

Assignment #1

Brandon Moretz

INTRODUCTION

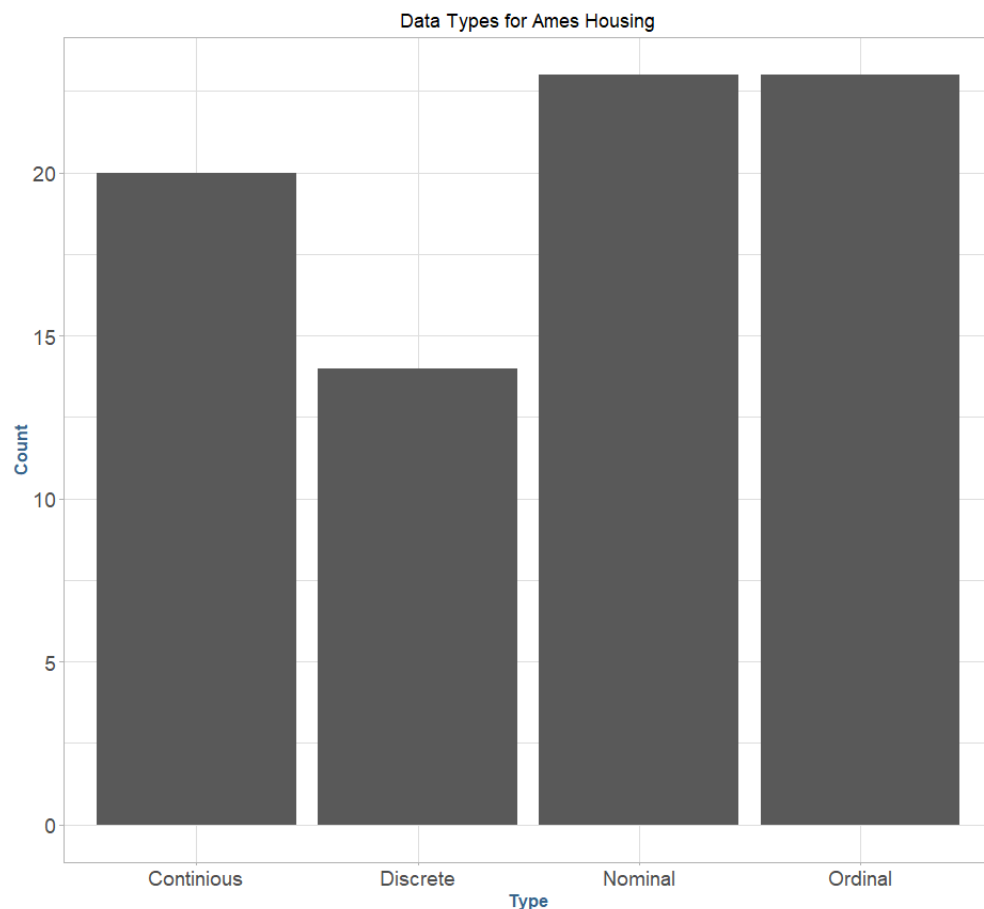
To accurately forecast the value of a home, we must find a relevant dataset that contains accurate information of comparable inventory so that we can explore the significant variables of a home which ultimately determine the sale price of the residence. Once we have explored the dataset, our task will 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

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.

We must use caution in this approach as generalizing

SAMPLE POPULATION

RESEARCH

CONCLUSION