**Introduction:**

This is the final project for the semester in MSDS 6371 Statistical Foundation for Data Science, and is a group project. It is based on the [Kaggle House Prices competition](https://www.kaggle.com/c/house-prices-advanced-regression-techniques). Fundamentally we are seeking to answer what will the sale price of a house be based on some combination of predictive attribute measures of it. The 2 specific questions below prescribe distinct approaches to answering this question.

**Data Description**

As mentioned above, the data comes from the [Kaggle House Prices competition](https://www.kaggle.com/c/house-prices-advanced-regression-techniques). The specific data sets we use for this project are the training data set (train.csv) to build the models, and test data set (test.csv) to cross-validate our models.

train.csv

This dataset is roughly 450k with 1460 observations of 81 variables. To read and understand more about this specific dataset please refer to the Kaggle website at <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>.

test.csv

This dataset, is roughly 441k with 1459 observation of 80 variables. This has one less observation: SalesPrices. This is due to the fact it is meant to be used for cross-validation, and does not require this variable because we are meant to predict that with our models.

**Analysis Question 1**

**The Problem**

Ames Century 21 real estate would like us to perform the following analysis of homes:

For the neighborhoods of: BrkSide, Edwards and NAmes, please find a predictive model for home SalePrice based on GrLivArea. Provide the model assumptions assessment with evidence, as well as data review and outlier analysis. Please provide estimate or estimates with confidence intervals, and a written conclusion of the relationship of GrLivArea and SalePrice within these neighborhoods.

**Primary Variables**

The following variables are fundamental to the answer for Question 1, do not contain missing values, and require no cleaning: \* SalePrice - the property’s sale price in dollars \* GrLivArea - Above grade (ground) living area square feet \* Neighborhood - Physical locations within Ames city limits

**Untransformed Data**

When looking at plots of the raw data, there is a clear linear relationship between GrLivArea and SalePrice. However, we can start to see there may be some issues with some of the points on Image 002, Image 003, and Image 004. These will be explored later.

Most of the parameter estimates look good. However the VIF for all estimates are high (See Image 005), the smallest being 22. This is another sign that outliers should be investigated.

When looking at the residual plots (Image 006) we can clearly see that Equal Variance is violated when looking athe the Residual and Studentized Residual. Also, the QQPlot and Histogram displays data that may not be normal.

We need to perform outlier analysis and peform transformations to address equal deviation and normality.

**Outlier Analysis**

**Transformed Data**

The following transformations were performed to attempt to address the issues found with the untransformed data: Normality and Equal Variance. Both Log-Linear and Linear-Log did not address these issues with the data. The final transformation Log-Log did.

Log-Log Transformation Analysis

Assumptions

Comparison of Competing Problems

Parameters

* Estimates
* Interpretation
* Confidence Intervals

**Analysis Question 1**

The Problem

Primary Variables

Forward Selection Model

Backward Selection Model

Stepwise Selection Model

Custom Model

|  |  |  |  |
| --- | --- | --- | --- |
| Predictive Model | Adjusted R^2 | CV Press | Kaggle Score |
| Forward |  |  |  |
| Backward |  |  |  |
| Stepwise |  |  |  |
| Custom |  |  |  |

**Conclusion:**

**Appendix**

Images