```python
In [3]: # IMPORTING VARIOUS LIBRARIES
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```python
In [6]: # UPLOADING DATA AND SEE TOP 5 ROWS
df= pd.read_csv("netflix.csv")
df.head(5)
```

Out[6]:

	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...

DATA INFO

```python
In [8]: 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
```

```python
In [77]: df.shape # THIS SHOWS DATA CONTAINS 8807 ROWS AND 12 COLUMNS
```

```
Out[77]: (8807, 12)
```

Analyse for Null values

```python
In [78]: df.isna().sum() # This shows data contains null values at director, cast, country, date_added,rating and duration column
```

```
Out[78]:
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
```

As the data shows - director, cast , date_added ,rating and duration column consist of null values. lets check for null values

```python
In [79]: df[df["director"].isna()]
```

Out[79]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
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...
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...
10	s11	TV Show	Vendetta: Truth, Lies and The Mafia	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, Docuseries, International TV S...	Sicily boasts a bold "Anti-Mafia" coalition. B...
14	s15	TV Show	Crime Stories: India Detectives	NaN	NaN	NaN	September 22, 2021	2021	TV-MA	1 Season	British TV Shows, Crime TV Shows, Docuseries	Cameras following Bengaluru police on the job ...
...	...	...	...	...	...	...	...	...	...	...	...	...
8795	s8796	TV Show	Yu-Gi-Oh! Arc-V	NaN	Mike Liscio, Emily Bauer, Billy Bob Thompson, ...	Japan, Canada	May 1, 2018	2015	TV-Y7	2 Seasons	Anime Series, Kids' TV	Now that he's discovered the Pendulum Summonin...
8796	s8797	TV Show	Yunus Emre	NaN	Gökhan Atalay, Payidar Tüfekçioglu, Baran Akbu...	Turkey	January 17, 2017	2016	TV-PG	2 Seasons	International TV Shows, TV Dramas	During the Mongol invasions, Yunus Emre leaves...
8797	s8798	TV Show	Zak Storm	NaN	Michael Johnston, Jessica Gee-George, Christin...	United States, France, South Korea, Indonesia	September 13, 2018	2016	TV-Y7	3 Seasons	Kids' TV	Teen surfer Zak Storm is mysteriously transpor...
8800	s8801	TV Show	Zindagi Gulzar Hai	NaN	Sanam Saeed, Fawad Khan, Ayesha Omer, Mehreen ...	Pakistan	December 15, 2016	2012	TV-PG	1 Season	International TV Shows, Romantic TV Shows, TV ...	Strong-willed, middle-class Kashaf and carefre...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...

2634 rows × 12 columns

```python
In [80]: df['director']= df['director'].fillna('unknown_director') # filling null values at director column with "unknown_director"
df
```

Out[80]:

	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	unknown_director	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	unknown_director	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	unknown_director	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...
...	...	...	...	...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	unknown_director	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero...
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty...

8807 rows × 12 columns

```python
In [81]: df.isna().sum()# director column now showing 0 null,as we have filled null values with "unknown"
```

```
Out[81]:
show_id            0
type               0
title              0
director           0
cast             825
country          831
date_added        10
release_year       0
rating             4
duration           3
listed_in          0
description        0
dtype: int64
```

```python
In [82]: for a in ['date_added','rating','duration']:
    df[a] = df[a].fillna(df[a].mode()[0]) #filling missing values with mode
```

```python
In [83]: df['country']=df['country'].fillna('unknown_country')
df['cast']=df['cast'].fillna('unknown_cast')
```

```python
In [84]: df.isna().sum() #as we have filled all the null values now data shows no null values and it is very much cleaned
```

```
Out[84]:
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
description     0
dtype: int64
```

Checking for unique values

```python
In [85]: df['director'].nunique() #director column contains 4529 unique director entries
```

```
Out[85]: 4529
```

```python
In [86]: df['title'].nunique() # title column contains 8807 unique entries
```

```
Out[86]: 8807
```

```python
In [87]: df['cast'].nunique() # cast column has 7693 unique entries
```

```
Out[87]: 7693
```

```python
In [88]: df['type'].unique() # type column has two unique entries as Movie and Tv Show
```

```
Out[88]: array(['Movie', 'TV Show'], dtype=object)
```

Checking for Data types

```python
In [89]: df.dtypes #all the data type seems correct except date_added. it should be timestamp
```

```
Out[89]:
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
```

```python
In [90]: df['date_added'] = df['date_added'].str.strip()
df['date_added'] = pd.to_datetime(df['date_added'], format='%B %d, %Y') #Changing date_added datatype to datetime
df.dtypes
```

```
Out[90]:
show_id                 object
type                    object
title                   object
director                object
cast                    object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration                object
listed_in               object
description             object
dtype: object
```

```python
In [91]: #Making of new columns day, month, year for better analysing
df['date']=df['date_added'].dt.day
df['month']=df['date_added'].dt.month
df['year']=df['date_added'].dt.year
df.dtypes
```

```
Out[91]:
show_id                 object
type                    object
title                   object
director                object
cast                    object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration                object
listed_in               object
description             object
date                     int64
month                    int64
year                     int64
dtype: object
```

```python
In [92]: #checking for value counts
df['director'].value_counts()
```

```
Out[92]:
unknown_director                  2634
Rajiv Chilaka                       19
Raúl Campos, Jan Suter              18
Suhas Kadav                         16
Marcus Raboy                        16
                                  ...
Raymie Muzquiz, Stu Livingston       1
Joe Menendez                         1
Eric Bross                           1
Will Eisenberg                       1
Mozez Singh                          1
Name: director, Length: 4529, dtype: int64
```

```python
In [93]: df['cast'].value_counts()
```

```
Out[93]:
unknown_cast
825
David Attenborough
19
Vatsal Dubey, Julie Tejwani, Rupa Bhimani, Jigna Bhardwaj, Rajesh Kava, Mousam, Swapnil
14
Samuel West
10
Jeff Dunham
7

...
Nick Lachey, Vanessa Lachey
1
Takeru Sato, Kasumi Arimura, Haru, Kentaro Sakaguchi, Takayuki Yamada, Kendo Kobayashi, Ken Yasuda, Arata Furuta, Suzuki Matsuo, Koichi Yamadera,
Arata Iura, Chikako Kaku, Kotaro Yoshida      1
Toyin Abraham, Sambasa Nzeribe, Chioma Chukwuka Akpotha, Chioma Omeruah, Chiwetalu Agu, Dele Odule, Femi Adebayo, Bayray McNwizu, Biodun Stephen
1
Neeraj Kabi, Geetanjali Kulkarni, Danish Husain, Sheeba Chaddha, Paras Priyadarshan, Anshul Chauhan, Anud Singh Dhaka, Shirin Sewani, Mihir Ahuja,
Vasundhara Rajput                             1
Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanana, Manish Chaudhary, Meghna Malik, Malkeet Rauni, Anita Shabdish, Chittaranjan Tripathy
1
Name: cast, Length: 7693, dtype: int64
```

```python
In [94]: df['type'].value_counts() #This shows Movie type has 6131 entry whereas Tv Show type has 2676 entry
```

```
Out[94]:
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

```python
In [95]: df['title'].value_counts()
```

```
Out[95]:
Dick Johnson Is Dead                     1
Ip Man 2                                 1
Hannibal Buress: Comedy Camisado         1
Turbo FAST                               1
Masha's Tales                            1
                                        ..
Love for Sale 2                          1
ROAD TO ROMA                             1
Good Time                                1
Captain Underpants Epic Choice-o-Rama    1
Zubaan                                   1
Name: title, Length: 8807, dtype: int64
```

```python
In [96]: df['country'].value_counts() #To check the value counts of country column
```

```
Out[96]:
United States                             2818
India                                      972
unknown_country                            831
United Kingdom                             419
Japan                                      245
                                          ...
Romania, Bulgaria, Hungary                   1
Uruguay, Guatemala                           1
France, Senegal, Belgium                     1
Mexico, United States, Spain, Colombia       1
United Arab Emirates, Jordan                 1
Name: country, Length: 749, dtype: int64
```

```python
In [97]: df['listed_in'].value_counts()
```

```
Out[97]:
Dramas, International Movies                          362
Documentaries                                         359
Stand-Up Comedy                                       334
Comedies, Dramas, International Movies                274
Dramas, Independent Movies, International Movies      252
                                                     ...
Kids' TV, TV Action & Adventure, TV Dramas              1
TV Comedies, TV Dramas, TV Horror                       1
Children & Family Movies, Comedies, LGBTQ Movies        1
Kids' TV, Spanish-Language TV Shows, Teen TV Shows      1
Cult Movies, Dramas, Thrillers                          1
Name: listed_in, Length: 514, dtype: int64
```

```python
In [98]: df['rating'].value_counts() # this column has anonyms value as 74 min, 84 min, 66min
```

```
Out[98]:
TV-MA       3211
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
74 min         1
84 min         1
66 min         1
Name: rating, dtype: int64
```

```python
In [99]: df['duration'].value_counts()
```

```
Out[99]:
1 Season     1796
2 Seasons     425
3 Seasons     199
90 min        152
94 min        146
             ...
16 min          1
186 min         1
193 min         1
189 min         1
191 min         1
Name: duration, Length: 220, dtype: int64
```

```python
In [100…: #Cleaning of country column
df['country']= df['country'].astype(str)
df['country']= df['country'].apply(lambda x: x.split(', ')[0])
```

```python
In [101…: df['country'].value_counts()
```

```
Out[101]:
United States      3210
India              1008
unknown_country     831
United Kingdom      626
Canada              271
                   ...
Namibia               1
Senegal               1
Cameroon              1
Syria                 1
Somalia               1
Name: country, Length: 90, dtype: int64
```

```python
In [102…: #Cleaning of rating column
df['rating'].unique()
```

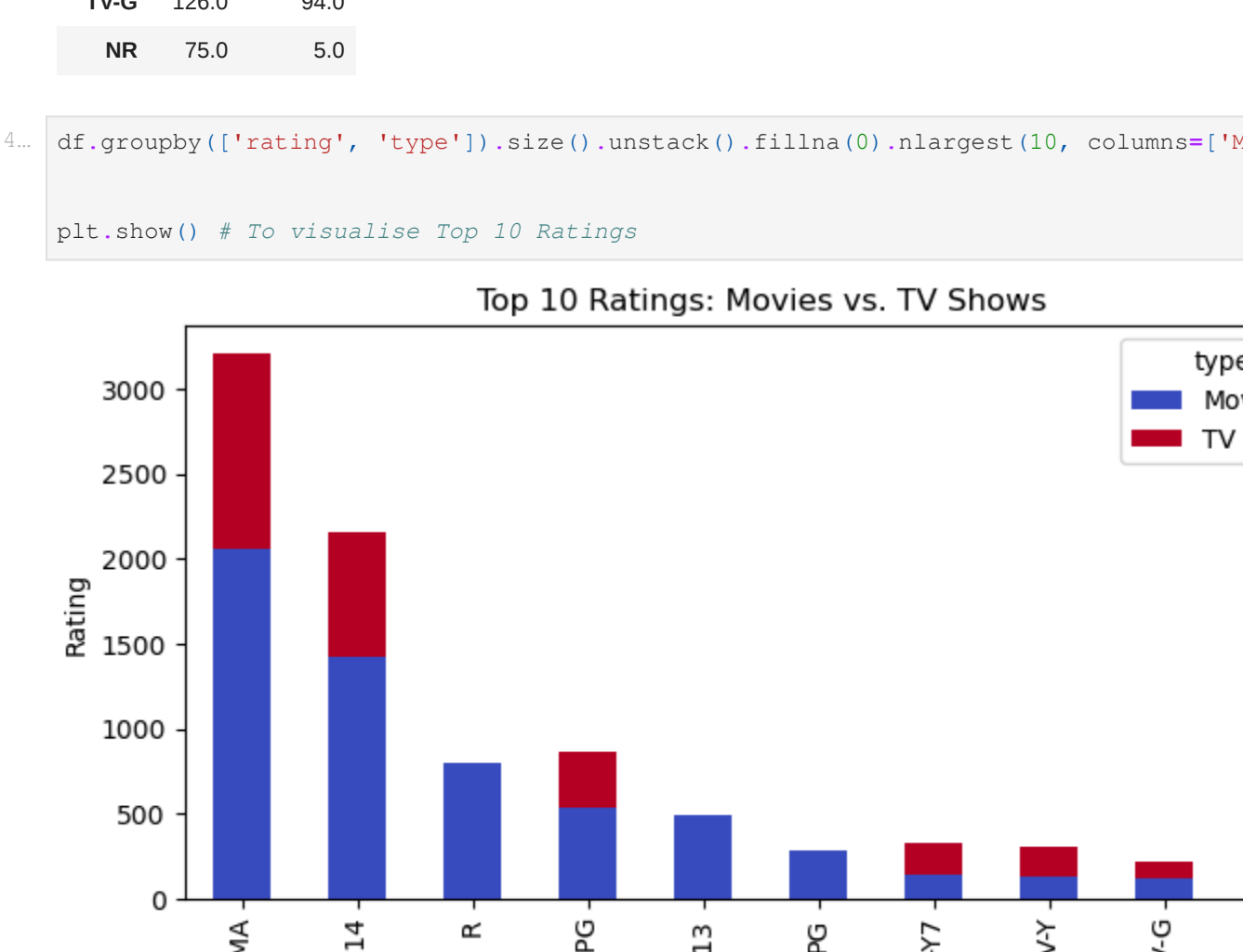
```
Out[102]: array(['PG-13', 'TV-MA', 'PG', 'TV-14', 'TV-PG', 'TV-Y', 'TV-Y7', 'R',
       'TV-G', 'G', 'NC-17', '74 min', '84 min', '66 min', 'NR',
       'TV-Y7-FV', 'UR'], dtype=object)
```

