Exploratory Data Analysis

Descriptive Statistics

There seems to be information on property listings in the "housing" dataset. Each property has different properties included in it. Let's examine the salient points: This dataset contains details on residential properties in Melbourne, Australia's Abbotsford area. Every row corresponds to a distinct property listing. The dataset contains information on the property's address, number of rooms, kind of property (home, for example), price, mode of sale, and selling agent.

It also gives specifics on the land size and parking places, as well as the features of the property, such as the number of bedrooms and bathrooms. The building's dimensions and year of construction are among the other details included in the dataset. Several of the properties have a fascinating historical background, having been established as early as the 1900s. The location of the property is disclosed, together with its latitude, longitude, postcode, and separation from the city center. The property is located in what is referred to as the "Northern Metropolitan" region.

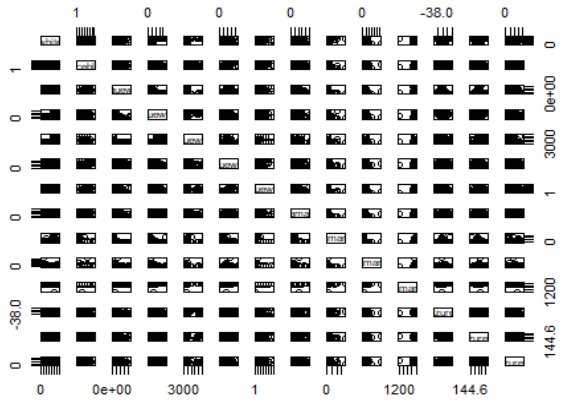
Finally, to give some information about the neighborhood and local area, the dataset includes the number of properties and the council area. This dataset appears to be useful for examining geographic patterns, property trends, and the dynamics of the Melbourne suburb of Abbotsford. Researchers interested in housing data and trends, data analysts, and real estate professionals may find it helpful.

```
housing =read.csv('data.csv')

head(housing)
```

```
##   X     Suburb         Address Rooms Type   Price Method SellerG      Date
## 1 1 Abbotsford 25 Bloomburg St     2    h 1035000      S  Biggin  4/2/2016
## 2 2 Abbotsford    5 Charles St     3    h 1465000     SP  Biggin  4/3/2017
## 3 3 Abbotsford     55a Park St     4    h 1600000     VB  Nelson  4/6/2016
## 4 4 Abbotsford    124 Yarra St     3    h 1876000      S  Nelson  7/5/2016
## 5 5 Abbotsford   98 Charles St     2    h 1636000      S  Nelson 8/10/2016
## 6 6 Abbotsford   10 Valiant St     2    h 1097000      S  Biggin 8/10/2016
##   Distance Postcode Bedroom2 Bathroom Car Landsize BuildingArea YearBuilt
## 1      2.5     3067        2        1   0      156           79      1900
## 2      2.5     3067        3        2   0      134          150      1900
## 3      2.5     3067        3        1   2      120          142      2014
## 4      2.5     3067        4        2   0      245          210      1910
## 5      2.5     3067        2        1   2      256          107      1890
## 6      2.5     3067        3        1   2      220           75      1900
##   CouncilArea Lattitude Longtitude           Regionname Propertycount
## 1       Yarra  -37.8079   144.9934 Northern Metropolitan          4019
## 2       Yarra  -37.8093   144.9944 Northern Metropolitan          4019
## 3       Yarra  -37.8072   144.9941 Northern Metropolitan          4019
## 4       Yarra  -37.8024   144.9993 Northern Metropolitan          4019
## 5       Yarra  -37.8060   144.9954 Northern Metropolitan          4019
## 6       Yarra  -37.8010   144.9989 Northern Metropolitan          4019
```

In a scatter plot, each data point represents a combination of two variables and is shown on a two-dimensional grid. It's an easy, visual method to look at how variables relate to one another and see trends or patterns. Because the data points are unconnected, they are perfect for spotting outliers and correlations.



The 'housing' dataset summary provides insightful information on Melbourne's housing stock. There are 22 variables and 6,830 observations in the dataset. The average house price is approximately 1,077,604 AUD, and the average number of rooms is roughly 3. These are the key statistics. The majority of properties can fit two automobiles and have one to two bathrooms. The average building space is 143.4 square meters, while the median land size is 404 square meters. With a median year of 1970, the dataset covers the years 1196 to 2018 in terms of construction. The aforementioned statistics offer a brief overview of the Melbourne housing market, rendering them an invaluable tool for real estate analysts and academics examining regional housing patterns.

