We are working with housing data from Kaggle for our project. Our data represents 79 explanatory variables that we will use to predict the sale prices of homes in Ames, Iowa. These variables cover features within the house, the yard, and the surrounding neighborhood. Within the dataset, we’ve identified 10 key variables to focus on while we conduct our analysis: number of bedrooms, number of bathrooms, lot size, year built, year sold, garage capacity, floor plan square footage, type of utilities available, neighborhood, and house style.

This data will be useful for determining the sale price of a house by offering a variety of factors that could attract home buyers to a certain price point. Square footage is included in the list of input variables because the size of a house is traditionally positively correlated with price. Square footage is what most people think of in terms of a house’s size and value (the bigger, the more expensive). However, square footage is just one of multiple facets of a house’s value.

We will also look at the utilization of the house’s space as a measure of value. To do so, we will study the number of bedrooms and bathrooms in the house, as well as the garage capacity. These variables are also indicators of the house’s size, as bigger houses tend to boast more bedrooms and bathrooms, along with more garage space. Additionally, lot size can be valuable when calculating the sale price of a house. A bigger space traditionally increases the value of a house, even if the house’s square footage does not increase.

We will also study the year the house was last sold, which could be an indicator of how modern the house’s foundation, construction, and appliances are. Modern appliances are often sought after by homebuyers, so more modern houses could be an indicator of the desirability, and therefore the price, of the house.

The style and features of the house are also potentially important to the final sale price. Homebuyers have preferences for certain house styles (modern, craftsman, Victorian, and more). Additionally, certain amenities may appeal more to audiences than others, which could make a house more appealing.

Lastly, potential buyers look for neighborhood amenities, such as nearby schools, pools, and more. The neighborhood that a house is located in can make the house more or less desirable, therefore affecting the price.

One challenge that we faced in using these variables is that many of them were split up into multiple variables in the dataset. For instance, the number of bathrooms in the house were represented as full baths above grade, basement full baths, half baths above grade, and basement half baths. We resolved this challenge by combining the variables into one. This will make it easier to work with the data in the future and obtain meaningful results.

An issue that we have yet to encounter but anticipate running into is the presence of confounding variables that are not included in our data set but will affect housing values. For example, events like the 2008 financial crisis and the 2020 COVID-19 pandemic likely would dramatically affect sale price but not necessarily be accounted for in the data that we are looking at. We will look out for strange patterns in the data that cannot be explained by the factors in our dataset to ensure that we find the cause of any break in normal patterns.