University of Strathclyde Glasgow

Regression Modelling Project

MM916 Data Analytics in R

Eitika Sharma

202352860

Contents

Table of Figures

1.Introduction

This study thoroughly analysis housing data related to 156 Chicago properties in order to investigate the detailed relationship between property values and a variety of associated elements. The dataset contains all the relevant data, such as the price (in USD 10K), the number of bedrooms and rooms, the size of the house (in square feet), the width of the lot, the amount of annual taxes (in USD), the number of bathrooms, the number of garages, and the house's condition (rated as 1 for good condition and 0 otherwise).

2.Objective

The main objective of the analysis, which employs regression modelling techniques, is to comprehend and forecast property values based on various inputs. Determine the essential elements that have a major impact on home prices. Assess the validity and reliability of the regression models as well as the accompanying hypotheses using the appropriate assessment tools.

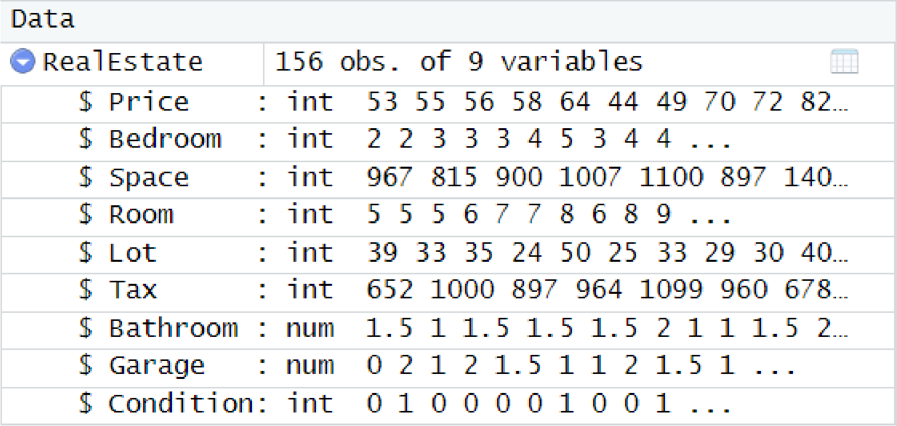
- Null Hypothesis: The target variable and any of the predictor variables do not have a linear relationship.

- First Hypothesis: It states that a home's size and space have a positive effect on its value, the greater homes with greater floor layouts and more rooms should see price hikes.

- Second Hypothesis: A house's condition has a significant impact on its price (Condition). Better-looking properties (Condition = 1) might sell for more money than worse-looking ones.

3.Dataset Exploration

The 'Real Estate.csv' file contains information on Chicago property prices. We read this csv file and saved it into a dataset called 'RealEstate'. This dataset has 156 observations organized into the 9 columns:



```
Data
RealEstate      156 obs. of 9 variables
   $ Price    : int  53 55 56 58 64 44 49 70 72 82...
   $ Bedroom  : int  2 2 3 3 3 4 5 3 4 4 ...
   $ Space    : int  967 815 900 1007 1100 897 140...
   $ Room     : int  5 5 5 6 7 7 8 6 8 9 ...
   $ Lot      : int  39 33 35 24 50 25 33 29 30 40...
   $ Tax      : int  652 1000 897 964 1099 960 678...
   $ Bathroom : num  1.5 1 1.5 1.5 1.5 2 1 1 1.5 2...
   $ Garage   : num  0 2 1 2 1.5 1 1 2 1.5 1 ...
   $ Condition: int  0 1 0 0 0 0 1 0 0 1 ...
```

Based on the boxplot depiction of every column in the dataset, we came to a conclusion that the Space and Tax columns include inconsistencies.

3.1 Importing Libraries and Dataset

The following list of libraries were required to dig in and explore the dataset properly: Tidyverse, ggplot2, GGally, ggfortify, car, rcompanion, leaps, MASS, corrplot, RColorBrewer.

3.2 Exploratory Data Analysis and Data Visualization

3.2.1 Box Plot of Housing Variables

This visualization shows a boxplot for each column in the dataset. It is noteworthy that columns such as Space and Tax have outlier values, indicating potential skewness or extreme values.

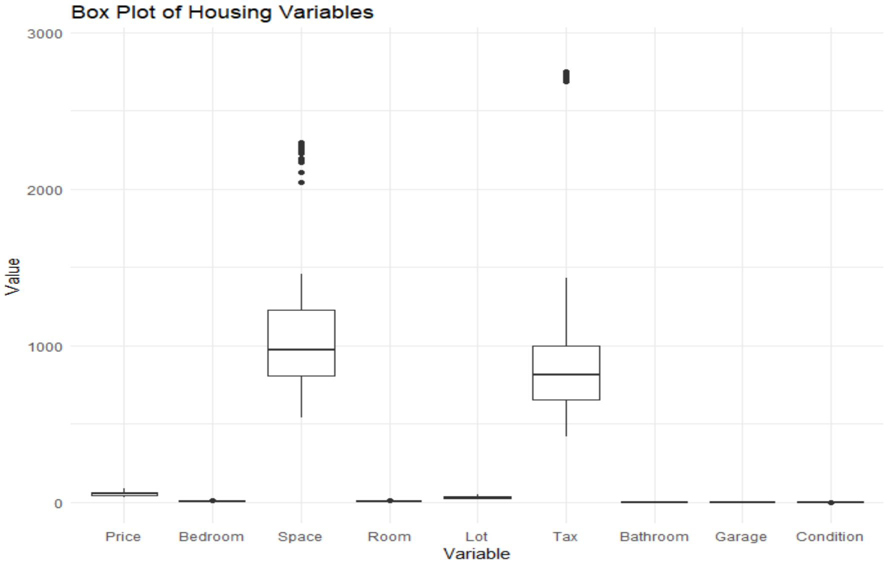


Figure 3.2.1: Box Plot of Housing Variables

3.2.2 Correlation Heatmap Matrix

The relationship between variables is depicted in a correlation matrix. The following include a few relevant findings:

- Price and Space, Room, and Condition have a possibly strong interconnect
- Bedroom, Space, and Room have an established relationship.

Using the ggpairs() technique provides useful predictor variable information. The distributions of the remaining correlate columns exhibit a noticeable upward bias when Condition is considered as a factor. To ensure fit for a linear model, Transformation-based normalisation is necessary.

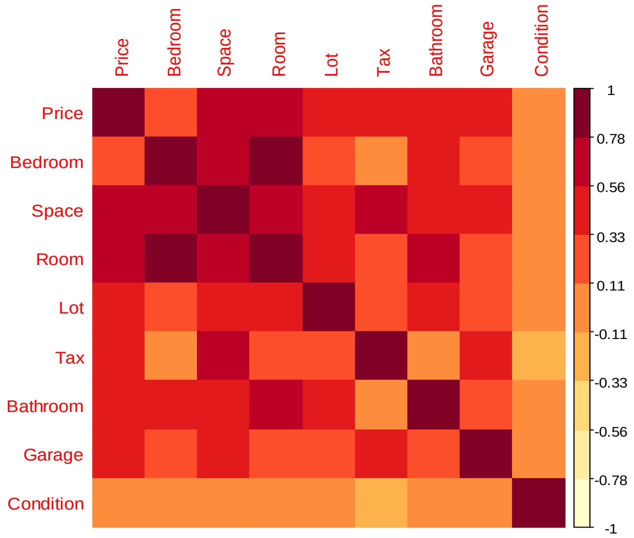


Figure 3.2.2: Correlation Heatmap Matrix

3.2.3 Scatterplot Matrix

A scatterplot matrix illustrating price, space, tax, bathroom, and garage pricing demonstrates the favourable correlations between these factors.

Each column has significant skewness indicated in it.

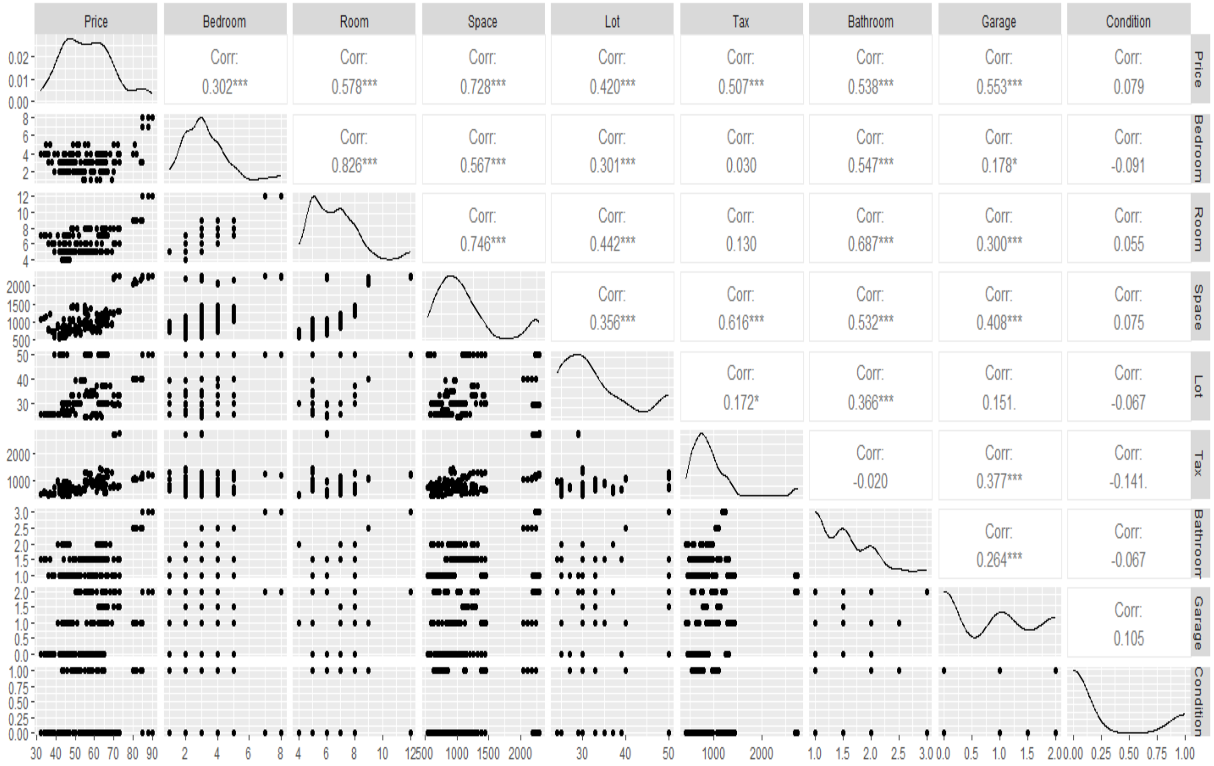


Figure 3.2.3: Scatterplot Matrix

3.2.4 Histogram Visualisation of Original Variables

It's necessary to explore histograms in order to comprehend the characteristics and distribution of every variable in the dataset. From Figure 3.2.4.a we can see most variables show a right skew, indicating a prevalence of properties with lower values for features like bedrooms, space, and taxes, while fewer have higher values, necessitating data transformation for analysis.

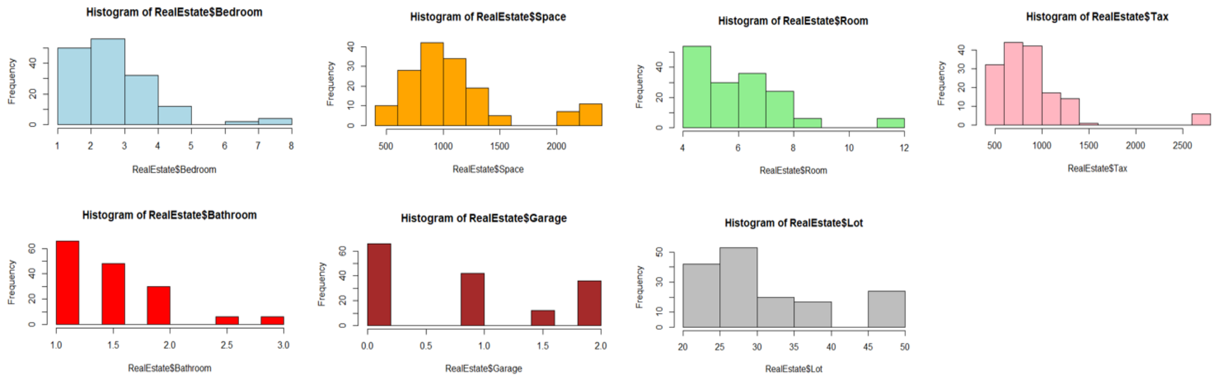


Figure 3.2.4.a: Histogram representation for all the original variables

To eliminate the skewness, I applied Tukey's transformation on few of the columns (Space, Tax, Bathroom, Garage)(Figure 3.2.4.b). By normalising the distribution of the predictor variables, the transformation aims to increase their relevance for a linear model.

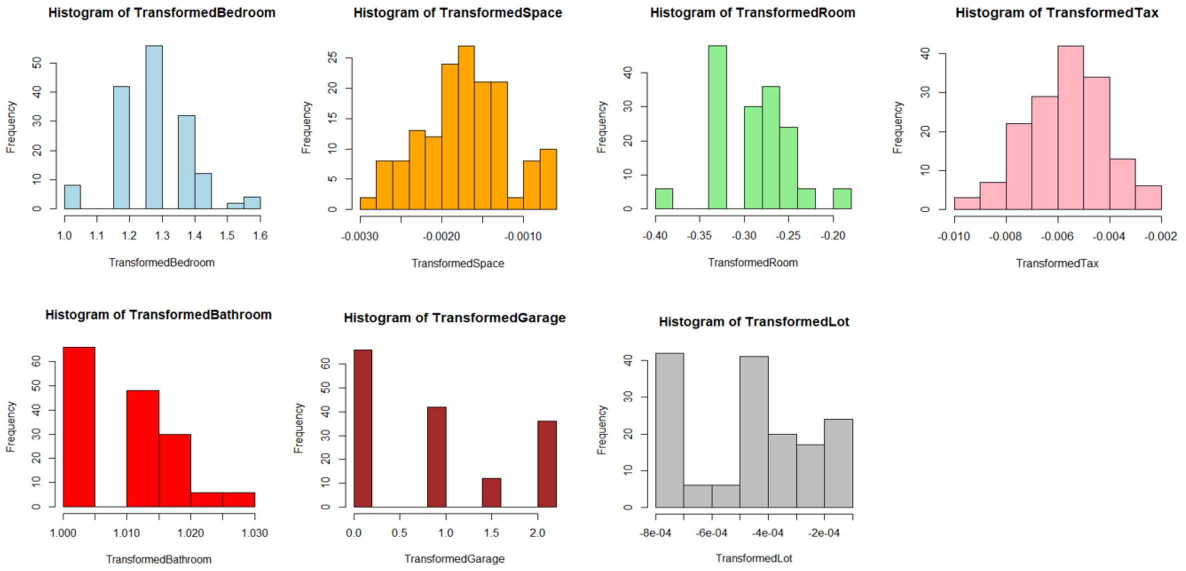
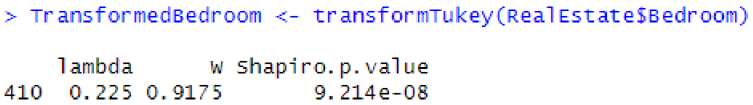


Figure 3.2.4.b: Histogram representation for all the transformed variables

4. Transformation on Individual Variables

4.1 Bedroom Variable

For Bedroom variable, below result was obtained after Tukey Transformation Technique was performed.



```
> TransformedBedroom <- transformTukey(RealEstate$Bedroom)

    lambda      W Shapiro.p.value
410  0.225 0.9175       9.214e-08
```

The Q-Q plot demonstrates (Figure 4.1) that the distribution has dominant tails, which points to a deviation from regularity. This is resolved by fitting a linear model to the data and examining the transformation's consequences.

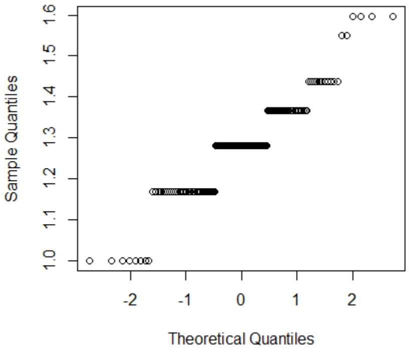
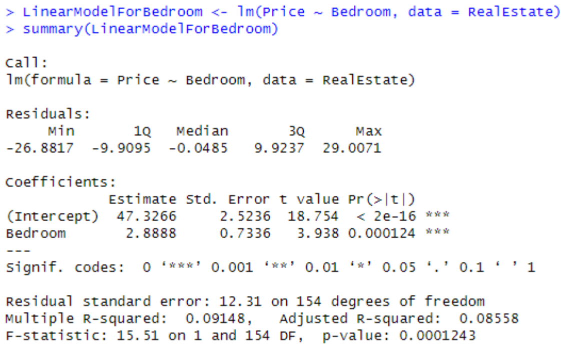
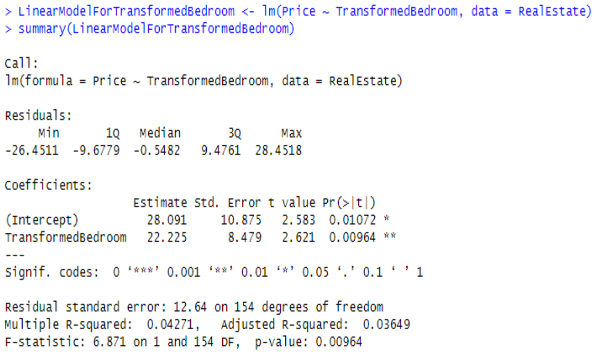


Figure 4.1: Q-Q Plot for Bedroom

It's interesting to note that the p-value for the Bedroom variable before transformation is 0.000124, greater than the typical significance limit of 0.05. However, after using the Tukey transformation with Lambda = 0.225, the p-value for the transformed Bedroom column decreases to 0.00964. The modification normalizes the data distribution, but it is still distinctly tailed. Bedroom has a moderate correlation coefficient (0.30245123), however the change in parameters alters the statistical significance. This demonstrates how important it is to consider both correlation and distribution features when modelling.



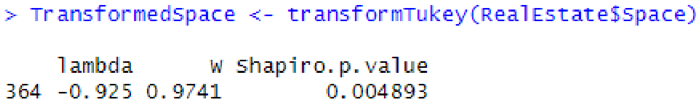
```
> LinearModelForBedroom <- lm(Price ~ Bedroom, data = RealEstate)
> summary(LinearModelForBedroom)

Call:
lm(formula = Price ~ Bedroom, data = RealEstate)

Residuals:
     Min       1Q   Median       3Q      Max 
-26.8817  -9.9095  -0.0485   9.9237  29.0071 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  47.3266     2.5236  18.754  < 2e-16 ***
Bedroom       2.8888     0.7336   3.938 0.000124 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.31 on 154 degrees of freedom
Multiple R-squared:  0.09148,	Adjusted R-squared:  0.08558 
F-statistic: 15.51 on 1 and 154 DF,  p-value: 0.0001243
```



```
> LinearModelForTransformedBedroom <- lm(Price ~ TransformedBedroom, data = RealEstate)
> summary(LinearModelForTransformedBedroom)

Call:
lm(formula = Price ~ TransformedBedroom, data = RealEstate)

Residuals:
     Min       1Q   Median       3Q      Max 
-26.4511  -9.6779  -0.5482   9.4761  28.4518 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)   
(Intercept)          28.091     10.875   2.583  0.01072 * 
TransformedBedroom   22.225      8.479   2.621  0.00964 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.64 on 154 degrees of freedom
Multiple R-squared:  0.04271,	Adjusted R-squared:  0.03649 
F-statistic: 6.871 on 1 and 154 DF,  p-value: 0.00964
```

4.2 Space Variable

For Space variable, below result was obtained after Tukey Transformation Technique was performed.



```
> TransformedSpace <- transformTukey(RealEstate$Space)

    lambda      W Shapiro.p.value
364 -0.925 0.9741        0.004893
```

Space variable has shown a clear indication of right skewness. By implementing Tukey's transformation with a lambda value of -0.925 (formula: TRANS = -1 * x ^ lambda), the data for the Space column is significantly normalized(Figure 4.2). The modification seeks to reduce the skewness to the right.

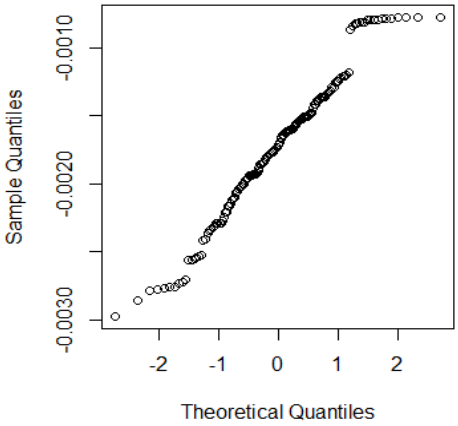
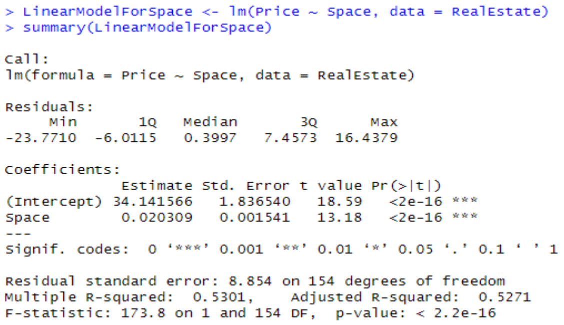
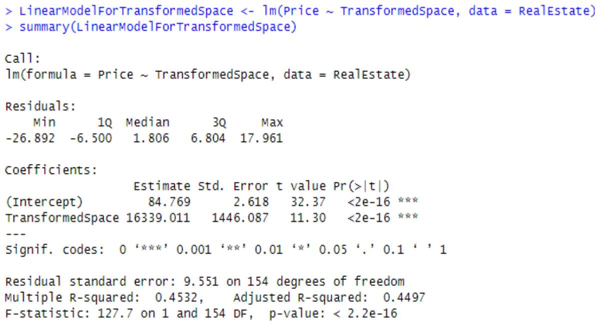


Figure 4.2: Q-Q Plot for Space

By applying the transformation to the Space variable, the data was successfully normalised, leading to an improved model fit. This improvement does, however, lessen the overall model significance, as the F-statistic score decreases from 173.8 to 127.7. Additionally, the standard error coefficients have expanded dramatically, from 0.001541 to 1446.087, indicating a wider distribution of the model. Despite the improved skewness, the decision to prioritize the least standard error above full normalization originates from the need to correctly read the linear model and understand the dependence with the dependent variable. This emphasis on accuracy ensures more accurate findings about the relationship, particularly when assessing the impact of the Space variable, which consistently has a high coefficient value.



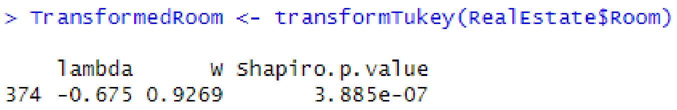
```
> LinearModelForSpace <- lm(Price ~ Space, data = RealEstate)
> summary(LinearModelForSpace)

Call:
lm(formula = Price ~ Space, data = RealEstate)

Residuals:
     Min       1Q   Median       3Q      Max 
-23.7710  -6.0115   0.3997   7.4573  16.4379 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.141566   1.836540   18.59   <2e-16 ***
Space        0.020309   0.001541   13.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.854 on 154 degrees of freedom
Multiple R-squared:  0.5301,	Adjusted R-squared:  0.5271 
F-statistic: 173.8 on 1 and 154 DF,  p-value: < 2.2e-16
```



```
> LinearModelForTransformedSpace <- lm(Price ~ TransformedSpace, data = RealEstate)
> summary(LinearModelForTransformedSpace)

Call:
lm(formula = Price ~ TransformedSpace, data = RealEstate)

Residuals:
    Min      1Q  Median      3Q     Max 
-26.892  -6.500   1.806   6.804  17.961 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         84.769      2.618   32.37   <2e-16 ***
TransformedSpace 16339.011   1446.087   11.30   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.551 on 154 degrees of freedom
Multiple R-squared:  0.4532,	Adjusted R-squared:  0.4497 
F-statistic: 127.7 on 1 and 154 DF,  p-value: < 2.2e-16
```

4.3 Room Variable

For Room variable, below result was obtained after Tukey Transformation Technique was performed.



```
> TransformedRoom <- transformTukey(RealEstate$Room)

    lambda      W Shapiro.p.value
374 -0.675 0.9269       3.885e-07
```

Even after applying Tukey's modification with a lambda value of -0.0675, the Q-Q plot(Figure 4.3) demonstrates that the Room variable is still heavily tailed. To have a more balanced distribution for Room, a few extra adjustments or considerations would be necessary.

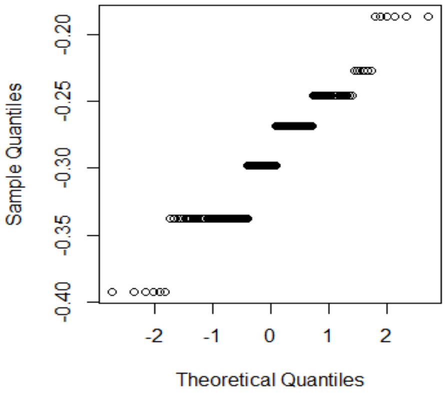
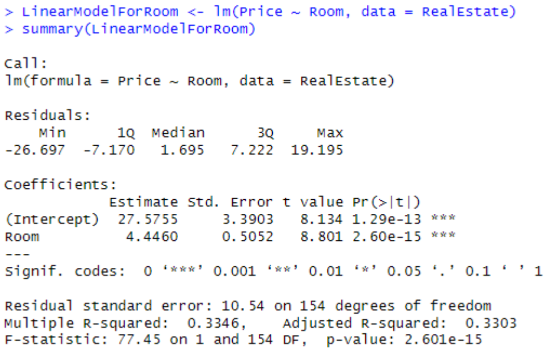


Figure 4.3: Q-Q Plot for Room

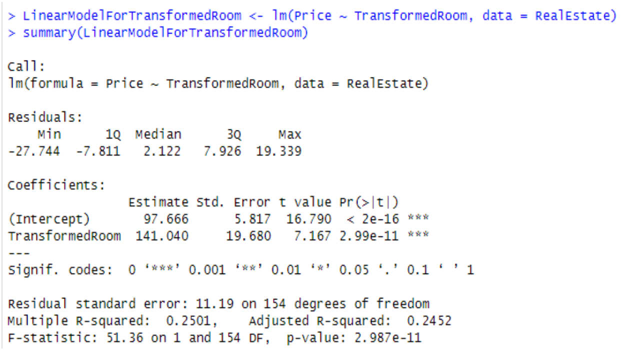
The comparison of the Price ~ Room and Price ~ transformed Room linear models produces the following conclusions:

The Room column remains firmly tailed, unchanged. However, the P-value does not fall substantially, implying that the modified 'Room' column is still relevant in the model.

As differences increases, the Std Error coefficient value rises, indicating fewer precise estimations. The F-statistic is still very significant despite the increase in standard error, indicating that the model is indeed relevant.



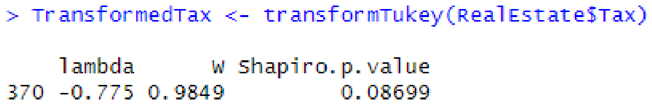
```
> LinearModelForRoom <- lm(Price ~ Room, data = RealEstate)
> summary(LinearModelForRoom)

Call:
lm(formula = Price ~ Room, data = RealEstate)

Residuals:
    Min      1Q  Median      3Q     Max 
-26.697  -7.170   1.695   7.222  19.195 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  27.5755     3.3903   8.134 1.29e-13 ***
Room          4.4460     0.5052   8.801 2.60e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.54 on 154 degrees of freedom
Multiple R-squared:  0.3346,	Adjusted R-squared:  0.3303 
F-statistic: 77.45 on 1 and 154 DF,  p-value: 2.601e-15
```



```
> LinearModelForTransformedRoom <- lm(Price ~ TransformedRoom, data = RealEstate)
> summary(LinearModelForTransformedRoom)

Call:
lm(formula = Price ~ TransformedRoom, data = RealEstate)

Residuals:
    Min      1Q  Median      3Q     Max 
-27.744  -7.811   2.122   7.926  19.339 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       97.666      5.817  16.790  < 2e-16 ***
TransformedRoom  141.040     19.680   7.167 2.99e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.19 on 154 degrees of freedom
Multiple R-squared:  0.2501,	Adjusted R-squared:  0.2452 
F-statistic: 51.36 on 1 and 154 DF,  p-value: 2.987e-11
```

4.4 Tax Variable

For Tax variable, below result was obtained after Tukey Transformation Technique was performed.



```
> TransformedTax <- transformTukey(RealEstate$Tax)

    lambda      W Shapiro.p.value
370 -0.775 0.9849         0.08699
```

Q-Q Plot was plotted for Tax variable.

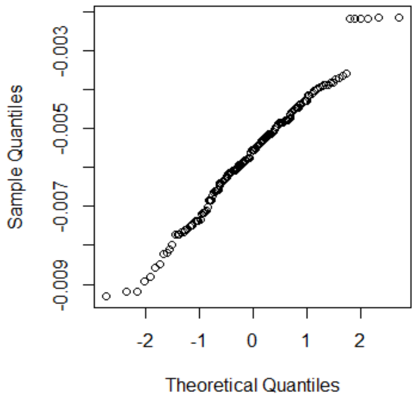
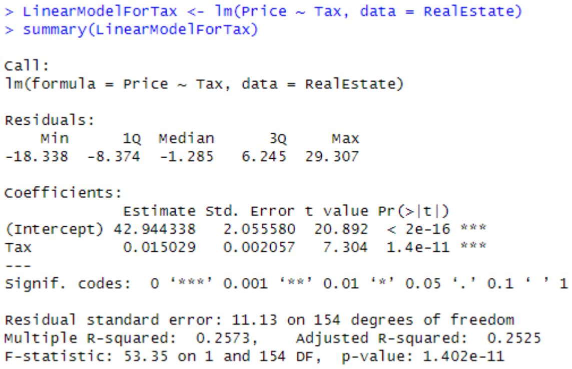
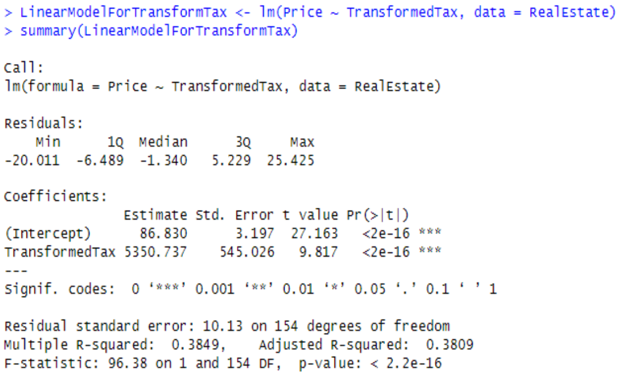


Figure 4.4: Q-Q Plot for Tax

Improved model significance was indicated by an increase in the F-statistic and P-value following the Tukey Transformation on Tax variable. As a result, the model fits TransformedTax more effectively, signalling that it is now of greater significance in explaining the link. as shown in Figure 4.4, causes the data for the Tax variable to become clearly more balanced.



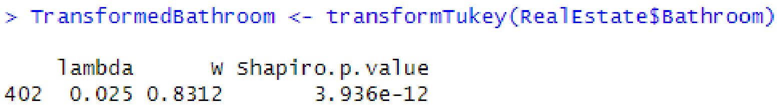
```
> LinearModelForTax <- lm(Price ~ Tax, data = RealEstate)
> summary(LinearModelForTax)

Call:
lm(formula = Price ~ Tax, data = RealEstate)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.338  -8.374  -1.285   6.245  29.307 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 42.944338   2.055580  20.892  < 2e-16 ***
Tax          0.015029   0.002057   7.304  1.4e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.13 on 154 degrees of freedom
Multiple R-squared:  0.2573,	Adjusted R-squared:  0.2525 
F-statistic: 53.35 on 1 and 154 DF,  p-value: 1.402e-11
```



```
> LinearModelForTransformTax <- lm(Price ~ TransformedTax, data = RealEstate)
> summary(LinearModelForTransformTax)

Call:
lm(formula = Price ~ TransformedTax, data = RealEstate)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.011  -6.489  -1.340   5.229  25.425 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      86.830      3.197  27.163   <2e-16 ***
TransformedTax 5350.737    545.026   9.817   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.13 on 154 degrees of freedom
Multiple R-squared:  0.3849,	Adjusted R-squared:  0.3809 
F-statistic: 96.38 on 1 and 154 DF,  p-value: < 2.2e-16
```

4.5 Bathroom Variable

For Bathroom variable, below result was obtained after Tukey Transformation Technique was performed.



```
> TransformedBathroom <- transformTukey(RealEstate$Bathroom)

    lambda      W Shapiro.p.value
402  0.025 0.8312       3.936e-12
```

Q-Q Plot was plotted for Bathroom variable.

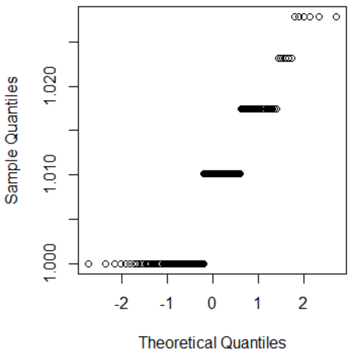
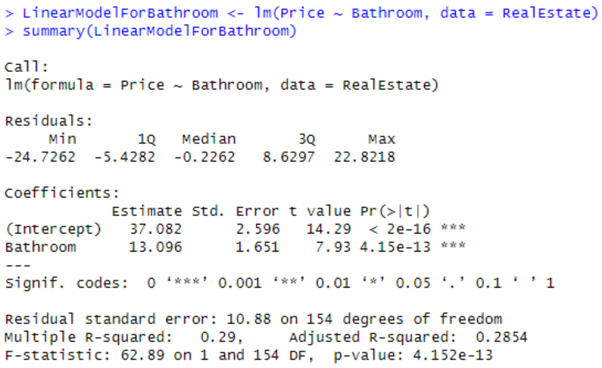
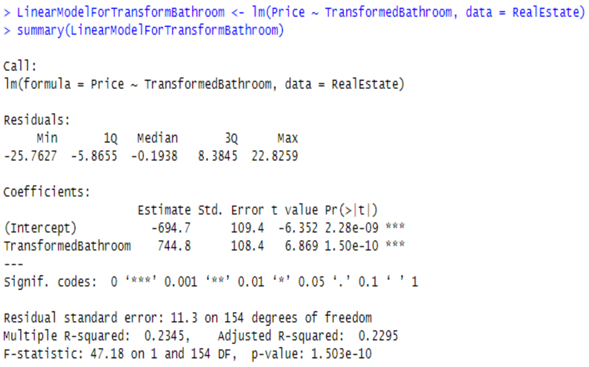


Figure 4.5: Q-Q Plot for Bathroom

As shown in Figure 4.5, Tukey's transformation on the Bedroom (lambda 0.025) lowered the P-value and F-statistic values and clearly normalized the data for the Bedroom variable. Greater with an upsurge in Std Error for the Bathroom upon normalisation, adjusted R square and R square values, together with a significant P-value, support the inclusion of the transformed Bathroom column in the variable selection phase.



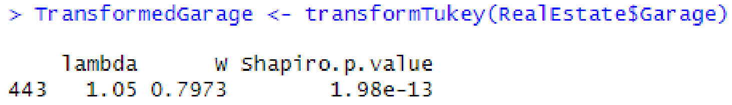
```
> LinearModelForBathroom <- lm(Price ~ Bathroom, data = RealEstate)
> summary(LinearModelForBathroom)

Call:
lm(formula = Price ~ Bathroom, data = RealEstate)

Residuals:
     Min       1Q   Median       3Q      Max 
-24.7262  -5.4282  -0.2262   8.6297  22.8218 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   37.082      2.596   14.29  < 2e-16 ***
Bathroom      13.096      1.651    7.93 4.15e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.88 on 154 degrees of freedom
Multiple R-squared:   0.29,	Adjusted R-squared:  0.2854 
F-statistic: 62.89 on 1 and 154 DF,  p-value: 4.152e-13
```



```
> LinearModelForTransformBathroom <- lm(Price ~ TransformedBathroom, data = RealEstate)
> summary(LinearModelForTransformBathroom)

Call:
lm(formula = Price ~ TransformedBathroom, data = RealEstate)

Residuals:
     Min       1Q   Median       3Q      Max 
-25.7627  -5.8655  -0.1938   8.3845  22.8259 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -694.7      109.4  -6.352 2.28e-09 ***
TransformedBathroom    744.8      108.4   6.869 1.50e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.3 on 154 degrees of freedom
Multiple R-squared:  0.2345,	Adjusted R-squared:  0.2295 
F-statistic: 47.18 on 1 and 154 DF,  p-value: 1.503e-10
```

4.6 Garage Variable

For Garage variable, below result was obtained after Tukey Transformation Technique was performed.



```
> TransformedGarage <- transformTukey(RealEstate$Garage)

    lambda      W Shapiro.p.value
443   1.05 0.7973        1.98e-13
```

Q-Q Plot was plotted for better understanding the Garage variable.

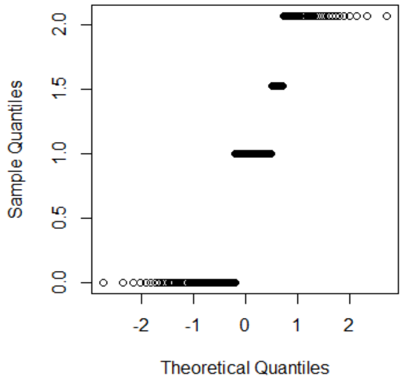
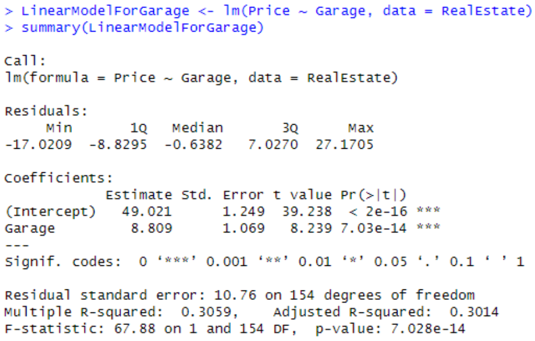
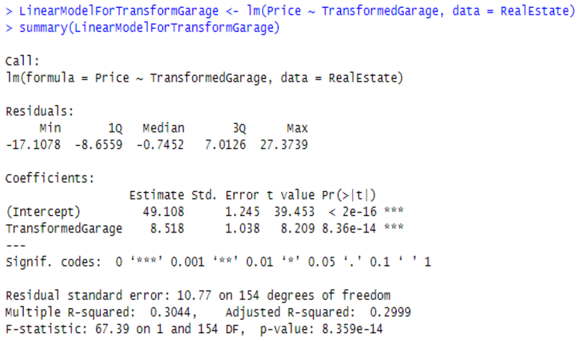


Figure 4.6: Q-Q Plot for Garage

The change that was made to the Garage variable, as shown in Figure 4.6, successfully reduced the dataset's skewness and produced a more noticeably adjusted dataset by normalizing the data for the Space column. In spite of this, the P-value and F-statistic values held steady, indicating that the model's relevance remained. The residual error changed relatively slightly, with values of 10.76 before transformation and 10.77 after transformation. This suggests that the change achieved normalisation without appreciably changing the model's overall fit.



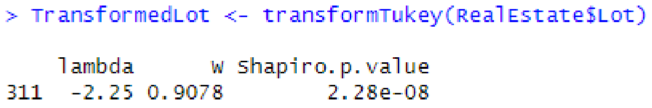
```
> LinearModelForGarage <- lm(Price ~ Garage, data = RealEstate)
> summary(LinearModelForGarage)

Call:
lm(formula = Price ~ Garage, data = RealEstate)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.0209  -8.8295  -0.6382   7.0270  27.1705 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   49.021      1.249  39.238  < 2e-16 ***
Garage         8.809      1.069   8.239 7.03e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.76 on 154 degrees of freedom
Multiple R-squared:  0.3059,	Adjusted R-squared:  0.3014 
F-statistic: 67.88 on 1 and 154 DF,  p-value: 7.028e-14
```



```
> LinearModelForTransformGarage <- lm(Price ~ TransformedGarage, data = RealEstate)
> summary(LinearModelForTransformGarage)

Call:
lm(formula = Price ~ TransformedGarage, data = RealEstate)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.1078  -8.6559  -0.7452   7.0126  27.3739 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         49.108      1.245  39.453  < 2e-16 ***
TransformedGarage    8.518      1.038   8.209 8.36e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.77 on 154 degrees of freedom
Multiple R-squared:  0.3044,	Adjusted R-squared:  0.2999 
F-statistic: 67.39 on 1 and 154 DF,  p-value: 8.359e-14
```

4.7 Lot Variable

For Lot variable, below result was obtained after Tukey Transformation Technique was performed.



```
> TransformedLot <- transformTukey(RealEstate$Lot)

    lambda      W Shapiro.p.value
311  -2.25 0.9078        2.28e-08
```

Q-Q Plot was plotted for Lot variable.

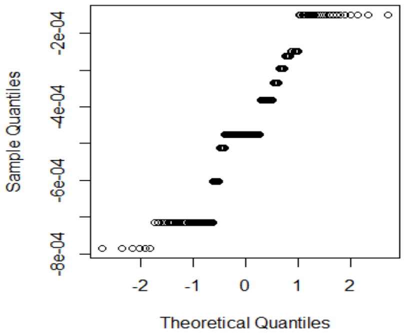
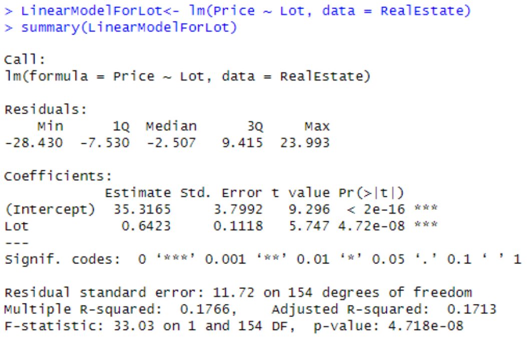
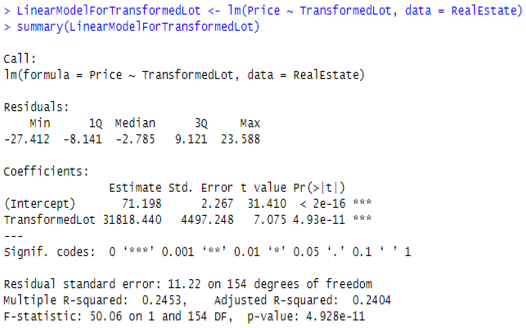


Figure 4.7: Q-Q Plot for Lot

More favourable outcomes were achieved by transforming the Lot variable with a lambda value of -2.25, as shown in Figure 4.7, which causes the data for the Space column to become noticeably balanced. The p-value changed from 10^-8 to 10^-11 as a result of this modification, indicating a higher degree of significance. Also, the F-statistic value increased significantly from 33.03 to 56.06, suggesting that the model's overall significance has improved. These modifications suggest that the transformation at lambda -2.25 had an advantageous effect on the Lot variable, which resulted in a stronger and statistically significant connections in the linear model.



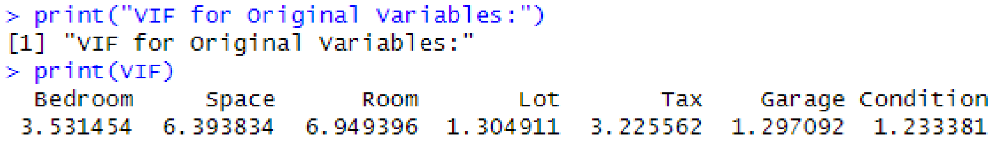
```
> LinearModelForLot<- lm(Price ~ Lot, data = RealEstate)
> summary(LinearModelForLot)

Call:
lm(formula = Price ~ Lot, data = RealEstate)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.430  -7.530  -2.507   9.415  23.993 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  35.3165     3.7992   9.296  < 2e-16 ***
Lot           0.6423     0.1118   5.747 4.72e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.72 on 154 degrees of freedom
Multiple R-squared:  0.1766,	Adjusted R-squared:  0.1713 
F-statistic: 33.03 on 1 and 154 DF,  p-value: 4.718e-08
```



```
> LinearModelForTransformedLot <- lm(Price ~ TransformedLot, data = RealEstate)
> summary(LinearModelForTransformedLot)

Call:
lm(formula = Price ~ TransformedLot, data = RealEstate)

Residuals:
    Min      1Q  Median      3Q     Max 
-27.412  -8.141  -2.785   9.121  23.588 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      71.198      2.267  31.410  < 2e-16 ***
TransformedLot 31818.440   4497.248   7.075 4.93e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.22 on 154 degrees of freedom
Multiple R-squared:  0.2453,	Adjusted R-squared:  0.2404 
F-statistic: 50.06 on 1 and 154 DF,  p-value: 4.928e-11
```

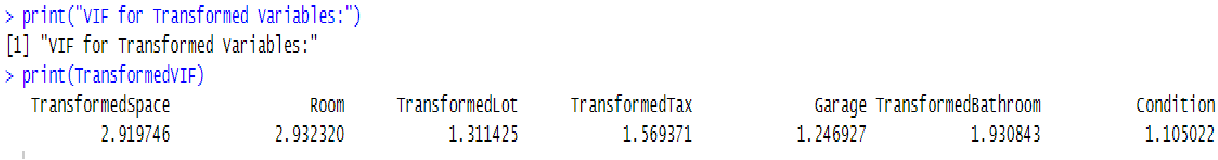
5.Check for Multicollinearity

Multicollinearity makes linear regression models more difficult to evaluate because it presupposes that independent variables are not highly linked. Due to the combination of their implications, multicollinearity makes it more difficult to differentiate between the unique effects of every variable that is independent on the one that is dependent. Bedroom, space, and room scores are high when the Variance Inflation Factor (VIF) is calculated using a linear model with all predictor variables. Reducing multicollinearity and enhancing the reliability of the model may require eliminating one of the Room and Space columns, according to iterative approaches that involve modifications to these columns.



```
> print("VIF for Original Variables:")
[1] "VIF for Original Variables:"
> print(VIF)
  Bedroom     Space      Room       Lot       Tax    Garage Condition 
 3.531454  6.393834  6.949396  1.304911  3.225562  1.297092  1.233381 
```

Based on a continuous iterative approach, the current mix of predictor variables is the optimal choice because it reduces collinearity and enhances the model's stability and dependability.



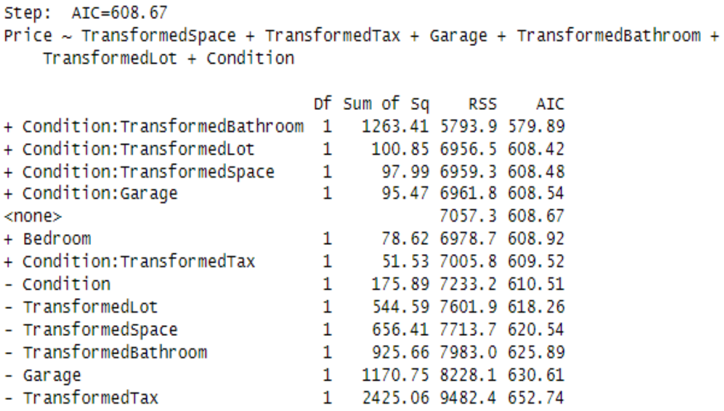
```
> print("VIF for Transformed Variables:")
[1] "VIF for Transformed Variables:"
> print(TransformedVIF)
  TransformedSpace               Room     TransformedLot     TransformedTax             Garage TransformedBathroom          Condition 
          2.919746           2.932320           1.311425           1.569371           1.246927           1.930843           1.105022 
```

6. Suggesting the Best Model

Once the selected predictor variables with lower collinearity have been determined, the next step is to maximize the Akaike Information Criterion (AIC) score in order to determine the best linear model. This approach makes it achievable to represent the link between the predictor and response variables in a dependable and cost-effective way by guaranteeing that the model strikes the appropriate balance between simplicity and goodness of fit.

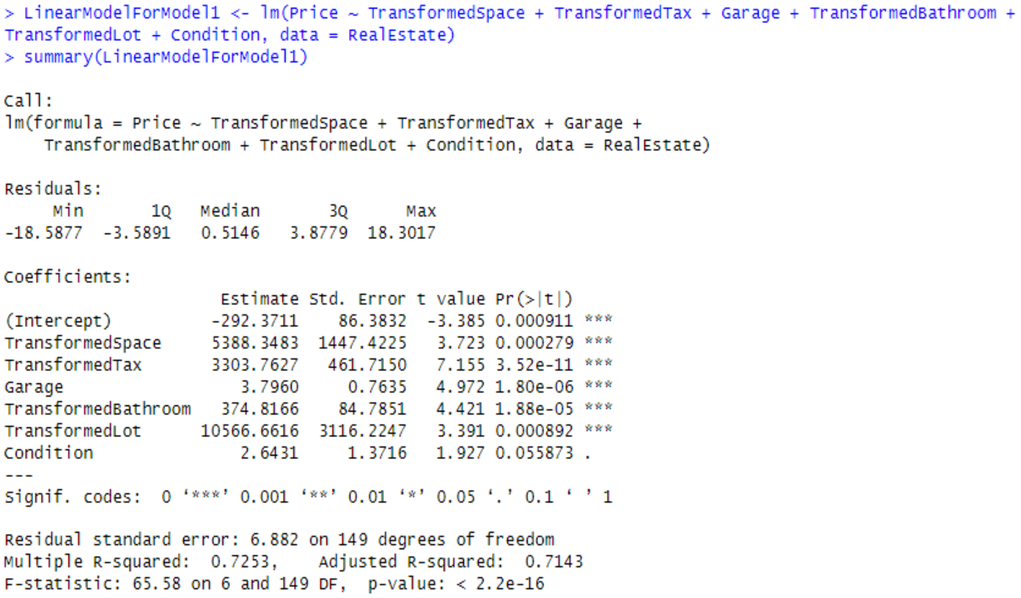
6.1 Model 1

With predictor variables Bedroom, TransformedSpace, TransformedLot, TransformedTax, Garage, Condition, and TransformedBathroom, the final model has an AIC score of 608.67. The process began with the base linear model (Price ~ 1) and used 'both' as the direction for the stepwise regression approach. In this basic model, condition is added as an independent predictor. The summary of this linear model provides significant insights into the manner in which each predictor influences the Price, allowing for a more lucid and understandable representation of the relationship.



```
Step:  AIC=608.67
Price ~ TransformedSpace + TransformedTax + Garage + TransformedBathroom + 
    TransformedLot + Condition

                              Df Sum of Sq    RSS    AIC
+ Condition:TransformedBathroom  1   1263.41 5793.9 579.89
+ Condition:TransformedLot       1    100.85 6956.5 608.42
+ Condition:TransformedSpace     1     97.99 6959.3 608.48
+ Condition:Garage               1     95.47 6961.8 608.54
<none>                                       7057.3 608.67
+ Bedroom                        1     78.62 6978.7 608.92
+ Condition:TransformedTax       1     51.53 7005.8 609.52
- Condition                      1    175.89 7233.2 610.51
- TransformedLot                 1    544.59 7601.9 618.26
- TransformedSpace               1    656.41 7713.7 620.54
- TransformedBathroom            1    925.66 7983.0 625.89
- Garage                         1   1170.75 8228.1 630.61
- TransformedTax                 1   2425.06 9482.4 652.74
```

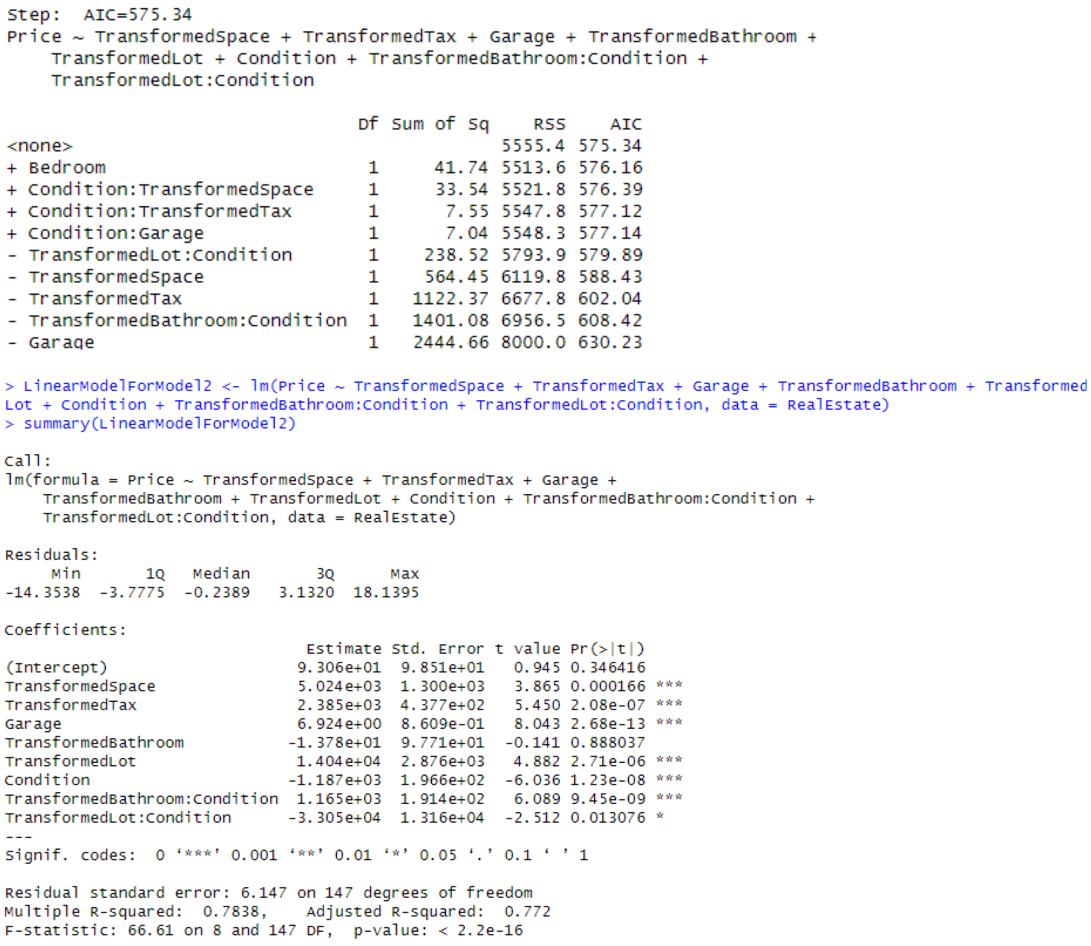
The resulting linear model has predictor variables with p-values that are significantly high, all over 0.05, with the exception of Condition, whose p-value is slightly higher than the critical threshold. However, the total p-value of the model remains below the critical value, indicating that the model remains statistically significant overall. The factors' combined significant contribution strengthens the model's overall explanatory power, even when the individual predictors may not be significant.



```
> LinearModelForModel1 <- lm(Price ~ TransformedSpace + TransformedTax + Garage + TransformedBathroom + 
TransformedLot + Condition, data = RealEstate)
> summary(LinearModelForModel1)

Call:
lm(formula = Price ~ TransformedSpace + TransformedTax + Garage + 
    TransformedBathroom + TransformedLot + Condition, data = RealEstate)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.5877  -3.5891   0.5146   3.8779  18.3017 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -292.3711    86.3832  -3.385 0.000911 ***
TransformedSpace     5388.3483  1447.4225   3.723 0.000279 ***
TransformedTax       3303.7627   461.7150   7.155 3.52e-11 ***
Garage                  3.7960     0.7635   4.972 1.80e-06 ***
TransformedBathroom   374.8166    84.7851   4.421 1.88e-05 ***
TransformedLot      10566.6616  3116.2247   3.391 0.000892 ***
Condition               2.6431     1.3716   1.927 0.055873 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.882 on 149 degrees of freedom
Multiple R-squared:  0.7253,	Adjusted R-squared:  0.7143 
F-statistic: 65.58 on 6 and 149 DF,  p-value: < 2.2e-16
```

6.2 Model 2

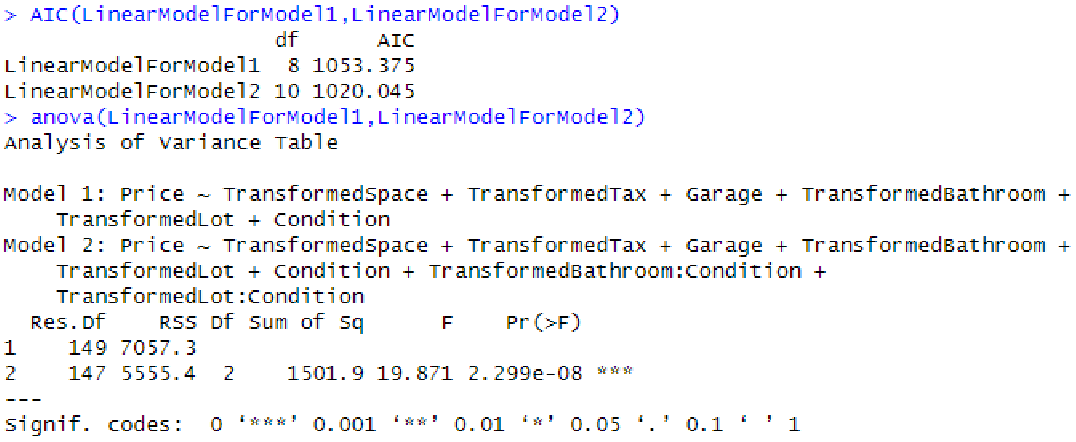
Taking the Condition variable to be considered as a categorical predictor in the stepwise regression with 'both' as the direction produces a more optimized model with an AIC score of 575.34. This resulting model's Condition variable has a substantial relationship with TransformedBathroom and TransformedLot, as demonstrated by the highly significant p-values for these variables. The Adjusted R-square and F-statistic values have both remarkably slightly improved in comparison to the previous model. An analysis of the ANOVA scores for both models demonstrates that the modified model with the categorical representation of the Condition variable performs better and is more significant.



```
Step:  AIC=575.34
Price ~ TransformedSpace + TransformedTax + Garage + TransformedBathroom +
    TransformedLot + Condition + TransformedBathroom:Condition +
    TransformedLot:Condition

                               Df Sum of Sq    RSS    AIC
<none>                                      5555.4 575.34
+ Bedroom                       1     41.74 5513.6 576.16
+ Condition:TransformedSpace    1     33.54 5521.8 576.39
+ Condition:TransformedTax      1      7.55 5547.8 577.12
+ Condition:Garage              1      7.04 5548.3 577.14
- TransformedLot:Condition      1    238.52 5793.9 579.89
- TransformedSpace              1    564.45 6119.8 588.43
- TransformedTax                1   1122.37 6677.8 602.04
- TransformedBathroom:Condition 1   1401.08 6956.5 608.42
- Garage                        1   2444.66 8000.0 630.23

> LinearModelForModel2 <- lm(Price ~ TransformedSpace + TransformedTax + Garage + TransformedBathroom + Transformed
Lot + Condition + TransformedBathroom:Condition + TransformedLot:Condition, data = RealEstate)
> summary(LinearModelForModel2)

Call:
lm(formula = Price ~ TransformedSpace + TransformedTax + Garage +
    TransformedBathroom + TransformedLot + Condition + TransformedBathroom:Condition +
    TransformedLot:Condition, data = RealEstate)

Residuals:
     Min       1Q   Median       3Q      Max
-14.3538  -3.7775  -0.2389   3.1320  18.1395

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   9.306e+01  9.851e+01   0.945 0.346416
TransformedSpace              5.024e+03  1.300e+03   3.865 0.000166 ***
TransformedTax                2.385e+03  4.377e+02   5.450 2.08e-07 ***
Garage                        6.924e+00  8.609e-01   8.043 2.68e-13 ***
TransformedBathroom          -1.378e+01  9.771e+01  -0.141 0.888037
TransformedLot                1.404e+04  2.876e+03   4.882 2.71e-06 ***
Condition                    -1.187e+03  1.966e+02  -6.036 1.23e-08 ***
TransformedBathroom:Condition  1.165e+03  1.914e+02   6.089 9.45e-09 ***
TransformedLot:Condition     -3.305e+04  1.316e+04  -2.512 0.013076 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.147 on 147 degrees of freedom
Multiple R-squared:  0.7838,	Adjusted R-squared:  0.772
F-statistic: 66.61 on 8 and 147 DF,  p-value: < 2.2e-16
```

6.3 Best Model

With a smaller Residual Sum of Squares (RSS) than the Model 1, Model 2 is more capable to explain the wide range of response variables. It also incorporates interactions between the Condition column and the Bathroom and Lot.



```
> AIC(LinearModelForModel1,LinearModelForModel2)
                     df      AIC
LinearModelForModel1  8 1053.375
LinearModelForModel2 10 1020.045
> anova(LinearModelForModel1,LinearModelForModel2)
Analysis of Variance Table

Model 1: Price ~ TransformedSpace + TransformedTax + Garage + TransformedBathroom +
    TransformedLot + Condition
Model 2: Price ~ TransformedSpace + TransformedTax + Garage + TransformedBathroom +
    TransformedLot + Condition + TransformedBathroom:Condition +
    TransformedLot:Condition
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    149 7057.3
2    147 5555.4  2    1501.9 19.871 2.299e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

6.4 Model Assumptions after Transformation

Below Figures illustrates the assumptions of a linear regression model after the modifications have been implemented to the data, allowing for the verification that the assumptions remain accurate.

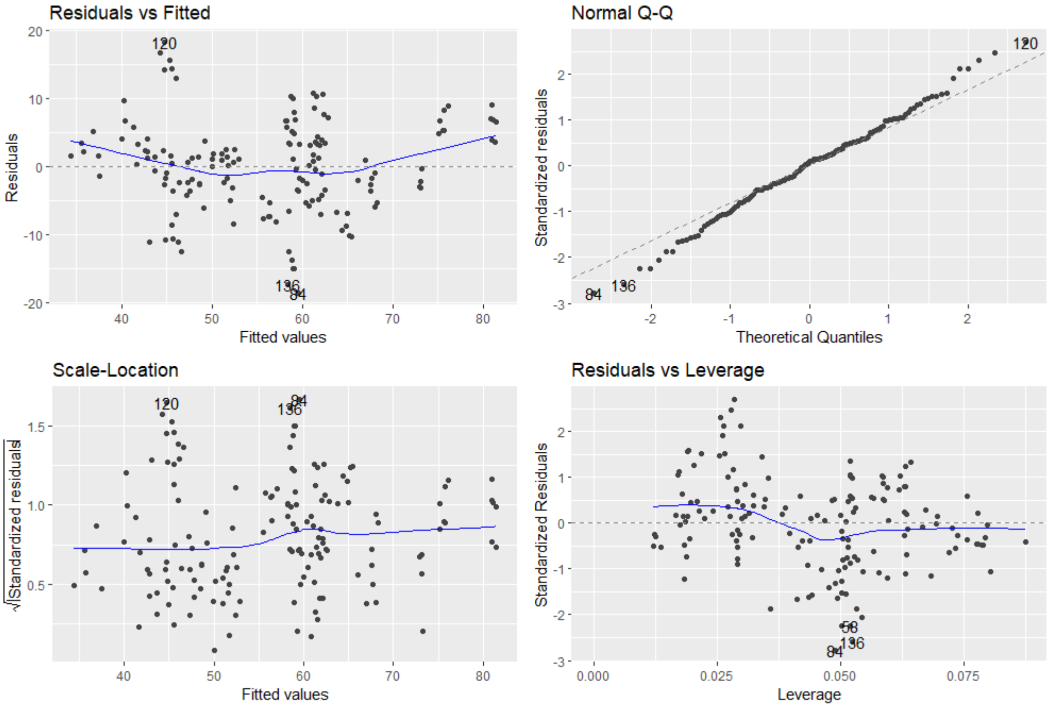


Figure 6.4.a: Linear Model For Model 1

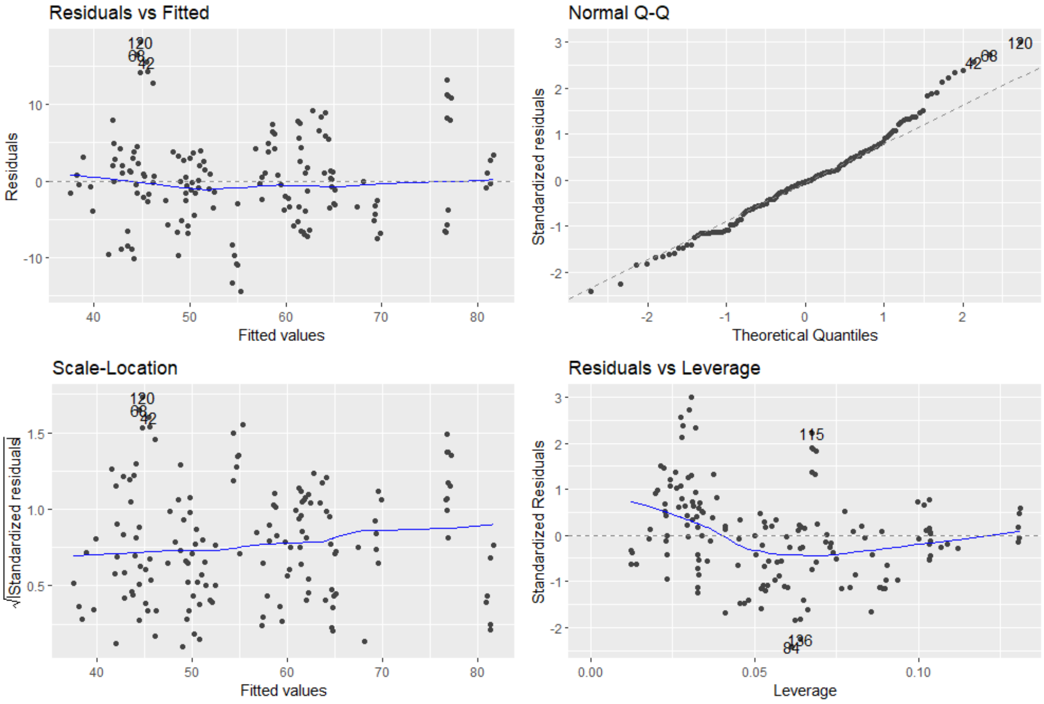
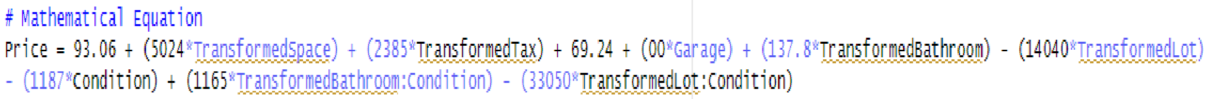


Figure 6.4.b: Linear Model For Model 2

7. Mathematical Equation

To estimate the property price, we will input the following equation:



```
# Mathematical Equation
Price = 93.06 + (5024*TransformedSpace) + (2385*TransformedTax) + 69.24 + (00*Garage) + (137.8*TransformedBathroom) - (14040*TransformedLot)
- (1187*Condition) + (1165*TransformedBathroom:Condition) - (33050*TransformedLot:Condition)
```

This displays the models' Residual vs. Fitted plot.

You should enter values for the TransformedSpace, TransformedTax, Garage, TransformedBathroom, TransformedLot, and Condition variables for the property you are interested in. The equation will then be used to predict the price of the property.

8. Conclusion

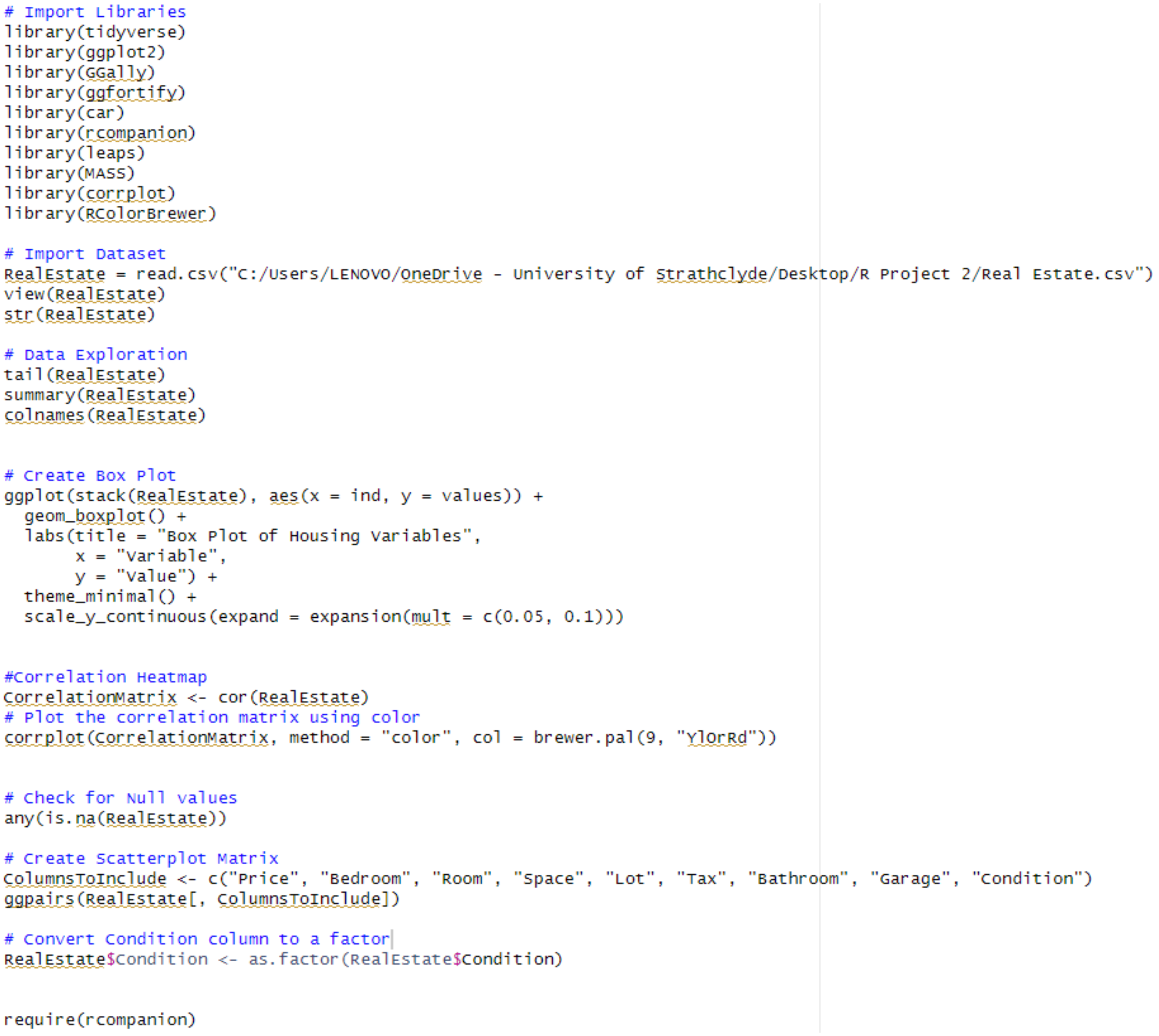
The resulting linear model, which predicts 78.38% of the variation in the Price variable, has an adequate level of explanation power with an R-squared value of 0.7838. The model seems appropriate to the dataset and captures quite a bit of the fluctuations in housing prices, as demonstrated by its strong R-squared value.

The variables that are highly connected with higher transformed pricing(TransformedPrice) are TransformedSpace, TransformedTax, Garage, and TransformedLot.

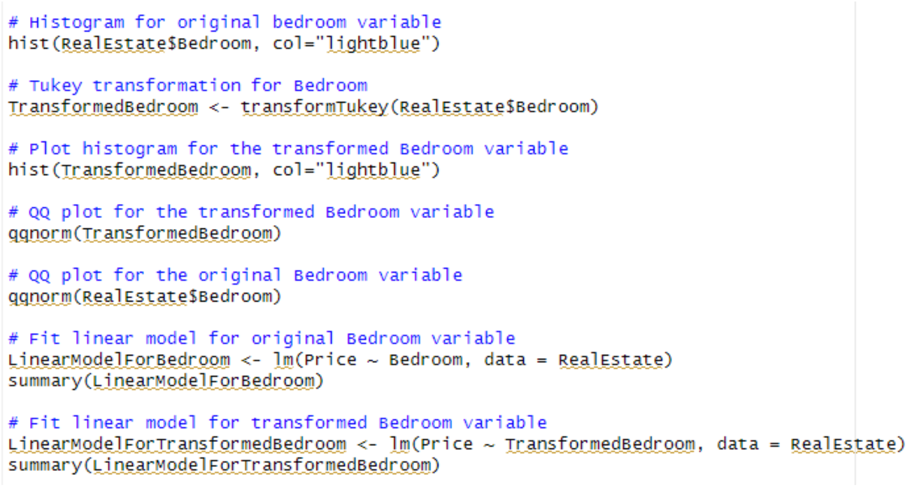
These elements positively affect the model, indicating how significant an impact they have on house values. The model also discovers two important interactions between Condition and other predictors, namely TransformedBathroom:Condition and TransformedLot:Condition. The fact that these interactions are statistically significant at the 0.05 level indicates how important they are to home prices.

A Appendices

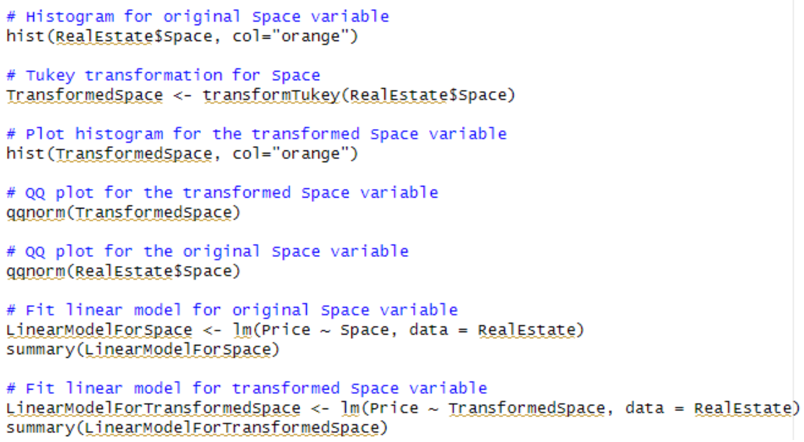
Code:



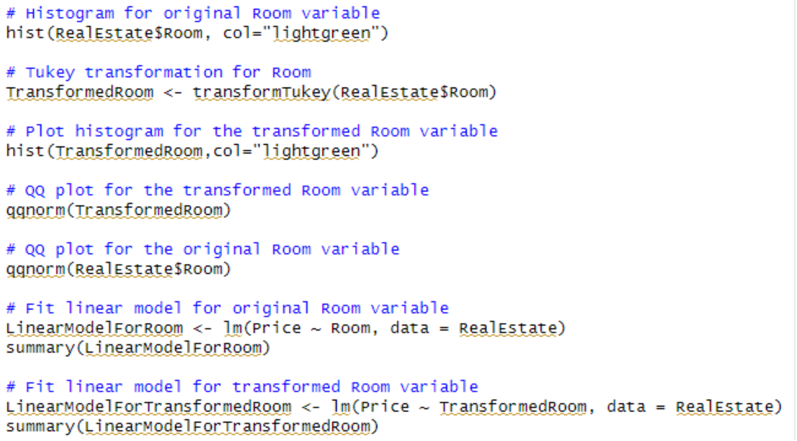
```
# Import Libraries
library(tidyverse)
library(ggplot2)
library(GGally)
library(ggfortify)
library(car)
library(rcompanion)
library(leaps)
library(MASS)
library(corrplot)
library(RColorBrewer)

# Import Dataset
RealEstate = read.csv("C:/Users/LENOVO/OneDrive - University of Strathclyde/Desktop/R Project 2/Real Estate.csv")
View(RealEstate)
str(RealEstate)

# Data Exploration
tail(RealEstate)
summary(RealEstate)
colnames(RealEstate)


# Create Box Plot
ggplot(stack(RealEstate), aes(x = ind, y = values)) +
  geom_boxplot() +
  labs(title = "Box Plot of Housing Variables",
       x = "Variable",
       y = "Value") +
  theme_minimal() +
  scale_y_continuous(expand = expansion(mult = c(0.05, 0.1)))


#Correlation Heatmap
CorrelationMatrix <- cor(RealEstate)
# Plot the correlation matrix using color
corrplot(CorrelationMatrix, method = "color", col = brewer.pal(9, "YlOrRd"))


# Check for Null values
any(is.na(RealEstate))

# Create Scatterplot Matrix
ColumnsToInclude <- c("Price", "Bedroom", "Room", "Space", "Lot", "Tax", "Bathroom", "Garage", "Condition")
ggpairs(RealEstate[, ColumnsToInclude])

# Convert Condition column to a factor
RealEstate$Condition <- as.factor(RealEstate$Condition)


require(rcompanion)
```



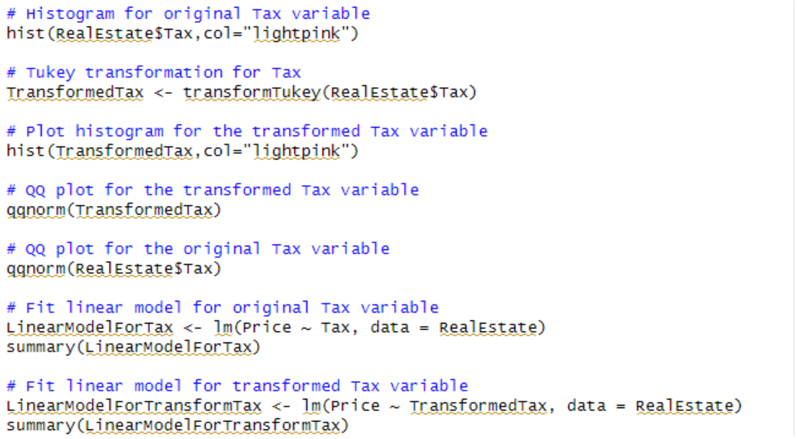
```
# Histogram for original bedroom variable
hist(RealEstate$Bedroom, col="lightblue")

# Tukey transformation for Bedroom
TransformedBedroom <- transformTukey(RealEstate$Bedroom)

# Plot histogram for the transformed Bedroom variable
hist(TransformedBedroom, col="lightblue")

# QQ plot for the transformed Bedroom variable
qqnorm(TransformedBedroom)

# QQ plot for the original Bedroom variable
qqnorm(RealEstate$Bedroom)

# Fit linear model for original Bedroom variable
LinearModelForBedroom <- lm(Price ~ Bedroom, data = RealEstate)
summary(LinearModelForBedroom)

# Fit linear model for transformed Bedroom variable
LinearModelForTransformedBedroom <- lm(Price ~ TransformedBedroom, data = RealEstate)
summary(LinearModelForTransformedBedroom)
```



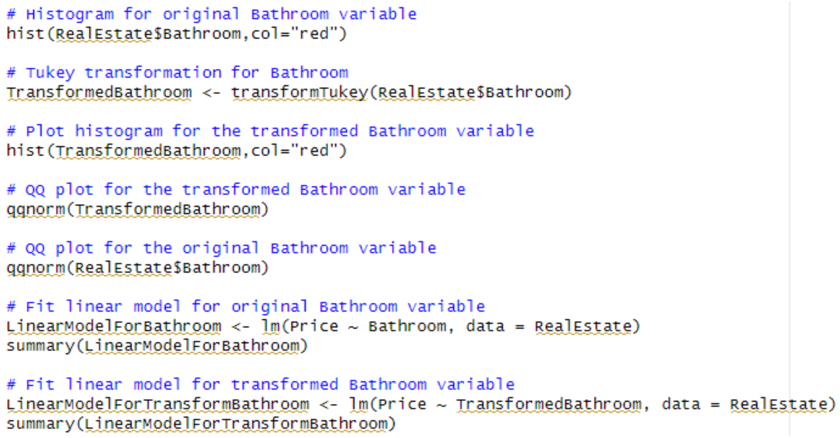
```
# Histogram for original Space variable
hist(RealEstate$Space, col="orange")

# Tukey transformation for Space
TransformedSpace <- transformTukey(RealEstate$Space)

# Plot histogram for the transformed Space variable
hist(TransformedSpace, col="orange")

# QQ plot for the transformed Space variable
qqnorm(TransformedSpace)

# QQ plot for the original Space variable
qqnorm(RealEstate$Space)

# Fit linear model for original Space variable
LinearModelForSpace <- lm(Price ~ Space, data = RealEstate)
summary(LinearModelForSpace)

# Fit linear model for transformed Space variable
LinearModelForTransformedSpace <- lm(Price ~ TransformedSpace, data = RealEstate)
summary(LinearModelForTransformedSpace)
```



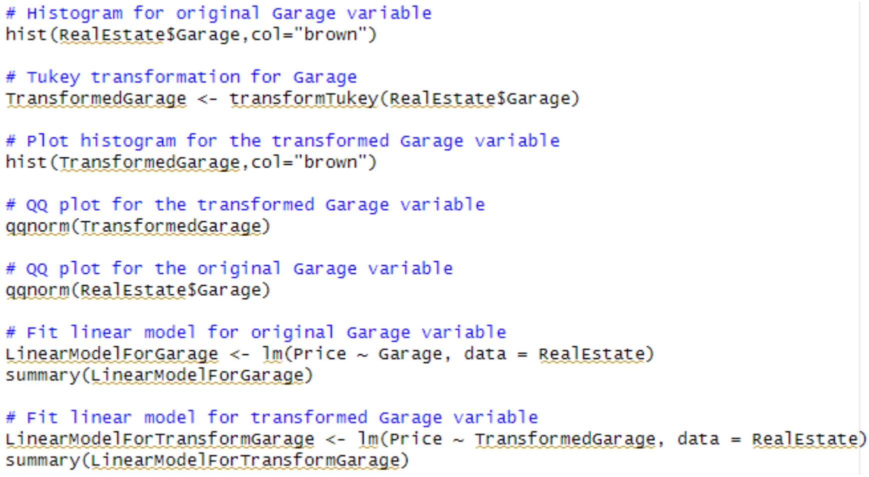
```
# Histogram for original Room variable
hist(RealEstate$Room, col="lightgreen")

# Tukey transformation for Room
TransformedRoom <- transformTukey(RealEstate$Room)

# Plot histogram for the transformed Room variable
hist(TransformedRoom,col="lightgreen")

# QQ plot for the transformed Room variable
qqnorm(TransformedRoom)

# QQ plot for the original Room variable
qqnorm(RealEstate$Room)

# Fit linear model for original Room variable
LinearModelForRoom <- lm(Price ~ Room, data = RealEstate)
summary(LinearModelForRoom)

# Fit linear model for transformed Room variable
LinearModelForTransformedRoom <- lm(Price ~ TransformedRoom, data = RealEstate)
summary(LinearModelForTransformedRoom)
```



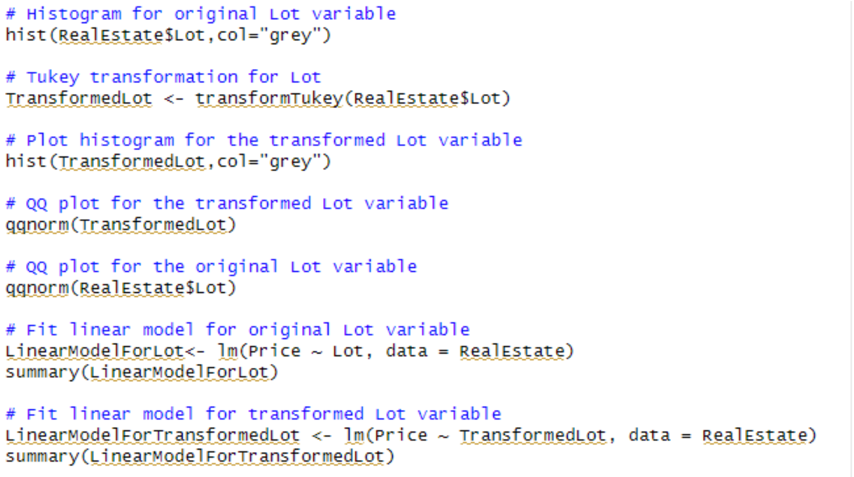
```
# Histogram for original Tax variable
hist(RealEstate$Tax,col="lightpink")

# Tukey transformation for Tax
TransformedTax <- transformTukey(RealEstate$Tax)

# Plot histogram for the transformed Tax variable
hist(TransformedTax,col="lightpink")

# QQ plot for the transformed Tax variable
qqnorm(TransformedTax)

# QQ plot for the original Tax variable
qqnorm(RealEstate$Tax)

# Fit linear model for original Tax variable
LinearModelForTax <- lm(Price ~ Tax, data = RealEstate)
summary(LinearModelForTax)

# Fit linear model for transformed Tax variable
LinearModelForTransformTax <- lm(Price ~ TransformedTax, data = RealEstate)
summary(LinearModelForTransformTax)
```



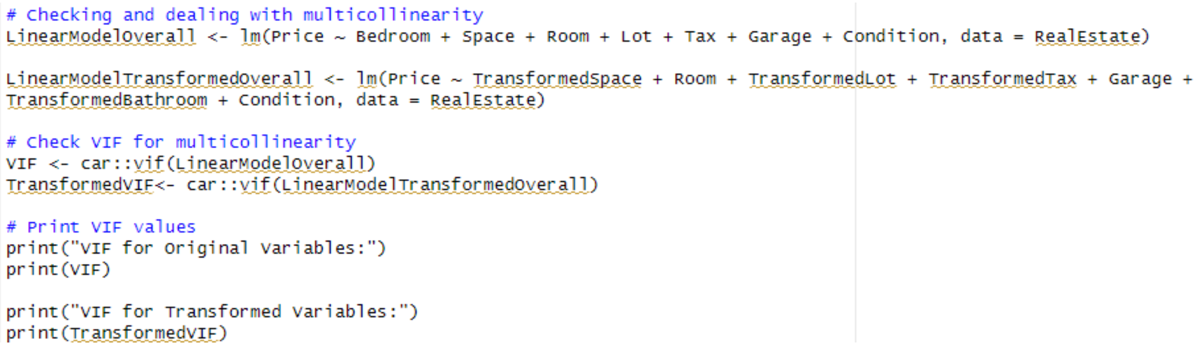
```
# Histogram for original Bathroom variable
hist(RealEstate$Bathroom,col="red")

# Tukey transformation for Bathroom
TransformedBathroom <- transformTukey(RealEstate$Bathroom)

# Plot histogram for the transformed Bathroom variable
hist(TransformedBathroom,col="red")

# QQ plot for the transformed Bathroom variable
qqnorm(TransformedBathroom)

# QQ plot for the original Bathroom variable
qqnorm(RealEstate$Bathroom)

# Fit linear model for original Bathroom variable
LinearModelForBathroom <- lm(Price ~ Bathroom, data = RealEstate)
summary(LinearModelForBathroom)

# Fit linear model for transformed Bathroom variable
LinearModelForTransformBathroom <- lm(Price ~ TransformedBathroom, data = RealEstate)
summary(LinearModelForTransformBathroom)
```



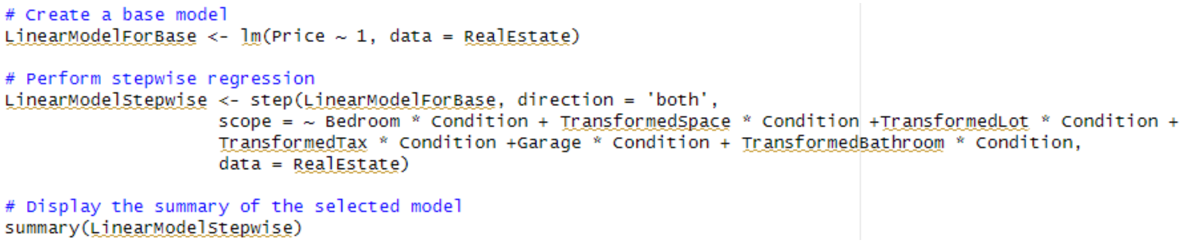
```
# Histogram for original Garage variable
hist(RealEstate$Garage,col="brown")

# Tukey transformation for Garage
TransformedGarage <- transformTukey(RealEstate$Garage)

# Plot histogram for the transformed Garage variable
hist(TransformedGarage,col="brown")

# QQ plot for the transformed Garage variable
qqnorm(TransformedGarage)

# QQ plot for the original Garage variable
qqnorm(RealEstate$Garage)

# Fit linear model for original Garage variable
LinearModelForGarage <- lm(Price ~ Garage, data = RealEstate)
summary(LinearModelForGarage)

# Fit linear model for transformed Garage variable
LinearModelForTransformGarage <- lm(Price ~ TransformedGarage, data = RealEstate)
summary(LinearModelForTransformGarage)
```



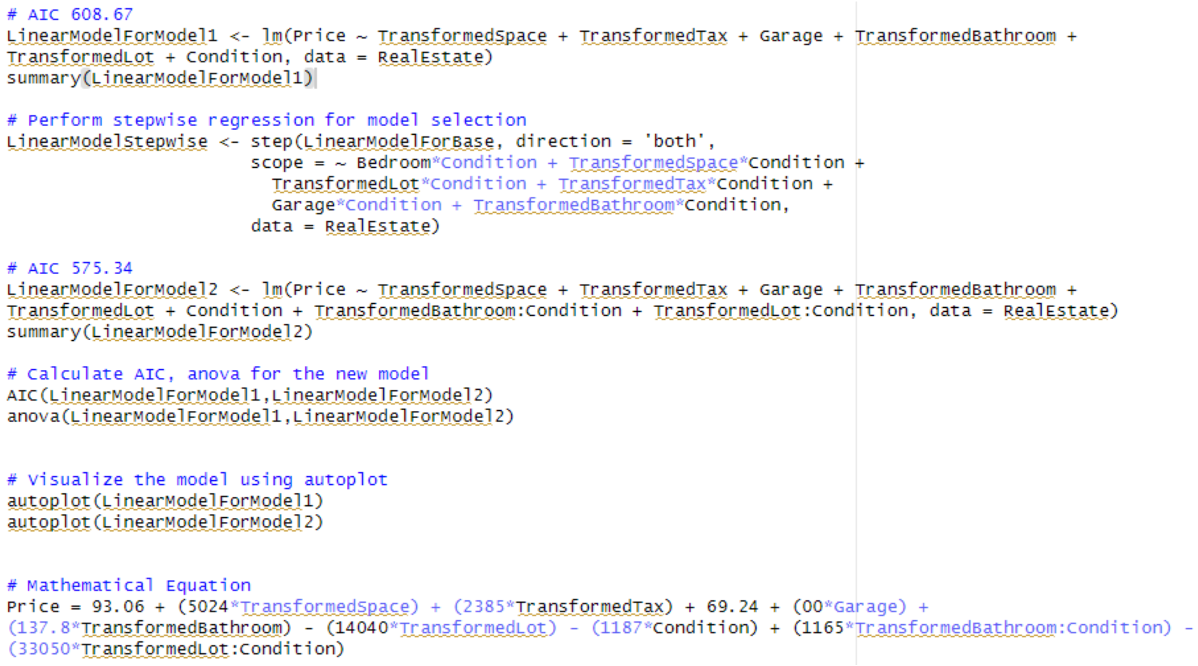
```
# Histogram for original Lot variable
hist(RealEstate$Lot,col="grey")

# Tukey transformation for Lot
TransformedLot <- transformTukey(RealEstate$Lot)

# Plot histogram for the transformed Lot variable
hist(TransformedLot,col="grey")

# QQ plot for the transformed Lot variable
qqnorm(TransformedLot)

# QQ plot for the original Lot variable
qqnorm(RealEstate$Lot)

# Fit linear model for original Lot variable
LinearModelForLot<- lm(Price ~ Lot, data = RealEstate)
summary(LinearModelForLot)

# Fit linear model for transformed Lot variable
LinearModelForTransformedLot <- lm(Price ~ TransformedLot, data = RealEstate)
summary(LinearModelForTransformedLot)
```



```
# Checking and dealing with multicollinearity
LinearModelOverall <- lm(Price ~ Bedroom + Space + Room + Lot + Tax + Garage + Condition, data = RealEstate)

LinearModelTransformedOverall <- lm(Price ~ TransformedSpace + Room + TransformedLot + TransformedTax + Garage +
TransformedBathroom + Condition, data = RealEstate)

# Check VIF for multicollinearity
VIF <- car::vif(LinearModelOverall)
TransformedVIF<- car::vif(LinearModelTransformedOverall)

# Print VIF values
print("VIF for Original Variables:")
print(VIF)

print("VIF for Transformed Variables:")
print(TransformedVIF)
```



```
# Create a base model
LinearModelForBase <- lm(Price ~ 1, data = RealEstate)

# Perform stepwise regression
LinearModelStepwise <- step(LinearModelForBase, direction = 'both',
                      scope = ~ Bedroom * Condition + TransformedSpace * Condition +TransformedLot * Condition +
                      TransformedTax * Condition +Garage * Condition + TransformedBathroom * Condition,
                      data = RealEstate)

# Display the summary of the selected model
summary(LinearModelStepwise)
```



```
# AIC 608.67
LinearModelForModel1 <- lm(Price ~ TransformedSpace + TransformedTax + Garage + TransformedBathroom +
TransformedLot + Condition, data = RealEstate)
summary(LinearModelForModel1)

# Perform stepwise regression for model selection
LinearModelStepwise <- step(LinearModelForBase, direction = 'both',
                       scope = ~ Bedroom*Condition + TransformedSpace*Condition +
                         TransformedLot*Condition + TransformedTax*Condition +
                         Garage*Condition + TransformedBathroom*Condition,
                       data = RealEstate)

# AIC 575.34
LinearModelForModel2 <- lm(Price ~ TransformedSpace + TransformedTax + Garage + TransformedBathroom +
TransformedLot + Condition + TransformedBathroom:Condition + TransformedLot:Condition, data = RealEstate)
summary(LinearModelForModel2)

# Calculate AIC, anova for the new model
AIC(LinearModelForModel1,LinearModelForModel2)
anova(LinearModelForModel1,LinearModelForModel2)


# Visualize the model using autoplot
autoplot(LinearModelForModel1)
autoplot(LinearModelForModel2)


# Mathematical Equation
Price = 93.06 + (5024*TransformedSpace) + (2385*TransformedTax) + 69.24 + (00*Garage) +
(137.8*TransformedBathroom) - (14040*TransformedLot) - (1187*Condition) + (1165*TransformedBathroom:Condition) -
(33050*TransformedLot:Condition)
```