Zach Barnhart  
DSC 530 Exploratory Data Analysis  
Final Project

The problem of evaluating the potential market value of homes is a question that concerns a diverse cross-section of society - from potential home buyers looking for a decent place to live all while securing their savings in the equity of their home, which they would like to see grow steadily over time, to large real estate speculators who make money on the arbitrage of thousands of homes daily across the entire US market, to home flippers who identify homes whose other characteristics could lend it value if condition of the house and property itself were to be improved. In this project, we turn our attention to this issue, in the hopes of gaining some insight into the most valuable predictors of home value, and use this insight to find potential undervalued houses in the markets which we will examine.

In our EDA, we discover several variables in a dataset of Ames, Iowa homes that have a strong relationship with prices of homes. Perhaps predictably, the strongest predictors of home value are related to lot size, living area size, and size of other various features of the home. Additionally (and also unsurprisingly), quality-related features also correlate somewhat with home value. We statistically demonstrate that certain intangible characteristics such as neighborhood have a measurable impact on home prices, even after controlling for the material differences between the houses in those neighborhoods. We build a linear model to predict home price (in our final model, log of sales prices) from a small subset of available features, achieving an R-squared of 0.854, in other words, a model that explains 85.4% of the variance in the target variable.

Our analysis, since we only looked at a subset of features, obviously lacks the power of a model that could be developed by fully integrated all features in the original dataset (as well as creating other features that may isolate pieces of information which are shared or split between two or more features currently). I also feel like a lot more could be done with data transformations. In reality, all area features, due to their strong right-skew, should probably have log applied to improve the predictive power of the model, but we decided to use as few transformations as possible so that the interpretability of the model did not suffer.

I feel like integrating more or the feature variables (fireplace, pool, etc) could have only helped our model (these features are generally seen as valuable in all cases). In addition, there are a number of condition variables we could have encoded and included, but this would have been rather tedious, as each of these variables would require a separate encoding. It is not obviously clear that some of the other variables would have helped – their possible usefulness would need to be explored.

One of the assumptions I made was that it was preferable to handle outliers by taking the log of our response variable, but there is an argument to be made that the appropriate approach is to leave the variable themselves untouched, but to clear outliers with a rather heavy hand so that the model is able to perform very well on most houses, at the risk of not performing well on unusual properties.

One challenge I faced what knowing when to stop. As EDA is a cyclical, iterative process, there is an impulse to clean data, process data, build model, transform data, build model again, create new feature, build model again, and on and on forever. At some point you have to know when you have achieved your basic goals and tie things up, as you can keep wrangling data, adjust hyperparameters, trying new algorithms, etc, essentially forever.