1. **Statistical Analysis and Data Exploration**
   - **Number of data points(houses)**
     - 506
   - **Number of features**
     - 13
   - **Minimum and maximum housing prices**
     - The minimum housing price is 5.0.
     - The maximum housing price is 50.0.
   - **Mean and median Boston housing prices**
     - The mean Boston housing price is 22.5328063241.
     - The median Boston housing price is 21.2.
   - **Standard deviation**
     - The standard deviation of Boston housing price is 9.18801154528.

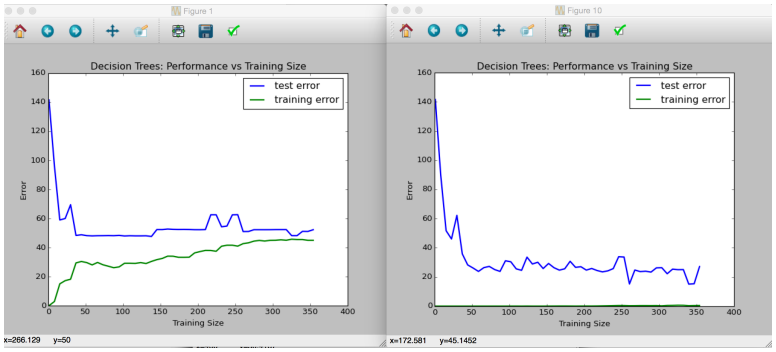
2. **Evaluation Model Performance**
   - **Which measure of model performance is best to use for predicting Boston housing data and analyzing the errors? Why do you think this measurement most appropriate? Why might the other measurements not be appropriate here?**
     - MSE.
     - MSE calculates the squared error, it tends to put more weight on large errors or outliers. MAE weights all errors equally hence it is less sensitive to outliers, but if there are many outliers or the outliers are extreme then it will affect the accuracy of the model. To prevent the influence of large errors, MSE is more appropriate.

   - **Why is it important to split the Boston housing data into training and testing data? What happens if you do not do this**?
     - Splitting data into two sets prevents overfitting. If we do not do splitting, the trained model would fit tightly with the whole data set. It is less accurate when we test the trained model with new data.

   - **What does grid search do and why might you want to use it?**
     - Grid search exhaustively searches through all possible values of a specified estimate and pick the optimal one.
     - We use grid search because it is simple and searches through all possible values.

   - **Why is cross validation useful and why might we use it with grid search?**
     - During k experiments, each data point is assigned to test set and training set, and each data point is validated once. It is useful because it prevents overfitting and generates a more accurate estimate of a model.
     - Using cross validation with grid search ensures an optimal result.

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?**



- Training error increases as training size increases.
- Testing error decreases as 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/ underfitting or high variance/overfitting?**
  - When max depth is 1, both testing error and training error are high. The model suffers from high bias/underfitting
  - When max depth is 10, the training error remains at 0 and the testing error decreases and converges gradually. Zero training error suggesting the model perfectly fits the training data, hence it is less fit to the test data. Also the gap between training error and testing error remains large. Therefore the model suffers from high variance/overfitting.

- **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?**

  - The training error decreases and finally converges at max depth = 11 as the model complexity increases. The test error drops and reaches the lowest point at max depth = 5, and then slightly increases.
  - The test error is the lowest at max depth = 5. After this point, the increasing error suggests the model is overfitting. Hence the max depth = 5 best generalizes the dataset.

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.**
  - The most common/reasonable price is 21.62974359 and max depth = 4.

- **Compare prediction to earlier statistics and make a case if you think it is a valid model.**
  - The price is closed to the mean price 22.5328063241 and the median price 21.2. It is a valid model.

[('[21.62974359]', 24), ('[20.76598639]', 15), ('[19.99746835]', 5), ('[18.81666667]', 2), ('[20.96776316]', 2), ('[19.32727273]', 1), ('[20.72]', 1)]
average: 21.004142323