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?
    - **First of all training data is partitioned into training and testing data using random partitioner. Randomizing helps in somewhat preventing similar kind of data being present into one part of the dataset(i.e training or testing). The testing data helps in validating the output of the model or choosing the best possible with least RMSE or MSE. If we do not divide the data then there can**

be chances that the model may not predict properly about the data which it has not seen i.e generalization from the current data (overfitting)

- What does grid search do and why might you want to use it?
    - **Gridsearch 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?
    - **In order to judge that you have modelled the input data correctly you need to have some data with you which can verify that whatever model or classifier you have built is predicted correctly. In order to do so, we divide data into test and training parts and feed the training part to build the model and test the model with the test part. As mentioned above, CV can prevent overfitting.**
    - **Grid search is a technique of finding the most optimal parameters for a model. When I say optimal, it means with least RMSE or MAE. Now for that the grid search needs to set aside some data to test for that. In grid search, K-fold is used which can further cross validate the model better. The number of folds can be provided by the user. I am taking the default value as 3. Depending on the number of input data we can further increase or decrease the folds**

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.**

    - **As per the training error, it increases gradually which actually makes sense, the more data is fed into the model, the model will try to accommodate the training data more, which can generally leads to gradual increase in error.**
- 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). As the model complexity increased, the training error came down to almost zero but the testing error gradually became linear. This means the model is exhibiting high variance.**
    - **When the max depth is 1, with increase in training data, both the training and test error seems to become stagnant. This seems to be scenario of high bias where the model is not complex enough to properly fit the training data.**
    - **When max depth is 10, high variance exists. See my comments above.**
- 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. Depth is : 7 or 8**

- Compare prediction to earlier statistics and make a case if you think it is a valid model.
    - **To verify the prediction, I calculated 10 nearest points to the input and then find out its mean and std. Here is the mean and std for that**
    - **mean : 21.52 and std : 10.30. The predicted value lies within this range.**