# Boston Housing Modeling

Release 2016.02.21

Russell Nakamura

March 01, 2016

## Contents

# 1 Statistical Analysis and Data Exploration

## 1.1 The Data

The data was taken from the sklearn.load_boston function, which itself cites the UCI Machine Learning Repository as their source for the data. The data gives values for various features of different suburbs of Boston as well as the median-value for homes in the suburbs. The features were chosen to reflect various aspects believed to influence the price of houses including the structure of the house (age and spaciousness), the quality of the neighborhood, transportation access to employment centers and highways, and pollution.

Here is the description of the data variables provided by sklearn.

Table 1: Attribute Information (in order)

| Variable Name | Description |
|---|---|
| CRIM | per capita crime rate by town |
| ZN | proportion of residential land zoned for lots over 25,000 sq.ft. |
| INDUS | proportion of non-retail business acres per town |
| CHAS | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) |
| NOX | nitric oxides concentration (parts per 10 million) |
| RM | average number of rooms per dwelling |
| AGE | proportion of owner-occupied units built prior to 1940 |
| DIS | weighted distances to five Boston employment centers |
| RAD | index of accessibility to radial highways |
| TAX | full-value property-tax rate per $10,000 |
| PTRATIO | pupil-teacher ratio by town |
| B | 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town |
| LSTAT | % lower status of the population |
| MEDV | Median value of owner-occupied homes in $1000's |

**Note:** The data comes from the 1970 U.S. Census so its values don't (necessarily) reflect current values.

## 1.2 Cleaning the Data

Since there are no missing data points, there isn't much to do to clean the data, but the odd variable names increase the likelihood of error so I'm going to expand them to full variable names.

Table 2: Variable Aliases

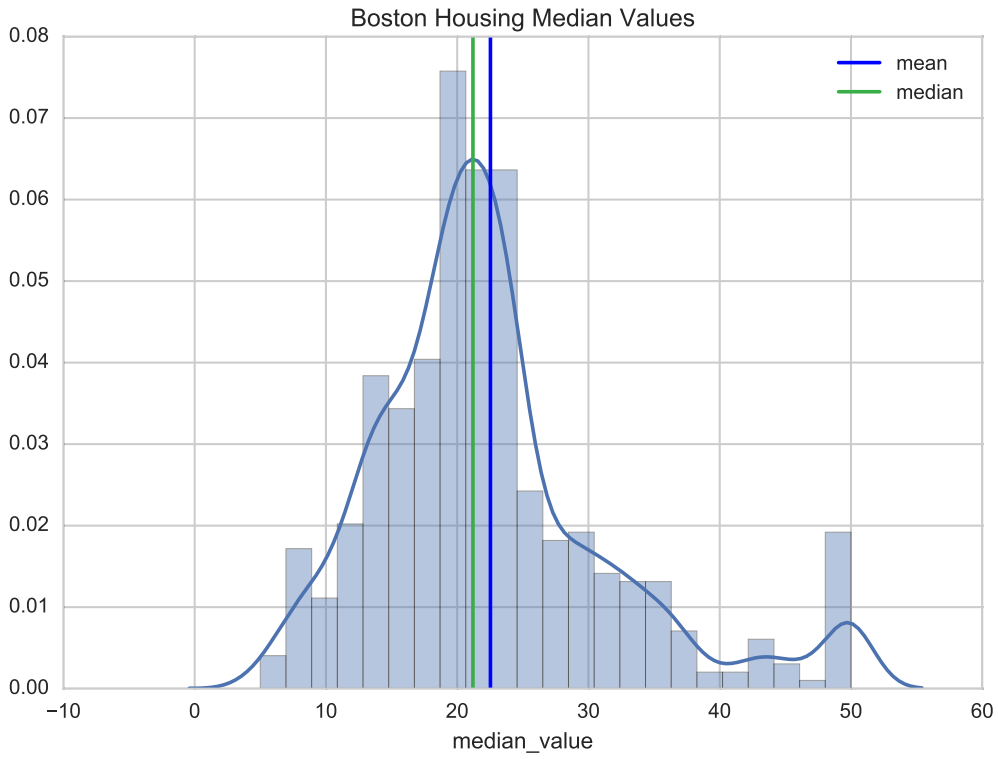
| Original Variable | New Variable |
|---|---|
| CRIM | crime_rate |
| ZN | large_lots |
| INDUS | industrial |
| CHAS | charles_river |
| NOX | nitric_oxide |
| RM | rooms |
| AGE | old_houses |
| DIS | distances |
| RAD | highway_access |
| TAX | property_taxes |
| PTRATIO | pupil_teacher_ratio |
| B | proportion_blacks |
| LSTAT | lower_status |

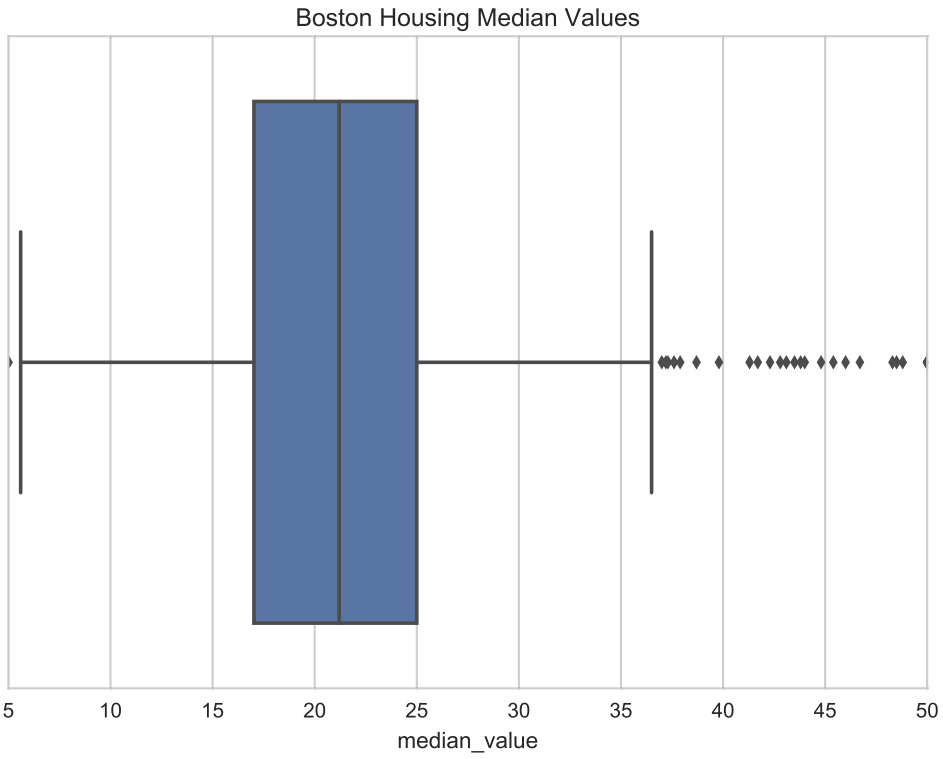
## 1.3 Summary Statistics

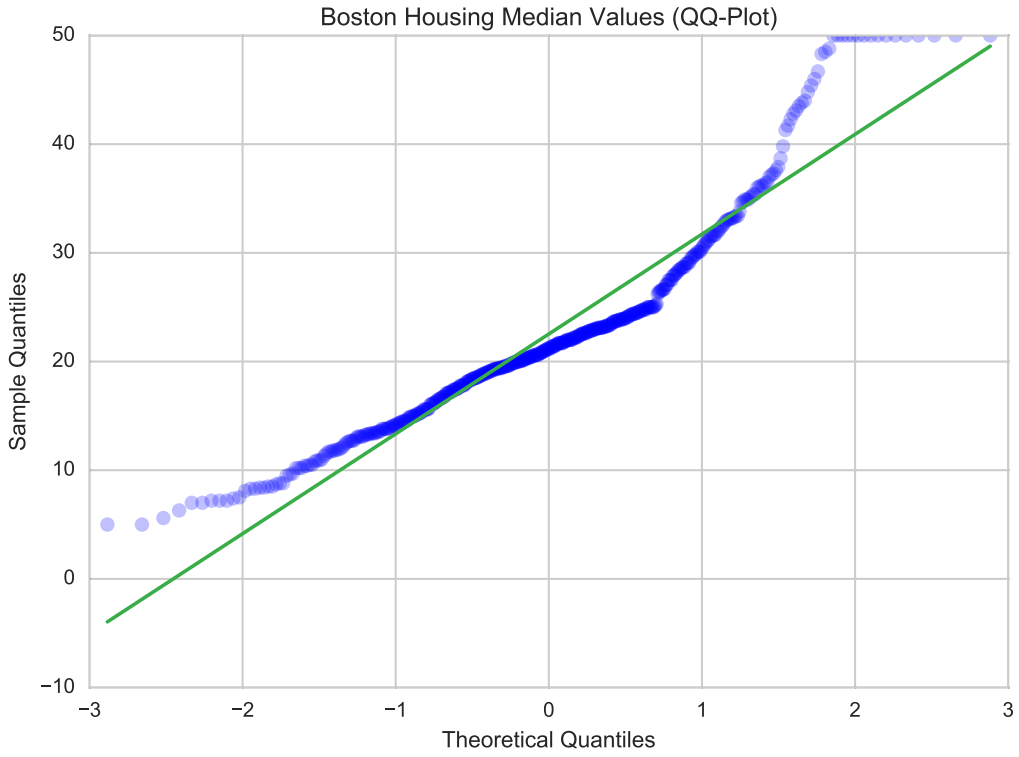
Table 3: Boston Housing data-set statistics (in $1000's)

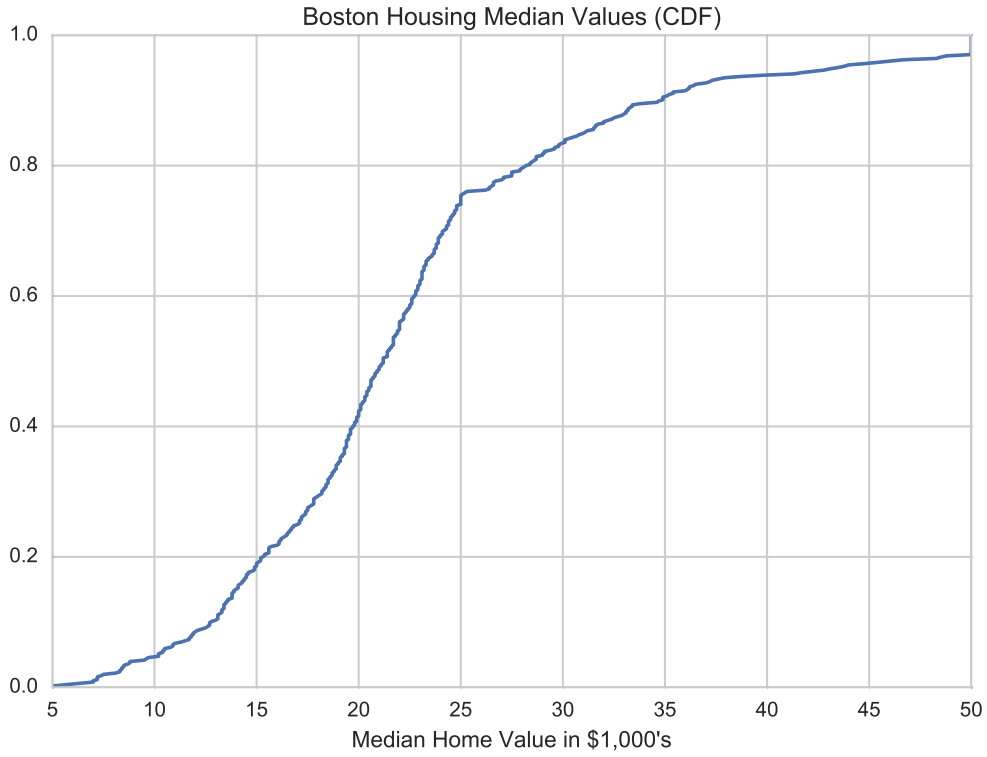
| Item | Value |
|---|---|
| Total number of instances | 506 |
| Total number of features | 13.0 |
| Minimum house price | 5.0 |
| Maximum house price | 50.0 |
| Mean house price | 22.53 |
| Median house price | 21.2 |
| Sample Standard deviation of house price | 9.19 |

## 1.4 Plots


Boston Housing Median Values
mean
median
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−10
0
10
20
30
40
50
60
median_value


