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 7](#_Toc25664966)

[2 Data Description 7](#_Toc25664970)

[3 Analysis Question 1 7](#_Toc25664975)

[3.1 Problem Statement 8](#_Toc25664976)

[3.2 Build and Fit the Model 8](#_Toc25664978)

[3.2.1 Assumptions and Corresponding Plots 8](#_Toc25664979)

[3.2.2 Comparing Competing Models 9](#_Toc25664980)

[3.3 Parameters 9](#_Toc25664981)

[3.3.1 Estimates 9](#_Toc25664982)

[3.3.2 Interpretation 9](#_Toc25664983)

[3.3.3 Confidence Intervals 9](#_Toc25664984)

[3.4 Conclusion 9](#_Toc25664985)

[3.4.1 A short summary of the analysis 9](#_Toc25664986)

[4 Analysis Question 2 9](#_Toc25664987)

[4.1 Problem Statement 9](#_Toc25664988)

[4.2 Build and Fit the Model 9](#_Toc25664989)

[4.2.1 Model Selection 9](#_Toc25664990)

[4.2.2 Assumptions and Corresponding Plots 9](#_Toc25664991)

[4.2.3 Comparing Competing Models 9](#_Toc25664992)

[4.3 Parameters 10](#_Toc25664993)

[4.3.1 Estimates 10](#_Toc25664994)

[4.3.2 Interpretation 10](#_Toc25664995)

[4.3.3 Confidence Intervals 10](#_Toc25664996)

[4.4 Conclusion 10](#_Toc25664997)

[4.4.1 A short summary of the analysis 10](#_Toc25664998)

[5 Appendix 11](#_Toc25664999)

[5.1 Analysis 1 11](#_Toc25665000)

[5.1.1 11](#_Toc25665001)

[5.1.2 11](#_Toc25665002)

[5.1.3 12](#_Toc25665003)

[5.1.4 12](#_Toc25665004)

[5.1.5 12](#_Toc25665005)

[5.1.6 13](#_Toc25665006)

[5.1.7 13](#_Toc25665007)

[5.1.8 13](#_Toc25665008)

[5.1.9 14](#_Toc25665009)

[5.1.10 14](#_Toc25665010)

[5.1.11 14](#_Toc25665011)

[5.1.12 14](#_Toc25665012)

[5.1.13 14](#_Toc25665013)

[5.1.14 14](#_Toc25665014)

[5.1.15 14](#_Toc25665015)

[5.1.16 14](#_Toc25665016)

[5.1.17 14](#_Toc25665017)

[5.1.18 14](#_Toc25665018)

[5.1.19 14](#_Toc25665019)

[5.1.20 14](#_Toc25665020)

[5.2 Analysis 2 14](#_Toc25665021)

[5.2.1 14](#_Toc25665022)

[5.2.2 14](#_Toc25665023)

[5.2.3 14](#_Toc25665024)

[5.2.4 14](#_Toc25665025)

[5.2.5 14](#_Toc25665026)

[5.2.6 14](#_Toc25665027)

[5.2.7 14](#_Toc25665028)

[5.2.8 14](#_Toc25665029)

[5.2.9 15](#_Toc25665030)

[5.2.10 15](#_Toc25665031)

[5.2.11 15](#_Toc25665032)

[5.2.12 15](#_Toc25665033)

[5.2.13 15](#_Toc25665034)

[5.2.14 15](#_Toc25665035)

[5.2.15 15](#_Toc25665036)

[5.2.16 15](#_Toc25665037)

[5.2.17 15](#_Toc25665038)

[5.2.18 15](#_Toc25665039)

[5.2.19 15](#_Toc25665040)

[5.2.20 15](#_Toc25665041)

[6 Codebook 15](#_Toc25665042)

[6.1 15](#_Toc25665043)

# 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 the data does reasonably fall along a straight (nonhorizontal) line and nearly passes through the origin. However, we do see evidence of three influential outliers in the data.

#### Outliers Influential point analysis (Studentized Residual, Cook’s D, and Leverage)

#### In Figure 5.1.2 we see outliers in our Studentized Residual Plot.

#### There are three outliers identified with Leverage in Figure 5.1.3 and Cook’s D in Figure 5.1.4. In reviewing the data these points are specifically identified as a ‘partial’ and ‘abnormal’ sales. They are featured houses with overly large extra living areas like garages and basements. We will remove these three observations from our analysis in order to create a more statistically sound analysis.

#### Linearity

#### Met with the original data post outlier drop set we see in Figure 5.1.5 the data does reasonably fall along a straight (nonhorizontal) line and nearly passes through the origin.

#### Normality

#### Not met with the original data after outlier drop, judging from the histogram and QQ-plots Figure 5.1.6 and 5.1.7, there is evidence against normality of sale price on fixed values of square footage and neighborhood. In order to move forward we perform a linear-log transformation on the data.

#### Judging from the histogram and QQ-plots Figure 5.1.8 and 5.1.9, once the log-linear 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 5.1.10, once the log-linear transformation has taken place, the residual plot resembles a random scatter of points around the 0 line. Therefore, there is no evidence against constant variance of sale price against constant variance of square footage and neighborhood.

#### Independence

#### Homes within the same neighborhoods are often linked in sale price. This is due to the fact that when selling/buying homes real-estate agents references the ‘comps’ (similar homes of living square footage and price) within a neighborhood in order to determine a market fair price for the home. This irrefutably means that homes within a neighborhood are not independent of each other therefore violating the assumption of independence.

#### We would stop here but likely fail this paper. Therefore, we will move forward with caution in this statistical linear modeling of home prices versus living square footage to compete in this Kaggle competition.

#### Multicollinearity

#### Note that multicollinearity is resolved with a Tolerance < 1 [Figure 5.1.11]

### Comparing Competing Models

#### Adj R2

#### Internal CV Press

## 

## Parameters

### Estimates

### 

### Interpretation

#### β0: The intercept in this model provides an estimate (11.44) of the sale price of the NAmes neighborhood (reference neighborhood) with a GrLivArea of zero. Of course, this is an extrapolation and does not have a clear, practical meaning.

#### β1: For each 100 (1\*100) increments increase in GrLiveArea of houses of the NAmes neighborhood, is associated with a multiplicative change in median sale price of [e^.0003 = 1 = 100\*1 =] $100 dollars.

#### β2: This is an adjustment of the intercept for the BrkSide neighborhood with respect to the NAmes neighborhood. For GrLivArea of zero, the BrkSide neighborhood has an estimated median sale price -.6517 dollars less than the NAmes neighborhood.

#### β3: This is an adjustment of the intercept for the Edwards neighborhood with respect to the NAmes neighborhood. For GrLivArea of zero, the Edwards neighborhood has an estimated median sale price -.4179 dollars less than the NAmes neighborhood.

#### β4: For each 100 (1\*100) increments increase in GrLivArea of houses in the BrkSide neighborhood, is associated with a multiplicative change in median sale price of [e^.0004 = 1 = 100\*1 =] $100 dollars from the change to the NAmes neighborhood.

#### β5: For each 100 (1\*100) increments increase in GrLivArea of houses in the Edwards neighborhood, is associated with a multiplicative change in median sale price of [e^.0002 = 1 = 100\*1 =] $100 dollars from the change to the NAmes neighborhood.

### Confidence Intervals

### .

### .

## 

## Conclusion

Given that the model looks to fit (see assumptions covered in section 3.2.1), we now move to interpret the parameter estimates found in section 3.3.1. All GrLivArea, BrkSide, and Edwards were found to be significant when neighborhood NAmes was a reference.

A one unit increase in square footage is associated with a $1 increase in median sale price. Differently stated, a 100 square footage increase in GrLivArea results in a median sale price change of $100.

A 95% confidence interval for the multiplicative increase GrLivArea is (e.0002, e.0003), (1,1) = 100%?

It is estimated that the GrLivArea (square footage) explains about 52.7% of the variation in sale price for house among the NAmes, BrkSide, and Edwards neighborhoods.

As suggestive as the findings may be, inferences that go beyond these data are unwise. The data were summarized from available house sales and cannot be representative of any wider population. As usual, no causal interpretation can be made from these observational data.

# 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

### A screenshot of a map Description automatically generated

### A screenshot of a cell phone Description automatically generated

### A close up of a map Description automatically generated

### A screenshot of a cell phone Description automatically generated

### A close up of a map Description automatically generated

### A close up of a logo Description automatically generated

### A close up of a map Description automatically generated

### A close up of a device Description automatically generated

### A close up of a map Description automatically generated

### A screenshot of a cell phone Description automatically generated

### A screenshot of a cell phone Description automatically generated

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

## Analysis 2

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

# Codebook

## 