Data Exploration -
Statistical Analysis

```
Minimum Value  CRIM 0.00632
Maximum Value  CRIM 88.9762
Calculate mean  CRIM 3.59376071146
Calculate median  CRIM 0.25651
Calculate standard deviation  CRIM 8.58828354765
Minimum Value  ZN 0.0
Maximum Value  ZN 100.0
Calculate mean  ZN 11.3636363636
Calculate median  ZN 0.0
Calculate standard deviation  ZN 23.2993956948
Minimum Value  INDUS 0.46
Maximum Value  INDUS 27.74
Calculate mean  INDUS 11.1367784561
Calculate median  INDUS 9.69
Calculate standard deviation  INDUS 6.85357058339
Minimum Value  CHAS 0.0
Maximum Value  CHAS 1.0
Calculate mean  CHAS 0.0691699604743
Calculate median  CHAS 0.0
Calculate standard deviation  CHAS 0.25374293496
Minimum Value  NOX 0.385
Maximum Value  NOX 0.871
Calculate mean  NOX 0.554695059289
Calculate median  NOX 0.538
Calculate standard deviation  NOX 0.115763115407
Minimum Value  RM 3.561
Maximum Value  RM 8.78
Calculate mean  RM 6.28463438735
Calculate median  RM 6.2085
Calculate standard deviation  RM 0.701922514335
Minimum Value  AGE 2.9
Maximum Value  AGE 100.0
Calculate mean  AGE 68.5749011858
Calculate median  AGE 77.5
Calculate standard deviation  AGE 28.1210325702
Minimum Value  DIS 1.1296
Maximum Value  DIS 12.1265
Calculate mean  DIS 3.79504269775
Calculate median  DIS 3.20745
Calculate standard deviation  DIS 2.10362835634
Minimum Value  RAD 1.0
Maximum Value  RAD 24.0
Calculate mean  RAD 9.54940711462
Calculate median  RAD 5.0
Calculate standard deviation  RAD 8.698651113779
Minimum Value  TAX 187.0
Maximum Value  TAX 711.0
Calculate mean  TAX 408.23715415
Calculate median  TAX 330.0
Calculate standard deviation  TAX 168.3704495039
Minimum Value  PTRATIO 12.6
Maximum Value  PTRATIO 22.0
Calculate mean  PTRATIO 18.4555335968
Calculate median  PTRATIO 19.05
Calculate standard deviation  PTRATIO 2.164280519148
Minimum Value  B 0.32
Maximum Value  B 396.9
Calculate mean  B 356.674032621
Calculate median  B 391.44
Calculate standard deviation  B 91.2046074522
Minimum Value  LSTAT 1.73
Maximum Value  LSTAT 37.97
Calculate mean  LSTAT 12.6530632411
Calculate median  LSTAT 11.36
Calculate standard deviation  LSTAT 7.13400163665
```

Evaluating Model Performance

Performance Metric:
To evaluate performance of model I am using R2 score. I am using this metric instead of mean squared error because R2 score is rescaling of MSE (relative to the dataset) and it ranges from 0-1.

Testing/Training Split:
I have split data set into **Testing/Training** to evaluate the model on the data it has not seen. I am holding out 40% for testing and retaining 60% for training.

I am using cross_validation train_test_split to separate the data set.

Cross Validation & Gridsearch

I use along side grid search to find an optimal parameter.

Analyzing Model Performance

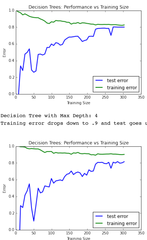
Learning Curves and Bias & Variance Analysis

After fitting a model with 60% of training data and testing with 40% data of data I tested `Decision Tree with Max Depth from 1 to 10.`

Learning Curves and Bias & Variance Analysis

```
First graph shows that  error of both the test and training sample is around .5 indicating that this depth of 1 is not predictive.
Decision Tree with Max Depth: 1
```

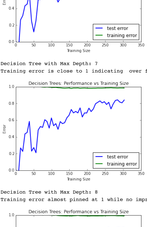

```
Decision Tree with Max Depth: 2
Does better than depth 1 but still is not optimal
```




```
Decision Tree with Max Depth:3
We can see that the accuracy drops off and levels off after about 100 training instances
```


```
Decision Tree with Max Depth: 4
Training error drops down to .9 and test goes up to .8. Still not perfect
```


```
Decision Tree with Max Depth: 5
Training error stays close to 1, and then it drops off indicating that this model approximates data well.
```

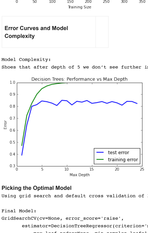



```
Decision Tree with Max Depth: 6
Training error stays close to 1 for 1/3 of set while test R2 decreases indicating that we are starting to over fit
```


```
Decision Tree with Max Depth: 7
Training error is close to 1 indicating  over fitting. No improvement in test error
```


```
Decision Tree with Max Depth: 8
Training error almost pinned at 1 while no improvement in test error. Overfitting
```

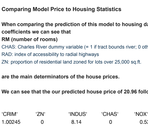

```
Decision Tree with Max Depth: 9
 Training error almost pinned at 1 while no improvement in test error. Overfitting
```




```
Decision Tree with Max Depth: 10
 Training error  pinned at 1 while no improvement in test error. Overfitting
```

Error Curves and Model Complexity

```
Model Complexity:
Shows that after depth of 5 we don't see further improvement with addition of complexity (depth)
```



Picking the Optimal Model
Using grid search and default cross validation of 3 we examine the affects of parameters on the model while re-fitting with each irritation.

```
Final Model:
GridSearchCV(cv=None, error_score='raise',
       estimator=DecisionTreeRegressor(criterion='mse', max_depth=None, max_features=None,
           max_leaf_nodes=None, min_samples_leaf=1, min_samples_split=2,
           min_weight_fraction_leaf=0.0, random_state=None,
           splitter='best'),
       fit_params={}, iid=True, loss_func=None, n_jobs=1,
       param_grid={'max_depth': (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)},
       pre_dispatch='2*n_jobs', refit=True, score_func=None,
       scoring=make_scorer(r2_score), verbose=0)

We find best  parameter max_depth = 5

Best regressor params: {'max_depth': 5}
```

Predicted Housing Price

```
House: [11.95, 0.0, 18.1, 0, 0.659, 5.609, 90.0, 1.385, 24, 680.0, 20.2, 332.09, 12.13]
Predictions: [ 20.96776316]
```

Comparing Model Price to Housing Statistics

When comparing the prediction of this model to housing data we can see that the prediction is inline with training data set. Looking at the training data and after obtaining coefficients we can see that
RM (number of rooms)
CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
RAD: index of accessibility to radial highways
ZN: proportion of residential land zoned for lots over 25,000 sq.ft.

are the main determinators of the house prices.

We can see that the our predicted house price of 20.96 follows a similar hose

'CRIM'	'ZN'	'INDUS'	'CHAS'	'NOX'	'RM'	'AGE'	'DIS'	'RAD'	'TAX'	'PTRATIO'	'B'	'LSTAT'	Prediction
1.00245	0	8.14	0	0.538	6.674	87.3	4.239	4	307	21	380.23	11.98	21