📊 Netflix EDA: Genre Trends, Release Years, and More

Welcome! This notebook contains an exploratory data analysis (EDA) of Netflix titles using Python. We'll explore genre popularity, content release trends over the years, rating distribution, and top-producing countries.

Tools used: pandas, matplotlib, seaborn.

Dataset: Netflix Titles (from Kaggle)

In [1]:

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

# read the data and preview first 5 lines of the data
df = pd.read_csv("/kaggle/input/netflix-shows/netflix_titles.csv")
df.head()
```

Out[1]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duratio
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 mi
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	Season
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Seaso
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Seaso
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	Season

🔍 Quick Look at the Dataset

Let's check what columns we have and how the data looks.

In [2]:
```
# to check the dataset structure
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
```

🧹 Data Cleaning

Before diving into exploration, we need to check for missing values and standardize certain columns (e.g., convert dates to datetime, split multiple genres, etc.). Cleaning ensures that our analysis will be accurate and consistent.

In [3]:
```
df.isnull().sum().sort_values(ascending=False)
```

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

In [4]:
```
# Drop rows with missing values in key fields
df_clean = df.dropna(subset=['rating', 'date_added'])

# Fill other less-critical columns (optional)
df_clean.loc[:, 'country'] = df_clean['country'].fillna('Unknown')
```

📊 Plot the data set

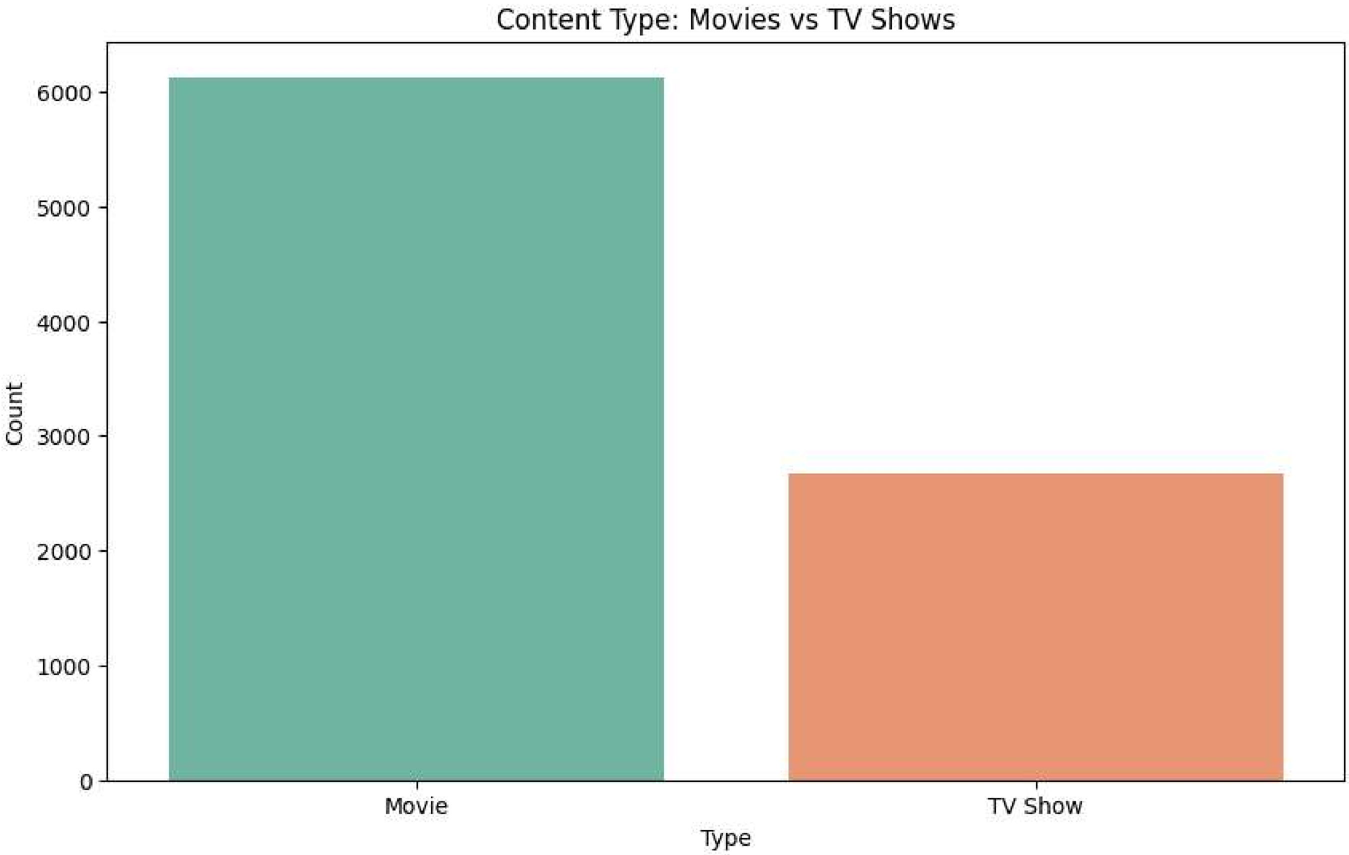
For gain the overview of the dataset, we can plot the data with some ways, bar plot, histogram, etc. Further analysis also can be done after we plot the dataset

1. 🎬 Content Type: Movies vs TV Shows

Let’s look at the distribution of content types available on Netflix. Is it dominated by movies or TV shows?

In [5]:

