Build and Fit the Model

We will proceed with two competing models: our first model uses a log-log transformation of the dataset and includes the full scope. Our second model shares the log-log transformation, however, we will be narrowing our data’s scope to only include square footage where GrLivArea >= 1000 and GrLivArea <= 3250 and SalePrice is SalePrice >= 75000 and SalePrice <= 150000.

Concept:

logSalePrice = B0 + B1 \* logGrLiv + B2 \*Edwards + B3 \* NAmes + B4 \* (logGrLiv \* Neighborhood Edwards) + B5 \* (logGrLiv \* Neighborhood NAmes)

Fitted Model:

log\_SalePrice = 5.9129 + 0.8196 \* logGrLiv + 2.0935 \* Edwards + 2.5798 \* NAmes - 0.2999 \* (logGrLiv \* Edwards) - 0.3466 \* (logGrLiv \* NAmes)

Broken down per neighborhood:

BrkSide (reference):

Predicted SalePrice = 5.9129 + 0.8196 \* logGrLiv

Edwards:

Predicted SalePrice = 5.9129 + 0.8196 \* logGrLiv + 2.0935 \* Edwards - 0.2999 \* (logGrLiv \* Edwards)

Factored out

Predicted SalePrice = 7.0064 + 0.5197 \* logGrLiv

NAmes:

Predicted SalePrice = 5.9129 + 0.8196 \* logGrLiv + 2.5798 \* NAmes - 0.3466 \* (logGrLiv \* NAmes)

Factored out

Predicted SalePrice = 8.4927 + 0.473 \* logGrLiv

**Checking Assumptions**

Residual Plots

A blue dot on a white background

Description automatically generated

A graph showing a number of blue dots

Description automatically generated

Influential point analysis (Cook’s D and Leverage)

A blue dotted diagram with white text

Description automatically generated with medium confidenceA blue dotted diagram with white text

Description automatically generated with medium confidence

Normality: Judging from the scatter plot, Q-Q plot, and the histogram of residuals we see a normal distribution with some residual outliers; however, we will move forward with the assumption of normality.

Linear Trend: We see a positive linearity across our pairwise plots for each individual neighborhood.

Equal SD: There is some evidence of heteroscedasticity from our residual plots, where a few outliers are slightly out of range for comfort, however, we will proceed with caution and attempt competing models to confirm this assumption further.

Independence: We will assume the observations are independent.

A graph of a neighborhood

Description automatically generated with medium confidence

A graph with blue dots

Description automatically generated

A graph of a number of numbers

Description automatically generated with medium confidence

A graph and diagram of a graph

Description automatically generated

Pre-restricted Data:

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

Normality: For the restricted data, we see similar normality to our first model (unrestricted data). We will assume normality.

Linear Trend: We see a positive linearity across our pairwise plots for each individual neighborhood. The Edwards neighborhood shows an outlier; however, this is less influential than the original dataset.

Equal SD: There is some evidence of heteroscedasticity from our residual plots again, however, it appears to be less in strength (judging from our residual plots and DFBetas).

Independence: We will assume the observations are independent.

A graph with numbers and points

Description automatically generated with medium confidence A graph with numbers and dots

Description automatically generated with medium confidence

A graph and diagram of a graph

Description automatically generatedA graph with numbers and points

Description automatically generated with medium confidence

A comparison of blue dots

Description automatically generated with medium confidence

A graph showing a number of blue dots

Description automatically generated

Post-restricted Data:

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated A screenshot of a graph

Description automatically generated

A graph of a graph showing the difference between a number of individuals

Description automatically generated with medium confidence

A diagram of a plot

Description automatically generated with medium confidence

**Restricted Model:**

log\_SalePrice = 8.4195 + 0.4607 \* logGrLiv + 0.7059 \* Edwards + 2.2418 \* NAmes - 0.1051 \* (logGrLiv \* Edwards) - 0.3050 \* (logGrLiv \* NAmes)

Edwards:

Estimated SalePrice = 8.4195 + 0.4607 \* logGrLiv + 0.7059 - 0.1051 \* logGrLiv

NAmes:

Estimated SalePrice = 8.4195 + 0.4607 \* logGrLiv + 2.2418 - 0.3050 \* logGrLiv

BrkSide:

Estimated SalePrice = 8.4195 + 0.4607 \* logGrLiv

Comparing Competing Models

Adj R2

Non-restricted model:

Restricted Model:

Internal CV Press

Parameters

Estimates

Interpretation

Confidence Intervals

Conclusion

A short summary of the analysis.

R Shiny: Price v. Living Area Chart