In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import plotly.express as px
```

In [9]:
```
df = pd.read_csv('C:/Users/ritvi/Desktop/data/movies.csv')
df.head(3)
```

Out[9]:

	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 [10]:
```
df.isnull().sum()
```

Out[10]:
```
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 [ ]:

In [11]:
```
df.drop(columns = ["id","imdb_id","homepage","cast","tagline","overview","budget_adj",
```

In [12]:
```
df.head()
```

Out[12]:

	popularity	budget	revenue	original_title	director	keywords	ru
0	32.985763	150000000	1513528810	Jurassic World	Colin Trevorrow	monster|dna|tyrannosaurus rex|velociraptor|island	
1	28.419936	150000000	378436354	Mad Max: Fury Road	George Miller	future|chase|post-apocalyptic|dystopia|australia	
2	13.112507	110000000	295238201	Insurgent	Robert Schwentke	based on novel|revolution|dystopia|sequel|dyst...	
3	11.173104	200000000	2068178225	Star Wars: The Force Awakens	J.J. Abrams	android|spaceship|jedi|space opera|3d	
4	9.335014	190000000	1506249360	Furious 7	James Wan	car race|speed|revenge|suspense|car	

In [18]:
```
df.isnull().sum()
```

Out[18]:
```
popularity              0
budget                  0
revenue                 0
original_title          0
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 [15]:
```
df.dropna(how = 'any',subset = ["genres","director"], inplace = True)
```

In [ ]:

In [17]:
```
df["production_companies"] = df["production_companies"].fillna(0)
df["keywords"] = df["keywords"].fillna(0)
```

In [ ]:

In [27]:
```
df["popularity"] = df["popularity"].round(2)
df["roi"] = df["roi"].round(2)
```

In [28]:
```
df
```

Out[28]:

	popularity	budget	revenue	profit	roi	original_title	director	
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	
1	28.42	150000000	378436354	228436354	1.52	Mad Max: Fury Road	George Miller	ap
2	13.11	110000000	295238201	185238201	1.68	Insurgent	Robert Schwentke	novel|revolu
3	11.17	200000000	2068178225	1868178225	9.34	Star Wars: The Force Awakens	J.J. Abrams	android|spa
4	9.34	190000000	1506249360	1316249360	6.93	Furious 7	James Wan	car race|s
...	...	...	...	...	...	...	...	
10861	0.08	0	0	0	NaN	The Endless Summer	Bruce Brown	
10862	0.07	0	0	0	NaN	Grand Prix	John Frankenheimer	
10863	0.07	0	0	0	NaN	Beregis Avtomobilya	Eldar Ryazanov	
10864	0.06	0	0	0	NaN	What's Up, Tiger Lily?	Woody Allen	
10865	0.04	19000	0	-19000	-1.00	Manos: The Hands of Fate	Harold P. Warren	fire|g

10801 rows × 16 columns

In [ ]:
```
```

In [23]:
```
df.insert(3, 'profit', df.revenue - df.budget)
```

In [25]:
```
df.insert(4, 'roi', df.profit / df.budget)
```

In [ ]:
```
df
```

In [38]:
```
df1 = df[['popularity','budget','revenue','profit','roi','vote_count','vote_average','
```

In [39]:
```
df.isnull().sum()
```

Out[39]:
```
popularity                 0
budget                     0
revenue                    0
profit                     0
roi                     5636
original_title             0
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 [40]:
```
df.roi.value_counts()
```

Out[40]:
```
-1.00     1350
-0.99       29
-0.98       27
-0.38       21
 0.20       19
           ...
 4.15        1
 24.90       1
 2.32        1
 6.24        1
 6.62        1
Name: roi, Length: 1074, dtype: int64
```

In [41]:
```
non_finite_values = ~np.isfinite(df['roi'])
```

In [42]:
```
non_finite_values.sum()
```

Out[42]:
```
5636
```

In [43]:
```
df['roi'] = df['roi'].replace([np.inf, -np.inf], np.nan)
```

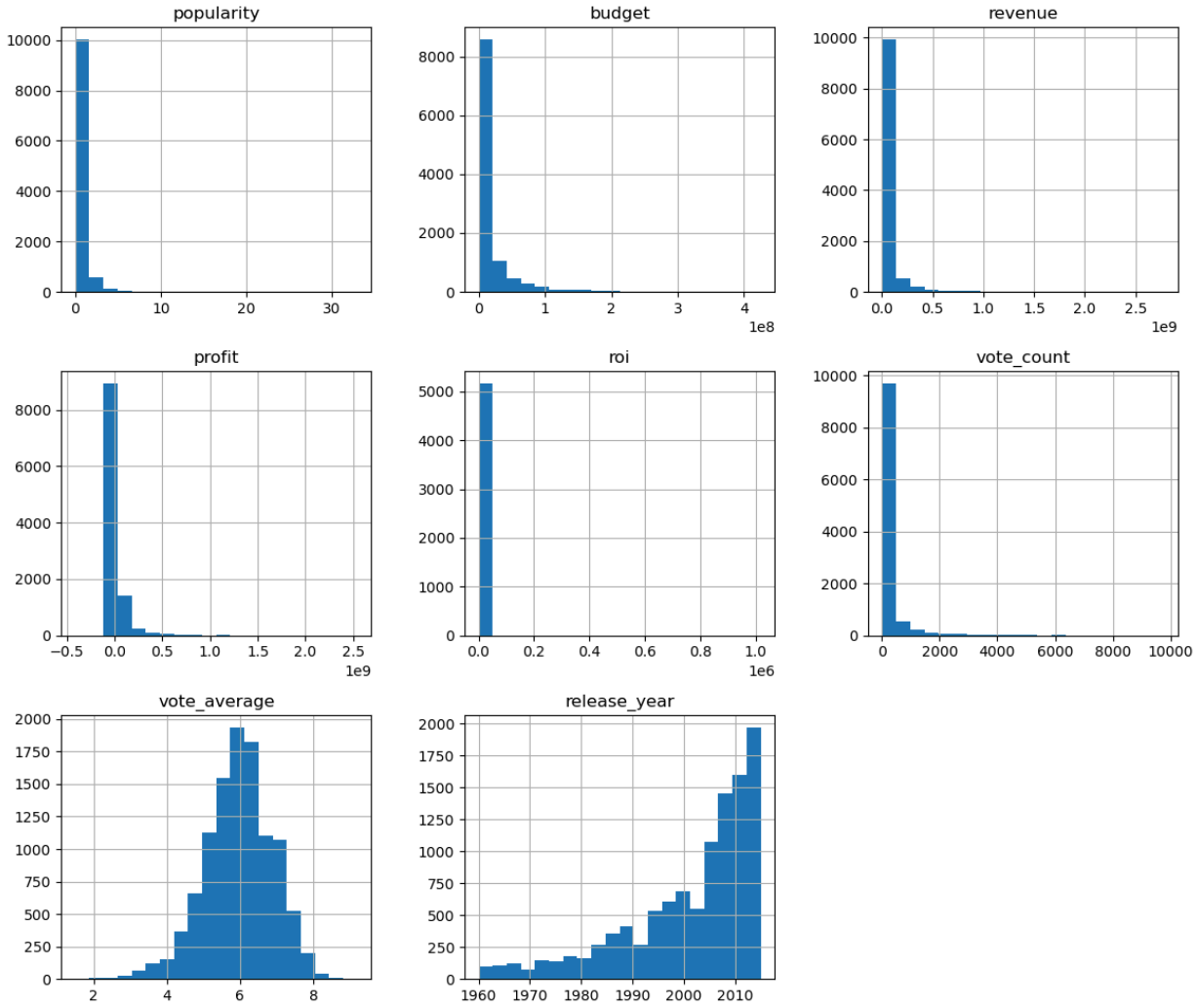
In [ ]:
```

```

In [ ]:
```

```

In [45]:
```
df1.hist(bins = 20,figsize = (14,12))
plt.show()
```



In [46]:
```
df.popularity.value_counts()
```

Out[46]:
```
0.14    193
0.28    190
0.21    186
0.25    182
0.20    179
       ...
5.81      1
5.08      1
3.83      1
3.74      1
2.68      1
Name: popularity, Length: 483, dtype: int64
```

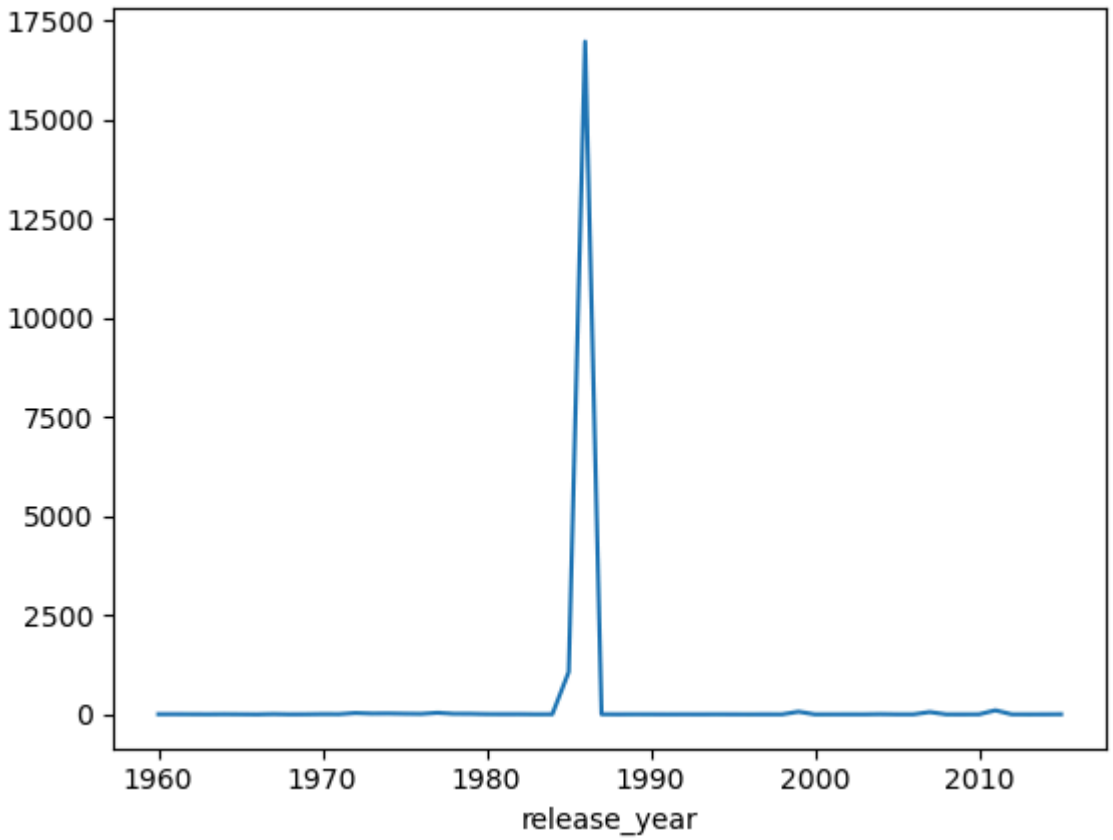
In [52]:
```
df.head(2)
```

Out[52]:

	popularity	budget	revenue	profit	roi	original_title	director	key
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyrannc rex|velocirapto
1	28.42	150000000	378436354	228436354	1.52	Mad Max: Fury Road	George Miller	future|chas apocalyptic|dystopia|a

In [54]:
```
df2 = df.groupby('release_year')['roi'].mean()
df2.plot(kind = 'line')
```

Out[54]:
```
<AxesSubplot:xlabel='release_year'>
```



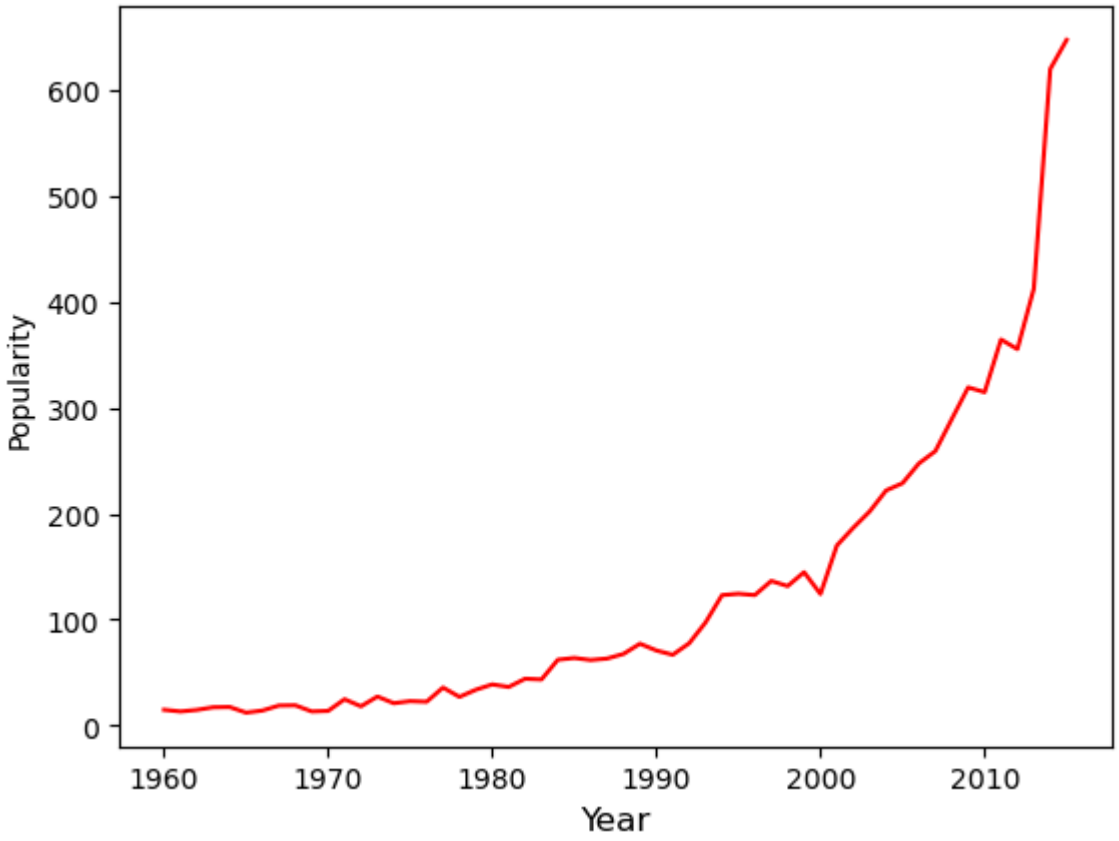
In [ ]:

In [ ]:

In [ ]:

In [55]:
```
df3 = df.groupby('release_year')['popularity'].sum()
df3.plot(kind = 'line',color = 'red')
plt.xlabel('Year',fontsize = 12)
plt.ylabel('Popularity')
```

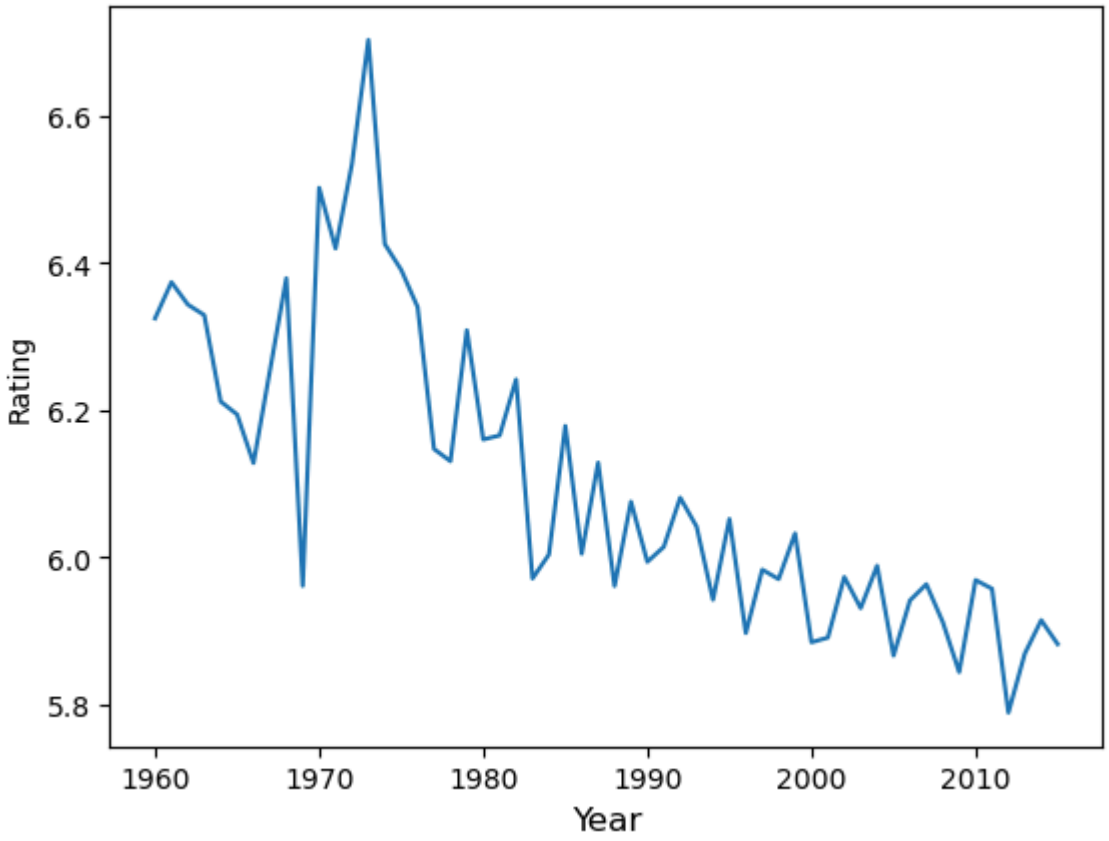
Out[55]:
```
Text(0, 0.5, 'Popularity')
```



In [56]:
```
#Rating
```

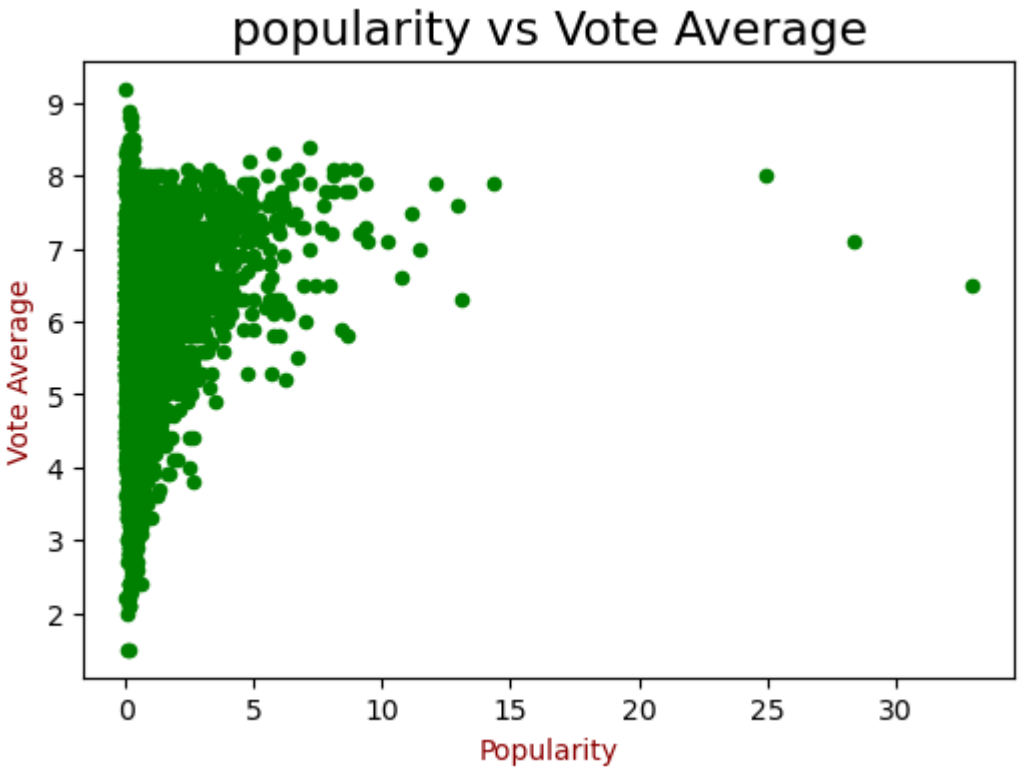
In [57]:
```
df4 = df.groupby('release_year')['vote_average'].mean()
df4.plot(kind = 'line')
plt.xlabel('Year',fontsize = 12)
plt.ylabel('Rating')
```

Out[57]:
```
Text(0, 0.5, 'Rating')
```



In [58]:
```
df5 = df.plot.scatter(x = 'popularity', y =  'vote_average', c = 'green',figsize=(6,4)
df5.set_xlabel('Popularity',color = 'DarkRed')
df5.set_ylabel('Vote Average', color = 'DarkRed')
df5.set_title('popularity vs Vote Average', fontsize = 17)
```

Out[58]:
```
Text(0.5, 1.0, 'popularity vs Vote Average')
```



In [ ]:

In [60]:
```
df.genres.value_counts()
```

Out[60]:
```
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 [61]:
```
df
```

Out[61]:

	popularity	budget	revenue	profit	roi	original_title	director	
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	
1	28.42	150000000	378436354	228436354	1.52	Mad Max: Fury Road	George Miller	ap
2	13.11	110000000	295238201	185238201	1.68	Insurgent	Robert Schwentke	novel|revolu
3	11.17	200000000	2068178225	1868178225	9.34	Star Wars: The Force Awakens	J.J. Abrams	android|spa
4	9.34	190000000	1506249360	1316249360	6.93	Furious 7	James Wan	car race|s
...	...	...	...	...	...	...	...	
10861	0.08	0	0	0	NaN	The Endless Summer	Bruce Brown	
10862	0.07	0	0	0	NaN	Grand Prix	John Frankenheimer	
10863	0.07	0	0	0	NaN	Beregis Avtomobilya	Eldar Ryazanov	
10864	0.06	0	0	0	NaN	What's Up, Tiger Lily?	Woody Allen	
10865	0.04	19000	0	-19000	-1.00	Manos: The Hands of Fate	Harold P. Warren	fire|g

10801 rows × 16 columns

In [62]:
```
split = ['genres']
for i in split:
    df[i] = df[i].apply(lambda x: x.split("|"))
df.head(3)
```

Out[62]:

	popularity	budget	revenue	profit	roi	original_title	director	
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|d rex|v
1	28.42	150000000	378436354	228436354	1.52	Mad Max: Fury Road	George Miller	f apocalyptic|
2	13.11	110000000	295238201	185238201	1.68	Insurgent	Robert Schwentke	novel|revolution|dysto

In [ ]:
```
```

In [63]:
```
df = df.explode('genres')
df
```

Out[63]:

	popularity	budget	revenue	profit	roi	original_title	director	
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna rex|vel
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna rex|vel
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna rex|vel
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna rex|vel
1	28.42	150000000	378436354	228436354	1.52	Mad Max: Fury Road	George Miller	fut apocalyptic|dy
...	...	...	...	...	...	...	...	
10863	0.07	0	0	0	NaN	Beregis Avtomobilya	Eldar Ryazanov	car|tro
10863	0.07	0	0	0	NaN	Beregis Avtomobilya	Eldar Ryazanov	car|tro
10864	0.06	0	0	0	NaN	What's Up, Tiger Lily?	Woody Allen	
10864	0.06	0	0	0	NaN	What's Up, Tiger Lily?	Woody Allen	
10865	0.04	19000	0	-19000	-1.00	Manos: The Hands of Fate	Harold P. Warren	fire|gun|drive|sa

26869 rows × 16 columns

In [ ]:

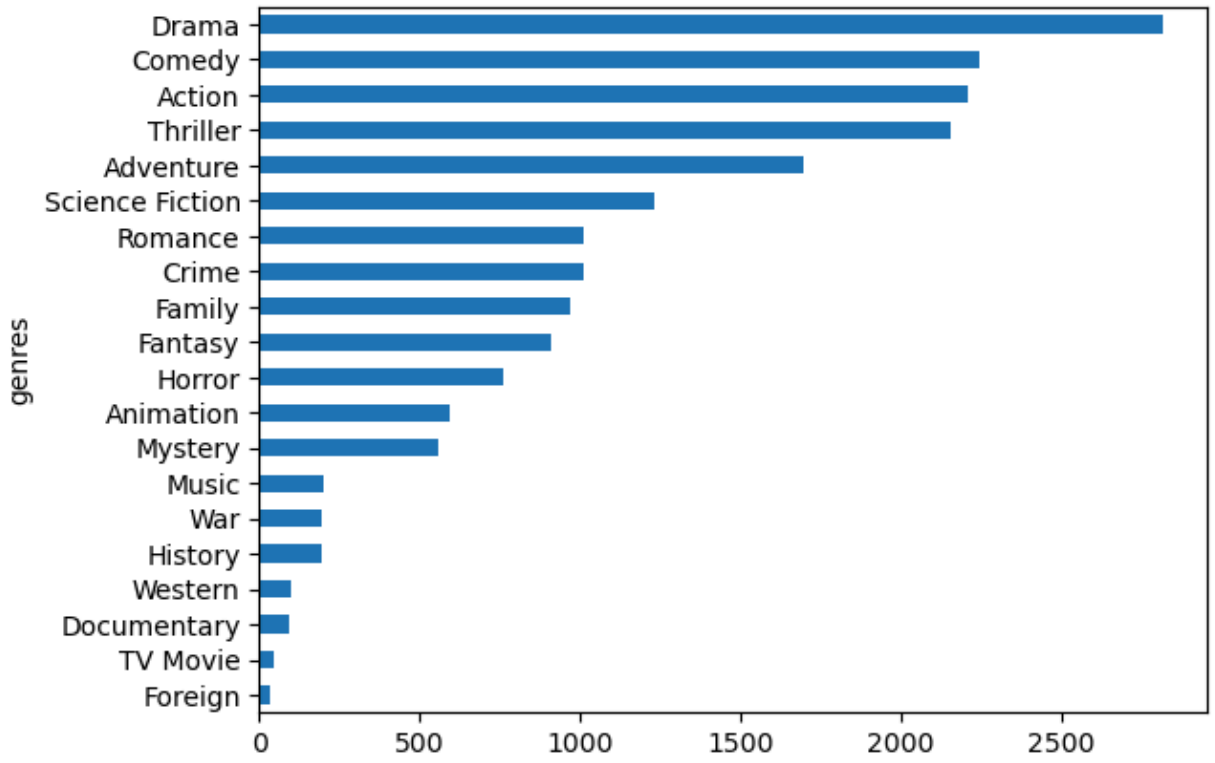
In [ ]:

In [64]:
```
df7 = df.groupby('genres')['popularity'].sum().sort_values(ascending = True)
df7
```

Out[64]:
```
genres
Foreign              35.24
TV Movie             44.03
Documentary          93.13
Western              97.42
History             192.35
War                 196.48
Music               198.15
Mystery             558.55
Animation           594.46
Horror              761.39
Fantasy             908.87
Family              967.06
Crime              1009.07
Romance            1013.21
Science Fiction    1230.41
Adventure          1697.11
Thriller           2155.90
Action             2208.08
Comedy             2246.25
Drama              2815.43
Name: popularity, dtype: float64
```

In [65]:
```
df7.plot.barh(x = 'genres', y = 'popularity')
```

Out[65]:
```
<AxesSubplot:ylabel='genres'>
```



In [66]:
```
df.head(1)
```

Out[66]:

	popularity	budget	revenue	profit	roi	original_title	director	keyw
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyrannosa rex|velociraptor|is

In [67]:
```python
df.dtypes
```

Out[67]:
```
popularity              float64
budget                    int64
revenue                   int64
profit                    int64
roi                     float64
original_title           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 [69]:
```python
df['release_date'] = pd.to_datetime(df['release_date'])
```

In [ ]:

In [70]:
```python
df['extracted_month'] = df['release_date'].dt.month
```

In [71]:
```python
df.head()
```

Out[71]:

	popularity	budget	revenue	profit	roi	original_title	director	key
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyranno rex|velocirapto
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyranno rex|velocirapto
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyranno rex|velocirapto
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyranno rex|velocirapto
1	28.42	150000000	378436354	228436354	1.52	Mad Max: Fury Road	George Miller	future|chas apocalyptic|dystopia|a

In [ ]:

```
```

In [72]:

```
df8 = df.groupby('extracted_month')['popularity'].sum()
```

In [73]:

```
df8
```

Out[73]:

```
extracted_month
1     1131.78
2     1092.93
3     1458.32
4     1191.81
5     1687.53
6     1936.84
7     1694.03
8     1432.59
9     1872.28
10    1811.91
11    1710.35
12    2002.22
Name: popularity, dtype: float64
```

In [74]:

```
df8.index
```

Out[74]:

```
Int64Index([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12], dtype='int64', name='extracted_mo
nth')
```

In [75]:

```
data = {
    'extracted_month': df8.index,
    'popularity': df8.values
}
df8 = pd.DataFrame(data)
```

In [76]:

```
df8
```

Out[76]:

	extracted_month	popularity
0	1	1131.78
1	2	1092.93
2	3	1458.32
3	4	1191.81
4	5	1687.53
5	6	1936.84
6	7	1694.03
7	8	1432.59
8	9	1872.28
9	10	1811.91
10	11	1710.35
11	12	2002.22

In [77]:
```
index_to_month = {
    1:'Jan', 2:'Feb',3:'Mar',4:'Apr',5:'May',6:'Jun',7:'Jul',8:'Aug',9:'Sep',10:'Oct',
}
```

In [78]:
```
df8.extracted_month = df8.extracted_month.map(index_to_month)
```

In [79]:
```
df8
```

Out[79]:

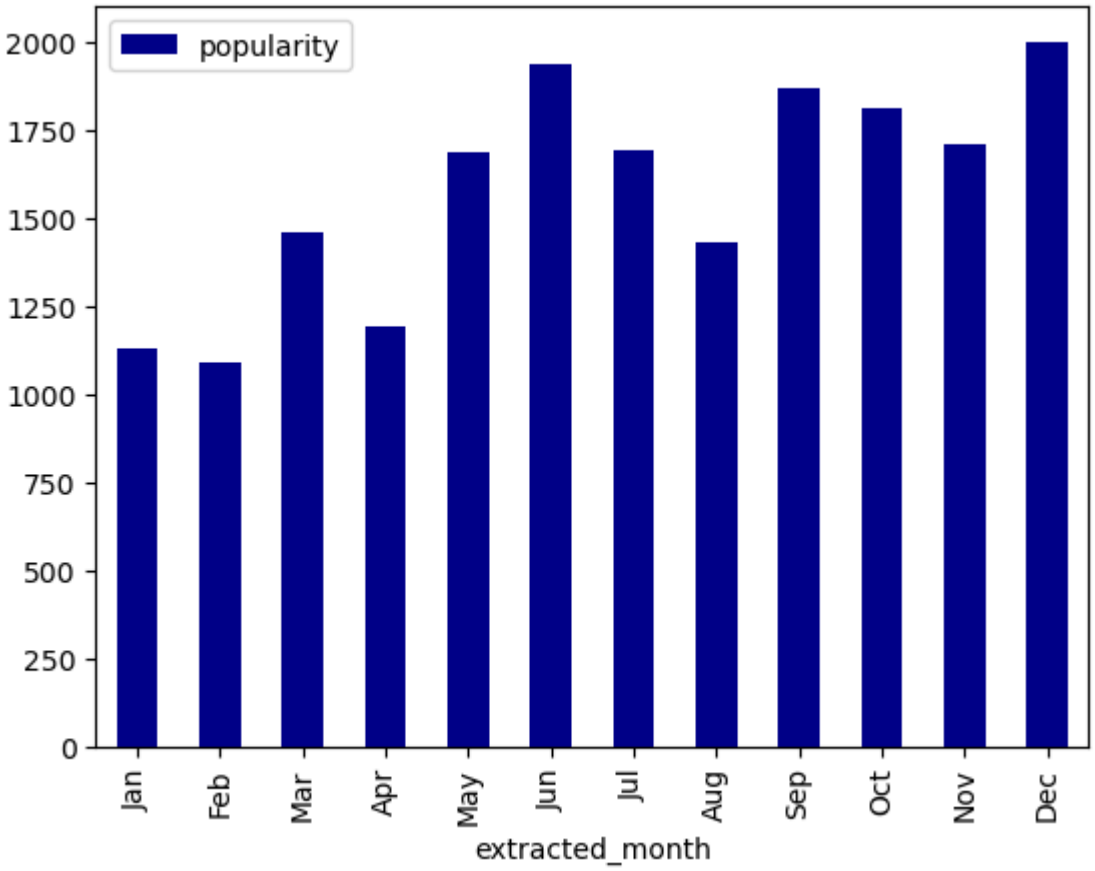
	extracted_month	popularity
0	Jan	1131.78
1	Feb	1092.93
2	Mar	1458.32
3	Apr	1191.81
4	May	1687.53
5	Jun	1936.84
6	Jul	1694.03
7	Aug	1432.59
8	Sep	1872.28
9	Oct	1811.91
10	Nov	1710.35
11	Dec	2002.22

In [ ]:
```

```

In [80]:
```
df8.plot(kind = 'bar', x = 'extracted_month', y = 'popularity', color = 'DarkBlue')
```

Out[80]:
```
<AxesSubplot:xlabel='extracted_month'>
```



In [ ]:

In [81]:
```
df9 = df.groupby('extracted_month')['revenue'].sum()
df9
```

Out[81]:
```
extracted_month
1      35873456579
2      54352852344
3      93669046441
4      77813179749
5     151475532493
6     193681776686
7     141947570995
8      71642408883
9      70379641581
10     84054172048
11    139176268899
12    164738399960
Name: revenue, dtype: int64
```

In [82]:
```
data = {
    'extracted_month': df9.index,
    'revenue': df9.values
}
df9 = pd.DataFrame(data)
```

In [83]:
```
index_to_month = {
    1:'Jan', 2:'Feb',3:'Mar',4:'Apr',5:'May',6:'Jun',7:'Jul',8:'Aug',9:'Sep',10:'Oct',
}
```

In [84]:
```
df9.extracted_month = df9.extracted_month.map(index_to_month)
```

In [85]:
```
df9
```

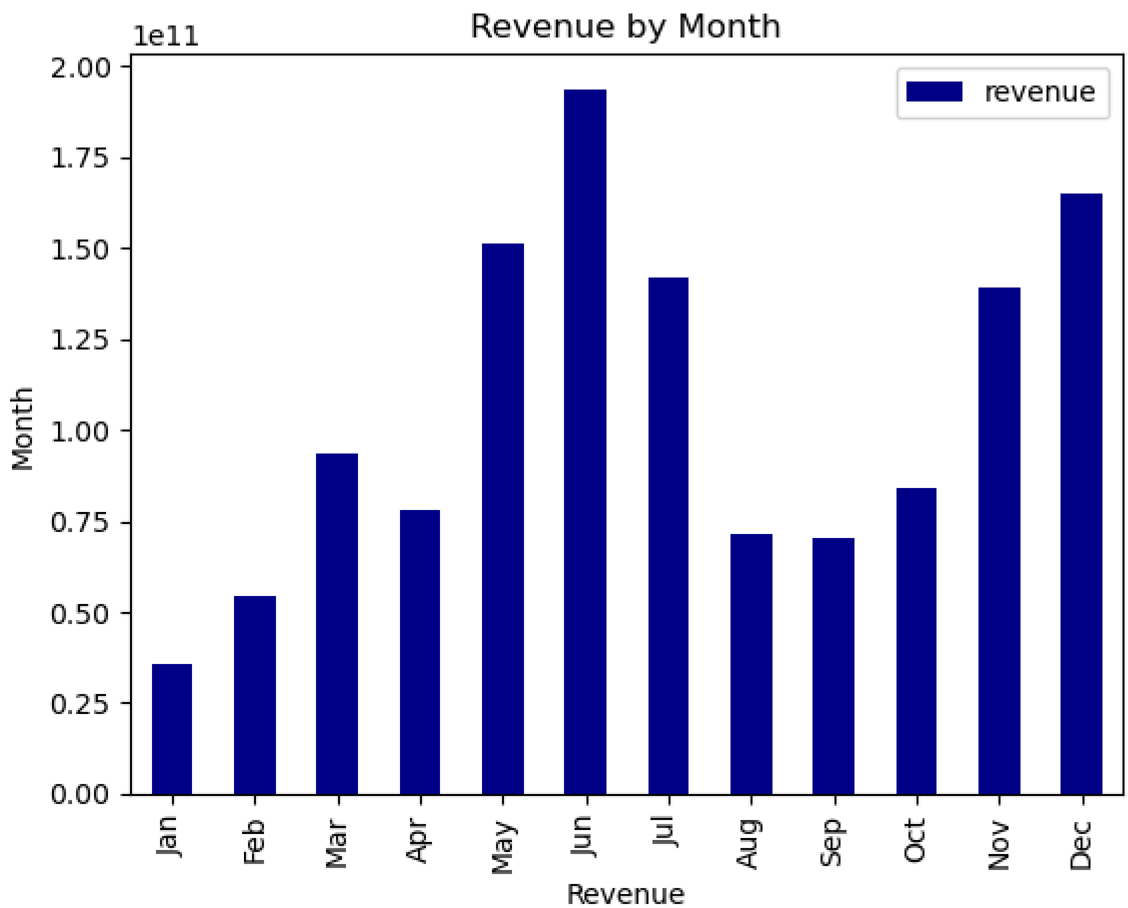
Out[85]:

	extracted_month	revenue
0	Jan	35873456579
1	Feb	54352852344
2	Mar	93669046441
3	Apr	77813179749
4	May	151475532493
5	Jun	193681776686
6	Jul	141947570995
7	Aug	71642408883
8	Sep	70379641581
9	Oct	84054172048
10	Nov	139176268899
11	Dec	164738399960

In [ ]:
```

```

In [86]:
```
df9.plot(kind = 'bar', x = 'extracted_month', y = 'revenue', color = 'DarkBlue')
plt.title('Revenue by Month')
plt.xlabel('Revenue')
plt.ylabel('Month')
plt.show()
```



In [ ]:

In [87]:
```
df.head()
```

Out[87]:

	popularity	budget	revenue	profit	roi	original_title	director	key
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyrannc rex|velocirapto
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyrannc rex|velocirapto
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyrannc rex|velocirapto
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyrannc rex|velocirapto
1	28.42	150000000	378436354	228436354	1.52	Mad Max: Fury Road	George Miller	future|chas apocalyptic|dystopia|a

In [ ]:

In [90]:
```
df10=df.groupby('original_title')['profit'].sum().sort_values(ascending = False).head(
df10
```

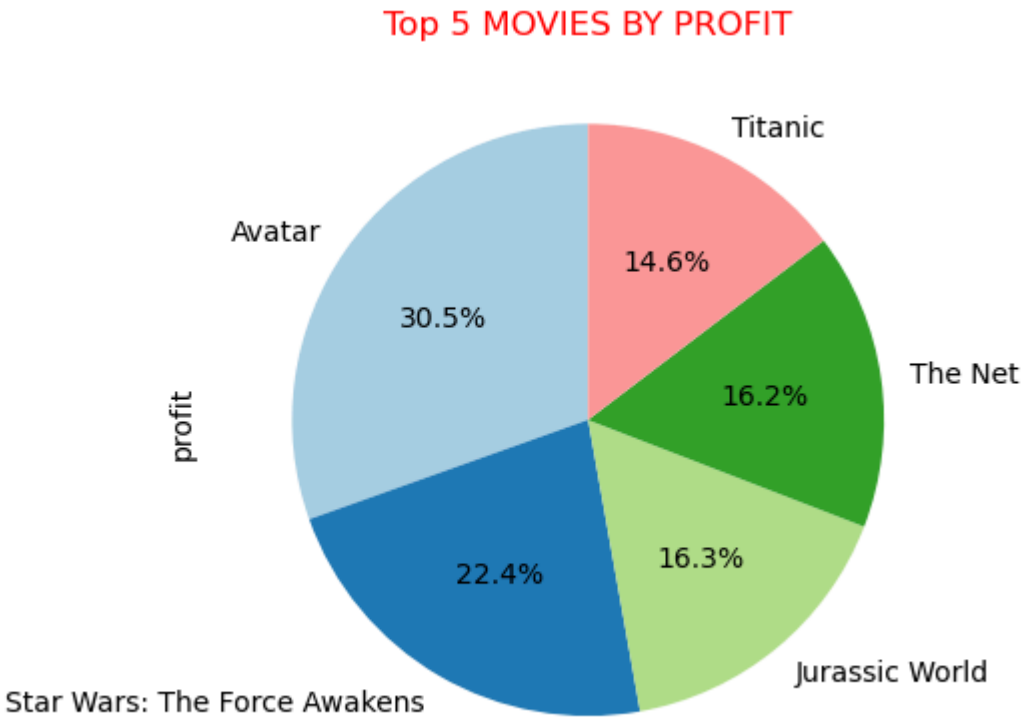
Out[90]:
```
original_title
Avatar                          10178023388
Star Wars: The Force Awakens     7472712900
Jurassic World                   5454115240
The Net                          5421398290
Titanic                          4896102564
Name: profit, dtype: int64
```

In [ ]:
```

```

In [93]:
```
df10.plot(kind = 'pie', autopct = '%1.1f%%',startangle = 90, colors = plt.cm.Paired.cc
plt.title('Top 5 MOVIES BY PROFIT',color = 'red')
```

Out[93]:
```
Text(0.5, 1.0, 'Top 5 MOVIES BY PROFIT')
```



In [96]:
```
df.head()
```

Out[96]:

	popularity	budget	revenue	profit	roi	original_title	director	key
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyrannc rex|velocirapto
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyrannc rex|velocirapto
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyrannc rex|velocirapto
0	32.99	150000000	1513528810	1363528810	9.09	Jurassic World	Colin Trevorrow	monster|dna|tyrannc rex|velocirapto
1	28.42	150000000	378436354	228436354	1.52	Mad Max: Fury Road	George Miller	future|chas apocalyptic|dystopia|a

In [97]:

```
df11 = df.production_companies.value_counts().head(5)
df11
```

Out[97]:

```
0                        2152
Paramount Pictures        404
Universal Pictures        352
Walt Disney Pictures      236
Warner Bros.              225
Name: production_companies, dtype: int64
```

In [102…

```
df11.index
```

Out[102]:

```
Index([                     0,    'Paramount Pictures',    'Universal Pictures',
       'Walt Disney Pictures',          'Warner Bros.'],
      dtype='object')
```

In [ ]:

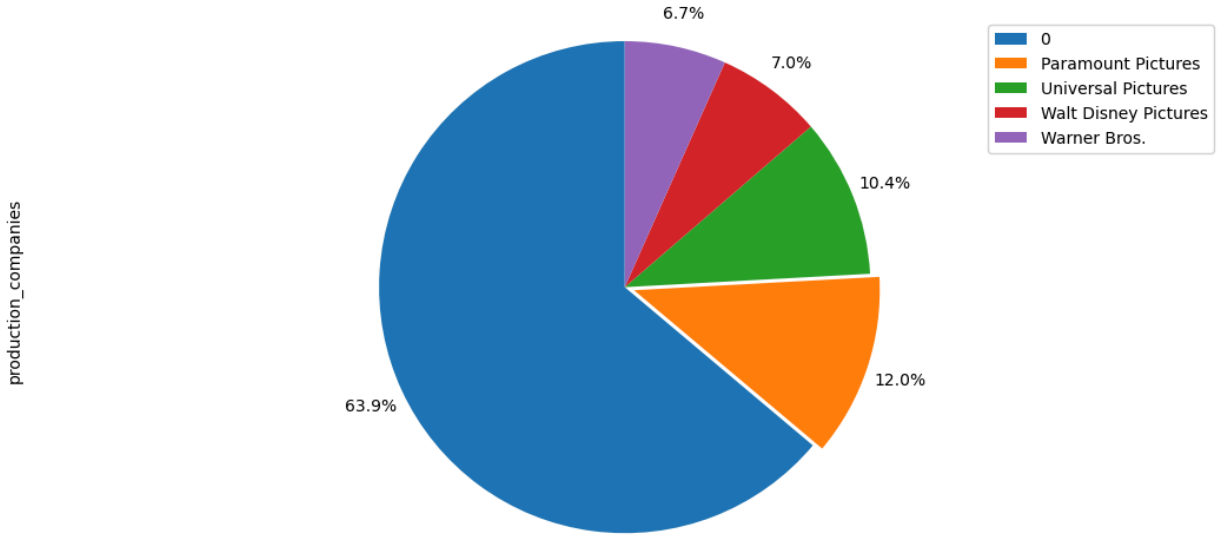
```

```

In [110…

```
explode_list = [0,0.04,0,0,0]
df11.plot(kind = 'pie', figsize = (13,6), autopct = '%1.1f%%',startangle = 90, labels
#plt.title('Top 5 MOVIES BY PROFIT',color = 'red')

plt.legend(labels = df11.index, loc = 'upper right')
plt.axis('equal')
plt.show()
```


production_companies
6.7%
7.0%
10.4%
12.0%
63.9%
0
Paramount Pictures
Universal Pictures
Walt Disney Pictures
Warner Bros.

In [ ]:

In [ ]:

In [ ]: