Using Linear Regression to predict Selling Price

The Problem

A homeowner with multiple properties is looking to sale some houses in Ames, IA and they want to know what prices to expect. They have enlisted me to to predict what prices their properties will sale for.

The Data

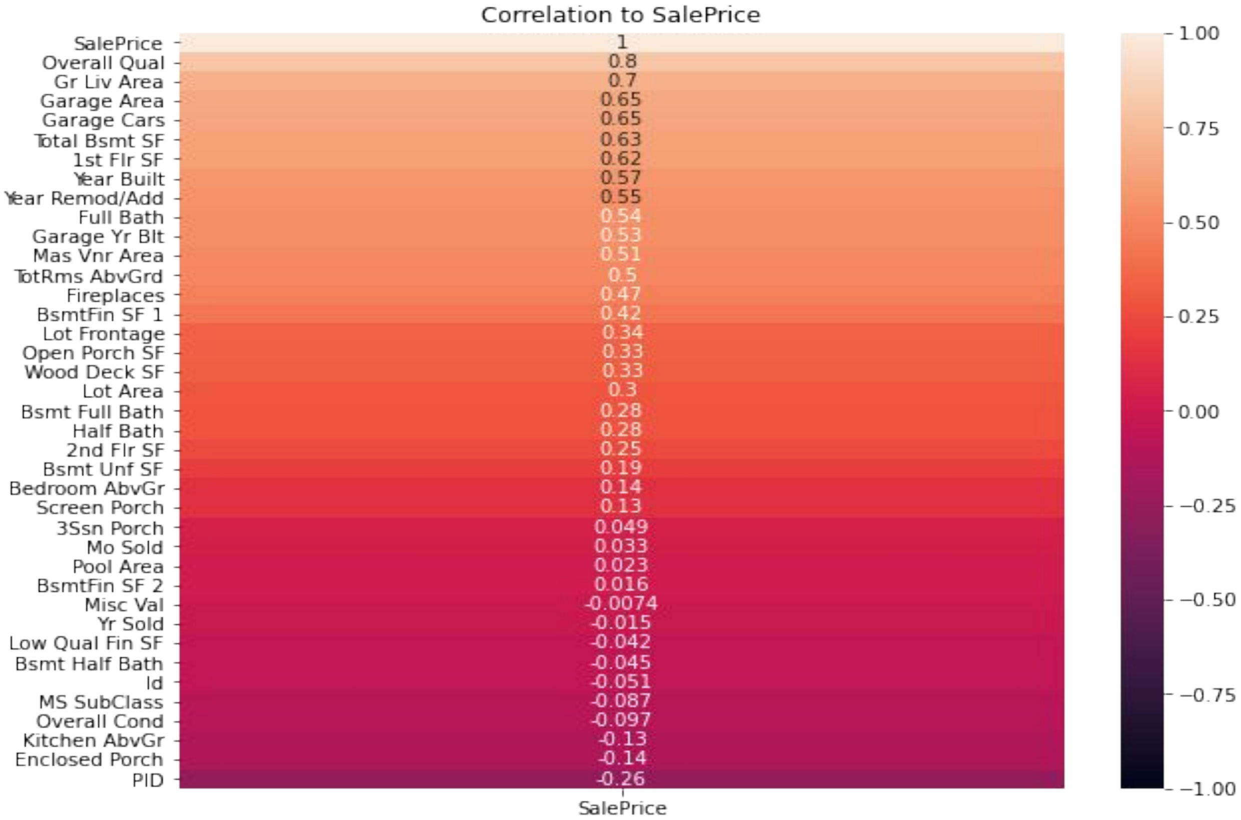
Data from the Ames Housing dataset was analyzed

- 2051 rows of data
- 81 columns of features

Catagories with missing values were dropped and final dataset for analysis consisted of:

- 2051 rows
- 55 columns of features

Selecting the features



SalesPrice is our feature of interest so we checked how all the others correlated to it

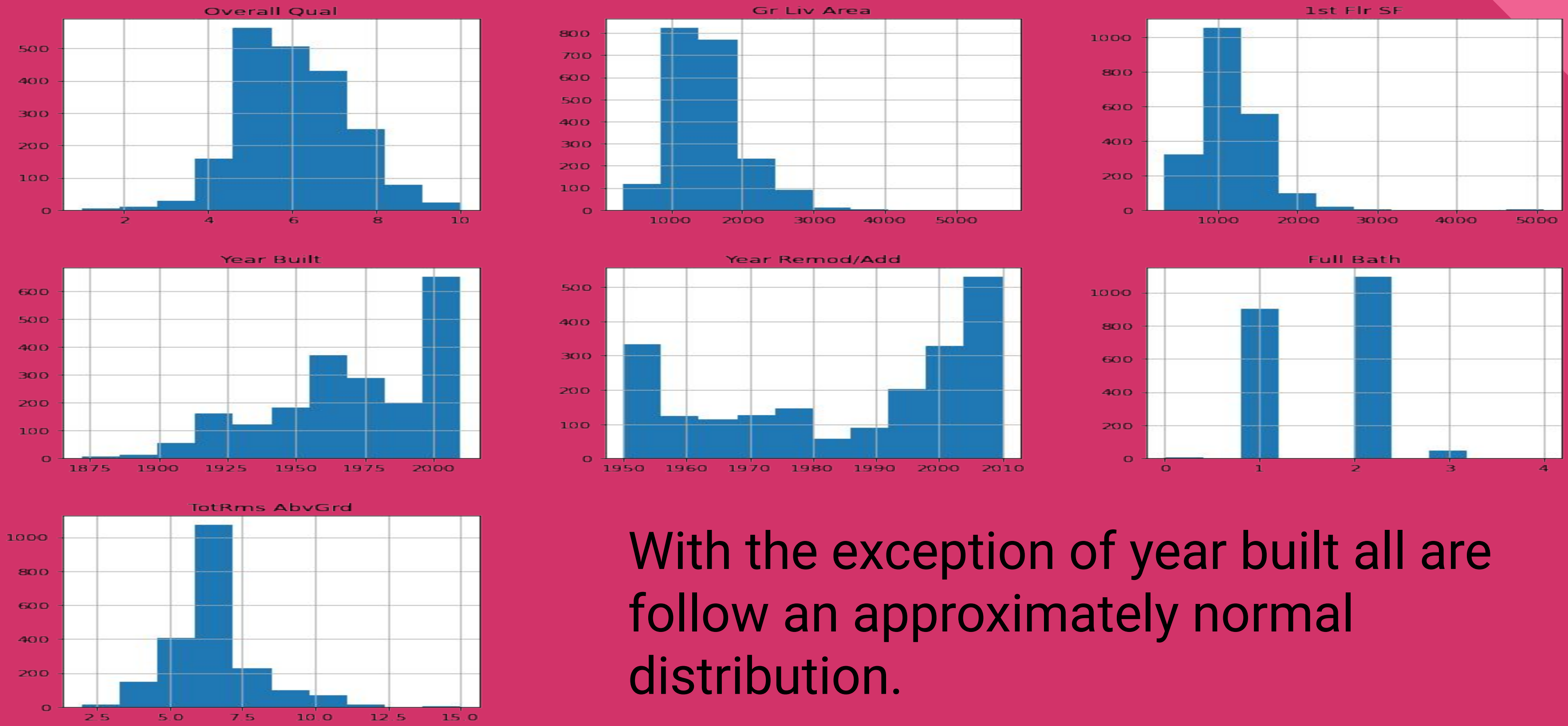
The Final Selection

The top 7 features most correlated to SalePrice were selected for the model

- Overall Qual
- Gr Living Area
- 1st Floor SF
- Year Built

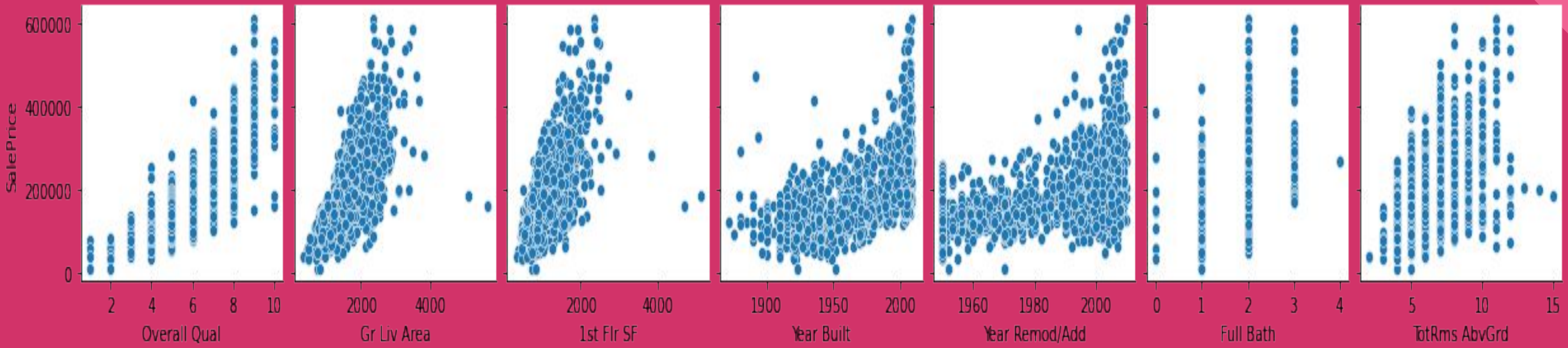
- Year Remod/Add
- Full Bath
- Total Rms Above Ground

Examining the Features



With the exception of year built all are follow an approximately normal distribution.

Linear Relationship to SalePrice

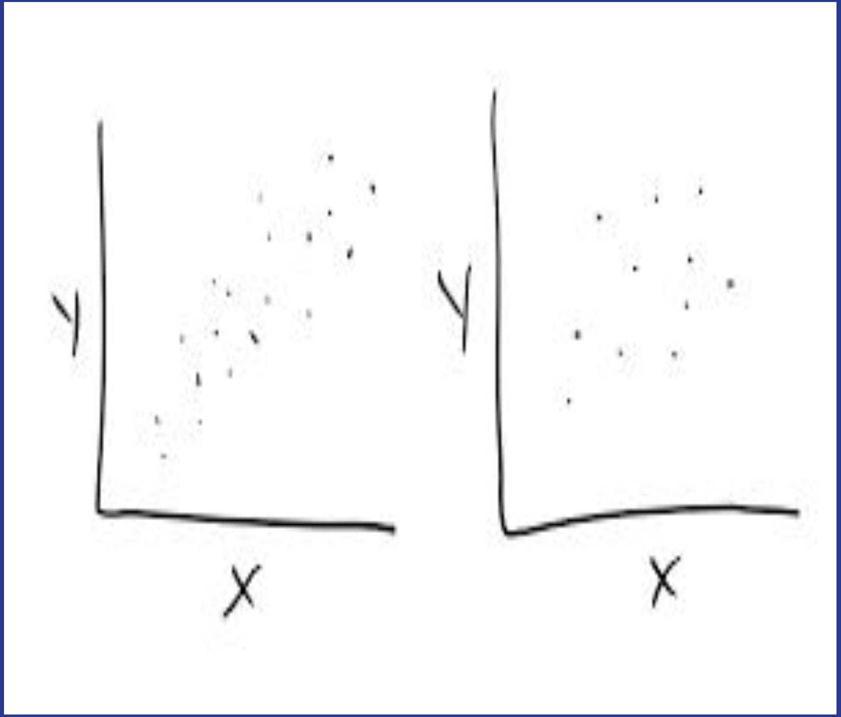


DONT FORGET TRAIN TEST SPLIT!

Before analysis and modeling, 70% our data set was split into a training set for modeling and 30% of the data was put into a testing set to test each model

Fitting our First Model and Initial Analysis

- The features were SCALED and fit to
  - Linear Regression
- Ridge Regularization Reg and
- Lasso Regularization Reg



Model 1 Results:

All three methods of fit produced the same r^2 scores for both sets of data

Training RSS = 0.759

Testing RSS = 0.819

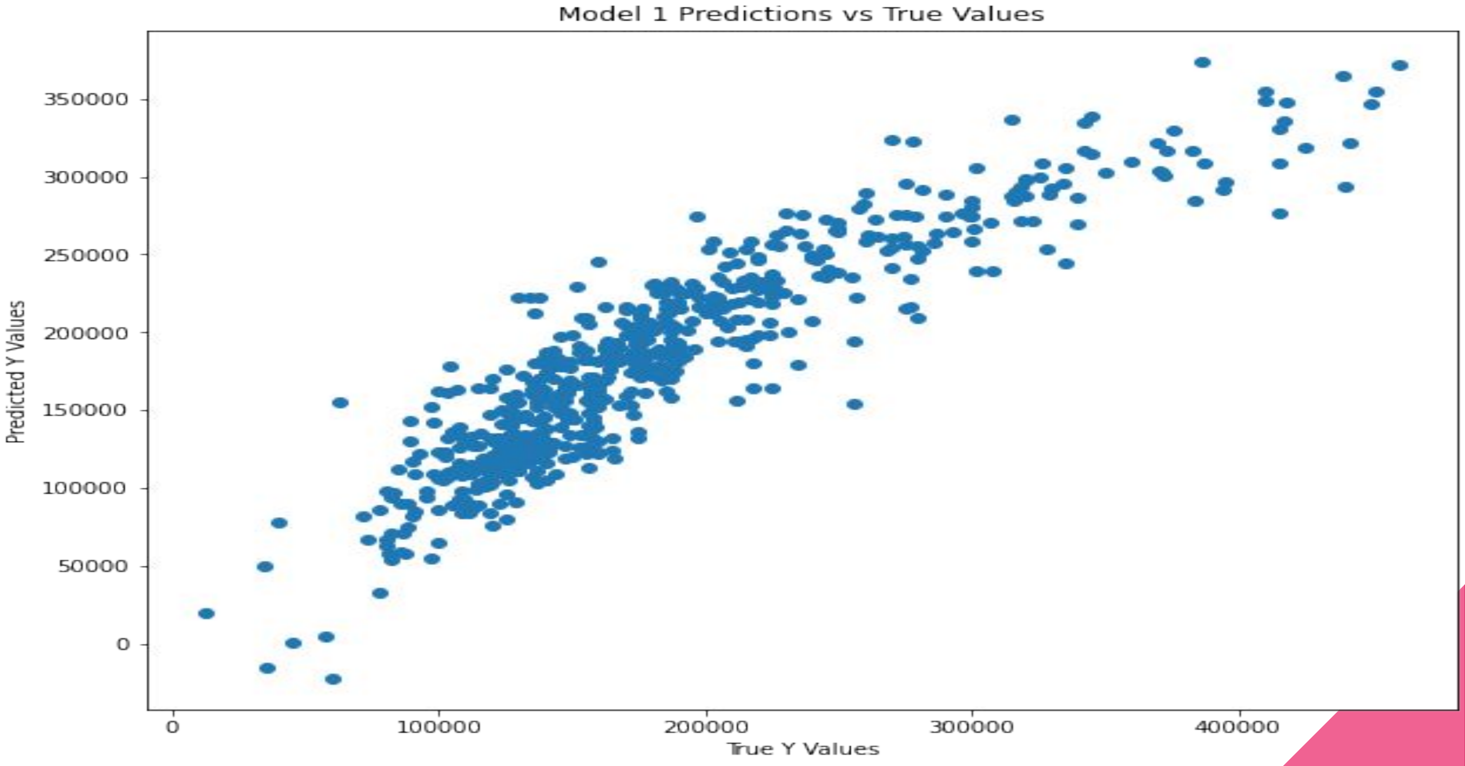
Training MSE = 154,1060,664.08

Testing MSE = 1,078,225,424.47

The model is Underfit and not a very strong fit for the data

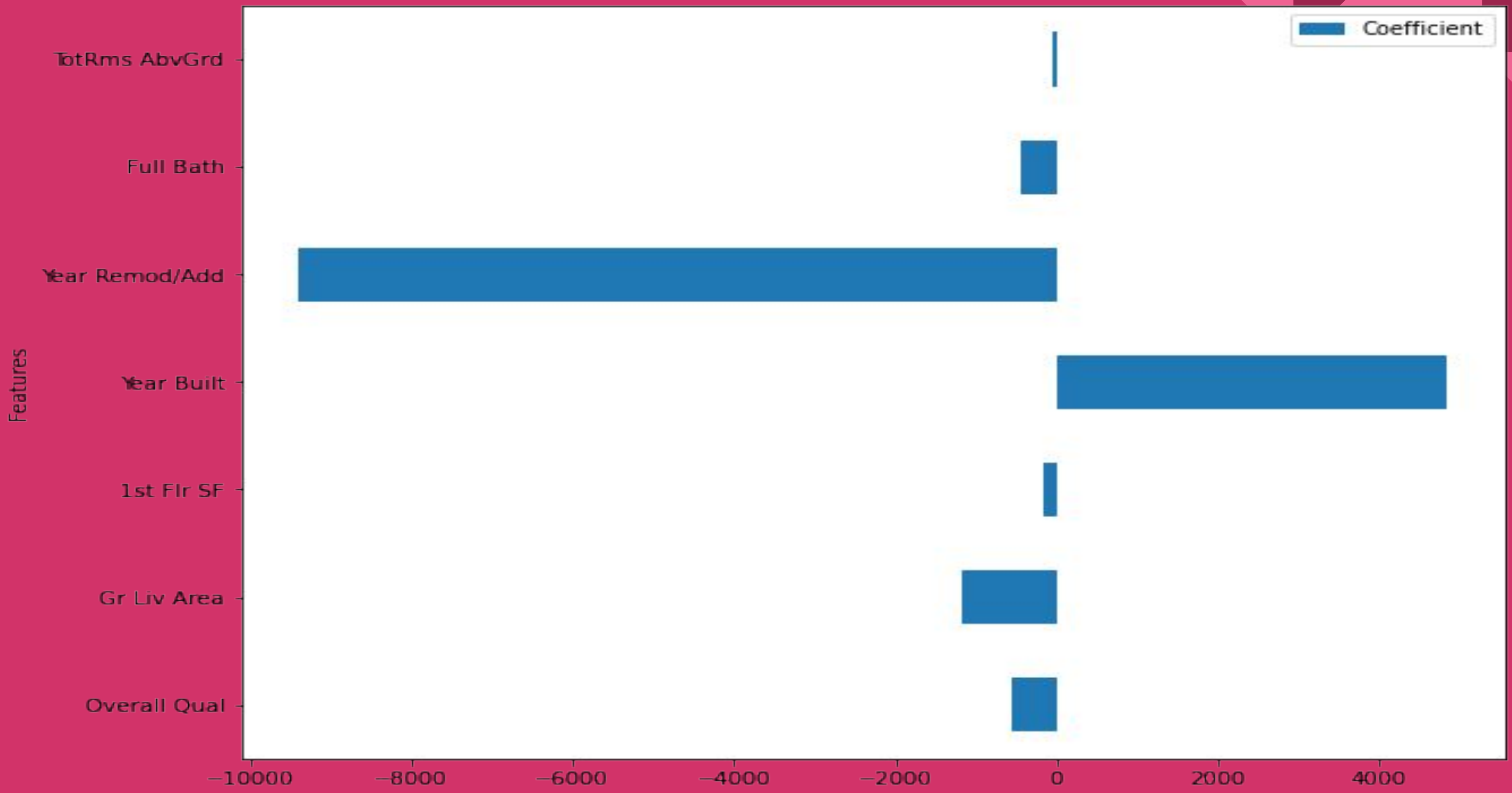
Model 1 results continued….

- Analysis based on the Ridge model:



Clear Homoscedasticity in the plot of the residuals

Model 1 Final Takeaways



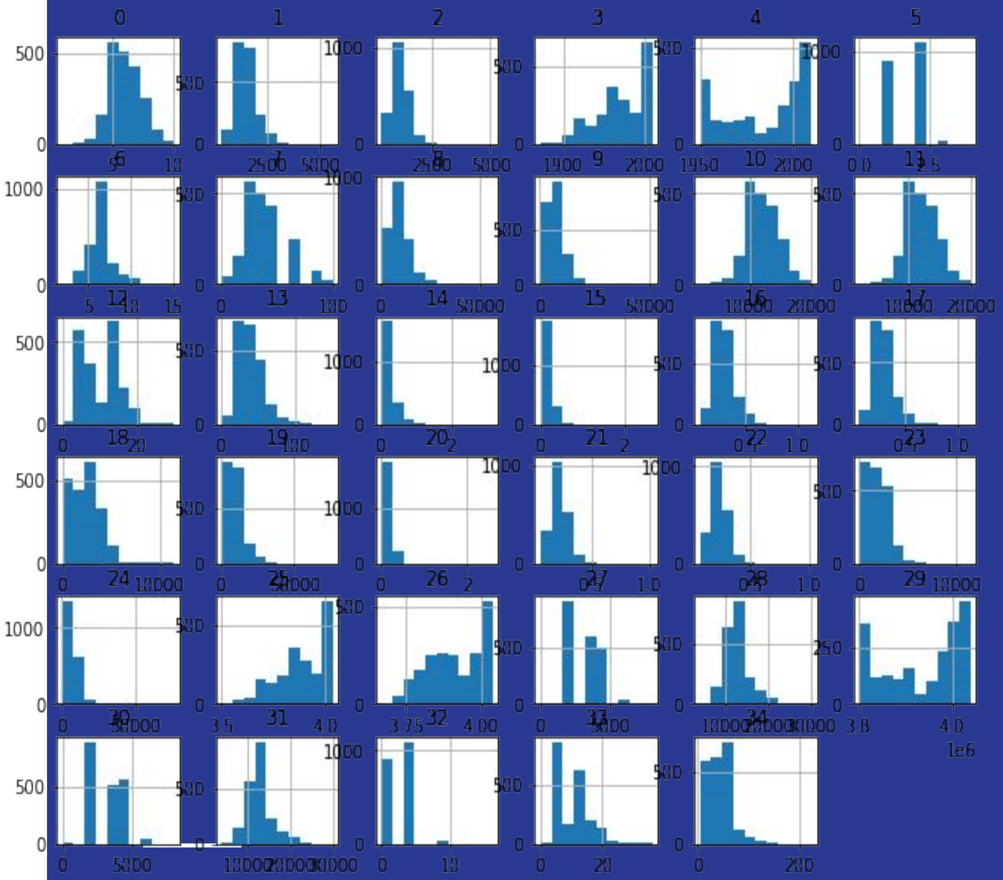
On to Model 2

Model 2

- Same Regressions were performed

BUT

- The features analyzed were transformed to also include interaction features.
- Total features went from 7 in model 1 to 35 in model 2.



Model 2 Results

Ridge Regularization produced the strongest model

Training RSS =0.865

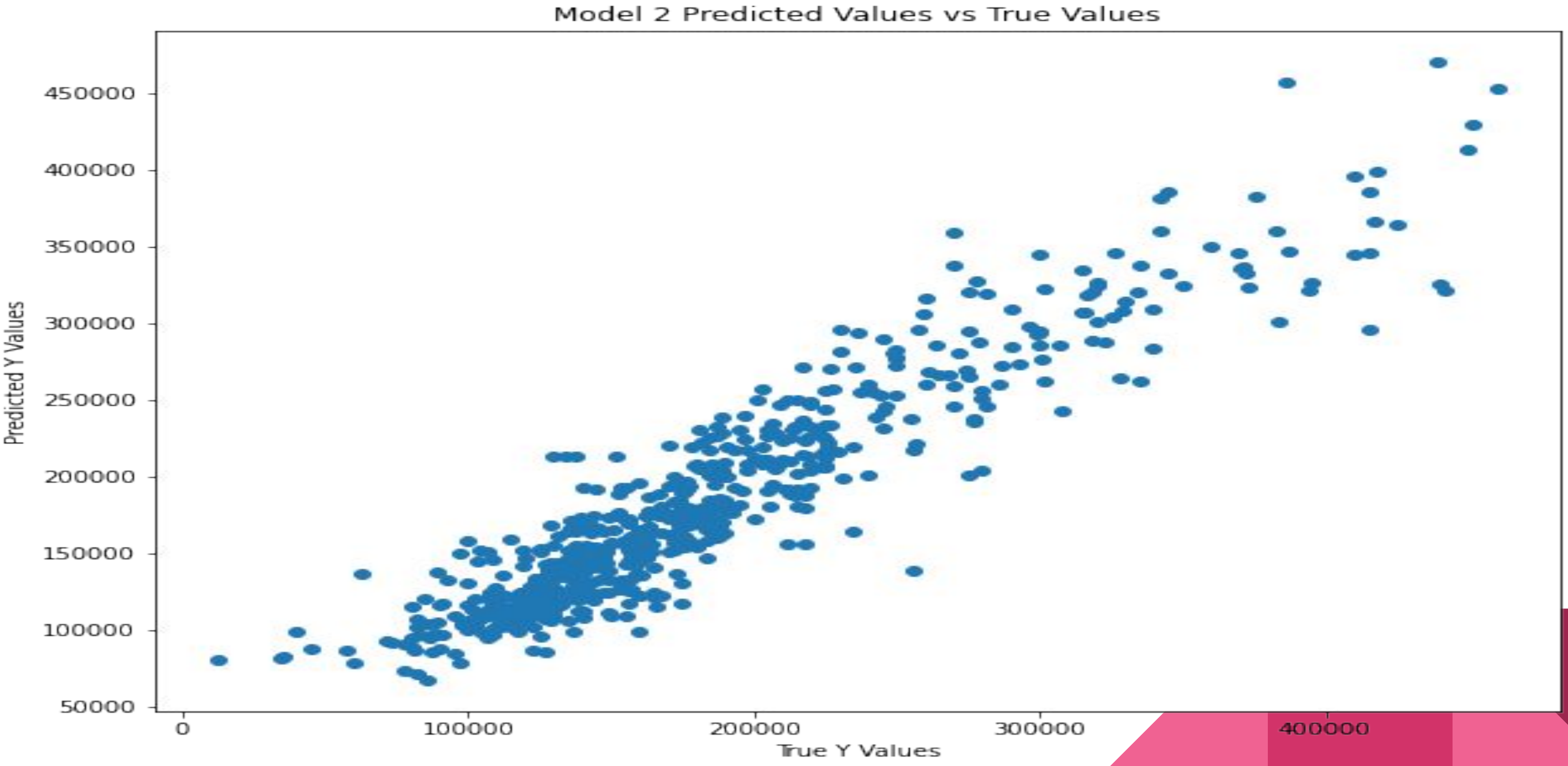
Testing RSS = 0.873

Training MSE = 848,291,059.5

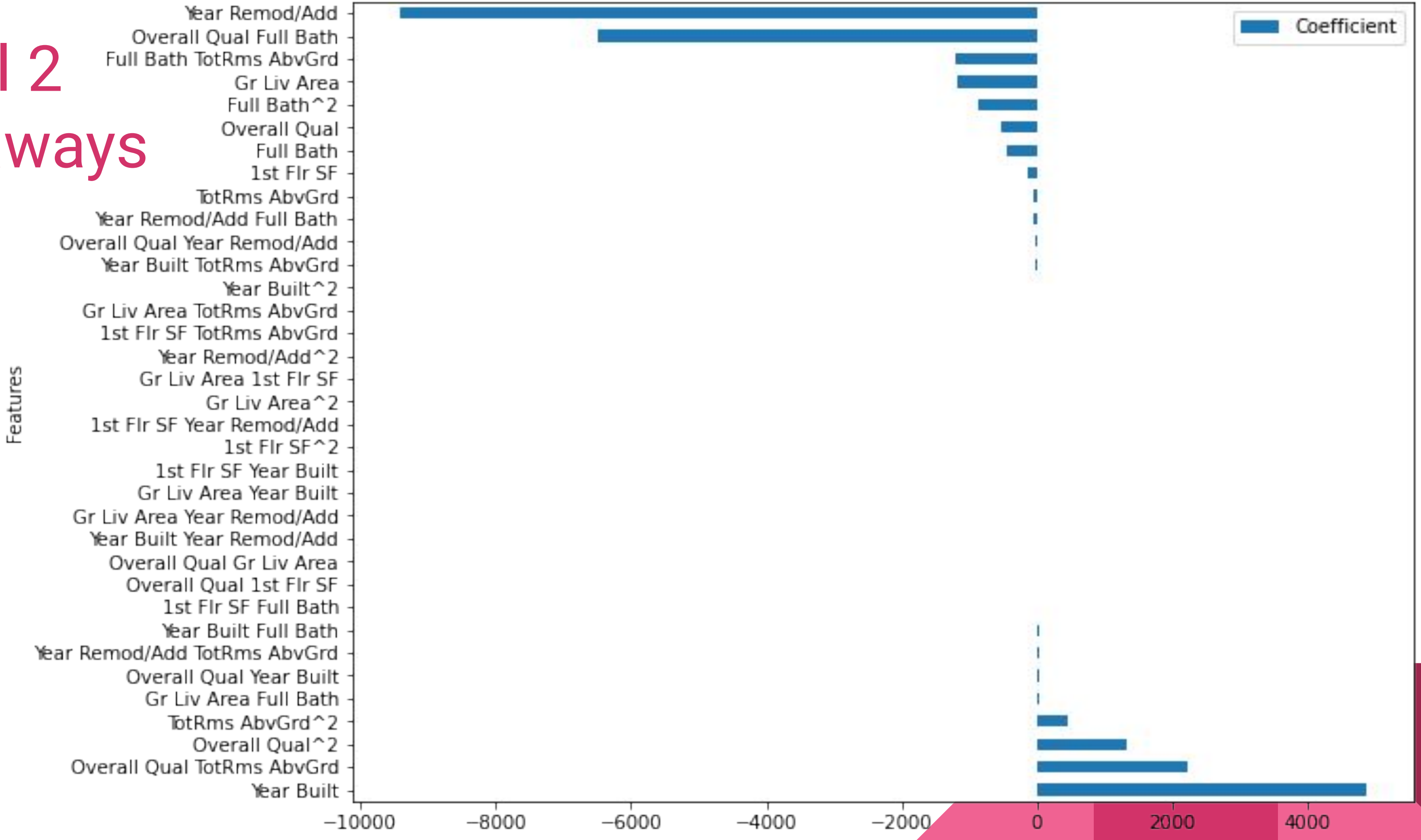
Testing MSE = 763,996,138.96

Again, this model is slightly underfit but a large improvement from model 1

Model 2 Results…..



Model 2 Takeaways



Model 2 was the better predictor and this is what it tells us:

When predicting selling price, the most important features are:

- Year Remod/Add : $ -9401.535226
- Overall Qual Full Bath: $ -6477.499430
- Year Built: $4850.314955
- Overall Qual TotRms AbvGrd: $2234.494347

Recommendations for further Analysis

- Neither model incorporated categorical features, so incorporating some could definitely add strength

- Imputing values instead of dropping could lead to a stronger model

- More research into interpreting interactions and how they work together will allow you to give the client more concise steps in case they want to add value.