Lab 2: Linear Regression & Data Visualization with Python

This lab is worth 100 points total. You must show code to back up your results i.e. if you report just a number, you will get zero credit. Submit your code in a Jupyter Notebook file to Blackboard. Clearly label each section of your code that responds to a particular question. There should be a labeled section for Question 1, 2, ...etc. For example, you could call the first section "Question 1: Reading in datasets"; the second section "Question 2: Merging train and properties datasets", etc.

Due date: 6/20 at 11:59PM

Background

The data you will be using is housing data from Zillow (same as Lab 1, but with one additional dataset) and includes attributes such as garage size, bathroom count, fireplace flag etc. A data dictionary is available under the Lab 2 assignment on Blackboard.

Problems

1.Read in train_2017.csv and properties_2017.csv **(5 points)**

2.Merge these two datasets using **parcelid** as the key. Use an inner join. Report the number of rows and columns in the merged dataset. Note that this will drop some rows, but for the purposes of this lab, we're only interested in those records where we have a match between these two datasets **(5 points).**

3. When building a model, it's a good idea to split your data into train and validation so that you can see how well the model performs on data it has not seen before. Use 70% of the data for training and 30% of the data for validation. **(5 points)**

4. In this lab, you will be using the **logerror** variable as the dependent variable i.e. you will be trying to predict this value from other independent

variables in the dataset. This variable is defined as the logarithm of the difference between Zillow's estimate for a home price vs. the actual selling price for the home. In other words, you can think of it as a measure of the error in Zillow's home price estimate. It's a good idea to explore your dataset before attempting modeling (particularly the label). In this step, start with getting the descriptive statistics of **logerror** in the training set. What is its mean, median, range, and standard deviation? **(10 points)**

5. Generate a histogram for each variable in the dataset **(10 points)**

6. Filter the variables in the dataset to a list of ones missing less than 20% of their values in the training set. For this problem, do not count discrete variables, like **poolcnt** as missing. Instead, impute the values for these. For example, for **poolcnt**, you can impute the missing values as 0. You may need to read through the data dictionary for reference. Return this list of variables. **(15 points)**

7. Based on the list of variables above, select 5 predictors of your choice (i.e. these 5 should not include the label, **logerror**). For these 5 variables, impute the missing values in the training set using the median of the each column's value. For the validation set, impute the missing values for these variables – also using the medians **based off the training set**. **(15 points)**

8. Create a heatmap of the correlation matrix for these 5 variables. Are there any issues with correlation? How might this affect a linear regression model? **(10 points)**

9. Train a linear regression model using these 5 variables as predictors and **logerror** as the response variable on the training set. Report the RMSE and MAE on the training data and validation data. Second, train a baseline linear regression model using the median (or mean) of the **logerror** variable. Compare the performance of the baseline vs. the model with the 5 variables. What else could you do to improve your model? **(25 points)**