

Exploratory Data Analytics on Netflix Data

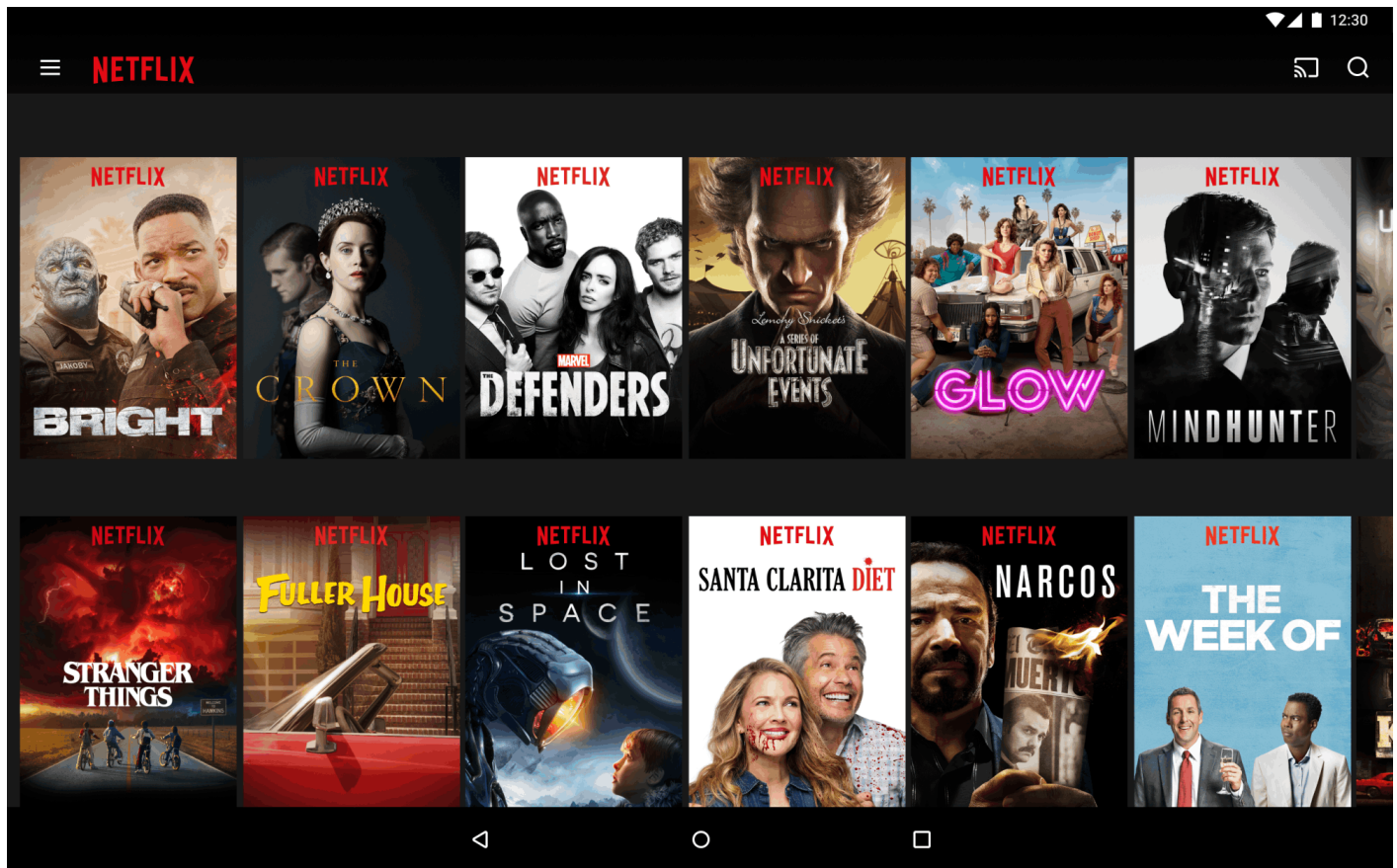


Table of Contents

1. [Data Description](#)
 - 1.1 [Introduction](#)
 - 1.2 [Data source and data set](#)
2. [Load the Packages and Data](#)
3. [Data Profiling](#)
 - 3.1 [Understanding the Dataset](#)
 - 3.2 [Pre Profiling](#)
 - 3.3 [Preprocessing](#)
 - 3.4 [Post Profiling](#)
4. [EDA Questions](#)
 - 4.1 [What different types of show or movie are uploaded on Netflix?](#)
 - 4.2 [What is the Correlation between the features?](#)
 - 4.3 [Most watched shows on the Netflix?](#)
 - 4.4 [Distribution of Ratings?](#)
 - 4.5 [Which has the highest rating Tv show or Movies?](#)
 - 4.6 [Finding the best Month for releasing content?](#)
 - 4.7 [Highest watched genres on Netflix?](#)
 - 4.8 [Released movie over the years?](#)
 - 4.9 [movies made on year basis?](#)
 - 4.10 [What is the show id and director for 'House of cards'?](#)
 - 4.11 [4.11 Show all the movies that were released in year 2000.](#)
 - 4.12 [Show only the title of all TV shows that were released in India only.](#)
 - 4.13 [Show top 10 director, who gave the highest number of TV shows & Movies to Netflix?](#)
 - 4.14 [In how many movies/ tv shows, 'tom Cruise' was cast?](#)
 - 4.15 [How many movies got the "TV-14" rating in the canada?](#)
5. [Conclusions](#)

1. Data Description:

This Netflix Dataset has information about the TV shows and Movies available on Netflix from the year 2008 to 2021. Netflix is an application that keeps growing exponentially whole around the world and it is the most famous streaming platform. This dataset also available on kaggle website for free.

This dataset collected from Netflix of different TV shows and Movies from the year 2008 to 2021.

- show_id: Gives the information about show id

- **type**: Gives information about 2 different unique values one is TV Show and another is Movie.
- **title**: Gives information about the title of Movie or TV Show.
- **director**: Gives information about the director who directed the Movie or TV Show.
- **cast**: Gives information about the cast who plays role in Movie or TV Show.
- **country**: Gives information about the Name of country.
- **date_added**: Gives information about the tv shows or movie released.
- **release_year**: Gives information about the year when Movie or TV Show was released.
- **rating**: Gives information about the Movie or TV Show are in which category (eg like the movies are only for students, or adults, etc).
- **duration**: Gives information about the duration of Movie or TV Show.
- **listed_in**: Gives information about the genre of Movie or TV Show.
- **description**: Gives information about the description of Movie or TV Show.

1.1. Introduction

This Exploratory Data Analysis is to practice Python skills learned till now on a structured data set including loading, inspecting, wrangling, exploring, and drawing conclusions from data. The notebook has observations with each step in order to explain thoroughly how to approach the data set. Based on the observation some questions also are answered in the notebook for the reference though not all of them are explored in the analysis.

1.2 Data source and data set

This Netflix Dataset has information about the TV shows and Movies available on Netflix from 2018 to 2021. This dataset collected from kaggle website for free.

- Dataset link: <https://www.kaggle.com/datasets/swatikhedekar/exploratory-data-analysis-on-netflix-data> (<https://www.kaggle.com/datasets/swatikhedekar/exploratory-data-analysis-on-netflix-data>)

2. Load the Packages and Data

Import libraries or modules

In [1]:

```
%%time
import numpy as np           # for numerical computing
import pandas as pd          # for dataframe and manipulation
import seaborn as sns        #for graphs and data visualization
from matplotlib import pyplot as plt
import pandas_profiling as prf
sns.set()
%matplotlib inline
```

CPU times: total: 8.23 s

Wall time: 22.8 s

Load the dataset

In [2]:

```
netflix = pd.read_csv(r"F:\SDA\live\python1\12_jan_netflix\netflix_titles_2021.csv")
```

3.Data Profiling (Data Exploration)

3.1 Understanding the Dataset(Basic information about dataset)

In [3]:

```
netflix.head() # for shows top 5 rows
```

Out[3]:

	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 Mabalané, Thabani...	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, Nabl...	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 l...

In [4]:

```
netflix.tail() # For bottom 5 rows
```

Out[4]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
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	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...
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...

In [5]:

```
netflix.shape # to show the total no. of columns and rows
```

Out[5]:

(8807, 12)

- This dataset contains 8807 rows and 12 columns.

In [6]:

```
netflix.size # to show the total no. of volume(elements)
```

Out[6]:

105684

In [7]:

```
netflix.columns # to shows each column name
```

Out[7]:

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

In [8]:

```
netflix.dtypes      # to shows data types each column name
```

Out[8]:

```
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
```

In [9]:

```
netflix.info()      # to shows indexes,data types each column name
```

```
<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
```

In [10]:

```
netflix.describe() # for shows statistical information
```

Out[10]:

```
release_year
count  8807.000000
mean   2014.180198
std     8.819312
min    1925.000000
25%    2013.000000
50%    2017.000000
75%    2019.000000
max     2021.000000
```

In [11]:

```
netflix.describe(include='all')
```

Out[11]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
count	8807	8807	8807	6173	7982	7976	8797	8807.000000	8803	8804	8807	8807
unique	8807	2	8807	4528	7692	748	1767	NaN	17	220	514	8775
top	s1	Movie	Dick Johnson Is Dead	Rajiv Chilaka	David Attenborough	United States	January 1, 2020	NaN	TV-MA	1 Season	Dramas, International Movies	Paranormal activity at a lush, abandoned prope...
freq	1	6131	1	19	19	2818	109	NaN	3207	1793	362	4
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2014.180198	NaN	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8.819312	NaN	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1925.000000	NaN	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2013.000000	NaN	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2017.000000	NaN	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2019.000000	NaN	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2021.000000	NaN	NaN	NaN	NaN

In [12]:

```
#Finding how many unique values are in the dataset
netflix.nunique()
```

Out[12]:

```
show_id      8807
type          2
title        8807
director     4528
cast         7692
country       748
date_added   1767
release_year  74
rating        17
duration     220
listed_in    514
description   8775
dtype: int64
```

Data Preperation(Data Cleaning)

- Check missing values
- Fill null values

In [13]:

```
netflix.isnull() # to shows where the null values is present
```

Out[13]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	False	False	False	False	True	False	False	False	False	False	False	False
1	False	False	False	True	False	False	False	False	False	False	False	False
2	False	False	False	False	False	True	False	False	False	False	False	False
3	False	False	False	True	True	True	False	False	False	False	False	False
4	False	False	False	True	False	False	False	False	False	False	False	False
...
8802	False	False	False	False	False	False	False	False	False	False	False	False
8803	False	False	False	True	True	True	False	False	False	False	False	False
8804	False	False	False	False	False	False	False	False	False	False	False	False
8805	False	False	False	False	False	False	False	False	False	False	False	False
8806	False	False	False	False	False	False	False	False	False	False	False	False

8807 rows × 12 columns

In [14]:

```
miss=netflix.isnull().sum() # to shows the count of null values
miss
```

Out[14]:

```
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
```

In [15]:

```
miss1 = (netflix.isnull().sum()/len(netflix))* 100
miss1
```

Out[15]:

```
show_id      0.000000
type         0.000000
title        0.000000
director     29.908028
cast         9.367549
country      9.435676
date_added   0.113546
release_year 0.000000
rating       0.045418
duration     0.034064
listed_in    0.000000
description  0.000000
dtype: float64
```

In [16]:

```
#missing values with %
m = pd.concat([miss,miss1],axis=1,keys=['Total', '%'])
m
```

Out[16]:

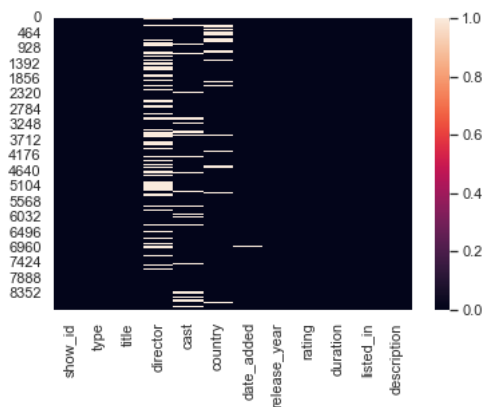
	Total	%
show_id	0	0.000000
type	0	0.000000
title	0	0.000000
director	2634	29.908028
cast	825	9.367549
country	831	9.435676
date_added	10	0.113546
release_year	0	0.000000
rating	4	0.045418
duration	3	0.034064
listed_in	0	0.000000
description	0	0.000000

In [17]:

```
sns.heatmap(netflix.isnull()) # using heatmap to shows the null value count
```

Out[17]:

<AxesSubplot:>



- From the above output we can see that **director** , **cast** ,**country** columns contains **maximum null values**. We will see how to deal with them.
- So, We Delete director and cast columns because they are not going to use those features right now.

In [18]:

```
# making copy of dataset for changes
netflix_copy = netflix.copy()
netflix_copy.head()
```

Out[18]:

	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 l...

In [19]:

```
netflix_copy =netflix_copy.dropna(how='any', subset=['director', 'cast'])
netflix_copy.head(2)
```

Out[19]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
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...
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	NaN	September 24, 2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brin...

In [20]:

```
netflix_copy.fillna({'country':'missing', 'rating':'missing'},inplace=True)
```

In [21]:

```
netflix_copy.isnull().sum()
```

Out[21]:

```
show_id      0
type         0
title        0
director     0
cast         0
country      0
date_added  0
release_year 0
rating       0
duration     3
listed_in    0
description  0
dtype: int64
```

3.2 Pre-Profiling

In [22]:

```
pip install -U pandas_profiling
```



```

Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: pandas_profiling in c:\users\user\appdata\roaming\python\python39\site-packages (3.6.2)
Collecting pandas_profiling
  Downloading pandas_profiling-3.6.3-py2.py3-none-any.whl (328 kB)
Requirement already satisfied: phik<0.13,>=0.11.1 in c:\users\user\appdata\roaming\python\python39\site-packages (from pandas_profiling) (0.12.3)
Requirement already satisfied: Jinja2<3.2,>=2.11.1 in f:\my apps\anaconda\conda\lib\site-packages (from pandas_profiling) (2.11.3)
Requirement already satisfied: HTMLmin==0.1.12 in c:\users\user\appdata\roaming\python\python39\site-packages (from pandas_profiling) (0.1.12)
Requirement already satisfied: pandas!=1.4.0,<1.6,>1.1 in f:\my apps\anaconda\conda\lib\site-packages (from pandas_profiling) (1.4.2)
Requirement already satisfied: PyYAML<6.1,>=5.0.0 in f:\my apps\anaconda\conda\lib\site-packages (from pandas_profiling) (6.0)
Requirement already satisfied: tqdm<4.65,>=4.48.2 in f:\my apps\anaconda\conda\lib\site-packages (from pandas_profiling) (4.64.0)
Requirement already satisfied: statsmodels<0.14,>=0.13.2 in f:\my apps\anaconda\conda\lib\site-packages (from pandas_profiling) (0.13.2)
Requirement already satisfied: numpy<1.24,>=1.16.0 in f:\my apps\anaconda\conda\lib\site-packages (from pandas_profiling) (1.21.5)
Requirement already satisfied: visions[type_image_path]==0.7.5 in c:\users\user\appdata\roaming\python\python39\site-packages (from pandas_profiling) (0.7.5)
Requirement already satisfied: requests<2.29,>=2.24.0 in f:\my apps\anaconda\conda\lib\site-packages (from pandas_profiling) (2.27.1)
Requirement already satisfied: typeguard<2.14,>=2.13.2 in c:\users\user\appdata\roaming\python\python39\site-packages (from pandas_profiling) (2.13.3)
Requirement already satisfied: seaborn<0.13,>=0.10.1 in f:\my apps\anaconda\conda\lib\site-packages (from pandas_profiling) (0.11.2)
Requirement already satisfied: pydantic<1.11,>=1.8.1 in c:\users\user\appdata\roaming\python\python39\site-packages (from pandas_profiling) (1.10.4)
Requirement already satisfied: scipy<1.10,>=1.4.1 in f:\my apps\anaconda\conda\lib\site-packages (from pandas_profiling) (1.7.3)
Requirement already satisfied: multimethod<1.10,>=1.4 in c:\users\user\appdata\roaming\python\python39\site-packages (from pandas_profiling) (1.9.1)
Requirement already satisfied: matplotlib<3.7,>=3.2 in f:\my apps\anaconda\conda\lib\site-packages (from pandas_profiling) (3.5.1)
Requirement already satisfied: networkx>=2.4 in f:\my apps\anaconda\conda\lib\site-packages (from visions[type_image_path]==0.7.5->pandas_profiling) (2.7.1)
Requirement already satisfied: tangled-up-in-unicode>=0.0.4 in c:\users\user\appdata\roaming\python\python39\site-packages (from visions[type_image_path]==0.7.5->pandas_profiling) (0.2.0)
Requirement already satisfied: attrs>=19.3.0 in f:\my apps\anaconda\conda\lib\site-packages (from visions[type_image_path]==0.7.5->pandas_profiling) (21.4.0)
Requirement already satisfied: imagehash in c:\users\user\appdata\roaming\python\python39\site-packages (from visions[type_image_path]==0.7.5->pandas_profiling) (4.3.1)
Requirement already satisfied: Pillow in f:\my apps\anaconda\conda\lib\site-packages (from visions[type_image_path]==0.7.5->pandas_profiling) (9.0.1)
Requirement already satisfied: MarkupSafe>=0.23 in f:\my apps\anaconda\conda\lib\site-packages (from Jinja2<3.2,>=2.11.1->pandas_profiling) (2.0.1)
Requirement already satisfied: fonttools>=4.22.0 in f:\my apps\anaconda\conda\lib\site-packages (from matplotlib<3.7,>=3.2->pandas_profiling) (4.25.0)
Requirement already satisfied: python-dateutil>=2.7 in f:\my apps\anaconda\conda\lib\site-packages (from matplotlib<3.7,>=3.2->pandas_profiling) (2.8.2)
Requirement already satisfied: pyparsing>=2.2.1 in f:\my apps\anaconda\conda\lib\site-packages (from matplotlib<3.7,>=3.2->pandas_profiling) (3.0.4)
Requirement already satisfied: kiwisolver>=1.0.1 in f:\my apps\anaconda\conda\lib\site-packages (from matplotlib<3.7,>=3.2->pandas_profiling) (1.3.2)
Requirement already satisfied: packaging>=20.0 in f:\my apps\anaconda\conda\lib\site-packages (from matplotlib<3.7,>=3.2->pandas_profiling) (21.3)
Requirement already satisfied: cycler>=0.10 in f:\my apps\anaconda\conda\lib\site-packages (from matplotlib<3.7,>=3.2->pandas_profiling) (0.11.0)
Requirement already satisfied: pytz>=2020.1 in c:\users\user\appdata\roaming\python\python39\site-packages (from pandas!=1.4.0,<1.6,>1.1->pandas_profiling) (2022.7.1)
Requirement already satisfied: joblib>=0.14.1 in f:\my apps\anaconda\conda\lib\site-packages (from phik<0.13,>=0.11.1->pandas_profiling) (1.1.0)
Requirement already satisfied: typing-extensions>=4.2.0 in c:\users\user\appdata\roaming\python\python39\site-packages (from pydantic<1.11,>=1.8.1->pandas_profiling) (4.4.0)
Requirement already satisfied: six>=1.5 in f:\my apps\anaconda\conda\lib\site-packages (from python-dateutil>=2.7->matplotlib<3.7,>=3.2->pandas_profiling) (1.16.0)
Requirement already satisfied: idna<4,>=2.5 in f:\my apps\anaconda\conda\lib\site-packages (from requests<2.29,>=2.24.0->pandas_profiling) (3.3)
Requirement already satisfied: certifi>=2017.4.17 in f:\my apps\anaconda\conda\lib\site-packages (from requests<2.29,>=2.24.0->pandas_profiling) (2021.10.8)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in f:\my apps\anaconda\conda\lib\site-packages (from requests<2.29,>=2.24.0->pandas_profiling) (1.26.9)
Requirement already satisfied: charset-normalizer~2.0.0 in f:\my apps\anaconda\conda\lib\site-packages (from requests<2.29,>=2.24.0->pandas_profiling) (2.0.4)
Requirement already satisfied: patsy>=0.5.2 in f:\my apps\anaconda\conda\lib\site-packages (from statsmodels<0.14,>=0.13.2->pandas_profiling) (0.5.2)
Requirement already satisfied: colorama in f:\my apps\anaconda\conda\lib\site-packages (from tqdm<4.65,>=4.48.2->pandas_profiling) (0.4.4)
Requirement already satisfied: PyWavelets in f:\my apps\anaconda\conda\lib\site-packages (from imagehash->visions[type_image_path]==0.7.5->pandas_profiling) (1.3.0)
Installing collected packages: pandas-profiling
  Attempting uninstall: pandas-profiling
    Found existing installation: pandas-profiling 3.6.2
    Uninstalling pandas-profiling-3.6.2:
      Successfully uninstalled pandas-profiling-3.6.2
  Successfully installed pandas-profiling-3.6.3
Note: you may need to restart the kernel to use updated packages.

```

WARNING: The script pandas_profiling.exe is installed in 'C:\Users\user\AppData\Roaming\Python\Python39\Scripts' which is not on PATH.
Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Now performing pandas profiling to understand data better.

In [23]:

```
netflix_profile = prf.ProfileReport(netflix)
netflix_profile
```

Summarize dataset: 100%23/23 [00:12<00:00, 3.06it/s, Completed]

Generate report structure: 100%1/1 [00:08<00:00, 8.30s/it]

Render HTML: 100%1/1 [00:04<00:00, 4.00s/it]

Overview

Dataset statistics

Number of variables	12
Number of observations	8807
Missing cells	4307
Missing cells (%)	4.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	825.8 KiB
Average record size in memory	96.0 B

Variable types

Categorical	11
Numeric	1

Alerts

show_id has a high cardinality: 8807 distinct values	High cardinality
title has a high cardinality: 8807 distinct values	High cardinality
director has a high cardinality: 4528 distinct values	High cardinality
cast has a high cardinality: 7692 distinct values	High cardinality
country has a high cardinality: 748 distinct values	High cardinality
date_added has a high cardinality: 1767 distinct values	High cardinality
duration has a high cardinality: 220 distinct values	High cardinality

Out[23]:

In [24]:

```
netflix_profile.to_file(output_file="netflix21_before_preprocessing.html")
```

Export report to file: 100%1/1 [00:00<00:00, 9.26it/s]

3.3 Preprocessing

- Dealing with duplicate rows
 - Find number of duplicate rows in the dataset.
 - Print the duplicate entries and analyze.
 - Drop the duplicate entries from the dataset.

In [25]:

```
netflix[netflix.duplicated()] # to shows the dupblicate rows
```

Out[25]:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
---------	------	-------	----------	------	---------	------------	--------------	--------	----------	-----------	-------------

In [26]:

```
netflix.duplicated().sum() # to shows the count of dupblicate rows
```

Out[26]:

0

- No duplicate rows are present in this dataset.

4.EDA Questions

4.1. What different types of show or movie are uploaded on Netflix? or

How many TV shows & Movies are in the dataset? Show with Bar Gragh?

In [27]:

```
netflix_copy.head(2)
```

Out[27]:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	missing	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...
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	missing	September 24, 2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brin...

