

## Data Wrangling Summary

The Ames Iowa housing data set had a number of missing sections as would any public data set. To make the data relevant to the project at and early on we needed to make sure the data was fully cleaned. Early on started off by looking for columns with missing values. I relied heavily on some basic pandas commands to locate the columns with missing values. For the most part, 90% of the columns were missing no values which were a pleasant surprise. Most of the missing values were in columns that I later removed or they were not relevant for the task at hand.

One important column that needed to be created was the column called Total Living Area. The data did not provide a column with the total size of the home, instead, the data set provided the size of the upstairs living area and the size of the basement but not both put together. After combining the two columns into the newly created Total Living Area column I removed the two columns to be sure that this info did not receive too much weight.

The final challenge to undertake during the data wrangling was the creation of the dummy variables for the categorical data. The first step that I took for locating the categorical data was to select all the data types and exclude int64 and float 64. After having a list of all the categorical data I converted the categorical data to categorical\_list. Finally, after takign on these steps, I had no missing values, and dummy variables which allowed me to move onto building the models.