CAPSTONE PROJECT

House Price Prediction

Abstract

This comprehensive analysis of housing data provides valuable insights and actionable recommendations for real estate stakeholders. Starting with exploratory data analysis (EDA), the study uncovers feature importance, market segmentation, and key factors influencing house prices. Univariate and bivariate analyses reveal intriguing patterns, such as the lack of a strong relationship between quality houses and coastal proximity. Further, the correlations between living area and ceiling area, lot sizes, and lot sizes in 2015 offer opportunities for tailored marketing strategies.

In the modelling phase, Random Forest and Boost models stand out for their accuracy, suggesting their adoption for price predictions. Ensemble modelling and hyperparameter tuning are suggested for enhanced performance. Continual monitoring and model maintenance are essential for reliable predictions over time.

These data-driven recommendations empower real estate professionals to make informed decisions regarding property pricing, market segmentation, and predictive modelling, ultimately improving customer satisfaction and business success.

Sudheendra K
Sudhi0404@gmail.com
PGPDSBA-OCT-B-22

Capstone Project

Contents

1) Introduction

a) Brief introduction about the problem statement

The problem statement is well-defined in the project description. It revolves around the need to accurately predict house prices using a variety of features beyond just location and square footage. The problem can be summarized as follows: "To predict house prices accurately, taking into account numerous properties features and characteristics."

b) The need of solving problem

The need for this study/project is clear. House prices are influenced by a multitude of factors, and having an accurate prediction model can benefit various stakeholders such as homeowners looking to sell, buyers trying to assess fair prices, and real estate professionals. This project addresses the need for a robust house price prediction model that incorporates diverse features to provide accurate valuations.

c) Understanding business/social opportunity.

The business/social opportunity here is substantial. Accurate house price predictions are essential for making informed real estate decisions. This project can benefit individuals

looking to buy or sell homes, real estate agents aiming to offer precise valuations, and potentially even policy-makers interested in understanding housing market dynamics. Additionally, it presents an opportunity to leverage data science and machine learning to solve a real-world problem with a potentially significant impact on people's lives.

Overall, the introduction effectively defines the problem, articulates the need for the study, and highlights the business and social opportunities associated with solving it.

d) Understanding how data was collected in terms of time, frequency and methodology.

The data pertains to house price prediction and includes information about the time range, total data points, and the distribution of data points across years and months. Let's break down the key details:

- *Time Range: The data was collected over a period from May 2, 2014, at 00:00:00 to May 27, 2015, at 00:00:00.*
- *Total Data Points: There are a total of 21,613 data points in our dataset.*
- *Data Points per Year: The data is divided into two years: 2014 and 2015. Here is the distribution of data points across these years:*
    1. *2014: 14,633 data points*
    2. *2015: 6,980 data points*
- *Data Points per Month: The data is further divided into months. Here is the distribution of data points across months:*
    1. *January (1): 978 data points*
    2. *February (2): 1,250 data points*
    3. *March (3): 1,875 data points*
    4. *April (4): 2,231 data points*
    5. *May (5): 2,414 data points*
    6. *June (6): 2,180 data points*
    7. *July (7): 2,211 data points*
    8. *August (8): 1,940 data points*
    9. *September (9): 1,774 data points*

10. October (10): 1,878 data points

11. November (11): 1,411 data points

12. December (12): 1,471 data points

Data Points per Year and Month: This table provides a more detailed breakdown of data points by both year and month. It shows how many data points are available for each combination of year and month. For example, in May 2014, there were 1,768 data points, while in January 2015, there were 978 data points.

This information is essential for understanding the temporal distribution of our dataset. It can be useful for time-series analysis and for making inferences about house price trends over this specific time period

e) Visual inspection of data (rows, columns, descriptive details).

Data Overview:

	cid	dayhours	price	room_bed	room_bath	living_measure	lot_measure	ceil	coast	sight	condition	quality	ceil_measure
0	3.876101e+09	20150427T000000	600000.0	4.0	1.75	3050.0	9440.0	1.0	0.0	0.0	3.0	8.0	1800.0
1	3.145600e+09	20150317T000000	190000.0	2.0	1.00	670.0	3101.0	1.0	0.0	0.0	4.0	6.0	670.0
2	7.129303e+09	20140820T000000	735000.0	4.0	2.75	3040.0	2415.0	2.0	1.0	4.0	3.0	8.0	3040.0
3	7.338220e+09	20141010T000000	257000.0	3.0	2.50	1740.0	3721.0	2.0	0.0	0.0	3.0	8.0	1740.0
4	7.950301e+09	20150218T000000	450000.0	2.0	1.00	1120.0	4590.0	1.0	0.0	0.0	3.0	7.0	1120.0

Snippet 1. 1 Head of the dataset

- ***The dataset consists of 21,613 rows and 23 columns.***

	cid	dayhours	price	room_bed	room_bath	living_measure	lot_measure	ceil	coast	sight	condition	quality	ceil_measure
21608	2.036006e+08	20150310T000000	685530.0	4.0	2.50	3130.0	60467.0	2.0	0.0	0.0	3.0	9.0	3130.0
21609	6.250493e+08	20140521T000000	535000.0	2.0	1.00	1030.0	4841.0	1.0	0.0	0.0	3.0	7.0	920.0
21610	4.240690e+08	20140905T000000	998000.0	3.0	3.75	3710.0	34412.0	2.0	0.0	0.0	3.0	10.0	2910.0
21611	7.258200e+09	20150206T000000	262000.0	4.0	2.50	1560.0	7800.0	2.0	0.0	0.0	3.0	7.0	1560.0
21612	8.805900e+09	20141229T000000	1150000.0	4.0	2.50	1940.0	4875.0	2.0	0.0	0.0	4.0	9.0	1940.0

