Habeeba Mohamed - 900191525

Mona Ahmad - 900191833

Sara Eldegwy - 900191860

Regression Project: House Prices

Problem Statement: The main aim of the project is to examine the features that contribute to the price of a property. A dataset of 4600 properties will be investigated using regression analysis. Before fitting any models or even graphing the data, variable type and description must be clearly and properly identified. The dataset includes 6 **Categorical** variables, 2 **Ordinal** variables and 10 **Quantitative** variables; making a total of 18 variables. We will consider the price as the response variable.

Source: Data set is obtained from the website: https://www.kaggle.com/datasets/shree1992/housedata?resource=download&select=data.dat

Variable Description

Variable Name (unit)	Variable Description	Variable Type
Date & time (m/dd/YYYY 0:00)	Date & time of the home sale.	Categorical
price($)	The property price (in dollars).	Quantitative
bedrooms	The number of bedrooms.	Quantitative
bathrooms	The number of bathrooms	Quantitative

sqft_living(sqft)	Square footage of the apartment's interior living space.	Quantitative
sqft_lot(sqft)	Square footage of the land space.	Quantitative
floors	The property floor.	Quantitative
waterfront	Whether the property has a waterfront or not. 0: Does not have a waterfront. 1: Does have a waterfront.	Binary
view	Rating of the property view from 0 to 4.tu6	Ordinal
condition	The condition of the property is given on a scale from 1 to 5.	Ordinal
sqft_above	The square footage of the interior housing space that is above ground level	Quantitative
sqft_basement	The sqrt of the basement.	Quantitative
yr_built	The year in which property is built in.	Quantitative
yr_renovated	The year in which property was renovated in.	Quantitative

street	The street in which the property is located in.	Nominal
city	The city where the property is located in.	Categorical
statezip	The statezip code where the property is located in.	Categorical
country	The country where the property is located in.	Categorical

Hypothesis about effect of each predictor on the response variable:

Figure 01 shows 4 scatter plots portraying bedrooms vs price, bathrooms vs price, sqft_living vs price vs sqft_lot. Starting with the number of bedrooms vs price,we expect that they would have a positive relationship, but there are significant outliers that could be affecting the scale and causing the graph to not clearly show the relationship. As for the second graph, it is a plot of the number of bathrooms and the price. It still shows two properties that are outliers than the other ones. More analysis is needed to confirm whether the same two properties are a source of outlier in all of the graphs. The graph representing sqft_living and price shows something very interesting which is that one property has significantly big living space but this does not reflect in the price. As for the fourth scatter plot, it touches on sqft_lot (square foot of the land space) and price. This also shows that one property has a large value on its x axis but this does not necessarily translate into higher price as one would expect.

Figure 02 shows four scatter plots. The first scatter plot represents sqft_above and the price represents a slightly positive relationship as we would expect, however, two properties could be

considered outliers as they do have higher prices but their sqft_above is less than the average. A very similar pattern could be identified for the graph sqft_basement and price. As for the third one, yr_built and price, we can see that newer properties do not have higher prices but rather almost the same prices. This might need further analysis as it was expected that newer houses will have higher prices. As for the last graph, yr_renovated and price don't seem to have a relationship but we can see that there are two properties that stand out and would need further investigation.

Figure 03 represents the categorical variables showing no relationship between each of them on its own and the response variable and it is still clear that two properties stand out from the others that might be the same in all graphs. We need to investigate more since we think that as the view and condition is enhanced the price of the house will increase and the same is applied for the waterfront and number of floors. *Figure 04* shows the boxplot of the standardized price (response variable). Since it is not appropriately shown, a histogram is made (Figure 05). The standardized histogram shows positive skewness, which might be due to the presence of outliers.

Data Analysis:

1.Data Processing

Upon reading the dataset, some data processing was considered. First, to avoid redundancy and repeated information some variables were dropped (street, city, statezip and country). Some properties that are included in the data set have a price of zero dollars, this is definitely a typo and thus these observations were removed. In addition, the date included in the dataset belonged to the same year but different month so to simplify

analysis in the next steps. So the format included only the month for simplification. Part of our data preparation before conducting any model fitting is to standardize the data.

2.Iterative Process

Iteration 01:

Graphs before fitting(*Figure 06 until Figure 22*)

The **first linear regression model fitted** showed that the coefficient of the predictor s*qft_basement* was NA. As we research this unusual result, we find that this is due to the variable in question being perfectly linearly related to the other variable. We went back to the description of the variables and found out that the sum of *sqft_basement* and sqft_above is sqft_living. This was confirmed relying on R language statistical tools. We added the two variables and calculated the correlation coefficient which was perfectly related to sqft_living.

Summary about the model in Iteration 01 (*Figure 23)*

Results: R2 adjusted is 0.234 and some coefficients are NA, which mean perfect collinearity.

Iteration 02:

Perfect collinearity between sqft_living and the summation of *sqft_basement* and sqft_above is clearly shown in the *Figure 24* of the appendix. Thus, we attempted to fit two models to make a decision about which of the variables we shall keep and which ones will be removed. The two models turned out to have the same exact p value (< 2.2e-16) so to have a simpler model, only sqft_living was kept while *sqft_basement* and sqft_above were removed. The model was fitted again after removing *sqft_basement* and sqft_above. The adjusted R2 adjusted is still 0.234, but there is no NA.

Iteration 03:

As it was clear from the graphs before fitting, there exist two very extreme outliers that exist in the price variable. Two models will be developed; one with the two outliers and the other models

will not contain the two outlier points (4351 and 4347). [*Figure 07* in the appendix shows those two points].Fitting the two models we receive two models with the same p value, thus we will be working on the data that does not have the two outliers points. In fact, we notice that R2 adjusted became 0.5799, which is a big improvement from the previous one 0.234.Not only is R2 adjusted improved but also the graphs after fitting (*Figure 25)*

Iteration 04:

Upon examining the fitted Values against the Residuals plot we will work on a proper power transformation for the response variable. Testing with different lambda (*Figure 26)* , we found out that 0.5 is the best lambda as the slope is 1 and it passes by the origin. In reality. We suspected three lambdas, so we constructed three models and R2 adjusted is only improved at lambda equal to 0.5, the R2 adjusted becomes 0.5929. In the graphs after fitting (*Figure 27),*not only is the Normal Q-Q plot improved but also the Histogram of the residual which is starting to look more normal and the Residual Index plot is bound between the zero with little outliers in comparison with the huge bulk around the zero and between 3 and -3.

Iteration 05:

Examining each predictor variable if it needs transformation. We looked at the graphs of lambdas for each of them. The only predictors that required transformation are sqft_living (*Figure 28)* and bathrooms (*Figure 29)* since the other predictors showed no difference in graphs for different lambdas (these are not in the appendix due to their insignificance)

After fitting the models, we only transformed the bathroom since the R2 after transforming sqft_living decreased. Fitting the model after transformation, R2 adjusted is 0.5903. When we plot the graphs after fitting (*Figure 30),*they were similar to the initial ones. In an attempt to modify the graphs and the model, we will investigate further.

Iteration 06:

We started by examining Cook's distance, leverage, Hadi' influence, and the potential residual plots (*Figure 31).* From the potential residual plot, there are 6 regression outliers and 5 high leverage points. We want to investigate if interactive variables will enhance the model. Intuitively, the number of bathrooms and bedrooms can enhance the model if added as an interactive variable. To be able to determine this, the graphs of bathrooms vs. residual *(Figure 32)* and bedrooms vs. residual *(Figure 33)* were displayed with the extreme outliers identified. There are common outliers. Thus, we will try the interactive variable bedroom*bathroom. We will fit the model with and without the interactive variable to see if the difference is significant. Without the interactive variable, the R2 adjusted is 0.5903. With the interactive variable, the R2 adjusted increase to 0.5941 and it has a significant coefficient showing that it is important to add it to the model, so it was added to the model.

Another interactive variable that we wanted to investigate is the yr_built*condition. We fit a model before and after adding this interactive variabel. Since R2 adjusted remained 0.5941 and the coefficient of the interactive variable was insignificant, we concluded that this interactive variable is not useful, so we did not add it to the model.

Iteration 07:

In an attempt to enhance the model, we decided to examine the problem of collinearity. Perfect collinearity was resolved in a previous iteration, but we need to make sure that there is no correlation between predictors. First,we had to change all variables in the data frame to be numeric

to construct the correlation matrix, eigenvalues and eigenvectors. We got the values of the condition indices to know whether there is collinearity in the data. Since one of the values was more than 10, we had to start running the Principal Component Regression to remove this collinearity. Fitting the response variable against the principal component, we found that W7 and W12 are insignificant and W6 had a p-value of 0.0515, so we decided to remove all three. The fitted model had a very high R2 adjusted of 96.7%.Again, we fit the graphs after fitting the model*(Figures 35 & 36)*, but it seems that there are still unresolved issues.

Conclusion

As clarified before, we have made several attempts to resolve the model, but there are still apparent problems in the graphs. We suppose that maybe the dataset is part of a larger dataset with more observations from different states or even countries. We were only able to improve the R2 adjusted to 96.7%, but there are still outliers and some graphs show that not all assumptions are verified.

Appendix A

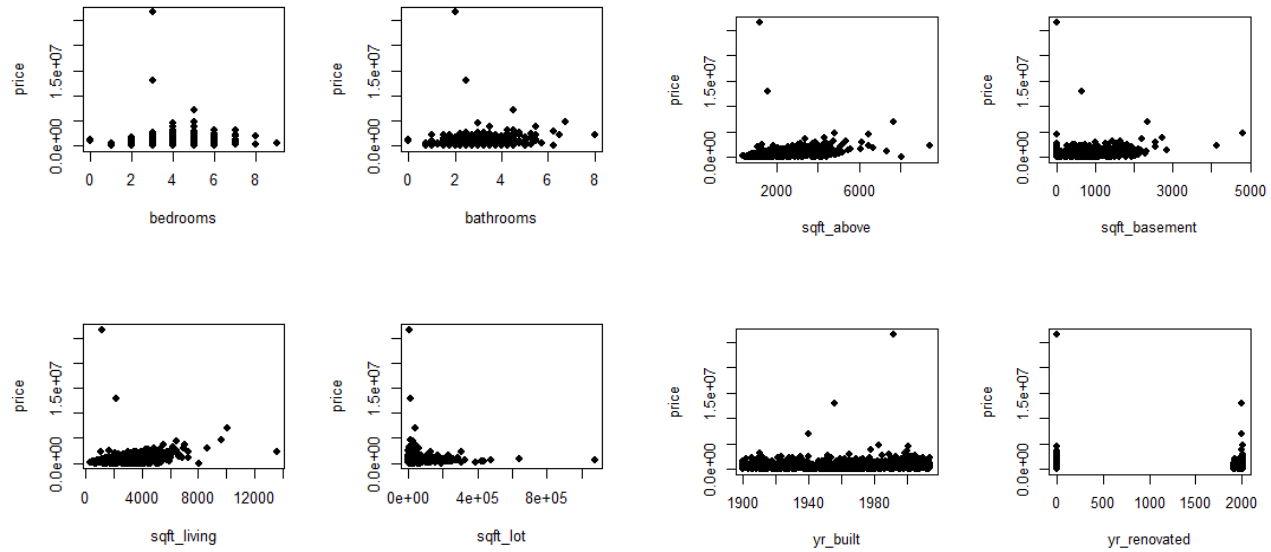


Figure 01

Figure 02

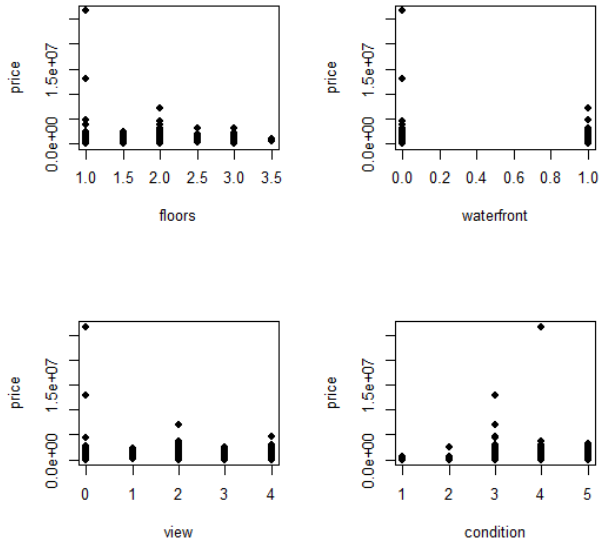


Figure 03

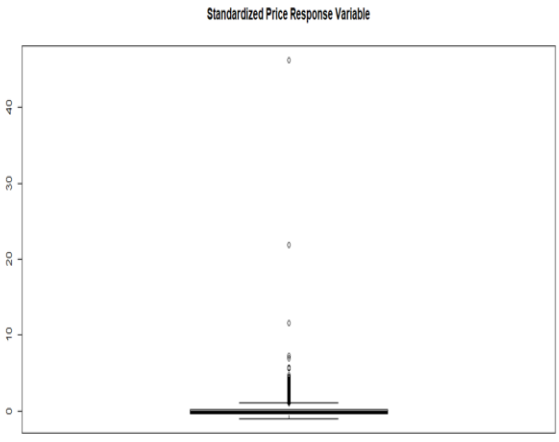


Figure 04

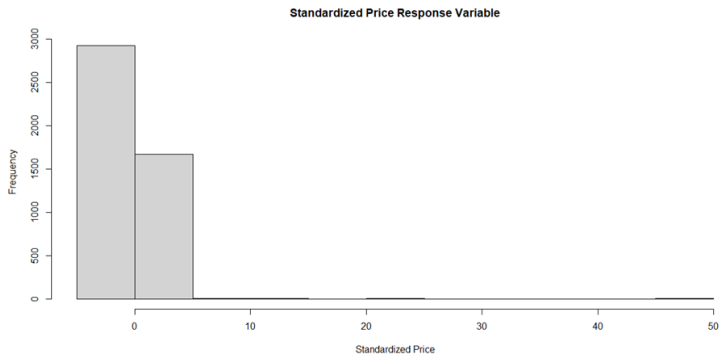

Standardized Price Response Variable
Frequency
0
500
1000
1500
2000
2500
3000
0
10
20
30
40
50
Standardized Price

Figure 05

Appendix B

Iteration 01:

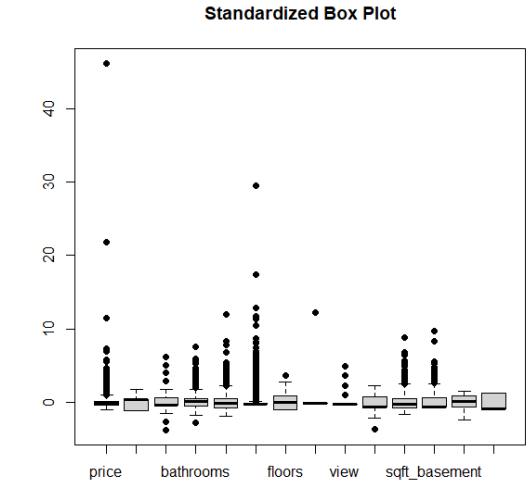
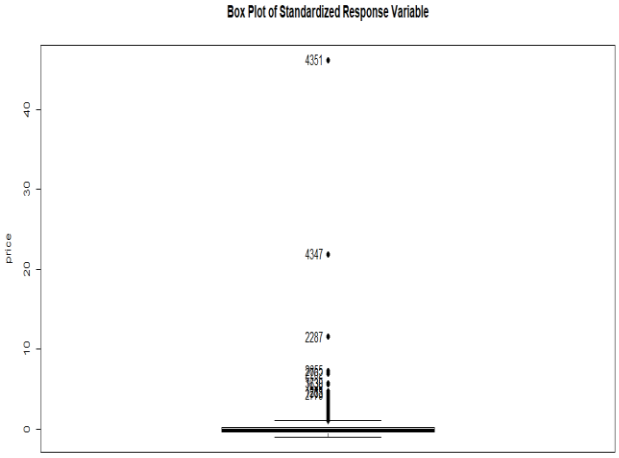


Figure 06



Fiaure 07

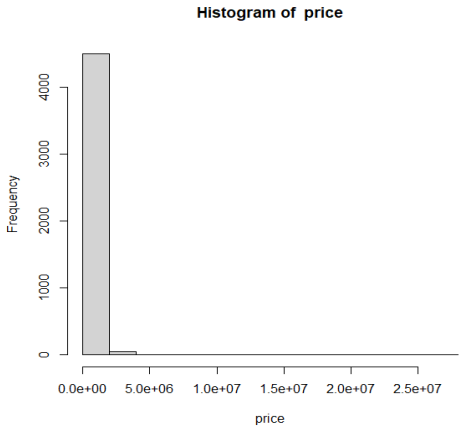


Figure 08

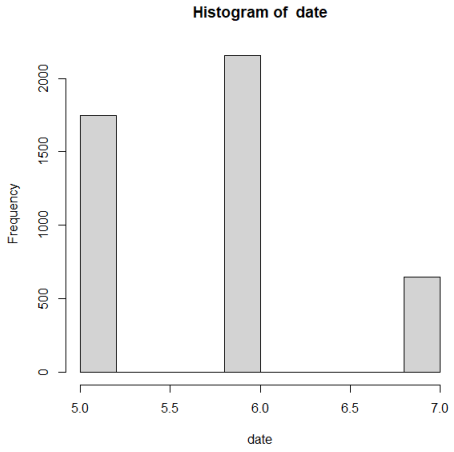


Figure 09

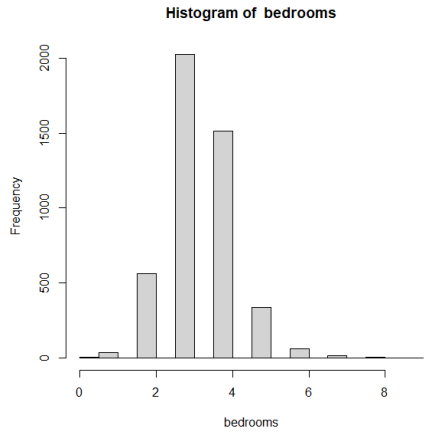


Figure 10

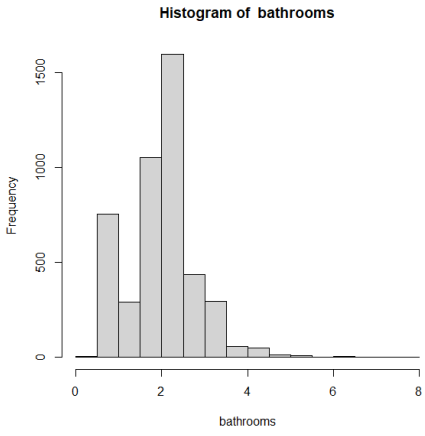


Figure 11

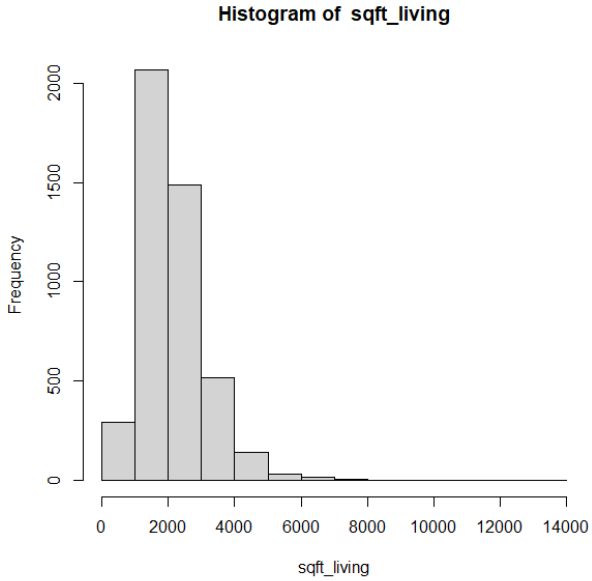


Figure 12

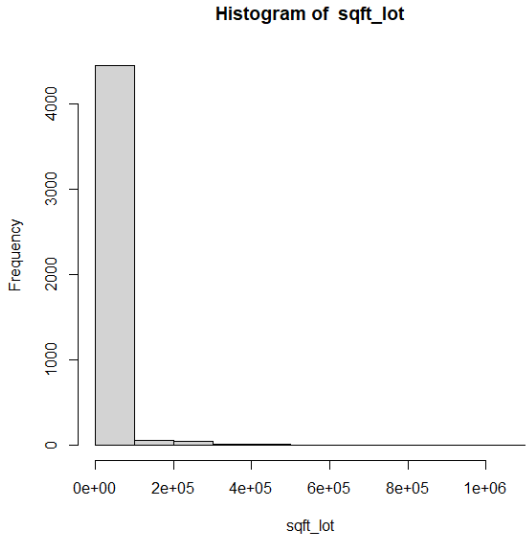


Figure 13

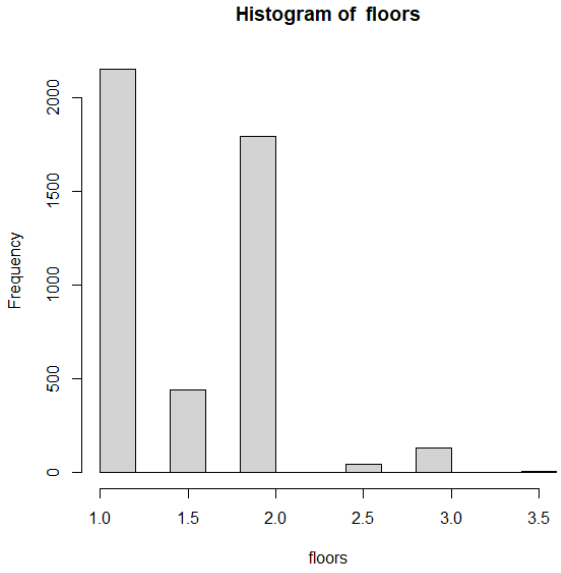


Figure 14

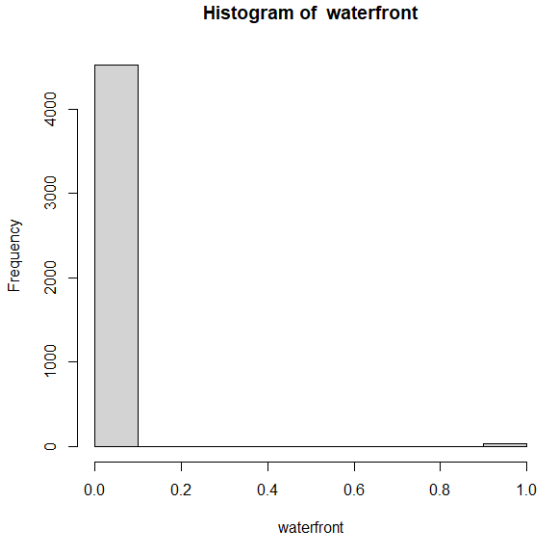


Figure 15

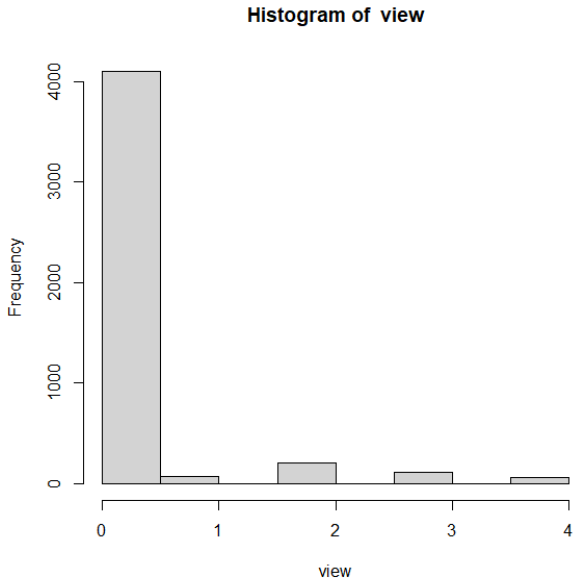


Figure 16

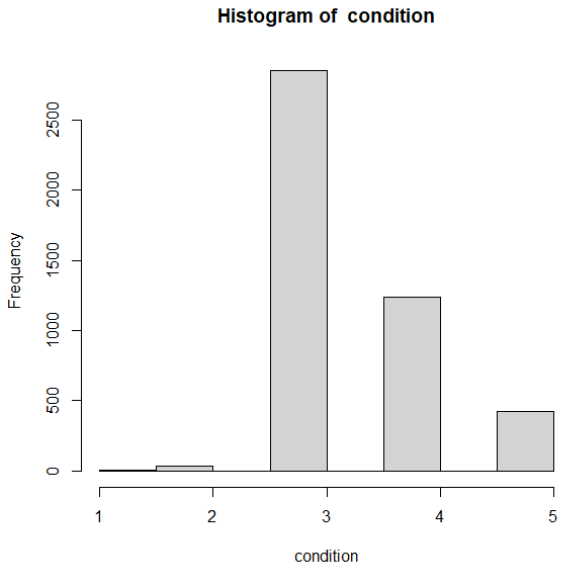


Figure 17

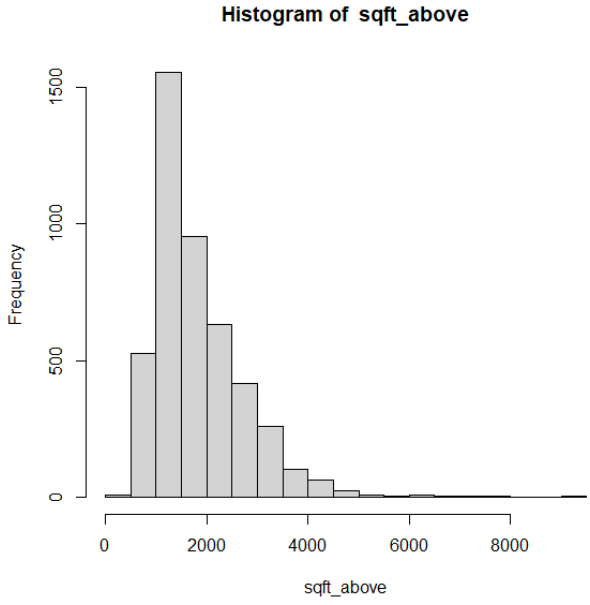


Figure 18

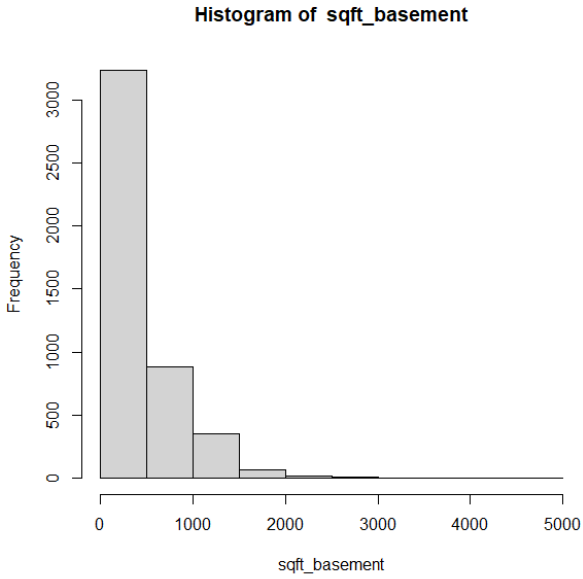


Figure 19

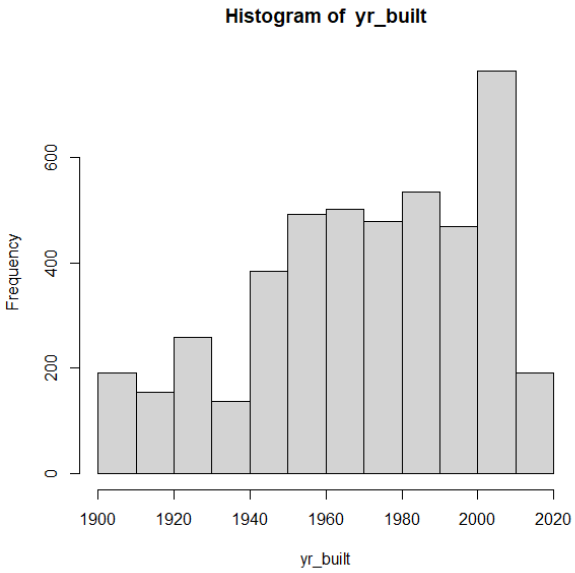


Figure 20

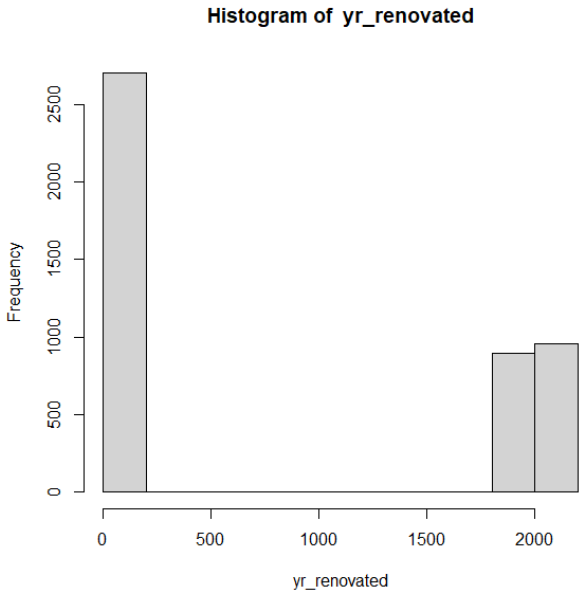


Figure 21

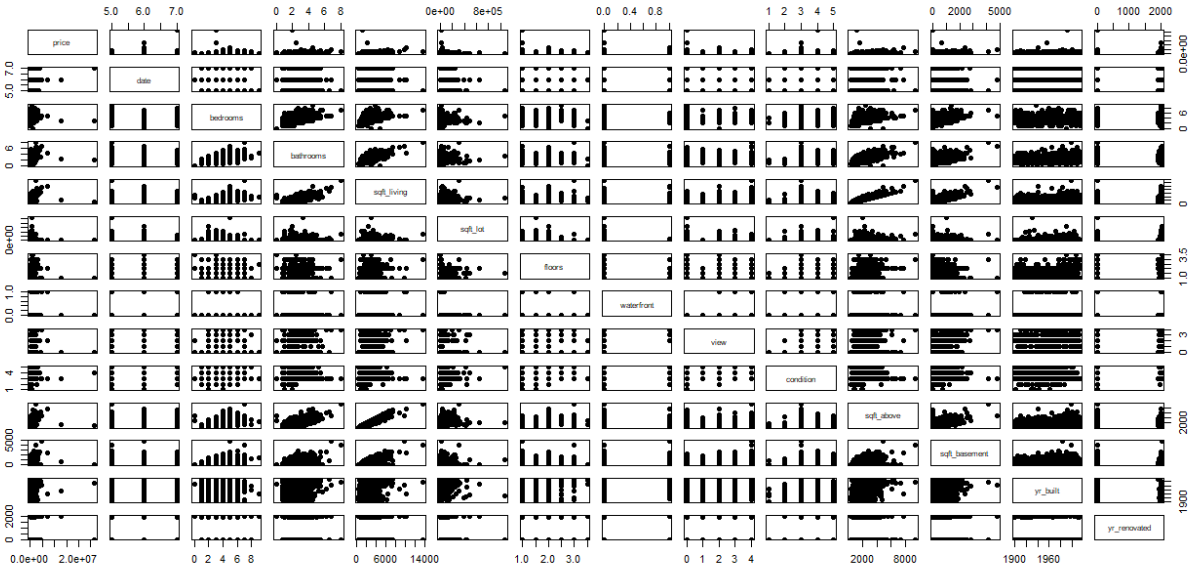


Figure 22

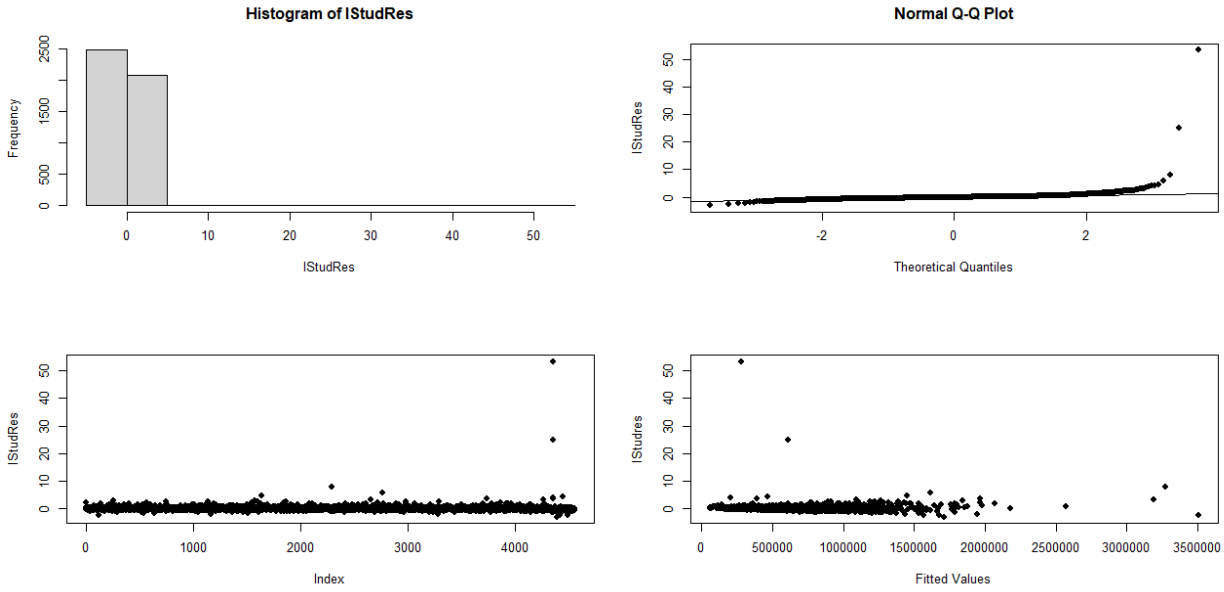


Figure 23

Iteration 02:

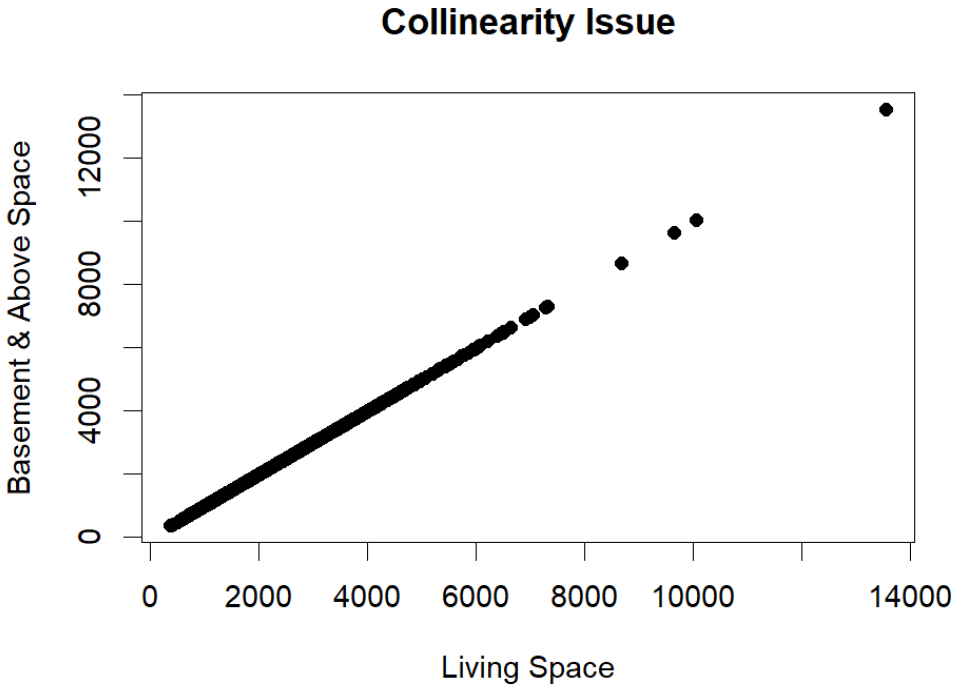


Figure 24

Iteration 03:

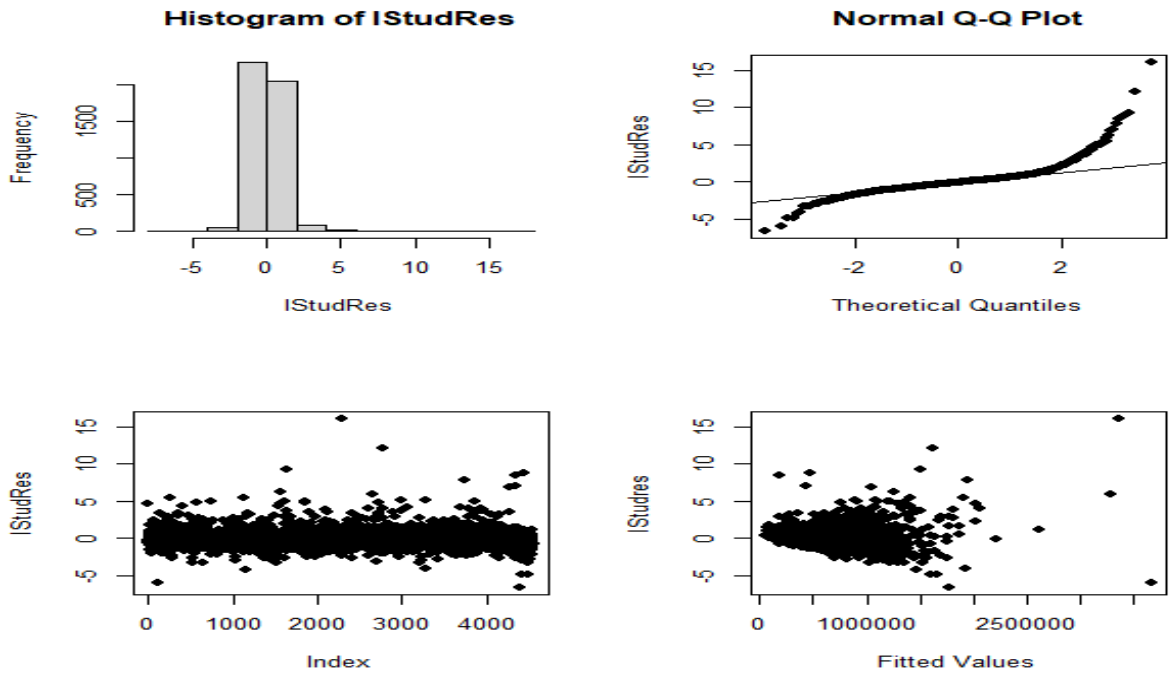


Figure 25

Iteration 04:

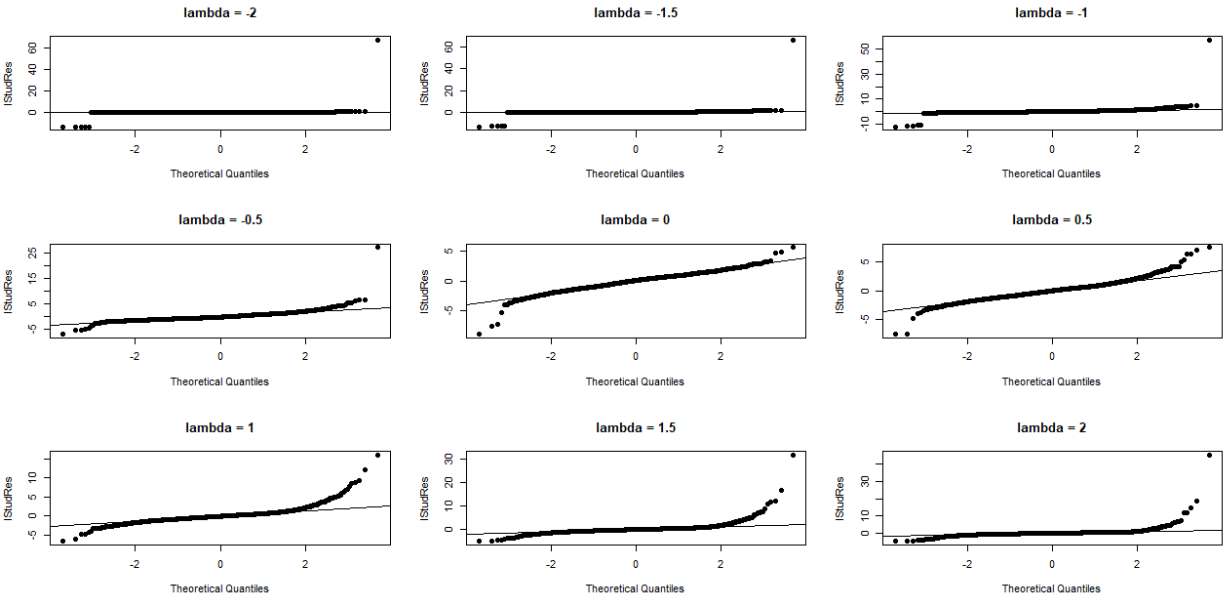


Figure 26

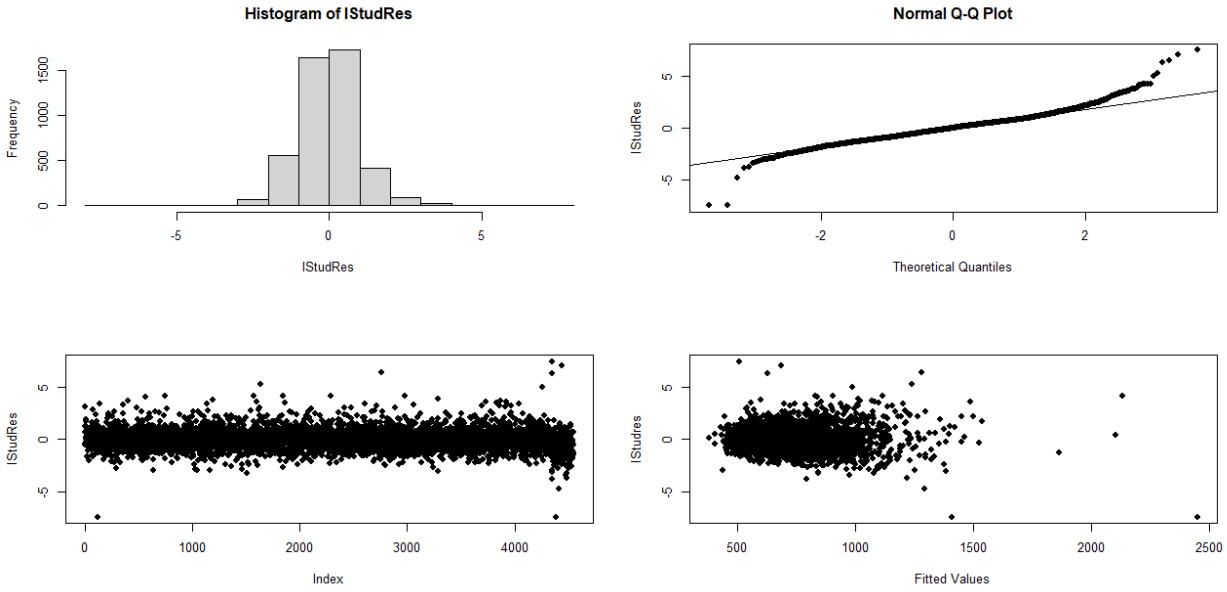


Figure 27

Iteration 05:

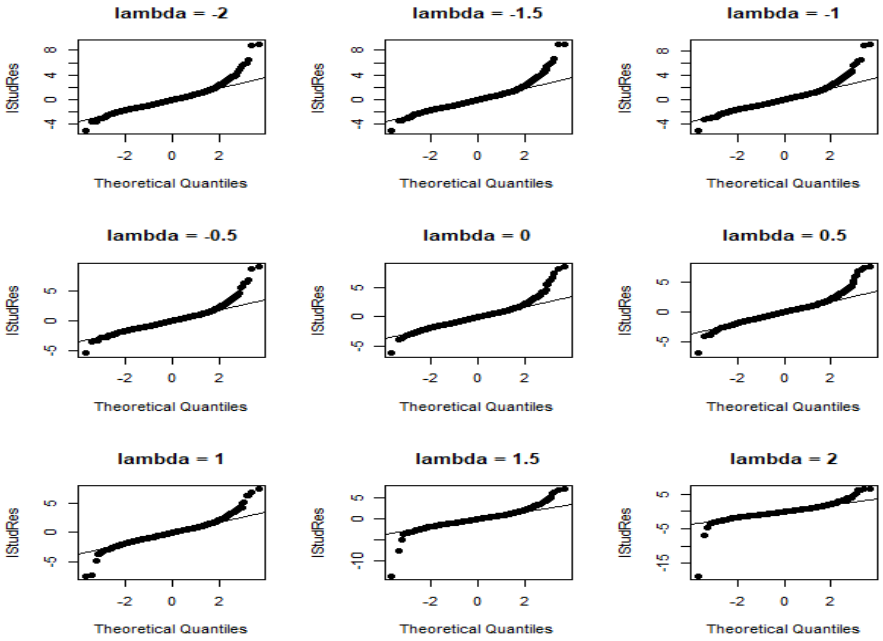


Figure 28

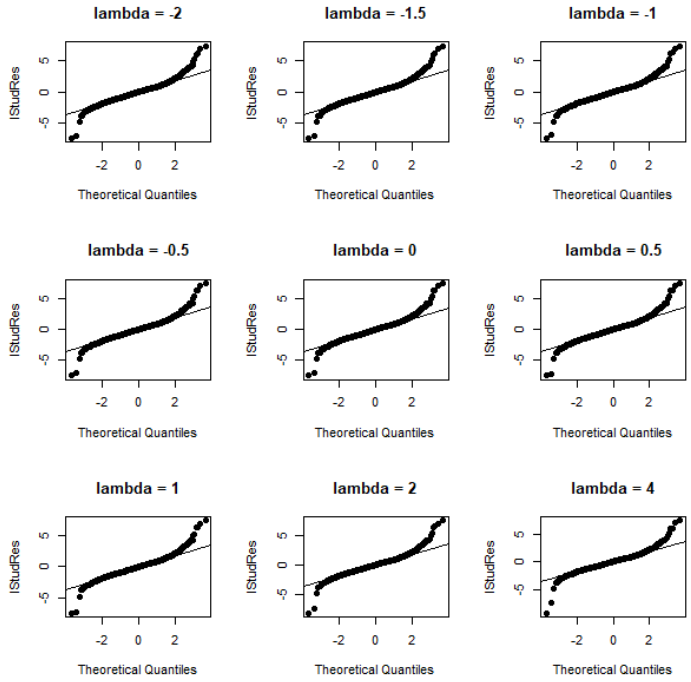


Figure 29

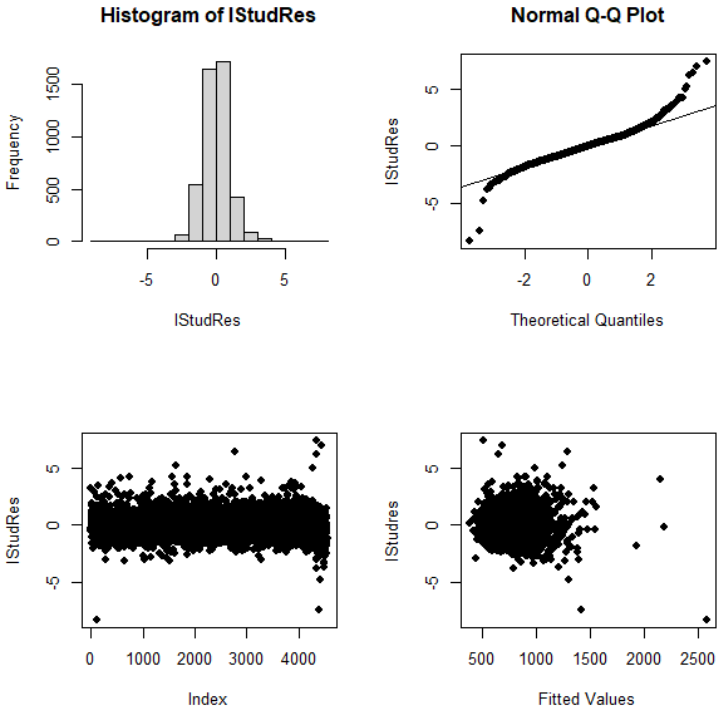


Figure 30

Iteration 06:

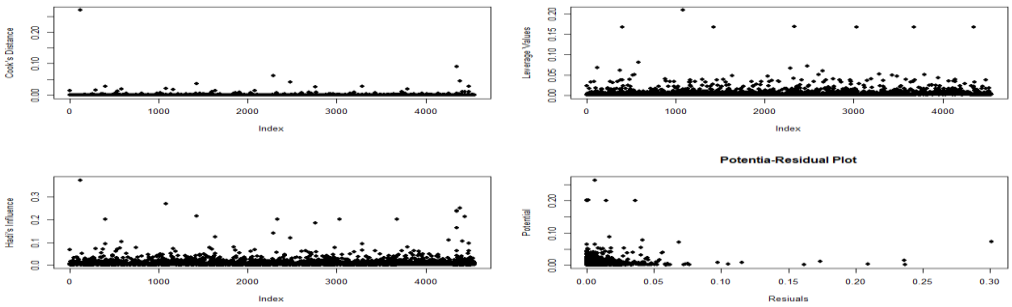


Figure 31

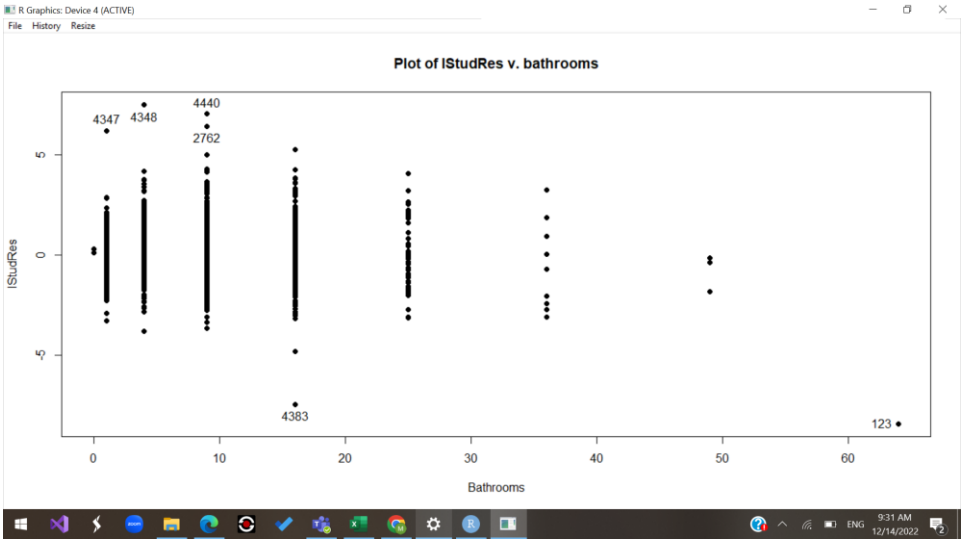


Figure 32

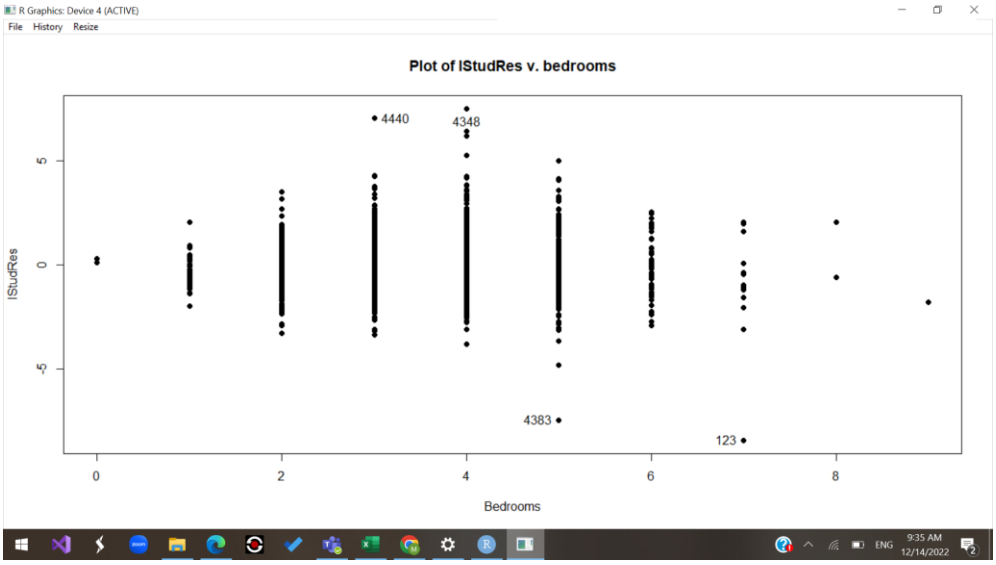


Figure 33

Iteration 07:

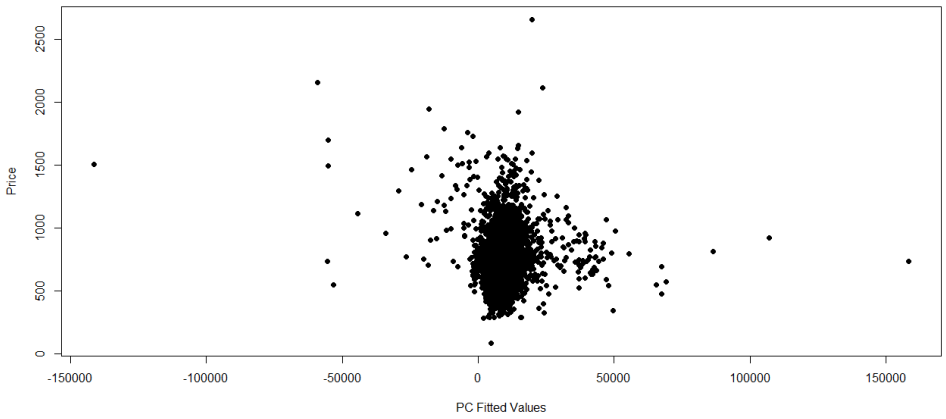


Figure 34. Graph after PC regression

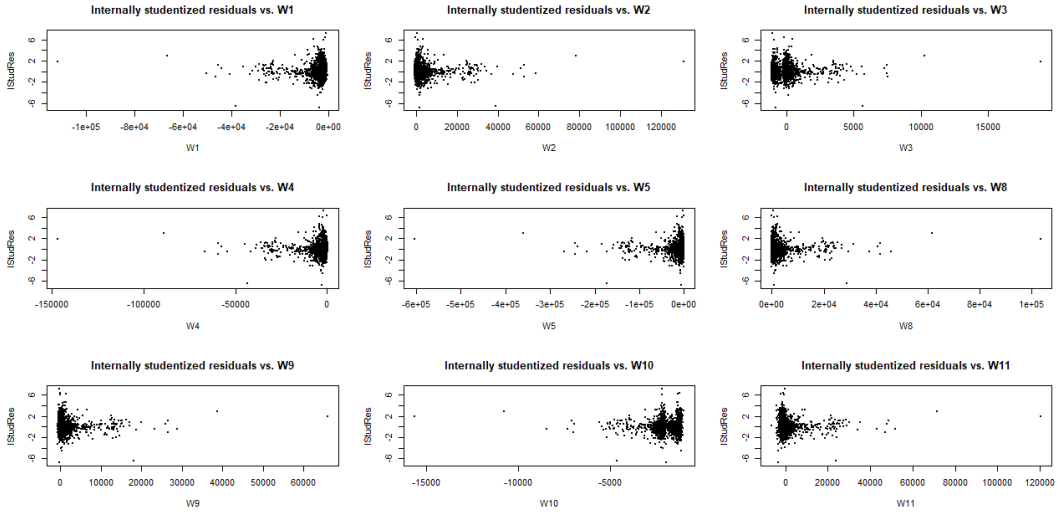


Figure 35

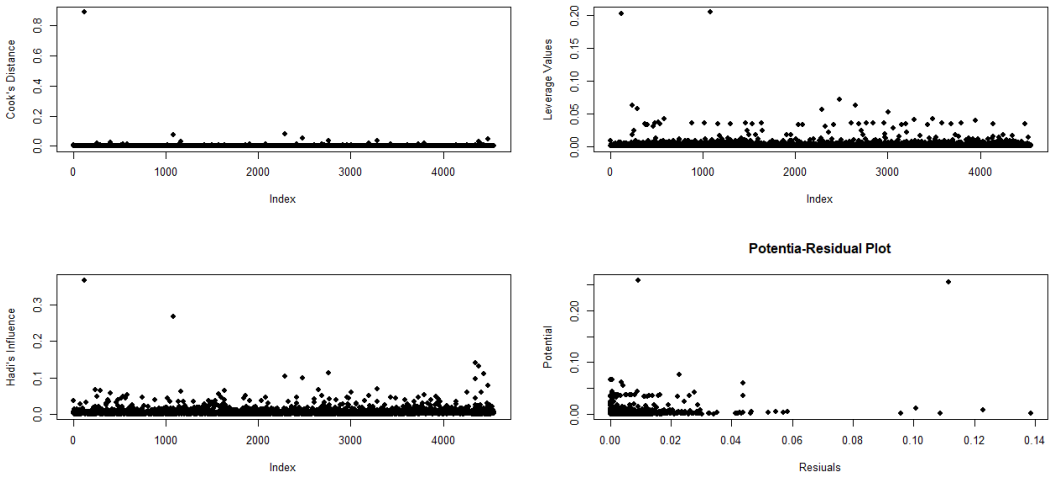


Figure 36