Simple linear regression - Exercise

You are given a real estate dataset.

Real estate is one of those examples that every regression course goes through as it is extremely easy to understand and there is a (almost always) certain causal relationship to be found.

The data is located in the file: 'real_estate_price_size.csv'.

You are expected to create a simple linear regression (similar to the one in the lecture), using the new data.

Apart from that, please:

- Create a scatter plot (with or without a regression line)
- Calculate the R-squared
- Display the intercept and coefficient(s)
- Using the model make a prediction about an apartment with size 750 sq.ft.

Note: In this exercise, the dependent variable is 'price', while the independent variable is 'size'.

Good luck!

Import the relevant libraries

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()

from sklearn.linear_model import LinearRegression
```

Load the data

In [2]:

```
data = pd.read_csv('real_estate_price_size.csv')
```

In [3]:
```
data.head()
```

Out[3]:

	price	size
0	234314.144	643.09
1	228581.528	656.22
2	281626.336	487.29
3	401255.608	1504.75
4	458674.256	1275.46

Create the regression

Declare the dependent and the independent variables

In [4]:
```
y = data['price']
x = data['size']
```

Explore the data

In [5]:
```
x.shape
```

Out[5]:
```
(100,)
```

In [6]:
```
y.shape
```

Out[6]:
```
(100,)
```

Transform the inputs into a matrix (2D object)

In [7]:
```
x_matrix = x.values.reshape(-1,1)
x_matrix.shape
```

Out[7]:
```
(100, 1)
```

Regression itself

In [8]:
```
reg = LinearRegression()
reg.fit(x_matrix,y)
```

Out[8]:
```
LinearRegression()
```

Calculate the R-squared

In [9]:

```
reg.score(x_matrix,y)
```

Out[9]:

```
0.7447391865847586
```

Find the intercept

In [10]:

```
reg.intercept_
```

Out[10]:

```
101912.60180122912
```

Find the coefficients

In [11]:

```
reg.coef_
```

Out[11]:

```
array([223.17874259])
```

Making predictions

You find an apartment online with a size of 750 sq.ft.

All else equal what should be its price according to the model?

In [12]:

```
new_data = pd.DataFrame(data=[656.22, 1504.75], columns=['size'])
new_data
```

Out[12]:

	size
0	656.22
1	1504.75

In [13]:

```
reg.predict(new_data)
```

Out[13]:

```
array([248366.95626666, 437740.81472046])
```

In [14]:

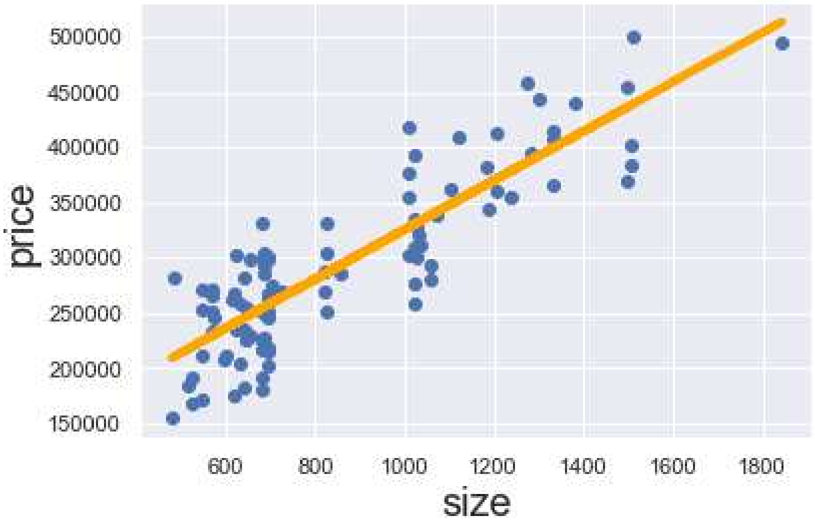
```
new_data['predicted_price'] = reg.predict(new_data)
new_data
```

Out[14]:

	size	predicted_price
0	656.22	248366.956267
1	1504.75	437740.814720

In [17]:

```
plt.scatter(x,y)
yhat = reg.coef_ * x_matrix + reg.intercept_
fig = plt.plot(x,yhat, lw=4, c='orange', label='regression line')
plt.xlabel('size', fontsize=20)
plt.ylabel('price', fontsize=20)
plt.show()
```


price
size