Snippet 1. 2 Tail of the dataset

- ***The columns include both numerical and categorical data.***

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 23 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   cid               21613 non-null  float64
 1   dayhours          21613 non-null  object
 2   price             21613 non-null  float64
 3   room_bed          21505 non-null  float64
 4   room_bath         21505 non-null  float64
 5   living_measure    21596 non-null  float64
 6   lot_measure       21571 non-null  float64
 7   ceil              21571 non-null  object
 8   coast             21612 non-null  object
 9   sight             21556 non-null  float64
 10  condition         21556 non-null  object
 11  quality           21612 non-null  float64
 12  ceil_measure      21612 non-null  float64
 13  basement          21612 non-null  float64
 14  yr_built          21612 non-null  object
 15  yr_renovated      21613 non-null  float64
 16  zipcode           21613 non-null  float64
 17  lat               21613 non-null  float64
 18  long              21613 non-null  object
 19  living_measure15  21447 non-null  float64
 20  lot_measure15     21584 non-null  float64
 21  furnished         21584 non-null  float64
 22  total_area        21584 non-null  object
dtypes: float64(16), object(7)
memory usage: 3.8+ MB
```

Snippet 1. 3 Dataset Info

Column Information:

- ***The key columns include 'cid' (a unique identifier), 'dayhours' (date and time of the data point), and 'price' (house price).***
- ***There are columns that represent various features related to the house, such as the number of bedrooms ('room_bed'), number of bathrooms ('room_bath'), living area ('living_measure'), lot size ('lot_measure'), and more.***
- ***Some columns contain categorical data, such as 'ceil' (ceiling type), 'coast' (proximity to the coast), 'condition' (house condition), and 'yr_built' (year built).***

Data Types:

- ***Most of the columns have numerical data types (e.g., float64), while some are of type object (likely indicating mixed data types or string values).***

Missing Values:

- ***Some columns have missing values. For example, 'room_bed', 'room_bath', 'living_measure', 'lot_measure', and others have missing data points.***

Descriptive Statistics (Numerical Columns):

Price (House Price):

- ***The mean house price is approximately $540,182, while the median price is $450,000. This suggests that the distribution of house prices is right-skewed, as the mean is greater than the median.***
- ***The interquartile range (IQR) is $323,050, indicating a wide spread of prices between the 25th and 75th percentiles.***
- ***There are extreme values in the dataset, as evidenced by the large maximum value of $7,700,000.***

Living Measure (Living Area of the House):

- ***The mean living area of houses is approximately 2,080 square feet, with a median of 1,910 square feet.***
- ***The IQR for living area is 1,120.75 square feet, indicating variability in the sizes of houses.***
- ***There are houses with very small living areas (minimum of 290 square feet) and some with very large living areas (maximum of 13,540 square feet).***

Lot Measure (Lot Size):

- ***The mean lot size is approximately 15,104 square feet, with a median of 7,618 square feet.***
- ***The IQR for lot size is 5,644.50 square feet, indicating variability in the sizes of lots.***
- ***There are lots with small sizes (minimum of 520 square feet) and some with very large sizes (maximum of 1,651,359 square feet).***

Ceil Measure (Ceiling Area of the House):

- ***The mean ceiling area is approximately 1,788 square feet, with a median of 1,560 square feet.***
- ***The IQR for ceiling area is 1,020 square feet.***
- ***Like living area, there are houses with a wide range of ceiling areas, from small to large.***

Basement Area:

- ***The mean basement area is approximately 292 square feet, but the median is 0 square feet, indicating that many houses have no basement.***
- ***The IQR for basement area is 560 square feet.***
- ***Some houses have large basements (maximum of 4,820 square feet), while many have none.***

Living Measure 15 (Living Area in 2015):

- ***The mean living area in 2015 is approximately 1,987 square feet, with a median of 1,840 square feet.***
- ***The IQR for living area in 2015 is 870 square feet.***
- ***This column likely represents changes in living area over time, with some houses expanding their living spaces.***

Lot Measure 15 (Lot Size in 2015):

- ***The mean lot size in 2015 is approximately 12,767 square feet, with a median of 7,620 square feet.***
- ***The IQR for lot size in 2015 is 4,987 square feet.***
- ***Similar to living area, this column represents changes in lot sizes over time.***

Total Area (Total Area of House and Lot):

- ***The mean total area is approximately 17,192 square feet, with a median of 9,575 square feet.***
- ***The IQR for total area is 5,968 square feet.***
- ***This column likely combines both living and lot sizes to represent the overall property size.***

These descriptive statistics provide a snapshot of the data's central tendency, spread, and potential outliers for key features related to house prices. Understanding these statistics can help guide further analysis and modelling efforts, such as identifying influential features or outliers that may need special consideration in predictive modelling.

Categorical Columns:

- ***Categorical columns like 'ceil', 'coast', 'condition', and 'yr_built' may need encoding or transformation for use in predictive modelling.***
- ***I'll address this in univariate analysis in detailed.***

f) Understanding of attributes (variable info, renaming if required).

- ***it's essential to review and potentially rename some columns for clarity and consistency. But as far as I don't want to get into renaming because I've understood the dataset.***

2. EDA and Business Implication

a) How my analysis is impacting the business?

Before proceeding with any analysis or modelling, we should consider addressing missing values, converting data types as needed, and preparing the data for our specific objectives, such as house price prediction. We must also want to perform exploratory data analysis (EDA) to gain more insights into the data distribution and relationships between variables

Visual Analysis

a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones).

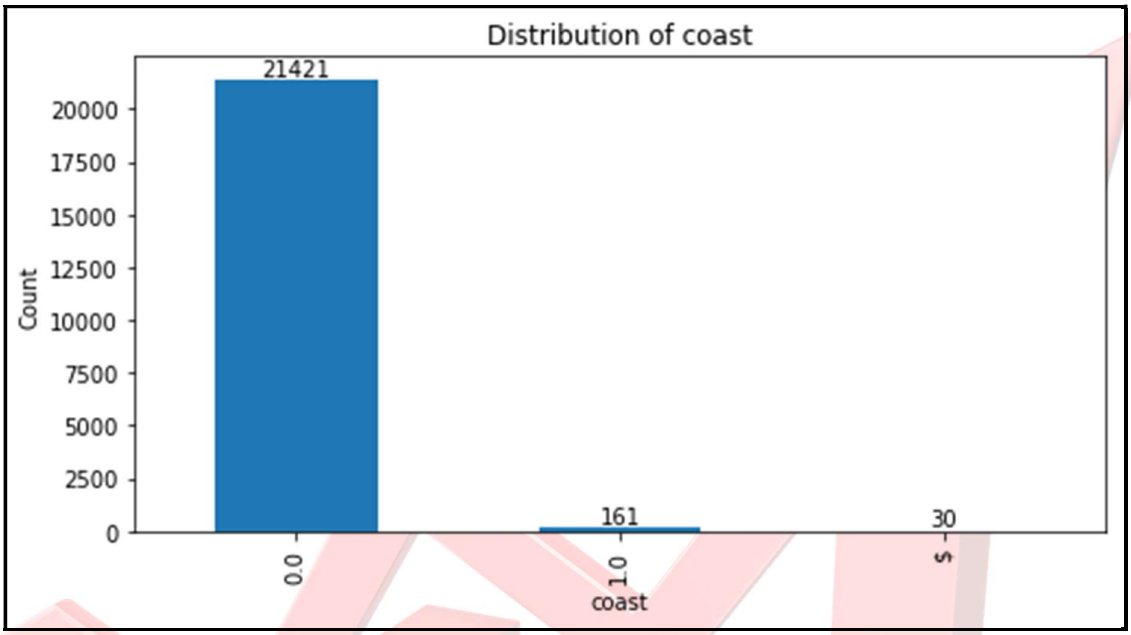

Figure 1 Distribution of Coast

'coast' (Proximity to Coast):

- ***Unique Values Count: 3***
- ***Most houses have a value of 0.0 (far from the coast), with a smaller number of houses having values of 1.0 and a few with '$' (unusual value).***

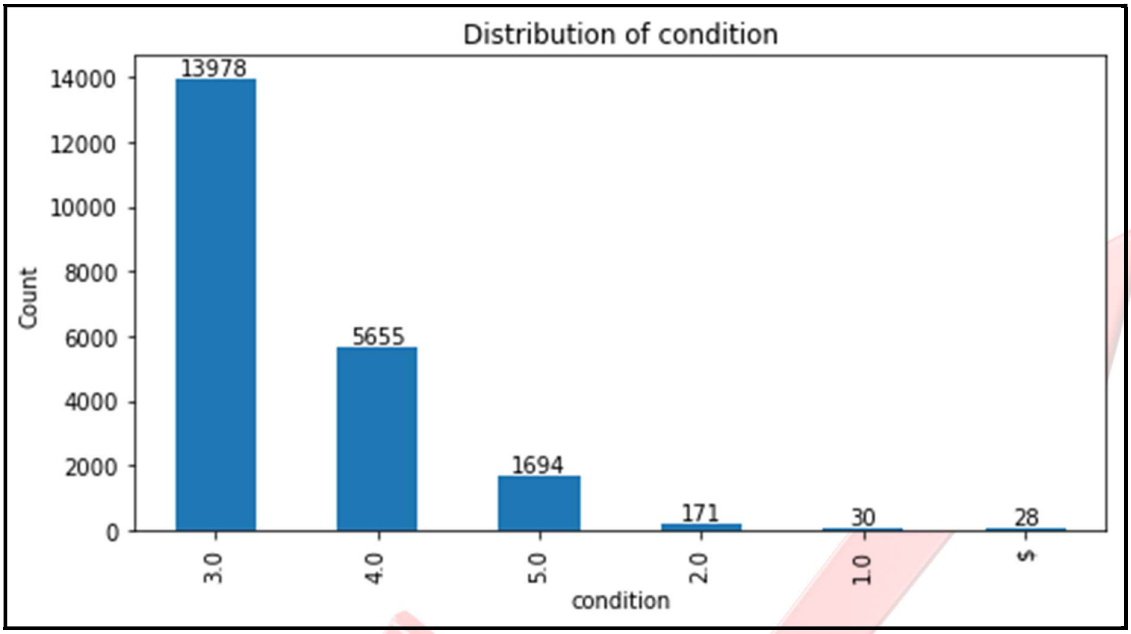

Figure 2 Distribution of condition

'condition' (House Condition):

- ***Unique Values Count: 6***
- ***Most houses are in condition 3.0, followed by 4.0. There are relatively few houses with conditions 1.0, 2.0, and '$' (unusual value).***

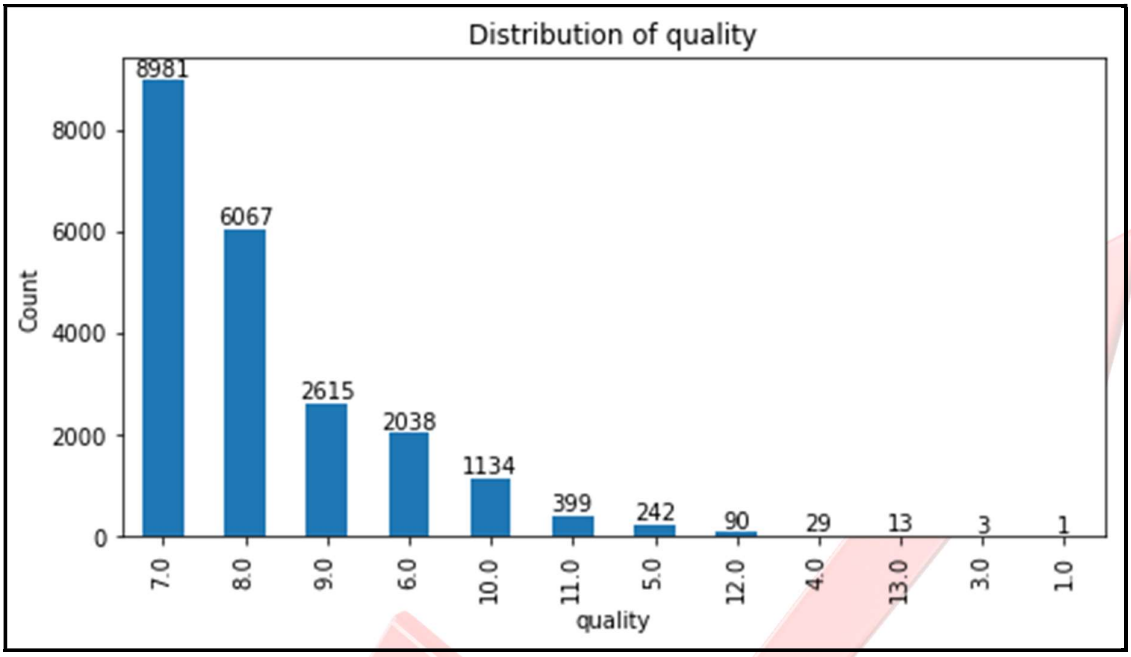

Figure 3 Distribution of quality

'quality' (House Quality):

- ***Unique Values Count: 12***
- ***The majority of houses have quality ratings of 7.0 and 8.0, followed by 9.0 and 6.0. There are relatively few houses with quality ratings below 6.0 or above 9.0***

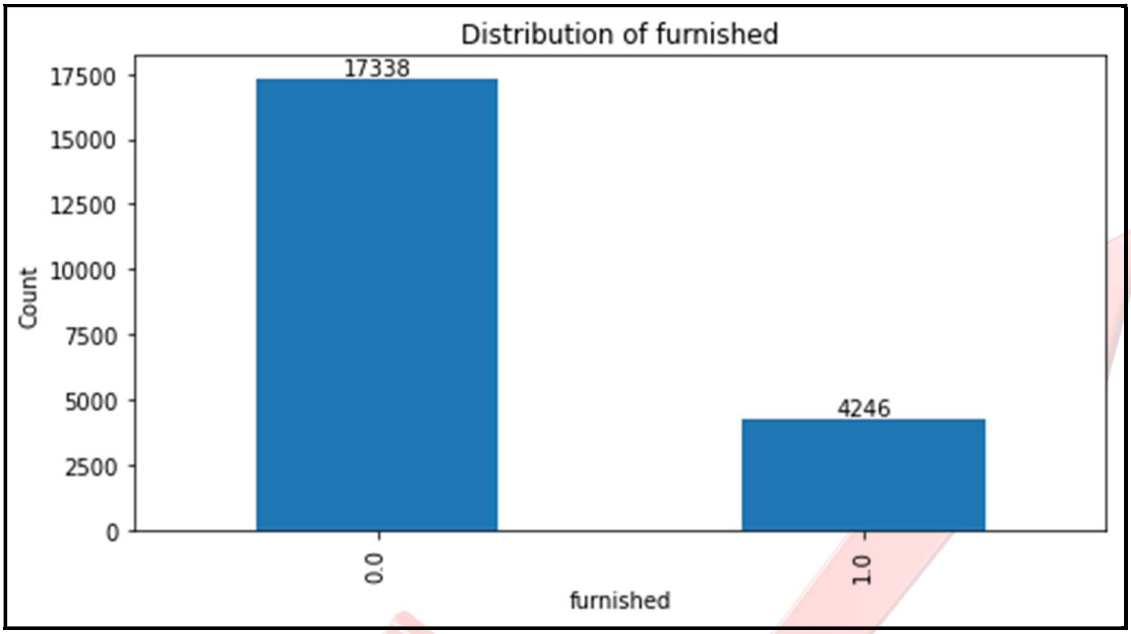

Figure 4 Distribution of Furnished

'furnished' (Furnished):

- ***Unique Values Count: 2***
- ***Most houses have a value of 0.0 (not furnished), while a smaller number are furnished (1.0).***

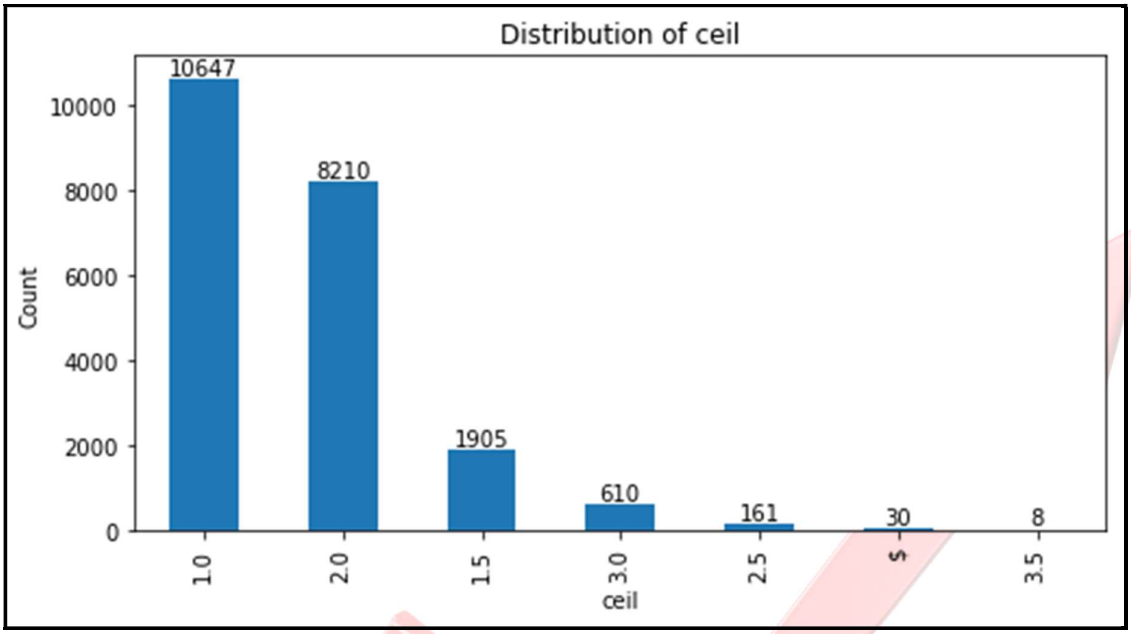

Figure 5 Distribution of ceil

'ceil' (Ceiling Type):

- ***Unique Values Count: 7***
- ***Most houses have ceiling types of 1.0 (single level) and 2.0 (two levels). There are smaller numbers of houses with other ceiling types, including 1.5, 3.0, 2.5, '$' (unusual value), and 3.5.***

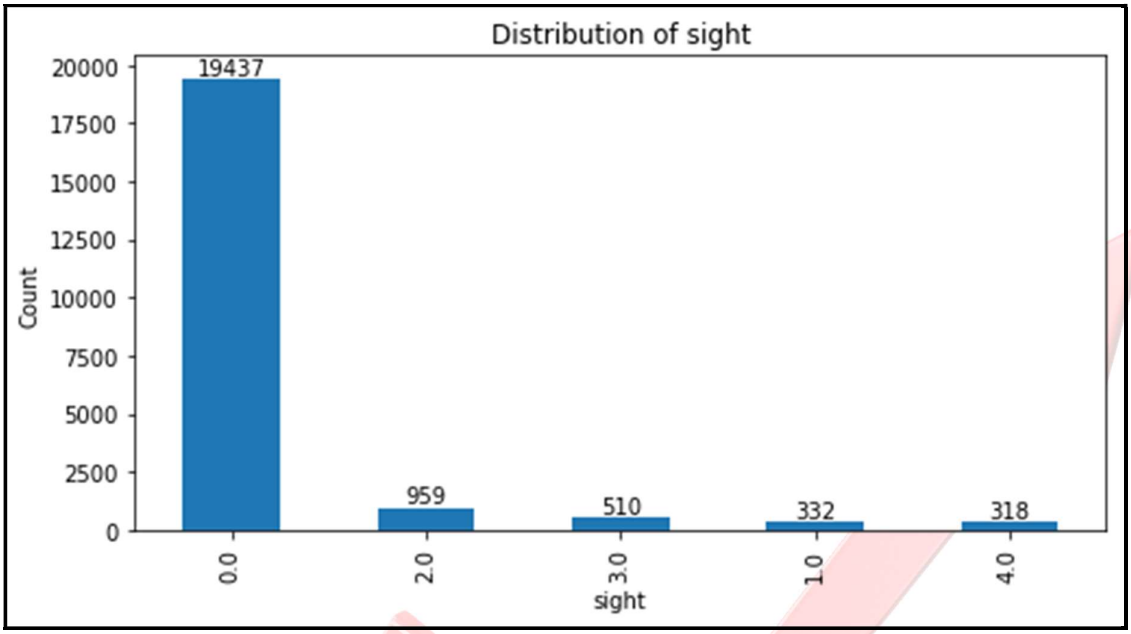

Figure 6 Distribution of sight

'sight' (Sight Rating):

- ***Unique Values Count: 5***
- ***The majority of houses have a sight rating of 0.0 (no special sight), followed by 2.0 and 3.0. There are smaller numbers of houses with sight ratings of 1.0 and 4.0.***

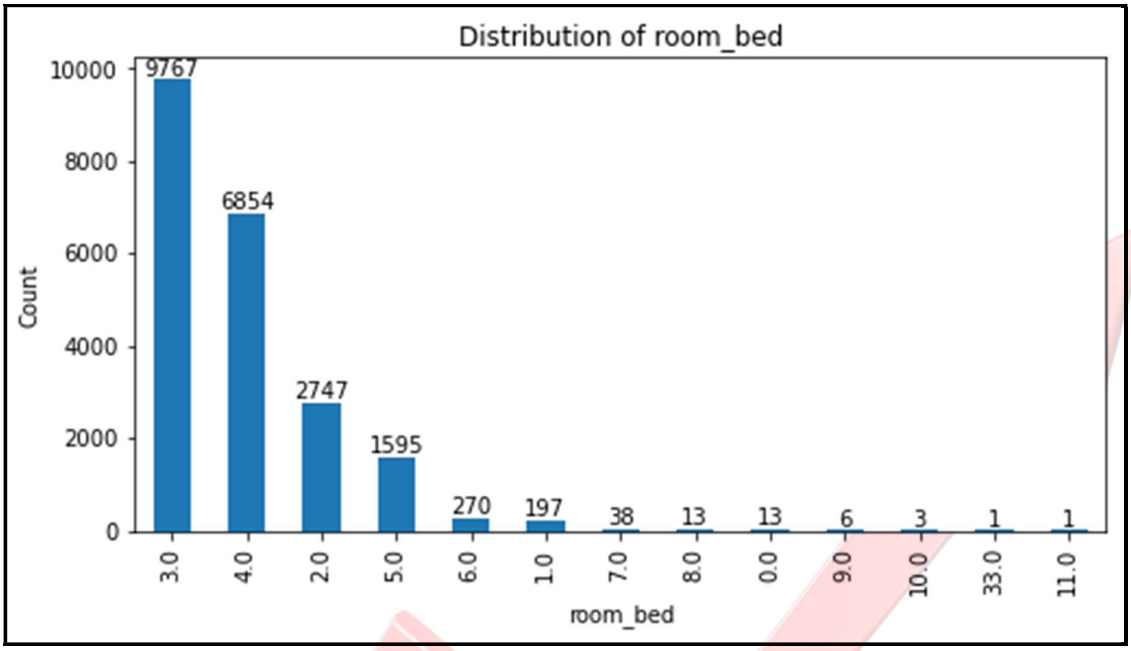

Figure 7 Distribution of No. of Bedrooms

'room_bed' (Number of Bedrooms):

- ***Unique Values Count: 13***
- ***Most houses have 3 bedrooms, followed by 4 bedrooms. There are relatively fewer houses with other bedroom counts, and a few with unusual values such as 33.0 and 11.0.***

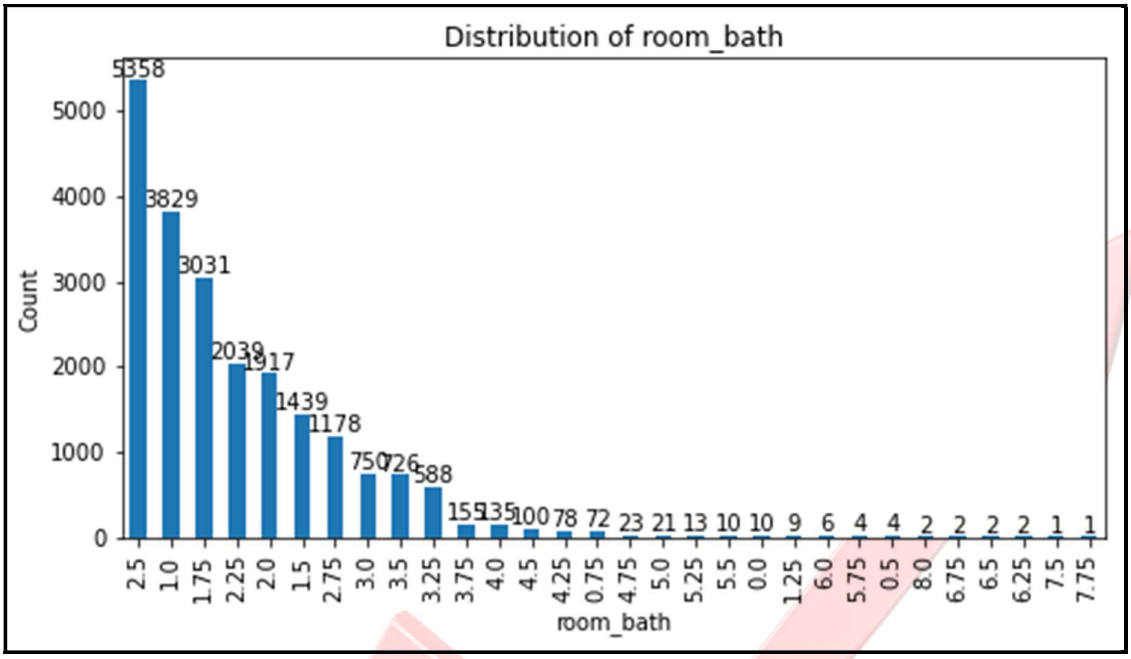

Figure 8 Distribution of No. of bathrooms

'room_bath' (Number of Bathrooms):

- ***Unique Values Count: 30***
- ***Houses commonly have 2.50 and 1.00 bathrooms. There are various bathroom counts, with some houses having unusual values such as 0.00, 8.00, 7.50, and others.***

Continuous Attributes:

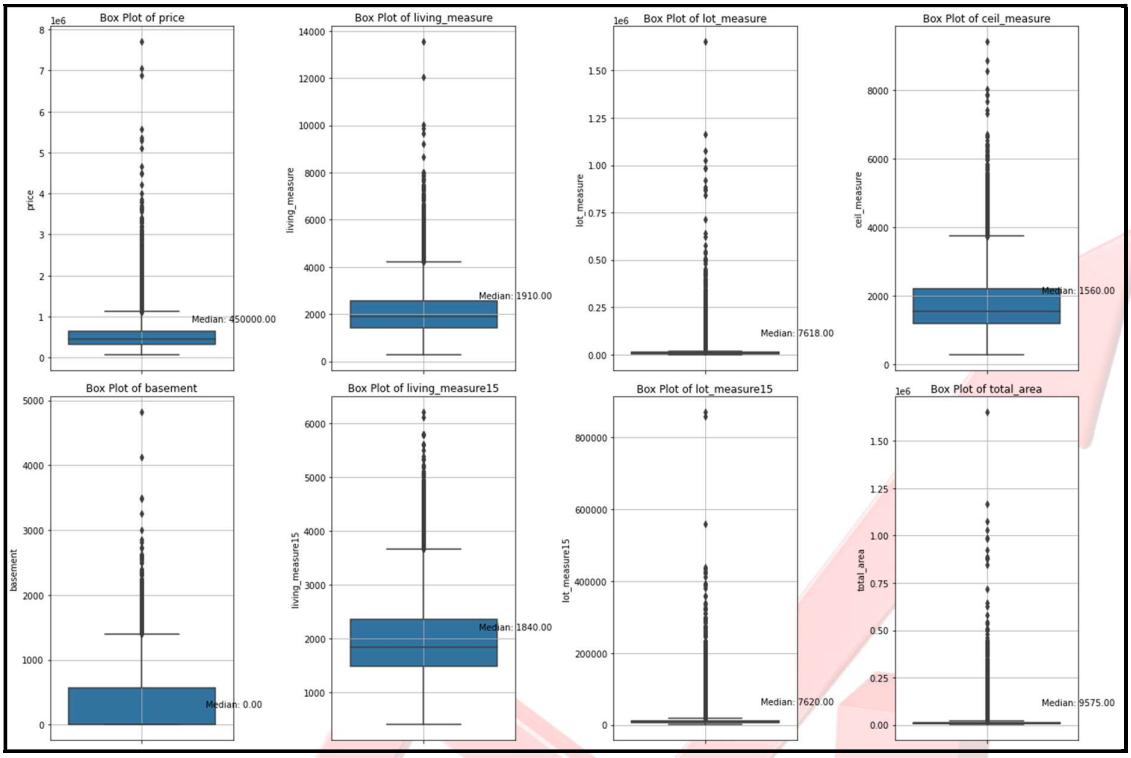

Figure 9 Box plots for all the numerical attributes

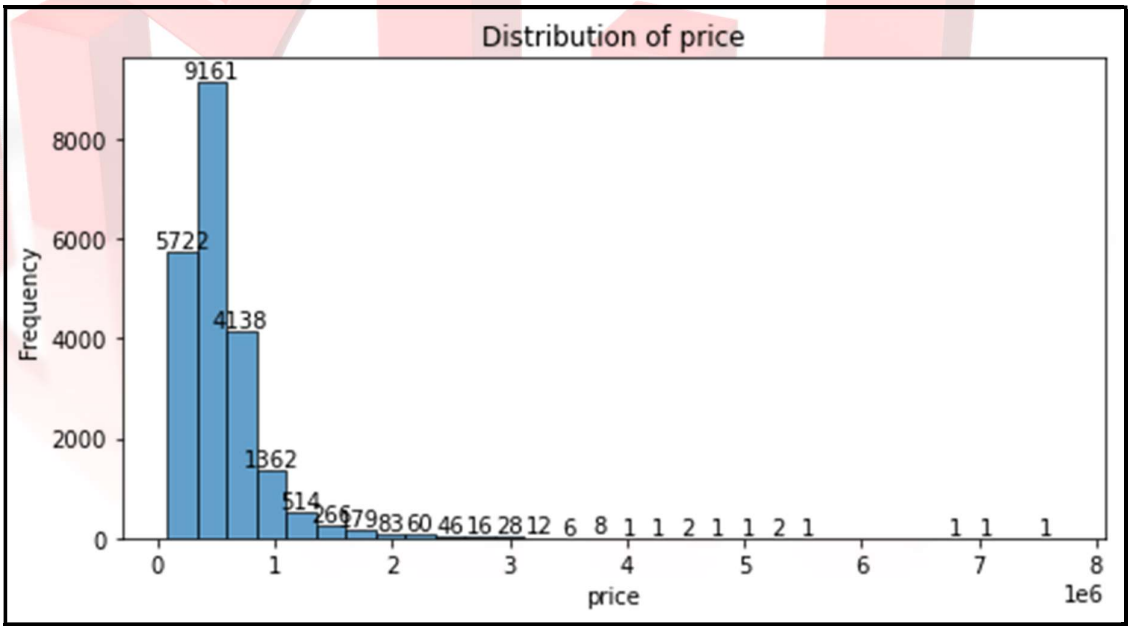

Figure 10 Distribution of price

'price' (House Price):

- ***Distribution: The distribution of house prices appears to be right-skewed, with a long tail towards higher prices. This is evident from the mean being greater than the median.***
- ***Spread: The prices range from a minimum of $75,000 to a maximum of $7,700,000.***

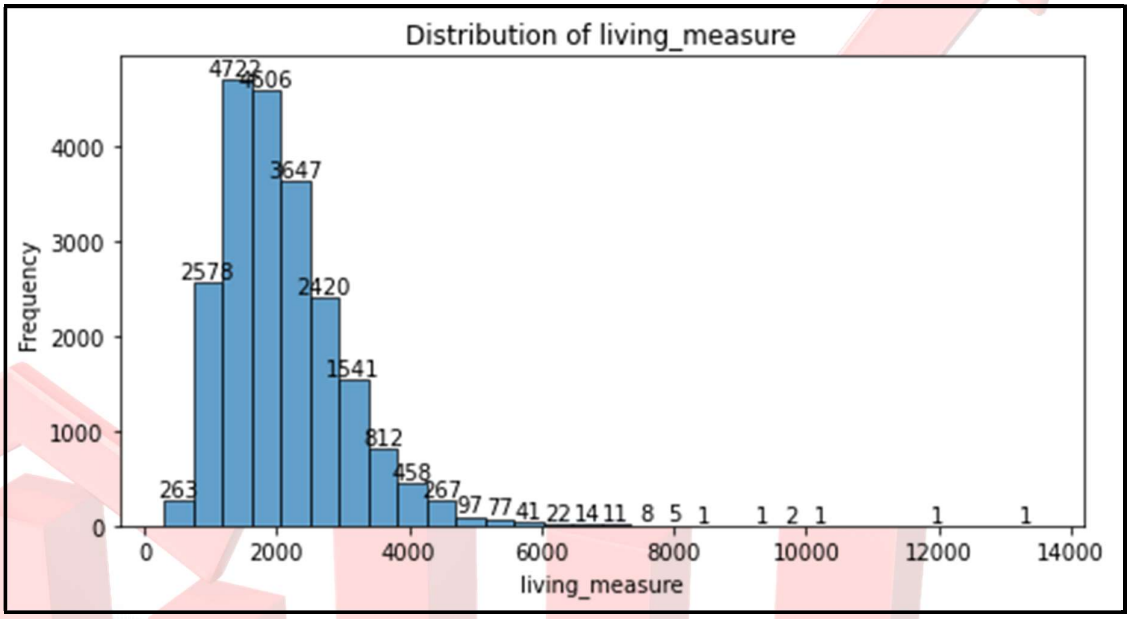

Figure 11 Distribution of Living measure

'living_measure' (Living Area of the House):

- ***Distribution: The distribution of living area sizes is approximately right-skewed, with a longer tail towards larger living areas.***
- ***Spread: Living areas range from a minimum of 290 square feet to a maximum of 13,540 square feet.***

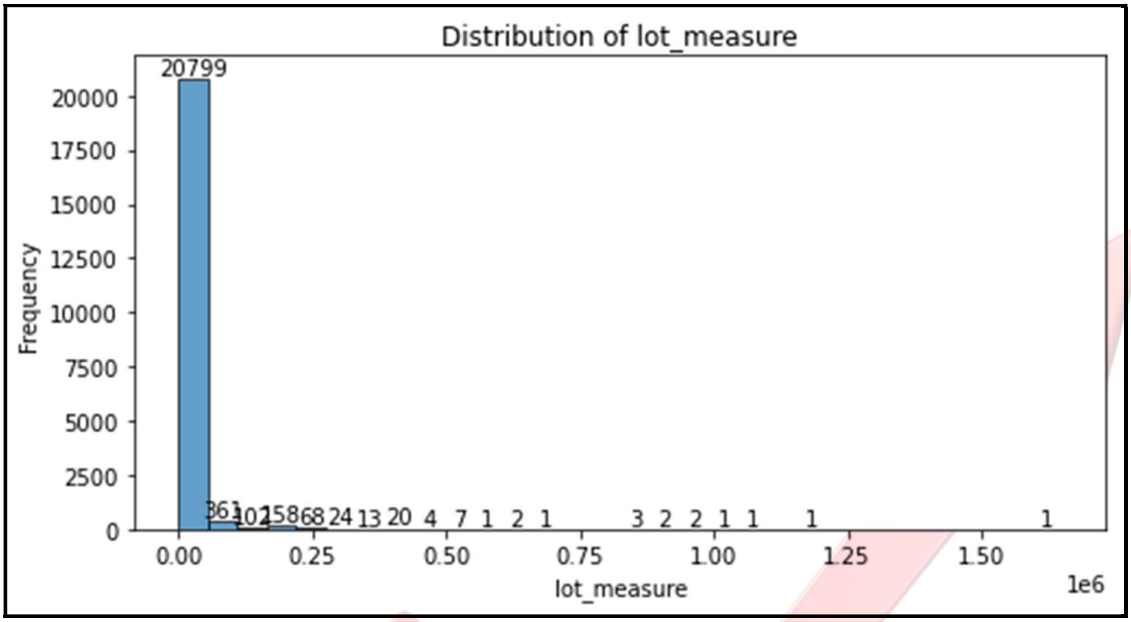

Figure 12 Distribution of lot measure

'lot_measure' (Lot Size):

- ***Distribution: The distribution of lot sizes is right-skewed, with a longer tail towards larger lots.***
- ***Spread: Lot sizes range from a minimum of 520 square feet to a maximum of 1,651,359 square feet.***

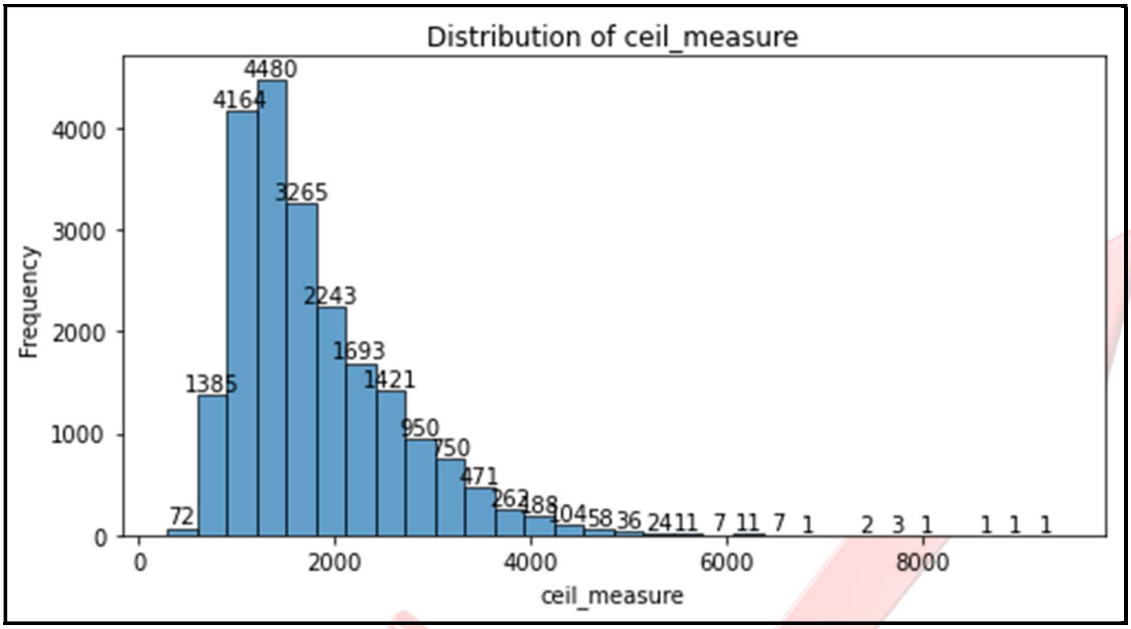

Figure 13 Distribution of ceil measure

'ceil_measure' (Ceiling Area of the House):

- ***Distribution: The distribution of ceiling areas appears to be right-skewed, with a longer tail towards larger ceiling areas.***
- ***Spread: Ceiling areas range from a minimum of 290 square feet to a maximum of 9,410 square feet.***

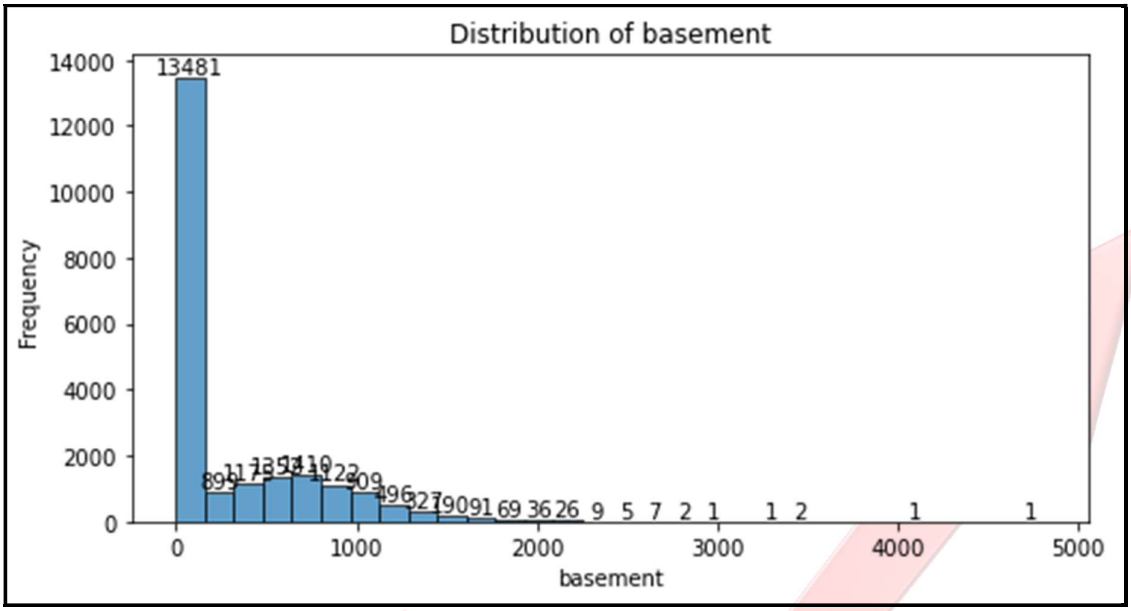

Figure 14 Distribution of basement

'basement' (Basement Area):

- ***Distribution: The distribution of basement areas is right-skewed, with a significant number of houses having no basements (value of 0).***
- ***Spread: Basement areas range from 0 square feet (no basement) to a maximum of 4,820 square feet.***

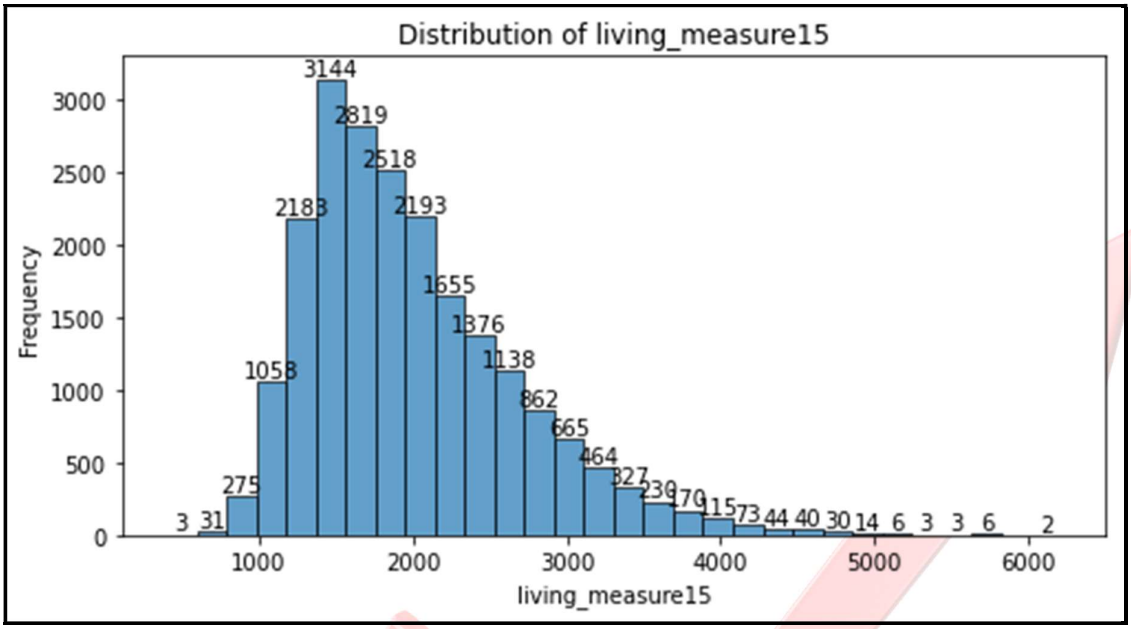

Figure 15 Distribution of Living measure15

'living_measure15' (Living Area in 2015)

- ***Distribution: The distribution of living areas in 2015 appears to be right-skewed, similar to the living area distribution.***
- ***Spread: Living areas in 2015 range from a minimum of 399 square feet to a maximum of 6,210 square feet.***

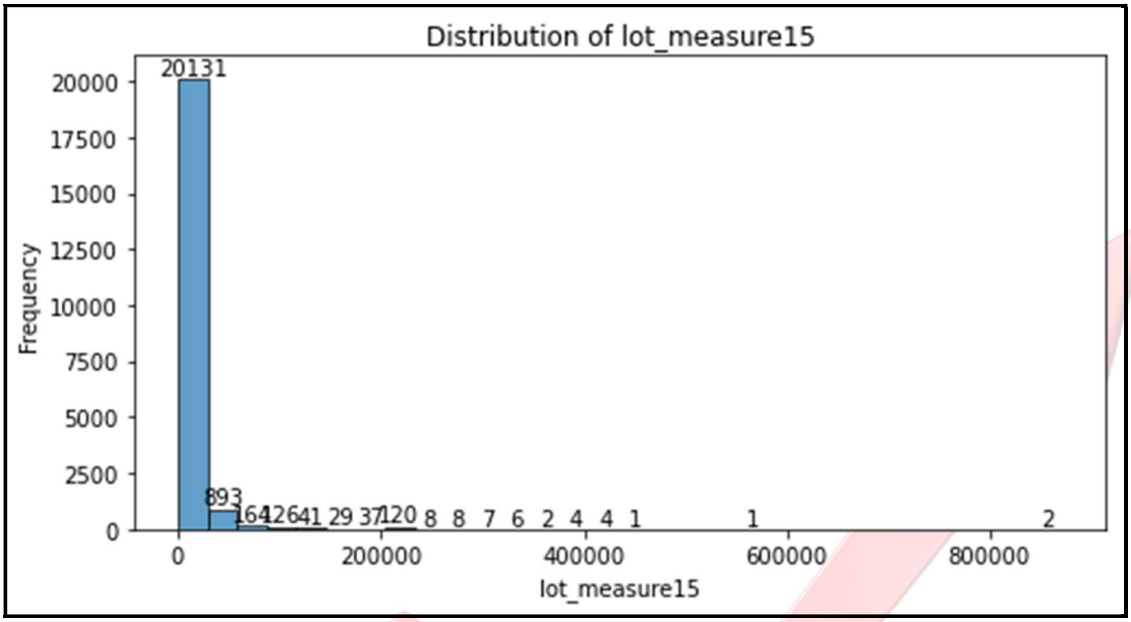

Figure 16 Distribution of lot measure15

'lot_measure15' (Lot Size in 2015):

