

```
# Import libraries
import pandas as pd
import numpy as np
import seaborn as sns


import matplotlib.pyplot as plt
import matplotlib.mlab as mlab
import matplotlib
plt.style.use('ggplot')
from matplotlib.pyplot import figure

%matplotlib inline
matplotlib.rcParams['figure.figsize']=(12,8)

pd.options.mode.chained_assignment=None

# Load the data
df =pd.read_csv("movies.csv")
```

```
# Overview of the data
df
```



	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross
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
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
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
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
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
...
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks	Joseph Ebanks	Shannon Bond	United States	7000.0	NaN



McAfee | WebAdvisor



Your download's being scanned.
We'll let you know if there's an issue.

```
# Remove missing values in the data
```

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

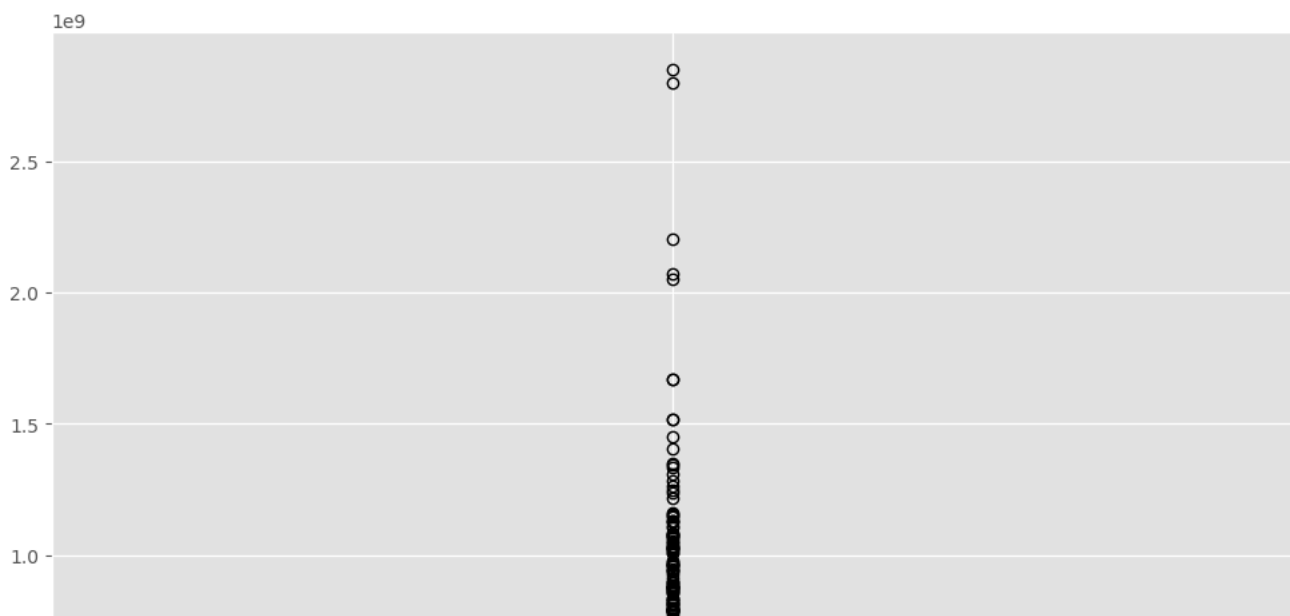
```
name - 0%
rating - 1%
genre - 0%
year - 0%
released - 0%
score - 0%
votes - 0%
director - 0%
writer - 0%
star - 0%
country - 0%
budget - 28%
gross - 2%
company - 0%
runtime - 0%
```

```
print(df.dtypes)
```

```
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
```

```
# Remove the outliers
df.boxplot(column='gross')
```

```
<Axes: >
```




Start coding or [generate](#) with AI.

```
df.drop_duplicates()
```




McAfee | WebAdvisor

Your download's being scanned.
We'll let you know if there's an issue.





	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross
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
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
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
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
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
...
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks	Joseph Ebanks	Shannon Bond	United States	7000.0	NaN



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







Your download's being scanned.
We'll let you know if there's an issue.

