

Exploratory  
Data  
analysis of  
Netflix

# NETFLIX

## Exploratory Data Analysis

By Lucky Bisht

N

# ABOUT NETFLIX

Netflix is a global leader in streaming entertainment, offering a vast library of TV shows, movies, documentaries, and original content to millions worldwide. Founded in 1997 by Reed Hastings and Marc Randolph as a DVD rental service, Netflix revolutionized media consumption by transitioning to online streaming in 2007. This shift popularized on-demand entertainment and binge-watching, making the platform a household name. Operating in over 190 countries, Netflix caters to a diverse audience with localized content. Its success is driven by advanced algorithms that provide personalized recommendations, enhancing user engagement. The company is also renowned for its original content. Since the release of "House of Cards" in 2013, Netflix Originals like "Stranger Things," "The Crown," and "Squid Game" have earned critical acclaim and loyal viewers.

Netflix has embraced innovation, offering 4K Ultra HD streaming and interactive storytelling, as seen in "Black Mirror: Bandersnatch." While it faces competition from Disney+, Amazon Prime Video, and Hulu, Netflix continues to lead by adapting to evolving consumer needs and exploring new revenue models, including ad-supported subscriptions. It remains a cultural phenomenon shaping the entertainment industry.



## ➤ Source Of The Dataset

This dataset, sourced from (Skill Circle), contains information collected from Netflix and resembles a typical movie and TV shows dataset. A quick examination reveals the presence of some missing values (NaN) in certain columns. The dataset includes 8,807 unique entries, representing a mix of TV shows and movies. Due to its structured format and the variety of data it offers, this dataset is commonly utilized by beginners for learning and practicing exploratory data analysis (EDA). Its blend of categorical and numerical data makes it an excellent resource for honing analytical skills and understanding trends within the entertainment industry.



## ➤ Goal Of The Project

The primary objective of this Netflix EDA project is to perform an in-depth analysis of Netflix's content dataset. The analysis involves exploring the dataset structure, addressing missing values and duplicates to ensure data integrity, and generating descriptive statistics. Additionally, the project focuses on visualizing content distribution across genres and release years, identifying temporal trends, and examining attributes such as ratings and duration. It also seeks to analyze audience engagement patterns. By extracting and synthesizing these insights, the project aims to uncover meaningful trends and provide actionable recommendations to improve Netflix's content offerings and enhance the overall user experience.



# ➤ Loading The Dataset

```
#Importing all the required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
```

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

⇒ Mounted at /content/drive
```

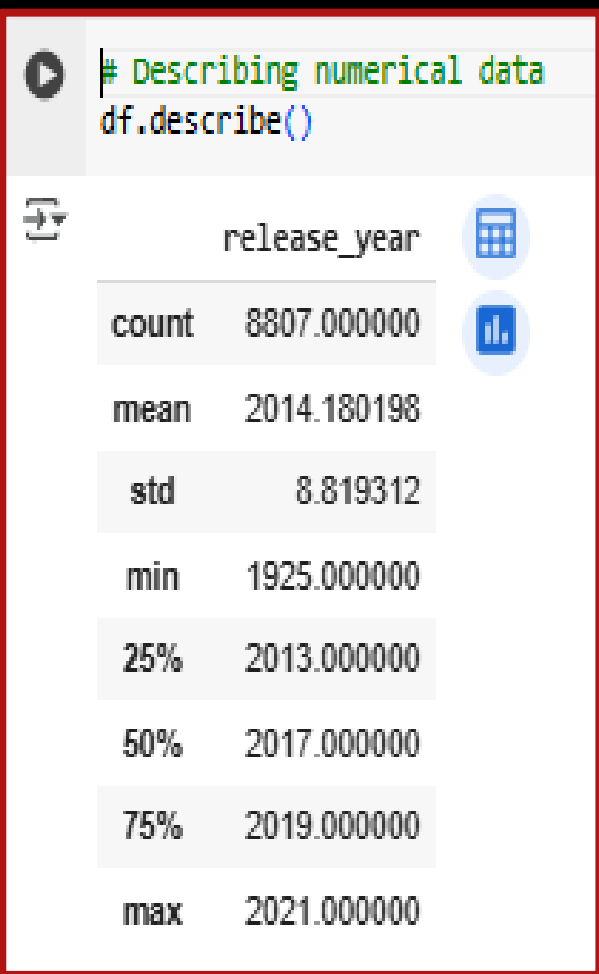
- ✓ I have conducted my work using google colab notebook.
- ✓ The dataset has been imported from google drive.
- ✓ As we begin our exploratory data analysis (EDA), i've named the dataset 'df'.
- ✓ The dataset comprises of 8807 rows and 12 columns.
- ✓ For data cleaning/visualization, I have utilized libraries like numpy, pandas, seaborn & plotly.
- ✓ Any duplicate entries that were found have also been removed

# ➤ Head function

```
# Loading the dataset from google drive
netflix = '/content/drive/MyDrive/netflix_titles.csv'
df = pd.read_csv(netflix)
df.head(10)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mababane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train l...
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	NaN	September 24, 2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brin...
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	NaN	September 24, 2021	2021	PG	91 min	Children & Family Movies	Equestria's divided. But a bright-eyed hero be...
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies	On a photo shoot in Ghana, an American model s...
8	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho...	United Kingdom	September 24, 2021	2021	TV-14	9 Seasons	British TV Shows, Reality TV	A talented batch of amateur bakers face off in...
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	September 24, 2021	2021	PG-13	104 min	Comedies, Dramas	A woman adjusting to life after a loss contend...

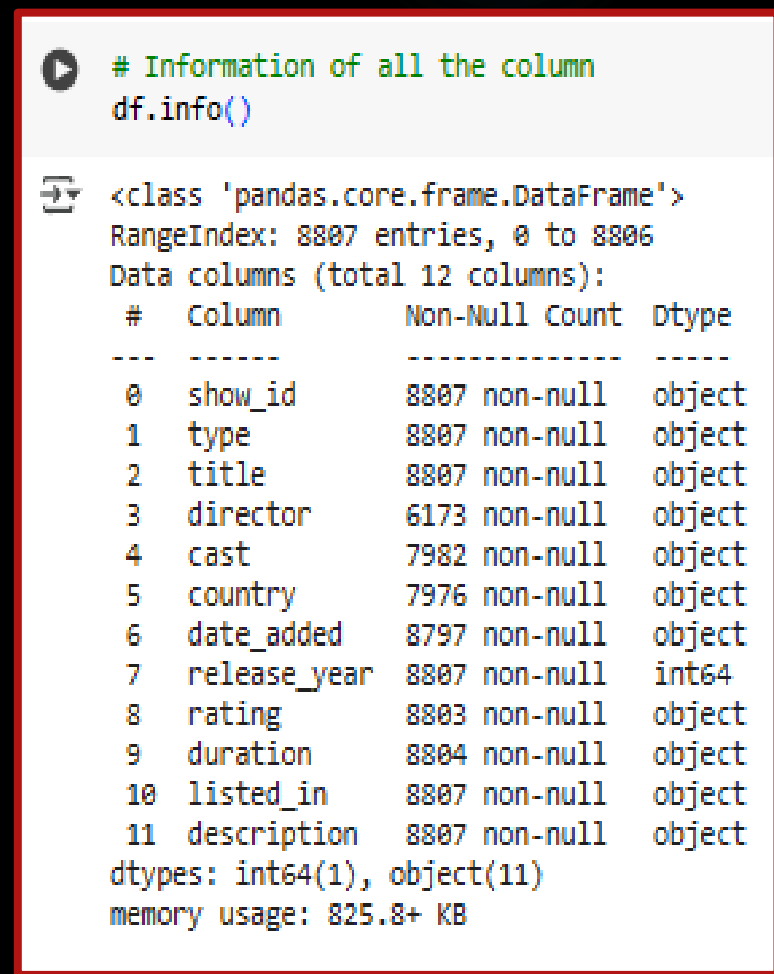
# ➤ Description Of The Dataset



# Describing



#Dtypes



# Information



## # Describing The Dataset

This dataset offers a comprehensive view of Netflix's content library, encompassing both movies and TV shows. It includes key details such as the title, director, cast, country of origin, release year, rating, duration, genres, and a brief description.

The dataset provides ample opportunities for analysis, enabling insights into various aspects of Netflix's offerings. Key areas of exploration include:

- Examining content distribution across genres, countries, and ratings.
- Analyzing trends in content additions and their popularity over time.
- Investigating relationships between variables, such as duration and rating.
- Assessing the diversity of content through unique genres and categories.
- Understanding the evolution of content characteristics over the years.

With its rich attributes, this dataset is ideal for uncovering patterns and trends within Netflix's vast content collection.

## # Dtypes of The Dataset

- This Dataset contain 12 columns.
- 11 columns are object and 1 is integer





# # Information of The Dataset

This dataset offers a detailed view of the content available on Netflix, providing key insights into the platform's content strategy and audience preferences.

The dataset includes the following main attributes:

- Show ID:** A unique identifier for each title.
- Type:** Classifies the content as either a "Movie" or "TV Show."
- Title:** The name of the movie or TV show.
- Director:** The director(s) associated with the content.
- Cast:** The list of actors featured in the movie or show.
- Country:** The country where the content originated.
- Date Added:** The date the content was made available on Netflix.
- Release Year:** The original release year of the content.
- Rating:** The content's rating (e.g., TV-MA, PG-13).
- Duration:** The length of movies or the number of seasons for TV shows.
- Listed In:** The genres and categories the content belongs to.
- Description:** A short synopsis of the movie or TV show.

By analyzing these features, we can derive valuable insights into Netflix's content offerings, audience preferences, and identify potential areas for strategic growth and optimization.



# ➤ Data Cleaning and Preparation

```
# Checking null values of the columns  
df.isnull().sum()
```

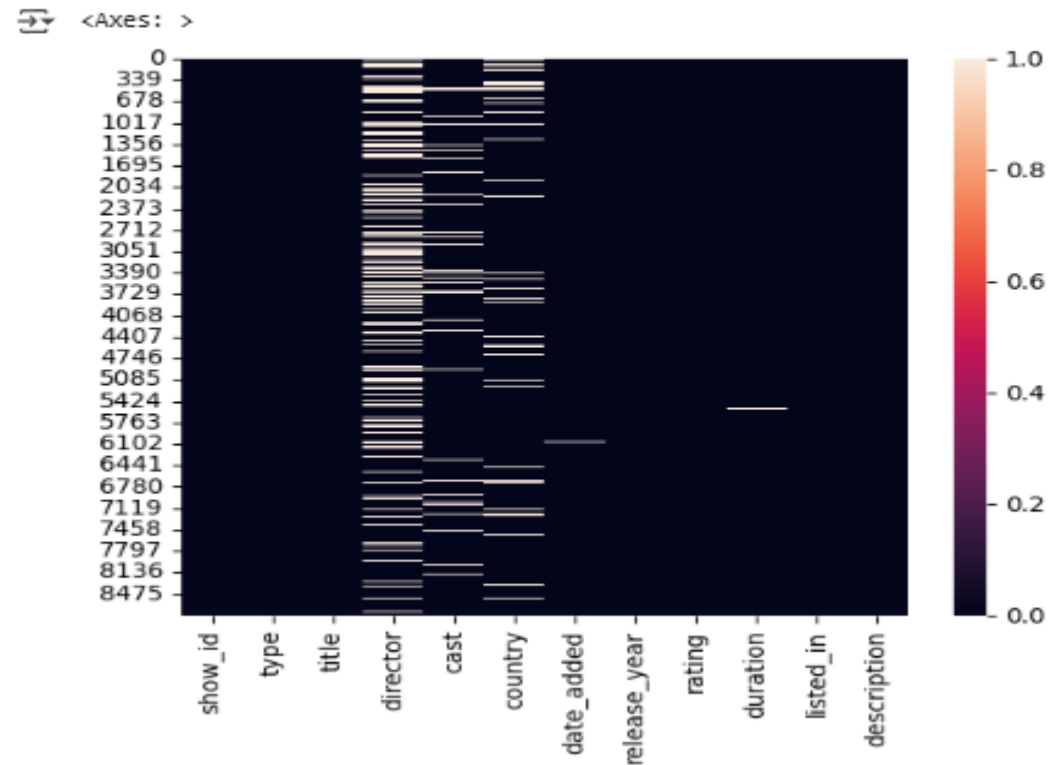
```
show_id    0  
type        0  
title       0  
director   2634  
cast       825  
country    831  
date_added  10  
release_year  0  
rating      4  
duration    3  
listed_in   0  
description  0
```

dtype: int64

```
[71] df.isnull().sum().sum()
```

```
4387
```

```
# Creating a heatmap to show null value  
# This will help us to check the null value present in Dataset  
sns.heatmap(df.isnull())
```



## ✓ The Dataset Contains A Total Of 4,307 Null Values.


- Columns with no null values:
  - Show\_id
  - Type
  - Title
  - Release\_year
  - Listed\_in
  - Description
- Columns with null values:
  - Director: 2634 null values
  - Cast: 825 null values
  - Country: 831 null values
  - Date\_added: 10 null values
  - Rating: 4 null values
  - Duration: 3 null values
- The image provides valuable information about the completeness of your dataset. Understanding the distribution of null values is crucial for making informed decisions about data cleaning and feature engineering steps in your analysis.
- Heatmap shows that “director”, “cast” and “country” has much null values.
- This heatmap was used to visualize these null values, highlighting areas that require attention. Addressing these missing values is essential before proceeding with further analysis.

## ➤ Filling and replacing null values

```
▶ # Filling and replacing null values with Descripted statistics
df['director'].fillna('director not found',inplace = True)
df['cast'].fillna('cast not found',inplace = True)
df['country'].fillna('Country not found',inplace = True)
df['date_added'].fillna('Date not added',inplace = True)
df['rating'].fillna('rating not given',inplace = True)
df['duration'].fillna('duration not given', inplace = True)
```

**Director:** For the 'Director' attribute, which has 2,634 missing values, one solution is to replace these null entries with a placeholder such as 'Unknown' or 'Director Not Found.' This ensures that the data remains intact while signaling the lack of director information. Alternatively, to enhance data accuracy, we can investigate and fill in the missing director details by consulting external sources or relevant databases for the specific movies or TV shows.

**Cast:** The 'Cast' attribute contains 825 missing values. To ensure data consistency, we can fill these gaps with a placeholder such as 'Cast Not Found.' Alternatively, we can enhance the accuracy of the dataset by using external sources like IMDb or other relevant databases to retrieve and populate the missing cast details for each movie or TV show.



**Country:** For the 'Country' attribute, which has 831 missing values, one option is to replace the null entries with the most frequent country of origin or a placeholder like 'Country Not Found.' Another approach could be to cross-reference the title or other metadata to deduce the country of production, using clues from the content's origin or the production company's details.

**Date:** Since there are only 10 missing values in the 'Date Added' attribute, addressing these gaps is straightforward. We can simply impute the missing dates by filling them with a placeholder such as 'Date Not Added.'

**Rating:** For the 'Rating' attribute, which has 4 missing values, I chose to replace the null entries with 'Rating Not Given.' This approach avoids making assumptions about the ratings, ensuring fairness and maintaining the integrity of the dataset

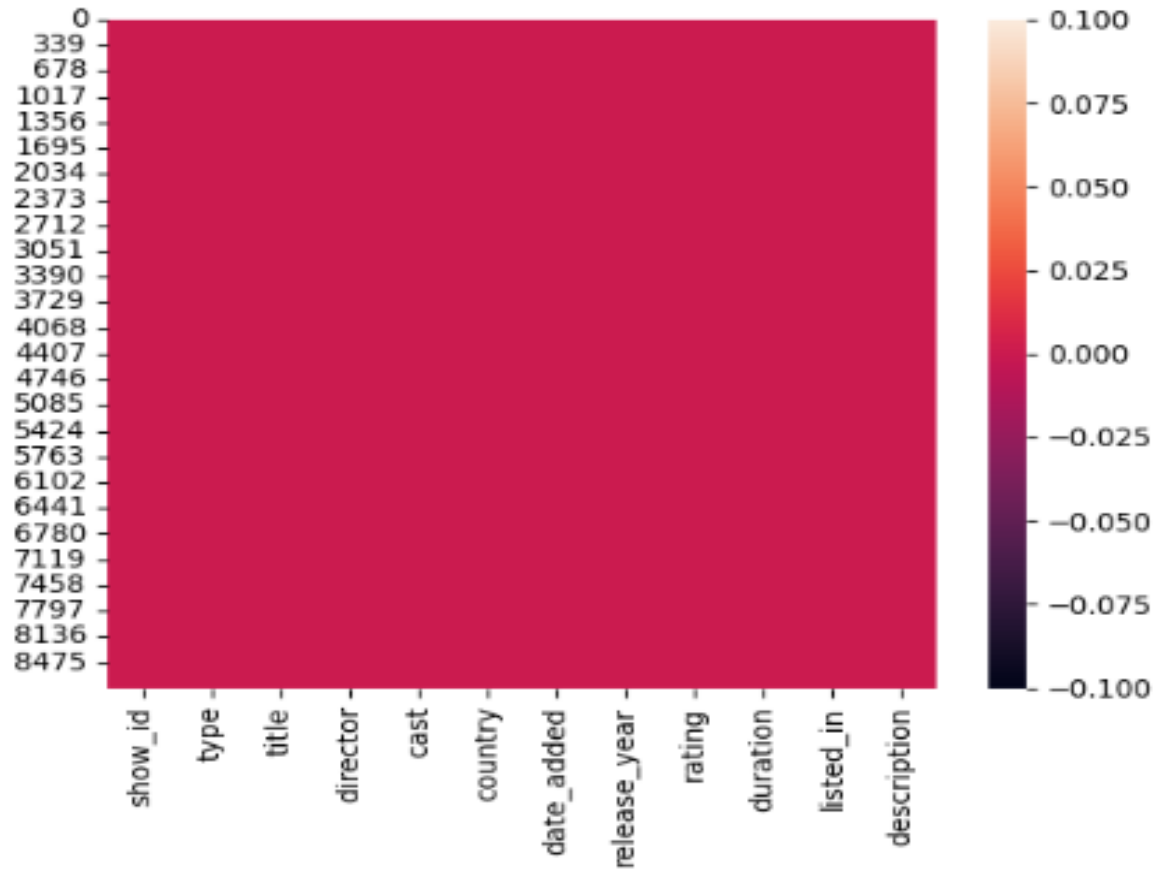
**Duration:** For the 'Duration' attribute, which has 3 missing values, I replaced the null entries with 'Duration Not Given.' This decision ensures that no assumptions are made about the duration, maintaining fairness and accuracy in the dataset.



# ➤ This Heatmap Shows That Dataset Is Clean

```
# creating heatmap to show that dataset is clear  
sns.heatmap(df.isnull())
```

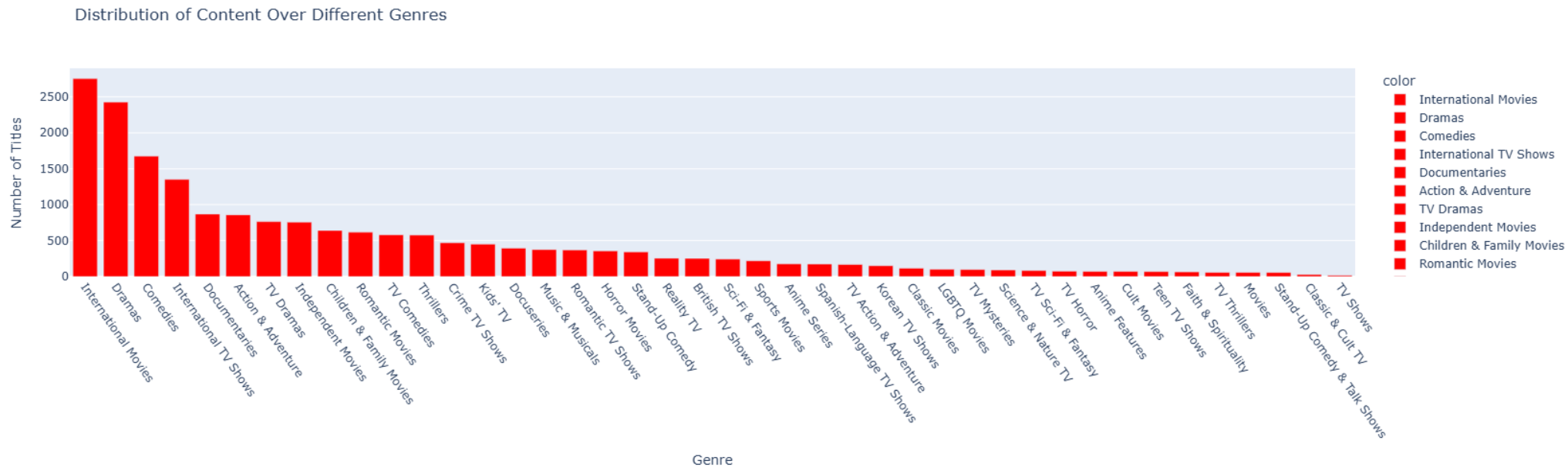
<Axes: >



# ➤ Data Visualization and Key Insights

```
# Distribution of content over different genres by using Barplot.  
genre_counts = df['listed_in'].str.split(', ').explode().value_counts()  
  
fig = px.bar(  
    x=genre_counts.index,  
    y=genre_counts.values,  
    labels={'x': 'Genre', 'y': 'Number of Titles'},  
    title='Distribution of Content Over Different Genres',  
    color=genre_counts.index,  
    color_discrete_sequence=['red']  
)  
  
fig.update_layout(xaxis_tickangle=55)  
fig.show()
```

# Distribution of content over different genres





## Key Insights:

- ✓ **Dominant Genres:** The bar chart reveals that "International Movies" (2,752 titles) and "Dramas" (2,427 titles) are the leading genres on Netflix. This indicates that a large portion of the content is concentrated in these categories.
- ✓ **Opportunities for Growth:** Genres with lower representation, such as "TV Shows" and "Classic TV," present potential areas for growth or further content acquisition. Targeting these genres could help balance Netflix's library.
- ✓ **Content Library Focus:** Netflix's content library is heavily focused on "International Movies," "Dramas," and "Comedies." This suggests these genres are the most popular among its viewers. However, niche genres like "Classic Movies" and "Independent Movies" also hold appeal, highlighting a demand for diverse content.

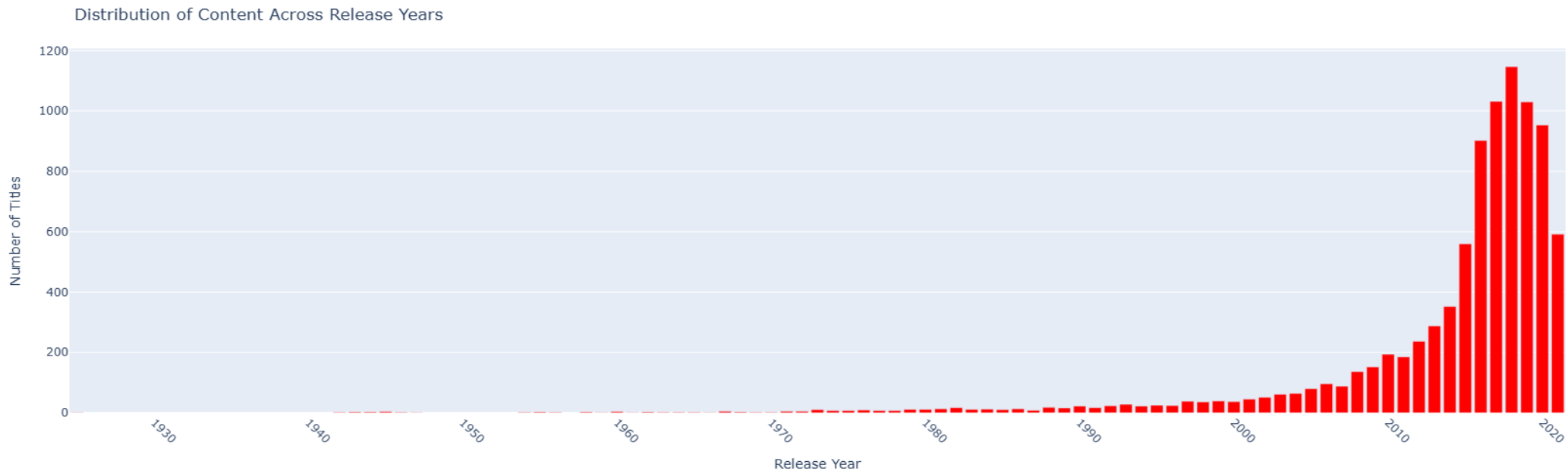


```
# creating barplot for distribution of content across release year
release_years = df['release_year'].value_counts().sort_index()

fig = px.bar(
    x=release_years.index,
    y=release_years.values,
    labels={'x': 'Release Year', 'y': 'Number of Titles'},
    color_discrete_sequence=['red']
)

fig.update_layout(title='Distribution of Content Across Release Years', xaxis_tickangle=45)
fig.show()
```

# Distribution of content Across  
release year



## ➤ Data Visualization and Key Insights

### Key Insights:

- ✓ **Dominance of Recent Content:** The bar chart reveals a clear increase in content volume in recent years, with a peak around 2019-2020. This trend suggests Netflix's emphasis on offering new and up-to-date content to its audience.
- ✓ **Content Library Expansion:** The steady upward trajectory highlights the ongoing growth of Netflix's content library, indicating a strategy of continual expansion.
- ✓ **Focus on New Releases:** Netflix's content library is heavily concentrated on recent releases, signaling a strategy designed to attract and retain viewers with fresh, current programming.



```

▶ # creating a barplot of the geographical distribution of content
# For more attraction of this project i have choosed countries colours in their barplot
country_counts = df['country'].str.split(',').explode().value_counts()

flag_colors = {
    'United States': 'blue',
    'India': 'orange',
    'United Kingdom': 'red',
    'Canada': 'red',
    'France': 'blue',
    'Germany': 'black',
    'Australia': 'green',
    'Japan': 'red',
    'South Korea': 'blue',
    'Italy': 'green'
}

# Map colors of each countries
top_10_countries = country_counts.head(10)
bar_colors = [flag_colors.get(country, 'gray') for country in top_10_countries.index]

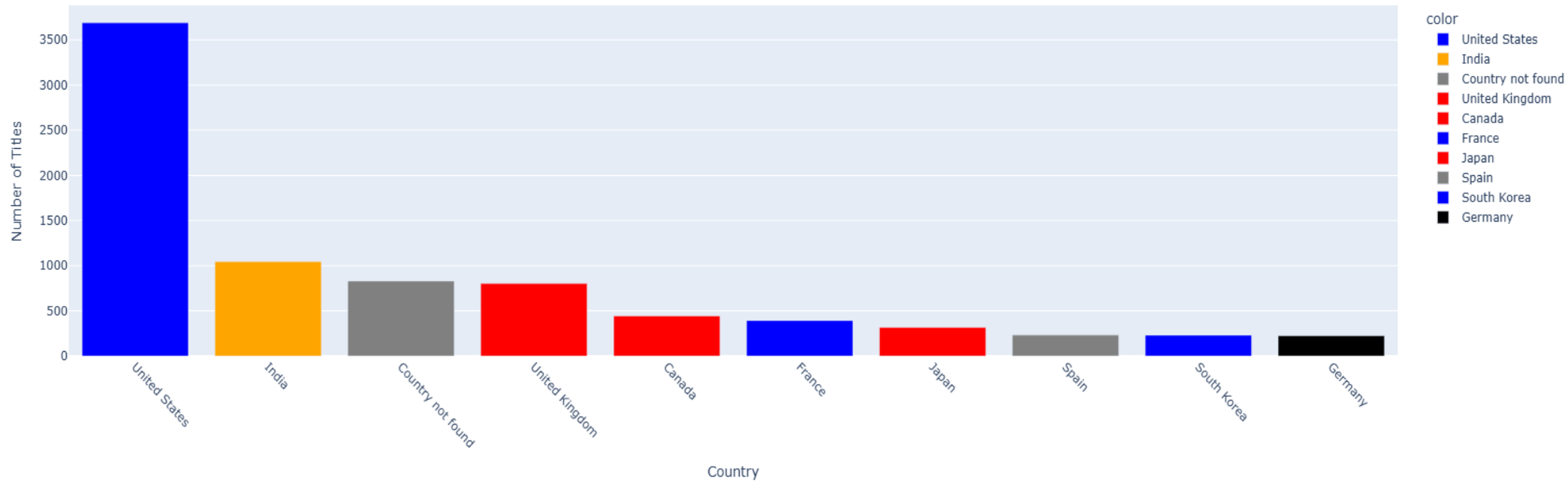
fig = px.bar(
    x=top_10_countries.index,
    y=top_10_countries.values,
    labels={'x': 'Country', 'y': 'Number of Titles'},
    color=top_10_countries.index,
    color_discrete_sequence=bar_colors
)

fig.update_layout(title='Geographical Distribution of Content (Top 10 Countries)', xaxis_tickangle=45)
fig.show()

```

# Geographical Distribution of content ( top 10 countries)

Geographical Distribution of Content (Top 10 Countries)



## Key Insights:

- ✓ **Content Origin Diversity:** The bar chart highlights the top countries contributing to Netflix's content library, showcasing a diverse range of origins.
- ✓ **Leading Contributors:** The United States stands as the largest contributor, with 3,689 titles, followed by India (1,046 titles) and the UK (804 titles).
- ✓ **Strategic Focus Areas:** Identifying countries with a high volume of titles could reveal key markets for Netflix's content acquisition and production strategies.
- ✓ **Growth of Emerging Markets:** Countries like the UK and Canada are increasingly becoming important providers of content, signaling the rise of emerging markets in Netflix's global library.



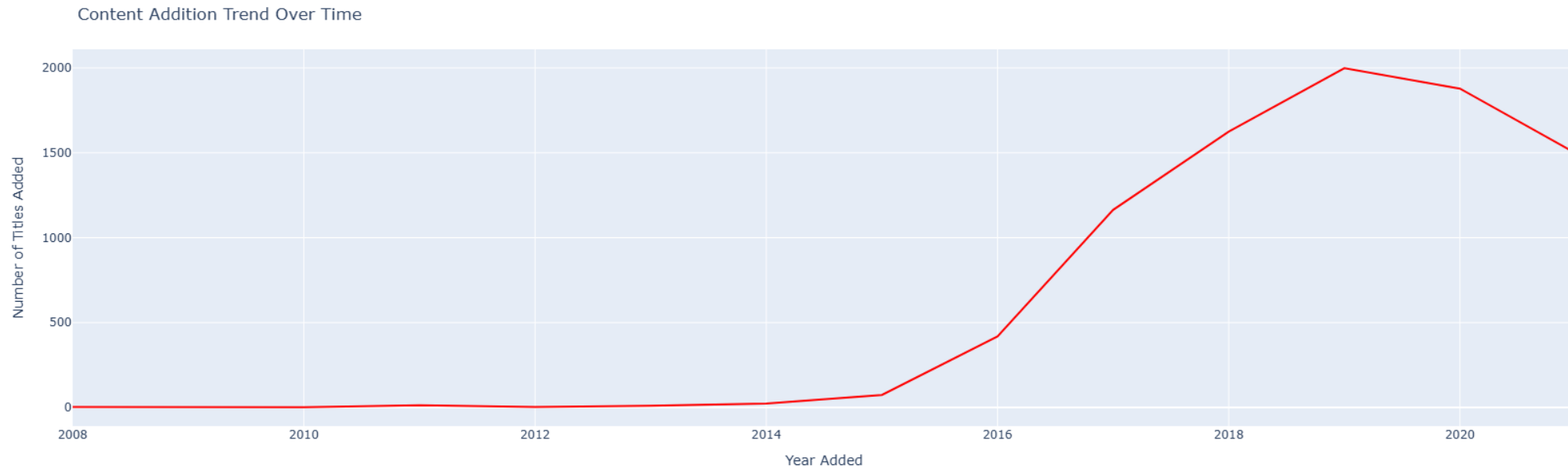
```
# creating a lineplot to check timr series analysis to identify trends and pattern over time
df['year_added'] = pd.to_datetime(df['date_added'], errors='coerce').dt.year
content_added_yearly = df['year_added'].value_counts().sort_index()

fig = px.line(
    x=content_added_yearly.index,
    y=content_added_yearly.values,
    labels={'x': 'Year Added', 'y': 'Number of Titles Added'},
    line_shape='linear'
)

fig.update_traces(line_color='red')
fig.update_layout(title='Content Addition Trend Over Time')

fig.show()
```

#Content addition trend over time



N

## Key Insights:

- ✓ **Growth Trends:** The line chart illustrates a notable rise in content added to Netflix from 2015 to around 2019, followed by a slight decline in 2020/2021. This shift could suggest that Netflix is prioritizing quality over quantity or facing challenges in content acquisition.
- ✓ **Seasonal Patterns:** Analyzing monthly data might uncover seasonal trends in content additions, potentially reflecting shifts in viewer behavior, if such data is available.
- ✓ **Content Library Expansion:** Despite the recent dip, Netflix continues to expand its content library, demonstrating a commitment to offering a broader selection to its audience.
- ✓ **Optimizing Content Strategy:** Gaining insights into potential seasonal patterns could inform more effective content release strategies, ensuring that new titles are launched at times that align with viewer demand.







# Distributing and analyzing the content rating by using Barplot

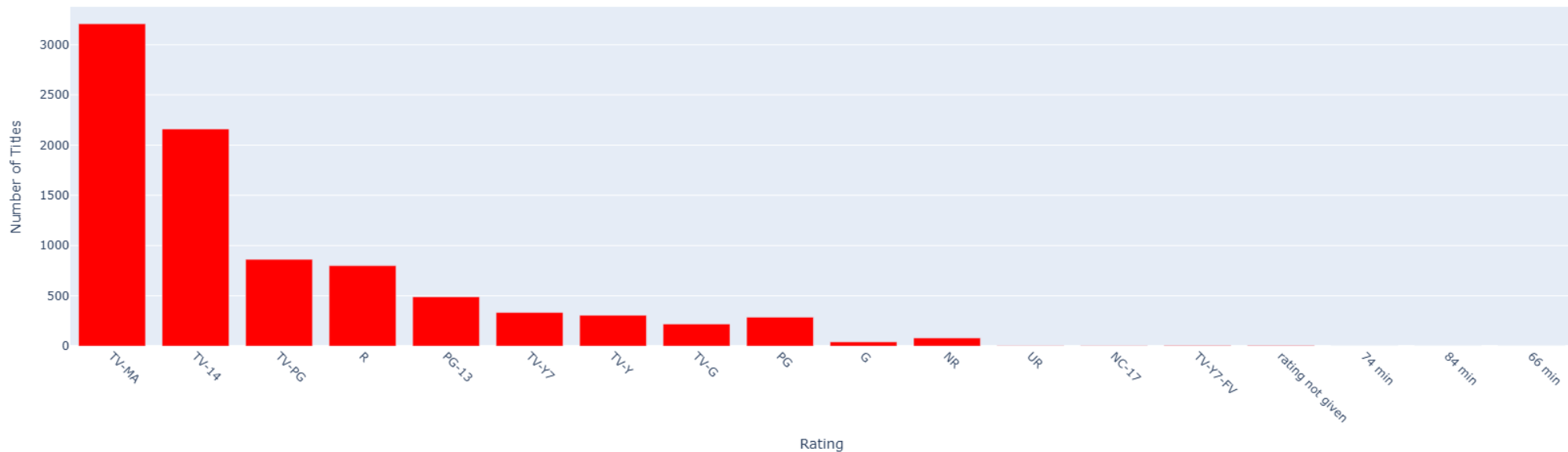
```
rating_counts = df['rating'].value_counts()
```

```
fig = px.bar(  
    x=rating_counts.index,  
    y=rating_counts.values,  
    labels={'x': 'Rating', 'y': 'Number of Titles'},  
    category_orders={'x': ['TV-MA', 'TV-14', 'TV-PG', 'R', 'PG-13', 'TV-Y7', 'TV-Y', 'TV-G', 'PG', 'G', 'NR', 'UR', 'NC-17']},  
    color_discrete_sequence=['red']  
)
```

```
fig.update_layout(title='Distribution of Content Ratings', xaxis_tickangle=45)  
fig.show()
```

## # Distribution of content rating

Distribution of Content Ratings



## Key Insights:

- ✓ **Dominance of Mature Content:** The bar chart highlights that a large portion of the content is rated TV-MA (mature audience) and TV-14 (parents strongly cautioned), suggesting a primary focus on adult and older teen demographics.
- ✓ **Family-Friendly Content:** There is a significant amount of content rated TV-PG, TV-Y7, and TV-Y, indicating that Netflix also caters to families and younger viewers.
- ✓ **Content Strategy:** The distribution of ratings reflects Netflix's strategy to serve a broad spectrum of audience preferences, ensuring content for various age groups and tastes.
- ✓ **Importance of Rating Distribution:** Understanding this distribution is essential for content creators and Netflix to tailor their content offerings to meet the needs and preferences of their target audiences.



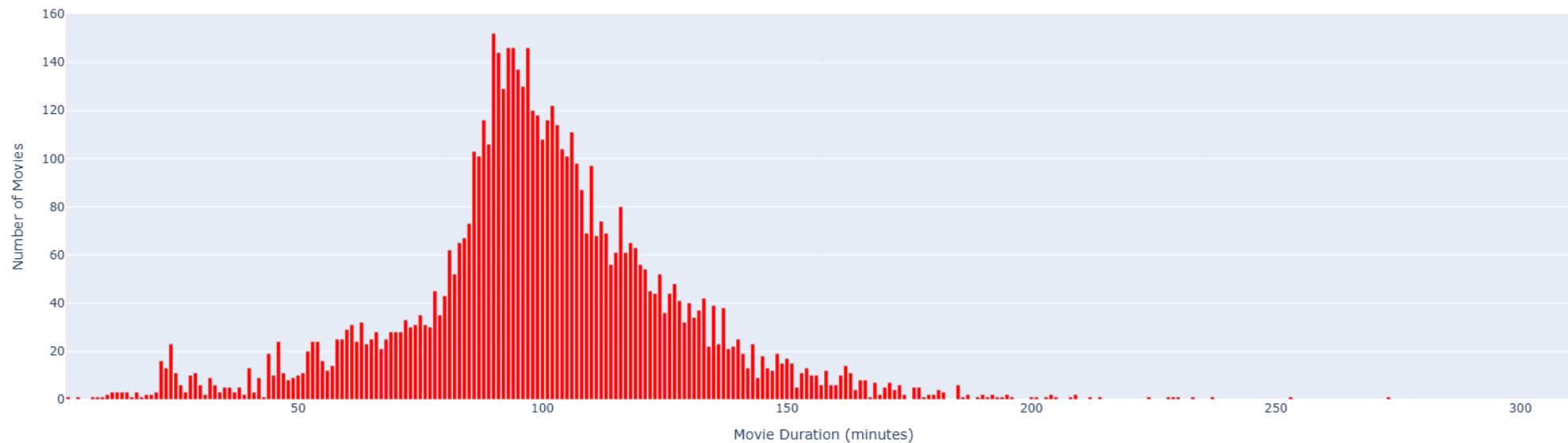


```
# Analyzing the duration of movies and TV shows to uncover any notable trends.
# Plotting bar plot for TV show durations
df['duration_type'] = df['duration'].str.extract('(\d+)').astype(float)
movies = df[df['type'] == 'Movie']
tv_shows = df[df['type'] == 'TV Show']

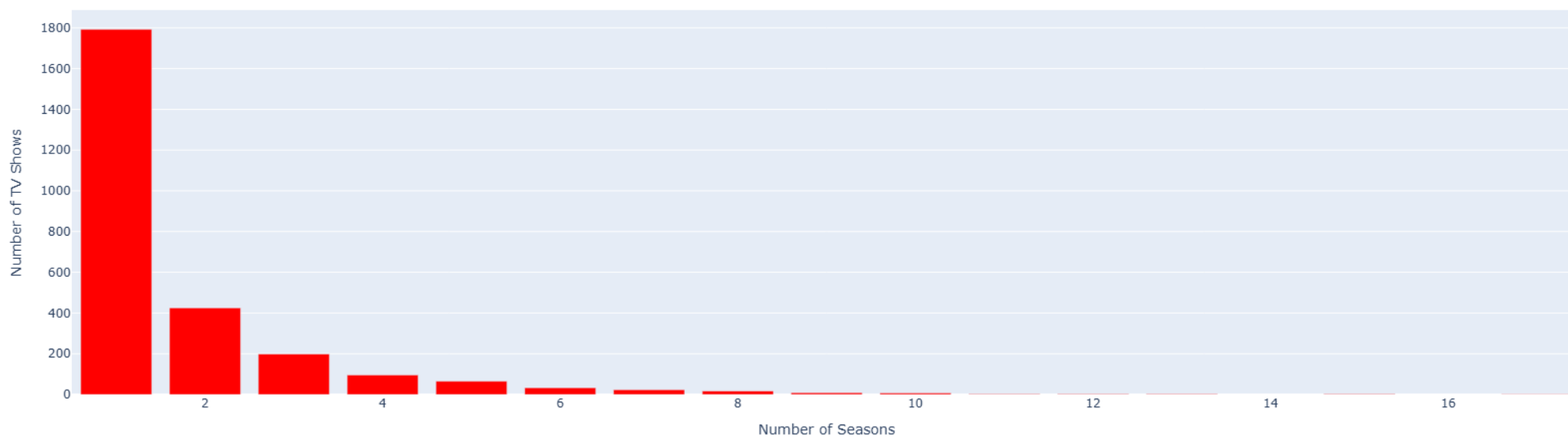
movies_duration_counts = movies['duration_type'].value_counts().sort_index()
fig_movies = px.bar(
    x=movies_duration_counts.index,
    y=movies_duration_counts.values,
    labels={'x': 'Movie Duration (minutes)', 'y': 'Number of Movies'},
    color_discrete_sequence=['red']
)
fig_movies.update_layout(title='Distribution of Movie Durations')
fig_movies.show()

# Plotting bar plot for TV show durations
tv_shows_duration_counts = tv_shows['duration_type'].value_counts().sort_index()
fig_tv_shows = px.bar(
    x=tv_shows_duration_counts.index,
    y=tv_shows_duration_counts.values,
    labels={'x': 'Number of Seasons', 'y': 'Number of TV Shows'},
    color_discrete_sequence=['red'] # Set bars to red
)
fig_tv_shows.update_layout(title='Distribution of TV Show Durations (Number of Seasons)')
fig_tv_shows.show()
```

Distribution of Movie Durations



Distribution of TV Show Durations (Number of Seasons)



## Key Insights:

- ✓ **Movie Duration:** The histogram for movies reveals a peak in content duration around 90-100 minutes, indicating a preference for standard feature film lengths. There is also a smaller yet notable number of movies under 90 minutes.
- ✓ **TV Show Seasons:** The histogram for TV shows shows that most series have 1-3 seasons, which may reflect the challenge of maintaining long-term viewer engagement.
- ✓ **Movie Duration Variety:** Netflix offers a wide range of movie lengths, catering to diverse viewer preferences. While most movies fall within the typical feature-length range, there are also options for those seeking shorter or longer viewing experiences.
- ✓ **TV Show Seasons:** The majority of TV shows on Netflix feature a limited number of seasons, likely due to the difficulty in sustaining audience interest over extended periods. This suggests a strategy of providing concise, impactful storytelling within a shorter timeframe.



```
[63] # Finding Top rated movies and TV shows based on user rating.
topRated_movies = df[df['type'] == 'Movie'].sort_values('rating', ascending=False).head(10)
topRated_tv_shows = df[df['type'] == 'TV Show'].sort_values('rating', ascending=False).head(10)

# printing TOP 10 movies with title and rating
print("Top 10 Rated Movies:")
print(topRated_movies[['title', 'rating']])
```

```
➡ Top 10 Rated Movies:
```

	title	rating
5989	13TH: A Conversation with Oprah Winfrey & Ava ...	rating not given
7537	My Honor Was Loyalty	rating not given
8790	You Don't Mess with the Zohan	UR
7058	Immoral Tales	UR
7988	Sex Doll	UR
7290	LEGO Ninjago: Masters of Spinjitzu: Day of the...	TV-Y7-FV
7292	Leo the Lion	TV-Y7-FV
7513	Motu Patlu: King of Kings	TV-Y7-FV
7317	Little Singham aur Kaal ka Mahajaal	TV-Y7-FV
6581	Dear Dracula	TV-Y7-FV

```
▶ # printing TOP 10 TV shows with title and rating
print("\nTop 10 Rated TV Shows:")
print(topRated_tv_shows[['title', 'rating']])
```

```
➡ Top 10 Rated TV Shows:
```

	title	rating
6827	Gargantia on the Verdurous Planet	rating not given
7312	Little Lunch	rating not given
7646	Oh No! It's an Alien Invasion	TV-Y7-FV
8803	Zombie Dumb	TV-Y7
7766	Power Rangers Lightspeed Rescue	TV-Y7
512	Code Lyoko	TV-Y7
7779	Power Rangers Super Samurai	TV-Y7
7777	Power Rangers Super Megaforce	TV-Y7
7773	Power Rangers Samurai	TV-Y7
7772	Power Rangers S.P.D.	TV-Y7

# TOP 10  
Movies  
and TOP  
10 TV  
Shows  
based on  
user  
rating

N

## Top 10 Rated Movies:

title	rating
13TH: A Conversation with Oprah Winfrey & Ava	rating not given
My Honor Was Loyalty	rating not given
You Don't Mess with the Zohan	UR
Immoral Tales	UR
Sex Doll	UR
LEGO Ninjago: Masters of Spinjitzu: Day of the...	TV-Y7-FV
Leo the Lion	TV-Y7-FV
Motu Patlu: King of Kings	TV-Y7-FV
Little Singham aur Kaal ka Mahajaal	TV-Y7-FV
Dear Dracula	TV-Y7-FV



## Top 10 Rated TV Shows:

title	rating
Gargantia on the Verdurous Planet	rating not given
Little Lunch	rating not given
Oh No! It's an Alien Invasion	TV-Y7-FV
Zombie Dumb	TV-Y7
Power Rangers Lightspeed Rescue	TV-Y7
Code Lyoko	TV-Y7
Power Rangers Super Samurai	TV-Y7
Power Rangers Super Megaforce	TV-Y7
Power Rangers Samurai	TV-Y7
Power Rangers S.P.D.	TV-Y7

# ➤ Data Visualization and Key Insights

```
▶ # analyzing the trends in the popularity of diffrents genres over time.

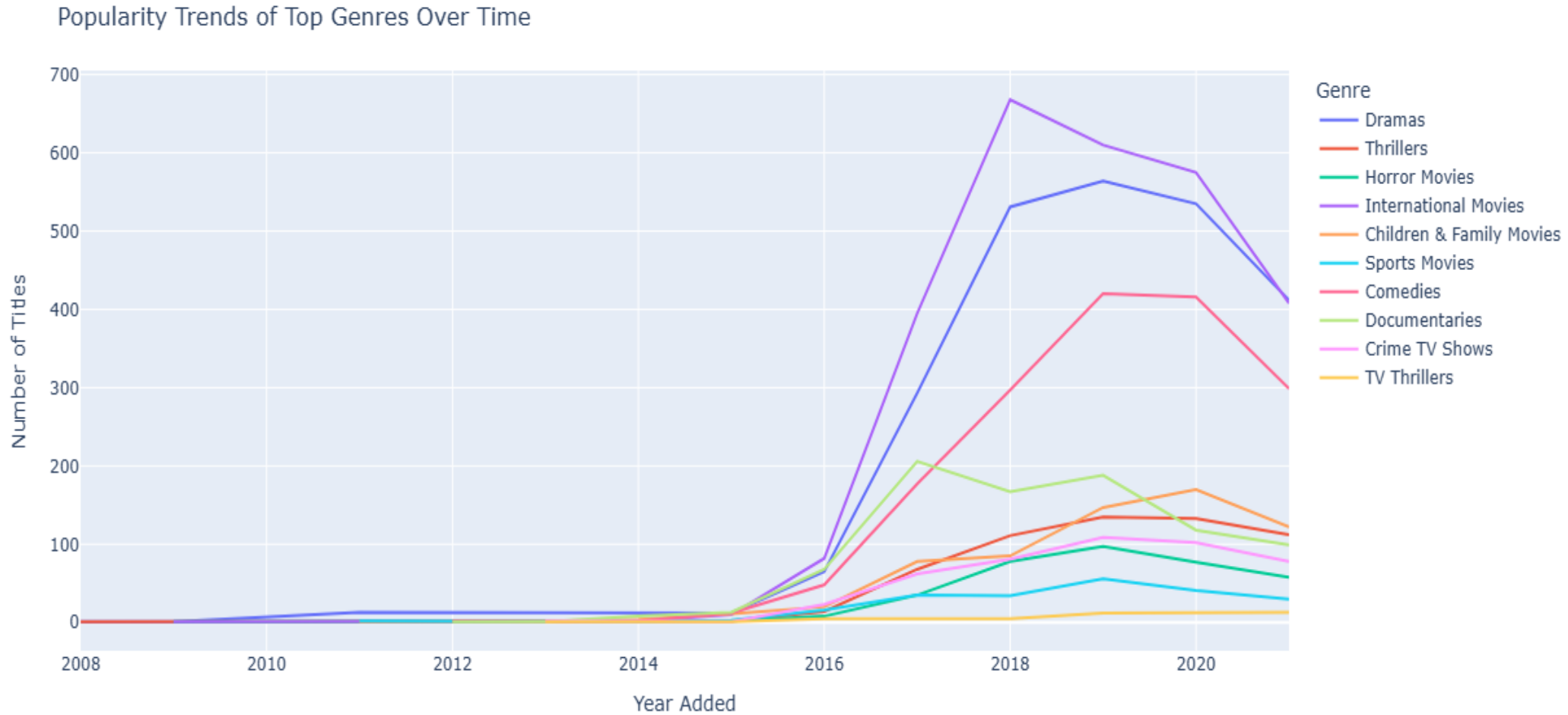
df['year_added'] = pd.to_datetime(df['date_added'], errors='coerce').dt.year
df['genres'] = df['listed_in'].str.split(', ')

# Counting the number of occurrences by year and genre.
genre_trends = df.explode('genres').groupby(['year_added', 'genres'])['show_id'].count().reset_index(name='count')

# Selecting top 10 genres for better visualization.
top_genres = genre_trends['genres'].value_counts().head(10).index.tolist()
genre_trends_top = genre_trends[genre_trends['genres'].isin(top_genres)]

# Plotting line chart using Plotly to show trends for top genres.
fig = px.line(genre_trends_top, x='year_added', y='count', color='genres',
              labels={'year_added': 'Year Added', 'count': 'Number of Titles', 'genres': 'Genre'})
fig.update_layout(title='Popularity Trends of Top Genres Over Time')
fig.show()
```

## # Popularity trends of TOP genres over time



## Key Insights:

**1.Overall Trend:** The graph shows a general upward trend in the number of titles added for most genres over the years 2008 to 2020. This suggests a growing library of content across different genres.

### 2.Genre Growth:

- Dramas** consistently have the highest number of titles added throughout the period, with a significant surge in the late 2010s.
- Comedies** also show a strong upward trend, with a peak around 2018.
- International Movies** and **Horror Movies** have seen a rapid increase in recent years, indicating a growing popularity.
- Children & Family Movies** have a relatively stable number of titles added.
- Sports Movies** and **TV Thrillers** have a less pronounced increase compared to other genres.

### 3.Genre Fluctuations:

- Some genres, like **Thrillers** and **Crime TV Shows**, exhibit fluctuations in their growth rates, with periods of rapid increase followed by slower growth or even decline.

## ✓ Possible Interpretations:

- **Changing Audience Preferences:** The increasing popularity of International Movies and Horror Movies could reflect a shift in viewer preferences towards diverse and more thrilling content.
- **Platform Strategy:** The platform might be strategically focusing on adding more titles in genres like Dramas and Comedies to cater to a wider audience.
- **Production Trends:** The growth in certain genres could be due to increased production and release of titles in those categories.

## ✓ Further Analysis:

- To gain deeper insights, it would be helpful to analyze the data for individual titles within each genre to understand their popularity and trends.
- Comparing this data with viewership or subscription data could provide a more comprehensive understanding of audience engagement with different genres.

## ➤ Data Visualization and Key Insights

```
# exploring the distribution of content across different countries and regions by creating piechart.

country_counts = df['country'].str.split(', ').explode().value_counts()

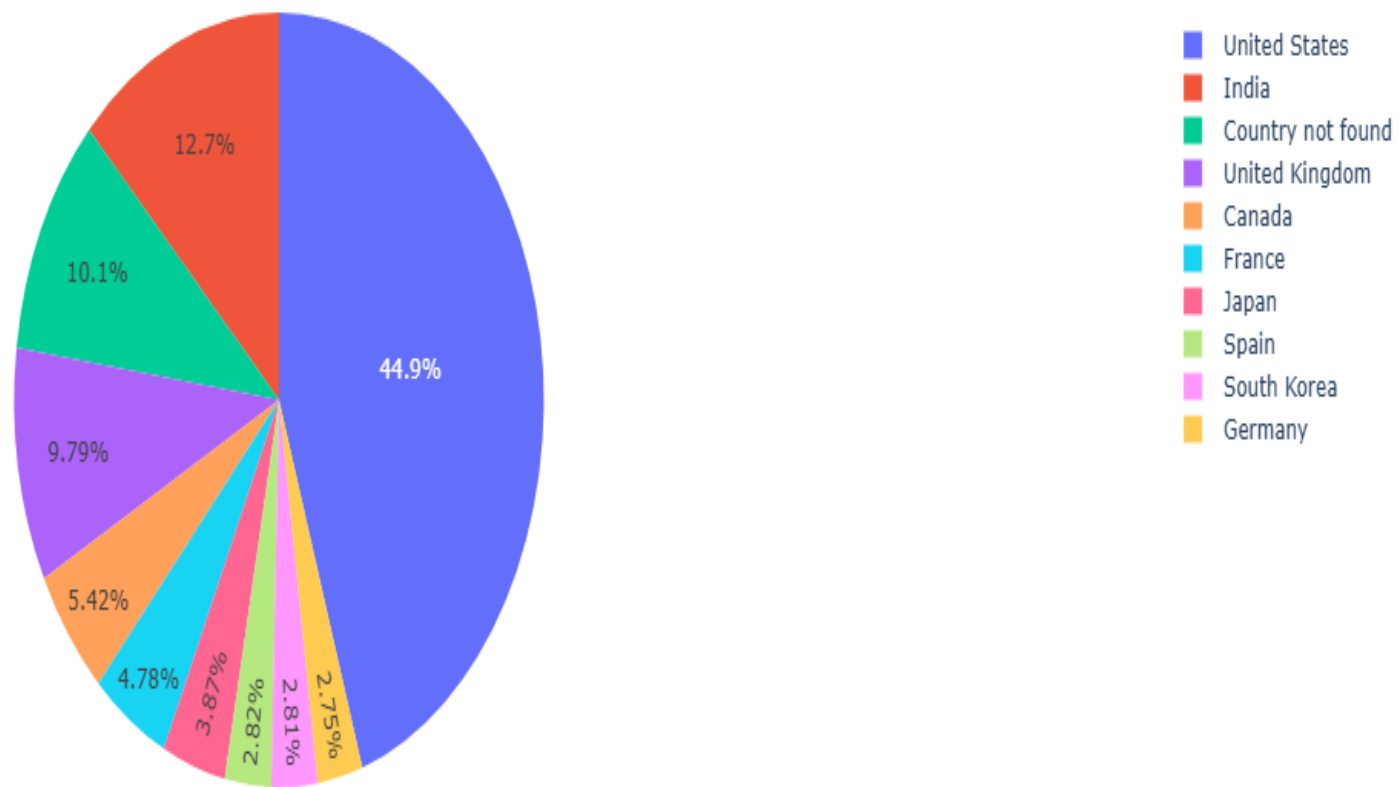
# Selecting top 10 countries for better visualization in pie chart.
top_10_countries = country_counts.head(10)
fig = px.pie(values=top_10_countries.values, names=top_10_countries.index, title='Geographical Distribution of Content (Top 10 Countries)')
fig.show()

# Now Grouping countries into Regions.
region_mapping = {'United States':'North America', 'India':'Asia', 'United Kingdom':'Europe'}
df['region'] = df['country'].map(region_mapping)
region_counts = df['region'].value_counts()

# Plotting pie chart for regional distribution
fig = px.pie(values=region_counts.values, names=region_counts.index, title='Regional Distribution of Content')
fig.show()
```

## #Geographical Distribution of content (TOP 10 countries)

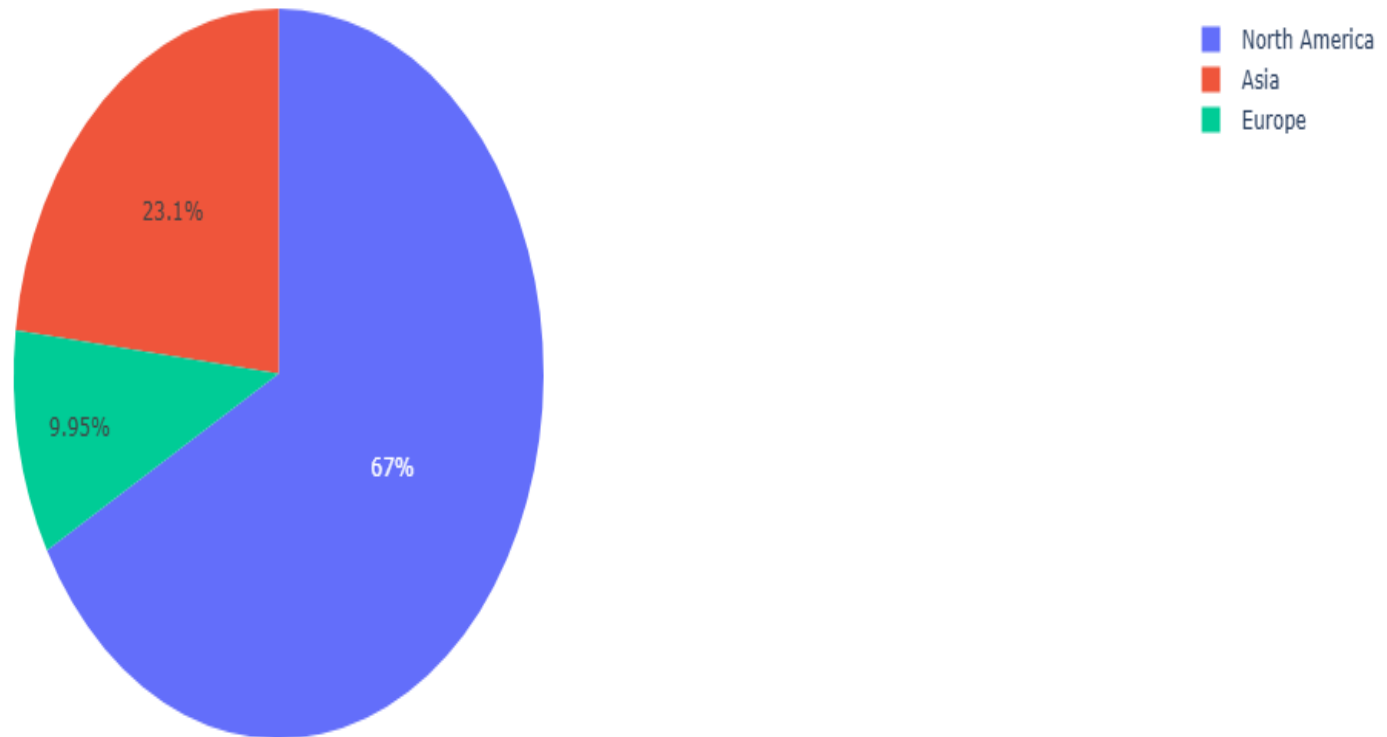
Geographical Distribution of Content (Top 10 Countries)





## # Regional Distribution of content

Regional Distribution of Content



## #Geographical Distribution of content (TOP 10 countries)

### Key Insights:

- ✓ **Dominance of the United States:** The chart shows a significant portion of the content (44.9%) originates from the United States. This suggests a strong presence and influence of US-based content creators or platforms.
- ✓ **India as a Major Contributor:** India follows closely with 12.7% of the content share, indicating a substantial contribution from Indian creators or platforms.
- ✓ **United Kingdom's Presence:** The United Kingdom holds the third position with 10.1% of the content share, showcasing its significant role in content creation.
- ✓ **Other Notable Countries:** Canada, France, Japan, South Korea, Spain, and Germany each contribute a smaller but noticeable percentage to the overall content distribution.

## # Regional Distribution of content

### Key Insights:

- ❑ **Europe:** 67% of the content comes from Europe.
- ✓ **European Dominance:** The chart clearly shows that Europe is the leading region in terms of content creation, accounting for a substantial 67% of the total.
- ❑ **North America:** 23.1% of the content originates from North America.
- ✓ **North America's Share:** North America holds the second-largest share with 23.1% of the content.
- ❑ **Asia:** 9.95% of the content is created in Asia.
- ✓ **Asia's Contribution:** Asia has the smallest share among the three regions, contributing 9.95% to the overall content.



# investigating the potential correlations between variables by using scatterplot.

```
plt.figure(figsize=(10, 5))
```

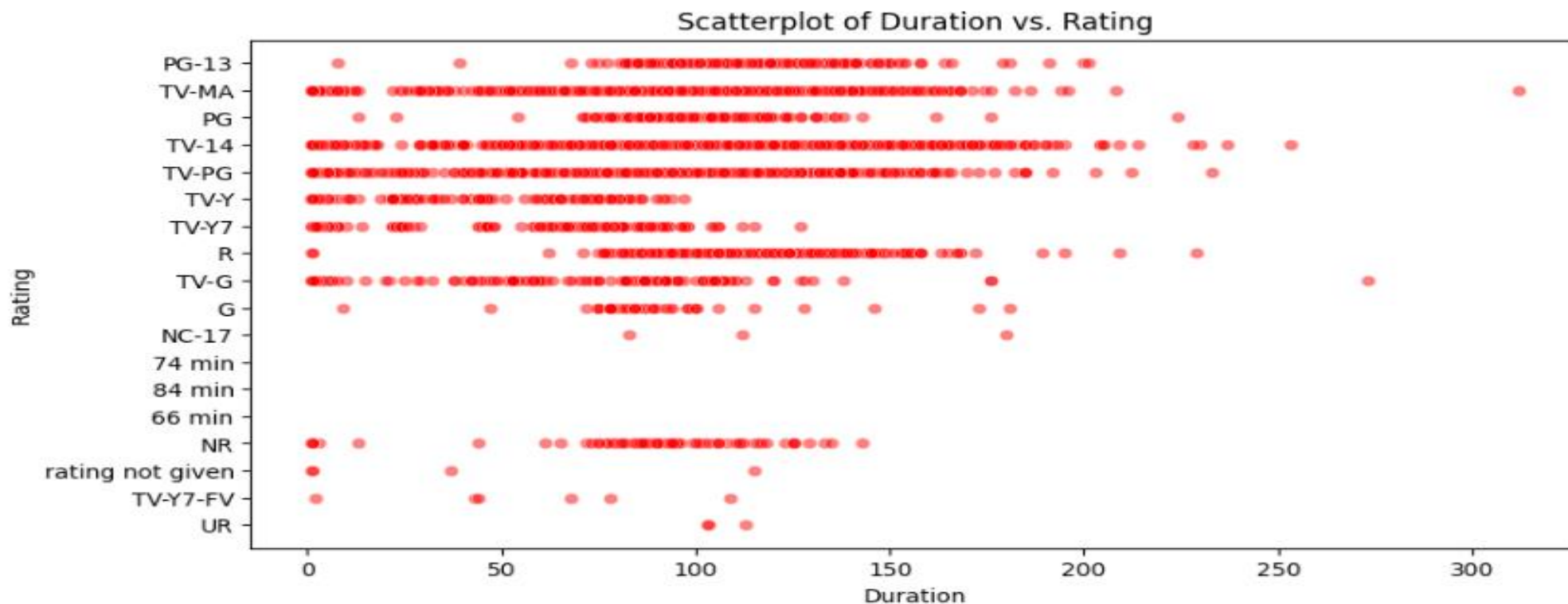
```
sns.scatterplot(data=df, x='duration_type', y='rating', color='red', alpha=0.5) # Set color to red
```

```
plt.title('Scatterplot of Duration vs. Rating')
```

```
plt.xlabel('Duration')
```

```
plt.ylabel('Rating')
```

```
plt.show()
```



## ➤ Data Visualization and Key Insights

### Key Insights:

- **No Strong Correlation:** The scatterplot indicates no clear linear relationship between duration and rating, suggesting that the length of a movie or TV show does not significantly influence its rating.
- **Scope for Further Analysis:** To uncover deeper insights, we could investigate correlations between ratings and other variables such as genre, release year, or country of origin. Statistical tests could also help measure the strength of any potential relationships.
- **Rating Distribution Across Durations:** Movies and TV shows of varying durations exhibit a broad range of ratings, highlighting that factors beyond duration are more influential in determining user ratings.

## ➤ Data Visualization and Key Insights

```
[ ] # Counting the unique genres and categories to measure content diversity.
unique_genres = df['listed_in'].str.split(', ').explode().unique()
num_unique_genres = len(unique_genres)
print("Number of unique genres:", num_unique_genres)

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

- ✓ Number of unique genres are 42
- ✓ Number of unique categories including genres are also 42

## ➤ Data Visualization and Key Insights

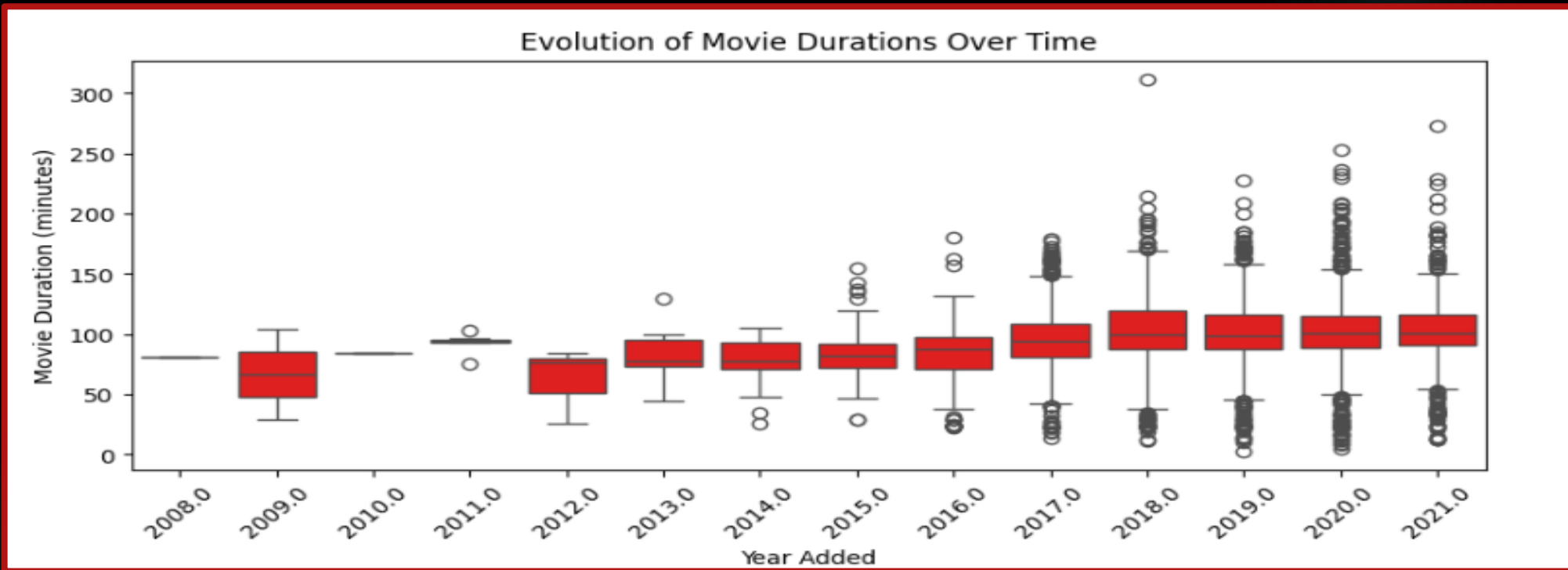


```
# Exploring the characteristics of content ("duration","rating ") have evolved over the years.

# Boxplot of evolution of movie duration over time
plt.figure(figsize=(10, 4))
sns.boxplot(data=movies, x='year_added', y='duration_type', color='red')
plt.title('Evolution of Movie Durations Over Time')
plt.xlabel('Year Added')
plt.ylabel('Movie Duration (minutes)')
plt.xticks(rotation=45)
plt.show()

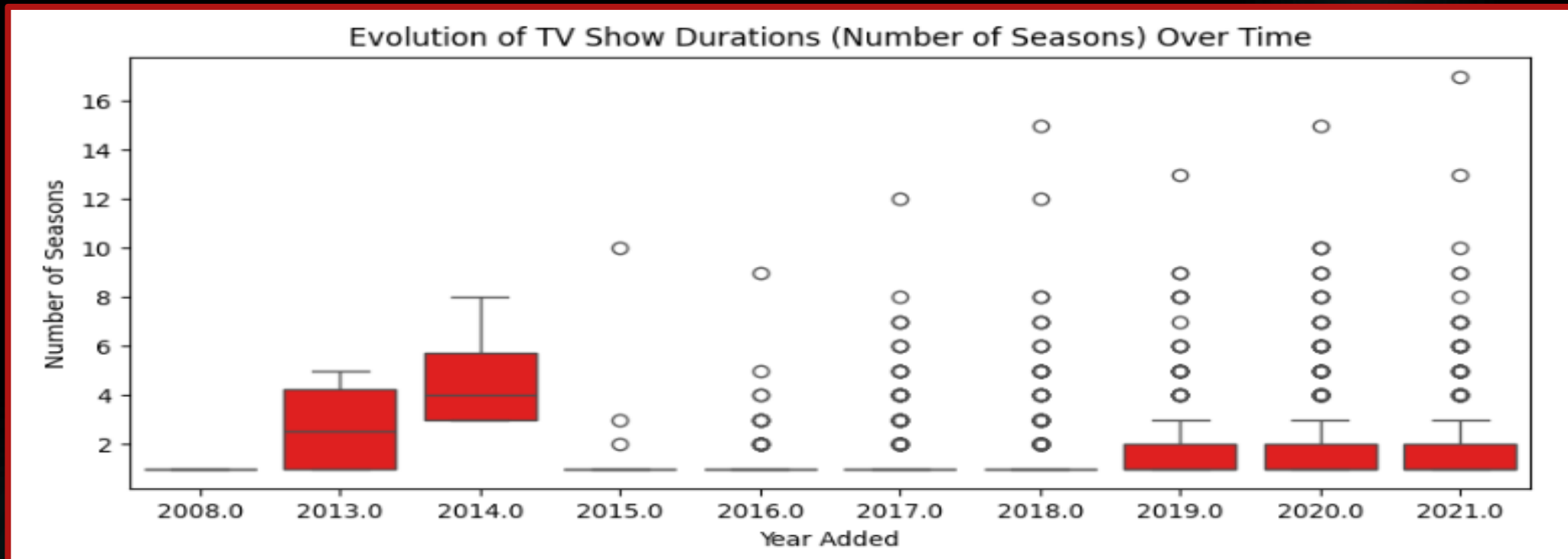
# Boxplot of evolution of TV shows duration over time
plt.figure(figsize=(10, 4))
sns.boxplot(data=tv_shows, x='year_added', y='duration_type', color='red')
plt.title('Evolution of TV Show Durations (Number of Seasons) Over Time')
plt.xlabel('Year Added')
plt.ylabel('Number of Seasons')

# lineplot of evolution of content rating over time
rating_trends = df.groupby(['year_added', 'rating'])['show_id'].count().reset_index(name='count')
plt.figure(figsize=(10, 4))
sns.lineplot(data=rating_trends, x='year_added', y='count', hue='rating')
plt.title('Evolution of Content Ratings Over Time')
plt.xlabel('Year Added')
plt.ylabel('Number of Titles')
plt.xticks(rotation=45)
plt.show()
```



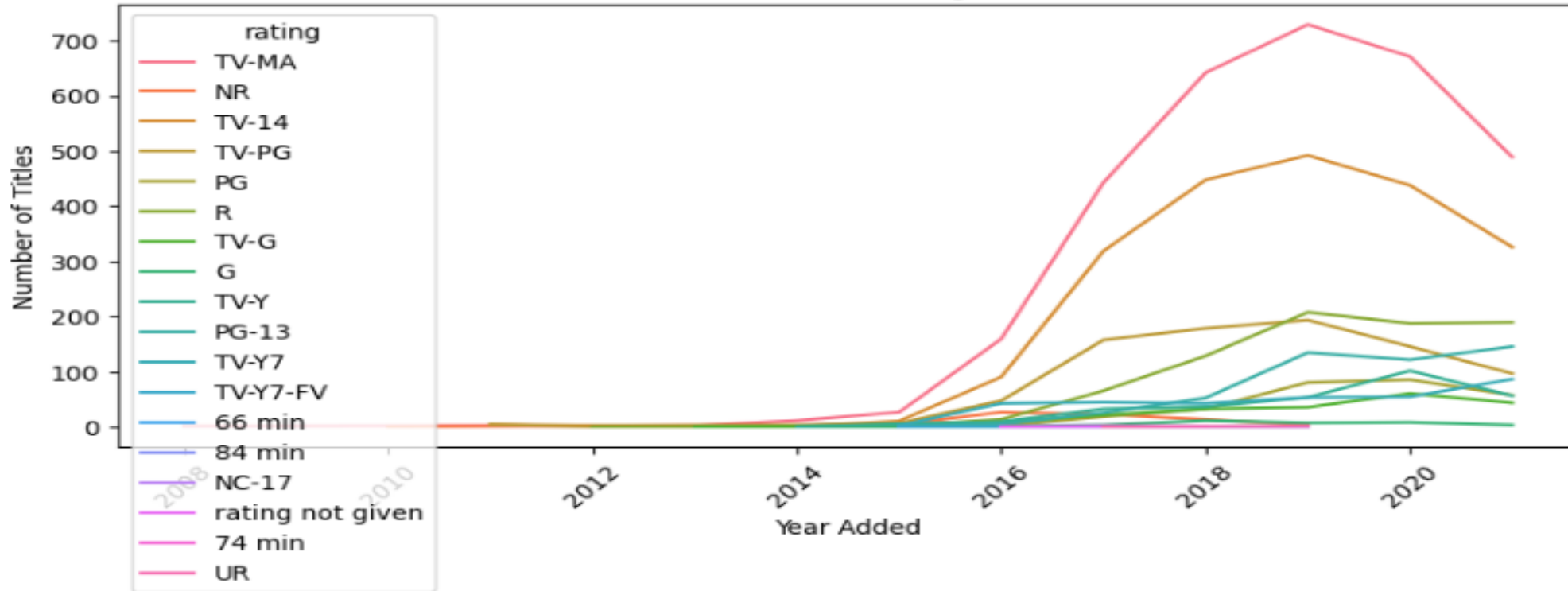
- ✓ **Median Duration:** The red line within each box represents the median duration for that year. We can see that the median duration has increased over time.
- ✓ **Spread of Durations:** The box itself shows the interquartile range (IQR), which represents the middle 50% of movie durations. The width of the box gives an idea of the spread of durations within each year.
- ✓ **Outliers:** The dots above and below the whiskers represent outliers, which are movies with unusually long or short durations compared to the rest of the movies in that year.





- ✓ **Median Increase:** The red line within each box represents the median number of seasons for that year. We can see a clear upward trend in the median value over time.
- ✓ **Spread of Durations:** The box itself shows the interquartile range (IQR), which represents the middle 50% of show durations. The width of the box gives an idea of the spread of durations within each year.
- ✓ **Outliers:** The dots above and below the whiskers represent outliers, which are shows with unusually long or short durations compared to the rest of the shows in that year.

Evolution of Content Ratings Over Time



- TV-MA:** This rating (likely for mature audiences) shows the most dramatic increase, with a steep rise, particularly between 2016 and 2019. It suggests a growing trend of content created for adult viewers.

- TV-14 and NR:** These ratings also show a significant increase, indicating a rise in content suitable for older teens and general audiences.

- Other Ratings:** Ratings like **PG**, **R**, and **TV-PG** show a moderate increase, while ratings like **G**, **TV-Y**, and **PG-13** show a relatively slower growth.

# FINAL REPORT

The analysis highlights key aspects of Netflix's content strategy and library. Dominant genres like "International Movies" and "Dramas" lead the platform's offerings, with content additions peaking around 2019-2020. Major contributors to the library include the United States, India, and the UK, while a significant portion of content is rated TV-MA and TV-14, reflecting a focus on mature and older teen audiences. Trends show rapid content growth until 2019, followed by a slight decline in 2020-2021, indicating potential strategic shifts. Most movies align with standard durations of 90-100 minutes, while the majority of TV shows consist of 1-3 seasons, demonstrating a preference for concise storytelling. A steady rise in "International Movies" and "Dramas" reflects evolving viewer preferences, with North America, India, and Europe dominating content sources. Netflix's library offers diversity, with many unique genres and categories. However, no clear correlation between duration and ratings suggests other factors influence viewer engagement. Data also reveals an increase in shorter movies and variety in TV show durations. Recommendations include analyzing correlations between ratings and factors like genre, leveraging user feedback to enhance personalization, and adapting content strategies to industry trends to align with audience demands and preferences effectively.



# CONCLUSION

Netflix caters to a global audience by offering diverse content, with a primary focus on adult and older teen demographics. The platform continuously expands its content library, emphasizing recent releases to keep its offerings fresh and relevant. Strategic attention is directed towards key markets like the United States, India, and the United Kingdom. Netflix's content strategy reflects adaptability to evolving viewer preferences, with a noticeable shift toward shorter formats that cater to changing consumption habits. To enhance user experience, Netflix can invest in improving its search and discovery features, enabling users to find content that aligns more closely with their interests. Regular reviews of the content library will ensure that the platform continues to meet users' evolving tastes and preferences. Additionally, deeper analysis of user data and viewing patterns can provide valuable insights into audience behavior, enabling Netflix to refine its content recommendations and maintain its competitive edge in the streaming industry.



Exploratory  
Data  
analysis of  
Netflix

# NETFLIX

# Thank You

Made by Lucky Bisht

Gmail: [Bishtlucky543@gmail.com](mailto:Bishtlucky543@gmail.com)

N