In [ ]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib
plt.style.use('ggplot')
from matplotlib.pyplot
```

In [ ]:
```
df = pd.read_csv(r'C:\Users\FATEEMADEEN\Desktop\movies.csv')
Load dataset
```

In [ ]:
```
for col in df.columns:
    pct_missing = np.mean(df[col].isnull())
    print('{} - {}%'.format(col, round(pct_missing*100)))
```

In [33]:
```
df = df.dropna()
```

In [35]:
```
for col in df.columns:
    pct_missing = np.mean(df[col].isnull())
    print('{} - {}%'.format(col, round(pct_missing*100)))
```

```
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%
```

In [37]:
```
df.dtypes
df['budget'] = df['budget'].astype('int64')
df['gross'] = df['gross'].astype('int64')
```

In [39]:
```
df = df.sort_values(by=['gross'], inplace=False, ascending=False)
```

In [43]:
```
df.drop_duplicates()
```

Out[43]:

	name	rating	genre	year	released	score	votes	director	writ
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	Jam Camer
7445	Avengers: Endgame	PG-13	Action	2019	April 26, 2019 (United States)	8.4	903000.0	Anthony Russo	Christoph Mark
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	Jam Camer
6663	Star Wars: Episode VII - The Force Awakens	PG-13	Action	2015	December 18, 2015 (United States)	7.8	876000.0	J.J. Abrams	Lawren Kasd
7244	Avengers: Infinity War	PG-13	Action	2018	April 27, 2018 (United States)	8.4	897000.0	Anthony Russo	Christoph Mark
...	...	...	...	...	...	...	...	...	
5640	Tanner Hall	R	Drama	2009	January 15, 2015 (Sweden)	5.8	3500.0	Francesca Gregorini	Tatiana v Fürstenbe
2434	Philadelphia Experiment II	PG-13	Action	1993	June 4, 1994 (South Korea)	4.5	1900.0	Stephen Cornwell	Wallace Benne
3681	Ginger Snaps	Not Rated	Drama	2000	May 11, 2001 (Canada)	6.8	43000.0	John Fawcett	Kar Walt
272	Parasite	R	Horror	1982	March 12, 1982 (United States)	3.9	2300.0	Charles Band	Alan Adl
3203	Trojan War	PG-13	Comedy	1997	October 1, 1997 (Brazil)	5.7	5800.0	George Huang	Andy Bu

5421 rows × 15 columns

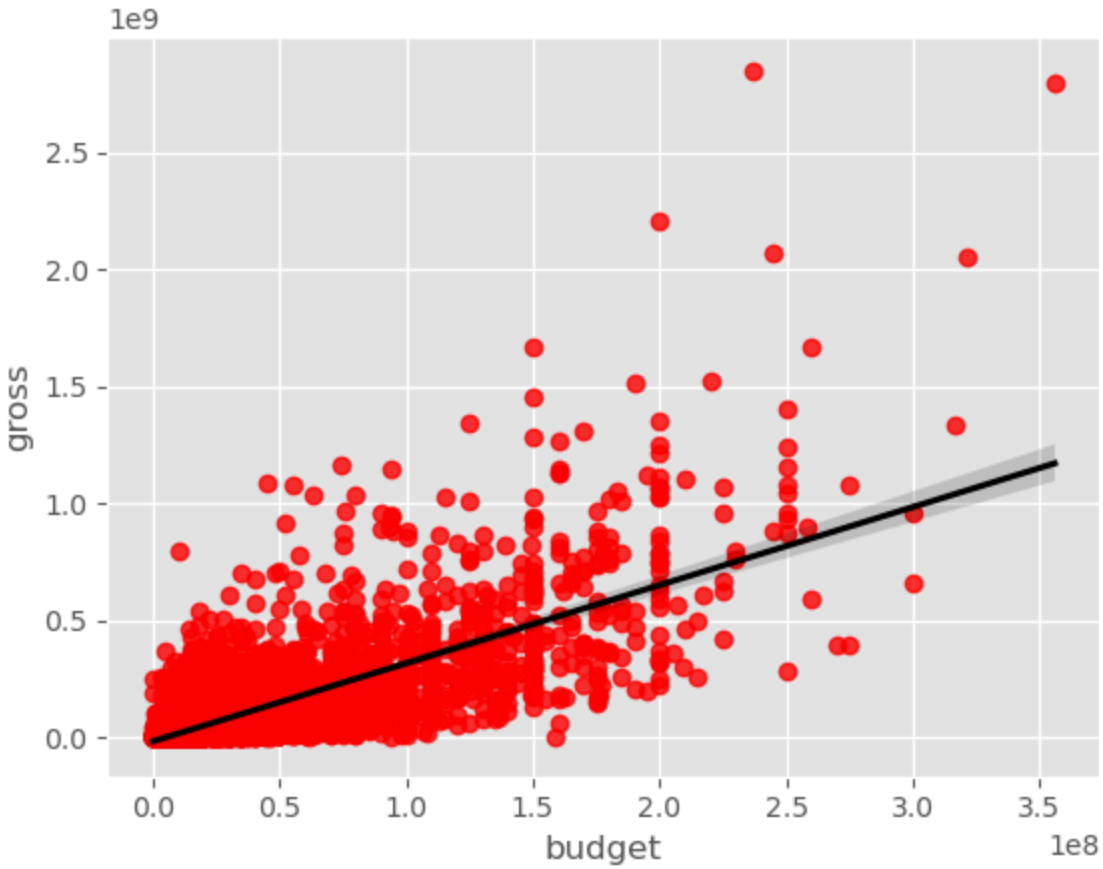
In [45]:
```
y = df['budget']
x = df['gross']
```

```
correlation = y.corr(x)
print('Correlation between Budget and Gross:', correlation)
```

```
Correlation between Budget and Gross: 0.7402465439219624
```

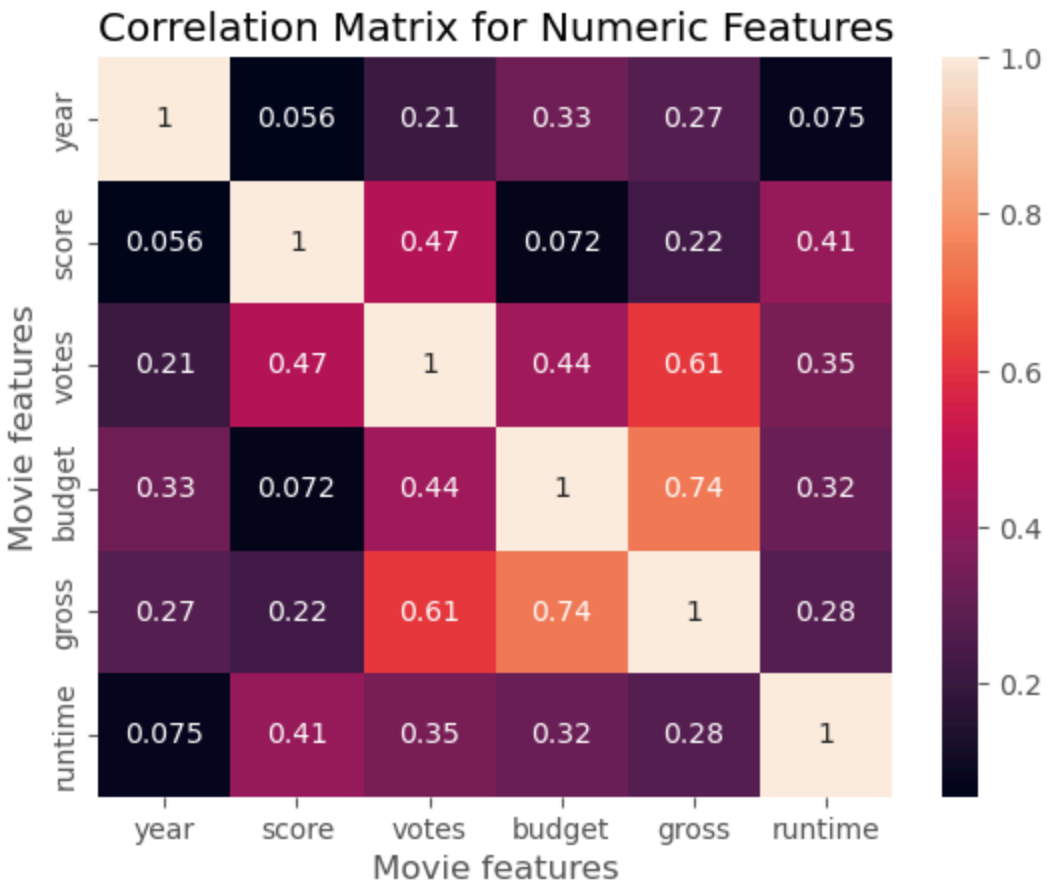
In [49]:
```
#Regression Plot: Added a regression line to the scatter plot for better analysis:
sns.regplot(x='budget', y='gross', data=df, scatter_kws={'color': 'red'}, line_kws=
```

Out[49]: <Axes: xlabel='budget', ylabel='gross'>



In [59]:
```
numeric_df = df.select_dtypes(include=['number'])
correlation_matrix = numeric_df.corr()

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
Movie features
1
0.056
0.21
0.33
0.27
0.075
0.056
1
0.47
0.072
0.22
0.41
0.21
0.47
1
0.44
0.61
0.35
0.33
0.072
0.44
1
0.74
0.32
0.27
0.22
0.61
0.74
1
0.28
0.075
0.41
0.35
0.32
0.28
1
1.0
0.8
0.6
0.4
0.2

In [ ]: