In [13]:
```
# Import libraries mport pandas as pd
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) # adjusts the configuration of the
plots we will create

# read in the data

df= pd.read_csv(r'C:\Users\coold\Downloads\movies1.csv')
```

In [14]:
```
# let's look at the data
df.head()
```

Out[14]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	19000000.0
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields	United States	4500000.0
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner	Leigh Brackett	Mark Hamill	United States	18000000.0
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3500000.0
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	6000000.0

In [20]:
```
# now we will look for missing data

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%
```

In [17]:
```
# Data types for our columns
df.dtypes
```

Out[17]:
```
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
```

In [31]:
```
df.sort_values (by=['gross'], inplace=False, ascending=False)
```

Out[31]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	bud
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	James Cameron	Sam Worthington	United States	2370000C
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	3560000C
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	James Cameron	Leonardo DiCaprio	United States	2000000C
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	2450000C
7244	Avengers: Infinity	PG-13	Action	2018	April 27, 2018	8.4	897000.0	Anthony	Christopher	Robert	United	3210000C

	War				(United States)			Russo	Markus	Downey Jr.	States	
...	...	...	...	...	...	...	...	...	...	...	...	
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks	Joseph Ebanks	Shannon Bond	United States	700
7664	Dream Round	NaN	Comedy	2020	February 7, 2020 (United States)	4.7	36.0	Dusty Dukatz	Lisa Huston	Michael Saquella	United States	N
7665	Saving Mbango	NaN	Drama	2020	April 27, 2020 (Cameroon)	5.7	29.0	Nkanya Nkwai	Lynno Lovert	Onyama Laura	United States	5875
7666	It's Just Us	NaN	Drama	2020	October 1, 2020 (United States)	NaN	NaN	James Randall	James Randall	Christina Roz	United States	150
7667	Tee em el	NaN	Horror	2020	August 19, 2020 (United States)	5.7	7.0	Pereko Mosia	Pereko Mosia	Siyabonga Mabaso	South Africa	N

7668 rows × 15 columns

In [71]:
```
pd.set_option('display.max.rows', 1000)
```

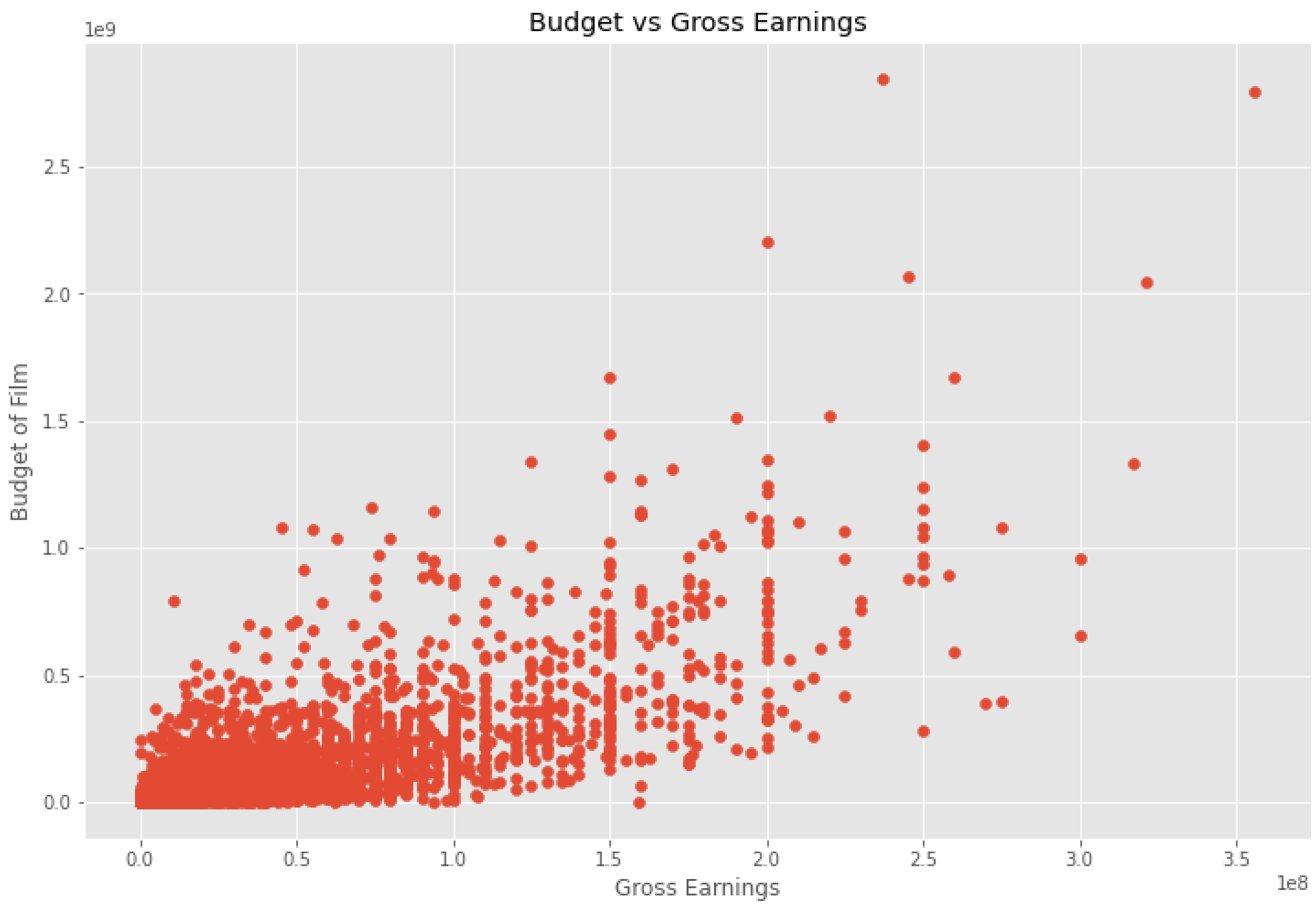
In [72]:
```
# Drop any duplicates

df['company'].drop_duplicates().sort_values(ascending=False)
```

Out[72]:
```
7129    2384
5664    2383
6412    2382
4007    2381
6793    2380
        ...
3748       3
3024       2
7525       1
4345       0
408       -1
Name: company, Length: 2386, dtype: int16
```

In [ ]:
```
# Budget will have a high correlation with increase gross
# Company high correlation with increase gross
```

In [36]:
```
# Scatter plot with budget vs gross

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

plt.title('Budget vs Gross Earnings')

plt.xlabel('Gross Earnings')
plt.ylabel('Budget of Film')

plt.show
```

Out[36]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



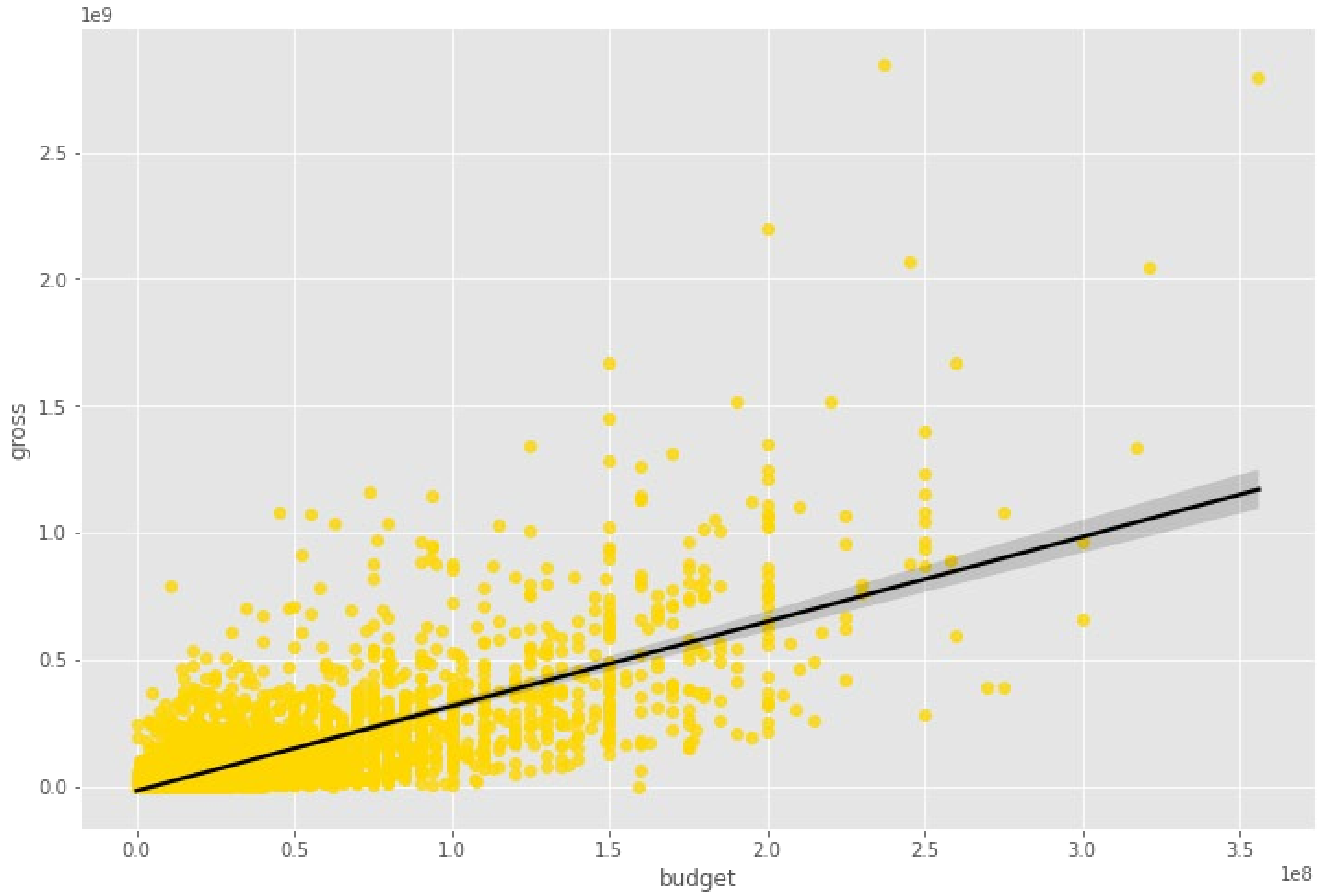
In [73]:
```
df.head()
```

Out[73]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	c
0	The Shining	6	6	1980	1705	8.4	927000.0	2589	4014	1047	54	19000000.0	46998772.0	
1	The Blue Lagoon	6	1	1980	1492	5.8	65000.0	2269	1632	327	55	4500000.0	58853106.0	
2	Star Wars: Episode V - The Empire Strikes Back	4	0	1980	1771	8.7	1200000.0	1111	2567	1745	55	18000000.0	538375067.0	
3	Airplane!	4	4	1980	1492	7.7	221000.0	1301	2000	2246	55	3500000.0	83453539.0	
4	Caddyshack	6	4	1980	1543	7.3	108000.0	1054	521	410	55	6000000.0	39846344.0	

In [41]:
```
# Plot budget vs Gross using seaborn

sns.regplot(x='budget', y='gross', data=df, scatter_kws={'color': 'gold'},
line_kws={'color':'black'})
```

Out[41]:
```
<AxesSubplot:xlabel='budget', ylabel='gross'>
```



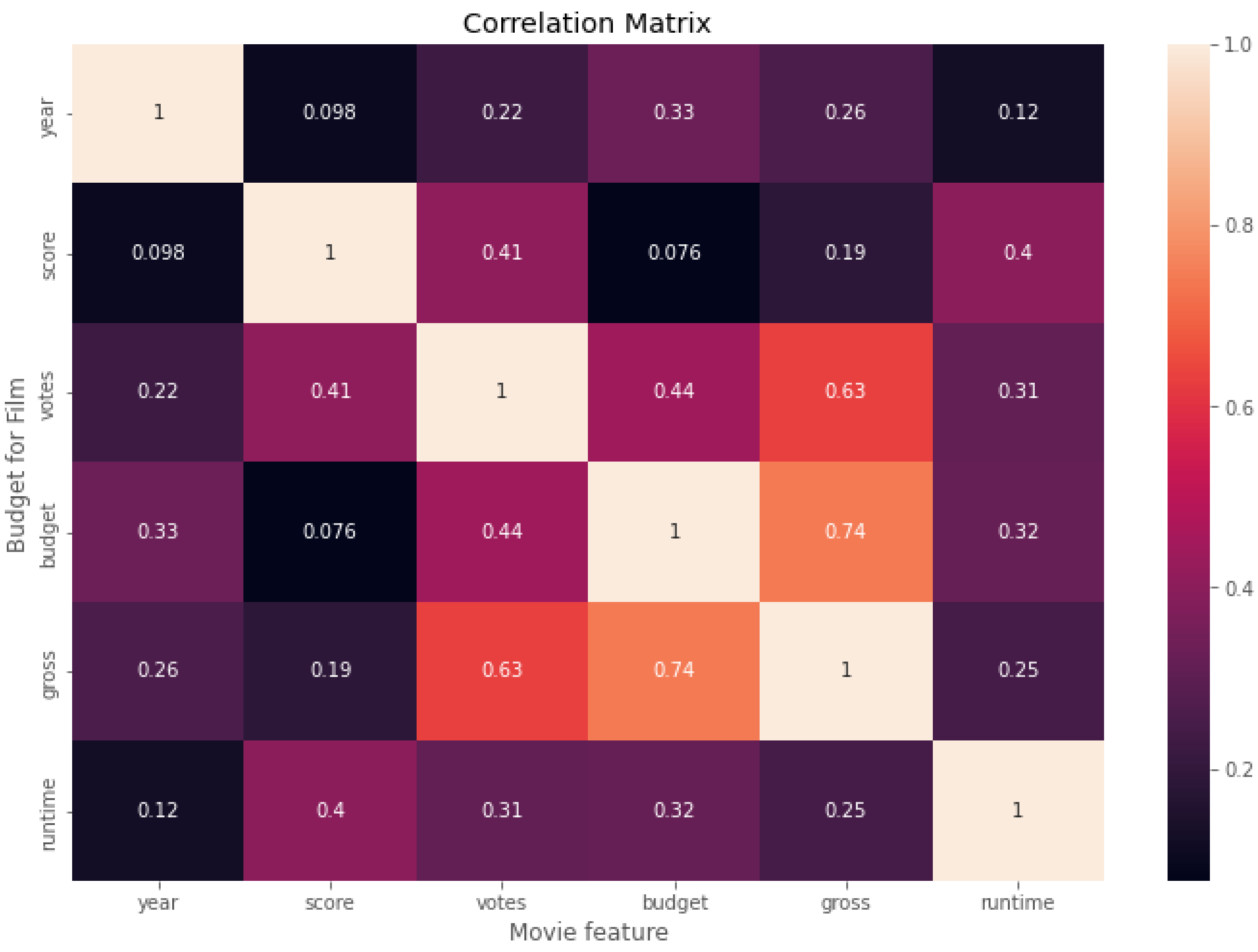
In [43]:
```
#Let's start looking at correlation
df.corr(method='pearson') # different correlation metrics pearson, kendall, spearman
```

Out[43]:

	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

In [44]:
```
# High Correlation between budget and gross
# my hypothesis was correct
```

In [50]:
```
correlation_matrix = df.corr()

sns.heatmap(correlation_matrix, annot=True)

plt.title('Correlation Matrix')

plt.xlabel('Movie feature')
plt.ylabel('Budget for Film')

plt.show()
```



	year	score	votes	budget	gross	runtime
year	1	0.098	0.22	0.33	0.26	0.12
score	0.098	1	0.41	0.076	0.19	0.4
votes	0.22	0.41	1	0.44	0.63	0.31
budget	0.33	0.076	0.44	1	0.74	0.32
gross	0.26	0.19	0.63	0.74	1	0.25
runtime	0.12	0.4	0.31	0.32	0.25	1

In [74]:
```
# we will now look at another factor which is Company

df.head()
```

Out[74]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	c
0	The Shining	6	6	1980	1705	8.4	927000.0	2589	4014	1047	54	19000000.0	46998772.0	
1	The Blue Lagoon	6	1	1980	1492	5.8	65000.0	2269	1632	327	55	4500000.0	58853106.0	
2	Star Wars: Episode V - The Empire Strikes Back	4	0	1980	1771	8.7	1200000.0	1111	2567	1745	55	18000000.0	538375067.0	
3	Airplane!	4	4	1980	1492	7.7	221000.0	1301	2000	2246	55	3500000.0	83453539.0	
4	Caddyshack	6	4	1980	1543	7.3	108000.0	1054	521	410	55	6000000.0	39846344.0	

In [75]:
```
# Now we have to assign data types that are not numbers a random number so we can
compare it with floats and int
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
```

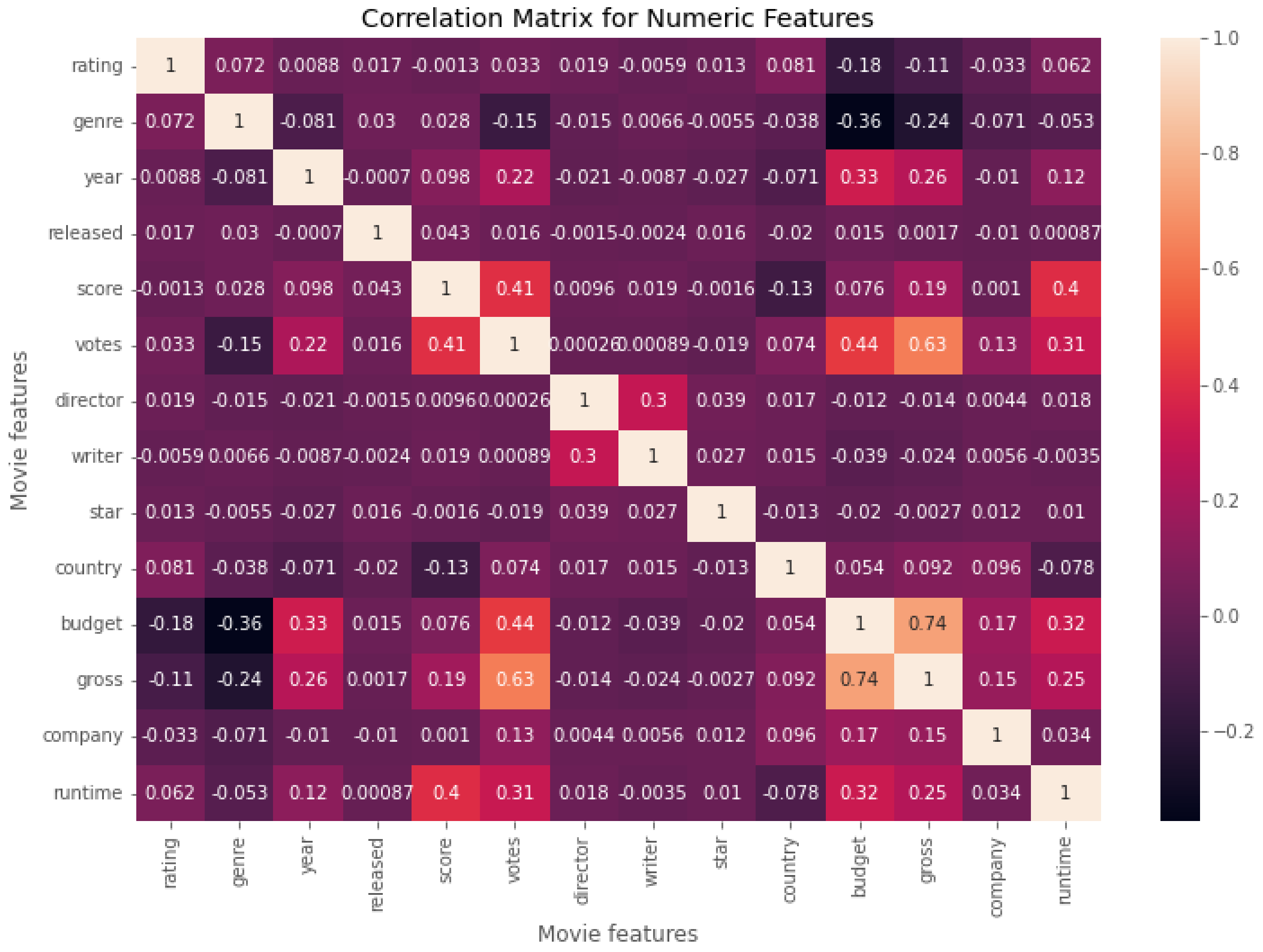
Out[75]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gros
0	The Shining	6	6	1980	1705	8.4	927000.0	2589	4014	1047	54	19000000.0	46998772.
1	The Blue Lagoon	6	1	1980	1492	5.8	65000.0	2269	1632	327	55	4500000.0	58853106.
2	Star Wars: Episode V - The Empire Strikes Back	4	0	1980	1771	8.7	1200000.0	1111	2567	1745	55	18000000.0	538375067.
3	Airplane!	4	4	1980	1492	7.7	221000.0	1301	2000	2246	55	3500000.0	83453539.
4	Caddyshack	6	4	1980	1543	7.3	108000.0	1054	521	410	55	6000000.0	39846344.
...	...	...	...	...	...	...	...	...	...	...	...	...	..
7663	More to Life	-1	6	2020	2964	3.1	18.0	1500	2289	2421	55	7000.0	Na
7664	Dream Round	-1	4	2020	1107	4.7	36.0	774	2614	1886	55	NaN	Na
7665	Saving Mbango	-1	6	2020	193	5.7	29.0	2061	2683	2040	55	58750.0	Na
7666	It's Just Us	-1	6	2020	2817	NaN	NaN	1184	1824	450	55	15000.0	Na
7667	Tee em el	-1	10	2020	391	5.7	7.0	2165	3344	2463	44	NaN	Na

7668 rows × 15 columns

In [57]:

```
df_numerized = df.corr()

sns.heatmap(df_numerized, annot=True)

plt.title('Correlation Matrix for Numeric Features')

plt.xlabel('Movie features')

plt.ylabel('Movie features')

plt.show()
```



	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime
rating	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
genre	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
year	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
released	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
score	-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
votes	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
director	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
writer	-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
star	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
country	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
budget	-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
gross	-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
company	-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
runtime	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

In [81]:
```
correlation_mat = df_numerized.corr()

corr_pairs = correlation_mat.unstack()
```

In [82]:
```
sorted_pairs = corr_pairs.sort_values()

sorted_pairs
```

Out[82]:
```
budget    genre      -0.356564
genre     budget     -0.356564
gross     genre      -0.235650
genre     gross      -0.235650
budget    rating     -0.176002
rating    budget     -0.176002
votes     genre      -0.145307
genre     votes      -0.145307
score     country    -0.133348
country   score      -0.133348
rating    gross      -0.107339
gross     rating     -0.107339
year      genre      -0.081261
genre     year       -0.081261
country   runtime    -0.078412
runtime   country    -0.078412
company   genre      -0.071067
genre     company    -0.071067
```

```
year      country    -0.070938
country   year       -0.070938
runtime   genre      -0.052711
genre     runtime    -0.052711
writer    budget     -0.039451
budget    writer     -0.039451
country   genre      -0.037615
genre     country    -0.037615
rating    company    -0.032943
company   rating     -0.032943
star      year       -0.027242
year      star       -0.027242
gross     writer     -0.023519
writer    gross      -0.023519
director  year       -0.020795
year      director   -0.020795
released  country    -0.020427
country   released   -0.020427
star      budget     -0.019589
budget    star       -0.019589
votes     star       -0.019282
star      votes      -0.019282
genre     director   -0.015258
director  genre      -0.015258
          gross      -0.014441
gross     director   -0.014441
star      country    -0.012998
country   star       -0.012998
budget    director   -0.012272
director  budget     -0.012272
released  company    -0.010474
company   released   -0.010474
year      company    -0.010431
company   year       -0.010431
year      writer     -0.008656
writer    year       -0.008656
rating    writer     -0.005921
writer    rating     -0.005921
star      genre      -0.005477
genre     star       -0.005477
runtime   writer     -0.003511
writer    runtime    -0.003511
star      gross      -0.002717
gross     star       -0.002717
released  writer     -0.002404
writer    released   -0.002404
score     star       -0.001609
star      score      -0.001609
released  director   -0.001478
director  released   -0.001478
score     rating     -0.001314
rating    score      -0.001314
released  year       -0.000695
year      released   -0.000695
votes     director    0.000260
director  votes       0.000260
released  runtime     0.000868
```

```
        runtime   released    0.000868
        votes     writer      0.000892
        writer    votes       0.000892
        company   score       0.001030
        score     company     0.001030
        released  gross       0.001659
        gross     released    0.001659
        director  company     0.004404
        company   director    0.004404
        writer    company     0.005646
        company   writer      0.005646
        genre     writer      0.006567
        writer    genre       0.006567
        year      rating      0.008779
        rating    year        0.008779
        score     director    0.009559
        director  score       0.009559
        runtime   star        0.010174
        star      runtime     0.010174
                  company     0.012442
        company   star        0.012442
        star      rating      0.013405
        rating    star        0.013405
        budget    released    0.014683
        released  budget      0.014683
        writer    country     0.015343
        country   writer      0.015343
        star      released    0.015777
        released  star        0.015777
                  votes       0.016097
        votes     released    0.016097
        rating    released    0.016613
        released  rating      0.016613
        director  country     0.017490
        country   director    0.017490
        director  runtime     0.017624
        runtime   director    0.017624
        writer    score       0.019416
        score     writer      0.019416
        rating    director    0.019483
        director  rating      0.019483
        star      writer      0.027245
        writer    star        0.027245
        score     genre       0.027965
        genre     score       0.027965
                  released    0.029822
        released  genre       0.029822
        rating    votes       0.033225
        votes     rating      0.033225
        company   runtime     0.034402
        runtime   company     0.034402
        star      director    0.039234
        director  star        0.039234
        released  score       0.042788
        score     released    0.042788
        country   budget      0.054063
        budget    country     0.054063
```

```
rating    runtime     0.062145
runtime   rating      0.062145
genre     rating      0.072423
rating    genre       0.072423
votes     country     0.073625
country   votes       0.073625
score     budget      0.076254
budget    score       0.076254
rating    country     0.081244
country   rating      0.081244
gross     country     0.092129
country   gross       0.092129
          company     0.095548
company   country     0.095548
year      score       0.097995
score     year        0.097995
year      runtime     0.120811
runtime   year        0.120811
votes     company     0.133204
company   votes       0.133204
          gross       0.154840
gross     company     0.154840
budget    company     0.173214
company   budget      0.173214
score     gross       0.186258
gross     score       0.186258
votes     year        0.222945
year      votes       0.222945
runtime   gross       0.245216
gross     runtime     0.245216
year      gross       0.257486
gross     year        0.257486
writer    director    0.299067
director  writer      0.299067
votes     runtime     0.309212
runtime   votes       0.309212
budget    runtime     0.320447
runtime   budget      0.320447
year      budget      0.329321
budget    year        0.329321
runtime   score       0.399451
score     runtime     0.399451
          votes       0.409182
votes     score       0.409182
budget    votes       0.442429
votes     budget      0.442429
gross     votes       0.630757
votes     gross       0.630757
budget    gross       0.740395
gross     budget      0.740395
rating    rating      1.000000
gross     gross       1.000000
budget    budget      1.000000
country   country     1.000000
star      star        1.000000
writer    writer      1.000000
director  director    1.000000
```

```
votes     votes       1.000000
score     score       1.000000
released  released    1.000000
year      year        1.000000
genre     genre       1.000000
company   company     1.000000
runtime   runtime     1.000000
dtype: float64
```

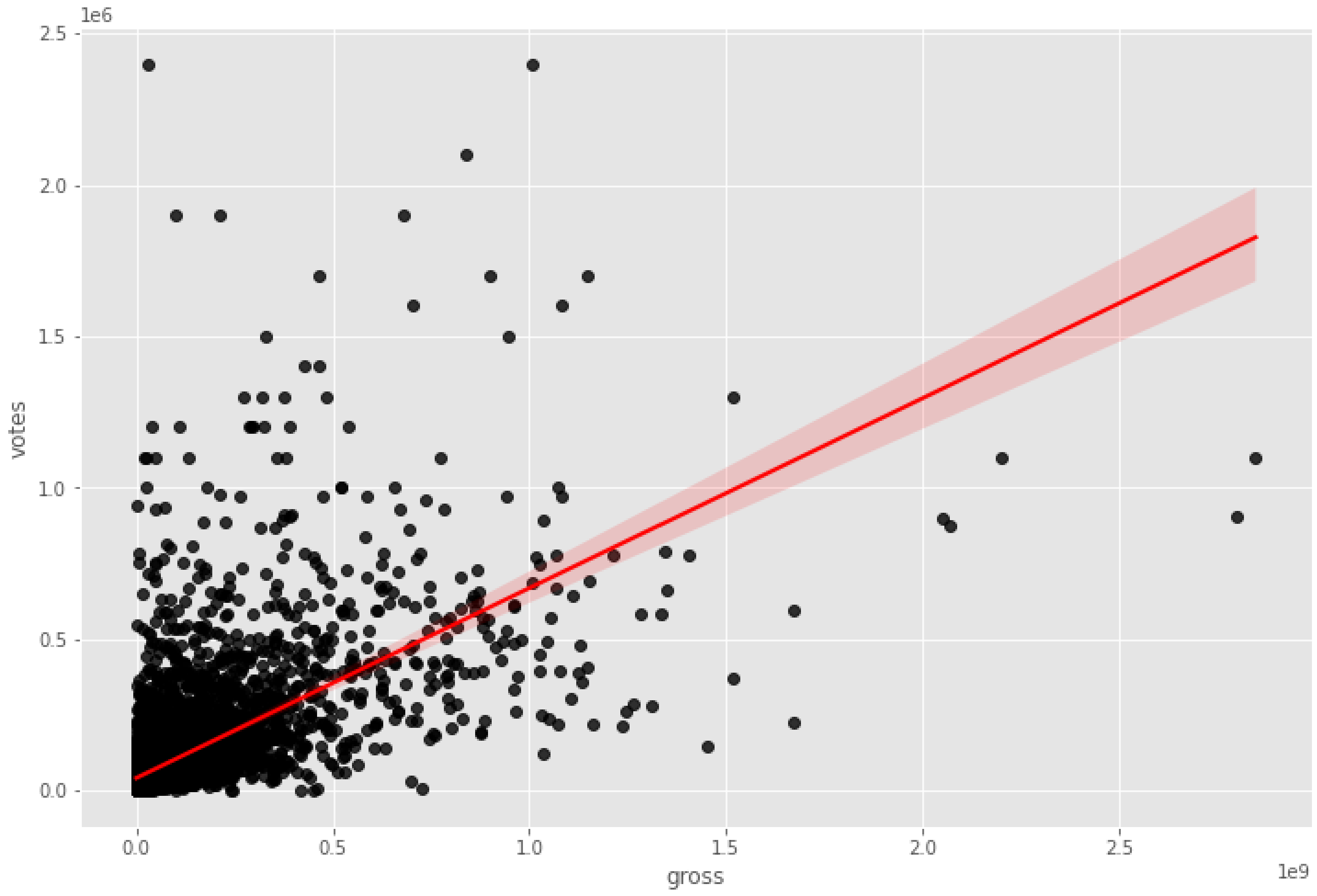
In [78]:
```
High_corr = sorted_pairs[(sorted_pairs)> 0.5]

High_corr
```

Out[78]:
```
gross     votes       0.630757
votes     gross       0.630757
budget    gross       0.740395
gross     budget      0.740395
rating    rating      1.000000
gross     gross       1.000000
budget    budget      1.000000
country   country     1.000000
star      star        1.000000
writer    writer      1.000000
director  director    1.000000
votes     votes       1.000000
score     score       1.000000
released  released    1.000000
year      year        1.000000
genre     genre       1.000000
company   company     1.000000
runtime   runtime     1.000000
dtype: float64
```

In [66]:
```
# Votes and budget Have the highest correlation to gross earnings
# Company has very litle correlation I was wrong
```

In [69]:
```
# here is the chart for gross vs vote
sns.regplot(x="gross", y="votes", data=df,scatter_kws={'color': 'black'},
line_kws={'color':'red'})
```

Out[69]:
```
<AxesSubplot:xlabel='gross', ylabel='votes'>
```


1e6
2.5
2.0
1.5
1.0
0.5
0.0
votes
0.0
0.5
1.0
1.5
2.0
2.5
gross
1e9