David Zamojda

Assignment 2

1. Prior to data exploration and analysis, I used the One Hot Encoding method in order to be able to use categorical predictors. Then, I replaced all NaN values with the most frequent value. This gave me a slightly better R squared value than using the mean or median.
2. Prior to making any linear regressions, I graphed some data in order to look at the predictors that I thought would correlate with a higher selling price. I graphed things such as the lot area, lot frontage, as well as garage area and found that in general, the larger these variables were, the higher the selling price. I predict that the information about the house, such as lot area and number of rooms, will account for much of the selling price. Also, I believe that the year of the house sold will be important, as the data contains houses that were sold before and after the housing crisis
3. I used several different techniques to obtain the best R squared value. First, I created my own list of predictors that I thought would give a decent R squared value. Second, I used the Lasso regression with different alpha values, and tested each one to see which one gave the best result. I used alpha values of 0.1, 0.2, 0.25, 0.5, 1.0, 2.5, and 5.0. Next one I tried was using Selector F, which chose the best 50% of values based on their f-scores. For the best performing model, I used the R squared value, median absolute error, mean squared error, and explained variance score to get a better understanding of the model.
4. When trying to predict the sale price using the base model on the second data set, the error metrics that came back were very weird. The base model returned an R squared value of -2.16085229505e+22, a mean absolute error of 4.01836520113e+15, a mean squared error of 1.21387294407e+32, and an explained variance value of -2.16085229505e+22. These odd values indicate that the base model that was made for the first data set does not work for the second data set. The results from the Lasso regression were much better, with the best Lasso regression R squared value being 0.908305551532 with a alpha of 30. My hand picked set of predictors did not fair as well as the Lasso regression, but was better than the first with an increase from .79896 to .81039. This indicates that my handpicked predictors were better predictors than in the first data set. I believe that my best model is a good model because the other error metrics indicate that the predicted values are relatively accurate.