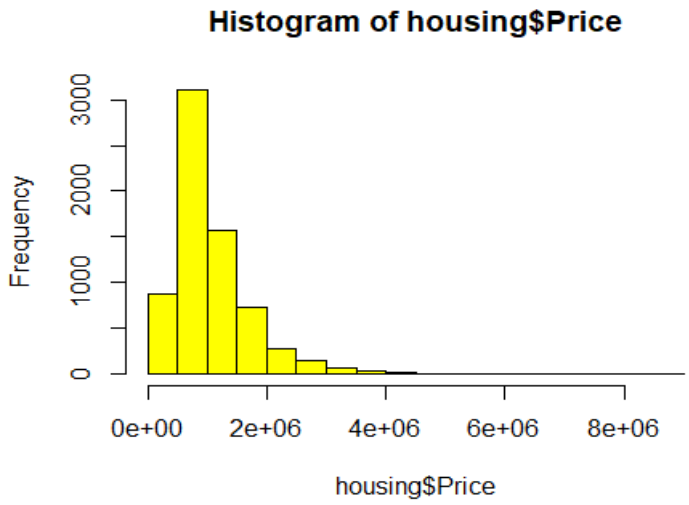
```
summary(housing)

##        X           Suburb            Address              Rooms
##  Min.   :   1   Length:6830        Length:6830        Min.   :1.000
##  1st Qu.:1708   Class :character   Class :character   1st Qu.:2.000
##  Median :3416   Mode  :character   Mode  :character   Median :3.000
##  Mean   :3416                                         Mean   :2.978
##  3rd Qu.:5123                                         3rd Qu.:4.000
##  Max.   :6830                                         Max.   :8.000
##      Type               Price            Method            SellerG
##  Length:6830        Min.   : 131000   Length:6830        Length:6830
##  Class :character   1st Qu.: 630000   Class :character   Class :character
##  Mode  :character   Median : 890000   Mode  :character   Mode  :character
##                     Mean   :1077604
##                     3rd Qu.:1334000
##                     Max.   :9000000
##      Date              Distance         Postcode       Bedroom2
##  Length:6830        Min.   : 0.00   Min.   :3000   Min.   :0.000
##  Class :character   1st Qu.: 6.10   1st Qu.:3044   1st Qu.:2.000
##  Mode  :character   Median : 9.20   Median :3083   Median :3.000
##                     Mean   :10.15   Mean   :3104   Mean   :2.951
##                     3rd Qu.:13.00   3rd Qu.:3147   3rd Qu.:4.000
##                     Max.   :47.40   Max.   :3977   Max.   :9.000
##     Bathroom          Car             Landsize        BuildingArea
##  Min.   :1.000   Min.   : 0.000   Min.   :    0.0   Min.   :   0.0
##  1st Qu.:1.000   1st Qu.: 1.000   1st Qu.:  167.0   1st Qu.:  93.0
##  Median :1.000   Median : 2.000   Median :  404.0   Median : 126.0
##  Mean   :1.594   Mean   : 1.607   Mean   :  487.5   Mean   : 143.4
##  3rd Qu.:2.000   3rd Qu.: 2.000   3rd Qu.:  641.0   3rd Qu.: 173.0
##  Max.   :8.000   Max.   :10.000   Max.   :37000.0   Max.   :3112.0
##    YearBuilt    CouncilArea          Lattitude        Longtitude
##  Min.   :1196   Length:6830        Min.   :-38.16   Min.   :144.5
##  1st Qu.:1940   Class :character   1st Qu.:-37.86   1st Qu.:144.9
##  Median :1970   Mode  :character   Median :-37.80   Median :145.0
##  Mean   :1964                      Mean   :-37.81   Mean   :145.0
##  3rd Qu.:2000                      3rd Qu.:-37.76   3rd Qu.:145.1
##  Max.   :2018                      Max.   :-37.41   Max.   :145.5
##   Regionname        Propertycount
##  Length:6830        Min.   :  389
##  Class :character   1st Qu.: 4381
##  Mode  :character   Median : 6567
##                     Mean   : 7434
##                     3rd Qu.:10175
##                     Max.   :21650
```

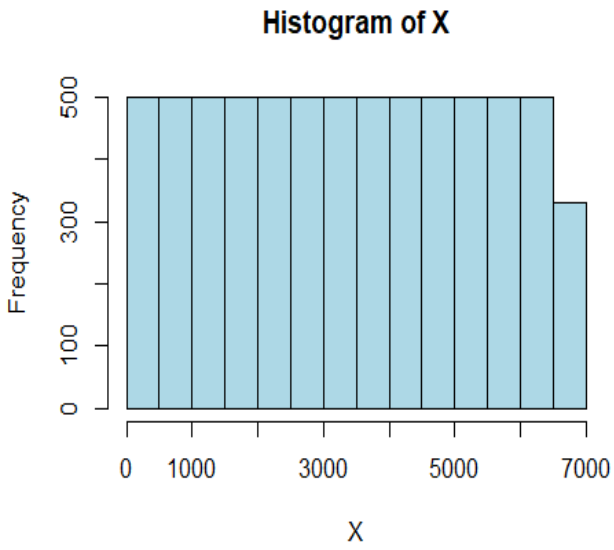
Visualization

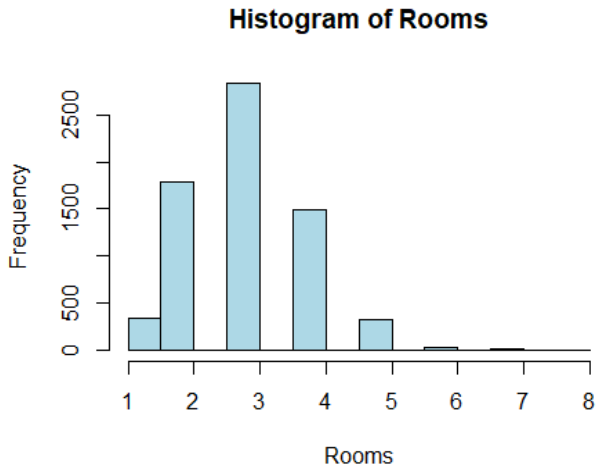
A histogram shows the frequency of values within predetermined intervals or bins and is a graphical representation of the data distribution. It is an essential tool for exploratory data analysis (EDA) since it gives a quick overview of data patterns like central tendency and dispersion visually.

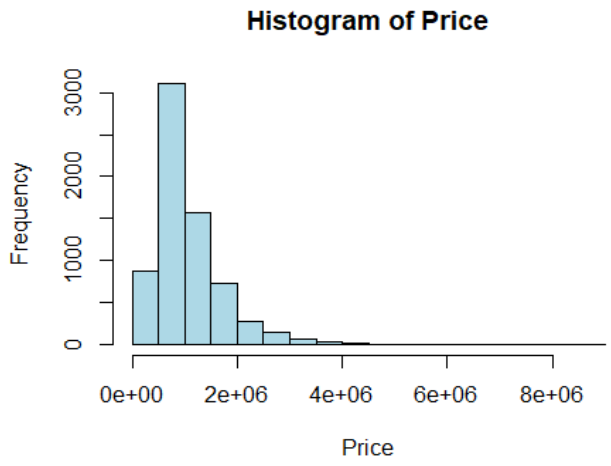
Histogram of our response variable

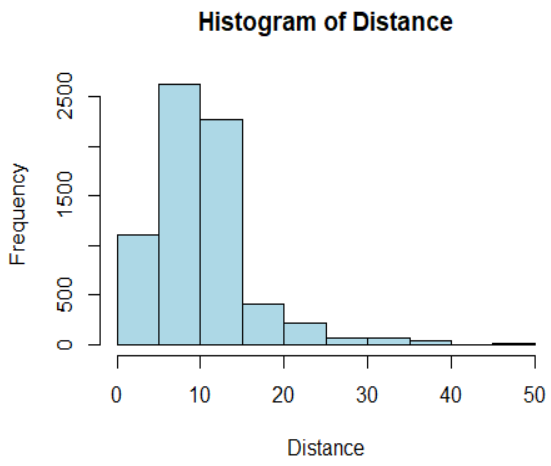


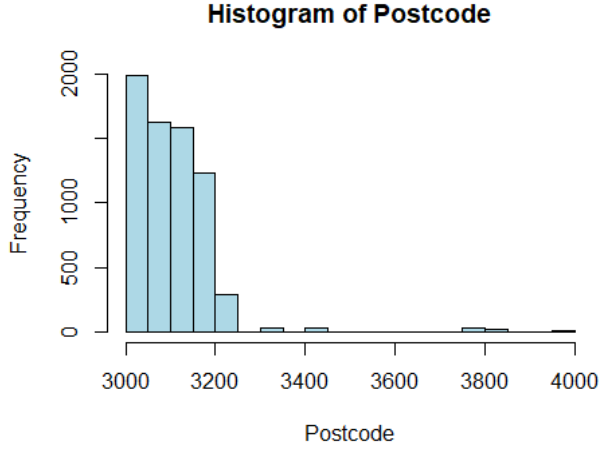
Histograms of our predict variables

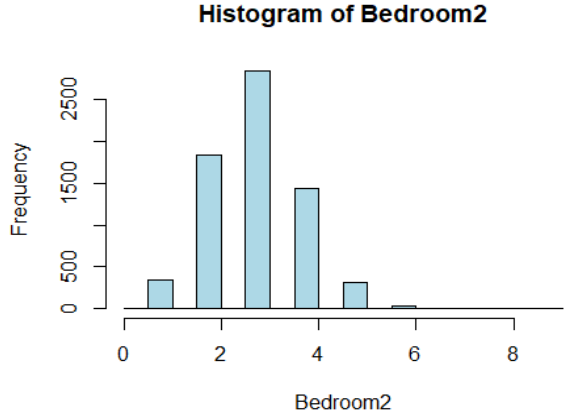


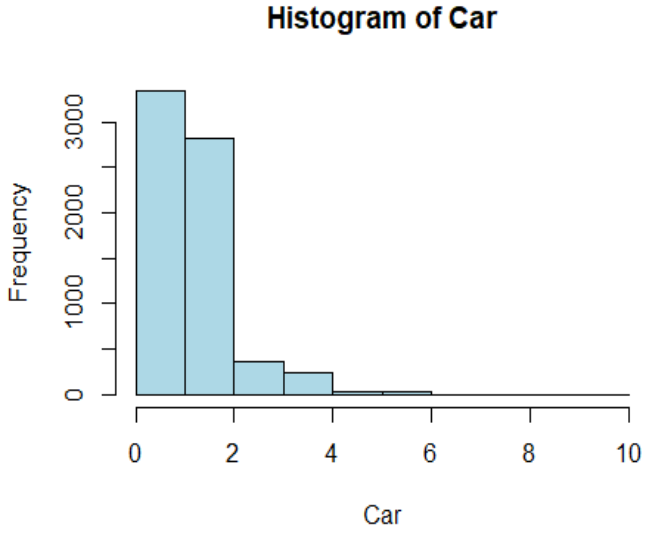


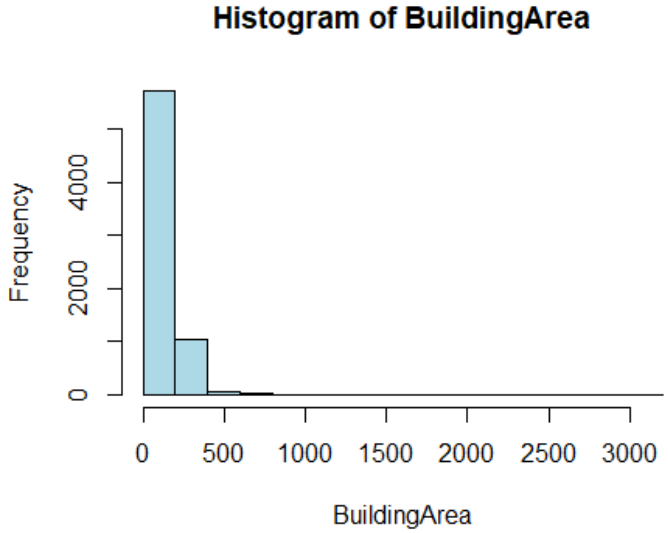

Histogram of Price
Frequency
0
1000
2000
3000
0e+00
2e+06
4e+06
6e+06
8e+06
Price

