Business Statistics Using R - Group Assignment (Final Report)

Abdul Hakim Bin Kamalur Rahman (24015257)
Annabel Ching Ke Xin (24002685)
Divani A/P Arumugam (19058908)
Harresh A/L Ragunathan (19076090)
Lai Woei Harng (20003158)

2024-12-16

1.0 Introduction

Housing prices play an important role in both personal and economic contexts. This is because a house not only fulfils the basic human needs for shelter but also serves as a foundation for individual and family's stability. It also acts as a critical asset that drives economic growth and influences market trends which in turn contribute to a community's financial wellbeing (Yao & Feng, 2023). Moreover, fluctuating house prices can have significant impacts on financial markets and economy as a whole. For example, in Auckland, the sharp rise in the median house price to median household income ratio from 6.4 in 2010 to 10.0 in 2016 created financial strain for households and led to an uncontrollable rise in house prices (Greenaway-McGrevy & Phillips, 2021).

House price determination is a complex process that is influenced by many factors. Traditional economic indicators like interest rates and employment levels often provide insightful information but they do not capture the complexity of housing markets completely (Rico-Juan & Taltavull de La Paz, 2021). Several factors that determine houses prices including economic and sociodemographic factors as well as role of writing style in real estate advertisements.

A study by Cellmer et al., (2020) states that determinants such as income, employment levels, migration rates, population density and the percentage of people in the workforce impact housing demand and prices. These factors differ across regions, highlighting housing markets' demographic and spatial diversity. Markowitz (2023) highlights that linguistic complexity in real estate advertisements such as using fewer common words, less readable text and more analytical writing is linked with higher housing prices. This reveals that using complex language can make buyers see the ad as having an added value leading to higher listing prices.

The House Price Index (HPI) is a common tool used to track changes in residential property prices and analyze housing market performance over time. A study by (Li, 2021) explained that HPI is influenced by multiple factors such as economic conditions, housing demand and location of the house which vary significantly across regions. There are few limitations identified in various studies.

Aliefendioğlu et al. (2022) highlighted that the HPI can be influenced by external factors such as the Covid-19 pandemic which led to uneven pricing behaviour across different regions. It affected the HPI in Turkey by causing both short-term and long-term uneven price changes. The pandemic took a toll on the country's economic activities and affected the real estate demand supply which ultimately caused a shift in house prices. Other than that, HPI's reliability can be affected significantly by data constraints which include challenges such as financial constraints that limit access to high-quality and extensive data sources. These limitations lead to data discrepancies, causing difficulties to maintain consistency and accuracy (Sipan et al., 2018). As a result of these, the overall precision and dependability of the HPI will be compromised.

To address the limitations of traditional methods like HPI, machine learning is an essential tool for predicting housing prices. Unlike the traditional approaches, machine learning has the ability to process complex datasets and reveal hidden patterns that traditional methods often overlook. By incorporating housing attributes with historical transaction data, these ML models offer more precise and accurate predictions (Fang, 2023).

Research Questions

1. How do structural factors such as house area, number of bedrooms, bathrooms, stories and the availability of guestroom, basement and parking space influence the market price of a property?
2. How do house condition factors such as hot water heating, air conditioning and furnishing status impact housing prices?
3. What is the effect of environmental factors, including proximity to main roads and location in preferred areas on the market value of a property?
4. How do interaction terms improve the accuracy of housing price prediction models?

Research Objectives

1. To examine the influence of structural factors including house area, number of bedrooms, bathrooms, stories and the availability of a guestroom, basement and parking space on the market price of a property.
2. To assess the impact of house condition factors such as hot water heating, air conditioning and furnishing status on housing prices.
3. To analyze the effect of environmental factors such as proximity to main roads and location in preferred areas on the market value of a property.
4. To evaluate the effectiveness of interaction terms in enhancing the accuracy of housing price prediction models.

Motivation of Research

The constantly increasing housing prices in the real estate market highlight the need for accurate and reliable prediction models. House developers face multiple challenges which include soaring construction costs, financial risks as well as market volatility (Khalid et al., 2018). Studies by Mac-Barango (2017) and Ariffin et al. (2018) found that mismanagement of these factors lead to abandonment of large housing projects which in turn resulted in financial losses for buyers and loss of credibility for developers. These critical issues emphasize the significance of developing robust prediction tools to estimate housing prices accurately to make sure projects remain financially sustainable and viable. Furthermore, both buyers and sellers face difficulties in navigating the housing market. For instance, buyers lack access to current market data, which leads to overpaying for properties (Wanjiku et al., 2021; Muñoz & Cueto, 2017). From a seller's perspective, they struggle to set competitive prices that attract buyers without compromising profits. Considering these situations, there is a clear need for advanced techniques to address and overcome these challenges. This research aims to fill these gaps by developing robust and improved predictive models that guide real estate developers and agents as well as individual house buyers to make better and informed decisions.

Data Description

```r
# Load the dataset
df <- read.csv("data/Housing.csv")
```

Table 1: Table: Description of Variables

Variable Name	Description	Data Type
area	The size of the house in square feet	Numeric
bedrooms	The number of bedrooms	Numeric
bathrooms	The number of bathrooms	Numeric
stories	The number of stories	Numeric
main road	The location of the house, whether it is on the main road (yes, no)	Binary
guestroom	Whether there is a guest room (yes, no)	Binary
basement	Whether there is a basement (yes, no)	Binary
hot water heating	Whether there is hot water heating (yes, no)	Binary
air conditioning	Whether there is air conditioning (yes, no)	Binary
parking	The number of parking spaces	Numeric
preferarea	Whether it is in a preferred area (yes, no)	Binary
furnishing status	The furnishing status (furnished, semi-furnished, unfurnished)	Categorical
price	The price of the house	Numeric

Literature Reveiw

Housing Structure

Housing structure is vital in determining property values with features such as area, the number of bedrooms and bathrooms and the presence of additional features like basements and parking spaces which contribute to housing prices. For instance, Li et al. (2015) highlight that positive correlation exists between house area and its corresponding prices in compact urban places like Hong Kong which emphasizes the role of floor area in determining the house's worth. Besides, another study by Nwankwo et al. (2023) suggests that house prices in Lagos are influenced by the number of bedrooms and bathrooms which add value to the property, but it does not indicate a significant relationship between parking spaces and house prices. Additionally, Xu and Nguyen (2022) found that basement features such as the size and type do not affect house prices significantly but other attributes like house area, tax amounts and number of bathrooms were identified as the most influential factors in predicting house prices in their study. These findings suggest that machine learning model has the potential to include various housing attributes which increases the reliability and accuracy of predictions. Therefore, some hypotheses have been made as presented below.

H1: The larger the house area, the higher the housing price.

H2: An increase in the number of bedrooms positively influence housing price.

H3: The greater the number of bathrooms, the higher the housing prices.

H4: Multi-storey houses have higher prices than single-storey houses.

H5: The availability of guestrooms has positive correlation with housing prices.

H6: Houses with basement are valued higher than those without.

H7: The larger the parking space of a house, the higher the housing price.

Housing Condition

Housing condition is another aspect that is instrumental in determining house prices and it comprises of factors such as air conditioning, hot water heating and furnishing status. Air conditioning systems, particularly central air conditioning have positive influence on house prices. Homes equipped with air conditioning are often viewed to be more comfortable and convenient, making them more attractive to buyers, especially in warmer temperatures (Imran et al. (2021). Furthermore, energy-efficient air conditioning systems can

increase house prices as they align with sustainable and cost-effective housing solutions. To support this claim, a study by Shen et al. (2021) demonstrates that energy-efficient installations improve comfort and contribute to higher house prices in real estate markets. Additionally, furnishing status can be categorized into three categories, which are fully furnished, semi-furnished or unfurnished and it is an attribute that affects house prices. To validate this, studies by Abdul-Rahman et al. (2021) and Cai and Zhao (2023) found that furnishing status has minimal impact on property value as indicated by its relatively low R2 value. These findings reveal that furnishing status is not a primary driver of house prices. Thus, hypotheses have been made as below.

H8: Houses with hot water heating systems and air condition have higher prices.

H9: Fully furnished houses are priced higher than semi-furnished or unfurnished houses.

Housing Environment

Proximity to main roads and highways influences housing prices in both positive and negative manners. Better transportation access makes commuting easier and reduces travel times and hence, increasing property values. To illustrate this, Levkovich et al. (2016) found that new highways tend to hike up housing prices surrounding the area due to improved connectivity. However, at the same time, homes located very close to the highways can also lose its value because of noise and traffic. Similarly, Theisen and Emblem (2021) portray that a new highway in Norway has increased house prices by 5% on average with the biggest gains noticed in towns closer to the highway. Research by Rahadi et al. (2015) observed that direct toll road access has an impact on housing prices. This is because it provides ease of accessibility to destinations and significantly impacts consumers' housing decisions by reducing commuting time. A study by Heyman and Sommervoll (2019) reveals that the location of a house has a significant impact on its price. The researchers explain that the closeness of the house to amenities such as metro stations, shops, parks and schools. Houses near useful places with facilities and amenities have higher values indicating that certain features are more essential than just the overall location. In short, both proximity to roads and location are key factors in determining housing prices. Hence, hypotheses have been presented as below.

H10: Proximity to main roads shows a positive relationship with housing prices.

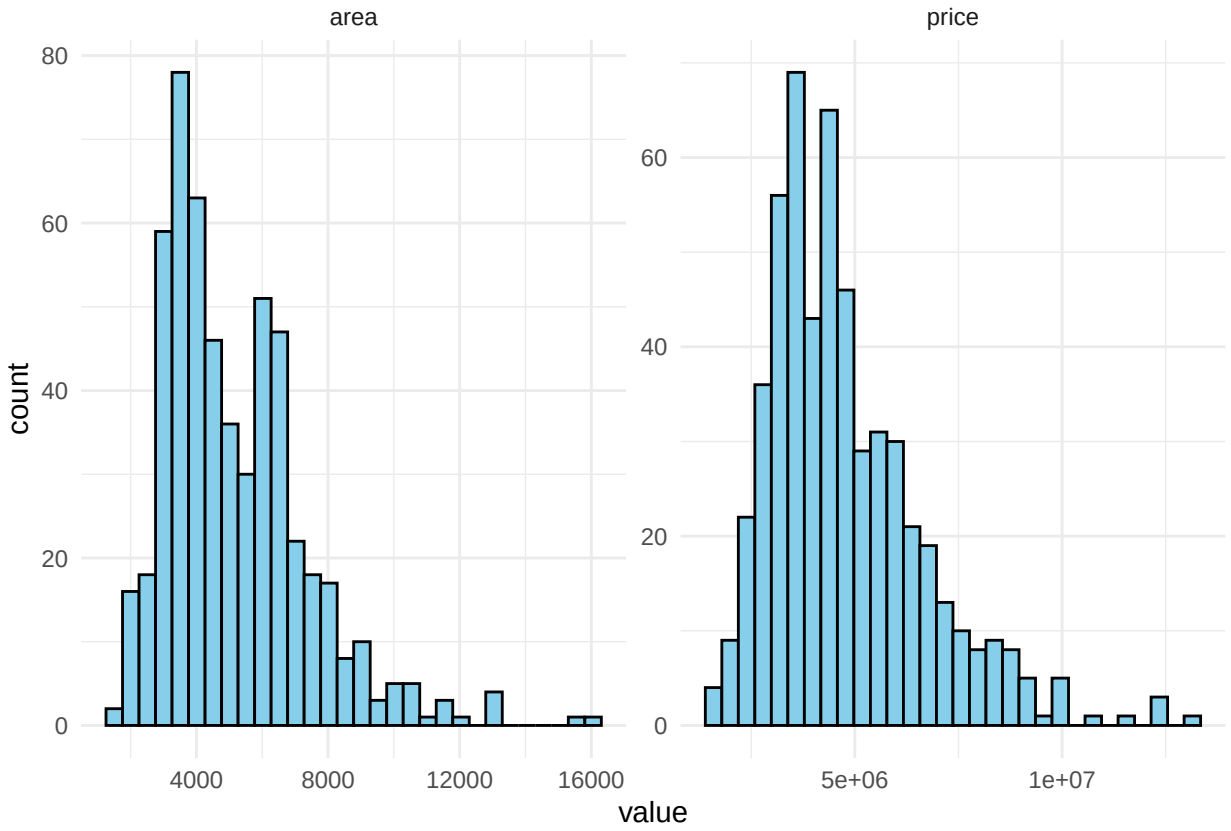
H11: Houses located in preferred areas are priced higher than those in non-preferred areas.

Exploratory Data Analysis

```
##
## ==========================================================
## Statistic  N      Mean        St. Dev.       Min       Max
## ----------------------------------------------------------
## price     545 4,766,729.000 1,870,440.000 1,750,000 13,300,000
## area      545   5,150.541     2,170.141     1,650     16,200
## bedrooms  545     2.965         0.738         1         6
## bathrooms 545     1.286         0.502         1         4
## stories   545     1.806         0.867         1         4
## parking   545     0.694         0.862         0         3
## ----------------------------------------------------------
```

The dataset provides insights into key housing characteristics. The average house price is 4,766,729 with significant variability, ranging from 1,750,000 to 13,200,000, reflecting differences in location, size, and amenities. Homes have an average area of 5,150 square feet, with sizes ranging from 1,650 to 16,200 square feet, indicating a mix of small and large properties. The typical house has about 3 bedrooms, with a range of 1 to 6, catering to different family sizes. On average, houses have 1.3 bathrooms, though some larger homes feature up to 4 bathrooms. Most homes are 1 or 2 stories high, with a few reaching up to 4 stories. Parking

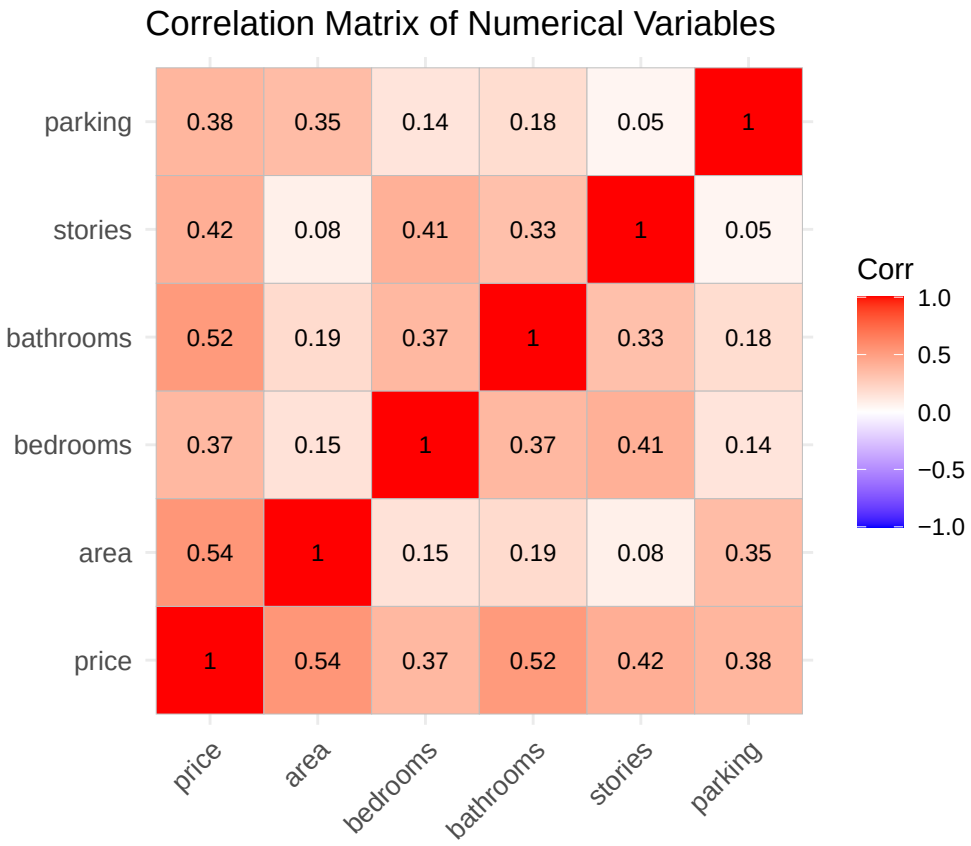
is limited, with an average of 0.7 spaces per house, but some homes offer up to 3 spots. Overall, the data reflects diverse property types, catering to various buyer needs and preferences.



The graph above shows histograms for the variables area and price which illustrate the frequency distribution. For area, the histogram reveals that most of the houses fall within the range of 3,000 to 8,000 square feet. Then, there is a sharp decline after above 10,000 square feet. The distribution of the histogram is rightly skewed, which suggests that while most of the houses are of moderate size, there are some exceptions that offered larger properties.

As for the house price, the distribution of the histogram is also rightly skewed. Most of the houses are priced between 3,000,000 to 7,000,000. However, there are some houses that are valued exceed 10,000,000, reflecting luxury housing or properties in prime locations.

Correlation Matrix



The correlation matrix reveals key relationships between variables in the dataset. Price is most strongly correlated with area (0.54), indicating that larger properties tend to be more expensive, which is consistent with market trends. Price also has a strong correlation with the number of bathrooms (0.52) and a moderate correlation with the number of stories (0.42), suggesting that homes with more bathrooms and multiple stories are generally priced higher. However, the number of bedrooms (0.37) and parking spaces (0.38) show weaker correlations with price, implying that these factors have a less significant impact on property prices compared to area, bathrooms, and stories.

2.0 Regression Analysis

For this analysis, this paper used a multiple linear regression model to predict house prices based on various predictors such as area, bedrooms, bathrooms, and additional features like parking and basement availability. Linear regression was chosen because it allows for straightforward interpretation of how each predictor affects the target variable which is the house price, and it fits the assumption of linear relationships between variables. We followed a model selection procedure that included exploratory data analysis (EDA) to identify significant predictors. We also considered adding interaction terms, such as area with stories and bedrooms with air conditioning, to explore potential combined effects of predictors. However, after evaluating the significance of these terms, we found that they did not significantly improve the model's predictive power, indicating that a simpler model might be more effective. Overall, the chosen model provides a good balance of interpretability and predictive accuracy. Additional work has been included in the additional section.

Final Model Used

Multiple Linear Regression using All Variables

```
##
## Call:
## lm(formula = price ~ area + bedrooms + bathrooms + stories +
##     guestroom + basement + parking + hotwaterheating + airconditioning +
##     mainroad + prefarea + furnishingstatus, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2619718  -657322   -68409   507176  5166695
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      42771.69  264313.31   0.162 0.871508
## area                               244.14      24.29  10.052  < 2e-16 ***
## bedrooms                         114787.56   72598.66   1.581 0.114445
## bathrooms                        987668.11  103361.98   9.555  < 2e-16 ***
## stories                          450848.00   64168.93   7.026 6.55e-12 ***
## guestroomyes                     300525.86  131710.22   2.282 0.022901 *
## basementyes                      350106.90  110284.06   3.175 0.001587 **
## parking                          277107.10   58525.89   4.735 2.82e-06 ***
## hotwaterheatingyes               855447.15  223152.69   3.833 0.000141 ***
## airconditioningyes               864958.31  108354.51   7.983 8.91e-15 ***
## mainroadyes                      421272.59  142224.13   2.962 0.003193 **
## prefareayes                      651543.80  115682.34   5.632 2.89e-08 ***
## furnishingstatussemi-furnished  -46344.62  116574.09  -0.398 0.691118
## furnishingstatusunfurnished    -411234.39  126210.56  -3.258 0.001192 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1068000 on 531 degrees of freedom
## Multiple R-squared:  0.6818, Adjusted R-squared:  0.674
## F-statistic: 87.52 on 13 and 531 DF,  p-value: < 2.2e-16
```

The linear regression model was constructed to analyze the relationship between various predictors and house price. The dependent variable was house price, while the independent variables included area, number of bedrooms, bathrooms, stories, presence of a guestroom, basement, parking space, hot water heating, air conditioning, proximity to the main road, location in a preferred area, and furnishing status.

Model Summary

The model explained approximately 68.18% of the variation in house prices since R-squared is 0.6818 and Adjusted R-squared is 0.674, indicating a strong relationship between these predictors and house price. The F-statistic was 87.52 and p value less than 0.05 showing that the model as a whole is statistically significant. The residual standard error was 1,068,000, indicating the variability of the observed house prices around the regression line.

Coefficients and Significance

The intercept of the model was not statistically significant (p = 0.871), suggesting that the baseline price without any predictors does not meaningfully explain the variation in house prices. Among the predictors, most variables were significant at various levels.

The regression analysis provided significant insights into the factors influencing housing prices, confirming many of the proposed hypotheses. Property area emerged as a key determinant, with larger properties showing a strong positive and significant impact on housing prices, validating the hypothesis that size drives value (H1). Similarly, the number of bathrooms and the presence of multiple stories were positively related to house prices, further supporting the hypotheses (H3 and H4) that these features contribute to increased property values.

The presence of additional features such as a guestroom, a basement, and parking space also significantly increased house prices. This validates the hypothesis (H5, H6 and H7) that these amenities make properties more valuable. Furthermore, modern conveniences like the availability of hot water heating system and air conditioning (H8) also played a significant role in driving prices upward, indicating the high demand for comfortable living conditions. Location factors also played an important role. Houses located near main roads or in preferred neighborhoods were shown to have significantly higher prices, confirming the hypotheses (H10 and H11) that proximity to amenities and desirable areas increases property value.

For furnishing status, semi-furnished houses did not show a significant price difference compared to fully furnished houses. However, unfurnished houses were found to have significantly lower prices, partially supporting the hypothesis (H9) that furnishing affects house prices.

Interestingly, the number of bedrooms did not have a significant impact on house prices, which contradicts the initial hypothesis (H2). This result aligns with prior study in Lyu (2024) that also found no significant relationship between bedrooms and house prices. A possible explanation for this is that the overall area of the property plays a more dominant role in determining its value. Even if a house's area increases, its price is more influenced by the total size rather than the number of bedrooms, especially if the number of bedrooms remains unchanged.

Multicollinearity Issue

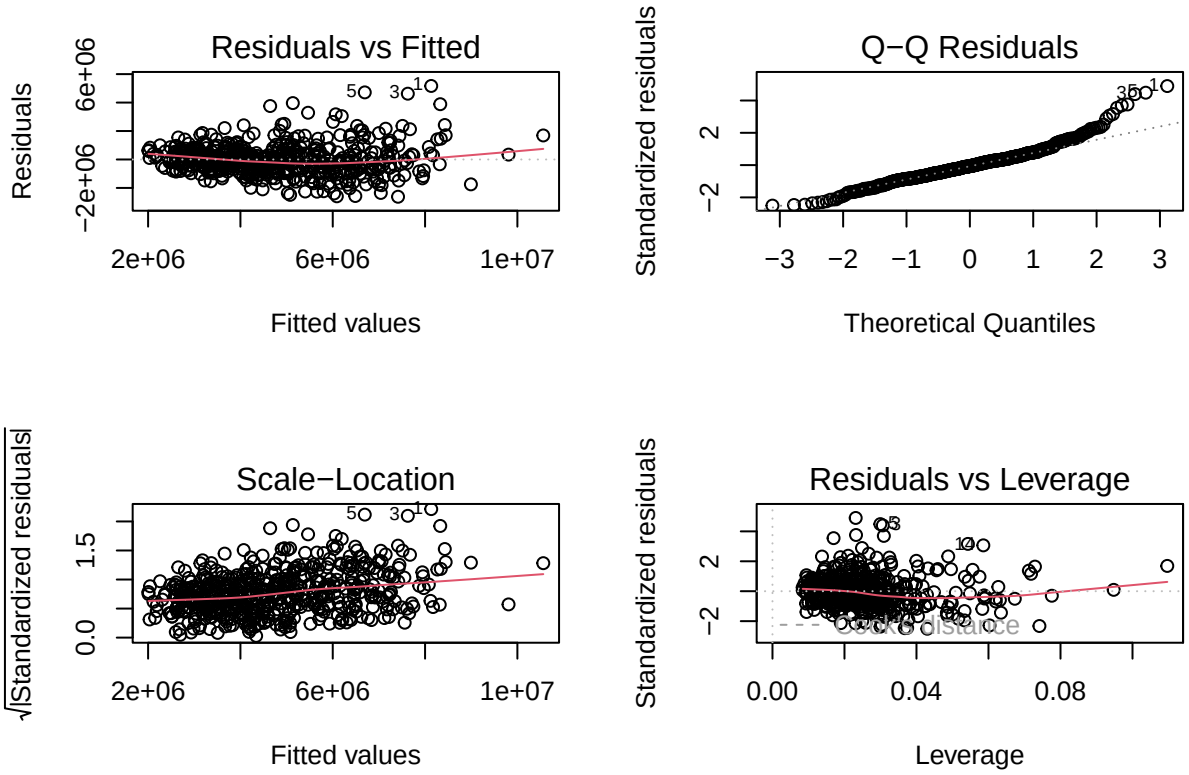
```
##                     GVIF Df GVIF^(1/(2*Df))
## area            1.325250  1        1.151195
## bedrooms        1.369477  1        1.170246
## bathrooms       1.286621  1        1.134293
## stories         1.478055  1        1.215753
## guestroom       1.212838  1        1.101289
## basement        1.323050  1        1.150239
## parking         1.212837  1        1.101289
## hotwaterheating 1.041506  1        1.020542
## airconditioning 1.211840  1        1.100836
## mainroad        1.172728  1        1.082926
## prefarea        1.149196  1        1.072006
## furnishingstatus 1.109639  2        1.026350
```

The Variance Inflation Factor (VIF) analysis was performed to assess multicollinearity among the predictors in the linear regression model. Multicollinearity occurs when two or more independent variables are highly correlated, potentially distorting the estimated coefficients and affecting the model's interpretability. A VIF value exceeding 10 is typically considered indicative of significant multicollinearity, while values under 5 are generally acceptable, and values close to 1 indicate minimal correlation.

In this analysis, all predictors show VIF values well below 2, suggest that multicollinearity is not a concern for this model, as none of the predictors exhibit a VIF high enough to indicate problematic correlations.

This implies that the predictors can be considered independent of one another, supporting the reliability of the estimated coefficients and the robustness of the model's outputs.

MLR Assumptions



The Residuals vs. Fitted plot assesses the linearity assumption by showing the residuals (errors) against the predicted (fitted) values. Ideally, the points should be randomly scattered around the horizontal line (zero residual line), indicating that the model captures the relationship between predictors and the response variable well. In this case, there are some patterns, especially at higher fitted values, suggesting potential non-linearity or heteroscedasticity.

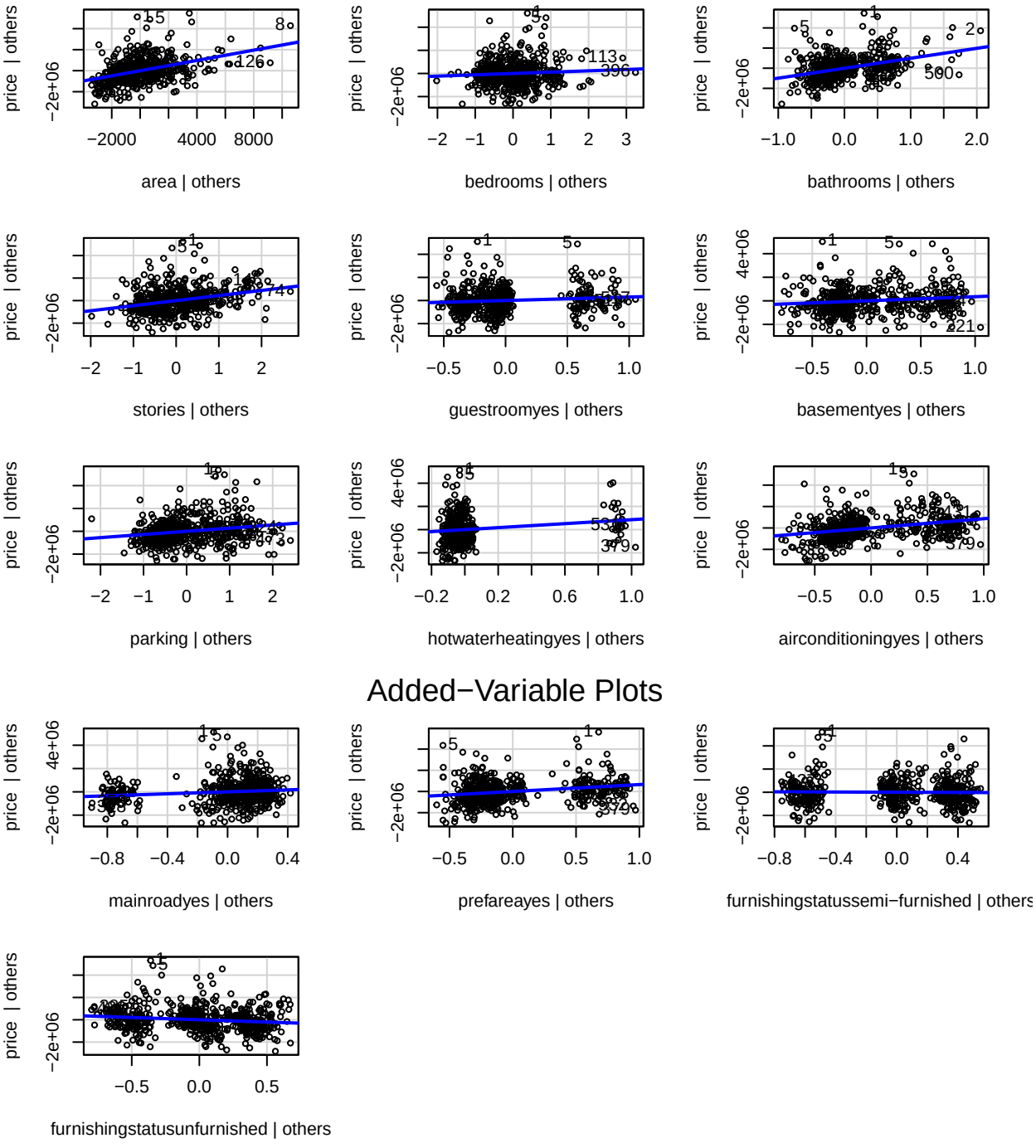
Normal Q-Q plot checks if the residuals follow a normal distribution. The points should ideally fall along the diagonal line. Here, the points deviate from the line, particularly at the tails, indicating that the residuals are not perfectly normally distributed. This deviation suggests potential outliers or a non-normal distribution of errors.

Scale-Location plot, also known as the spread-location plot, shows the square root of the standardized residuals versus the fitted values. It helps assess homoscedasticity. A horizontal line with evenly spread points indicates homoscedasticity. In this plot, a slight upward trend can be observed, which may indicate heteroscedasticity, suggesting that the variance of the residuals increases with fitted values.

Residuals vs. Leverage plot helps identify influential observations that have a significant impact on the model. Points outside the "Cook's distance" lines could be influential data points. Here, there are a few points that are farther from the main cluster and may indicate potentially influential observations that could disproportionately affect the model's results.

Overall, the plots suggest that while the model meets some linear regression assumptions, there are indications of non-normality of residuals, possible heteroscedasticity, and a few influential data points.

Added Variable Plots



The added-variable plots provide an insightful visual assessment of the unique contribution of each predictor variable to the response variable (price) while accounting for the effects of all other predictors in the linear regression model. Each plot displays the relationship between price and an individual predictor by plotting the residuals of the predictor after being regressed on other variables against the residuals of price after removing the effect of other predictors. A positive trend in the blue line suggests that the predictor has a significant and positive contribution to the price, while a negative trend indicates a negative contribution. In these plots, predictors such as area, bathrooms, and stories exhibit clear positive trends, suggesting that these variables play an essential role in influencing the price. On the other hand, variables like mainroad and furnishingstatus show flatter trend lines, indicating a lesser impact on the response variable. The scatter pattern around the trend line should ideally appear random, systematic or curved patterns may suggest non-linearity or issues with the assumptions of the model. Notable outliers marked with numbers highlight influential data points that may need further investigation, as they could affect the interpretation and stability of the model. Overall, these plots help to verify the significance and linearity of the predictors in contributing to the model's performance.

3.0 Discussion

Based on the testing, the best regression-based model, using various predictors to predict housing prices is Multiple Linear Regression. Exploring further, it may be seen that the majority of the variables are significant, the most significant being 'area', bathrooms', 'airconditioningyes', 'stories', 'prefareayes', 'parking', and 'hotwaterheatingyes'. These variables affect the model the most due to their p-values being less than 0.001. Furthermore, the variables 'furnishingstatusunfurnished', 'mainroadyes', and 'basementyes' are also significant as they have p-values less than 0.01. Additionally, the 'guestroomyes' variable is also significant as it has a p-value less than 0.05.

Based on the tested model, the area of a house is the most important factor in determining a house's predicted price. This implies that a larger house would result in an increase in its price, which may be justified by its spaciousness, flexibility in customization, and overall, being more accommodating. From a buyer's perspective, they are able to make easier decisions by being able to compare the estimated prices of properties of various sizes. From the standpoint of sellers, they are able to gauge an estimated price of their property based on its size and use it as a benchmark for how they should price their properties. Looking further into variables related to the structure of a house, the number of bathrooms and stories of a house also greatly affects its price. The number of bathrooms and stories of a house may be determined by the number of occupants in that house. It may be implied that houses with a higher number of bathrooms and stories are bought by buyers with a higher standard of living. For example, bathrooms may be perceived as a luxury due to the convenience and privacy they provide. Using this information, future property developers may want to prioritize constructing more bathrooms for future properties as a higher number of bathrooms would lead to an increase in the property's price. In the case of properties with more stories, a higher number of stories is desirable to larger families as it provides more living space in the property. Multi-story properties are especially desirable in urban areas, as it is harder to find wide spaces of land. Property developers should capitalize on this by prioritizing the construction of multi-story houses as opposed to single-story, especially in locations with constraints in area.

Aside from variables regarding the house's structure, the possession of various amenities, such as air conditioning, hot water heating, parking, a basement, and a guest room is also an important factor in determining a house's predicted price. Buyers may value these amenities when it comes to purchasing a house due to their various functional capabilities. For example, buyers in particularly hot climates may value houses that have air conditioning. Additionally, in locations with colder climates, buyers may value houses that have hot water heating. In urban locations, where driving is essential for commuting, a house that has parking would be valuable to buyers. Lastly, a house with a basement or guest room offers buyers more space and flexibility, as they are able to utilize these spaces based on what they desire (eg. storage room, office, or recreational room). Depending on location, property developers should prioritize implementing various amenities to meet buyers' needs and desires, in turn, increasing the price of houses.

Finally, the house's location, whether it is in a preferred area and located near a main road, is also important in determining a house's predicted price. A preferred area of a property may differ based on region. Generally speaking, in the eyes of a buyer, a property located in a preferred area may have a more desirable environment in terms of aesthetic and safety. Furthermore, a property in a preferred area may also be located near a centralized area (city or town), where it is in close proximity with offices, shopping malls, and schools. These locations are desirable to buyers as it makes it convenient for them to have access to various nearby amenities and landmarks. Property developers should plan to construct future properties in preferred areas, or in areas that have potential to be in close proximity to centralized areas, which in turn would increase the value of their properties. Along with the property being located in a preferred area, accessibility to a main road is also useful to homeowners, increasing the convenience of being able to commute. However, one drawback is that main roads tend to produce noise from cars passing by and traffic, which may be disruptive towards homeowners. In order to combat this drawback, property developers should implement sound proofing to mitigate this noise pollution.

The previously discussed variables all affect the model positively, meaning an increase in these variables would lead to an increase in predicted price. However, if a house is found unfurnished, it would lead to a decrease in predicted price. Overall, an unfurnished house is significantly less attractive toward buyers, both aesthetically and functionally. Due to the unappealing nature of unfurnished houses, buyers are less likely to purchase the house, in turn, decreasing the property value. In order to prevent this, property developers must focus on furnishing future properties to increase their property value, as it will be more appealing to buyers.

4.0 Limitations

Evidently, the dataset utilized regarding housing prices was proved reliable as a Multiple Linear Regression model was successfully created, where a strong relationship between the predictors and house price was demonstrated. Furthermore, the significant variables were able to be identified. However, it must be acknowledged that this study can be improved in the future as the dataset used has some limitations.

First of all, the sample size of the dataset is relatively small, 545 observations. Using a small dataset results in a less reliable and valid model compared to a larger dataset. In general, it is difficult to make generalizations from a small dataset, as it does not provide an accurate depiction of the variability that occurs in the real world. Since the sample size is small, the model will only learn from that specific sample, which would limit its application capabilities, as it will only apply to that specific sample. The model may be applied to other samples, but it will not be accurate. For example, if a dataset may only include a few locations, it may be difficult to make conclusions that apply to other locations that are not included in that sample. Furthermore, a small sample size could lead to overfitting, as the model may take note of the dataset's noise and variations. As a result, the model may perform well when using the training data, but poorly on new data. These issues can be mitigated in the future by obtaining a larger sample size, resulting in a more reliable and valid model.

Additionally, the dataset utilized could benefit from more features. By using a limited amount of features to create a model, it provides a less accurate prediction compared to a model with more features. The main features of the dataset only describe the house's structure, amenities, and location. These features are acceptable to create a model, however, it lacks variety in variables that capture other characteristics. For example, some variables that may be useful for the model are the age of the building, the cost of construction, and the neighborhood's safety.

5.0 Conclusion

This study successfully identified key predictors influencing housing prices and demonstrated that Multiple Linear Regression (MLR) is an effective model for predicting these prices based on various structural,

locational, and amenity-related factors. Significant predictors included house size (area), the number of bathrooms, the presence of amenities like air conditioning and parking, and locational factors such as proximity to main roads and preferred areas. Among these, the house's size emerged as the most influential factor, reinforcing the importance of space in determining property value.

The insights derived from this analysis have practical implications for both buyers and sellers in the housing market. Buyers can make informed decisions by comparing property prices based on key features, while sellers and property developers can strategically prioritize features like multi-story layouts, additional bathrooms, and essential amenities to enhance property value. Locational factors, particularly properties in centralized or preferred areas, further highlight opportunities for real estate investments.

Despite the success of the MLR model, the study is not without limitations. The small sample size of 545 observations restricts the model's generalizability, and the limited features of the dataset reduce the scope of its predictions. Future research could address these limitations by using larger datasets with more diverse features, such as construction costs, property age, and neighborhood characteristics. Additionally, exploring advanced machine learning models could provide deeper insights and improve predictive accuracy.

In conclusion, this research contributes to understanding housing price determinants and provides actionable recommendations for stakeholders in the real estate market. Expanding the dataset and incorporating additional predictors could further enhance the reliability and applicability of the findings, paving the way for more robust housing price prediction models in future studies.

6.0 Additional Work

Additional EDA

```
##
## ===============================================================
## Statistic  N      Mean        St. Dev.       Min       Max
## ---------------------------------------------------------------
## price     545 4,766,729.000 1,870,440.000 1,750,000 13,300,000
## area      545   5,150.541     2,170.141     1,650     16,200
## bedrooms  545     2.965         0.738         1         6
## bathrooms 545     1.286         0.502         1         4
## stories   545     1.806         0.867         1         4
## parking   545     0.694         0.862         0         3
## ---------------------------------------------------------------
```

The table above shows the summary of the overall key numerical in the datasets. These numbers provide useful information for the distribution and spread of the data which helps to understand the general characteristics of each dataset. The breakdown of each statistic is explained below.

The average house price in the dataset is approximately 4,766,729 and a standard deviation of 1,870,440. By looking at the numbers, there is high variability in house prices, ranging from a minimum value of 1,750,000 to a maximum value of 13,200,000. The high variability in house prices may be reflected in differences in location, size, and other facilities.

For the area, the average housing area is 5,150 square feet with a standard deviation of 2,170 square feet. The house with the smallest area is 1,650 square feet while the house with biggest area is 16,200 square feet. The wide range of sizes indicate that the dataset includes both small and large houses.

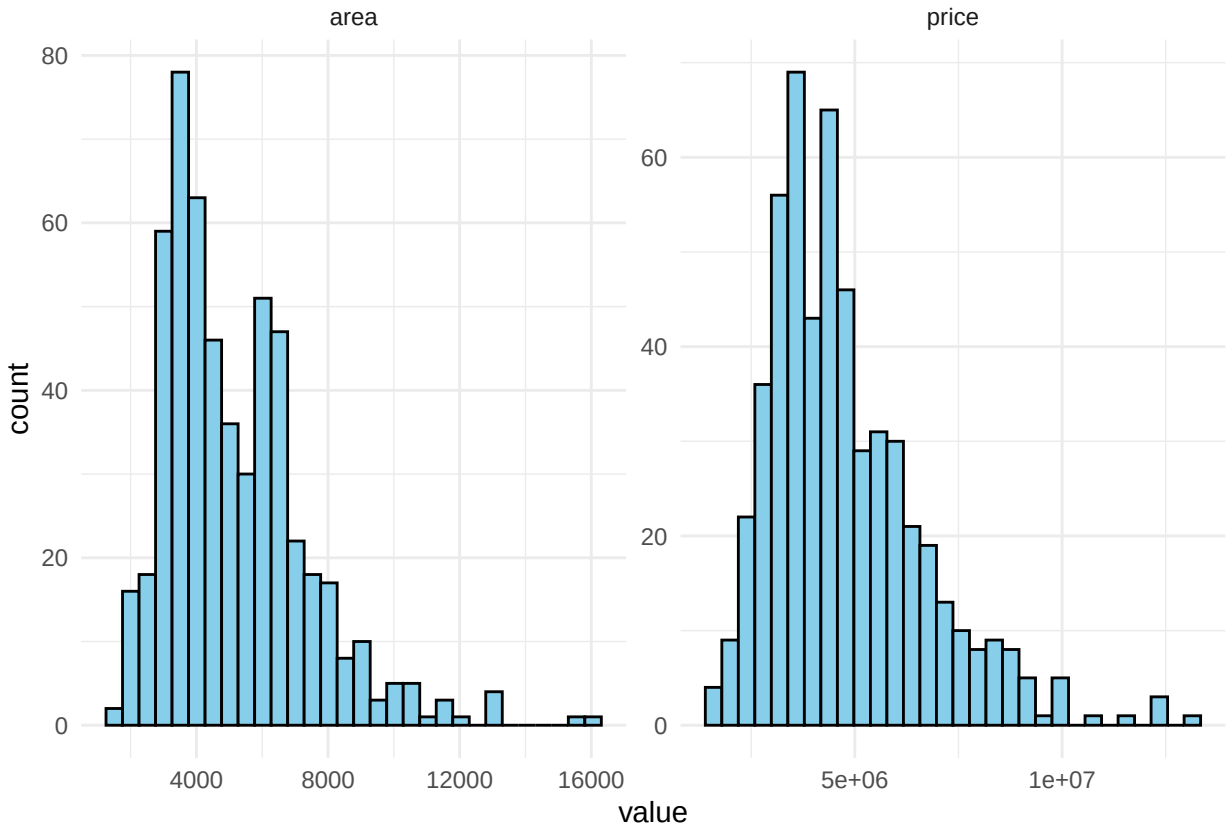
The number of bedrooms across the houses shows an average of approximately 3 bedrooms per property. By looking at the number of bedrooms, it suggests that the majority of the houses are designed for small to medium-sized families. The range of bedrooms from 1 to 6 indicates the diversity in the housing size. Houses with 1 or 2 bedrooms are likely to cater for individuals or couples. Properties with 3 bedrooms are common, which reflects a standard size for family homes in most housing markets. Houses with up to 6 bedrooms suggest a very niche for larger families or premium properties which is likely aimed at affluent buyers or those who needed more living spaces.

For bathrooms, the mean number is approximately 1.3 which indicates that most of the houses have at least one bathroom. Properties with up to 4 bathrooms are less common and are likely part of the higher end market. The distribution of the bathrooms reflects the housing design properties based on economic segments and household size.

The average number of stories is 1.8 which shows that most of the houses are single or two stories building. The highest number of stories is 4 which indicates a specific architectural or demographic need. The variations in the number of stories reflect differences in property design based on location and lifestyle presence.

For parking, the average parking space is 0.7 spots per house which indicates that many properties either lack parking altogether or offer limited parking spaces. However, there are some houses that provide up to 3 parking spots, but these are relatively rare. With this data, potential house owners can use this information to plan buying the house with or without parking spaces as many of the people prefer public transportation and owning a vehicle will add cost to their monthly expenditure.

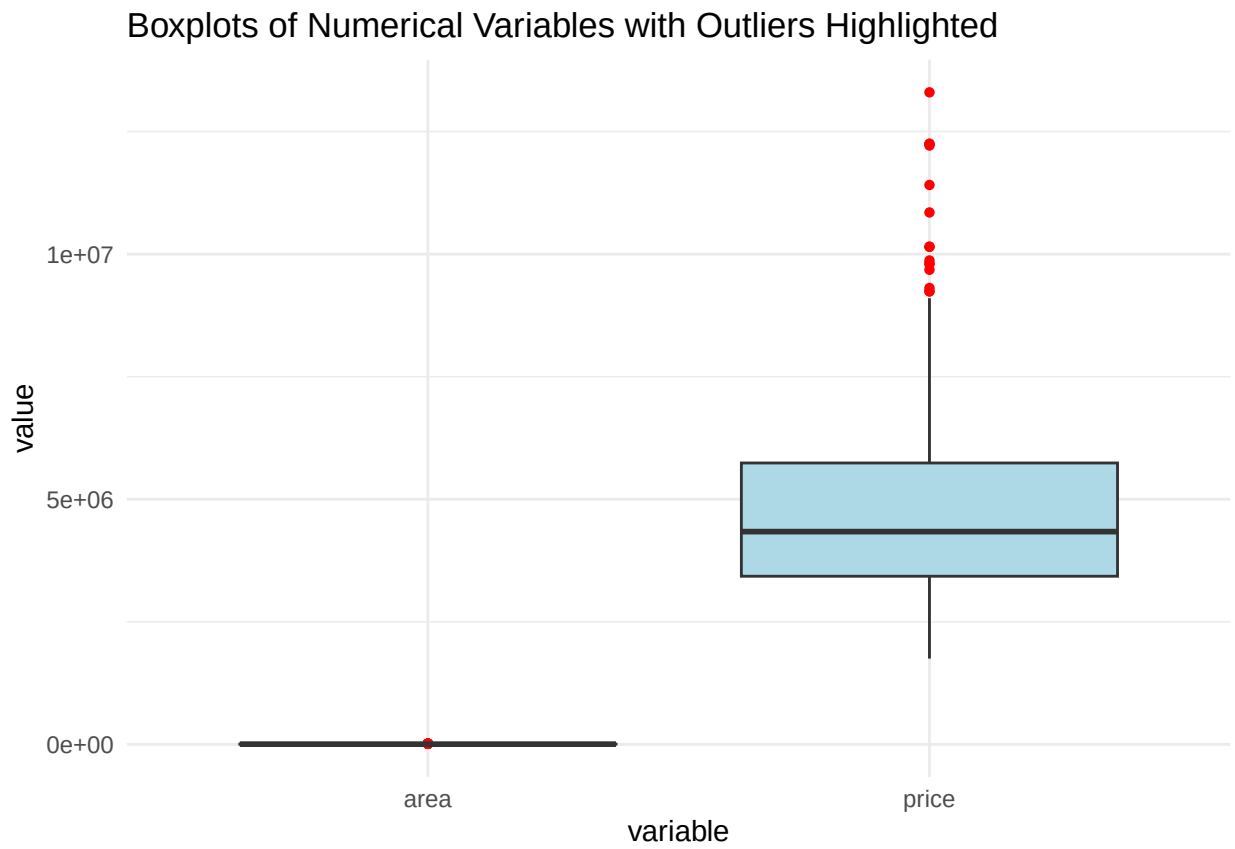
Overall, the data shown in the table is important for understanding the central tendencies and variations across the key variables.



The graph above shows histograms for the variables area and price which illustrate the frequency distribution. For area, the histogram reveals that most of the houses fall within the range of 3,000 to 8,000 square feet. Then, there is a sharp decline after above 10,000 square feet. The distribution of the histogram is rightly skewed, which suggests that while most of the houses are of moderate size, there are some exceptions that offered larger properties.

As for the house price, the distribution of the histogram is also rightly skewed. Most of the houses are priced between 3,000,000 to 7,000,000. However, there are some houses that are valued exceed 10,000,000, reflecting luxury housing or properties in prime locations.

These distributions for both histograms highlight the presence of outliers which will be further analyzed using boxplot.

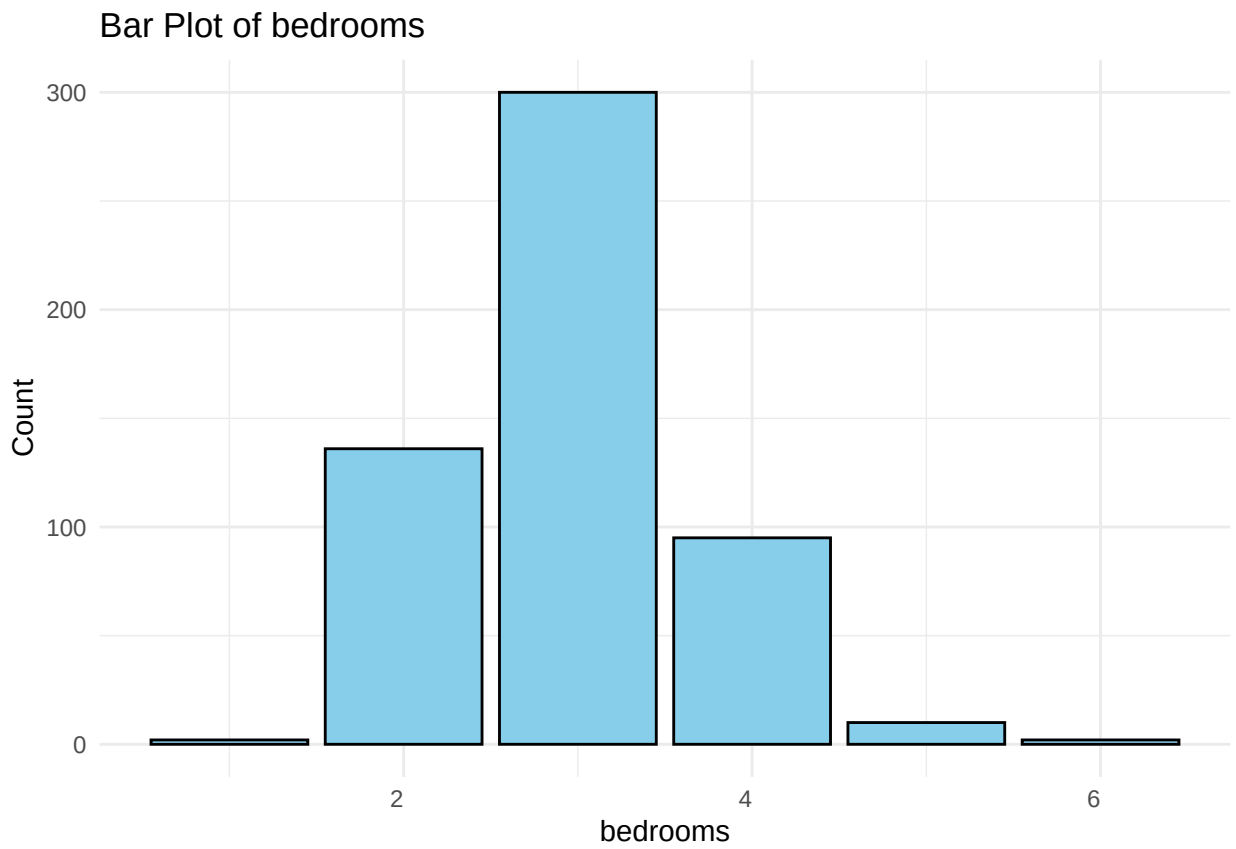


The boxplot visualizes the distribution of both price and area variables which highlight potential outliers. The outliers are highlighted with red dots to indicate clearly the presence of the outliers' figures in the boxplot.

For area, the values are concentrated near the lower end of the scale and almost close to 0 which suggests that there are no visible outliers for this variable as it has lower variance.

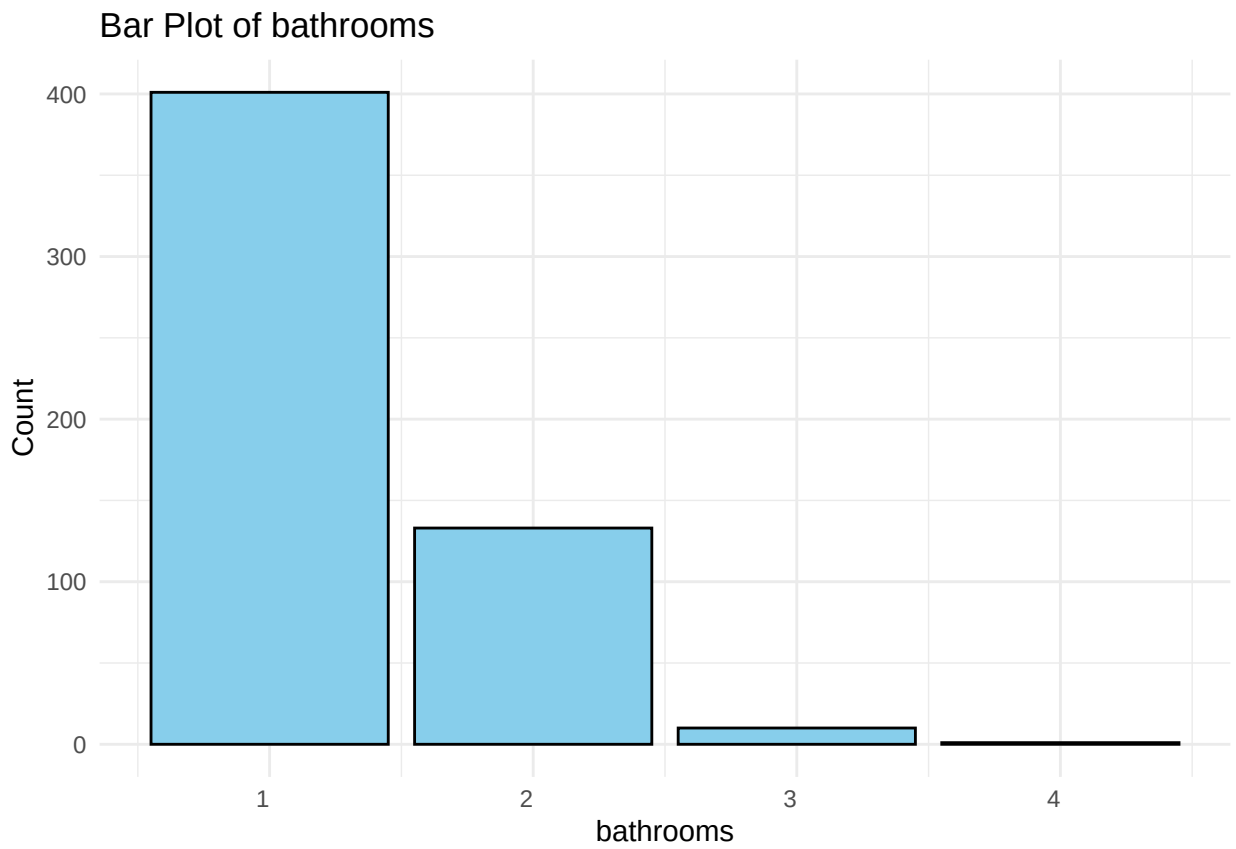
For price, the distribution is wider, with a higher median and value extending towards 5,000,000 and beyond. There are also several outliers indicating a significant number of unusually high price values.

The key insight from this boxplot is that price has a much broader range of values compared to area, and it also includes outliers. This shows that variability in prices is possibly due to factors like locations, market trends or differences in property characteristics. As for area, it has a relatively consistent range with no extreme variations highlighting the uniformity in the size of properties in this dataset.



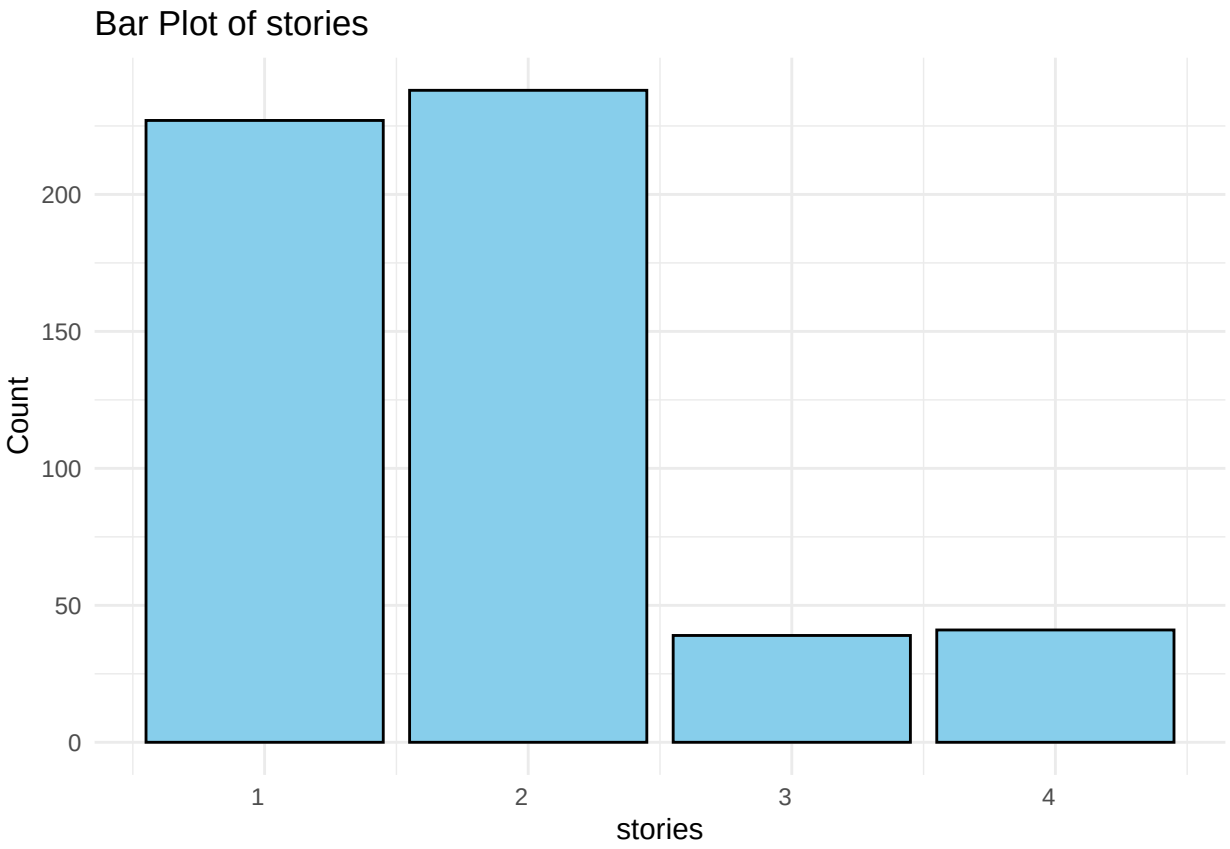
The bar plot above shows the number of houses by the number of bedrooms. The most common number of bedrooms is 3 which indicates the tallest bar showing the highest frequency (approximately 300 properties). The second and third most common of the number of bedrooms are 2 and 4 respectively. Properties with 1 or 6 bedrooms are not common as their count is very low (less than 100 properties).

The distribution is heavily skewed towards properties with 3 bedrooms which may suggest that most properties in this dataset are designed for medium-sized families. Larger properties with more than 4 bedrooms are very less, potentially because they cater to a niche market or are less affordable. Smaller properties with only 1 bedroom are also uncommon which suggests that single-person households may not be the main target for this dataset.



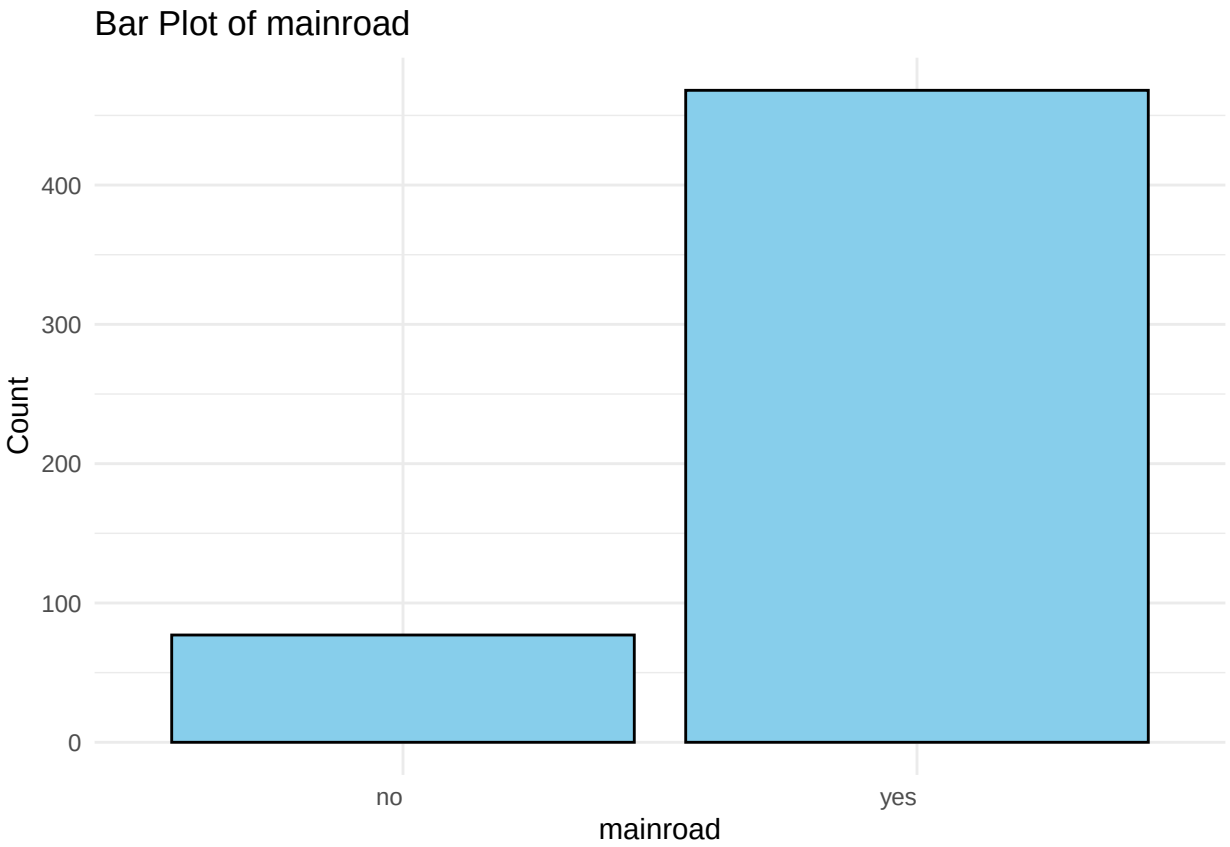
For the number of bathrooms, the properties with 1 bathroom are the most common with around 400 properties. The second most frequent is the house with 2 bathrooms, but their count is significantly lower than those with 1 bathroom with around 100 to 150 properties. Properties with 3 and 4 bathrooms are rare with a very low frequency.

Most properties have 1 bathroom which indicates that the dataset primarily represents smaller or more affordable properties. Houses with 2 bathrooms are also somewhat common which likely targeting slightly larger families or properties with additional facilities. The properties with 3 and 4 bathrooms are uncommon which indicates these are either luxury options or not widely available in the dataset.



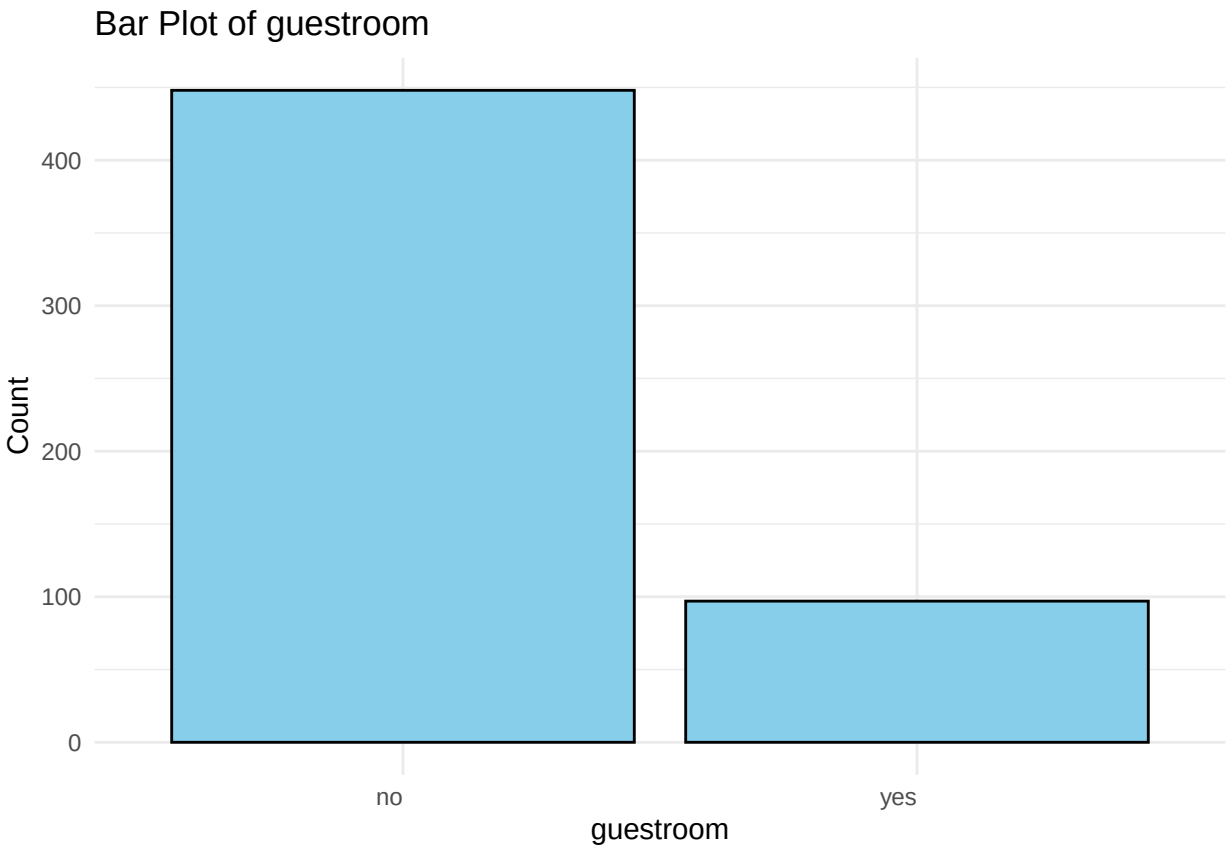
Properties with 1 and 2 stories are the most common, each with a frequency of about 200 to 250. Properties with 3 and 4 stories are uncommon with each having frequency of approximately 50 to 75 properties. Single and two stories house dominate the dataset which suggests that they are standard or most preferred for properties in this dataset.

Properties with 3 or more stories are uncommon which could indicate that these types of properties are luxurious. The similarity in the numbers for 1 and 2 stories suggest that both are popular choices for families or individuals which correlated with the preferences or budget.



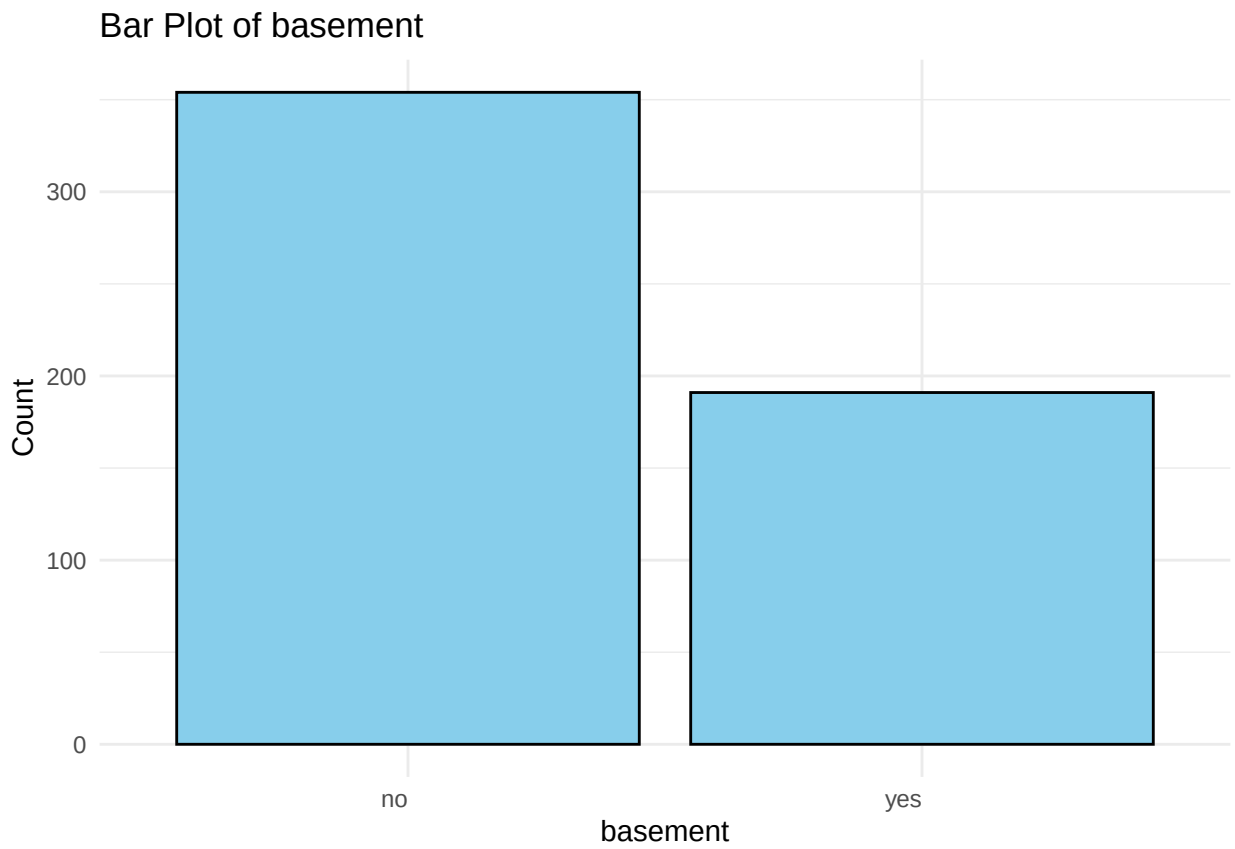
The bar plot above shows the distribution of properties based on whether the houses are located near a main road which are categorized by 'yes' and 'no'. Most of the houses are located on the main road with a frequency of over 400. A less proportion of properties are not on the main road with a frequency of around 100.

Houses which are near to main road are significantly more common in the dataset which suggests a preference for locations with better accessibility or visibility. On the other hand, the relatively lower count for properties not on a main road indicates less demand for these locations which is potentially due to less accessibility. The location of the properties also influences property prices, with those near a main road possibly having higher premium prices.



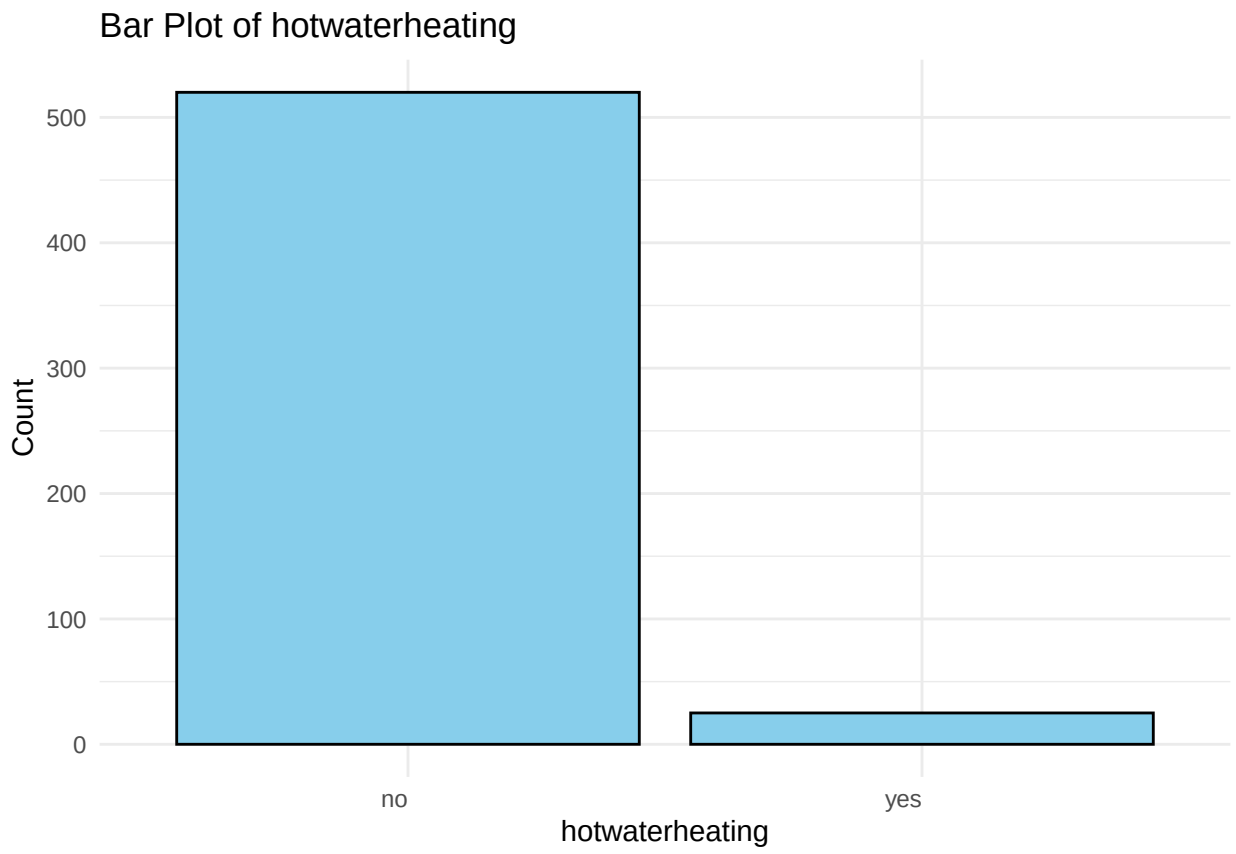
The above bar plot shows the distribution of properties whether the houses occupied with guestroom. From the graph above, the majority of houses do not have a guestroom with a frequency of 400 whereas the properties with guestrooms are only a small proportion with frequency of around 100.

By looking at the graph, properties without guestroom are significantly more common indicating that guestrooms are considered a luxury feature. Another reason could be that guestrooms are not a priority for most buyers in this dataset. Hence, the relatively lower count of houses with guestrooms might indicate that these are typically found in larger or more expensive properties.



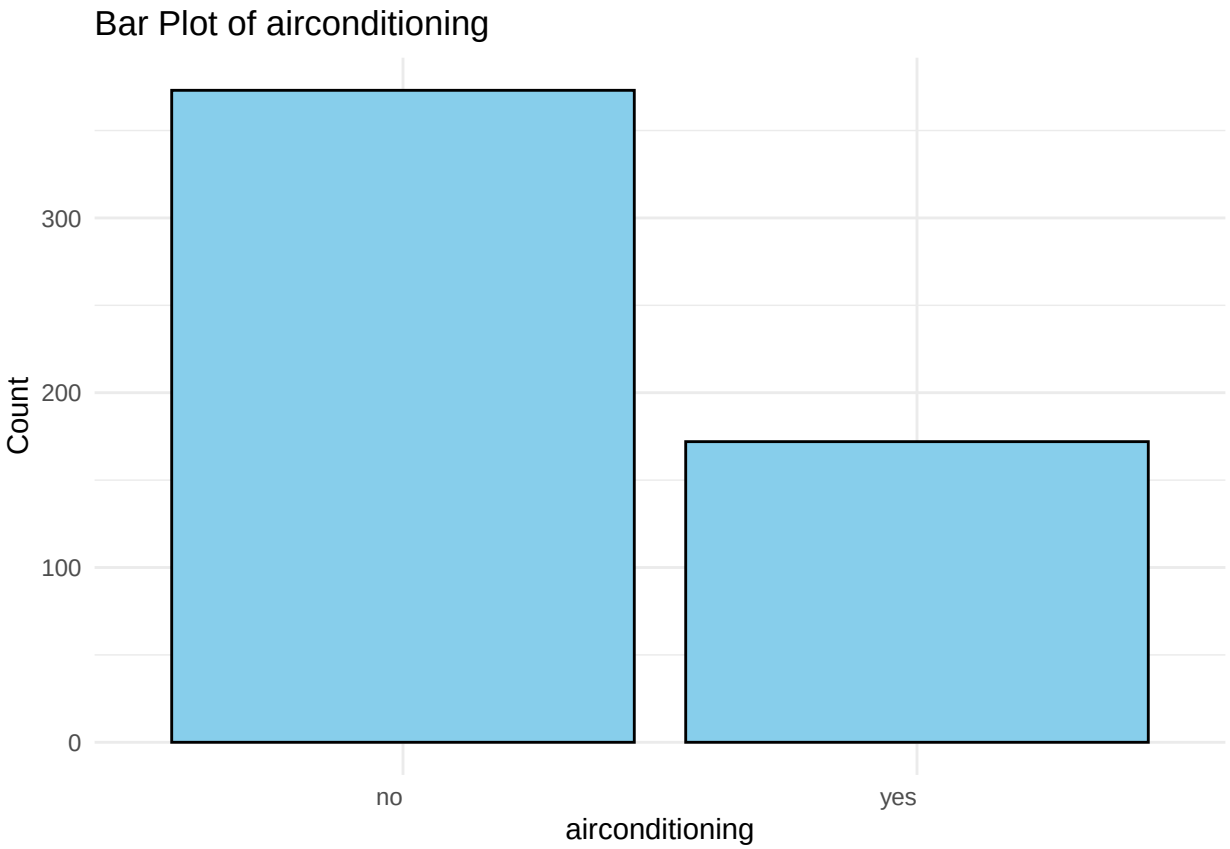
By looking at the graph, most properties do not have a guestroom with a frequency of over 400. A less proportion of houses do have a guestroom with a frequency of around 100.

As the graph illustrated, houses without a guestroom are significantly more common which suggests guestrooms are either considered a luxury feature or a feature that are not priority for most buyers in the dataset. The relatively lower frequency of properties with guestrooms might indicate that these are typically found in bigger houses or more expensive homes.



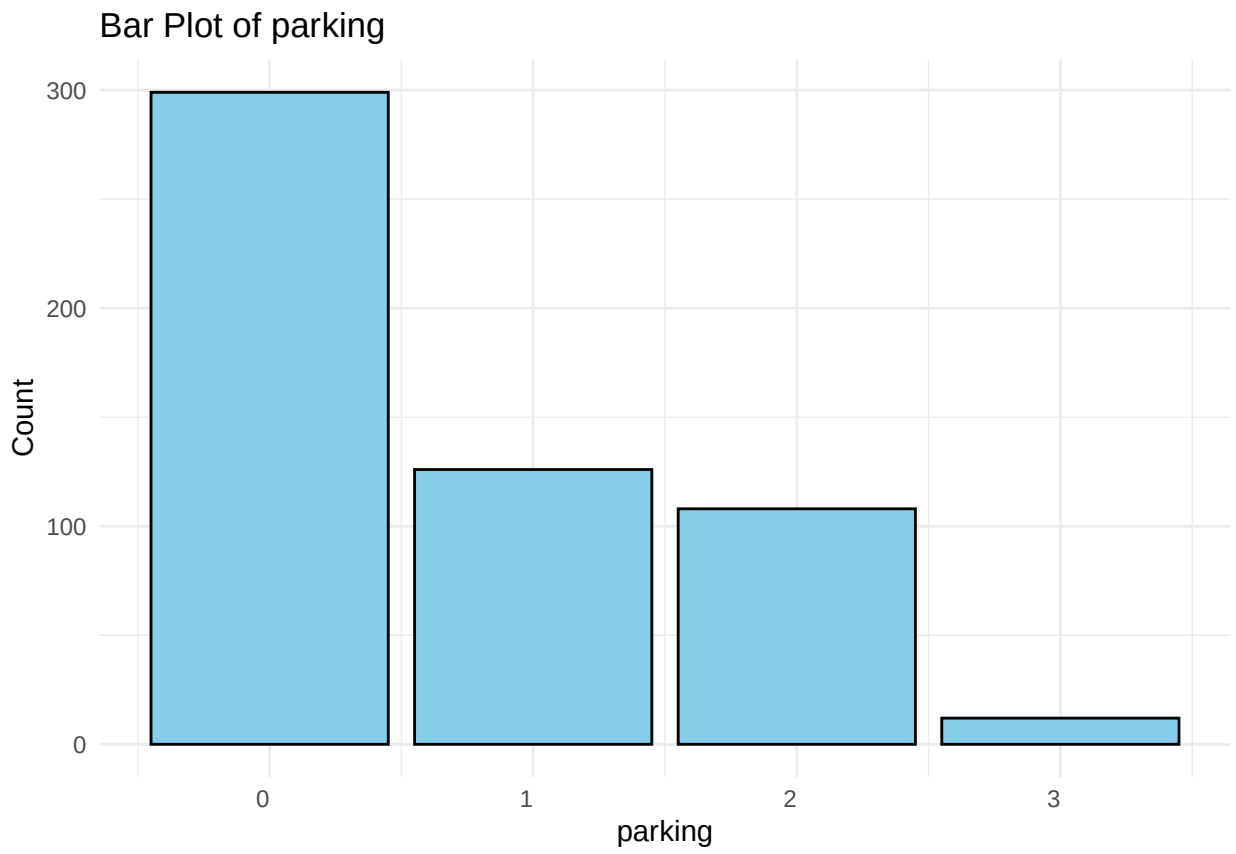
The above bar plot graph shows that the vast majority of properties which are over 500 houses do not have hot water heating and a very small number of properties which are less than 100 houses have water heating.

