INSY- 5377

Web and Social Analysis

Analyzing Trends and User Engagement on Netflix

Professor: Prof. Riyaz Sikora

Presented by Group 2

Sharwari Pathak

Siddhesh Karle

Pratiksha Mohite

Raj Panchal

Dibya Chudal

Objective:

The main objective of this project is to leverage comprehensive data analysis techniques to uncover and understand trends, preferences, and predictive factors in online streaming content on Netflix. Through our meticulous steps involving data collection, preprocessing, exploratory and predictive analysis, we aim to provide actionable insights into viewer engagement and content strategy. This will facilitate content creators and marketers in making informed decisions to enhance viewer satisfaction and maximize engagement across various demographics and regions. Our analyses span various dimensions such as genre popularity, content trends over the years, and the effectiveness of different content types across diverse markets, providing a holistic view of the streaming landscape.

Data Description

- Categorical Variables: show_id, type, title, director, cast, country, description, listed_in

- Continuous Variables: date_added, release_year, duration.

```
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
```

Exploratory Data Analysis (EDA)

Introduction to EDA:

In this phase, we applied various statistical and visual techniques to understand the underlying patterns of the dataset. EDA is crucial as it allows us to see trends, patterns, and outliers before applying any machine learning or predictive techniques.

Key Steps Undertaken:

1.Data Visualization:

2.Analysis of Content Distribution:

3.Trend Over Years:

4.Duration Analysis:

5.Genre Popularity and Demographics:

Key Findings:

1. A significant increase in content production over the last decade, highlighting Netflix's expansion strategy.
2. A diverse range of genres with specific genres peaking in popularity in certain age groups, indicating targeted content strategies.
3. The average duration of movies has shown slight variations, suggesting a stable market expectation in movie length.

Methodology

A. Data Cleaning and Preparation

1. Handling Missing Values:

- Identified missing values across different columns such as 'director', 'cast', and 'country'.
- Applied appropriate strategies like filling missing values with placeholder or the most frequent value, or sometimes dropping rows where essential data was missing.

```
# Fill missing values or drop rows with missing data
netflix_data = netflix_data.dropna(subset=['director', 'cast', 'country', 'rating','duration','year_added'])

# Fill missing 'date_added' with a placeholder or the most frequent value
netflix_data['date_added'] = netflix_data['date_added'].fillna('Unknown')

# Check again for missing values
print(netflix_data.isnull().sum())
```

Before

```
show_id            0
type               0
title              0
director        2634
cast             825
country          831
date_added        10
release_year       0
rating             4
duration           3
listed_in          0
description        0
dtype: int64
```

After

```
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
description     0
year_added      0
age_group       0
dtype: int64
```

2. Correcting Data Formats:

Ensured all data types were correct for analysis.

Normalized formats for categorical data to eliminate variations caused by typos or inconsistent labeling.

Normalized Data formats

```
show_id                 object
type                    object
title                   object
director                object
cast                    object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration               float64
listed_in               object
description             object
dtype: object
```

3. Removing Duplicates:

- Checked for and removed any duplicate entries to prevent skewed analysis results.

```
Number of duplicate rows: 0
Data after removing duplicates:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
None
```

Removed duplicate rows

B. Tools Used:

•**Python:** Primary programming language.

•**Libraries:**

- **Pandas:** For data manipulation and cleaning.
- **Matplotlib/Seaborn:** For creating initial visualizations to identify outliers and errors.

C. Analysis Techniques:

Descriptive statistics to understand central tendencies and dispersions.

Visualization techniques to detect outliers and patterns in the data.

Linear Regression quantified trends like the relationship between release year and content duration, providing insights into linear correlations.

Random Forest identified key predictors of viewer engagement and content popularity, improving accuracy through multiple decision trees.

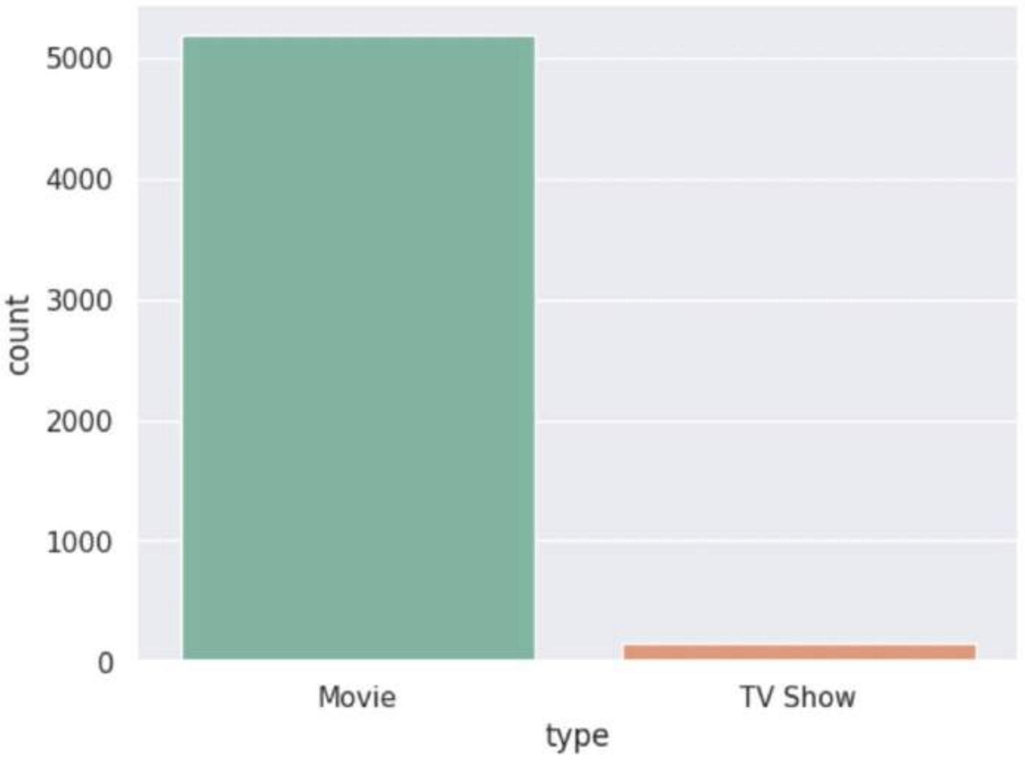
Time Series Analysis predict the future Trend

Visualisation Results

Number of Movies vs TV Shows

This chart shows that Netflix offers significantly more movies, around 4,500 titles, compared to approximately 500 TV shows.

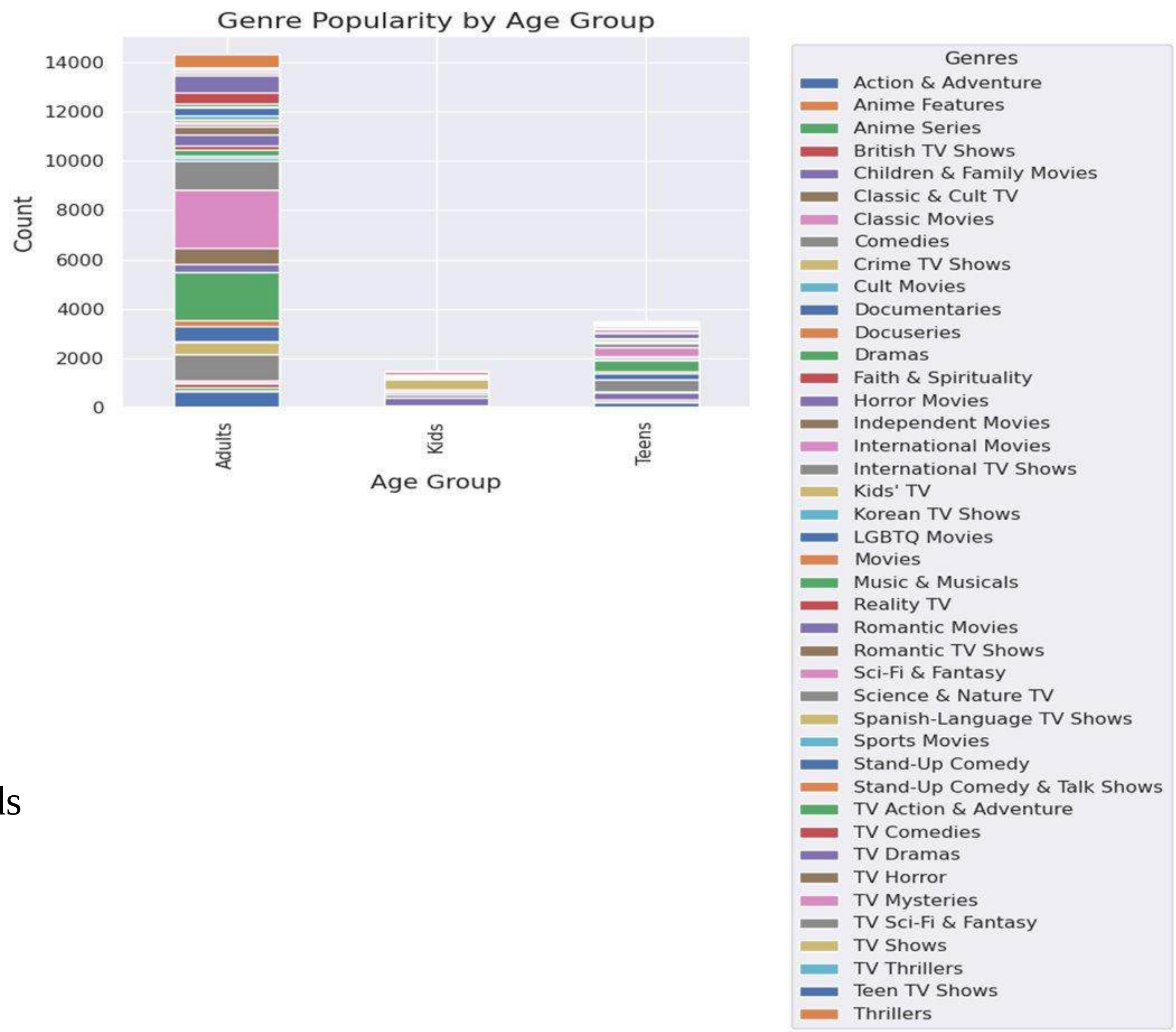
The predominance of movies highlights Netflix's strategy to cater to a wide variety of film preferences



Q. Popular genre among different age group

•This bar chart is depicting the popularity of different genres by age group.

•The legend on the right side lists various genres.



Insights:

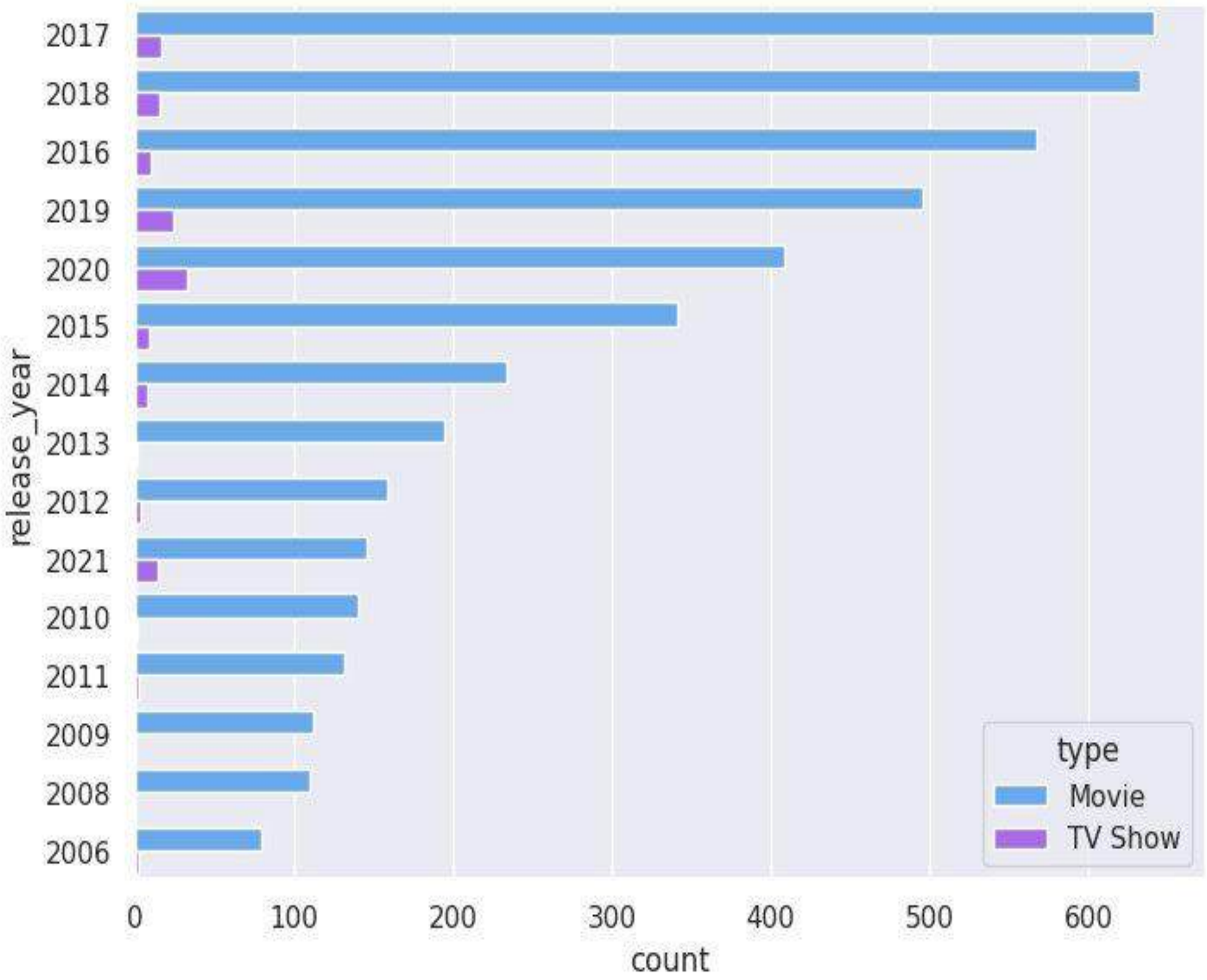
•Adults: The highest number of genres are consumed by adults, with a significant count for each genre.

•Kids: Fewer genres are popular compared to adults.

•Teens: More genres are popular among teens than kids but fewer than adults.

Yearly Analysis of Content

•This is a bar chart showing the count of movies and TV shows released each year.
•The legend at the bottom right corner indicates the type of
content such as Movies or TV Show.

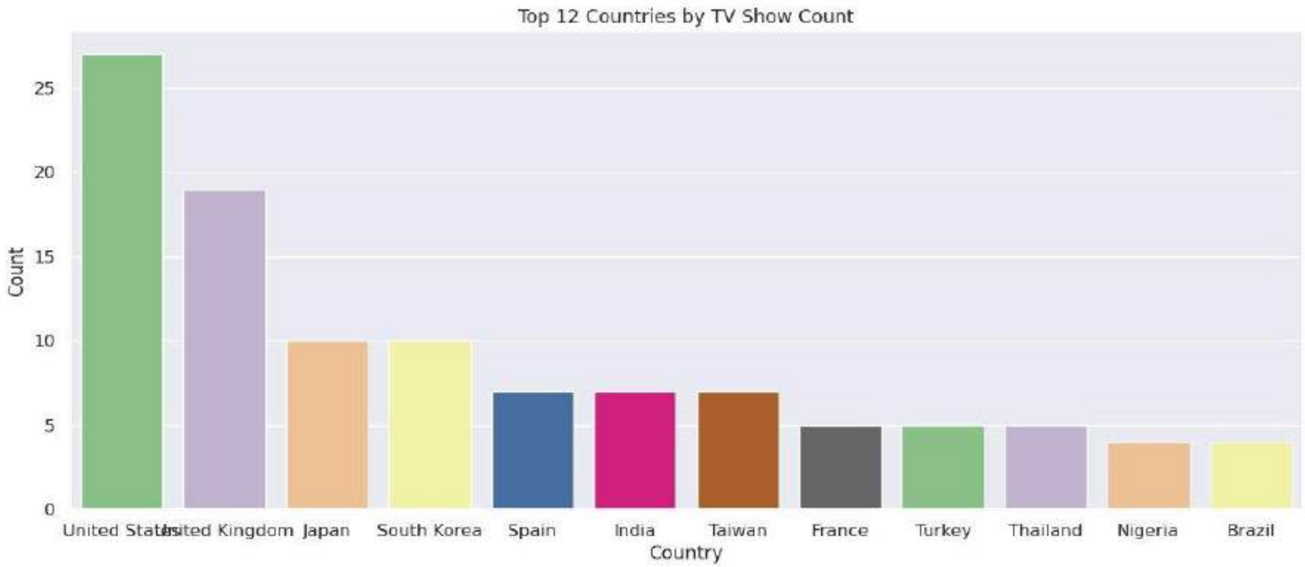


Insights:

- This is a bar chart showing the count of movies and TV shows released each year.
- The legend at the bottom right corner indicates the type of
content such as Movies or TV Show.

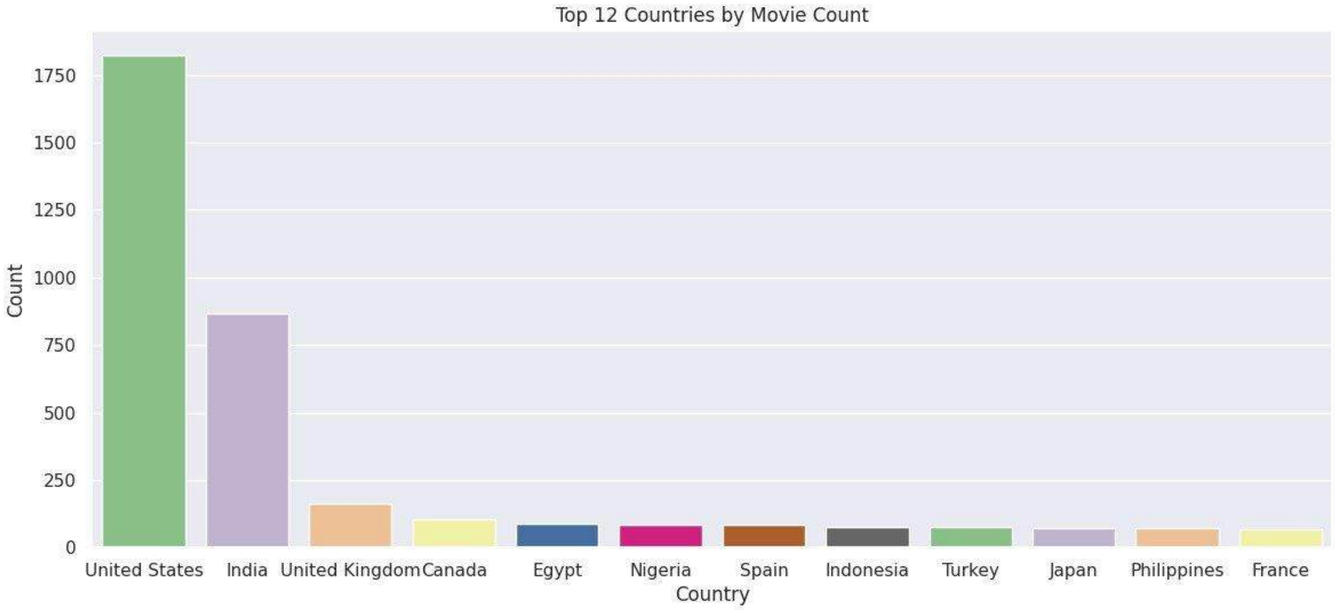
Q. Which countries produce the most content consumed internationally?

•These bar chart shows us the number of movies and TV shows produced in a specific country.
•Each bar represents Movies and TV shows produced in a specific country.



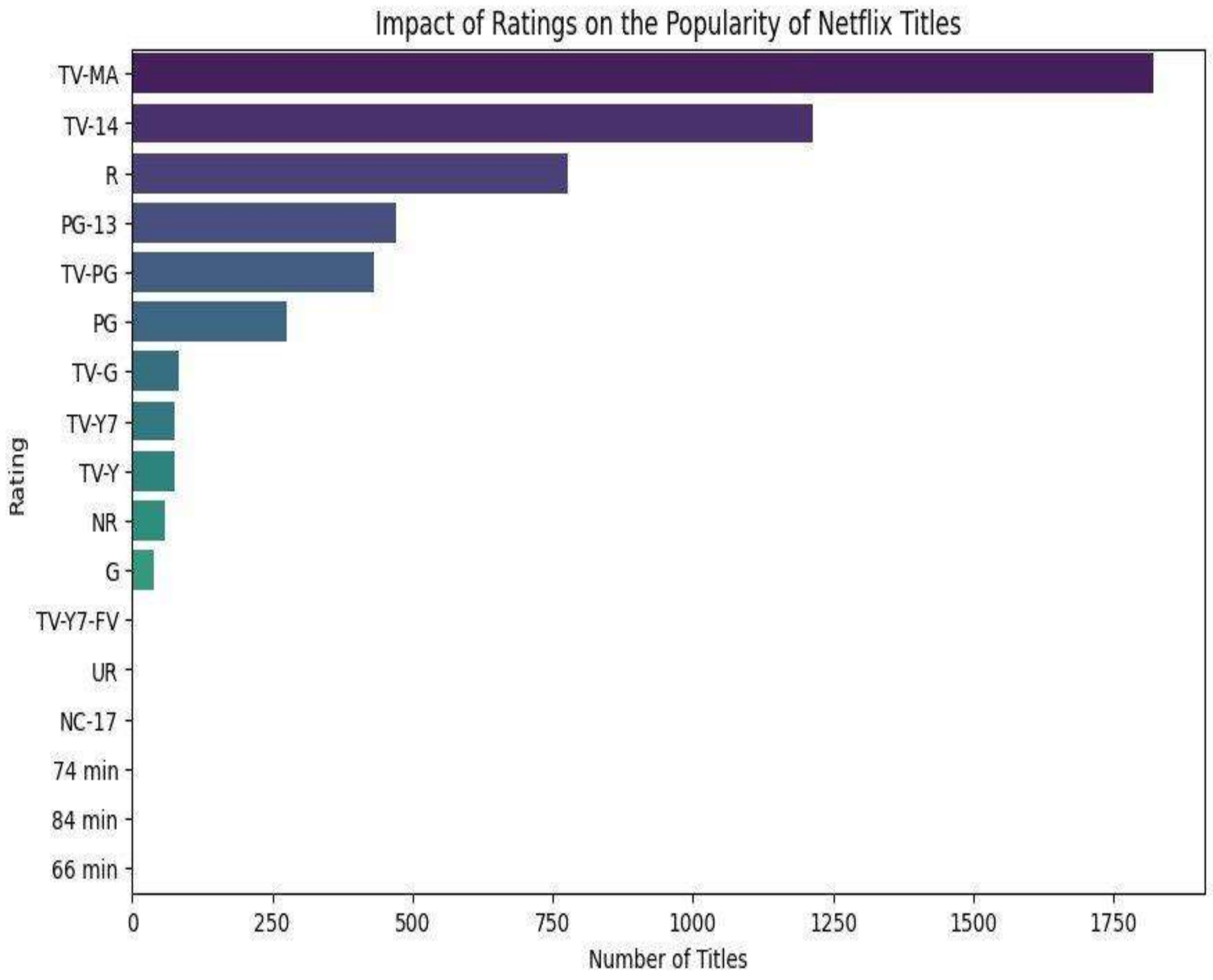
Insights
•Movies: The United States has the highest count of movies by a significant margin. India, the United Kingdom, and Canada follow.

