In [2]:

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

In [11]:

```
df=pd.read_csv("C:/Users/hp/Downloads/movies.csv")
df.head(5)
```

Out[11]:

	id	imdb_id	popularity	budget	revenue	original_title	cast	
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.th
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	http://\
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	

5 rows × 21 columns

In [13]:

```
df.columns
```

Out[13]:

```
Index(['id', 'imdb_id', 'popularity', 'budget', 'revenue', 'original_title',
       'cast', 'homepage', 'director', 'tagline', 'keywords', 'overview',
       'runtime', 'genres', 'production_companies', 'release_date',
       'vote_count', 'vote_average', 'release_year', 'budget_adj',
       'revenue_adj'],
      dtype='object')
```

In [16]:

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

Out[16]:

```
id                         0
imdb_id                   10
popularity                 0
budget                     0
revenue                    0
original_title             0
cast                      76
homepage                7930
director                  44
tagline                 2824
keywords                1493
overview                   4
runtime                    0
genres                    23
production_companies    1030
release_date               0
vote_count                 0
vote_average               0
release_year               0
budget_adj                 0
revenue_adj                0
dtype: int64
```

In [27]:

```
#Drop unnecessary columns
df.drop(columns=['id','imdb_id','homepage','tagline','overview','budget_adj'],inplace=
```

In [28]:

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

Out[28]:

```
popularity                0
budget                    0
revenue                   0
original_title            0
cast                     69
director                  0
keywords               1449
runtime                   0
genres                    0
production_companies    994
release_date              0
vote_count                0
vote_average              0
release_year              0
revenue_adj               0
dtype: int64
```

In [29]:

```
#remove any rows in the DataFrame df where at least one of the values in the genres or
df.dropna(how='any',subset=['genres','director'],inplace=True)
```

In [30]:

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

Out[30]:

```
popularity                 0
budget                     0
revenue                    0
original_title             0
cast                      69
director                   0
keywords                1449
runtime                    0
genres                     0
production_companies     994
release_date               0
vote_count                 0
vote_average               0
release_year               0
revenue_adj                0
dtype: int64
```

In [34]:

```
#fill null with 0
df['production_companies']=df['production_companies'].fillna(0)
df['keywords']=df['keywords'].fillna(0)
```

In [35]:

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

Out[35]:

	popularity	budget	revenue	original_title	cast	director	
0	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	monster|dn rex|vel
1	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	fut apocalyptic|dy
2	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	novel|revolution|dystop
3	11.173104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams	android|spaceship|jedi|
4	9.335014	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James Wan	car race|speed|reven

In [38]:

```
#round after decimal value to 2 decimal values
df['popularity']=df['popularity'].round(2)
df.head(5)
```

Out[38]:

	popularity	budget	revenue	original_title	cast	director	
0	32.99	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	monster|dna rex|vel
1	28.42	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	fut apocalyptic|dy
2	13.11	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	novel|revolution|dystop
3	11.17	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	J.J. Abrams	android|spaceship|jedi|
4	9.34	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	James Wan	car race|speed|reven

In [42]:

```python
#inserted the profit column at third place
df.insert(3,'profit',df.revenue-df.budget)
```

In [138…

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

Out[138]:

	popularity	budget	revenue	profit	ROI	original_title	cast	director	
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	mo
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	mo
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	mo
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	mo
1	28.42	150000000	378436354	228436354	1.52	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	apoca

In [48]:

```
#inserted the rate_of_interest column at 4th place
df.insert(4,'ROI',df.profit/df.budget)
```

In [49]:

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

Out[49]:

	popularity	budget	revenue	profit	ROI	rate_of_interest	original_title	
0	32.99	150000000	1513528810	1363528810	9.090192	9.09	Jurassic World	Chris Pratt Howarc Kh
1	28.42	150000000	378436354	228436354	1.522909	1.52	Mad Max: Fury Road	Hardy|Cl Theron I Byrn
2	13.11	110000000	295238201	185238201	1.683984	1.68	Insurgent	Sh Woodley Jame Winslet|A
3	11.17	200000000	2068178225	1868178225	9.340891	9.34	Star Wars: The Force Awakens	Ha Forc Hamill Fisher|Ada
4	9.34	190000000	1506249360	1316249360	6.927628	6.93	Furious 7	Vin Dies Walker Statham|M

In [50]:

```
df['ROI']=df['ROI'].round(2)
```

In [73]:

```
df.head(3)
```

Out[73]:

	popularity	budget	revenue	profit	ROI	original_title	cast	director	
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	
1	28.42	150000000	378436354	228436354	1.52	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	
2	13.11	110000000	295238201	185238201	1.68	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	nove

In [67]:

```
df1=df[['popularity','budget','revenue','profit','ROI','vote_count','vote_average','re
df1
```

Out[67]:

	popularity	budget	revenue	profit	ROI	vote_count	vote_average	release_year
0	32.99	150000000	1513528810	1363528810	9.09	5562	6.5	2015
1	28.42	150000000	378436354	228436354	1.52	6185	7.1	2015
2	13.11	110000000	295238201	185238201	1.68	2480	6.3	2015
3	11.17	200000000	2068178225	1868178225	9.34	5292	7.5	2015
4	9.34	190000000	1506249360	1316249360	6.93	2947	7.3	2015
...	...	...	...	...	...	...	...	...
10861	0.08	0	0	0	NaN	11	7.4	1966
10862	0.07	0	0	0	NaN	20	5.7	1966
10863	0.07	0	0	0	NaN	11	6.5	1966
10864	0.06	0	0	0	NaN	22	5.4	1966
10865	0.04	19000	0	-19000	-1.00	15	1.5	1966

10801 rows × 8 columns

In [68]:

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

Out[68]:

```
popularity                0
budget                    0
revenue                   0
profit                    0
ROI                    5636
original_title            0
cast                     69
director                  0
keywords                  0
runtime                   0
genres                    0
production_companies      0
release_date              0
vote_count                0
vote_average              0
release_year              0
revenue_adj               0
dtype: int64
```

In [61]:

```
#inf means infinity values are there
df.ROI.value_counts()
```

Out[61]:

```
-1.00     1350
 inf       995
-0.99       29
-0.98       27
-0.38       21
          ... 
 4.15        1
 24.90       1
 2.32        1
 6.24        1
 6.62        1
Name: ROI, Length: 1075, dtype: int64
```

In [65]:

```
#the total number of non-finite values (infinite or NaN) in the ROI column
non_finite_values=~np.isfinite(df['ROI'])
```

In [64]:

```
non_finite_values.sum()
```

Out[64]:

```
5636
```

In [66]:

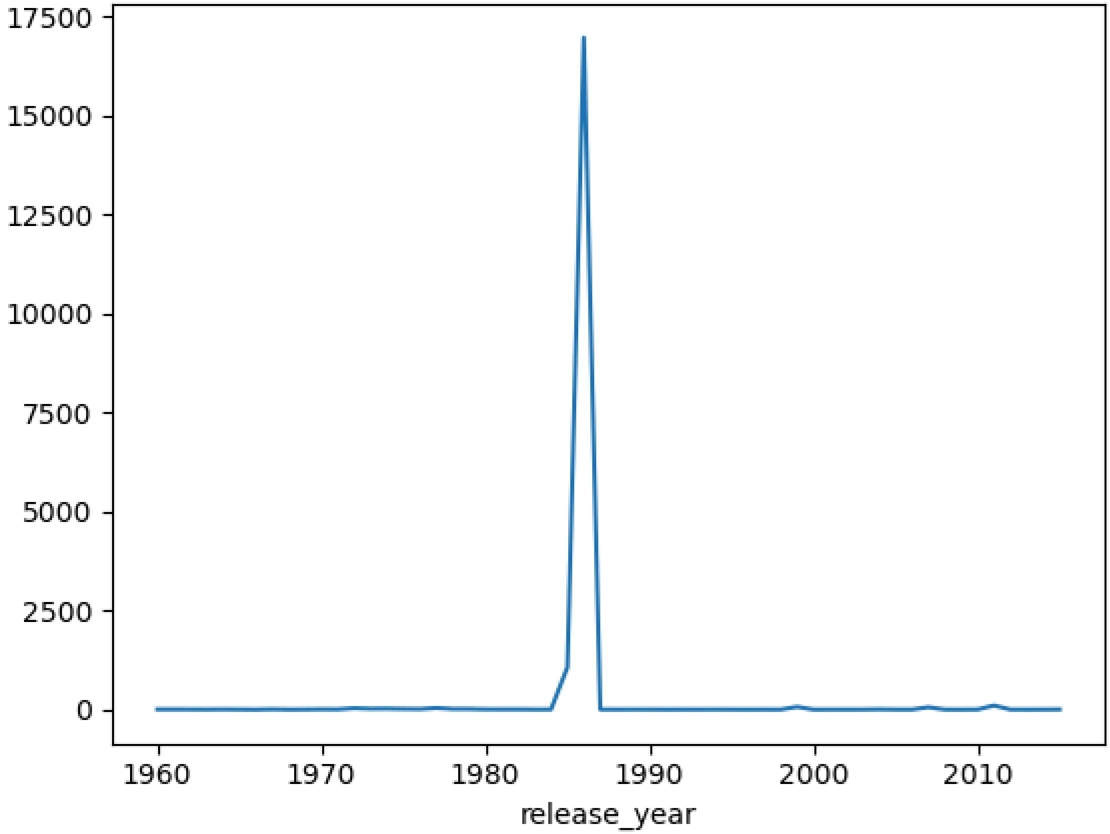
```
df['ROI']=df['ROI'].replace([np.inf,-np.inf],np.nan)
```

In [75]:

```
df2=df.groupby('release_year')['ROI'].mean()
df2.plot(kind='line')
```

Out[75]:

```
<Axes: xlabel='release_year'>
```



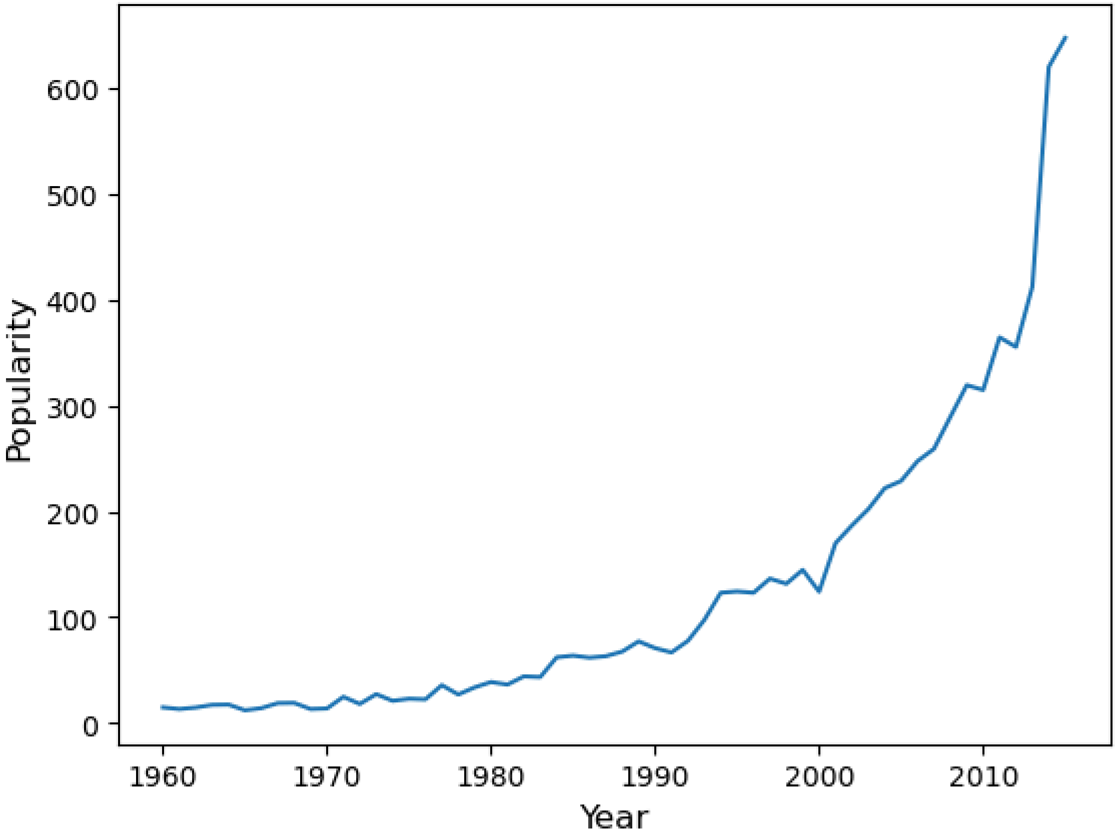
In [80]:

```
#Popularity over the years
df3=df.groupby('release_year')['popularity'].sum()
df3.plot(kind='line')
```

```
plt.xlabel('Year',fontsize=12)
plt.ylabel('Popularity',fontsize=12)
```

Out[80]:

```
Text(0, 0.5, 'Popularity')
```

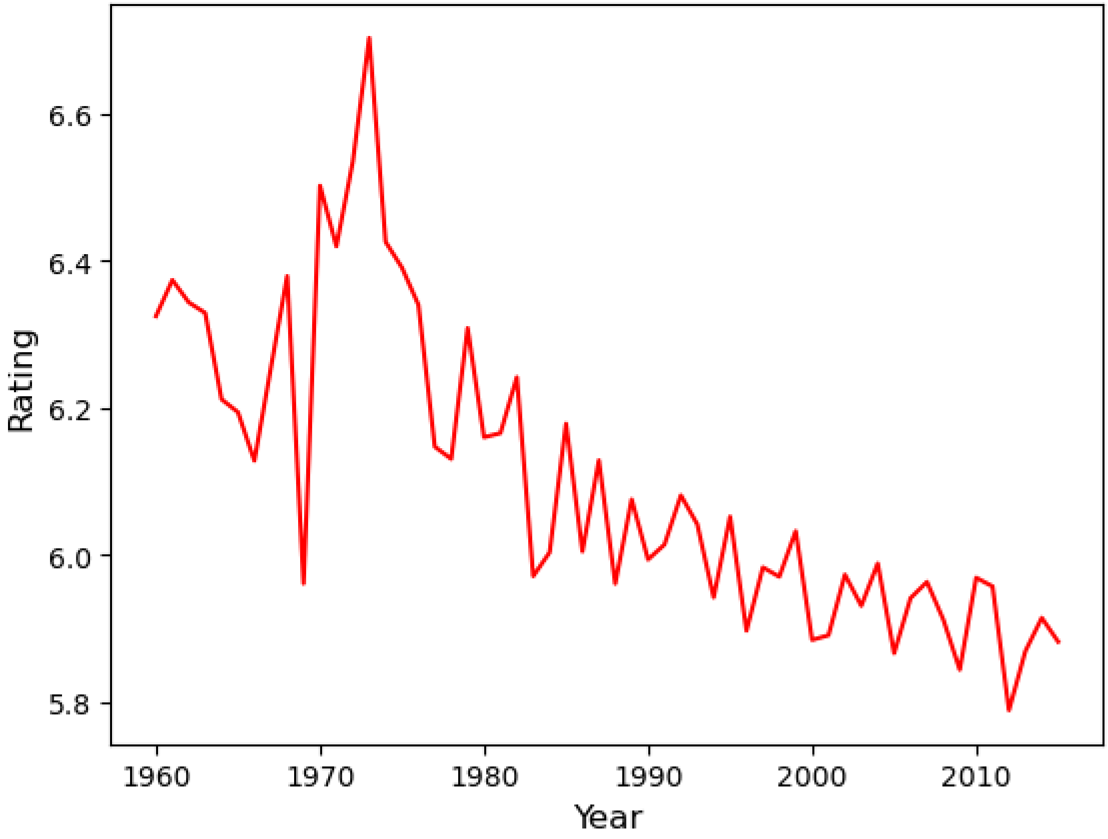


In [82]:

```
#Vote average over the years
df3=df.groupby('release_year')['vote_average'].mean()
df3.plot(kind='line',color='red')
plt.xlabel('Year',fontsize=12)
plt.ylabel('Rating',fontsize=12)
```

Out[82]:

```
Text(0, 0.5, 'Rating')
```



In [83]:
```
df.head(3)
```

Out[83]:

	popularity	budget	revenue	profit	ROI	original_title	cast	director	
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	
1	28.42	150000000	378436354	228436354	1.52	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	
2	13.11	110000000	295238201	185238201	1.68	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	nove

In [85]:
```
df.genres.value_counts()
```

Out[85]:

```
Drama                                      711
Comedy                                     707
Documentary                                306
Drama|Romance                              289
Comedy|Drama                               280
                                          ... 
Science Fiction|Horror|Action|Thriller       1
Action|Thriller|Science Fiction|Mystery      1
Comedy|Music|Romance|Foreign                 1
Documentary|Drama|Comedy                     1
Mystery|Science Fiction|Thriller|Drama       1
Name: genres, Length: 2031, dtype: int64
```

In [86]:

```
#to split the genres (turns the string into a list of substrings),
#'Action|Adventure|Sci-Fi', it will be transformed into the list ['Action', 'Adventure

split=['genres']
for i in split:
    df[i]=df[i].apply(lambda x:x.split('|'))
df.head(3)
```

Out[86]:

	popularity	budget	revenue	profit	ROI	original_title	cast	director	
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	
1	28.42	150000000	378436354	228436354	1.52	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	
2	13.11	110000000	295238201	185238201	1.68	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	Robert Schwentke	nove

In [184…

```
#Using df.explode('genres') is a great way to handle a DataFrame where the genres colu
#Exploding this column will transform each genre in the lists into its own row
df=df.explode('genres')
df
```

Out[184]:

	id	imdb_id	popularity	budget	revenue	original_title	cast	
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.th
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	http://\
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	

5 rows × 21 columns

In [94]:

```
df7=df.groupby('genres')['popularity'].sum().sort_values(ascending=False)
df7
```

Out[94]:

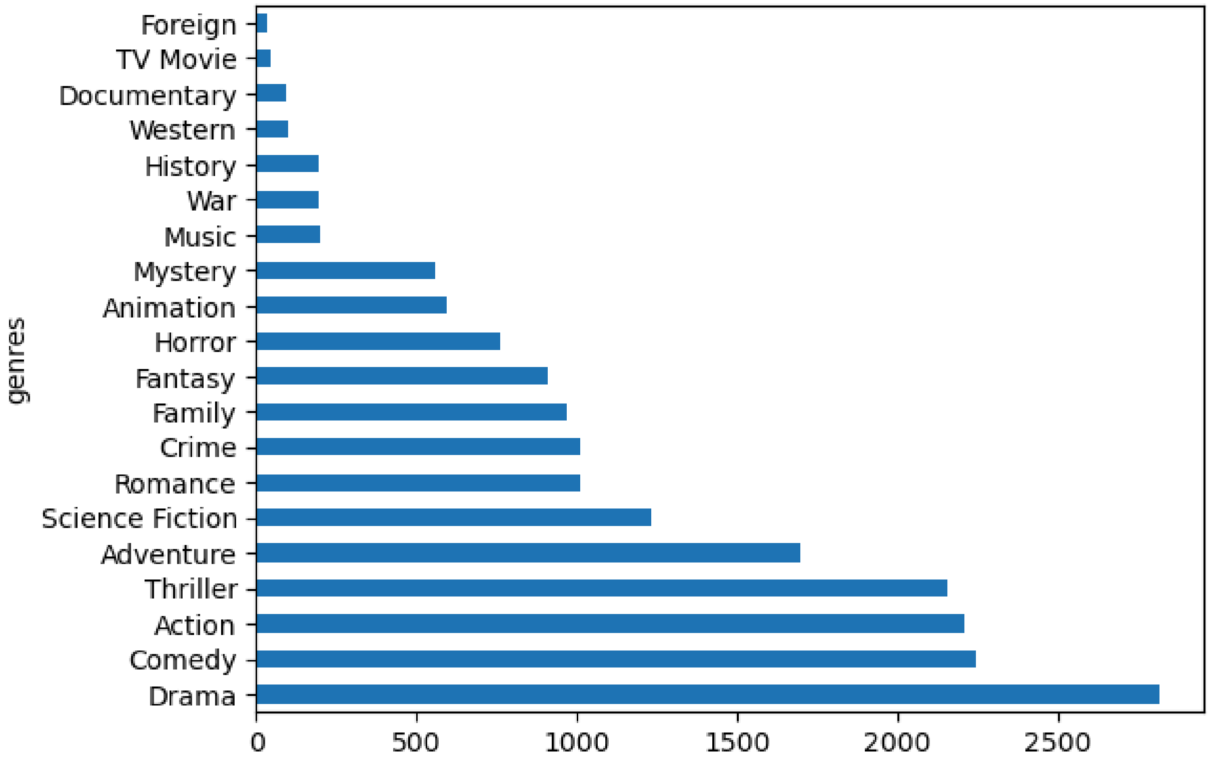
```
genres
Drama              2815.43
Comedy             2246.25
Action             2208.08
Thriller           2155.90
Adventure          1697.11
Science Fiction    1230.41
Romance            1013.21
Crime              1009.07
Family              967.06
Fantasy             908.87
Horror              761.39
Animation           594.46
Mystery             558.55
Music               198.15
War                 196.48
History             192.35
Western              97.42
Documentary          93.13
TV Movie             44.03
Foreign              35.24
Name: popularity, dtype: float64
```

In [99]:

```
#Total movies according to genres
df7.plot.barh(x='genres',y='popularity') #barh is for horizontal visual
```

Out[99]:

```
<Axes: ylabel='genres'>
```



In [103…

```
df.dtypes
```

Out[103]:

```
popularity              float64
budget                    int64
revenue                   int64
profit                    int64
ROI                     float64
original_title           object
cast                     object
director                 object
keywords                 object
runtime                   int64
genres                   object
production_companies     object
release_date             object
vote_count                int64
vote_average            float64
release_year              int64
revenue_adj             float64
dtype: object
```

In [104…

```
#change release_date column datatype to datetime
df['release_date']=pd.to_datetime(df['release_date'])
```

In [105…

```
df.dtypes
```

Out[105]:

```
popularity                     float64
budget                           int64
revenue                          int64
profit                           int64
ROI                            float64
original_title                  object
cast                            object
director                        object
keywords                        object
runtime                          int64
genres                          object
production_companies            object
release_date            datetime64[ns]
vote_count                       int64
vote_average                   float64
release_year                     int64
revenue_adj                    float64
dtype: object
```

In [106…

```
df.head(3)
```

Out[106]:

	popularity	budget	revenue	profit	ROI	original_title	cast	director	
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	monst r
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	monst r
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	monst r

In [108…

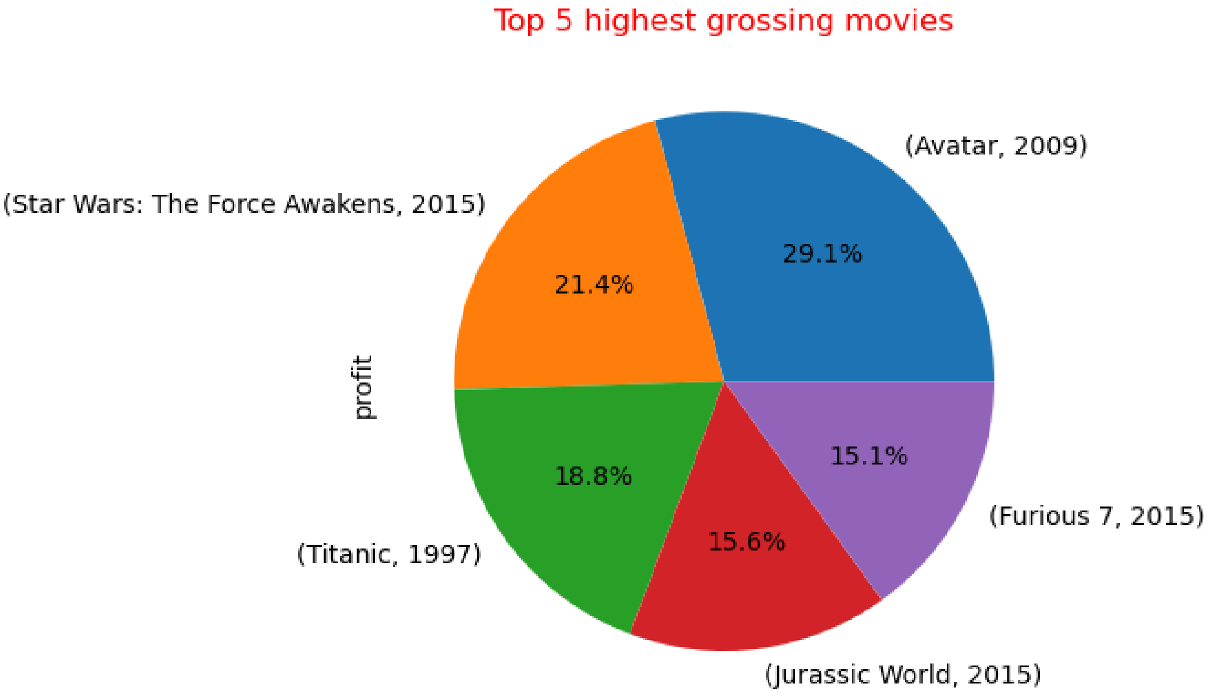
```
df['months']=df['release_date'].dt.month
```

In [122…

```
#Top 5 highest grossing movies
df6=df.groupby(['original_title','release_year'])['profit'].max().sort_values(ascendir
```

In [183…

```
df6.plot(kind='pie',autopct='%1.1f%%')
plt.title('Top 5 highest grossing movies',color='red')
plt.show()
```



In [153…

```
df=pd.read_csv("C:/Users/hp/Downloads/movies.csv")
df.head(3)
```

Out[153]:

	id	imdb_id	popularity	budget	revenue	original_title	cast	
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.thed

3 rows × 21 columns

In [163…

```
#Top 5 production companies
df8=df.production_companies.value_counts().head(5)
df8
```

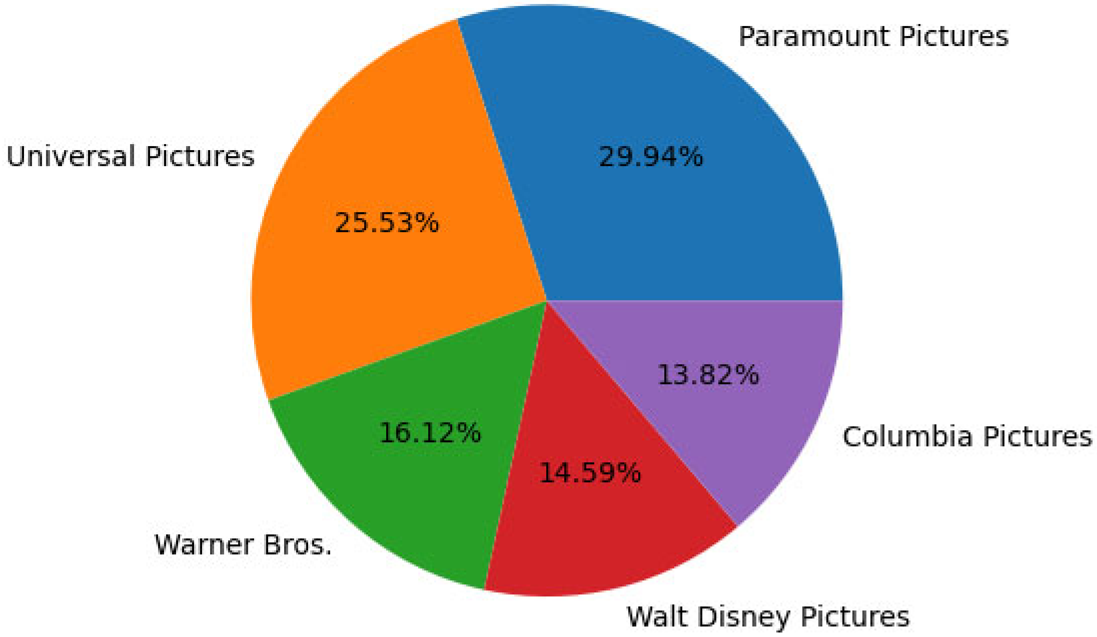
Out[163]:

```
Paramount Pictures      156
Universal Pictures      133
Warner Bros.             84
Walt Disney Pictures     76
Columbia Pictures        72
Name: production_companies, dtype: int64
```

In [167…

```
df8 = df.production_companies.value_counts().head(5)
df8.plot(kind='pie', autopct='%1.2f%%')
plt.ylabel('')  # This removes the y-label

plt.show()
```



In [14]:
```
#Top 5 movies with high budget
df[['original_title','budget']].sort_values(by='budget',ascending=False).head(5)
```

Out[14]:

	original_title	budget
2244	The Warrior's Way	425000000
3375	Pirates of the Caribbean: On Stranger Tides	380000000
7387	Pirates of the Caribbean: At World's End	300000000
14	Avengers: Age of Ultron	280000000
6570	Superman Returns	270000000

Top 5 voted movies

In [4]:
```
#Top 5 voted movies
top_5_movies = df[['original_title', 'vote_count']].sort_values(by='vote_count', ascer

# Reset the index and drop the old index
top_5_movies = top_5_movies.reset_index(drop=True)
top_5_movies
```

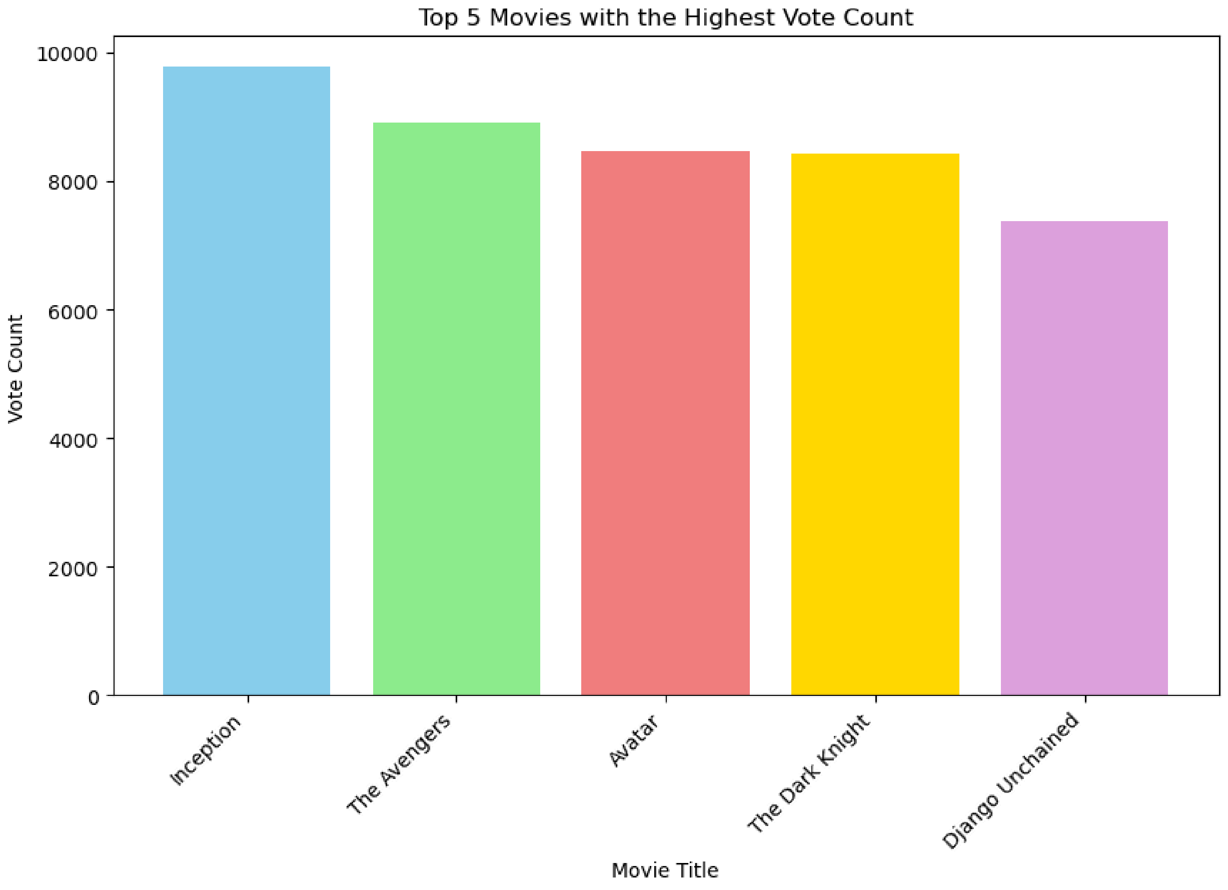
Out[4]:

	original_title	vote_count
0	Inception	9767
1	The Avengers	8903
2	Avatar	8458
3	The Dark Knight	8432
4	Django Unchained	7375

In [8]:

```python
#a list of colors for each bar
colors = ['skyblue', 'lightgreen', 'lightcoral', 'gold', 'plum']

# Plotting
plt.figure(figsize=(10,6))
plt.bar(top_5_movies['original_title'], top_5_movies['vote_count'], color=colors)
plt.xlabel('Movie Title')
plt.ylabel('Vote Count')
plt.title('Top 5 Movies with the Highest Vote Count')
plt.xticks(rotation=45, ha='right')
plt.show()
```


Top 5 Movies with the Highest Vote Count
Vote Count
Inception
The Avengers
Avatar
The Dark Knight
Django Unchained
Movie Title