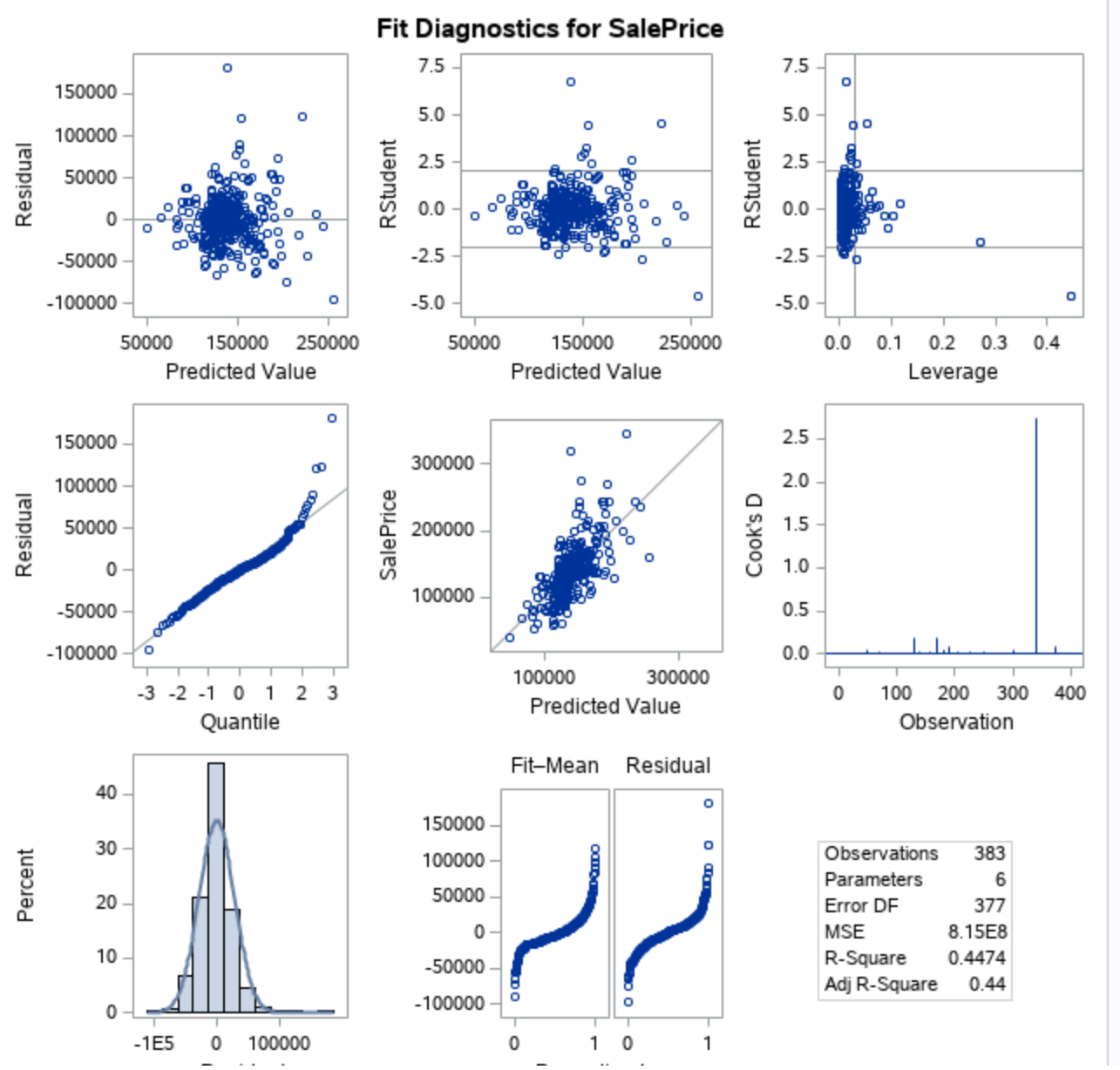
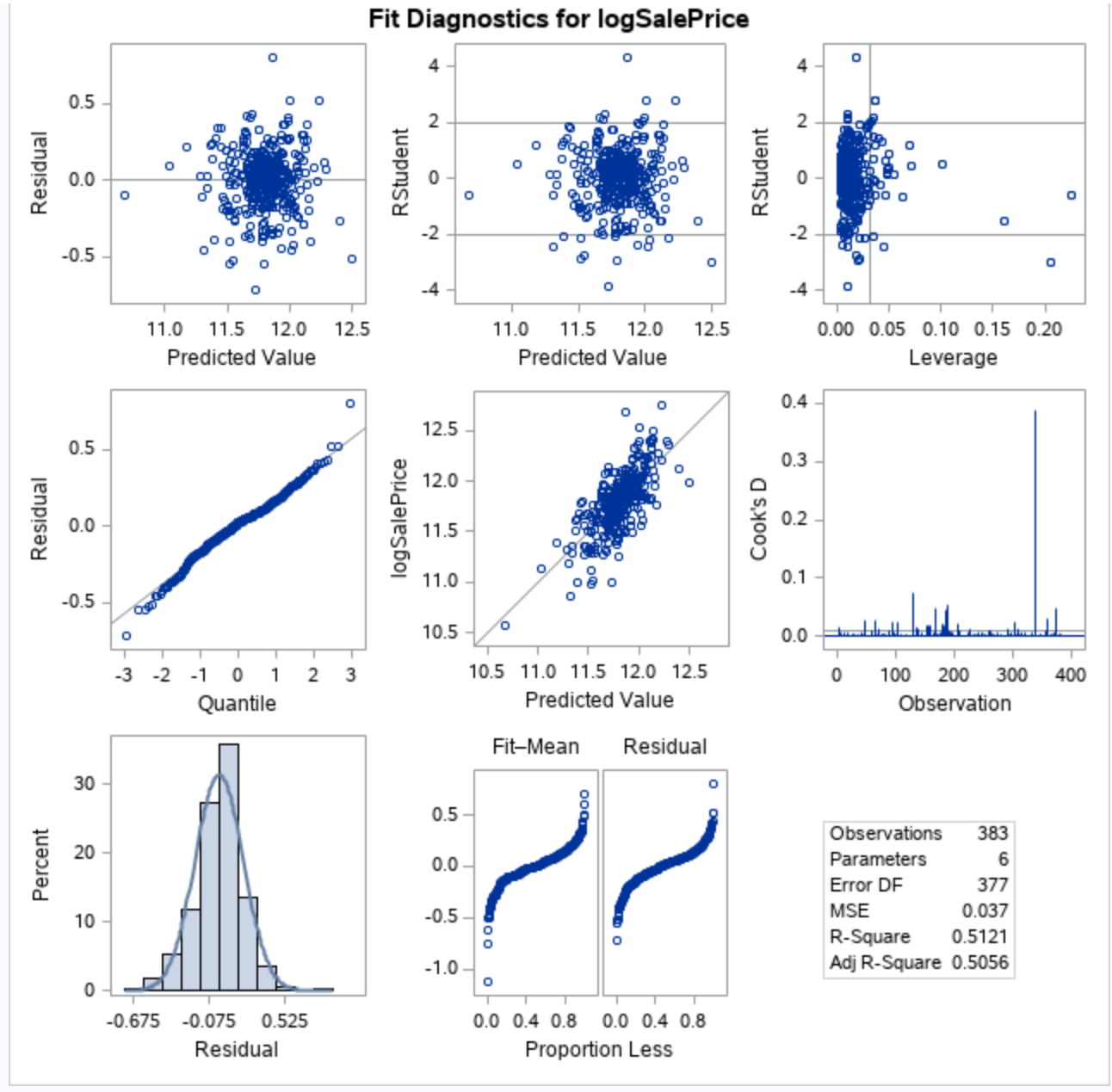
1. **Executive Summary**//I figure here we can include a quick, down and dirty writeup of what we found. Maybe just for Q1?
2. **Introduction**Century 21 Ames, a real estate company in Ames, Iowa requested analysis on data which they collected, to provide insight into price trends in the neighborhoods they sell houses in. The following writeup describes the methodology, analysis, and results of our analysis of this data. The first half of this writeup describes the analysis performed to answer specific questions Century 21 Ames asked- What is the sale price per hundred square feet on average in the three neighborhoods we sell in? How can we predict sales price based on square footage? The second half of this writeup details our efforts to more accurately model the sale price of a given house, provided with more variables to help explain the sales price.
3. **Data Description**The data provided by our client was compiled by Dean De Cock and has been published into the public domain for educational use. 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>.   
     
   For our initial analysis, 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. The variables used to build our second more complex model will be detailed in the latter half of this writing.
4. **Analysis 1: Correlating sales price, neighborhood, and square footage.**
   1. **Analysis 1 Summary**

As discussed briefly above, Centure 21 Ames requested that we build a model which can be used to estimate sales price, given the neighborhood and square footage of the home. To perform this analysis, multiple linear regression was performed, using square footage as the variable to explain the sales price, and using neighborhood as a “grouping” variable, to help better estimate the prices in each neighborhood. The data was found to require a transformation prior to analysis, as the relationship between square footage and sales price was found to not be linear and normally distributed. We opted to take the logarithm of both variables, which resulted in a satisfactory distribution of data for analysis. SAS was used to perform the analysis, the results of which are detailed below.

* 1. **Analysis Assumptions**

Initial analysis was performed on the raw data, with no transformation or outliers addressed. The fit analytics are shown below, which raise some concerns.  
  
Specifically, the residual (difference between modeled and actual values) plots show that the residuals are not normally distributed, and the data appears to be nonlinearly correlated. For these reasons, we opted to transform the data by taking the logarithm of sales prices and square footage. The fit diagnostics for this are below and are much more encouraging.



The residuals here are normally distributed, appear to be linearly correlated, and the overall distribution of the data looks a lot better. Specifically, the residual plot appears to be a random cloud centered over the model line. There is some evidence of non-constant variance still, but it is dramatically reduced, and should be safe for analysis. The studentized residuals plot, a normalized version of the residual plot, indicates a few points which are of some concern as possible outliers. This may also be evidence of non-constant variance. However, examining the Cook’s-D chart, it appears that only one point is a significant outlier with leverage to skew the model, and it still wouldn’t affect the model significantly. The histogram of residuals suggests that the residuals are normally distributed; this is a large improvement over the initial evaluation. The Q-Q plot of residuals provides no evidence against the residuals being normally distributed. These observations will be assumed to be independent. Again, as the charts and discussion above display, the model is a reasonable fit for this data. It is worth addressing again that there are some slight outliers in this data set, but as they do appear to be true recordings, we will include them. They do not violate the assumptions required for linear regression inferences to be made. We will proceed with the Log(x)/Log(y) transformed model.

* 1. **Models**

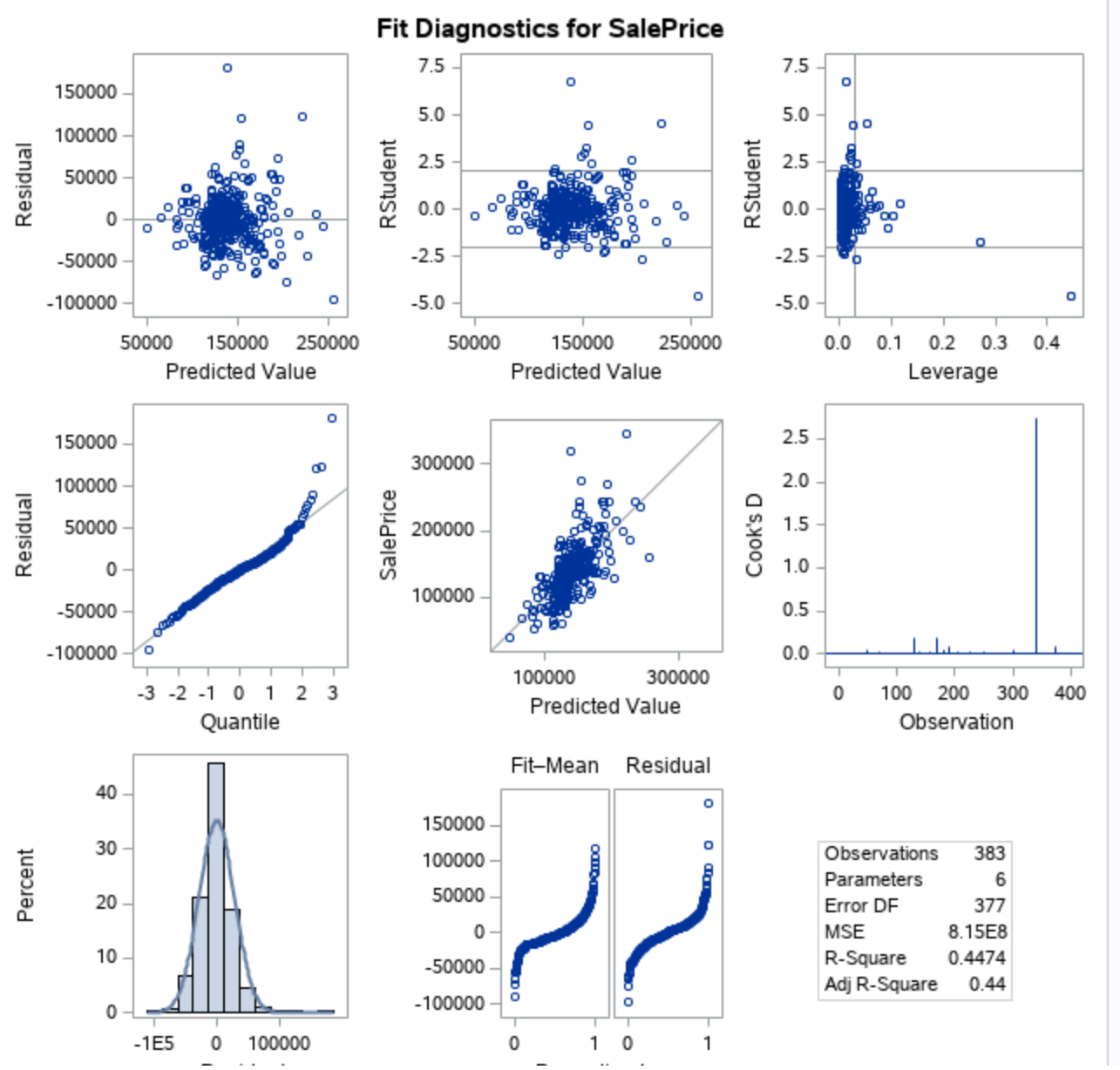
Three separate models were fitted to this data- One for each neighborhood.   
**//////////pausing here for the night, will finish up conclusions tomorrow**

**Build and Fit of the Model:**

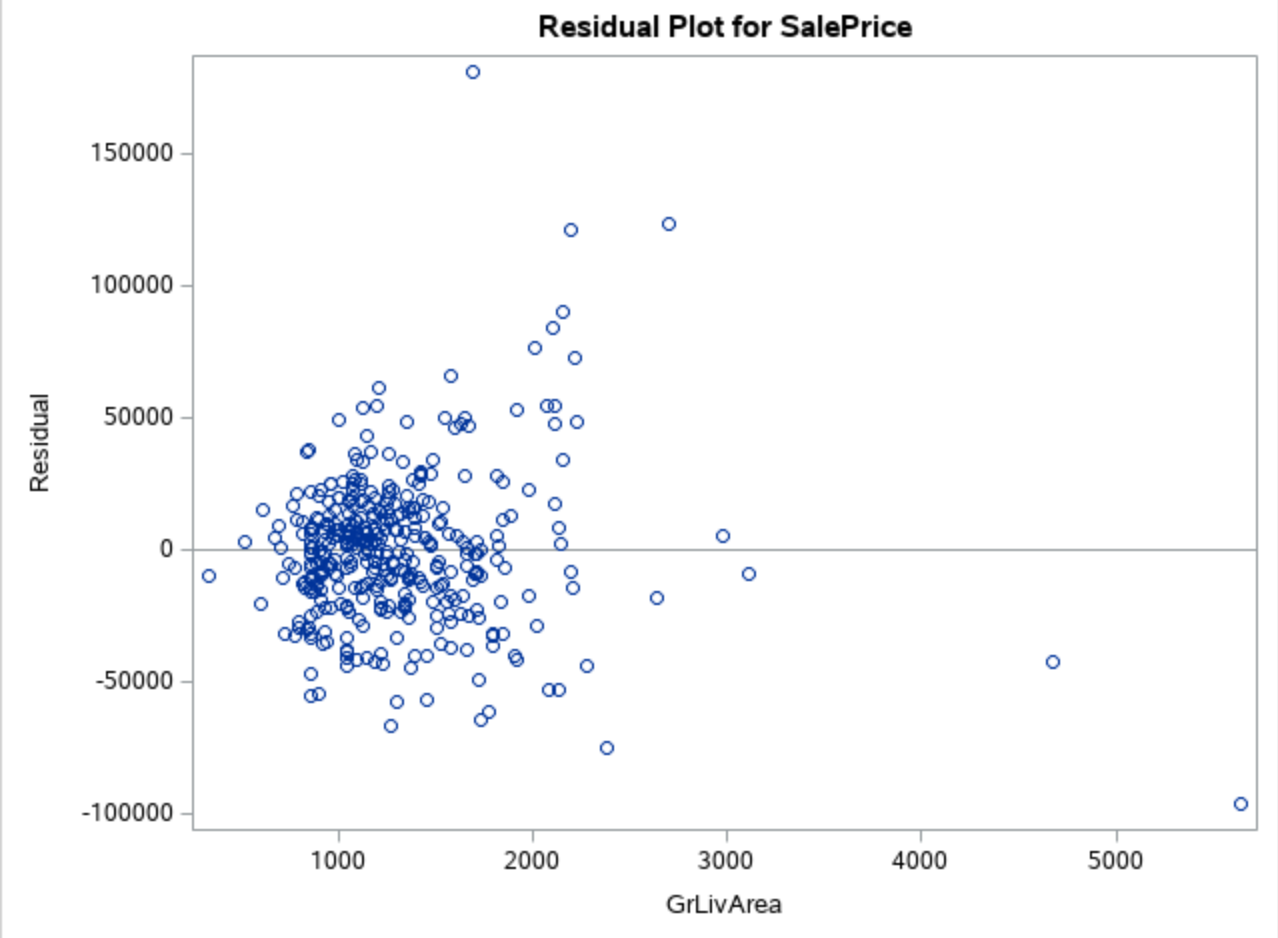
Build and fit a model that will answer this question, keeping in mind that realtors prefer to talk about living area in increments of 100 sq. ft. Provide your client with the estimate (or estimates if it varies by neighborhood) as well as confidence intervals for any estimate(s) you provide. – We can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



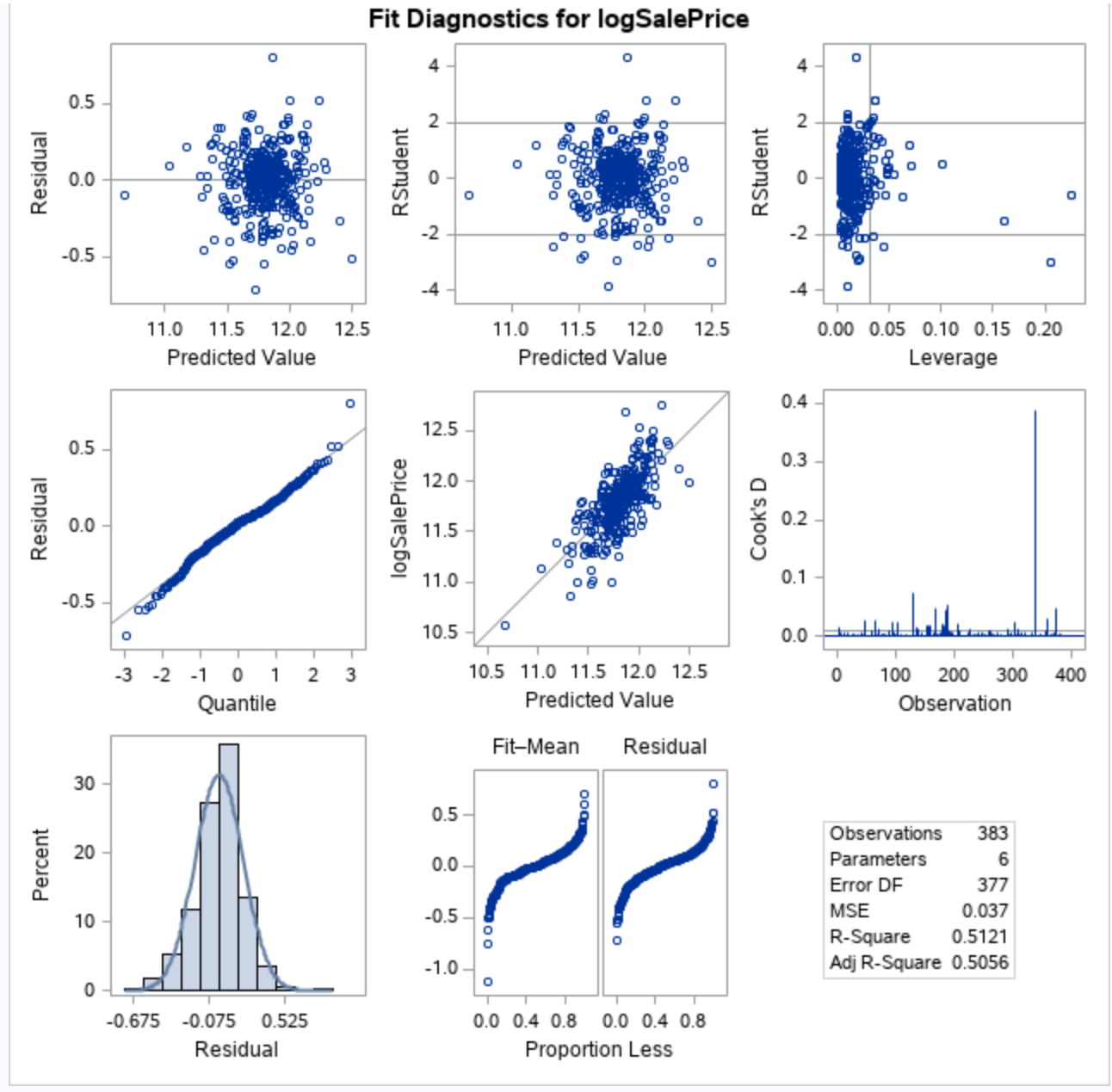
**Comparing Competing Models:**

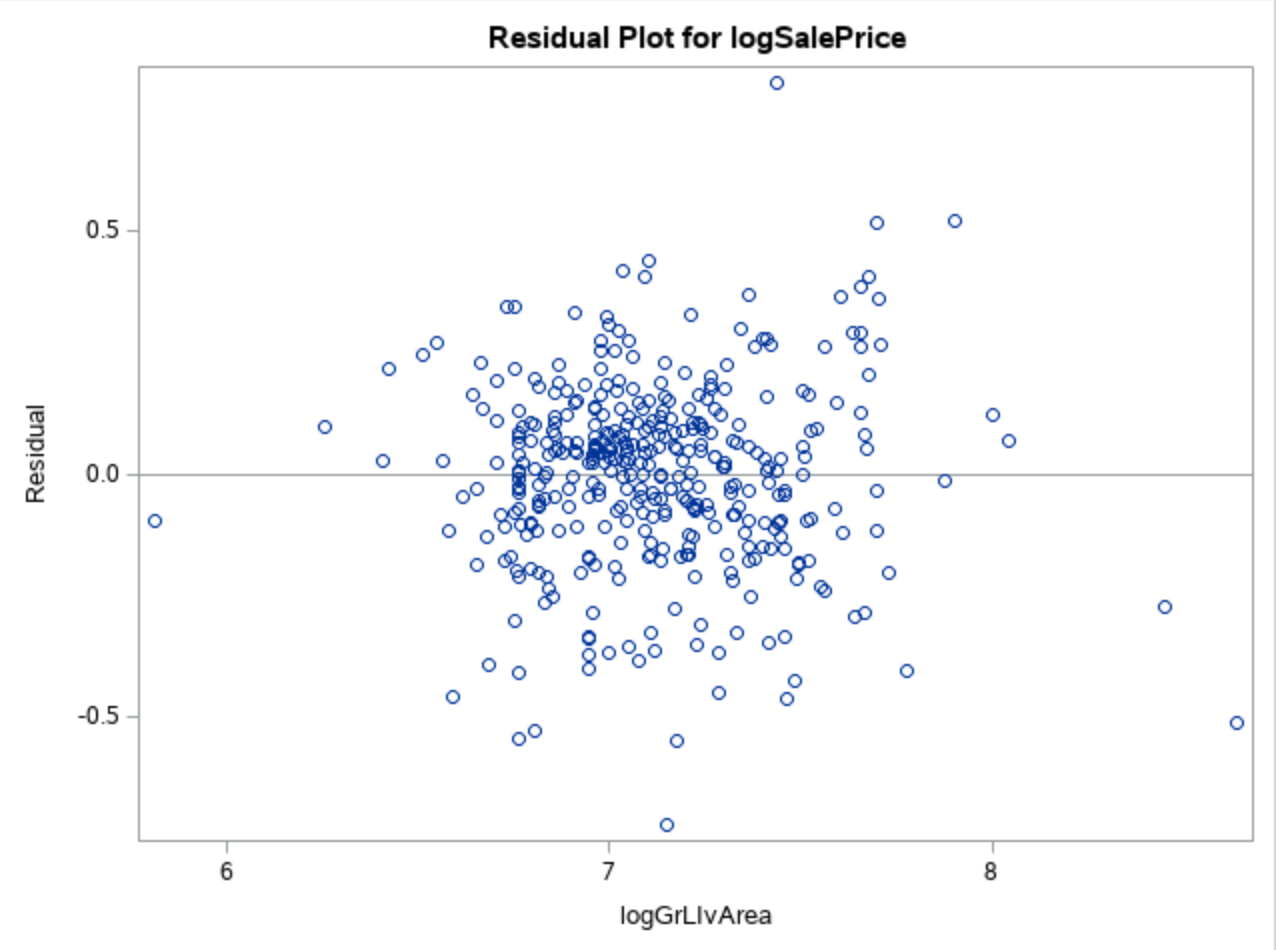
Adj R2

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)

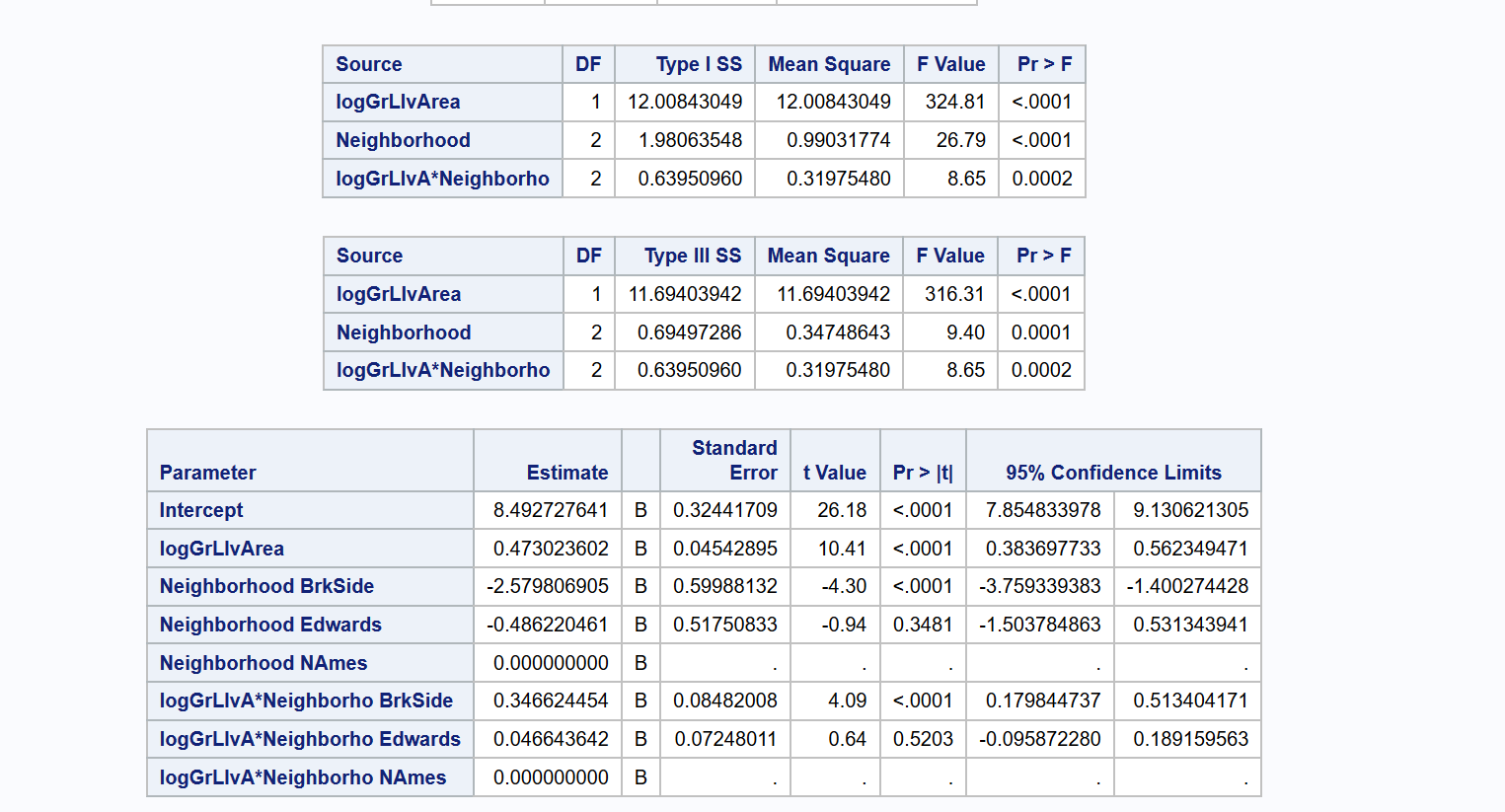
Observed significant Cook’s D observation on all models except the linear model which only has one spike at the tail end of the observations

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?   
**Seconded- You used more full/clear English in your code (which I also didn’t do), I like it.**

**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;