Linear Regression Case Study Analysis

Neerja Doshi, Sri Santhosh Hari, Ker-Yu Ong, Nicha Ruchirawat

I. Data Cleansing

The Iowa housing dataset contains 1460 rows and 81 variables. Most of the variables are categorical - both numeric and character types - and only a handful are continuous. The response variable for our analysis is SalePrice, and the remaining 79 variables (excluding the record ID column) are considered potential predictor variables. Checking the data dictionary, we found the following distribution for the predictor variables:

- 49 categorical
- 19 continuous
- 11 discrete

As we explore % of NAs across these predictor variables, we first substituted NAs that are actually meaningful with 0s. According to the data dictionary, these NAs indicated non-applicability or a lack of the feature rather than missing data. We then re-check the count and percentage of NAs per variable left in the dataset as shown below:

Variable	Number of NA	Percentage of NA
LotFrontage	259	17.74 %
GarageYrBlt	81	5.55 %
MasVnrType	8	0.55 %
MasVnrArea	8	0.55 %
Electrical	1	0.07 %

We impute NAs in these variables with:

- Lot Frontage: mean of the data since it is a continuous variable
- GarageYrBlt: median of the data since it is discrete
- MasVnrType, Electrical: mode of the data since they are categorical variables
- MasVnrArea: since Masonry veneer area (MasVnrArea) is directly related to MasVnrType, we impute for area based on the mode of MasVnrType, which is None.

Having dealt with the NAs in our dataset, we use the model.matrix() function from the glmnet package to convert each categorical variable into an appropriate set of binary indicators: for a categorical variable that takes k levels, model.matrix() produces k-1 binary indicators. We then re-append our response vector SalePrice to the resulting data frame for use to run our regression.

II. Exploratory Data Analysis and Visualization

With our clean dataset, we perform exploratory data visualization of the distribution of key measures such as volume and sale price of houses by what we hypothesize to be key predictor variables.

To begin with, we check the distribution of sale prices for different neighborhoods using a box-plot:

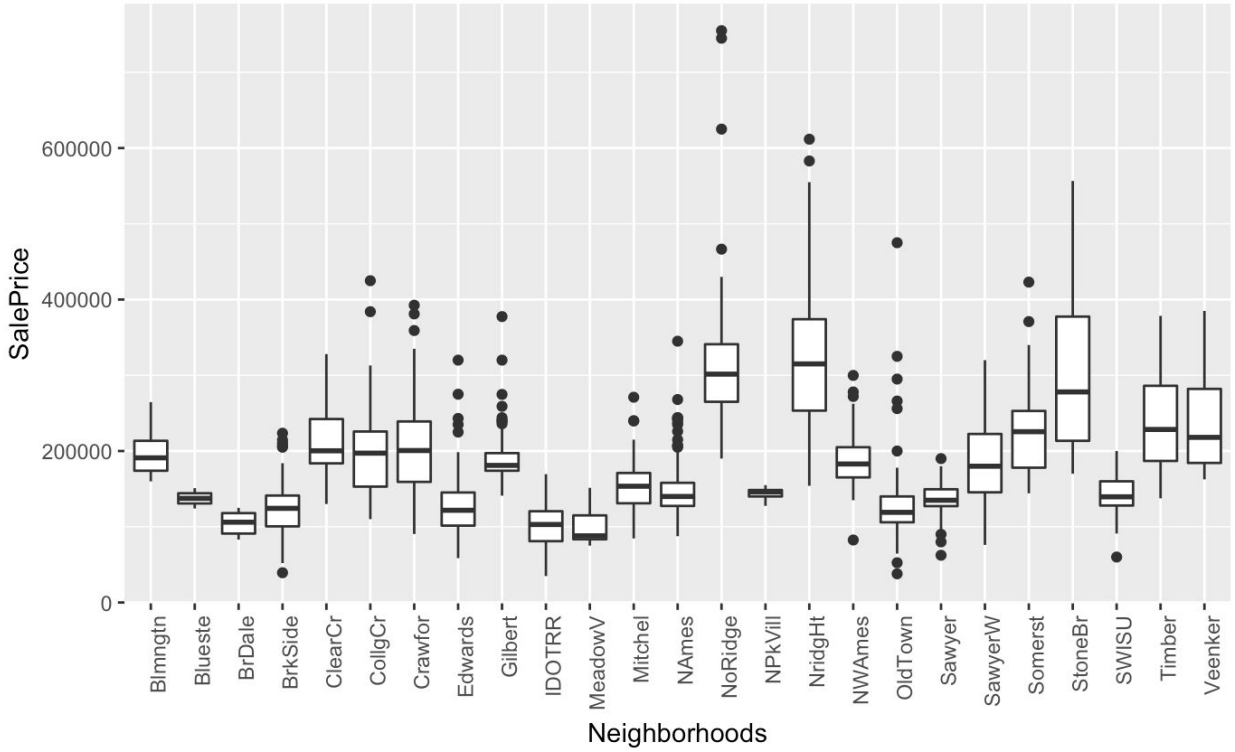

Figure 1: SalePrice distribution per neighborhood

From Figure 1, we can observe that Brookside and Meadow Vista have the lowest median house price while Northridge and Northridge Height have the highest median house price as well as several outliers.

Looking at figures 2 and 3, we can infer the following about the properties captured in the dataset:

- Most of the houses are located in `Medium/Low Density` residential areas
- Most of the houses have paved road access, do not have alleys
- Most of the houses have all public utilities(E,G,W,& S)
- Most of the properties are regular or slightly irregular in shape,
- Most of the properties are built on level surfaces with gentle slope.

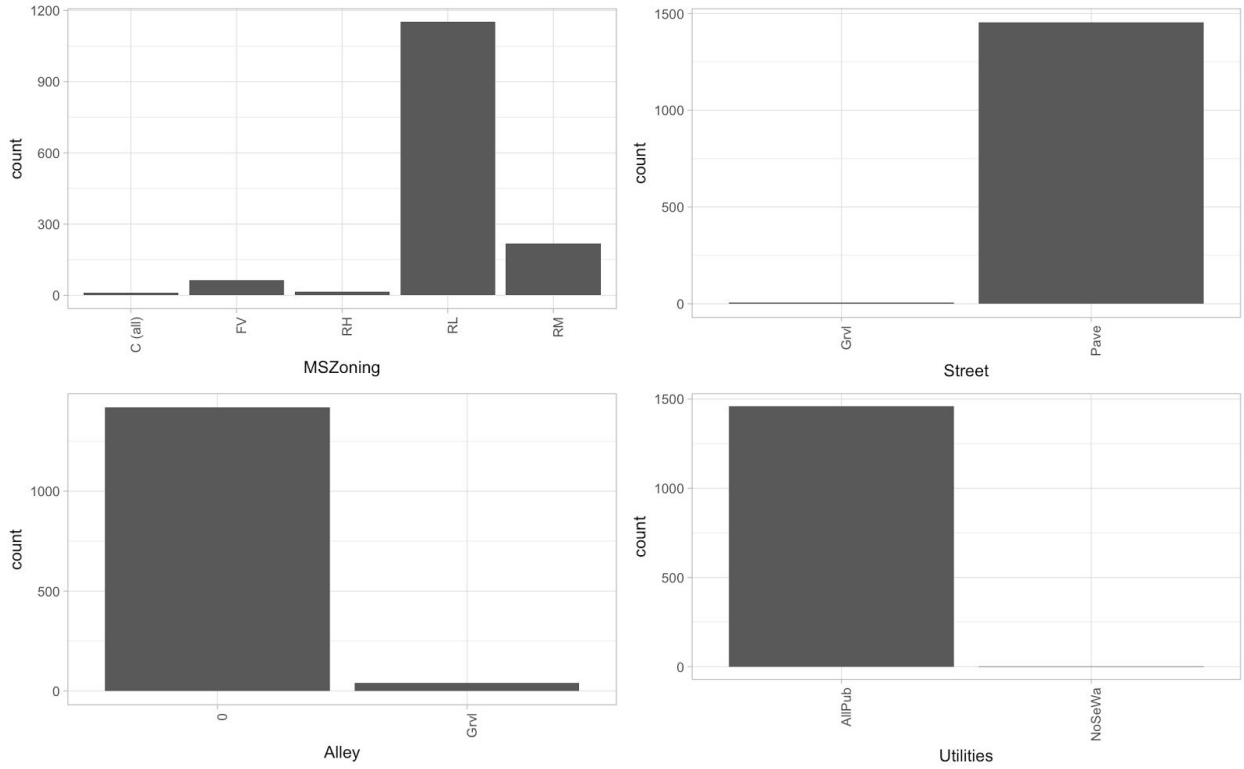

Figure 2: Distribution of Locality, access, utility features

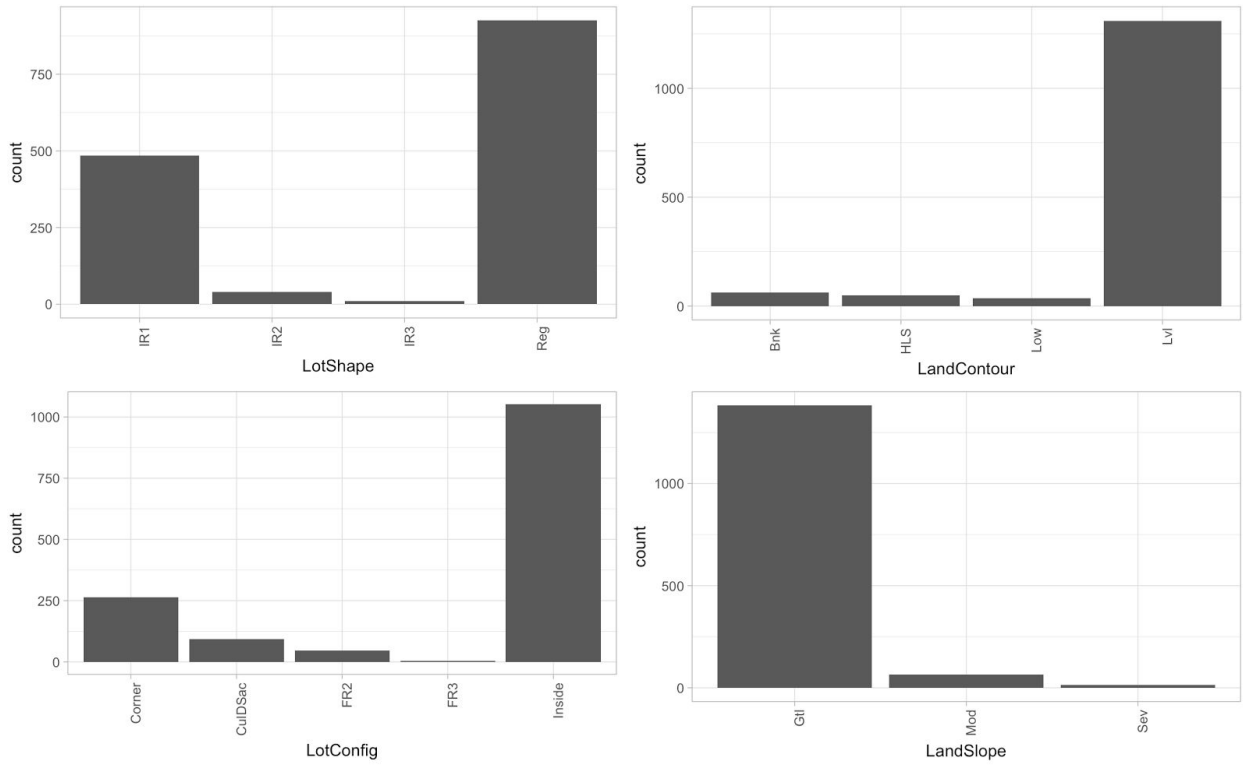

Figure 3: Distribution of Lot/Land features

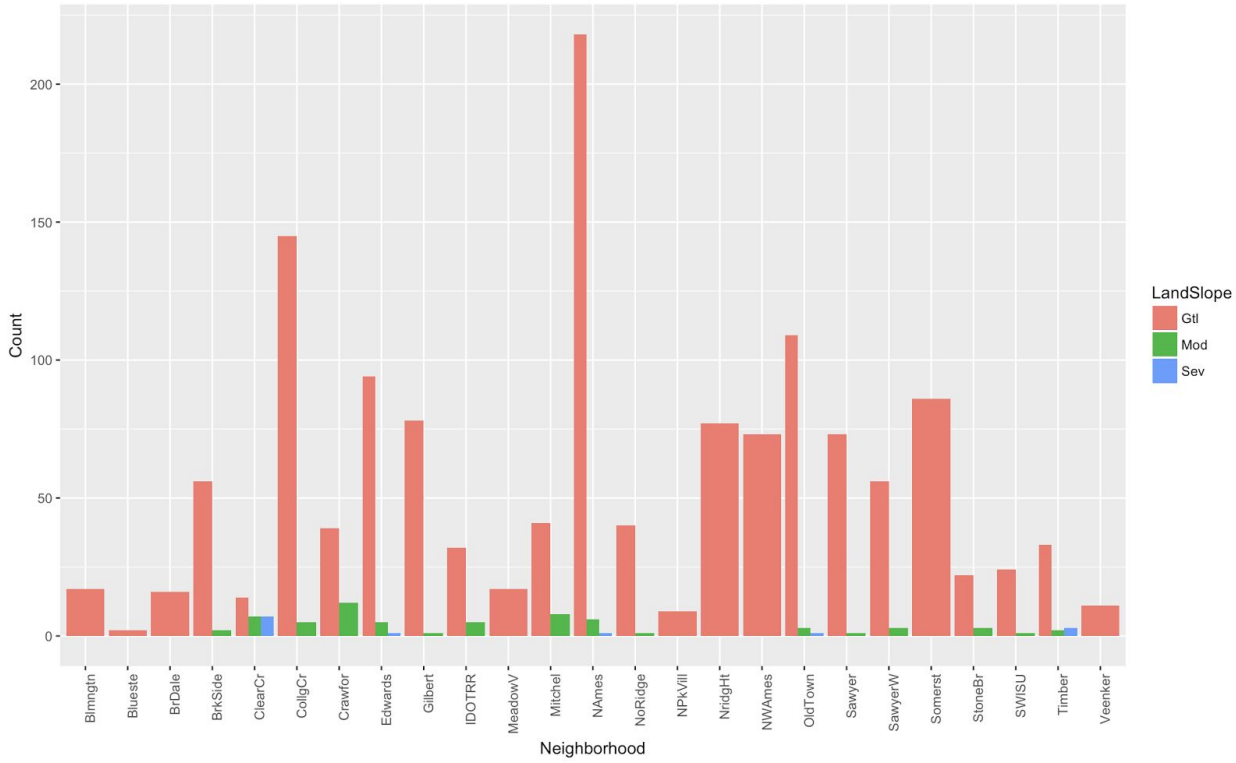

Figure 4: Neighborhood level slope distribution

From Figure 4, we can see that houses with severe slope are located only in Clear Creek and Timberland while more than 10 neighborhoods have properties with moderate slope.

<u>Following 10 have high positive correlation with the response variable (SalePrice):</u>

- OverallQual (overall quality of material and finish of the house)
- YearBuilt (year built)
- YearRemodAdd (remodel date)
- MasvnrArea (masonry veneer area)
- BsmtFinSF1 (rating of basement finished area)
- TotalBsmtSF (total square ft of basement area)
- 1stFlrSF (1st floor square ft)
- GrLiveArea (Above grade (ground) living area square feet)
- FullBath (number of full bathrooms above grade)
- TotRmsAbvGrd (number of total rooms above grade)

Figure 5 shows the scatter plot between SalePrice and the ten variables listed above. Corresponding R^2 values have also been listed. This shows that there is a linear relationship between response variable and the predictors. We therefore use BIC as our criteria for backward subset selection (discussed in next section).

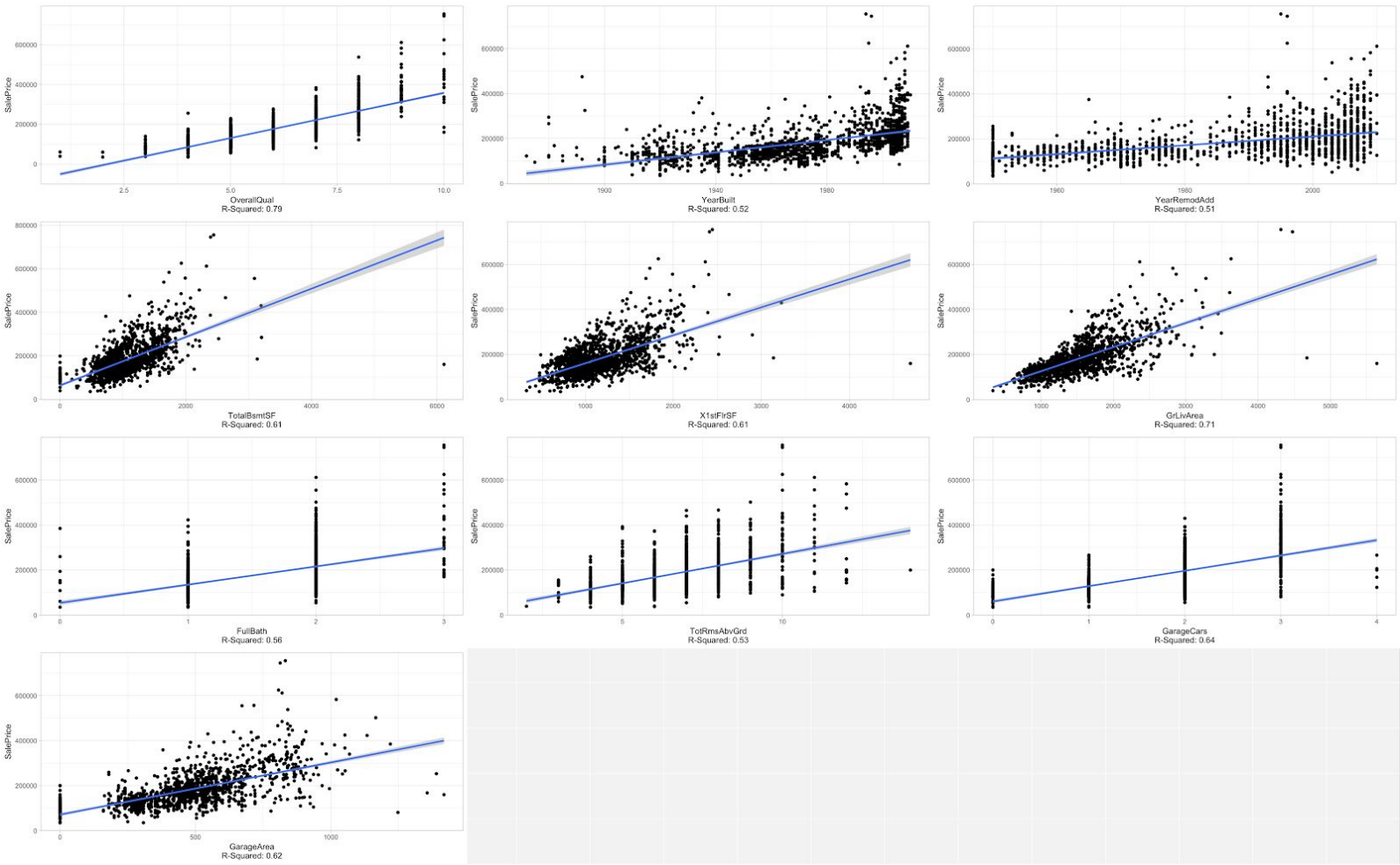

Figure 5: Scatter plot of variables showing high positive linear relationship with SalePrice

III. Explanatory Modelling

<u>Testing for Influential Points</u>

In looking for influential points, we tested observations for influence according to the DFFITS diagnostic. Note that according to the criterion of threshold t = 2*sqrt(n/p) = 0.88, the DFFITS plot shows a large number of influential observations. We also plot the standardized and studentized residuals to check these observations for being outliers and/or points of leverage respectively. Our plots of standardized (Figure 7) and studentized (Figure 8) residuals indicate that all observations are both points of leverage and outliers. We remove them from our dataset.

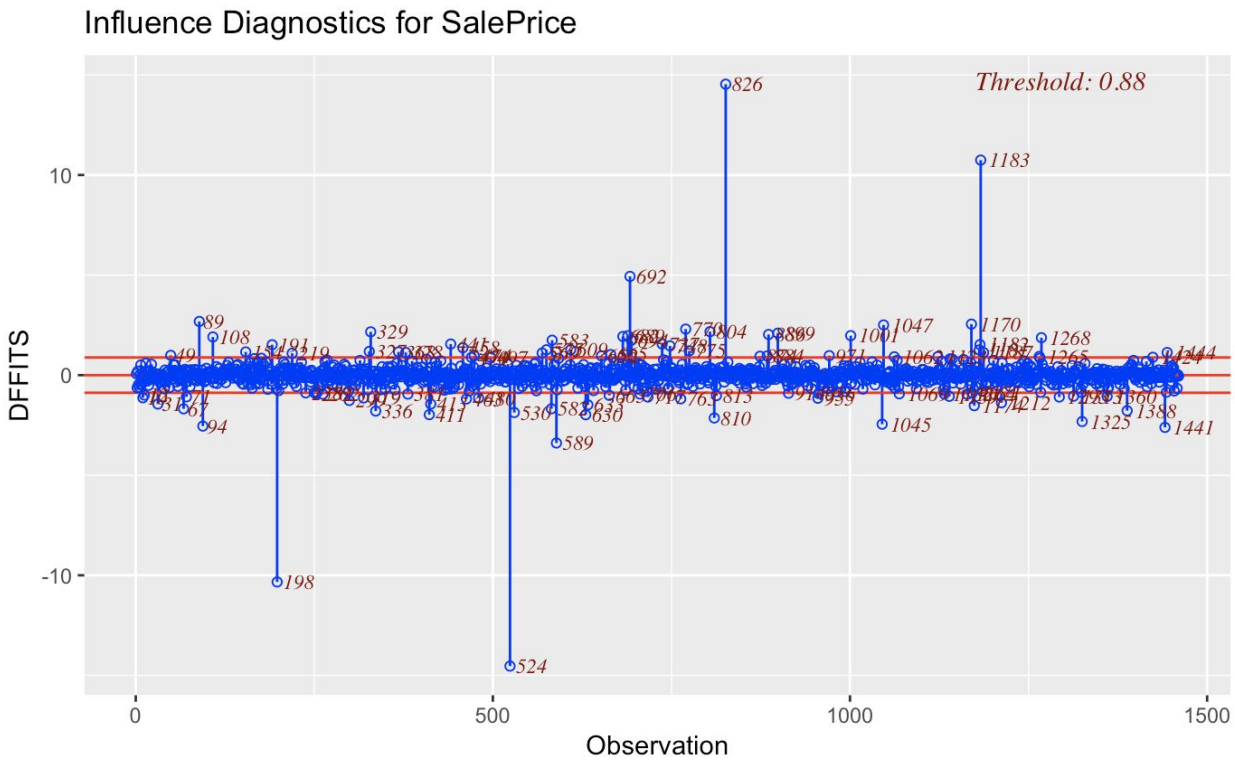

Figure 6: Difference in Fits for all data points calculated for saturated OLS model

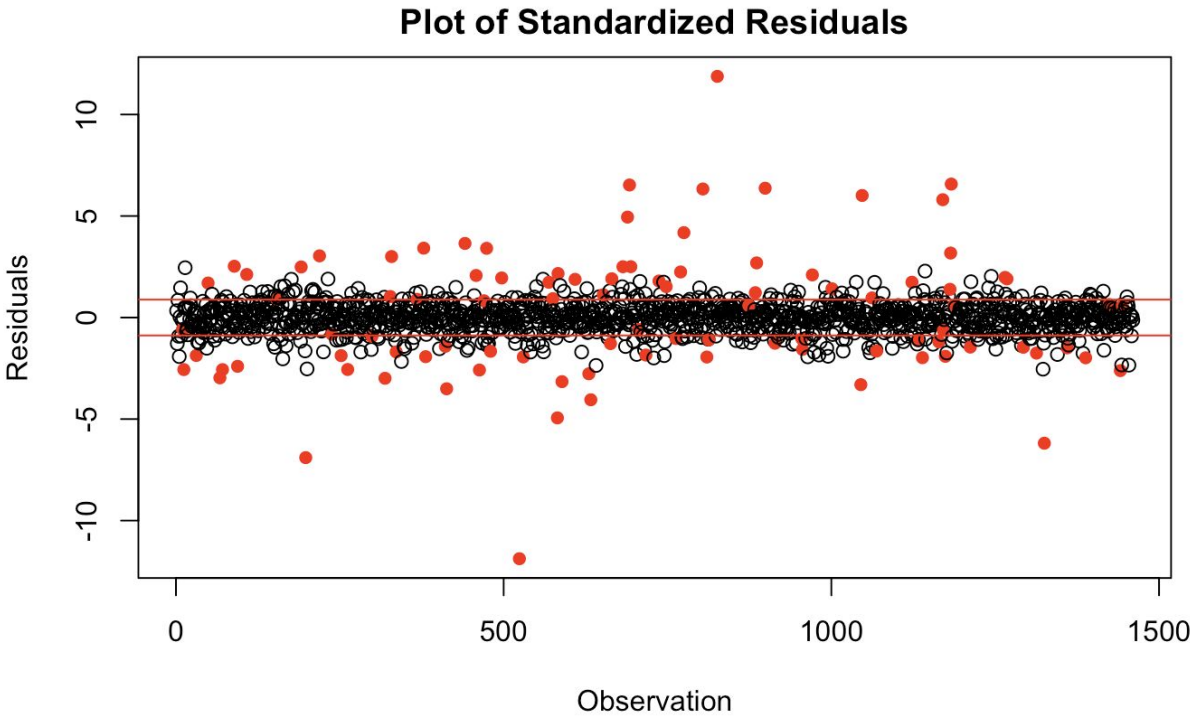

Figure 7: Standardized residuals for the saturated OLS model

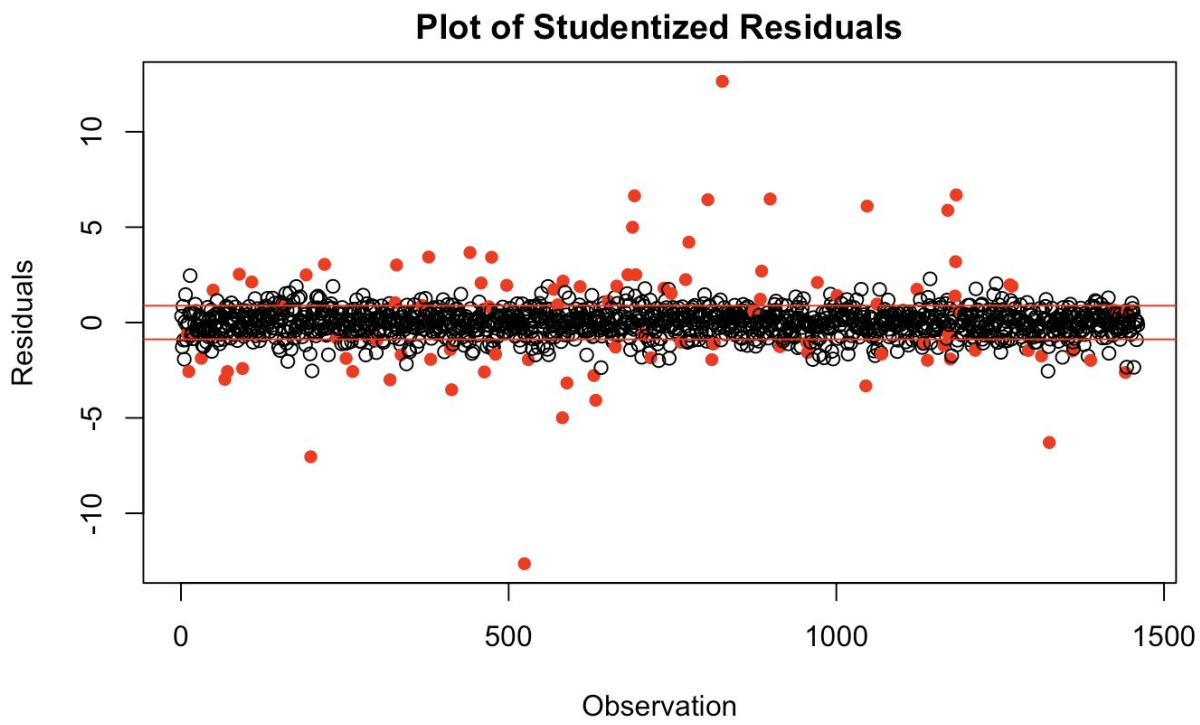

Figure 8: Studentized residuals for the saturated OLS model

Model Selection

For the purposes of variable selection, we refer to the saturated OLS model created above and perform stepwise model selection according to BIC criterions. We take only the most significant variables with p-value < 0.1. We perform OLS regression with our subset of 64 significant variables. We then check for multicollinearity in our model by checking for variables with Variance Inflation Factors that are over threshold value of 10 and verifying these by checking the Singular Value Criteria for multicollinearity. Based on the results of VIF, we drop one of the two variables (GarageQual, GarageCond) that are collinear. This removed 4 dummy variables. Therefore, our OLS regression is left with 60 significant variables. This yielded the multiple R-squared value of 0.9551, indicating that 95.51% of the variability in SalePrice around its mean is explained by the mode, i.e. by the predictor variables that have been included. This suggests a high-performing explanatory model. However, we need to validate the linearity and normality assumptions of our model by checking our residuals.

The plot of residuals against fitted values shows that for the most part, the residuals are evenly distributed across the y = 0 line. However, we see that as Sale Price increases, the residuals start to deviate homoscedasticity. More specifically, we see this deviation happen at approximately Sale Price = $350K, which our earlier summary showed to be between the variable's 3rd quartile and maximum. This suggests that for the last quartile of high-priced houses, the fitted regression model is not as adequate as it is for the reset of the population. The normal probability (QQ) plot corroborates this finding: it shows deviance from linearity at both tail ends of the residual rang, which suggests a heavy tailed distribution. This occurs

at both ends of the distribution, i.e. both extremely low-priced houses and extremely high-priced houses are pulling the distribution away from normality.

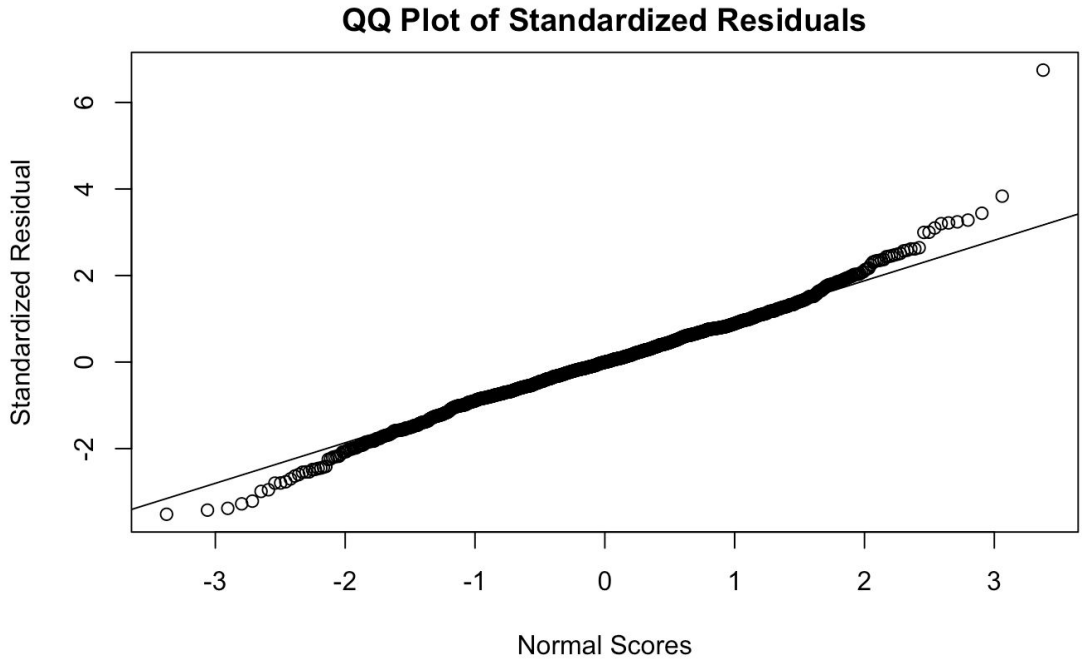

Figure 9: QQ Plot of standardized residuals of the OLS model on subset of variables

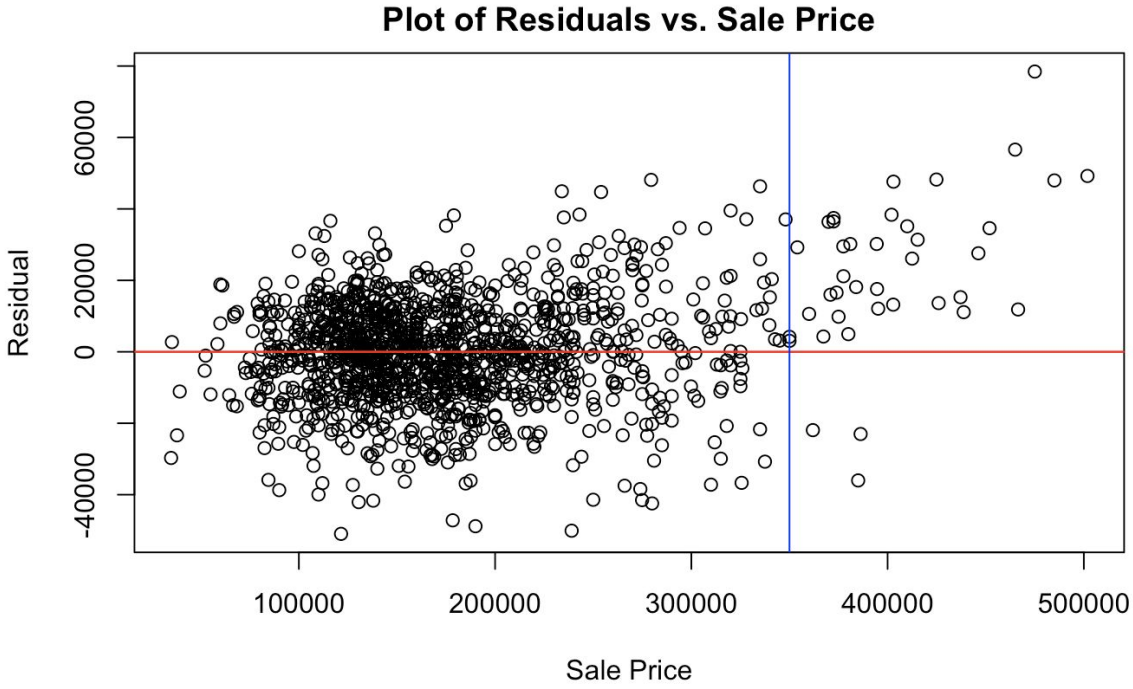

Figure 10: Plot of residuals vs response for the OLS model on subset of variables

As a remedial measure, we consider performing a transformation on Sale Price. We see that lambda = 1 is not captured in the 95% CI of lambdas, indicated by the three vertical dashed lines. This means a transformation is necessary. We choose lambda = ~ 0.5 since this is an interpretable transformation value. Therefore, we apply a square root transformation to Sale Price, refit the model, and validate our model assumptions with residual plots.

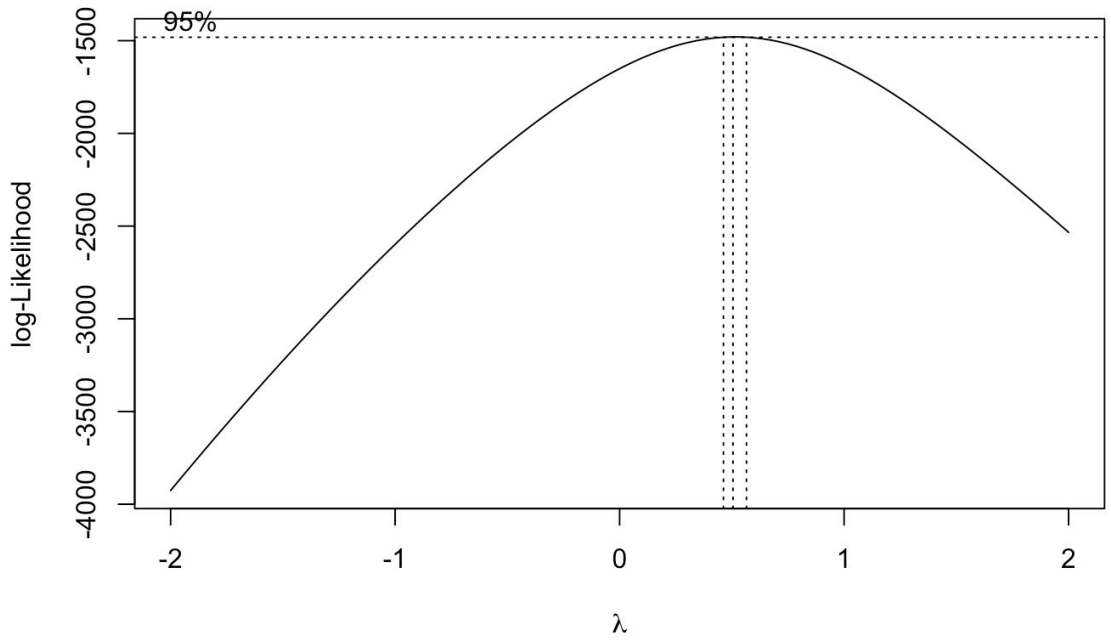

Figure 11: Plot of log-likelihood for different power parameter (lambda)

Our new model produces an R-squared value of 0.9573, indicating that 95.73% of the variance in Sale Price is captured by the model. Our residuals vs. fitted values plot shows a better pattern of homoscedasticity, which suggests that our linear regression model adequately captures the trend in the log-transformed data. The normal probability plot of the standardized residuals also shows better adherence to a linear pattern, which suggests that our assumption of normality is better. Figures 12 and 13 prove the validity of normality assumptions. We also perform the two sample Kolmogorov-Smirnov test to check goodness of the fit, which validates normality assumption.

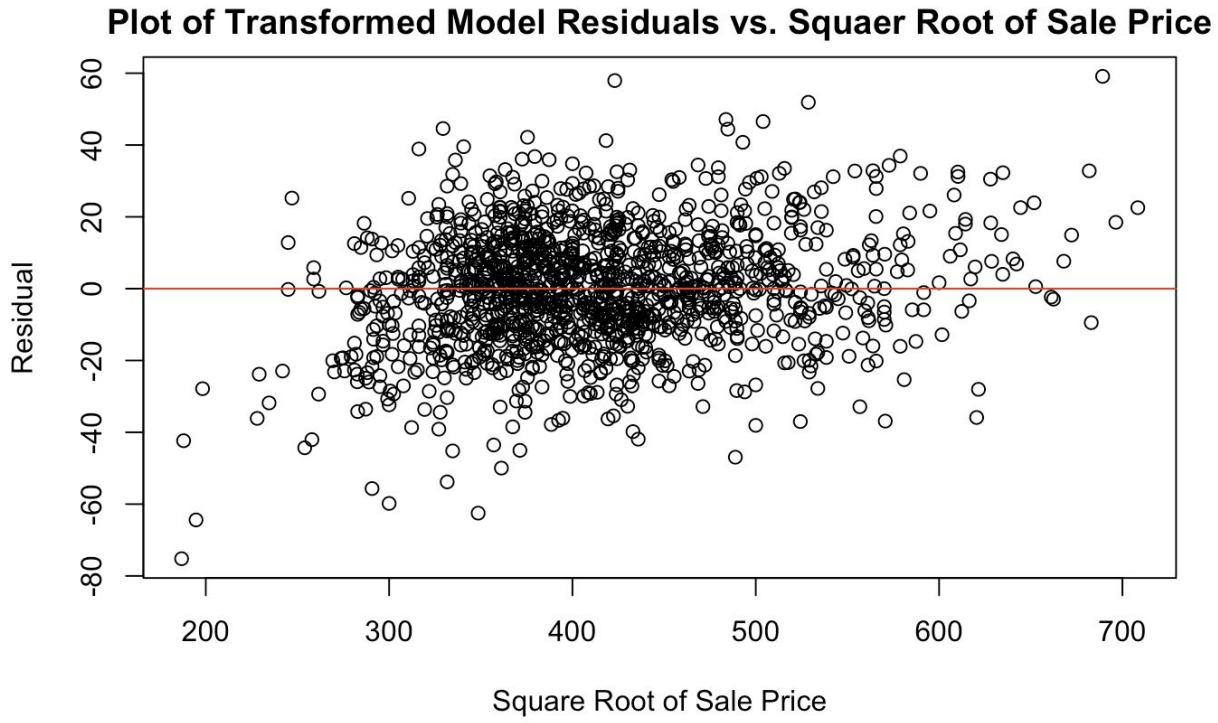

Figure 12: Plot of residuals vs response for the transformed OLS model on subset of variables

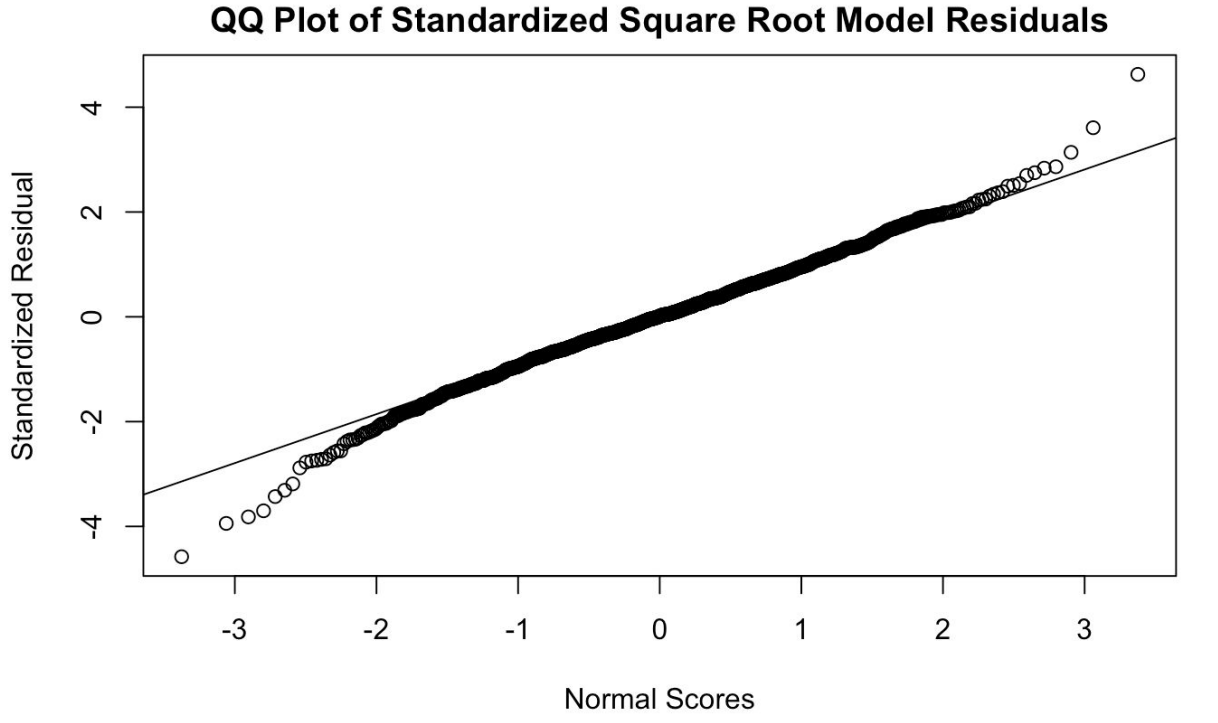

Figure 13: QQ Plot of standardized residuals of the transformed OLS model on subset of variables

Features from the final transformed model that affect house price include:

- Floating Village Residential or Residential Low Density zoning (-)
- Lot area (+)
- Paved street (+)
- Cul-de-sac lot configuration (+)
- Severe land slope (-)
- Neighborhood (+ or - depending on neighborhood)
- Overall quality of material and finish of the house (+)
- Overall condition of the house (+)
- House age (year built) (+ the newer the house, the higher the price)
- Certain roof materials (+)
- Brick face exterior (+)
- Having no masonry veneer type or stone veneer (+)
- Masonry veneer area (+)
- Quality of exterior material
- Basement quality
- Basement exposure (+)
- Floor size (+)
- Number of bedrooms/kitchen above grade (-)
- Having typical functionality (+)
- Garage characteristics
- Wood deck, screen porch, and pool area (+)
- Home being just constructed and sold (+)
- Normal sales condition (+)

Note: (+) means feature increases the price, (-) means feature decreases the price

Making recommendations for Morty's House

To make a recommendation to Morty regarding the selling price of his house, we leverage the model built above. We see that the 95% CI for the predicted Sale Price of Morty's house, based on selected attributes, is $153,613 to $168,014. As such, we recommend that Morty sell his house at a maximum of $168k. Figure 14 shows the coefficients of 25 significant variables in the transformed OLS model used to estimate the price of Morty's house

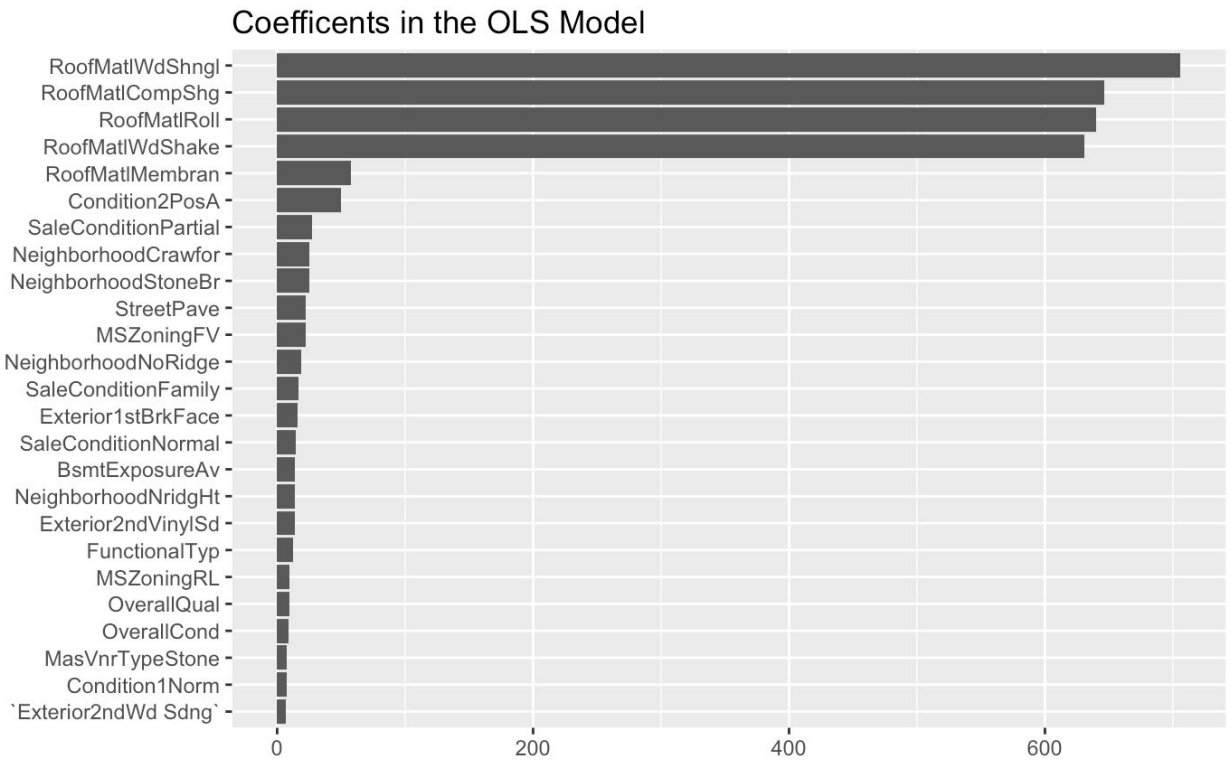

Figure 14: Top 25 Coefficients of significance in the transformed OLS Model

Coefficients of the linear model can be interpreted as the change in response variable for unit change in corresponding predictor variable, keeping all other variables constant. In current scenario, 10 unit increase in a predictor value will result in $ $(10*\text{coefficient})^2$ increase in the house price estimate (transformed model). Below are the three key recommendations and potential increase in selling price corresponding to the change:

1. Exterior remodelling with brick face can increase the selling price of house upto $196,890
2. Remodel basement to increase exposure (Av) can increase the selling price upto $180,093
3. Stone masonry veneer can increase the price upto $174,890

Recommendation	Minimum Price Estimate	Maximum Price Estimate
Exterior remodelling (Brick Face)	$ 172,170	$ 196,890
Basement Exposure	$ 164,217	$ 180,093
Masonry Veneer (Stone)	$ 158,929	$ 174,890

IV. Predictive Modelling

We explored 3 predictive models in order to find the model with the best prediction accuracy based on lowest MSPE:

1. Ridge Regression
2. Lasso Regression
3. Elastic Net

We fitted the models with 75% of the data being training set and calculated MSPE on the remaining 25% data as test set. For Ridge and Lasso regression, we iterate across all lambda values using cv.glmnet to choose the best lambda yielding lowest MSPE. For Elastic Net, we first iterate across alpha values between 0 and 1 in 0.1 increments to decide alpha value that yields the lowest MSPE. Then, using that alpha value, we iterate across all lambda values using cv.glmnet again to choose the best lambda. Given above procedure, we obtained the following results:

	Model Description	**MSPE**
Ridge Regression	263 variables $\lambda = 52607.68$	945,229,885
Lasso Regression	94 variables $\lambda = 1032.71$	894,807,404
Elastic Net	97 variables $\alpha = 0.6$ $\lambda = 1721.184$	886,861,924

It appears from above table that the Elastic Net model built performs the best in prediction, given it has the lowest MSPE. The selected alpha is 0.6 and lambda is 1721.184, with 97 variables used as predictors.