Business Statistics Using R - Group Assignment (Analysis)

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

Research Questions

1. How do structural factors such as house area, number of bedrooms, bathrooms, and stories influence the market price of a property?
2. What is the effect of house condition factors (e.g., hot water heating, air conditioning, and furnishing status) on housing prices?
3. How do environmental factors, including proximity to main roads and preferred areas, impact property value?
4. Can interaction terms improve the accuracy of predicting housing prices?

Modeling Objective

The primary objective is prediction, aiming to develop a regression model that accurately forecasts housing prices based on structural, condition, and environmental factors. In addition, an inference component will identify key predictors and quantify their effects on housing prices, providing actionable insights for real estate developers, buyers, and policymakers.

Dataset Overview

The dataset contains 545 observations and 13 variables describing various features of houses and their corresponding prices. This includes structural, condition-related, and environmental attributes.

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

Variable Type	Examples of Variables	Description
Structural Factors	Area, bedrooms, bathrooms, stories, guestroom, basement availability, parking availability	Describe the physical attributes of the house
Condition Factors	Hot water heating, air conditioning, furnishing status	Reflect the amenities and overall condition of the house

Variable Type	Examples of Variables	Description
Environmental Factors	Main road, preferarea	Capture location-specific attributes that influence desirability
Response Variable	Price	Represents the market price of the house in monetary terms.

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 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.

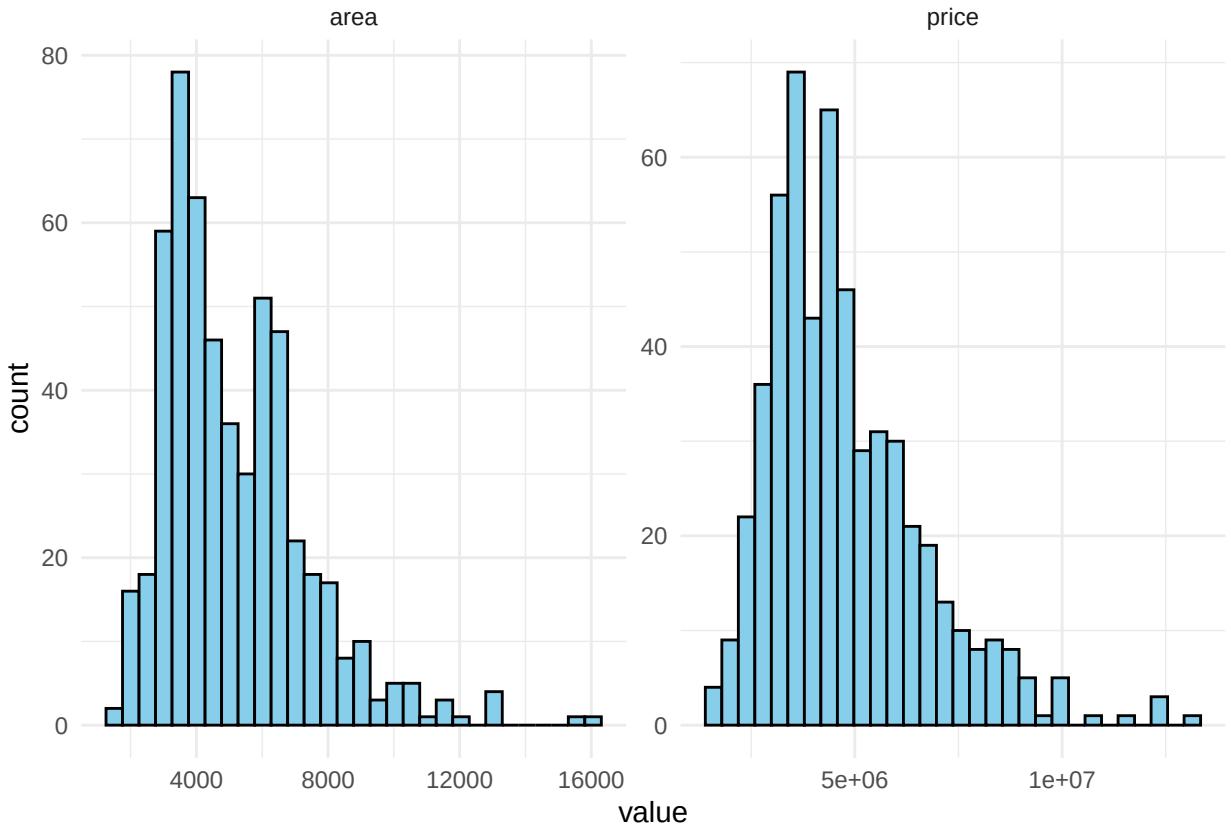
For bathrooms, the mean number is approximately 1.3 which indicates that most of the houses have at least one bathroom. Properties with up to 4 bathrooms are less common and are likely part of the higher end market. The distribution of the bathrooms reflects the housing design properties based on economic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

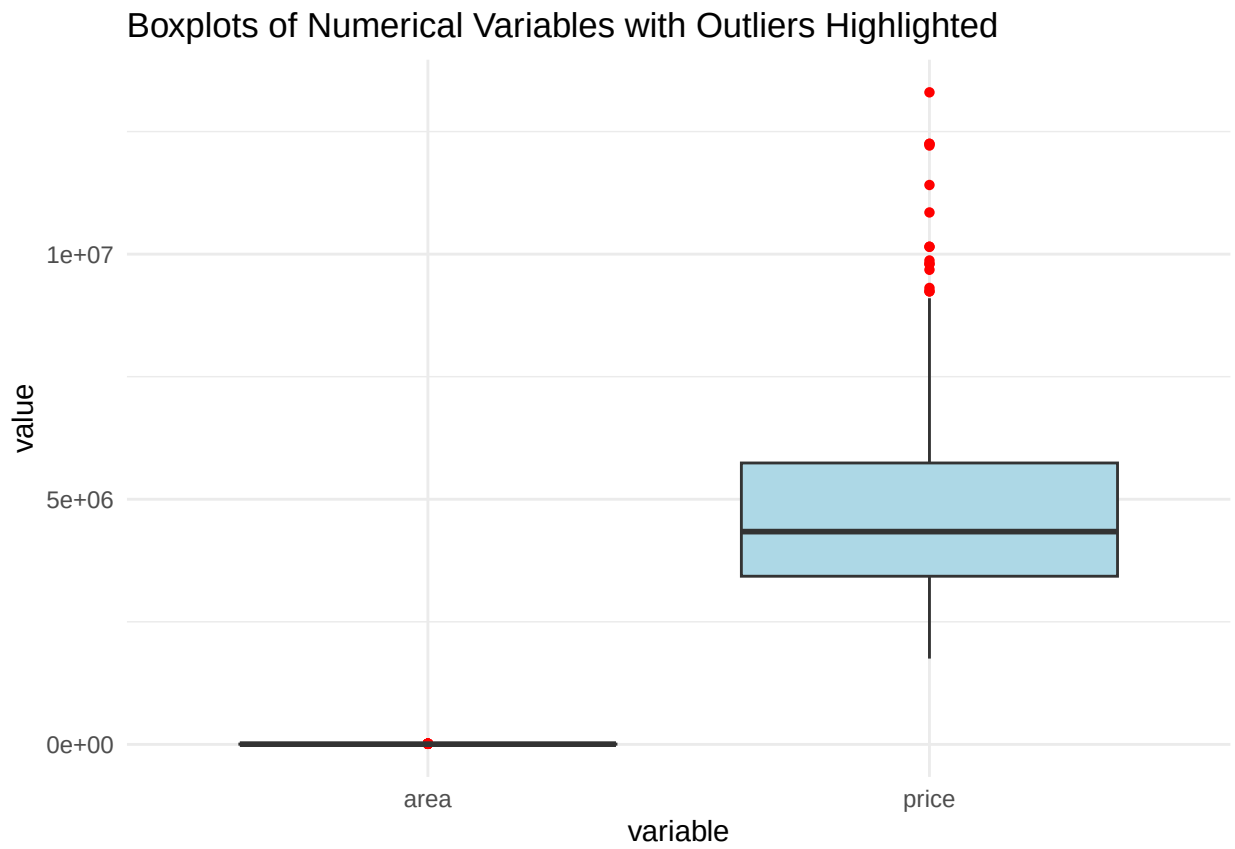
Overall, the data shown in the table is important for understanding the central tendencies and variations across the key variables.



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.



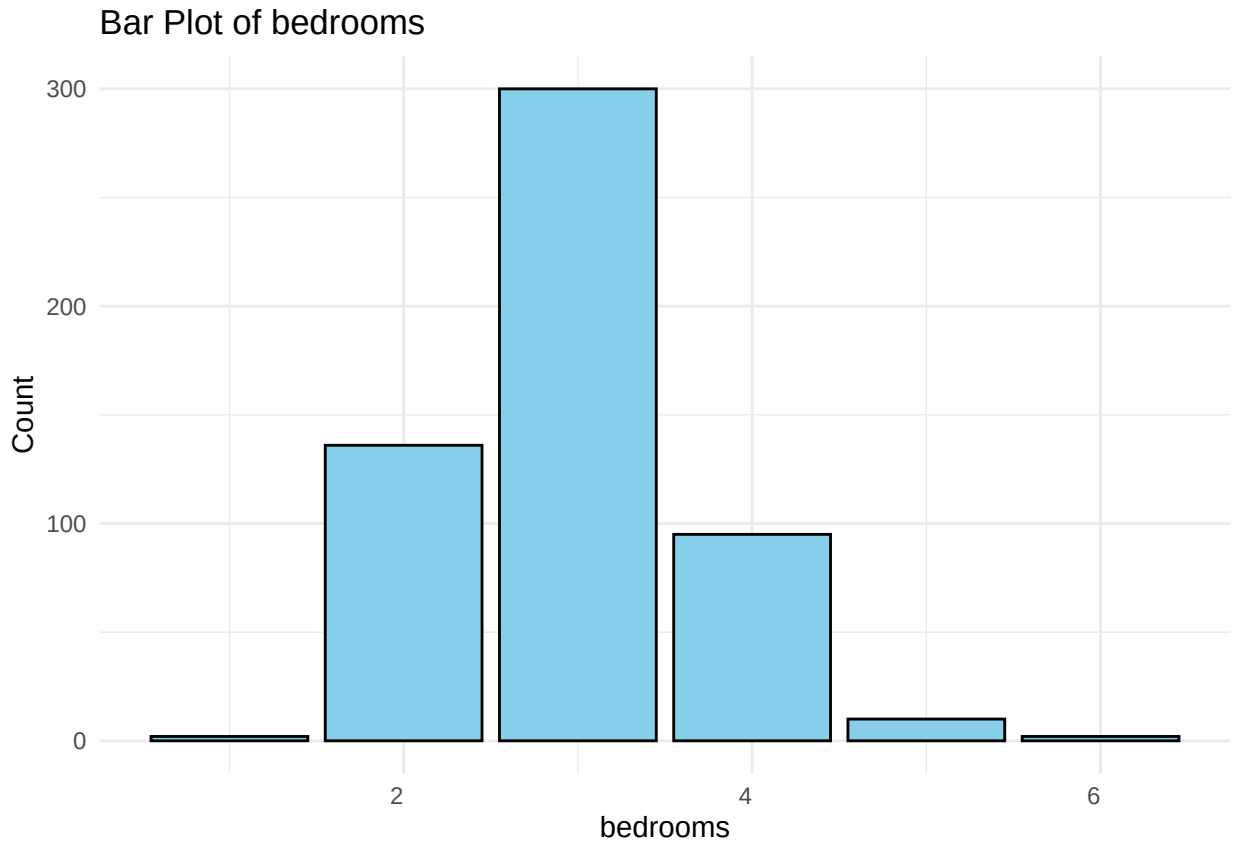
The boxplot visualizes the distribution of both price and area variables which highlight potential outliers. The outliers are highlighted with red dots to indicate clearly the presence of the outliers' figures in the boxplot.

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.

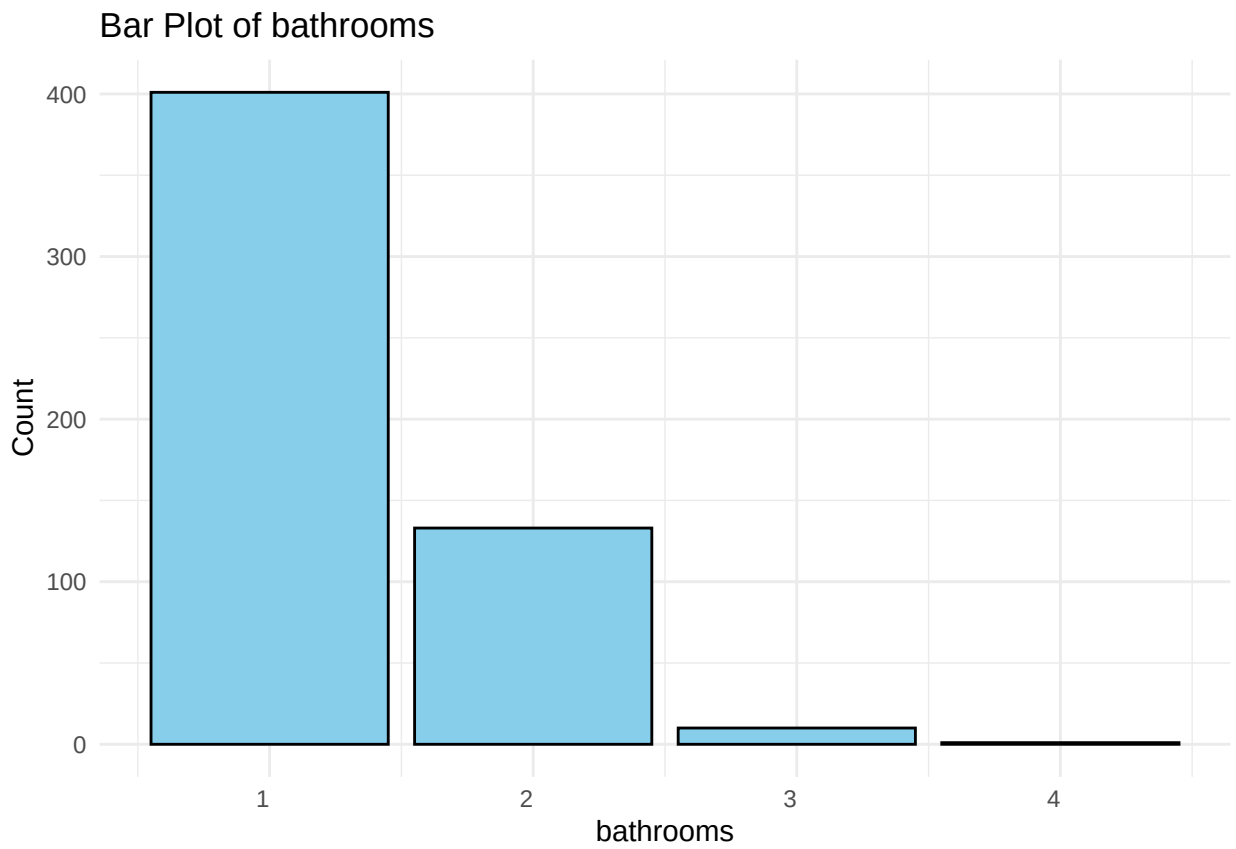
The key insight from this boxplot is that price has a much broader range of values compared to area, and it also includes outliers. This shows that variability in prices is possibly due to factors like locations, market trends or differences in property characteristics. As for area, it has a relatively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

Univariate Analysis



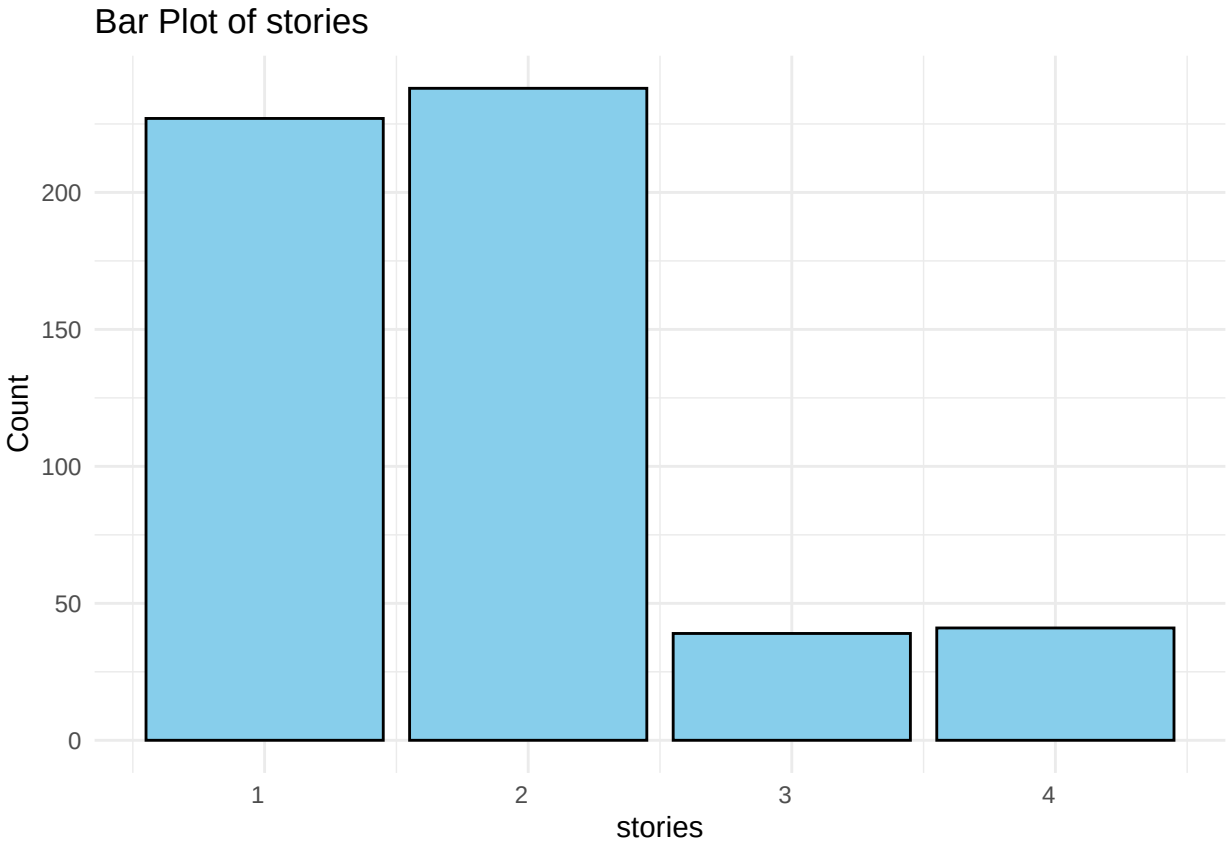
The bar plot above shows the number of houses by the number of bedrooms. The most common number of bedrooms is 3 which indicates the tallest bar showing the highest frequency (approximately 300 properties). The second and third most common of the number of bedrooms are 2 and 4 respectively. Properties 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.



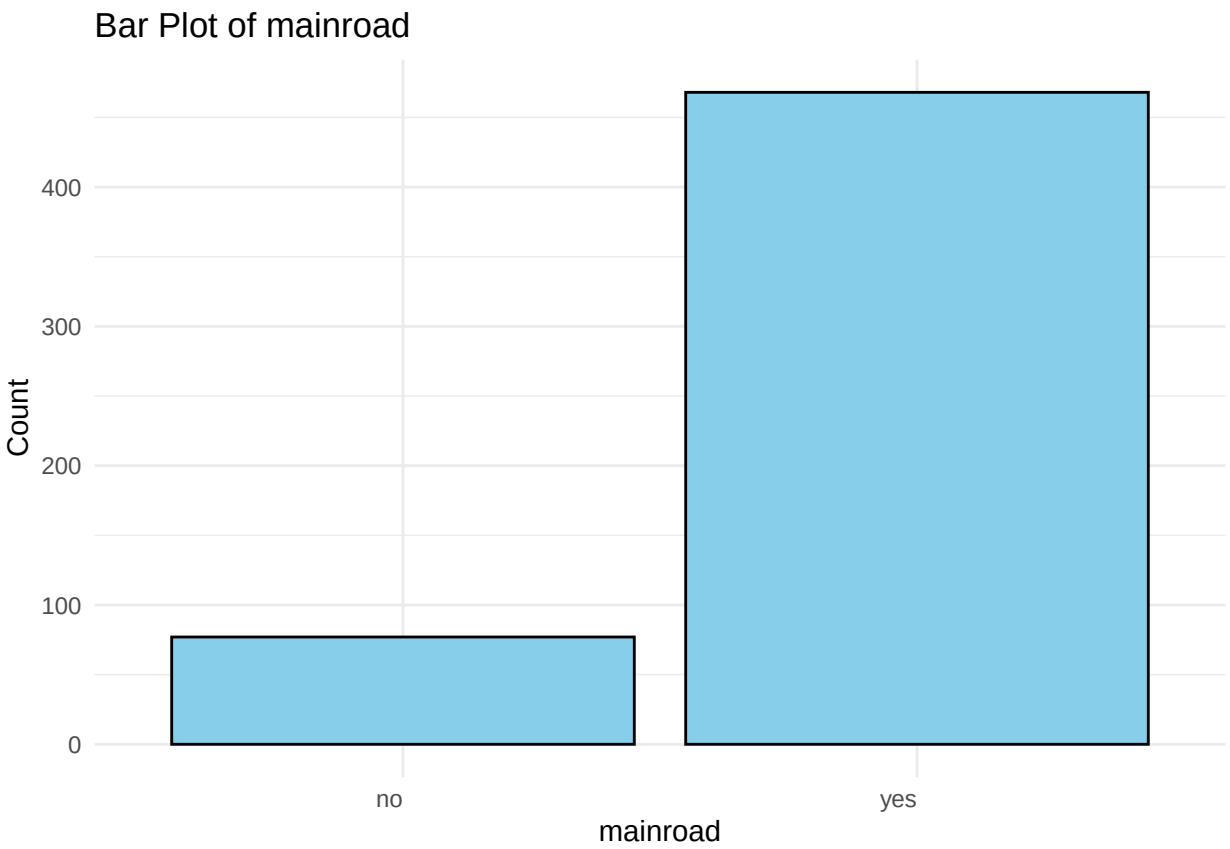
For the number of bathrooms, the properties with 1 bathroom are the most common with around 400 properties. The second most frequent is the house with 2 bathrooms, but their count is significantly lower than those with 1 bathroom with around 100 to 150 properties. Properties with 3 and 4 bathrooms are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



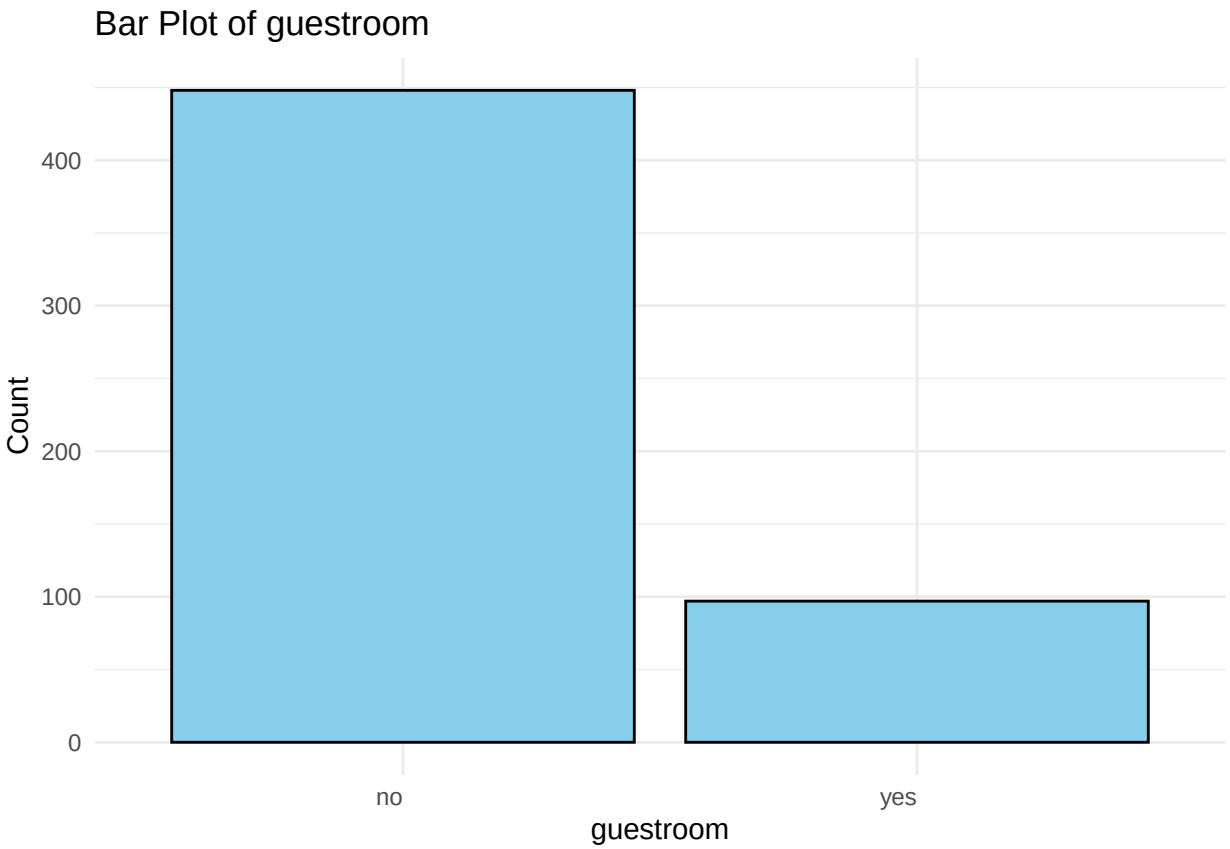
Properties with 1 and 2 stories are the most common, each with a frequency of about 200 to 250. Properties with 3 and 4 stories are uncommon with each having frequency of approximately 50 to 75 properties. Single and two stories house dominate the dataset which suggests that they are standard or most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



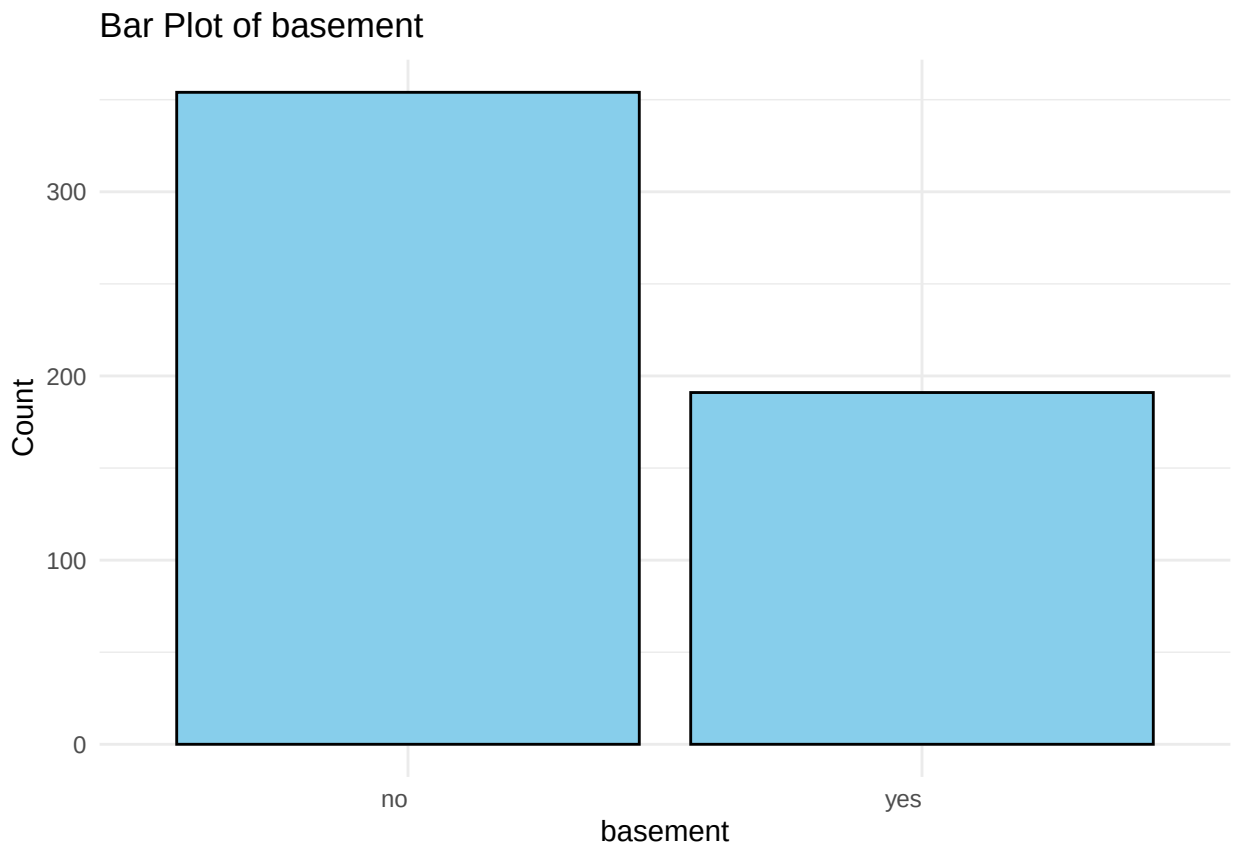
The bar plot above shows the distribution of properties based on whether the houses are located near a main road which are categorized by 'yes' and 'no'. Most of the houses are located on the main road with a frequency of over 400. A less proportion of properties are not on the main road with a frequency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



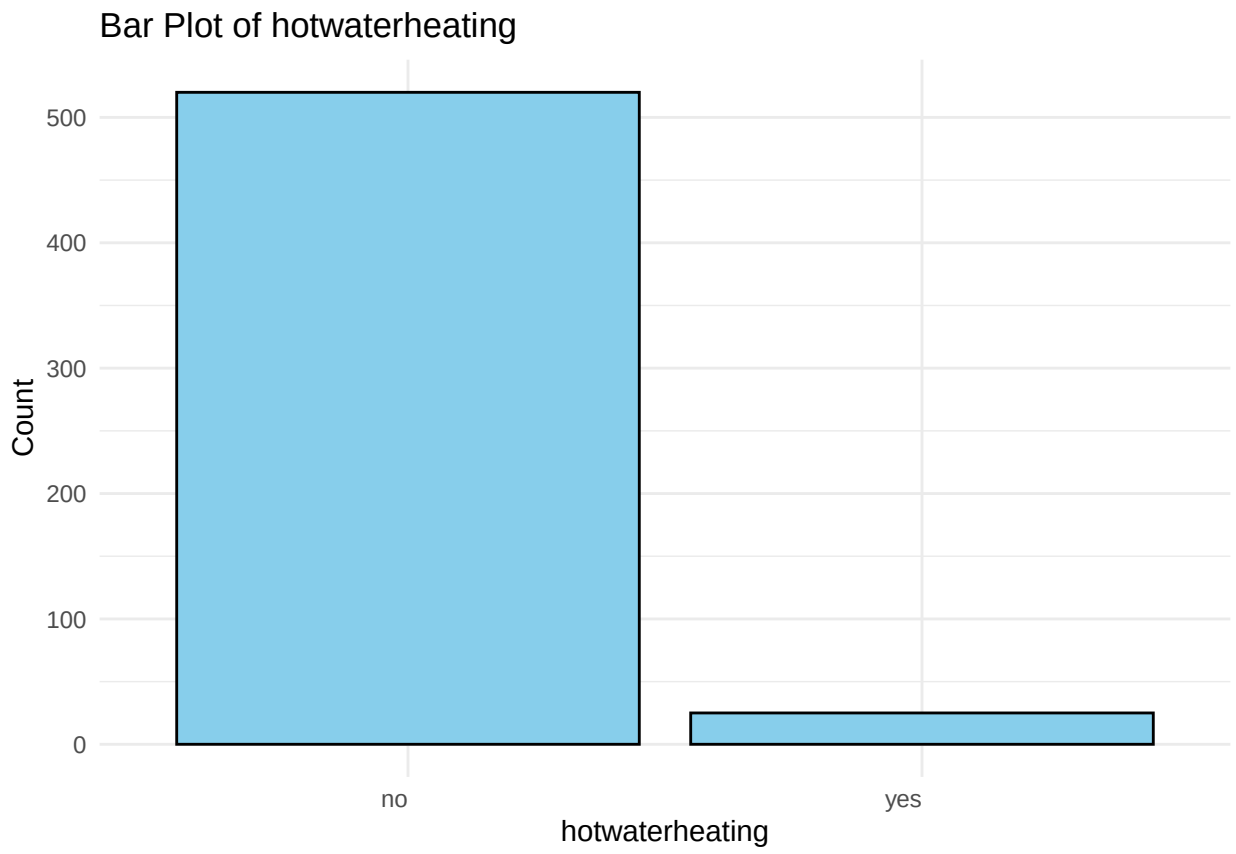
The above bar plot shows the distribution of properties whether the houses occupied with guestroom. From the graph above, the majority of houses do not have a guestroom with a frequency of 400 whereas the properties with guestrooms are only a small proportion with frequency of around 100.

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.



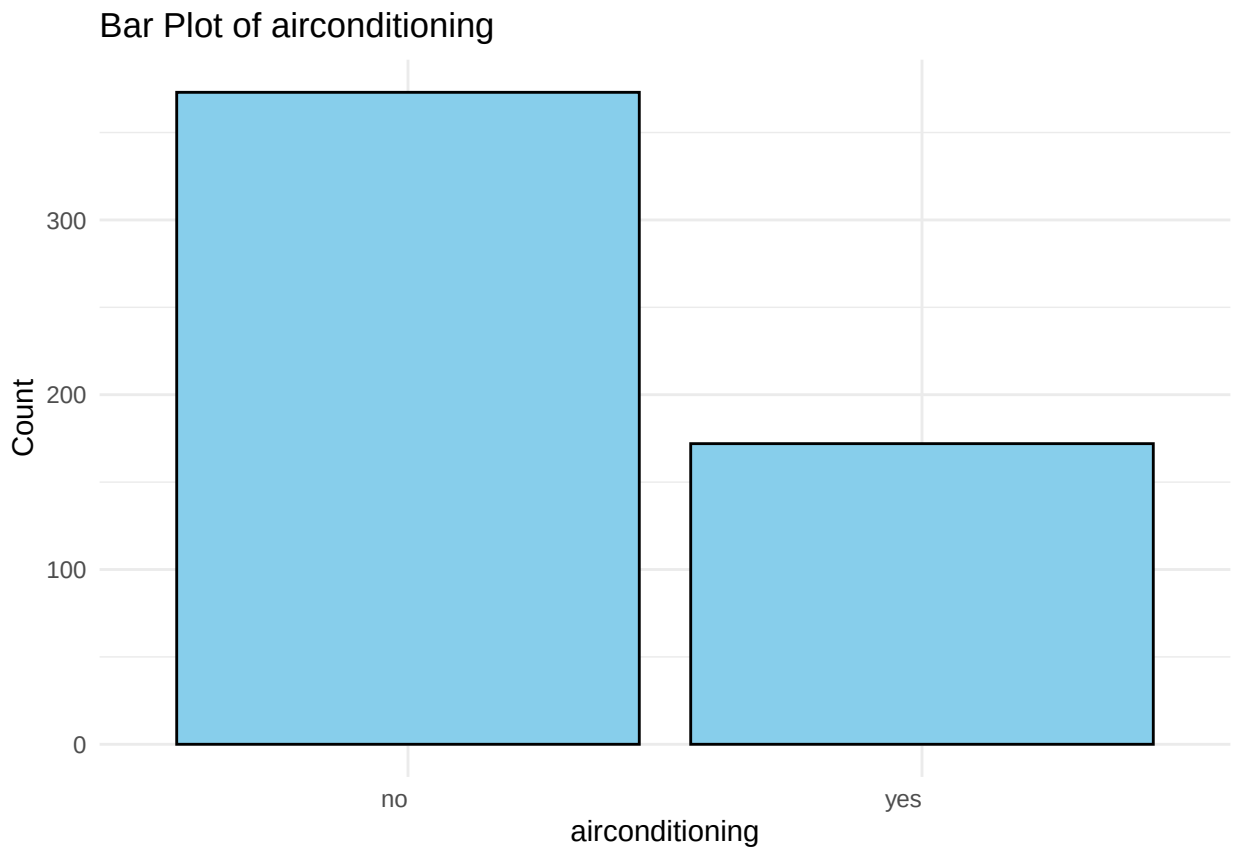
By looking at the graph, most properties do not have a guestroom with a frequency of over 400. A less proportion of houses do have a guestroom with a frequency of around 100.

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.



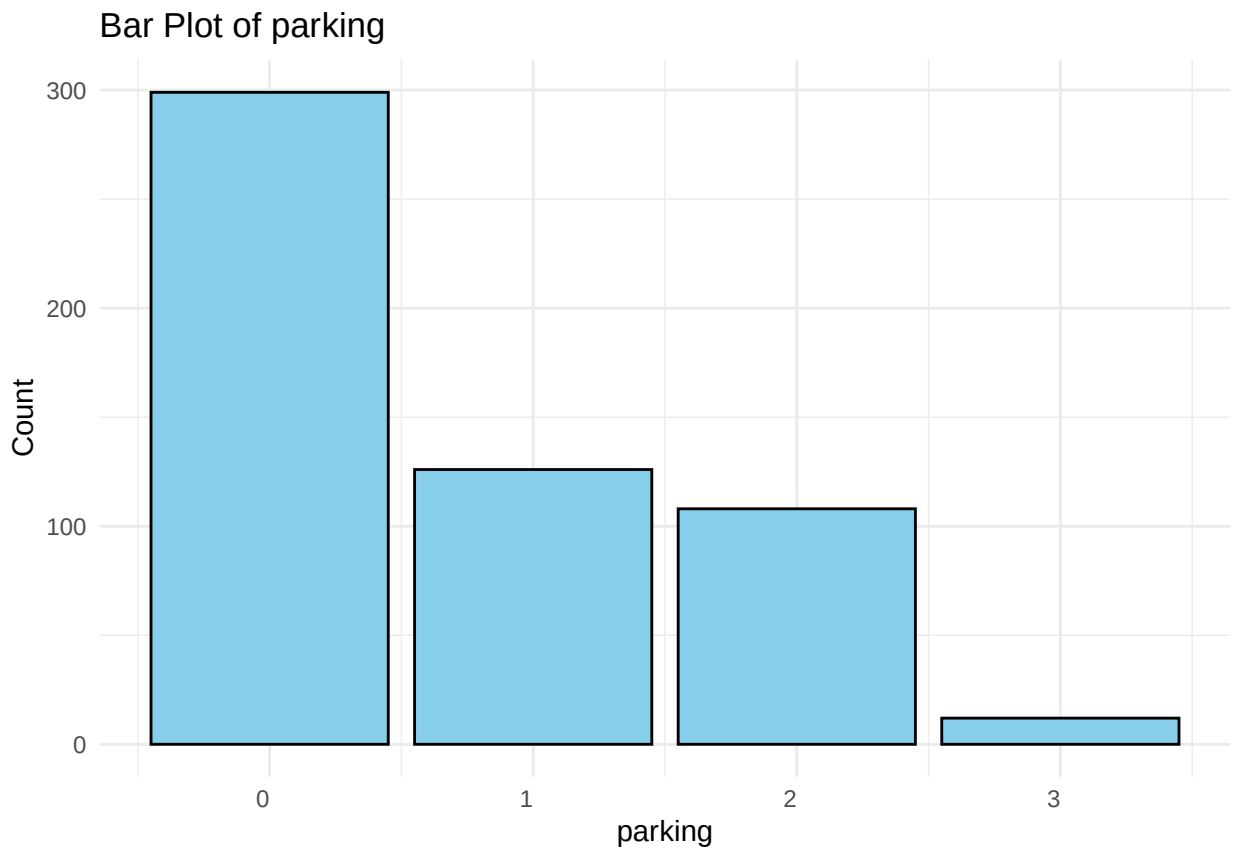
The above bar plot graph shows that the vast majority of properties which are over 500 houses do not have hot water heating and a very small number of properties which are less than 100 houses have water heating.

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.



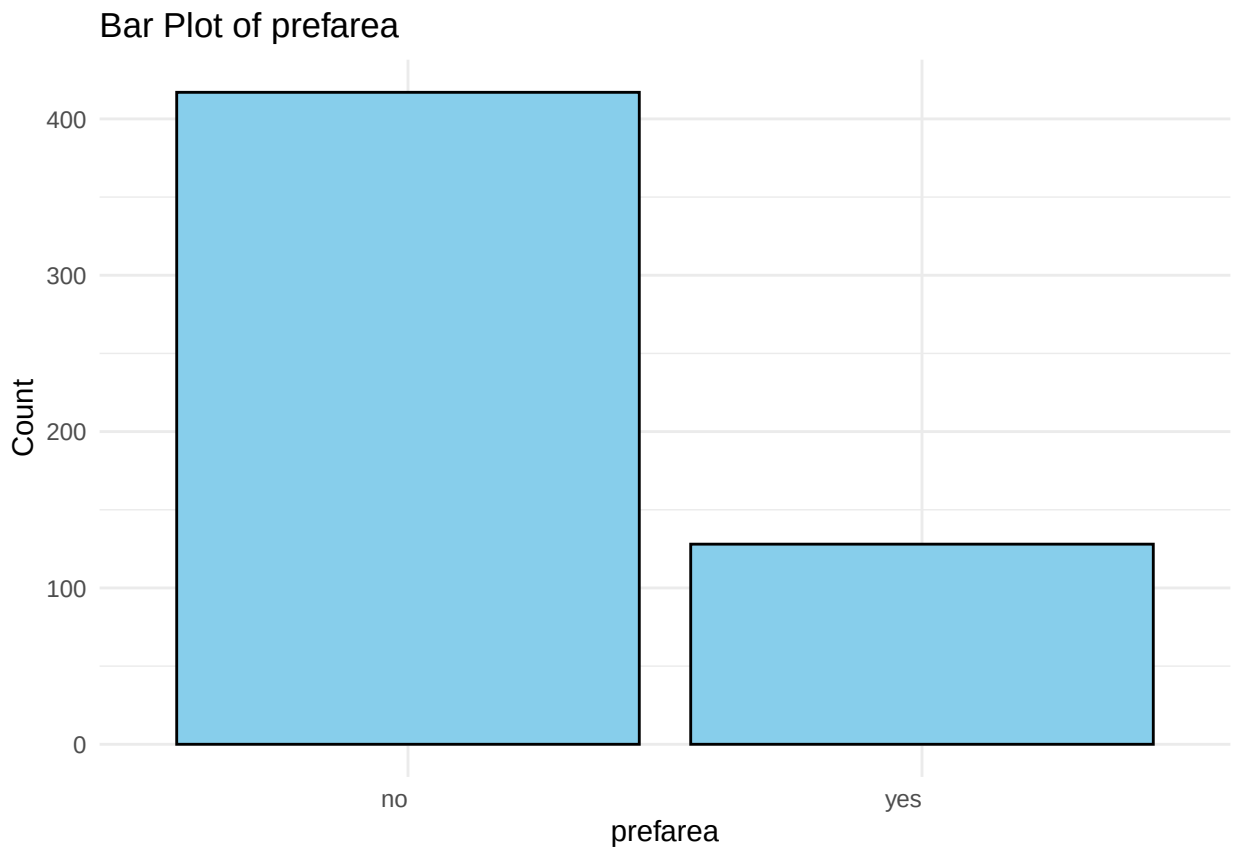
Most houses do not provide air conditioning with the count are more than 300 whereas a smaller proportion of houses provide air conditioning with a count between 100 to 200. The absence of air conditioning in most properties could be explained by the climate or buyers' preferences in the region of interest.

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.

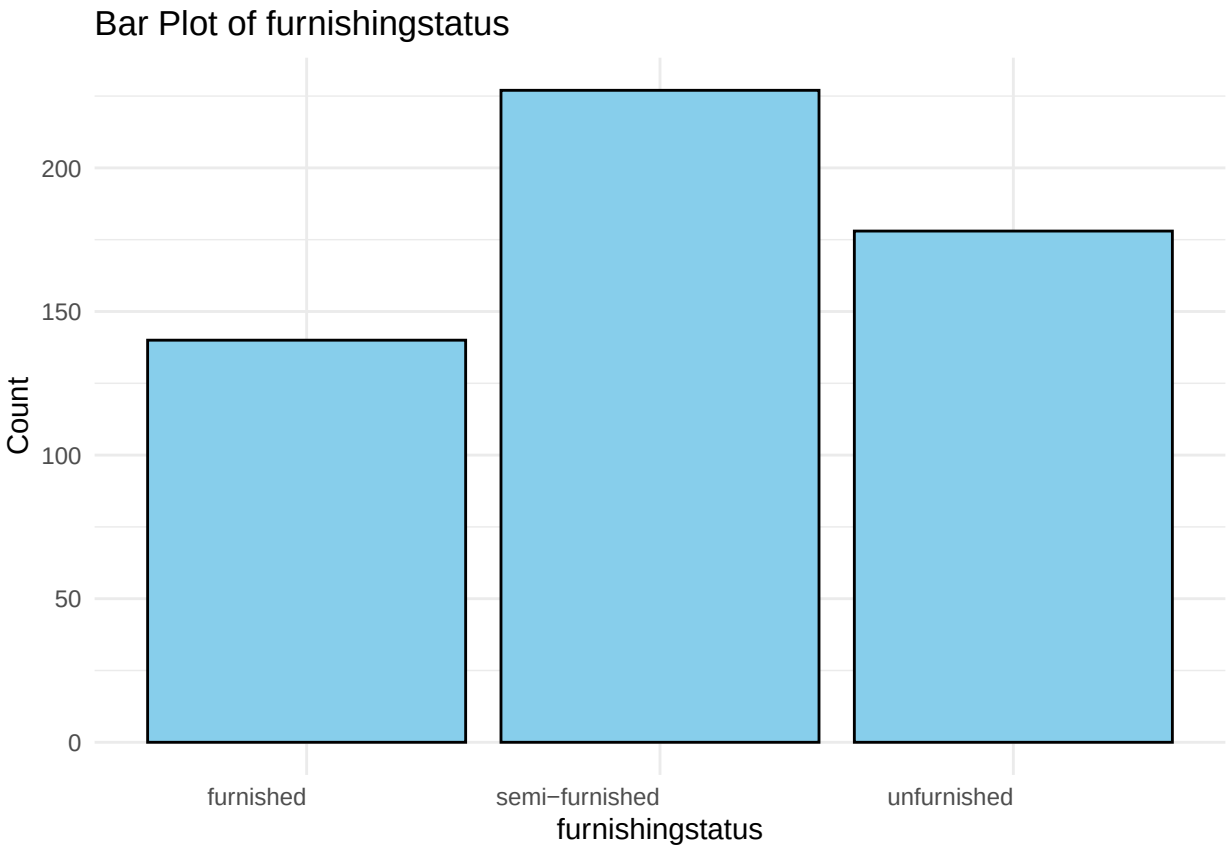
Properties with 1 and 2 parking spaces have almost an equal number of approximately 150 houses. These properties likely cater to households with smaller families or individuals with fewer vehicles. For houses with 3 parking spaces, it shows the least number with fewer properties accommodate for this 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.

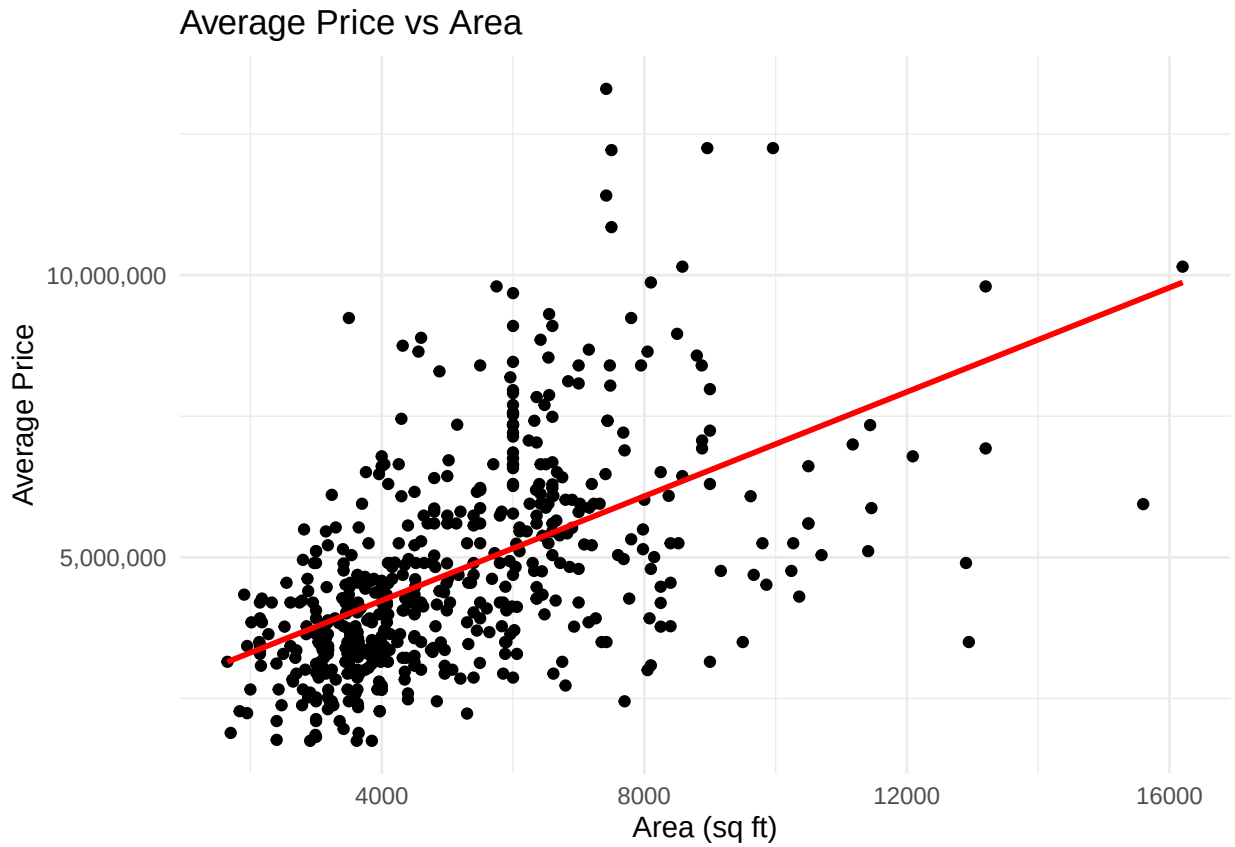


The bar plot above illustrates the furnishing status of properties categorized into semi-furnished, unfurnished, and furnished. It shows that semi-furnished properties are the most common with a count exceeding 200, followed by unfurnished properties with around 175 counts. This suggests that a wide 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.

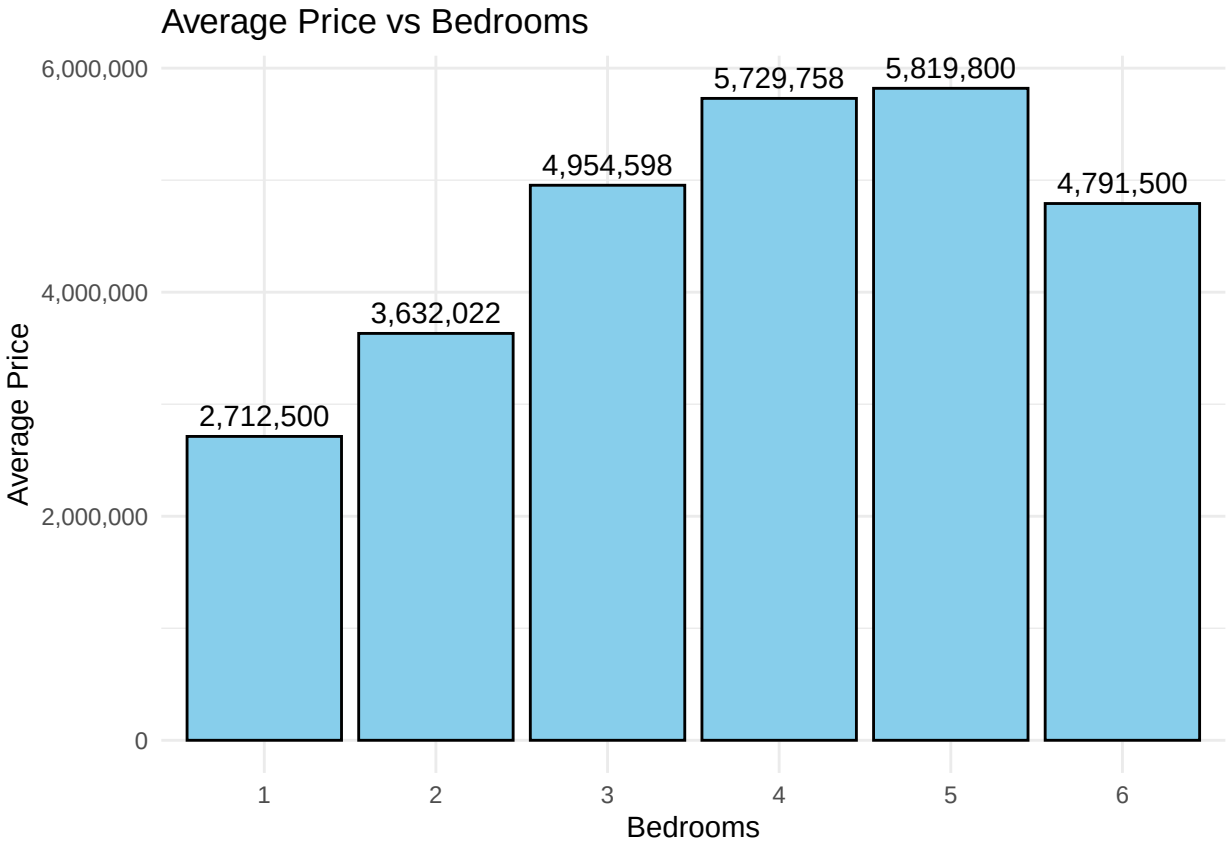
Bivariate Analysis



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.

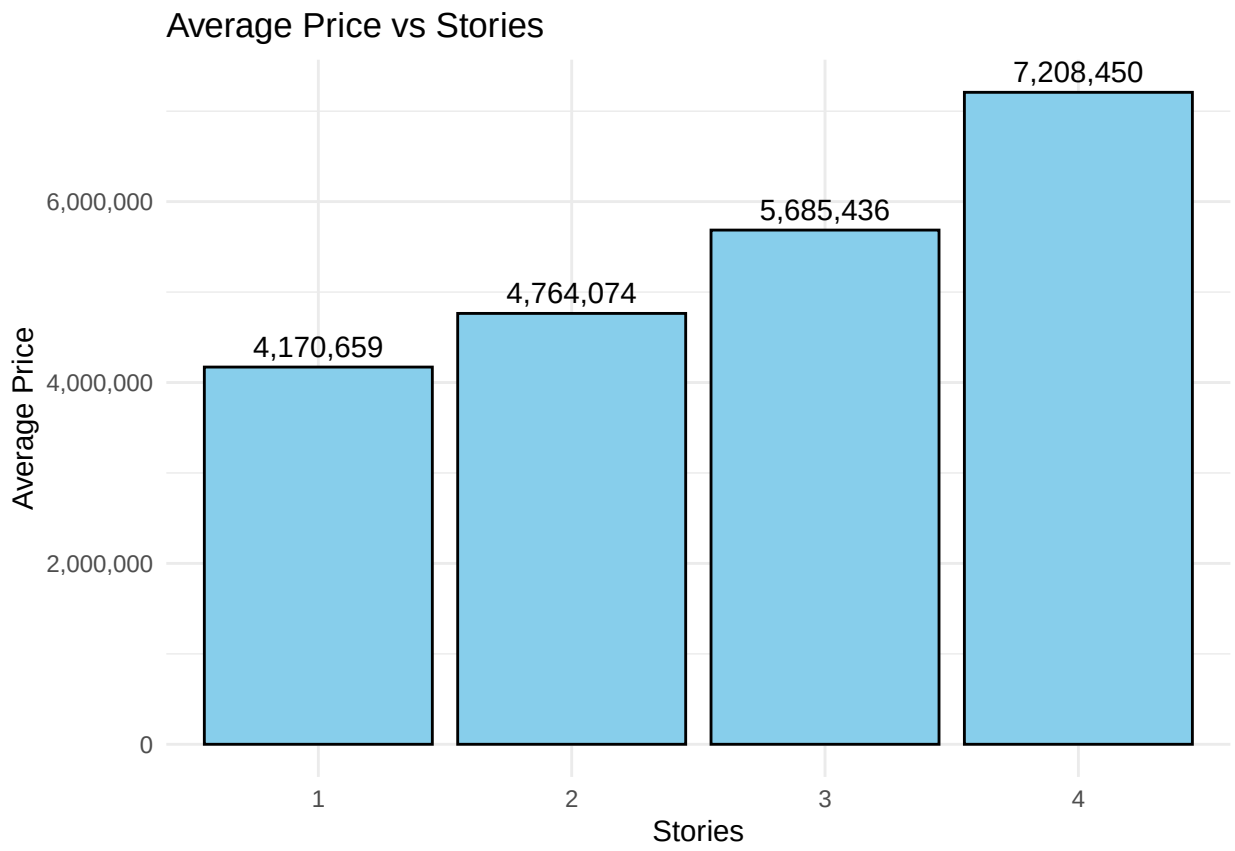
Notably, 6 bedrooms properties have a lower average price than properties with 3 to 5 bedrooms. This could possibly be influence by other factors such as location or amenities, which affecting housing price.

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.

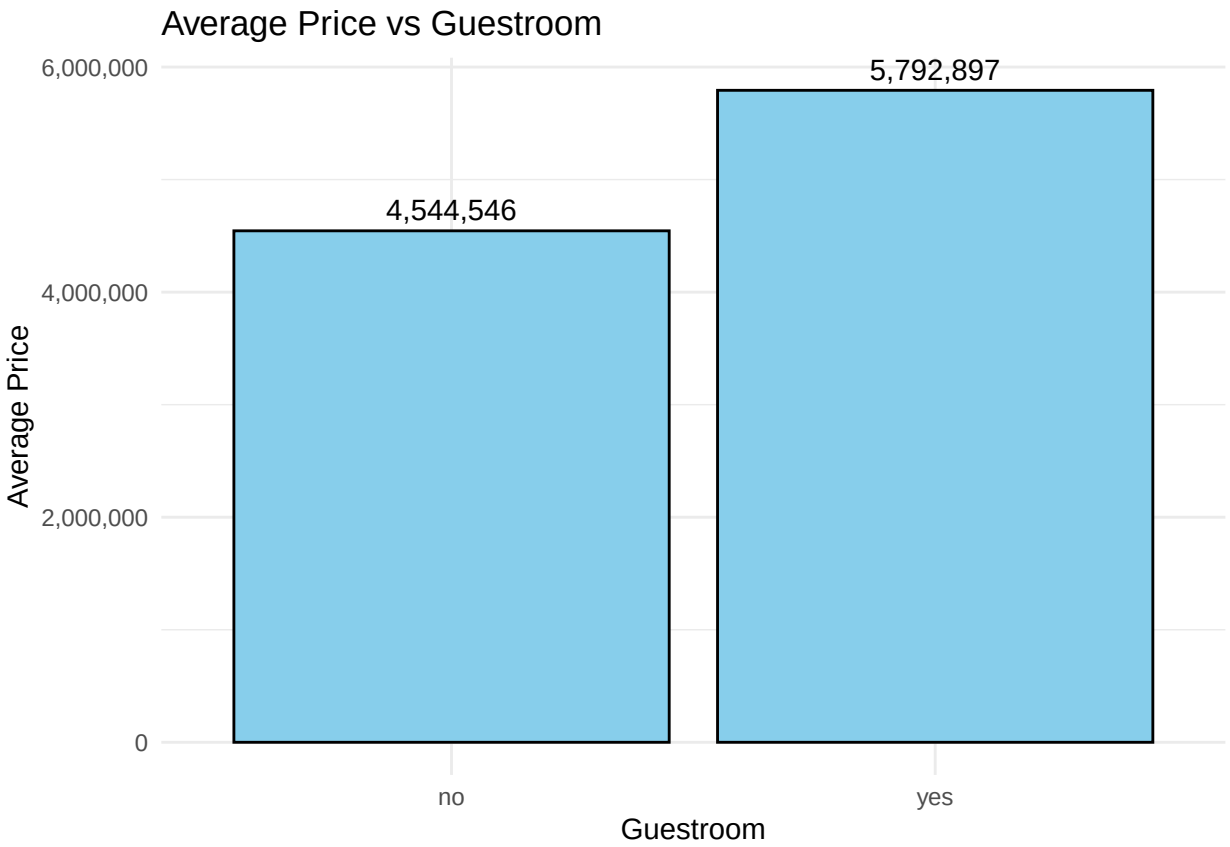
Overall, the bar plot shows a positive relationship between housing price and number of bathrooms, suggesting that number of bathrooms serve as a significant factor of housing price.



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.

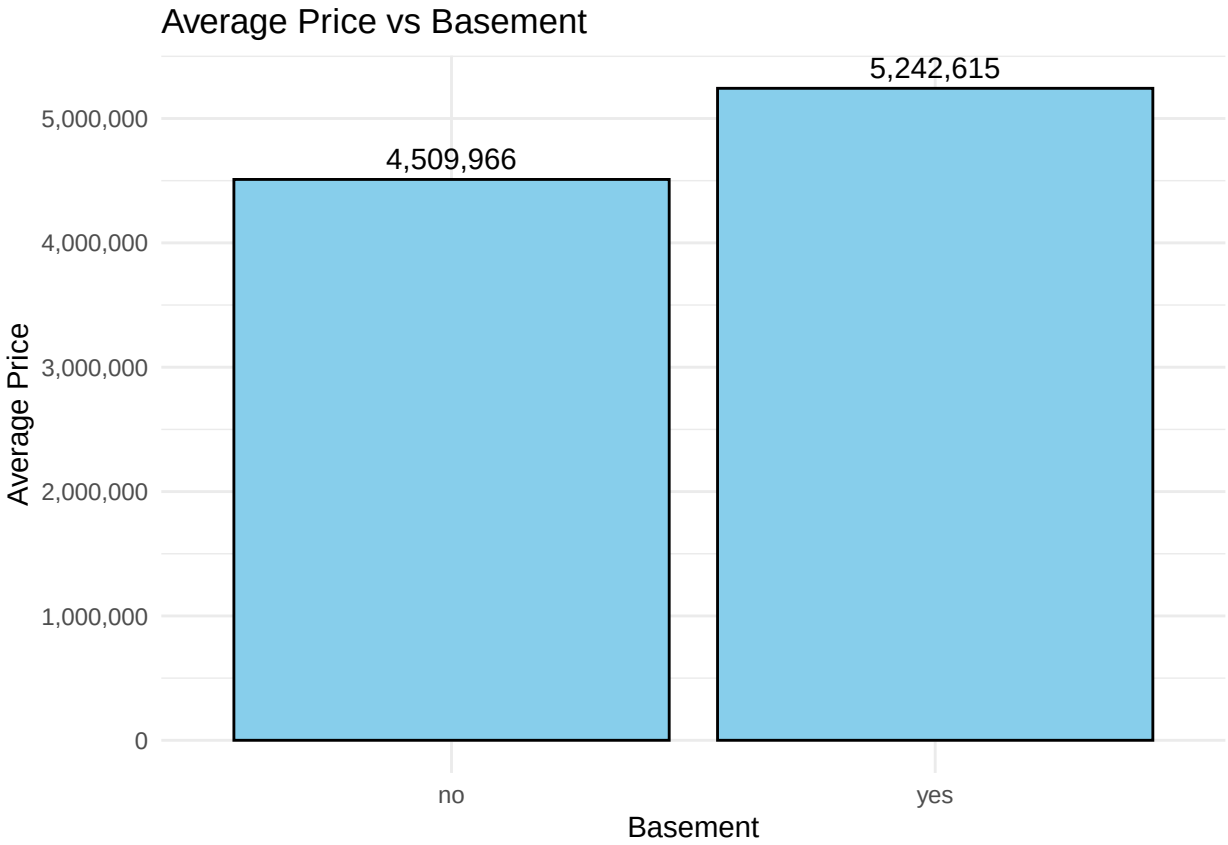
Properties with 4 stories have the highest average price at 7,208,450 followed by 3 stories properties with an average price of 5,685,436. The premium prices suggest a strong correlation between housing price and number of stories, which mainly targeting luxury properties market.

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.



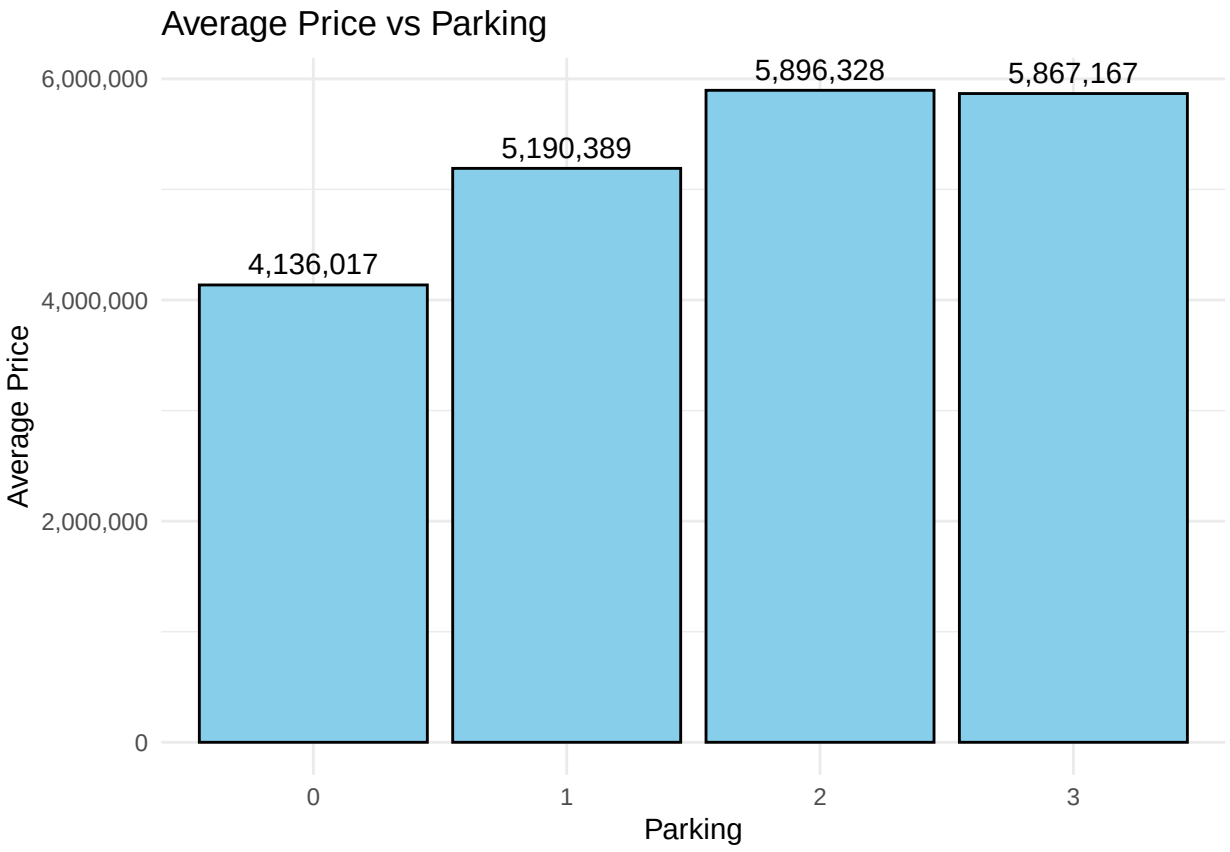
The bar plot above illustrates the relationship between the average price of properties and the presence of guestrooms. It shows that properties with guestroom have a higher average price of 5,792,697 which is slightly higher than the average price of 4,544,546 for properties without a guestroom. This 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.



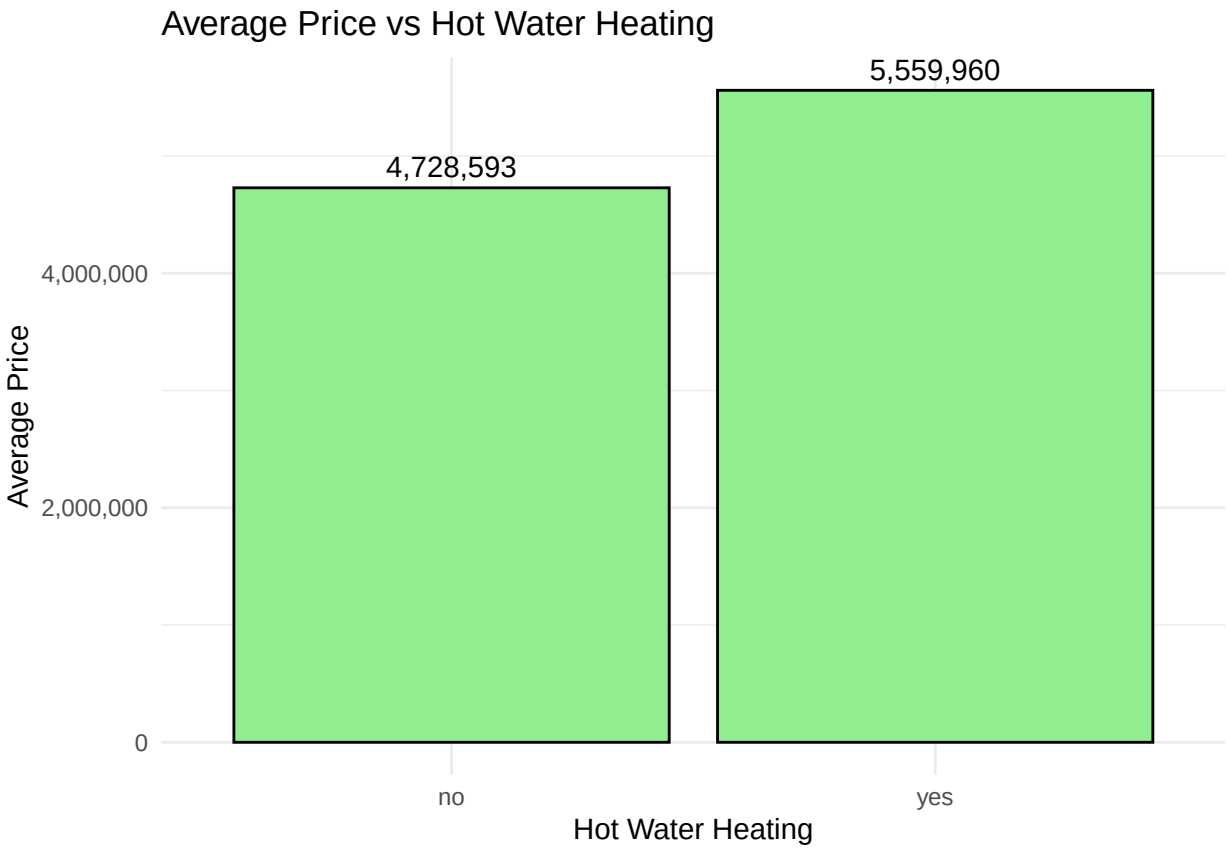
The bar plot illustrates the relationship between the average price of properties and the presence of basements. It indicates that properties with basements have a higher average price at 5,242,615, while properties without basement have a lower average price of 4,509,966. This reveals the added value 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.

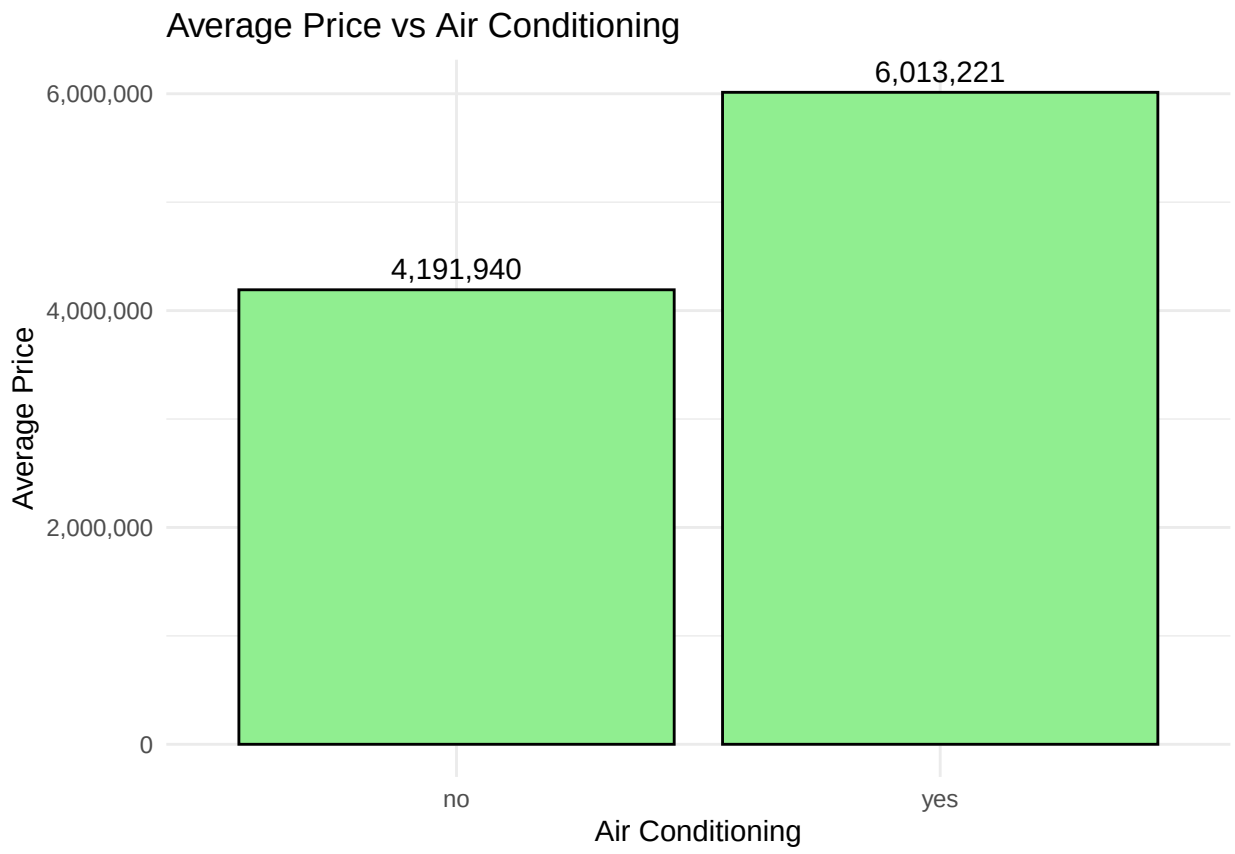


The bar plot shows the relationship between the average price of properties and the number of parking. Properties with 3 parking have the highest average price at 5,896,328, followed by 4 parking properties with an average price of 5,867,167. Additionally, Properties with 1 and 2 parking have an average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

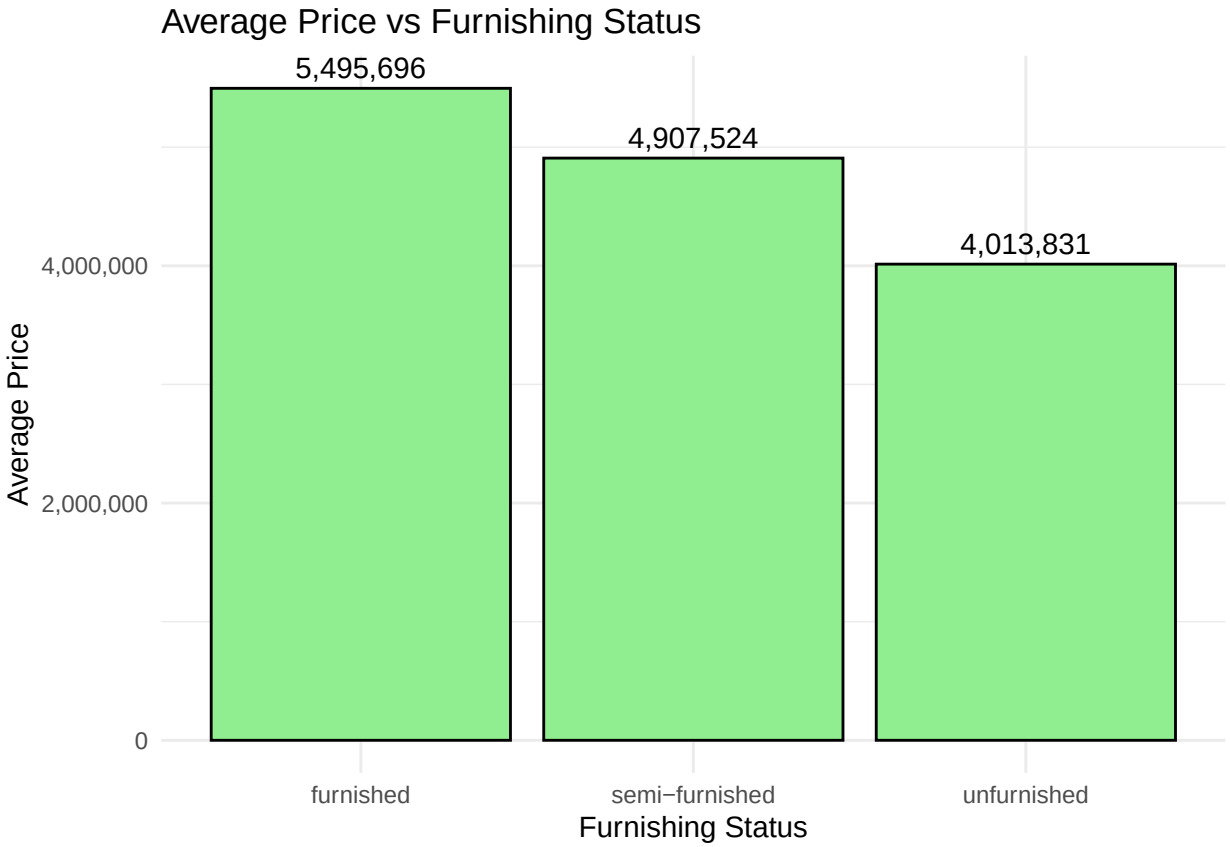
Overall, the bar plot demonstrates a positive relationship between the number of parking and the average price. This suggests that parking is a valuable factors of housing price, which tend to be more expensive as the number of parking spaces increases.



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.

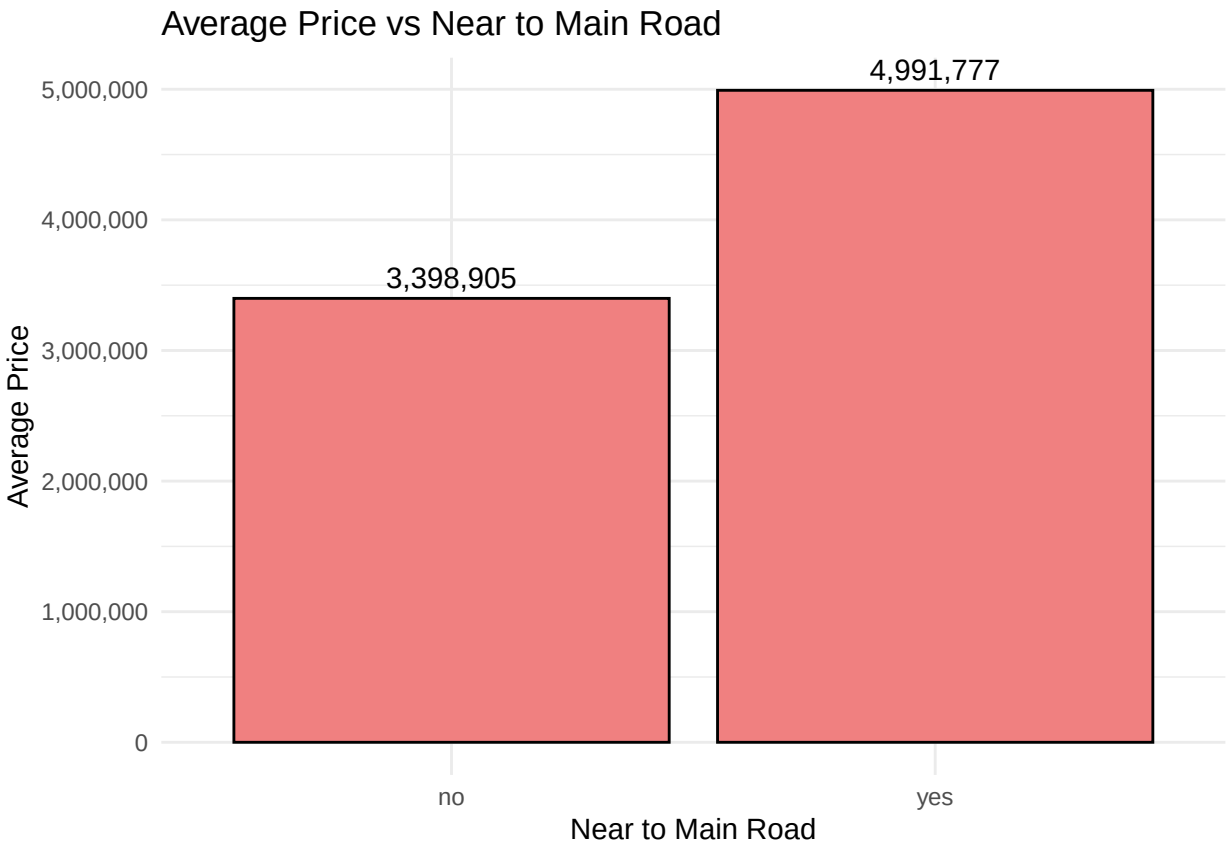


The bar plot shows the relationship between the average price of properties and the presence of air conditioning. Properties with air conditioning have a higher average price of 6,013,221, while properties without air conditioning have a lower average price at 4,191,940. This suggests that air conditioning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

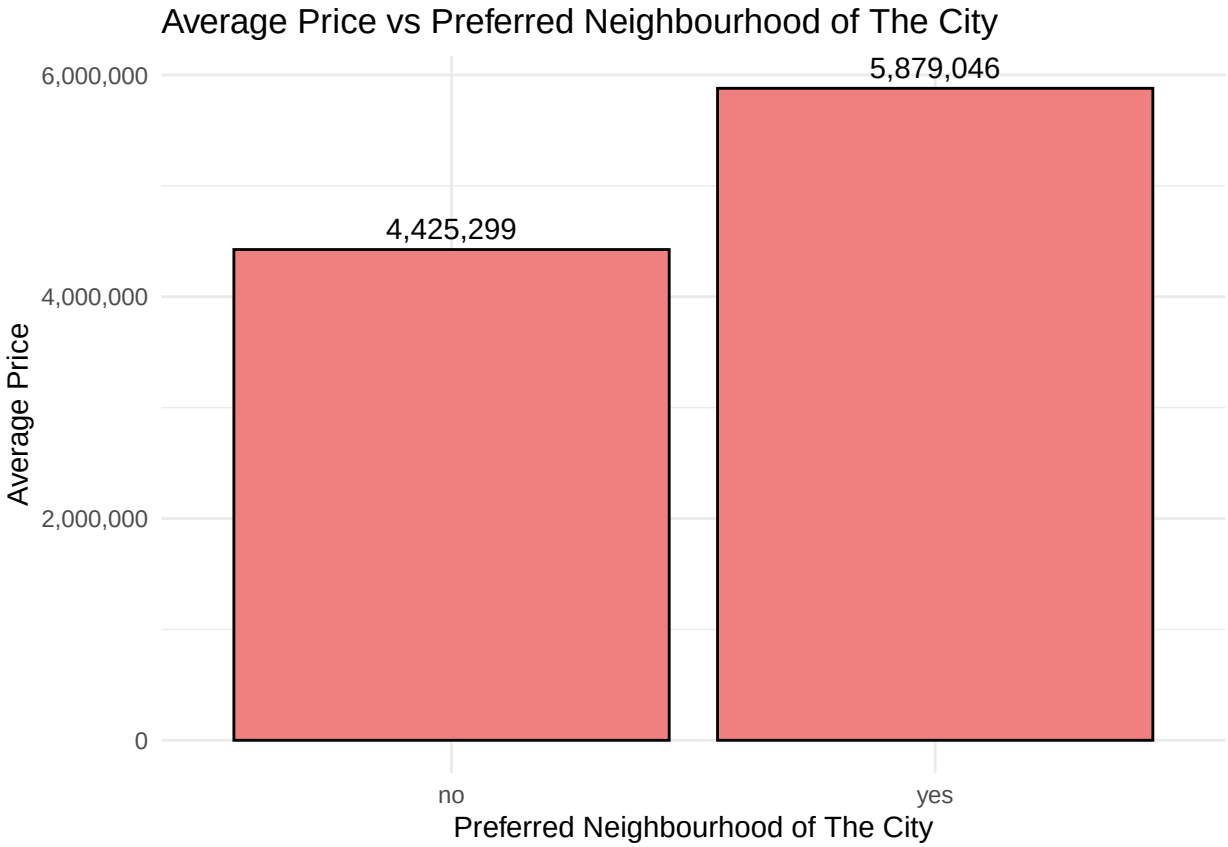


The bar plot shows the relationship between the average price of properties and the furnishing status. It reveals that furnished properties have the highest average price at 5,495,696. While semi-furnished properties have an average price of 4,907,524, unfurnished properties have the lowest average price 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.

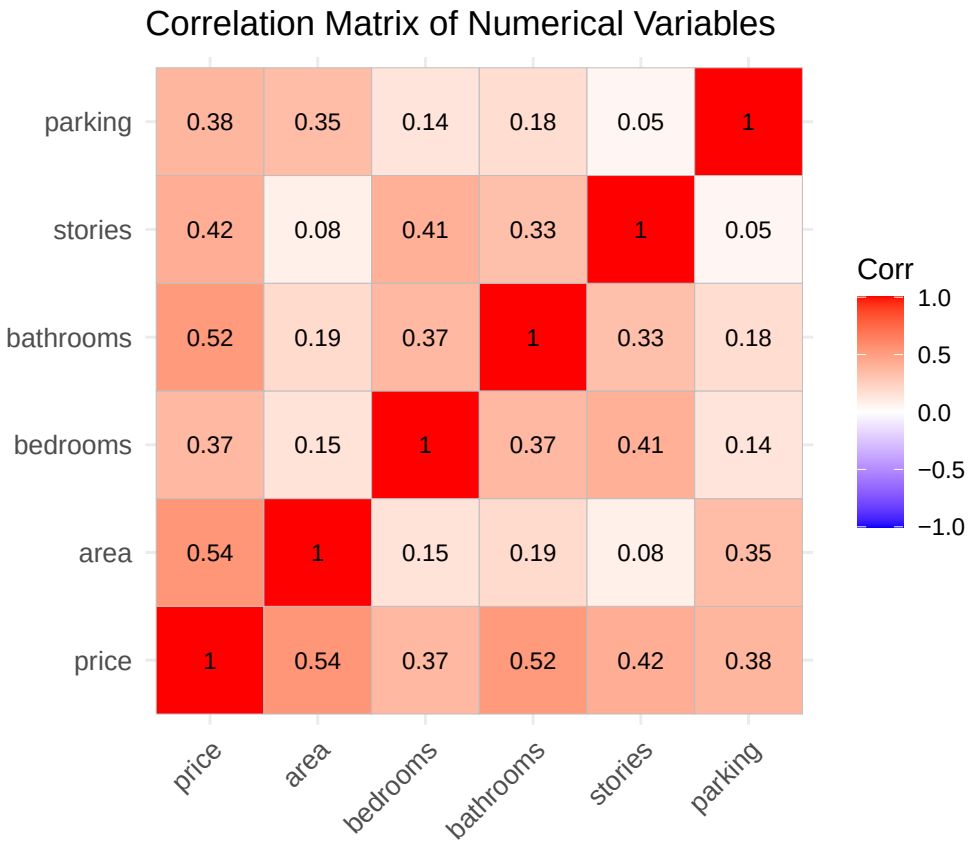


The bar plot illustrates the relationship between the average price of properties and the distance to main road. It shows that properties near to the main road have a significantly higher average price of 4,991,777, while properties not near the main road have a lower average price of 3,398,905. This 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.

Correlation Matrix



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.

Modeling approach

Model type

This study employed multiple linear regression to predict house prices based on predictors like area, bedrooms, bathrooms, stories, and additional property features such as parking and basement availability. Linear

regression was chosen due to its simplicity, interpretability, and ability to model linear relationships between predictors and the target variable. The modeling process began with exploratory data analysis (EDA) to identify significant variables, followed by testing interaction terms like area with stories, bedrooms with air conditioning to evaluate combined effects. While these interactions added complexity, they often did not significantly improve the model's predictive accuracy, highlighting the utility of simpler models.

Model selection

Model1 – Multiple Linear Regression (All Variables):

The initial model used all predictors, achieving an adjusted R-squared value of 0.674, explaining 67.4% of the variance in house prices. Key predictors such as area, bathrooms, stories, air conditioning, and preferred area significantly impacted house prices. For example, air conditioning added approximately $864,958 to a house's price, while location in a preferred area contributed $651,543 on average. Variance Inflation Factor (VIF) analysis indicated no multicollinearity concerns, as all predictors exhibited VIF values below 2. Model1 emerged as a robust baseline due to its interpretability and predictive performance.

Model_cat – Area Categorized into Quartiles:

This variation categorized the continuous area variable into three levels (Small, Mid, Big) based on quartiles. While this approach simplified interpretation, properties in smaller categories exhibited significant negative impacts on price. Adjusted R-squared dropped slightly, making Model1 the preferred choice for its superior explanatory power and flexibility.

Model_interaction – Area * Bedrooms:

Adding an interaction between area and bedrooms increased the adjusted R-squared to 0.676 but introduced multicollinearity, with VIF values exceeding acceptable thresholds for area and interaction terms. Although interaction effects were significant, multicollinearity diminished interpretability, favoring Model1 for practical application.

Model_interaction2 – Area * Stories:

This model included an interaction between area and stories, achieving a slightly higher adjusted R-squared of 0.676. The interaction term was significant ($p = 0.0346$), but high multicollinearity persisted (VIF = 19.39). The model's complexity outweighed its marginal performance gains, favoring simpler alternatives.

Model_interaction3 – Area * Stories + Bedrooms * Air Conditioning:

Incorporating two interaction terms yielded an adjusted R-squared of 0.677, the highest among all models. However, multicollinearity remained problematic, with interaction term VIF values exceeding 20. These issues undermined the reliability of coefficient estimates, making Model1 the optimal choice.

Final Model: Model1

Model1 emerged as the preferred model due to its balance of simplicity, robustness, and interpretability. It avoids multicollinearity while maintaining strong statistical significance for key predictors, offering actionable insights into house price determinants.

Model Assumptions

The assumptions of multiple linear regression including linearity, independence, normality of residuals, and homoscedasticity were evaluated and largely satisfied. Residual plots indicated no major deviations, and VIF analysis confirmed the absence of multicollinearity.

Key Findings

```
##
## 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 regression analysis revealed that property size, the number of bathrooms, multiple stories, and additional amenities such as a guestroom, basement, and parking space significantly contributed to higher house prices, supporting the corresponding hypotheses. Modern features like hot water heating systems and air conditioning also had a positive impact on prices, reflecting the demand for comfort. Location factors, including proximity to main roads and desirable neighborhoods, further increased house values. While furnishing status had a limited effect, with unfurnished houses showing lower prices, the number of bedrooms did not significantly influence house prices, contradicting initial expectations. This finding suggests that property area is a more decisive factor in determining value than bedroom count.

Summary

While interaction terms offered nuanced insights, their introduction often led to multicollinearity, complicating interpretation. The simpler Model1 provided a robust and interpretable framework, explaining 67.4% of the variance in house prices and identifying significant predictors that aligned with market expectations. Additional models and findings are discussed in the supplementary section.