Data Analysis: Netflix

1. Aims, objectives and background

1.1 Introduction

Data analysis has been used by many companies for a long time. With the main objective of looking for trends and answering questions using data visualization techniques.

As one of the top streaming service in 2022, Netflix has been a staple to many households. From Tv shows to movies, there are over 17,000 titles worldwide. As an avid fan of Netflix, I have spent countless hours binging on the latest hits. Using this opportunity, I want to explore Netflix and its content.

Being a huge company, analysing its data is will allow for Informed decision-making based on facts, Reduce costs by looking for improvement opportunities, trends, and patterns in their data and plan their strategies accordingly, provide a better user recommendations by analysing their interest and behaviours.

1.2 Aims and objectives

Through this project, i would like to find out:

- How much content there actually is
- When did Netflix peaked
- Look for any trends and relationships

The aims for this project proposal are to:

- Find a dataset that allows for data cleaning and analysis
- Clean the data to make it usable
- Visualizing the data using graph plotting techniques
- Analyzing the data for trends and answering questions

1.3 Data

1.3.1 Data acquisition and reliability

After doing some research, I have found out the official Netflix API has been shutdown in 2014, This lead me to searching for datasets that are available on the internet. Which brought me to Kaggle, an online community platform for data scientists, that allows users to collaborate with other users, find and publish datasets.

Based on the posting, the dataset was regularly updated up till September 2021 and based on the usability rating it scored a 10/10, showing this can be a trusted and reliable dataset.

1.3.2 Data type and suitability

The dataset is in the CSV format allowing it to be easily used with Pandas. It consists of the main information such as title, country, date added etc, this allows us to answer the questions that i have posed.

1.3.3 Time period

The years provided in the dataset span from 1925 - 2021. However, having only a small quantity of titles (<10) in that particular year, it would not make a significant impact on the changes of the data. Hence, we will be focusing mainly on 2000 - 2021.

1.4 Ethical consideration

1.4.1 copyrights

Based on Kaggle's terms, I am allowed to use it as long as it is for personal and non-commercial. For specific copyrights, they will be displayed on the dataset page.

After looking through, I found the license on the dataset's page, CC0: Public Domain, which states "The person who associated a work with this deed has dedicated the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law. You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission." Thus, the data is dafe to use.

1.4.2 Reusage of data

The legality of the dataset has been confimed as of writing. Anyone planning on using the dataset provided may refer to the link in the references for the terms and conditions to prevent any infringement of the law.

1.4.2 Potential impacts

As this project is used for personal and educational use, any conclusions from this project will be based on assumptions after looking at the graphs and should not be used as a professional analytical judgement.

2. Importing Data and libraries

importing the required libraries for data visulization

In [1]:

```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import plotly.graph_objects as pgo
```

Applying error handling methods, in case a file is not available.

In [2]:

```
try:
    netflix = pd.read_csv('netflix_titles.csv')
except FileNotFoundError:
    print("File does not exist")
```

Original dataset with missing values

In [3]:

```
netflix
```

Out[3]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descripti
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As h father nea the end his li filmn
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	Af crossi paths a party Cape To
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 prote his fam fron power drug lo
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feu flirtatio and toi talk 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 coachi cente known train
...	...	...	...	...	...	...	...	...	...	...	...	
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 politi cartooni a cri repor and a
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 livi alone ir spoo town 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 survive i world tak 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	Dragg from civili life forn 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 scrap but po boy worl his way ir a t

8807 rows × 12 columns

3. Data cleaning and processing

Before plotting any graphs, I need to ensure that the dataset does not have any missing data values by:

1. scanning through the data set and calculating the percentage of missing values
2. Filling up missing data with as much relevant details
3. Remaining missing data will be dropped

3.1 Calculation of missing data in percentage

In [4]:

```
for i in netflix.columns:
    nullPercentage = netflix[i].isna().sum() / len(netflix)*100
    if nullPercentage > 0 :
        print(f"{i} NaN rate is {round(nullPercentage,2)}%")
```

```
director NaN rate is 29.91%
cast NaN rate is 9.37%
country NaN rate is 9.44%
date_added NaN rate is 0.11%
rating NaN rate is 0.05%
duration NaN rate is 0.03%
```

There are 6 columns that have missing values.

- Director and cast's information is missing a big portion, although it is not important for the purpose of data analysis,it is vital so that it will not be dropped later on.
- Country's information is vital for our data analysis, hence, it is vital to fill it up with the most repeated country. This will not impact the the data analysis results later on.
- date_added, rating and duration will be dropped as their amount will have little impact on the dataset.

In [5]:

```
netflix["cast"].fillna("No Data", inplace=True)
netflix["director"].fillna("No Data", inplace=True)
netflix["country"].fillna(netflix["country"].mode()[0])
netflix.dropna(inplace= True)
netflix.drop_duplicates(inplace=True) #drop duplicate rows
```

Double checking all missing values has been filled or removed.

In [6]:

```
for i in netflix.columns:
    nullPercentage = netflix[i].isna().sum() / len(netflix)*100
    if nullPercentage > 0 :
        print(f"{i} NaN rate is {round(nullPercentage,2)}%")
    else:
        print("no missing values")
```

```
no missing values
no missing values
no missing values
no missing values
no missing values
no missing values
no missing values
no missing values
no missing values
no missing values
no missing values
no missing values
```

Dataset after cleaning

In [7]:

```
netflix
```

Out[7]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	descr
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	No Data	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	father the h f
1	s2	TV Show	Blood & Water	No Data	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	cr pat p Cape
4	s5	TV Show	Kota Factory	No Data	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 co c kn t
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies	On a sl Gha Am mo
8	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Ho...	United Kingdom	September 24, 2021	2021	TV-14	9 Seasons	British TV Shows, Reality TV	A ta ba ar baker c
...	...	...	...	...	...	...	...	...	...	...	...	
8801	s8802	Movie	Zinzana	Majid Al Ansari	Ali Suliman, Saleh Bakri, Yasa, Ali Al-Jabri, ...	United Arab Emirates, Jordan	March 9, 2016	2015	TV-MA	96 min	Dramas, International Movies, Thrillers	Reco alo Talal up in
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 p carto a re a
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Look survi world over b
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	Dr from super
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 s bu boy his w

7961 rows × 12 columns

4. Data analysis

4.1 Movies vs TV

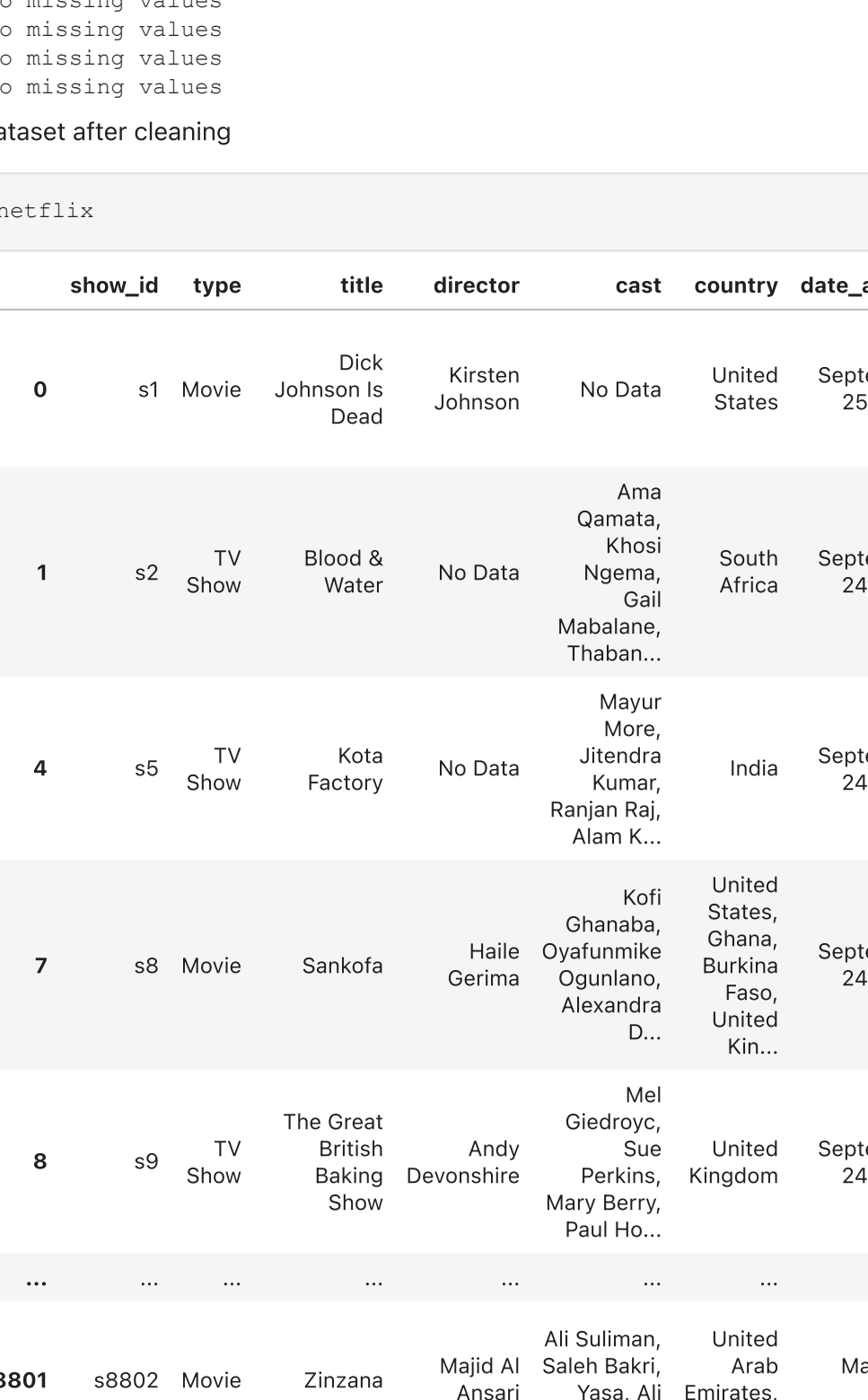
We will be comparing the number of Movies to TV shows on Netflix

In [8]:

```
netflixShows=netflix[netflix['type']=='TV Show']
netflixMovies=netflix[netflix['type']=='Movie']
```

In [9]:

```
plt.figure(figsize=(10,10))
sns.set(style="darkgrid")
graph = sns.countplot(x="type", data = netflix, palette= "Set1")
```



Here, we can see there are almost double the number of movies to TV shows.

