In [1]:
```
# 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 configuration of the plots we will create

# Read in the data

df = pd.read_csv(r"C:\Users\Hussain\Desktop\Project 3 - Correlation in Python\movies.csv")
```

In [2]:
```
# Looking at the data

df.head()
```

Out[2]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	r
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	46998772.0	Warner Bros.	
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	58853106.0	Columbia Pictures	
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	538375067.0	Lucasfilm	
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3500000.0	83453539.0	Paramount Pictures	
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	39846344.0	Orion Pictures	

Data Cleaning

In [3]:
```
# 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 - 0.00026082420448617 63%
score - 0.00039123630672926 45%
votes - 0.00039123630672926 45%
director - 0.0%
writer - 0.00039123630672926 45%
star - 0.00013041210224308815%
country - 0.00039123630672926 45%
budget - 0.2831246739697444%
gross - 0.02464788732394366%
company - 0.002217005738132499%
runtime - 0.00052164840897235 26%
```

In [4]:
```
df.dropna(inplace = True)
df
```

Out[4]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	co
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	46998772.0	Warn
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	58853106.0	C
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	538375067.0	Lu
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3500000.0	83453539.0	Pa
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	39846344.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
7648	Bad Boys for Life	R	Action	2020	January 17, 2020 (United States)	6.6	140000.0	Adil El Arbi	Peter Craig	Will Smith	United States	90000000.0	426505244.0	C
7649	Sonic the Hedgehog	PG	Action	2020	February 14, 2020 (United States)	6.5	102000.0	Jeff Fowler	Pat Casey	Ben Schwartz	United States	85000000.0	319715683.0	Pa
7650	Dolittle	PG	Adventure	2020	January 17, 2020 (United States)	5.6	53000.0	Stephen Gaghan	Stephen Gaghan	Robert Downey Jr.	United States	175000000.0	245487753.0	U

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	co
7651	The Call of the Wild	PG	Adventure	2020	February 21, 2020 (United States)	6.8	42000.0	Chris Sanders	Michael Green	Harrison Ford	Canada	135000000.0	111105497.0	20th
7652	The Eight Hundred	Not Rated	Action	2020	August 28, 2020 (United States)	6.8	3700.0	Hu Guan	Hu Guan	Zhi-zhong Huang	China	80000000.0	461421559.0	Bei Entert

5421 rows × 15 columns

In [5]:
```
# Data Types

df.dtypes
```

Out[5]:
```
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 [6]:
```
# Changing data type of columns

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

df
```

Out[6]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	comp
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	19000000	46998772	Warner B
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	58853106	Colur Pict
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	538375067	Lucas
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3500000	83453539	Param Pict
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	6000000	39846344	C Pict
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
7648	Bad Boys for Life	R	Action	2020	January 17, 2020 (United States)	6.6	140000.0	Adil El Arbi	Peter Craig	Will Smith	United States	90000000	426505244	Colur Pict
7649	Sonic the Hedgehog	PG	Action	2020	February 14, 2020 (United States)	6.5	102000.0	Jeff Fowler	Pat Casey	Ben Schwartz	United States	85000000	319715683	Param Pict

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	comp
7650	Dolittle	PG	Adventure	2020	January 17, 2020 (United States)	5.6	53000.0	Stephen Gaghan	Stephen Gaghan	Robert Downey Jr.	United States	175000000	245487753	Unive Pict
7651	The Call of the Wild	PG	Adventure	2020	February 21, 2020 (United States)	6.8	42000.0	Chris Sanders	Michael Green	Harrison Ford	Canada	135000000	111105497	20th Cer Stu
7652	The Eight Hundred	Not Rated	Action	2020	August 28, 2020 (United States)	6.8	3700.0	Hu Guan	Hu Guan	Zhi-zhong Huang	China	80000000	461421559	Beijing Yinx Entertainr

5421 rows × 15 columns

In [7]:
```
# Create correct Year column

df['yearcorrect'] = df['released'].str.extract(pat = '([0-9]{4})').astype(int)
df
```

Out[7]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	19000000	46998772	Warner Bros
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	58853106	Columbia Picture
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	538375067	Lucasfilm
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3500000	83453539	Paramour Picture
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	6000000	39846344	Orio Picture
...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
7648	Bad Boys for Life	R	Action	2020	January 17, 2020 (United States)	6.6	140000.0	Adil El Arbi	Peter Craig	Will Smith	United States	90000000	426505244	Columbia Picture
7649	Sonic the Hedgehog	PG	Action	2020	February 14, 2020 (United States)	6.5	102000.0	Jeff Fowler	Pat Casey	Ben Schwartz	United States	85000000	319715683	Paramour Picture
7650	Dolittle	PG	Adventure	2020	January 17, 2020 (United States)	5.6	53000.0	Stephen Gaghan	Stephen Gaghan	Robert Downey Jr.	United States	175000000	245487753	Universa Picture

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	compan
7651	The Call of the Wild	PG	Adventure	2020	February 21, 2020 (United States)	6.8	42000.0	Chris Sanders	Michael Green	Harrison Ford	Canada	135000000	111105497	20th Centur Studio
7652	The Eight Hundred	Not Rated	Action	2020	August 28, 2020 (United States)	6.8	3700.0	Hu Guan	Hu Guan	Zhi-zhong Huang	China	80000000	461421559	Beijing Dic Yinxian Entertainmer

5421 rows × 16 columns

In [8]:
```
df.sort_values(by = ['gross'], inplace = False, ascending = False)
```

Out[8]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	com
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	James Cameron	Sam Worthington	United States	237000000	2847246203	Twe Centur
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	356000000	2797501328	M St
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	James Cameron	Leonardo DiCaprio	United States	200000000	2201647264	Twe Centur
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	245000000	2069521700	Luca
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	321000000	2048359754	M St
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5640	Tanner Hall	R	Drama	2009	January 15, 2015 (Sweden)	5.8	3500.0	Francesca Gregorini	Tatiana von Fürstenberg	Rooney Mara	United States	3000000	5073	Two F Le
2434	Philadelphia Experiment II	PG-13	Action	1993	June 4, 1994 (South Korea)	4.5	1900.0	Stephen Cornwell	Wallace C. Bennett	Brad Johnson	United States	5000000	2970	Tri Pic
3681	Ginger Snaps	Not Rated	Drama	2000	May 11, 2001 (Canada)	6.8	43000.0	John Fawcett	Karen Walton	Emily Perkins	Canada	5000000	2554	Copper Entertain
272	Parasite	R	Horror	1982	March 12, 1982 (United States)	3.9	2300.0	Charles Band	Alan J. Adler	Robert Glaudini	United States	800000	2270	Emb Pic

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	com
3203	Trojan War	PG-13	Comedy	1997	October 1, 1997 (Brazil)	5.7	5800.0	George Huang	Andy Burg	Will Friedle	United States	15000000	309	Dayb

5421 rows × 16 columns

In [9]:
```
pd.set_option('display.max_rows', None)
```

In [10]:
```
df = df.sort_values(by = ['gross'], inplace = False, ascending = False)
```

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

df.drop_duplicates()
```

Out[11]:

	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:	PG-13	Action	2018	April 27, 2018 (United	8.4	897000.0	Anthony Russo	Christopher	Robert Downey	United	321

In [12]:
```
pd.set_option('display.max_rows', 500)
```

In [13]:
```
df
```

Out[13]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	com
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	James Cameron	Sam Worthington	United States	237000000	2847246203	Twe Centur
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	356000000	2797501328	M St
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	James Cameron	Leonardo DiCaprio	United States	200000000	2201647264	Twe Centur
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	245000000	2069521700	Luca
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	321000000	2048359754	M St
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5640	Tanner Hall	R	Drama	2009	January 15, 2015 (Sweden)	5.8	3500.0	Francesca Gregorini	Tatiana von Fürstenberg	Rooney Mara	United States	3000000	5073	Two F Le
2434	Philadelphia Experiment II	PG-13	Action	1993	June 4, 1994 (South Korea)	4.5	1900.0	Stephen Cornwell	Wallace C. Bennett	Brad Johnson	United States	5000000	2970	Tri Pic
3681	Ginger Snaps	Not Rated	Drama	2000	May 11, 2001 (Canada)	6.8	43000.0	John Fawcett	Karen Walton	Emily Perkins	Canada	5000000	2554	Copper Entertain
272	Parasite	R	Horror	1982	March 12, 1982 (United States)	3.9	2300.0	Charles Band	Alan J. Adler	Robert Glaudini	United States	800000	2270	Emb Pic

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	com
3203	Trojan War	PG-13	Comedy	1997	October 1, 1997 (Brazil)	5.7	5800.0	George Huang	Andy Burg	Will Friedle	United States	15000000	309	Dayb

