

Project 2

Ames Housing Data and Kaggle Challenge

Presented by:

Boon Jun



Ames is a city in Story County, Iowa, United States approximately 30 miles north of Des Moines in central Iowa. It is best known as the home of Iowa State University, with leading Agriculture, Design, Engineering, and Veterinary Medicine colleges.

INTRODUCTION

In this project, datasets obtained from the Ames Assessor's Office (through Kaggle) are used to create a regression model that predicts the price of houses in Ames, IA.



PROBLEM STATEMENT

To build a regression model with the
lowest error
to predict Sales Price of houses sold in Ames

DATASETS

Data set contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010.

Source: <https://www.kaggle.com/c/dsi-us-6-project-2-regression-challenge/>

Train.csv

2051

Observations

81

variables

Test.csv

879

Observations

80

variables

DATASETS

Train.csv	23	21	20	17
	Ordinal	Nominal	Continuous	Discrete

For model selection & fitting

Test.csv	23	21	19	17
	Ordinal	Nominal	Continuous	Discrete

For prediction of house price to submit to Kaggle

WORKFLOW



Data Cleaning

- Null handling
- Combine/remove
- Outlier removal
- EDA



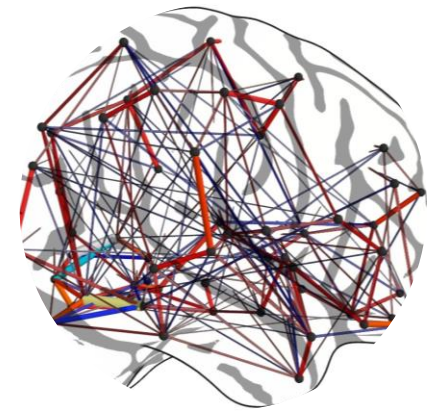
One-Hot Encoding

- Encode category variable
- Ensure same shape for Train & Test



Feature Engineering

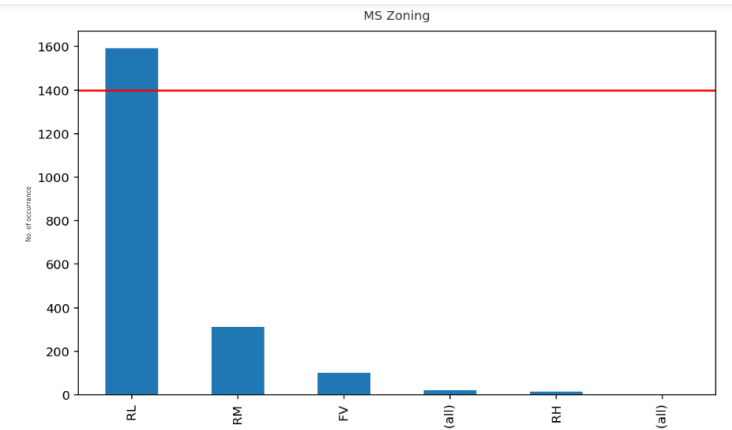
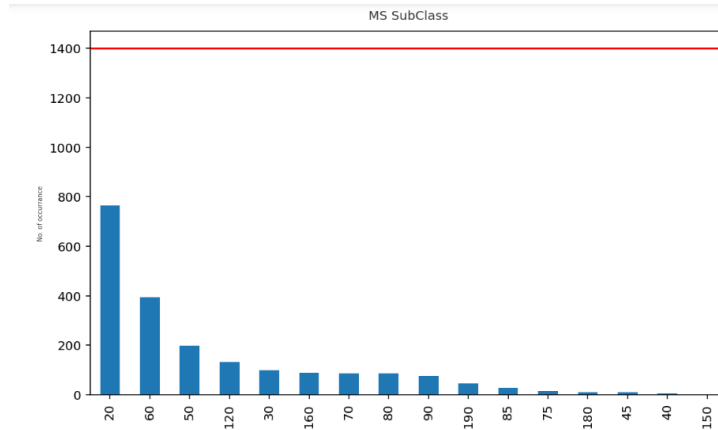
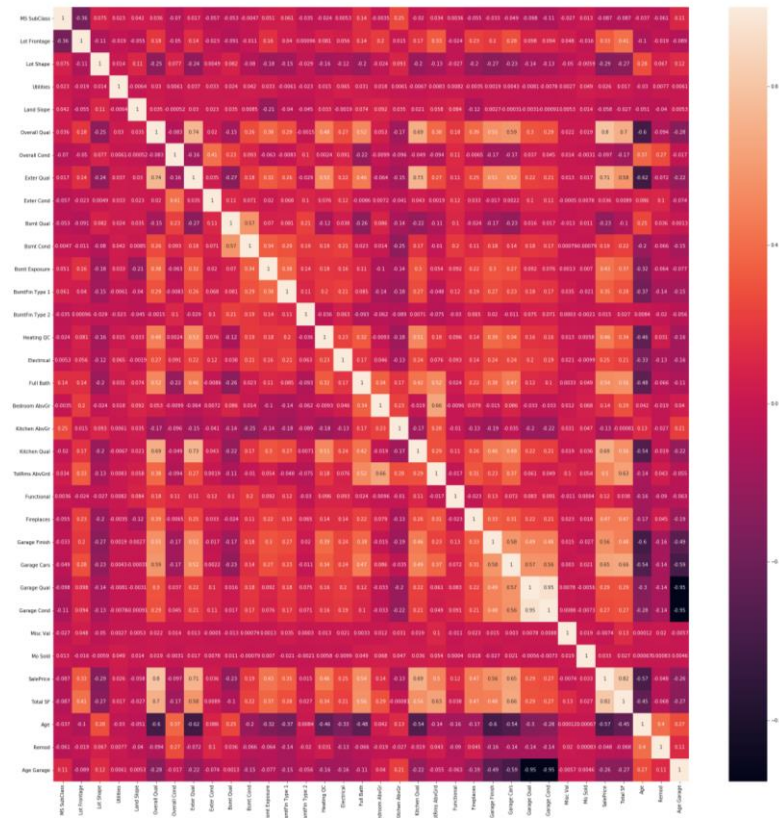
- Lasso Selection
- 30 variables



Modeling & Prediction

- 4 model: LR, Lasso, Ridge, Elastic
- Predict with LR

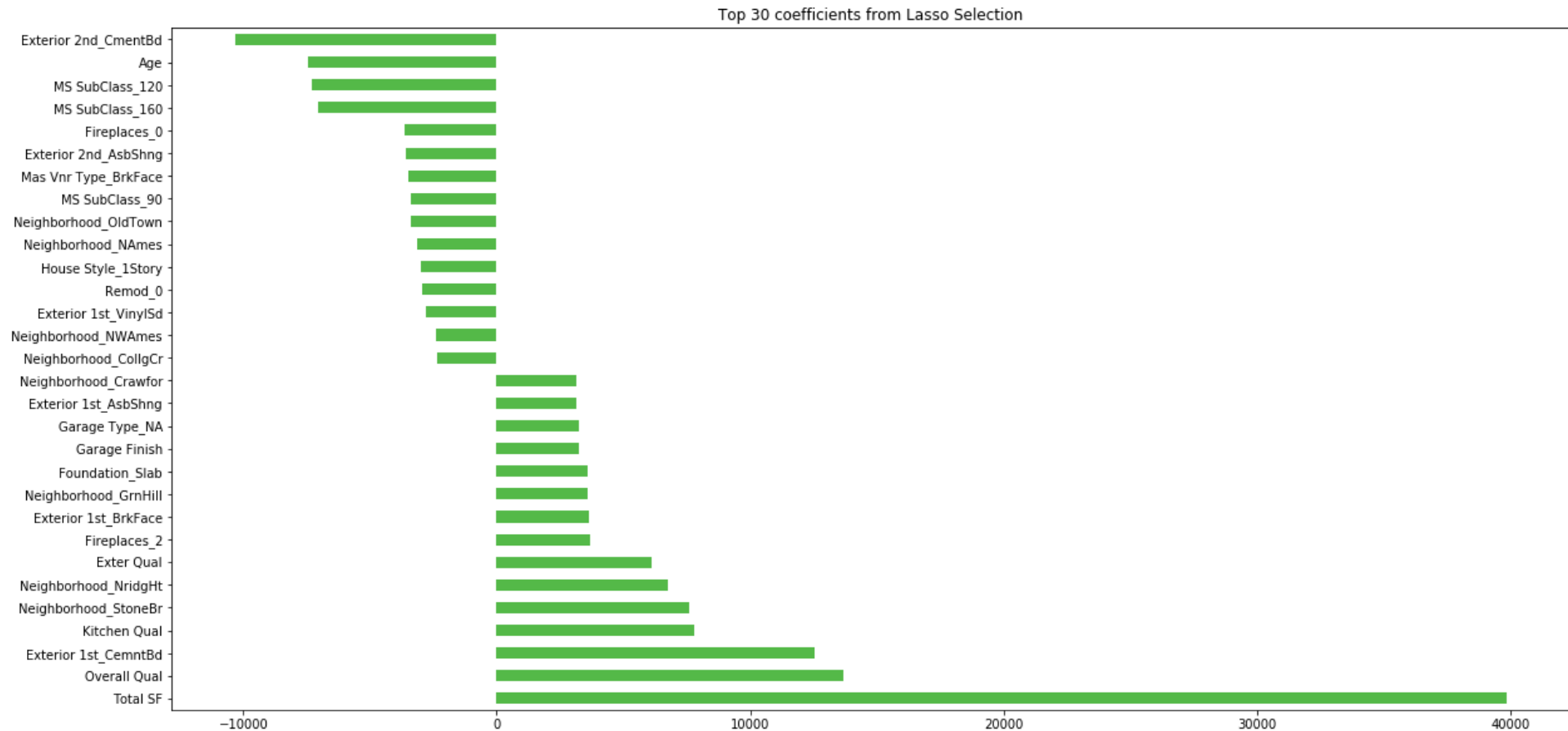




Plotting distribution heatmap to eliminate

skewed distributed variables

FEATURES ENGINEERING



Using Lasso to select top 15 +ve & -ve coefficients as final variables for model selection

MODEL SELECTION



Linear Regression

$$\text{minimize: } RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left(y_i - \left(\beta_0 + \sum_{j=1}^p \beta_j x_{ij} \right) \right)^2$$

Elastic Net

$$\text{minimize: } RSS + \text{Ridge} + \text{Lasso} = \sum_{i=1}^n \left(y_i - \left(\beta_0 + \sum_{j=1}^p \beta_j x_{ij} \right) \right)^2 + \alpha \rho \sum_{j=1}^p |\beta_j| + \alpha(1 - \rho) \sum_{j=1}^p \beta_j^2$$

Ridge

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = RSS + \lambda \sum_{j=1}^p \beta_j^2,$$

where $\lambda \geq 0$ is a *tuning parameter*, to be determined separately.

Lasso

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = RSS + \lambda \sum_{j=1}^p |\beta_j|.$$

MODEL SELECTION



1. Train/Test Split: 0.25 test size
2. Validation of model by comparing scores of 4 models

Model	R2 Score
Linear Regression	0.8852429133130981
Ridge	0.8851807633561328
Lasso	0.8852429122211832
Elastic Net	0.8727541829079618

3. Select Lasso Regression and fit X, y (Before split data)
4. Predict with test data set

PREDICTION WITH LR

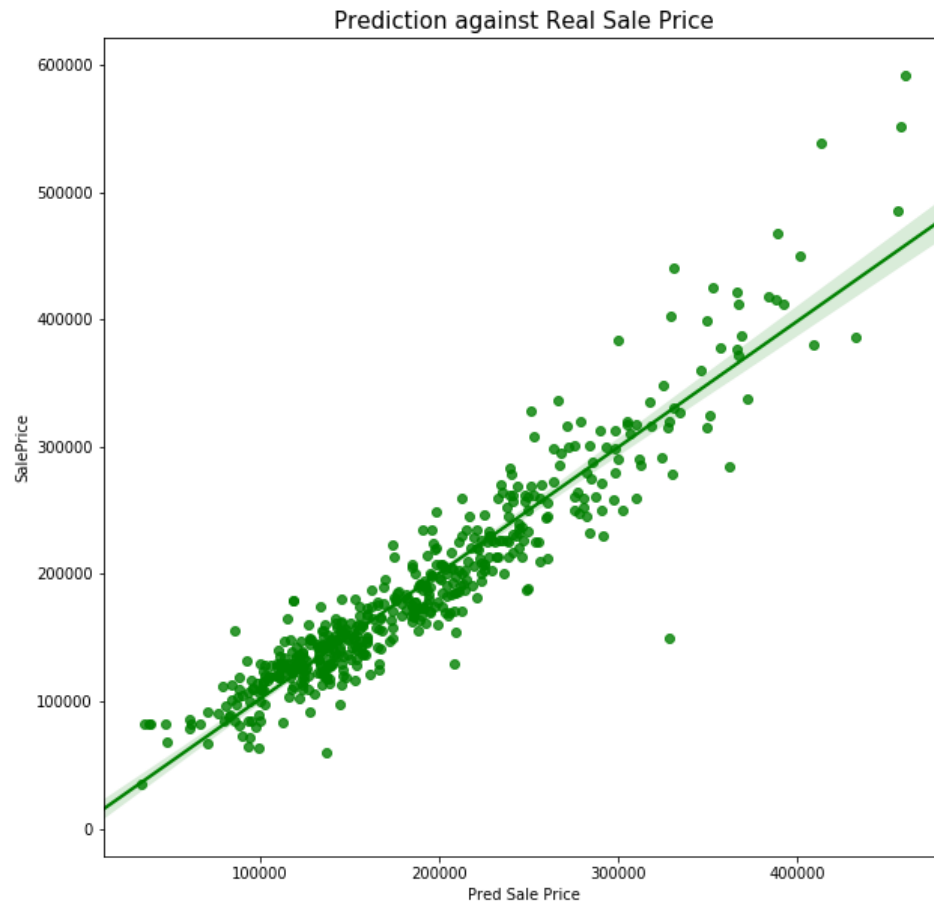


Your most recent submission				
Name	Submitted	Wait time	Execution time	Score
ames_predictions2.csv	just now	0 seconds	0 seconds	31610.23062
Complete				
Jump to your position on the leaderboard ▼				

R2 Score

0.887

SUMMARY

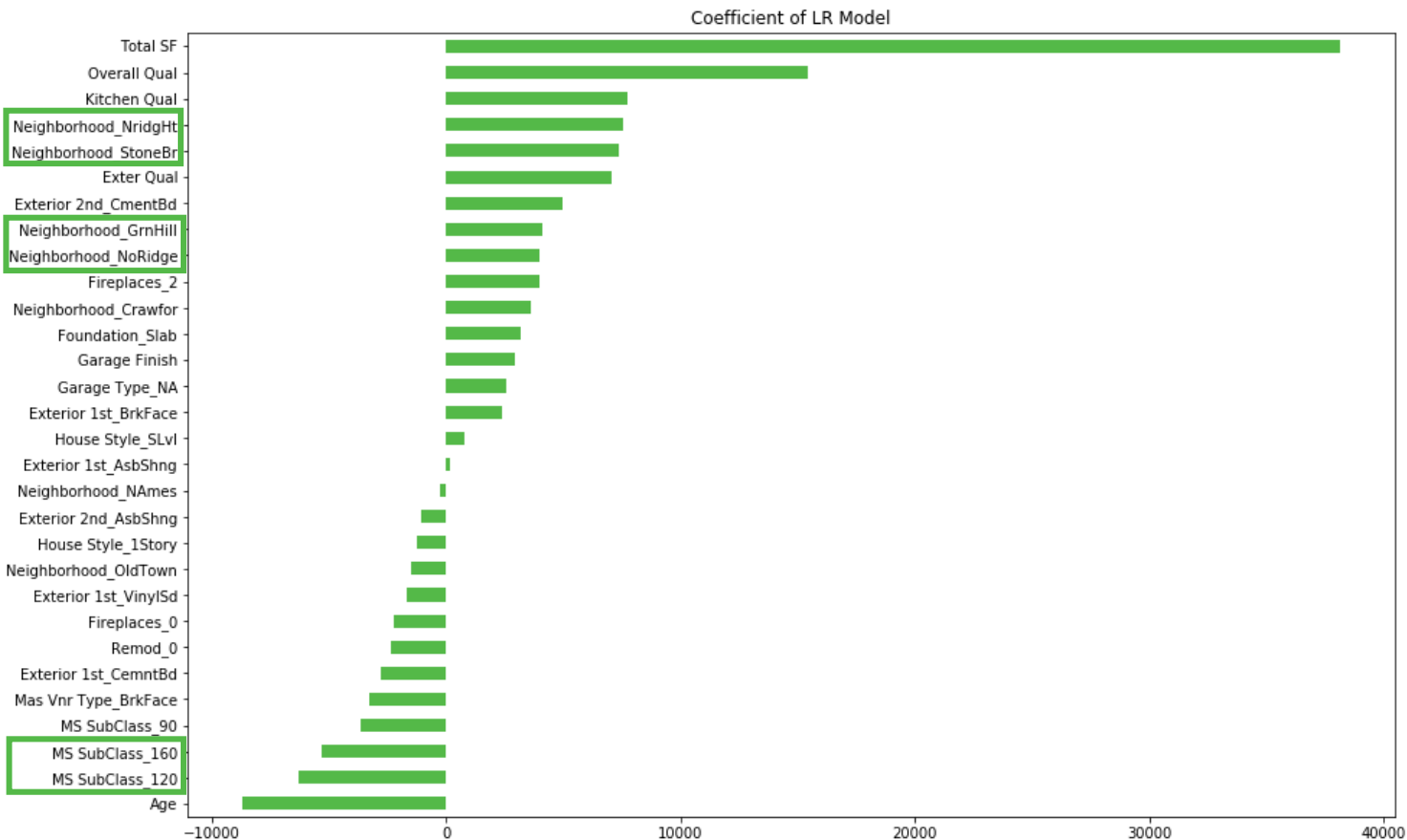


Covers 88.7% of the dataset

Limitation

- Many variables were drop because of skewed data/null values
- Model will improve if the data collection is more comprehensive
- Errors increases when predicting sale price of higher range
- External Unknown variables not included

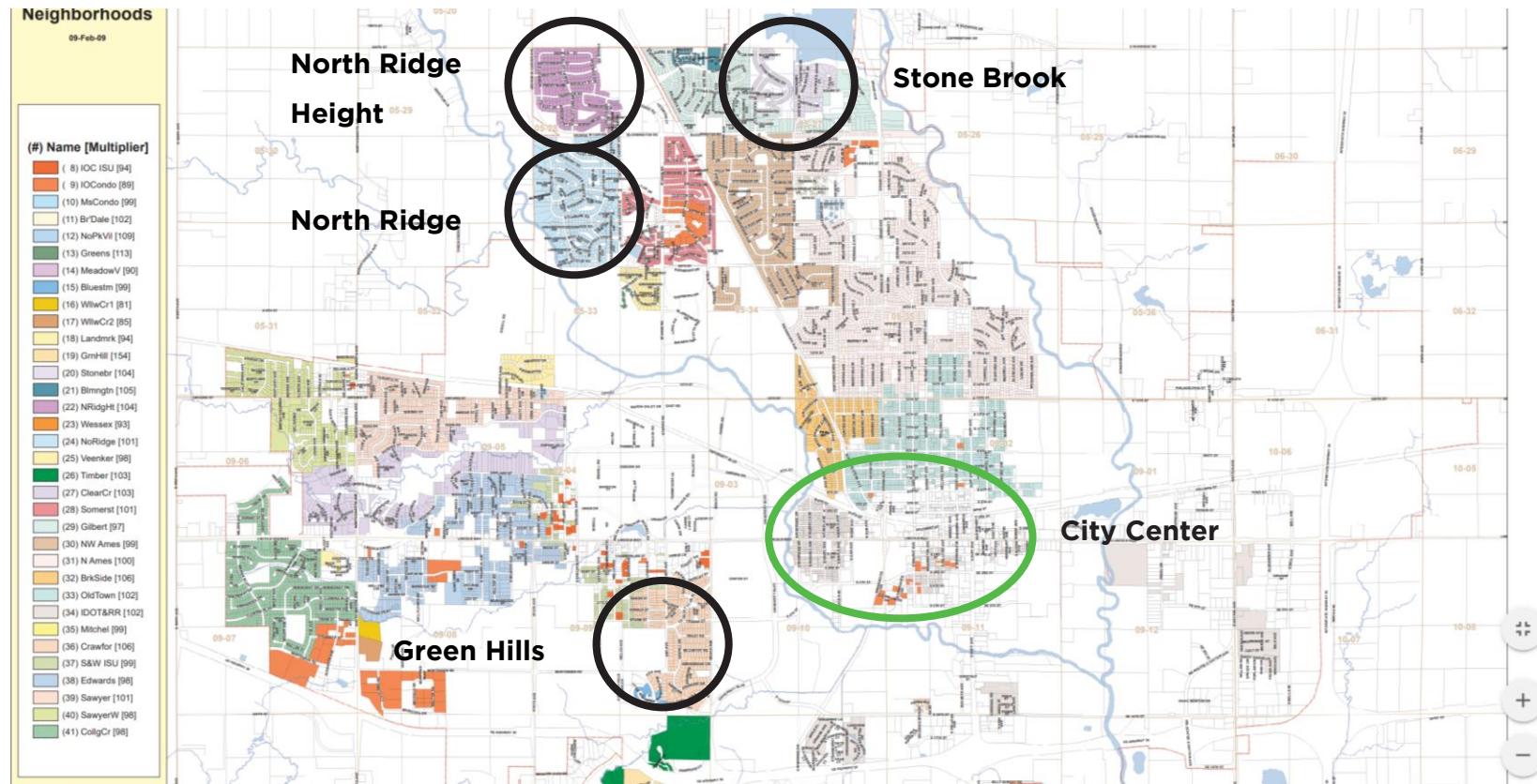
SUMMARY



- Total SF is the most significant variables that will affect the house price, followed by Overall Quality and Kitchen Quality
- Green Hills, Stone Brook, North Ridge Height, North Ridge neighborhood houses affect sale prices the most among others in Ames
- Planned Unit Development (PUD) houses will decrease the sale price (MS SubClass 160 & 120)
- Age of house has the most negative effect on Sale Price

SUMMARY

- Green Hills close to Iowa University
- North Ridge, North Ridge Height and Stone Bridge are in upper class neighborhood



THANK YOU

