Statistical Inference Project

Student Name:
Pouya Haji Mohammadi Gohari

SID:810102113

Date of deadline
Thursday 16th February, 2023

Dept. of Computer Engineering

University of Tehran

Contents

1 Introduction(Number of Results Are Here)

The paper is about house price prediction using machine algorithem.People are very careful when they want to buy a new house with market strategies and their budgets. The objective of this paper is to predict the house prices for non- house holders based on their aspirations and financial provisions. By analyzing different parameters like area of the house, square feet of the house, no of floors in the house etc. This research work has utilized the dataset from Kaggle.. An analysis is performed by applying advanced machine learning regression techniques such as Linear regression, KNN Regression, Random Forest Regression, Decision Tree Regression, Extra Trees Regression etc. to attain the most efficient and least error driven regression technique. From the analysis performed, an observation has been made that Catboost Regression Algorithm has outperformed other algorithms. The model predicts the final output with respect to correlated attributes in the dataset.
The dataset description are as follow:

- Price: The selling price of the house. This is usually the dependent variable that you try to predict in a house pricing model.
- Area: The size of the property in square units (e.g., square feet or square meters).
- Bedrooms: The number of bedrooms in the house.
- Bathrooms: The number of bathrooms in the house.
- Stories: The number of levels or floors the house has.
- Mainroad: A categorical variable indicating whether the house is on the main road. This is typically binary, with 'yes' indicating proximity to the main road.
- Guestroom: Indicates whether the house has a guest room. Again, this is a binary variable with 'yes' or 'no' values.
- Basement: Specifies whether the house has a basement, with 'yes' or 'no' as possible values.
- Hotwaterheating: Indicates if the house is equipped with a hot water heating system. This is a binary variable.
- Airconditioning: A binary variable indicating whether the house has air conditioning.
- Parking: The number of parking spaces available with the property.
- Prefarea: A categorical variable indicating whether the house is in a preferred area, often associated with desirability factors like better schools, views, lower crime rates, etc.
- Furnishingstatus: Indicates the furnishing status of the house, which could be categorical with values such as 'furnished', 'semi-furnished', or 'unfurnished'.

Each of these features can affect the price of a house and are often used in regression analysis to build predictive models for real estate pricing. Categorical variables such as 'mainroad', 'guestroom', 'basement', 'hotwaterheating', 'airconditioning', 'prefarea', and 'furnishingstatus' would typically be converted into numerical form through encoding before they can be used in a regression model.
The work summary:

a. Visualizaton:(8 Different Visualization Types)

- Histogram
- Boxplots
- Scatter plot
- Correlation matrix
- Q-Q plots
- bar plots for categorical features
- Heatmap
- Mosaic

b. Parametric Inferences and estimations:(11 Different Results)

- Shapiro Test For Normality Assumptions
- Transformed Data and Test Shapiro
- t-tests (Two different test)
- Plot distribution of parameters using Bootstrap
- Conduct Percentile Interval For CI
- Fit Normal With MLE[1]
- Fit Gamma Dist With MLE
- Conduct CI for Nomral Parameters(Exact Method)
- Conduct CI for Gamma(Fisher Information with Approximate Method) parameters

c. Hypothesis Testing and Statistical Analysis:(11 Different Results)

- Sign
- Wilcoxon
- CI For Median
- Pearson
- Spearman
- One-Way ANOVA
- Tukey's Method
- The Kruskal-Wallis
- Dunn's Method
- Contingency(2 Different Result)

d. Regression Analysis:(8 Different Results)

[1]Maximum Likelihood Estimation

- 2D Regression Model
- Plot Residuals
- KS test
- Transform Data,Do same procedure
- Bootstrap For Slope
- CI With Percentile Method
- 3D Regression Model
- Encode All Categorical Cols
- Fit Regression Model with All Data

2 Visualization

In this section we are going to visualization of data using several methods in the plots.
From plotting histograms, we will observe and get intiution about the distribution of each feature and detect if they behvae like the distributions we might know or not.At next step plotting scatter plot(pair plot) will illustrate if any variables have correlation between each other or not.

2.1 Histogram of Numerical Features

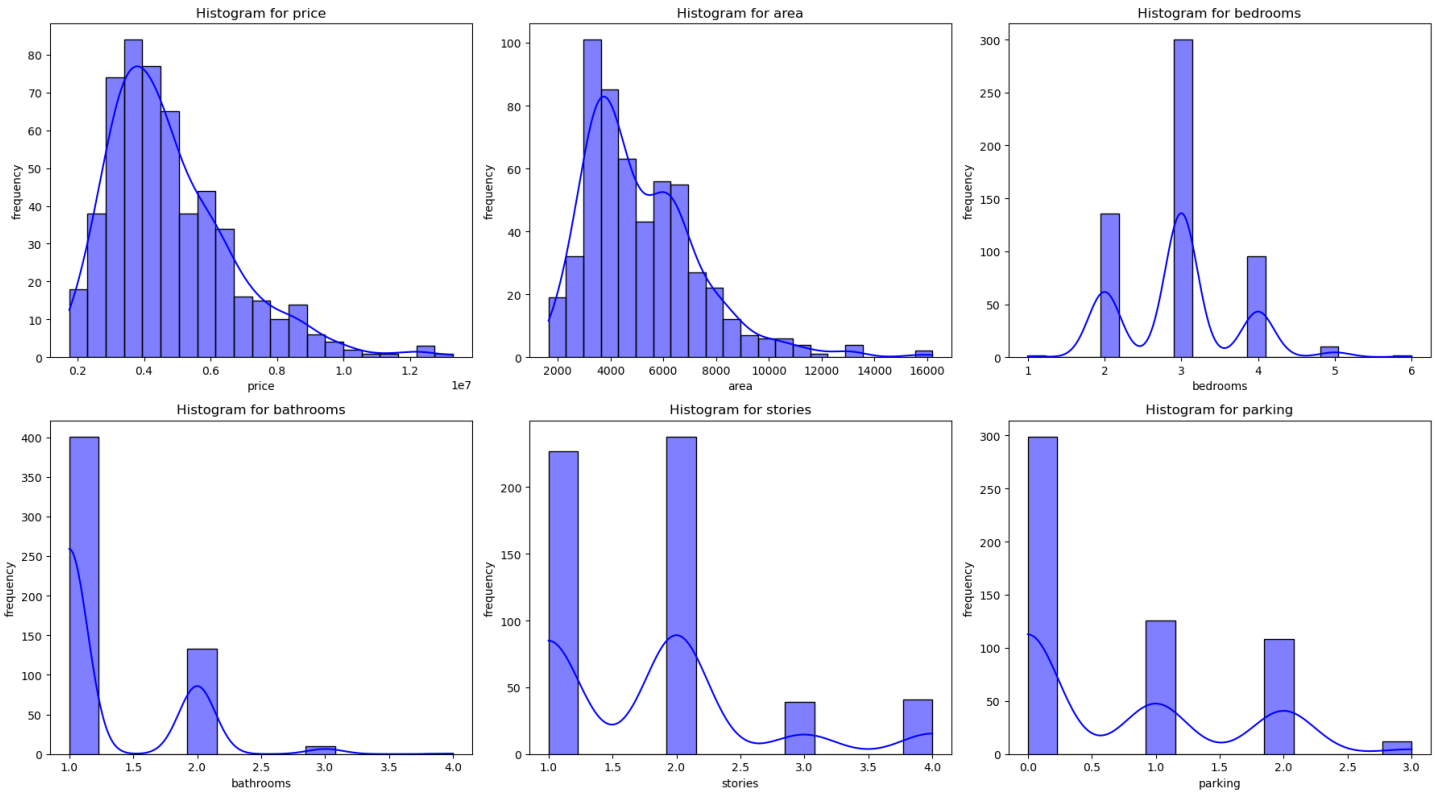

Figure 1: H

As illustrated in figure 1 the price and area features are likely to be a normal distribution but other features seems like are not following any particular distributions.
We can see the mean of price distribtuion is around $0.4 * 10^7$, houses having 3 bedrooms in average, approximately 400 houses out of 545 houses(74%) has single bathrooms, approximately all of them has either 1 or 2 stories and about 60% of the houses has no parking spot.
More detail about the histograms:

- Histogram for price: This histogram shows the distribution of house prices. The distribution appears to be right-skewed, meaning there are a few houses with much higher prices than the rest. Most of the data is concentrated on the lower end of the price spectrum. The skewness suggests that there are outliers in the higher price ranges or that luxury homes with high prices are less common. They are typical characteristics of a log-normal distribution often seen in real estate prices where a few properties are significantly more expensive than the rest.

- Histogram for area: The 'area' distribution also appears to be right-skewed, with most of the houses having a smaller area, while fewer houses have a significantly larger area. This pattern is typical in real estate, where smaller homes are more common than larger ones.This could again suggest a log-normal distribution where most properties have a smaller area, with fewer properties having a significantly larger area.
- Histogram for bedrooms: The 'bedrooms' histogram shows a concentration of data around houses with a specific number of bedrooms (most likely 3, as it is the most common setup for houses). The distribution has a peak, suggesting a mode around that value, and tails off for smaller and larger numbers of bedrooms.
- Histogram for bathrooms: Similar to bedrooms, there is a concentration of houses with a specific number of bathrooms. The data might suggest common configurations such as one or two bathrooms.
- Histogram for stories: The histogram for 'stories' seems to indicate that single-story and two-story homes are most common. Fewer homes have three or more stories.
- Histogram for parking: This histogram suggests that most houses have space for at least one or two cars, with a significant drop in frequency as the parking capacity increases.

2.2 Box plots

Box plots will give us very good information where how data vary, and also gives us a good intiution about outliers.Since some tests are very sensitive to them (eg, correlation test). The box plot are illustrated in figure 2:

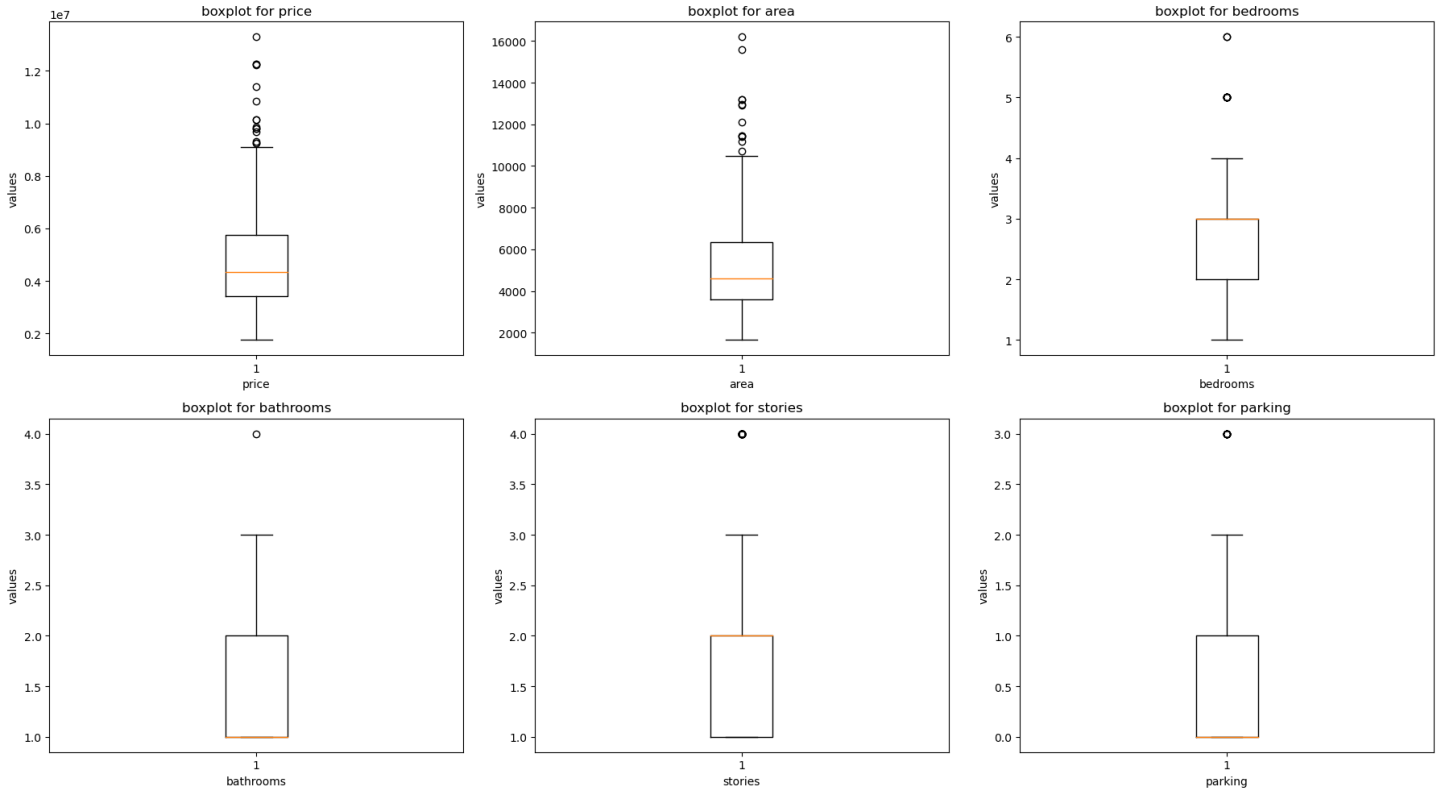

Figure 2: Box plots

As you can see we can calculate the third and first quartiles and see the IQR where can calculated as follows:

$$IQR = Q_3 - Q_1 \tag{1}$$

The second quartile is median of data. As we can see in figure 2, price and area features has so much outliers where we should give an attention to this problem if we are going to conduct a correlation test in future.
More details:

- Boxplot for price:

  I) The median price (indicated by the line inside the box) is below the half-way mark of the y-axis range, suggesting that more than half of the houses are priced below the median value.

  II) The interquartile range (IQR - the box height) is relatively small compared to the full range of the data, indicating that the middle 50% of prices are clustered within a narrower range.

  III) There are several outliers on the higher end (indicated by the separate dots), showing that there are some houses significantly more expensive than the rest.

- Boxplot for area:

I) The median area is somewhat central within the box, which suggests a more symmetrical distribution of values in the middle 50% of the dataset.

II) The IQR is somewhat large, indicating variability in the sizes of the properties.

III) Multiple outliers are present above the upper whisker, indicating some properties with significantly larger areas than the general trend.

- Boxplot for bedrooms:

I) The median number of bedrooms appears to be between 2 and 3, which is typical for many residential houses.

II) The IQR is small, suggesting that most houses have a similar number of bedrooms.

III) There is at least one outlier, indicating a house with a much higher number of bedrooms than usual.

- Boxplot for bathrooms:

I) The median number of bathrooms is around 2, and the IQR is narrow, which means most houses have a similar number of bathrooms.

II) There are outliers, which may represent houses with a significantly larger number of bathrooms.

- Boxplot for stories:

I) The median number of stories is 2, with a small IQR, indicating that most houses are either 1 or 2 stories tall.

II) There is an outlier, suggesting there are a few houses with more stories than is typical.

- Boxplot for parking:

I) The median number of parking spaces is close to 1, with a small IQR, suggesting limited variation in parking space availability among most houses.

II) There is at least one outlier indicating a house with substantially more parking space.

2.3 Scatter Plots and Correlation Matrix

In this section we are going to create pairplots to see if any particular variabales has correlation to each other or not. The pair plot has been provided for you in figure 3:

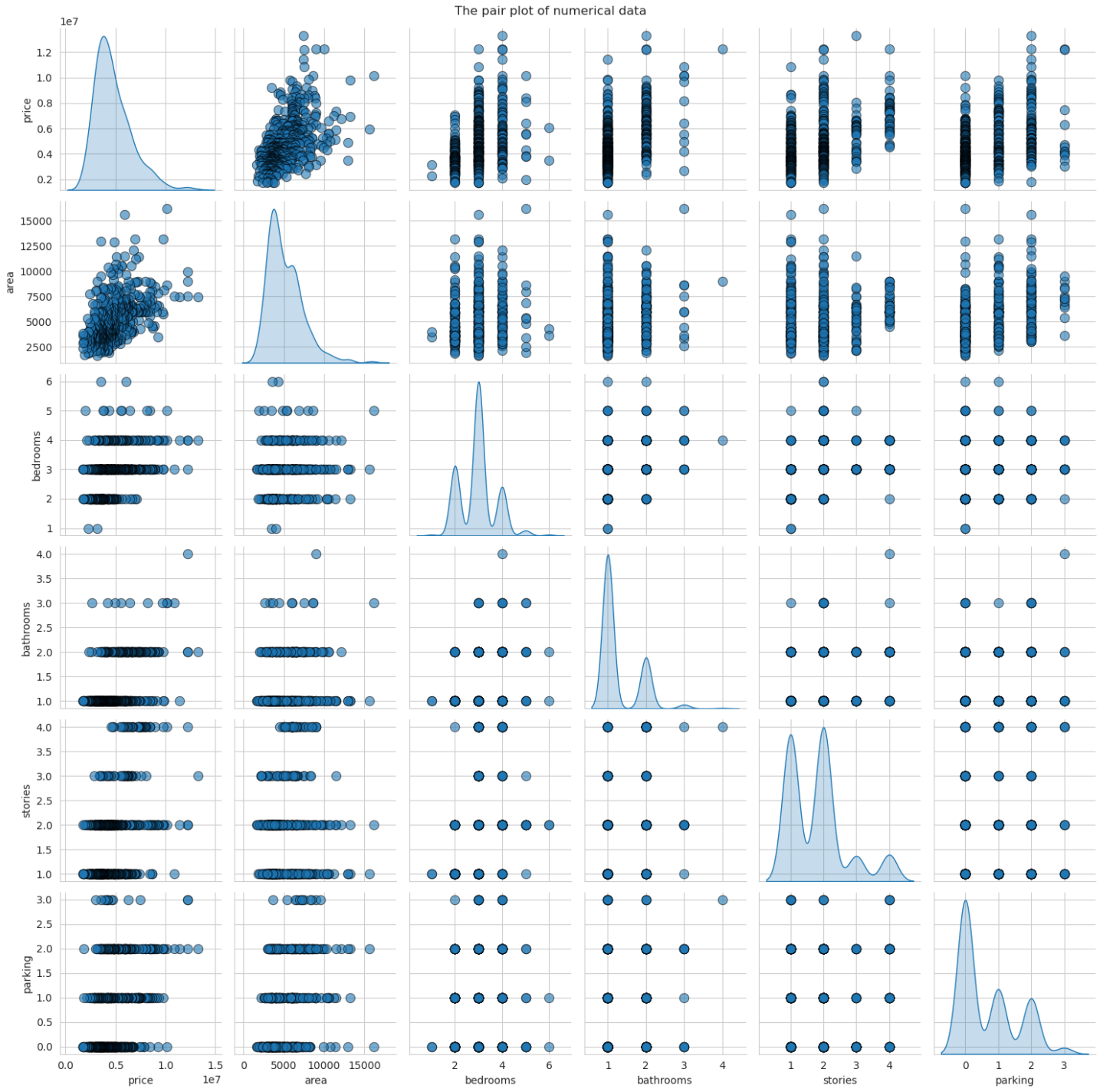

Figure 3: Pair Plot For data

As decipted in figure 3 we can see that there is good accossiation between price and area.As it was expected, they are correrlated because ”area” of a house is a criucial factor on the ”price” of it.
But we can not say anything about other factors where can have an impact on the ”price” and we need to test whether with fitting regression line or correlation tests (contingency tables can be conducted as well).
We have provided correlation matirx as well in order to see how correlated the variables are where can be seen in figure 4:

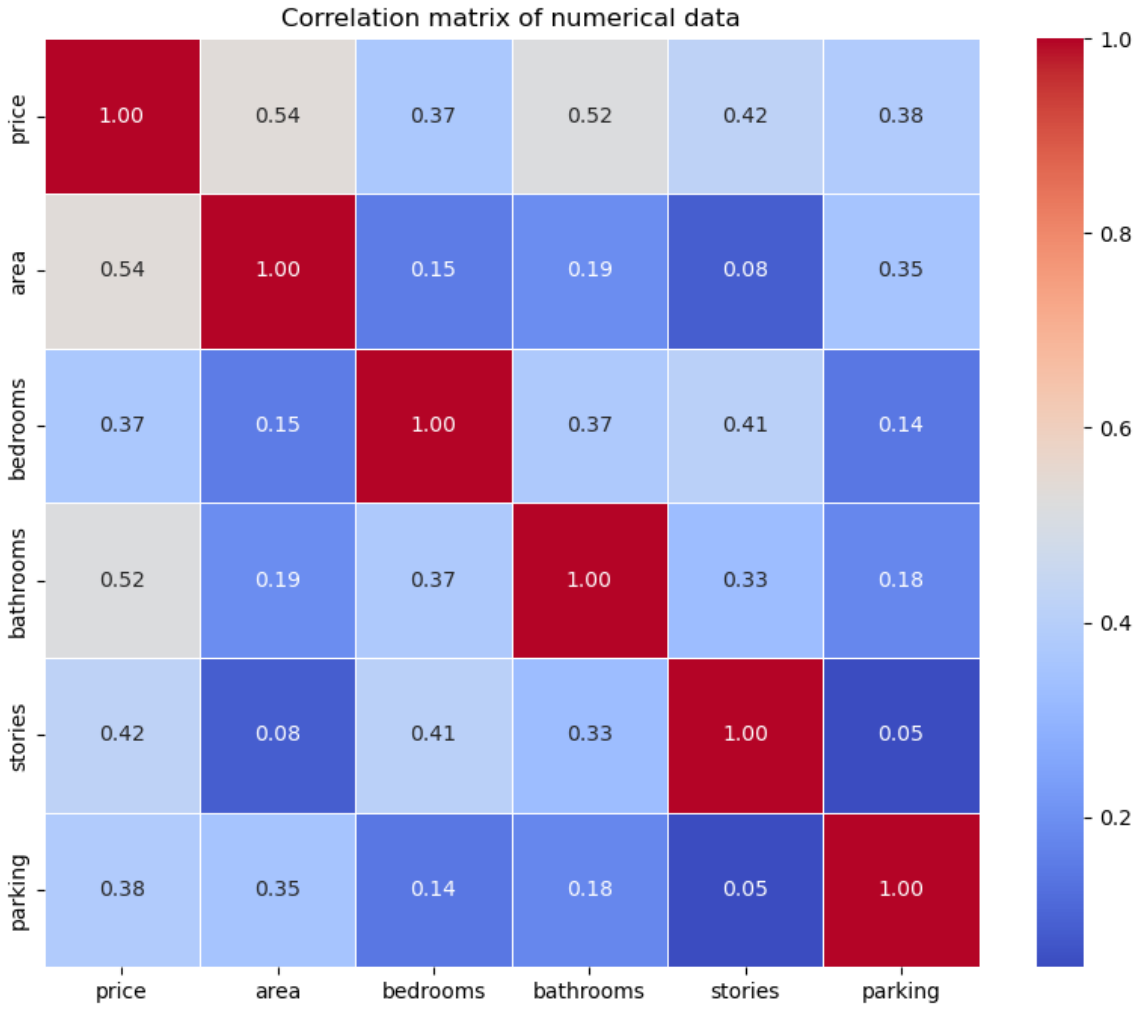

Figure 4: Correlation Matrix

There are three different variables where have an impact on the ”price”. They are ”bathrooms”, ”area”, ”stories”. We know whenever ”stories” of house increases ”price” will increase as well. More details:
There is a moderate positive correlation between price and area, which is expected as larger houses generally cost more.
These also show a positive correlation with the price, suggesting that more bedrooms, bathrooms, and stories typically increase a house’s value. However, the correlation is not as strong as with the area, indicating other factors may also significantly influence the price.
The diagonal histograms in the pair plot show the distribution of each variable.The relationships between some variables, such as price and area, show a roughly linear trend, which is suitable for linear regression models.

2.4 Q-Q plots

Generally Q-Q plots are very good tool in order to assess if a data comes from some theoritical distribtuion such as Normal and etc.Since the parametric tests required data comes from normal distribution we can use Q-Q plots to confirm that we are allowed to use parametric tests or not.
Some of intiution that Q-Q plots will provide for us:

- Normal Distribution Comparison:If the data is normally distributed, the points on the Q-Q plot will lie on a straight diagonal line that typically runs from the bottom left to the top right of the plot.
- Deviations from Normality:If the points systematically deviate from the straight line in a certain pattern, it indicates that the data does not follow a normal distribution. If the points deviate in an upward curve (concave up), the data distribution has heavier tails than a normal distribution (i.e., more data in the tails than expected and also inverse of sentence is true)
- Skewness: If the points form an S-shaped curve, the data is skewed. If the left tail is lower than the diagonal line and the right tail is higher than the line, the data is right-skewed (positive skew). If the left tail is higher than the diagonal line and the right tail is lower, the data is left-skewed (negative skew).

We have provided Q-Q plots between ”price, area” and Normal distribution. and also provide if they are more likely to be t-distirbutions or not.You can see provided Q-Q plots in figure 5:

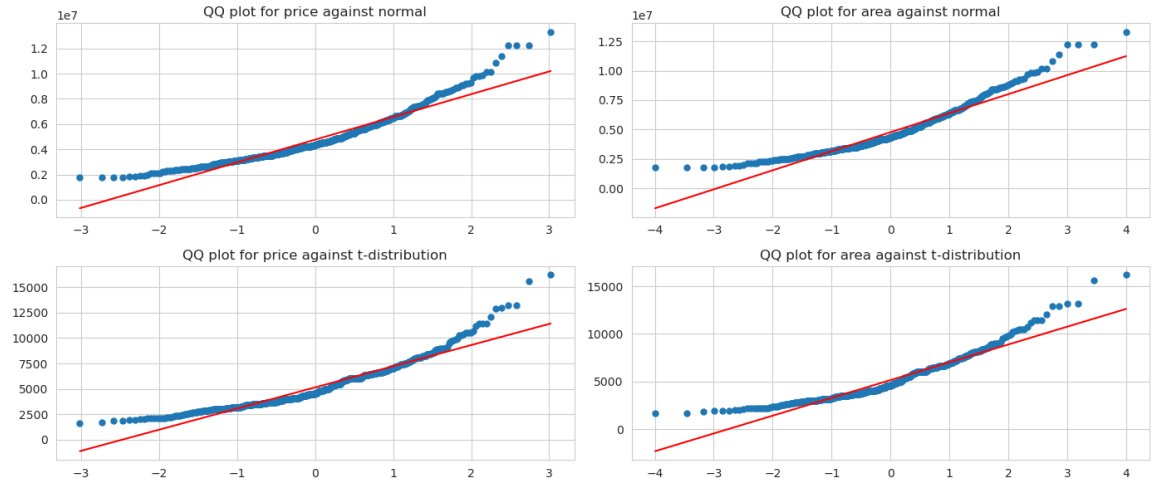

Figure 5: Q-Q plots

The interpret the Q-Q plots:(Note that in ”t-distribution” we assume the degree of freedom 10)
For price and normal:The data points curve upward away from the line at both ends, which suggests that the distribution of ’price’ has heavier tails than the normal distribution. This indicates a right-skewed distribution with a concentration of data on the lower end and some high-value outliers.
For price and T-distribution:The data points curve upward away from the line at both ends, which suggests that the distribution of ’price’ has heavier tails than the normal distribution. This indicates a right-skewed distribution with a concentration of data on the lower end and some high-value outliers.
For area and normal: Similar to the ’price’ plot, the ’area’ plot against the normal distribution shows an upward curvature at both ends. This suggests that the data is right-skewed with more extreme values than you would expect with a normal distribution.

2.5 Categorical Visualization

For categorical visualization we have multiple methods to plot them like barplots and count plots.Since we have more than 8 results already, therefore we avoid to create more plots.The provided bar plots for each categorical features are depicted in figure 6:

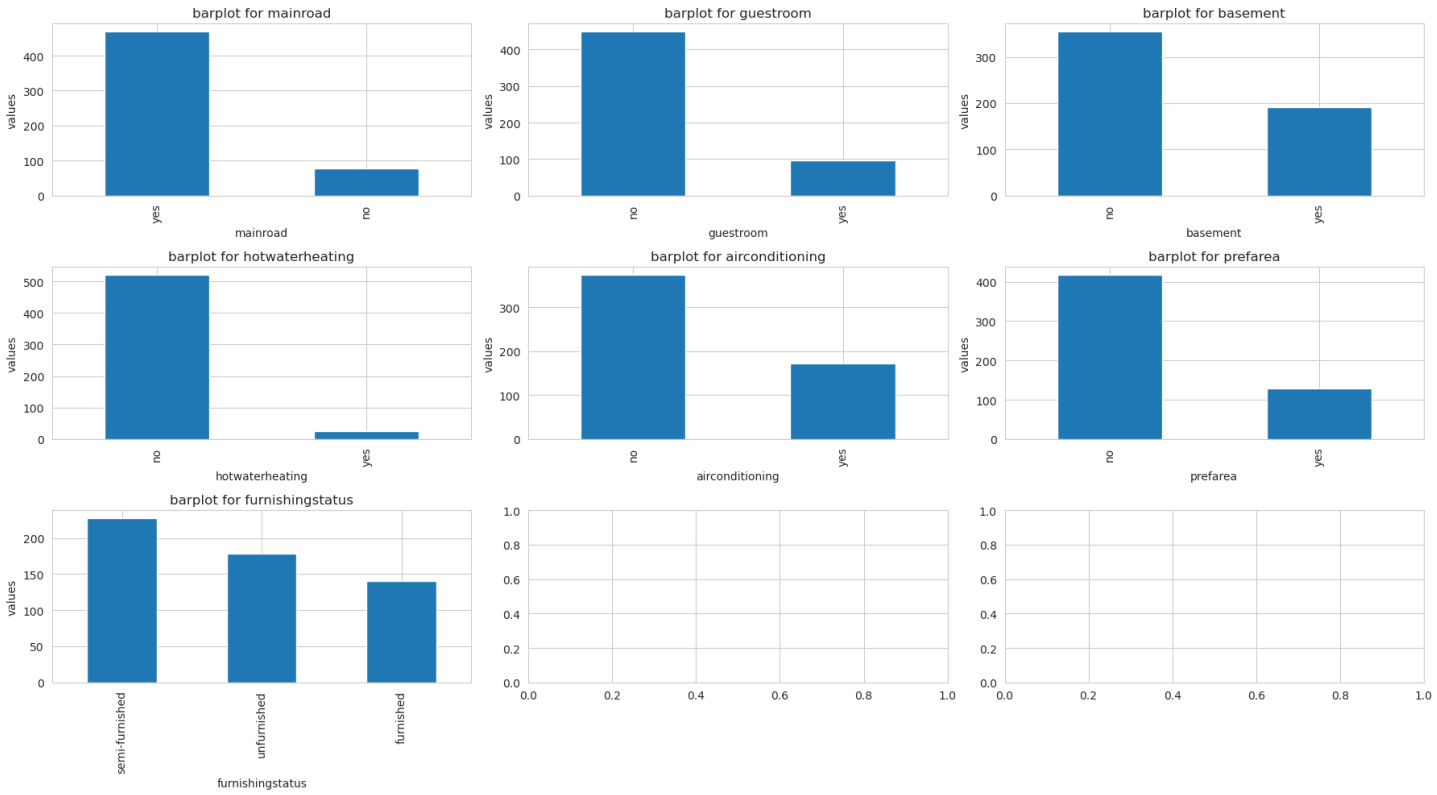

Figure 6: Bar plots for categorical features

The overall details:

1. Barplot for mainroad: The majority of properties are located on the main road, as indicated by the much taller bar for ”yes” compared to ”no”.

2. Barplot for guestroom:Fewer properties have a guestroom than those that do not. The bar for ”no” is substantially taller, showing that guestrooms are less common in this dataset.

3. Barplot for basement:More properties do not have a basement than those that do. The ”no” bar is taller, suggesting basements are not a standard feature in most properties in the dataset.

4. Barplot for hotwaterheating: Hot water heating appears to be an uncommon feature, as very few properties have it, shown by the much smaller ”yes” bar.

5. Barplot for airconditioning:A significant number of properties do not have airconditioning, although the feature is fairly common, with the ”yes” bar being substantial but not as tall as ”no”.

6. Barplot for prefarea: A majority of properties are not in the preferred area, as the bar for ”no” is taller than ”yes”, indicating that being in a preferred area is a less common attribute.

7. Barplot for furnishingstatus:The distribution among furnishing status is relatively more balanced compared to other features. ”Semi-furnished” and ”unfurnished” categories have a similar number of properties, with ”furnished” being slightly less common.

2.6 Heatmap

A heatmap is a graphical representation of data where individual values contained in a matrix are represented as colors. The primary usage of heatmap includes:

- Visualization of Distributions: Heatmaps can show the distribution of a variable across two other variables. For example, in a matrix of two categorical variables, a heatmap can display the count or proportion of observations for each combination of categories.
- Correlation Analysis: One of the most common uses of heatmaps is to visualize the correlation matrix between variables. In such a heatmap, each cell shows the correlation coefficient between two variables, helping to identify which pairs of variables are most positively or negatively correlated

In this part we are going to plot heatmap between two categorical variables ”basement” and ”prefarea”. It is decipted in figure 7:

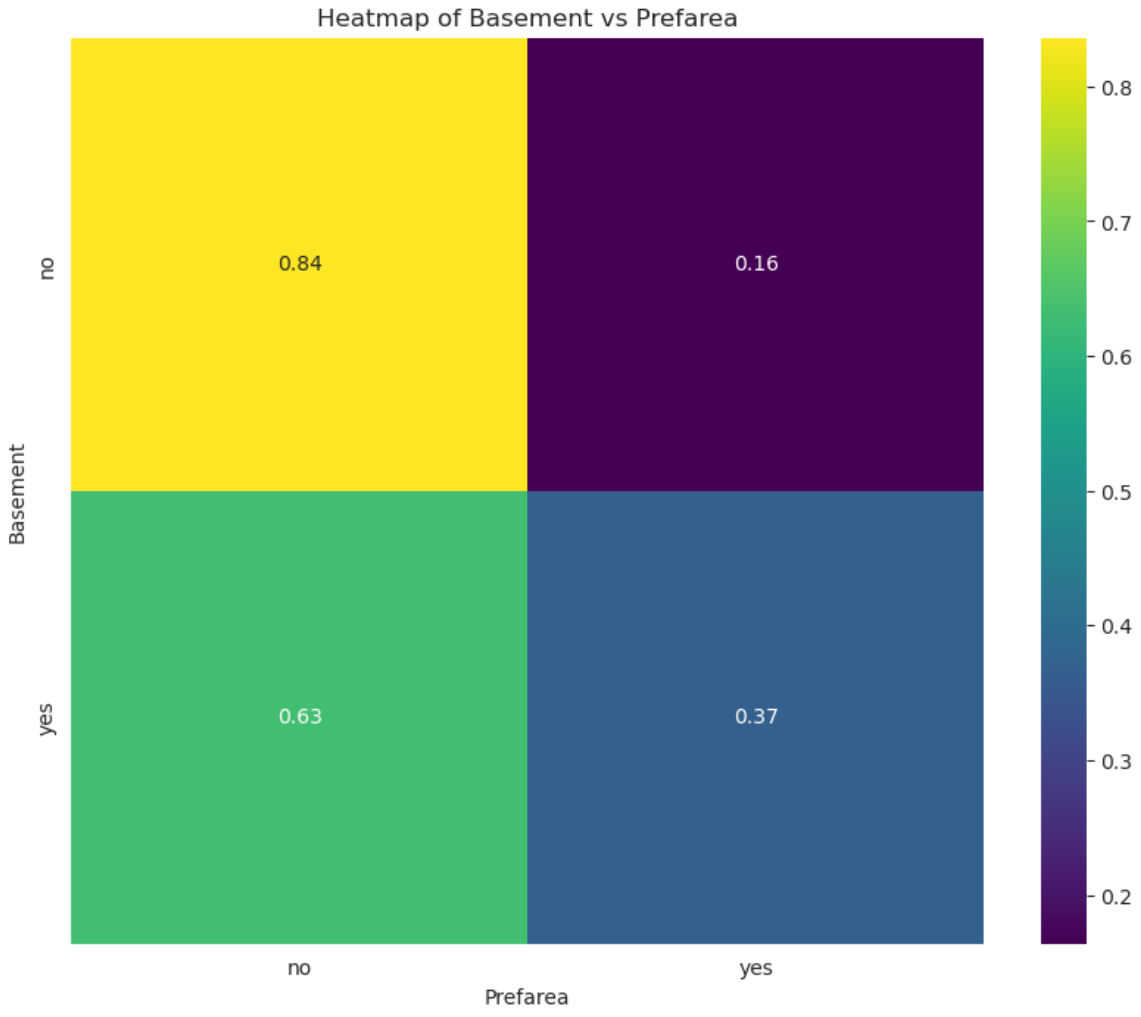

Figure 7: Heatmap for basement and prefarea

Interpretation of decipted heatmap:

- Columns: The categories ”no” and ”yes” under ”prefarea” suggest that this variable indicates whether a preferred area feature is present or not.
- Rows: The ”Basement” variable has two categories: ”no” and ”yes”, indicating the absence or presence of a basement.

Each cells:

- The top left yellow cell with the value ”0.84” indicates that 84% of the observations do not have a basement and are not in the preferred area.
- The top right purple cell with the value ”0.16” suggests that 16% of the observations do not have a basement but are in the preferred area.
- The bottom left green cell with the value ”0.63” indicates that 63% of the observations have a basement and are not in the preferred area.
- The bottom right blue cell with the value ”0.37” suggests that 37% of the observations have a basement and are also in the preferred area.

Summary of heatmap: It seems that a large proportion of houses have not basement and preferred area as well.This can indicate that basements are more common in less preffered areas.

2.7 Mosaic

A mosaic plot is a graphical method of diplaying two-way of higher-way tables. The primary reasons for using a mosaic plot include:

- Visualizing Categorical Data: Mosaic plots are specifically designed to visualize relationships between two or more categorical variables. They provide a quick and intuitive picture of how different categories interact.
- They help in comparing the proportions of categories and the joint distribution of two categorical variables. Each rectangle's size is proportional to the frequency or count of the category, making it easy to compare the relative sizes of each category. Detecting Patterns: Mosaic plots can reveal patterns, interactions, and relationships between categorical variables that might not be immediately obvious from raw data tables.

In this part we are going to plot mosaic between two categorical variables "furnishingstatus" and "guestroom". The mosiac plot corresponding to the cataegorical variables are illustrated in figure 8: The interpret mosaic plot:

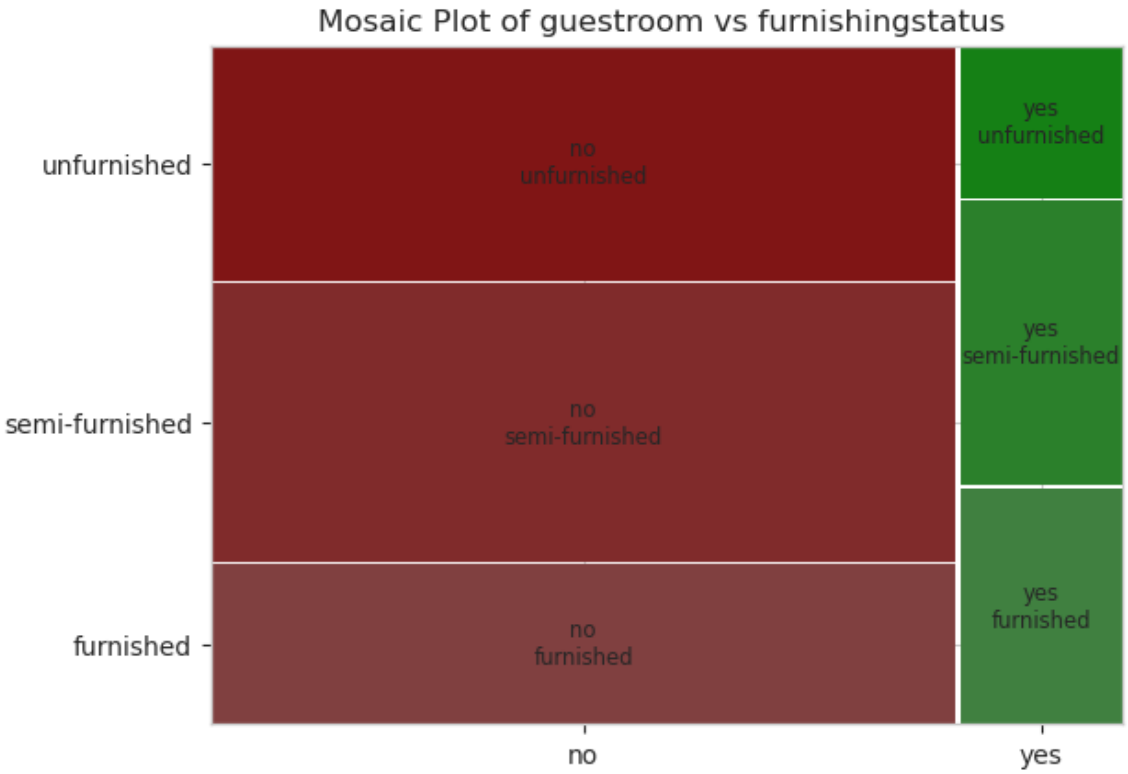

Figure 8: Mosaic plot for furnishingstatus and guestroom

- The x-axis represent the presence of a guestroom ('no' on the left, 'yes' on the right), while the y-axis represents the furnishing status of the property.
- The width of the columns corresponds to the proportion of properties with and without a guestroom. If the widths are equal, it would suggest an equal number of properties with and without guestrooms. If one is wider, that category has a higher count.

- The height of the rows corresponds to the proportion of properties within each furnishing status category.
- The size of each colored rectangle indicates the count or proportion of properties that fall into both categories (for example, ’furnished’ and ’yes’ for having a guestroom).
- The largest rectangle appears to be for ’no guestroom’ and ’unfurnished’, indicating that this is the most common category combination within the dataset.

3 Parametric Inference and Estimation

In this section we are going to conduct some test over parametric of data and also see if MLE is appliable in order to fit normal distirbutions on the data conduct CI for the parameters.

3.1 Shapiro Test For Normality Assumptions

In order to conduct approximately any parametric test, we should check if normality assumptions are met. For ”price” and ”area” features we going to use shapiro test in order to see if assumptions whether satisfiable or not. The result of shapiro test are in table 1: As it has been illustrated in table 1 the p-value is close, indicating that data are

Table 1: Shapiro test on ”price” and ”area”

	Price	Area
p-values	3.1e-16	2.6e-17

not coming from normal distribution. So we must use some transformations like log and boxcox transformations where the result of shapiro test is in table 2:

Table 2: Shapiro test on log transformed ”price” and ”area”

	Price	Area
p-values	0.21	0.011

The log transformation of ”price” column will be normal eventhough ”area” feature will be rejected since the p-value is lower than 0.05.Careful the null hypothesis for shapiro test is:

$$\begin{aligned} H_0 &: \text{Data comes from normal distribution} \\ H_1 &: \text{Data will not comes from normal distribution} \end{aligned} \tag{2}$$

We tranform data with boxcox as well where the results are illustrated in tabel 3:

Table 3: Shapiro test on boxcox transformed ”price” and ”area”

	Price	Area
p-values	0.5	0.028

From any transformation, we can not see the transformed ”area” to be normal distribution but we made ”price” to be like normal distribution.

3.2 Parametric Tests on Parameters of Data

For this purpose we define a function where help us to get sample from data with any particular feature. We know that if we want to test the mean of each sample we must sure that data is normal.But if the sample size are large enough therefore we can use CLT[2] where normaility assumptions will met.
For conducting t-test, at first we are going to generate two different samples from particular feature (with large sample size) based on the one balanced categorical column like "basement".The test is to see if these two samples where one of them has "basement" and another has not having same mean or not. The null hypothesis and alternative hypothesis for t-test are as follows:

$$\begin{aligned} H_0 &: \mu_0 = \mu_1 \\ H_1 &: \mu_0 \neq \mu_1 \end{aligned} \tag{3}$$

The result is:

$$\begin{aligned} \text{p-value of t-test} &= 9.114493979954155e-06 \\ \text{statistic} &= 4.437180316396756 \end{aligned} \tag{4}$$

The null hypothesis will rejected since p-value is less than significant level of 0.05.
Another t-test can conducted when we want to see the mean of "area" of houses where they have "basement" are equal to the mean of "area" of houses not having "basement".
So the result is:

$$\begin{aligned} \text{p-value of t-test} &= 0.2686520802379405 \\ \text{statistic} &= 1.1061719985940541 \end{aligned} \tag{5}$$

Therefore the null hypotehsis will not rejected since p-value is higher than significant level of 0.05.

[2]Centeral Limit Theorem

3.3 Estimation MLE

In this section we are going to use MLE approach to fit a normal distribution for both ”price” and ”area” data. First we discuss how to estimate parameters for normal distributions using MLE:
If $X_1, X_2, \cdots, X_n$ are i.i.d with $N(\mu, \sigma^2)$,their joint density is the product of their marginal densities:

$$f(x_1, x_2, \ldots, x_n|\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right) \tag{6}$$

The log likelihood ratio is:

$$l(\mu, \sigma) = -n\log(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{7}$$

The partials with the respect to μ, σ are:

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) \\ \frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \sigma^{-3}\sum_{i=1}^{n}(X_i - \mu)^2 \end{aligned} \tag{8}$$

From setting this partials to zero we can have MLE parametrs:

$$\begin{aligned} \hat{\mu}_{\text{MLE}} &= \overline{X} \\ \hat{\sigma}_{\text{MLE}} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \end{aligned} \tag{9}$$

So we have fit the MLE parameters to the ”price” where the result is in figure 9:

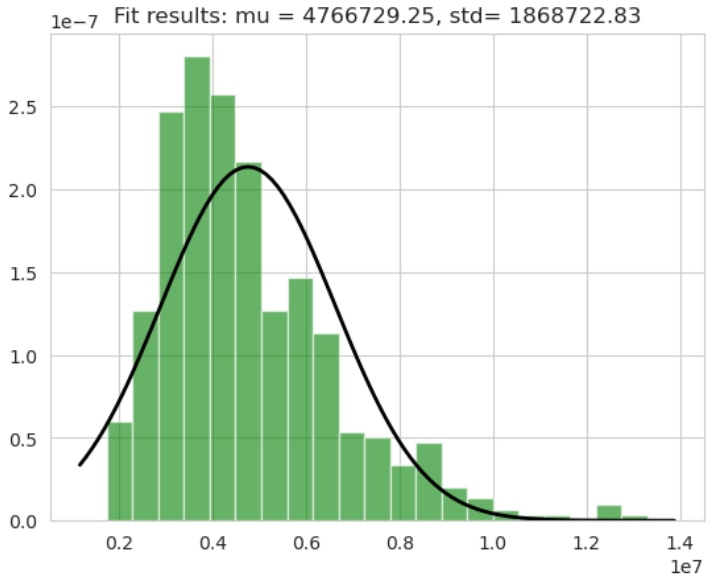

Figure 9: Real ”price” data with fitted normal

And for ”area” data the figure 10 will show the result:

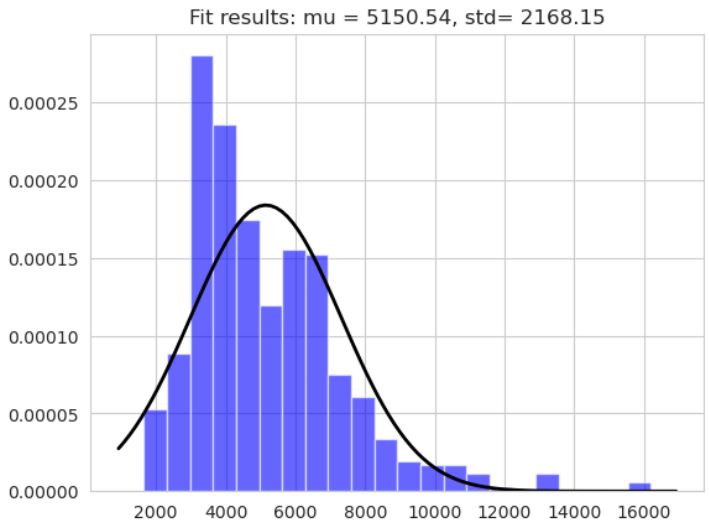

Figure 10: Real ”area” data with fitted normal

We know that data didn't comes from normal but we perform MLE to fit the best normal distributon to these data. The confidence interval for μ can be formulated as follow:

$$\hat{\mu}_{\text{MLE}} \pm \frac{1}{\sqrt{n-1}} \hat{\sigma}_{\text{MLE}} t_{n-1}(\alpha/2) \tag{10}$$

For σ we have:

$$\left(\frac{n\hat{\sigma}_{\text{MLE}^2}}{\chi^2(\alpha/2)}, \frac{n\hat{\sigma}_{\text{MLE}^2}}{\chi^2_{n-1}(1-\alpha/2)} \right) \tag{11}$$

Where the CI of both ”price” and ”area” are:

$$\begin{aligned} \text{CI for mean price} &= (4609839.44, 4923619.056) \\ \text{CI for sigma price} &= (1763980, 1986789.98) \\ \text{CI for mean area} &= (4968.51, 5332.57) \\ \text{CI for sigma area} &= (2046.62, 2305.13) \end{aligned} \tag{12}$$

where the real mean and std for ”price” data are 4766729, 1870440.And also for ”area” data are 5150.54, 2170.141023. Therefore the true parameters are in the CI's.
We have also used MLE method for fitting gamma distribution to ”price” where it has been illustrated in figure 11: The parameters are $(\alpha_{\text{MLE}}, \theta_{\text{MLE}}) = (9.178406870555618, 511752.83398873534)$

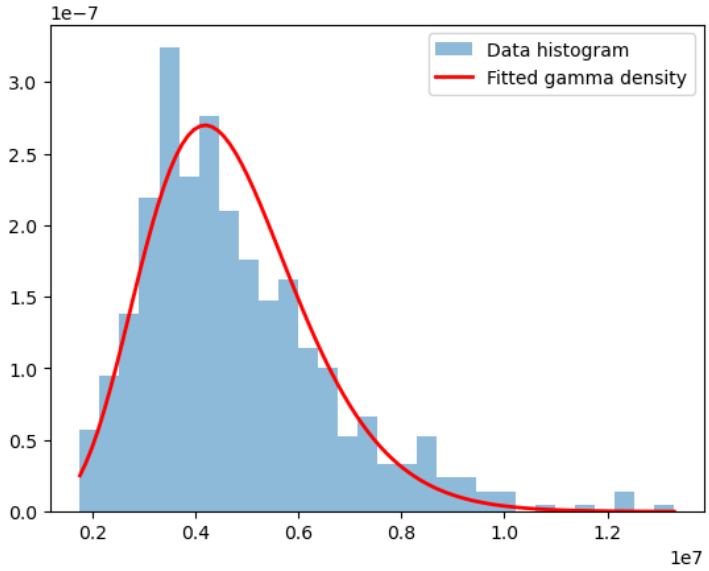

Figure 11: Fitted gamma with MLE on ”price” data

we can use the asymptotic properties of the MLE in order to conduct CI for the parameters Confidence interval for the parameters are in table 4:

	left CI	right CI
α	7.20321701	110138.204
θ	511750.859	621881.860

Table 4: CI for gamma prameters

3.4 Bootstrap

We will simulate 100000 times getting a sample from ”area” data and calculate each mean and plot it in order to see the distribution for mean of ”area” where can be seen in figure 12: In this figure 12 we use the percentile

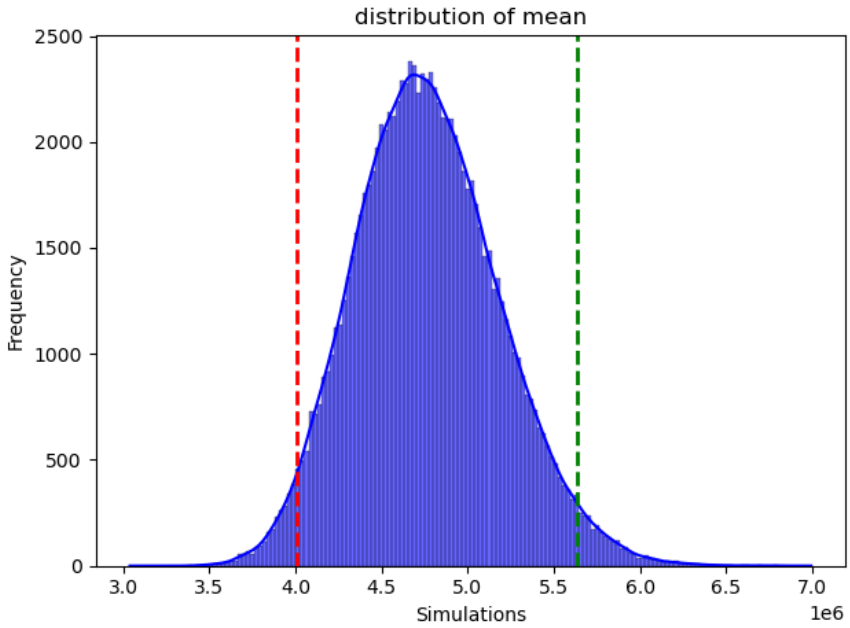

Figure 12: Distribution of mean using Bootstrap

interval in order to see the confidence interval of bootstrap where can be calculated as follow: Let $\hat{\theta}^{*(\alpha/2)}$ and $\hat{\theta}^{*(1-\alpha/2)}$ be the $\alpha/2$ and $1-\alpha/2$ quantiles of the simulate values of $\hat{\theta}^*$.(If we performed B bootstrap simulations, these are the $(\alpha/2B)^{\text{th}}$ and $((1-\alpha/2)B)^{\text{th}}$ ordered valuse of $\hat{\theta}^*$).Construct a $100(1-\alpha)\%$ confidence interval as:

$$[\hat{\theta}^{*(\alpha/2)}, \hat{\theta}^{*(1-\alpha/2)}] \tag{13}$$

So the percentile interval in our scenario is:

$$\text{CI for mean of area} = [4262.14625, 6176.10125] \tag{14}$$

Where the real mean is 5150.54.Thus our CI will cover the parameter mean of the particular data(”area”).

4 Hypothesis Testing

In this section, we examined the data using different statistical tests to understand the relationships between various factors. For instance, when we wanted to find out if two things were related in a way that wasn't just due to chance, we used the Spearman test. This test doesn't need the data to be spread out in a particular way, which is helpful when we're not sure about the data's pattern. But when we could assume that the data followed a normal pattern, we used the Pearson test to see how strongly two things were linked.
When it came to comparing averages across different groups, we used the ANOVA test, which works well when data is spread out normally. If the data didn't spread out this way, we used the Kruskal-Wallis test instead, which is a good alternative that doesn't need the data to be normal. For situations where the same thing was measured more than once or the data was in pairs, we chose the Friedman's test, which is like the ANOVA test but for data that doesn't follow a normal spread.
We also looked at whether certain categories, like house prices and the number of bedrooms, or the number of bedrooms and parking spaces, had any influence on each other. We did this by making special tables called contingency tables. If the data was in pairs, we used methods like the sign test and the Wilcoxon test, which are good when the data doesn't have a normal spread. These methods helped us understand our data better and gave us clear answers to our questions.

4.1 Sign and Wilcoxon

We will perform one Sign test for meidan of "price" and one Wilcoxon test for median of "area".Later we will conduct a confidence interval for meidan as well and interpret them.
Another strategy is when we are going to conduct the null hypothesis, we assume the null hypothesis is equal to mean. This can provide us very good understanding if the mean of data is equal to mean and see even if the distribution is either symmetric or not. The null hypothesis for the meidan of "price" of sign test:

$$\begin{aligned} H_0 &: median = \mu \\ H_1 &: median \neq \mu \end{aligned} \tag{15}$$

The result of sign test is in table 5: The p-value is approximately zero so the null hypothesis will be rejected and

Table 5: Result of Sign Test

P-value	Statistic
1.7466750325988145e-05	0.4073394495412844

therefore the median of "price" data is not equal to mean of it, therefore we can say that this distribution is not symmetric without plotting the histogram.
For the median of "area", we will conduct Wilcoxon test therefore the null hypothesis:

$$\begin{aligned} H_0 &: median = \text{real median} \\ H_1 &: median = \mu \end{aligned} \tag{16}$$

The result of Wilcoxon test: But there is a bug here we set the null hypothesis to the real mean but why the p-value is so low and therefore we can reject null hypothesis but why?
The answer is ono of the assumptions for the Wilcoxon test is when the data comes from a continuous symmetric

Table 6: Result of Sign Test

P-value	Statistic
4.9868918908861984e-05	58321.5

random variable.The particluar data (”area”) is continuous but it is not symmetric therefore we can not perform Wilcoxon test for this data. We will use Sign test again for ”area”, the null hypothesis agains alternative hypothesis is:

$$\begin{aligned} H_0 &: median = \mu \\ H_1 &: median \neq \mu \end{aligned} \tag{17}$$

The result of Sign test is:

Table 7: Result of Sign Test

P-value	Statistic
0.0003124001938012276	0.42201834862385323

As you can see the table of result 7, we conclude that the median is not equal to the mean of ”area” therefor this data is not symmetric and it is indeed considered to be skewed. The 96%confidence interval for median of each data(”price” or ”area”) is:

$$CI = (248.0, 296.0) \tag{18}$$

Careful for sign test a $100P\%$ confidence interval of median in sign test

4.2 Correlation Tests

It is very neccessary to see if the any particular features are correlated to one another or not. We can use the result of this test to fit regression line on the future or extract the features where they are not correlated to each other. Firstly we are going to conduct a Pearson test to see if the ”area” and ”price” are correlated or not. The null hypothesis against alternative:

$$\begin{aligned} H_0 &: \rho_{XY} = 0 \\ H_1 &: \rho_{XY} \neq 0 \end{aligned} \tag{19}$$

The statistic is:

$$R = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \tag{20}$$

Under the null hypothesis, $R \sim t_{n-2}$,i.e, R is distributed as Student-t distribution with n-2 degrees of freedom. The Pearson result is table 8:

Table 8: Result of Pearson

R-statistic	t-value	p-value
14.80	1.96	7.38e-42

The ”area” and ”price” data are not normal therefore we can conclude that they are correlated or not with even low p-value so we can conduct a non-parametric test like Spearman test where the procedure for computing the Spearman correlation test:

- Order and rank the x's and y's. With each pair (x_i, y_i) we will have two ranks (R_i^x, R_i^y).
- Compute the absulote different of two ranks $d_i = |R_i^x - R_i^y|$.
- Compute the sum of the squared differences:

$$D = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} |R_i^x - R_i^y|^2$$

- The Spearman's rank correlation coeffient is:

$$r_s = 1 - \frac{6D}{n(n-1)^2}$$

Having computed r_s we can now conduct the hypothesis test:

$$\begin{aligned} H_0 &: \text{There is no accossiation between the ranks, against} \\ H_1 &: \text{There is accossiation between the ranks} \end{aligned} \tag{21}$$

The result of Spearman test between ”price” and ”area” is:

Table 9: Result

P-value	Correlation Coeffient
3.12e-55	0.602

As you can see the result in table 9, we reject null hypothesis for significant level of 0.05.Thus there is a accossiation between ”price” and ”area”.

4.3 ANOVA and Kruskal-Wallis

If the data comes from normal distribution, we can perform ANOVA test.As a result of previous part we know our data does not come from normal distribution,Therefore we might use some transformation like boxcox as implemented before(you can see the result in table 3).
Some strategy that can help us to use ANOVA test effciently and make groups is when we use categorical data in order to see the differences of means in each group.For example in ”furnishingstatus” we can make three groups where data is distinguised from thier ”furnishingstatus” values.
But a naive idea is when we split data into groups where in each group we get 100 samples therefore from CLT their mean's will be equal to each other and therefore it can not be a good idea for testing ANOVA.
The procedure of ANOVA test is:

- The statistic model for the ANOVA:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- The analysis of variance is based on the following identity:

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i.})^2 + J\sum_{i=1}^{I}(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

  where $\bar{Y}_{i.}$ and $\bar{Y}_{..}$ are:

$$\begin{aligned} \bar{Y}_{i.} &= \frac{1}{J}\sum_{j=1}^{J} Y_{ij} \\ \bar{Y}_{..} &= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij} \end{aligned} \tag{22}$$

  We can write them in another notation where:

$$\begin{aligned} SS_{\text{TOT}} &= \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{..})^2 \\ SS_W &= \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i.})^2 \\ SS_B &= J\sum_{i=1}^{I}(\bar{Y}_{i.} - \bar{Y}_{..})^2 \\ SS_{\text{TOT}} &= SS_W + SS_B \end{aligned} \tag{23}$$

- The statistic is:

$$F = \frac{SS_B/(I-1)}{SS_W/\left[I(J-1)\right]}$$

- The null hypothesis is mean of each group are equal to each other while alternative hypothesis is, there exist at least one pair of groups that has difference in their means.

The ANOVA test result for ”price” of a house splited by the ”furnishingstatus” feature are shown in the following table 10:

Table 10: Result of the ANOVA test

Statistics	P-value
545624.13	0

As you can see the result in table 10, we reject null hypothesis since the p-value is approximately zero. We concluded that the means of measurements from different labs are not all equal, but the test gives no information about how they differ, in particular about which pairs are significantly different.
So we can use the Tukey’s method to solve this problem where the result of Tukey is in table 11:

Table 11: Multiple Comparison of Means Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
furnished	semi-furnished	-6.6989	0.0	-6.7171	-6.6806	True
furnished	unfurnished	-8.0597	0.0	-8.0789	-8.0404	True
semi-furnished	unfurnished	-1.3608	0.0	-1.3779	-1.3438	True

We can see all the means are diffenet to one another so the mean of house with different ”furnishingstatus” status has different mean values.More detail about the result of tukey:

1. Furnished vs. Semi-furnished:
    - The mean difference is approximately -6.6989.
    - The p-value adjusted for multiple comparisons (p-adj) is 0.0, which is significant at the 0.05 level.
    - The confidence interval ranges from -6.7171 to -6.6806.
    - The reject column is True, which means the furnished and unfurnished groups’ means are significantly different.
2. Furnished vs. Unfurnished:
    - The mean difference is approximately -8.0597.
    - The p-adj is 0.0, again indicating a significant difference at the 0.05 level.
    - The confidence interval for the mean difference is between -8.0789 and -8.0404.
    - The reject column is True, indicating that the null hypothesis (that the means are the same) is rejected. This means there is a statistically significant difference between the furnished and semi-furnished groups.
3. Semi-furnished vs. Unfurnished:
    - The mean difference is about -1.3608.
    - The p-adj is 0.0, which suggests a significant difference.
    - The confidence interval for the mean difference is from -1.3779 to -1.3438.

- The reject column is True, indicating that there is a significant difference in means between the semi-furnished and unfurnished groups.

We can perform the Kruskal-Wallis test on the real data(not transformed of the data) to see if the real means are different or not.The Kruskal-Wallis test prodecure is:

- The observations are pooled together and ranked

$$R_{ij} = \text{the rank of } Y_{ij} \text{ in the combined sample}$$

Let $\bar{R}_{i.}$ and $\bar{R}_{..}$ are:

$$\begin{aligned} \bar{R}_{i.} &= \frac{1}{J_i} \sum_{j=1}^{J_i} R_{ij} \\ \bar{R}_{..} &= \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{J_i} R_{ij} \\ &= \frac{N+1}{2} \end{aligned} \tag{24}$$

where N is the total number of observations.

The result for the Kruskal-Wallis test for the ”price” data based on categroical feature ”furnishingstatus” is: Due to

Table 12: Resut of Kruskal-Wallis

Statistic	P-value
69.58	7.7e-16

p-value is less than significant level of 0.05 therefore it will reject the null hypothesis.So there is a atleast difference between the mean of each group.We know that Tukey’s method is for ANOVA testing but there is an alternative method for Kruskal-Wallis test where is Dunn’s test where the reuslt is in table 13:

Table 13: Dunn’s method result for Kruskal-Wallis

	furnished	semi-furnished	unfurnished
furnished	1.000000e+00	1.373095e-01	3.385291e-14
semi-furnished	1.373095e-01	1.000000e+00	1.499534e-10
unfurnished	3.385291e-14	1.499534e-10	1.000000e+00

The interpretation of the result in table 13:

1. Furnished vs. Semi-furnished:
   - The p-value is 1.373095e-01.
   - This value is above the common alpha level of 0.05, which suggests that the difference in median ranks between furnished and semi-furnished is not statistically significant.
2. Furnished vs. Unfurnished:

- The p-value is 3.385291e-14 (which is an extremely small number, practically zero).
- This very small p-value indicates a statistically significant difference in median ranks between furnished and unfurnished groups

3. Semi-furnished vs. Unfurnished:

- The p-value is 1.499534e-10 (also a very small number, approaching zero).
- This indicates a statistically significant difference in median ranks between semi-furnished and unfurnished groups.

In summary, based on the p-values between each groups(Bonferroni used to adjust multiple comparisons) are indicating we can not say anything for the furnished houses versus semi-furnished groups.However, between other groups there is a statistically significant difference in median ranks.

4.4 Two-Way ANOVA

A two-way ANOVA is typically used when you have two independent categorical variables and one continuous dependent variable, and you want to understand if there is an interaction effect between the two categorical variables on the dependent variable. Now consider two categorical data in our dataset like ”mainroad” and ”airconditioning” and we want to see if they have an interaction effect on the continuous data variable(”price”) or not.
Therefore we can conduct two-way ANOVA where the reuslt is as follows:

Table 14: Result of two-way ANOVA test

	sum of square	df	F	PR(>F)
C(mainroad)	0.099965	1.0	62.452983	1.537529e-14
C(airconditioning)	0.214557	1.0	134.044066	7.560201e-28
C(mainroad):C(airconditioning)	0.001940	1.0	1.212193	2.713878e-01
Residual	0.865950	541.0	-	-

We are going to provide some explanintation for the result(in table 14):

1. Mainroad:
    - The sum of squares associated with ’mainroad’ is approximately 0.099965, which represents the variation due to ’mainroad’.
    - The F-statistic is approximately 62.452983, which is a measure of the variance due to ’mainroad’ relative to the residual (error) variance.
    - The p-value (PR(>F)) is very small (1.537529e-14), suggesting that the presence or absence of a main road has a statistically significant effect on the ”price” variable.
2. Airconditioning:
    - The sum of squares is about 0.214557, which represents the variation due to ’airconditioning’.
    - The F-statistic is approximately 134.044066, which suggests a high variance due to ’airconditioning’ compared to the residual variance.
    - The p-value is extremely small (7.56020e-28), indicating a statistically significant effect of air conditioning on the dependent variable.
3. Interaction:
    - The sum of squares is 0.001940, which represents the variation due to the interaction between ’mainroad’ and ’airconditioning’.
    - The F-statistic is 1.212193, which compares the interaction variance to the residual variance.
    - The p-value is 0.271387e-01 (or approximately 0.271 when adjusted for scientific notation), which is above conventional significance levels such as 0.05 or 0.01. This suggests that the interaction between ’mainroad’ and ’airconditioning’ is not statistically significant.
4. Residual:

- The sum of squares is approximately 0.865950, representing the variation within the groups that is not explained by the main effects or interaction.
- The degrees of freedom for residuals is 541.0, which is derived from the total number of observations minus the number of groups defined by the factors.

In conclusion, both ’mainroad’ and ’airconditioning’ individually have statistically significant effects on the dependent variable, but their interaction does not significantly affect the dependent variable.
In order to see the intercation plot to convince the interaction does not significantly effect the ”price” variable we have provided interaction plot where depicted in figure 13:

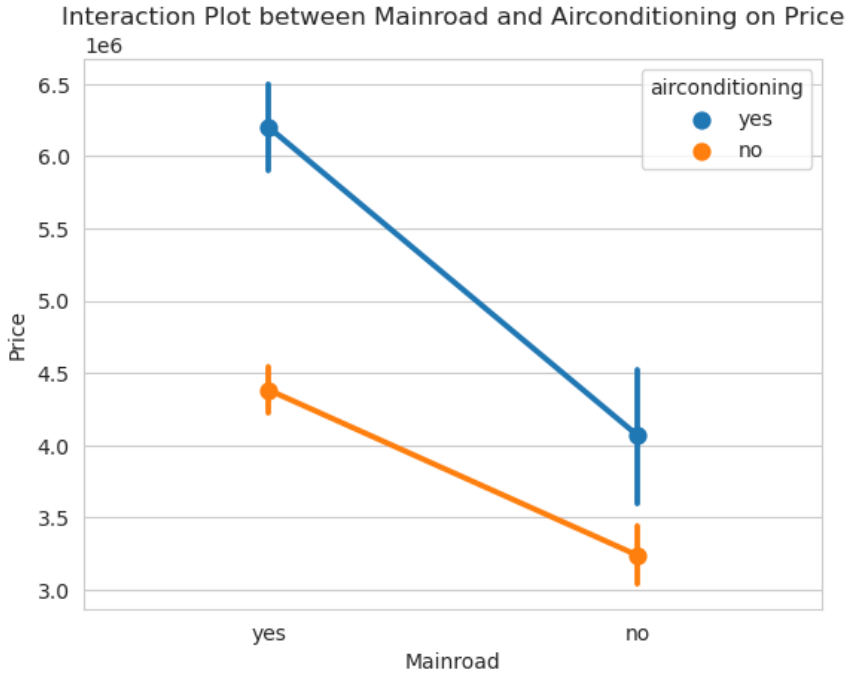

Figure 13: Interaction plot for ”main road” and ”airconditioning”

As you can see the intercation plot will prove where the intercation between ”mainroad” and ”airconditioning” does not effect the ”price” variable.

4.5 Contingency Table

Contingency tables are used to perform the Chi-square test of independence, which determines whether there is a significant association between two categorical variables. If the test indicates that there is no association, the variables are considered to be independent of each other.We will create two contingency tables in order to see if the variables in table are dependent or not. The procedure of test is:

1. Assume θ_{ij} is the probability that an item will fall into the cell belonging to the ith row and the jth column.
2. $\theta_{i.}$ is the probability that an item will fall into the ith row
3. $\theta_{.j}$ is the probability that an item will fall into the jth column
4. Null hypothesis:
$$\theta_{ij} = \theta_{i.} * \theta.j$$
5. Estimation of $\theta_{i.}$ and $\theta_{.j}$:
$$\hat{\theta}_{i.} = \frac{f_{i.}}{f}$$
$$\hat{\theta}_{.j} = \frac{f_{.j}}{f} \tag{25}$$
Under the null hypothesis of independence we get:
$$e_{ij} = \hat{\theta}_{i.} * \hat{\theta}_{.j} * f$$
6. Once we calculate e_{ij} we can decide if null hypothesis will be rejected or not with χ^2 statistic:
$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$
We will reject the null hypothesis if it exceeds $\chi^2_{\alpha,(r-1)(c-1)}$

The contingency table between ”price” and ”bedrooms” where the ”price” data are labeled as ”low”,”medium” and ”high” and ”bedrooms” variate from 1 to 6 bedrooms in our dataset where this contingency table is in table 15:

Table 15: Contingency table between ”price” and ”bedrooms”

	bedrooms	1	2	3	4	5	6
price	low	2	83	83	16	1	1
	medium	0	46	100	28	3	0
	high	0	7	117	51	6	1

The result for independency test is in table 16:

Table 16: Result test between ”price” and ”bedrooms”

P-value	Degree of freedom	statistic
2.24e-16	10	96.93

The interpretation of the reuslt: The chi-square statistic is 96.93, and the p-value is 2.24e-16. This means that there is statistically significant evidence to reject the null hypothesis of independence between the two variables. In other words, there is a relationship between the price of a home and the number of bedrooms it has.
The contingency table between ”area” and ”stories” has been shown in table 17:

Table 17: Contingency table between ”area” and ”stories”

	stories	1	2	3	4
area	low	78	97	9	0
	medium	69	88	12	10
	high	80	53	18	31

The result of this test, has been shown in table 18:

Table 18: Result test between ”area” and ”storie”

P-value	Degree of freedom	statistic
7.16e-10	6	54.06

The reuslt’s interpretation:The chi-square statistic is 54.06, and the p-value is 7.16e-10. This means that there is statistically significant evidence to reject the null hypothesis of independence between the two variables. In other words, there is a relationship between the area of a house and the number of stories it has.

5 Regression Analysis

This part looks at the relationships between a dependent variable and several independent variables using multiple regression analysis method. In detail, we will fit various regression lines to the data, and carry out the interpretation for each model. Along with that, we will make use of the bootstrap method to represent the distribution of the slope on the regression line, and simultaneously learn of its variability. The residuals of each model will then be scrutinized, and the model's goodness-of-fit will be examined and shortcomings were identified, if any. Ultimately, I will present my comprehensive report, which will be a summary of the main results of my statistical testing, analysis and interpretation during this study.

5.1 Regression With One Variable

We will consider two variables to fit regression line on these data.The explanatory variable is "area" while the dependent variable is "price"(since scatter plot was already plotted in figure 3, We know that these variables are correlated to each other.)
The procedure of regression line as follows:

- Let y_i be the observed value of the random variable Y_i dependes on X_i

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Linear Least Square:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{26}$$

We must minimize the LLS function, therefore we can estimate β_0, β_1:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned} \tag{27}$$

So apply the algorithem for our scenario the result for the regression line between "price" and "area" is shown in the table 19:

Table 19: OLS result

	coef	std err	t	P>|t|	[0.025 , 0.975]
const	2.387e+06	1.74e+05	13.681	0.000	2.04e+06 , 2.73e+06
area	461.9749	31.226	14.795	0.000	400.637 , 523.313

The result interpretation:

- The model is statistically significant, with an F-statistic of 218.9 and a p-value less than 0.001. This means that the independent variable "area" has a statistically significant impact on the dependent variable "price".
- The R-squared value is 0.287, which means that the model explains 28.7% of the variance in price. This is a relatively low R-squared value, suggesting that there are other factors that influence price besides area.

- The coefficient for area is positive and statistically significant, with a p-value less than 0.001. This means that there is a positive relationship between area and price. In other words, as the area of a house increases, the predicted price also increases.

The fitted line beside the data has been depicted in figure 14, note that the distribution of the residuals are also has been shown in the figure:

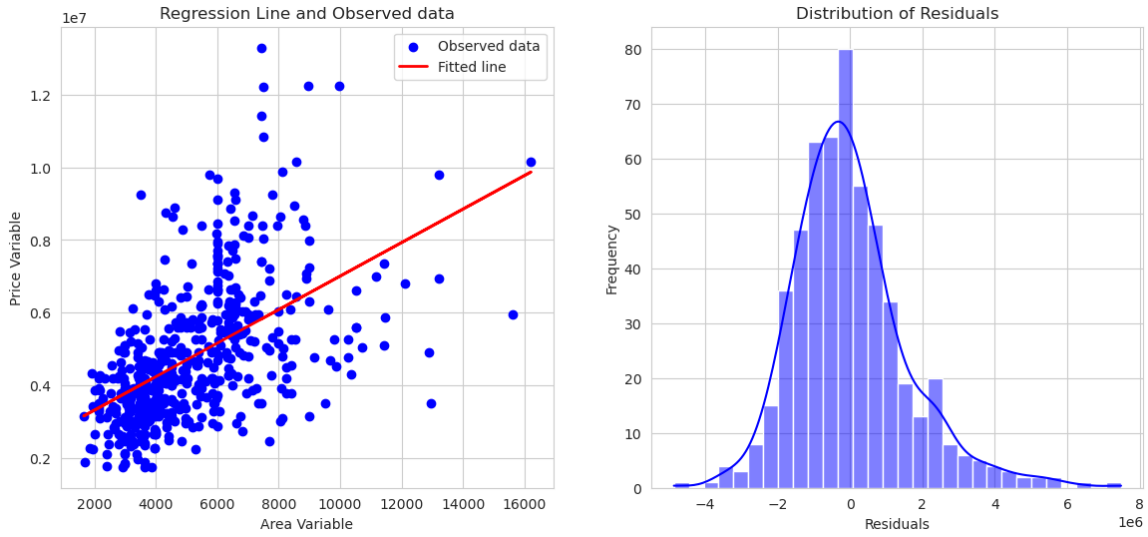

Figure 14: Fitted line between ”area” and ”price”

We must check if the residuals are coming from normal distribution or not, since normality of residuals is one of the key assumptions for optimal performance and validity of statistical tests related to the model. In order to check this we consider conduct Kolmogorov-Smirnov test.The Kolmogorov-Smirnov test is a non-parametric approach to see if data comes from particular distribution or not. The procedure of the KS test is as follows:

- Null Hypothesis is:

$$\begin{aligned} H_0 &: \text{The samples come from P distribution} \\ H_1 &: \text{The samples do not come from P} \end{aligned} \tag{28}$$

- The statistic:

$$D_n = \max_x |F_{exp}(x) - F_{obs}(x)| \tag{29}$$

- If D_n exceeds the critical value we will reject the null hypothesis.

The result of the KS test is:

Table 20: KS Result

Statistic	P-value
0.084	0.000807389239895906

Therefore we may reject null hypothesis for 0.05 significant level.We shall consider the transforming data like boxcox transformation and etc.
The OLS result for transformed ”price” and ”area” are:

Table 21: OLS result for transformed data

	coef	std err	t	P>|t|	[0.025 , 0.975]
const	5.4501	0.062	87.823	0.000	5.328 , 5.572
area	0.1962	0.012	16.603	0.000	0.173 , 0.219

The result interpretation for boxcox transformated ”price” and ”area”:

- R-square: 0.337 indicates that approximately 33.7% of the variance in the dependent variable ”price” can be explained by the independent variable ”area”. And also we can see this metric has improved from the raw data where was just about 28.7% of the variance in price.
- The F-statistic value is 275.6 with a very low p-value (practically 0, indicated by 0.000), suggesting that the model is statistically significant and that the explained variance is not due to chance.
- Coef for ”area”: The coefficient for ”area” is 0.1962, meaning that for every one-unit increase in ”area”, the dependent variable ”price” is expected to increase by an average of 0.1962 units.

In summary, the model is statistically significant and the independent variable ”area” is a significant predictor of the dependent variable ”price”. (Note that in here we are speaking of transformed data not the real ”price” and the real ”area”).
The residuals distirbution and fitted line for transformated data are depicted in the figure 15:

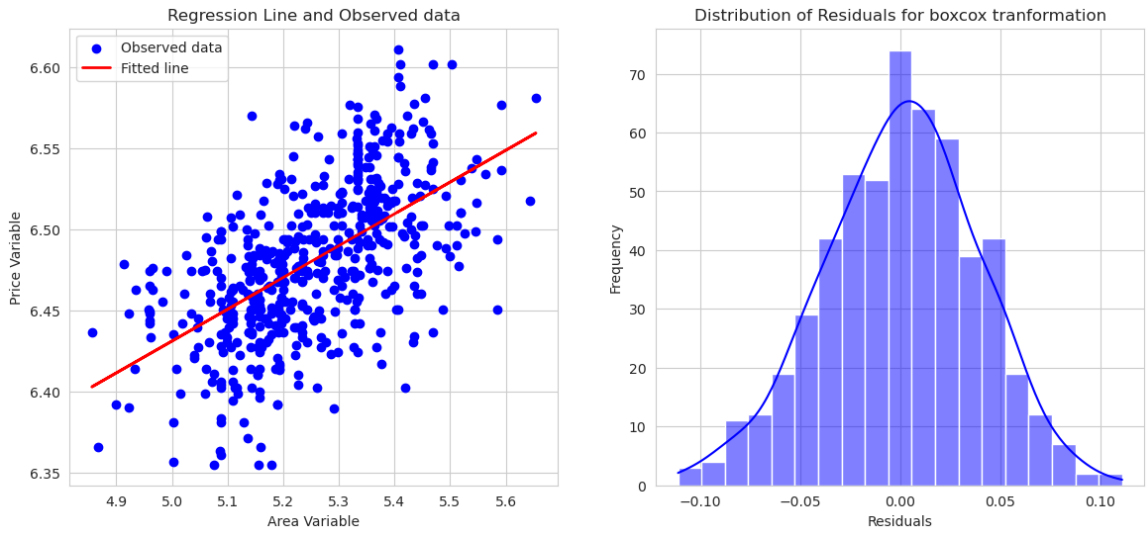

Figure 15: Fitted line between transformed ”area” and ”price”

The p-value(0.7908622855412032) of Kolmogorov-Smirnov test also indicates that the residuals are normaly distributed.

5.2 Bootstrap Approach

We know from MLE that $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_xx})$ where S_{xx} is:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{30}$$

And we can estimate σ^2 as follows:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \tag{31}$$

Where e_i's are the residuals.Now we can use bootstrap method to see if the distibution of slope is normal or not.We will simulate 100000 times where in each we get 30 sample and fit regression line and captures the slop of fitted line where the distribution of slopes are in figure 16:

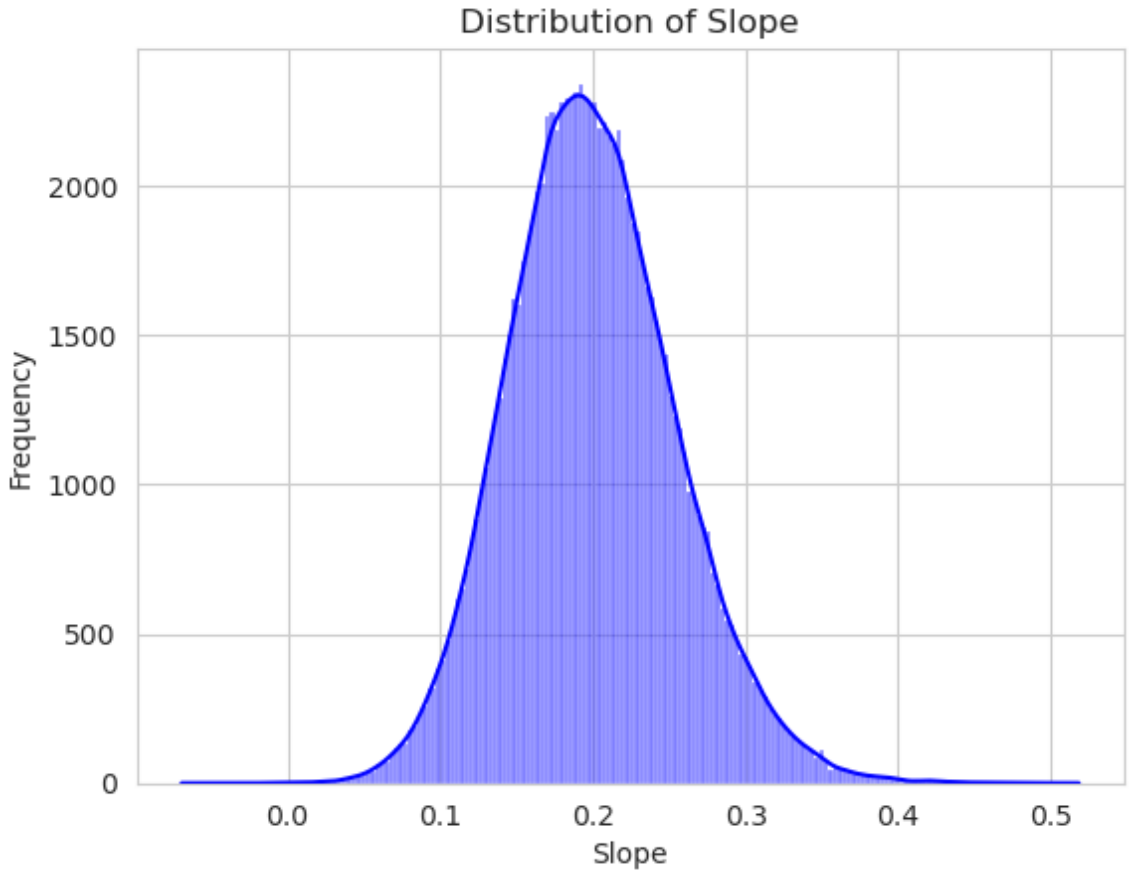

Figure 16: Distribution of β_1

As illustrated in figure 16 distribution of slope is symmetric around 0.2 where the real value of slope was also 0.2! The emprical std and mean of slopes are in table 22:

Table 22: Mean and std

Mean	STD
0.199	0.054

From calculations we have:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\ &= 2.0891413254387468 \end{aligned} \tag{32}$$

Therefore the theoretical std is:

$$\begin{aligned} \text{estimated std for } \beta_1 &= \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \\ &= 0.02667738590180805 \end{aligned} \tag{33}$$

As you can see the emprical std and theoretical std are so close to each other.We can also conduct confidence interval as well (We will use percentile interval for CI). 95% CI for β_1 is:

$$\text{CI} = (0.09874744, 0.3140147) \tag{34}$$

As you can see, the real slope is 0.1962 where CI will cover the real β_1(slope).

5.3 Regression With Two variables

We know from the correlation matrix that ”bathrooms” and ”price” variable are correlated to each other as well.Also the coefficient between (price,area) and (price, bathrooms) approximately equal to one another.So we can fit a regression line between the (”area”, ”bathrooms”) as explanatory variables and the ”price” as dependent variable. The procdeure of the multiple regression model in matrix notation:

$$\mathrm{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The y_i's their observed values andϵ_i's random errors.Respectively for all n observations:

$$\mathrm{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

the (k + 1) x 1 vectors of unknown model parameters and their LS estimates:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

If the inverse of the matrix $X'X^{-1}$ exists, then the solution is given by:(X' denotes X^T)

$$\hat{\beta} = (X'X)^{-1}X'y \tag{35}$$

Now we can fit a plane to our 3D model, where the result can be seen in table 23:

Table 23: Result of 3D regression model for (”boxcox-area”,”bathrooms”) and ”boxcox-price”

	coef	std err	t	P>|t|	[0.025 , 0.975]
const	5.5397	0.056	99.133	0.000	5.430 , 5.649
boxcox-area	0.1704	0.011	15.831	0.000	0.149 , 0.192
bathrooms	0.0355	0.003	11.840	0.000	0.030 , 0.041

The interpretation of the result:

1. coef (Coefficient):
   - ”Const”: 5.5397, this is the boxcox-price-intercept, meaning when ”boxcox-area” and ”bathrooms” are 0, the expected value of ”boxcox-price” is approximately 5.5397.
   - ”Boxcox-area”: 0.1704, this implies that for each unit increase in ”boxcox-area”, the ”boxcox-price” is expected to increase by 0.1704 units, holding all else constant.

- ”Bathrooms”: 0.0355, for each additional ”bathroom”, the ”boxcox-price” increases by 0.0355 units, holding all else constant.

2. Standard Error: Reflects the average distance that the observed values fall from the regression line. For ”boxcox-area”, the standard error is 0.011, and for bathrooms, it is 0.003. Smaller standard errors suggest more precise estimates.

3. P-value: Indicates the probability of observing the data or something more extreme if the null hypothesis is true. Here, p-values for both ”boxcox-area” and ”bathrooms” are 0.000, suggesting strong evidence against the null hypothesis, leading to the conclusion that these coefficients are significantly different from zero.

4. Confidence Interval:

   - For const, the 95% CI ranges from 5.430 to 5.649, meaning we are 95% confident that the true value of the intercept lies within this range.
   - For ”boxcox-area”, the 95% CI is from 0.149 to 0.192.
   - For bathrooms, the 95% CI is from 0.030 to 0.041.

5. Prob(JB) (Jarque-Bera Test Probability): At 0.068, this is the p-value for the Jarque-Bera test, which tests the null hypothesis that the residuals are normally distributed. A value greater than 0.05 typically suggests that we fail to reject the null hypothesis, indicating normal distribution of residuals.

6. R-squared:At 0.473, this value indicates that approximately 47.3% of the variability in ”boxcox-price” can be explained by the model. It is a measure of the goodness of fit.(Where we fitted regression line in previous part this metric was about 33.7% therefore this 3D model is more effient than just a regression line between ”boxcox-area” and ”boxcox-price”!)

The regression model alonside observed values are and residuals has been illustrated in figure 17:

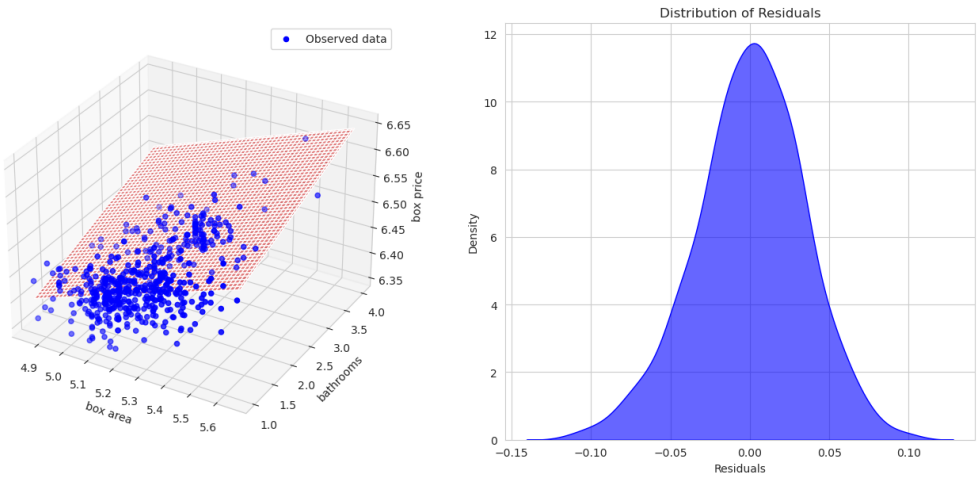

Figure 17: Residuals and fitted model

5.4 Regression Model With All Variabls

In the process of constructing a regression model, it is imperative to work with numerical data. However, categorical variables such as ”guestrooms” and ”furnishing status” may also influence on the dependent variable ”price”. To incorporate these categorical variables into the regression analysis, it is necessary to convert them into numerical form. This conversion enables the evaluation of their significance in the model, which can be determined by examining the p-values associated with these transformed categorical predictors.
So we will encode the categorical predictors into numerical form and use them in regression analysis. The result of OLS is in table 24:

Table 24: Result of regression model with all data

	coef	std err	t	P>|t|	[0.025 , 0.975]
const	5.8119	0.049	119.666	0.000	5.716 , 5.907
bedrooms	0.0042	0.002	2.325	0.020	0.001 , 0.008
bathrooms	0.0198	0.003	7.774	0.000	0.015 , 0.025
stories	0.0113	0.002	7.132	0.000	0.008 , 0.014
parking	0.0052	0.001	3.631	0.000	0.002 , 0.008
boxcox-area	0.1088	0.010	11.414	0.000	0.090 , 0.128
encode-furnishingstatus	0.0080	0.002	5.128	0.000	0.005 , 0.011
encode-prefarea	0.0164	0.003	5.765	0.000	0.011 , 0.022
encode-airconditioning	0.0197	0.003	7.349	0.000	0.014 , 0.025
encode-hotwaterheating	0.0216	0.006	3.932	0.000	0.011 , 0.032
encode-guestroom	0.0065	0.003	1.997	0.046	0.000 , 0.013
encode-mainroad	0.0126	0.004	3.529	0.000	0.006 , 0.020
encode-basement	0.0129	0.003	4.719	0.000	0.008 , 0.018

Some of important results obtained from the OLS:

1. Coefficient: This represents the estimated change in the dependent variable (”boxcox-price”) for a one-unit change in the predictor variable, holding all other variables constant.
    - Const: The intercept (5.8119) is the expected value of ”boxcox-price” when all other variables are zero.
    - For the other variables, such as ”bedrooms”, ”bathrooms”, ”stories”, etc., their coefficients indicate how much the ”boxcox-price” is expected to increase (or decrease, if the coefficient were negative) with each additional unit of these variables.
2. p-value:This value tests the null hypothesis that a coefficient is equal to zero (no effect). A low p-value (< 0.05) suggests that you can reject the null hypothesis and that the corresponding variable is statistically significant in predicting ”boxcox-price”.
    - For example, ”bedrooms” has a p-value of 0.020, which is less than the standard 0.05 threshold, suggesting that the number of ”bedrooms” is a statistically significant predictor of the ”boxcox-price”.
    - Variables like encode-guestroom with a p-value of 0.046 are also statistically significant but close to the threshold, indicating a weaker evidence of its impact on ”boxcox-price”.

3. R-squared: At 0.696, this value indicates that approximately 69.6% of the variability in ”boxcox-price” is explained by the model. It is a measure of the model’s goodness of fit where it is a very good fit against previous models.

4. Standard Error: This measures the average amount that the coefficient estimates vary from the actual average value of the dependent variable. A smaller standard error suggests that the estimate of the coefficient is more precise.

    - For instance, ”const” has a standard error of 0.049, which is quite small relative to the coefficient itself, suggesting that the estimate of the intercept is precise.

In conclusion, the model explains substantial portion of the variance in ”boxcox-price”, with several predictor showing statistical significance in affecting the ”price”. Variables with p-values and relatively low standard errors are likely to be reliable predictors in the model.

5.5 Comprehensive Report

The comprehensive analysis outlined in the provided sections offers a holistic overview of the multifaceted approach used to understand and predict house prices

5.5.1 Executive Summary

This report combines classical statistical methods and machine learning techniques to analyze the factors influencing house prices. Through a series of sophisticated statistical analyses, it provides actionable insights for stakeholders in the real estate sector, emphasizing the value of data-driven decision-making.

5.5.2 Introduction

The report sets the stage by highlighting the importance of predicting house prices, not only for individual buyers and sellers but also for broader economic planning and policy-making. It outlines the objectives: to identify significant predictors of house prices and assess the predictive power of various statistical and machine learning models.

5.5.3 Methodology

A detailed methodology section explains the array of statistical tools and visualization techniques employed:

- Data Visualization: Utilization of histogram analysis, box plots, scatter plots, Q-Q plots, and categorical visualization to explore data distribution and relationships.
- Statistical Testing: Application of Shapiro-Wilk tests, t-tests, and ANOVA to test for normality and significance of variables.
- Regression Analysis: Implementation of regression analyses, including MLE for parameter estimation and bootstrapping methods for assessing estimate stability.
- Machine Learning: Deployment of machine learning algorithms, including Catboost Regression(On the paper), to predict house prices effectively.

5.5.4 Results and Analysis

This section presents the core findings from the statistical tests and predictive models:

- Identification of key variables significantly impacting house prices, such as area, number of bathrooms, and the presence of a basement.
- Evaluation of machine learning models, highlighting the effectiveness of specific algorithms in predicting house prices.(Paper and my Report)

5.5.5 Discussion and Implications

The report discusses the implications of these findings, suggesting practical applications for various stakeholders:

- For Homebuyers and Sellers: Insights into how specific features affect property values.
- For Investors and Market Strategists: Data-driven foundation for assessing market trends.
- For Policymakers: Understanding of housing market dynamics to inform regulations and development strategies.

5.5.6 Conclusions and Future Directions

It concludes by synthesizing the report's findings, underlining the significant portion of variance in house prices explained by the models. The report calls for future research to explore these variables in different contexts and further refine predictive models, contributing to the field of real estate economics.

5.5.7 Significance for Stakeholders

The comprehensive analysis underscores the necessity of incorporating data-driven methodologies in real estate economics, providing a basis for informed decision-making and highlighting the role of significant predictors in housing price determination.
Attention:The concept of generalizability is crucial in research. Given the unspecified origin of the sample, conclusions drawn from the analysis can only be tentatively applied to the sample's demographic. To enhance generalizability, it is advisable to source a diverse array of samples across different regions or countries. This would allow for a more comprehensive prediction of house prices influenced by a variety of factors. Incorporating a broader spectrum of data could potentially refine the general applicability of the findings.