5421 rows × 16 columns

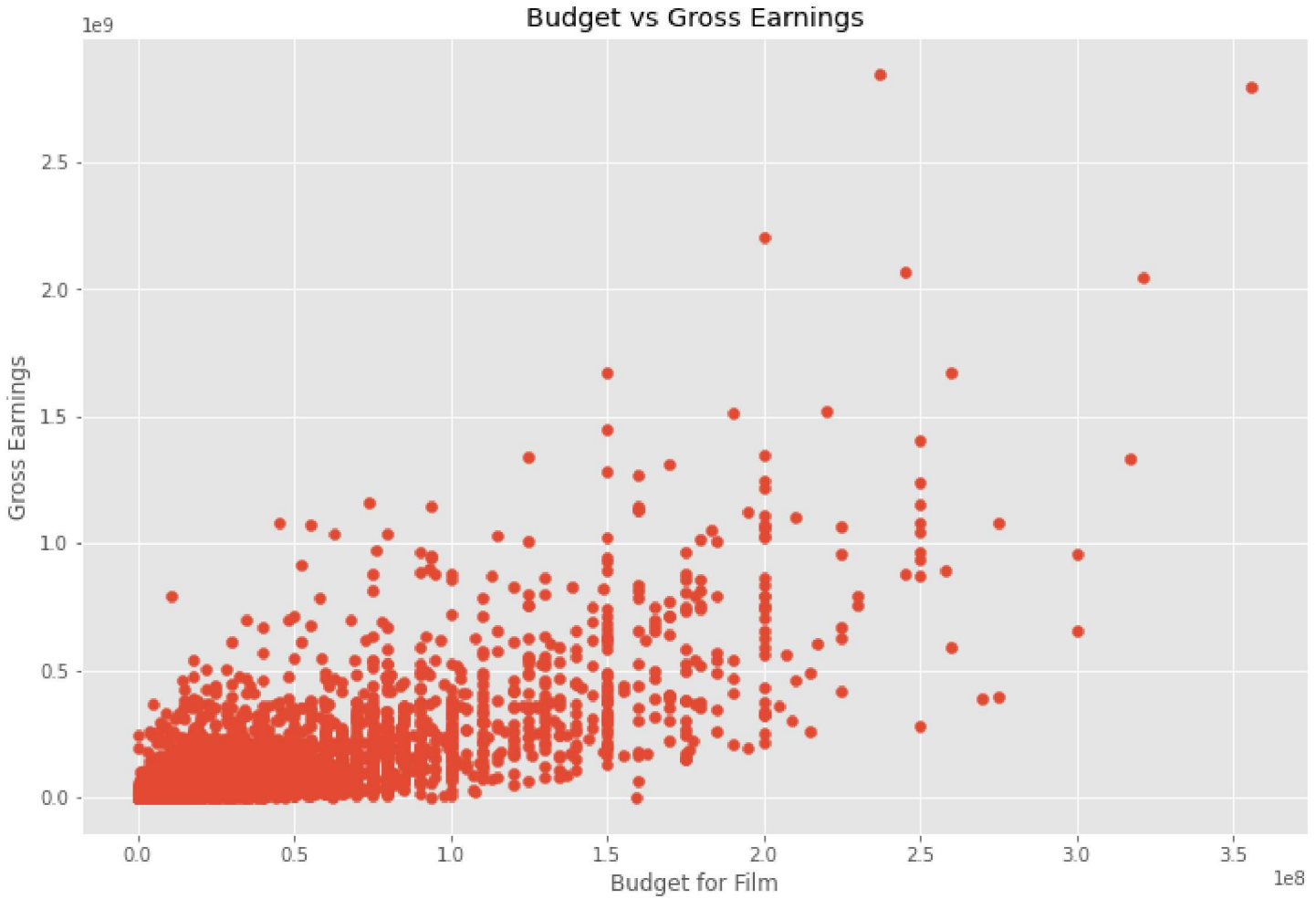
Correlation

In [15]:
```
# Scatter plot with budget vs gross

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

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

plt.show()
```



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

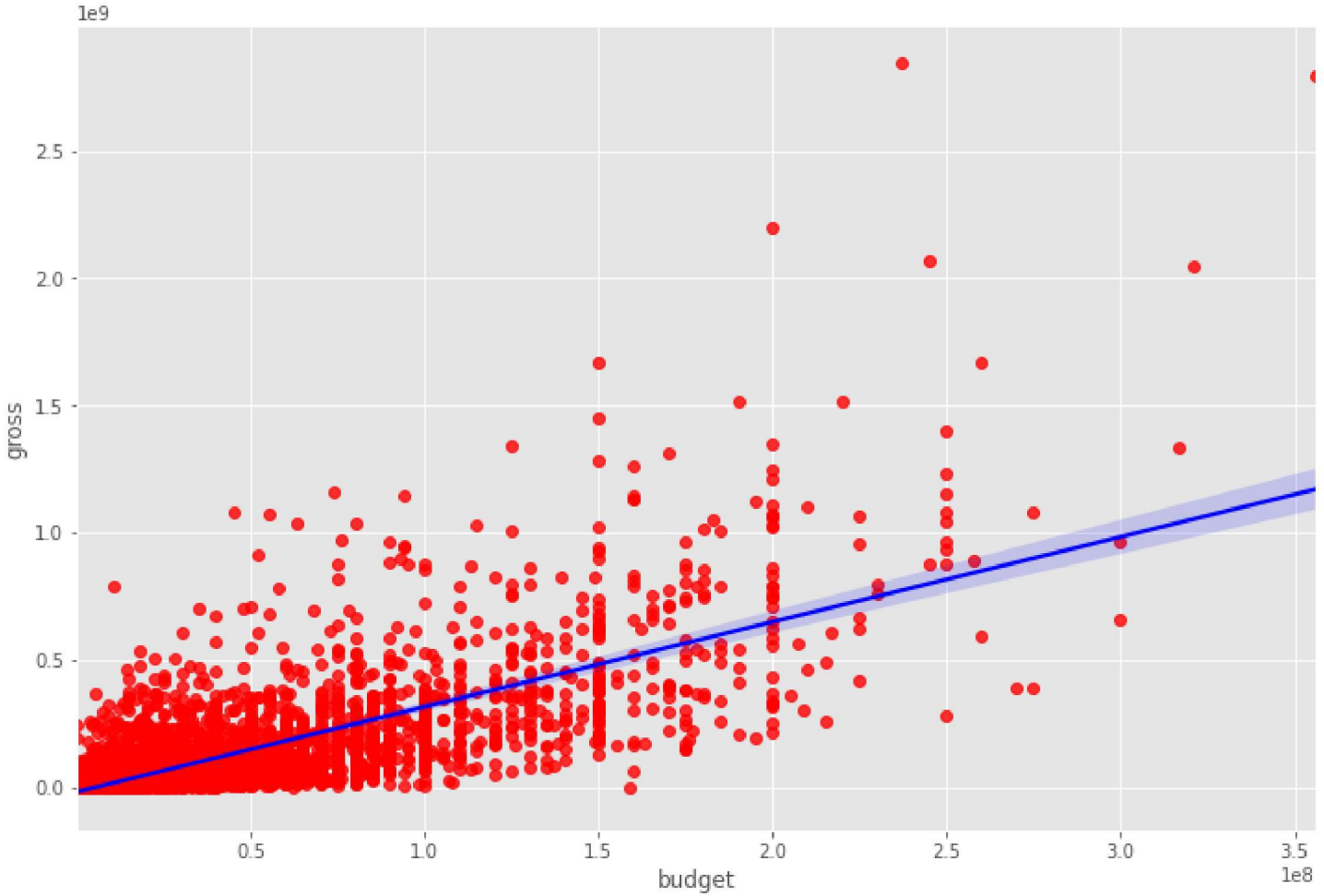
Out[16]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	ru
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	James Cameron	Sam Worthington	United States	237000000	2847246203	Twentieth Century Fox	
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	356000000	2797501328	Marvel Studios	
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	James Cameron	Leonardo DiCaprio	United States	200000000	2201647264	Twentieth Century Fox	
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	245000000	2069521700	Lucasfilm	
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	321000000	2048359754	Marvel Studios	

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

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

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



In [18]:
```
# Looking at Correlation

df.corr(method = 'pearson') #pearson, kendall, spearman
```

Out[18]:

	year	score	votes	budget	gross	runtime	yearcorrect
year	1.000000	0.056386	0.206021	0.327722	0.274321	0.075077	0.998726
score	0.056386	1.000000	0.474256	0.072001	0.222556	0.414068	0.061923
votes	0.206021	0.474256	1.000000	0.439675	0.614751	0.352303	0.203098
budget	0.327722	0.072001	0.439675	1.000000	0.740247	0.318695	0.320312
gross	0.274321	0.222556	0.614751	0.740247	1.000000	0.275796	0.268721
runtime	0.075077	0.414068	0.352303	0.318695	0.275796	1.000000	0.075294
yearcorrect	0.998726	0.061923	0.203098	0.320312	0.268721	0.075294	1.000000

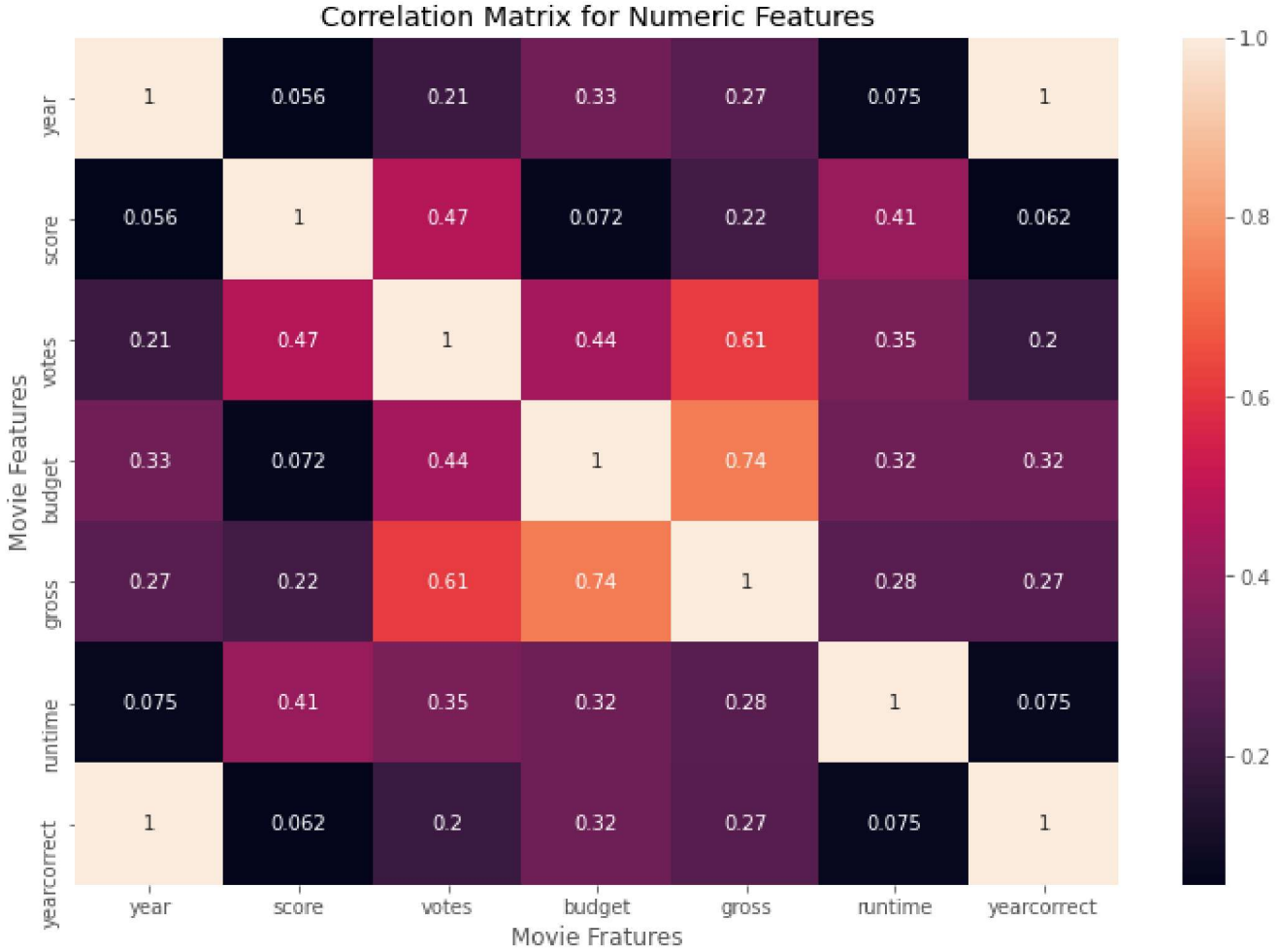
In [19]:
```
# High Correlation between budget and gross
```

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

sns.heatmap(correlation_matrix, annot = True)

plt.title('Correlation Matrix for Numeric Features')
plt.xlabel('Movie Fratures')
plt.ylabel('Movie Features')

plt.show()
```



In [21]:
```
# Looking at Company

df.head()
```

Out[21]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	ru
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	James Cameron	Sam Worthington	United States	237000000	2847246203	Twentieth Century Fox	
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	356000000	2797501328	Marvel Studios	
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	James Cameron	Leonardo DiCaprio	United States	200000000	2201647264	Twentieth Century Fox	
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	245000000	2069521700	Lucasfilm	
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	321000000	2048359754	Marvel Studios	

In [24]:
```
df_numerised = df

for col_name in df_numerised.columns:
    if(df_numerised[col_name].dtype == 'object'):
        df_numerised[col_name] = df_numerised[col_name].astype('category')
        df_numerised[col_name] = df_numerised[col_name].cat.codes

df_numerised
```

Out[24]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime	yearcorrec
5445	386	5	0	2009	527	7.8	1100000.0	785	1263	1534	47	237000000	2847246203	1382	162.0	200
7445	388	5	0	2019	137	8.4	903000.0	105	513	1470	47	356000000	2797501328	983	181.0	201
3045	4909	5	6	1997	534	7.8	1100000.0	785	1263	1073	47	200000000	2201647264	1382	194.0	199
6663	3643	5	0	2015	529	7.8	876000.0	768	1806	356	47	245000000	2069521700	945	138.0	201
7244	389	5	0	2018	145	8.4	897000.0	105	513	1470	47	321000000	2048359754	983	149.0	201
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
5640	3794	6	6	2009	890	5.8	3500.0	585	2924	1498	47	3000000	5073	1385	96.0	201
2434	2969	5	0	1993	1467	4.5	1900.0	1805	3102	186	47	5000000	2970	1376	97.0	199
3681	1595	3	6	2000	1721	6.8	43000.0	952	1683	527	6	5000000	2554	466	108.0	200
272	2909	6	9	1982	1525	3.9	2300.0	261	55	1473	47	800000	2270	582	85.0	198
3203	4966	5	4	1997	2152	5.7	5800.0	651	161	1811	47	15000000	309	504	85.0	199

5421 rows × 16 columns

In [23]:
```
df
```

Out[23]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	com
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	James Cameron	Sam Worthington	United States	237000000	2847246203	Twe Centur
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	356000000	2797501328	M St
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	James Cameron	Leonardo DiCaprio	United States	200000000	2201647264	Twe Centur
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	245000000	2069521700	Luca
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	321000000	2048359754	M St
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5640	Tanner Hall	R	Drama	2009	January 15, 2015 (Sweden)	5.8	3500.0	Francesca Gregorini	Tatiana von Fürstenberg	Rooney Mara	United States	3000000	5073	Two F Le
2434	Philadelphia Experiment II	PG-13	Action	1993	June 4, 1994 (South Korea)	4.5	1900.0	Stephen Cornwell	Wallace C. Bennett	Brad Johnson	United States	5000000	2970	Tri Pic
3681	Ginger Snaps	Not Rated	Drama	2000	May 11, 2001 (Canada)	6.8	43000.0	John Fawcett	Karen Walton	Emily Perkins	Canada	5000000	2554	Copper Entertain
272	Parasite	R	Horror	1982	March 12, 1982 (United States)	3.9	2300.0	Charles Band	Alan J. Adler	Robert Glaudini	United States	800000	2270	Emb Pic

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	com
3203	Trojan War	PG-13	Comedy	1997	October 1, 1997 (Brazil)	5.7	5800.0	George Huang	Andy Burg	Will Friedle	United States	15000000	309	Dayb

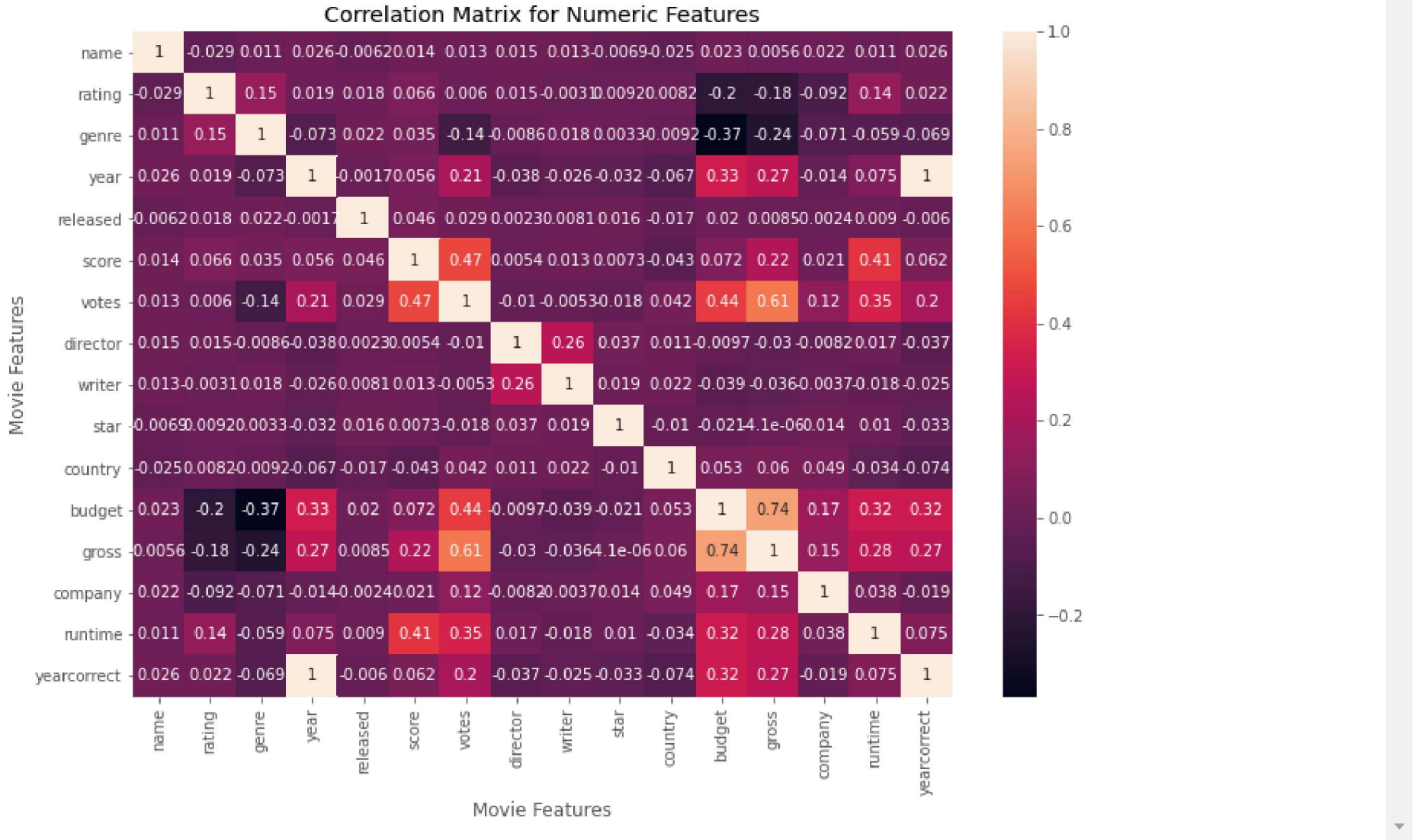
5421 rows × 16 columns

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

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
Movie Features
name 1 -0.029 0.011 0.026 -0.0062 0.014 0.013 0.015 0.013 -0.0069 -0.025 0.023 0.0056 0.022 0.011 0.026
rating -0.029 1 0.15 0.019 0.018 0.066 0.006 0.015 -0.0031 0.0092 0.0082 -0.2 -0.18 -0.092 0.14 0.022
genre 0.011 0.15 1 -0.073 0.022 0.035 -0.14 -0.0086 0.018 0.0033 -0.0092 -0.37 -0.24 -0.071 -0.059 -0.069
year 0.026 0.019 -0.073 1 -0.0017 0.056 0.21 -0.038 -0.026 -0.032 -0.067 0.33 0.27 -0.014 0.075 1
released -0.0062 0.018 0.022 -0.0017 1 0.046 0.029 0.0023 0.0081 0.016 -0.017 0.02 0.0085 0.0024 0.009 -0.006
score 0.014 0.066 0.035 0.056 0.046 1 0.47 0.0054 0.013 0.0073 -0.043 0.072 0.22 0.021 0.41 0.062
votes 0.013 0.006 -0.14 0.21 0.029 0.47 1 -0.01 -0.0053 -0.018 0.042 0.44 0.61 0.12 0.35 0.2
director 0.015 0.015 -0.0086 -0.038 0.0023 0.0054 -0.01 1 0.26 0.037 0.011 -0.0097 -0.03 -0.0082 0.017 -0.037
writer 0.013 -0.0031 0.018 -0.026 0.0081 0.013 -0.0053 0.26 1 0.019 0.022 -0.039 -0.036 -0.0037 -0.018 -0.025
star -0.0069 0.0092 0.0033 -0.032 0.016 0.0073 -0.018 0.037 0.019 1 -0.01 -0.021 4.1e-06 0.014 0.01 -0.033
country -0.025 0.0082 -0.0092 -0.067 -0.017 -0.043 0.042 0.011 0.022 -0.01 1 0.053 0.06 0.049 -0.034 -0.074
budget 0.023 -0.2 -0.37 0.33 0.02 0.072 0.44 -0.0097 -0.039 -0.021 0.053 1 0.74 0.17 0.32 0.32
gross 0.0056 -0.18 -0.24 0.27 0.0085 0.22 0.61 -0.03 -0.0364 4.1e-06 0.06 0.74 1 0.15 0.28 0.27
company 0.022 -0.092 -0.071 -0.014 -0.0024 0.021 0.12 -0.0082 0.0037 0.014 0.049 0.17 0.15 1 0.038 -0.019
runtime 0.011 0.14 -0.059 0.075 0.009 0.41 0.35 0.017 -0.018 0.01 -0.034 0.32 0.28 0.038 1 0.075
yearcorrect 0.026 0.022 -0.069 1 -0.006 0.062 0.2 -0.037 -0.025 -0.033 -0.074 0.32 0.27 -0.019 0.075 1
name rating genre year released score votes director writer star country budget gross company runtime yearcorrect
Movie Features
1.0 0.8 0.6 0.4 0.2 0.0 −0.2

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

Out[27]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gro
name	1.000000	-0.029234	0.010996	0.025542	-0.006152	0.014450	0.012615	0.015246	0.012880	-0.006882	-0.025490	0.023392	0.005
rating	-0.029234	1.000000	0.147796	0.019499	0.018083	0.065983	0.006031	0.014656	-0.003149	0.009196	0.008230	-0.203946	-0.181
genre	0.010996	0.147796	1.000000	-0.073167	0.022142	0.035106	-0.135990	-0.008553	0.017578	0.003341	-0.009164	-0.368523	-0.244
year	0.025542	0.019499	-0.073167	1.000000	-0.001740	0.056386	0.206021	-0.038354	-0.025908	-0.032157	-0.066748	0.327722	0.274
released	-0.006152	0.018083	0.022142	-0.001740	1.000000	0.045874	0.028833	0.002308	0.008072	0.015706	-0.017228	0.019952	0.008
score	0.014450	0.065983	0.035106	0.056386	0.045874	1.000000	0.474256	0.005413	0.012843	0.007296	-0.043051	0.072001	0.222
votes	0.012615	0.006031	-0.135990	0.206021	0.028833	0.474256	1.000000	-0.010376	-0.005316	-0.017638	0.041551	0.439675	0.614
director	0.015246	0.014656	-0.008553	-0.038354	0.002308	0.005413	-0.010376	1.000000	0.261735	0.036593	0.011133	-0.009662	-0.029
writer	0.012880	-0.003149	0.017578	-0.025908	0.008072	0.012843	-0.005316	0.261735	1.000000	0.018520	0.022488	-0.039466	-0.035
star	-0.006882	0.009196	0.003341	-0.032157	0.015706	0.007296	-0.017638	0.036593	0.018520	1.000000	-0.009990	-0.021473	-0.000
country	-0.025490	0.008230	-0.009164	-0.066748	-0.017228	-0.043051	0.041551	0.011133	0.022488	-0.009990	1.000000	0.052977	0.060
budget	0.023392	-0.203946	-0.368523	0.327722	0.019952	0.072001	0.439675	-0.009662	-0.039466	-0.021473	0.052977	1.000000	0.740
gross	0.005639	-0.181906	-0.244101	0.274321	0.008501	0.222556	0.614751	-0.029560	-0.035885	-0.000004	0.060078	0.740247	1.000
company	0.021697	-0.092357	-0.071334	-0.014333	-0.002407	0.020656	0.118470	-0.008223	-0.003697	0.014082	0.048569	0.170235	0.149
runtime	0.010850	0.140792	-0.059237	0.075077	0.008975	0.414068	0.352303	0.017433	-0.017561	0.010108	-0.034477	0.318695	0.275
yearcorrect	0.025542	0.022021	-0.069147	0.998726	-0.005989	0.061923	0.203098	-0.037371	-0.025495	-0.032687	-0.073569	0.320312	0.268

In [28]:
```
correlation_mat = df_numerised.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
```

Out[29]:
```
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
country      yearcorrect   -0.073569
yearcorrect  country       -0.073569
year         genre         -0.073167
genre        year          -0.073167
             company       -0.071334
company      genre         -0.071334
genre        yearcorrect   -0.069147
```

In [31]:
```
high_corr = sorted_pairs[(sorted_pairs) > 0.5]

high_corr
```

Out[31]:
```
gross        votes          0.614751
votes        gross          0.614751
gross        budget         0.740247
budget       gross          0.740247
year         yearcorrect    0.998726
yearcorrect  year           0.998726
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
```

In [ ]:
```

```