Business Case: Netflix - Data Exploration and Visualisation

About NETFLIX

Netflix is one of the most popular media and video streaming platforms. They have over 10000 movies or tv shows available on their platform, as of mid-2021, they have over 222M Subscribers globally. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.

Business Problem

Analyze the data and generate insights that could help Netflix ijn deciding which type of shows/movies to produce and how they can grow the business in different countries

```
!gdown
https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940
/original/netflix.csv -O netflix.csv
```

```
Downloading...
From:
https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940
/original/netflix.csv
To: /content/netflix.csv
   0% 0.00/3.40M [00:00<?, ?B/s]  31% 1.05M/3.40M [00:00<00:00,
10.2MB/s] 100% 3.40M/3.40M [00:00<00:00, 24.8MB/s]
```

```
# Importing necessary libraries
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv('netflix.csv')

# Displaying first few Rows of Data
df.head()
```

```
  show_id     type                  title         director  \
0      s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show          Blood & Water              NaN
2      s3  TV Show              Ganglands  Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans              NaN
4      s5  TV Show           Kota Factory              NaN

                                                cast        country  \
0                                                NaN  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
```

```
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                                NaN            NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India

           date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min  
1  September 24, 2021          2021  TV-MA  2 Seasons  
2  September 24, 2021          2021  TV-MA   1 Season  
3  September 24, 2021          2021  TV-MA   1 Season  
4  September 24, 2021          2021  TV-MA  2 Seasons  

                                           listed_in  \
0                                      Documentaries   
1    International TV Shows, TV Dramas, TV Mysteries   
2  Crime TV Shows, International TV Shows, TV Act...   
3                             Docuseries, Reality TV   
4  International TV Shows, Romantic TV Shows, TV ...   

                                         description  
0  As her father nears the end of his life, filmm...  
1  After crossing paths at a party, a Cape Town t...  
2  To protect his family from a powerful drug lor...  
3  Feuds, flirtations and toilet talk go down amo...  
4  In a city of coaching centers known to train I...  
```

```
# Displaying from last few Rows of Data
df.tail()
```

```
     show_id     type        title         director  \
8802   s8803    Movie       Zodiac    David Fincher   
8803   s8804  TV Show  Zombie Dumb              NaN   
8804   s8805    Movie   Zombieland  Ruben Fleischer   
8805   s8806    Movie         Zoom     Peter Hewitt   
8806   s8807    Movie       Zubaan      Mozez Singh   

                                                   cast        country  
\
8802  Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...  United States   

8803                                                NaN            NaN   

8804  Jesse Eisenberg, Woody Harrelson, Emma Stone, ...  United States   

8805  Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...  United States   

8806  Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...          India   


             date_added  release_year rating   duration  \
8802  November 20, 2019          2007      R    158 min   
8803       July 1, 2019          2018  TV-Y7  2 Seasons   
```

```
8804   November 1, 2019          2009      R     88 min
8805   January 11, 2020          2006     PG     88 min
8806      March 2, 2019          2015  TV-14    111 min

                                           listed_in  \
8802                  Cult Movies, Dramas, Thrillers  
8803          Kids' TV, Korean TV Shows, TV Comedies  
8804                         Comedies, Horror Movies  
8805              Children & Family Movies, Comedies  
8806  Dramas, International Movies, Music & Musicals  

                                            description
8802  A political cartoonist, a crime reporter and a...
8803  While living alone in a spooky town, a young g...
8804  Looking to survive in a world taken over by zo...
8805  Dragged from civilian life, a former superhero...
8806  A scrappy but poor boy worms his way into a ty...
```

```
# Shape of Dataset
df.shape
```

```
(8807, 12)
```

```
# Data types of attributes
df.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
```

The Dataset Consist of 8807 Enrties and 12 attributes

The dataset attributes are:

- 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: 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

```
#Describe for numerical columns basic metrics
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
```

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

```
#Describe for Categorical Important columns type, country, rating
basic metrics
df[['type','country','rating']].describe(include=['object'])

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

Categorical Attributes type, country and rating has this basic metrics

- Count
- Unique Values
- Most Frequent
- Frequency

Observation from abover Numerical and Categorical Attributes Basic matrics:

- The content on Netflix is most frequency is Movies compared to TV Shows.
- The Netflix has most of the content in recent decade as per Average Release year is 2014 and median of release year is 2017
- United States is the top most producer of content on Netflix
- Rating TV-MA is most frequent so it suggesting focus on Mature Audience Only

```
#Coverted Category Columns to Category Data type
df[['type','country','rating']]=df[['type','country','rating']].astype
('category')

#Converted date_added to Date Data type
df['date_added'] = pd.to_datetime(df['date_added'])

df.dtypes
```

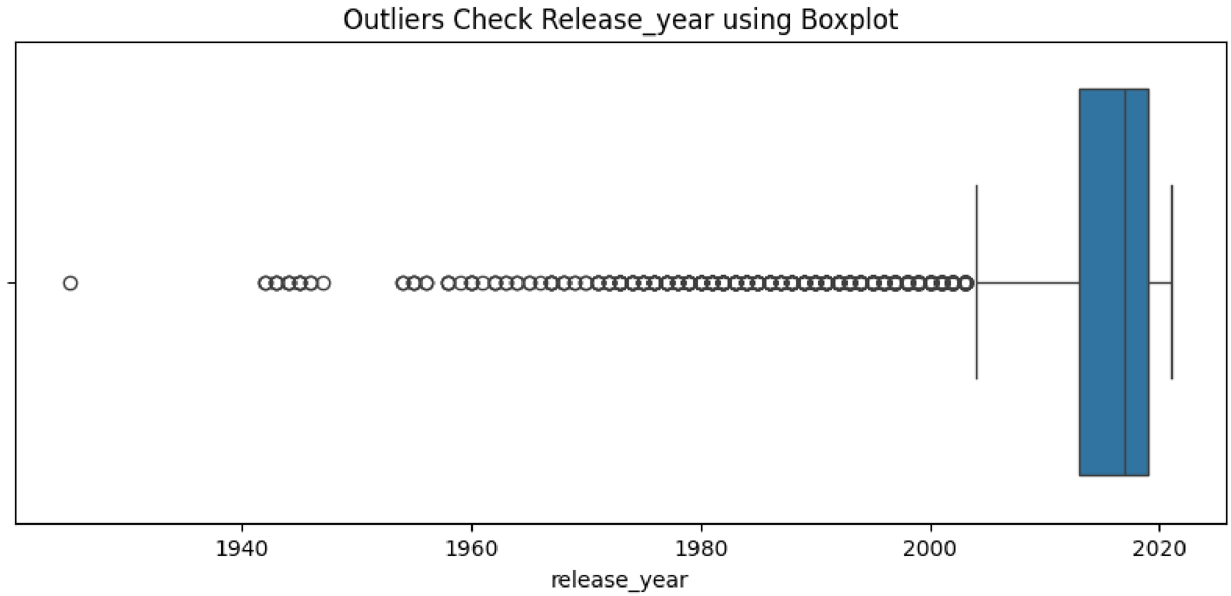
```
show_id                 object
type                  category
title                   object
director                object
cast                    object
country               category
date_added      datetime64[ns]
release_year             int64
rating                category
duration                object
listed_in               object
description             object
dtype: object
```

```
#Missing values Detection
missing_values=df.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
```

For our analysis, these missing values may not impact the outcome. But still we will Replace with. Unkown in Catogerical and Zero in Numerical but in neumerical release year has no null values

```
# Replace missing value
df['director']=df['director'].fillna('Unkown')

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



Above chat Shows no significant outliers, indicating that the data for this attribute is consistent.

```
df['cast']=df['cast'].fillna('Unkown')
```

Non-Graphical Analysis: Value Counts and Unique Attributes

```
#Value Counts
Value_count_type = df['type'].value_counts()
Value_count_country = df['country'].value_counts()
Value_count_rating = df['rating'].value_counts()
Value_count_release_year = df['release_year'].value_counts()

#unique values
Unique_type = df['type'].unique()
Unique_country = df['country'].unique()
Unique_rating = df['rating'].unique()
Unique_release_year = df['release_year'].unique()

Value_count_type,Value_count_country,Value_count_rating,Value_count_re
lease_year,
Unique_type, Unique_country, Unique_rating, Unique_release_year
```

```
(['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]))
```

Value Counts

Type of Content (Movies vs. TV Shows)

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

Top 10 Countries Producing Content

- United States: 2,818
- India: 972
- United Kingdom: 419 ...etc

Ratings

- TV-MA: 3,207
- TV-14: 2,160 ....etc

Top 10 Release Years

- 2018: 1,147
- 2017: 1,032 ....etc

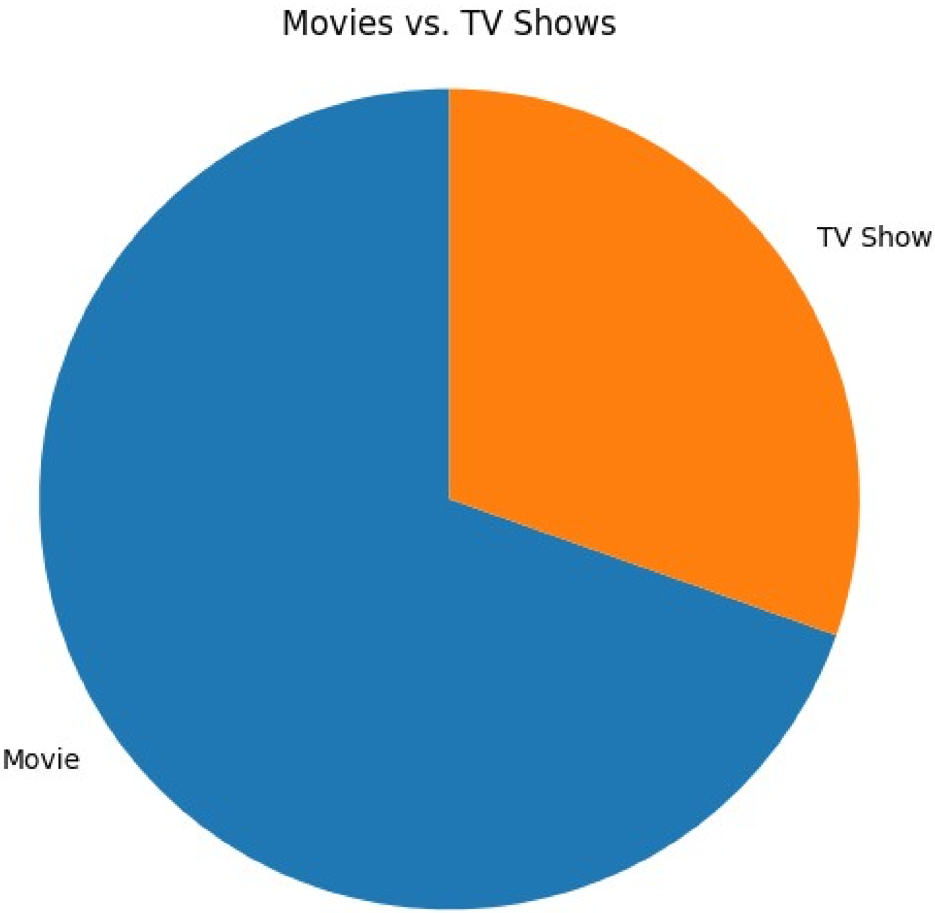
Unique Attributes

- Type: 2 unique values (‘Movie’, ‘TV Show’)
- Country: 748 unique values
- Rating: 17 unique values
- Release Year: Ranges from 1925 to 2021

Observations

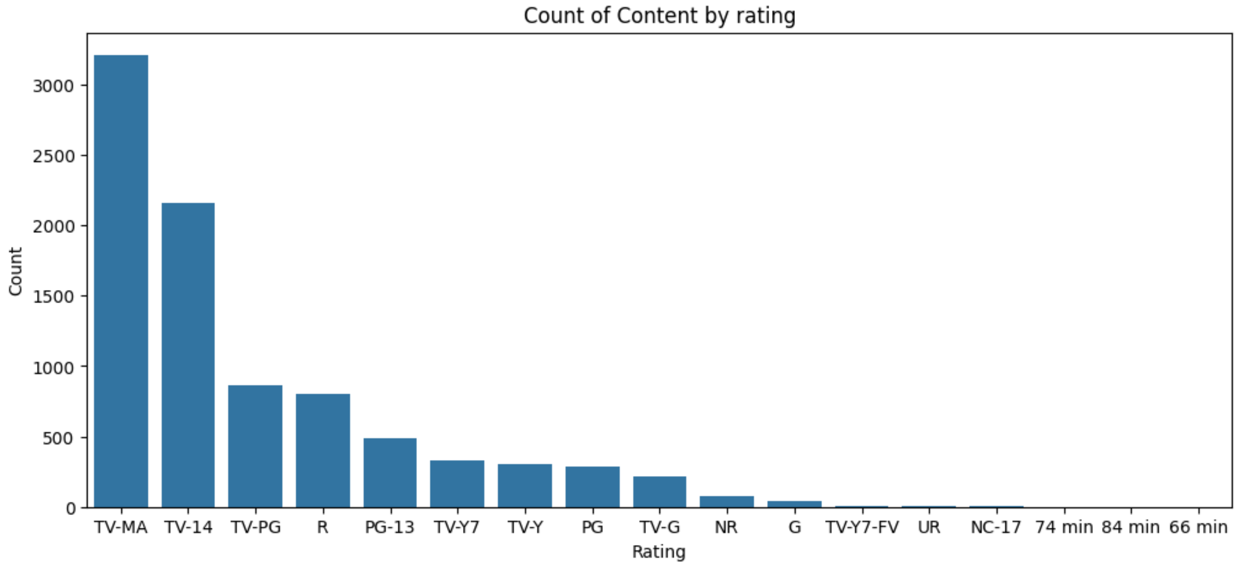
- Most of the content released in 2018,2017,2019 focus on recent content and also most of content rating is TV-MA and TV-14 focus on Mature and teen audience
- United States, India, United Kingdom are Top three producers of Content, where the most of the content is movies which is twice as TV Shows

```
#Univariate Analysis
#Type of Content (Movies vs. TV Shows)
#Movies: 6,131
#TV Shows: 2,676

type_counts = df['type'].value_counts()
labels = type_counts.index
sizes = type_counts.values

plt.figure(figsize=(6, 6))
plt.pie(sizes, labels=labels, startangle=90)
plt.title('Movies vs. TV Shows')
plt.axis('equal')
plt.show()
```



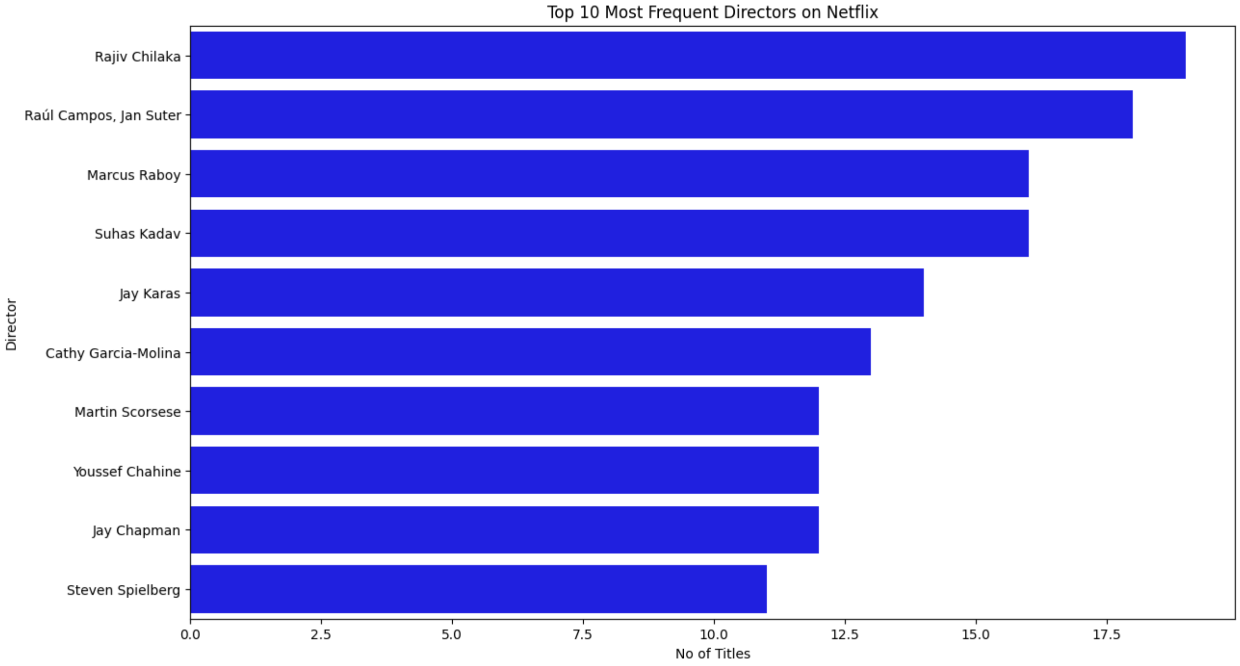
Here from above Pie chart the most of the content is movies which is twice as TV Shows

```python
#Count Plot for Rating
plt.figure(figsize=(12,5))
sns.countplot(x='rating', data=df,
order=df['rating'].value_counts().index)
plt.title("Count of Content by rating")
plt.xlabel("Rating")
plt.ylabel("Count")
plt.show()
```

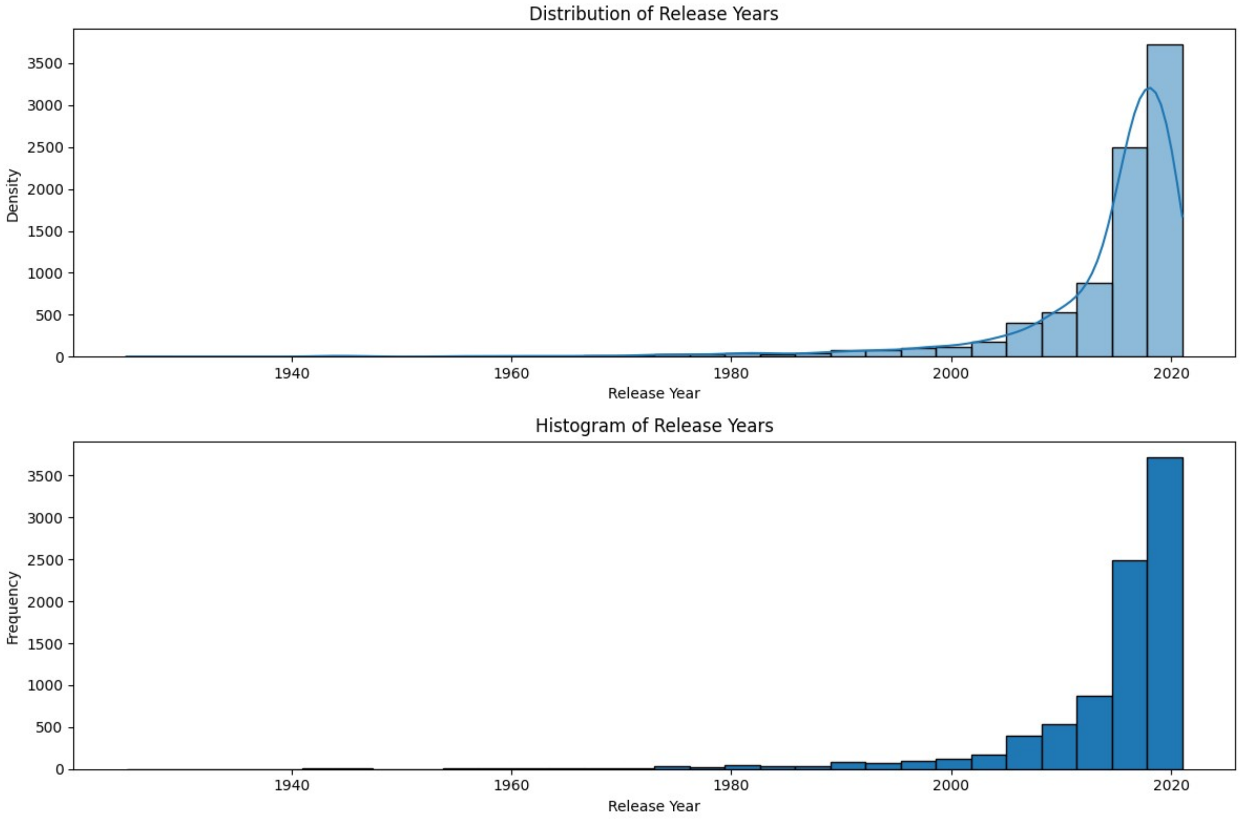


From Above bar chart Most of content rating is TV-MA and TV-14 focus on Mature and teen audience

```
top_directors=df[df['director']!='Unkown']
['director'].value_counts().head(10)

#Top 10 Directors on Netflix
#Below Bar chart Indicates Top 10 Directors on Netflix with Most
Titles
plt.figure(figsize=(14, 8))
sns.barplot(y=top_directors.index, x=top_directors.values, color='b')
plt.title('Top 10 Most Frequent Directors on Netflix')
plt.xlabel('No of Titles')
plt.ylabel('Director')
plt.show()
```

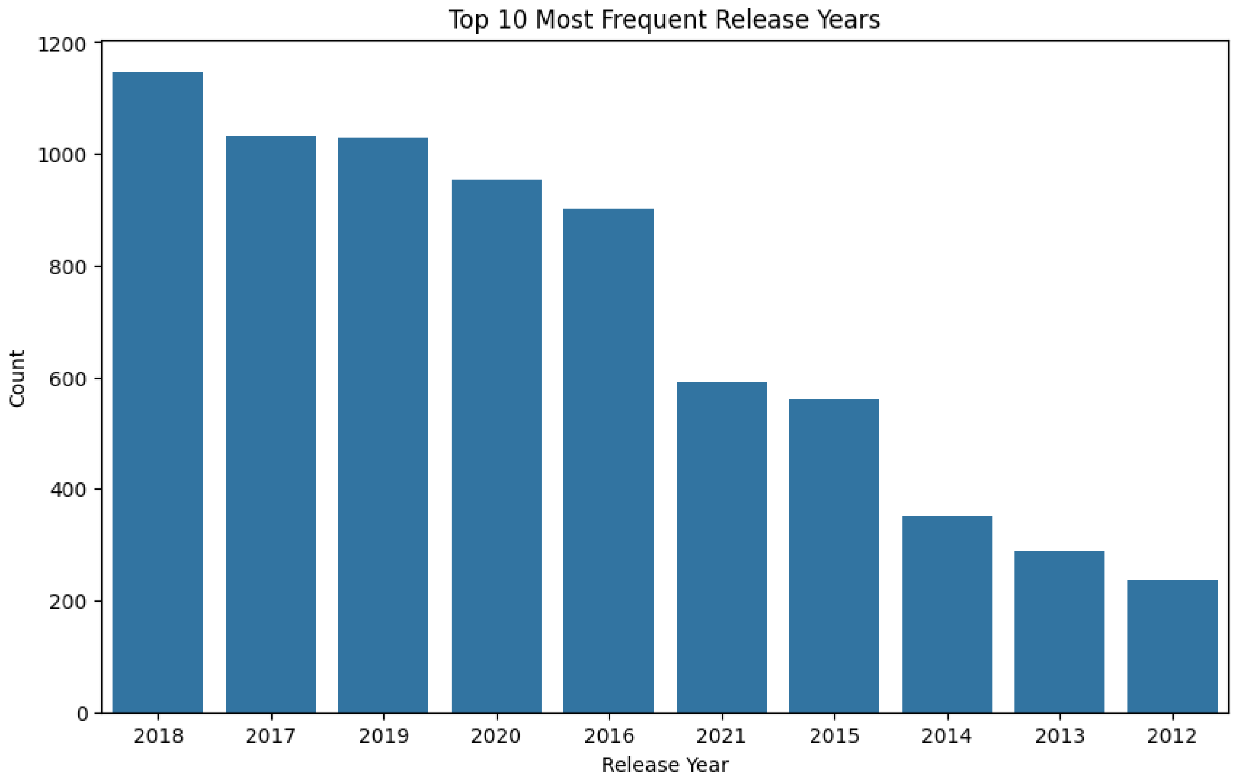


```
fig, axes =plt.subplots(2,1, figsize=(12, 8))
# Distribution plot for release_year

sns.histplot(df['release_year'], kde=True, bins=30, ax=axes[0])
axes[0].set_title('Distribution of Release Years')
axes[0].set_xlabel('Release Year')
axes[0].set_ylabel('Density')

# Histogram for release_year

axes[1].hist(df['release_year'], bins=30, edgecolor='black')
axes[1].set_title('Histogram of Release Years')
axes[1].set_xlabel('Release Year')
axes[1].set_ylabel('Frequency')

plt.tight_layout()
plt.show()
```



From Above chart Most of the content on Netflix is new, with most of the content released in the last decade.

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



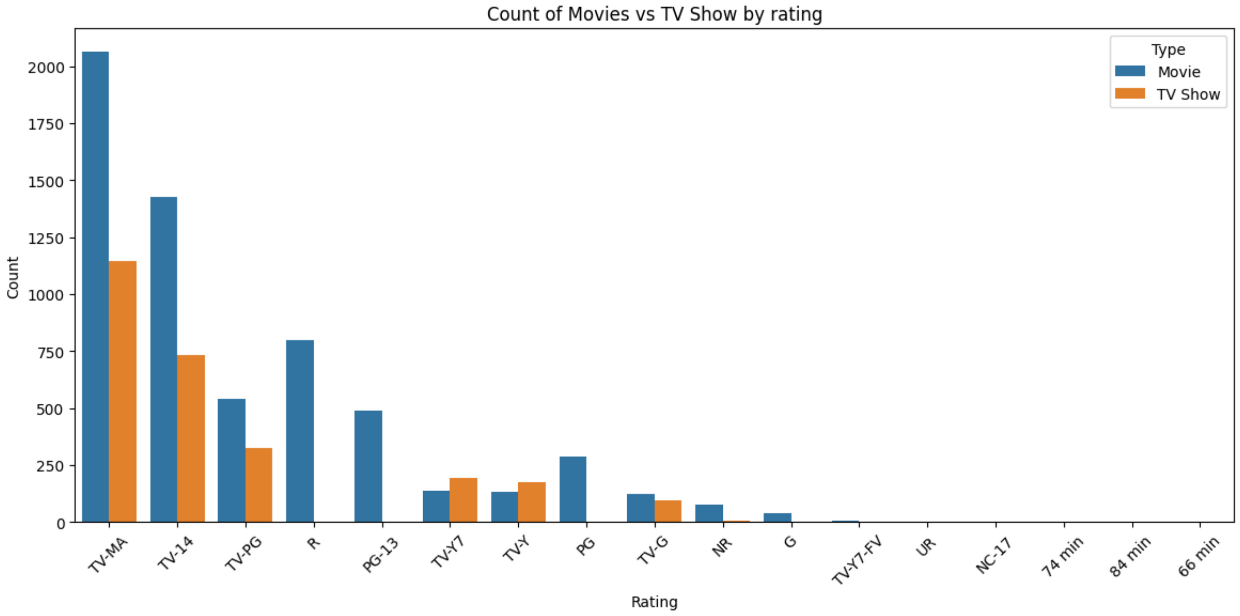
Most of the content released in 2018 and in recent decade the up trend of movies is seen if we observe the chart from right to left mostly is in upwards direction, the production is increasing as time passes

Bivariate Analysis

Relation between Type and Rating

Where in Type Movie Vs TV Show

```
# Countplot for Type vs Rating

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



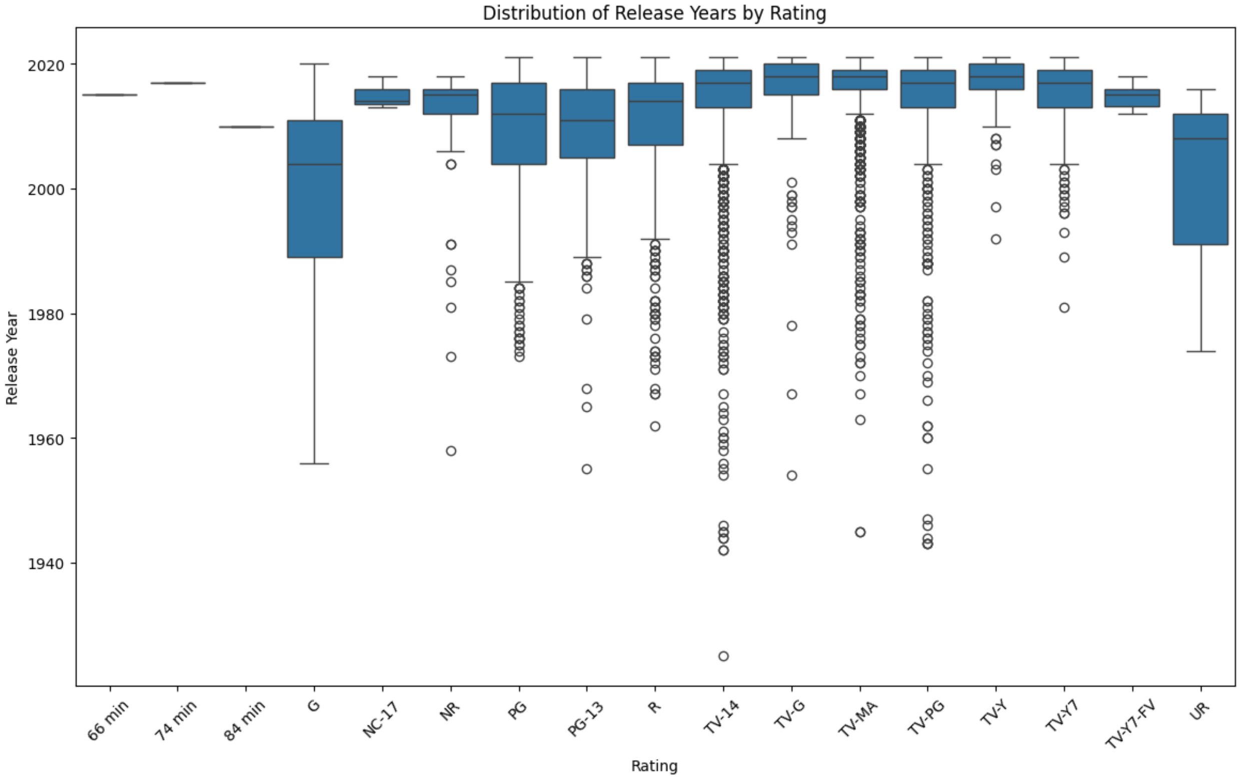
From above chart we can say the rating TV-MA and TV-14 is most of content from Movie and TV Show is for Teen and Mature audience

```
df.head()
```

```
  show_id     type                  title         director  \
0      s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show          Blood & Water           Unkown
2      s3  TV Show              Ganglands  Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans           Unkown
4      s5  TV Show           Kota Factory           Unkown

                                                cast        country  \
0                                             Unkown  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...            NaN
3                                             Unkown            NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...          India

  date_added  release_year rating   duration  \
0 2021-09-25          2020  PG-13     90 min
1 2021-09-24          2021  TV-MA  2 Seasons
2 2021-09-24          2021  TV-MA   1 Season
3 2021-09-24          2021  TV-MA   1 Season
4 2021-09-24          2021  TV-MA  2 Seasons

                                           listed_in  \
0                                      Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act...
3                             Docuseries, Reality TV
```

```
4  International TV Shows, Romantic TV Shows, TV ...

                                         description
0  As her father nears the end of his life, filmm...
1  After crossing paths at a party, a Cape Town t...
2  To protect his family from a powerful drug lor...
3  Feuds, flirtations and toilet talk go down amo...
4  In a city of coaching centers known to train I...
```

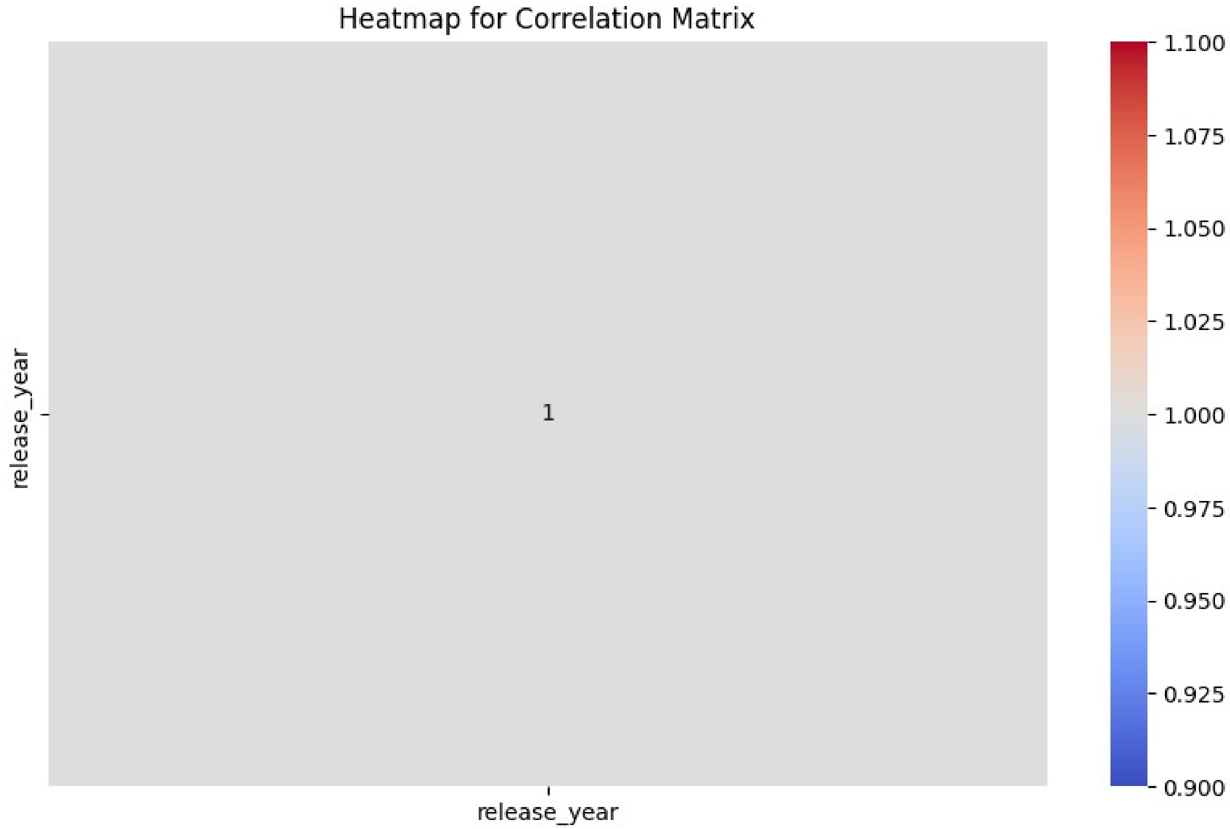
```
# Boxplot for rating vs. release_year
plt.figure(figsize=(14, 8))
sns.boxplot(x='rating', y='release_year', data=df)
plt.title('Distribution of Release Years by Rating')
plt.xlabel('Rating')
plt.ylabel('Release Year')
plt.xticks(rotation=45)
plt.show()
```



The Above boxplot shows that the median release year for most rating is recent

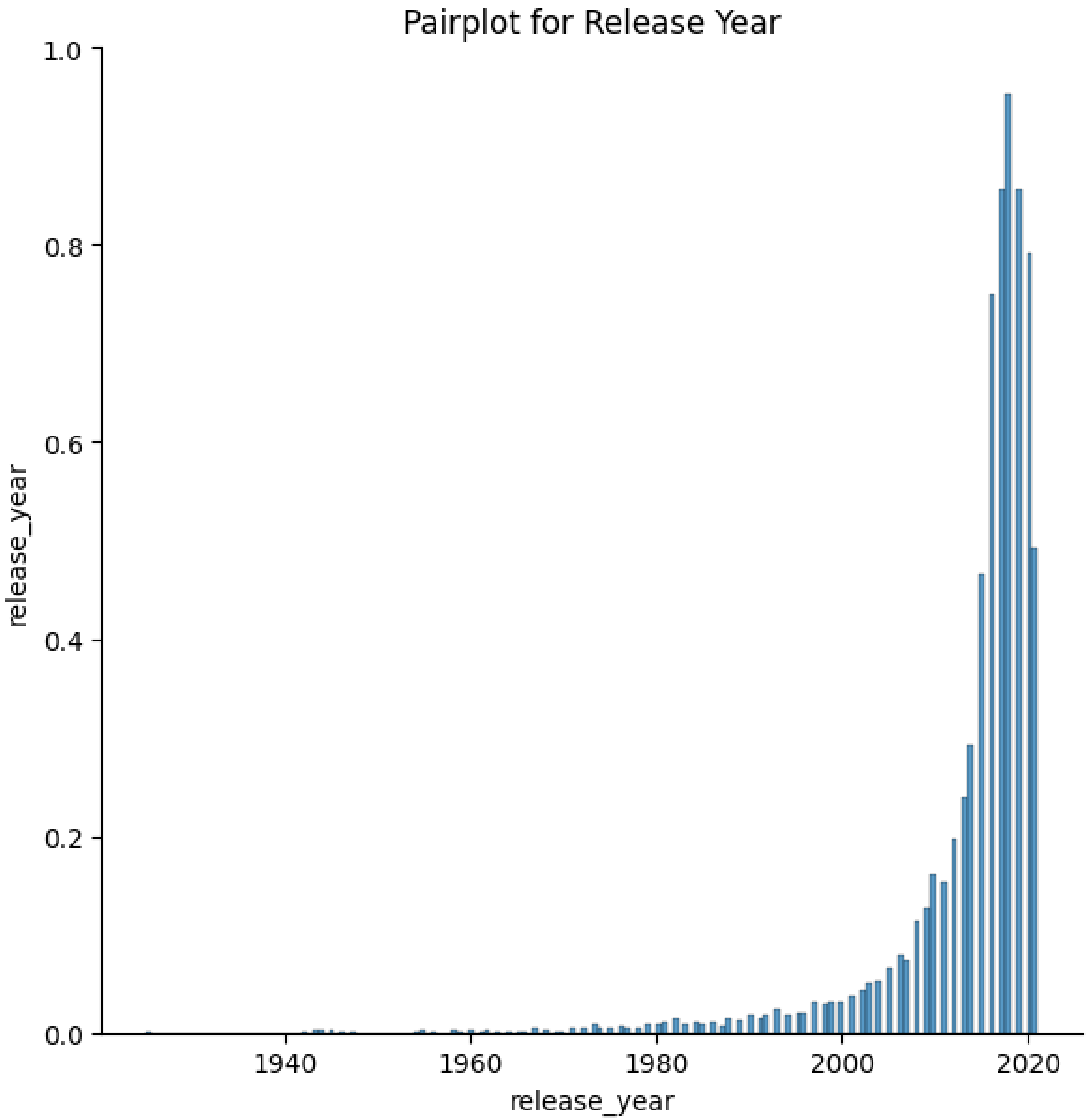
```
# Correlation Analysis
# Heatmap for correlation matrix
correlation_matrix = df.corr()

# Create a heatmap for the correlation matrix
plt.figure(figsize=(10, 6))
```

```
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Heatmap for Correlation Matrix')
plt.show()

<ipython-input-27-cef29990b263>:3: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it
will default to False. Select only valid columns or specify the value
of numeric_only to silence this warning.
  correlation_matrix = df.corr()
```



Only 1 continuous variable so heat map is not much informative, the diagonal elements are always 1 because any variable is perfectly correlated with itself.

```
# Pairplot for Continuous Variables which is Release year only one
Continuous variable
sns.pairplot(df[['release_year']], kind='scatter', height=6)
plt.title('Pairplot for Release Year')
plt.show()
```



Pairplot for Continuous Variables which is Release year only one Continuous variable same as heat map doesnt provide much information

Business Insights for Netflix Content Strategy

Insight- 1 Content

- Content on Netflix catalog is the most diversified with 748 countries, and the top three are the United States (2,818 titles), India (972 titles), and the United Kingdom (419 titles)
- This diversification helps diverse genres and audiences, which will help in enhancing penetration of more content from other regions.

Insight- 2 Rating

- In Rating of Netflix content 'TV-MA' and 'TV-14' dominate, comprising 61.2% of all titles (3,207 and 2,160 titles)

- These ratings suggest Netflix's focus on mature and teen audiences. Tailoring content strategies to these is likely to get more successful outcomes in customer retention and attracting new customers.

Insight- 3 Release Year

- The (36.4%) part of Netflix's content is from recent years, with 2018, 2017, and 2019 contributing 3,209 titles. TV Shows have a more recent median release year compared to Movies.
- Prioritizing newer content from new talents and regions keeps with viewer preferences as per trent and freshness according to the market, indicating Netflix's commitment to keeping its content up-to-date to maintain subscriber interest and also attracts newer subscribers.

Recommendations

- Adding Regional and Local Content: Content from the United States, India, and the United Kingdom makes up nearly 50% of the entire Netflix catalog. Content available from 748 different countries, Netflix has the opportunity to further expand its offerings based on regional popularity and local content encouraging Good popularity over time. This could lead to attracting local customers for subscription and customer satisfaction from various regions.

- Focus on other Geners to Attract various audience: Ratings ‘TV-MA’ and ‘TV-14’ account for 61.2% of all content. Genres like Documentaries and Children’s Movies or TV Shows are less frequent in the catalog and focus more on Teen and Mature Audience Genres as per numbers. But Netflix could diversify its portfolio by exploring underrepresented and also unappreciated genres to enrich and ratings to attract a more diversified audience like Kids and old age people with this Netflix can also change plans like College Students and family has different Plans this can be done with pricing data.

- Continue Older TV Shows or Remake Old Movies: The median release year for TV Shows is more recent compared to Movies. Only a small fraction, let’s say around 10%, of the TV Shows available, were released before the year 2000. Given this focus on newer TV Shows, Netflix could consider adding more classic old TV Shows and Movies to its catalog to attract a broader age group, including older adults who may have nostalgia for older series and also Remake Old Movies or TV Shows so even New Audience will be attracted.