Netflix Project

August 9, 2024

[1]:

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

[2]:

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

[3]:

```
df.head()
```

[3]:

```
  Show_Id Category  Title           Director  \
0      s1  TV Show     3%                NaN
1      s2    Movie  07:19  Jorge Michel Grau
2      s3    Movie  23:59       Gilbert Chan
3      s4    Movie      9        Shane Acker
4      s5    Movie     21     Robert Luketic

                                                Cast        Country  \
0  João Miguel, Bianca Comparato, Michel Gomes, R…         Brazil
1  Demián Bichir, Héctor Bonilla, Oscar Serrano, …         Mexico
2  Tedd Chan, Stella Chung, Henley Hii, Lawrence …      Singapore
3  Elijah Wood, John C. Reilly, Jennifer Connelly…  United States
4  Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar…  United States

        Release_Date Rating   Duration  \
0    August 14, 2020  TV-MA  4 Seasons
1  December 23, 2016  TV-MA     93 min
2  December 20, 2018      R     78 min
3  November 16, 2017  PG-13     80 min
4    January 1, 2020  PG-13    123 min

                                                Type  \
0  International TV Shows, TV Dramas, TV Sci-Fi &…
1                       Dramas, International Movies
2                Horror Movies, International Movies
3  Action & Adventure, Independent Movies, Sci-Fi…
4                                             Dramas

                                         Description
```

```
0  In a future where the elite inhabit an island …
1  After a devastating earthquake hits Mexico Cit…
2  When an army recruit is found dead, his fellow…
3  In a postapocalyptic world, rag-doll robots hi…
4  A brilliant group of students become card-coun…
```

[4]:

```
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
```

1 Is there any null value, if yes, show with heatmap and handle it accordingly

[5]:

```
df.isnull().sum()
```

[5]:

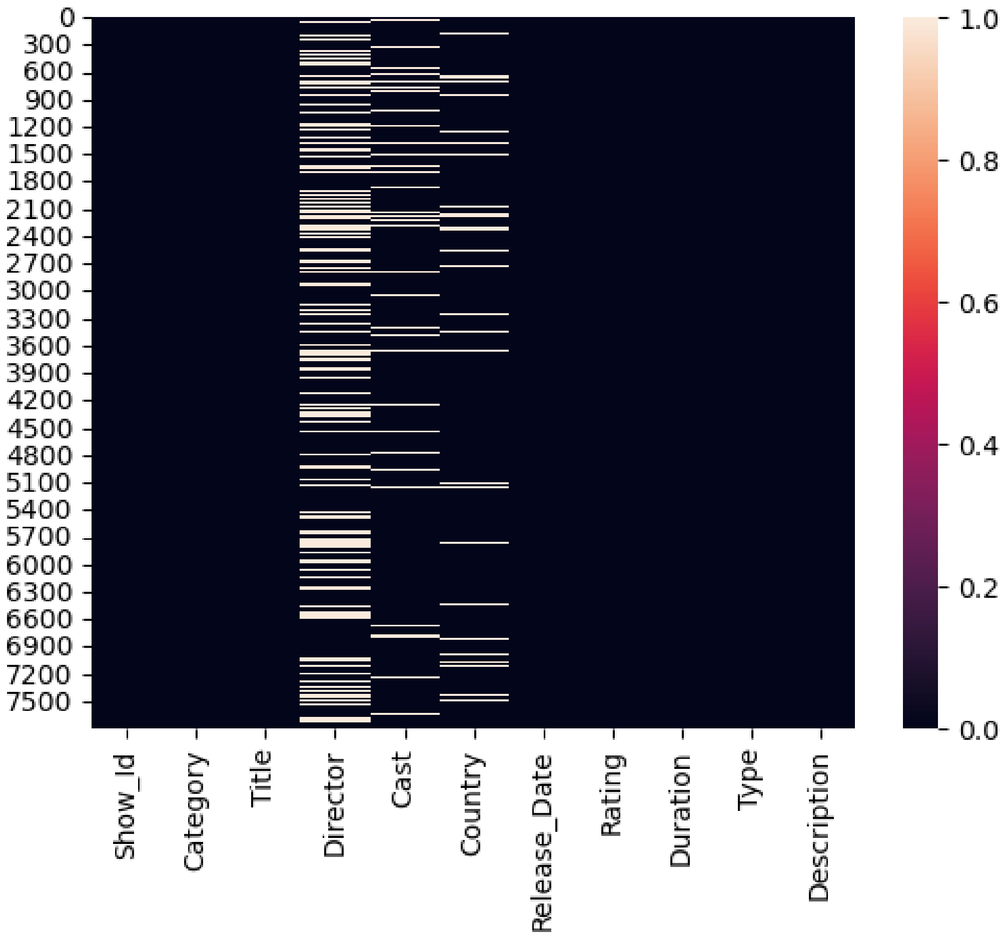
```
Show_Id            0
Category           0
Title              0
Director        2388
Cast             718
Country          507
Release_Date      10
Rating             7
Duration           0
Type               0
Description        0
dtype: int64
```

[6]:

```
sns.heatmap(df.isnull())
```

[6]:

```
<Axes: >
```



[7]:

```python
df["Director"]=df["Director"].fillna("Not available")
```

[8]:

```python
df["Cast"]=df["Cast"].fillna("Not available")
```

[9]:

```python
df["Country"]=df["Country"].fillna("Not available")
```

[10]:

```python
df=df.dropna()
```

2 Is there any duplicate value, if yes, remove it from the data

[11]:

```python
df.duplicated().sum()
```

[11]:

```
2
```

[12]:

```python
df=df.drop_duplicates()
```

3 For “House of cards” who is the show director and what is the show id

[13]:
```
df2=df[df["Title"]=="House of Cards"]
df2[["Director", "Show_Id"]]
```

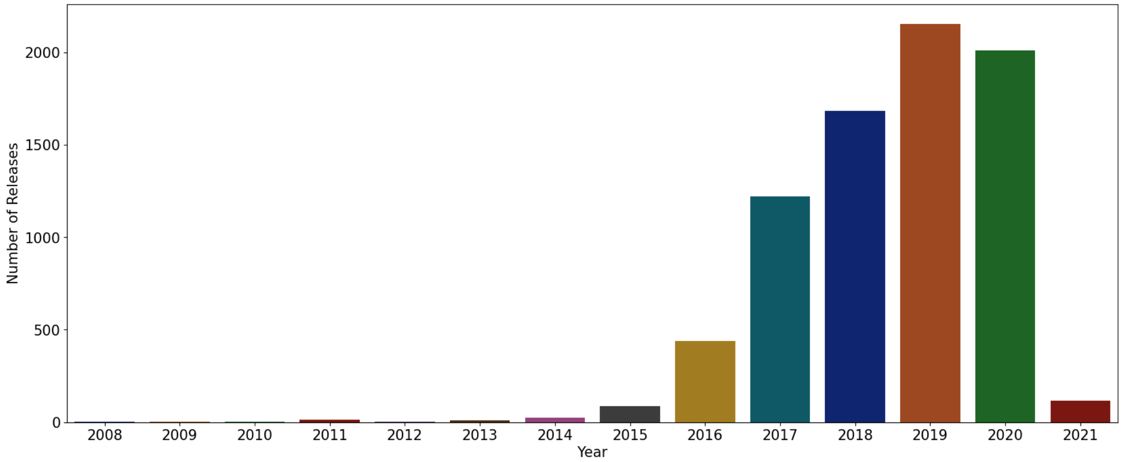
[13]:
```
                                               Director Show_Id
2832  Robin Wright, David Fincher, Gerald McRaney, J…   s2833
```

4 In which year the highest number of TV Shows and Movies were released?

[14]:
```
df['Date']=pd.to_datetime(df['Release_Date'])
```

[15]:
```
df["Year"] = df["Date"].dt.year
```

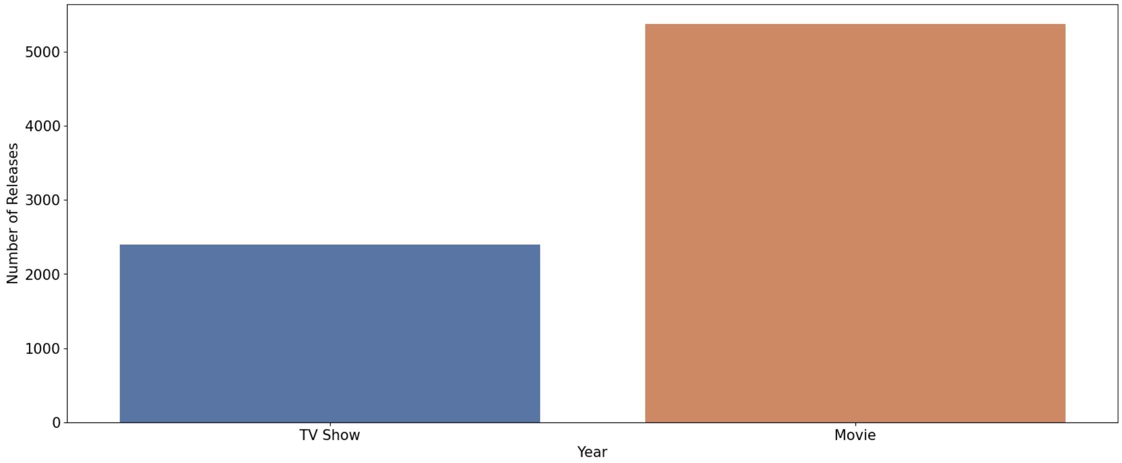
[16]:
```
plt.figure(figsize=(20,8))
sns.countplot(x=df["Year"], palette='dark')
plt.xlabel('Year', fontsize=15)
plt.ylabel('Number of Releases', fontsize=15)
plt.xticks(size=15)
plt.yticks(size=15)
plt.show()
```



[17]:
```
#Conclusion: From graph we can see that highest number of movies or shows were␣
↪released in year 2019, followed by 2020 and 2018.
```

5 How many movies and shows are there in dataset?

```
[18]: plt.figure(figsize=(20,8))
      sns.countplot(x=df["Category"], palette='deep')
      plt.xlabel('Year', fontsize=15)
      plt.ylabel('Number of Releases', fontsize=15)
      plt.xticks(size=15)
      plt.yticks(size=15)
      plt.show()
```



6 Show all the movies that were released in year 2020

```
[19]: df.loc[15:20].head()
```

```
[19]:    Show_Id Category   Title            Director  \
      15     s16    Movie  Oct-01      Kunle Afolayan
      16     s17  TV Show  Feb-09       Not available
      17     s18    Movie  22-Jul     Paul Greengrass
      18     s19    Movie  15-Aug  Swapnaneel Jayakar
      19     s20    Movie     '89       Not available

                                                     Cast  \
      15  Sadiq Daba, David Bailie, Kayode Olaiya, Kehin…
      16  Shahd El Yaseen, Shaila Sabt, Hala, Hanadi Al-…
      17  Anders Danielsen Lie, Jon Øigarden, Jonas Stra…
      18  Rahul Pethe, Mrunmayee Deshpande, Adinath Koth…
      19                  Lee Dixon, Ian Wright, Paul Merson

                               Country       Release_Date Rating  Duration  \
      15                       Nigeria  September 1, 2019  TV-14   149 min
```

```
16                   Not available     March 20, 2019  TV-14  1 Season
17  Norway, Iceland, United States   October 10, 2018      R   144 min
18                           India     March 29, 2019  TV-14   124 min
19                  United Kingdom       May 16, 2018  TV-PG    87 min

                                              Type  \
15  Dramas, International Movies, Thrillers
16        International TV Shows, TV Dramas
17                       Dramas, Thrillers
18     Comedies, Dramas, Independent Movies
19                           Sports Movies

                                          Description       Date  Year
15  Against the backdrop of Nigeria's looming inde… 2019-09-01  2019
16  As a psychology professor faces Alzheimer's, h… 2019-03-20  2019
17  After devastating terror attacks in Norway, a … 2018-10-10  2018
18  On India's Independence Day, a zany mishap in … 2019-03-29  2019
19  Mixing old footage with interviews, this is th… 2018-05-16  2018
```

[20]:

```
df2 = df[(df["Year"] == 2020) & (df["Category"] == 'Movie')]
df2[["Title", "Duration"]]
```

[20]:

```
                                         Title Duration
4                                           21  123 min
6                                          122   95 min
14                                        3022   91 min
27                                      #Alive   99 min
28                #AnneFrank - Parallel Stories   95 min
…                                            …        …
7762                                 Zaki Chan  109 min
7783                                      Zoom   88 min
7784                                      Zozo   99 min
7786                        Zulu Man in Japan   44 min
7788  ZZ TOP: THAT LITTLE OL' BAND FROM TEXAS   90 min

[1312 rows x 2 columns]
```

7 Show top 10 directors who gave most tv shows and movies to netflix

[21]:

```
df5 = df.groupby(['Director'])[['Director']].value_counts().
 ↪reset_index(name="No. of Shows")
df5 = df5.rename(columns={'Director': "Name of Director"})
df5 = df5.sort_values(by="No. of Shows", ascending= False)
df5[df5["Name of Director"]!= "Not available"].head(10)
```

```
[21]:             Name of Director  No. of Shows
      3077  Raúl Campos, Jan Suter            18
      2319              Marcus Raboy            16
      1606                 Jay Karas            14
      623        Cathy Garcia-Molina            13
      1603               Jay Chapman            12
      2386            Martin Scorsese           12
      4007            Youssef Chahine           12
      3599           Steven Spielberg           10
      874                David Dhawan            9
      3182           Robert Rodriguez            8
```

8 In how many movies there was Tom Cruise?

```
[22]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7770 entries, 0 to 7788
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Show_Id       7770 non-null   object
 1   Category      7770 non-null   object
 2   Title         7770 non-null   object
 3   Director      7770 non-null   object
 4   Cast          7770 non-null   object
 5   Country       7770 non-null   object
 6   Release_Date  7770 non-null   object
 7   Rating        7770 non-null   object
 8   Duration      7770 non-null   object
 9   Type          7770 non-null   object
 10  Description   7770 non-null   object
 11  Date          7770 non-null   datetime64[ns]
 12  Year          7770 non-null   int64
dtypes: datetime64[ns](1), int64(1), object(11)
memory usage: 1.1+ MB
```

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

```
[23]:      Show_Id Category     Title              Director  \
      3860   s3861    Movie  Magnolia  Paul Thomas Anderson
      5071   s5071    Movie  Rain Man        Barry Levinson

                                                    Cast        Country  \
      3860  Jeremy Blackman, Tom Cruise, Melinda Dillon, A…  United States
      5071  Dustin Hoffman, Tom Cruise, Valeria Golino, Ge…  United States
```

```
         Release_Date Rating Duration                        Type  \
3860  January 1, 2020      R  189 min  Dramas, Independent Movies
5071     July 1, 2019      R  134 min      Classic Movies, Dramas

                                            Description       Date  Year
3860  Through chance, human action, past history and… 2020-01-01  2020
5071  A fast-talking yuppie is forced to slow down w… 2019-07-01  2019
```

9 What are the different ratings defined by Netflix

```
[24]: df["Rating"].nunique()
```

```
[24]: 14
```

```
[25]: df["Rating"].unique()
```

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

10 Which individual country has the highest number of Shows or movies

```
[45]: df4=df.groupby(["Country"])[['Country']].value_counts().reset_index(name= 'No.␣
       ↪of Shows/Movies')
      df4.sort_values(by="No. of Shows/Movies", ascending=False).head(1)
```

```
[45]:            Country  No. of Shows/Movies
      550  United States                 2546
```