# Linear Regression

# Data Cleaning

The first step in preparing our data for linear regression was to one-hot encode the categorical variables. The resulting dataset has 364 features.

Next, we randomly split the data into training and test sets, assigning 80% of the observations to the training set and 20% of the observations to the test set.

Subsequently, we imputed missing values using the median value of the feature in the training set. Only three features (lot frontage, masonry veneer area, and bike score) were missing values, and no one feature was missing more than 18% of its observations.

Finally, we normalized the features in the training and test sets to have a mean of zero and a standard deviation of one. We normalized the test set based on the normalization parameters of the training set.

# Results

# Basic Linear Model

Our first model was a basic OLS linear regression model using all 83 features in our dataset.

[INSERT RESULTS]

# Regularized Linear Model

For our second model, we applied Lasso, Ridge, and Elastic Net regression to our dataset. We performed a grid search and used 5-fold cross validation to determine the best alpha hyper-parameter ranging from 0.1 to 10,000.

[INSERT RESULTS]

# Linear Model with Polynomial Features

For our third model, we applied an OLS linear regression model with polynomial features. We performed a grid search and used 5-fold cross validation to determine whether the optimal number of degrees for the polynomial features was two or three degrees.

When we first attempted to run this model, it ran unsuccessfully due to inadequate memory. As a result, we decided to exclude the features that were dropped by the Lasso regression model (i.e. the features that the Lasso model determined to be the worst predictors of a home’s sale price). We chose to drop features based on the Lasso model, because the Elastic Net and Ridge models only dropped five features each. The Lasso model, on the other hand, dropped [INSERT NUMBER] features, so we dropped these same features in order to ensure that our pruned dataset was small enough for Jupyter Notebook to handle.

[INSERT RESULTS]

# Regularized Linear Model with Polynomial Features

For our fourth and final linear model, we regularized the best-performing linear model with polynomial features (i.e. the model with two degrees). We applied Lasso, Ridge, and Elastic Net regression to this model and used 5-fold cross validation and grid search to determine the best alpha hyper-parameter ranging from 0.1 to 10,000.

[INSERT RESULTS]