STAT R 503 Final Project

Location, Location, Location:

Predicting Residential Housing Prices

Thomas Wright

Introduction

Commercial real estate appraisal is simple – anyone can appraise a building in an afternoon. You identify the "Highest and Best Usage" of the lot and base the value off of that theoretical building. As such zoning and location are the two highest indicators of value – what can you do with the lot and where is it? Standing structures are often tallied as a liability at demolition cost, rather than the positive value that a residential home has.

Residential real estate appraisal is tough. The multitude of fixtures, finishes and building characteristics force appraisers to rely on their deep domain knowledge to identity features of value.

Previously, this steep learning curve manifested itself in a much longer apprenticeship period than its commercial counterpart. Consumers where previously left relying on real estate agents and using rough aggregate of similar house sales.

Big data and machine learning methods have now unlocked much faster and less variable estimates of housing prices, with Zillow using its algorithms to produce estimates of housing prices, funnel consumers to real estate agents and near $6B market cap.

This study seeks to create two models to predict housing price based on 79 characteristics gathered on residential housing prices from Ames Iowa, using the lasso and random forest methodology.

In addition to the main predictive focus, the study will also examine which variables have a large impact on Housing Sale Price to begin to explore which individual features drive final value the most.

The study expects to generate two models with highly significant variables associated with neighborhood, zoning and Age of the house – with Age possibly having a quadratic effect. Additionally, those sales to family members (or other non arms-length transactions) should also have a significantly lower sale value and generate large residuals – due to the property likely being sold at below market rate.

Methods

Materials:

The initial dataset was provided on Kaggle.com, split into training and test CSV files. The Ames Housing dataset was compiled by Dean De Cock for use in data science education as an alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.[1]

The data describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing home values[2].

As the test dataset did not include the response variable, it was discarded. The remaining data set has dimensions of 1460 and 81 – meaning there are 79 recorded variables (81 – response – ID) for 1460 houses.

Data Cleaning

The dataset contained several factor variables with an NA response coded if the feature was no present – example: the variable corresponding to type of garage had an NA if there was no garage. To overcome this, the study coded in new factor levels for NX with X being the first initial of whatever the factor variable was (ex: NG). This would also be a useful step for an inferential study, as it would allow comparing different types of features against the base level of not having said feature.

Data Manipulation

Several variables were classed as integers when they are better modeled as a categorical predictor as they either do not have discrete intervals or the ordinal ranking does not pertain to the ranking of the variables levels.

- MSSubClass is a numerical score standing in for the type of building class and shouldn't be represented by an ordinal scale.
- OverallQual and OverallCond are subject measures of quality without discrete intervals.
- MoSold represents the month sold and is best suited with factor levels
- YrSold represents the year sold and could be modeled as a factor or a numerical vector. The study chose factor to see if any of the particular years had a significant individual impact on sale price.

---

[1] https://www.kaggle.com/c/house-prices-advanced-regression-techniques
[2] https://ww2.amstat.org/publications/jse/v19n3/decock.pdf

Several other variables were linear combinations of other covariates, leading to a very high level of collinearity in the sample. As a result, new variables were created and the linear combinations were removed:

- The study combined the amount of basement bathrooms into one variable (full + half) as well as the total bathrooms in the rest of the house.
- General Living Area is a linear combination of 1st floor SF and 2nd floor SF. The study removed the 1st and 2nd floor measures, adding in a dummy variable for whether the house has a second floor or not.
- Total Rooms was a linear combination of bedrooms and kitchens. The study created a new variable for total rooms (to account for the cases where Total Rooms> bedrooms + kitchens) and deleted total rooms. This allows a more granular prediction based on the layout of a house, rather than just the total number of rooms.
- Total Finished Basement Area is a linear combination of the finished SF in the 1st and 2nd feature. As there is already a dummy variables for whether and what each of those features area, the study delete their variables measuring SF.
- The Utilities Variables also caused the errors to occur in the study – leading to its exclusion.

The study has created new variables for the age of the house and the time since remodel, deleting the year sold and year remodeled variables. While there are vintage decade for house buying, there are not vintage years like wine. An age variable can capture the potentially quadratic nature of the year sold – with newer houses and those built before the 1950s being worth more. Additionally, the year remodeled doesn't seem to be a key value driver – there are no vintage or collectors years for remodels. However, the amount of time since it has been remodeled could offset the age of the house.

Transformations:

When looking at the pairs plot for the numeric only data, several variables appear to be irregularly shaped or skewed.

Both the Sale Price and Lot Area (size of the property) are all greater than zero, have a large variance in value, and a lot of values clustered together with other spread further out – prime candidates for long transformation. Indeed, both of their qqplots improve after transformation.

The rest of the variables in question are related to the SF of a feature. If a house has a deck, it is 1000 SF (for example). However, all the houses that don't take decks have 0 as their SF value. When preforming a log transformation (with 0.01 added to make the zero values transformable), all of the qqplots worsened. As such, the study left those variables alone.

Final Data Size

After the rows still containing missing values were deleted, the study had a final dataset with 955 observations and 73 explanatory variables.

Splitting the Data:

The dataset was then split following the 80/20 rule with random sampling – with 20% of the data randomly split into the test dataset and 80% remaining in the training dataset.

Methodology:

Data Visualization:

After cleaning and manipulating the data, the study proceeded with a preliminary visual examination. The shapes of all of the covariate distribution was examined with pairs.plot. However, due to the large nature of the dataset, a correlation plot had to be used to visualize all of the covariate correlations.

Using the shapes of the distribution as a guide for transformations and the correlation plot as a guide for data manipulation, the study generated the final dataset. From there, the same pairs plot and correlation plots were run – with the correlation plot serving as an early indicator of which covariates might dominate the results models.

Lasso:

For its preliminary analysis, the study employed the Lasso shrinkage method. The model uses least-squares criterion to shrink the Betas on non-significant covariates to zero – performing model selection and fitting.

The method accepts two primary tuning parameters – alpha and lambda. Alpha is one by definition and lambda was tuned by the study, with a large range of lambda being tested for to find the one that produced the model that minimized the Mean Squared Error (MSE). A 10-fold Cross-Validation (CV) was used in fitting the model (as a resampling method).

Once the "best lambda" was identified, the corresponding model was compared to the "next best fit" model – the model whose lambda lies within 1 Standard Error (SE) of the "best lambda".

The lambda.1se model was much more parsimonious than the lambda.min model (269 covariates against 18). As such, the lambda.1se model was selected and presented and in the results section.

The winning model was used to predict on the test dataset, with MSE being presented.

Random Forest:

Random Forests are a machine learning technique that can be used either as a classifer or as a regressor. The method uses sampling with replacement to create a large number of regression trees to average across.

The study first performed tuning on the ntree argument, how big this forest resampling size should be. A vector of possible sample sizes was created, ranging from 200 – 15000. However, both the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) dropped off sharply – suggesting that fewer forests were needed. The study then repeated the study with a smaller range of ntree values, before settling on a conservative value of 100 for ntree.

Next, the study performed 2D tuning on two other parameters – node size and the number of parameters that should be included in the model – using parallel processing and ntree set to 100. Both the models calculated with RMSE and MAE selected a node size of 4 and 20 parameters.

The model with the winning parameter was used to predict on the test dataset, with MSE being presented.

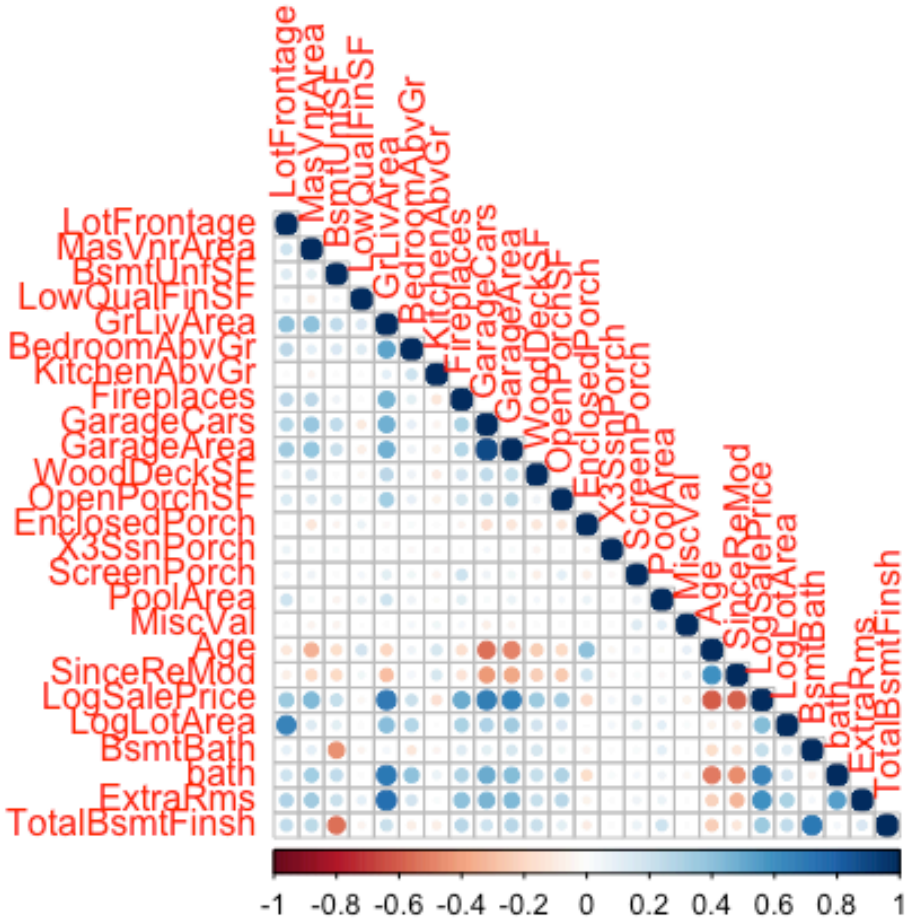
Results

Data Visualization:

After manipulating the data to account for variables that were linearly correlated, the study examined the correlation plot of all of the numeric variables (covariates must be numeric to calculate correlation).

From this plot it appears that LogSalePrice is correlated with:

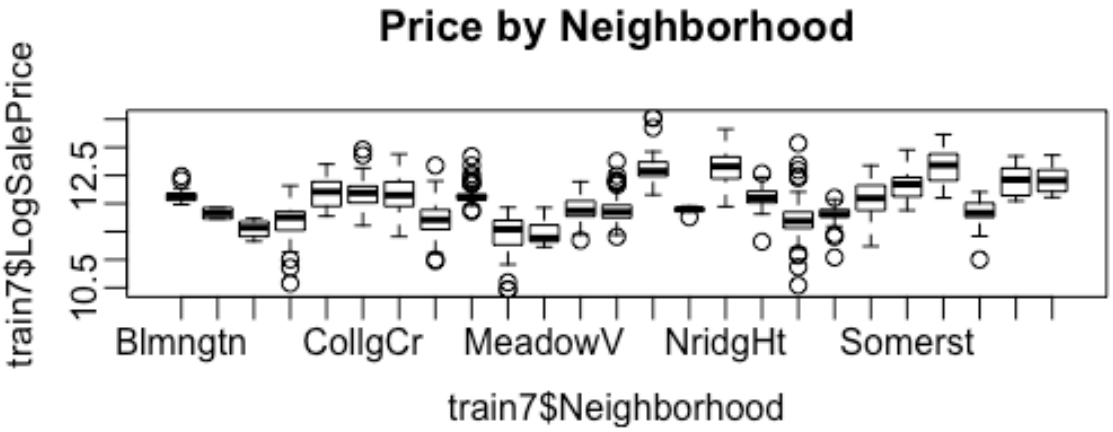
- General Living Area – positively
- Fireplaces – positively
- Area of the Garage - positively
- # of Car Garage – positively
- Age of House – Negatively
- Time since remodel – Negatively



With Age, Garage Area and General Living Area having the largest correlation (not necessarily the largest impact).

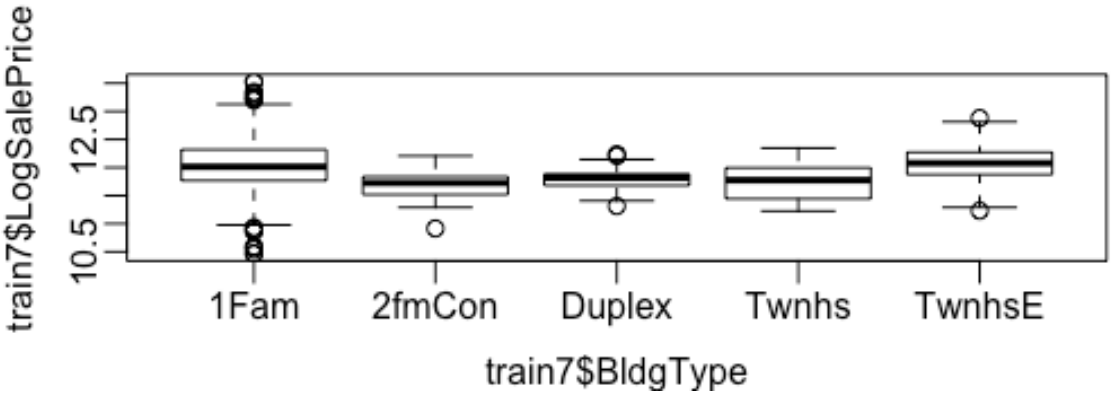
Turning to the categorical variables – the study has previously identified neighborhood, zoning, type of sale, and building type as key value predictors.

Looking at the Price by Neighborhood, there is clearly large variation between the prices of neighborhoods – backing the location, location, location premise. However, there appears to be a rough mean with only a few neighborhoods significantly above it and none beneath it – suggesting that both this is a sample of homogenous neighborhoods and also that also a few of the variables levels will survive as predictors (if most aren't significantly different than the mean).



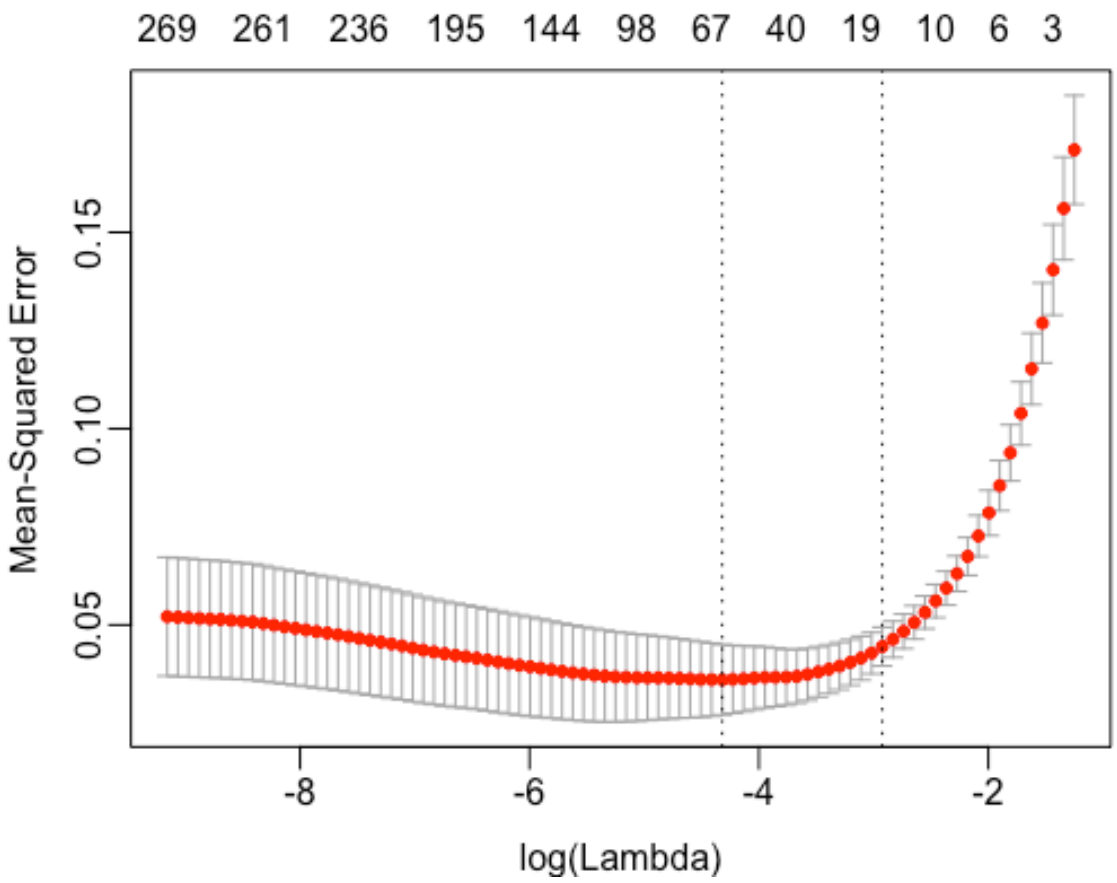
Turning to Sale condition, it does not appear that any type of sale is significantly different than the others. While Normal has quite a bit more outliers, it also has a much larger volume. This plot dis-spells one of the studies hypothesis – that family sales would be significantly different than normal ones.

Looking at the building type, there is a lot more variability for a Single family home – which makes sense as they probably have the highest variance in features. Townhomes are also on the rise – suggesting the trendy building craze has reached Ames, Iowa.



Lasso:

The study generated a lamba tuning curve that suggested 20 to 50 variables be selected:



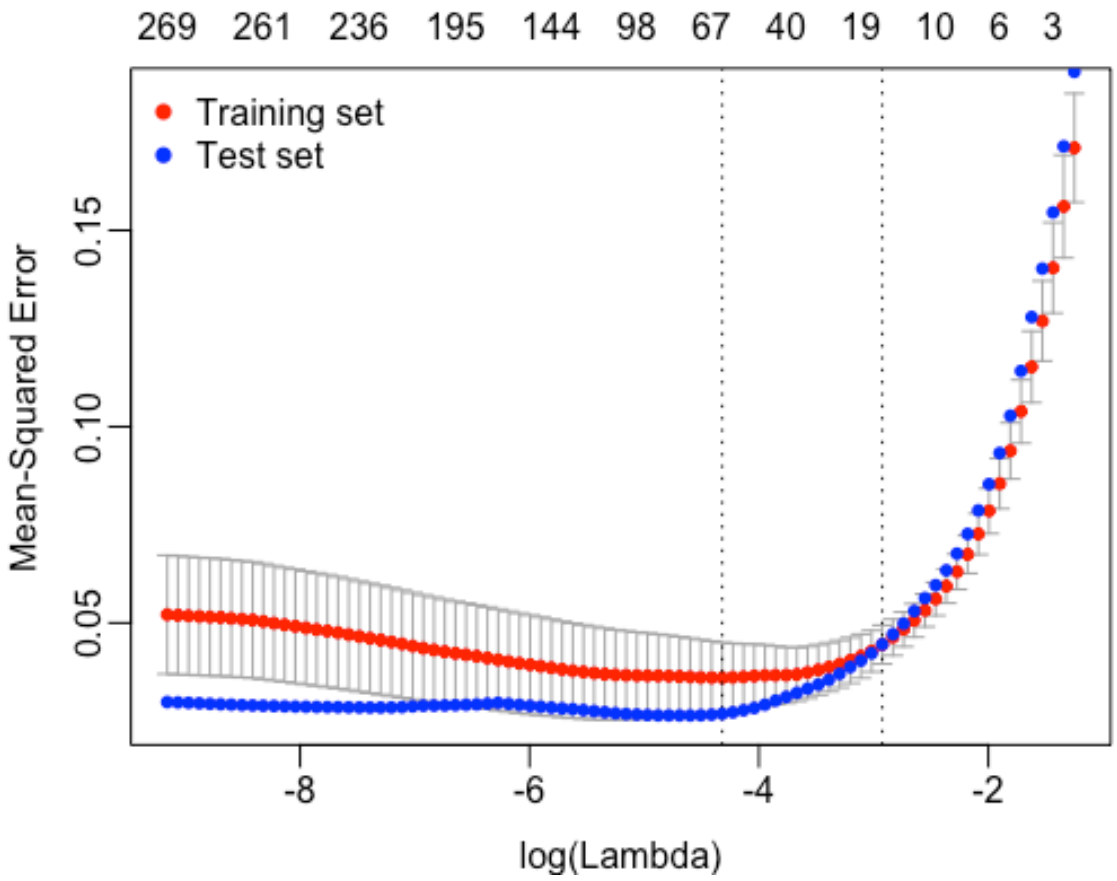
The study selected the more parsimonious model (see Lasso appendix for full, annotated code including the lambda.min analysis). With lambda set to 'lambda.1se' or 0.05367992, the 10-Fold CV glmnet gives a model with the following predictors:

1. MSZoningRL
2. OverallQual8
3. OverallQual9
4. ExterQualTA
5. BsmtFinType1GLQ
6. CentralAirY
7. GrLivArea
8. KitchenQualTA
9. Fireplaces
10. FireplaceQuNF
11. GarageTypeAttchd
12. GarageCars
13. Age
14. SinceReMod
15. LogLotArea

16. Bath
17. TotalBsmtFinsh

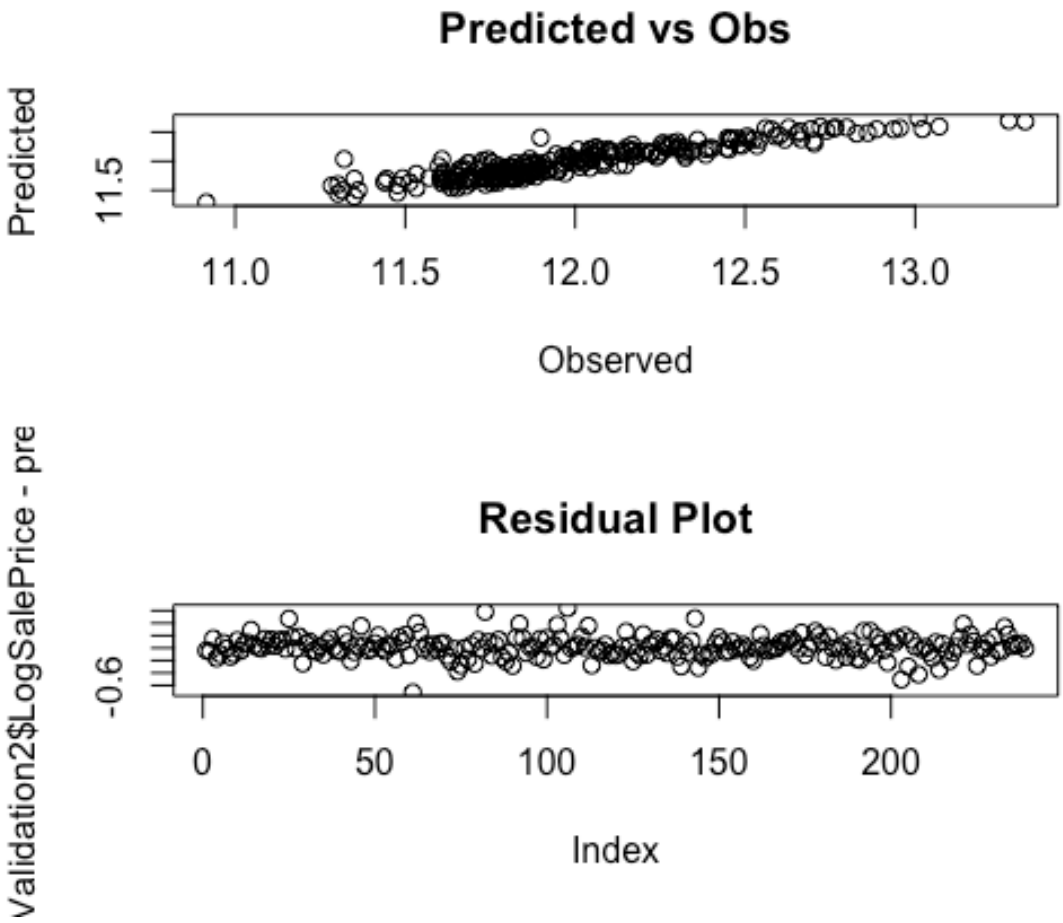
* No coefficients are supplied as this is an predictive test.

The “winning model” was then used to predict on the test data set, generating a MSE of 0.04452791 (very low). Comparing the training to the test curves:



Here we see that the two curve perform similarly towards the lower end of the optimal variable range – with 19 to 30 looking very similar. This is a strong result and validates selecting the more parsimonious model.

Looking at two further diagnostic plots – the predicted vs observed and the residual plot. The predicted vs observed seem to follow a nice 45 degree line while the residual plot is centered at 0 with no discerning pattern. This again backs the strength of the predictive model.

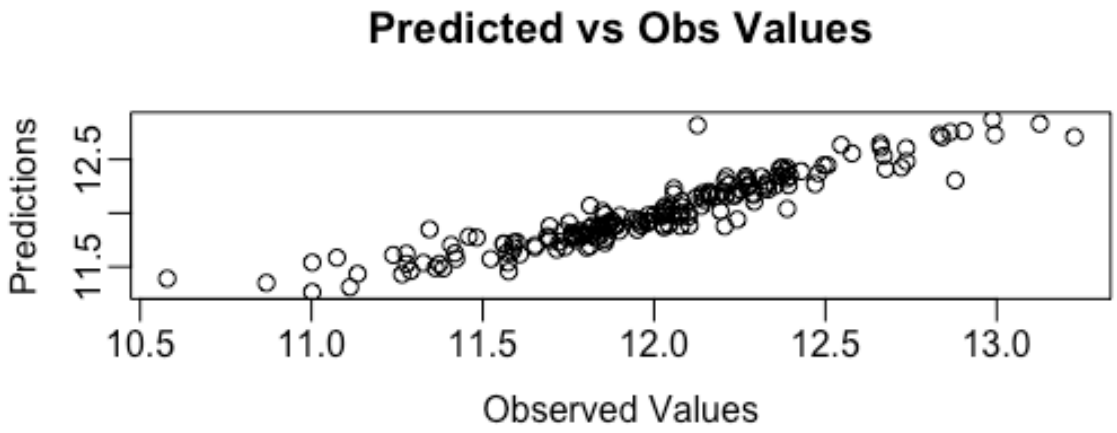


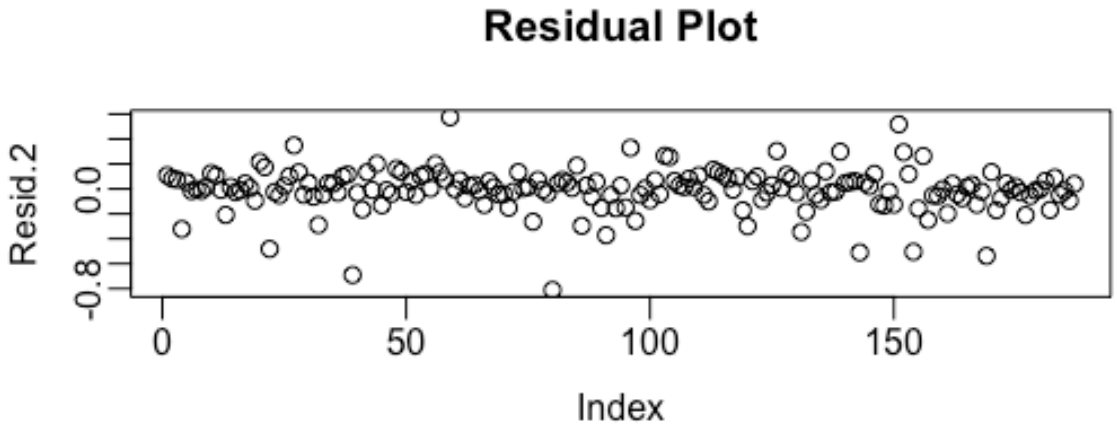
Random Forest:

The final Random Forest model had the arguments ntree = 100 (number of resampling regression trees), detail = 4 (nodesize), mtry = 20 (number of parameters to include).

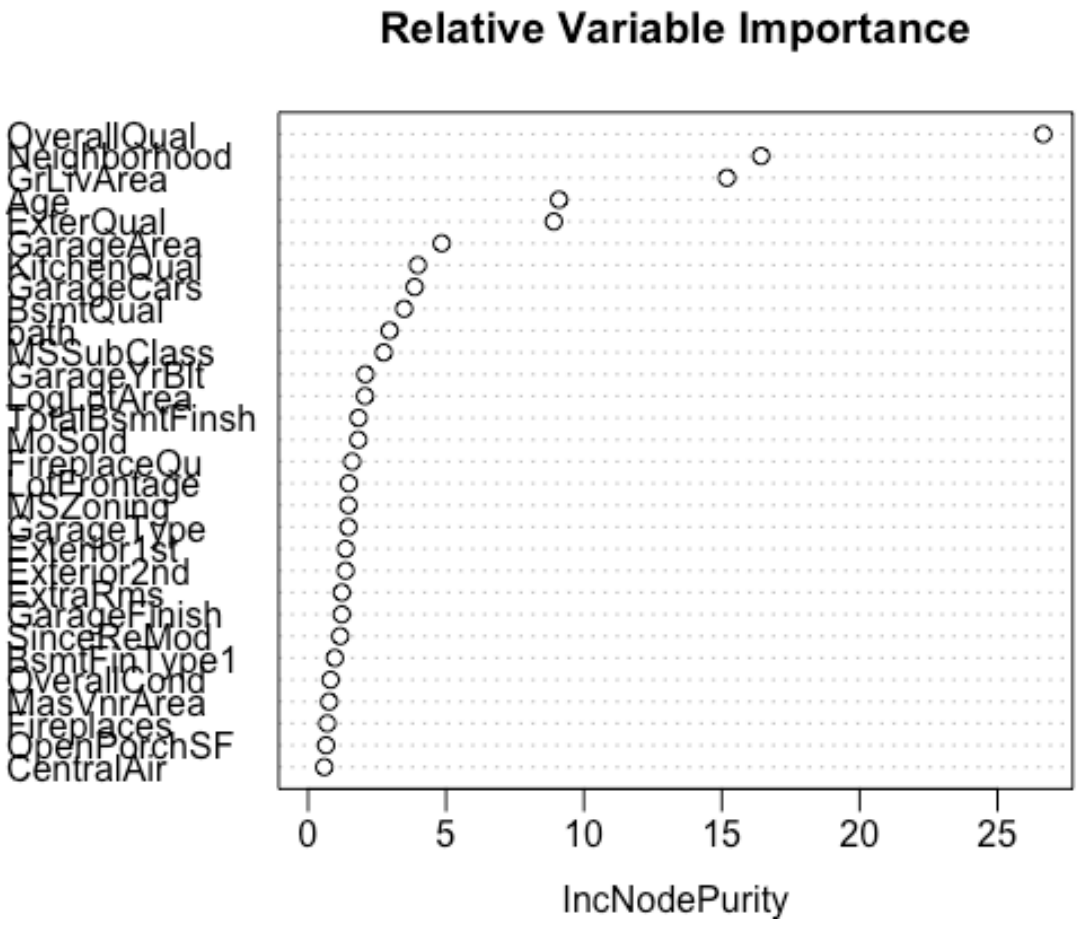
After predicting on the test data set, the "winning" random forest model produced a MSE of 0.0300127 – even lower than Lasso.

Looking at the same diagnostic plots as Lasso yields a similarly strong fit:





Lastly turning to the Variable Importance plot – which shows which variables give the most predictive power to the model:



While very crowded, it shows that:

- Overall Quality is the most important
- Neighborhood and size are on the next tier
- Age and Exterior quality are in the next tranche
- The rest of the Covariates drop off after that

Discussion

The study set out to produce a robust and accurate model for predicting House Sale Prices in Ames, Iowa. There was a secondary question of a light inferential investigating to determine which parameters were significant driver of Sale Price.

A Random Forest model (ntree=100, mtry = 20, detail = 4) outperforms a Lasso regression in predicting the Sale Price. Both models performed well with low MSE (Lasso – 0.04 and RF – 0.03) and suggested parsimonious models.

Moreover, the robust diagnostic plots of both methods show that the data manipulation and modeling methods were able to handle the covariate correlation – providing accurate predictions on both the training and test data.

Due to the relative low sample size and concentration of Sales (Ames, Iowa), both of the methods most likely have poor out-of-sample predictive power. If a reader is interested in starting their own Zillow, the author recommends creating a "Big Data" data-frame containing values from other cities in the country. Unless you are localized to Ames, Iowa – then use this model.

The study had original theorized that Age, Neighborhood and Zoning would have significantly impacts – with Zoning losing out in the Random Forest Variable Importance Plot.

Interestingly, the importance of variables varied across analysis method. With the most naïve analysis, the visualizations suggested that time since remodel, Neighborhood and zoning all looking to be important.

While the "best fit" Lasso contained all of these values (and most of the variables in the dataset), the "winning" Lasso left out Neighborhood and Exterior Quality – two variables that scored highly on the Random Forest Variable Importance plot.

In conclusion, a Random Forest method should be used to predict this data set with Overall Quality, Neighborhood, General Living Area, Age of House and Exterior quality all having a large impact on value.

If the reader is looking for a way to improve their house value without major constructions, change your exterior finish and decorate your yard. If the reader wants to know how much their house in Ames, Iowa is worth – plug it into the Random Forest model.

Graphs Appendix

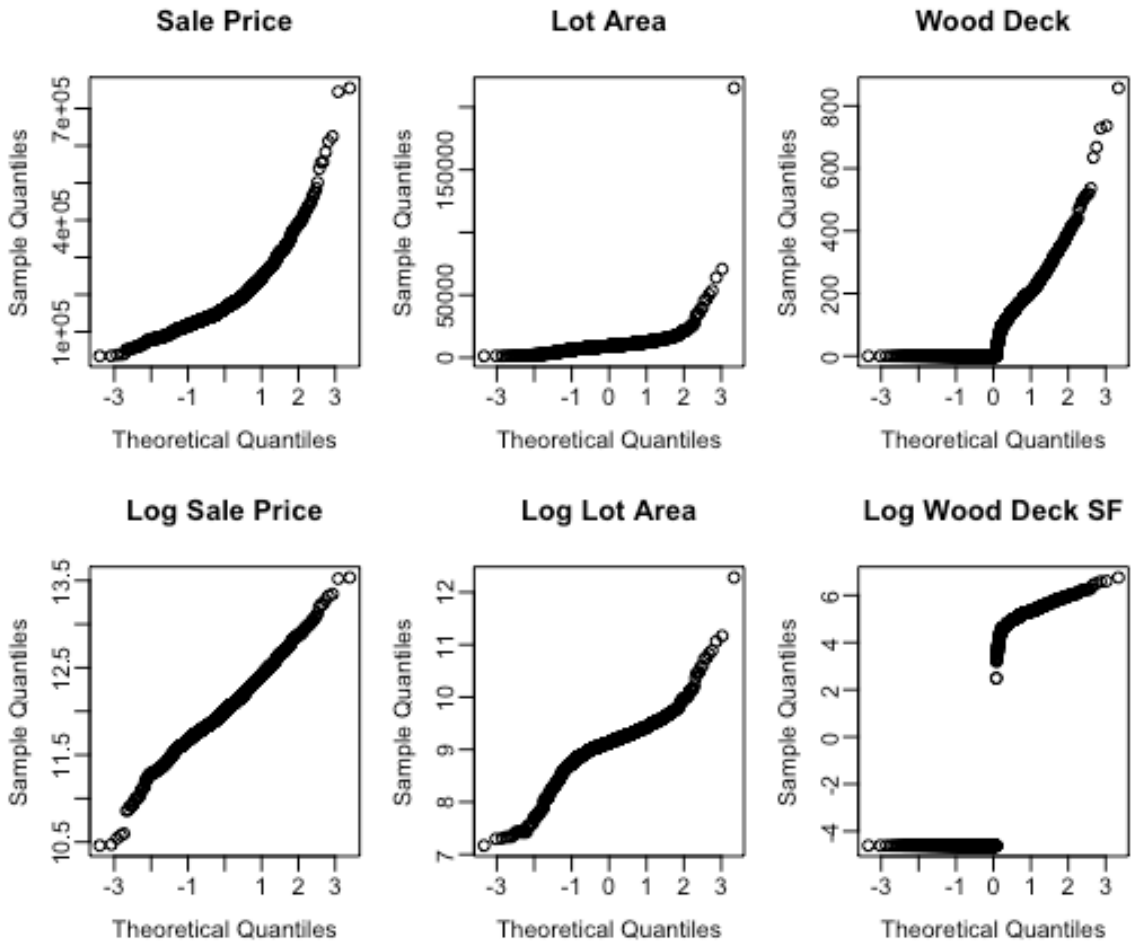
Additional graphs for the interested.

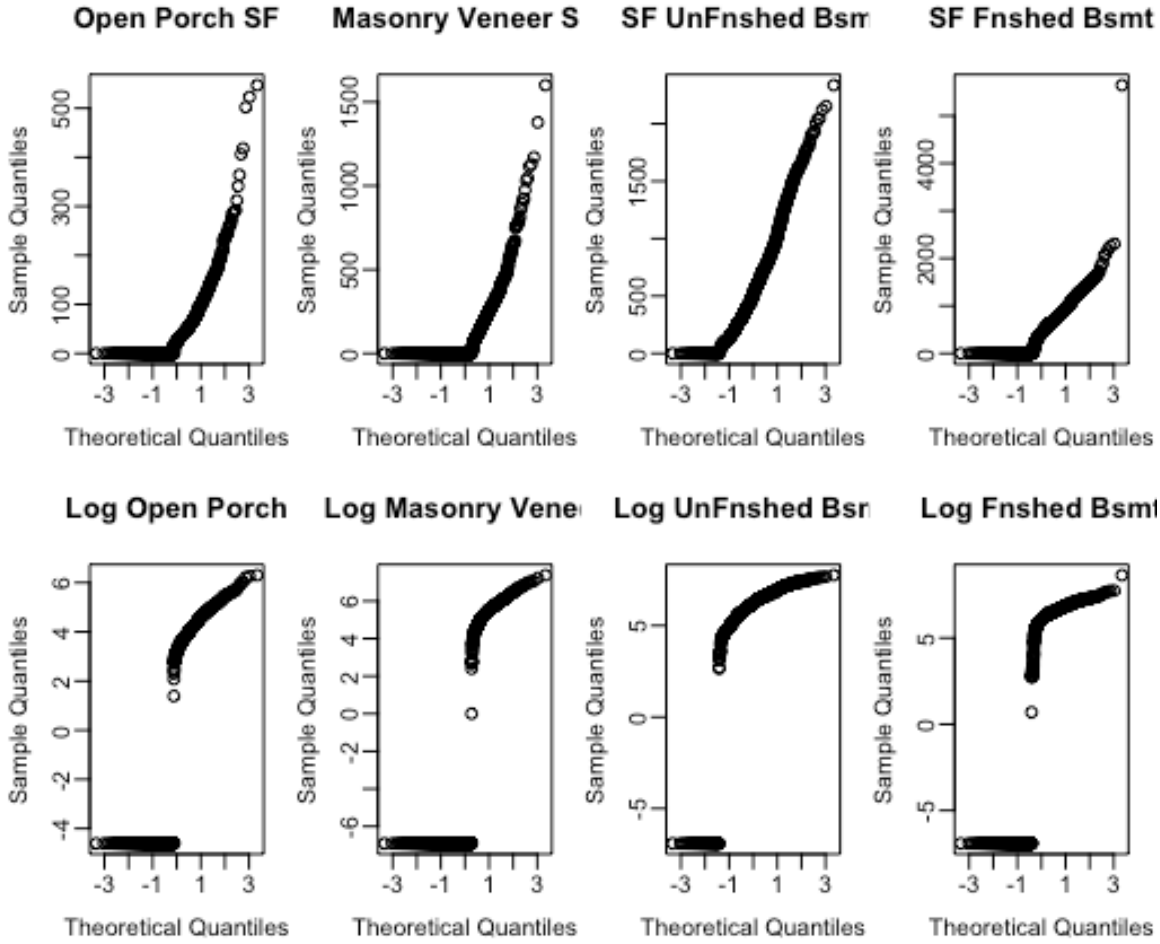
Preliminary Diagnostic Plots

Code for pairs plots:

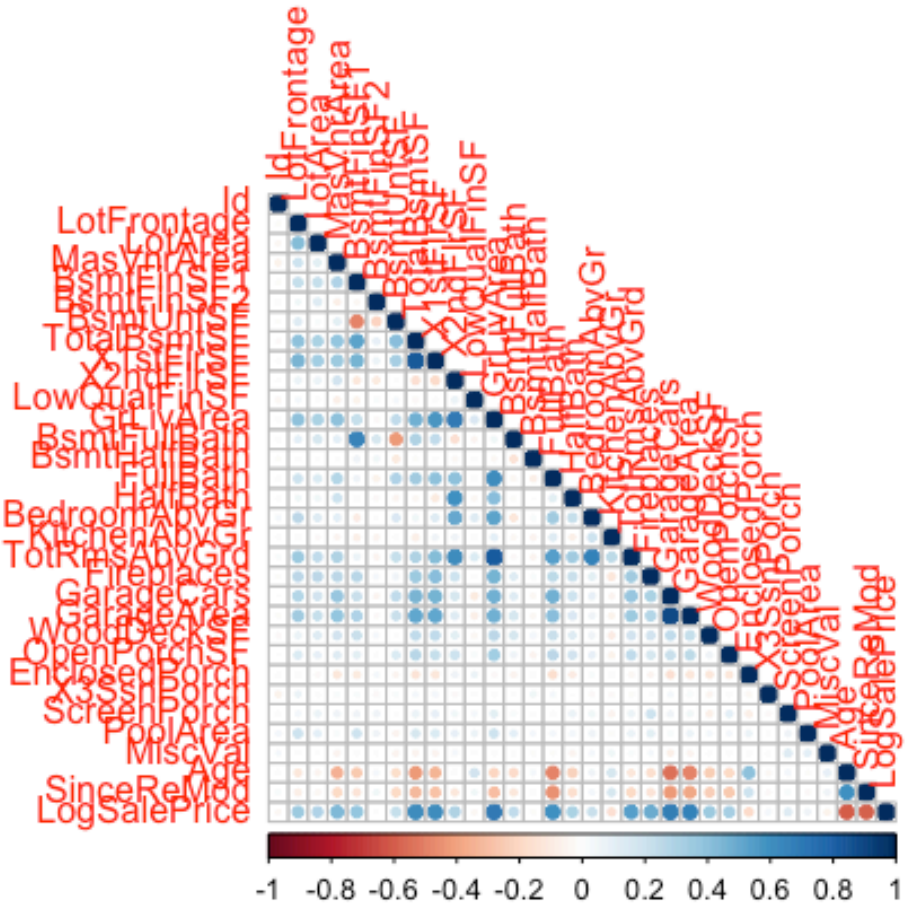
```
Num<-sapply(train6,is.numeric)
Num<-train6[,Num]

pairs.panels(Num[,c(1:10,31)], scale = TRUE)
pairs.panels(Num[,c(11:20,31)], scale = TRUE)
pairs.panels(Num[,c(21:30,31)], scale = TRUE)
```

QQ Plots





Correlation Plot:



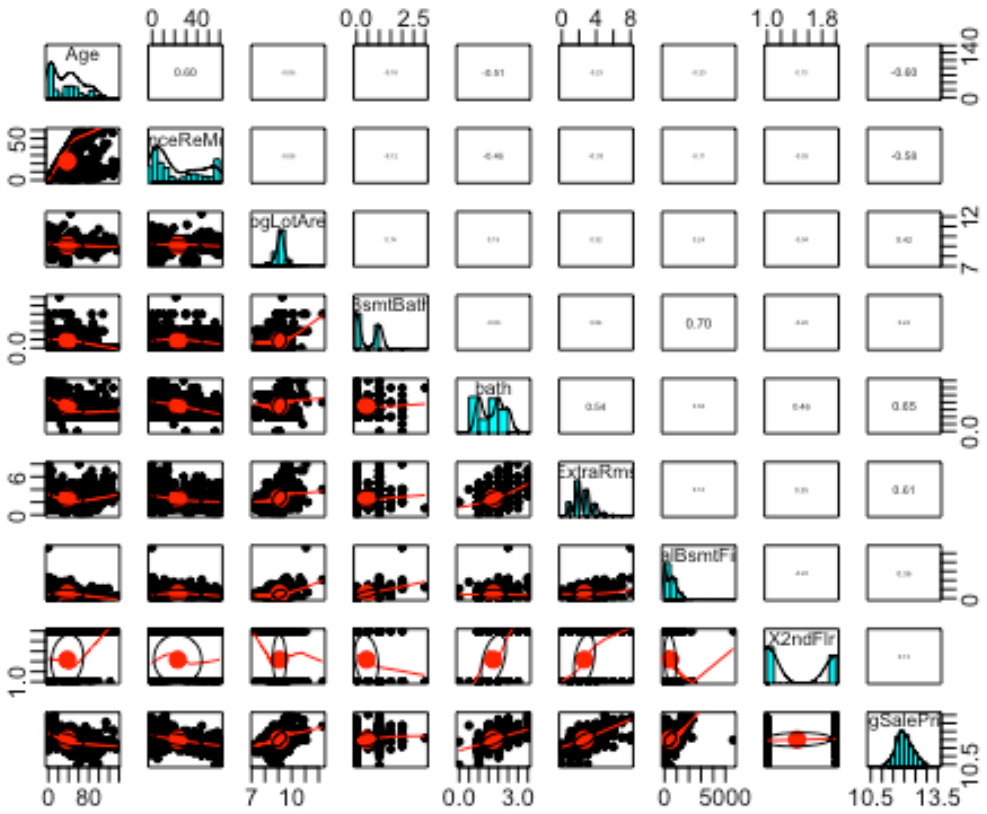
Code For Correlation Plot

```
require(corrplot)

par(mfrow=c(1,1))
correlations<- cor(Num,use="everything")
corrplot(correlations, method="circle", type="lower",  sig.level = 0.01, insig =
"blank")
```

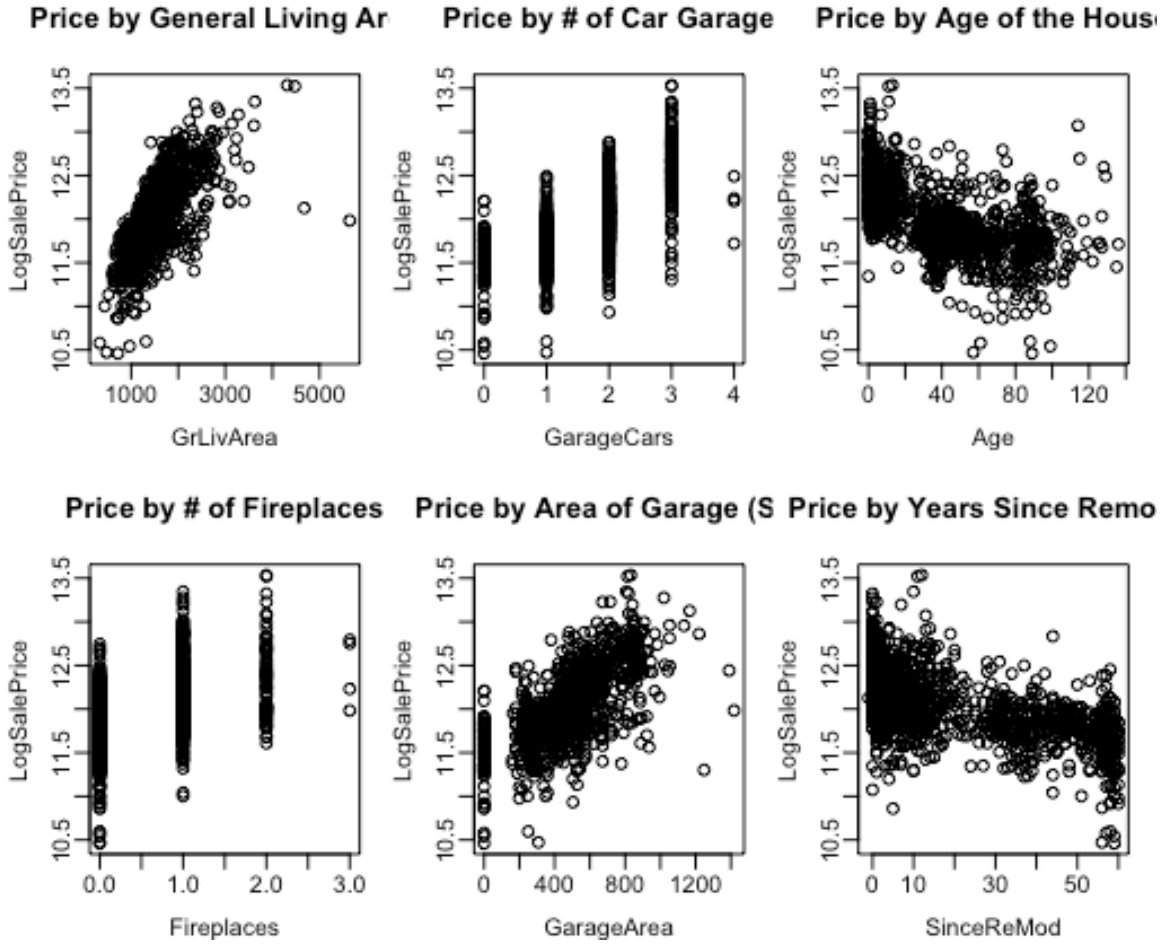
Post Cleaning Visualizations

Code for Pairs Plot with one produced:

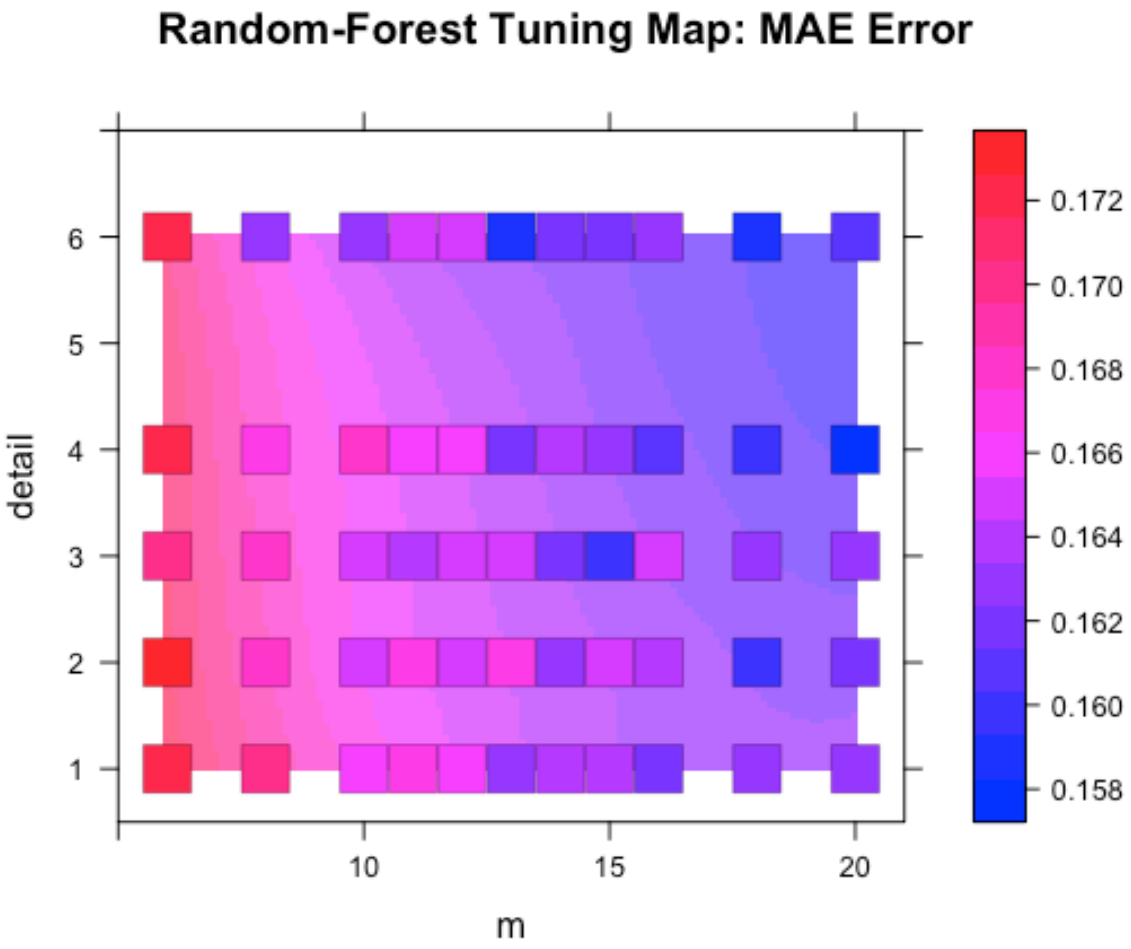
```
pairs.panels(Num3[,c(1:8,20)], scale = TRUE)
pairs.panels(Num3[,c(9:17,20)], scale = TRUE)
pairs.panels(Num3[,c(18:19,21:26,20)], scale = TRUE)
```

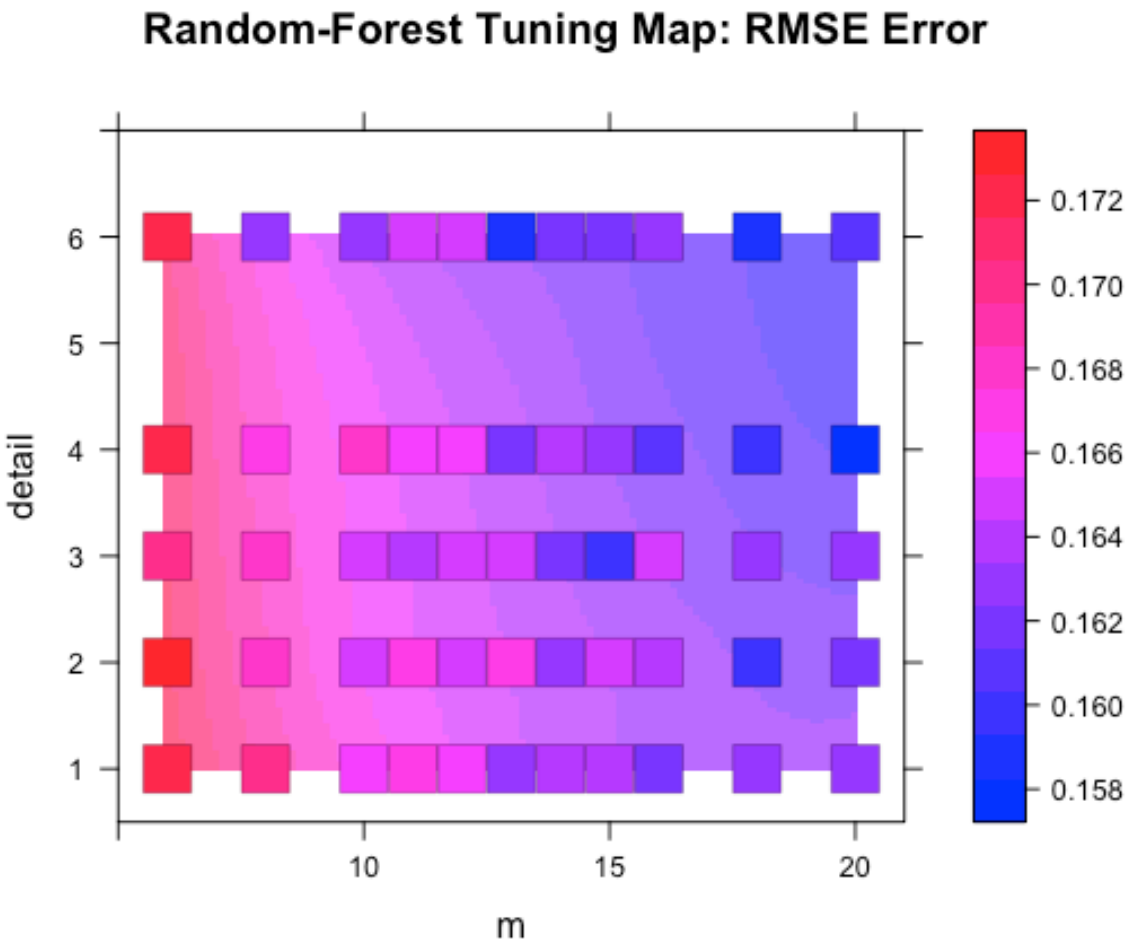


Numeric Variable Highly Correlated with LogSalePrice:



2D Random Forest Tuning Surfaces


Random-Forest Tuning Map: MAE Error
detail
1
2
3
4
5
6
10
15
20
m
0.172
0.170
0.168
0.166
0.164
0.162
0.160
0.158


Random-Forest Tuning Map: RMSE Error
detail
1
2
3
4
5
6
10
15
20
m
0.172
0.170
0.168
0.166
0.164
0.162
0.160
0.158