Boston Housing Regression Analysis

Liping Gu, Zhezhi Hou, Ahmad Jarara

Binghamton University

December 5, 2016

Overview

1. High Level Overview
2. Model Instantiation
3. Model Verification

High Level Overview

- Variable Analysis
- Model Trimming
- Featured Variables
- Neighborhood Analysis
- Model Selection

Variable Analysis

- 1stFlrSF
- 2ndFlrSF
- WoodDeckSF
- OpenPorchSF
- YearBuilt
- BsmtSF
- Neighborhood
- Street

Model trimming

We eliminate the Street attribute from our model.

The aim of this analysis is not to overfit the formula to the little data we have, but instead have our model be the best at accurately predicting new elements.

Featured Variables

- X2ndFlrSF
- WoodDeckSF
- OpenPorchSF
- BasementSF

Model trimming

Special emphasis on neighborhood.

This removes degrees of freedom from the analysis and allows for more emphasis on other attributes.
However, this methodology is significantly flawed!

Model Selection

Very small sample size suggests a lax model.

Pitfalls of a complex analysis:

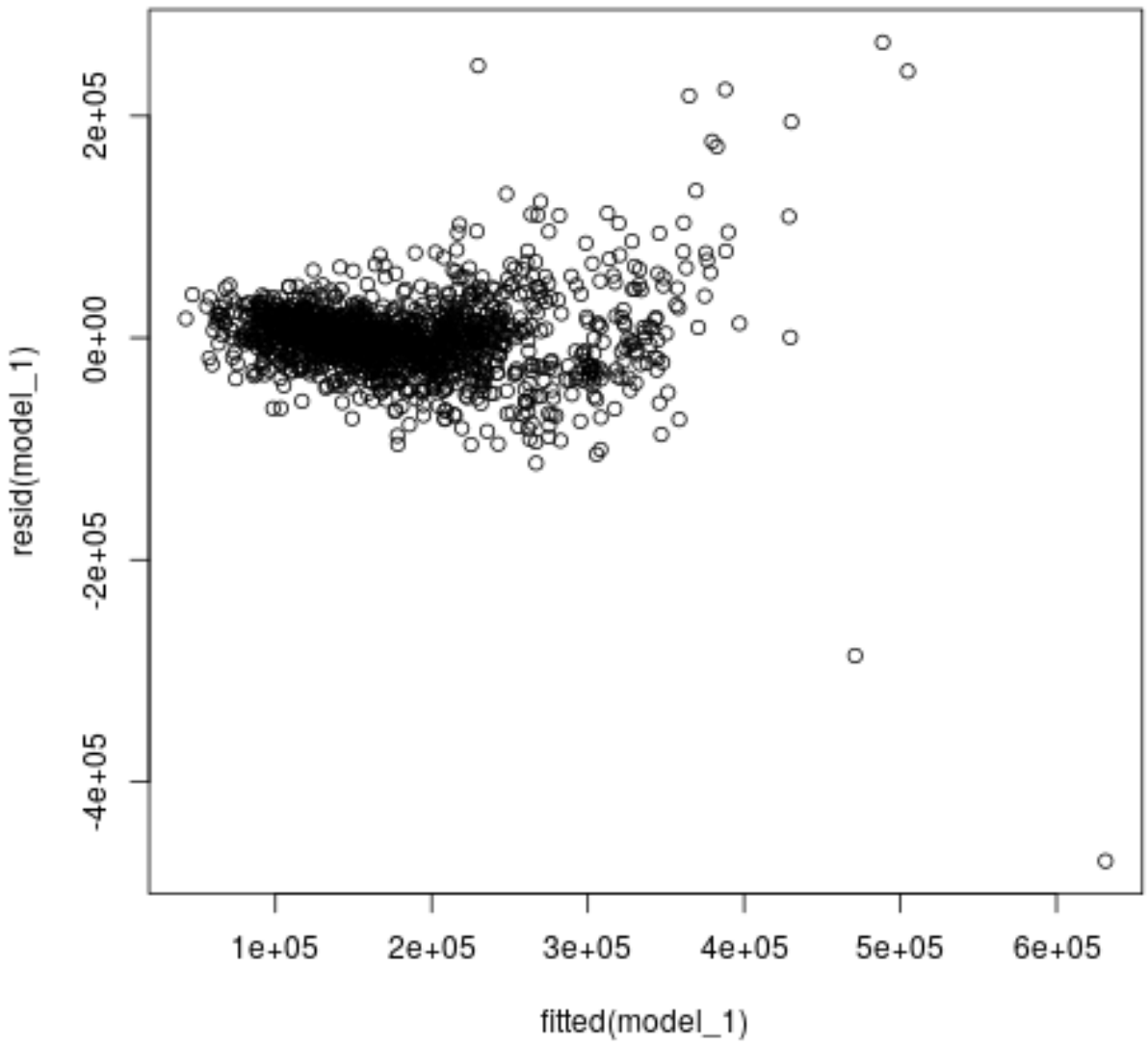
- Rigid model
- Simpson's Paradox

Then a simple linear regression is most likely the correct one.

Model Instantiation

- Initial Linear Regression Model
- Neighborhood Segregation
- Correlation Study
- Applying Transformations
- Intermediary Models and Stepwise Analysis
- Final Model

Initial Linear Regression Model

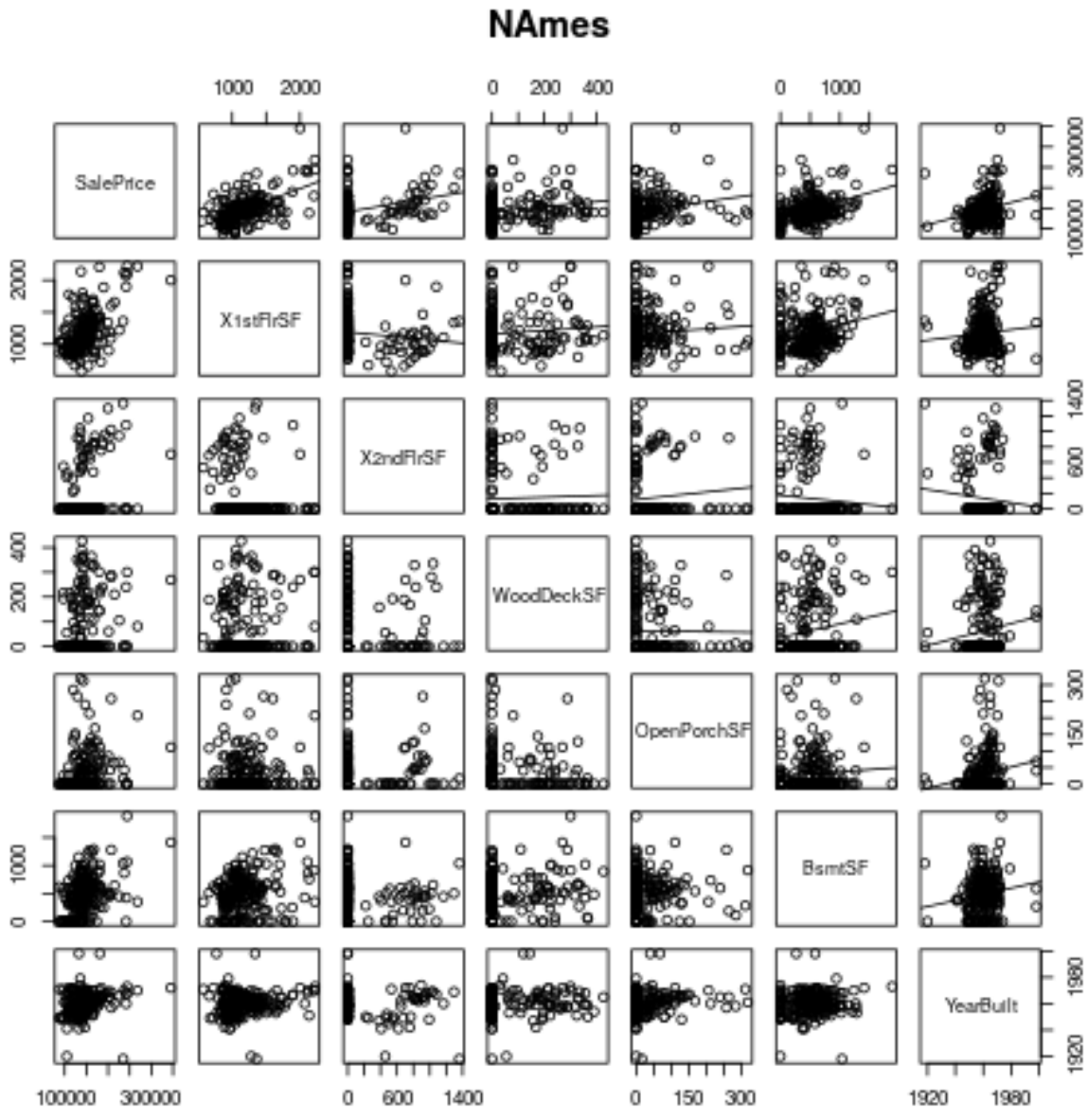


Neighborhood Segregation

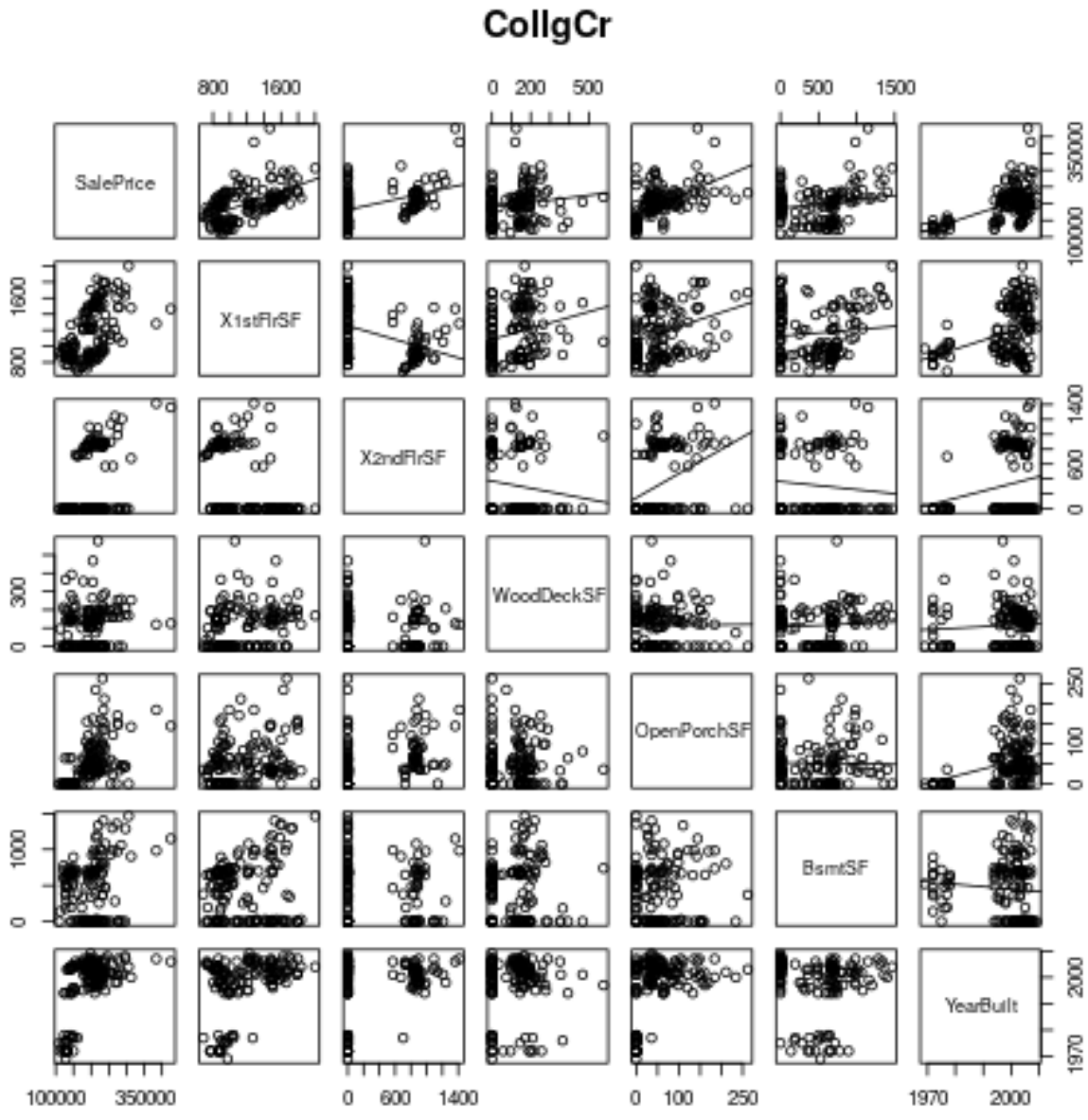
Neighborhood	Count	mean(SalePrice)
NAmes	225	145847.1
CollgCr	150	197965.8
OldTown	113	128225.3
NWAmes	73	189050.1
Timber	38	242247.4
IDOTRR	37	100123.8
ClearCr	28	212565.4
StoneBr	25	310499
SWISU	25	142591.4
Blmngtn	17	194870.9
MeadowV	17	98576.47
NPkVill	9	142694.4
Blueste	2	137500

Table: (Omission of other neighborhoods for brevity)

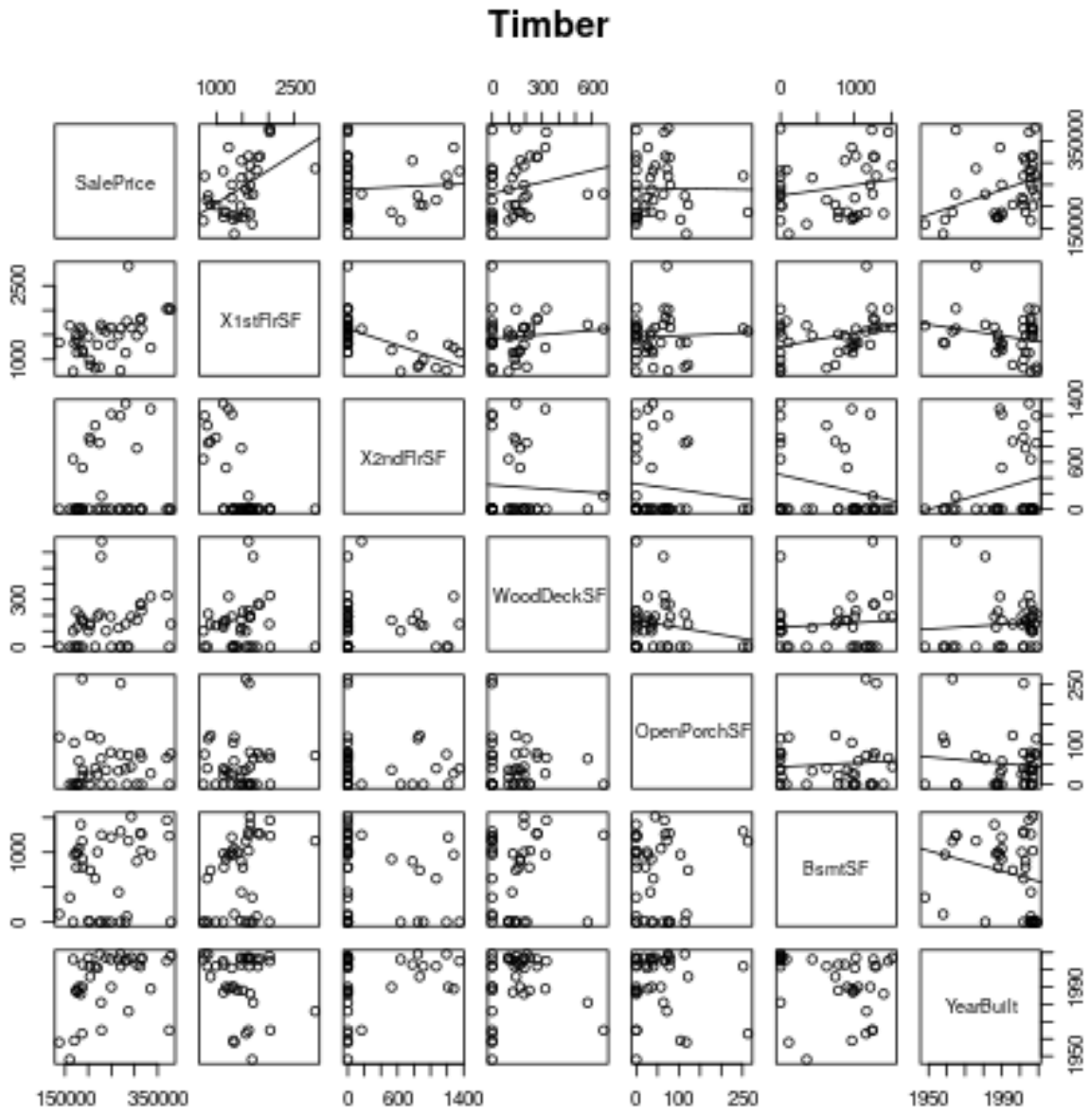
Neighborhood Segregation Pt. 2



Neighborhood Segregation Pt. 2

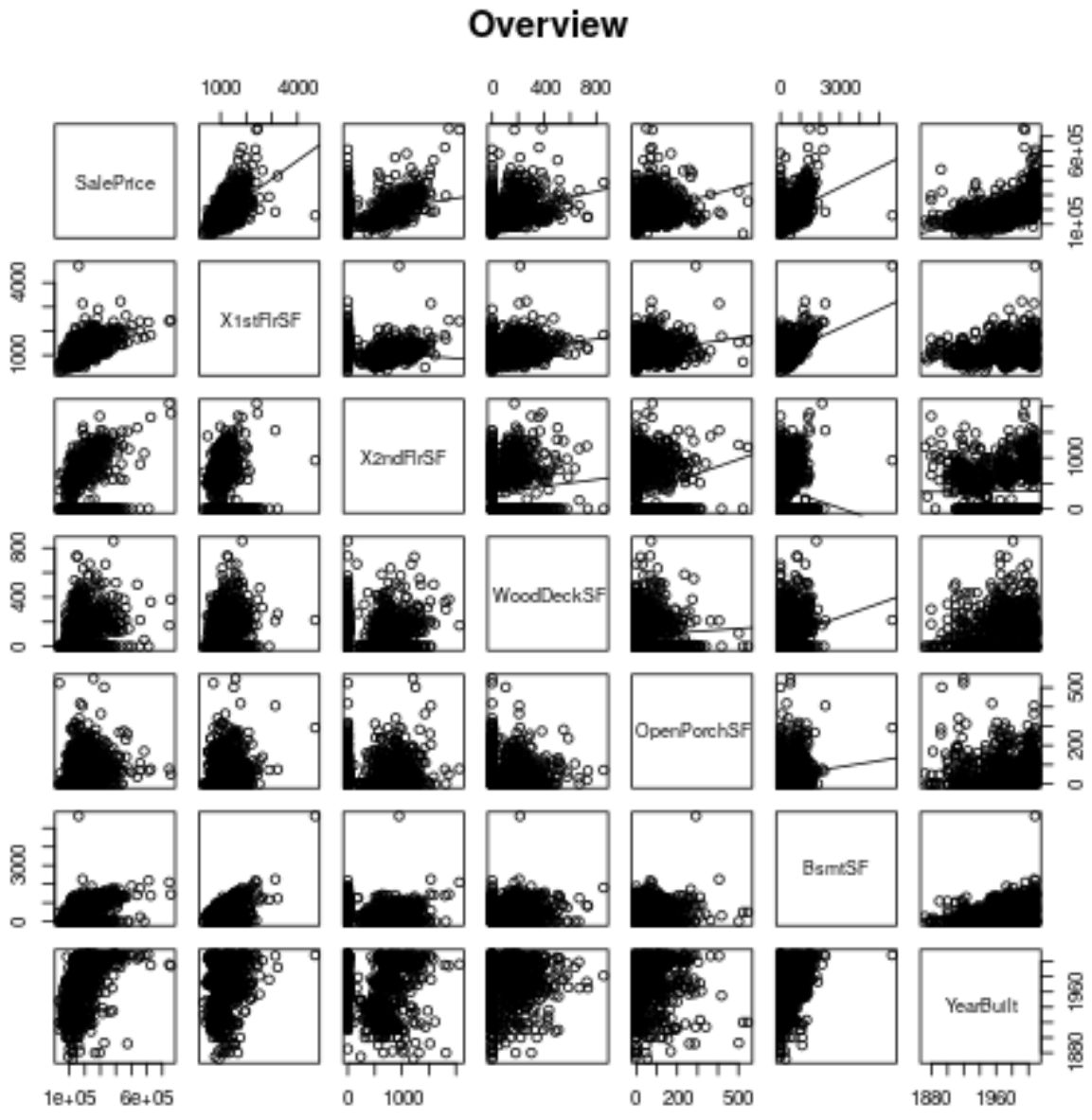


Neighborhood Segregation Pt. 2



Ignore all that craziness...

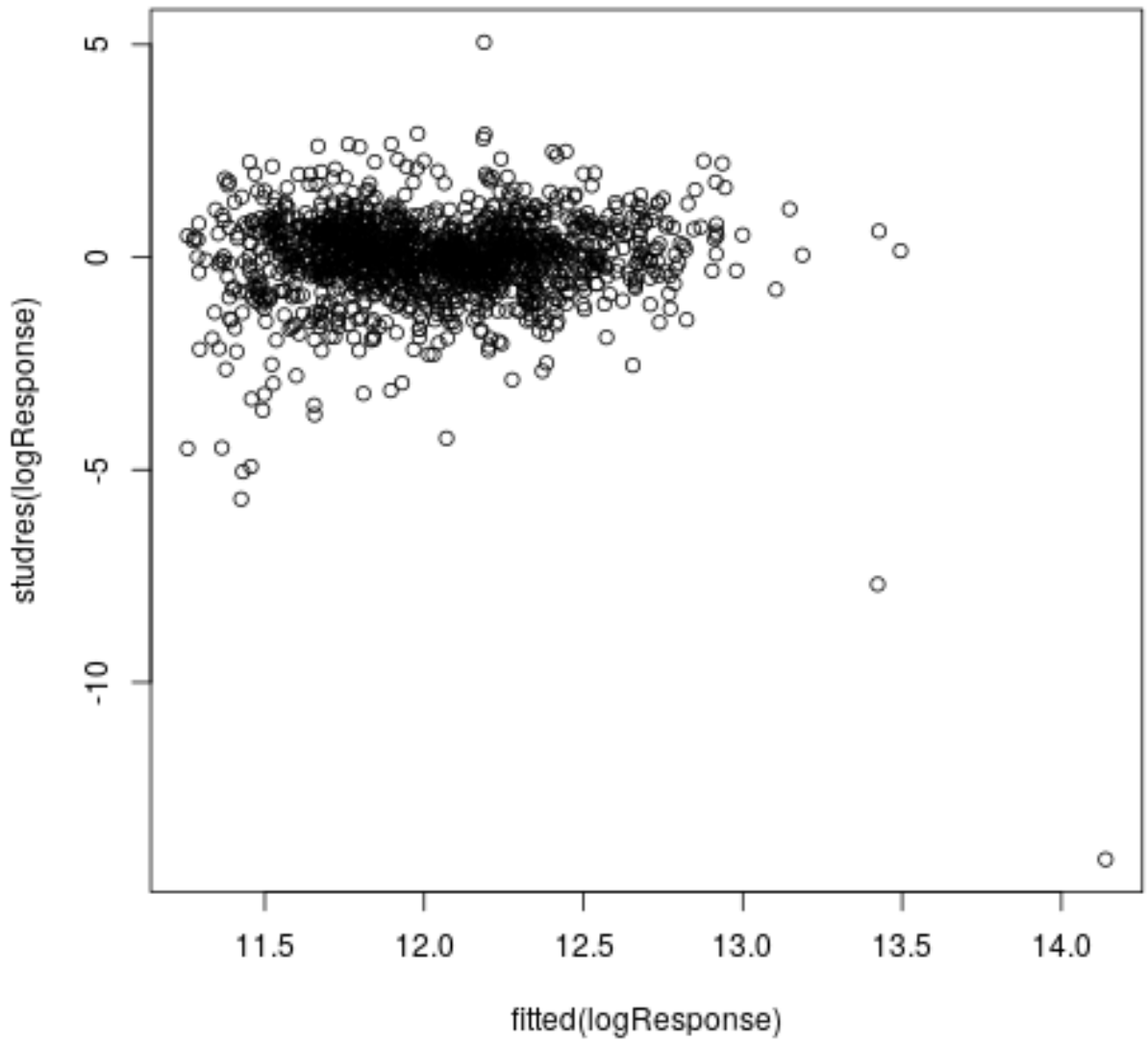
Correlation Study



Pesky Heteroscedasticity

- Removing neighborhoods stands to lose too much
- Large amount of variance means an inaccurate regression.
- Solution? Apply log transform!

Applying Transformations



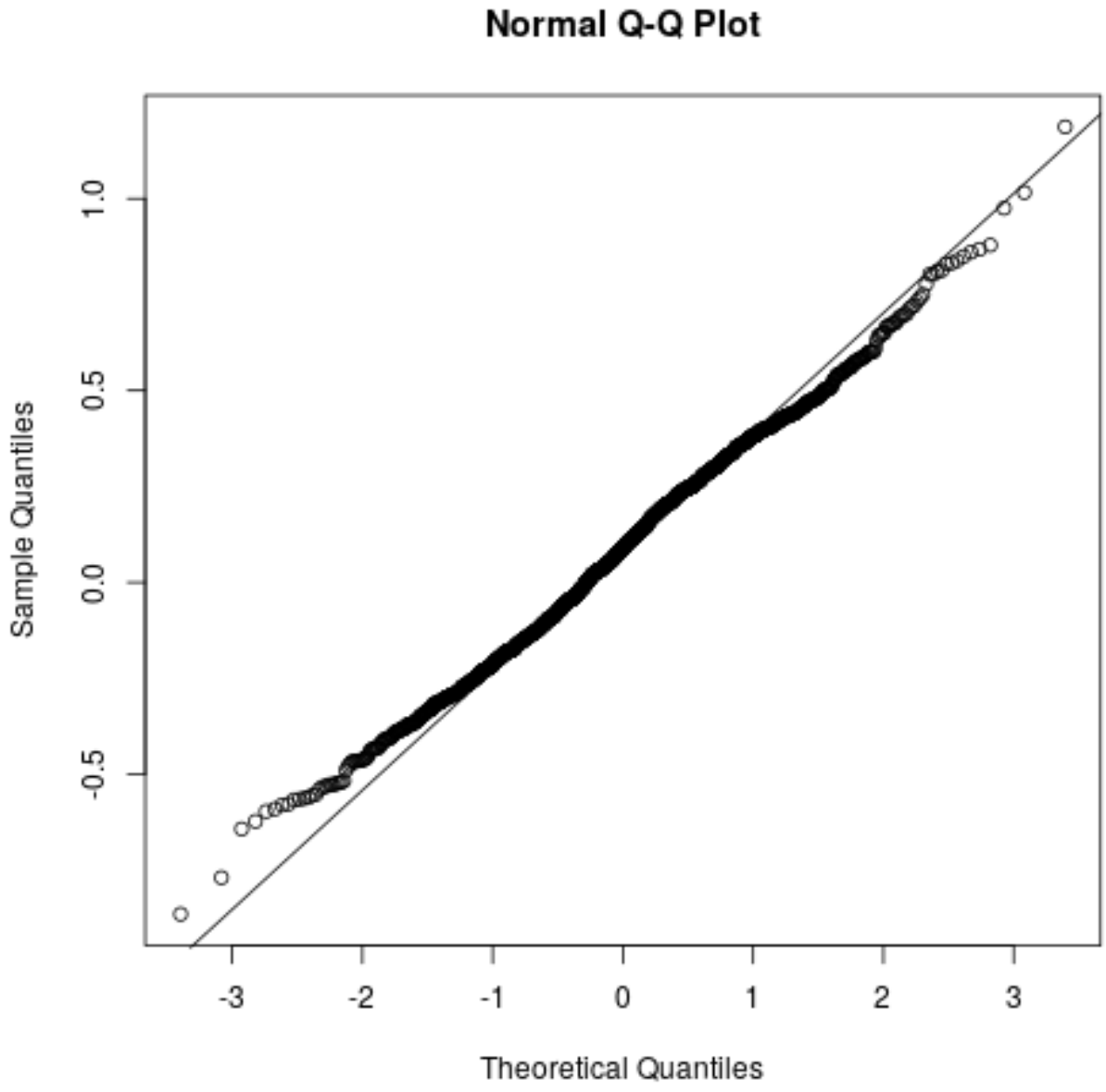
Applying Transformations Pt. 2

The residuals do not follow normal distribution.

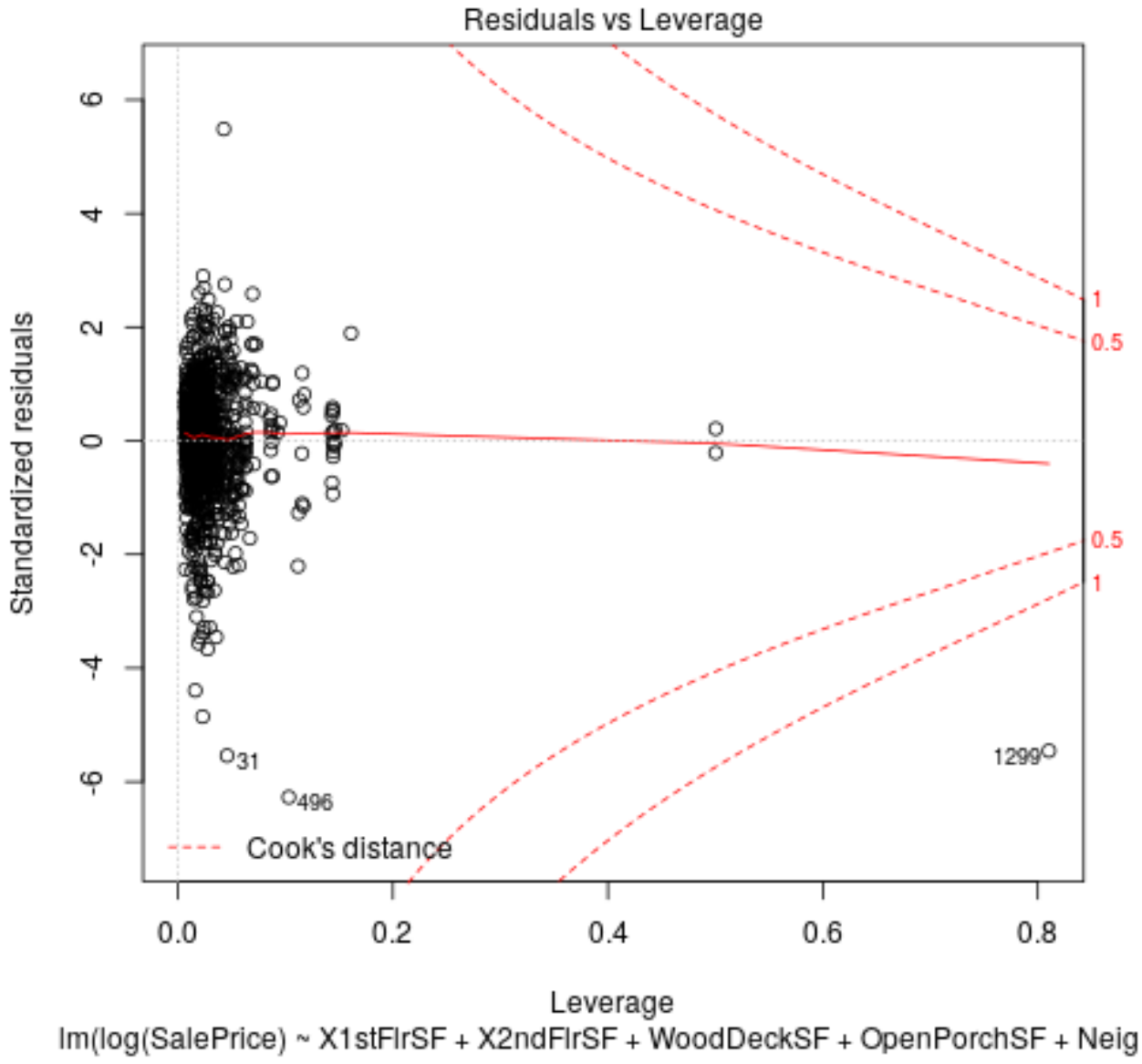
The solution to this is to use the generalized least squares estimator to establish correlation and coefficients for our model.

We do this by applying the weight of $1/\epsilon_i^2$.

Applying Transformations Pt. 3



Applying Transformations Pt. 4



Intermediary Models and Stepwise Analysis

Relevant variables include:

- X1stFlrSF
- X2ndFlrSF
- WoodDeckSF
- BsmtSF
- YearBuilt

We observed a strong quadratic trend in our correlation analysis, so we included X1stFlrSF^2 in our model.
Further, we used a stepAIC in both directions, and the only variable it came up with was an interaction variable between the first and second floors. We used that as well.

Final Model

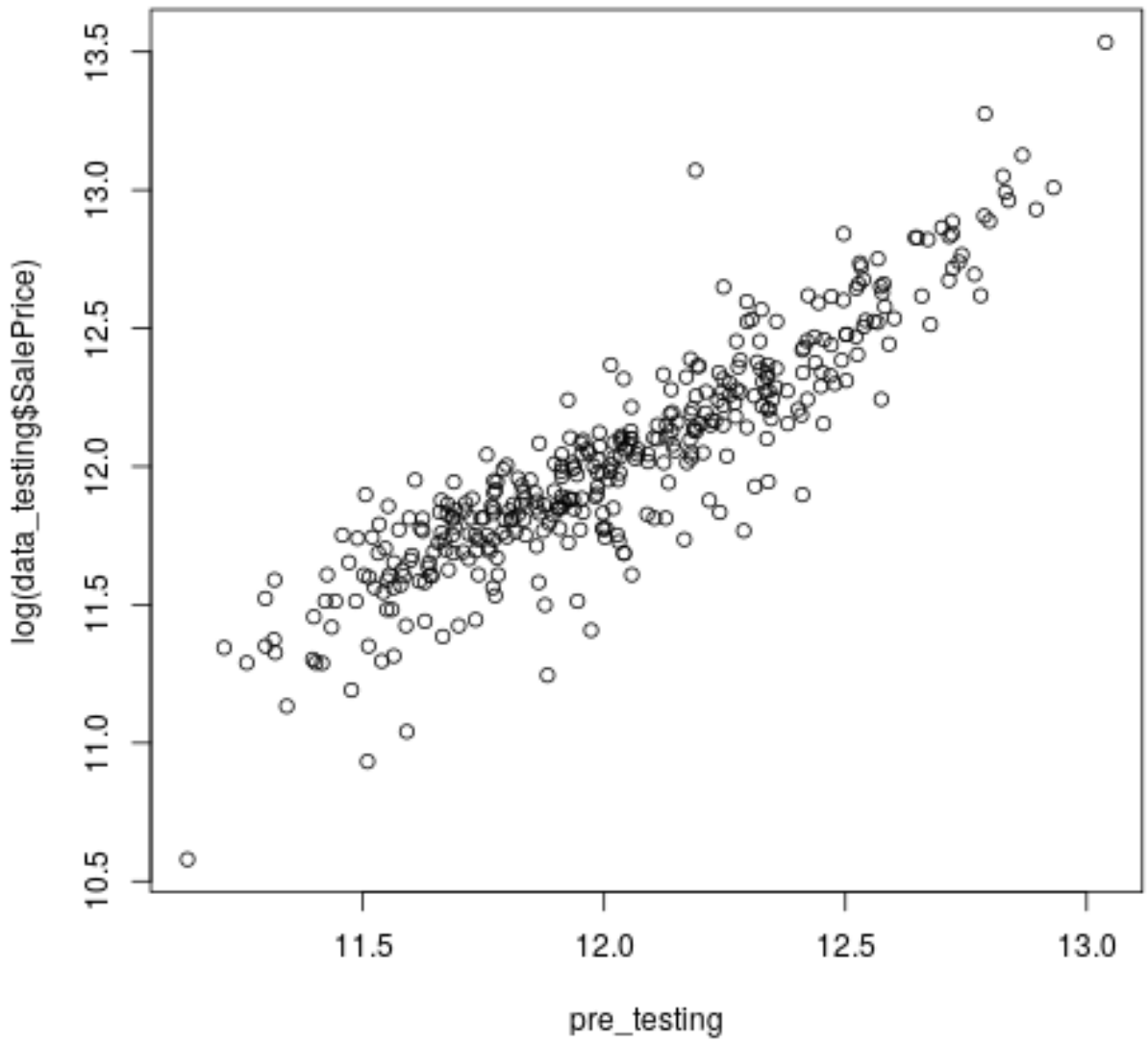
Attribute	**Coefficient**
X1stFlrSF	0.0008023042
X2ndFlrSF	0.0005162776
WoodDeckSF	0.000210046
OpenPorchSF	0.0003655512
BsmtSF	0.0001368659
YearBuilt	0.004020196
$X1stFlrSF^2$	-1.173194e-07
X1stFlrSF:X2ndFlrSF	-1.504619e-07

Table: Neighborhoods omitted for brevity

Model Verification

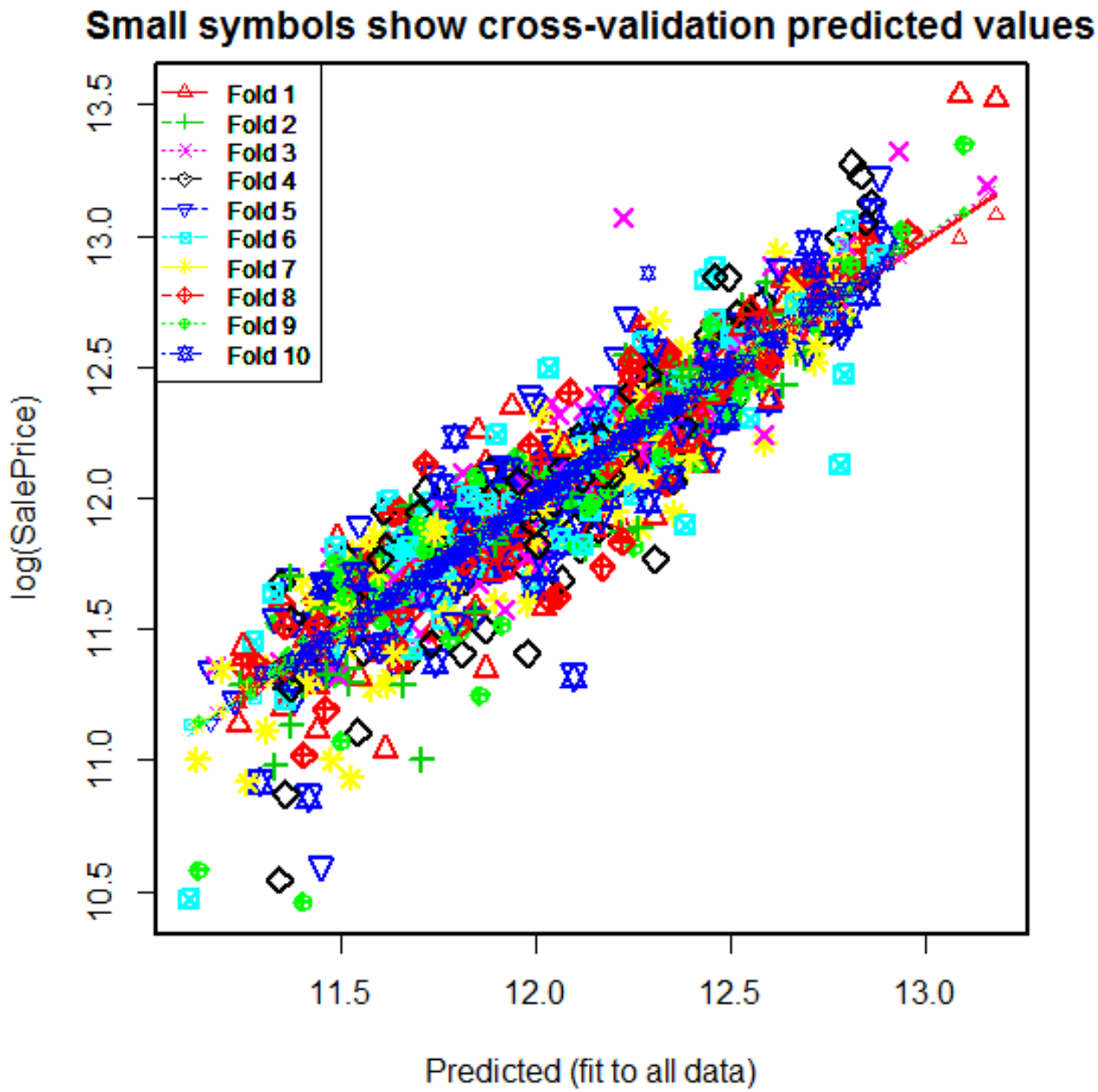
- Predictions
- Graphical Confirmation
- Cross Validation
- Additional Musings

Predictions



Drumroll...

Cross Validation

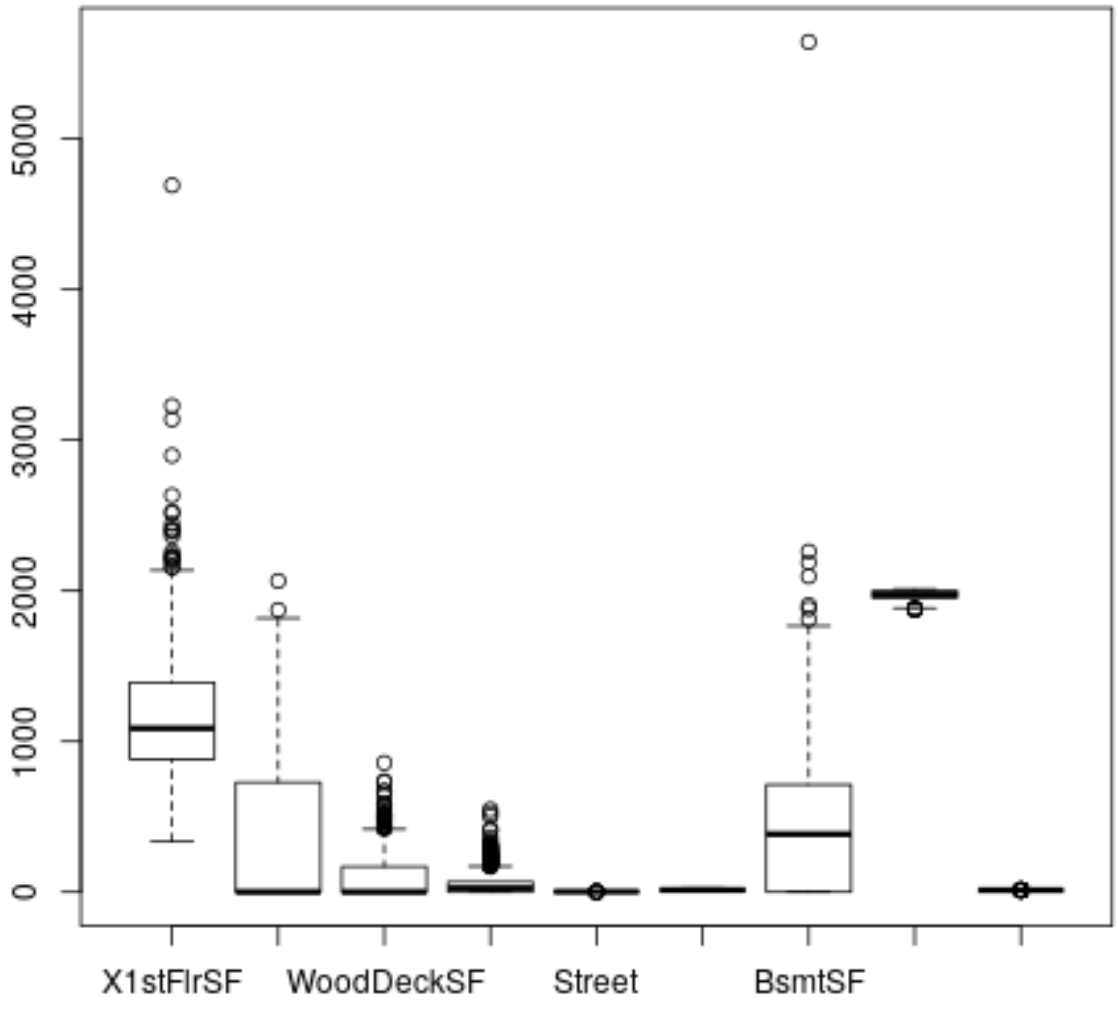


Additional Musings

Training data proportionally selected from neighborhood

- Is there a better model by selecting by neighborhood?
- No guarantees without guarantees

Outlier Treatment



Outlier Treatment Pt. 2

Parameter	**Old Coef**	**New Coef**	**Percentage**
X1stFlrSF	0.0008023042	0.000972	21.151055
X2ndFlrSF	0.0005162776	0.000545	5.5633636
WoodDeckSF	0.000210046	0.000216	2.8346172
OpenPorchSF	0.0003655512	0.000265	-27.506735
BsmtSF	0.0001368659	0.000138	0.82862130
YearBuilt	0.004020196	0.00361	-10.203383
$X1stFlrSF^2$	-1.173194e-07	-1.74e-07	48.313067
X1stFlrSF:X2ndFlrSF	-1.504619e-07	-1.58e-07	5.0099726

Thank you!