EXPLORATORY DATA ANALYSIS

Project by
KUSHAGR SAXENA

ABOUT NETFLIX

Netflix is an American media streaming OTT platform that operates in nearly every country. Launched on 29th August 1997, it was one of the pioneers in streaming industry, transitioning in 2007. With hundreds of millions of subscribers worldwide, Netflix offers best-in-class TV series, documentaries, feature films, and games.

INTRODUCTION

This dataset is sourced from Skill Circle and contains data collected from Netflix. After a quick view of the Dataset, it looks like a typical Movie/TV shows data frame. We can see that there are NaN values present in some columns as well.

It contains 8807 unique TV Shows and Movies.
This Dataset is widely used by beginners to learn EDA.

PURPOSE OF THE PROJECT:

The goal of the Netflix EDA project is to conduct a comprehensive exploration and analysis of Netflix's content dataset. This includes understanding the data structure, ensuring data integrity by handling missing values and duplicates, deriving descriptive statistics, and visualizing content distribution across genres and release years. Additionally, the project aims to identify temporal trends, analyze content attributes like ratings and duration, and assess audience engagement metrics. By synthesizing these insights, the project aims to draw meaningful conclusions and provide actionable recommendations to enhance Netflix's content offerings and user experience.

Description of Dataset:

- I have conducted my work using Google Colab Notebook.
- The dataset has been imported from Google Drive.
- As we begin our Exploratory Data Analysis (EDA), I've named the dataset ‘df’. ▪
- The dataset comprises of 8807 Rows and 12 Columns.
- For data cleaning, I have utilized libraries like Pandas, Seaborn & Plotly.
- Any duplicate entries that were found have also been removed.

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
```

```
[ ] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
[ ] data = '/content/drive/MyDrive/EDA Netflix.csv'
    df = pd.read_csv(data)
```

Description of Dataset:

This dataset provides a rich overview of Netflix's content library, including movies and TV shows. It features details like title, director, cast, country, release year, rating, duration, genres, and a brief description. This dataset enables a wide range of analyses, including:

- Content distribution across genres, countries, and ratings.
- Trends in content addition and popularity over time.
- Correlations between variables like duration and rating.
- Diversity of content based on unique genres and categories.
- Evolution of content characteristics over the years.

```
'''Descriptive Statistics about our dataset'''

df.describe()
```

	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

Description of Dataset:

- This dataset provides a comprehensive overview of the content available on Netflix, offering valuable insights into the platform's content strategy and viewer preferences.

- Key features include:

- Show ID: Unique identifier for each title.
- Type: Categorizes content as "Movie" or "TV Show."
- Title: The name of the movie or TV show.
- Director: Name of the director(s).
- Cast: List of actors involved.
- Country: Country of origin.
- Date Added: Date when the content was added to Netflix.
- Release Year: Year of original release.
- Rating: Content rating (e.g., TV-MA, PG-13).
- Duration: Length of movies & number of seasons for TV shows.
- Listed In: Genres and categories associated with the content.
- Description: Brief synopsis of the plot.

```
df.info()
```

```
<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
```

By exploring these aspects, we can gain a deeper understanding of Netflix's content strategy, viewer preferences, and potential areas for growth and improvement.

Data Cleaning & Pre-Processing:

- **Cast** : As 825 null values are present in the 'Cast' attribute, Filling the missing values with ‘Cast not defined' can help maintain data completeness. Also, we can leverage external databases or IMDb (Internet Movie Database) to populate missing cast information for each movie or TV show.

```
df['cast']=df['cast'].fillna('Cast not defined')
```

- **Country** : For the 'Country' attribute with 831 null values, filling the missing values with the most common country of production or ‘Country not defined' can be a feasible approach. Another strategy is to cross-refer with the title or other metadata to infer the country of production based on the content's origin or production company.

```
df['country']=df['country'].fillna('Country not defined')
```

```
[ ]  '''Total Number of Null values in our dataset'''
     df.isnull().sum().sum()

     4307
```

```
[ ]  '''Let's find Null/Missing values in our dataset(Column-wise)'''
     df.isnull().sum()

     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
```

Data Cleaning & Pre-Processing:

- **Director** : For the 'Director' attribute with 2634 null values, one approach is to fill these missing values with a placeholder such as 'Unknown' or ‘Director not defined'. This allows retaining the data records while indicating the absence of director information. Alternatively, for more accurate data, we can research and populate missing director information by referencing external sources or databases related to the movies or TV shows.

```
df['director']=df['director'].fillna('Director not defined')
```

- **Rating** : For the 'Rating' attribute with 4 null values . I replaced the values with ‘Rating not defined’ because I don’t want to assume the ratings because it will be unfair with the dataset.

```
df['rating']=df['rating'].fillna('Rating not defined')
```

- **Date Added** : As only 10 null values are present in 'Date Added' attribute, filling these missing values can be cakewalk. We can impute the missing dates by filling ‘Date not defined’.

```
df['date_added']=df['date_added'].fillna('Date not defined')
```

Data Cleaning & Pre-Processing:

- **Duration** : For the ‘Duration' attribute with 3 null values, I had replaced the values with ‘Duration not defined’ because I don’t want to assume the duration because it will be unfair with the dataset.

```
df['duration']=df['duration'].fillna('Duration not defined')
```

Checking Duplicate entries:

As per the observation, we see no duplicate entries in our dataset. However if we did have found any duplicate entries, those would have removed certainly.

```
[ ]  '''Let's drop any duplicate entries and check the shape of our dataset'''
     df.drop_duplicates()
     df.shape
```

```
(8807, 12)
```

```
df.isnull().sum()
```

```
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
dtype: int64
```

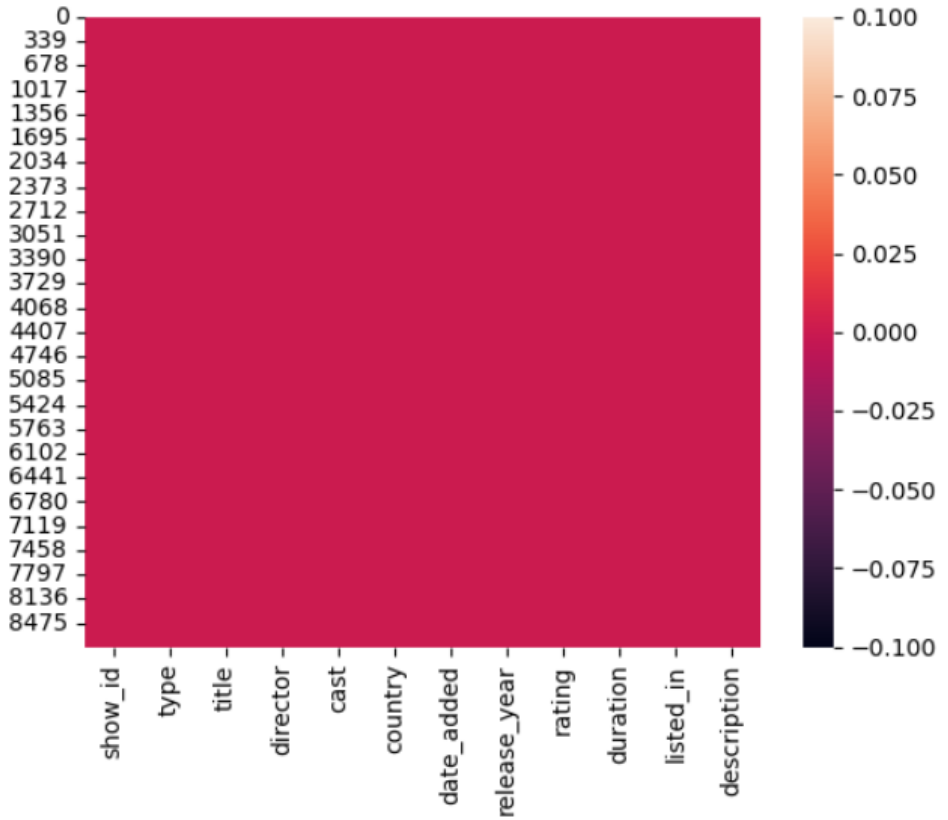
Data Cleaning & Pre-Processing:

Conclusion : In conclusion, dealing with null values requires a systematic approach based on the nature of the data and the specific attributes. By employing the described strategies, we can effectively handle and fill the missing values in the dataset, ensuring data completeness, integrity, and reliability for subsequent analysis and insights derivation.

Also, we can now proceed further to analyze our dataset as all the Null/Missing/Invalid values has been handled accordingly.

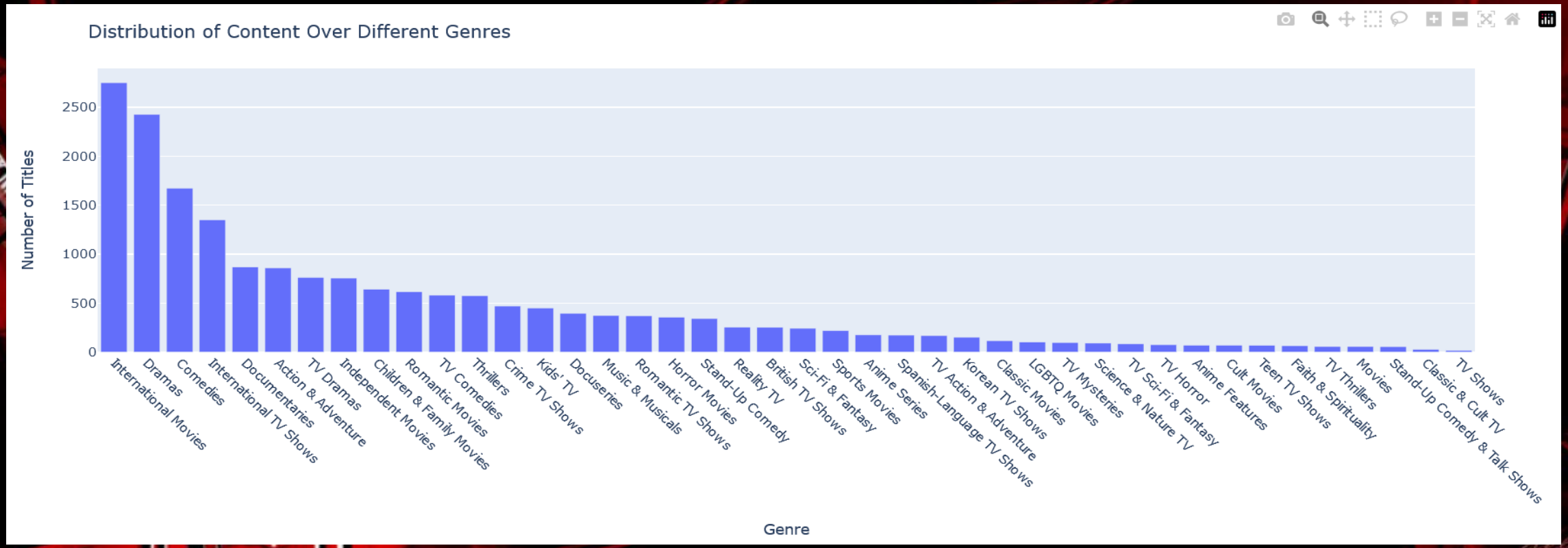
```
[ ]  '''Before moving on to explore dataset, let's draw a HEATMAP of cleaned dataset'''
     '''As all the Null/missing values has been handled, we can now analyse our dataset'''

sns.heatmap(df.isnull())
```

```
<Axes: >
```



Data Visualization and Insights

Representing the Distribution of content over different genres :



Data Visualization and Insights

Representing the Distribution of content over different genres :

Key insights

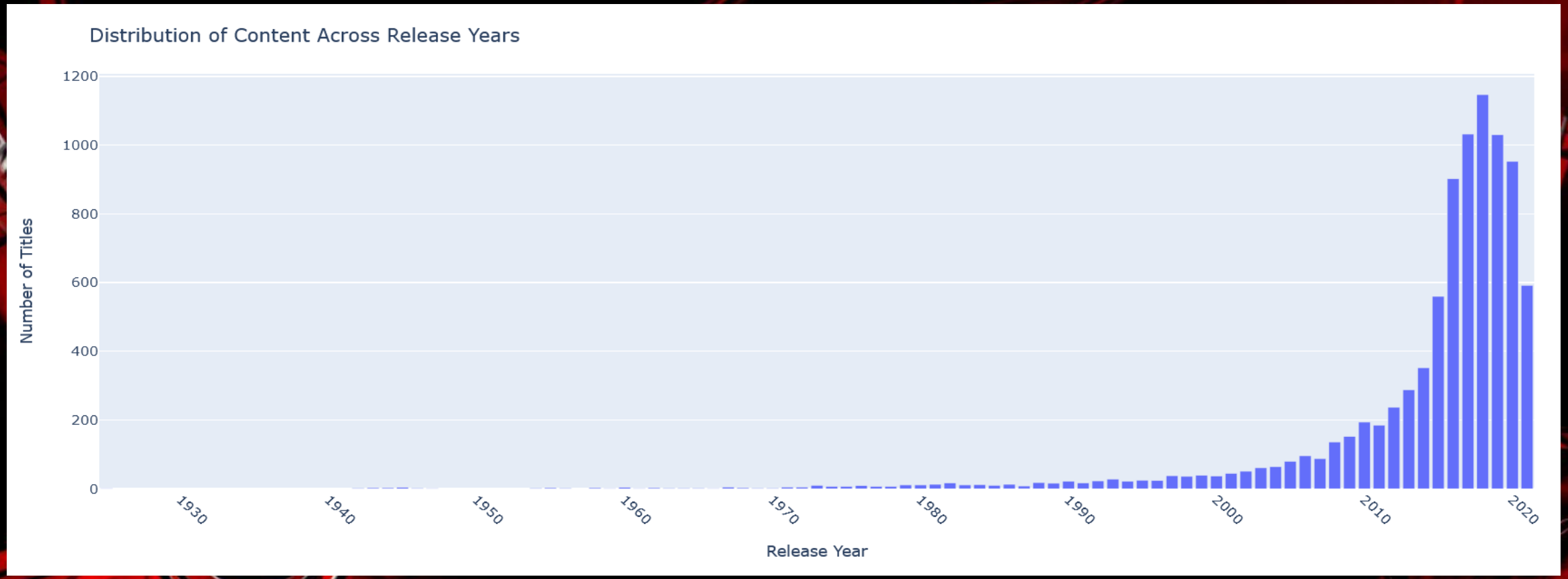
- **Dominant Genres:** The bar chart clearly shows "International Movies"(2752 titles) and "Dramas”(2427 titles)" are among the top genres.
- It indicates a significant portion of content falls under these categories.
- **Potential Areas for Growth:** Identifying genres with lower representation like "TV Shows" & "Classic TV" could highlight areas for potential expansion or content acquisition.

Verdict

- ✓ Netflix's content library is heavily skewed towards international movies, dramas, and comedies.
- ✓ This suggests that these genres are most popular among their audience. However, there is also a demand for niche genres, as indicated by the presence of categories like "Classic Movies" and "Independent Movies".

Data Visualization and Insights

Visualizing Distribution of content across release years :



Data Visualization and Insights

Visualizing Distribution of content across release years :

Key insights

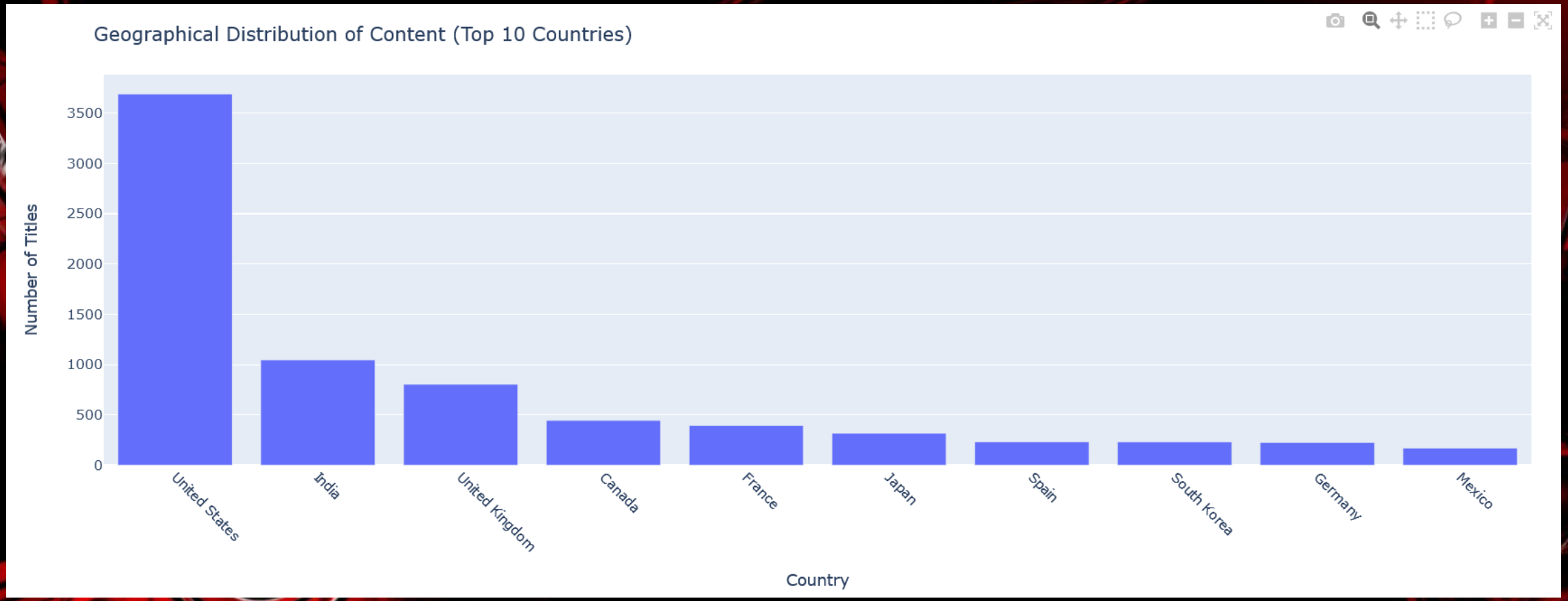
- **Recent Content Dominance:** The bar chart shows a clear trend of increasing content volume in recent years, peaking around 2019-2020. This suggests a focus on providing fresh content to viewers.
- **Content Library Growth:** The upward trend also indicates a continuous expansion of the Netflix content library over time.

Verdict

- ✓ Netflix's content library is heavily focused on recent releases, suggesting a strategy to attract viewers with fresh and current programming.
- ✓ However, the presence of older titles caters to a wider audience and provides a diverse selection.

Data Visualization and Insights

Exploring Geographical Distribution of content across countries :



Data Visualization and Insights

Exploring Geographical Distribution of content across countries :

Key insights

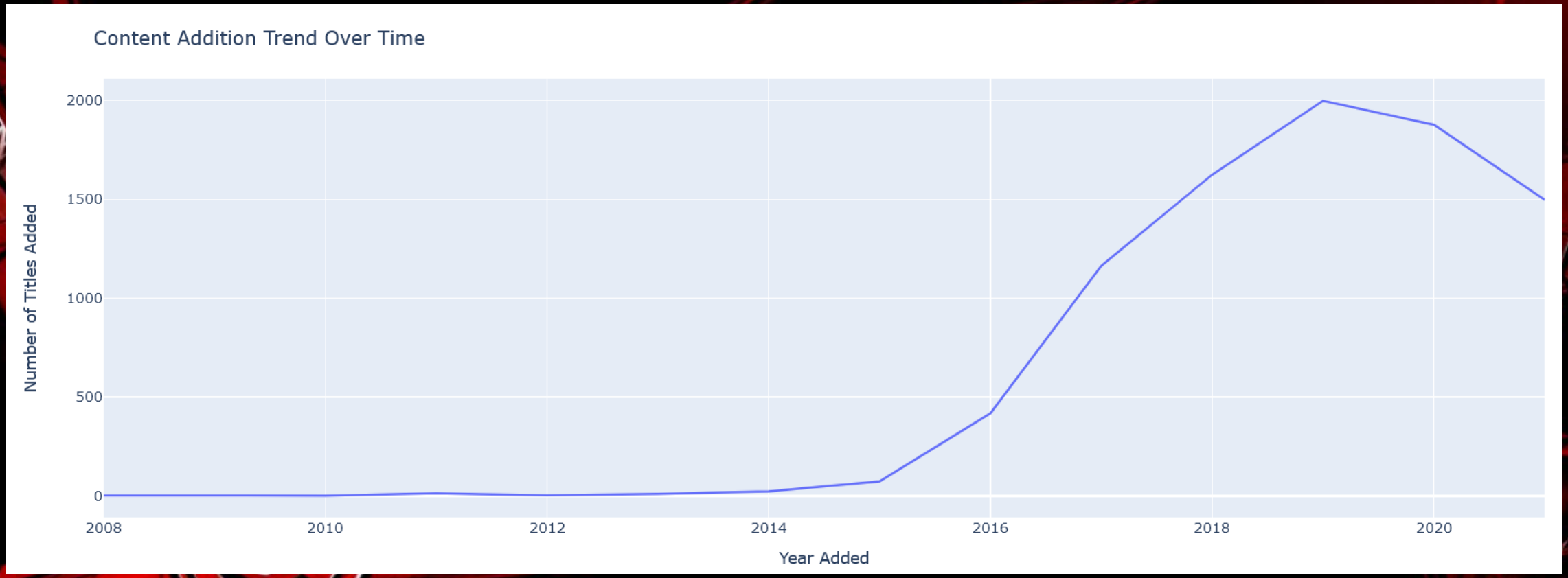
- **Content Origin Diversity:** The bar chart reveals the top countries contributing to Netflix's content library.
- The US is the primary contributor of content on Netflix with 3,689 titles, followed by India(1046 titles) and UK(804 titles).
- **Strategic Focus Areas:** Identifying countries with a high number of titles might indicate key markets for Netflix's content acquisition and production strategies.
- **Emerging markets are contributing:** Countries like UK, Canada etc. are becoming significant content providers.

Verdict

- ✓ Netflix's strategy of diversifying its content to include productions from different countries and regions reflects its goal of becoming a leading global entertainment platform.

Data Visualization and Insights

Performing Time Series analysis to identify trends and patterns over time:



Data Visualization and Insights

Performing Time Series analysis to identify trends and patterns over time:

Key insights

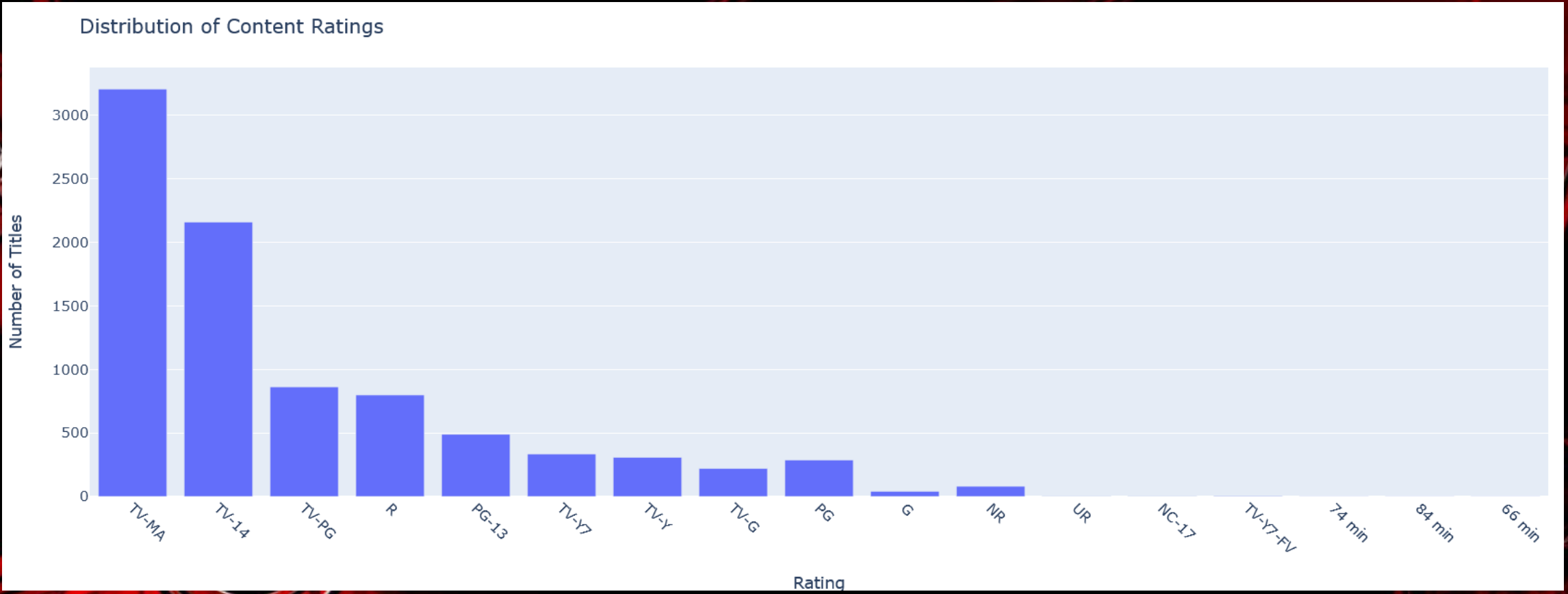
- **Growth:** The line chart shows a significant increase in content added to Netflix from Year 2015 until around 2019, followed by a slight decrease in 2020/2021.
- This could indicate that either Netflix is potentially focusing on quality over quantity, or facing challenges in content acquisition.
- **Seasonal Trends:** By analyzing monthly data could reveal seasonal patterns in content additions, which might be linked to viewer behavior, if available.

Verdict

- ✓ Netflix continues to grow its content library, showing dedication to offering viewers more choices.
- ✓ Understanding potential seasonality could help optimize content release strategies.
- ✓ Investigating the reasons behind significant yearly fluctuations could provide insights into market trends or internal strategic decisions.

Data Visualization and Insights

Analyze the Distribution of content ratings :



Data Visualization and Insights

Analyze the Distribution of content ratings :

Key insights

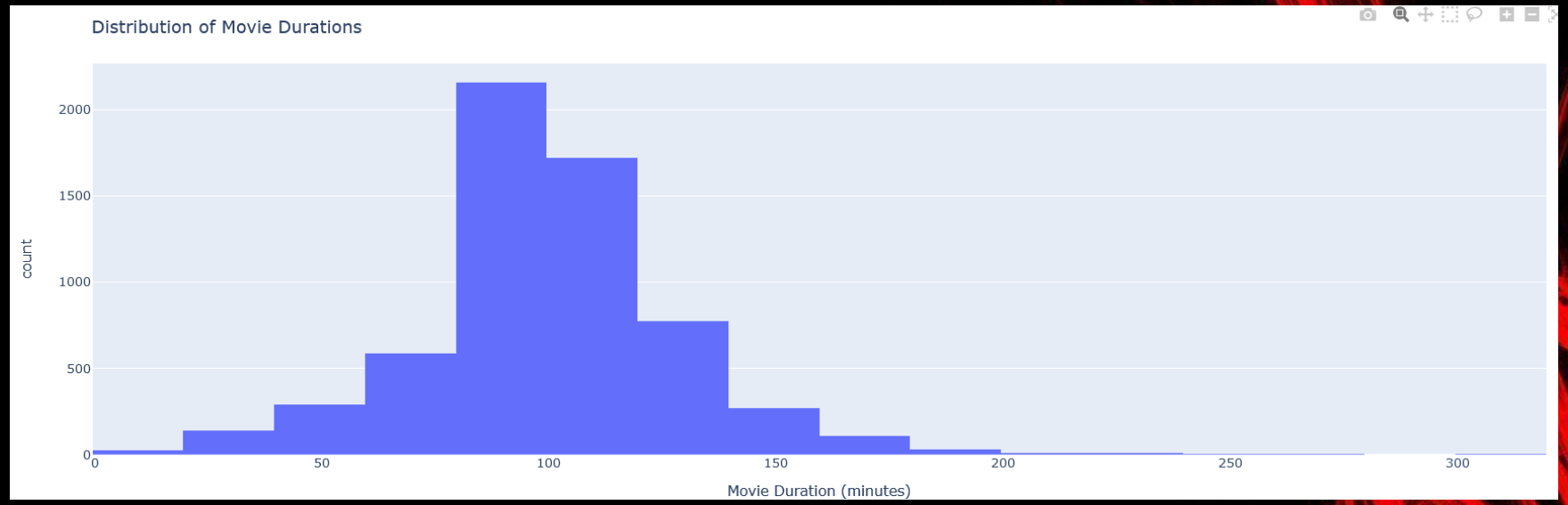
- **Mature Content Dominance:** The bar chart shows a giant portion of content is rated TV-MA (Mature Audience) and TV-14 (Parents Strongly Cautioned), indicating a focus on adult or older teen demographics.
- **Family-Friendly Content:** There's a decent amount of content with ratings like TV-PG, TV-Y7, and TV-Y, catering to families and younger audiences.
- **Content Strategy:** The distribution of ratings reflects Netflix's strategy to cater to a wide range of audience preferences.

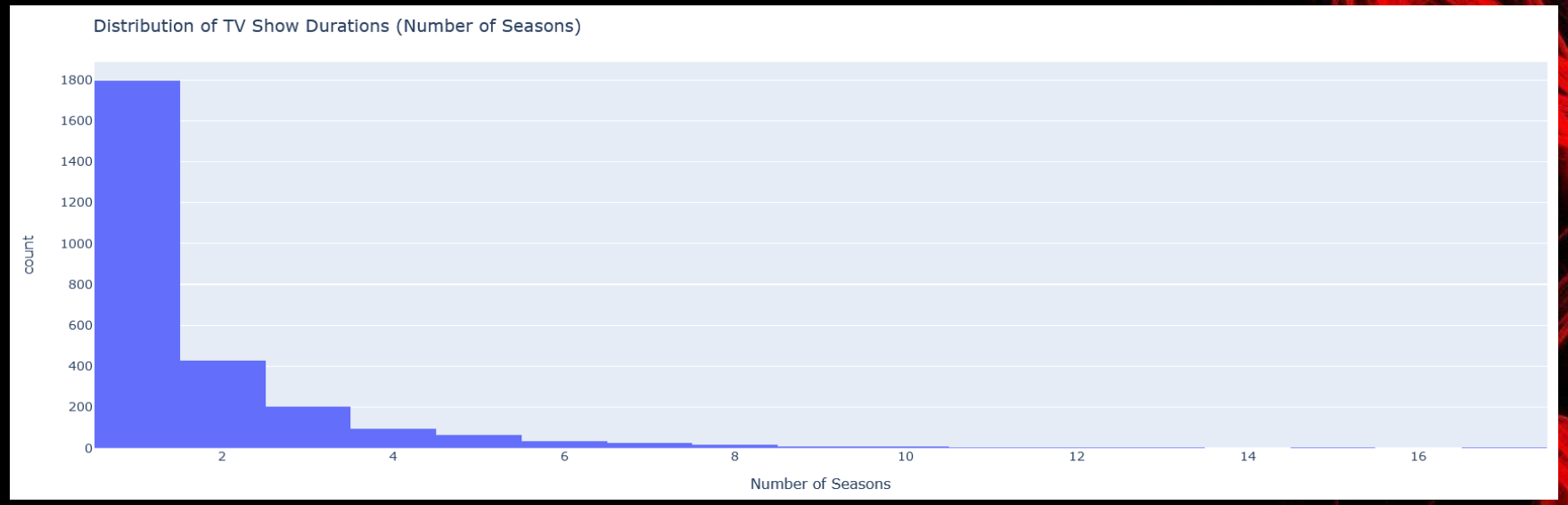
Verdict

- ✓ Understanding this distribution is crucial for content creators and Netflix to tailor their offerings to their target audience.

Data Visualization and Insights

Exploring the length of Movies/TV Shows and identifying trends, if any :





Data Visualization and Insights

Exploring the length of Movies/TV Shows and identifying trends, if any :

Key insights

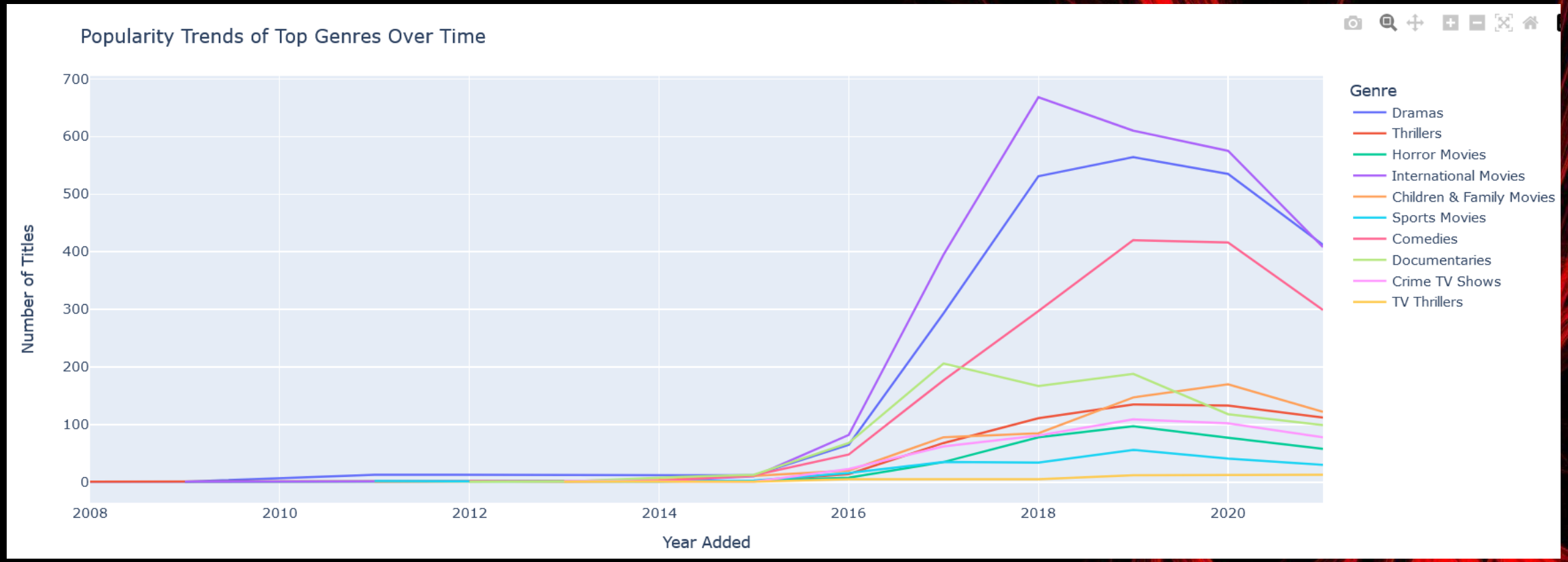
- **Movie Duration:** The histogram for movies shows a peak around 90-100 minutes, suggesting a preference for standard feature film lengths.
- There's a smaller but notable presence of shorter movies under 90 minutes.
- **TV Show Seasons:** The histogram for TV shows indicates that a majority of shows have 1-3 seasons. This could reflect the challenges of maintaining viewer engagement over many seasons.

Verdict

- ✓ **Movie Duration:** Netflix offers a diverse range of movie lengths, catering to different viewer preferences. The majority of movies fall within the typical feature film length, but there are options for those seeking shorter or longer viewing experiences.
