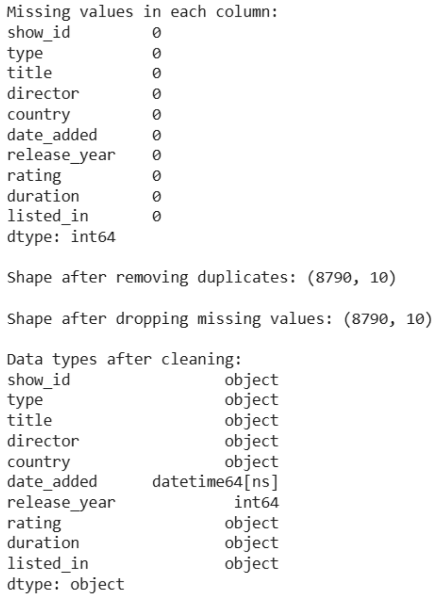
- The date_added column was converted from string to datetime format using pd.to_datetime() for accurate temporal analysis.

4. Feature Extraction

- From the date_added column, new time-based features were extracted:
    - year, month, and day of addition

5. Genre Column Transformation

- The listed_in column was split into a list of genres per title.
- A new column, genres, was created to facilitate genre frequency analysis.



```
Missing values in each column:
show_id         0
type            0
title           0
director        0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
dtype: int64

Shape after removing duplicates: (8790, 10)

Shape after dropping missing values: (8790, 10)

Data types after cleaning:
show_id                 object
type                    object
title                   object
director                object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration                object
listed_in               object
dtype: object
```

Figure 2 : *data cleaning*

Exploratory Data Analysis (EDA)

With a cleaned dataset, a series of visualizations were created to uncover patterns, trends, and insights related to the content available on Netflix.

1. Content Type Distribution

- A comparison between **Movies** and **TV Shows** showed that movies dominate the platform, comprising approximately 70% of the total content.
- Visualized using count plots and pie charts.

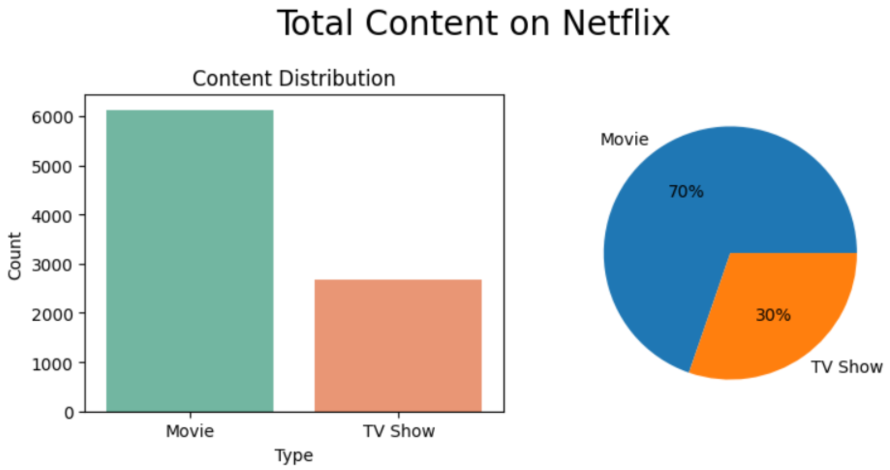


Figure 3 : *total content on netflix*

2. Genre Analysis

- The `listed_in` column was processed to extract individual genres.
- The top 10 most common genres were identified, with **Dramas**, **International Movies**, and **Comedies** leading the list.
- Displayed using horizontal bar plots for clarity.

