Ames Housing Analysis

## Introduction

This research is conducted on behalf of Century 21 Ames. We will be analyzing housing data for Ames, Iowa collected between 2006 and 2010.

The original data set, collected by Dean De Cock, contains 2930 observations and a large variety of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing property values.

In the first analysis, we will be looking at the relationship between sale price and total square footage for each home in three specific neighborhoods: North Ames, Eastwood, and Brookside neighborhoods.

For the second analysis, we will be looking at all neighborhoods in Ames to predict the sale price of homes.

## Question 1

For this analysis we will be assessing the relationship between the square footage (sqft) of a property and the sale price in the North Ames, Eastwood and Brookside neighborhoods. Specifically, we will be examining a subset from the larger data pool, using 381 observations from the 3 neighborhoods of interest. Our model, taking in to account the neighborhoods as the interaction term, looks like this:

After initial assessment of the data (Figures 1 & 2), we deemed it best to perform a log of both sale price (log[sale]) and property square footage (log[sqft]). The results improved the normality and linearity of the data (Figures 1 & 3).

Even after log transformation, however, we found that there were still a small number of potential outliers, specifically in the Edwards neighborhood. We omitted these properties based on their studentized residual leverage score and Cook’s D and found that the linearity and normality of the data improved even more (Figure 4); we determined that the degree of improvement in the model were worth the omissions. We ran the regression, using North Ames as reference, and found interaction between all terms to be significant at the p = 0.05 level (Table 1 and Figure 5).

Using these results, our model now looks like this:

This model further gives us three equations for sale price, one for each neighborhood:

{logsale | Brookside} = 5.91 + 0.82 logsqft

{logsale | Edwards} = 6.92 + 0.67 logsqft

{logsale | North Ames} = 8.49 + 0.47 logsqft

We will address our findings and the implications of these formulae for each of the neighborhoods in question, in turn.

### Brookside

For Brookside, the model indicates that a doubling of the median square footage of a property in Brookside is associated with a 2.82 = 1.765 multiplicative change in the median sale price. The predicted median for the sale price of a property in this location increases by 76.5% for each doubling of square footage, with a 95% confidence interval for this relationship being in the range of (2.65 = 1.56, 2.98 = 1.97), or 56% and 97%.

Using our formula, a baseline home with 100 square feet would have a median sale price of $16,092; each doubling of the square footage would multiply this by 1.765 (thus, 200 square feet would be $28,403; 400, $50,131; and so forth).

### Edwards

For Edwards, the model indicates that a doubling of the median square footage of a property in Edwards is associated with a 2.67 = 1.59 multiplicative change in the median sale price. The predicted median for the sale price of a property in this location increases by 59% for each doubling of square footage, with a 95% confidence interval for this relationship being in the range of (2.508 = 1.42, 2.83 = 1.78), or 42% and 78%.

Using our formula, a baseline home with 100 square feet would have a median sale price of $22,144; each doubling of the square footage would multiply this by 1.59 (thus, 200 square feet would be $35,209; 400, $55,983; and so forth).

### North Ames

For North Ames, the model indicates that a doubling of the median square footage of a property in North Ames is associated with a 2.47 = 1.38 multiplicative change in the median sale price. The predicted median for the sale price of a property in this location increases by 38% for each doubling of square footage, with a 95% confidence interval for this relationship being in the range of (2.38 = 1.3, 2.56 = 1.47), or 30% and 47%.

Using our formula, a baseline home with 100 square feet would have a median sale price of $42,376; each doubling of the square footage would multiply this by 1.38 (thus, 200 square feet would be $58,479; 400, $80,701; and so forth).

### Conclusions

This was an observational study, thus we cannot make inferences to causation. There could be many other variables that contribute to the relationship between property square footage and sale price, and given the number of variables, most likely are.

Using our models to determine sales prices of some of the homes within the data provided, these formulae are rough estimates: this should be understood especially given the range of the confidence intervals, provided earlier. In addition, the omission of all factors other than square footage would suggest that the results from these estimates should be taken as precisely that – estimates – with further analysis required to produce more accurate results.

## Question 2

In this analysis, we will assess 4 different predictive models using data for all the neighborhoods in Ames.

## References

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

<https://ww2.amstat.org/publications/jse/v19n3/decock.pdf>

## Appendix

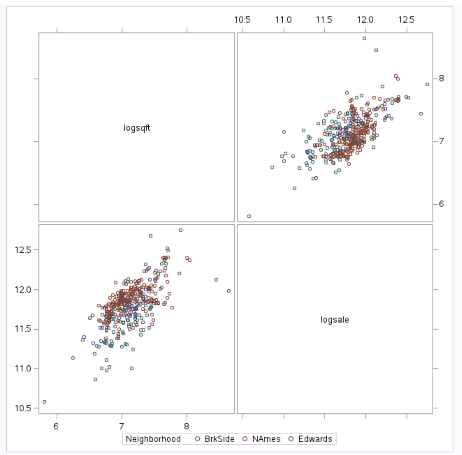
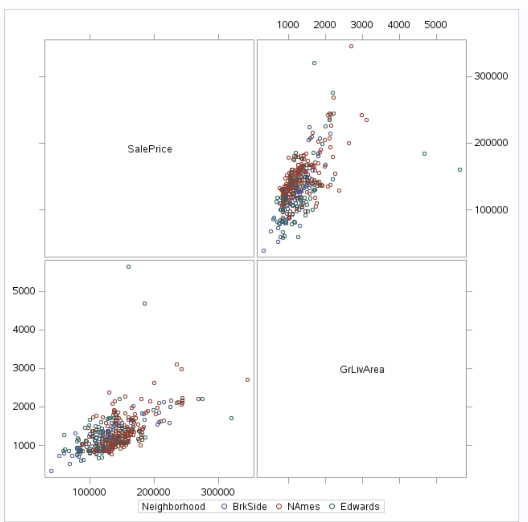


Figure 1 - Analysis 1 - EDA Pre and Post Log Log Transformation

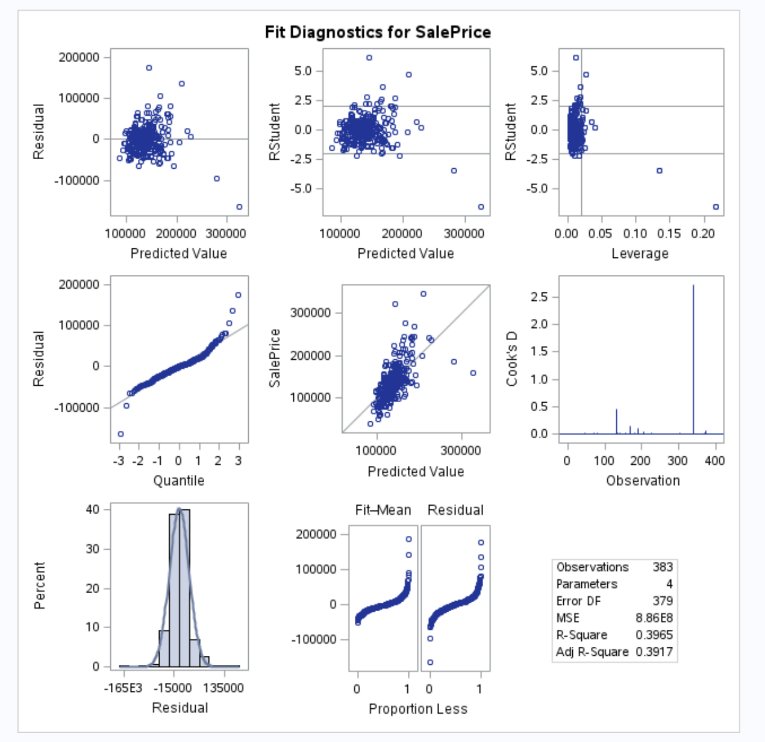


Figure 2 - Analysis 1 - Initial Fit Diagnostics

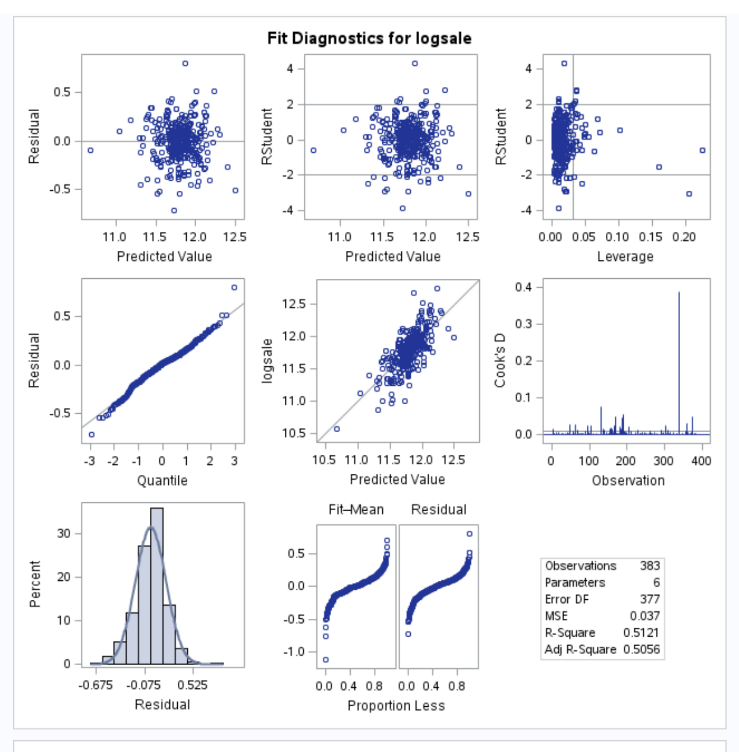


Figure 3- Analysis 1 - Post Log Log Fit Diagnostics

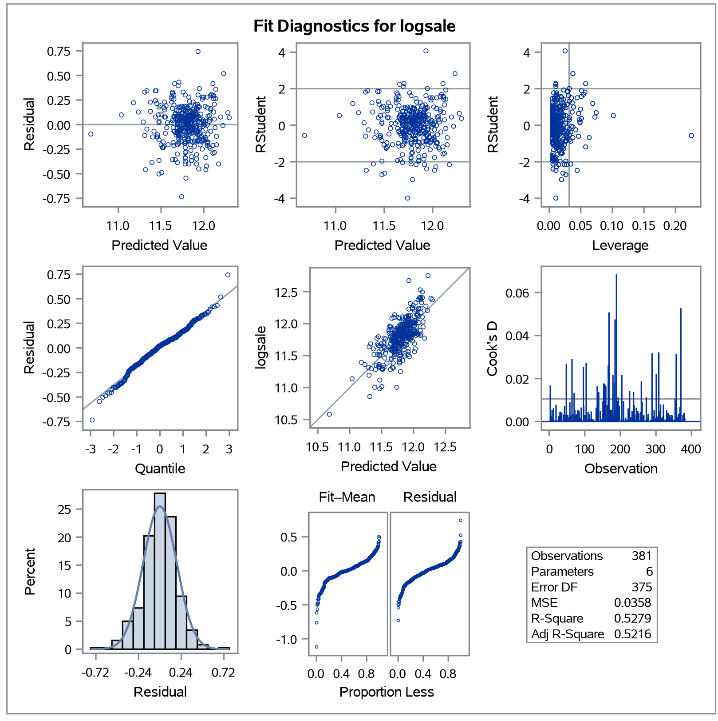


Figure 4 – Analysis 1 – Post Removal of Outliers

Table 1 - Question 1 - Result of analysis

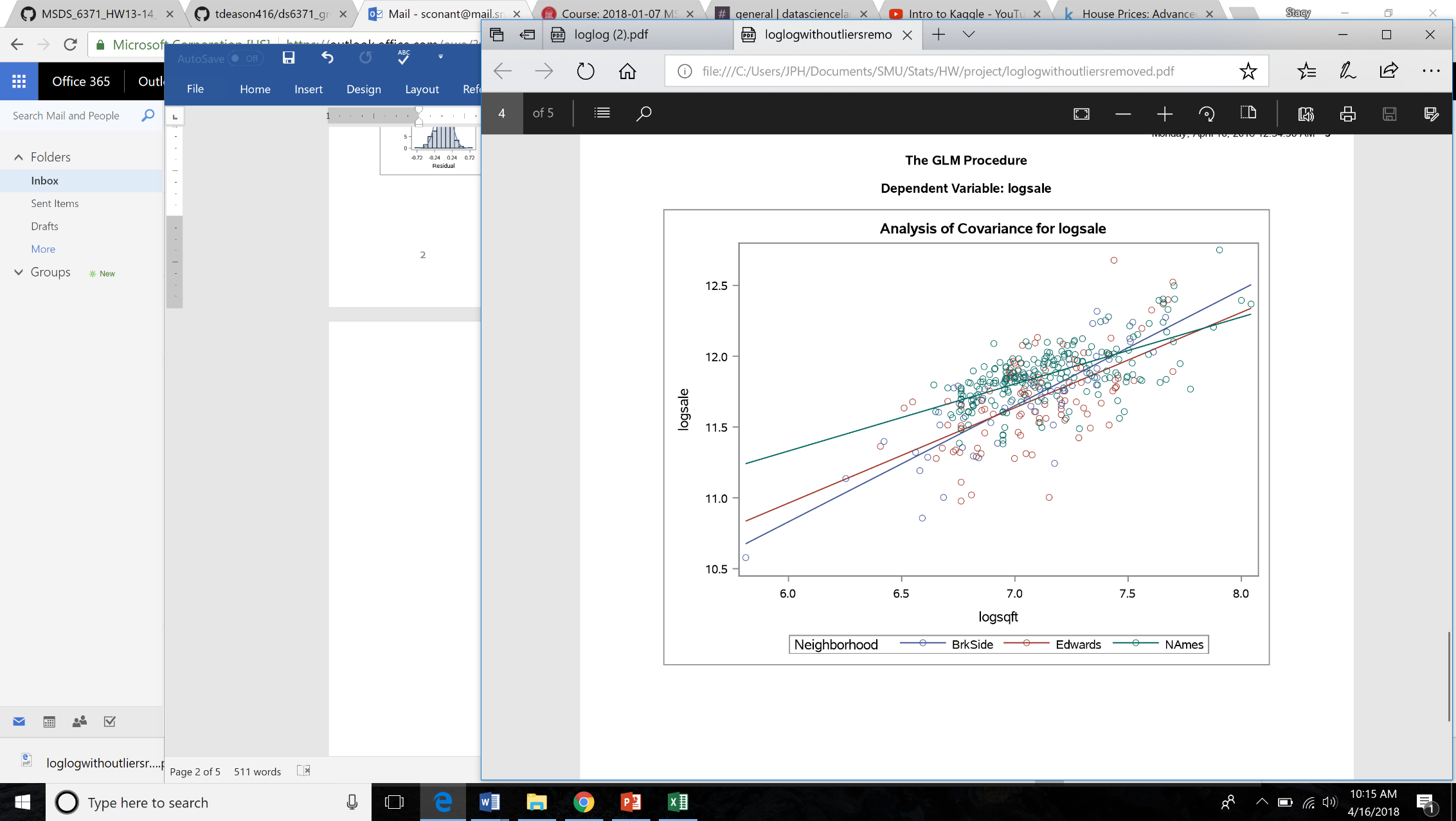
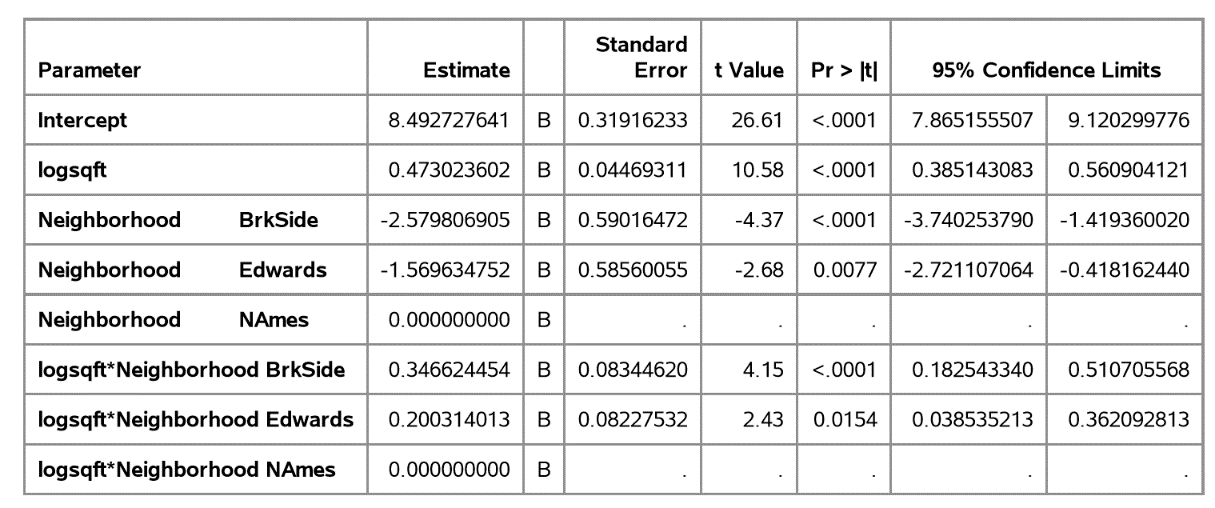


Figure 5 - Question 1 - Analysis Results

Code for Question 1:

\*uploading data;

data train1;

infile '/home/sconant0/sasuser.v94/train.csv' firstobs= 2 dlm=",";

input Id MSSubClass MSZoning$ LotFrontage LotArea Street$ Alley$ LotShape$ LandContour$ Utilities$ LotConfig$ LandSlope$ Neighborhood$ Condition1$ Condition2$ BldgType$ HouseStyle$ OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle$ RoofMatl$ Exterior1st$ Exterior2nd$ MasVnrType$ MasVnrArea ExterQual$ ExterCond$ Foundation$ BsmtQual$ BsmtCond$ BsmtExposure$ BsmtFinType1$ BsmtFinSF1 BsmtFinType2$ BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating$ HeatingQC$ CentralAir$ Electrical$ FirstFlrSF SecondFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual$ TotRmsAbvGrd Functional$ Fireplaces FireplaceQu$ GarageType$ GarageYrBlt GarageFinish$ GarageCars GarageArea GarageQual$ GarageCond$ PavedDrive$ WoodDeckSF OpenPorchSF EnclosedPorch ThreeSsnPorch ScreenPorch PoolArea PoolQC$ Fence$ MiscFeature$ MiscVal MoSold YrSold SaleType$ SaleCondition$ SalePrice;

run;

\*uploading data;

data test1;

infile '/home/sconant0/sasuser.v94/test.csv' firstobs = 2 dlm=",";

input Id MSSubClass MSZoning$ LotFrontage LotArea Street$ Alley$ LotShape$ LandContour$ Utilities$ LotConfig$ LandSlope$ Neighborhood$ Condition1$ Condition2$ BldgType$ HouseStyle$ OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle$ RoofMatl$ Exterior1st$ Exterior2nd$ MasVnrType$ MasVnrArea ExterQual$ ExterCond$ Foundation$ BsmtQual$ BsmtCond$ BsmtExposure$ BsmtFinType1$ BsmtFinSF1 BsmtFinType2$ BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating$ HeatingQC$ CentralAir$ Electrical$ FirstFlrSF SecondFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual$ TotRmsAbvGrd Functional$ Fireplaces FireplaceQu$ GarageType$ GarageYrBlt GarageFinish$ GarageCars GarageArea GarageQual$ GarageCond$ PavedDrive$ WoodDeckSF OpenPorchSF EnclosedPorch ThreeSsnPorch ScreenPorch PoolArea PoolQC$ Fence$ MiscFeature$ MiscVal MoSold YrSold SaleType$ SaleCondition$;

run;

\*adding sale price to test set;

data test2;

set test1;

SalePrice = .;

;

\*combining data set;

data test3;

set train1 test2;

run;

proc print data=test3; run;

\*subsetting data for question 1 with only neighborhoods of interest;

data test4;

set test3;

if Neighborhood = "NAmes" or Neighborhood = "Edwards" or Neighborhood = "BrkSide"; run;

proc print data=test4; run;

\*exporting cvs;

proc export data=test4

outfile='/home/sconant0/sasuser.v94/test4.csv'

dbms=csv; run;

\*EDA of data set for sale price and sqft;

proc sgscatter data = test4;

matrix SalePrice GrLivArea / group=Neighborhood; run;

\*run glm on raw data;

proc glm data=test4 plots=ALL;

class Neighborhood;

model SalePrice = GrLivArea | Neighborhood / solution clparm; run;

\*log transform both variables;

data test4log;

set test4;

logsale = log(SalePrice);

logsqft = log(GrLivArea); run;

\*re-run EDA on log log;

proc sgscatter data = test4log;

matrix logsqft logsale / group=neighborhood; run;

\*remove 2 outliers in Edwards: Id 524 and 1299;

data test4log;

set test4;

logsale = log(SalePrice);

logsqft = log(GrLivArea);

if \_n\_ = 131 or \_n\_ = 339 then delete; run;

proc print data=test4log; run;

\*re-run glm on log log data with outliers removed;

proc glm data=test4log plots=ALL;

class Neighborhood;

model logsale = logsqft | Neighborhood / solution clparm;