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? **Min : 5, Max: 50**
- Mean and median Boston housing prices? **Mean : 22.53, Median : 21.2**
- Standard deviation? **Std: 9.18**

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 used both mean square error and mean absolute error. Both predicted the same housing rent. Since there are not outliers, min max range is not huge, therefore I guess both the evaluation metrics provided a fair result and any one can be used here.**
- Why is it important to split the Boston housing data into training and testing data? What happens if you do not do this?
  - **In order to verify the model it is important to split the data into training and testing data. The model needs to be evaluated with the testing data as well as choose the best possible parameters for the model.**
- What does grid search do and why might you want to use it?
  - **Gridserarch helps in finding the best parameter for the model that we use. If we do not use it, its a very recursive and tiring way of finding the best parameters.**

Gridsearch helps to avoid the manual brute force technique of finding the best of the parameters for a model

- Why is cross validation useful and why might we use it with grid search?
  - **Cross valdiation is important inorder to make our model more robust and predict better. In grid search we can specify the number of folds the data should be divided. A particular type of data can be present in a particular fold. So K-fold technique can help us predict the output better**

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 the training size increased, the error rate for the test data gradually reduced. So we can say that as the model is fed more data, it learned more and predicted better and hence test error declined**
- 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 model is fully trained, it suffers from 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?
  - **Increasing model complexity, reduces the training error but more of stagnates the test error. Model with depth 6 best generalizes the dataset as at this point the test error shows off a linear change or becomes stagnant.**

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.
    - **Yes, it gives me 2 different results, 20.76 and 21.62**
- Compare prediction to earlier statistics and make a case if you think it is a valid model.
    - **What is this earlier statistics? please advise**