Houses with hot water could be targeted towards a niche market which appeals to premium buyers or those staying in colder climates where this feature adds value.



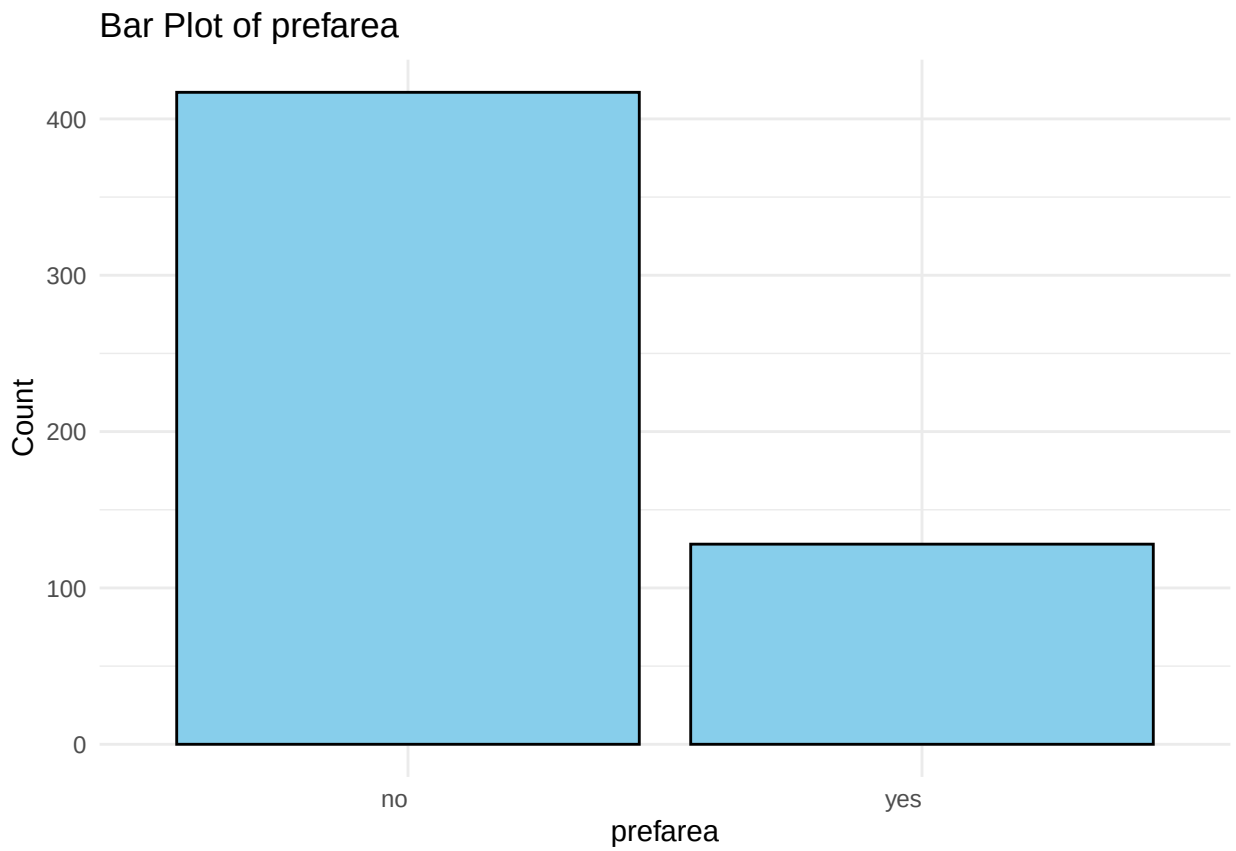
Most houses do not provide air conditioning with the count are more than 300 whereas a smaller proportion of houses provide air conditioning with a count between 100 to 200. The absence of air conditioning in most properties could be explained by the climate or buyers' preferences in the region of interest.

For instance, in areas with lower climate, air conditioning might not be seen as necessary. Properties with air conditioning could appeal to buyers who are looking for more comfort, particularly in hotter climates or during summer months.



The bar plot graph above shows that properties with 0 parking space are dominant, accounting for over 300 houses. The absence of parking area may indicate that the houses located in areas where parking is either shared, unavailable or unnecessary.

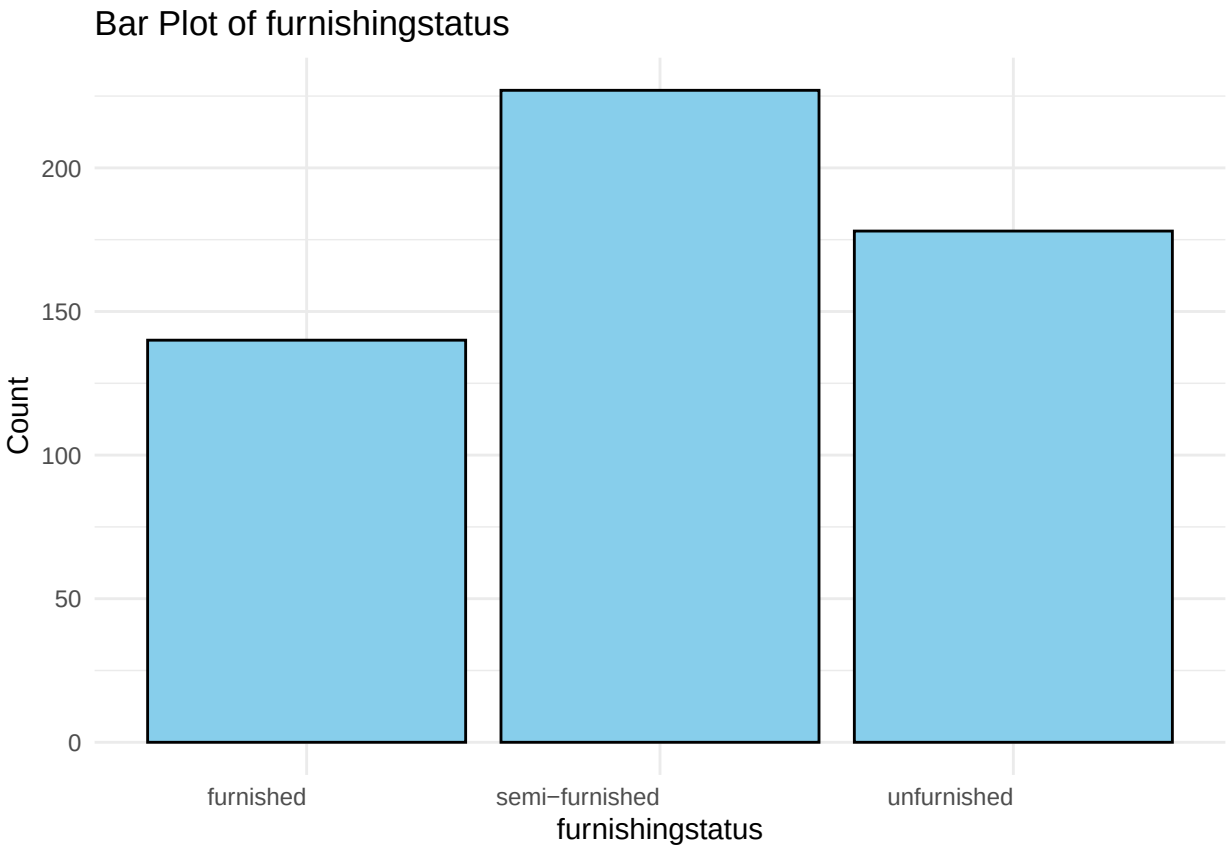
Properties with 1 and 2 parking spaces have almost an equal number of approximately 150 houses. These properties likely cater to households with smaller families or individuals with fewer vehicles. For houses with 3 parking spaces, it shows the least number with fewer properties accommodate for this amenity. These houses are likely more expensive and located in regions where lager land sizes permit more parking spots.



The graph above illustrates the distribution of houses based on whether they are in a preferred area or not. By looking at the graph, most of the properties fall into the non-preferred area with a count exceeding 400. This shows that most houses in the dataset are in regions that are not in a premium area thus the price is more affordable.

A small portion of houses are in a preferred area with a count of 100 to 150. This highlights that the properties which are located in these areas are less preferable due to higher cost and other restrictive factors.

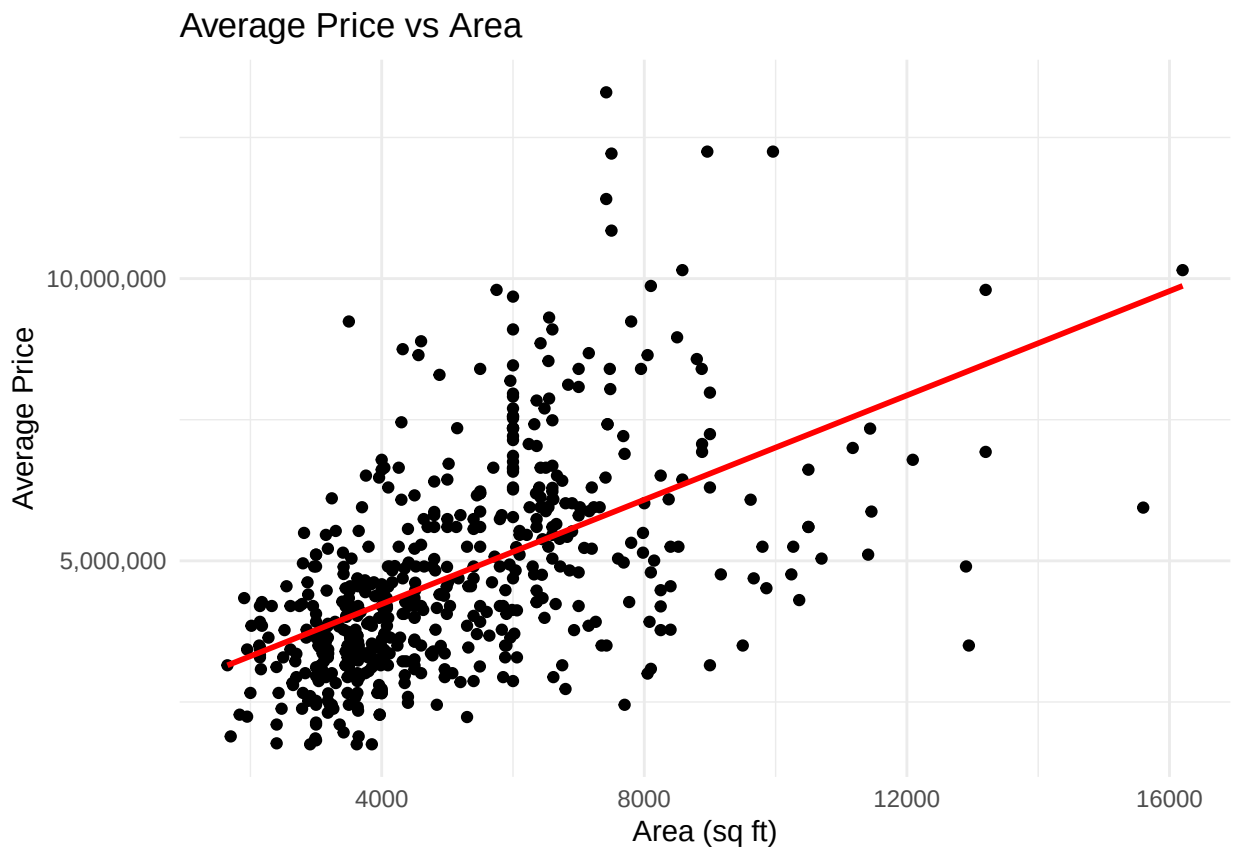
The relatively low number of houses in the preferred area indicates that buyers might prioritize affordability over locations. Non-preferred areas can offer more affordable prices with and still offer larger space area and other decent amenities.



The bar plot above illustrates the furnishing status of properties categorized into semi-furnished, unfurnished, and furnished. It shows that semi-furnished properties are the most common with a count exceeding 200, followed by unfurnished properties with around 175 counts. This suggests that a wide range of buyers or renters like to have some flexibility in customizing their properties.

Moreover, furnished properties are the least common with a count of approximately 145. These properties are mainly targeting certain consumers that seeking immediate move in options. The lower count also indicates that these fully furnished properties are only focused on the target customer base, as the costs associated with fully furnished properties are likely to be higher.

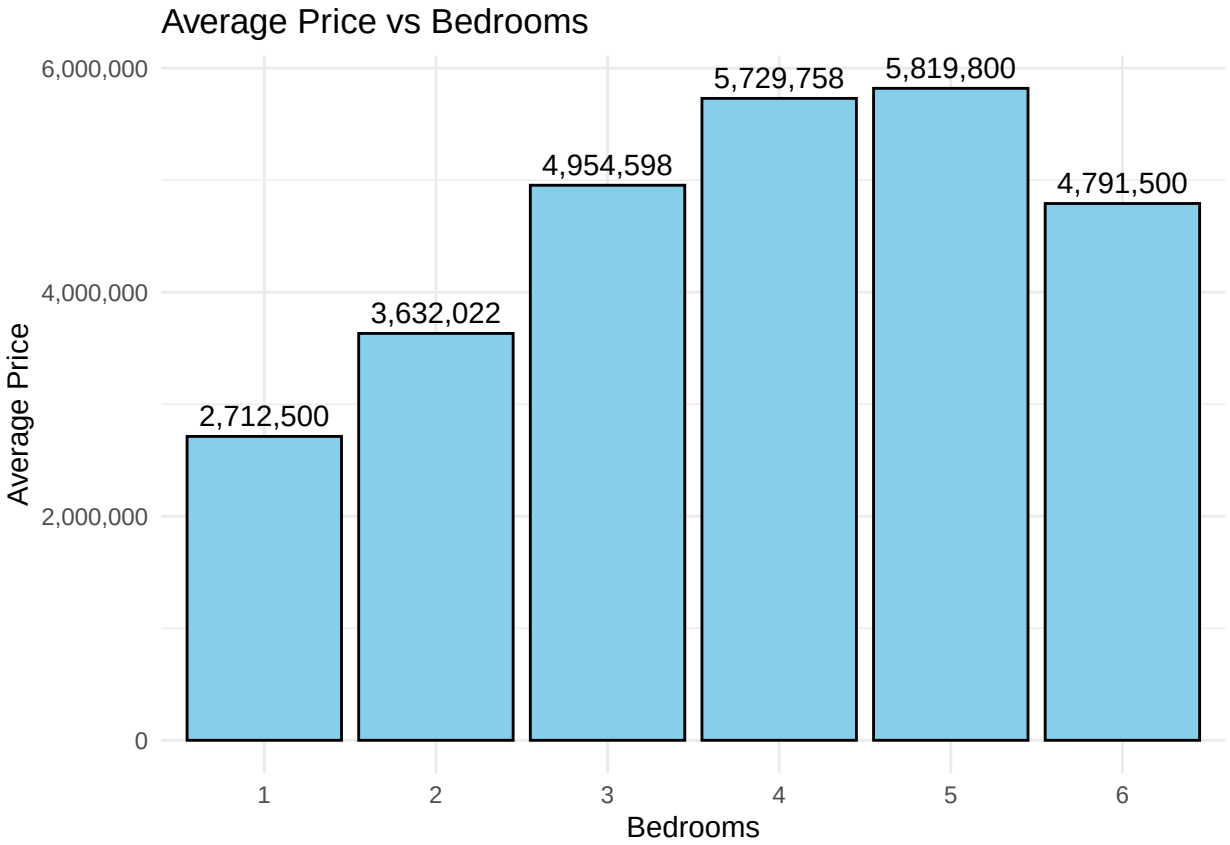
Overall, this distribution highlights the diverse needs and preferences in the housing market, with semi-furnished properties leading among the options and appealing to a wider audience.



The scatter plot above visualizes the relationship between average price and area (sq ft), with a fitted regression line highlighting the trend. The plot indicates a positive linear relationship, which indicating that the average price increases with the size of the property. This positive linear relationship aligns with the trend of real estate market, where larger properties generally have higher prices.

The plot also shows that data points are mainly clustered between 2,000 and 8,000 square feet and below average prices of 7,500,000. This suggests that most properties fall within this size and price range, suggesting that this range is a major market for properties. While a few data points at higher area and price levels represent the premium or luxury properties.

Overall, this scatter plot reveals that area is a significant factors of housing price, while other potential variables like bathrooms or stories can lead to price changes.



The bar plot illustrates the relationship between the average price of properties and the number of bedrooms. Properties with 5 bedrooms have the highest average price of approximately 5,819,800, followed closely by 4-bedroom properties at 5,729,758. This indicates that properties with more bedrooms are likely to fall in a higher price range. Moreover, properties with only 1 bedroom have the lowest average price at 2,712,500. These properties are generally smaller and more affordable, targeting buyers that are more sensitive to price.

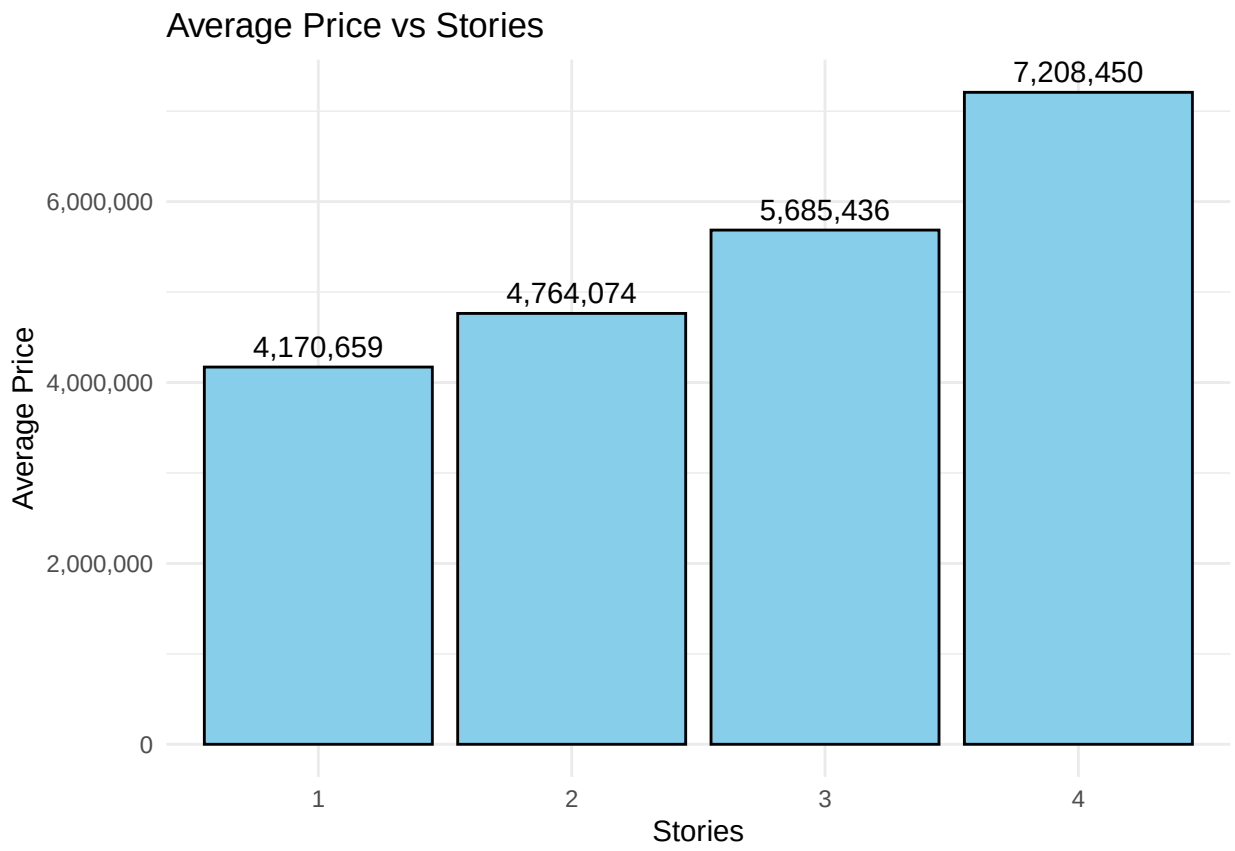
Notably, 6 bedrooms properties have a lower average price than properties with 3 to 5 bedrooms. This could possibly be influence by other factors such as location or amenities, which affecting housing price.

Overall, the bar plot indicates a positive relationship between the number of bedrooms and average price, with 5-bedroom properties being the most expensive on average. However, the abnormal price of 6-bedroom properties suggests that additional factors other than the number of bedrooms could possibly affect the housing price.

The bar plot shows the relationship between the average price of properties and the number of bathrooms. Properties with 4 bathrooms have the highest average price at 12,250,000, followed by 3 bathrooms properties with an average price of 7,282,100. This huge gap indicates a strong correlation between luxury properties and multiple bathrooms.

Furthermore, properties with 1 and 2 bathrooms, priced at an average of 6,209,206 and 4,206,913 respectively. These properties fall within a more affordable pricing range, making them accessible to a broader market.

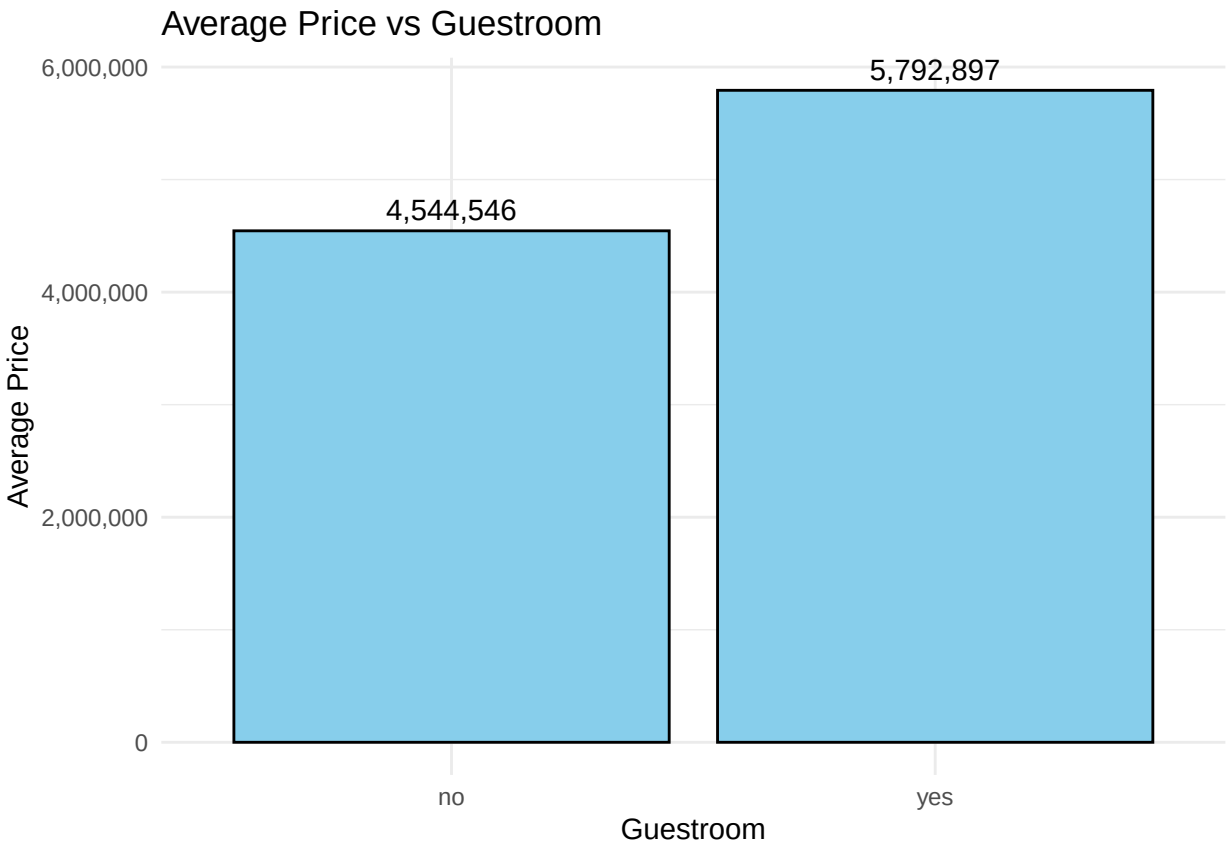
Overall, the bar plot shows a positive relationship between housing price and number of bathrooms, suggesting that number of bathrooms serve as a significant factor of housing price.



The bar plot shows the relationship between the average price of properties and the number of stories. It indicates a positive relationship between number of stories and average price, which properties with more stories generally have higher price.

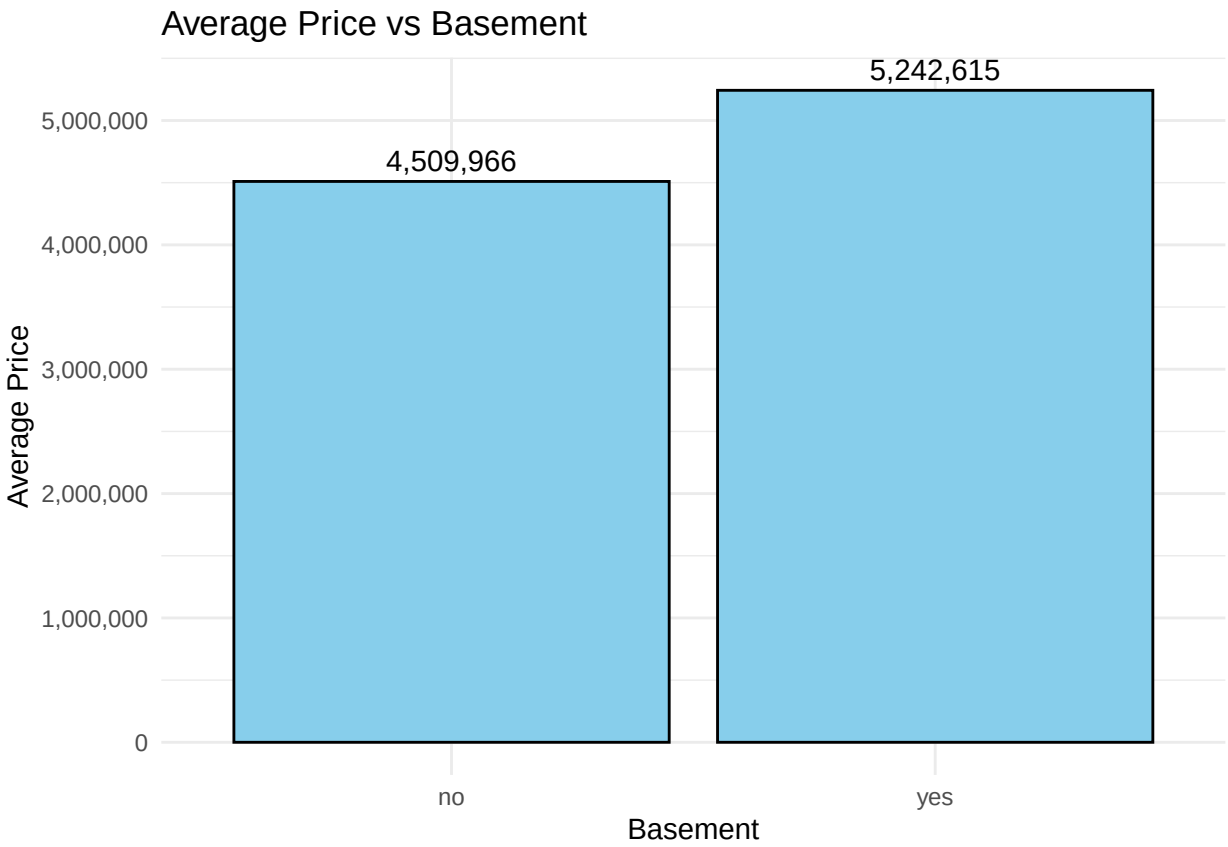
Properties with 4 stories have the highest average price at 7,208,450 followed by 3 stories properties with an average price of 5,685,436. The premium prices suggest a strong correlation between housing price and number of stories, which mainly targeting luxury properties market.

Moreover, properties with 1 and 2 stories have an average price of 4,764,074 and 4,170,659 respectively. These properties fall within a more affordable pricing range, mainly targeting buyers or renters that are more sensitive to housing price.



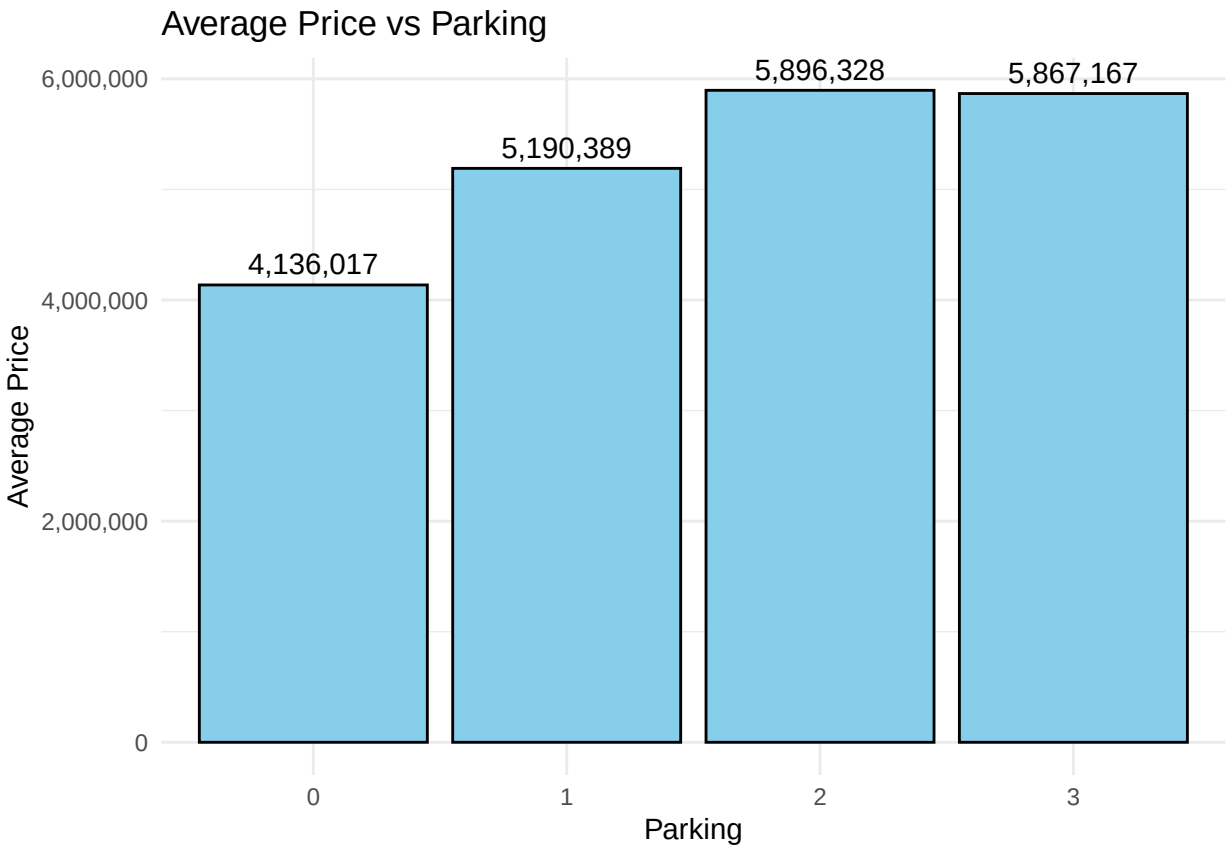
The bar plot above illustrates the relationship between the average price of properties and the presence of guestrooms. It shows that properties with guestroom have a higher average price of 5,792,697 which is slightly higher than the average price of 4,544,546 for properties without a guestroom. This shows that having a guestroom is a desirable feature that have added value to a house, which will significantly affecting the house price.

The trend highlights the impact of guestroom on property housing, which shows that by having a guestroom, it will increase the property price. This shows that the buyers are willing to pay higher price for a property with guestroom as it adds substantial value to a house.



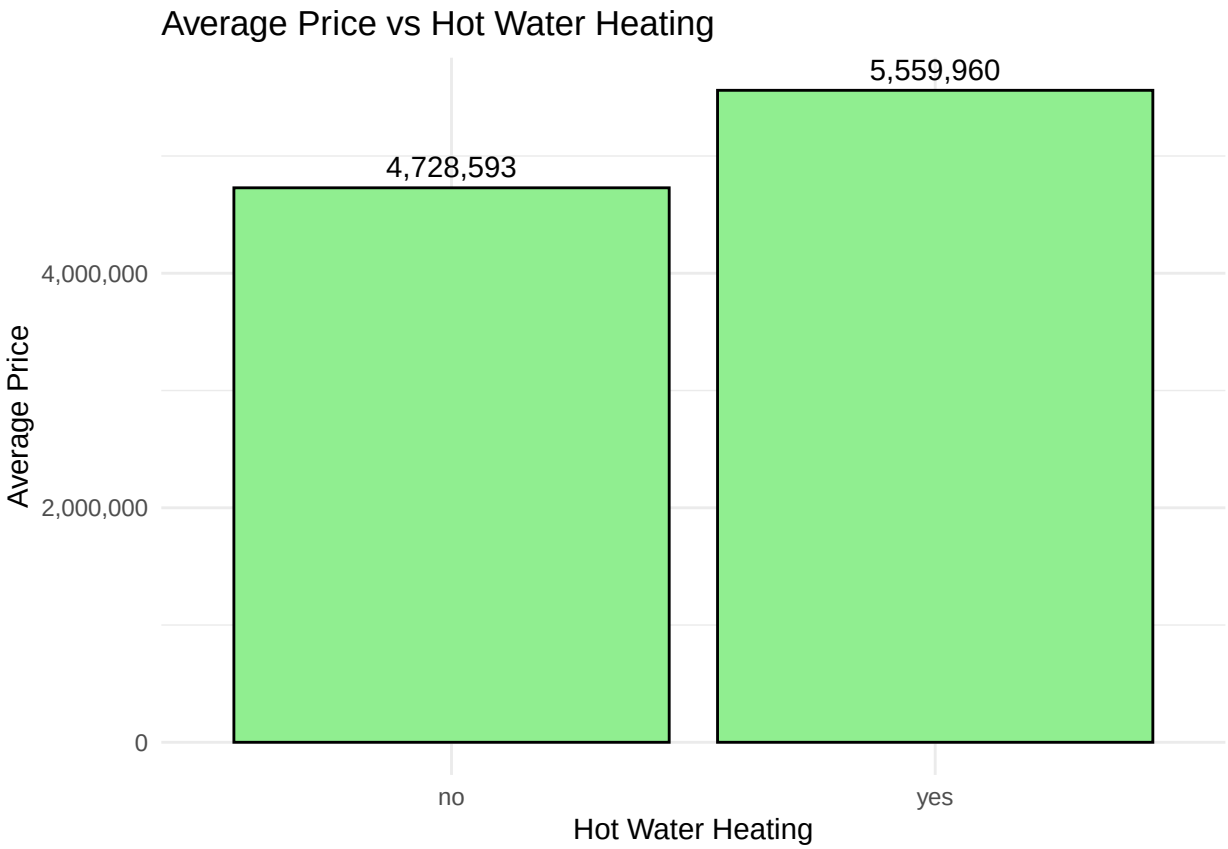
The bar plot illustrates the relationship between the average price of properties and the presence of basements. It indicates that properties with basements have a higher average price at 5,242,615, while properties without basement have a lower average price of 4,509,966. This reveals the added value and functionality of basements, such as additional storage or space, which significantly affecting the housing price.

The trend emphasizes the impact of basements on property pricing, which additional basements enhancing the value of properties. This suggests that consumers are willing to pay more for a property with a basement due to the added value.

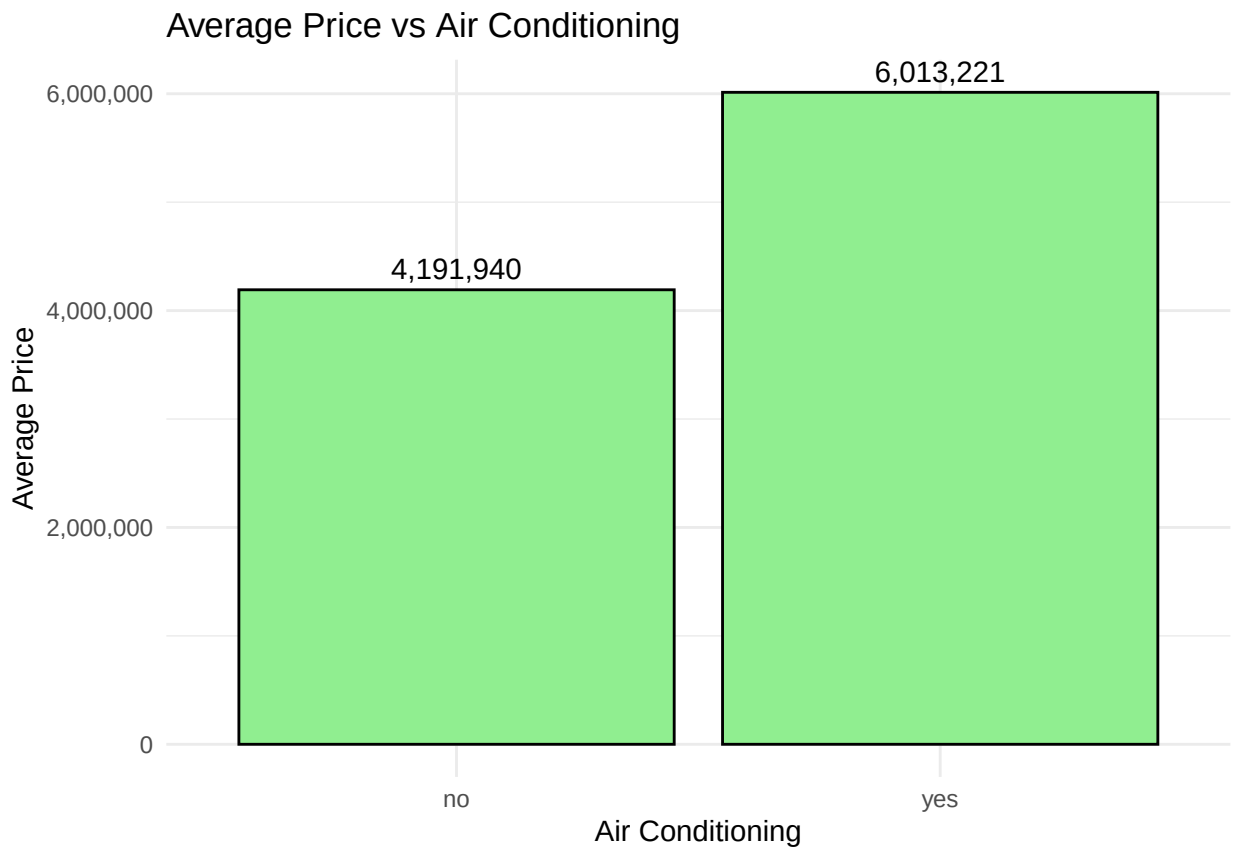


The bar plot shows the relationship between the average price of properties and the number of parking. Properties with 3 parking have the highest average price at 5,896,328, followed by 4 parking properties with an average price of 5,867,167. Additionally, Properties with 1 and 2 parking have an average price of 4,136,017 and 5,190,389 respectively. This suggests that properties with more parking tend to be more expensive.

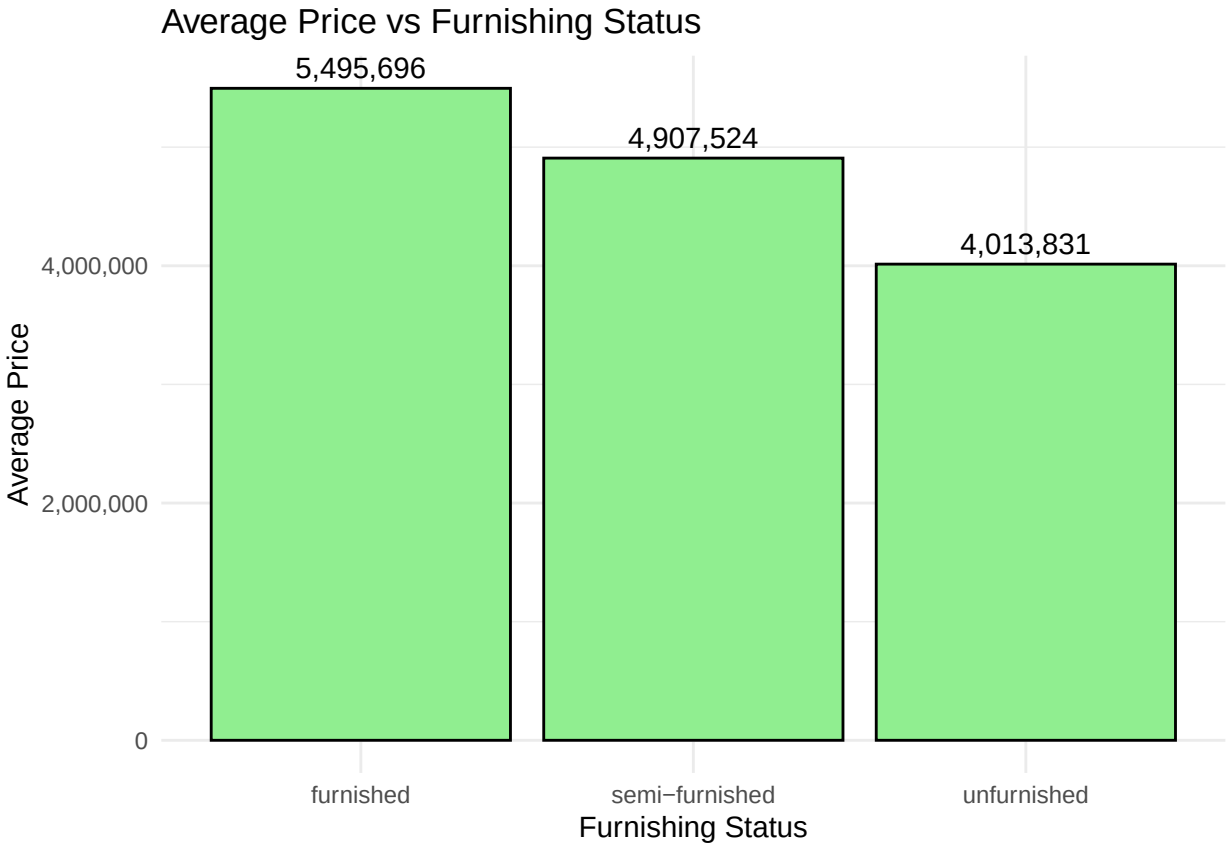
Overall, the bar plot demonstrates a positive relationship between the number of parking and the average price. This suggests that parking is a valuable factors of housing price, which tend to be more expensive as the number of parking spaces increases.



The bar plot illustrates the relationship between the average price of properties and the presence of water heating. Properties with water heating have a higher average price of 5,559,960, while properties without water heating have a lower average price of 4,728,593. This suggests that water heating is a valuable feature that can significantly affect the owner's experiences, resulting in a significant difference in housing prices.

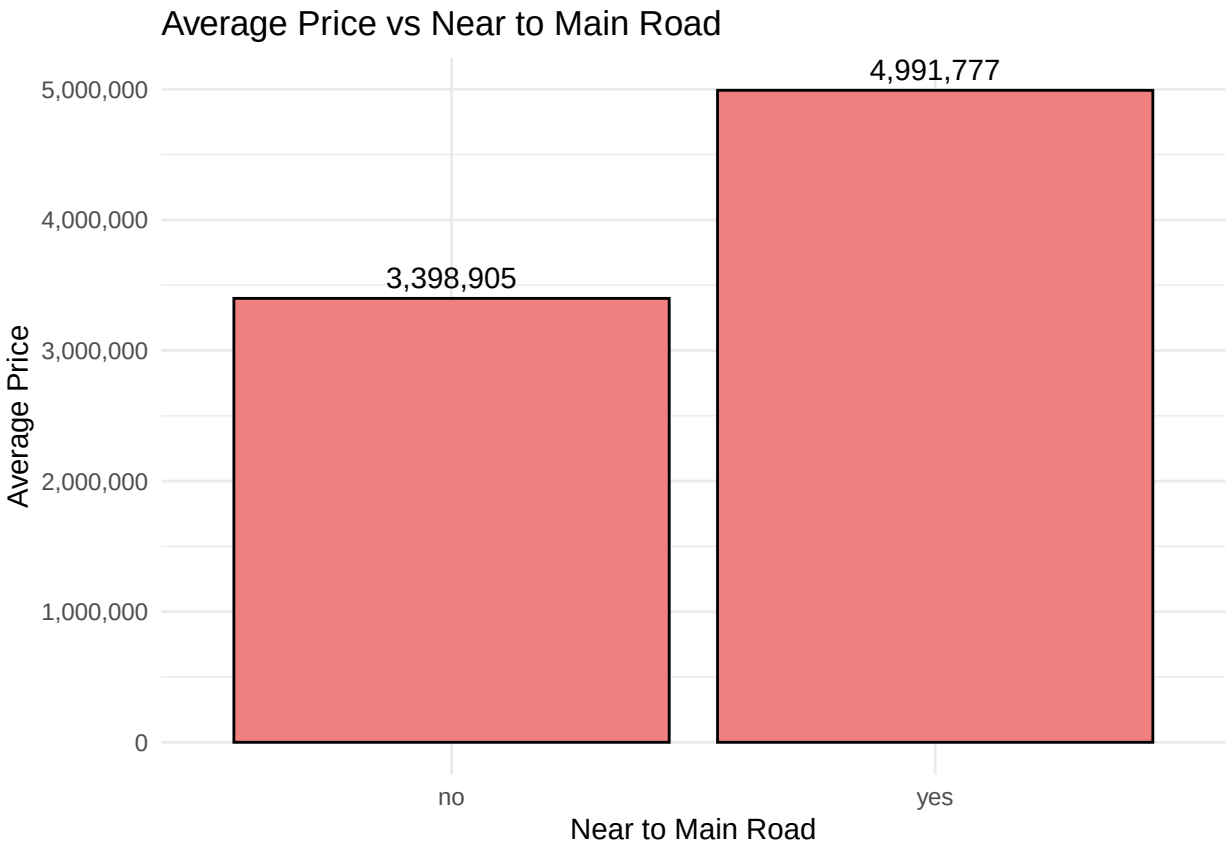


The bar plot shows the relationship between the average price of properties and the presence of air conditioning. Properties with air conditioning have a higher average price of 6,013,221, while properties without air conditioning have a lower average price at 4,191,940. This suggests that air conditioning is an valuable feature that can significantly affect the living quality, which properties with air conditioning generally fall in a higher price range.

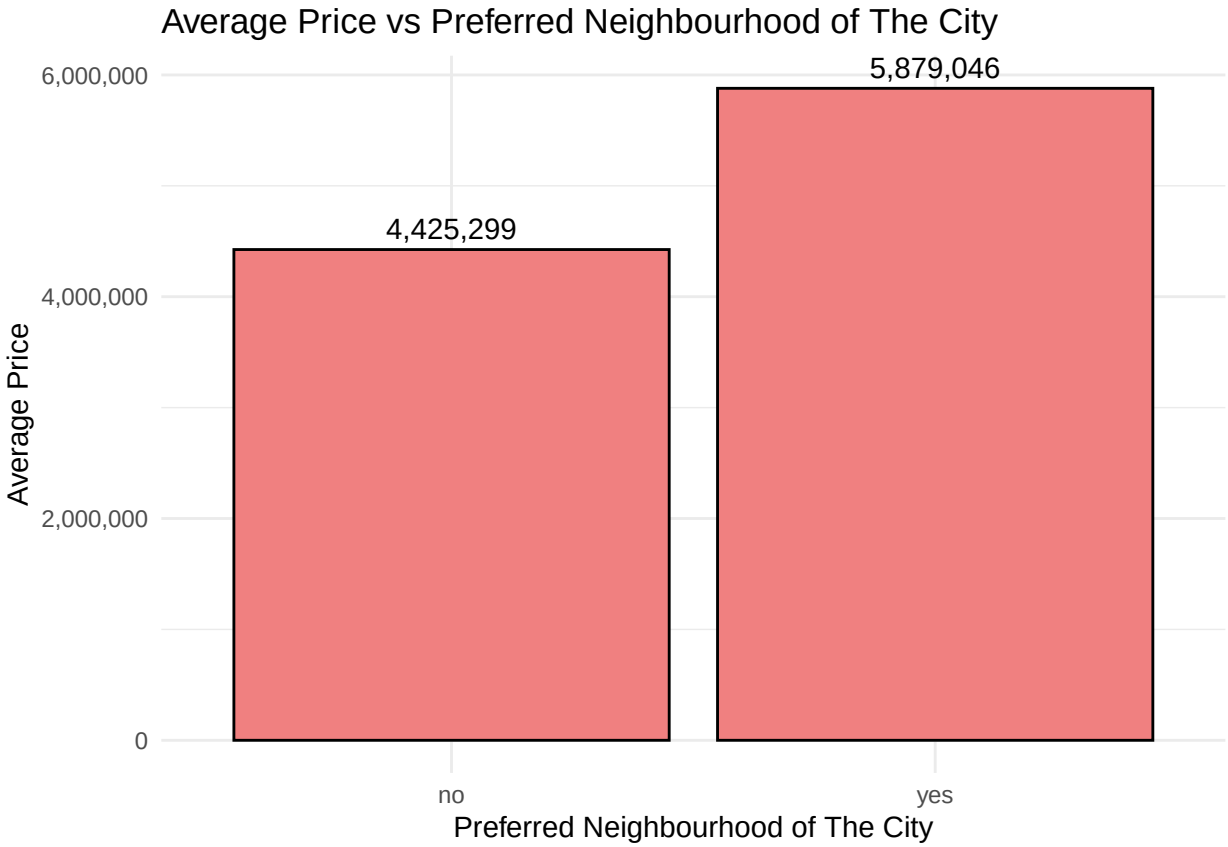


The bar plot shows the relationship between the average price of properties and the furnishing status. It reveals that furnished properties have the highest average price at 5,495,696. While semi-furnished properties have an average price of 4,907,524, unfurnished properties have the lowest average price of 4,013,831. This suggests that fully furnished and semi-furnished properties are perceived as more valuable due to higher cost of the properties.

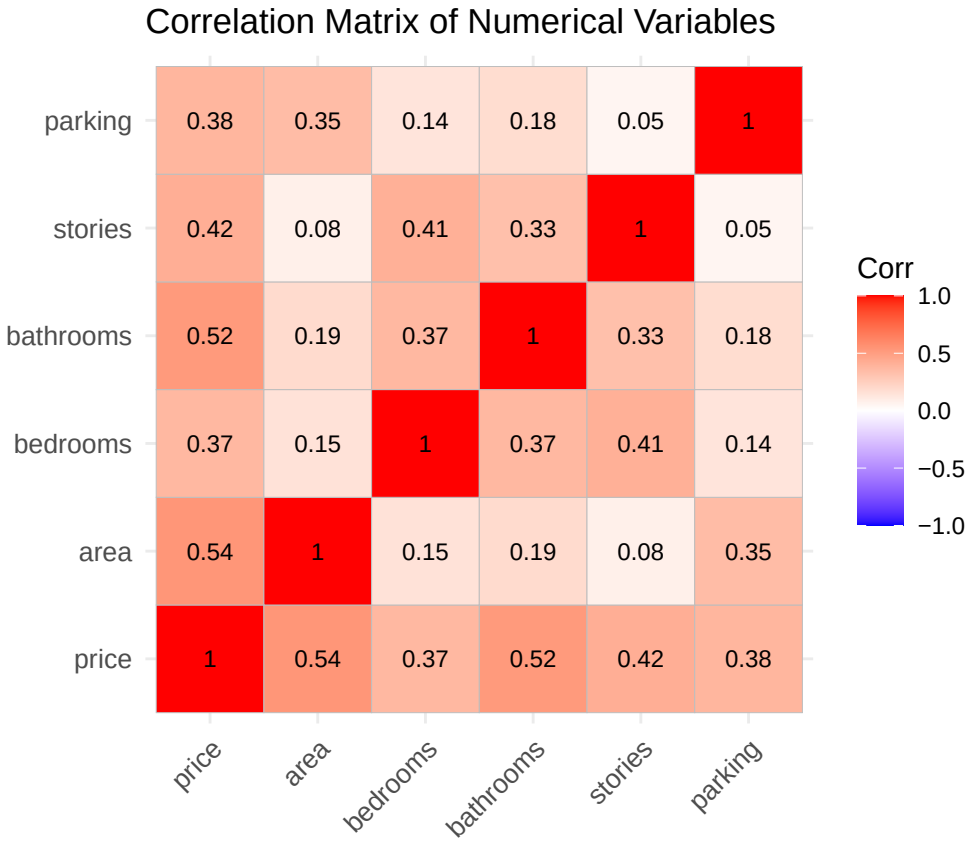
In other words, there is a positive correlation between cost and price of the properties. Furnished and semi-furnished are often indicative of a higher cost of the property, resulting in higher pricing.



The bar plot illustrates the relationship between the average price of properties and the distance to main road. It shows that properties near to the main road have a significantly higher average price of 4,991,777, while properties not near the main road have a lower average price of 3,398,905. This suggests that nearer distance to a main road is a valuable feature for buyers, likely due to the convenience of accessibility in the area. In this case, properties that are near to the main road tend to have higher average price compared to properties that are not.



The bar plot shows the relationship between the average price of properties and whether they are located in a preferred neighbourhood of the city. It indicates that properties located in preferred neighbourhoods have a higher average price of 5,879,046, while properties that are not located in preferred neighbourhoods have a lower average price at 4,425,299. This suggests that these areas are more popular, possibly due to factors such as a more convenient and safer environment. Location of a property has a significant influence on the housing price, which buyers are willing to pay more for a better location.



The correlation matrix above indicates the relationships between key numerical variables in the dataset. The analysis reveals that price has the strongest correlation with area (0.54), indicating that properties with larger areas tend to be more expensive. This aligns with the typical market trend that size can significantly impact property prices, making area a critical factor in pricing.

Moreover, price also has a strong correlation with bathrooms (0.52). This suggests that properties with more bathrooms tend to be more expensive, reflecting the relationship between the number of bathrooms and housing prices. Similarly, the positive correlation between price and stories (0.42) suggests that multi-story properties generally fall into a higher price range.

The number of bedrooms (0.37) and parking (0.38) shows a weaker correlation with Price. This relatively lower correlation suggests that additional bedrooms and parking do not have a significant impact on property prices like important factors such as area and bathrooms.

In summary, the correlation matrix reveals that area as the most significant factor on property prices, followed by bathrooms and stories. Although other variables like bedrooms and parking also contribute, their correlation is relatively weaker.

Model Selection

Model 1 - Multiple Linear Regression (all variables)

```
##
## Call:
## lm(formula = price ~ area + bedrooms + bathrooms + stories +
##     guestroom + basement + parking + hotwaterheating + airconditioning +
##     mainroad + prefarea + furnishingstatus, data = df)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2619718  -657322   -68409   507176  5166695
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       42771.69  264313.31   0.162 0.871508
## area                                244.14      24.29  10.052  < 2e-16 ***
## bedrooms                         114787.56   72598.66   1.581 0.114445
## bathrooms                        987668.11  103361.98   9.555  < 2e-16 ***
## stories                          450848.00   64168.93   7.026 6.55e-12 ***
## guestroomyes                     300525.86  131710.22   2.282 0.022901 *
## basementyes                      350106.90  110284.06   3.175 0.001587 **
## parking                          277107.10   58525.89   4.735 2.82e-06 ***
## hotwaterheatingyes               855447.15  223152.69   3.833 0.000141 ***
## airconditioningyes               864958.31  108354.51   7.983 8.91e-15 ***
## mainroadyes                      421272.59  142224.13   2.962 0.003193 **
## prefareayes                      651543.80  115682.34   5.632 2.89e-08 ***
## furnishingstatussemi-furnished  -46344.62  116574.09  -0.398 0.691118
## furnishingstatusunfurnished    -411234.39  126210.56  -3.258 0.001192 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1068000 on 531 degrees of freedom
## Multiple R-squared:  0.6818, Adjusted R-squared:  0.674
## F-statistic: 87.52 on 13 and 531 DF,  p-value: < 2.2e-16
```

The linear regression model aims to predict house prices using a range of features, including area, number of bedrooms, bathrooms, stories, and other factors related to property characteristics and amenities. The analysis reveals a strong relationship between these predictors and house prices, indicated by a high F-statistic 87.52 and p value less than 0.05 suggesting the overall model is statistically significant. The adjusted R-squared value of 0.674 indicates that approximately 67.4% of the variability in house prices is explained by the model.

Among the predictors, area, bathrooms, stories, guestroom, basement, parking, hot water heating, air conditioning, main road, preferred area, and furnishing status are included. The coefficients indicate the direction and magnitude of each predictor's effect on house price. Significant predictors include area, number of bathrooms, stories, and others such as air conditioning and preferred area, implying these have substantial effects on house prices. For instance, having air conditioning adds an estimated $864,958 to the price, while being in a preferred area contributes an additional $651,543 on average.

```
##                      GVIF Df GVIF^(1/(2*Df))
## area             1.325250  1        1.151195
## bedrooms         1.369477  1        1.170246
## bathrooms        1.286621  1        1.134293
## stories          1.478055  1        1.215753
## guestroom        1.212838  1        1.101289
## basement         1.323050  1        1.150239
## parking          1.212837  1        1.101289
## hotwaterheating  1.041506  1        1.020542
## airconditioning  1.211840  1        1.100836
## mainroad         1.172728  1        1.082926
## prefarea         1.149196  1        1.072006
## furnishingstatus 1.109639  2        1.026350
```

In this analysis, all predictors show VIF values well below 2, suggest that multicollinearity is not a concern for this model, as none of the predictors exhibit a VIF high enough to indicate problematic correlations. This implies that the predictors can be considered independent of one another, supporting the reliability of the estimated coefficients and the robustness of the model's outputs.

Model Cat - change "area" into categorical based on Q1, Q2, and Q3

```
##
## Call:
## lm(formula = price ~ area_category + bedrooms + bathrooms + stories +
##     guestroom + basement + parking + hotwaterheating + airconditioning +
##     mainroad + prefarea + furnishingstatus, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2787988  -664626   -48516   523674  5049977
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1953479     297557   6.565 1.24e-10 ***
## area_categoryMid                 -713489     126618  -5.635 2.85e-08 ***
## area_categorySmall              -1340598     149690  -8.956  < 2e-16 ***
## bedrooms                          133803      73784   1.813 0.070330 .
## bathrooms                         997959     105209   9.485  < 2e-16 ***
## stories                           443323      65345   6.784 3.13e-11 ***
## guestroomyes                      249511     135102   1.847 0.065329 .
## basementyes                       325546     112271   2.900 0.003891 **
## parking                           300392      59683   5.033 6.62e-07 ***
## hotwaterheatingyes                838004     227102   3.690 0.000247 ***
## airconditioningyes                875536     110577   7.918 1.43e-14 ***
## mainroadyes                       437629     146986   2.977 0.003041 **
## prefareayes                       664779     120202   5.531 5.02e-08 ***
## furnishingstatussemi-furnished    -66166     119203  -0.555 0.579082
## furnishingstatusunfurnished      -462491     128831  -3.590 0.000362 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1087000 on 530 degrees of freedom
## Multiple R-squared:  0.671,  Adjusted R-squared:  0.6624
## F-statistic: 77.23 on 14 and 530 DF,  p-value: < 2.2e-16
```

The linear regression model analyzed the relationship between housing prices and various predictors, including the newly categorized variable area_category (Small, Mid, Big), along with bedrooms, bathrooms, stories, and several binary variables such as guestroom, basement, parking, hotwaterheating, airconditioning, mainroad, prefarea, and furnishingstatus. The model revealed that properties in the "Mid" area category had a significant negative impact on price, with an estimate of -713,489, while properties in the "Small" category had an even greater negative impact, with an estimate of -1,340,598, both statistically significant at $p < 0.001$. Bathrooms, air conditioning, and stories showed the most substantial positive associations with price, with bathroom addition increasing price by 997,959 and air conditioning adding 875,536. Variables like basement and parking also had a notable positive impact, with estimates of 325,546 and 300,392, respectively. The model demonstrated that housing features significantly influence prices, with area size playing a critical role in determining value. Based on the comparison of adjusted R-squared values, model1 outperforms model_cat, making it the preferred choice.

```
##                       GVIF Df GVIF^(1/(2*Df))
## area_category     1.462164  2        1.099636
## bedrooms          1.365717  1        1.168639
## bathrooms         1.286997  1        1.134459
## stories           1.479808  1        1.216474
## guestroom         1.232045  1        1.109975
## basement          1.323805  1        1.150567
## parking           1.217734  1        1.103510
## hotwaterheating   1.041449  1        1.020514
## airconditioning   1.218478  1        1.103847
## mainroad          1.209325  1        1.099693
## prefarea          1.197897  1        1.094485
## furnishingstatus 1.120097  2        1.028760
```

The Variance Inflation Factor (VIF) analysis shows that multicollinearity among the predictors in the model is not a significant concern, as all VIF values are well below the commonly accepted threshold of 5. Overall, the predictors in the model are well-suited for linear regression analysis without any major issues of multicollinearity affecting the reliability of the coefficient estimates.

Model_interaction - area*bedrooms

```
##
## Call:
## lm(formula = price ~ area + bedrooms + area * bedrooms + bathrooms +
##     stories + guestroom + basement + parking + hotwaterheating +
##     airconditioning + mainroad + prefarea + furnishingstatus,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2808412  -625920   -63939   493283  5061408
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     938269.87  506723.28   1.852 0.064633 .
## area                               70.84      87.19   0.813 0.416842
## bedrooms                         -164459.72  153149.92  -1.074 0.283380
## bathrooms                         948747.45  104747.19   9.057  < 2e-16 ***
## stories                           449217.88   63976.47   7.022 6.75e-12 ***
## guestroomyes                      280577.49  131658.74   2.131 0.033541 *
## basementyes                       364648.64  110169.37   3.310 0.000997 ***
## parking                           278167.13   58348.18   4.767 2.41e-06 ***
## hotwaterheatingyes                871171.65  222596.31   3.914 0.000103 ***
## airconditioningyes                882671.33  108360.07   8.146 2.72e-15 ***
## mainroadyes                       419007.71  141791.03   2.955 0.003265 **
## prefareayes                       655372.80  115341.48   5.682 2.20e-08 ***
## furnishingstatussemi-furnished   -46509.39  116215.67  -0.400 0.689171
## furnishingstatusunfurnished     -424844.69  125994.32  -3.372 0.000801 ***
## area:bedrooms                      56.40      27.26   2.069 0.039033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1065000 on 530 degrees of freedom
```

```
## Multiple R-squared:  0.6844, Adjusted R-squared:  0.676
## F-statistic: 82.08 on 14 and 530 DF,  p-value: < 2.2e-16
```

The regression model with interaction terms provides insights into the relationship between various factors and house prices. The R-squared value of 0.6844 indicates that approximately 68.4% of the variance in house prices is explained by the model, which is a strong fit. Key predictors, including bathrooms, stories, guestroom, basement, parking, hotwaterheating, airconditioning, mainroad, and prefarea, are highly significant, with p-values less than 0.05. For instance, bathrooms and stories have large positive coefficients, suggesting that additional bathrooms or stories significantly increase the price of a house. Guestroom, basement, parking, hotwaterheating, airconditioning, mainroad, and prefarea all show significant positive effects on price, indicating that homes with these features tend to have higher prices. The variable furnishingstatusunfurnished has a significant negative coefficient, suggesting that unfurnished homes have lower prices compared to fully furnished homes. The interaction term area with bedrooms is also significant, indicating that the effect of area on price depends on the number of bedrooms. The model's residual standard error of 1,065,000 indicates some level of unexplained variation in the data, but the overall model fit is strong, as indicated by the F-statistic of 82.08 and a p-value of less than 0.05, confirming the model's statistical significance.

```
##                       GVIF Df GVIF^(1/(2*Df))
## area             17.182258  1        4.145149
## bedrooms          6.132052  1        2.476298
## bathrooms         1.329501  1        1.153040
## stories           1.478279  1        1.215845
## guestroom         1.219377  1        1.104254
## basement          1.328457  1        1.152587
## parking           1.212931  1        1.101331
## hotwaterheating   1.042721  1        1.021137
## airconditioning   1.219452  1        1.104288
## mainroad          1.172798  1        1.082958
## prefarea          1.149492  1        1.072144
## furnishingstatus  1.114288  2        1.027423
## area:bedrooms    24.681814  1        4.968080
```

In this case, the VIF values for area and the interaction term area with bedrooms are notably high (17.18 and 24.68, respectively), indicating that these variables are strongly correlated with other predictors in the model. This suggests potential multicollinearity issues that could distort the interpretation of the model. A high VIF means that the variable's effect may be difficult to distinguish from the effects of other variables due to shared variance.

On the other hand, other predictors like bedrooms, bathrooms, stories, and guestroom show moderate VIF values (ranging from 1.22 to 6.13), indicating lower but still some degree of multicollinearity. The furnishingstatus variable, with two categories, has a VIF of 1.11 to 1.12, suggesting minimal multicollinearity.

Although model_interaction has a higher adjusted R-squared value (0.676) compared to model1 (0.674), the presence of multicollinearity makes model1 more suitable for interpretability.

Model_interaction2 – area*stories

```
##
## Call:
## lm(formula = price ~ area + bedrooms + area * stories + bathrooms +
##     stories + guestroom + basement + parking + hotwaterheating +
##     airconditioning + mainroad + prefarea + furnishingstatus,
##     data = df)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2993286  -651993   -63610   500780  5068299
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     629357.27  382170.62   1.647 0.100193
## area                               138.58      55.39   2.502 0.012656 *
## bedrooms                        135832.83   73039.87   1.860 0.063481 .
## stories                          78932.43  186825.12   0.422 0.672836
## bathrooms                       943931.27  105071.80   8.984  < 2e-16 ***
## guestroomyes                    290325.33  131367.87   2.210 0.027531 *
## basementyes                     369184.09  110291.66   3.347 0.000874 ***
## parking                         265955.85   58571.50   4.541 6.95e-06 ***
## hotwaterheatingyes              893018.24  223128.89   4.002 7.17e-05 ***
## airconditioningyes              860830.90  108017.84   7.969 9.84e-15 ***
## mainroadyes                     447231.03  142287.62   3.143 0.001765 **
## prefareayes                     654911.72  115315.10   5.679 2.23e-08 ***
## furnishingstatussemi-furnished  -29761.47  116456.29  -0.256 0.798390
## furnishingstatusunfurnished    -425457.06  125976.92  -3.377 0.000786 ***
## area:stories                        63.40      29.92   2.119 0.034577 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1064000 on 530 degrees of freedom
## Multiple R-squared:  0.6845, Adjusted R-squared:  0.6761
## F-statistic: 82.12 on 14 and 530 DF,  p-value: < 2.2e-16
```

The results from the second interaction model, which includes the interaction term area*stories, show that area ($p = 0.0127$) and the interaction term area with stories ($p = 0.0346$) are statistically significant predictors of price. Specifically, the positive coefficient for area suggests that as the area increases, so does the price, while the interaction term indicates that the relationship between area and price depends on the number of stories in the property. Other variables such as bathrooms, guestroom, basement, parking, hotwaterheating, airconditioning, mainroad, and prefarea are also significant, suggesting their strong influence on the price. The coefficients for furnishingstatus where semi-furnished have $p = 0.7984$ and unfurnished have $p = 0.000786$ indicate that the unfurnished status negatively affects price, but the semi-furnished status does not have a statistically significant impact. The model explains approximately 68.45% of the variance in the dependent variable (R-squared = 0.6845), and the F-statistic (82.12) with a p-value < 0.05 suggests that the model is statistically significant. However, the variable stories itself is not significant ($p = 0.6728$), indicating that the number of stories alone does not have a strong impact on price without considering its interaction with area.

```
##                      GVIF Df GVIF^(1/(2*Df))
## area             6.937849  1        2.633980
## bedrooms         1.395281  1        1.181220
## stories         12.611181  1        3.551222
## bathrooms        1.338276  1        1.156839
## guestroom        1.214469  1        1.102030
## basement         1.331927  1        1.154091
## parking          1.222710  1        1.105762
## hotwaterheating  1.048126  1        1.023780
## airconditioning  1.212234  1        1.101015
## mainroad         1.181488  1        1.086963
## prefarea         1.149415  1        1.072108
```

```
## furnishingstatus  1.129823  2        1.030986
## area:stories     19.388656  1        4.403255
```

The Variance Inflation Factor (VIF) results for the interaction model indicate potential multicollinearity issues, especially with the interaction term area with stories (GVIF = 19.39), suggesting that this term may be highly correlated with other predictors in the model. Area and stories also have relatively high VIF values (GVIF = 6.94) for area, GVIF = 12.61) for stories, indicating some degree of multicollinearity between these two variables. Multicollinearity can lead to unreliable coefficient estimates and inflated standard errors, making it difficult to assess the individual impact of predictors. Other variables, such as bedrooms, bathrooms, guestroom, basement, parking, hotwaterheating, airconditioning, mainroad, prefarea, and furnishingstatus, exhibit lower VIF values, which suggests they do not have significant multicollinearity concerns.

Model_interaction2 also exhibits a higher adjusted R-squared value (0.676) compared to model1 (0.674). However, due to the presence of multicollinearity, model1 is preferred for better interpretability.

Model_interaction3 – area*stories + bedrooms*airconditioning

```
##
## Call:
## lm(formula = price ~ area + bedrooms + area * stories + bathrooms +
##     stories + guestroom + basement + parking + airconditioning +
##     hotwaterheating + bedrooms * airconditioning + mainroad +
##     prefarea + furnishingstatus, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3116155  -662539   -46049   475482  4939298
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     699663.38  384097.09   1.822 0.069083 .
## area                               154.91      56.23   2.755 0.006076 **
## bedrooms                         75816.34   81927.21   0.925 0.355174
## stories                         123747.05  188616.81   0.656 0.512060
## bathrooms                       949245.73  104967.04   9.043  < 2e-16 ***
## guestroomyes                    276402.66  131457.29   2.103 0.035972 *
## basementyes                     371329.95  110135.14   3.372 0.000802 ***
## parking                         264942.14   58487.49   4.530 7.30e-06 ***
## airconditioningyes              126952.49  468978.99   0.271 0.786728
## hotwaterheatingyes              904701.66  222914.35   4.059 5.68e-05 ***
## mainroadyes                     442772.52  142102.32   3.116 0.001934 **
## prefareayes                     647083.39  115245.88   5.615 3.18e-08 ***
## furnishingstatussemi-furnished  -14103.93  116689.50  -0.121 0.903842
## furnishingstatusunfurnished    -403787.90  126508.74  -3.192 0.001498 **
## area:stories                        55.27      30.30   1.824 0.068697 .
## bedrooms:airconditioningyes     240379.08  149494.69   1.608 0.108444
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1063000 on 529 degrees of freedom
## Multiple R-squared:  0.686,  Adjusted R-squared:  0.6771
## F-statistic: 77.05 on 15 and 529 DF,  p-value: < 2.2e-16
```

The summary of the model with interaction terms (model_interaction3) reveals that several variables significantly influence the house price. The area ($p = 0.006$) and bathrooms ($p < 0.05$) variables have strong positive relationships with price, indicating that larger areas and more bathrooms lead to higher prices. Guestroom, basement, parking, hotwaterheating, mainroad, prefarea, and furnishingstatusunfurnished are also significant predictors, with all but airconditioning ($p = 0.79$) having strong impacts. The interaction terms, such as area with stories ($p = 0.0687$) and bedrooms with airconditioning ($p = 0.1084$), are not as strongly significant, though area approaches statistical significance, suggesting there may be a slight interactive effect between these variables and house price. The model's R-squared value is 0.686, indicating that the model explains 68.6% of the variance in house prices. However, the relatively high residual standard error (around 1.06 million) and the non-significant interaction terms suggest that further model refinement, possibly addressing multicollinearity, could improve prediction accuracy.

```
##                                GVIF Df GVIF^(1/(2*Df))
## area                       7.171846  1        2.678030
## bedrooms                   1.760740  1        1.326929
## stories                   12.892682  1        3.590638
## bathrooms                  1.339604  1        1.157413
## guestroom                  1.219761  1        1.104428
## basement                   1.332122  1        1.154176
## parking                    1.222852  1        1.105826
## airconditioning           22.919213  1        4.787401
## hotwaterheating            1.049241  1        1.024324
## mainroad                   1.181938  1        1.087170
## prefarea                   1.151470  1        1.073066
## furnishingstatus           1.143507  2        1.034093
## area:stories              19.943028  1        4.465762
## bedrooms:airconditioning  24.352613  1        4.934837
```

The VIF results for model_interaction3 indicate some potential multicollinearity issues, particularly with the interaction terms and certain predictors. The airconditioning variable (VIF = 22.92) and the interaction term bedrooms with airconditioning (VIF = 24.35) have notably high VIFs, suggesting that these variables may be highly collinear with others in the model, which could inflate the standard errors and affect the reliability of the coefficient estimates. Similarly, the area with stories interaction term (VIF = 19.94) also shows high collinearity. While other variables, like bedrooms, bathrooms, and parking, have moderate VIFs, the high values of the interaction terms imply that these variables could be contributing to multicollinearity, which may reduce the precision of the model's estimates and affect the interpretation of the relationships between predictors and the target variable which is house price.

Model_interaction3 demonstrates a higher adjusted R-squared value (0.677) compared to model1 (0.674). Nonetheless, the presence of multicollinearity makes model1 a more suitable choice for interpretability.

Summary

The model that introduces interaction terms are not strongly significant and lead to multicollinearity issues. Although it provides a reasonable R-squared value, the high residual standard error and non-significant interaction terms suggest that the added complexity may not offer much improvement over simpler models. The previous model, which is model1, avoids multicollinearity problems and focuses on key predictors with strong significance, likely to offer better predictive accuracy and interpretability. Therefore, the previous model is more reliable and efficient due to their stronger statistical significance and lower multicollinearity.

References

Abdul-Rahman, S., Mutalib, S., Alam, S., Nor, M., Zulkifley, H., & Ibrahim, M. I. (2021). Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur. In IJACSA) International Journal of Advanced Computer Science and Applications (Vol. 12, Issue 12). www.ijacsa.thesai.org

Aliefendioğlu, Y., Tanrivermis, H., & Salami, M. A. (2022). House price index (HPI) and Covid-19 pandemic shocks: evidence from Turkey and Kazakhstan. International Journal of Housing Markets and Analysis, 15(1), 108–125. https://doi.org/10.1108/IJHMA-10-2020-0126

Ariffin, N. F., Jaafar, M. F. M., Ali, M. a. H., Ramli, N. I., Muthusamy, K., & Lim, N. H. a. S. (2018). Investigation on factors that contribute to the abandonment of building in construction industry in Malaysia. E3S Web of Conferences, 34, 01025. https://doi.org/10.1051/e3sconf/20183401025

Cai, Z., & Zhao, Y. (2023). House Rent Analysis with Linear Regression Model-A Case Study of Six Cities in India. In Highlights in Science, Engineering and Technology TPCEE (Vol. 2022). https://www.magicbricks.com/

Cellmer, R., Cichulska, A., & Bełej, M. (2020). Spatial analysis of housing prices and market activity with the geographically weighted regression. ISPRS International Journal of Geo-Information, 9(6). https://doi.org/10.3390/ijgi9060380

Fang, L. (2023). Machine learning models for house price prediction. Applied and Computational Engineering, 4(1), 409–415. https://doi.org/10.54254/2755-2721/4/20230505

Greenaway-McGrevy, R., & Phillips, P. C. B. (2021). House prices and affordability. In New Zealand Economic Papers (Vol. 55, Issue 1, pp. 1–6). Routledge. https://doi.org/10.1080/00779954.2021.1878328

Heyman, A. V., & Sommervoll, D. E. (2019). House prices and relative location. Cities, 95. https://doi.org/10.1016/j.cities.2019.06.004

Imran, Zaman, U., Waqar, M., & Zaman, A. (2021). Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data. In Soft Computing and Machine Intelligence Journal (Issue 1). www.kaggle.com

Khalid, S. S., Ahmad, M., Khalid, L. S., & Odimegwu, T. C. (2018). Problems and factors affecting property developers performance in the Dubai construction industry. International Journal of Scientific and Research Publications, 8(11). https://doi.org/10.29322/ijsrp.8.11.2018.p8312

Levkovich, O., Rouwendal, J., & van Marwijk, R. (2016). The effects of highway development on housing prices. Transportation, 43(2), 379–405. https://doi.org/10.1007/s11116-015-9580-7

Li, L. H., Cheung, D., & Sun, H. (2015). Does size matter? The dynamics of housing sizes and prices in Hong Kong. Journal of Housing and the Built Environment, 30(1), 109–124. https://doi.org/10.1007/s10901-014-9398-1

Li, Z. (2021). Prediction of house price index based on machine learning methods. Proceedings - 2021 2nd International Conference on Computing and Data Science, CDS 2021, 472–476. https://doi.org/10.1109/CDS52072.2021.00087

Lyu, Z. (2024). Research on the Relationship between Influencing Factors and House Prices Changes. Highlights in Science Engineering and Technology, 93, 1–7. https://doi.org/10.54097/2nz8xe97

Mac-Barango, D. (2017). Construction project abandonment: An appraisal of causes, effects and remedies. World Journal of Innovation and Modern Technology, 1(1), 1-10.

Markowitz, D. M. (2023). Words for Sale: Linguistic Complexity Associates with Higher Housing Prices in Online Realty Advertisements.

Muñoz, S. F., & Cueto, L. C. (2017). What has happened in Spain? The real estate bubble, corruption and housing development: A view from the local level. Geoforum, 85, 206–213. https://doi.org/10.1016/j.geoforum.2017.08.002

Nwankwo, M. P., Onyeizu, N. M., Asogwa, E. C., Ejike, C. O., & Obulezi, O. J. (2023). Prediction of House Prices in Lagos-Nigeria Using Machine Learning Models. European Journal of Theoretical and Applied Sciences, 1(5), 313–326. https://doi.org/10.59324/ejtas.2023.1(5).22

Rahadi, R. A., Wiryono, S. K., Koesrindartoto, D. P., & Syamwil, I. B. (2015). Factors influencing the price of housing in indonesia. International Journal of Housing Markets and Analysis, 8(2), 169–188. https://doi.org/10.1108/IJHMA-04-2014-0008

Rico-Juan, J. R., & Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. Expert Systems with Applications, 171. https://doi.org/10.1016/j.eswa.2021.114590

Shen, X., Liu, P., Qiu, Y. (Lucy), Patwardhan, A., & Vaishnav, P. (2021). Estimation of change in house sales prices in the United States after heat pump adoption. Nature Energy, 6(1), 30–37. https://doi.org/10.1038/s41560-020-00706-4

Sipan, I., Mar Iman, A. H., & Razali, M. N. (2018). Spatial–temporal neighbourhood-level house price index. International Journal of Housing Markets and Analysis, 11(2), 386–411. https://doi.org/10.1108/IJHMA-03-2017-0027

Theisen, T., & Emblem, A. W. (2021). The Road to Higher Prices: Will Improved Road Standards Lead to Higher Housing Prices? Journal of Real Estate Finance and Economics, 62(2), 258–282. https://doi.org/10.1007/s11146-020-09751-y

Wanjiku, S., Bosire, J., & Matanda, J. (2021). EFFECT OF MACRO-ECONOMIC FACTORS ON FINANCIAL PERFORMANCE IN KENYA OF REGISTERED REAL ESTATE INVESTMENTS TRUSTS. International Journal of Finance and Accounting (Nairobi), 6(1), 72–92. https://doi.org/10.47604/ijfa.1381

Xu, K., & Nguyen, H. (2022). Predicting housing prices and analyzing real estate markets in the Chicago suburbs using machine learning. https://github.com/

Yao, S., & Feng, J. (2023). Research on changes in housing price and exploration of its influencing factors. Theoretical and Natural Science, 26(1), 187–191. https://doi.org/10.54254/2753-8818/26/20241061