1) Statistical Analysis and Data Exploration

- Number of data points (houses)?

  ```
  Size of data: 506
  ```

- Number of features?

  ```
  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: 22.5328063241
  ```

  ```
  Median Price: 21.2
  ```

- Standard deviation?

  ```
  Standard deviation: 9.18801154528
  ```

2) Evaluating Model Performance

- Which measure of model performance is best to use for predicting Boston housing data and analysing the errors? Why do you think this measurement most appropriate? Why might the other measurements not be appropriate here?

  Mean Squared Error is the metric use to measure model performance. Benefits of using this metric are:-

  1. It automatically converts all the errors to Positive.
  2. From calculus it allows us to find minimum and maximum values.

- Why is it important to split the Boston housing data into training and testing data? What happens if you do not do this?

  It's important to split training and testing data because

  1. It gives us an estimate of performance on an independent subset which is never used on the model before.
  2. It serves as a check on the problem of overfitting.

  If we don't split the data into training and testing then:

  1. We are unable to measure the performance of the model on a data that is never used on it before.
  2. Our model can be highly sensitive for training set i.e. Overfitting. It means error in testing data can be much more than that of training data.

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

  Grid Search is used to systematically tune the model with different combination of parameters when they are not directly learnt by estimators.

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

  Cross validation is an approach in which training set is split into K smaller sets, and the following procedure is followed for each of the k-folds.
  1. A model is trained for k-1 subset
  2. The trained model is tested for the remaining set.

  The performance measured reported by the k-folds cross validation Is then averaged.

  Cross validation is used with Grid search because:-
  1. It reduces the chance of overfitting. If we limit gridsearch to single training set, we may accidentally overfit our model if the training set is somehow imbalanced. Using cross validation the parameters will be optimized on the entire data set and any random anomalies due to random splitting will be removed.
  2. Maximizing data usage: When a dataset is limited in size cross validation becomes extremely useful as it allows for an extensive exploitation of available data allowing assessing the real potential of our algorithm in terms of performance metrics.

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 training size increases error in testing data decreases and error in training set 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?

  At Depth 1:- As Training and Testing Error converges and are quite high at max Depth 1 it means model suffers from **high bias /underfitting**. No matter how much data we feed it, the model cannot represent underlying relationship.

  At Depth 10:- As gap between testing and training data is high and training error is quite low. The model suffers from **high variance/ overfitting**. It can be fixed by providing more data.

- 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?

  With increase in model complexity training data keeps on decreasing as it should but after a certain depth testing error Plateaus or slightly increases.

  I think in my plots max depth of 4 best generalizes the dataset because after this depth only training error decreases while testing error remains almost same.

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.

  I think most common/reasonable price is 21.12 at model complexity of 5

- Compare prediction to earlier statistics and make a case if you think it is a valid model.

  I took mean and standard deviation of 10 nearest neighbours of the input using kneighbor's method which I got 21.52 and 10.31. The predicted values lies within the range hence I think it's a valid model