4.2 Amount of content by country

Under the country column, there are some rows where there are multiple countries name, so to resolve that we will take the first country in that row.

In [10]:

```
#lambda function to split the countries by " " and taking the first country
netflix["firstCountry"] = netflix["country"].apply(lambda x: x.split(",")[0])

netflix['counter'] = 1
data = netflix.groupby('firstCountry')['counter'].sum().sort_values(ascending=False)[:8]
```

In [11]:

```
#Draw figure/ setting colour
color = ['#FFFFFF' for i in range(10)]
color[0] = color[1]= color[2] = '#0000FF'
figure,axis = plt.subplots(1,1,figsize=(12,6))

#Annotate figure
axis.bar(data.index,data,width=0.5,color=color)
for i in data.index:
    axis.annotate(f"{data[i]}",xy=(i,data[i]+50),va='center',ha='center',fontweight='light',fontfamily='Times N

#Add text
figure.text(0.125,0.9,'Top 8 countries on Netflix by amount content',fontsize=20,fontweight='bold',fontfamily='
```

Out[11]:

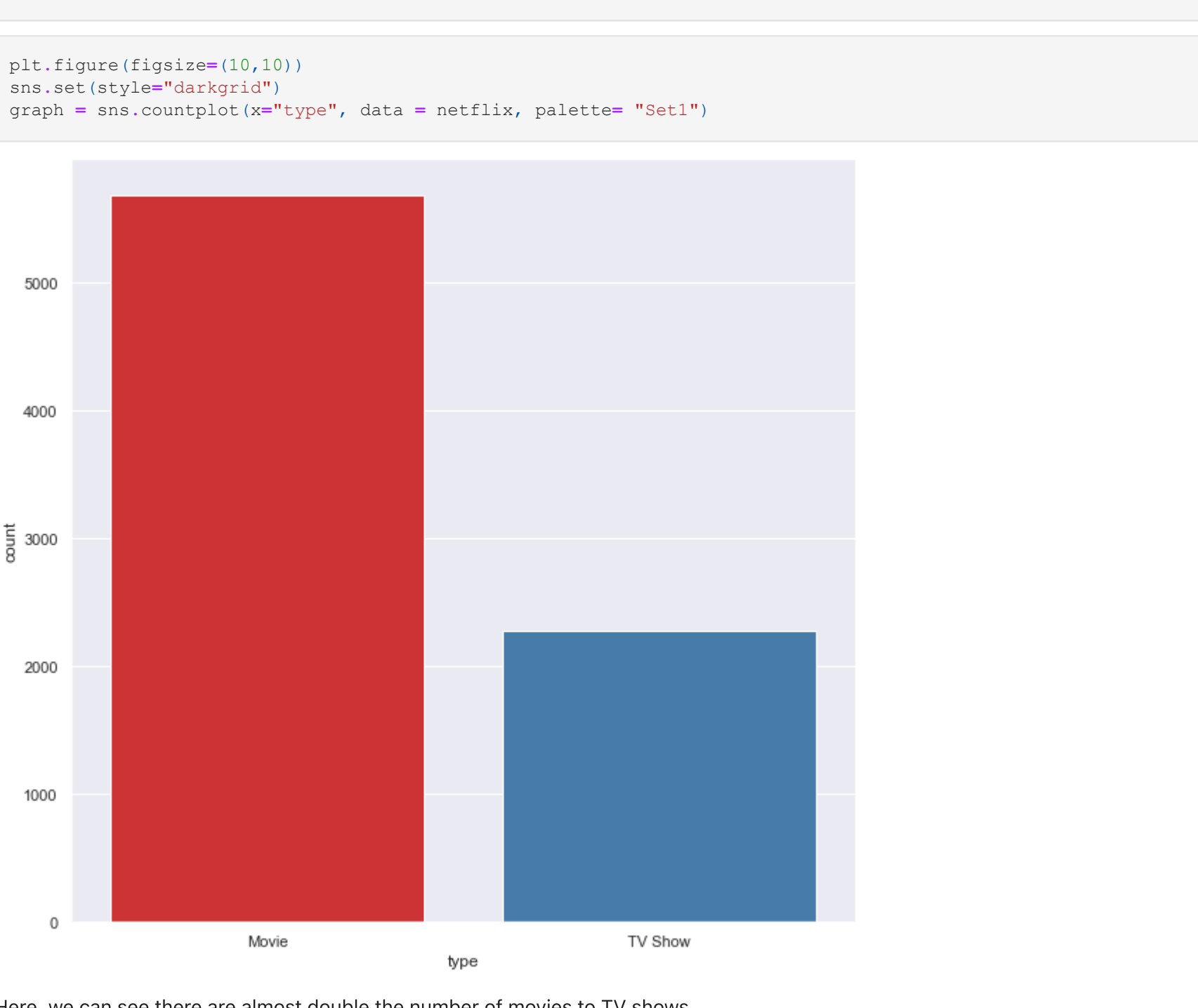
```
Text(0.125, 0.9, 'Top 8 countries on Netflix by amount content')
```

Most of the content are made in the US, this could be due to the fact that Netflix is from the Us.

4.3.1 Movie ratings analysis

In [12]:

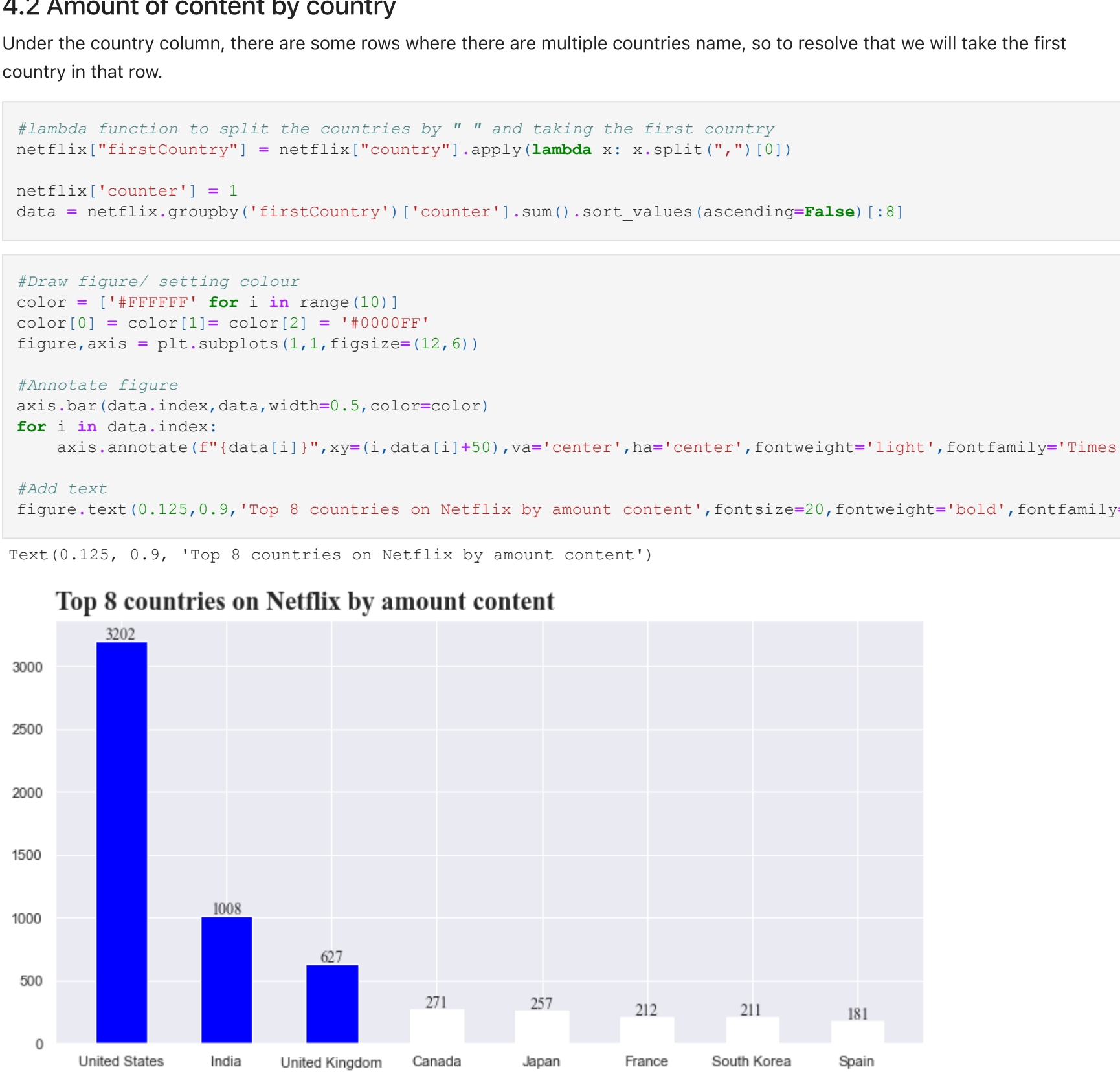
```
plt.figure(figsize=(15,10))
graph = sns.countplot(x="rating", data=netflixMovies, order=netflixMovies['rating'].value_counts().index)
```



4.3.2 TV show ratings analysis

In [13]:

```
plt.figure(figsize=(15,10))
graph = sns.countplot(x="rating", data=netflixShows, order=netflixShows['rating'].value_counts().index)
```

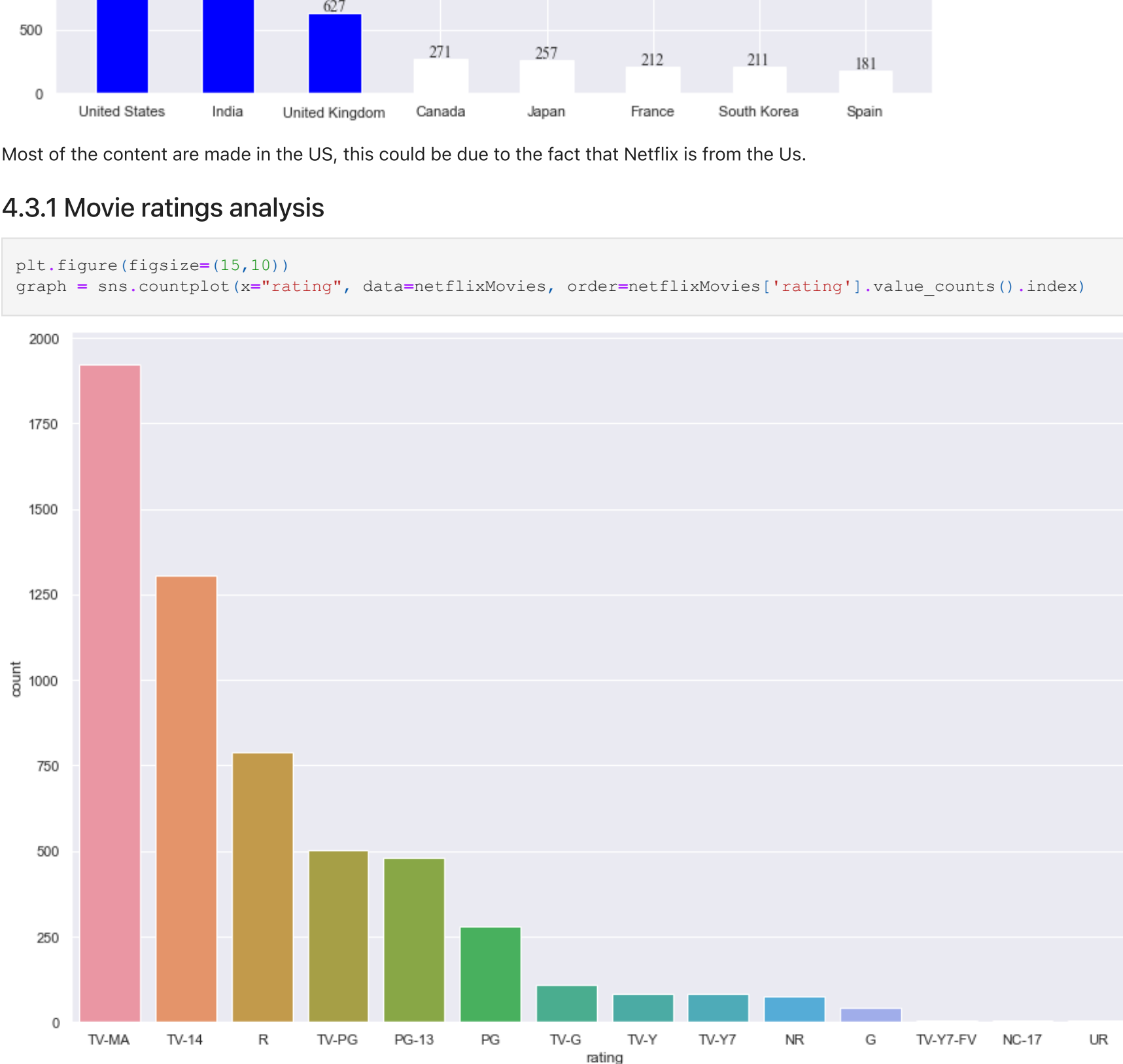


Both movies and TV shows have the most mature audience ratings. This tells me that netflix content is more driven towards 17 and above

4.4.1 No. of Movie released

In [14]:

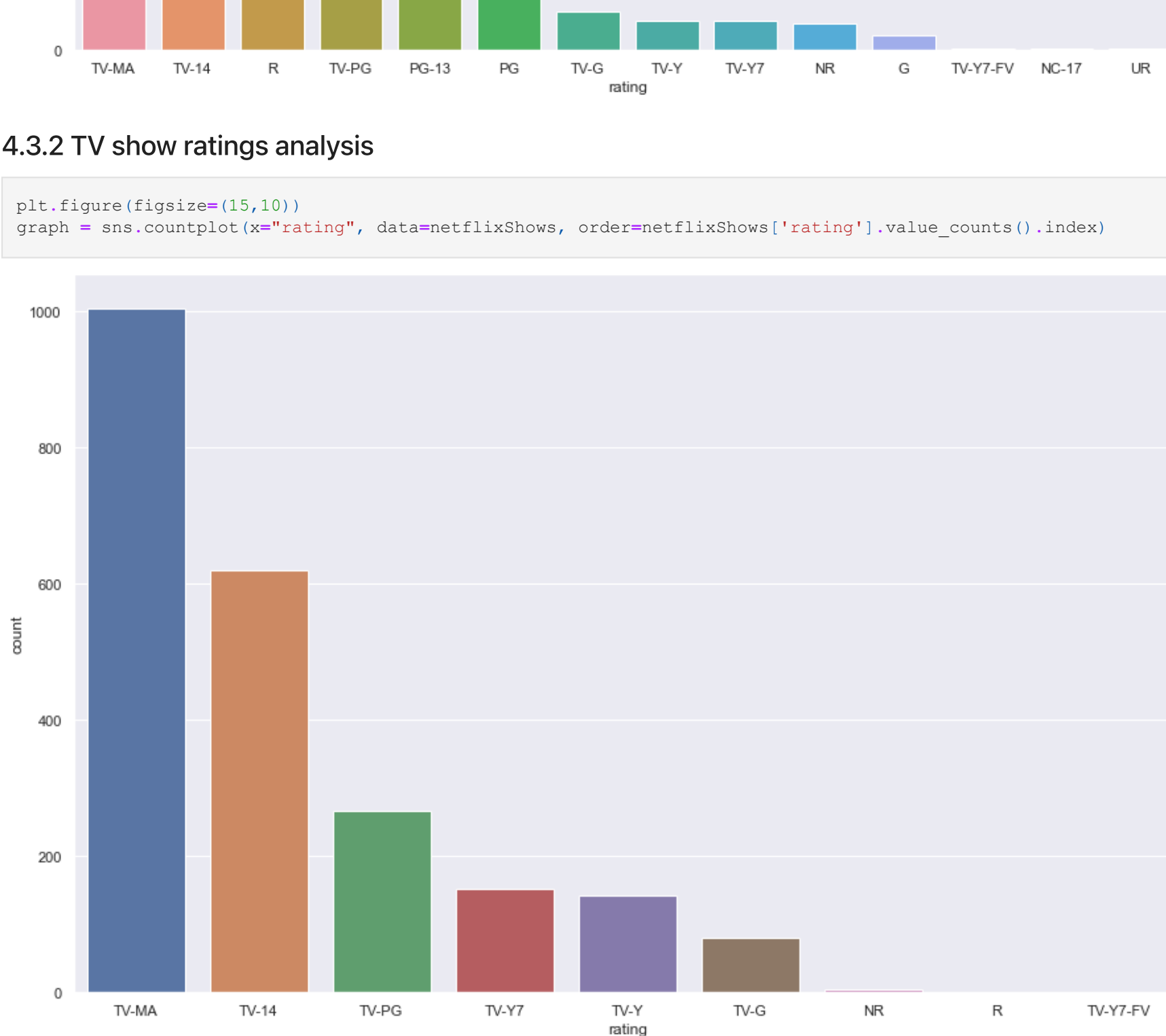
```
plt.figure(figsize=(15,10))
graph = sns.countplot(y="release_year", data=netflixMovies, palette="tab10", order=netflixMovies['release_year'
```



4.4.2 No. of TV shows released

In [15]:

```
plt.figure(figsize=(15,10))
graph = sns.countplot(y="release_year", data=netflixShows, palette="tab10", order=netflixShows['release_year'].
```



2017 was the year with most movies released and 2020 with the most TV shows. I can also assume those are the years Netflix produced the most amount of movies and TV shows respectively.

4.5 TV shows with longest season

In [16]:

```
#replace min in "duration" with " "
netflixMovies['duration']=netflixMovies['duration'].str.replace(' min','')

#convert string to integer
netflixMovies['duration']=netflixMovies['duration'].astype(str).astype(int)
```

```
/var/folders/63/48hcdx8x3_s9lh_x2yvfm0kc0000gn/T/ipykernel_4234/1815439773.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  netflixMovies['duration']=netflixMovies['duration'].str.replace(' min','')
/var/folders/63/48hcdx8x3_s9lh_x2yvfm0kc0000gn/T/ipykernel_4234/1815439773.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  netflixMovies['duration']=netflixMovies['duration'].astype(str).astype(int)
```

In [17]:

```
showLength = netflixShows[['title','duration']]
#replace Season and s in with " "
showLength['noOfSeasons'] = showLength['duration'].str.replace('Season','')
showLength['noOfSeasons'] = showLength['noOfSeasons'].str.replace('s','')

#convert string to integer
showLength['noOfSeasons'] = showLength['noOfSeasons'].astype(str).astype(int)
```

```
/var/folders/63/48hcdx8x3_s9lh_x2yvfm0kc0000gn/T/ipykernel_4234/4089118336.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  showLength['noOfSeasons'] = showLength['duration'].str.replace('Season','')
/var/folders/63/48hcdx8x3_s9lh_x2yvfm0kc0000gn/T/ipykernel_4234/4089118336.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  showLength['noOfSeasons'] = showLength['noOfSeasons'].str.replace('s','')
/var/folders/63/48hcdx8x3_s9lh_x2yvfm0kc0000gn/T/ipykernel_4234/4089118336.py:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ret
urning-a-view-versus-a-copy
  showLength['noOfSeasons'] = showLength['noOfSeasons'].astype(str).astype(int)
```

In [18]:

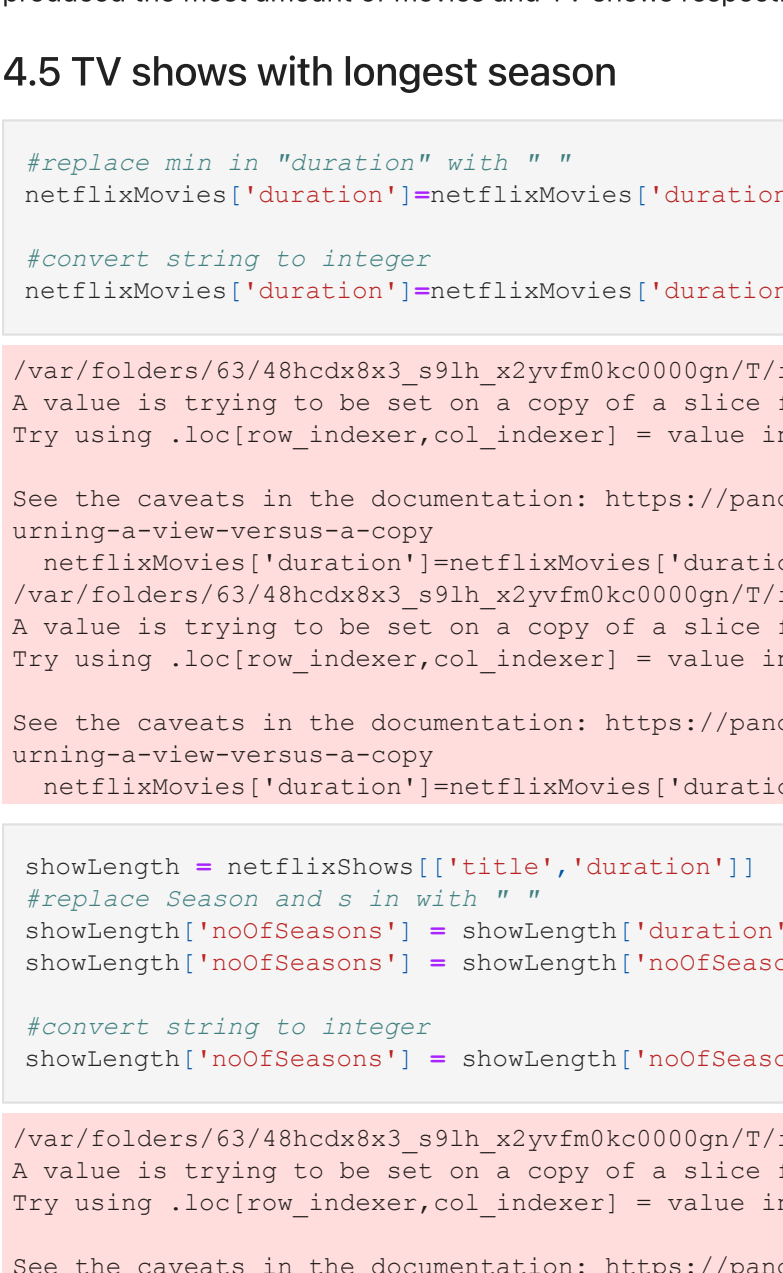
```
topShows = showLength[['title','noOfSeasons']]
topShows = topShows.sort_values(by='noOfSeasons', ascending=False)
```

In [19]:

```
first10 = topShows[0:10]
plt.figure(figsize = (15,10))
first10.plot(kind ='bar',x ='title',y = 'noOfSeasons', color='blue')
```

Out[19]:

```
<AxesSubplot:xlabel='title'>
<Figure size 1080x720 with 0 Axes>
```



The bar graph shows the top 10 TV shows with the longest season with each show having at least 10 seasons.

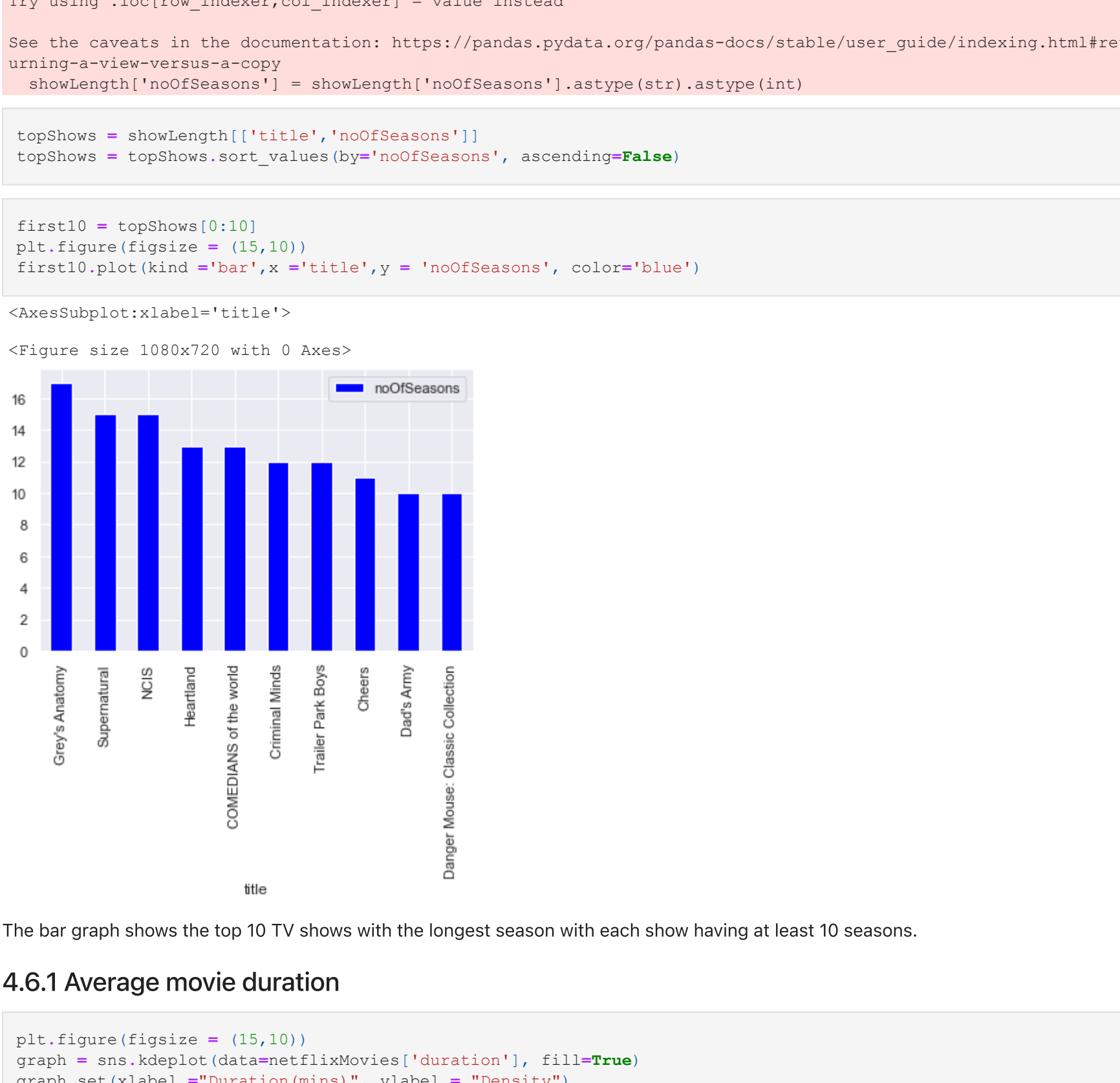
4.6.1 Average movie duration

In [20]:

```
plt.figure(figsize = (15,10))
graph = sns.kdeplot(data=netflixMovies['duration'], fill=True)
graph.set(xlabel ="Duration(mins)", ylabel = "Density")
```

Out[20]:

```
[Text(0.5, 0, 'Duration(mins)'), Text(0, 0.5, 'Density')]
```



4.6.2 Average TV shows duration

In [21]:

```
plt.figure(figsize = (15,10))
graph = sns.kdeplot(data = showLength['noOfSeasons'], fill=True)
graph.set(xlabel ="Seasons", ylabel = "Density")
```

Out[21]:

```
[Text(0.5, 0, 'Seasons'), Text(0, 0.5, 'Density')]
```

Based on the kernel density estimate plot, an average movie length lasts around 90-100 mins and an average TV Show last 1 season. I can infer that only Tv shows with high ratings will have more than 1 season which is rare.

5. Conclusion

Based on our data analysis, I was able to infer the relationships and trends based on the data. If I were given more data such as content ratings, such as 1 -5 stars I would be able to delve in deeper and get a clearer analysis.

6. References

seaborn

- https://seaborn.pydata.org/api.html
- NeuralNine (17 Apr 2022) Seaborn Is The Easier Matplotlib [online] Avialable: https://www.youtube.com/watch?v=ooqXQ37XHMM
- Kimberly Fessel (30 Jun 2020) What is kernel density estimation? And how to build a KDE plot in Python? | Seaborn KDEplot [online] Avialable: https://www.youtube.com/watch?v=DCgPRalDYXA&t=384s

Matplotlib

- https://matplotlib.org/stable/api/index

plotly

- https://plotly.com/python-api-reference/generated/plotly.graph_objects.Figure.html

netflix data set

- SHIVAM BANSAL (Oct 2021) Netflix Movies and TV Shows [online] Avialable: https://www.kaggle.com/datasets/shivamb/netflix-shows

In [ ]: