**To:** Professor Yu Du

**From:** Micah Bray

**Subject:** Final Project

**Date:** 5/12/2019

**PROBLEM STATEMENT**

This summary seeks to find the best linear regression model to predict the Selling Price of homes in ski towns.

**EXECUTIVE SUMMARY**

**Major Finding(s)**

*86.46% of the variation in home Sale Price in Ames, Iowa is explained by the variation in the overall material and finish quality, log of the total square feet of the house, size of garage in car capacity, sum of year built and remodel date, number of fireplaces, log of the lot size in square feet, total porch area in square feet, and unfinished basement area in square feet, taking into account the sample size and number of independent variables*

**Recommendation(s) for Action**

*Given my findings, my recommendation is to utilize the suggested model for the average home. Not necessarily a small or large home or a home of below or above average overall quality. Since the predicted values start to vary widely with Sale Prices at the tip and tail of the Sale Price distribution, it could be beneficial to attempt a random forest or machine learning approach to improve predicted values.*

**Analytical Overview**

A quick normality summary of the data set and log transformation of right-skewed variables was added to data set. I identified variables with a large number of nulls and some variables with outliers. Then, I created a scatter matrix (Exhibit F) to find correlation to SalesPrice, identify any other possible transformations that might not appear in a normality summary (quadratic, exponential), as well as identify any potential collinearity that may arise. Best subsetting and stepwise techniques were utilized in conjunction with correlation for variable selection. Finally, an iterative process (model fitting, model diagnosis and cross-validation) was utilized to locate the highest adjusted R squared, while trying to eliminate colinearity and minimize RMSE.

**APPENDIX**

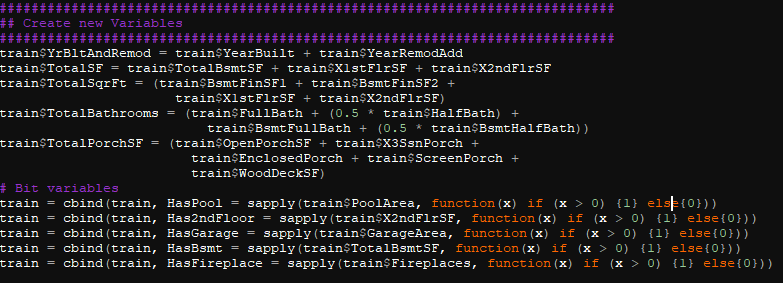
**Model Fitting/Selection**

*First, I created some variables that totaled variations of square feet and indicated whether a home feature was present (Exhibit A). Next, adjustments were needed for skewed variables, so a log transformation was completed (Exhibit B shows SalePrice approaching normality with a log transformation). Outliers were identified in LotArea\_log and TotalSF\_log (Exhibit D and E). However, outliers would exist in the “test” data set, so outliers were only removed to see how they could improve the adjusted r-squared (Exhibit R). Dummy variables for categorical features (text) were created to the n-1 level. Then, a scatter/correlation matrix of the dummy variables to SalePrice was completed. The matrix showed that the significantly correlated dummy variables pertain to quality, which comprise the variable OverallQual (Exhibit Y), but have a mild correlation of less than 0.6. Since OverallQual is highly correlated to SalePrice and we want to reduce the likelihood of collinearity with the mildly correlated dummy variables, we are going to ultimately use numerical variables. Because they could cause perfect colinearity in the subsetting approach, I utilized the correlation matrix of numerical values (to SalePrice) to help eliminate variables before subsetting (Exhibit T). The correlation matrix showed some heavily correlated (to SalePrice) variables such as: OverallQual, TotalSF\_log, GrLiveArea\_log (Exhibit H). Best subsetting identied some additional variables that could add value like: GarageCars, YrBltAndRemod, Fireplaces, LotArea\_log, and TotalPorchesSF (Exhibit I, J, K, and L). Best subsetting identified a model with only 9 variables (Exhibit I shows that model performance decreases with additional variables), but TotalSF\_log and GrLivArea\_log were highly colinear (Exhibit P). Model 3 was used instead of Model 1, due to colinearity, and the adjusted r-squared only shifted slightly from 0.866 to 0.8646. K-Folds cross-validation was utilized to illustrate that Model 3 was the best overall choice in terms of adjusted r-squared and RMSE (Exhibit U). Some final graphing was completed for fun (Exhibit V – Exhibit X).*

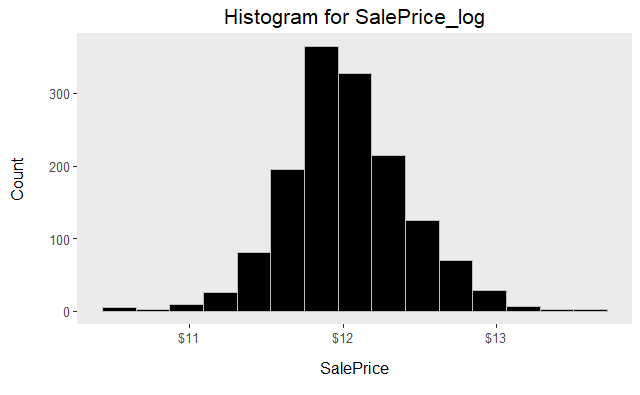
**Model Validation**

*The model has linearity. This was tested by creating a plot showing the predictive vs. actual values. The points on the plot were symmetrical around a diagonally fitted line, suggesting linearity (Exhibit N). Independence of errors was tested through a plot of the residuals (Exhibit M). There doesn’t seem to be a discernable pattern in the residual plot, so we can safely assume independence of errors and an equal variance in those errors. In a QQ Plot of the residuals (Exhibit O), the residuals do not form a perfectly straight line. The line is virtually straight through the center, with variation at the tip and tail. To me, this so closely approaches normal that it passes our validation test. The variations are likely caused by extremes in housing prices that haven’t been accounted for in the data. The Durbin-Waston test illustrates a lack homoscedasticity (Exhibit P). Collinearity, as measured by the VIF function, was initially relatively high (above 5) in Model 1 between TotalSF\_log and GrLivArea\_log. However, this was eliminated in Model 3 with the removal of the variable GrLivArea\_log (Exhibit P). This suggests that there isn’t collinearity in the model.*

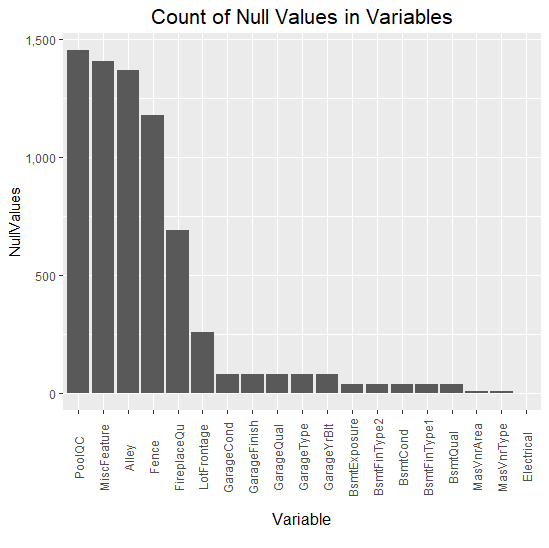
**Exhibit A**

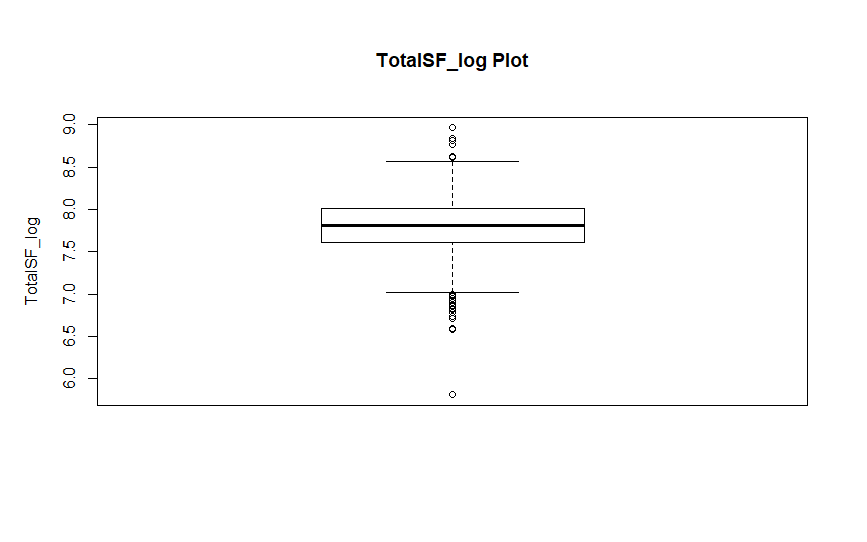


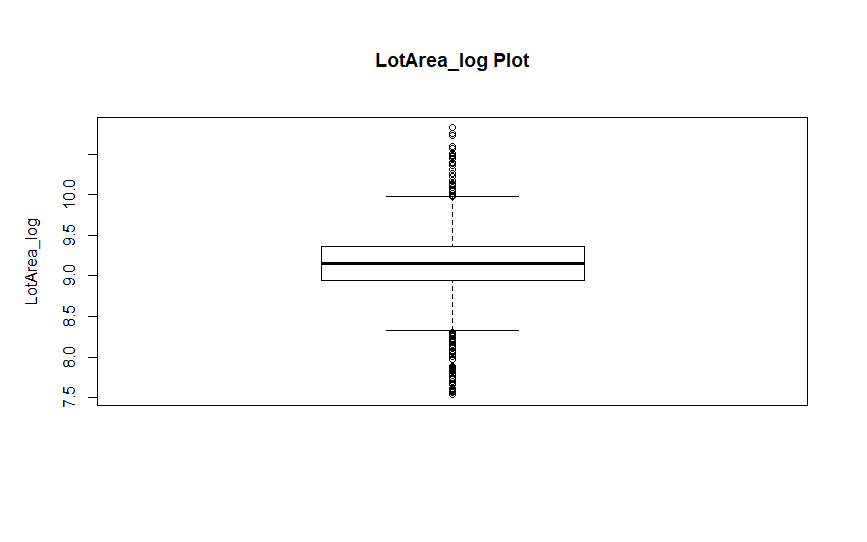
**Exhibit B**

****

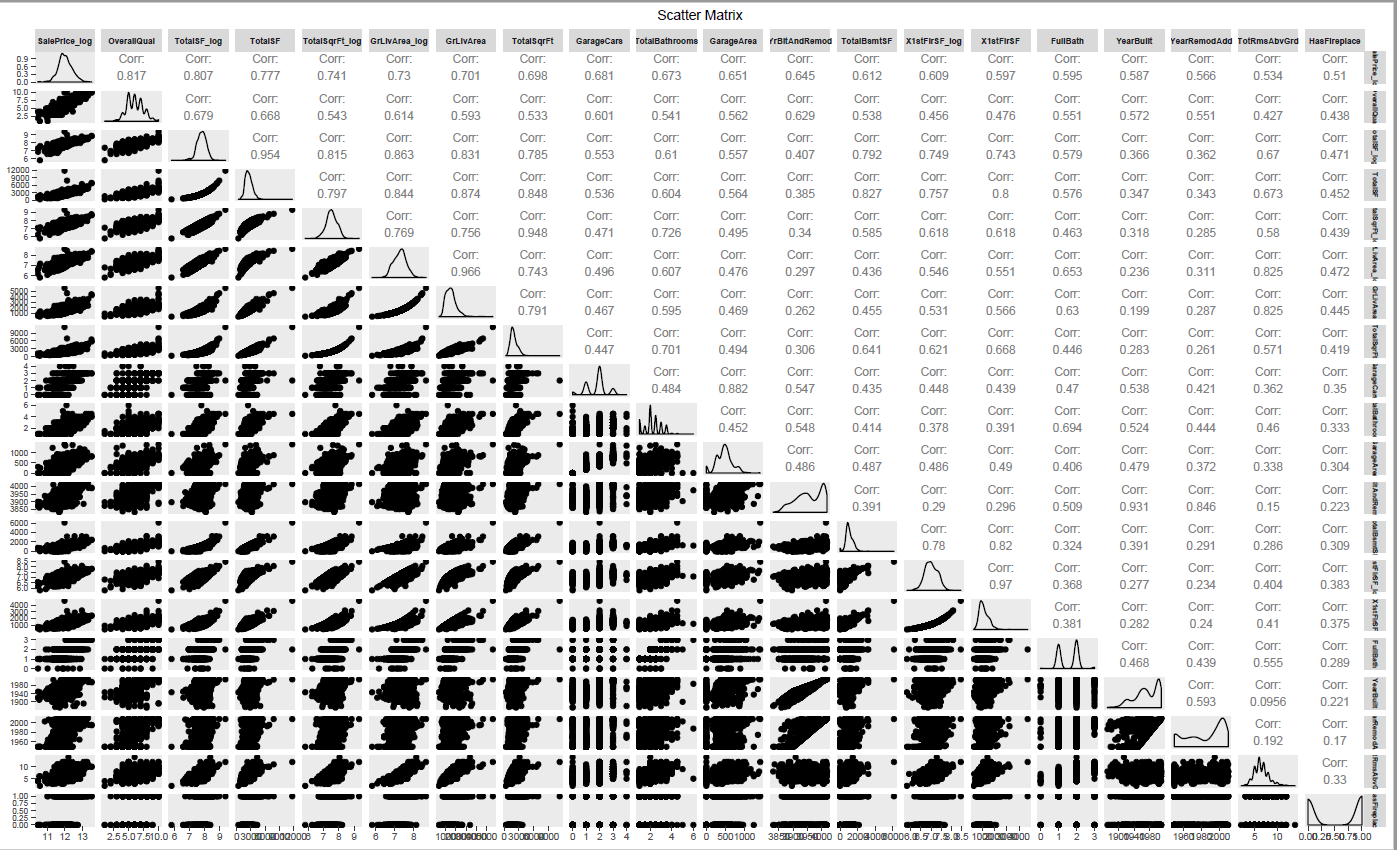
**Exhibit C**

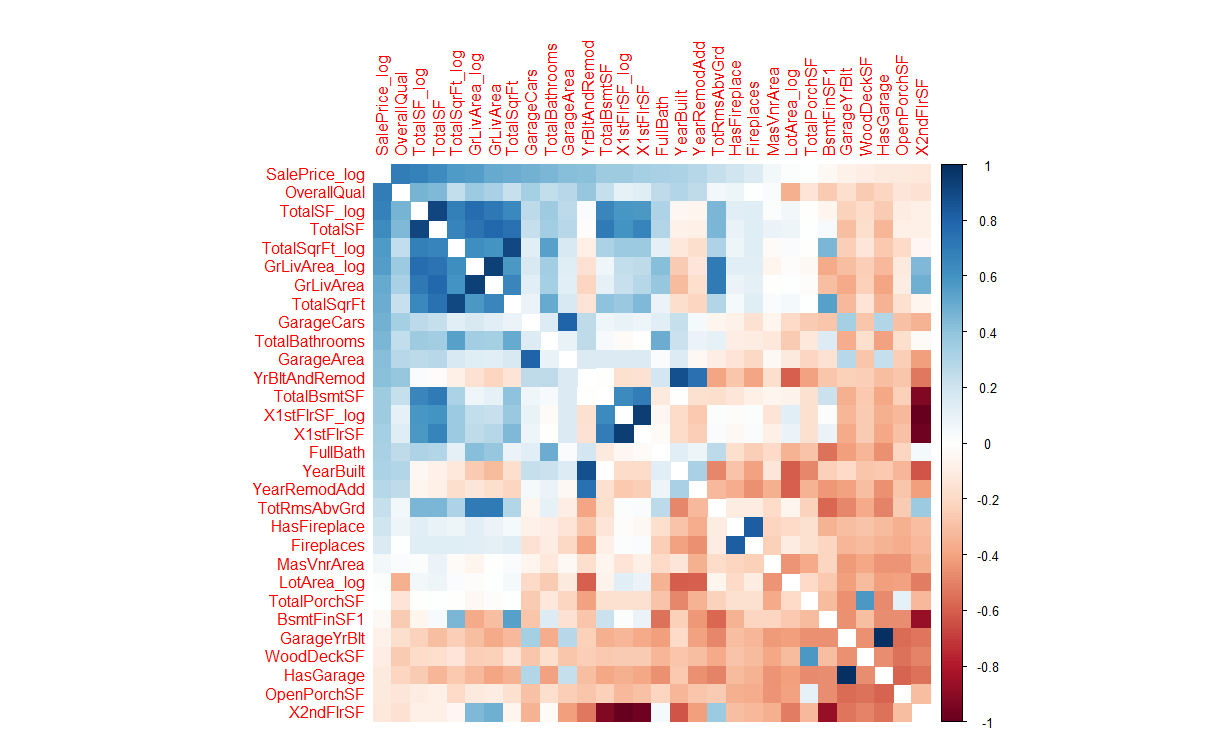
****

**Exhibit D**

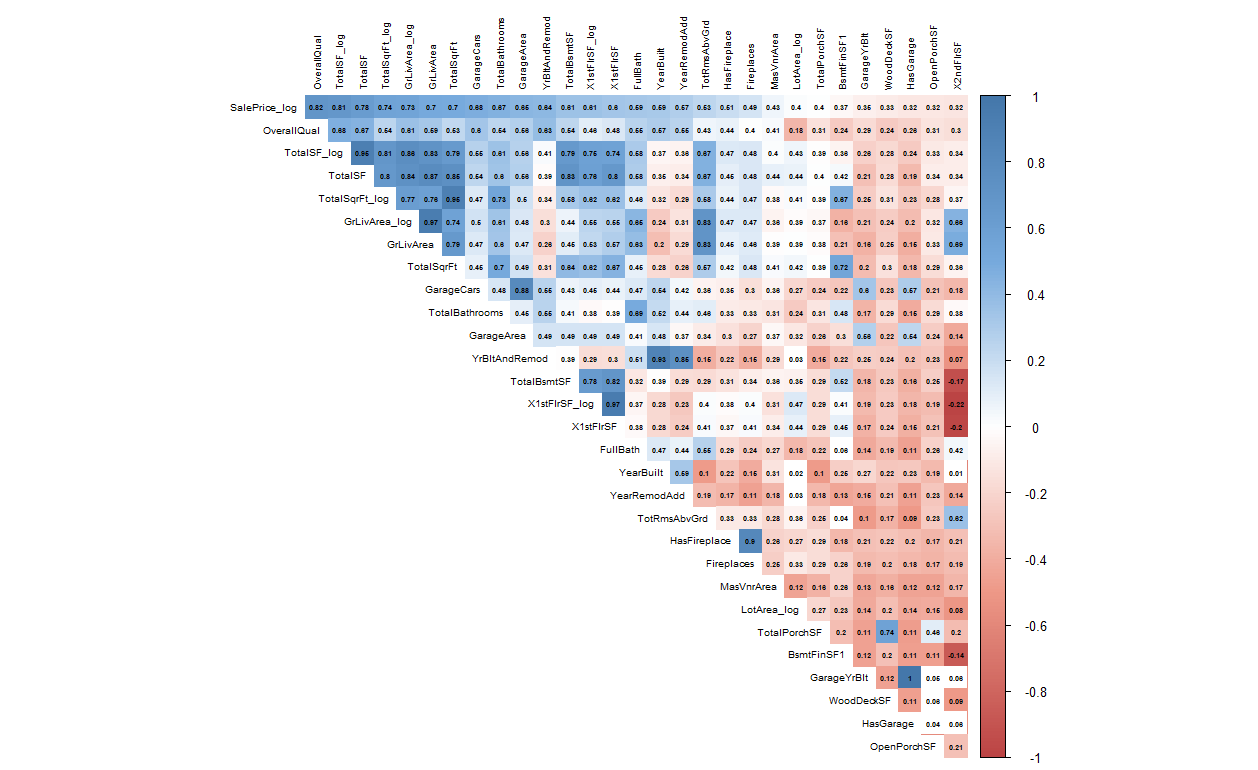
**Exhibit E**

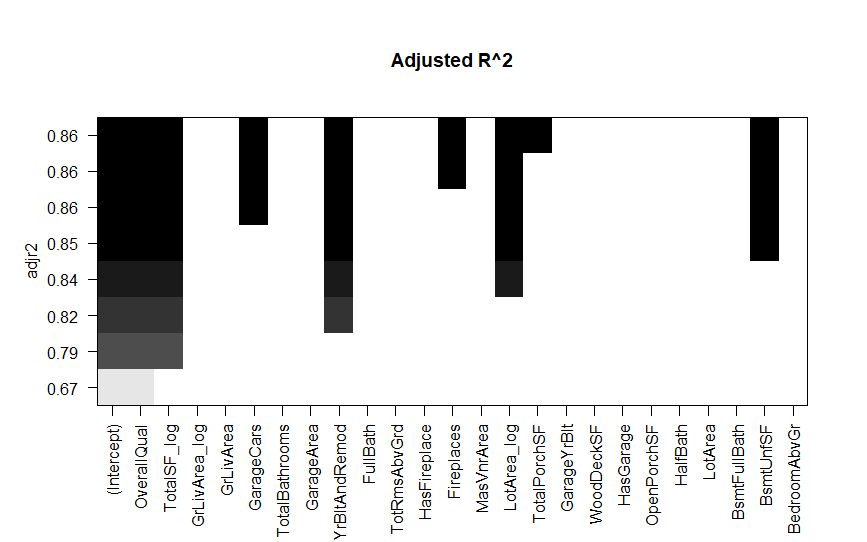
**Exhibit F**

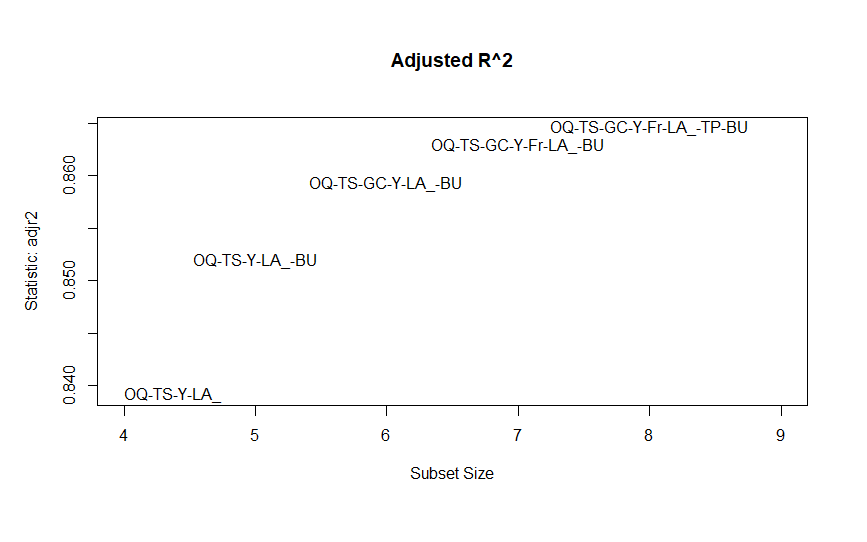


**Exhibit G**

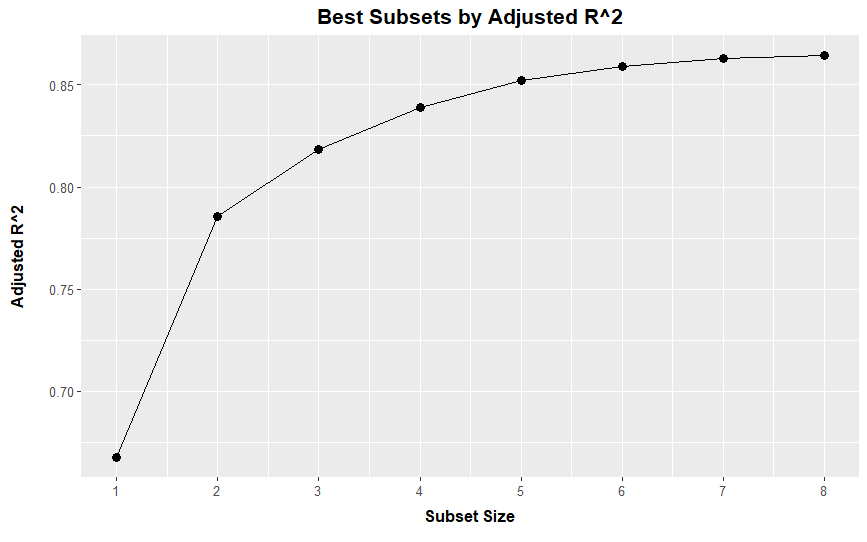
**Exhibit H**



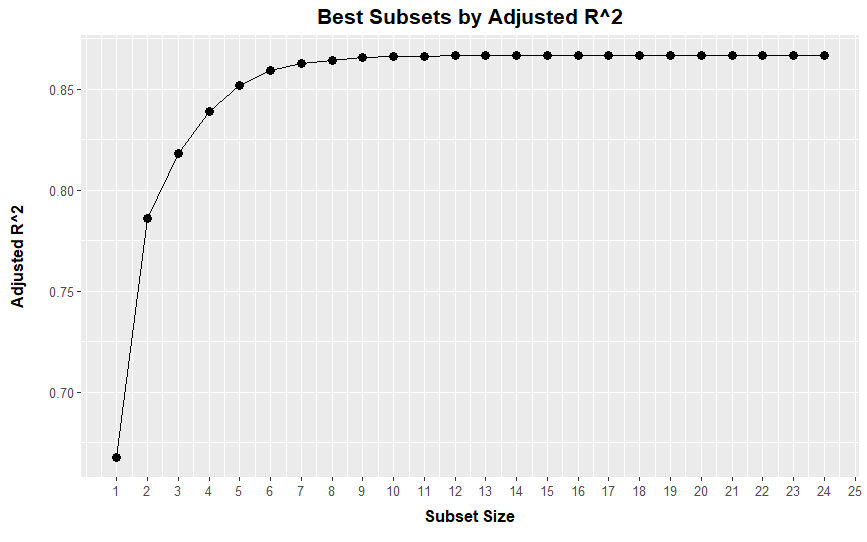
**Exhibit I**

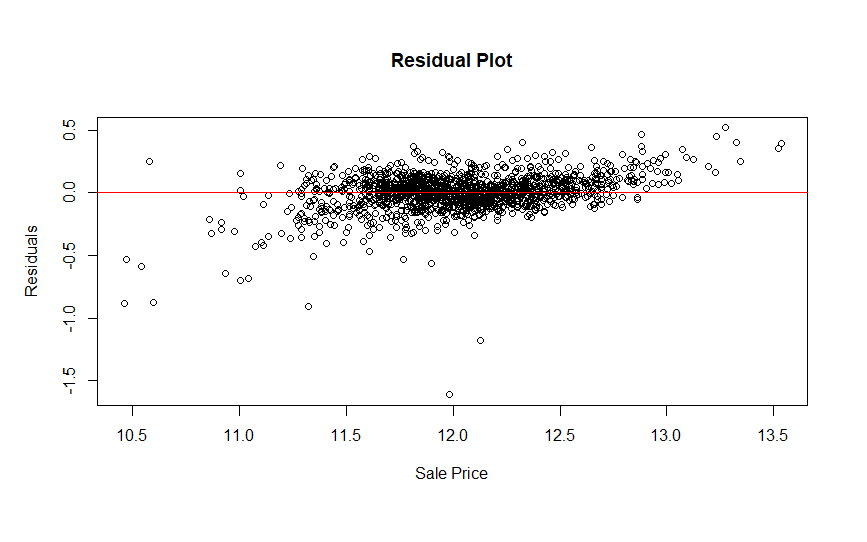
**Exhibit J**

**Exhibit K**

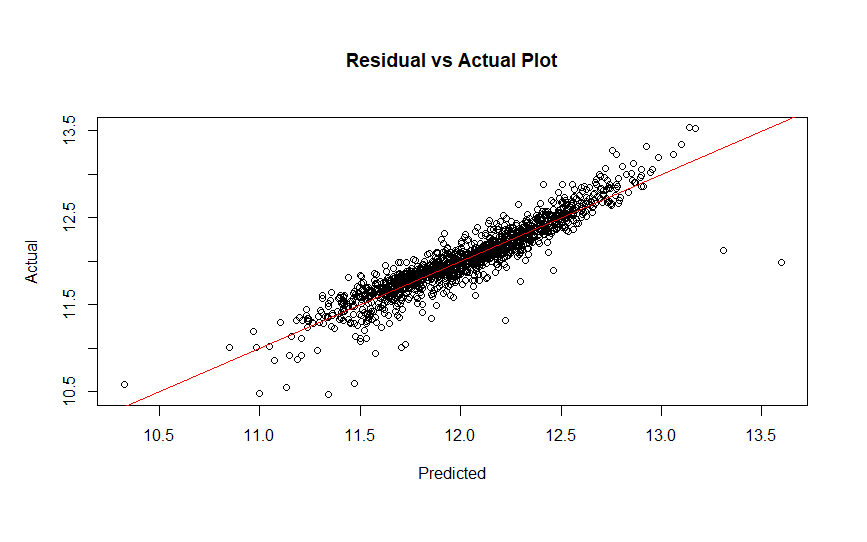


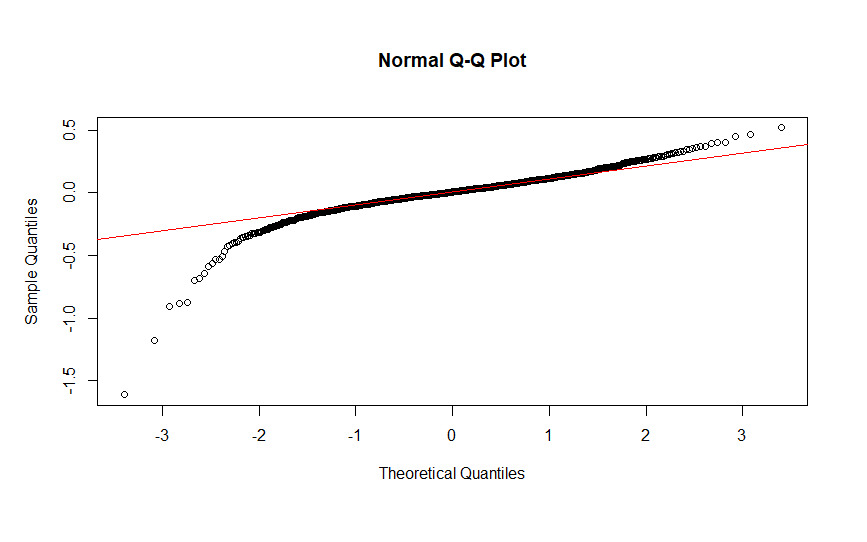
**Exhibit L**



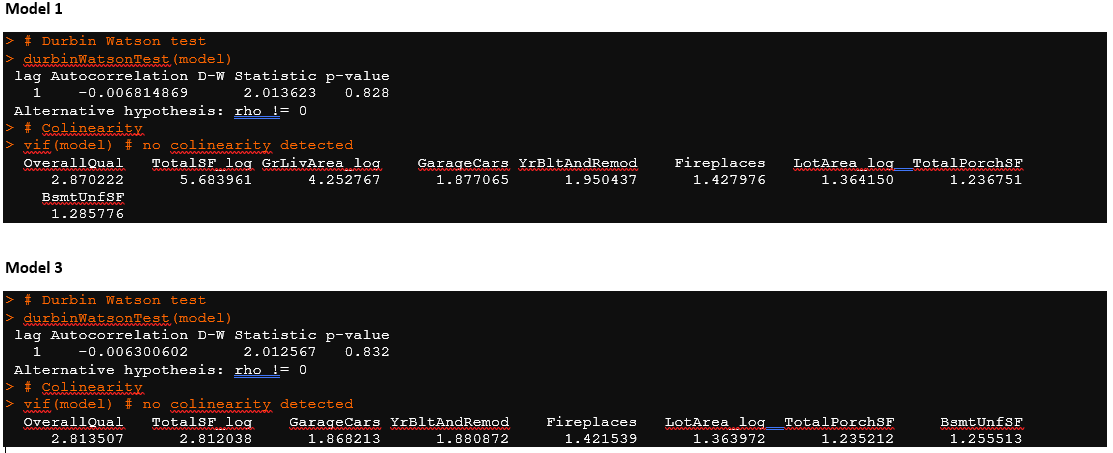
**Exhibit M**

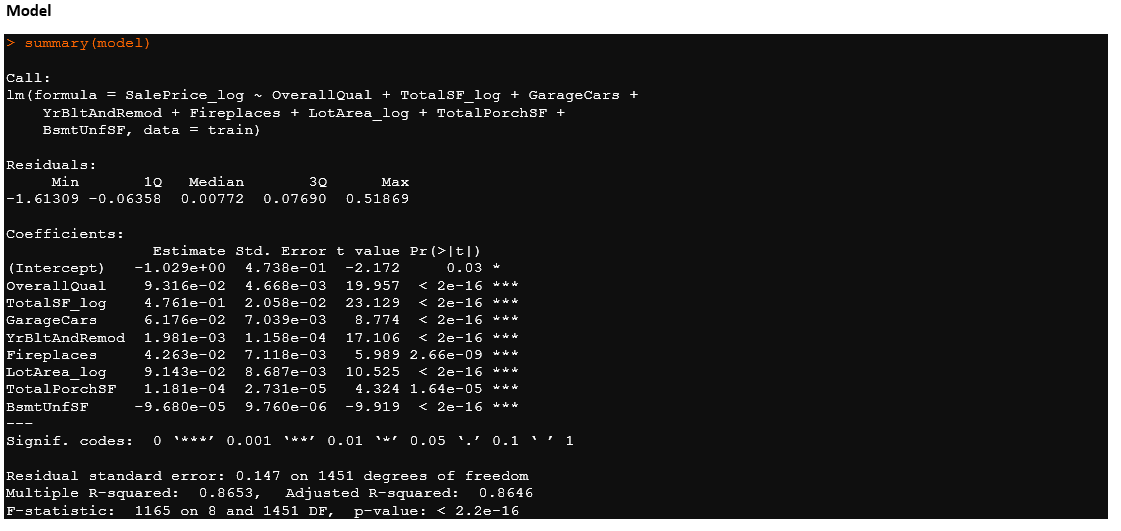
**Exhibit N**



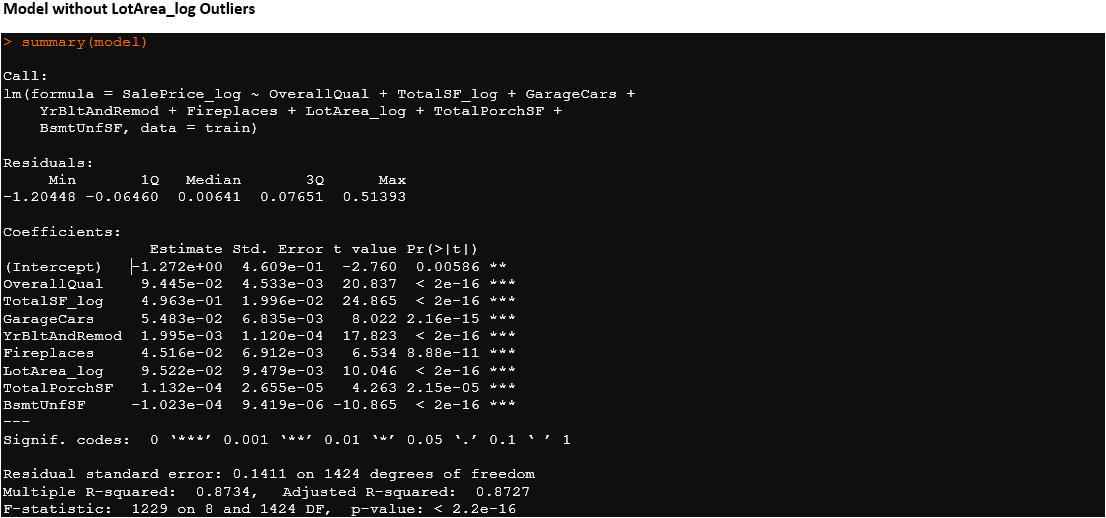
**Exhibit O**

**Exhibit P**

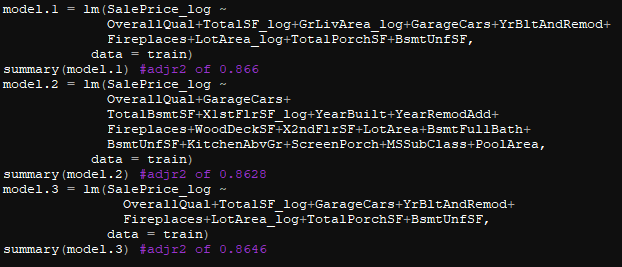


**Exhibit Q**

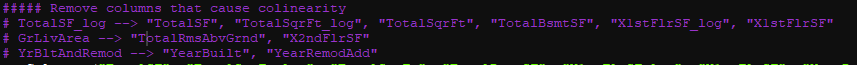
**Exhibit R**



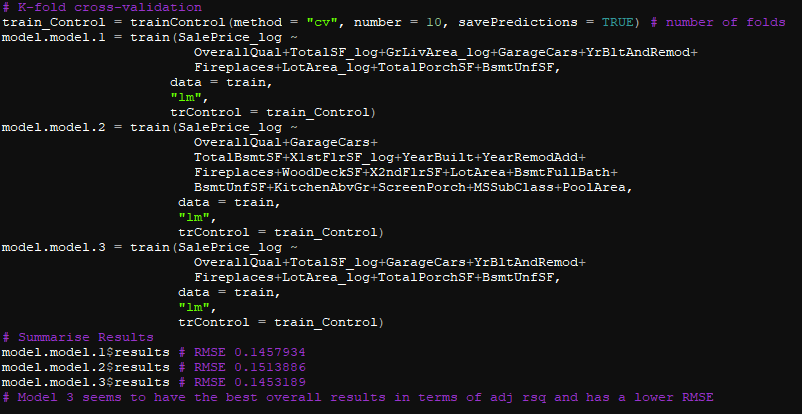
**Exhibit S**



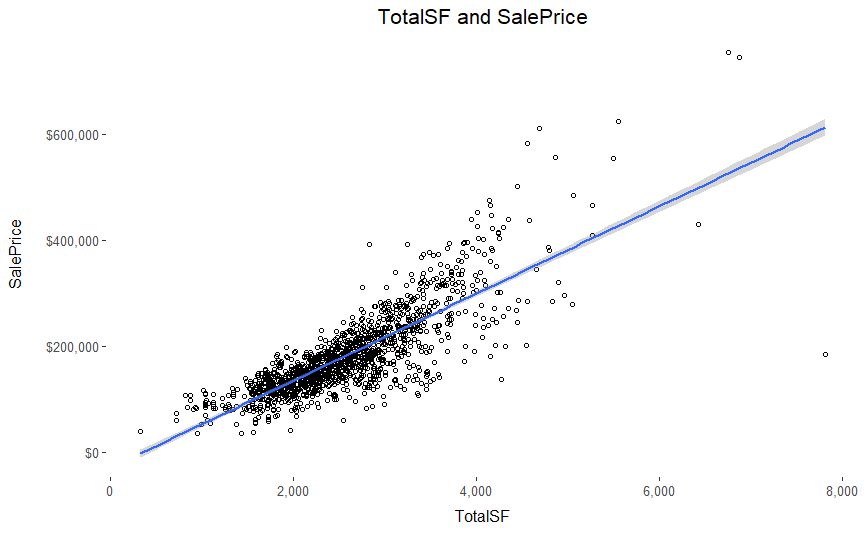
**Exhibit T**



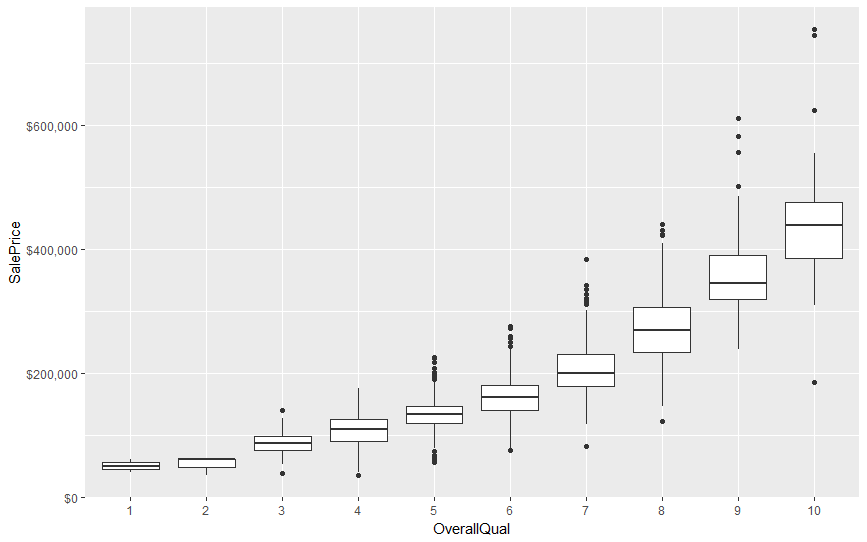
**Exhibit U**



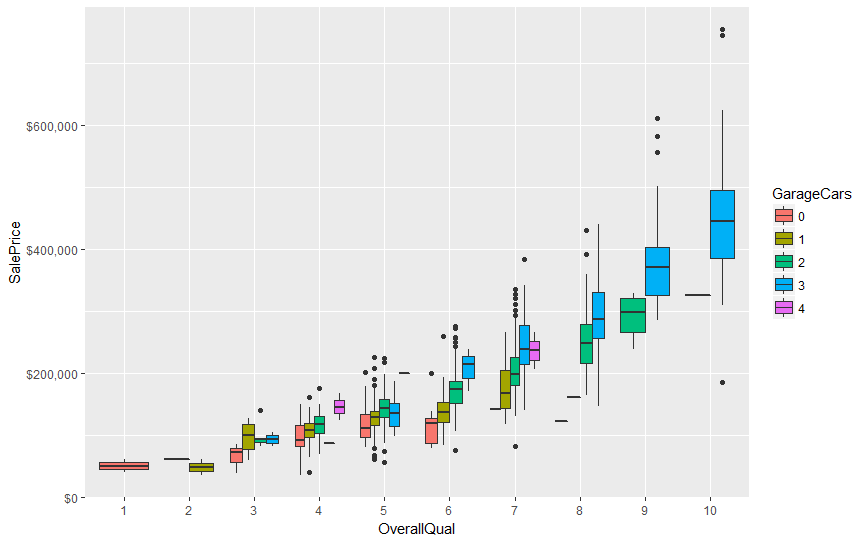
**Exhibit V**



**Exhibit W**



**Exhibit X**



**Exhibit Y**

