Assignment 2: Regression Modeling & Prediction

For this assignment you will develop a regression-based model to predict housing prices in Ames, IA. You will find two data files accompanying this document:

1) AmesHousingSetA.csv
2) AmesHousingSetB.csv

You will use dataset A to build your model (split this into training and test sets) and then validate your model on dataset B by making predictions on it. In this assignment you will submit a Jupyter notebook containing the following:

1) A GitHub link to your notebook which performs this analysis
2) Your notebook will include written responses to the questions below, in addition to just the code itself.

I. Data Preparation Questions

1) What specific data transforms did you perform prior to exploration and analysis, and why did you choose these?

II. Exploratory Analysis Questions

Perform an exploratory analysis on your data by visualizing and/or applying other means of data exploration.

1) What (if any) insights jump out at you?
2) Do you have any hypotheses about relationship of certain variables to the price?

III. Model Building

First construct a baseline model (containing all predictors) to predict the price. Then build the best model you can devise. In this part use ONLY dataset A and DO NOT TOUCH dataset B. You will want to split this into training and test sets and apply error metrics/compare models only on the test data.

1) What approach did you use to arrive at the best model? Why did you select this approach?
2) Which error metric(s) are you using to compare performance? What is the value(s) of the error metric(s) for the baseline model and your best model?

IV. Predicting and Validating

Run your baseline and best models on dataset B. DO NOT do any further training. Remember to apply all transforms you used in building the model to this set (use the transform function on the preprocessors you created in part I).

1) What are the respective error metric values for each model on this set? How did your best model do on this data as compared to the baseline?
2) Is your best model a good model? Why or why not