•TV Shows: The United States has the highest count of TV shows. The United Kingdom, Japan, and South Korea follow.



Impact of Ratings on the Popularity of Netflix Titles

- This bar chart implies us the impact of rating on the popularity of titles.
- Each bar represents the count of titles for specific rating.



Insights:

- TV-MA (Mature Audience) has the highest number of titles.
- TV-14 and R ratings follow.
- Ratings like PG-13, TV-PG, PG, and TV-G have a moderate number of titles.
- Other ratings like TV-Y7, TV-Y, NR, G, TV-Y7-FV, UR, NC-17, and some duration labels (74 min, 84 min, 66 min) have fewer titles.

How does the length of movies and series correlate with viewer engagement

•This scatter plot shows the correlation between Length of Movies and Viewer Engagement.

•Each point represents a movie, with color indicating its rating.

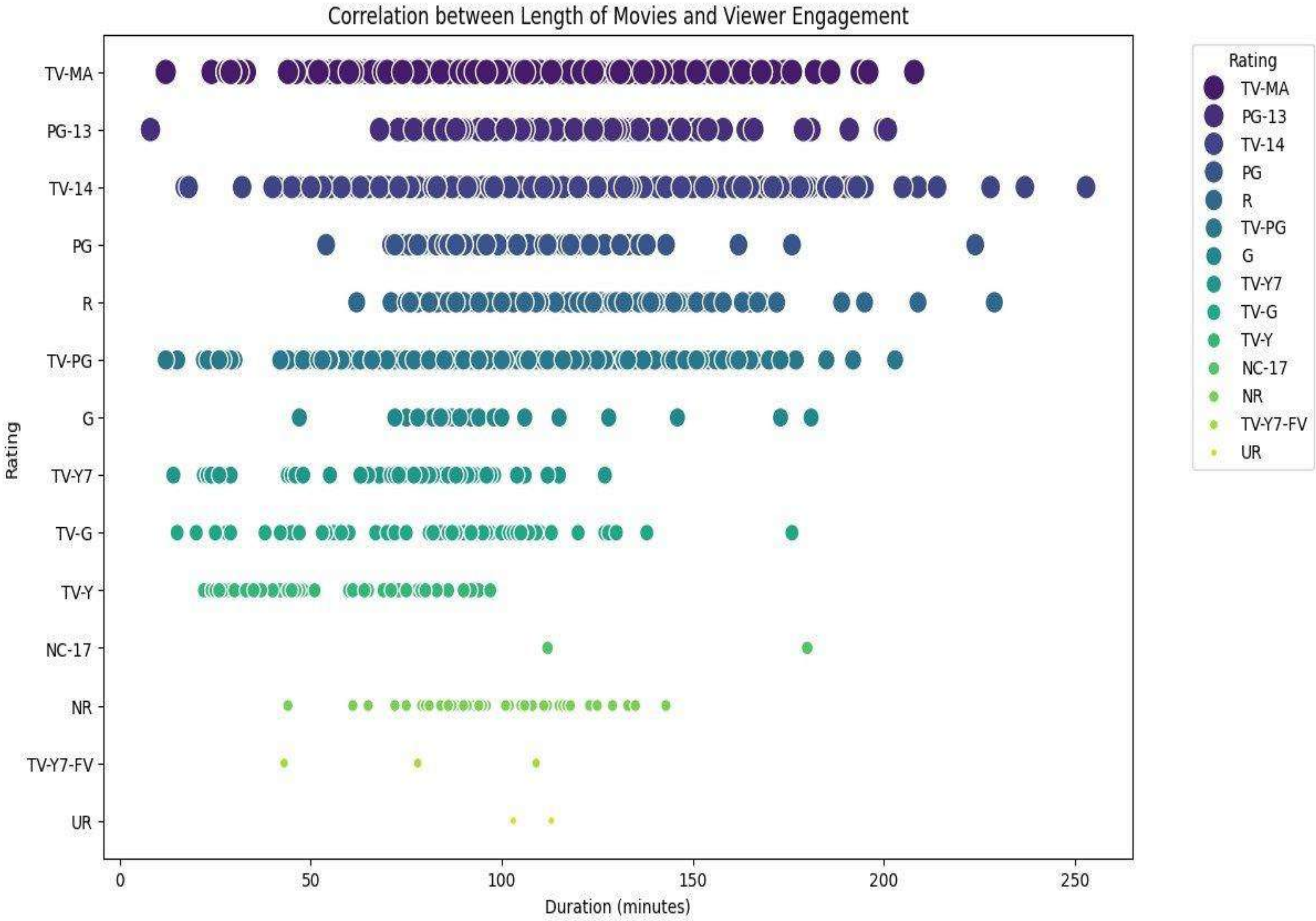
Insights

•There is a spread of movie durations across all ratings.

•TV-MA and TV-14 rated movies appear frequently across a range of durations.

•G, TV-Y, and other kid-friendly ratings tend to have shorter durations.

•The plot helps in identifying how movie duration varies with different ratings.



Movies and TV Shows Rating Analysis

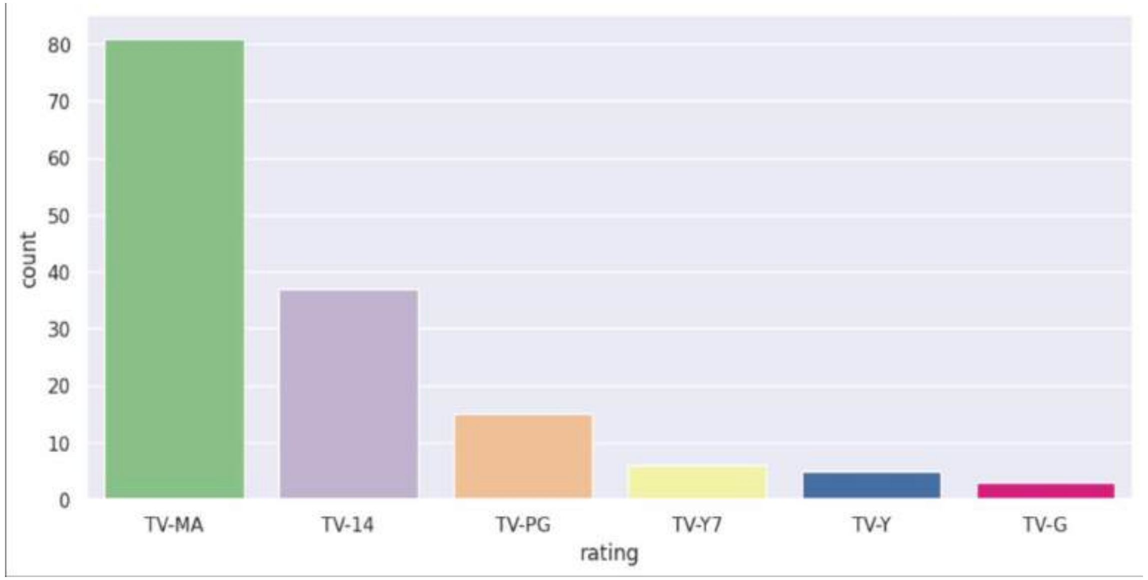
•This bar chart show the distribution of ratings for Movies and TV shows.

•Each bar represents the count of titles for a specific rating.

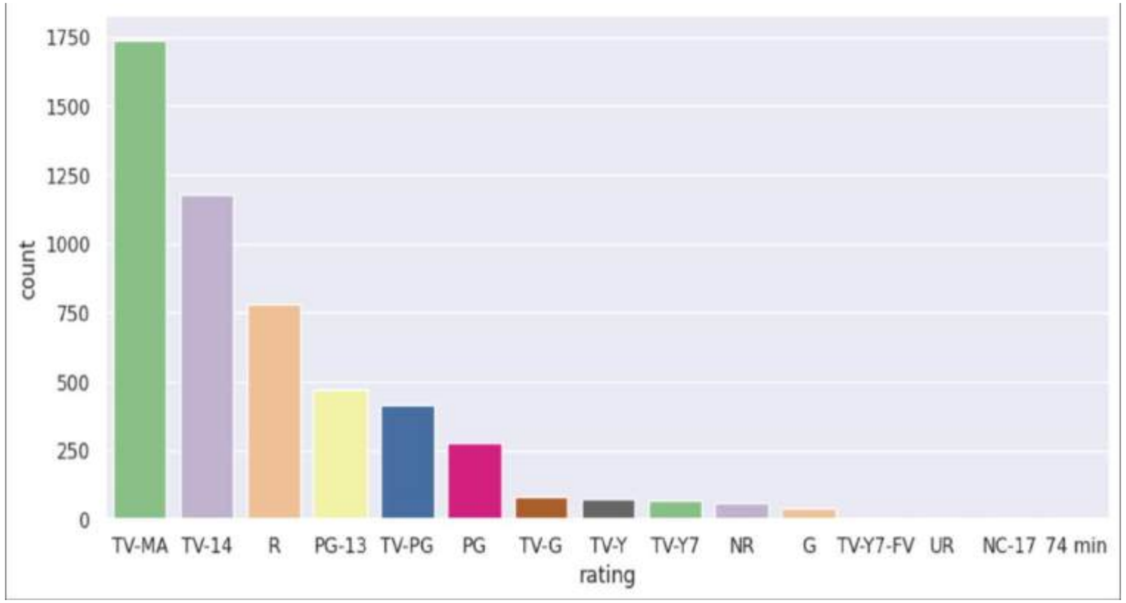
Insights

•Movies: TV-MA has the highest count. Other ratings like TV-14, TV-PG, TV-Y7, TV-Y, and TV-G follow.

•TV shows: TV-MA has the highest count, similar to the top chart. TV-14, R, PG-13, and TV-PG have significant counts as well. Other ratings have fewer titles, indicating lower prevalence.

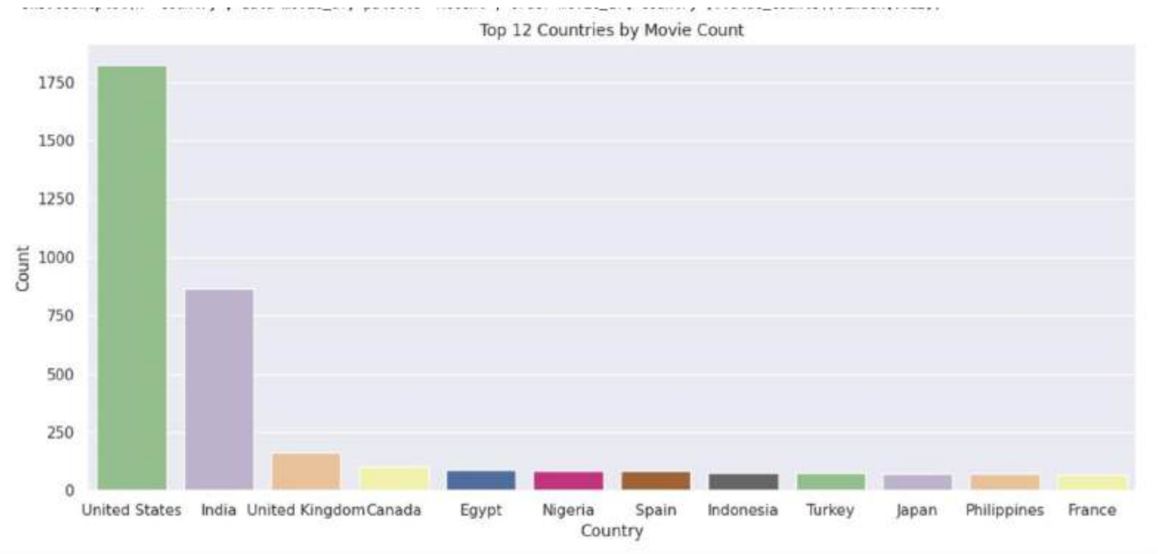


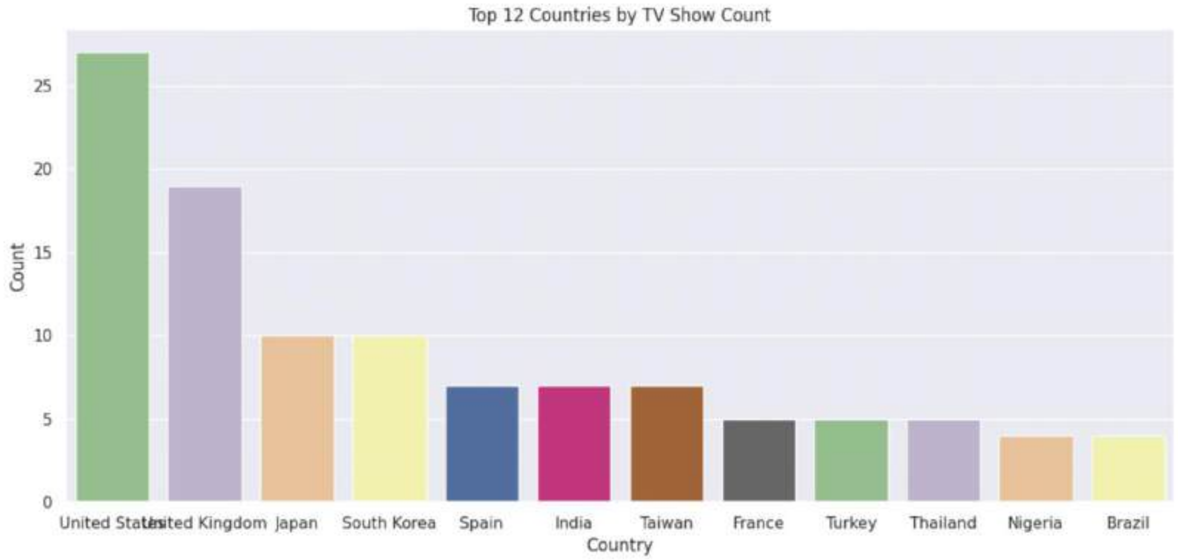
Movies rating



TV shows rating

Movie and TV show counts across the top 12 countries.

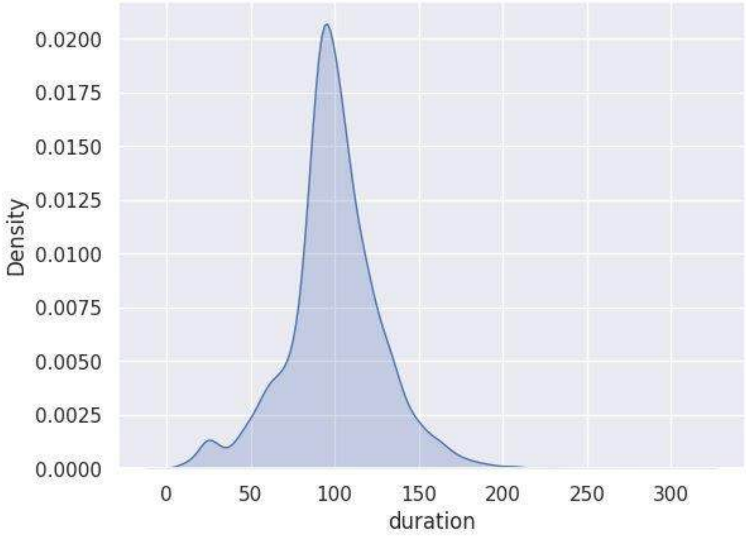


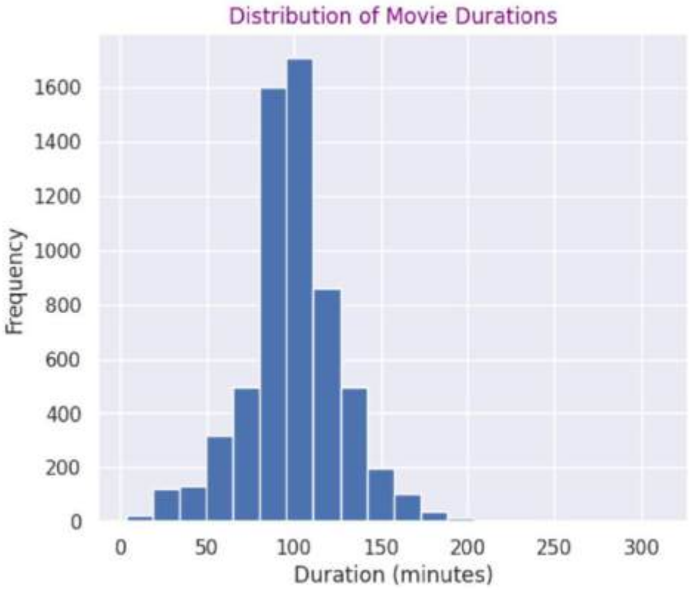


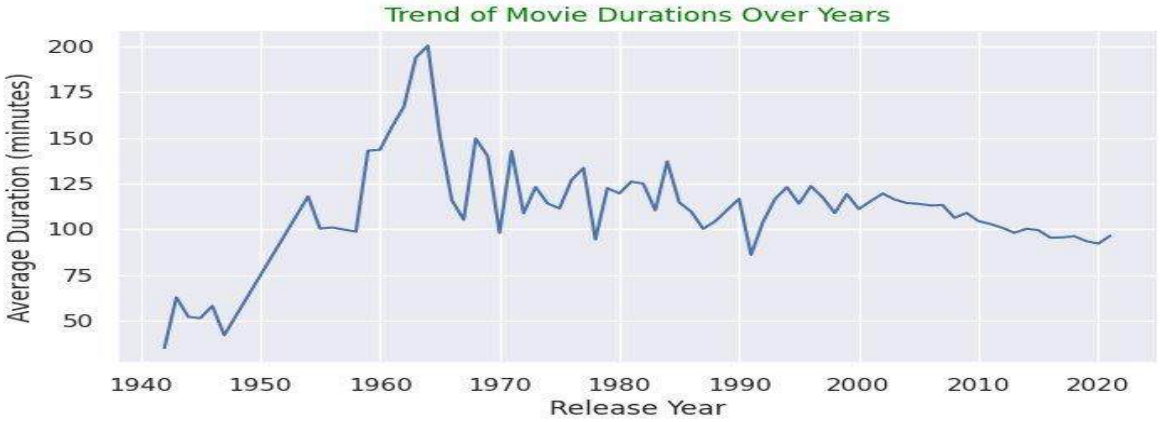
Insights:

- **Dominance in Content Production:** The United States dominates both charts, highlighting its significant role in producing content for Netflix, particularly in movies.
- **Diverse Global Representation:** While the U.S. leads, there's notable content production from countries like India and the United Kingdom, indicating Netflix's diverse global content strategy.

Analysis of Movie Duration

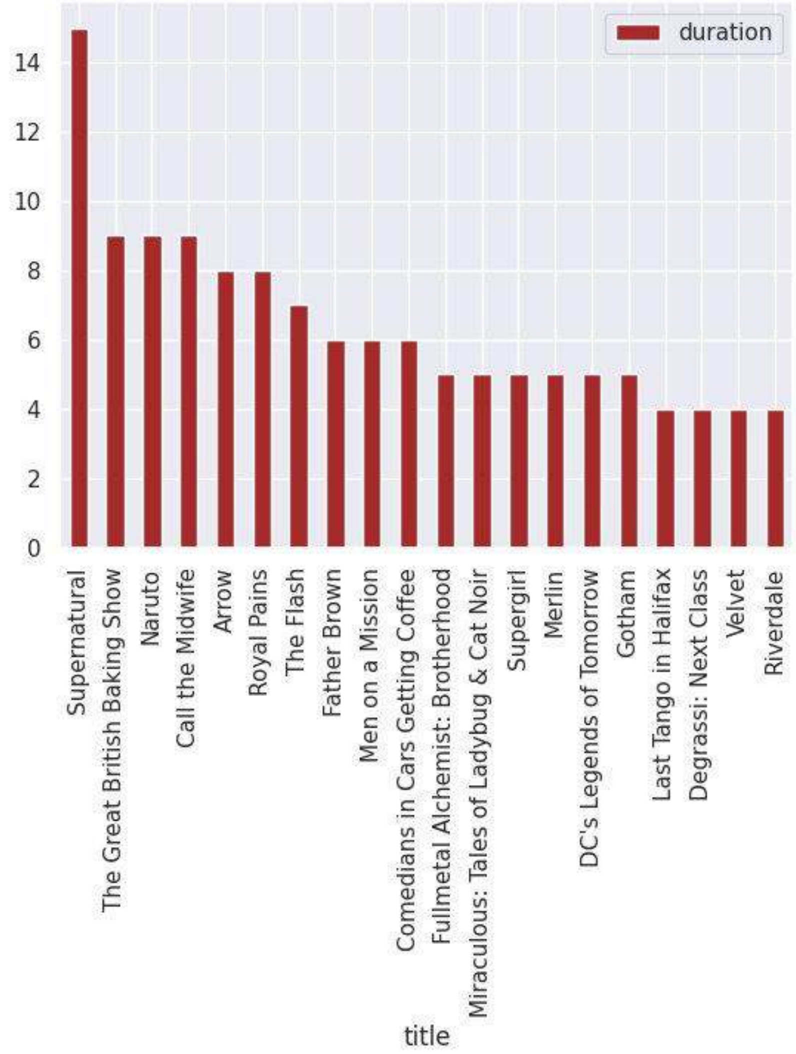


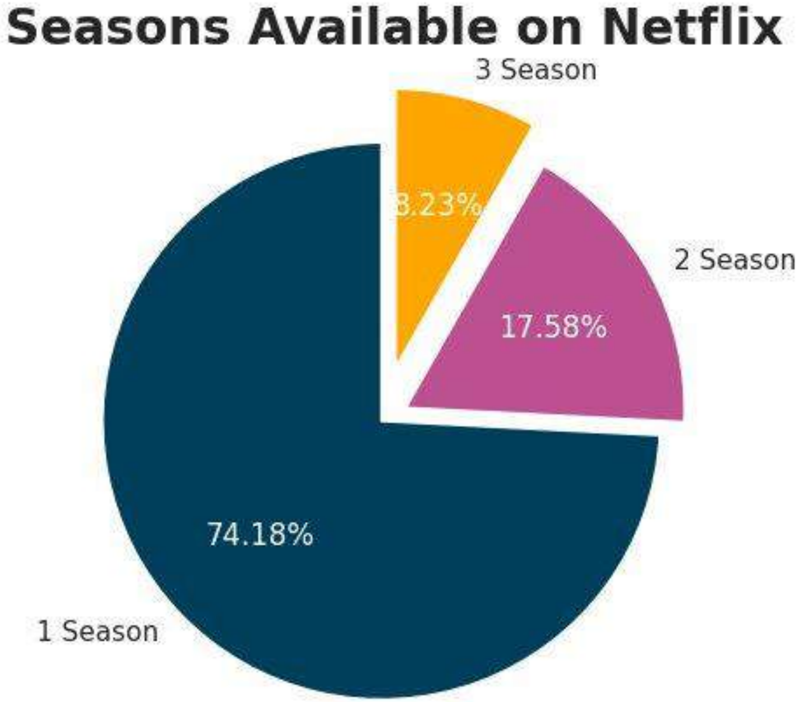




count	6128.000000
mean	99.577187
std	28.290593
min	3.000000
25%	87.000000
50%	98.000000
75%	114.000000
max	312.000000

Analysis of TV Shows with the most number of seasons

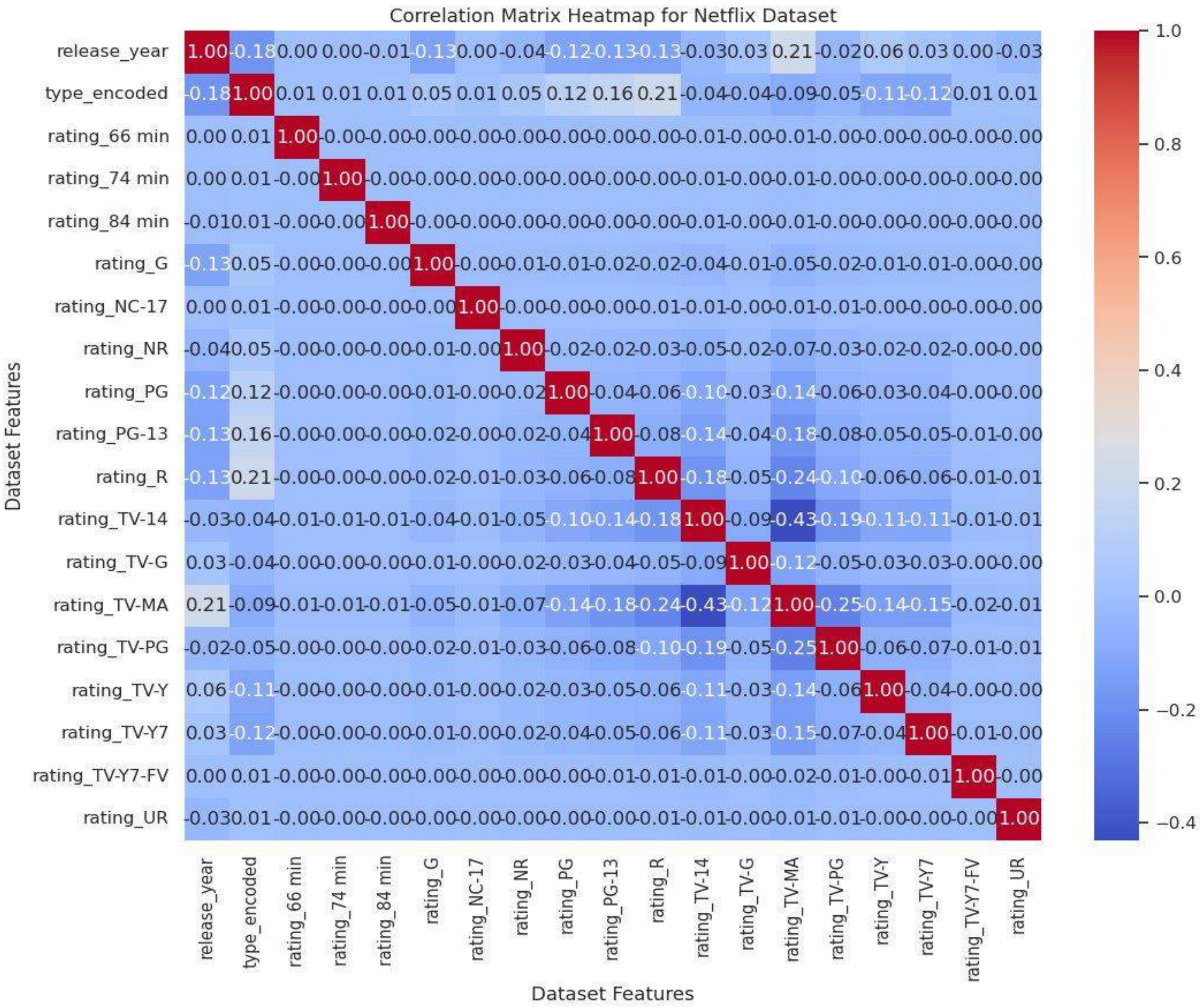




Correlation Matrix Heatmap for Netflix Dataset

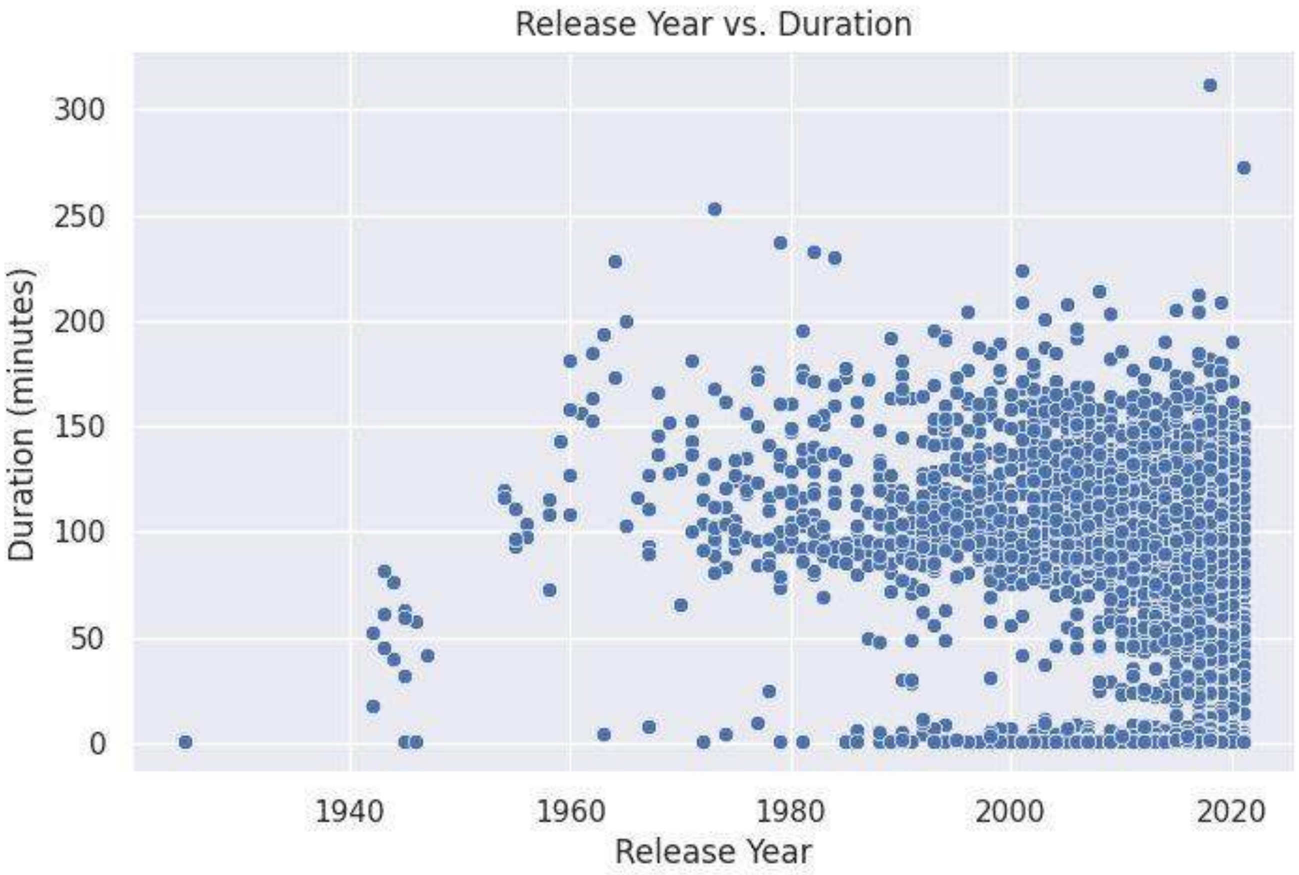
•This heat map shows us the correlation matrix for Netflix.
•Both axes list dataset features (e.g., release_year, type_encoded, different ratings). The heatmap colors indicate the correlation strength between features

Insights:
•Strong positive correlations can be observed along the diagonal, as a feature is perfectly correlated with itself.
•Other correlations help in understanding relationships between features like ratings, release year, and type.



Q. How has the duration of Netflix titles changed over the years, and are there any notable trends or patterns in movie or TV show durations?

•This scatter plot
shows us the comparison between release year
 and duration.
•Each point represents a movie, showing its
release year and duration.

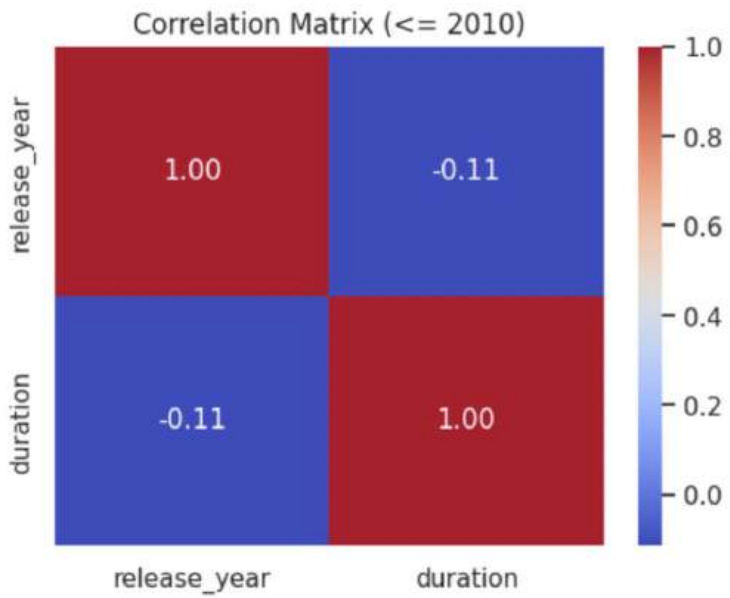


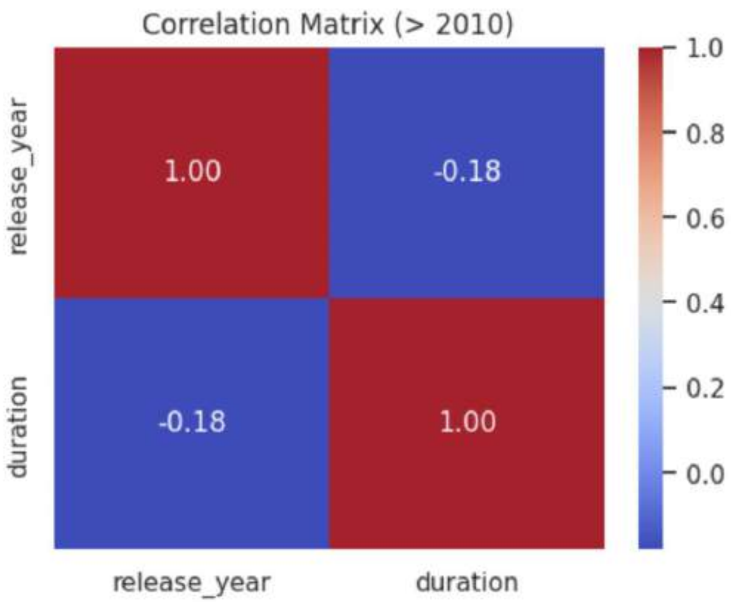
Insights

•Older movies tend to have shorter durations.
•As the years progress, the duration of movies increases, with a larger spread in recent years.
•This plot helps in understanding trends in movie durations over time.

How correlations change over time by segmenting the data into different time periods.

Display the relationship between release year and duration of content on Netflix, segmented into two time periods: before 2010 and from 2010 onwards.



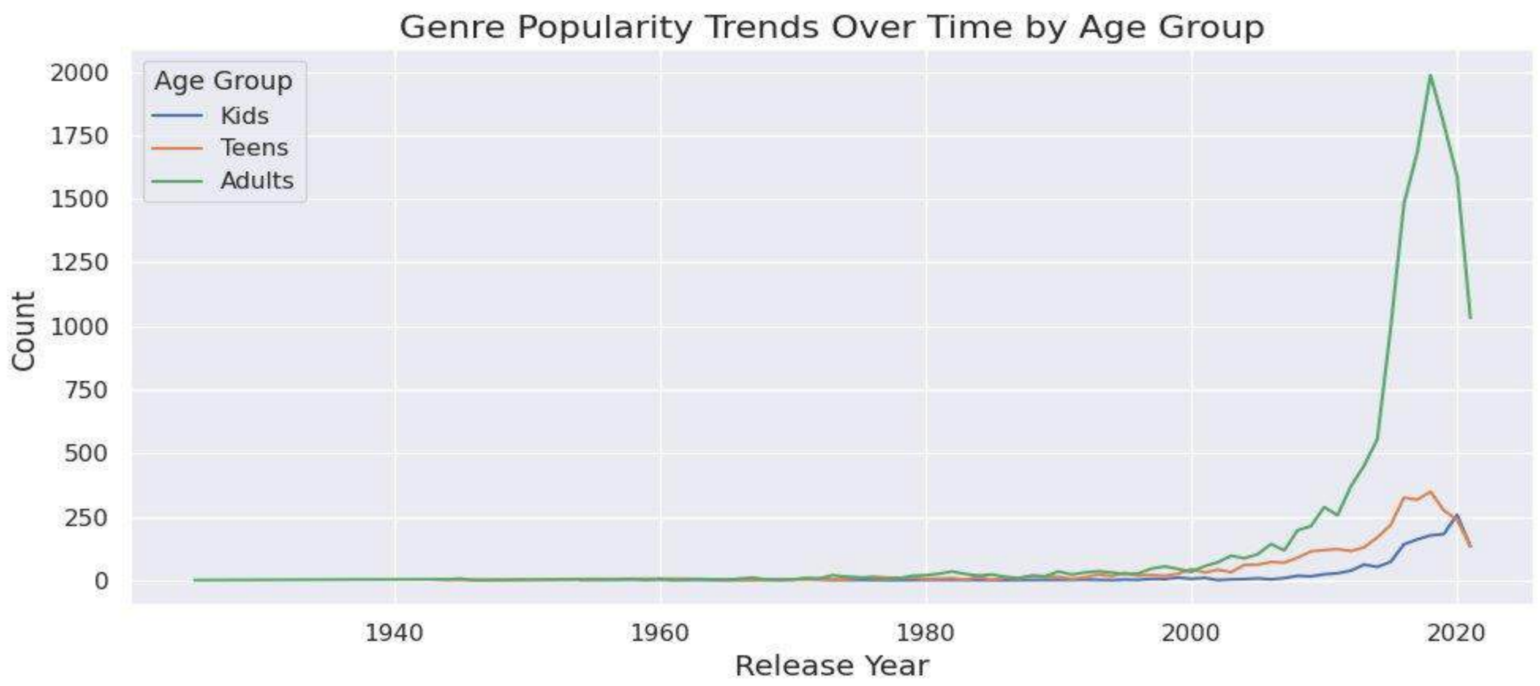


Insights:

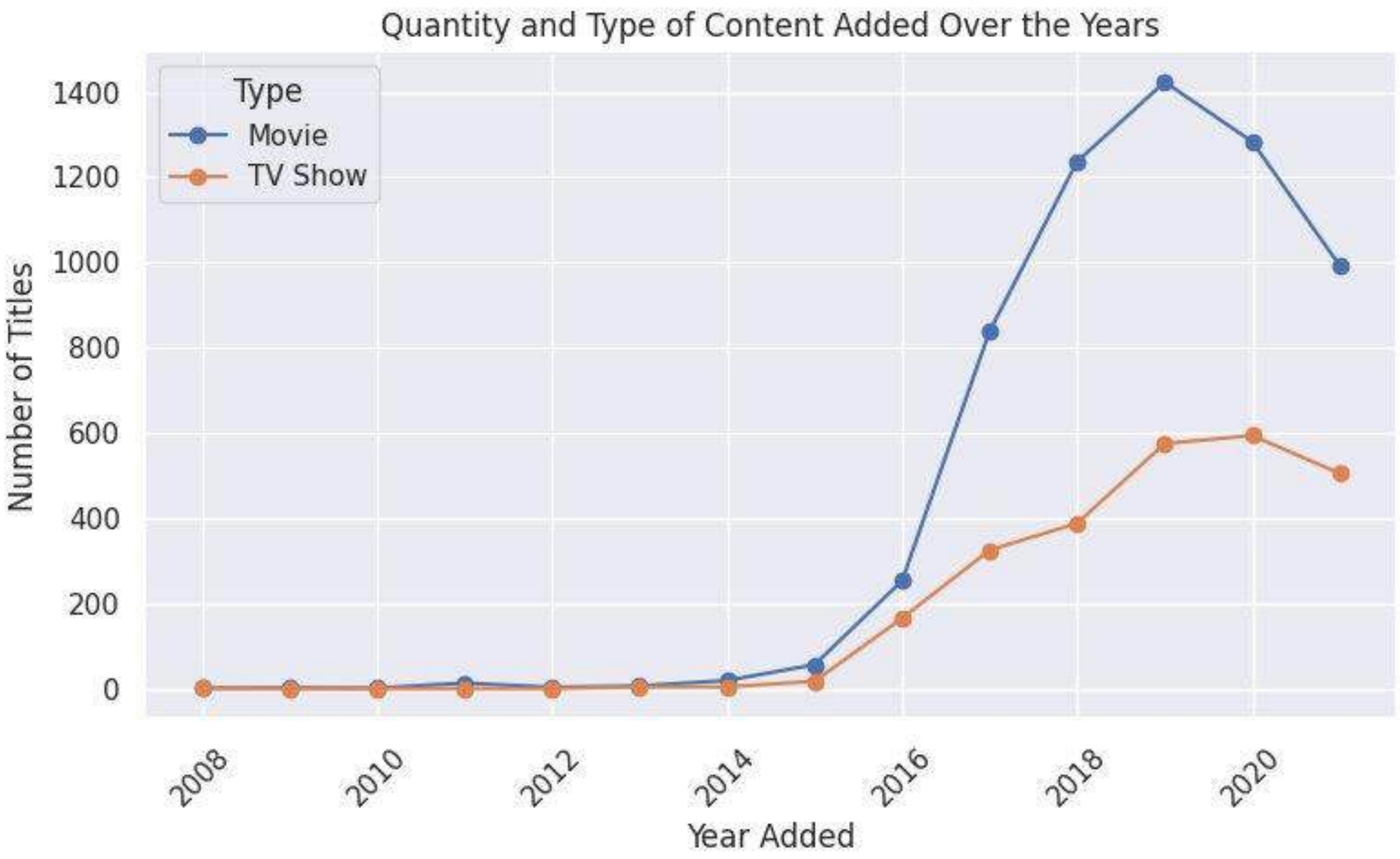
- **Increasing Correlation Strength Post-2010:** The correlation coefficient decreases from -0.11 to -0.18 when comparing before and after 2010. This indicates that the trend towards shorter content has become more pronounced in recent years.

Q. Can we predict the popularity of a genre based on historical trends?

- This line graph illustrates the historical trends in genre popularity across different age groups on Netflix, showing a significant increase in content tailored for adults in recent years.
- The sharp rise in adult content, particularly after 2010, suggests that adult genres have become increasingly popular.



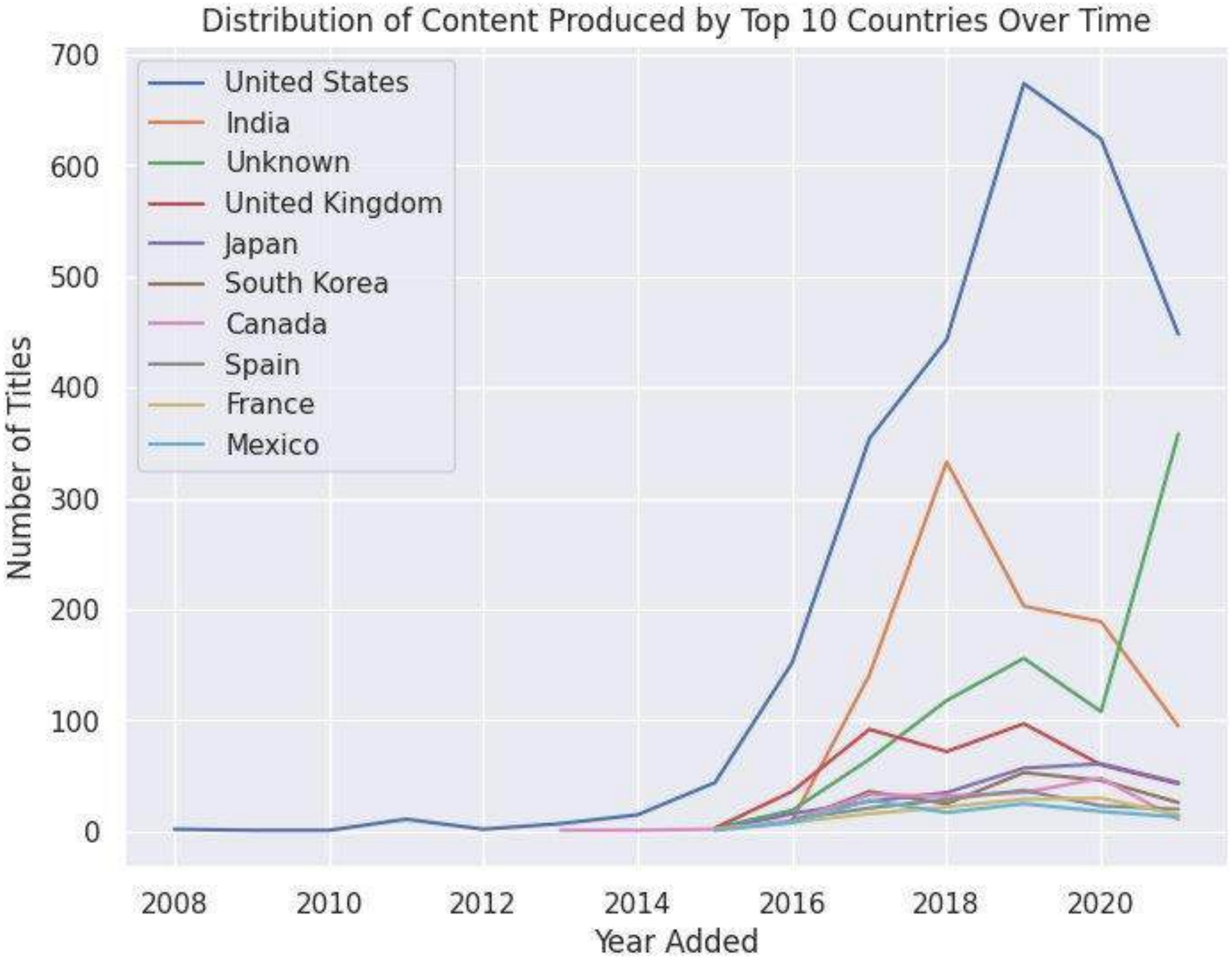
Q. How has the quantity and type of content changed over the years?



- This line graph shows the dramatic increase in both movies and TV shows added to Netflix from 2008 to 2020.
- While the addition of movies has seen a substantial peak around 2018 before slightly declining, TV shows have experienced a steadier, more moderate growth.

Q. What is the distribution of content produced by different countries, and how has this changed over time?

- This line graph illustrates the distribution and growth of Netflix content production across the top 10 contributing countries from 2008 to 2020.
- The United States has consistently led in content production, with a notable surge starting around 2014, while countries like India and South Korea show significant increases in contributions in the later years.



Predictive Modeling and Forecasting

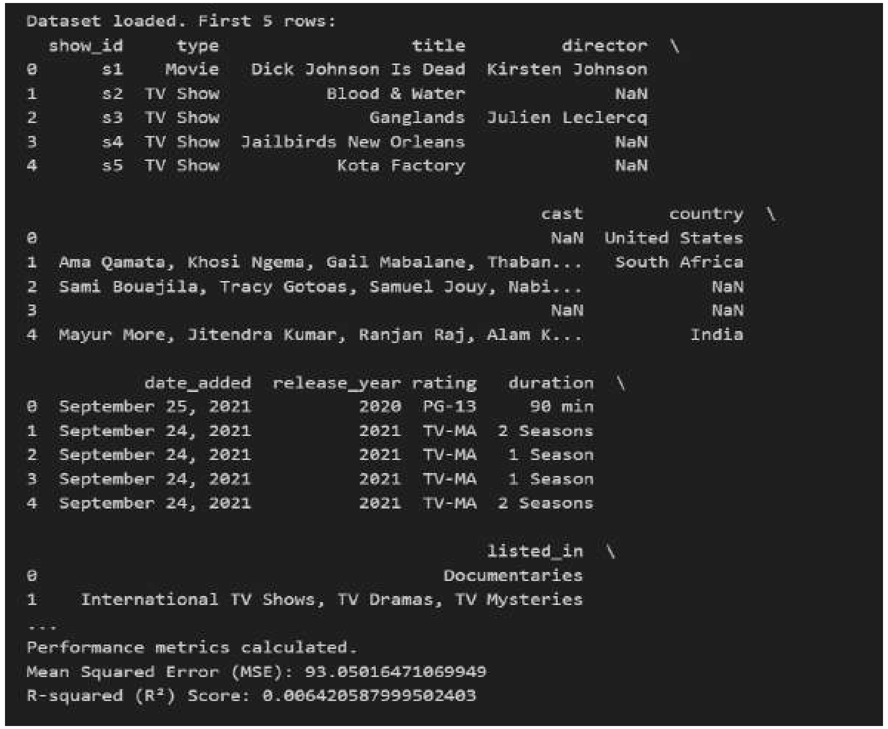
We used techniques that collectively enhanced our analytical capabilities, offering precise and predictive insights into the streaming content landscape. These methods provided a comprehensive analysis of how content characteristics have evolved over time on Netflix.

Linear Regression: To determine straightforward linear relationships and trends, allowing us to assess how content duration impacts release timelines.

Random Forest: to handle more complex patterns and interactions within the data, offering a more nuanced understanding and higher accuracy by leveraging an ensemble of decision trees.

Time Series Analysis (Forecast number of title added): We used Time Series Analysis with seasonal decomposition and ARIMA modeling to dissect and predict monthly patterns in Netflix's content additions. This method clarifies trends and seasonal fluctuations, enabling accurate future trend forecasting and strategic content planning.

Linear regression



```
Dataset loaded. First 5 rows:
  show_id     type                  title         director  \
0      s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show          Blood & Water              NaN
2      s3  TV Show              Ganglands  Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans              NaN
4      s5  TV Show           Kota Factory              NaN

                                                cast        country  \
0                                                NaN  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                NaN            NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India

           date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min
1  September 24, 2021          2021  TV-MA  2 Seasons
2  September 24, 2021          2021  TV-MA   1 Season
3  September 24, 2021          2021  TV-MA   1 Season
4  September 24, 2021          2021  TV-MA  2 Seasons

                                             listed_in  \
0                                        Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
...
Performance metrics calculated.
Mean Squared Error (MSE): 93.05016471069949
R-squared (R²) Score: 0.006420587999502403
```

Variables Used:

- **Feature (X): Categorical variables** such as type (e.g., movie, TV show), rating (e.g., PG, PG-13), and genres (action, comedy, etc.) are transformed using one-hot encoding.
- **Numerical variables** such as the release year.

•**Target (y):** Duration

Model Training:

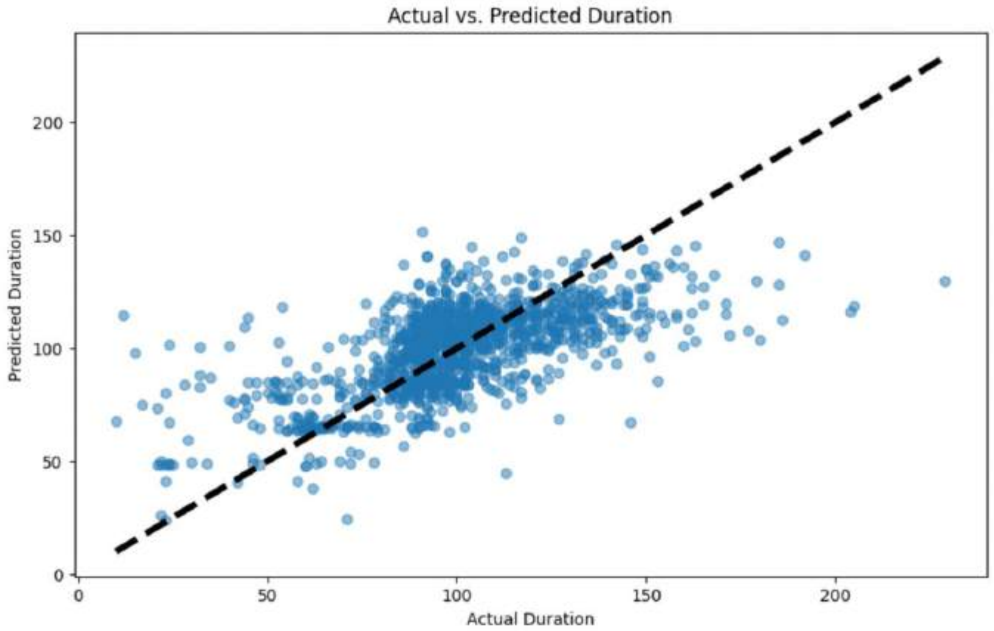
•The data was split into training (80%) and testing (20%) sets.

•A Linear Regression model was trained using duration_numeric as the predictor and release_year as the response variable.

Performance Metrics:

- **Mean Squared Error (MSE) of 423.75**: This indicates the average of the squares of the errors—that is, the average squared difference between the actual and predicted durations. A lower MSE would indicate a better fit.

- **R² Score of 0.4014**: This score tells you that approximately 40.14% of the variance in the duration of Netflix titles is explained by your model. The closer this value is to 1, the better the explanatory power of the model.



Regression Equation:

$$\boxed{\text{Release\_year} = (-0.015 \times \text{duration\_in\_minutes}) + 2015.74}$$

Where,
Independent Variable (x): Duration in minutes (duration_numeric). This is the variable you manipulate or change to observe how it affects the release year.
Dependent Variable (y): Release year of the titles (release_year). This is the variable you are trying to predict using the duration.

Insights:

The linear regression analysis indicates that while there is a slight tendency for newer Netflix titles to be shorter, the duration of titles alone does not strongly predict their release years. This might suggest the influence of other factors not included in the model, such as genre, type of content (movie vs. TV show), or production considerations that could better explain the trends in release years. Given the weak relationship indicated by the linear model, exploring more complex models or additional predictors could provide deeper insights.

Random Forest

Variables Used:

•**Feature (X):** 'type', 'rating', 'listed_in',' 'release_year'.

•**Target (y):** Duration

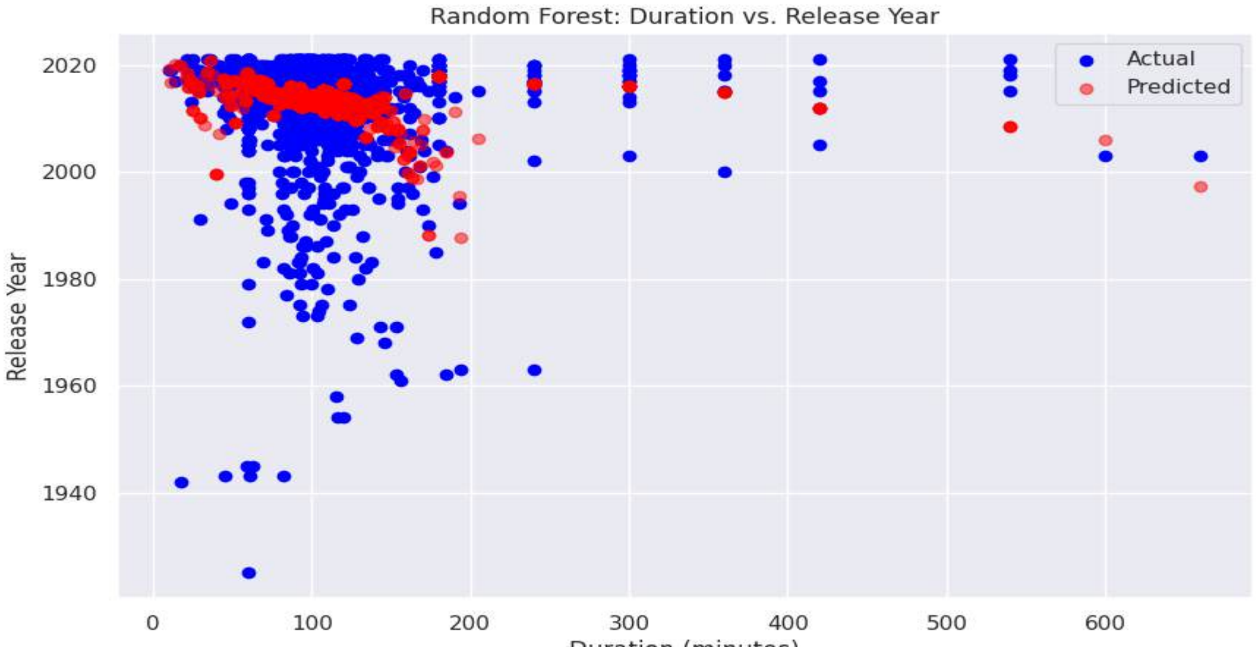
```
Dataset loaded. First 5 rows:
  show_id     type                 title         director  \
0      s1    Movie  Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show         Blood & Water              NaN
2      s3  TV Show             Ganglands  Julien Leclercq
3      s4  TV Show Jailbirds New Orleans              NaN
4      s5  TV Show          Kota Factory              NaN

                                                cast        country  \
0                                                NaN  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                NaN            NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India

          date_added  release_year rating   duration  \
0 September 25, 2021          2020  PG-13     90 min
1 September 24, 2021          2021  TV-MA  2 Seasons
2 September 24, 2021          2021  TV-MA   1 Season
3 September 24, 2021          2021  TV-MA   1 Season
4 September 24, 2021          2021  TV-MA  2 Seasons

                                      listed_in  \
0                                 Documentaries
1   International TV Shows, TV Dramas, TV Mysteries
...
R-squared (R²) Score: 0.032018196134712906
Mean Absolute Error (MAE): 5.326721829732138
```

Model Training:

•**Data Split:** The dataset was divided into training (80%) and testing (20%) sets.

•**Random Forest Model:** Trained using duration as the predictor with the following settings:

- **Number of Trees:** 100
- **Maximum Depth:** None (allowing trees to grow until all leaves are pure or until all leaves contain less than min_samples_split samples)
- **Number of Features:** 1.0 (using all available features for splitting at each node)



Model Performance:
•**Mean Squared Error (MSE):** 90.65, indicating the model's average squared error in predicting the release year.

•**R-squared (R^2):** 0.032, showing the model explains about 3.2% of the variance in release years based on durations—a slight improvement over the linear model but still low.

•**Mean Absolute Error (MAE):** 5.33, representing the average absolute difference between predicted and actual release years.

Insights from Random Forest Regression:
•**Improved Fit Over Linear Model:** Despite the low R^2 value, the random forest model shows a slight improvement over the linear regression model in fitting the data.
•**Complex Relationships:** The use of multiple decision trees allows the model to capture more complex non-linear relationships than a simple linear regression model.

Conclusion:
The Random Forest model, while more complex, only marginally improves prediction accuracy over the linear model. This suggests that duration alone may not be sufficient to predict release years accurately and that other factors could play significant roles. Further analysis with additional features or different modeling techniques might yield better predictive performance.

Time Series Analysis : Forecast Number of Title added

Data Preparation:

•Resampled data monthly to analyze trends in content addition.

Forecasting with ARIMA:

•Utilized ARIMA(1,1,1) to predict future content additions.

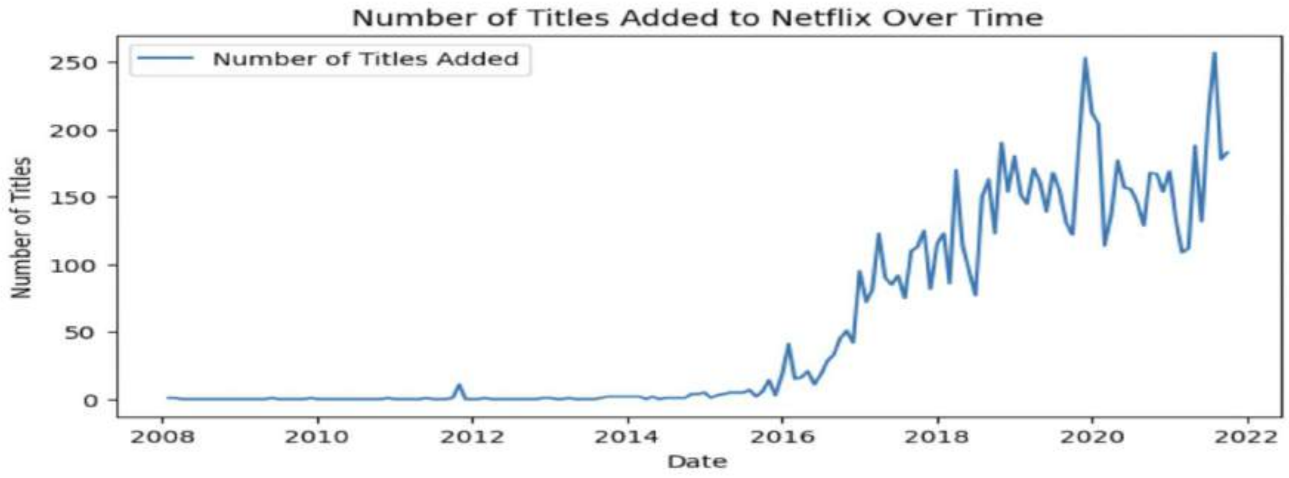
•Forecast for the next 12 months, showing expected increase and variability.

Time Series Decomposition:

•Decomposed into trend, seasonal, and residual components.

Key Observations:

- Increasing trend indicates Netflix's library expansion.
- Seasonal fluctuations highlight peak times for content addition.



Visual Insights:

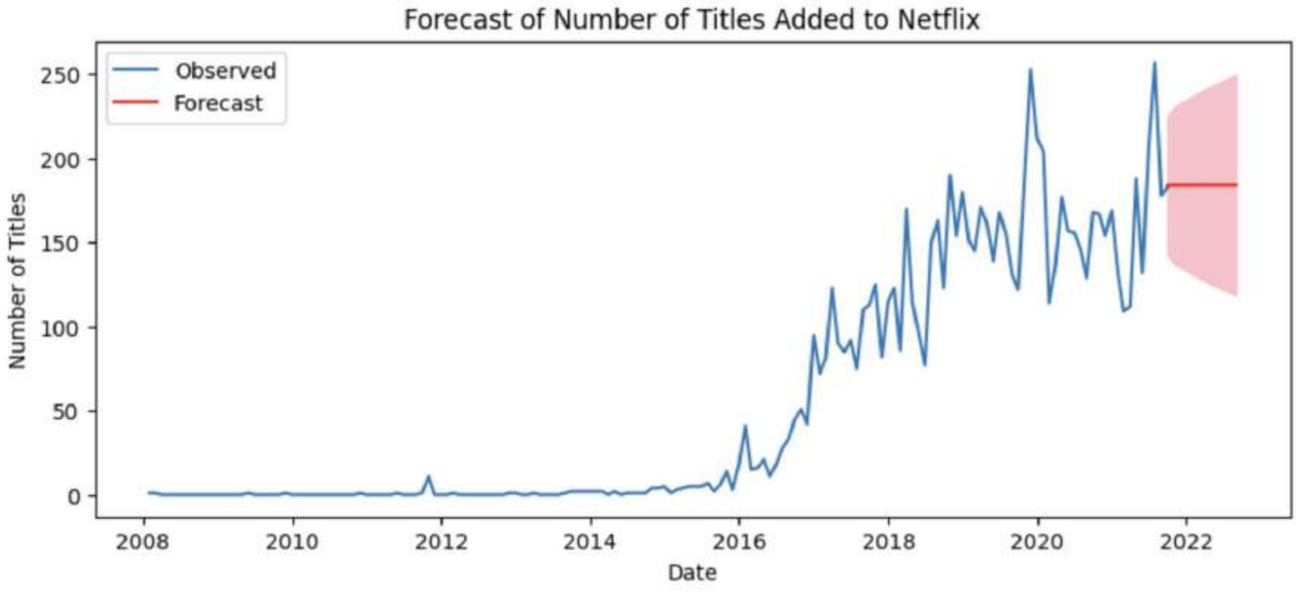
•Plots display historical data, trends, seasonal patterns, and forecasts.

•Helps in strategic planning and resource allocation for upcoming content.

Conclusion:

•Analysis reveals Netflix's growth patterns and seasonal peaks.

•Forecast aids in anticipating and planning for future content needs.



Conclusion

- **Key Insights**: The project effectively demonstrated the evolving landscape of Netflix's content strategy through advanced analytical techniques. It highlighted significant trends, including an increasing volume of content over time, diverse content strategies tailored to different demographics, and the importance of seasonality in content addition.
- **Impact on Strategic Planning**: The insights gained from the predictive models and time series analysis enable better forecasting of content trends, aiding Netflix in strategic content planning and resource allocation to meet viewer demands.

Future Scope

- **Enhanced Predictive Models**: The presentation suggests exploring more complex predictive models or incorporating additional data features to improve the accuracy and depth of insights. This could include more granular demographic data or viewer engagement metrics.
- **Broader Content Analysis**: Future analyses could expand to include a wider array of content sources or compare Netflix's strategy with other streaming platforms to identify unique competitive advantages or areas for improvement.
- **Real-time Data Utilization**: Incorporating real-time data analytics to dynamically adjust content strategies based on current viewer preferences and global market trends.