- ***Distribution: The distribution of lot sizes in 2015 is right-skewed, similar to the lot size distribution.***
- ***Spread: Lot sizes in 2015 range from a minimum of 651 square feet to a maximum of 871,200 square feet.***

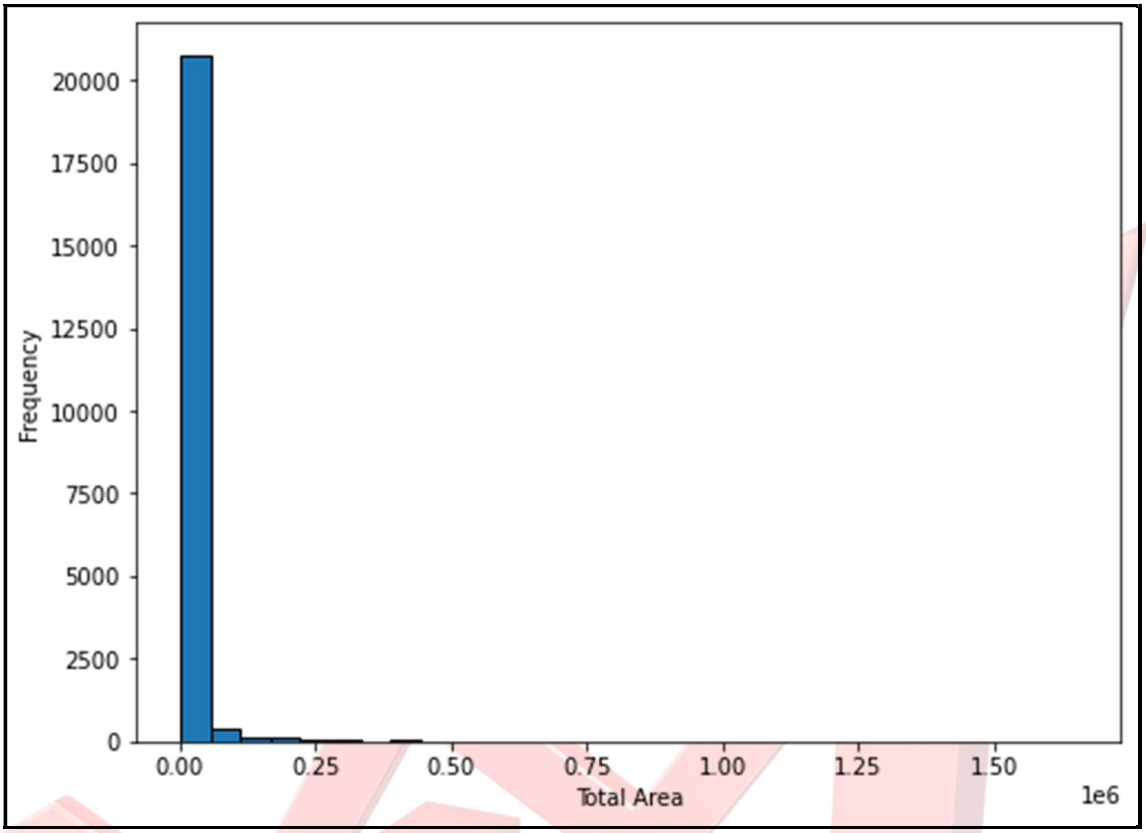

Figure 17 Distribution of Total Area

'total_area' (Total Area of House and Lot):

- ***Distribution: The distribution of total areas combines both living and lot sizes, resulting in a right-skewed distribution.***
- ***Spread: Total areas range from a minimum of 1,423 square feet to a maximum of 1,652,659 square feet.***

b) Bivariate analysis (relationship between different variables, correlations).

Non Visual

Understanding the data and its analysis which could impact on business

The analysis conducted has a significant impact on the real estate business, providing valuable insights and data-driven recommendations that influence various aspects of the industry:

1. Property Valuation and Pricing:

- ***The analysis reveals strong correlations between property attributes (e.g., living area size, ceiling area, and living area in 2015) and house prices.***
- ***Business Impact: This information allows real estate professionals to accurately price properties, considering the specific features that drive value. Pricing strategies can be fine-tuned to maximize profitability and meet market demand.***

2. Marketing and Positioning:

- ***The analysis identifies relationships between attributes like sight rating, quality, and proximity to the coast, shedding light on the desirability of certain property features.***
- ***Business Impact: Real estate agents can use this data to create compelling marketing materials, emphasizing the unique selling points of properties. This approach can attract more potential buyers and lead to quicker sales.***

3. Property Investment and Renovation:

- ***The analysis highlights the impact of property condition, quality, and furnished status on prices.***
- ***Business Impact: Investors and property owners can make informed decisions on renovation or improvement projects based on the expected increase in property value. This data-driven approach minimizes the risk associated with investment decisions.***

4. Segmentation and Targeting:

- ***The analysis provides insights into how various attributes (e.g., bedroom count, bathroom count) are distributed across different property conditions and quality ratings.***
- ***Business Impact: Real estate professionals can segment the market effectively and target specific customer groups based on their preferences and needs. This results in more targeted marketing efforts and better customer satisfaction.***

5. Consistency and Historical Trends:

- ***The analysis indicates the consistency of certain property attributes over time, such as lot size and living area in 2015.***
- ***Business Impact: This information helps in assessing the long-term value of properties and understanding how specific attributes have remained consistent, which can be reassuring to buyers.***

6. Competitive Positioning:

- ***The analysis of correlations between attributes like living area size and price can inform competitive positioning.***
- ***Business Impact: Understanding how certain features impact pricing allows businesses to position properties effectively in the competitive landscape, offering attractive pricing and appealing features.***

7. Informed Decision-Making:

- ***The data-driven insights and recommendations guide business owners and stakeholders in making informed decisions about property valuation, marketing, and investment.***
- ***Business Impact: This analysis ensures that decisions are backed by quantifiable data, reducing risks, improving profitability, and enhancing the overall competitiveness of the business in the real estate market.***

In summary, the analysis has a profound impact on the real estate business by enabling more accurate property valuation, better marketing strategies, and informed decision-

making. It empowers the industry to meet the diverse needs of buyers and sellers, enhance the customer experience, and thrive in a competitive marketplace.

3. Data Cleaning and Pre-Processing

Approach used for identifying and treating missing values and outlier treatment (and why)

The approach used for identifying and treating missing values and outlier treatment is as follows:

1. Identifying Missing Values:

- ***The analysis identifies missing values in the dataset by computing the count and percentage of missing values for each column.***
- ***The approach is to quantify the extent of missing data in each column to make informed decisions about how to handle them.***

2. Removal of Unwanted Variables:

- ***The decision to remove these columns is based on the principle of maintaining data integrity and ensuring that missing values do not significantly affect the analysis and modelling.***
- ***Columns with a very high percentage of missing values, such as "dayhours" and "yr_renovated," are identified and deemed as potentially not providing meaningful information due to the high proportion of missing data.***

3. Outlier Treatment:

- ***The IQR (Interquartile Range) method is chosen to identify and treat outliers in selected numerical columns.***
- ***Outliers are capped to the lower and upper bounds determined by the IQR.***
- ***This method is used to mitigate the impact of outliers on subsequent analyses or modelling tasks.***

Here's an overview of the before and after treatment of outliers

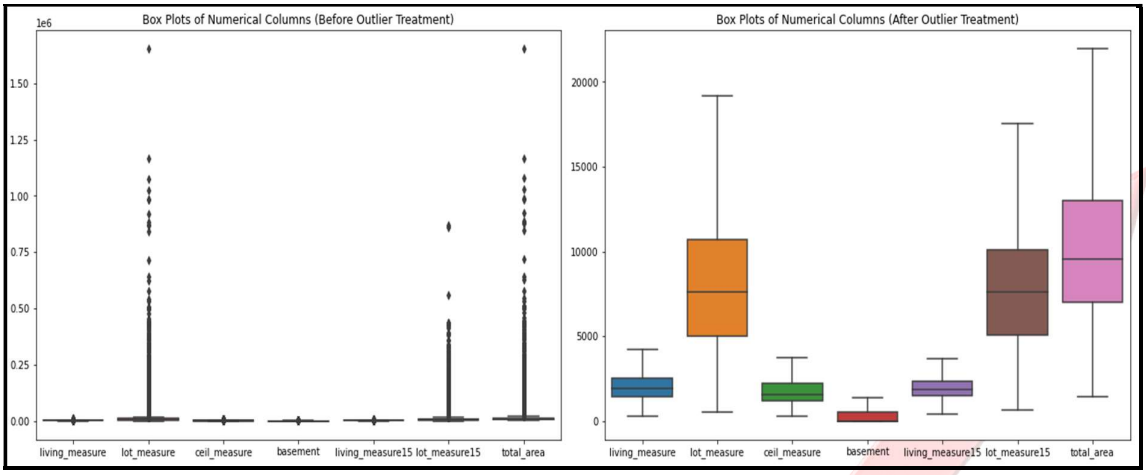

Figure 25 Before and after treatment of outliers

Here are the key findings regarding missing values in our dataset:

- ***Total Missing Values in the Entire Dataset: 43,176***
- ***Percentage of Missing Values in the Entire Dataset: 6.66%***

Missing Values per Column:

- ***dayhours: All 21,613 values (100%) are missing in this column.***
- ***yr_renovated: 20,699 values (95.77%) are missing.***
- ***living_measure15: 166 values (0.77%) are missing.***
- ***room_bed: 108 values (0.50%) are missing.***
- ***room_bath: 108 values (0.50%) are missing.***
- ***condition: 85 values (0.39%) are missing.***
- ***ceil: 72 values (0.33%) are missing.***
- ***total_area: 68 values (0.31%) are missing.***
- ***sight: 57 values (0.26%) are missing.***
- ***lot_measure: 42 values (0.19%) are missing.***
- ***long: 34 values (0.16%) are missing.***
- ***coast: 31 values (0.14%) are missing.***

- ***furnished: 29 values (0.13%) are missing.***
- ***lot_measure15: 29 values (0.13%) are missing.***
- ***living_measure: 17 values (0.08%) are missing.***
- ***yr_built: 15 values (0.07%) are missing.***
- ***ceil_measure: 1 value (0.0046%) is missing.***
- ***basement: 1 value (0.0046%) is missing.***
- ***quality: 1 value (0.0046%) is missing.***

4. Missing Value Treatment:

- ***For columns with missing values, numerical values are imputed with median values, while categorical columns are imputed with mode values.***
- ***Imputation is carried out after outlier treatment to ensure that the imputed columns are not influenced by extreme values.***

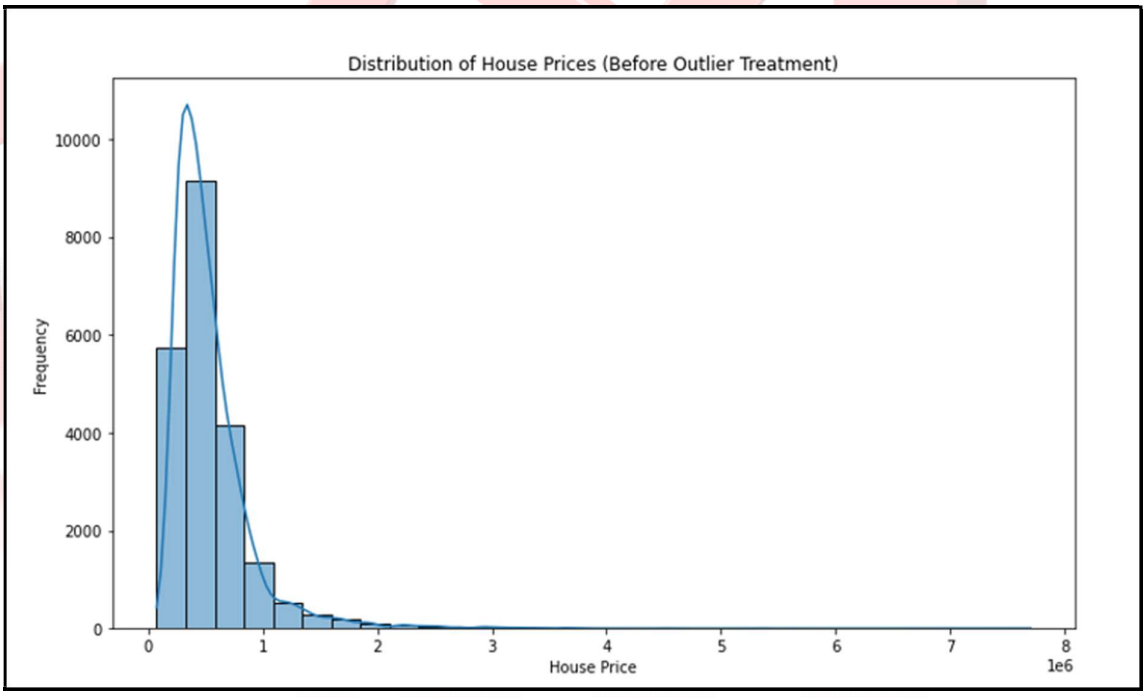

Figure 26 Distribution of Price Before Outlier Treatment

Outlier treatment, if applied, can help mitigate the impact of extreme values on our analysis and modelling.

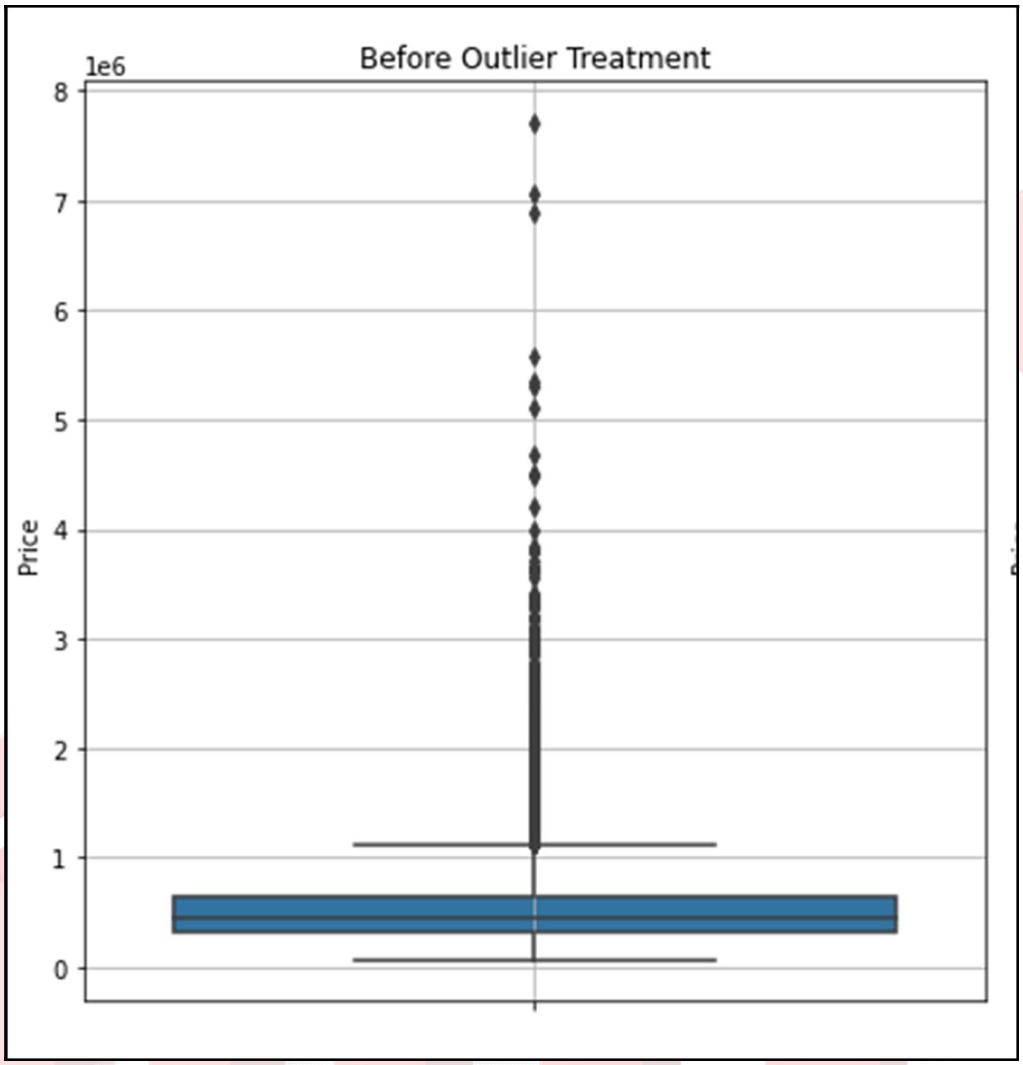

Figure 27 Box plot for Price Attribute

Here are the summary statistics for house prices before outlier treatment:

- ***Count: 21,613***
- ***Mean: $540,182.20***
- ***Standard Deviation: $367,362.20***
- ***Minimum Price: $75,000.00***
- ***25th Percentile (Q1): $321,950.00***
- ***Median (Q2 or 50th Percentile): $450,000.00***

- *75th Percentile (Q3): $645,000.00*
- *Maximum Price: $7,700,000.00*

Additionally, the statistical measures of skewness and kurtosis for house prices are as follows:

- *Skewness: 4.0217*
- *Kurtosis: 34.5224*

The positive skewness value (greater than 0) indicates that the distribution of house prices is right-skewed, meaning that it has a longer right tail and a concentration of values on the left side of the distribution. The high kurtosis value indicates that the distribution has heavy tails and may have outliers or extreme values.

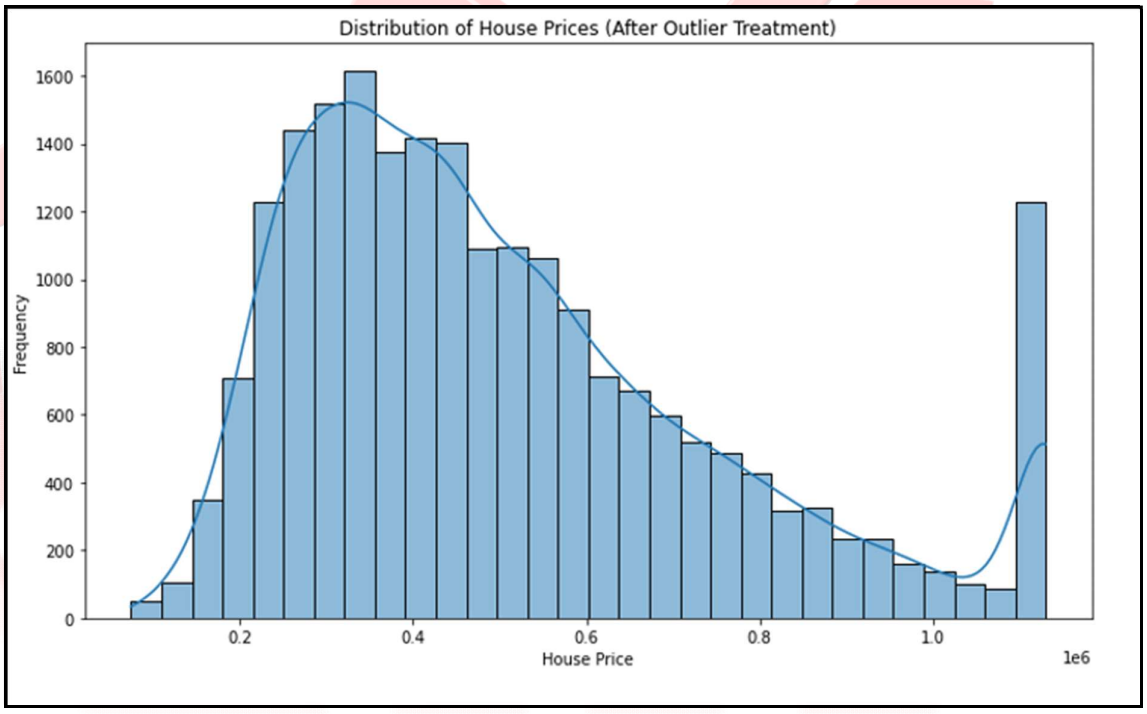

Figure 28 Distribution of Price After Outlier Treatment

Outlier treatment has helped in reducing the impact of extreme values and in making the distribution of house prices more suitable for analysis and modelling.

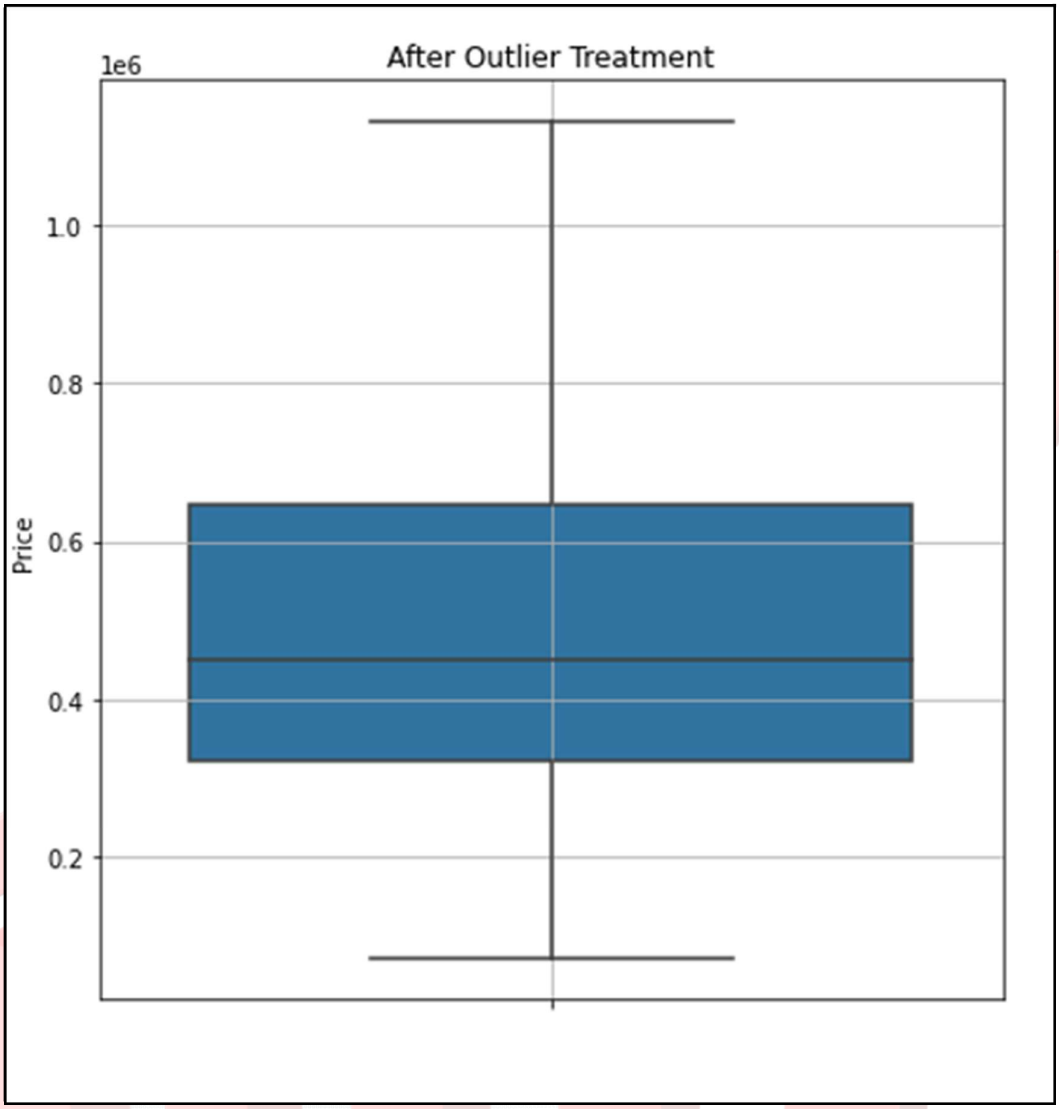

Figure 29 Box plot after outlier treatment

Here are the summary statistics for house prices after outlier treatment:

- ***Count: 21,613***
- ***Mean: $511,607.60***
- ***Standard Deviation: $250,047.90***
- ***Minimum Price: $75,000.00***
- ***25th Percentile (Q1): $321,950.00***
- ***Median (Q2 or 50th Percentile): $450,000.00***
- ***75th Percentile (Q3): $645,000.00***

- ***Maximum Price (After Outlier Treatment): $1,129,575.00***

Additionally, the statistical measures of skewness and kurtosis for house prices after outlier treatment are as follows:

- ***Skewness: 0.9371***
- ***Kurtosis: 0.2001***

Compared to the skewness and kurtosis values before outlier treatment, the skewness is significantly reduced after treatment, indicating that the distribution of house prices is closer to a normal distribution. The lower kurtosis value also suggests that the distribution has lighter tails and is less prone to extreme values.

Need for variable transformation (if any)

- ***The "dayhours" column, which contains date and time information, is planned for transformation into a datetime data type.***
- ***Additionally, there is an acknowledgment of the existence of derived columns related to date, year, month, day, hour, minute, and second.***
- ***These transformations are made to prepare the data for time-series analysis and enhance the usability of time-related features in the dataset.***
- ***Categorical variables are encoded to convert them into numerical format, which is necessary for many machine learning algorithms.***

Variables removed or added and why (if any)

- ***Columns with a very high percentage of missing values, such as "dayhours" and "yr_renovated," are identified and deemed as potentially not providing meaningful information due to the high proportion of missing data.***
- ***The decision to remove these columns is based on the principle of maintaining data integrity and ensuring that missing values do not significantly affect the analysis and modeling.***

4.Model building

a) Clear on why was a particular model(s) chosen.

Approach for model building

1.Selecting Target and Features:

The first step in model building involves identifying the target variable, which, in this context, is likely 'house prices.' Additionally, features that can significantly influence the target variable need to be selected. These features might include attributes like living measure, lot measure, ceil measure, and others that were explored and analysed during the data pre-processing and EDA phases.

2.Train-Test Split:

To evaluate the performance of our models, the dataset is divided into training and testing sets.

