Simple linear regression

It is a simple real-estate sample data about price and size of houses in a particular city.

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

A simple linear regression is created using the data.

In this exercise, I have taken the dependent variable as 'price', while the independent variables is 'size'. The causal relationship I am looking for is that price is dependent upon the size of the building purchased. Let's checkout the relationship.

Importing the relevant libraries

In [10]:
```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
```

Loading the data

In [2]:
```
data = pd.read_csv('real_estate_price_size.csv')
```

In [3]:
```
data
```

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
5	245050.280	575.19
6	265129.064	570.89
7	175716.480	620.82
8	331101.344	682.26
9	218630.608	694.52
10	279555.096	1060.36
11	494778.992	1842.51
12	215472.104	694.52
13	418753.008	1009.25
14	444192.008	1300.96
15	440201.616	1379.72
16	248337.600	690.54
17	234178.160	623.94
18	225451.984	681.07
19	299416.976	1027.76
20	268125.080	620.71
21	171795.240	549.69
22	412569.472	1207.45

	price	size
23	183459.488	518.38
24	168047.264	525.81
25	362519.720	1103.30
26	271793.312	570.89
27	406852.304	1334.10
28	297760.440	681.07
29	368988.432	1496.36
...	...	...
70	276875.632	1021.95
71	181587.576	643.41
72	298926.496	656.22
73	211724.096	549.80
74	228313.024	685.48
75	286161.600	685.48
76	382120.152	1183.46
77	365863.936	1334.10
78	251560.040	682.26
79	342988.456	1188.62
80	180307.216	681.07
81	408637.816	1122.34
82	190909.056	681.07
83	282683.544	643.09
84	303597.216	685.48

	price	size
85	376253.808	1009.25
86	154282.128	479.75
87	327252.112	1028.41
88	211904.536	601.66
89	354512.112	1236.93
90	251140.656	694.52
91	338078.168	1071.55
92	298170.880	694.52
93	266684.248	698.29
94	262477.856	698.29
95	252460.400	549.80
96	310522.592	1037.44
97	383635.568	1504.75
98	225145.248	648.29
99	274922.856	705.29

100 rows × 2 columns

Descriptive Analytics of the data provided, by using pandas library

In [4]:
```
data.describe()
```

Out[4]:

	price	size
count	100.000000	100.000000
mean	292289.470160	853.024200
std	77051.727525	297.941951
min	154282.128000	479.750000
25%	234280.148000	643.330000
50%	280590.716000	696.405000
75%	335723.696000	1029.322500
max	500681.128000	1842.510000

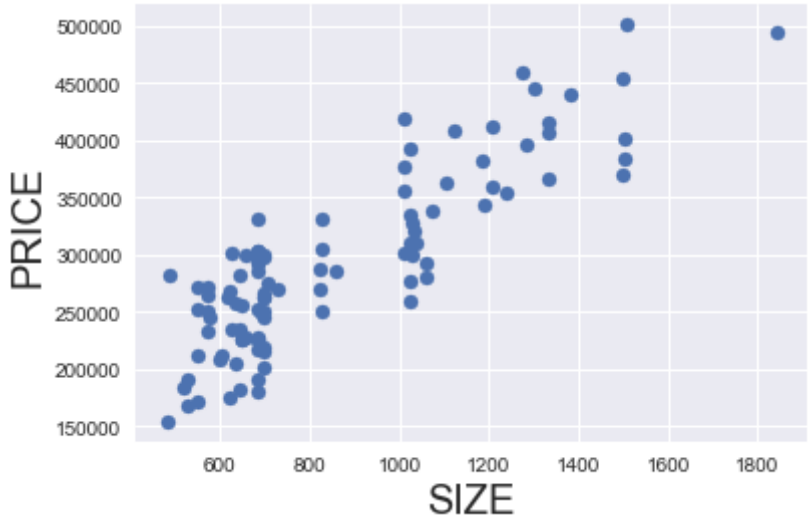
Creating the regression

Dependent and the independent variables

In [5]:
```
x1 = data['size']
y = data['price']
```

Scaterplot to visualize the data points

In [6]:
```
plt.scatter(x1, y)
plt.xlabel('SIZE', fontsize = 20)
plt.ylabel('PRICE', fontsize = 20)
plt.show()
```



The graph shows that there is a pattern to the data and price tends to increase with size of the building purchased.

Regression Analysis

In [7]:
```
x = sm.add_constant(x1)
results = sm.OLS(y,x).fit()
results.summary()
```

Out[7]:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.745
Model:	OLS	**Adj. R-squared:**	0.742
Method:	Least Squares	**F-statistic:**	285.9
Date:	Thu, 12 Mar 2020	**Prob (F-statistic):**	8.13e-31
Time:	23:02:27	**Log-Likelihood:**	-1198.3
No. Observations:	100	**AIC:**	2401.
Df Residuals:	98	**BIC:**	2406.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	1.019e+05	1.19e+04	8.550	0.000	7.83e+04	1.26e+05
size	223.1787	13.199	16.909	0.000	196.986	249.371

Omnibus:	6.262	**Durbin-Watson:**	2.267
Prob(Omnibus):	0.044	**Jarque-Bera (JB):**	2.938
Skew:	0.117	**Prob(JB):**	0.230
Kurtosis:	2.194	**Cond. No.**	2.75e+03

The constant value is 101900 and the coefficient of independent variable is 223.18. So the equation of regression line is:

y = 101900 + (223.18*x)

R-squared value is the variability of the data that is explained by the regression model. The value is considerably high (0.745). So the amount of error (or the amount of variability that is unexplained) is less. This shows that the causal relationship assumed is strong and holds.

Plot the regression line on the initial scatter

In [8]:
```
plt.scatter(x1, y)
yhat = 223.1787*x1 + 101900
fig = plt.plot(x1,yhat, lw=4, c='orange', label ='regression line')
plt.xlabel('SIZE', fontsize = 20)
plt.ylabel('PRICE', fontsize = 20)
plt.show()
```