# Introduction

There are many factors that determine the value of a home in the real estate market. We have been provided a dataset containing explanatory variables and the final sale prices of homes. We will attempt to predict the final sales price of a home for Ames, Iowa using regression.

# Data Description

The data was provided from the following educational kaggle competition:

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

The data consists of 79 explanatory variables and the final Sales Price of a home in Ames, Iowa. The [Ames Housing dataset](http://www.amstat.org/publications/jse/v19n3/decock.pdf) was compiled by Dean De Cock for use in data science education.

For analysis question 1, we are only interested in the following variables:

GrLIvArea

SalesPrice

Neighborhoods (NWAmes, Edwards and BrkSide specifically)

For analysis question 2, we are interest in.

???????????????

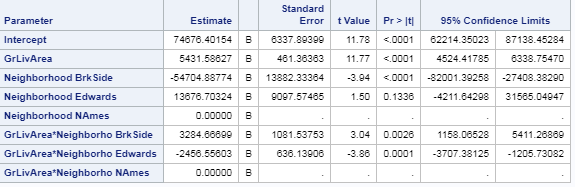
# Analysis Question 1:

## Restatement of Problem

Century 21 Ames only sells houses in the NAmes, Edwards and BrkSide neighborhoods and would like to simply get an estimate of how the SalePrice of the house is related to the square footage of the living area of the house (GrLIvArea) and if the SalesPrice (and its relationship to square footage) depends on which neighborhood the house is located in.

## Build and Fit the Model

**The following model was built using glm regression (code referenced in Appendix).**



|  |  |
| --- | --- |
|  | Note that the Edwards neighborhood regression line seems to be heavily influenced by two points with a GrLivArea greater than 30. |

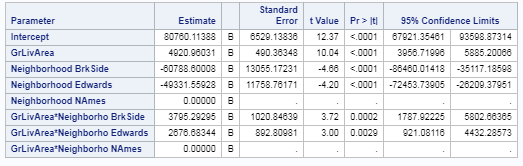
## Checking Assumptions

|  |  |
| --- | --- |
|  | **Normality:** The data appears to be normal given the bell shaped distribution in the histogram highlighted in the red box of illustration. We also see evidence of normality in the qqplot in the red box as well.  **Linear Trend:** The pairwise scatter plot in orange box indicates a strong linear trend.  **Equal SD:** There is little evidence from the scatter plots of heteroscedasticity and the Residual plot in the blue box shows a nice random cloud.  **Independence:**  We will assume the observations are independent.  **Influential point analysis:**  We see that we have two points circled in green that have high leverage over the rest of the data. I believe these are the points above 30 in GrLivArea.  Also, we see that the Cook’s D has a very high point, circled in purple, at 2.5 that is concerning. |

## Comparing Competing Models

An R-squared of .4785 was obtained as well as an adjusted R-squared of .4714.

## Parameters



Estimates

Interpretation

Confidence Intervals

## Conclusion

A short summary of the analysis.

# Analysis Question 2

## Restatement of Problem

## Model Selection

Type of Selection

Stepwise

Forward

Backward

Custom

## Checking Assumptions

Residual Plots

Influential point analysis (Cook’s D and Leverage)

Make sure to address each assumption

## Comparing Competing Models

Adj R2

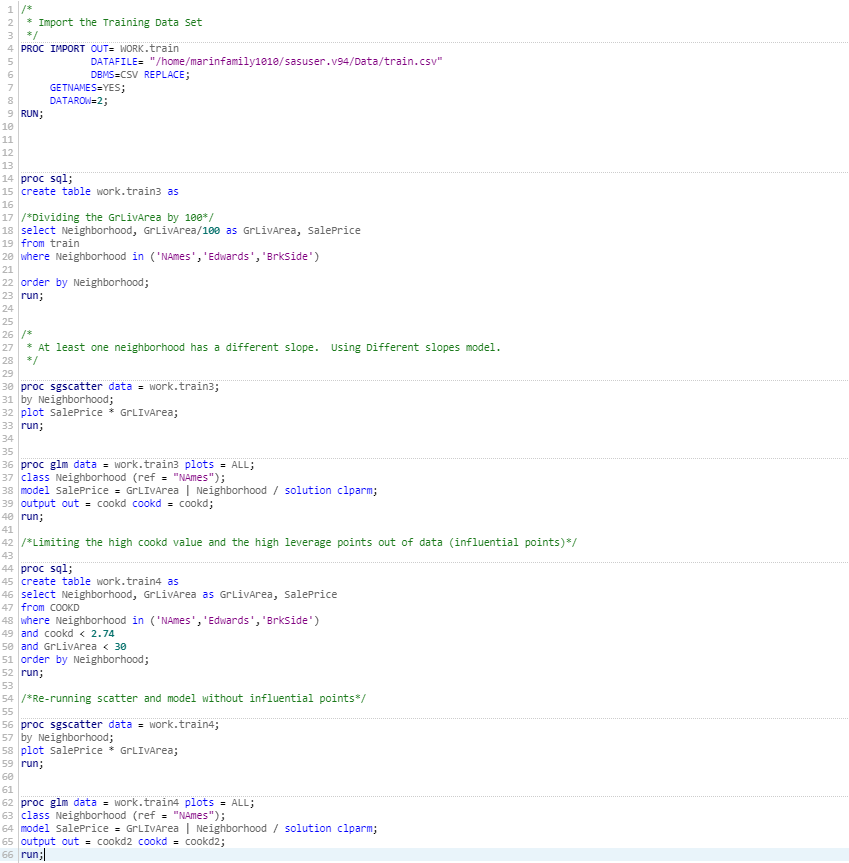
Internal CV Press

Kaggle Score

## Conclusion: A short summary of the analysis.

# Appendix

## Question 1 Code:



## Question 2 Code: