In [11]:

```
import pandas
from pandas import DataFrame
import matplotlib.pyplot as plt

from sklearn.linear_model import  LinearRegression
```

In [7]:

```
data = pandas.read_csv('cost_revenue_clean.csv')
```

In [8]:

```
data.describe()
```

Out[8]:

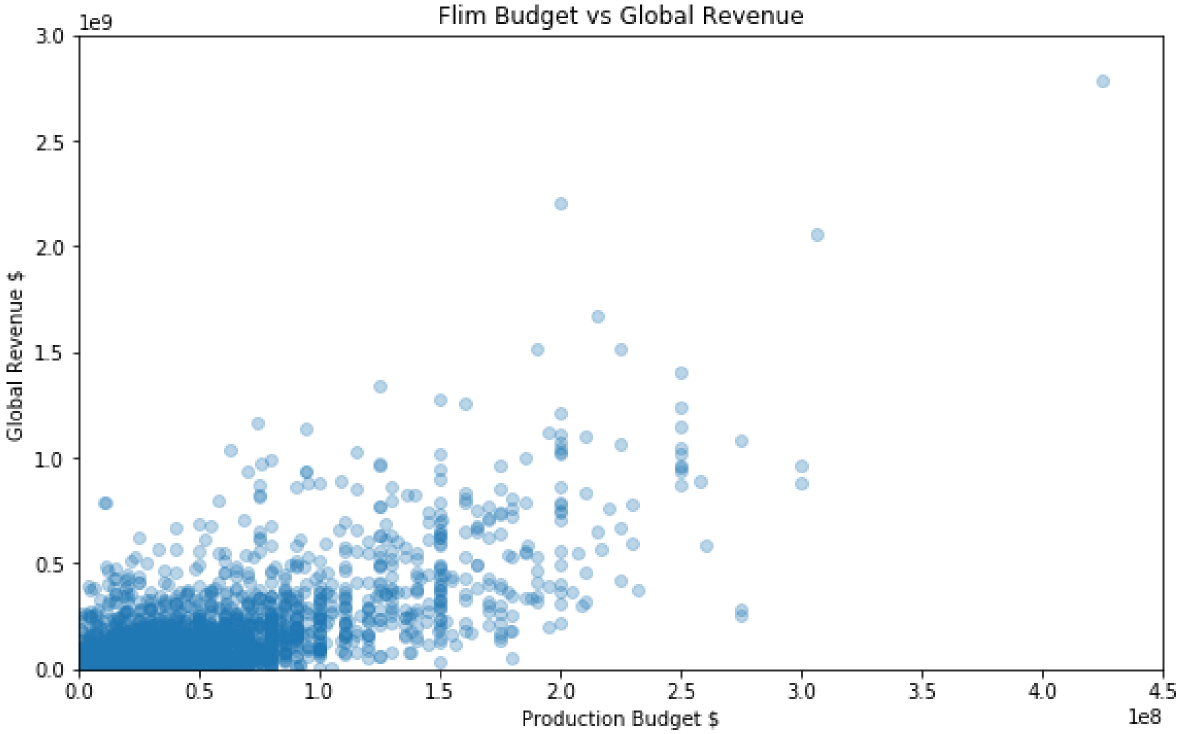
	production_budget_usd	worldwide_gross_usd
count	5.034000e+03	5.034000e+03
mean	3.290784e+07	9.515685e+07
std	4.112589e+07	1.726012e+08
min	1.100000e+03	2.600000e+01
25%	6.000000e+06	7.000000e+06
50%	1.900000e+07	3.296202e+07
75%	4.200000e+07	1.034471e+08
max	4.250000e+08	2.783919e+09

In [9]:

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

In [10]:

```
plt.figure(figsize=(10,6))
plt.scatter(X, y, alpha = 0.3)
plt.title('Flim Budget vs Global Revenue')
plt.xlabel('Production Budget $')
plt.ylabel('Global Revenue $')
plt.ylim(0, 3000000000)
plt.xlim(0, 450000000)
plt.show()
```



In [14]:

```
regression = LinearRegression()
regression.fit(X, y)
```

Out[14]:

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=Fals
e)
```

Slope Cofficient:

In [15]:

```
regression.coef_  # theta_1
```

Out[15]:

```
array([[3.11150918]])
```

In [16]:

```
#Intercept
regression.intercept_ # theta_0
```

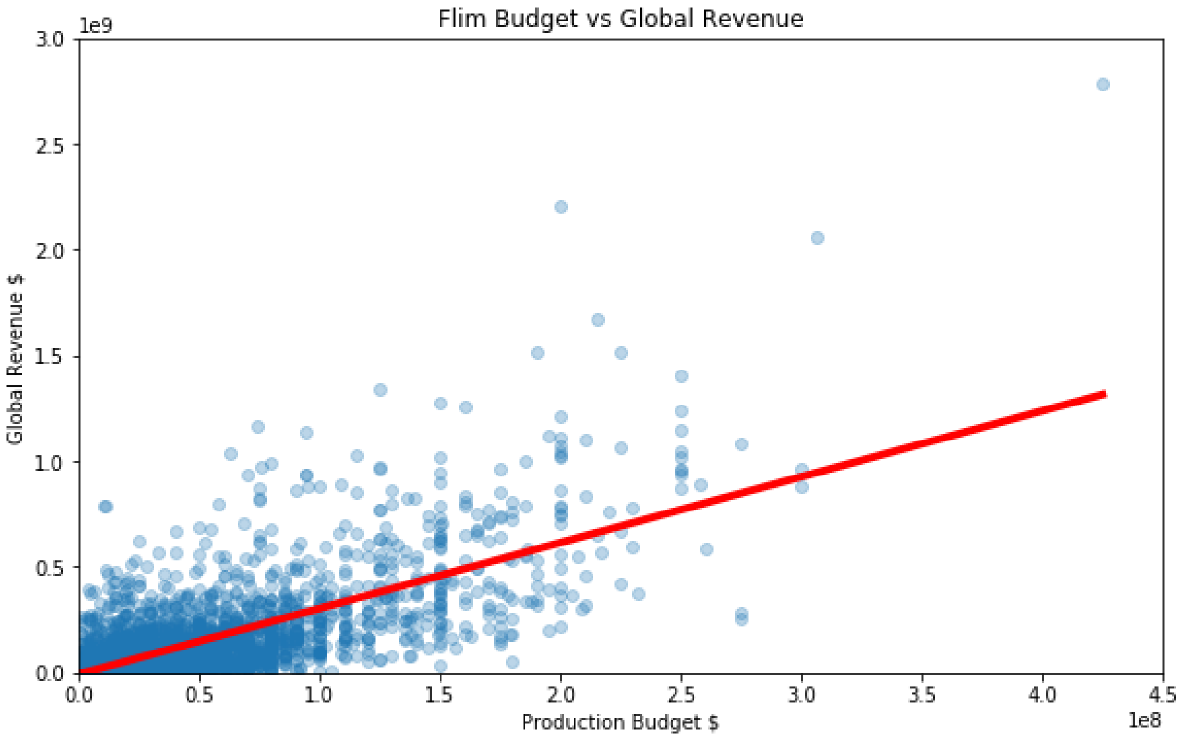
Out[16]:

```
array([-7236192.72913958])
```

In [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('Flim Budget vs Global Revenue')
plt.xlabel('Production Budget $')
plt.ylabel('Global Revenue $')
plt.ylim(0, 3000000000)
plt.xlim(0, 450000000)
plt.show()
```



Goodness of Fit r^2 or R^2

In [20]:

```
regression.score(X, y)
```

Out[20]:

```
0.5496485356985729
```

In [ ]:

```

```