# NETFILIX

# EXPLORATORY DATA ANALYSIS CAPSTONE PROJECT

# INTRODUCTION

This project involves analyzing a Netflix dataset using the Google Colab environment. The dataset contains information about movies and TV shows available on Netflix, including titles, release years, ratings, genres, and more.

The analysis aims to uncover insights into various aspects of the dataset such as:

- **Content Distribution:** Explore the distribution of content across genres, release years, ratings, and countries.
- **Trends:** Identify trends in content release over time and analyze the popularity of different genres and ratings.
- **Diversity:** Evaluate the diversity of content based on genre and country.
- **Potential correlations:** investigate relationships between variables like duration and rating.
- **Language analysis:** understand the language distribution of Netflix content.

# OBJECTIVE

1. **Data Cleaning and Preprocessing:** The code starts by cleaning the data by removing duplicates and handling missing values. This ensures that the analysis is based on accurate and reliable data.
2. **Content Distribution:** The code explores the distribution of content across different genres, release years, countries, and ratings. This provides insights into the types of content available on Netflix and their popularity.
3. **Content Trends:** The analysis looks at trends in the popularity of genres over time. This can help understand the evolution of content preferences and identify emerging trends.

# OBJECTIVE

**4.Content Length and Ratings:** The code explores the relationship between content duration and ratings. This provides insights into user preferences for content length and the relationship with ratings.

**5.User Reviews and Sentiment:** The code also tries to analyze user sentiment based on descriptions if available, helping understand user feedback and preferences.

**6.Content Diversity:** The code evaluates the variety of content by examining the number of unique genres and categories. This provides insights into the range of content offered by Netflix.

# IMPORT PYTHON LYBRARY

## NumPy

NumPy (Numerical Python) is a powerful library for working with arrays and numerical data.

It is used for mathematical operations, statistical analysis, and handling large datasets efficiently.

## Pandas

Pandas is a Python library used for data analysis and manipulation.

It is mainly used for working with structured data like tables, spreadsheets, or databases.

It provides Data Frames, which are like Excel tables but with much more flexibility and power

## Seaborn

Seaborn is a data visualization library used to create beautiful and professional charts.

It allows you to create Bar Charts, Line Charts, Scatter Plots, Box Plots, and more.

Seaborn makes data representation easier to understand.

## Matplotlib

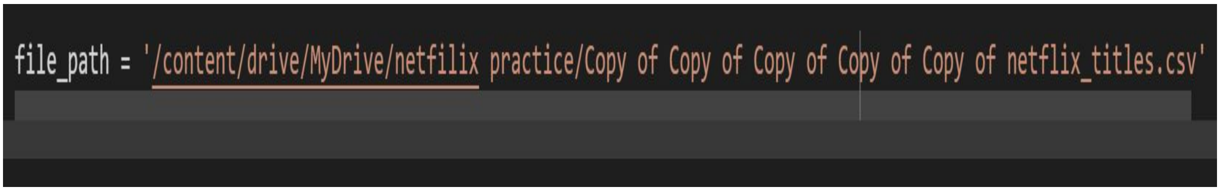
Matplotlib is another charting library but gives more customization options.

It is often used with Seaborn to modify or save graphs.

Seaborn internally uses Matplotlib to generate graphs.

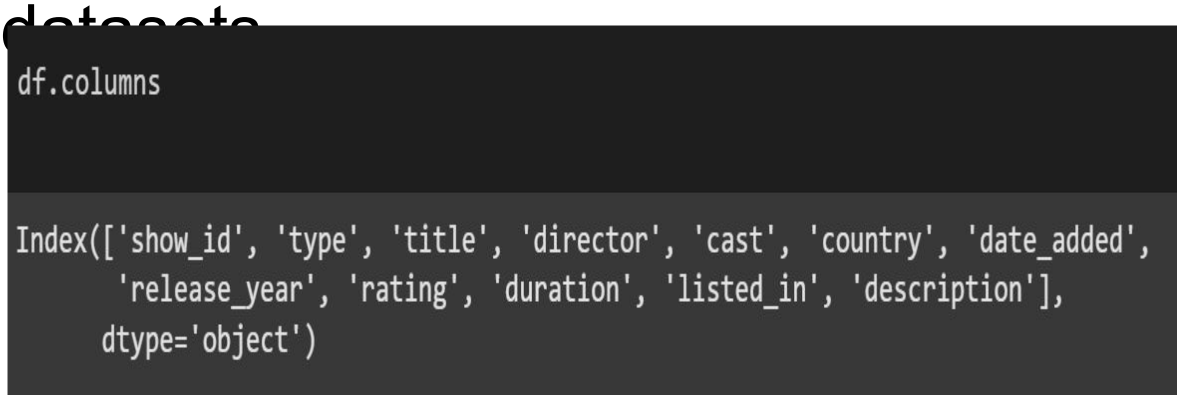
LOADING DATASET

# Loading Dataset



```
file_path = '/content/drive/MyDrive/netfilix practice/Copy of Copy of Copy of Copy of Copy of netflix_titles.csv'
```

I've load dataset  named ‘netfilix_titles.csv’ into a variable name ‘df’ in order to import  the datasets, I’ve  used pandas Libraries in which I’ve used ‘pd.read_csv’ command to import the respective datasets.



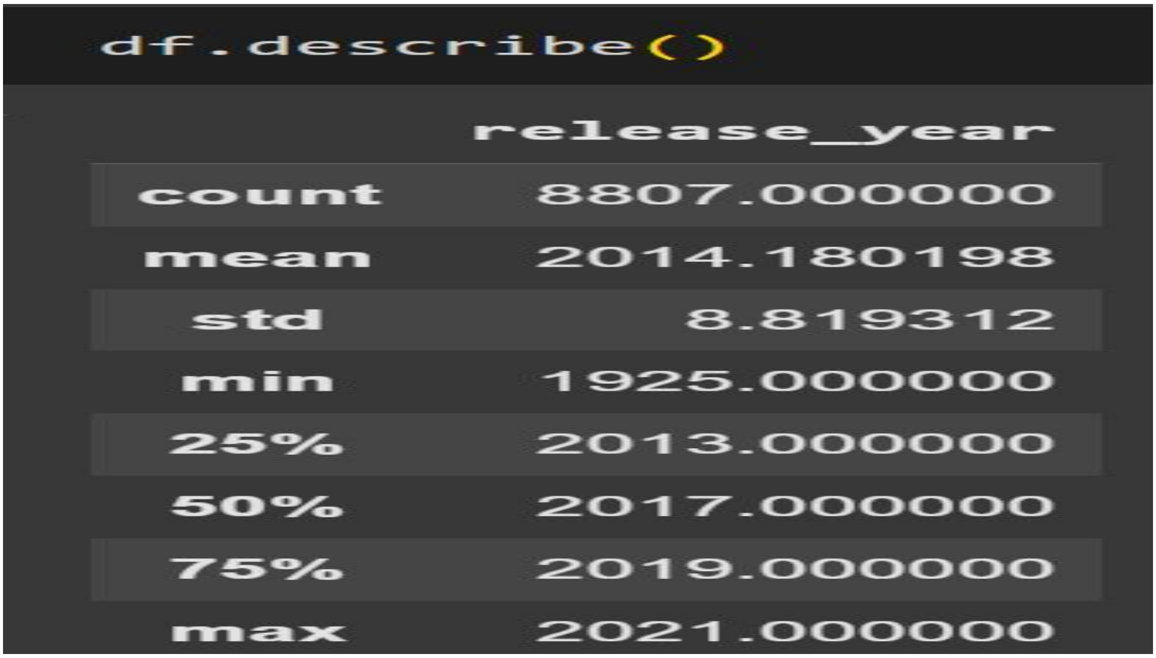
```
df.columns
```

```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

I’ve accessed and displayed the column names of the loaded dataset using  columns. These columns  encompass  various aspects of ‘Indian restaurants’ information, providing  detailed overview of the dataset.

DATA OVERVIEW

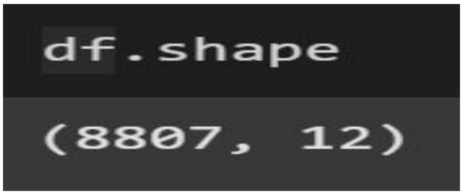
# Basic Composition of data