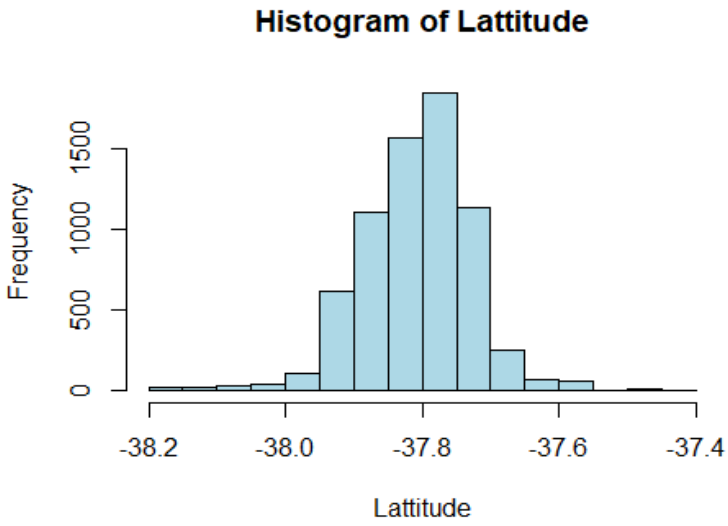

Histogram of Distance
Frequency
0
500
1500
2500
0
10
20
30
40
50
Distance

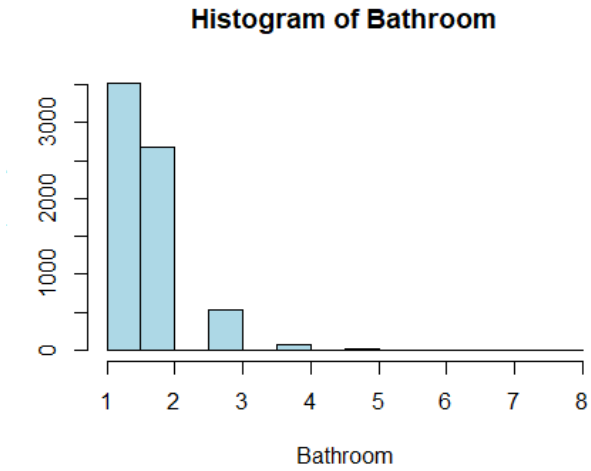

Histogram of Postcode
Frequency
0
500
1000
2000
3000
3200
3400
3600
3800
4000
Postcode

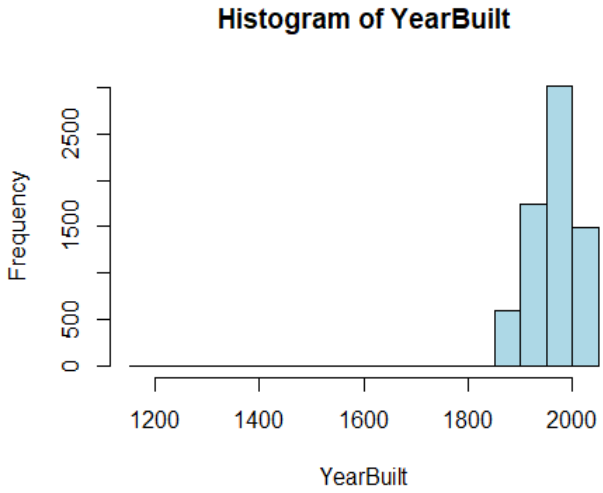

Histogram of Bedroom2
Frequency
0
500
1500
2500
0
2
4
6
8
Bedroom2

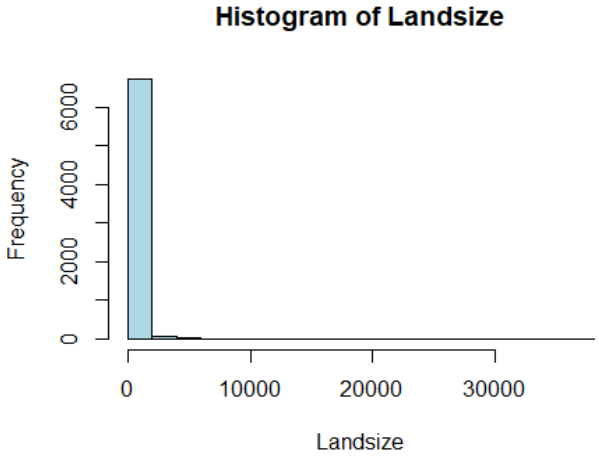

Histogram of Car
Frequency
0
1000
2000
3000
0
2
4
6
8
10
Car

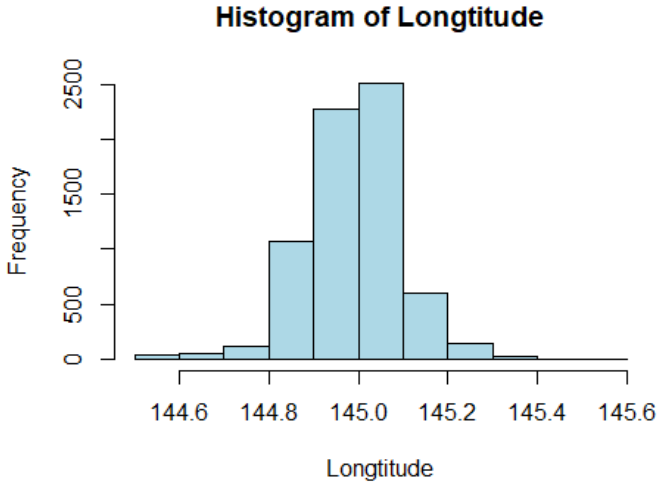

Histogram of BuildingArea
Frequency
0
2000
4000
0
500
1000
1500
2000
2500
3000
BuildingArea

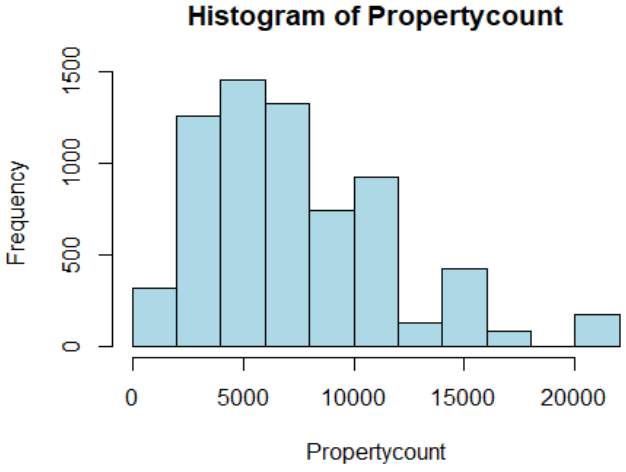

Histogram of Lattitude
Frequency
0
500
1000
1500
-38.2
-38.0
-37.8
-37.6
-37.4
Lattitude


Histogram of Bathroom
0
1000
2000
3000
1
2
3
4
5
6
7
8
Bathroom

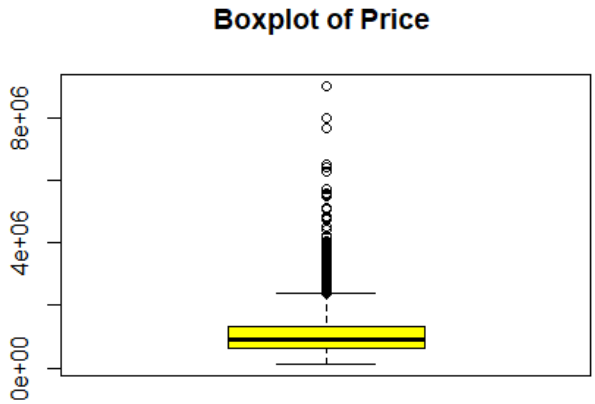

Histogram of YearBuilt
Frequency
0
500
1500
2500
1200
1400
1600
1800
2000
YearBuilt


Histogram of Landsize
Frequency
0
2000
4000
6000
0
10000
20000
30000
Landsize

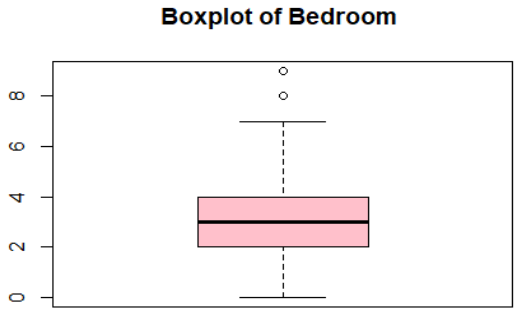

Histogram of Longtitude
Frequency
0
500
1500
2500
144.6
144.8
145.0
145.2
145.4
145.6
Longtitude

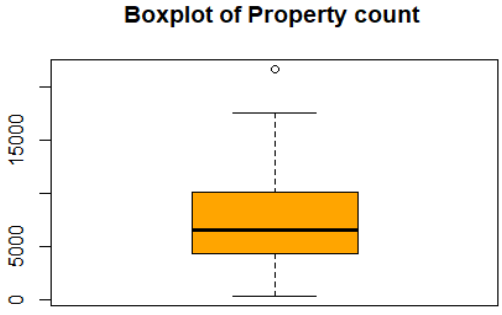

Histogram of Propertycount
Frequency
0
500
1000
1500
0
5000
10000
15000
20000
Propertycount

A boxplot, sometimes referred to as a box-and-whisker plot, is a statistical and data analysis graphical representation. The interquartile range (IQR) of the data is shown by a rectangular box that shows the distribution of a dataset, with whiskers extending to the minimum and maximum values. Boxplots are a useful tool for exploratory data analysis because they can effectively visualize data distribution, central tendency, and find outliers (EDA).



The price distribution of a property is displayed in a boxplot. The center 50% of the data is represented by the central box, and a horizontal line indicates the median value. The interquartile range (IQR), or the difference between the third and first quartiles, is 1.5 times the whiskers' maximum and lowest values. Outlier values are those that fall outside of the whiskers. Most prices, as the boxplot illustrates, are in the $4 million to $8 million range. The price is $6 million on average. A small number of outliers fall between $4 million and $8 million. With a median price of $6 million, a boxplot of prices reveals that most values are in the $4 million to $8 million range.There are a few outliers below $4 million and above $8 million.





Analysis and Model Building

Correlation

The correlation matrix that was previously displayed looks at the connections between a number of variables and the response variable, "Price," in a dataset that may be connected to Melbourne real estate. The correlation coefficient, which has a range of -1 to 1, sheds light on the direction and intensity of these connections.

1. Rooms (0.518): A property's price and number of rooms have a somewhat positive association, meaning that larger properties typically cost more.

2. Bedroom2 (0.500): There is also a strong positive association between the number of bedrooms and the price, indicating that higher-priced residences typically have more bedrooms.

3. BuildingArea (0.520): This indicates that larger properties typically have higher price tags, as there is a positive association between the building area and price.

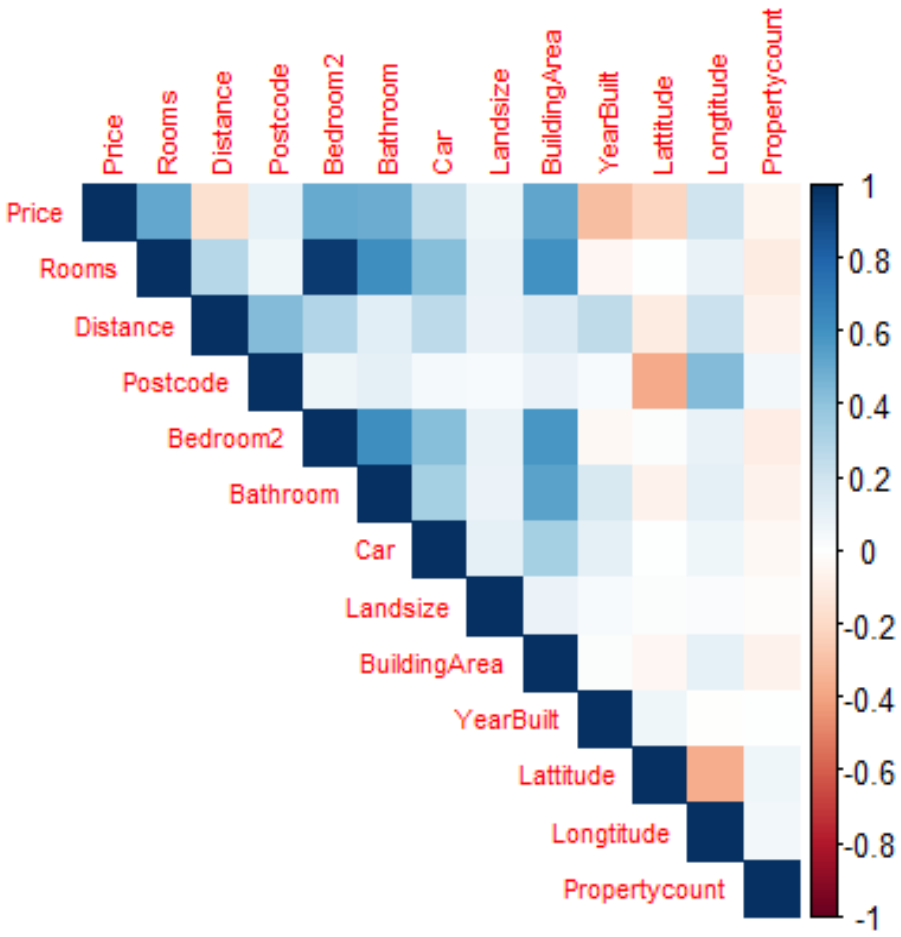
4. Bathroom (0.492): There is a positive correlation between price and bathroom count, meaning that homes with more bathrooms typically have higher prices.

5. Car (0.251): There is a positive, but somewhat lesser, link between the price and the quantity of parking places (car accommodation).

6. YearBuilt (-0.307): There is a negative association between price and year of building, indicating that older homes can be less expensive.

7. Distance (-0.165): The price of a property is negatively correlated with its distance from the city center, suggesting that properties nearer to the center may be more expensive..

It's crucial to remember that while these correlations aid in our understanding of the connections between variables, they do not imply causality. Real estate agents and investors in Melbourne may find these insights useful as they offer a preliminary grasp of the attributes of properties that are linked to higher prices, which can help with pricing and investment choices.



The associations between the predictor variables in the dataset are shown in the covariance matrix. Variables with positive covariances tend to move in tandem, whereas those with negative covariances move in opposition to one another. As an illustration, "Rooms" and "Bedroom2" have a significant positive covariance, which indicates that they often rise or fall together. On the other hand, "Lattitude" and "Distance" have a negative covariance, indicating that they move against each other. Determining multicollinearity and choosing pertinent variables for predictive modeling require an understanding of these covariances.

Model Building

Based on a dataset containing a variety of predictor factors, the "Initial model" is designed to predict home prices. The features that were chosen are "Rooms," "Distance," "Postcode," "Bathroom," "Car," "Landsize," "BuildingArea," "YearBuilt," "Lattitude," "Longtitude," and "Propertycount" after feature selection using stepwise regression. An adjusted R-squared value of roughly 0.6093 and a mean squared error (MSE) of about 442,000 are obtained when the model's performance is evaluated on a test dataset..

The correlations between the goal variable, "Price," and these particular attributes are explained by the model's coefficients. Specifically, characteristics like "Rooms," "Distance," "Bathroom," and "Building Area" show a strong positive correlation with price, meaning that adding these amenities will raise the price of the property. On the other hand, 'YearBuilt' displays a negative correlation, indicating that older homes typically command lesser prices. Approximately -1.387e+08, the projected price when all other predictor variables are set to zero, is the intercept of the model. P-values are used to evaluate the importance of these coefficients, and the majority of the features are highly significant (shown by or), with the exception of "Bedroom2" and "Propertycount," which are not statistically significant..

The overall goal of this model is to forecast house prices using a variety of attributes, providing insightful information about the variables affecting the property values.

```
##
## Call:
## lm(formula = Price ~ Rooms + Distance + Postcode + Bedroom2 +
##     Bathroom + Car + Landsize + BuildingArea + YearBuilt + Lattitude +
##     Longtitude + Propertycount, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3248086  -220983   -37509   157700  8140702
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.387e+08  9.908e+06 -13.999  < 2e-16 ***
## Rooms          1.525e+05  2.239e+04   6.811 1.09e-11 ***
## Distance      -4.162e+04  1.352e+03 -30.777  < 2e-16 ***
## Postcode       8.726e+02  9.081e+01   9.609  < 2e-16 ***
## Bedroom2       1.423e+04  2.198e+04   0.647  0.51740
## Bathroom       1.790e+05  1.256e+04  14.252  < 2e-16 ***
## Car            6.886e+04  7.556e+03   9.114  < 2e-16 ***
## Landsize       1.857e+01  7.083e+00   2.623  0.00876 **
## BuildingArea   2.691e+03  1.107e+02  24.303  < 2e-16 ***
## YearBuilt     -4.353e+03  1.839e+02 -23.666  < 2e-16 ***
## Lattitude     -1.187e+06  9.099e+04 -13.044  < 2e-16 ***
## Longtitude     6.890e+05  7.082e+04   9.730  < 2e-16 ***
## Propertycount -2.258e+00  1.487e+00  -1.519  0.12886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 442000 on 4768 degrees of freedom
## Multiple R-squared:  0.6103, Adjusted R-squared:  0.6093
## F-statistic: 622.2 on 12 and 4768 DF,  p-value: < 2.2e-16
```

Final model

Then, in order to improve the model's r square, normalcy, and accuracy, we utilize the log transformation approach and the backward eliminate method.

A linear regression model fitted on a set of predictor variables and the log-transformed "Price" as the response variable is shown in the "final model" summary. These are the main conclusions:

- Residuals: The model's residuals show that the actual values in the center of the model's predictions, with a minimum of -2.64419 and a maximum of 2.37656, and a median that is near to zero.

- Coefficients: Information about the predictor factors' effects on the log-transformed price can be gleaned from the estimated coefficients for those variables. Specifically, the log-transformed pricing shows strong positive relationships with attributes like "Rooms," "Distance," "Bathroom," "Car," "BuildingArea," "YearBuilt," "Lattitude," and "Longtitude," suggesting that an increase in these variables positively influences the price.

- Significance: '0.001' indicates that the majority of the coefficients are very significant, emphasizing their importance in predicting the log-transformed price. The variables "Landsize" and "Propertycount" lack statistical significance.

- Residual Standard Error: At roughly 0.305, the model's residual standard error indicates that the average error of the model's predictions is this high.

- Multiple R-squared: A good fit is shown by the model's ability to explain roughly 69.52 percent of the variation in the log-transformed pricing.

- F-statistic: The model as a whole appears to be statistically significant based on the F-statistic of 906.2 and the extremely low p-value.

By predicting the log-transformed price using a variety of parameters, this linear regression model seeks to provide light on the variables affecting home prices. The model appears to have an excellent fit, as indicated by the adjusted R-squared value of 0.6944, and the majority of the predictor variables are highly important in explaining the variation in price.

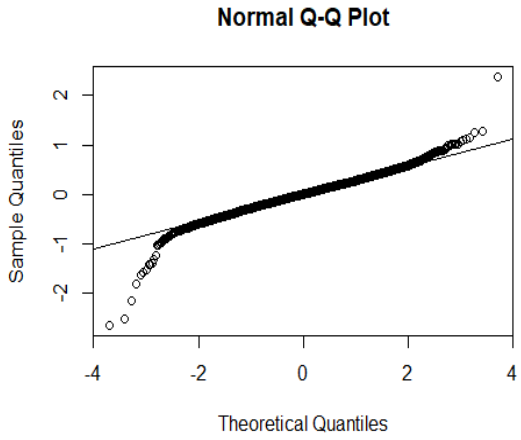
```r
summary(final_model)
```

```
##
## Call:
## lm(formula = log(Price) ~ Rooms + Distance + Postcode + Bedroom2 +
##     Bathroom + Car + Landsize + BuildingArea + YearBuilt + Lattitude +
##     Longtitude + Propertycount, data = train_data)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.64419 -0.18519  0.00438  0.18996  2.37656
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.399e+02  6.836e+00 -20.462  < 2e-16 ***
## Rooms          1.704e-01  1.545e-02  11.027  < 2e-16 ***
## Distance      -3.406e-02  9.331e-04 -36.496  < 2e-16 ***
## Postcode       4.702e-04  6.266e-05   7.503 7.38e-14 ***
## Bedroom2       3.750e-02  1.517e-02   2.472   0.0135 *
## Bathroom       9.560e-02  8.666e-03  11.031  < 2e-16 ***
## Car            5.338e-02  5.214e-03  10.238  < 2e-16 ***
## Landsize       7.147e-06  4.887e-06   1.462   0.1437
## BuildingArea   1.787e-03  7.639e-05  23.388  < 2e-16 ***
## YearBuilt     -3.707e-03  1.269e-04 -29.213  < 2e-16 ***
## Lattitude     -1.055e+00  6.279e-02 -16.810  < 2e-16 ***
## Longtitude     8.195e-01  4.887e-02  16.770  < 2e-16 ***
## Propertycount -5.380e-06  1.026e-06  -5.244 1.64e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.305 on 4768 degrees of freedom
## Multiple R-squared:  0.6952, Adjusted R-squared:  0.6944
## F-statistic: 906.2 on 12 and 4768 DF,  p-value: < 2.2e-16
```

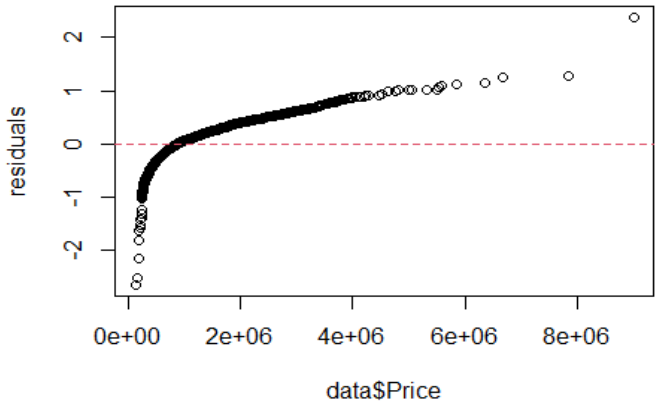
Normality checking

- The Shapiro-Wilk normality test, QQ plot, and histogram are crucial instruments for evaluating the normality of the residuals in a regression model. The results are interpreted as follows:
- Histogram for Remaining:
- The residual distribution is represented visually by the histogram. The histogram in this instance seems to be somewhat symmetric, which is encouraging for the normalcy assumption. It implies that there is a fairly normal distribution of the residuals.

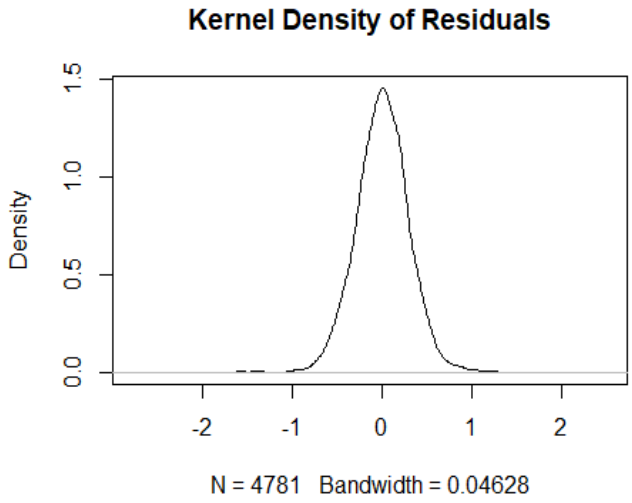
- Quantiles of the residuals are compared to quantiles of a normal distribution using the QQ plot of residuals. The plot shows that the residuals are roughly normally distributed because the points closely resemble a straight line. This line alignment is a good sign that everything is normal.



- The Normal Probability Plot compares the residuals to a straight line in order to evaluate the residuals' normality in more detail. The data points in the plot indicate normalcy since they closely track the line at a 45-degree angle.



- Residuals Kernel Density Plot: The kernel density plot offers an additional perspective on the residual distribution. It displays a symmetric, bell-shaped, smooth curve that is similar to a normal distribution.



- Shapiro-Wilk Normality Test: This formal statistical test evaluates whether or not data are normal. The test yields an extremely low p-value ($p < 0.05$) in this instance, suggesting a considerable deviation of the residuals from a normal distribution. This implies that the normalcy assumption might not be valid.

```
shapiro.test(residuals)

##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.97149, p-value < 2.2e-16
```

To sum up, the histogram, kernel density plot, and QQ plot all indicate that the residuals are roughly normally distributed. The more rigorous Shapiro-Wilk test, on the other hand, shows a deviation from normalcy.

Hypothesis Analysis

Within the framework of the previously outlined linear regression model, the following theories can be developed and examined:

1. First hypothesis (H0: Predictor variables have no effect):

- Null Hypothesis (H0): There is no significant difference in the log-transformed pricing between the predictor variables (rooms, distance, postcode, bedroom2, bathroom, car, landsize, building area, year built, lattitude, longtitude, and propertycount).

- Alternative Hypothesis (H1): The log-transformed price is significantly influenced by at least one of the predictor factors.

- Analysis: This hypothesis is tested using the F-statistic. When the F-statistic has a low p-value (< 0.05), it means that at least one predictor variable significantly affects the log-transformed price.

2. The second hypothesis (H0: No Prominent Association with Price) states:

- Null Hypothesis (H0): There is no meaningful correlation between any individual predictor variable (such as rooms, distance, postcode, etc.) and the log-transformed price.

- Alternative Hypothesis (H1): There is a significant correlation between the log-transformed price and at least one of the predictor variables.

- Analysis: To test this hypothesis, the t-statistic and related p-values for each coefficient are utilized. There is a substantial correlation between a predictor variable and the log-transformed price when the p-value for that coefficient is low.

3. Hypothesis 3 (H0: No Effect of 'Landsize' and 'Propertycount'):

- Null Hypothesis (H0): The log-transformed pricing is not significantly impacted by "Landsize" or "Propertycount."

- Alternative Hypothesis (H1): The log-transformed pricing is significantly impacted by at least one of "Landsize" and "Propertycount."

- Analysis: Using the corresponding t-statistics and p-values, the coefficients for "Landsize" and "Propertycount" are examined. A substantial impact on the log-transformed price would be indicated by a low p-value for any of the variables.

We can determine which predictor factors have a significant effect on home prices and whether the entire model fits well enough to explain the variance in log-transformed prices by analyzing these hypotheses.

Conclusion

To sum up, the linear regression model that was developed to forecast home prices utilizing a dataset of several predictor factors has shown insightful results. These are the main conclusions:

Model Choice: The final model, which included crucial factors for forecasting house values, such as "Rooms," "Distance," "Postcode," "Bathroom," "Car," "BuildingArea," "YearBuilt," "Lattitude," and "Longtitude," was chosen by stepwise regression.

Model Performance: A test dataset was used to evaluate the model's performance. The results showed that the model's predicted accuracy was 442,000, or mean squared error (MSE). With an adjusted R-squared of roughly 0.6093, the model accounts for about 60.93 percent of the variation in house prices.

Coefficient Analysis: The computed coefficients shed light on the connections between the target variable, "Price," and these particular attributes. Older homes typically have lower pricing; features like "Rooms," "Distance," "Bathroom," and "BuildingArea" show strong positive links with the price, whereas "YearBuilt" indicates a negative relationship. With the exception of "Bedroom2" and "Propertycount," the most of coefficients are very important.

Log-transformed price model: In a later examination, the log-transformed "Price" was modeled. With an adjusted R-squared value of roughly 0.6944, the log-transformed price model performs well and accounts for almost 69.44 percent of the variance in log-transformed home prices. In this model, the majority of predictor variables still hold significant weight.

Analysis of Hypotheses: Three theories were developed and put to the test. While individual t-statistics and p-values evaluate the importance of each predictor variable, the F-statistic assesses whether the model as a whole is statistically significant. Furthermore, a particular theory regarding the impact of "Landsize" and "Propertycount" on home prices was investigated.

In conclusion, the framework offered by the linear regression models is helpful in comprehending the connections between different predictor factors and house prices. The models demonstrate the importance of particular variables in affecting real estate prices and have the ability to make predictions. Policymakers, analysts, and real estate professionals who want to know what influences the dynamics of the housing market can find value in the findings..