In [57]:
```
import pandas
from pandas import DataFrame
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
```

In [58]:
```
data = pandas.read_csv('cost_revenue_clean.csv')
```

In [59]:
```
data
```

Out[59]:

	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 × 2 columns

In [60]:
```
data.describe()
```

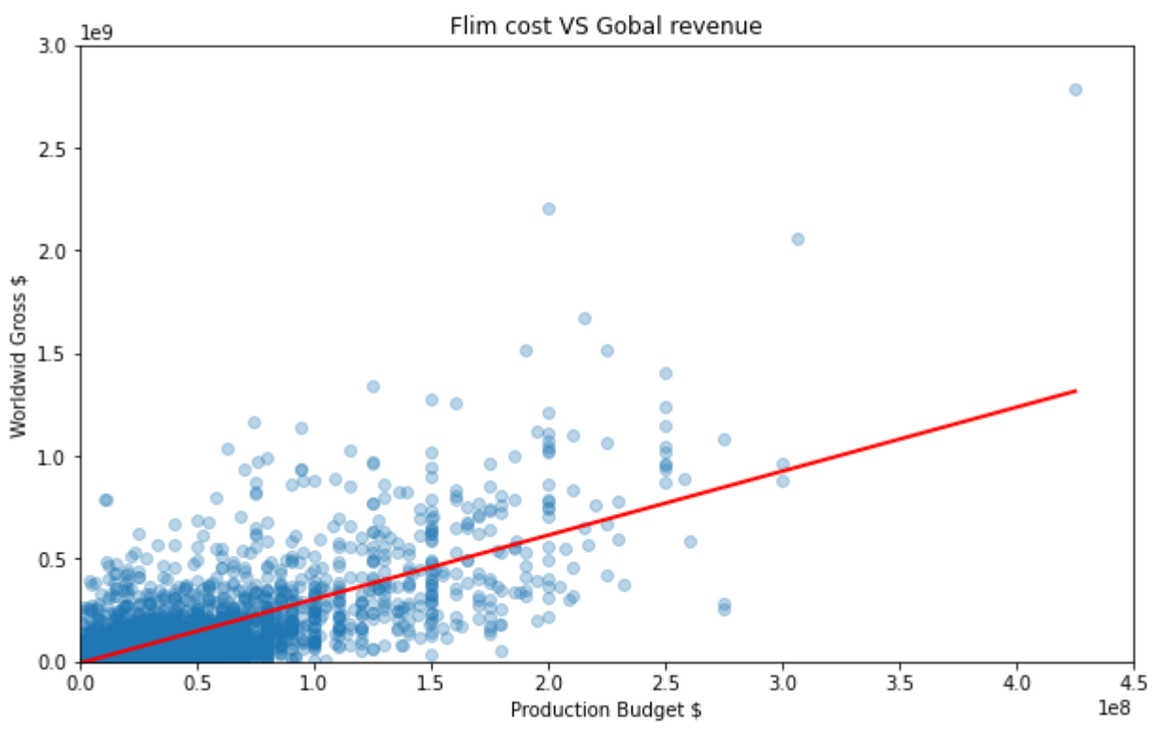
Out[60]:

	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

In [61]:
```
X= DataFrame(data,columns=['production_budget_usd'])
Y= DataFrame(data,columns=['worldwide_gross_usd'])
```

In [66]:
```
plt.figure(figsize = (10,6))
plt.scatter(X, Y, alpha=0.3)
plt.plot(X, regression.predict (X), color = 'red', linewidth =2)
plt.title('Flim cost VS Gobal revenue')
plt.xlabel('Production Budget $')
plt.ylabel('Worldwid Gross $')
plt.ylim(0, 3000000000)
plt.xlim(0, 450000000)
```

Out[66]:
```
(0.0, 450000000.0)
```



In [47]:
```
regression = LinearRegression()
regression.fit(X, Y)
```

Out[47]:
```
LinearRegression()
```

Slope coeffcient

In [49]:
```
regression.coef_#theta 0
```

Out[49]:
```
array([[3.11150918]])
```

In [ ]:
```

```

Intercept

In [50]:
```
regression.intercept_#theta 1
```

Out[50]:
```
array([-7236192.72913963])
```

In [77]:
```
regression.score(X, Y)
```

Out[77]:
```
0.5496485356985727
```

PreditRevenue for 50 million budget

f(x)=theta 0 + theta 1 (x)

f(x)=-7236192.72913963 + 3.11150918(50)

This revenue we will get 148,338,807=(-7236192.72913963) + 3.11150918(50,000,000)