```
import pandas as pd
import numpy as np
import seaborn as sns

import matplotlib.pyplot as plt
import matplotlib.mlab as mlab
import matplotlib
plt.style.use('ggplot')
from matplotlib.pyplot import figure

%matplotlib inline
matplotlib.rcParams['figure.figsize'] = (12,8)

pd.options.mode.chained_assignment = None



# Now we need to read in the data
df = pd.read_csv(r'C:\Users\Hi\Downloads\movies.csv')

df
```

```
                                                name rating      genre
year  \
0                                        The Shining      R      Drama
1980
1                                    The Blue Lagoon      R  Adventure
1980
2     Star Wars: Episode V - The Empire Strikes Back     PG     Action
1980
3                                          Airplane!     PG     Comedy
1980
4                                         Caddyshack      R     Comedy
1980
...                                              ...    ...        ...
...
7663                                    More to Life    NaN      Drama
2020
7664                                     Dream Round    NaN     Comedy
2020
7665                                  Saving Mbango    NaN      Drama
2020
7666                                    It's Just Us    NaN      Drama
2020
7667                                     Tee em el    NaN     Horror
2020

                            released  score      votes
director  \
0        June 13, 1980 (United States)    8.4   927000.0  Stanley
```

```
Kubrick
1         July 2, 1980 (United States)    5.8    65000.0   Randal
Kleiser
2        June 20, 1980 (United States)    8.7  1200000.0   Irvin
Kershner
3         July 2, 1980 (United States)    7.7   221000.0     Jim
Abrahams
4        July 25, 1980 (United States)    7.3   108000.0     Harold
Ramis
...                                ...    ...        ...
...
7663  October 23, 2020 (United States)    3.1       18.0    Joseph
Ebanks
7664  February 7, 2020 (United States)    4.7       36.0     Dusty
Dukatz
7665         April 27, 2020 (Cameroon)    5.7       29.0     Nkanya
Nkwai
7666   October 1, 2020 (United States)    NaN        NaN    James
Randall
7667   August 19, 2020 (United States)    5.7        7.0     Pereko
Mosia

                       writer              star         country
budget  \
0                Stephen King    Jack Nicholson  United Kingdom
19000000.0
1     Henry De Vere Stacpoole    Brooke Shields   United States
4500000.0
2              Leigh Brackett       Mark Hamill   United States
18000000.0
3                Jim Abrahams       Robert Hays   United States
3500000.0
4          Brian Doyle-Murray       Chevy Chase   United States
6000000.0
...                       ...               ...             ...
...
7663            Joseph Ebanks      Shannon Bond   United States
7000.0
7664              Lisa Huston  Michael Saquella   United States
NaN
7665             Lynno Lovert      Onyama Laura   United States
58750.0
7666            James Randall     Christina Roz   United States
15000.0
7667             Pereko Mosia  Siyabonga Mabaso    South Africa
NaN

            gross                    company  runtime
0      46998772.0               Warner Bros.    146.0
```

```
1      58853106.0          Columbia Pictures    104.0
2     538375067.0                  Lucasfilm    124.0
3      83453539.0         Paramount Pictures     88.0
4      39846344.0            Orion Pictures      98.0
...           ...                        ...      ...
7663          NaN                        NaN     90.0
7664          NaN  Cactus Blue Entertainment     90.0
7665          NaN           Embi Productions      NaN
7666          NaN                        NaN    120.0
7667          NaN                PK 65 Films    102.0

[7668 rows x 15 columns]
```

```
for col in df.columns:
    pct_missing = np.mean(df[col].isnull())
    print('{} - {}%'.format(col, round(pct_missing*100)))
```

```
name - 0%
rating - 1%
genre - 0%
year - 0%
released - 0%
score - 0%
votes - 0%
director - 0%
writer - 0%
star - 0%
country - 0%
budget - 28%
gross - 2%
company - 0%
runtime - 0%
```

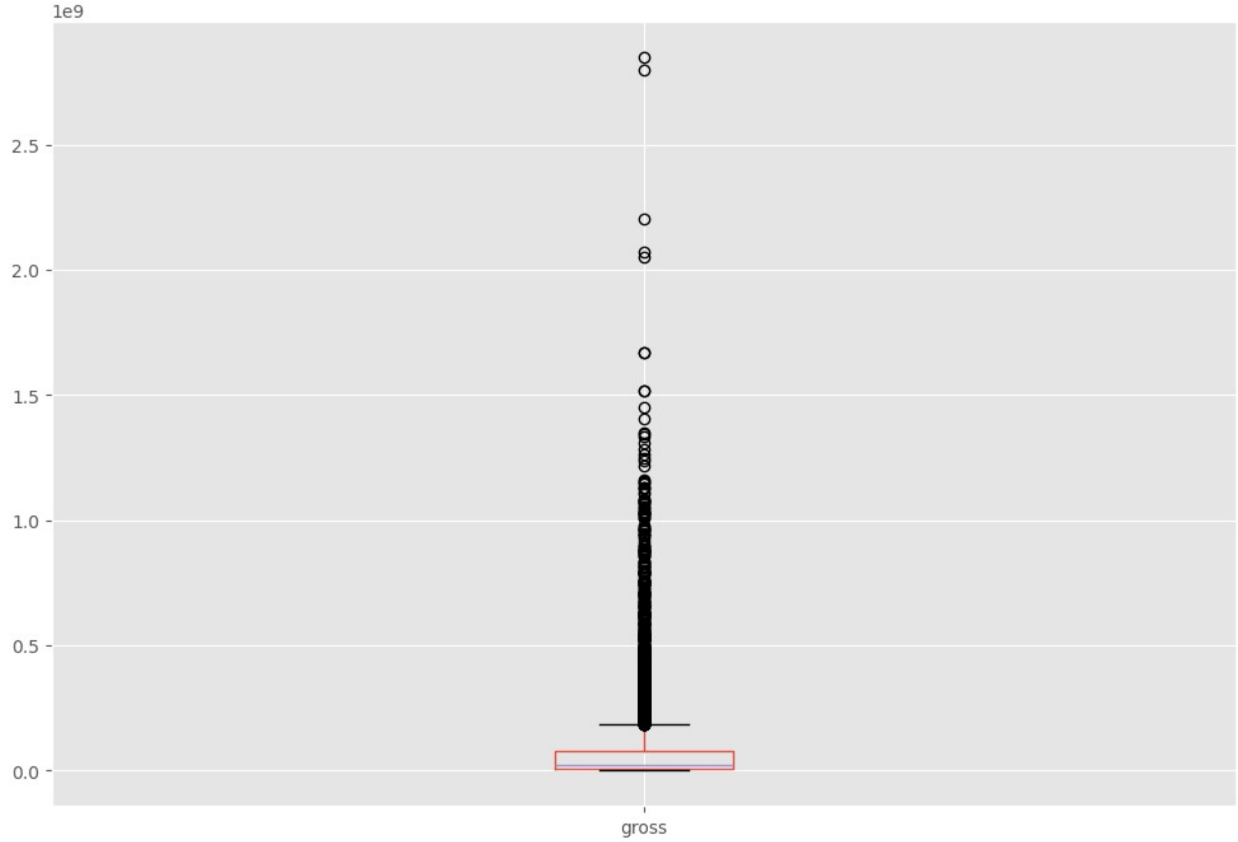
```
print(df.dtypes)
```

```
name         object
rating       object
genre        object
year          int64
released     object
score       float64
votes       float64
director     object
writer       object
star         object
country      object
budget      float64
gross       float64
company      object
```

```
runtime     float64
dtype: object
```

```
df.boxplot(column=['gross'])
```

```
<Axes: >
```



```
df.drop_duplicates()
```

```
                                                   name rating      genre  
year  \
0                                           The Shining      R      Drama  
1980   
1                                       The Blue Lagoon      R  Adventure  
1980   
2        Star Wars: Episode V - The Empire Strikes Back     PG     Action  
1980   
3                                             Airplane!     PG     Comedy  
1980   
4                                            Caddyshack      R     Comedy  
1980   
...                                                 ...    ...        ...  
...   
7663                                       More to Life    NaN      Drama  
```

```
2020
7664                                     Dream Round    NaN     Comedy
2020
7665                                   Saving Mbango    NaN      Drama
2020
7666                                    It's Just Us    NaN      Drama
2020
7667                                       Tee em el    NaN     Horror
2020

                              released  score      votes
director  \
0        June 13, 1980 (United States)    8.4   927000.0  Stanley
Kubrick
1         July 2, 1980 (United States)    5.8    65000.0   Randal
Kleiser
2        June 20, 1980 (United States)    8.7  1200000.0   Irvin
Kershner
3         July 2, 1980 (United States)    7.7   221000.0     Jim
Abrahams
4        July 25, 1980 (United States)    7.3   108000.0     Harold
Ramis
...                                ...    ...        ...
...
7663  October 23, 2020 (United States)    3.1       18.0    Joseph
Ebanks
7664  February 7, 2020 (United States)    4.7       36.0     Dusty
Dukatz
7665         April 27, 2020 (Cameroon)    5.7       29.0     Nkanya
Nkwai
7666   October 1, 2020 (United States)    NaN        NaN    James
Randall
7667   August 19, 2020 (United States)    5.7        7.0     Pereko
Mosia

                       writer              star         country
budget  \
0                Stephen King    Jack Nicholson  United Kingdom
19000000.0
1     Henry De Vere Stacpoole    Brooke Shields   United States
4500000.0
2              Leigh Brackett       Mark Hamill   United States
18000000.0
3                Jim Abrahams       Robert Hays   United States
3500000.0
4          Brian Doyle-Murray       Chevy Chase   United States
6000000.0
...                       ...               ...             ...
...
```

```
7663            Joseph Ebanks      Shannon Bond   United States
7000.0
7664              Lisa Huston  Michael Saquella   United States
NaN
7665             Lynno Lovert      Onyama Laura   United States
58750.0
7666            James Randall     Christina Roz   United States
15000.0
7667             Pereko Mosia  Siyabonga Mabaso    South Africa
NaN

            gross                    company  runtime
0      46998772.0               Warner Bros.    146.0
1      58853106.0          Columbia Pictures    104.0
2     538375067.0                  Lucasfilm    124.0
3      83453539.0         Paramount Pictures     88.0
4      39846344.0             Orion Pictures     98.0
...           ...                        ...      ...
7663          NaN                        NaN     90.0
7664          NaN  Cactus Blue Entertainment     90.0
7665          NaN           Embi Productions      NaN
7666          NaN                        NaN    120.0
7667          NaN                PK 65 Films    102.0

[7668 rows x 15 columns]
```

```
df.sort_values(by=['gross'], inplace=False, ascending=False)
```

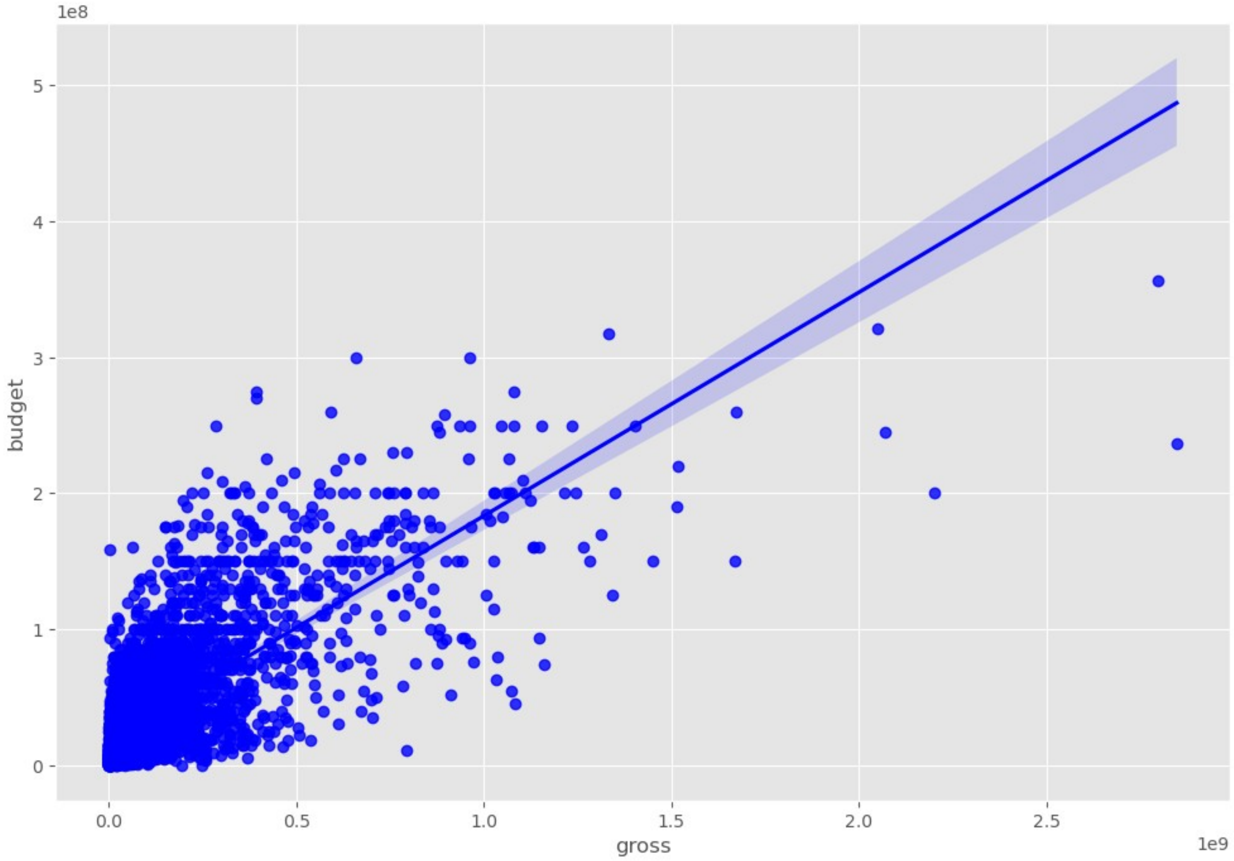
```
                                            name rating   genre  year
\
5445                                      Avatar  PG-13  Action  2009

7445                           Avengers: Endgame  PG-13  Action  2019

3045                                     Titanic  PG-13   Drama  1997

6663  Star Wars: Episode VII - The Force Awakens  PG-13  Action  2015

7244                      Avengers: Infinity War  PG-13  Action  2018

...                                          ...    ...     ...   ...

7663                                More to Life    NaN   Drama  2020

7664                                 Dream Round    NaN  Comedy  2020

7665                               Saving Mbango    NaN   Drama  2020

7666                                It's Just Us    NaN   Drama  2020

7667                                   Tee em el    NaN  Horror  2020
```

```
                               released  score      votes
director  \
5445  December 18, 2009 (United States)    7.8  1100000.0  James
Cameron
7445     April 26, 2019 (United States)    8.4   903000.0  Anthony
Russo
3045  December 19, 1997 (United States)    7.8  1100000.0  James
Cameron
6663  December 18, 2015 (United States)    7.8   876000.0    J.J.
Abrams
7244     April 27, 2018 (United States)    8.4   897000.0  Anthony
Russo
...                                ...    ...        ...             .
..
7663   October 23, 2020 (United States)    3.1       18.0  Joseph
Ebanks
7664   February 7, 2020 (United States)    4.7       36.0   Dusty
Dukatz
7665          April 27, 2020 (Cameroon)    5.7       29.0   Nkanya
Nkwai
7666    October 1, 2020 (United States)    NaN        NaN  James
Randall
7667    August 19, 2020 (United States)    5.7        7.0   Pereko
Mosia

                  writer               star        country
budget  \
5445       James Cameron    Sam Worthington  United States
237000000.0
7445  Christopher Markus  Robert Downey Jr.  United States
356000000.0
3045       James Cameron  Leonardo DiCaprio  United States
200000000.0
6663     Lawrence Kasdan       Daisy Ridley  United States
245000000.0
7244  Christopher Markus  Robert Downey Jr.  United States
321000000.0
...                  ...                ...            ...          ..
.
7663       Joseph Ebanks       Shannon Bond  United States
7000.0
7664         Lisa Huston   Michael Saquella  United States
NaN
7665        Lynno Lovert       Onyama Laura  United States
58750.0
7666       James Randall      Christina Roz  United States
15000.0
7667        Pereko Mosia   Siyabonga Mabaso   South Africa
```

```
NaN

             gross                      company  runtime
5445  2.847246e+09         Twentieth Century Fox    162.0
7445  2.797501e+09                Marvel Studios    181.0
3045  2.201647e+09         Twentieth Century Fox    194.0
6663  2.069522e+09                     Lucasfilm    138.0
7244  2.048360e+09                Marvel Studios    149.0
...            ...                          ...      ...
7663           NaN                          NaN     90.0
7664           NaN  Cactus Blue Entertainment     90.0
7665           NaN             Embi Productions      NaN
7666           NaN                          NaN    120.0
7667           NaN                  PK 65 Films    102.0

[7668 rows x 15 columns]
```

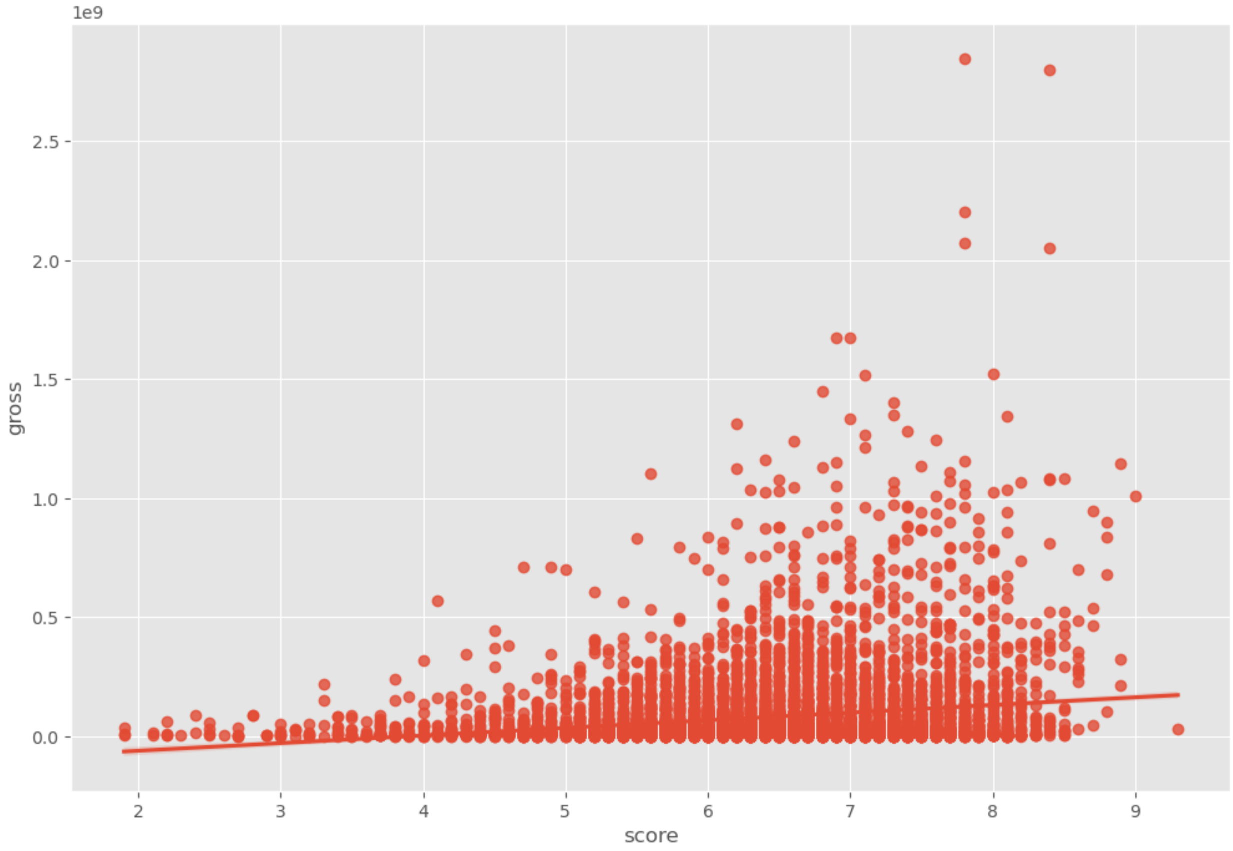
```
sns.regplot(x="gross", y="budget", data=df, color='blue')
```

```
<Axes: xlabel='gross', ylabel='budget'>
```



```
sns.regplot(x="score", y="gross", data=df)
```

```
<Axes: xlabel='score', ylabel='gross'>
```



```
df
                                             name rating      genre
year  \
0                                     The Shining      R      Drama
1980
1                                 The Blue Lagoon      R  Adventure
1980
2     Star Wars: Episode V - The Empire Strikes Back     PG     Action
1980
3                                       Airplane!     PG     Comedy
1980
4                                      Caddyshack      R     Comedy
1980
...                                           ...    ...        ...
...
7663                                 More to Life    NaN      Drama
2020
7664                                  Dream Round    NaN     Comedy
2020
7665                               Saving Mbango    NaN      Drama
```

```
2020
7666                                    It's Just Us    NaN      Drama
2020
7667                                       Tee em el    NaN     Horror
2020

                              released  score      votes
director  \
0        June 13, 1980 (United States)    8.4   927000.0  Stanley
Kubrick
1         July 2, 1980 (United States)    5.8    65000.0   Randal
Kleiser
2        June 20, 1980 (United States)    8.7  1200000.0   Irvin
Kershner
3         July 2, 1980 (United States)    7.7   221000.0     Jim
Abrahams
4        July 25, 1980 (United States)    7.3   108000.0     Harold
Ramis
...                                ...    ...        ...
...
7663  October 23, 2020 (United States)    3.1       18.0    Joseph
Ebanks
7664  February 7, 2020 (United States)    4.7       36.0     Dusty
Dukatz
7665         April 27, 2020 (Cameroon)    5.7       29.0     Nkanya
Nkwai
7666   October 1, 2020 (United States)    NaN        NaN    James
Randall
7667   August 19, 2020 (United States)    5.7        7.0     Pereko
Mosia

                       writer              star         country
budget  \
0                Stephen King    Jack Nicholson  United Kingdom
19000000.0
1     Henry De Vere Stacpoole    Brooke Shields   United States
4500000.0
2              Leigh Brackett       Mark Hamill   United States
18000000.0
3                Jim Abrahams       Robert Hays   United States
3500000.0
4          Brian Doyle-Murray       Chevy Chase   United States
6000000.0
...                       ...               ...             ...
...
7663            Joseph Ebanks      Shannon Bond   United States
7000.0
7664              Lisa Huston  Michael Saquella   United States
NaN
```

```
7665             Lynno Lovert      Onyama Laura   United States
58750.0
7666            James Randall     Christina Roz   United States
15000.0
7667             Pereko Mosia  Siyabonga Mabaso    South Africa
NaN

            gross                    company  runtime
0      46998772.0               Warner Bros.    146.0
1      58853106.0          Columbia Pictures    104.0
2     538375067.0                  Lucasfilm    124.0
3      83453539.0         Paramount Pictures     88.0
4      39846344.0             Orion Pictures     98.0
...           ...                        ...      ...
7663          NaN                        NaN     90.0
7664          NaN  Cactus Blue Entertainment     90.0
7665          NaN           Embi Productions      NaN
7666          NaN                        NaN    120.0
7667          NaN                PK 65 Films    102.0

[7668 rows x 15 columns]
```

```
print(df.head())
```

```
                                             name rating      genre
year  \
0                                     The Shining      R      Drama
1980
1                                 The Blue Lagoon      R  Adventure
1980
2  Star Wars: Episode V - The Empire Strikes Back     PG     Action
1980
3                                       Airplane!     PG     Comedy
1980
4                                      Caddyshack      R     Comedy
1980

                        released  score      votes         director  \
0  June 13, 1980 (United States)    8.4   927000.0  Stanley Kubrick
1   July 2, 1980 (United States)    5.8    65000.0   Randal Kleiser
2  June 20, 1980 (United States)    8.7  1200000.0   Irvin Kershner
3   July 2, 1980 (United States)    7.7   221000.0     Jim Abrahams
4  July 25, 1980 (United States)    7.3   108000.0     Harold Ramis

                    writer            star         country      budget
\
0             Stephen King  Jack Nicholson  United Kingdom  19000000.0

1  Henry De Vere Stacpoole  Brooke Shields   United States   4500000.0
```

```
2           Leigh Brackett     Mark Hamill   United States  18000000.0

3             Jim Abrahams     Robert Hays   United States   3500000.0

4       Brian Doyle-Murray     Chevy Chase   United States   6000000.0


         gross             company  runtime
0   46998772.0        Warner Bros.    146.0
1   58853106.0   Columbia Pictures    104.0
2  538375067.0           Lucasfilm    124.0
3   83453539.0  Paramount Pictures     88.0
4   39846344.0      Orion Pictures     98.0
```

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7668 entries, 0 to 7667
Data columns (total 15 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   name      7668 non-null   object 
 1   rating    7591 non-null   object 
 2   genre     7668 non-null   object 
 3   year      7668 non-null   int64  
 4   released  7666 non-null   object 
 5   score     7665 non-null   float64
 6   votes     7665 non-null   float64
 7   director  7668 non-null   object 
 8   writer    7665 non-null   object 
 9   star      7667 non-null   object 
 10  country   7665 non-null   object 
 11  budget    5497 non-null   float64
 12  gross     7479 non-null   float64
 13  company   7651 non-null   object 
 14  runtime   7664 non-null   float64
dtypes: float64(5), int64(1), object(9)
memory usage: 898.7+ KB
None
```

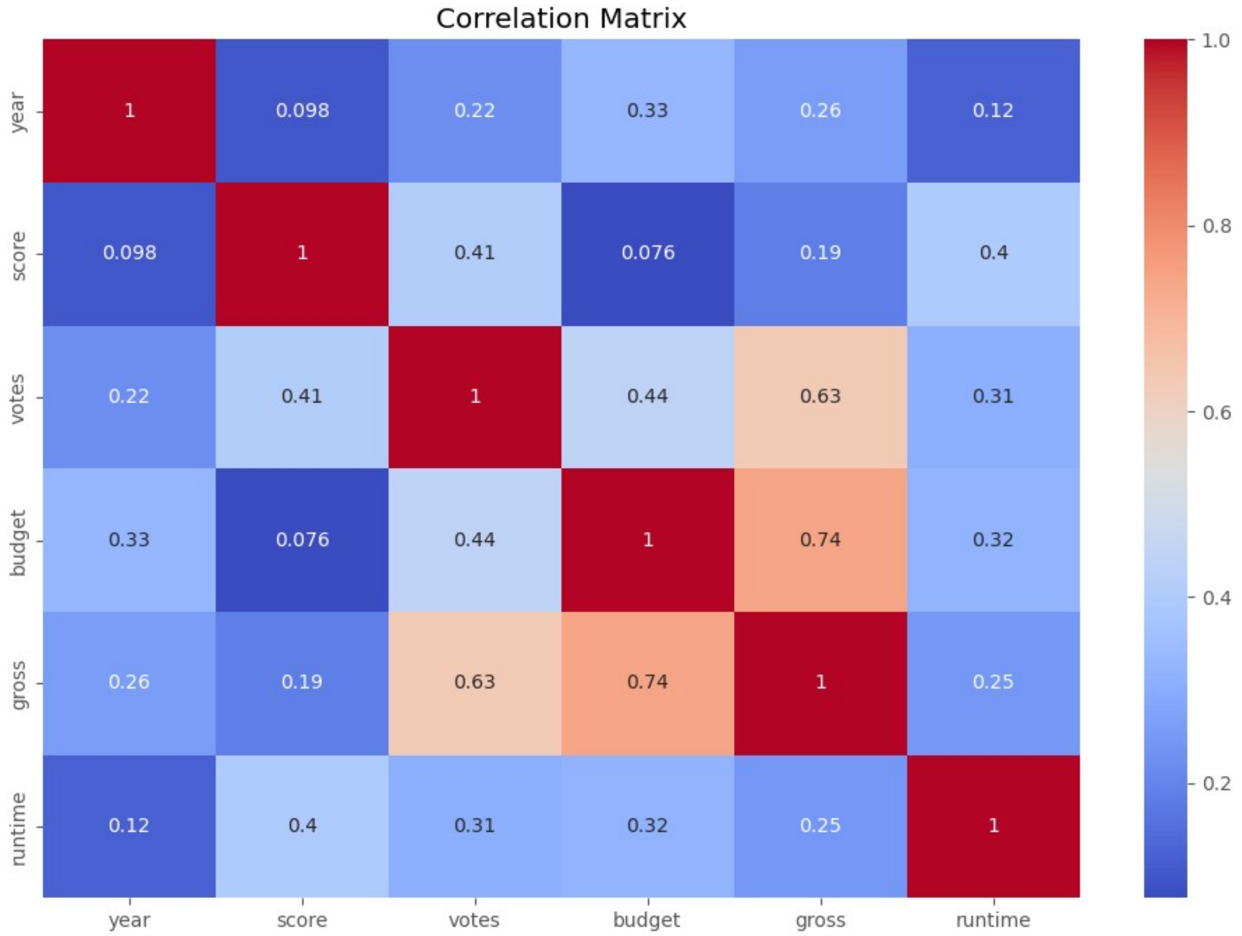
```
numerical_columns = df.select_dtypes(include=['float64',
'int64']).columns
print("Numerical Columns:", numerical_columns)
```

```
Numerical Columns: Index(['year', 'score', 'votes', 'budget', 'gross',
'runtime'], dtype='object')
```

```
correlation_matrix = df[numerical_columns].corr()

print(correlation_matrix)
```

```
             year     score     votes    budget     gross   runtime
year     1.000000  0.097995  0.222945  0.329321  0.257486  0.120811
score    0.097995  1.000000  0.409182  0.076254  0.186258  0.399451
votes    0.222945  0.409182  1.000000  0.442429  0.630757  0.309212
budget   0.329321  0.076254  0.442429  1.000000  0.740395  0.320447
gross    0.257486  0.186258  0.630757  0.740395  1.000000  0.245216
runtime  0.120811  0.399451  0.309212  0.320447  0.245216  1.000000
```

```
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



```
numeric_df = df.select_dtypes(include=[float, int])
correlation_matrix = numeric_df.corr(method='pearson')
print(correlation_matrix)
```

```
             year     score     votes    budget     gross   runtime
year     1.000000  0.097995  0.222945  0.329321  0.257486  0.120811
score    0.097995  1.000000  0.409182  0.076254  0.186258  0.399451
votes    0.222945  0.409182  1.000000  0.442429  0.630757  0.309212
```

```
budget   0.329321  0.076254  0.442429  1.000000  0.740395  0.320447
gross    0.257486  0.186258  0.630757  0.740395  1.000000  0.245216
runtime  0.120811  0.399451  0.309212  0.320447  0.245216  1.000000
```

```python
numeric_df = df.select_dtypes(include=[float, int])
correlation_matrix = numeric_df.corr(method='kendall')
print(correlation_matrix)
```

```
             year     score     votes    budget     gross   runtime
year     1.000000  0.067652  0.331465  0.224120  0.200618  0.097184
score    0.067652  1.000000  0.300115 -0.000566  0.086046  0.283611
votes    0.331465  0.300115  1.000000  0.353702  0.548899  0.198240
budget   0.224120 -0.000566  0.353702  1.000000  0.512637  0.235483
gross    0.200618  0.086046  0.548899  0.512637  1.000000  0.168933
runtime  0.097184  0.283611  0.198240  0.235483  0.168933  1.000000
```

```python
numeric_df = df.select_dtypes(include=[float, int])
correlation_matrix = numeric_df.corr(method='spearman')
print(correlation_matrix)
```

```
             year     score     votes    budget     gross   runtime
year     1.000000  0.099045  0.469829  0.317336  0.293084  0.142977
score    0.099045  1.000000  0.428138 -0.001403  0.126116  0.399857
votes    0.469829  0.428138  1.000000  0.502466  0.742050  0.290159
budget   0.317336 -0.001403  0.502466  1.000000  0.693670  0.336370
gross    0.293084  0.126116  0.742050  0.693670  1.000000  0.246243
runtime  0.142977  0.399857  0.290159  0.336370  0.246243  1.000000
```

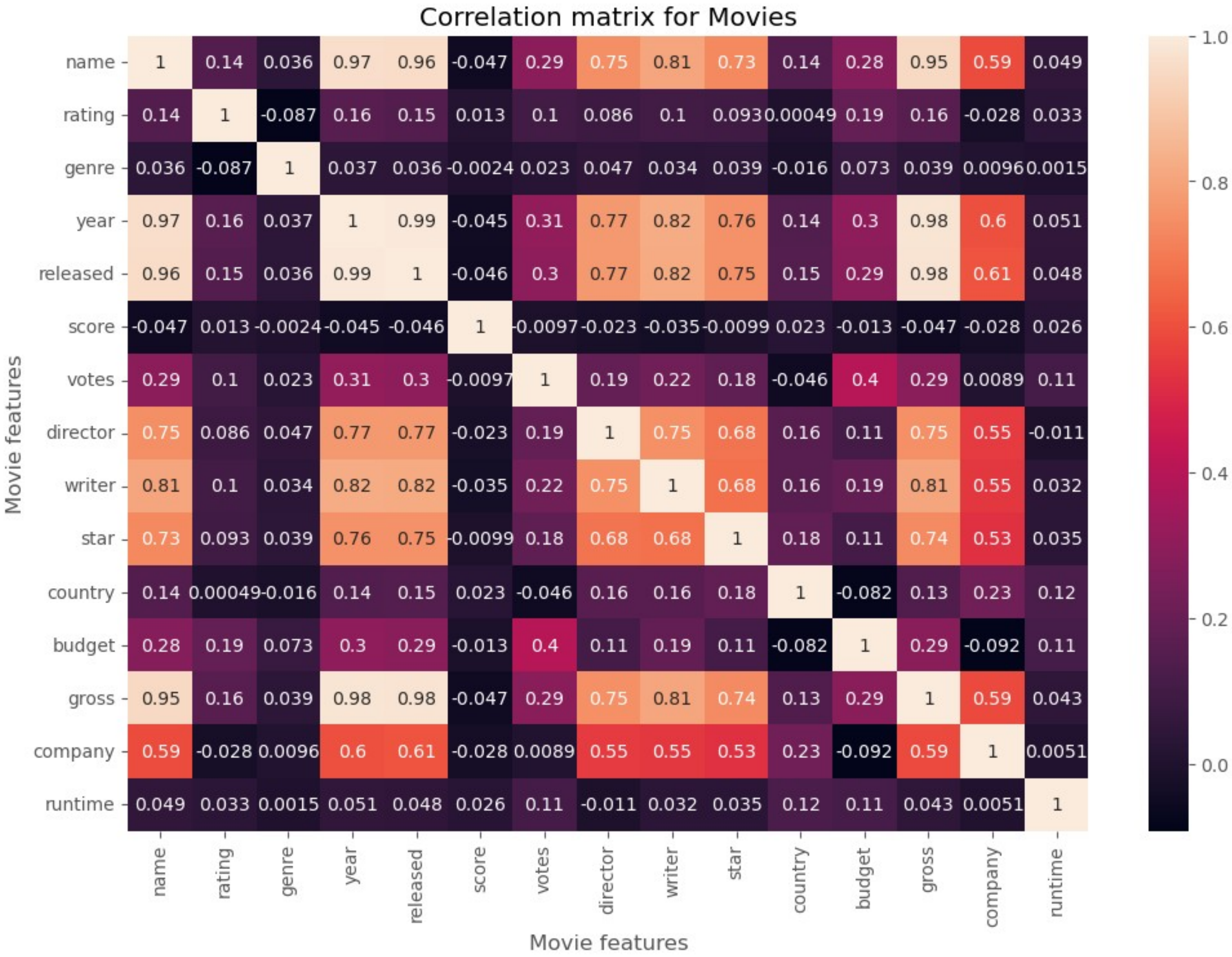
```python
df.apply(lambda x: x.factorize()[0]).corr(method='pearson')
```

```
              name    rating     genre      year  released
score  \
name      1.000000  0.143938  0.036367  0.965761  0.959015 -0.046733

rating    0.143938  1.000000 -0.086723  0.156713  0.146606  0.012595

genre     0.036367 -0.086723  1.000000  0.037184  0.035940 -0.002437

year      0.965761  0.156713  0.037184  1.000000  0.993190 -0.044981

released  0.959015  0.146606  0.035940  0.993190  1.000000 -0.045761

score    -0.046733  0.012595 -0.002437 -0.044981 -0.045761  1.000000

votes     0.287776  0.099972  0.023285  0.312401  0.299905 -0.009749

director  0.745905  0.085520  0.047288  0.770497  0.770876 -0.022687

writer    0.805211  0.103623  0.033688  0.824770  0.819617 -0.034685

star      0.731565  0.093116  0.038649  0.756400  0.754468 -0.009896
```

```
country   0.142828  0.000494 -0.015795  0.140216  0.148468  0.023097

budget    0.277488  0.193353  0.073008  0.300621  0.285691 -0.012642

gross     0.947324  0.158582  0.038616  0.980873  0.976423 -0.047041

company   0.591667 -0.028035  0.009566  0.601571  0.607954 -0.028432

runtime   0.048955  0.032741  0.001462  0.050647  0.048235  0.026436


             votes  director    writer      star   country  
budget  \
name      0.287776  0.745905  0.805211  0.731565  0.142828  0.277488

rating    0.099972  0.085520  0.103623  0.093116  0.000494  0.193353

genre     0.023285  0.047288  0.033688  0.038649 -0.015795  0.073008

year      0.312401  0.770497  0.824770  0.756400  0.140216  0.300621

released  0.299905  0.770876  0.819617  0.754468  0.148468  0.285691

score    -0.009749 -0.022687 -0.034685 -0.009896  0.023097 -0.012642

votes     1.000000  0.192220  0.224122  0.179601 -0.045914  0.398519

director  0.192220  1.000000  0.748340  0.682385  0.155471  0.106617

writer    0.224122  0.748340  1.000000  0.675685  0.157202  0.187238

star      0.179601  0.682385  0.675685  1.000000  0.182045  0.107991

country  -0.045914  0.155471  0.157202  0.182045  1.000000 -0.082082

budget    0.398519  0.106617  0.187238  0.107991 -0.082082  1.000000

gross     0.286180  0.750911  0.805576  0.735680  0.133982  0.285832

company   0.008900  0.552258  0.546151  0.527116  0.226346 -0.092249

runtime   0.106024 -0.011070  0.032264  0.035392  0.124154  0.112097


             gross   company   runtime
name      0.947324  0.591667  0.048955
rating    0.158582 -0.028035  0.032741
genre     0.038616  0.009566  0.001462
year      0.980873  0.601571  0.050647
released  0.976423  0.607954  0.048235
score    -0.047041 -0.028432  0.026436
```

```
votes     0.286180  0.008900  0.106024
director  0.750911  0.552258 -0.011070
writer    0.805576  0.546151  0.032264
star      0.735680  0.527116  0.035392
country   0.133982  0.226346  0.124154
budget    0.285832 -0.092249  0.112097
gross     1.000000  0.588156  0.042978
company   0.588156  1.000000  0.005137
runtime   0.042978  0.005137  1.000000
```

```
correlation_matrix = df.apply(lambda x: x.factorize()
[0]).corr(method='pearson')

sns.heatmap(correlation_matrix, annot = True)

plt.title("Correlation matrix for Movies")

plt.xlabel("Movie features")

plt.ylabel("Movie features")

plt.show()
```



	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime
name	1	0.14	0.036	0.97	0.96	-0.047	0.29	0.75	0.81	0.73	0.14	0.28	0.95	0.59	0.049
rating	0.14	1	-0.087	0.16	0.15	0.013	0.1	0.086	0.1	0.093	0.00049	0.19	0.16	-0.028	0.033
genre	0.036	-0.087	1	0.037	0.036	-0.0024	0.023	0.047	0.034	0.039	-0.016	0.073	0.039	0.0096	0.0015
year	0.97	0.16	0.037	1	0.99	-0.045	0.31	0.77	0.82	0.76	0.14	0.3	0.98	0.6	0.051
released	0.96	0.15	0.036	0.99	1	-0.046	0.3	0.77	0.82	0.75	0.15	0.29	0.98	0.61	0.048
score	-0.047	0.013	-0.0024	-0.045	-0.046	1	-0.0097	-0.023	-0.035	-0.0099	0.023	-0.013	-0.047	-0.028	0.026
votes	0.29	0.1	0.023	0.31	0.3	-0.0097	1	0.19	0.22	0.18	-0.046	0.4	0.29	0.0089	0.11
director	0.75	0.086	0.047	0.77	0.77	-0.023	0.19	1	0.75	0.68	0.16	0.11	0.75	0.55	-0.011
writer	0.81	0.1	0.034	0.82	0.82	-0.035	0.22	0.75	1	0.68	0.16	0.19	0.81	0.55	0.032
star	0.73	0.093	0.039	0.76	0.75	-0.0099	0.18	0.68	0.68	1	0.18	0.11	0.74	0.53	0.035
country	0.14	0.00049	-0.016	0.14	0.15	0.023	-0.046	0.16	0.16	0.18	1	-0.082	0.13	0.23	0.12
budget	0.28	0.19	0.073	0.3	0.29	-0.013	0.4	0.11	0.19	0.11	-0.082	1	0.29	-0.092	0.11
gross	0.95	0.16	0.039	0.98	0.98	-0.047	0.29	0.75	0.81	0.74	0.13	0.29	1	0.59	0.043
company	0.59	-0.028	0.0096	0.6	0.61	-0.028	0.0089	0.55	0.55	0.53	0.23	-0.092	0.59	1	0.0051
runtime	0.049	0.033	0.0015	0.051	0.048	0.026	0.11	-0.011	0.032	0.035	0.12	0.11	0.043	0.0051	1

```
correlation_mat = df.apply(lambda x: x.factorize()[0]).corr()

corr_pairs = correlation_mat.unstack()

print(corr_pairs)
```

```
name     name        1.000000
         rating      0.143938
         genre       0.036367
         year        0.965761
         released    0.959015
                       ...
runtime  country     0.124154
         budget      0.112097
         gross       0.042978
         company     0.005137
         runtime     1.000000
Length: 225, dtype: float64
```

```
sorted_pairs = corr_pairs.sort_values(kind="quicksort")

print(sorted_pairs)
```

```
budget   company   -0.092249
company  budget    -0.092249
genre    rating    -0.086723
rating   genre     -0.086723
budget   country   -0.082082
                      ...
year     year       1.000000
genre    genre      1.000000
rating   rating     1.000000
company  company    1.000000
runtime  runtime    1.000000
Length: 225, dtype: float64
```

```
strong_pairs = sorted_pairs[abs(sorted_pairs) > 0.5]

print(strong_pairs)
```

```
star      company    0.527116
company   star       0.527116
          writer     0.546151
writer    company    0.546151
director  company    0.552258
                       ...
year      year       1.000000
genre     genre      1.000000
rating    rating     1.000000
company   company    1.000000
runtime   runtime    1.000000
Length: 71, dtype: float64
```

```
CompanyGrossSum = df.groupby('company')[["gross"]].sum()

CompanyGrossSumSorted = CompanyGrossSum.sort_values('gross', ascending
= False)[:15]

CompanyGrossSumSorted = CompanyGrossSumSorted['gross'].astype('int64')


CompanyGrossSumSorted
```

```
company
Warner Bros.                 56491421806
Universal Pictures           52514188890
Columbia Pictures            43008941346
Paramount Pictures           40493607415
Twentieth Century Fox        40257053857
```

```
Walt Disney Pictures         36327887792
New Line Cinema              19883797684
Marvel Studios               15065592411
DreamWorks Animation         11873612858
Touchstone Pictures          11795832638
Dreamworks Pictures          11635441081
Metro-Goldwyn-Mayer (MGM)     9230230105
Summit Entertainment          8373718838
Pixar Animation Studios       7886344526
Fox 2000 Pictures             7443502667
Name: gross, dtype: int64
```

```
df['Year'] = df['released'].astype(str).str[:4]
df
```

```
                                                name rating      genre
year  \
0                                        The Shining      R      Drama
1980
1                                    The Blue Lagoon      R  Adventure
1980
2     Star Wars: Episode V - The Empire Strikes Back     PG     Action
1980
3                                          Airplane!     PG     Comedy
1980
4                                         Caddyshack      R     Comedy
1980
...                                              ...    ...        ...
...
7663                                    More to Life    NaN      Drama
2020
7664                                     Dream Round    NaN     Comedy
2020
7665                                   Saving Mbango    NaN      Drama
2020
7666                                    It's Just Us    NaN      Drama
2020
7667                                       Tee em el    NaN     Horror
2020

                              released  score      votes
director  \
0         June 13, 1980 (United States)    8.4   927000.0  Stanley
Kubrick
1          July 2, 1980 (United States)    5.8    65000.0   Randal
Kleiser
2         June 20, 1980 (United States)    8.7  1200000.0   Irvin
Kershner
3          July 2, 1980 (United States)    7.7   221000.0     Jim
Abrahams
```

```
4        July 25, 1980 (United States)    7.3   108000.0     Harold 
Ramis  
...                                ...    ...        ...           
...  
7663  October 23, 2020 (United States)    3.1       18.0    Joseph 
Ebanks  
7664  February 7, 2020 (United States)    4.7       36.0     Dusty 
Dukatz  
7665         April 27, 2020 (Cameroon)    5.7       29.0     Nkanya 
Nkwai  
7666   October 1, 2020 (United States)    NaN        NaN    James 
Randall  
7667   August 19, 2020 (United States)    5.7        7.0     Pereko 
Mosia  

                       writer              star         country  
budget  \
0                Stephen King    Jack Nicholson  United Kingdom  
19000000.0   
1     Henry De Vere Stacpoole    Brooke Shields   United States   
4500000.0   
2              Leigh Brackett       Mark Hamill   United States  
18000000.0   
3                Jim Abrahams       Robert Hays   United States   
3500000.0   
4          Brian Doyle-Murray       Chevy Chase   United States   
6000000.0   
...                       ...               ...             ...          
...   
7663            Joseph Ebanks      Shannon Bond   United States      
7000.0   
7664              Lisa Huston  Michael Saquella   United States          
NaN   
7665             Lynno Lovert      Onyama Laura   United States     
58750.0   
7666            James Randall     Christina Roz   United States     
15000.0   
7667             Pereko Mosia  Siyabonga Mabaso    South Africa          
NaN   

            gross                    company  runtime  Year  
0      46998772.0               Warner Bros.    146.0  June  
1      58853106.0          Columbia Pictures    104.0  July  
2     538375067.0                  Lucasfilm    124.0  June  
3      83453539.0         Paramount Pictures     88.0  July  
4      39846344.0             Orion Pictures     98.0  July  
...           ...                        ...      ...   ...  
7663          NaN                        NaN     90.0  Octo  
7664          NaN  Cactus Blue Entertainment     90.0  Febr  
```

```
7665          NaN           Embi Productions      NaN  Apri
7666          NaN                        NaN    120.0  Octo
7667          NaN                PK 65 Films    102.0  Augu

[7668 rows x 16 columns]
```

```
df.groupby(['company', 'year'])[["gross"]].sum()
```

```
                                                   gross
company                                    year
"DIA" Productions GmbH & Co. KG            2003   44350926.0
"Weathering With You" Film Partners        2019  193457467.0
.406 Production                            1996      10580.0
1+2 Seisaku Iinkai                         2000    1196218.0
10 West Studios                            2010     814906.0
...                                                      ...
i am OTHER                                 2015   17986781.0
i5 Films                                   2001   10031529.0
iDeal Partners Film Fund                   2013     506303.0
micro_scope                                2010    7099598.0
thefyzz                                    2017   62198461.0

[4536 rows x 1 columns]
```

```
CompanyGrossSum = df.groupby(['company', 'year'])[["gross"]].sum()

CompanyGrossSumSorted =
CompanyGrossSum.sort_values(['gross','company','year'], ascending =
False)[:15]

CompanyGrossSumSorted = CompanyGrossSumSorted['gross'].astype('int64')


CompanyGrossSumSorted
```

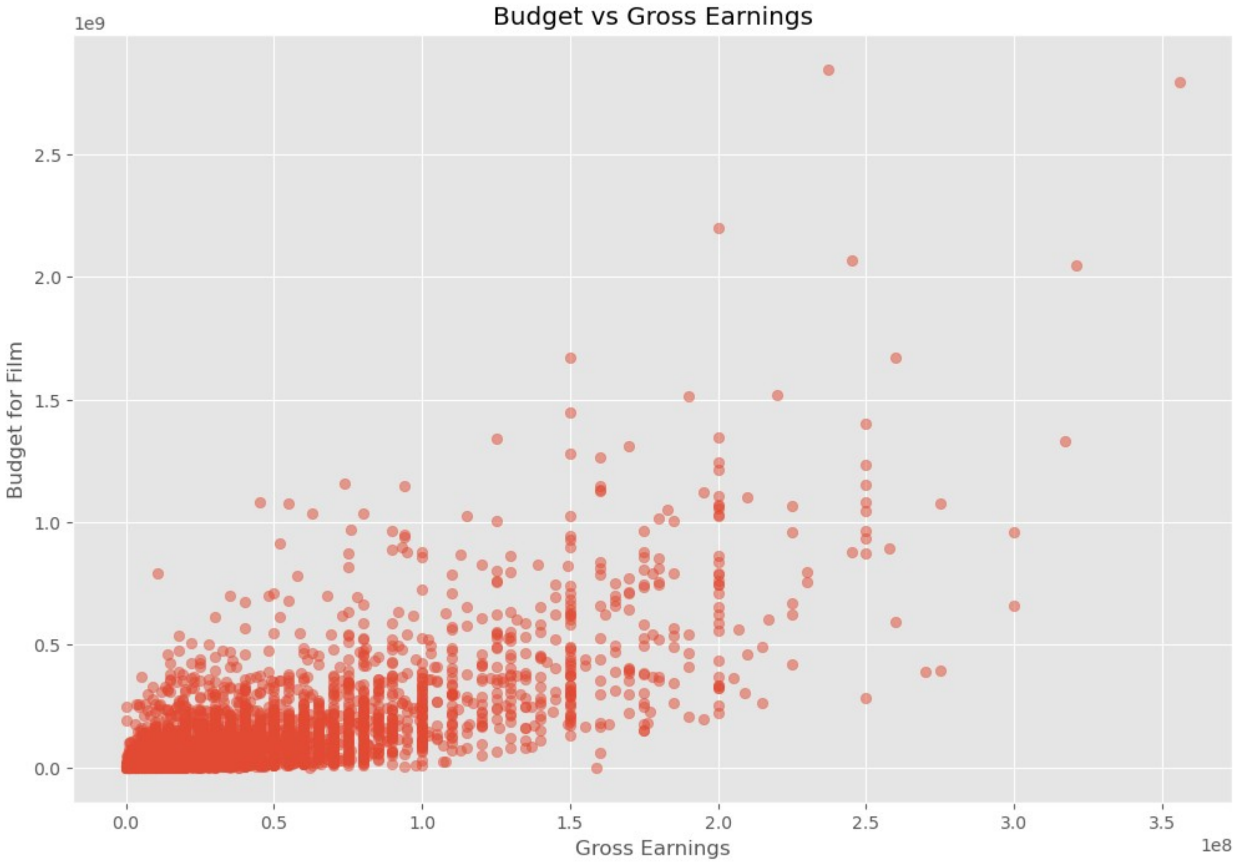
```
company                year
Walt Disney Pictures   2019    5773131804
Marvel Studios         2018    4018631866
Universal Pictures     2015    3834354888
Twentieth Century Fox  2009    3793491246
Walt Disney Pictures   2017    3789382071
Paramount Pictures     2011    3565705182
Warner Bros.           2010    3300479986
                       2011    3223799224
Walt Disney Pictures   2010    3104474158
Paramount Pictures     2014    3071298586
Columbia Pictures      2006    2934631933
                       2019    2932757449
Marvel Studios         2019    2797501328
Warner Bros.           2018    2774168962
```

```
Columbia Pictures      2011    2738363306
Name: gross, dtype: int64
```

```
CompanyGrossSum = df.groupby(['company'])[["gross"]].sum()

CompanyGrossSumSorted =
CompanyGrossSum.sort_values(['gross','company'], ascending = False)
[:15]

CompanyGrossSumSorted = CompanyGrossSumSorted['gross'].astype('int64')


CompanyGrossSumSorted
```

```
company
Warner Bros.                 56491421806
Universal Pictures           52514188890
Columbia Pictures            43008941346
Paramount Pictures           40493607415
Twentieth Century Fox        40257053857
Walt Disney Pictures         36327887792
New Line Cinema              19883797684
Marvel Studios               15065592411
DreamWorks Animation         11873612858
Touchstone Pictures          11795832638
Dreamworks Pictures          11635441081
Metro-Goldwyn-Mayer (MGM)     9230230105
Summit Entertainment          8373718838
Pixar Animation Studios       7886344526
Fox 2000 Pictures             7443502667
Name: gross, dtype: int64
```

```
plt.scatter(x=df['budget'], y=df['gross'], alpha=0.5)
plt.title('Budget vs Gross Earnings')
plt.xlabel('Gross Earnings')
plt.ylabel('Budget for Film')
plt.show()
```



```
df
```

```
                                                 name rating      genre
year  \
0                                         The Shining      R      Drama
1980
1                                     The Blue Lagoon      R  Adventure
1980
2     Star Wars: Episode V - The Empire Strikes Back     PG     Action
1980
3                                           Airplane!     PG     Comedy
1980
4                                          Caddyshack      R     Comedy
1980
...                                               ...    ...        ...
...
7663                                     More to Life    NaN      Drama
2020
7664                                      Dream Round    NaN     Comedy
2020
7665                                    Saving Mbango    NaN      Drama
2020
7666                                     It's Just Us    NaN      Drama
```

```
2020
7667                                       Tee em el    NaN     Horror
2020

                              released  score      votes
director  \
0        June 13, 1980 (United States)    8.4   927000.0  Stanley
Kubrick
1         July 2, 1980 (United States)    5.8    65000.0   Randal
Kleiser
2        June 20, 1980 (United States)    8.7  1200000.0   Irvin
Kershner
3         July 2, 1980 (United States)    7.7   221000.0     Jim
Abrahams
4        July 25, 1980 (United States)    7.3   108000.0     Harold
Ramis
...                                ...    ...        ...
...
7663  October 23, 2020 (United States)    3.1       18.0    Joseph
Ebanks
7664  February 7, 2020 (United States)    4.7       36.0     Dusty
Dukatz
7665         April 27, 2020 (Cameroon)    5.7       29.0     Nkanya
Nkwai
7666   October 1, 2020 (United States)    NaN        NaN    James
Randall
7667   August 19, 2020 (United States)    5.7        7.0     Pereko
Mosia

                       writer              star         country
budget  \
0                Stephen King    Jack Nicholson  United Kingdom
19000000.0
1     Henry De Vere Stacpoole    Brooke Shields   United States
4500000.0
2              Leigh Brackett       Mark Hamill   United States
18000000.0
3                Jim Abrahams       Robert Hays   United States
3500000.0
4          Brian Doyle-Murray       Chevy Chase   United States
6000000.0
...                       ...               ...             ...
...
7663            Joseph Ebanks      Shannon Bond   United States
7000.0
7664              Lisa Huston  Michael Saquella   United States
NaN
7665             Lynno Lovert      Onyama Laura   United States
58750.0
```

```
7666            James Randall     Christina Roz   United States
15000.0
7667             Pereko Mosia  Siyabonga Mabaso    South Africa
NaN

            gross                    company  runtime  Year
0      46998772.0               Warner Bros.    146.0  June
1      58853106.0          Columbia Pictures    104.0  July
2     538375067.0                  Lucasfilm    124.0  June
3      83453539.0         Paramount Pictures     88.0  July
4      39846344.0             Orion Pictures     98.0  July
...           ...                        ...      ...   ...
7663          NaN                        NaN     90.0  Octo
7664          NaN  Cactus Blue Entertainment     90.0  Febr
7665          NaN           Embi Productions      NaN  Apri
7666          NaN                        NaN    120.0  Octo
7667          NaN                PK 65 Films    102.0  Augu

[7668 rows x 16 columns]
```

```
df_numerized = df


for col_name in df_numerized.columns:
    if(df_numerized[col_name].dtype == 'object'):
        df_numerized[col_name]=
df_numerized[col_name].astype('category')
        df_numerized[col_name] = df_numerized[col_name].cat.codes

df_numerized
```

```
      name  rating  genre  year  released  score      votes  director
writer  \
0     6587       6      6  1980      1705    8.4   927000.0      2589
4014
1     5573       6      1  1980      1492    5.8    65000.0      2269
1632
2     5142       4      0  1980      1771    8.7  1200000.0      1111
2567
3      286       4      4  1980      1492    7.7   221000.0      1301
2000
4     1027       6      4  1980      1543    7.3   108000.0      1054
521
...    ...     ...    ...   ...       ...    ...        ...       ...
...
7663  3705      -1      6  2020      2964    3.1       18.0      1500
2289
7664  1678      -1      4  2020      1107    4.7       36.0       774
2614
7665  4717      -1      6  2020       193    5.7       29.0      2061
```

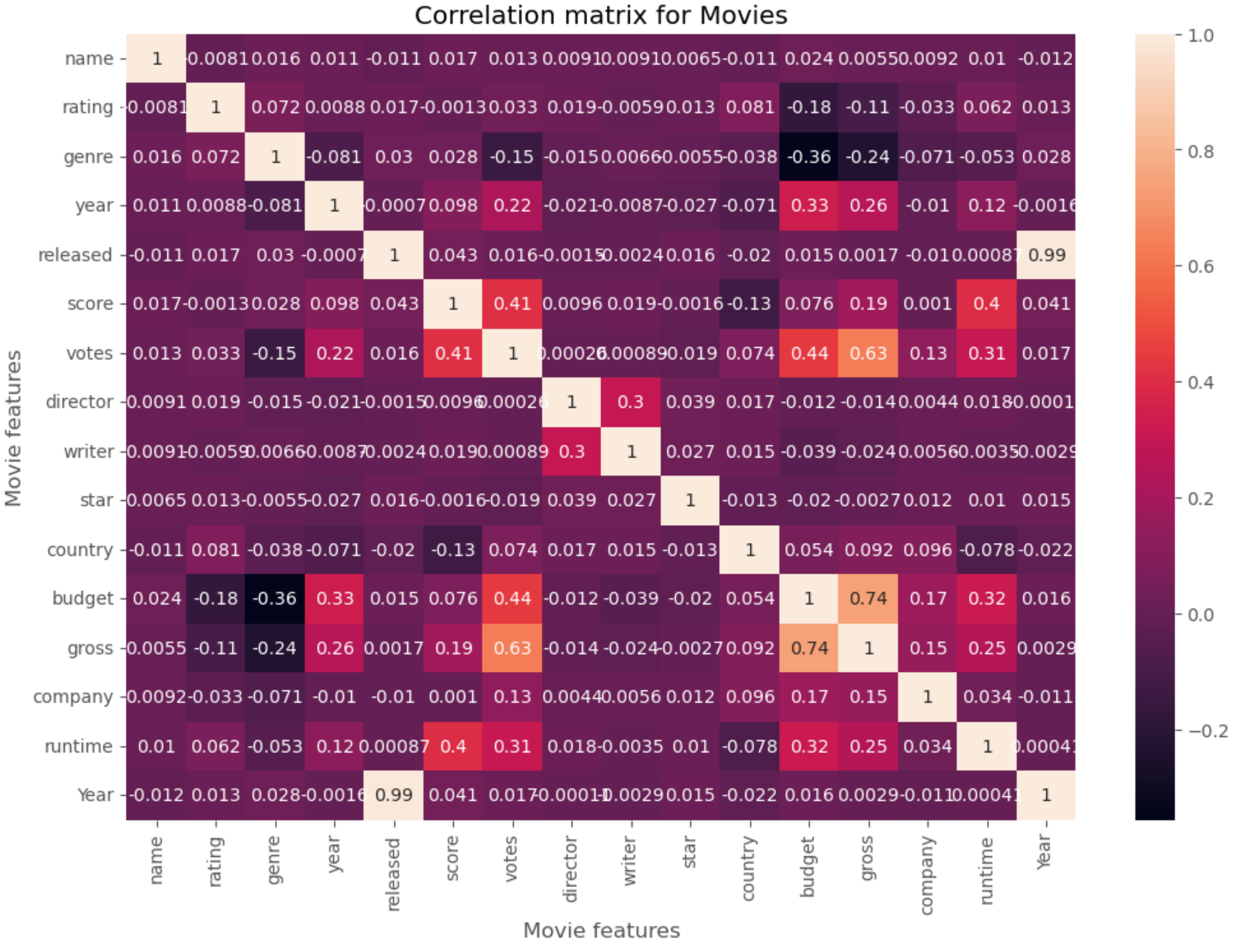
```
2683
7666  2843      -1      6  2020      2817    NaN        NaN      1184
1824
7667  5394      -1     10  2020       391    5.7        7.0      2165
3344

      star  country       budget        gross  company  runtime  Year
0     1047       54  19000000.0   46998772.0     2319    146.0    14
1      327       55   4500000.0   58853106.0      731    104.0    13
2     1745       55  18000000.0  538375067.0     1540    124.0    14
3     2246       55   3500000.0   83453539.0     1812     88.0    13
4      410       55   6000000.0   39846344.0     1777     98.0    13
...    ...      ...         ...          ...      ...      ...   ...
7663  2421       55      7000.0          NaN       -1     90.0    18
7664  1886       55         NaN          NaN      539     90.0    11
7665  2040       55     58750.0          NaN      941      NaN     8
7666   450       55     15000.0          NaN       -1    120.0    18
7667  2463       44         NaN          NaN     1787    102.0     9

[7668 rows x 16 columns]
```

```
df_numerized.corr(method='pearson')
```

```
              name    rating     genre      year  released
score  \
name      1.000000 -0.008069  0.016355  0.011453 -0.011311  0.017097

rating   -0.008069  1.000000  0.072423  0.008779  0.016613 -0.001314

genre     0.016355  0.072423  1.000000 -0.081261  0.029822  0.027965

year      0.011453  0.008779 -0.081261  1.000000 -0.000695  0.097995

released -0.011311  0.016613  0.029822 -0.000695  1.000000  0.042788

score     0.017097 -0.001314  0.027965  0.097995  0.042788  1.000000

votes     0.013088  0.033225 -0.145307  0.222945  0.016097  0.409182

director  0.009079  0.019483 -0.015258 -0.020795 -0.001478  0.009559

writer    0.009081 -0.005921  0.006567 -0.008656 -0.002404  0.019416

star      0.006472  0.013405 -0.005477 -0.027242  0.015777 -0.001609

country  -0.010737  0.081244 -0.037615 -0.070938 -0.020427 -0.133348

budget    0.023970 -0.176002 -0.356564  0.329321  0.014683  0.076254

gross     0.005533 -0.107339 -0.235650  0.257486  0.001659  0.186258
```

```
company   0.009211 -0.032943 -0.071067 -0.010431 -0.010474  0.001030

runtime   0.010392  0.062145 -0.052711  0.120811  0.000868  0.399451

Year     -0.011725  0.013475  0.028397 -0.001562  0.993694  0.040993


             votes  director    writer      star   country
budget  \
name      0.013088  0.009079  0.009081  0.006472 -0.010737  0.023970

rating    0.033225  0.019483 -0.005921  0.013405  0.081244 -0.176002

genre    -0.145307 -0.015258  0.006567 -0.005477 -0.037615 -0.356564

year      0.222945 -0.020795 -0.008656 -0.027242 -0.070938  0.329321

released  0.016097 -0.001478 -0.002404  0.015777 -0.020427  0.014683

score     0.409182  0.009559  0.019416 -0.001609 -0.133348  0.076254

votes     1.000000  0.000260  0.000892 -0.019282  0.073625  0.442429

director  0.000260  1.000000  0.299067  0.039234  0.017490 -0.012272

writer    0.000892  0.299067  1.000000  0.027245  0.015343 -0.039451

star     -0.019282  0.039234  0.027245  1.000000 -0.012998 -0.019589

country   0.073625  0.017490  0.015343 -0.012998  1.000000  0.054063

budget    0.442429 -0.012272 -0.039451 -0.019589  0.054063  1.000000

gross     0.630757 -0.014441 -0.023519 -0.002717  0.092129  0.740395

company   0.133204  0.004404  0.005646  0.012442  0.095548  0.173214

runtime   0.309212  0.017624 -0.003511  0.010174 -0.078412  0.320447

Year      0.017337 -0.000105 -0.002892  0.015406 -0.022277  0.015682


             gross   company   runtime      Year
name      0.005533  0.009211  0.010392 -0.011725
rating   -0.107339 -0.032943  0.062145  0.013475
genre    -0.235650 -0.071067 -0.052711  0.028397
year      0.257486 -0.010431  0.120811 -0.001562
released  0.001659 -0.010474  0.000868  0.993694
score     0.186258  0.001030  0.399451  0.040993
votes     0.630757  0.133204  0.309212  0.017337
director -0.014441  0.004404  0.017624 -0.000105
```

```
writer   -0.023519  0.005646 -0.003511 -0.002892
star     -0.002717  0.012442  0.010174  0.015406
country   0.092129  0.095548 -0.078412 -0.022277
budget    0.740395  0.173214  0.320447  0.015682
gross     1.000000  0.154840  0.245216  0.002946
company   0.154840  1.000000  0.034402 -0.010726
runtime   0.245216  0.034402  1.000000  0.000410
Year      0.002946 -0.010726  0.000410  1.000000

correlation_matrix = df_numerized.corr(method='pearson')

sns.heatmap(correlation_matrix, annot = True)

plt.title("Correlation matrix for Movies")

plt.xlabel("Movie features")

plt.ylabel("Movie features")

plt.show()
```



	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime	Year
name	1	-0.0081	0.016	0.011	-0.011	0.017	0.013	0.0091	0.0091	0.0065	-0.011	0.024	0.0055	0.0092	0.01	-0.012
rating	-0.0081	1	0.072	0.0088	0.017	-0.0013	0.033	0.019	-0.0059	0.013	0.081	-0.18	-0.11	-0.033	0.062	0.013
genre	0.016	0.072	1	-0.081	0.03	0.028	-0.15	-0.015	0.0066	-0.0055	-0.038	-0.36	-0.24	-0.071	-0.053	0.028
year	0.011	0.0088	-0.081	1	-0.0007	0.098	0.22	-0.021	-0.0087	-0.027	-0.071	0.33	0.26	-0.01	0.12	-0.0016
released	-0.011	0.017	0.03	-0.0007	1	0.043	0.016	-0.0015	0.0024	0.016	-0.02	0.015	0.0017	-0.01	0.00087	0.99
score	0.017	-0.0013	0.028	0.098	0.043	1	0.41	0.0096	0.019	-0.0016	-0.13	0.076	0.19	0.001	0.4	0.041
votes	0.013	0.033	-0.15	0.22	0.016	0.41	1	0.00026	0.00089	-0.019	0.074	0.44	0.63	0.13	0.31	0.017
director	0.0091	0.019	-0.015	-0.021	-0.0015	0.0096	0.00026	1	0.3	0.039	0.017	-0.012	-0.014	0.0044	0.018	-0.0001
writer	0.0091	-0.0059	0.0066	-0.0087	0.0024	0.019	0.00089	0.3	1	0.027	0.015	-0.039	-0.024	0.0056	-0.0035	-0.0029
star	0.0065	0.013	-0.0055	-0.027	0.016	-0.0016	-0.019	0.039	0.027	1	-0.013	-0.02	-0.0027	0.012	0.01	0.015
country	-0.011	0.081	-0.038	-0.071	-0.02	-0.13	0.074	0.017	0.015	-0.013	1	0.054	0.092	0.096	-0.078	-0.022
budget	0.024	-0.18	-0.36	0.33	0.015	0.076	0.44	-0.012	-0.039	-0.02	0.054	1	0.74	0.17	0.32	0.016
gross	0.0055	-0.11	-0.24	0.26	0.0017	0.19	0.63	-0.014	-0.024	-0.0027	0.092	0.74	1	0.15	0.25	0.0029
company	0.0092	-0.033	-0.071	-0.01	-0.01	0.001	0.13	0.0044	0.0056	0.012	0.096	0.17	0.15	1	0.034	-0.011
runtime	0.01	0.062	-0.053	0.12	0.00087	0.4	0.31	0.018	-0.0035	0.01	-0.078	0.32	0.25	0.034	1	0.00041
Year	-0.012	0.013	0.028	-0.0016	0.99	0.041	0.017	-0.0001	-0.0029	0.015	-0.022	0.016	0.0029	-0.011	0.00041	1

```
import pandas as pd
import numpy as np
import seaborn as sns

import matplotlib.pyplot as plt
import matplotlib.mlab as mlab
import matplotlib
plt.style.use('ggplot')
from matplotlib.pyplot import figure

%matplotlib inline
matplotlib.rcParams['figure.figsize'] = (12,8)

pd.options.mode.chained_assignment = None



# Now we need to read in the data
df = pd.read_csv(r'C:\Users\Hi\Downloads\movies.csv')

df

                                                name rating      genre
year  \
0                                        The Shining      R      Drama
1980
1                                    The Blue Lagoon      R  Adventure
1980
2     Star Wars: Episode V - The Empire Strikes Back     PG     Action
1980
3                                          Airplane!     PG     Comedy
1980
4                                         Caddyshack      R     Comedy
1980
...                                              ...    ...        ...
...
7663                                    More to Life    NaN      Drama
```

```
2020
7664                                     Dream Round    NaN     Comedy
2020
7665                                   Saving Mbango    NaN      Drama
2020
7666                                    It's Just Us    NaN      Drama
2020
7667                                       Tee em el    NaN     Horror
2020

                              released  score      votes
director  \
0        June 13, 1980 (United States)    8.4   927000.0  Stanley
Kubrick
1         July 2, 1980 (United States)    5.8    65000.0   Randal
Kleiser
2        June 20, 1980 (United States)    8.7  1200000.0   Irvin
Kershner
3         July 2, 1980 (United States)    7.7   221000.0     Jim
Abrahams
4        July 25, 1980 (United States)    7.3   108000.0     Harold
Ramis
...                                ...    ...        ...
...
7663  October 23, 2020 (United States)    3.1       18.0    Joseph
Ebanks
7664  February 7, 2020 (United States)    4.7       36.0     Dusty
Dukatz
7665         April 27, 2020 (Cameroon)    5.7       29.0     Nkanya
Nkwai
7666   October 1, 2020 (United States)    NaN        NaN    James
Randall
7667   August 19, 2020 (United States)    5.7        7.0     Pereko
Mosia

                       writer               star         country
budget  \
0                Stephen King    Jack Nicholson  United Kingdom
19000000.0
1     Henry De Vere Stacpoole    Brooke Shields   United States
4500000.0
2              Leigh Brackett       Mark Hamill   United States
18000000.0
3                Jim Abrahams       Robert Hays   United States
3500000.0
4          Brian Doyle-Murray       Chevy Chase   United States
6000000.0
...                       ...               ...             ...
...
```

```
7663             Joseph Ebanks      Shannon Bond   United States
7000.0
7664               Lisa Huston  Michael Saquella   United States
NaN
7665              Lynno Lovert      Onyama Laura   United States
58750.0
7666             James Randall     Christina Roz   United States
15000.0
7667              Pereko Mosia  Siyabonga Mabaso    South Africa
NaN

            gross                     company  runtime
0      46998772.0                Warner Bros.    146.0
1      58853106.0           Columbia Pictures    104.0
2     538375067.0                   Lucasfilm    124.0
3      83453539.0          Paramount Pictures     88.0
4      39846344.0             Orion Pictures     98.0
...           ...                         ...      ...
7663          NaN                         NaN     90.0
7664          NaN  Cactus Blue Entertainment     90.0
7665          NaN            Embi Productions      NaN
7666          NaN                         NaN    120.0
7667          NaN                 PK 65 Films    102.0

[7668 rows x 15 columns]
```

```python
for col_name in df.columns:
    if(df[col_name].dtype == 'object'):
        df[col_name]= df[col_name].astype('category')
        df[col_name] = df[col_name].cat.codes

df
```

```
      name  rating  genre  year  released  score      votes  director
writer  \
0     6587       6      6  1980      1705    8.4   927000.0      2589
4014
1     5573       6      1  1980      1492    5.8    65000.0      2269
1632
2     5142       4      0  1980      1771    8.7  1200000.0      1111
2567
3      286       4      4  1980      1492    7.7   221000.0      1301
2000
4     1027       6      4  1980      1543    7.3   108000.0      1054
521
...    ...     ...    ...   ...       ...    ...        ...       ...
...
7663  3705      -1      6  2020      2964    3.1       18.0      1500
2289
7664  1678      -1      4  2020      1107    4.7       36.0       774
```

```
2614
7665  4717      -1      6  2020       193    5.7         29.0       2061
2683
7666  2843      -1      6  2020      2817    NaN          NaN       1184
1824
7667  5394      -1     10  2020       391    5.7          7.0       2165
3344

      star  country       budget        gross  company  runtime
0     1047       54   19000000.0   46998772.0     2319    146.0
1      327       55    4500000.0   58853106.0      731    104.0
2     1745       55   18000000.0  538375067.0     1540    124.0
3     2246       55    3500000.0   83453539.0     1812     88.0
4      410       55    6000000.0   39846344.0     1777     98.0
...    ...      ...          ...          ...      ...      ...
7663  2421       55       7000.0          NaN       -1     90.0
7664  1886       55          NaN          NaN      539     90.0
7665  2040       55      58750.0          NaN      941      NaN
7666   450       55      15000.0          NaN       -1    120.0
7667  2463       44          NaN          NaN     1787    102.0

[7668 rows x 15 columns]
```

```
cat_columns = df.select_dtypes(include='object').columns
df[cat_columns] = df[cat_columns].apply(lambda x:
x.astype('category').cat.codes)

df
```

```
      name  rating  genre  year  released  score        votes  director
writer  \
0     6587       6      6  1980      1705    8.4     927000.0       2589
4014
1     5573       6      1  1980      1492    5.8      65000.0       2269
1632
2     5142       4      0  1980      1771    8.7    1200000.0       1111
2567
3      286       4      4  1980      1492    7.7     221000.0       1301
2000
4     1027       6      4  1980      1543    7.3     108000.0       1054
521
...    ...     ...    ...   ...       ...    ...          ...        ...
...
7663  3705      -1      6  2020      2964    3.1         18.0       1500
2289
7664  1678      -1      4  2020      1107    4.7         36.0        774
```

```
2614
7665  4717       -1       6  2020        193    5.7        29.0       2061
2683
7666  2843       -1       6  2020       2817    NaN         NaN       1184
1824
7667  5394       -1      10  2020        391    5.7         7.0       2165
3344

      star  country      budget        gross  company  runtime
0     1047       54  19000000.0   46998772.0     2319    146.0
1      327       55   4500000.0   58853106.0      731    104.0
2     1745       55  18000000.0  538375067.0     1540    124.0
3     2246       55   3500000.0   83453539.0     1812     88.0
4      410       55   6000000.0   39846344.0     1777     98.0
...    ...      ...         ...          ...      ...      ...
7663  2421       55      7000.0          NaN       -1     90.0
7664  1886       55         NaN          NaN      539     90.0
7665  2040       55     58750.0          NaN      941      NaN
7666   450       55     15000.0          NaN       -1    120.0
7667  2463       44         NaN          NaN     1787    102.0

[7668 rows x 15 columns]
```