Using SAS with over 80 variables and 1400 rows of data this team puts together two multilinear models to predict a home’s sell price in Ames, Iowa.

Kaggle Project

Century 21 Ames: Home Sell Price Prediction Model

Jaclyn Coate

Josh Eysenbach

Andrew Leppla

Branum Stephan

Table of Contents

[1 Introduction 4](#_Toc25148027)

[2 Data Description 4](#_Toc25148031)

[3 Analysis Question 1 4](#_Toc25148036)

[3.1 Problem Statement 5](#_Toc25148037)

[3.2 Build and Fit the Model 5](#_Toc25148039)

[3.2.1 Assumptions and Corresponding Plots 5](#_Toc25148040)

[3.2.2 Comparing Competing Models 5](#_Toc25148041)

[3.3 Parameters 5](#_Toc25148042)

[3.3.1 Estimates 5](#_Toc25148043)

[3.3.2 Interpretation 5](#_Toc25148044)

[3.3.3 Confidence Intervals 5](#_Toc25148045)

[3.4 Conclusion 5](#_Toc25148046)

[3.4.1 A short summary of the analysis 5](#_Toc25148047)

[4 Analysis Question 2 5](#_Toc25148048)

[4.1 Problem Statement 5](#_Toc25148049)

[4.2 Build and Fit the Model 6](#_Toc25148050)

[4.2.1 Model Selection 6](#_Toc25148051)

[4.2.2 Assumptions and Corresponding Plots 6](#_Toc25148052)

[4.2.3 Comparing Competing Models 6](#_Toc25148053)

[4.3 Parameters 6](#_Toc25148054)

[4.3.1 Estimates 6](#_Toc25148055)

[4.3.2 Interpretation 6](#_Toc25148056)

[4.3.3 Confidence Intervals 6](#_Toc25148057)

[4.4 Conclusion 6](#_Toc25148058)

[4.4.1 A short summary of the analysis 6](#_Toc25148059)

[5 Appendix 7](#_Toc25148060)

[5.1 Analysis 1 7](#_Toc25148061)

[5.1.1 7](#_Toc25148062)

[5.1.2 7](#_Toc25148063)

[5.1.3 7](#_Toc25148064)

[5.1.4 7](#_Toc25148065)

[5.1.5 7](#_Toc25148066)

[5.1.6 7](#_Toc25148067)

[5.1.7 7](#_Toc25148068)

[5.1.8 7](#_Toc25148069)

[5.1.9 7](#_Toc25148070)

[5.1.10 7](#_Toc25148071)

[5.1.11 7](#_Toc25148072)

[5.1.12 7](#_Toc25148073)

[5.1.13 7](#_Toc25148074)

[5.1.14 7](#_Toc25148075)

[5.1.15 7](#_Toc25148076)

[5.1.16 7](#_Toc25148077)

[5.1.17 7](#_Toc25148078)

[5.1.18 7](#_Toc25148079)

[5.1.19 7](#_Toc25148080)

[5.1.20 7](#_Toc25148081)

[5.2 Analysis 2 7](#_Toc25148082)

[5.2.1 7](#_Toc25148083)

[5.2.2 7](#_Toc25148084)

[5.2.3 7](#_Toc25148085)

[5.2.4 7](#_Toc25148086)

[5.2.5 7](#_Toc25148087)

[5.2.6 7](#_Toc25148088)

[5.2.7 7](#_Toc25148089)

[5.2.8 7](#_Toc25148090)

[5.2.9 7](#_Toc25148091)

[5.2.10 7](#_Toc25148092)

[5.2.11 7](#_Toc25148093)

[5.2.12 7](#_Toc25148094)

[5.2.13 7](#_Toc25148095)

[5.2.14 7](#_Toc25148096)

[5.2.15 7](#_Toc25148097)

[5.2.16 8](#_Toc25148098)

[5.2.17 8](#_Toc25148099)

[5.2.18 8](#_Toc25148100)

[5.2.19 8](#_Toc25148101)

[5.2.20 . 8](#_Toc25148102)

# Introduction

## When someone starts the processes of buying a home it is usually filled with daydreams of must haves and a few needs. This investigation is going to examine [through multilinear regression] the different components that contribute to the pricing of house buying and what Century 21 Ames can do to predict the pricing for home buyers in Ames, Iowa.

## With conservative methods our first analysis will provide a house sell price prediction model based solely on square footage and sell prices from only active Century 21 Ames neighborhoods [NAmes, Edwards, BrkSide]. With this initial model Century 21 Ames will be armed with a model to help them predict the sell prices of homes in the neighborhoods in which they are actively working.

## Using four different model selections we will be completing a second analysis that will build a predictor model for the selling price of house across all of Ames, Iowa. With this model analysis, Century 21 Ames, will have a strong predictor of all variables that highly affect sell prices across all the neighborhoods in Ames, Iowa; effectively empowering them to expand from their three active neighborhoods to much more.

# Data Description

## The data in this evaluation contains over 1400 rows and 80 different variables that could contribute to the selling price of a home in Ames, Iowa. The data was collected across 25 different neighborhoods from houses that have been built between 1872 – 2010.

## In our first analysis we will be using the following variables to predict sell price: *GrLivArea* [living area square footage], *SalesPrice* [sales prices of homes in neighborhoods], and *Neighborhoods* [NAmes, Edwards, BrkSide].

## In our second analysis we will be building a predictive model for sales prices of all the homes in Ames, Iowa. This will include the follow variables: XXX [] XXX [] XXX [] XXX [] XXX [] XXX [] XXX []XXX []XXX [].

## To find out more about this data and the definitions of all individual variables you can visit the Kaggle competition website [here](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data) [https://www.kaggle.com/c/house-price-advanced-regression-techniques/data].

# Analysis Question 1

## Problem Statement

## Century 21 Ames wants an estimate of how the sale price of a house is related to the square footage of the living area of the home in their three active neighborhoods: NAmes, Edwards, and BrkSide.

## Build and Fit the Model

### Assumptions and Corresponding Plots

#### Linearity

#### Met with the original data set. We see in Figure 5.1.1 that judging by the scatter plot alone the data does reasonably fall along a straight (nonhorizontal) line and nearly passes through the origin.

#### Met with the log-log transformed data set. We see in Figure 5.1.2 that judging by the scatter plot alone the data does reasonably fall along a straight (nonhorizontal) line and nearly passes through the origin.

#### Normality

#### Not met with the original data judging from the histogram and QQ-plots Figure 5.1.3 and 5.1.4, there is evidence against normality of sale price on fixed values of square footage and neighborhood.

#### Judging from the histogram and QQ-plots Figure 5.1.5 and 5.1.6, once the log-log transformation has taken place, there is no evidence against normality of sale price on fixed values of square footage and neighborhood.

#### Equal variances

#### Judging from the residuals scatter plot Figure 5.1.7, once the log-log transformation has taken place, there is no evidence against constant variance of sale price against constant variance of square footage and neighborhood.

#### Independence

#### We will assume the sales prices of homes gathered are independent.

#### Outliers: Influential point analysis (Cook’s D and Leverage)

#### There was an outlier identified with Leverage in Figure 5.1.8 and Cook’s D in Figure 5.1.9 that is showing a ‘partial sale’. As this is not a completed sale thus the dollars being spent not actualized. We will remove this observation from our analysis.

#### Multicollinearity

#### Note that multicollinearity is resolved with a VIFS < 1 [Table 5.1.10]

### Comparing Competing Models

#### Adj R2

#### Internal CV Press

## 

## Parameters

### Estimates

### Interpretation

### Confidence Intervals

## 

## Conclusion

### A short summary of the analysis

# Analysis Question 2

## Problem Statement

## Build and Fit the Model

### Model Selection

#### Stepwise

#### Forward

#### Backward

#### Custom

### Assumptions and Corresponding Plots

#### Linearity

#### Normality

#### Equal variances

#### Independence

#### Outliers: Influential point analysis (Cook’s D and Leverage)

#### Multicollinearity.

### Comparing Competing Models

#### Adj R2

#### Internal CV Press

#### Kaggle Score.

## Parameters

### Estimates

### Interpretation

### Confidence Intervals

## 

## Conclusion

### A short summary of the analysis

# Appendix

## Analysis 1

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

## Analysis 2

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### .