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross
5445	Avatar	PG-13	Action	2009	December 18, 2009 (United States)	7.8	1100000.0	James Cameron	James Cameron	Sam Worthington	United States	237000000.0	2.847246e+09
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	356000000.0	2.797501e+09
3045	Titanic	PG-13	Drama	1997	December 19, 1997 (United States)	7.8	1100000.0	James Cameron	James Cameron	Leonardo DiCaprio	United States	200000000.0	2.201647e+09
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	245000000.0	2.069522e+09
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	321000000.0	2.048360e+09
...
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks	Joseph Ebanks	Shannon Bond	United States	7000.0	NaN
February 7,													

```
sns.regplot (x="gross",y="budget",data =df)
```

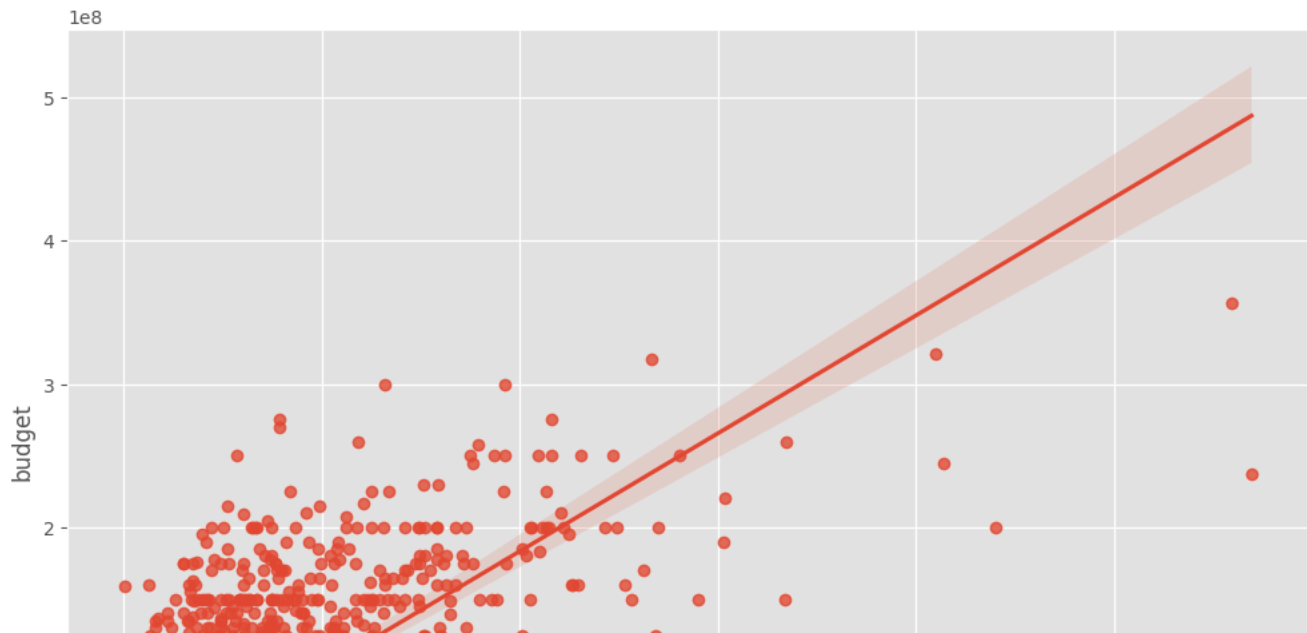


McAfee | WebAdvisor



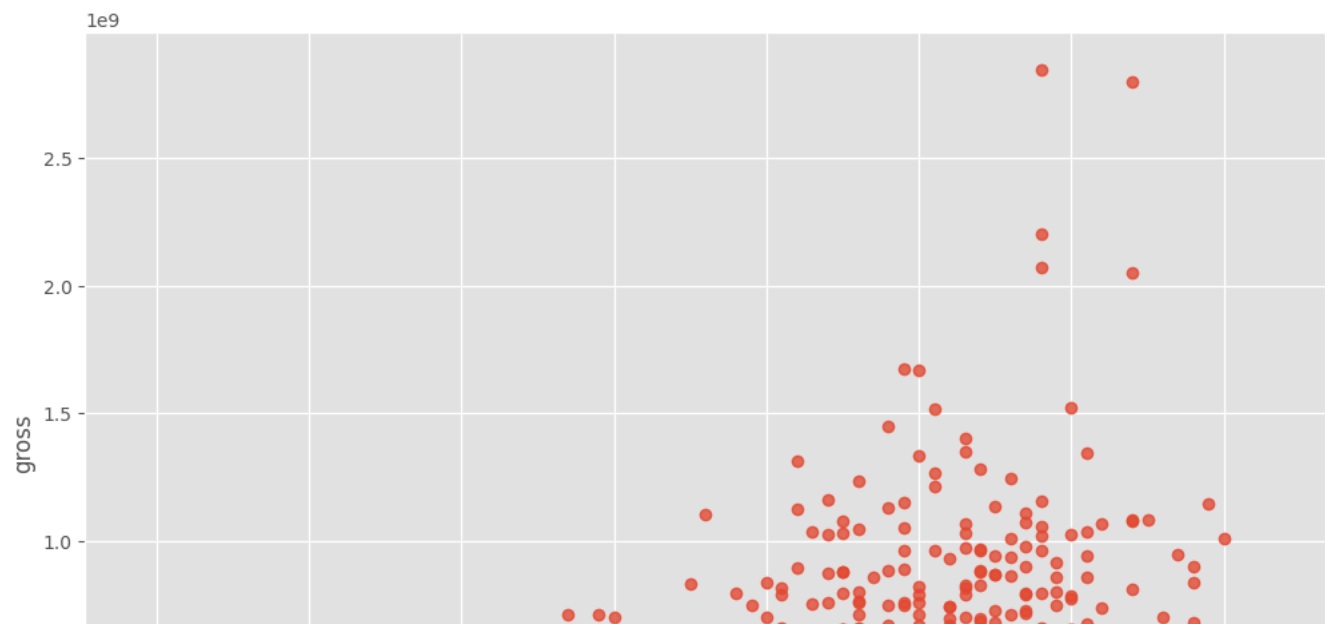
Your download's being scanned.
We'll let you know if there's an issue.

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



```
sns.regplot(x="score",y="gross",data =df)
```

 <Axes: xlabel='score', ylabel='gross'>



```
# Correlation matrix between numeric values
```

```
numeric_df = df.select_dtypes(include=['int64', 'float64'])
```

```
correlation = numeric_df.corr(method='pearson')
```



```
numeric_df = df.select_dtypes(include=['int64', 'float64'])
```

```
correlation = numeric_df.corr(method='kendall')
```



```
numeric_df = df.select_dtypes(include=['int64', 'float64'])
```

```
correlation = numeric_df.corr(method='spearman')
```



```
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")
```

