

## exercise01

November 14, 2024

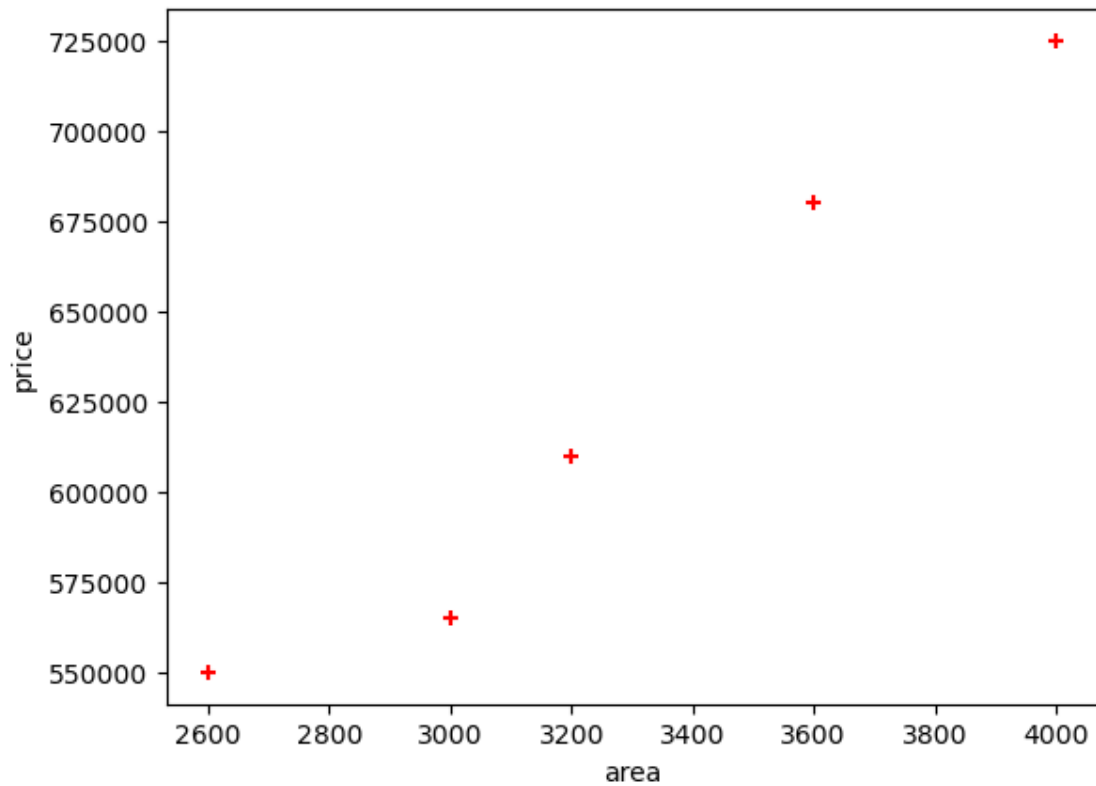
```
[15]: import pandas as pd
import numpy as np
from sklearn import linear_model
import matplotlib.pyplot as plt

df = pd.read_csv('homeprices.csv')
df
```

```
[15]:   area  price
0  2600  550000
1  3000  565000
2  3200  610000
3  3600  680000
4  4000  725000
```

```
[16]: %matplotlib inline
plt.xlabel('area')
plt.ylabel('price')
plt.scatter(df.area,df.price,color='red',marker='+')
```

```
[16]: <matplotlib.collections.PathCollection at 0x1673eaac500>
```



```
[17]: new_df = df.drop('price',axis='columns')
      new_df
```

```
[17]:   area
0  2600
1  3000
2  3200
3  3600
4  4000
```

```
[18]: price = df.price
      price
```

```
[18]: 0    550000
1    565000
2    610000
3    680000
4    725000
Name: price, dtype: int64
```

```
[19]: # Create linear regression object
reg = linear_model.LinearRegression()
reg.fit(new_df,price)
```

```
[19]: LinearRegression()
```

(1) Predict price of a home with area = 3300 sqr ft

```
[20]: reg.predict([[3300]])
```

```
d:\Projects\Anaconda_Installed\Lib\site-packages\sklearn\base.py:493:
UserWarning: X does not have valid feature names, but LinearRegression was
fitted with feature names
  warnings.warn(
```

```
[20]: array([628715.75342466])
```

```
[21]: reg.coef_
```

```
[21]: array([135.78767123])
```

```
[22]: reg.intercept_
```

```
[22]: 180616.43835616432
```

$Y = m * X + b$  (m is coefficient and b is intercept)

```
[23]: 3300*135.78767123 + 180616.43835616432
```

```
[23]: 628715.7534151643
```

(1) Predict price of a home with area = 5000 sqr ft

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

```
d:\Projects\Anaconda_Installed\Lib\site-packages\sklearn\base.py:493:
UserWarning: X does not have valid feature names, but LinearRegression was
fitted with feature names
  warnings.warn(
```

```
[24]: array([859554.79452055])
```

Generate CSV file with list of home price predictions

```
[25]: area_df = pd.read_csv("areas.csv")
area_df.head(3)
```

```
[25]:   area
0  1000
1  1500
2  2300
```

```
[26]: p = reg.predict(area_df)
      p
```

```
[26]: 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])
```

```
[27]: area_df['prices']=p
      area_df
```

```
[27]:
```

	area	prices
0	1000	3.164041e+05
1	1500	3.842979e+05
2	2300	4.929281e+05
3	3540	6.613048e+05
4	4120	7.400616e+05
5	4560	7.998082e+05
6	5490	9.260908e+05
7	3460	6.504418e+05
8	4750	8.256079e+05
9	2300	4.929281e+05
10	9000	1.402705e+06
11	8600	1.348390e+06
12	7100	1.144709e+06

```
[28]: area_df.to_csv("prediction.csv")
```