**MSDS 6371 Group Project - D&B All-Stars**

Vanessa Torres

Michael Landrum

Travis Deason

Brian Coari

# Introduction

Introduction to the paper

# Data Description

(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

Mr. Mason,

Thank you for contacting DABAS LLC; we feel confident we can answer your questions about the Ames, Iowa housing market. We were able to track down the most recent 2,930 home sales in Ames as well as 79 characteristics of each house. Since you are only interested in three specific neighborhoods, North Ames (NA), Eastwood (EW) and Brookside (BS), and how square footage (SF) relates to their sale price (SP, in dollars), we will restrict our model to only focus on those variables.

By examining the scatterplot (see Appendix Plot 1A), we concluded we would need to take the log of the sales price and square footage to create a linear relationship. Also, the slope of the regression lines of different neighborhoods would be different, so we included interaction terms. This means our model would look like this:

ln(SP) = B0 + B1 \* ln(SF) + B2 \* BS + B3 \* EW + B4 \* NA + B5 \* ln(SF) \* BS + B6 \* ln(SF) \* EW + B7 \* ln(SF) \* NA

To test our model, we used a null hypothesis that all Bi are zero, or that there is no correlation between sales price and any of the other variables. And our alternative hypothesis is that at least one Bi is not zero, or that there is a correlation between at least one of the variables and sales price. We received a p-value well below 0.05 (it actually was below 0.001, see Appendix Table 1B column 6), so we rejected the null hypothesis and concluded there was a relationship between the sales price and the other variables.

We then needed to test whether each individual variable was needed in our model. For each variable, our null hypothesis was that given the other variables contribution to the model, that particular variable did not help our model (or that the coefficient in front of that variable was 0). And our alternative hypothesis was that the variable contributed to the model (or that the coefficient in front of the variable was not 0). For each variable aside from BS, including the interaction terms, we received p-value below 0.05 (see Appendix Table 1C column 6), which means for each variable we rejected the null hypothesis and concluded that they were all beneficial to our model. Since the interaction term was significant that included BS, we kept BS in the model. The full regression equation is:

**ln(SP) = 5.16 + 0.95 ln(SF) + 1.04 BS + 2.40 EW + 3.33 NA – 0.17 BS ln(SF) - 0.36 EW ln(SF) - 0.48 NA ln(SF)**

Note: A zero for all categorical variables (BS, EW, NA) would represent the houses outside those three neighborhoods, but still in Ames. We decided to use them in an effort to reduce the overall variance to our model.

This model has an r-squared of 0.5649 and an adjusted r-squared of 0.5368. The r-squared means 56.49% of the variation in sales price, can be explained the variation in neighborhood and square footage. As always, these coefficients are just estimates. For instance, we are 95% confident the true coefficient for ln(Square Footage) is in the interval (0.914,0.984) (see Appendix 1C, row 3 columns 7 and 8. The other confidence intervals for the rest of the coefficients can be found in the same table.).

For us to use this model, we had to assume independence even though that likely isn't completely true. Also, this model gave us relatively normal residuals (slightly curved and skewed left, but with 2000+ observations, the Central Limit Theorem should kick in), but it did show some outliers and leverage points (see Appendix Plot 1D). We concluded the outliers were all accurate measurements so we had to include them. The model created after removing the leverage point gave very similar results to our model, so we decided to keep our model and include the leverage points **(**all confidence intervals for the coefficients contain the coefficients from our full model. See Appendix Table 1E).

We can simplify this model a little by separating it by neighborhood. This leaves us with three different equations you can look at depending on which neighborhood you are working with.

**{ ln(SP) | Brookside } = 6.20 + 0.78 ln(SF)**

**{ ln(SP) | Edwards } = 7.56 + 0.59 ln(SF)**

**{ ln(SP) | North Ames } = 8.49 + 0.47 ln(SF)**

In general, when you double the size of a house in Brookside, the sale price increases by a factor of 2.16. When you double the size of a house in Edwards, the sale price increases by a factor of 1.79. And when you double the size of a house in North Ames, the sale price increases by a factor of 1.56.

Note: The model created by removing the other neighborhoods produced the same regression equations for the individual neighborhoods (See Appendix Table 1F and Reference 1G).

Due to the nature of this study (it was an observational study by nature), we cannot infer causation and say the square footage or the neighborhood causes a change in the sales price of the house; there are obviously many confounding variables that play a part. But, we can say there is a correlation between the sales price and the square footage and neighborhood.

# Analysis Question 2

**Restatement of Problem**

**Model Selection**

Type of Selection

Stepwise

Forward

Backward

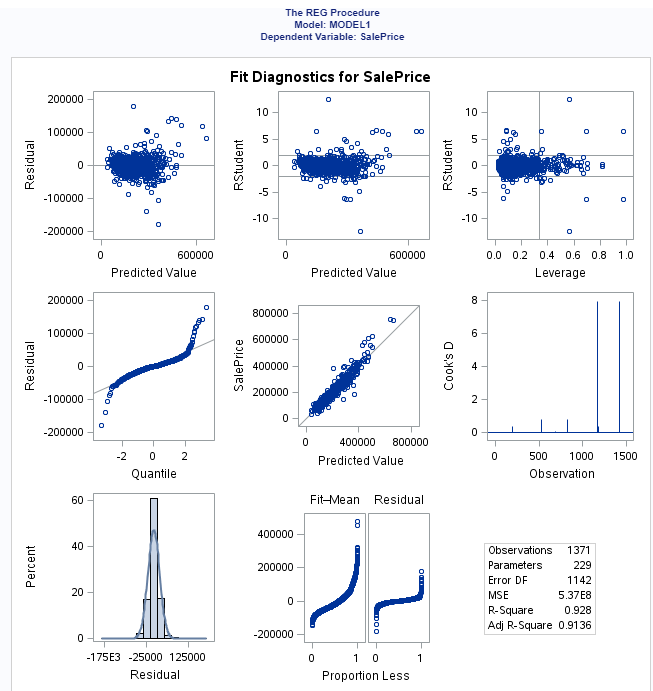
CUSTOM

**Overall Data Observations:**

* Number of Observations Read: 1460
* Number of Observations with Missing Values: 339
* Number of Observations Used: 1121

**Checking Assumptions:**

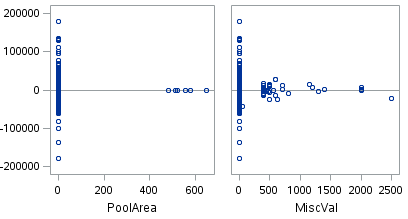
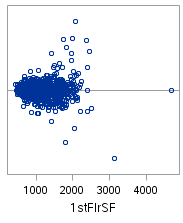
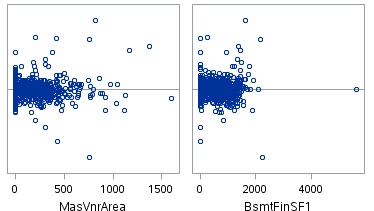
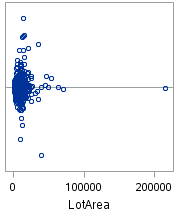
First we will look at the distribution of our response variable, the Sale Price using all of our identified continuous and indicator variables through proc reg. Here are the plots for SalePrice:

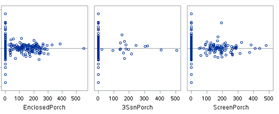
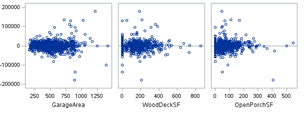


Here we see a few worrying things. The residual plot seems clustered tightly together with a few outliers.

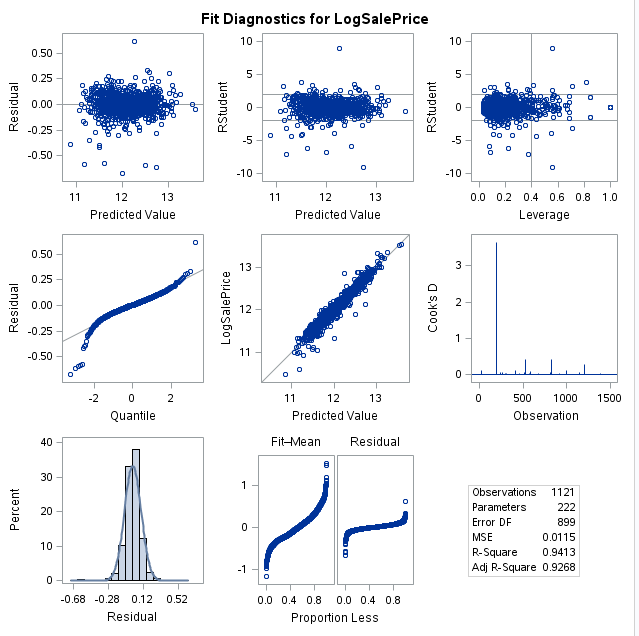
1. **Normality**: Judging from the qq plot, there seems to be a curvature at the tail ends of the data, which violates normality.
2. **Constant variance:** Judging from the scatter plots we do not see a large concern with variance. There are many outliers that are concerning, but with so many observations that itself wouldn’t be out of the question… still, given the violation in normality we should try a transform and hope this gets better.
3. **Linear trend:** From the predicted value line and an adjusted R-Square of .91 our linear trend looks good here, so we will compare this to our transformed data.
4. We will assume all observations are **independent**.

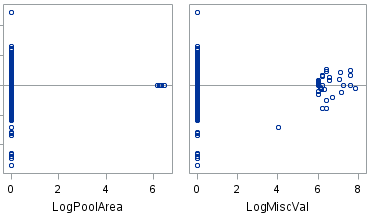
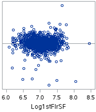
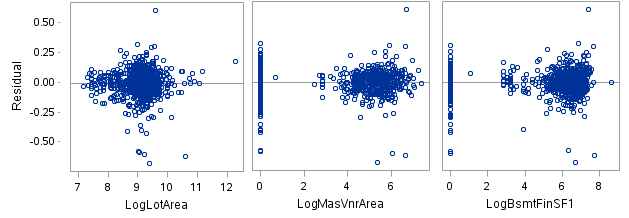
Viewing the scatterplots for the individual variables we see many variables with uneven distributions and distant outliers. These data points might also benefit from a log transformation:

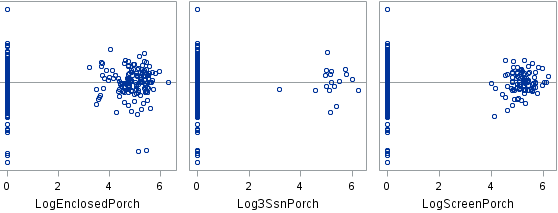
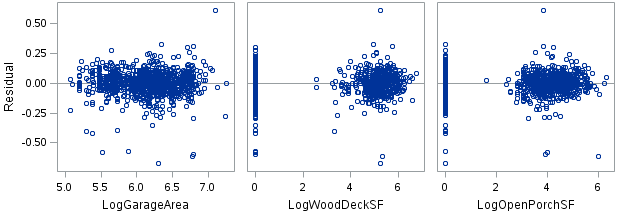




After taking the logs we see a few improvements in the model:







After the transformations we see many improvements in the shape of our data. In general the data has become less affected by outliers, and closer to a random cloud.

1. **Normality**: Judging from the qq plot, the curvature at both ends is still present, but far less pronounced. The tails are mostly going to be formed by our large number of categorical variables that have a small number of discrete values. In the middle we see a straight line indicating normality, and the histogram shows a much more normal distribution.
2. **Constant variance:** Across the board the scatter plots are showing data that are more cloud-like and normal in their variance. We still have a large number of variables with values of zero which negatively affects our constant variance, but we do not see a way around this, we will accept it and proceed.
3. **Linear trend:** From the predicted value line we see a slight improvement here, which is also reflected by an increase in our adjusted R-Square to .93. Our linear trend looks even better than before the transformation.
4. We will assume all observations are **independent**.

**Comparing Competing Models**

Adj R2

Interval CVPress

Kaggle Score

**Conclusion**:

A short summary of the analysis.

# Appendix

Plot 1A:

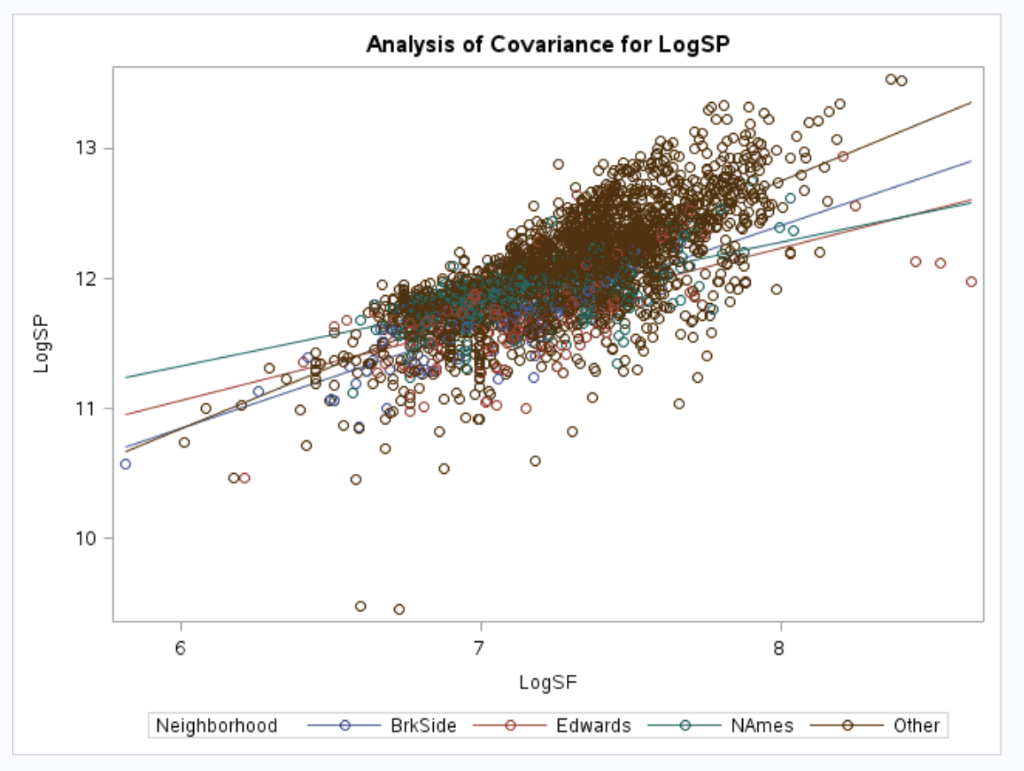


Table 1B:

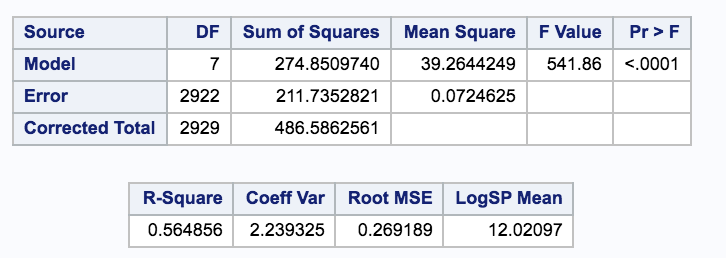
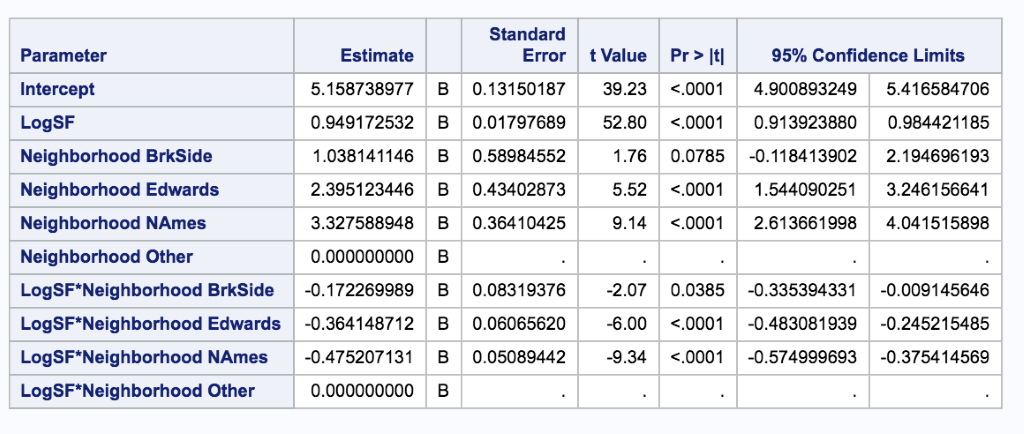


Table 1C:



Plot 1D:

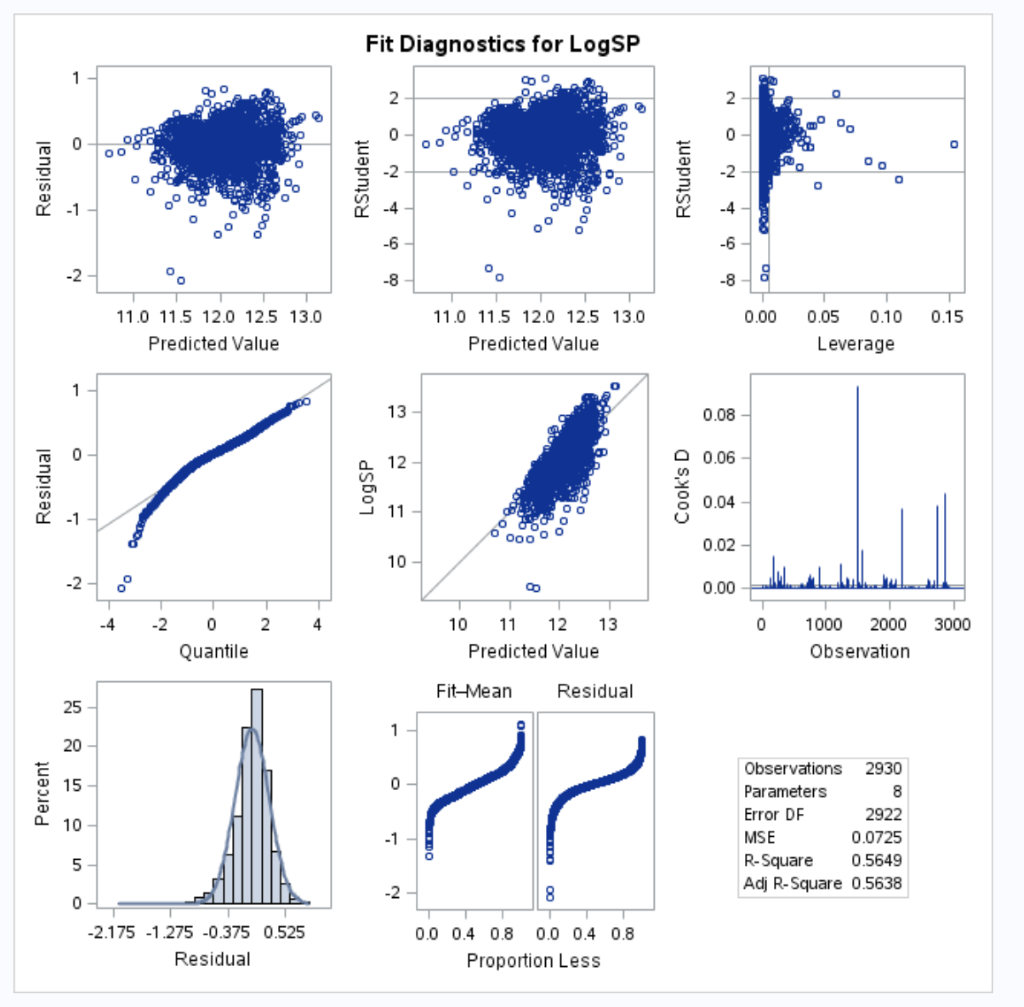


Table 1E:

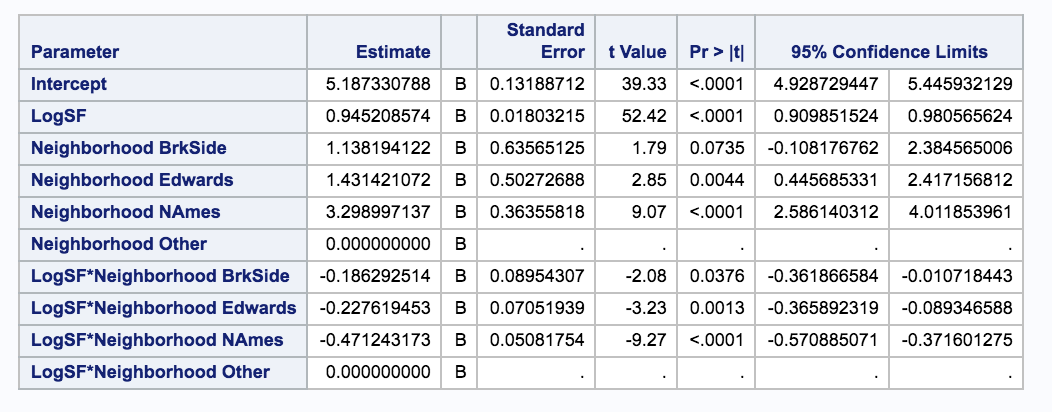
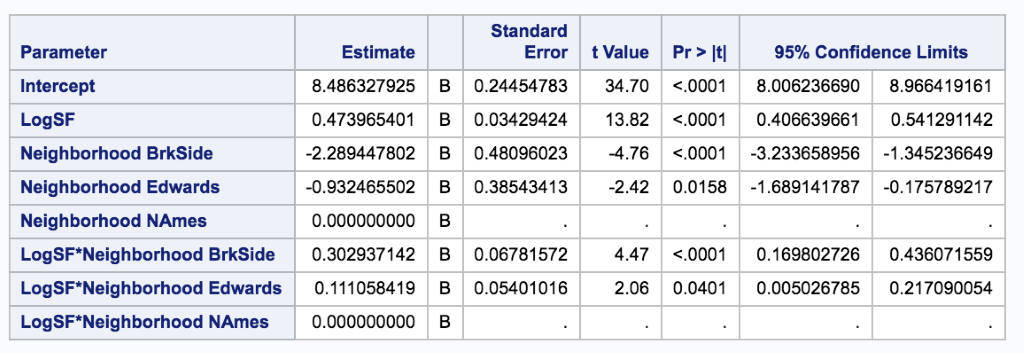


Table 1F:



Reference 1G:

Separated regression equations:

{ ln(SP) | Brookside } = 6.19 + 0.77 ln(SF)

{ ln(SP) | Edwards } = 7.55 + 0.58 ln(SF)

{ ln(SP) | North Ames } = 8.48 + 0.47 ln(SF)

**Question 1 Code:**

R code to turn all other neighborhoods into "other", removed unwanted variables to help with speed of SAS and created log variables:

ah <- read.csv("~/Documents/AmesHousing.csv")

q1 <- ah[,c("Neighborhood","Gr.Liv.Area","SalePrice")] ## removes unwanted variables

colnames(q1) <- c("Neighborhood","SquareFeet","SalePrice") ## Renamed Gr.Liv.Area to Sqaure Feet

q1$LogSF <- log( q1$SquareFeet )

q1$LogSP <- log( q1$SalePrice ) ## create log variables for square feet and sale price

q1$Neighborhood[ Q1$Neighborhood != "BrkSide" & Q1$Neighborhood != "Edwards" & q1$Neighborhood = "NAmes" ] <- "Other" ## turns all the other neighborhoods into "Other"

write.csv(q1,"q1all.csv")

R code to remove leverage points:

sorted <- q1[ order(q1$LogSF,decreasing=TRUE ), ]

row.names(sorted) <- 1:nrow(sorted) ## sorted and changed row names to easily find highest/lowest

sorted2 <- sorted[-c(1,2,3,4,5,2930),] ## 5 highest were above 4000 sq ft. Lowest was also a leverage pt

write.csv(sorted2,"nl.csv") ## that's an L, for "no leverage"

R code to remove all other neighborhoods:

q1some <- q1[ q1$Neighborhood=="NAmes" | q1$Neighborhood=="Edwards" | q1$Neighborhood == "BrkSide" , ]

write.csv(q1some,"q1some.csv")

All SAS code was the same, but with different DATANAME

proc glm data=DATANAME plots=all;

class Neighborhood;

model LogSP = LogSF | Neighborhood / solution clparm;

run;

Plot 1A, Table 1B, Table 1C, and plot for 1D were generated using "q1all". Table 1E used "nl". Table 1F used "q1some".