**Assignment #1**

Brandon Moretz

### Introduction

To accurately forecast the value of a home, we must find a relevant dataset that contains accurate information of comparable inventory so that we can explore the significant variables of a home which ultimately determine the sale price of the residence. Once we have explored the dataset, our task will be to create a multivariate regression model that leverages these key indicators in the data to predict the value of a home given based upon its features.

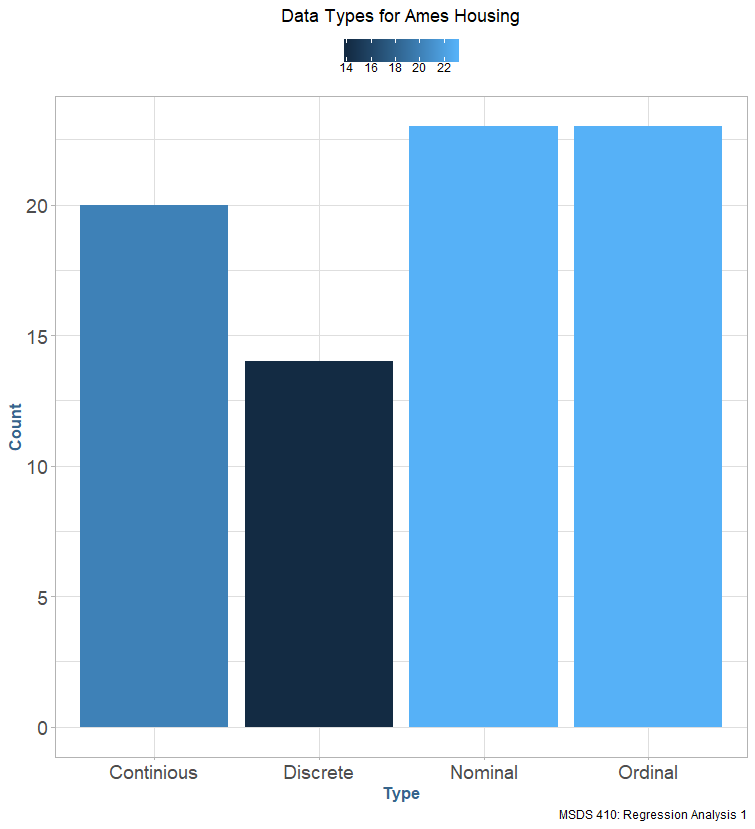
In this report, we will use the Ames dataset which is an alternative to the famous Boston housing data to perform exploratory data analysis through variable derivation, validation, selection and visualization to measure the relevance of these indicators as they pertain to the value of the home in terms of a dollar estimate.

### Data Survey

This data is from the Ames Iowa Assessor’s Office and contains characteristics regarding residential properties sold in Ames from 2006 to 2010.

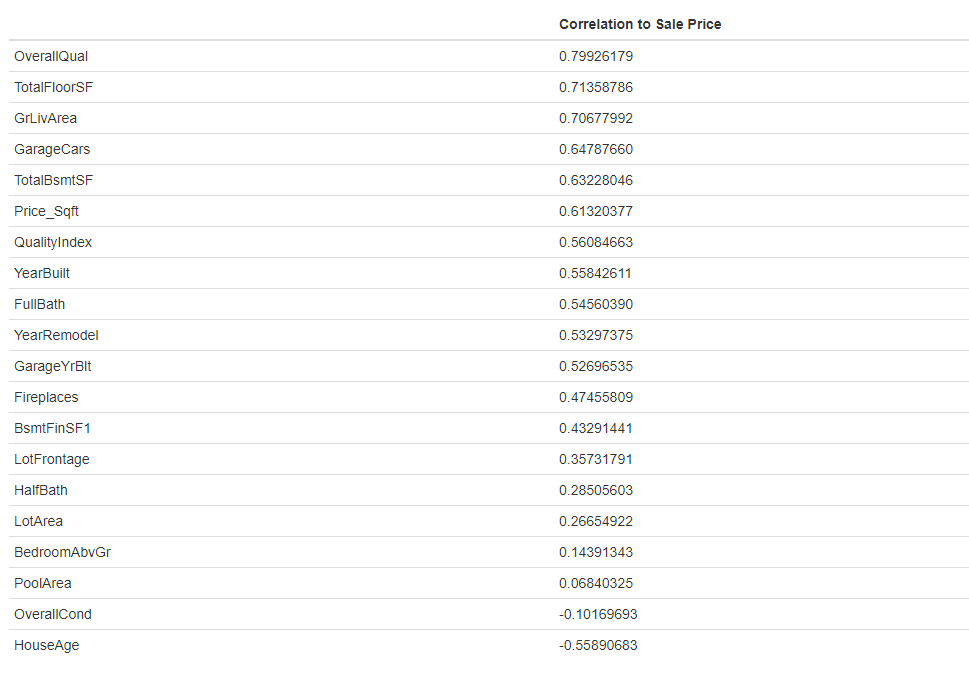
The Ames housing dataset contains approximately three-thousand observations of eighty-two variables collected from the Ames Assessor’s Office specifically for the purpose of assessing value of individual residential properties sold in Ames, Iowa from 2006 to 2010. Given that this data was collected for the purpose of assessing home values, it should be an ideal source of information for our observational study and resulting regression model.

Taking a deeper dive into the eighty-two characteristics of each property, we can classify twenty-three as nominal, twenty-three as ordinal, fourteen discrete and twenty continuous.



In addition to the variables provided in the dataset, we can also derive our own calculated and derived variables for use in our predictive model. Given that our overall goal is to predict the sale price we can calculate the total square footage of the house by combining the square footage of all the floors and break each property down to a price per square foot. This will provide us with a generalized common denominator for each property so that we can assess the impact of features such as house style, neighborhood and quality on a per-square foot basis.

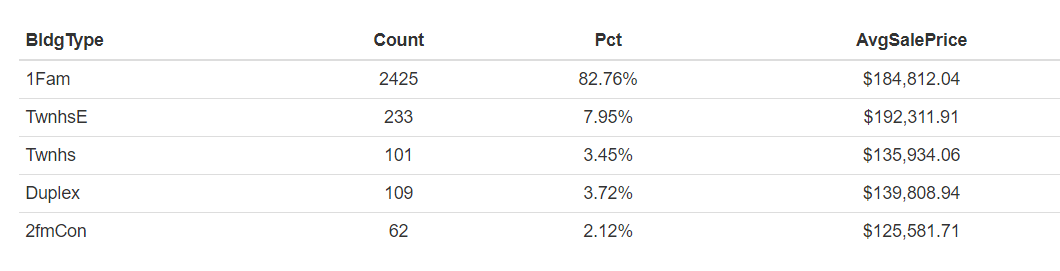
Since we are interested in building a linear model to predict a homes sale price given its features, we should consider which variables are correlated to the sale price. In the following figure we can see the correlation of several variables to the sale price, which we will note here and explore more in-depth in the preceding sections. The following variables were chosen from the total dataset based on initial intuition around housing prices and what drives them.



These variables will form the baseline for our model. The simple correlation does not tell us explicitly how useful a given variable will be in building a model, but it is useful to note that intuitively overall quality, total square-footage, and living are have a highly positive correlation, and house age is a strong negative correlation as we would suspect with our simple intuition.

### Sample population

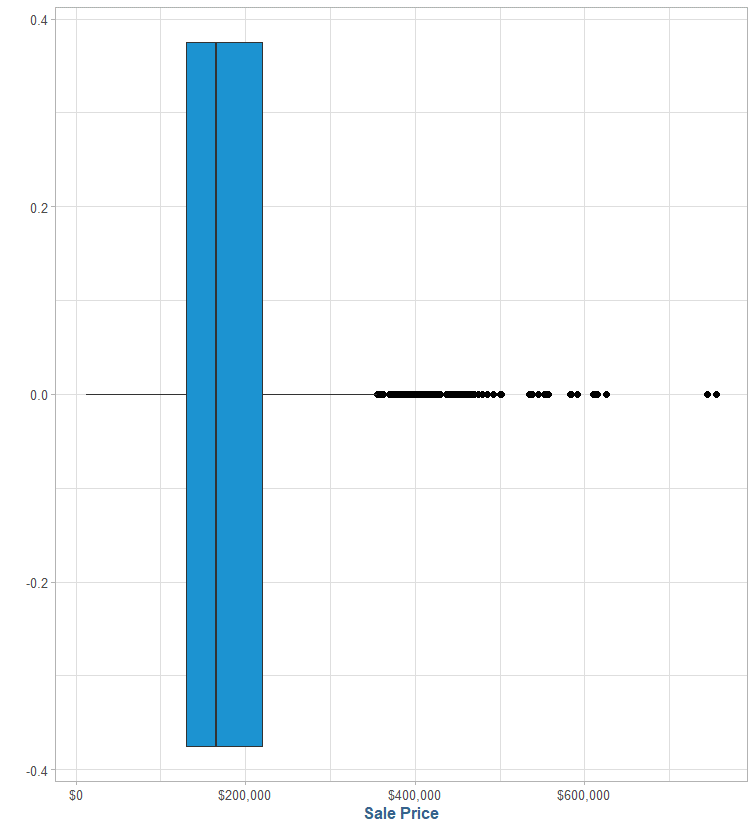
To check the similarity to other property types we can examine the average sale price by building type in the following table:



In the preceding table we can clearly see that single-family homes make up most of the data. Given that many homes are of type single-family, and the other building types make up such a small portion of the observed data it might over complicate the model in order to accommodate the special cases for these homes. Therefore, we will exclude any homes not of type single-family from the proceeding analysis in order to meet the objective of predicting a ‘typical’ home price.

### DATA QUALITY CHECKS

We will perform some standard data quality checks against the primary target variable, sale price. No negative values or non-applicable values are reflected in the population. However, there are some cases of relatively extreme outliers in the sale price as we can see in a boxplot for the variable in the following figure.



The values on the far right represent homes with valuations over $700,000 which is far greater than the average value which is the objective of this analysis, therefore they will be excluded from the rest of this analysis. The boxplot also shows a considerable amount of kurtosis in the distribution of sale prices, so special attention will be given to monitor the impact on the residuals.

### Research

### Conclusion