Introduction

This project will investigate if there is a strong correlation between the budget movies are given and gross revenue they receive.

The goals are to prepare the data, clean the data, followed by analysis with plots, and seek to explain the findings from the study.

Here are a few questions that this project will seek to answer:

Data sources

Data Source: https://www.kaggle.com/danielgrijalvas/movies This data was scraped from IMDb.

Information about the data...

There are 6820 movies in the dataset (220 movies per year, 1986-2016).

In [4]:
```
# Import libraries

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

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

%matplotlib inline
matplotlib.rcParams['figure.figsize'] = (12,8) # Adjusts the configurations of t
```

In [21]:
```
# Read in and look at the data

df = pd.read_csv('movies.csv')
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
```

Cleaning the Data

In [6]:

```
# Checking to see if there is any missing data.
# Loop through columns to check for missing data

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

```
name - 0.0%
rating - 0.010041731872717789%
genre - 0.0%
year - 0.0%
released - 0.00026082420448661763%
score - 0.00039123630672926445%
votes - 0.00039123630672926445%
director - 0.0%
writer - 0.00039123630672926445%
star - 0.00013041210224308815%
country - 0.00039123630672926445%
budget - 0.2831246739697444%
gross - 0.02464788732394366%
company - 0.002217005738132499%
runtime - 0.0005216484089723526%
```

In [7]:

```
# Data types for our columns

df.dtypes
```

Out[7]:

```
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 [9]:
```
# change data type of columns to make data cleaner

df['budget'] = df.budget.fillna(0)
df['budget'] = df['budget'].astype(int)

df['gross'] = df.gross.fillna(0)
df['gross'] = df['gross'].astype(int)

df['runtime'] = df.runtime.fillna(0)
df['runtime'] = df['runtime'].astype(int)

df.head()
```

Out[9]:

	name	rating	genre	year	released	score	votes	director	writer	st
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Ja Nichols
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Broo Shiel
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	Ma Ham
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robe Ha
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Che Cha

In [10]:
```
# Ordering the data by the gross revenue
df = df.sort_values(by=['gross'], inplace=False, ascending=False)
```

In [75]:
```
# Allows me to be able to scroll through df to see more info
pd.set_option('display.max_rows', None)
```

In [11]:
```
# Drop any duplicates

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

Out[11]:
```
7129                                thefyzz
5664                           micro_scope
6412    iDeal Partners Film Fund
4007                                i5 Films
6793                              i am OTHER
                        ...
3748                        1+2 Seisaku Iinkai
```

