NETFLIX DATA ANALYSIS REPORT

By- Drishti Khosla

About Netflix

- Netflix is one of the world's leading streaming entertainment services, offering a wide variety of TV shows, movies, documentaries, and more.
- Founded in 1997 by Reed Hastings and Marc Randolph, Netflix started as a DVD rental service before transforming into a global streaming giant.
- Today, Netflix operates in over 190 countries and has millions of subscribers worldwide.
- It provides content across multiple genres and languages, catering to a diverse global audience.
- Netflix's success lies in its ability to produce original content and use data analytics to understand viewer preferences.

Project Overview

This project analyzes Netflix’s dataset to uncover insights about its movies and TV shows. Using Exploratory Data Analysis (EDA), it explores trends in genres, release years, content types, and country distributions. The goal is to understand how Netflix’s content has evolved over time and identify key patterns that influence its global content strategy. Tools like Pandas, Matplotlib, and Seaborn are used for data cleaning, visualization, and analysis.

Project Objectives

1. Understand Netflix’s content composition (movies vs TV shows).
2. Analyze release year trends and content growth.
3. Explore genre and country distribution.
4. Study content duration and evolution trends.
5. Examine ratings and correlation between variables.
6. Provide insights and recommendations for Netflix’s future content strategy

Tools and Libraries Used

- Python
- Pandas
- Seaborn
- Matplotlib
- Google Collab
- Canva (for presentation)

Dataset Description

Attribute	Description
show_id	Unique ID for each title
type	Movie or TV Show
title	Name of the content
director	Director of the content
cast	List of main actors
country	Country of origin
date_added	When it was added to Netflix
release_year	Year the content was release
rating	Content rating (e.g., PG, R, T
duration	Duration in minutes or seaso
listed_in	Genre
description	Short summary

Data Cleaning

- Handled missing values (filled with mode or dropped nulls).

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

	0
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

```
[ ]  df['director'].fillna('Not Available', inplace = False)
```

	director
0	Kirsten Johnson
1	Not Available
2	Julien Leclercq
3	Not Available
4	Not Available
...	...
8802	David Fincher
8803	Not Available
8804	Ruben Fleischer
8805	Peter Hewitt
8806	Mozez Singh

8807 rows × 1 columns

dtype: object

```
[ ]
```

Data Cleaning

- Removed duplicate entries.
- Converted date_added column to datetime format.
- Extracted year_added and month_added.
- Created new column main_genre from listed_in.

```
[5]
✓ 0s
df['date_added'].fillna(df['date_added'].mode()[0], inplace = True)
```

```
[ ]
df['rating'].fillna(df['rating'].mode()[0], inplace = False)
```

	rating
0	PG-13
1	TV-MA
2	TV-MA
3	TV-MA
4	TV-MA
...	...
8802	R
8803	TV-Y7
8804	R
8805	PG
8806	TV-14

8797 rows × 1 columns

dtype: object

Descriptive Statistics

```
[ ]  df.describe()
```

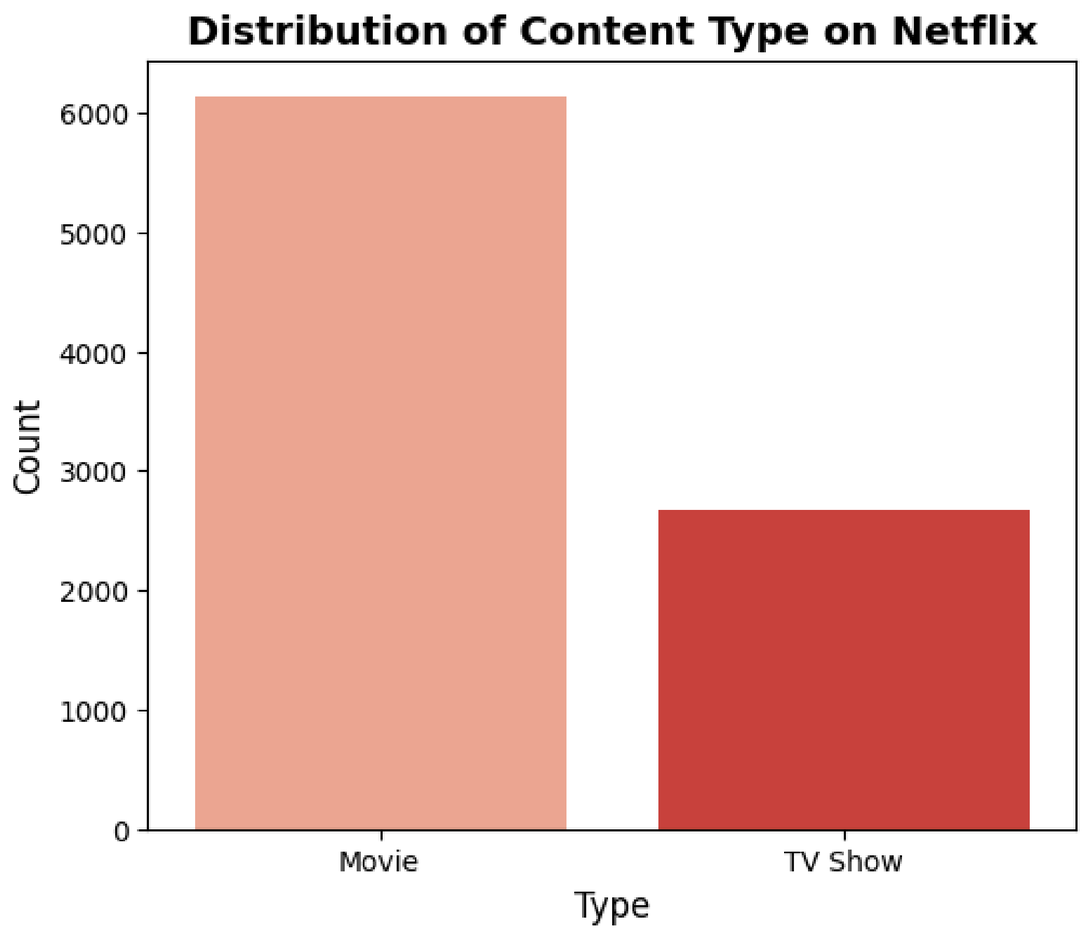
	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

- Count: 8,807 total titles in the dataset.
- Mean: Average release year is 2014, showing a focus on modern content.
- Min–Max: Titles range from 1925 to 2021.
- Median: Half of the titles were released in or after 2017.
- Trend: Most content was released between 2013 and 2021, indicating Netflix’s major growth period.

Distribution of content over different genres

```
plt.figure(figsize=(6,5))
sns.countplot(data = df, x = 'type', palette = 'Reds')
plt.title('Distribution of Content Type on Netflix', fontsize=14, fontweight='bold')
plt.xlabel('Type', fontsize=12)
plt.ylabel('Count', fontsize=12)
plt.show()
```

```
/tmp/ipython-input-2638211487.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set

  sns.countplot(data = df, x = 'type', palette = 'Reds')
```

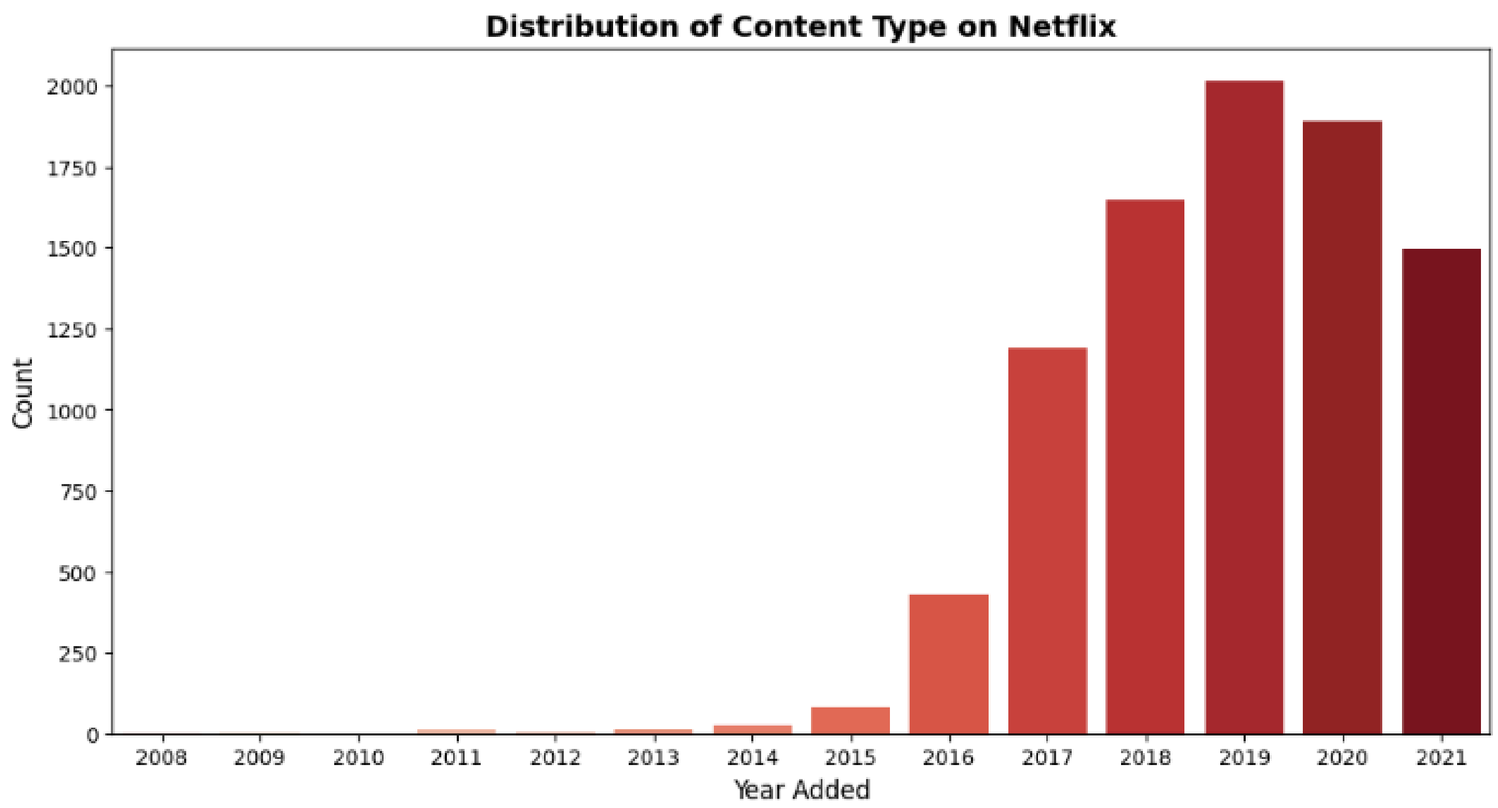


The graph illustrates the distribution of content types available on Netflix, comparing the number of movies and TV shows. It is evident that movies dominate the platform's library, with a much higher count than TV shows. This suggests that Netflix primarily focuses on offering a wide variety of movies, while TV shows make up a smaller portion.

Distribution of content across release years

```
plt.figure(figsize=(12,6))
sns.countplot(data = df, x = 'year_added', order = sorted(df['year_added'].dropna().unique()), palette = 'Reds')
plt.title('Distribution of Content Type on Netflix', fontsize = 14, fontweight = 'bold')
plt.xlabel('Year Added', fontsize = 12)
plt.ylabel('Count', fontsize = 12)
plt.show()
```

```
/tmp/ipython-input-1708756741.py:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(data = df, x = 'year_added', order = sorted(df['year_added'].dropna().unique()), palette = 'Reds')
```

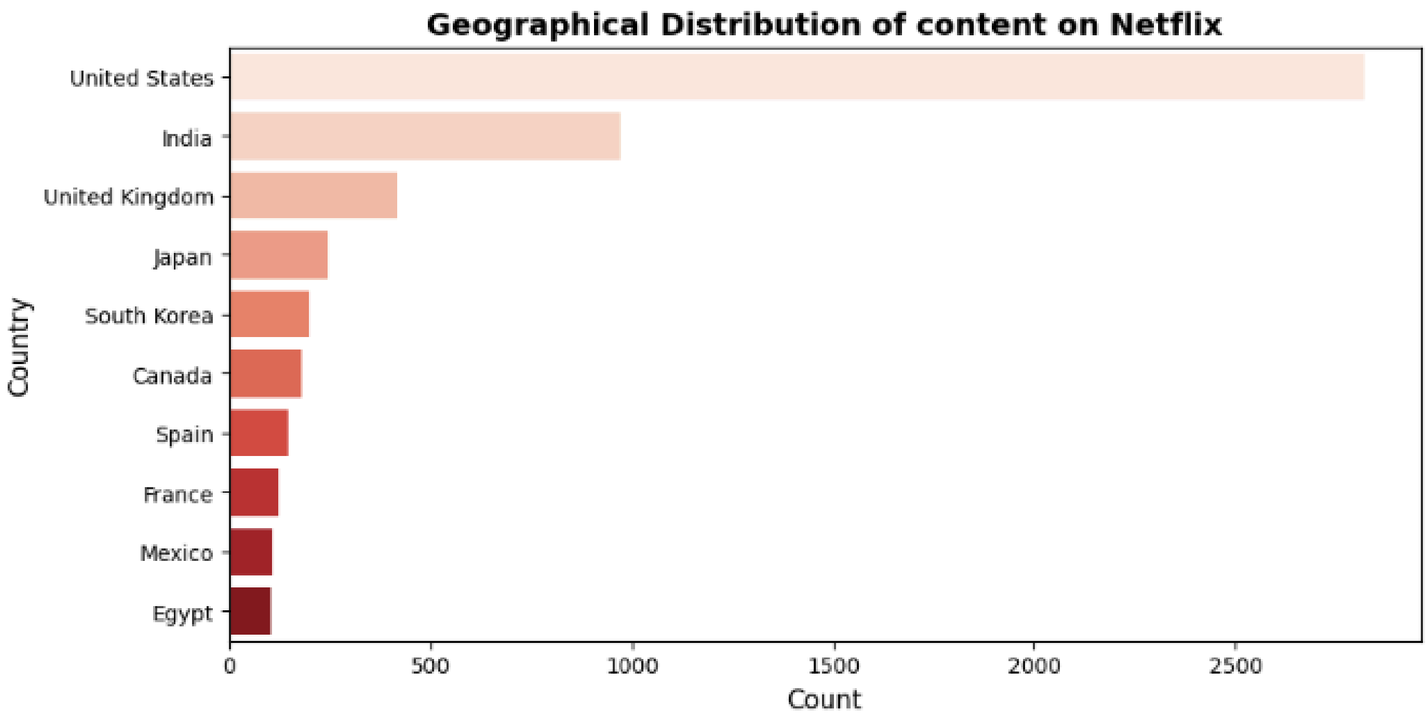


- The chart shows the yearly distribution of content added to Netflix from 2008 to 2021.
- A sharp rise in titles is seen from 2015 onwards.
- The highest additions occurred around 2018–2019.
- A slight decline in 2020–2021 suggests a slowdown, possibly due to the pandemic.

Geographical distribution of content

```
top_countries = df['country'].value_counts().head(10).index
plt.figure(figsize=(10,5))
sns.countplot(data = df[df['country'].isin(top_countries)], y = 'country', order = top_countries, palette = 'Reds')
plt.title('Geographical Distribution of content on Netflix', fontsize = 14, fontweight = 'bold')
plt.xlabel('Count', fontsize = 12)
plt.ylabel('Country', fontsize = 12)
plt.show()
```

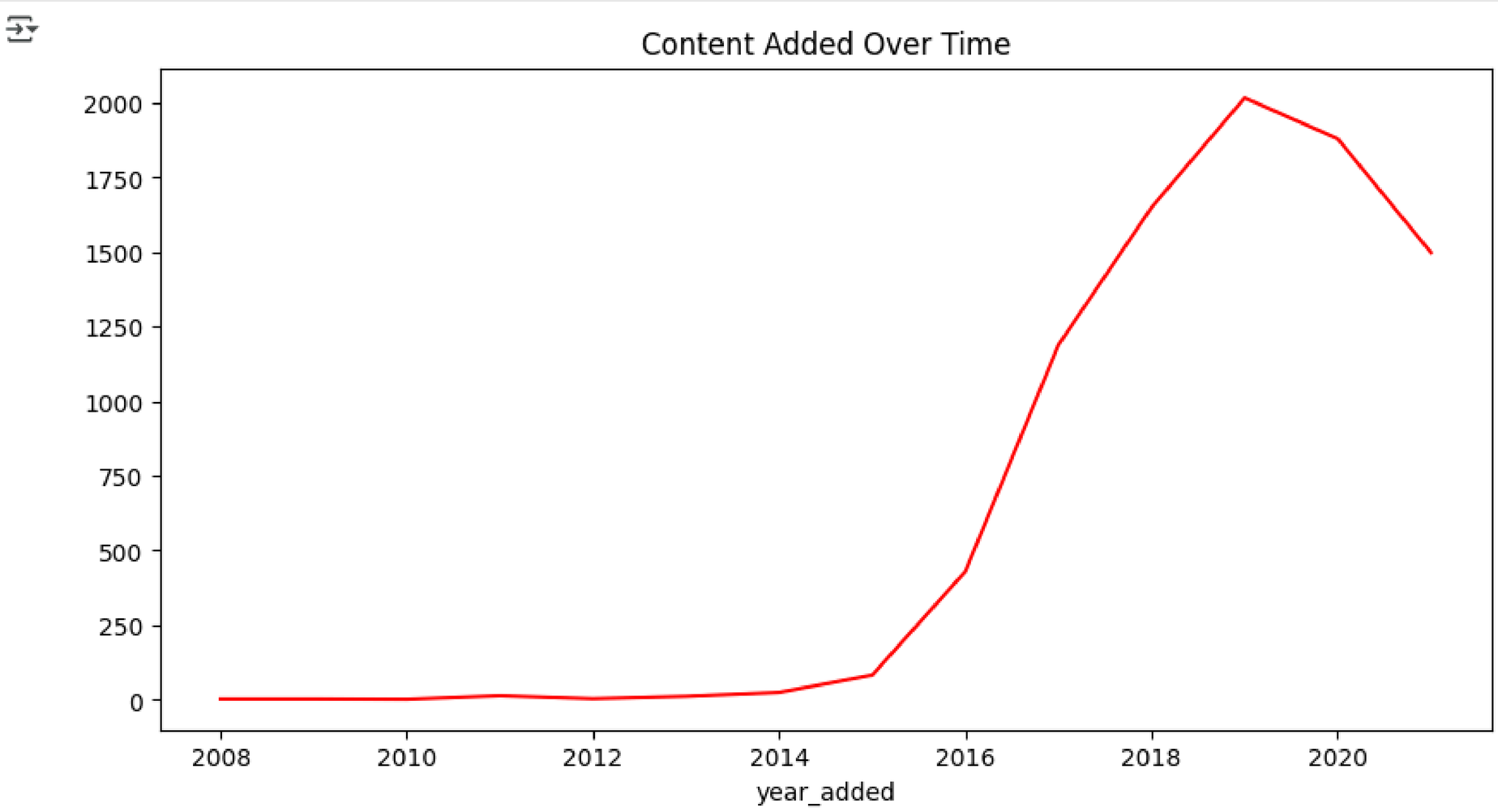
```
/tmp/ipython-input-496939976.py:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(data = df[df['country'].isin(top_countries)], y = 'country', order = top_countries, palette = 'Reds')
```



- The chart shows the top countries contributing to Netflix’s content library.
- The United States has the highest number of titles.
- India and the United Kingdom are the next major contributors.
- Other key countries include Japan, South Korea, and Canada.
- This indicates Netflix’s strong global presence across multiple regions.

Time series analysis

```
[ ]
df['year_added'].value_counts().sort_index().plot(kind='line', figsize=(10,5), color='red')
plt.title('Content Added Over Time')
plt.show()
```

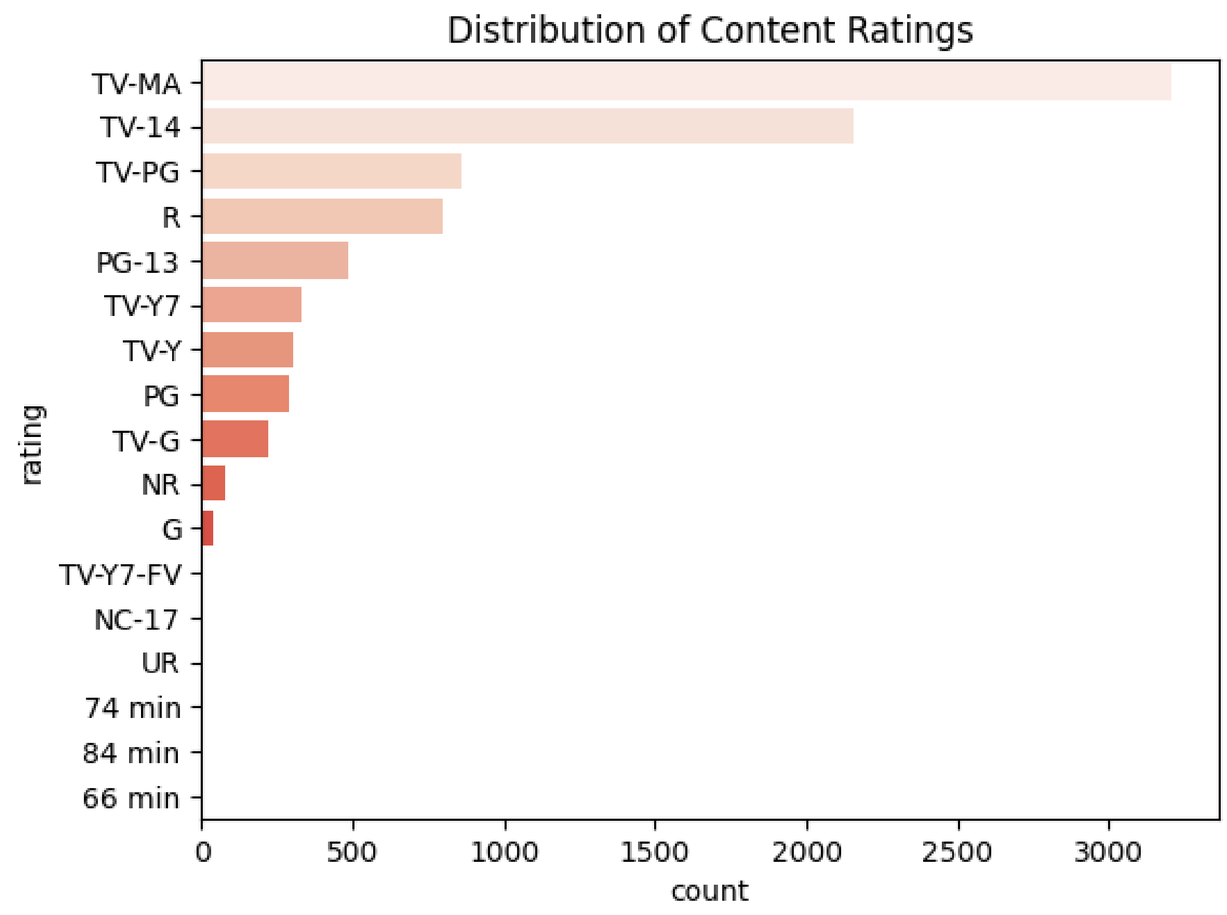


1. Content additions stayed low until 2014.
2. A sharp increase began around 2015.
3. Additions peaked in 2019.
4. Content added declined slightly after 2019.
5. Overall trend shows strong growth over the years.

Distribution of content ratings

```
[ ]
sns.countplot(y='rating', data=df, order=df['rating'].value_counts().index, palette='Reds')
plt.title('Distribution of Content Ratings')
plt.show()
```

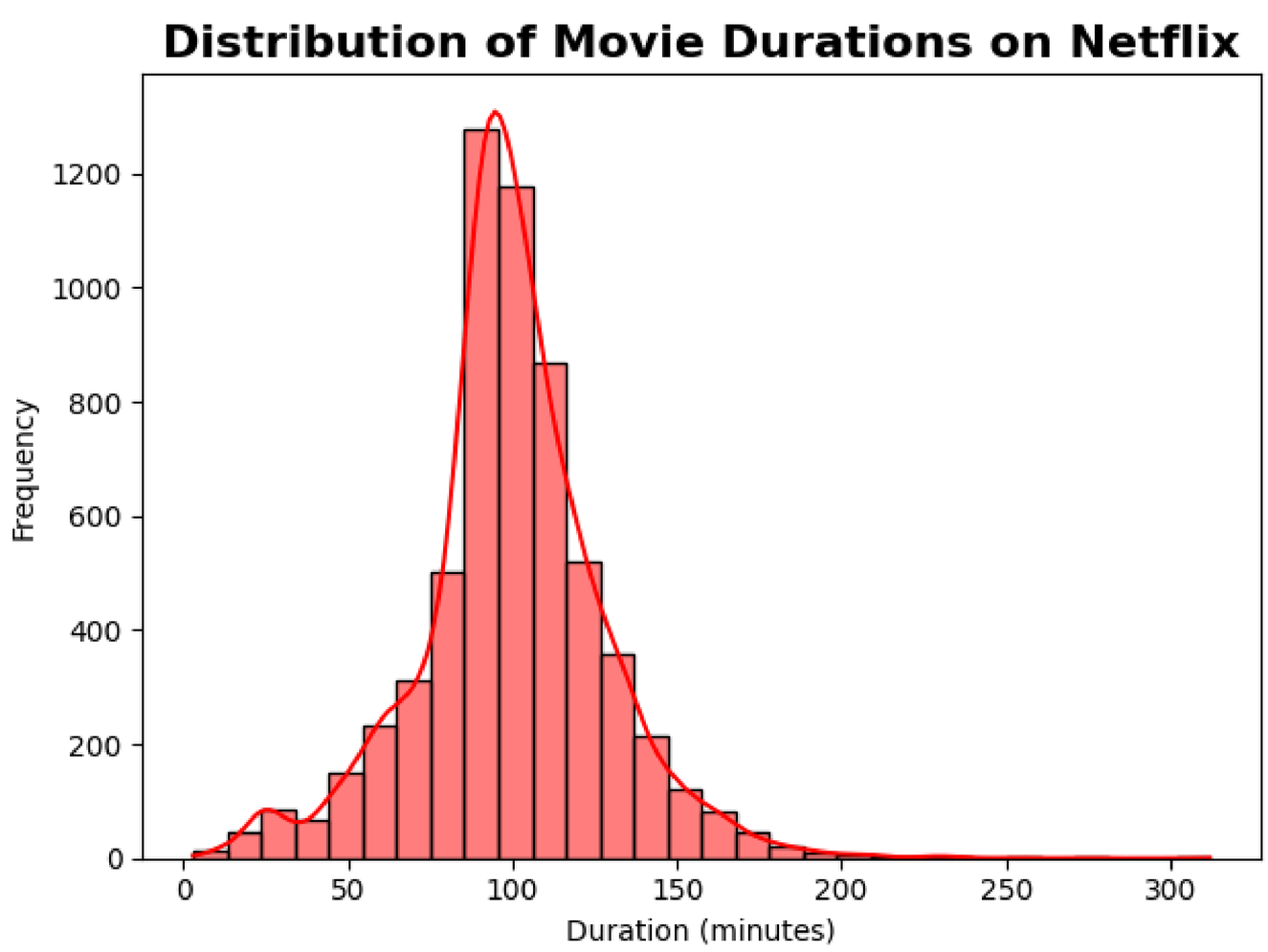
```
/tmp/ipython-input-2368202667.py:1: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set

  sns.countplot(y='rating', data=df, order=df['rating'].value_counts().index, palette='Reds')
```



1. TV-MA is the most common rating on Netflix, indicating a large amount of mature content.
2. TV-14 and TV-PG follow, showing strong availability of teen-friendly content.
3. Ratings like R, PG-13, and TV-Y7 appear in moderate numbers.
4. Kids' ratings such as TV-Y and G are present but much less common.

Distribution of movie durations

```
sns.histplot(data=movies, x='duration_int', bins=30, kde=True, color='red')
plt.title('Distribution of Movie Durations on Netflix', fontsize=16, fontweight='bold')
plt.xlabel('Duration (minutes)')
plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```

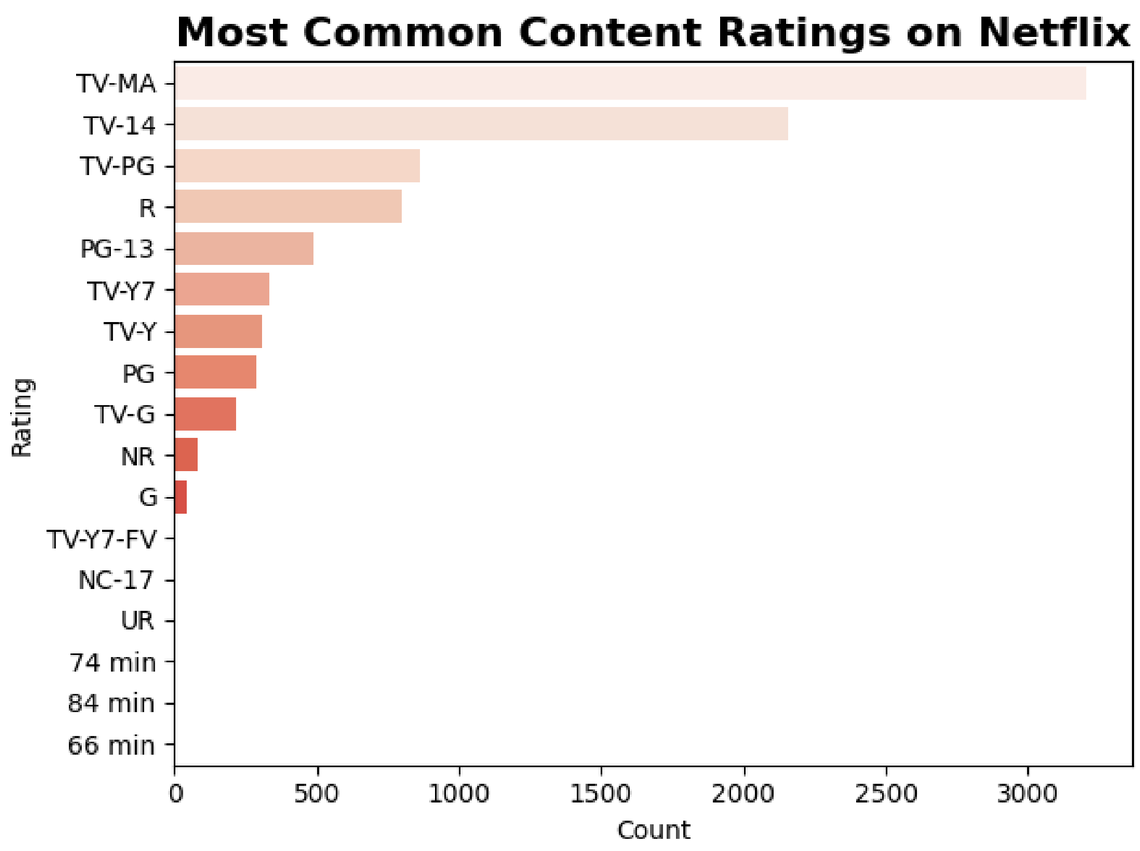


- Most Netflix movies have durations between 80 and 110 minutes, forming a strong peak in this range.
- Very short movies (under 50 minutes) and very long ones (over 150 minutes) are much less common.
- The distribution is right-skewed, meaning a small number of movies are significantly longer than average.

Most common content ratings

```
[ ]
sns.countplot(data=df, y='rating', order=df['rating'].value_counts().index, palette='Reds')
plt.title('Most Common Content Ratings on Netflix', fontsize=16, fontweight='bold')
plt.xlabel('Count')
plt.ylabel('Rating')
plt.tight_layout()
plt.show()
```

```
/tmp/ipython-input-1985240965.py:1: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=

  sns.countplot(data=df, y='rating', order=df['rating'].value_counts().index, palette='Reds')
```

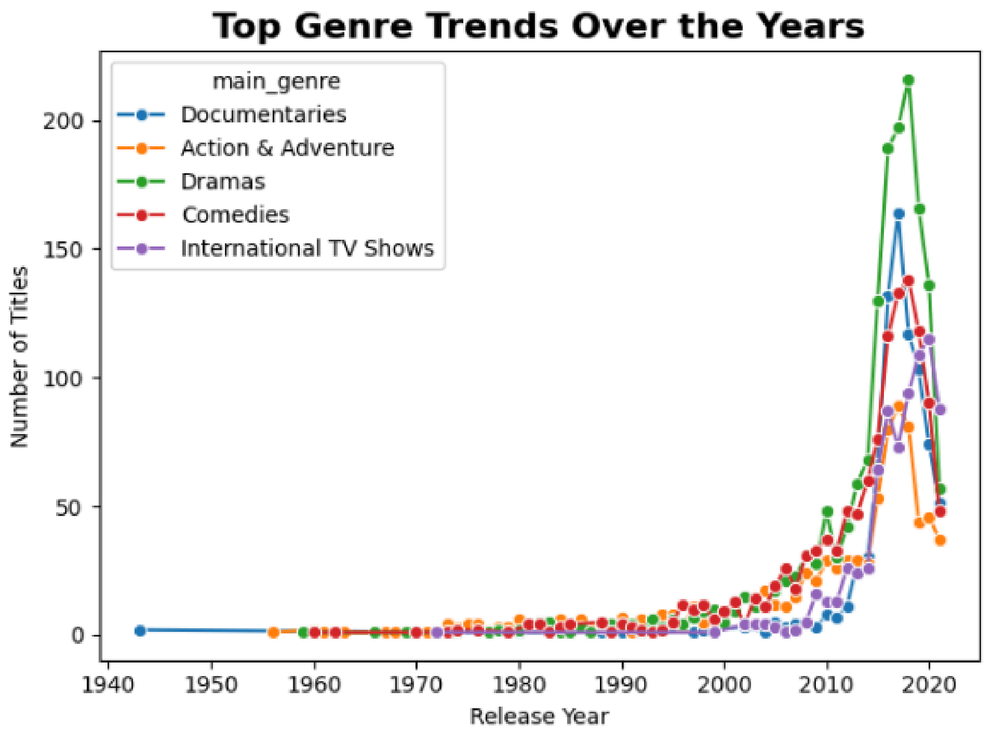


1. TV-MA is the most common rating on Netflix, showing a strong focus on mature content.
2. TV-14 and TV-PG follow, indicating lots of teen-friendly shows and movies.
3. Kid-friendly ratings like TV-Y, PG, and G appear much less frequently.
4. Overall, Netflix’s catalog leans heavily toward adult and teen audiences.

Top genre trends over the years

```
df['main_genre'] = df['listed_in'].apply(lambda x: str(x).split(',')[0])
genre_trend = df.groupby(['release_year', 'main_genre']).size().reset_index(name='Count')

top_genres = df['main_genre'].value_counts().head(5).index
sns.lineplot(data=genre_trend[genre_trend['main_genre'].isin(top_genres)],
             x='release_year', y='Count', hue='main_genre', marker='o', palette='tab10')
plt.title('Top Genre Trends Over the Years', fontsize=16, fontweight='bold')
plt.xlabel('Release Year')
plt.ylabel('Number of Titles')
plt.tight_layout()
plt.show()
```

```
/tmp/ipython-input-1019680688.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['main_genre'] = df['listed_in'].apply(lambda x: str(x).split(',')[0])
```

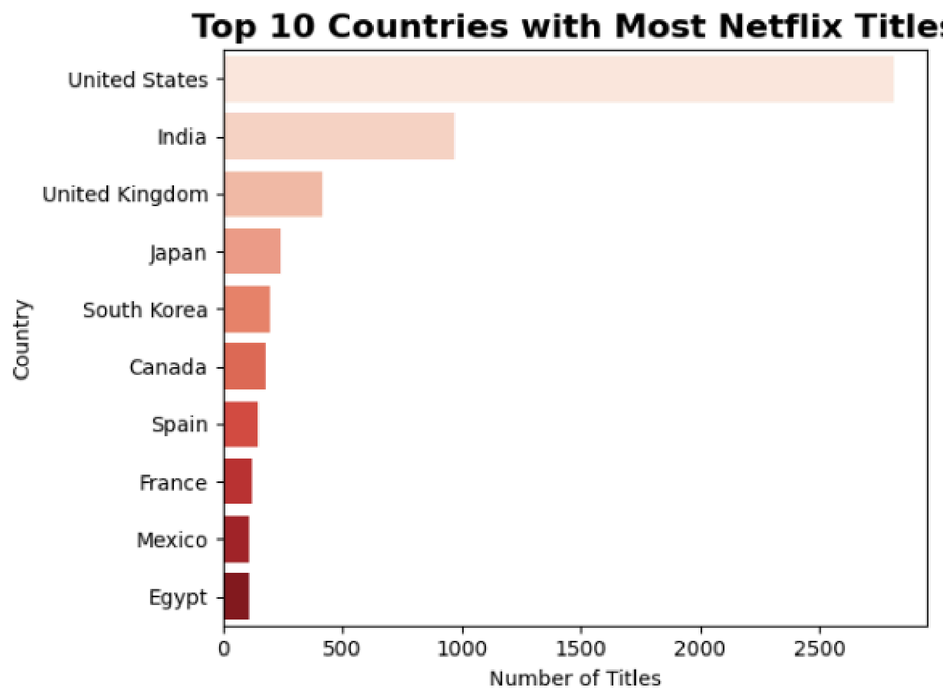


1. All major genres show very little activity before 2010, followed by a sharp rise afterward.
2. International TV Shows and Documentaries see the biggest surge in recent years.
3. Dramas and Comedies also grow steadily but at a slightly lower rate.
4. Overall, Netflix dramatically increased content across all top genres after 2015.

Geographical analysis

```
top_countries = df['country'].value_counts().head(10).index
sns.countplot(data=df[df['country'].isin(top_countries)], y='country', order=top_countries, palette='Reds')
plt.title('Top 10 Countries with Most Netflix Titles', fontsize=16, fontweight='bold')
plt.xlabel('Number of Titles')
plt.ylabel('Country')
plt.tight_layout()
plt.show()
```

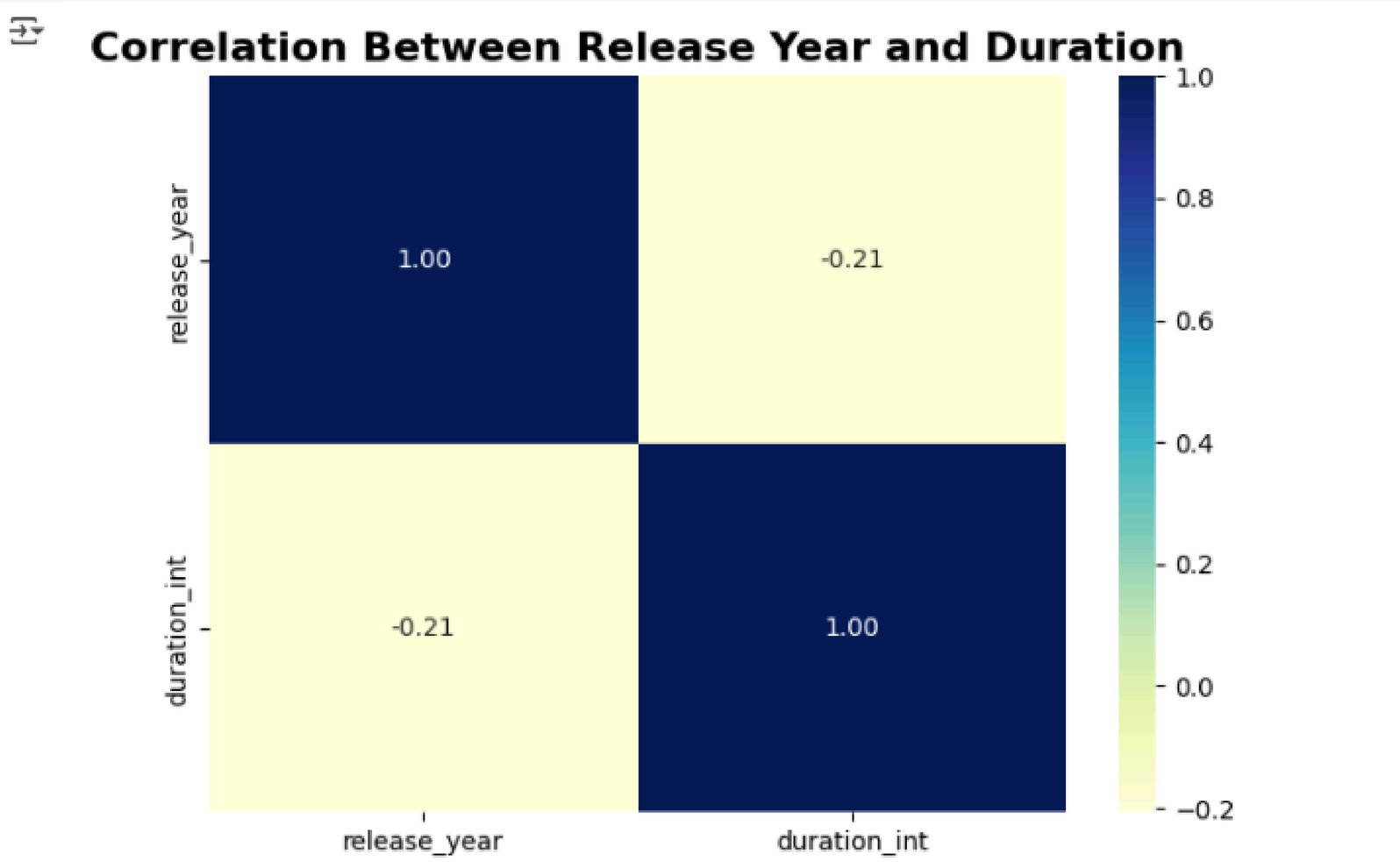
```
/tmp/ipython-input-995609631.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(data=df[df['country'].isin(top_countries)], y='country', order=top_countries, palette='Reds')
```



This code identifies the top 10 countries with the most Netflix titles and visualizes them using a horizontal bar chart. It shows that the United States leads by a large margin, followed by India and the United Kingdom. The chart uses a red color palette for better visual appeal.

Correlation between variables

```
[ ]
numeric_df = movies[['release_year', 'duration_int']].dropna()

sns.heatmap(numeric_df.corr(), annot=True, cmap='YlGnBu', fmt='.2f')
plt.title('Correlation Between Release Year and Duration', fontsize=16, fontweight='bold')
plt.tight_layout()
plt.show()
```

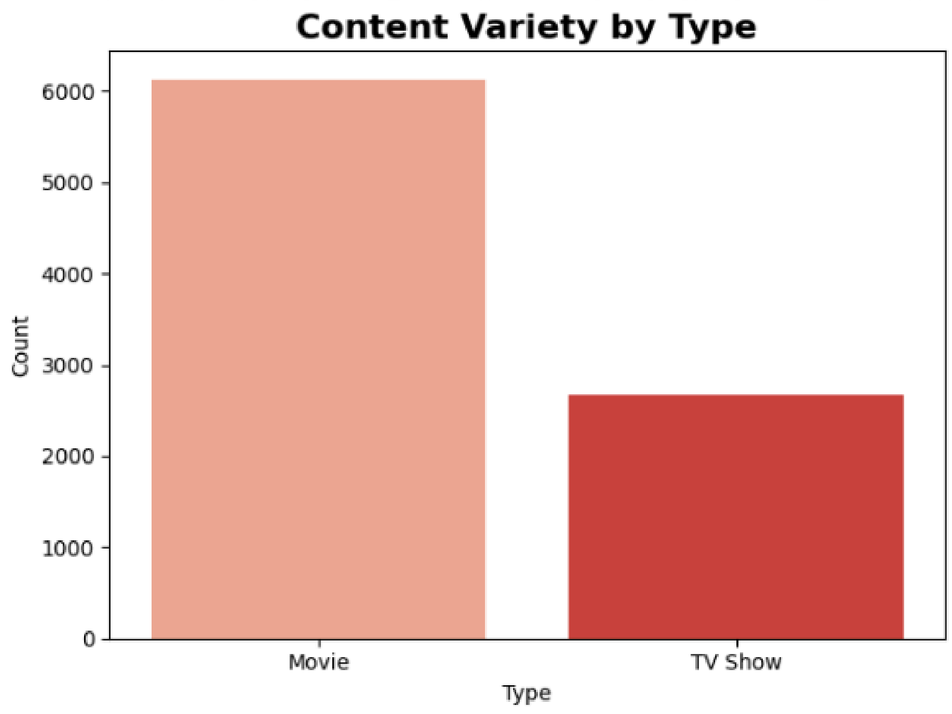


- The heatmap shows the correlation between a movie's release year and its duration.
- The correlation value is -0.21, indicating a weak negative relationship.
- This means newer movies tend to have slightly shorter durations than older ones.
- The color scale helps visualize the strength and direction of the correlation.

Content variety by type

```
[ ]
genre_counts = df['main_genre'].nunique()
print(f"Netflix offers {genre_counts} unique genres.")

sns.barplot(x=df['type'].value_counts().index, y=df['type'].value_counts().values, palette='Reds')
plt.title('Content Variety by Type', fontsize=16, fontweight='bold')
plt.xlabel('Type')
plt.ylabel('Count')
plt.tight_layout()
plt.show()
```

```
Netflix offers 36 unique genres.
/tmp/ipython-input-3277821270.py:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=df['type'].value_counts().index, y=df['type'].value_counts().values, palette='Reds')
```

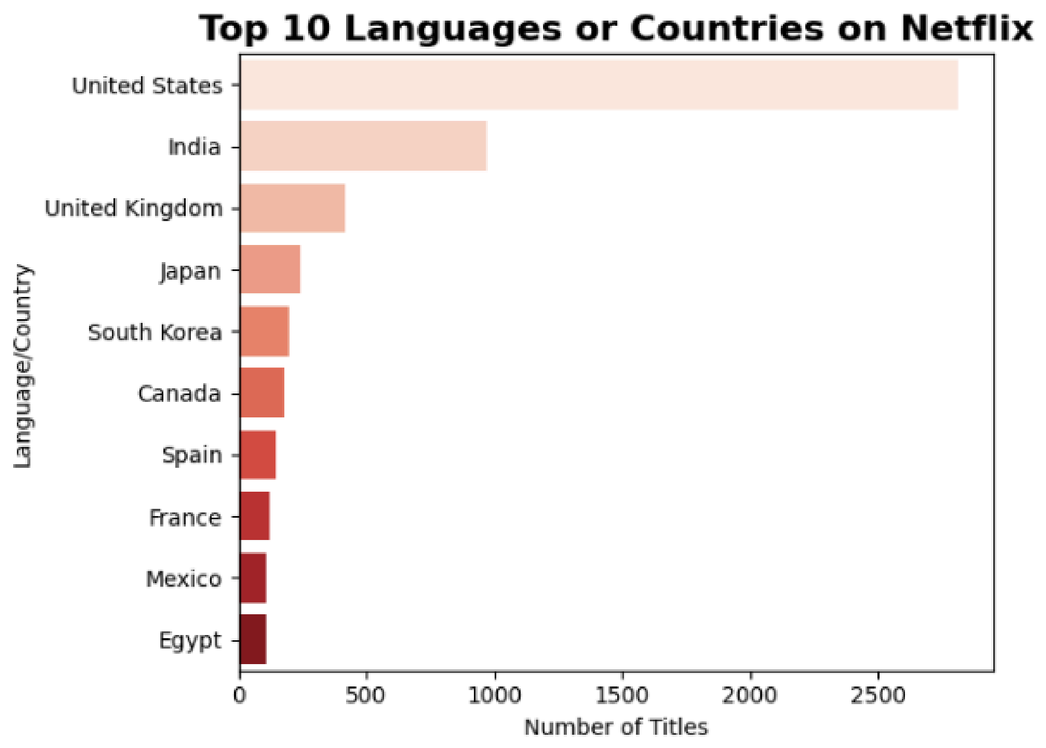


- The chart shows the content variety on Netflix by type — Movies and TV Shows.
- Movies dominate the platform with a much higher count than TV shows.
- The dataset includes 36 unique genres, offering diverse entertainment options.
- A red color palette is used for better visual representation.

Language analysis

```
[ ]
language_df = df['country'].value_counts().head(10).reset_index()
language_df.columns = ['Country', 'Count']

sns.barplot(data=language_df, x='Count', y='Country', palette='Reds')
plt.title('Top 10 Languages or Countries on Netflix', fontsize=16, fontweight='bold')
plt.xlabel('Number of Titles')
plt.ylabel('Language/Country')
plt.tight_layout()
plt.show()
```

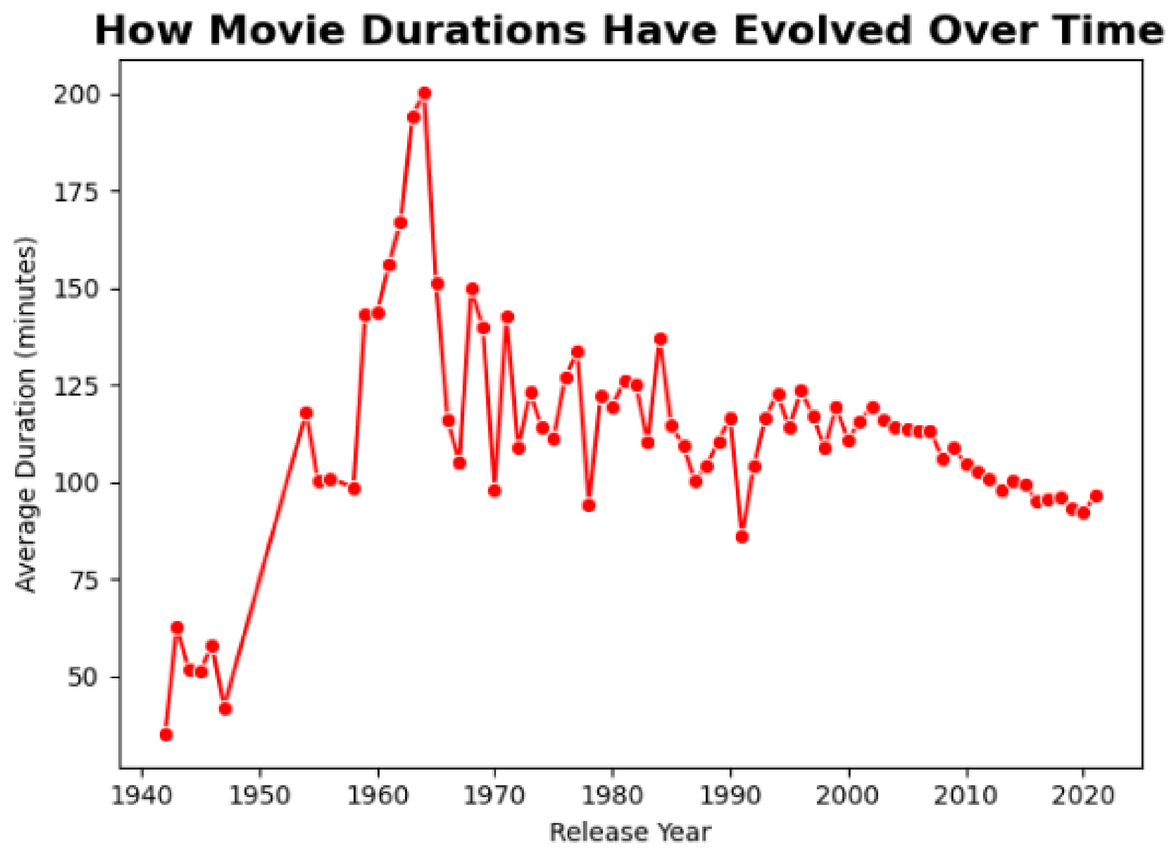
```
/tmp/ipython-input-3027894686.py:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(data=language_df, x='Count', y='Country', palette='Reds')
```



- The bar chart displays the top 10 countries or languages with the most Netflix titles.
- The United States ranks first, followed by India and the United Kingdom.
- This shows Netflix's major content production and availability regions.
- A red color palette is used for clear and appealing visualization.

Content evolution over time

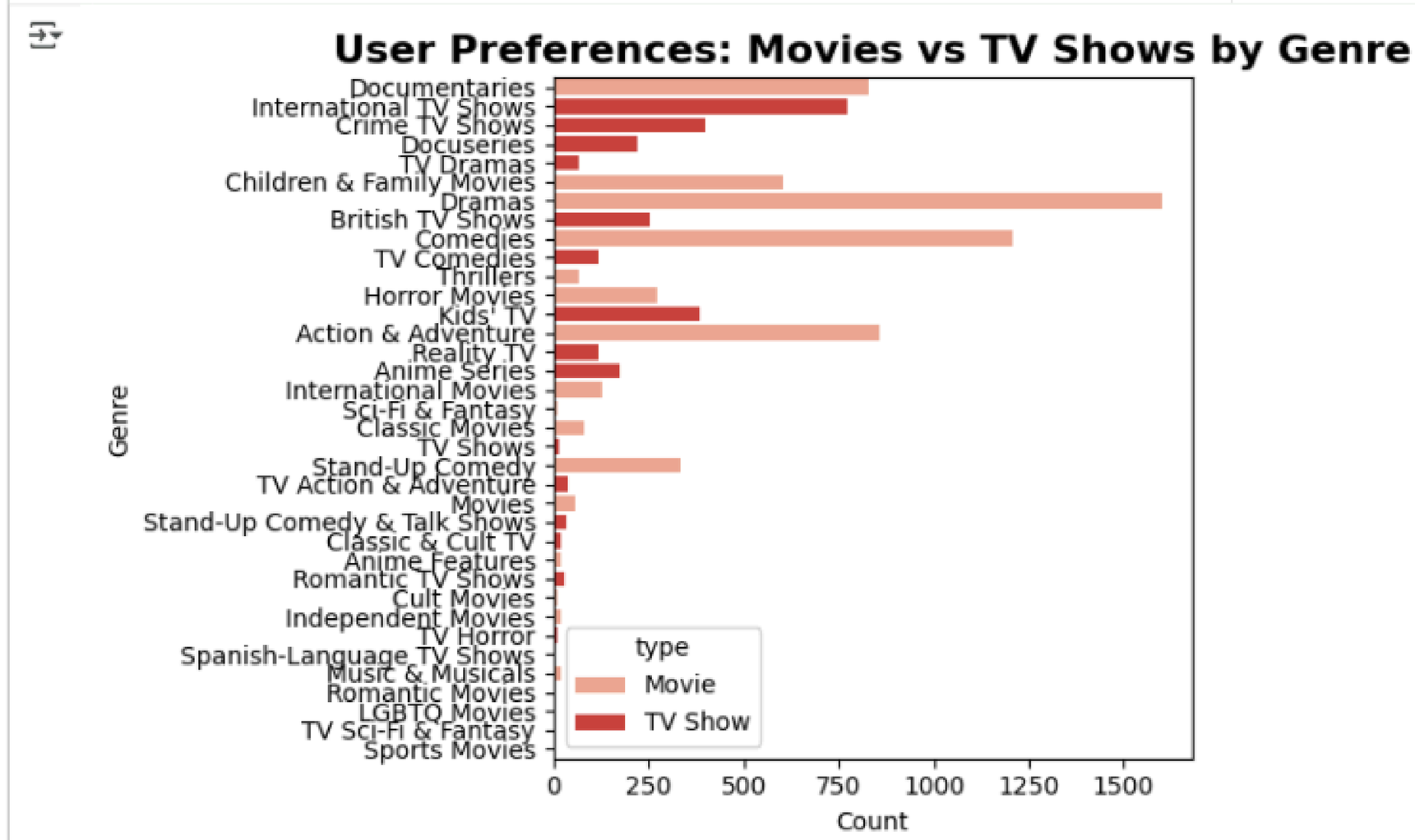
```
[ ]
evolution = movies.groupby('release_year')['duration_int'].mean().reset_index()

sns.lineplot(data=evolution, x='release_year', y='duration_int', color='red', marker='o')
plt.title('How Movie Durations Have Evolved Over Time', fontsize=16, fontweight='bold')
plt.xlabel('Release Year')
plt.ylabel('Average Duration (minutes)')
plt.tight_layout()
plt.show()
```



- The line chart shows how average movie durations have changed over the years.
- Movie lengths peaked around the 1960s but gradually declined after 2000.
- This indicates a trend toward shorter movies in recent years.
- The red line with dots makes the time-based trend clear and easy to follow.

User preferences

```
[ ]  sns.countplot(data=df, y='main_genre', hue='type', palette='Reds')
     plt.title('User Preferences: Movies vs TV Shows by Genre', fontsize=16, fontweight='bold')
     plt.xlabel('Count')
     plt.ylabel('Genre')
     plt.tight_layout()
     plt.show()
```



- The chart compares user preferences for Movies vs TV Shows across different genres.
- Dramas, comedies, and documentaries are the most popular genres overall.
- Movies dominate most genres, while TV shows are strong in international and children’s content.
- The red color shades highlight the contrast between movie and TV show counts.

Conclusion

- Netflix’s content library has grown rapidly since 2015, showing its strong global expansion.
- Movies make up most of the catalog, but TV shows are increasing steadily.
- Drama, Comedy, and Documentaries are the most popular genres.
- The U.S. leads in content production, followed by India and the U.K.
- Most content is rated TV-MA, focusing on mature audiences.
- Average movie duration is around 90–100 minutes, indicating shorter storytelling trends.
- Netflix’s content is becoming more diverse and globally inclusive each year.

Recommendations

- Invest more in regional and non-English content to reach wider audiences.
- Produce more family-friendly and balanced-rated shows.
- Continue focusing on Drama, Thriller, and Documentary genres.
- Encourage international collaborations and cultural storytelling.
- Experiment with short-form and interactive content to boost engagement.
- Use data-driven insights to tailor recommendations and content strategies.

THANK YOU!

Click here for the google collab file