Hello, Century 21!

(Restatement of problem)

We hear you need some answers to some very important questions with respect to your business! We can provide the answers you are looking for concerning the housing market in Ames, Iowa.

We were able to get our hands on a dataset containing 1460 home sales, and 80 associated variables. While the dataset is quite big, you’ve narrowed things down to a few specific neighborhoods you’re interested in (NAmes, Edwards, and BrkSide), leaving us with 383 houses to study. This number according to statistical methodology will be sufficient to draw conclusions on for the neighborhoods. In short, you are interested in how specific neighborhoods and square footage are associated in price.

To begin, we need to look at the requested data and determine if it is suitable for analysis in its current form, and based on the non-normality of the histograms of sale price and square footage (Plot 1, Plot 2, respectively), these variables will need to be log-transformed to address the normality assumption (Plot 5, Plot 6 respectively).

(Build and fit the model)

At this point I believe we are ready to build a model and evaluate results.

Log(SalePrice)= B0 + B1 \* log(SquareFootage) + B2 \* Edwards + B3 \* NAmes + B4 \* log(SquareFootage) \* Edwards + B5  \* log(SquareFootage) \* NAmes

Log(SalePrice)= 6.33 + 7.480e^-1 \* log(SquareFootage) + 2.393e^-1\* Edwards + 4.315e^-1\* NAmes + -2.146e^-3\* log(SquareFootage) \* Edwards + .2.483^e-4 \* log(SquareFootage) \* NAmes

(Checking assumptions)

For linear regression we have some assumptions to meet before we can make inference on a model. Based on our evaluations from this model, the assumptions are met after applying a log transformation to sales price and square footage, and because of this we are going to make inference on medians now. Those assumptions are Normality of data, homogeneity of variance, and independence. We can examine the normality assumption from plot 5 and plot 6 that the data follows the trend of normality. The quantile-quantile (plot 9 and plot 11) looks roughly normal, and is about as good as we can get with a quantile-quantile plot of real data. Homogeneity of variance appears to be met (plot 10, plot 10), as there is no obvious trend to the data, it seems to be spread evenly between the two sides of the line. Finally, we have the independence assumption, we will have to make this assumption for this dataset, although realistically we cannot see this as being true, houses will be related to each other.

Additionally, the residuals looked very good given the dataset, and since we have a large number of observations the central limit theorem will come into play. The data did show some outliers and leverage points (plots 7,plot 8, and plot 9). The outliers were all investigated and we concluded that there we no measurement errors and decided to keep the outliers in the data. The model was created again with outliers removed and we determined the results were very similar to the model with the outliers included. The plots of this model are 14,15,16,17,18,19,20.

(Comparing competing models)

The model we created has an r-squared value of 0.5178, meaning that 51.78% of the variance of the sales price variable is explained by the variation in neighborhood and square footage.

(Parameter Estimates)

Beta 0: Estimate =6.334, Standard Error =7.060e^-1, t-value=8.972, P-value: <0.0001, significant for model

Beta 1: Estimate =7.480^e-01, Standard Error =1.176e^-01, t-value=6.359, P-value: <0.0001, significant for model

Beta 2: Estimate =2.393^e-01, Standard Error =1.12e^-01, t-value=2.136, P-value: 0.03332, significant for model

Beta 3: Estimate =4.315^e-01, Standard Error =9.929e^-02, t-value=4.346, P-value: <0.0001, significant for model

Beta 4: Estimate =-2.146^e-04, Standard Error =8.655e^-05, t-value=-2.480, P-value:0.01359, significant for model

Beta 5: Estimate =-2.483^e-04, Standard Error =7.704e^-5, t-value=-3.223, P-value: <0.00138, significant for model

DF: 376

95% CI for Beta 0: (4.945,7.721)

95% CI for Beta 1: (0.516,9.793)

95% CI for Beta 2: (0.019,4.596)

95% CI for Beta 3: (0.236,6.268)

95% CI for Beta 4: (-0.0004,-4.444)

95% CI for Beta 5: (-0.0004,-9.685)

For each term in the model, the p-values were all below the standard significance level of 0.05, and thus for each variable we will reject the null hypothesis and conclude that there is a significant difference and that each parameter is useful in the model.

To more easily interpret the model, we can take a look at the equation for each neighborhood and how sale price is related to square footage.

Log(squarefootage)|Brkside = 6.334 + 0.748log(squarefootage)

Log(squarefootage)|Edwards = 6.573 + 0.74779log(squarefootage)

Log(squarefootage)|NAmes = 6.766 +0.74775log(squarefootage)

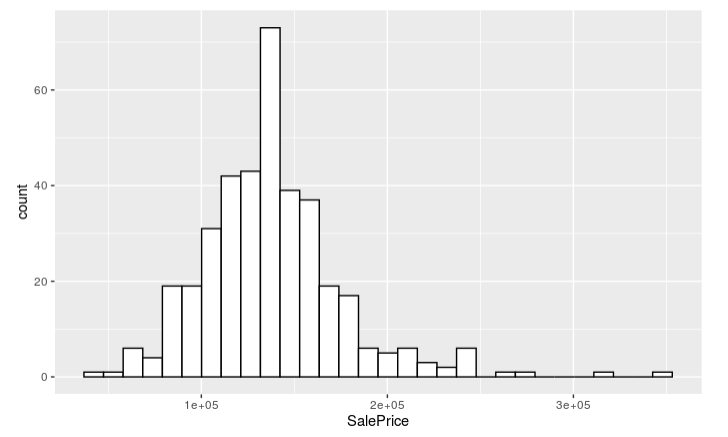
Interpretation of model:

Since this model is log transformed we will speak in terms of medians. When we increase the size of a house in a given neighborhood (since its log transformed we will say the size is doubled).

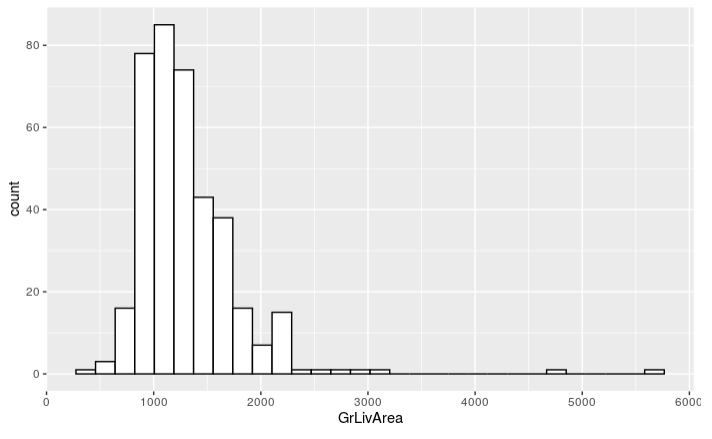
However, since this is an observational study, we cannot infer causation and say with certainty that the square footage or neighborhood can cause a change in sale price. It can be inferred that there are many more variables at play here, and we can conclude by saying that there is an association between the selling price of a house, it’s square footage, and its neighborhood.

Appendix:

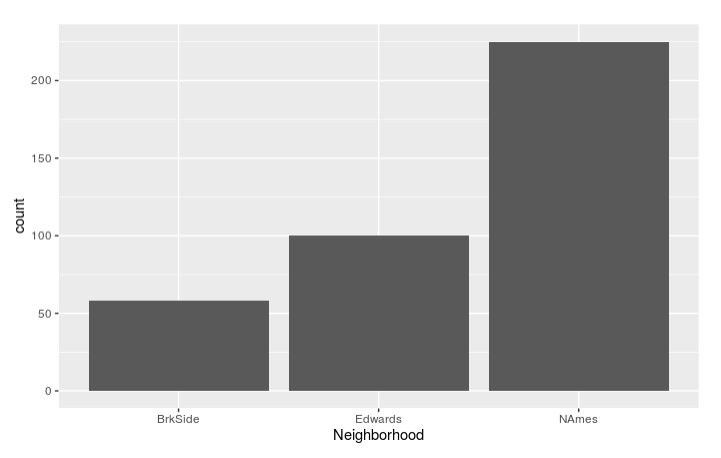
Plot 1, non-log transformed histogram of sale price



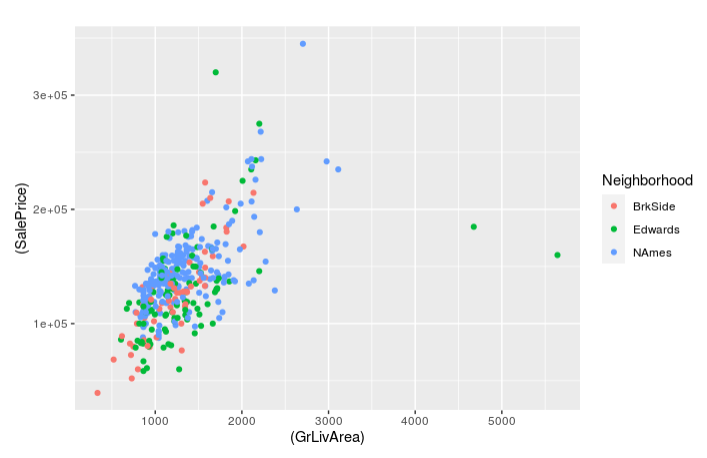
Plot 2, non-log transformed histogram of square footage



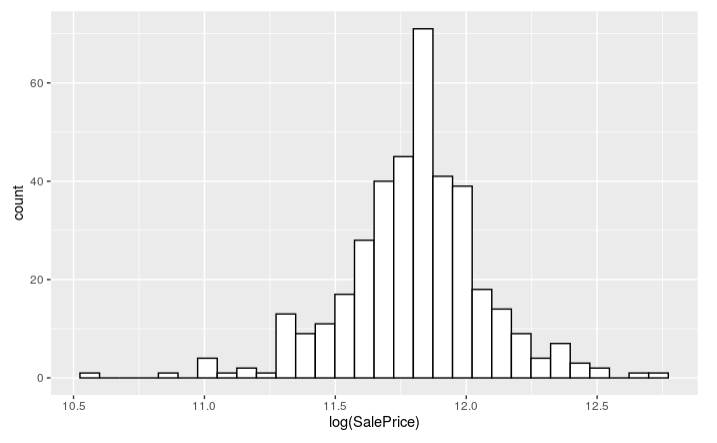
Plot 3, bar plot of house counts in each neighborhood



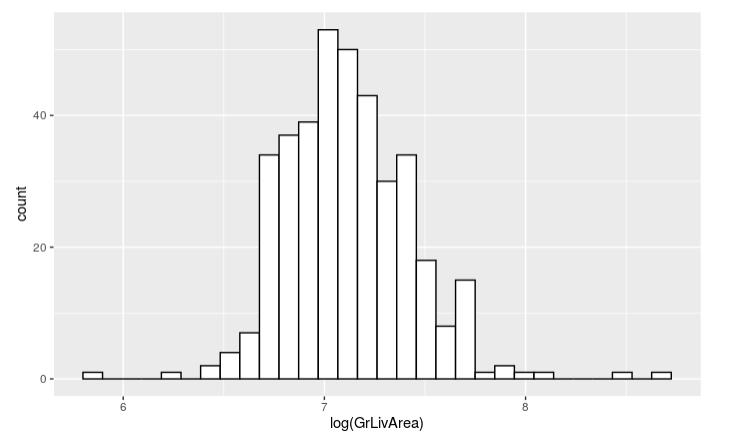
Plot 4, scatter plot of non-transformed data of sale price with neighborhood and square footage



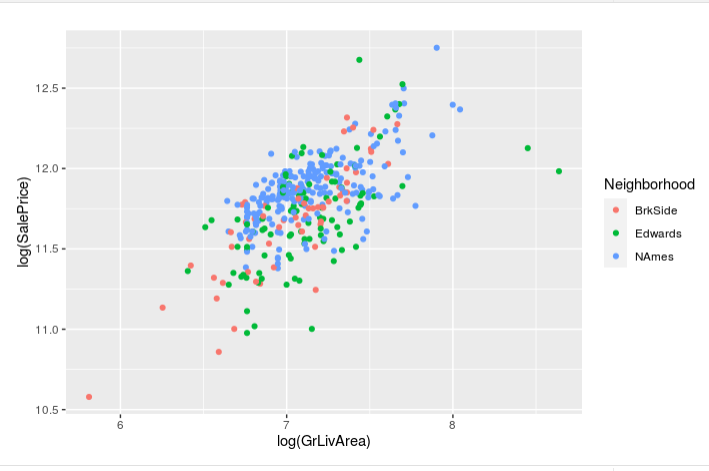
Plot 5, histogram of log-transformed data of house prices



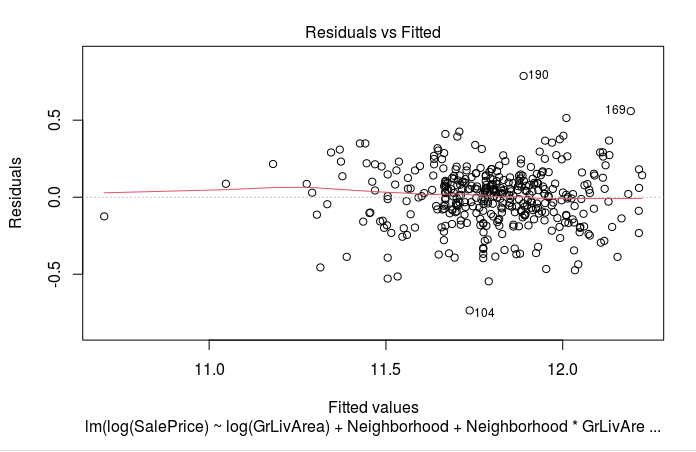
Plot 6, histogram of log-transformed square footage



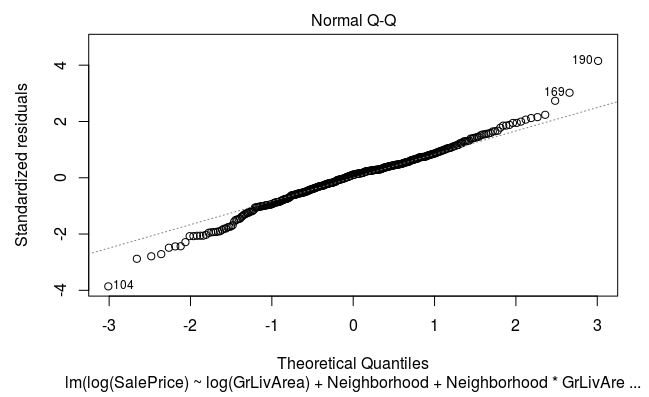
Plot 7, scatter plot of log-transformed data of sale price, neighborhood, and square footage



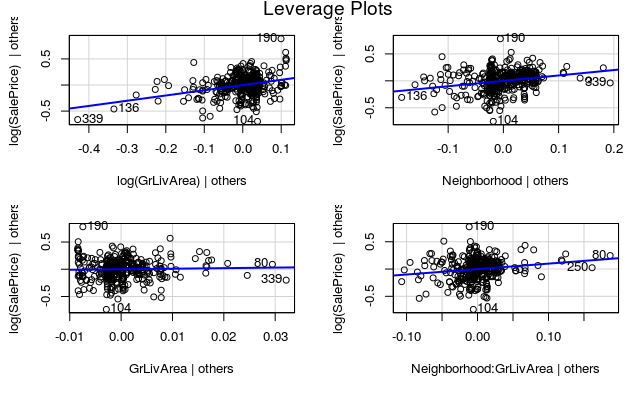
Plot 8, residual plot for LM model with outliers included



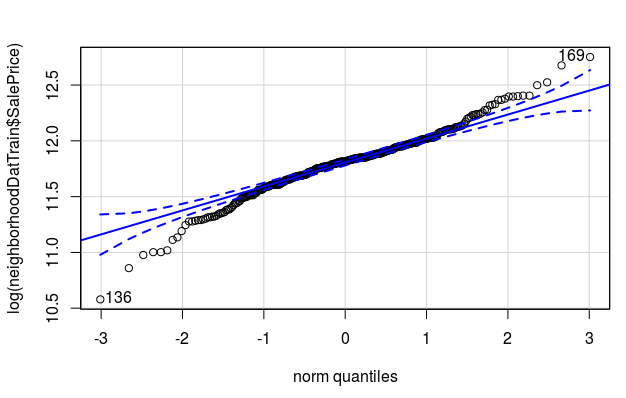
Plot 9, Quantile-Quantile plot with outliers included



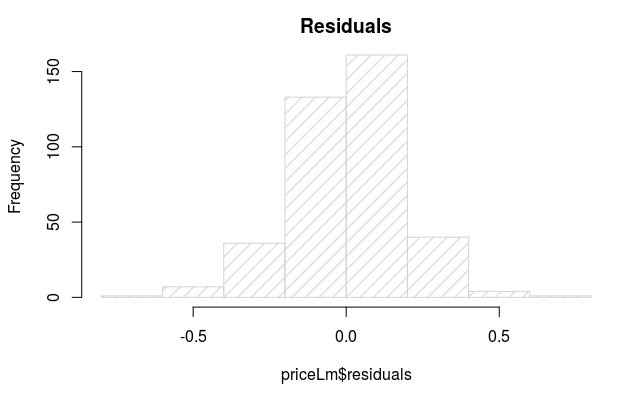
Plot 10, Leverage plots for model with outliers



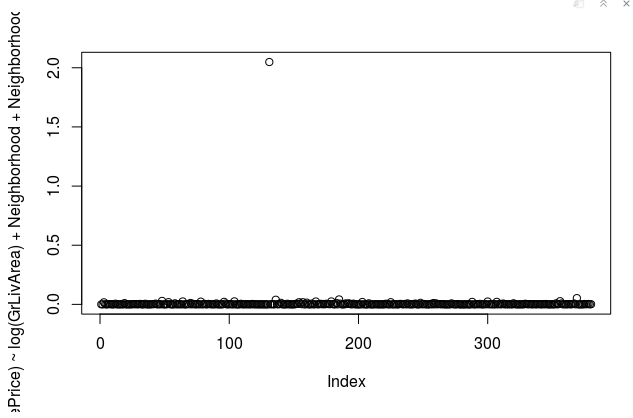
Plot 11, More detailed Quantile-Quantile plot with outliers included



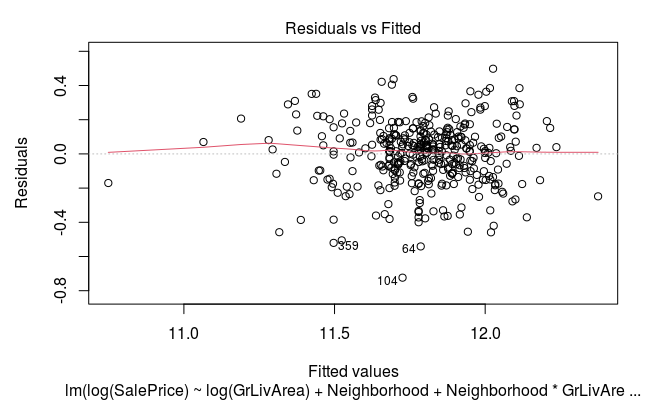
Plot 12, Residual Plot



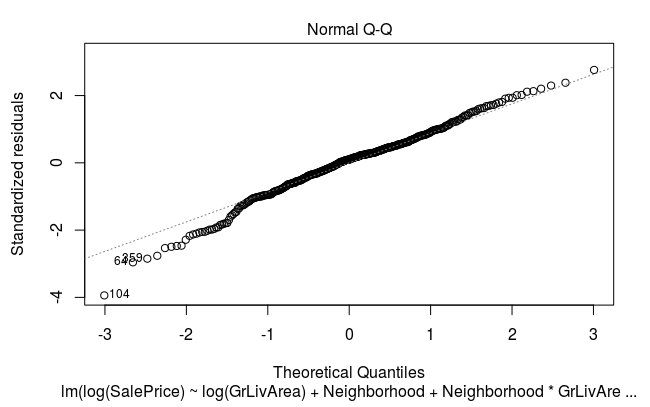
Plot 13, Cook’s Distance of non outlier removed data



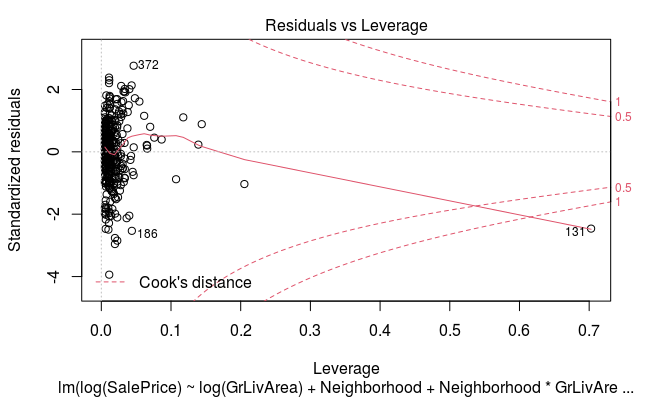
Plot 14, Residuals vs fitted on data with outliers removed



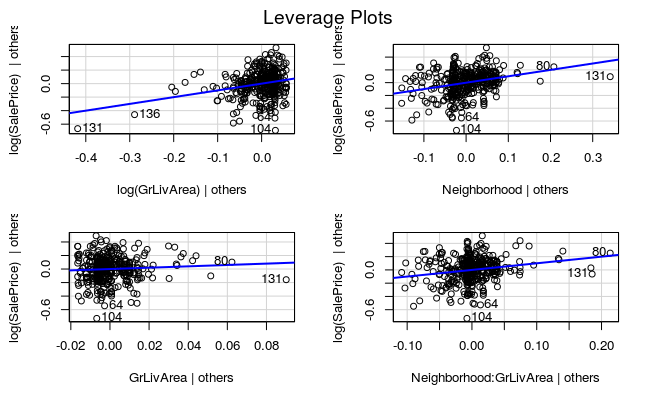
Plot 15, Quantile-Quantile plots of data with outliers removed



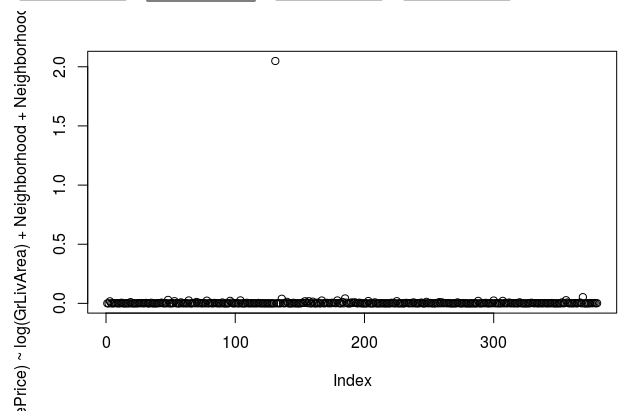
Plot 16, Cook’s distance plotted with data without outliers



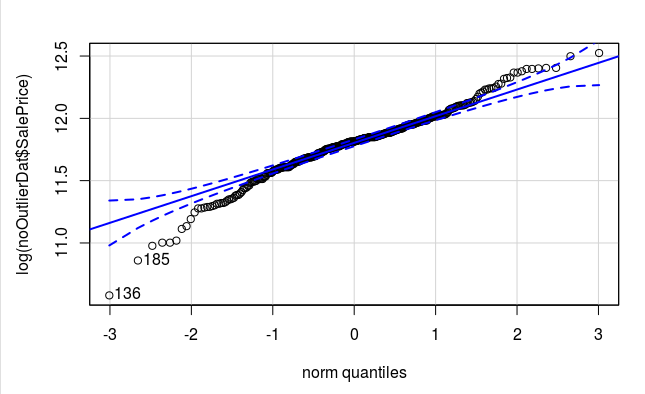
Plot 17, Leverage plots with outliers removed



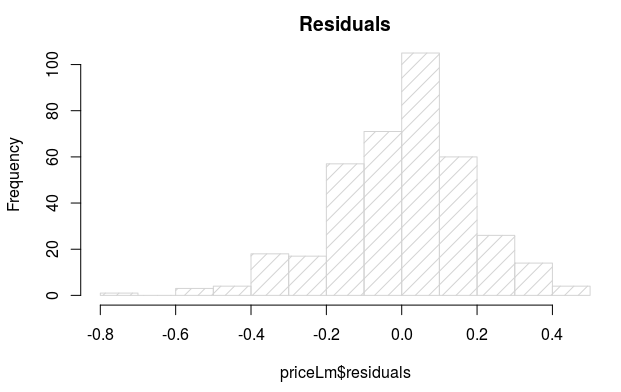
Plot 18, Cook’s distance (again) with outliers removed



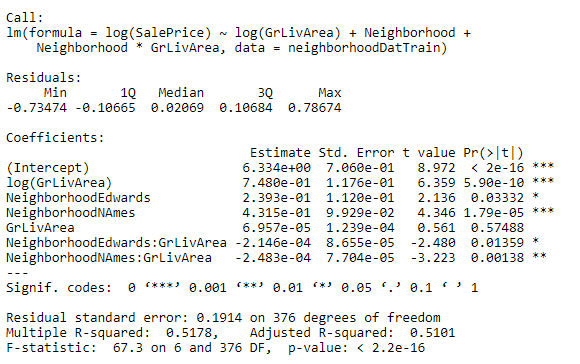
Plot 19, A more detailed quantile-quantile plot with outliers removed



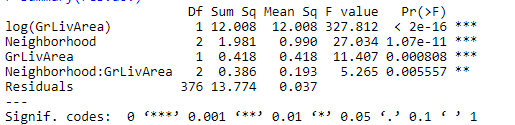
Plot 20, Residual plot of data with removed outliers



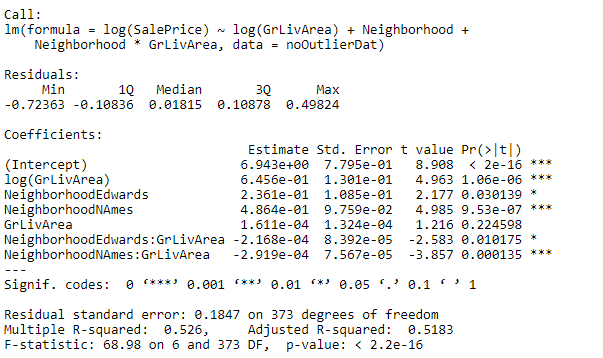
Plot 21, Model without outliers removed:



Plot 22, ANOVA table with outliers



Plot 23, Model with outliers removed



Plot 23, ANOVA table with outliers removed