```
(21613, 27)
(21613,)
```

Snippet 1. 4 Shape of X & y split

Initial Dataset:

- ***(21613, 27) indicates that the original dataset has 21,613 samples and 27 features in training data***
- ***(21613,) signifies a target vector with 21,613 labels corresponding to the samples.***

```
(17290, 27)
(4323, 27)
(17290,)
(4323,)
```

Snippet 1. 5 Train-Test Split

Train-Test Split:

- *(17290, 27) reveals that the training set consists of 17,290 samples and retains the original 27 features.*
- *(4323, 27) indicates that the testing set contains 4,323 samples, maintaining the 27 original features.*
- *(17290,) denotes the target vector for the training set with 17,290 labels.*
- *(4323,) signifies the target vector for the testing set, comprising 4,323 labels.*

This approach helps assess how well the models generalize to new, unseen data. The training set is used to train the models, while the testing set provides an unbiased evaluation of their predictive capabilities.

3.Clear on Choosing the regression models

The choice of regression models is a critical step in predictive modeling. Each model has its strengths and weaknesses, and the selection should be based on the specific characteristics and goals of the dataset. Here's a clear explanation of why particular regression models were chosen:

1.Linear Regression:

- *Why Chosen: Linear regression is a fundamental and interpretable model that serves as a good starting point. It is chosen when the relationship between the target variable and predictor variables can be approximated linearly. It provides insights into the importance and direction of each predictor's impact on the target variable. It's a simple, transparent model that can serve as a baseline for regression tasks.*

2.Random Forest:

- *Why Chosen: Random Forest is an ensemble model that is chosen for its ability to handle complex relationships between variables. It can capture non-linearities and interactions between features. Random Forest is robust to outliers and can handle a*

mix of numerical and categorical data. It is often selected when the dataset is large, and there's a need for high predictive accuracy.

3.Decision Tree:

- ***Why Chosen: Decision trees are chosen for their simplicity and interpretability. They are particularly useful when there are non-linear relationships in the data and can handle both numerical and categorical data. Decision trees are often used as base models in ensemble methods like Random Forest, and they can provide insights into feature importance and decision-making.***

4.Lasso Regression:

- ***Why Chosen: Lasso regression is a type of linear regression that is chosen when there are many features, and feature selection is crucial. It adds L1 regularization to the linear regression, which can shrink coefficients to zero, effectively performing feature selection. Lasso helps in identifying the most important predictors and reducing overfitting.***

5.Ridge Regression:

- ***Why Chosen: Ridge regression is another type of linear regression that is chosen when multicollinearity (high correlation between predictors) is a concern. It adds L2 regularization to linear regression, which helps in reducing the impact of multicollinearity and stabilizing coefficient estimates. It can improve model robustness.***

6.ElasticNet:

- ***Why Chosen: ElasticNet is a combination of Lasso and Ridge regression. It is chosen when there is a need for both feature selection (like Lasso) and handling multicollinearity (like Ridge). ElasticNet balances the effects of L1 and L2 regularization, offering a flexible approach to regression modeling.***

The choice of these models represents a diverse set of techniques that can address different aspects of the dataset, including linearity, non-linearity, feature importance, and regularization. The decision on which model to use or whether to combine multiple models should be driven by the data's characteristics and the specific objectives of the regression

task. It's common to try multiple models and evaluate their performance to select the one that best suits the problem at hand.

b) Effort to improve model performance.

The efforts made to improve model performance involve several techniques, including ensemble modeling and model tuning measures. Here's how these efforts were implemented based on the provided context:

1.Ensemble Modelling:

- ***Ensemble modeling is employed to combine multiple machine learning models to enhance overall performance and make more accurate predictions. Two common ensemble methods, bagging and boosting, are mentioned in the context.***

2.Random Forest:

- ***Random Forest is used as an ensemble of Decision Trees. It reduces overfitting by averaging the results of multiple decision trees. This ensemble method is applied***

for regression tasks, and the specific implementation used is RandomForestRegressor.

3.Gradient Boosting:

- *Gradient Boosting is another ensemble method that builds trees sequentially, with each tree correcting the errors of the previous one. This technique is used for regression tasks, and the specific implementation is GradientBoostingRegressor.*

4.XGBoost:

- *XGBoost is introduced as an efficient and scalable implementation of gradient boosting. It is a powerful ensemble technique that is known for its high performance.*

After training the models, hyperparameter tuning is mentioned as a model tuning measure. Techniques like grid search and random search can be used to find the best set of hyperparameters for the XGBoost model.

These efforts to improve model performance focus on combining multiple models through ensemble methods like Random Forest and Gradient Boosting. These ensemble methods reduce overfitting and enhance predictive accuracy. Additionally, the utilization of XGBoost, a highly efficient gradient boosting implementation, further enhances model performance by allowing for hyperparameter tuning.

The key idea behind these efforts is to leverage the strengths of various models and fine-tune their parameters to achieve the best possible predictive performance, which is essential for the success of regression tasks.

5.Model Evaluation

a) How was the model validated?

Model Validation Metrics Comparison:

Here's a detailed comparison of the training and test results for each evaluation metric:

1.Mean Absolute Error (MAE):

- ***Training MAE: Linear model achieved a MAE of approximately 129,602, which indicates the average absolute difference between predicted and actual values in the training data.***
- ***Test MAE: The test MAE for the linear model is about 126,249, suggesting that it generalizes well to unseen data.***

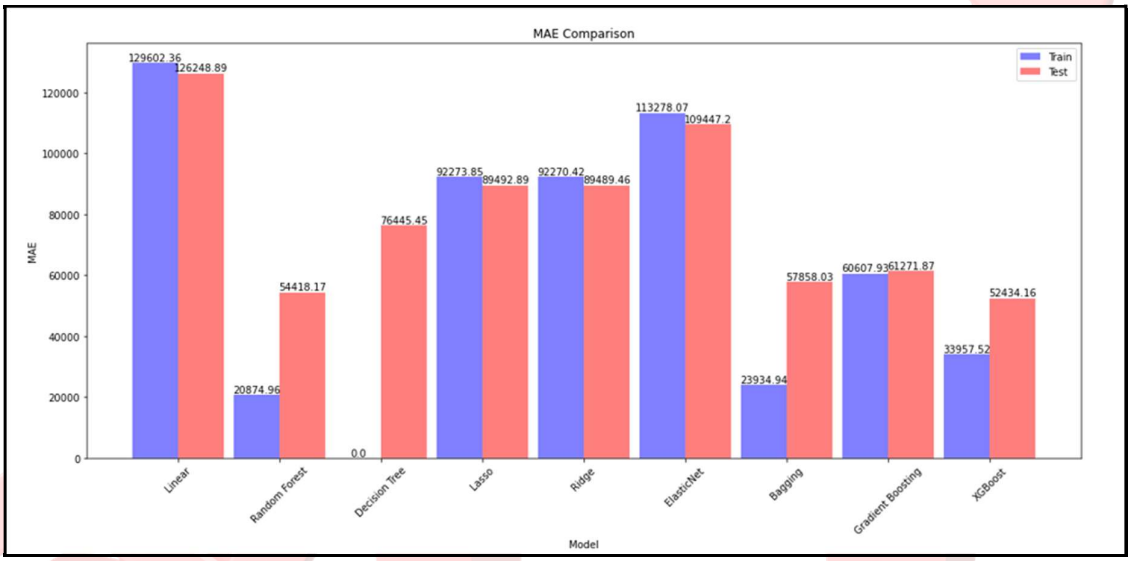

Figure 30 MAE Comparison

2.Mean Squared Error (MSE):

- ***Training MSE: The linear model has a training MSE of around 2.72e+10, reflecting the average squared difference between predicted and actual values in the training data.***
- ***Test MSE: In the test dataset, the linear model's MSE is approximately 2.60e+10.***

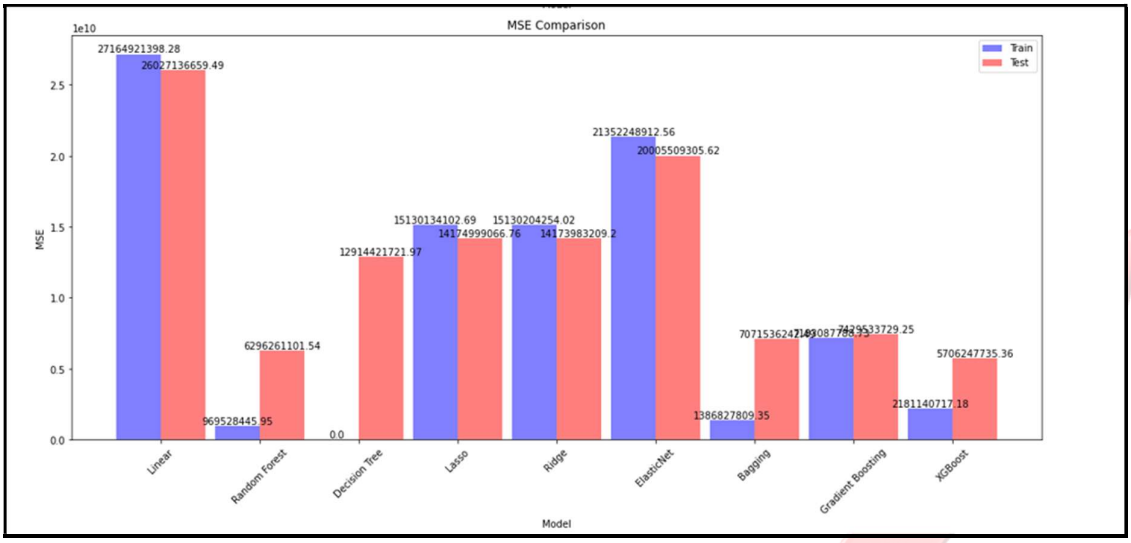

Figure 31 MSE Comparison

3.Root Mean Squared Error (RMSE):

- ***Training RMSE: The linear model's training RMSE is approximately 164,818, providing a measure of the standard deviation of prediction errors in the training dataset.***
- ***Test RMSE: The test RMSE for the linear model is around 161,329.***

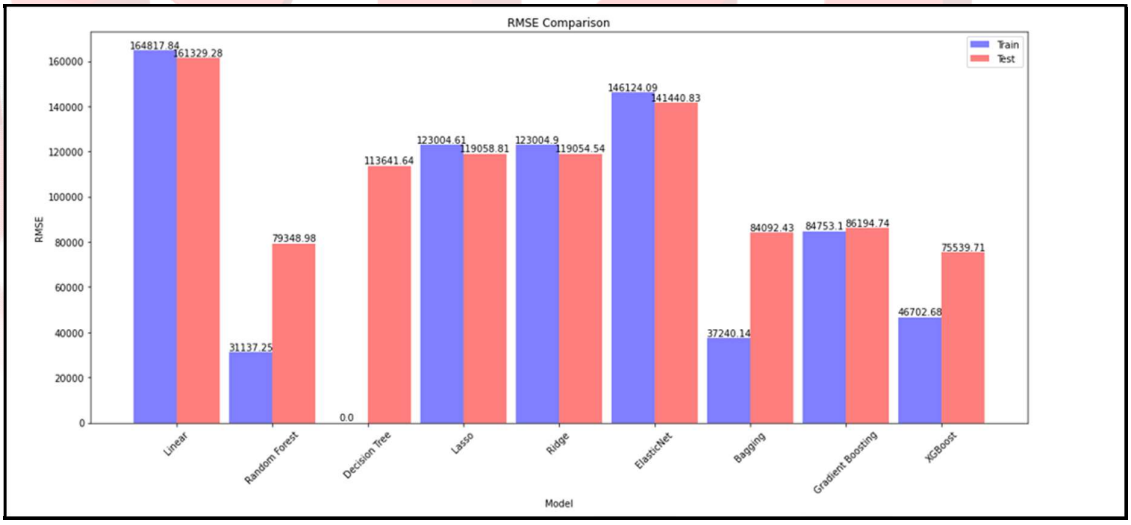

Figure 32 RMSE Comparison

4. R-squared (R2) Score:

- ***Training R2 Score: The linear model achieved an R2 score of approximately 0.568, indicating that it explains 56.8% of the variance in the training data.***
- ***Test R2 Score: In the test data, the linear model's R2 score is approximately 0.573.***

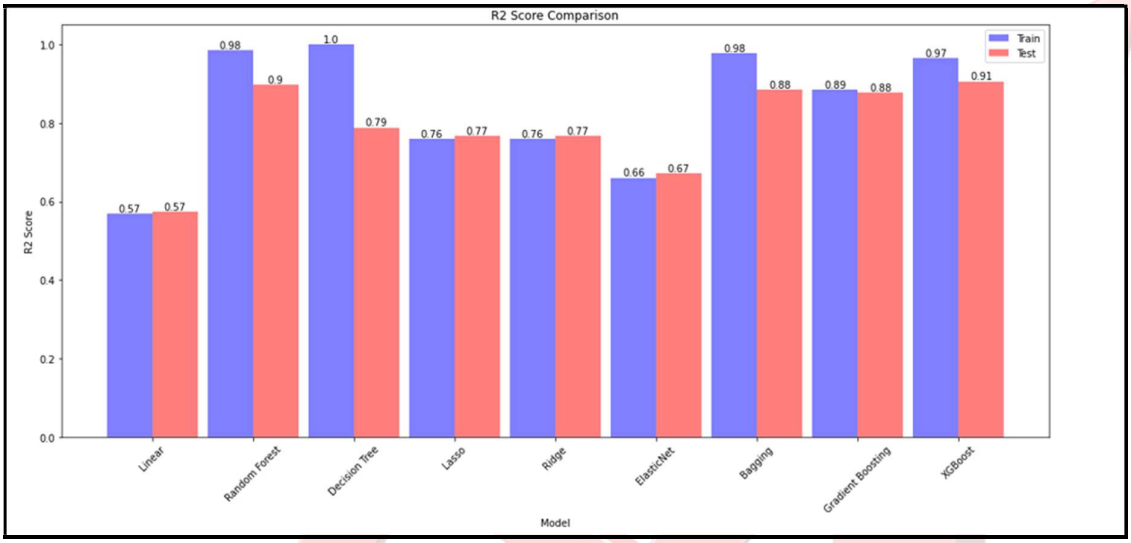

Figure 33 R2 Score Comparison

b) Just accuracy, or anything else too?

The model was not validated based on accuracy alone; a range of additional metrics were used to assess its performance. These metrics include:

Model	Train MAE	Test MAE	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2 Score	Test R2 Score
XGBoost	33958	52434	2181141000	5706248000	46703	75540	97%	91%
Random Forest	20875	54418	969528400	6296261000	31137	79349	98%	90%
Bagging	23935	57858	1386828000	7071536000	37240	84092	98%	88%
Gradient Boosting	60608	61272	7183088000	7429534000	84753	86195	89%	88%
Decision Tree	0	76445	0	12914420000	0	113642	100%	79%
Ridge	92270	89489	15130200000	14173980000	123005	119055	76%	77%
Lasso	92274	89493	15130130000	14175000000	123005	119059	76%	77%
ElasticNet	113278	109447	21352250000	20005510000	146124	141441	66%	67%
Linear Regression	129602	126249	27164920000	26027140000	164818	161329	57%	57%

Snippet 1. 6 Train Test Score Table

1.Mean Absolute Error (MAE)

- ***MAE measures the absolute differences between predicted and actual values. A lower MAE indicates better performance in terms of prediction accuracy.***

2.Mean Squared Error (MSE)

- ***MSE calculates the squared differences between predicted and actual values. Lower MSE values represent a better fit of the model to the data.***

3.Root Mean Squared Error (RMSE)

- ***RMSE is the square root of MSE and provides a more interpretable measure of prediction error. It quantifies the error in the same units as the target variable.***

4.R-squared (R2) Score

- ***R2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R2 score signifies a better fit of the model to the data.***

These metrics offer a comprehensive evaluation of the model's performance, taking into account various aspects of accuracy, precision, and the model's ability to capture variance in the data. While accuracy is an essential metric, using multiple evaluation criteria provides a more complete understanding of the model's effectiveness and limitations.

6.Final interpretation / recommendation

a) Detailed recommendations for the management/client based on the analysis done.

1. Exploratory Data Analysis (EDA)

- ***Feature Importance: Identify the most important features influencing house prices. Based on the modeling results, it's evident that living area size, ceiling area size, and living area size in 2015 are crucial predictors.***
- ***Market Segmentation: Understand market segments by analyzing the distribution of house conditions, qualities, and other categorical variables.***

Recommendation:

- ***Focus on these key features when assessing property values and determining listing prices.***
- ***Tailor marketing and pricing strategies to specific market segments. For example, consider different approaches for houses in excellent condition versus those in fair condition.***

2.Univariate Analysis

- ***Coast Proximity and Quality: The analysis reveals that higher-quality houses are not necessarily clustered near the coast.***
- ***Bedroom and Bathroom Counts: Most houses have 3 bedrooms and 2.5 bathrooms. However, there are outliers with unusual values.***

Recommendation:

- ***Consider diversifying real estate investments by looking for high-quality properties in both coastal and inland areas.***
- ***Understand customer preferences by analysing which bedroom and bathroom configurations are in high demand. Adjust property listings accordingly.***

3. Bivariate Analysis

- ***Living Area vs. Ceiling Area: There is a very strong positive correlation between living area size and ceiling area size.***
- ***Lot Size vs. Lot Size in 2015: The strong correlation between lot size and lot size in 2015 suggests that lot sizes tend to be consistent over time.***

Recommendation:

- ***Consider using this information to offer properties with matching living and ceiling areas, which could be an attractive feature for potential buyers.***
- ***Highlight the stability of lot sizes in your marketing, emphasizing that what buyers see is what they get in terms of outdoor space.***

4. Modelling:

- ***Model Selection: Based on the modelling results, Random Forest and XGBoost demonstrate the highest predictive accuracy.***
- ***Ensemble Modelling: Ensemble methods like Random Forest and XGBoost can further boost predictive accuracy.***
- ***Model Tuning: Fine-tuning the selected model using techniques like grid search or random search.***
- ***Monitoring and Maintenance: Implement a monitoring system to regularly assess model performance and retrain as needed.***

Recommendation:

- ***Utilize Random Forest or XGBoost for price predictions. Ensure that the selected model is incorporated into your online platform for property valuation.***
- ***Explore ensembling techniques by combining multiple models to enhance prediction performance.***
- ***Continuously optimize the model's hyperparameters to improve prediction accuracy.***
- ***Monitor model performance over time and retrain as market conditions change to maintain accurate predictions.***

5. Generic Recommendation:

- ***Incorporate Point of Interests (POIs) Data: Consider integrating data on the proximity of essential amenities such as malls, hospitals, transportation hubs, and the overall neighbourhood quality. These factors can significantly impact house prices, and their inclusion can enhance prediction accuracy.***
- ***Amenities and Neighbourhood Insights: Gather more detailed information on the specific amenities available in the neighbourhood and how they influence pricing. Analyze the relationship between amenities, neighbourhood characteristics, and house values to enrich your predictive model.***
- ***Establish a feedback mechanism to gather user input for ongoing model improvement.***

- ***Monitor the real estate market closely to adapt to changing trends.***
- ***Futuristic Considerations: Include data on ongoing or planned infrastructure developments in the area. This forward-looking approach can help anticipate potential changes in property values due to upcoming projects or improvements.***
- ***Legal and Historical Data: Investigate the legal aspects of the properties, such as land titles and survey numbers. Additionally, historical price data for the buildings can offer valuable insights into price trends and market dynamics.***
- ***Demographic and Economic Trends: Consider incorporating demographic data and economic indicators that may affect property values, such as population growth, employment rates, pollution metrics, and income levels in the area.***
