Assignment 2

Part 1

1. I first started off by deleting the PID because I felt it was useless information. I then transformed the data by converting the categorical data into dummy variables and the numerical data with no value to be set to the mean of each column. I did both these transformations so that the categorical data would represented as a number and could be used by the algorithm and so that the numerical data’s mean wouldn’t be too heavily influenced by using a 0.

Part 2

1. I deleted various column and plotted them to check to see how the correlation changed between each test. Most of them had an impact on how to sales price would be.
2. If the houses were in certain neighborhoods or if they had fences increased the price. As well as whether the house had more than the mean square footage of all the houses. The condition of the house as well as the time it takes to sell the house will be an impacting predictor as well.

Part 3

1. I used the Lasso method to train my model, which allowed it to select variables based on the predictors with coefficients that equaled zero. This will remove most of the features that won’t be useful for predicting. I found that the best alpha value was 0.25 for me because it offered the highest R2 of 0.989, a MSE of 72,422, and a MAE of 474. It outperformed my base model which had an R2 of only 0.95.

Part 4

1. My model didn’t do as well with the test data as it did with the training data. So I’d say it was a decent model, maybe it would have done better with some slight manipulation of the test data.