Analysis Q1 of KAGGLE project

Jodi Pafford, Tori Wheelis, Sara Zaheri

**Introduction**

This is where the introduction will go. We will add some sort of overview of our purpose and the outcome desired.

**Data Description**

This is where the data description will go. (Where did the data come from?, How big is it?, How many observations?, Where can we find out more?, What are the specific variables that we need to know with respect to your analysis?)

Analysis Question 1:

**Restatement of Problem**

This is where we will restate the problem in question 1.

**Build and Fit the Model**

In order to fit the model, we tried 2 variations of 4 different data sets:

* Model 1: No transformation, no outliers, and no interactions
* Model 2: Log-Linear transformation model taking the log of SalePrice only, no outliers, and no interactions
* Model 3: Linear-Log transformation, taking the log of GrLIvARea only, no outliers, and no interactions
* Model 4: Log-Log model, taking the log of SalePrice and GrLIvArea, no outliers, no interactions

In doing so, we noticed that there appear to be some outliers with the data. Therefore an analysis was done to evaluate the log-log plot for outliers. There were two outliers found that had a studentized residual higher than |5|. These data were removed and the top two models were run again (No Transformation model and Log-Log model).

Things to include per assignment paper:

Checking Assumptions

Residual Plots

Influential point analysis (Cook’s D and Leverage)

**Comparing Competing Models**

Things to include per assignment paper:

Adj R squared

Internal CV Press

**Parameters**

Things to include per assignment paper:

Estimates

Interpretation

Confidence Intervals

**Conclusion**

This is where we will summarize our findings from question 1.

Figure 3, Figure 4, Figure 5, and Figure 6 are presenting the results for the above models, respectively:

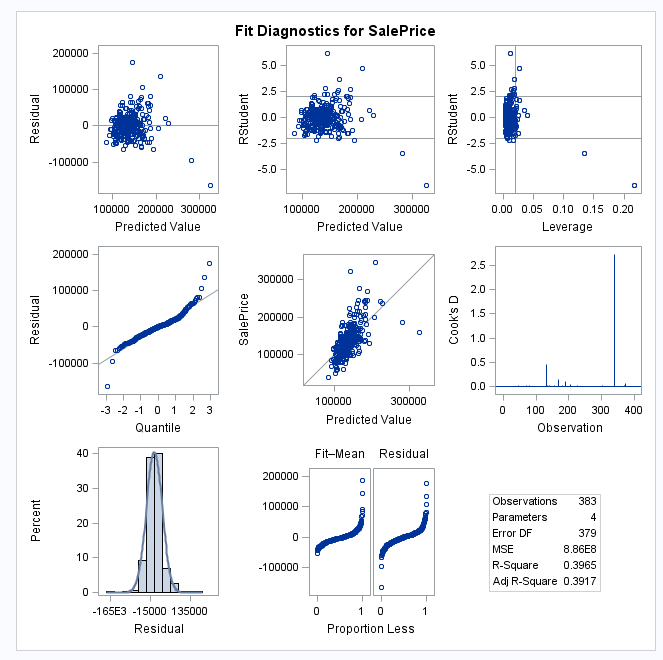
 

Figure 3. Result of the model with no transformation, no outlier, and no interaction effect.

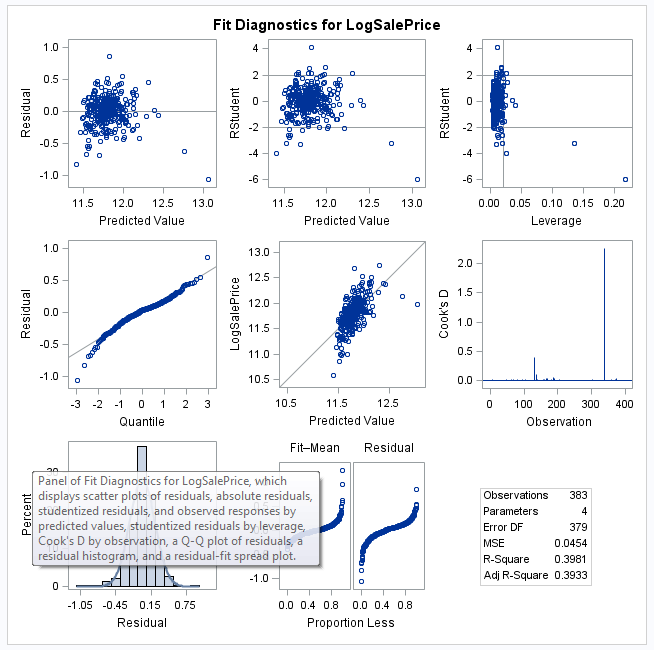
 

Figure 4. Result of the Log-Linear model, no outlier, and no interaction effect.

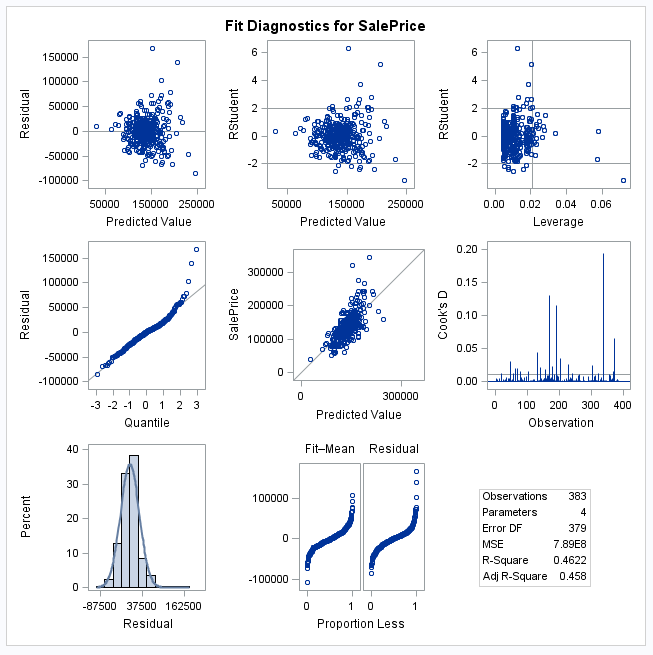
 

Figure 5. Result of the Linear-Log model, no outlier, and no interaction effect.

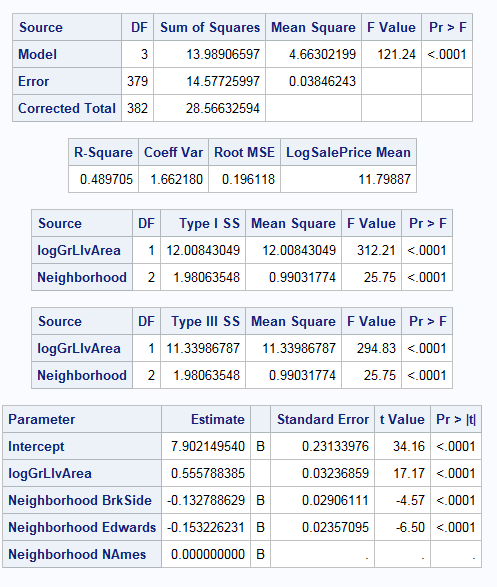
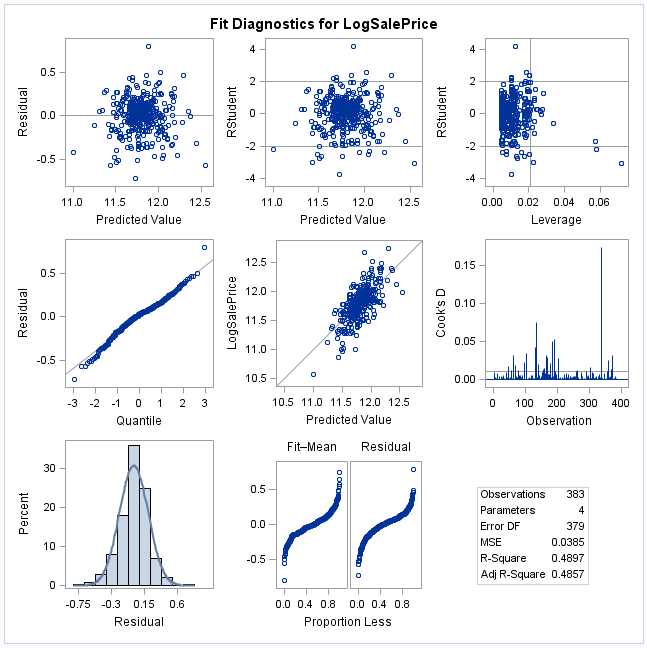
 

Figure 6. Result of the Log-Log model, no outlier, and no interaction effect.

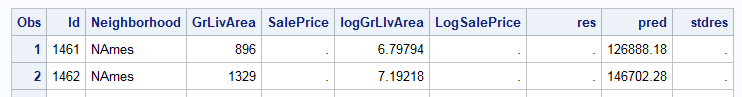
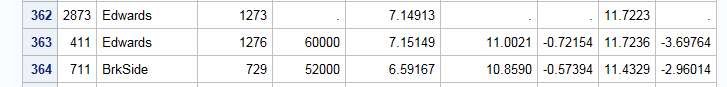
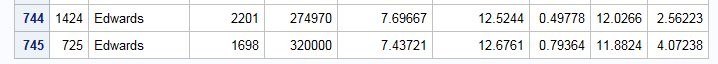
  

Figure 7. Screening possible outlier in data.

These results indicate at least two possible outliers with studentized residuals higher |5|, (id=411, and id =725). Please note that I used the output of the Log-Log model for distinguishing the outliers.

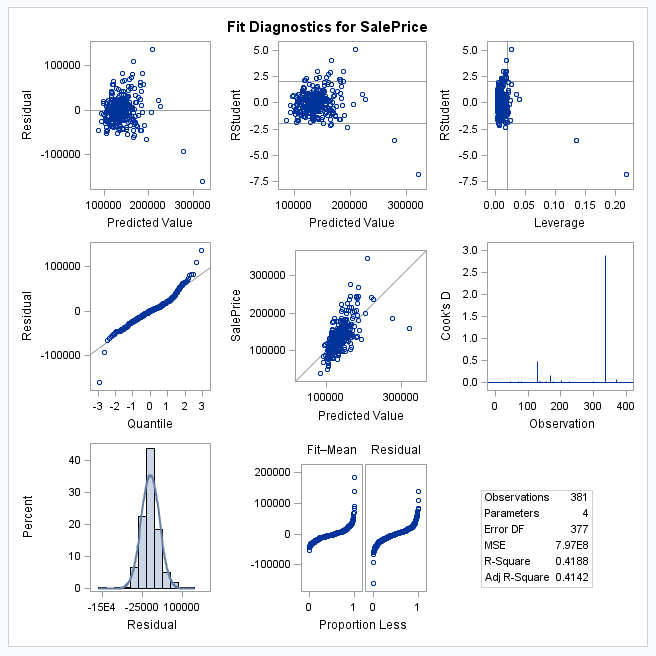


Figure 8. output of no transformation model with removed two outliers and no interaction.

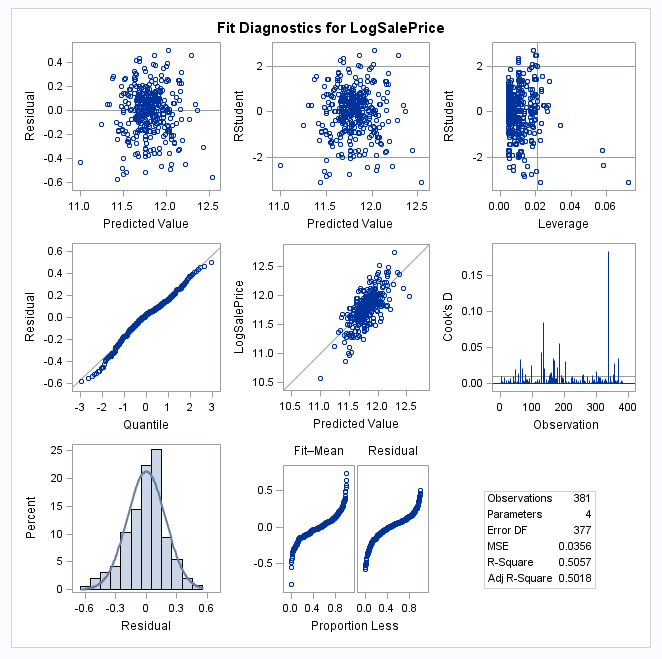
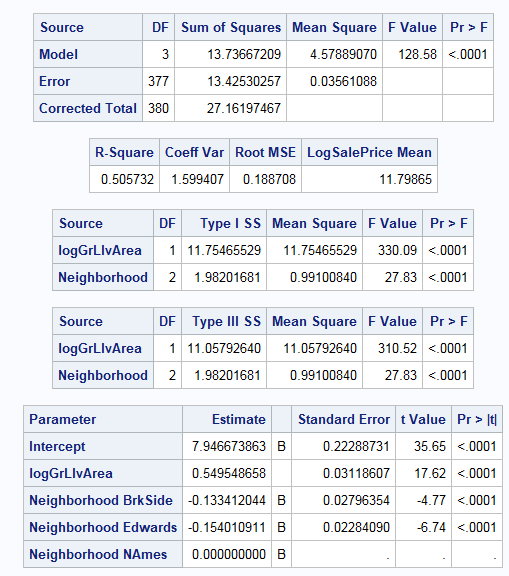


Figure 8. Output of Log-Log transformation model with removed two outliers and no interaction.

Residual in these two models are propperly fitted with normal distribution and show no more outliers, but there is still a great problem with these two models which is low r-sqaure. However the r-square in log-log model is higher than the original data modeling, but this r-square is however low. Therefore, I tested these models with interactions according to the following code:

\*no transformation & removed outlier = 2 & Interaction;

**proc** **glm** data=Q1\_2 plots=all;

class Neighborhood ;

model SalePrice = GrLIvArea | Neighborhood / solution clparm;

output out = Results1 r=res p=pred student=stdres;

**run**;

\*log-log model & removed outlier = 2 & Interaction;

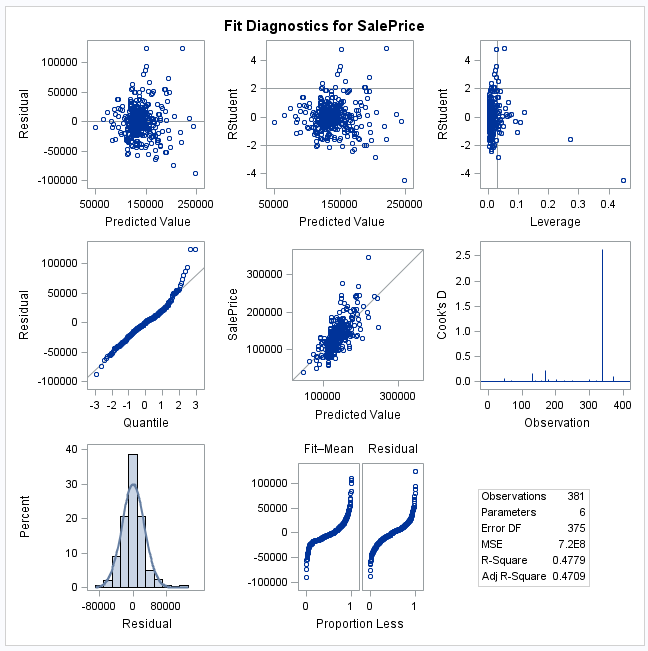
**proc** **glm** data=Q1\_2 plots=all;

class Neighborhood ;

model LogSalePrice = LogGrLIvArea | Neighborhood / solution clparm ;

output out = Results r=res p=pred student=stdres;

**run**;

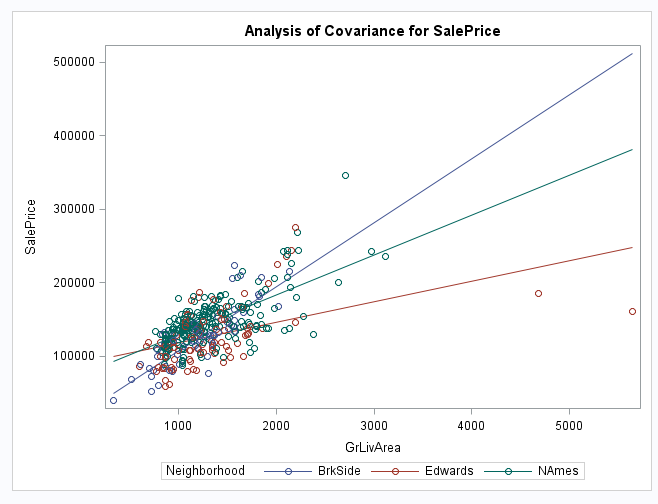


Figure 9. No transformation model with two removed outliers and interaction effects.

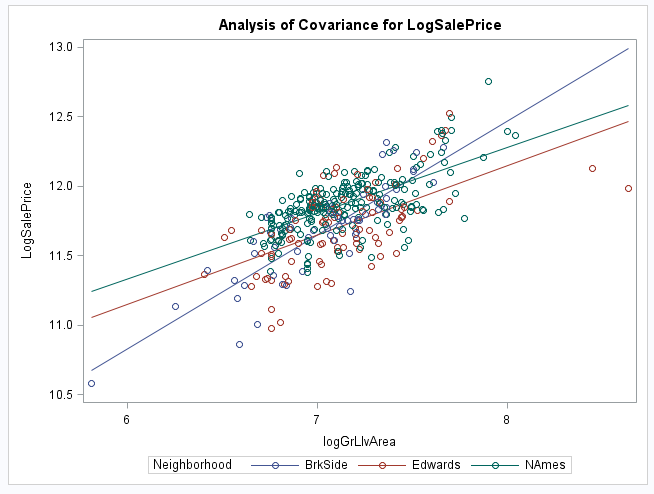
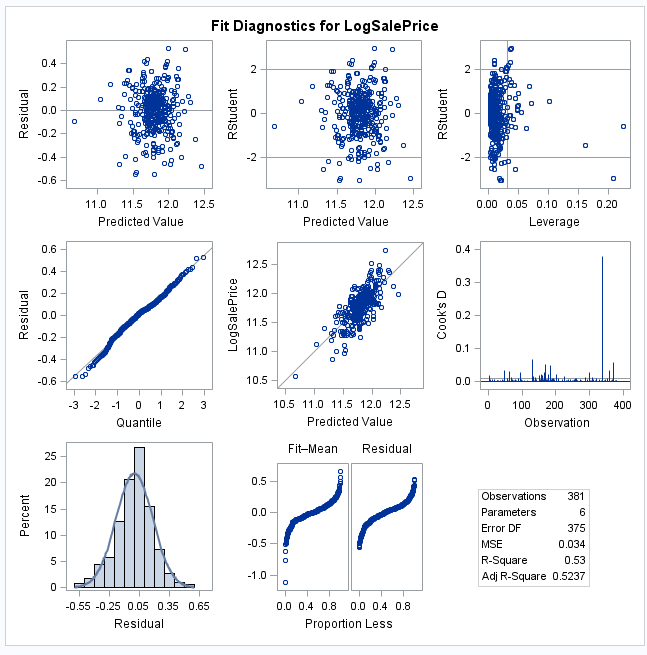
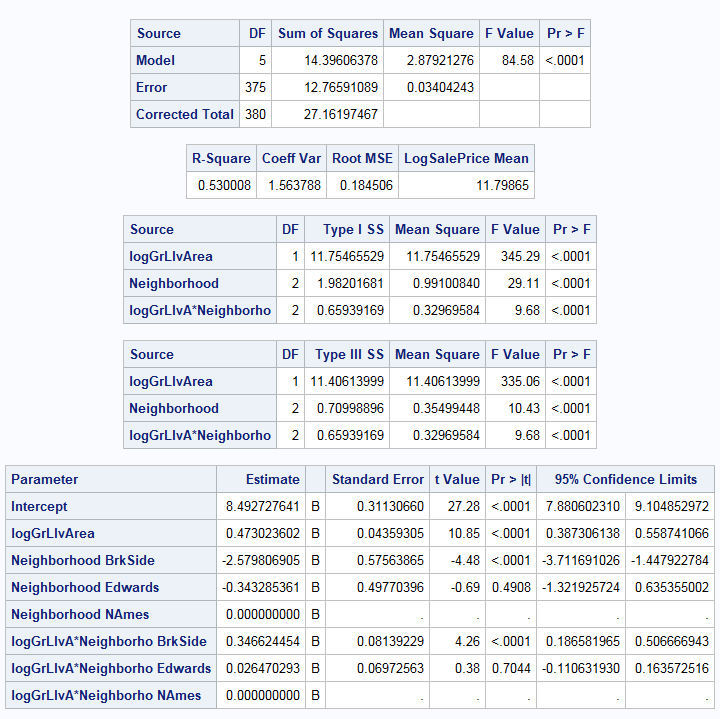


Figure 10. Log-Log transformation model with two removed outliers and interaction effects.

The output of the no transformation and log-log models indicates that interaction effect should be included in the models because the lines for each Neighborhood crossed one another. Also, residuals showed no problem with distriution or outlier or other possible probels. The r-square in log-log moel is 0.53 which is clealry hgiher than that for original data.

All in all, I think that the best model for modeling these data is Log-Log transformed method, with two removed outliers (id =411, and id=725) and including interaction efects in the model.

In this stage, we can go farther and predict all SalePrice in the test set by simply using following code:

\*log-log model & removed outlier = 2 & Interaction;

**proc** **glm** data=Q1\_2 plots=all;

class Neighborhood ;

model LogSalePrice = LogGrLIvArea | Neighborhood / solution clparm ;

output out = Results r=res p=pred student=stdres;

**run**;

**proc** **print** data = Results;

**run**;