**Predicting Home Prices in Ames, Iowa**

**Sashank Bandhakavi**

**Ben Chittick**

**Ankit Kumar**

**Jacob Wehlman**

**SCMT 650:600**

**Group 1**

Contents

[Introduction 3](#_Toc500282789)

[Missing Data 3](#_Toc500282790)

[Covariance 4](#_Toc500282791)

[Exploratory Data Analysis 5](#_Toc500282792)

[Univariate Analysis 5](#_Toc500282793)

[Log Transformation 5](#_Toc500282794)

[Bivariate Analysis 9](#_Toc500282795)

[Outlier Detection 9](#_Toc500282796)

[Ordinal Correlation 11](#_Toc500282797)

[Combining Variables 12](#_Toc500282798)

[Category Reduction 13](#_Toc500282799)

[Multivariate Analysis 15](#_Toc500282800)

[Models 17](#_Toc500282801)

[Stepwise Selection Method 17](#_Toc500282802)

[Linear regression 18](#_Toc500282803)

[SVM 19](#_Toc500282804)

[Decision Tree 20](#_Toc500282805)

[Random Forest 21](#_Toc500282806)

[Bootstrap Sampling and Bagging 21](#_Toc500282807)

[Boosted Random Forest 22](#_Toc500282808)

[Penalized Models 23](#_Toc500282809)

[Ridge Regression 23](#_Toc500282810)

[Lasso Regression 25](#_Toc500282811)

[Model Summary 27](#_Toc500282812)

[Conclusions 28](#_Toc500282813)

[Flipping Houses 28](#_Toc500282814)

[What should the target list price be on homes 28](#_Toc500282815)

[Renovations 29](#_Toc500282816)

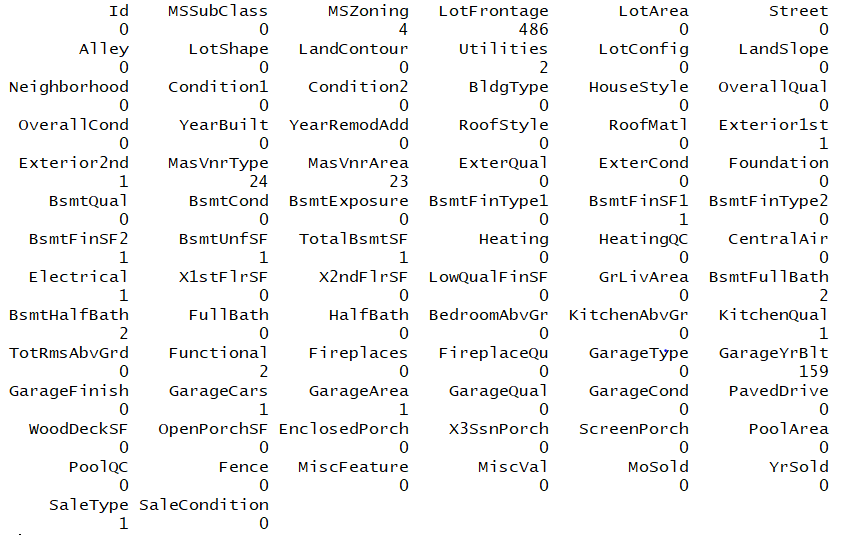
# Introduction

Our data set came from a Kaggle case competition. The data was collected on homes all over Ames from 2006 to 2010. The training set provided by Kaggle contained 1,453 observations, which we further broke into a validation set of 436 observations. This is a 70/30 split of the training set. The test set, also provided by Kaggle, contained 1,459 observations. The data set describes almost every aspect of home with 79 explanatory variables and a single response variable of sale price. The explanatory variables range from the livable square footage of the home to descriptors of the property and neighborhood. The breakdown of explanatory variables is as follows:

* 14 discrete
* 20 continuous
* 23 categorical
* 23 ordinal

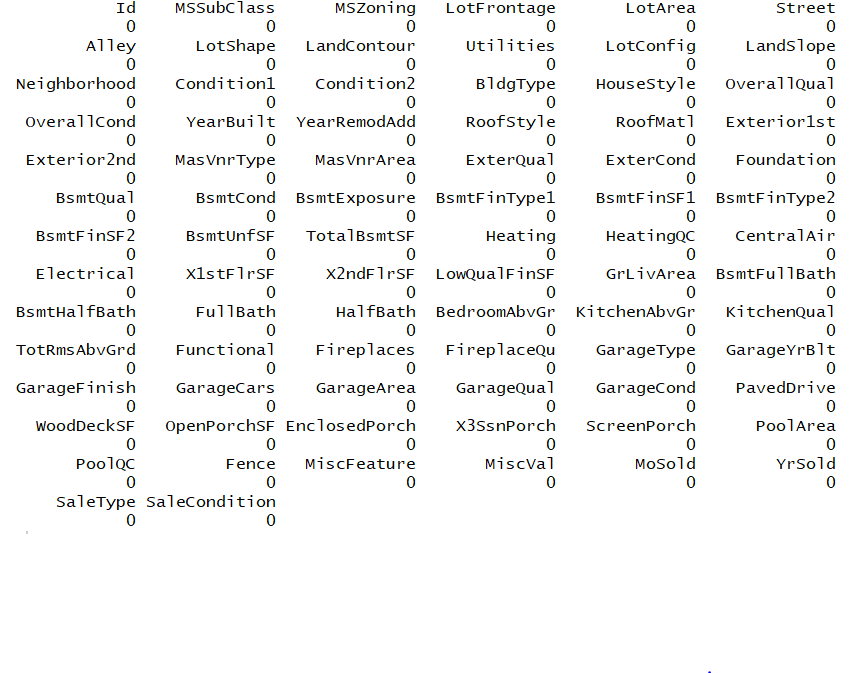
# Missing Data

Our data set did contain some missing data, and we used the R-package “Matrix” to generate a count of the number of missing values in each variable (**Figure 1.1**).



**Figure 1.1 Missing values in the data set**

Once we identified the missing values, we used the R-package “Mice” for imputation of the missing values. After running this function, we filled in all missing data (**Figure 1.2**):

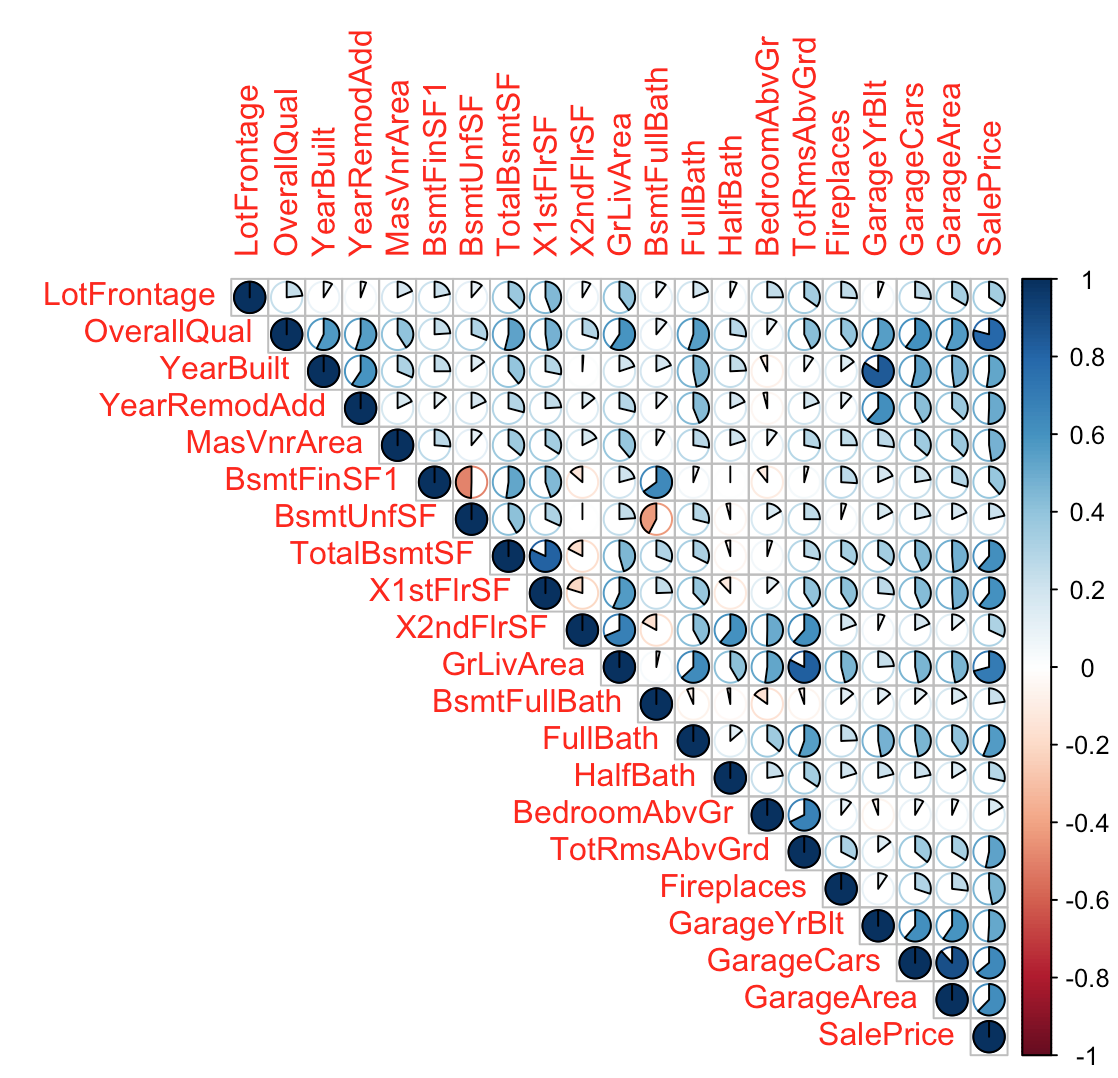


**Figure 1.2 Data set after imputation**

# Covariance

Our first step in understanding our data, after fixing any data issues, was to see if any of the variables were correlated with each other or with sales price. We generated several matrices like the one shown in **Figure 1.3** to gauge the covariance of our variables. What we were able to derive from the matrix is as follows:

* All of the square footage variables were highly correlated
  + TotalBsmtSF: Total square footage of basement area
  + 1stFloorSF: First floor square footage
  + 2ndFloorSF: Second floor square footage
  + GRLiveArea: Above-ground living area
* Some of the variables were highly correlated with sales price
  + GRLivingArea
  + FullBath: Number of full bathrooms
  + TotalBsmtSF
  + OverallQual: Ordinal rating of the house’s finish
  + OverallCond: Ordinal rating of the house’s condition



**Figure 1.3 Covariance plot of numeric variables**

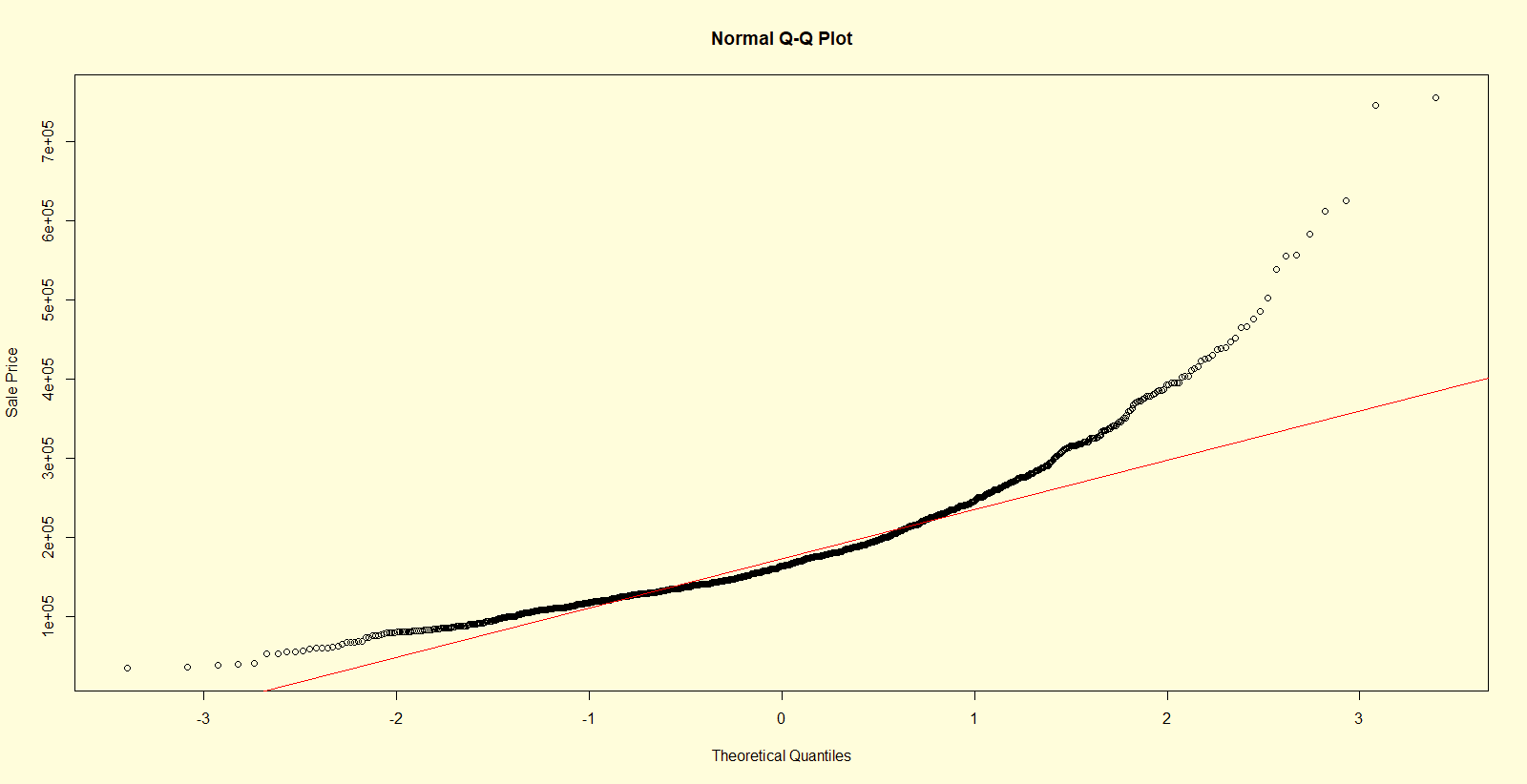
# Exploratory Data Analysis

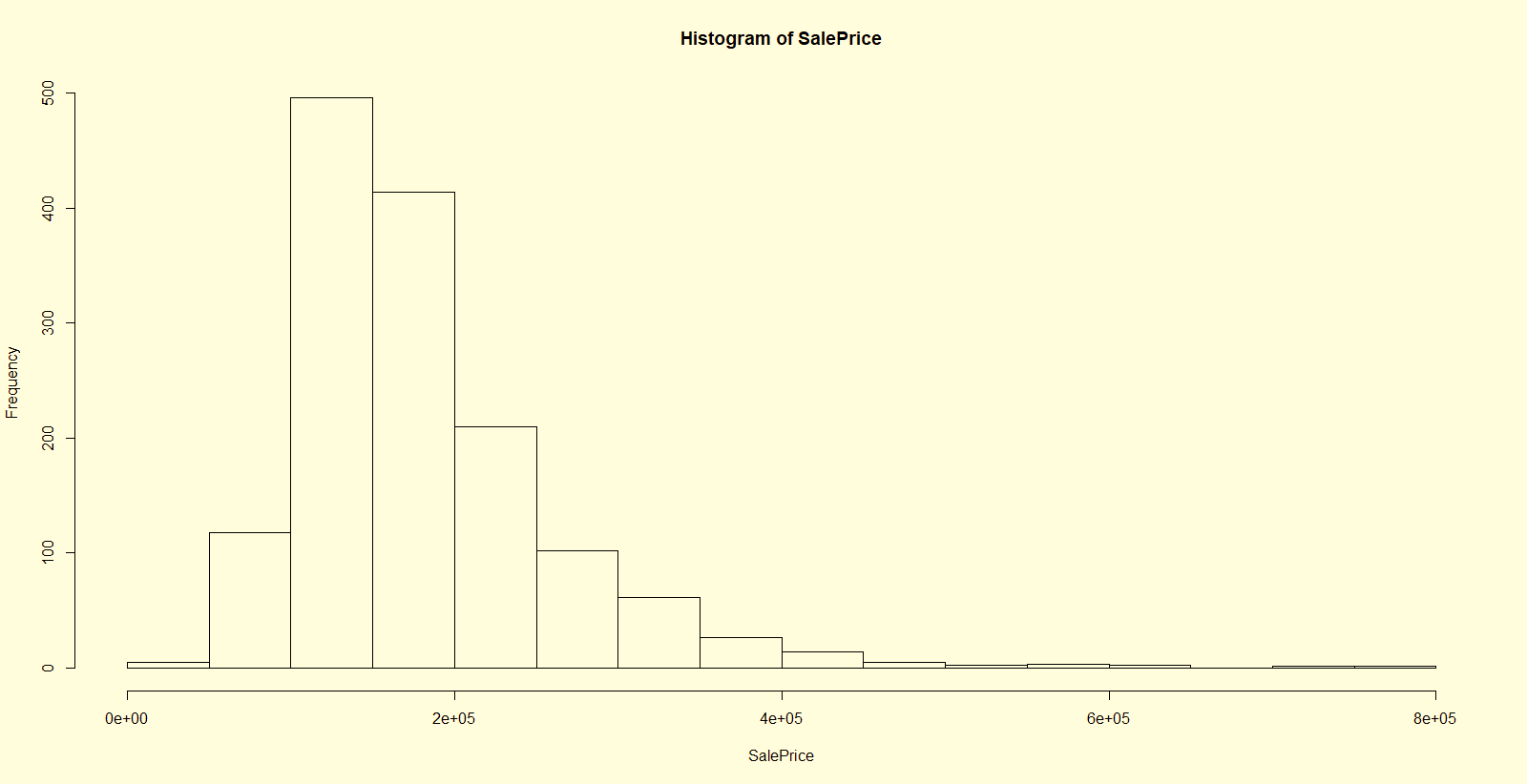
## Univariate Analysis

### Log Transformation

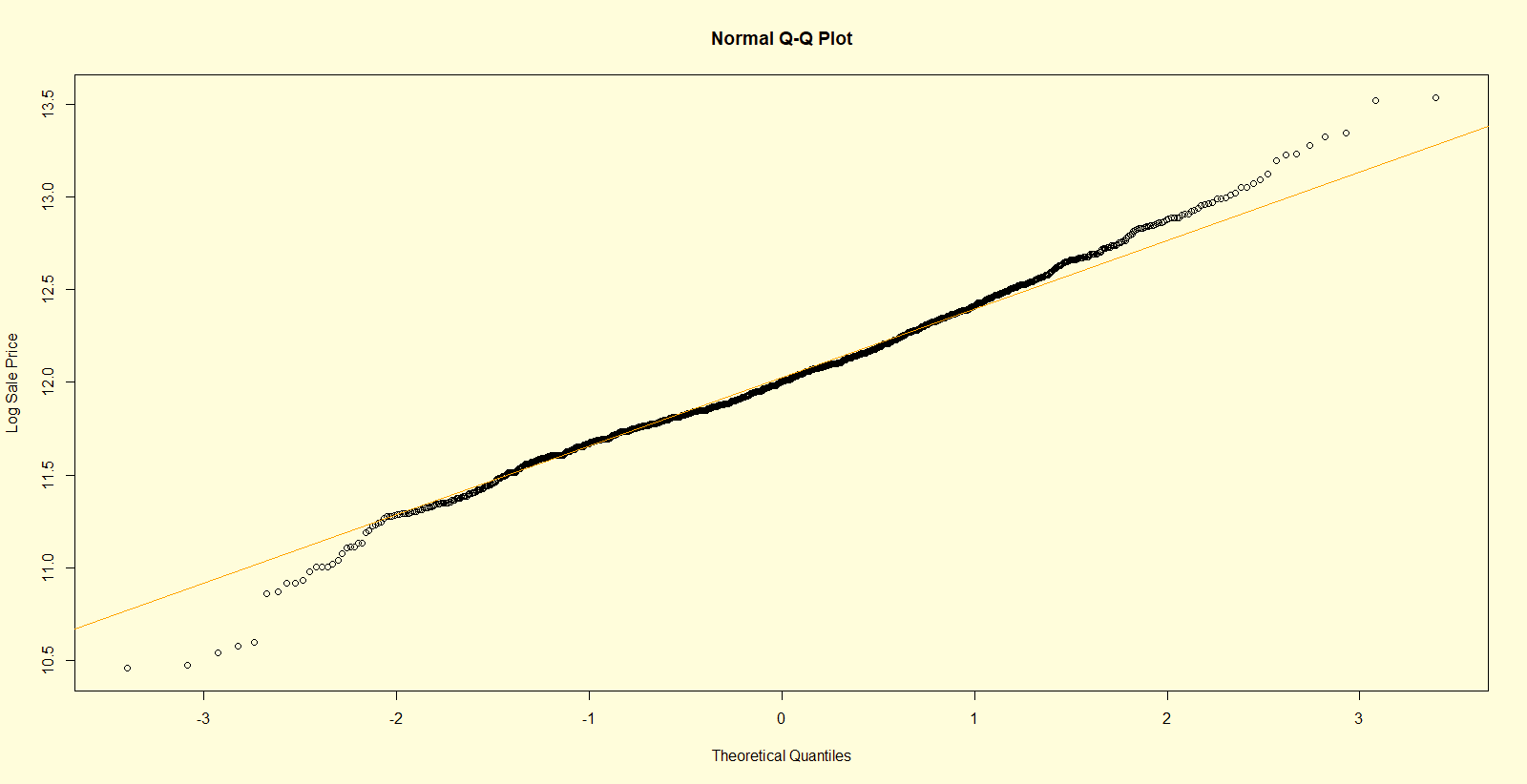
The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics.

SalePrice variable has high skewness and high kurtosis. Even though it’s just the SalePrice that is not normal, we can also say that entire data is not multivariate normal. Multivariate normal is one of the assumptions of linear regression. Using SalePrice as it is may lead to poor accuracy while performing linear regression. Hence a log transformation is necessary to make SalePrice a normal distribution and also meet other assumption of linear regression such as linearity and homoscedasticity. The same transformation was followed for variables basement square footage and above ground living area. As shown in **figures 2.2, 2.5, and 2.6,** each of those variables have skewed distributions on their histograms and high kurtosis values. From **figures 2.1 and 2.3,** the Q-Q plots for each of those variables were skewed from the straight line expected in a normal Q-Q plot. **Figure 2.4, 2.5, and 2.6** show distribution of log transformed variables. They clearly have low skewness and kurtosis.

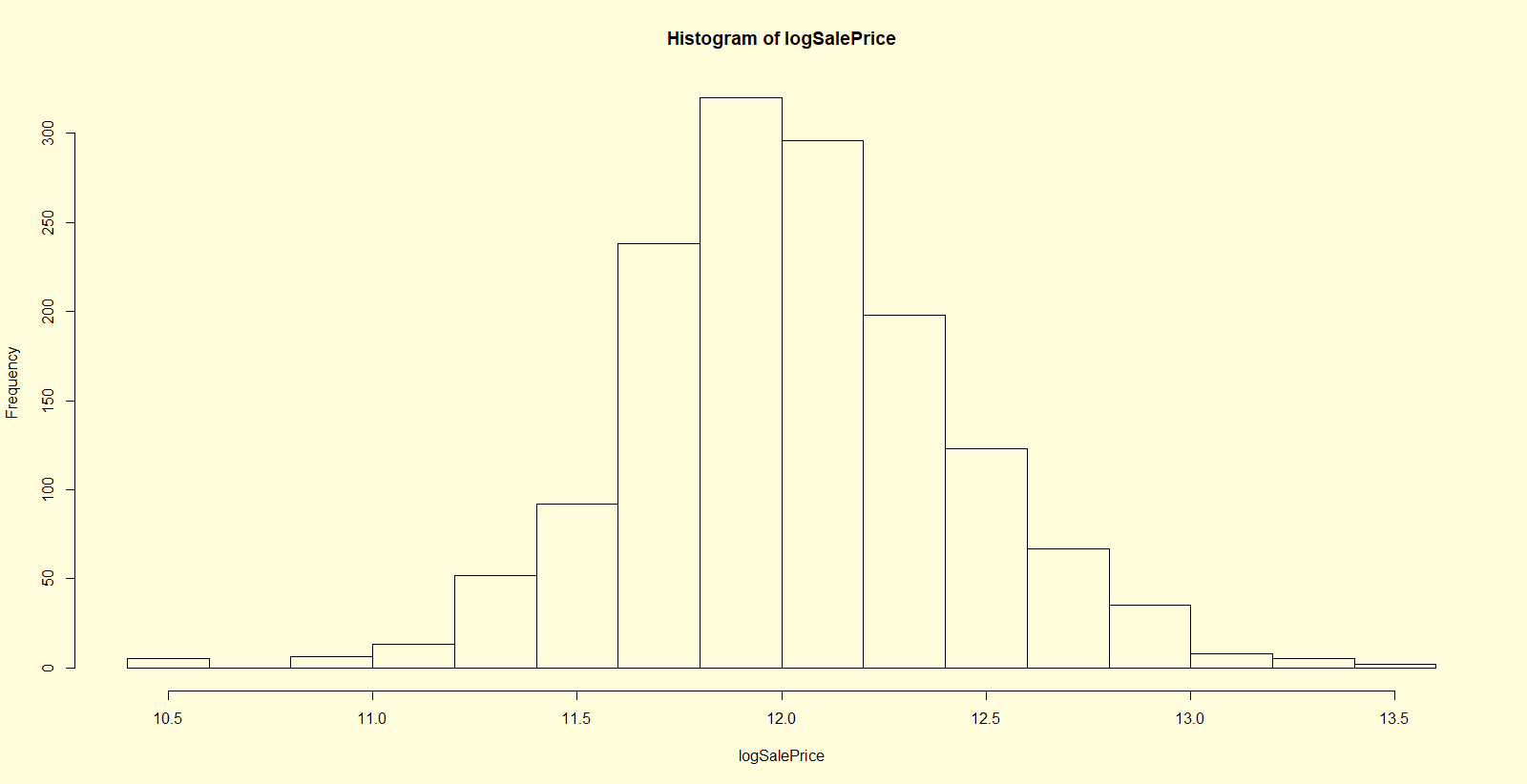


**Figure 2.1 Q-Q plot of SalePrice**

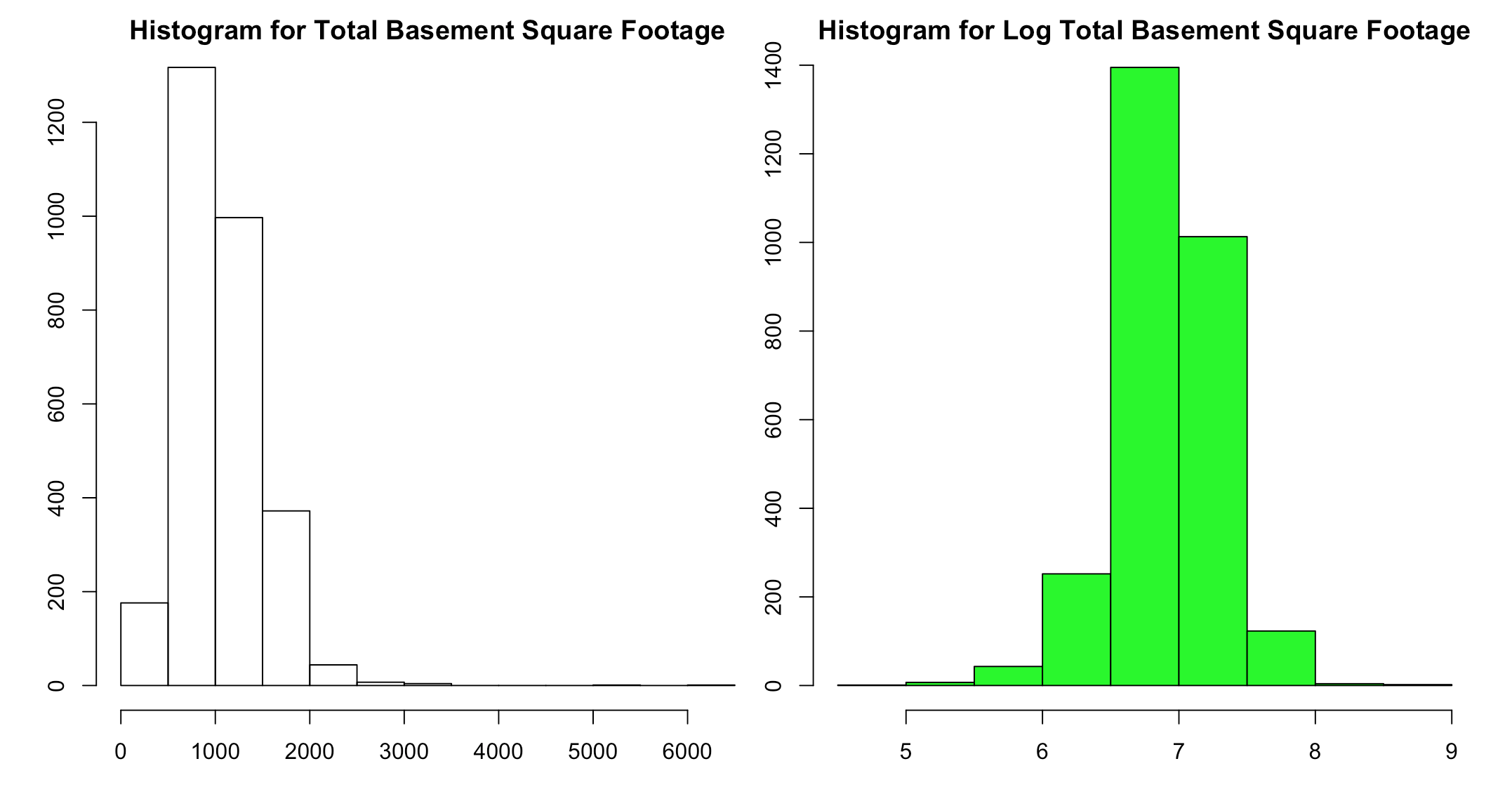
**Figure 2.2 Histogram of SalePrice**



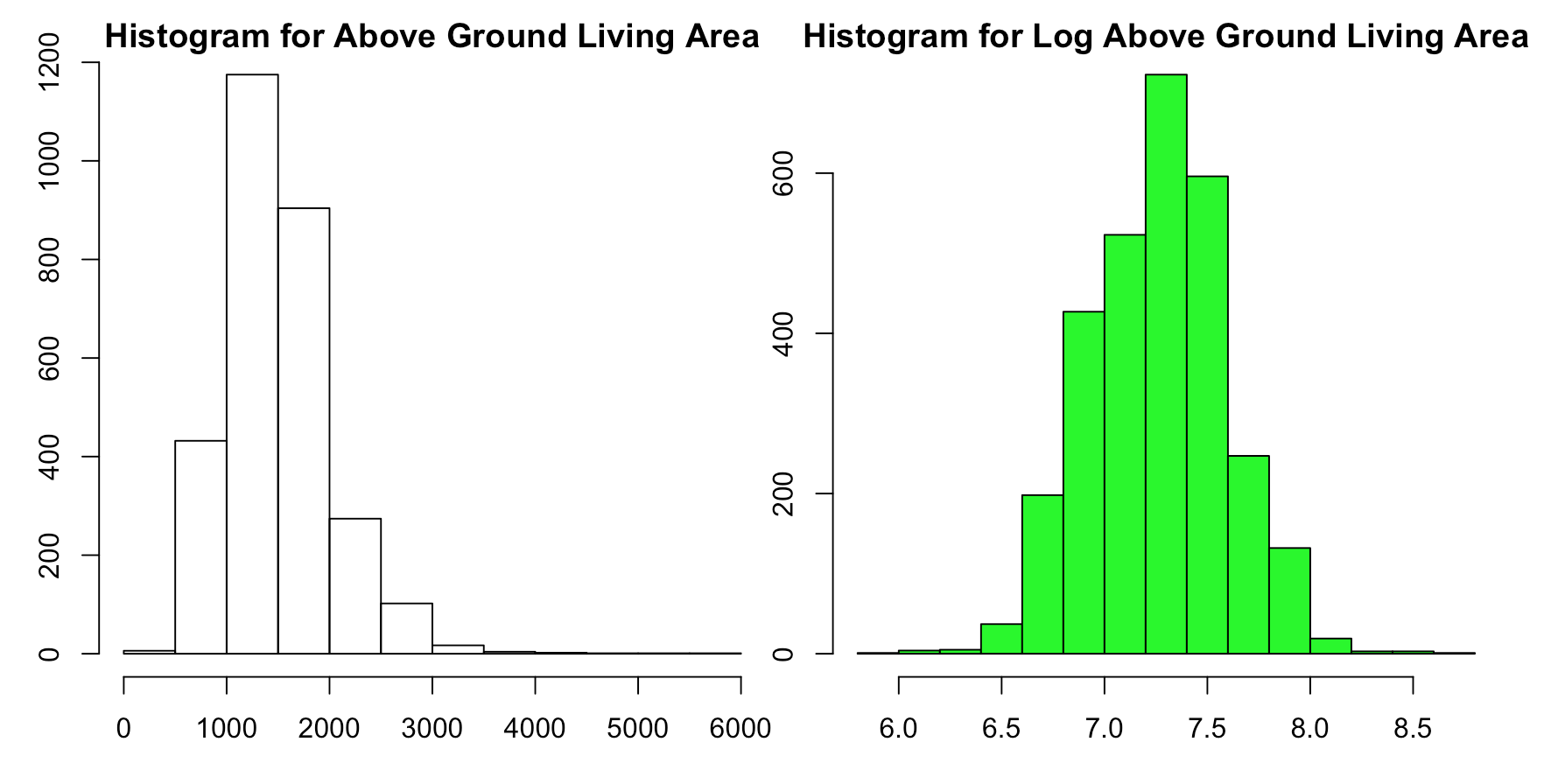
**Figure 2.3 Q-Q plot of logSalePrice**



**Figure 2.4 Histogram of logSalePrice**



**Figure 2.5 Histogram Total Basement Square Footage and its Log values**



**Figure 2.6 Histogram Above Ground Living Area and its Log values**

# Bivariate Analysis

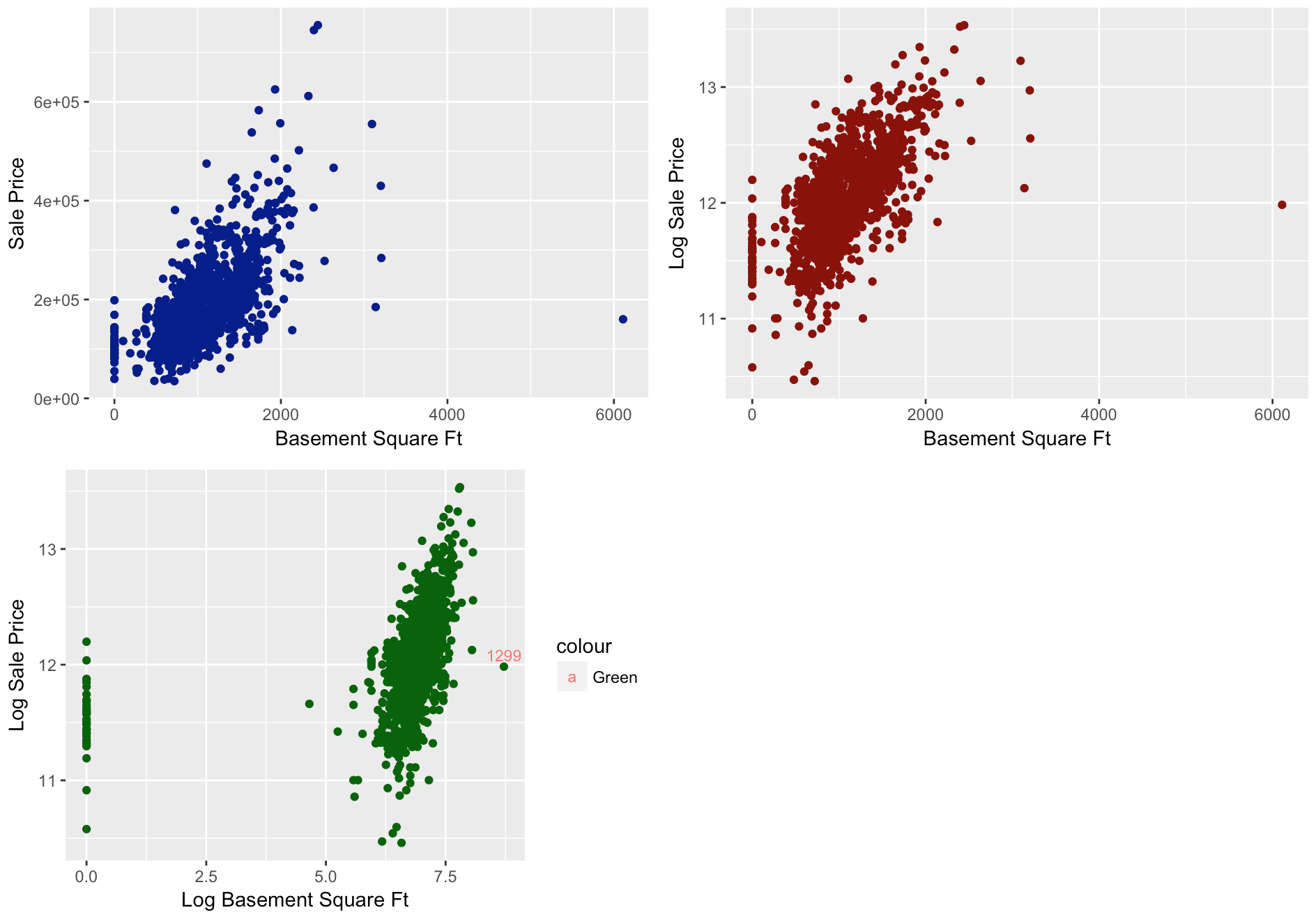
**Figures 2.7, 2.8, 2.9, 2.10** highlight the advantages of doing log transformations. Before transformations all the scatter plots have a satellite shape, which indicates heteroscedasticity. After log transformation, the scatter plots seem more linear.

## Outlier Detection

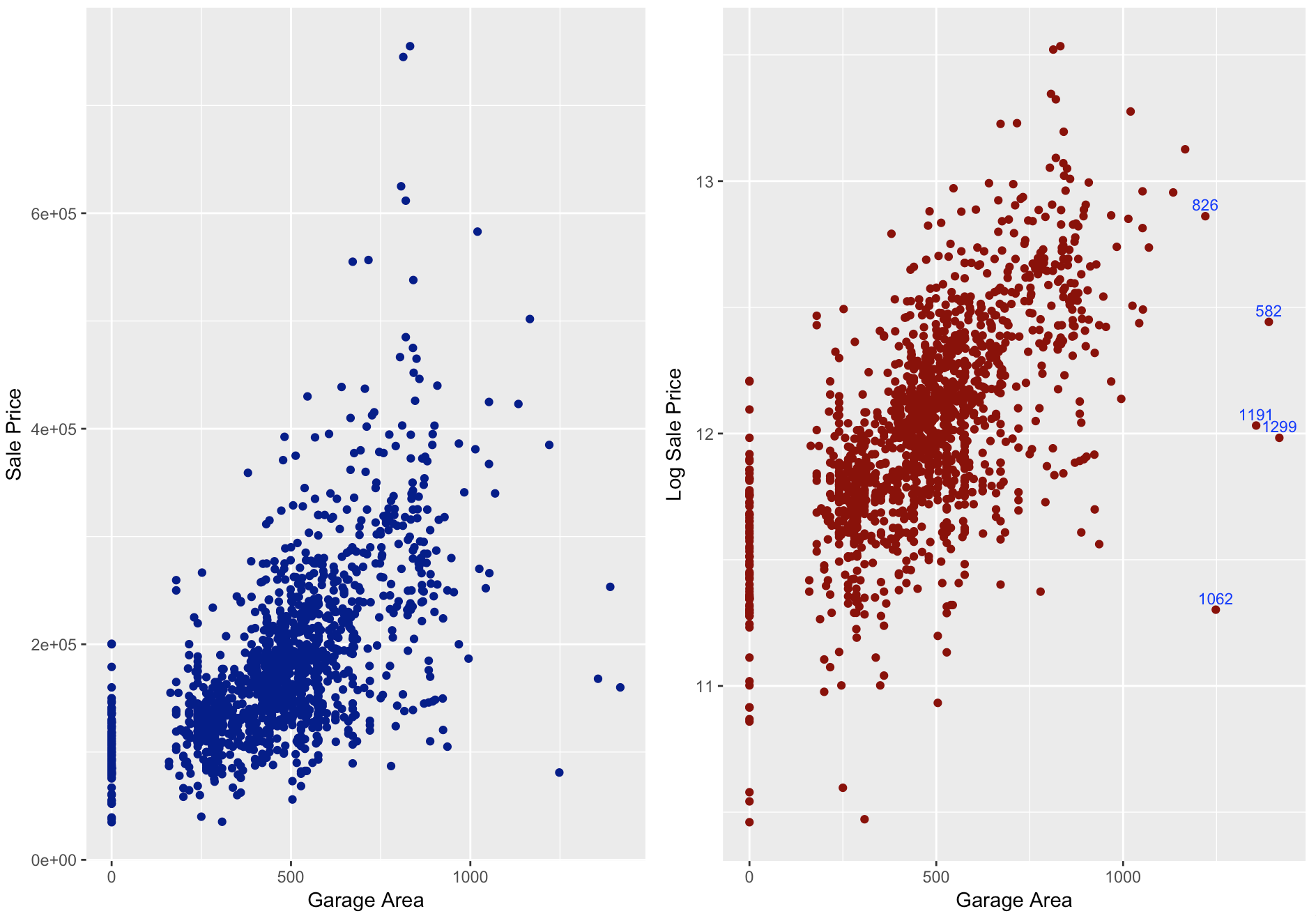
**Figure 2.7, 2.8, 2.9, 2.10** indicate that data rows with rows ids 1191, 1299, 582, 1191, 1062, 186, and 1454 are possible outliers. We removed these rows from the data since outliers act as leverage points.



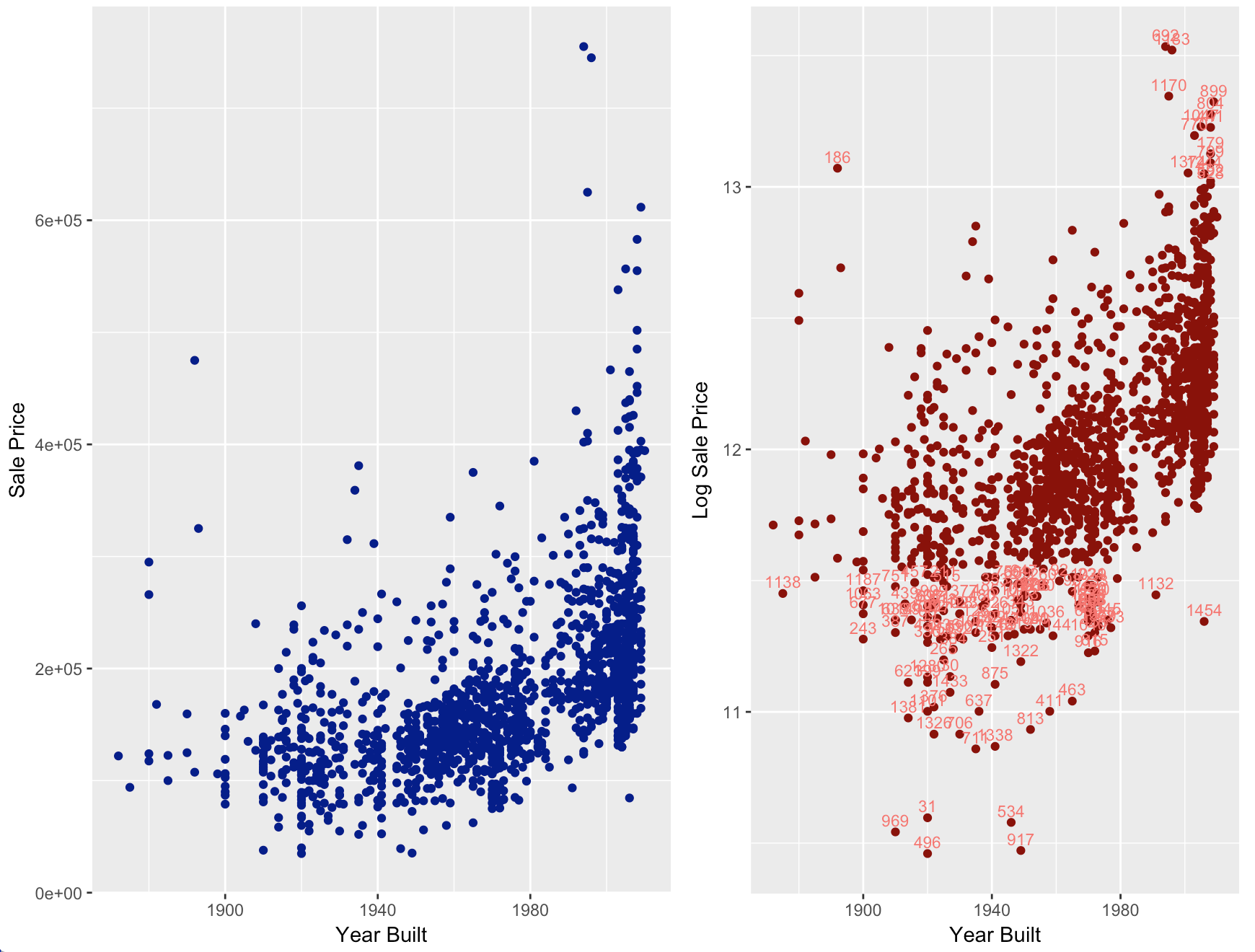
**Figure 2.7 Scatter plots of Sale Price Vs Above grade living area, Log Sale Price Vs Above grade living area, and Log Sale Price Vs Log Above grade living area**



**Figure 2.8 Scatter plots of Sale Price Vs Basement Square Ft, Log Sale Price Vs Basement Square Ft, and Log Sale Price Vs Log Basement Square Ft**



**Figure 2.9 Scatter plots of Sale Price Vs Garage Area and Log Sale Price Vs Garage Area**

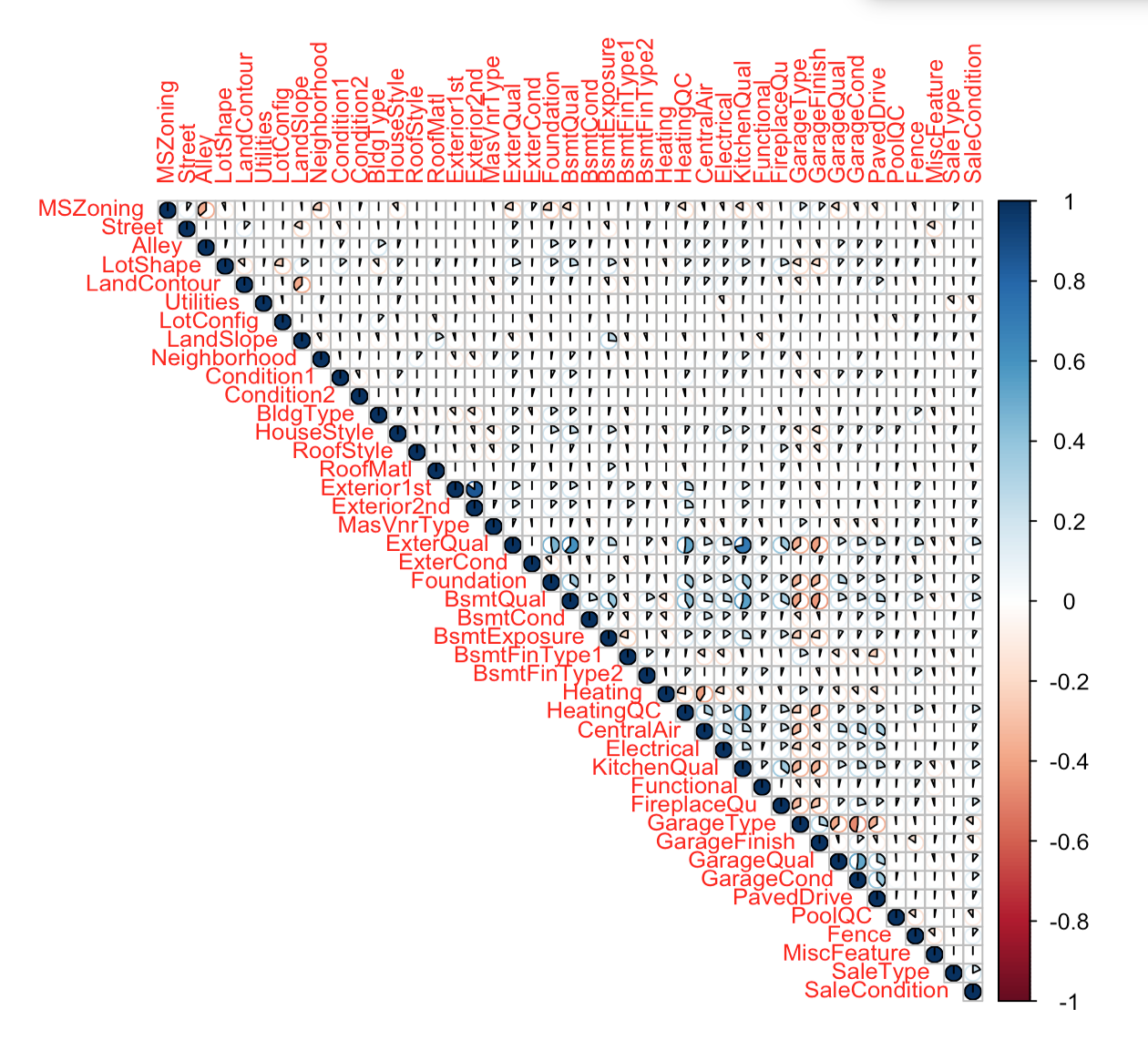


**Figure 2.10 Scatter plots of Sale Price Vs Year Built and Log Sale Price Vs Year Built**

## Ordinal Correlation

**Figure 2.11** represents a correlation plot showing correlations less than 0.3 among categorical variables in the data set. This matrix was plotted after converting all the categories to numeric values. The variables in the plot when used together will not lead to multicollinearity issues.

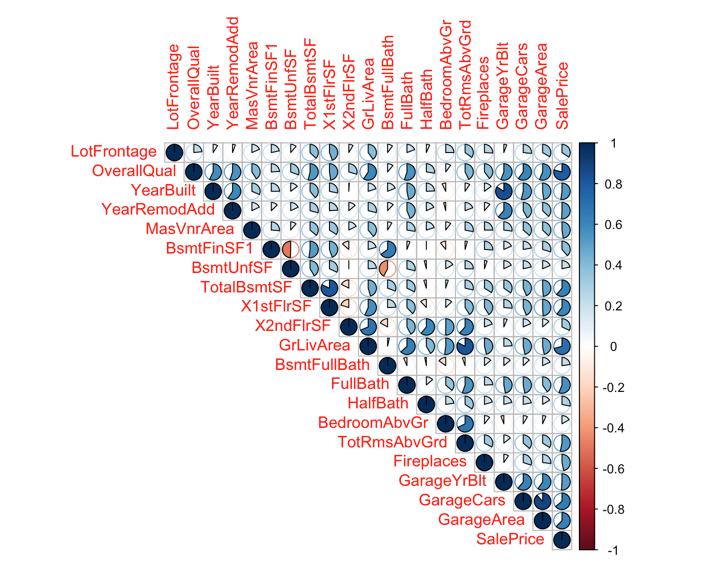
Some of the categorical variables have a natural order. These are called ordinal variables. After thorough research by going through discussion forms and reading articles in internet, the following ordinal variables were picked: ExterQual (Exterior Quality), PoolQC (Pool Quality), SaleCondition, Street, LotShape, LandContour, Utilities, LotConFigure (Lot Configuration), RoofMatl (Roof Material), Condition1.



**Figure 2.11 Correlation plot showing correlations less than 0.3 among categorical variables**

## Combining Variables

**Figure 2.12** indicates that OverallQual variable is correlated with several other variables in the dataset. Because of this, either OverallQual should be used or other variables should be used while performing data modelling. But by excluding variables there is a chance that the overall explanatory power many reduce. Hence certain variables such as YearBuilt, YearRemodAdd, logTotalBsmtSF, logGrLivArea, and FullBath were all combined with OverallQual and new variables were created. We combined variables by multiplying one variable with the other.



**Figure 2.12 Correlation plot showing correlations greater than 0.4 among numerical variable**

## Category Reduction

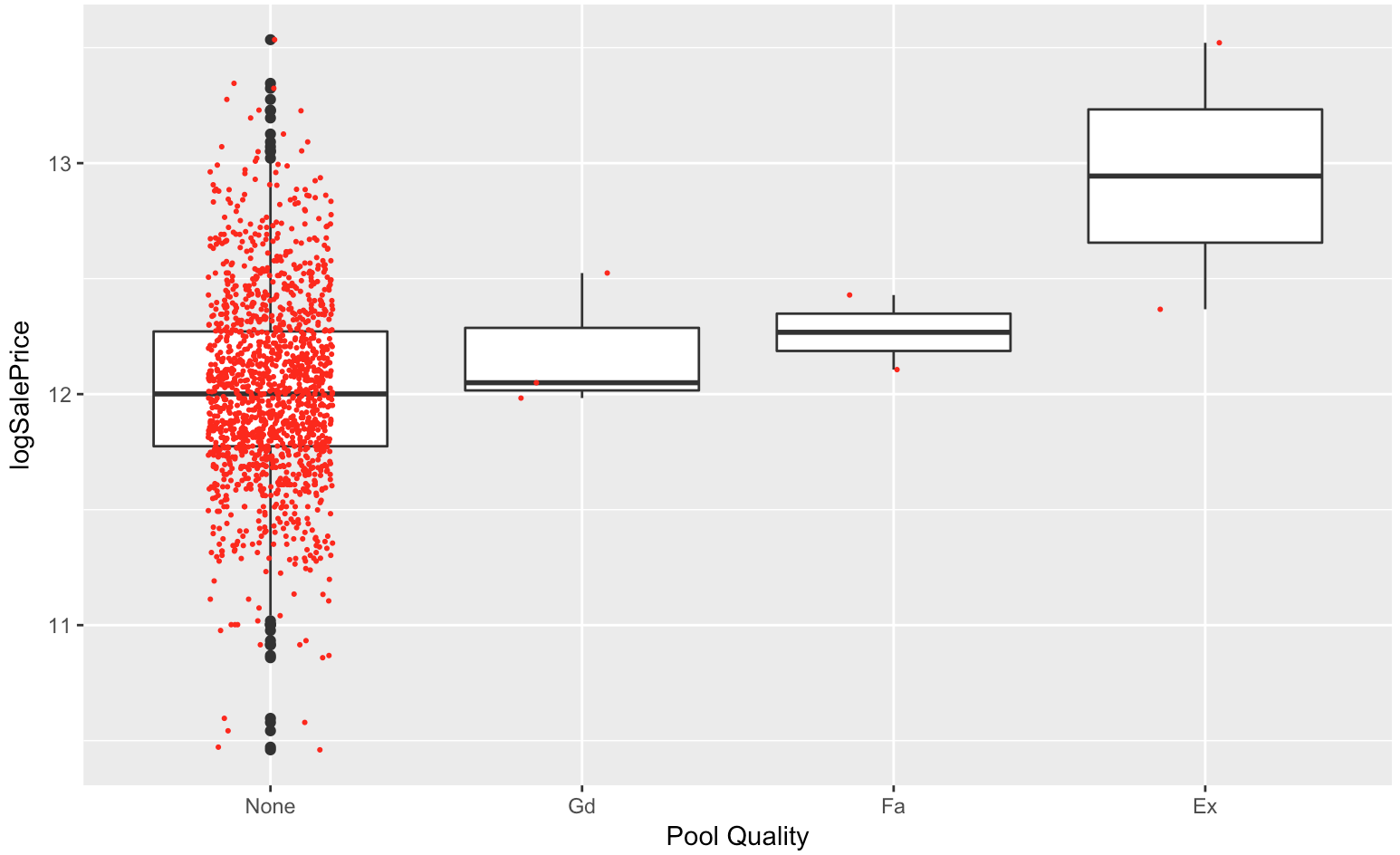
If a variable had more than 4 categories it was converted to a simple variable containing 2-4 categories. This was one keeping in mind that simple data is more comprehendible to human eye. By doing so we also identified good sources of our predictive probabilities. Prediction accuracy obtained during the data model corroborates this point. **Figures 2.13, 2.14** show variables whose categories were reduced and they were converted to binary variables. The same process was followed to update other variables such as Street, LotShape, LandContour, Functional, Utilities, LotConFigure (lot configuration), RoofMatl (roof material), and Condition1. All the old variables were deleted while data modelling.

**Old Variable**: PoolQC

**New Variable** : pool\_good 1 or 0

**Old Variable**: SaleCondition

**New Variable** : Sale\_cond Partial <- 4 Normal, Alloca<- 3 Family,Abnorml<- 2 AdjLand <- 1



**Figure 2.13 Box Jitter Plot of Pool Quality vs Log Sale Price**

A picture containing sky, wall

Description generated with high confidence

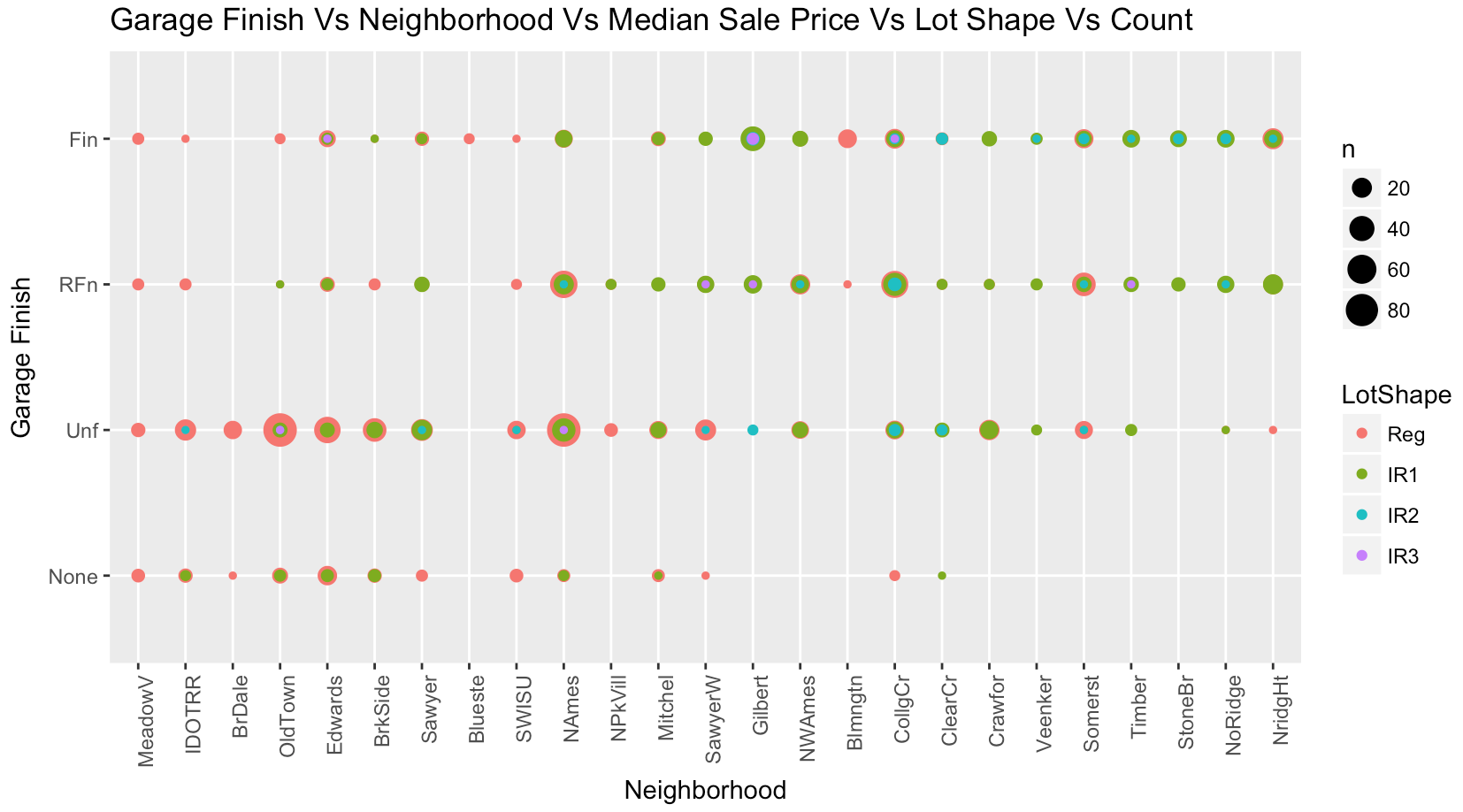
**Figure 2.14 Box Jitter Plot of Sale Condition vs Log Sale Price**

# Multivariate Analysis

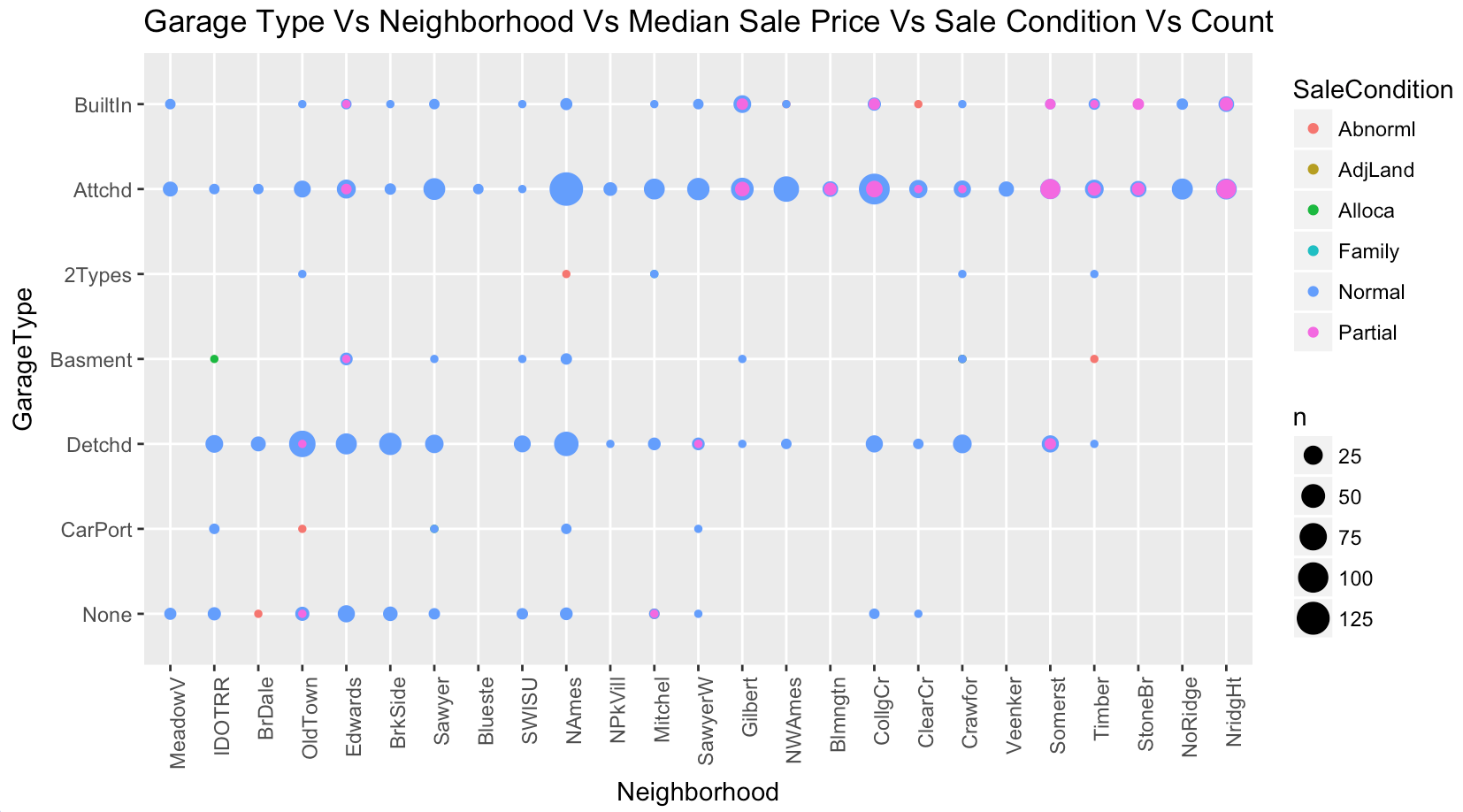
The **figures 2.15, 2.16, and 2.17** are multivariate representation of data in two dimensions. In **Figure 2.15**, the x axis is represented by Neighborhood and Y axis is represented by the Garage Finish variable. Both the X and Y axis are arranged in an ascending order based on median SalePrice. The color of the circles is represented by LotShape variable. By looking at figure 2.15 it can be said that Rich Neighborhoods have more IR2 (Moderately Irregular) lot shapes. Their garages are either finished or roughly finished.

In **Figure 2.16,** x axis is represented by Neighborhood and Y axis represented by Garage Type variable. Both X and Y axis are arranged in an ascending order based on median SalePrice. The color of the circles is represented by sale condition. By looking at figure 2.16 it can be said that Rich neighborhoods have more partial sale conditions. Rich neighborhoods have more built-in and attached garages.

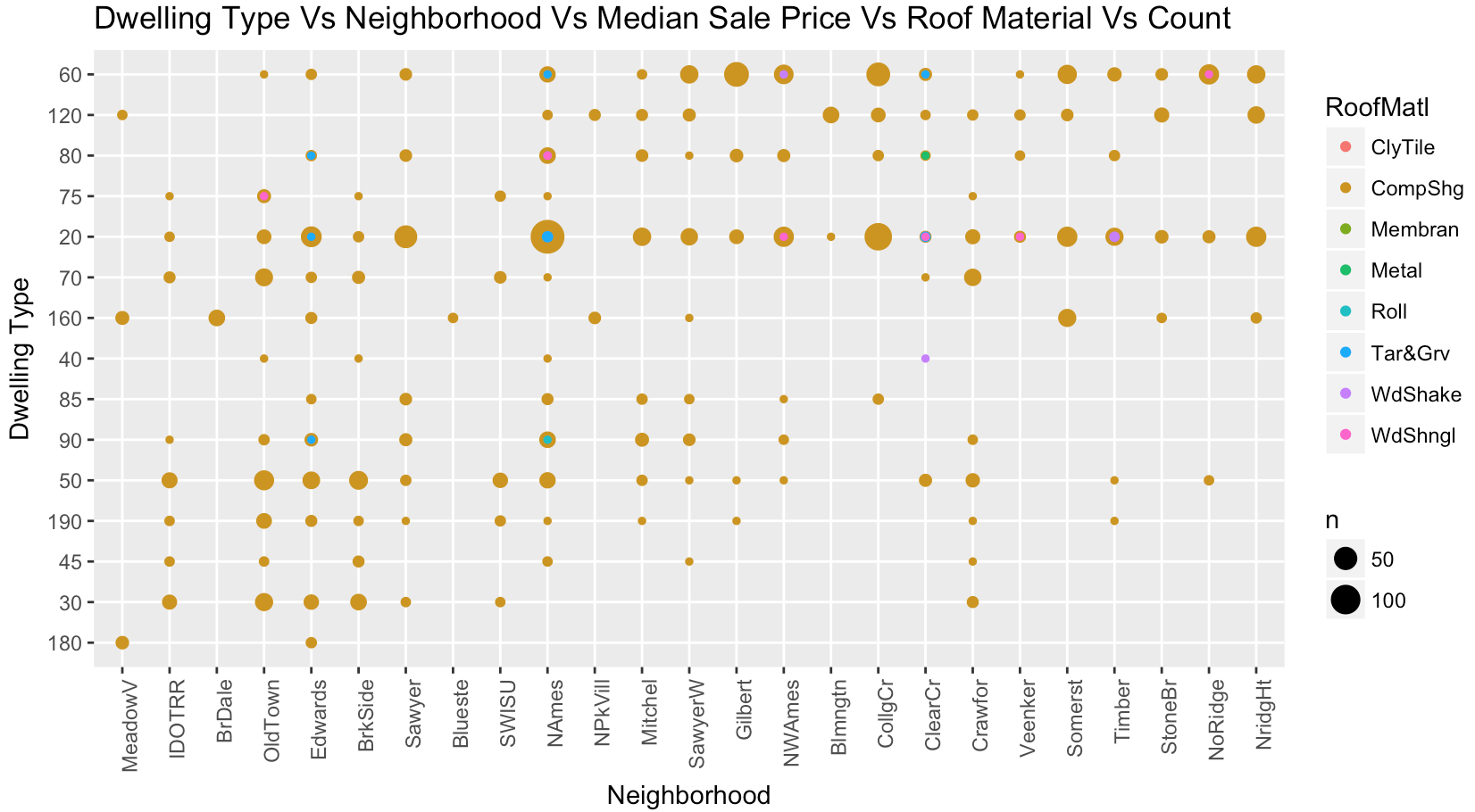
In **Figure 2.17**, x axis is represented by Neighborhood and Y axis represented by Dwelling Type variable. Both X and Y axis are arranged in an ascending order based on median SalePrice. The color of the circles is represented by roof material. By looking at **Figure 2.17** it can be said that Rich neighborhoods have more ClyTile (Clay Tile) and WdShake (Wood Shake) roof materials. Rich neighborhoods also have more 2 story homes.



**Figure 2.15 2D representation of a multivariate plot**



**Figure 2.16 2D representation of a multivariate plot**



**Figure 2.17 2D representation of a multivariate plot**

All figures in this section give a hint that a new variable indicating if a neighborhood is rich or not has to be created. Hence a new variable called RichNeighborhood was created. Figure 2.18 explains on what condition it was created. All the neighborhoods with median sale price on are after the vertical red line are considered rich and before the vertical line are considered not so rich neighborhoods.



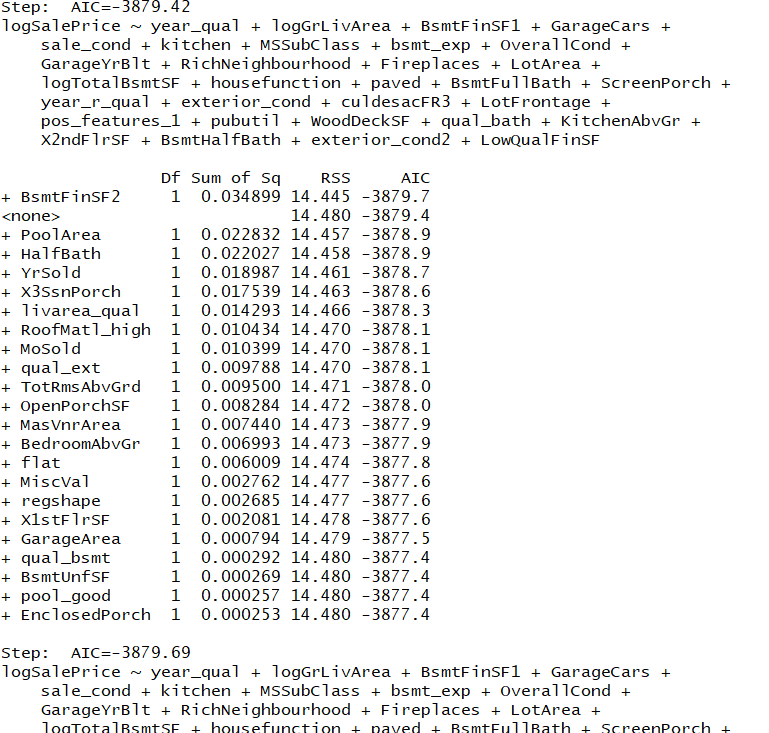
**Figure 2.18 Dot plot of Neighborhood vs Median Sale Price**

# Models

## Stepwise Selection Method

Once we got a reduced and enhanced data set (“train set”) after feature selection from the exploratory analysis, we decided to try stepwise regression to see if we can get a smaller subset with better AIC (lower value). We used the library “Mass” for running stepwise (with function ‘stepAIC’). The best model result from this exercise is shown in **Figure 3.1**. By using the best model from this, we couldn’t improve our final linear regression result as expected with the validation and test data. For example, after running linear regression with stepwise we got result: “Multiple R-squared: 0.9026, Adjusted R-squared: 0.8991” whereas with the ‘train set’, we got “Multiple R-squared: 0.9055, Adjusted R-squared: 0.8998”.

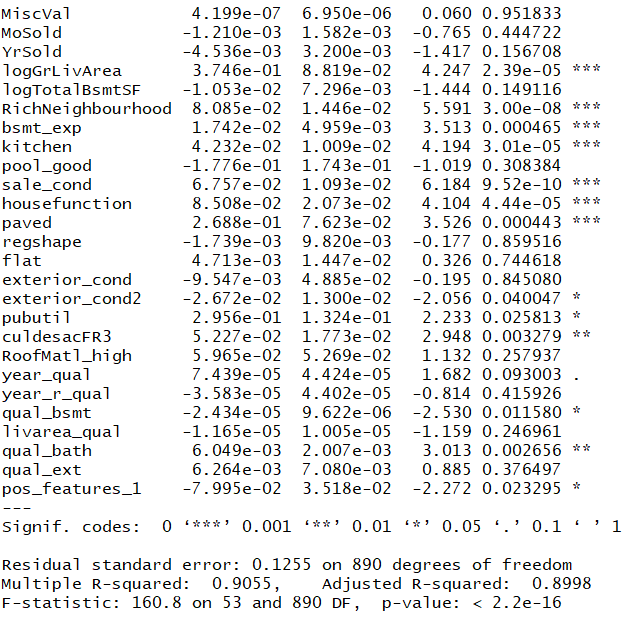
Hence, we decided to stick to the ‘train set’ for further analysis.



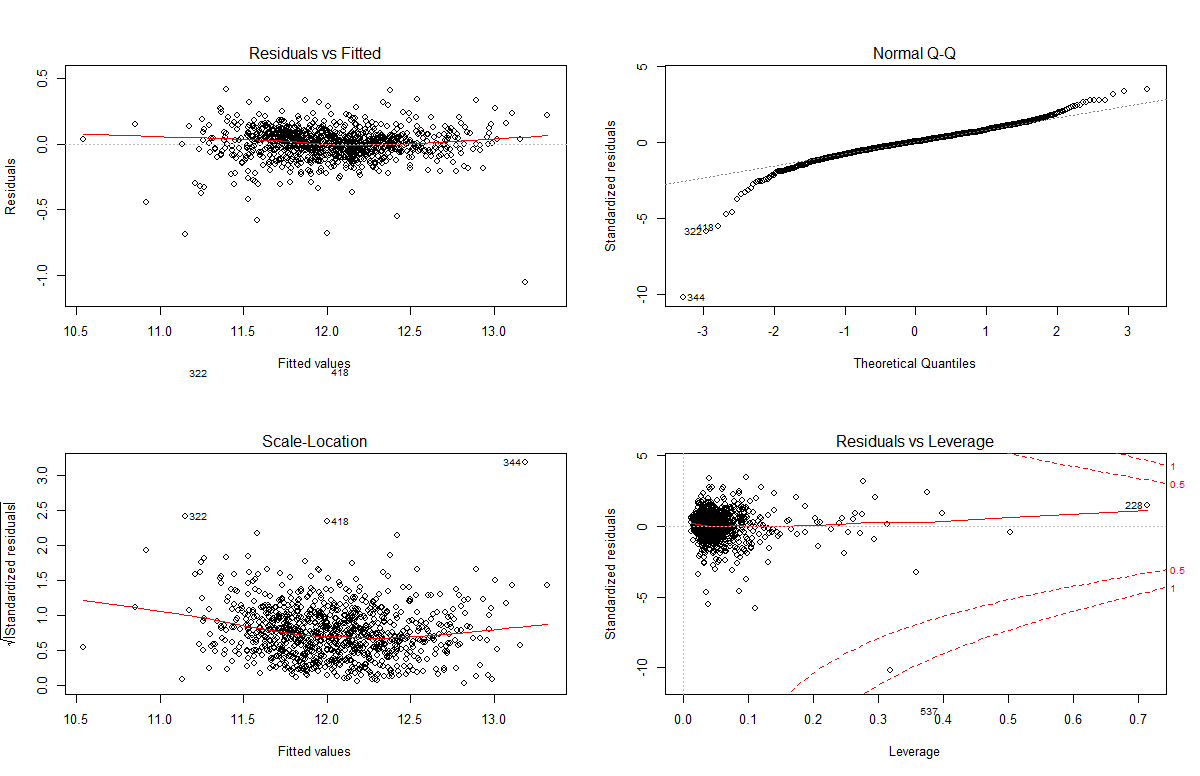
**Figure 3.1: Stepwise regression final result**

## Linear regression

With train set, we ran linear regression model in which we got the Adjusted R Square result of 0.8998. With linear regression, we got a RMSLE value of 0.1302 which was a good starting point. As we see from the coefficient value snapshot from **Figure 3.2,** some of the variables (like ‘MiscVal’,’MoSold’) has high (>0.01 p values). Hence, we considered them as not a significant variable for the final variable selection for prediction. **Figure 3.3** gives a sense of residuals and the validity of linear regression assumptions through residual plots. As we can see from residual vs fitted plot, it proves the homoscedasticity. Similarly, the q-q plot approves the normality of the residuals. We also don’t see any outliers outside the cook’s distance in the residual vs leverage graph, which is also a positive result. Though we see a curve pattern in the scale location plot between standardized residuals and fitted values, it seems to be minimalistic.



**Figure 3.2: Linear regression result**



**Figure 3.3: Linear regression residual plots**

## SVM

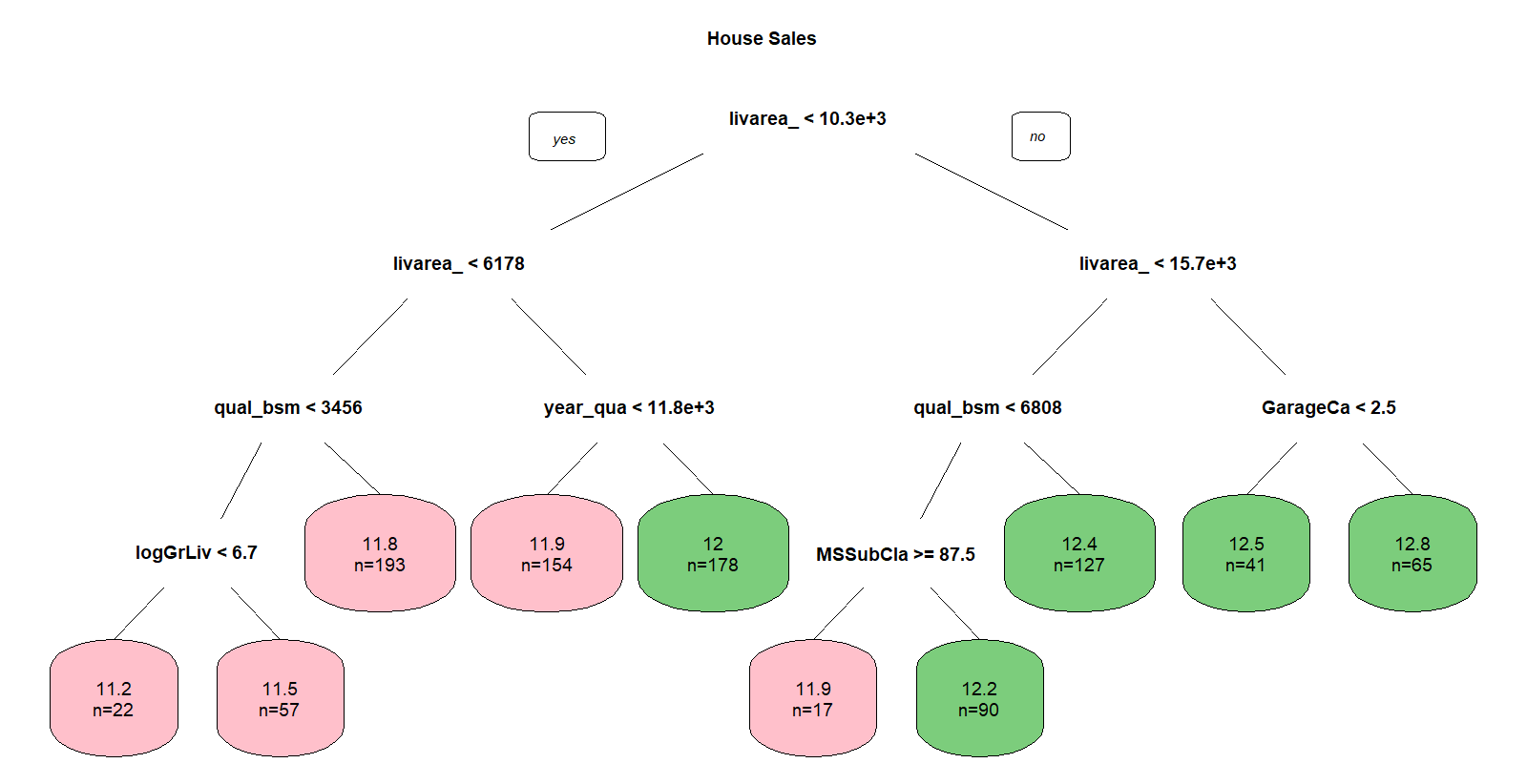
Since linear regression looked to be good, we wanted to verify if there is any kind of non-linearity associated with the data set. Hence, we ran linear, radial and polynomial kernel SVM models after tuning them for cost, gamma and degree parameters. **Table 3.1** shows the values of these best tuned models of “train data” for linear, radial and polynomial SVM. And we can clearly see that linear model is far better in terms of accuracy. Hence with this, we rule out the possibility of non-linear data set.

|  |  |  |  |
| --- | --- | --- | --- |
| **SVM MODEL TYPE** | **LINEAR** | **RADIAL** | **POLYNOMIAL** |
| **COST** | 0.001 | 5 | 0.1 |
| **GAMMA** | 0.01886792 | 0.5 | 0.02631579 |
| **SUPPORT VECTORS** | 611 | 866 | 754 |
| **DEGREES** |  |  | 3 |
| ***RMSLE*** | ***0.1408446*** | ***0.3932813*** | ***0.185484*** |

**Table 3.1: Table showing** **the best model parameters - cost, gamma, support vectors, degrees and RMSLE values for different SVM models**

## Decision Tree

Though we know that decision trees do not provide a reliable result when compared to random forest, we wanted to make sure we see the results of a single tree before attempting random forest. With different packages, “tree” and “rpart”, we tried running decision tree with different cost parameter values for a more accurate pruned tree. The result was the best decision tree model with 10 leaf nodes (see **Figure 3.4**) and RMSLE value of 0.22 (which is much lower than linear regression/SVM linear model). Though variables (like living area, quality of basement etc.) match the most significant ones from our previous analysis (linear and data exploratory), we are not considering the decision tree conditions for our predictions because of lower RMSLE values.



**Figure 2.4: Decision tree with 10 nodes after doing cost based tree pruning**

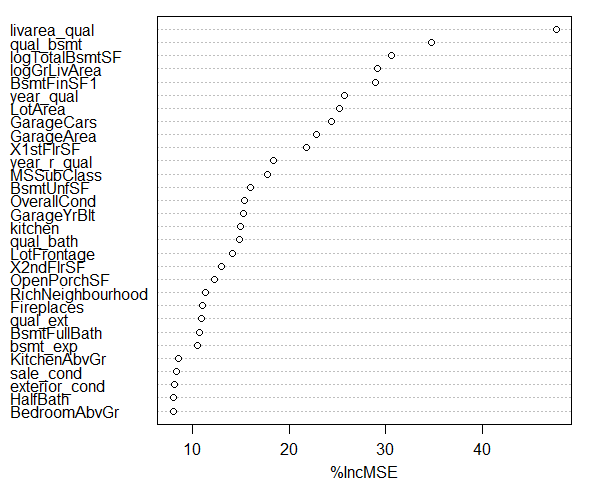
## Random Forest

With the thought that many variables will be correlated to each other and many interactions may need to be accounted for, we thought we would try out some Random Forest models. We tried both a Bagging model and a Boosted model.

### Bootstrap Sampling and Bagging

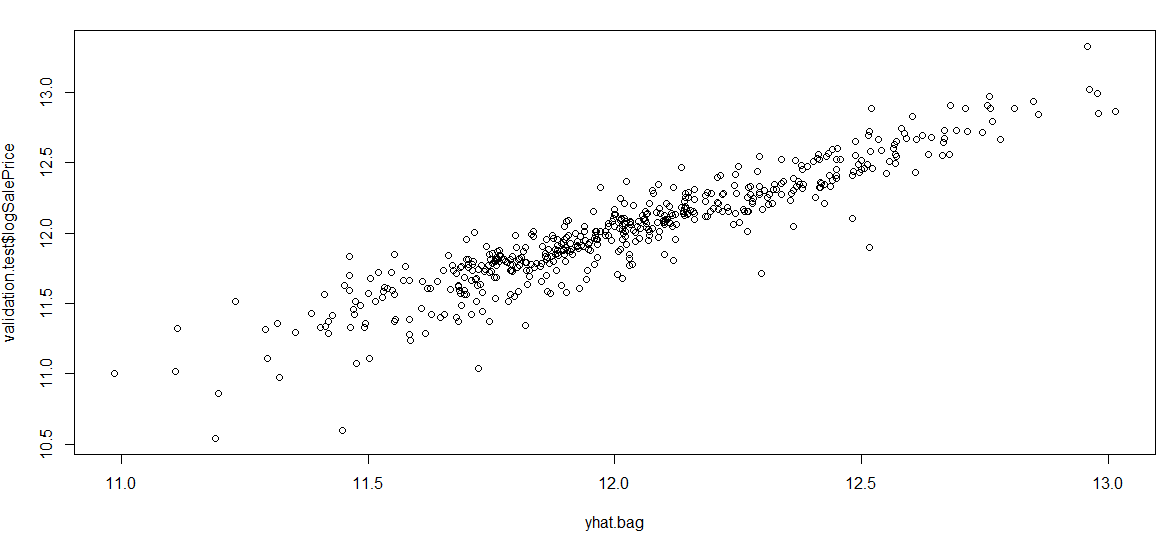
The initial model included all variables thus is a bagging model. The bagging model randomly selects subsets of observations to fit x number of trees to the data to account for a major pitfall in decision trees; high variance. We then tried running models with subsets of the variables as well as subsets of the observations.

The first thing we did was to run a base model. This was to give us an idea of relative importance of each variable. We then looked at the variable importance plot to determine which variables were most important to the model (**Figure 3.5**).



**Figure 3.5: Variable Importance Plot for Random Forest Model: Shows how much each variable contributes to the Mean Square Error. Livable area is the most important variables to predicting home value.**

After we had our base model we looked into tuning the model. In the Random Forest Package we can adjust “mtry”- the number of variables to try for each tree, and “ntree”- the number of trees in the forest. We can also choose to take some of the least important variables out of the model based on our variable importance plot. After tuning, the model we got was a Random Forest model with 1000 trees and mtry= 28 (28 variables sampled at a time). This model explained 87.09% of the variance and has a RMSLE of .1422. It layman's terms we were off by 14% on average (**Figure 3.6**).

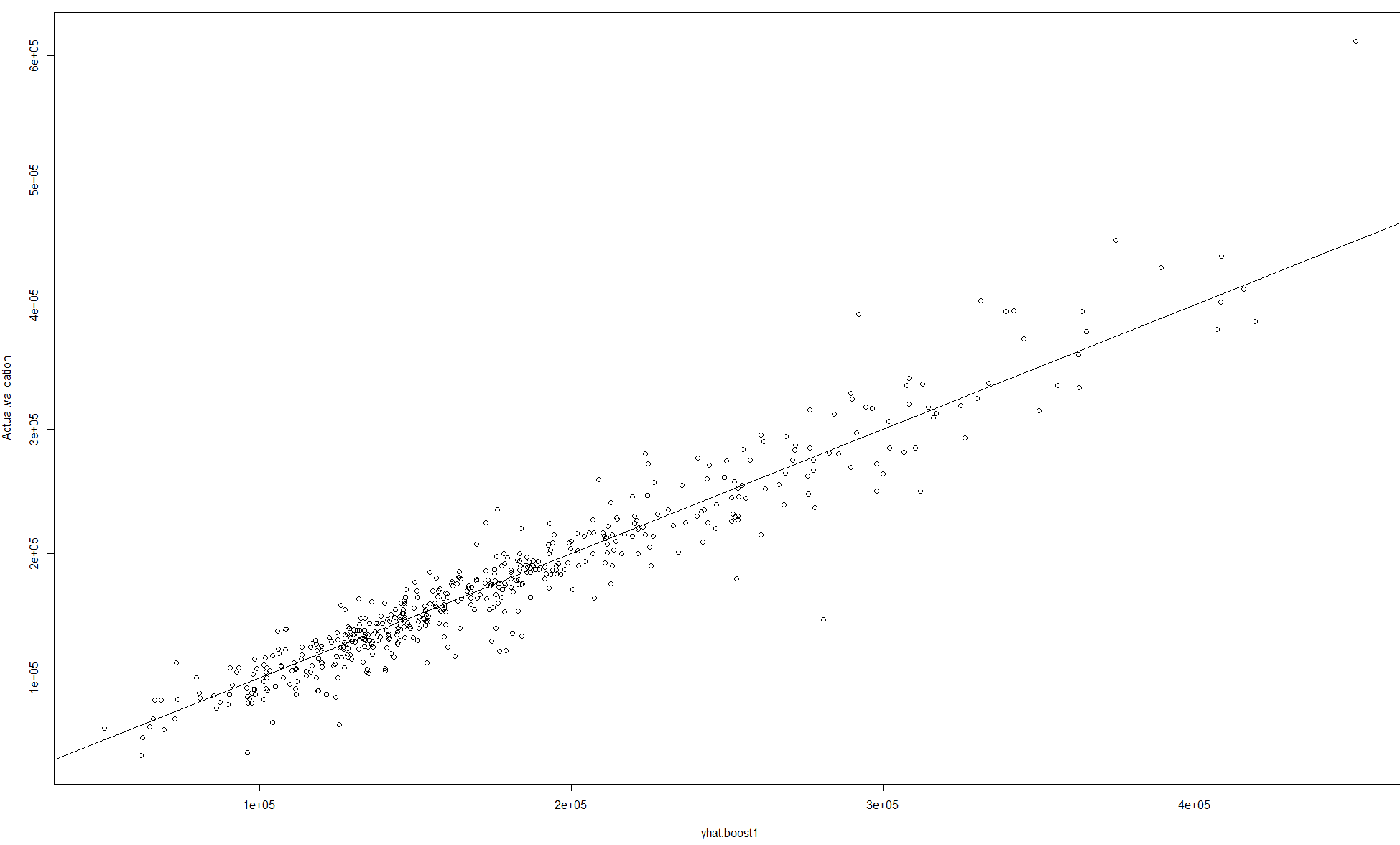
****

**Figure 3.6: Actual vs. Predicted Plot from best Bagging Random Forest Model: Values are in log SalePrice. RMSLE= .1422**

### Boosted Random Forest

We also tried a Boosting Random Forest Method. Boosting is a machine learning method which acts similarly to a bagging model. It samples and reuses data but learns from previous trees in the model. A penalty (lambda) is assigned to the model to correct and reduce residuals. We can choose the number of trees, the size of the penalty and the number of times the model should iterate. High penalties result in overfitting models so we tried to make sure that we tested small values of lambda.

We found we got better performance after removing the 7 lowest performing variables for the model. We then tested different values of lambda, interaction depth and number of trees. The optimal model here had 3000 trees, an interaction depth of 4, and a lambda value of .01. This model performed slightly better than the previous Random Forest model with a RMSLE of .1311 (**Figure 3.7**).

****

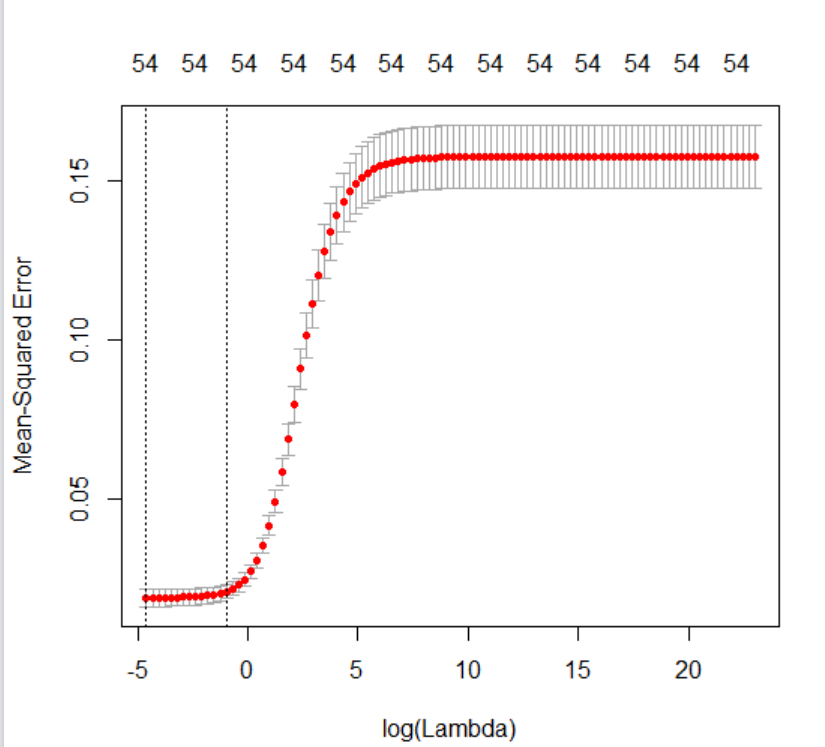
**Figure 3.7: Actual vs Predicted Plot from the best Boosted Random Forest Method: Values in logSalePrice. RMSLE=.131. The graph displays a slightly better fit than a Bagging Random Forest model, especially on the tails.**

## Penalized Models

We tried using two penalized models to predict sales, Ridge Regression and Lasso Regression. Thee goal of these two models is to assign some penalty to the model to control for the size of the coefficients. The model tries to move coefficients to, or towards 0. This is an effort to make a more robust model and reduce variance in the model. However this introduces bias into the model.

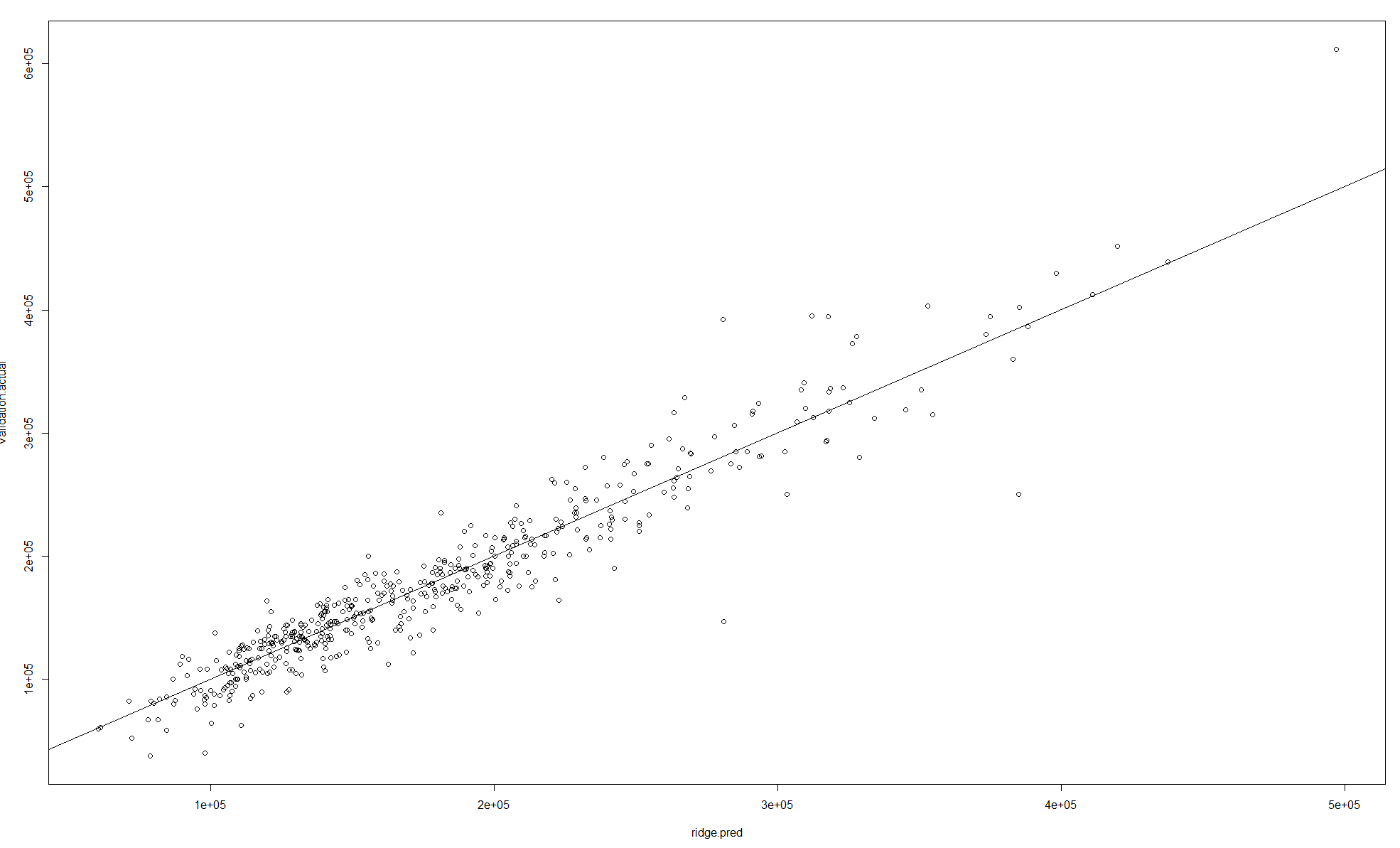
### Ridge Regression

Ridge Regression deals with correlated variables by assessing a penalty and shrinkage the coefficient of poor performing predictors close to zero. The first step is tune the model for optimal value of lambda (the penalty). We did this and then assessed how the penalty value performed by cross validating. The optimal Lambda value was .01 (**Figure 3.8**).

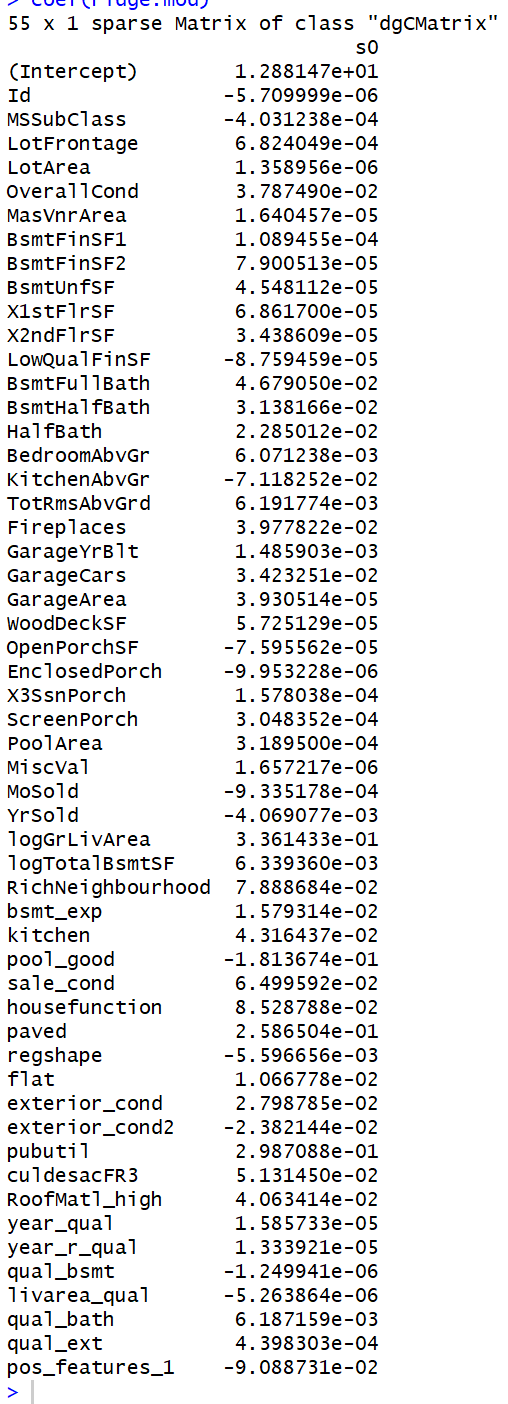
****

**Figure 3.8: Value of Lambda plotted against MSE of cross validation. We can see that smaller values of lambda are better. The optimal value we tested was .01.**

We tested a number of lambdas ranging from 10^10 to 10e-3. We found this model to be the best so far. It had an RSMLE of .1311 (**Figure 3.9** and **Figure 3.10**). Coefficients of variables are also provided.

****

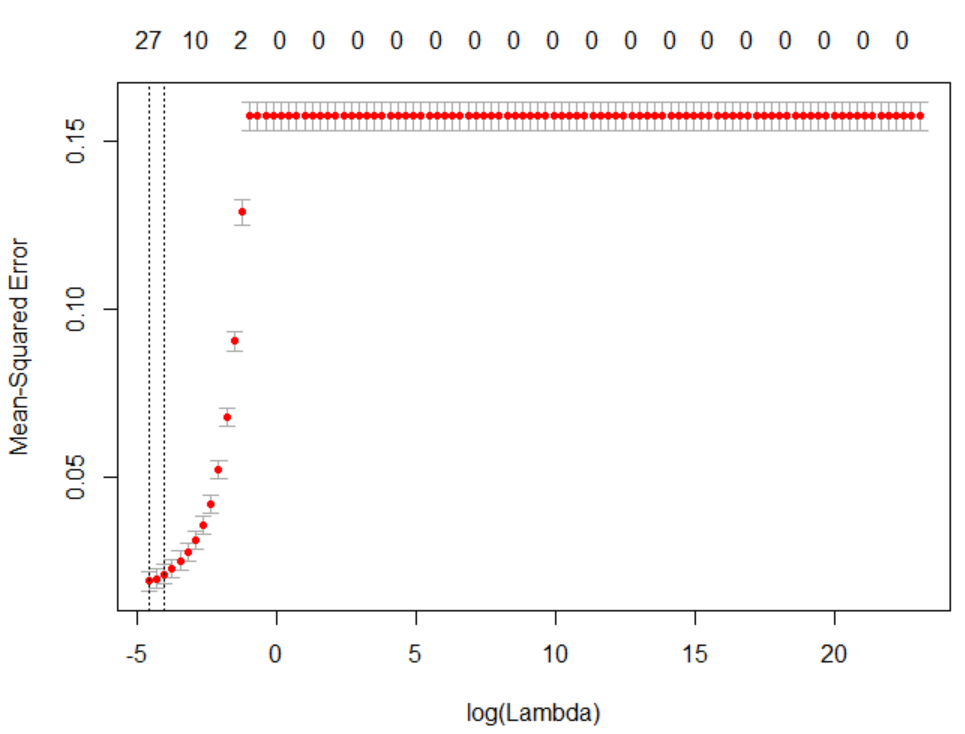
**Figure 3.9: Ridge Regression Model actual vs. Predicted Plot. RMSLE= .131, this has been the best model thus far.**

****

**Figure 3.10: Coefficients of Ridge Regression Model. It can be seen that no coefficient is at zero but there are variables that are very close.**

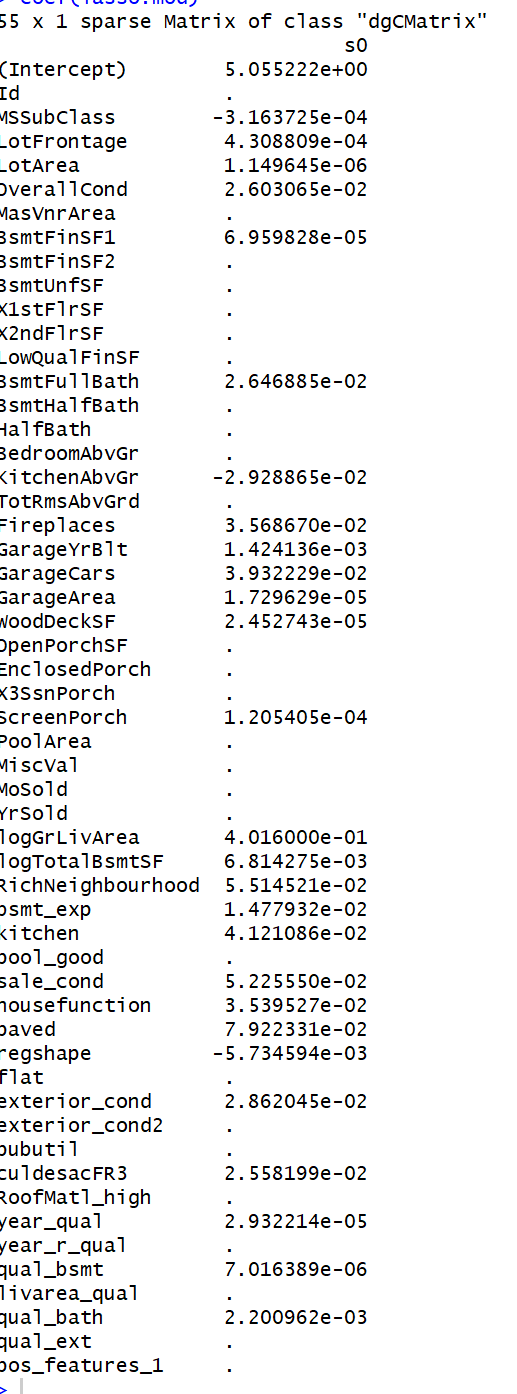
### Lasso Regression

Lasso works very similarly to Ridge regression except for the fact the coefficient of ineffective predictors is shrunken to zero whereas Ridge brings the coefficients towards zero. The optimal lambda turned out to be the same as in Ridge Regression, .01. You can see the plots look very similar (**Figure 3.11**).

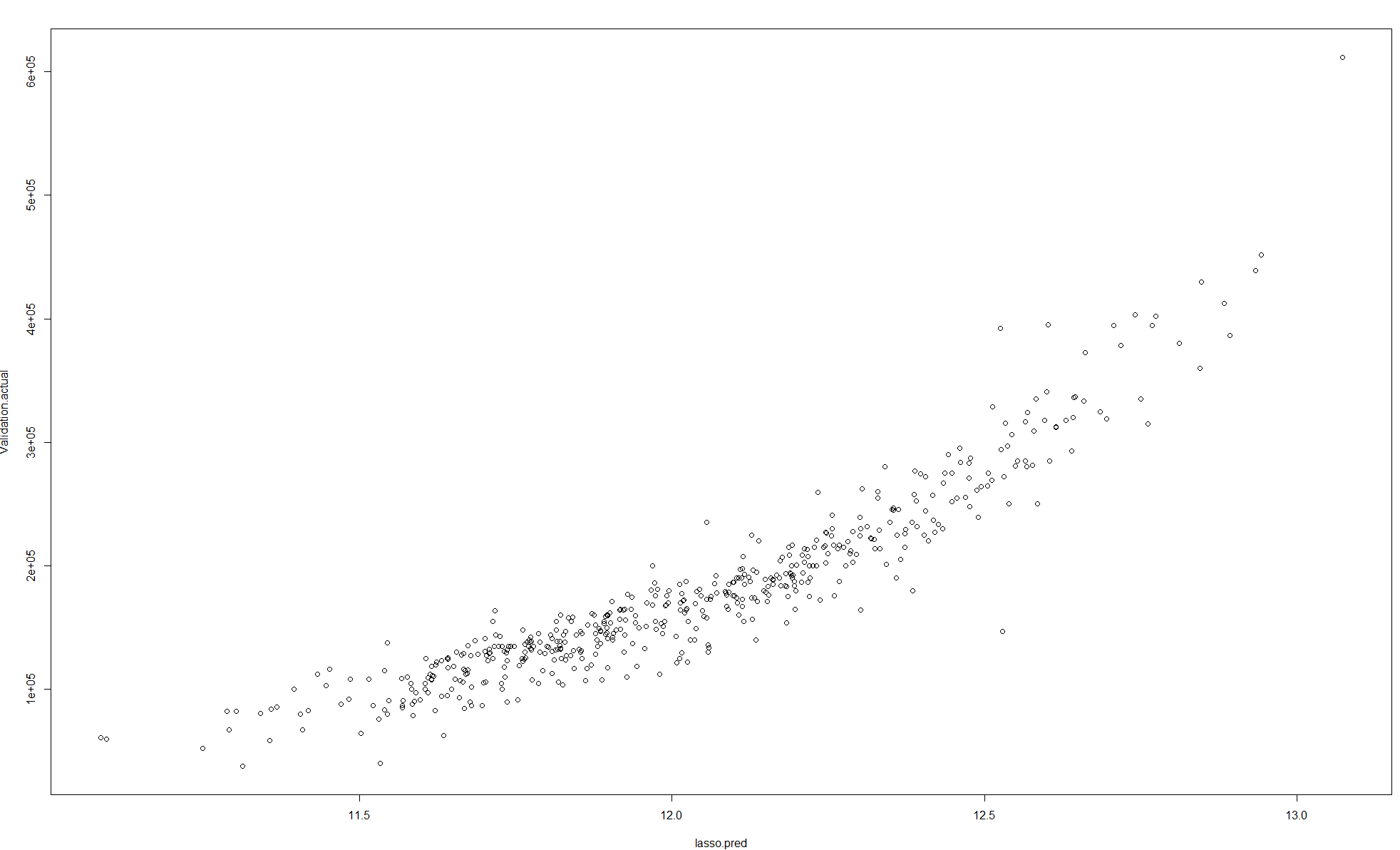
****

**Figure 3.11: Lasso Regression MSE plotted against lambda values. The optimal lambda value in at .01**

You can see the Lasso Regression brings about half of the predictor values to zero (**Figure 3.12** and **Figure 3.13**).

****

**Figure 3.12: Lasso Outputs: A list of coefficients for Lasso regression (notice many variables brought to zero).**

****

**Figure 3.13: Actual vs Predicted Plot for the validation set. Also notice the curved shape on the fit curve.**

This model performed slightly worse than the Ridge model but with half as many predictors. The RMSLE was .136. The Actual vs Predicted plot is U-shaped, however, indicating we may be missing variables that would be good predictors.

### Model Summary

After creating tuned models for each of these methods, summarized in **Table 3.2**, we found that our most accurate model was a combined penalized model. The results of this model were found by averaging the results of the penalized models, Ridge and Lasso. Though this model had our lowest RMSL error, it is also one of least interpretable models. Depending on the intended use of the model, discussed further in the conclusion section, it is worth noting that for a relatively small increase in RMSLE, a large increase in interpretability can be gained by using a linear or decision tree model. For the Kaggle case, the only factor that was considered for choosing the “best” model was RMSLE, so that is where we focused our efforts. In practice, however, a highly interpretable model that is fairly accurate may be desirable.

|  |  |
| --- | --- |
| **MODEL** | **ACCURACY PREDICTOR- RMSLE** |
| LINEAR | 0.1302 |
| RIDGE | 0.131 |
| LASSO | 0.136 |
| COMBINED PENALIZED (AVG RIDGE, LASSO) | 0.129 |
| DECISION TREE | 0.2196 |
| SVM (LINEAR) | 0.1408 |
| SVM (RADIAL) | 0.3932 |
| SVM (POLYNOMIAL) | 0.1855 |
| RANDOM FOREST – BAGGING | 0.1422 |
| RANDOM FOREST – BOOSTING | 0.1311 |

**Table 3.2: Summary of model accuracy**

# Conclusions

From analysis and creation of prediction models, several different ways were though to apply the insights in a real-world setting.

## Flipping Houses

If someone uses our best model to predict the price of a home that is currently for sale, they may find an opportunity where the home is listed for a price lower than predicted. If that is the case, there is the opportunity of “flipping” the house. That is, buying a home for the sole purpose of quickly selling it for a higher price. By using our model, investors could go into negotiations with a clear idea of how much they should be willing to bid on the home. The prediction will also give them an idea of how much profit they can expect from each home, assuming they can negotiate the sale price well.

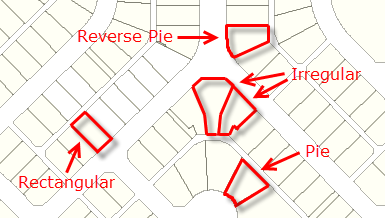
## What should the target list price be on homes

First, the prediction model can be used by banks and government agencies to set the price on foreclosed or seized homes. In these cases, the selling entity may not be too concerned about getting the best price for the home, but are instead concerned with quickly selling the home. By using the model to predict the home price, they can minimize the cost they put into evaluating the property.

The second scenario where the home value prediction may be used is with realtors and tax assessor-collectors that need a baseline valuation of a home before further analysis. The realtor or assessor-collector may want to quickly establish a baseline value of a home before a thorough analysis or inspection of the property. By using the model, either group could quickly establish a home value for initial taxation or listing purposes.

## Renovations

A quick google search yielded 1,000s of articles for the best things you can do for a home before selling it. Part of our analysis was to determine what variables, that is, what aspects of a home, are most influential on the sale price. Though some of the most important variables were things that can’t easily be changed by the seller (neighborhood, year built), there are some important variables that can be manipulated by the seller. For instance, the quality of the kitchen, bathrooms, and home finish are all positively correlated with sales price. The lowest correlation with sales price amongst these three variables is kitchen, with a correlation of .6. Some variables are negatively correlated with sales price, meaning things the homeowner should avoid, if possible. Both irregular lots and enclosed porches were shown to have negative correlations with sales price. To understand why lot shape is negatively correlated with sale price, we performed some research into the official Multiple Listing Service (MLS) lot definitions. **Figure 4.1** shows what some of the general MLS lot shapes are:



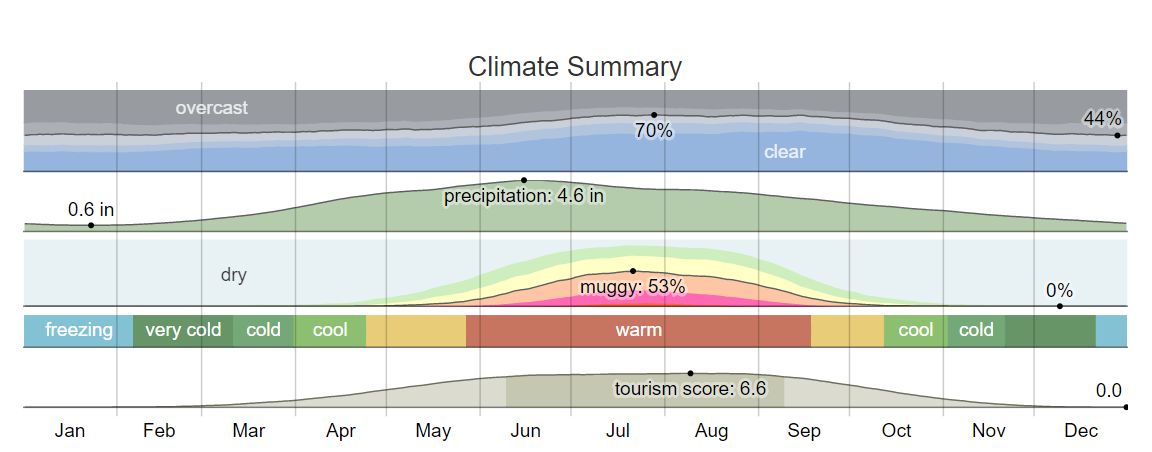
**Figure 4.1:** **MLS Lot Shape Examples[[1]](#footnote-1)**

In our data set, the researchers collecting the data only classified lots as regular or irregular. Irregular lot shapes are more difficult to sell since they are costlier to landscape or make any additions to. For instance, if someone wanted to add a shed to their property, it is more difficult to design and build within an irregular lot than it is with a regular lot shape. As such, as the degree of lot irregularity goes up, the sales price goes down. Our findings support this conclusion, as can be seen in **Figure 2.11,** the correlation plot of categorical variables.

To understand why enclosed porches were negatively correlated with sales price, we again had to conduct research. We came up with two possible explanations of why our data shows a negative correlation between sales price and square footage of homes:

1. Enclosed porches have gone out of style
2. Ames, Iowa, isn’t the most appropriate place for an enclosed porch

According to this article from bankrate.com[[2]](#footnote-2), young homebuyers are interested in homes that are energy efficient and maximize open floor plans. An enclosed porch is detrimental to both of those items. An enclosed porch is essentially taking away the total livable square footage of the home while adding a room that is expensive, or impossible, to keep heated or cooled.

The second reason we could think of why the inclusion of an enclosed porch would reduce sales price is that Ames, Iowa, does not have the best climate for porches. According to the average temperature data of Ames, shown in **Figure 4.2,** Ames is only considered “warm” for four months out of the year. 

**Figure 4.2: Climate Summary of Ames, Iowa, since 1980[[3]](#footnote-3)**

This means that an enclosed porch would be relatively cold, compared to the rest of the house, for the majority of the year. This may be seen as a waste of square footage and energy by homebuyers.

1. http://www.gimme-shelter.com/lot-shape-50068/ [↑](#footnote-ref-1)
2. http://www.bankrate.com/finance/real-estate/must-haves-to-sell-to-young-homebuyers-1.aspx [↑](#footnote-ref-2)
3. https://weatherspark.com/y/10339/Average-Weather-in-Ames-Iowa-United-States-Year-Round [↑](#footnote-ref-3)