University of Strathclyde Glasgow

DEPARTMENT OF
MATHEMATICS AND STATISTICS

MM916 Regression Modelling Project

by

202490978

MSc Data Analytics
2024/25

Contents

List of Figures

List of Tables

1 Introduction

This project aims to build a multiple linear regression model to predict house prices in Chicago using a dataset of 157 houses. The dataset includes variables such as price, number of bedrooms, total rooms, house size (space), lot width, annual tax, number of bathrooms, number of garages, and house condition.

The primary objectives of this study are to answer the following questions:

1. Which predictors have significant effects on the house prices?
2. Does any predictor have nonlinear effects on house price?
3. Is there a significant interaction between Condition and any other predictor on the house price?

Based on these questions, the following hypotheses will be tested:

- H1: Predictors, such as house size, tax, and bathroom, have significant effects on house prices.
- H2: Some predictors exhibit non-linear relationships with house prices.
- H3: Interaction effects significantly influence house prices.

The analysis involves data cleaning, exploratory data analysis to understand relationships between variables, and statistical modelling techniques for model selection and evaluation. By interpreting the results, we seek to provide insights into how different property features influence house prices in Chicago.

2 Methods

In this section, we outline the approach used to analyse the dataset, including exploratory techniques and model building strategies.

2.1 Exploratory Data Analysis

The dataset was first explored to understand its structure and relationships between variables. Key steps included:

- Correlation Analysis:
  - A correlation matrix was generated to identify strong linear relationships between potential predictors and Price, as well as among predictors themselves. This helped to detect potential multicollinearity issues.
- Scatterplots and Boxplots:
  - Scatterplots were used to visualise relationships between Price and continuous predictors like Tax and Space.
  - Boxplots were employed to examine Price differences based on categorical variables like Condition.

2.2 Model Building

2.2.1 Leaps and Bounds

The first model is constructed using the Leaps and Bounds regression technique, which evaluates possible subsets of predictors to identify the combination that optimally explains house prices. This method ranks models based on adjusted R-squared, ensuring a balance between model complexity and goodness-of-fit. Adjusted R-squared will penalise the addition of unnecessary variables, avoiding overfitting.

As the leaps package in R does not support interaction terms, this model is limited to main effects only. Despite this restriction, Leaps and Bounds provides an efficient way to explore the predictive power of different subsets of variables. This analysis will look at several candidate models with varying numbers of predictors, ranked by their adjusted-R-squared.

This process serves as a strong starting point, identifying the most important main effect predictors driving house prices whilst maintaining simplicity.

2.2.2 Stepwise Selection

The second model is developed using the Stepwise Selection technique, which refines the predictors included in the model by adding or removing variables based on their contribution to model performance. This approach uses the Akaike Information Criterion (AIC) to evaluate model performance, balancing goodness-of-fit with model simplicity, where a lower AIC value indicates a better model.

Stepwise selection combines both of the following methods of selection:

- Forward Selection: Variables are included iteratively, starting with an empty model, and included only if they improve the model's AIC.
- Backward Selection: All predictors are initially included, and variables are removed iteratively based on their impact on AIC.

Also, Stepwise Selection allows for the inclusion of interaction terms, enabling a deeper exploration of how predictors interact to influence house prices. By utilising Stepwise Selection, this ensures that both individual predictor effects and their combined influence are accounted for in the final regression model.

2.3 Statistical Software and Assumptions

All analyses were conducted in R using packages such as ggplot2, MASS, leaps, corrplot and tidyverse. Diagnostic checks for linear regression assumptions included:

- Residuals vs. Fitted plots to check the linearity and constant variance assumption.
- Q-Q plots to check the normality assumption.
- Scale-Location plots to check the constant variance assumption.
- Residuals vs. Leverage plots to check for influential observations.

The significance level for statistical tests was set at $\alpha = 0.05$, and transformations were applied where necessary to meet model assumptions. This robust approach ensured that the final model was both statistically valid and interpretable.

3 Results

3.1 Exploratory Data Analysis

A correlation matrix was computed as the first step in our exploratory data analysis (EDA) to visualise the linear relationships between continuous variables, providing a foundation for further analysis.

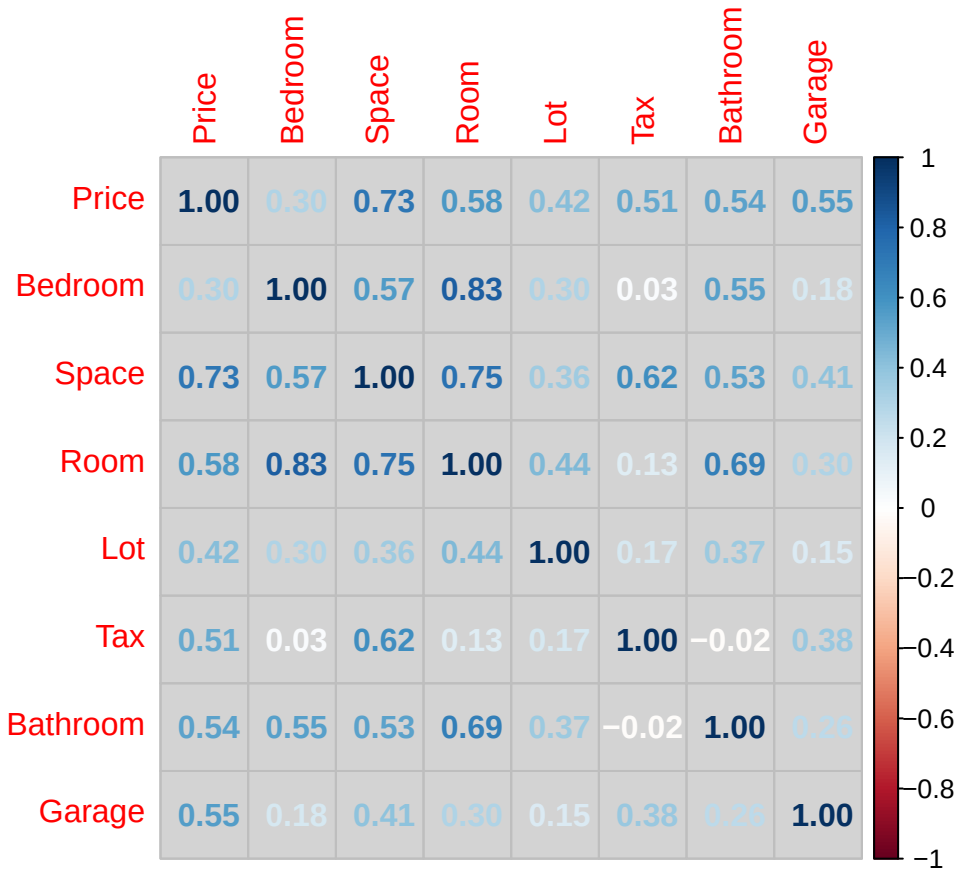

Figure 1: Correlation Matrix of Variables

From Figure 1, there is a strong positive correlation of 0.75 between the number of rooms and the space of a house. This indicates that as the number of rooms increases, the total space tends to increase proportionally. This relationship is expected, as larger houses typically have more rooms. Also, the correlation between the number of bedrooms and rooms is exceptionally high at 0.83, suggesting that the number of bedrooms is a significant component of the total rooms in a house.

The high correlation between Room and Space suggests potential redundancy. To avoid overfitting or instability in model coefficients, one of these variables may need to be excluded. Similarly, the high correlation between Bedroom and Room also requires the same careful consideration. Therefore, for the purposes of this analysis the number of rooms will not be considered as a predictor variable in our model building process.

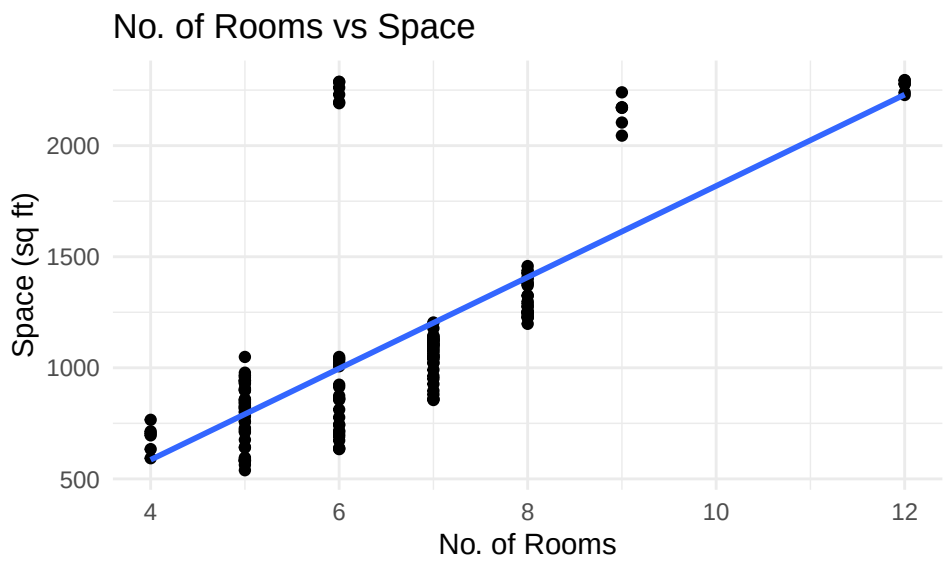

Figure 2: Scatterplot of No. of Rooms vs Space

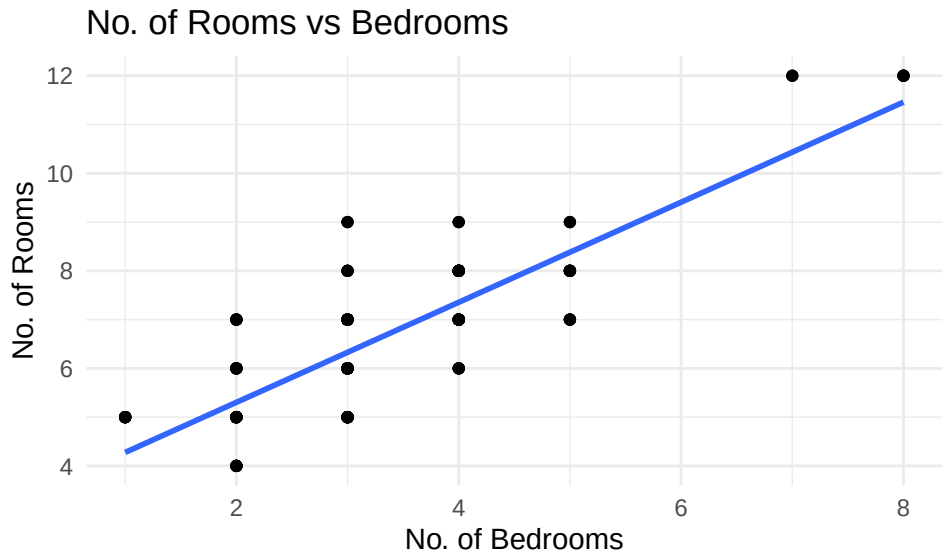

Figure 3: Scatterplot of No. of Rooms vs Bedrooms

Figures 2 and 3 provide additional evidence of the strong linear relationships among the predictors, particularly between the number of rooms and space, and the number of rooms and bedrooms. The scatterplot of Rooms vs. Space (Figure 2) demonstrates a clear linear trend, indicating that as the number of rooms increases, the total space of the house increases proportionally. Similarly, Figure 3 illustrates the relationship between the number of bedrooms and total rooms, confirming their high correlation and suggesting that bedrooms are a key component of the total number of rooms in a house. These findings highlight the importance of addressing multicollinearity in the modeling process, as these variables may introduce redundancy if both are included without adjustment

Now, to evaluate potential non-linear relationships between Price and other predictors we looked at scatterplots of potential Predictors against Price:

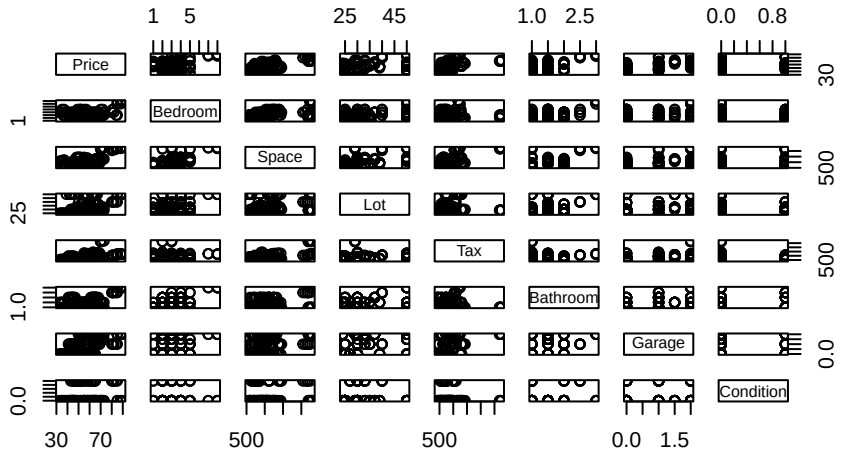

Figure 4: Scatterplots of Relationships of Predictors

From Figure 4, we observed a slight indication of non-linearity in the scatterplot between Price and Tax, particularly at higher tax values where the variability in Price increased. This prompted a closer examination of this relationship:

