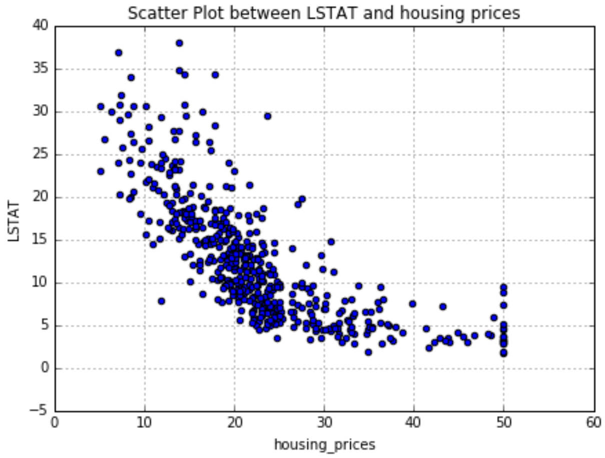
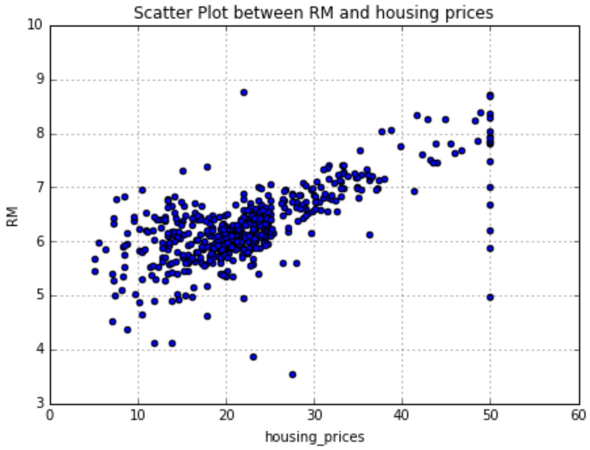
Statistical Analysis and Data Exploration

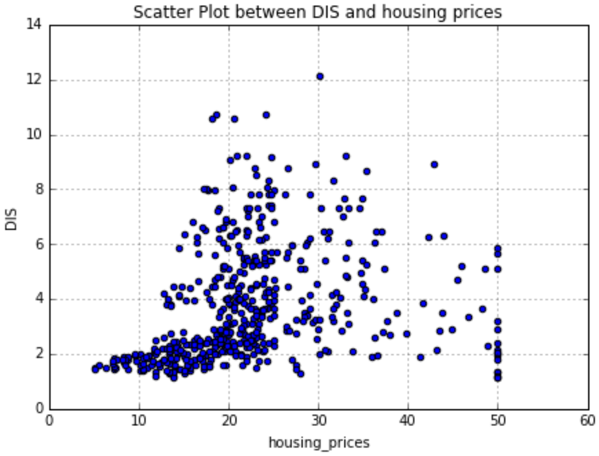
1) Of the available features for a given home, choose three you feel are significant and give a brief description for each of what they measure.

After correlating all features against houses prices, we conclude that there are 3 significate features

1. RM: average number of rooms per dwelling
2. DIS: weighted distances to five Boston employment centers
3. LSTAT: % lower status of the population







2) Using your client’s feature set `CLIENT_FEATURES` in the template code, which values correspond to the chosen features?

From the feature above we can see that:

LSTAT: % lower status of the population = 12.13%

RM: average number of rooms per dwelling = 5.609

DIS: weighted distances to five Boston employment centers = 1.385

Evaluating Model Performance

3) Why do we split the data into training and testing subsets?

Splitting the data into training and testing subsets, give us estimates of performance on an independent dataset (training set), on top of that having separate training and testing serve as check on overfitting by testing the generated model on the training set, if it's perfectly matches the training data that means the model has failed to generalize its predictions to the larger population and we need to change the way we split the data, maybe by doing shuffling or any other mechanism

4) Which performance metric below is most appropriate for predicting housing prices and analyzing error? Why?

- Accuracy
- Precision
- Recall
- F1 score
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)

The first four options used mainly for Classification Metrics, Boston housing project has a continuous data so it's a regression model , with that being said we have to choose between MSE & MAS, I think MSE is a better choice since it emphasizes larger errors rather than smaller errors.

5) What is the grid search algorithm and when is it applicable?

It's a parameter optimization method, which is work through multiple combinations of parameter tunes, where user gives the function a ranges of parameters, then function will interpolates them making a grid and evaluates each point of this grid

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

It’s an evaluation of a model by of taking a sample of data into number of ‘folds’, performing the analysis on group of folds, and validating the analysis on the other folds. So to reduce variability, multiple rounds of different partitions, and the validation results are averaged over the rounds. In order to pick the model which generates the lowest error, cross validation is very help when using grid search because it help us find the optimal model to be used on the testing set

It’s an evaluation function used to evaluate model performance, it achieves this by randomly splitting the dataset to training & testing sub set, it repeat this technique with enormous ‘folds’, and optimize the model using performance metrics for each random split of training and testing set, then it measures the model performance based on average errors resulted from predicting label values from the test set over all randomly split versions of the dataset.

CV is very helpful for grid search algorithm, since it provides enhanced measures of model performance for parameter values tested in the grid, the only alternative we have for CV is single train/test split, where grid search would possibly identify the selected features as optimal features and we falls into an overfitting problem, and eventually the model is not generalized enough

Analyzing Model Performance

7) Choose one of the learning curve graphs your code creates. What is the max depth for the model? As the size of the training set increases, what happens to the training error? Describe what happens to the testing error.

The max depth is 3, we notice that as the size of the training set increases the training error increases as well, and testing error decreases. From the graph above, we can see that with slight increase in the training set, there is a drastically sudden decrease in testing error, after training set pass the point 50, no significate improvement on the testing error graph as we increase the training set.

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?

For max depth of 1: as we use one feature only (or 1d), it suffers from high bias, which means the training error curve is having a high level of error, no matter how we increase the number of data points in the training set.

For max depth of 10: since this graph represent a very complex model which use 10 features (10d), this makes the curve suffers from high variance problem (overfitting), we can identify a case as high variance when the difference between the testing error and the training set error is high

9) From the model complexity graph, 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?

From the graphs above we can notice that as the model complexity increases, the testing error (blue line) decreases. In this case example I think the best model complexity is with max_depth = 4, because it’s simplest model complexity which give us the optimal testing error value, this complexity level would keep the model generalized enough to predict future data, without falling into the curse of dimensionality

Model Prediction

10) Using grid search, what is the optimal max depth for your model? How does this result compare to your initial intuition?

The max_depth as per reg function is 4, which is match my initial selected dimensions

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 statistics you calculated on the dataset?

The predicted value of the home is $ 21,630. This price is perfectly fall within the acceptable range of the basic statistics range mean+-(1sd), which is between 13,345 & 31,721 USD

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 Boston area.

I would use this model, because as mentioned above the predicted value falls between the acceptable range on mean and standard deviation which can successfully cover 97% of the houses in Greater Boston area.