**Report For:**

**Century 21**

**Sales Prediction Report**

**August 11, 2018**

Report By:

Quinton Nixon

Brock Friedrich

Grant Bourzikas

# Table of Contents

[Table of Contents 1](#_Toc521358711)

[Introduction 1](#_Toc521358712)

[Data Description 1](#_Toc521358713)

[Analysis Question 1: 1](#_Toc521358714)

[Restatement of Problem 1](#_Toc521358715)

[Build and Fit the Model 1](#_Toc521358716)

[Checking Assumptions 1](#_Toc521358717)

[Comparing Competing Models 1](#_Toc521358718)

[Conclusion 1](#_Toc521358719)

[Analysis Question 2 1](#_Toc521358720)

[Restatement of Problem 1](#_Toc521358721)

[Model Selection 1](#_Toc521358722)

[Checking Assumptions 1](#_Toc521358723)

[Comparing Competing Models 1](#_Toc521358724)

[Conclusion: A short summary of the analysis. 1](#_Toc521358725)

[Appendix 1](#_Toc521358726)

# 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. However, it is essential to review the data because it proves that there are many other influences in price negotiations than the number of bedrooms or a white-picket fence.

The purpose of this report is to review the dataset of residential homes in Ames Iowa and predict the final price of each home.

# Data Description

The Ames House dataset was compiled by Dean De Cock and contains 79 explanatory variables describing almost every aspect of residual home in Ames Iowa from 2006 to 2010.

The data set contains 2930 observations involved in assessing home values.

# Analysis Question 1:

## Restatement of Problem

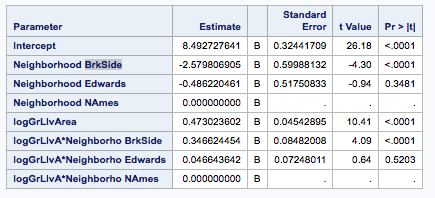
Century 21 has commissioned Nixon, Friedrich, and Bourzikas to perform a study to assist them in answering some critical questions they have regarding homes prices in Ames. Since they only sell homes in the “NAmes”, “Edwards”, and “BrkSide” neighborhoods, they would like to get an estimate of how the Sales Price of the house is related to the square footage of the living area of the house. Additionally, they would like to understand the relationship if the Sales Price in relation to the square footage depends on which neighborhood the house is located.

## Build and Fit the Model

In order to build and fit a model, an analysis is performed to determine when one of the predictor variables impacts how another predictor variable is related to the dependent variable. A multiple linear regression model in which the mean of the LogLivingArea depends linearly on the important of the LogSalesPrice and all three Neighborhoods, allowing for different slopes and intercepts, is as follow:

* Ames^SalesPrice = B0 + B1\*BrkSide + B2\*Edwards + B3\*NAmes + B4(LogLivingArea\*BrkSide) + B5(LogLivingArea\*Edwards)
* Ames^SalesPrice = 8.49 + (-2.58\*BrkSide) + (-0.49\*Edwards) + (0.0 \* NAmes) + (B3\*0.0) + B4(0.47\*BrkSide) + B5(0.47\*Edwards)

An Intercept and slopes for the Ames dataset are not significantly different (p values = <0.001, <0.001, respectively)



As intercept and slopes for the Neighborhood BrkSide is not significantly different (P values < 0.001, but the Edwards neighborhood is significantly different (P values 0.35)

*Note to Grant and Brock: I don’t like to create extra work, but I think question 1 is a bit of a trick question in that we have to answer two questions before addressing the business question. What I mean is that we might need to do two six step analyses before providing our predictive model: A) does sale price depend on living area and B) does sale price depend on location. Part B is easy because we have three discreet neighborhoods. For Part A, we’d need to “bucket” houses into groups bases on size (perhaps small, medium, and large). Once we can say definitively that the average sale price is different in different neighborhoods and different for different sized homes, we’ll have a more compelling argument that our model for predicting price is comprehensive and correct.*

I could be wrong.

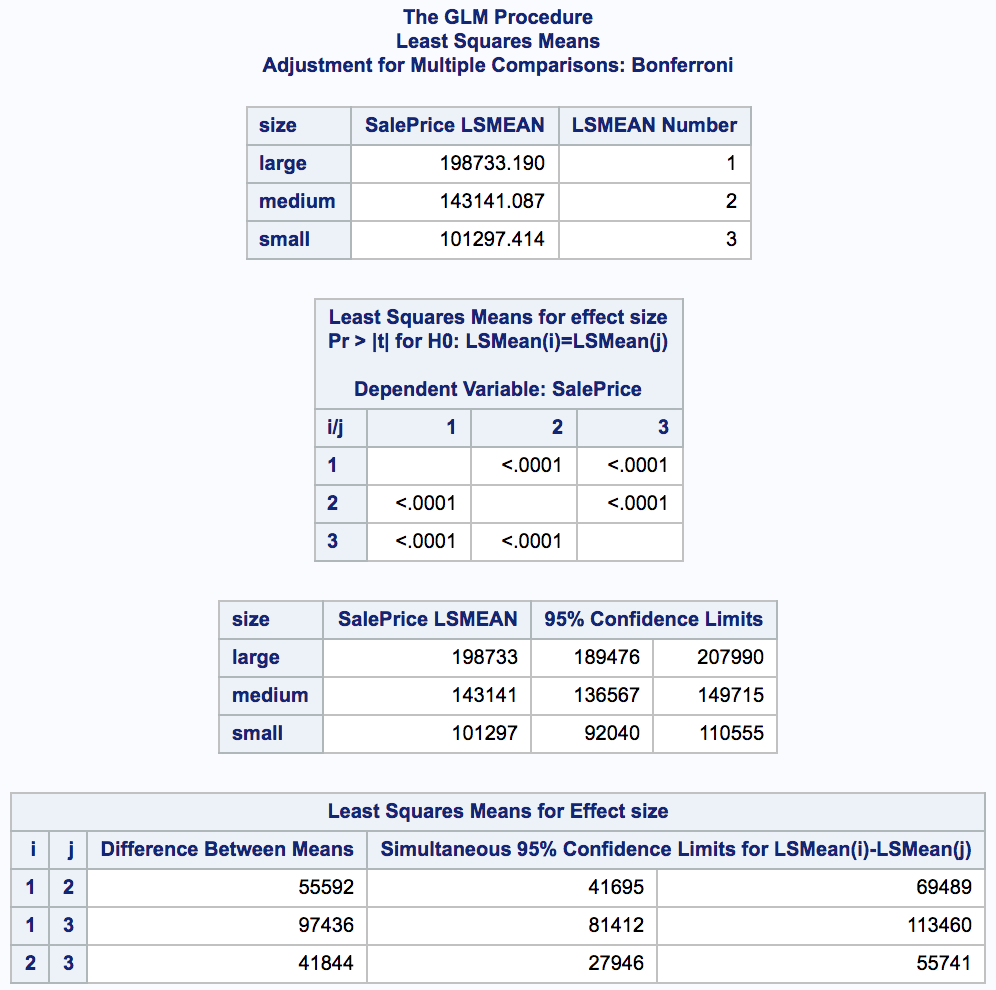
Part A) Is selling price a function of living space?

H0: Average selling price is the same for small, medium, and large houses.

Ha: Average selling price is different for small, medium, or large houses.

At the alpha = 0.05 significance level, we reject the null hypothesis.

Based on multiple comparison of means, there is significant evidence (p-vaule < .001) that the mean selling price for home varies by size. The difference in mean price between large and medium homes is $55,592 with a 95% confidence interval of [$41,695:$69,489]. The difference between large homes and small homes is $97,436 with a 95% confidence interval of [$81,412:$113,460].



Part B) Is selling price dependent on neighborhood?

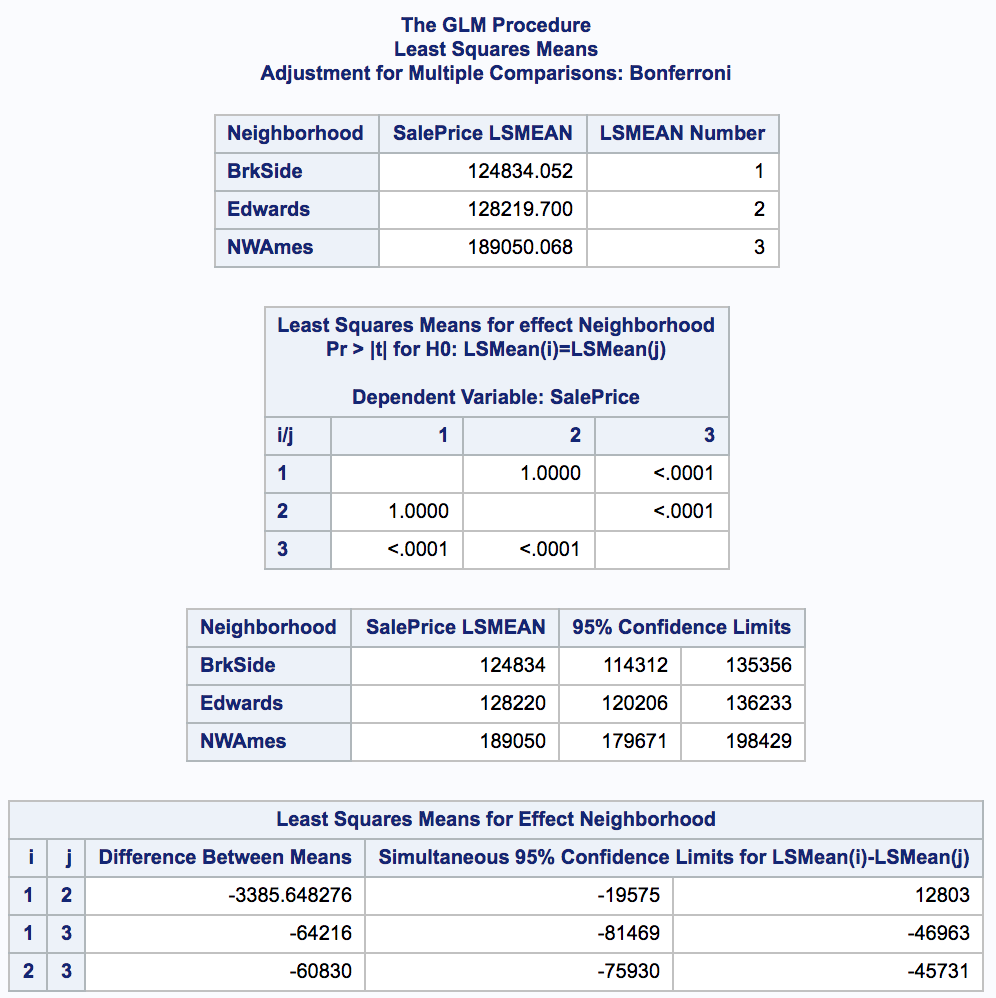
H0: Average selling price is the same for all three neighborhoods.

Ha: Average selling price is different for “NAmes”, “Edwards”, and “BrkSide”.

At the alpha = 0.05 significance level we reject the null hypothesis.

There is significant evidence (p-value < .0001) that the mean selling price differs by neighborhood. Specifically, the mean selling price for a home in Names is $64,216 greater than BrkSide with a 95% confidence interval of [$46,963:$81,469]. The mean selling price for Names is also greater than Edwards ($60,830 with a 95% confidence interval of [$45,731:$75,930].

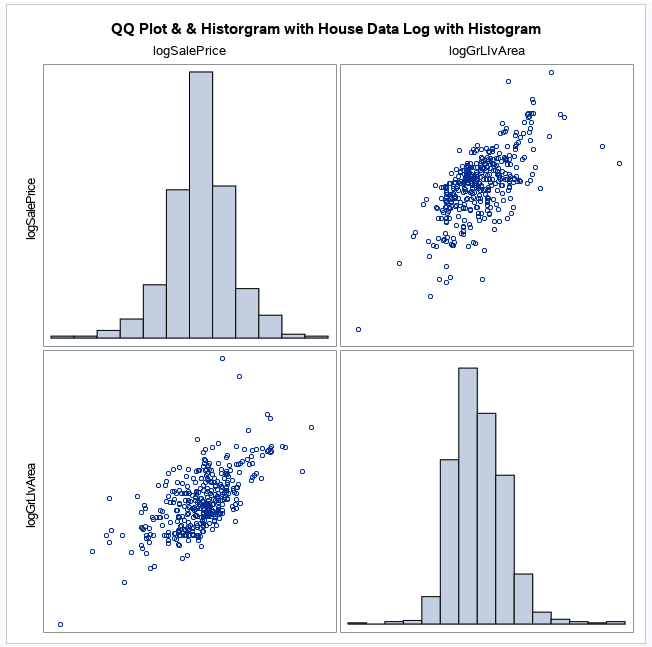
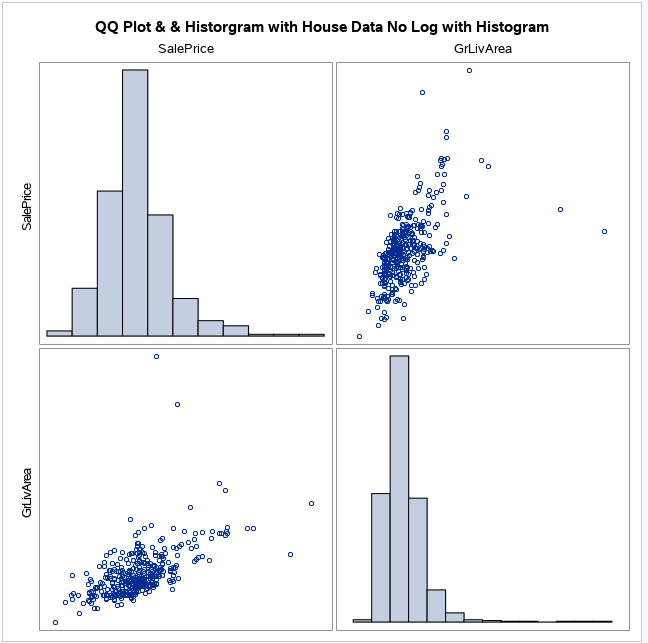
Knowing that selling price varies by size and location we can build a model that will predict the average selling price for a home based on size and neighborhood.



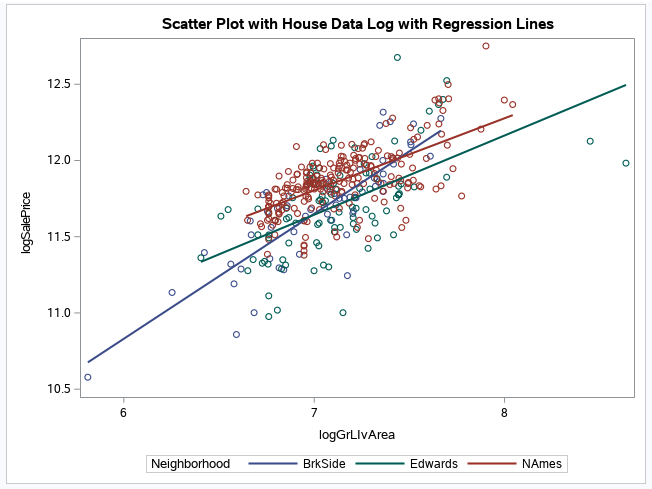
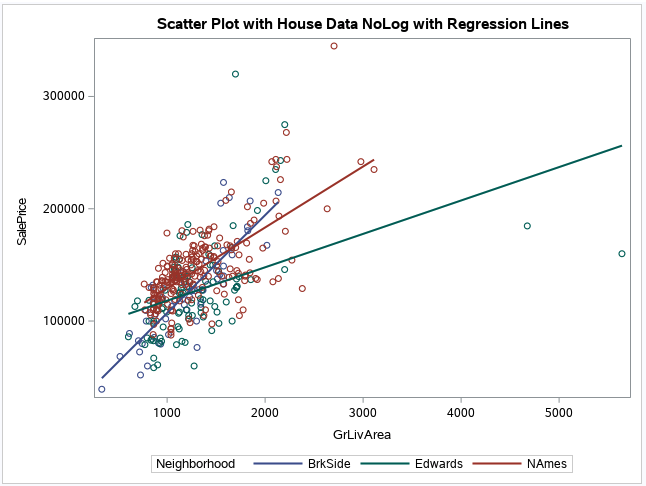
1. State Hypothesis
2. Critical Values
3. Find T Value
4. Find P Value
5. Decision
6. Conclusion

## Checking Assumptions

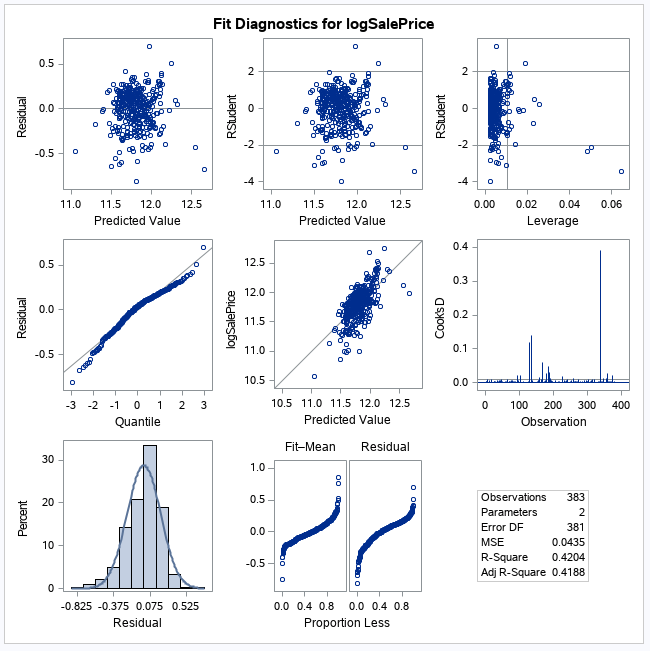
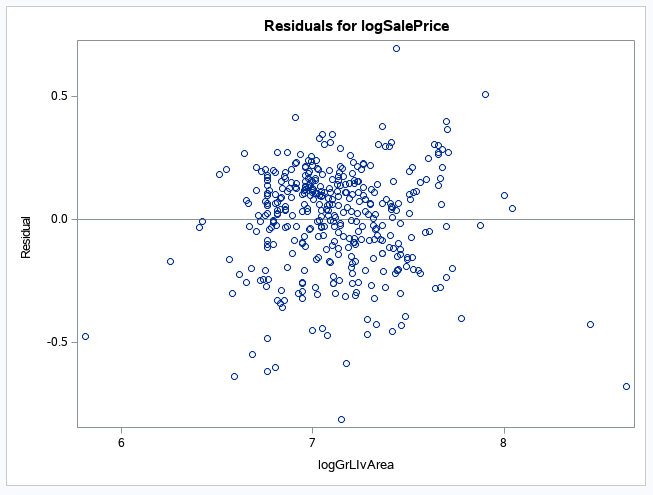
In reviewing the data, an analysis was performed using QQ Plots and Histograms, the linearity of the data is not in question due to the sample size, the data is right skewed and is not normally distributed as depicted, the data does not have equal standard deviations, and the data is independent of each other.



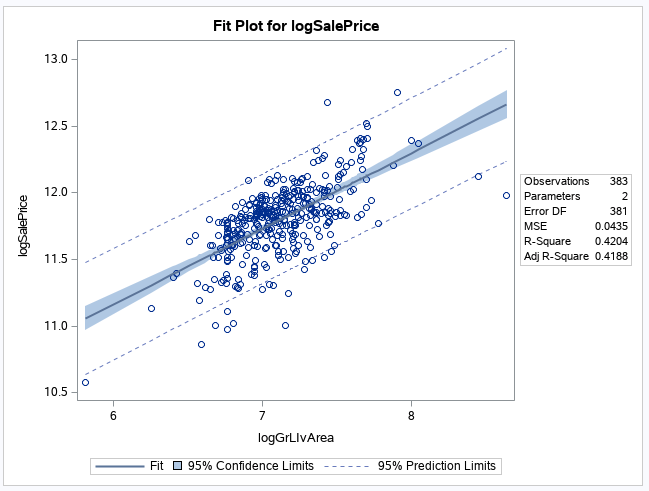
Because the assumptions in the data do not support evidence that will allow the study to continue due to data, transformation of the data was perfromed using the Log of the Sales Price and Log of the GrLivArea. Upon this transforamtion, the linearity of the data is not in question, that is not strong evidence against normalizty of residuals looking at the histogram and QQ plot, the standard deviasion appear to be equel, and the data is indpependent.



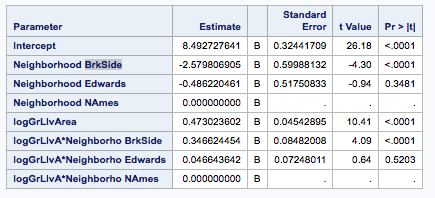
Upon reviewing the regression fit analysis of the logSalePrice and logGrLIvArea, Cooks D for leverage and studentized residuals a had a high degree of leverage and the residuals are small.

One more test was performed as part of the regression analysis to determine the confidence and prediction intervals at a 95% and the data strongly supports a good fit of the Ames data.



In Assessing the Fit, the coefficient is interpreted by the following models utilizing the base formula:

* Ames^SalesPrice = β0 + β1\*BrkSide + β1\*Edwards + β3:\*NAmes + β4 (LogLivingArea\*BrkSide) + β5(LogLivingArea\*Edwards)
* β0: The intercept in this model provides an estimate 8.49 of the logGrLivArwea (reference NAmes) with a logGrLivArwea of zero. Of course, this is extrapolation and does not have a clear, practical meaning.
* β1: This is the adjustment of the intercept for a Neighborhood BrkSide with respect to a NAmes Neighborhood. For a Living room of zero, the Neighborhood BrkSide has an estimated Sale Price Increases of -5. 16 (2^-2.58 back transformation) dollars per square foot less than the NAmes Livingroom.
* β1 This is the adjustment of the intercept for a Neighborhood Edwards with respect to a NAmes Neighborhood. For a Living room of zero, the Neighborhood Edwards has an estimated Sale Price Increases of -1.40 (2^-0.49 back transformation) dollars per square foot less than the NAmes Livingroom.
* β3: For each 1 unit increase in the Living Room of a NAmes, the estimated Sale Price increases 2^0.47 units
* β4: For each 1 unit increase in the Living Room Size of BrkSide, the estimated Sale Price increases 8.16 (2^0.35 Back transformed) dollars per square foot from the change with the NAmes.
* β5: For each 1 unit increase in the Living Room Size of Edwards, the estimated Sale Price increases 1.04 (2^0.05 Back transformed) dollars per square foot from the change with the NAmes.



## Comparing Competing Models

After reviewing the models after and before the transformation, the R-Square is (0.51, 0.45, respectively) with a RMSE of (0.19, 28552.30, respectively) and a Coefficient Variance of (1.63, 20.68, respectively) ensuring the better fit is with the transformed data.

## Conclusion

To interpret the model, a change in Living Room Square Feet Is a 10-fold increase. For the neighborhood with approximately the same mass, it is estimate that a 10-fold increase in the Living Room Square feet is associated with a (10^-2.58 = 7.42) which is a 742% increase in the median Sales Price of the neighborhood. (P value < 0.001). At a 95% confidence intervals for the increase in sales price of (10^-3.76, 10^-1.4) = (6.24, 8.6) which equates to an estimated increase between 625% and 860%

# 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

\*\* URL for Kickoff \*\*

https://www.youtube.com/watch?v=0QJtczDPxZQ

;

\*\* Review Data;

\*\*proc print data = train;

\*\*run;

\*\*proc print data = test;

\*\*run;

\*\* Add SalePrice to Dataset and Review;

data test;

set test;

SalePrice = .;

run;

\*\* Consolidate Train & Test Files Together and Review Data;

data train2;

set train test;

run;

\*\* Transform Data;

data train3;

set train2;

rGrLIvArea = round(GrLIvArea, 100);

log2GrlivArea = log(GrLIvArea / 100);

logGrLIvArea = log(GrLIvArea);

logSalePrice = log(SalePrice);

run;

\*\*proc print data = train3;

\*\*var GrLIvArea rGrLIvArea;

\*\*run;

\*\* Select only Century 21 Homes - NAmes, Edwards, BrkSide;

data train4;

set train3;

if Neighborhood in ("NAmes", "Edwards", "BrkSide");

run;

\*\*Print Header to get Field Names;

\*\*proc print date = train4 (obs=15);

\*\*run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Begin Analysis of only Century 21 Data

Data Summary

Train4 - Century 21 Only with Log

Train3 - Rounded Data for GrLIvArea

Train2 - Consolidated Data

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

proc sgscatter data = train4;

title "QQ Plot & & Historgram with House Data Log with Histogram";

matrix logSalePrice logGrLIvArea / diagonal=(histogram);

run;

proc sgscatter data = train4;

title "QQ Plot & & Historgram with House Data No Log with Histogram";

matrix SalePrice GrLIvArea / diagonal=(histogram);

run;

proc sgplot data = train4;

title "Scatter Plot with House Data NoLog with Regression Lines";

reg x = GrLIvArea y=SalePrice / group=Neighborhood;

run;

proc sgplot data = train4;

title "Scatter Plot with House Data Log with Regression Lines";

reg x = logGrLIvArea y=logSalePrice / group=Neighborhood;

run;

proc reg data = train4;

title "Pro Reg House Data Log";

model logSalePrice = logGrLIvArea / vif;

run;

proc reg data = train4;

title "Pro Reg House No Data Log";

model SalePrice = GrLIvArea / vif;

run;

proc univariate data=train4; var logSaleprice; run;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*Question #1

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\* Model Attempt #1- Log Ames Data ;

proc glm data = train4 plots = all;

class Neighborhood(ref='NAmes');

model logSalePrice = Neighborhood | logGrLIvArea / CLPARM solution clm;;

output out = results p = Predict;

run;

\*\* Model Attempt #1- No Ames Data ;

proc glm data = train4 plots = all;

class Neighborhood(ref='NAmes');

model SalePrice = Neighborhood | GrLIvArea / CLPARM solution clm;;

output out = results p = Predict;

run;