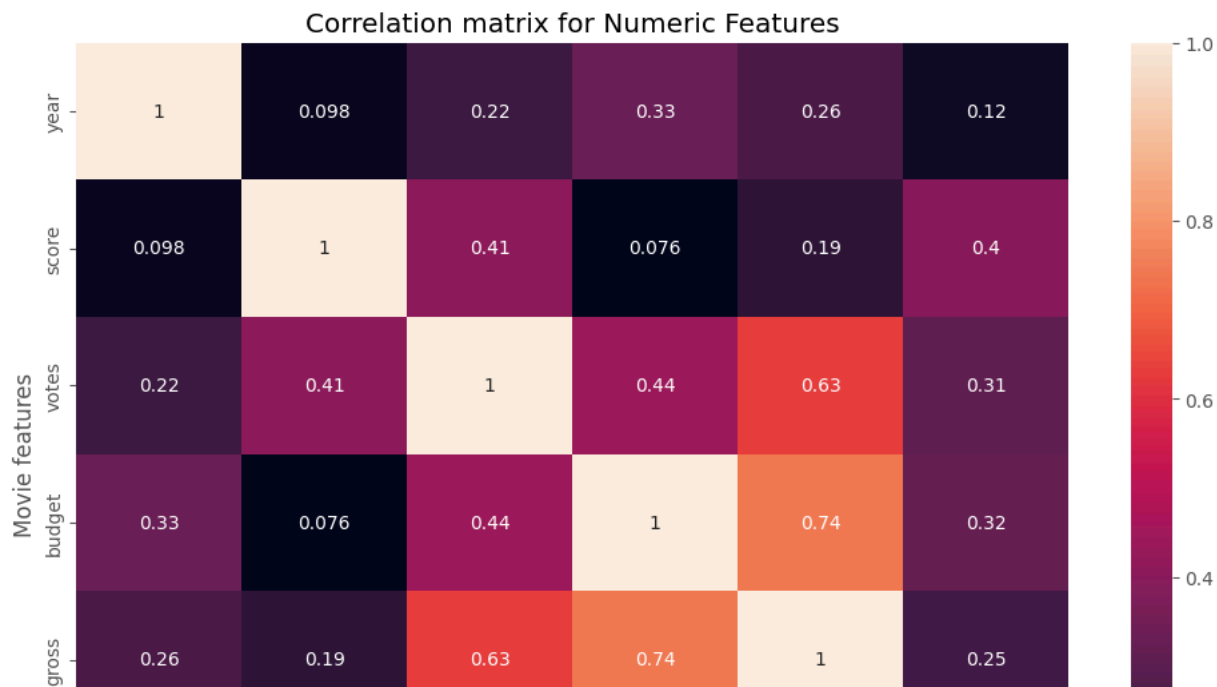


McAfee | WebAdvisor



Your download's being scanned.
We'll let you know if there's an issue.

```
plt.show()
```



```
# Using factorize - this assigns a random numeric value for each unique categorical value
```

```
df.apply(lambda x: x.factorize()[0]).corr(method='pearson')
```



	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gr
name	1.000000	0.143938	0.036367	0.965761	0.959015	-0.046733	0.287776	0.745905	0.805211	0.731565	0.142828	0.277488	0.947
rating	0.143938	1.000000	-0.086723	0.156713	0.146606	0.012595	0.099972	0.085520	0.103623	0.093116	0.000494	0.193353	0.156
genre	0.036367	-0.086723	1.000000	0.037184	0.035940	-0.002437	0.023285	0.047288	0.033688	0.038649	-0.015795	0.073008	0.036
year	0.965761	0.156713	0.037184	1.000000	0.993190	-0.044981	0.312401	0.770497	0.824770	0.756400	0.140216	0.300621	0.980
released	0.959015	0.146606	0.035940	0.993190	1.000000	-0.045761	0.299905	0.770876	0.819617	0.754468	0.148468	0.285691	0.976
score	-0.046733	0.012595	-0.002437	-0.044981	-0.045761	1.000000	-0.009749	-0.022687	-0.034685	-0.009896	0.023097	-0.012642	-0.047
votes	0.287776	0.099972	0.023285	0.312401	0.299905	-0.009749	1.000000	0.192220	0.224122	0.179601	-0.045914	0.398519	0.286
director	0.745905	0.085520	0.047288	0.770497	0.770876	-0.022687	0.192220	1.000000	0.748340	0.682385	0.155471	0.106617	0.750
writer	0.805211	0.103623	0.033688	0.824770	0.819617	-0.034685	0.224122	0.748340	1.000000	0.675685	0.157202	0.187238	0.805
star	0.731565	0.093116	0.038649	0.756400	0.754468	-0.009896	0.179601	0.682385	0.675685	1.000000	0.182045	0.107991	0.735
country	0.142828	0.000494	-0.015795	0.140216	0.148468	0.023097	-0.045914	0.155471	0.157202	0.182045	1.000000	-0.082082	0.133
budget	0.277488	0.193353	0.073008	0.300621	0.285691	-0.012642	0.398519	0.106617	0.187238	0.107991	-0.082082	1.000000	0.285
gross	0.947324	0.158582	0.038616	0.980873	0.976423	-0.047041	0.286180	0.750911	0.805576	0.735680	0.133982	0.285832	1.000
company	0.591667	-0.028035	0.009566	0.601571	0.607954	-0.028432	0.008900	0.552258	0.546151	0.527116	0.226346	-0.092249	0.588
runtime	0.048955	0.032741	0.001462	0.050647	0.048235	0.026436	0.106024	-0.011070	0.032264	0.035392	0.124154	0.112097	0.042

```
correlation_matrix = df.apply(lambda x: x.factorize()[0]).corr(method='pearson')
```

```
sns.heatmap(correlation_matrix, annot = True)
```

```
plt.title("Correlation matrix for Movies")
```

```
plt.xlabel("Movie features")
```

```
plt.ylabel("Movie features")
```

```
plt.show()
```

