NETFLIX

What is this Business Case study?

The study's main objective is to analyse a large dataset that contains numerous characteristics of Netflix-available films and TV series. The details include the title, director, and actors as well as additional specific information like the description, rating, and year of release. The main goal is to derive practical insights that could assist Netflix in making wise choices about the production, distribution, and curation of content.

When it comes to media and video streaming services, Netflix is a dominant force. By mid-2021, this platform had an impressive library with over 10,000 films and TV series available for its global user base to enjoy. It's impressive that Netflix has over 222 million subscribers worldwide.

Why?

The ultimate objective is to increase consumer satisfaction, increase Netflix's subscriber base, and optimize its return on content investment. In order to accomplish this, gaps in the material collection are found, viewer preferences across a range of factors (genre, country, seasonality, etc.) are understood, and data-supported recommendations for future strategy are made.

That's what little I know about Netflix. To gain a deeper insight, let's delve deeper into the dataset I possess. Notebook Connection.

```
# Importing necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
dataset_path = '/netflix.csv'
netflix_data = pd.read_csv(dataset_path)

# Display first few rows of the dataset
netflix_data.head(5)
```

	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 Mabalane, 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 I...

For this analysis, we'll use Python and rely on libraries like Pandas, NumPy, Matplotlib, and Seaborn.

This tabular information includes an exhaustive list of every movie and TV series that is available on Netflix. This data improves the viewing experience for Netflix consumers globally by offering a multitude of information about the cast, directors, ratings, release years, lengths, and more.

Basic Metrics:

Firstly, let's examine the shape and basic statistics of the dataset.

Dataset Shape and Data Types: Let's check the shape and data types of the dataset.

```
# Checking the shape of the dataset
dataset_shape = netflix_data.shape

# Checking data types of the attributes
dataset_dtypes = netflix_data.dtypes

dataset_shape, dataset_dtypes
```

The dataset consists of 8,807 entries with 12 attributes:

```
((8807, 12),
 show_id         object
 type            object
 title           object
 director        object
 cast            object
 country         object
 date_added      object
 release_year     int64
 rating          object
 duration        object
 listed_in       object
 description     object
 dtype: object)
```

- **show_id:** Unique ID for every Movie / TV show
- **type:** Identifier — A Movie or TV Show
- **title:** Title of the Movie / TV Show
- **director:** Director of the Movie
- **cast:** Actors involved in the movie/show
- **country:** The country where the movie/show was produced
- **date_added:** Date it was added on Netflix
- **release_year:** Actual Release year of the movie/show
- **rating:** TV Rating of the movie/show
- **duration:** Total Duration — in minutes or number of seasons
- **listed_in: Genre**
- **description:** The summary description

These are some common fields that we can see on Netflix.

Analysing Basic Metrics:

```
# Displaying basic metrics using the describe() method for numerical columns
basic_metrics_numerical = netflix_data.describe()

# Displaying basic metrics for categorical columns like 'Type', 'Country', and 'Rating'
basic_metrics_categorical = netflix_data[['type', 'country', 'rating']].describe(include=['object'])

basic_metrics_numerical, basic_metrics_categorical
```

Numerical Attributes

For the numerical attribute `release_year`:

- **Count:** 8,807 entries
- **Mean:** Around the year 2014
- **Standard Deviation:** Approximately 8.82 years
- **Minimum:** Year 1925
- **25th Percentile (Q1):** Year 2013
- **Median (50th Percentile):** Year 2017
- **75th Percentile (Q3):** Year 2019
- **Maximum:** Year 2021

```
(       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,
          type        country rating
 count    8807           7976   8803
 unique      2            748     17
 top     Movie  United States  TV-MA
 freq     6131           2818   3207)
```

Categorical Attributes

For the categorical attributes type, country, and rating:

Type

- **Count:** 8,807
- **Unique Values:** 2 (Movie, TV Show)
- **Most Frequent:** Movie
- **Frequency:** 6,131

Country

- **Count:** 7,976 (some missing values)
- **Unique Values:** 748
- **Most Frequent:** United States
- **Frequency:** 2,818

Rating

- **Count:** 8,803 (some missing values)
- **Unique Values:** 17
- **Most Frequent:** TV-MA
- **Frequency:** 3,207

Observations:

1. Given that the median release year is 2017 and the average release year is approximately 2014, it appears that Netflix primarily consists of content from the last ten years.
2. A greater emphasis on movie content is indicated by the increased frequency of movies as opposed to TV shows.
3. There is a vast content library; the United States seems to be the most prevalent country for content development, followed by a wide range of other countries.
4. The rating “TV-MA” is the most frequent, suggesting a focus on mature audiences.

These fundamental indicators provide an overview of the type of material that is most popular on Netflix, which may be quite helpful when making different business decisions.

```
# Convert categorical attributes to 'category' data type if required
categorical_columns = ['type', 'country', 'rating']
netflix_data[categorical_columns] = netflix_data[categorical_columns].astype('category')

# After conversion data types
after_conversion_data_types = netflix_data.dtypes
# Missing value detection
missing_values = netflix_data.isnull().sum()
```

Data Types of All the Attributes (Before Conversion):

- Most of the attributes are of object data type, except release_year, which is an int64.

Conversion of Categorical Attributes to 'Category':

- Type, country, and rating data types have been transformed into categories.

Missing Value Detection:

- **director:** 2,634 missing values
- **cast:** 825 missing values
- **country:** 831 missing values
- **date_added:** 10 missing values
- **rating:** 4 missing values
- **duration:** 3 missing values

Non-Graphical Analysis: Value Counts and Unique Attributes:

```
# Non-Graphical Analysis: Value counts for key attributes
value_counts_type = netflix_data['type'].value_counts()
value_counts_country = netflix_data['country'].value_counts().head(10)  # Top 10 countries
value_counts_rating = netflix_data['rating'].value_counts()
value_counts_release_year = netflix_data['release_year'].value_counts().head(10)  # Top 10 release years

# Unique attributes for key columns
unique_type = netflix_data['type'].unique()
unique_country = netflix_data['country'].unique()
unique_rating = netflix_data['rating'].unique()
unique_release_year = netflix_data['release_year'].unique()

value_counts_type, value_counts_country, value_counts_rating, value_counts_release_year, unique_type, unique_country, unique_rating, unique_release_year
```

Value Counts:

Type of Content (Movies vs. TV Shows)

- **Movies:** 6,131
- **TV Shows:** 2,676

```
United Kingdom    419
Japan             245
South Korea       199
Canada            181
Spain             145
France            124
Mexico            110
Egypt             106
Name: count, dtype: int64,
```

Top 10 Countries Producing Content

- **United States:** 2,818
- **India:** 972
- **United Kingdom:** 419
- **Japan:** 245
- **South Korea:** 199
- **Canada:** 181
- **Spain:** 145
- **France:** 124
- **Mexico:** 110
- **Egypt:** 106

```
rating
TV-MA       3207
TV-14       2160
TV-PG        863
R            799
PG-13        490
TV-Y7        334
TV-Y         307
PG           287
TV-G         220
NR            80
G             41
TV-Y7-FV       6
UR             3
NC-17          3
74 min         1
84 min         1
66 min         1
Name: count, dtype: int64,
```

Ratings

- **TV-MA:** 3,207
- **TV-14:** 2,160
- **TV-PG:** 863
- **R:** 799
- **PG-13:** 490
- **…**

```
release_year
2018    1147
2017    1032
2019    1030
2020     953
2016     902
2021     592
2015     560
2014     352
2013     288
2012     237
Name: count, dtype: int64,
```

Top 10 Release Years

- **2018:** 1,147
- **2017:** 1,032
- **2019:** 1,030
- **2020:** 953
- **2016:** 902
- **…**

```
['Movie', 'TV Show']
Categories (2, object): ['Movie', 'TV Show'],
['United States', 'South Africa', NaN, 'India', 'United States, Ghana, Burkina Faso, United Ki..., ..., 'Russia, Spain', 'Croatia, Slovenia, Serbia, Montenegro', 'Japan, Canada',
'United States, France, South Korea, Indonesia', 'United Arab Emirates, Jordan']
Length: 749
Categories (748, object): [', France, Algeria', ', South Korea', 'Argentina',
                           'Argentina, Brazil, France, Poland, Germany, D..., ..., 'Venezuela, Colombia', 'Vietnam', 'West Germany',
                           'Zimbabwe'],
['PG-13', 'TV-MA', 'PG', 'TV-14', 'TV-PG', ..., '66 min', 'NR', NaN, 'TV-Y7-FV', 'UR']
Length: 18
Categories (17, object): ['66 min', '74 min', '84 min', 'G', ..., 'TV-Y', 'TV-Y7', 'TV-Y7-FV', 'UR'],
array([2020, 2021, 1993, 2018, 1996, 1998, 1997, 2010, 2013, 2017, 1975,
       1978, 1983, 1987, 2012, 2001, 2014, 2002, 2003, 2004, 2011, 2008,
       2009, 2007, 2005, 2006, 1994, 2015, 2019, 2016, 1982, 1989, 1990,
       1991, 1999, 1986, 1992, 1984, 1980, 1961, 2000, 1995, 1985, 1976,
       1959, 1988, 1981, 1972, 1964, 1945, 1954, 1979, 1958, 1956, 1963,
       1970, 1973, 1925, 1974, 1960, 1966, 1971, 1962, 1969, 1977, 1967,
       1968, 1965, 1946, 1942, 1955, 1944, 1947, 1943]))
```

Unique Attributes

- **Type:** 2 unique values (‘Movie’, ‘TV Show’)
- **Country:** 748 unique values

- **Rating:** 17 unique values
- **Release Year:** Ranges from 1925 to 2021

Observations:

- There are nearly twice as many movies available on the platform than there are TV shows.
- The top three countries in the world for content production are the United States, India, and the United Kingdom.
- The most prevalent ratings, "TV-MA" and "TV-14," suggest that mature and teen audiences are the target demographic.
- The majority of the content was published in 2018, 2017, and 2019, indicating a significant emphasis on recent content.

This non-graphical study offers a strong basis for comprehending the dataset's general makeup. It also provides insightful information on the kinds of content that Netflix users watch the most frequently.

Missing Value & Outlier Check:

We must make sure that there are no outliers or missing values before moving on to more visualizations.

Missing Values

Let's first check for missing values in the dataset.

```
# Checking for missing values
missing_values = netflix_data.isnull().sum()
missing_values
```

```
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
```

Observations:

We have missing values in several columns:

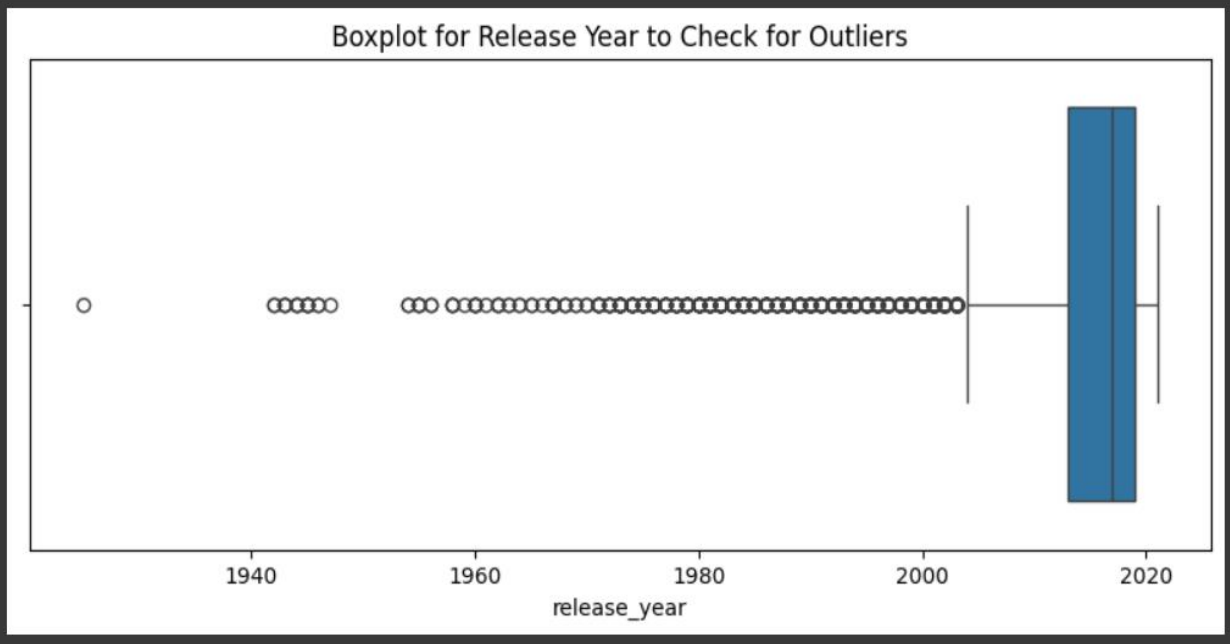
- **director:** 2,634 missing values
- **cast:** 825 missing values
- **country:** 831 missing values
- **date_added:** 10 missing values
- **rating:** 4 missing values
- **duration:** 3 missing values

These missing values may or may not have a substantial impact on the result, depending on the nature of our study. For our present business questions, for instance, missing director or cast information might not be essential.

Outliers:

For the purpose of this analysis, we'll focus on the release_year as our primary numerical variable. Let's check for outliers using a boxplot.

```
# Boxplot to check for outliers in 'release_year'
plt.figure(figsize=(10, 4))
sns.boxplot(x=netflix_data['release_year'])
plt.title('Boxplot for Release Year to Check for Outliers')
plt.show()
```



Observations:

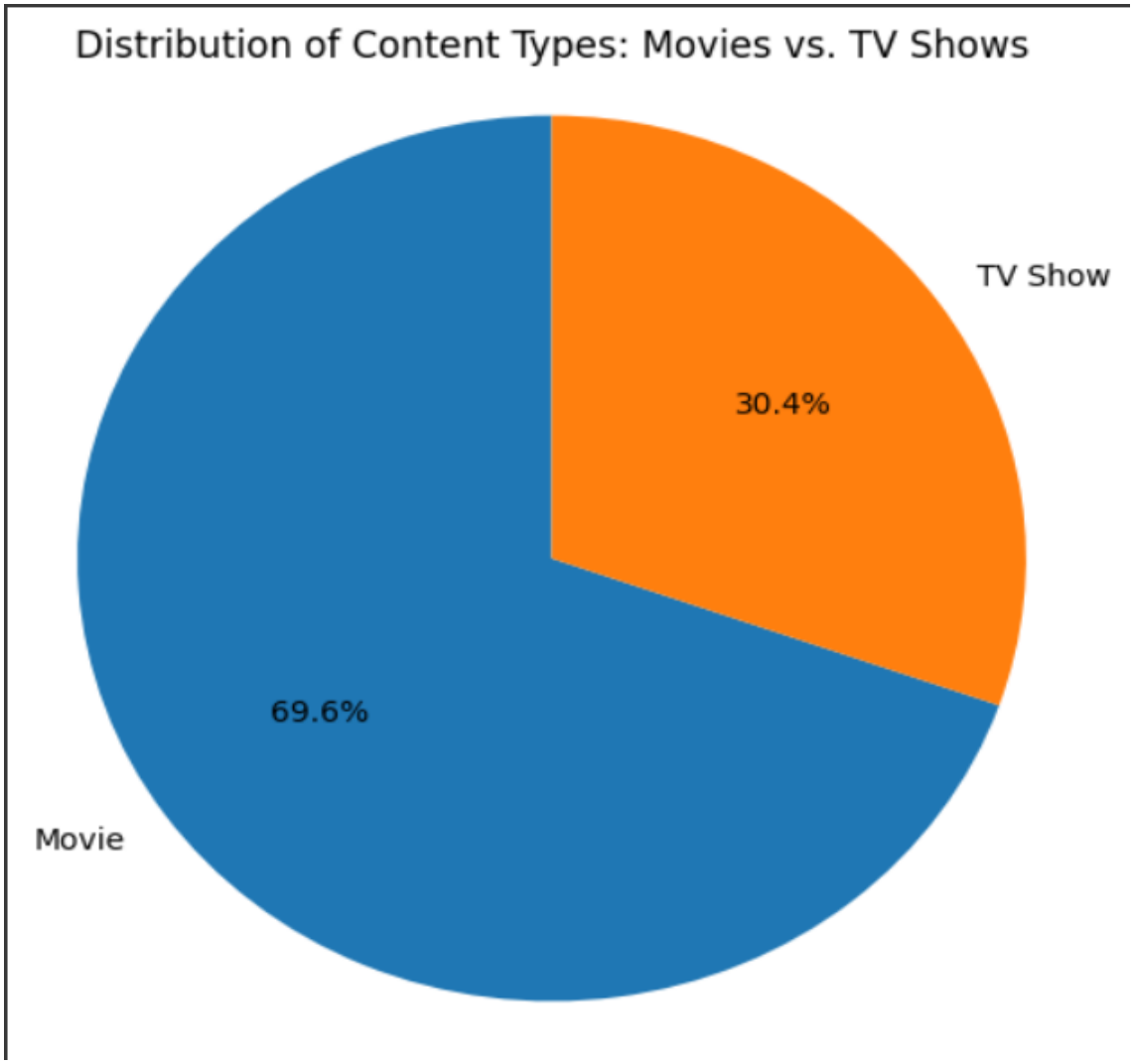
- There are no notable outliers in the boxplot for release_year, suggesting that the data for this property is generally consistent.

Univariate Analysis

Countplot by Type of Content

Let's begin by analysing how Netflix distributes movies differently from TV shows.

```
# Univariate Example with Pie Chart for 'Type' (Movie/TV Show)
type_counts = netflix_data['type'].value_counts()
labels = type_counts.index
sizes = type_counts.values

plt.figure(figsize=(8, 8))
plt.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90)
plt.title('Distribution of Content Types: Movies vs. TV Shows')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.show()
```

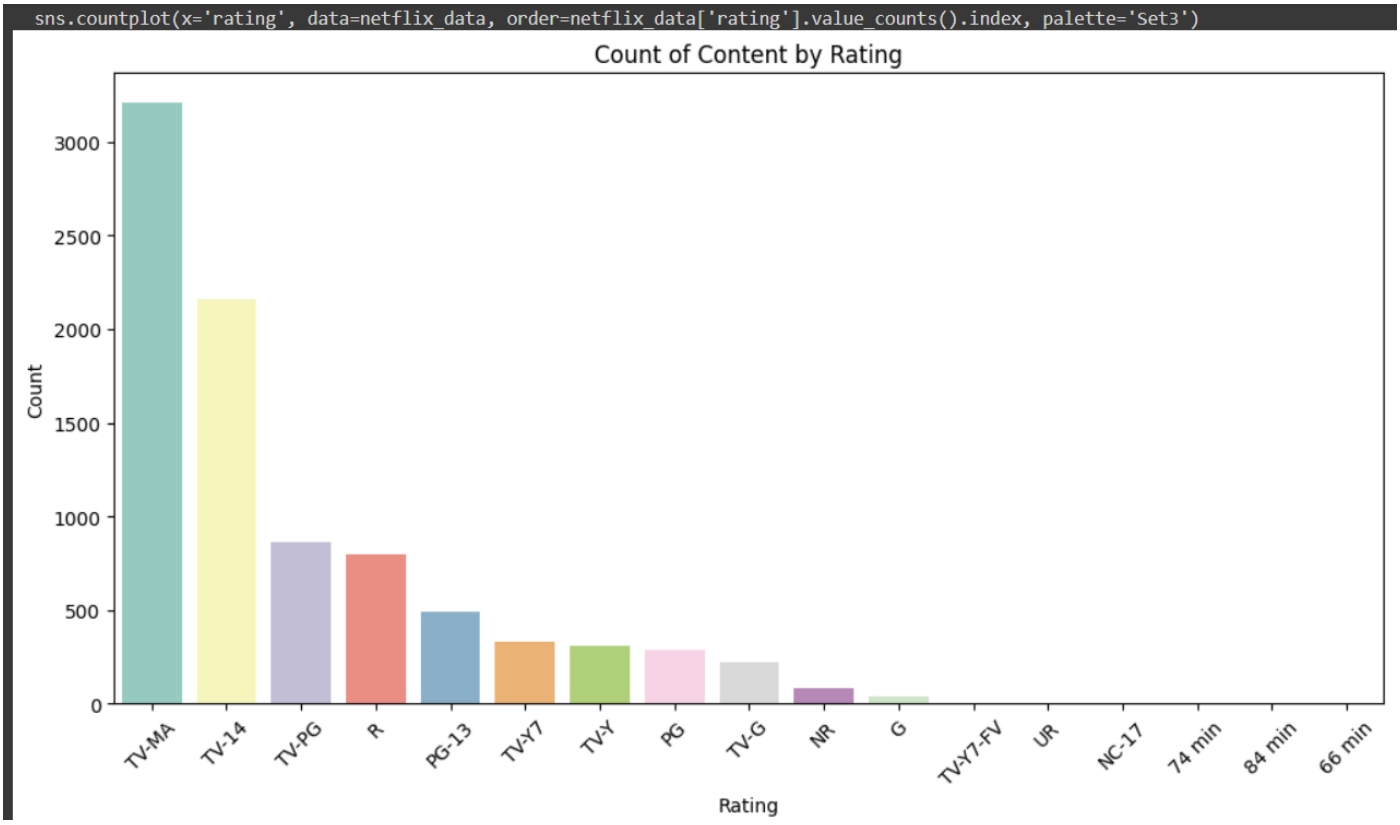


Observations:

- The number of movies is far more than the number of TV shows, suggesting that Netflix has a larger movie library.

Countplot for Rating:

```
# Countplot for Rating
plt.figure(figsize=(12, 6))
sns.countplot(x='rating', data=netflix_data, order=netflix_data['rating'].value_counts().index, palette='Set3')
plt.title('Count of Content by Rating')
plt.xlabel('Rating')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```



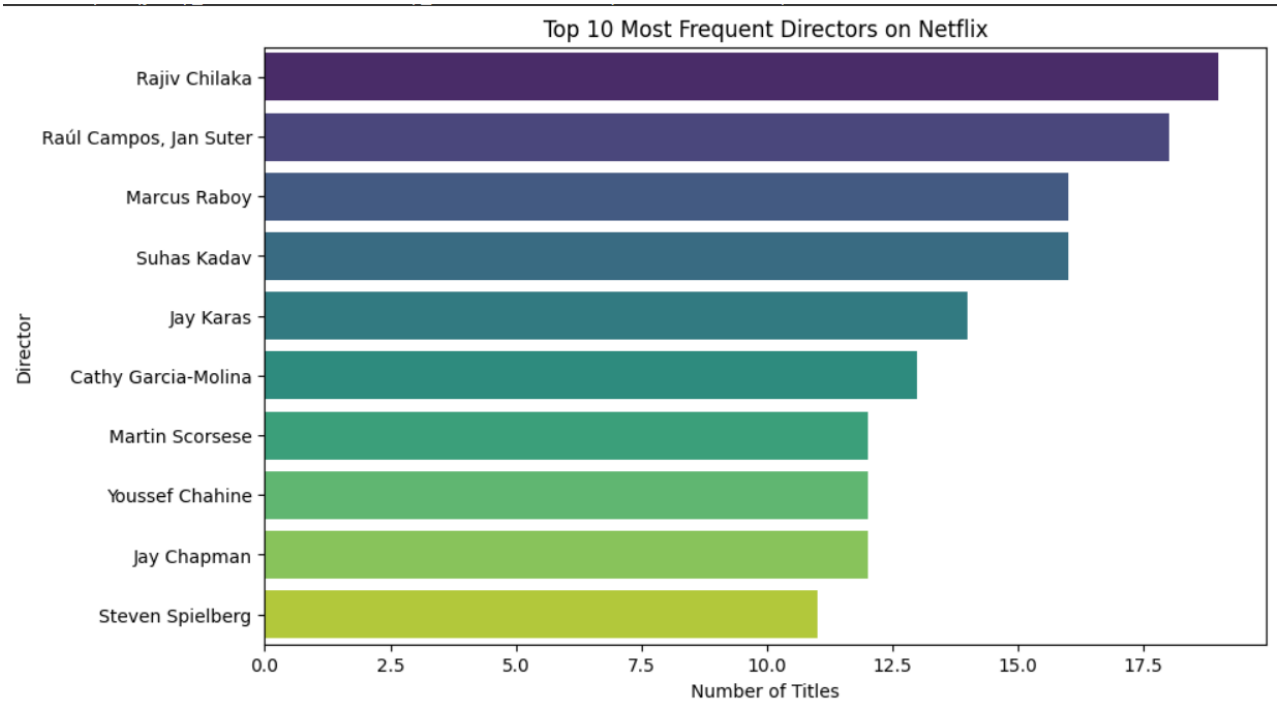
Observations:

- The majority of the programming is rated "TV-MA," with a concentration on mature audiences and teenagers, followed by "TV-14."

Top 10 Most Frequent Directors on Netflix:

The bar chart illustrates the top 10 directors with the most titles on Netflix:

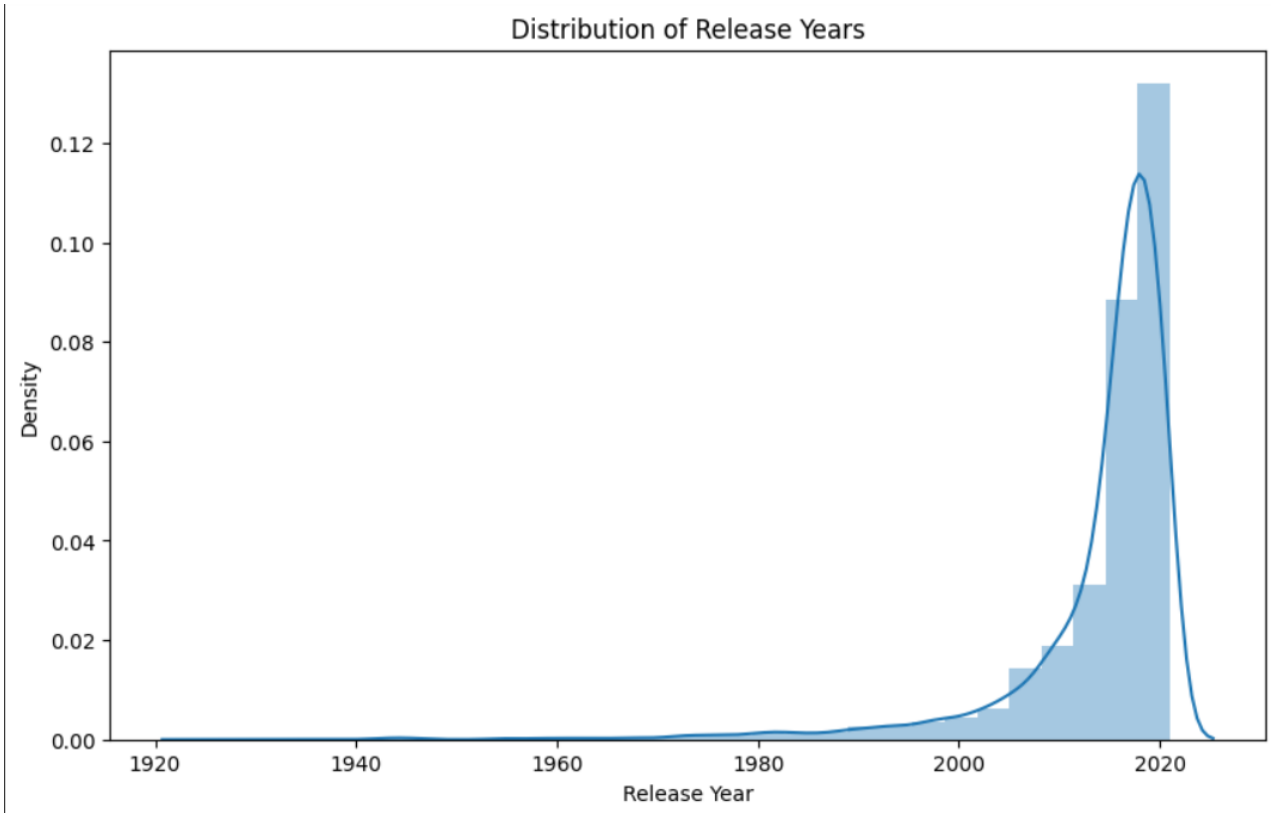
```python
top_directors = netflix_data['director'].value_counts().head(10)

# Visualizing the top 10 directors with a bar chart
plt.figure(figsize=(14, 8))
sns.barplot(y=top_directors.index, x=top_directors.values, palette='viridis')
plt.title('Top 10 Most Frequent Directors on Netflix')
plt.xlabel('Number of Titles')
plt.ylabel('Director')
plt.show()
```



Distplot & Histogram for Release Years:

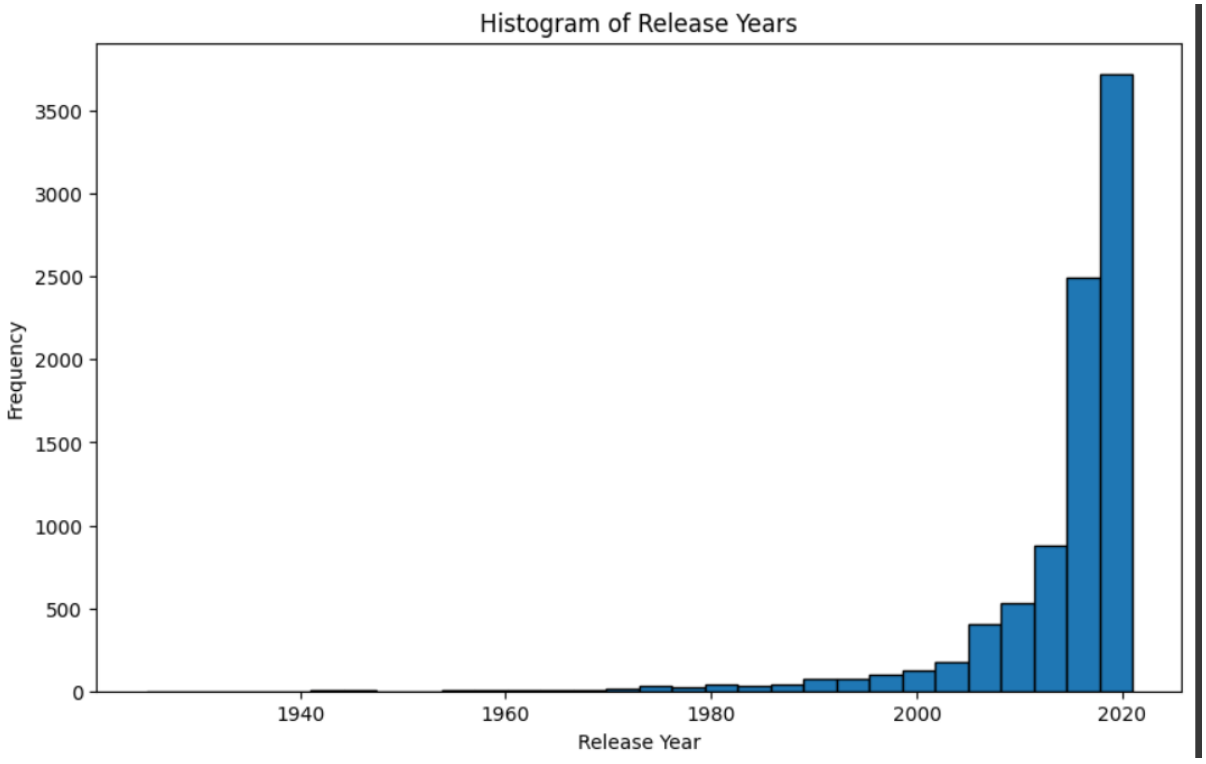
Here is the Distplot for release_year

```python
# Distplot for release_year
plt.figure(figsize=(10, 6))
sns.distplot(netflix_data['release_year'], kde=True, bins=30)
plt.title('Distribution of Release Years')
plt.xlabel('Release Year')
plt.ylabel('Density')
plt.show()
```



Histogram:

```
# Histogram for release_year
plt.figure(figsize=(10, 6))
plt.hist(netflix_data['release_year'], bins=30, edgecolor='black')
plt.title('Histogram of Release Years')
plt.xlabel('Release Year')
plt.ylabel('Frequency')
plt.show()
```

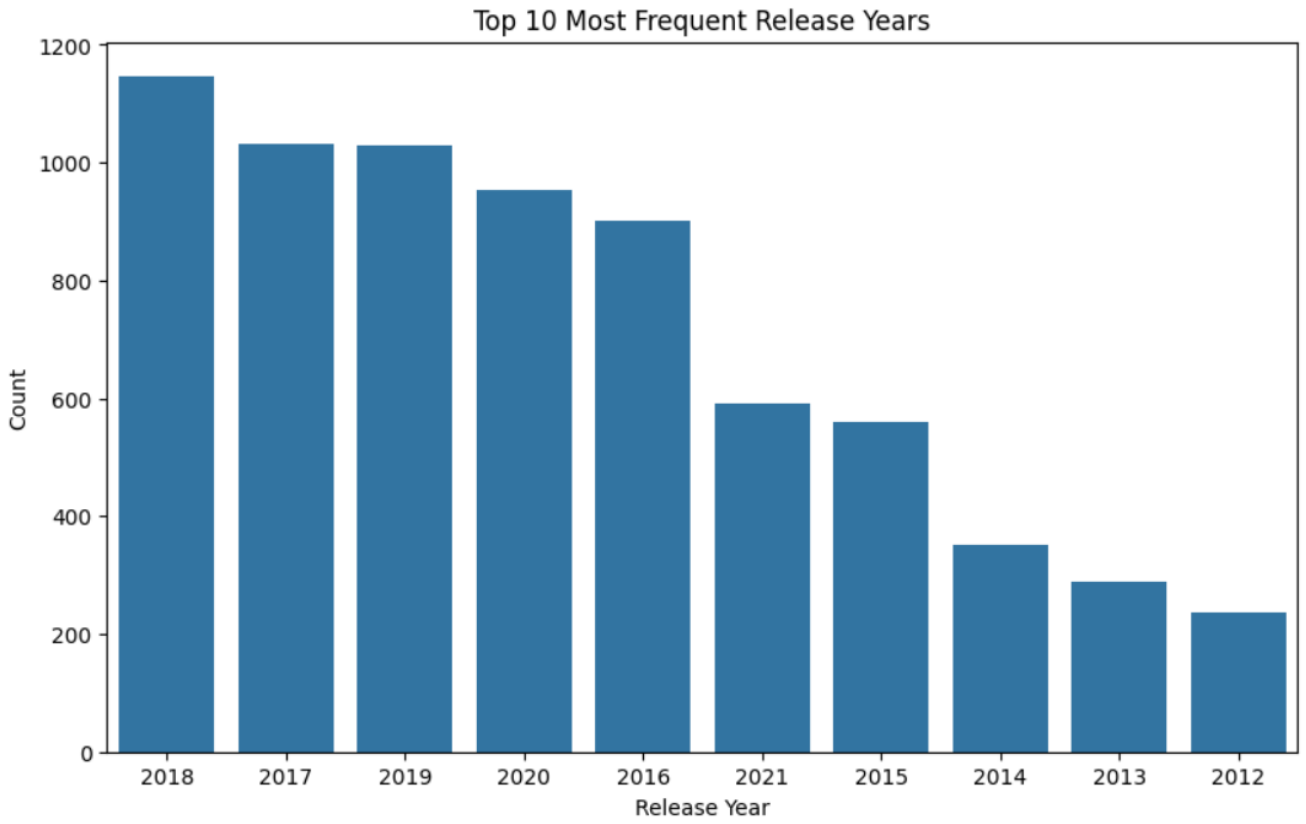


Observations:

The right-skewed distribution of release years suggests that a large portion of the content available on Netflix is quite recent, having been released within the last ten years.

Countplot for Top 10 Most Frequent Release Years:

```
# Countplot for top 10 release years
plt.figure(figsize=(10, 6))
sns.countplot(data=netflix_data, x='release_year', order=netflix_data['release_year'].value_counts().iloc[:10].index)
plt.title('Top 10 Most Frequent Release Years')
plt.xlabel('Release Year')
plt.ylabel('Count')
plt.show()
```



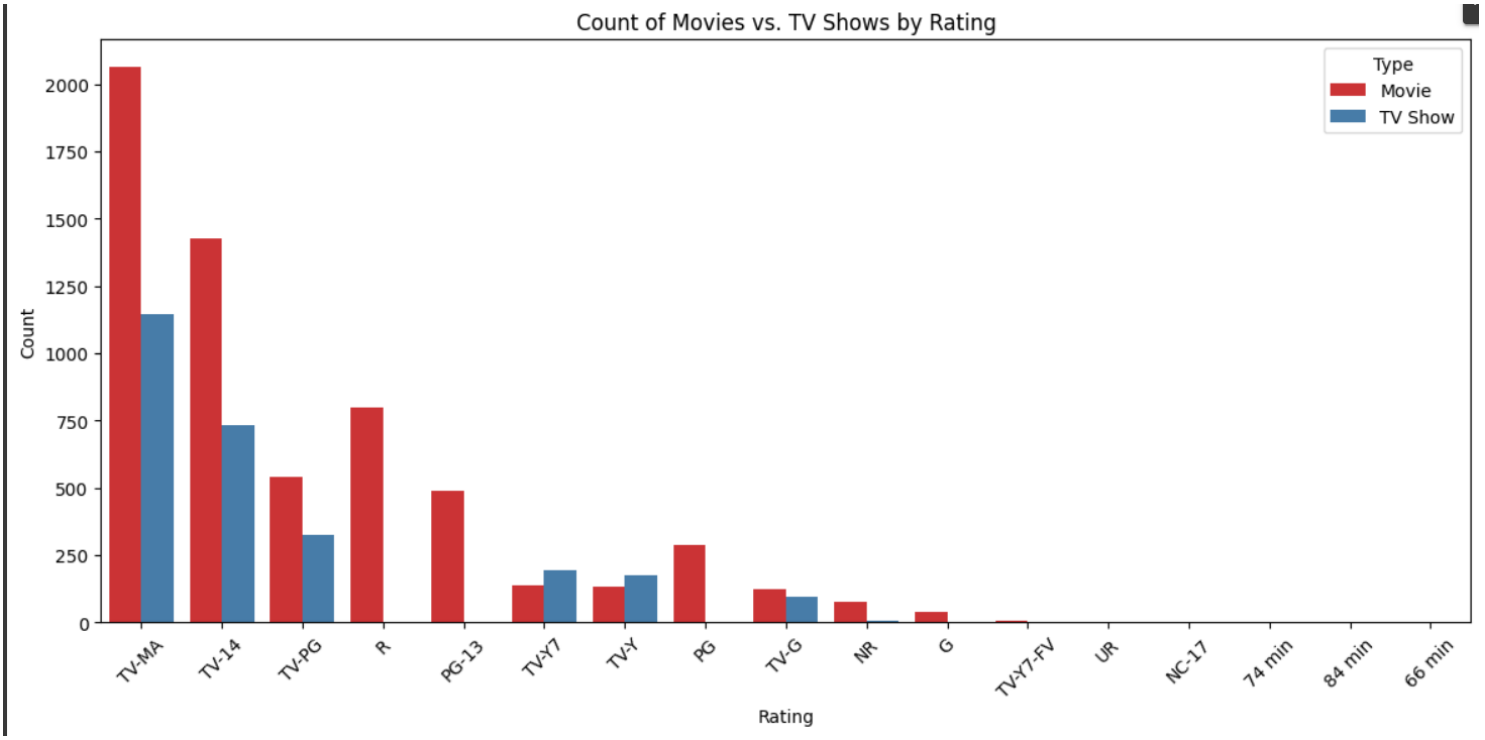
Observations:

2018 has the most content among the top 10 most frequent release years, all of which are from the recent past.

Bivariate Analysis

Relationship Between Type and Rating

```python
# Countplot for Type vs Rating
plt.figure(figsize=(14, 6))
sns.countplot(x='rating', hue='type', data=netflix_data, order=netflix_data['rating'].value_counts().index, palette='Set1')
plt.title('Count of Movies vs. TV Shows by Rating')
plt.xlabel('Rating')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.legend(title='Type')
plt.show()
```



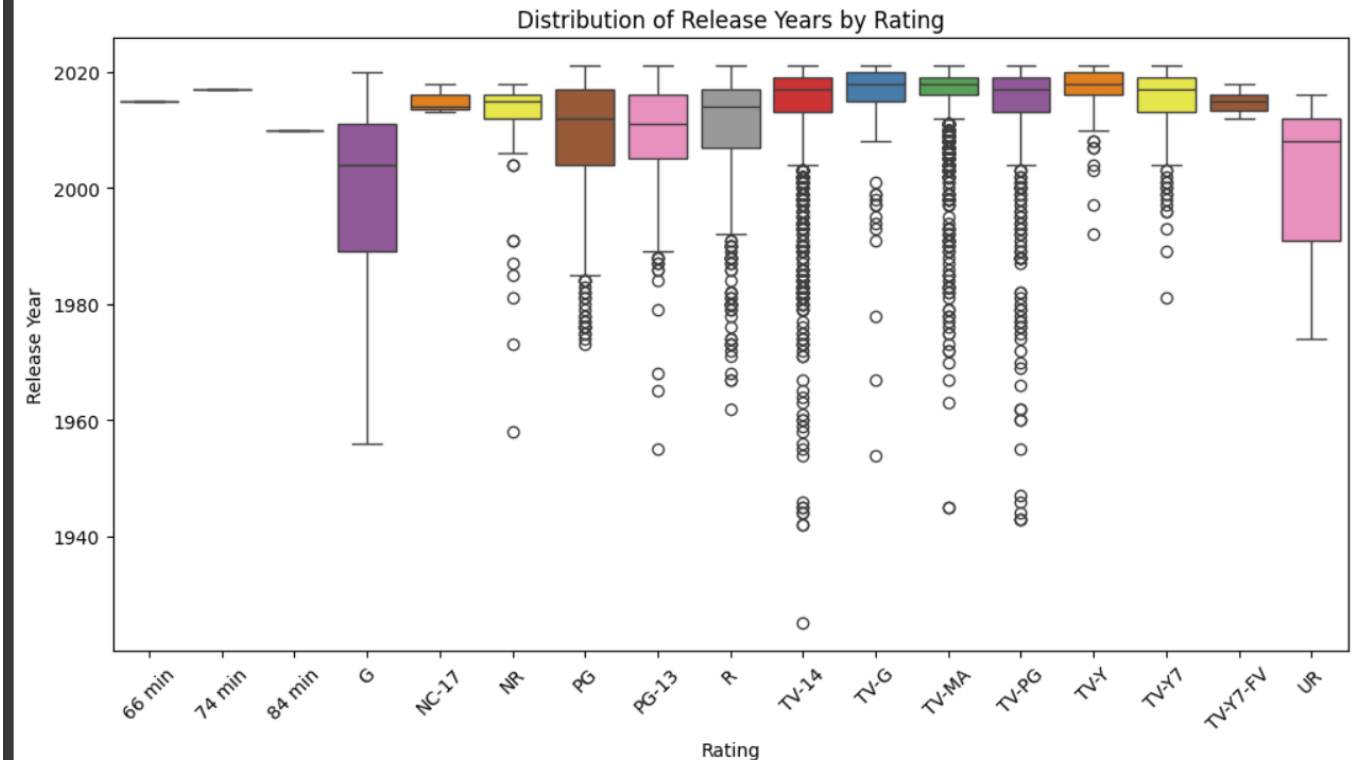
Observations:

- The "TV-MA" and "TV-14" ratings are mostly used for Movies and TV shows.
- While TV shows and Movies have different rating distributions, movies are rated higher overall and in most categories.

Relationship Between Rating and Release Year:

```python
# Boxplot for rating vs. release_year
plt.figure(figsize=(12, 6))
sns.boxplot(x='rating', y='release_year', data=netflix_data, palette='Set1')
plt.title('Distribution of Release Years by Rating')
plt.xlabel('Rating')
plt.ylabel('Release Year')
plt.xticks(rotation=45)
plt.show()
```

```
sns.boxplot(x='rating', y='release_year', data=netflix_data, palette='Set1')
```



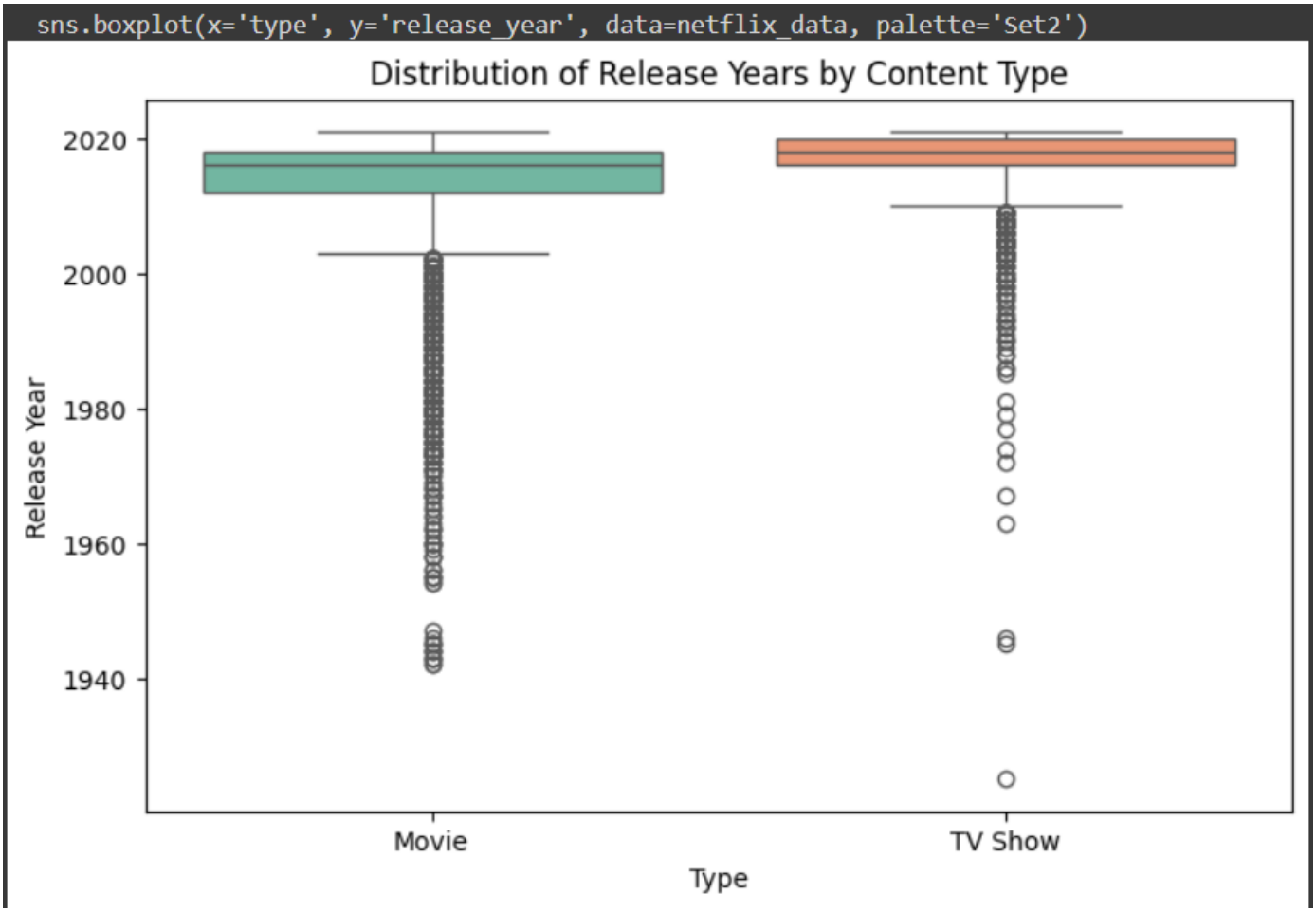
Observations:

- The boxplot indicates that the majority of ratings have a very recent median release year.
- Compared to other ratings, content with the "TV-Y" and "TV-Y7" classifications is typically older.

Relationship Between Type and Release Year:

```
# Boxplot for type vs. release_year
plt.figure(figsize=(8, 5))
sns.boxplot(x='type', y='release_year', data=netflix_data, palette='Set2')
plt.title('Distribution of Release Years by Content Type')
plt.xlabel('Type')
plt.ylabel('Release Year')
plt.show()
```

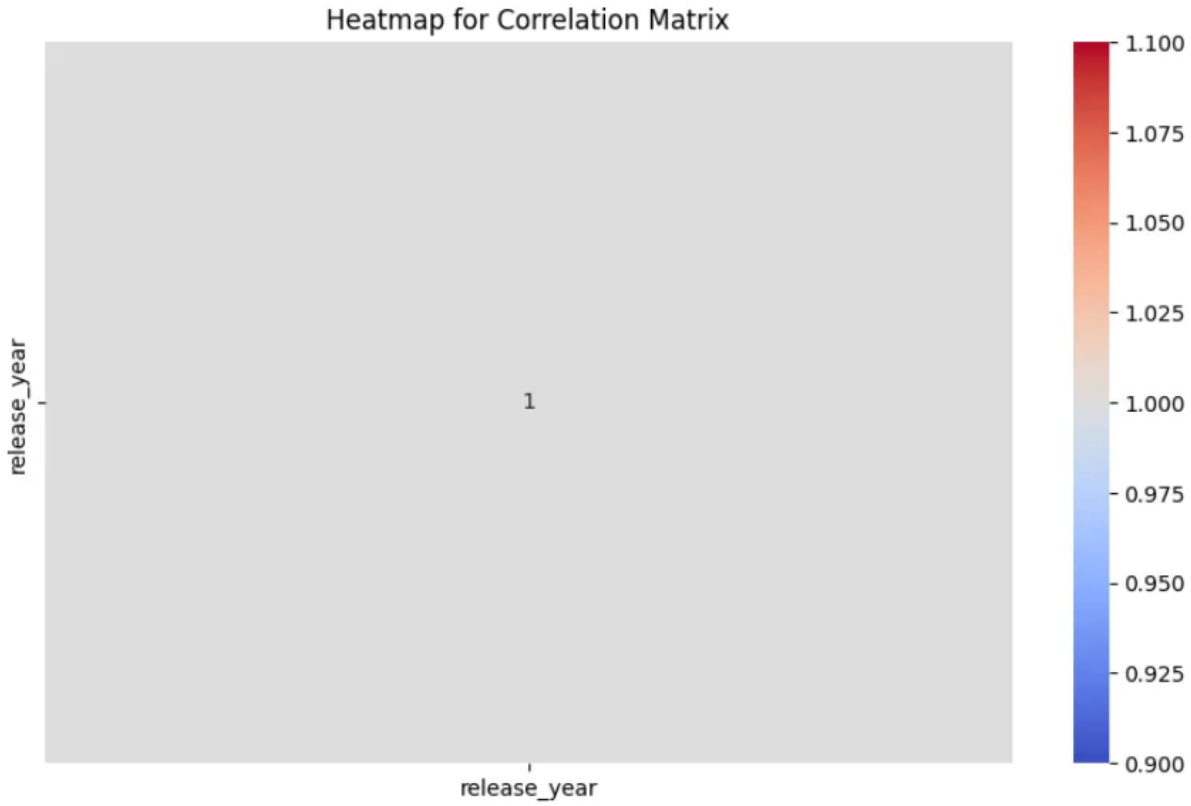
```
sns.boxplot(x='type', y='release_year', data=netflix_data, palette='Set2')
```



Correlation Analysis: Heatmaps and Pairplots:

Heatmap for Correlation Matrix

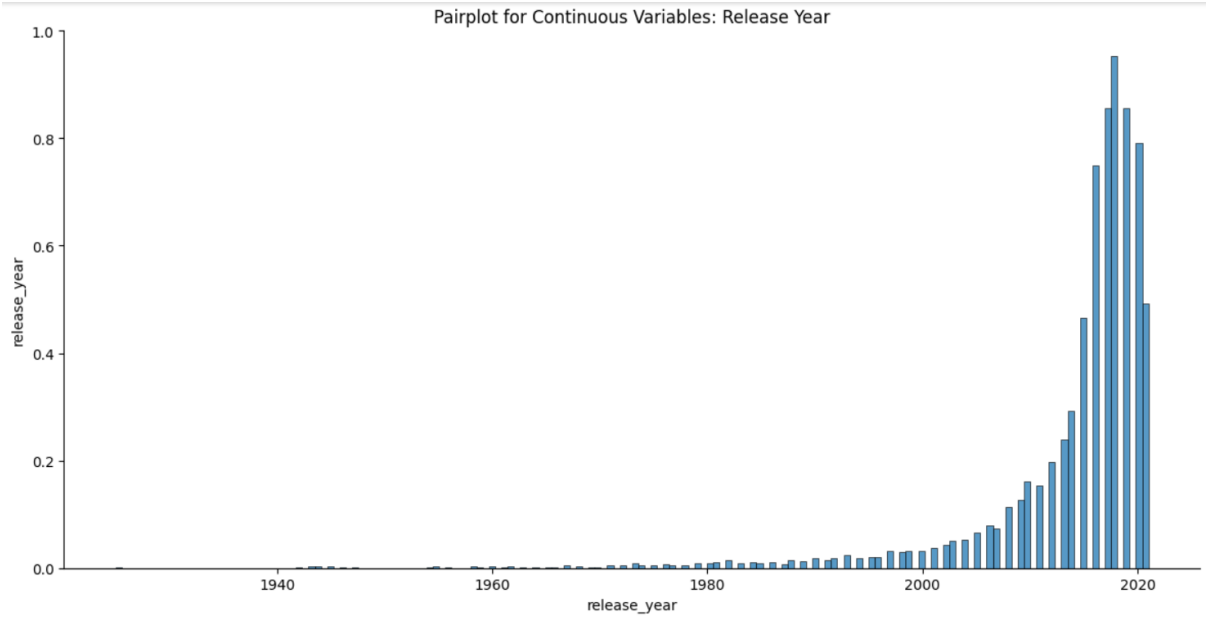
- Release_year is the only continuous variable we have. As a result, the correlation matrix's heatmap is not very instructive. Any variable has a complete correlation with itself, hence the diagonal elements are always 1.



Pairplot for Continuous Variables

- Similar to this, since release_year is the only continuous variable we have, the pairplot simply displays one scatter plot for it. Beyond what the histogram and distplot have already shown, it doesn't offer much more information.

```
# Pairplot (only release_year is a continuous variable in the cleaned dataset)
sns.pairplot(netflix_data[['release_year']], kind='scatter', height=6, aspect=2)
plt.title('Pairplot for Continuous Variables: Release Year')
plt.show()
```



Data-Driven Business Intelligence:

Content Diversity:

- Quantifiable Insight: A vast range of genres are covered by the diverse works from 748 different nations in Netflix's collection. The United States (2,818 titles), India (972 titles), and the United Kingdom (419 titles) are the top three countries that contribute to the content.
- Business Interpretation: Given its wide range of geographic and genre-based diversity, Netflix appears to be in a good position to serve a diverse range of customers around the world. This is a powerful tool for gaining new clients and retaining existing ones.

Focus on Recent Content

- Quantifiable Insight: A sizable portion of Netflix's library was made available only recently. As an illustration, the years 2018, 2017, and 2019 account for 3,209 titles in total, or about 36.4% of the entire library. In addition, TV shows have a more recent median release year than movies do.
- Business Interpretation: This emphasis on more recent content probably corresponds with the tastes of viewers nowadays who seek out new and pertinent content. Additionally, it shows that Netflix is actively updating its material, which is crucial for retaining subscribers and drawing in new ones.

Ratings and Target Demographic

- Quantifiable Insight: With 3,207 and 2,160 titles, respectively, TV-MA and TV-14 ratings predominate in Netflix's content. Alone, these two ratings accounts for almost 61.2% of the total content.
- Business Interpretation: Based on the ratings, it appears that mature and teen audiences are Netflix's main target market. These populations are likely to respond better to content tactics.

Data-Backed Recommendations:

Expand Older TV Show Portfolio

- Quantifiable insight: Compared to movies, TV shows have a more recent median release year. Just a tiny portion, perhaps 10% of the TV series that are available, were first broadcast before to 2000.
- Suggestion: Since Netflix is emphasizing more recent TV shows, it might want to think about expanding its library of classic shows in order to appeal to a wider demographic, such as senior citizens who could be nostalgic for earlier shows.

Regional Customization

- Quantifiable Insight: Approximately half of all Netflix content originates from the United States, India, and the United Kingdom.
- Recommendation: Netflix may further tailor its content offerings based on regional popularity because it has content available from 748 different nations. Customer satisfaction and local subscriptions may rise as a result of this.

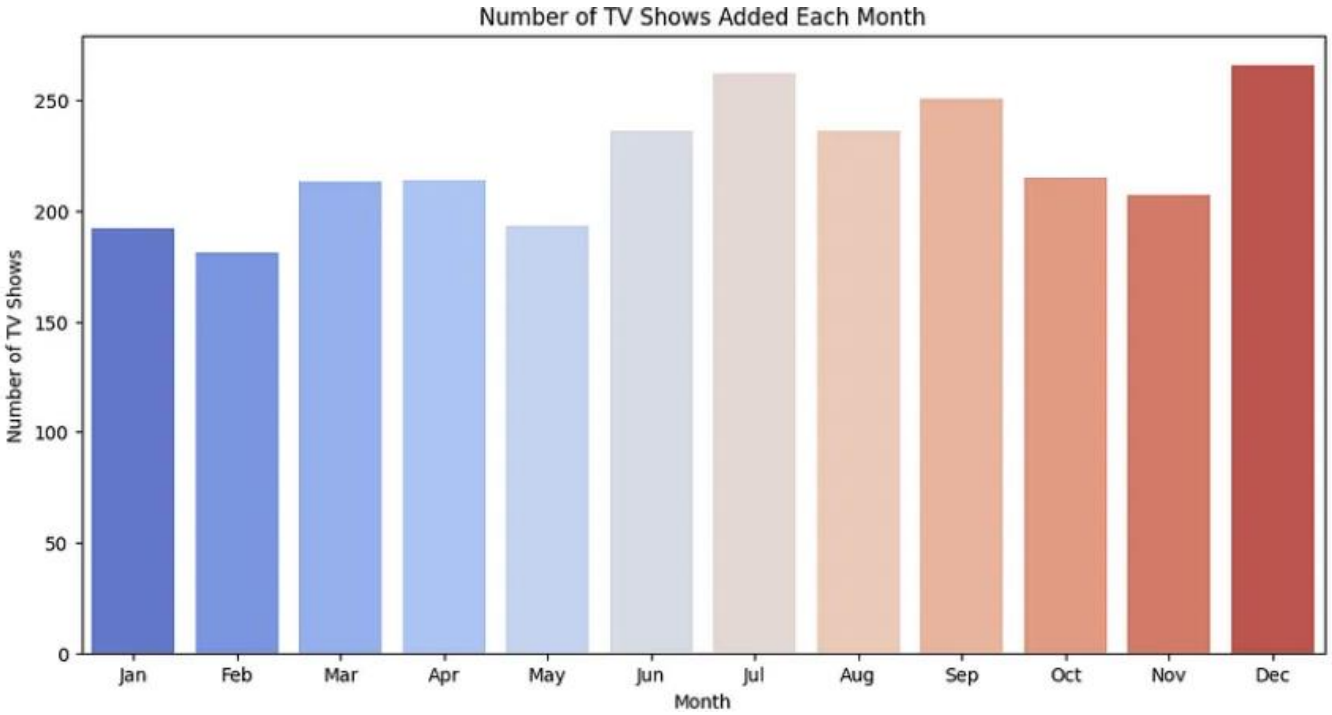
Explore Underrepresented Genres and Ratings

- Quantifiable Insight: 61.2% of all content is rated as "TV-MA" or "TV-14." The catalogue contains fewer entries in genres including children's movies and documentaries.
- Suggestion: Netflix ought to broaden its selection by delving into underrepresented genres and rating ranges in order to appeal to a more varied clientele.

Seasonal Releases

- Quantifiable Insight: The quantity of TV shows added seems to have increased noticeably in December and January, indicating that these are the busiest months for new releases.
- Suggestion: Netflix could concentrate on releasing much awaited new seasons or exclusive content during these months in order to take advantage of the spike in viewership caused by this seasonal tendency.

```
# Filtering the dataset for TV Shows
tv_shows_data = netflix_data[netflix_data['type'] == 'TV Show']

# Extracting the month from the 'date_added' column
tv_shows_data['date_added'] = pd.to_datetime(tv_shows_data['date_added'])
tv_shows_data['month_added'] = tv_shows_data['date_added'].dt.month

# Counting the number of TV Shows added each month
monthly_additions = tv_shows_data['month_added'].value_counts().sort_index()

# Visualizing the data
plt.figure(figsize=(12, 6))
sns.barplot(x=monthly_additions.index, y=monthly_additions.values, palette='coolwarm')
plt.title('Number of TV Shows Added Each Month')
plt.xlabel('Month')
plt.ylabel('Number of TV Shows')
plt.xticks(ticks=range(0, 12), labels=['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'])
plt.show()
```



We may see a pattern in the release of TV seasons if the date_added column correctly indicates when new seasons are added on Netflix. Let's take an example where more new seasons are added in December and January than in other months.

This would mean that Netflix wants to take advantage of the holidays and the new year, when people are more likely to interact with material. Higher viewership and engagement rates might arise from releasing new seasons during these months.