BUSINESS CASE: NETFLIX

Defining Problem Statement and Analysing basic metrics

Problem Statement:

- Netflix is a global streaming platform that offers a vast library of movies and TV shows.
- We can analyse Netflix content to gain insights into trends, preferences, and patterns.
- By analysing the data, we can determine the proportion of movies and TV shows in the dataset, and how content releases have changed over the years.
- We can gain insights into user preferences, content production strategies, and potential gaps in the Netflix library.

Basic Metrics:

Metrics will help us to gain a deeper understanding of the content and its characteristics.

- Types (Movies vs. TV Shows)
- Release Trends Over the Years
- Duration Distribution
- Production Countries
- Top Genres
- Rating (Movies vs. TV Shows)

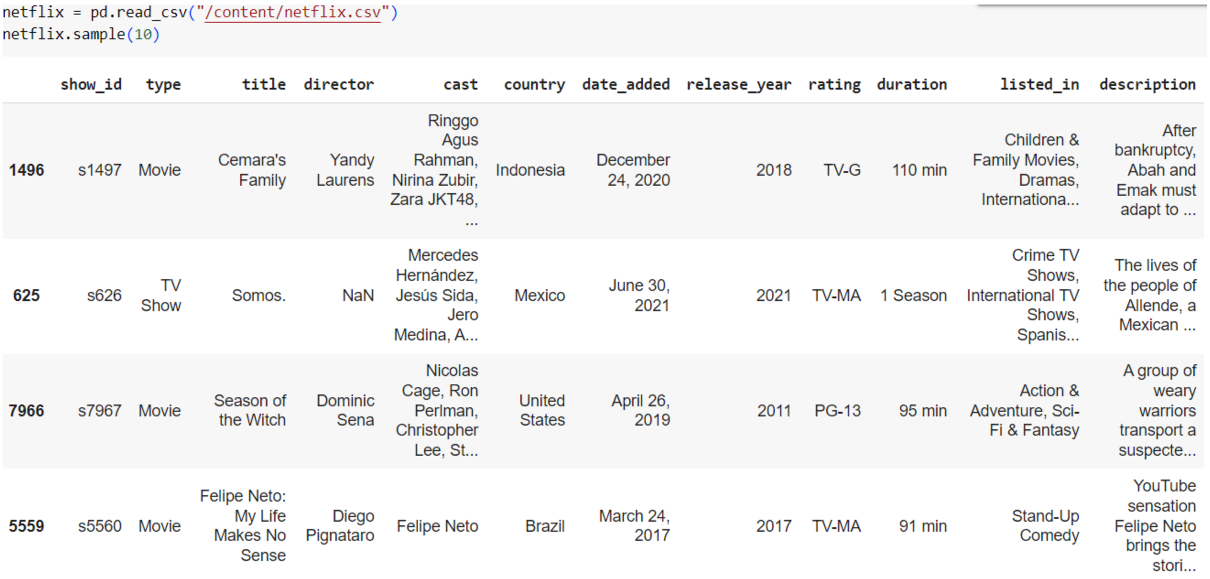
Observations on the shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), missing value detection, statistical summary

The dataset contains information about movies and TV shows available on Netflix.

Key attributes:

- show_id
- type
- title
- director
- cast
- country

- date_added
- release_year
- rating
- duration
- listed_in
- description

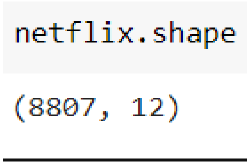


```
netflix = pd.read_csv("/content/netflix.csv")
netflix.sample(10)
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
1496	s1497	Movie	Cemara's Family	Yandy Laurens	Ringgo Agus Rahman, Nirina Zubir, Zara JKT48, ...	Indonesia	December 24, 2020	2018	TV-G	110 min	Children & Family Movies, Dramas, Internationa...	After bankruptcy, Abah and Emak must adapt to ...
625	s626	TV Show	Somos.	NaN	Mercedes Hernández, Jesús Sida, Jero Medina, A...	Mexico	June 30, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, Spanis...	The lives of the people of Allende, a Mexican ...
7966	s7967	Movie	Season of the Witch	Dominic Sena	Nicolas Cage, Ron Perlman, Christopher Lee, St...	United States	April 26, 2019	2011	PG-13	95 min	Action & Adventure, Sci-Fi & Fantasy	A group of weary warriors transport a suspecte...
5559	s5560	Movie	Felipe Neto: My Life Makes No Sense	Diego Pignataro	Felipe Neto	Brazil	March 24, 2017	2017	TV-MA	91 min	Stand-Up Comedy	YouTube sensation Felipe Neto brings the stori...

Shape of Data:

The dataset has a total of 8807 rows (entries) and 12 columns (attributes).

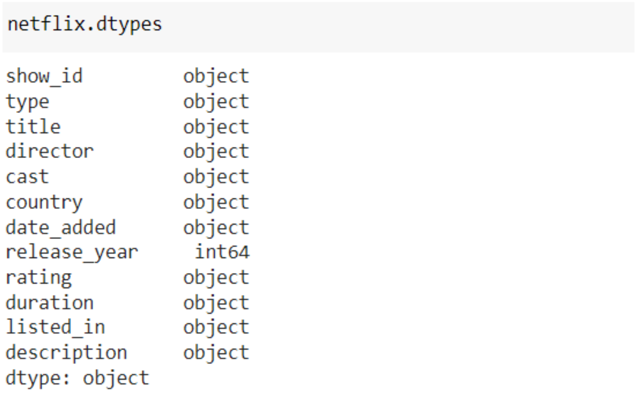


```
netflix.shape
```

```
(8807, 12)
```

Datatypes:

Datatype is object type for all attributes except release_year.



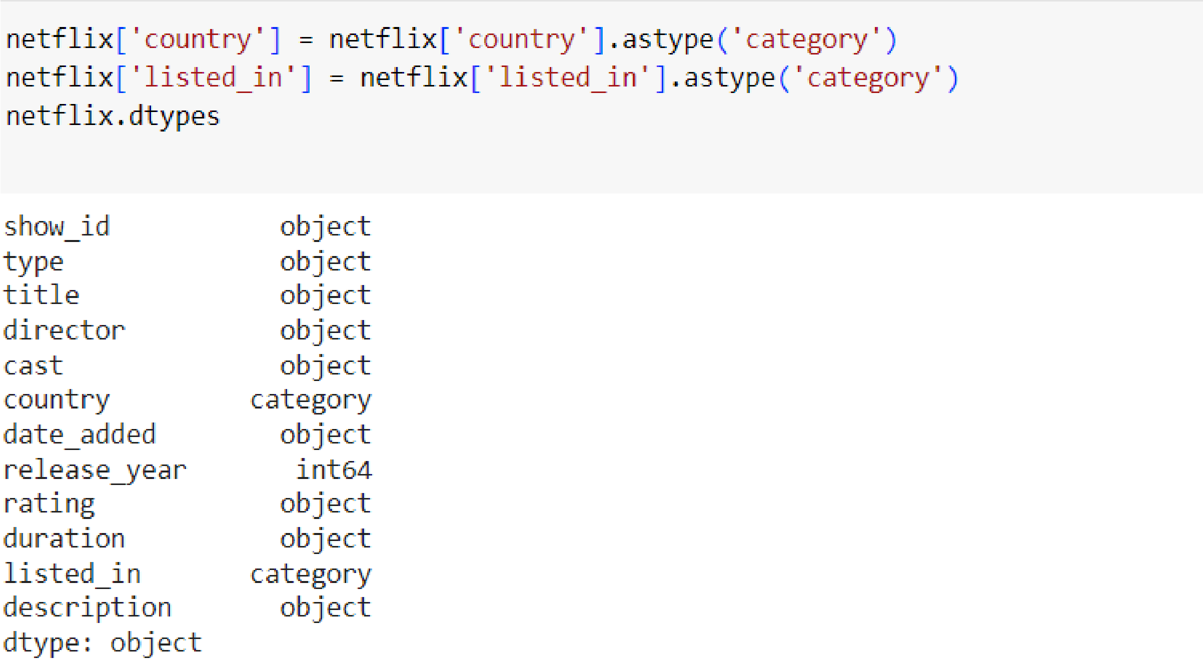
```
netflix.dtypes
```

```
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
```

Conversion of categorical attributes to 'category':

The main purpose of converting categorical attributes to the category data type is to optimize memory usage and improve performance during data analysis.

For example, in Netflix data we can convert Country, Listed_in (Genre) attributes to category.

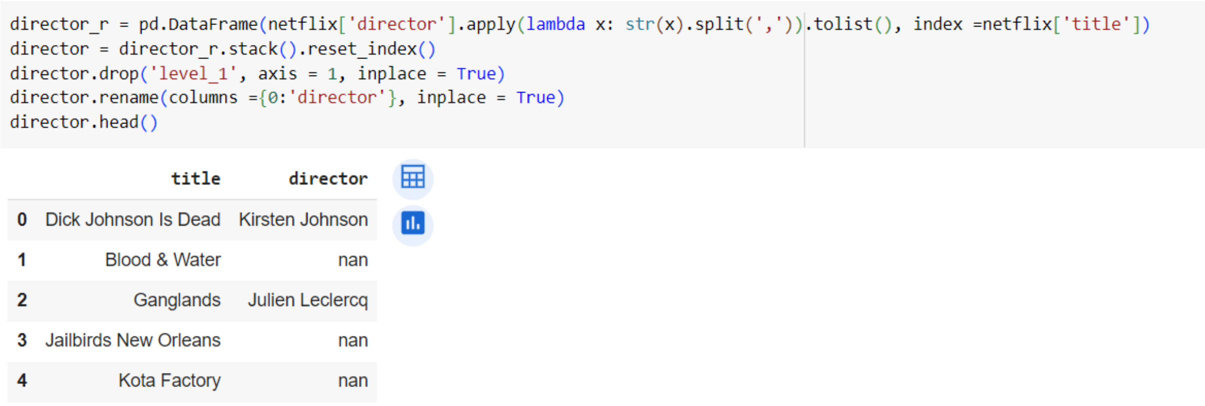


```
netflix['country'] = netflix['country'].astype('category')
netflix['listed_in'] = netflix['listed_in'].astype('category')
netflix.dtypes
```

```
show_id           object
type              object
title             object
director          object
cast              object
country         category
date_added        object
release_year       int64
rating            object
duration          object
listed_in       category
description       object
dtype: object
```

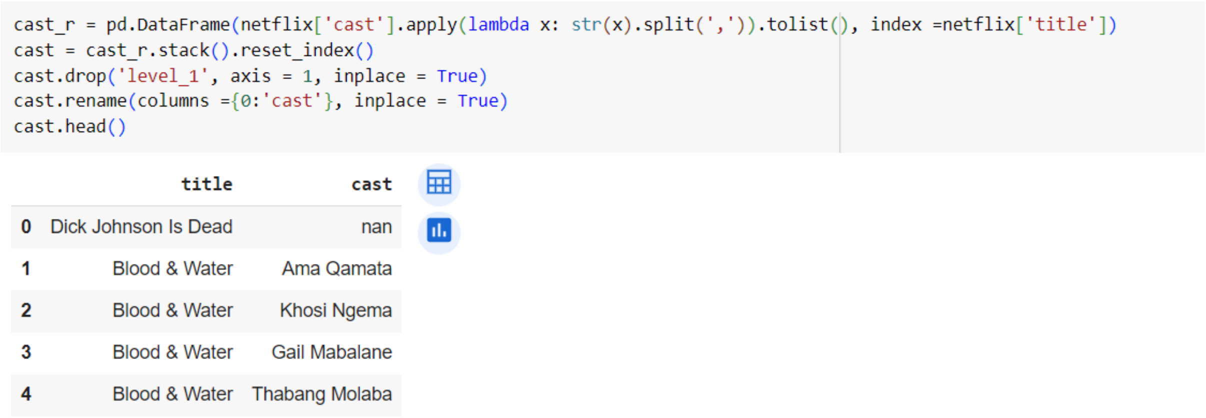
Transforming a single column containing lists (or nested lists) into multiple columns:

In Netflix data frame attributes director, cast, country and listed_in contains data in list.



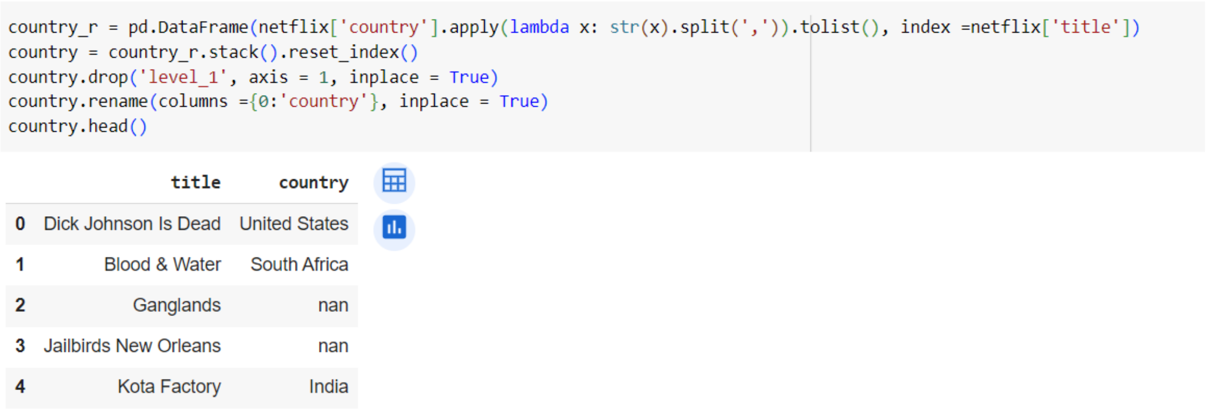
```
director_r = pd.DataFrame(netflix['director'].apply(lambda x: str(x).split(',')).tolist(), index =netflix['title'])
director = director_r.stack().reset_index()
director.drop('level_1', axis = 1, inplace = True)
director.rename(columns ={0:'director'}, inplace = True)
director.head()
```

	title	director
0	Dick Johnson Is Dead	Kirsten Johnson
1	Blood & Water	nan
2	Ganglands	Julien Leclercq
3	Jailbirds New Orleans	nan
4	Kota Factory	nan



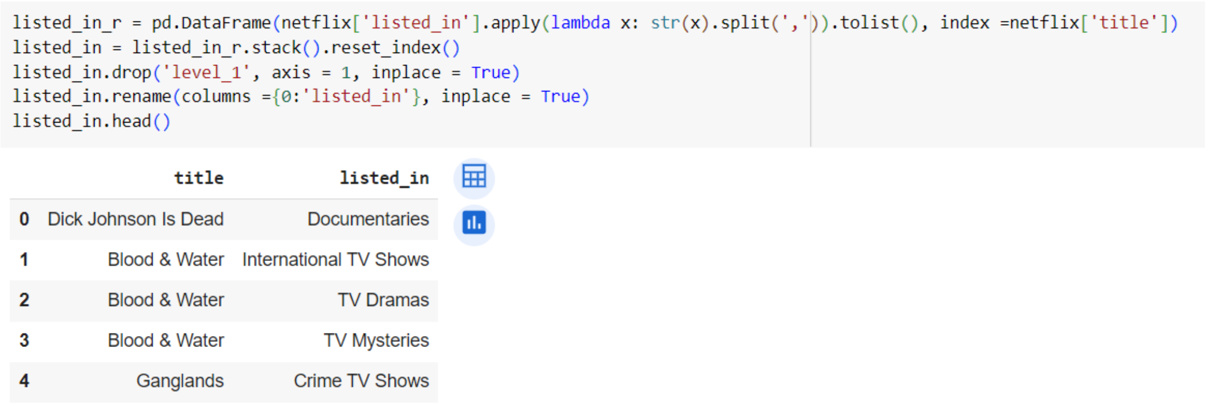
```
cast_r = pd.DataFrame(netflix['cast'].apply(lambda x: str(x).split(',')).tolist(), index =netflix['title'])
cast = cast_r.stack().reset_index()
cast.drop('level_1', axis = 1, inplace = True)
cast.rename(columns ={0:'cast'}, inplace = True)
cast.head()
```

	title	cast
0	Dick Johnson Is Dead	nan
1	Blood & Water	Ama Qamata
2	Blood & Water	Khosi Ngema
3	Blood & Water	Gail Mabalane
4	Blood & Water	Thabang Molaba



```
country_r = pd.DataFrame(netflix['country'].apply(lambda x: str(x).split(',')).tolist(), index =netflix['title'])
country = country_r.stack().reset_index()
country.drop('level_1', axis = 1, inplace = True)
country.rename(columns ={0:'country'}, inplace = True)
country.head()
```

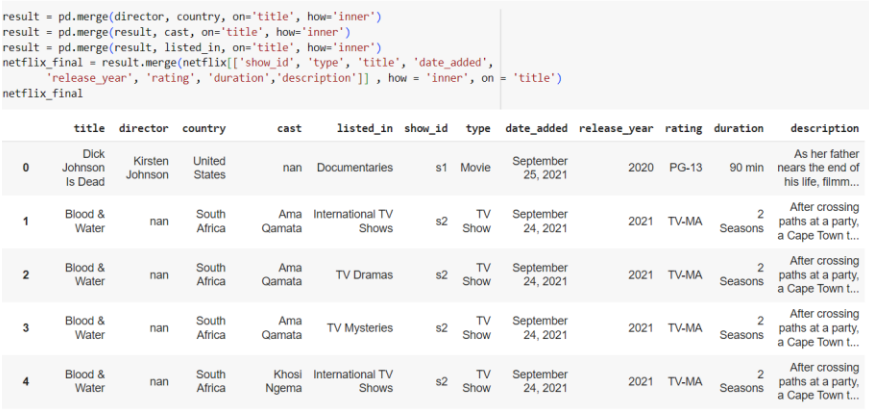
	title	country
0	Dick Johnson Is Dead	United States
1	Blood & Water	South Africa
2	Ganglands	nan
3	Jailbirds New Orleans	nan
4	Kota Factory	India



```
listed_in_r = pd.DataFrame(netflix['listed_in'].apply(lambda x: str(x).split(',')).tolist(), index =netflix['title'])
listed_in = listed_in_r.stack().reset_index()
listed_in.drop('level_1', axis = 1, inplace = True)
listed_in.rename(columns ={0:'listed_in'}, inplace = True)
listed_in.head()
```

	title	listed_in
0	Dick Johnson Is Dead	Documentaries
1	Blood & Water	International TV Shows
2	Blood & Water	TV Dramas
3	Blood & Water	TV Mysteries
4	Ganglands	Crime TV Shows

Merge all attributes and form a new dataset:

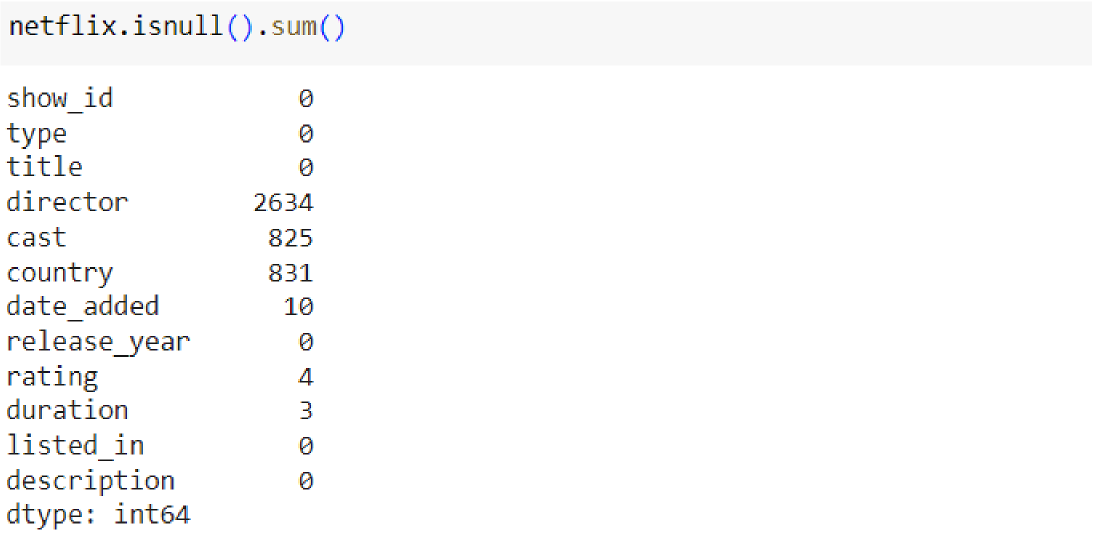


```
result = pd.merge(director, country, on='title', how='inner')
result = pd.merge(result, cast, on='title', how='inner')
result = pd.merge(result, listed_in, on='title', how='inner')
netflix_final = result.merge(netflix[['show_id', 'type', 'title', 'date_added',
      'release_year', 'rating', 'duration','description']] , how = 'inner', on = 'title')
netflix_final
```

	title	director	country	cast	listed_in	show_id	type	date_added	release_year	rating	duration	description
0	Dick Johnson Is Dead	Kirsten Johnson	United States	nan	Documentaries	s1	Movie	September 25, 2021	2020	PG-13	90 min	As her father nears the end of his life, filmm...
1	Blood & Water	nan	South Africa	Ama Qamata	International TV Shows	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...
2	Blood & Water	nan	South Africa	Ama Qamata	TV Dramas	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...
3	Blood & Water	nan	South Africa	Ama Qamata	TV Mysteries	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...
4	Blood & Water	nan	South Africa	Khosi Ngema	International TV Shows	s2	TV Show	September 24, 2021	2021	TV-MA	2 Seasons	After crossing paths at a party, a Cape Town t...

Missing value detection:

Checking for null values:



```
netflix.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
```

The following columns have null values that need to be cleaned:

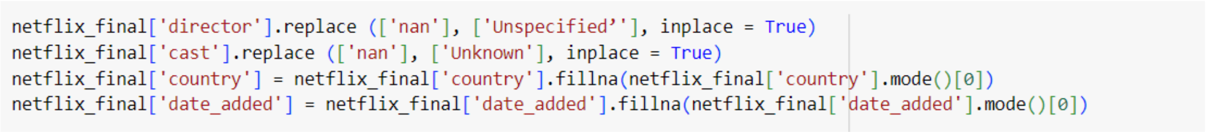
- director: 2634
- cast: 825
- country: 831
- date_added: 10
- rating: 4
- duration: 3

director: Filled missing values with 'Unspecified'

cast: Filled missing values with 'Unknown'

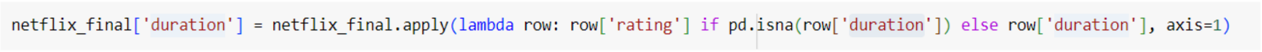
country: Filled missing values with the mode (most frequent value) of the column.

date_added: Filled missing values with the of the column.



```
netflix_final['director'].replace (['nan'], ['Unspecified’'], inplace = True)
netflix_final['cast'].replace (['nan'], ['Unknown'], inplace = True)
netflix_final['country'] = netflix_final['country'].fillna(netflix_final['country'].mode()[0])
netflix_final['date_added'] = netflix_final['date_added'].fillna(netflix_final['date_added'].mode()[0])
```

duration: Filled missing values with the corresponding row rating column.



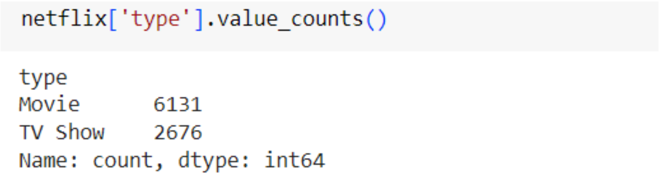
```
netflix_final['duration'] = netflix_final.apply(lambda row: row['rating'] if pd.isna(row['duration']) else row['duration'], axis=1)
```

Non-Graphical Analysis: Value counts and unique attributes

Value Counts:

Type: The dataset contains both movies and TV shows. value counts for each type:

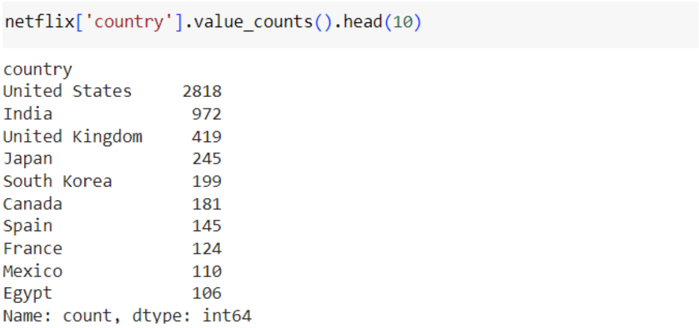
- Movies: 6131
- TV Shows: 2676



```
netflix['type'].value_counts()
```

```
type
Movie      6131
TV Show    2676
Name: count, dtype: int64
```

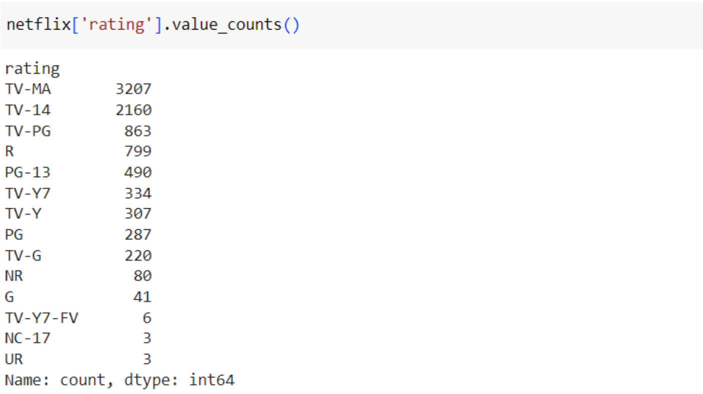
Country: Value counts for each top 10 countries:



```
netflix['country'].value_counts().head(10)
```

```
country
United States     2818
India              972
United Kingdom     419
Japan              245
South Korea        199
Canada             181
Spain              145
France             124
Mexico             110
Egypt              106
Name: count, dtype: int64
```

Rating: Value counts for each rating:



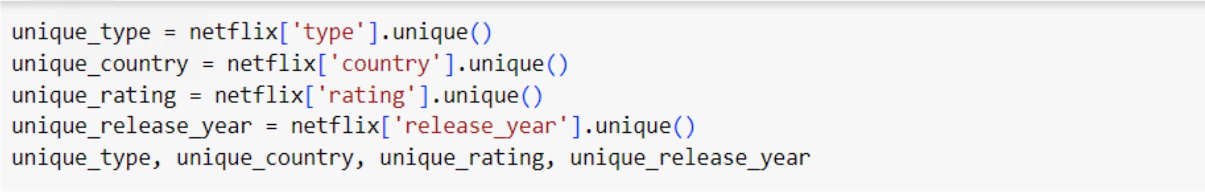
```
netflix['rating'].value_counts()
```

```
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
NC-17          3
UR             3
Name: count, dtype: int64
```

Netflix can make informed decisions based on value counts and unique attributes like:

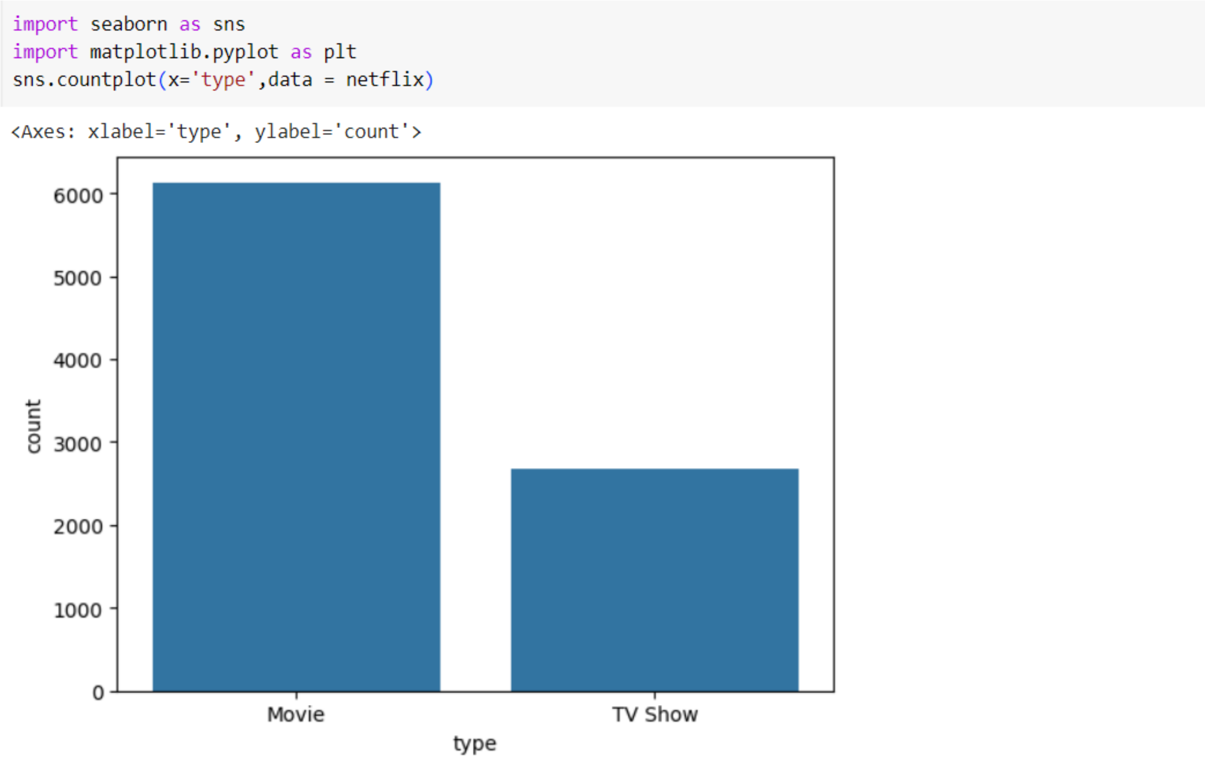
- Consider focusing on producing more content in top countries.
- Pay attention to viewer preferences based on ratings (e.g., TV-MA, TV-14).

Unique Attributes:



```
unique_type = netflix['type'].unique()
unique_country = netflix['country'].unique()
unique_rating = netflix['rating'].unique()
unique_release_year = netflix['release_year'].unique()
unique_type, unique_country, unique_rating, unique_release_year
```

Visual Analysis - Univariate, Bivariate after pre-processing of the data



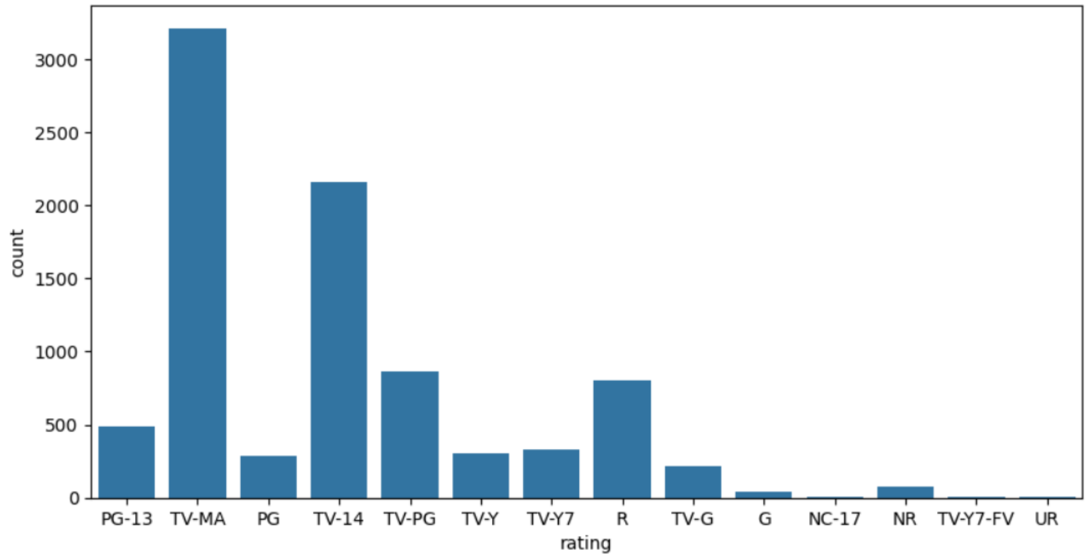
```
import seaborn as sns
import matplotlib.pyplot as plt
sns.countplot(x='type',data = netflix)
```

```
<Axes: xlabel='type', ylabel='count'>
```

The plot shows that there are more movies than TV shows in the dataset.

```
plt.figure(figsize = (10,5))
sns.countplot(x='rating',data = netflix)
```

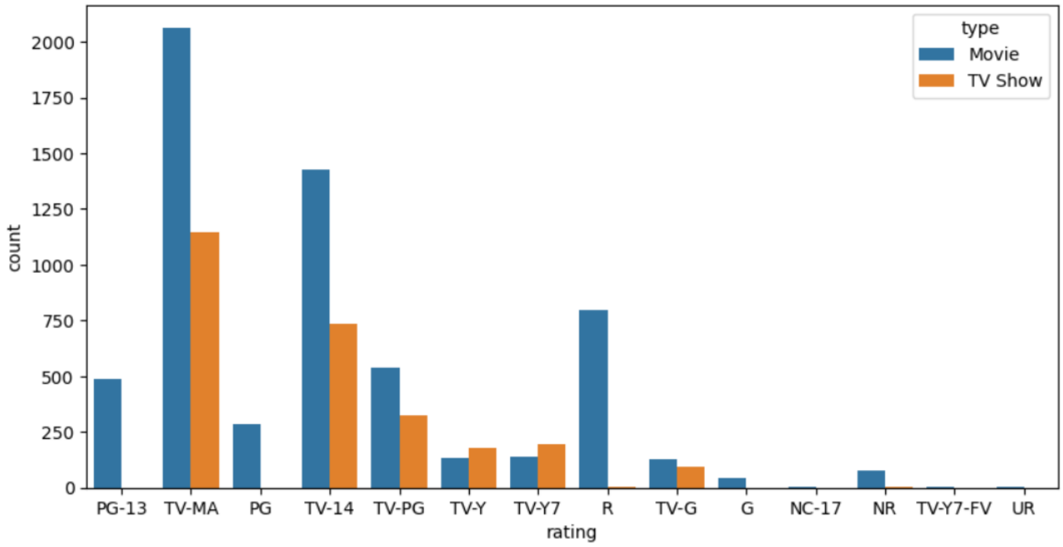
```
<Axes: xlabel='rating', ylabel='count'>
```



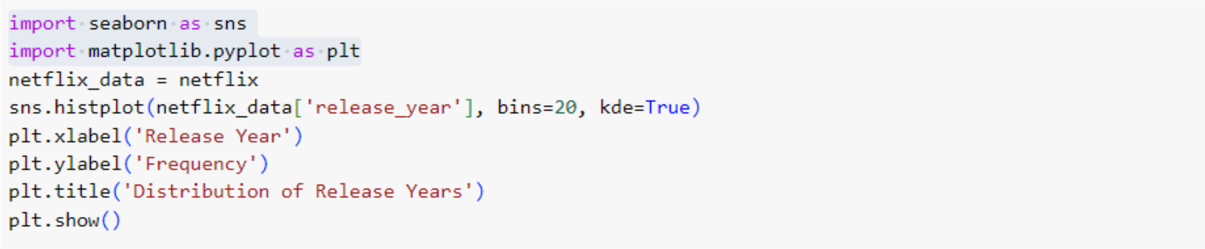
The plot shows that the most frequent rating in the dataset is TV-MA, followed by TV-14 and TV-PG.

```
plt.figure(figsize = (10,5))
sns.countplot(x='rating',data = netflix,hue='type')
```

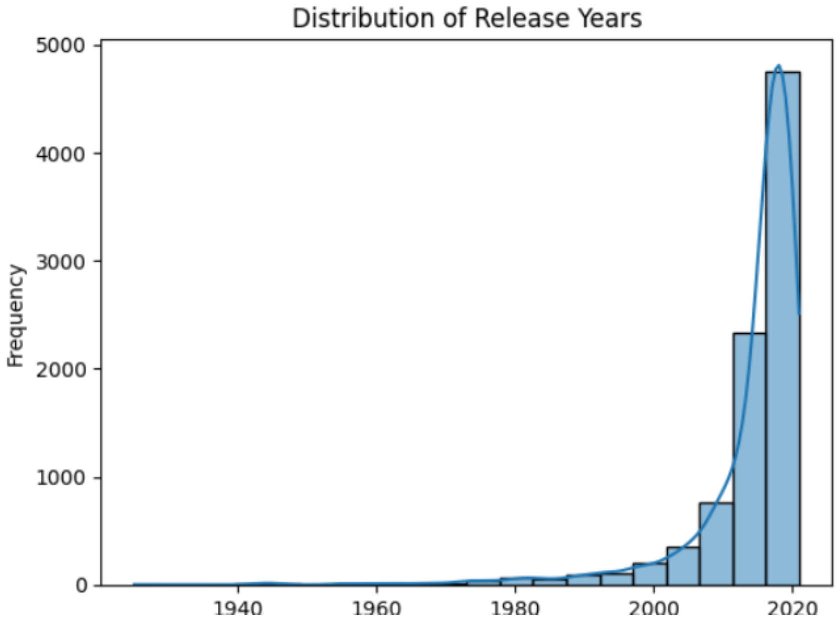
```
<Axes: xlabel='rating', ylabel='count'>
```



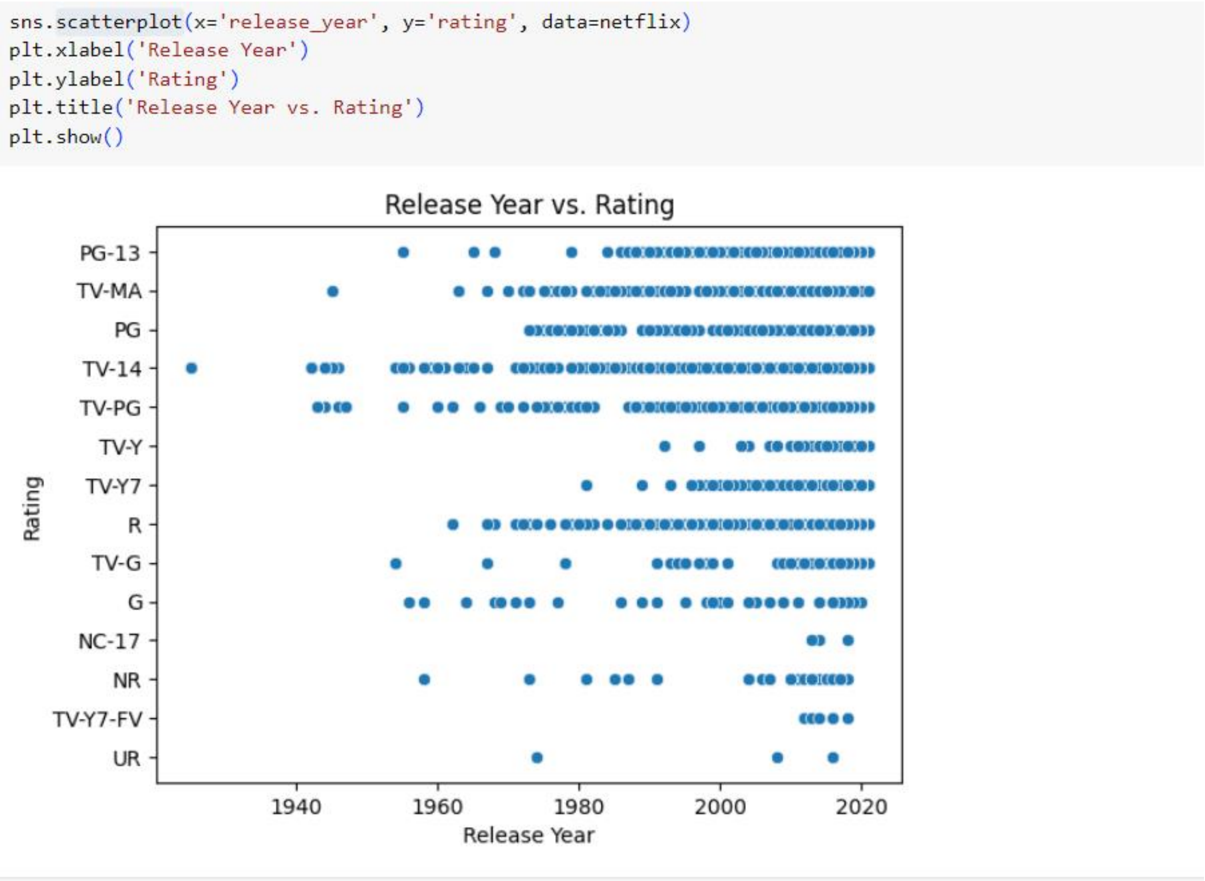
Rating TV-MA is more in movies than in TV, while the rating PG-13 is more common in movies than in TV shows.



```
import seaborn as sns
import matplotlib.pyplot as plt
netflix_data = netflix
sns.histplot(netflix_data['release_year'], bins=20, kde=True)
plt.xlabel('Release Year')
plt.ylabel('Frequency')
plt.title('Distribution of Release Years')
plt.show()
```



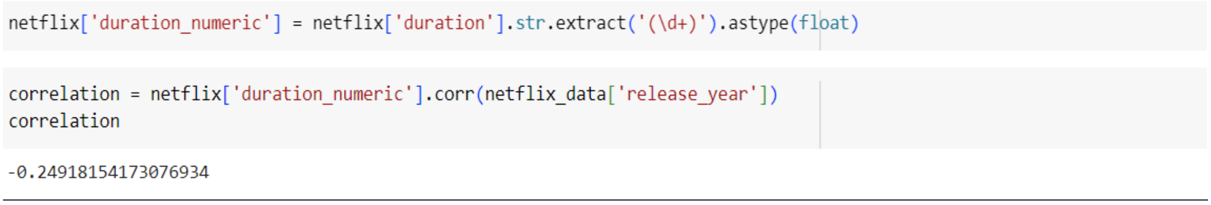
This plot shows the frequency of the distribution of release years available on Netflix.



```
sns.scatterplot(x='release_year', y='rating', data=netflix)
plt.xlabel('Release Year')
plt.ylabel('Rating')
plt.title('Release Year vs. Rating')
plt.show()
```

This plot shows the relationship between release year and rating of movies and tv shows available on Netflix.

Conversion of duration into numeric value:

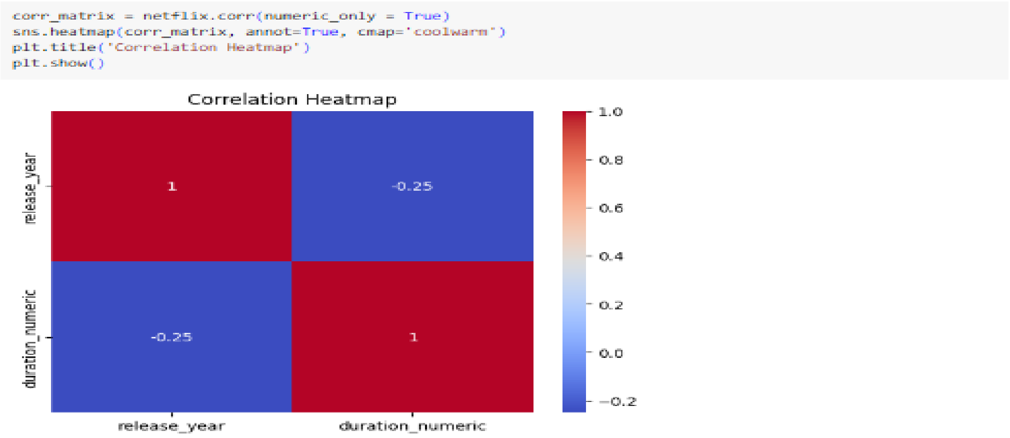


```
netflix['duration_numeric'] = netflix['duration'].str.extract('(\d+)').astype(float)
```

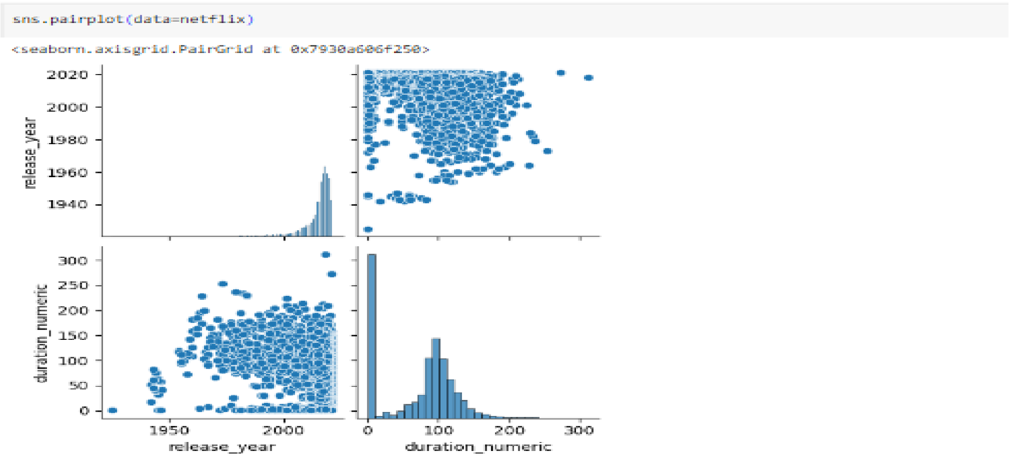
```
correlation = netflix['duration_numeric'].corr(netflix_data['release_year'])
correlation
```

```
-0.24918154173076934
```

Plots for correlation of the numeric values:



```
corr_matrix = netflix.corr(numeric_only = True)
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



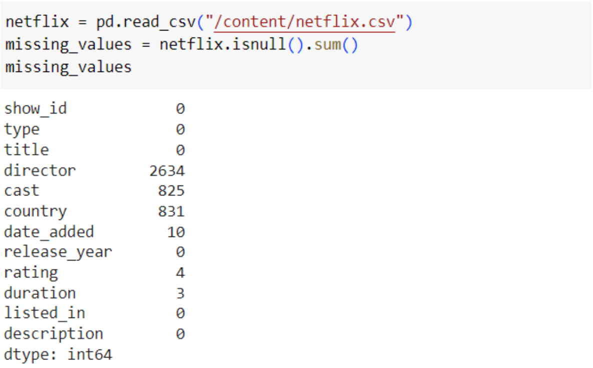
```
sns.pairplot(data=netflix)
```

```
<seaborn.axisgrid.PairGrid at 0x7930a606f250>
```

Missing Value & Outlier check

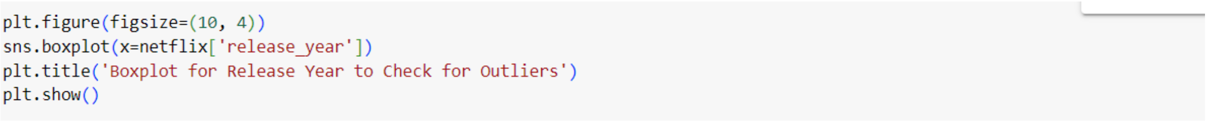
We have missing values in several columns:

- 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

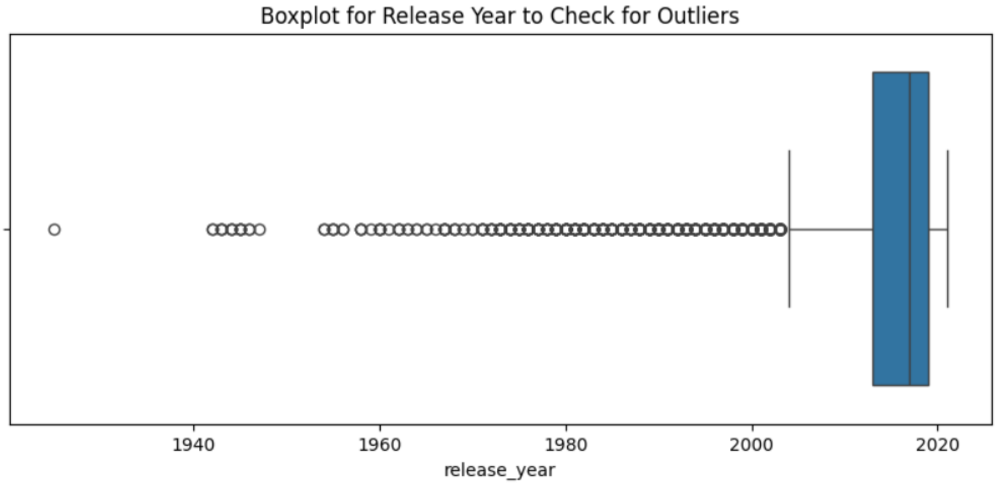


```
netflix = pd.read_csv("/content/netflix.csv")
missing_values = netflix.isnull().sum()
missing_values
```

```
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
```



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



The boxplot for release_year shows no significant outliers, indicating that the data for this attribute is relatively consistent.

Insights based on Non-Graphical and Visual Analysis

- The platform on the whole offer's movies, almost twice as many as TV Shows.
- Most of the content was released in the years 2018, 2017, and 2019, showing a strong focus on recent content.
- The United States is the leading country in producing content, followed by India and the United Kingdom.
- The distribution of release years is right-skewed, indicating that most of the content on Netflix is relatively new, with a significant amount released in the last decade.
- Both Movies and TV Shows predominantly fall under the "TV-MA" and "TV-14" ratings.
- The distribution of ratings between Movies and TV Shows is somewhat similar, though Movies have a higher count in most rating categories.

Business Insights

- Netflix's data is wide-ranging with productions from 749 unique countries and covers a wide array of genres. The top three countries contributing to the content are the United States, India, and the United Kingdom.

- Ratings 'TV-MA' and 'TV-14' dominate the content on Netflix, with 3,207 and 2,160 titles respectively. These two ratings alone make up around 61.2% of all content
- A significant chunk of Netflix's content has been released in recent years. For instance, the years 2018, 2017, and 2019 collectively account for 3,209 titles, making up approximately 36.4% of the total catalogue.

Recommendations

- Given this seasonal trend, Netflix could focus on releasing highly anticipated new seasons or exclusive content during these months to capitalize on increased viewership.
- With content available from 749 different countries, Netflix has the opportunity to further customize its offerings based on regional popularity. This could lead to an increase in local subscriptions and customer satisfaction.