**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**

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 columns that had factors that had problems with their levels (EX: Utilities had two levels, 1459 of the rows were of one level and the final row was of the other level). Removal of columns that had too much missing data and the consolidation of redundant columns were next. We examined the correlation plots and removed a few highly correlated ones that were describing similar attributes. At this point the data was clean so we moved on to the model building analysis.

*Transformations, Pairs plots, correlation plots, removal of columns, adding baths, put in charts here.*

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

…and the overall approach to solve it

1. **Build and Fit Models**

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

**Stepwise**

Do initial plotting

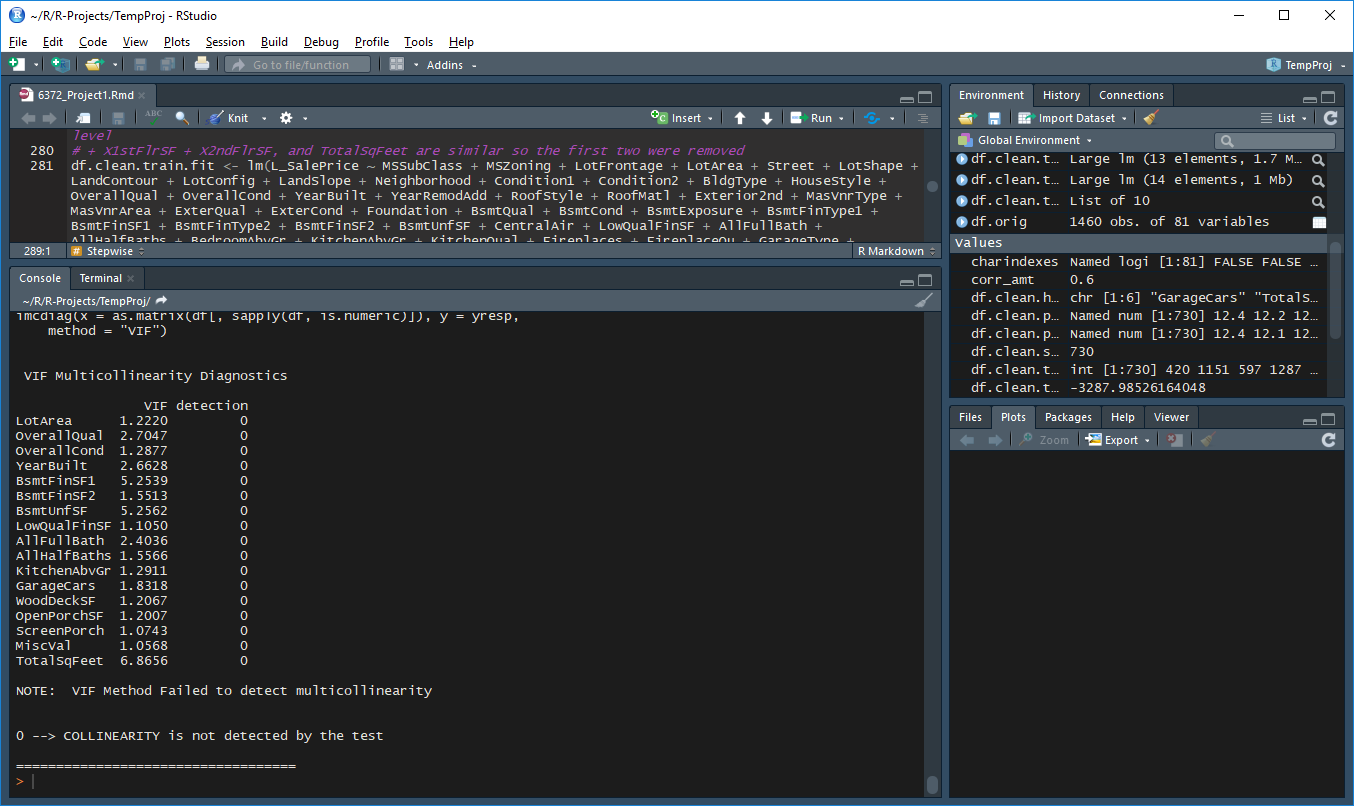
Check plots

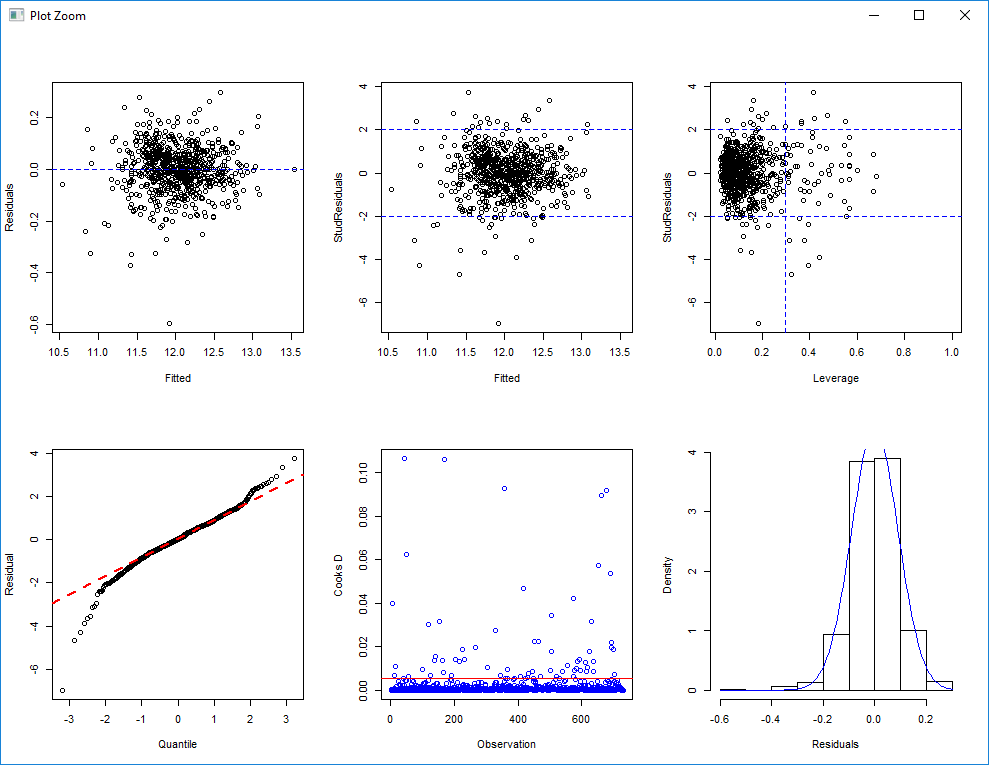
Outliers are not recording error, we do not wish to restrict value, so we run with/without

(due to low number of leverage points, high number of observations in the dataset, we leave the points in the dataset)

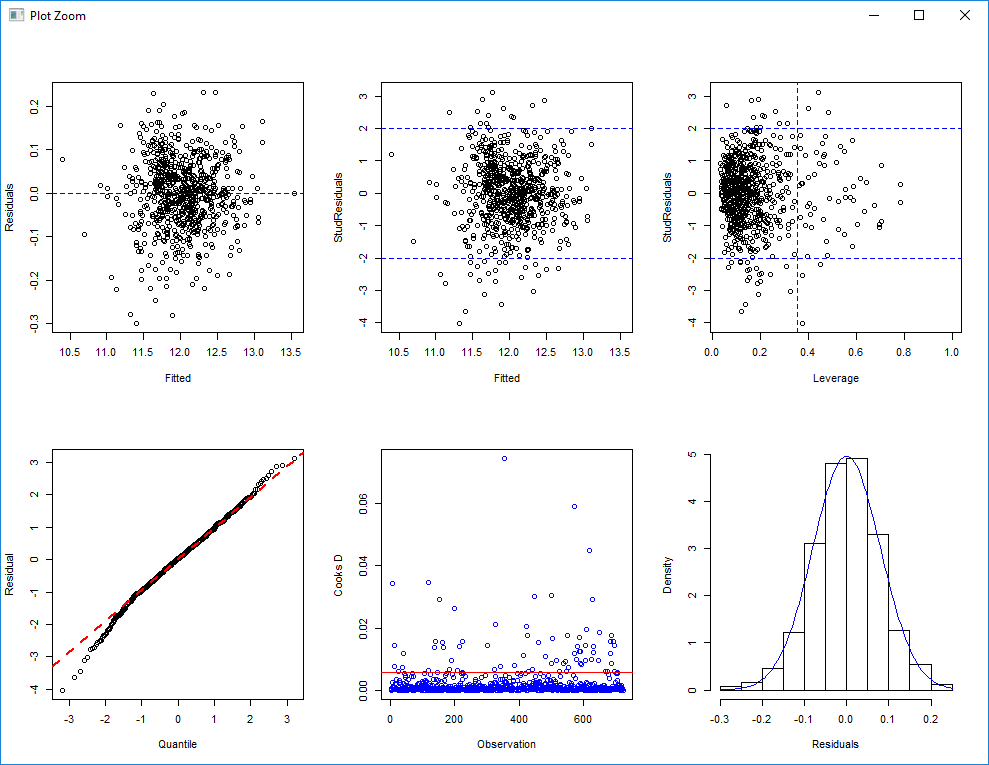
VIF’s

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. In the end, the stepwise model selected 35 of the variables and 1 interaction. Those variables are: MSZoning, LotArea, Street, LandContour, LandSlope, Neighborhood, Condition1, BldgType, OverallQual, OverallCond, YearBuilt, RoofMatl, ExterQual, Foundation, BsmtQual, BsmtExposure, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, CentralAir, LowQualFinSF, AllFullBath, AllHalfBaths, KitchenAbvGr, KitchenQual, FireplaceQu, GarageCars, GarageQual, GarageCond, WoodDeckSF, OpenPorchSF, ScreenPorch, MiscVal, SaleCondition, TotalSqFeet, and one the interaction term GarageCars\*TotalSqFeet.



RMSE: 35743.03, Adj-R2: 0.82

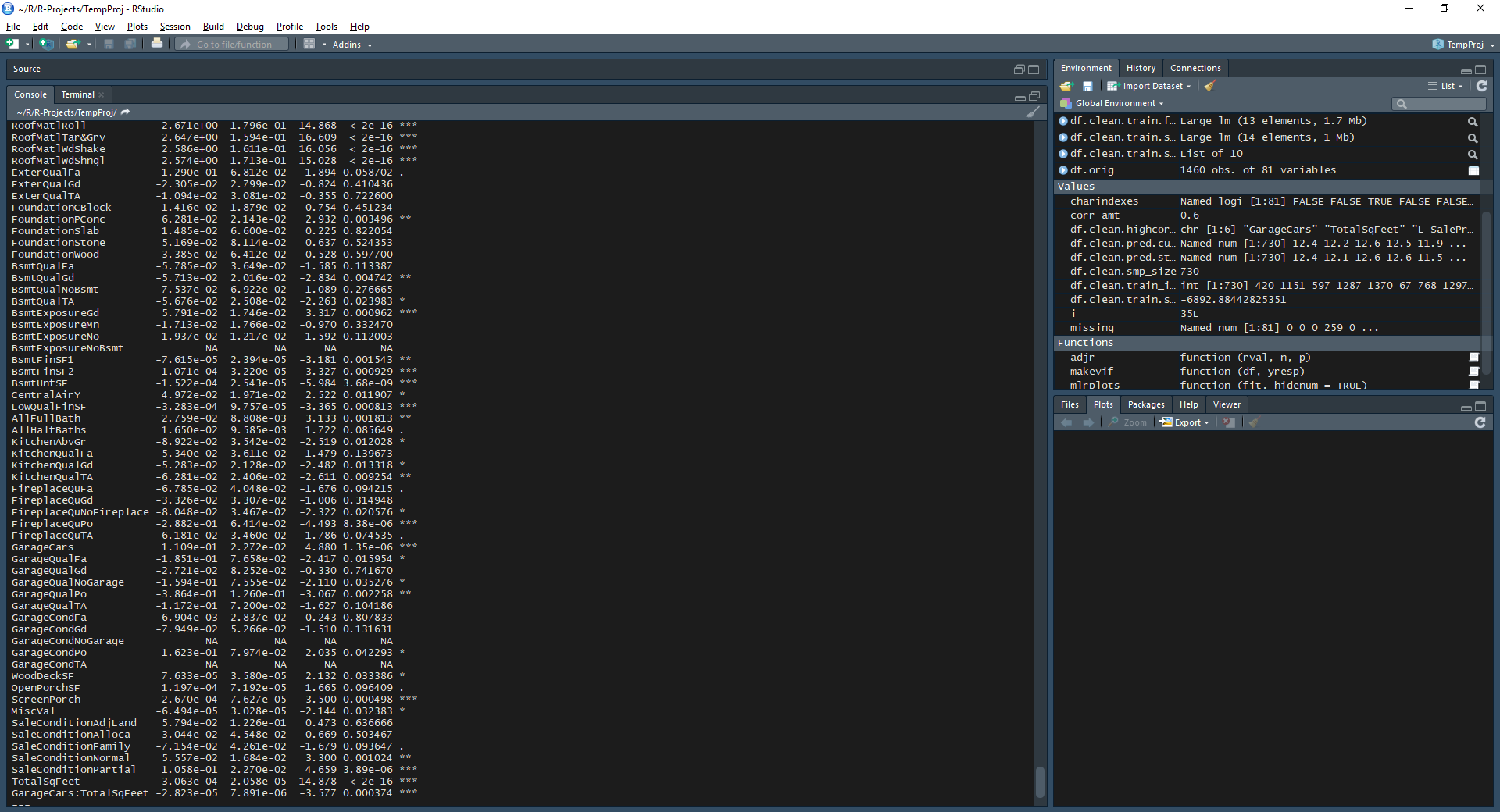
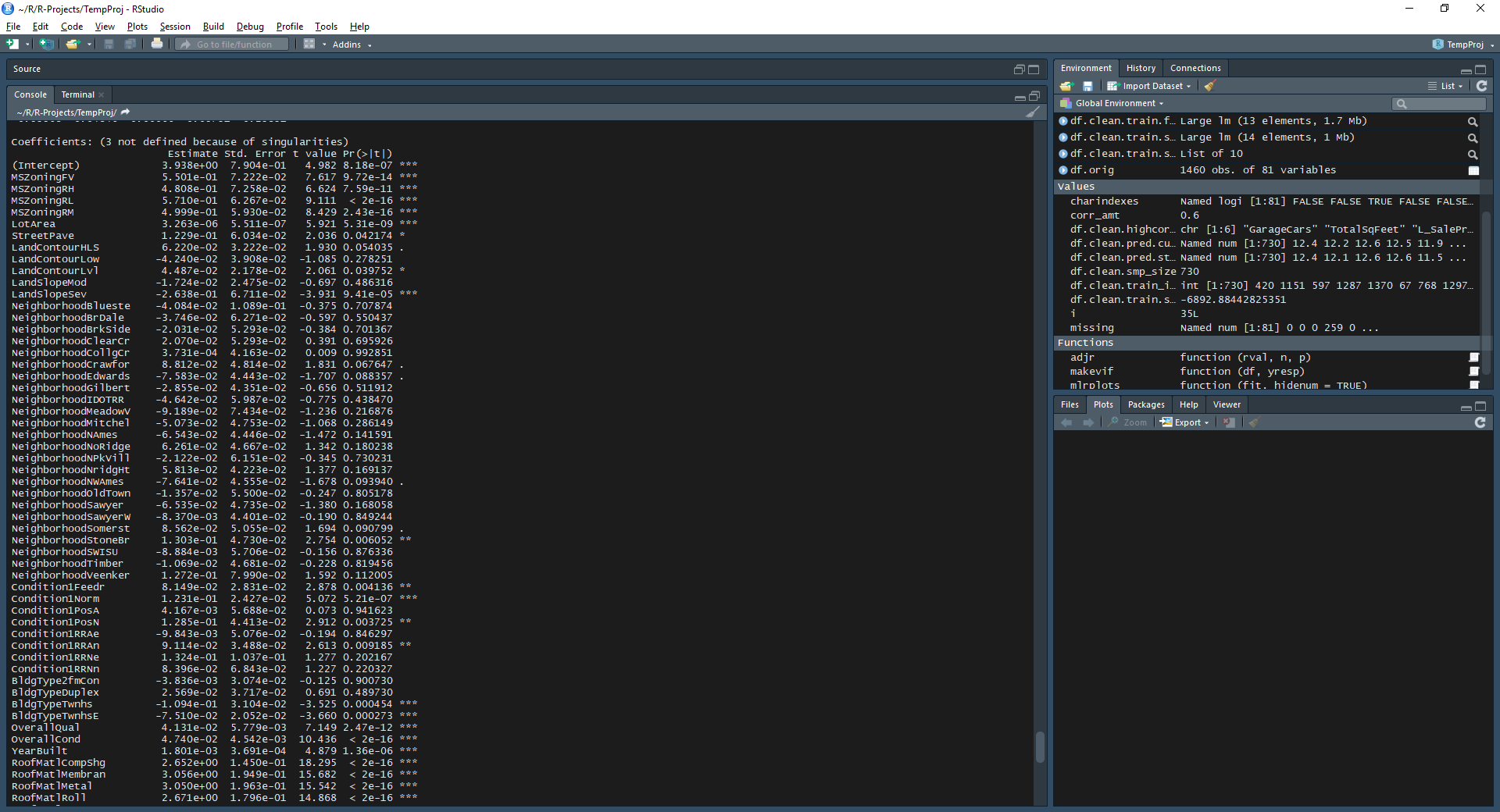
The model seems to have a problem with normality with a few outliers. Here is the result of removing the five offending points.



RMSE: 36421.56, Adj-R2: 0.81

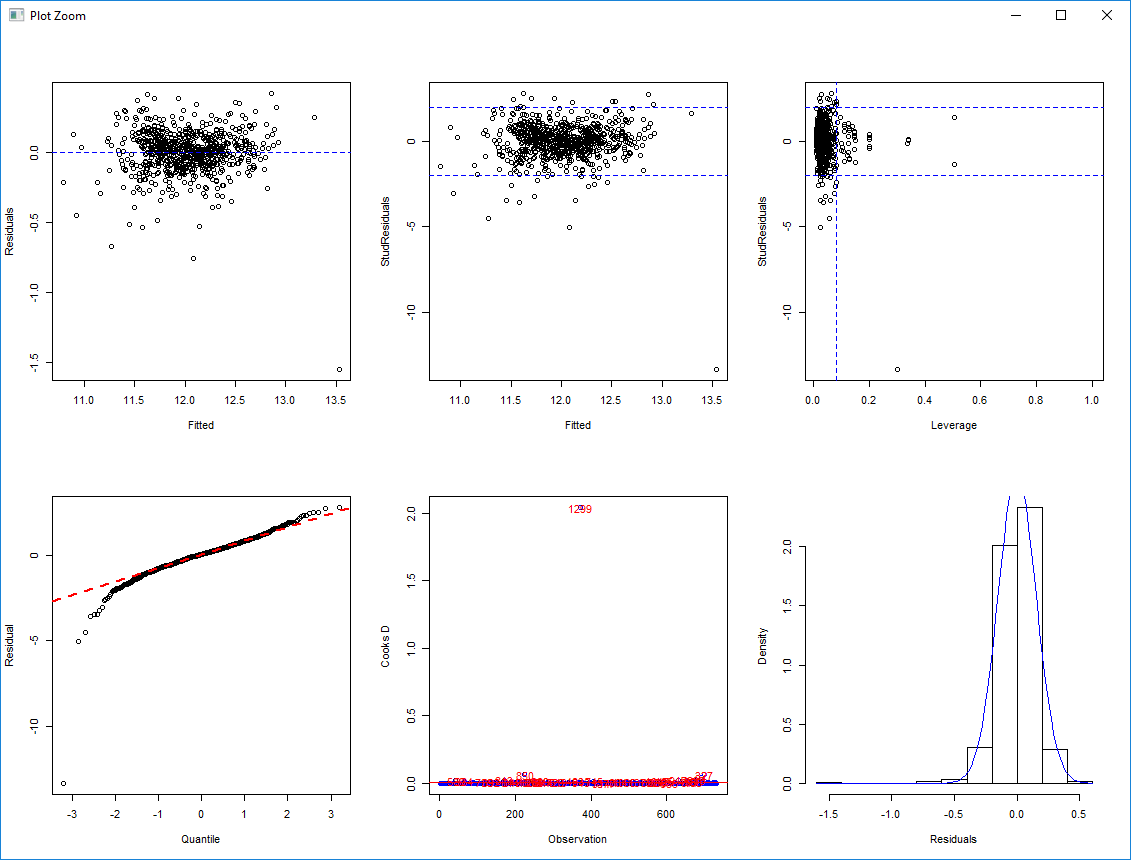
Without the outlying points the model is much more normal and the residual plot is a little better, however the RMSE is worse. Therefore we will continue with the points included.

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 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 are a few minor outliers. However because these outliers are minor and out of 1480 total, their influence is low so we continued our analysis with the point included.

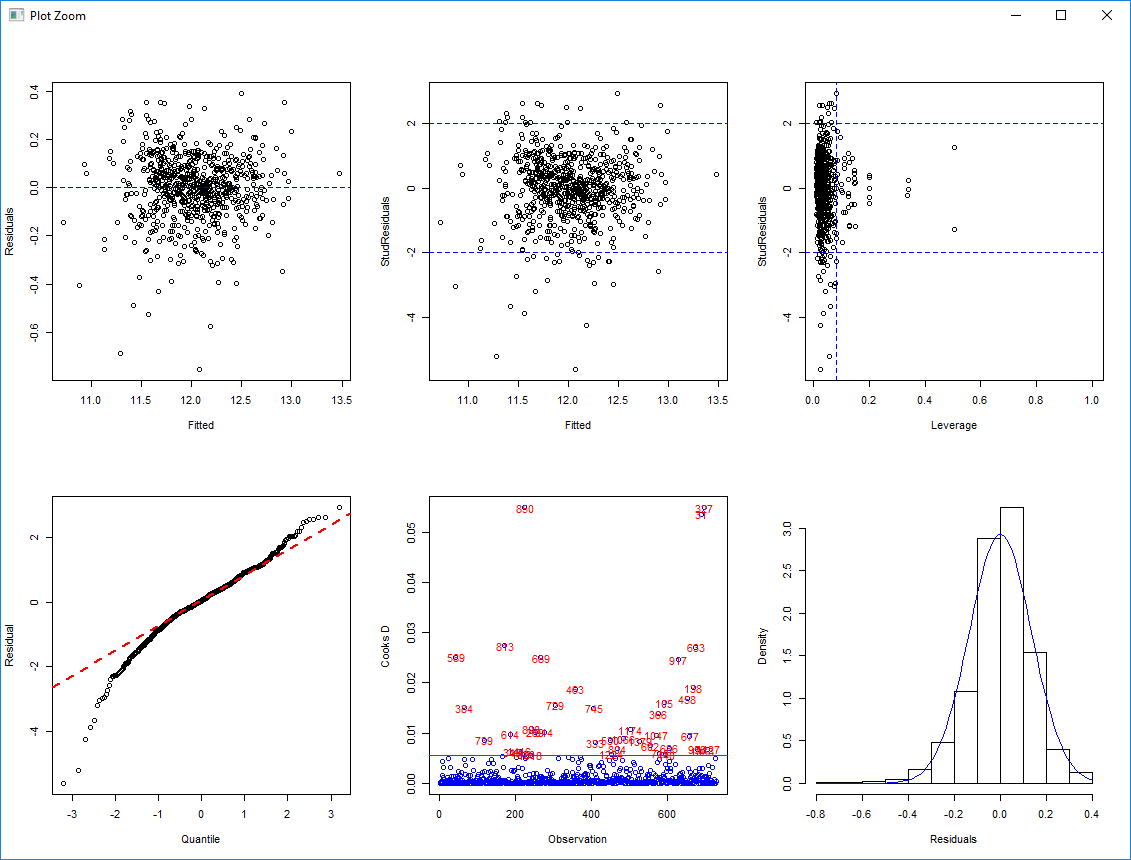


**Custom**

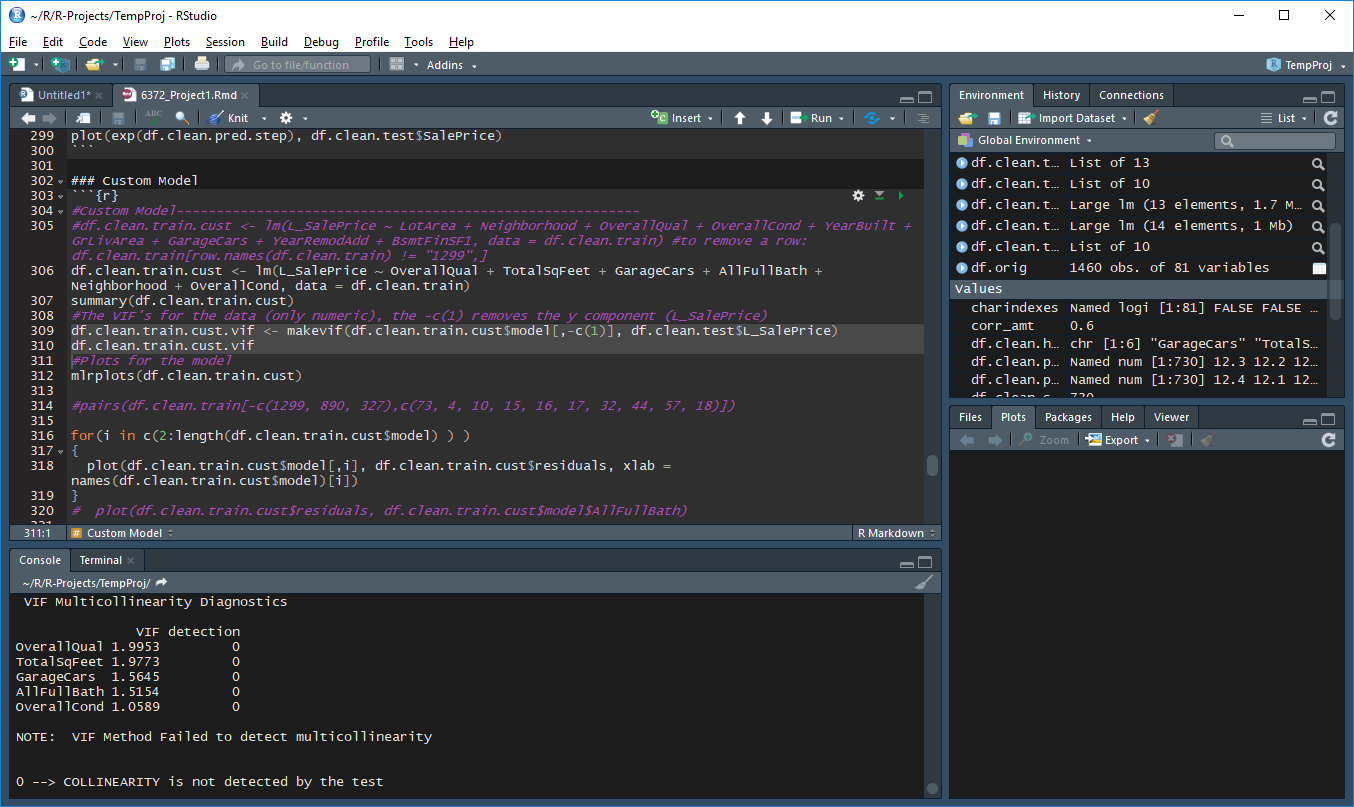
The custom model was to be one that is simple to understand and interpret. For this model we took the variables that were the most significantly correlated with L\_SalePrice and also a few sensible ones. The model consists of OverallQual, TotalSqFeet, GarageCars, AllFullBath, Neighborhood, and OverallCond.

RMSE: 31489.01, Adj-R2: 0.86

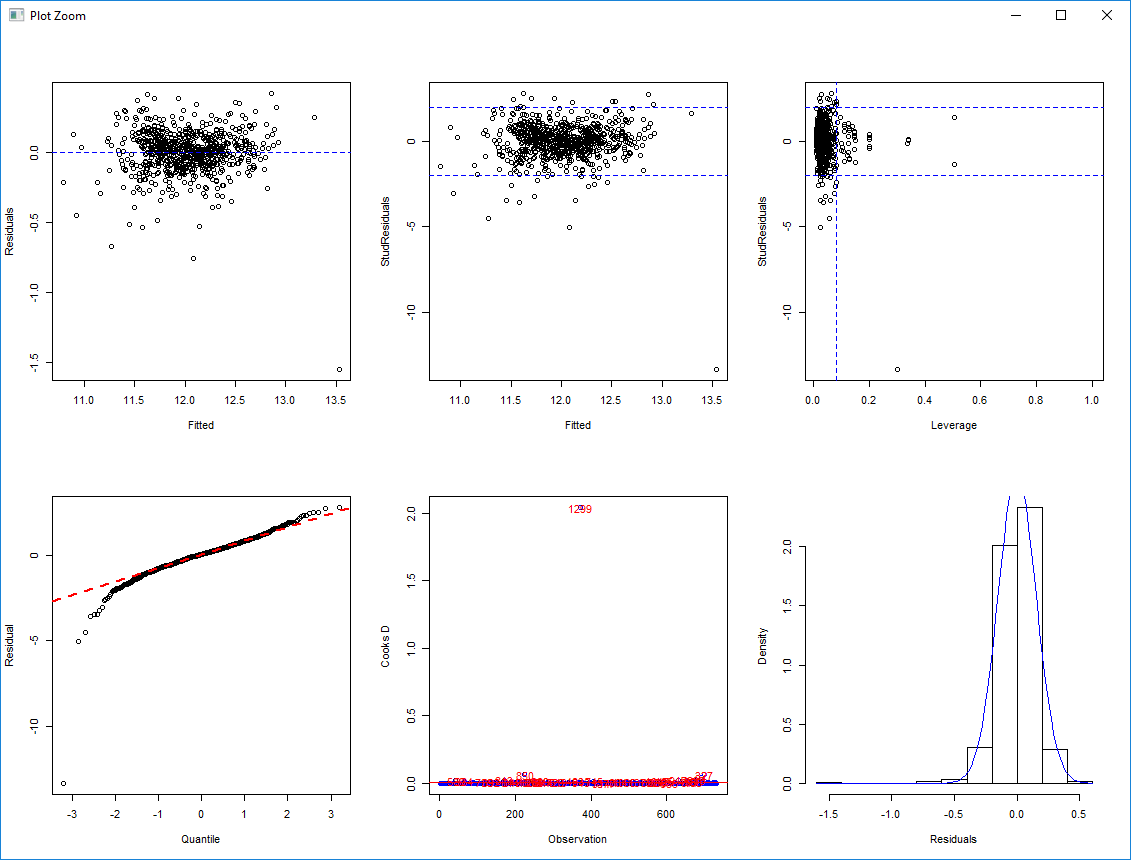
The initial model has a severely outlying value. After analyzing the row there is no error in that recording so we proceed to do the test without it to see the result.

RMSE: 32573.45, Adj-R2: 0.85

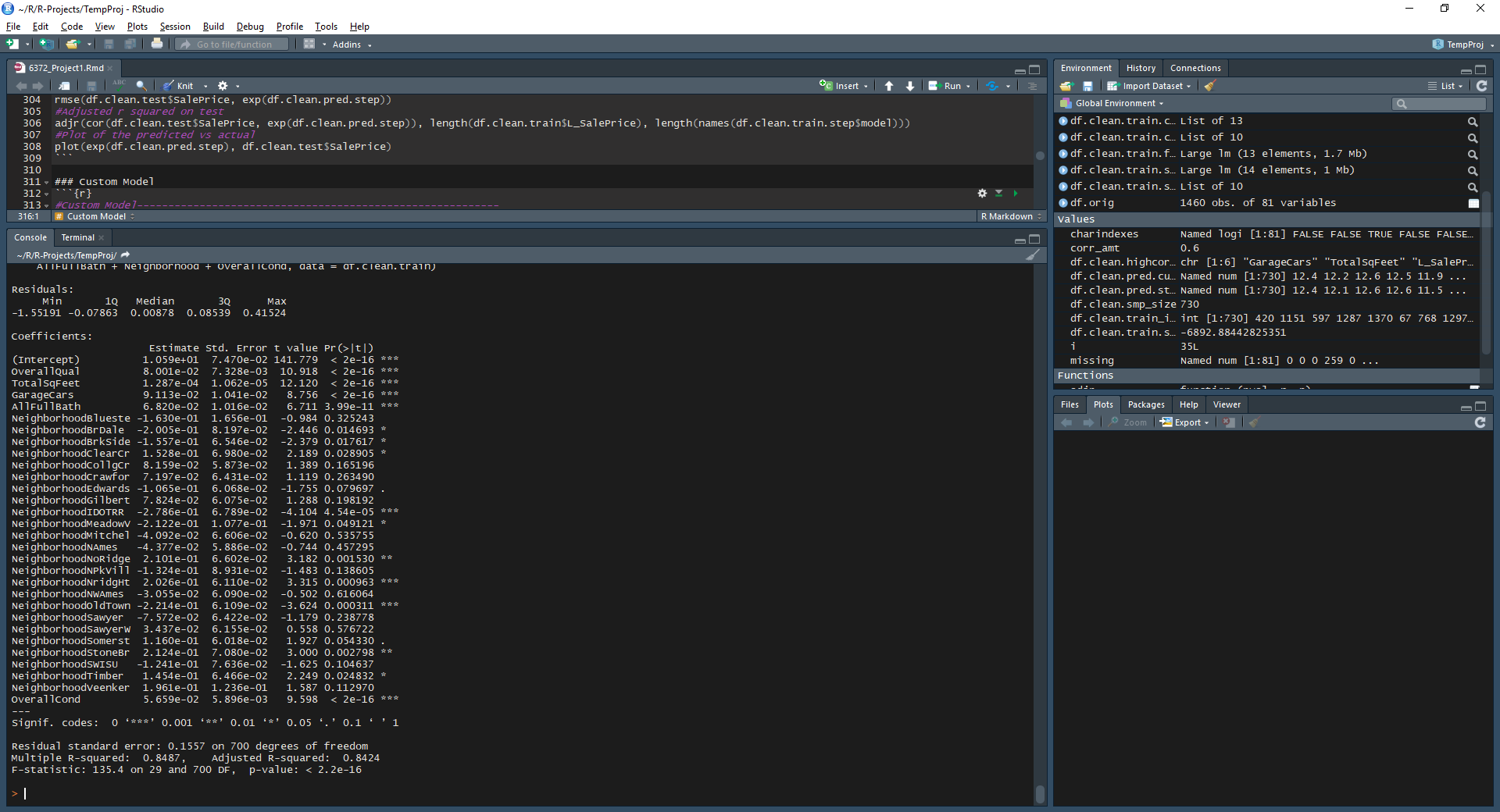
Without the point normality appears worse but Cook’s D and Leverage appear to be better. The results from the test set appear worse without the point. Because this is just one point out of 1480, its influence is low so we will continue our analysis with the point included.



As seen the VIF’s are very low and nearly the same as each other, which indicates that there is no multicollinearity.



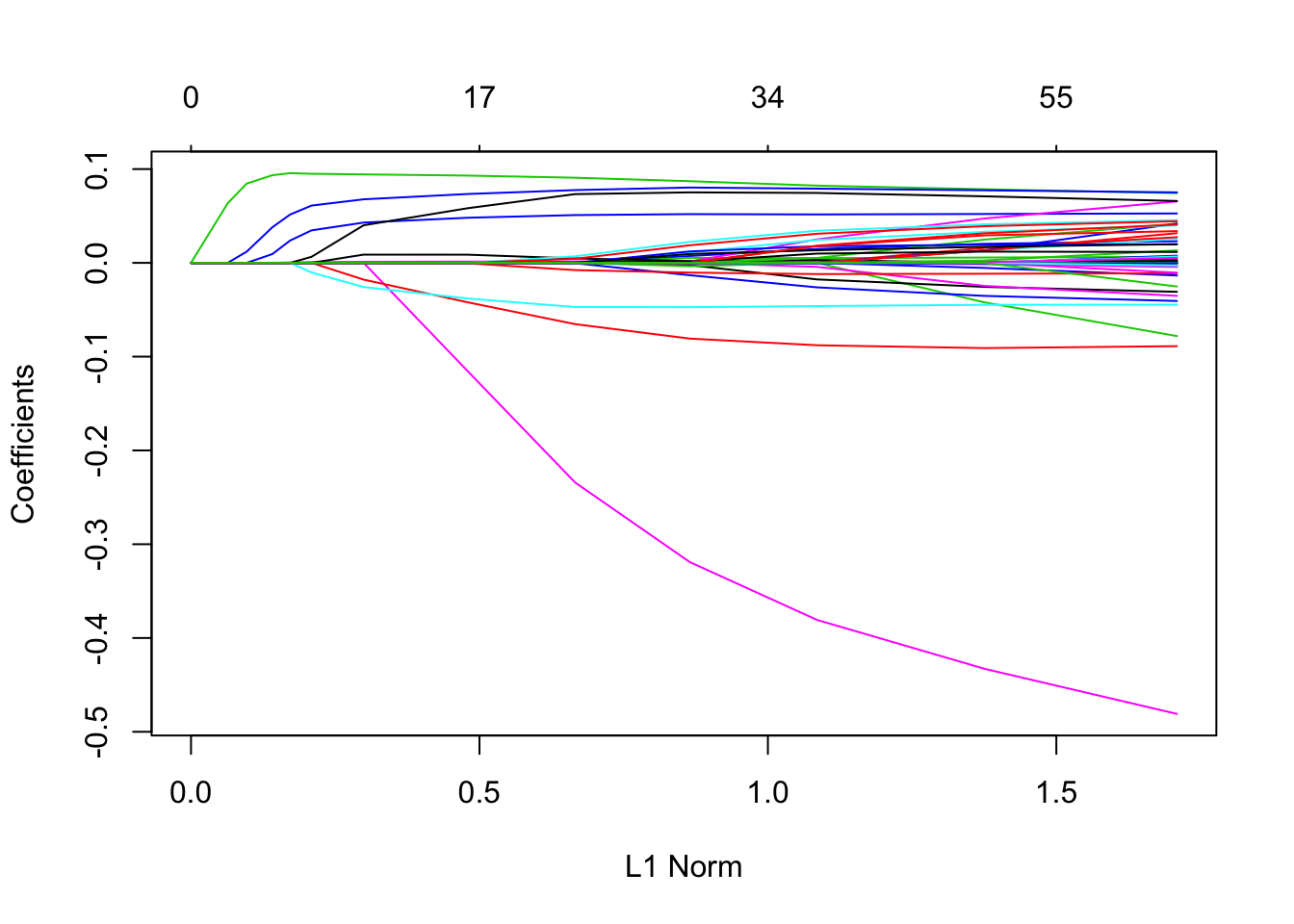
The residual plot indicates that except for a few outliers (which we decided to keep in the initial stage of analyzing this model) there is constant variance. The data is nearly normally distributed, with a bit of a tail at the beginning. This is fine because (as seen in the histogram) there few outliers. We will assume independence, although due to the nature of house pricing this is in suspect. As mentioned earlier, there is one major outlier. However this is just one point out of 1480, its influence is low so we continued our analysis with the point included.



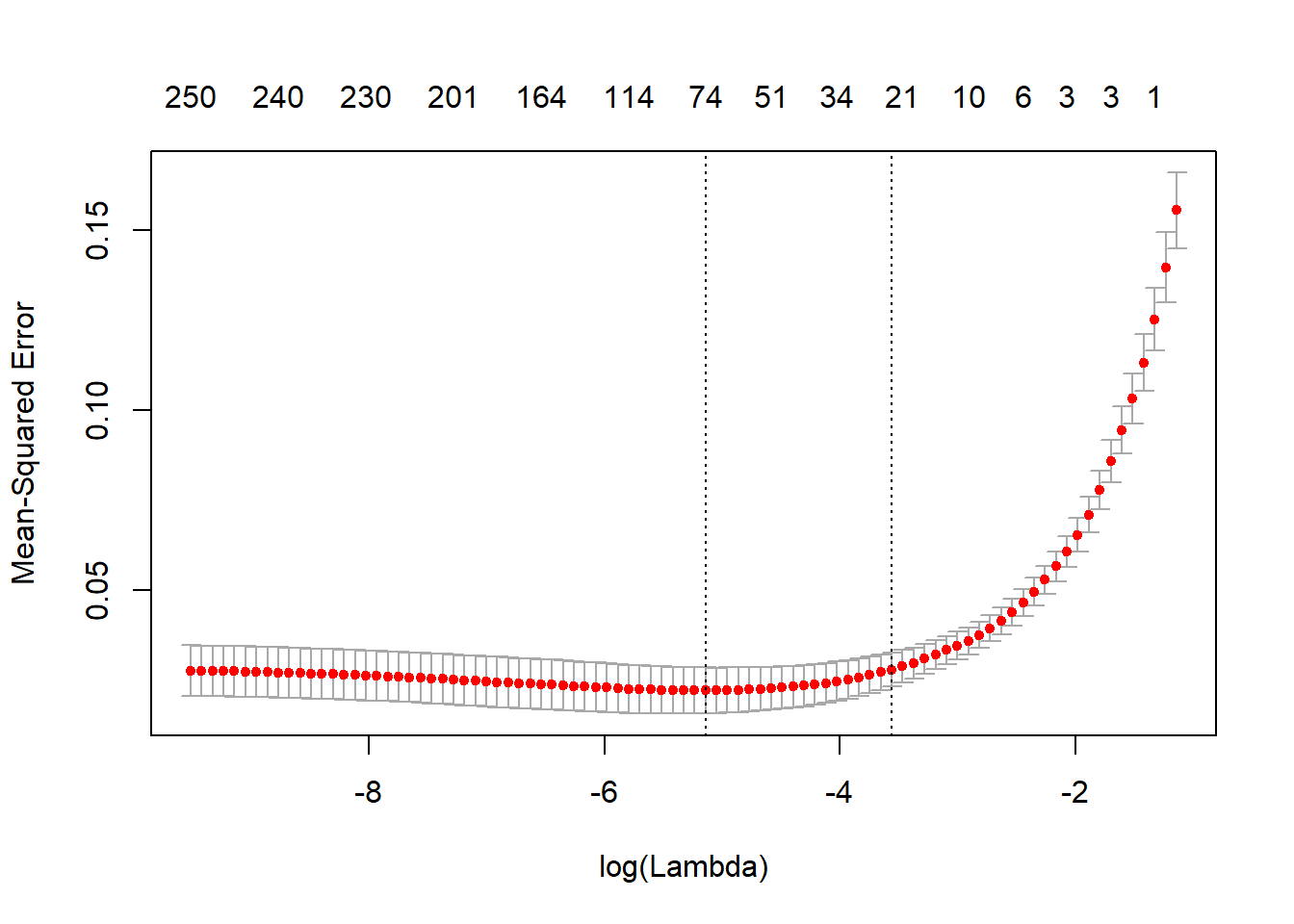
**Lasso**

Since 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 colinearity. Lasso penalizes the model coefficient estimates, using Lambda tuning parameter, to reduce their variance which result in a better model fitting.

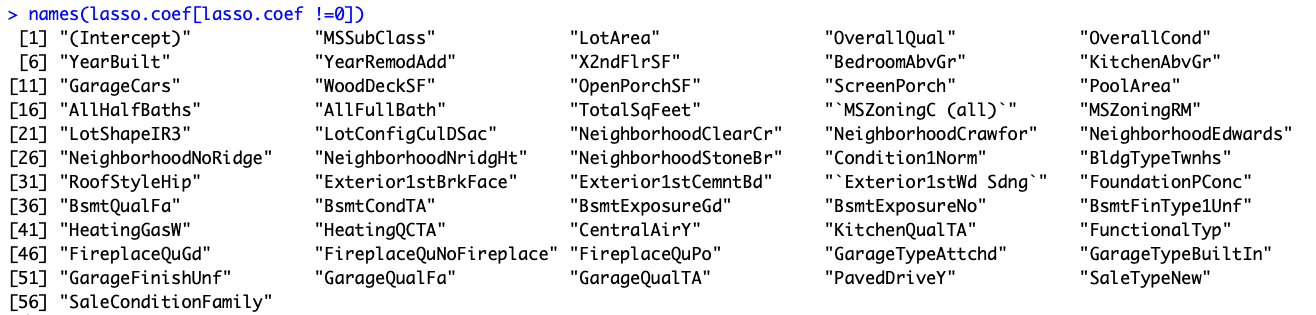
Lasso's GLM-NET performs 10-fold CV to determine an optimal penalty parameter. The coefficients are easy to extract and making predictions are straight forward.



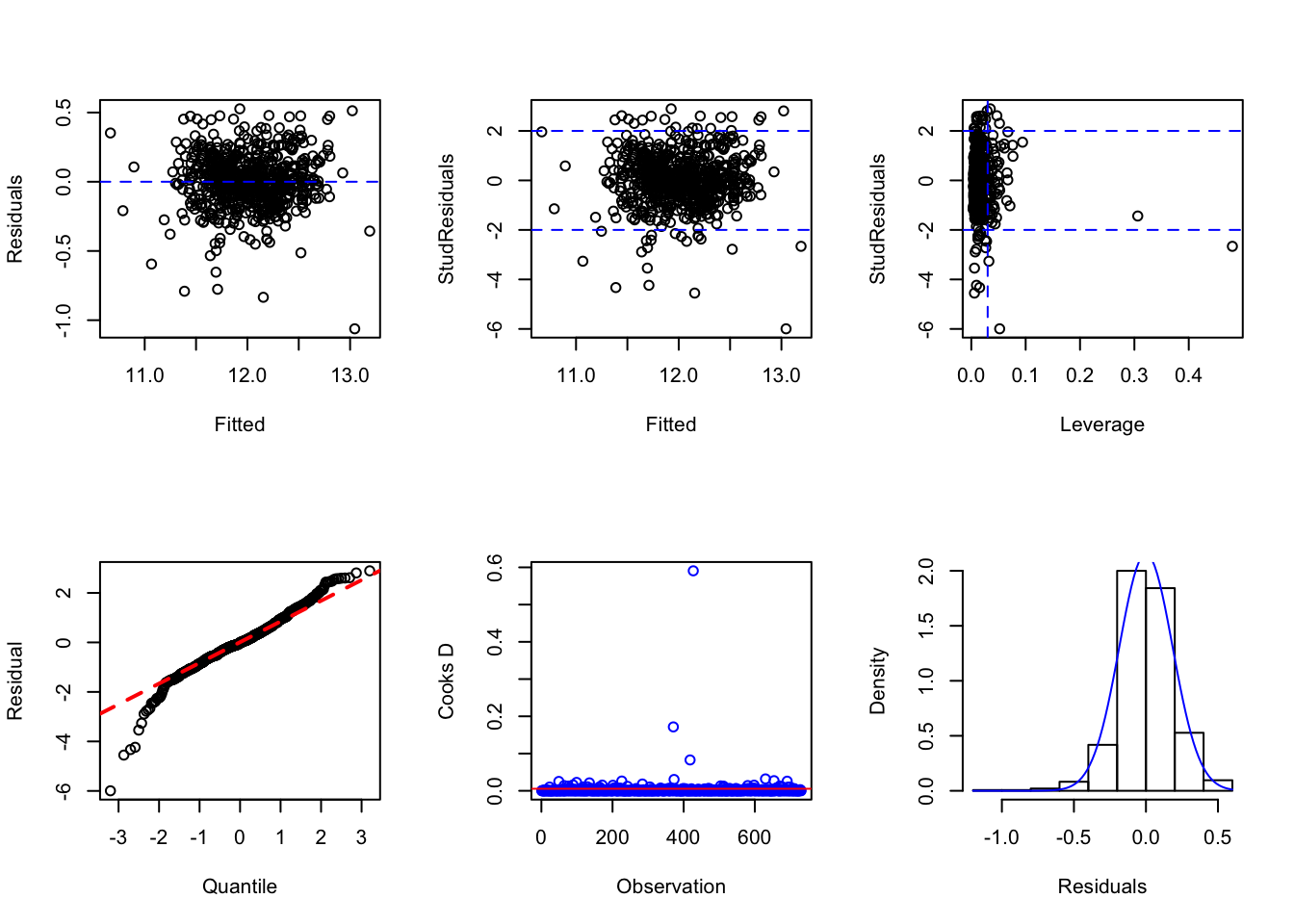
We can see from the coefficient plot that depending on the choice of tuning parameter, some of the coefficients will be exactly equal to zero. Let’s perform cross validation and see the test error.



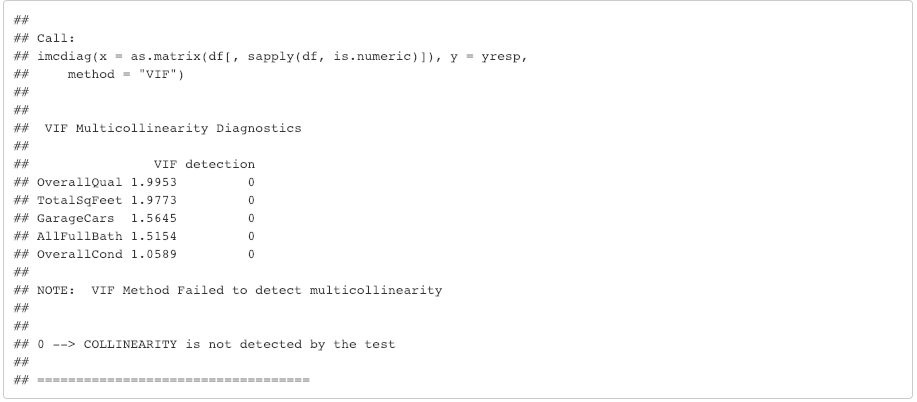
The RMSE resulted from the Lasso Regression model is 0.1423346. Number of coefficients that the Lasso Regression picked (not zero) is 56.



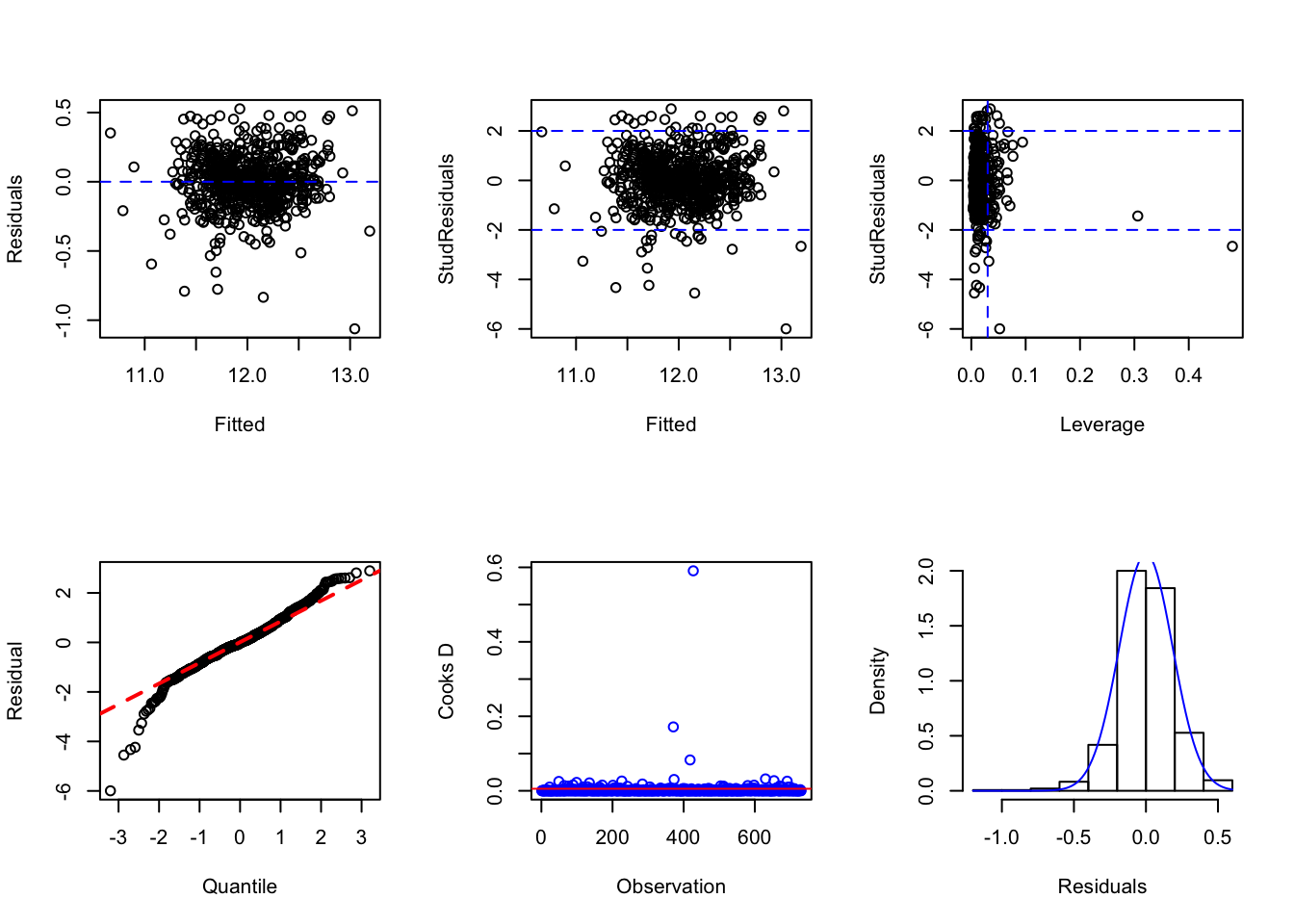
From this model we took the top ten variables that were the most significantly correlated with L\_SalePrice. The model consists of MSSubClass, LotArea, OverallQual, OverallCond, YearBuilt, YearRemodAdd, X2ndFlrSF, BedroomAbvGr, KitchenAbvGr, GarageCars



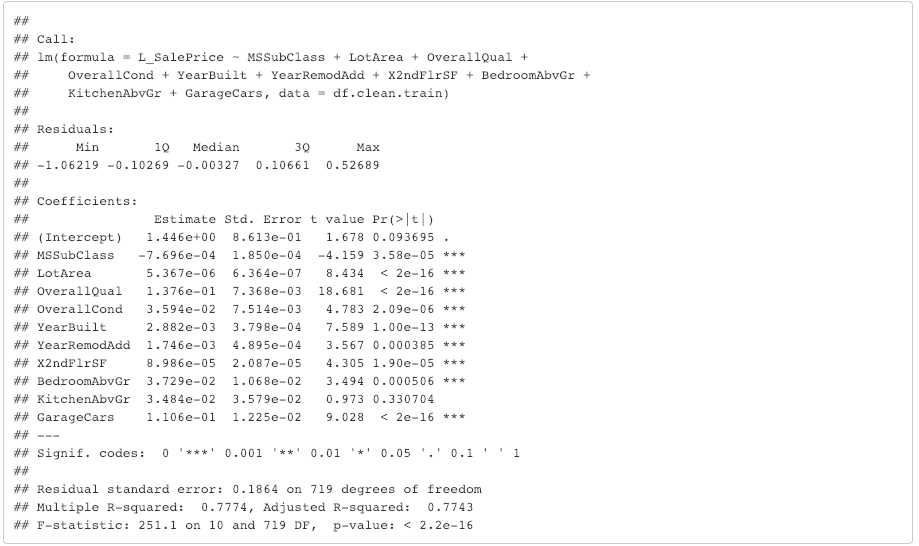
RMSE: 40740.31, Adj-R2: 0.7642468



As seen the VIF’s are very low and nearly the same as each other, 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 outlier. However this is just one point out of 1480, its influence is low so we continued our analysis with the point included.



1. **Checking Assumptions**

Residual Plots

Influential Points

1. **Comparing Models**

RMSE or whatever

1. **Parameter Interpretation**

Interpretation

Confidence Intervals

1. **Conclusion**

Conclusions, insights, concerns, what to do better next time?

*Do better*: deal with zero inflation aka create dummy variables

1. **Objective 2**
2. **Goal of 2way ANOVA**

State what route you are going to take 2way ANOVA or Time series and summarize the goal.

1. **Analysis of 2way ANOVA**
2. **Conclusion/Discussion Required**

The conclusion should reprise the questions and conclusions of objective 2.

1. **Appendix**

Contains code and extra charts