```
3024                          .406 Production
7525    "Weathering With You" Film Partners
4345        "DIA" Productions GmbH & Co. KG
7657                                    NaN
Name: company, Length: 2386, dtype: object
```

What columns are most correlated to gross revenue?

Budget? Company?

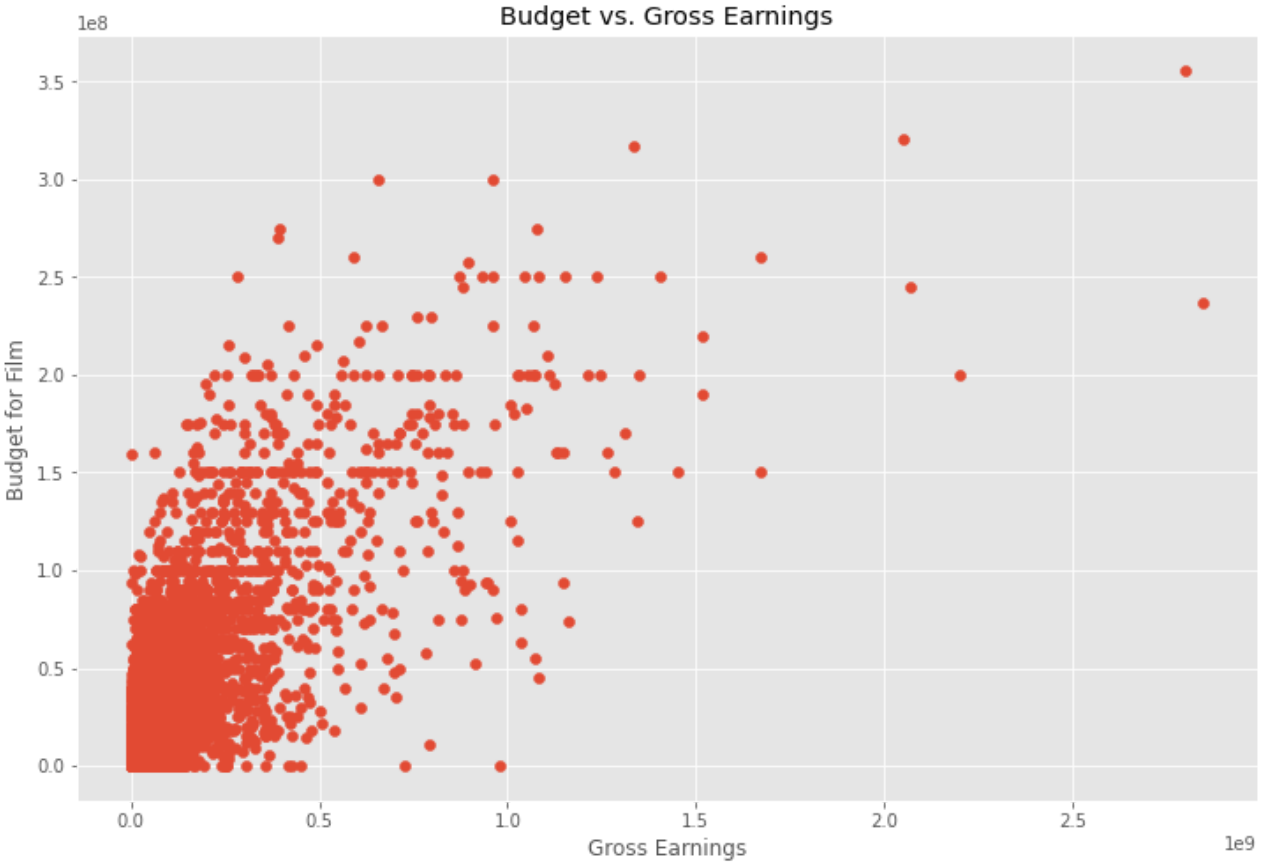
In [12]:

```
# Scatter plot with budget vs gross revenue

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

plt.title("Budget vs. Gross Earnings")
plt.xlabel("Gross Earnings")
plt.ylabel("Budget for Film")

plt.show()
```



In [15]:

```
df.head()
```

Out[15]:

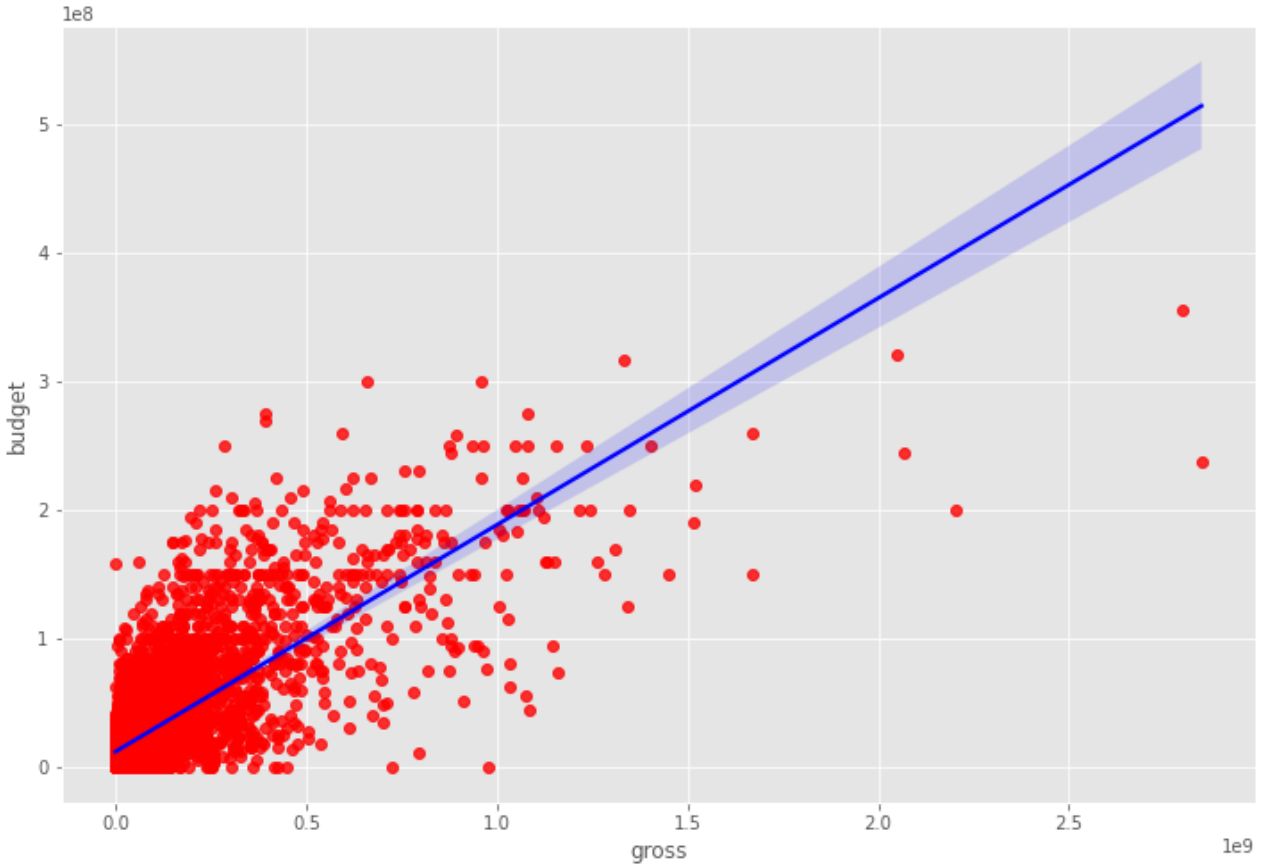
	name	rating	genre	year	released	score	votes	director	writer

	name	rating	genre	year	released	score	votes	director	writer	
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	James Cameron	Worthin
7445	Avengers: Endgame	PG-13	Action	2019	April 26, 2019 (United States)	8.4	903000.0	Anthony Russo	Christopher Markus	Ro Downe
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	James Cameron	Leon DiCa
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 Ri
7244	Avengers: Infinity War	PG-13	Action	2018	April 27, 2018 (United States)	8.4	897000.0	Anthony Russo	Christopher Markus	Ro Downe

In [16]:
```
# Plot budget vs gross using seaborn

sns.regplot(x = 'gross', y = 'budget', data=df, scatter_kws={"color": "red"}, li
```

Out[16]:
```
<AxesSubplot:xlabel='gross', ylabel='budget'>
```



Taking a closer look at correlation for all columns

In [17]:
```
df.corr(method='pearson') #pearson, kendall, spearman
```

Out[17]:

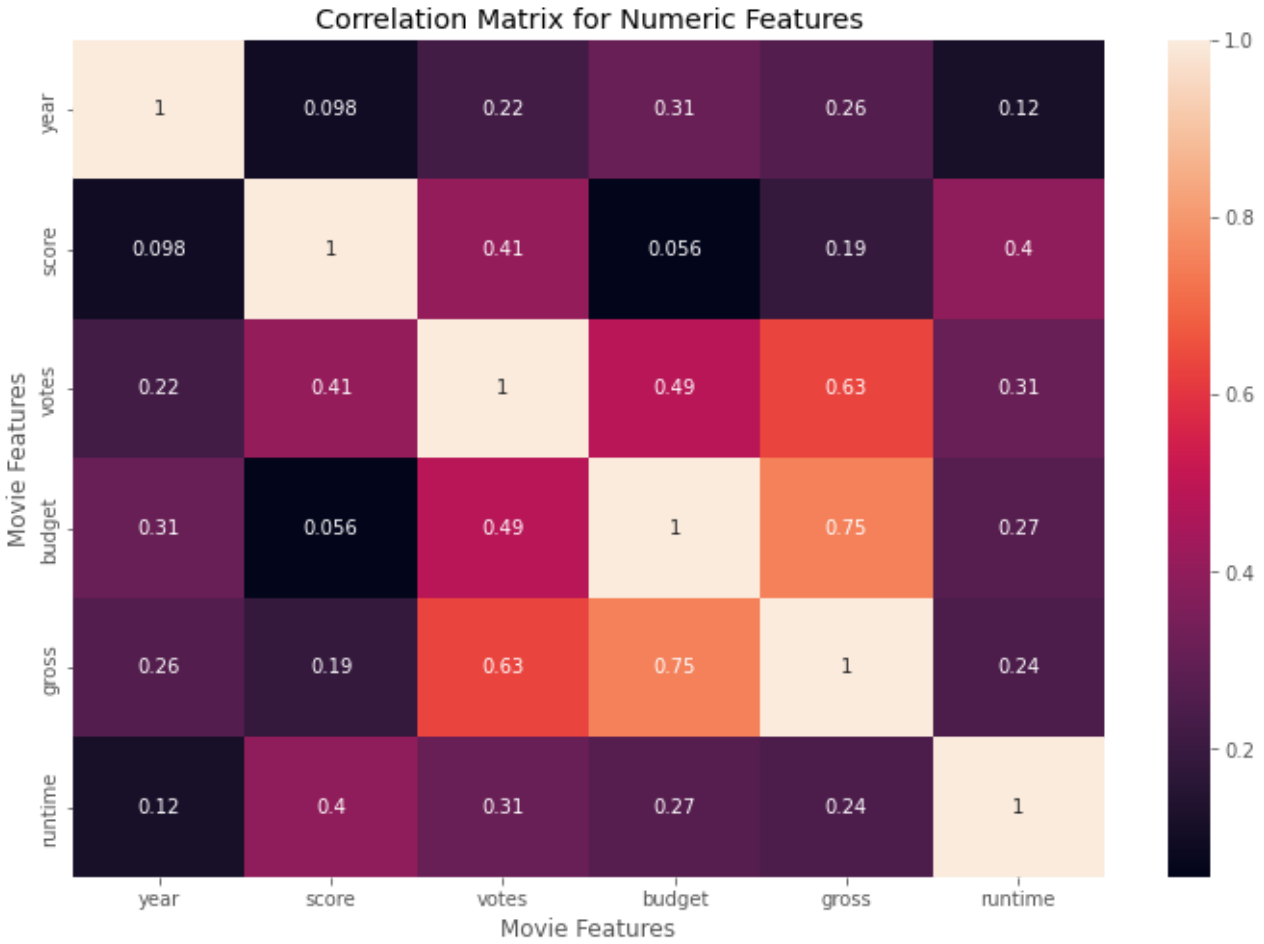
	year	score	votes	budget	gross	runtime
year	1.000000	0.097995	0.222945	0.309212	0.261900	0.116358
score	0.097995	1.000000	0.409182	0.055665	0.186392	0.398387
votes	0.222945	0.409182	1.000000	0.486862	0.632834	0.307074
budget	0.309212	0.055665	0.486862	1.000000	0.750157	0.268372
gross	0.261900	0.186392	0.632834	0.750157	1.000000	0.244339
runtime	0.116358	0.398387	0.307074	0.268372	0.244339	1.000000

In [ ]:
```
# High correlation between budget and gross
```

In [18]:
```
correlation_matrix = df.corr(method='pearson')

sns.heatmap(correlation_matrix, annot = True)
plt.title("Correlation Matrix for Numeric Features")
plt.xlabel("Movie Features")
plt.ylabel("Movie Features")
```

Out[18]:
```
Text(87.0, 0.5, 'Movie Features')
```



In [19]:

```
# Giving number values to category names
print(df.head())
```

```
                                      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

                         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

                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

          gross                company  runtime
5445  2847246203  Twentieth Century Fox      162
7445  2797501328         Marvel Studios      181
3045  2201647264  Twentieth Century Fox      194
```

```
6663  2069521700            Lucasfilm      138
7244  2048359754        Marvel Studios     149
```

In [20]:

```
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.head()
```

Out[20]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	
5445	533	5	0	2009	696	7.8	1100000.0	1155	1778	2334	55	2
7445	535	5	0	2019	183	8.4	903000.0	162	743	2241	55	3
3045	6896	5	6	1997	704	7.8	1100000.0	1155	1778	1595	55	2
6663	5144	5	0	2015	698	7.8	876000.0	1125	2550	524	55	2
7244	536	5	0	2018	192	8.4	897000.0	162	743	2241	55	3

In [22]:

```
df.head()
```

Out[22]:

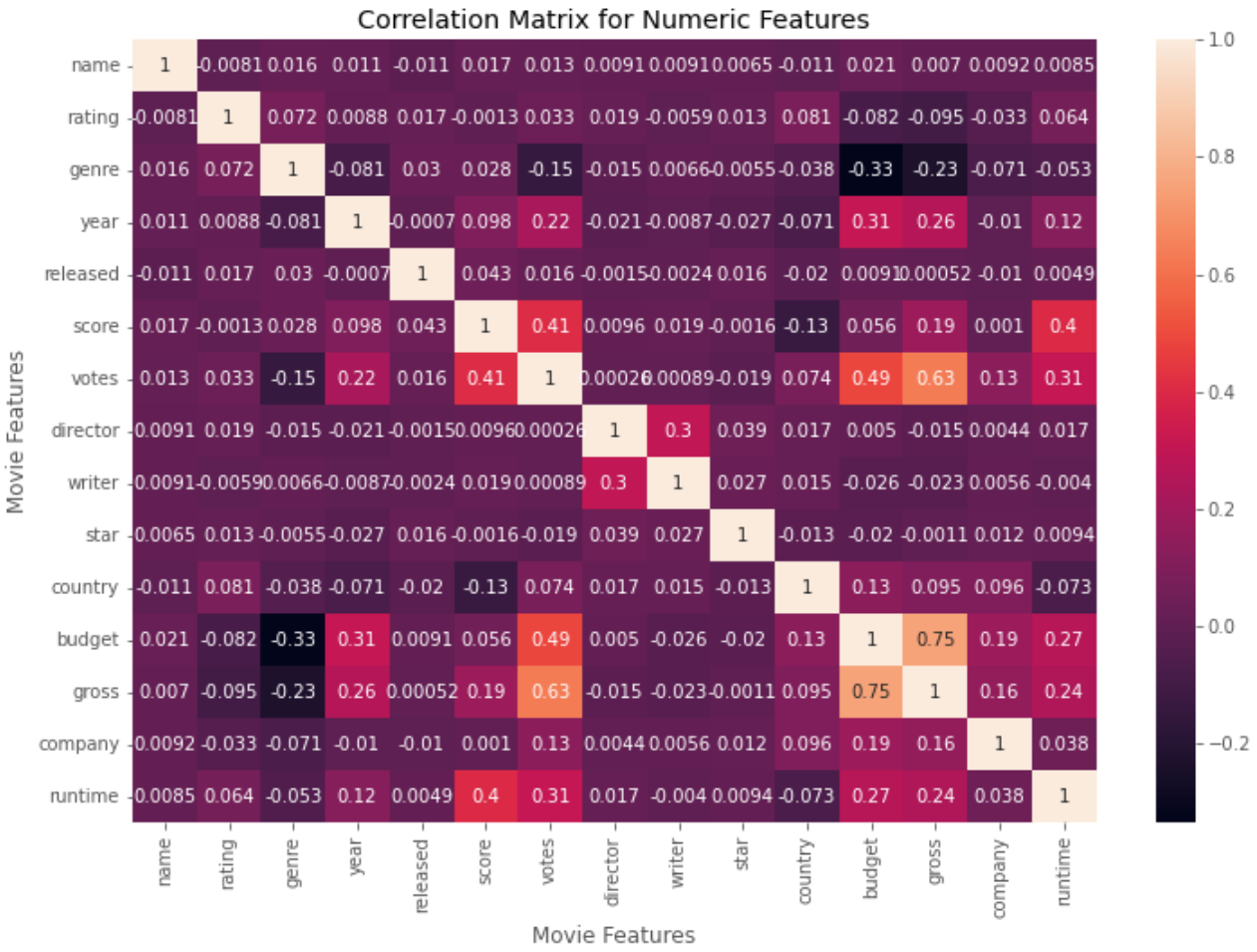
	name	rating	genre	year	released	score	votes	director	writer	st
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Ja Nichols
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Broo Shiel
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner	Leigh Brackett	Ma Ham
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robe Ha
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Che Cha

In [23]:

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

sns.heatmap(correlation_matrix, annot = True)
plt.title("Correlation Matrix for Numeric Features")
plt.xlabel("Movie Features")
plt.ylabel("Movie Features")
```