```python
In [110…: df.groupby(['rating', 'type']).size().unstack().fillna(0).nlargest(10, columns=['Movie', 'TV Show'])
```

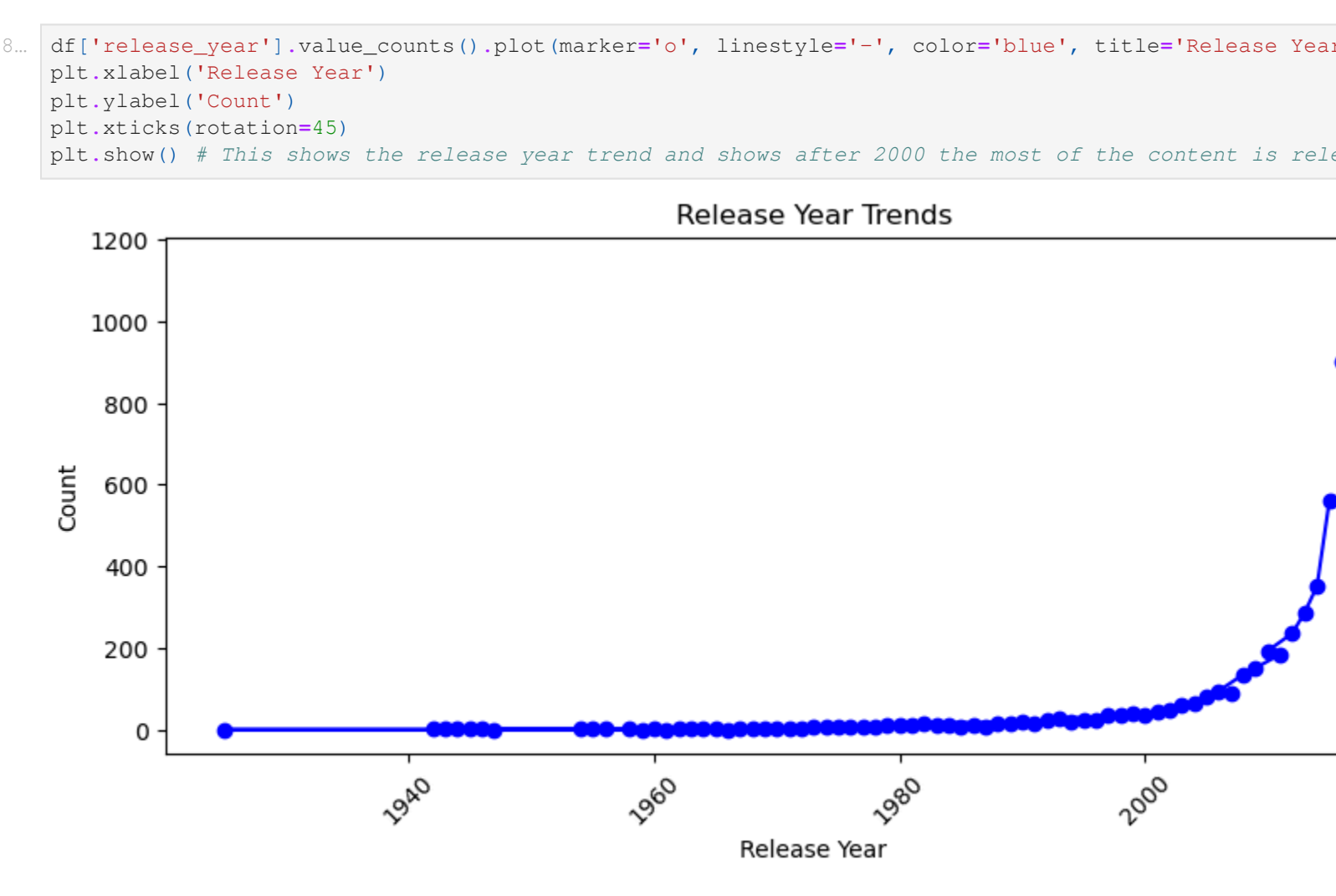
Out[110]:

type	Movie	TV Show
rating		
TV-MA	2064.0	1147.0
TV-14	1427.0	733.0
R	797.0	2.0
TV-PG	540.0	323.0
PG-13	490.0	0.0
PG	287.0	0.0
TV-Y7	139.0	195.0
TV-Y	131.0	176.0
TV-G	126.0	94.0
NR	75.0	5.0

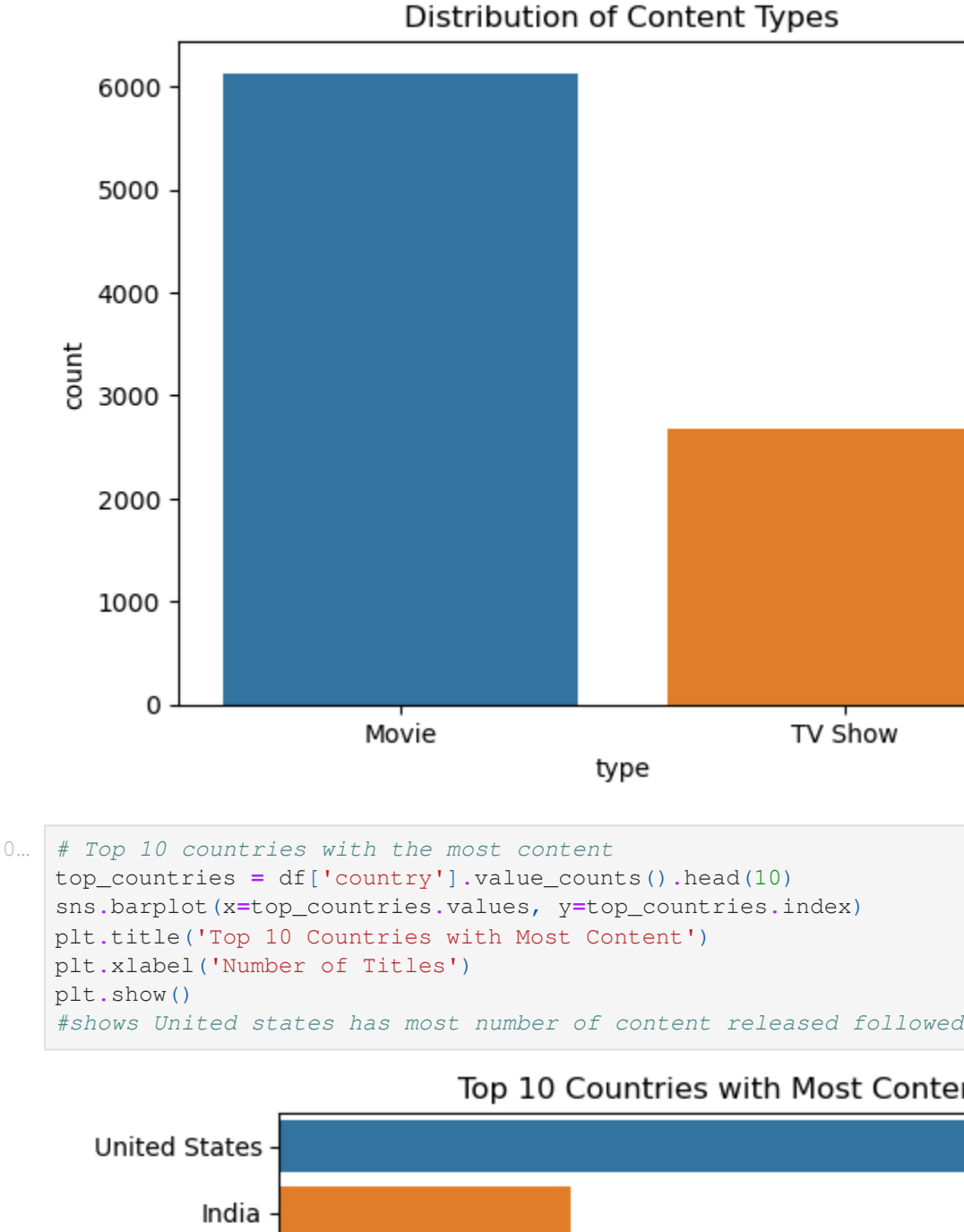
```python
In [114…: df.groupby(['rating', 'type']).size().unstack().fillna(0).nlargest(10, columns=['Movie', 'TV Show']).plot(kind='bar', stacked=True, colormap='coolw
                                                                                                   figsize=(8, 4),
                                                                                                   title='Top 10 Ratings: Movies vs. TV Show
plt.show() # To visualise Top 10 Ratings
```



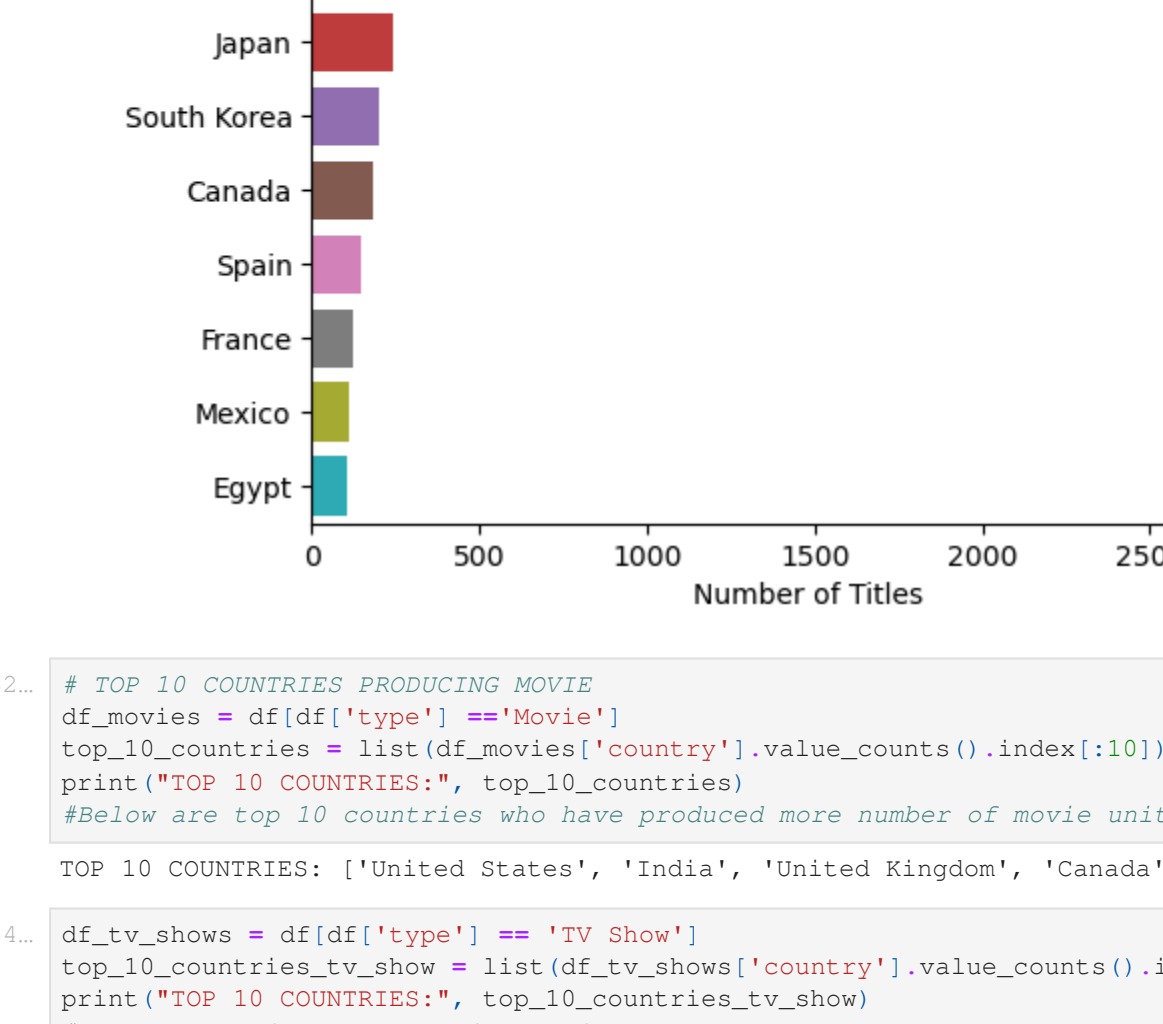

```python
In [118…: df['release_year'].value_counts().plot(marker='o', linestyle='-', color='blue', title='Release Year Trends' , figsize= (10,4))
plt.xlabel('Release Year')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show() # This shows the release year trend and shows after 2000 the most of the content is released
```




```python
In [121…: sns.countplot(data=df, x='type')
plt.title('Distribution of Content Types')
plt.show() # distribution of content shows type Movies are released more than that of Tv Show
```




```python
In [130…: # Top 10 countries with the most content
top_countries = df['country'].value_counts().head(10)
sns.barplot(x=top_countries.values, y=top_countries.index)
plt.title('Top 10 Countries with Most Content')
plt.xlabel('Number of Titles')
plt.show()
#shows United states has most number of content released followed by India
```




```python
In [132…: # TOP 10 COUNTRIES PRODUCING MOVIE
df_movies = df[df['type'] =='Movie']
top_10_countries = list(df_movies['country'].value_counts().index[:10])
print("TOP 10 COUNTRIES:", top_10_countries)
#Below are top 10 countries who have produced more number of movie united states are on the top
```

```
TOP 10 COUNTRIES: ['United States', 'India', 'United Kingdom', 'Canada', 'Spain', 'Egypt', 'Nigeria', 'Indonesia', 'Turkey', 'Japan']
```

```python
In [134…: df_tv_shows = df[df['type'] == 'TV Show']
top_10_countries_tv_show = list(df_tv_shows['country'].value_counts().index[:10])
print("TOP 10 COUNTRIES:", top_10_countries_tv_show)
# Below are list of countries having more number of Tv Show released. United states here also on maintaining the top position
```

```
TOP 10 COUNTRIES: ['United States', 'United Kingdom', 'Japan', 'South Korea', 'India', 'Taiwan', 'Canada', 'France', 'Australia', 'Spain']
```

No of distinct titles on the basis of genre

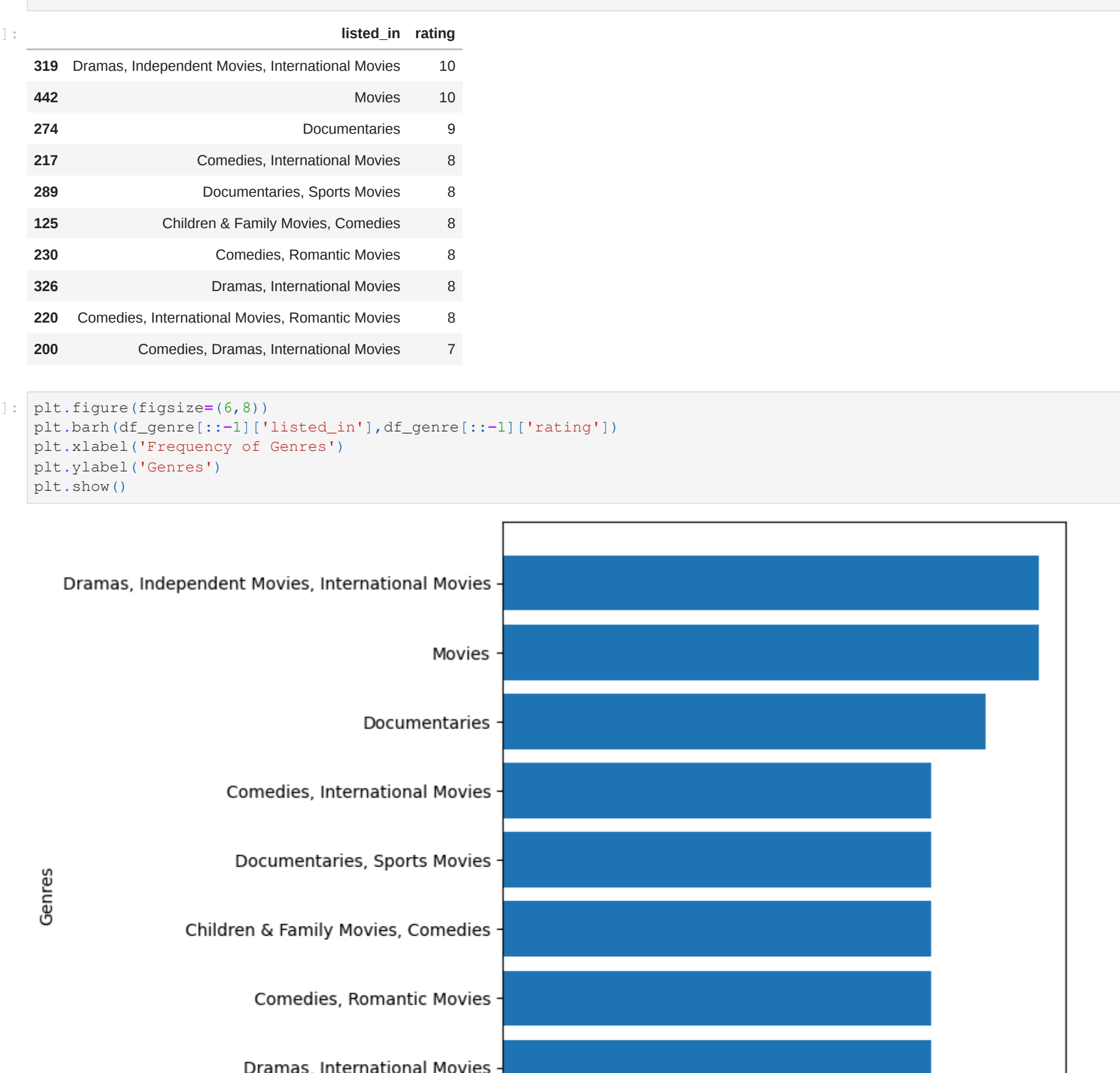
```python
In [9]: df_genre = df.groupby(['listed_in']).agg({'rating':"nunique"}).reset_index().sort_values(by=['rating'],ascending = False)[:10]
```

```python
In [10]: df_genre  #Top 10 genres as per the rating
```

Out[10]:

	listed_in	rating
319	Dramas, Independent Movies, International Movies	10
442	Movies	10
274	Documentaries	9
217	Comedies, International Movies	8
289	Documentaries, Sports Movies	8
125	Children & Family Movies, Comedies	8
230	Comedies, Romantic Movies	8
326	Dramas, International Movies	8
220	Comedies, International Movies, Romantic Movies	8
200	Comedies, Dramas, International Movies	7

```python
In [12]: plt.figure(figsize=(6,8))
plt.barh(df_genre[::-1]['listed_in'],df_genre[::-1]['rating'])
plt.xlabel('Frequency of Genres')
plt.ylabel('Genres')
plt.show()
```




```
In [ ]:
In [ ]:
In [ ]:
In [ ]:
```