Netflix Dataset Analysis

This Netflix dataset has information about the TV Shows and Movies on Netflix till 2021. This dataset is collected from flixable which is a third-party Netflix search engine and available on Kaggle website for free.

In [4]:
```
import pandas as pd
import os
```

In [5]:
```
file_path = r"D:\Shikha files\Data Analytics\netflix_titles_nov_2019.csv\netfl
df = pd.read_csv(file_path)
df.head()
```

Out[5]:

	show_id	title	director	cast	country	date_added	release_year	rating	durati
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	Seas
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	November 30, 2019	2019	TV-G	67 r
2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	November 30, 2019	2019	TV-14	135 r
3	81082007	Atlantics	Mati Diop	Mama Sane, Amadou Mbow, Ibrahima Traore, Nicol...	France, Senegal, Belgium	November 29, 2019	2019	TV-14	106 r
4	80213643	Chip and Potato	NaN	Abigail Oliver, Andrea Libman, Briana Buckmast...	Canada, United Kingdom	NaN	2019	TV-Y	Seasc

In [6]:
```
df.tail()
```

Out[6]:

	show_id	title	director	cast	country	date_added	release_year	rating	du
5832	70141644	Mad Ron's Prevues from Hell	Jim Monaco	Nick Pawlow, Jordu Schell, Jay Kushwara, Micha...	United States	November 1, 2010	1987	NR	8
5833	70127998	Splatter	Joe Dante	Corey Feldman, Tony Todd, Tara Leigh, Erin Way...	United States	November 18, 2009	2009	TV-14	2
5834	70084180	Just Another Love Story	Ole Bornedal	Anders W. Berthelsen, Rebecka Hemse, Nikolaj L...	Denmark	May 5, 2009	2007	NR	10
5835	70157452	Dinner for Five	NaN	NaN	United States	February 4, 2008	2007	TV-MA	S
5836	70053412	To and From New York	Sorin Dan Mihalcescu	Barbara King, Shaana Diya, John Krisiukenas, Y...	United States	January 1, 2008	2006	NR	8

In [7]:
```
df.shape  # to show no of rows and columns
```

Out[7]:
```
(5837, 12)
```

In [8]:
```
df.size  # to show no. of total values(elements) in the dataset
```

Out[8]:
```
70044
```

In [9]:
```
df.columns
```

Out[9]:
```
Index(['show_id', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description',
       'type'],
      dtype='object')
```

In [10]:
```
df.dtypes
```

Out[10]:
```
show_id          int64
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
duration        object
listed_in       object
description     object
type            object
dtype: object
```

In [11]:
```
df.info()  # to show indexes, columns,types of each column,memory at once
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5837 entries, 0 to 5836
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       5837 non-null   int64 
 1   title         5837 non-null   object
 2   director      3936 non-null   object
 3   cast          5281 non-null   object
 4   country       5410 non-null   object
 5   date_added    5195 non-null   object
 6   release_year  5837 non-null   int64 
 7   rating        5827 non-null   object
 8   duration      5837 non-null   object
 9   listed_in     5837 non-null   object
 10  description   5837 non-null   object
 11  type          5837 non-null   object
dtypes: int64(2), object(10)
memory usage: 547.3+ KB
```

Task 1. Is there any Duplicate Record in this dataset? If yes, then remove the duplicate records.

duplicate()

In [12]:
```
df.shape
```

Out[12]:
```
(5837, 12)
```

In [13]:
```
df.duplicated()  # To check row wise and detect the duplicate rows
```

Out[13]:
```
0       False
1       False
2       False
3       False
4       False
        ...
5832    False
5833    False
5834    False
5835    False
5836    False
Length: 5837, dtype: bool
```

In [14]:
```
df[df.duplicated()]   # To check row wise and detect the duplicate rows.
```

Out[14]:

show_id	title	director	cast	country	date_added	release_year	rating	duration	listed_in	de

In [15]:
```
# data.drop_duplicates(inplace = True)  To remove the duplicate rows permanent
```

Task 2. Is there any Null Value present in any column ? Show with Heat-map.

isnull()

In [16]:
```
 # To show where Null value is present
```

In [17]:
```
df.head()
```

Out[17]:

	show_id	title	director	cast	country	date_added	release_year	rating	durati
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	Seas
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	November 30, 2019	2019	TV-G	67 r
2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	November 30, 2019	2019	TV-14	135 r
3	81082007	Atlantics	Mati Diop	Mama Sane, Amadou Mbow, Ibrahima Traore, Nicol...	France, Senegal, Belgium	November 29, 2019	2019	TV-14	106 r
4	80213643	Chip and Potato	NaN	Abigail Oliver, Andrea Libman, Briana Buckmast...	Canada, United Kingdom	NaN	2019	TV-Y	Seasc