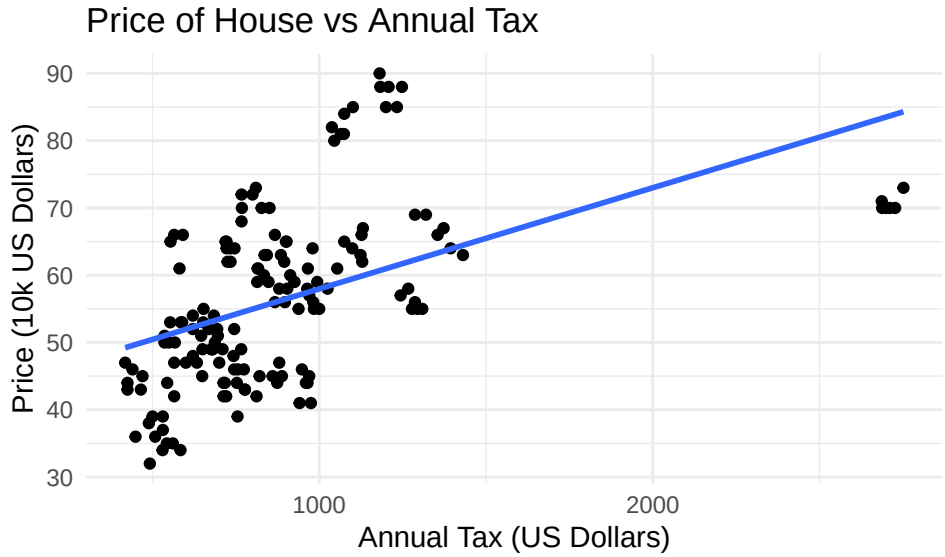

Figure 5: Scatterplot of Price of House vs Annual Tax

Figure 5 illustrates the scatterplot of Price against Annual Tax, highlighting a generally positive relationship. However, closer inspection reveals increasing variability in Price as Tax values rise. This suggests potential non-linearity in the relationship, with higher Tax values disproportionately influencing Price. Additionally, the clustering of outliers at the upper end of Tax values further emphasises the need for corrective measures to ensure the regression model accurately captures the relationship.

Diagnostics plots were then examined:

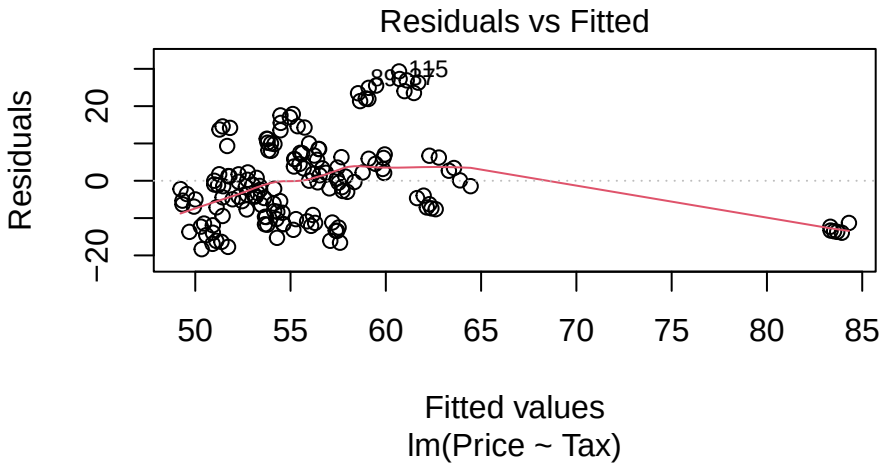

Figure 6: Diagnostic Plots of Annual Tax Model

From the diagnostic plot analysis the residuals vs. fitted plot in Figure 6 provides additional evidence supporting the observations from Figure 5. The red smoothing line in the plot demonstrates a noticeable curvature, indicating a violation of the linearity assumption.

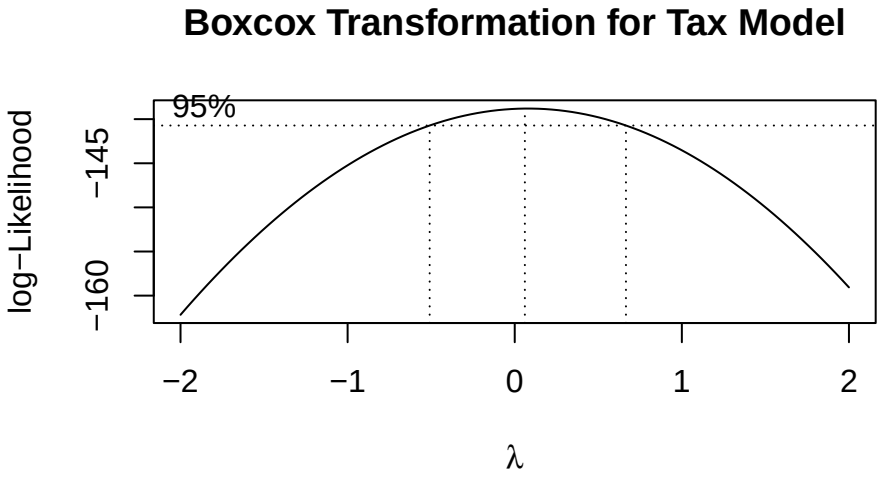

Figure 7: Boxcox Transformation for Tax Model

Figure 7 shows the results of a Box-Cox transformation analysis applied to the Tax variable. The log-likelihood curve peaks at $\lambda = 0$, indicating that a logarithmic transformation is the most appropriate adjustment for Tax.

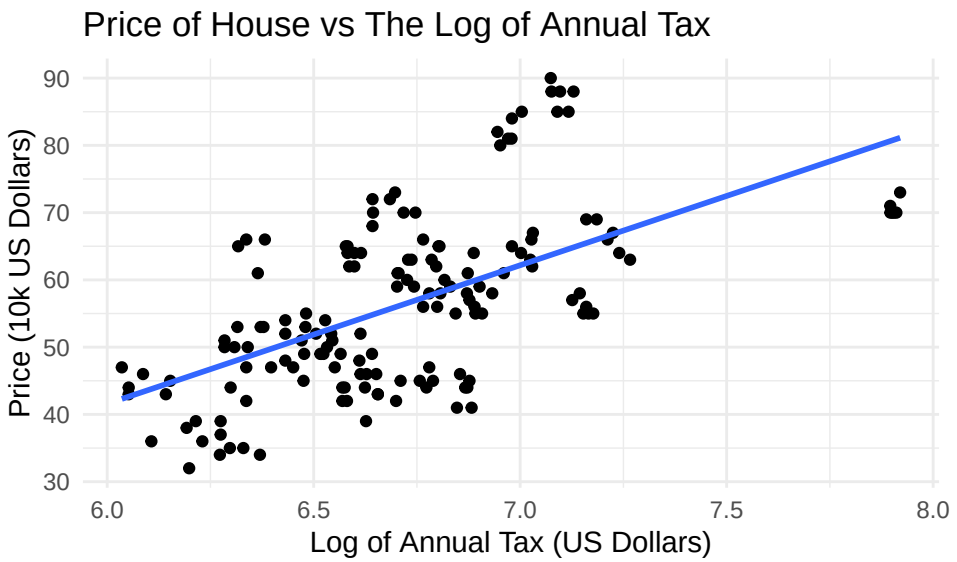

Figure 8: Scatterplot of Price of House vs Annual Tax

Figure 8 shows the scatterplot of Price against the logarithm of Annual Tax after applying the logarithmic transformation to the Tax variable. The transformation successfully addresses the issues observed in the original Price vs. Tax relationship (Figure 5). The scatterplot now demonstrates a clearer linear trend, with reduced variability at higher Tax values. The outliers that previously distorted the relationship are also better integrated into the overall trend, further validating the effectiveness of the logarithmic transformation in stabilising the relationship between Tax and Price

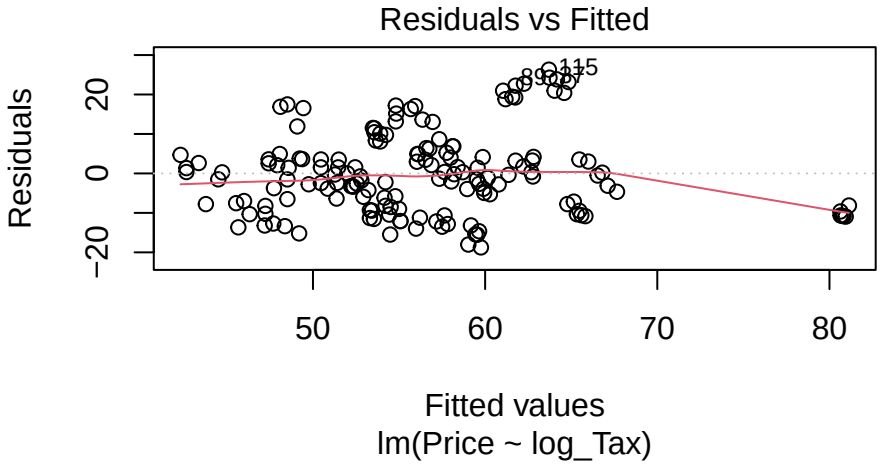

Figure 9: Diagnostic Plot of Log of Annual Tax Model

The residuals vs. fitted plot in Figure 9 illustrates the diagnostic results after including the log-transformed Tax variable in the regression model. Compared to the original residuals vs. fitted plot (Figure 6), there is less noticeable curvature in the smoothing line. This improvement indicates that the logarithmic transformation successfully mitigates the non-linearity previously observed in the non-transformed model.

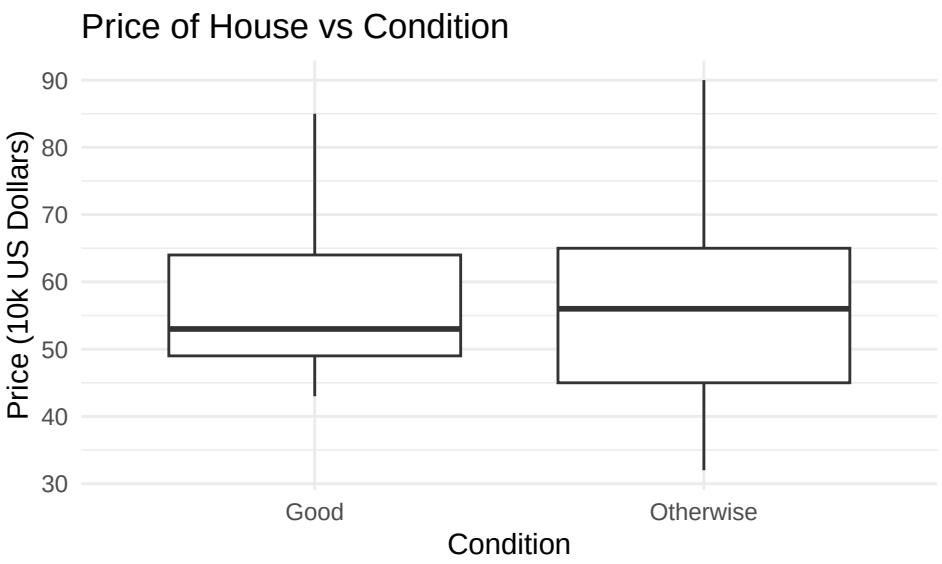

Figure 10: Boxplot of Price by Condition of House

Figure 10 illustrates the relationship between Price and the categorical variable Condition ("Good" vs. "Otherwise") using a boxplot. From this pot there does not appear to be much difference in the price of houses in terms of their condition. However, it is important to explore whether the condition interacts with other predictors to influence Price.

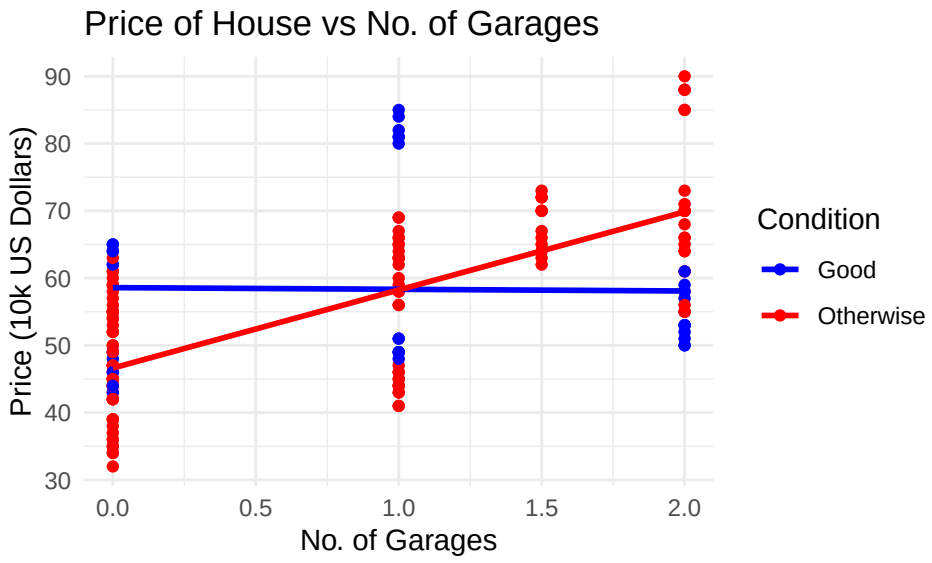

Figure 11: Scatterplot of No. of Garages vs Price by Condition of House

The effect of condition was examined for all predictors. Notably, in figure 11 the relationship between Price and the number of garages, with houses categorised by Condition ("Good" vs. "Otherwise") reveals a distinct interaction between Condition and garages. For houses classified as "Otherwise," there is a positive relationship, where additional garages are associated with higher house prices. Conversely, for houses in "Good" condition, the number of garages has little to no effect, as indicated by the nearly flat blue trend line.

This contrast in slope and intercept highlights the importance of including interaction terms in the analysis. The effect of number of garages on Price depends significantly on the Condition of the house, and failing to account for this interaction would overlook an important driver of house prices. Therefore, interaction terms should be considered in our model building process.

3.2 Leaps and Bounds

Building on our EDA, we can now utilise Leaps and Bounds regression to identify the best subsets of predictors for modelling house prices.

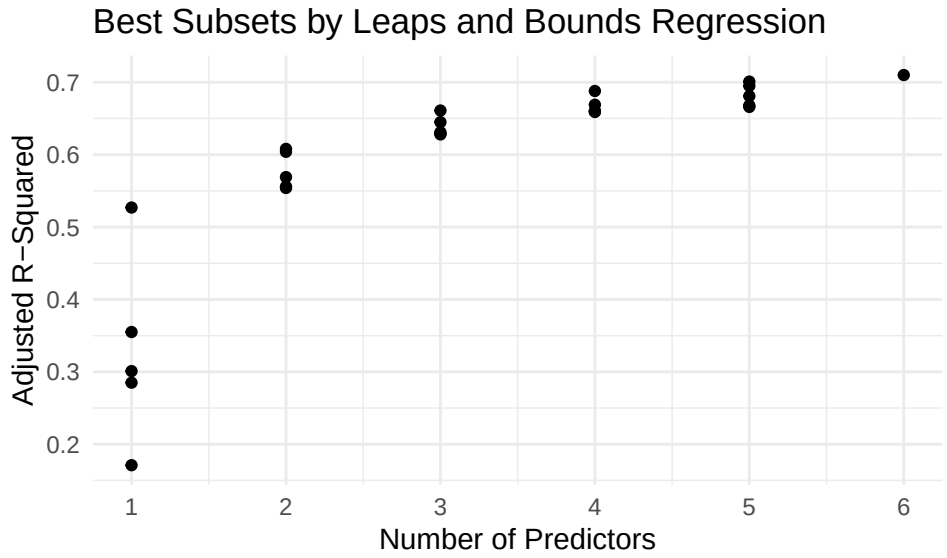

Figure 12: Scatterplot of Best Subsets by Leaps and Bounds Regression

From figure 12, we can see a trend where the adjust r-squared has a positive relationship with the number of predictors and that as we increase the number of predictors. Now lets look at these models:

Table 1: Best Subsets by Leaps and Bounds Regression

No. of predictors	Adjusted R-Squared	Bedroom	Space	Bathroom	Garage	log(Tax)	Lot
6	0.710	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
5	0.701	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
5	0.695	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
4	0.688	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE
5	0.681	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
4	0.669	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE

Table 1 displays the best subsets identified through leaps and bounds, ranked by adjusted-r-squared. It shows the combinations of predictors included in each subset, helping us evaluate the trade-offs

between model complexity and predictive accuracy. From this we can see, with an adjusted r-squared of 0.710, the best model includes all six predictors: Bedroom, Space, Bathroom, Garage, log(Tax), and Lot

From this we get the following regression equation:

$$\text{Price} = \beta_0 + \beta_1\text{Bedroom} + \beta_2\text{Space} + \beta_3\text{Bathroom} + \beta_4\text{Garage} + \beta_5\log(\text{Tax}) + \beta_6\text{Lot} + \epsilon$$

Where:

- β_0 is the intercept
- $\beta_1, \beta_2 ... \beta_6$ are the coefficients of the predictors
- ϵ is the error term

3.3 Stepwise Selection

We began the stepwise selection process with an initial null model that included only the intercept. Using both forward and backward selection, predictors and interactions were evaluated to identify the optimal combination of variables for explaining Price. The scope of this analysis included variables established in the Leaps and Bounds process but were also able to include the categorical variable, Condition, as well as interaction terms between the continuous variables and Condition.

Through this process, stepwise selection iteratively added and removed predictors based on the AIC. The resulting regression equation is as follows:

$$\begin{aligned}\text{Price} = \beta_0 &+ \beta_1\text{Space} + \beta_2\text{Garage} + \beta_3\text{Lot} + \beta_4\log(\text{Tax}) + \beta_5\text{Bathroom} + \beta_6\text{Condition} \\ &+ \beta_7(\text{Bathroom} \times \text{Condition}) + \beta_8(\text{Lot} \times \text{Condition}) + \epsilon\end{aligned}$$

Where:

- β_0 is the intercept
- $\beta_1, \beta_2 ... \beta_8$ are the coefficients of the predictors
- ϵ is the error term

This stepwise model demonstrates the importance of combining both main effects and interactions to capture the nuanced relationships between predictors and Price

3.4 Model Selection

To evaluate the performance of the two models derived through our methods of Leaps and Bounds Regression and Stepwise Selection, we compared them based on their adjusted-r-squared and AIC.

Table 2: Comparison of Leaps and Bound Model to Stepwise Model

Model	Adjusted R-Squared	AIC
Stepwise	0.742	1039.133
Leaps and Bounds	0.710	1055.436

From table 2 we can see the Stepwise model achieved a higher adjusted r-squared of 0.742 compared to 0.710 for the Leaps and Bounds model. This indicates that the Stepwise model explains a greater proportion of the variability in Price after adjusting for the number of predictors. In terms of AIC, the Stepwise model also performed better, with a lower AIC value of 1039.133 compared to 1055.436 for the Leaps and Bounds model. A lower AIC value indicates a model that strikes a better balance between goodness-of-fit and complexity.

Overall, the Stepwise model outperforms the Leaps and Bounds model on both metrics, making it the preferred model for predicting Price. Its incorporation of interaction terms, such as Bathroom x Condition and Lot x Condition, likely accounts for its superior performance by capturing the nuanced ways in which property condition affects other predictors and predicting house price.

3.5 Final Model

From our model selection we get our final model with the following regression equation:

$$\begin{aligned}\text{Price} = &-19.29 + 0.01 \cdot \text{Space} + 6.38 \cdot \text{Garage} - 0.69 \cdot \text{Lot} + 8.96 \cdot \log(\text{Tax}) + 19.26 \cdot \text{Bathroom} \\ &- 7.98 \cdot \text{Condition} - 16.84 \cdot (\text{Bathroom} \times \text{Condition}) + 0.93 \cdot (\text{Lot} \times \text{Condition})\end{aligned}$$

Table 3 presents the coefficients, standard errors, and p-values for each predictor in the final model:

Table 3: Final Model Output

Variable	Coefficient	Std. Error	P-value
(Intercept)	-19.29	17.918	2.83e-01
Space	0.01	0.002	3.69e-03
Garage	6.38	0.940	2.56e-10
Lot	-0.69	0.409	9.21e-02
log(Tax)	8.96	2.143	4.98e-05
Bathroom	19.26	2.877	4.24e-10
Condition	-7.98	10.499	4.48e-01
Bathroom x Condition	-16.84	3.548	4.87e-06
Lot x Condition	0.93	0.418	2.74e-02

- (Intercept): The intercept represents the baseline house price when all predictors are zero. In this case, the negative value indicates that, without features like space, bathrooms, or a lot, there is effectively no house to assign a price. This makes sense because a property with no attributes would not exist in reality and therefore would not have a price. The intercept is not significant at the 5% level ($p = 0.283$), further emphasising that it is not meaningful to interpret a price of a house that does not exist.
- Space: Each additional square foot of space increases the property Price by `$100`, indicating a positive relationship between space and house price. This predictor is statistically significant at the 5% level ($p = 0.00369$).
- Garage: For each additional garage adds `$63800` of value to a house, showing a substantial positive relationship between house price and number of garage. This predictor is statistically significant at the 5% level ($p < 0.00001$).

- Lot: The lot size has a small negative effect ($\beta = -0.69$), but it is not statistically significant at the 5% level ($p = 0.0921$), indicating that its effect on house price may not be consistent across properties.
- log(Tax): A one-unit increase in the logarithm of taxes increases the value of a house by `$89600`. This predictor is statistically significant at the 5% level ($p = 0.0000498$).
- Bathroom: Each additional bathroom increases house price by `$192600`, making it one of the most influential features in the model. This predictor is statistically significant at the 5% level ($p < 0.00001$).
- Condition: The main effect of condition ($\beta = -7.98$) is not statistically significant at the 5% level ($p = 0.448$), suggesting its direct contribution to house price is minimal. Its importance is primarily through interaction terms.
- Bathroom x Condition: For properties in "otherwise" condition (where Condition = 1), the positive impact of additional bathrooms decreases by `$168,400`. This interaction is statistically significant at the 5% level ($p = 0.00000487$), highlighting that the condition of a house strongly influences the value that additional bathrooms add.
- Lot x Condition: Larger lots increase house price more significantly for properties in an 'otherwise' condition ($\beta = 0.93$). This interaction is statistically significant at the 5% level ($p = 0.0274$). This suggests that the value of larger lots is more pronounced when properties are in worse condition, potentially because the lot size compensates for the lower quality of the house.

This analysis displays how house price is and the various predictors, focusing on both their direct effects and the influence of condition as a moderating factor. Key insights include the significant impact of space, garages, and bathrooms on house prices, as well as the intricate role condition plays through its interactions with bathrooms and lot size.

Also looking at the model fit:

- Residual Standard Error (6.535): The residual standard error indicates the average deviation of the predicted house prices from the actual prices.
- Adjusted R-squared (0.7424): The model explains 74.24% of the variance in house prices. This adjusted measure suggests the model is robust and does not overfit.
- F-statistic (56.83) and Overall Significance ($p < 2.2e-16$): The F-statistic indicates that the overall model is highly significant, meaning the included predictors collectively explain a substantial portion of the variability in house prices.

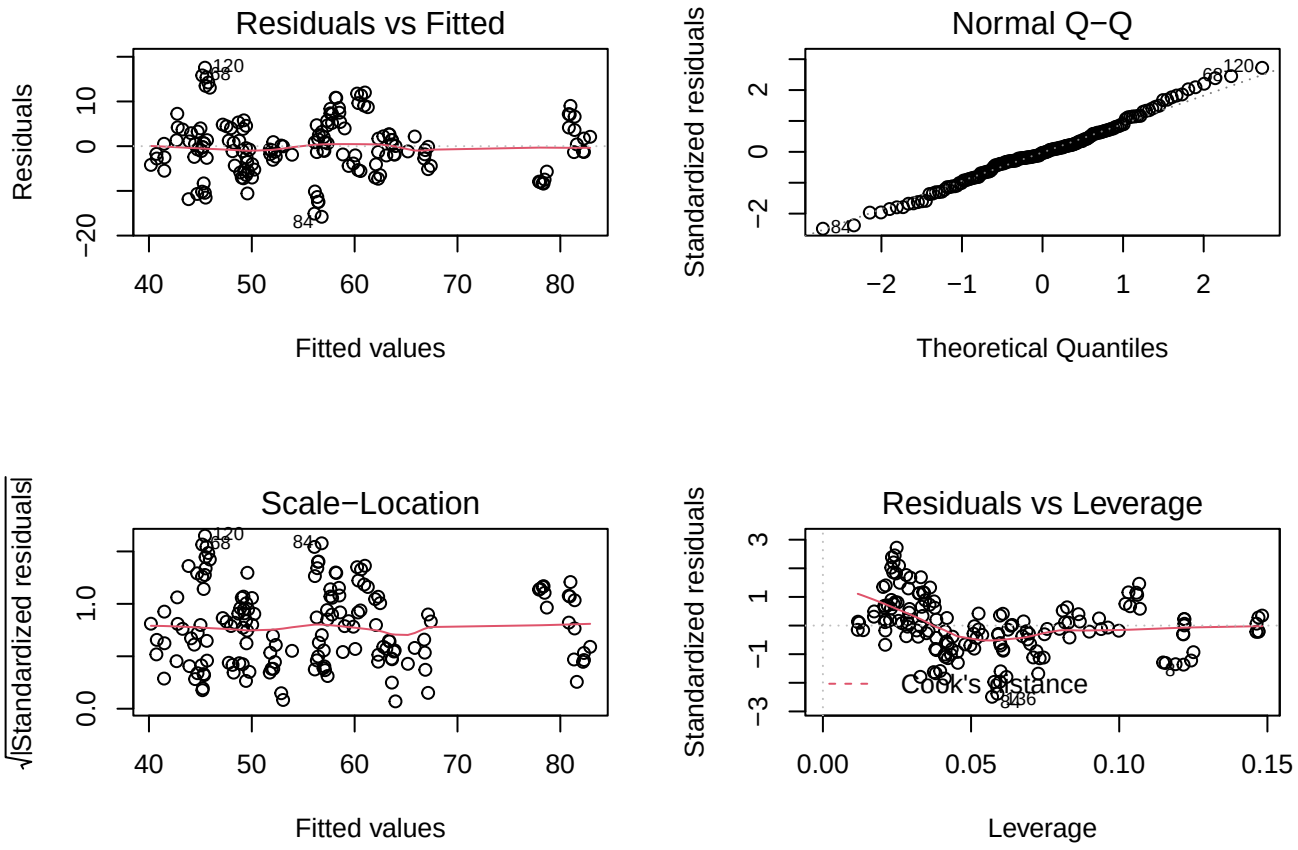

Figure 13: Diagnostic Plots of Final Model

Figure 13, displays the diagnostic plots of our final model and indicatea that the regression model meets the key assumptions of linearity, normality, constant variance, and independence of residuals. There are no significant patterns or outliers affecting the model, and it performs well in terms of prediction and stability.

4 Discussion

From our final model, we were able to explain 74.24% of the variability of house prices in Chicago, indicating, relatively strong predictive power. However, for future work we could address the following limitations of our model:

- Small Sample Size: The dataset only consisted of 157 observations, which may limit the robustness of the findings. A larger sample size would provide more statistical power and be able to establish a more nuanced view of the relationships that predict house prices.
- Broad Definition of "Otherwise": The binary classification of house condition into "Good" and "Otherwise" is overly simplistic. Grouping all non-"Good" conditions into a single category does not account for the variability and extremities of how bad the condition of a property is. This lack of granularity could reduce the model's ability to accurately capture the role of condition in predicting house price.
- Property Types: The dataset does not distinguish between property types, such as detached

houses, bungalows, or apartments. Different property types often have unique pricing dynamics, which could influence the relationships between predictors and house prices.

In future research we could hope to expand the dataset to incorporate these limitations where their incorporation would hope to improve the applicability and accuracy of the predictive model, providing a more comprehensive view into the factors affecting house prices in Chicago.

5 Conclusion

This study aimed to develop a predictive model for house prices in Chicago, whilst addressing the significance of predictors, non-linear relationships, and interaction effects. Using multiple linear regression, key factors influencing house prices were identified, and the model's performance was evaluated with a predictive accuracy of 74.24%.

From this we can answer the following hypotheses:

- H1: Predictors, such as house size, tax, and bathroom, have significant effects on house prices.
  - We accept this hypothesis as Space, log(Tax), and Bathroom were all significant predictors.
- H2: Some predictors exhibit non-linear relationships with house prices.
  - We accept this hypothesis as a non-linear relationship between Tax and house prices was identified, and a log-transformation of Tax successfully addressed this issue, reducing variability and improving model fit.
- H3: Interaction effects significantly influence house prices.
  - We accept this hypothesis as significant interactions were observed, Bathroom $\times$ Condition and Lot $\times$ Condition. These interactions highlighted the complex relationships between predictors and their influence on house price, with the condition of the property being a key moderating factor.

Overall, this analysis identifies Space, log(Tax), and Bathroom as significant predictors of house prices, with interaction effects and non-linear relationships further enhancing the model's explanatory power. However, limitations such as the small sample size, the broad classification of "Otherwise" for Condition and the lack of property type differentiation may limit the explanatory power of our findings.

Appendix

Code:

```
# Setting working directory

setwd("~/Library/CloudStorage/OneDrive-UniversityofStrathclyde/Fifth Year/MM916
      /Project 2")

# Loading necessary packages

library(ggplot2)
library(MASS)
library(tidyverse)
library(corrplot)
library(leaps)

# Loading in data

real <- read.csv("real_est.csv")
head(real)
view(real)
summary(real)

# Exploratory data analysis

par(mfrow = c(1,1))
corrplot(cor(real%>%dplyr::select(-Condition)), method = 'number',
         bg = 'lightgrey') # Correlation plot

# Addressing potential multicollinearity

ggplot(realestate, aes(x = Room, y = Space)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'Space (sq ft)',
       x = 'No. of Rooms',
       title = 'No. of Rooms vs Space') +
  theme_minimal()

ggplot(realestate, aes(x = Bedroom, y = Room)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'No. of Rooms',
       x = 'No. of Bedrooms',
       title = 'No. of Rooms vs Bedrooms') +
  theme_minimal()
```

```r
# Room is not considered as a potential predictor

# Looking at relationship of potential predictors

noroom <- real %>%
  dplyr::select(!Room)

plot(noroom)

# Converting condition to a categorical variable

realestate <- real %>%
  mutate(Condition = as.factor(ifelse(Condition == 1, "Good", "Otherwise")))
str(realestate)

# Boxplot of Price vs Condition

ggplot(realestate, aes(x = Condition, y = Price, group = Condition)) +
  geom_boxplot() +
  stat_boxplot(geom ='errorbar', width = 0.3) +
  labs(y = 'Price (10k US Dollars)',
       x = 'Condition',
       title = 'Price of House vs Condition') +
  theme_minimal()

# Comparing properties in good condition and otherwise

table(realestate$Condition)

# Now looking at each of the variables we will be looking at scatterplots and
# diagnostic plots

# Space

ggplot(realestate, aes(x = Space, y = Price, colour = Condition)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'Price (10k US Dollars)',
       x = 'Space (sq ft)',
       title = 'Price of House vs Size of House') +
  scale_colour_manual(values = c("Good" = "blue", "Otherwise" = "red")) +
  theme_minimal()

z <- lm(Price ~ Space, data = realestate)
par(mfrow = c(2,2))
plot(z)
```

```r
# Transforming space with log transformation

realestate <- realestate %>%
  mutate(log_Space = log(Space))

ggplot(realestate, aes(x = log_Space, y = Price, colour = Condition)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'Price (10k US Dollars)',
       x = 'Log of Space',
       title = 'Price of House vs The Log of Space') +
  scale_colour_manual(values = c("Good" = "blue", "Otherwise" = "red")) +
  theme_minimal()

z2 <- lm(Price ~ log_Space, data = realestate)
par(mfrow = c(2,2))
plot(z2)

# Garage

ggplot(realestate, aes(x = Garage, y = Price, colour = Condition)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'Price (10k US Dollars)',
       x = 'No. of Garages',
       title = 'Price of House vs No. of Garages') +
  scale_colour_manual(values = c("Good" = "blue", "Otherwise" = "red")) +
  theme_minimal()

z3 <- lm(Price ~ Garage, data = realestate)
par(mfrow = c(2,2))
plot(z3)

# Bathroom

ggplot(realestate, aes(x = Bathroom, y = Price, colour = Condition)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'Price (10k US Dollars)',
       x = 'No. of Bathrooms',
       title = 'Price of House vs No. of Bathrooms') +
  scale_colour_manual(values = c("Good" = "blue", "Otherwise" = "red")) +
  theme_minimal()

z4 <- lm(Price ~ Bathroom, data = realestate)
par(mfrow = c(2,2))
```

```r
plot(z4)

# Tax

ggplot(realestate, aes(x = Tax, y = Price)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'Price (10k US Dollars)',
       x = 'Annual Tax (US Dollars)',
       title = 'Price of House vs Annual Tax') +
  theme_minimal()


ggplot(realestate, aes(x = Tax, y = Price, colour = Condition)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'Price (10k US Dollars)',
       x = 'Annual Tax (US Dollars)',
       title = 'Price of House vs Annual Tax') +
  scale_colour_manual(values = c("Good" = "blue", "Otherwise" = "red")) +
  theme_minimal()

z5 <- lm(Price ~ Tax, data = realestate)
par(mfrow = c(2,2))
plot(z5)

par(mfrow = c(1,1))
boxcox(z5)

# Transforming tax variable with log transformation

realestate <- realestate %>%
  mutate(log_Tax = log(Tax))

ggplot(realestate, aes(x = log_Tax, y = Price)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'Price (10k US Dollars)',
       x = 'Annual Tax (US Dollars)',
       title = 'Price of House vs Annual Tax') +
  theme_minimal()

ggplot(realestate, aes(x = log_Tax, y = Price, colour = Condition)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'Price (10k US Dollars)',
```

```
       x = 'Log of Annual Tax (US Dollars)',
       title = 'Price of House vs The Log of Annual Tax') +
  scale_colour_manual(values = c("Good" = "blue", "Otherwise" = "red")) +
  theme_minimal()

z6 <- lm(Price ~ log_Tax, data = realestate)
par(mfrow = c(2,2))
plot(z6)

summary(z6)

# Removing tax outliers

taxless_realestate <- realestate %>%
  filter(Tax < 2000)

ggplot(taxless_realestate, aes(x = Tax, y = Price, colour = Condition)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'Price (10k US Dollars)',
       x = 'Annual Tax (US Dollars)',
       title = 'Price of House vs Annual Tax') +
  scale_colour_manual(values = c("Good" = "blue", "Otherwise" = "red")) +
  theme_minimal()

zz <- lm(Price ~ Tax, data = taxless_realestate)
par(mfrow = c(2,2))
plot(zz)

# Lot

ggplot(realestate, aes(x = Lot, y = Price, colour = Condition)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'Price (10k US Dollars)',
       x = 'Width of Lot',
       title = 'Price of House vs Width of Lot') +
  scale_colour_manual(values = c("Good" = "blue", "Otherwise" = "red")) +
  theme_minimal()

z7 <- lm(Price ~ Lot, data = realestate)
par(mfrow = c(2,2))
plot(z7)

# Bedroom
```

```r
ggplot(realestate, aes(x = Bedroom, y = Price, colour = Condition)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(y = 'Price (10k US Dollars)',
       x = 'No. of Bedrooms',
       title = 'Price of House vs No. of Bedrooms') +
  scale_colour_manual(values = c("Good" = "blue", "Otherwise" = "red")) +
  theme_minimal()

z8 <- lm(Price ~ Bedroom, data = realestate)
par(mfrow = c(2,2))
plot(z8)

# Filtering to only include relevant predictors for leaps and bounds

leapsdata <- realestate %>%
  dplyr::select('Price', 'Bedroom', 'Space', 'Bathroom', 'Garage', 'log_Tax',
                'Lot')
str(leapsdata)

# Leaps and bounds regression

best_subset <- leaps(x = leapsdata[,2:7], y = leapsdata[,1],
                     nbest=5, method="adjr2",
                     names=names(leapsdata)[-1])

# Converting to data frame

fit <- data.frame(Size = best_subset$size,
                  `Adjusted R-Squared` = round(best_subset$adjr2, 3),
                  best_subset$which, row.names = NULL)

# Plot of models by size and adjusted r squared

ggplot(fit, aes(Size, Adjusted.R.Squared)) +
  geom_point() +
  labs(title = "Best Subsets by Leaps and Bounds Regression",
       x = "Number of Predictors",
       y = "Adjusted R-Squared") +
  theme_minimal()

# Displaying the best models to be displayed as table

fit <- fit %>%
  arrange(desc(`Adjusted.R.Squared`)) %>%
  mutate('No. of predictors' = Size - 1) %>%
```

```
  rename('log(Tax)' = 'log_Tax',
         'Adjusted R-Squared' = 'Adjusted.R.Squared') %>%
  dplyr::select('No. of predictors', 'Adjusted R-Squared', 'Bedroom', 'Space',
                'Bathroom', 'Garage', 'log(Tax)', 'Lot')
head(fit)

# Creating model from leaps

leaps_model <- lm(Price ~ Bedroom + Space + Bathroom + Garage + log_Tax + Lot,
                  data = realestate)

# Summary and diagnostic plots for leaps

summary(leaps_model)
par(mfrow = c(2,2))
plot(leaps_model)

# Initial model setup for step wise selection

step_model <- lm(Price ~ 1, data = realestate)

summary(step_model)

# Step wise selection

best_step_model <- step(step_model, scope = ~ Bedroom + Space + Bathroom +
                        Garage + log_Tax + Lot + Condition + Bedroom:Condition +
                        Space:Condition + Bathroom:Condition + Garage:Condition
                        + log_Tax:Condition + Lot:Condition, direction = "both")

step_model <- lm(Price ~ Space + Garage + Lot + log_Tax + Bathroom + Condition +
                   Bathroom:Condition + Lot:Condition, data = realestate)

summary(step_model)
par(mfrow = c(2,2))
plot(step_model)

# Comparing models

leaps <- summary(leaps_model)
step <- summary(step_model)

# Creating data frame for comparison

df1 <- data_frame(Model = 'Stepwise',
                  `Adjusted R-squared` = round(step$adj.r.squared, 3),
```

```
                  `AIC` = AIC(step_model), row.names = NULL)

df2 <- data_frame(Model = 'Leaps and Bounds',
                  `Adjusted R-squared` = round(leaps$adj.r.squared, 3),
                  `AIC` = AIC(leaps_model), row.names = NULL)

comparison <- rbind(df1, df2)
comparison


# Creating table for final model output

final <- summary(step_model)

coef <- final$coefficients

final_df <- data.frame(
  Variable = rownames(coef),
  Coefficient = round(coef[, "Estimate"], 2),
  `Std. Error` = round(coef[, "Std. Error"], 2),
  `P-value` = coef[, "Pr(>|t|)"], row.names = NULL
)

final_df <- final_df %>%
  rename(`Std. Error` = Std..Error,
         `P-value` = P.value) %>%
  mutate(Variable = ifelse(Variable == "log_Tax", "log(Tax)", Variable),
         Variable = ifelse(Variable == "ConditionOtherwise", "Condition",
                           Variable),
         Variable = ifelse(Variable == "Bathroom:ConditionOtherwise",
                           "Bathroom x Condition", Variable),
         Variable = ifelse(Variable == "Lot:ConditionOtherwise",
                           "Lot x Condtion", Variable))

kable(final_df, caption = "Final Model Output")


# From Residuals vs fitted we see these observations

par(mfrow = c(1,1))
plot(step_model, which = 1)

outliers <- realestate[c(68,84,120),] %>%
  rename(`Log Tax` = log_Tax)

kable(outliers, caption = "Observations of High-leverage Points")
```