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 config of plots

#read in data
df = pd.read_csv(r'C:\Users\Liam\Downloads\archive(1)\movies.csv')
```

In [2]:

```
#look at data
df
```

Out[2]:

	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	Ni
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	
...	...	...	...	...	...	...	...	...	...	
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks	Joseph Ebanks	S
7664	Dream Round	NaN	Comedy	2020	February 7, 2020 (United States)	4.7	36.0	Dusty Dukatz	Lisa Huston	S
7665	Saving Mbango	NaN	Drama	2020	April 27, 2020 (Cameroon)	5.7	29.0	Nkanya Nkwai	Lynno Lovert	C
7666	It's Just Us	NaN	Drama	2020	October 1, 2020 (United States)	NaN	NaN	James Randall	James Randall	C
7667	Tee em el	NaN	Horror	2020	August 19, 2020 (United States)	5.7	7.0	Pereko Mosia	Pereko Mosia	Siy

7668 rows × 15 columns

In [3]:

```
#look for missing data

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

```
name - 0.0%
rating - 0.010041731872717789%
genre - 0.0%
year - 0.0%
released - 0.0002608242044861763%
score - 0.0003912363067292645%
votes - 0.0003912363067292645%
director - 0.0%
writer - 0.0003912363067292645%
star - 0.0001304121022430815%
country - 0.0003912363067292645%
budget - 0.2831246739697444%
gross - 0.02464788732394366%
company - 0.002217005738132499%
runtime - 0.0005216484089723526%
```

In [4]:
```
#drop missing data

df = df.dropna().copy()
#.copy() is needed to avoid pandas error
```

In [5]:
```
#check our work for missing data

for col in df.columns:
    pct_missing = np.mean(df[col].isnull())
    print('{} - {}%'.format(col, pct_missing))
# results show no more duplicates
#7668 rows before removal
#5421 after duplicates removed
#70.7% of the data remaining
```

```
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]:
```
#look for duplicate values

new_output = df[df.duplicated()]
print("duplicated values", new_output)
#no duplicates so we proceed without needing to update our df
```

```
duplicated values Empty DataFrame
Columns: [name, rating, genre, year, released, score, votes, director, writer, sta
r, country, budget, gross, company, runtime]
Index: []
```

In [7]:

```python
# data types of 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 [8]:

```python
#some columns don't match "year of release" and "release date"
#creating a new column that matches

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

Out[8]:

	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	Nicho
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Brc Shi
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	N Ha
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Rc I
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Cl Cl
...	...	...	...	...	...	...	...	...	...	
7648	Bad Boys for Life	R	Action	2020	January 17, 2020 (United States)	6.6	140000.0	Adil El Arbi	Peter Craig	S
7649	Sonic the Hedgehog	PG	Action	2020	February 14, 2020 (United States)	6.5	102000.0	Jeff Fowler	Pat Casey	Schw
7650	Dolittle	PG	Adventure	2020	January 17, 2020 (United States)	5.6	53000.0	Stephen Gaghan	Stephen Gaghan	Rc Dov
7651	The Call of the Wild	PG	Adventure	2020	February 21, 2020 (United States)	6.8	42000.0	Chris Sanders	Michael Green	Harr
7652	The Eight Hundred	Not Rated	Action	2020	August 28, 2020 (United States)	6.8	3700.0	Hu Guan	Hu Guan	zh Hu

5421 rows × 16 columns

In [9]:

```
#setting max rows higher and sorting by gross column

pd.set_option('display.max_rows', 200)
pd.set_option('display.min_rows', 50)

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

Out[9]:

	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
7445	Avengers: Endgame	PG-13	Action	2019	April 26, 2019 (United States)	8.4	903000.0	Anthony Russo	Christopher Markus
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	James Cameron
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
7480	The Lion King	PG	Animation	2019	July 19, 2019 (United States)	6.9	222000.0	Jon Favreau	Jeff Nathanson
6653	Jurassic World	PG-13	Action	2015	June 12, 2015 (United States)	7.0	593000.0	Colin Trevorrow	Rick Jaffa
6043	The Avengers	PG-13	Action	2012	May 4, 2012 (United States)	8.0	1300000.0	Joss Whedon	Joss Whedon
6646	Furious 7	PG-13	Action	2015	April 3, 2015 (United States)	7.1	370000.0	James Wan	Chris Morgan
7494	Frozen II	PG	Animation	2019	November 22, 2019 (United States)	6.8	148000.0	Chris Buck	Jennifer Lee
6644	Avengers: Age of Ultron	PG-13	Action	2015	May 1, 2015 (United States)	7.3	777000.0	Joss Whedon	Joss Whedon
7247	Black Panther	PG-13	Action	2018	February 16, 2018 (United States)	7.3	661000.0	Ryan Coogler	Ryan Coogler