- ✓ **TV Show Seasons:** Most TV shows on Netflix have a limited number of seasons, likely due to the difficulty of sustaining viewership over extended periods. This suggests a focus on delivering concise and impactful storytelling within a shorter timeframe.

Data Visualization and Insights

Analyzing trends in the popularity of different genres over time :



Data Visualization and Insights

Analyzing trends in the popularity of different genres over time :

Key insights

- **Genre Popularity Shifts:** The line chart shows how the popularity of different genres has changed over time.
- # For example, "International Movies" and "Dramas" have seen a consistent rise in recent years.

- **Emerging Trends:** We might observe genres that have gained popularity more recently, indicating potential shifts in viewer preferences.

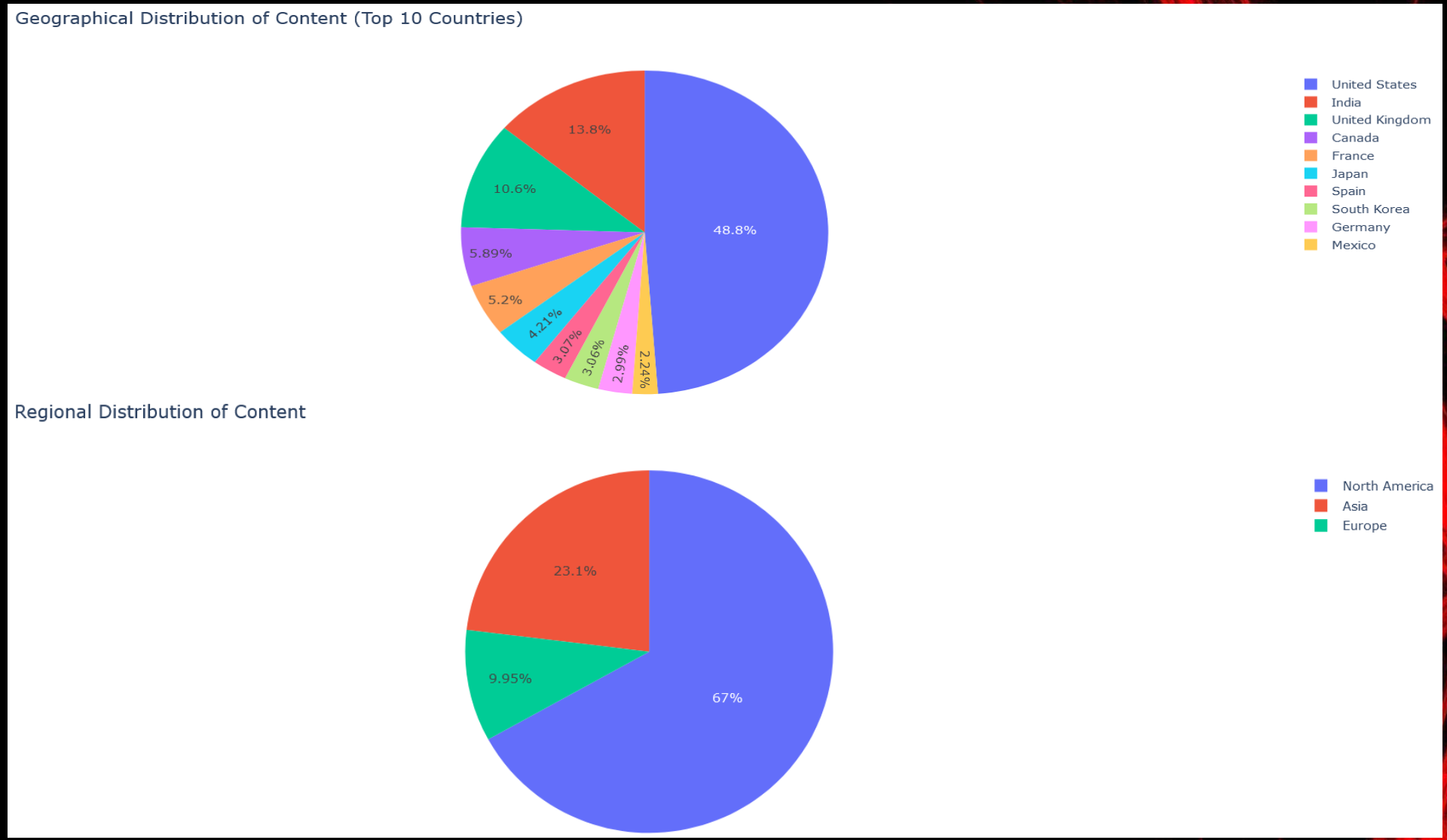
- **Content Strategy Alignment:** This analysis can help Netflix understand if their content strategy aligns with evolving viewer tastes and identify areas for potential genre expansion or reduction.

Verdict

- ✓ The popularity of genres on Netflix fluctuates and is shaped by factors such as global trends, viewer preferences, and the platform's content strategy. Insight into these trends aids Netflix in making well-informed choices regarding acquiring and producing content.

Data Visualization and Insights

Exploring the distribution of content across different countries & regions :



Data Visualization and Insights

Exploring the distribution of content across different countries & regions :

Key insights

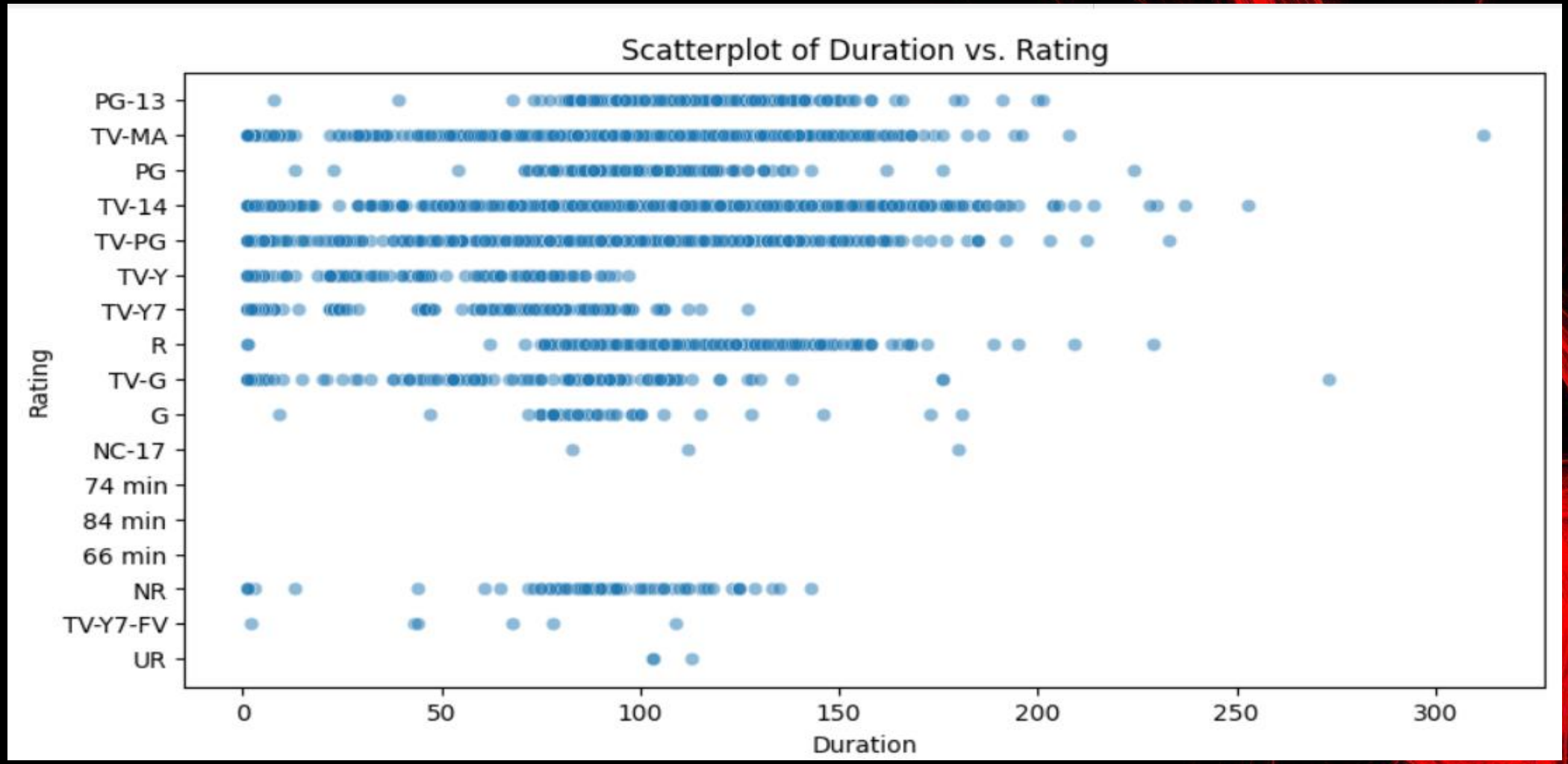
- **Global Reach:** The pie charts visualize the distribution of content across countries and regions. It highlights Netflix's efforts to cater to a global audience by sourcing content from various parts of the world.
- **Key Markets:** Identifying countries or regions with a significant share of content can indicate key markets for Netflix's growth and investment.
- **Content Localization Strategy:** A significant portion of content originates from North America, India and Europe.
- **Expansion Opportunities:** Analyzing underrepresented regions could reveal potential areas for Netflix to expand its content library and reach new audiences.

Verdict

- ✓ Netflix's content library predominantly features Western productions, highlighting its origins and focus on key markets.
- ✓ The platform is actively broadening its international content selection to appeal to a diverse global audience.
- ✓ There is an opportunity for additional diversification of content, particularly from regions that are currently underrepresented.

Data Visualization and Insights

Investigating potential correlations between variables 'ratings' & 'duration'.



Data Visualization and Insights

Investigating potential correlations between variables ‘ratings’ & ‘duration’.

Key insights

- **No Clear Correlation:** The scatterplot doesn't show a strong linear relationship between duration and rating. This suggests that the length of a movie or TV show doesn't necessarily dictate its rating.

- **Further Analysis:** To gain deeper insights, we could explore correlations between rating and other variables like genre, release year, or country of origin. We could also use statistical tests to quantify the strength of any potential relationships.

Verdict

- ✓ **Rating Distribution Across Durations:** We can observe that movies and TV shows of various durations receive a wide range of ratings. This indicates that factors other than duration play a significant role in determining user ratings.

Data Visualization and Insights

Evaluating the diversity of content by analyzing the number of unique genres and categories :

```
'''Question 11 - Evaluate the diversity of content by analyzing the number of unique genres and categories.'''

# Calculate the number of unique genres.
unique_genres = df['listed_in'].str.split(', ').explode().unique()
num_unique_genres = len(unique_genres)
print("Number of unique genres:", num_unique_genres)

# Assuming 'listed_in' column contains both genres and categories.
unique_categories = df['listed_in'].str.split(', ').explode().unique()
num_unique_categories = len(unique_categories)
print("Number of unique categories (including genres):", num_unique_categories)
```

```
Number of unique genres: 42
Number of unique categories (including genres): 42
```

Data Visualization and Insights

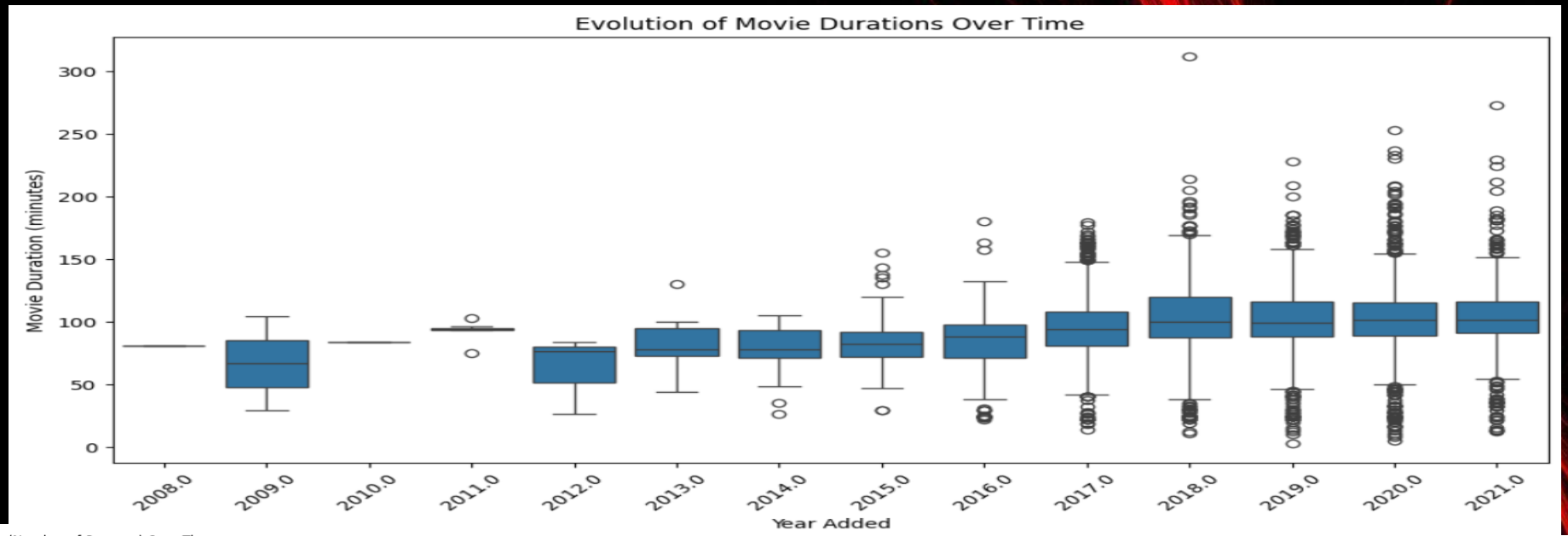
Evaluating the diversity of content by analyzing the number of unique genres and categories :

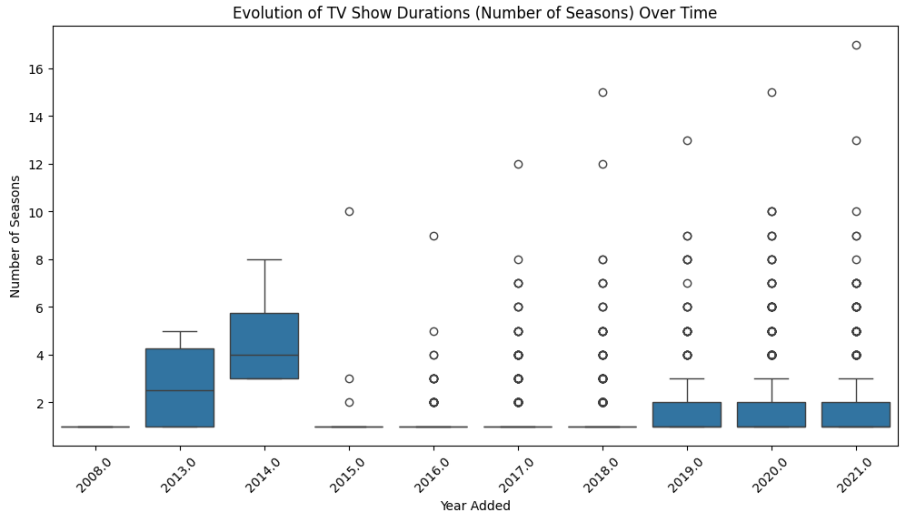
Key insights

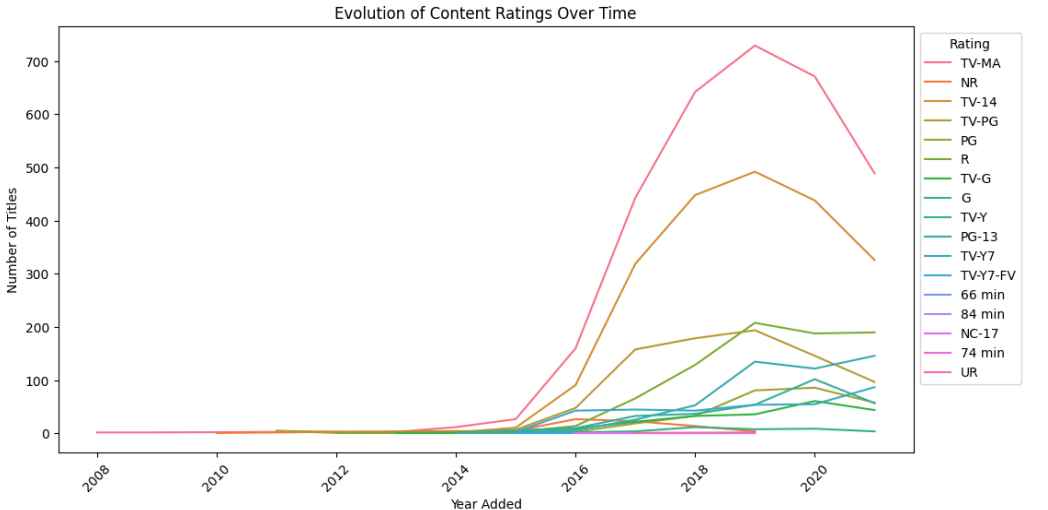
- **Content Diversity:** The number of unique genres and categories reflects the diversity of content available on Netflix. A higher number indicates a wider range of options for viewers.
- **Niche Content:** The presence of numerous unique categories suggests that Netflix caters to various niche interests, potentially attracting a broader audience.
- **Content Strategy:** This analysis can help Netflix evaluate the effectiveness of their content diversification strategy and identify areas for potential expansion.

Data Visualization and Insights

Exploring how the characteristics of content (duration & ratings) have evolved over the years.







Data Visualization and Insights

Exploring how the characteristics of content (duration & ratings) have evolved over the years.

Key insights

- **Movie Duration Trends:** The box plots for movie durations show how the distribution of movie lengths has changed over time. We might observe a trend towards shorter movies in recent years, or a wider range of durations being offered.
- **TV Show Season Trends:** Similarly, the box plots for TV show durations (number of seasons) reveal trends in the length of TV series. We might see a shift towards shorter series, or a greater variety in the number of seasons offered.
- **Content Strategy Adaptation:** These analyses provide insights into how Netflix's content strategy has adapted to changing viewer preferences and industry trends.
- # It can help them identify areas for potential adjustments to their content acquisition and production strategies.

Verdict

- ✓ **Rating Trends:** The line chart for content ratings shows how the distribution of ratings has evolved over time. We might observe an increase in the proportion of mature content, or a more balanced distribution across different rating categories.
- ✓ The analysis suggests that content characteristics on Netflix have evolved to adapt to changing viewer preferences and market trends.

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :

KEY FINDINGS :

❑ Content Distribution :

- ***- Dominant genres are International Movies, Dramas.***
- ***- Recent content dominance: Peak content volume around 2019-2020.***
- ***- Geographical distribution: US, India, UK as major contributors.***
- ***- Content ratings: Majority rated TV-MA and TV-14.***

❑ Trends and Patterns :

- ***- Content addition: Significant increase until 2019, slight decrease in 2020/2021.***
- ***- Movie durations: Peak around 90-100 minutes.***
- ***- TV show durations: Majority with 1-3 seasons.***
- ***- Genre popularity: Rise of "International Movies" and "Dramas".***
- ***- Regional distribution: North America, India, and Europe as major content sources.***

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :

KEY FINDINGS :

- Correlations and Diversity :

- ***- No strong correlation between duration and rating.***
- ***- High number of unique genres and categories indicate diverse content library.***

- Evolution of Content :

- ***- Potential trend towards shorter movies in recent years.***
- ***- Greater variety in TV show durations.***
- ***- Possible increase in the proportion of mature content.***

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :

Recommendations :

- Content Diversification :

- *- Explore genres with lower representation (e.g., "TV Shows", "Classic TV") for potential expansion.*
- *- Consider increasing content from underrepresented regions to reach new audiences.*

- Content Strategy Refinement :

- *- Continue monitoring genre popularity trends to align content acquisition and production with viewer preferences.*
- *- Evaluate the impact of shorter movie and TV show formats on viewer engagement.*

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :

Recommendations :

- Data-Driven Decision Making :

- ***- Leverage further analysis (e.g., correlations between rating and other variables) to inform content decisions.***
- ***- Utilize user ratings and feedback to personalize content recommendations and improve user experience.***

- Continuous Monitoring and Adaptation :

- ***- Stay abreast of industry trends and viewer behavior to proactively adjust content strategies.***
- ***- Regularly evaluate the effectiveness of content initiatives and make data-driven adjustments.***

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis :

Conclusions :

- *- Netflix caters to a global audience with diverse content, focusing on adult and older teen demographics.*
- *- The platform continuously expands its content library, primarily with recent releases.*
- *- Strategic focus on key markets like the US, India, and UK.*
- *- Content strategy adapts to evolving viewer preferences, with a potential shift towards shorter formats.*
- *To improve user experience, Netflix can invest in enhancing its search and discovery features to help users find content that matches their interests more easily.*
- *The platform can conduct regular reviews of its content library to ensure it remains fresh and relevant to users' evolving tastes and preferences.*
- *The platform can further analyze user data and viewing patterns to gain deeper insights into user preferences and tailor content recommendations accordingly.*

THANK YOU FOR READING

FOR CODING PART, KINDLY VISIT:-

https://github.com/SaxenaKushagr/EDA_Netflix.git