Linear Regression

Problem statement : Predict home prices in monroe, new jersey (USA) from the given data.

Name : Harshvardhan Singh

Roll number : 1019161

importing useful libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import linear_model
```

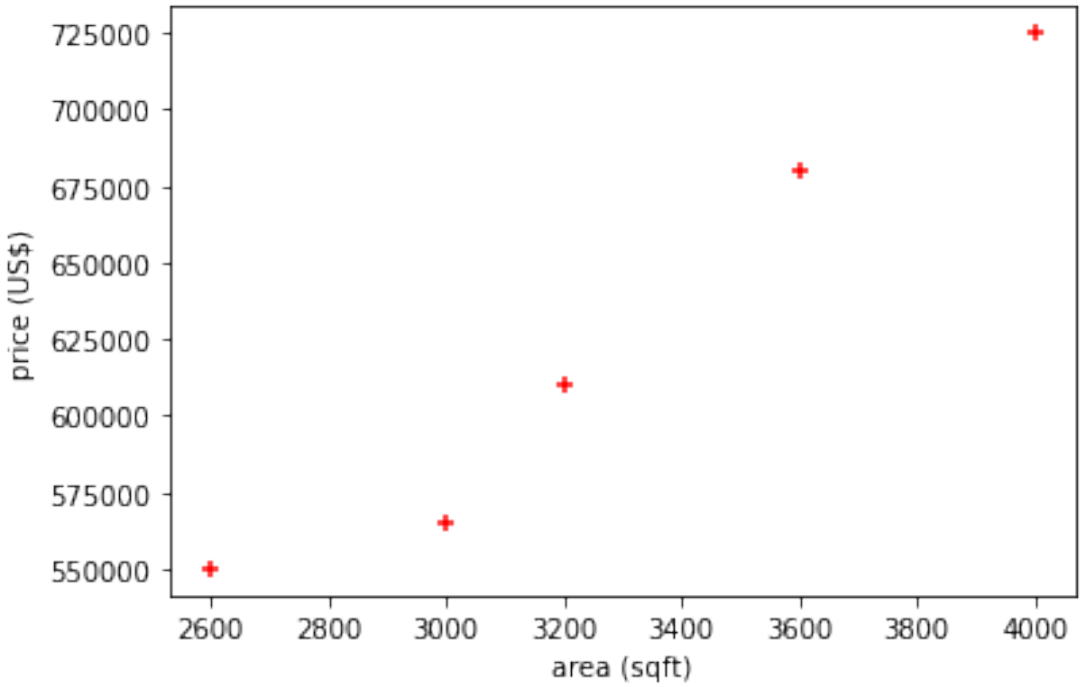
Loading Dataframe

```
df = pd.read_csv("homeprices.csv")
df.head(4)
```

```
   area   price
0  2600  550000
1  3000  565000
2  3200  610000
3  3600  680000
```

our aim is to find the best fit for the following graph:

```
plt.xlabel("area (sqft)")
plt.ylabel("price (US$)")
plt.scatter(df.area,df.price,color="red",marker="+")
plt.show()
```




```
new_df = df.drop('price',axis='columns')
new_df.head(4)
```

```
   area
0  2600
1  3000
2  3200
3  3600
```

```
price = df.price
price
```

```
0    550000
1    565000
2    610000
3    680000
4    725000
Name: price, dtype: int64
```

Creating a model , and fitting the data frame

```
reg = linear_model.LinearRegression()
reg.fit(new_df,price)
```

```
LinearRegression()
```

testing

```
reg.predict([[5000]])
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:451:
UserWarning: X does not have valid feature names, but LinearRegression
was fitted with feature names
  "X does not have valid feature names, but"

array([859554.79452055])
```

loading data for which we want to predict values

```
area_df = pd.read_csv("areas.csv")
area_df.head(4)
```

```
   area
0  1000
1  1500
2  2300
3  3540
```

listing all possible predictions

```
p=reg.predict(area_df)
p
```

```
array([ 316404.10958904,  384297.94520548,  492928.08219178,
        661304.79452055,  740061.64383562,  799808.21917808,
        926090.75342466,  650441.78082192,  825607.87671233,
        492928.08219178, 1402705.47945205, 1348390.4109589 ,
       1144708.90410959])
```

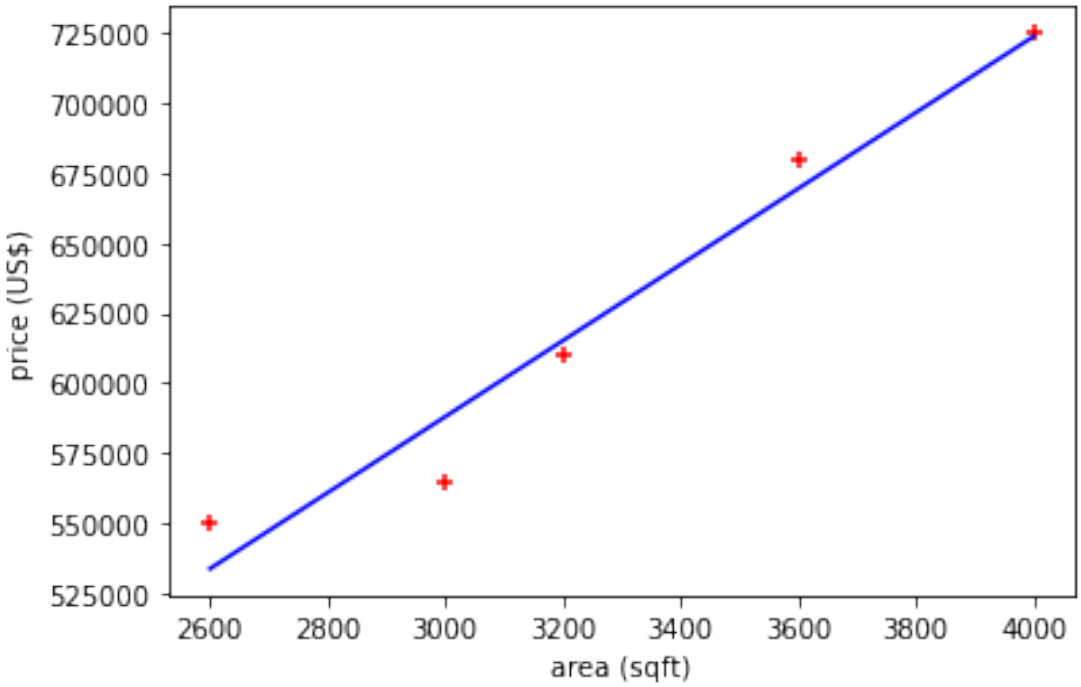
These are our Predicted values

```
area_df['prices']=p
area_df.head(5)
```

```
   area         prices
0  1000  316404.109589
1  1500  384297.945205
2  2300  492928.082192
3  3540  661304.794521
4  4120  740061.643836
```

```
area_df.to_csv("prediction.csv",index=False)
```

```
plt.xlabel("area (sqft)")
plt.ylabel("price (US$)")
plt.scatter(df.area,df.price,color="red",marker="+")
plt.plot(df.area,reg.predict(df[['area']]),color="blue")
plt.show()
```



Hence we have found the best fit for our graph