M916 – Project 2

Hrutuja Patkar

MM916: Data Analytics in R

12/4/23

Table of Contents

Table of Figures

List of Tables

1. Introduction:

The realm of real estate pricing is greatly influenced by interplay of intricate variables and external factors. This study embarks on an exploratory journey through a dataset encompassing detailed information about housing prices in Chicago and the various factors that are consider having an influence in the real estate landscape.

Within this dataset containing records of 157 houses in Chicago, various aspects such as property size, room numbers, amenities, tax figures, and the property condition of the houses are investigated. The main objective of the study is to construct an optimal linear model capable of predicting house prices in Chicago by different statistical methodology and data modelling. It delves into finding out the variables that plays pivotal role in determining the house prices, their individual and collective impacts and thereby using statistical techniques to prove its significance.

It further investigates to explore any non-linear relationship between different variables and house pricing as linear model might not adequately capture the intricate patterns between them. Furthermore, the study seeks to unravel any significant interactions between the condition of a house and other predictors, potentially shaping the overarching influence on housing prices. These hypotheses are statistically proven using appropriate statistically measures.

2. Materials and methods:

a. Descriptive Analysis

Table 1: Data Description

Variables	Name
Target	Price (numeric)
Independent	Bedroom (numeric), Space (numeric), Room (numeric), Lot (numeric), Tax(numeric), Bathroom(numeric), Garage(numeric), Condition (categorical)

b. Exploratory Analysis:

Exploring relationships between target and independent variables:

The Figure 1 assess the relationship between numerical independent variables and the target variable(Price).

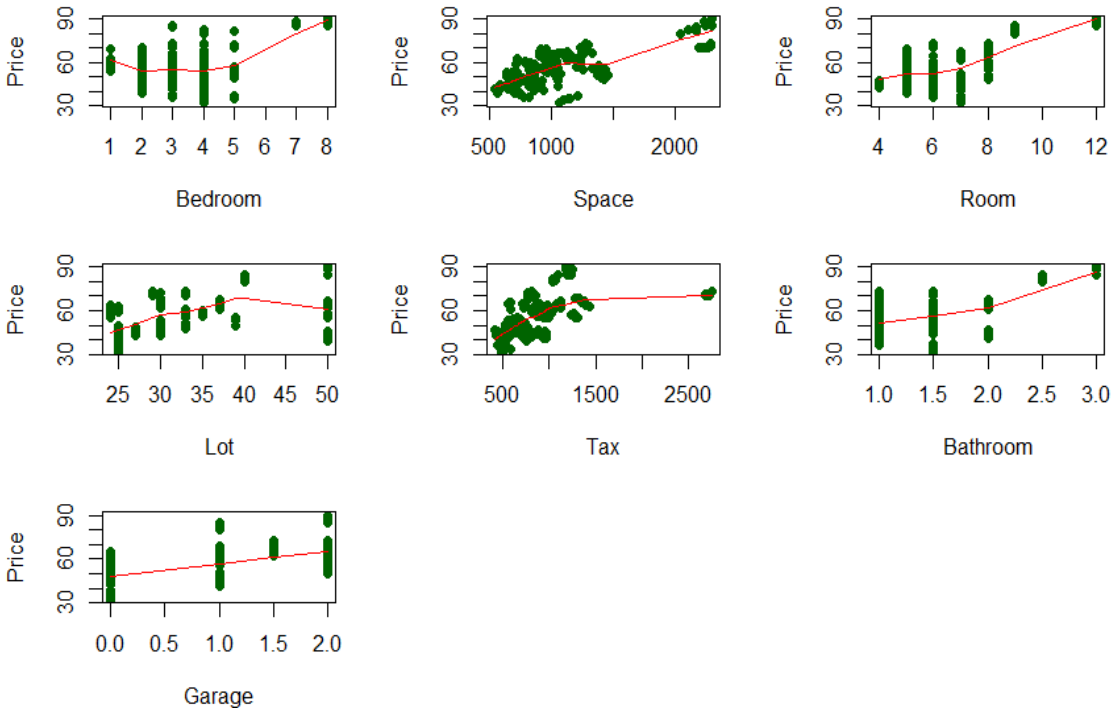

Figure 1: Target vs Independent Variables

Evaluating each predictor's relationship with target variable reveals a linear pattern among them. The Space and Room variable exhibits strong linear relationship with the target variable. It strongly suggests the absence of non-linear relationship between target variables and any other variables visualized.

The below Figure 2 depicts the interaction between the categorical variable "Condition" (0 for bad and 1 for good condition) and the independent numeric variables indicating how the variables differ based on house condition. The condition of the house seems to have no effect on the Bedroom and Room variable, as both depict similar density curve. However, to validate this hypothesis rigorously, statistical analysis is conducted to determine whether these variables indeed show no significant difference between the two conditions.

It also illustrates how the Price target variable is affected by the condition of the house.

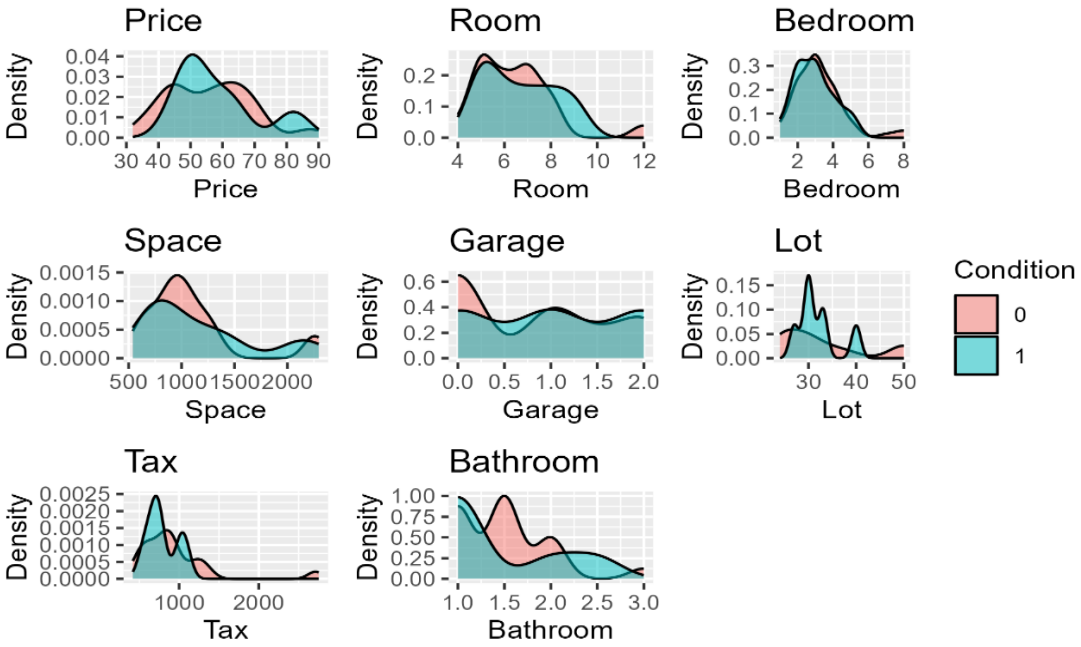

Figure 2: Density plot of all the variables

2.3 Correlation Analysis:

The correlation matrix provides a snapshot of the relationships between independent variables and target, showcasing the strength and direction of their linear associations.

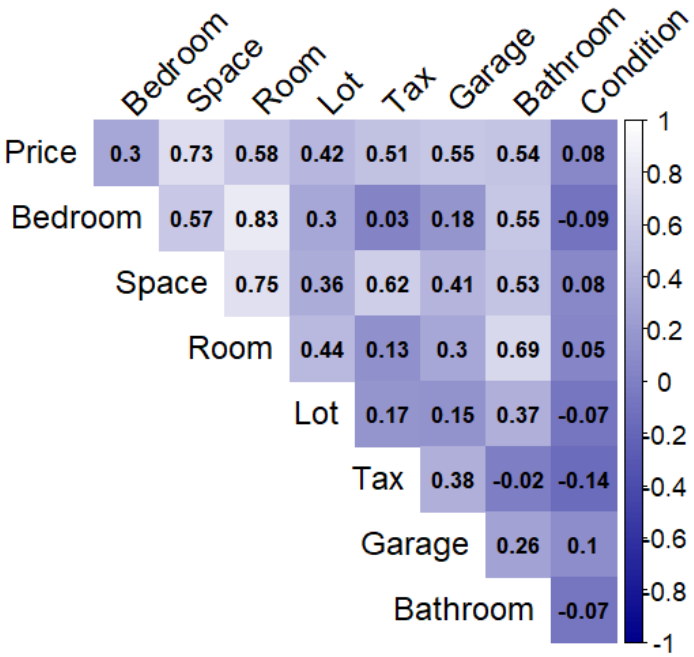

Figure 3: Correlation Matrix of the dataset

All the numeric variables show strong relationships (>0.5) with the target variable. Space variables have the strongest correlation with the Price, which generally is observed to be true in real scenarios.

The categorical variable Condition don't directly display accurate relationship with the target variable in a correlation matrix.

Also, it is observed that the Bedroom and Room, Space and Room, Room and Bathroom, Space and Tax are highly correlated pairs, which, when high, might signal multicollinearity issues, influencing model stability and interpretability.

Using this correlation matrix, a linear model is fitted considering all variables having strong correlation (>0.5).

```
real_est_mod <- lm(Price~Bedroom+Space+Room+Lot+Tax+Bathroom+Garage+Condition, data=real_est)
```

2.4 Model Assumptions:

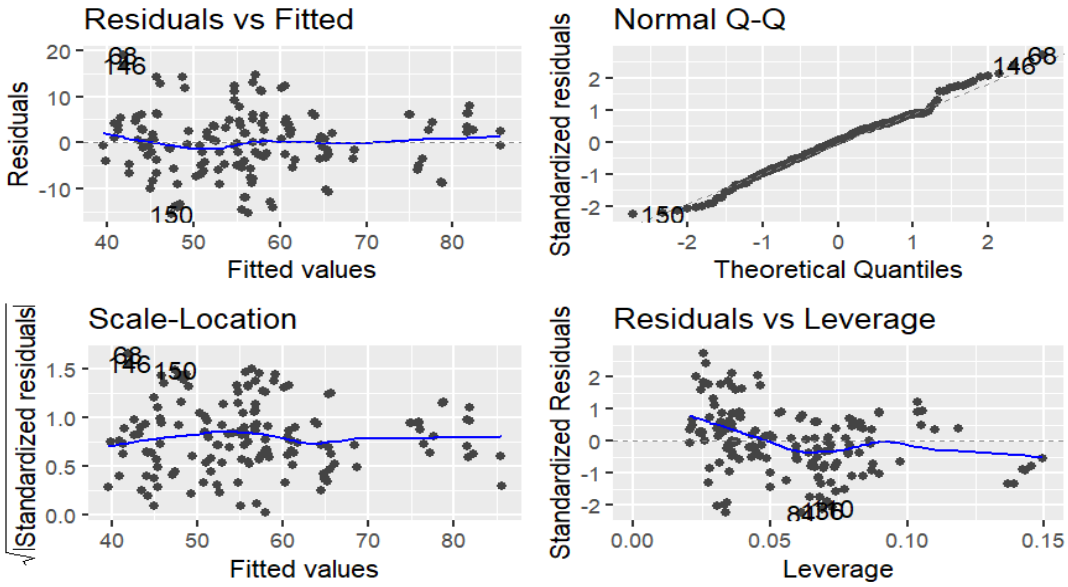

Figure 4: Model assumption plot

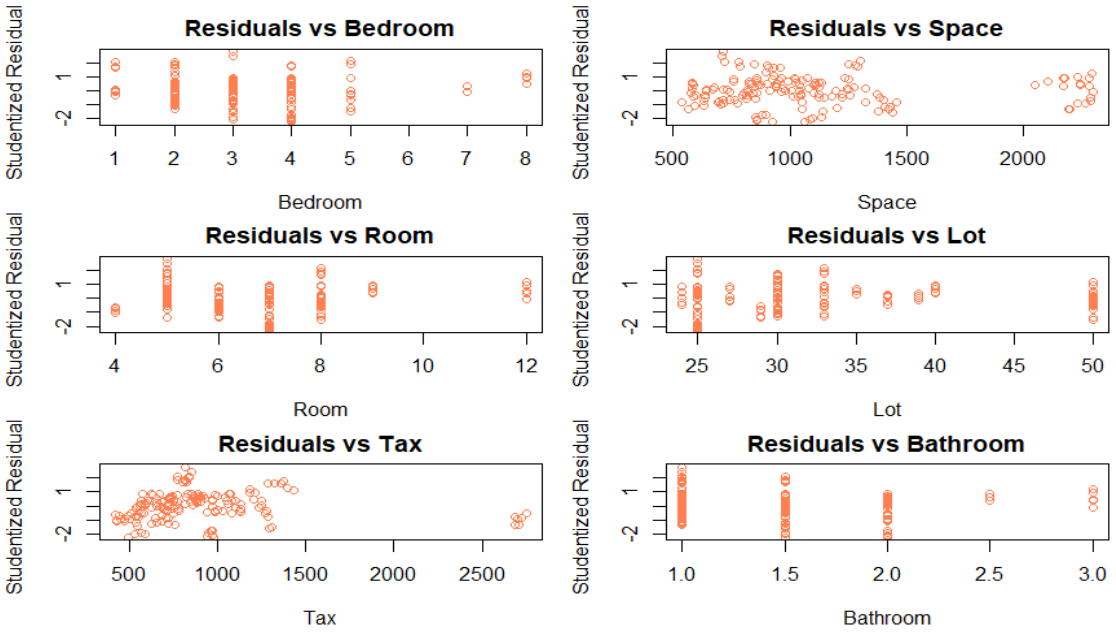

Figure 5: Residuals vs all the independent variables

The assumptions are:

• normality of errors: The normal Q-Q plot in Figure 5 how a perfect straight line, and suggest that the model covers normality.

• Constant error variance: The Residuals vs Fitted in Figure 5 plots suggests no perfect spread of the points, and hence the variance of error is same throughout (referred as homoscedasicity). Also, there is no trend observed in the Scale-Location plot.

• Independence of errors: The Figure 5 suggests non-random patterns visible in the independent variables, suggesting independence of errors. In order to satisfy the assumptions necessary transformation are taken using the Box-Cox transformation technique.

2.5 Transformation of the independent variables:

In this analysis, a Box-Cox transformation was employed to address issues associated with varying scale among the independent variables. The Box-Cox transformation applies a power transformation to the data finding optimal lambda values, enabling the model to better adhere to the assumptions of linear regression.

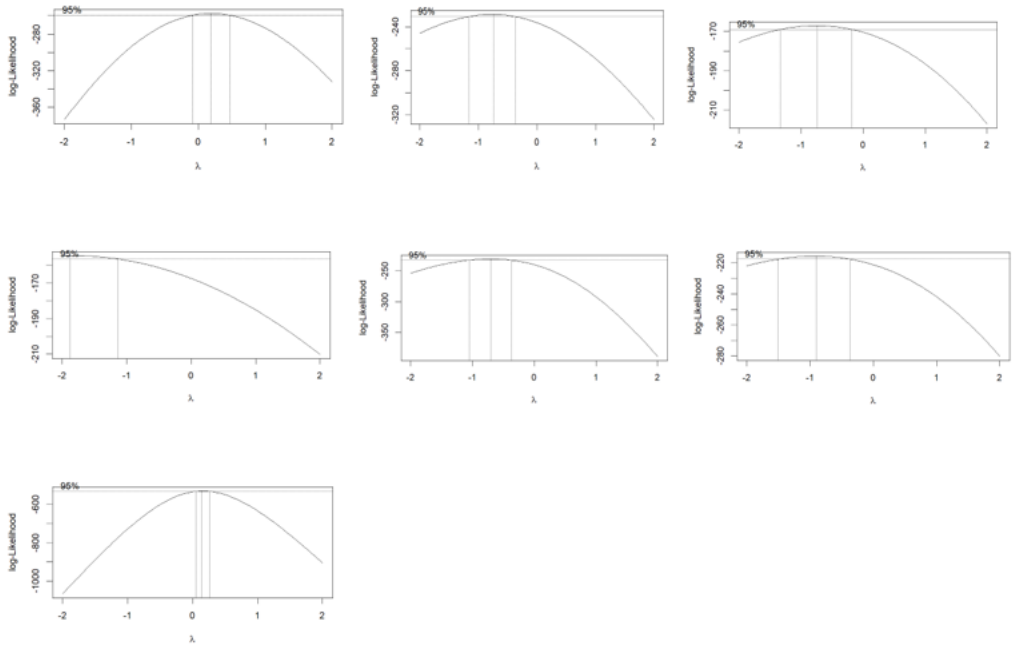

Figure 6: Box-Cox plot for the independent variables

Table 2: Lambda values and transformations

Variable	Optimal Lambda value	transformation
Bedroom	0.2	log(y)
Space	-0.7	y -1
Room	-0.7	y -2
Lot	-1.9	y -1
Tax	-0.7	y -1
Bathroom	-0.9	y -1
Garage (a constant 0.01 is used to handle non-positive values)	0.1	log(y + 0.01)

The transformed variables offer a more normalized distribution and improved linearity, aligning with the assumptions of linear regression and potentially resulting in a more robust predictive model. However, the variable transformation of Garage does not improve its distribution.

After the assumptions, the distribution of the variables and their residuals plot was checked again to ensure that the robust model building.

Also, multicollinearity is assessed using VIF factor, and ensured that no variables are causing multicollinearity issues.

```
library(car)

vif_results             <-             car::vif(transformed_real_est1_mod)
vif_results

##    Bedroom      Space       Room        Lot        Tax   Bathroom     Garage
Condition
##   2.243345   3.041366   3.866373   1.280855   1.590211   1.770577   1.286242
1.184875
```

2.6 Categorical Interaction

Categorical variables, such as the 'Condition' variable in our dataset, required special treatment to incorporate into regression models.

The approach involves exploring interactions between categorical and numerical predictors to capture nuanced relationships that might affect house prices. The preliminary exploratory analysis suggests the other variables which might have interactions with the Condition. However, to achieve more accurate model it is necessary to study its interaction using necessary statistical techniques. This is achieved by creating interaction terms between 'Condition' and other numeric predictors (e.g., 'Space', 'Lot', 'Tax', 'Bathroom', 'Garage') within the regression model.

```
categorical_full_mod <- lm(Price ~ Bedroom + Space + Room + Lot + Tax +
Bathroom + Garage + Condition + Bedroom:Condition + Space:Condition +
Room:Condition +   Lot:Condition + Tax:Condition + Bathroom:Condition
+Garage:Condition, data = transformed_real_est1)
```

A model is further developed removing variables removing the insignificant variables from the categorical model above.

```
categorical_full_mod2 <- lm(Price ~ Bedroom + Space + Room + Lot + Tax +
Bathroom   +   Garage   +   Condition   +     Bathroom:Condition,   data   =
transformed_real_est1)
```

In the context of linear regression models, testing the hypothesis of interaction between categorical and numeric variables is examined by the significance of the interaction terms in the model.

2.7 Variable Selection:

When constructing predictive models, variable selection plays a pivotal role in determining the most influential predictors for the target variable, in this case, house prices in Chicago. Various methods are employed to sift through the predictors and discern the most pertinent ones for model construction. One such method utilized here is backward elimination.

2.7.1 Backward elimination:

Backward elimination is a systematic approach used for variable selection in regression analysis. In context of this study, backward elimination is implemented after successfully finding significant categorical interaction between the variables.

```
full_mod <- lm(Price ~ Bedroom + Space + Room + Lot + Tax + Bathroom +
Garage + Condition +  Bathroom:Condition, data = transformed_real_est1)
drop1(full_mod, scope = ~Bedroom + Space + Room + Lot + Tax + Bathroom +
Garage + Condition +  Bathroom:Condition, data = transformed_real_est1,
test = "F")
```

```
## Single term deletions
##
## Model:
## Price ~ Bedroom + Space + Room + Lot + Tax + Bathroom + Garage +
##     Condition + Bathroom:Condition
##                    Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                          5451.4 574.39
## Bedroom             1    201.72 5653.2 578.06  5.4024 0.0214885 *
## Space               1    391.76 5843.2 583.22 10.4922 0.0014844 **
## Room                1     15.20 5466.6 572.82  0.4071 0.5244254
## Lot                 1    502.87 5954.3 586.15 13.4677 0.0003395 ***
## Tax                 1   1560.66 7012.1 611.66 41.7975 1.426e-09 ***
## Bathroom            1    238.89 5690.3 579.08  6.3979 0.0124871 *
## Garage              1   1763.26 7214.7 616.11 47.2236 1.705e-10 ***
## Condition           1    483.56 5935.0 585.65 12.9507 0.0004372 ***
## Bathroom:Condition  1    686.22 6137.7 590.89 18.3783 3.274e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By testing the significance of each variable, it is found that the Room variable does not have significant (p-value > 0.05) impact and hence it is dropped. In the second step, all variables are found to be significant (p-value<0.05) and hence the process is ended.

```
#removing insignificant Room variable
back_mod1 <- lm(Price ~ Bedroom + Space  + Lot + Tax + Bathroom + Garage
+ Condition +  Bathroom:Condition, data = transformed_real_est1)
drop1(back_mod1, scope = ~Bedroom + Space  + Lot + Tax + Bathroom +
Garage + Condition +  Bathroom:Condition, data = transformed_real_est1,
test = "F")
```

```
## Single term deletions
##
## Model:
## Price ~ Bedroom + Space + Lot + Tax + Bathroom + Garage + Condition +
##     Bathroom:Condition
##                    Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                          5466.6 572.82
## Bedroom             1    215.35 5682.0 576.85  5.7909 0.0173508 *
## Space               1    636.27 6102.9 588.00 17.1095 5.913e-05 ***
## Lot                 1    521.30 5987.9 585.03 14.0181 0.0002592 ***
## Tax                 1   1571.38 7038.0 610.24 42.2552 1.170e-09 ***
## Bathroom            1    251.24 5717.9 577.83  6.7559 0.0102963 *
## Garage              1   1756.11 7222.7 614.28 47.2225 1.674e-10 ***
## Condition           1    481.34 5948.0 583.99 12.9435 0.0004379 ***
## Bathroom:Condition  1    697.51 6164.1 589.56 18.7563 2.735e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This method aids in simplifying the model without compromising its predictive accuracy and interpretability.

2.7.2 Stepwise method:

Stepwise regression, as implemented in the step function, is another variable selection technique used to refine a regression model by adding or removing predictors based on certain statistical criteria that best explains the variation in the target variable while avoiding overfitting.

The model output displays the final set of predictors chosen by the stepwise procedure.

```
library(car)
step_mod <- step(categorical_full_mod2)
```

```
## Start:  AIC=574.39
## Price ~ Bedroom + Space + Room + Lot + Tax + Bathroom + Garage +
##     Condition + Bathroom:Condition
##
##                      Df Sum of Sq    RSS    AIC
## - Room                1     15.20 5466.6 572.82
## <none>                            5451.4 574.39
## - Bedroom             1    201.72 5653.2 578.06
## - Space               1    391.76 5843.2 583.22
## - Lot                 1    502.87 5954.3 586.15
## - Bathroom:Condition  1    686.22 6137.7 590.89
## - Tax                 1   1560.66 7012.1 611.66
## - Garage              1   1763.26 7214.7 616.11
##
## Step:  AIC=572.82
## Price ~ Bedroom + Space + Lot + Tax + Bathroom + Garage + Condition +
```

```
##     Bathroom:Condition
##
##                       Df Sum of Sq    RSS    AIC
## <none>                               5466.6 572.82
## - Bedroom              1    215.35 5682.0 576.85
## - Lot                  1    521.30 5987.9 585.03
## - Space                1    636.27 6102.9 588.00
## - Bathroom:Condition   1    697.51 6164.1 589.56
## - Tax                  1   1571.38 7038.0 610.24
## - Garage               1   1756.11 7222.7 614.28
```

The model obtained from both the process are equivalent as they contain same set of variables, suggesting their accuracy.

3. Results and Discussion:

3.1 Model Summary

3.1.1 Initial Model:

In the initial model obtained through correlation analysis, relationships among variables to gauge their potential impact on house prices was explored and an initial linear model was developed.

```
summary(real_est_mod)
```

```
## Call:
## lm(formula = Price ~ Bedroom + Space + Room + Lot + Tax + Bathroom +
##     Garage + Condition, data = real_est)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4413  -4.6347   0.2271   4.0636  19.1376
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.645736   3.593389   4.911 2.39e-06 ***
## Bedroom     -3.337156   0.800508  -4.169 5.21e-05 ***
## Space        0.007354   0.003334   2.205 0.028973 *
## Room         2.631539   0.915874   2.873 0.004665 **
## Lot          0.178596   0.077876   2.293 0.023247 *
## Tax          0.006290   0.002602   2.417 0.016877 *
## Bathroom     6.221515   1.672201   3.721 0.000282 ***
## Garage       3.746902   0.819471   4.572 1.02e-05 ***
```

```
## Condition    1.195937   1.573975   0.760 0.448579
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.093 on 147 degrees of freedom
## Multiple R-squared:  0.7121, Adjusted R-squared:  0.6965
## F-statistic: 45.46 on 8 and 147 DF,  p-value: < 2.2e-16
```

The model itself portrays a commendable explanatory power, elucidating approximately 71.21% (Multiple R-squared) of the variability in house prices. This statistical strength is validated by the highly significant 'F-statistic' and its associated p-value, affirming the model's robustness and relevance.

While this model serves as a good starting point, further refinement, considering model assumptions, variable importance, the categorical variable – condition interaction and model diagnostics, might enhance its predictive power.

3.1.2 Transformed Model:

```
summary(transformed_real_est1_mod)
```

```
##
## Call:
## lm(formula = Price ~ Bedroom + Space + Room + Lot + Tax + Bathroom +
##     Garage + Condition, data = transformed_real_est)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.1945  -4.1567   0.3208   4.4134  18.3093
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.071e+02  5.614e+00  19.073  < 2e-16 ***
## Bedroom     -5.079e+00  2.022e+00  -2.512 0.013096 *
## Space       -1.061e+04  2.900e+03  -3.657 0.000354 ***
## Room        -3.792e+01  3.001e+01  -1.264 0.208297
## Lot         -5.291e+03  1.522e+03  -3.476 0.000669 ***
## Tax         -1.045e+04  1.675e+03  -6.240 4.43e-09 ***
## Bathroom    -9.629e+00  3.307e+00  -2.912 0.004153 **
## Garage       1.189e+00  2.618e-01   4.542 1.15e-05 ***
## Condition    1.759e+00  1.464e+00   1.202 0.231468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.077 on 147 degrees of freedom
## Multiple R-squared:  0.7135, Adjusted R-squared:  0.6979
## F-statistic: 45.76 on 8 and 147 DF,  p-value: < 2.2e-16
```

The transformed model showcases notable improvements with enhanced explanatory power (Adjusted R-squared: 0.6979) and statistical significance in 'Bedroom,' 'Space,' 'Lot,' 'Tax,' and 'Garage' variables. While 'Room' and 'Condition' categorical variable exhibit less impact on house prices, indicating potential areas for further exploration. The model continues to refine our understanding, although further adjustments may be beneficial to enhance its predictive capability.

3.1.3 Categorical Interaction model:

```
summary(categorical_full_mod)
## Call:
## lm(formula = Price ~ Bedroom + Space + Room + Lot + Tax + Bathroom +
##     Garage + Condition + Bedroom:Condition + Space:Condition +
##     Room:Condition + Lot:Condition + Tax:Condition +
Bathroom:Condition +
##     Garage:Condition, data = transformed_real_est1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7199  -2.9679  -0.1006   2.6348  17.5687
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            80.856      6.212  13.015  < 2e-16 ***
## Bedroom                -4.302      2.177  -1.976 0.050119 .
## Space               -8804.856   2902.992  -3.033 0.002886 **
## Room                  -18.662     30.895  -0.604 0.546775
## Lot                 -5484.633   1371.313  -4.000 0.000102 ***
## Tax                 -8800.938   1597.426  -5.509 1.67e-07 ***
## Bathroom                2.923      1.690   1.729 0.085926 .
## Garage                  5.948      1.017   5.847 3.37e-08 ***
## Condition             -21.068     37.583  -0.561 0.575989
## Bedroom:Condition       3.595      4.356   0.825 0.410567
## Space:Condition     16942.577  15850.082   1.069 0.286942
## Room:Condition        -94.348    143.127  -0.659 0.510856
## Lot:Condition        8359.952  22913.667   0.365 0.715777
## Tax:Condition       -7209.601  10884.137  -0.662 0.508807
## Bathroom:Condition     13.400      5.694   2.353 0.019994 *
## Garage:Condition       -1.778      5.141  -0.346 0.730000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.095 on 140 degrees of freedom
## Multiple R-squared:  0.7976, Adjusted R-squared:  0.7759
## F-statistic: 36.78 on 15 and 140 DF,  p-value: < 2.2e-16
```

The inclusion of categorical variables, notably Condition and its interactions with other variables, in the regression model significantly enhances its predictive capacity. The adjusted R-

squared value, a measure of model goodness-of-fit, notably elevates from 0.6965 in the initial model to 0.7759 in the enhanced categorical model. This substantial increase indicates that approximately 77.59% of the variance in the Price variable is now explained by the predictors, showcasing a more robust and improved fit.

The hypothesis that the categorical variable has significant interaction with the numeric variables is tested.

The coefficients of Bedroom: Condition , Space: Condition , Room: Condition, Lot: Condition , Tax: Condition, Garage: Condition have p-value > 0.05 and hence acts as a evidence to reject the null hypothesis, indicating a significant interaction between these numeric variable and the categorical variable 'Condition'.

However, Bathroom: Condition depicts significant interaction between bathroom and condition of house (p-value<0.05).

```
summary(categorical_full_mod2)

##
## Call:
## lm(formula = Price ~ Bedroom + Space + Room + Lot + Tax + Bathroom +
##     Garage + Condition + Bathroom:Condition, data =
transformed_real_est1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5126  -3.1212  -0.2117   2.8973  16.9065
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            78.7179     5.6311  13.979  < 2e-16 ***
## Bedroom                -4.2158     1.8138  -2.324 0.021488 *
## Space               -8088.2057  2497.0057  -3.239 0.001484 **
## Room                  -16.4498    25.7803  -0.638 0.524425
## Lot                 -4872.0106  1327.5815  -3.670 0.000340 ***
## Tax                 -9337.8839  1444.3537  -6.465 1.43e-09 ***
## Bathroom                3.8621     1.5269   2.529 0.012487 *
## Garage                  5.3626     0.7804   6.872 1.70e-10 ***
## Condition             -13.5504     3.7654  -3.599 0.000437 ***
## Bathroom:Condition     10.4020     2.4264   4.287 3.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.111 on 146 degrees of freedom
## Multiple R-squared:  0.7878, Adjusted R-squared:  0.7748
## F-statistic: 60.24 on 9 and 146 DF,  p-value: < 2.2e-16
```

Furthermore, removing the insignificant variables, anova test is used to compare both the models, and a p-value > 0.05 is observed, suggesting that the additional interaction terms not significantly improve the models fit.

```
anova(categorical_full_mod, categorical_full_mod2)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Bedroom + Space + Room + Lot + Tax + Bathroom +
Garage +
##     Condition + Bedroom:Condition + Space:Condition + Room:Condition +
##     Lot:Condition + Tax:Condition + Bathroom:Condition +
Garage:Condition
## Model 2: Price ~ Bedroom + Space + Room + Lot + Tax + Bathroom +
Garage +
##     Condition + Bathroom:Condition
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    140 5200.1
## 2    146 5451.4 -6   -251.29 1.1275 0.3494
```

3.1.4 Variable Selection model:

```
summary(back_mod1)
```

```
##
## Call:
## lm(formula = Price ~ Bedroom + Space + Lot + Tax + Bathroom +
##     Garage + Condition + Bathroom:Condition, data =
transformed_real_est1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3454  -3.4630  -0.2461   3.0903  16.6619
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           76.1229     3.8870  19.584  < 2e-16 ***
## Bedroom               -3.5423     1.4720  -2.406 0.017351 *
## Space              -8893.6329  2150.1086  -4.136 5.91e-05 ***
## Lot                -4943.0541  1320.2340  -3.744 0.000259 ***
## Tax                -9365.6622  1440.7832  -6.500 1.17e-09 ***
## Bathroom               3.9459     1.5181   2.599 0.010296 *
## Garage                 5.3500     0.7785   6.872 1.67e-10 ***
## Condition            -13.5181     3.7574  -3.598 0.000438 ***
## Bathroom:Condition    10.4754     2.4188   4.331 2.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6.098 on 147 degrees of freedom
## Multiple R-squared:  0.7872, Adjusted R-squared:  0.7757
## F-statistic: 67.99 on 8 and 147 DF,  p-value: < 2.2e-16



#Step Model

summary(step_mod)

##
## Call:
## lm(formula = Price ~ Bedroom + Space + Lot + Tax + Bathroom +
##     Garage + Condition + Bathroom:Condition, data =
transformed_real_est1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3454  -3.4630  -0.2461   3.0903  16.6619
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           76.1229     3.8870  19.584  < 2e-16 ***
## Bedroom               -3.5423     1.4720  -2.406 0.017351 *
## Space              -8893.6329  2150.1086  -4.136 5.91e-05 ***
## Lot                -4943.0541  1320.2340  -3.744 0.000259 ***
## Tax                -9365.6622  1440.7832  -6.500 1.17e-09 ***
## Bathroom               3.9459     1.5181   2.599 0.010296 *
## Garage                 5.3500     0.7785   6.872 1.67e-10 ***
## Condition            -13.5181     3.7574  -3.598 0.000438 ***
## Bathroom:Condition    10.4754     2.4188   4.331 2.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.098 on 147 degrees of freedom
## Multiple R-squared:  0.7872, Adjusted R-squared:  0.7757
## F-statistic: 67.99 on 8 and 147 DF,  p-value: < 2.2e-16
```

In the pursuit of refining the model, a backward elimination approach and stepwise selection was employed, systematically removing non-significant variables and a model with same variables is obtained. The R-squared is not found to be improved. However, while the adjusted R-squared is a valuable metric for evaluating model improvement, its increase isn't guaranteed with every iteration in variable selection. Instead, the primary goal is to enhance the model's interpretability and accuracy by removing non-significant variables, even if the adjusted R-squared remains consistent or slightly decreases. The main aim is to refine the model by excluding irrelevant predictors, improving its efficiency and practicality in predicting the target variable.

3.2 Model Evaluation

Table 3: Model Metrics

Model	AIC	Multiple R-squared	Adjusted R-squared	F-statistic	Residual SE	Decision
Initial Model	1064.696	0.7121	0.6965	45.46	7.093	Baseline model, higher AIC, less desirable
Transformation Model	1063.957	0.7135	0.6979	45.76	7.077	Slight improvement, but similar to initial model
Categorical Model	1019.098	0.7878	0.7748	60.24	6.11	Good improvement, lower AIC, handles categoricals
Backward Elimination Model	1017.533	0.7872	0.7757	67.99	6.09	Similar to step-wise, reduced predictors
Step-wise Model	1017.533	0.7872	0.7757	67.99	6.09	Similar to backward elimination, fewer predictors

AIC: Lower AIC values indicate a better balance between goodness of fit and model complexity. Both the Backward Elimination Model and the Step-wise Model have the lowest AIC scores, indicating that these models perform better in explaining the variance while considering the number of predictors involved.

Multiple R-squared: Represents the proportion of variance in the dependent variable explained by the independent variables. The higher the value, the better the model explains the variance in the data. The categorical and final models show the highest values.

Adjusted R-squared: Like R-squared but penalizes for additional variables. It's a better measure when comparing models with different numbers of predictors. All models perform close, but the categorical and final models have slightly higher values.

F-statistic: Determines if there's a significant relationship between the independent variables and the dependent variable. A higher value indicates a better relationship. The categorical, backward elimination, and stepwise models exhibit similar and high values.

Residual Standard Error (Residual SE): Indicates the average distance that the observed values fall from the regression line. Lower values indicate better fit. All models, except the initial, have reduced error compared to the initial model.

Hence, considering all the above metrics the final model considered is the model obtained after variable selection.

3.3 Final Linear Model:

```
summary(final_mod)

##
## Call:
## lm(formula = Price ~ Bedroom + Space + Lot + Tax + Bathroom +
##     Garage + Condition + Bathroom:Condition, data =
transformed_real_est1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3454  -3.4630  -0.2461   3.0903  16.6619
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            76.1229     3.8870  19.584  < 2e-16 ***
## Bedroom                -3.5423     1.4720  -2.406 0.017351 *
## Space               -8893.6329  2150.1086  -4.136 5.91e-05 ***
## Lot                 -4943.0541  1320.2340  -3.744 0.000259 ***
## Tax                 -9365.6622  1440.7832  -6.500 1.17e-09 ***
## Bathroom                3.9459     1.5181   2.599 0.010296 *
## Garage                  5.3500     0.7785   6.872 1.67e-10 ***
## Condition             -13.5181     3.7574  -3.598 0.000438 ***
## Bathroom:Condition     10.4754     2.4188   4.331 2.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.098 on 147 degrees of freedom
## Multiple R-squared:  0.7872, Adjusted R-squared:  0.7757
## F-statistic: 67.99 on 8 and 147 DF,  p-value: < 2.2e-16
```

The model, with an adjusted R-squared of 0.7757, elucidates that approximately 77.57% of the variability in 'Price' is accounted for by these variables. The residual standard error of about 6.098 reflects the typical deviation between actual 'Price' and the model's predicted values, underscoring the precision of the model in approximating property prices based on these features.

3.3.1 Model assumptions

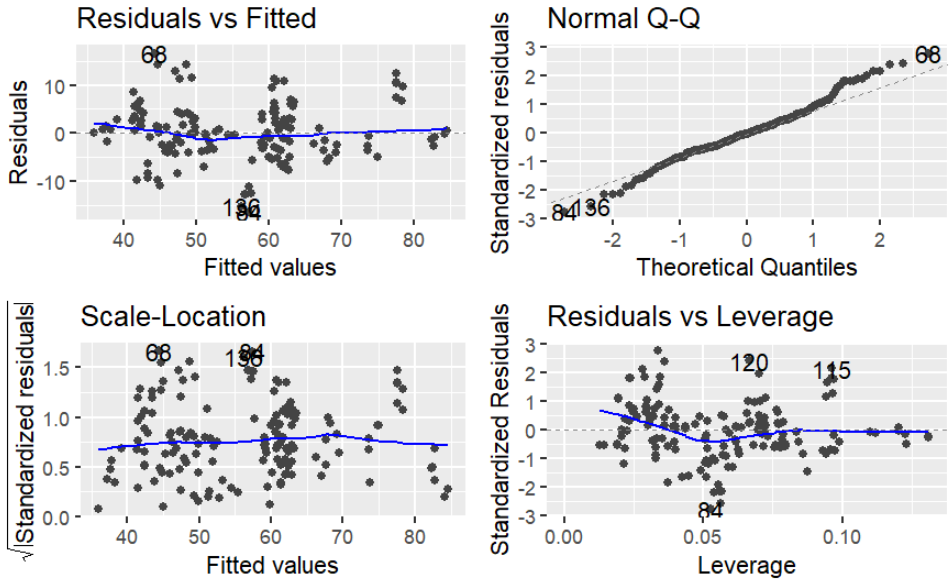

Figure 7: Model assumptions plot for the final model

```
vif_results

##             Bedroom               Space                 Lot
Tax
##            1.600685            2.251411            1.297564
1.584782
##            Bathroom              Garage           Condition
##            2.692316            1.651180           10.513169
10.82

        Bathroom: Condition

                  8043
```

The assumptions are:

• normality of errors: The normal Q-Q plot is a perfect straight line and suggest that the model covers normality.

• Constant error variance: The Residuals vs Fitted in Figure 7 plots suggests no perfect spread of the points, and hence the variance of error is same throughout (referred as homoscedasticity). Also, there is no trend observed in the Scale-Location plot.

• Independence of errors: The Residuals vs Fitted plot does not show any trend and suggests independence of errors.

• Multicollinearity: VIF values >= 5 or 10 are often considered concerning, indicating a high degree of collinearity between that variable and the other predictors. Higher VIF values imply higher multicollinearity. The VIF factor results for the final model show no such values except the condition interaction which is expected to be high.

3.4 Mathematical model and its interpretation:

The mathematical equation of the final model derived from the regression analysis is:

*Price = -3.54225527783279 * Bedroom + -8893.63293813347 * Space + -4943.05413148442 * Lot + -9365.66224351182 * Tax + 3.94593065721 * Bathroom + 5.34997713562311 * Garage + -13.5180623457835 * Condition + 10.4754073719921 * Bedroom + 76.1228968524806*

- Intercept (Constant): The base price of properties when all other predictor variables are zero is $76.12 10k US Dollars.
- Bedroom: Each additional 'Bedroom' decreases the 'Price' by $3.54 10k US Dollars, holding all other variables constant.
- Space: An increase of one unit in 'Space' reduces the 'Price' by $8893.63 10k US Dollars, assuming other factors remain constant.
- Lot: A one-unit increase in 'Lot' results in a decrease of $4943.05 10k US Dollars in 'Price.'
- Tax: 'Price' decreases by $9365.66 10k US Dollars for every one-unit increase in 'Tax.'
- Bathroom: An additional 'Bathroom' leads to an increase of $3.95 10k US Dollars in 'Price.'
- Garage: Each additional 'Garage' increases the 'Price' by $5.35 10k US Dollars.
- Condition: 'Condition' negatively impacts the 'Price,' with an increment of $13.52 10k US Dollars leading to a decrease.
- Interaction Term - Bathroom:Condition: The combined effect of 'Bathroom' and 'Condition' results in a $10.48 10k US Dollars increase in 'Price.'

4. Conclusion:

The analysis demonstrates several key findings regarding predictors influencing house prices:

Predictors with Significant Effects: Variables such as 'Space,' 'Lot,' 'Tax,' 'Bathroom,' 'Garage,' and 'Condition' exhibit substantial effects on house prices. These factors showcase statistically significant relationships which can be accessed via the p-values of the variables in the final linear model, impacting the pricing structure.

Nonlinear Effects: All the predictors exhibit linear relationships with house prices. Its impact on house prices varies, potentially influenced by additional factors beyond a linear association.

Significant Interaction Effects: A noteworthy observation involves the interaction between 'Condition' and 'Bathroom.' This interaction demonstrates a significant influence on house prices, indicating that the combined effect of these variables plays a crucial role in determining property prices which has been statistically proved.

In conclusion, the analysis underscores the significance of various predictors in determining house prices. While linear relationships dominate, nonlinear effects and interactions between specific predictors demand closer examination to comprehensively understand the pricing landscape.

5. Appendix:

5.1 R code:

Summary of dataset:

```
summary(real_est)
```

Finding the relationship between numeric predictor variables and Price:

```
par(mfrow=c(3, 3))
par(mar=c(4, 4, 2, 2))
variables <- c('Bedroom', 'Space', 'Room', 'Lot', 'Tax', 'Bathroom',
'Garage')
for (var in variables) {
 plot(real_est[[var]], real_est$Price,
 ylab = 'Price',
 xlab = var,
 col = 'darkgreen',
 pch = 19)
 lines(lowess(real_est[[var]], real_est$Price), col = 'red')
}
ggsave("plot.png", width = 6, height = 4)
```

Finding the relationship between categorical variable with other variables and target variable:

```
library(ggplot2)
library(patchwork)
options(repr.plot.width = 14, repr.plot.height = 12, repr.plot.res = 100)
options(repr.plot.width = 20, repr.plot.height = 12, repr.plot.res = 100)
numeric_vars <- c("Price", "Room", "Bedroom", "Space", "Garage", "Lot",
"Tax", "Bathroom")
density_plots <- lapply(numeric_vars, function(var) {
  ggplot(real_est, aes(x = !!as.name(var), fill = factor(Condition))) +
    geom_density(alpha = 0.5) +
    labs(title = paste( var), x = var, y = "Density") +
    scale_fill_discrete(name = "Condition")
})
plots_arranged <- wrap_plots(density_plots, nrow = 3) +
  plot_layout(guides = "collect", widths = c(6, 6, 4))
plots_arranged
```

Correlation analysis:

```
library(corrplot)
correlation_matrix <- cor(real_est)
```

```
desired_order <- c("Price", "Bedroom", "Space", "Room", "Lot", "Tax",
"Garage", "Bathroom", "Condition")
reordered <- correlation_matrix[desired_order, desired_order]
corrplot(
  reordered,
  method = "color",
  type = "upper",
  col = colorRampPalette(c("darkblue", "white"))(100),  # Divergent color
scale
  tl.col = "black",
  tl.srt = 45,
  addCoef.col = "black",
  number.cex = 0.7,
  order = "original",
  addrect = 2,
  diag = FALSE,
  mar = c(0, 0, 2, 0)
)
```

Model Assumptions:

```
par(mfrow=c(2,2), mar=c(4.5,4,2,2))
plot(real_est_mod)

require(ggfortify)

autoplot(real_est_mod)
```

Independence of Error assumption:

```


par(mfrow = c(3, 2), mar = c(4, 4, 2, 1))  # Increase bottom margin to 2

# Plotting studentized residuals against each independent variable
for (i in names(real_est)[2:7]) {
  plot(real_est[[i]], rstudent(real_est_mod),
       xlab = i, ylab = "Studentized Residuals",
       main = paste("Residuals vs", i),
       col = "coral")
}
```

BoxCox Transformation:

```
library(MASS)

bc_real_est <- boxcox(real_est_mod)

bc_bedroom <- boxcox(lm(Bedroom ~ 1, data = real_est))

bc_room <- boxcox(lm(Room ~ 1, data = real_est))

bc_space <- boxcox(lm(Space ~ 1, data = real_est))

bc_lot <- boxcox(lm(Lot ~ 1, data = real_est))
```

```r
bc_tax <- boxcox(lm(Tax ~ 1, data = real_est))

bc_bathroom <- boxcox(lm(Bathroom ~ 1, data = real_est))

bc_garage <- boxcox(lm((Garage+0.01) ~ 1, data = real_est))

optimal_lambda_bedroom <- bc_bedroom$x[which.max(bc_bedroom$y)]
optimal_lambda_room <- bc_room$x[which.max(bc_room$y)]
optimal_lambda_space <- bc_space$x[which.max(bc_space$y)]
optimal_lambda_lot <- bc_lot$x[which.max(bc_lot$y)]
optimal_lambda_tax <- bc_tax$x[which.max(bc_tax$y)]
optimal_lambda_bathroom <- bc_bathroom$x[which.max(bc_bathroom$y)]
optimal_lambda_garage <- bc_garage$x[which.max(bc_garage$y)]

# Display optimal lambda values
optimal_lambda_values <- data.frame(
  Variable = c("Bedroom", "Room", "Space", "Lot", "Tax", "Bathroom",
"Garage"),
  Optimal_Lambda = c(
    optimal_lambda_bedroom,
    optimal_lambda_room,
    optimal_lambda_space,
    optimal_lambda_lot,
    optimal_lambda_tax,
    optimal_lambda_bathroom,
    optimal_lambda_garage
  )
)

optimal_lambda_values$Optimal_Lambda <-
round(optimal_lambda_values$Optimal_Lambda,1)
optimal_lambda_values
```

```
##   Variable Optimal_Lambda
## 1  Bedroom            0.2
## 2     Room           -0.7
## 3    Space           -0.7
## 4      Lot           -1.9
## 5      Tax           -0.7
## 6 Bathroom           -0.9
## 7   Garage            0.1
```

```r
#transforming the variables according the lambda values
transformed_real_est <- real_est

transformed_real_est$Bedroom <- log(real_est$Bedroom)
transformed_real_est$Room <- real_est$Room^(-1)
transformed_real_est$Space <- real_est$Space^(-1)
transformed_real_est$Lot <- real_est$Lot^(-2)
transformed_real_est$Tax <- real_est$Tax^(-1)
transformed_real_est$Bathroom <- real_est$Bathroom^(-1)
transformed_real_est$Garage <- log((real_est$Garage+0.01))
```

Checking transformed data model Assumptions again:

```
par(mfrow=c(2, 2))
variables <- c('Price', 'Space', 'Garage', 'Lot', 'Bedroom', 'Bathroom',
'Room', 'Tax')
for (var in variables) {
  hist(transformed_real_est[[var]], main="", xlab=var, ylab="Frequency",
col="skyblue", border="black", breaks=20)
}

par(mfrow=c(2,2), mar=c(4.5,4,2,2))
plot(transformed_real_est1_mod)

require(ggfortify)
autoplot(transformed_real_est1_mod)
```

AIC metrics for all models:

```
AIC(categorical_full_mod2)

AIC(full_mod)

AIC(back_mod1)

AIC(real_est_mod)

AIC(transformed_real_est1_mod)

AIC(categorical_full_mod2)

AIC(back_mod1)

AIC(step_mod)
```

Final Model:

```
final_mod <- step_mod
```

Checking its Model Assumptions:

```
par(mfrow=c(2,2), mar=c(4.5,4,2,2))
plot(final_mod)
```

```
require(ggfortify)
autoplot(final_mod)
```

Checking multicollinearty:

```
vif_results <- car::vif(final_mod)
```

Mathematical formula:

```
final_model <- lm(Price ~ Bedroom + Space + Lot + Tax + Bathroom + Garage
+ Condition + Bathroom:Condition, data = transformed_real_est1)

coefficients <- coef(final_model)
formula <- as.formula(final_model)

# Getting variable names
variables <- all.vars(formula)[-1]  # Excluding the intercept

# Constructing the equation with variable names
equation <- paste("Price =", paste(coefficients[2:length(coefficients)],
variables, sep = " * ", collapse = " + "), collapse = " ")
intercept <- coefficients[1]  # Intercept value
equation_with_intercept <- paste(equation, intercept, sep = " + ")
equation_with_intercept
```

```
## [1] "Price = -3.54225527783279 * Bedroom + -8893.63293813347 * Space +
-4943.05413148442 * Lot + -9365.66224351182 * Tax + 3.94593065721 *
Bathroom + 5.34997713562311 * Garage + -13.5180623457835 * Condition +
10.4754073719921 * Bedroom + 76.1228968524806"
```