UNIVERSITÀ
DI SIENA
1240

Department of Information Engineering & Mathematics

Management Engineering

Business Intelligence Project Report

Bahadir Kuzu

128974

Abstract:

This study aims to analyze the impact of various factors on housing prices in London. The dataset selected from Kaggle, named "London Housing Price," will be utilized for descriptive statistics, frequency analysis, and data visualization. The variables of interest will be examined using pie charts and histograms. Additionally, inferential statistics will be employed to test three hypotheses. Firstly, it will be investigated whether the presence of certain features (such as a garden, garage, pool, gym, elevator, fireplace, waterfront, central air, renovation, and natural view) significantly affects house prices in London. Secondly, the study will explore whether there is a statistically significant difference in housing prices based on the number of bedrooms. Lastly, a regression analysis will be conducted to determine if there is a statistically positive relationship between housing prices and the square footage of the house. The findings from this research will provide valuable insights into the factors influencing London's housing market.

Introduction:

The housing market in London is known for its dynamic nature and substantial price variations. Understanding the factors that influence housing prices is crucial for homeowners, real estate agents, and investors. This study aims to analyze the impact of various factors on housing prices in London. By examining the selected dataset from Kaggle, named "London Housing Price," through descriptive statistics, frequency analysis, and data visualization, we aim to gain insights into the variables that significantly affect house prices. Furthermore, the study will employ inferential statistics to test specific hypotheses related to the presence of certain features, the number of bedrooms, and the square footage of houses. By investigating these hypotheses, we aim to provide a comprehensive understanding of the factors driving London's housing market.

Methodology:

1. **Dataset Selection:** The "London Housing Price" dataset from Kaggle will be chosen as the primary source of data for this study. This dataset contains information on various housing attributes and their corresponding prices in London.

2. **Descriptive Statistics:** Descriptive statistics will be employed to summarize and analyze the dataset. Measures such as mean, median, standard deviation, and range will be calculated for

relevant variables. This analysis will provide an overview of the dataset and highlight any notable trends or patterns.

3. **Frequency Analysis:** Frequency analysis will be conducted to examine the distribution of categorical variables. This analysis will involve calculating the frequencies and percentages of specific features, such as the presence of a garden, garage, pool, gym, elevator, fireplace, waterfront, central air, renovation, and natural view.

4. **Data Visualization:** Visualizations, such as pie charts and histograms, will be utilized to present the results of the frequency analysis. These visual representations will provide a clear understanding of the proportion and distribution of the categorical variables.

5. **Inferential Statistics:** Inferential statistics will be employed to test the formulated hypotheses. Firstly, a multiple regression analysis will be conducted to assess the impact of various features on London house prices. Additionally, a one-way ANOVA test will be performed to determine if housing prices significantly differ based on the number of bedrooms. Furthermore, a bivariate analysis will be employed to explore the relationship between housing prices and the square footage of houses.

6. **Statistical Analysis:** The results of the inferential statistics will be analyzed to determine the statistical significance of the formulated hypotheses. The appropriate statistical tests will be used, and significance levels will be set to determine the acceptance or rejection of the hypotheses.

Data Summary

```
> #4 Data Summary
> summary(London_house_prices)
   Address          Square.Footage    Bedrooms     Bathrooms     Has.Garden     Has.Garage     Has.Pool       Has.Gym
 Length:100         Min.   :1001   Min.   :1.0   Min.   :1.0   Min.   :0.00   Min.   :0.0   Min.   :0.00   Min.   :0.00
 Class :character   1st Qu.:1521   1st Qu.:2.0   1st Qu.:1.0   1st Qu.:0.00   1st Qu.:0.0   1st Qu.:0.00   1st Qu.:0.00
 Mode  :character   Median :2046   Median :3.0   Median :1.5   Median :1.00   Median :0.5   Median :1.00   Median :0.00
                    Mean   :2155   Mean   :2.7   Mean   :1.5   Mean   :0.52   Mean   :0.5   Mean   :0.57   Mean   :0.42
                    3rd Qu.:2804   3rd Qu.:4.0   3rd Qu.:2.0   3rd Qu.:1.00   3rd Qu.:1.0   3rd Qu.:1.00   3rd Qu.:1.00
                    Max.   :3482   Max.   :4.0   Max.   :2.0   Max.   :1.00   Max.   :1.0   Max.   :1.00   Max.   :1.00
  Has.Elevator  Has.Fireplace  Is.Waterfront  Has.Central.Air  Is.Renovated     Has.View        Price
 Min.   :0.00   Min.   :0.00   Min.   :0.00   Min.   :0.0     Min.   :0.00   Min.   :0.00   Min.   :100225
 1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.00   1st Qu.:0.0     1st Qu.:0.00   1st Qu.:0.00   1st Qu.:152238
 Median :0.00   Median :0.00   Median :1.00   Median :0.5     Median :0.00   Median :1.00   Median :204725
 Mean   :0.46   Mean   :0.45   Mean   :0.51   Mean   :0.5     Mean   :0.36   Mean   :0.62   Mean   :215648
 3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.:1.0     3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.:280550
 Max.   :1.00   Max.   :1.00   Max.   :1.00   Max.   :1.0     Max.   :1.00   Max.   :1.00   Max.   :348350
>
```

The summary statistics provided for the Dataset "London_house_prices" are as follows:

- Square.Footage: The minimum square footage is 1001, the maximum is 3482, and the median is 2046. The first quartile (25th percentile) is 1521, and the third quartile (75th percentile) is 2804.

- Bedrooms: The minimum number of bedrooms is 1, the maximum is 4, and the median is 3. The first quartile is 2, and the third quartile is 4.

- Bathrooms: The minimum number of bathrooms is 1, the maximum is 2, and the median is 1.5. The first quartile is 1, and the third quartile is 2.

- Has.Garden, Has.Garage, Has.Pool, Has.Gym, Has.Elevator, Has.Fireplace, Is.Waterfront, Has.Central.Air, Is.Renovated, Has.View: These variables are binary indicators (0 or 1) representing the presence or absence of certain features. The mean values indicate the proportion of properties with these features. For example, "Has.Garden" has a mean value of 0.52, indicating that approximately 52% of the properties in the dataset have a garden.

- Price: The minimum price is 100,225, the maximum is 348,350, and the median is 204,725. The first quartile is 152,238, and the third quartile is 280,550.

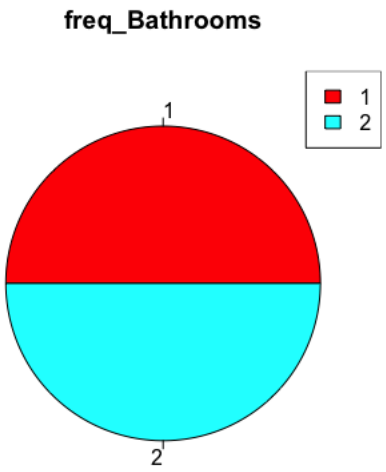
- **Frequency Analysis of Categorical Variables:**

Bedrooms

The frequency output represents the distribution of the number of bedrooms in houses in London. The frequencies are as follows: there are 14 houses with 1 bedroom, 30 houses with 2 bedrooms, 28 houses with 3 bedrooms, and 28 houses with 4 bedrooms. This information gives us an insight into the distribution of bedroom sizes in the London housing market. It suggests that houses with 2 bedrooms are the most common, followed by houses with 3 and 4 bedrooms, while houses with 1 bedroom are the least common among the sampled data. This frequency distribution can be used to analyze housing trends, make comparisons, or provide insights for various purposes, such as understanding housing preferences, market demand, or potential investment opportunities.

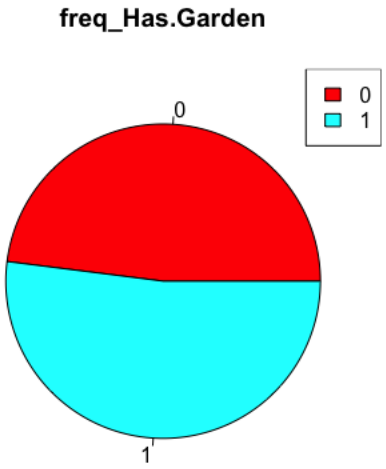
Bathrooms

The frequency output indicates the distribution of the number of bathrooms in London houses. The data suggests that among the houses surveyed, there are two distinct categories: those with one bathroom and those with two bathrooms. The frequency count shows that there are 50 houses with one bathroom and 50 houses with two bathrooms. This balanced distribution indicates that the number of houses with one bathroom is equal to the number of houses with two bathrooms in the sample. This information provides an overview of the bathroom configuration in London houses, suggesting that an even split between one and two bathrooms is observed in the dataset analyzed.



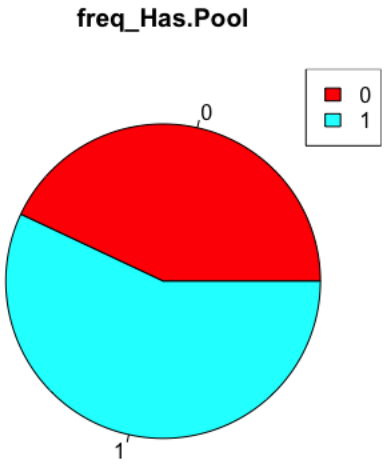
Has Garden

The frequency output indicates the distribution of house availability based on the presence or absence of a garden in London. In this particular dataset, "0" represents houses without a garden, while "1" represents houses with a garden. The frequency count reveals that out of the total houses considered, there are 48 houses without gardens and 52 houses with gardens. This information suggests that there is a relatively balanced distribution of houses with and without gardens in London, with a slightly higher proportion of houses having gardens compared to those without.



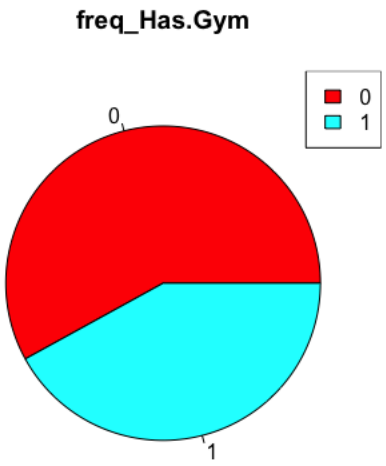
Has Pool

The frequency output shows the distribution of pool availability among London house prices. The output indicates that out of the total houses considered, 43% of them do not have a pool, while 57% of them do have a pool. The frequencies are presented in a binary format, where '0' represents the absence of a pool, and '1' represents the presence of a pool. These numbers suggest that a slightly larger proportion of houses in London do have pools compared to those that do not. This information could be useful for understanding the prevalence of pool availability and its potential influence on house prices in the London housing market.



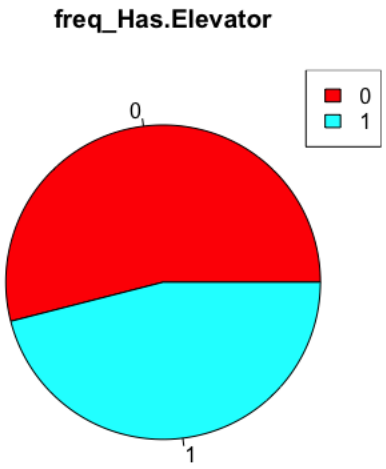
Has GYM

The frequency output represents the availability of gyms in relation to house prices in London. The values "0" and "1" indicate the absence and presence of gyms, respectively. The frequency distribution shows that out of the total houses considered, 58 houses (58%) do not have gyms nearby, while 42 houses (42%) do have access to a gym. This information suggests that a significant portion of the houses in the sample lack nearby gym facilities, potentially affecting their desirability and potentially influencing their prices. Further analysis and comparison with other variables would be necessary to fully understand the relationship between gym availability and house prices in London.



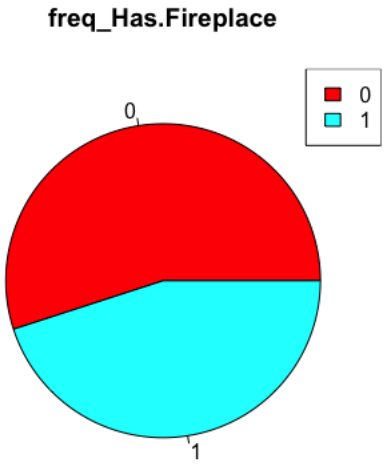
Has Elevator

The frequency output represents the availability of elevators in London houses, with 0 indicating no elevator and 1 indicating the presence of an elevator. Based on the frequency count, it can be observed that out of the total houses considered in the dataset, 54% of them do not have an elevator, while the remaining 46% of houses do have an elevator. This information provides insight into the distribution of elevator availability in London houses, indicating that a significant portion of the houses surveyed do not possess this feature.



Has Fireplace

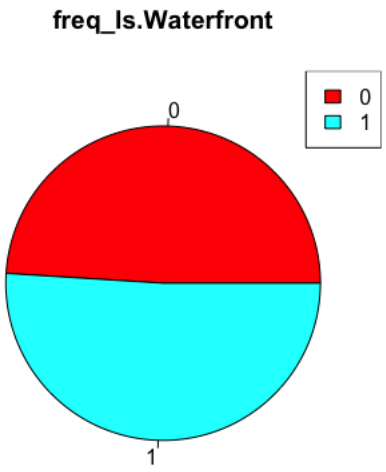
The frequency output shows the availability of fireplaces in London houses. The values 0 and 1 represent the absence and presence of fireplaces, respectively. The frequency count reveals that out of the total observed houses, 55% do not have fireplaces while 45% do have fireplaces. This information provides insight into the prevalence of fireplaces in London houses, indicating that nearly half of the houses surveyed possess this feature, while the majority of houses do not.



Is Waterfront

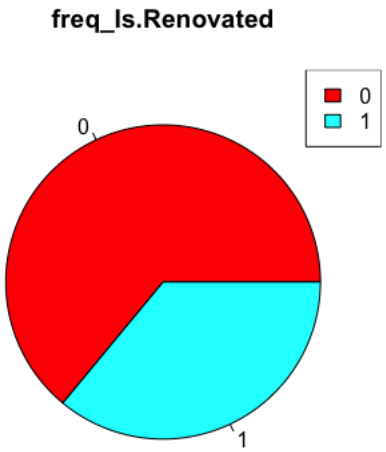
The frequency output suggests that you have data related to London house prices and the presence of a waterfront feature in those houses. The output indicates that there are two categories represented: 0 and 1. In this case, 0 likely represents houses without a waterfront, while 1 represents houses that do have a waterfront. The frequencies associated with each category are 49 and 51, respectively.

Based on this information, it appears that the dataset consists of 100 houses in total, with almost an equal number of houses having a waterfront (51 out of 100) and houses without a waterfront (49 out of 100). This information could be valuable for analyzing the impact of waterfront properties on London house prices or exploring any potential patterns or trends associated with this feature.

