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?
    - **22.5328 21.2**
- Standard deviation?
    - **9.18801154528**

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?
    - mean squared error
    - MSE is good at penalizing high errors which can provide stabler prediction, and no outlier exist in dataset, so MSE is the best metric for this case.
    - Definitely I won't choose other categories metric but regression metric, those measurements are not suitable for regression case. R_2 score is not a error metric, it outputs values between 0-1, sometime can be negative, which can't use in compute the total error. Mean absolute error outputs the turly error between predications and actual values, but it provides error that smaller than mean squared error, which i think as not good enough for model training like previous answer. And the median squared error, it is robust to outlier, but my dataset do not have any outlier, so this metric may be not suitable.

- Why is it important to split the Boston housing data into training and testing data? What happens if you do not do this?
    - We need a test set because we need to know whether our learning algorithm can generalize to new data or not. A learning model that ‘remember’ train data too deeply or just ‘forget’ almost all train data a.k.a. overfit and underfit are useless. By scoring the train error and test error, and plot them into a curve graph, we can know if our model suffer from high bias or high variance. First of all, we need a test set.
    - Just like previous block, a model without test will be esay to suffer from underfit or overfit. Model like that can’t generalize to new data, that means useless.

- What does grid search do and why might you want to use it?
    - Grid search will exhaustive search over specified parameter values for an estimator. [According to the page in sklearn]
    - Grid search is one of the methods for parameters auto-tuning.If we need to fit a new dataset, grid serach will get the job done automatically, that can save our power(do not need to write the code again).

- Why is cross validation useful and why might we use it with grid search?
    - Cross validation is a method that ues in compare models with difference, it will randomly generate training and test set multiple times(intend to avoid the possible bias), and map different training/test splits to different models, and compare different pairs of training and test error. Without cross validation, it is impossible to tell which model is better or worse than others though observe models and theirs parameters only.
    - What grid serach do is parameters auto-tuning, that’s the reason why we use cross validation with it(because same models with different parameters are different to each other).Also their combination can reduce the chance of overfiting and maximize data usage.

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 slowly as training size increases.
    - Test error gets less rapidly at first, and then decrease very slow generally 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?
    - The decision tree regressor with max depth 1 was suffer from high bias/underfitting, because training error curve is close to test error curve and both have a great error.
    - The decision tree regressor with max depth 10 was suffer from high variance/overfitting, because there is a gap between training error curve and test error curve, and the training error is almost 0.

- 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?
    - When the model complexity increase, the training error is getting less and close to 0, the test error is getting less but it won't continue go down after some point, the test error will stay high.
    - I think model with max depth 4 are the best two generalizes of the dataset. The reason is they both get lower training error and test error than others, due to the

model with max depth smaller than 3 is suffer form underfit and the model with max depth higher than 4 is suffer form overfit.

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 price is about 21.63
    - The most reasonable model complexity reported from grid search is desision tree regressor with max depth 4.

- Compare prediction to earlier statistics and make a case if you think it is a valid model.
    - Prediction in code:
        - features:[11.95,0,18.1,0,0.659,5.609,90.0,1.385,24,680.0,20.2,332.09, 12.13]
        - prediction: 21.63
        - prediction is 0.098 standard deviations below from mean, and there is 0.0468 standard deviations above from median.
        - After use sklearn find out the nearest neighbours(n=10), and compute the mean price of them, the value is 21.52
        - prediction is 0.012 standard deviations above from the average of nearest neighbours.
        - so the prediction should be reasonable.

    - Try to predict some new case:
        - features: [35,55,40,0,0.9,5,99,1.9,25,500,10,430,18]
        - prediction: 15.48

- for the new prediction, it is 0.7676 standard deviations below from mean, and 0.6226 standard deviation below from median.
- the value of the nearest neighbours(n=10) is 17.36
- prediction is 0.2046 standard deviations below from the average of nearest neighbours.
- so the prediction should be reasonable.