In [1]:
```
import pandas as pd
import os
```

In [2]:
```
os.chdir(r"D:\Training\Data analyst\real data set")
print(os.getcwd())
```

```
D:\Training\Data analyst\real data set
```

In [3]:
```
df=pd.read_csv(r'netflex dataset.csv')
```

In [4]:
```
df.shape
```

Out[4]:
```
(7788, 11)
```

In [5]:
```
df.size
```

Out[5]:
```
85679
```

In [6]:
```
df.head(2)
```

Out[6]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi & ...	In a future where the elite inhabit an island ...
1	s2	Movie	07:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7789 entries, 0 to 7788
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   Show_Id       7789 non-null   object
 1   Category      7789 non-null   object
 2   Title         7789 non-null   object
 3   Director      5401 non-null   object
 4   Cast          7071 non-null   object
 5   Country       7282 non-null   object
 6   Release_Date  7779 non-null   object
 7   Rating        7782 non-null   object
 8   Duration      7789 non-null   object
 9   Type          7789 non-null   object
 10  Description   7789 non-null   object
dtypes: object(11)
memory usage: 669.5+ KB
```

In [8]:
```
df.columns
```

Out[8]:
```
Index(['Show_Id', 'Category', 'Title', 'Director', 'Cast', 'Country',
       'Release_Date', 'Rating', 'Duration', 'Type', 'Description'],
      dtype='object')
```

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

Out[9]:
```
Show_Id         object
Category        object
Title           object
Director        object
Cast            object
Country         object
Release_Date    object
Rating          object
Duration        object
Type            object
Description     object
dtype: object
```

In [10]:
```
df.isnull().sum()
```

Out[10]:
```
Show_Id         0
Category        0
Title           0
Director     2388
Cast          718
Country       507
Release_Date   10
Rating          7
Duration        0
Type            0
Description     0
dtype: int64
```

In [11]:
```
# is there any duplicates in the data set ,if yes then remove duplicates
df.duplicated()
```

Out[11]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description
6300	s684	Movie	Backfire	Dave Patten	Black Dennis, Byron "Squally" Vinson, Dominic ...	United States	April 5, 2019	TV-MA	97 min	Dramas, Independent Movies, Thrillers	When two would-be robbers accidentally kill a ...
6622	s6621	Movie	The Lost Okoroshi	Abba T. Makama	Seun Ajayi, Judith Audu, Tope Tedela, Ifu Ekwa...	Nigeria	September 6, 2020	TV-MA	94 min	Comedies, Dramas, Independent Movies	A disillusioned security guard transforms into ...

In [12]:
```
df.drop_duplicates(inplace=True)
```

In [13]:
```
# is there any null values ,if yes then show with heat map
df.isnull().sum()
```

Out[13]:
```
Show_Id         0
Category        0
Title           0
Director     2388
Cast          718
Country       507
Release_Date   10
Rating          7
Duration        0
Type            0
Description     0
dtype: int64
```

In [14]:
```
import seaborn as sn
import matplotlib.pyplot as plt
```

In [15]:
```
sn.heatmap(df.isnull())
# plt.show()
```

Out[15]:
```
<AxesSubplot:>
```

In [16]:
```
# for 'house of cards' what is the show id and who is the director of the film
df[df.Title.isin(['House of Cards'])]
```

Out[16]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description
2832	s2833	TV Show	House of Cards	Robin Wright, David Fincher, Gerald McRaney, J...	Kevin Spacey, Robin Wright, Kate Mara, Corey S...	United States	November 2, 2018	TV-MA	6 Seasons	TV Dramas, TV Thrillers	A ruthless politician will stop at nothing to ...

In [17]:
```
df[df.Title.str.contains('House of Cards')]
```

Out[17]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description
2832	s2833	TV Show	House of Cards	Robin Wright, David Fincher, Gerald McRaney, J...	Kevin Spacey, Robin Wright, Kate Mara, Corey S...	United States	November 2, 2018	TV-MA	6 Seasons	TV Dramas, TV Thrillers	A ruthless politician will stop at nothing to ...

In [18]:
```
df.dtypes
```

Out[18]:
```
Show_Id         object
Category        object
Title           object
Director        object
Cast            object
Country         object
Release_Date    object
Rating          object
Duration        object
Type            object
Description     object
dtype: object
```

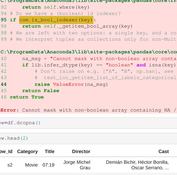
In [19]:
```
# in which year highest number of tv shows and movies where released? show in bar graph
df['date_n']=pd.to_datetime(df['Release_Date'])
```

In [20]:
```
df['date_n'].dt.year.value_counts()
```

Out[20]:
```
2019.0    2153
2020.0    2009
2018.0    1685
2017.0    1225
2016.0     443
2021.0     117
2015.0      88
2014.0      25
2011.0      13
2013.0      11
2012.0       3
2008.0       2
2009.0       2
2010.0       1
Name: date_n, dtype: int64
```

In [21]:
```
df['date_n'].dt.year.value_counts().plot(kind="bar")
```

Out[21]:
```
<AxesSubplot:>
```



In [22]:
```
# how many movies and tv shows are in the dataset ,show in bargraph
df.groupby('Category').Category.count().plot(kind='bar')
```

Out[22]:
```
<AxesSubplot:xlabel='Category'>
```

In [23]:
```
sn.countplot(df.Category)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[23]:
```
<AxesSubplot:xlabel='Category', ylabel='count'>
```

In [24]:
```
df.columns
```

Out[24]:
```
Index(['Show_Id', 'Category', 'Title', 'Director', 'Cast', 'Country',
       'Release_Date', 'Rating', 'Duration', 'Type', 'date_n'],
      dtype='object')
```

In [25]:
```
# show all the movies that were released in the year 2000
df[df['year']==df['date_n'].dt.year
```

In [26]:
```
df[(df['Category']=='Movie') & (df['year']==2020)].head(2)
```

Out[26]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description	date_n	year
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	PG-13	123 min	Dramas	A brilliant group of students become card-coun...	2020-01-01	2020.0
6	s7	Movie	122	Yasir Al Yasiri	Amina Khalil, Ahmed Dawood, Tarek Lotfy, Ahmed...	Egypt	June 1, 2020	TV-MA	95 min	Horror Movies, International Movies	After an awful accident, a couple admitted to ...	2020-06-01	2020.0

In [27]:
```
# show only the titles of tv shows that are released in India
df[(df['Category']=='TV Show') & (df['Country']=='India')].head(2)
```

Out[27]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description	date_n	year
86	s87	TV Show	21 Sarfarosh: Saragarhi 1897	NaN	Luke Kenny, Mohit Raina, Mukul Dev	India	December 1, 2018	TV-14	1 Season	International TV Shows, TV Dramas	In one of history's greatest last stands, a ba...	2018-12-01	2018.0
132	s133	TV Show	7 (Seven)	Nizar Shah	Rahman, Havish, Regina Cassandra, Nandita Swet...	India	July 30, 2019	TV-14	1 Season	TV Shows	Multiple women report their husbands as missin...	2019-07-30	2019.0

In [28]:
```
# show the top 10 directors who gave the highest number of tv shows and movie to netflix
df['Director'].value_counts().head(10)
```

Out[28]:
```
Rajiv Chilaka             19
Marcus Raboy              16
Jay Karas                 14
Cathy Garcia-Molina       13
Jay Chapman               12
Youssef Chahine           12
Martin Scorsese           12
Steven Spielberg          10
David Dhawan               9
Hakan Algül                8
Name: Director, dtype: int64
```

In [29]:
```
# in how many movies/shows Tom Cruise was cast
df[df['Cast']=='Tom Cruise']
```

Out[29]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description	date_n	year

In [30]:
```
df[df.Cast.str.contains('Tom Cruise')]

#as there as Nan values IN Cast columns, str.contains function give an error,so we have to drop the Nan from Cast column.
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Input In [30], in <cell line: 1>()
----> 1 df[df.Cast.str.contains('Tom Cruise')]

File C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\frame.py:3495, in DataFrame.__getitem__(self, key)
   3492     return self._getitem_slice(key)
   3494 # Do we have a (boolean) 1d indexer?
-> 3495 if com.is_bool_indexer(key):
   3496     return self._getitem_bool_array(key)
   3498 # We are left with two options: a single key, and a collection of keys,
   3499 # We interpret tuples as collections only for non-MultiIndex

File C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\common.py:144, in is_bool_indexer(key)
    140     na_msg = "Cannot mask with non-boolean array containing NA / NaN values"
    141     if lib.infer_dtype(key_array) == "boolean" and isna(key_array).any():
    142         # Don't raise on e.g. ["A", "B", np.nan], see
    143         #  test_loc_getitem_list_of_labels_categoricalindex_with_na
--> 144         raise ValueError(na_msg)
    145     return False
    146 return True

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

In [31]:
```
df_new=df.dropna()
```

In [32]:
```
df_new.head(2)
```

Out[32]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description	date_n	year
1	s2	Movie	07:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...	2016-12-23	2016.0
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...	2018-12-20	2018.0

In [33]:
```
df_new[df_new.Cast.str.contains('Tom Cruise')]
```

Out[33]:

	Show_Id	Category	Title	Director	Cast	Country	Release_Date	Rating	Duration	Type	Description	date_n	year
3860	s3861	Movie	Magnolia	Paul Thomas Anderson	Jeremy Blackman, Tom Cruise, Melinda Dillon, A...	United States	January 1, 2020	R	189 min	Dramas, Independent Movies	Through chance, human action, past history and...	2020-01-01	2020.0
5071	s5071	Movie	Rain Man	Barry Levinson	Dustin Hoffman, Tom Cruise, Valeria Golino, Ge...	United States	July 1, 2019	R	134 min	Classic Movies, Dramas	A fast-talking yuppie is forced to slow down w...	2019-07-01	2019.0

In [34]:
```
df_new.Rating.nunique()
```

Out[34]:
```
14
```

In [35]:
```
df_new.Rating.unique()
```

Out[35]:
```
array(['TV-MA', 'R', 'PG-13', 'TV-14', 'TV-G', 'TV-PG', 'NR', 'PG', 'G',
       'TV-Y7', 'TV-Y', 'TV-Y7-FV', 'NC-17', 'UR'], dtype=object)
```

In [36]:
```
# what is the maximum duration of a movie/show in netflix
df_new['Min/units']=df_new['Duration'].str.split(" ",expand=True)
```

```
C:\Users\lakshmi.panireddy\AppData\Local\Temp\ipykernel_26520\2144682247.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#returning-a-view-versus-a-copy
  df_new['Min/units']=df_new['Duration'].str.split(" ",expand=True)
C:\Users\lakshmi.panireddy\AppData\Local\Temp\ipykernel_26520\2144682247.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#returning-a-view-versus-a-copy
  df_new['Min/units']=df_new['Duration'].str.split(" ",expand=True)
```

In [37]:
```
df_new.Min.max()
```

Out[37]:
```
'99'
```

In [ ]:
```

```