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? 5 and 50
- Mean and median Boston housing prices?22.53 and 21.2
- Standard deviation?9.19

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?**

This is a regression problem. The best test is root mean squared error(RMSE), I feel. This metric gives us an intuition about the euclidean distance between the vector of true scores and the vector of predicted scores, averaged by square root of the number of data points.

Explained variance score - would measure the proportion of to which the model accounts for the variation in a data set which would not be so useful here.

Mean and median absolute errors do not generalize very well unless they are averaged

Coefficient of determination(R^2) would mainly be used when one would want to fit a line or curve on the data.

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

  Splitting the data into training and testing is needed to assess model generalizability on unseen data. Not doing this will result in overfitting in which the model would perform very well on the training data but will have a lot of errors on the training data.

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

  Grid search does an exhaustive optimization of parameters using manually specified p parameter space using cross-validation on the training set. This will help us achieve highest score for whatever error metric we choose.

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

  Cross-validation is useful because just relying on one set of testing and training might be a little biased. Cross-validation overcomes this by randomly partitioning the data into random pairs of training sets and the results are averaged over them. The squared error may be averaged over the different sets of testing data..

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, training error increases and testing error decreases

- **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?**

For the ‘max depth 1’ one notes the high error value that the two curves converge into This can be because of overfitting

For the ‘max depth 10’, one notes the low error for the training curve and a small distance between the training and testing curve. This could be because of underfitting.

- **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 model complexity increases, training and testing errors decrease although the testing error seem to plateaus off after 5. The training error has almost reached zero at higher depths. Max depth of 5 best generalizes the dataset because the test error plateaus off after that. If we base on training errors, there may be a chance of overfitting.

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

  I think it is a valid model because the predicted price is not far from the dataset mean(within one standard deviation of the mean of the dataset).