IMDB_MOVIE_EDA_PROJECT_VIVEK_CHAUHAN

In [1]:

```
# upload the necessary libraries

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

In [2]:

```
# forward / is used for path.
data = pd.read_csv("C:/Users/VIVEK CHAUHAN/Desktop/eda-projects (1)/5-eda-project/IMDB-Movie-Data.csv")
data
```

Out[2]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02
...	...	...	...	...	...	...	...	...	...	...	..
995	996	Secret in Their Eyes	Crime,Drama,Mystery	A tight-knit team of rising investigators, alo...	Billy Ray	Chiwetel Ejiofor, Nicole Kidman, Julia Roberts...	2015	111	6.2	27585	NaN
996	997	Hostel: Part II	Horror	Three American	Eli Roth	Lauren German, Heather Matarazzo,	2007	94	5.5	73152	17.54

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)
				college students studying abroa...		Bijou Philli...					
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008	98	6.2	70699	58.01
998	999	Search Party	Adventure,Comedy	A pair of friends embark on a mission to reuni...	Scot Armstrong	Adam Pally, T.J. Miller, Thomas Middleditch,Sh...	2014	93	5.6	4881	NaN
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016	87	5.3	12435	19.64

1000 rows × 12 columns

In [3]:
```
# display top10 rows of the dataset
data.head(10)
```

Out[3]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02
5	6	The Great Wall	Action,Adventure,Fantasy	European mercenaries searching for black powde...	Yimou Zhang	Matt Damon, Tian Jing, Willem Dafoe, Andy Lau	2016	103	6.1	56036	45.13

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)
6	7	La La Land	Comedy,Drama,Music	A jazz pianist falls for an aspiring actress i...	Damien Chazelle	Ryan Gosling, Emma Stone, Rosemarie DeWitt, J....	2016	128	8.3	258682	151.06
7	8	Mindhorn	Comedy	A has-been actor best known for playing the ti...	Sean Foley	Essie Davis, Andrea Riseborough, Julian Barrat...	2016	89	6.4	2490	NaN
8	9	The Lost City of Z	Action,Adventure,Biography	A true-life drama, centering on British explor...	James Gray	Charlie Hunnam, Robert Pattinson, Sienna Mille...	2016	141	7.1	7188	8.01
9	10	Passengers	Adventure,Drama,Romance	A spacecraft traveling to a distant colony pla...	Morten Tyldum	Jennifer Lawrence, Chris Pratt, Michael Sheen,...	2016	116	7.0	192177	100.01

In [4]:
```
# # display bottom 10 rows of the dataset
data.tail(10)
```

Out[4]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	M
990	991	Underworld: Rise of the Lycans	Action,Adventure,Fantasy	An origins story centered on the centuries-old...	Patrick Tatopoulos	Rhona Mitra, Michael Sheen, Bill Nighy, Steven...	2009	92	6.6	129708	45.80	
991	992	Taare Zameen Par	Drama,Family,Music	An eight-year-old boy is thought to be a lazy ...	Aamir Khan	Darsheel Safary, Aamir Khan, Tanay Chheda, Sac...	2007	165	8.5	102697	1.20	
992	993	Take Me Home Tonight	Comedy,Drama,Romance	Four years after graduation, an awkward high s...	Michael Dowse	Topher Grace, Anna Faris, Dan Fogler, Teresa P...	2011	97	6.3	45419	6.92	
993	994	Resident Evil: Afterlife	Action,Adventure,Horror	While still out to destroy the evil Umbrella C...	Paul W.S. Anderson	Milla Jovovich, Ali Larter, Wentworth Miller,K...	2010	97	5.9	140900	60.13	
994	995	Project X	Comedy	3 high school seniors throw a birthday party t...	Nima Nourizadeh	Thomas Mann, Oliver Cooper, Jonathan Daniel Br...	2012	88	6.7	164088	54.72	
995	996	Secret in Their Eyes	Crime,Drama,Mystery	A tight-knit team of rising investigators, alo...	Billy Ray	Chiwetel Ejiofor, Nicole Kidman, Julia Roberts...	2015	111	6.2	27585	NaN	

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	M
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	94	5.5	73152	17.54	
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008	98	6.2	70699	58.01	
998	999	Search Party	Adventure,Comedy	A pair of friends embark on a mission to reuni...	Scot Armstrong	Adam Pally, T.J. Miller, Thomas Middleditch,Sh...	2014	93	5.6	4881	NaN	
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016	87	5.3	12435	19.64	

In [5]:

```
# find shape of our dataset(number of rows & columns)

data.shape
```

Out[5]:

```
(1000, 12)
```

In [6]:

```
# print the number of rows in the dataset

print("no of rows in the dataset",data.shape[0])
```

```
no of rows in the dataset 1000
```

In [7]:

```
# print the number of columns in the dataset

print("no of columns in the dataset",data.shape[1])
```

```
no of columns in the dataset 12
```

In [8]:

```
# print all the information of our dataset

data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Rank                1000 non-null   int64  
 1   Title               1000 non-null   object 
 2   Genre               1000 non-null   object 
 3   Description         1000 non-null   object 
 4   Director            1000 non-null   object 
 5   Actors              1000 non-null   object 
 6   Year                1000 non-null   int64  
 7   Runtime (Minutes)   1000 non-null   int64  
 8   Rating              1000 non-null   float64
 9   Votes               1000 non-null   int64  
 10  Revenue (Millions)  872 non-null    float64
 11  Metascore           936 non-null    float64
dtypes: float64(3), int64(4), object(5)
memory usage: 93.9+ KB
```

In [9]:

```
# check null values in the dataset

data.isnull()
```

Out[9]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
0	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...
995	False	False	False	False	False	False	False	False	False	False	True	False
996	False	False	False	False	False	False	False	False	False	False	False	False
997	False	False	False	False	False	False	False	False	False	False	False	False
998	False	False	False	False	False	False	False	False	False	False	True	False
999	False	False	False	False	False	False	False	False	False	False	False	False

1000 rows × 12 columns

In [10]:

```
# let's count how many null values in the dataset column wise

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

Out[10]:

```
Rank                    0
Title                   0
Genre                   0
Description             0
Director                0
Actors                  0
Year                    0
Runtime (Minutes)       0
Rating                  0
Votes                   0
Revenue (Millions)    128
Metascore              64
dtype: int64
```

In [11]:

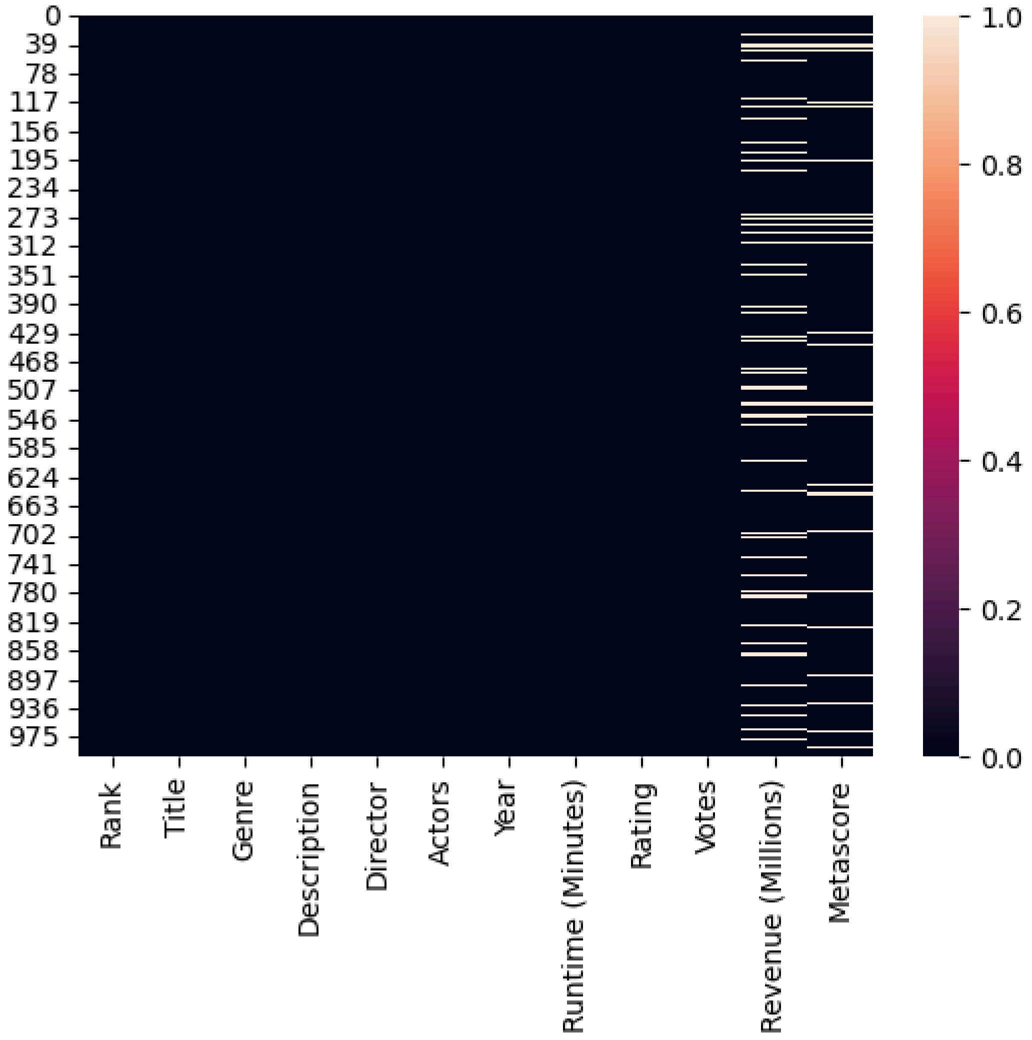
```
# let's count how many null values in the dataset row wise

data.isnull().sum(axis=0)
```

Out[11]:

```
Rank                    0
Title                   0
Genre                   0
Description             0
Director                0
Actors                  0
Year                    0
Runtime (Minutes)       0
Rating                  0
Votes                   0
Revenue (Millions)    128
Metascore              64
dtype: int64
```

In [12]:

```
# let's visualise how many null values is present in the dataset via column wise

sns.heatmap(data.isnull()) # white lines indicates the presence of null values
plt.show()
```



In [13]:

```
# drop null values in the dataset if any

data.dropna()
```

Out[13]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02
...	...	...	...	...	...	...	...	...	...	...	...
993	994	Resident Evil: Afterlife	Action,Adventure,Horror	While still out to destroy the evil Umbrella C...	Paul W.S. Anderson	Milla Jovovich, Ali Larter, Wentworth Miller,K...	2010	97	5.9	140900	60.13
994	995	Project X	Comedy	3 high school	Nima Nourizadeh	Thomas Mann, Oliver Cooper,	2012	88	6.7	164088	54.72

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)
				seniors throw a birthday party t...		Jonathan Daniel Br...					
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	94	5.5	73152	17.54
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008	98	6.2	70699	58.01
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016	87	5.3	12435	19.64

838 rows × 12 columns

In [14]:

```
# another method to remove the null values in the dataset by row wise

data = data.dropna(axis=0) # for permanent saved removed null values in the original dataset
data
```

Out[14]:

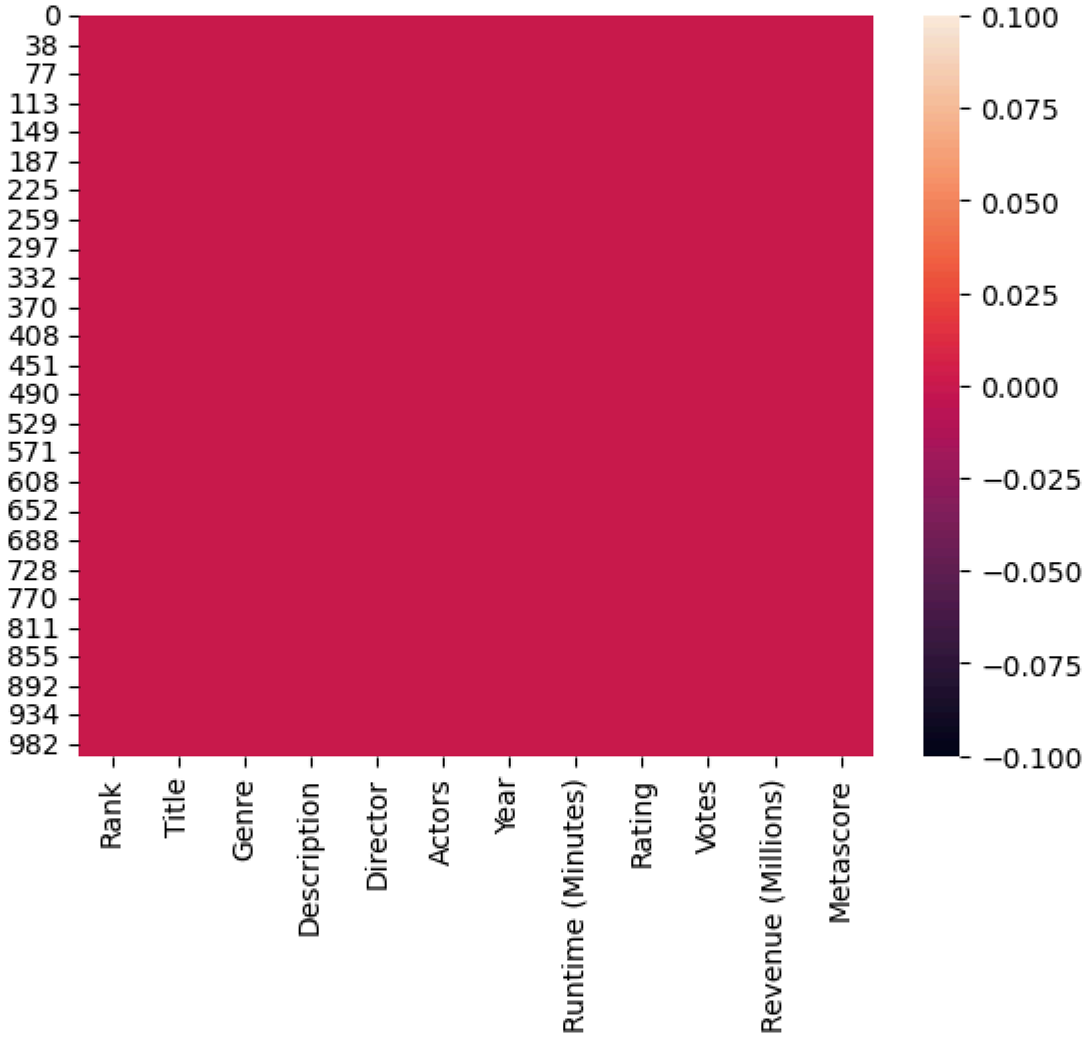
	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02
...	...	...	...	...	...	...	...	...	...	...	...
993	994	Resident Evil: Afterlife	Action,Adventure,Horror	While still out to destroy the evil Umbrella C...	Paul W.S. Anderson	Milla Jovovich, Ali Larter, Wentworth Miller,K...	2010	97	5.9	140900	60.13
994	995	Project X	Comedy	3 high school	Nima Nourizadeh	Thomas Mann, Oliver Cooper,	2012	88	6.7	164088	54.72

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)
				seniors throw a birthday party t...		Jonathan Daniel Br...					
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	94	5.5	73152	17.54
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008	98	6.2	70699	58.01
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016	87	5.3	12435	19.64

838 rows × 12 columns

In [15]:

```
# again visualize & cross check to check null values is removed or not?

sns.heatmap(data.isnull()) # no white lines indicates null values is permanently removed
plt.show()
```



In [16]:
```
# let's check is there any duplicate values is present or not?

dup_data = data.duplicated().any()
dup_data
```

Out[16]:
```
False
```

In [17]:
```
# fill the null values with the median cause if we fill average values in the null cell then average also represent wrong anal

data["Revenue (Millions)"] = data["Revenue (Millions)"].fillna(data["Revenue (Millions)"].median())
```

In [18]:
```
# print the dataset after the fill median values in the blank cell in the Revenue(Millions) column
data
```

Out[18]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02
...	...	...	...	...	...	...	...	...	...	...	..
993	994	Resident Evil: Afterlife	Action,Adventure,Horror	While still out to destroy the evil Umbrella C...	Paul W.S. Anderson	Milla Jovovich, Ali Larter, Wentworth Miller,K...	2010	97	5.9	140900	60.13
994	995	Project X	Comedy	3 high school	Nima Nourizadeh	Thomas Mann, Oliver Cooper,	2012	88	6.7	164088	54.72

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)
				seniors throw a birthday party t...		Jonathan Daniel Br...					
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	94	5.5	73152	17.54
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008	98	6.2	70699	58.01
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016	87	5.3	12435	19.64

838 rows × 12 columns

In [19]:

```
# let's check is there is null value still present in the revenue column

data["Revenue (Millions)"].isnull()
```

Out[19]:

```
0      False
1      False
2      False
3      False
4      False
       ...  
993    False
994    False
996    False
997    False
999    False
Name: Revenue (Millions), Length: 838, dtype: bool
```

In [20]:

```
# check if the median values is 0 or not for the revenue column

data["Revenue (Millions)"].median()
```

Out[20]:

```
48.150000000000006
```

In [21]:

```
# let's check still any blank cell in the metascore columns

data["Metascore"].hasnans
```

Out[21]:

```
False
```

In [22]:

```
# let's fill the blank cell of metascore with median values.

data["Metascore"] = data["Metascore"].fillna(data["Metascore"].median())
```

In [23]:

```
# let's check again there is any nan or blank values is present

data["Revenue (Millions)"].hasnans
```

Out[23]:

```
False
```

In [24]:

```
# Getting information about our dataset like total number of rows,
# columns, datatype of each column & memory requirement

data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 838 entries, 0 to 999
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Rank                838 non-null    int64  
 1   Title               838 non-null    object 
 2   Genre               838 non-null    object 
 3   Description         838 non-null    object 
 4   Director            838 non-null    object 
 5   Actors              838 non-null    object 
 6   Year                838 non-null    int64  
 7   Runtime (Minutes)   838 non-null    int64  
 8   Rating              838 non-null    float64
 9   Votes               838 non-null    int64  
 10  Revenue (Millions)  838 non-null    float64
 11  Metascore           838 non-null    float64
dtypes: float64(3), int64(4), object(5)
memory usage: 85.1+ KB
```

In [25]:

```
# get statistics about the dataframe

data.describe()
```

Out[25]:

	Rank	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
count	838.000000	838.00000	838.000000	838.000000	8.380000e+02	838.000000	838.000000
mean	485.247017	2012.50716	114.638425	6.814320	1.932303e+05	84.564558	59.575179
std	286.572065	3.17236	18.470922	0.877754	1.930990e+05	104.520227	16.952416
min	1.000000	2006.00000	66.000000	1.900000	1.780000e+02	0.000000	11.000000
25%	238.250000	2010.00000	101.000000	6.300000	6.127650e+04	13.967500	47.000000
50%	475.500000	2013.00000	112.000000	6.900000	1.368795e+05	48.150000	60.000000
75%	729.750000	2015.00000	124.000000	7.500000	2.710830e+05	116.800000	72.000000
max	1000.000000	2016.00000	187.000000	9.000000	1.791916e+06	936.630000	100.000000

In [26]:
```
# to describe overall statistics of the data

data.describe(include="all")
```

Out[26]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	R (M
count	838.000000	838	838	838	838	838	838.00000	838.000000	838.000000	8.380000e+02	838
unique	NaN	837	189	838	524	834	NaN	NaN	NaN	NaN	
top	NaN	The Host	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	Ridley Scott	Jennifer Lawrence, Josh Hutcherson, Liam Hemsw...	NaN	NaN	NaN	NaN	
freq	NaN	2	50	1	8	2	NaN	NaN	NaN	NaN	
mean	485.247017	NaN	NaN	NaN	NaN	NaN	2012.50716	114.638425	6.814320	1.932303e+05	84
std	286.572065	NaN	NaN	NaN	NaN	NaN	3.17236	18.470922	0.877754	1.930990e+05	104
min	1.000000	NaN	NaN	NaN	NaN	NaN	2006.00000	66.000000	1.900000	1.780000e+02	0
25%	238.250000	NaN	NaN	NaN	NaN	NaN	2010.00000	101.000000	6.300000	6.127650e+04	13
50%	475.500000	NaN	NaN	NaN	NaN	NaN	2013.00000	112.000000	6.900000	1.368795e+05	48
75%	729.750000	NaN	NaN	NaN	NaN	NaN	2015.00000	124.000000	7.500000	2.710830e+05	116
max	1000.000000	NaN	NaN	NaN	NaN	NaN	2016.00000	187.000000	9.000000	1.791916e+06	936

In [27]:
```
# How many movie have exceptiona# ly long runtimes <=180 runtimes

a = data["Runtime (Minutes)"]>=180
a.value_counts()
```

Out[27]:

```
Runtime (Minutes)
False    835
True       3
Name: count, dtype: int64
```

In [28]:

```python
# which movie have exceptionaly long runtimes <=180 runtimes

a = data["Runtime (Minutes)"]>=180
b = data["Title"]

ans = b.where(a).dropna()
ans
```

Out[28]:

```
82     The Wolf of Wall Street
88           The Hateful Eight
311             La vie d'Adèle
Name: Title, dtype: object
```

In [29]:

```python
# other method which movie have exceptionaly long runtimes <=180 runtimes

data[data["Runtime (Minutes)"]>=180]["Title"]
```

Out[29]:

```
82     The Wolf of Wall Street
88           The Hateful Eight
311             La vie d'Adèle
Name: Title, dtype: object
```

In [30]:

```python
# which film has max voting

a = data["Votes"].max()
print("Maximum Voting Of the Film:",a)
b = data["Title"][data["Votes"]==a]
b
```

```
Maximum Voting Of the Film: 1791916
```

Out[30]:

```
54    The Dark Knight
Name: Title, dtype: object
```

In [31]:

```python
# which film has min voting
```

```
a = data["Votes"].min()
print("Minimum Voting Of the Film:",a)
b = data["Title"][data["Votes"]==a]
b
```

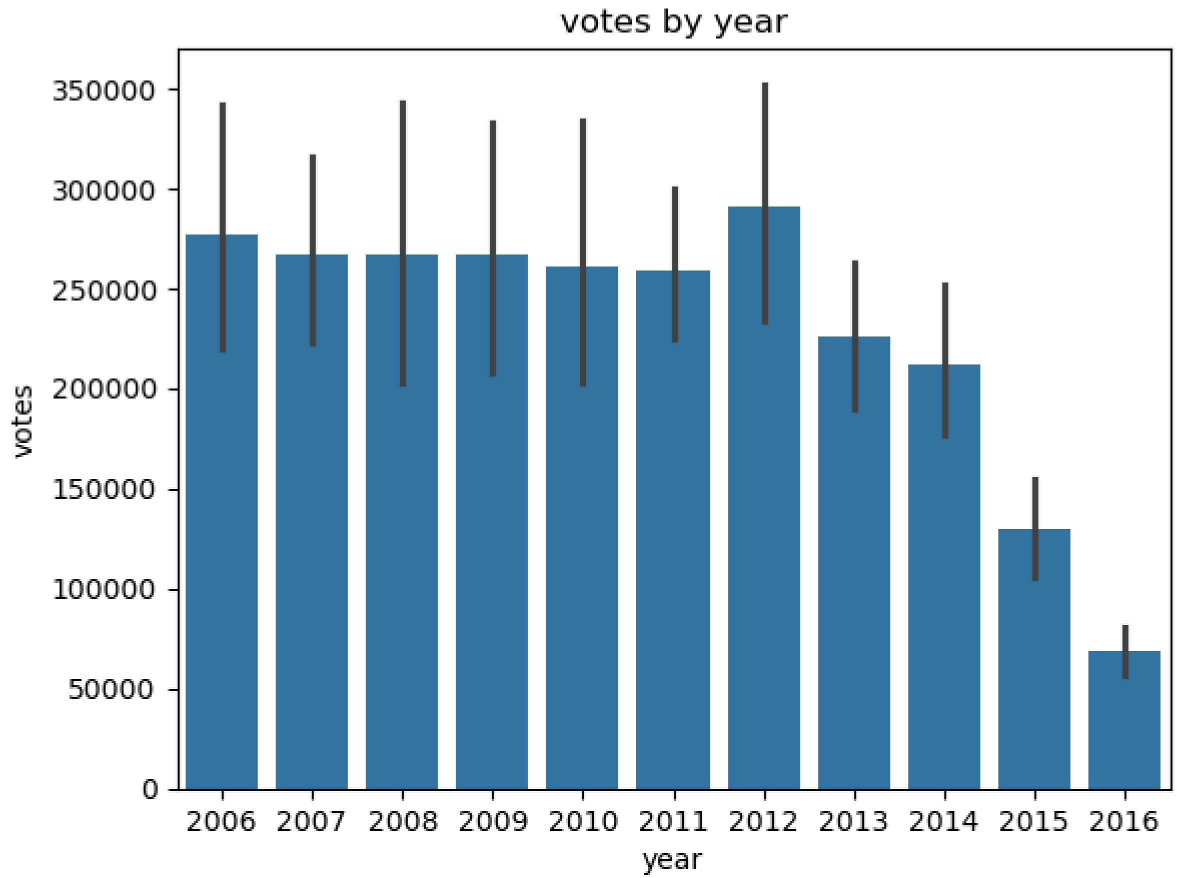
```
Minimum Voting Of the Film: 178
```

Out[31]:

```
250    Bonjour Anne
Name: Title, dtype: object
```

In [32]:

```
# year wise votes

sns.barplot(x="Year",y="Votes",data=data)
plt.title("votes by year")
plt.xlabel("year")
plt.ylabel("votes")
plt.show()
```



In [33]:

```
# which movie has highest revenue

a = data["Revenue (Millions)"].max()
print("Maximum Revenue Of the Film:",a)
b = data["Title"][data["Revenue (Millions)"]==a]
b
```

```
Maximum Revenue Of the Film: 936.63
```

Out[33]:

```
50    Star Wars: Episode VII - The Force Awakens
Name: Title, dtype: object
```

In [34]:

```
# other method which movie has highest revenue
```

```
data[data["Revenue (Millions)"].max()==data["Revenue (Millions)"]]["Title"]
```

Out[34]:

```
50    Star Wars: Episode VII - The Force Awakens
Name: Title, dtype: object
```

In [35]:

```
# which movie has lowest revenue

a = data["Revenue (Millions)"].min()
print("Minimum Revenue Of the Film:",a)
b = data["Title"][data["Revenue (Millions)"]==a]
b
```

```
Minimum Revenue Of the Film: 0.0
```

Out[35]:

```
231    A Kind of Murder
Name: Title, dtype: object
```

In [36]:

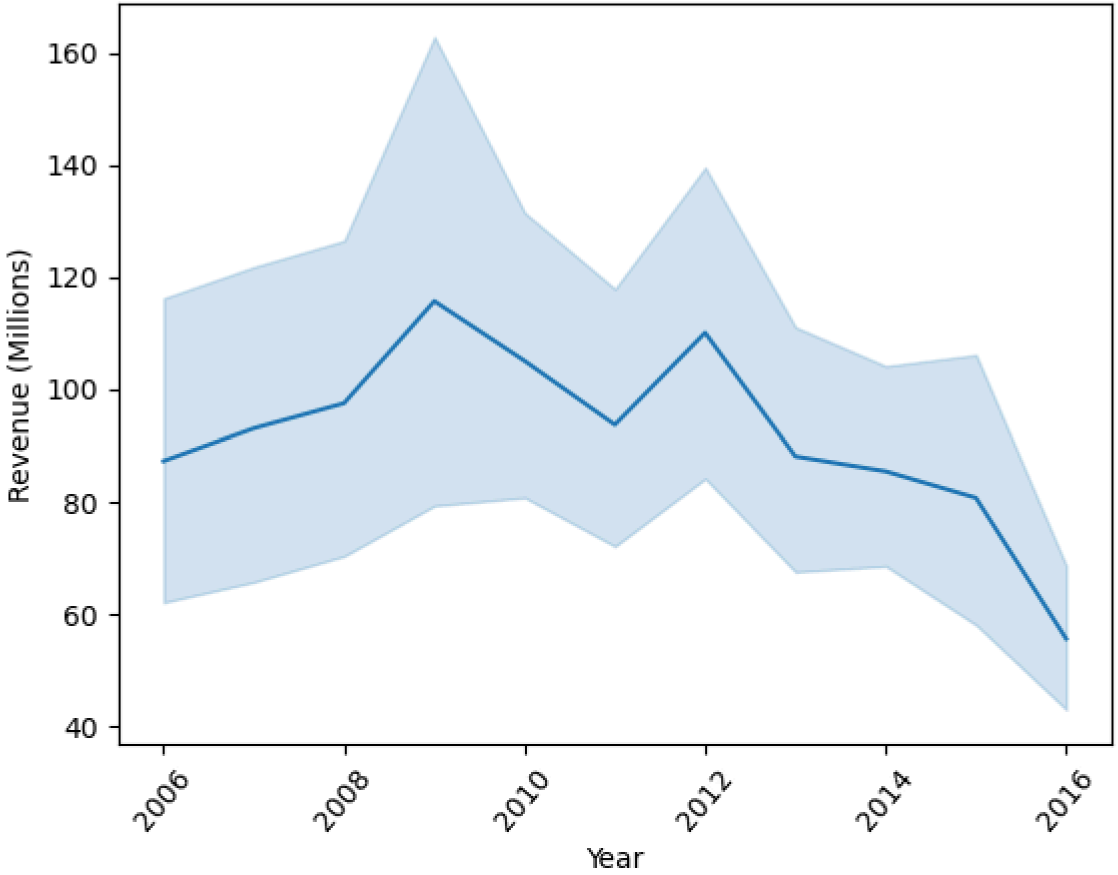
```
# other method which movie has highest revenue

data[data["Revenue (Millions)"].min()==data["Revenue (Millions)"]]["Title"]
```

Out[36]:

```
231    A Kind of Murder
Name: Title, dtype: object
```

In [37]:

```
# barplot for the categorical data so we are going to check the genre wise revenue

sns.lineplot(x="Year",y="Revenue (Millions)",data=data)
plt.xticks(rotation=50)
plt.show()
```



In [38]:

```
# which movie has highest Metascore

a = data["Metascore"].max()
print("Maximum Metascore Of the Film:",a)
b = data["Title"][data["Metascore"]==a]
b
```

```
Maximum Metascore Of the Film: 100.0
```

Out[38]:

```
656    Boyhood
Name: Title, dtype: object
```

In [39]:

```
# which movie has lowest Metascore
```

```
a = data["Metascore"].min()
print("Minimum Metascore Of the Film:",a)
b = data["Title"][data["Metascore"]==a]
b
```

```
Minimum Metascore Of the Film: 11.0
```

Out[39]:

```
999    Nine Lives
Name: Title, dtype: object
```

In [40]:

```
# which year saws highest no of votes

a = data["Votes"].max()
print("Highes no.of votes is :",a)
b = data["Year"][data["Votes"]==a]
b
```

```
Highes no.of votes is : 1791916
```

Out[40]:

```
54    2008
Name: Year, dtype: int64
```

In [41]:

```
# which year get more revenue over the time

a = data["Revenue (Millions)"].max()
print("Highes no.of revenue is :",a)
b = data["Year"][data["Revenue (Millions)"]==a]
b
```

```
Highes no.of revenue is : 936.63
```

Out[41]:

```
50    2015
Name: Year, dtype: int64
```

In [42]:

```
# average rating by movie genre

a = data["Rating"].mean()
print("Average ratings is :",a)
b = data["Genre"][data["Rating"]==a]
b
```

```
Average ratings is : 6.814319809069212
```

Out[42]:

```
Series([], Name: Genre, dtype: object)
```

In [43]:

```python
# which genre type of movie get high metascore

a = data["Metascore"].max()
b = data["Genre"][data["Metascore"]==a]
b
```

Out[43]:

```
656    Drama
Name: Genre, dtype: object
```

In [44]:

```python
# which runtime movie get high metascore

a = data["Runtime (Minutes)"]
b = data["Title"][data["Metascore"]==a].max()
b
```

Out[44]:

```
nan
```

In [45]:

```python
# what is the average runtime of movie get high metascore

a = data["Runtime (Minutes)"].mean()
print("Average runtime movie is :",a)
b = data["Title"][data["Metascore"]==a].max()
b
```

```
Average runtime movie is : 114.63842482100239
```

Out[45]:

```
nan
```

In [46]:

```python
# top 10 highest ratings wise movie name,revenue,ratings,runtime,director

a = data[["Title","Revenue (Millions)","Runtime (Minutes)","Rating","Director"]].sort_values(by = "Rating",ascending=False)
a.head(10)
```

Out[46]:

	Title	Revenue (Millions)	Runtime (Minutes)	Rating	Director
54	The Dark Knight	533.32	152	9.0	Christopher Nolan
80	Inception	292.57	148	8.8	Christopher Nolan
36	Interstellar	187.99	169	8.6	Christopher Nolan
249	The Intouchables	13.18	112	8.6	Olivier Nakache
96	Kimi no na wa	4.68	106	8.6	Makoto Shinkai
124	The Dark Knight Rises	448.13	164	8.5	Christopher Nolan
991	Taare Zameen Par	1.20	165	8.5	Aamir Khan
133	Whiplash	13.09	107	8.5	Damien Chazelle
99	The Departed	132.37	151	8.5	Martin Scorsese
476	The Lives of Others	11.28	137	8.5	Florian Henckel von Donnersmarck

