Regression Carats vs. Price

In this notebook, you will perform a similar analysis to the one you did in the previous notebook, but using a dataset holding the weight of a diamond in carats, and the price of the corresponding diamond in dollars.

To get started, let's read in the necessary libraries and the dataset.

In [1]:
```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
%matplotlib inline

df = pd.read_csv('./carats.csv', header= None)
df.columns = ['carats', 'price']
df.head()
```

```
/opt/conda/lib/python3.6/site-packages/statsmodels/compat/pandas.py:56:
FutureWarning: The pandas.core.datetools module is deprecated and will
be removed in a future version. Please use the pandas.tseries module in
stead.
  from pandas.core import datetools
```

Out[1]:

	carats	price
0	0.17	355
1	0.16	328
2	0.17	350
3	0.18	325
4	0.25	642

1. Similar to the last notebook, fit a simple linear regression model to predict price based on the weight of a diamond. Use your results to answer the first question below. Don't forget to add an intercept.

In [3]:
```
df['intercept'] = 1

lm = sm.OLS(df['price'], df[['intercept', 'carats']])
results = lm.fit()
results.summary()
```

Out[3]:

OLS Regression Results

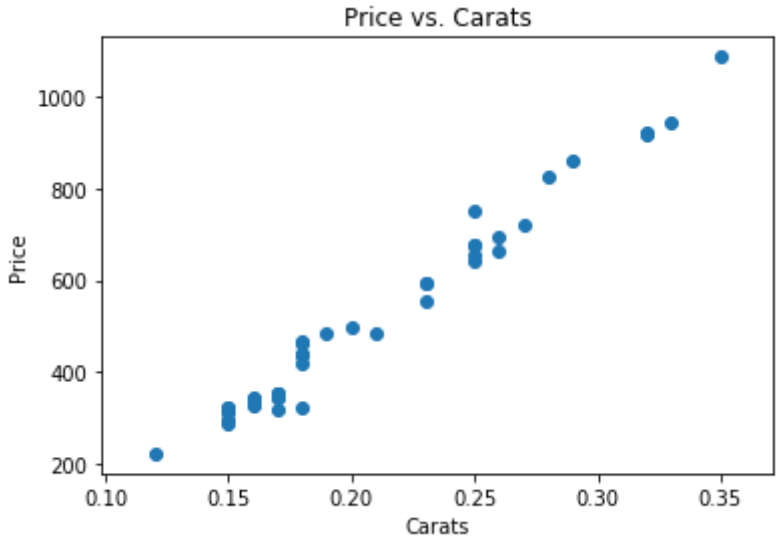
Dep. Variable:	price	**R-squared:**	0.978
Model:	OLS	**Adj. R-squared:**	0.978
Method:	Least Squares	**F-statistic:**	2070.
Date:	Tue, 14 Apr 2020	**Prob (F-statistic):**	6.75e-40
Time:	02:39:16	**Log-Likelihood:**	-233.20
No. Observations:	48	**AIC:**	470.4
Df Residuals:	46	**BIC:**	474.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	-259.6259	17.319	-14.991	0.000	-294.487	-224.765
carats	3721.0249	81.786	45.497	0.000	3556.398	3885.651

Omnibus:	0.739	**Durbin-Watson:**	1.994
Prob(Omnibus):	0.691	**Jarque-Bera (JB):**	0.181
Skew:	0.056	**Prob(JB):**	0.913
Kurtosis:	3.280	**Cond. No.**	18.5

2. Use scatter (https://matplotlib.org/gallery/lines_bars_and_markers/scatter_symbol.html?highlight=scatter%20symbol) to create a scatterplot of the relationship between price and weight. Then use the scatterplot and the output from your regression model to answer the second quiz question below.

In [4]:
```
plt.scatter(df['carats'], df['price']);
plt.xlabel('Carats');
plt.ylabel('Price');
plt.title('Price vs. Carats');
```



In [ ]: