STAT501 Project Report

Dataset:

https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction

Introduction

The real estate rental market is diverse and dynamic, influenced by various factors such as property size, location, and furnishing status. Understanding these factors is crucial for both renters and landlords to make informed decisions. This project aims to analyze a comprehensive dataset of residential rental properties, exploring the relationships between rent and various property attributes. By employing statistical analysis techniques, we aim to uncover patterns and provide insights that can aid in predicting rental prices.

Research Questions

1. How much does each numerical variable impact the rent? This question aims to determine which variables, such as area, number of bedrooms, and number of bathrooms, have the most significant influence on rent. Understanding these relationships will help in making predictions or approximations about rent values based on these key factors.

2. What impact does the area of the property have on the rent? We want to explore the relationship between the size of a property (in square feet) and its rental price to assess if the area can be used as a reliable metric for estimating rent. This analysis will help to quantify how changes in property size affect rental costs.

3. Is there a relationship between having an air conditioning system and the furnishing status of a property? This question investigates whether properties with different furnishing statuses (furnished, semi-furnished, unfurnished) are more or less likely to have air conditioning. Understanding this trend can provide insights into how amenities are distributed across different types of properties.

4. Is there a relationship between the number of bathrooms and the number of bedrooms in a property? We aim to understand the correlation between the number of bathrooms and bedrooms. This insight can be valuable for real estate developers and planners to inform design and marketing strategies by understanding how bathroom counts typically scale with bedroom counts.
5. What types of properties typically have a guest room? Is there any relationship between the presence of a guestroom with the number of bedrooms in a property? This analysis can reveal if certain types of properties are more likely to feature a guest room, which may be influenced by factors like property size or design preferences.

6. What is the distribution of rent values across the dataset, and what are its features and ranges? By examining the distribution of rent values, we aim to understand the economic diversity within the rental market. This includes identifying the range, central tendencies, and any significant patterns or outliers in rental prices, providing a comprehensive overview of rent variability.

Motivation

The motivation behind this project is to provide a data-driven understanding of the rental market. With the rental market being a significant component of the housing sector, it is essential to identify the key factors that influence rent prices. This analysis can help landlords set competitive prices and assist renters in finding properties that meet their needs and budget. Additionally, understanding regional preferences for tenant types can provide insights into the socio-economic dynamics of different areas.

Dataset Description

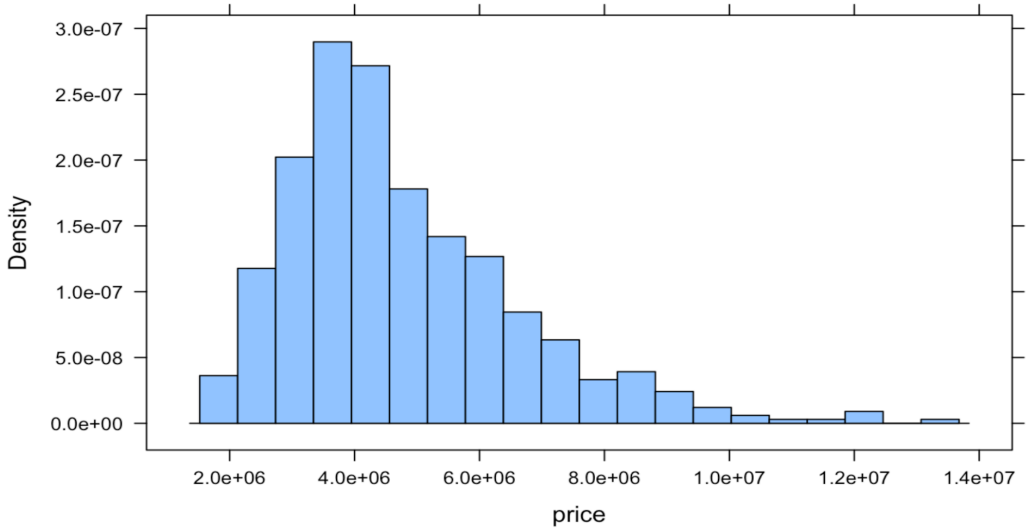
This dataset has been sourced from **Kaggle** and provides a comprehensive collection of information on over **545** available residential properties, including furnished, semi-furnished and unfurnished properties offered for rent. It offers insights into various attributes crucial for renters and landlords alike, such as the number of bedrooms (BHK), rental rates, property size, etc. This dataset serves as a valuable resource for individuals and entities involved in the housing market, offering insights into the diverse range of residential properties available for rent.

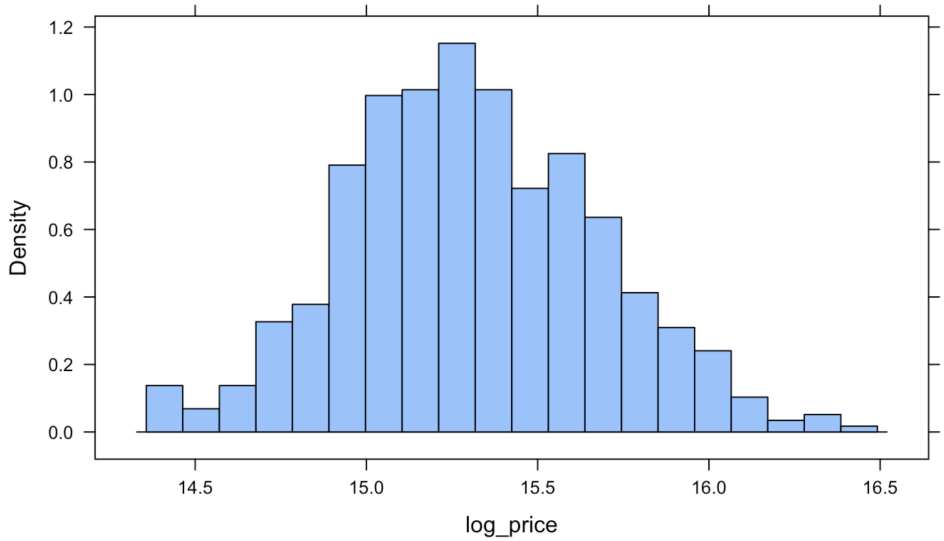
Columns Description:

- **Price:** Rent of the Houses/Apartments/Flats.
- **Area:** Area of the Property in Square Feet.
- **Bedrooms:** Number of Bedrooms
- **Bathrooms:** Number of Bathrooms
- **Stories:** Number of stories in the house.
- **Mainroad:** Whether the house is connected to the main road (Yes/No)
- **Guestroom:** Whether the house has a guest room
- **Basement:** Whether the house has a basement
- **Hotwaterheating:** Whether the house has a hot water heating system
- **Airconditioning:** Whether the house has an air conditioning system

Data Transformation:

We first performed a log transformation on the price column because it was skewed. This helped to stabilize the variance, normalize the distribution, and make the data more suitable for linear regression analysis, leading to better model performance.





Analysis:

1. ***Simple Linear Regression (SLR)***
   SLR is a statistical method that helps us understand the linear relationship between two variables by fitting a line through the data points in a way that minimizes the distance between each data point and the line. This model is helpful because it provides insights into how significant prediction variables affect the outcome variable. In this project, we

conducted a Simple Linear Regression (SLR) analysis to understand the relationship between the area of a house (**Area**) and its price (**Price**).

Assumptions:

1. The relationship between the predictor (independent variable) and the response (dependent variable) must be linear.

2. Residuals from the model should not exhibit any correlation.

Regression Model Summary:

```
Call:
lm(formula = log_price ~ area, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96476 -0.20560  0.00421  0.19609  0.88583

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.483e+01  3.453e-02  429.41   <2e-16 ***
area        9.316e-05  6.179e-06   15.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3128 on 543 degrees of freedom
Multiple R-squared:  0.2951,    Adjusted R-squared:  0.2938
F-statistic: 227.3 on 1 and 543 DF,  p-value: < 2.2e-16
```

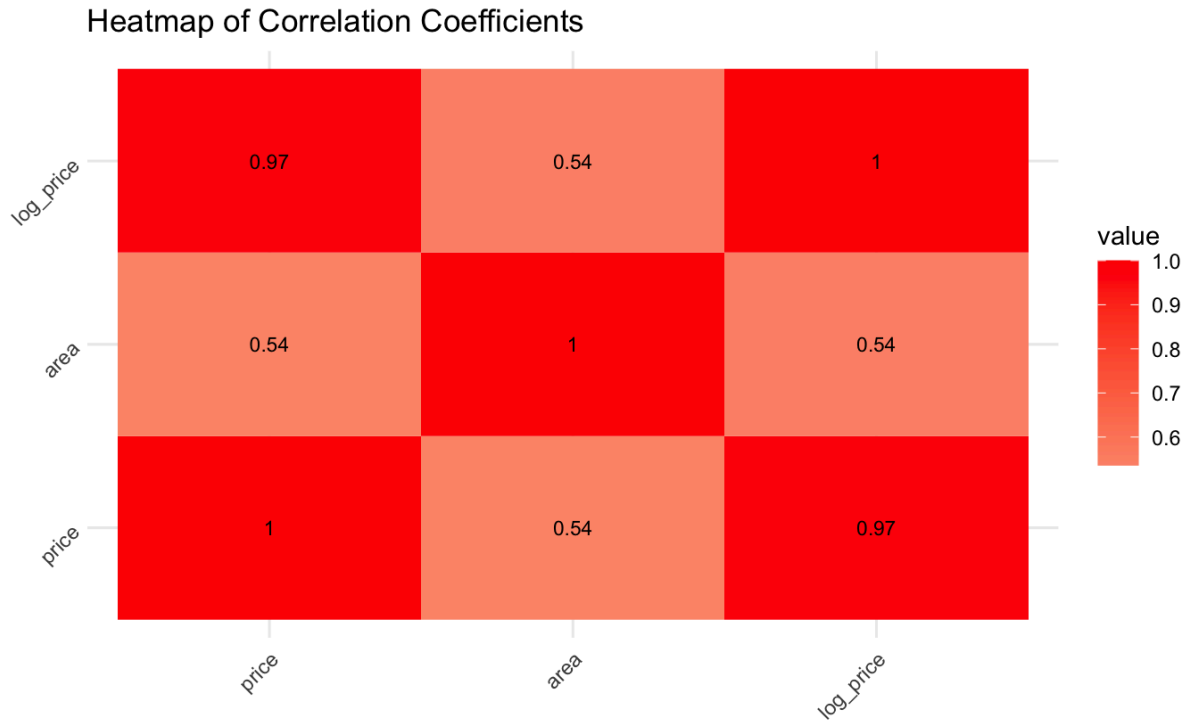
Equation of the Model:
The model can be expressed as
Log_price = 1.483e+01 + (9.316e−05 × Area)

log_price=1.483e+01+(9.316e−05×Area), where the intercept is approximately 14.38, indicating the estimated log price when the area is zero. The slope coefficient, approximately 0.00009316, suggests that with each additional square foot, the log of the house price increases by approximately 0.00009316 units.

Statistical Significance:
Both the intercept and the slope have extremely low p-values (p < 2e-16), signifying a strong linear relationship between area and the log of price. This implies significant effects of area on the price at any conventional significance level.

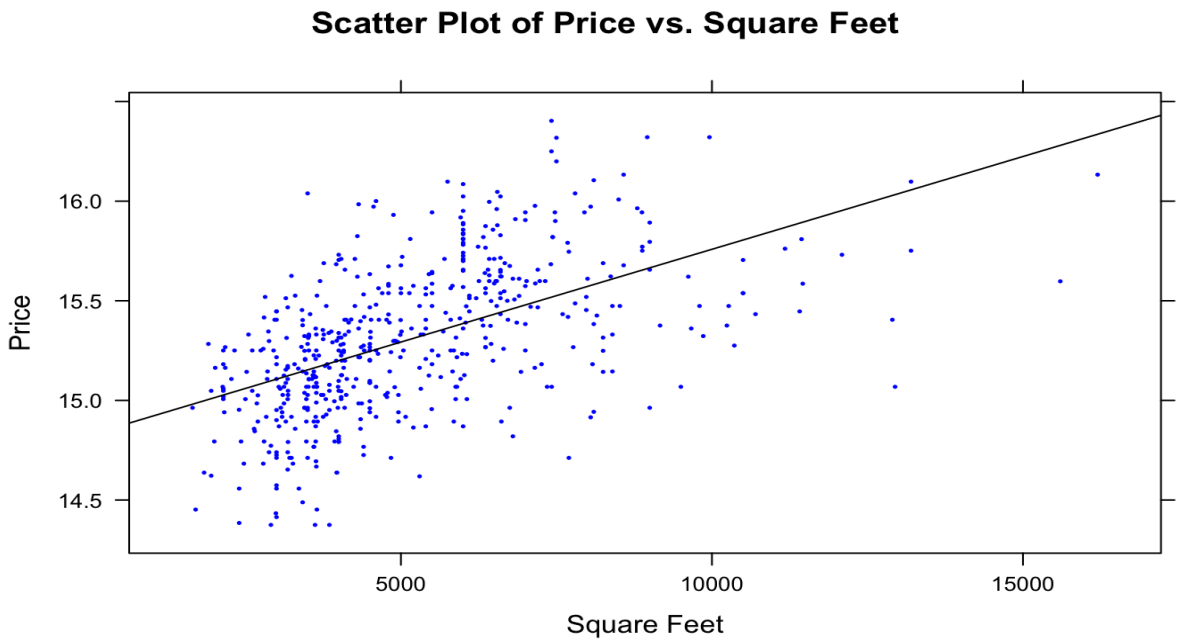
Fit Quality:
The model's R-squared is 0.2951, indicating that about 29.51% of the variability in the log of house prices is explained by the area. While this shows a moderate fit, it suggests that other factors might also be important in predicting house prices.

Heatmap of Correlation Coefficients



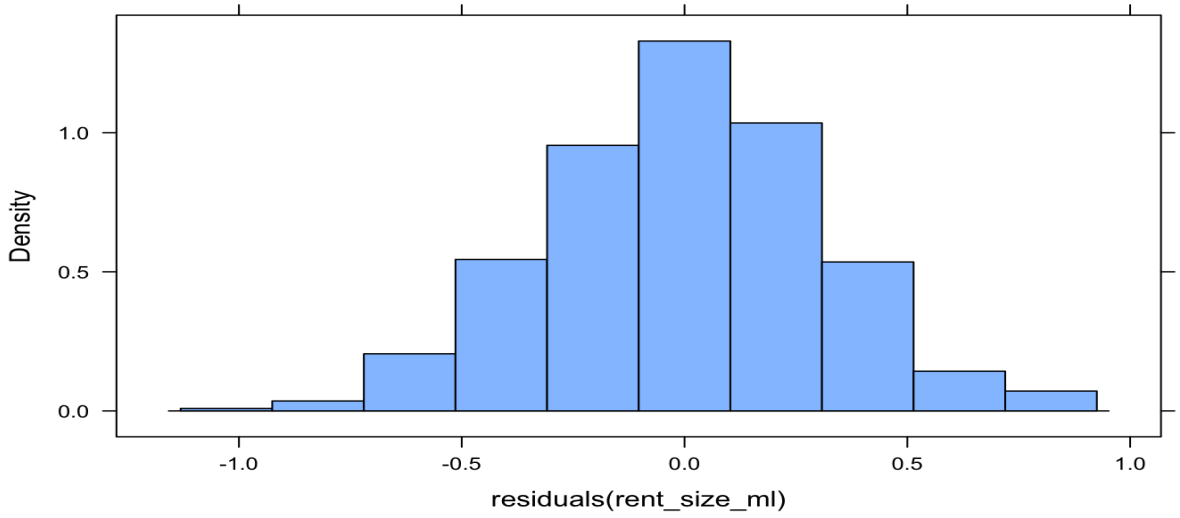
We created a heatmap between the 2 numerical values in the data - price and area as we wanted to verify the strength of the correlation between the two. The plot shows a moderately strong, and positive correlation of 0.54 between Price and Area, supporting the choice of Area as a predictor in the SLR model. The table contained both log_price and price, but we only consider one and discard the other for simplicity.

Scatter Plot of Price vs. Square Feet



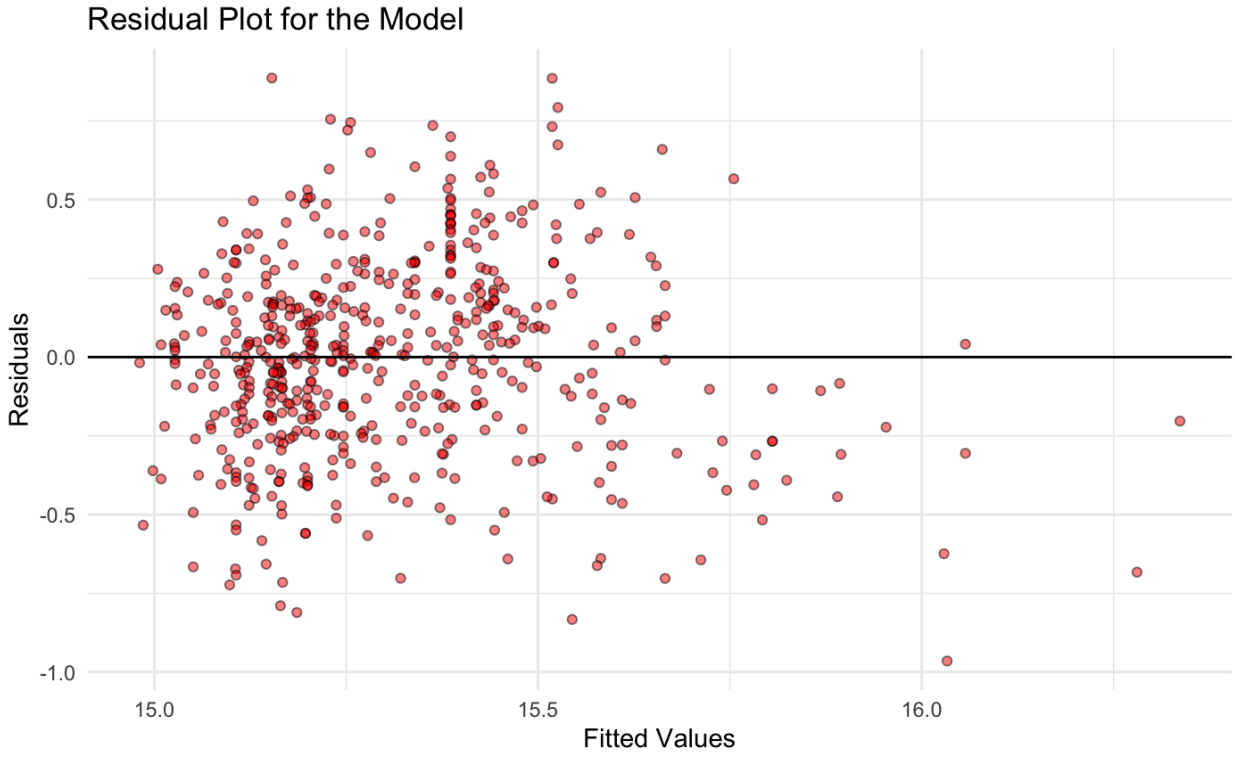
This graph shows a positive linear relationship between Price and Area. The fitted regression line confirms that as the area increases, the price tends to increase as well. However, the spread of data points around the line indicates variability that the model doesn't capture, suggesting other factors may influence prices.There seem to be several outliers, but there seems to be a healthy distribution of the data..

Histogram of Residuals



We also plotted the distribution of residuals which tell the differences between observed and predicted prices. In an ideal scenario, this distribution would be a perfect normal distribution, that is centered around 0. From this graph, we can see an almost perfect normal distribution of the residuals that is distributed around 0- the presence of extreme values suggests outliers which has caused the residuals histogram to skew a little, however it is mostly normal.

Residual Plot for the Model



We want our residuals to not show a pattern but rather be randomly distributed in roughly equal numbers around the line, so that we can say that the expected value of our residuals is zero (to maximize correctness) and assume constant variance. This plot reveals some structure but mostly randomly distributed/scattered residuals which are present in roughly similar numbers above and below the line. We can therefore say that the expected value of our residuals is close to zero.

2. *Chi-square*

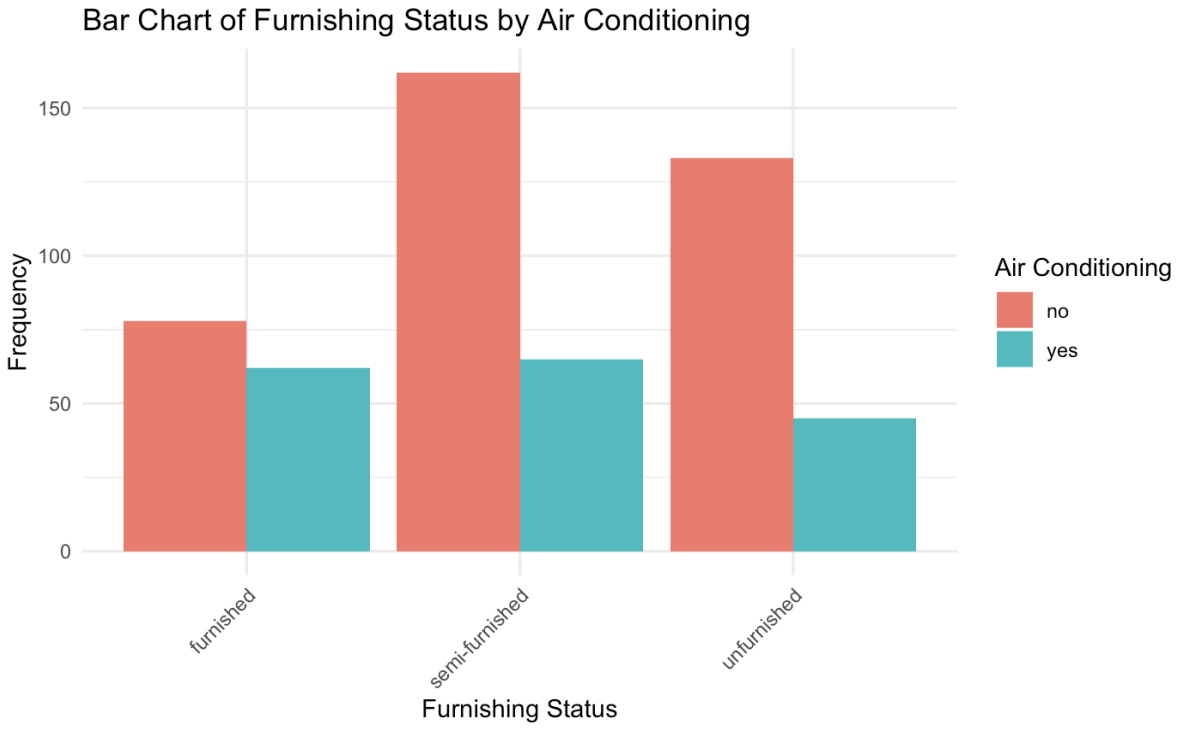
The objective of this analysis was to investigate the relationships between various categorical property attributes.

Assumptions:

- The samples are independent.
- The data are categorical.
- Expected frequencies in each cell of the contingency table are at least 5.

H_0: There is no relationship between the attributes
H_a: There is a relationship between the attributes

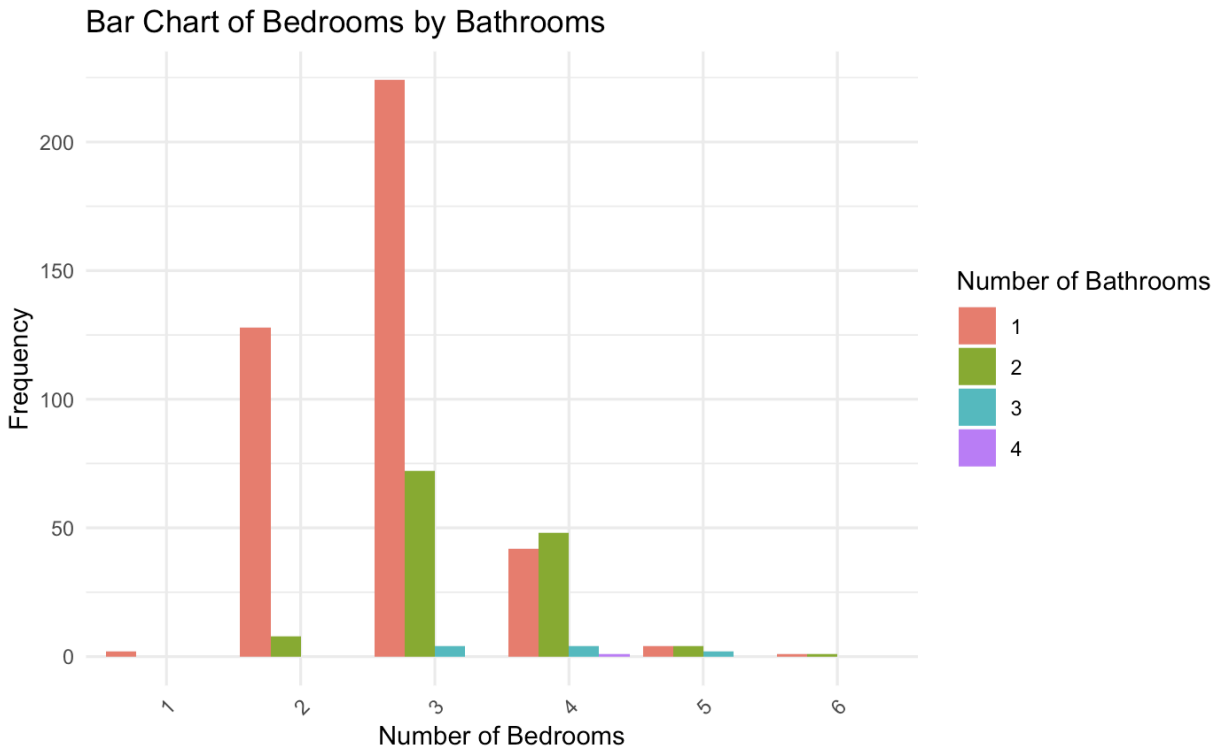
Air Conditioning and Furnishing Status:(Chi-sq value = 14.645, df = 2,p-value = 0.006)
Since p-value<0.05, we reject H_0.
We conclude that there is a statistically significant relationship between air conditioning and furnishing status.
Properties that are semi-furnished are more likely to have air conditioning, while unfurnished properties are less likely to have it. This insight suggests a potential trend where more furnished properties are better equipped with amenities such as air conditioning.




```
        Pearson's Chi-squared test

data:  contingency_table_1
X-squared = 14.645, df = 2, p-value = 0.0006604
```

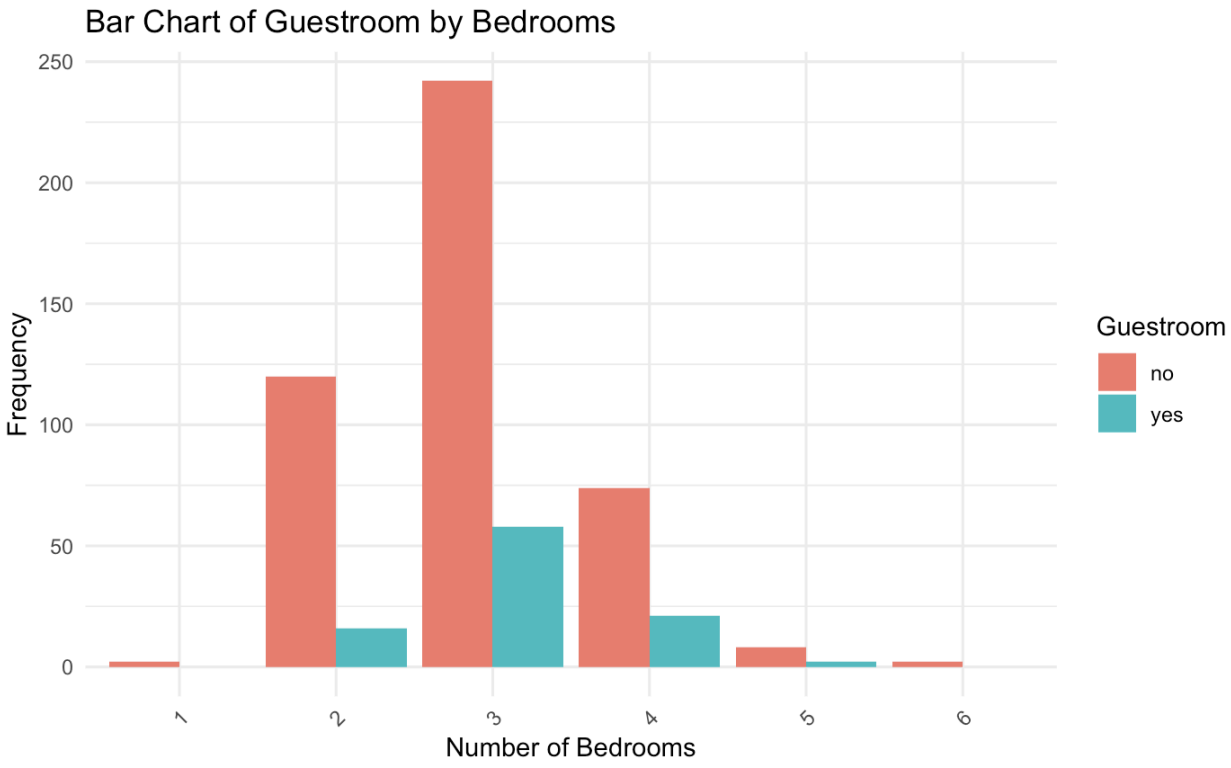
No. of Bathrooms and No. of Bedrooms:(Chi-sq value = 97.201, df = 15, p-value < 0.0001)
Since p-value<0.05, we reject H_0.
We conclude that there is a statistically significant relationship between the number of bathrooms and the number of bedrooms.
Properties with more bedrooms tend to have a higher number of bathrooms. This insight is useful for real estate developers and planners in understanding how bathroom counts typically scale with bedroom counts, which can inform design and marketing strategies.




```
        Pearson's Chi-squared test

data:  contingency_table_2
X-squared = 97.201, df = 15, p-value = 4.415e-14
```

Presence of Guestroom and No. of Bedrooms: (Chi-sq value = 5.971, df = 5, p-value = 0.309)
Since p-value>0.05, we do not reject H_0.
We conclude that there is no statistically significant relationship between the presence of a guest room and the number of bedrooms.
The presence or absence of a guestroom does not vary significantly with the number of bedrooms, as one would assume. This could indicate that other factors, such as lifestyle preferences or specific design choices, play a more critical role in the inclusion of guestrooms.

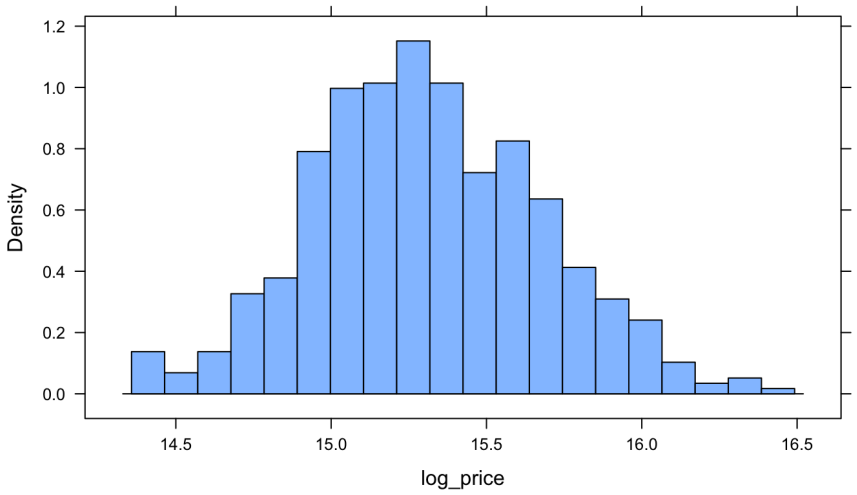



```
        Pearson's Chi-squared test

data:  contingency_table_3
X-squared = 5.9709, df = 5, p-value = 0.3091
```

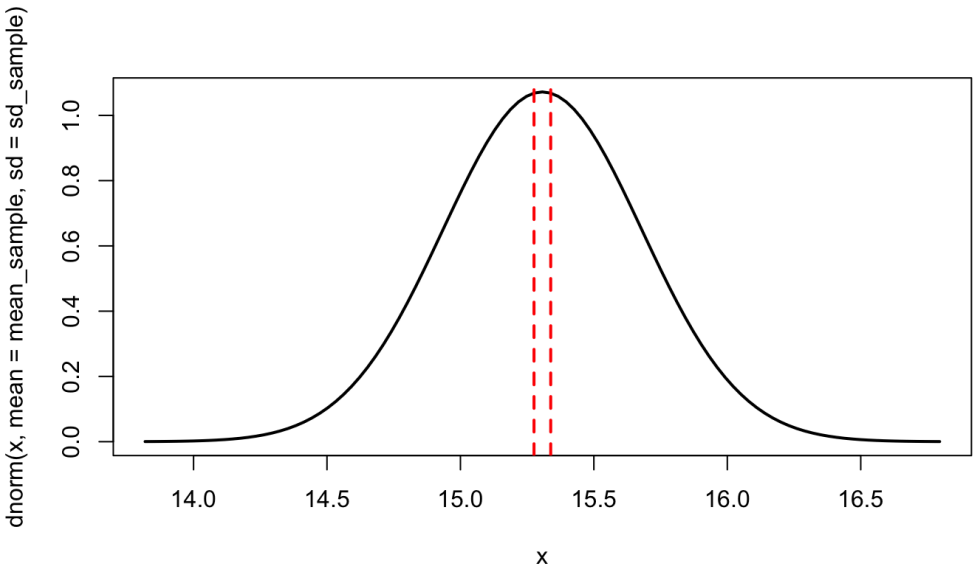
3. Confidence Interval

Here we're examining a sample from the population to infer the average rent for the entire population. This involves conducting a t-test on the rent distribution and subsequently constructing a 95% confidence interval to estimate the mean. However, before proceeding with the t-test, certain prerequisites need to be met. Firstly, the distribution of rent should exhibit normality, ensuring that the data follows a bell-shaped curve. Additionally, the observations should be independent of one another, meaning that the values do not influence each other's occurrence. Upon verifying these conditions, as illustrated in the distribution below where data points cluster around the mean symmetrically, a t-test can be aptly employed to make inferences about the population mean.




```
        One Sample t-test

data:  log_data
t = 0, df = 544, p-value = 1
alternative hypothesis: true mean is not equal to 15.30699
95 percent confidence interval:
 15.27567 15.33830
sample estimates:
mean of x
 15.30699
```



Above the graph shows a confidence interval based on the values of the price.

After calculating the exponents of the confidence intervals.

```{r}
print(c(exp(ci_lower), exp(ci_upper)))
```

```
[1] 4306654 4585006
```

We are 95% confident that the price should lie between 4306654 and 4585006.

Conclusion

This project has provided a comprehensive analysis of the residential rental market using a dataset of over 545 properties. Through our investigation, we identified key factors influencing rental prices, including the area of the property, number of bedrooms and bathrooms, and presence of amenities such as air conditioning. Our Simple Linear Regression model revealed a significant, albeit moderate, correlation between property size and rent, suggesting that while area is a crucial determinant, other factors also play substantial roles.

The chi-square tests highlighted notable relationships between categorical variables, such as the strong association between air conditioning and furnishing status, and between the number of bedrooms and bathrooms. These insights are invaluable for real estate developers and landlords in designing and marketing properties that meet market demand. However, the lack of a significant relationship between the presence of a guestroom and the number of bedrooms suggests that other lifestyle factors may influence the inclusion of such features.

Overall, our analysis underscores the complexity of the rental market, driven by a myriad of interrelated factors. By leveraging statistical methods and data-driven insights, stakeholders can make more informed decisions, from pricing strategies to property development. This study not only enhances our understanding of rental dynamics but also serves as a foundational tool for further research and practical applications in the real estate sector.