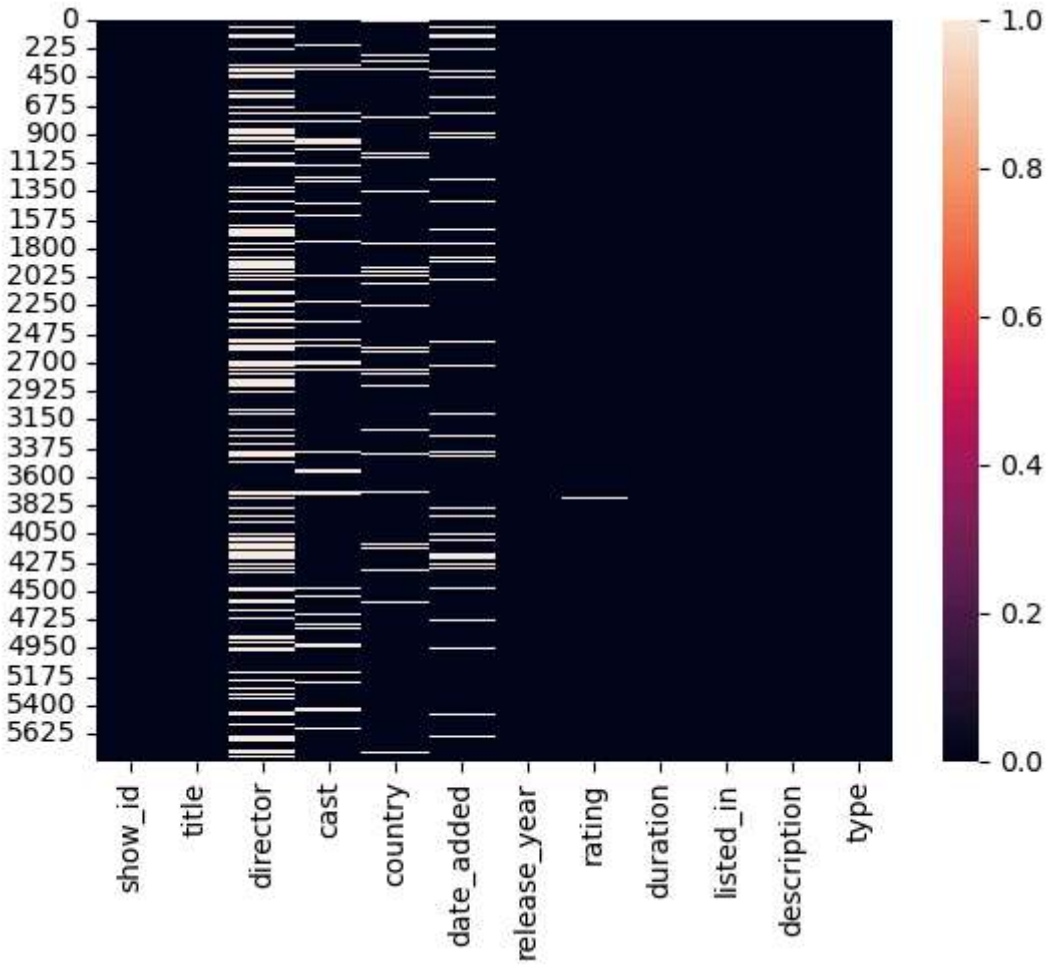
In [18]:
```
df.isnull().sum()
```

Out[18]:
```
show_id          0
title            0
director      1901
cast           556
country        427
date_added     642
release_year     0
rating          10
duration         0
listed_in        0
description      0
type             0
dtype: int64
```

In [19]:
```
import seaborn as sns
```

In [20]:
```
sns.heatmap(df.isnull())
```

Out[20]: <Axes: >



1. For 'House of Cards', what is the Show Id and Who is the Director of this show?

In [23]:
```
df.head(2)
```

Out[23]:

	show_id	title	director	cast	country	date_added	release_year	rating	duration
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	1 Season
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	November 30, 2019	2019	TV-G	67 min

In [21]:
```
# Using two methods. isin(),
```

In [22]:
```
df.title
```

Out[22]:
```
0                                  Chocolate
1       Guatemala: Heart of the Mayan World
2                           The Zoya Factor
3                                 Atlantics
4                           Chip and Potato
                        ...
5832            Mad Ron's Prevues from Hell
5833                             Splatter
5834                Just Another Love Story
5835                        Dinner for Five
5836                   To and From New York
Name: title, Length: 5837, dtype: object
```

In [23]:
```
df[df['title'].isin(['House of Cards'])] # To show all records of a particular
```

Out[23]:

	show_id	title	director	cast	country	date_added	release_year	rating	duration	li
2435	70178217	House of Cards	NaN	Kevin Spacey, Robin Wright, Kate Mara, Corey S...	United States	NaN	2018	TV-MA	6 Seasons	[

str.contains()

In [24]:
```
# To show all records of a particular string in any column
```

In [25]:
```
df[df['title'].str.contains('House of Cards')]
```

Out[25]:

	show_id	title	director	cast	country	date_added	release_year	rating	duration	li
2435	70178217	House of Cards	NaN	Kevin Spacey, Robin Wright, Kate Mara, Corey S...	United States	NaN	2018	TV-MA	6 Seasons	[

2. In which year the highest number of the TV Shows and Movies were released ? Show with Bar Graph.

In [26]:
```
df.dtypes
```

Out[26]:
```
show_id          int64
title           object
director        object
cast            object
country         object
date_added      object
release_year     int64
rating          object
duration        object
listed_in       object
description     object
type            object
dtype: object
```

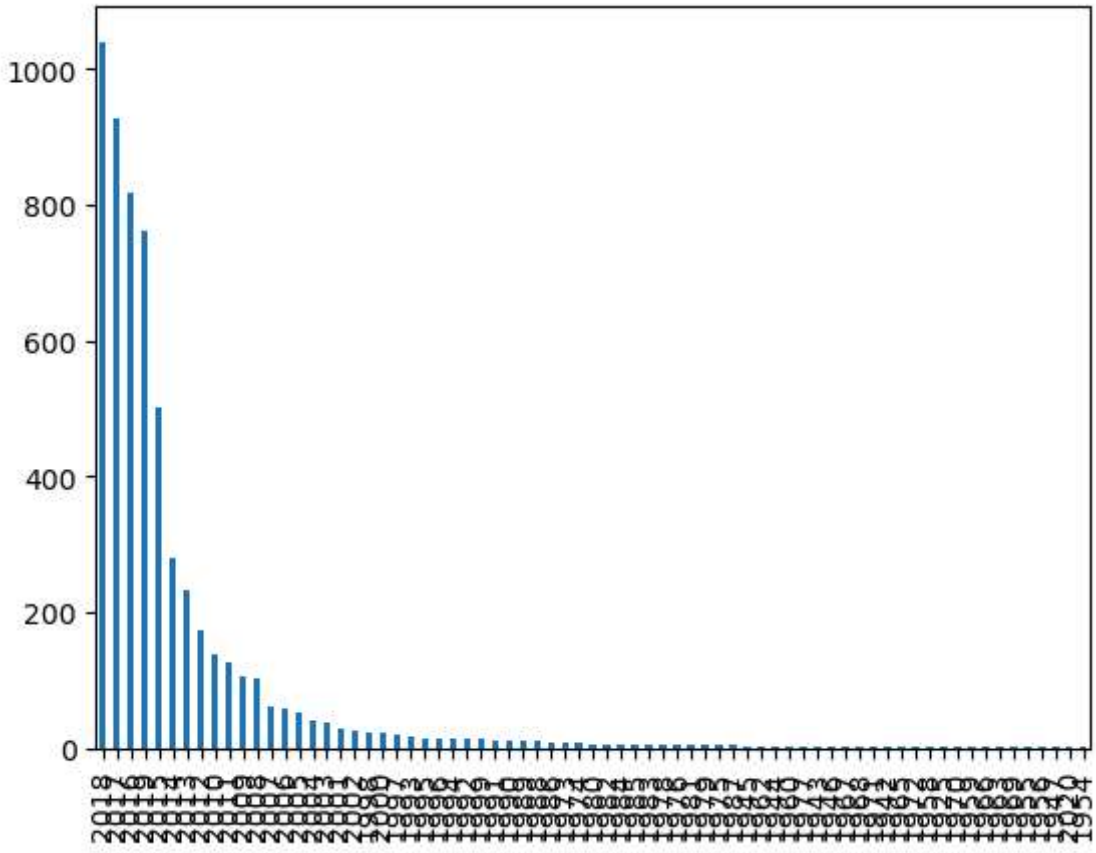
In [27]:
```
df.release_year.value_counts() # It counts the occurrence of all individual Ye
```

Out[27]:
```
2018    1040
2017     928
2016     818
2019     762
2015     502
        ...
1955       1
1956       1
1947       1
2020       1
1954       1
Name: release_year, Length: 71, dtype: int64
```

Bar Graph

In [29]:
```
df[df.release_year.value_counts().plot(kind='bar')]
```

```
---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call last)
File ~\anaconda3\Lib\site-packages\pandas\core\indexes\base.py:3802, in Inde
x.get_loc(self, key, method, tolerance)
   3801 try:
-> 3802     return self._engine.get_loc(casted_key)
   3803 except KeyError as err:

File ~\anaconda3\Lib\site-packages\pandas\_libs\index.pyx:138, in pandas._lib
s.index.IndexEngine.get_loc()

File ~\anaconda3\Lib\site-packages\pandas\_libs\index.pyx:165, in pandas._lib
s.index.IndexEngine.get_loc()

File pandas\_libs\hashtable_class_helper.pxi:5745, in pandas._libs.hashtable.
PyObjectHashTable.get_item()

File pandas\_libs\hashtable_class_helper.pxi:5753, in pandas._libs.hashtable.
PyObjectHashTable.get_item()

KeyError: <Axes: >

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call last)
Cell In[29], line 1
----> 1 df[df.release_year.value_counts().plot(kind='bar')]

File ~\anaconda3\Lib\site-packages\pandas\core\frame.py:3807, in DataFrame.__
getitem__(self, key)
   3805 if self.columns.nlevels > 1:
   3806     return self._getitem_multilevel(key)
-> 3807 indexer = self.columns.get_loc(key)
   3808 if is_integer(indexer):
   3809     indexer = [indexer]

File ~\anaconda3\Lib\site-packages\pandas\core\indexes\base.py:3804, in Inde
x.get_loc(self, key, method, tolerance)
   3802     return self._engine.get_loc(casted_key)
   3803 except KeyError as err:
-> 3804     raise KeyError(key) from err
   3805 except TypeError:
   3806     # If we have a listlike key, _check_indexing_error will raise
   3807     #  InvalidIndexError. Otherwise we fall through and re-raise
   3808     #  the TypeError.
   3809     self._check_indexing_error(key)

KeyError: <Axes: >
```



3. How many Movies and TV Shows are in the dataset? Show with Bar Graph.

groupby()

In [30]:
```
df.head(2)
```

Out[30]:

	show_id	title	director	cast	country	date_added	release_year	rating	duration
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	1 Season
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	November 30, 2019	2019	TV-G	67 min

In [31]:
```
df.groupby('type').type.count()
```

Out[31]:
```
type
Movie      3939
TV Show    1898
Name: type, dtype: int64
```

countplot()

In [32]:
```
import seaborn as sns
sns.countplot(df['type'])
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[32], line 2
      1 import seaborn as sns
----> 2 sns.countplot(df['type'])

File ~\anaconda3\Lib\site-packages\seaborn\categorical.py:2943, in countplot
(data, x, y, hue, order, hue_order, orient, color, palette, saturation, widt
h, dodge, ax, **kwargs)
   2940 elif x is not None and y is not None:
   2941     raise ValueError("Cannot pass values for both `x` and `y`")
-> 2943 plotter = _CountPlotter(
   2944     x, y, hue, data, order, hue_order,
   2945     estimator, errorbar, n_boot, units, seed,
   2946     orient, color, palette, saturation,
   2947     width, errcolor, errwidth, capsize, dodge
   2948 )
   2950 plotter.value_label = "count"
   2952 if ax is None:

File ~\anaconda3\Lib\site-packages\seaborn\categorical.py:1530, in _BarPlotte
r.__init__(self, x, y, hue, data, order, hue_order, estimator, errorbar, n_bo
ot, units, seed, orient, color, palette, saturation, width, errcolor, errwidt
h, capsize, dodge)
   1525 def __init__(self, x, y, hue, data, order, hue_order,
   1526              estimator, errorbar, n_boot, units, seed,
   1527              orient, color, palette, saturation, width,
   1528              errcolor, errwidth, capsize, dodge):
   1529     """Initialize the plotter."""
-> 1530     self.establish_variables(x, y, hue, data, orient,
   1531                              order, hue_order, units)
   1532     self.establish_colors(color, palette, saturation)
   1533     self.estimate_statistic(estimator, errorbar, n_boot, seed)

File ~\anaconda3\Lib\site-packages\seaborn\categorical.py:516, in _Categorica
lPlotter.establish_variables(self, x, y, hue, data, orient, order, hue_order,
units)
    513     plot_data = data
    515 # Convert to a list of arrays, the common representation
--> 516 plot_data = [np.asarray(d, float) for d in plot_data]
    518 # The group names will just be numeric indices
    519 group_names = list(range(len(plot_data)))

File ~\anaconda3\Lib\site-packages\seaborn\categorical.py:516, in <listcomp>
(.0)
    513     plot_data = data
    515 # Convert to a list of arrays, the common representation
--> 516 plot_data = [np.asarray(d, float) for d in plot_data]
    518 # The group names will just be numeric indices
    519 group_names = list(range(len(plot_data)))

File ~\anaconda3\Lib\site-packages\pandas\core\series.py:893, in Series.__arr
ay__(self, dtype)
    846 def __array__(self, dtype: npt.DTypeLike | None = None) -> np.ndarra
y:
    847     """
    848     Return the values as a NumPy array.
```

```
    849
   (...)
    891           dtype='datetime64[ns]')
    892     """
--> 893     return np.asarray(self._values, dtype)

ValueError: could not convert string to float: 'TV Show'
```

4. Show all the Movies that were released in year 2000.

In [33]:
```
df.head()
```

Out[33]:

	show_id	title	director	cast	country	date_added	release_year	rating	durati
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	Seas
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	November 30, 2019	2019	TV-G	67 r
2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	November 30, 2019	2019	TV-14	135 r
3	81082007	Atlantics	Mati Diop	Mama Sane, Amadou Mbow, Ibrahima Traore, Nicol...	France, Senegal, Belgium	November 29, 2019	2019	TV-14	106 r
4	80213643	Chip and Potato	NaN	Abigail Oliver, Andrea Libman, Briana Buckmast...	Canada, United Kingdom	NaN	2019	TV-Y	Seasc

In [34]:
```
df[(df['type'] == 'Movie') & (df['release_year'] == 2000)]  # Filtering
```

Out[34]:

	show_id	title	director	cast	country	date_added	release_year	rating	d
85	60003508	The Gift	Sam Raimi	Cate Blanchett, Giovanni Ribisi, Keanu Reeves,...	United States	November 20, 2019	2000	R	
452	60003242	Charlie's Angels	McG	Cameron Diaz, Drew Barrymore, Lucy Liu, Bill M...	United States, Germany	October 1, 2019	2000	PG-13	
486	60003540	Rugrats in Paris: The Movie	Stig Bergqvist, Paul Demeyer	Elizabeth Daily, Tara Strong, Cheryl Chase, Ch...	Germany, United States	October 1, 2019	2000	G	
500	60000921	The Flintstones in Viva Rock Vegas	Brian Levant	Mark Addy, Stephen Baldwin, Kristen Johnston, ...	United States	October 1, 2019	2000	PG	
640	60000861	American Psycho	Mary Harron	Christian Bale, Willem Dafoe, Jared Leto, Rees...	United States, Canada	September 1, 2019	2000	R	
836	60001363	Space Cowboys	Clint Eastwood	Clint Eastwood, Tommy Lee Jones, Donald Suther...	United States	August 1, 2019	2000	PG-13	
1009	60000415	Scream 3	Wes Craven	David Arquette, Neve Campbell, Courteney Cox, ...	United States	July 1, 2019	2000	R	
1754	60002907	Crouching Tiger, Hidden Dragon	Ang Lee	Chow Yun Fat, Michelle Yeoh, Zhang Ziyi, Chang...	Taiwan, Hong Kong, United States, China	March 1, 2019	2000	PG-13	
1913	60003378	Billy Elliot	Stephen Daldry	Jamie Bell, Gary Lewis, Stuart Wells, Billy Fa...	United Kingdom, France	February 1, 2019	2000	R	
2618	70118859	Monty Python: Before the Flying Circus	Will Yapp	Graham Chapman, Eric Idle, Terry Jones, Michae...	United Kingdom	October 2, 2018	2000	TV-14	

	show_id	title	director	cast	country	date_added	release_year	rating	d
2959	60003290	Fiza	Khalid Mohamed	Karisma Kapoor, Jaya Bhaduri, Hrithik Roshan, ...	India	August 2, 2018	2000	TV-14	
3461	60033787	An American Tail: The Treasures of Manhattan I...	Larry Latham	Thomas Dekker, Dom DeLuise, Pat Musick, Nehemi...	United States	April 1, 2018	2000	G	
3482	60002111	Kya Kehna	Kundan Shah	Preity Zinta, Saif Ali Khan, Anupam Kher, Fari...	India	April 1, 2018	2000	TV-PG	
3491	60000722	Phir Bhi Dil Hai Hindustani	Aziz Mirza	Shah Rukh Khan, Juhi Chawla, Paresh Rawal, Sat...	India	April 1, 2018	2000	TV-PG	
3623	60003401	Hamara Dil Aapke Paas Hai	Satish Kaushik	Anil Kapoor, Aishwarya Rai Bachchan, Sonali Be...	India	March 1, 2018	2000	TV-14	
3649	60037397	Pukar	Rajkumar Santoshi	Anil Kapoor, Madhuri Dixit, Namrata Shirodkar,...	India	March 1, 2018	2000	TV-14	
4010	60002808	Papa the Great	Bhagyaraj	Krishan Kumar, Nagma, Satya Prakash, Master Bo...	India	December 8, 2017	2000	TV-PG	
4675	60000901	How the Grinch Stole Christmas	Ron Howard	Jim Carrey, Taylor Momsen, Jeffrey Tambor, Chr...	United States	June 1, 2017	2000	PG	
5830	60003155	Joseph: King of Dreams	Rob LaDuca, Robert C. Ramirez	Ben Affleck, Mark Hamill, Richard Herd, Mauree...	United States	September 27, 2011	2000	TV-PG	

5. Show only the Titles of all TV Shows that were released in India only.

In [35]:
```
df[(df['type'] == 'TV Show') & (df['country'] == 'India')]
```

Out[35]:

	show_id	title	director	cast	country	date_added	release_year	r
142	81011159	Little Things	NaN	Dhruv Sehgal, Mithila Palkar	India	NaN	2019	
361	81020066	Mighty Little Bheem: Diwali	NaN	Samriddhi Shukla, Nishka Raheja, Arannya Kaur,...	India	October 18, 2019	2019	
456	81113888	College Romance	NaN	Manjot Singh, Apoorva Arora, Keshav Sadhna, Sh...	India	October 1, 2019	2018	
		Engineering		Barkha Singh,		October 1		

6. Show top 10 director who gave the highest number of TV Shows and Movies to Netflix?

In [36]:
```
# data['director'].value_counts().head()
```

In [37]:
```
df.head(2)
```

Out[37]:

	show_id	title	director	cast	country	date_added	release_year	rating	duration
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	1 Season
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	November 30, 2019	2019	TV-G	67 min

In [38]:
```
df['director'].value_counts().head(10)
```

Out[38]:
```
Raúl Campos, Jan Suter    18
Marcus Raboy              14
Jay Karas                 13
Jay Chapman               12
Steven Spielberg           8
Johnnie To                 8
Martin Scorsese            8
Hakan Algül                7
S.S. Rajamouli             7
Ryan Polito                7
Name: director, dtype: int64
```

7. Show all the Records where "Category" is Movie and Type is "Comedies" of "Country" is United Kingdom".

Filtering (And, Or Operators)

In [43]:
```
df.head(5)
```

Out[43]:

	show_id	title	director	cast	country	date_added	release_year	rating	durati
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	Seas
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	November 30, 2019	2019	TV-G	67 r
2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	November 30, 2019	2019	TV-14	135 r
3	81082007	Atlantics	Mati Diop	Mama Sane, Amadou Mbow, Ibrahima Traore, Nicol...	France, Senegal, Belgium	November 29, 2019	2019	TV-14	106 r
4	80213643	Chip and Potato	NaN	Abigail Oliver, Andrea Libman, Briana Buckmast...	Canada, United Kingdom	NaN	2019	TV-Y	Seasc

In [41]:
```
(df['type'] == 'Movie')
```

Out[41]:
```
0       False
1        True
2        True
3        True
4       False
        ...
5832     True
5833     True
5834     True
5835    False
5836     True
Name: type, Length: 5837, dtype: bool
```

In [42]:
```
#data[(data['Category'] == 'Movie') & (data['type'] == 'Comedies')]
```

In [47]:
```
df[(df['type'] == 'Movie') & (df['listed_in'].str.contains('Comedies')) | (df[
```

