Project 2

Ames Housing Data and Kaggle Challenge

Presented by: Aw Boon Jun

CITY OF Ames™

INTRODUCTION

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.

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

CITY OF ames

PROBLEM STATEMENT

To build a regression model with the

lowest error

to predict Sales Price of houses sold in Ames

CITY OF 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

CITY OF Ames

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

CITY OF Ames™

WORKFLOW

Data Cleaning

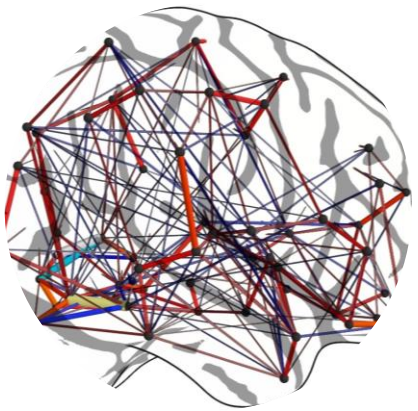
- Null handling
- Combine/remove
- Outlier removal
- EDA

One-Hot Encoding

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

Feature Engineering

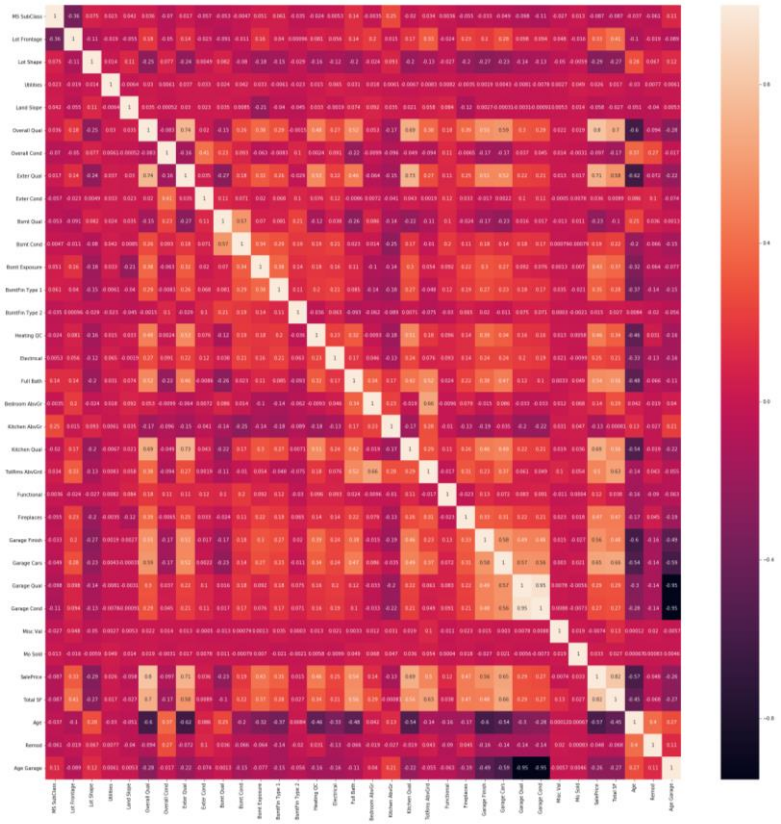
- Lasso Selection
- 30 variables



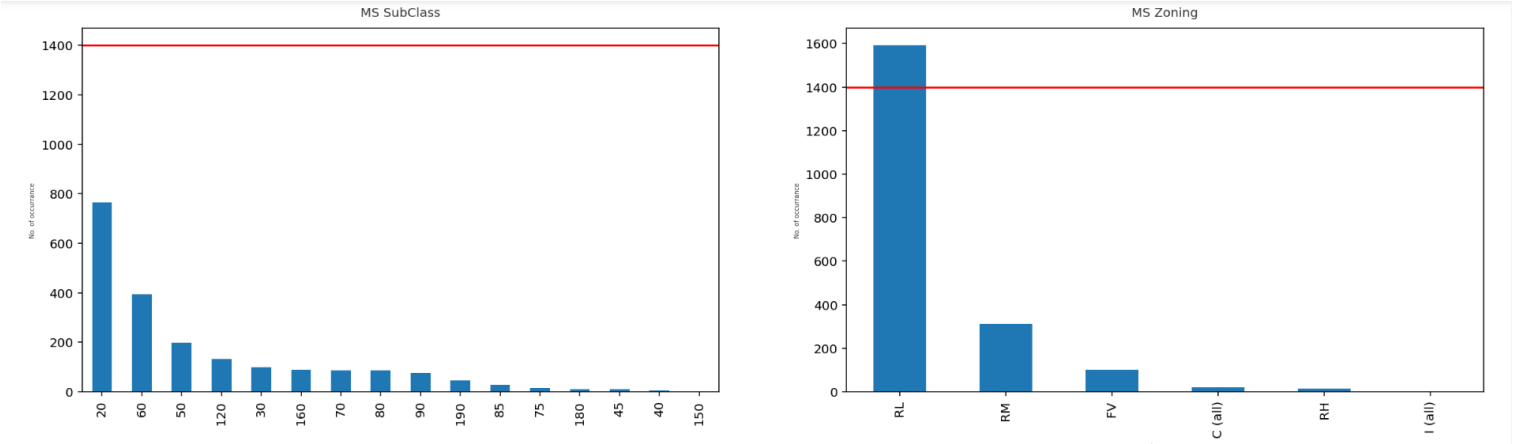
Modeling & Prediction

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

EDA



Correlation heatmap to eliminate variables



Plotting distribution histogram to eliminate skewed distributed variables

WORKFLOW

Data Cleaning

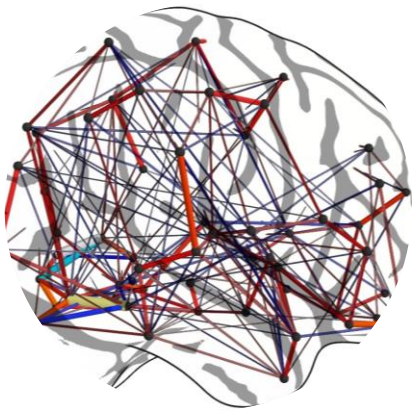
- Null handling
- Combine/remove
- Outlier removal
- EDA

One-Hot Encoding

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

Feature Engineering

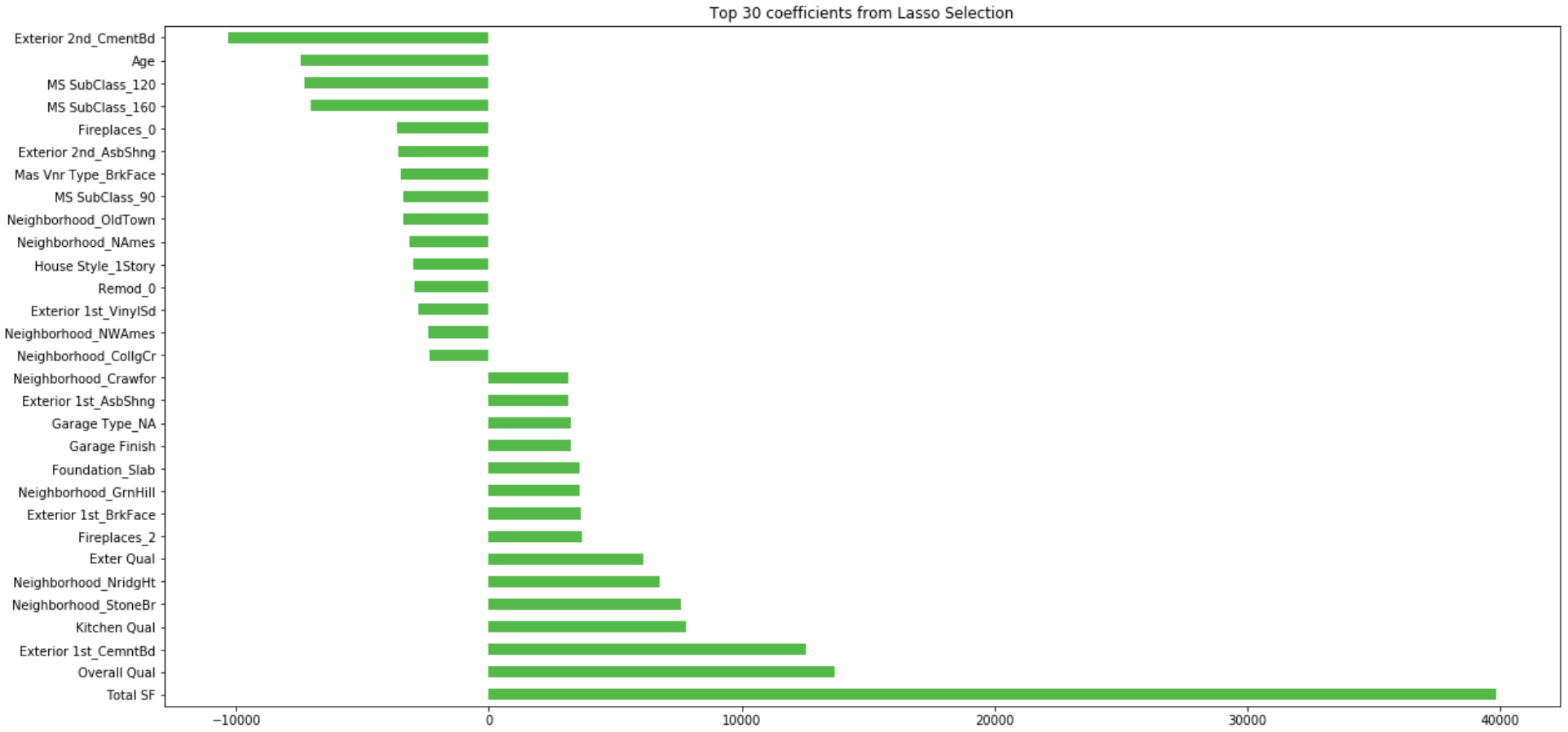
- Lasso Selection
- 30 variables



Modeling & Prediction

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

FEATURES ENGINEERING



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

WORKFLOW

Data Cleaning

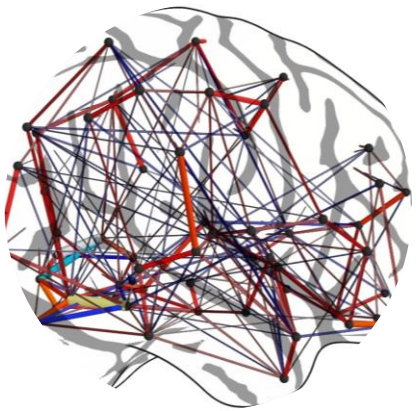
- Null handling
- Combine/remove
- Outlier removal
- EDA

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

MODEL SELECTION

Linear Regression

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

Ridge

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

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

Elastic Net

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

Lasso

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

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	**0.885242913313098 1**
Ridge	0.885180763356132 8
Lasso	0.885242912221183 2
Elastic Net	0.872754182907961 8

3. Select Linear 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
target.csv	just now	0 seconds	0 seconds	275706.49149

Complete

Jump to your position on the leaderboard

R2 Score	0.8871977269985987

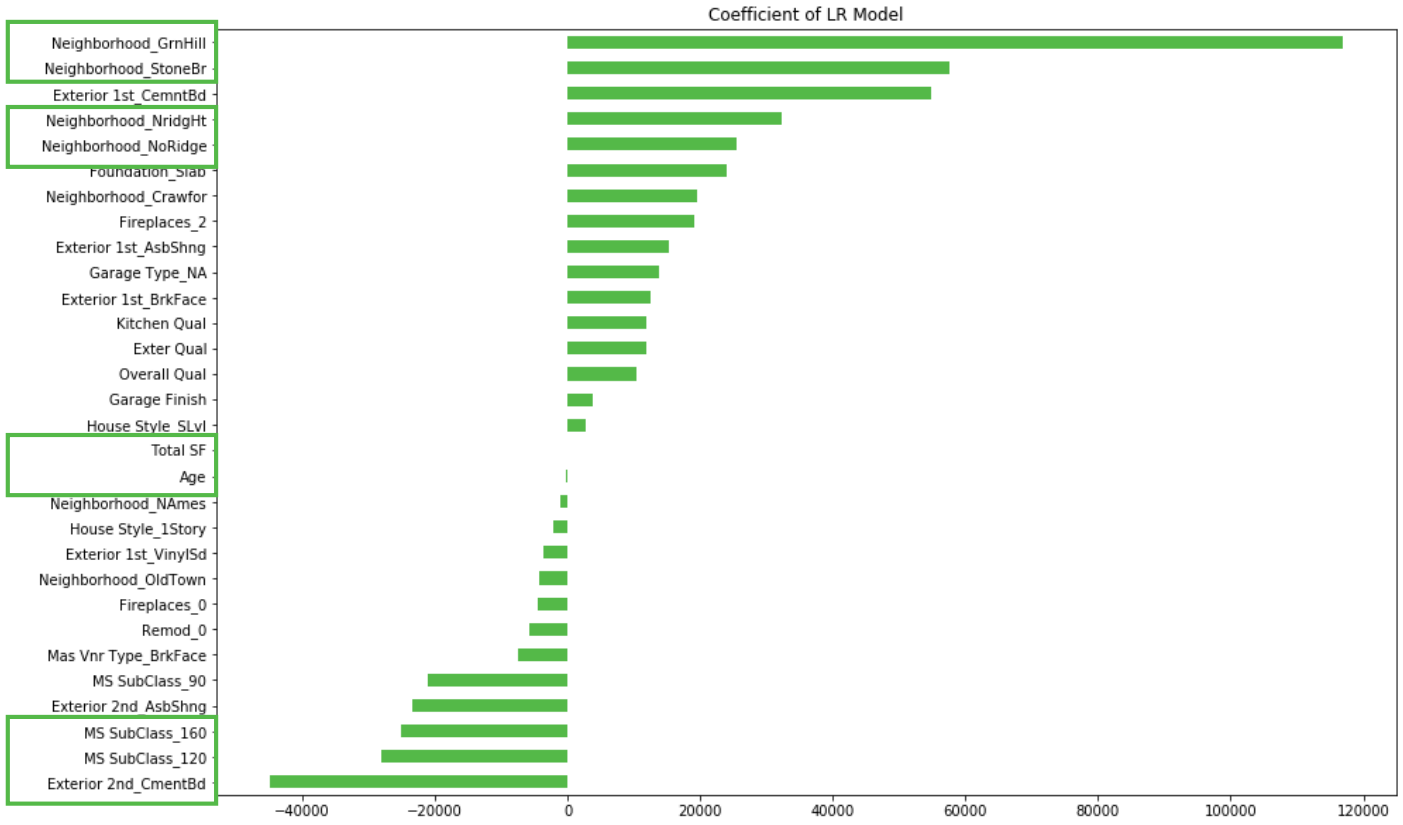
SUMMARY

Model	R2 Score
Linear Regression	**0.8852429133130981**
Lasso	0.8852429122211832

Small differences in R2 score with lasso = model is not over-fitted

Covers 88.5% of the dataset

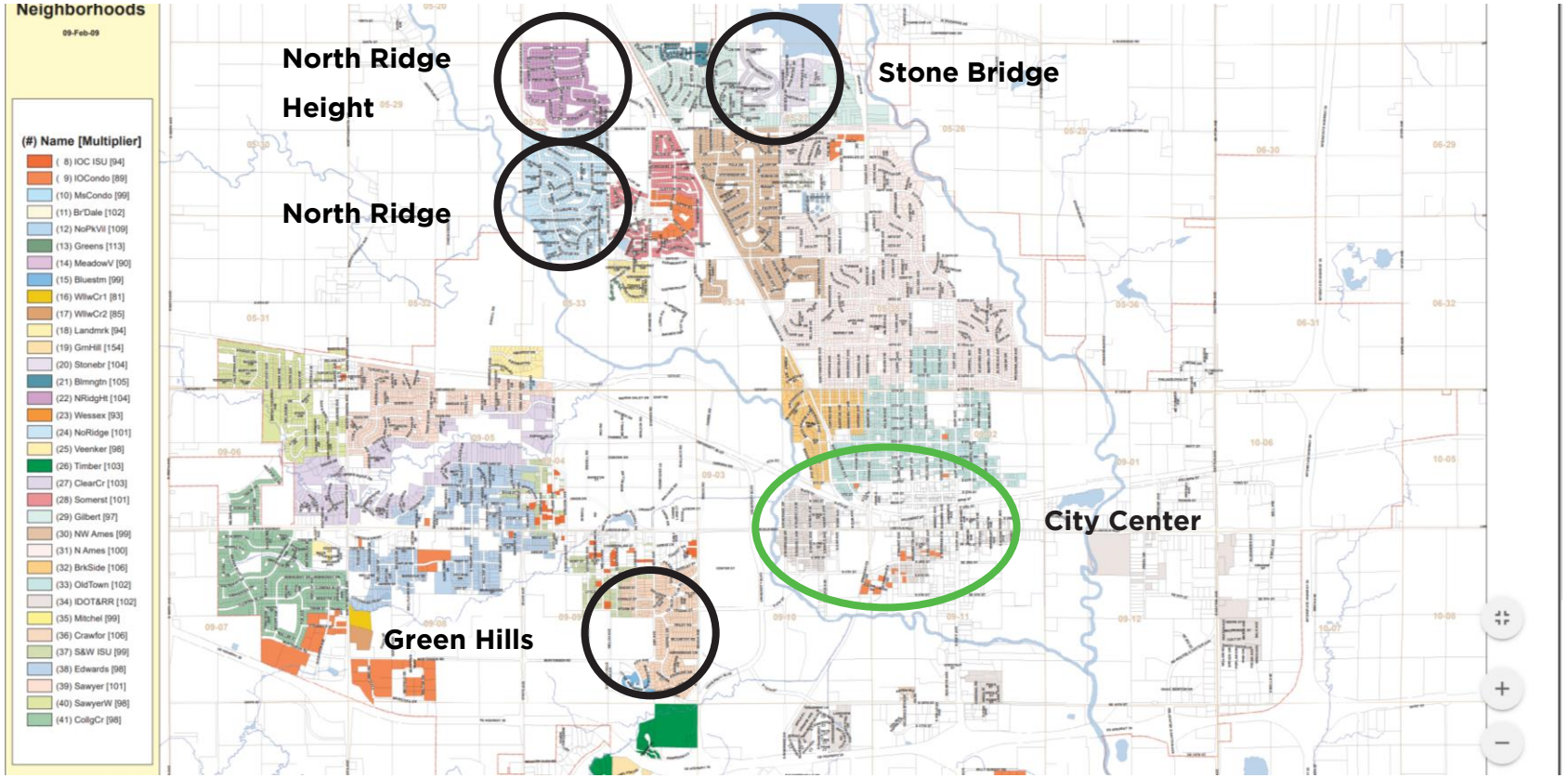
SUMMARY



- Being in the neighborhood Green Hills will increase the Sale Price by USD 16,899
- Having house exterior covered with cement board (Exterior 2nd_CmentBd) will decrease the prices by USD 44,892
- **Total Square Feet & Age of House** will not affect house sale price as much
- **Green Hills, Stone Bridge, 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)

SUMMARY

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



THANK YOU

CITY OF Ames™