```
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
```

I’ve generated descriptive stastic for the loaded dataset stored in the variable ‘df’ using the describe method. This Pandas function provides a summary of stastical measure , minimum , 25th percentile, median(50th percentile), 75th percentile and maximum values for each numeric column in the dataset . This summary aids in in understanding the central tendency ,dispersion , and distribution of the numeric feature in the dataset.



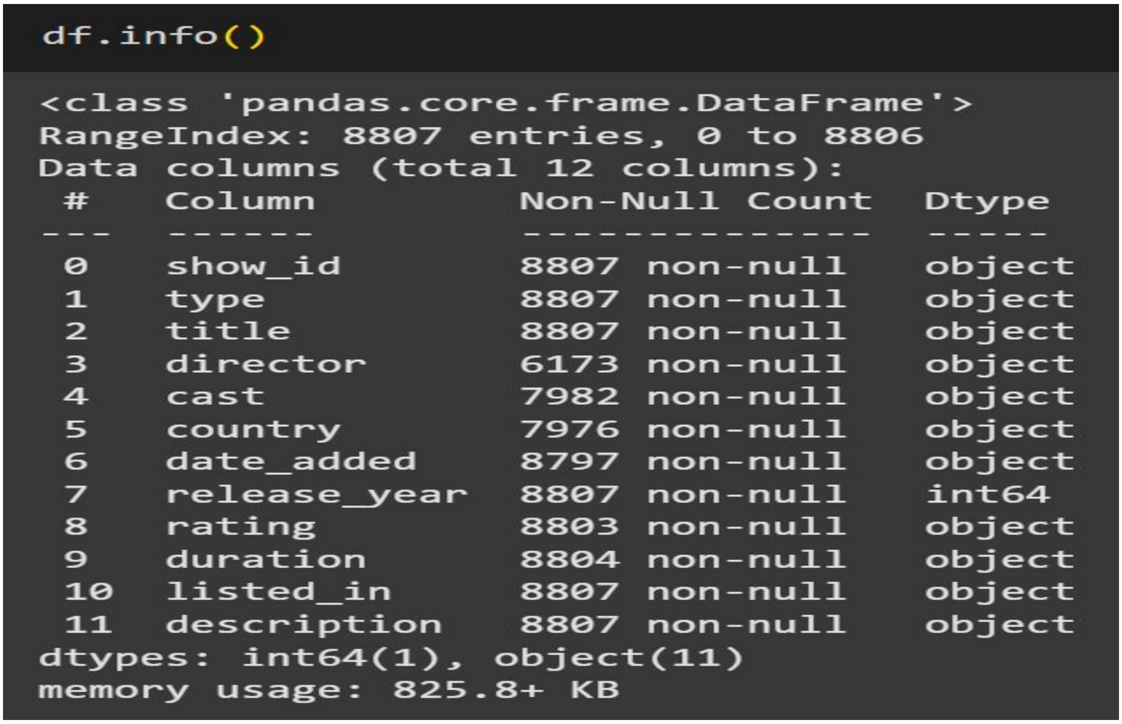
```
df.shape

(8807, 12)
```

df. Shape , revealing that it consists of 8807rows and 12 column

DATA OVERVIEW

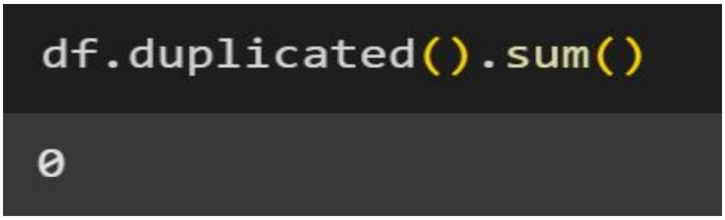
## Checking  Information



```
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
```

I’ve  obtained information about the dataset  using df.info(), This method  provide a  concise summary , including the total number of entries ,the data types of each column ,and the count of non-nun values



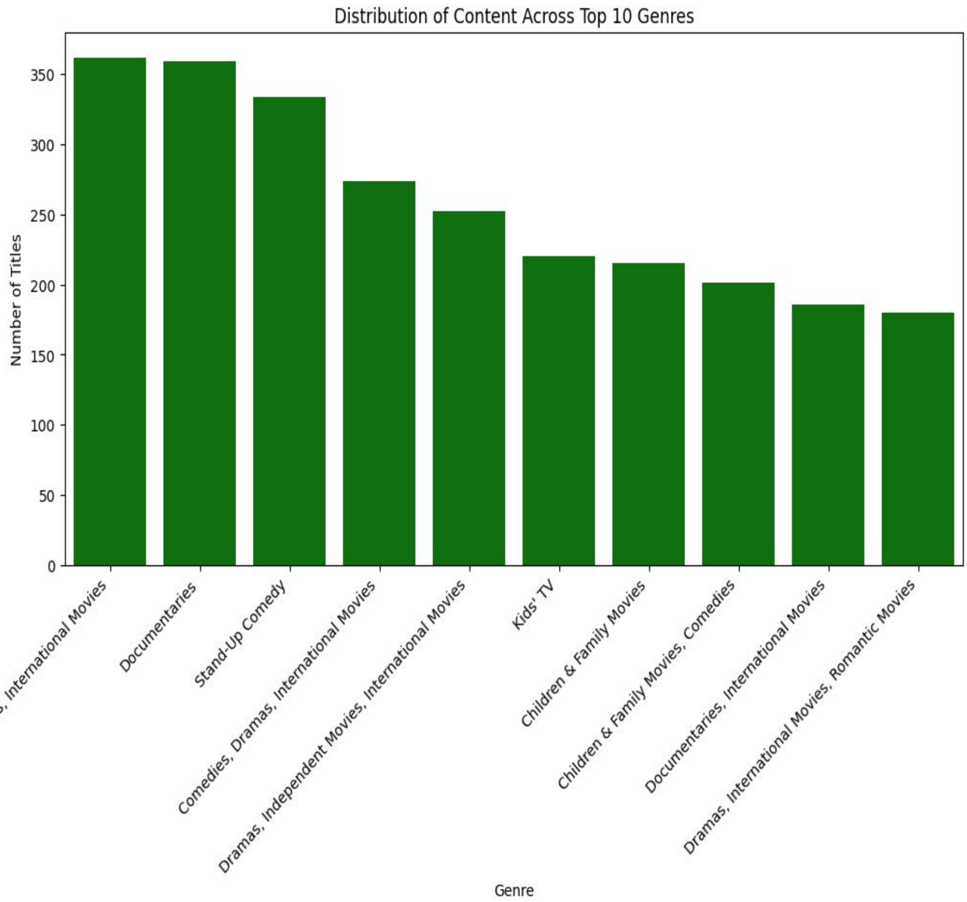
```
df.duplicated().sum()
```

```
0
```

Check my dataset no duplicate value present

DATA VISUALIZATION

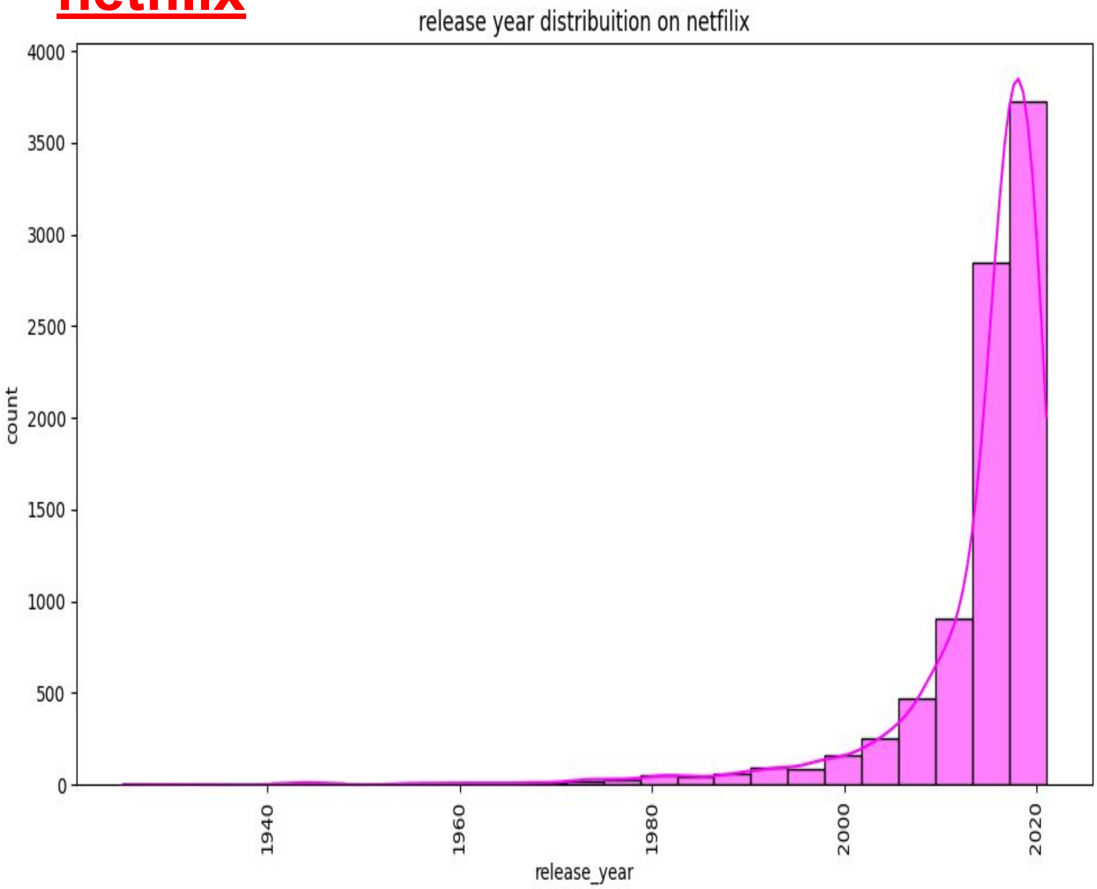
# Distribuition of Content Across Top 10 Genres



- **Top Genres:** Dramas & Documentaries lead.
- **Comedy Popular:** Stand-Up & Comedies rank high.
- **Family Content:** Kids TV & Family Movies present.
- **Mixed Genres:** Many genres overlap.
- **Trend:** Dramas & International Movies dominate.

DATA VISUALIZATION

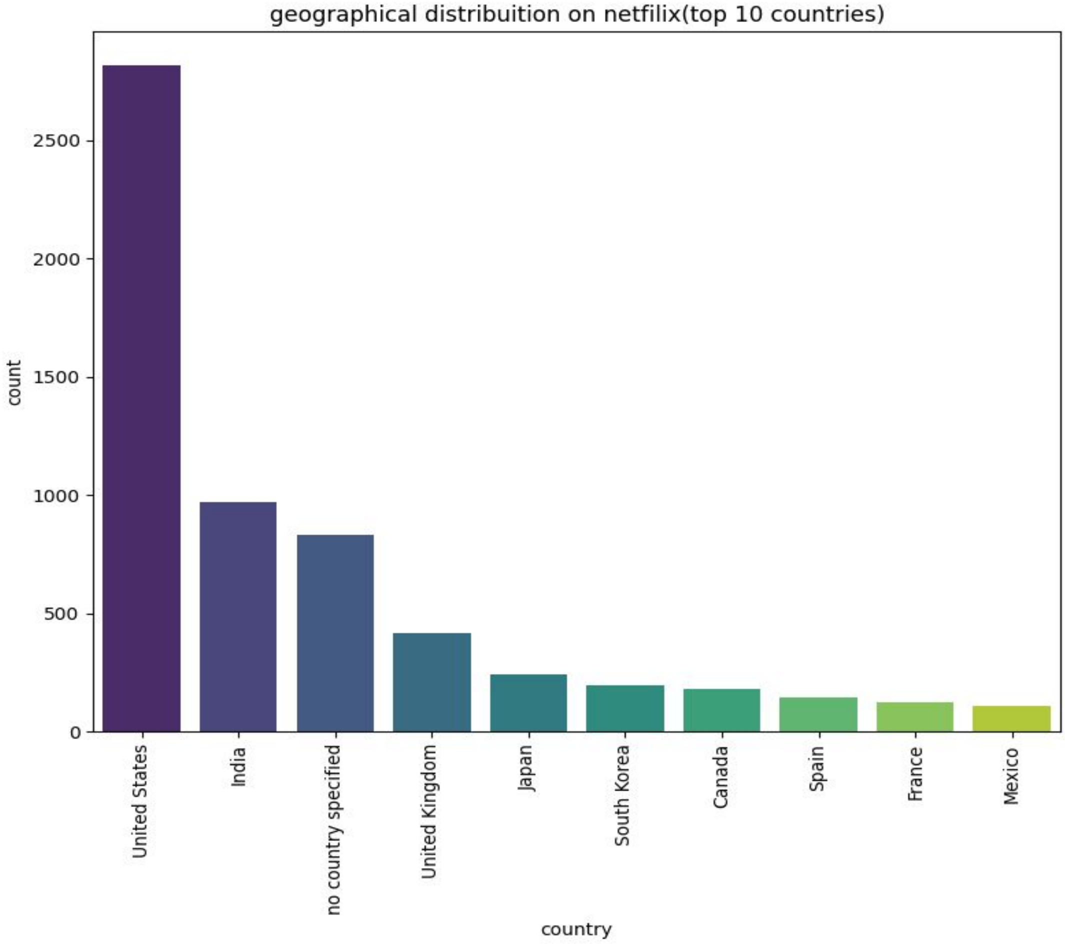
## Release year distribuition on netfilix



- **Recent Content Dominates** – Most titles are from the 2000s, especially after 2015.
- **Steady Growth** – The number of releases gradually increased over time.
- **Peak in 2020** – The highest number of titles were released around 2020.
- **Old Content is Rare** – Very few titles are from before the 1980s.
- **Sharp Drop After 2020** – The number of new releases decreases post-2020.

DATA VISUALIZATION

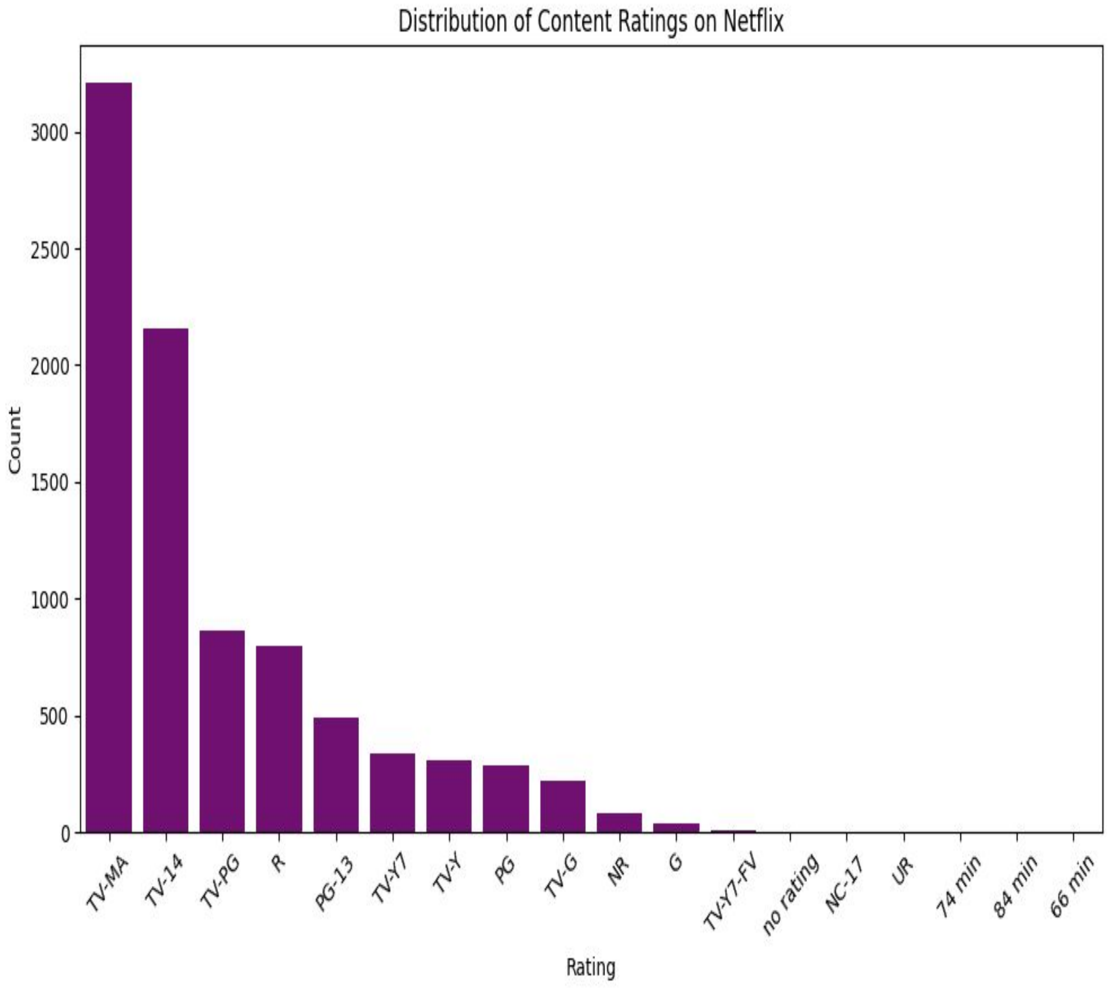
# Geographical distribuition on Netfilix 10 Countries



- **United States Dominates –** Highest content count.
- **India & No Country Specified –** Significant share.
- **United Kingdom & Japan –** Moderate presence.
- **Other Countries (South Korea, Canada, Spain, France, Mexico) –** Lower count.

DATA VISUALIZATION

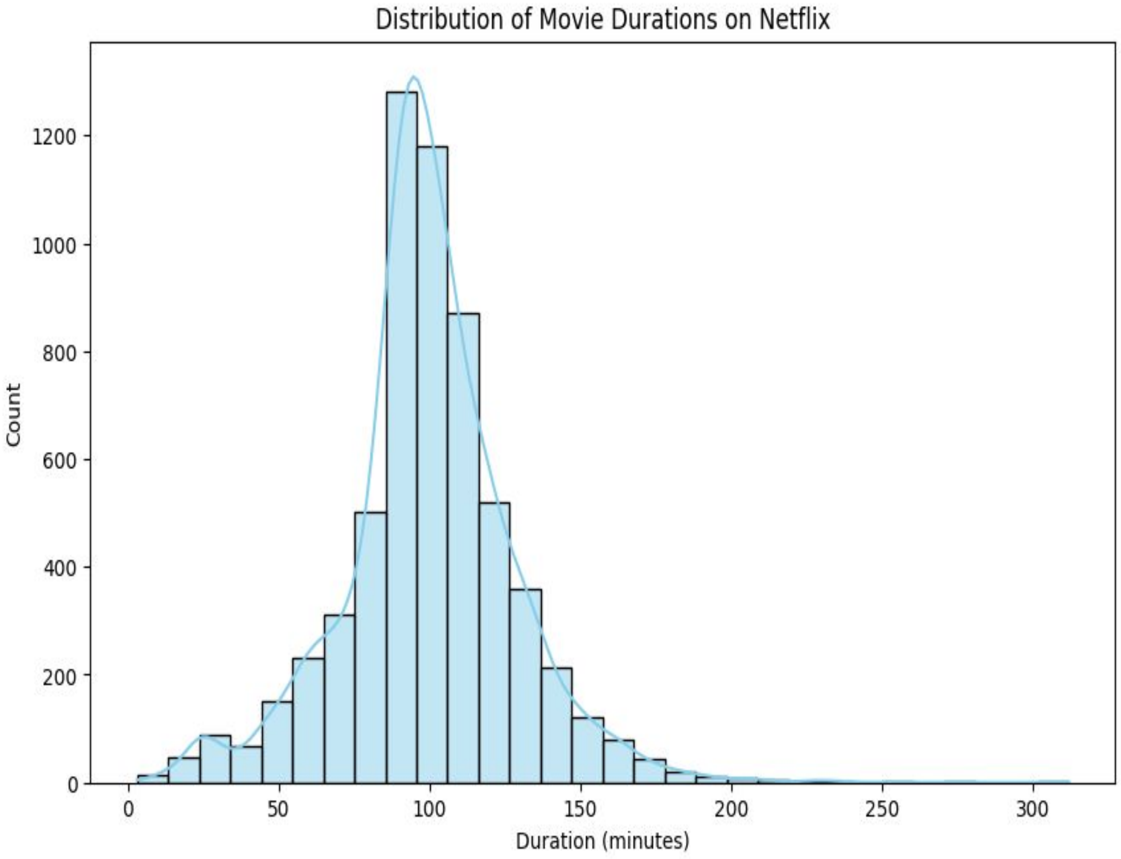
# Distribuition of Context Rating on Netfilix



- **TV-MA Leads –** Most content is for mature audiences.
- **TV-14 & TV-PG –** Significant share, family-friendly.
- **R & PG-13 –** Moderate presence.
- **Other Ratings (TV-Y, PG, TV-G, etc.) –** Less common.
- **Few Unrated & Odd Labels –** Some data inconsistencies.

DATA VISUALIZATION

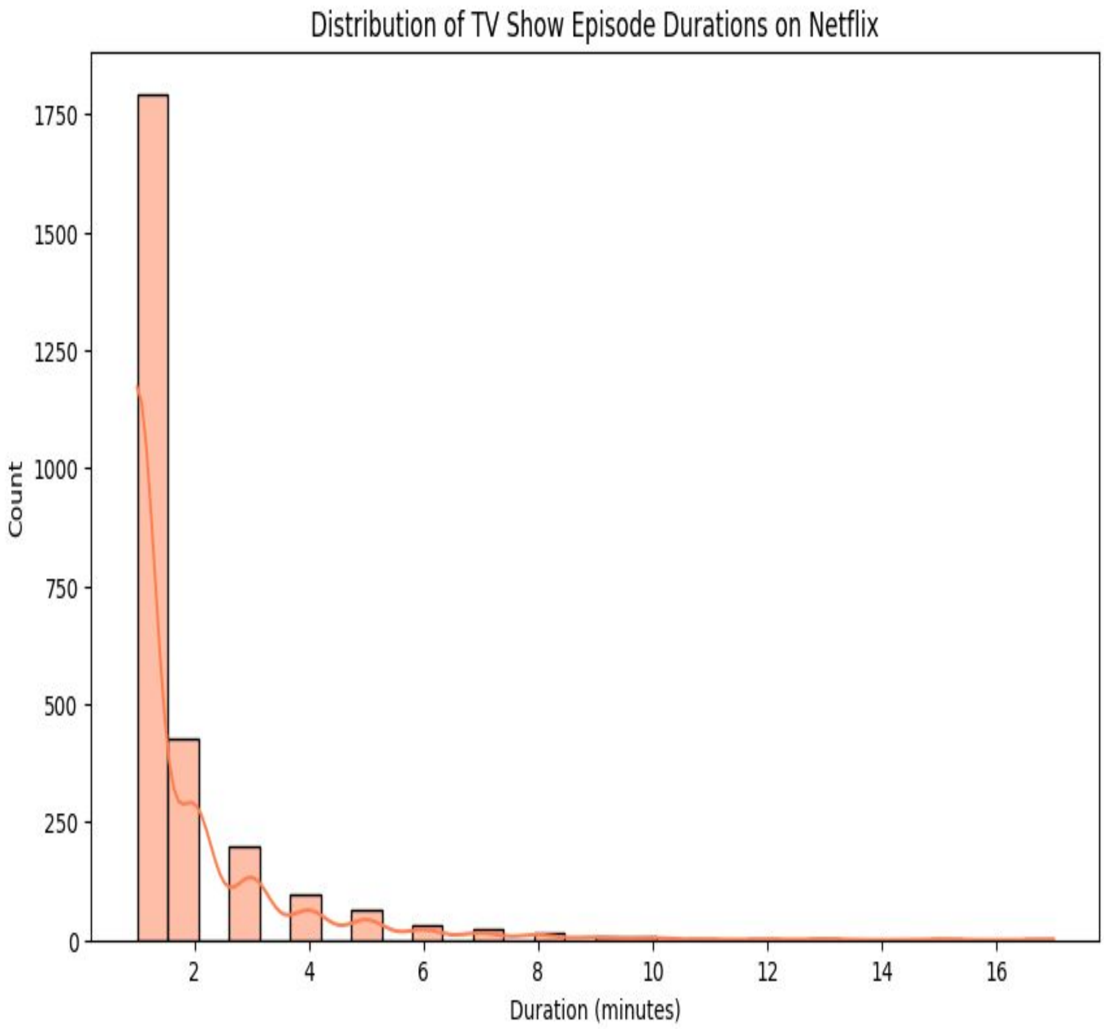
# Distribuition of Moviie Duration on Netfilix



- **Typical Duration:** Most movies on Netflix have durations between 90 to 120 minutes.
- **Average Length:** The distribution peaks around 100 minutes, indicating that this is the most common or average movie length.
- **Longer Films are Less Frequent:** There are fewer movies with durations exceeding 150 minutes, and very few go beyond 200 minutes.

DATA VISUALIZATION

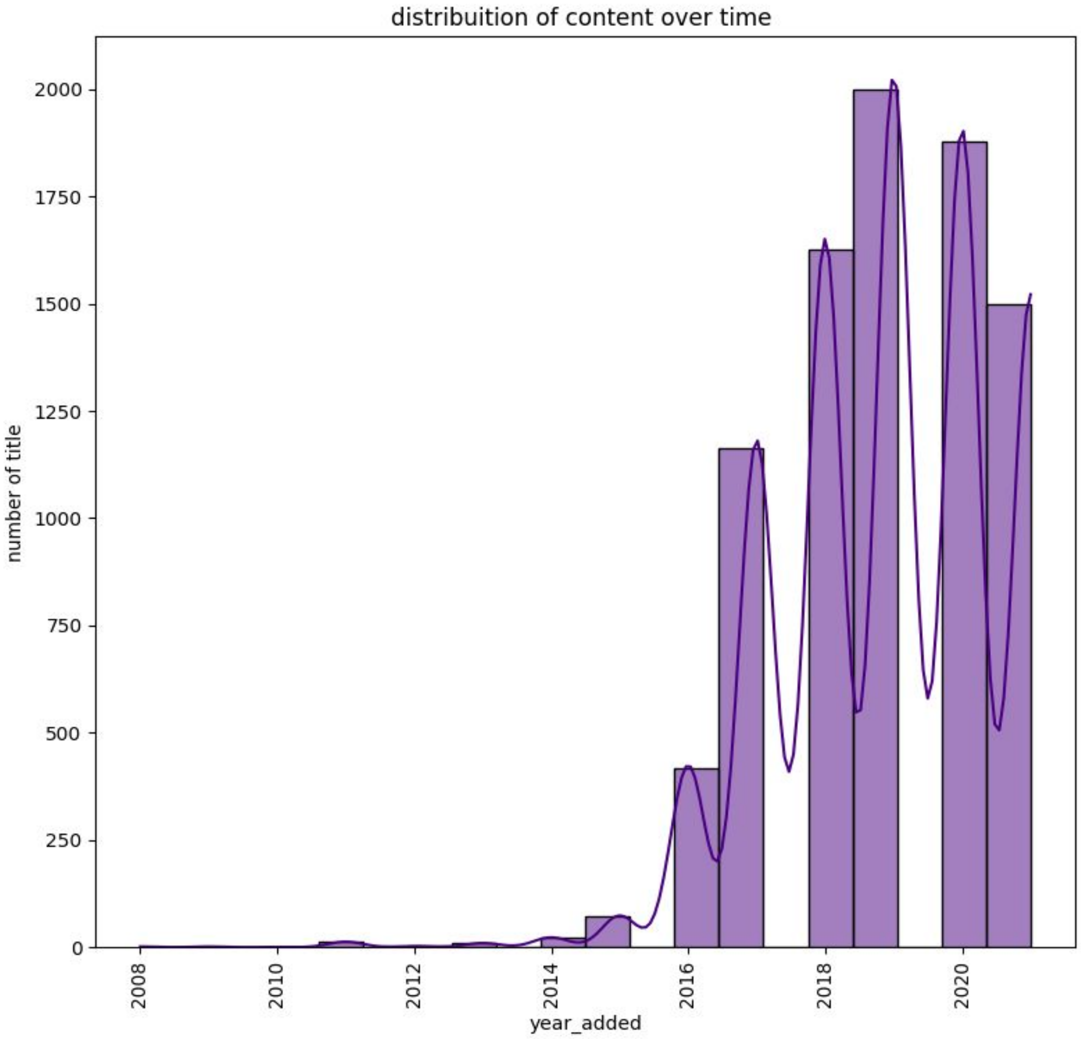
# Distribuition of TV Show Episode Duaration on Netfilix



- **Typical Length:** Most TV show episodes on Netflix have durations between 20 to 60 minutes.
- **Peak Duration:** The most common episode length is around 40 minutes.
- **Limited Longer Episodes:** Episodes exceeding 60 minutes are less frequent, and those over 120 minutes are rare.

DATA VISUALIZATION

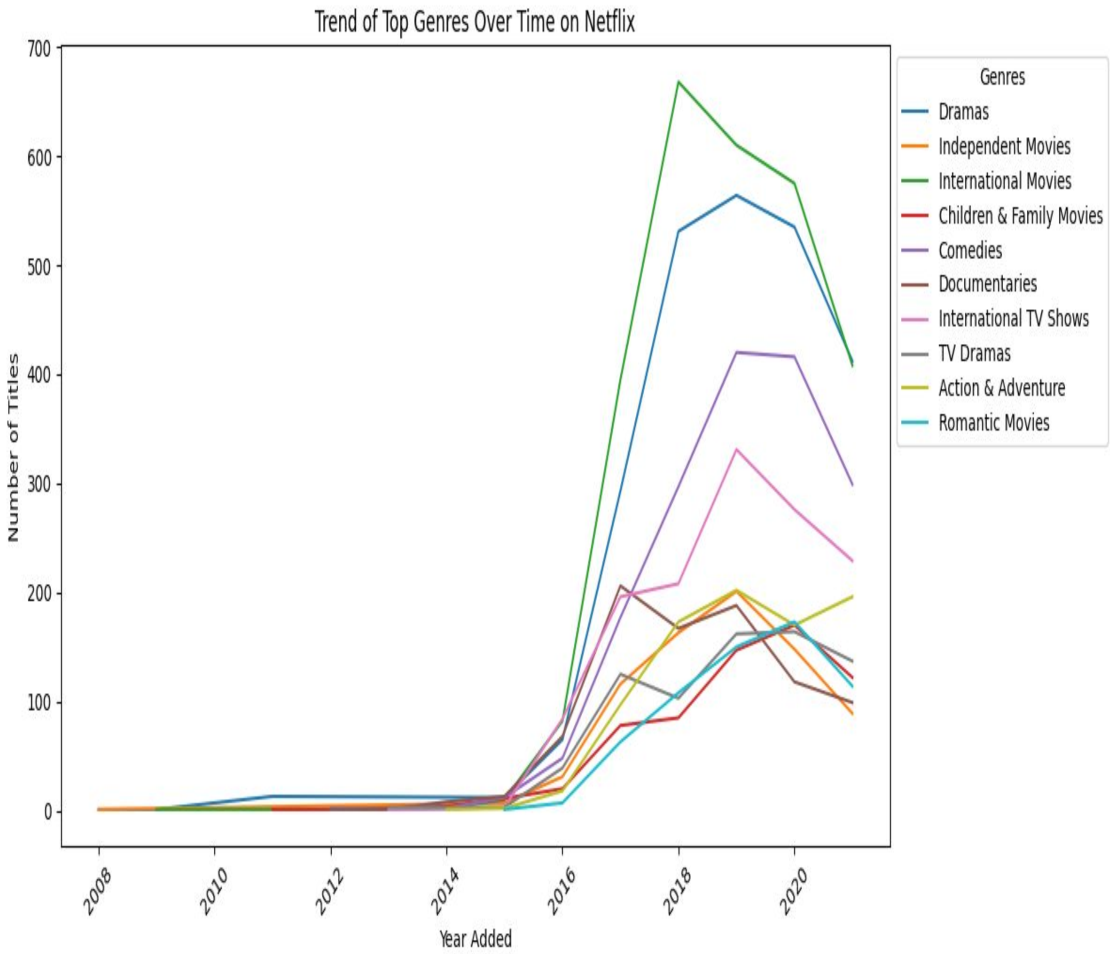
# Distribuition of content over time



- **Growth Surge –** Content increased rapidly after 2015.
- **Peak Years –** Highest additions between 2017-2020.
- **Fluctuations –** Yearly variations in content additions.
- **Low Before 2015 –** Very few titles added earlier.

DATA VISUALIZATION

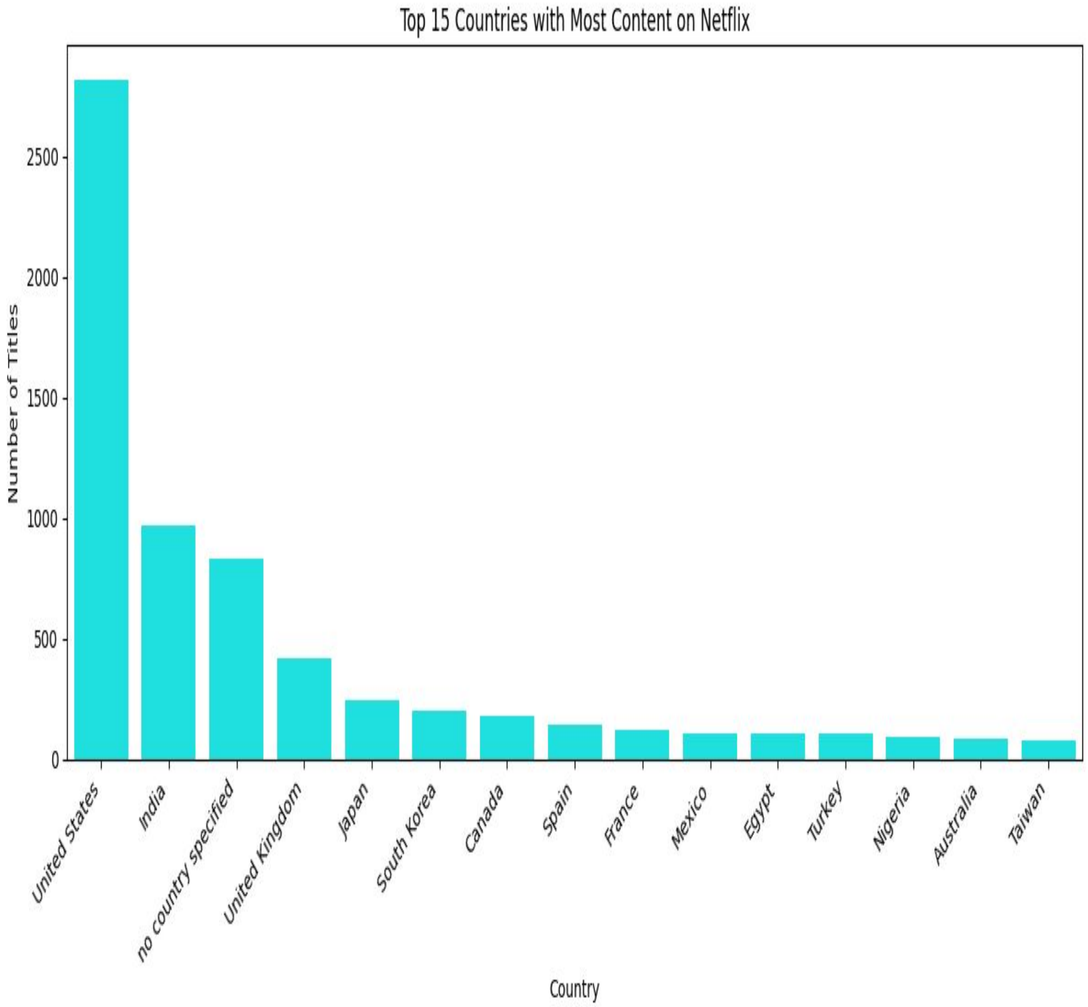
# Trend of Genres over Time on Netfilix



- **Overall Content Growth:** Netflix's content library has expanded across most genres.
- **International Movies:** This genre is becoming increasingly popular.
- **Dramas:** Remain a consistently popular genre.
- **Comedies:** Showing notable growth in recent years.
- **Documentaries:** Experiencing a significant surge in popularity.

DATA VISUALIZATION

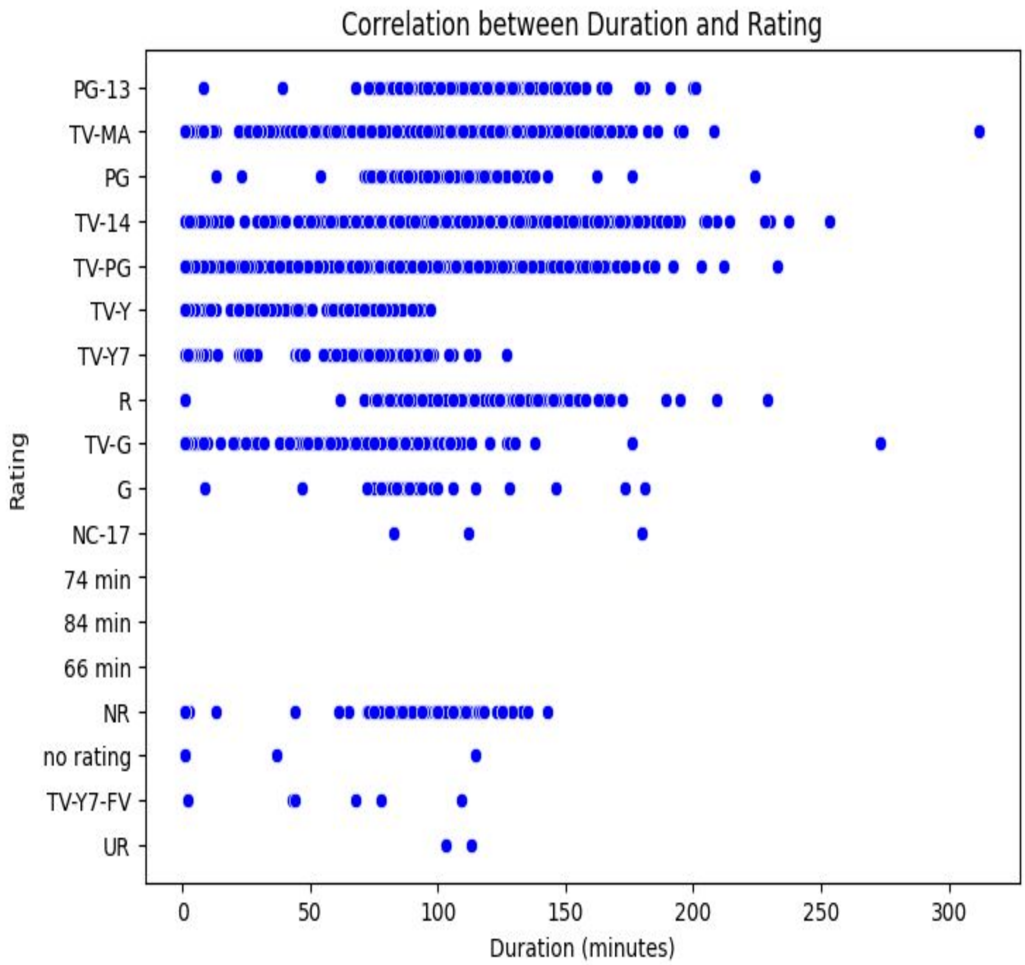
# Top 15 Countries With Most Content on Netfilix



- **US Dominates:** The US has the most Netflix content by far.
- **India is Second:** India is the second largest content source.
- **Global Variety:** Content comes from many countries, but the US and India contribute the most.
- **Unspecified Countries:** There's a sizable amount of content with no specified country.

DATA VISUALIZATION

# Correlation between Duration and Rating



**Wide Duration Range :** Content varies from **short (0-50 min)** to **long (250+ min).**

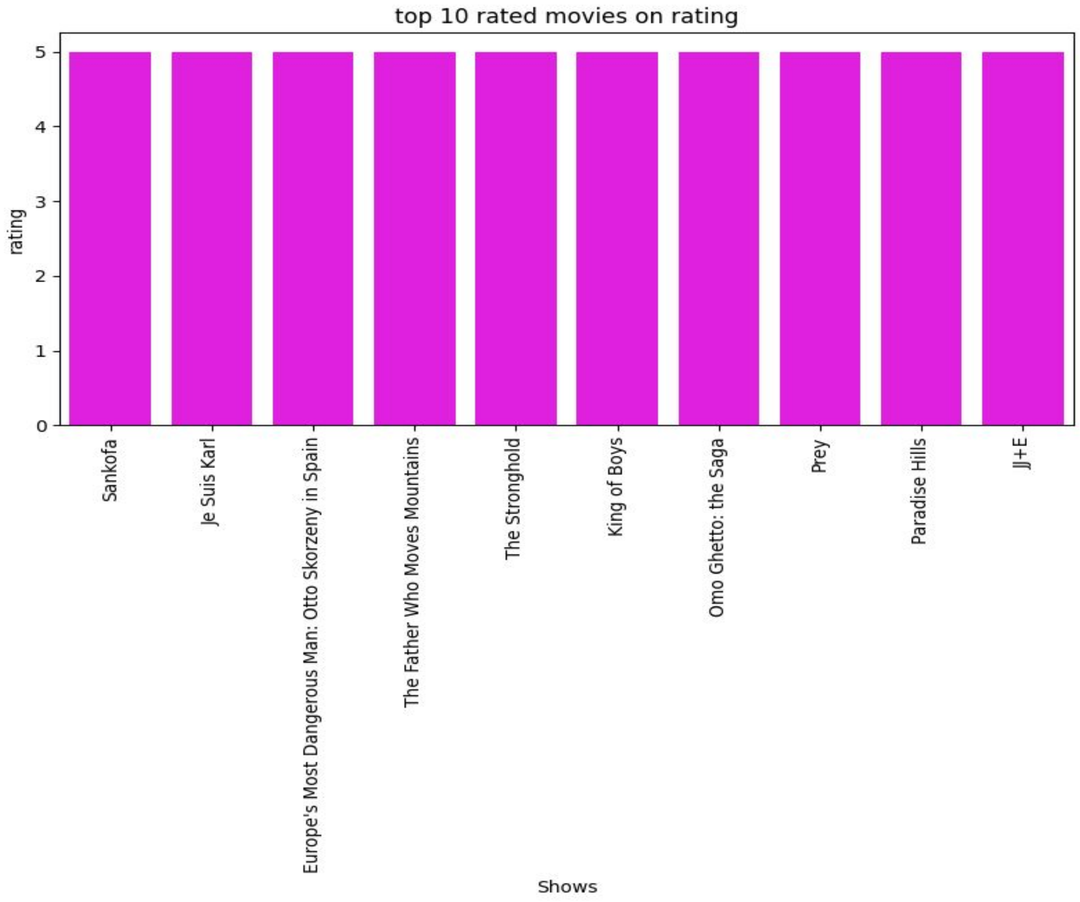
**Common Ratings :** *TV-MA, TV-14, PG-13* have the most content.

**Kids vs. Mature Content :** *TV-Y, TV-PG* content is shorter, while *R, TV-MA* shows have longer durations.

**Unrated Titles Exist :** Some

DATA VISUALIZATION

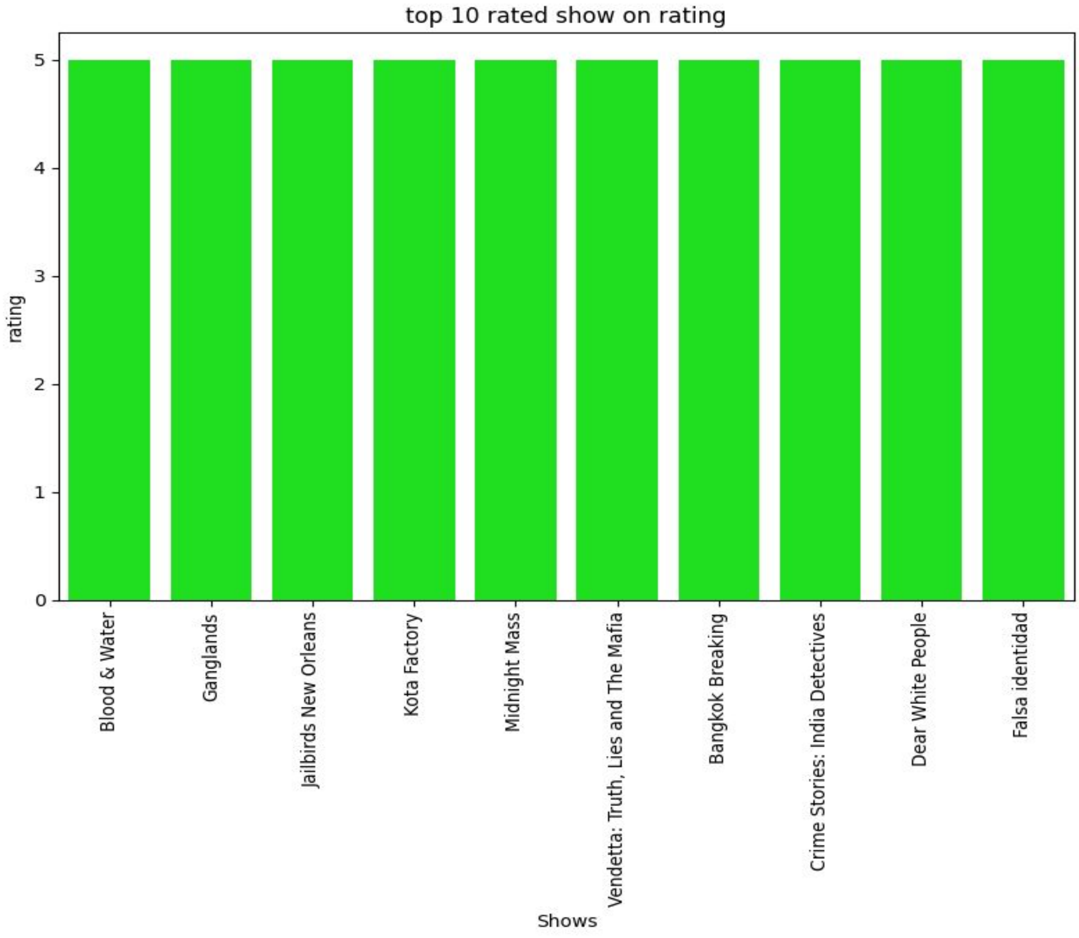
# Top 15 rated movies on rating



- **All Ratings are 5 –** Every movie has a top rating of 5.
- **Diverse Titles –** Includes different genres and origins.
- **Even Distribution –** No variation in ratings among top 10.
- **Popular Picks –** Titles likely chosen based on user reviews.
- **Vertical Labels –** Some movie names are long and rotated.

DATA VISUALIZATION

# Top 15 rated show on rating



**Top-Rated Shows :** All shows have a perfect **5-star rating**.

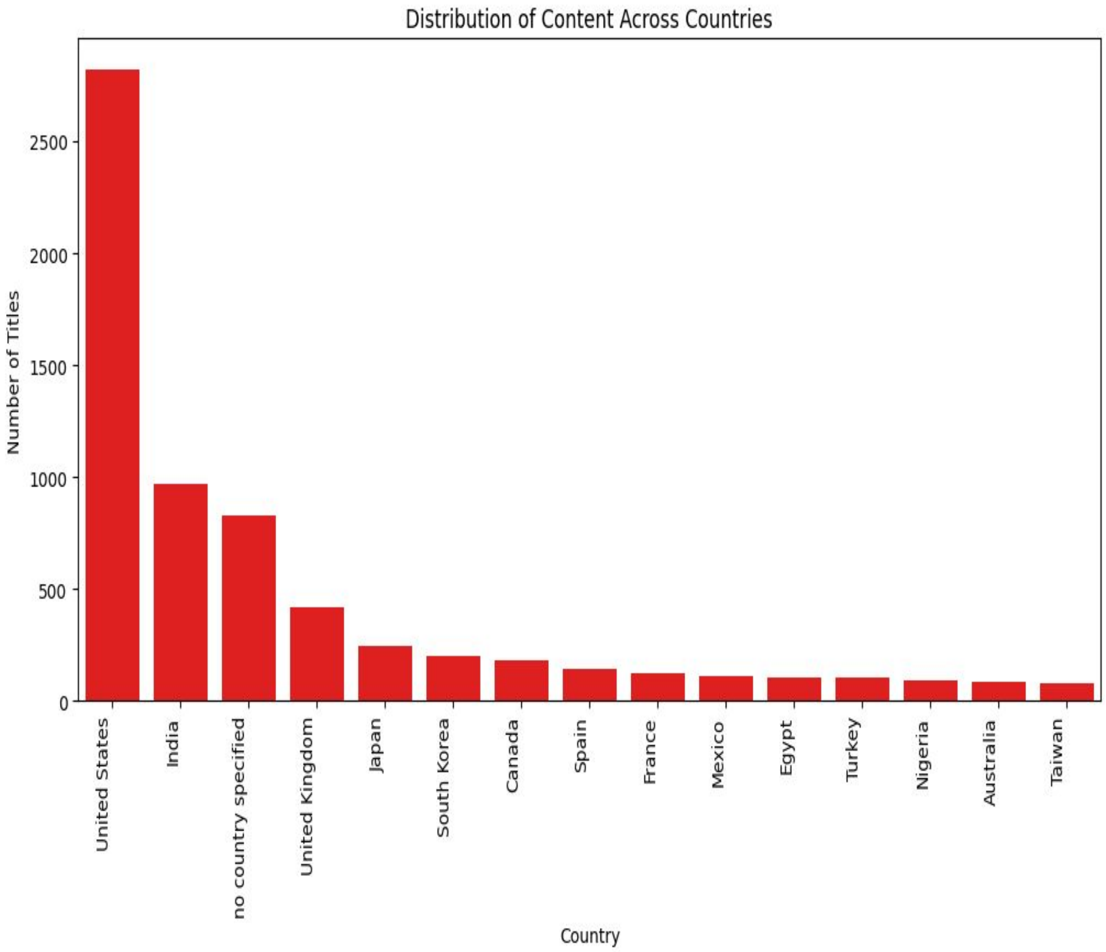
**Diverse Content:** Includes crime, drama, thriller, and investigative genres.

**Global Mix :** Shows from different regions, reflecting audience diversity.

**High Viewer Satisfaction :** Indicates strong content quality and popularity.

DATA VISUALIZATRION

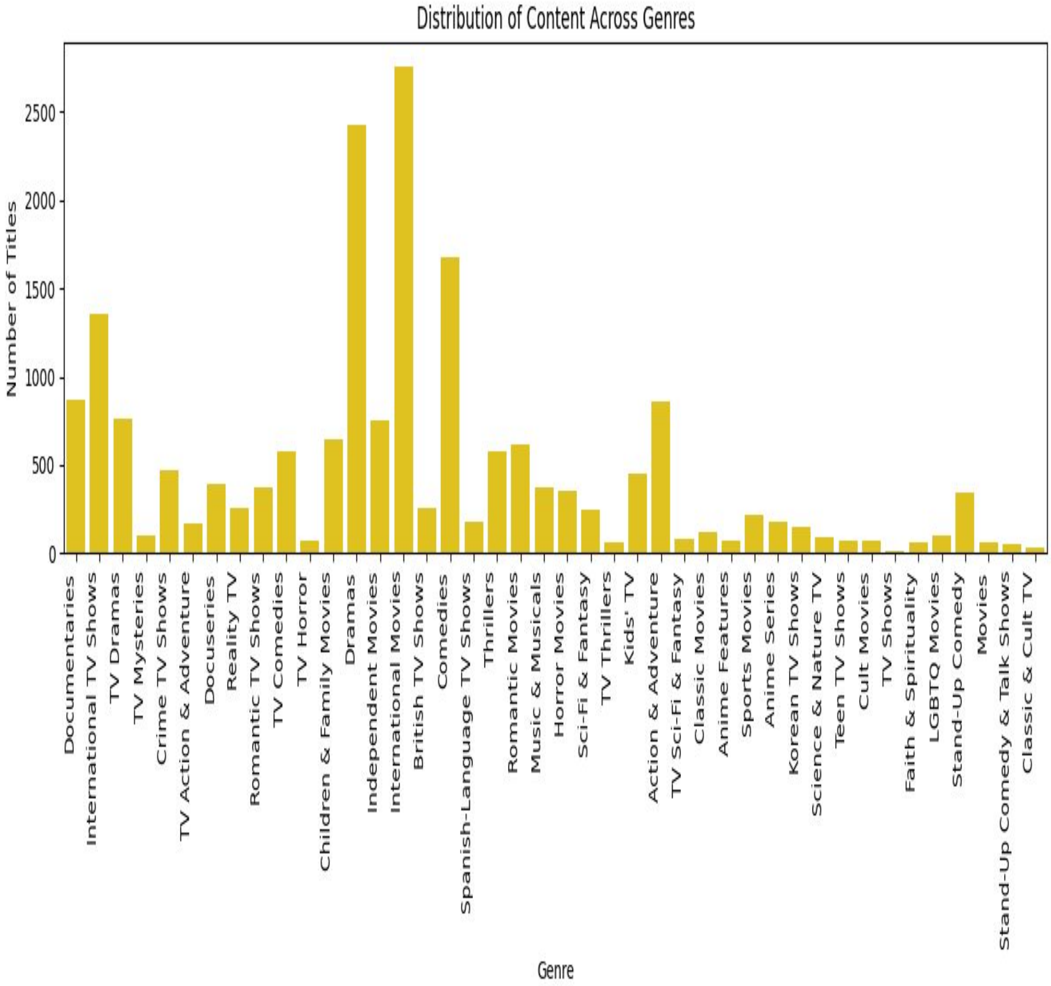
# Distribuition of Content Across Countries



- **USA leads** with the highest content count.
- **India ranks 2nd,** contributing significantly.
- **Many titles have no country specified.**
- **UK, Japan, South Korea** have moderate content.
- **Other countries (Canada, Spain, France, etc.)** contribute less.

DATA VISUALIZATION

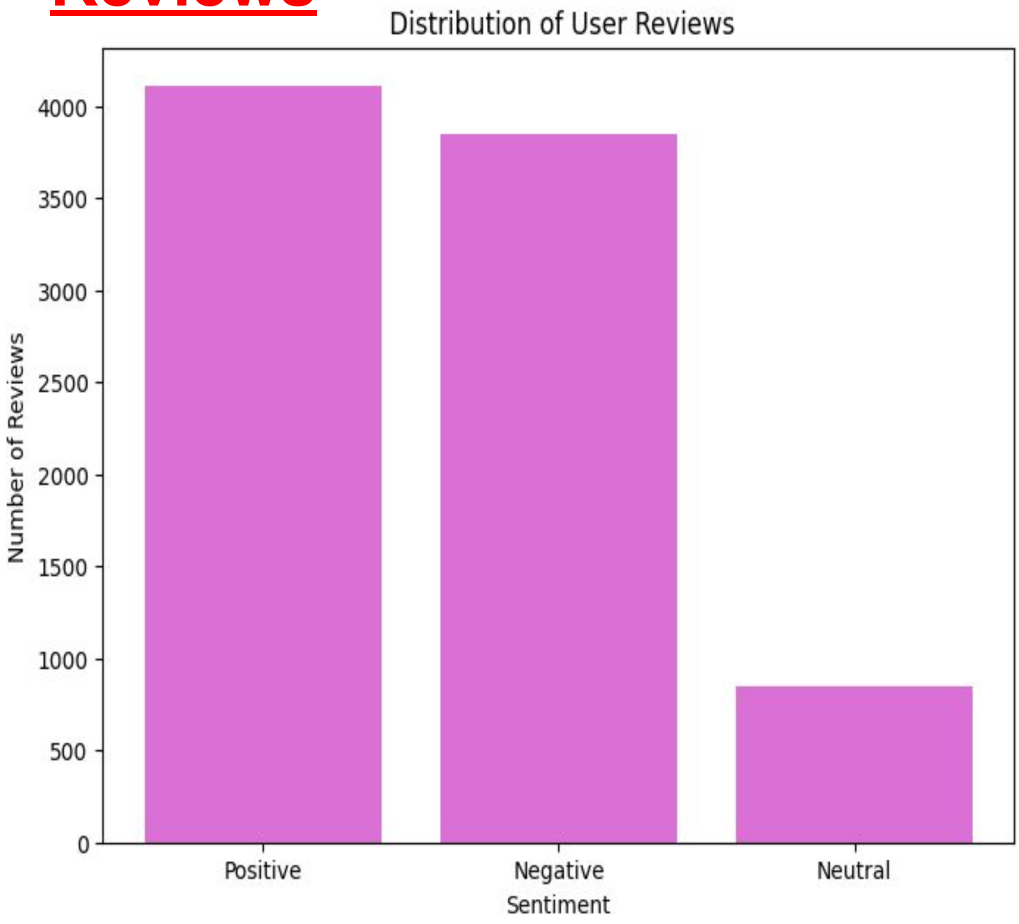
# Distribuition of Content Across Genres



- **Dramas & Comedies** have the highest number of titles.
- **Independent Movies** also dominate.
- **Documentaries & International TV Shows** have strong representation.
- **Action & Adventure** has a moderate presence.

DATA VISUALIZATION

## Distribuition of Users Reviews



- **Positive Reviews Lead –** Highest number of user reviews are positive.
- **Negative Reviews Close Behind –** Almost as many as positive reviews.
- **Neutral Reviews Fewest –** Least number of reviews fall in the neutral category.

# RECOMMANDATIONS

- **Focus on Recent Content –** Since most content is from 2015 onwards, Netflix should prioritize new and trending releases.
- **Revive Classic Content –** Older movies and shows (pre-2000s) are minimal; adding more classic content may attract nostalgia-driven viewers.
- **Maintain Release Consistency –** sharp drop after 2020 suggests external factors; Netflix should ensure a steady flow of new releases.
- **Expand Genre Variety –** Increasing diversity in top genres can attract a broader audience.
- **Improve Content Discovery –** With a high number of recent titles, better recommendations and categorization can enhance user experience.

# CONCLUSION

- **Content Diversity:** Netflix offers a vast and diverse library, with a wide range of genres and content from various countries. However, there's a notable dominance of content from the United States, followed by India. This suggests a focus on catering to these major markets.
- **Genre Preferences:** The most popular genres include dramas, international movies, documentaries, and stand-up comedy. This indicates user preferences for diverse content types, including both scripted and unscripted programming.
- **Content Ratings:** TV-MA and TV-14 are the most common ratings, suggesting a focus on mature audiences and teenagers. This aligns with the general trend of streaming services offering more mature content.

# CONCLUSION

- **Release Trends:** There's been a significant increase in content released after 2000, with a peak between 2015 and 2020. This highlights Netflix's commitment to constantly expanding its library with new and recent releases.
- **Movie and TV Show Durations:** Movie durations typically range between 90 to 120 minutes, while TV show episodes are shorter, mostly falling between 20 to 60 minutes. This aligns with standard industry practices for content length.
- **User Sentiment:** Overall, user sentiment is mostly positive, with a substantial number of negative reviews as well. This indicates a mixed reception to Netflix's content, but with a predominantly positive outlook.

# THANX FOR READING

## FOR CODE

https://github.com/Vishudeshwal12/ZOMATO