DS6371 Project

**Data Description**

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. We have 1460 observations with 79 explanatory variables and it is collected from various neighborhoods in Ames, Iowa.

**Analysis Question #1**

1. **Problem Statement:** Century 21 Ames only sells houses in the NAmes, Edwards and BrkSide neighborhoods and would like to simply get an estimate of how the sale price of the house is related to the square footage of the living area of the house and if it depends on which neighborhood the house is located in.
2. **Build and Fit the model:**

**log(SalePrice) = b0 + b1log(sqft) + b2(BrkSide) + b3(Edwards) + b4(log(sqft)\*BrkSide) + b5(log(sqft)\*Edwards)**

1. **Checking assumptions:**
   1. When plotting the scatterplot (appendix: 1a) of Sale Price vs Square Footage, we observed a large cluster of points between 1000-2000 square feet which provided visual evidence against normality assumption. As values over 2,000, we observed higher spread between subpopulations. Two homes out of the 383 homes observed, stood out at >4,000 sqft. We examined Cook’s D and residual plots prior to moving forward with the restricted range approach (appendix: 1b). In conclusion, we decided to move forward with a Log/Log transformation of both Sale Price and Square Footage to deal with the concerns of normality, heteroscedasticity, and influential points.
   2. The plot (appendix: 1c) is the initial scatterplot after a Log/Log transformation was imposed. Now we don’t observe any tight clusters around certain focal points. The level of heteroscedasticity has decreased and allows our team to move forward with the assumption of constant variance across each subpopulation of LogSqft; however, it is important to visually note the appearance of the two outliers in the Edwards neighborhoods that still seem to be apparent. Further investigation with residual and Cook’s D plots will be needed on those 2 influential points once our model is built to see how we need to address those concerns.
2. Model Build
   1. **log(SalePrice) = b0 + b1log(sqft) + b2(BrkSide) + b3(Edwards) + b4(log(sqft)\*BrkSide) + b5(log(sqft)\*Edwards)**
      1. We included interaction terms to give linear regression model flexible slopes and intercepts for each neighborhood the house is located in.
3. Fit Model with 2 influential points included (code, output and discussion)

See Appendix 1d for code and output

**log(SalePrice) = 8.493 + 0.473\*log(sqft) – 2.580\*(BrkSide) – 0.486\*(Edwards) + 0.347(log(sqft)\*BrkSide) + 0.047(log(sqft)\*Edwards)**

Note here that both the slope and intercept for Edwards and NAmes neighborhoods don’t have enough evidence to suggest they are statistically different (p-values = 0.3481,0.5203, respectively). To overcome this, we could be to rerun the data combining those two areas. Alternatively, we can observe the influential points and see if the collinearity between those two groups can be explained by filtering out the outliers.

1. Checking Assumptions with 2 influential points (residuals, influential point analysis)

See Appendix 1e for plot visuals regarding assumption analysis below

**Residual Plot**: A nice scattered cloud around 0 doesn’t show outliers or lack of homoscedasticity

**Studentized Residual Plot**: Identifies 2 concerning points with predicted values 11.75 and 11.80. We will explore further as this may provide some evidence against the normality assumption.

**Histogram of Residuals**: The histogram of residuals 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 normality

**Leverage:** Shows 2-3 observations with high residual and high leverage that may be affecting our model. There is one point in particular near leverage = 0.20 that requires dire examination.

**Cook’s D:** The Cook’s D has one extreme point near 0.4. This home needs further examination.

**The model is a reasonable fit but needs to be re-ran to deal with the 1-2 opposing outliers that seem to be showing a lot of effect on our model. We created a restricted range of homes between 0-4000 square feet and reran our model.**

1. Fit Model without 2 influential points (code, output and discussion)

See Appendix 1f for code and output

**log(SalePrice) = 8.493 + 0.473\*log(sqft) – 2.580\*(BrkSide) – 1.570\*(Edwards) + 0.347(log(sqft)\*BrkSide) + 0.200(log(sqft)\*Edwards)**

It is important to note here that both the slope and intercept for Edwards and NAmes neighborhoods now have enough evidence to suggest they are statistically different (p-value = 0.0077,0.0154, respectively) while all other variables still maintain their statistical significance. This further corroborates our initial claim that those 2 outlying points may have explained a level of collinearity between the Edwards and NAmes neighborhoods.

1. Checking Assumptions without 2 influential points (influential point analysis)

See Appendix 1g for plot visuals regarding assumption analysis below

**Leverage:** The leverage plot looks much better. Not much argument for any serious outlying observations that may be considered egregious enough to do further analysis on.

**Cook’s D:** The Cook’s D chart also looks much better. First thing to note is how the scale changed from 0.1 increments to 0.01 increments. Creating that range allowed for a much more balanced chart on both leverage and residuals.

**This restricted range model is the most reasonable fit and will move forward as such.**

1. Comparing Competing Models (with and without 2 outlying observations)

See Appendix 1h and 1i for parts a. and b. for competing model outputs respectively

* 1. With 2 observations
     1. Adj r^2 = 0.5056
     2. Root MSE = 0.19228
  2. Without 2 observations (restricted range model)
     1. Adj r^2 = 0.5216
     2. Root MSE = 0.18916

1. Parameters (estimates/interpretation, confidence intervals)
   1. Parameter Estimates

**Log(SalePrice) = 8.493 + 0.473\*log(sqft) – 2.580\*(BrkSide) – 1.570\*(Edwards) + 0.347(log(sqft)\*BrkSide) + 0.200(log(sqft)\*Edwards)**

**Individual Regression Equations**

* + 1. NAmes (When BrkSide = 0 and Edwards = 0)

Log(SalePrice) = 8.493 + 0.473\*log(sqft)

* + 1. BrkSide(When Edwards = 0 and BrkSide = 1)

Log(SalePrice) = 5.913 + 0.820\*log(sqft)

* + 1. Edwards(When BrkSide = 0 and Edwards = 1)

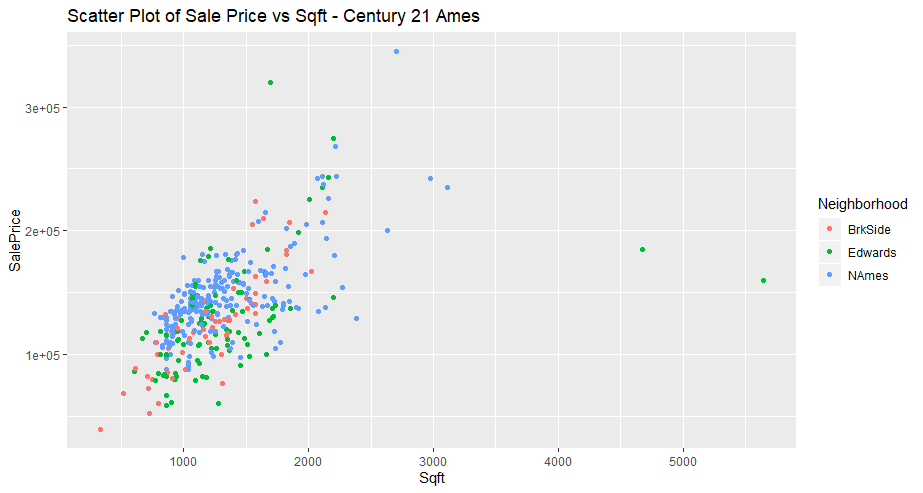
Log(SalePrice) = 6.923 + 0.673\*log(sqft)

* 1. Interpretation
     1. NAmes: The log sale price of NAmes homes will have an intercept of 8.493 with a living space of 1sqft. A doubling in square feet for NAmes homes is associated with a 2^0.473 or 139% increase in median price, holding all variables fixed.
     2. BrkSide: The log sale price of BrkSide homes will have an intercept of 5.913 with a living space of 1sqft. A doubling in square feet for BrkSide homes is associated with a 2^0.820 or 177% increase in median sales price, relative to NAmes homes, holding all variables fixed.
     3. Edwards: The log sale price of Edward homes will have an intercept of 6.923 with a living space of 1sqft. A doubling in square feet for Edward homes is associated with a 2^0.673 or 159% increase in median sales price, relative to NAmes homes, holding all variables fixed.
  2. Confidence Intervals – See Appendix 1j
     1. The 95% confidence intervals for the coefficients are as follows:
        1. log(sqft) = (0.385, 0.561)
        2. BrkSide = (-3.740, -1.419)
        3. Edwards = (-2.721, -0.418)
        4. log(sqft)\*BrkSide = (0.183, 0.511)
        5. log(sqft)\*Edwards = (0.039, 0.362)
        6. intercept = (8.865, 9.120)

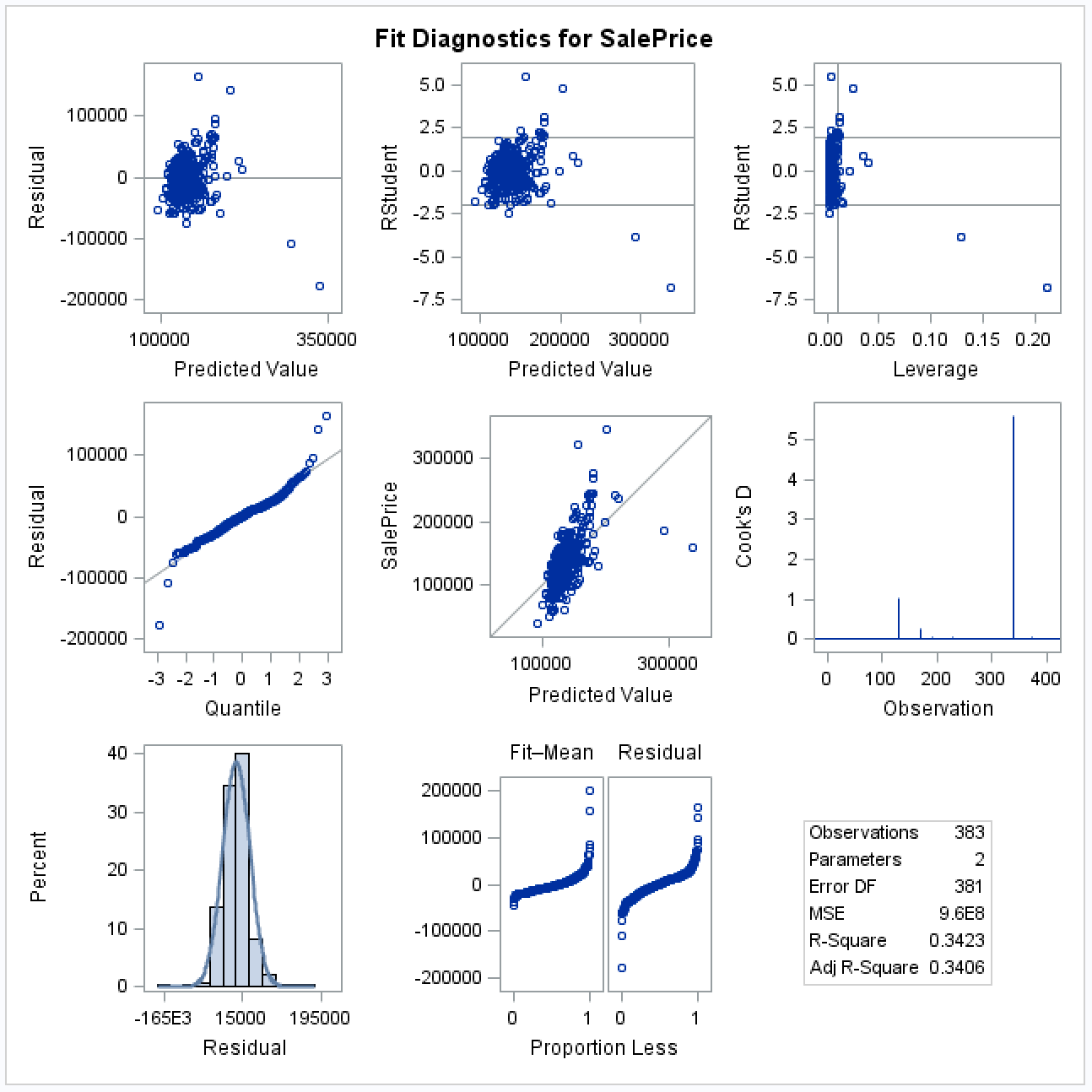
1. Conclusion
   1. An important point in this analysis were the homes with greater than 4000 square feet. Since we had 381/383 observations all in one subarea of square feet <4000 and only 2 in the outlying area, we moved forward to restrict those 2 data points from the set. When we re-ran our models with the restricted range, we saw that both the slopes and intercepts of Edwards and NAmes neighborhoods turned to be statistically significant regarding their terms being different (p-value = 0.0077,0.0154, respectively). This explains a level of collinearity that was being explained in the original model (including both points) between the NAmes and Edwards neighborhoods. For these reasons above, we decided to move forward with our restricted range log/log model. Moving forward, we found that log square feet was found to be significant when each of the 3 neighborhoods were added into the model. This indicates that each doubling of square feet, holding all variables fixed, results in an associated 1.39 multiplicative increase in median sales price for NAmes homes. For BrkSide homes, for every doubling of square feet, holding all other homes fixed, results in a 1.77 multiplicative increase in median sale price. Lastly, for every doubling of square feet in Edward homes, holding all other homes fixed, results in a 1.59 multiplicative increase in median sale price. Please refer to the “interpretation” point above to see the percent interpretation.

Appendix:

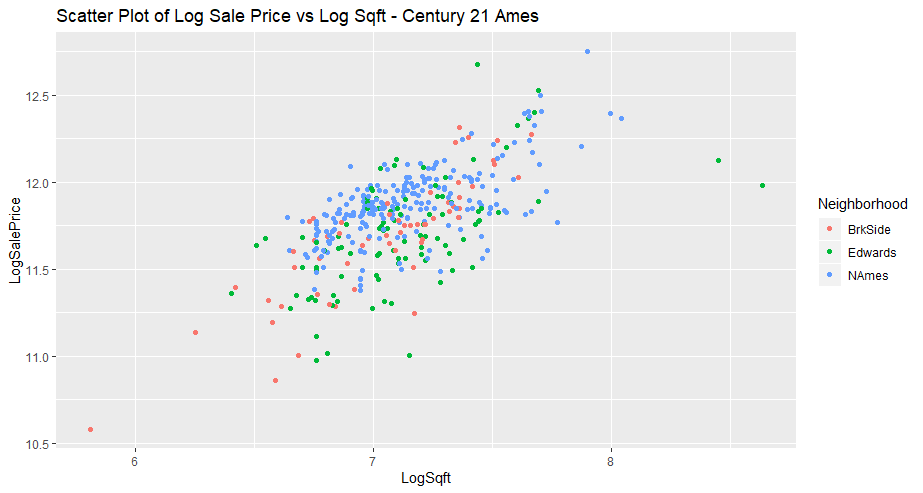
1a)



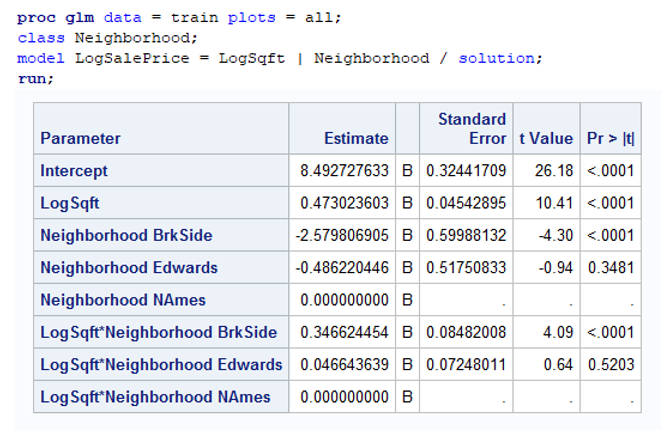
1b)



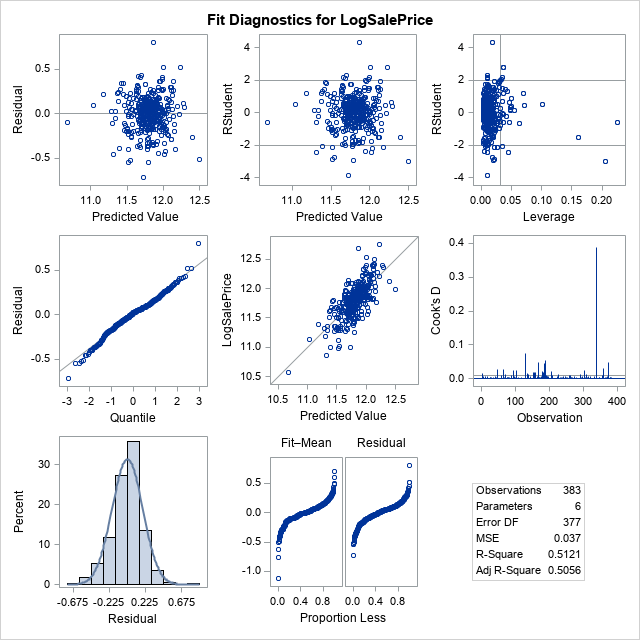
1c)



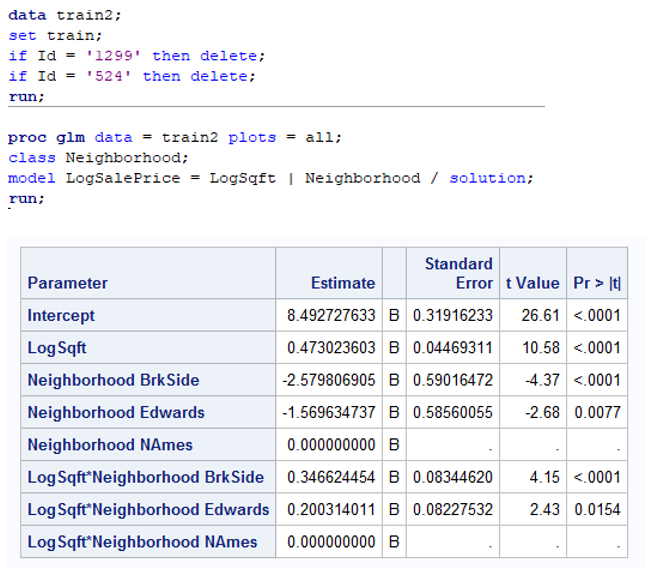
1d)



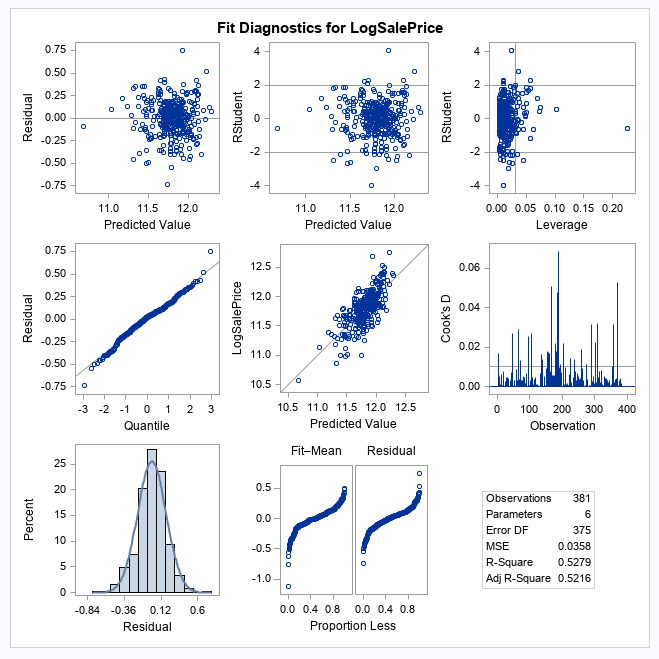
1e)



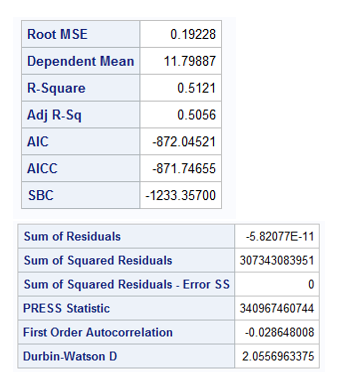
1f)



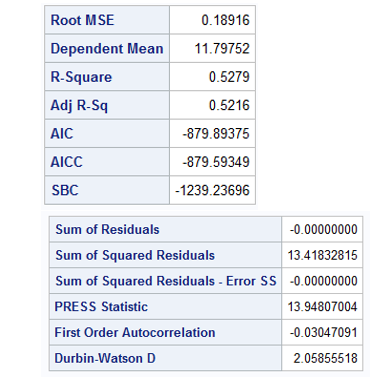
1g)



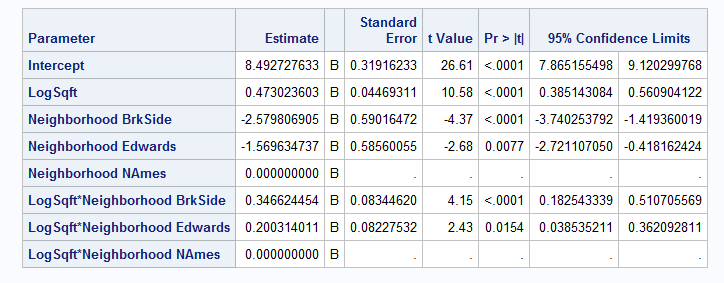
1h)



1i)



1j)



SAS OUTPUT

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*End Term Kaggle Project\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*Train Data Set\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**PROC** **IMPORT** OUT = WORK.train

DATAFILE = "U:\Apps.SMU\Documents\DS6371\_datafiles\KaggleProject\train.csv"

DBMS = CSV REPLACE;

DATAROW = **2**;

**RUN**;

**data** trainfiltered;

SET train;

IF (Neighborhood = 'NAmes') OR (Neighborhood = 'Edwards') OR (Neighborhood = 'BrkSide');

**RUN**;

/\*\*\*\*\*\*\*\*\*Test Data Set\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**PROC** **IMPORT** OUT = WORK.test

DATAFILE = "U:\Apps.SMU\Documents\DS6371\_datafiles\KaggleProject\test.csv"

DBMS = CSV REPLACE;

DATAROW = **2**;

**RUN**;

/\*proc GLM to check assumptions\*/

**proc** **glm** data = trainfiltered plots = all;

model SalePrice = GrLivArea;

**run**;

/\*Applying log-log transformation to check because observations are not very well distributed\*/

**data** trainfilteredlog;

set trainfiltered;

lGrLivArea=log(GrLivArea);

lSalePrice=log(SalePrice);

**run**;

/\*proc GLM to check assumptions again\*/

**proc** **glm** data = trainfilteredlog plots = all;

model SalePrice = GrLivArea;

**run**;

/\*Two extreme values found in leverage graph and cook's D graph which does not look like regular observation. Filtering those 2 values to check result again.\*/

**data** trainfiltered\_noOutlier;

SET trainfiltered;

IF (GrLivArea < **4000**);

**RUN**;

/\*proc GLM to check assumptions\*/

**proc** **glm** data = trainfiltered\_noOutlier plots = all;

model SalePrice = GrLivArea;

**run**;

/\*Applying log-log transformation (after extreme value removal) to check\*/

**data** trainfiltered\_outlierlog;

set trainfiltered\_noOutlier;

lGrLivArea=log(GrLivArea);

lSalePrice=log(SalePrice);

**run**;

/\*Adding interaction variable using Neighborhood \*/

**proc** **glm** data = traintestfilteredlog plots = all;

class Neighborhood (REF = "NAmes");

model lSalePrice = Neighborhood | lGrLivArea;

**run**;

/\*Now merging train and test data set\*/

**data** test;

set test;

SalePrice = **.**;

**run**;

**data** traintest;

set train test;

**run**;

/\*Merging complete\*/