In [1]:

```
# Importing relevent libraries:

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')

path = './movies.csv'
data = pd.read_csv(path)
data.head()
```

Out[1]:

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTime	Gross
0	Blood Red Sky	(2021)	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\n| \n Star...	21,062	121.0	NaN
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870	25.0	NaN
2	The Walking Dead	(2010– 2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	885,805	44.0	NaN
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	414,849	23.0	NaN
4	Army of Thieves	(2021)	\nAction, Crime, Horror	NaN	\nA prequel, set before the events of Army of ...	\n Director:\nMatthias Schweighöfer\n| \n ...	NaN	NaN	NaN

In [2]:

```
# DATA CLEANING :-
# Dropping duplicates with respect to MOVIES column:

data = data.drop_duplicates(subset = ["MOVIES"], keep = False)
data
```

Out[2]:

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTime	Gross
0	Blood Red Sky	(2021)	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\n| \n Star...	21,062	121.0	NaN
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870	25.0	NaN
2	The Walking Dead	(2010– 2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	885,805	44.0	NaN
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	414,849	23.0	NaN
4	Army of Thieves	(2021)	\nAction, Crime, Horror	NaN	\nA prequel, set before the events of Army of ...	\n Director:\nMatthias Schweighöfer\n| \n ...	NaN	NaN	NaN

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTime	Gross
...	...	...	...	...	...	...	...	...	...
9845	Disparu à jamais	(2021)	\nCrime, Drama, Mystery	NaN	\nAdd a Plot\n	\n Director:\nJuan Carlos Medina\n| \n S...	NaN	NaN	NaN
9901	Modern Family	(2009– 2020)	\nComedy, Drama, Romance	8.2	\nJay must adapt to his young new wife, Gloria...	\n Director:\nJason Winer\n| \n Stars:\n...	3,404	23.0	NaN
9993	Totenfrau	(2022–)	\nDrama, Thriller	NaN	\nAdd a Plot\n	\n Director:\nNicolai Rohde\n| \n Stars:...	NaN	NaN	NaN
9995	Arcane	(2021–)	\nAnimation, Action, Adventure	NaN	\nAdd a Plot\n	\n	NaN	NaN	NaN
9996	Heart of Invictus	(2022–)	\nDocumentary, Sport	NaN	\nAdd a Plot\n	\n Director:\nOrlando von Einsiedel\n| \n ...	NaN	NaN	NaN

6398 rows × 9 columns

In [3]:

```
# Removing '\n' from GENRE, ONE-LINE and STARS:

for column in ['GENRE','ONE-LINE','STARS']:
    data[column]=data[column].str.replace('\n',' ')

# Replacing unwanted strings and converting datatype of VOTES:

data['ONE-LINE']=data['ONE-LINE'].str.replace('Add a Plot','Not Specified')
data['VOTES']=data['VOTES'].str.replace(',','')
data['VOTES']=pd.to_numeric(data['VOTES'])
data
```

Out[3]:

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTime	Gross
0	Blood Red Sky	(2021)	Action, Horror, Thriller	6.1	A woman with a mysterious illness is forced i...	Director: Peter Thorwarth | Stars: P...	21062.0	121.0	NaN
1	Masters of the Universe: Revelation	(2021–)	Animation, Action, Adventure	5.0	The war for Eternia begins again in what may ...	Stars: Chris Wood, Sarah Mi...	17870.0	25.0	NaN
2	The Walking Dead	(2010– 2022)	Drama, Horror, Thriller	8.2	Sheriff Deputy Rick Grimes wakes up from a co...	Stars: Andrew Lincoln, Norm...	885805.0	44.0	NaN
3	Rick and Morty	(2013–)	Animation, Adventure, Comedy	9.2	An animated series that follows the exploits ...	Stars: Justin Roiland, Chri...	414849.0	23.0	NaN
4	Army of Thieves	(2021)	Action, Crime, Horror	NaN	A prequel, set before the events of Army of t...	Director: Matthias Schweighöfer | St...	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...
9845	Disparu à jamais	(2021)	Crime, Drama, Mystery	NaN	Not Specified	Director: Juan Carlos Medina | Star:...	NaN	NaN	NaN
9901	Modern Family	(2009– 2020)	Comedy, Drama, Romance	8.2	Jay must adapt to his young new wife, Gloria ...	Director: Jason Winer | Stars: Ed O'...	3404.0	23.0	NaN
9993	Totenfrau	(2022–)	Drama, Thriller	NaN	Not Specified	Director: Nicolai Rohde | Stars: Fel...	NaN	NaN	NaN
9995	Arcane	(2021–)	Animation, Action, Adventure	NaN	Not Specified		NaN	NaN	NaN
						Director: Orlando			

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTime	Gross
9996	Heart of Invictus	(2022–)	Documentary, Sport	NaN	Not Specified	von Eingiedel | St...	NaN	NaN	NaN

6398 rows × 9 columns

In [4]:

```
# Seperating Directors and Stars into different columns from STARS:

def seperate_director(director):
    if 'Director' in director or 'Directors' in director:
        director = director.strip().split('|')[0]
        return director.split(":")[1].strip()
    else:
        return ''

def seperate_stars(stars):
    if 'Star' not in stars or 'Stars' not in stars:
        return ''
    else:
        return stars.split(":")[-1].strip()

data['Director'] = data['STARS'].apply(lambda d: seperate_director(d))
data['Stars'] = data['STARS'].apply(lambda s: seperate_stars(s))
data.head()
```

Out[4]:

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTime	Gross	Director	Star
0	Blood Red Sky	(2021)	Action, Horror, Thriller	6.1	A woman with a mysterious illness is forced i...	Director: Peter Thorwarth | Stars: P...	21062.0	121.0	NaN	Peter Thorwarth	Pe Baumeiste Carl Anto Koch Alexander .
1	Masters of the Universe: Revelation	(2021–)	Animation, Action, Adventure	5.0	The war for Eternia begins again in what may ...	Stars: Chris Wood, Sarah Mi...	17870.0	25.0	NaN		Chris Wooc Sara Michell Gellar, Len Head.
2	The Walking Dead	(2010– 2022)	Drama, Horror, Thriller	8.2	Sheriff Deputy Rick Grimes wakes up from a co...	Stars: Andrew Lincoln, Norm...	885805.0	44.0	NaN		Andrev Lincol Norma Reedus Meliss McBri.
3	Rick and Morty	(2013–)	Animation, Adventure, Comedy	9.2	An animated series that follows the exploits ...	Stars: Justin Roiland, Chri...	414849.0	23.0	NaN		Justi Roiland, Chri Parnel Spence Gramm.
4	Army of Thieves	(2021)	Action, Crime, Horror	NaN	A prequel, set before the events of Army of t...	Director: Matthias Schweighöfer | St...	NaN	NaN	NaN	Matthias Schweighöfer	Matthia Schweighöfe Nathali Emmanue Ru.

In [5]:

```
# Dropping Gross and STARS columns:

data = data.drop(columns='Gross')
data = data.drop(columns='STARS')
data.head()
```

Out[5]:

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	VOTES	RunTime	Director	Stars
0	Blood Red Sky	(2021)	Action, Horror, Thriller	6.1	A woman with a mysterious illness is forced i...	21062.0	121.0	Peter Thorwarth	Peri Baumeister, Carl Anton Koch, Alexander ...
1	Masters of the Universe: Revelation	(2021–)	Animation, Action, Adventure	5.0	The war for Eternia begins again in what may ...	17870.0	25.0		Chris Wood, Sarah Michelle Gellar, Lena Head...
2	The Walking Dead	(2010–2022)	Drama, Horror, Thriller	8.2	Sheriff Deputy Rick Grimes wakes up from a co...	885805.0	44.0		Andrew Lincoln, Norman Reedus, Melissa McBri...
3	Rick and Morty	(2013–)	Animation, Adventure, Comedy	9.2	An animated series that follows the exploits ...	414849.0	23.0		Justin Roiland, Chris Parnell, Spencer Gramm...
4	Army of Thieves	(2021)	Action, Crime, Horror	NaN	A prequel, set before the events of Army of t...	NaN	NaN	Matthias Schweighöfer	Matthias Schweighöfer, Nathalie Emmanuel, Ru...

In [6]:

```
# Checking for null values:

data.isnull().sum()
```

Out[6]:

```
MOVIES          0
YEAR          438
GENRE          70
RATING        918
ONE-LINE        0
VOTES         918
RunTime      1350
Director        0
Stars           0
dtype: int64
```

In [7]:

```
# Filling null values:

for column in ['YEAR','GENRE']:
    data[column]=data[column].fillna('Not Specified')

# Filling null values with mean value of the column:

for column in ['RATING','VOTES','RunTime']:
    data[column]=data[column].fillna(int(data[column].mean()))
data
```

Out[7]:

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	VOTES	RunTime	Director	Stars
0	Blood Red Sky	(2021)	Action, Horror, Thriller	6.1	A woman with a mysterious illness is forced i...	21062.0	121.0	Peter Thorwarth	Peri Baumeister, Carl Anton Koch, Alexander ...
									Chris Wood,

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	VOTES	RunTime	Director	Stars
1	Marvel Universe: Revelation	(2021–)	Animation, Action, Adventure	5.0	The war for Earth begins again in what may ...	17870.0	25.0		Michelle Gellar, Lena Head...
2	The Walking Dead	(2010– 2022)	Drama, Horror, Thriller	8.2	Sheriff Deputy Rick Grimes wakes up from a co...	885805.0	44.0		Andrew Lincoln, Norman Reedus, Melissa McBri...
3	Rick and Morty	(2013–)	Animation, Adventure, Comedy	9.2	An animated series that follows the exploits ...	414849.0	23.0		Justin Roiland, Chris Parnell, Spencer Gramm...
4	Army of Thieves	(2021)	Action, Crime, Horror	6.0	A prequel, set before the events of Army of t...	21652.0	79.0	Matthias Schweighöfer	Matthias Schweighöfer, Nathalie Emmanuel, Ru...
...	...	...	...	...	...	...	...	...	...
9845	Disparu à jamais	(2021)	Crime, Drama, Mystery	6.0	Not Specified	21652.0	79.0	Juan Carlos Medina	
9901	Modern Family	(2009– 2020)	Comedy, Drama, Romance	8.2	Jay must adapt to his young new wife, Gloria ...	3404.0	23.0	Jason Winer	Ed O'Neill, Sofía Vergara, Julie Bowen, Ty ...
9993	Totenfrau	(2022–)	Drama, Thriller	6.0	Not Specified	21652.0	79.0	Nicolai Rohde	Felix Klare, Romina Küper, Anna Maria Mühe, ...
9995	Arcane	(2021–)	Animation, Action, Adventure	6.0	Not Specified	21652.0	79.0		
9996	Heart of Invictus	(2022–)	Documentary, Sport	6.0	Not Specified	21652.0	79.0	Orlando von Einsiedel	

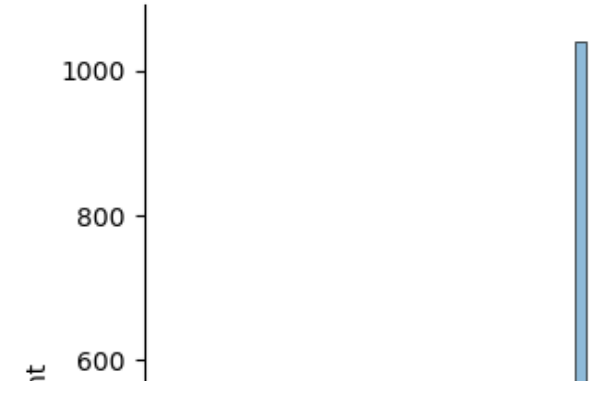
6398 rows × 9 columns

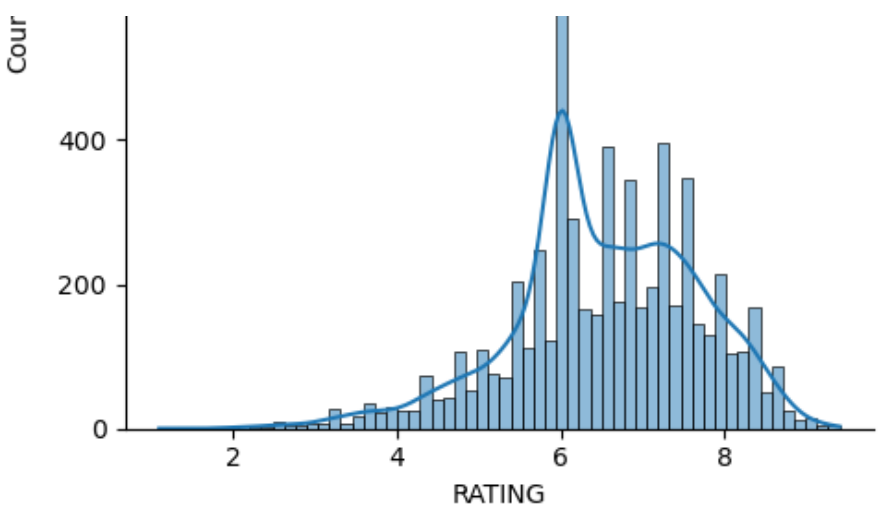
In [8]:

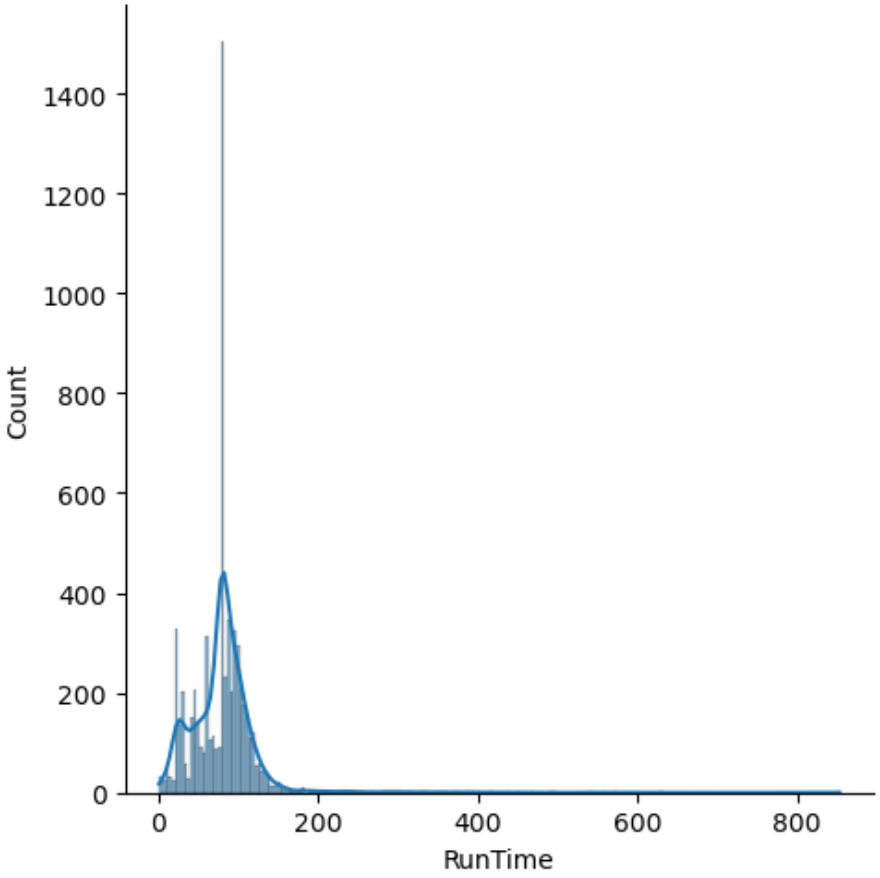
```
# Converting clean data to csv file:

data.to_csv(r'C:\Users\samve\Documents\Data science\movies_clean.csv')
```

In [9]:

```
# DATA VISUALIZATION :-
# Distribution plots for RATING and RunTime columns:

for column in ['RATING','RunTime']:
    sns.displot(data[column], kind='hist', kde= True)
plt.show()
```

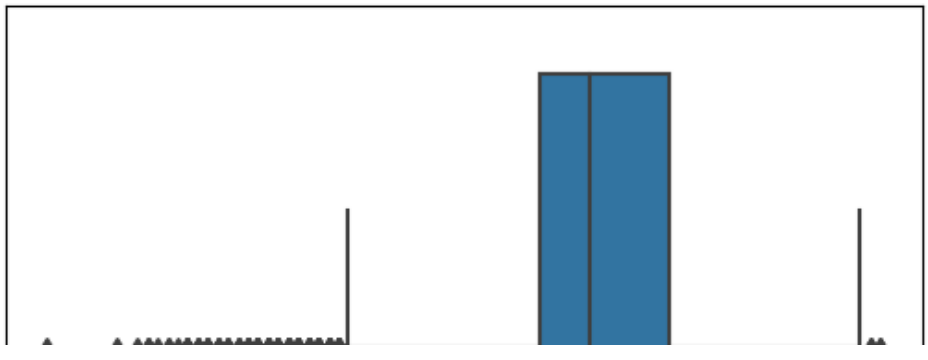


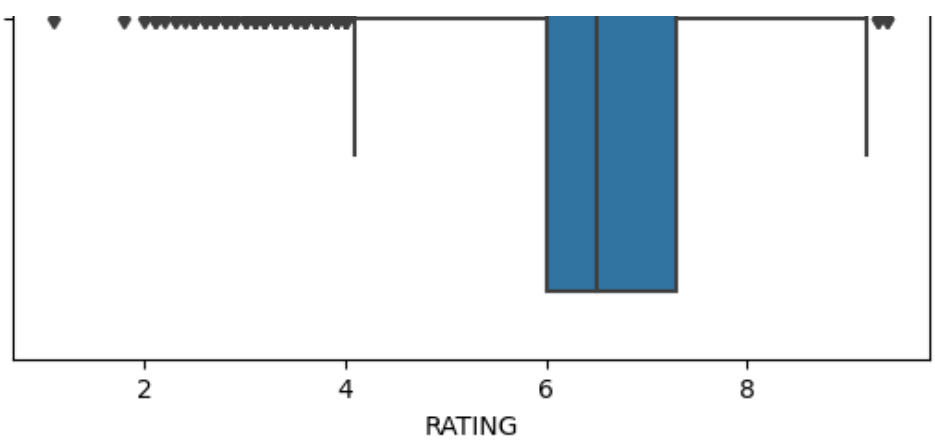


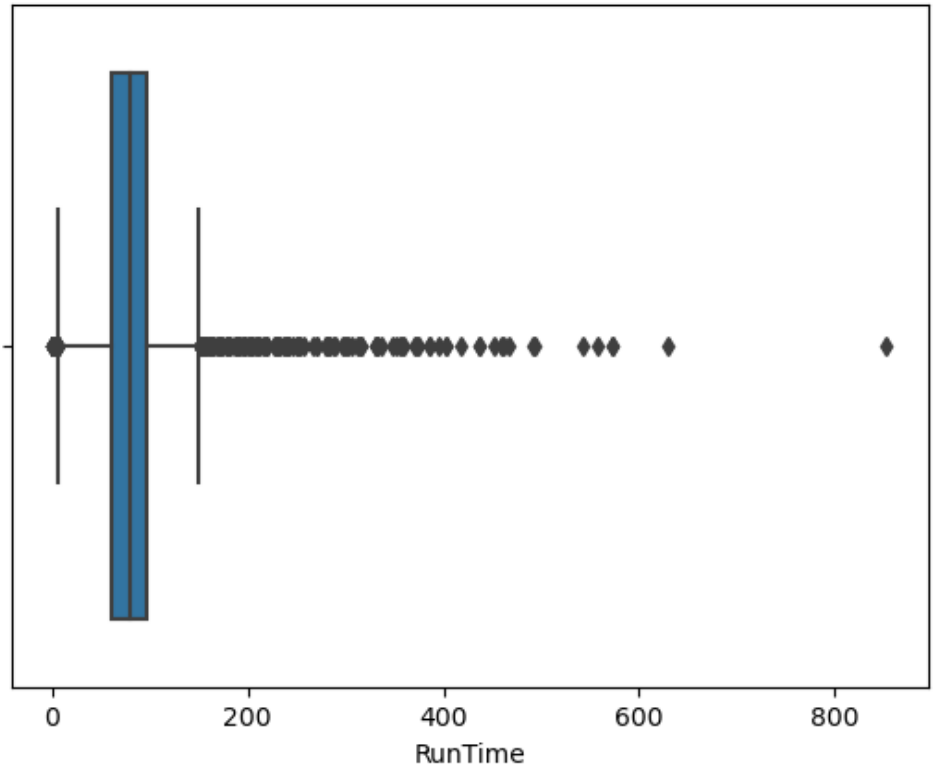


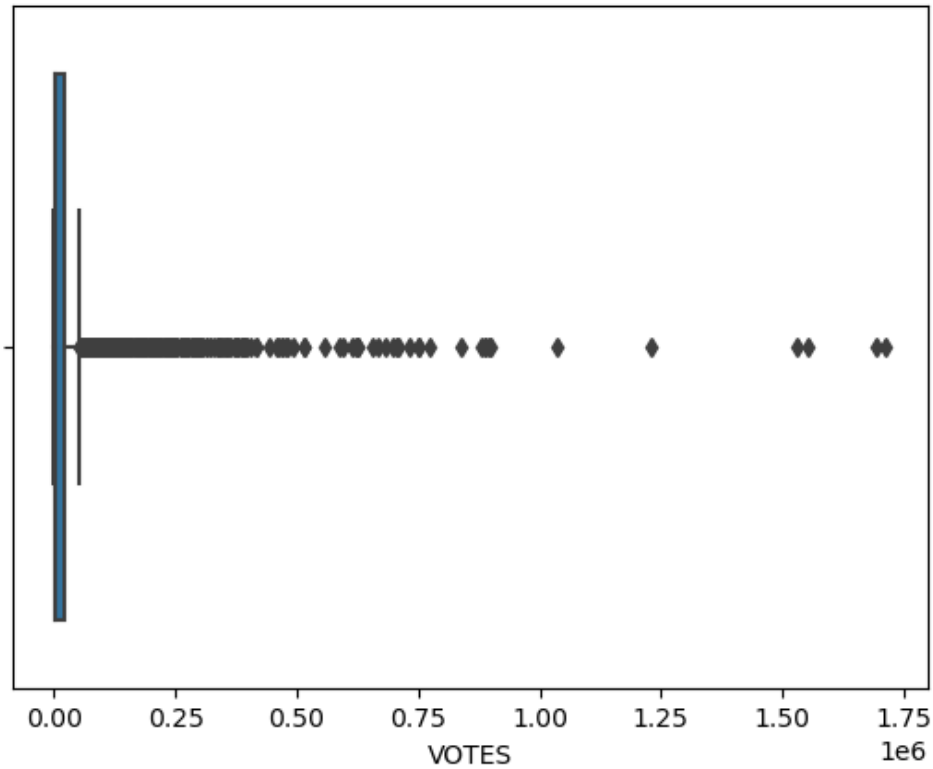
In [10]:

```
# Boxplots for RATING, RunTIme and VOTES:

sns.boxplot(data['RATING'])
plt.show()

sns.boxplot(data['RunTime'])
plt.show()

sns.boxplot(data['VOTES'])
plt.show()
```





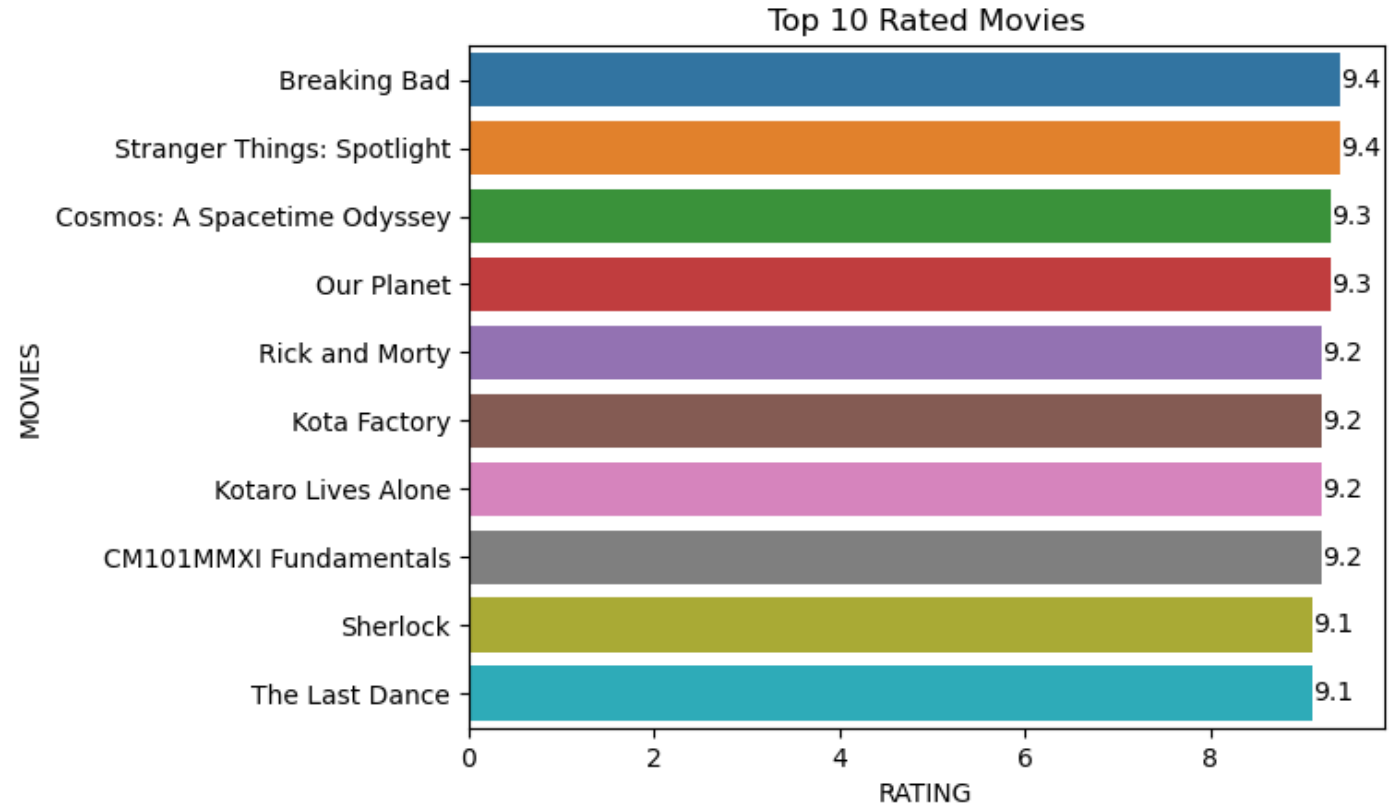


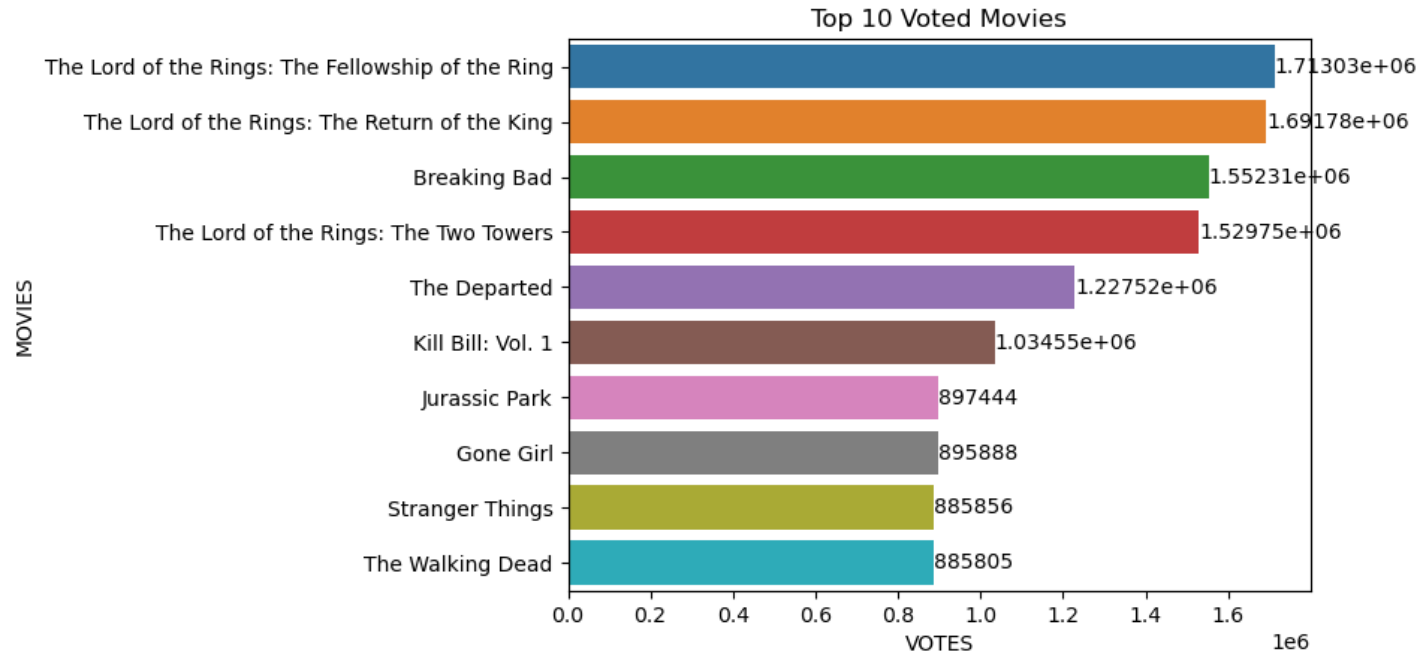


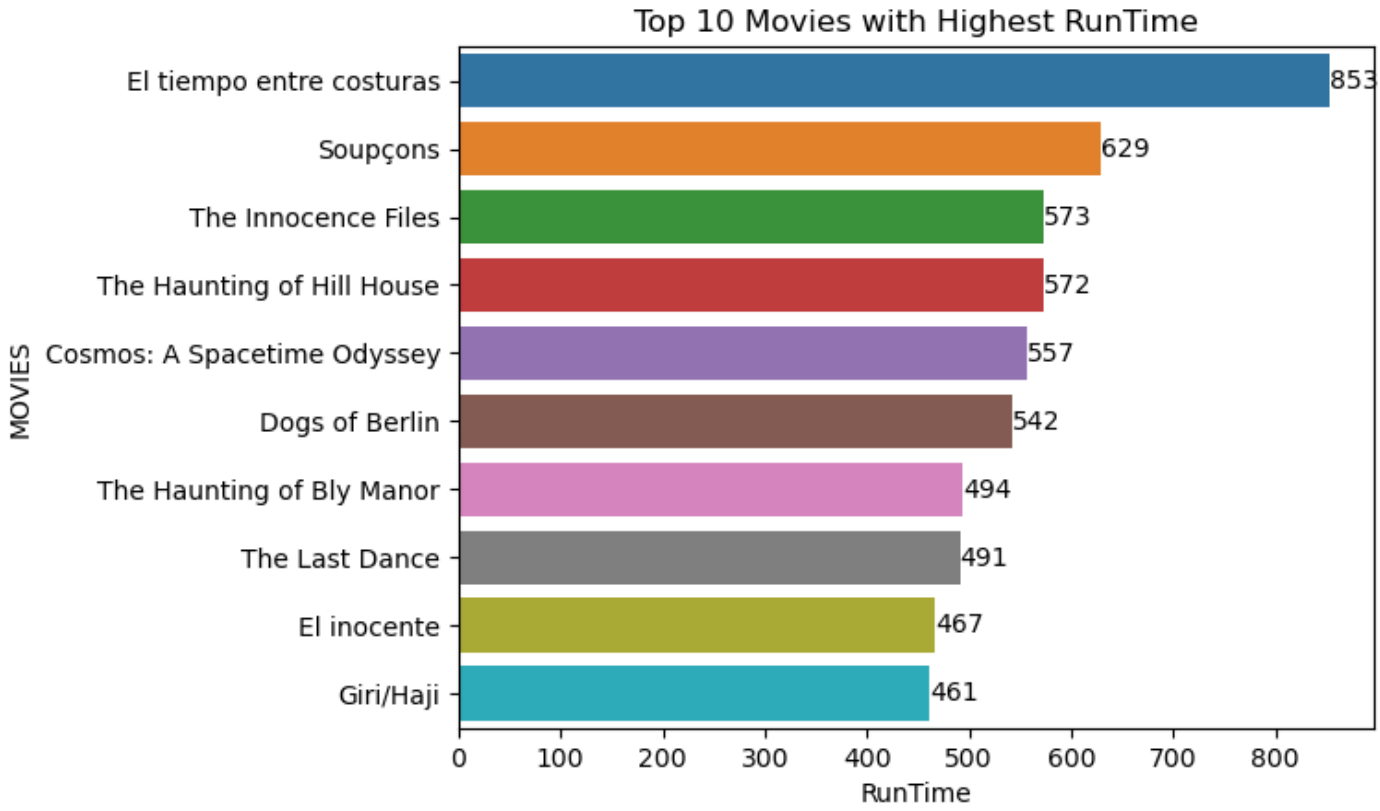
In [11]:

```
# Depicting Top 10:
```

```
data_R10 = data.nlargest(10, ['RATING'])
plot= sns.barplot(data = data_R10, y= 'MOVIES', x= 'RATING')
for var in plot.containers:
    plot.bar_label(var)
plt.title('Top 10 Rated Movies')
plt.show()

data_V10 = data.nlargest(10, ['VOTES'])
plot= sns.barplot(data = data_V10, y= 'MOVIES', x= 'VOTES')
for var in plot.containers:
    plot.bar_label(var)
plt.title('Top 10 Voted Movies')
plt.show()

data_RT10 = data.nlargest(10, ['RunTime'])
plot= sns.barplot(data = data_RT10, y= 'MOVIES', x= 'RunTime')
for var in plot.containers:
    plot.bar_label(var)
plt.title('Top 10 Movies with Highest RunTime')
plt.show()
```

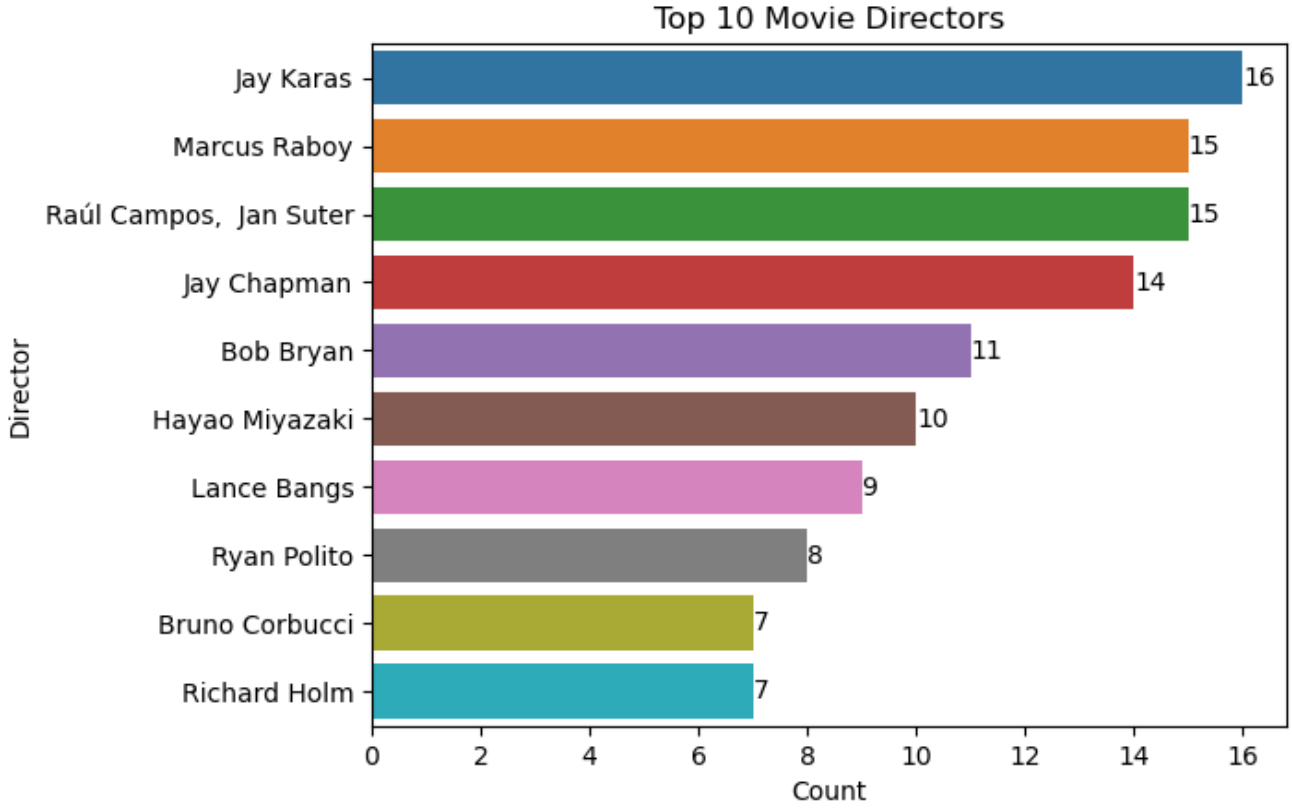






In [12]:

```
data_direc = data[~(data['Director'] == "")]['Director'].value_counts().reset_index().he
ad(10)
plot = sns.barplot(data = data_direc, x = 'Director', y = 'index')
plt.xlabel('Count')
plt.ylabel('Director')
plt.title('Top 10 Movie Directors')
for var in plot.containers:
    plot.bar_label(var)
plt.show()
```



In [13]:

```
from collections import Counter
```

```
genre = data['GENRE'].to_list()
genre_list = list()

for genres in genre:
    genres = genres.split(', ')
    for g in genres:
        genre_list.append(g)


data_genre = pd.DataFrame.from_dict(Counter(genre_list), orient = 'index').rename(column
s = {0:'Count'})
data_G = data_genre.sort_values(by = 'Count', ascending = False).head(10)
data_G
```

Out[13]:

	Count
Animation	786
Comedy	750
Drama	735
Action	735
Drama	696
Comedy	672
Drama	665
Romance	599
Documentary	553
Thriller	545

In [14]:

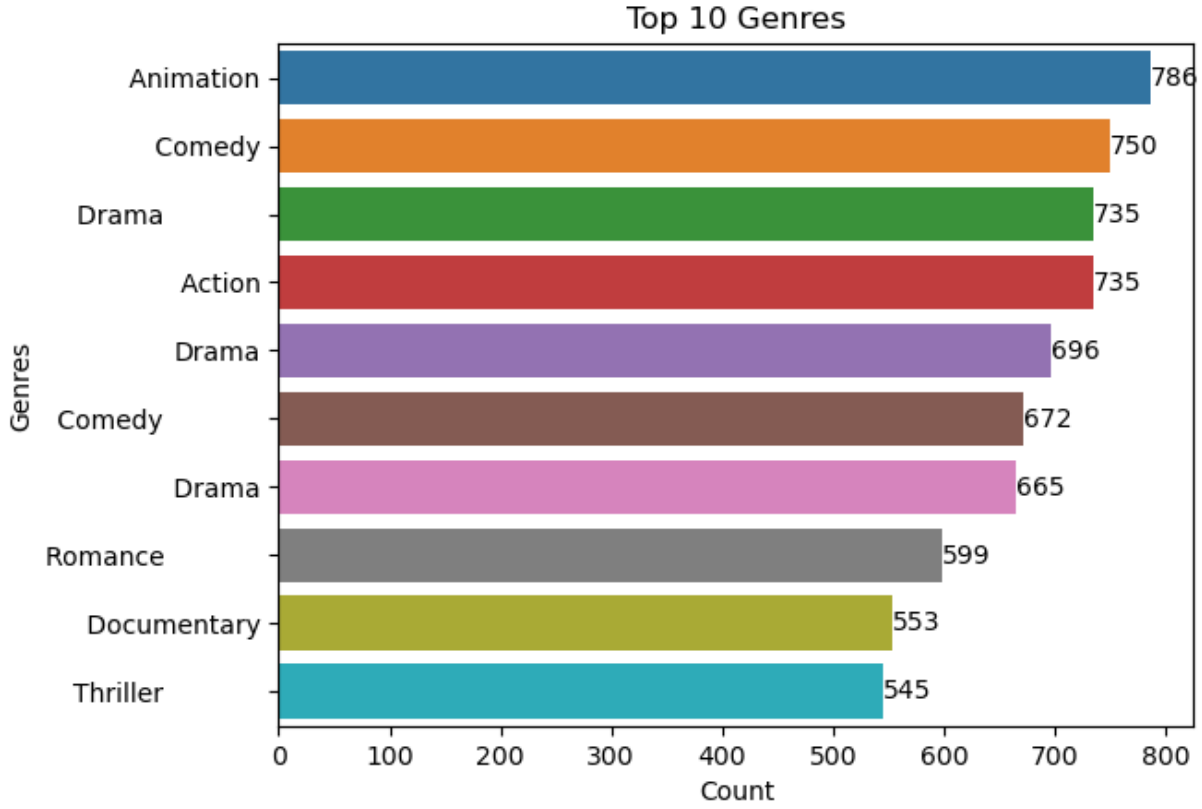
```
STARS = data[~(data['Stars'] == '')]['Stars'].to_list()
stars_list = list()

for stars in STARS:
    stars = stars.split(', ')
    for s in stars:
        stars_list.append(s)

data_stars = pd.DataFrame.from_dict(Counter(stars_list), orient = 'index').rename(column
s = {0:'Count'})
data_S = data_stars.sort_values(by = 'Count', ascending = False).head(10)
data_S
```

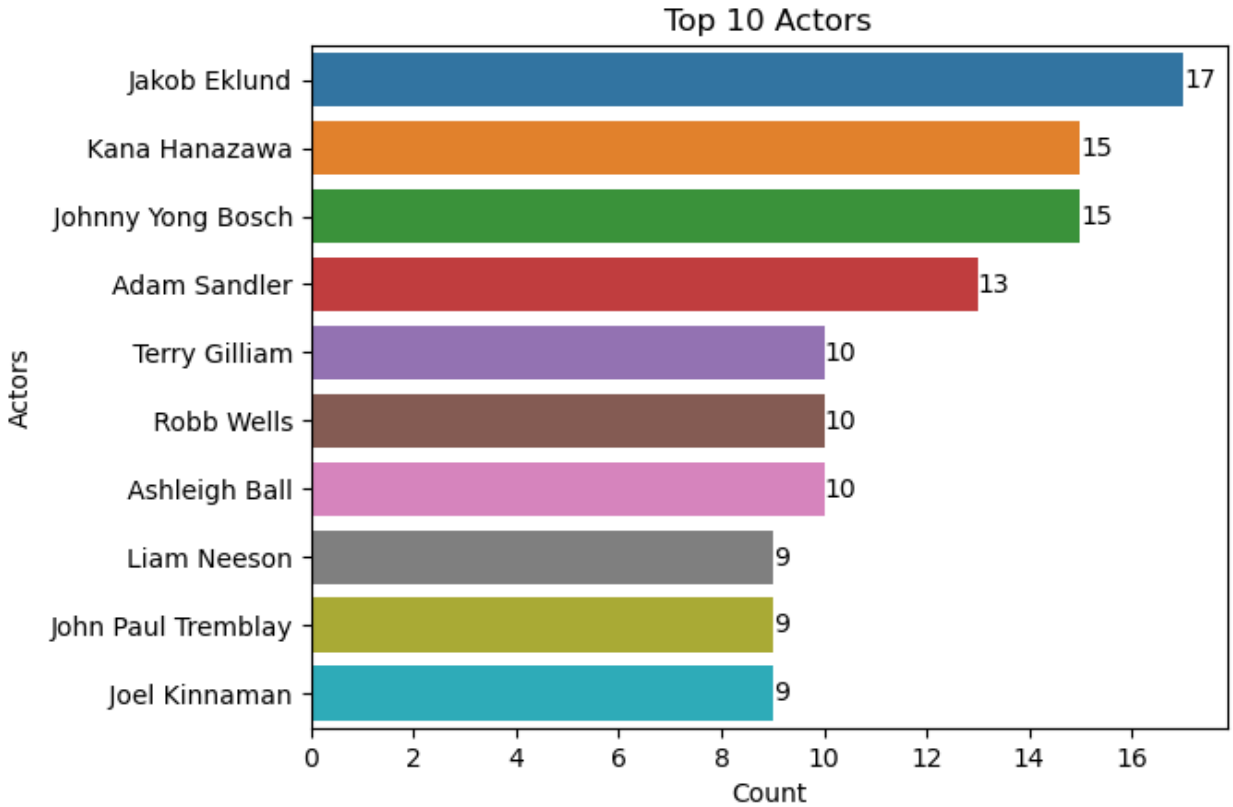
Out[14]:

	Count
Jakob Eklund	17
Kana Hanazawa	15
Johnny Yong Bosch	15
Adam Sandler	13
Terry Gilliam	10
Robb Wells	10
Ashleigh Ball	10
Liam Neeson	9
John Paul Tremblay	9
Joel Kinnaman	9

In [15]:

```
plot = sns.barplot(data = data_G, x = 'Count', y = data_G.index)
plt.ylabel('Genres')
for var in plot.containers:
    plot.bar_label(var)
plt.title('Top 10 Genres')
plt.show()

plot = sns.barplot(data = data_S, x = 'Count', y = data_S.index)
plt.ylabel('Actors')
for var in plot.containers:
    plot.bar_label(var)
plt.title('Top 10 Actors')
plt.show()
```





In [ ]: