**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

**Restatement of Problem**

**Specify the Model**

**Checking Assumptions**

Residual Plots

Influential point analysis (Cook’s D and Leverage)

Make sure and address each assumption.

**Comparing Competing Models**

adj R2

Interval CVPress

**Parameter Interpretation**

Interpretation

Confidence Intervals

**Conclusion**

A short summary of the analysis.

# Analysis Question 2

**Restatement of Problem**

**Model Selection**

Type of Selection

Stepwise

Forward

Backward

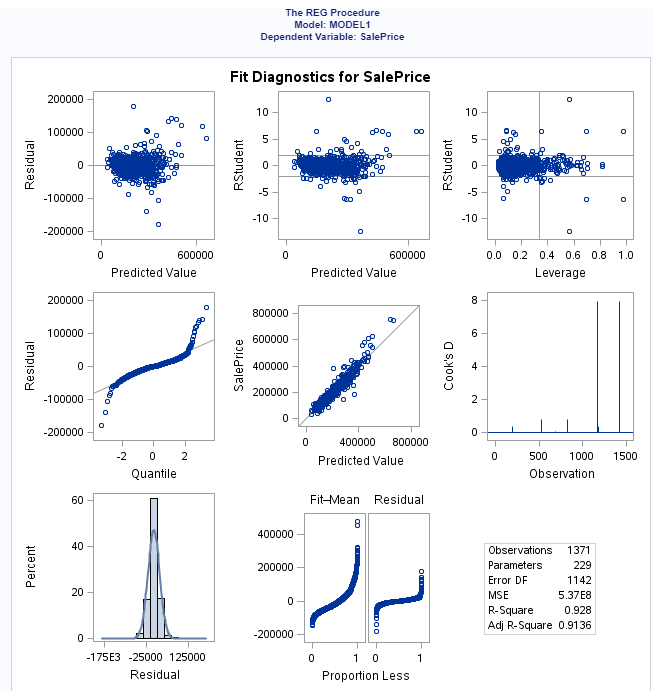
CUSTOM

**Overall Data Observations:**

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

**Checking Assumptions:**

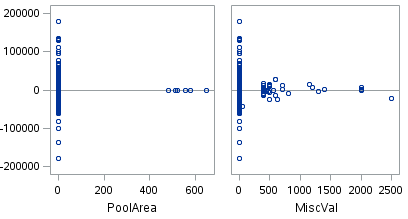
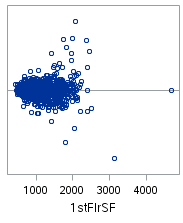
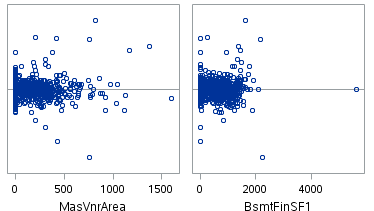
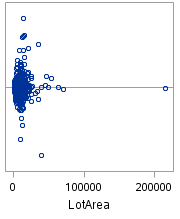
First we will look at the distribution of our response variable, the Sale Price using all of our identified continuous and indicator variables through proc reg. Here are the plots for SalePrice:

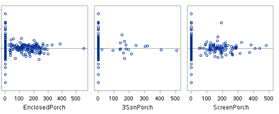
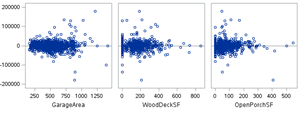


Here we see a few worrying things. The residual plot seems clustered tightly together with a few outliers.

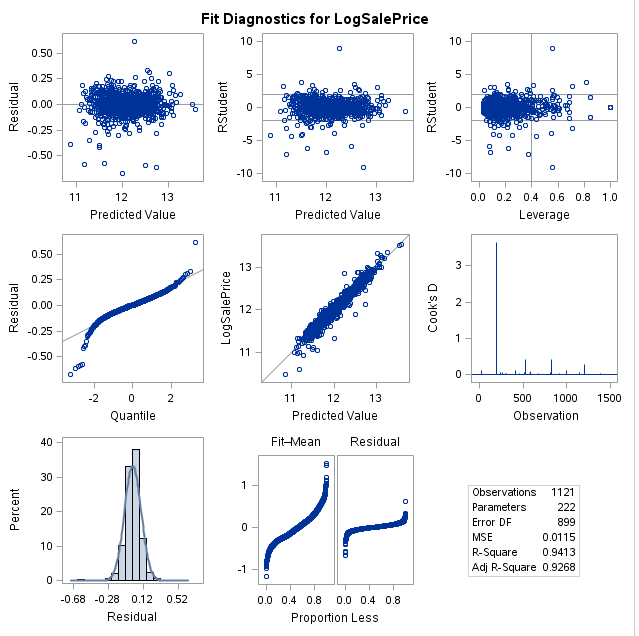
1. **Normality**: Judging from the qq plot, there seems to be a curvature at the tail ends of the data, which violates normality.
2. **Constant variance:** Judging from the scatter plots we do not see a large concern with variance. There are many outliers that are concerning, but with so many observations 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**.

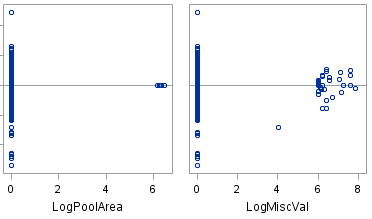
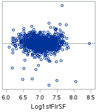
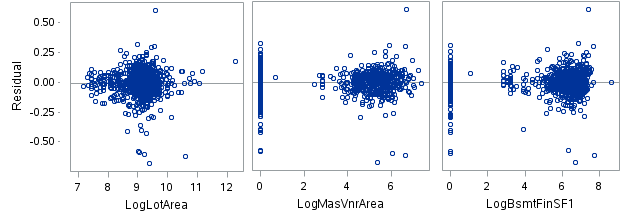
Viewing the scatterplots for the individual variables we see many variables with uneven distributions and distant outliers. These data points might also benefit from a log transformation:

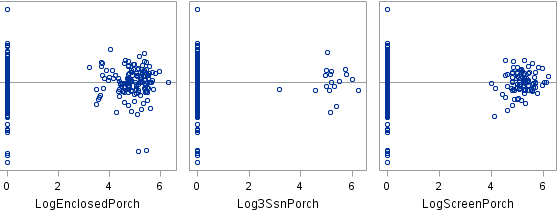
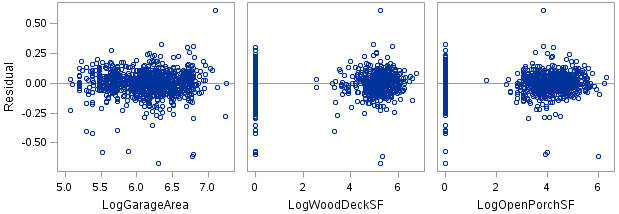




After taking the logs we see a few improvements in the model:







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

Well commented SAS Code for Analysis 1 and 2