Out[47]:

	show_id	title	director	cast	country	date_added	release_year	rating	du
2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	November 30, 2019	2019	TV-14	1
4	80213643	Chip and Potato	NaN	Abigail Oliver, Andrea Libman, Briana Buckmast...	Canada, United Kingdom	NaN	2019	TV-Y	Se
5	81172754	Crazy people	Moses Inwang	Ramsey Nouah, Chigul, Sola Sobowale, Ireti Doy...	Nigeria	November 29, 2019	2018	TV-14	1
9	81172841	Lagos Real Fake Life	Mike Ezuruonye	Nonso Diobi, Mike Ezuruonye, Mercy Aigbe, Rex ...	NaN	November 29, 2019	2018	TV-14	1
10	81172899	Payday	Cheta Chukwu	Baaj Adebule, Ebiye Victor, Meg Otanwa, Bisola...	Nigeria	November 29, 2019	2018	TV-MA	1
...	...	...	...	...	...	...	...	...	
5800	70253398	Shrek the Musical	Jason Moore	Brian d'Arcy James, Daniel Breaker, Sutton Fos...	United States	December 29, 2013	2013	TV-G	1
5808	70157231	The 4400	NaN	Joel Gretsch, Jacqueline McKenzie, Patrick Joh...	United States, United Kingdom	NaN	2007	TV-14	Se
5812	70253397	Kung Fu Panda: Holiday	Tim Johnson	Jack Black, Angelina Jolie, Dustin Hoffman, Ja...	United States	December 1, 2012	2010	TV-PG	
5813	70213466	Casa de mi Padre	Matt Piedmont	Will Ferrell, Gael García Bernal, Diego Luna, ...	United States, Mexico	November 14, 2012	2012	R	

	show_id	title	director	cast	country	date_added	release_year	rating	du
5817	60003082	White Christmas	Michael Curtiz	Bing Crosby, Danny Kaye, Rosemary Clooney, Ver...	United States	July 15, 2012	1954	TV-G	1

1498 rows × 12 columns

8. In how many movies/shows, Tom Cruise was cast ?

In [51]:
```
#data.head()
# data[data['cast'] == 'Tom Cruise']  filtering
# dat[data['cast'].str.contains('Tom Cruise')]
```

In [54]:
```
df.head()
```

Out[54]:

	show_id	title	director	cast	country	date_added	release_year	rating	durati
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	Seas
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	November 30, 2019	2019	TV-G	67 r
2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	November 30, 2019	2019	TV-14	135 r
3	81082007	Atlantics	Mati Diop	Mama Sane, Amadou Mbow, Ibrahima Traore, Nicol...	France, Senegal, Belgium	November 29, 2019	2019	TV-14	106 r
4	80213643	Chip and Potato	NaN	Abigail Oliver, Andrea Libman, Briana Buckmast...	Canada, United Kingdom	NaN	2019	TV-Y	Seasc

In [56]:
```
df[df['cast'] == ('Tom Cruise')]
```

Out[56]:

show_id	title	director	cast	country	date_added	release_year	rating	duration	listed_in	de

In [57]:
```
df[df['cast'].str.contains('Tom Cruise')]
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[57], line 1
----> 1 df[df['cast'].str.contains('Tom Cruise')]

File ~\anaconda3\Lib\site-packages\pandas\core\frame.py:3797, in DataFrame.__
getitem__(self, key)
   3794     return self.where(key)
   3796 # Do we have a (boolean) 1d indexer?
-> 3797 if com.is_bool_indexer(key):
   3798     return self._getitem_bool_array(key)
   3800 # We are left with two options: a single key, and a collection of key
s,
   3801 # We interpret tuples as collections only for non-MultiIndex

File ~\anaconda3\Lib\site-packages\pandas\core\common.py:135, in is_bool_inde
xer(key)
    131     na_msg = "Cannot mask with non-boolean array containing NA / NaN
values"
    132     if lib.infer_dtype(key_array) == "boolean" and isna(key_array).an
y():
    133         # Don't raise on e.g. ["A", "B", np.nan], see
    134         #  test_loc_getitem_list_of_labels_categoricalindex_with_na
--> 135         raise ValueError(na_msg)
    136     return False
    137 return True

ValueError: Cannot mask with non-boolean array containing NA / NaN values
```

Creating new data-frame

In [58]:
```
# data_new = data.dropna()   # It drops the rows that contains all or any miss
```

In [59]:
```
# data_new.head(2)
```

In [60]:
```
# data_new[data_new['cast'].str.contains('Tom Cruise')]
```

In [64]:
```
df_new = df.dropna()
```

In [65]:
```
df_new.head(2)
```

Out[65]:

	show_id	title	director	cast	country	date_added	release_year	rating	duration
2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	November 30, 2019	2019	TV-14	135 min
3	81082007	Atlantics	Mati Diop	Mama Sane, Amadou Mbow, Ibrahima Traore, Nicol...	France, Senegal, Belgium	November 29, 2019	2019	TV-14	106 min

In [66]:
```
# df_new[df_new['cast'].str.contains('Tom Cruise')]
```

In [67]:
```
df_new[df_new['cast'].str.contains('Tom Cruise')]
```

Out[67]:

	show_id	title	director	cast	country	date_added	release_year	rating	duration	li
1006	60029369	Rain Man	Barry Levinson	Dustin Hoffman, Tom Cruise, Valeria Golino, Ge...	United States	July 1, 2019	1988	R	134 min	

9. What are the different Ratings defined by Netflix ?

In [68]:
```
# nunique = returns the no. of unique values for each column
```

In [69]:
```
df['rating'].nunique()
```

Out[69]:
```
14
```

unique

In [71]:
```
df['rating'].unique()
```

Out[71]:
```
array(['TV-14', 'TV-G', 'TV-Y', 'TV-MA', 'TV-PG', 'R', 'TV-Y7', 'PG', 'G',
       'PG-13', 'TV-Y7-FV', 'NR', 'UR', 'NC-17', nan], dtype=object)
```

9.1 How many Movies got the 'TV-14' rating in Canada ?

In [73]:
```
df[(df['type'] == 'Movie') & (df['rating'] == 'TV-14')].shape
```

Out[73]: (955, 12)

In [74]:
```
df[(df['type'] == 'Movie') & (df['rating'] == "TV-14") & (df['country'] == 'Ca
```

Out[74]: (13, 12)

9.2 How many TV Show got the 'R' rating, after year 2018 ?

In [75]:
```
df[(df['type'] == 'TV Show') & (df['rating'] == 'R')].shape
```

Out[75]: (2, 12)

In [76]:
```
# data[(data['category'] == 'TV show') & (data['rating'] == 'R') & (data['Year
```

In [77]:
```
df[(df['type'] == 'TV Show') & (df['rating'] == 'R') & (df['release_year'] > 2
```

Out[77]:

show_id	title	director	cast	country	date_added	release_year	rating	duration	listed_in	de

10. What is the maximum duration of a Movie/Show on Netflix ?

In [79]:
```
# data.head()
# data['Duration'].unique()
# data.Duration.dtypes
```

In [80]:
```
df.head()
```

Out[80]:

	show_id	title	director	cast	country	date_added	release_year	rating	durati
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	Seas
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	November 30, 2019	2019	TV-G	67 r
2	81213894	The Zoya Factor	Abhishek Sharma	Sonam Kapoor, Dulquer Salmaan, Sanjay Kapoor, ...	India	November 30, 2019	2019	TV-14	135 r
3	81082007	Atlantics	Mati Diop	Mama Sane, Amadou Mbow, Ibrahima Traore, Nicol...	France, Senegal, Belgium	November 29, 2019	2019	TV-14	106 r
4	80213643	Chip and Potato	NaN	Abigail Oliver, Andrea Libman, Briana Buckmast...	Canada, United Kingdom	NaN	2019	TV-Y	Seasc

In [82]:
```
df['duration'].unique()
```

Out[82]:
```
array(['1 Season', '67 min', '135 min', '106 min', '2 Seasons', '107 min',
       '81 min', '118 min', '110 min', '104 min', '93 min', '94 min',
       '124 min', '137 min', '134 min', '69 min', '209 min', '86 min',
       '24 min', '46 min', '117 min', '87 min', '92 min', '114 min',
       '121 min', '109 min', '96 min', '97 min', '56 min', '119 min',
       '3 Seasons', '8 Seasons', '138 min', '111 min', '88 min', '73 min',
       '116 min', '85 min', '102 min', '101 min', '28 min', '103 min',
       '131 min', '166 min', '105 min', '82 min', '84 min', '112 min',
       '89 min', '136 min', '129 min', '158 min', '78 min', '100 min',
       '74 min', '60 min', '143 min', '98 min', '54 min', '59 min',
       '95 min', '5 Seasons', '61 min', '4 Seasons', '123 min', '44 min',
       '68 min', '99 min', '91 min', '40 min', '90 min', '108 min',
       '200 min', '133 min', '115 min', '55 min', '153 min', '185 min',
       '127 min', '120 min', '139 min', '122 min', '36 min', '141 min',
       '65 min', '126 min', '63 min', '14 min', '20 min', '6 Seasons',
       '52 min', '83 min', '66 min', '140 min', '22 min', '62 min',
       '151 min', '154 min', '70 min', '76 min', '77 min', '45 min',
       '147 min', '58 min', '64 min', '125 min', '79 min', '163 min',
       '42 min', '38 min', '146 min', '130 min', '152 min', '182 min',
       '171 min', '80 min', '157 min', '7 Seasons', '9 Seasons',
       '142 min', '10 min', '128 min', '149 min', '113 min', '47 min',
       '167 min', '72 min', '145 min', '164 min', '11 Seasons', '177 min',
       '57 min', '161 min', '32 min', '53 min', '26 min', '48 min',
       '176 min', '15 min', '15 Seasons', '71 min', '12 min', '3 min',
       '30 min', '159 min', '150 min', '165 min', '14 Seasons', '148 min',
       '49 min', '168 min', '170 min', '132 min', '75 min', '162 min',
       '51 min', '50 min', '144 min', '13 Seasons', '34 min', '29 min',
       '312 min', '27 min', '35 min', '155 min', '23 min', '11 min',
       '205 min', '190 min', '25 min', '214 min', '156 min', '31 min',
       '196 min', '33 min', '203 min', '160 min', '19 min', '169 min',
       '173 min', '195 min', '10 Seasons', '12 Seasons', '41 min',
       '187 min', '193 min', '192 min', '224 min', '172 min', '18 min',
       '179 min', '37 min', '180 min', '43 min'], dtype=object)
```

In [84]:
```
df.duration.dtypes
```

Out[84]:
```
dtype('O')
```

str.split()

In [85]:
```
# data[['Minutes','Unit']] = data['duration'].str.split(' ', expand = True)
```

In [87]:
```
df.head(2)
```

Out[87]:

	show_id	title	director	cast	country	date_added	release_year	rating	duration
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	1 Season
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	November 30, 2019	2019	TV-G	67 min

In [89]:
```
df[['Minutes', 'Unit']] = df['duration'].str.split(' ', expand = True)
```

In [91]:
```
df.head(2)
```

Out[91]:

	show_id	title	director	cast	country	date_added	release_year	rating	duration
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	1 Season
1	81197050	Guatemala: Heart of the Mayan World	Luis Ara, Ignacio Jaunsolo	Christian Morales	NaN	November 30, 2019	2019	TV-G	67 min

max()

In [92]:
```
df.Minutes.max()
```

Out[92]:
```
'99'
```

In [93]:
```
df.Minutes.min()
```

Out[93]:
```
'1'
```

In [94]:
```
df.Minutes.mean()
```

Out[94]:
```
inf
```

11. Which individual country has the highest No. of TV shows?

In [95]:
```
# data_tvshow = data[data['Category'] == 'TV Show']

df_tvshow = df[df['type'] == 'TV Show']
```

In [96]:
```
df_tvshow.head(2)
```

Out[96]:

	show_id	title	director	cast	country	date_added	release_year	rating	duration
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	1 Season
4	80213643	Chip and Potato	NaN	Abigail Oliver, Andrea Libman, Briana Buckmast...	Canada, United Kingdom	NaN	2019	TV-Y	2 Seasons

In [97]:
```
# data_tvshow.Country.value_counts()

df_tvshow.country.value_counts()
```

Out[97]:
```
United States                                                 532
United Kingdom                                                174
Japan                                                         124
South Korea                                                   101
Taiwan                                                         65
                                                             ...
Italy, South Africa, West Germany, Australia, United States     1
Spain, United Kingdom                                           1
United States, Colombia                                         1
United Arab Emirates                                            1
United States, Ireland                                          1
Name: country, Length: 152, dtype: int64
```

In [98]:
```
df_tvshow.country.value_counts().head(1)
```

Out[98]:
```
United States    532
Name: country, dtype: int64
```

12. How can we sort the dataset by Year ?

In [99]:
```
# data.head()
# data.sort_values(by ='Year', ascending=False).head(2)
```

In [101]:
```
df.sort_values(by = 'release_year')
```

Out[101]:

	show_id	title	director	cast	country	date_added	release_year	rating	d
2154	81030762	Pioneers: First Women Filmmakers*	NaN	NaN	NaN	December 30, 2018	1925	TV-PG	
4929	60027945	Prelude to War	Frank Capra	NaN	United States	March 31, 2017	1942	TV-PG	
4931	60027942	The Battle of Midway	John Ford	Henry Fonda, Jane Darwell	United States	March 31, 2017	1942	TV-G	
4946	70022548	WWII: Report from the Aleutians	John Huston	NaN	United States	March 31, 2017	1943	NR	
4943	70013050	Why We Fight: The Battle of Russia	Frank Capra, Anatole Litvak	NaN	United States	March 31, 2017	1943	TV-14	
...	...	...	...	...	...	...	...	...	
1632	80220655	Stay Tuned!	NaN	Kyoko Yoshine, Hiroki Iijima, Kanako Miyashita...	Japan	March 21, 2019	2019	TV-14	
1630	81049949	Antoine Griezmann: The Making of a Legend	NaN	NaN	NaN	March 21, 2019	2019	TV-PG	
1628	80044950	The OA	NaN	Brit Marling, Jason Isaacs, Emory Cohen, Scott...	United States	NaN	2019	TV-MA	S
0	81193313	Chocolate	NaN	Ha Ji-won, Yoon Kye-sang, Jang Seung-jo, Kang ...	South Korea	November 30, 2019	2019	TV-14	
133	81034946	Maradona in Mexico	NaN	Diego Armando Maradona	Argentina, United States, Mexico	November 13, 2019	2020	TV-MA	

5837 rows × 14 columns

In [102]:
```
df.sort_values(by = 'release_year', ascending = False)
```

Out[102]:

	show_id	title	director	cast	country	date_added	release_year	rating	d
133	81034946	Maradona in Mexico	NaN	Diego Armando Maradona	Argentina, United States, Mexico	November 13, 2019	2020	TV-MA	
850	80213715	Whitney Cummings: Can I Touch It?	Marcus Raboy	Whitney Cummings	NaN	July 30, 2019	2019	TV-MA	
2009	80117498	Marvel's The Punisher	NaN	Jon Bernthal, Ebon Moss-Bachrach, Ben Barnes, ...	United States	NaN	2019	TV-MA	S
2006	80134721	IO	Jonathan Helpert	Margaret Qualley, Anthony Mackie, Danny Huston	United States	January 18, 2019	2019	TV-14	
2005	80017537	Grace and Frankie	NaN	Jane Fonda, Lily Tomlin, Martin Sheen, Sam Wat...	United States	NaN	2019	TV-MA	S
...	...	...	...	...	...	...	...	...	
4943	70013050	Why We Fight: The Battle of Russia	Frank Capra, Anatole Litvak	NaN	United States	March 31, 2017	1943	TV-14	
4941	80119186	Undercover: How to Operate Behind Enemy Lines	John Ford	NaN	United States	March 31, 2017	1943	TV-PG	
4929	60027945	Prelude to War	Frank Capra	NaN	United States	March 31, 2017	1942	TV-PG	
4931	60027942	The Battle of Midway	John Ford	Henry Fonda, Jane Darwell	United States	March 31, 2017	1942	TV-G	

	show_id	title	director	cast	country	date_added	release_year	rating	c
2154	81030762	Pioneers: First Women Filmmakers*	NaN	NaN	NaN	December 30, 2018	1925	TV-PG	

5837 rows × 14 columns

13. Find all the instances where:

Category IS 'Movie' and Type is 'Dramas'

or

Category is 'TV Show' & Type is 'Kids' 'TV'

In [ ]:
```
df[(df['type'] == 'TV Dramas') & (df[])]
```