**Project 1**

**By: Lizzy Sterling and Scott Herford**

Table of Contents

[Introduction 1](#_Toc505873931)

[Data Description 1](#_Toc505873932)

[Exploratory Analysis 2](#_Toc505873933)

[Analysis Question 1 2](#_Toc505873934)

[The Problem 2](#_Toc505873935)

[The Model 2](#_Toc505873936)

[Analysis Question 2 2](#_Toc505873937)

[The Problem 2](#_Toc505873938)

[The Model & Assumptions 2](#_Toc505873939)

[Comparing Competing Models 2](#_Toc505873940)

[Conclusion 2](#_Toc505873941)

[Appendix 2](#_Toc505873942)

# Introduction

“Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this Kaggle competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.” This project, investigates factors that influence sale prices of houses in Ames, Iowa. We will conduct two analyses for this project to examine the influence that neighborhood and square footage have on sale prices of homes as well as other factors that may be significant on the sale price of homes. This project could benefit realators and homebuyers in Ames that may want to know what the most influential factors are on home prices.

# Data Description

This data is from the Kaggle competition “House Prices: Advanced Regression Techniques” (see more at <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>). The data consists of 79 explanatory variables that describe elements of homes in Ames, Iowa. The training data set consists of 1460 homes with the test data set providing 1459 more observations, a total of 2919 homes.

***Note: All figures and graphs have been placed in the appendix for readability and space conservation purposes. Please refer to the*** [***appendix***](#_Figures) ***to see figures referenced in this project.***

# Exploratory Analysis

To begin this analysis, we need to consider 3 main assumptions: **Normality**, **Equality of Variance**, and **Independence**. We started to look at this by comparing our results variable, SalePrice, side-by-side with every other quantifiable variable. Looking at [figure 1](#_Exploratory_Analysis), the data seems to be very skewed and non-linear, so we tried logging SalePrice, the results of which can be found in [figure 2](#_Exploratory_Analysis_1). We are going to keep this variable logged for the remainder of our analysis.

Logging other variables

Because each model is unique, we are going to conduct further analysis as we create new models.

# Analysis Question 1

## The Problem

We want to create a model that will help buyers, real estate agents, and contractors to gain insight into what factors influence housing prices in Ames, Iowa. By doing this, we are going to perform three basic methods of model selection: **Forward**, **Backward**, and **Stepwise** selection.

## Forward Model

We began our model by including all qualitative and quantitative variables into a glmselect procedure. After running the forward selection procedure a few times, we ended up deciding that this model ([figure 3](#_Forward_Selection_7)) was the best fit. Now we are going to check the assumptions on this model in particular:

**Normality:**

**Equality of Variance:**

**Independence:**

Now that we have adjusted our model as necessary, we can add it to our table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Adjusted R2 | AIC | ASE | Kaggle Score |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

## Backward Model

We began our model by including all qualitative and quantitative variables into a glmselect procedure. After running the backward selection procedure a few times, we ended up deciding that this model ([figure 4](#_Backward_Selection)) was the best fit. Now we are going to check the assumptions on this model in particular:

**Normality:**

**Equality of Variance:**

**Independence:**

Now that we have adjusted our model as necessary, we can add it to our table:

## Stepwise Model

We began our model by including all qualitative and quantitative variables into a glmselect procedure. After running the forward selection procedure a few times, we ended up deciding that this model ([figure 5](#_Stepwise_Selection)) was the best fit. Now we are going to check the assumptions on this model in particular:

**Normality:**

**Equality of Variance:**

**Independence:**

Now that we have adjusted our model as necessary, we can add it to our table:

# Analysis Question 2

## The Problem

We would like to build the most predictive model for sale prices of homes in all of Ames, Iowa. We will use four model selection techniques to analyze the influential factors on sale prices of homes in Ames.

## The Model & Assumptions

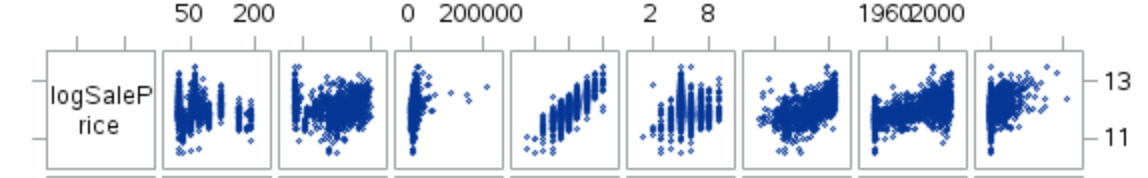
## Comparing Competing Models

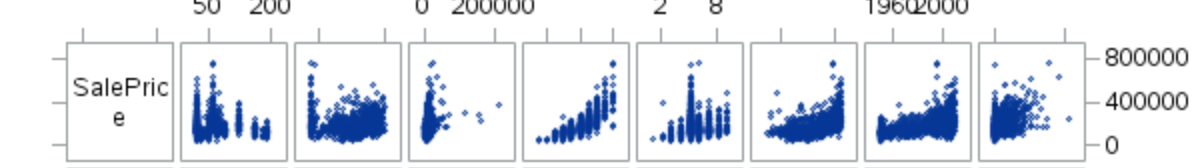
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kaggle Competition Models | Adjusted R2 | AIC | ASE | Kaggle Score |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

# Conclusion

## Appendix

### Exploratory Analysis

*Figure 1 Figure 2*



### Analysis 1

#### Forward Selection

#### Backward Selection

#### Stepwise Selection

***General Notes from Class – not going to be included in final document, just for reference:***

Scatterplots: LotArea, LotFrontage, BsmtFinSF1, BsmtFinSF2, LowQualFinSF, TotalBsmtSF

Outliers: LotArea, LotFrontage, BsmtFinSF1, GrLivArea, LowQualFinSF, TotalBsmtSF, PoolArea, MiscVal

Normality:

High Leverage: LotArea, LogSalePrice, BsmtFinSF1, GrLivArea, TotalBsmtSF

Cooks D: LotArea, TotalBsmtSF

Constant Variance: GrLivArea, LowQualFinSF, TotalBsmtSF, PoolArea, MiscVal