Linear regression

January 17, 2021

[8]:
```
import pandas
from pandas import DataFrame
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
```

[9]:
```
data = pandas.read_csv("cost_revenue_clean.csv")
```

[10]:
```
data
```

[10]:

	production_budget_usd	worldwide_gross_usd
0	1000000	26
1	10000	401
2	400000	423
3	750000	450
4	10000	527
…	…	…
5029	225000000	1519479547
5030	215000000	1671640593
5031	306000000	2058662225
5032	200000000	2207615668
5033	425000000	2783918982

[5034 rows x 2 columns]

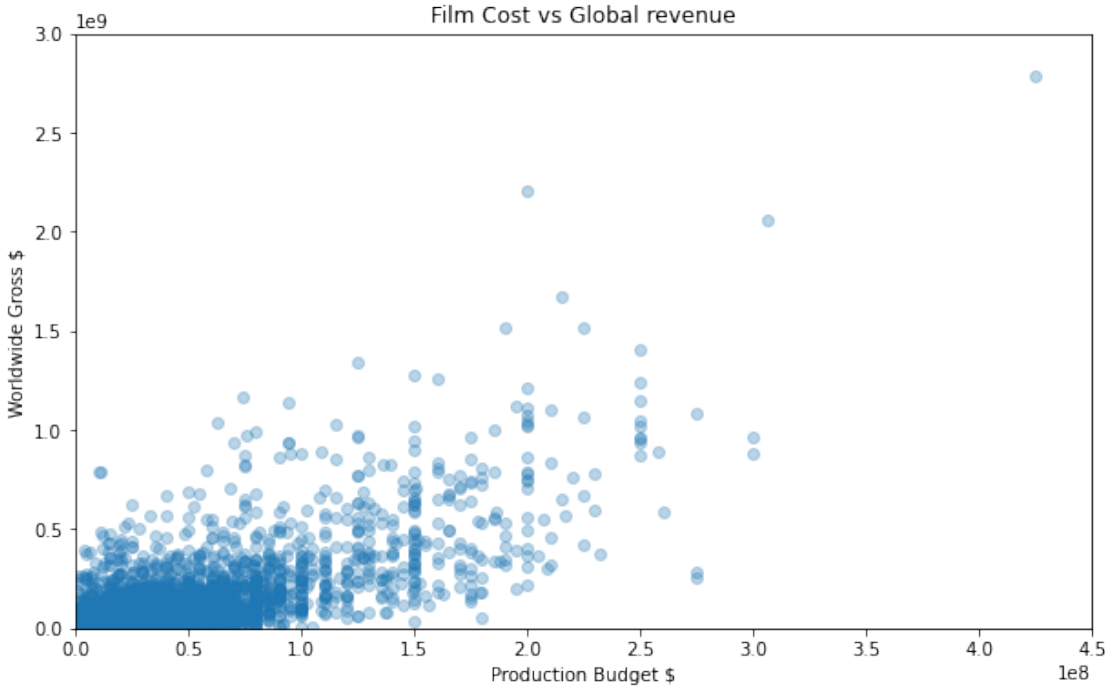
[11]:
```
data.describe()
```

[11]:

	production_budget_usd	worldwide_gross_usd
count	5.034000e+03	5.034000e+03
mean	3.290784e+07	9.515685e+07
std	4.112589e+07	1.726012e+08
min	1.100000e+03	2.600000e+01
25%	6.000000e+06	7.000000e+06
50%	1.900000e+07	3.296202e+07
75%	4.200000e+07	1.034471e+08
max	4.250000e+08	2.783919e+09

[12]:
```
X = DataFrame(data,columns=['production_budget_usd'])
y = DataFrame(data,columns=['worldwide_gross_usd'])
```

[13]:
```
plt.figure(figsize=(10,6))
plt.scatter(X,y,alpha=0.3)
plt.title('Film Cost vs Global revenue')
plt.xlabel('Production Budget $')
plt.ylabel('Worldwide Gross $')
plt.ylim(0,3000000000)
plt.xlim(0,450000000)
plt.show()
```



[16]:
```
regression=LinearRegression()
regression.fit(X, y)
```

[16]:
```
LinearRegression()
```

Slope coefficient

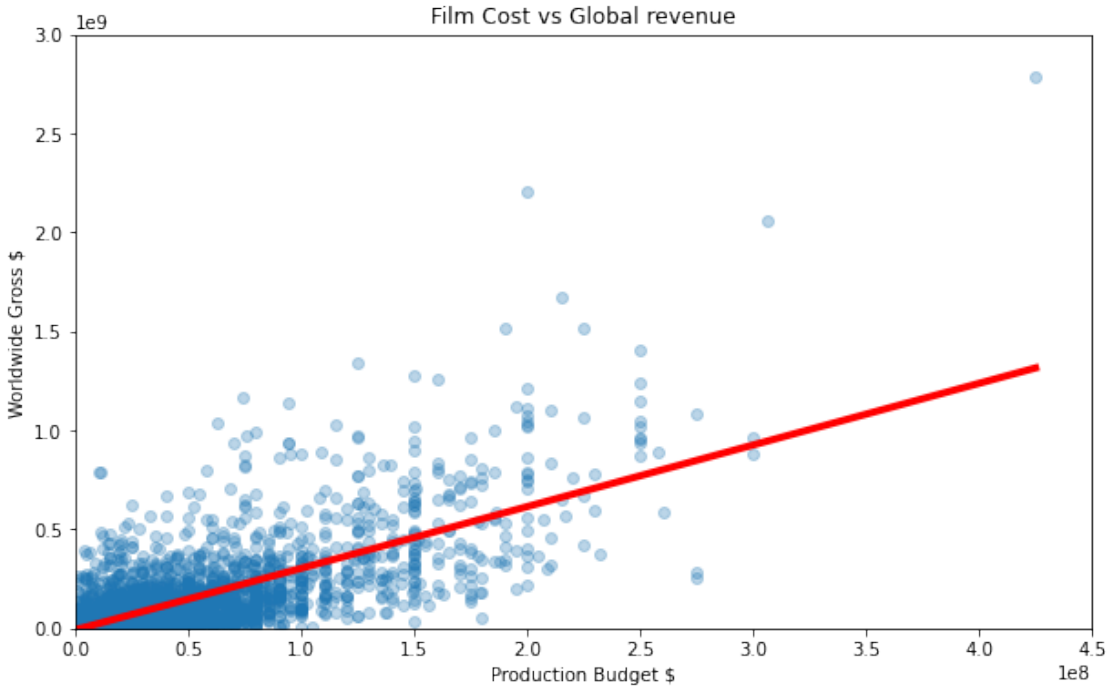
[17]:
```
regression.coef_ # theta_1
```

[17]:
```
array([[3.11150918]])
```

[18]:
```
regression.intercept_ # intercept
```

[18]:
```
array([-7236192.72913963])
```

[19]:
```
plt.figure(figsize=(10,6))
plt.scatter(X,y,alpha=0.3)
plt.plot(X,regression.predict(X),color='red',linewidth=4)
plt.title('Film Cost vs Global revenue')
plt.xlabel('Production Budget $')
plt.ylabel('Worldwide Gross $')
plt.ylim(0,3000000000)
plt.xlim(0,450000000)
plt.show()
```



[20]:
```
regression.score(X,y)
```

[20]: 0.5496485356985727

[21]:
```
type(data)
```

[21]: pandas.core.frame.DataFrame

[22]:
```
type(pandas)
```

[22]: module

[23]:
```
type(regression.intercept_)
```

[23]: numpy.ndarray

[ ]: