What Determines House Prices in Ames, Iowa

Bryan Klee and Joe Lucas

July Xth, 2022

1 Introduction

Background and where data comes from
Maybe some other articles?

1.1 Intro Para

Houses are important m'kay

1.2 Data

We found a data set of house sales in Ames, Iowa between 2006 and 2010 gathered from the Ames Assessor's Office. The data includes almost any kind of attribute a house can have, ranging from how many rooms there are and the square footage of the house, to if there is a pool or garage and what type of road goes to the home. In total there were 2,930 homes sold between 2006 and 2010, each having 79 descriptive attributes. This amount of data is both helpful and hindering. With the large amount of houses and attributes it should give more accurate results, but will require more effort in cleaning the data overall and in narrowing down the data set to only the most important factors.

2 Materials and Methods

The housing data set we have has 80 variables for each house, which include 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables. In order to learn more about our data we must start looking into it. We first want to look at the categorical variables to remove unwanted houses and variables that may not be helpful to us, for example if they have a lot of null or empty values. Some of the houses or rows of data we don't want are "houses" that are zoned as anything but residential. The data set includes properties that were sold between the time frame that weren't exactly residential homes, for instance some of them were zoned commercial and since our target is residential homes we will remove those not zoned as such. There were also 4 variables which had a null value

for over half of the houses. A bar graph of houses by zone and variable by null counts are below:

PUT SOME GRAPHS HERE

After removing rows that aren't houses and slimming down variables that dont have information in them, we can start to take a look at the composition of our data. Keeping on with the categorical variables, seeing large amount of null values, we take a closer look at the distribution of the categorical variables and the variables themselves. One of the first things noticed is that some of the categorical variables have related numerical variables, making them redundant. For example there is a variable Pool which has either yes or no for if there is a pool, but there is also a variable called PoolSQ which has a number for the square footage of the pool, with 0 if there is no pool. This means we can remove the pool variable without losing anything. Now looking at distribution we can see that some of the variables are very skewed one way

PUT PIE CHARTS HERE

Moving onto our numerical variables; we want to test for normality?

Finally we have our target variable. Sale Price for the house is our target variable, as that is the ultimate outcome when looking at all the variables together. It is a pretty self explanatory variable being the price the house was sold at, as a numerical variable. Since it is our target, we want to inspect it more closely and better understand and prepare it for analysis. We can see from the bar chart that the sale price is pretty heavily skewed to the right and will need some adjustments to be made.

PuT MORE Charts HERE

What we do to Sale Price

Maybe talk about some stuff we did before MVA?

First look at data / pre-processing

exploratory methods

maybe final method

3 Results

maybe discussion as its own section?

4 Conclusion