In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
```

In [2]:

```
dataset = pd.read_excel("Linear Regression.xlsx",sheet_name=0)
```

In [3]:

```
dataset.head(10)
```

Out[3]:

	price	sqft_living	bedrooms	bathrooms	floors
0	221900	1180	3	1.00	1.0
1	538000	2570	3	2.25	2.0
2	180000	770	2	1.00	1.0
3	604000	1960	4	3.00	1.0
4	510000	1680	3	2.00	1.0
5	1225000	5420	4	4.50	1.0
6	257500	1715	3	2.25	2.0
7	291850	1060	3	1.50	1.0
8	229500	1780	3	1.00	1.0
9	323000	1890	3	2.50	2.0

1. Independent Variable(sqft_living) & dependent Variable(price)

In [4]:

```
y = dataset.price
x = dataset.sqft_living
x1 = sm.add_constant(x)
simple = sm.OLS(y,x1)
result = simple.fit()
result.summary()


## change in independent variable (x)-(sqft_living) value of dependent variable(y)-(price)
## change in one varaible 49.3 % changes  in another variable.
## If P value is less than 0.05 means Null Value is rejected and Alternative value is accep
## If P value is more than 0.05 means Null value is accepted and Alternative Value is rejec
## In this case P value is less than 0.05 so Null value is rejected and Alternative value i
## There is a Clausal effect relationship between dependent variable and independent variab
```

Out[4]:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.493
Model:	OLS	**Adj. R-squared:**	0.493
Method:	Least Squares	**F-statistic:**	2.100e+04
Date:	Wed, 31 Mar 2021	**Prob (F-statistic):**	0.00
Time:	12:44:21	**Log-Likelihood:**	-3.0027e+05
No. Observations:	21613	**AIC:**	6.005e+05
Df Residuals:	21611	**BIC:**	6.006e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-4.358e+04	4402.690	-9.899	0.000	-5.22e+04	-3.5e+04
sqft_living	280.6236	1.936	144.920	0.000	276.828	284.419

Omnibus:	14832.490	**Durbin-Watson:**	1.983
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	546444.709
Skew:	2.824	**Prob(JB):**	0.00
Kurtosis:	26.977	**Cond. No.**	5.63e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.63e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

2. Independent variable(sqft_living) & dependent variable(bathroom)

In [5]:

```
y = dataset.bathrooms
x = dataset.sqft_living
x1 = sm.add_constant(x)
simple = sm.OLS(y,x1)
result = simple.fit()
result.summary()



## change in independent variable (x)-(sqft_living) value of dependent variable(y)-(bathroo
## change in one varaible 57 % changes  in another variable.
## If P value is less than 0.05 means Null Value is rejected and Alternative value is accep
## If P value is more than 0.05 means Null value is accepted and Alternative Value is rejec
## In this case P value is less than 0.05 so Null value is rejected and Alternative value i
## There is a Clausal effect relationship between dependent variable and independent variab
```

Out[5]:

OLS Regression Results

Dep. Variable:	bathrooms	**R-squared:**	0.570
Model:	OLS	**Adj. R-squared:**	0.569
Method:	Least Squares	**F-statistic:**	2.859e+04
Date:	Wed, 31 Mar 2021	**Prob (F-statistic):**	0.00
Time:	12:44:22	**Log-Likelihood:**	-15914.
No. Observations:	21613	**AIC:**	3.183e+04
Df Residuals:	21611	**BIC:**	3.185e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.7985	0.009	93.843	0.000	0.782	0.815
sqft_living	0.0006	3.74e-06	169.089	0.000	0.001	0.001

Omnibus:	774.429	**Durbin-Watson:**	1.876
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1445.779
Skew:	0.282	**Prob(JB):**	0.00
Kurtosis:	4.134	**Cond. No.**	5.63e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.63e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

3. independent variable(sqft_living) & dependent variable(bedrooms)

In [6]:

```
y = dataset.bedrooms
x = dataset.sqft_living
x1 = sm.add_constant(x)
simple = sm.OLS(y,x1)
result = simple.fit()
result.summary()


## change in independent variable (x)-(sqft_living) value of dependent variable(y)-(bedroom
## change in one varaible 33.3% changes  in another variable.
## If P value is less than 0.05 means Null Value is rejected and Alternative value is accep
## If P value is more than 0.05 means Null value is accepted and Alternative Value is rejec
## In this case P value is less than 0.05 so Null value is rejected and Alternative value i
## There is a Clausal effect relationship between dependent variable and independent variab
```

Out[6]:

OLS Regression Results

Dep. Variable:	bedrooms	**R-squared:**	0.333
Model:	OLS	**Adj. R-squared:**	0.333
Method:	Least Squares	**F-statistic:**	1.077e+04
Date:	Wed, 31 Mar 2021	**Prob (F-statistic):**	0.00
Time:	12:44:22	**Log-Likelihood:**	-24731.
No. Observations:	21613	**AIC:**	4.947e+04
Df Residuals:	21611	**BIC:**	4.948e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.1562	0.013	168.517	0.000	2.131	2.181
sqft_living	0.0006	5.63e-06	103.766	0.000	0.001	0.001

Omnibus:	19263.322	**Durbin-Watson:**	1.964
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	11535842.412
Skew:	3.289	**Prob(JB):**	0.00
Kurtosis:	115.990	**Cond. No.**	5.63e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.63e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

4. independent variable(sqrt_living & dependent variable(floors)

In [7]:

```
y = dataset.floors
x = dataset.sqft_living
x1 = sm.add_constant(x)
simple = sm.OLS(y,x1)
result = simple.fit()
result.summary()


## change in independent variable (x)-(sqft_living) value of dependent variable(y)-(floors)
## change in one varaible 12.5% changes  in another variable.
## If P value is less than 0.05 means Null Value is rejected and Alternative value is accep
## If P value is more than 0.05 means Null value is accepted and Alternative Value is rejec
## In this case P value is less than 0.05 so Null value is rejected and Alternative value i
## There is a Clausal effect relationship between dependent variable and independent variab
```

Out[7]:

OLS Regression Results

Dep. Variable:	floors	**R-squared:**	0.125
Model:	OLS	**Adj. R-squared:**	0.125
Method:	Least Squares	**F-statistic:**	3095.
Date:	Wed, 31 Mar 2021	**Prob (F-statistic):**	0.00
Time:	12:44:22	**Log-Likelihood:**	-15902.
No. Observations:	21613	**AIC:**	3.181e+04
Df Residuals:	21611	**BIC:**	3.182e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.0615	0.009	124.812	0.000	1.045	1.078
sqft_living	0.0002	3.74e-06	55.634	0.000	0.000	0.000

Omnibus:	2258.231	**Durbin-Watson:**	1.738
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3129.668
Skew:	0.837	**Prob(JB):**	0.00
Kurtosis:	3.821	**Cond. No.**	5.63e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.63e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Multiple Regression

5. Independent Variable (sqft_living,bathrooms,bedrooms,floors) & dependent variable(Price)

In [8]:

```
y = dataset.price
x = dataset[['sqft_living','bedrooms','bathrooms','floors']]
x1 = sm.add_constant(x)
multiple = sm.OLS(y,x1)
result=multiple.fit()
result.summary()

## change in independent variable (x)-(sqft_living,bedrooms,bathrooms,floors) value of depe
## change in one varaible 50.7% changes  in another variable.
## If P value is less than 0.05 means Null Value is rejected and Alternative value is accep
## If P value is more than 0.05 means Null value is accepted and Alternative Value is rejec
## In this case P value is less than 0.05 so Null value is rejected and Alternative value i
## There is a Clausal effect relationship between dependent variable(price) and independent
## Floors is a independent variable P value is more than 0.05 means Null value is accecpted
## change in independent variable in floor doesnot have clause effect on dependent variable
```

Out[8]:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.507
Model:	OLS	**Adj. R-squared:**	0.507
Method:	Least Squares	**F-statistic:**	5554.
Date:	Wed, 31 Mar 2021	**Prob (F-statistic):**	0.00
Time:	12:44:22	**Log-Likelihood:**	-2.9996e+05
No. Observations:	21613	**AIC:**	5.999e+05
Df Residuals:	21608	**BIC:**	6.000e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.467e+04	7679.122	9.724	0.000	5.96e+04	8.97e+04
sqft_living	309.3932	3.087	100.228	0.000	303.343	315.444
bedrooms	-5.785e+04	2347.323	-24.644	0.000	-6.24e+04	-5.32e+04
bathrooms	7853.5235	3814.223	2.059	0.040	377.365	1.53e+04
floors	200.4943	3775.505	0.053	0.958	-7199.774	7600.763

Omnibus:	14450.413	**Durbin-Watson:**	1.985
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	494760.938
Skew:	2.739	**Prob(JB):**	0.00
Kurtosis:	25.790	**Cond. No.**	1.04e+04

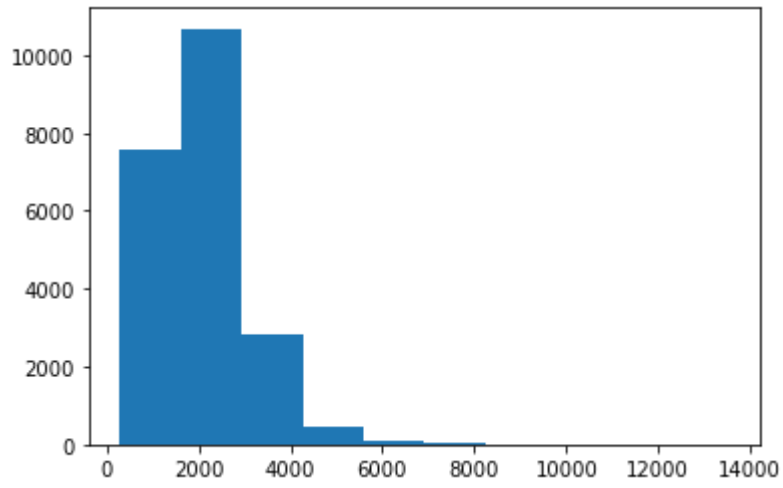
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.04e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

In [9]:

```
plt.hist(dataset.sqft_living)
```

Out[9]:

```
(array([7.5690e+03, 1.0681e+04, 2.8140e+03, 4.4100e+02, 7.7000e+01,
        2.4000e+01, 2.0000e+00, 3.0000e+00, 1.0000e+00, 1.0000e+00]),
 array([  290.,  1615.,  2940.,  4265.,  5590.,  6915.,  8240.,  9565.,
        10890., 12215., 13540.]),
 <a list of 10 Patch objects>)
```

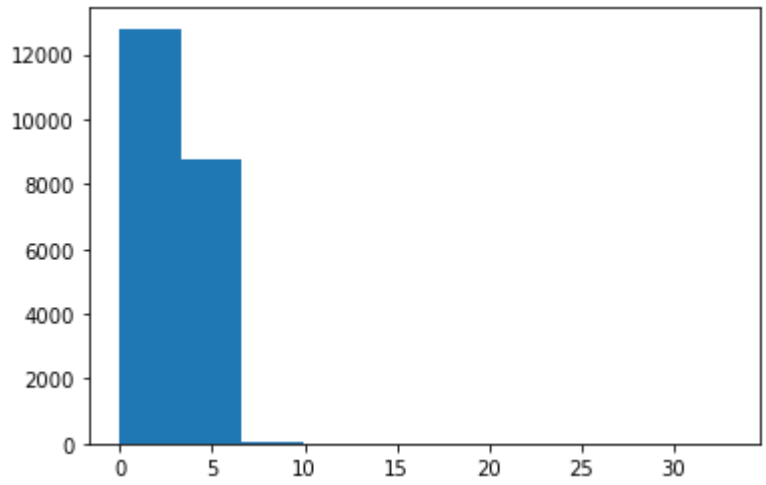


In [10]:

```
plt.hist(dataset.bedrooms)
```

Out[10]:

```
(array([1.2796e+04, 8.7550e+03, 5.7000e+01, 4.0000e+00, 0.0000e+00,
        0.0000e+00, 0.0000e+00, 0.0000e+00, 0.0000e+00, 1.0000e+00]),
 array([ 0. ,  3.3,  6.6,  9.9, 13.2, 16.5, 19.8, 23.1, 26.4, 29.7, 33. ]),
 <a list of 10 Patch objects>)
```

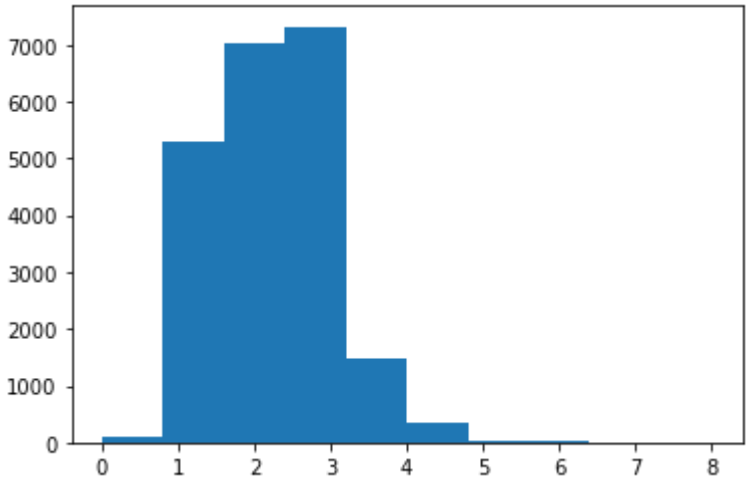


In [11]:

```
plt.hist(dataset.bathrooms)
```

Out[11]:

```
(array([8.600e+01, 5.307e+03, 7.025e+03, 7.318e+03, 1.475e+03, 3.380e+02,
        4.400e+01, 1.200e+01, 4.000e+00, 4.000e+00]),
 array([0. , 0.8, 1.6, 2.4, 3.2, 4. , 4.8, 5.6, 6.4, 7.2, 8. ]),
 <a list of 10 Patch objects>)
```

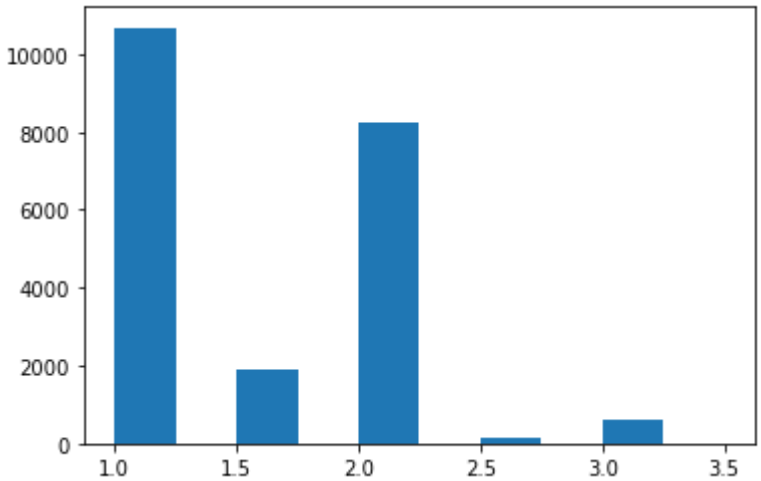


In [13]:

```
plt.hist(dataset.floors)
```

Out[13]:

```
(array([1.068e+04, 0.000e+00, 1.910e+03, 0.000e+00, 8.241e+03, 0.000e+00,
        1.610e+02, 0.000e+00, 6.130e+02, 8.000e+00]),
 array([1.  , 1.25, 1.5 , 1.75, 2.  , 2.25, 2.5 , 2.75, 3.  , 3.25, 3.5 ]),
 <a list of 10 Patch objects>)
```



In [15]:

```
from scipy.stats import pearsonr
```

In [16]:

```
stats,p=pearsonr(dataset.price,dataset.sqft_living)
```

In [17]:

```
print(stats,p)
```

```
0.7020350524336835 0.0
```

In [18]:

```
stats,p=pearsonr(dataset.price,dataset.bedrooms)
print(stats,p)
```

```
0.30834959788482247 0.0
```

In [19]:

```
stats,p=pearsonr(dataset.price,dataset.bathrooms)
print(stats,p)
```

```
0.5251375045796025 0.0
```

In [20]:

```
stats,p=pearsonr(dataset.price,dataset.floors)
print(stats,p)
```

```
0.256793884063341 1.6e-322
```

All this combination of dependnt and indepedent varaible is passed this is autocorreated

In [21]:

```
dataset.corr()
```

Out[21]:

	price	sqft_living	bedrooms	bathrooms	floors
price	1.000000	0.702035	0.308350	0.525138	0.256794
sqft_living	0.702035	1.000000	0.576671	0.754665	0.353949
bedrooms	0.308350	0.576671	1.000000	0.515884	0.175429
bathrooms	0.525138	0.754665	0.515884	1.000000	0.500653
floors	0.256794	0.353949	0.175429	0.500653	1.000000

In this project dependent variable (sqft_living,bedrooms,bathrooms) are significant to dependent variable(price)

variable(price)
Floors are not significant to dependent variable.

In [ ]: