Project Proposal

Zack Brodtman

April 2023

Introduction

The sale price of a residential home is influenced by a wide range of factors, including the physical characteristics of the home, the location of the home, and various economic and demographic factors. Understanding the factors that influence house prices is essential for homeowners, real estate agents, and policymakers who want to make informed decisions about buying, selling, and investing in residential real estate.

What are the key factors that influence the sale price of residential homes in Ames, Iowa, and how accurately can we predict the sale price of homes using regression models?

Developing an accurate and robust regression model for predicting house prices in Ames, Iowa, can have important practical implications. In this project, we will focus on predicting house prices in Ames, Iowa, using multiple linear regression analysis. Ames is a small city located in the heart of Iowa, with a population of approximately 66,000 people. Despite its small size, Ames is home to Iowa State University, which has a large student population and a significant impact on the local economy. The city has a diverse range of housing options, from historic homes in the downtown area to modern developments on the outskirts of the city. Understanding the factors that influence house prices in Ames can provide insights into the broader factors that influence housing markets in small and mid-sized cities across the United States. Policymakers can use the results to develop policies and programs that can promote affordable housing and improve access to housing for low- and middle-income families.

The Dataset

The Ames House Prices dataset contains a total of 2930 observations, with 2930 residential homes sold between 2006 and 2010 in Ames, Iowa. The dataset includes a total of 80 explanatory variables (features) and a target variable (SalePrice), which represents the sale price of each house. The dataset includes a variety of features related to the physical characteristics of the homes, such as the size of the lot, the number of bedrooms and bathrooms, the overall quality and condition of the home, and the year the home was built or remodeled. The dataset also includes information on the location of the home, such as the neighborhood and proximity to various amenities. The dataset contains several categorical variables, such as the type of dwelling, the style of the home, and the type of foundation, which will need to be converted into numerical variables with dummy encoding before they can be used in a regression analysis. Below is a small sample of the data.

```
##         PID Lot.Area Overall.Qual Full.Bath Year.Built Heating SalePrice
## 1 526301100    31770            6         1       1960    GasA    215000
## 2 526350040    11622            5         1       1961    GasA    105000
## 3 526351010    14267            6         1       1958    GasA    172000
## 4 526353030    11160            7         2       1968    GasA    244000
## 5 527105010    13830            5         2       1997    GasA    189900
## 6 527105030     9978            6         2       1998    GasA    195500
```

Plan

We will use techniques learned in the linear models class to develop and refine the regression model. We will use feature selection techniques to identify the most significant predictors of sale price and use model checking methods to ensure that the model is well-specified and that the assumptions of linear regression are met. We will also use shrinkage methods, such as ridge regression and LASSO, to improve the accuracy and robustness of the regression model.

Our analysis will use multiple linear regression to model the relationship between the sale price of a house and various predictor variables. We will start with a simple linear regression model that predicts the sale price based on one predictor variable, such as the overall quality of the house. We will then gradually add more predictor variables to the model to create a multiple linear regression model that incorporates the most important factors influencing house prices. As the dataset comes with 80 variables, clearly there will be a lot of colinearity and unnecessary variables so selecting useful variables will be very important.

Some of the predictor variables we will consider include:

- Lot size: The size of the lot on which the house is located
- Year built: The year the house was built
- Overall condition: A rating of the overall condition of the house
- Number of bedrooms and bathrooms: The number of bedrooms and bathrooms in the house
- Neighborhood: The location of the house in one of the 25 different neighborhoods in Ames

In addition to these variables, we will also consider interactions between variables, such as the interaction between lot size and neighborhood, to capture more complex relationships.

To ensure the validity of our regression model, we will check for multicollinearity between predictor variables, which can occur when two or more predictor variables are highly correlated with each other. We will also test for heteroscedasticity, which occurs when the variance of the errors in the model is not constant across all levels of the predictor variables. To address this issue, we will use weighted least squares regression or robust regression if necessary.

Finally, we will use various model selection and shrinkage techniques, such as forward stepwise regression or ridge regression, to identify the best subset of predictor variables for our final model. We will also use cross-validation techniques to estimate the out-of-sample predictive performance of our model and compare it to other regression models, such as random forest or gradient boosting models.

Through this comprehensive regression analysis plan, we aim to identify the key predictors of house prices in Ames and develop a regression model that accurately predicts the sale price of residential homes.