Import the required libraries we need for the lab.

In [2]:
```
import piplite
await piplite.install(['numpy'],['pandas'])
await piplite.install(['seaborn'])
```

In [3]:
```
import pandas as pd import pandas as pd
import    seaborn    as    sns    import
matplotlib.pyplot   as   pyplot   import
scipy.stats
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

```
<ipython-input-3-b3fdaf15785b>:1: DeprecationWarning:
Pyarrow will become a required dependency of pandas in the next major releas
e of pandas (pandas 3.0),
(to allow more performant data types, such as the Arrow string type, and bet
ter interoperability with other libraries)
but was not found to be installed on your system.
If this would cause problems for you,
please provide us feedback at https://github.com/pandas-dev/pandas/issues/54
466

  import pandas as pd
```

Read the dataset in the csv file from the URL

In [4]:
```
from js import fetch
import io

URL = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/
resp = await fetch(URL)
boston_url = io.BytesIO((await resp.arrayBuffer()).to_py())
```

In [5]:
```
boston_df=pd.read_csv(boston_url)
```

Add your code below following the instructions given in the course to complete the peer graded assignment

In [6]:
```
boston_df.head()
```

Out[6]:

	Unnamed: 0	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX
0	0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0

1	1	0.0273	0.	7.0	0.	0.46	6.42	78.	4.967	2.0	242.
2	2	1	0	7	0	9	1	9	1	2.0	0
3	3	0.0272	0.	7.0	0.	0.46	7.18	61.	4.967	3.0	242.
4	4	9	0	7	0	9	5	1	1	3.0	0
		0.0323	0.	2.1	0.	0.45	6.99	45.	6.062		222.
		7	0	8	0	8	8	8	2		0
		0.0690	0.	2.1	0.	0.45	7.14	54.	6.062		222.
		5	0	8	0	8	7	2	2		0

In [34]:
```
boston_df.describe()
```

Out[34]:

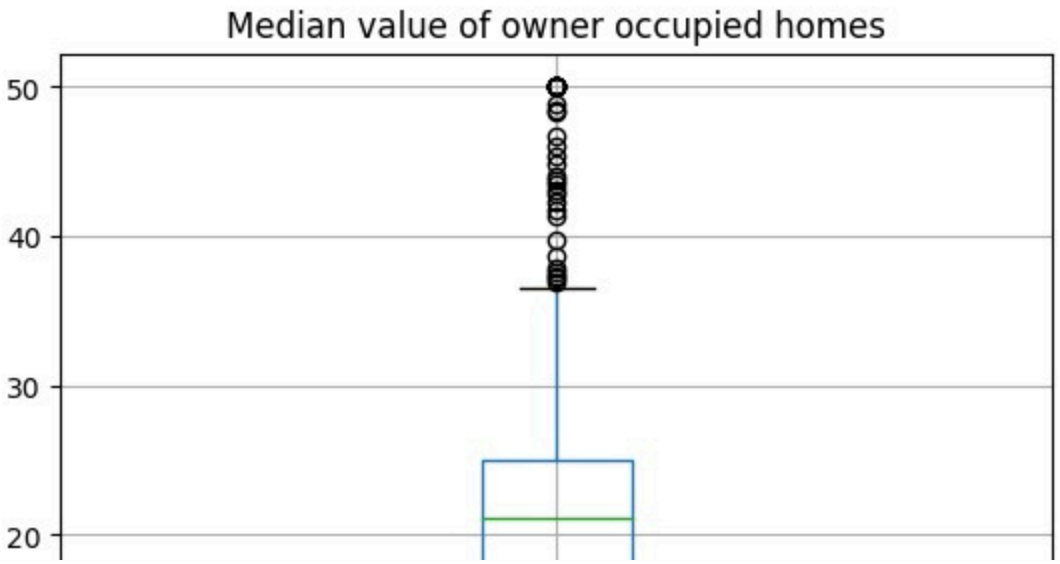
	Unnamed: 0	CRIM	ZN	INDUS	CHAS	NOX	
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	50
mean	252.500000	3.613524	11.363636	11.136779	0.069170	0.554695	
std	146.213884	8.601545	23.322453	6.860353	0.253994	0.115878	
min	0.000000	0.006320	0.000000	0.460000	0.000000	0.385000	
25%	126.250000	0.082045	0.000000	5.190000	0.000000	0.449000	
50%	252.500000	0.256510	0.000000	9.690000	0.000000	0.538000	
75%	378.750000	3.677083	12.500000	18.100000	0.000000	0.624000	
max	505.000000	88.976200	100.000000	27.740000	1.000000	0.871000	

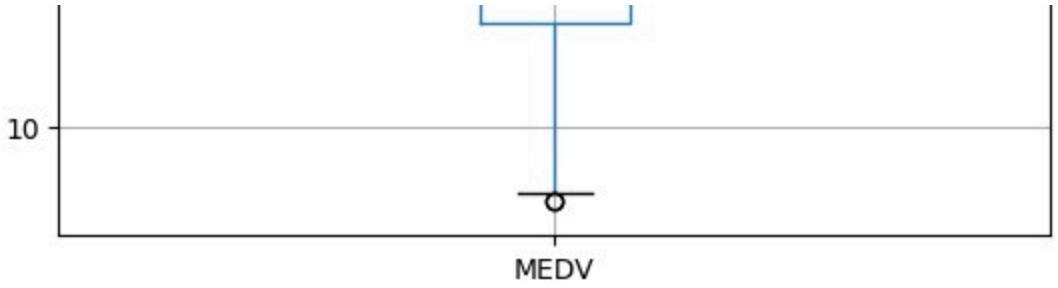
Task 2: Visualizations

For the "Median value of owner-occupied homes" provide a boxplot

In [24]:
```
boston_df.boxplot(column='MEDV')
pyplot.title('Median value of owner occupied homes')
```

Out[24]:
```
Text(0.5, 1.0, 'Median value of owner occupied homes')
```



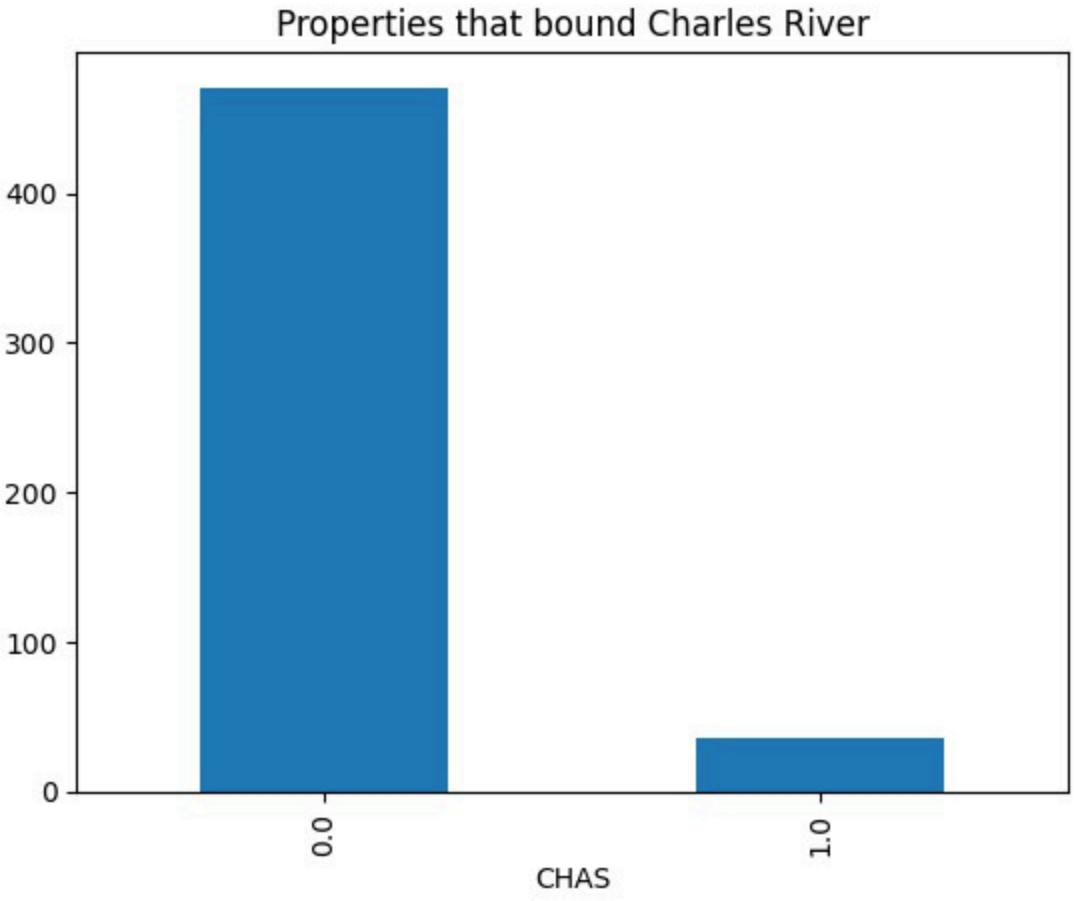


Explanation: We see that the median value is a little more than 20, but there are many outliers above the 75 quantile, meaning the population is right skewed.

Provide a bar plot for the Charles river variable

In [30]:
```
boston_df['CHAS'].value_counts().plot(kind='bar')
pyplot.title('Properties that bound Charles River')
```

Out[30]:
```
Text(0.5, 1.0, 'Properties that bound Charles River')
```



Explanation: The majority of the houses does not bound the river, as we would expect.

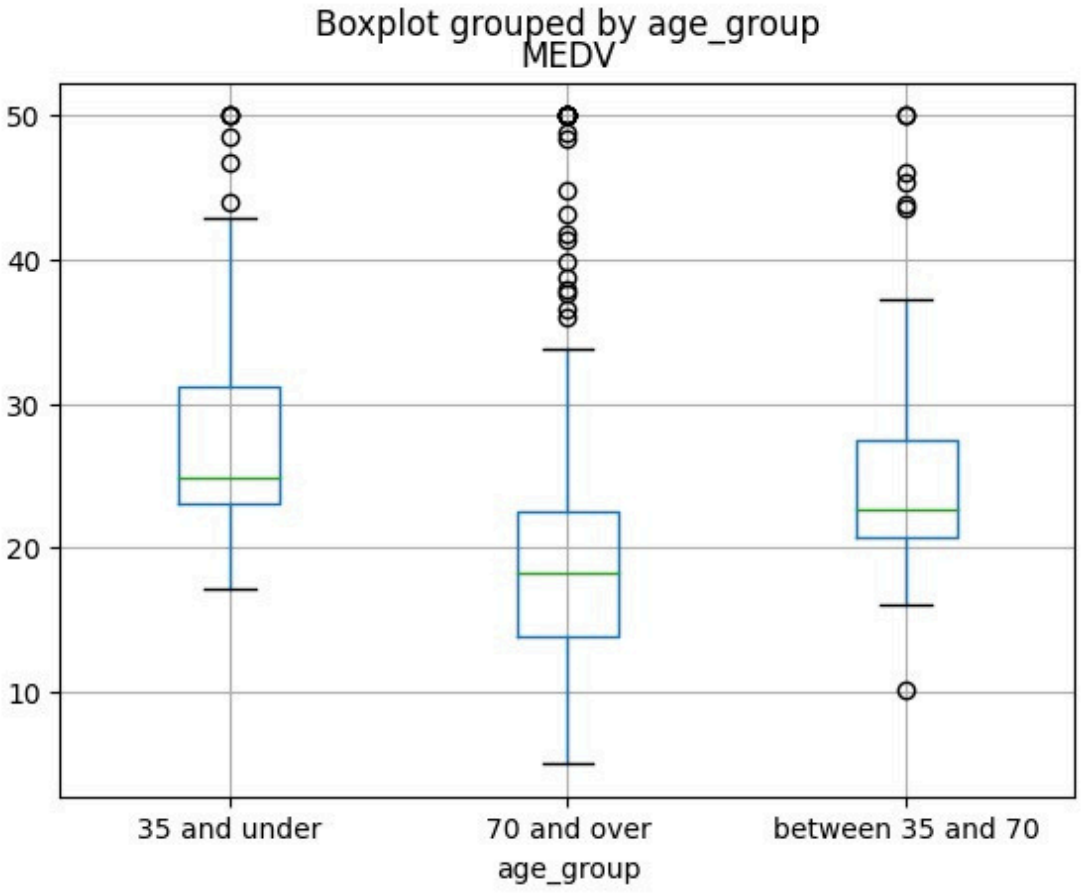
Provide a boxplot for the MEDV variable vs the AGE variable. (Discretize the age variable into three groups of 35 years and younger, between 35 and 70 years and 70 years and

older)

In [36]:
```
boston_df.loc[boston_df['AGE'] <= 35, 'age_group'] = '35 and under'
boston_df.loc[(boston_df['AGE'] > 35) & (boston_df['AGE'] < 70), 'age_grou
boston_df.loc[boston_df['AGE'] >= 70, 'age_group'] = '70 and over'

boston_df.boxplot(column='MEDV', by='age_group')
```

Out[36]: <AxesSubplot:title={'center':'MEDV'}, xlabel='age_group'>

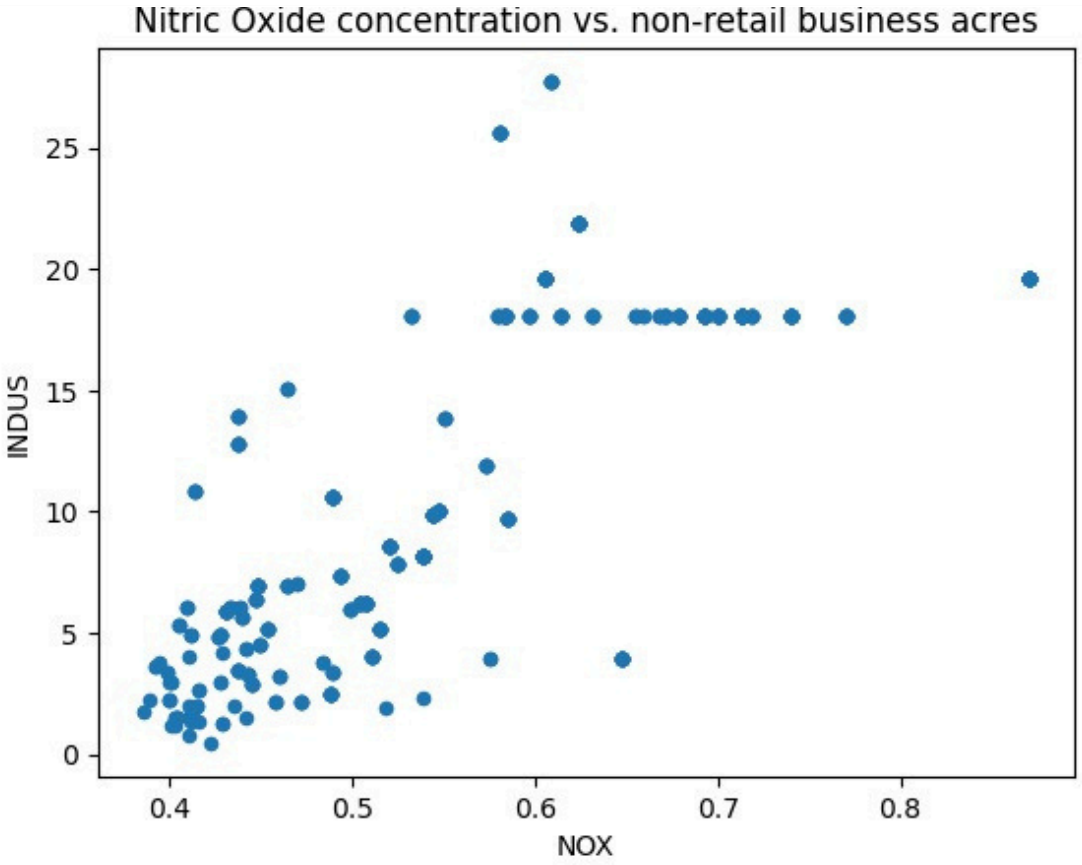


Explanation: The towns with the highest population of old houses (70% and over) have the lowest median house value, and the towns with the lowest population of old houses have the highest median house value. But we can see that all three groups have outliers with a higher value, so the maximum house price is similar for all three groups.

Provide a scatter plot to show the relationship between Nitric oxide concentrations and the proportion of non-retail business acres per town. What can you say about the relationship?

In [43]:
```
boston_df.plot.scatter(x='NOX', y='INDUS')
pyplot.title('Nitric Oxide concentration vs. non-retail business acres' )
```

Out[43]: Text(0.5, 1.0, 'Nitric Oxide concentration vs. non-retail business acres')

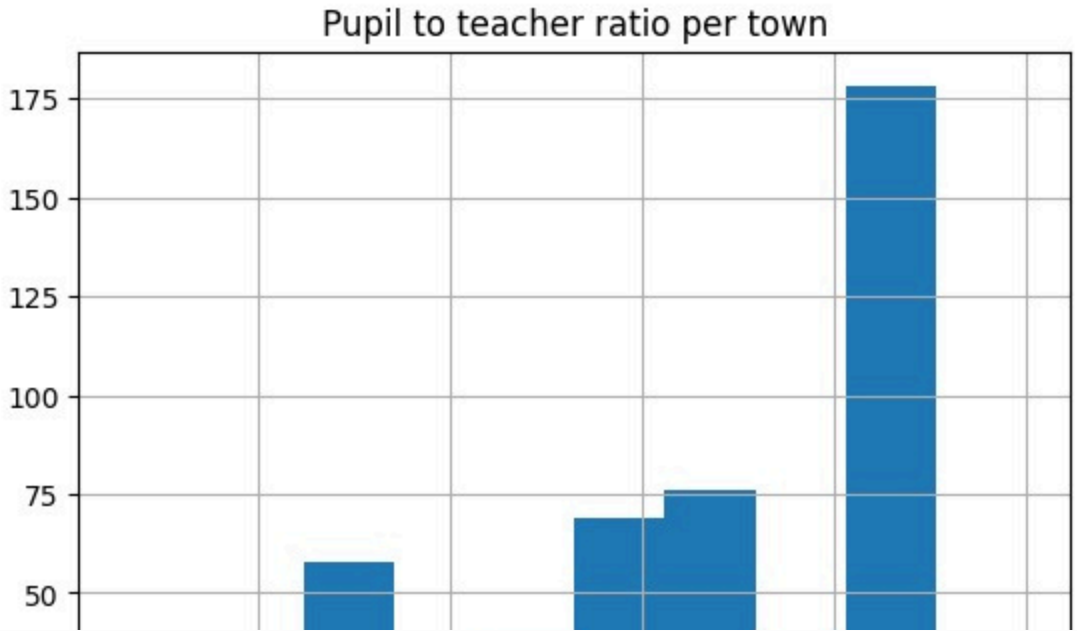


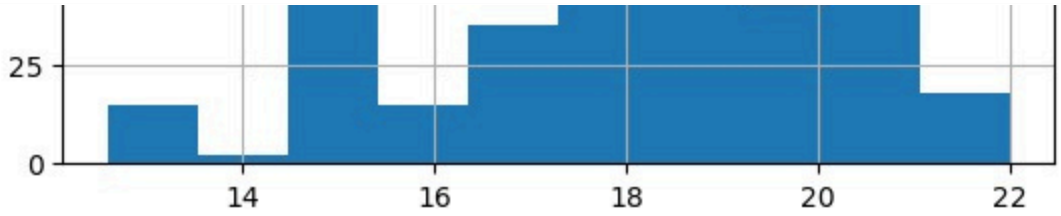
Explanation: There seems to be a correlation between the NOX and INDUS variable, where when NOX increases, INDUS also increases (this will be confirmed later using the Pearson test).

Create a histogram for the pupil to teacher ratio variable

In [48]:
```
boston_df['PTRATIO'].hist()
pyplot.title('Pupil to teacher ratio per town')
```

Out[48]:
```
Text(0.5, 1.0, 'Pupil to teacher ratio per town')
```





Explanation: There are many towns with a pupil to teacher ratio of 21 (mode), the lowest pupil to teacher ratio is 13 and the highest is 22.

Task 3: Use the appropriate tests to answer the questions provided.

Is there a significant difference in median value of houses bounded by the Charles river or not? (T-test for independent samples)

State the hypothesis

- Null hypothesis: $H0$:μ1=μ2 ("there is no difference in median value of houses bounded by the Charles river and those that are not")
- Alternative hypothesis: $H1$:$\mu 1 \neq \mu 2$ ("the median value of the houses bouded by the river and not differ")

Starting with the Levene test to check for equality of variance:

In [51]:
```
scipy.stats.levene(boston_df[boston_df['CHAS'] == 1.0]['MEDV'],
                   boston_df[boston_df['CHAS'] == 0.0]['MEDV'], center='me
```

Out[51]:
```
LeveneResult(statistic=8.75190489604598, pvalue=0.003238119367639829)
```

The p-value is smaller than 0.05, thus we cannot assume equal variances.

In [54]:
```
scipy.stats.ttest_ind(boston_df[boston_df['CHAS'] == 1.0]['MEDV'],
                   boston_df[boston_df['CHAS'] == 0.0]['MEDV'], equal_var
```

Out[54]:
```
TtestResult(statistic=3.113291312794837, pvalue=0.003567170098137517, df=3
6.876408797611994)
```

Conclusion: The p-value is very small (smaller than our α=0.05) so we reject the null hypothesis and conclude that there is statistical evidence that the medan value of these two groups is not the same.

Is there a difference in Median values of houses (MEDV) for each proportion of owner occupied units built prior to 1940 (AGE)? (ANOVA)

State the hypothesis

- Null hypothesis: $H0$:$\mu1=\mu2=\mu3$ ("there is no difference in median value of houses for each proportion of owner-occupied units built prior to 1940")
- Alternative hypothesis: $H1$: At least one of the means differ

In [61]:
```
boston_df.loc[boston_df['AGE'] <= 35, 'age_group'] = '35 and under'
boston_df.loc[(boston_df['AGE'] > 35) & (boston_df['AGE'] < 70), 'age_grou
boston_df.loc[boston_df['AGE'] >= 70, 'age_group'] = '70 and over'
```

Starting with the Levene test to check for equality of variance:

In [63]:
```
scipy.stats.levene(boston_df[boston_df['age_group'] == '35 and under']['MED
                   boston_df[boston_df['age_group'] == 'between 35 and 70'
                   boston_df[boston_df['age_group'] == '70 and over']['MEDV
                   center='mean')
```

Out[63]:
```
LeveneResult(statistic=2.7806200293748304, pvalue=0.06295337343259205)
```

The p-value is greater than 0.05, thus we can assume equal variances.

In [64]:
```
thirtyfive_lower = boston_df[boston_df['age_group'] == '35 and under']['MED
thirtyfive_seventy = boston_df[boston_df['age_group'] == 'between 35 and 70
seventy_more = boston_df[boston_df['age_group'] == '70 and over']['MEDV']
```

In [65]:
```
f_statistic, p_value = scipy.stats.f_oneway(thirtyfive_lower, thirtyfive_se
print("F_Statistic: {0}, P-Value: {1}".format(f_statistic,p_value))
```

```
F_Statistic: 36.40764999196599, P-Value: 1.7105011022702984e-15
```

Conclusion: Since the p-value is less than our α=0.05, we will reject the null hypothesis as there is significant evidence that at least one of the means differ.

Can we conclude that there is no relationship between Nitric oxide concentrations and proportion of non-retail business acres per town? (Pearson Correlation)

State the hypothesis:

- $H0$: Nitric oxide concentrations is not correlated with proportion of non-retail business acres per town
- $H1$: Nitric oxide concentrations is correlated with proportion of non-retail business acres per town

In [57]:
```
scipy.stats.pearsonr(boston_df['NOX'], boston_df['INDUS'])
```

Out[57]:
```
PearsonRResult(statistic=0.7636514469209192, pvalue=7.913361061210442e-98)
```

Conclusion: Since the p-value < 0.05, we reject the Null hypothesis and conclude that there exists a relationship between nitric oxide concentration and proportion of non-retail business acres per town.

What is the impact of an additional weighted distance to the five Boston employment centres on the median value of owner occupied homes? (Regression analysis)

State the hypothesis

- Null hypothesis: $\beta 1=0$ (distance has no effect on median value)
- Alternative hypothesis: $\beta 1 \neq 0$ (distance has an effect on median value)

In [59]:

```
## X is the input variables (or independent variables)
X = boston_df['DIS']
## y is the target/dependent variable
y = boston_df['MEDV']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```