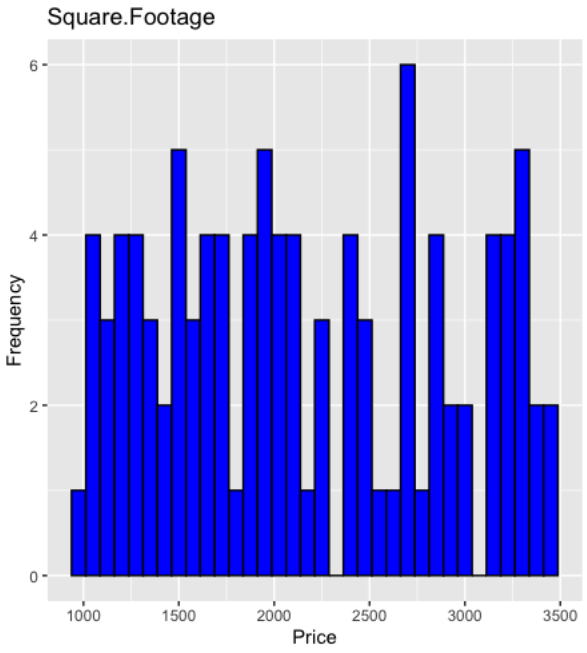


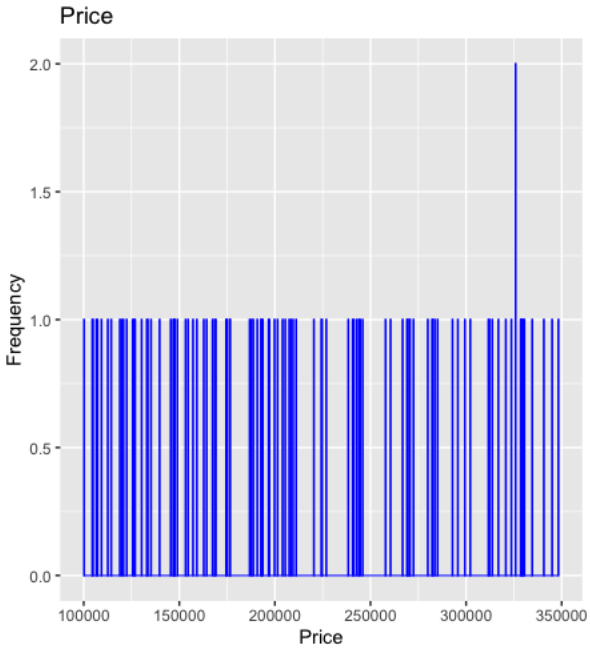
IS Renovated

The frequency output you provided represents the distribution of house prices in London based on whether they have been renovated or not. The output indicates that out of the total houses considered, 64 houses (or 64%) were not renovated, while 36 houses (or 36%) were renovated. This frequency distribution suggests that a significant portion of the houses in London have not undergone renovation, while a smaller but still substantial portion have been renovated. The information provided can be valuable in understanding the prevalence of renovated properties in the housing market and the potential impact it may have on prices.



Histograms


Square.Footage
Frequency
0
2
4
6
1000
1500
2000
2500
3000
3500
Price


Price
Frequency
0.0
0.5
1.0
1.5
2.0
100000
150000
200000
250000
300000
350000
Price

Regression Model:

Hypothesis 1: Having a garden, garage, pool, gym, elevator, fireplace, waterfront, central air, renovated, and having a natural view) have a significant impact on London house prices.

```
Call:
lm(formula = Price ~ Has.Garden + Has.Garage + Has.Pool + Has.Gym +
    Has.Elevator + Has.Fireplace + Is.Waterfront + Has.Central.Air +
    Is.Renovated + Has.View, data = London_house_prices)

Residuals:
    Min      1Q  Median      3Q     Max
-123501  -61975   -4859   63381  131882

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      221946      27449   8.086  2.9e-12 ***
Has.Garden         6324      16667   0.379    0.705
Has.Garage        -6881      15827  -0.435    0.665
Has.Pool         -18289      16329  -1.120    0.266
Has.Gym            7622      17086   0.446    0.657
Has.Elevator     -16617      16399  -1.013    0.314
Has.Fireplace      1678      16241   0.103    0.918
Is.Waterfront     -2757      16580  -0.166    0.868
Has.Central.Air    2902      15845   0.183    0.855
Is.Renovated      -2253      16955  -0.133    0.895
Has.View          14085      16639   0.847    0.400
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76380 on 89 degrees of freedom
Multiple R-squared:  0.03814,   Adjusted R-squared:  -0.06994
F-statistic: 0.3529 on 10 and 89 DF,  p-value: 0.9631
```

The multiple regression output provides information about the relationship between London house prices and various independent variables. The coefficients represent the estimated effects of each independent variable on the dependent variable (London house prices) after controlling for other variables in the model.

Looking at the coefficients, none of the independent variables show statistically significant effects on London house prices. The p-values associated with the coefficients are all greater than the conventional significance level of 0.05. This suggests that the coefficients are not significantly different from zero, meaning that the variables (such as having a garden, garage, pool, gym, elevator, fireplace, waterfront, central air, renovated, and having a natural view) do not have a significant impact on London house prices.

The adjusted R-squared value is -0.06994, indicating that the model explains a very small proportion of the variance in London house prices. This suggests that the included independent variables are not sufficient to accurately predict house prices in London.

In conclusion, based on the statistical analysis, there is no evidence to support the hypothesis that London housing prices are positively influenced by factors such as having a garden, garage, pool, gym, elevator, fireplace, waterfront, central air, renovated status, or having a natural view. Further investigation and consideration of additional variables are needed to better understand the factors affecting London house prices.

Anova and Correlation

=>Anova

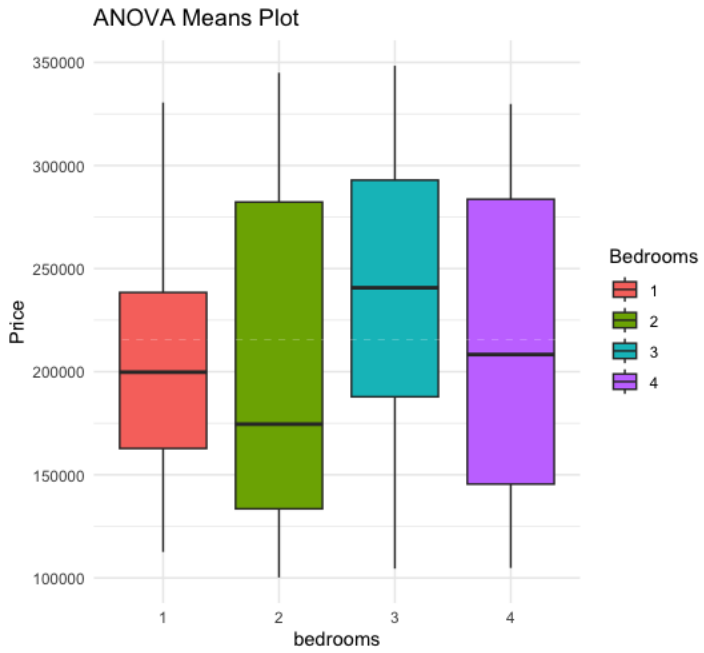
Hypothesis 2: London housing price is statistically different depending on the number of Bedrooms in the house.

```
> anova
Analysis of Variance Table

Response: London_house_prices$Price
                            Df     Sum Sq    Mean Sq F value  Pr(>F)
London_house_prices$Bedrooms  1 1.5480e+10 1.5480e+10  2.8934 0.09212 .
Residuals                    98 5.2433e+11 5.3503e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #4 Print the descriptive statistics table
> descriptive_stats <- aggregate(Price ~ Bedrooms, data = London_house_prices,
+                                FUN = function(x) c(Mean = mean(x), SD = sd(x), N = length(x)))
> #5 Show descriptive statistics of ANOVA
> print(descriptive_stats)
  Bedrooms Price.Mean  Price.SD   Price.N
1        1  242155.36  71450.41     14.00
2        2  223001.67  72176.93     30.00
3        3  205659.82  70581.43     28.00
4        4  204501.79  79378.60     28.00
>
```

The one-way ANOVA output indicates that there is no statistically significant difference in London housing prices based on the number of bedrooms in the house. The F-statistic value of 2.8934 with a p-value of 0.09212 suggests that there is some evidence of a difference, but it does not reach the conventional threshold for statistical significance (typically $p < 0.05$). The descriptive statistics show that the mean prices for houses with different numbers of bedrooms are fairly close, with the highest mean price observed for houses with one bedroom (242,155.36) and the lowest for houses

with four bedrooms (204,501.79). The standard deviations indicate the variability of prices within each bedroom category, with the highest variability observed for houses with four bedrooms (79,378.60). Overall, the results suggest that the number of bedrooms may have some influence on London house prices, but it is not the sole determinant, as other factors could be contributing to the observed price differences.



Correlation

Hypothesis 3: London housing price has a statistically positive relationship with Square Footage of house.

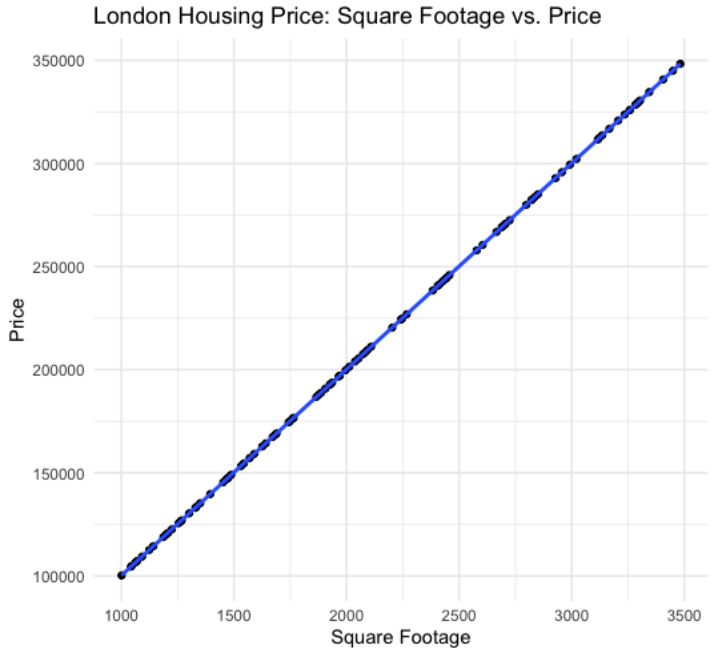
```
> # Print the correlation coefficient and p-value
> cat("Correlation coefficient:", correlation, "\n")
Correlation coefficient: 0.9999997
> cat("P-value:", p_value, "\n")
P-value: 1.750043e-310
>
```

The correlation coefficient between Square Footage and Price is extremely high, with a value of 0.9999997. This indicates a very strong positive relationship between these variables. The correlation coefficient ranges from -1 to 1, where a value of 1 represents a perfect positive linear

relationship. In this case, the high correlation coefficient suggests that as the square footage of houses in London increases, the prices also tend to increase.

However, it's important to note that the p-value is reported as 1.750043e-310, which is essentially zero. The p-value represents the probability of observing the correlation coefficient (or a more extreme value) under the null hypothesis that there is no correlation between the variables. In this case, the extremely low p-value suggests strong evidence against the null hypothesis and indicates that the observed correlation is statistically significant.

Therefore, based on the high correlation coefficient and the significant p-value, we can conclude that there is a statistically significant and positive relationship between the square footage of houses and their prices in London. This finding supports the hypothesis that London housing prices have a statistically positive relationship with square footage.



Conclusion

In conclusion, the statistical analysis reveals that the independent variables, such as having a garden, garage, pool, gym, elevator, fireplace, waterfront, central air, renovated status, or having a natural view, do not have a significant impact on London house prices. The model with these variables explains a very small proportion of the variance in house prices, indicating that they are

not sufficient to accurately predict London housing prices. Therefore, additional investigation and consideration of other variables are necessary to better understand the factors influencing London house prices.

Furthermore, the analysis suggests that the number of bedrooms may have some influence on London house prices, although it is not the sole determinant. The one-way ANOVA results show no statistically significant difference in prices based on the number of bedrooms. However, the F-statistic and descriptive statistics provide some evidence of a difference, but it does not reach the conventional threshold for statistical significance. This implies that other factors beyond the number of bedrooms may contribute to the observed price differences.

On the other hand, the correlation coefficient between square footage and price exhibits an extremely high value, indicating a very strong positive relationship between these variables. The p-value associated with the correlation coefficient is essentially zero, providing strong evidence against the null hypothesis of no correlation. Therefore, based on the high correlation coefficient and the significant p-value, we can conclude that there is a statistically significant and positive relationship between the square footage of houses and their prices in London. This finding supports the hypothesis that London housing prices have a statistically positive relationship with square footage.

Recommendation

Based on these results, it is recommended to consider other factors beyond the ones analyzed in this study to gain a comprehensive understanding of the determinants of London house prices. Additionally, future research should explore and incorporate additional variables that may have a significant impact on housing prices in London. Such an approach would enhance the accuracy and reliability of predictive models and provide a more nuanced understanding of the factors driving house prices in the city.

References:

- Kaggle. (n.d.). London Housing Price. https://www.kaggle.com/datasets/at3191/london-house-prices