Out[23]: Text(87.0, 0.5, 'Movie Features')



In [24]:
```
df_numerized.corr()
```

Out[24]:

	name	rating	genre	year	released	score	votes	directo
name	1.000000	-0.008069	0.016355	0.011453	-0.011311	0.017097	0.013088	0.00907
rating	-0.008069	1.000000	0.072423	0.008779	0.016613	-0.001314	0.033225	0.01948
genre	0.016355	0.072423	1.000000	-0.081261	0.029822	0.027965	-0.145307	-0.01525
year	0.011453	0.008779	-0.081261	1.000000	-0.000695	0.097995	0.222945	-0.02079
released	-0.011311	0.016613	0.029822	-0.000695	1.000000	0.042788	0.016097	-0.00147
score	0.017097	-0.001314	0.027965	0.097995	0.042788	1.000000	0.409182	0.00955
votes	0.013088	0.033225	-0.145307	0.222945	0.016097	0.409182	1.000000	0.00026
director	0.009079	0.019483	-0.015258	-0.020795	-0.001478	0.009559	0.000260	1.00000
writer	0.009081	-0.005921	0.006567	-0.008656	-0.002404	0.019416	0.000892	0.29906
star	0.006472	0.013405	-0.005477	-0.027242	0.015777	-0.001609	-0.019282	0.03923
country	-0.010737	0.081244	-0.037615	-0.070938	-0.020427	-0.133348	0.073625	0.01749
budget	0.020548	-0.081939	-0.334021	0.309212	0.009145	0.055665	0.486862	0.00497
gross	0.006989	-0.095450	-0.234297	0.261900	0.000519	0.186392	0.632834	-0.01491
company	0.009211	-0.032943	-0.071067	-0.010431	-0.010474	0.001030	0.133204	0.00440

	name	rating	genre	year	released	score	votes	directo
runtime	0.008483	0.064133	-0.052914	0.116358	0.004852	0.398387	0.307074	0.01706

In [25]:
```
# Organize to see specific rows from correlation

correlation_mat = df_numerized.corr()

corr_pairs = correlation_mat.unstack()

corr_pairs
```

Out[25]:
```
name     name         1.000000
         rating      -0.008069
         genre        0.016355
         year         0.011453
         released    -0.011311
                          ...
runtime  country     -0.073319
         budget       0.268372
         gross        0.244339
         company      0.037537
         runtime      1.000000
Length: 225, dtype: float64
```

In [26]:
```
sorted_pairs = corr_pairs.sort_values()

sorted_pairs
```

Out[26]:
```
genre    budget    -0.334021
budget   genre     -0.334021
gross    genre     -0.234297
genre    gross     -0.234297
votes    genre     -0.145307
                       ...
year     year       1.000000
genre    genre      1.000000
rating   rating     1.000000
company  company    1.000000
runtime  runtime    1.000000
Length: 225, dtype: float64
```

In [27]:
```
high_corr = sorted_pairs[((sorted_pairs) > 0.5)]
high_corr
```

Out[27]:
```
votes     gross       0.632834
gross     votes       0.632834
          budget      0.750157
budget    gross       0.750157
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
```

Conclusion

Votes and budget have the highest correlation to gross earnings
Company name has a low correlation