# correlation w pandas

May 30, 2024

Movie Industry

[7]:
```
# First let's import the packages we will use in this project
# You can do this all now or as you need them
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('movies.csv')
```

[37]:
```
df
```

[37]:

| | 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]
```

[39]:
```python
# We need to see if we have any missing data
# Let's loop through the data and see if there is anything missing

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%
```

[41]:
```
# We need to see if we have any missing data
# Let's loop through the data and see if there is anything missing

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

[41]:
```
name           0
rating        77
genre          0
year           0
released       2
score          3
votes          3
director       0
writer         3
star           1
country        3
budget      2171
gross        189
company       17
runtime        4
dtype: int64
```

[43]:
```
# We need to see if we have any missing data
# Let's loop through the data and see if there is anything missing


missing_data = df.isnull()
missing_data.head(5)
```

```
for column in missing_data.columns.values.tolist():
    print(column)
    print (missing_data[column].value_counts())
    print("")
```

```
name
name
False    7668
Name: count, dtype: int64

rating
rating
False    7591
True       77
Name: count, dtype: int64

genre
genre
False    7668
Name: count, dtype: int64

year
year
False    7668
Name: count, dtype: int64

released
released
False    7666
True        2
Name: count, dtype: int64

score
score
False    7665
True        3
Name: count, dtype: int64

votes
votes
False    7665
True        3
Name: count, dtype: int64

director
director
```

```
False    7668
Name: count, dtype: int64

writer
writer
False    7665
True        3
Name: count, dtype: int64

star
star
False    7667
True        1
Name: count, dtype: int64

country
country
False    7665
True        3
Name: count, dtype: int64

budget
budget
False    5497
True     2171
Name: count, dtype: int64

gross
gross
False    7479
True      189
Name: count, dtype: int64

company
company
False    7651
True       17
Name: count, dtype: int64

runtime
runtime
False    7664
True        4
Name: count, dtype: int64
```

[68]:
```
#drop all rows with null values because we can t replace them

df = df.dropna()
```

[70]:
```
df
```

[70]:

```
                                          name     rating      genre  \
0                                The Shining          R      Drama
1                            The Blue Lagoon          R  Adventure
2     Star Wars: Episode V - The Empire Strikes Back         PG     Action
3                                   Airplane!         PG     Comedy
4                                   Caddyshack          R     Comedy
…                                          …        …         …
7652                    The Eight Hundred  Not Rated     Action
7653                           The Quarry          R      Crime
7656                                Tulsa      PG-13     Comedy
7658          Black Wall Street Burning          R      Drama
7659                            I Am Fear  Not Rated     Horror

      year                         released  score      votes  \
0     1980     June 13, 1980 (United States)    8.4   927000.0
1     1980      July 2, 1980 (United States)    5.8    65000.0
2     1980     June 20, 1980 (United States)    8.7  1200000.0
3     1980      July 2, 1980 (United States)    7.7   221000.0
4     1980     July 25, 1980 (United States)    7.3   108000.0
…     …                                  …    …        …
7652  2020   August 28, 2020 (United States)    6.8     3700.0
7653  2020           April 17, 2020 (Mexico)    5.4     2400.0
7656  2020      June 3, 2020 (United States)    5.0      294.0
7658  2020  February 7, 2020 (United States)    6.6       35.0
7659  2020     March 3, 2020 (United States)    3.4      447.0

             director                  writer             star  \
0     Stanley Kubrick             Stephen King   Jack Nicholson
1      Randal Kleiser  Henry De Vere Stacpoole   Brooke Shields
2      Irvin Kershner           Leigh Brackett      Mark Hamill
3        Jim Abrahams             Jim Abrahams      Robert Hays
4        Harold Ramis       Brian Doyle-Murray      Chevy Chase
…                 …                        …                …
7652          Hu Guan                  Hu Guan  Zhi-zhong Huang
7653      Scott Teems              Scott Teems     Shea Whigham
7656      Scott Pryor              Scott Pryor      Scott Pryor
7658    Marcus Brown          Dekoven Riggins      Dan Belcher
7659    Kevin Shulman            Kevin Shulman   Kristina Klebe

             country    budget      gross  \
0     United Kingdom  19000000   46998772
```

```
1      United States   4500000   58853106
2      United States  18000000  538375067
3      United States   3500000   83453539
4      United States   6000000   39846344
…                …         …         …
7652           China  80000000  461421559
7653   United States         0       3661
7656   United States         0     413378
7658   United States      5000          0
7659   United States         0      13266

                                company  runtime
0                          Warner Bros.    146.0
1                       Columbia Pictures  104.0
2                               Lucasfilm  124.0
3                      Paramount Pictures   88.0
4                          Orion Pictures   98.0
…                                     …      …
7652  Beijing Diqi Yinxiang Entertainment  149.0
7653                     Prowess Pictures   98.0
7656                  Pryor Entertainment  120.0
7658                        Notis Studio   78.0
7659                        Roxwell Films   87.0

[7574 rows x 15 columns]
```

[72]:
```
#checking again to see. Everything should be zero
df.isnull().sum()
```

[72]:
```
name        0
rating      0
genre       0
year        0
released    0
score       0
votes       0
director    0
writer      0
star        0
country     0
budget      0
gross       0
company     0
runtime     0
dtype: int64
```

[45]:
```
# Data Types for our columns

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
```

[47]:
```
#change data type of columns

# df['budget'] = df['budget'].astype('int64') this ussually works but it didnt,␣
 ↪it brought anerror

# df['gross'] = df['gross'].astype('int64') this ussually works but it didnt␣
 ↪for some reason, it brought an error

# If anyone else is having issues due to IntCastingNanError, I advise to try␣
 ↪the following:

df['budget'] = pd.to_numeric(df['budget'], errors='coerce').fillna(0).
 ↪astype(int)
df['gross'] = pd.to_numeric(df['gross'], errors='coerce').fillna(0).astype(int)
```

[49]:
```
df
```

[49]:

| | 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
1     Henry De Vere Stacpoole    Brooke Shields   United States   4500000
2              Leigh Brackett       Mark Hamill   United States  18000000
3                Jim Abrahams       Robert Hays   United States   3500000
4           Brian Doyle-Murray       Chevy Chase   United States   6000000
…                          …                 …               …       …
7663            Joseph Ebanks      Shannon Bond   United States      7000
7664              Lisa Huston  Michael Saquella   United States         0
7665             Lynno Lovert      Onyama Laura   United States     58750
7666            James Randall     Christina Roz   United States     15000
7667             Pereko Mosia  Siyabonga Mabaso    South Africa         0

          gross                    company  runtime
0      46998772               Warner Bros.    146.0
1      58853106          Columbia Pictures    104.0
2     538375067                  Lucasfilm    124.0
3      83453539         Paramount Pictures     88.0
4      39846344             Orion Pictures     98.0
…          …                          …      …
7663          0                        NaN     90.0
7664          0  Cactus Blue Entertainment     90.0
7665          0           Embi Productions      NaN
7666          0                        NaN    120.0
7667          0                PK 65 Films    102.0

[7668 rows x 15 columns]
```

[59]:
```
# Order our Data a little bit to see
```

```
#just checking to see which movie makes the most money, that will be interms of␣
 ↪gross
#gross is the only word in this data that means revenue
# We dont want to save the df like this so inplace=false
#if you put ascending= true youll see the movie that made the least amount of␣
 ↪money

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

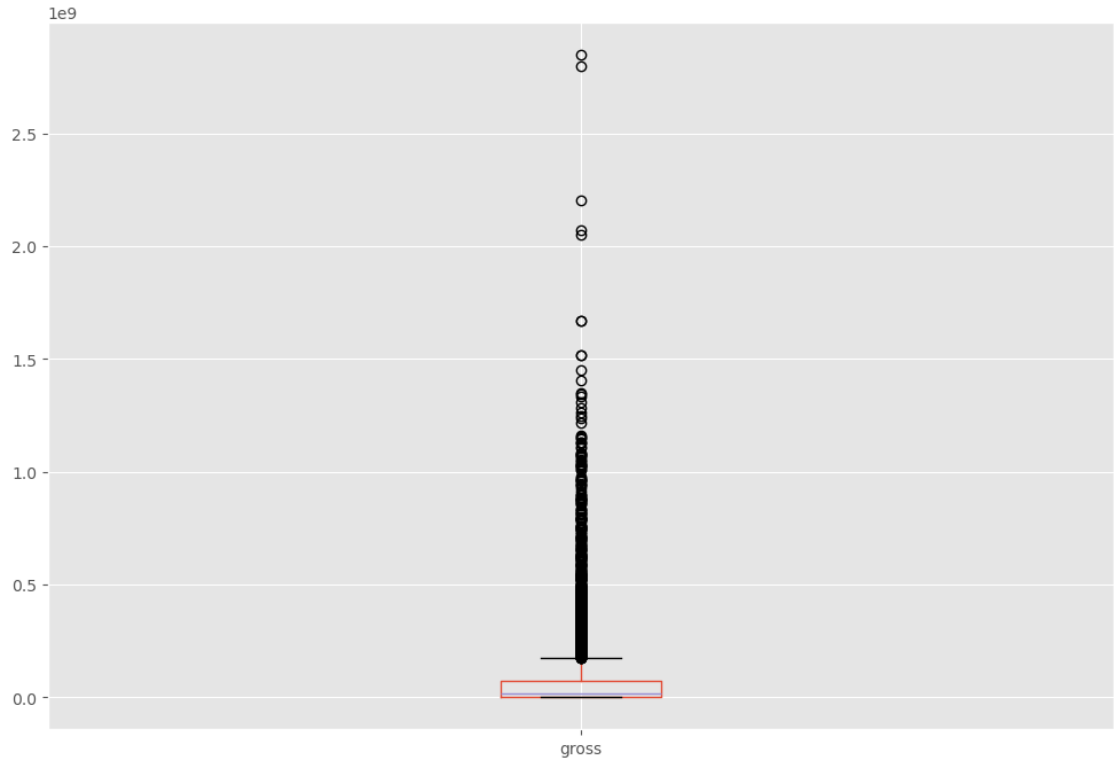
[59]:

| | name | rating | genre | year | released | score | votes | director | writer | star | country | budget |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5445 | Avatar | PG-13 | Action | 2009 | December 18, 2009 (United States) | 7.8 | 1100000.0 | James Cameron | James Cameron | Sam Worthington | United States | 237000000 |
| 7445 | Avengers: Endgame | PG-13 | Action | 2019 | April 26, 2019 (United States) | 8.4 | 903000.0 | Anthony Russo | Christopher Markus | Robert Downey Jr. | United States | 356000000 |
| 3045 | Titanic | PG-13 | Drama | 1997 | December 19, 1997 (United States) | 7.8 | 1100000.0 | James Cameron | James Cameron | Leonardo DiCaprio | United States | 200000000 |
| 6663 | Star Wars: Episode VII - The Force Awakens | PG-13 | Action | 2015 | December 18, 2015 (United States) | 7.8 | 876000.0 | J.J. Abrams | Lawrence Kasdan | Daisy Ridley | United States | 245000000 |
| 7244 | Avengers: Infinity War | PG-13 | Action | 2018 | April 27, 2018 (United States) | 8.4 | 897000.0 | Anthony Russo | Christopher Markus | Robert Downey Jr. | United States | 321000000 |
| … | … | … | … | … | … | … | … | … | … | … | … | … |
| 1866 | Sex and Zen | R | Comedy | 1991 | November 30, 1991 (Hong Kong) | 5.6 | 2200.0 | Michael Mak | Alexander Lee | Lawrence Ng | Hong Kong | 0 |
| 1837 | La discrète | NaN | Drama | 1990 | November 21, 1990 (France) | 7.0 | 930.0 | Christian Vincent | Christian Vincent | Fabrice Luchini | France | 0 |
| 1838 | Heaven and Earth | PG-13 | Action | 1990 | February 8, 1991 (United States) | 7.0 | 958.0 | Haruki Kadokawa | Haruki Kadokawa | Takaaki Enoki | Japan | 42000000 |
| 1842 | Archangel | Not Rated | Comedy | 1990 | September 1, 1990 (Canada) | 6.5 | 1300.0 | Guy Maddin | John B. Harvie | Michael Gottli | Canada | 0 |
| 1814 | Boiling Point | Not Rated | Action | 1990 | November 19, 1999 (United States) | 6.8 | 6300.0 | Takeshi Kitano | Takeshi Kitano | Takeshi Kitano | Japan | 0 |

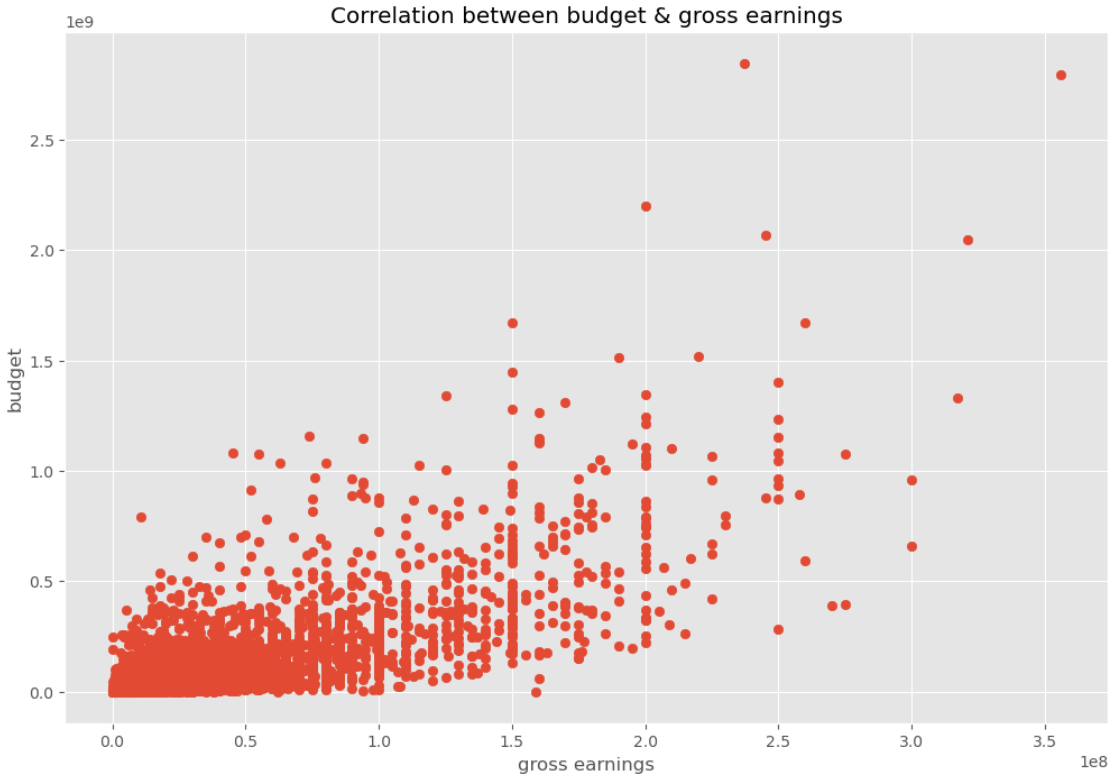
```
           gross                  company  runtime
5445  2847246203   Twentieth Century Fox    162.0
7445  2797501328          Marvel Studios    181.0
3045  2201647264   Twentieth Century Fox    194.0
6663  2069521700               Lucasfilm    138.0
7244  2048359754          Marvel Studios    149.0
…          …                        …      …
1866           0  Golden Harvest Company     99.0
1837           0         France 3 Cinéma     94.0
1838           0   Haruki Kadokawa Films    125.0
1842           0               Cinephile     90.0
1814           0   Bandai Visual Company     96.0

[7668 rows x 15 columns]
```

[61]:
```
#any outliers?
#are there any movies making a lot more money?
#yes as you can see

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

[61]: <Axes: >

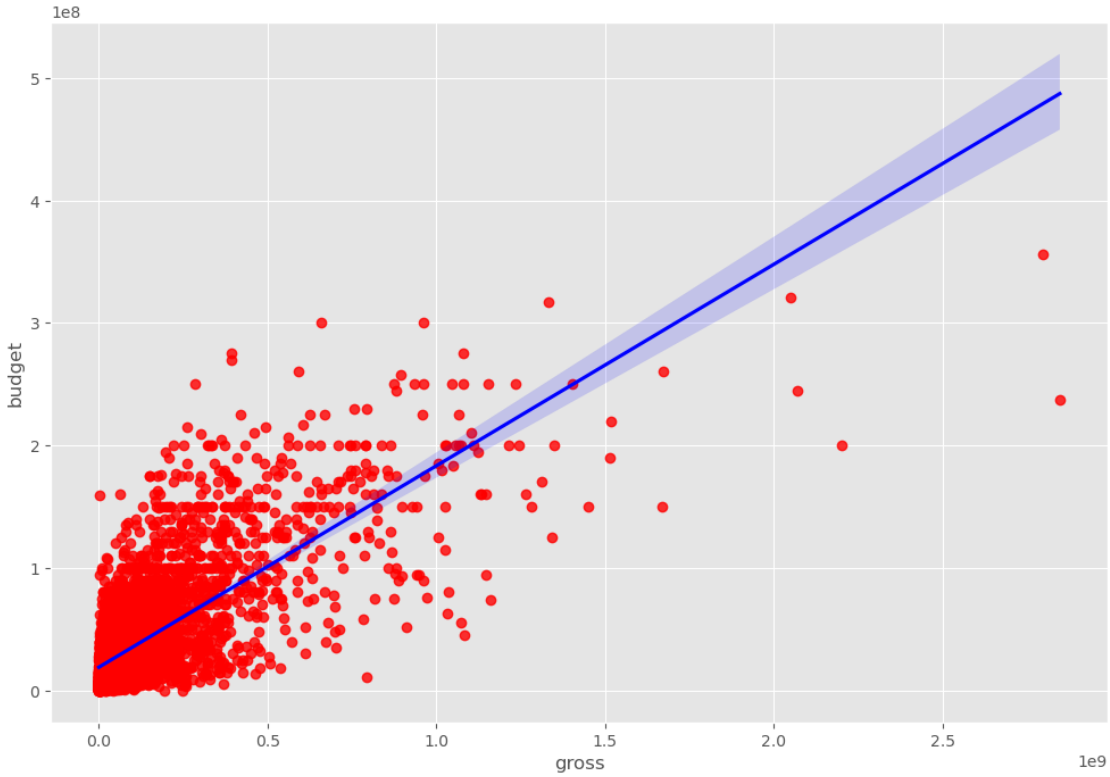


[13]:
```
#what is the relationship between budget and gross?
#correlation?
#assumption is that there is a strong positive correlation such that the more␣
 ↪inside the budget, the greater the returns

plt.scatter(x=df['budget'], y=df['gross'])

plt.title("Correlation between budget & gross earnings")

plt.xlabel("gross earnings")

plt.ylabel("budget")

plt.show()
```




[19]:
```
# add the line using regplot()
#change the color


sns.regplot(x="gross", y="budget", data=df, scatter_kws={'color': 'red'},␣
 ↪line_kws={'color': 'blue'})
```

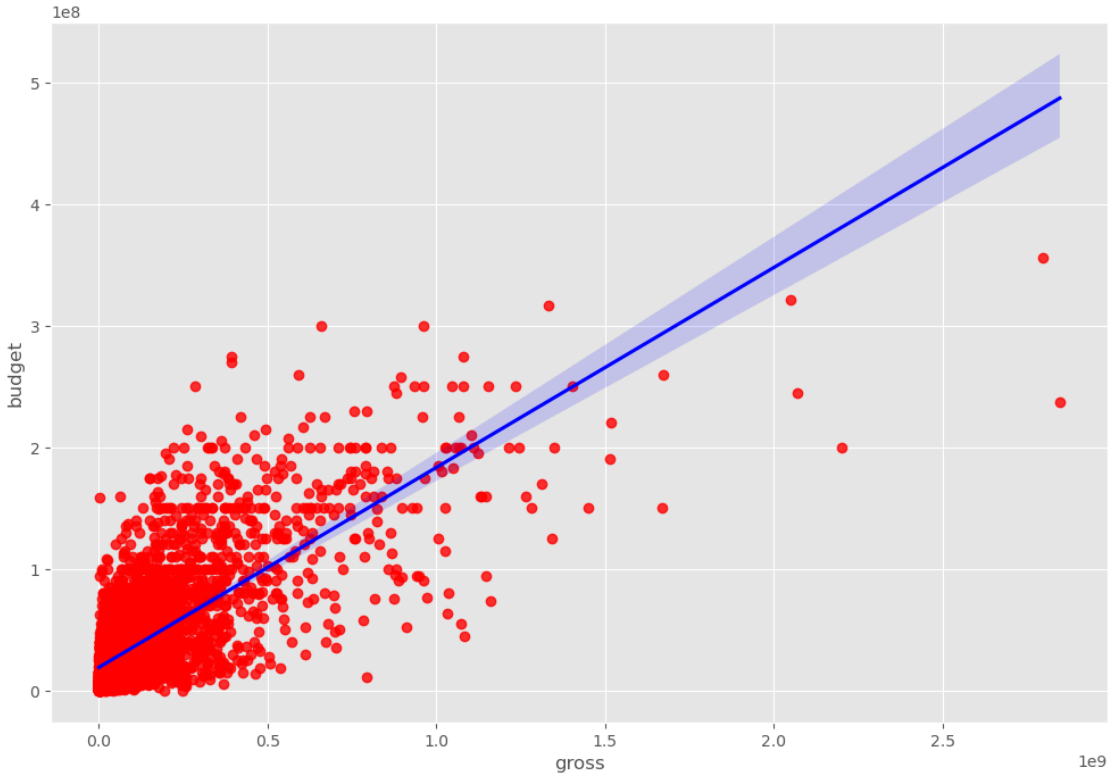
[19]: <Axes: xlabel='gross', ylabel='budget'>




[ ]:

[23]:
```
#short cut of doing everything we just did above
#plot the scatter
#add the line


sns.regplot(x="gross", y="budget", data=df, scatter_kws={'color': 'red'},␣
 ↪line_kws={'color': 'blue'})
```

[23]: <Axes: xlabel='gross', ylabel='budget'>




[33]:
```
correlation_matrix = df.corr(method='pearson', numeric_only=True)
correlation_matrix


#pearson, kendall, spearman #three different types of correlation
#pearson is ussually default
#as you can see, there is a strong posiitive correlation between budget and␣
 ↪gross of 0.740
#our asumption was correct



# sns.heatmap(correlation_matrix, annot=True)

# plt.show()
```

[33]:

| | 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
```

[29]:
```
df.corr(method='kendall', numeric_only=True)
```

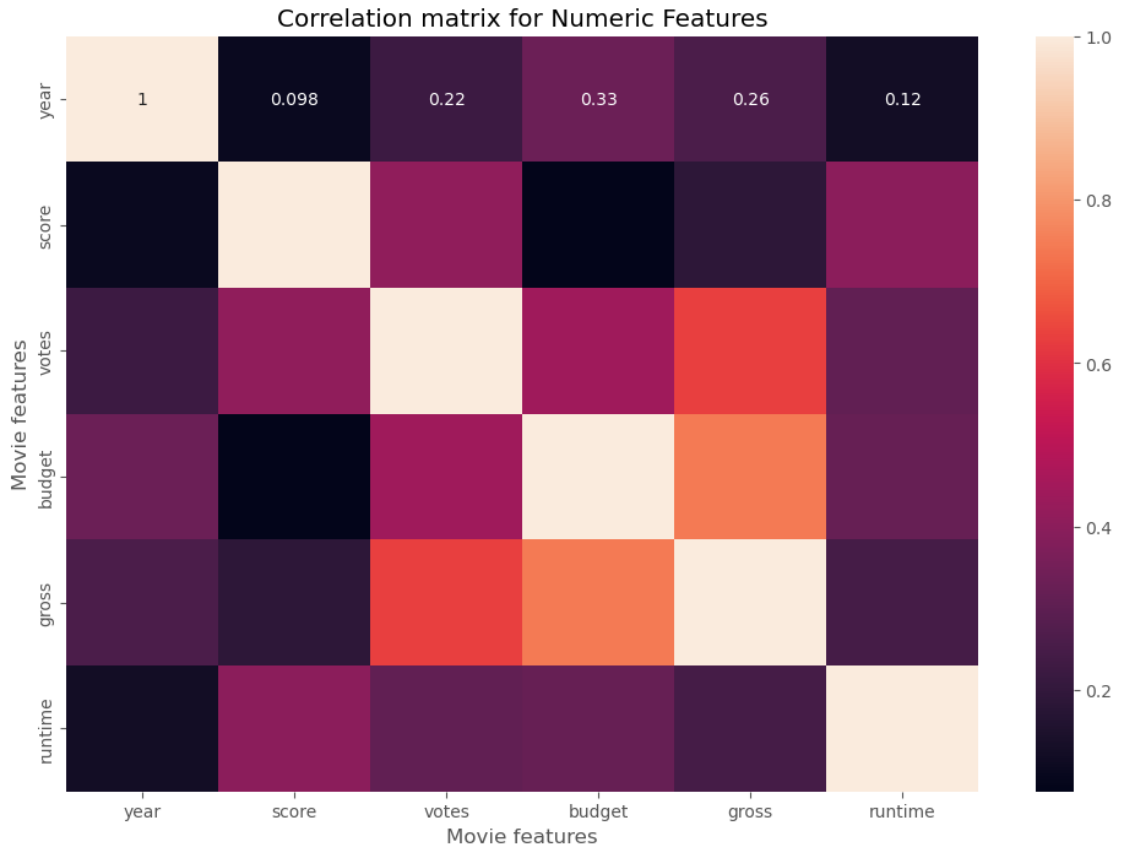
[29]:
```
             year     score     votes    budget     gross   runtime
year     1.000000  0.067652  0.331465  0.224120  0.200618  0.097184
score    0.067652  1.000000  0.300115 -0.000566  0.086046  0.283611
votes    0.331465  0.300115  1.000000  0.353702  0.548899  0.198240
budget   0.224120 -0.000566  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
```

[31]:
```
df.corr(method='spearman', numeric_only=True)
```

[31]:
```
             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
```

[ ]:
```

```

[41]:
```
sns.heatmap(correlation_matrix, annot=True)

plt.title("Correlation matrix for Numeric Features")

plt.xlabel("Movie features")

plt.ylabel("Movie features")

plt.show()
```


Correlation matrix for Numeric Features
year
score
votes
budget
gross
runtime
1
0.098
0.22
0.33
0.26
0.12
Movie features
year
score
votes
budget
gross
runtime
Movie features
1.0
0.8
0.6
0.4
0.2

[ ]:

[ ]:

[ ]:

[ ]: