Movie Correlation Analysis

Objective

- To analyze the correlation between movie features and gross earning.
- Whether the popularity of brand has impact on gross earning?

Data Source

link to the dataset

Key findings

- Strong positive correlation between budget and gross earnings
- Votes and budget have the highest correlation on gross earnings

Import Libraries

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

Set matplotlib style and figure size

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

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

Get to know the data

```
df=pd.read_csv(r"movies.csv")
df.head(5)
```

	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	N
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	

Handle missing values

```
df.isna()

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

```
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 - 0.0005216484089723526%
```

```
df=df.dropna()
```

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

```
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%
```

Handle data types

```
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
```

Create column for year from released date

```
df['correct_year']=df['released'].str.extract(pat='([0-9]{4})').astype('int64')
df['correct_year'].head()
```

```
0    1980
1    1980
2    1980
3    1980
4    1980
Name: correct_year, dtype: int64
```

Sort by gross, desc

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

```
5445    2.847246e+09
7445    2.797501e+09
3045    2.201647e+09
6663    2.069522e+09
7244    2.048360e+09
Name: gross, dtype: float64
```

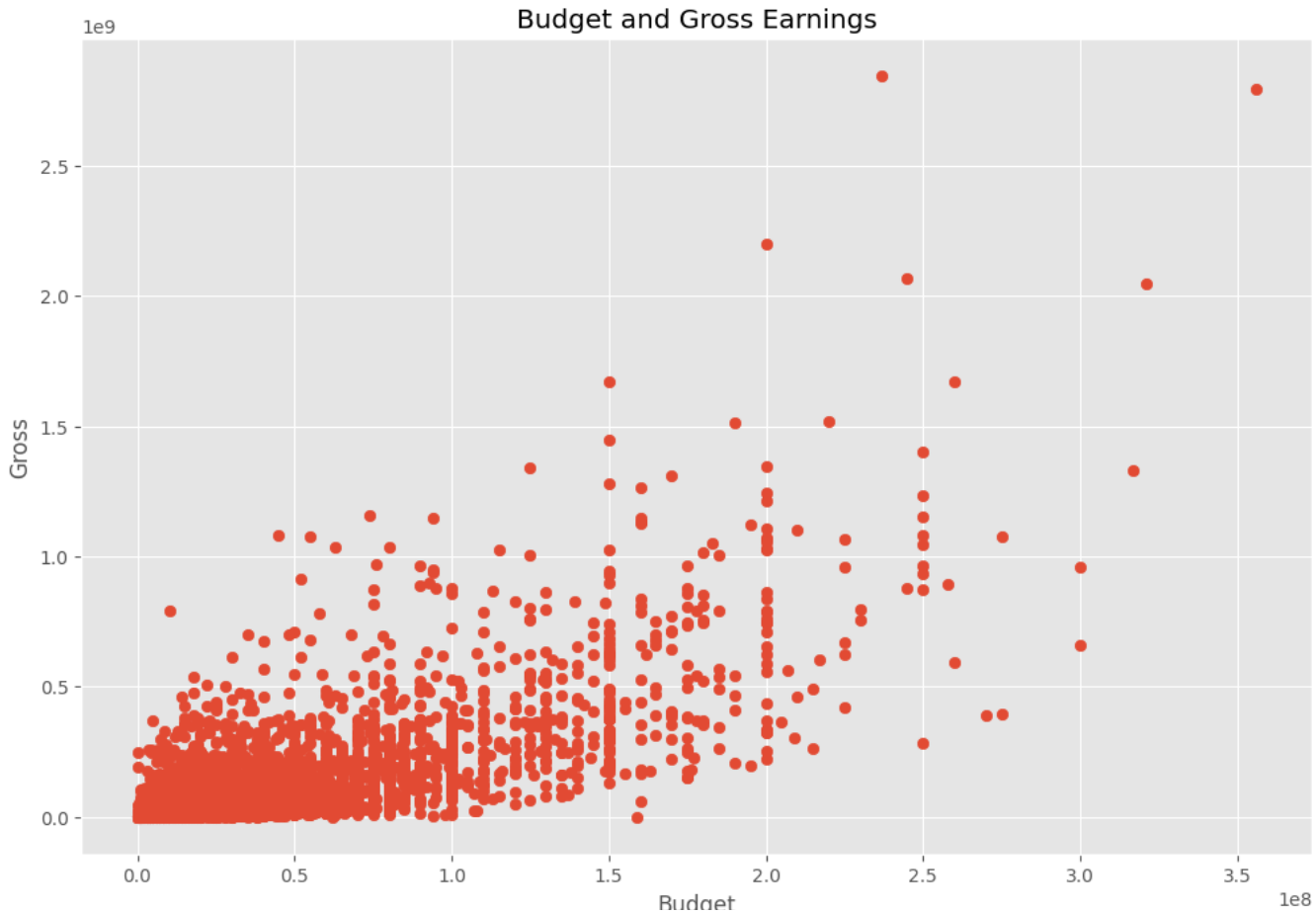
Delete Duplicates

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

	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	W

Scatter plot, relationship Budget-Gross

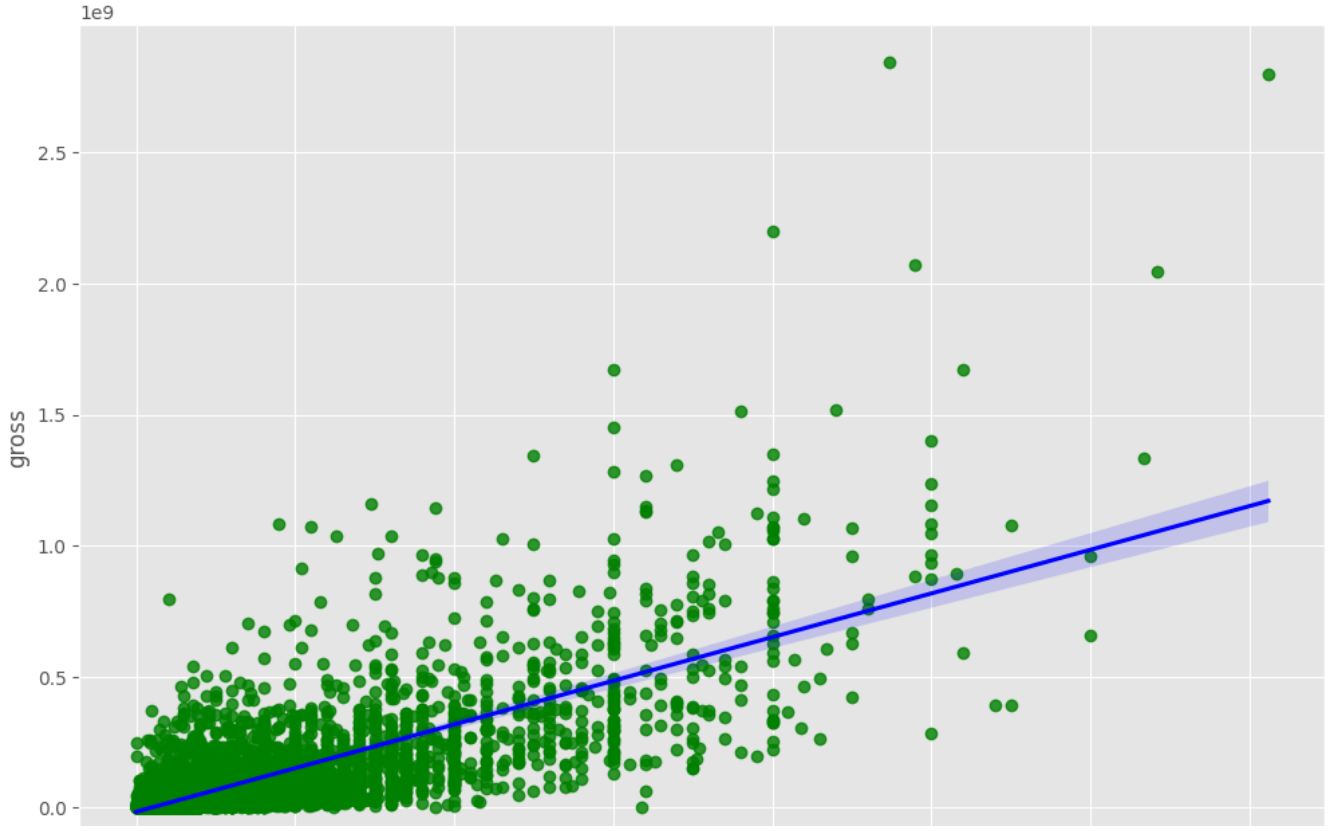
```
plt.scatter(x=df['budget'],y=df['gross'])
plt.title('Budget and Gross Earnings')
plt.xlabel('Budget')
plt.ylabel('Gross')
plt.show()
```



Regression plot, quantify relationship Budget-Gross

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

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



Numerizing Columns for creating correlation matrix

```
df_numerized=df
for col in df_numerized.columns:
  if(df_numerized[col].dtype=='object'):
    df_numerized[col]=df_numerized[col].astype('category')
    df_numerized[col]=df_numerized[col].cat.codes
df_numerized.head()
```

	name	rating	genre	year	released	score	votes	director	writer	star	coun
5445	386	5	0	2009	527	7.8	1100000.0	785	1263	1534	
7445	388	5	0	2019	137	8.4	903000.0	105	513	1470	
3045	4909	5	6	1997	534	7.8	1100000.0	785	1263	1073	
6663	3643	5	0	2015	529	7.8	876000.0	768	1806	356	
7244	389	5	0	2018	145	8.4	897000.0	105	513	1470	

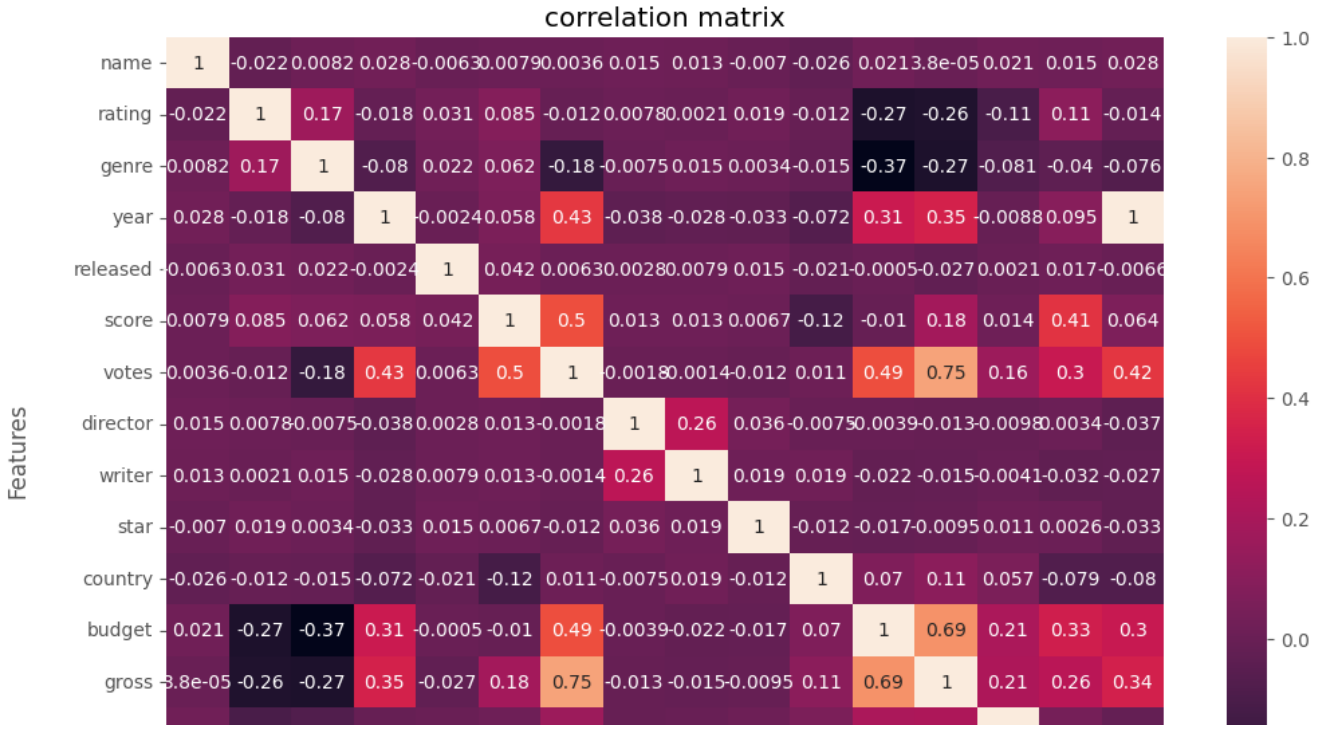
Create cor matrix

```
cor_matrix=df_numerized.corr(method='spearman')
cor_matrix
```

	name	rating	genre	year	released	score	votes	dire
name	1.000000	-0.021980	0.008213	0.027766	-0.006341	0.007866	0.003615	0.01
rating	-0.021980	1.000000	0.167778	-0.018206	0.031301	0.085237	-0.011871	0.00
genre	0.008213	0.167778	1.000000	-0.080105	0.022254	0.061615	-0.182682	-0.00
year	0.027766	-0.018206	-0.080105	1.000000	-0.002404	0.057741	0.427623	-0.03
released	-0.006341	0.031301	0.022254	-0.002404	1.000000	0.042145	0.006280	0.00
score	0.007866	0.085237	0.061615	0.057741	0.042145	1.000000	0.495409	0.01
votes	0.003615	-0.011871	-0.182682	0.427623	0.006280	0.495409	1.000000	-0.00
director	0.014933	0.007809	-0.007466	-0.037591	0.002797	0.013366	-0.001819	1.00
writer	0.013023	0.002124	0.015393	-0.027646	0.007941	0.013441	-0.001398	0.26
star	-0.007027	0.019408	0.003449	-0.032760	0.015392	0.006735	-0.011716	0.03
country	-0.026431	-0.011824	-0.015225	-0.072272	-0.021012	-0.124916	0.010930	-0.00
budget	0.021395	-0.267486	-0.372729	0.312886	-0.000495	-0.009971	0.493461	-0.00
gross	0.000038	-0.256014	-0.268314	0.351045	-0.027079	0.183192	0.745793	-0.01
company	0.021247	-0.108557	-0.080808	-0.008798	0.002086	0.013694	0.159554	-0.00
runtime	0.014849	0.110151	-0.040119	0.095444	0.017166	0.412155	0.300621	0.00
correct_year	0.027590	-0.013863	-0.075633	0.998694	-0.006623	0.063674	0.422988	-0.03

Visualize the cor_matrix

```
sns.heatmap(cor_matrix, annot=True)
plt.title('correlation matrix')
plt.xlabel('Features')
plt.ylabel('Features')
plt.show()
```



Pair and Display correlation between cols

```
corr_pairs = cor_matrix.unstack()
corr_pairs.head()
```

```
name  name        1.000000
      rating     -0.021980
      genre       0.008213
      year        0.027766
      released   -0.006341
dtype: float64
```

Strong correlation pairs

```
high_corr = corr_pairs[(abs(corr_pairs) > 0.5) & (corr_pairs != 1)]
high_corr
```

```
year          correct_year    0.998694
votes         gross           0.745793
budget        gross           0.692958
gross         votes           0.745793
              budget          0.692958
correct_year  year            0.998694
dtype: float64
```