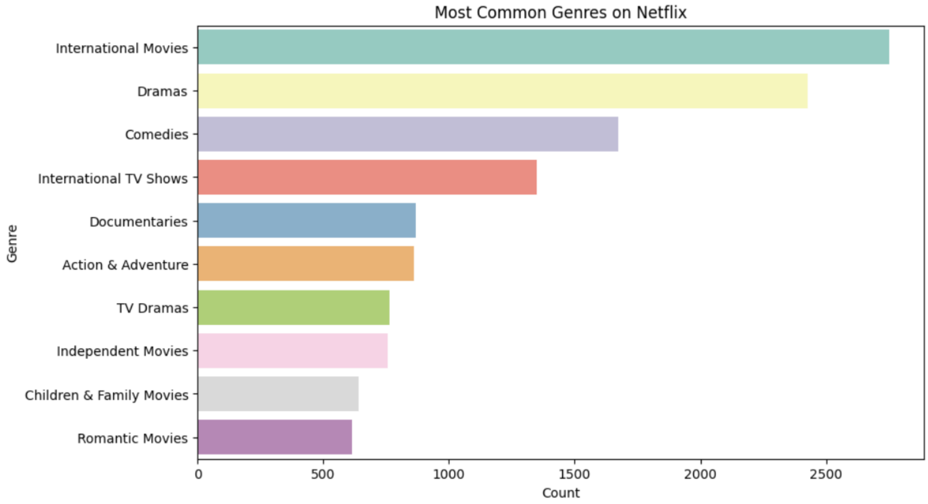


Figure 4 : *genre analysis*

3. Content Addition Over Time

- By extracting the year and month from `date_added`, trends in content acquisition were examined.
- A significant spike was observed between 2017 and 2020, highlighting Netflix's aggressive expansion period.
- Year-wise and month-wise additions were plotted using count plots and line graphs.

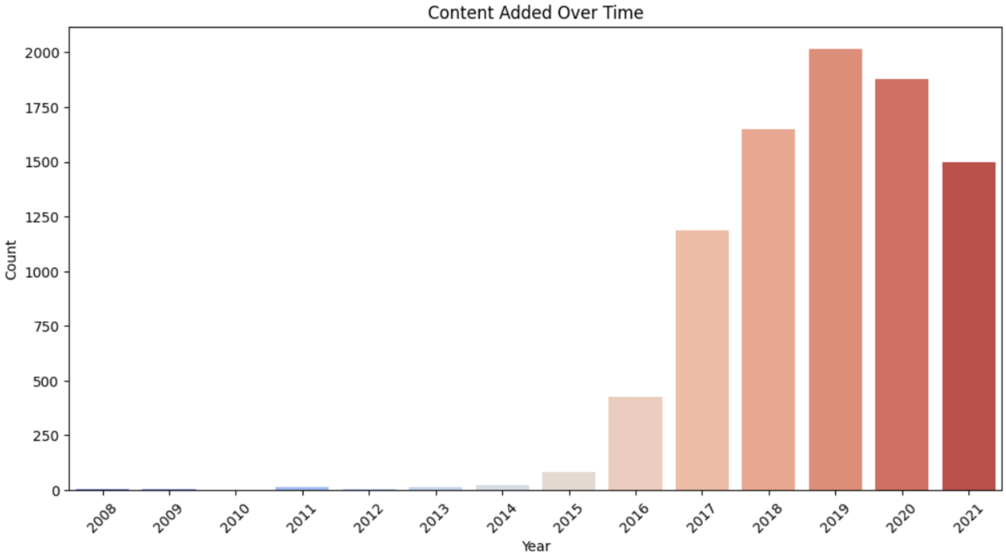
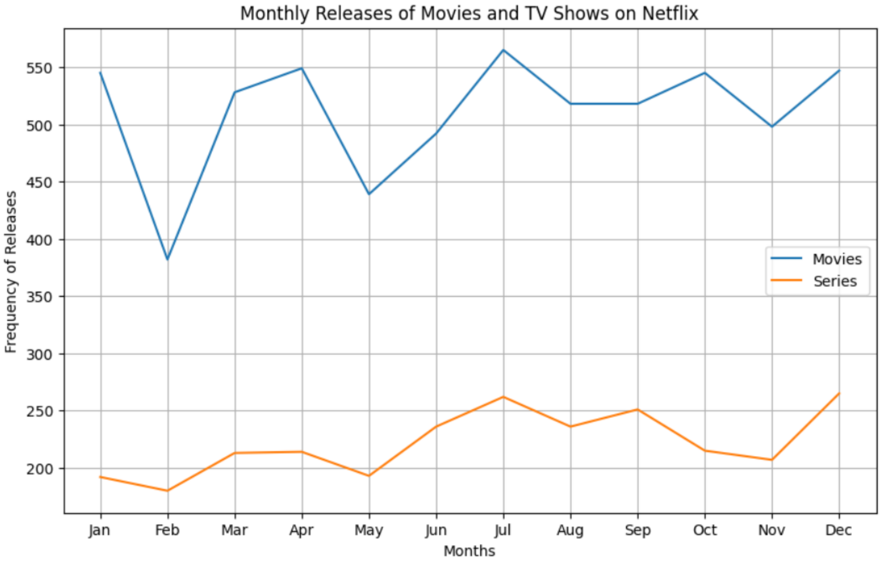


Figure 5 : *content added over time*

4. Monthly and Yearly Releases by Type

- Separate plots for movies and TV shows showed seasonal and annual fluctuations in new content additions.
- Monthly trends revealed consistent releases throughout the year, with minor peaks in March and December.



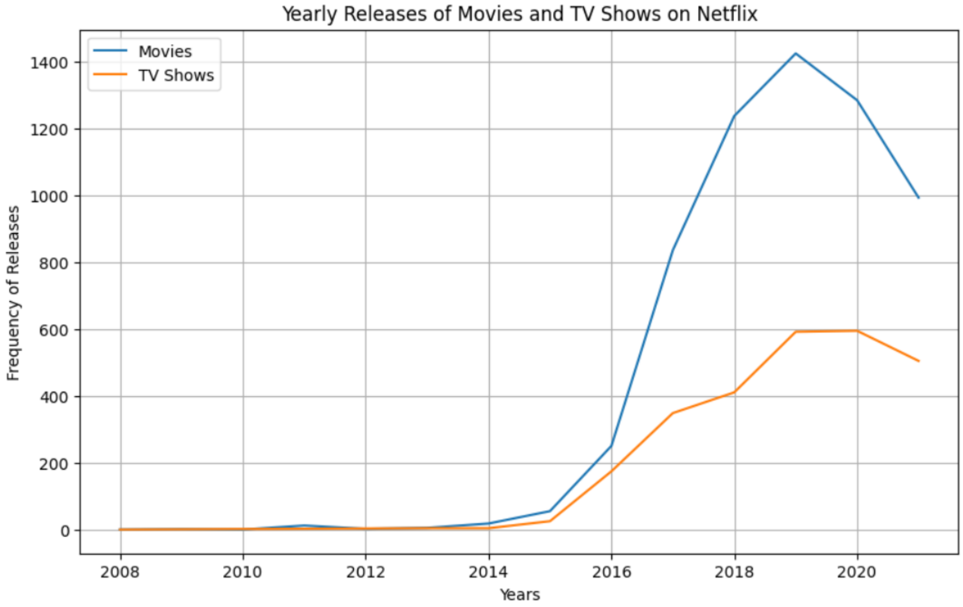


Figure 6 : *Monthly and Yearly Releases by Type*

5. Director Frequency

- A bar chart of the top 10 directors revealed names with the most contributions to Netflix's catalog, such as **Rajiv Chilaka** and **Alastair Fothergill**.

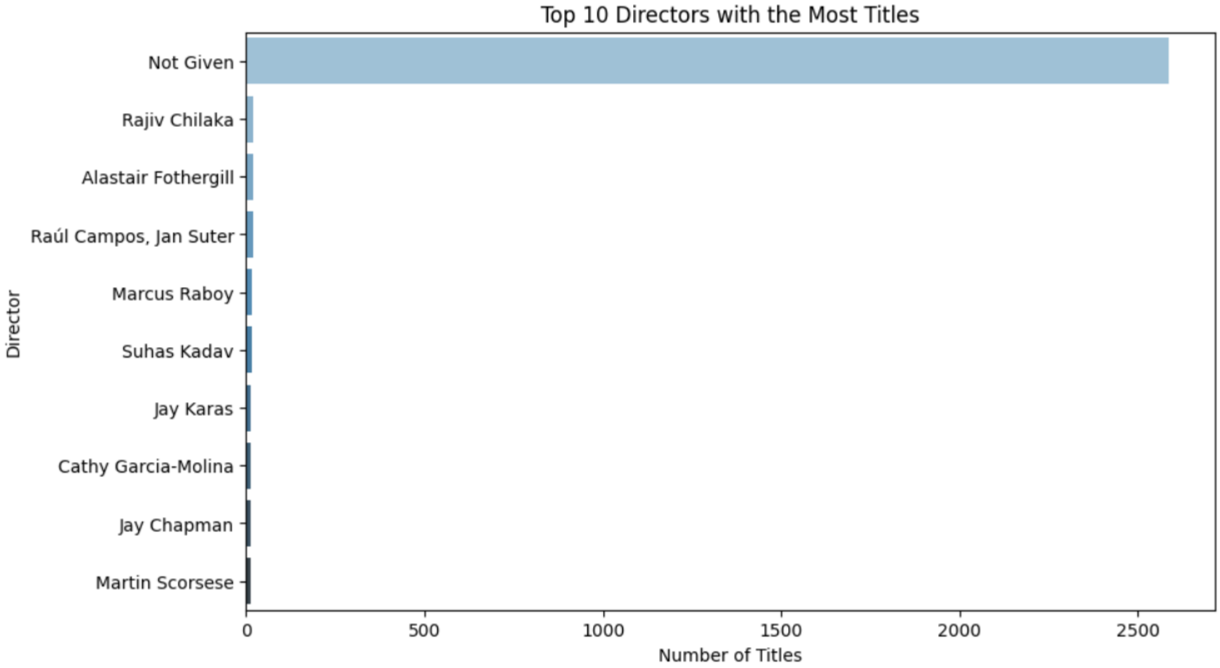


Figure 7 : *top 10 directors with most titles*

6. Country-wise Content Distribution

- The United States, India, and the United Kingdom were the top three contributors in terms of content origin.
- The top 10 countries were visualized using bar charts.

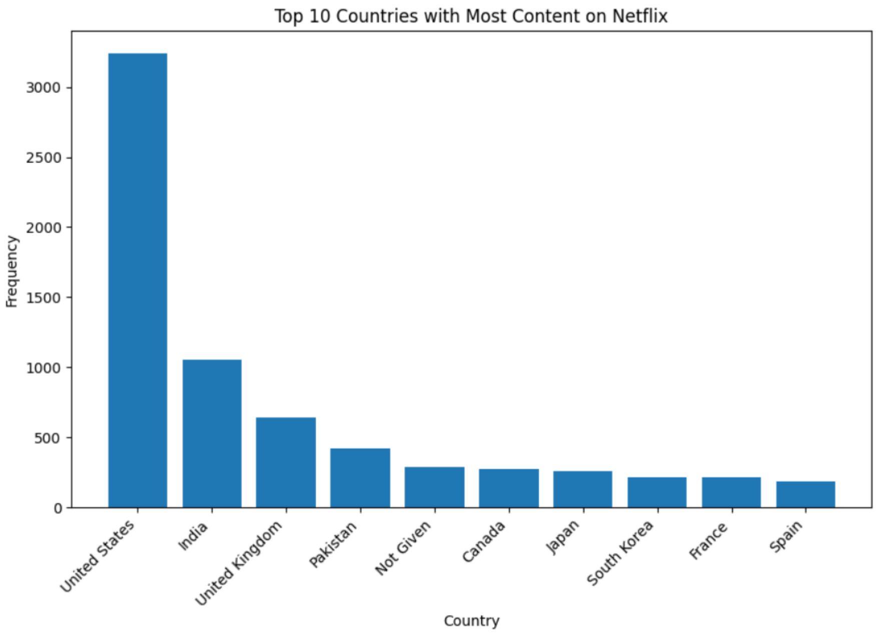


Figure 8 : *top 10 countries with most content*

7. Rating Distribution

- The most frequent content ratings included **TV-MA**, **TV-14**, and **TV-PG**.
- Both bar charts and pie charts were used to highlight distribution across rating categories.