```
plt.figure(figsize=(10,6))
sns.countplot(data=df_clean, x='type', palette='Set2')
plt.title("Content Type: Movies vs TV Shows")
plt.xlabel("Type")
plt.ylabel("Count")
plt.show()
```



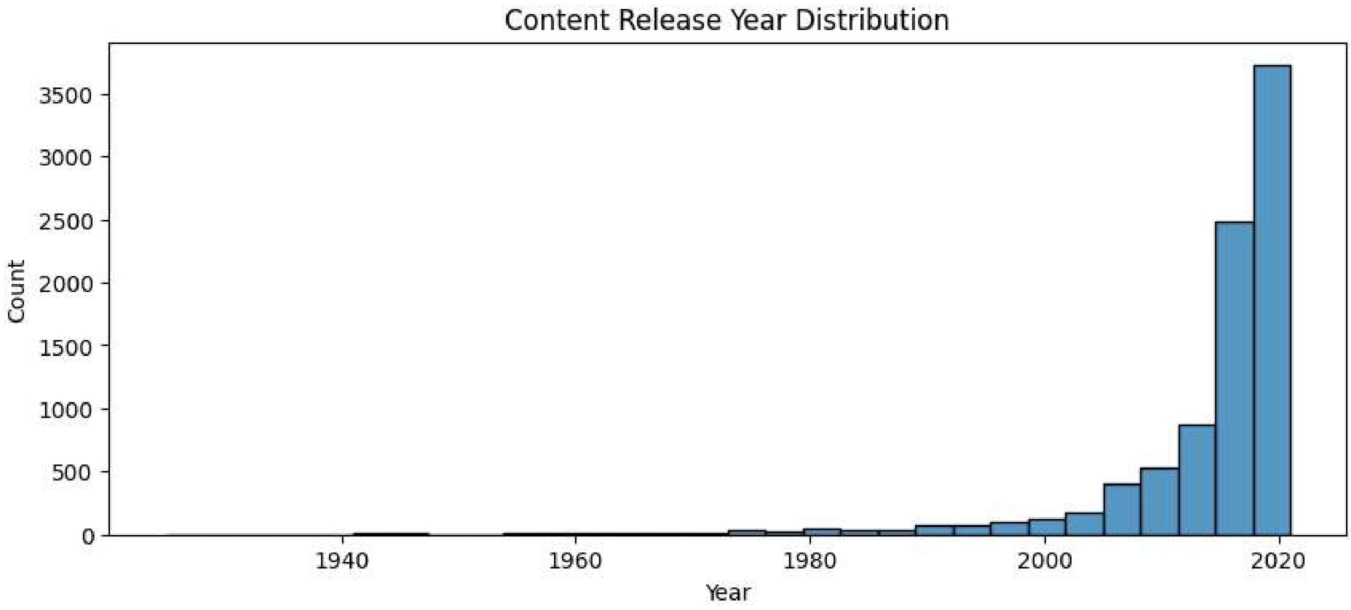
2. 📈 Trend Over the Years

How has Netflix's content library evolved over the years? We’ll explore the number of releases by year to see trends in new content being added.

In [6]:

```
plt.figure(figsize=(10,4))
sns.histplot(df_clean['release_year'], bins=30, kde=False)
plt.title("Content Release Year Distribution")
plt.xlabel("Year")
plt.ylabel("Count")
plt.show()
```

```
/usr/local/lib/python3.11/dist-packages/seaborn/_oldcore.py:1119: FutureWarning: use_inf_as_na
option is deprecated and will be removed in a future version. Convert inf values to NaN before
operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



3. 🎭 Top Genres on Netflix

Which genres are most common on Netflix?

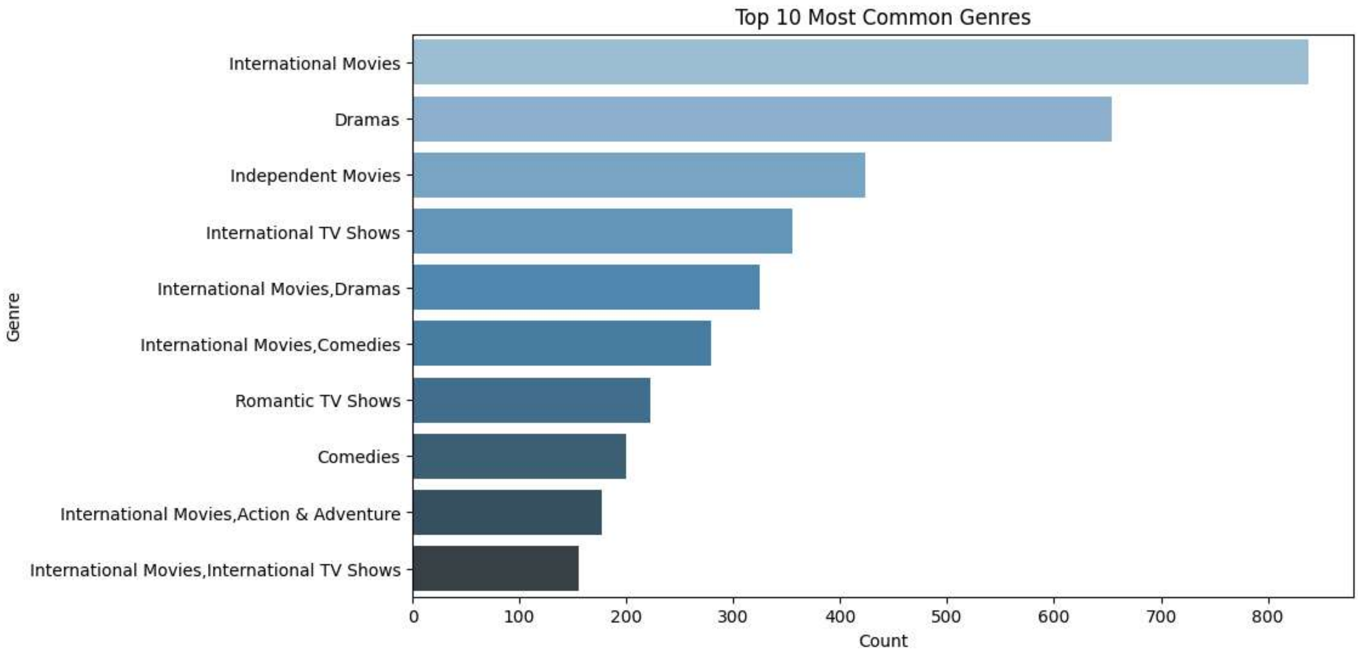
> 🔍 **Insight:** Netflix has a strong focus on dramas and international content — especially content classified under "Documentaries" or "Comedies".

In [7]:
```
from collections import Counter

all_genres = ','.join(df_clean['listed_in'].dropna()).split(', ')
genre_counts = Counter(all_genres)

genres_df = pd.DataFrame(genre_counts.items(), columns=['Genre', 'Count']).sort_values(by='Co

plt.figure(figsize=(10,6))
sns.barplot(data=genres_df.head(10), x='Count', y='Genre', palette='Blues_d')
plt.title("Top 10 Most Common Genres")
plt.show()
```

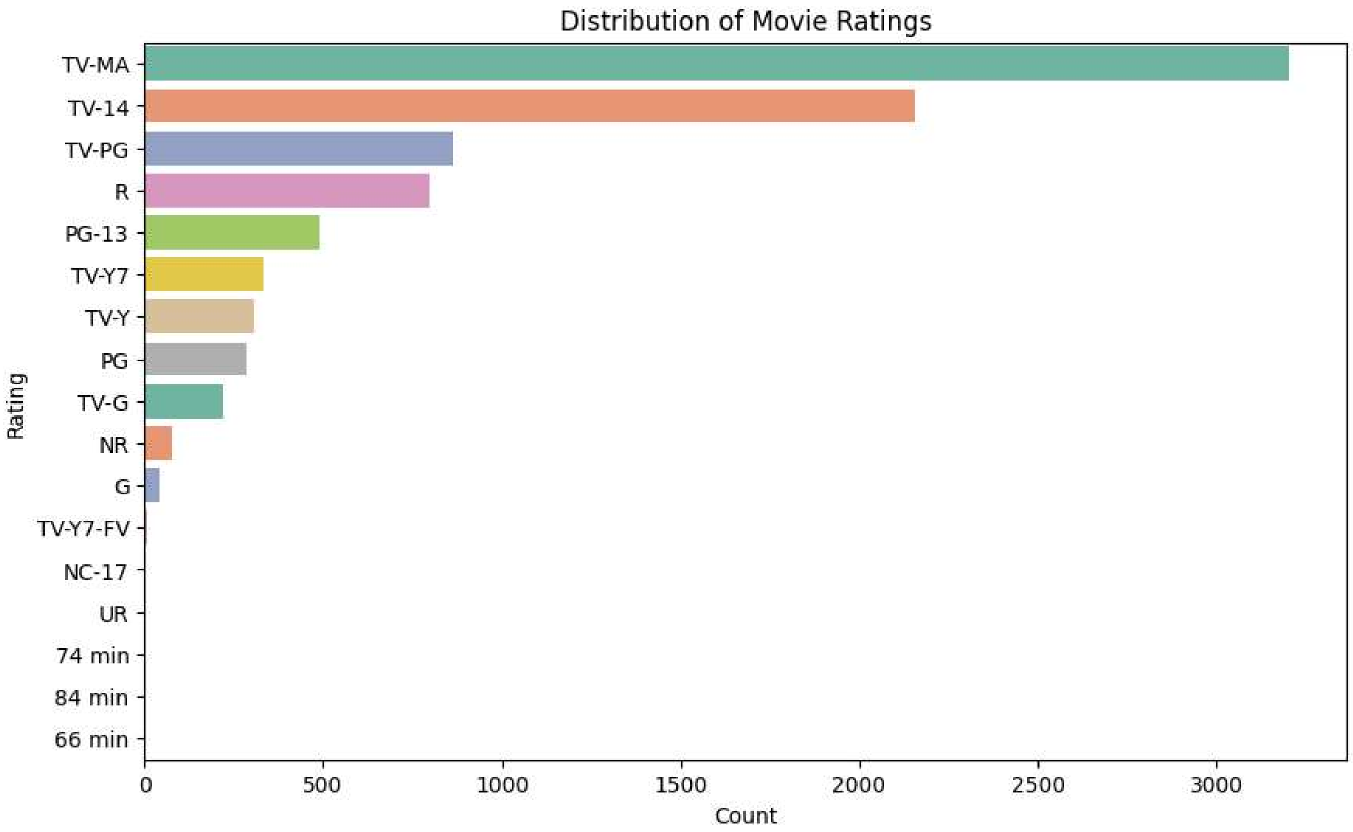


4. 🔹 Rating Distribution

This chart shows how content is distributed across different rating categories, such as TV-MA, TV-14, PG, etc.

In [8]:

```
plt.figure(figsize=(10,6))
sns.countplot(data=df_clean, y='rating', order=df_clean['rating'].value_counts().index, palet
plt.title('Distribution of Movie Ratings')
plt.xlabel('Count')
plt.ylabel('Rating')
plt.show()
```



Let's look at how Netflix content is rated. Most shows are suitable for mature audiences (TV-MA), followed by PG and TV-14. This tells us about the platform’s demographic focus.

5. 🌍 Country-wise Content Production

Which countries contribute the most content on Netflix?

We analyze the number of titles produced by country, using the cleaned `country` column. Since some entries contain multiple countries, only the first listed country is used for consistency.

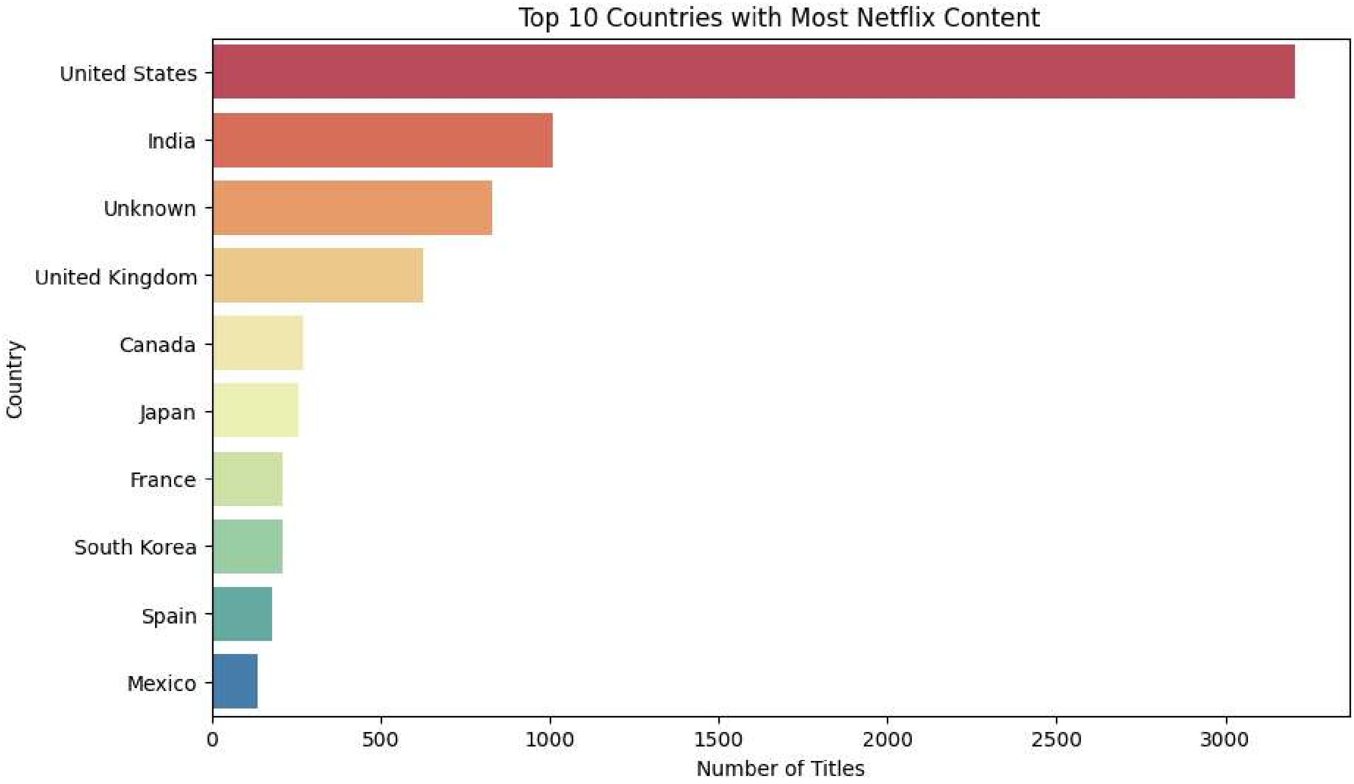
> "Spoiler: the US still dominates — but which countries follow?"

In [10]:

```
# Extract only the first country if multiple are listed
df_clean.loc[:, 'country_main'] = df_clean['country'].apply(lambda x: x.split(',')[0].strip()

# Plot top 10 countries
top_countries = df_clean['country_main'].value_counts().head(10)

plt.figure(figsize=(10,6))
sns.barplot(x=top_countries.values, y=top_countries.index, palette='Spectral')
plt.title('Top 10 Countries with Most Netflix Content')
plt.xlabel('Number of Titles')
plt.ylabel('Country')
plt.show()
```



📌 Insights & Summary

- Netflix has more **Movies** than TV Shows in this dataset.
- Most content was released in recent years (especially after 2015).
- **Dramas**, and **Comedies** are the most common genres.
- There's a steady increase in content added over the years until around 2019–2020.
- The United States is the leading country in terms of content production, followed by India.

📎 This is a simple starter EDA – more advanced filtering and modeling can follow in the next steps!

🧠 What's Next?

- Try to analyze the rating distribution
- Group content by country or director
- Explore co-occurrence of genres using NLP (multi-label)

Follow me for updates!