Name: Kunwoo Park (bywords.kor@gmail.com)

1) Statistical Analysis and Data Exploration

- Number of data points (houses)? 506
- Number of features? 13
- Minimum and maximum housing prices? 5.0 / 50.0
- Mean and median Boston housing prices? 22.533 / 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?

  : I think mean squared error is the most appropriate metric for boston housing data. Predicting boston housing prices is a regression problem, which aims to predict continuous values, so we need to measure differences between actual prices and predicted prices. Metrics such as accuracy and precision are not proper here, because those metrics are calculated based on distinct number of labels. Thus, they are more appropriate for classification problems. Mean absolute error can be another good candidate here.
- Why is it important to split the Boston housing data into training and testing data? What happens if you do not do this?

  : It is crucially important to split data to test performances of models after training them. If we test a model's performance on the same data from which we trained the model, the model would have high variances and low generalizability because the model will be trained in a way to minimize errors on training data. In other words, it would suffer from the overfitting.
- What does grid search do and why might you want to use it?

  : Grid search trains models for given parameters and searches for the best model with regard to a metric. Without grid search, we cannot know which parameters are the best and should manually test performances by training models separately. Therefore, I want to use it.

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

  : Cross validation helps us to find the best set of parameters having generalizability. Without cross validation, grid search returns parameters to fit best for training data. Thus, we should use cross validation for grid search.

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 and testing error generally decrease as training size increases. When depth of trees is shallow, both training and testing error seems to decrease monotonically. However, when depth of trees is quite deep (e.g., 10), test errors fluctuate when the training size is big.

- 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 the max depth is low, it suffers from high bias. The model does not fully explain the data. On the contrary, when the max depth becomes high, it suffers from high variance / overfitting. Testing errors does not decrease continuously as max depth increases.

- 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?

  : By increasing model complexity (increasing max depth), training errors continuously decrease. However, testing errors stop to decrease when max depth is around 4. This trend indicates that if max depth is higher than 4, models may have high variances. For that reason, decision trees of max depth 4 best generalize 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. Compare prediction to earlier statistics and make a case if you think it is a valid model.
  : By running reg.best_params_, I found it is highly likely to return 4 as the best maximum depth to minimize minimum squared errors. This result is same as the inferences from the above statistics on bias and variance.