1) Statistical Analysis and Data Exploration

- Number of data points (houses): **506**
- Number of features? **13**
- Minimum and maximum housing prices? **5 - 50**
- Mean and median Boston housing prices? **22.532 – 21.2**
- Standard deviation? **9.188**

2) Evaluating Model Performance

- **Which measure of model performance is best to use for predicting Boston housing data and analyzing the errors? Why do you think this measurement most appropriate? Why might the other measurements not be appropriate here?**

  The prediction of Boston Housing data is a regression problem. The metrics that help in evaluating the performance of a model are

  a) Mean Squared Error
  b) Mean Absolute Error and
  c) R2 Score
  d) Explained Variance Score

a and b provide a measure of how close forecasts or predictions are to the eventual outcomes, c and d give a measure of the proportion to which a model accounts for the variation of the data set.

I have chosen Mean Squared Error as the model performance metric because it takes into account the variance and the bias of the estimator although MSE punishes outliers heavily due to squaring. Mean Absolute Error also could be used.

R2 Score and Explained Variance Score account for the variance of the data more than the model parameters. *R* Score will never decrease as variables are added and will probably experience an increase due to chance alone. Making the model complex due to more variables but a low R2 score.

- **Why is it important to split the Boston housing data into training and testing data? What happens if you do not do this?**

  We separate training and testing sets so we can get a better idea whether the model can generalize to unseen data rather than fit to the data just seen. If we do not do this, then we will over fit the model to all training data and thereby fail on new data.

- **What does grid search do and why might you want to use it?**

  Grid search exhaustively generates candidates from a grid of parameter values that helps in training the model with variations of the model parameters and finally helping in deciding the best model parameters to be used.

- **Why is cross validation useful and why might we use it with grid search?**

  Cross Validation helps us to split the data into training and testing sets while also maximizing the data available to train the model without overfitting the model. Cross validation helps in fine tuning the model by holding out some data for testing and then with randomizing the training data by running several folds of train – test splits.

  Cross validation along with grid search can help in fine tuning the model performance.

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?**

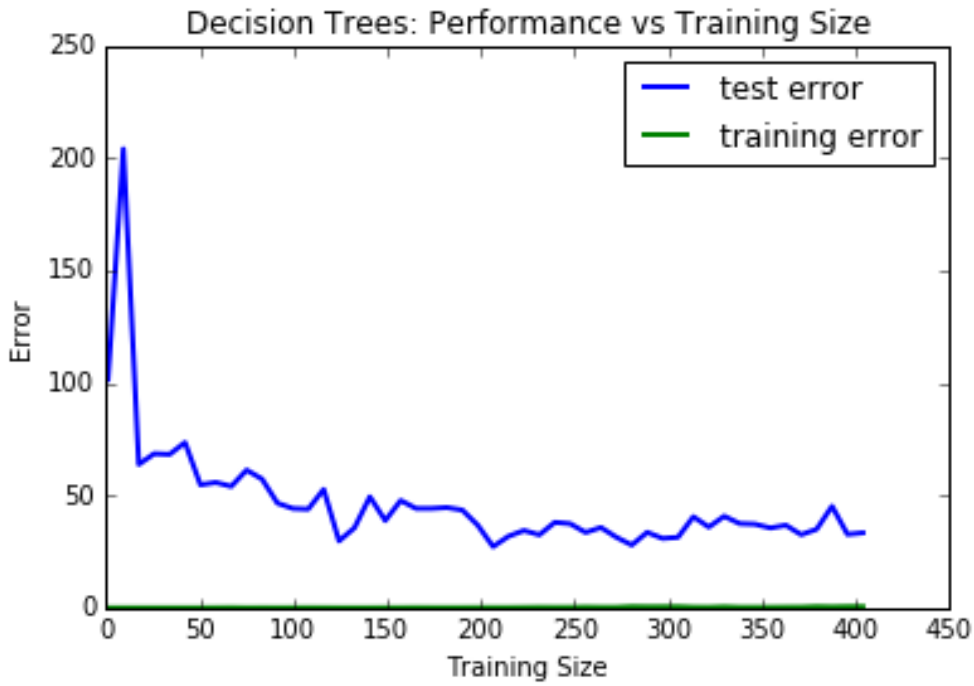
As training size increases the training and testing errors plateaus. As the sizes increase there is hardly any difference in the change of values.

- **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/underfitting or high variance/overfitting?**

  The Learning Curve with depth 1.

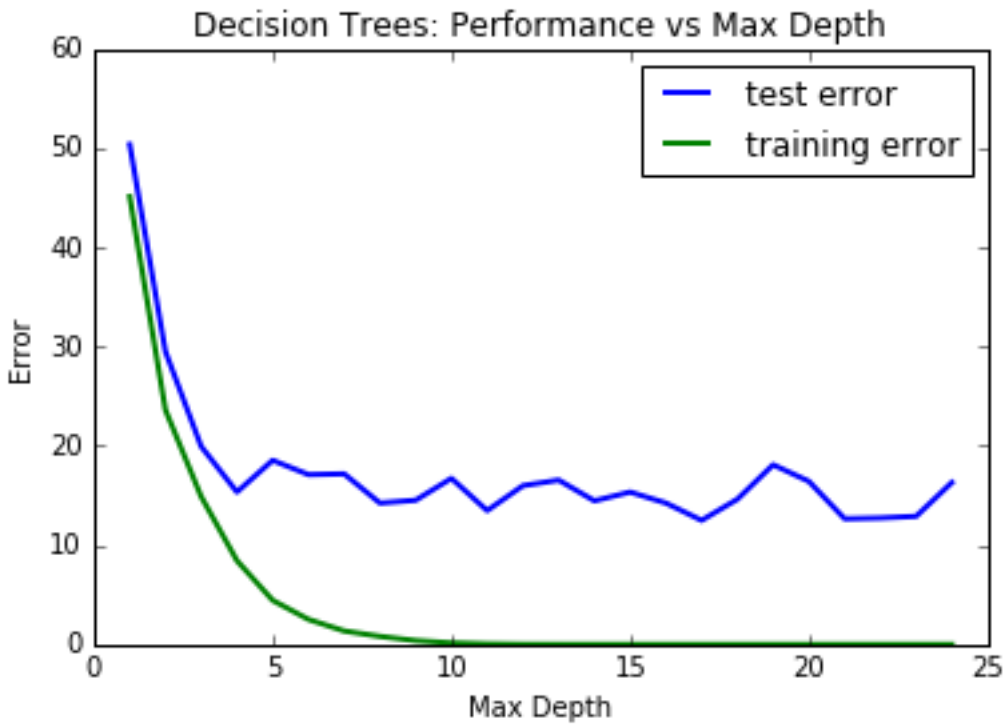
We see that the training and testing errors converge and the errors are quite high. We can infer that the model is biased. Even if we train this model further it will fil to predict the underlying relationship and therefore will fail.

Learning Curve with Depth 10



We have a situation where the training error is zero and we have a large gap between the training and testing error. We can infer that the model suffers from high variance.

- **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?**



We can see that as the Max depth increases the test error and training error plateaus. Which means that more features do not increase the accuracy of this model.
The ultimate goal for a model is one that has low errors and generalizes pretty well for unseen data (testing data). We can see this when both the testing and training curves converge and where the error is extremely low. From this graph a **Depth of 4** has the best performance.

4) Model Prediction

- **Compare prediction to earlier statistics and make a case if you think it is a valid model**.

After several runs the model makes a Prediction which is close to the Median of the data set. The Best Estimator also gives the model parameter of max depth = 4.

Last run Prediction: [21.62974359].