**1) Statistical Analysis and Data Exploration**

* Number of data points (houses)? 506
* Number of features? 13
* Minimum and maximum housing prices? Minimum Price: 5.0 Maximum Price: 50.0
* Mean and median Boston housing prices? Mean Price: 22.5228 Median Price: 21.2
* Standard deviation? 9.18801

**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?

Find the best model performance metric for regressions and understand its downsides.

* 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 sample is vital to train and evaluate the performance of our models. Training samples are used to train the Machine Learning algorithms and test samples are used to evaluate the performance of the model on ‘new’ data points. With the test sample we are able to compare the predicted label vs the real label so we can check how did the model perform on data points that we not used in the learning process. This process is called cross validation.

The main goal of cross validation to evaluate the predicted power of a model and to avoid over fitting to the training sample. We may be able to get high accuracy on training samples but that doesn’t imply high accuracy on new unknown data points. In this sense, splitting the data and using a particular sample to allows us to evaluate the performance of the model on new data points. If cross validation is not used, training errors might lead us to wrong conclusions.

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

Each Machine Learning algorithm has its own set of parameters that helps us tune their learning process. For example, in decision trees we can set up the maximum depth, the minimum size for splits or the minimum size for a leaf. In order to find the optimal set of parameters we use grid search. This process consists in training the model several times with different parameters and check with which configuration we achieved the best performance.

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

As stated before, cross validation is useful to evaluate the performance of a particular model with new data points (not used in the learning process) and to avoid overfitting. Both objectives should be considered by grid search in order to determinate which parameter configuration is the optimal.

**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?
* 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?
* 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?

**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.