In [47]:

```
# other method to get top 10 rating wise title,rating,director,revenue,runtime

a = data.nlargest(10,"Rating")[["Title","Rating","Director","Revenue (Millions)","Runtime (Minutes)"]]
a.set_index("Title") # for title wise get the data
```

Out[47]:

	Rating	Director	Revenue (Millions)	Runtime (Minutes)
Title				
The Dark Knight	9.0	Christopher Nolan	533.32	152
Inception	8.8	Christopher Nolan	292.57	148
Interstellar	8.6	Christopher Nolan	187.99	169
Kimi no na wa	8.6	Makoto Shinkai	4.68	106
The Intouchables	8.6	Olivier Nakache	13.18	112
The Prestige	8.5	Christopher Nolan	53.08	130
The Departed	8.5	Martin Scorsese	132.37	151
The Dark Knight Rises	8.5	Christopher Nolan	448.13	164
Whiplash	8.5	Damien Chazelle	13.09	107
The Lives of Others	8.5	Florian Henckel von Donnersmarck	11.28	137

In [48]:

```python
# director wise average ratings

data.groupby("Director")["Rating"].mean().sort_values(ascending=False)
```

Out[48]:

```
Director
Christopher Nolan                   8.68
Olivier Nakache                     8.60
Makoto Shinkai                      8.60
Florian Henckel von Donnersmarck    8.50
Aamir Khan                          8.50
                                    ...
Sam Taylor-Johnson                  4.10
Joey Curtis                         4.00
George Nolfi                        3.90
James Wong                          2.70
Jason Friedberg                     1.90
Name: Rating, Length: 524, dtype: float64
```

In [49]:

```
# director wise metascore and rating and revenue

director = data[["Director","Revenue (Millions)","Rating","Metascore"]].groupby(by = "Director")
director.head()
```

Out[49]:

	Director	Revenue (Millions)	Rating	Metascore
0	James Gunn	333.13	8.1	76.0
1	Ridley Scott	126.46	7.0	65.0
2	M. Night Shyamalan	138.12	7.3	62.0
3	Christophe Lourdelet	270.32	7.2	59.0
4	David Ayer	325.02	6.2	40.0
...	...	...	...	...
991	Aamir Khan	1.20	8.5	42.0
994	Nima Nourizadeh	54.72	6.7	48.0
996	Eli Roth	17.54	5.5	46.0
997	Jon M. Chu	58.01	6.2	50.0
999	Barry Sonnenfeld	19.64	5.3	11.0

832 rows × 4 columns

In [50]:

```
# year wise revenue,rating,votes,metascore
# which year get highest film releases over the year
# movie production is increases or decreases by year or not?

g1 = data.groupby(by = "Year")
print(g1)
g1.count()
```

```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x000002280AFF9D90>
```

Out[50]:

	Rank	Title	Genre	Description	Director	Actors	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
Year											
2006	41	41	41	41	41	41	41	41	41	41	41
2007	44	44	44	44	44	44	44	44	44	44	44
2008	48	48	48	48	48	48	48	48	48	48	48
2009	45	45	45	45	45	45	45	45	45	45	45
2010	57	57	57	57	57	57	57	57	57	57	57
2011	57	57	57	57	57	57	57	57	57	57	57
2012	62	62	62	62	62	62	62	62	62	62	62
2013	84	84	84	84	84	84	84	84	84	84	84
2014	93	93	93	93	93	93	93	93	93	93	93
2015	109	109	109	109	109	109	109	109	109	109	109
2016	198	198	198	198	198	198	198	198	198	198	198

In [51]:

```
# genure wise revenue,rating,votes,metascore
# # which type of movie trends in term of runtime, genre,director & their audience

g1 = data.groupby(by = "Genre")
print(g1)
g1.count()
```

```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x000002280B0371D0>
```

Out[51]:

	Rank	Title	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
Genre											
Action	1	1	1	1	1	1	1	1	1	1	1
Action,Adventure	3	3	3	3	3	3	3	3	3	3	3
Action,Adventure,Biography	2	2	2	2	2	2	2	2	2	2	2
Action,Adventure,Comedy	14	14	14	14	14	14	14	14	14	14	14
Action,Adventure,Crime	6	6	6	6	6	6	6	6	6	6	6
...	...	...	...	...	...	...	...	...	...	...	...
Mystery,Sci-Fi,Thriller	3	3	3	3	3	3	3	3	3	3	3
Mystery,Thriller	6	6	6	6	6	6	6	6	6	6	6
Sci-Fi	1	1	1	1	1	1	1	1	1	1	1
Sci-Fi,Thriller	1	1	1	1	1	1	1	1	1	1	1
Thriller	1	1	1	1	1	1	1	1	1	1	1

189 rows × 11 columns

In [52]:

```
# print all the columns in the dataframe

data.columns
```

Out[52]:

```
Index(['Rank', 'Title', 'Genre', 'Description', 'Director', 'Actors', 'Year',
       'Runtime (Minutes)', 'Rating', 'Votes', 'Revenue (Millions)',
       'Metascore'],
      dtype='object')
```

In [53]:

```
# which genre film is frequently releases

data["Genre"].mode()
```

```
Out[53]: 0    Action,Adventure,Sci-Fi
         Name: Genre, dtype: object
```

```
In [54]: # which actors pairs/combinations is frequently in the films

         data["Actors"].mode()
```

```
Out[54]: 0    Daniel Radcliffe, Emma Watson, Rupert Grint, M...
         1    Gerard Butler, Aaron Eckhart, Morgan Freeman,A...
         2    Jennifer Lawrence, Josh Hutcherson, Liam Hemsw...
         3    Shia LaBeouf, Megan Fox, Josh Duhamel, Tyrese ...
         Name: Actors, dtype: object
```

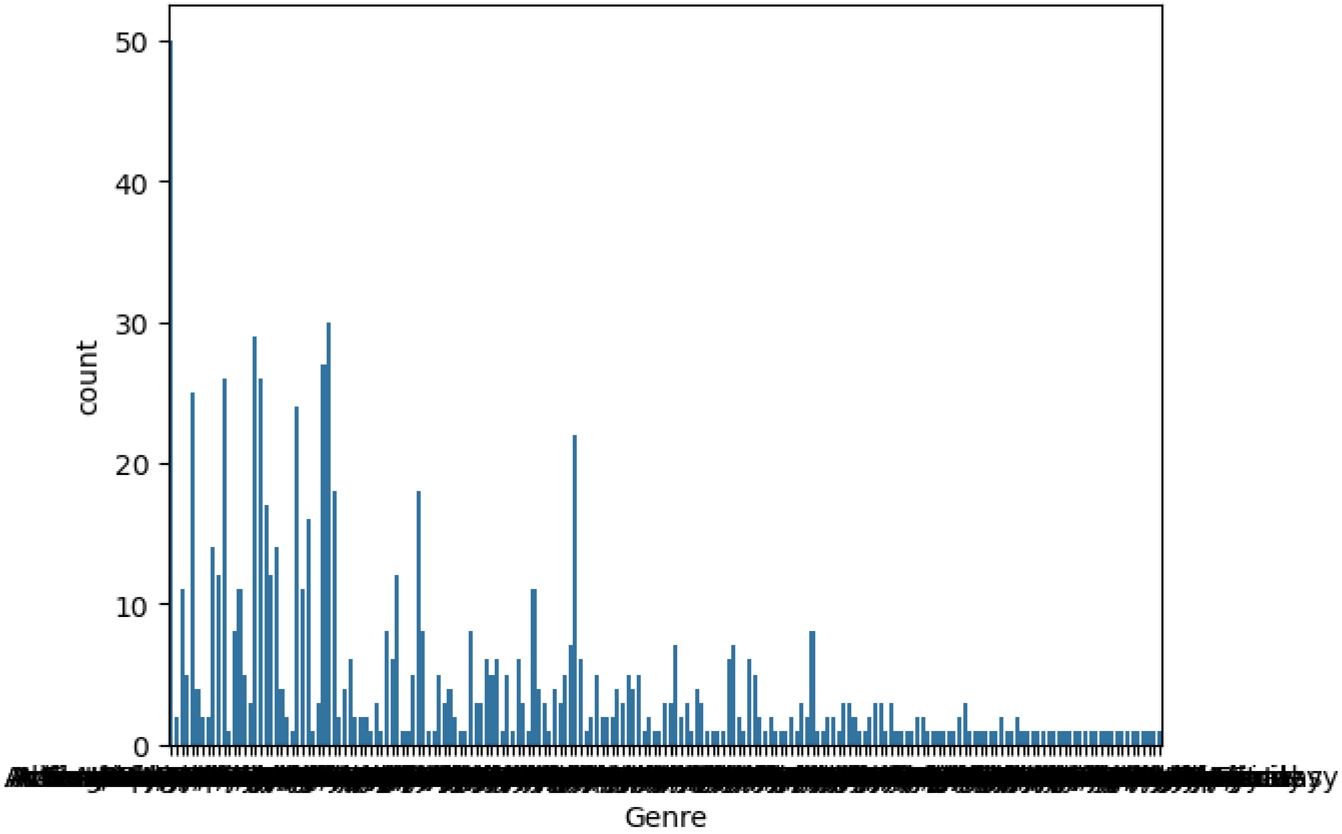
```
In [55]: # which director makes the frequently films and releases

         data["Director"].mode()
```

```
Out[55]: 0    Ridley Scott
         Name: Director, dtype: object
```

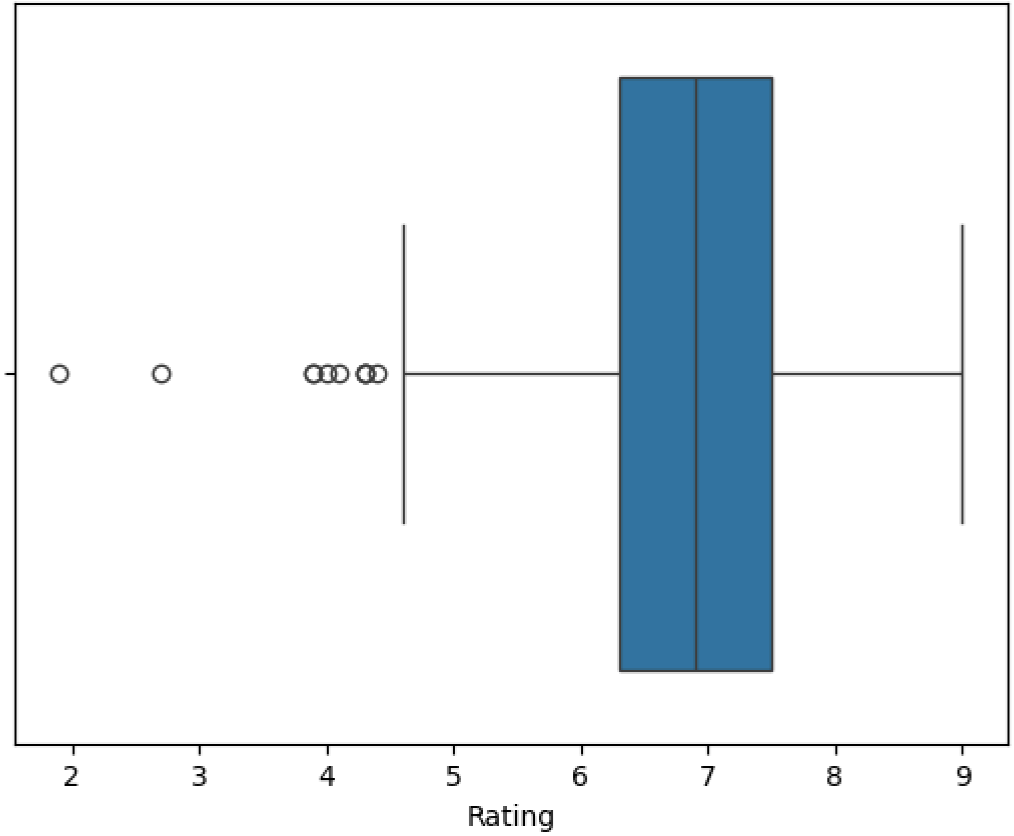
```
In [56]: # variance  & mean of the dataframe

         print("mean of the runtime in minutes",data["Runtime (Minutes)"].mean())
         print("variance of the runtime in minutes",data["Runtime (Minutes)"].var())
         diff = data["Runtime (Minutes)"].var() - data["Runtime (Minutes)"].mean()
         print("the spreadness of the data means the difference of between var and mean values",diff)
```

```
mean of the runtime in minutes 114.63842482100239
variance of the runtime in minutes 341.17496143460437
the spreadness of the data means the difference of between var and mean values 226.53653661360198
```

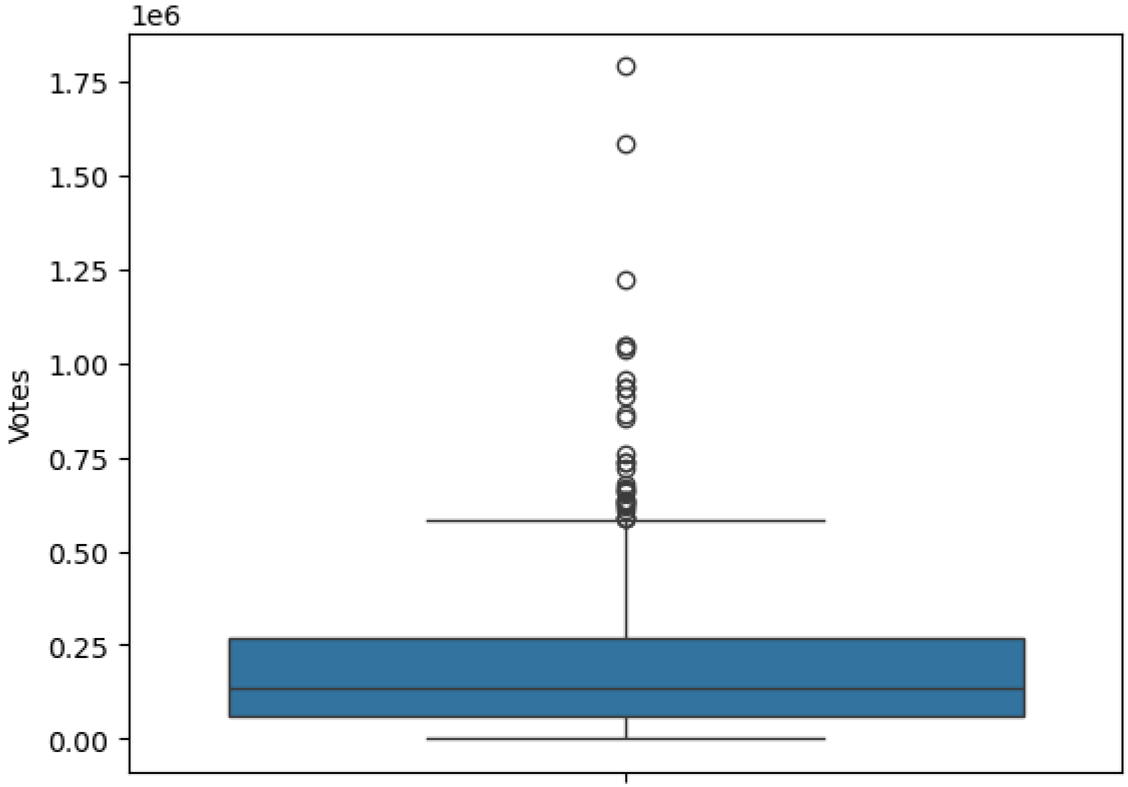
```
In [57]: # count plot for the genre

         sns.countplot(x="Genre",data=data)
         plt.show()
```



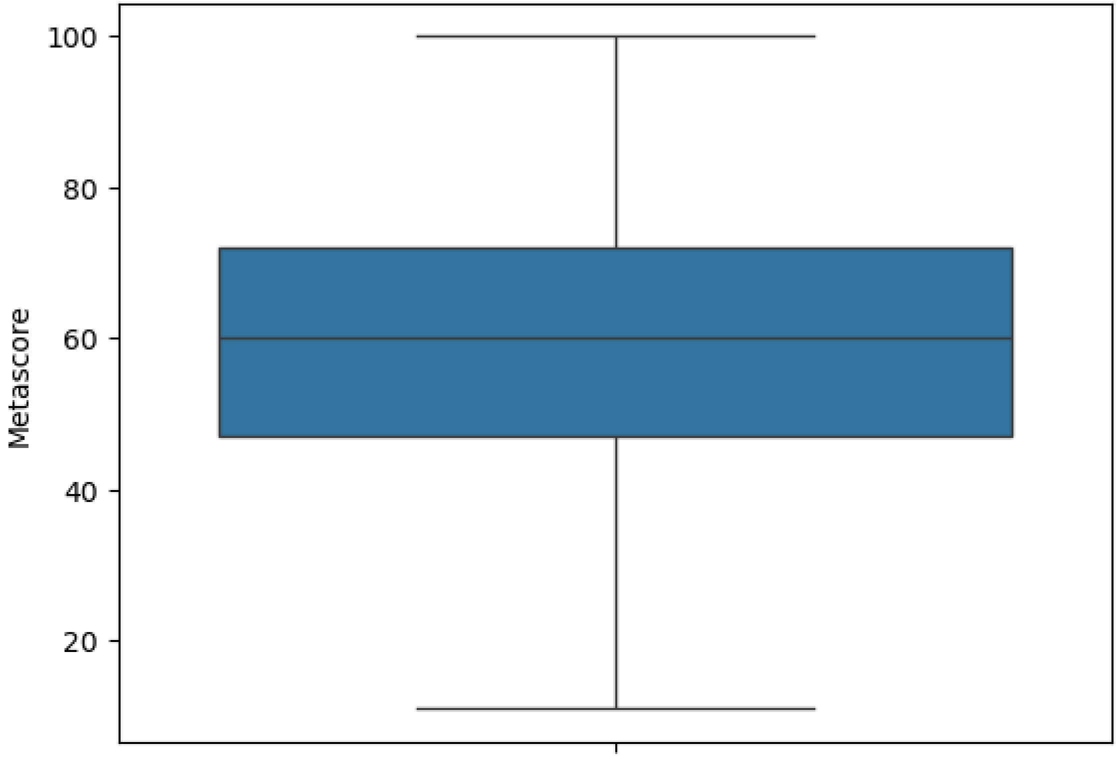
In [58]:
```
# for checking the outliers of the Rating data

sns.boxplot(x="Rating",data=data)
plt.show()
```



In [59]:
```
# for checking the outliers of the votes data

sns.boxplot(y="Votes",data=data)
plt.show()
```

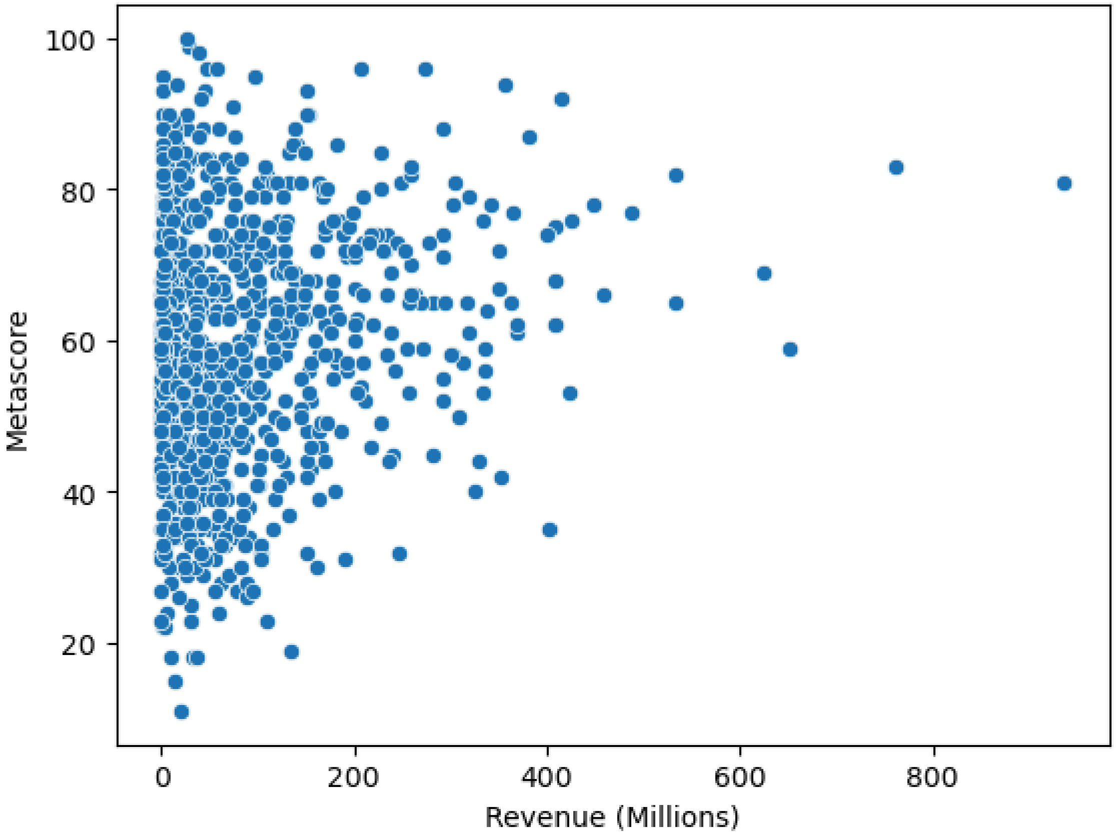


In [60]:
```
# for checking the outliers of the metascore data

sns.boxplot(y="Metascore",data=data)
plt.show()
```

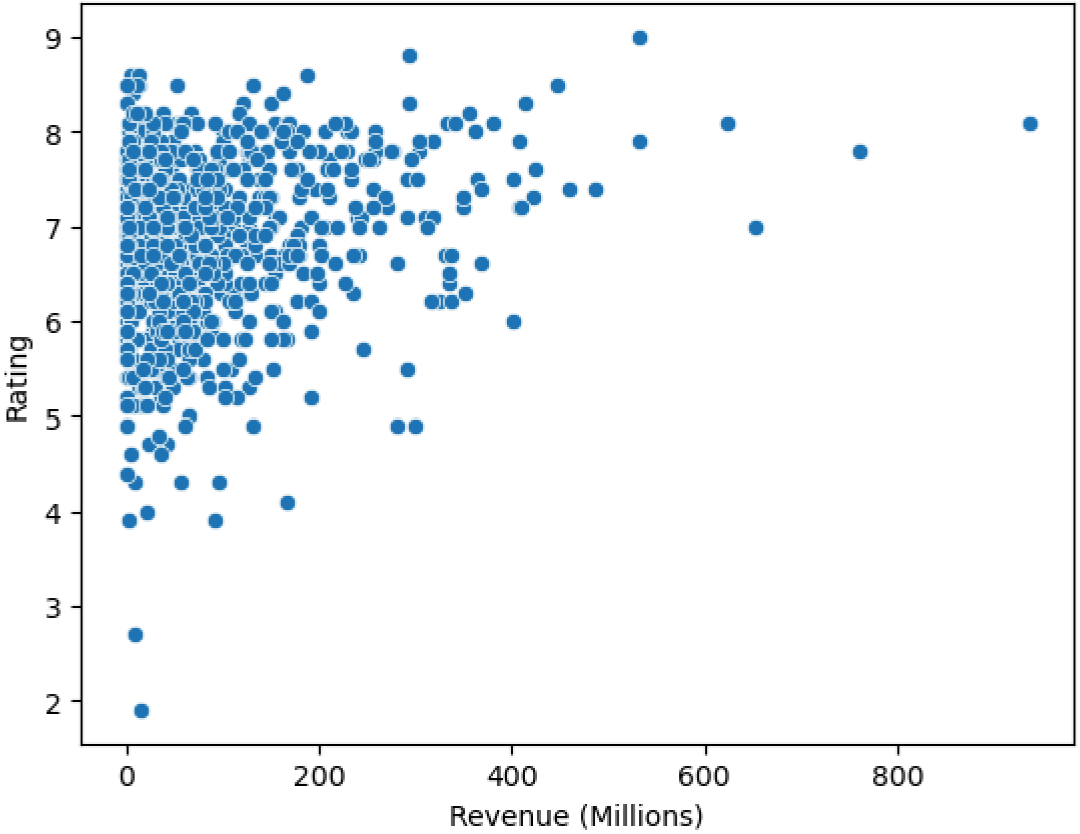


In [61]:

```
# check the relation between two variables by using scatterplot

sns.scatterplot(x="Revenue (Millions)",y="Metascore",data=data)
plt.show()
```

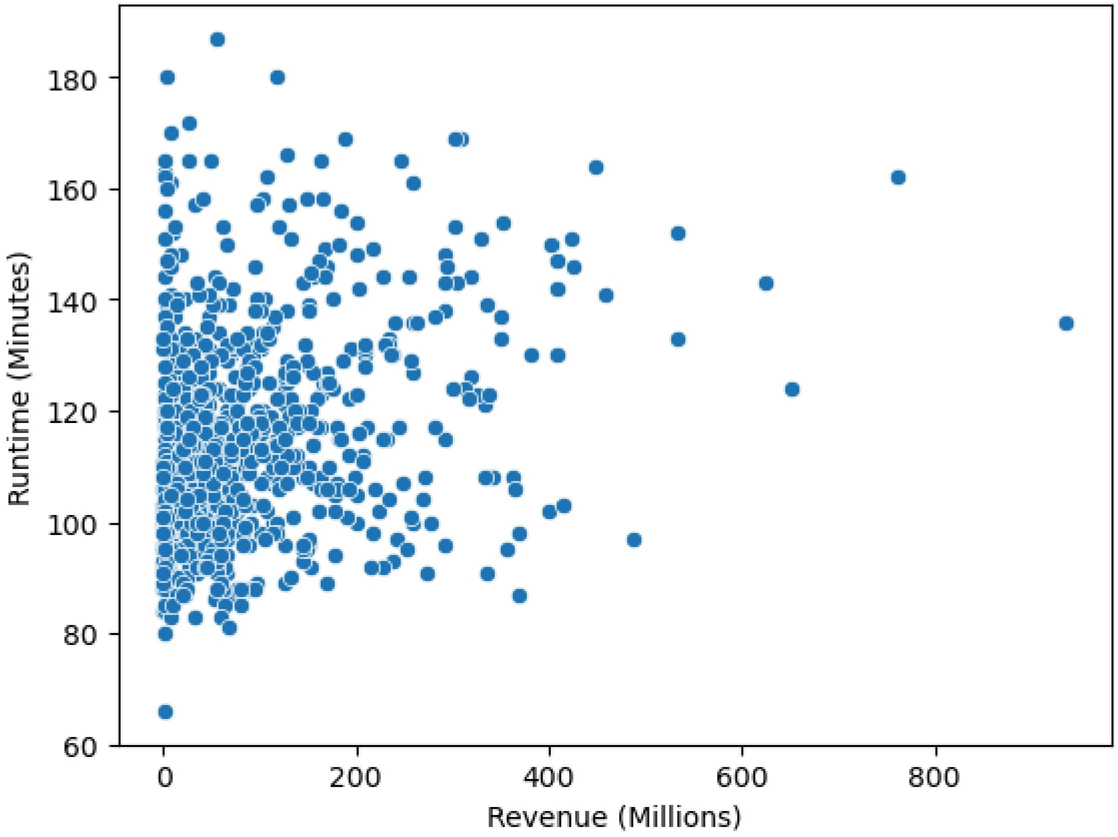


In [62]:
```
# check the relation between two variables by using scatterplot

sns.scatterplot(x="Revenue (Millions)",y="Rating",data=data)
plt.show()
```

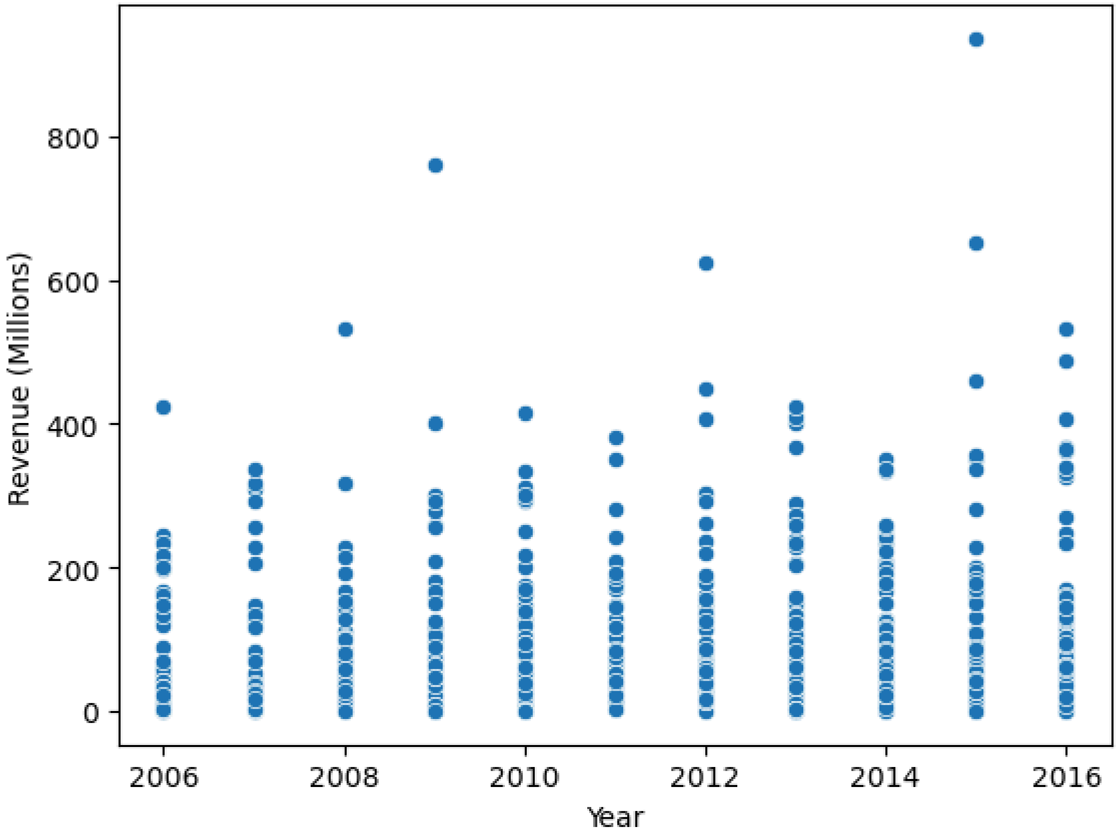


In [63]:

```
# check the relation between two variables by using scatterplot

sns.scatterplot(x="Revenue (Millions)",y="Runtime (Minutes)",data=data)
plt.show()
```

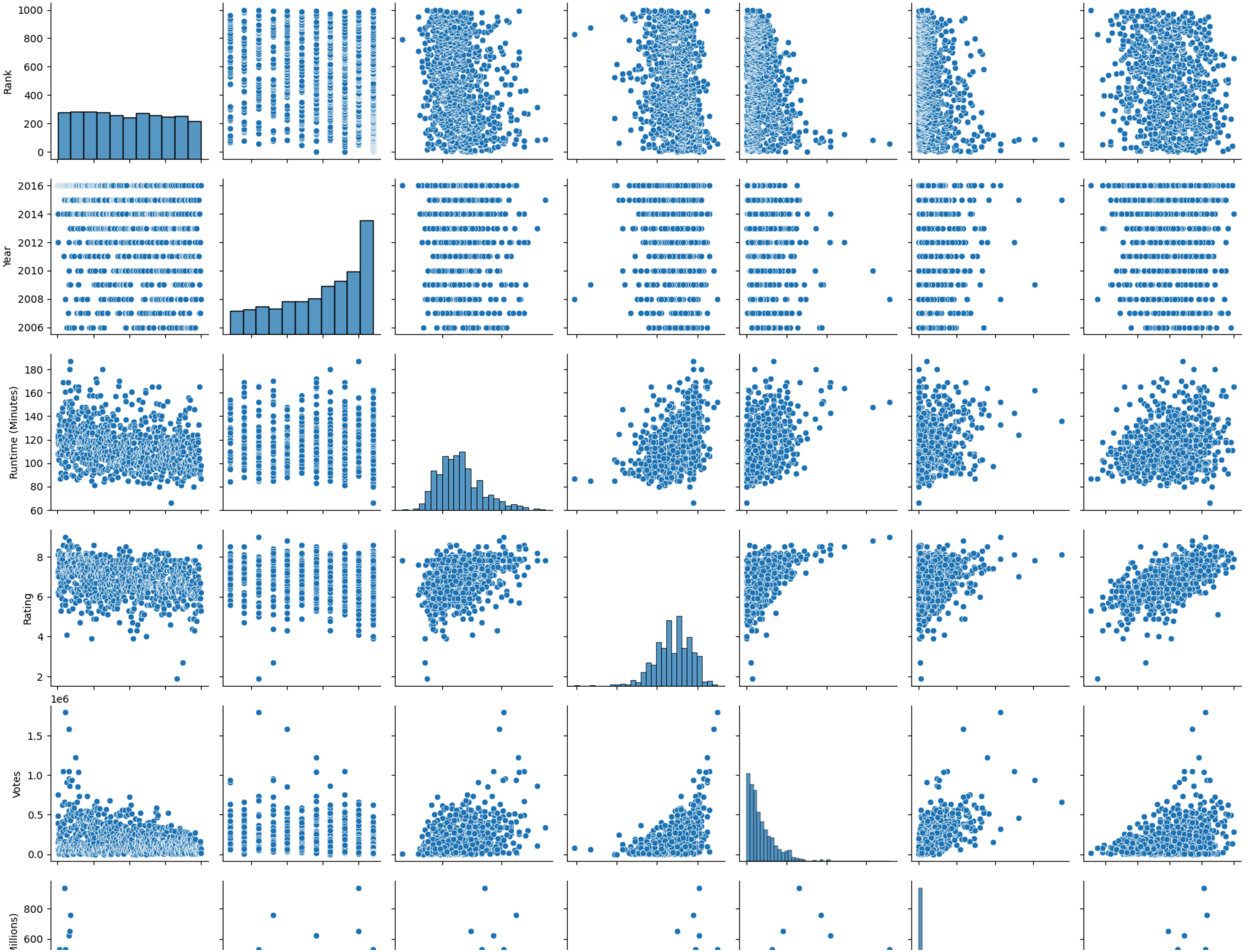


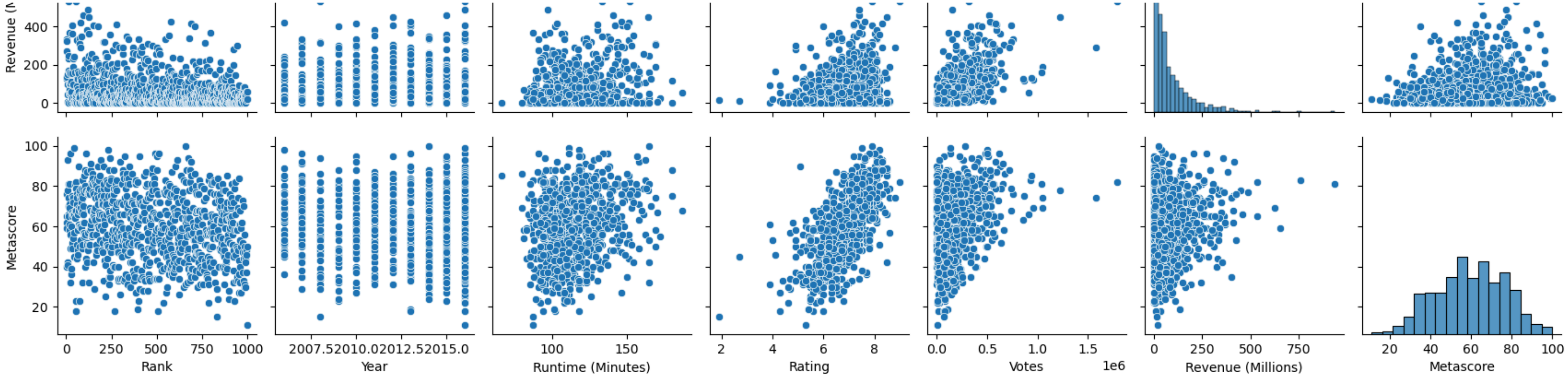
In [64]:
```
# check the relation between two variables by using scatterplot

sns.scatterplot(y="Revenue (Millions)",x="Year",data=data)
plt.show()
```



In [65]:

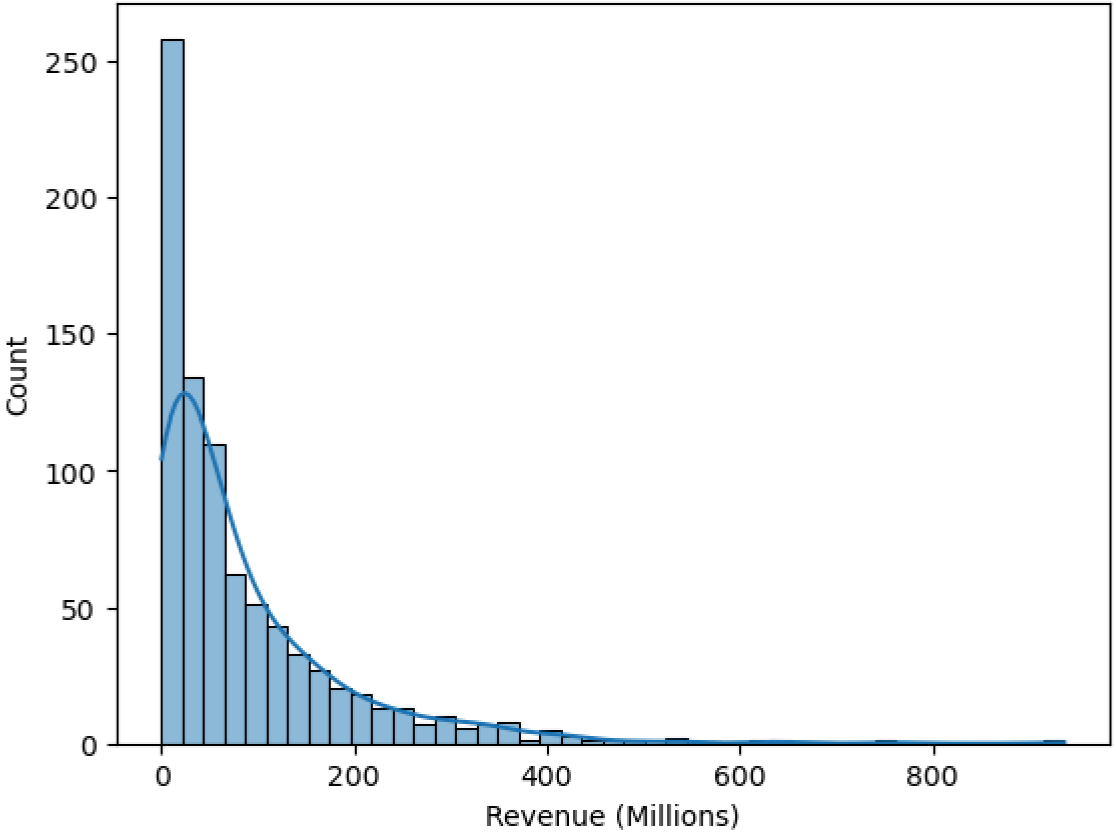
```
# create a pairplot

sns.pairplot(data)
plt.show()
```


Rank
1000
800
600
400
200
0
Year
2016
2014
2012
2010
2008
2006
Runtime (Minutes)
180
160
140
120
100
80
60
Rating
8
6
4
2
Votes
1e6
1.5
1.0
0.5
0.0
illions)
800
600

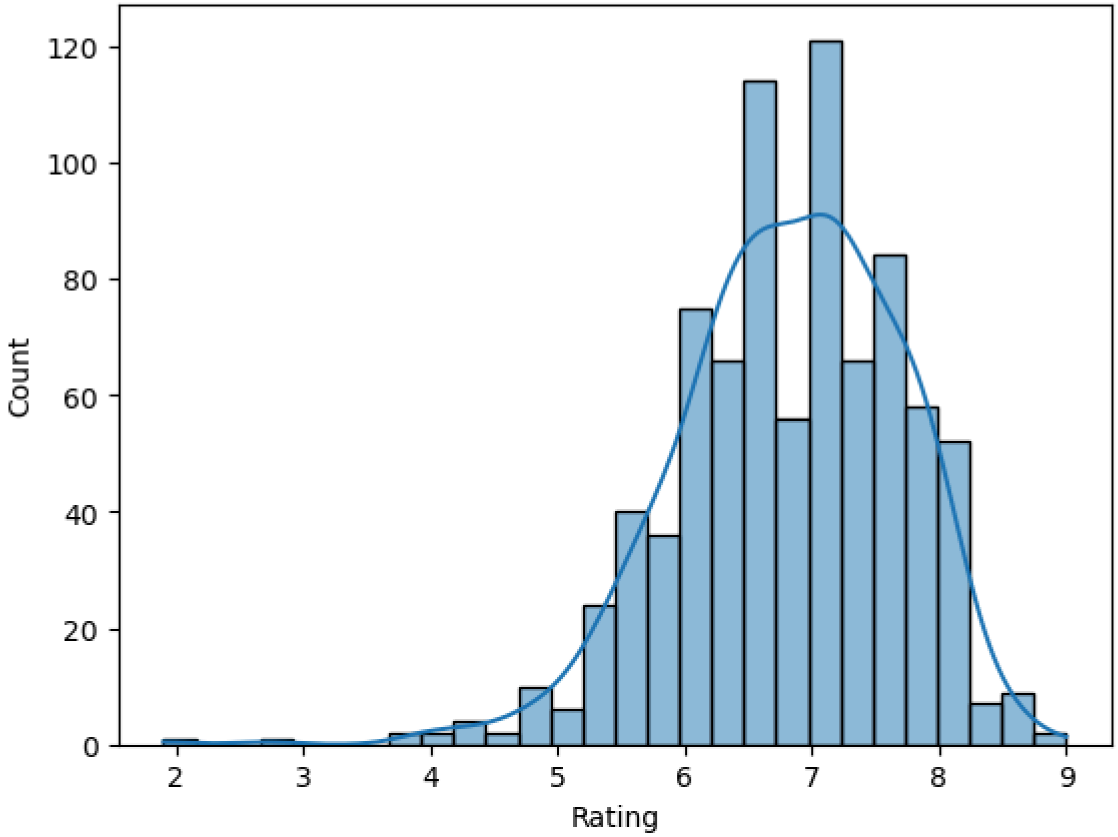


In [66]:

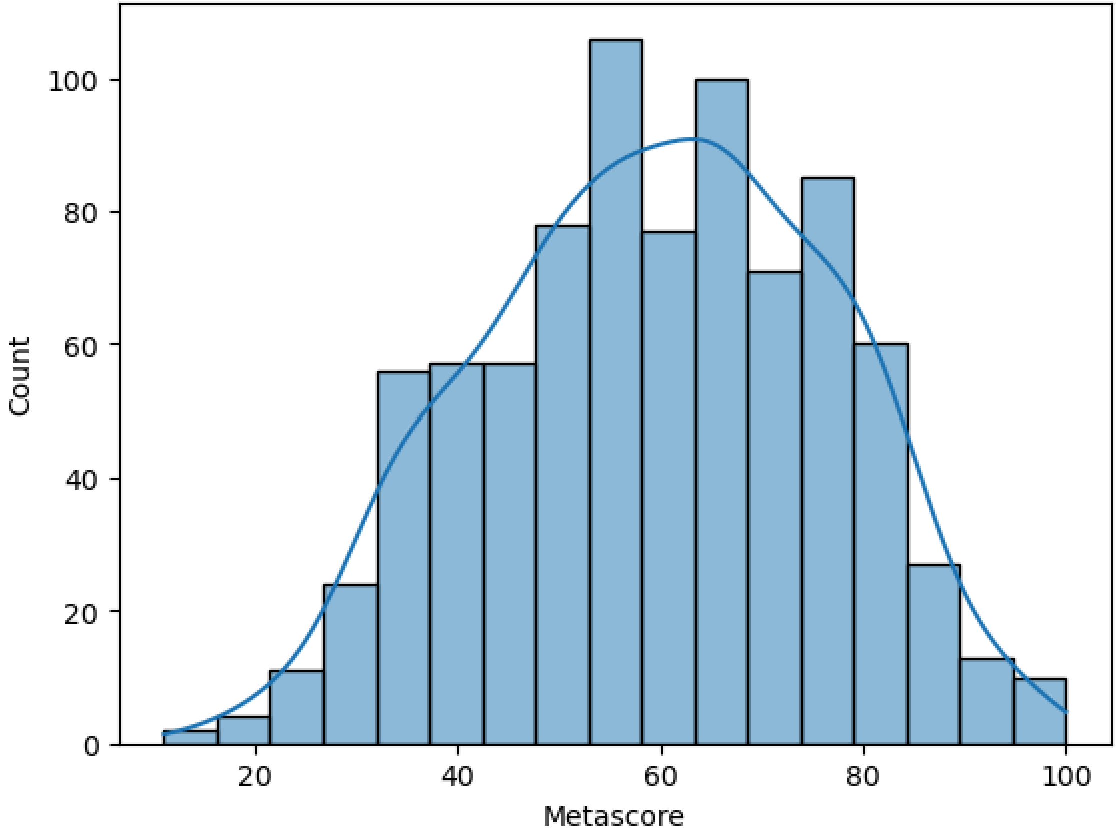
```
# to check the distribution

sns.histplot(x="Revenue (Millions)",data=data,kde=True)
plt.show()
```



In [67]:
```
# to check the distribution

sns.histplot(x="Rating",data=data,kde=True)
plt.show()
```

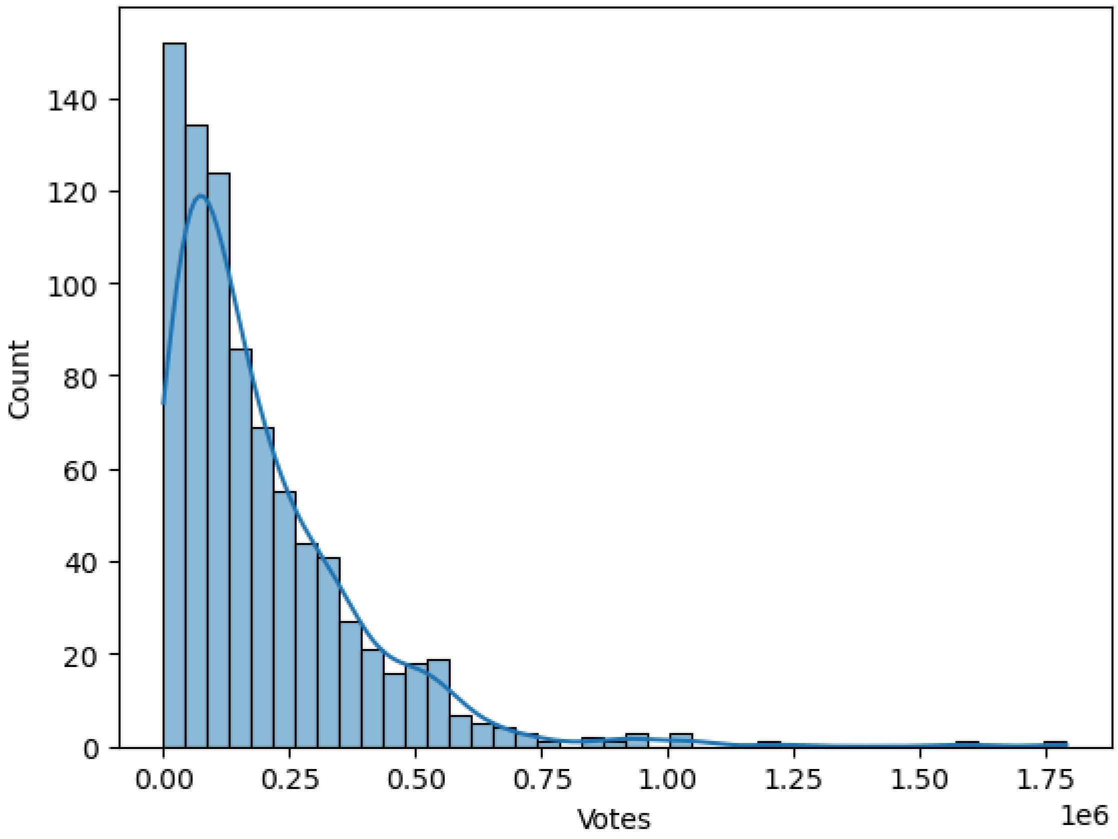


In [68]:
```
# to check the distribution

sns.histplot(x="Metascore",data=data,kde=True)
plt.show()
```



In [69]:

```
# to check the distribution

sns.histplot(x="Votes",data=data,kde=True)
plt.show()
```



In [70]:

```
# column wise standard deviations

data.std(axis=1,skipna=True,numeric_only=True)
```

```
Out[70]: 0      285987.140710
         1      183476.958712
         2       59426.314556
         3       22740.119443
         4      148658.129981
                    ...
         993     53058.592161
         994     61822.077119
         996     27459.556103
         997     26529.404756
         999      4565.302594
         Length: 838, dtype: float64
```

```
In [71]: # row wise standard deviations

         data.std(axis=0,skipna=True,numeric_only=True)
```

```
Out[71]: Rank                       286.572065
         Year                         3.172360
         Runtime (Minutes)           18.470922
         Rating                       0.877754
         Votes                   193099.005104
         Revenue (Millions)         104.520227
         Metascore                   16.952416
         dtype: float64
```

```
In [72]: # column wise variance

         data.var(axis=1,skipna=True,numeric_only=True)
```

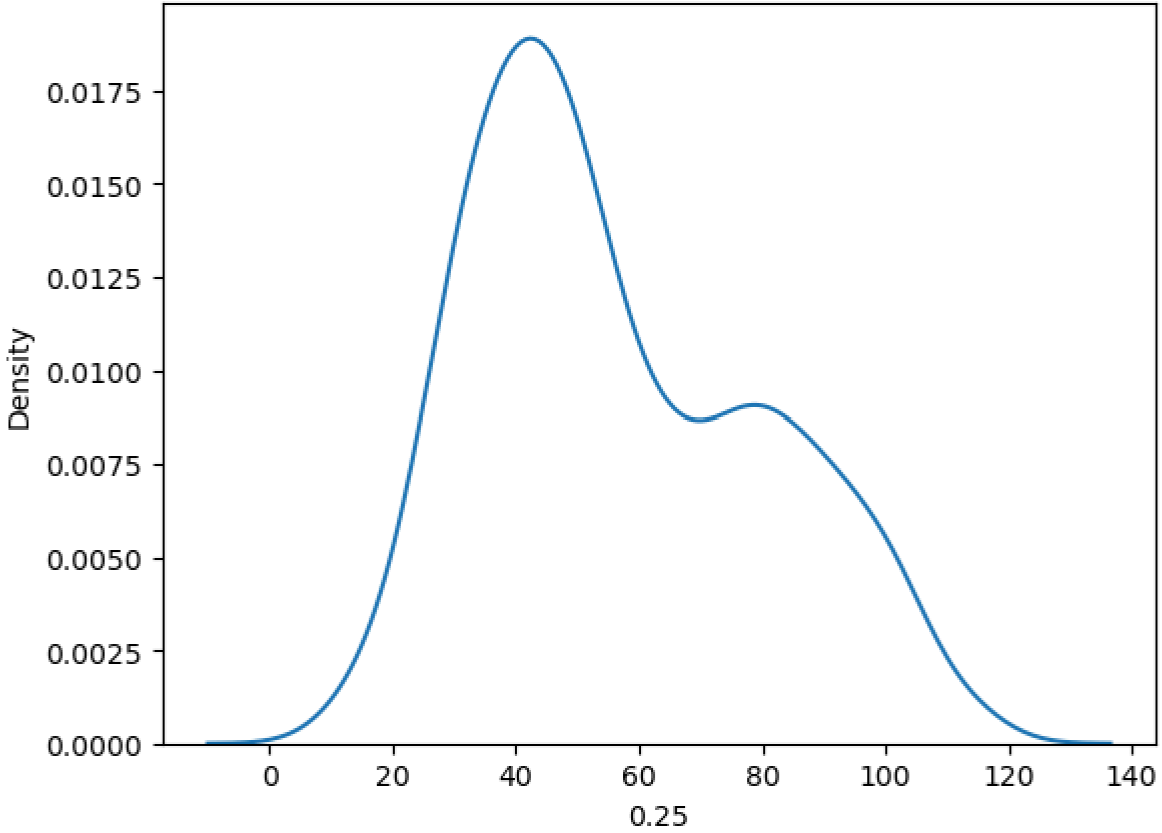
```
Out[72]: 0      8.178864e+10
         1      3.366379e+10
         2      3.531487e+09
         3      5.171130e+08
         4      2.209924e+10
                    ...
         993    2.815214e+09
         994    3.821969e+09
         996    7.540272e+08
         997    7.038093e+08
         999    2.084199e+07
         Length: 838, dtype: float64
```

```
In [73]: # row wise variance

         data.var(axis=0,skipna=True,numeric_only=True)
```

```
Out[73]: Rank                    8.212355e+04
         Year                    1.006387e+01
         Runtime (Minutes)       3.411750e+02
         Rating                  7.704518e-01
         Votes                   3.728723e+10
         Revenue (Millions)      1.092448e+04
         Metascore               2.873844e+02
         dtype: float64
```

```
In [74]: # let's check the bell shape curve means proper distribution curve is made by our dataset or not? row wise

         a = data.quantile(q=0.25,axis=1,numeric_only=True)

         sns.kdeplot(a)
```

```
Out[74]: <Axes: xlabel='0.25', ylabel='Density'>
```



In [75]:

```
# which year saws highest no of votes

a = data["Votes"].max()
print("Highes no.of votes is :",a)
b = data["Year"][data["Votes"]==a]
b
```

```
Highes no.of votes is : 1791916
```

Out[75]:

```
54    2008
Name: Year, dtype: int64
```

In [76]:

```
a = data.Votes.max()
data[data["Votes"]==a]
```

Out[76]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
54	55	The Dark Knight	Action,Crime,Drama	When the menace known as the Joker wreaks havo...	Christopher Nolan	Christian Bale, Heath Ledger, Aaron Eckhart,Mi...	2008	152	9.0	1791916	533.32	82.0

In [77]:

```python
# let's get the categorise rating like excellent,good,average

def rating(rating):
    if rating>=7.0:
        return "Excellent"
    elif rating>=6.0:
        return "Good"
    else:
        return "Average"
```

In [78]:

```python
# get the ratings

data["Rating_category"] = data["Rating"].apply(rating)
data["Rating_category"]
```

Out[78]:

```
0       Excellent
1       Excellent
2       Excellent
3       Excellent
4            Good
          ...
993       Average
994          Good
996       Average
997          Good
999       Average
Name: Rating_category, Length: 838, dtype: object
```

In [79]:

```
# print the data

data
```

Out[79]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02
...	...	...	...	...	...	...	...	...	...	...	...
993	994	Resident Evil: Afterlife	Action,Adventure,Horror	While still out to destroy the evil Umbrella C...	Paul W.S. Anderson	Milla Jovovich, Ali Larter, Wentworth Miller,K...	2010	97	5.9	140900	60.13
994	995	Project X	Comedy	3 high school	Nima Nourizadeh	Thomas Mann, Oliver Cooper,	2012	88	6.7	164088	54.72

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)
				seniors throw a birthday party t...		Jonathan Daniel Br...					
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	94	5.5	73152	17.54
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie Ventura,...	2008	98	6.2	70699	58.01
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016	87	5.3	12435	19.64

838 rows × 13 columns