Statistical Analysis of Factors Influencing House Prices in India

Sumanth Polati

2024-10-05


```
# Load required libraries
library(readr) # For reading data files
# Load the dataset
data <- read.csv("C:/Users/Sumanth/Downloads/housing_price_dataset.csv")
```

1) Introduction:

The real estate market is a fascinating world, full of houses, apartments, and neighborhoods each with its own story. Ever wondered what makes a house more expensive than another? That's what we're here to find out!

In our analysis, we're diving into a dataset with information about houses – things like how big they are, how many bedrooms and bathrooms they have, when they were built, and where they're located. By looking at all this data, we hope to figure out what factors influence how much a house costs.

It's like solving a puzzle. We'll use statistics and graphs to see if there are any patterns or trends that can help us understand why some houses are more valuable than others. Our goal is to make sense of the complexities of the real estate market and uncover the secrets behind property prices.

Data set source : Kaggle

2) Overview Of Data

```
# Load required libraries
library(readr) # For reading data files

# Load the dataset
data <- read.csv("C:/Users/Sumanth/Downloads/housing_price_dataset.csv")

str(data)
```

```
## 'data.frame':    50000 obs. of  6 variables:
##  $ SquareFeet  : int  2126 2459 1860 2294 2130 2095 2724 2044 2638 1121
...
##  $ Bedrooms    : int  4 3 2 2 5 2 2 4 4 5 ...
##  $ Bathrooms   : int  1 2 1 1 2 3 1 3 3 2 ...
##  $ Neighborhood: chr  "Rural" "Rural" "Suburb" "Urban" ...
##  $ YearBuilt   : int  1969 1980 1970 1996 2001 2020 1993 1957 1959 2004
...
##  $ Price       : num  215355 195014 306891 206787 272436 ...

head(data)

##   SquareFeet Bedrooms Bathrooms Neighborhood YearBuilt    Price
## 1       2126        4         1        Rural      1969 215355.3
## 2       2459        3         2        Rural      1980 195014.2
## 3       1860        2         1       Suburb      1970 306891.0
## 4       2294        2         1        Urban      1996 206786.8
## 5       2130        5         2       Suburb      2001 272436.2
## 6       2095        2         3       Suburb      2020 198208.8

missing_values <- colSums(is.na(data))
print(missing_values)

##   SquareFeet     Bedrooms    Bathrooms Neighborhood    YearBuilt
Price
##            0            0            0            0            0
0

# Check for duplicates
duplicate_rows <- data[duplicated(data), ]
print(duplicate_rows)

## [1] SquareFeet   Bedrooms     Bathrooms    Neighborhood YearBuilt
## [6] Price
## <0 rows> (or 0-length row.names)
```

3) RESEARCH QUESTION

"How do change in factors such as square footage, number of bedrooms and bathrooms, neighborhood characteristics, and year built influence house prices?"

Research Objectives:

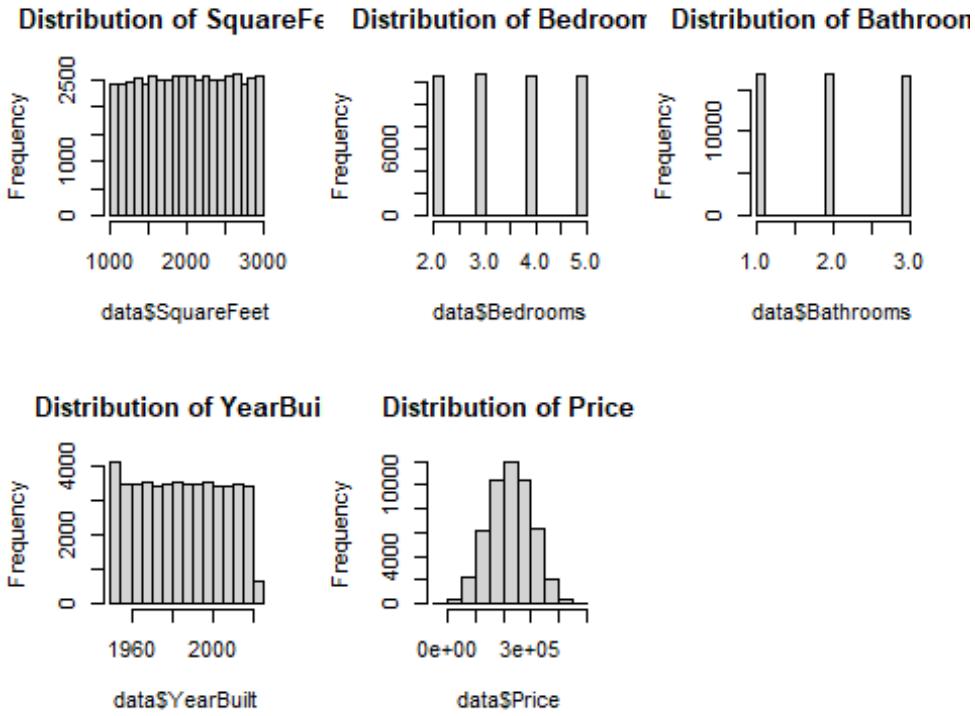
1) Explore the relationships between different variables, such as square footage, number of bedrooms and bathrooms, year built, neighborhood characteristics, and house prices.

2) Investigate the distributions of square footage, number of bedrooms and bathrooms, neighborhood characteristics, and year built to understand their spread and central tendencies.

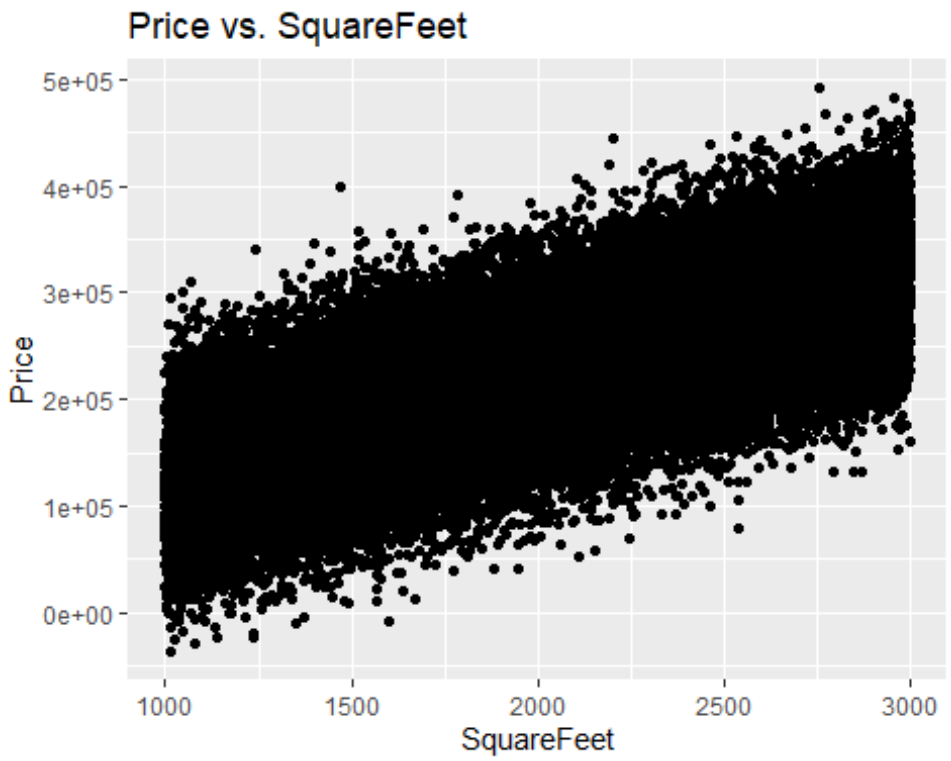
3) Determining the relative importance of each factor and their comibined impact the house prices.

4) Exploratory Data Analysis

```
# Summary statistics for SquareFeet
summary(data$SquareFeet)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    1513    2007    2006    2506    2999

#Summary statistics for Bedrooms
summary(data$Bedrooms)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.000   3.000   3.499   4.000   5.000

#Summary statistics Bathrooms
summary(data$Bathrooms)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.995   3.000   3.000

#Summary statistics for Price
summary(data$Price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -36588  169956  225052  224827  279374  492195

# Counts and frequencies for Neighborhood
table(data$Neighborhood)

##
##  Rural Suburb  Urban
##  16676  16721  16603

par(mfrow = c(2, 3))
hist(data$SquareFeet, main = "Distribution of SquareFeet")
hist(data$Bedrooms, main = "Distribution of Bedrooms")
hist(data$Bathrooms, main = "Distribution of Bathrooms")
hist(data$YearBuilt, main = "Distribution of YearBuilt")
hist(data$Price, main = "Distribution of Price")
```

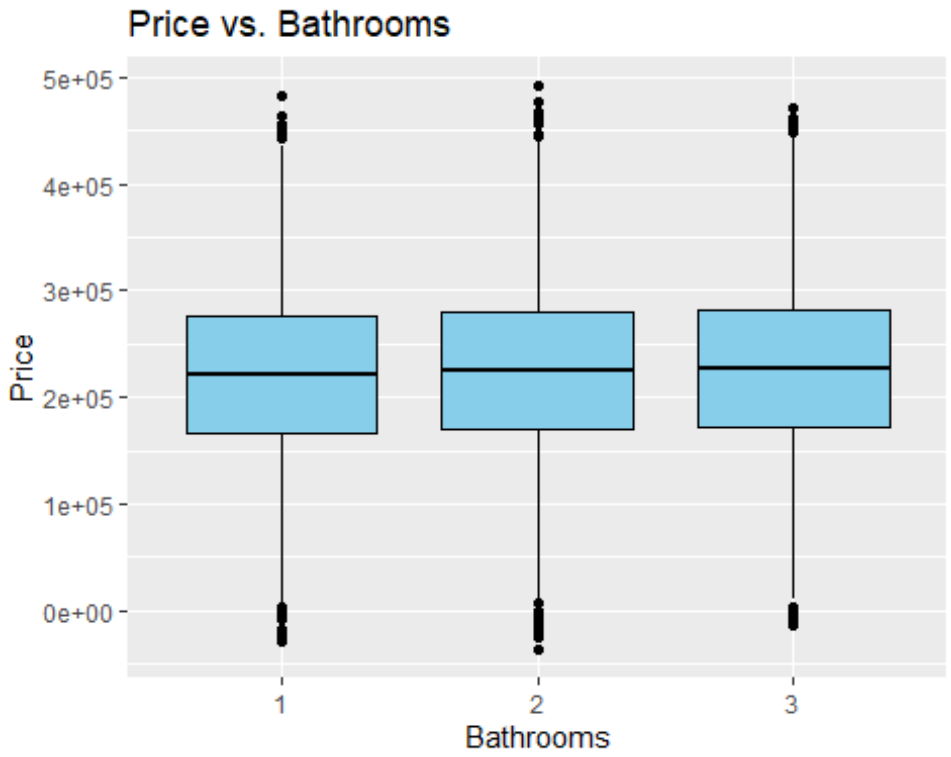



```r
# Visualize the relationship between SquareFeet and Price
library(ggplot2)
ggplot(data, aes(x = SquareFeet, y = Price)) +
  geom_point() +
  labs(title = "Price vs. SquareFeet",
       x = "SquareFeet",
       y = "Price")
```

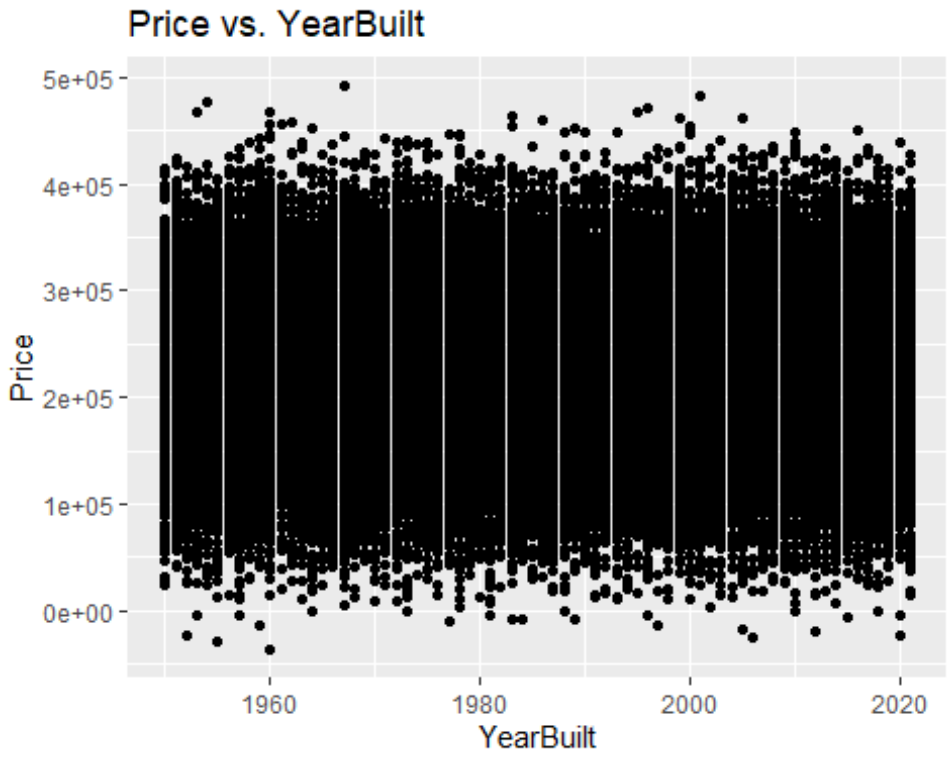



```r
# Visualize the relationship between Bedrooms and Price
ggplot(data, aes(x = factor(Bedrooms), y = Price)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Price vs. Bedrooms",
       x = "Bedrooms",
       y = "Price")
```


```r
# Visualize the relationship between Bathrooms and Price
ggplot(data, aes(x = factor(Bathrooms), y = Price)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Price vs. Bathrooms",
       x = "Bathrooms",
       y = "Price")
```




```r
# Visualize the relationship between YearBuilt and Price
ggplot(data, aes(x = YearBuilt, y = Price)) +
  geom_point() +
  labs(title = "Price vs. YearBuilt",
       x = "YearBuilt",
       y = "Price")
```



5)Statistical Analysis

1)Does the mean house price vary significantly across different combinations of bedrooms and bathrooms?

2)Does the mean of the house price vary significantly across different Neighborhoods ?

3)Is there a significant difference in the mean house prices between houses built before and after a certain year?

4)Is there a significant difference in the mean price of properties between urban and suburban neighborhoods?

5)What is the probability of a property being sold above a certain price threshold?

6)Do properties located in urban neighborhoods have a significantly higher probability of being sold at prices above a certain threshold compared to properties in suburban and rural neighborhoods?

```
# Calculate correlation matrix
correlation_matrix <- cor(data[, c("SquareFeet", "Bedrooms", "Bathrooms",
"Price")])

# Print correlation matrix
print(correlation_matrix)
```

```
##              SquareFeet     Bedrooms    Bathrooms      Price
## SquareFeet  1.000000000 -0.002638119 -0.003274733 0.75071979
```

```
## Bedrooms    -0.002638119  1.000000000  0.007405043 0.07262393
## Bathrooms   -0.003274733  0.007405043  1.000000000 0.02841765
## Price        0.750719786  0.072623932  0.028417648 1.00000000
```

Regression Model

Results:

1)Coefficients: SquareFeet: For every additional square foot in a house, the expected increase in price is approximately ₹99.34, holding all other variables constant.

Bedrooms: Each additional bedroom is associated with an expected increase in price of approximately ₹5,074.44, all else being equal.

Bathrooms: Each additional bathroom is associated with an expected increase in price of approximately ₹2,833.84, holding all other variables constant.

YearBuilt: However, the year the house was built (YearBuilt) does not appear to have a statistically significant effect on the price, as the p-value is greater than the typical significance level of 0.05.

Neighborhood Suburb and Neighborhood Urban: Compared to a reference neighborhood, houses in urban neighborhoods have an expected increase in price of approximately ₹1,550.09, while houses in suburban neighborhoods have a decrease in price of approximately ₹675.49, though the latter is not statistically significant.

2)Significance Levels:

The coefficients for SquareFeet, Bedrooms, Bathrooms, and NeighborhoodUrban are all highly significant, indicated by '***', suggesting strong evidence that these variables have a meaningful impact on house prices.

YearBuilt and NeighborhoodSuburb are not statistically significant predictors of house prices, indicated by '.'.

3)Residuals: The residuals represent the differences between the observed prices and the prices predicted by the model. They provide insights into the variability of the model's predictions.

4)Multiple R-squared and Adjusted R-squared: These metrics indicate how well the overall model fits the data. A value of 0.5702 for both metrics suggests that the model explains approximately 57% of the variability in house prices.

5)F-statistic and p-value: The F-statistic tests the overall significance of the regression model. The extremely low p-value (< 2.2e-16) indicates that the regression model as a whole is highly significant, suggesting that at least one of the predictors significantly explains the variability in house prices.

```
model <- lm(Price ~ SquareFeet + Bedrooms + Bathrooms + YearBuilt + ., data =
data)
summary(model)
```

```
##
## Call:
## lm(formula = Price ~ SquareFeet + Bedrooms + Bathrooms + YearBuilt +
##     ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -189150  -34015    -204   33724  227758
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         23431.407  21423.752   1.094  0.27409
## SquareFeet             99.340      0.388 256.062  < 2e-16 ***
## Bedrooms             5074.435    199.995  25.373  < 2e-16 ***
## Bathrooms            2833.835    273.654  10.356  < 2e-16 ***
## YearBuilt             -10.887     10.775  -1.010  0.31232
## NeighborhoodSuburb   -675.494    546.335  -1.236  0.21631
## NeighborhoodUrban    1550.088    547.328   2.832  0.00463 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49920 on 49993 degrees of freedom
## Multiple R-squared:  0.5702, Adjusted R-squared:  0.5702
## F-statistic: 1.106e+04 on 6 and 49993 DF,  p-value: < 2.2e-16
```

Anova Test:

1)Does the mean house price vary significantly across different combinations of bedrooms and bathrooms?

Bedrooms: The p-value for the Bedrooms factor is less than 0.05 ($p < 0.05$), indicating that the number of bedrooms has a significant effect on house prices.

Bathrooms: Similarly, the p-value for the Bathrooms factor is less than 0.05 ($p < 0.05$), suggesting that the number of bathrooms also has a significant effect on house prices.

Interaction (Bedrooms:Bathrooms): However, the p-value for the interaction between Bedrooms and Bathrooms (Bedrooms:Bathrooms) is greater than 0.05 ($p > 0.05$), indicating that the interaction effect between the number of bedrooms and bathrooms is not statistically significant. This means that the combined effect of bedrooms and bathrooms on house prices is not significantly different from what would be expected based on the individual effects of bedrooms and bathrooms.

In summary, while the individual factors of Bedrooms and Bathrooms have significant effects on house prices, their interaction does not significantly influence house prices.

```r
# Perform ANOVA
anova_result <- aov(Price ~ Bedrooms * Bathrooms, data = data)
summary(anova_result)
```

```
##                       Df    Sum Sq   Mean Sq F value   Pr(>F)
## Bedrooms               1 1.529e+12 1.529e+12 265.303  < 2e-16 ***
## Bathrooms              1 2.253e+11 2.253e+11  39.101 4.06e-10 ***
## Bedrooms:Bathrooms     1 7.685e+09 7.685e+09   1.334    0.248
## Residuals          49996 2.881e+14 5.763e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2)Does the mean of the house price vary significantly across different Neighborhoods ?

The p-value for the Neighborhood factor is less than 0.05 ($p < 0.05$), indicating that the neighborhood variable has a significant effect on house prices.

Mean house price varies significantly across different neighborhoods. The significant p-value suggests that the neighborhood variable is an important factor in determining house prices.

```r
# Perform ANOVA
anova_result <- aov(Price ~ Neighborhood, data = data)

# Summary of ANOVA
summary(anova_result)
```

```
##                 Df    Sum Sq   Mean Sq F value   Pr(>F)
## Neighborhood     2 1.422e+11 7.109e+10   12.27 4.72e-06 ***
## Residuals    49997 2.897e+14 5.795e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypothesis Testing

1)Is there a significant difference in the mean house prices between houses built before and after a certain year?

Null Hypothesis (H0): There is no significant difference in the mean house prices between houses built before and after a certain year.

Alternative Hypothesis (H1): There is a significant difference in the mean house prices between houses built before and after a certain year.

The test yielded a t-value of 0.17606 with a corresponding degrees of freedom (df) of 28475.

The calculated p-value was found to be 0.8602.

The 95% confidence interval for the difference in means between houses built after and before the specified year ranged from -1327.002 to 1588.924.

The sample mean house price for houses built after the specified year was estimated to be 224918.7, while the sample mean house price for houses built before the specified year was estimated to be 224787.8.

Based on the below results p-value of 0.8602, we fail to reject the null hypothesis.

Therefore, there is insufficient evidence to conclude that there is a significant difference in the mean house prices between houses built before and after the specified year.

```
# Create a binary variable for houses built before or after the specified
year

cutoff_year <- 2000

data$YearGroup <- ifelse(data$YearBuilt < cutoff_year, "Before", "After")

# Perform t-test
t_test_result <- t.test(Price ~ YearGroup, data = data)
t_test_result

##
##  Welch Two Sample t-test
##
## data:  Price by YearGroup
## t = 0.17606, df = 28475, p-value = 0.8602
## alternative hypothesis: true difference in means between group After and
group Before is not equal to 0
## 95 percent confidence interval:
##  -1327.002  1588.924
## sample estimates:
##  mean in group After mean in group Before
##             224918.7             224787.8
```

2)Is there a significant difference in the mean price of properties between urban and suburban neighborhoods?"

The Welch Two Sample t-test results indicate a significant difference in mean prices between urban and suburban neighborhoods. The p-value is 2.461e-06, which is much less than the significance level of 0.05, suggesting strong evidence against the null hypothesis. Therefore, we reject the null hypothesis and conclude that there is a significant difference in mean prices between urban and suburban neighborhoods.

```
# Subset the data for urban and suburban neighborhoods
urban_prices <- data$Price[data$Neighborhood == "Urban"]
suburban_prices <- data$Price[data$Neighborhood == "Suburb"]

# Perform t-test
t_test_result <- t.test(urban_prices, suburban_prices)
t_test_result

##
##  Welch Two Sample t-test
##
## data:  urban_prices and suburban_prices
```

```
## t = 4.7122, df = 33321, p-value = 2.461e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2296.486 5567.529
## sample estimates:
## mean of x mean of y
##  227166.2  223234.2
```

Probability Test

1)What is the probability of a property being sold above a certain price threshold?

Results :

The probability of a property being sold above a certain price threshold, as indicated by the result of 0.38154, suggests that approximately 38.15% of the properties in the dataset are sold at prices exceeding the specified threshold.

```
# Define the price threshold
threshold <-  250000

# Calculate the proportion of properties sold above the threshold
probability_above_threshold <- sum(data$Price > threshold) / nrow(data)

# Print the probability
print(probability_above_threshold)

## [1] 0.38154
```

2)Do properties located in urban neighborhoods have a significantly higher probability of being sold at prices above a certain threshold compared to properties in suburban and rural neighborhoods?

Result:

Number of successes (properties sold at prices above the threshold) in urban neighborhoods: 2996 Number of trials (total properties sold at prices above the threshold): 8671 The p-value is less than 2.2e-16, indicating strong evidence against the null hypothesis. The alternative hypothesis suggests that the true probability of success (properties sold at prices above the threshold) is not equal to the proportion of properties in suburban and rural neighborhoods combined. The estimated probability of success (properties sold at prices above the threshold) in urban neighborhoods is approximately 0.3455.

These results suggest a significant difference in the probability of properties being sold at prices above the threshold between urban neighborhoods and suburban/rural neighborhoods combined. Specifically, properties in urban neighborhoods have a higher probability of being sold at prices exceeding the threshold compared to properties in suburban and rural areas.

```r
# Set the threshold price
threshold_price <- 300000  # Example threshold price

# Filter properties sold at prices above the threshold
properties_above_threshold <- subset(data, Price > threshold_price)

# Calculate the proportion of properties in urban neighborhoods sold at
prices above the threshold
proportion_urban <- mean(properties_above_threshold$Neighborhood == "Urban")

# Calculate the proportion of properties in suburban and rural neighborhoods
combined sold at prices above the threshold
proportion_suburban_rural <- mean(properties_above_threshold$Neighborhood !=
"Urban")

# Perform a binomial probability test
binom_test_result <- binom.test(sum(properties_above_threshold$Neighborhood
== "Urban"), nrow(properties_above_threshold), p = proportion_suburban_rural,
alternative = "two.sided")

# Output the result
binom_test_result
```

```
##
##  Exact binomial test
##
## data:  sum(properties_above_threshold$Neighborhood == "Urban") and
nrow(properties_above_threshold)
## number of successes = 2996, number of trials = 8671, p-value < 2.2e-16
## alternative hypothesis: true probability of success is not equal to
0.6544805
## 95 percent confidence interval:
##  0.335506 0.355637
## sample estimates:
## probability of success
##              0.3455195
```

6) Conclusion

Several important trends emerged from our detailed research of the elements that influence housing values. First and foremost, we discovered a consistent association between specific property qualities and market prices. Specifically, we discovered that increases in square footage, bedrooms, and bathrooms were highly related with higher property prices. Interestingly, while neighborhood type played an important significance, with urban homes demanding greater prices than suburban ones, the year a property was built had no statistically significant effect on its price. Furthermore, our examination into the interaction of bedrooms and bathrooms indicated that, while these characteristics have an individual impact on pricing, their combined effect has no meaningful effect on house values. Additionally, probabilistic research revealed the possibility of properties being sold

beyond a specific price level, with urban properties having a significantly greater probability than their suburban and rural equivalents. Overall, these findings provide useful insights into the complex dynamics of the real estate market, emphasizing the diverse character of property pricing.