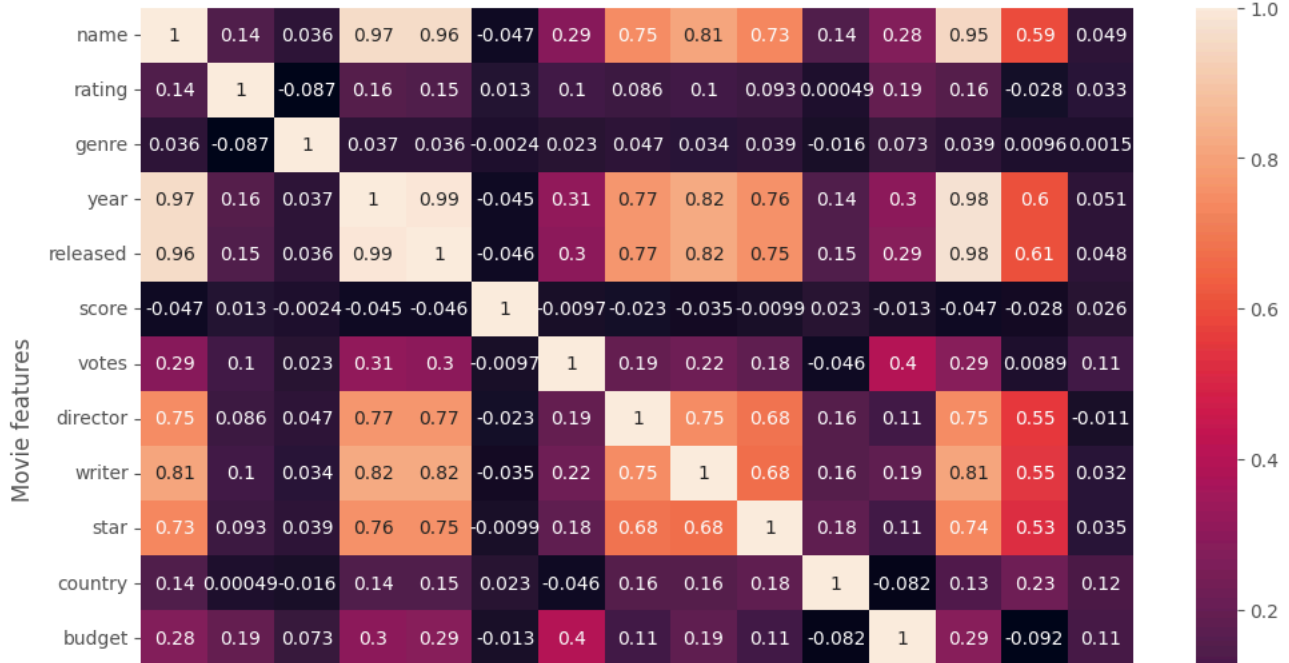


McAfee | WebAdvisor

Your download's being scanned.
We'll let you know if there's an issue.



Correlation matrix for Movies



```
correlation_mat = df.apply(lambda x: x.factorize()[0]).corr()
```

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

```
print(corr_pairs)
```

```
name    name    1.000000
        rating  0.143938
        genre   0.036367
        year    0.965761
        released 0.959015
        ...
runtime country 0.124154
runtime budget  0.112097
runtime gross   0.042978
runtime company 0.005137
runtime runtime 1.000000
Length: 225, dtype: float64
```

```
sorted_pairs = corr_pairs.sort_values(kind="quicksort")
```

```
print(sorted_pairs)
```

```
budget company -0.092249
company budget -0.092249
genre rating  -0.086723
rating genre   -0.086723
budget country -0.082082
        ...
year year      1.000000
genre genre    1.000000
rating rating  1.000000
company company 1.000000
runtime runtime 1.000000
Length: 225, dtype: float64
```

```
strong_pairs = sorted_pairs[abs(sorted_pairs) > 0.5]
```

```
print(strong_pairs)
```

```
star company 0.527116
company star  0.527116
        writer 0.546151
writer company 0.546151
director company 0.552258
        ...
year year      1.000000
genre genre    1.000000
```



McAfee | WebAdvisor



Your download's being scanned.
We'll let you know if there's an issue.

```
rating    rating    1.000000
company    company    1.000000
runtime    runtime    1.000000
Length: 71, dtype: float64
```

```
# Looking at the top 15 compaes by gross revenue
```

```
CompanyGrossSum = df.groupby('company')[["gross"]].sum()
```

```
CompanyGrossSumSorted = CompanyGrossSum.sort_values('gross', ascending = False)[:15]
```

```
CompanyGrossSumSorted = CompanyGrossSumSorted['gross'].astype('int64')
```

```
CompanyGrossSumSorted
```



	gross
company	
Warner Bros.	56491421806
Universal Pictures	52514188890
Columbia Pictures	43008941346
Paramount Pictures	40493607415
Twentieth Century Fox	40257053857
Walt Disney Pictures	36327887792
New Line Cinema	19883797684
Marvel Studios	15065592411
DreamWorks Animation	11873612858
Touchstone Pictures	11795832638
Dreamworks Pictures	11635441081
Metro-Goldwyn-Mayer (MGM)	9230230105
Summit Entertainment	8373718838
Pixar Animation Studios	7886344526
Fox 2000 Pictures	7443502667

```
df['Year'] = df['released'].astype(str).str[:4]
df
```



McAfee | WebAdvisor



Your download's being scanned.
We'll let you know if there's an issue.

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross
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
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
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
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
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
...
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks	Joseph Ebanks	Shannon Bond	United States	7000.0	NaN
7664	Dream Round	NaN	Comedy	2020	February 7, 2020 (United States)	4.7	36.0	Dusty Dukatz	Lisa Huston	Michael Saquella	United States	NaN	NaN

```
df.groupby(['company', 'year'])["gross"].sum()
```

	company	year	gross
	"DIA" Productions GmbH & Co. KG	2003	44350926.0
	"Weathering With You" Film Partners	2019	193457467.0
	.406 Production	1996	10580.0
	1+2 Seisaku linkai	2000	1196218.0
	10 West Studios	2010	814906.0

	i am OTHER	2015	17986781.0
	i5 Films	2001	10031529.0
	iDeal Partners Film Fund	2013	506303.0
	micro_scope	2010	7099598.0
	thefyzz	2017	62198461.0

4536 rows x 1 columns



McAfee | WebAdvisor



Your download's being scanned.
We'll let you know if there's an issue.

```
CompanyGrossSum = df.groupby(['company', 'year'])["gross"].sum()

CompanyGrossSumSorted = CompanyGrossSum.sort_values(['gross', 'company', 'year'], ascending = False)[:15]

CompanyGrossSumSorted = CompanyGrossSumSorted['gross'].astype('int64')

CompanyGrossSumSorted
```



		gross
company	year	
Walt Disney Pictures	2019	5773131804
Marvel Studios	2018	4018631866
Universal Pictures	2015	3834354888
Twentieth Century Fox	2009	3793491246
Walt Disney Pictures	2017	3789382071
Paramount Pictures	2011	3565705182
Warner Bros.	2010	3300479986
	2011	3223799224
Walt Disney Pictures	2010	3104474158
Paramount Pictures	2014	3071298586
Columbia Pictures	2006	2934631933
	2019	2932757449
Marvel Studios	2019	2797501328
Warner Bros.	2018	2774168962
Columbia Pictures	2011	2738363306

dtype: int64

```
CompanyGrossSum = df.groupby(['company'])["gross"].sum()

CompanyGrossSumSorted = CompanyGrossSum.sort_values(['gross', 'company'], ascending = False)[:15]

CompanyGrossSumSorted = CompanyGrossSumSorted['gross'].astype('int64')

CompanyGrossSumSorted
```



McAfee | WebAdvisor



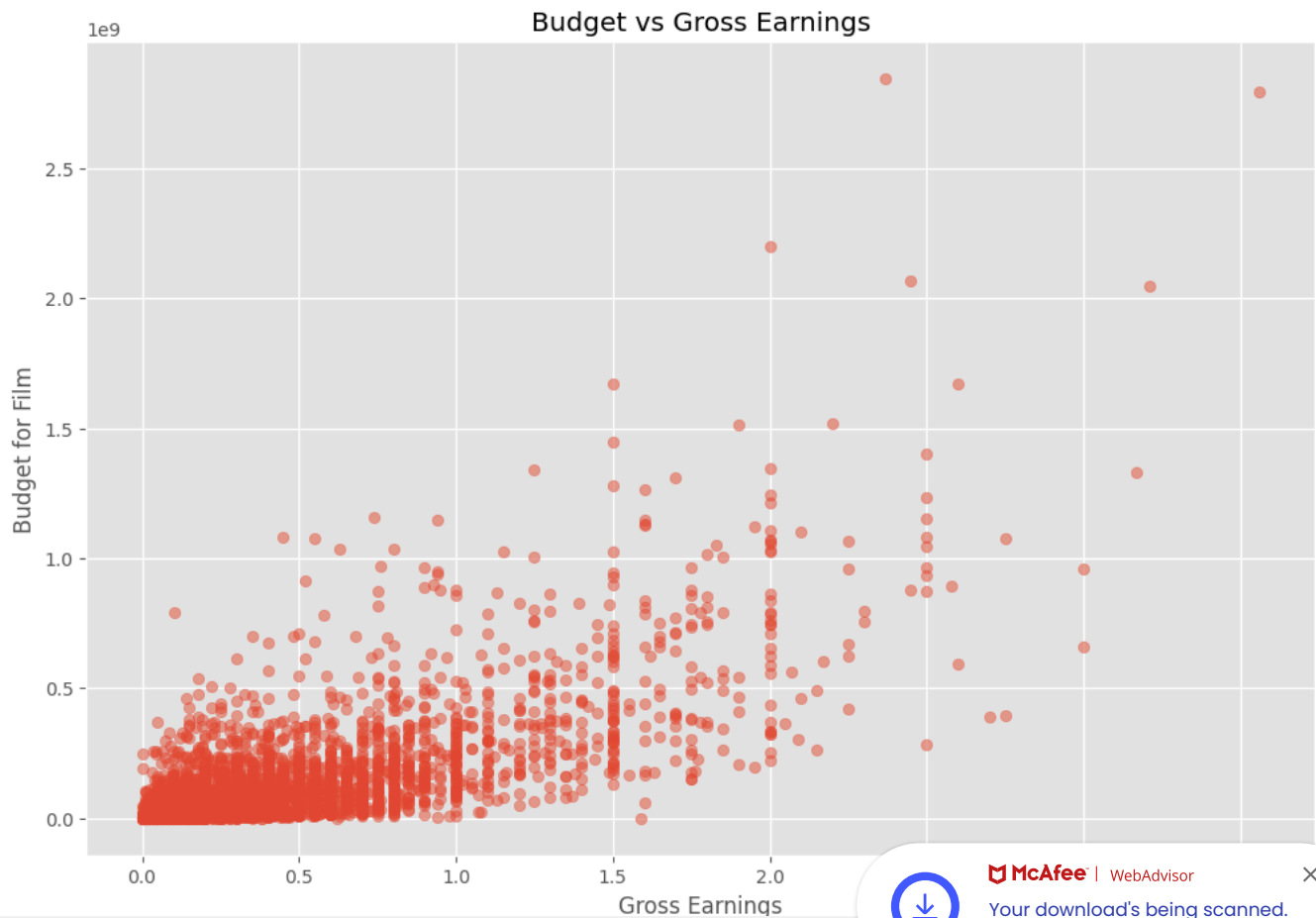
Your download's being scanned.
We'll let you know if there's an issue.



