1. Statistical Analysis and Data Exploration

   - There are **506 data points**.
   - There are **13 features**.
   - The **minimum** housing price is **5.0** and the **maximum** price is **50.0**.
   - The **mean** Boston housing price is **22.53** and the **median** Boston housing price is **21.2**.
   - The **standard deviation** of the data is **9.188**.

2. Evaluating Model Performance

   - Mean squared error is the best measure of model performance to use for regression and predicting Boston housing data because the result will always give a positive value and emphasize larger differences. This is most appropriate for regression because the goal of regression is to minimize error about the mean. To minimize error about the mean, we could use the mean absolute error as our metric; however, this measure does not penalize as harshly for larger deviations about the mean. Furthermore, we could use the coefficient of determination,r^2, which would detail how well the model fits. In general, r^2 is related to the MSE (seen by the adjusted r^2 below).

     $$r^2_{adj} = 1 - \frac{MSE}{\sigma_y^2}$$

   - It is important to split the data into training and testing data to help us assess if there is overfitting. Furthermore, it provides you known sample data to use to verify that the model is accurate. This provides a means for cross validation, which is an iterative process where you split data into train and test, train your data on the training set, test data on a portion of the testing set, go back and re-train, test on another portion of the testing set, etc. Lastly, in the `DecisionTreeRegressor`, the only supported `criterion` is `"mse"` for mean squared error.
   - The most appropriate cross validation technique in our case would likely be K-fold, which consists of splitting the data into k groups of train and test subsets. This will allow us to assess if there is overfitting and help us mitigate overfitting, as we will be able to test on known, pre-labeled data.
   - Grid search is a type of hyperparameter optimization, where given a set of parameters relevant to the learning algorithm, this technique will run exhaustively over the set of all values for each parameter. The goal of grid search is to optimize each parameter with respect to the scoring method (in our case, mean squared error). We might want to use this to further decrease overall error. Furthermore, another technique for hyperparameter optimization is ‘RandomizedSearchCV‘, which does not exhaustively search the parameter space, thereby saving time. Instead, ‘RandomizedSearchCV‘ samples from the parameter space following specified distributions. In our case, the data is relatively small, and therefore saving time may not be necessary.

3. Analyzing Model Performance

   - After observing the learning curve graphs provided, training error tends to increase at a seemingly logarithmic rate as training size increases. Test error tends to decrease and “smooth” (i.e., decrease in variance) as the training size increases, until overfitting begins to occur, which is when test error starts to increase in magnitude and variance.
   - When the model is fully trained, with a max depth of 1, the model appears to suffer from underfitting (high bias). With a max depth of 10, the model appears to suffer from overfitting (high variance). It’s important to note that Decision Trees in general tend to suffer from overfitting.
   - From the model complexity graph, the training error decreases at an exponential rate as the model grows more complex (the max depth increases). The test error decreases similarly as the max depth increases, until overfitting begins to occur, which causes the test error to vary at a much greater magnitude. Based on this relationship, the model (max depth) that best generalizes the dataset appears to be a max depth of 6. At a max depth of 6, both the training and test errors appear to stabilize to their seemingly asymptotic optimal values. Therefore, we avoid further overfitting, while ensuring accuracy.

4. Model Prediction

   - The model's predicted housing price is 19.934. The best model parameter from the GridSearch was 6.
   - The model's prediction, 19.934, is very close to both the mean and the median of the Boston housing price data. Furthermore, the prediction falls well within the standard deviation of the data.