**MSDS 6372 Project 1**

*Zackary Gill, Tej Tenmattam, Limin Zheng*

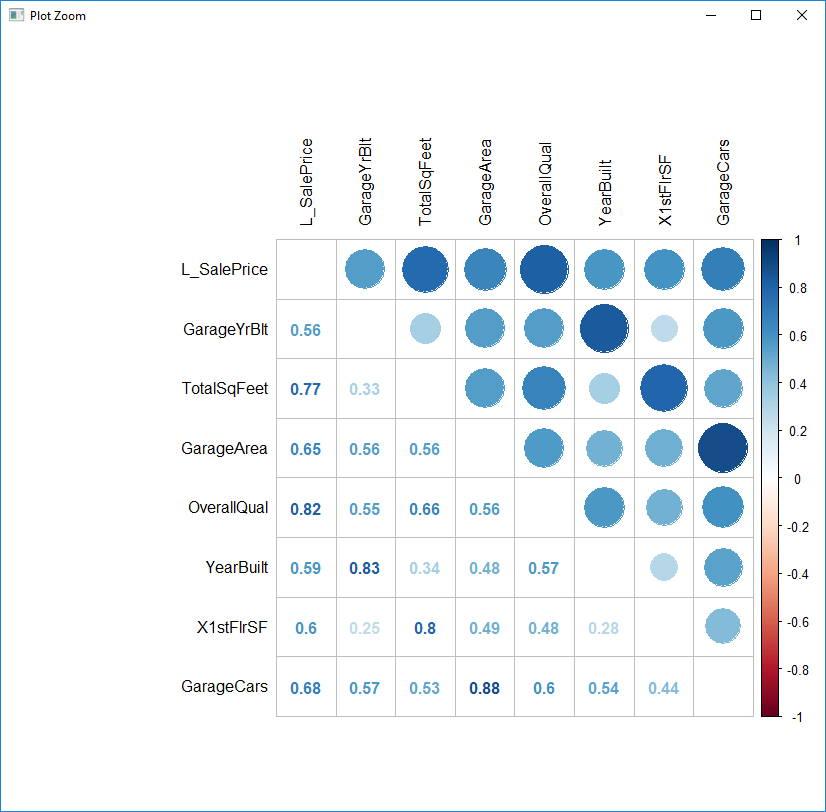
1. **Introduction**

With the large number of houses available on the market it is difficult to have a professional go through a home and come up with a reasonable price for a home. Our goal is to create a few models that will allow houses to be quickly and accurately priced. To accomplish this goal, we have worked on some detailed EDA and many different modeling techniques to identify an algorithm that performs better with a train/test sets RMSE-score. Two of our models will be complex with the third being easy to explain in order to allow people to quickly see what the most important things are that relate to the SalePrice of their home.

1. **Data Description**

This dataset is from the Ames area of Idaho and contains 1460 observations with 79 explanatory variables and one response variable called “SalePrice”. Each one of these explanatory variables describes nearly every aspect of the residential homes in that area. For more information about the dataset go to Kaggle’s website (https://www.kaggle.com/c/house-prices-advanced-regression-techniques).

1. **Exploratory Analysis**



[**CHART 3.1**]

The initial examination of the data resulted in finding approximately 19 columns have missing data. We examined each of these and fixed those with logical values (EX: with Fence being NA, it is assumed that there is no fence). Next we removed categorical columns that had problems with their levels (EX: Utilities has two levels, 1459 of the rows were of one level and the final row was of the other level). Examining a scatterplot of the variables led us to take a log transform the response variable ‘SalePrice’. Removal of columns that had too much missing data and the consolidation of redundant columns were next. We examined the correlation plots **[CHART 3.1]** and removed a few highly correlated columns that were describing similar attributes. At this point the data was clean so we moved on to the model building analysis.

1. **Objective 1**
2. **Problem Restatement**

This report presents the analysis and modeling done on the House Pricing dataset provided from Kaggle for advanced regression technique. It starts by acquiring and defining input data, data cleaning, feature engineering, and finally by model training, selection, and prediction. The end result is in the creation of a model that explains/predicts the SalePrice of a home.

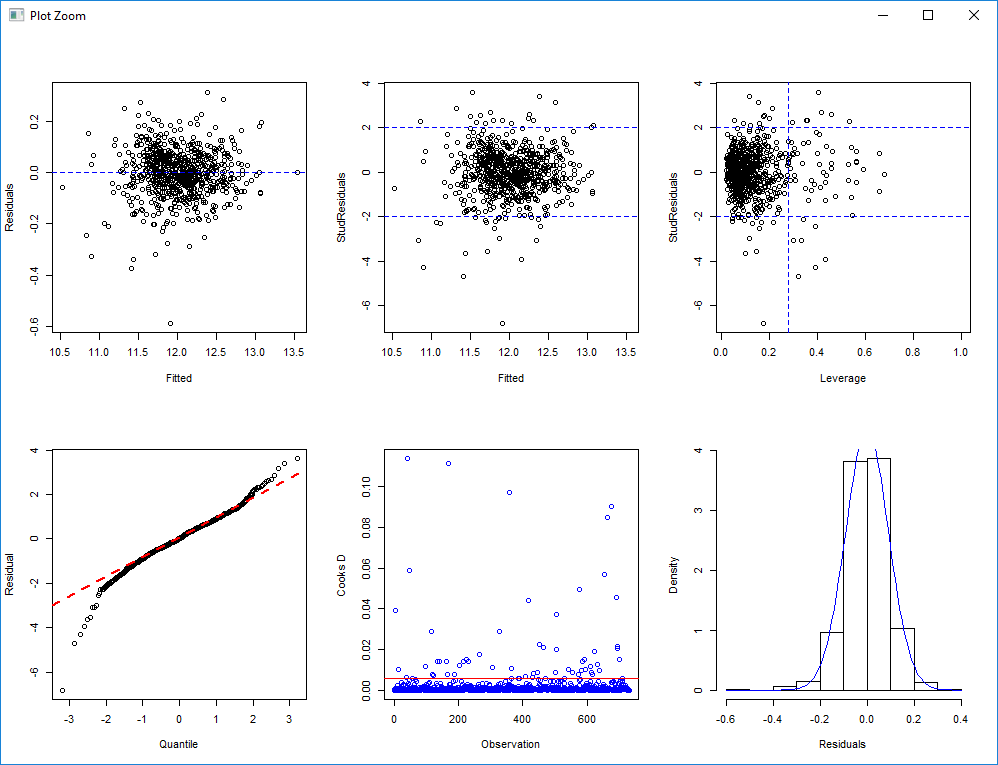
1. **Build and Fit Models**

During our analysis we built three different models. Two of them were built using automatic selection algorithms: stepwise and lasso. The other model was custom and was designed to be easily interpretable.

**Stepwise**

For the creation of the stepwise model we removed a few categorical parameters that did not have enough of certain levels to do testing upon them. Also we included the interaction terms: Neighborhood\*LotArea, Neighborhood\*GarageCars, TotalSqFeet\*FullBath, TotalSqFeet \*GarageCars, TotalSqFeet \*BedroomAbvGr, and Neighborhood\*OverallCond. When doing the selection we also ran into an issue with the categorical variables BsmtQual/BsmtExposure being highly correlated and GarageQual/GarageCond being highly correlated. We removed BsmtQual and GarageCond to fix this. In the end the stepwise model selected 33 of the variables and 1 interaction. Those variables are: MSZoning, LotArea, Street, LandContour, LandSlope, Neighborhood, Condition1, BldgType, OverallQual, OverallCond, YearBuilt, RoofMatl, ExterQual, Foundation, BsmtExposure, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, CentralAir, LowQualFinSF, AllFullBath, AllHalfBaths, KitchenAbvGr, KitchenQual, FireplaceQu, GarageCars, GarageQual, WoodDeckSF, OpenPorchSF, ScreenPorch, MiscVal, SaleCondition, TotalSqFeet, and one interaction term GarageCars\*TotalSqFeet.

The initial model seemed to have a small problem with normality with an outlier. We ran it again without those few outliers and looked at the results. Normality and constant variance were slightly better, but the Adjusted R-Squared value was slightly worse. We decided to leave the points in and continue with our analysis.

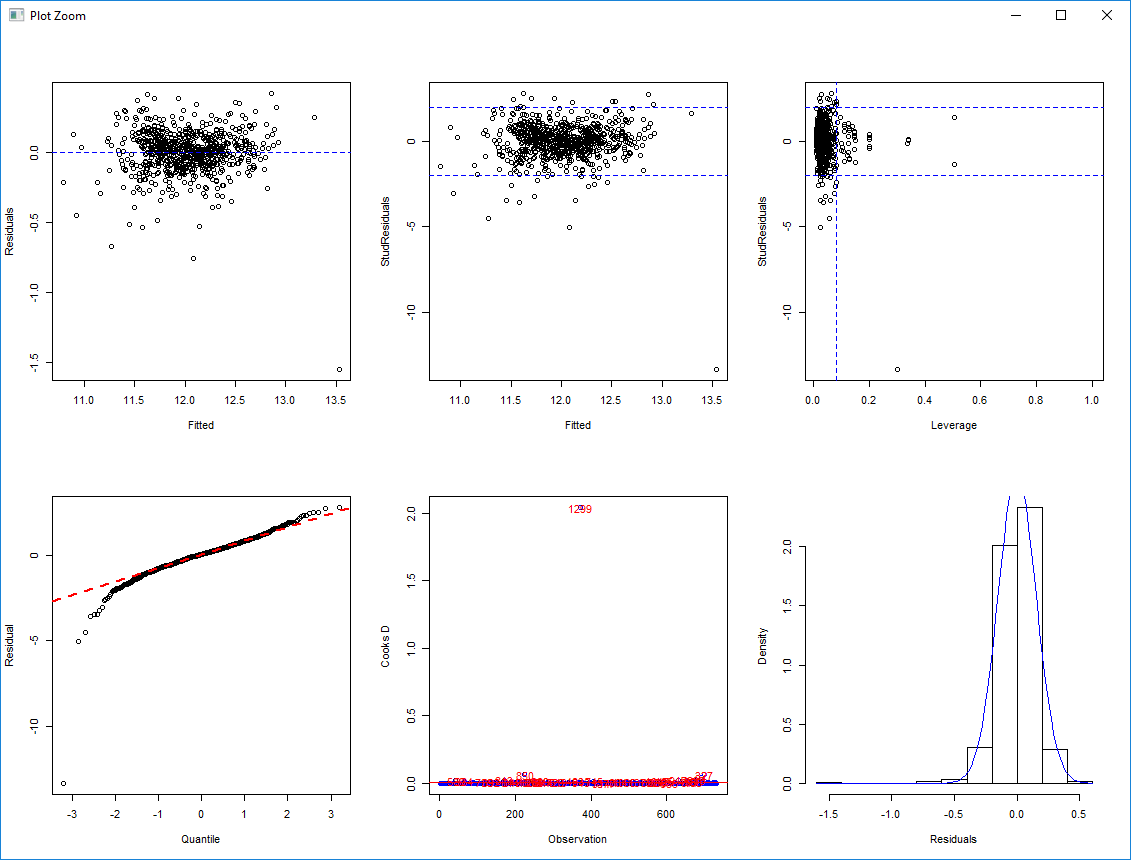


**[CHART 4.b.1]**

Checking the VIF’s resulted in the values all being between 1.06 and 6.9 which indicated no multicollinearity. The residual plot indicates that except for a few outliers (which we decided to keep in the initial stage of analyzing this model) there is constant variance. The data is a bit non-normally distributed with a small tail at the beginning. This is fine because (as seen in the histogram) there few outliers and plenty of observations. We will assume independence, although due to the nature of house pricing this is in suspect. As mentioned earlier there are a few minor outliers. However because these outliers are minor and out of 1480 total points, their influence is low so we included them.

**Custom**

The custom model was designed to be simple to understand and interpret. For this model we took the variables that were the most significantly correlated with L\_SalePrice and also a few sensible ones. The model consists of OverallQual, TotalSqFeet, GarageCars, AllFullBath, Neighborhood, and OverallCond. The initial model had a severely outlying value. After analyzing the row there is no error in that recording so we proceeded to do the test without it to see the result. Without the point normality and the adjusted r-squared appeared worse but Cook’s D and Leverage was be better. Because this is just a single point out of 1480, its influence is low so we continued our analysis with the point included.

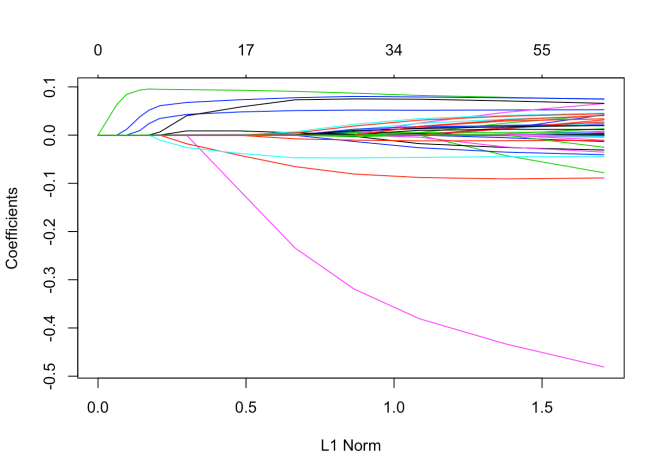


**[CHART 4.b.2]**

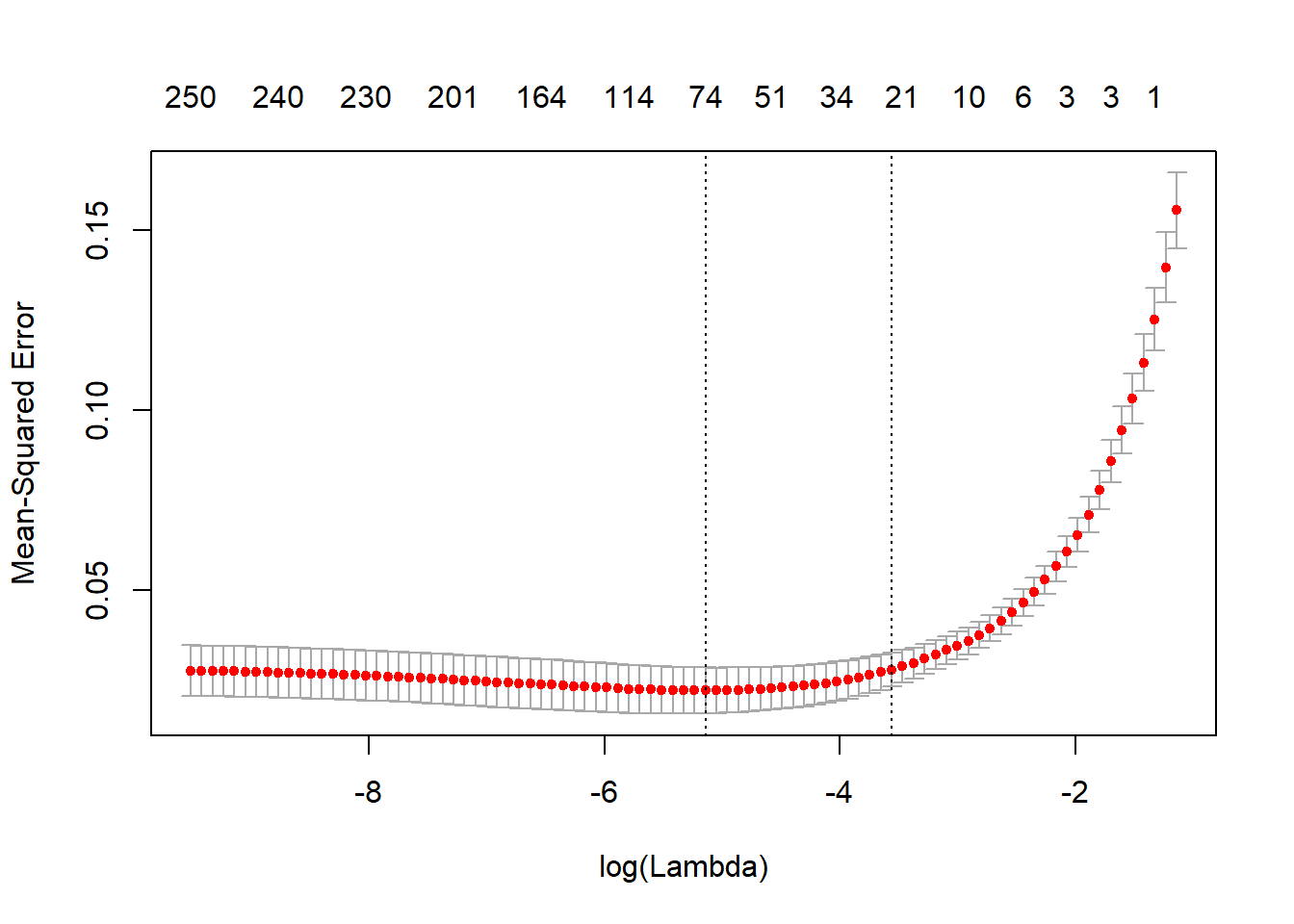
The VIF’s are very low and nearly identical with their values ranging from 1.06 to 2.0. This indicates that there is no multicollinearity present. The residual plot indicates that except for a few outliers (which we decided to keep in the initial stage of analyzing this model) there is constant variance. The data is nearly normally distributed, with a bit of a tail at the beginning. This is fine because (as seen in the histogram) there few outliers. We will assume independence, although due to the nature of house pricing this is in suspect. As mentioned earlier there is one major and a few minor outliers. However these points are out of 1480 total, their influence is low so we continued our analysis with them included.

**Lasso**

Because we are dealing with high-dimensional data set, we have used the shrinkage model Lasso to obtain the model with the least effect of predictors variance and collinearity. Lasso penalizes the model coefficient estimates using Lambda tuning parameter to reduce their variance which results in better model fitting. Lasso's GLM-NET performs 10-fold CV to determine an optimal penalty parameter. The coefficients are easy to extract and making predictions are straight forward.

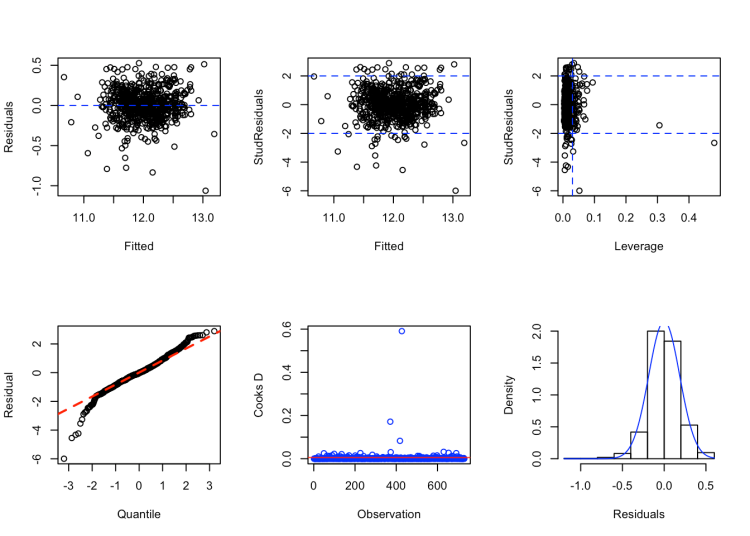


**[CHART 4.b.3]**



**[CHART 4.b.4]**

We can see from the coefficient plot **[CHART 4.b.3]** that depending on the choice of tuning parameter, some of the coefficients will be exactly equal to zero. We performed cross validation and examined the test error. The number of non-zero coefficients that the Lasso Regression picked is 56 **[CHART 4.b.4]**. From this list of coefficients we selected the top ten variables that were the most significantly correlated with L\_SalePrice and ran a LM model on those. The model consists of MSSubClass, LotArea, OverallQual, OverallCond, YearBuilt, YearRemodAdd, X2ndFlrSF, BedroomAbvGr, KitchenAbvGr, GarageCars.

**[CHART 4.b.5]**

On this model the VIF’s are very low and nearly identical (between 1.06 and 2.0), which indicates that there is no multicollinearity. The residual plot indicates that except for a few outliers (which we decided to keep in the initial stage of analyzing this model) there is constant variance. The data is nearly normally distributed, with a bit of a tail at the beginning. This is fine because (as seen in the histogram) there are few outliers. We will assume independence, although due to the nature of house pricing this is in suspect. As mentioned earlier there is one major and one minor outlier. However this is out of 1480 total rows which makes its influence low. Because of this so we continued our analysis with the points included.

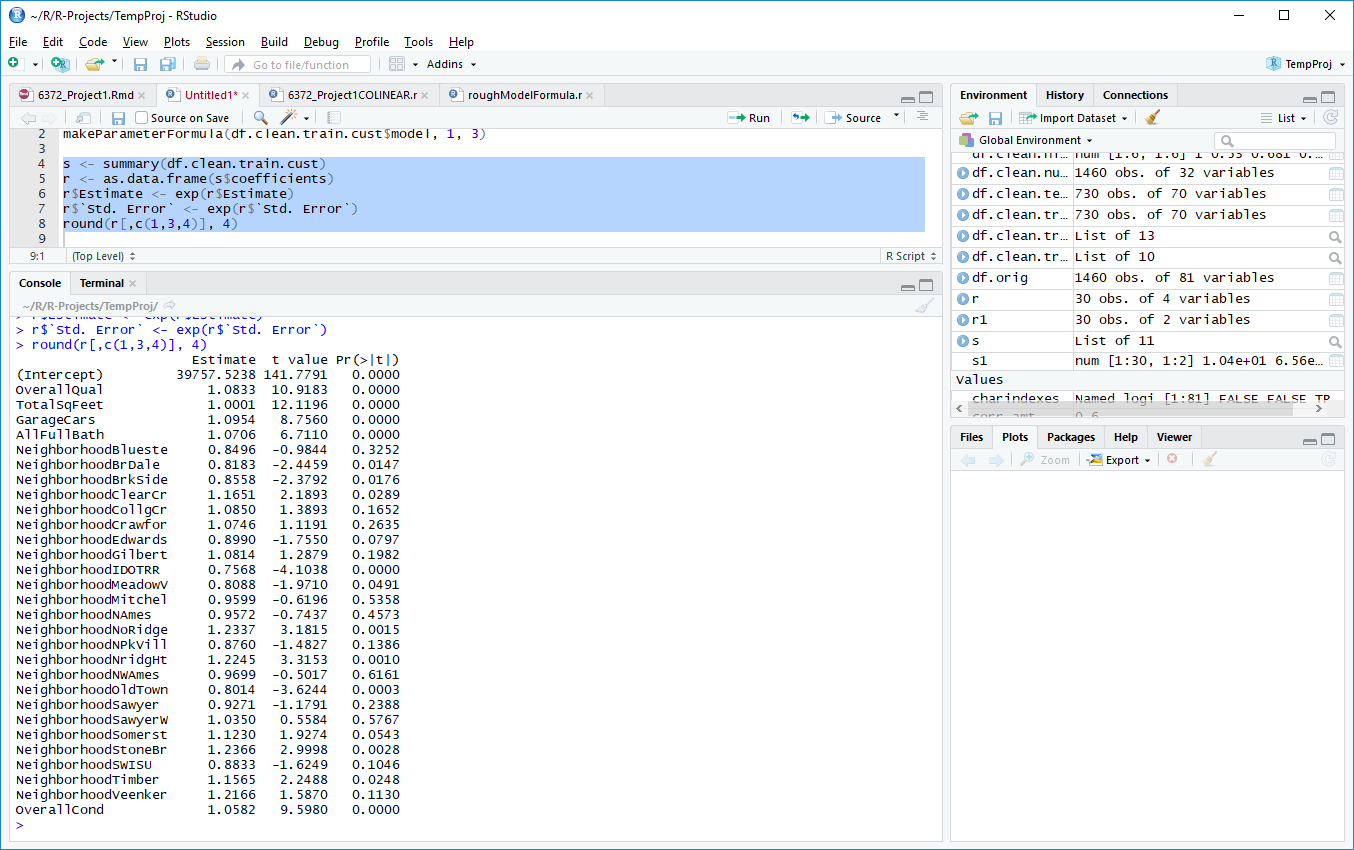
1. **Comparing Models**

|  |  |  |
| --- | --- | --- |
|  | **RMSE** | **Adj-R2** |
| **Stepwise** | 36687 | 0.813 |
| **Lasso** | 40740 | 0.76 |
| **Custom** | 31489 | 0.86 |

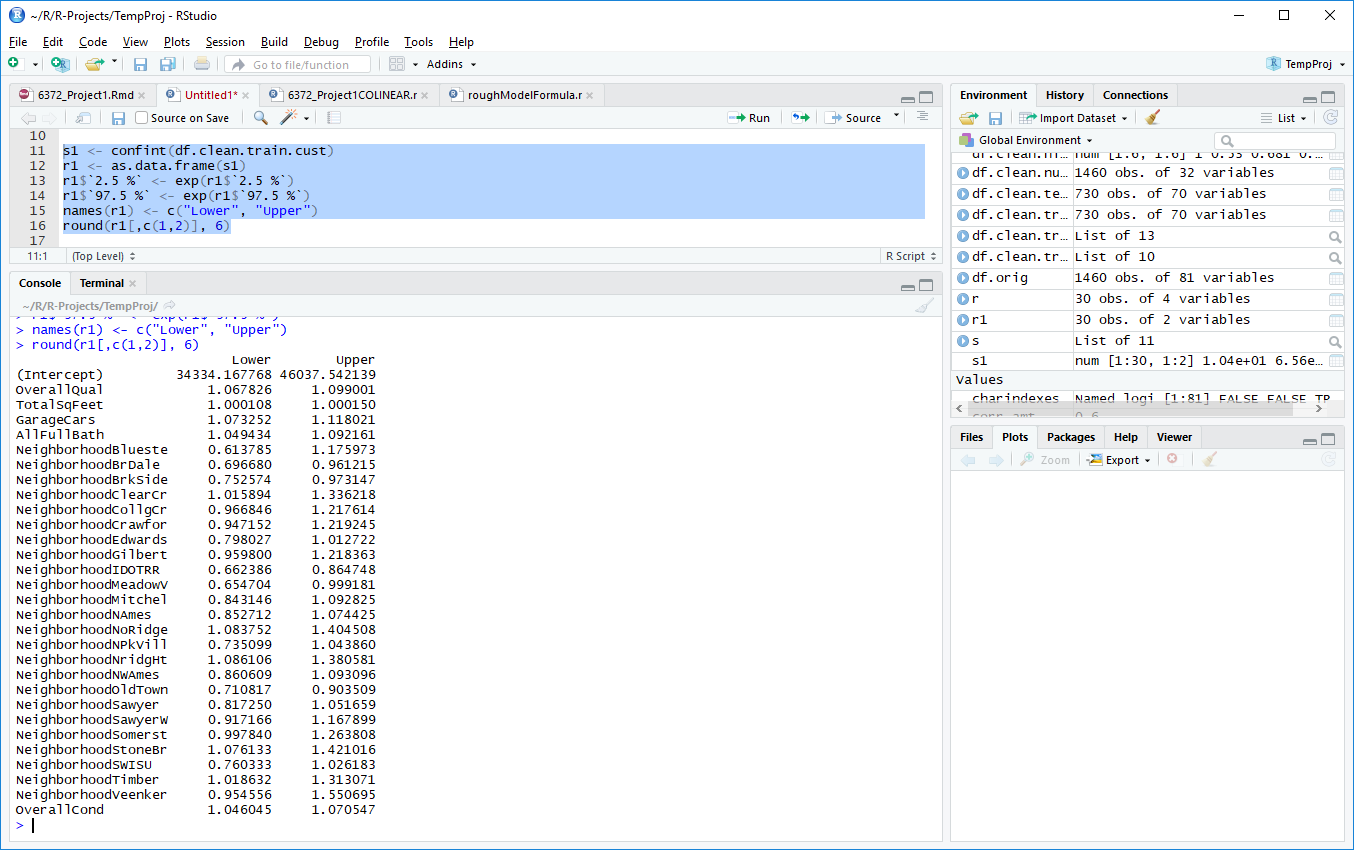
**[CHART 4.c.1]**

For each of the models we generated an adjusted r-squared (Adj-R2) and the root mean squared error (RMSE) values. On the left **[CHART 4.c.1]** are those results. As you can see the Custom model performed best (RMSE: 31489). This is most likely due to our domain knowledge.

1. **Parameter Interpretation**



**[CHART 4.d.1]**



**[CHART 4.d.2]**

As seen in **[CHART 4.c.1]** the most accurate model is the Custom model. The Custom model is also the simplest and therefore it is also the one that is the easiest to interpret. The model consists of OverallQual, TotalSqFeet, GarageCars, AllFullBath, Neighborhood, and OverallCond with the response variable being log(SalePrice). The results from this model **[CHARTS 4.d.1 and 4.d.2]** have been converted back from the log form and can be interpreted as follows. A one unit increase in TotalSqFeet is associated with a multiplicative median change of 1.01 **[CHART 4.d.1]** in the SalePrice (in essence an increase) with everything else held constant (p-value < 0.0001). The 95% confidence interval **[CHART 4.d.2]** for this median multiplicative change is between [1.0108, 1.015]. All of the other numerical parameters can be interpreted accordingly.

Neighborhood is the only categorical parameter. The intercept value is in reference to Neighborhood ‘Blmngtn’. Being in the Neighborhood ‘IDOTRR’ is associated with a **[CHART 4.d.1]** median multiplicative change of 0.76 in SalePrice (in essence a decrease) relative to Neighborhood ‘Blmngtn’ (p-value < 0.0001). The 95% confidence interval **[CHART 4.d.2]** for this median multiplicative change is between [0.66, 0.84]. All of the other Neighborhoods can be interpreted accordingly.

1. **Conclusion**

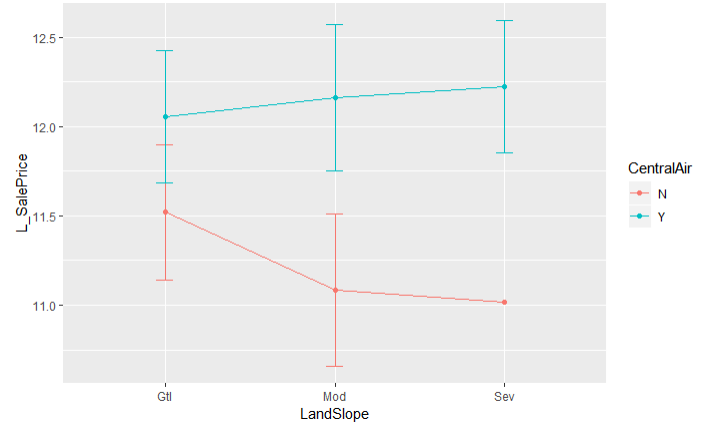
This report elaborated on the details of how the house prices were predicted. Missing data was first imputed, a combination of Stepwise, Lasso and Custom models were used. Of the three models, custom model performed better with an RMSE of 31489. Often, ensemble technique outperforms the single best model. However in this instance, a creating a model using domain knowledge enabled our Custom model to be a better predict the SalePrice.

To improve upon this project and to give the automatic selection algorithms a chance to do better than a custom model, we would need to do a bit more on front end of the data analysis. There are many aspects of homes (such as having a basement) that are not required for a home to be worth more. If we had dealt with that zero inflation before fitting those parameters in our model it is likely that there would have been a few more highly significant variables. This could have allowed the automatic selection techniques model’s to beat out our Custom model.

1. **Objective 2**
2. **Goal of Two Way ANOVA**

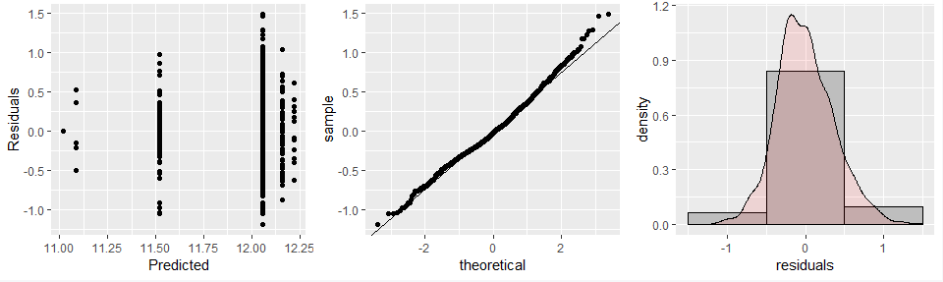
For this objective we will pretend LandSlope and CentralAir are the only two categorical variables and perform a Two Way ANOVA analysis. We will compare the mean differences between the combination of groups to understand if there is an interaction between the two independent categorical variables.

1. **Analysis of Two Way ANOVA**



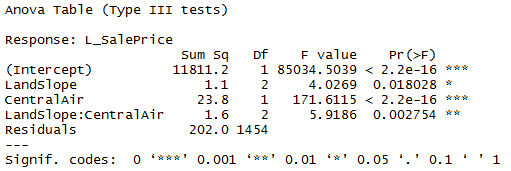
**[CHART 5.b.1]**

First of all we checked if the two independent variables are interacting by using the mean profile plot **[CHART 5.b.1]** with standard deviation as the error bar. From the profile plot it is obvious that the lines are not parallel which means the model is non-additive. Therefore there is likely an interaction between LandSlope and CentralAir.



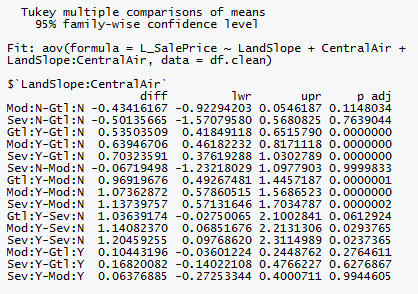
**[CHART 5.b.2]**

Next we created the two way ANOVA model and checked the diagnostic plots **[CHART 5.b.2]** to see if there were any violation of the assumptions. The residual plot shows the standard deviations are roughly the same. The qq-plot and histogram do not violate the assumption of normally distribution of the residual. Thus, the residual diagnostics do not provide any concern about the assumptions of a two way ANOVA analysis.

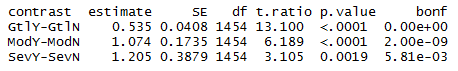


**[CHART 5.b.3]**

Examining the type-III sums of squares F table **[CHART 5.b.3]** shows there is a statistically significant interaction between LandSlope and CentralAir (p-value = 0.0028).



**[CHART 5.b.4]**



**[CHART 5.b.5]**

Thirdly we adjust for multiple tests using Tukey’s procedure. The Tukey method can be used to compare all combination of groups [CHART 5.b.4]. However, including all possible combinations of comparisons is too much and of little interest to the actual study. In order to compare a specific group of interest, we can use the Bonforroni adjustment. For example, we are interested to see if the SalePrice will be different with and without CentralAir in each LandSlope level. In this scenario, we only need to compare groups of "GtlY-GtlN", "ModY-ModN", "SevY-SevN". The result [CHART 5.b.5] shows that influence of CentralAir to SalePrice is significant in each specific LandSlope level, with all p-values < 0.05.

1. **Conclusion/Discussion**

The two way ANOVA analysis shows there is a statistically significant interaction between the two independent categorical variables CentralAir and LandSlope. The mean profile plot and the type III sum of squares F table can both exam the existing of interaction. The residual diagnostics do not provide any concern about the assumptions of a two way ANOVA analysis. Tukey adjustment can be used to perform comparison of all the combination of groups. However, we do not need to To check if the SalePrice will be different with and without CentralAir in each LandSlope level, we only need to compare groups of "GtlY-GtlN","ModY-ModN","SevY-SevN". Bonforroni adjustment can serve the purpose to compare each specific groups. The result shows that influence of CentralAir to SalePrice is significant in each specific LandSlope level.

1. **Appendix**

Contains code and extra charts