Capstone Project 1

Data Wrangling: Prediction of House Price by House Features

By: Neil Sonalkar
12/05/19

Data

The data was taken from the Kaggle competition: House Prices: Advanced Regression Techniques and downloaded as a .csv file. The data is on details of various homes in the city of Ames, IA

Data Wrangling Process

The data started off with 1460 rows and 81 columns, each column is a descriptive feature that could potentially be used to create a prediction model. Each row is a different house record that contains the values of the descriptive features for that specific record. Example, row 1 would consist of the house id, sale price, neighborhood, etc. Out of the 81 columns, I narrowed the data down to 34 columns. I removed 47 columns for multiple reasons including, entire column of missing values, entire column of only one value, columns with repetitive information, and columns with not enough data values. Because each row is an independent record and each column is a different feature, missing values could not be filled in from surrounding data and subsequently that row was dropped, leaving 1103 records left from the starting 1460.

Next, the data needed to be checked for outliers. For each column/descriptive feature a value was calculated that served as a threshold between outlier and non outlier data points. For the outliers, their value was reassigned to be the threshold value, essentially getting rid of all the outlier values.