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| (id=411 and id=725). These data were removed, and the top two models were run again (No Transformation model and Log).

Residuals in these two models are properly fitted with normal distribution and show no more outliers, but there is still a great problem with the two models due to the low r-square. Due to this, we tested these models with interactions. The output for these 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 distribution or outlier or other possible problems. The r-square in log-log model is 0.53 which is higher than that for the original data.

Based on the above information, we will proceed with the Log-Log transformed model with two removed outliers (id=411, and id=725) and include interaction effects.

Things to include per assignment paper:

Checking Assumptions **(done but may need rewording to suite Dr. South)**

Residual Plots **(Done)**

Influential point analysis (Cook’s D and Leverage)- ***Has this been addressed?***

**Comparing Competing Models**

Things to include per assignment paper:

Adj R squared ***(We included this above??)***

Internal CV Press ***– Has this been addressed?***

**Parameters**

Things to include per assignment paper:

***(Need to add all info below – can probably do some summary tables and such)***

Estimates

Interpretation

Confidence Intervals

**Conclusion**

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