	gross
company	
Warner Bros.	56491421806
Universal Pictures	52514188890
Columbia Pictures	43008941346
Paramount Pictures	40493607415
Twentieth Century Fox	40257053857
Walt Disney Pictures	36327887792
New Line Cinema	19883797684
Marvel Studios	15065592411
DreamWorks Animation	11873612858
Touchstone Pictures	11795832638
Dreamworks Pictures	11635441081
Metro-Goldwyn-Mayer (MGM)	9230230105
Summit Entertainment	8373718838
Pixar Animation Studios	7886344526
Fox 2000 Pictures	7443502667

dtype: int64

```
plt.scatter(x=df['budget'], y=df['gross'], alpha=0.5)
plt.title('Budget vs Gross Earnings')
plt.xlabel('Gross Earnings')
plt.ylabel('Budget for Film')
plt.show()
```



McAfee | WebAdvisor

Your download's being scanned.
We'll let you know if there's an issue.

```
# Review the data set after adjustment
df
```

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross
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
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
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
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
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
...
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks	Joseph Ebanks	Shannon Bond	United States	7000.0	NaN

```
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
```

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime
0	6587	6	6	1980	1705	8.4	927000.0	2589	4014	1047	54	19000000.0	46998772.0	2319	146.0
1	5573	6	1	1980	1492	5.8	65000.0	2269	1632	327	55	4500000.0	58853106.0	731	104.0
2	5142	4	0	1980	1771	8.7	1200000.0	1111	2567	1745	55	18000000.0	538375067.0	1540	124.0
3	286	4	4	1980	1492	7.7	221000.0	1301	2000	2246	55	3500000.0	83453539.0	1812	88.0
4	1027	6	4	1980	1543	7.3	108000.0	1054	521	410	55	6000000.0	39846344.0	1777	98.0
...
7663	3705	-1	6	2020	2964	3.1	18.0	1500	2289	2421	55	7000.0	NaN	-1	90.0
7664	1678	-1	4	2020	1107	4.7	36.0	774	2614	1886	55	NaN	NaN	539	90.0
7665	4717	-1	6	2020	193	5.7	29.0	2061	2683	2040	55	58750.0	NaN	941	NaN
7666	2843	-1	6	2020	2817	NaN	NaN	1184	1824	450	55	15000.0	NaN	-1	120.0
7667	5394	-1	10	2020	391	5.7	7.0	2165	3344	2463	44	NaN	NaN	1787	102.0

7668 rows × 15 columns

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



McAfee | WebAdvisor



Your download's being scanned.
We'll let you know if there's an issue.



	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gr
name	1.000000	-0.008069	0.016355	0.011453	-0.011311	0.017097	0.013088	0.009079	0.009081	0.006472	-0.010737	0.023970	0.005
rating	-0.008069	1.000000	0.072423	0.008779	0.016613	-0.001314	0.033225	0.019483	-0.005921	0.013405	0.081244	-0.176002	-0.107
genre	0.016355	0.072423	1.000000	-0.081261	0.029822	0.027965	-0.145307	-0.015258	0.006567	-0.005477	-0.037615	-0.356564	-0.235
year	0.011453	0.008779	-0.081261	1.000000	-0.000695	0.097995	0.222945	-0.020795	-0.008656	-0.027242	-0.070938	0.329321	0.257
released	-0.011311	0.016613	0.029822	-0.000695	1.000000	0.042788	0.016097	-0.001478	-0.002404	0.015777	-0.020427	0.014683	0.001
score	0.017097	-0.001314	0.027965	0.097995	0.042788	1.000000	0.409182	0.009559	0.019416	-0.001609	-0.133348	0.076254	0.186
votes	0.013088	0.033225	-0.145307	0.222945	0.016097	0.409182	1.000000	0.000260	0.000892	-0.019282	0.073625	0.442429	0.630
director	0.009079	0.019483	-0.015258	-0.020795	-0.001478	0.009559	0.000260	1.000000	0.299067	0.039234	0.017490	-0.012272	-0.014
writer	0.009081	-0.005921	0.006567	-0.008656	-0.002404	0.019416	0.000892	0.299067	1.000000	0.027245	0.015343	-0.039451	-0.025



McAfee | WebAdvisor



Your download's being scanned.
We'll let you know if there's an issue.