In [2]:

```
#Import libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
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

#Read in the data
df = pd.read_csv(r"C:\Users\ogumus\OneDrive\Desktop\movies.csv")
```

In [3]:

```
df.head()
```

Out[3]:

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

In [4]:

```
#Dropping NA values

df = df.dropna()
```

In [5]:

```
#Verifying if there's any missing data in the dataset

for col in d-F.columns:
    pct_missing = np.mean(d-F[col].isnull())
    print("{} - {}%".-Format(col,pct_missing))
```

```
name - 0.0%
rating - 0.0%
genre - 0.0%
year - 0.0%
released - 0.0%
score - 0.0%
votes - 0.0%
director - 0.0%
writer - 0.0%
star - 0.0%
country - 0.0%
budget - 0.0%
gross - 0.0%
company - 0.0%
runtime - 0.0%
```

In [6]:

```
#Datatypes for our columns

d-F.dtypes
```

Out[6]:

```
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 [7]:

```
d f
```

Out[7]:

	name	rating	genre	year	released	score	votes	director	writer
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole
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
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray
...	...	...	...	...	...	...	...	...	...
7648	Bad Boys for Life	R	Action	2020	January 17, 2020 (United States)	6.6	140000.0	Adil El Arbi	Peter Craig
7649	Sonic the Hedgehog	PG	Action	2020	February 14, 2020 (United States)	6.5	102000.0	Jeff Fowler	Pat Casey
7650	Dolittle	PG	Adventure	2020	January 17, 2020 (United States)	5.6	53000.0	Stephen Gaghan	Stephen Gaghan
7651	The Call of the Wild	PG	Adventure	2020	February 21, 2020 (United States)	6.8	42000.0	Chris Sanders	Michael Green
7652	The Eight Hundred	Not Rated	Action	2020	August 28, 2020 (United States)	6.8	3700.0	Hu Guan	Hu Guan

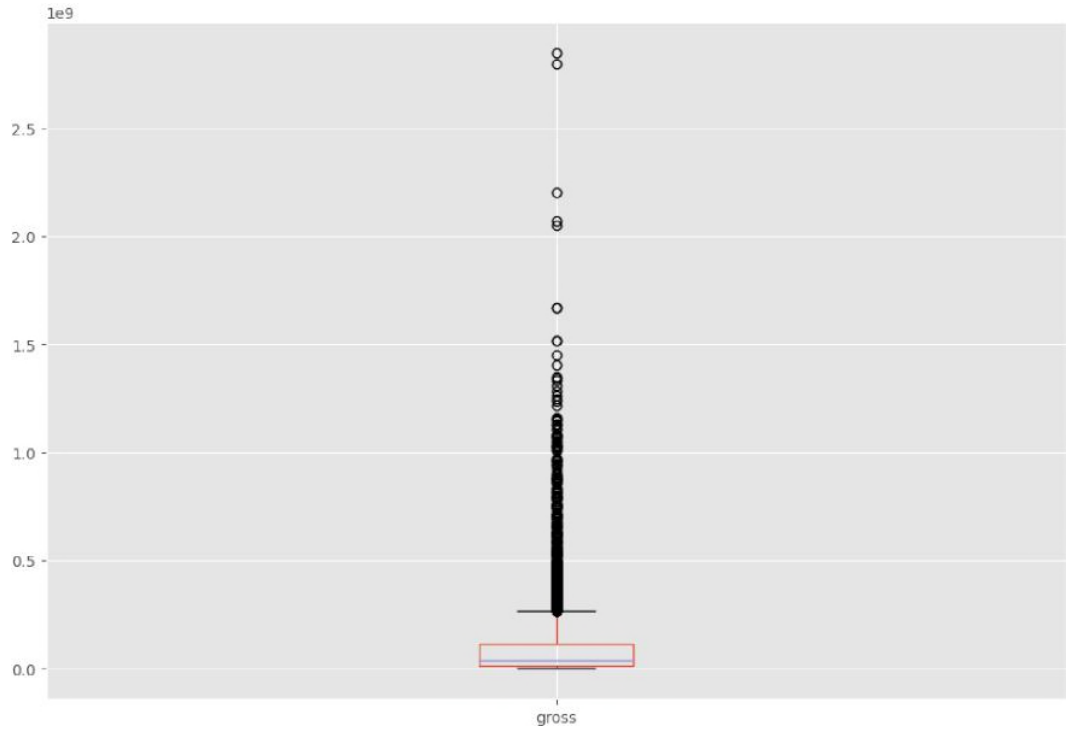
5421 rows × 15 columns

In [8]:

```
#Are there any outliers?

d-F.boxplot(column=["gross"])
```

Out[8]:

<Axes: >



In [9]:

```
#Creating correct year column

d-F['yearcorrect'] = d-F['released'].astype(str).str.split().str[2]
```

In [10]:

```
d f
```

Out[10]:

	name	rating	genre	year	released	score	votes	director	writer
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole
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
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray
...	...	...	...	...	...	...	...	...	...
7648	Bad Boys for Life	R	Action	2020	January 17, 2020 (United States)	6.6	140000.0	Adil El Arbi	Peter Craig
7649	Sonic the Hedgehog	PG	Action	2020	February 14, 2020 (United States)	6.5	102000.0	Jeff Fowler	Pat Casey
7650	Dolittle	PG	Adventure	2020	January 17, 2020 (United States)	5.6	53000.0	Stephen Gaghan	Stephen Gaghan
7651	The Call of the Wild	PG	Adventure	2020	February 21, 2020 (United States)	6.8	42000.0	Chris Sanders	Michael Green
7652	The Eight Hundred	Not Rated	Action	2020	August 28, 2020 (United States)	6.8	3700.0	Hu Guan	Hu Guan

5421 rows × 16 columns

In [11]:

```
#Ordering the data
df = df.sort_values("gross",ascending=False)
```

In [12]:

```
#Displaying the whole dataset
pd.set_option("display.max_rows",None)
```

In [13]:

```
df
```

Out[13]:

	name	rating	genre	year	released	score	votes	director	
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	Ca
7445	Avengers: Endgame	PG-13	Action	2019	April 26, 2019 (United States)	8.4	903000.0	Anthony Russo	Chris M
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	Ca
6663	Star Wars: Episode VII - The Force Awakens	PG-13	Action	2015	December 18, 2015 (United States)	7.8	876000.0	J.J. Abrams	Law K

In [14]:

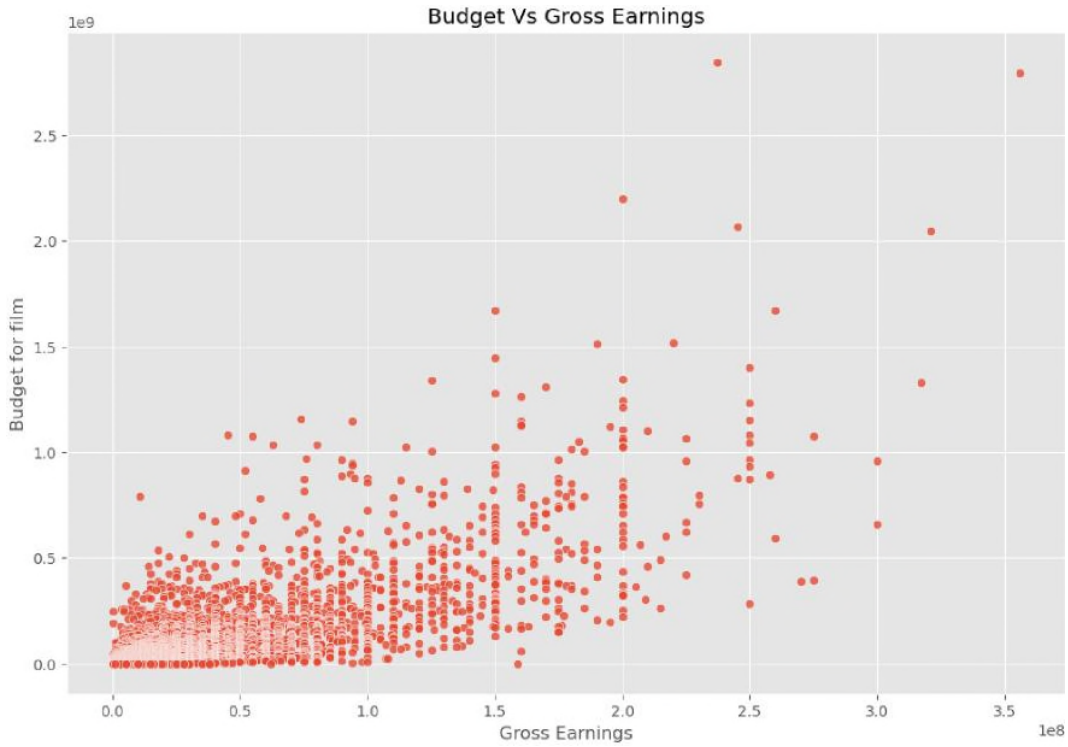
```
#Drop any duplicates

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

Out[14]:

```
7129                                    thefyzz
5664                                micro_scope
4007                                   i5 Films
6793                                i am OTHER
6420                                       erbp
3776                                double A Films
3330                Zucker Brothers Productions
520                            Zoetrope Studios
2213                          Zeta Entertainment
3698                     Zentropa Entertainments
1180                        Zenith Entertainment
5180                              Zazen Produções
1321                    Zanuck/Brown Productions
1329                Zacharias-Buhai Productions
789                   Young Sung Production Co.
5125                 Young Hannibal Productions
5499                                Yellow Bird
4618                                Yash Raj Films
```

In [31]:

```
#Scatter plot with budget Vs gross

sns.scatterplot(data=df,x="budget",y="gross",alpha=0.8)
plt.title("Budget Vs Gross Earnings")
plt.xlabel("Gross Earnings")
plt.ylabel("Budget for film")
plt.show()
```



In [16]:

```
df.head()
```

Out[16]:

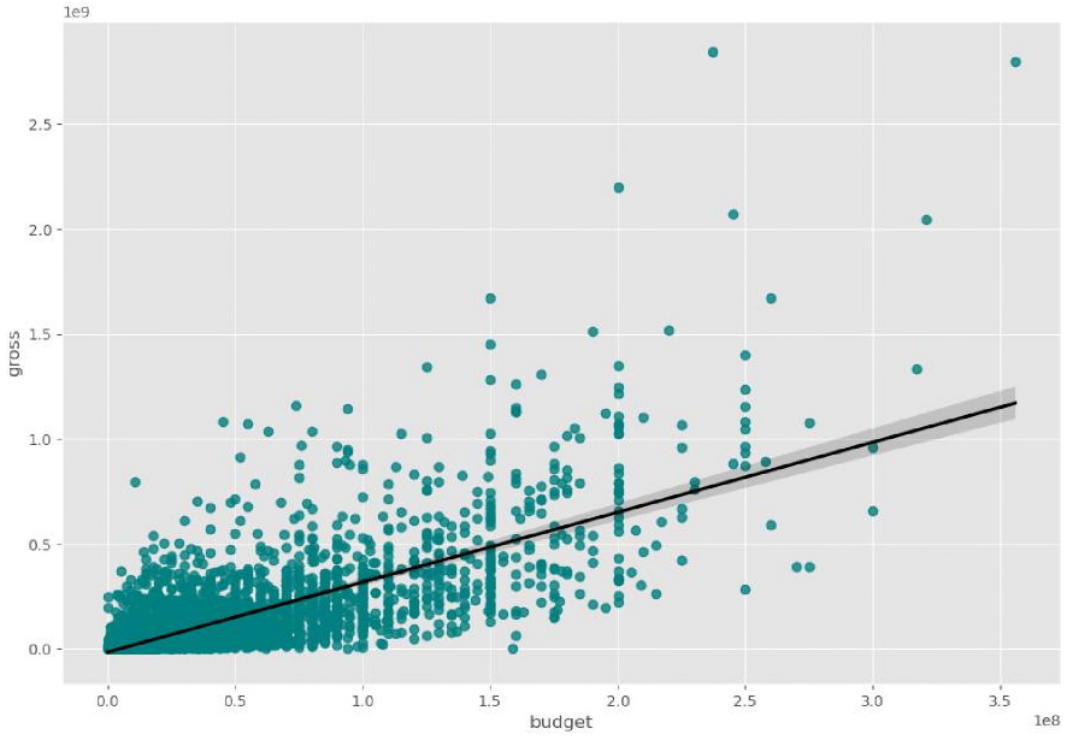
	name	rating	genre	year	released	score	votes	director	writer	
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	James Cameron	Wo
7445	Avengers: Endgame	PG-13	Action	2019	April 26, 2019 (United States)	8.4	903000.0	Anthony Russo	Christopher Markus	Do
3045	Titanic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

In [17]:

```
#Plot budget Vs Gross using Seaborn

sns.regplot(data=df,x="budget",y="gross",scatter_kws={"color":"teal"},line_kws={"color":
```

Out[17]:

```
<Axes: xlabel='budget', ylabel='gross'>
```



In [18]:

```
#Correlation matrix between all numeric columns

df.corr(method="pearson", numeric_only=True)
```

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]:

```
df.corr(method="kendall", numeric_only=True)
```

Out[19]:

	year	score	votes	budget	gross	runtime
year	1.000000	0.039389	0.296512	0.220833	0.239539	0.064824
score	0.039389	1.000000	0.350185	-0.006406	0.124943	0.292254
votes	0.296512	0.350185	1.000000	0.346274	0.553625	0.205344
budget	0.220833	-0.006406	0.346274	1.000000	0.512057	0.231278
gross	0.239539	0.124943	0.553625	0.512057	1.000000	0.176979
runtime	0.064824	0.292254	0.205344	0.231278	0.176979	1.000000

In [20]:

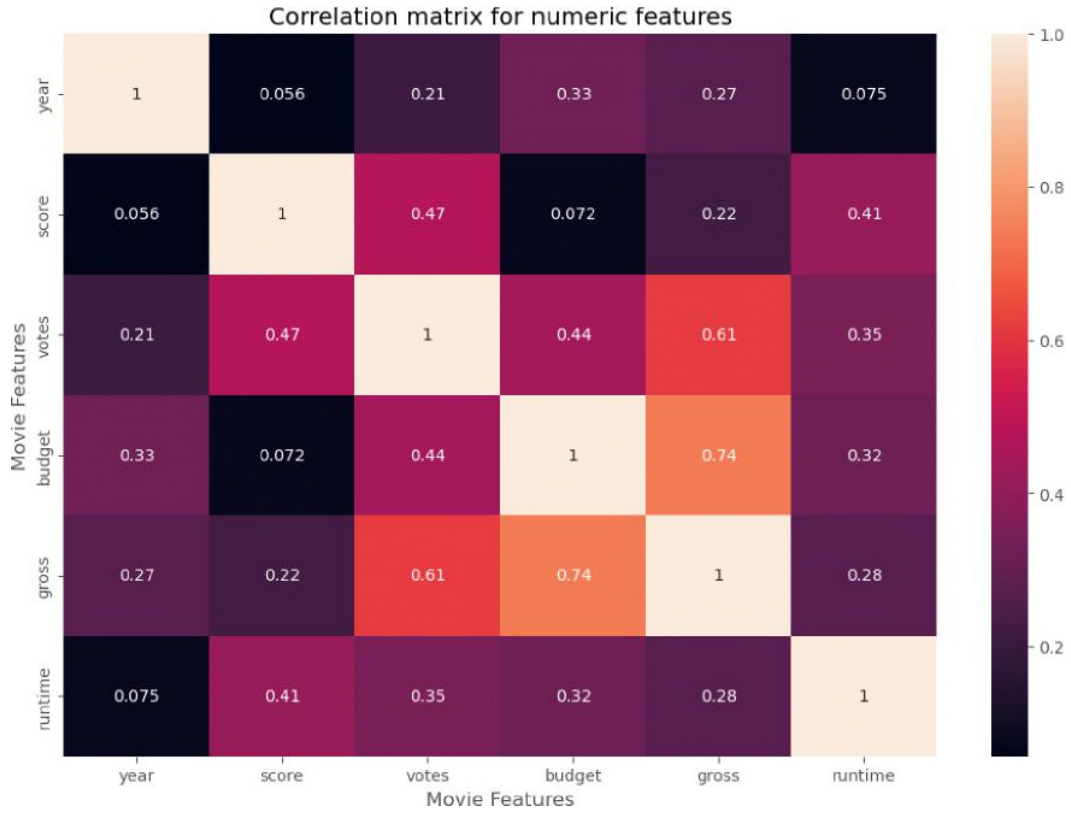
```
df.corr(method="spearman", numeric_only=True)
```

Out[20]:

	year	score	votes	budget	gross	runtime
year	1.000000	0.057741	0.427623	0.312886	0.351045	0.095444
score	0.057741	1.000000	0.495409	-0.009971	0.183192	0.412155
votes	0.427623	0.495409	1.000000	0.493461	0.745793	0.300621
budget	0.312886	-0.009971	0.493461	1.000000	0.692958	0.330794
gross	0.351045	0.183192	0.745793	0.692958	1.000000	0.257400
runtime	0.095444	0.412155	0.300621	0.330794	0.257400	1.000000

In [21]:

```
correlation_matrix = df.corr(method="pearson", numeric_only=True)
sns.heatmap(correlation_matrix,annot=True)
plt.title("Correlation matrix for numeric features")
plt.xlabel("Movie Features")
plt.ylabel("Movie Features")
plt.show()
```



In [22]:

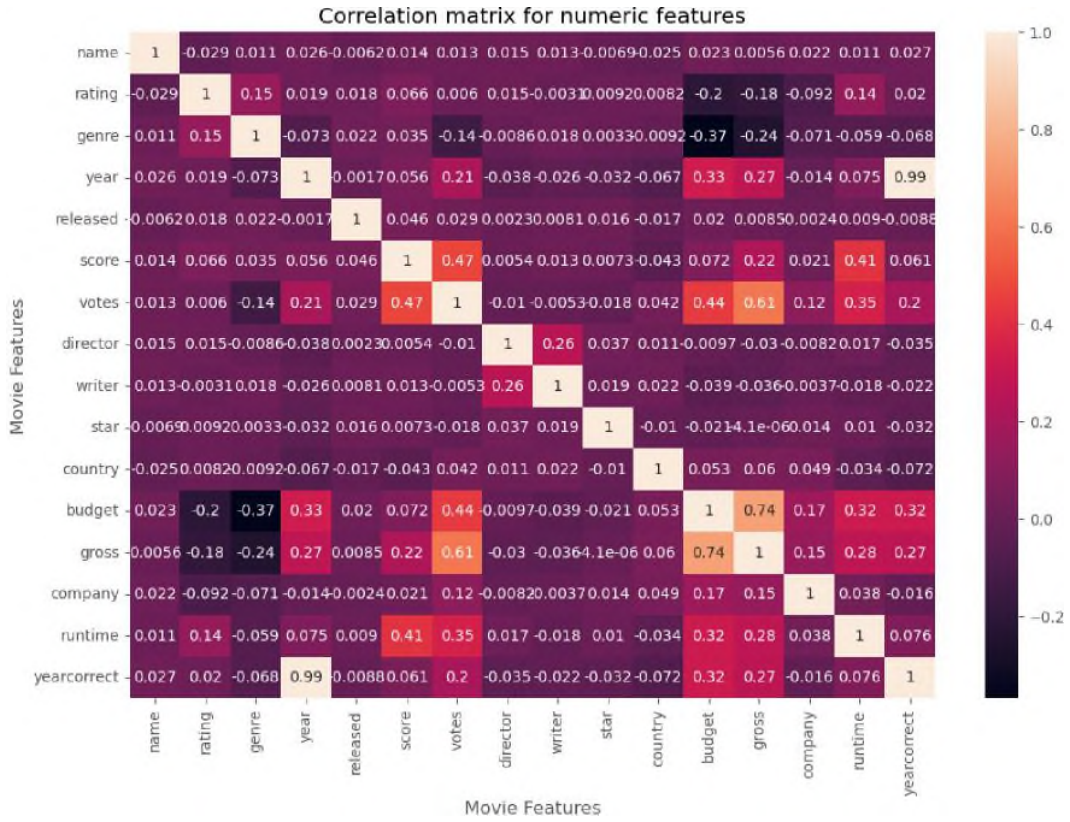
```
#Assigning random numeric value for each unique categorical value

d-F_numerized = d-F

for col_name in d-F_numerized.columns:
    if(d-F_numerized[col_name].dtype=="object"):
        d-F_numerized[col_name] = d-F_numerized[col_name].astype("category")
        d-F_numerized[col_name] = d-F_numerized[col_name].cat.codes

d-F_numerized
```

Out[22]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget
5445	386	5	0	2009	527	7.8	1100000.0	785	1263	1534	47	237000000.0
7445	388	5	0	2019	137	8.4	903000.0	105	513	1470	47	356000000.0
3045	4909	5	6	1997	534	7.8	1100000.0	785	1263	1073	47	200000000.0
6663	3643	5	0	2015	529	7.8	876000.0	768	1806	356	47	245000000.0
7244	389	5	0	2018	145	8.4	897000.0	105	513	1470	47	321000000.0
7480	4388	4	2	2019	1126	6.9	222000.0	1012	1361	457	47	260000000.0
6653	2117	5	0	2015	1303	7.0	593000.0	335	2523	293	47	150000000.0
6043	3878	5	0	2012	1899	8.0	1300000.0	1060	1646	1470	47	220000000.0
6646	1541	5	0	2015	165	7.1	370000.0	809	481	1785	47	190000000.0
7494	1530	4	2	2019	2053	6.8	148000.0	277	1383	1036	47	150000000.0

In [23]:

```
correlation_matrix = df_numerized.corr(method="pearson", numeric_only=True)
sns.heatmap(correlation_matrix,annot=True)
plt.title("Correlation matrix for numeric features")
plt.xlabel("Movie Features")
plt.ylabel("Movie Features")
plt.show()
```



In [24]:

```
correlation_matrix = df_numerized.corr()
correlation_pairs = correlation_matrix.unstack()
correlation_pairs
```

Out[24]:

```
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    0.026784
rating       name          -0.029234
             rating         1.000000
```

In [25]:

```
sorted_pairs = correlation_pairs.sort_values()
sorted_pairs
```

Out[25]:

```
genre        budget        -0.368523
budget       genre         -0.368523
gross        genre         -0.244101
genre        gross         -0.244101
rating       budget        -0.203946
budget       rating        -0.203946
rating       gross         -0.181906
gross        rating        -0.181906
votes        genre         -0.135990
genre        votes         -0.135990
company      rating        -0.092357
rating       company       -0.092357
year         genre         -0.073167
genre        year          -0.073167
country yearcorrect -0.071611
yearcorrect country        -0.071611
company      genre         -0.071334
genre        company       -0.071334
```

In [26]:

```
high_correlation = sorted_pairs[(sorted_pairs)>0.5]
high_correlation
```

Out[26]:

```
votes        gross          0.614751
gross        votes          0.614751
             budget         0.740247
budget       gross          0.740247
yearcorrect year            0.994821
year         yearcorrect    0.994821
name         name           1.000000
company      company        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
rating       rating         1.000000
runtime      runtime        1.000000
yearcorrect yearcorrect 1.000000
dtype: float64
```