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

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

```
--2023-11-15 04:27:11--  https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv
Resolving d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)... 18.172.139.61, 18.172.139.46, 18.172.139.210, ...
Connecting to d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)|18.172.139.61|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3399671 (3.2M) [text/plain]
Saving to: ‘netflix.csv’

netflix.csv         100%[===================>]   3.24M  2.37MB/s    in 1.4s

2023-11-15 04:27:13 (2.37 MB/s) - ‘netflix.csv’ saved [3399671/3399671]
```

```
netflix= pd.read_csv("netflix.csv")
```

```
netflix.shape
```

```
(8807, 12)
```

```
netflix.head()
```

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021
					Sami Bouajila,			

Inference from the above data

The dataset comprises more than 8,807 titles and 12 descriptions It resembles a typical movie and TV show dataset, notably lacking any ratings. Furthermore, some columns contain missing values (NaN).

```
netflix.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
```

```
print(netflix.isnull().any())
```

```
show_id         False
type            False
title           False
director         True
```

```
cast             True
country          True
date_added       True
release_year    False
rating           True
duration         True
listed_in       False
description     False
dtype: bool
```

Inference from the above data

From the info, we know that there are 8807 entries and 12 columns to work with for this EDA. There are few null values in the data set, we observe that columns like - "director", "cast", "country", "date_addded", "ratings" and "duration" have null values.

```
null_values = netflix.T.apply(lambda x: x.isnull().sum(), axis = 1)
```

```
null_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
```

```
null_values.sum()
```

```
4307
```

Inference from the above data

Based on the observations, it is evident that the dataset contains a total of **4,307** null values. Specifically, "director" has **2,634** , "cast" has **825**, and "country" has 831 null values. We will have to handle all null data points before we can dive into EDA and modelling.

```
netflix.director.fillna("No Director", inplace=True)
netflix.cast.fillna("No Cast", inplace=True)
netflix.country.fillna("Country Unavailable", inplace=True)
netflix.dropna(subset= ["date_added","duration", "rating"], inplace=True)
```

```
print(netflix.isnull().any())
```

```
show_id         False
type            False
title           False
director        False
cast            False
country         False
date_added      False
release_year    False
rating          False
duration        False
listed_in       False
description     False
dtype: bool
```

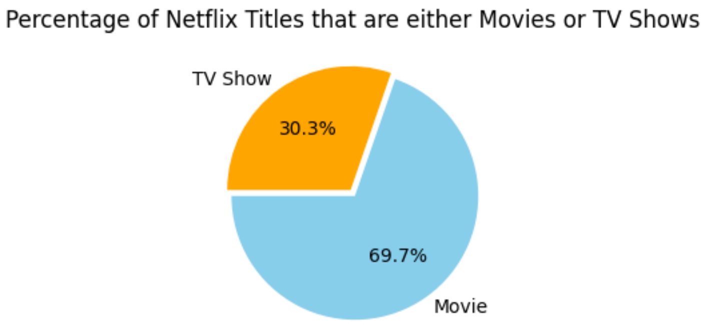
Inference from the above data

We addressed the issue of missing values by employing the imputation method. In particular, columns such as 'director,' 'cast,' and 'country' contained a significant number of missing values that couldn't be simply dropped without a loss of valuable data. To mitigate this, we utilized the .fillna function to replace the missing values with meaningful placeholders. Additionally, we performed data cleaning by removing rows with relatively fewer missing values in 'date_added,' 'duration,' and 'rating.' These steps allowed us to handle the null values effectively while preserving the integrity of the dataset."

```
netflix.head()
```

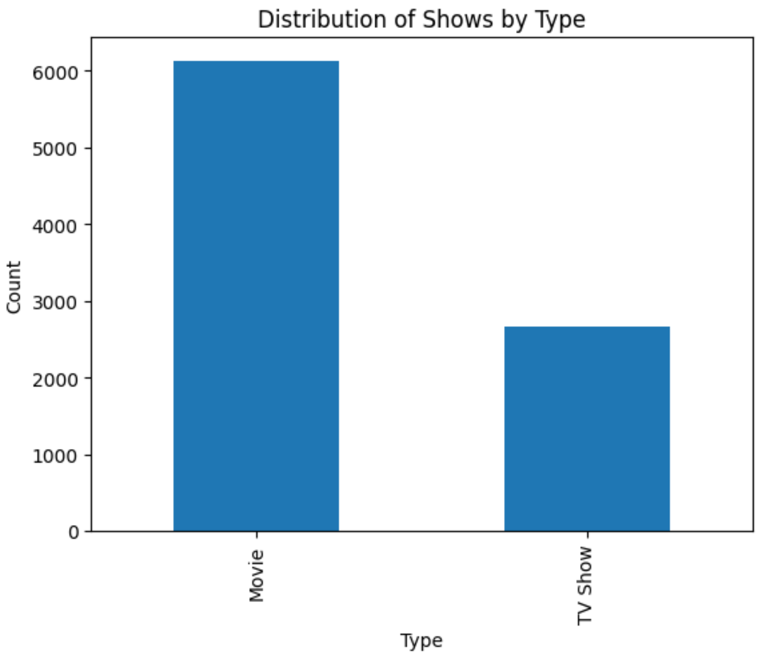
	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020
1	s2	TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021
2	s3	TV	Ganglands	Julien	Sami Bouajila, Tracy Gotoas...	Country	September	2021

```
plt.figure(figsize=(6,3))
plt.title("Percentage of Netflix Titles that are either Movies or TV Shows")
g=plt.pie(netflix.type.value_counts(),explode=(0.03,0.03),
labels=netflix.type.value_counts().index, colors=['skyblue','orange'],autopct='%1.1f%%',
startangle=180)
plt.show()
```



```
netflix['type'].value_counts().plot(kind='bar')
plt.title('Distribution of Shows by Type')
plt.xlabel('Type')
plt.ylabel('Count')
plt.show()

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



```
Movie      6126
TV Show    2664
Name: type, dtype: int64
```

Inference from the above data

Observing the pie chart above, it's evident that movies make up the majority, constituting approximately **69.7% (6126)** of the content, while TV shows account for the remaining **30.3%. (2664)**

```
Director = netflix.iloc[:,3]
```

```
grouped_data = netflix.groupby('director')
type_counts = grouped_data['type'].count()
```

```
top_directors= type_counts.sort_values(ascending= False).head(11)
```

```
top_directors
```

```
     director
     No Director               2621
     Rajiv Chilaka               19
     Raúl Campos, Jan Suter      18
     Marcus Raboy                16
     Suhas Kadav                 16
     Jay Karas                   14
     Cathy Garcia-Molina         13
     Jay Chapman                 12
     Youssef Chahine             12
     Martin Scorsese             12
     Steven Spielberg            11
     Name: type, dtype: int64
```

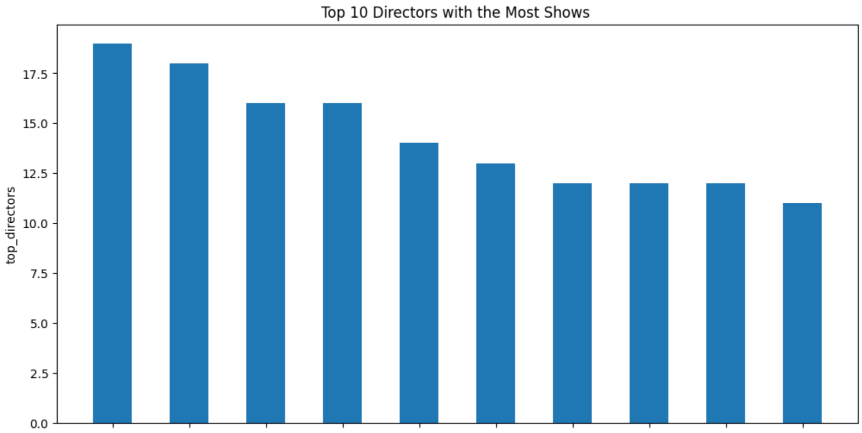
```
top_directors = top_directors.drop(top_directors.index[0])
```

```
top_directors
```

```
     director
     Rajiv Chilaka             19
     Raúl Campos, Jan Suter    18
     Marcus Raboy              16
     Suhas Kadav               16
     Jay Karas                 14
     Cathy Garcia-Molina       13
     Jay Chapman               12
     Youssef Chahine           12
     Martin Scorsese           12
     Steven Spielberg          11
     Name: type, dtype: int64
```

```
plt.figure(figsize =(12,6))
plt.bar(x=top_directors.index , height = top_directors,width = 0.5)
plt.title ("Top 10 Directors with the Most Shows ")
plt.xticks(rotation = 90)
plt.xlabel("director")
plt.ylabel("top_directors")
```

```
plt.show()
```



ANALYSIS ON MOVIES AND TV SHOWS

```
type_groupby = netflix.groupby(by ="type")
```

```
Movie_group = type_groupby.get_group("Movie")
Tv_show_group = type_groupby.get_group("TV Show")
```

```
Movie_group.head()
```

	show_id	type	title	director	cast	country	date_added	release_
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	Country Unavailable	September 24, 2021	
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United	September 24, 2021	

Inference from the above data

We have segregated movies and TV shows for our analysis, recognizing the inherent differences between the two. This separation allows us to gain more accurate and meaningful insights from our data, as each category warrants distinct examination and evaluation.

```
Movie_group.head()
```

	show_id	type	title	director	cast	country	date_added	release_
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsden, ...	Country Unavailable	September 24, 2021	
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United	September 24, 2021	

```
duration =Movie_group.iloc[: , [9]]
duration = Movie_group['duration'].str.split(" ").str.get(0).astype(int)
```

```
duration.mean().round(1)
```

```
99.6
```

Inference from the above data

"Based on the above analysis , it is evident that the average duration of movies on the platform is approximately 99.6 minutes.

```
Tv_show_group.head()
```

	show_id	type	title	director	cast	country	date_added	release_year
1	s2	TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Country Unavailable	September 24, 2021	2021

```
duration =Tv_show_group.iloc[: , [9]]
```

```
duration =Tv_show_group['duration'].str.split(" ").str.get(0).astype(int)
```

```
duration.mean().round(1)
```

```
1.8
```

Inference from the above data

"Based on the above analysis , it is evident that the average duration of TV Shows on the platform is approximately 1.8 seasons.

```
netflix.head()
```

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020
1	s2	TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021
2	s3	TV	Ganglands	Julien	Sami Bouajila, Tracy	Country	September	2021

```
country_groupby = netflix.groupby(by ="country")
```

```
Show_count_by_country = country_groupby.size()
```

```
Show_count_by_country =Show_count_by_country.sort_values(ascending= False)
```

```
Show_count_by_country
```

```
country
United States                                                                           2809
India                                                                                    972
```

```
Country Unavailable                                                             829
United Kingdom                                                                  418
Japan                                                                           243
                                                                               ...
Ireland, Canada, Luxembourg, United States, United Kingdom, Philippines, India    1
Ireland, Canada, United Kingdom, United States                                    1
Ireland, Canada, United States, United Kingdom                                    1
Ireland, France, Iceland, United States, Mexico, Belgium, United Kingdom, Hong Kong    1
Zimbabwe                                                                          1
Length: 749, dtype: int64
```

```
Top_10_highest_Show_count_by_country = Show_count_by_country.head(11)
```

```
Top_10_highest_Show_count_by_country = Top_10_highest_Show_count_by_country.drop(Top_10_highest_Show_count_by_country.index[2])
```

```
Top_10_highest_Show_count_by_country = Top_10_highest_Show_count_by_country.reset_index()
```

```
Top_10_highest_Show_count_by_country
```

	country	0
0	United States	2809
1	India	972
2	United Kingdom	418
3	Japan	243
4	South Korea	199
5	Canada	181
6	Spain	145
7	France	124
8	Mexico	110
9	Egypt	106

```
countries = Top_10_highest_Show_count_by_country['country']
movie_counts = Top_10_highest_Show_count_by_country[0]

plt.figure(figsize=(12, 6))
bars = plt.bar(countries, movie_counts)

plt.title("Top 10 Countries with the Most Shows on Netflix")
plt.xlabel("Country")
plt.ylabel("Number of Shows")
plt.xticks(rotation=45)

for bar, count in zip(bars, movie_counts):
    plt.text(bar.get_x() + bar.get_width() / 2, bar.get_height(), count, ha='center', va='bottom')

plt.show()
```

Inference from the above data

From the data, it's evident that the following countries have a significant presence in Netflix's movie library. These countries stand out as the top contributors to Netflix's diverse movie collection. This format provides a clear, organized presentation of the top countries and their corresponding movie counts.

```
netflix.head()
```

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020
1	s2	TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021
2	s3	TV	Ganglands	Julien	Sami Bouajila, Tracy	Country	September	2021

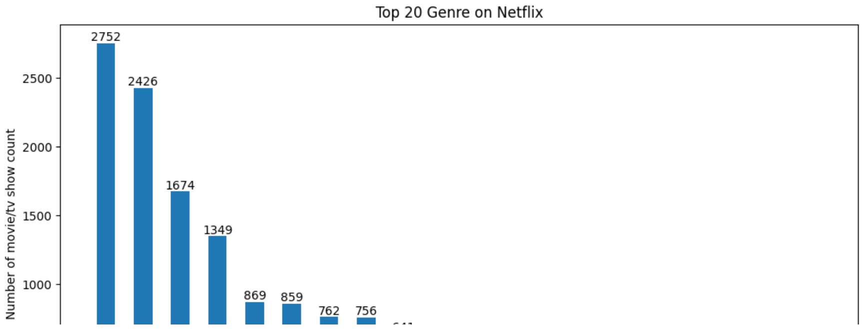
```
filtered_genres = netflix.set_index('title').listed_in.str.split(', ',
expand=True).stack().reset_index(level=1, drop=True);
```

```
genre = filtered_genres.value_counts().head(20)
```

```
plt.figure(figsize=(12, 6))
bars = plt.bar(genre.index,genre,width = .5)

plt.title("Top 20 Genre on Netflix")
plt.xlabel("Genre")
plt.ylabel("Number of movie/tv show count")
plt.xticks(rotation=90)
plt.show
for bar, count in zip(bars, genre):
    plt.text(bar.get_x() + bar.get_width() / 2, bar.get_height(), count, ha='center', va='bottom')

plt.show()
```



Key Insights :

Top 3 genres on Netflix, based on the dataset:

International Movies (2752 titles) Drama (2426 titles) Comedies (1674 titles) Recommendation: Netflix can maximize success by focusing investments on these popular genres, catering to diverse viewer preferences and enhancing the content library. Periodic analysis is advised for staying attuned to changing audience trends.

```
netflix.head()
```

	show_id	type	title	director	cast	country	date_added	release_year
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Cast	United States	September 25, 2021	2020
1	s2	TV Show	Blood & Water	No Director	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021
2	s3	TV	Ganglands	Julien	Sami Bouajila, Tracy Gotoas	Country	September	2021

```
date_added = netflix.iloc[:,[6]]
```

```
date_added
```

	date_added
0	September 25, 2021
1	September 24, 2021
2	September 24, 2021
3	September 24, 2021
4	September 24, 2021
...	...
8802	November 20, 2019
8803	July 1, 2019
8804	November 1, 2019
8805	January 11, 2020
8806	March 2, 2019

8790 rows × 1 columns

```
netflix['date_added'] = pd.to_datetime(netflix['date_added'].str.strip(), format='%B %d, %Y')
netflix['added_year'] = netflix['date_added'].dt.year
netflix['added_month'] = netflix['date_added'].dt.month_name()

date = netflix[['date_added', 'added_year', 'added_month']]
```

```
month_groups = netflix.groupby(by="added_month")
month_counts = month_groups.size()
month_counts = month_counts.sort_values(ascending=False)
```

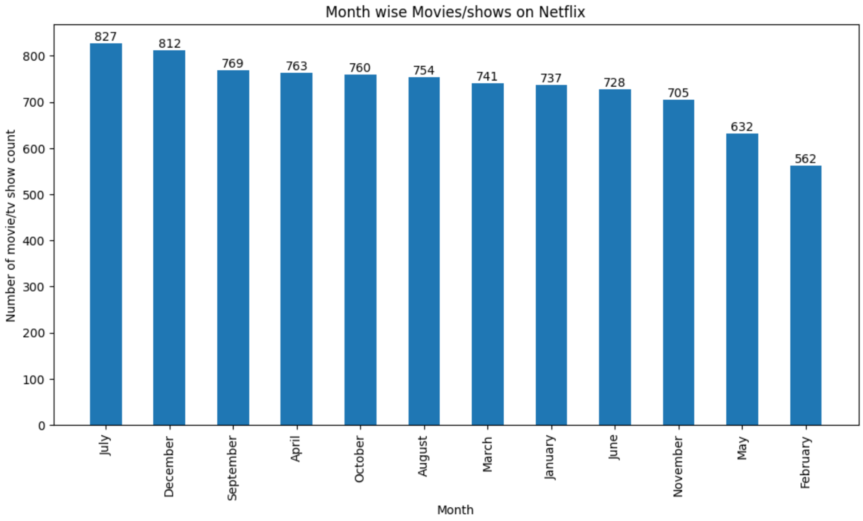
```
month_counts
```

```
added_month
July         827
December     812
September    769
April        763
October      760
August       754
March        741
January      737
June         728
November     705
May          632
February     562
dtype: int64
```

```
plt.figure(figsize=(12, 6))
bars = plt.bar(month_counts.index,month_counts,width = .5)

plt.title("Month wise Movies/shows on Netflix")
plt.xlabel("Month")
plt.ylabel("Number of movie/tv show count")
plt.xticks(rotation=90)
plt.show
for bar, count in zip(bars, month_counts):
    plt.text(bar.get_x() + bar.get_width() / 2, bar.get_height(), count, ha='center', va='bottom')

plt.show()
```



Insights from Netflix Monthly Additions:

Seasonal Peaks: High additions in July and December suggest seasonal peaks, possibly during holidays.

Consistent Releases: Months like September, April, and October show steady content additions, indicating a consistent release schedule.

**Variable Engagement:* * Lower counts in May, November, and February may imply variable viewer engagement, influenced by factors like holidays or events.

```
Year_groups = netflix.groupby(by="added_year")
Year_counts = Year_groups.size()

Year_counts = Year_counts.sort_values(ascending=False)
```

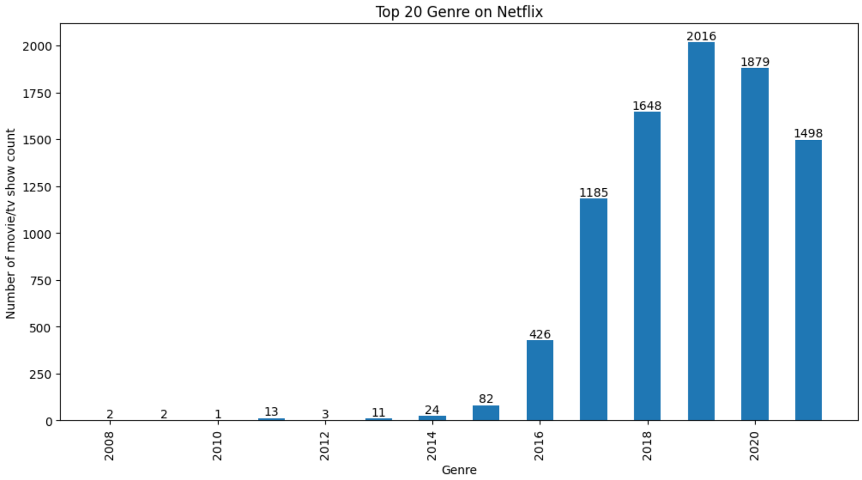
```
Year_counts

    added_year
    2019    2016
    2020    1879
    2018    1648
    2021    1498
    2017    1185
    2016     426
    2015      82
    2014      24
    2011      13
    2013      11
    2012       3
    2008       2
    2009       2
    2010       1
    dtype: int64
```

```
plt.figure(figsize=(12, 6))
bars = plt.bar(Year_counts.index,Year_counts,width = .5)

plt.title("Top 20 Genre on Netflix")
plt.xlabel("Genre")
plt.ylabel("Number of movie/tv show count")
plt.xticks(rotation=90)
plt.show
for bar, count in zip(bars, Year_counts):
    plt.text(bar.get_x() + bar.get_width() / 2, bar.get_height(), count, ha='center', va='bottom')

plt.show()
```



Insights from Netflix Yearly Additions:

Yearly Growth: Netflix experienced substantial growth in content additions over the years, with a significant increase from 2016 to 2020.

Consistent Expansion: The years 2018, 2019, and 2020 witnessed the highest content additions, indicating a consistent effort to expand the streaming library.

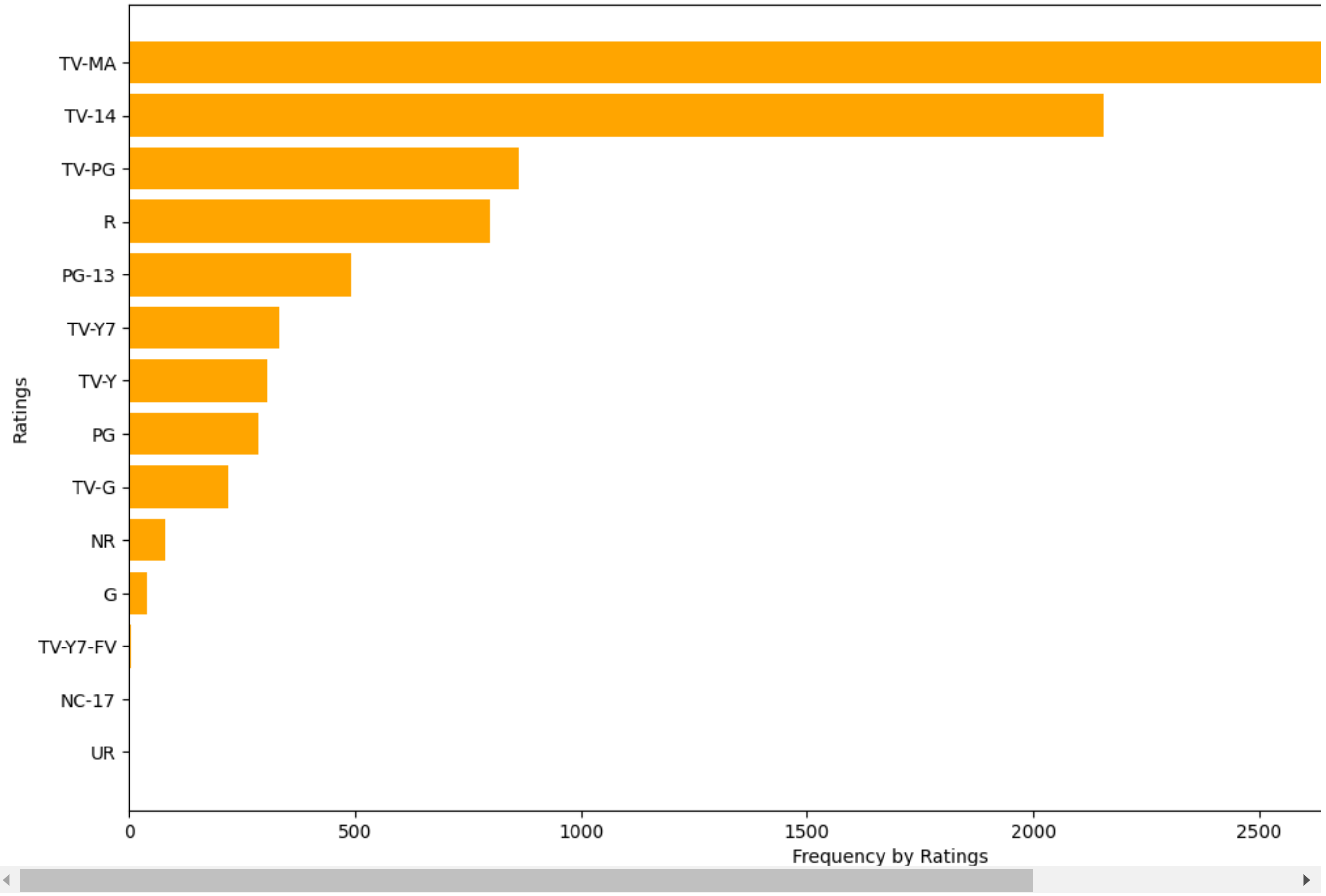
Establishment Phase: Earlier years (2011 to 2015) show relatively lower counts, suggesting Netflix's establishment phase and gradual ramp-up of content.

```
cast = netflix['cast']
```

```
netflix.groupby(['rating']).agg({"title":"nunique"})
```

rating	title
G	41
NC-17	3
NR	79
PG	287
PG-13	490
R	799
TV-14	2157
TV-G	220
TV-MA	3205
TV-PG	861
TV-Y	306
TV-Y7	333
TV-Y7-FV	6

```
df_rating=netflix.groupby(['rating']).agg({"title":"nunique"}).reset_index().sort_values(by=['title'],ascending=False)[:15]
plt.figure(figsize=(15,8))
plt.barh(df_rating[::-1]['rating'], df_rating[::-1]['title'],color=['orange'])
plt.xlabel('Frequency by Ratings')
plt.ylabel('Ratings')
plt.show()
```


TV-MA
TV-14
TV-PG
R
PG-13
TV-Y7
TV-Y
PG
TV-G
NR
G
TV-Y7-FV
NC-17
UR
Ratings
Frequency by Ratings