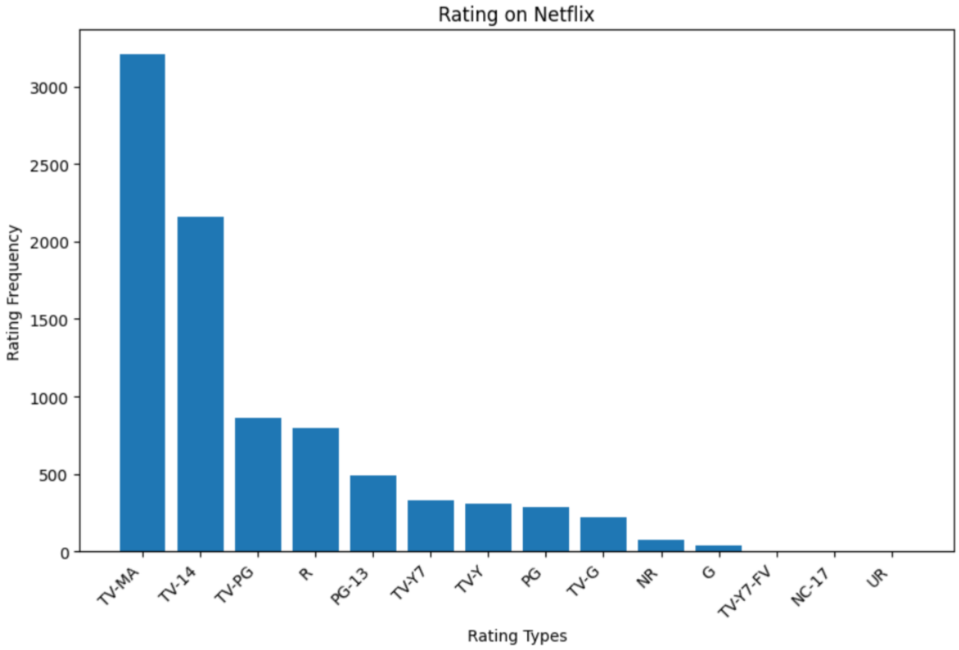


Figure 9 : *rating on netflix*

8. Word Cloud of Movie Titles

- A word cloud was generated from movie titles to give a visual overview of common terms and naming patterns.

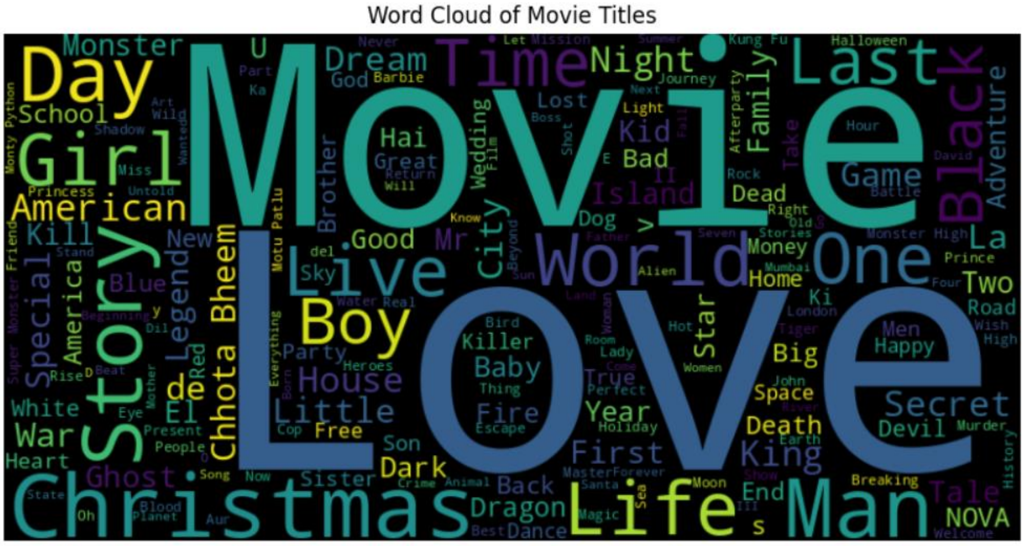


Figure 10 : *world cloud of movie titles*

9. Genre Trends by Content Type

- Separate analyses for **Movies** and **TV Shows** identified genre popularity within each category.
- Bar plots displayed the top 10 genres for both.

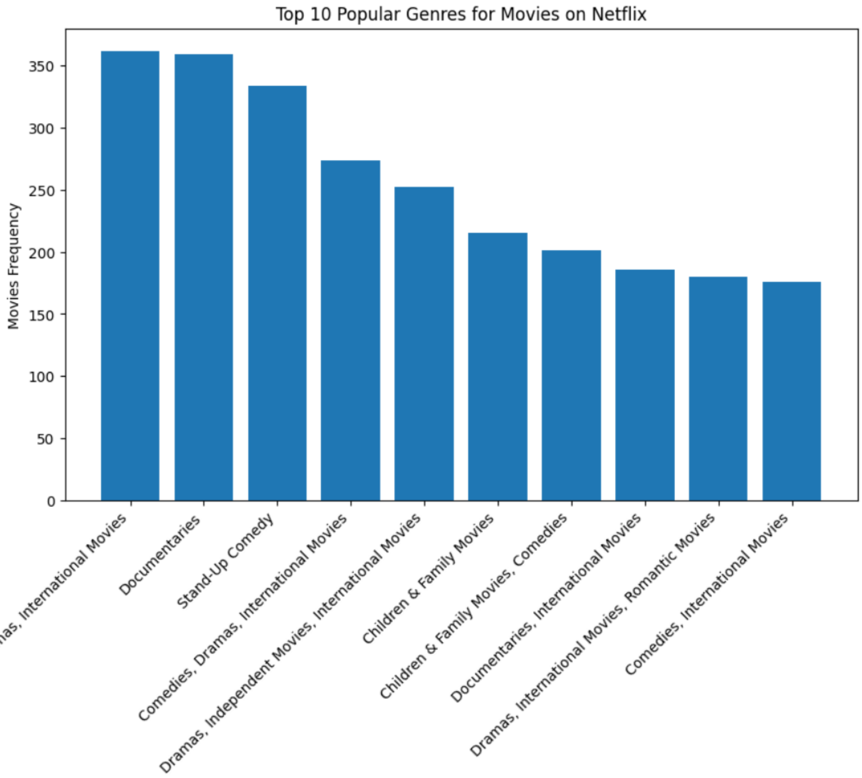


Figure 10 : *top10 popular genres for movies on netflix*

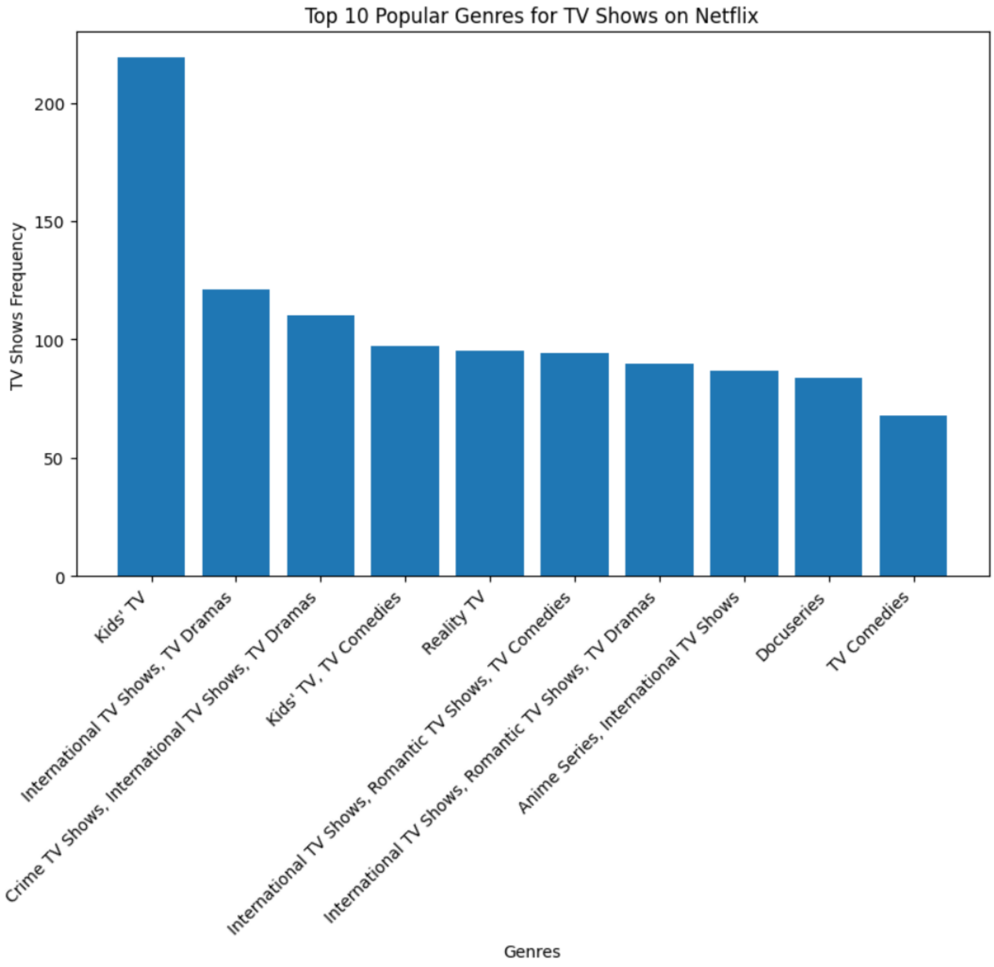


Figure 11 : *top10 popular genres for tv shows on netflix*

CONCLUSIONS AND INSIGHTS

This project involved cleaning, analyzing, and visualizing a dataset of Netflix content to derive meaningful insights about its structure, trends, and distribution. The following conclusions were drawn from the analysis:

1. **Content Composition**
    - Movies make up the majority of the Netflix catalog, accounting for approximately 70% of the total content.
2. **Genre Distribution**
    - The most prevalent genres include Dramas, International Movies, and Comedies, indicating a strong global and story-driven focus in Netflix's offerings.
3. **Temporal Trends**
    - A steady rise in content additions was observed from 2015, peaking between 2018 and 2020. This reflects Netflix's strategy to expand its library and market presence.
4. **Geographic Distribution**
    - The United States leads in content contribution, followed by India and the United Kingdom. This supports the platform's concentration in English-speaking and high-population markets.
5. **Top Contributors**
    - Directors like Rajiv Chilaka and Alastair Fothergill appear frequently, suggesting collaborations with specific creators or production houses.
6. **Rating Patterns**
    - TV-MA and TV-14 are the most common ratings, highlighting the platform's focus on mature audiences.
7. **Seasonality and Scheduling**
    - Slight peaks in content releases during March and December indicate strategic timing, possibly aligned with holiday seasons or fiscal cycles.

The analysis demonstrates how data exploration can uncover critical business insights and viewer behavior patterns, serving as a foundation for deeper predictive modeling and recommendation systems.

Next Steps

Building on the insights obtained from this analysis, several opportunities exist for further exploration and development:

1. **Feature Engineering**
    - Introduce new variables such as:
        - Genre count per title
        - Duration converted to numeric values (e.g., minutes or episodes)
        - Title length or keyword analysis for content classification
2. **Machine Learning Applications**
    - Use the cleaned dataset to build:
        - Content recommendation systems based on genre, country, or user preferences
        - Trend prediction models to forecast future content acquisition patterns
        - Clustering algorithms to group similar shows or movies
3. **Advanced Visualization**
    - Develop interactive dashboards using tools like Tableau or Plotly for real-time exploration of Netflix content trends.
4. **Natural Language Processing (NLP)**
    - Analyze title or description text for thematic insights or sentiment analysis.

This project serves as a foundation for a wide range of data science and business intelligence applications in the streaming media domain.

REFERNCES

Asfand Ali et al. (2025). *Data-Driven Insights: Machine Learning Approaches for Netflix Content Analysis and Visualization*. Journal of Engineering Research and Reports.
https://www.journaljerr.com/index.php/JERR/article/view/1471

Wanqi Zhang (2022). *Understanding the Development of Netflix during Recent Years through Data Visualization*. BCP Business & Management.
https://bcpublication.org/index.php/BM/article/view/386

Muhammad Zahriel Ismail et al. (2024). *An Analysis of Big Data Adoption: A Case Study of Netflix*. Advances on Business, Management and Accounting.
https://www.pablis.org/index.php/abma/article/view/4

Syed Reshma Banu, B. Sravani (2023). *Netflix Movies and TV Shows Data Analysis*. International Journal for Research in Applied Science and Engineering Technology (IJRASET).
https://www.ijraset.com/research-paper/netflix-movies-and-tv-shows-data-analysis

Srivatsa Maddodi, Krishna Prasad K. (2024). *Netflix Big Data Analytics: The Emergence of Data-Driven Recommendation*. International Journal of Case Studies in Business, IT, and Education (IJCSBE).
https://supublication.com/index.php/ijcsbe/article/view/1495

Sunitha B. K. et al. (2024). *A Study on Data Analytics Techniques Used by Netflix to Personalize Recommendations*. International Journal of Scientific Research and Engineering Management (IJSREM).
https://ijsrem.com/download/a-study-on-data-analytics-techniques-used-by-netflix-to-personalize-recommendations/

Faheena Thesni, Goutham Raj, Alwin Poulose (2024). *Visualizing the Global Streaming Landscape: An In-Depth Analysis of Netflix Content Distribution Using Tableau Data Visualization Tool*. 2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT).
https://www.researchgate.net/publication/382161165_Visualizing_the_Global_Streaming_Landscape_An_In-Depth_Analysis_of_Netflix_Content_Distribution_Using_Tableau_Data_Visualization_Tool

Chinmay Samir Kulkarni (2024). *Netflix Data Visualization Guide*. Sri Balaji University.
https://www.researchgate.net/publication/385650785_Netflix_Data_Visualization_Guide_Prof_Minal_Dutta