Questions and Report Structure

1) Statistical Analysis and Data Exploration

- Number of data points (houses)?
- there are 6578 total data points

- Number of features?
- there are 13 features.

- Minimum and maximum housing prices?
- the minimum is $5000 and the maximum is $50,000

- Mean and median Boston housing prices?
- the mean of the prices is $22,532.8 and the median is $21,200

- Standard deviation?
- the standard deviation is 9.188 or $9188.01 from the mean

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?
- I have chosen mean absolute error (MAE) in this case because housing prices are continuous values so the problem is a regression problem. I have counted 40 outliers so MAE is best to model these data without be swayed by the outliers, and at the same time giving equal emphasis on all the data points. Mean squared error is an option but it's not as optimal because it places heavy emphasis on large errors, so the errors from outliers will influence the model much more making it not as accurate. Media absolute error is another possibility because it measures the errors from the median of errors instead of the average, but in this case the outliers account for 8% of the data, so I still want to take them into consideration, but only as much weight as normal data. In this problem if we plot the errors on a bar graph we should see high frequencies for large errors and low frequencies for small errors, making it a left skewed graph. The model should then learn from those errors, and the best model for that in this case is MAE giving equal weight to large and small errors.

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

- It's important to use the test data to see if the model really works after fitting to the training data. If we do not do this then the model could be overfitting to the training data and unable to predict new data.

- What does grid search do and why might you want to use it?
- Grid search looks through all the parameters and finds the one with best performance. We want to use it so we can go through parameter combinations and optimize our models automatically.

- Why is cross validation useful and why might we use it with grid search?
- Cross validation is useful because it maximizes the usage of all given data by utilizing them for both training and testing data and protect against overfitting. We use it with grid search so we can use all the data to find the best parameter combinations. I originally picked 100 fold CV arbitrarily to see the benefit versus computation time, and it turns out it takes too much time so I reduced the folds by half to 50. This way it still allows the algorithm to go over the dataset 50 times over in a relatively short period of time. It maybe more optimal to have an even lower CV like 10 fold and still be sufficient but I'm leaving it at 50 in this case to ensure greater consistency over multiple runs. I have tried setting CV to the sample size of 506 and found no change in the results.

3) Analyzing Model Performance

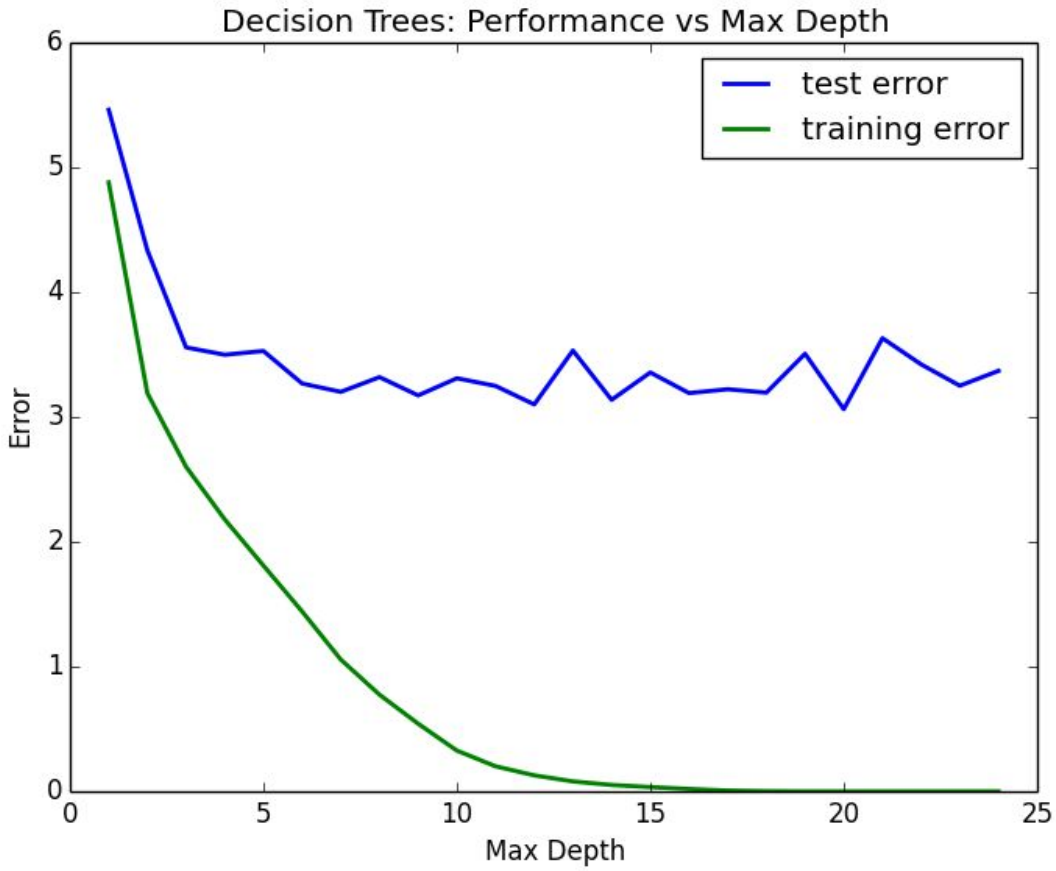
- Look at all learning curve graphs provided. What is the general trend of training and testing error as training size increases?
- As training size increases the training error increases then stays approximately the same, and testing error decreases and stays approximately the same.

- 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?
- Max depth 1:

- At max depth of 1 the training and testing error converge and they hover around 5 and the curves are relatively flat. This means as the data size increases the error count stays relatively constant meaning the model is not able to learn from new data, and the model is biased towards existing parameter, a form of underfitting.
- Max depth 10:

- At max depth of 10 the model shows a gap between test error and training error with the test error hovering just below 4 on a downward trend. This means the model has high variance and it's overfitting. The errors jump up and down with new data so the model is trying too hard.

- 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?



- As the model gotten more complex the gap between test error and training error increased, meaning a persistent high variance and overfitting issue. Looking at the graph a max depth of 4 would be best because it has the lowest test error with the least amount of gap between the two error curves.

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.
- I ran grid search 30 times and found the average prediction to be $20,824.72 and the most common max depth is at 4.

- Compare prediction to earlier statistics and make a case if you think it is a valid model.
- Without comparing each of the 13 features of this house individually it's difficult to say if this house is an outlier to this dataset. Nevertheless the house's predicted value is $1708.08 from the mean and $375.28 from the median, within 1 standard deviation of $9188.01 of the mean.