Import Libraries

In [41]:

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

Load Data

In [42]:

```
netflix_df = pd.read_csv('https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/
netflix_df
```

Out[42]:

	show_id	type	title	director	cast	country	date_added	release_year	rating
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA
...	...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	R

	show_id	type	title	director	cast	country	date_added	release_year	rating
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	PG
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-14

8807 rows × 12 columns

2. Observations on the shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), missing value detection, statistical summary (10 Points)

shape of data

In [43]:

```
netflix_df.shape
```

Out[43]:

```
(8807, 12)
```

In [44]:

```
netflix_df['type'].value_counts()
```

Out[44]:

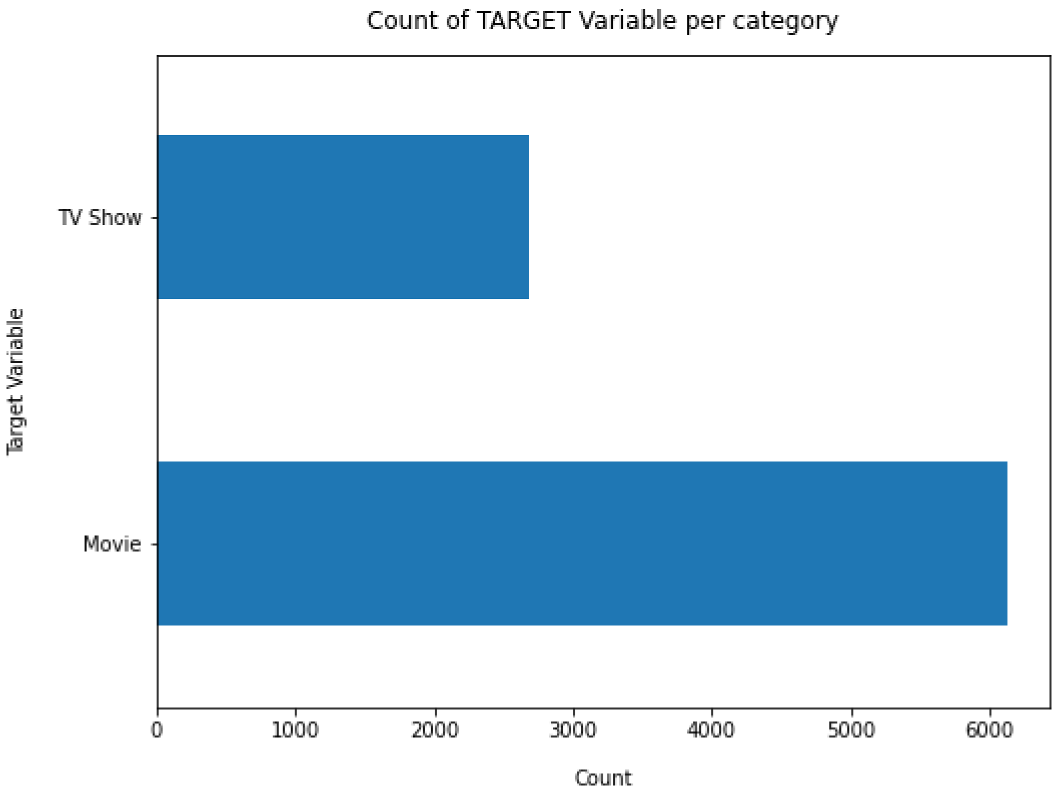
```
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

In [45]:

```
netflix_df['type'].value_counts().plot(kind='barh', figsize=(8, 6))
plt.xlabel("Count", labelpad=14)
plt.ylabel("Target Variable", labelpad=14)
plt.title("Count of TARGET Variable per category", y=1.02)
```

Out[45]:

```
Text(0.5, 1.02, 'Count of TARGET Variable per category')
```



data types of all the attributes

In [46]:
```
netflix_df.info()
```

```
<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
```

missing value detection

In [47]:
```
netflix_df.isnull().sum()
```

Out[47]:

```
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
```

statistical summary

In [48]:

```
netflix_df.describe()
```

Out[48]:

	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 [49]:

```
netflix_df.describe(include= 'object')
```

Out[49]:

	show_id	type	title	director	cast	country	date_added	rating	duration	
count	8807	8807	8807	6173	7982	7976	8797	8803	8804	
unique	8807	2	8807	4528	7692	748	1767	17	220	
top	s1	Movie	Dick Johnson Is Dead	Rajiv Chilaka	David Attenborough	United States	January 1, 2020	TV-MA	1 Season	Int
freq	1	6131	1	19	19	2818	109	3207	1793	

3. Non-Graphical Analysis: Value counts and unique attributes

In [50]:

```
netflix_df.value_counts()
```

```
Out[50]: show_id  type   title                director            cast
         country                                                  date_added          release_y
         ear  rating  duration  listed_in                                           description
         s10      Movie  The Starling         Theodore Melfi      Melissa McCarthy, Chris O'D
         owd, Kevin Kline, Timothy Olyphant, Daveed Diggs, Skyler Gisondo, Laura Harrier, Rosa
         lind Chao, Kimberly Quinn, Loretta Devine, Ravi Kapoor         United States
         September 24, 2021  2021          PG-13   104 min   Comedies, Dramas
         A woman adjusting to life after a loss contends with a feisty bird that's taken over
         her garden — and a husband who's struggling to find a way forward.     1
         s655     Movie  #Selfie              Cristina Jacob      Flavia Hojda, Crina Semciu
         c, Olimpia Melinte, Sali Levent, Vlad Logigan, Alex Călin, Alina Chivulescu, Răzvan V
         asilescu                                                         Romania
         June 21, 2021       2014          TV-MA   125 min   Comedies, Dramas, International M
         ovies            Two days before their final exams, three teen girls make a seaside g
         etaway to have the time of their lives.
         1
         s6548    Movie  Dabbe 6: The Return  Hasan Karacadağ     Sema Şimşek, Nilay Gök, Vol
         kan Ünal, Fehmi Karaarslan, Elçin Atamgüç, Ömer Duran, Murat Seviş, Aybike Turan, Bur
         ak Çimen                                                        Turkey
         April 17, 2019      2015          TV-MA   153 min   Horror Movies, International Movi
         es               A cardiologist tries to pinpoint the cause of her mother's sudden de
         ath as her sister, who witnessed it, claims malevolent demons are at play.
         1
         s6547    Movie  Daagdi Chaawl        Chandrakant Kanse    Ankush Chaudhari, Makrand D
         eshpande, Pooja Sawant, Sanjay Khapre, Yatin Karyekar, Kamlesh Sawant, Sandeep Gaikwa
         d, Digvijay Rohidas                                             India
         February 15, 2018   2015          TV-14   118 min   Action & Adventure, Dramas, Inter
         national Movies  A simple man's peaceful life is complicated when an incident brings
         him in contact with a gangster and launches his journey into the underworld.
         1
         s6546    Movie  Cutie and the Boxer  Zachary Heinzerling  Noriko Shinohara, Ushio Shi
         nohara
         United States                                            June 14, 2018       2013
         R       82 min    Documentaries                                      A 2014 Oscar nomi
         nee for Best Documentary Feature, this film explores the symbiotic relationship of ar
         tists Ushio and Noriko Shinohara.                         1

         ..
         s390     Movie  The Operative        Yuval Adler          Diane Kruger, Martin Freema
         n, Cas Anvar, Rotem Keinan, Yohanan Herson
         France, Israel, Germany, United States, United Kingdom  July 27, 2021       2019
         TV-MA   117 min   Dramas, International Movies, Thrillers           Working as a Moss
         ad spy assigned to Tehran, Rachel reaches her breaking point. Now it's dangerously ha
         rd to tell whose side she's on.                           1
         s39      Movie  Birth of the Dragon  George Nolfi         Billy Magnussen, Ron Yuan,
         Qu Jingjing, Terry Chen, Vanness Wu, Jin Xing, Philip Ng, Xia Yu, Yu Xia
         China, Canada, United States                             September 16, 2021  2017
         PG-13   96 min    Action & Adventure, Dramas                         A young Bruce Lee
         angers kung fu traditionalists by teaching outsiders, leading to a showdown with a Sh
         aolin master in this film based on real events.      1
         s3895    Movie  A Fortunate Man      Bille August         Esben Smed, Katrine Rosenth
         al, Benjamin Kitter, Julie Christiansen, Tommy Kenter, Tammi Øst, Rasmus Bjerg, Ole L
         emmeke, Sarah Viktoria Bjerregaard, Anders Hove, Jens Albinus  Denmark
         April 20, 2019      2018          TV-MA   168 min   Dramas, International Movies
         A gifted engineer flees his austere roots to pursue wealth and success among Copenhag
         en's elite, but the pride propelling him threatens to be his ruin.     1
         s3891    Movie  Njan Prakashan       Sathyan Anthikad     Fahadh Faasil, Sreenivasan,
         Nikhila Vimal, Devika Sanjay, Anju Kurian, K.P.A.C. Lalitha
         India                                                    April 26, 2019      2018
         TV-PG   125 min   Comedies, Dramas, International Movies            Yearning for a la
```

```
vish life abroad, an entitled, lazy sexist crafts a scam to ditch his thankless nursi
ng job and find a wealthy spouse to secure a visa.    1
s997     Movie  HOMUNCULUS          Takashi Shimizu      Go Ayano, Ryo Narita, Yukin
o Kishii, Anna Ishii, Seiyo Uchino
Japan                                                   April 22, 2021      2021
TV-MA   116 min   Horror Movies, International Movies, Thrillers    Truth and illusio
n blurs when a homeless amnesiac awakens from an experimental medical procedure with
the ability to see people's innermost traumas.        1
Length: 5332, dtype: int64
```

In [51]:

```python
netflix_df.nunique()
```

Out[51]:

```
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
```

4. Visual Analysis - Univariate, Bivariate after pre-processing of the data

Pre-processing (Unnesting)

In [52]:

```python
netflix_df.dropna(inplace=True)
```

In [53]:

```python
# df1 = netflix_df['cast'].str.split(',', expand=True)
# cast_columns_names = ['cast-' + str(col) for col in df1.columns]
# df1.columns  = cast_columns_names
# cast_temp_df = pd.concat( [netflix_df, df1], axis=1)


# df1 = cast_temp_df.melt(id_vars = 'cast', value_vars = cast_columns_names)
# df1.dropna(inplace=True)
# netflix_df = netflix_df.merge(df1, how='inner', on='cast')
# netflix_df.drop(columns='cast', inplace=True)
# netflix_df
```

In [54]:

```python
# df1 = netflix_df['director'].str.split(',', expand=True)
# director_columns_names = ['director-' + str(col) for col in df1.columns]
# df1.columns  = director_columns_names
# temp_df = pd.concat( [netflix_df, df1], axis=1)
# df1 = temp_df.melt(id_vars = 'director', value_vars = director_columns_names)
# df1.dropna(inplace=True)
# netflix_df = netflix_df.merge(df1, how='inner', on='director')
# netflix_df.drop(columns='director', inplace=True)
# netflix_df
```

In [55]:

```python
netflix_df
```

Out[55]:

	show_id	type	title	director	cast	country	date_added	release_year	rati
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	T N
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-
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, T...	United States	September 24, 2021	2021	PG-
12	s13	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler, Jannis Niewöhner, Milan Peschel, ...	Germany, Czech Republic	September 23, 2021	2021	T N
24	s25	Movie	Jeans	S. Shankar	Prashanth, Aishwarya Rai Bachchan, Sri Lakshmi...	India	September 21, 2021	1998	TV-
...	...	...	...	...	...	...	...	...	
8801	s8802	Movie	Zinzana	Majid Al Ansari	Ali Suliman, Saleh Bakri, Yasa, Ali Al-Jabri, ...	United Arab Emirates, Jordan	March 9, 2016	2015	T N
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	November 1, 2019	2009	

	show_id	type	title	director	cast	country	date_added	release_year	rati
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	January 11, 2020	2006	
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	March 2, 2019	2015	TV-

5332 rows × 12 columns

In [56]:

```
# df1 = netflix_df['country'].str.split(',', expand=True)
# country_columns_names = ['country-' + str(col) for col in df1.columns]
# df1.columns  = country_columns_names
# temp_df = pd.concat( [netflix_df, df1], axis=1)
# df1 = temp_df.melt(id_vars = 'country', value_vars = country_columns_names)
# df1.dropna(inplace=True)
# netflix_df = netflix_df.merge(df1, how='inner', on='country')
# netflix_df.drop(columns='country', inplace=True)
# netflix_df
```

In [57]:

```
listed_in_df = netflix_df['listed_in'].str.split(',', expand=True)
listed_in_columns_names = ['listed_in-' + str(col) for col in listed_in_df.columns]
listed_in_df.columns  = listed_in_columns_names
temp_df = pd.concat( [netflix_df, listed_in_df], axis=1)
listed_in_df = temp_df.melt(id_vars = 'listed_in', value_vars = listed_in_columns_name
listed_in_df.dropna(inplace=True)
listed_in_df
netflix_df = netflix_df.merge(listed_in_df, how='inner', on='listed_in')
netflix_df.drop(columns=['listed_in', 'variable'], inplace=True)
netflix_df.rename(columns = {'value':'listed_in'}, inplace = True)
netflix_df
```

Out[57]:

	show_id	type	title	director	cast	country	date_added	release_year	ra
0	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	
1	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	
2	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	
3	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	
4	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993	
...	...	...	...	...	...	...	...	...	
1338601	s8780	Movie	Yes or No 2.5	Kirati Nakintanon	Supanart Jittaleela, Pimpakan Bangchawong, Cha...	Thailand	November 8, 2018	2015	TV
1338602	s8780	Movie	Yes or No 2.5	Kirati Nakintanon	Supanart Jittaleela, Pimpakan Bangchawong, Cha...	Thailand	November 8, 2018	2015	TV
1338603	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	

	show_id	type	title	director	cast	country	date_added	release_year	ra
1338604	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	
1338605	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	

1338606 rows × 12 columns

In [58]:

```
## Make a Season Column an Integer
netflix_df["Season"]=""
netflix_df["Season"] = netflix_df[netflix_df["duration"].str.contains("Season")]["dura

netflix_df["Season"] = netflix_df["Season"].fillna("0")
netflix_df["Season"] = netflix_df["Season"].str.replace("Season", "").str.replace("s",
netflix_df["Season"] = netflix_df["Season"].astype(str).astype(int)

## Make a duration column an Integer
netflix_df["duration"]= netflix_df.duration.str.replace('^(\d+)(.Seasons*)$', "0") ##
netflix_df["duration"]= netflix_df["duration"].str.replace(" min", "") ## Remove min
netflix_df["duration"]= netflix_df["duration"].astype(int) ## Convert to Integer

netflix_df
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:10: FutureWarning: The
default value of regex will change from True to False in a future version.
  # Remove the CWD from sys.path while we load stuff.
```

Out[58]:

	show_id	type	title	director	cast	country	date_added	release_year
0	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993
1	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993
2	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993
3	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993
4	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandra D...	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	1993
...	...	...	...	...	...	...	...	...
1338601	s8780	Movie	Yes or No 2.5	Kirati Nakintanon	Supanart Jittaleela, Pimpakan Bangchawong, Cha...	Thailand	November 8, 2018	2015
1338602	s8780	Movie	Yes or No 2.5	Kirati Nakintanon	Supanart Jittaleela, Pimpakan Bangchawong, Cha...	Thailand	November 8, 2018	2015
1338603	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007

	show_id	type	title	director	cast	country	date_added	release_year
1338604	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007
1338605	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007

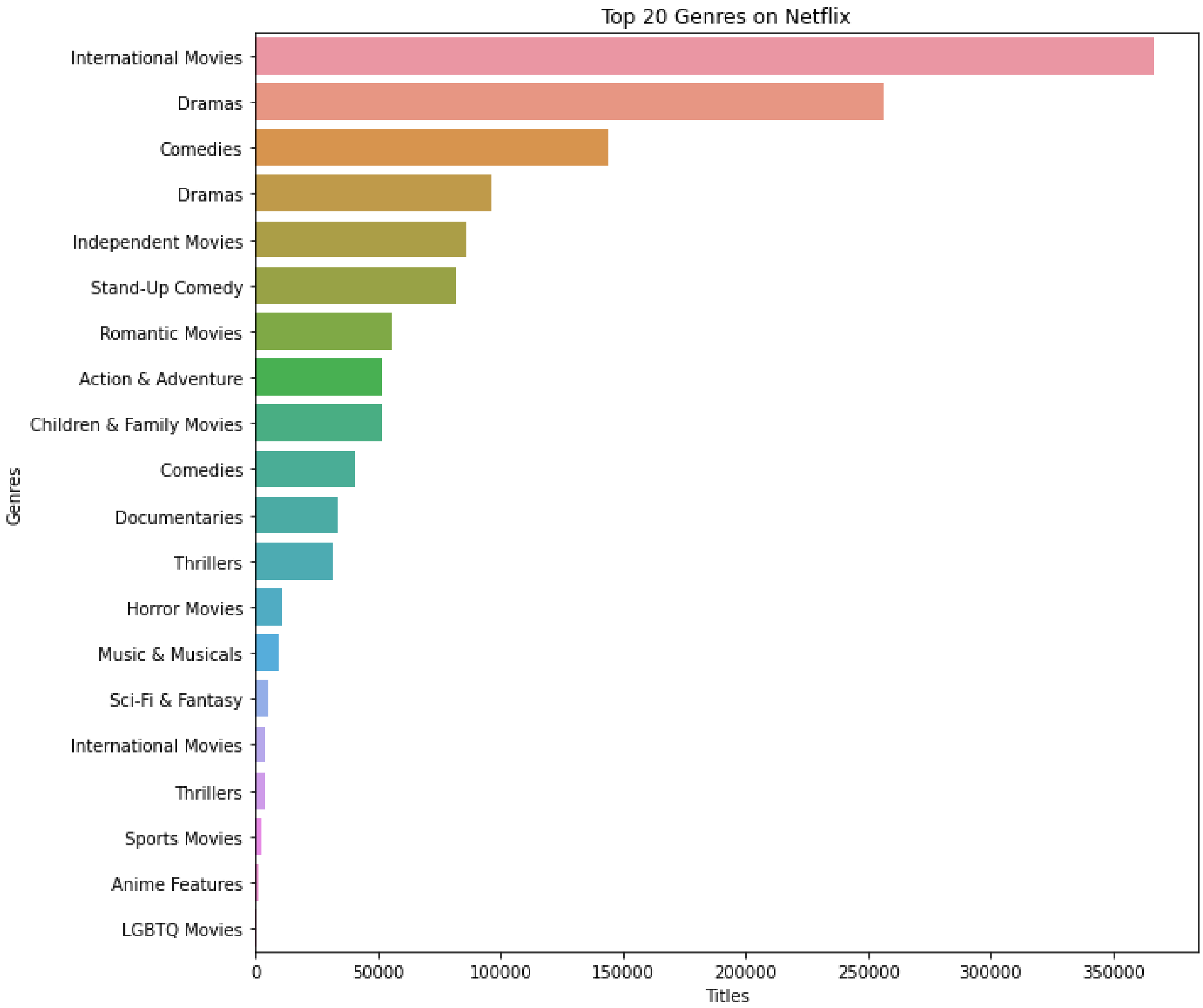
1338606 rows × 13 columns

Univariate analysis

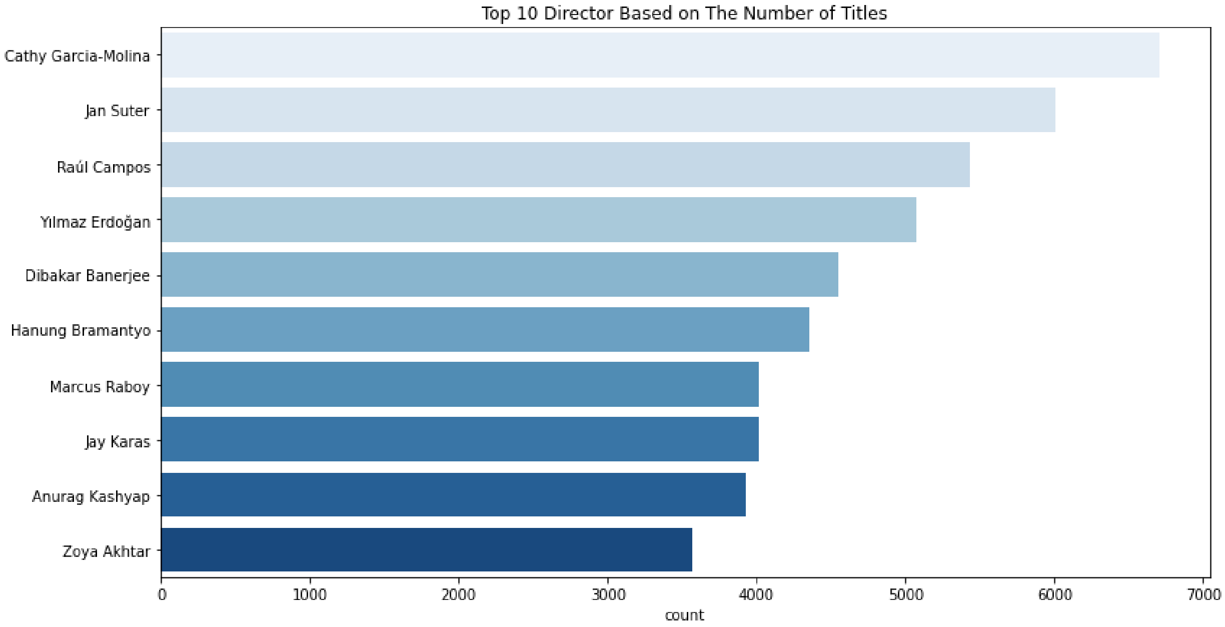
Count plot

In [59]:

```
plt.figure(figsize=(10,10))
g = sns.countplot(y = netflix_df['listed_in'], order=netflix_df['listed_in'].value_cou
plt.title('Top 20 Genres on Netflix')
plt.xlabel('Titles')
plt.ylabel('Genres')
plt.show()
```

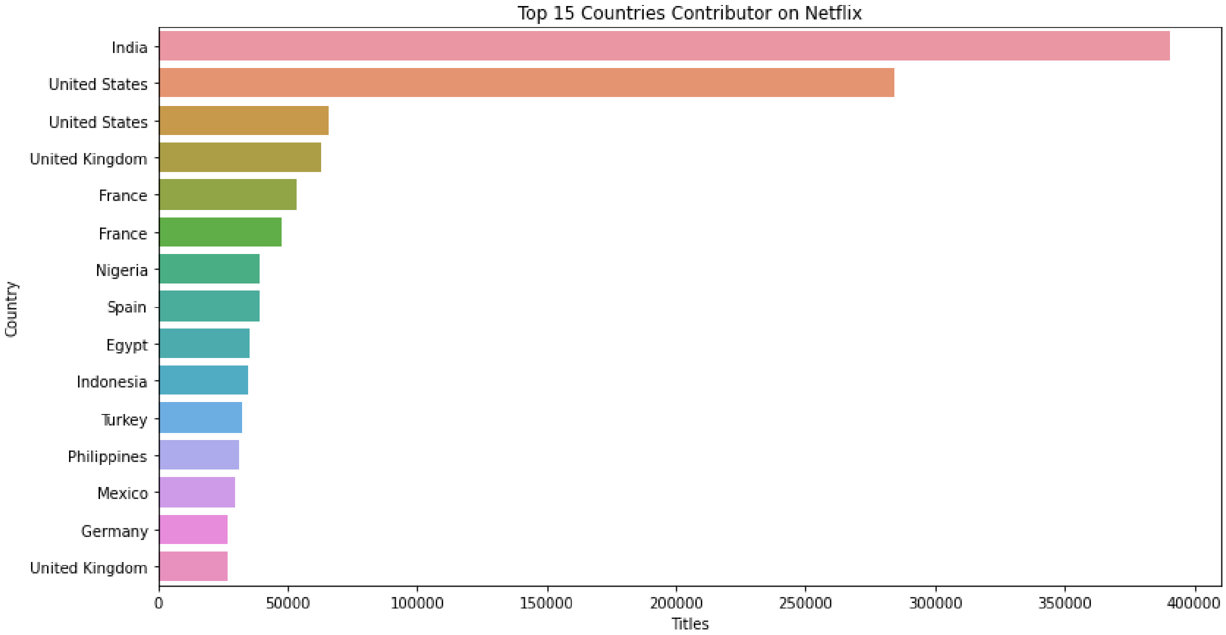


```
In [60]: filtered_directors = netflix_df[netflix_df.director != 'No Director'].set_index('title
         plt.figure(figsize=(13,7))
         plt.title('Top 10 Director Based on The Number of Titles')
         sns.countplot(y = filtered_directors, order=filtered_directors.value_counts().index[:1
         plt.show()
```



By Country

In [61]:
```
filtered_countries = netflix_df.set_index('title').country.str.split(',' , expand=True
filtered_countries = filtered_countries[filtered_countries != 'Country Unavailable']
plt.figure(figsize=(13,7))
g = sns.countplot(y = filtered_countries, order=filtered_countries.value_counts().inde
plt.title('Top 15 Countries Contributor on Netflix')
plt.xlabel('Titles')
plt.ylabel('Country')
plt.show()
```

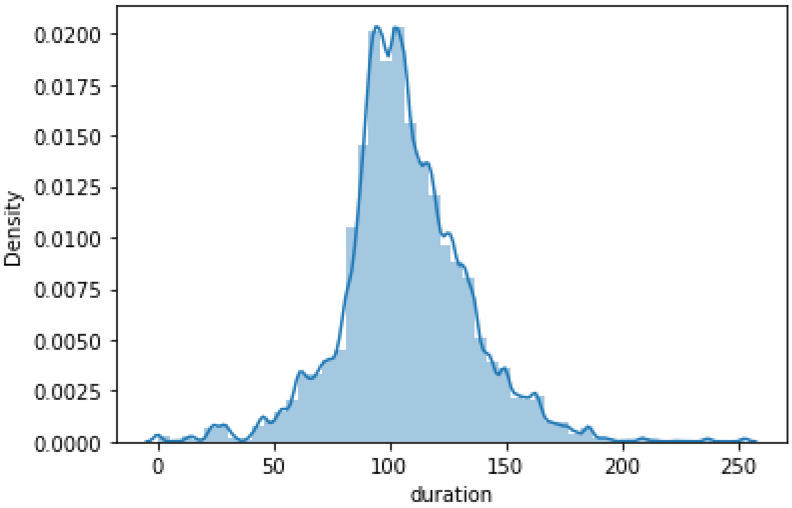


Dist plot

In [24]:
```
sns.distplot(netflix_df['duration'])
```

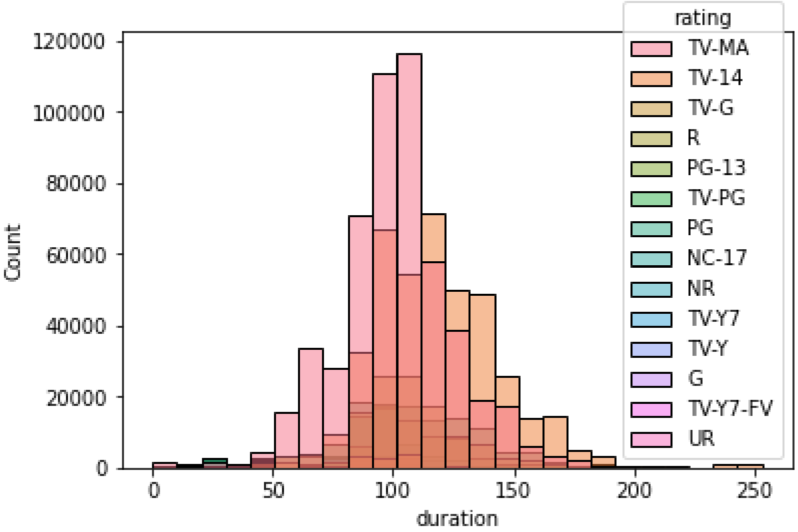
```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning:
`distplot` is a deprecated function and will be removed in a future version. Please a
dapt your code to use either `displot` (a figure-level function with similar flexibil
ity) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[24]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc305aee310>
```



histogram

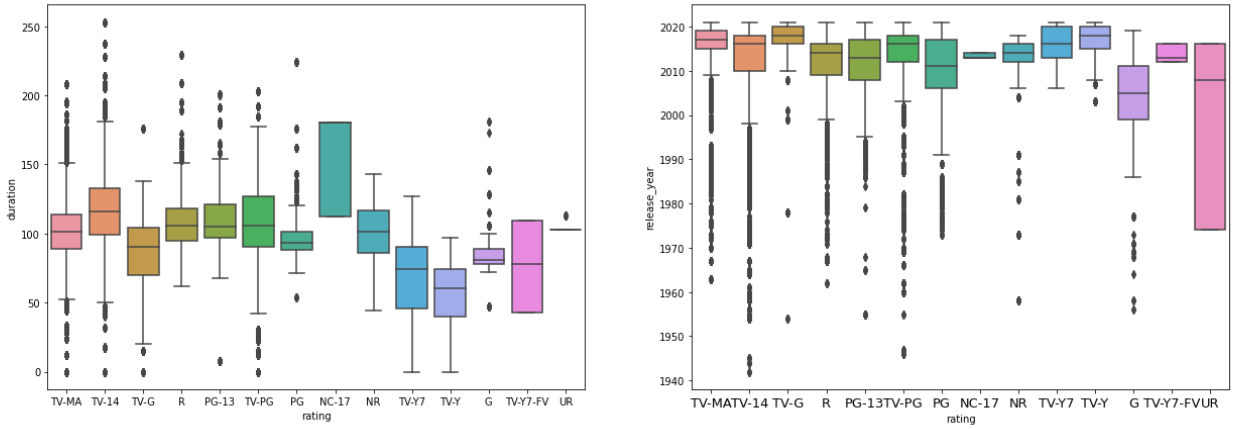
In [25]:
```
sns.histplot(data=netflix_df, x='duration', bins=25, hue = 'rating')
plt.show()
```



Boxplot

In [26]:
```
fig, (ax1, ax2) = plt.subplots(ncols=2, figsize=(21,7))
sns.boxplot(x = 'rating', y = 'duration', data=netflix_df, ax=ax1)
sns.boxplot(x = 'rating', y = 'release_year', data=netflix_df, ax=ax2)
# sns.boxplot(x = 'type', y = 'Season', data=netflix_df, ax=ax3)
plt.xticks(fontsize= 13)
# plt.title('Box plot of numerical columns', fontsize=16);
```

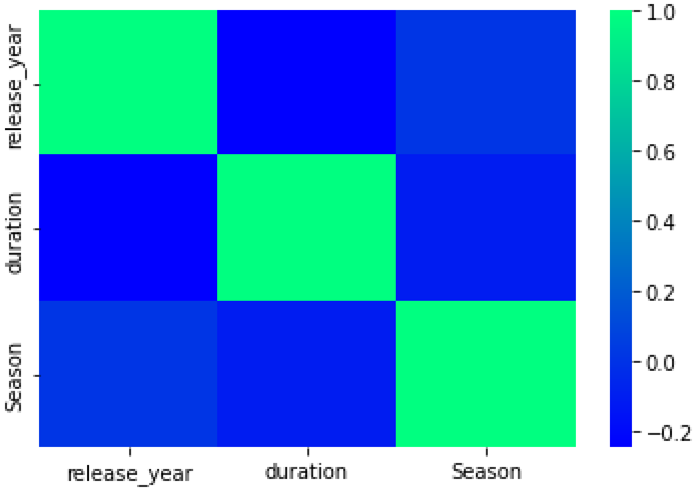
Out[26]:
```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13]),
 <a list of 14 Text major ticklabel objects>)
```



HeatMap

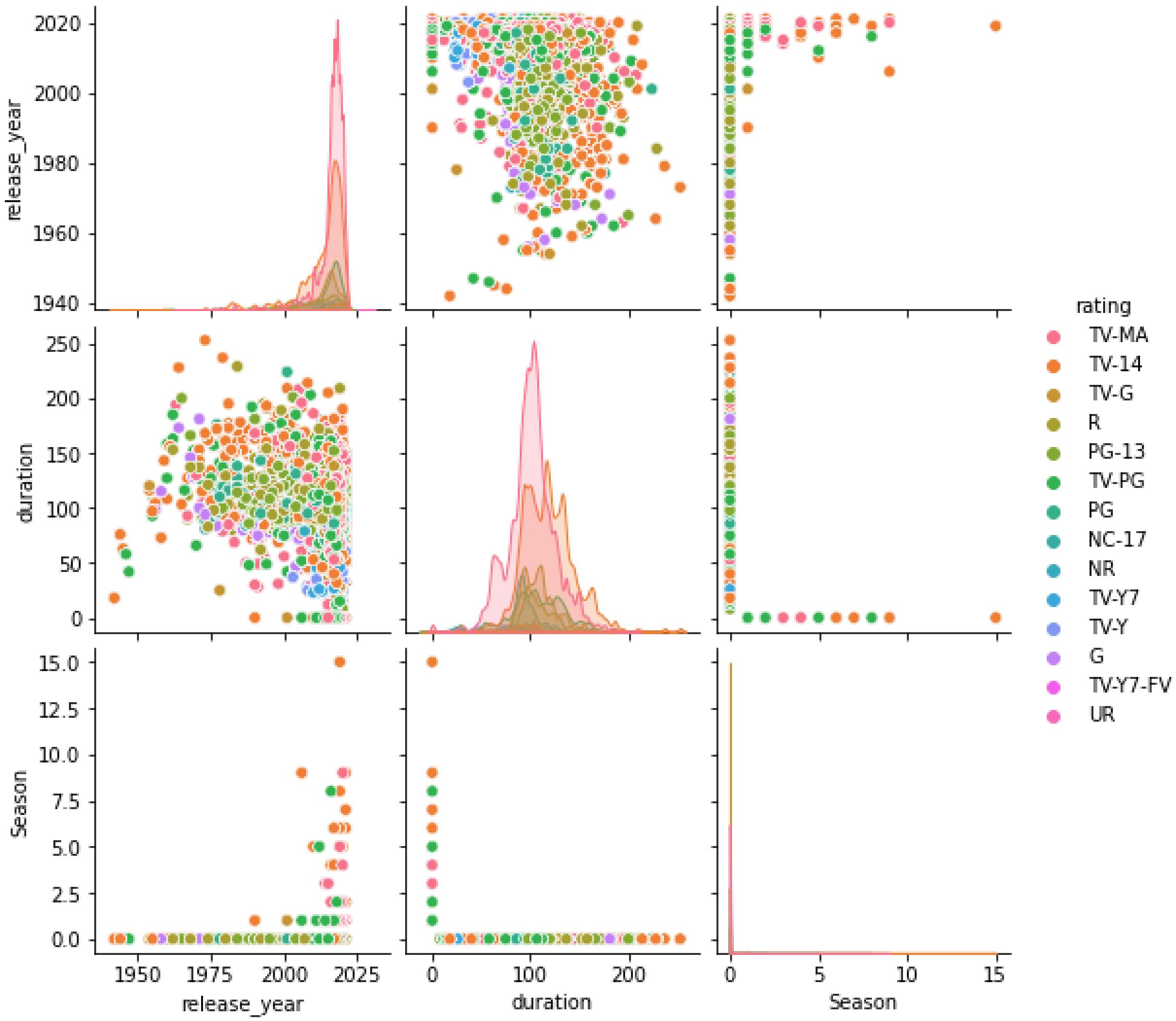
In [27]:
```
corr = netflix_df.corr()
sns.heatmap(corr, cmap='winter')
```

Out[27]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc301cdd7d0>
```



Pairplot

In [28]:
```
sns.pairplot(data=netflix_df, hue='rating')
plt.show()
```

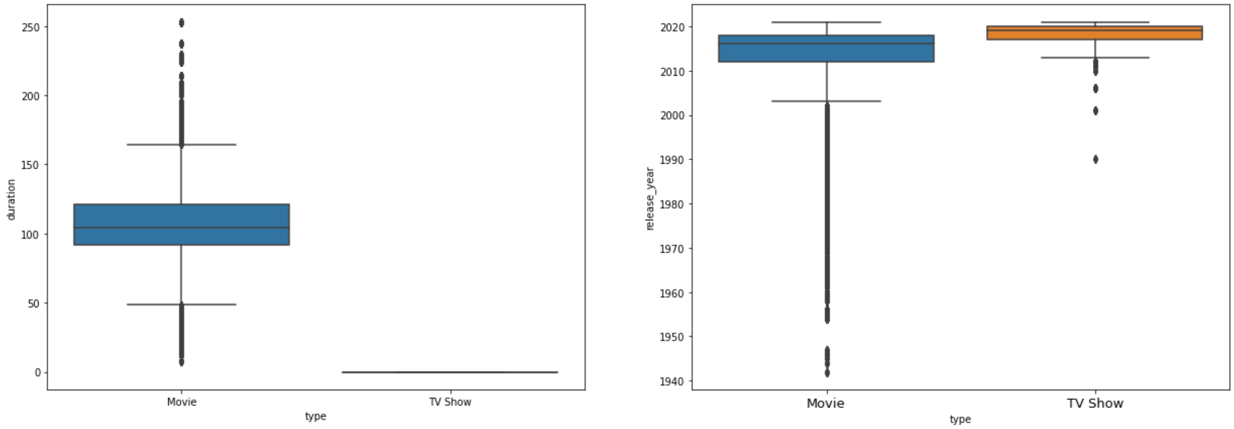


5. Missing Value & Outlier check

In [21]:
```
fig, (ax1, ax2) = plt.subplots(ncols=2, figsize=(21,7))
sns.boxplot(x = 'type', y = 'duration', data=netflix_df, ax=ax1)
sns.boxplot(x = 'type', y = 'release_year', data=netflix_df, ax=ax2)
# sns.boxplot(x = 'type', y = 'Season', data=netflix_df, ax=ax3)
plt.xticks(fontsize= 13)
# plt.title('Box plot of numerical columns', fontsize=16);
```

Out[21]:

```
(array([0, 1]), <a list of 2 Text major ticklabel objects>)
```


duration
Movie
TV Show
type
release_year
Movie
TV Show
type