In [1]:
```
#Import the 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) #Adjusting the configuration of the plots we

#Read the data
df = pd.read_csv(r'C:\Users\punee\OneDrive\Documents\Datasets\movies.csv')
```

In [2]:
```
df.head()
```

Out[2]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	1900
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	450
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	1800
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	350
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	600

In [3]:
```
df.isnull().sum()
```

Out[3]:
```
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
```

In [4]:
```
for col in df.columns:
    missing = np.mean(df[col].isnull())
    print(missing)
```

```
0.0
0.010041731872717789
0.0
0.0
0.0002608242044861763
0.0003912363067292645
0.0003912363067292645
0.0
0.0003912363067292645
0.00013041210224308815
0.0003912363067292645
0.2831246739697444
0.02464788732394366
0.002217005738132499
0.0005216484089723526
```

In [5]:
```
df = df.dropna()
```

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
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
```

In [7]:
```
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 [8]:
```
#Change the data type

df['budget'] = df['budget'].astype('int64')
df['gross'] = df['gross'].astype('int64')
```

In [9]:
```
df.head()
```

Out[9]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	bu
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	1900
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	450
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	1800
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	350
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	600

In [10]:
```
# create correct year column
df['yearcorrect'] = df['year'].astype(str).str[:4]
```

In [11]:
```
df.head()
```

Out[11]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	bu
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	1900
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	450
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	1800
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	350
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	600

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

Out[12]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	James Cameron	Sam Worthington	United States	237(
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	356(
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	James Cameron	Leonardo DiCaprio	United States	200(
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	245(
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	321(

In [13]:
```
pd.set_option('display.max_rows', None)
```

In [14]:
```
#drop as duplicates

df.drop_duplicates().head()
```

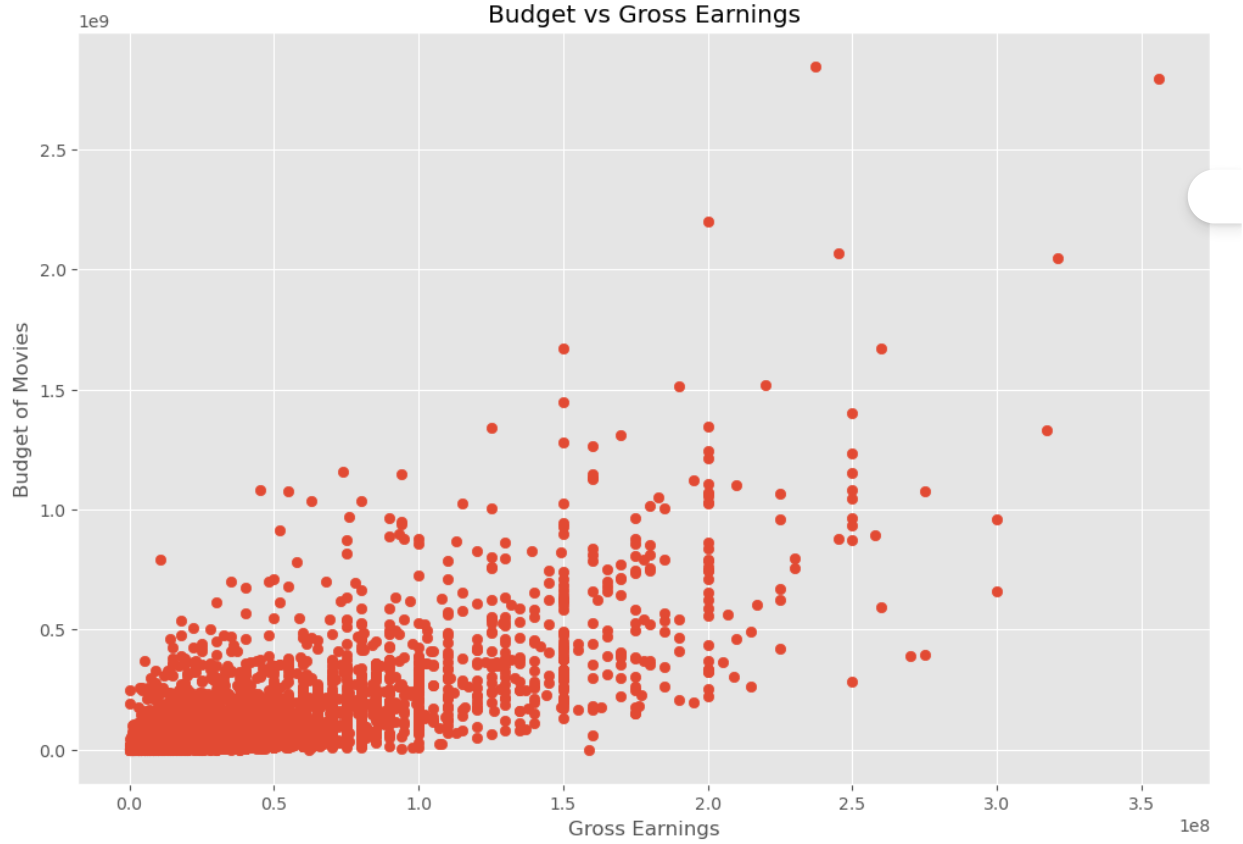
Out[14]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	bu
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	1900
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	450
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	1800
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	350
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	600

In [15]:
```
#scatter plot budget vs gross revenue

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

plt.title('Budget vs Gross Earnings')
plt.xlabel('Gross Earnings')
plt.ylabel('Budget of Movies')
plt.show()
```



In [16]:
```
df.head()
```

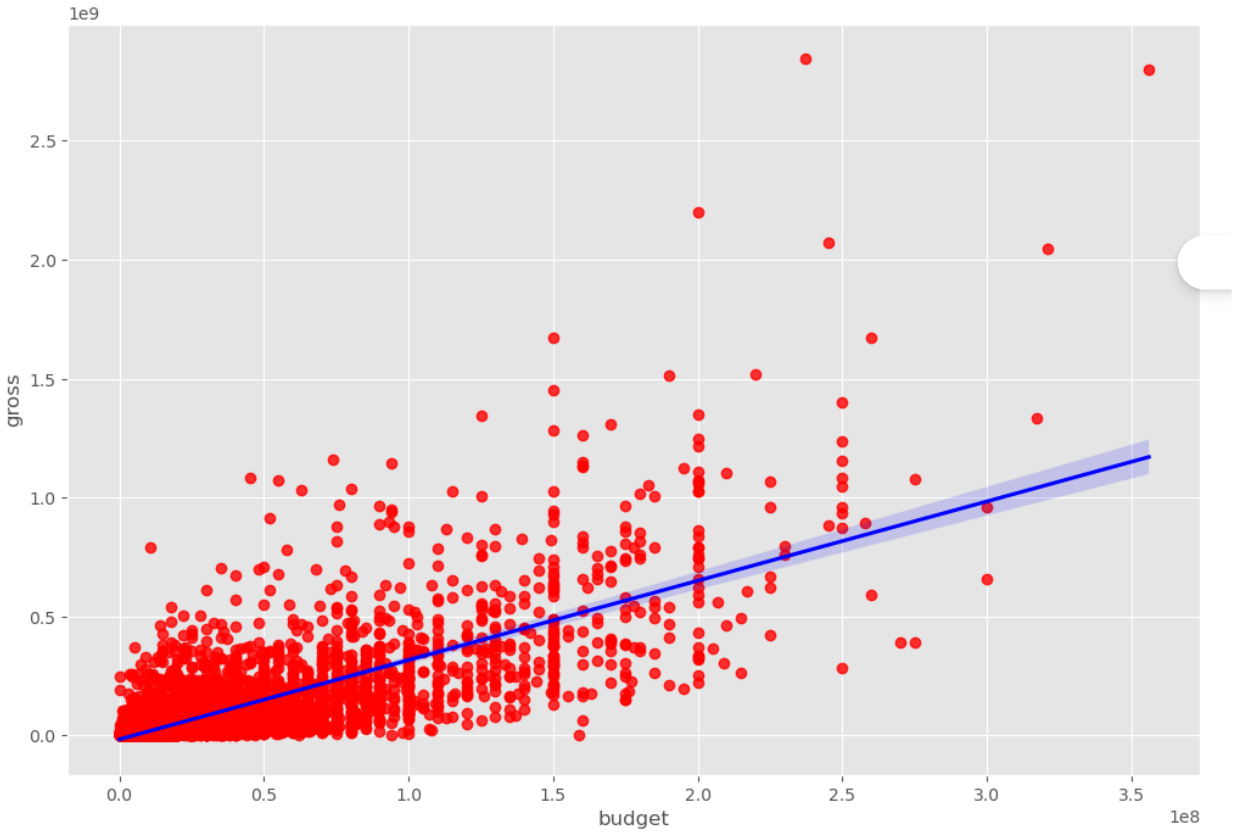
Out[16]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	bu
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	1900
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	450
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	1800
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	350
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	600

In [17]:
```
#plot the budget vs gross using seaborn

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

Out[17]: <Axes: xlabel='budget', ylabel='gross'>



In [18]:
```
# lets start looking at correlations
df.corr(method='pearson')  #pearson , kendall , spearman
```

```
C:\Users\punee\AppData\Local\Temp\ipykernel_9424\4278171647.py:2: FutureWarning: The defau
lt value of numeric_only in DataFrame.corr is deprecated. In a future version, it will def
ault to False. Select only valid columns or specify the value of numeric_only to silence t
his warning.
  df.corr(method='pearson')  #pearson , kendall , spearman
```

Out[18]:

	year	score	votes	budget	gross	runtime
year	1.000000	0.056386	0.206021	0.327722	0.274321	0.075077
score	0.056386	1.000000	0.474256	0.072001	0.222556	0.414068
votes	0.206021	0.474256	1.000000	0.439675	0.614751	0.352303
budget	0.327722	0.072001	0.439675	1.000000	0.740247	0.318695
gross	0.274321	0.222556	0.614751	0.740247	1.000000	0.275796
runtime	0.075077	0.414068	0.352303	0.318695	0.275796	1.000000

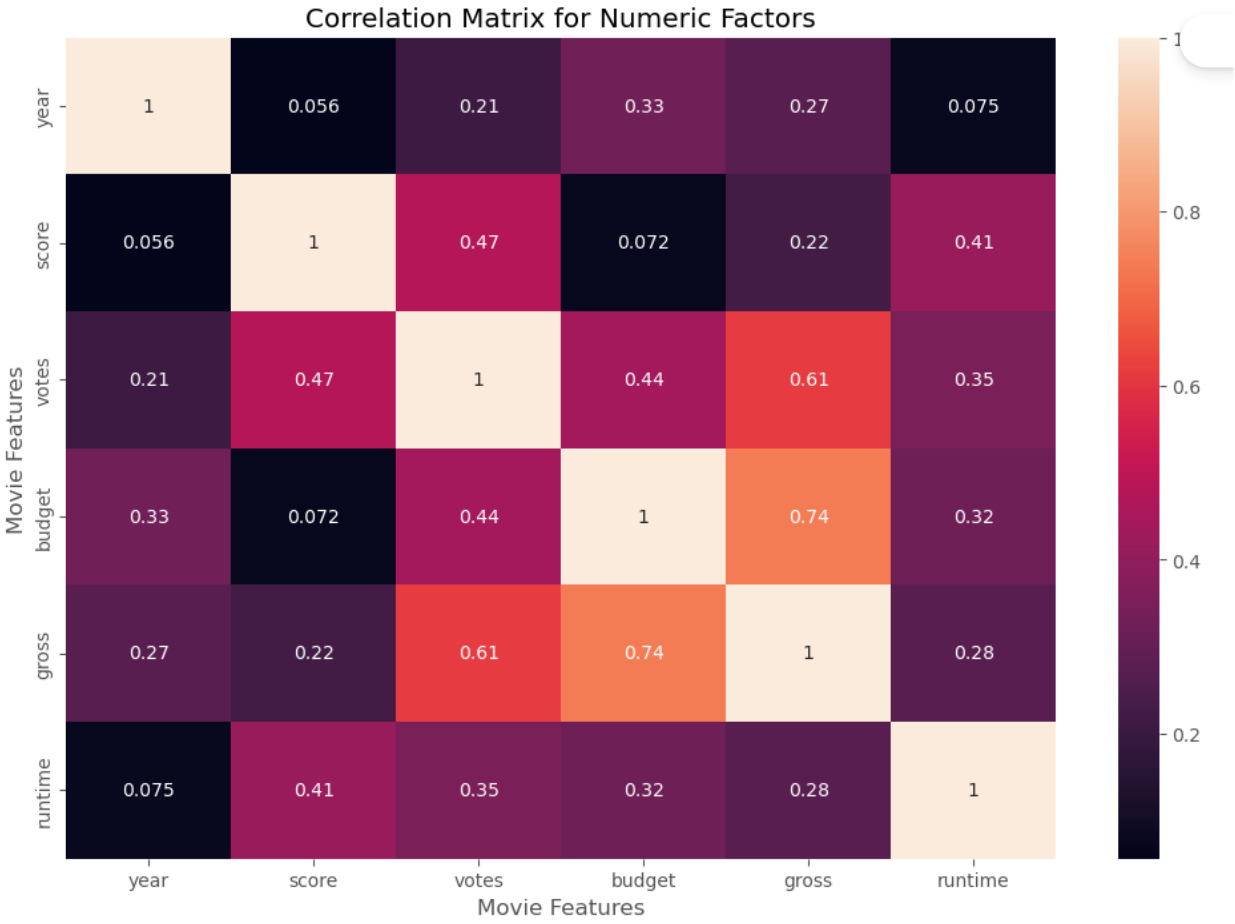
In [19]:
```
#high corr betwween budget and correlation
```

In [20]:
```
correlation_matrix = df.corr(method='pearson')

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

```
C:\Users\punee\AppData\Local\Temp\ipykernel_9424\2693371779.py:1: FutureWarning: The defau
lt value of numeric_only in DataFrame.corr is deprecated. In a future version, it will def
ault to False. Select only valid columns or specify the value of numeric_only to silence t
his warning.
  correlation_matrix = df.corr(method='pearson')
```

Out[20]:
```
Text(120.72222222222221, 0.5, 'Movie Features')
```



In [21]:
```
#Looks at the company

df.head()
```

Out[21]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	bu
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	1900
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	450
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	1[illegible]
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	350
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	600

In [23]:
```
#Numerizing all the variables
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[23]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	co
0	4692	6	6	1980	1304	8.4	927000.0	1795	2832	699	46	19000000	46998772	
1	3929	6	1	1980	1127	5.8	65000.0	1578	1158	214	47	4500000	58853106	
2	3641	4	0	1980	1359	8.7	1200000.0	757	1818	1157	47	18000000	538375067	
3	204	4	4	1980	1127	7.7	221000.0	889	1413	1474	47	3500000	83453539	
4	732	6	4	1980	1170	7.3	108000.0	719	351	271	47	6000000	39846344	

In [24]:
```
df = df.sort_values(by=['gross'],inplace=False,ascending=False)
```

In [25]:
```
df.head()
```

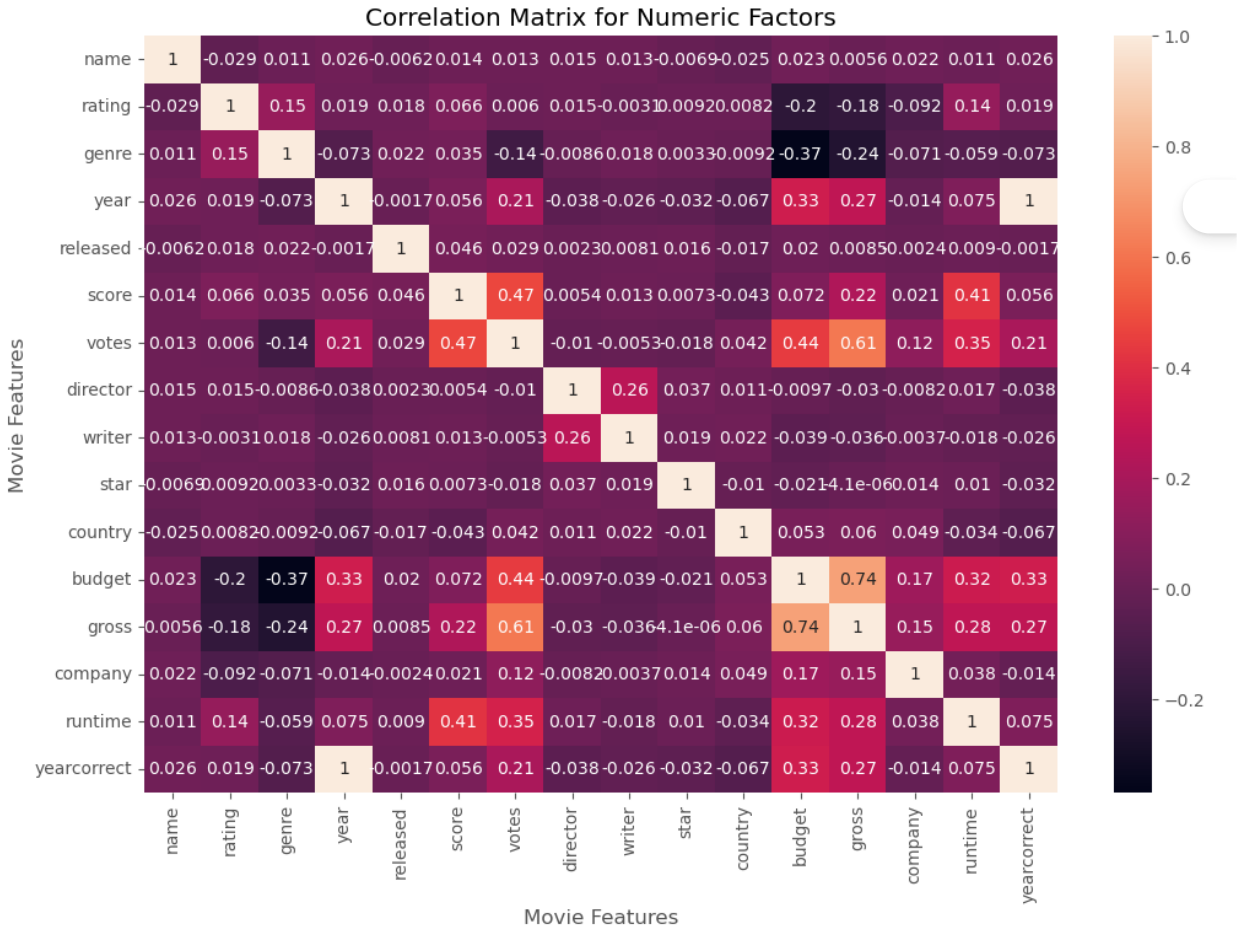
Out[25]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gros
5445	386	5	0	2009	527	7.8	1100000.0	785	1263	1534	47	237000000	28472462(
7445	388	5	0	2019	137	8.4	903000.0	105	513	1470	47	356000000	279750132
3045	4909	5	6	1997	534	7.8	1100000.0	785	1263	1073	47	200000000	220164726
6663	3643	5	0	2015	529	7.8	876000.0	768	1806	356	47	245000000	20695217(
7244	389	5	0	2018	145	8.4	897000.0	105	513	1470	47	321000000	204835975

In [26]:
```
correlation_matrix = df.corr(method='pearson')

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

Out[26]:
```
Text(120.72222222222221, 0.5, 'Movie Features')
```



In [27]:
```
df_numerized.corr()
```

Out[27]:

	name	rating	genre	year	released	score	votes	director	writer	
name	1.000000	-0.029234	0.010996	0.025542	-0.006152	0.014450	0.012615	0.015246	0.012880	-0.006
rating	-0.029234	1.000000	0.147796	0.019499	0.018083	0.065983	0.006031	0.014656	-0.003149	0.009
genre	0.010996	0.147796	1.000000	-0.073167	0.022142	0.035106	-0.135990	-0.008553	0.017578	0.003
year	0.025542	0.019499	-0.073167	1.000000	-0.001740	0.056386	0.206021	-0.038354	-0.025908	-0.032
released	-0.006152	0.018083	0.022142	-0.001740	1.000000	0.045874	0.028833	0.002308	0.008072	0.015
score	0.014450	0.065983	0.035106	0.056386	0.045874	1.000000	0.474256	0.005413	0.012843	0.007
votes	0.012615	0.006031	-0.135990	0.206021	0.028833	0.474256	1.000000	-0.010376	-0.005316	-0.017
director	0.015246	0.014656	-0.008553	-0.038354	0.002308	0.005413	-0.010376	1.000000	0.261735	(
writer	0.012880	-0.003149	0.017578	-0.025908	0.008072	0.012843	-0.005316	0.261735	1.000000	0.018
star	-0.006882	0.009196	0.003341	-0.032157	0.015706	0.007296	-0.017638	0.036593	0.018520	1.000
country	-0.025490	0.008230	-0.009164	-0.066748	-0.017228	-0.043051	0.041551	0.011133	0.022488	-0.009
budget	0.023392	-0.203946	-0.368523	0.327722	0.019952	0.072001	0.439675	-0.009662	-0.039466	-0.021
gross	0.005639	-0.181906	-0.244101	0.274321	0.008501	0.222556	0.614751	-0.029560	-0.035885	-0.000
company	0.021697	-0.092357	-0.071334	-0.014333	-0.002407	0.020656	0.118470	-0.008223	-0.003697	0.014
runtime	0.010850	0.140792	-0.059237	0.075077	0.008975	0.414068	0.352303	0.017433	-0.017561	0.010
yearcorrect	0.025542	0.019499	-0.073167	1.000000	-0.001740	0.056386	0.206021	-0.038354	-0.025908	-0.032

In [28]:
```
correlation_mat = df_numerized.corr()
corr_pairs = correlation_mat.unstack()

corr_pairs
```

Out[28]:
```
name         name           1.000000
             rating        -0.029234
             genre          0.010996
             year           0.025542
             released      -0.006152
             score          0.014450
             votes          0.012615
             director       0.015246
             writer         0.012880
             star          -0.006882
             country       -0.025490
             budget         0.023392
             gross          0.005639
             company        0.021697
             runtime        0.010850
             yearcorrect    0.025542
rating       name          -0.029234
             rating         1.000000
             genre          0.147796
```

In [29]:
```
sorted_pairs = corr_pairs.sort_values()
sorted_pairs
```

```
year         votes          0.206021
score        gross          0.222556
gross        score          0.222556
director     writer         0.261735
writer       director       0.261735
year         gross          0.274321
gross        year           0.274321
             yearcorrect    0.274321
yearcorrect  gross          0.274321
runtime      gross          0.275796
gross        runtime        0.275796
budget       runtime        0.318695
runtime      budget         0.318695
yearcorrect  budget         0.327722
budget       yearcorrect    0.327722
year         budget         0.327722
budget       year           0.327722
runtime      votes          0.352303
votes        runtime        0.352303
score        runtime        0.414068
```

In [30]:
```
high_corr = sorted_pairs[(sorted_pairs)>0.4]
high_corr
```

Out[30]:
```
score        runtime        0.414068
runtime      score          0.414068
votes        budget         0.439675
budget       votes          0.439675
score        votes          0.474256
votes        score          0.474256
gross        votes          0.614751
votes        gross          0.614751
budget       gross          0.740247
gross        budget         0.740247
name         name           1.000000
company      company        1.000000
rating       rating         1.000000
genre        genre          1.000000
year         year           1.000000
released     released       1.000000
score        score          1.000000
runtime      runtime        1.000000
votes        votes          1.000000
writer       writer         1.000000
star         star           1.000000
country      country        1.000000
budget       budget         1.000000
gross        gross          1.000000
director     director       1.000000
yearcorrect  yearcorrect    1.000000
             year           1.000000
year         yearcorrect    1.000000
dtype: float64
```

In [31]:
```
#votes and budget has a positive correlation on gross eanings
#scores andd runtime has a positive correlation on the company's earnings
```