Boston Housing Median Values
5
10
15
20
25
30
35
40
45
50
median_value



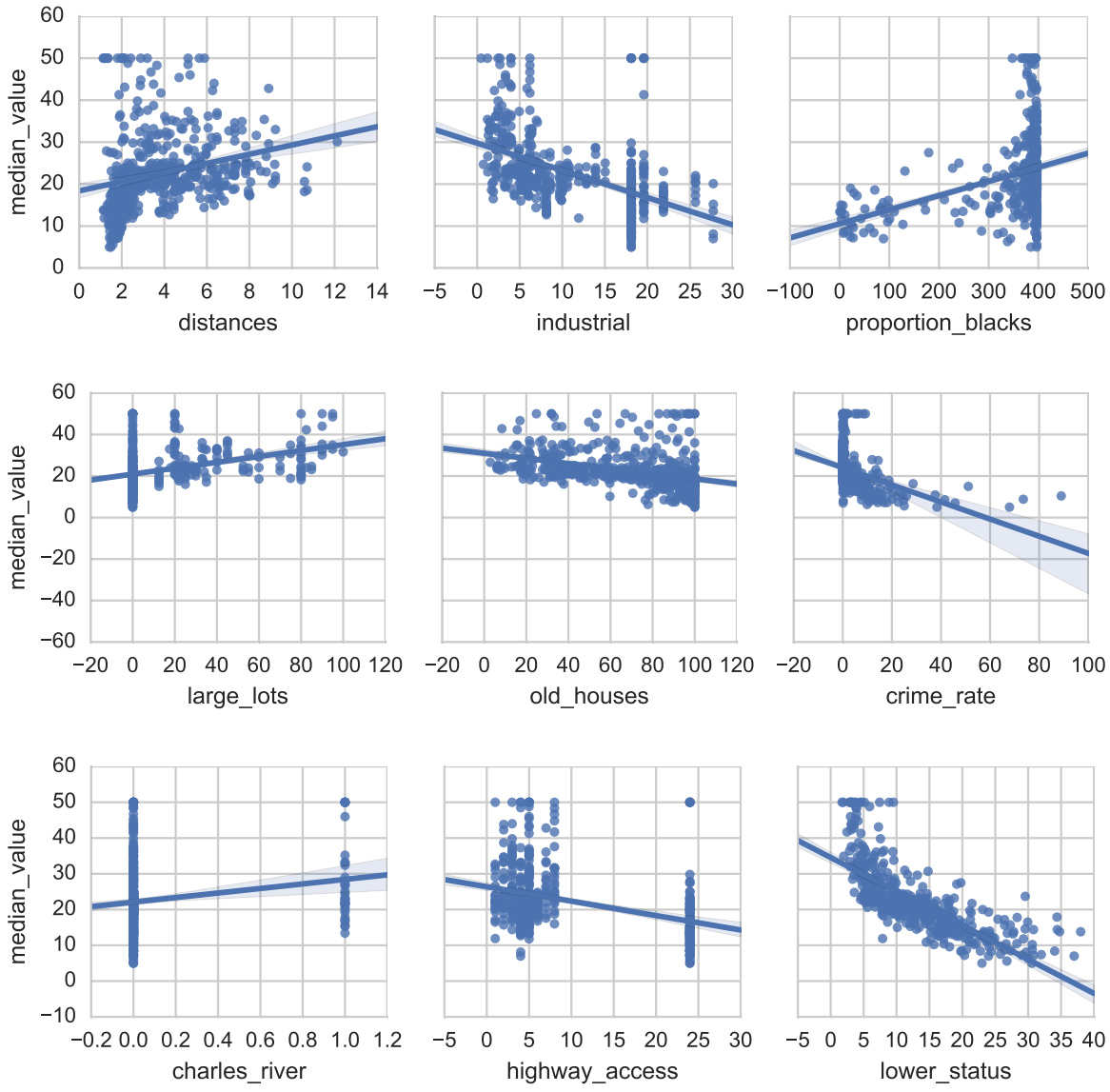

Looking at the distribution (histogram and KDE plot) and box-plot the median-values for the homes appear to be

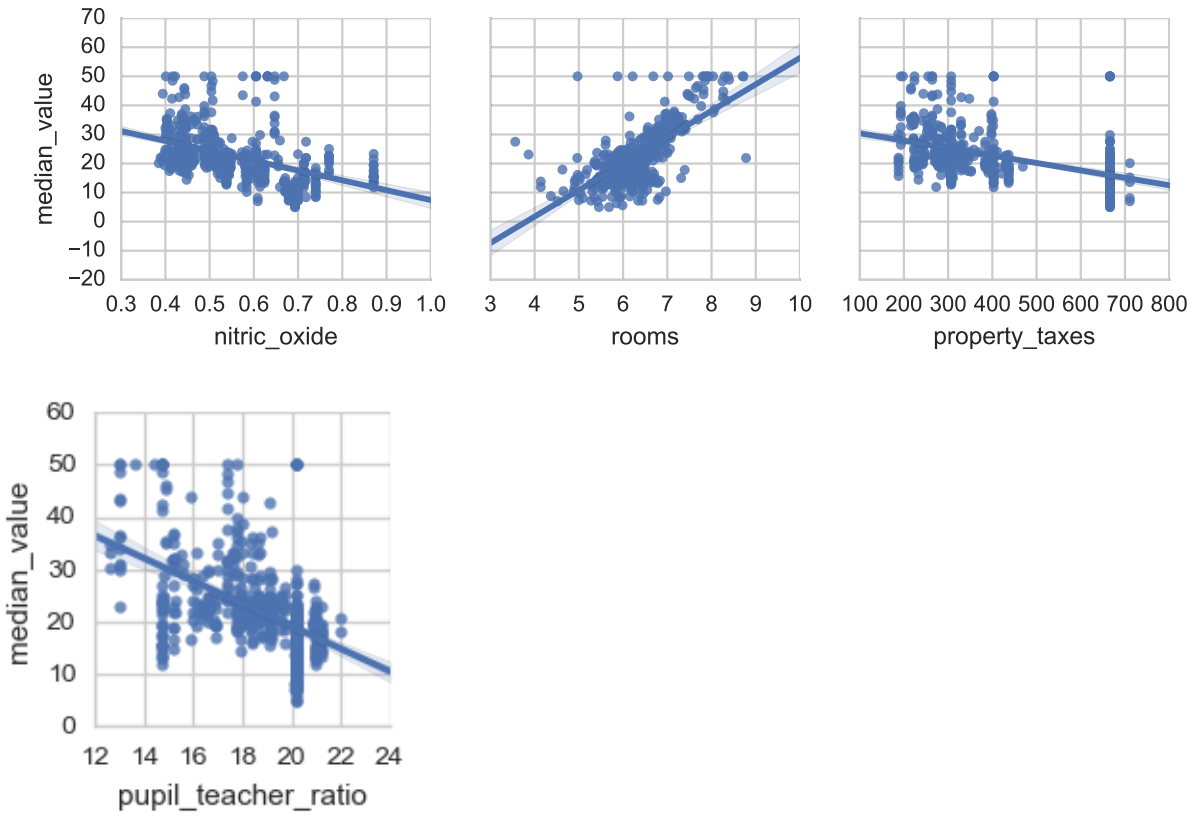
right-skewed. The CDF shows that about 90% of the homes are $35,000 or less (the 90th percentile for median-value is 34.8). The qq-plot and the other plots show that the median-values aren't normally distributed.

## 1.5 Question 1

*Of the features available for each data point, choose three that you feel are significant and give a brief description for each of what they measure.*

To get an idea of how the features are related to the median-value, I'll plot some linear-regressions.






Looking at the plots, the three features that I think are the most significant are *lower_status (LSTAT)*, *nitric_oxide (NOX)*, and *rooms (RM)*. The *lower_status* variable is the percent of the population of the town that is of 'lower status' which is defined in this case as being an adult with less than a ninth-grade education or a male worker that is classified as a laborer. The *nitric_oxide* variable represents the annual average parts per million of nitric-oxide measured in the air and is thus a stand-in for pollution. *rooms* is the average number of rooms per dwelling, representing the spaciousness of houses in the suburb (Harrison, 1978).

## 1.6 Question 2

*Using your client's feature set "CLIENT_FEATURES", which values correspond with the features you've chosen above?*

Table 4: Client Features

| Variable | Value |
|---|---|
| lower_status | 12.13 |
| nitric_oxide | 0.66 |
| rooms | 5.61 |

And now some summary statistics for the same variables in the boston-housing data set.

Table 5: Variables Summaries

| Variable | Min | Q1 | Median | Q3 | Max | Mean | Std |
|---|---|---|---|---|---|---|---|
| lower_status | 1.73 | 6.95 | 11.36 | 16.96 | 37.97 | 12.65 | 7.14 |
| nitric_oxide | 0.39 | 0.45 | 0.54 | 0.62 | 0.87 | 0.55 | 0.12 |
| rooms | 3.56 | 5.89 | 6.21 | 6.62 | 8.78 | 6.28 | 0.70 |

Comparing the values for the client to the median values for the data set as a whole shows that the client has a higher ratio of lower-status adults, more pollution and fewer rooms than the median for the suburbs so I would expect that the predicted value will be lower than the median median-value for the suburbs.

# 2 Evaluating Model Performance

## 2.1 Question 3

*Why do we split the data into training and testing subsets for our model?*

We split the data into training and testing subsets so that we can assess the model using a different data-set than what it was trained on, thus reducing the likelihood of overfitting the model to the training data and increasing the likelihood that it will generalize to other data.

## 2.2 Question 4

*Which performance metric below did you find was most appropriate for predicting housing prices and analyzing the total error. Why? - Accuracy - Precision - Recall - F1 Score - Mean Squared Error (MSE) - Mean Absolute Error (MAE)*

I chose *Mean Squared Error* as the most appropriate performance metric for predicting housing prices because we are predicting a numeric value (a regression problem) and while Mean Absolute Error could also be used, the MSE emphasizes larger errors more and so I felt it would be preferable.

## 2.3 Step 4 (Final Step)

## 2.4 Question 5

*What is the grid search algorithm and when is it applicable?*

The GridSearchCV algorithm exhaustively works through the parameters it is given to find the parameters that create the best model using cross-validation. Because it is exhaustive it is appropriate when the model-creation is not excessively computationally intensive, otherwise its run-time might be infeasible.

## 2.5 Question 6

*What is cross-validation, and how is it performed on a model? Why would cross-validation be helpful when using grid search?*

Cross-validation is a method of testing a model by partitioning the data into subsets, with each subset taking a turn as the test set while the data not being used as a test-set is used as the training set. This allows the model to be tested against all the data-points, rather than having some data reserved exclusively as training data and the remainder exclusively as testing data.

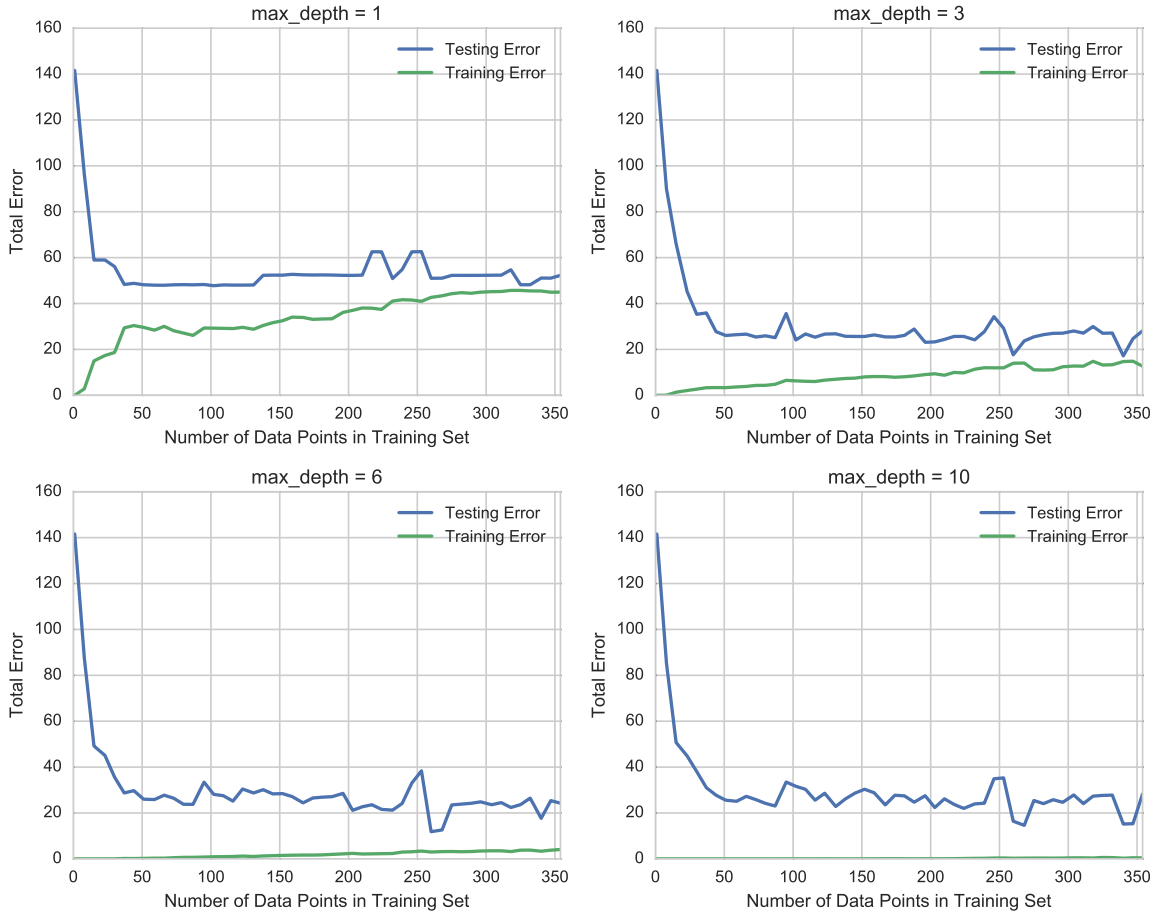
Because grid-search attempts to find the optimal parameters for a model, it's advantageous to use the same training and testing data in each case (case meaning a particular permutation of the parameters) so that the comparisons are

equitable. One could simply perform an initial train-validation-test split and use this throughout the grid search, but this then risks the possibility that there was something in the initial split that will bias the outcome. By using all the partitions of the data as both test and training data, as cross-validation does, the chance of a bias in the splitting is reduced and at the same time all the parameter permutations are given the same data to be tested against.

# 3 Analyzing Model Performance

## 3.1 Learning Curves

The learning curves for different max-depth parameters are plotted.




## 3.2 Question 7

*Choose one of the learning curve graphs that are created above. What is the max depth for the chosen model? As the size of the training set increases, what happens to the training error? What happens to the testing error?*

Looking at the model with max-depth of 3, as the size of the training set increases, the training error gradually increases. The testing error initially decreases, then seems to more or less stabilize around a MSE of about 20.

## 3.3 Question 8

*Look at the learning curve graphs for the model with a max depth of 1 and a max depth of 10. When the model is using the full training set, does it suffer from high bias or high variance when the max depth is 1? What about when the max depth is 10?*

The training and testing plots for the model with max-depth 1 move toward convergence with an error near 50, indicating a high bias (the model is too simple, and the additional data isn't improving the generalization of the model).

For the model with max-depth 10, the curves haven't converged, and the training error remains near 0, indicating that it suffers from high variance, and should be improved with more data.

## 3.4 Model Complexity


## 3.5 Question 9

*From the model complexity graph above, describe the training and testing errors as the max depth increases. Based on your interpretation of the graph, which max depth results in a model that best generalizes the dataset? Why?*

As max-depth increases the training error improves, while the testing error decreases up until a depth of 5 and then begins a slight increase as the depth is increased. Based on this I would say that the max-depth of 5 created the model that best generalized the data set, as it minimized the testing error.

# 4 Model Prediction

## 4.1 Question 10

*Using grid search on the entire data set, what is the optimal "max_depth" parameter for your model? How does this result compare to your initial intuition?*

To find the 'best' model I ran the *fit_model* function 1,000 times and took the *best_params_* (max-depth) and *best_score_* (negative MSE) for each trial.

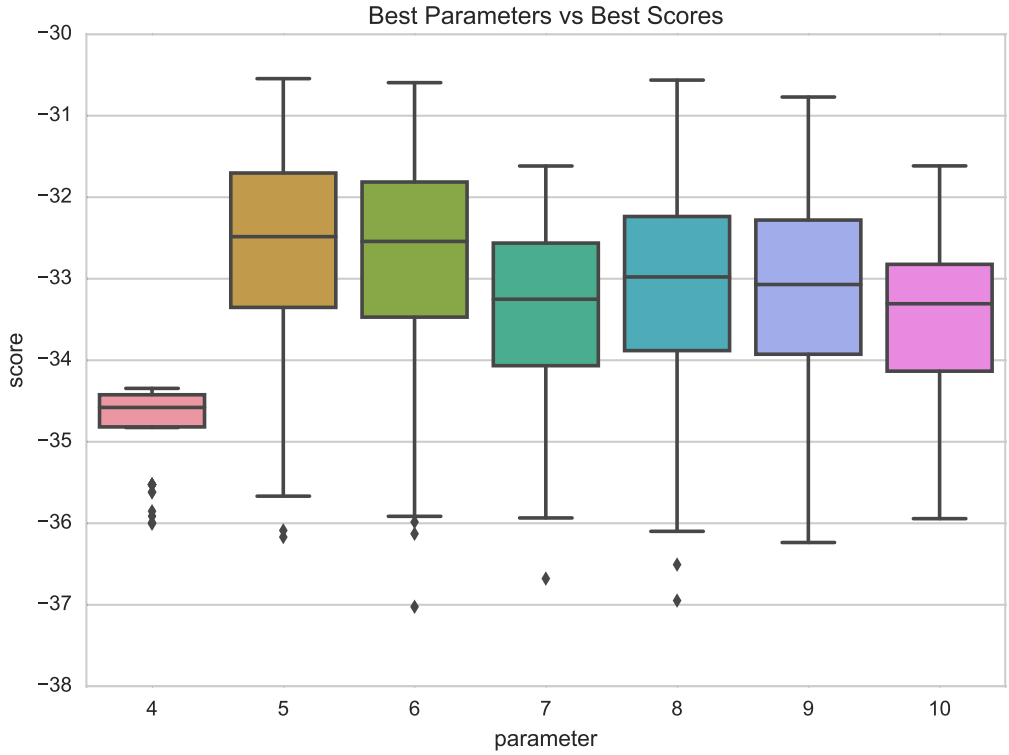



Table 6: Parameter Counts

| Max-Depth | Count |
|---|---|
| 4 | 295 |
| 5 | 231 |
| 7 | 161 |
| 6 | 141 |
| 8 | 102 |
| 9 | 70 |

Table 7: Median Scores

| Max-Depth | Median Score |
|---|---|
| 4 | -34.58 |
| 5 | -32.48 |
| 6 | -32.54 |
| 7 | -33.25 |
| 8 | -32.98 |
| 9 | -33.07 |
| 10 | -33.31 |

Table 8: Max Scores

| Max-Depth | Max Score |
|---|---|
| 4 | -34.35 |
| 5 | -30.55 |
| 6 | -30.59 |
| 7 | -31.62 |
| 8 | -30.56 |
| 9 | -30.77 |
| 10 | -31.62 |

**Note:** Since the *GridSearchCV* normally tries to maximize the output of the scoring-function, but the goal in this case was to minimize it, the scores are negations of the MSE, thus the higher the score, the lower the MSE.

While a max-depth of 4 was the most common best-parameter, the max-depth of 5 was the median max-depth, had the highest median score, and had the highest overall score, so I will say that the optimal *max_depth* parameter is 5. This is in line with what I had guessed, based on the Complexity Performance plot.

## 4.2 Question 11

*With your parameter-tuned model, what is the best selling price for your client's home? How does this selling price compare to the basic statistics you calculated on the dataset?*

Table 9: Predicted Price

| | |
|---|---|
| Predicted value of client's home | $20,967.76 |
| Median for all suburbs - predicted value | $232.24 |

My three chosen features (*lower_status*, *nitric_oxide*, and *rooms*) seemed to indicate that the client's house might be a lower-valued house, and the predicted value was about $232 less than the median median-value, so our model predicts that the client has a below-median-value house.

## 4.3 Question 12

*In a few sentences, discuss whether you would use this model or not to predict the selling price of future clients' homes in the Greater Boston area.*

I think that this model seems reasonable for the given data (Boston Suburbs in 1970), but I think that I might be hesitant to predict the value for a specific house using it, given that we are using aggregate-values for entire suburbs, not values for individual houses. I would also think that separating out the upper-class houses would give a better model for certain clients, given the right-skew of the data. Also, the median MSE for the best model was $\sim 32$ so

taking the square root of this gives an 'average' error of about $5,700, which seems fairly high, given the low median-values for the houses. I think that the model gives a useful ball-park-figure estimate, but I think I'd have to qualify the certainty of prediction for future clients, noting also the age of the data and not extrapolating much beyond 1970.