In [28]:

```
# method-1:
netflix_copy.groupby('type')['title'].count().sort_values(ascending = False)
```

Out[28]:

```
type
Movie      5522
TV Show     178
Name: title, dtype: int64
```

In [29]:

```
# method-2:
netflix_copy.type.value_counts().to_frame('Value_count')
```

Out[29]:

	Value_count
Movie	5522
TV Show	178

In [30]:

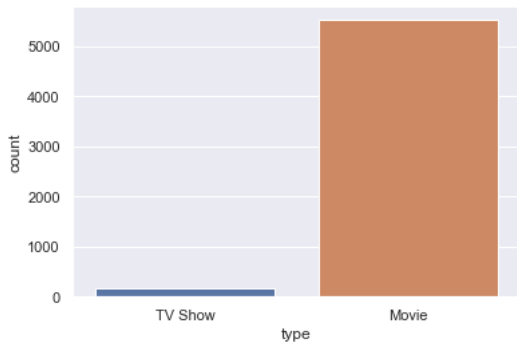
```
sns.countplot(netflix_copy['type'])# to show the count of all unique item of a any column in the form of bargrapp
```

F:\my apps\anaconda\conda\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[30]:

```
<AxesSubplot:xlabel='type', ylabel='count'>
```



Observation: There are **5522-types** of movies and **178-types** of tv shows are uploaded on Netflix.

4.2 What is the Correlation between the features?

In [31]:

```
netflix_copy.dtypes
```

Out[31]:

```
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
```

- Converting object into datetime formate of "date_added"

In [32]:

```
netflix_copy["date_added"]=pd.to_datetime(netflix_copy['date_added'])
netflix_copy.head(2)
```

Out[32]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	missing	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	missing	2021-09-24	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brin...

- Now ,we create two features year_added and month_added from date_added column.

In [33]:

```
netflix_copy['year_added'] = netflix_copy["date_added"].dt.year
netflix_copy['month_added'] = netflix_copy["date_added"].dt.month
netflix_copy.head(2)
```

Out[33]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	year_added	month_added
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	missing	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...	2021	9
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	missing	2021-09-24	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brin...	2021	9

In [34]:

```
netflix_copy.dtypes
```

Out[34]:

```
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
year_added   int64
month_added  int64
dtype: object
```

In [35]:

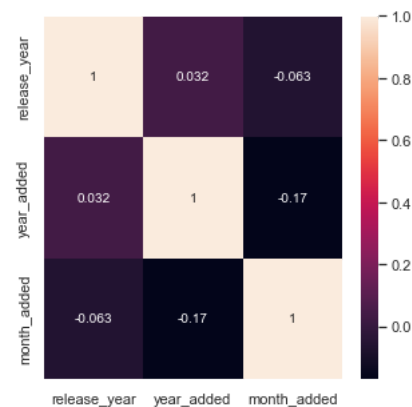
```
netflix_copy.corr()
```

Out[35]:

	release_year	year_added	month_added
release_year	1.000000	0.032265	-0.062903
year_added	0.032265	1.000000	-0.170078
month_added	-0.062903	-0.170078	1.000000

In [36]:

```
plt.subplots(figsize=(5,5))
sns.heatmap(netflix_copy.corr(),annot=True)
plt.show()
```



Observation : Above Heatmap shows correlation between release_year, year_added & month_added.

4.3 Most watched shows on the Netflix?

In [37]:

```
netflix_copy.type.value_counts().to_frame('Value_count')
```

Out[37]:

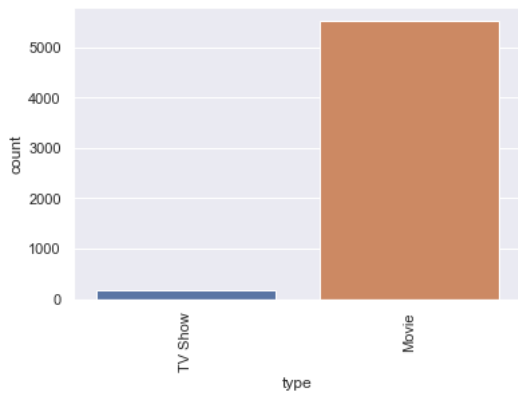
	Value_count
Movie	5522
TV Show	178

In [38]:

```
sns.countplot(x=netflix_copy['type'],orient='v')
plt.xticks(rotation=90)
```

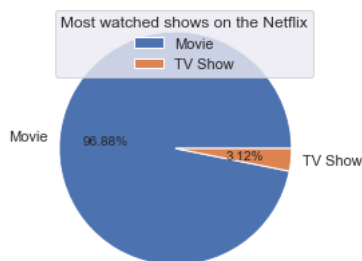
Out[38]:

```
(array([0, 1]), [Text(0, 0, 'TV Show'), Text(1, 0, 'Movie')])
```



In [39]:

```
# by using Pie chart
type_show = ['Movie', 'TV Show']
Value_count = [5522, 178]
plt.pie(Value_count, labels=type_show, autopct="%2.2f%%")
plt.legend(title='Most watched shows on the Netflix')
plt.show()
```



Observation: Audience likes to watched mostly **movies(96.88%)** over **TV shows(3.12%)**.

4.4 Distribution of Ratings? &

What are the different rating defined by Netflix?

In [40]:

```
netflix_copy.head(2)
```

Out[40]:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	year_added	month_added	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	missing	2021-09-24	2021	TV-MA	Season 1	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...	2021	9
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	missing	2021-09-24	2021	TV-MA	Season 1	TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brin...	2021	9

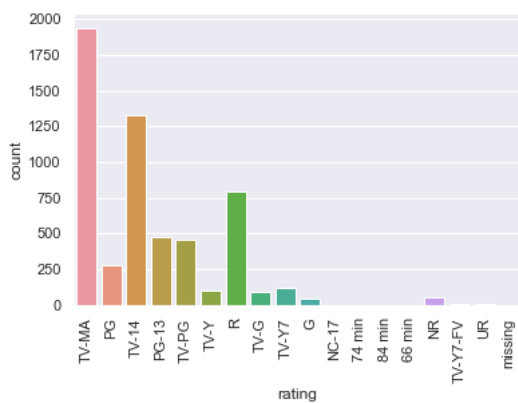
In [41]:

```
#countplot for distrubution
```

```
sns.countplot(x=netflix_copy['rating'],orient='v')
plt.xticks(rotation=90)
```

Out[41]:

```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
        17]),
 [Text(0, 0, 'TV-MA'),
  Text(1, 0, 'PG'),
  Text(2, 0, 'TV-14'),
  Text(3, 0, 'PG-13'),
  Text(4, 0, 'TV-PG'),
  Text(5, 0, 'TV-Y'),
  Text(6, 0, 'R'),
  Text(7, 0, 'TV-G'),
  Text(8, 0, 'TV-Y7'),
  Text(9, 0, 'G'),
  Text(10, 0, 'NC-17'),
  Text(11, 0, '74 min'),
  Text(12, 0, '84 min'),
  Text(13, 0, '66 min'),
  Text(14, 0, 'NR'),
  Text(15, 0, 'TV-Y7-FV'),
  Text(16, 0, 'UR'),
  Text(17, 0, 'missing')])
```



In [42]:

```
netflix['rating'].nunique()
```

Out[42]:

17

Obeservation :

- Audience prefers mostly TV-MA & TV-14 and less prefers NC-17 as rating.
- There are 17 types of ratings in the netflix.

4.5 Which has the highest rating Tv show or Movies?

In [43]:

```
netflix_copy.head(2)
```

Out[43]:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	year_added	month_added
2	s3 TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	missing	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...	2021	9
5	s6 TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	missing	2021-09-24	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brin...	2021	9

In [44]:

```
netflix.groupby("type")["rating"].agg(pd.Series.mode)
```

Out[44]:

```
type
Movie      TV-MA
TV Show    TV-MA
Name: rating, dtype: object
```

Adding new feature

```
new_genre (Genre 1,Genre 2,Genre 3)
```

In [45]:

```
new_genre = netflix_copy['listed_in'].str.split(",",2)
new_genre
```

Out[45]:

```
2      [Crime TV Shows, International TV Shows, TV ...
5      [TV Dramas, TV Horror, TV Mysteries]
6      [Children & Family Movies]
7      [Dramas, Independent Movies, International M...
8      [British TV Shows, Reality TV]
...
8801   [Dramas, International Movies, Thrillers]
8802   [Cult Movies, Dramas, Thrillers]
8804   [Comedies, Horror Movies]
8805   [Children & Family Movies, Comedies]
8806   [Dramas, International Movies, Music & Music...
Name: listed_in, Length: 5700, dtype: object
```

In [46]:

```
netflix_copy['Genre 1'] = new_genre.str.get(0)
netflix_copy['Genre 2'] = new_genre.str.get(1)
netflix_copy['Genre 3'] = new_genre.str.get(2)
netflix_copy.head(2)
```

Out[46]:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	year_added	month_added
2	s3 TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	missing	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...	2021	9
5	s6 TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	missing	2021-09-24	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	The arrival of a charismatic young priest brin...	2021	9

In [47]:

```
netflix_copy['Genre 1'].describe(include=all)
```

Out[47]:

```
count      5700
unique       33
top      Dramas
freq      1586
Name: Genre 1, dtype: object
```

In [48]:

```
netflix_copy['Genre 2'].describe(include=all)
```

Out[48]:

```
count      4485
unique       35
top  International Movies
freq      1545
Name: Genre 2, dtype: object
```

In [49]:

```
netflix_copy['Genre 3'].describe(include=all)
```

Out[49]:

```
count      2434
unique       29
top  International Movies
freq       881
Name: Genre 3, dtype: object
```

- drop listed_in column because not need

In [50]:

```
netflix_copy.drop('listed_in',axis = 1,inplace = True)
netflix_copy.head(2)
```

Out[50]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	description	year_added	month_added	Genre
														1
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	missing	2021-09-24	2021	TV-MA	Season 1	To protect his family from a powerful drug lor...	2021	9	Crime TV Shows
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	missing	2021-09-24	2021	TV-MA	Season 1	The arrival of a charismatic young priest brin...	2021	9	TV Dramas

In [51]:

```
netflix_copy.duplicated().sum()
```

Out[51]:

0

In [52]:

```
netflix_copy.groupby(['Genre 1'])['release_year'].count().sort_values(ascending = False)
```

Out[52]:

Genre 1	
Dramas	1586
Comedies	1184
Action & Adventure	848
Children & Family Movies	548
Documentaries	397
Stand-Up Comedy	309
Horror Movies	274
International Movies	122
Classic Movies	73
Thrillers	65
International TV Shows	59
Crime TV Shows	37
Movies	30
British TV Shows	21
Independent Movies	20
Anime Features	19
Music & Musicals	18
Sci-Fi & Fantasy	13
Cult Movies	12
Kids' TV	11
Anime Series	11
Docuseries	10
TV Shows	9
Stand-Up Comedy & Talk Shows	8
Romantic Movies	3
TV Comedies	3
Classic & Cult TV	2
TV Action & Adventure	2
TV Dramas	2
LGBTQ Movies	1
Romantic TV Shows	1
TV Horror	1
Reality TV	1

Name: release_year, dtype: int64

In [53]:

```
netflix_copy.groupby(['Genre 2'])['release_year'].count().sort_values(ascending = False)
```

Out[53]:

Genre 2	
International Movies	1545
Dramas	744
Independent Movies	572
Comedies	440
Thrillers	232
Romantic Movies	170
Music & Musicals	119
Sci-Fi & Fantasy	109
Horror Movies	71
Sports Movies	68
Anime Features	49
Faith & Spirituality	49
Cult Movies	44
International TV Shows	41
Children & Family Movies	31
Classic Movies	30
LGBTQ Movies	30
TV Dramas	23
Romantic TV Shows	22
Documentaries	21
TV Comedies	20
Korean TV Shows	9
Stand-Up Comedy	9
Crime TV Shows	8
Docuseries	7
TV Action & Adventure	7
TV Horror	3
Classic & Cult TV	2
Kids' TV	2
Reality TV	2
TV Sci-Fi & Fantasy	2
Spanish-Language TV Shows	1
Stand-Up Comedy & Talk Shows	1
TV Mysteries	1
Science & Nature TV	1

Name: release_year, dtype: int64

In [54]:

```
netflix_copy.groupby(['Genre 3'])['release_year'].count().sort_values(ascending = False)
```

Out[54]:

```
Genre 3
International Movies      881
Romantic Movies          431
Thrillers                 278
Music & Musicals          191
Independent Movies        159
Sci-Fi & Fantasy          115
Sports Movies             95
Dramas                   68
LGBTQ Movies             52
TV Dramas                36
TV Comedies              16
Cult Movies              14
Comedies                 12
Faith & Spirituality      11
Spanish-Language TV Shows 10
TV Mysteries             10
TV Action & Adventure      7
International TV Shows    7
Horror Movies             7
Classic Movies            6
Romantic TV Shows         5
TV Horror                 5
Children & Family Movies  4
Stand-Up Comedy & Talk Shows 3
TV Sci-Fi & Fantasy        3
TV Thrillers              3
Teen TV Shows             3
Kids' TV                  1
Korean TV Shows           1
Name: release_year, dtype: int64
```

4.6 Finding the best Month for releasing content?

In [55]:

```
netflix_copy.head(2)
```

Out[55]:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	description	year_added	month_added	Genre 1	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	missing	2021-09-24	2021	TV-MA	Season 1	To protect his family from a powerful drug lor...	2021	9	Crime TV Shows
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	missing	2021-09-24	2021	TV-MA	Season 1	The arrival of a charismatic young priest brin...	2021	9	TV Dramas

In [56]:

```
# converting month number to month name
netflix_copy['month_final'] = netflix_copy['month_added'].replace({1:'Jan', 2:'Feb', 3:'Mar', 4:'Apr', 5:'May', 6:'June', 7:'July', 8:'Aug'})
netflix_copy.head(2)
```

Out[56]:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	description	year_added	month_added	Genre 1	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	missing	2021-09-24	2021	TV-MA	Season 1	To protect his family from a powerful drug lor...	2021	9	Crime TV Shows
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	missing	2021-09-24	2021	TV-MA	Season 1	The arrival of a charismatic young priest brin...	2021	9	TV Dramas

In [57]:

```
netflix_copy.month_final.value_counts().to_frame('Value_count')
```

Out[57]:

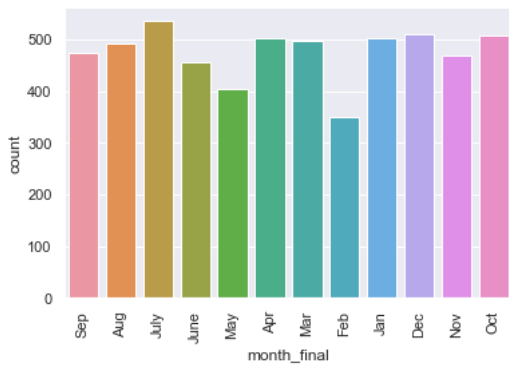
	Value_count
July	535
Dec	511
Oct	507
Jan	503
Apr	501
Mar	497
Aug	492
Sep	475
Nov	468
June	457
May	405
Feb	349

In [58]:

```
sns.countplot(x=netflix_copy['month_final'],orient='v')
plt.xticks(rotation=90)
```

Out[58]:

```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11]),
 [Text(0, 0, 'Sep'),
  Text(1, 0, 'Aug'),
  Text(2, 0, 'July'),
  Text(3, 0, 'June'),
  Text(4, 0, 'May'),
  Text(5, 0, 'Apr'),
  Text(6, 0, 'Mar'),
  Text(7, 0, 'Feb'),
  Text(8, 0, 'Jan'),
  Text(9, 0, 'Dec'),
  Text(10, 0, 'Nov'),
  Text(11, 0, 'Oct')])
```



Observation: Its clearly shows that **july** month has maximum number of movies released.july is the best Month for releasing content.

4.7 Highest watched genres on Netflix?

In [59]:

```
netflix_copy.head(2)
```

Out[59]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	description	year_added	month_added	Genre 1
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	missing	2021-09-24	2021	TV-MA	1 Season	To protect his family from a powerful drug lor...	2021	9	Crime TV Shows
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	missing	2021-09-24	2021	TV-MA	1 Season	The arrival of a charismatic young priest brin...	2021	9	TV Dramas

In [60]:

```
netflix_copy['Genre 1'].describe(include=all)
```

Out[60]:

```
count      5700
unique       33
top      Dramas
freq      1586
Name: Genre 1, dtype: object
```

In [61]:

```
netflix_copy['Genre 2'].describe(include=all)
```

Out[61]:

```
count      4485
unique       35
top      International Movies
freq      1545
Name: Genre 2, dtype: object
```

In [62]:

```
netflix_copy['Genre 3'].describe(include=all)
```

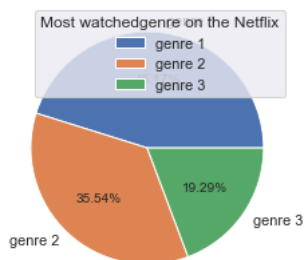
Out[62]:

```
count      2434
unique       29
top      International Movies
freq       881
Name: Genre 3, dtype: object
```

In [63]:

```
# by using Pie chart
genre_types = ['genre 1', 'genre 2', 'genre 3']

Value_count = [5700, 4485, 2434]
plt.pie(Value_count, labels = genre_types, autopct="%2.2f%")
plt.legend(title='Most watchedgenre on the Netflix')
plt.show()
```



Obeservation: Above Pi-chart Shows that Genre1 has watched maximum times.

4.8 Released movie over the years?

In [64]:

```
netflix_copy.head(2)
```

Out[64]:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	description	year_added	month_added	Genre
2	s3 TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	missing	2021-09-24	2021	TV-MA	1 Season	To protect his family from a powerful drug lor...	2021	9	Crime TV Shows
5	s6 TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	missing	2021-09-24	2021	TV-MA	1 Season	The arrival of a charismatic young priest brin...	2021	9	TV Dramas

In [65]:

```
# no. of movies released by every year
netflix_copy.groupby(['release_year'])['release_year'].count().sort_values(ascending = False)
```

Out[65]:

release_year	
2018	691
2017	684
2016	584
2019	570
2020	499
...	
1944	1
1947	1
1946	1
1945	1
1942	1

Name: release_year, Length: 72, dtype: int64

In [66]:

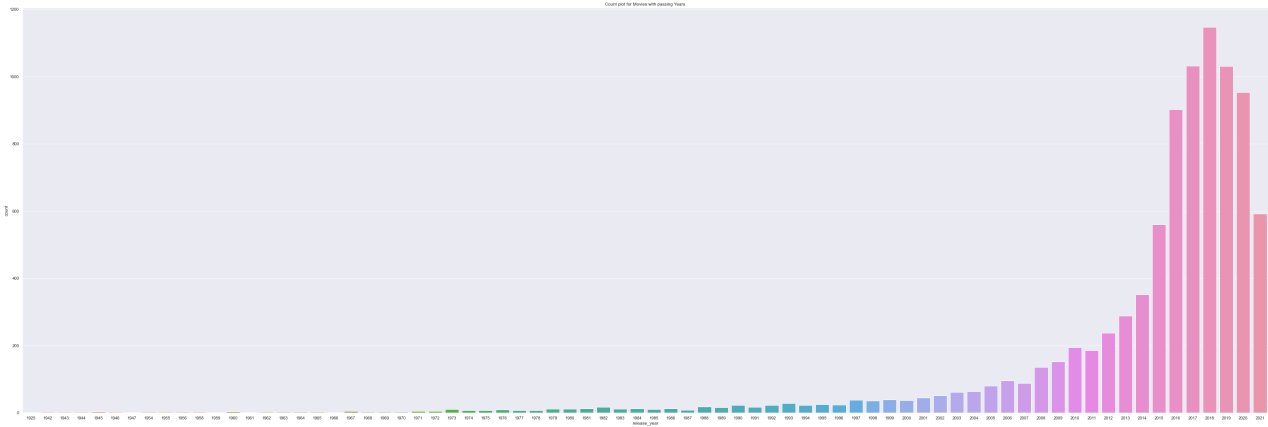
```
netflix_copy.release_year.value_counts().to_frame('Value_count')
```

Out[66]:

	Value_count
2018	691
2017	684
2016	584
2019	570
2020	499
...	...
1946	1
1959	1
1942	1
1947	1
1944	1

72 rows × 1 columns

```
In [68]:
sns.countplot(x='release_year', data=netflix).set_title('Count plot for Movies with passing Years.')
sns.set(rc={'figure.figsize':(40,20)})
plt.show()
```



Observation:

From the above we can see that with passing years more movies are being made year by year

The above data shows that there is a sudden increase in the creation of movies in year 2016 as compared to year 2015

4.9 movies made on year basis? or

In which year highest number of the TV shows & Movies were released? show with Bar Graph.

```
In [69]:
netflix_copy.head(2)
```

Out[69]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	description	year_added	month_added	Genre
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	missing	2021-09-24	2021	TV-MA	1 Season	To protect his family from a powerful drug lor...	2021	9	Crime TV Shows
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, H...	missing	2021-09-24	2021	TV-MA	1 Season	The arrival of a charismatic young priest brin...	2021	9	TV Dramas

In [70]:

```
netflix_copy.year_added.value_counts().to_frame('Value_count')
```

Out[70]:

	Value_count
2019	1323
2020	1248
2018	1125
2021	962
2017	744
2016	206
2015	51
2014	14
2011	13
2013	7
2012	3
2009	2
2008	1
2010	1

In [71]:

```
netflix_copy.dtypes
```

Out[71]:

```
show_id      object
type         object
title        object
director     object
cast         object
country      object
date_added   datetime64[ns]
release_year  int64
rating       object
duration     object
description   object
year_added   int64
month_added  int64
Genre_1      object
Genre_2      object
Genre_3      object
month_final  object
dtype: object
```

In [72]:

```
netflix_copy["date_added"].dt.year.value_counts()
```

Out[72]:

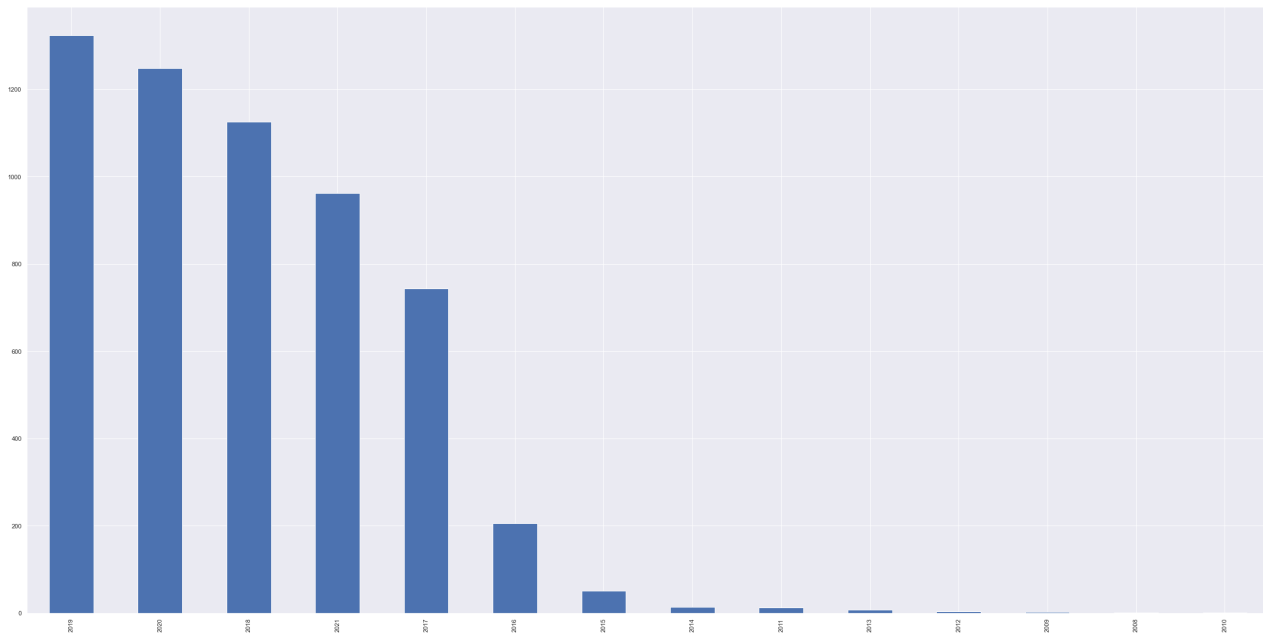
```
2019    1323
2020    1248
2018    1125
2021     962
2017     744
2016     206
2015      51
2014      14
2011      13
2013       7
2012       3
2009       2
2008       1
2010       1
Name: date_added, dtype: int64
```


In [73]:

```
netflix_copy["date_added"].dt.year.value_counts().plot(kind='bar')
```

Out[73]:

<AxesSubplot:>



Obeservation: Its clearly shows that maximum movie made on yr 2019.

4.10 What is the show id and director for 'House of cards'?

In [74]:

```
netflix.head(2)
```

Out[74]:

	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 Mabalanane, 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...

In [75]:

```
# method 1:
netflix[netflix['title'].isin(["House of Cards"])] # to shows all records of a particular item in any column
```

Out[75]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
1059	s1060	TV Show	House of Cards	NaN	Kevin Spacey, Robin Wright, Kate Mara, Corey S...	United States	April 14, 2021	2018	TV-MA	6 Seasons	TV Dramas, TV Thrillers	A ruthless politician will stop at nothing to ...

In [76]:

```
# method 2:
netflix[netflix['title'].str.contains("House of Cards")]
```

Out[76]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
1059	s1060	TV Show	House of Cards	NaN	Kevin Spacey, Robin Wright, Kate Mara, Corey S...	United States	April 14, 2021	2018	TV-MA	6 Seasons	TV Dramas, TV Thrillers	A ruthless politician will stop at nothing to ...

Obeservation: For "House of cards" the show id is s1059 and No director.

4.11 Show all the movies that were released in year 2000.

In [77]:

```
netflix.head(2)
```

Out[77]:

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...

In [78]:

```
netflix[(netflix['type'] == 'Movie' ) & (netflix['release_year'] == 2000)]
```

Out[78]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
180	s181	Movie	The Nutty Professor II: The Klumps	Peter Segal	Eddie Murphy, Janet Jackson, Larry Miller, Joh...	NaN	September 1, 2021	2000	PG-13	107 min	Comedies, Romantic Movies	After getting engaged, Sherman Klump prepares ...
350	s351	Movie	Space Cowboys	Clint Eastwood	Clint Eastwood, Tommy Lee Jones, Donald Suther...	United States	August 1, 2021	2000	PG-13	130 min	Action & Adventure, Dramas, Sci-Fi & Fantasy	A retired engineer agrees to help NASA prevent...
359	s360	Movie	The Original Kings of Comedy	Spike Lee	Steve Harvey, D.L. Hughley, Cedric the Enterta	United States	August 1, 2021	2000	R	111 min	Stand-Up Comedy	Comedians Steve Harvey, Cedric the Entertainer.

In [79]:

```
netflix[(netflix['type'] == 'Movie' ) & (netflix['release_year'] == 2000)].shape
```

Out[79]:

(33, 12)

Obeservation: There are 33 movies released in year 2000.

4.12 Show only the title of all TV shows that were released in India only.

In [80]:

```
netflix.head(2)
```

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	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...

In [81]:

```
netflix[(netflix['type'] == 'TV Show' ) & (netflix['country'] == 'India')]["title"]
```

Out[81]:

4 Kota Factory
39 Chhota Bheem
50 Dharmakshetra
66 Raja Rasoi Aur Anya Kahaniyan
69 Stories by Rabindranath Tagore
...
8173 Thackeray
8235 The Calling
8321 The Golden Years with Javed Akhtar
8349 The House That Made Me
8775 Yeh Meri Family
Name: title, Length: 79, dtype: object

In [82]:

```
netflix[(netflix['type'] == 'TV Show') & (netflix['country'] == 'India')]['title'].shape
```

Out[82]:

(79,)

Observation: There are **79 tv shows** that were released in India only.

4.13 Show top 10 director, who gave the highest number of TV shows & Movies to Netflix?

In [83]:

```
netflix.head(2)
```

Out[83]:

	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...

In [84]:

```
netflix["director"].value_counts().head(10)
```

Out[84]:

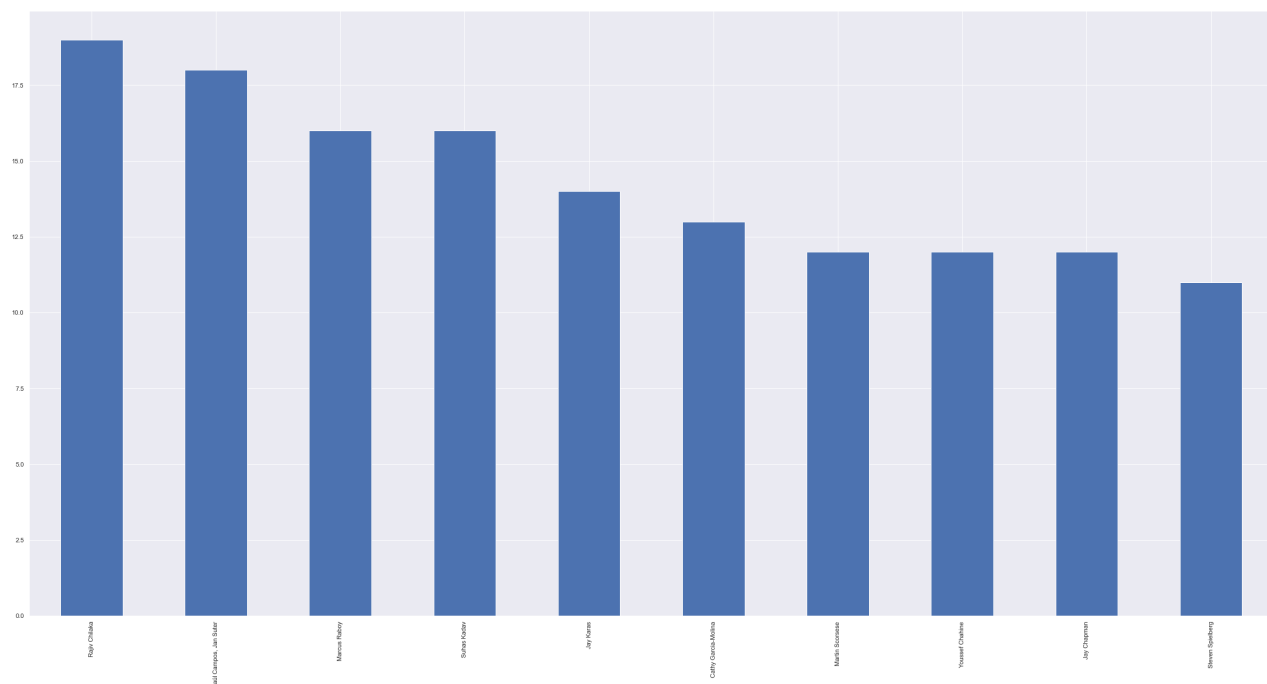
```
Rajiv Chilaka      19
Raúl Campos, Jan Suter  18
Marcus Raboy       16
Suhas Kadav        16
Jay Karas          14
Cathy Garcia-Molina 13
Martin Scorsese     12
Youssef Chahine     12
Jay Chapman         12
Steven Spielberg    11
Name: director, dtype: int64
```

In [85]:

```
# shown in bar graph
netflix["director"].value_counts().head(10).plot(kind='bar')
```

Out[85]:

<AxesSubplot:>



Observation : Top 10 director list is shown .Rajiv Chilaka he released maximum no. of Tv shows in india

4.14 In how many movies/ tv shows, 'tom Cruise' was cast?

In [86]:

```
netflix.head(2)
```

Out[86]:

	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 Mabalané, 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...

In [87]:

```
#method1:
netflix[netflix['cast'] == 'Tom Cruise']
```

Out[87]:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
---------	------	-------	----------	------	---------	------------	--------------	--------	----------	-----------	-------------

In [88]:

```
#method2:
netflix[netflix['cast'].str.contains('Tom Cruise')]
```

ValueError Traceback (most recent call last)

Input In [88], in <cell line: 3>()

```
1 #method2:
----> 3 netflix[netflix['cast'].str.contains('Tom Cruise')]
```

File F:\my apps\anaconda\conda\lib\site-packages\pandas\core\frame.py:3495, in DataFrame.__getitem__(self, key)

```
3492 return self.where(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 F:\my apps\anaconda\conda\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) == "boolean" and isna(key).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

Nan values are present, first fill null values

- create new dataframe

In [89]:

```
netflix_copy = netflix.dropna()
netflix_copy.head(2)
```

Out[89]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
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 photo shoot in Ghana, an American model s...
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 talented batch of amateur bakers face off in...

In [90]:

```
netflix_copy[netflix_copy['cast'].str.contains('Tom Cruise')]
```

Out[90]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
341	s342	Movie	Magnolia	Paul Thomas Anderson	John C. Reilly, Philip Baker Hall, Tom Cruise,...	United States	August 1, 2021	1999	R	189 min	Dramas, Independent Movies	Through chance, history and divine interventio...
1254	s1255	Movie	Rain Man	Barry Levinson	Dustin Hoffman, Tom Cruise, Valeria Golino, Ge...	United States	March 1, 2021	1988	R	134 min	Classic Movies, Dramas	Motivated by money, a selfish workaholic seeki...

Obeservation: There are two movies casted Mr.Tom Cruise -"Magnolia" & "Rain Man" .

4.15 How many movies got the "TV-14" rating in the canada?

In [91]:

```
netflix.head(2)
```

Out[91]:

	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 Mablane, 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...

In [92]:

```
netflix[(netflix['type'] == 'Movie') & (netflix['rating'] == 'TV-14')].shape
```

Out[92]:

(1427, 12)

In [93]:

```
netflix[(netflix['type'] == 'Movie') & (netflix['rating'] == 'TV-14') & (netflix['country'] == 'Canada')]
```

Out[93]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
157	s158	Movie	Level 16	Danishka Esterhazy	Katie Douglas, Celina Martin, Peter Outerbridge...	Canada	September 1, 2021	2018	TV-14	102 min	Sci-Fi & Fantasy, Thrillers	In a bleak academy that teaches girls the virt...
1279	s1280	Movie	Made You Look: A True Story About Fake Art	Barry Avrich	NaN	Canada	February 23, 2021	2020	TV-14	90 min	Documentaries, International Movies	A woman walks into a New York gallery with a c...
2769	s2770	Movie	There's Something in the Water	Ellen Page, Ian Daniel	Ellen Page	Canada	March 27, 2020	2019	TV-14	72 min	Documentaries, International Movies	This documentary spotlights the struggle of mi...
3424	s3425	Movie	Martin Matte: La vie, la mort...eh la la..!	Jean-François Blais	Martin Matte	Canada	October 14, 2019	2019	TV-14	62 min	Stand-Up Comedy	Quebecois comedy star Martin Matte serves up e...
5972	s5973	Movie	#cats_the_mewvie	Michael Margolis	NaN	Canada	February 5, 2020	2020	TV-14	90 min	Documentaries, International Movies	This pawesome documentary explores how our fel...
5973	s5974	Movie	#Roxy	Michael Kennedy	Jake Short, Sarah Fisher, Booboo Stewart, Dann...	Canada	April 10, 2019	2018	TV-14	105 min	Comedies, Romantic Movies	A teenage hacker with a huge nose helps a cool...
6192	s6193	Movie	At First Light	Jason Stone	Stefanie Scott, Theodore Pellerin, Percy Hynes...	Canada	December 27, 2018	2018	TV-14	91 min	Dramas, Sci-Fi & Fantasy	After mysterious lights imbue her with extraor...
6537	s6538	Movie	Counterfeiting in Suburbia	Jason Bourque	Sarah Butler, Larissa Albuquerque, Kayla Walla...	Canada	July 1, 2018	2018	TV-14	88 min	Dramas, Thrillers	Two teenagers print counterfeit money in their...
7334	s7335	Movie	Long Time Running	Jennifer Baichwal, Nicholas de Pencier	Gord Downie, Paul Langlois, Gord Sinclair, Rob...	Canada	November 12, 2017	2017	TV-14	96 min	Documentaries, International Movies, Music & M...	This documentary follows the 2016 tour of belo...
7916	s7917	Movie	Sadie's Last Days on Earth	Michael Seater	Morgan Taylor Campbell, Clark Backo, Ricardo H...	Canada	August 22, 2017	2016	TV-14	94 min	Comedies, Sci-Fi & Fantasy	High schooler Sadie Mitchell believes the worl...
7936	s7937	Movie	Sardaar ji	Rohit Jugraj	Diljit Dosanjh, Neeru Bajwa, Mandy Takhar, Jas...	Canada	December 1, 2017	2015	TV-14	140 min	Comedies, International Movies, Music & Musicals	When a palatial wedding venue turns out to be ...
8178	s8179	Movie	The Accountant of Auschwitz	Matthew Shoychet	NaN	Canada	July 1, 2019	2018	TV-14	78 min	Documentaries, International Movies	Decades after WWII, a former SS officer stands...
8566	s8567	Movie	The Wrong Babysitter	George Mendeluk	Daphne Zuniga, Seann Gallagher, Lisa Marie Car...	Canada	June 18, 2019	2017	TV-14	89 min	Thrillers	A widow searches for her missing teenage daugh...

In [94]:

```
netflix[(netflix['type'] == 'Movie') & (netflix['rating'] == 'TV-14') & (netflix['country'] == 'Canada')].shape
```

Out[94]:

(13, 12)

Insights: There are 13 movies got the "TV-14" rating in the caneda.

Post profiling

In [95]:

```
netflix_cleaned = prf.ProfileReport(netflix_copy)
netflix_cleaned.to_file(output_file="netflix21_post_preprocessing.html")
```

C:\Users\user\AppData\Roaming\Python\Python39\site-packages\pandas_profiling\utils\dataframe.py:137: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy)
df.rename(columns={"index": "df_index"}, inplace=True)

Summarize dataset: 100% 22/22 [00:06<00:00, 2.84it/s, Completed]

Generate report structure: 100% 1/1 [00:08<00:00, 8.60s/it]

Render HTML: 100% 1/1 [00:03<00:00, 3.32s/it]

Export report to file: 100% 1/1 [00:00<00:00, 6.02it/s]

In [96]:

```
netflix_cleaned.to_file(output_file="netflix21_post_preprocessing.html")
```

Export report to file: 100% 1/1 [00:00<00:00, 5.81it/s]

5. Conclusions:

We explore the Netflix dataset and saw how to clean the data and then jump into how to visualize the data with **Exploratory Data Analysis**. We saw some basic and advanced level charts of seaborn and matplotlib like Heatmap, Pie-chart, Bar chart, Countplot. And also visualize a data with **Pandas_Profiling :Preprofiling & Post_profiling**.

Thank You! Happy to get any suggestions or feedbacks.