	name	rating	genre	year	released	score	votes	director	writer
5845	Harry Potter and the Deathly Hallows: Part 2	PG-13	Adventure	2011	July 15, 2011 (United States)	8.1	790000.0	David Yates	Steve Kloves
7075	Star Wars: Episode VIII - The Last Jedi	PG-13	Action	2017	December 15, 2017 (United States)	7.0	581000.0	Rian Johnson	Rian Johnson
7271	Jurassic World: Fallen Kingdom	PG-13	Action	2018	June 22, 2018 (United States)	6.2	277000.0	J.A. Bayona	Derek Connolly
6262	Frozen	PG	Animation	2013	November 27, 2013 (United States)	7.4	585000.0	Chris Buck	Jennifer Lee
7072	Beauty and the Beast	PG	Family	2017	March 17, 2017 (United States)	7.1	283000.0	Bill Condon	Stephen Chbosky
7281	Incredibles 2	PG	Animation	2018	June 15, 2018 (United States)	7.6	263000.0	Brad Bird	Brad Bird
7055	The Fate of the Furious	PG-13	Action	2017	April 14, 2017 (United States)	6.6	214000.0	F. Gary Gray	Gary Scott Thompson
6244	Iron Man 3	PG-13	Action	2013	May 3, 2013 (United States)	7.1	779000.0	Shane Black	Drew Pearce
6688	Minions	PG	Animation	2015	July 10, 2015 (United States)	6.4	218000.0	Kyle Balda	Brian Lynch
6846	Captain America: Civil War	PG-13	Action	2016	May 6, 2016 (United States)	7.8	694000.0	Anthony Russo	Christopher Markus
7250	Aquaman	PG-13	Action	2018	December 21, 2018 (United States)	6.9	404000.0	James Wan	David Leslie Johnson-McGoldrick
4245	The Lord of the Rings: The Return of the King	PG-13	Action	2003	December 17, 2003 (United States)	8.9	1700000.0	Peter Jackson	J.R.R. Tolkien

	name	rating	genre	year	released	score	votes	director	writer
7458	Spider-Man: Far from Home	PG-13	Action	2019	July 2, 2019 (United States)	7.5	359000.0	Jon Watts	Chris McKenna
...	...	...	...	...	...	...	...	...	...
5412	Pontypool	Not Rated	Fantasy	2008	September 18, 2009 (Turkey)	6.6	30000.0	Bruce McDonald	Tony Burgess
3465	The Boondock Saints	R	Action	1999	January 21, 2000 (Canada)	7.8	230000.0	Troy Duffy	Troy Duffy
405	Rock & Rule	PG	Animation	1983	July 24, 1987 (United States)	6.5	3400.0	Clive Smith	Patrick Loubert
800	O.C. and Stiggs	R	Comedy	1985	1985 (United States)	5.4	1200.0	Robert Altman	Tod Carroll
1898	The Lovers on the Bridge	R	Drama	1991	July 2, 1999 (United States)	7.6	13000.0	Leos Carax	Leos Carax
2342	Freaked	PG-13	Comedy	1993	March 31, 1994 (Australia)	6.4	6700.0	Tom Stern	Tim Burns
3618	Best Laid Plans	R	Crime	1999	May 14, 1999 (United Kingdom)	6.1	7400.0	Mike Barker	Ted Griffin
467	My Brother's Wedding	Not Rated	Drama	1983	March 1985 (United States)	7.2	826.0	Charles Burnett	Charles Burnett
5840	Passion Play	R	Drama	2010	July 2, 2011 (Taiwan)	4.6	7400.0	Mitch Glazer	Mitch Glazer
3777	The Isle	Not Rated	Drama	2000	April 22, 2000 (South Korea)	7.0	13000.0	Kim Ki-duk	Kim Ki-duk
6512	Honeymoon	R	Drama	2014	September 12, 2014 (United States)	5.7	25000.0	Leigh Janiak	Phil Graziadei

	name	rating	genre	year	released	score	votes	director	writer
2401	Deadfall	R	Crime	1993	October 8, 1993 (United States)	4.0	3000.0	Christopher Coppola	Christopher Coppola
714	Smooth Talk	PG-13	Drama	1985	November 15, 1985 (United States)	6.5	2200.0	Joyce Chopra	Joyce Carol Oates
6616	Barefoot	PG-13	Comedy	2014	September 4, 2014 (Israel)	6.6	24000.0	Andrew Fleming	Stephen Zotnowski
3413	Savior	R	Drama	1998	November 20, 1998 (United States)	7.3	11000.0	Predrag Antonijevic	Robert Orr
3830	The Specials	R	Action	2000	September 18, 2000 (United States)	5.8	2200.0	Craig Mazin	James Gunn
3438	Hell's Kitchen	R	Crime	1998	January 19, 2001 (Italy)	4.7	2500.0	Tony Cinciripini	Tony Cinciripini
3024	Schizopolis	Not Rated	Comedy	1996	April 9, 1997 (United States)	6.8	5300.0	Steven Soderbergh	Steven Soderbergh
6147	About Cherry	R	Drama	2012	August 9, 2012 (United States)	4.8	10000.0	Stephen Elliott	Stephen Elliott
760	Crimewave	PG-13	Comedy	1985	April 25, 1986 (United States)	5.7	5300.0	Sam Raimi	Ethan Coen
5640	Tanner Hall	R	Drama	2009	January 15, 2015 (Sweden)	5.8	3500.0	Francesca Gregorini	Tatiana von Fürstenberg
2434	Philadelphia Experiment II	PG-13	Action	1993	June 4, 1994 (South Korea)	4.5	1900.0	Stephen Cornwell	Wallace C. Bennett
3681	Ginger Snaps	Not Rated	Drama	2000	May 11, 2001 (Canada)	6.8	43000.0	John Fawcett	Karen Walton
272	Parasite	R	Horror	1982	March 12, 1982 (United States)	3.9	2300.0	Charles Band	Alan J. Adler

	name	rating	genre	year	released	score	votes	director	writer
3203	Trojan War	PG-13	Comedy	1997	October 1, 1997 (Brazil)	5.7	5800.0	George Huang	Andy Burg

5421 rows × 16 columns

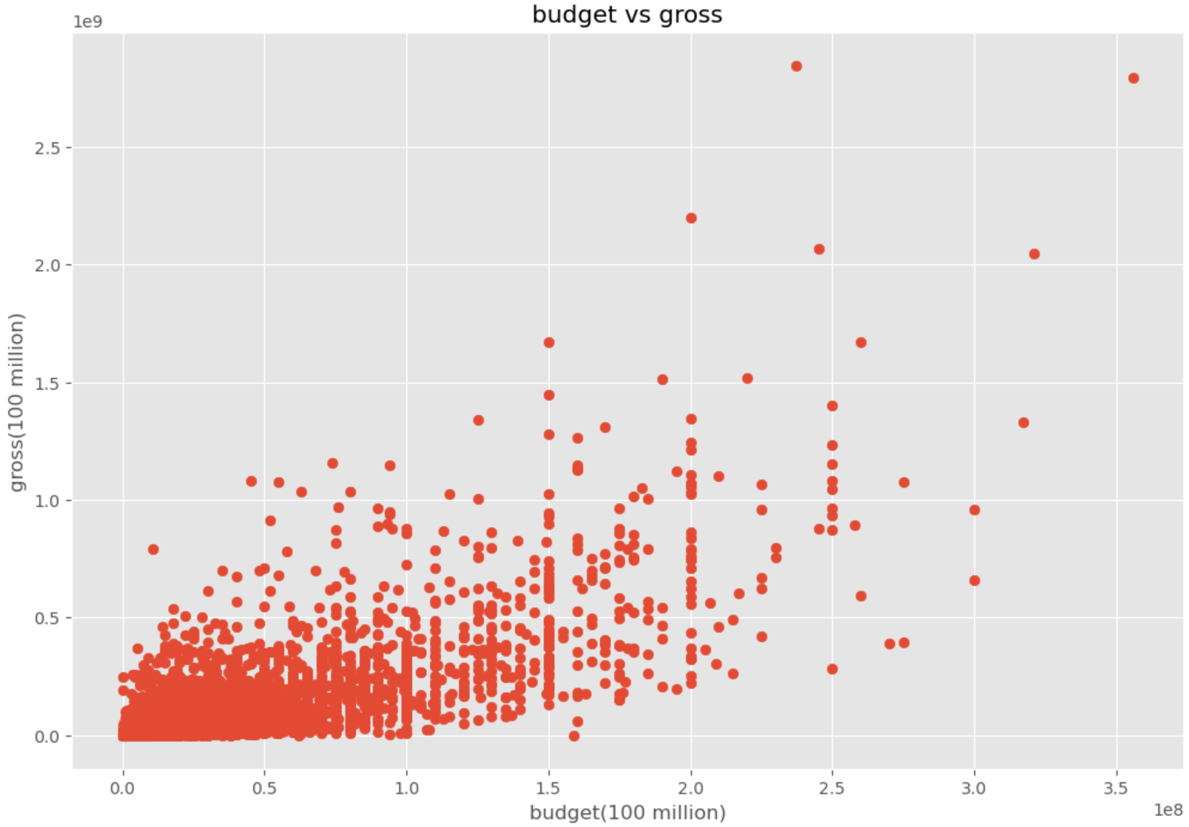
In [10]:

```
#scatter plot

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

plt.title('budget vs gross')
plt.xlabel('budget(100 million)')
plt.ylabel('gross(100 million)')

plt.show()
#results look positively correlated at first glance
```



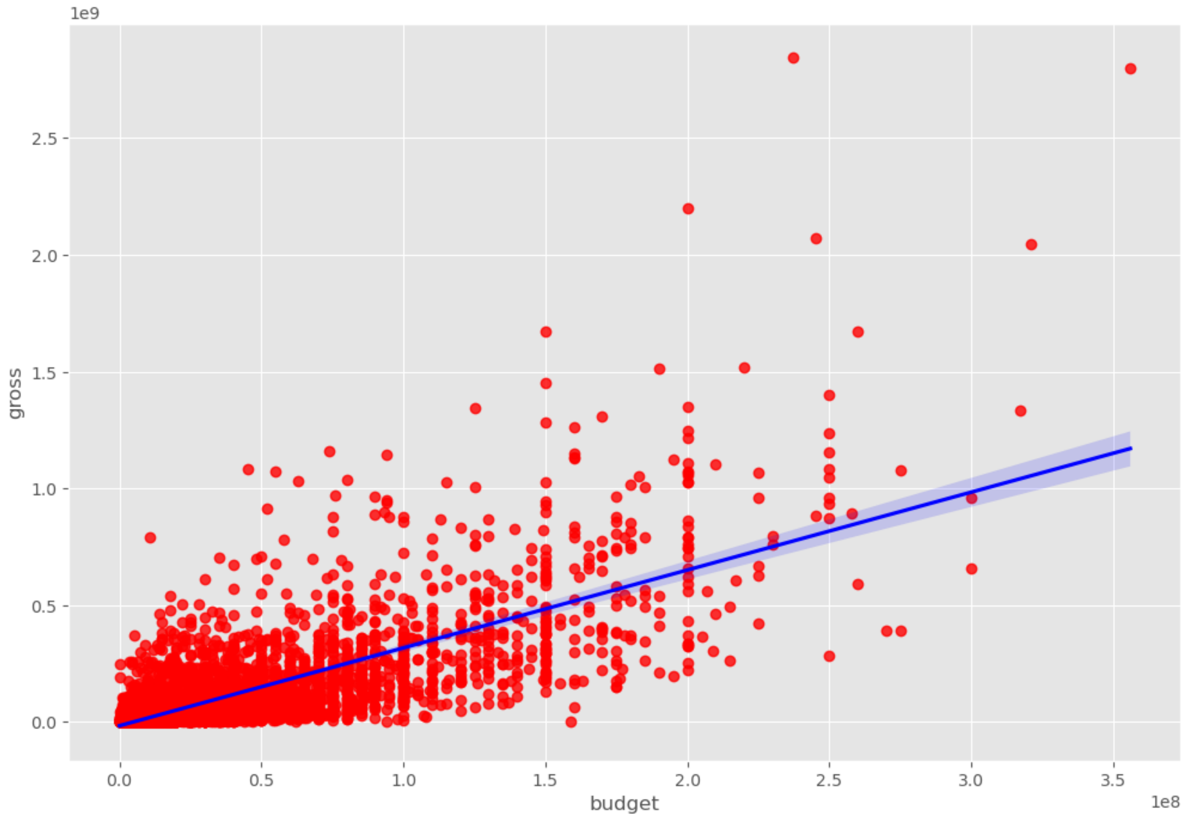
In [11]:

```
#now creating a regression line to see if our initial impression was correct

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

Out[11]:

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



In [12]:

```
df.corr(method='pearson') #pearson, kendall, spearman corr options
#across all 3 methods gross is most correlated with votes and budget
```

Out[12]:

	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

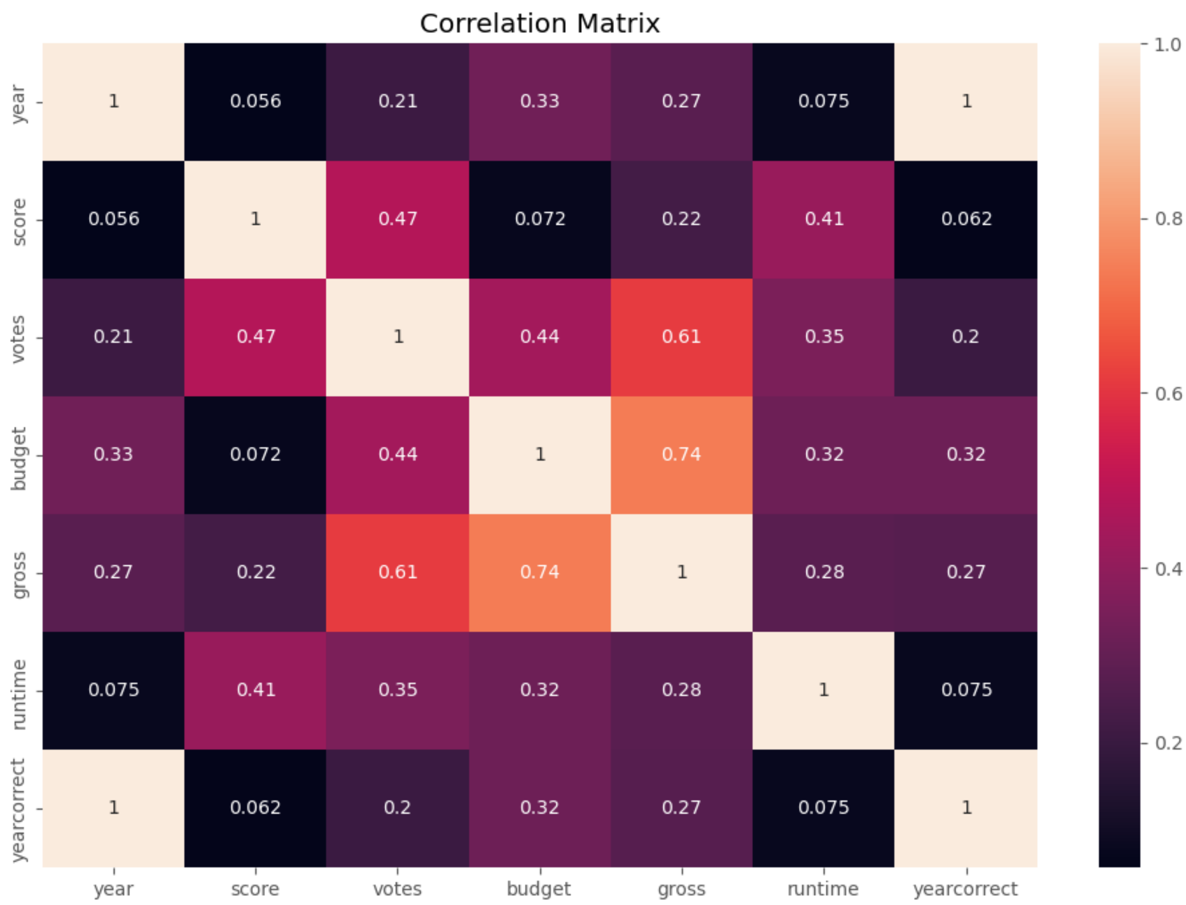
In [13]:

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

sns.heatmap(correlation_matrix, annot=True)

plt.title('Correlation Matrix')

plt.show()
```



