1.) Statistics Analysis and Data Exploration
Number of data points: 506
Number of features: 13
Minimum and Maximum house prices: 5.0, 50.0
Mean and Median house prices: 22.532806324110698, 21.199999999999999
Standard Deviation: 9.1880115452782061

2.) Evaluating Model Performance
The dataset is made of real / continuos numbers – Which suggests that it is a regression problem. Two common metrics are used to measure performance in a regression model. Mean Squared Error and Mean Absolute Error.
The M**ean Squared Error** is an appropriate measure for measuring the performance of the model selected

1. It makes larger differences more evident
2. It converts all errors to positives

Other rubrics for measuring model performance are
Mean Absolute Error, Recall, Precision, Accuracy and F-Score

It is important to split the Boston housing data in order to get an idea of how well the model is able to predict data it has not been trained on.

Grid Search is a technique optimize parameters for an estimator. The goal of grid search is to find the best parameters that minimize errors across all the data specified.

Cross Validation is a technique used to test how well a model performs on a given set of data. It involves splitting a dataset in to n folds then iteratively using each of the folds as test data.

We can combine Grid Search with Cross Validation to

1. Reduce the chance of overfitting. We may accidentally overfit our model if we have an imbalanced testing set. Using cross validation will
2. Maximize data usage. When a dataset is limited in size, cross validation becomes useful as it allows you explore the available dataset allowing you to access how well the selected algorithm performs using the selected performance metric.

3.) Analyzing Model Performance
As training size increases, training error increases while testing error decreases.

At Max Depth 1: The model shows high bias – It does not explore many features to properly generalize the model. In a case where the feature selected does not show high correlation with the expected result, the model will return a large error values.

At Max Depth 10: The model shows suffers from high variance / overfitting. This is when the model considers all features and tries to make use of the them. This leads to the model performing well on training data and poorly on test data. The model at max depth 10 does not do well in predicting the test data

The model with Max Depth = 4 best generalizes the dataset, Because at that time, the testing error is at a point where adding complexity to the model does not necessarily reduce errors. It is at this point that and training error balances themselves well.

4.) Model Prediction
The most reasonable price reported by grid search is 21.629
The model also reported the best model parameters to be at max depth 4.

According to the earlier statistic, this appears to be a reasonable model since the predicted value falls within the standard deviation. Also, the nearest neighbors to the house we are predicting (using the nearest neighbor algorithm at k=10) is 21.52. That looks like we had a pretty good prediction.