Regression Tree Details

For building the regression tree model, I used the R package *tree*. Surprisingly, the full tree only included 10 splits. Since we had such a wide dataset, I expected much more than 10. The variables involved in the split were (in descending order of importance) OverallQual, GrLivArea, TotalBsmtSF, CentralAirY, YearBuilt , and GarageCars. The choice of variables is notable because all together, they provide a somewhat comprehensive view of the house. GrLivArea is the above ground square footage of the house, so that combined with TotalBsmtSF gives an idea of how big the house is. I then used cross validation provided by the package to see the tree’s performance on the training set as the number of splits varied. The resulting deviance graph clearly showed that the more splits, the better the performance, and this turned out to be true for performance on the test set as well. The more pruning was done to the full tree, the more the prediction error increased on the test set. The regression tree model seems to be stable, as it chooses the same variables for splits, and roughly the same split locations each time. As a result, between multiple runs of the same model, results do not change very much.