	year	score	votes	budget	gross	runtime	yearcorrect
year	1	0.056	0.21	0.33	0.27	0.075	1
score	0.056	1	0.47	0.072	0.22	0.41	0.062
votes	0.21	0.47	1	0.44	0.61	0.35	0.2
budget	0.33	0.072	0.44	1	0.74	0.32	0.32
gross	0.27	0.22	0.61	0.74	1	0.28	0.27
runtime	0.075	0.41	0.35	0.32	0.28	1	0.075
yearcorrect	1	0.062	0.2	0.32	0.27	0.075	1

In [14]:

```
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[14]:

	name	rating	genre	year	released	score	votes	director	writer	star	country	buc
0	4692	6	6	1980	1304	8.4	927000.0	1795	2832	699	46	190000
1	3929	6	1	1980	1127	5.8	65000.0	1578	1158	214	47	45000
2	3641	4	0	1980	1359	8.7	1200000.0	757	1818	1157	47	180000
3	204	4	4	1980	1127	7.7	221000.0	889	1413	1474	47	35000
4	732	6	4	1980	1170	7.3	108000.0	719	351	271	47	60000

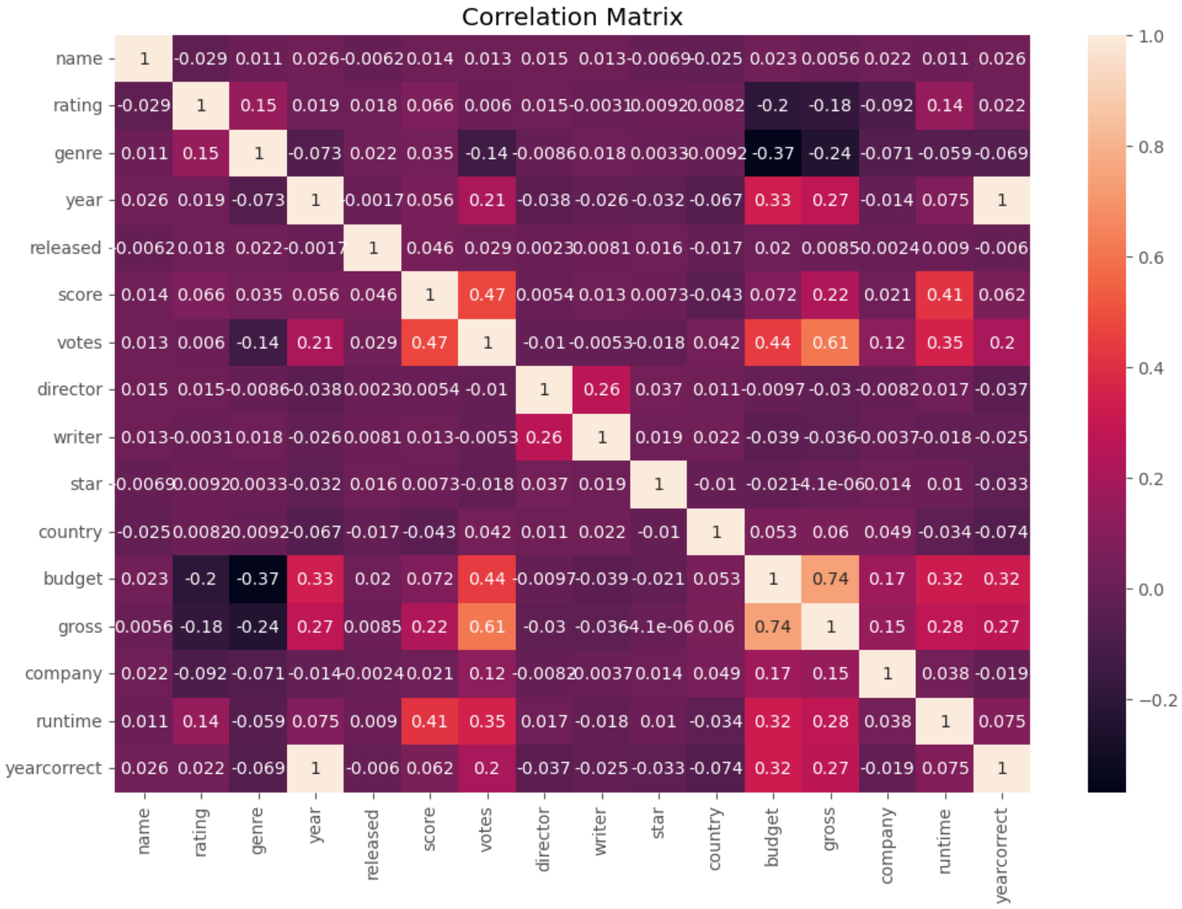
In [15]:

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

sns.heatmap(correlation_matrix, annot=True)

plt.title('Correlation Matrix')

plt.show()
#now that our heatmap includes other categories we can see that gross is still most
```



	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime	yearcorrect
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	-0.036	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

```
In [16]: df_numerized.corr()
```

Out[16]:

	name	rating	genre	year	released	score	votes	director
name	1.000000	-0.029234	0.010996	0.025542	-0.006152	0.014450	0.012615	0.015246
rating	-0.029234	1.000000	0.147796	0.019499	0.018083	0.065983	0.006031	0.014656
genre	0.010996	0.147796	1.000000	-0.073167	0.022142	0.035106	-0.135990	-0.008553
year	0.025542	0.019499	-0.073167	1.000000	-0.001740	0.056386	0.206021	-0.038354
released	-0.006152	0.018083	0.022142	-0.001740	1.000000	0.045874	0.028833	0.002308
score	0.014450	0.065983	0.035106	0.056386	0.045874	1.000000	0.474256	0.005413
votes	0.012615	0.006031	-0.135990	0.206021	0.028833	0.474256	1.000000	-0.010376
director	0.015246	0.014656	-0.008553	-0.038354	0.002308	0.005413	-0.010376	1.000000
writer	0.012880	-0.003149	0.017578	-0.025908	0.008072	0.012843	-0.005316	0.261735
star	-0.006882	0.009196	0.003341	-0.032157	0.015706	0.007296	-0.017638	0.036593
country	-0.025490	0.008230	-0.009164	-0.066748	-0.017228	-0.043051	0.041551	0.011133
budget	0.023392	-0.203946	-0.368523	0.327722	0.019952	0.072001	0.439675	-0.009662
gross	0.005639	-0.181906	-0.244101	0.274321	0.008501	0.222556	0.614751	-0.029560
company	0.021697	-0.092357	-0.071334	-0.014333	-0.002407	0.020656	0.118470	-0.008223
runtime	0.010850	0.140792	-0.059237	0.075077	0.008975	0.414068	0.352303	0.017433
yearcorrect	0.025542	0.022021	-0.069147	0.998726	-0.005989	0.061923	0.203098	-0.037371

In [17]:

```
correlation_mat = df_numerized.corr()
corr_pairs = correlation_mat.unstack()
corr_pairs
```

Out[17]:

```
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
             year           0.019499
             released       0.018083
             score          0.065983
             votes          0.006031
             director       0.014656
             writer        -0.003149
                              ...
runtime      director       0.017433
             writer        -0.017561
             star           0.010108
             country       -0.034477
             budget         0.318695
             gross          0.275796
             company        0.037585
             runtime        1.000000
             yearcorrect    0.075294
yearcorrect  name           0.025542
             rating         0.022021
             genre         -0.069147
             year           0.998726
             released      -0.005989
             score          0.061923
             votes          0.203098
             director      -0.037371
             writer        -0.025495
             star          -0.032687
             country       -0.073569
             budget         0.320312
             gross          0.268721
             company       -0.018806
             runtime        0.075294
             yearcorrect    1.000000
Length: 256, dtype: float64
```

In [18]:

```
sorted_pairs = corr_pairs.sort_values()

sorted_pairs
```

```
Out[18]: 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
         yearcorrect  genre         -0.069147
         year         country       -0.066748
         country      year          -0.066748
         genre        runtime       -0.059237
         runtime      genre         -0.059237
         score        country       -0.043051
                                       ...
         budget       votes          0.439675
         score        votes          0.474256
         votes        score          0.474256
         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
         Length: 256, dtype: float64
```

```
In [19]: high_corr = sorted_pairs[(sorted_pairs) > 0.5]
         high_corr
         #the non 1.0 values clearly show us the correlations with gross. Votes and budget a
```

```
Out[19]: 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
```