Anurag Jaiswal

12017460

INT 353CA3

Dataset- IMDB Rating

In [3]:

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

Key takeaways from this dataset

Basic use of Pandas and Exploratory Data Analysis(EDA) which includes cleaning, combining, reshaping, slicing, dicing, and transforming data for analysis purpose.

Plotting graphs using pandas like countplot, factorplot, swarmplot, barplot, violinplot, hexaplot, piechart, kdeplot, distplot, pairplot etc.

Using packages like matplotlib and seaborn to develop better insights about the data.

Create new features which will help in better prediction on hidden aspects of data.

Pandas Profiling to get an overall statistical knowldge of the data like any missing values and irregualities present in the data so as to normalize the data for better analysis.

Get to know coorelation b/w different variables present in the data which might have an impact on overall finding.

Drawing final conclusion on the problem at hand.

Introduction:

Analysis of last 10 yrs. (i.e. from 2006-2016) movies data on IMDB and come up with the success factors of any movie and their correlation.

Data description and loading the dataset:

The dataset contains 1000 observations of movies data hosted on IMDB. IMDB (Internet Movie Database) is an online database of information related to films, television programs, home videos and video games, and internet streams, including cast, production crew and personnel biographies, plot summaries, trivia, and fan reviews and ratings. Users registered on this site are invited to rate any film on a scale of 1 to 10, and the totals are converted into a weighted mean-rating that is displayed beside each title. It also displays the Metascore of each title. Metascore is the rating given by another movie rating company called Metacritic. However, unlike IMDB, they get ratings from registered well known rating agencies and calculates a weighted average of those ratings.

Data Dictionary

Rank - Movie rank order
Title - The title of the film

Genre - A comma-separated list of genres used to classify the film

Description - Brief one-sentence movie summary

Director - The name of the film's director

Actors - A comma-separated list of the main stars of the film

Year - The year that the film released as an integer.

Runtime (Minutes) - The duration of the film in minutes.

Rating - User rating for the movie 0-10

Votes - Number of votes

Revenue (Millions) - Movie revenue in millions

Metascore - An aggregated average of critic scores. Values are between 0 and 100. Higher scores represent positive reviews.

What are the libraries used to read the file?

Ans -pandas libraries are used to read the file .

In [4]:

```
df =pd.read_csv("IMDB-Movie-Data.csv")
```

In [5]:

```
df
```

0	1	of the Galaxy	Action,Adventure,Sci-Fi	intergalactic criminals are forced ...	James Gunn	Diesel, Bradley Cooper, Zoe S...	2014	1
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	1
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	1
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	1
				A secret				

Display top 5 rows of Data set

This will disply top 5 rows of a dataset.

In [7]:

```
df.head(5)
```

Out[7]:

	Rank	Title	Genre	Description	Director	Actors	Yea
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
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
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
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016
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

Display last 5 rows of Data Set

This will return last 5 row of dataset.

In [8]:

```
df.tail(5)
```

Out[8]:

	Rank	Title	Genre	Description	Director	Actors	Year	Run (Minu
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	
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	
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	
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	
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2016	

Shape of Data set(Rows and column)

In [9]:

```
df.shape
```

Out[9]:

```
(1000, 12)
```

In [10]:

```
print("Number of Rows ", df.shape[0])
print("Number of columns",df.shape[1])
```

```
Number of Rows  1000
Number of columns 12
```

Check rows, columns , datatype ,memory usage

Check rows, columns , datatype ,memory usage

In [11]:

```
df.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
```

Check and return true or false if any null value present in set

The isnull() method returns a DataFrame object where all the values are replaced with a Boolean value True for NULL values, and otherwise False.

In [12]:

```
df.isnull().values.any()
```

Out[12]:

```
True
```

In [13]:

```
df.isnull()
```

Out[13]:

	Rank	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Rever (Millio
0	False	False	False	False	False	False	False	False	False	False	Fa
1	False	False	False	False	False	False	False	False	False	False	Fa
2	False	False	False	False	False	False	False	False	False	False	Fa
3	False	False	False	False	False	False	False	False	False	False	Fa
4	False	False	False	False	False	False	False	False	False	False	Fa
...	...	...	...	...	...	...	...	...	...	...	
995	False	False	False	False	False	False	False	False	False	False	T
996	False	False	False	False	False	False	False	False	False	False	Fa
997	False	False	False	False	False	False	False	False	False	False	Fa
998	False	False	False	False	False	False	False	False	False	False	T
999	False	False	False	False	False	False	False	False	False	False	Fa

1000 rows × 12 columns

In [14]:

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

Out[14]:

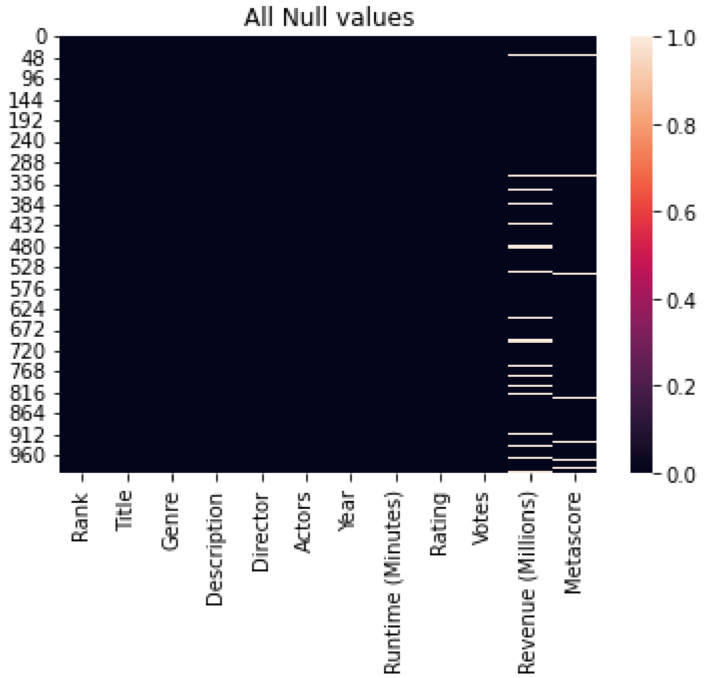
```
Rank                    0
Title                   0
Genre                   0
Description             0
Director                0
Actors                  0
Year                    0
Runtime (Minutes)       0
Rating                  0
Votes                   0
Revenue (Millions)    128
Metascore 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

The function dataframe. isnull(). sum(). sum() returns the number of missing values in the data set.

Visualising mising values by seaborn heatmap

In [15]:

```
sns.heatmap(df.isnull())
plt.title("All Null values")
plt.show()
```



In [16]:

```
miss_null_perct =df.isnull().sum()*100/len(df)
miss_null_perct
```

Out[16]:

```
Rank                   0.0
Title                  0.0
Genre                  0.0
Description            0.0
Director               0.0
Actors                 0.0
Year                   0.0
Runtime (Minutes)      0.0
Rating                 0.0
Votes                  0.0
Revenue (Millions)    12.8
Metascore              6.4
dtype: float64
```

Above code give total missing value in % i,e total missing values in revenue is 12.8%

Drop all Missing values

In [17]:

```
df.dropna(axis=0,)
```

Out[17]:

	Rank	Title	Genre	Description	Director	Actors
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...
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...
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...
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...
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...
...	...	...	...	...	...	...
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...
994	995	Project X	Comedy	3 high school seniors throw a birthday party t...	Nima Nourizadeh	Thomas Mann, Oliver Cooper, Jonathan Daniel Br...
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, 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,...
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...

838 rows × 12 columns

Duplicate data if any ?

In [18]:

```
df_dup =df.duplicated().any()
df_dup
```

Out[18]:

```
False
```

In [19]:

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

Out[19]:

	Rank	Title	Genre	Description	Director	Actors	
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...	2
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...	2
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...	2
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2
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...	2
...	...	...	...	...	...	...	
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...	2
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2
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,...	2
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...	2
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	2

1000 rows × 12 columns

Get overall statistics of Data frame ?

In [20]:

```
df.describe() # statistics for only numerical column
```

Out[20]:

	Rank	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metasco
count	1000.000000	1000.000000	1000.000000	1000.000000	1.000000e+03	872.000000	936.0000
mean	500.500000	2012.783000	113.172000	6.723200	1.698083e+05	82.956376	58.9850
std	288.819436	3.205962	18.810908	0.945429	1.887626e+05	103.253540	17.1947
min	1.000000	2006.000000	66.000000	1.900000	6.100000e+01	0.000000	11.0000
25%	250.750000	2010.000000	100.000000	6.200000	3.630900e+04	13.270000	47.0000
50%	500.500000	2014.000000	111.000000	6.800000	1.107990e+05	47.985000	59.5000
75%	750.250000	2016.000000	123.000000	7.400000	2.399098e+05	113.715000	72.0000
max	1000.000000	2016.000000	191.000000	9.000000	1.791916e+06	936.630000	100.0000

In [21]:

```
df.describe(include = "all") #statistics for both numerical and categorical data
```

Out[21]:

	Rank	Title	Genre	Description	Director	Actors	Year
count	1000.000000	1000	1000	1000	1000	1000	1000.000000
unique	NaN	999	207	1000	644	996	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
freq	NaN	2	50	1	8	2	NaN
mean	500.500000	NaN	NaN	NaN	NaN	NaN	2012.783000
std	288.819436	NaN	NaN	NaN	NaN	NaN	3.205962
min	1.000000	NaN	NaN	NaN	NaN	NaN	2006.000000
25%	250.750000	NaN	NaN	NaN	NaN	NaN	2010.000000
50%	500.500000	NaN	NaN	NaN	NaN	NaN	2014.000000
75%	750.250000	NaN	NaN	NaN	NaN	NaN	2016.000000
max	1000.000000	NaN	NaN	NaN	NaN	NaN	2016.000000

Display columns of a data set

In [20]:

```
df.columns
```

Out[20]:

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

Display titles of movie having movie Runtime > = 180 minutes

In [22]:

```
df[df['Runtime (Minutes)'] >=180]['Title']
```

Out[22]:

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

Display title of movie having rating >= 5

In [23]:

```
df[df['Rating'] >= 5]['Title']
```

Out[23]:

```
0      Guardians of the Galaxy
1                   Prometheus
2                        Split
3                         Sing
4                Suicide Squad
                ...           
995       Secret in Their Eyes
996            Hostel: Part II
997     Step Up 2: The Streets
998               Search Party
999                 Nine Lives
Name: Title, Length: 957, dtype: object
```

In which year there was highest avg voting ?

In [24]:

```
df.groupby('Year')['Votes'].mean().sort_values(ascending = False)
```

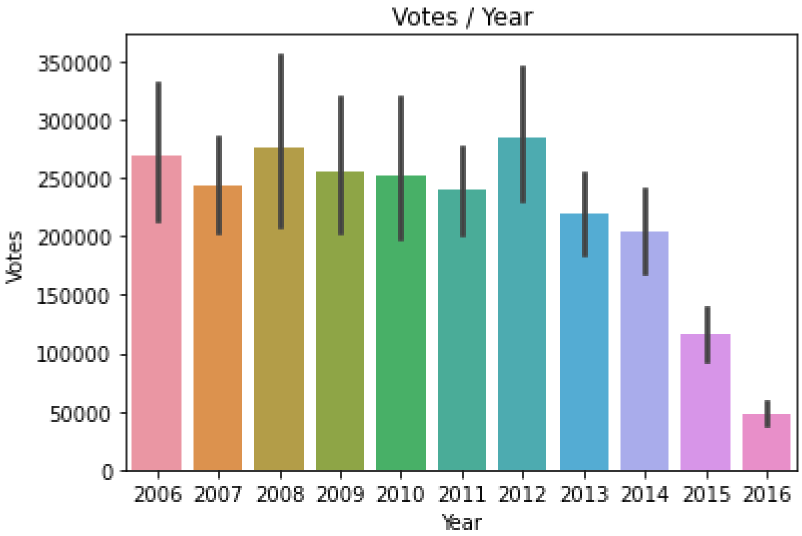
Out[24]:

```
Year
2012    285226.093750
2008    275505.384615
2006    269289.954545
2009    255780.647059
2010    252782.316667
2007    244331.037736
2011    240790.301587
2013    219049.648352
2014    203930.224490
2015    115726.220472
2016     48591.754209
Name: Votes, dtype: float64
```

In which year there was highest avg voting ?

In [25]:

```
sns.barplot(x ='Year', y ='Votes',data =df)
plt.title("Votes / Year")
plt.show()
```



The Average highest voting is calculated by the fraction of numbers of votes by numbers of years. As it is visible that highest Average vote is in year 2012 and the minimum average voting is in year 2016 . Votes are on y axis and year are on x axis.

In which year there was highest Average revenue?

In [26]:

```
df.columns
```

Out[26]:

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

df.column is used to find all the names of the columns in the data.

In [27]:

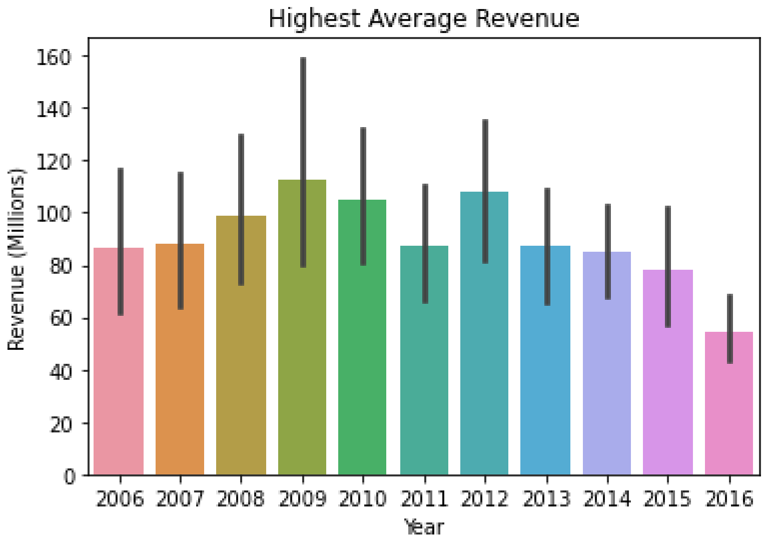
```
df.groupby('Year')['Revenue (Millions)'].mean().sort_values(ascending =False)
```

Out[27]:

```
Year
2009    112.601277
2012    107.973281
2010    105.081579
2008     99.082745
2007     87.882245
2011     87.612258
2013     87.121818
2006     86.296667
2014     85.078723
2015     78.355044
2016     54.690976
Name: Revenue (Millions), dtype: float64
```

In [28]:

```
sns.barplot(x='Year',y='Revenue (Millions)',data =df)
plt.title("Highest Average Revenue")
plt.show()
```



Q Average Rating of each Director ?

In [29]:

```
df.groupby('Director')['Rating'].mean()
```

Out[29]:

```
Director
Aamir Khan             8.50
Abdellatif Kechiche    7.80
Adam Leon              6.50
Adam McKay             7.00
Adam Shankman          6.30
                       ...
Xavier Dolan           7.55
Yimou Zhang            6.10
Yorgos Lanthimos       7.20
Zack Snyder            7.04
Zackary Adler          5.10
Name: Rating, Length: 644, dtype: float64
```

Display top 10 lenghty movie title and runtime ?

In [30]:

```
df.columns
```

Out[30]:

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

In [31]:

```
top10len =df.nlargest(10,'Runtime (Minutes)')[ ['Title', 'Runtime (Minutes)']]\
.set_index('Title')
```

In [32]:

```
top10len
```

Out[32]:

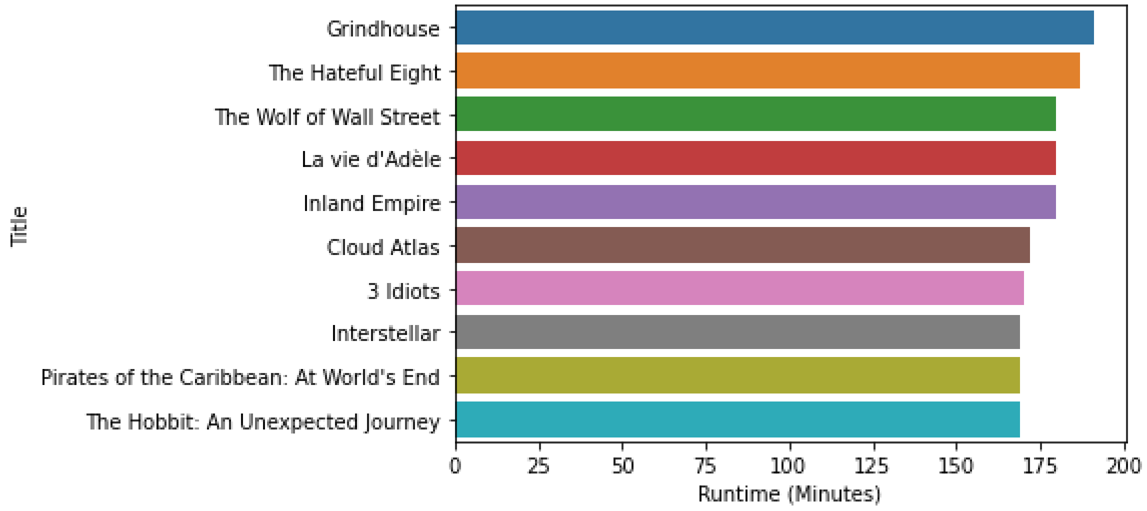
	Runtime (Minutes)
Title	
Grindhouse	191
The Hateful Eight	187
The Wolf of Wall Street	180
La vie d'Adèle	180
Inland Empire	180
Cloud Atlas	172
3 Idiots	170
Interstellar	169
Pirates of the Caribbean: At World's End	169
The Hobbit: An Unexpected Journey	169

In [33]:

```
sns.barplot(x='Runtime (Minutes)', y =top10len.index, data =top10len)
```

Out[33]:

```
<AxesSubplot:xlabel='Runtime (Minutes)', ylabel='Title'>
```



Display number of movies per year

In [34]:

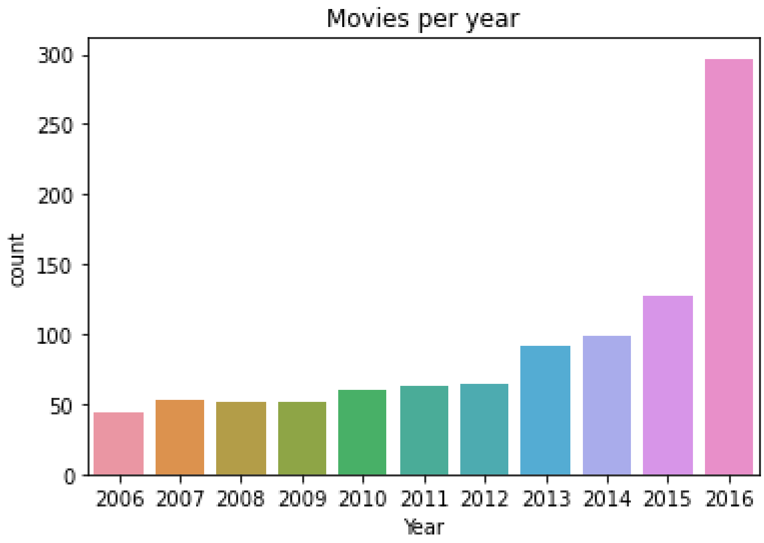
```
df['Year'].value_counts()
```

Out[34]:

```
2016    297
2015    127
2014     98
2013     91
2012     64
2011     63
2010     60
2007     53
2008     52
2009     51
2006     44
Name: Year, dtype: int64
```

In [35]:

```
sns.countplot(x ='Year', data =df)
plt.title("Movies per year")
plt.show()
```



It counts maximum movies per year , we get in year 2016 there are maximum movies in a year but in year there are least movies in year 2016.

Runtime/duration of movies shrink over a period of time

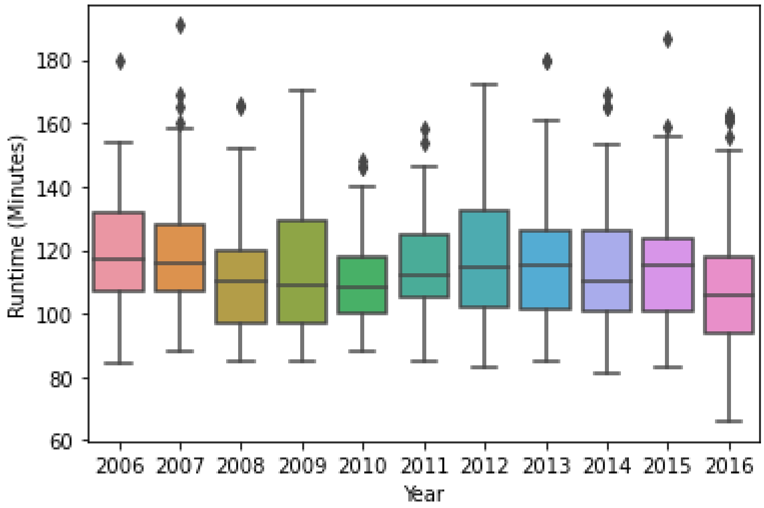
In [37]:

```
sns.boxplot('Year', 'Runtime (Minutes)', data = df)
```

```
C:\Users\Anurag Jaiswal\Anaconda\lib\site-packages\seaborn\_decorators.py:3
6: FutureWarning: Pass the following variables as keyword args: x, y. From v
ersion 0.12, the only valid positional argument will be `data`, and passing
other arguments without an explicit keyword will result in an error or misin
terpretation.
  warnings.warn(
```

Out[37]:

```
<AxesSubplot:xlabel='Year', ylabel='Runtime (Minutes)'>
```



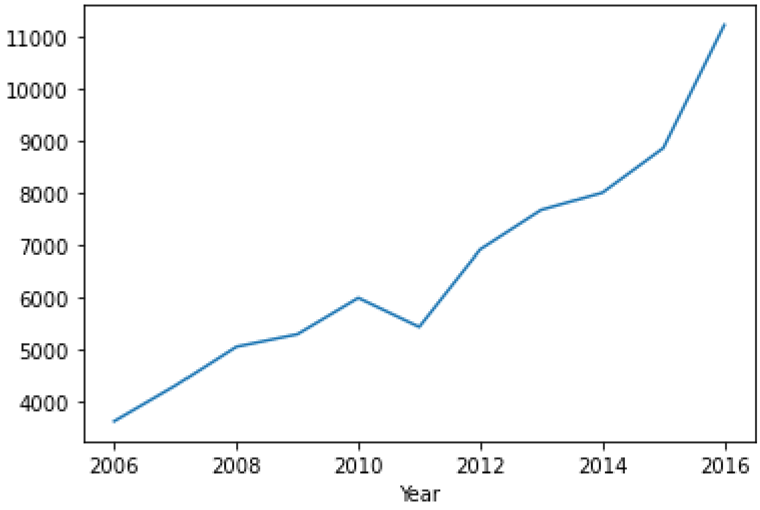
There is an upward Revenue trend due to higher number of movies getting released

In [38]:

```
df.groupby('Year')['Revenue (Millions)'].sum().sort_index().plot.line()
```

Out[38]:

```
<AxesSubplot:xlabel='Year'>
```



Display top 10 high rated movies and their Directors

In [39]:

```
top10rat =df.nlargest(10,'Rating')[['Title','Rating','Director']]\
.set_index("Director")
```

In [40]:

```
top10rat
```

Out[40]:

Director	Title	Rating
Christopher Nolan	The Dark Knight	9.0
Christopher Nolan	Inception	8.8
Nitesh Tiwari	Dangal	8.8
Christopher Nolan	Interstellar	8.6
Makoto Shinkai	Kimi no na wa	8.6
Olivier Nakache	The Intouchables	8.6
Christopher Nolan	The Prestige	8.5
Martin Scorsese	The Departed	8.5
Christopher Nolan	The Dark Knight Rises	8.5
Damien Chazelle	Whiplash	8.5

Display 10 highest revenue Movies

In [41]:

```
df.nlargest(10,'Revenue (Millions)')['Title']
```

Out[41]:

```
50     Star Wars: Episode VII - The Force Awakens
87                                         Avatar
85                                 Jurassic World
76                                   The Avengers
54                                The Dark Knight
12                                      Rogue One
119                                  Finding Dory
94                        Avengers: Age of Ultron
124                         The Dark Knight Rises
578               The Hunger Games: Catching Fire
Name: Title, dtype: object
```

In [42]:

```
top_10=df.nlargest(10,'Revenue (Millions)')[['Title','Revenue (Millions)']].\
set_index('Title')
```

In [43]:

```
top_10
```

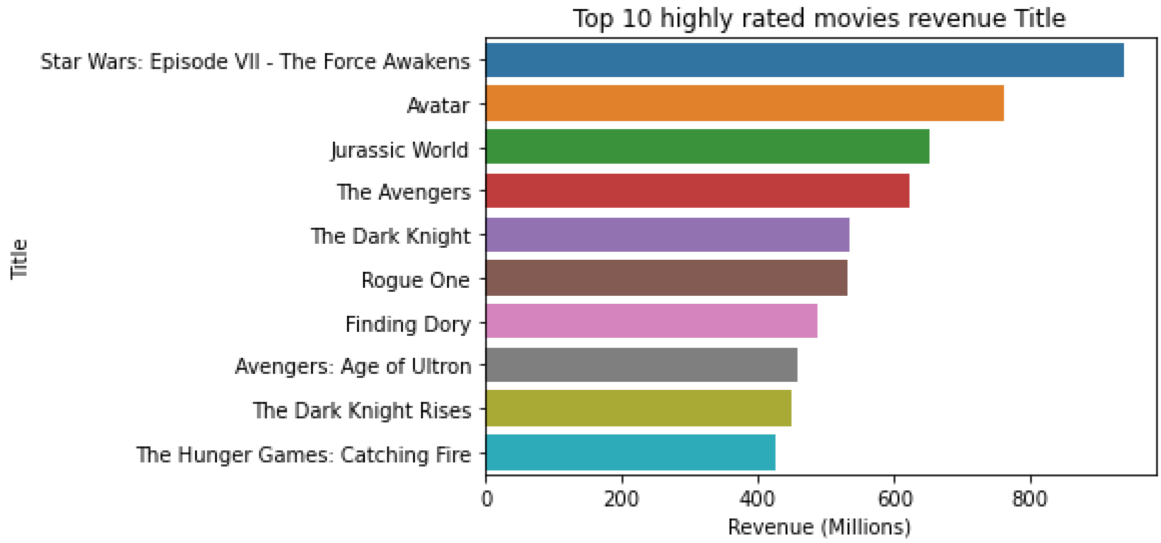
Out[43]:

	Revenue (Millions)
Title	
Star Wars: Episode VII - The Force Awakens	936.63
Avatar	760.51
Jurassic World	652.18
The Avengers	623.28
The Dark Knight	533.32
Rogue One	532.17
Finding Dory	486.29
Avengers: Age of Ultron	458.99
The Dark Knight Rises	448.13
The Hunger Games: Catching Fire	424.65

Top 10 highly rated movies revenue Title

In [44]:

```
sns.barplot(x ='Revenue (Millions)' ,y  =top_10.index ,data = top_10)
plt.title("Top 10 highly rated movies revenue Title")
plt.show()
```



Top highly rated movie is StarWars EpisodeVII - The Force Awakens and it has crosed the budget of 800 million whereas on the other hand the minimum earning by movies is near about 400 million which is Hunger Games: Cathing Fire

How rating Has affected the Revenue of a Movie

In [45]:

```
df.columns
```

Out[45]:

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

How Rating is dependent to Revenue

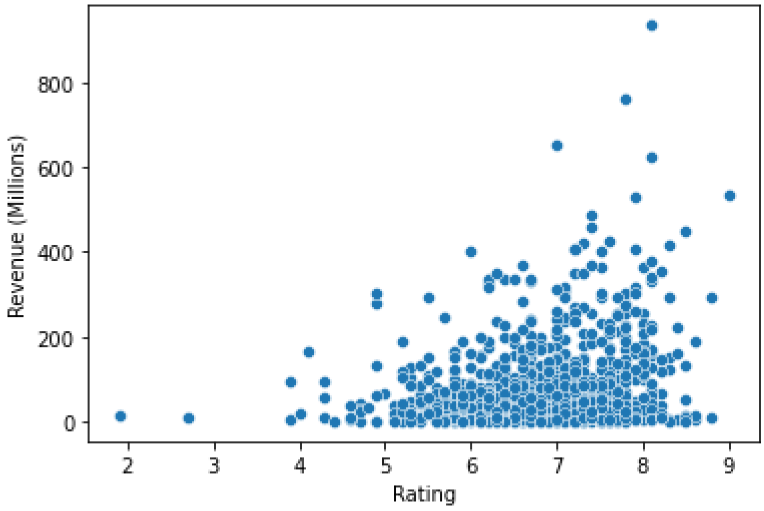
Revenue is directly dependent on Ratings , as the rating will increase then Revenue will grow.
As we can see that Most rating is between the range of 5 - 8 and all the movies are made approx under 400 million

In [46]:

```
sns.scatterplot(x ='Rating' , y ='Revenue (Millions)', data =df)
```

Out[46]:

```
<AxesSubplot:xlabel='Rating', ylabel='Revenue (Millions)'>
```



Highest revenue and year

In [47]:

```
df.columns
```

Out[47]:

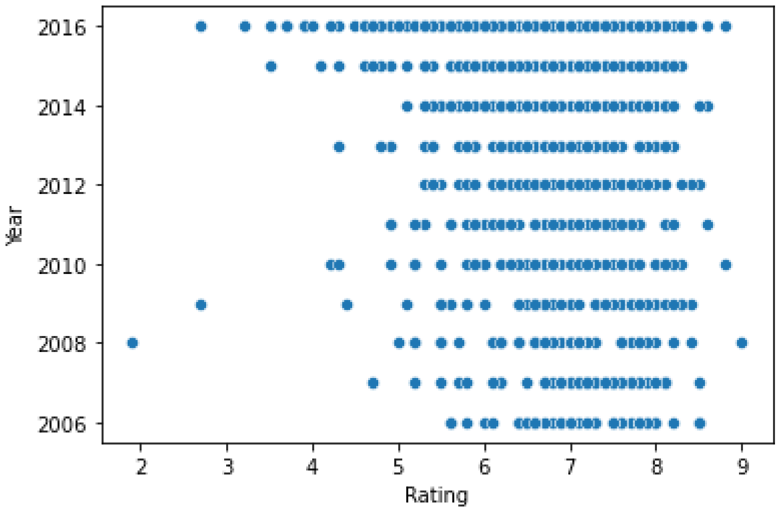
```
Index(['Rank', 'Title', 'Genre', 'Description', 'Director', 'Actors', 'Yea
r',
       'Runtime (Minutes)', 'Rating', 'Votes', 'Revenue (Millions)',
       'Metascore'],
      dtype='object')
```

In [42]:

```
sns.scatterplot(x ='Rating' , y= 'Year', data =df)
```

Out[42]:

```
<AxesSubplot:xlabel='Rating', ylabel='Year'>
```



Classify on ratings on the basis of [Excellent , Good , Average]

In [48]:

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

In this scenario we have classified movies on the basis of their ratings and put them into a categories like Excellent , Good , Average.
If rating is more than 7 then it comes under Excellent category or else if it is more than 6 but less than 7 then it comes under Good Categories
else it automatically comes in the Average categories.

In [49]:

```
df['rating_cat']=df['Rating'].apply(rating)
```

In [50]:

```
df.head()
```

Out[50]:

	Rank	Title	Genre	Description	Director	Actors	Yea
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
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
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
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016
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

Sort the data in ascending order of Runtime of movies

In [51]:

```
df.columns
```

Out[51]:

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

In [52]:

```
df.sort_values(by='Runtime (Minutes)')
```

Out[52]:

	Rank	Title	Genre	Description	Director	Actors	Yea
793	794	Ma vie de Courgette	Animation,Comedy,Drama	After losing his mother, a young boy is sent t...	Claude Barras	Gaspard Schlatter, Sixtine Murat, Paulin Jacco...	201
42	43	Don't Fuck in the Woods	Horror	A group of friends are going on a camping trip...	Shawn Burkett	Brittany Blanton, Ayse Howard, Roman Jossart,N...	201
819	820	Wolves at the Door	Horror,Thriller	Four friends gather at an elegant home during ...	John R. Leonetti	Katie Cassidy, Elizabeth Henstridge, Adam Camp...	201
711	712	La tortue rouge	Animation,Fantasy	A man is shipwrecked on a deserted island and ...	Michael Dudok de Wit	Emmanuel Garijo, Tom Hudson, Baptiste Goy, Axe...	201
949	950	Kicks	Adventure	Brandon is a 15 year old whose dream is a pair...	Justin Tipping	Jahking Guillory, Christopher Jordan Wallace,C...	201
...	...	...	...	...	...	...	..
82	83	The Wolf of Wall Street	Biography,Comedy,Crime	Based on the true story of Jordan Belfort, fro...	Martin Scorsese	Leonardo DiCaprio, Jonah Hill, Margot Robbie,M...	201
965	966	Inland Empire	Drama,Mystery,Thriller	As an actress starts to adopt the persona of h...	David Lynch	Laura Dern, Jeremy Irons, Justin Theroux, Karo...	200
311	312	La vie d'Adèle	Drama,Romance	Adèle's life is changed when she meets Emma, a...	Abdellatif Kechiche	Léa Seydoux, Adèle Exarchopoulos, Salim Kechio...	201
88	89	The Hateful Eight	Crime,Drama,Mystery	In the dead of a Wyoming winter, a bounty hunt...	Quentin Tarantino	Samuel L. Jackson, Kurt Russell, Jennifer Jaso...	201
828	829	Grindhouse	Action,Horror,Thriller	Quentin Tarantino and Robert Rodriguez's homag...	Robert Rodriguez	Kurt Russell, Rose McGowan, Danny Trejo, Zoë Bell	200

1000 rows × 13 columns

In [53]:

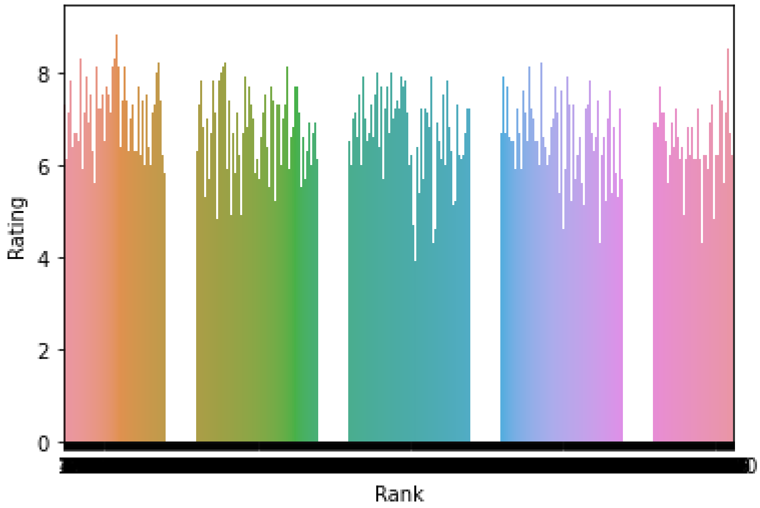
```
df.columns
```

Out[53]:

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

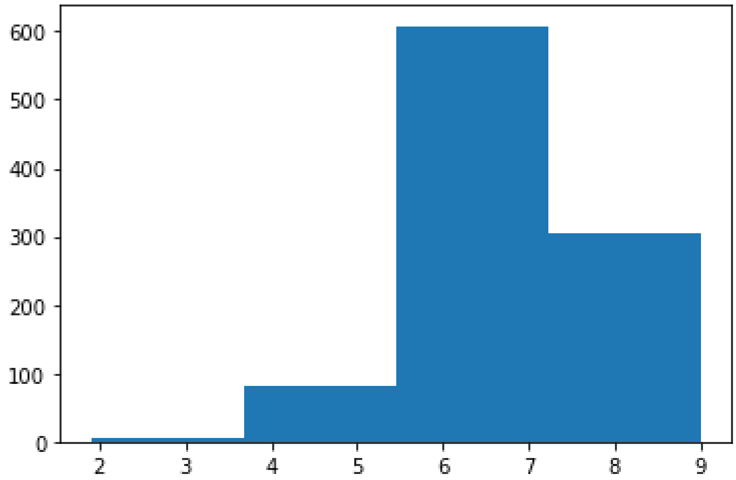
In [54]:

```
sns.barplot(x = 'Rank',y = 'Rating',data = df)
plt.show()
```



In [55]:

```
plt.hist(df['Rating'],bins=4)
plt.show()
```



In [56]:

```
dataset = df.head(20)
```

In [57]:

```
dataset = df.head(20)
```

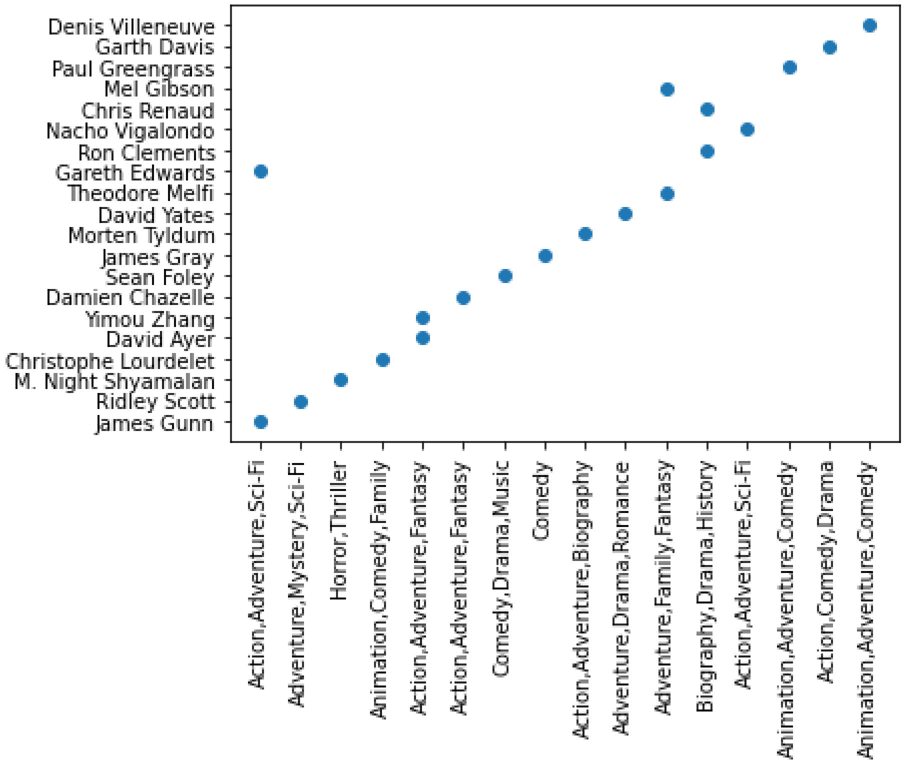
Multivariate Analysis

Multivariate analysis is conceptualized by tradition as the statistical study of experiments in which multiple measurements are made on each experimental unit and for which the relationship among multivariate measurements and their structure are important to the experiment's understanding.

In [58]:

```
import warnings
warnings.filterwarnings("ignore")
import matplotlib.pyplot as plt

fig,ax=plt.subplots()
ax.scatter(dataset.Genre,dataset['Director'])
ax.set_xticklabels(dataset.Genre,rotation=90);
```



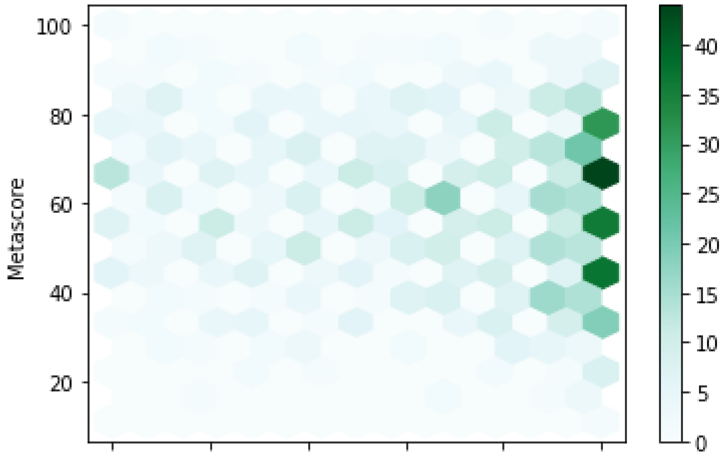
Critics gave better scores to movies released during recent years

In [59]:

```
df.plot.hexbin(x='Year', y='Metascore', gridsize=14)
```

Out[59]:

```
<AxesSubplot:xlabel='Year', ylabel='Metascore'>
```



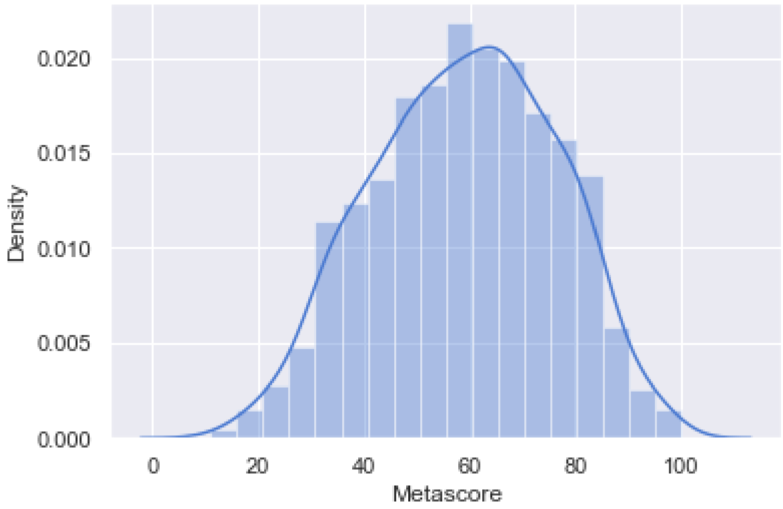
Metascore and Revenue earning is high for movies with 3 genres. Metascore with 55 score and genre 1 has Poor revenues while with genre 2 & 3 have Metascore of 65

In [60]:

```
sns.set(color_codes=True)
sns.set_palette(sns.color_palette('muted'))
sns.distplot(df['Metascore'].dropna())
```

Out[60]:

```
<AxesSubplot:xlabel='Metascore', ylabel='Density'>
```



Plotting the Heatmap on overall parameters on their numerical correlation. Findings:

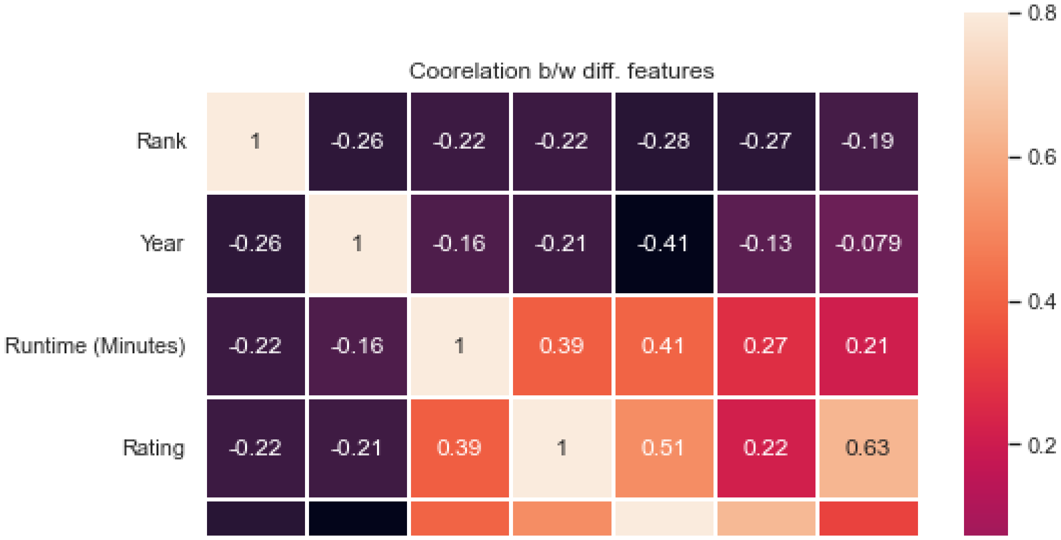
There were more number of movies getting produced in later years almost 5 times than that of initial year, 2006 Though no. of movies produced have increased but their 'Runtime' has reduced over a period of time significantly Movies have higher 'Metascore' in 2016 as compared to previous years As visible, movies having >75% Metascore have almost similar trend as no. of movies being produced Strange to see that average 'Revenue' has come down drastically in last 10 years. Net Revenue has increased due to more movies getting produced Average vaue of Rating is almost constant over the years

In [61]:

```
corr = df.corr()
plt.figure(figsize=(8,8))
sns.heatmap(corr, vmax=.8, linewidth=.01, square=True, annot=True)
plt.title('Coorelation b/w diff. features')
```

Out[61]:

```
Text(0.5, 1.0, 'Coorelation b/w diff. features')
```



There is high correlation b/w Rating & Metascore (critic's rating) Movies rated higher have earned more revenues People have voted movies with high Runtime more Highre Runtime means better rating as well and earned more Revenues as well Votes are directly proportional to movie rating Votes, Rating, Revenue, Runtime, Metascore have direct correlation with each other though in different proportions

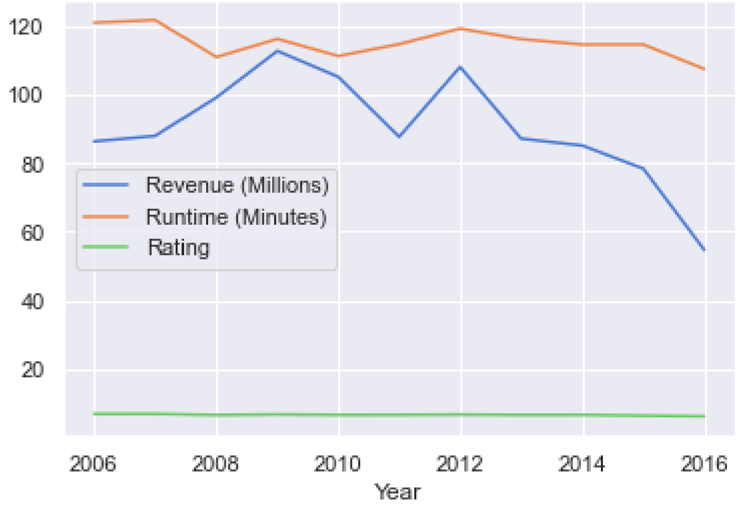
Mean value of Revenue and Runtime saw a dip in recent times

In [62]:

```
df.groupby('Year')['Revenue (Millions)', 'Runtime (Minutes)', 'Rating'].mean().plot.line()
```

Out[62]:

```
<AxesSubplot:xlabel='Year'>
```

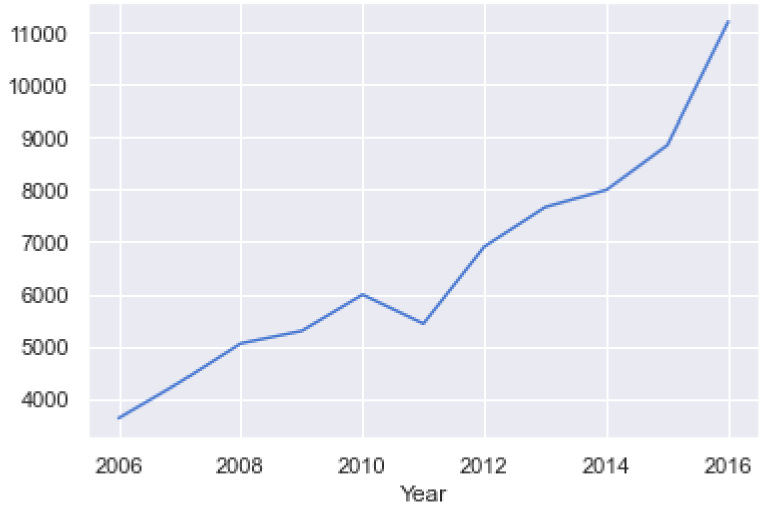


In [63]:

```
df.groupby('Year')['Revenue (Millions)'].sum().sort_index().plot.line()
```

Out[63]:

```
<AxesSubplot:xlabel='Year'>
```

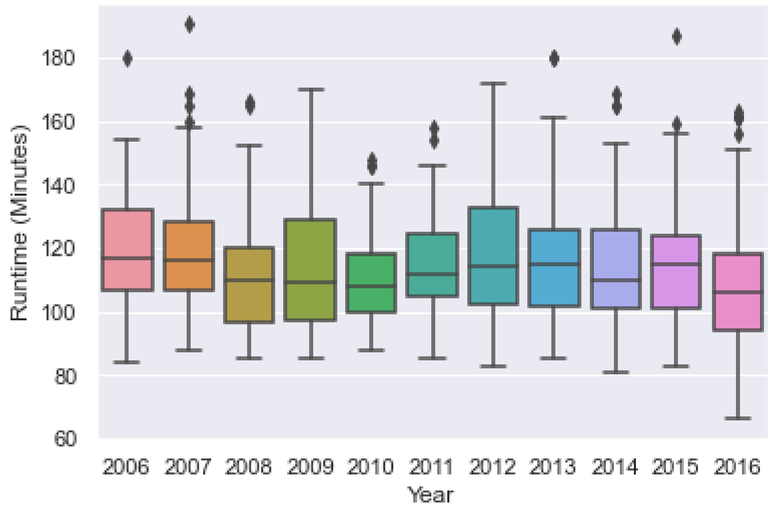


In [64]:

```
sns.boxplot('Year', 'Runtime (Minutes)', data = df)
```

Out[64]:

```
<AxesSubplot:xlabel='Year', ylabel='Runtime (Minutes)'>
```


Runtime (Minutes)
Year