**Brook – problem statement says introduction and data description come before analysis 1.**

**Introduction:**

Real estate company Century 21 Ames, selling homes in Ames, Iowa has come to us with a data set and some questions pertinent to their business. In this paper, we will address Century 21 Ames’ specific questions in our first analysis, before proceeding with a second analysis exploring the data on our own to predict sales prices of houses in all of Ames Iowa.

**Data Description**:

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. We have used this dataset throughout our analysis, it is contained in the files “train.csv” and   
“test.csv.” This dataset gives the sale price of homes in all of Ames, Iowa from 2006 to 2010. The dataset also provides 79 variables that were identified as possible explanatory variables for assessing the sale price of a home in Ames. There are 2921 observations in the dataset, and they have been split almost evenly between our model training csv file and model testing csv file.

Further information about the data can be found at <http://jse.amstat.org/v19n3/decock.pdf>

Our model for Century 21 Ames’ problem aims to predict the variable “SalePrice.” Our client has specifically asked us to only focus on the relationship of the explanatory variables referring to square footage of the living area of the house and the neighborhood the house is set in. These variables are labeled as “GrLivArea” and “Neighborhood” respectively.

* Here we will address the variables we end up using for analysis 2

**ANALYSIS 1:**

**Restatement of Problem:**

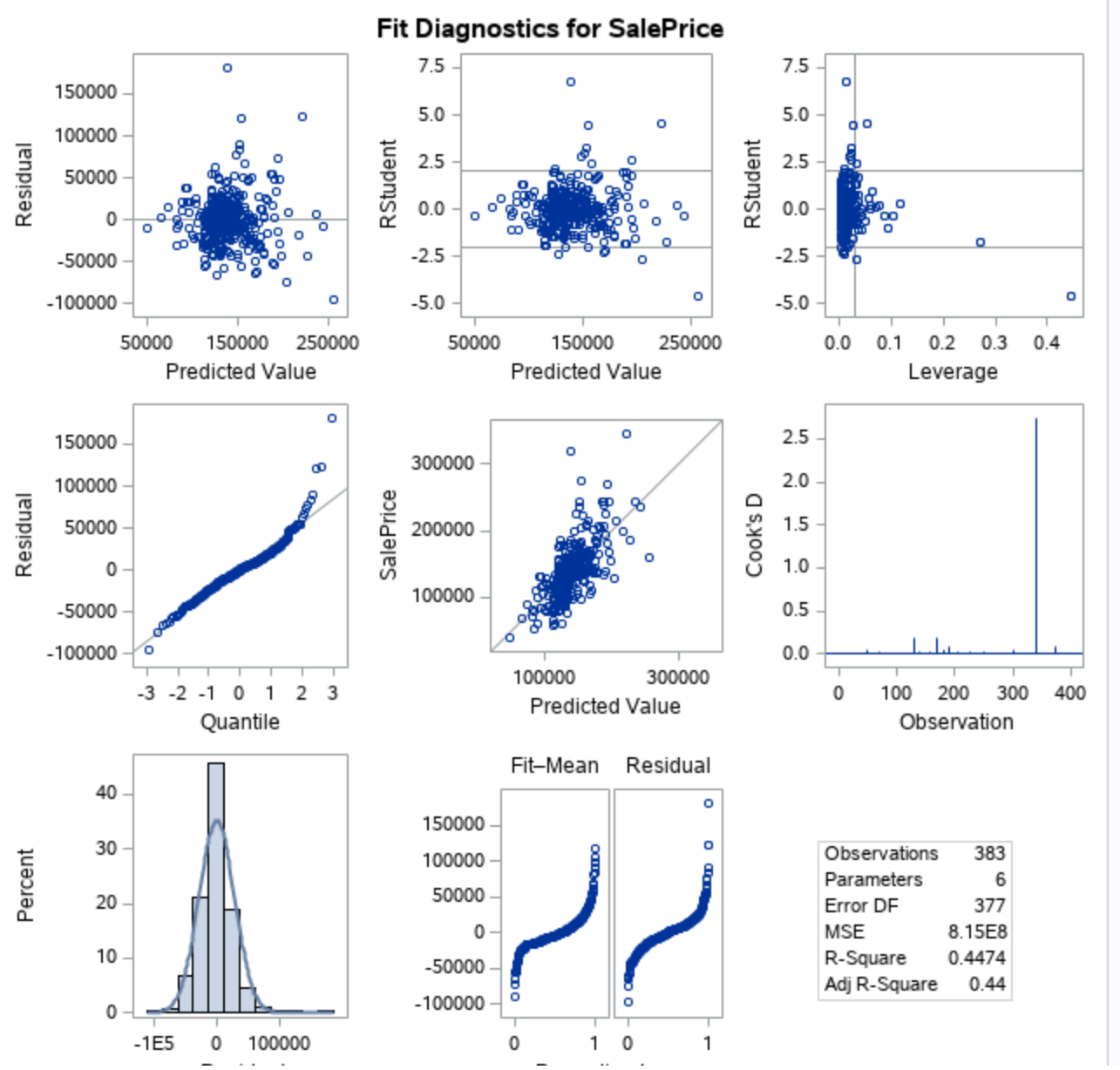
Our client, Century 21 Ames, sells houses in three neighborhoods of Ames, Iowa. These neighborhoods are Brookside, Edwards and N. Ames. Century 21 Ames would like us to estimate how the sale price of a house in one of their neighborhoods is related to the square footage of the living area of the house. They would also like to know if the sale price and aforementioned relationship with the living area depend on the neighborhood the house is being sold in.

**Build and Fit of the Model:**

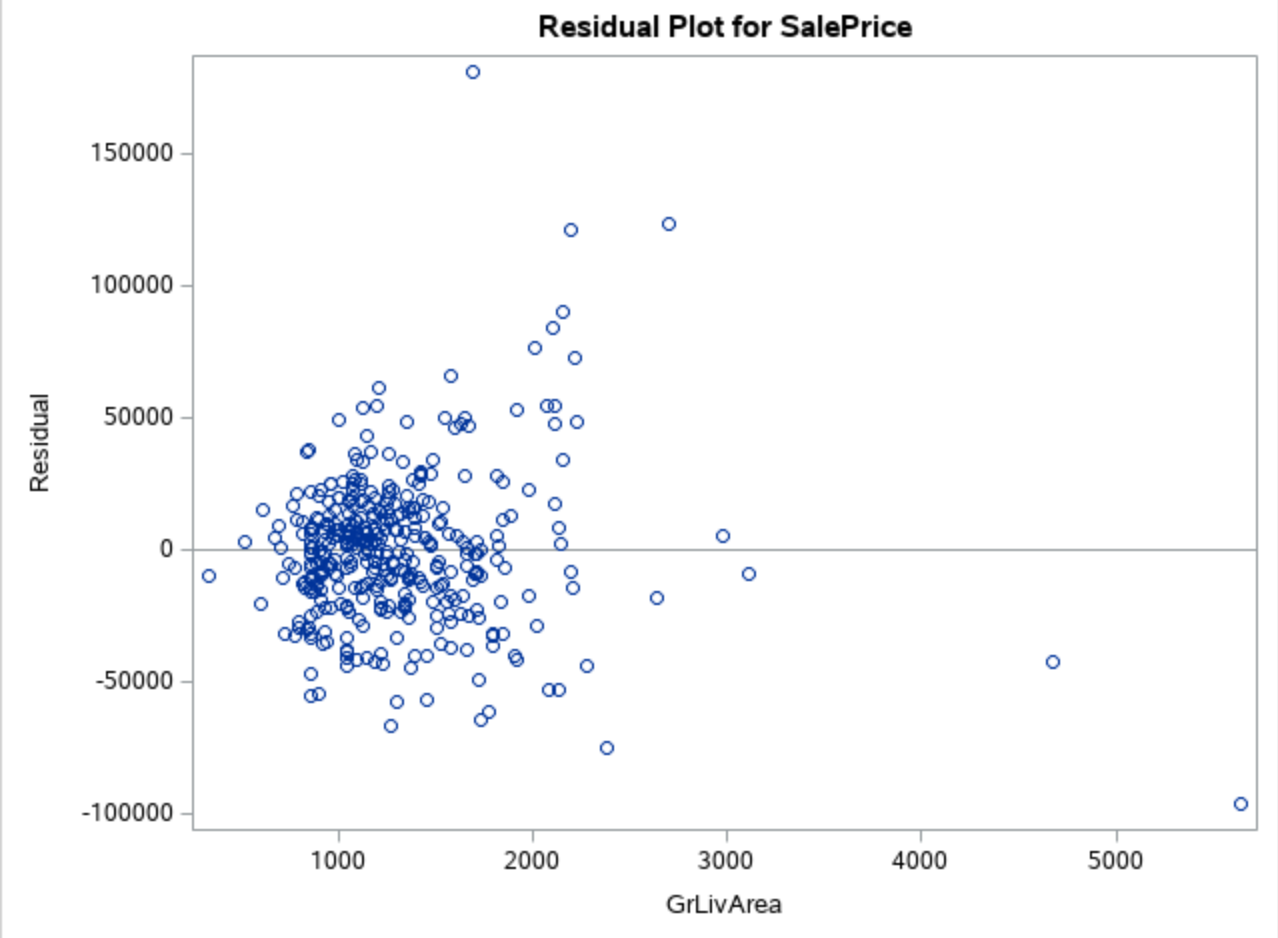
Build and fit a model that will answer this question, keeping in mind that realtors prefer to talk about living area in increments of 100 sq. ft. Provide your client with the estimate (or estimates if it varies by neighborhood) as well as confidence intervals for any estimate(s) you provide. – We can delete this after we’re done, good to have as a reference until finished

fit diagnostics for different cases

Linear – linear model



**The residual plot also presented as**



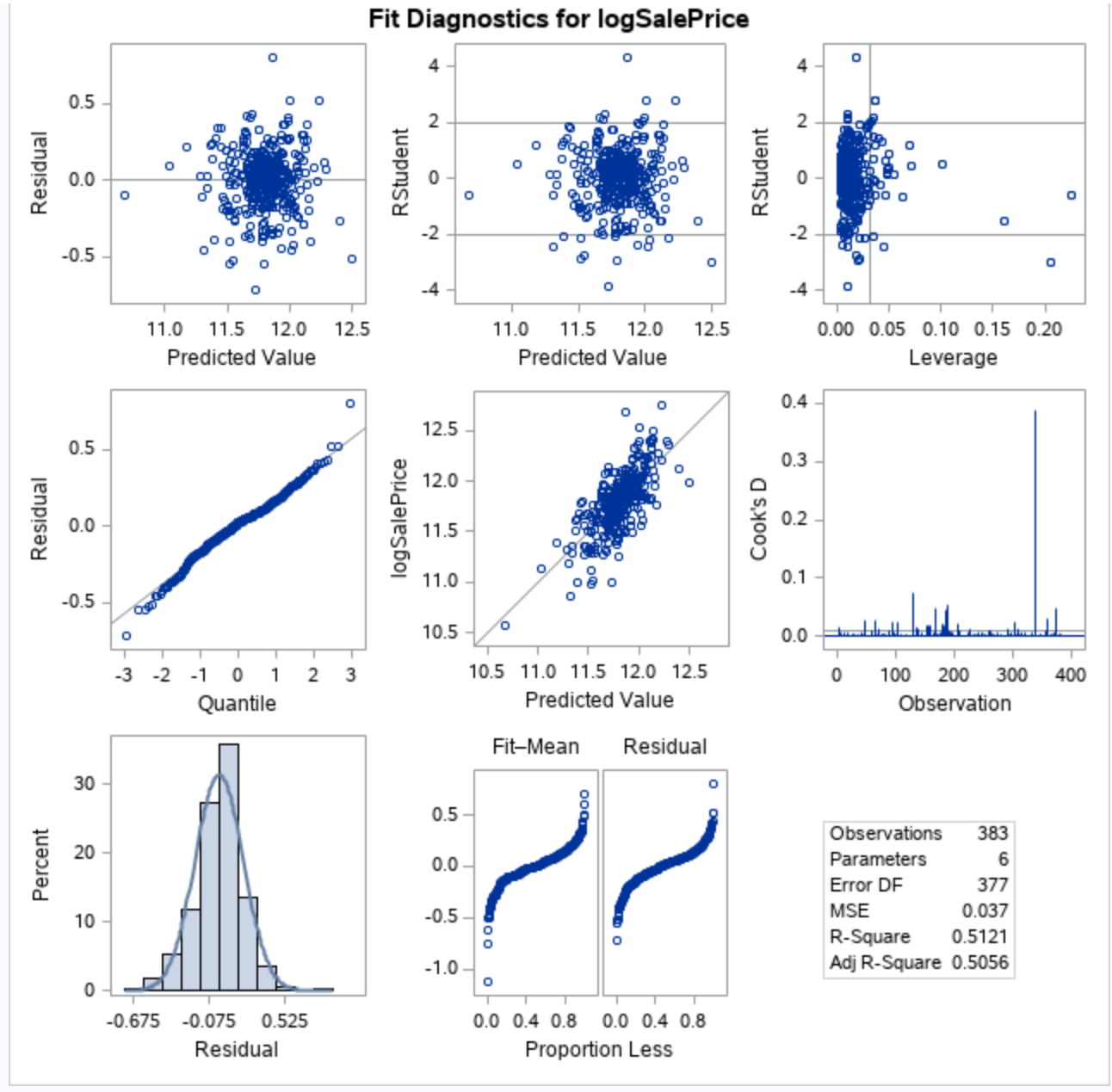
**Comparing Competing Models:**

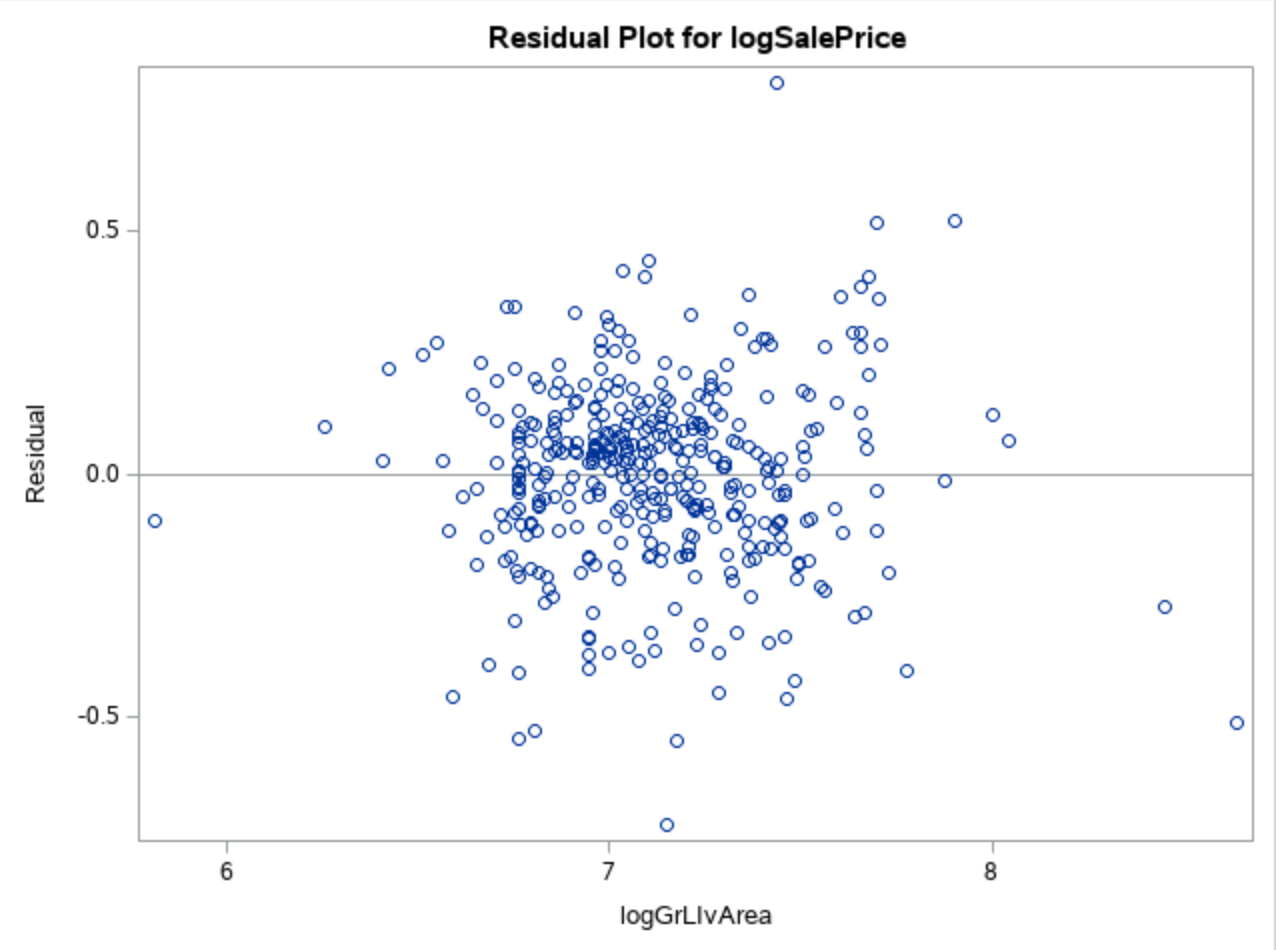
Adj R2

Internal CV Press

Brook – I’ve removed the linear-log section, I don’t think there’s a need to include it if it is not going to be used, unused EDA doesn’t have to be presented

**Log log model**





**Checking Assumptions:**

* Residual Plot: The residual plot resembles somewhat of a random scatter of points around the 0 line, although there is a slight suspicion of non-constant variance judging from the dense cloud around.
* Studentized Residual Plot: This plot is very similar to the residual plot, although this plot identifies potential outlying observations. This plot identifies a potentially very outlying point with a predicted value of 15. This may provide some evidence against the normality assumption and this point should be examined further.
* Histogram of Residuals: The histogram of residuals displayed does not provide strong evidence that the residuals are not normally distributed.
* Q-Q Plot of Residuals: The Q-Q Plot of residuals provides no evidence against the residuals being normally distributed.
* The model is a reasonable fit with transformations, although transformations may be investigated to handle the possible problem with equal standard deviations.

**A discussion supporting the use of the model you chose (support that the assumptions are met).**

* Linearity: Met with original and log-log model. (view scatterplots)
* Normality: Log-log model looks slightly better. (view histograms)
* Equal standard deviations: Log-log model looks much better. (view residual scatter plots)
* Independence: We will assume independence, although not much is known about how these species were chosen.
* Outliers: There are some outliers at the tail end of the data set but the log log model looks better.
* We will proceed to make inferences on a log – log model.

Residual Plots

Influential point analysis (Cook’s D and Leverage)

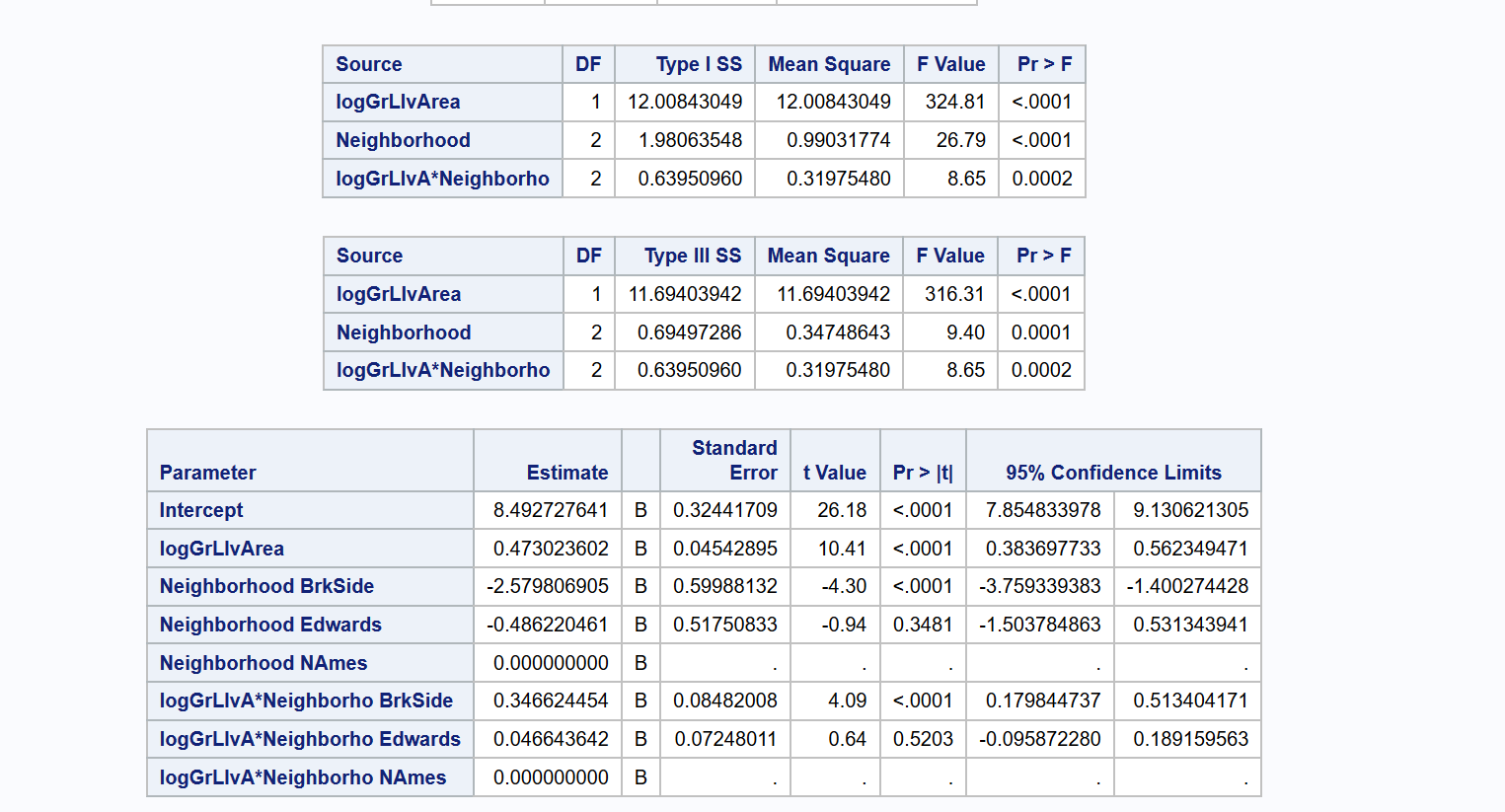
Observed significant Cook’s D observation on all models except the linear model which only has one spike at the tail end of the observations

Leverage seems normal for all the three models

**Parameters:**

Estimates

From the log – log model ;



Interpretation

Confidence Intervals

**Conclusion:**

A short summary of the analysis.

Brook – I have uploaded my code below yours, I think mine is a little more clear.. what do you think John?

**Appendix:**

Code and additional screen shot of plots and graphs

proc import datafile='/folders/myfolders/testCleaned.csv'

DBMS=csv out=work.testCleaned replace;

run;

proc print data=work.testCleaned;

run;

proc import datafile='/folders/myfolders/train.csv'

DBMS=csv out=work.train replace;

run;

proc print data=work.train;

run;

\* Build a Model;

\*

\* Fit plot assesment on log log model ;

proc glm data = logtrain plot = all;

class neighborhood;

model SalePrice = GrLivArea |neighborhood / solution;

run;

\*log - log model ;

proc glm data = logTrain plots = all alpha = 0.05;

class neighborhood;

model logSalePrice = logGrLivArea |neighborhood / solution clparm;

run;

\* linear log model ;

proc glm data = logTrain plots = all alpha = 0.05;

class neighborhood;

model SalePrice = logGrLivArea |neighborhood / solution clparm;

run;

\*log - linear model;

proc glm data = logTrain plots = all alpha = 0.05;

class neighborhood;

model logSalePrice = GrLivArea |neighborhood / solution clparm;

run;

proc reg data = logTrain ;

model logSalePrice = logGrLivArea /VIF ;

run;

\*Coded scatter plot ;

proc sgplot data=train;

reg x=GrLivArea y=SalePrice / group=neighborhood clm cli;

proc sgplot data=train;

where neighborhood ne contains ('NAmes', 'Edwards' ,'BrkSide') ;

reg x=GrLivArea y=SalePrice / group=neighborhood clm cli;

# Create our data set to answer Century 21 Ames’ questions. Select only the neighborhoods they care about, and create log values for our numerical variables in case we determine a transformation is necessary. Also, sort by neighborhood for clarity when looking at the raw data:

data Century21;  
 set train;   
 if Neighborhood = 'BrkSide' OR Neighborhood ='NAmes' OR Neighborhood = 'Edwards';  
 logGrLivArea = log(GrLIvArea);  
 logSalePrice = log(SalePrice);  
 run;

proc sort data = Century21;

by Neighborhood;

run;

# Create our linear regression model for initial analysis:

proc glm data = Century21;

class Neighborhood;

model SalePrice = GrLivArea | Neighborhood;

run;

# Convert to a log-log model after checking assumption:

proc glm data = Century21;  
class Neighborhood;   
model logSalePrice = logGrLIvArea| Neighborhood / solution clparm;  
run;