Project 1: Predict the Boston House Price
Name: Chinonso Maduakolam

Questions and Report Structure
1) Statistical Analysis and Data Exploration

Number of data points (houses)?
506

Number of features?
13

Minimum and maximum housing prices?
Min:5.0 Max:50.0

Mean and median Boston housing prices?
Mean:22.53 Median:21.2

Standard deviation?
9.18

2) Evaluating Model Performance

Which measure of model performance is best to use for predicting Boston housing data and analyzing the errors?
MSE

Why do you think this measurement most appropriate?
This measure is more appropriate because we are concerned about large errors whose consequence are much bigger than equivalent smaller ones. Hence, it penalizes the errors that are farther away from the mean. It also represents our errors in the same same unit as our data which allows us to individually square the error each observation and take the square root of the mean

Why might the other measurements not be appropriate here?
MAE is not an appropriate measurement because it less sensitive to the occasional very large errors in the calculation. And in a real world scenario, we want to minimize very large errors compared to smaller errors.

Why is it important to split the Boston housing data into training and testing data?
It is important because we want to assess the quality of a model on out-of-sample date where we don't know the response values. We want to see how close our predictions are to the real outcome and finally we want to see how large our margins of error when are predictions are incorrect.

What happens if you do not do this?
If we don't do this, then we can not qualitatively asses the questions above.

What does grid search do and why might you want to use it?
Grid search is a more computationally efficient method of searching a parameter space (grid of parameters) to find the best fit for a given model.

We might want to use it to find the parameter where a model best performances

Why is cross validation useful and why might we use it with grid search?
Cross validation is useful because it provides a more accurate estimate on out-of-sample data accuracy.

It is a more efficient in the use of data because every observation is used for both training and testing.

This might be used in conjunction with grid search because you have a set of model which differ from each other based on their parameter values which lie on a grid. After each has been subsequently trained and evaluated using cross validation, you can select the one that performed the best. It is essentially an automated way to train and test your model.

3) Analyzing Model Performance

Look at all learning curve graphs provided. What is the general trend of training and testing error as training size increases?
The training and test error graphs trend in opposite directions. The test error graph is trending downwards while the test error graph is trending upwards as the training size increases.

Look at the learning curves for the decision tree regressor with max depth 1 and 10 (first and last learning curve graphs). When the model is fully trained does it suffer from either high bias/under-fitting or high variance/over-fitting?

Looking at graph 1, it is obvious that the model suffers from under-fitting/high-bias and this is due to the fact that the model performs poorly on both the train and test data. Even though we see a drop in the number of test errors, they are still very high. Where as the training errors on the other hand keep rising at the training size increases. It is obvious that the model being too simple is the cause of its inability to pick out trends in the data.

Looking at graph 10, the model suffers from high variance/over-fitting. Looking at the nature of the test and train graph, the train graph assumes a linear progression as the model becomes more complex while the test graph show a lot of peaks and drops. Also the error for the training set is at and sometimes near 0 which shows that the model has learned the noise in the training data and as a result it does a very good job of classifying the training observation but it fails to generalize. Hence,it performs poorly on the test data. When we look at the test error graph, see a precipitous drop in the number of errors but as the training size increases, we notice the recurring trend of the errors beginning to rise and fall.

Look at the model complexity graph. How do the training and test error relate to increasing model complexity? Based on this relationship, which model (max depth) best generalizes the dataset and why?

After running the model multiple times, a max depth of 5 showed to be the point where the model best generalizes because it is the point in the graph where the training and test error are at their lowest, when we go past 5, the test errors begin to rise while the train error continues to approach 0.

4) Model Prediction

Model makes predicted housing price with detailed model parameters (max depth) reported using grid search. Note due to the small randomization of the code it is recommended to run the program several times to identify the most common/reasonable price/model complexity.

After running at program at least 50 times, the mmost common price was 23.34980 and most common model complexity is a {'max_depth':2}

Compare prediction to earlier statistics and make a case if you think it is a valid model.
The model looks valid, when we compare the model prediction with summary statistcal data from the explor_citydata() we see a mean price of 22.532 and a median price of 21.2. So a predicted price of 23.349 doesn't appear unrealisticly low or high.

Sources
How to compare models(Duke university)

Retrieved from:http://people.duke.edu/~rnau/compare.htm

Why Data Scientists Split Data into Train and Test (Dan Steinberg,March,3,2014)
Retreieved From: http://info.salford-systems.com/blog/bid/337783/Why-Data-Scientists-Split-Data-into-Train-and-Test