Movie Correlation Project

[2]:
```
# This Python program finds the highest correlation factors with gross earnings␣
 ↪using movie data.

# import libraries
import numpy as np
import pandas as pd
import seaborn as sns

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

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

pd.options.mode.chained_assignment = None

# Read in data
df = pd.read_csv('/Users/jan/desktop/data/jy/future/codesamples/python/python1/
 ↪movies.csv')

df.head()
```

[2]:

	name	rating	genre	year	released	score	votes	director
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams
4	Caddyshack	R	Comedy	1980				

```
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
```

[8]:
```
# Check for missing data
for col in df.columns:
    pct_missing = np.mean(df[col].isnull())
    print('{} - {}%'.format(col, pct_missing))

# Check data types
df.dtypes
```

```
name - 0.0%
rating - 0.010041731872717789%
genre - 0.0%
year - 0.0%
released - 0.0002608242044861763%
score - 0.00039123630672926445%
votes - 0.00039123630672926445%
director - 0.0%
writer - 0.00039123630672926445%
star - 0.00013041210224308815%
country - 0.00039123630672926445%
budget - 0.0%
gross - 0.0%
company - 0.002217005738132499%
runtime - 0.0005216484089723526%
```

[8]:
```
name         object
rating       object
genre        object
year          int64
released     object
score       float64
votes       float64
```

```
director     object
writer       object
star         object
country      object
budget        int64
gross         int64
company      object
runtime     float64
dtype: object
```

[6]:
```
# Change NaN(non-data type) to zero
df['budget'] = df['budget'].fillna(0).astype('int64')
df.gross = df.gross.fillna(0).astype('int64')
```

[4]:
```
df.sort_values(by=['gross'], inplace=False, ascending=False)
```

[4]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	James Cameron	Sam Worthington	United States	237000000.0	2847246203	Twentieth Century Fox	162.0
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	356000000.0	2797501328	Marvel Studios	181.0
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	James Cameron	Leonardo DiCaprio	United States	200000000.0	2201647264	Twentieth Century Fox	194.0
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	245000000.0	2069521700	Lucasfilm	138.0
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	321000000.0	2048359754	Marvel Studios	149.0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks	Joseph Ebanks	Joseph Ebanks	United States	0.0	0	0	0.0
7664	Dream Round	NaN	Comedy	2020	February 7, 2020 (United States)	4.7	36.0	Dusty Dukatz	Dusty Dukatz	Dusty Dukatz	United States	0.0	0	0	0.0
7665	Saving Mbango	NaN	Drama	2020	April 27, 2020 (Cameroon)	5.7	29.0	Nkanya Nkwai	Nkanya Nkwai	Nkanya Nkwai	Cameroon	0.0	0	0	0.0
7666	It's Just Us	NaN	Drama	2020	October 1, 2020 (United States)	NaN	NaN	James Randall	James Randall	James Randall	United States	0.0	0	0	0.0
7667	Tee em el	NaN	Horror	2020	August 19, 2020 (United States)	5.7	7.0	Pereko Mosia	Pereko Mosia	Pereko Mosia	United States	0.0	0	0	0.0

```
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]
```

[11]:
```
pd.set_option('display.max_rows', 10)
```

[12]:
```
df.sort_values(by=['gross'], inplace=False, ascending=False)
```

[12]:
```
                                      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
…                                       …    …       …   …
1617                    Mortal Passions      R    Crime  1989
1614                     Edge of Sanity      R   Horror  1989
1606                          I, Madman      R  Fantasy  1989
1601               My Twentieth Century    NaN   Comedy  1989
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
…                                  …    …        …              …
```

```
1617   January 26, 1990 (United States)    4.5      274.0     Andrew Lane
1614     April 14, 1989 (United States)    5.2     1300.0  Gérard Kikoïne
1606      April 7, 1989 (United States)    6.0     2900.0    Tibor Takács
1601           January 13, 1990 (Japan)    7.1     1500.0   Ildikó Enyedi
7667    August 19, 2020 (United States)    5.7        7.0    Pereko Mosia

                 writer               star         country     budget  \
5445       James Cameron    Sam Worthington   United States  237000000
7445  Christopher Markus  Robert Downey Jr.   United States  356000000
3045       James Cameron  Leonardo DiCaprio   United States  200000000
6663     Lawrence Kasdan       Daisy Ridley   United States  245000000
7244  Christopher Markus  Robert Downey Jr.   United States  321000000
…                   …                 …             …         …
1617      Alan Moskowitz      Zach Galligan   United States          0
1614          J.P. Félix    Anthony Perkins  United Kingdom          0
1606       David Chaskin       Jenny Wright          Canada          0
1601       Ildikó Enyedi       Dorota Segda         Hungary          0
7667        Pereko Mosia   Siyabonga Mabaso    South Africa          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
…          …                                …      …
1617           0          Gibraltar Entertainment     92.0
1614           0                     Allied Vision     85.0
1606           0  Trans World Entertainment (TWE)     89.0
1601           0         Budapest Stúdió Vállalat    104.0
7667           0                      PK 65 Films    102.0

[7668 rows x 15 columns]
```

[7]:
```
pd.set_option('display.max_rows', 20)
```

[8]:
```
df
```

[8]:
```
                                          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]
```

[9]:
```
df['company'].drop_duplicates().sort_values(ascending=True)
```

```
[9]: 4345        "DIA" Productions GmbH & Co. KG
     7525    "Weathering With You" Film Partners
     3024                        .406 Production
     3748                     1+2 Seisaku Iinkai
     5813                        10 West Studios
                             …
     4007                                i5 Films
     6412                iDeal Partners Film Fund
     5664                             micro_scope
     7129                                 thefyzz
     408                                      NaN
     Name: company, Length: 2386, dtype: object
```

```
[10]: df['company'].sort_values(ascending=True)
```

```
[10]: 4345        "DIA" Productions GmbH & Co. KG
      7525    "Weathering With You" Film Partners
      3024                        .406 Production
      3748                     1+2 Seisaku Iinkai
      5813                        10 West Studios
                              …
      7599                                     NaN
      7657                                     NaN
      7662                                     NaN
      7663                                     NaN
      7666                                     NaN
      Name: company, Length: 7668, dtype: object
```

```
[11]: df['company'].drop_duplicates()
```

```
[11]: 0              Warner Bros.
      1          Columbia Pictures
      2                  Lucasfilm
      3         Paramount Pictures
      4             Orion Pictures
                    …
      7658            Notis Studio
      7660    Abominable Pictures
      7661         Dow Jazz Films
      7665       Embi Productions
      7667             PK 65 Films
      Name: company, Length: 2386, dtype: object
```

```
[12]: df
```

```
[12]:                                        name rating      genre  year  \
      0                                The Shining      R      Drama  1980
```

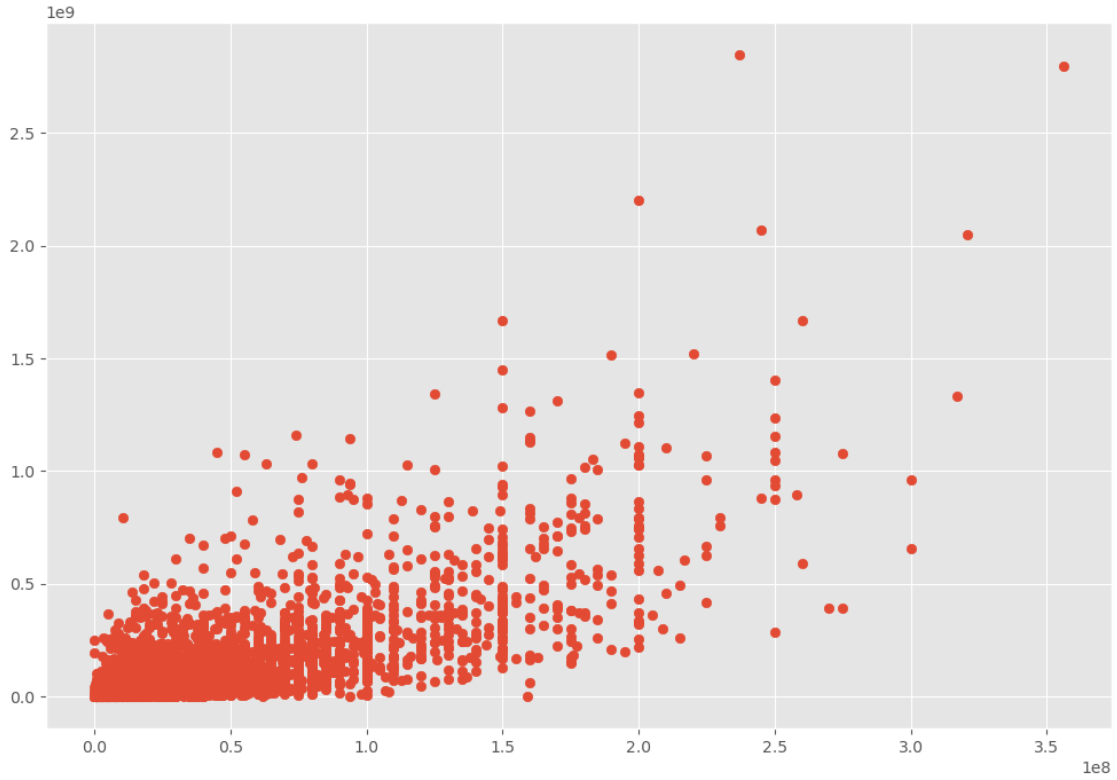
```
1                                    The Blue Lagoon      R  Adventure  1980
2     Star Wars: Episode V - The Empire Strikes Back     PG     Action  1980
3                                         Airplane!     PG     Comedy  1980
4                                       Caddyshack      R     Comedy  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]
```

[9]:
```
# Hypothesis 1: budget has high correlation with gross earnings
# Hypothesis 2: movie company has high correlation with gross earnings
```

[10]:
```
# Scatter plot with budget vs gross
```

[15]:
```
plt.scatter(x=df['budget'], y=df['gross'])
plt.show()
```

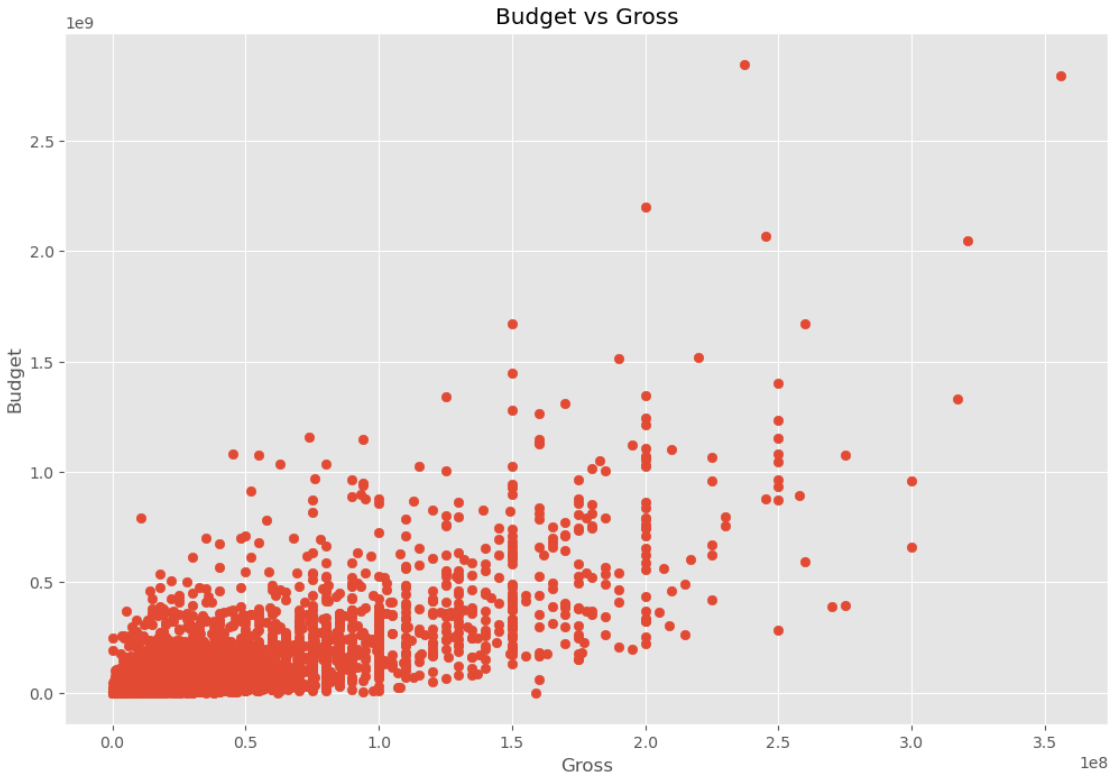


[16]:
```
df.head
```

[16]:
```
<bound method NDFrame.head of
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
```

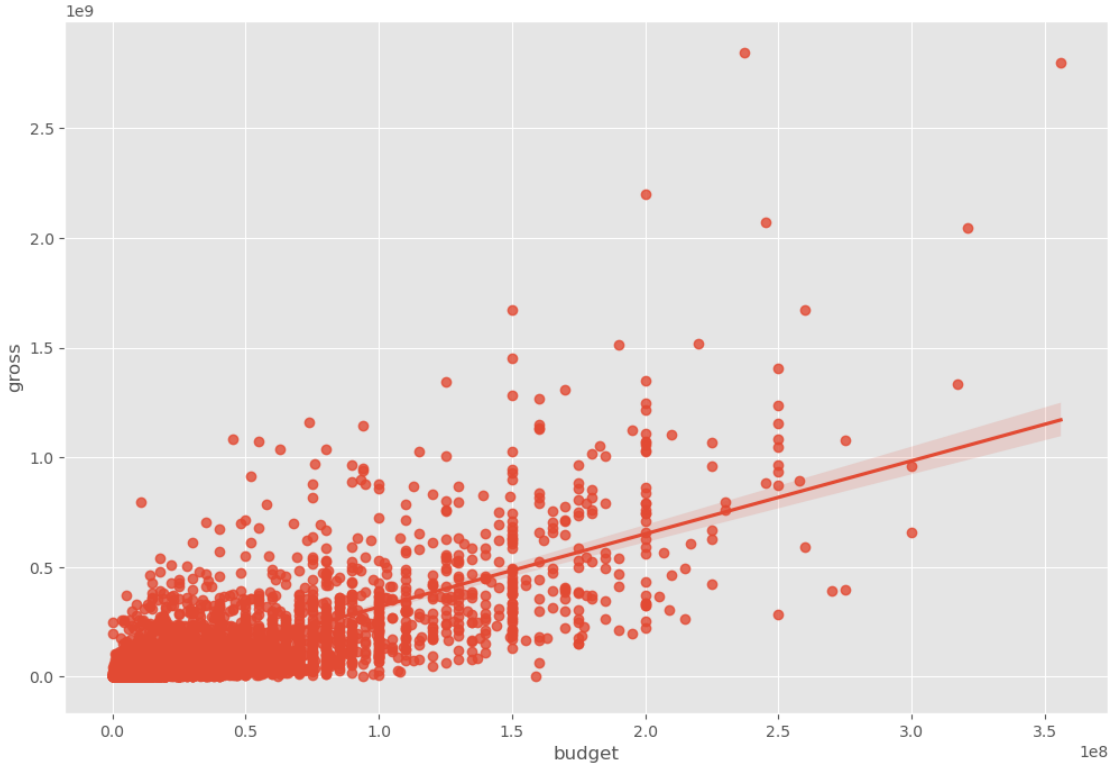
```
…                                                …    …        …   …
7663                                  More to Life    NaN      Drama  2020
7664                                   Dream Round    NaN     Comedy  2020
7665                                  Saving Mbango    NaN      Drama  2020
7666                                   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]>
```

[17]:
```
plt.scatter(x=df['budget'], y=df['gross'])
plt.title('Budget vs Gross')
plt.xlabel('Gross')
plt.ylabel('Budget')
plt.show()
```



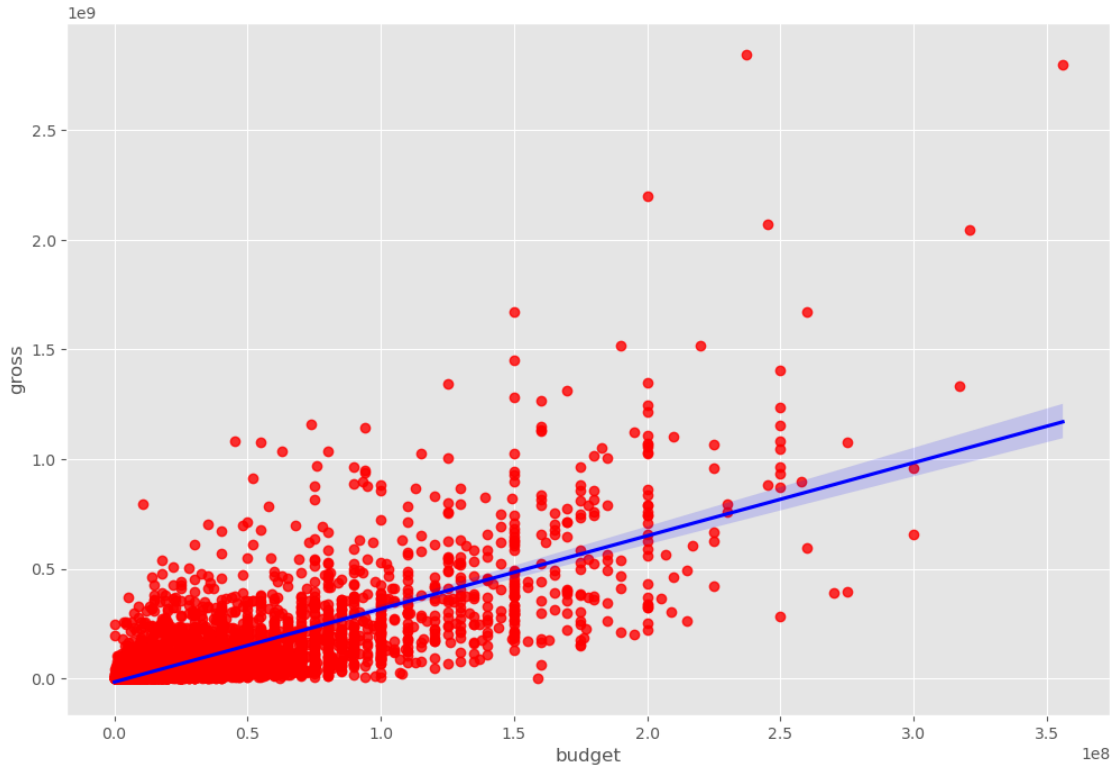
[18]:
```
sns.regplot(x='budget', y='gross', data=df)
```

[18]:
```
<AxesSubplot:xlabel='budget', ylabel='gross'>
```



[22]:
```
sns.regplot(x='budget', y='gross', data=df, scatter_kws={"color": "red"},␣
↪line_kws={"color": "blue"})
```

[22]: <AxesSubplot:xlabel='budget', ylabel='gross'>



[23]:
```
df.corr()
```

[23]:

	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

[24]:
```
df.corr(method='pearson')
```

[24]:

	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

[25]:
```
df.corr(method='kendall')
```

[25]:

	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

[26]:
```
df.corr(method='spearman')
```

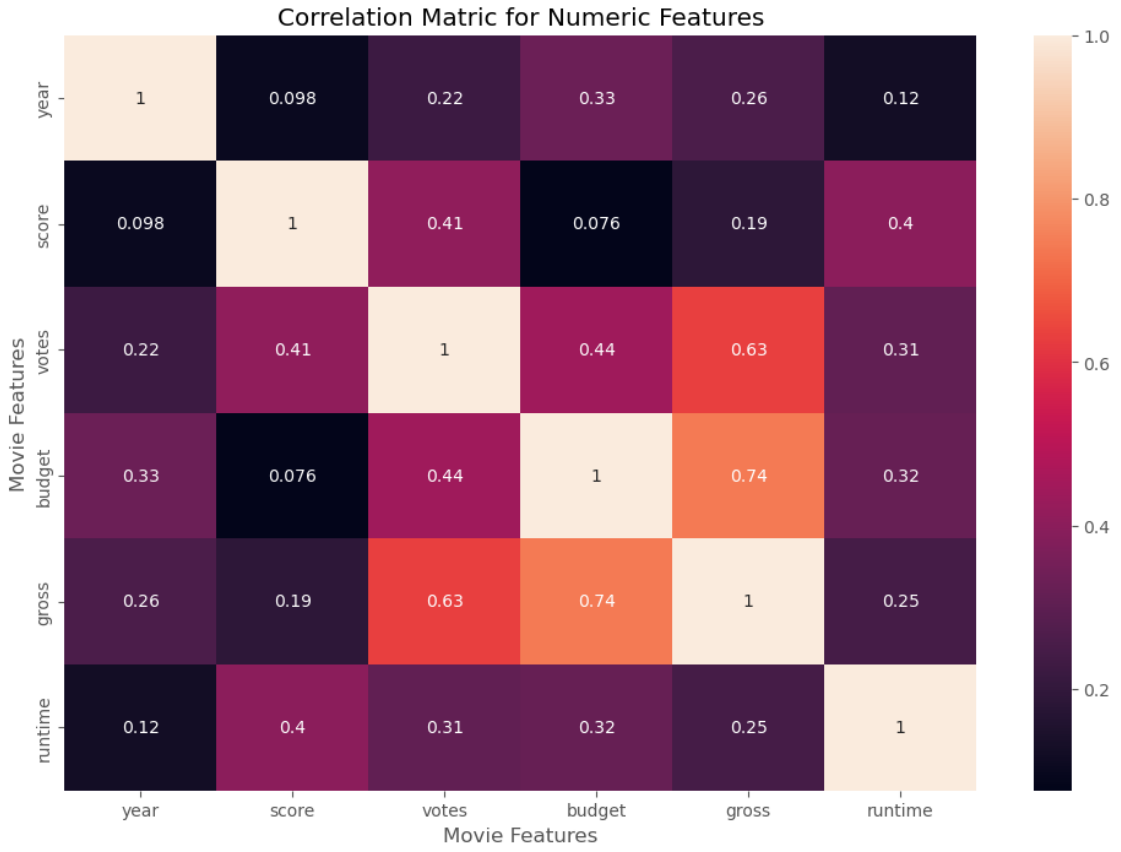
[26]:

	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

[27]:
```
correlation_matrix = df.corr(method='pearson')
```

[28]:
```
sns.heatmap(correlation_matrix, annot=True)
plt.show()
```

[29]:
```
sns.heatmap(correlation_matrix, annot=True)
plt.title('Correlation Matric for Numeric Features')
plt.xlabel('Movie Features')
plt.ylabel('Movie Features')
plt.show()
```



[30]:
```
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
```

[30]:

	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]
```

[42]:
```
df_numerized.corr()
```

[42]:
```
                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

               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

              gross   company   runtime
name       0.005533  0.009211  0.010392
rating    -0.107339 -0.032943  0.062145
genre     -0.235650 -0.071067 -0.052711
year       0.257486 -0.010431  0.120811
released   0.001659 -0.010474  0.000868
score      0.186258  0.001030  0.399451
votes      0.630757  0.133204  0.309212
director  -0.014441  0.004404  0.017624
writer    -0.023519  0.005646 -0.003511
star      -0.002717  0.012442  0.010174
country    0.092129  0.095548 -0.078412
budget     0.740395  0.173214  0.320447
gross      1.000000  0.154840  0.245216
company    0.154840  1.000000  0.034402
runtime    0.245216  0.034402  1.000000
```

[43]:
```
correlation_mat = df_numerized.corr()
corr_pairs = correlation_mat.unstack()
corr_pairs
```

[43]:
```
name     name         1.000000
         rating      -0.008069
         genre        0.016355
         year         0.011453
         released    -0.011311
                         …
runtime  country     -0.078412
         budget       0.320447
         gross        0.245216
         company      0.034402
         runtime      1.000000
Length: 225, dtype: float64
```

[44]:
```
sorted_paris = corr_pairs.sort_values()
sorted_paris
```

[44]:
```
budget    genre     -0.356564
genre     budget    -0.356564
          gross     -0.235650
gross     genre     -0.235650
rating    budget    -0.176002
                      ...
year      year       1.000000
genre     genre      1.000000
rating    rating     1.000000
company   company    1.000000
runtime   runtime    1.000000
Length: 225, dtype: float64
```

[45]:
```
high_corr = sorted_paris[sorted_paris > 0.5]
high_corr
```

[45]:
```
gross     votes       0.630757
votes     gross       0.630757
budget    gross       0.740395
gross     budget      0.740395
name      name        1.000000
director  director    1.000000
gross     gross       1.000000
budget    budget      1.000000
country   country     1.000000
star      star        1.000000
writer    writer      1.000000
votes     votes       1.000000
score     score       1.000000
released  released    1.000000
year      year        1.000000
genre     genre       1.000000
rating    rating      1.000000
company   company     1.000000
runtime   runtime     1.000000
dtype: float64
```

[3]:
```
# Votes and budget have the highest correlation to gross earnings.
# Company has no low correlation.
```

[ ]:
```

```