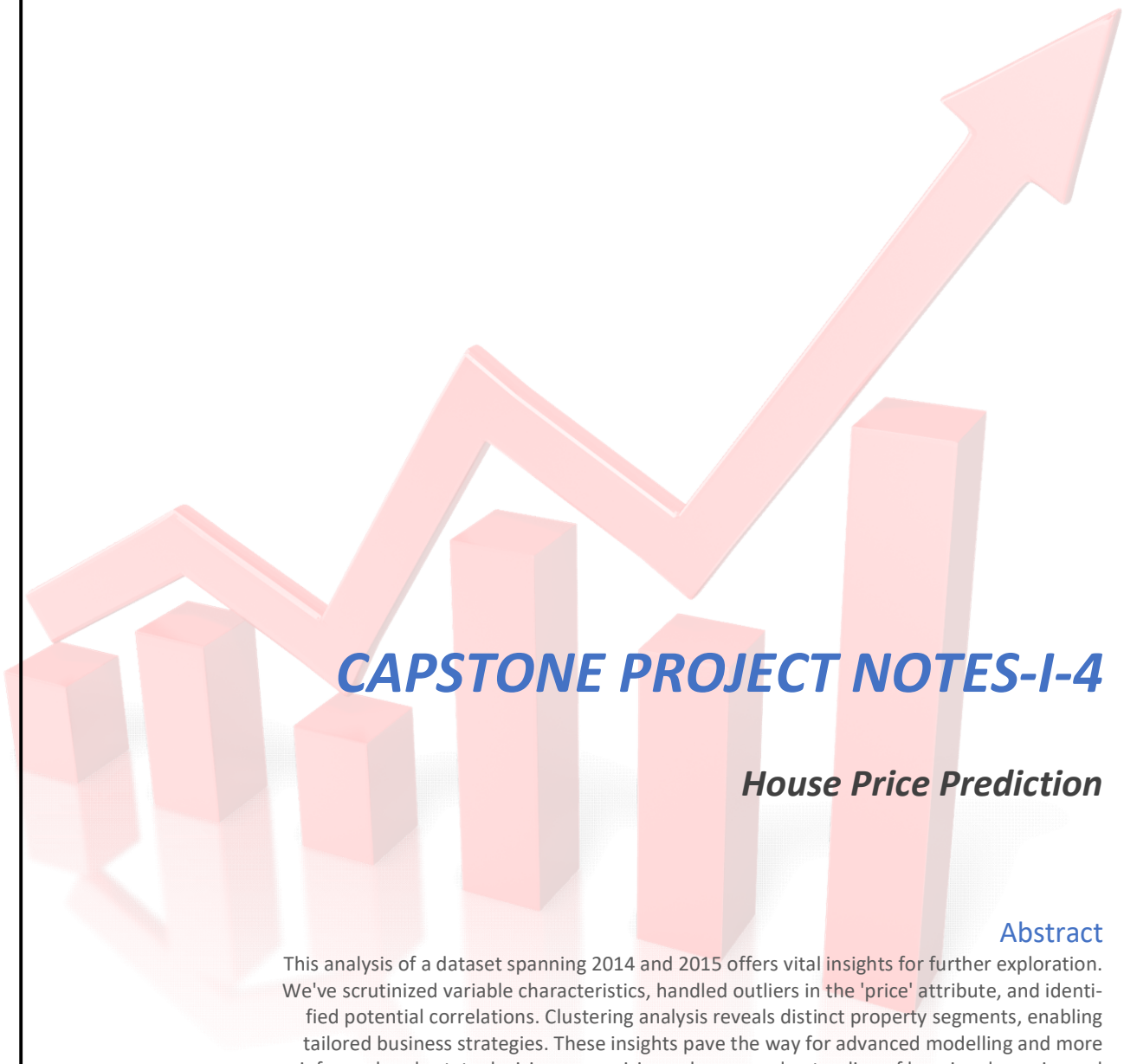


CAPSTONE PROJECT NOTES-I-4

House Price Prediction

Abstract

This analysis of a dataset spanning 2014 and 2015 offers vital insights for further exploration. We've scrutinized variable characteristics, handled outliers in the 'price' attribute, and identified potential correlations. Clustering analysis reveals distinct property segments, enabling tailored business strategies. These insights pave the way for advanced modelling and more informed real estate decisions, promising a deeper understanding of housing dynamics and enhancing decision-making in the real estate sector.

Sudheendra K

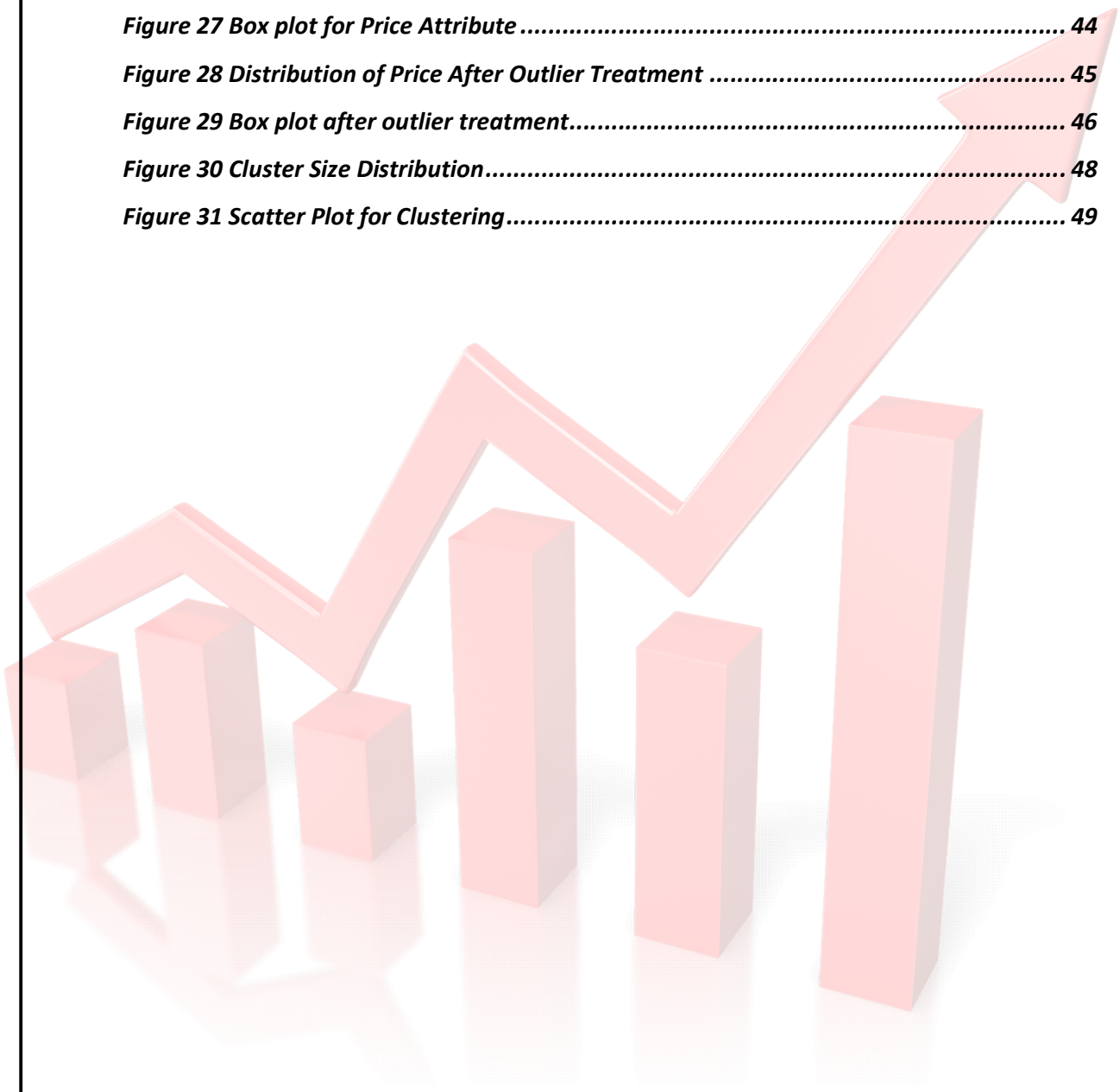
Sudhi0404@gmail.com

PGPDSBA-OCT-B-22

Capstone Project notes-I-4

Contents

Capstone Project notes-I-4



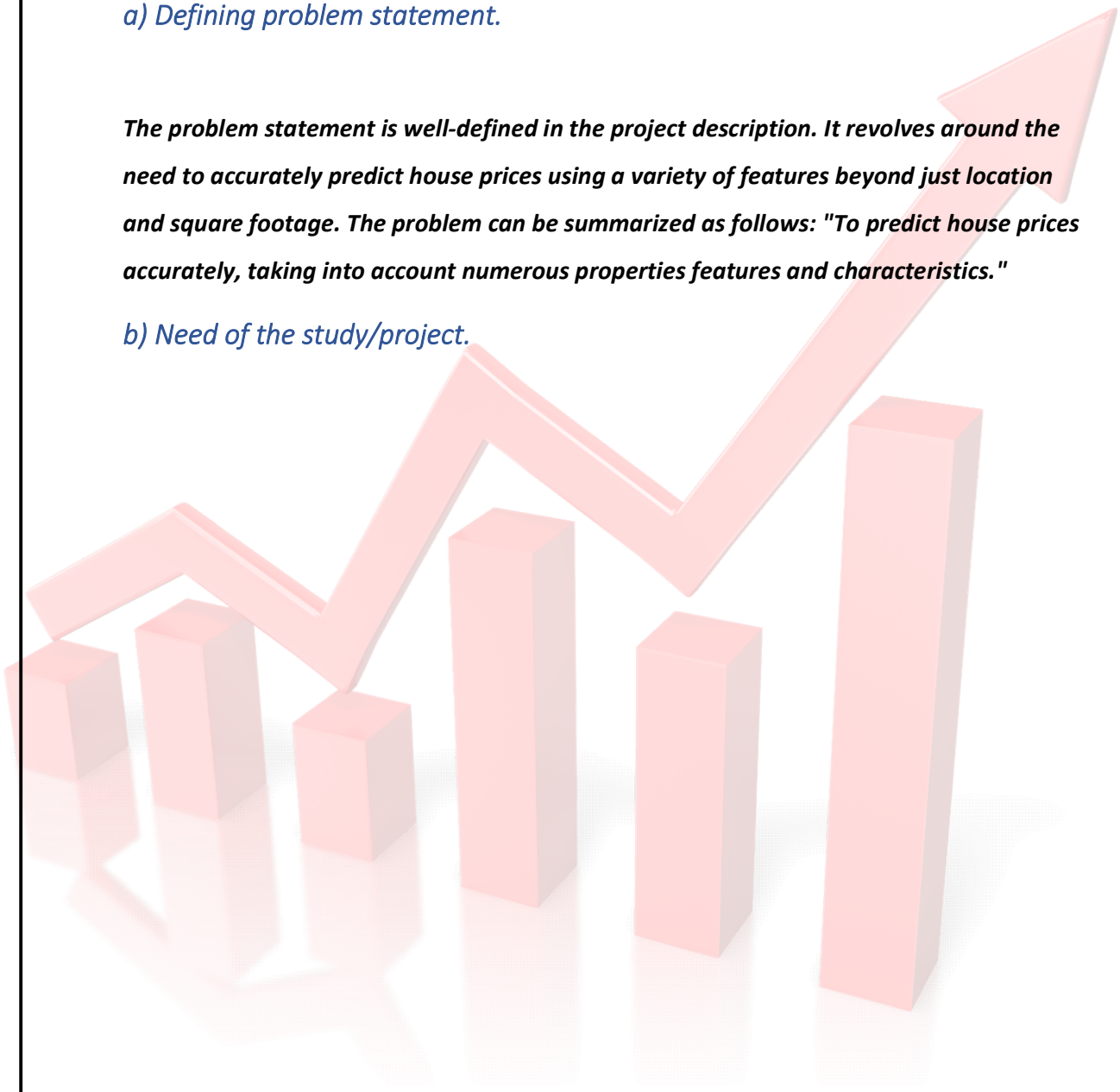
Capstone Project notes-I-4

1) Introduction of the business problem

a) Defining 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) Need of the study/project.



Capstone Project notes-I-4

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.

2)Data Report

a) 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.***

Capstone Project notes-I-4

- ***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

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

Capstone Project notes-I-4

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).***

Capstone Project notes-I-4

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).***

Capstone Project notes-I-4

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.***

Additional Columns:

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

c) 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.***

Capstone Project notes-I-4

3) Exploratory data analysis

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

Categorical Attributes:

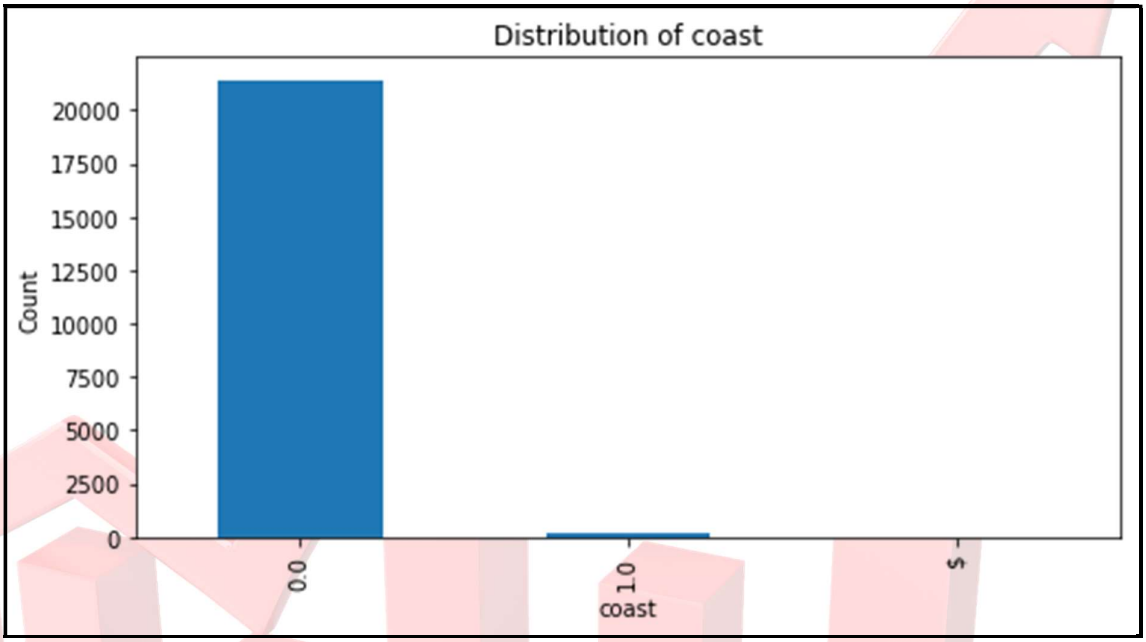


Figure 1 Distribution of Coast

'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).***

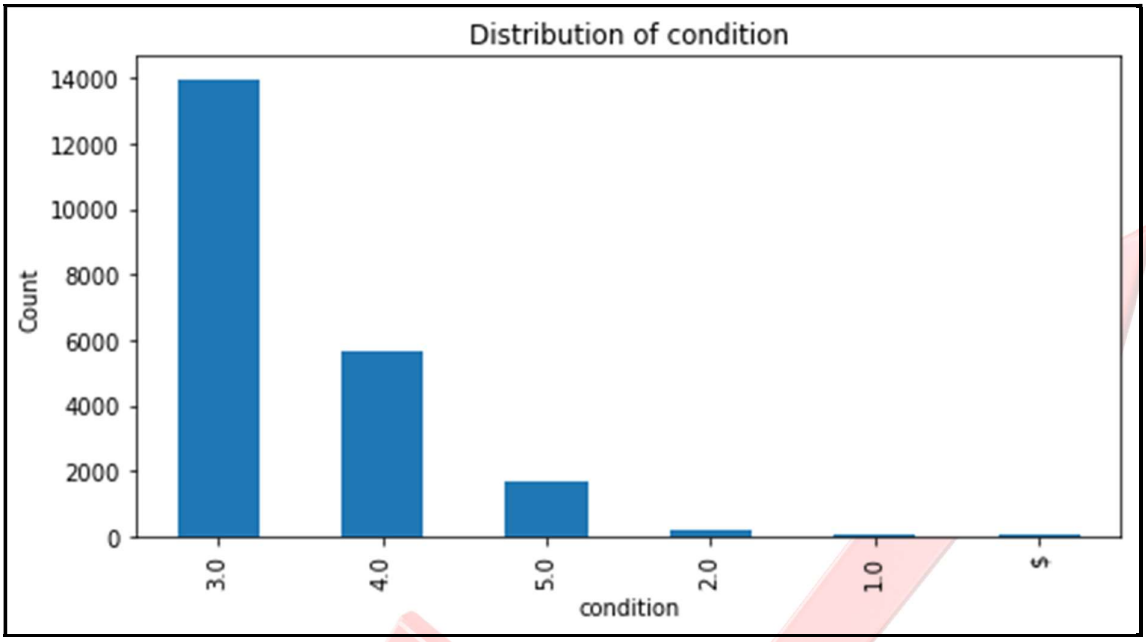


Figure 2 Distribution of condition

'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).***

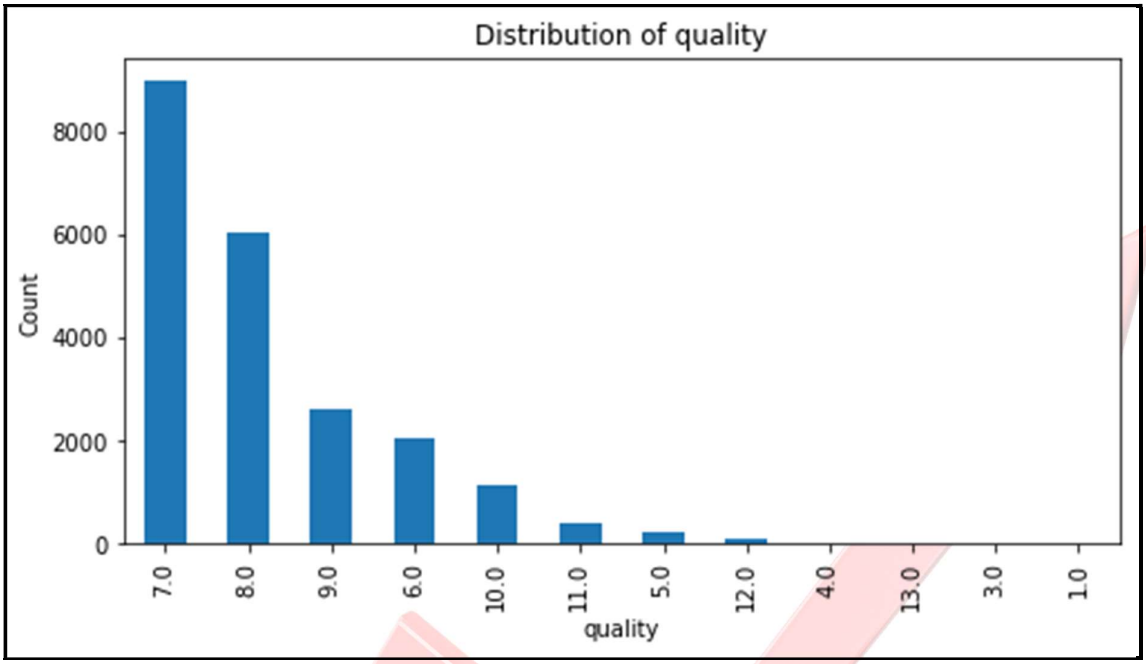


Figure 3 Distribution of quality

'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***

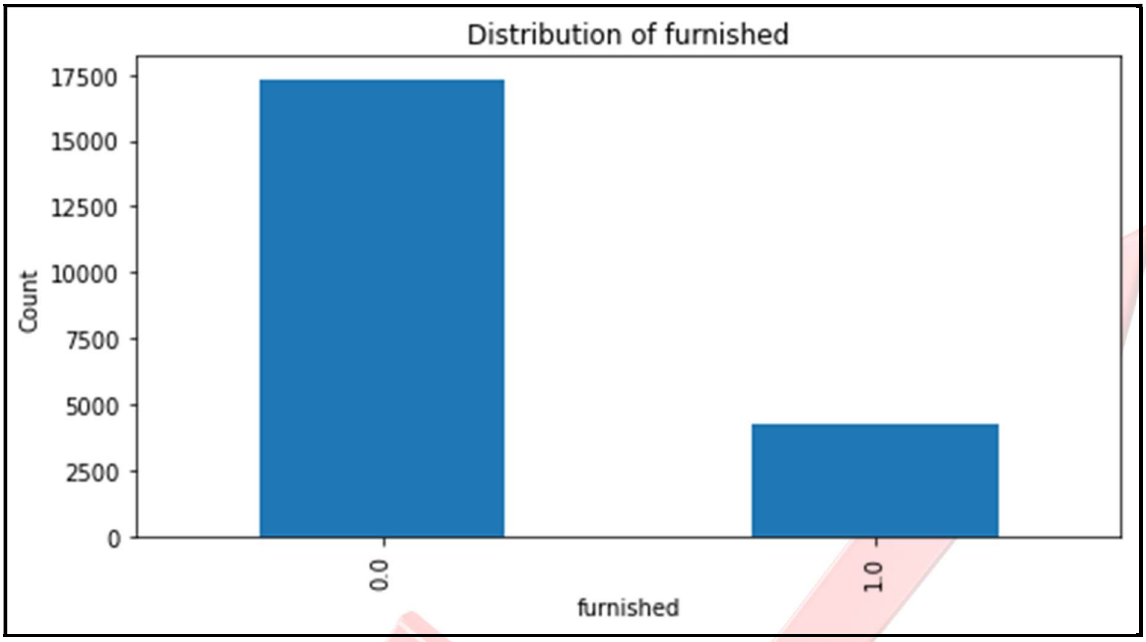


Figure 4 Distribution of Furnished

'furnished' (Furnished):

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

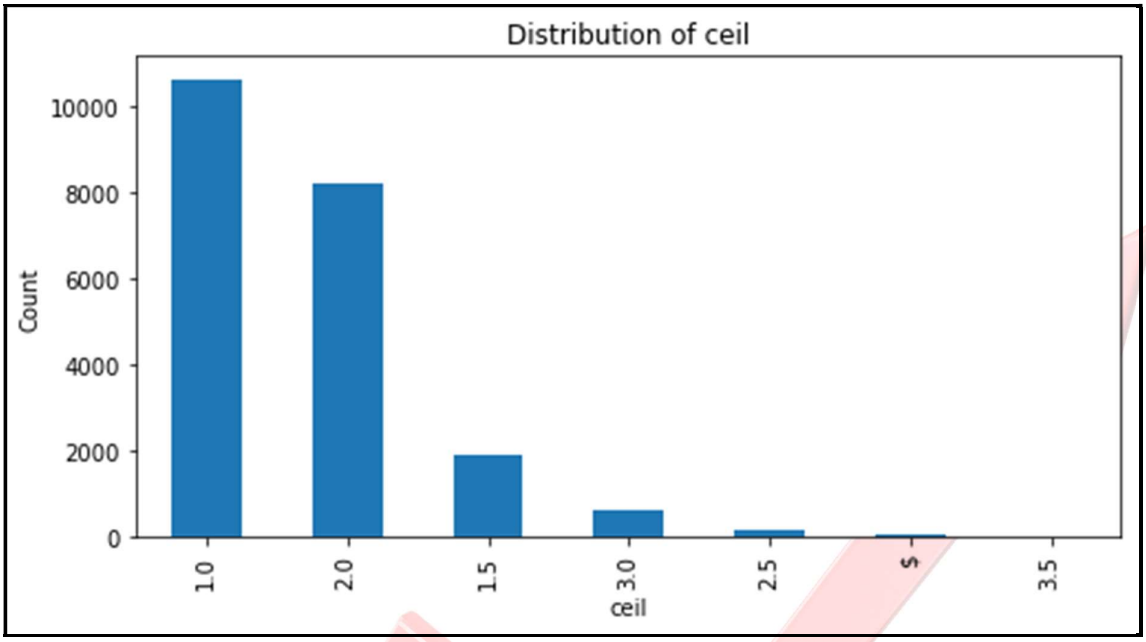


Figure 5 Distribution of ceil

'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.***

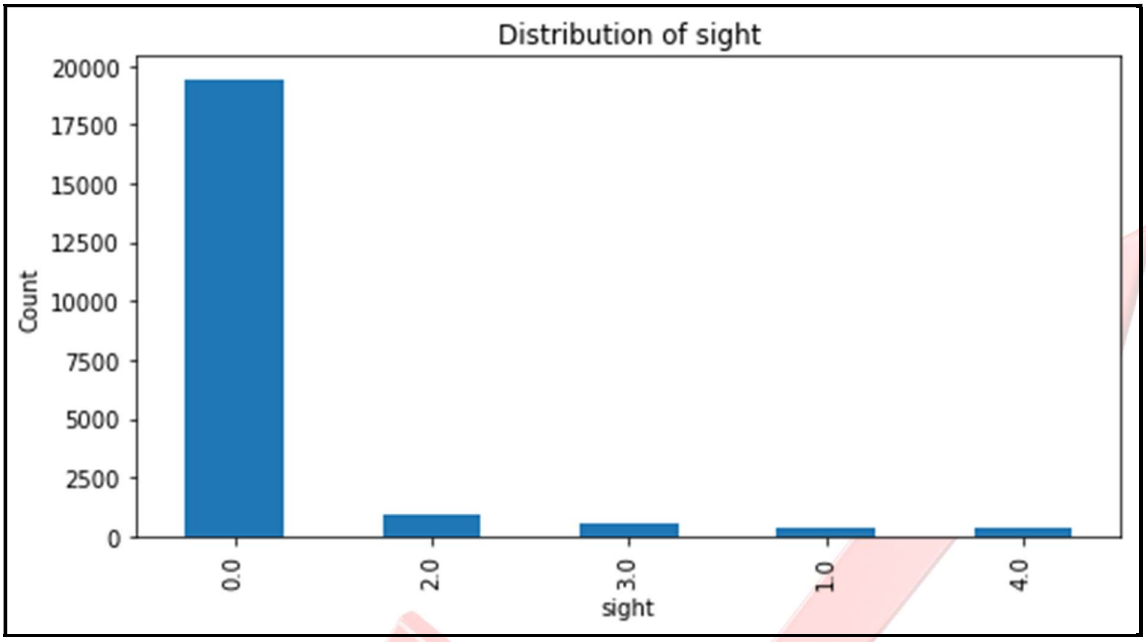


Figure 6 Distribution of sight

'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.***

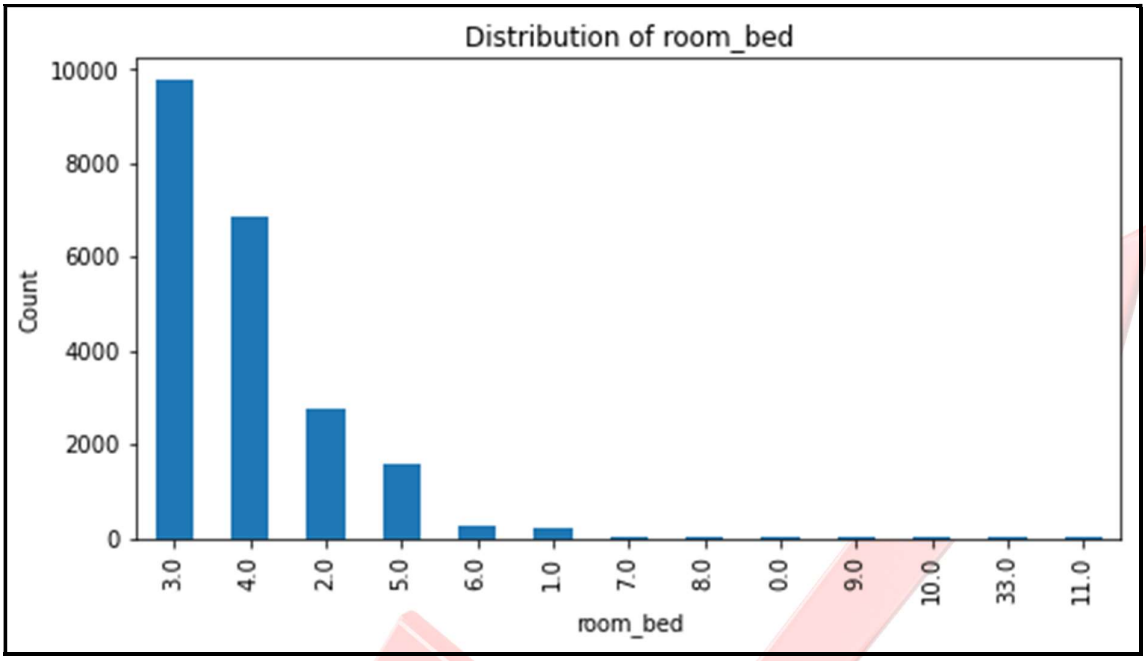


Figure 7 Distribution of No. of Bedrooms

'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.***

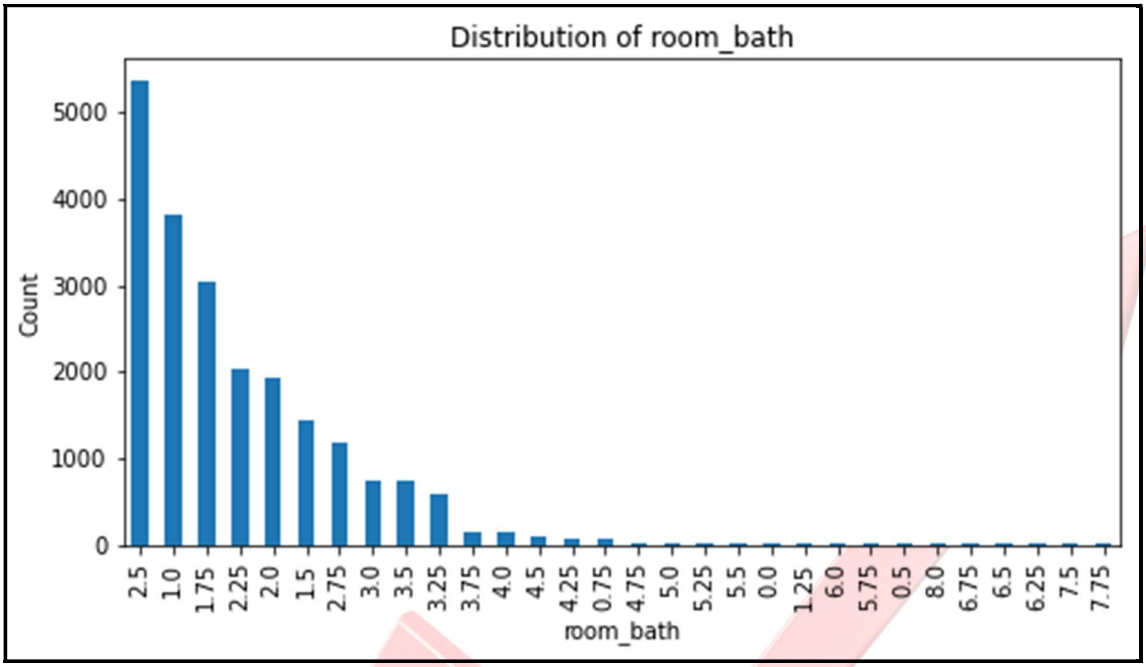


Figure 8 Distribution of No. of bathrooms

'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:

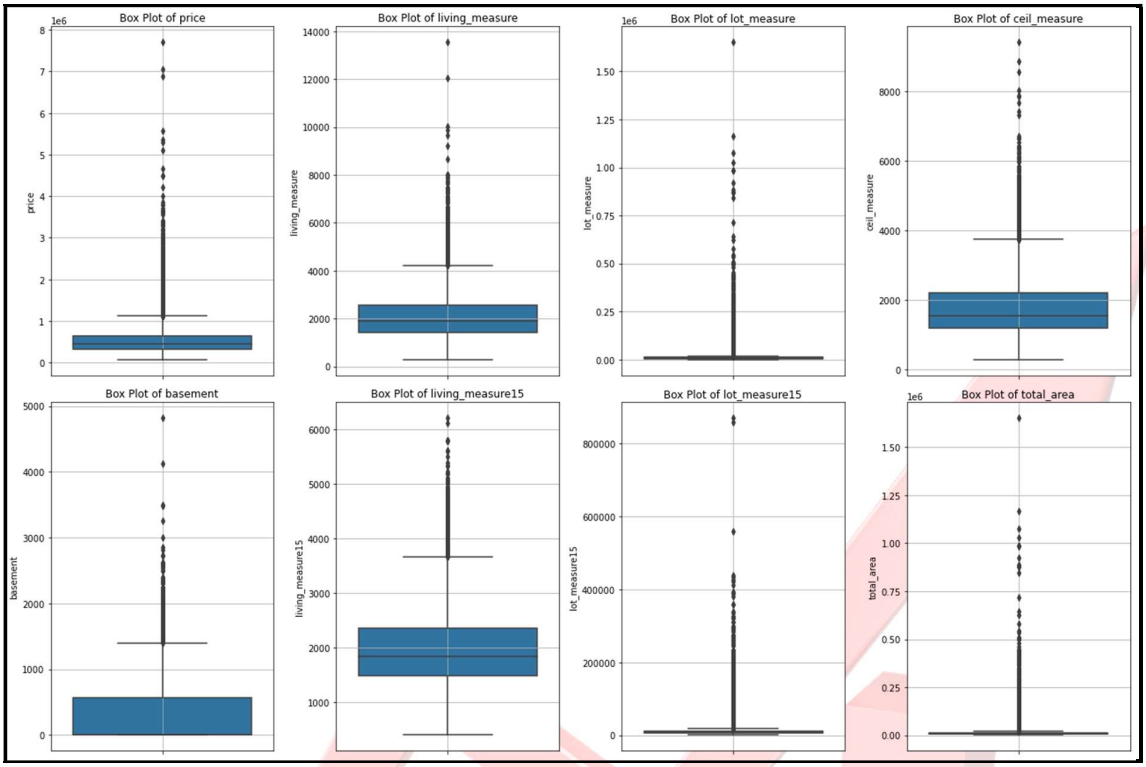


Figure 9 Box plots for all the numerical attributes

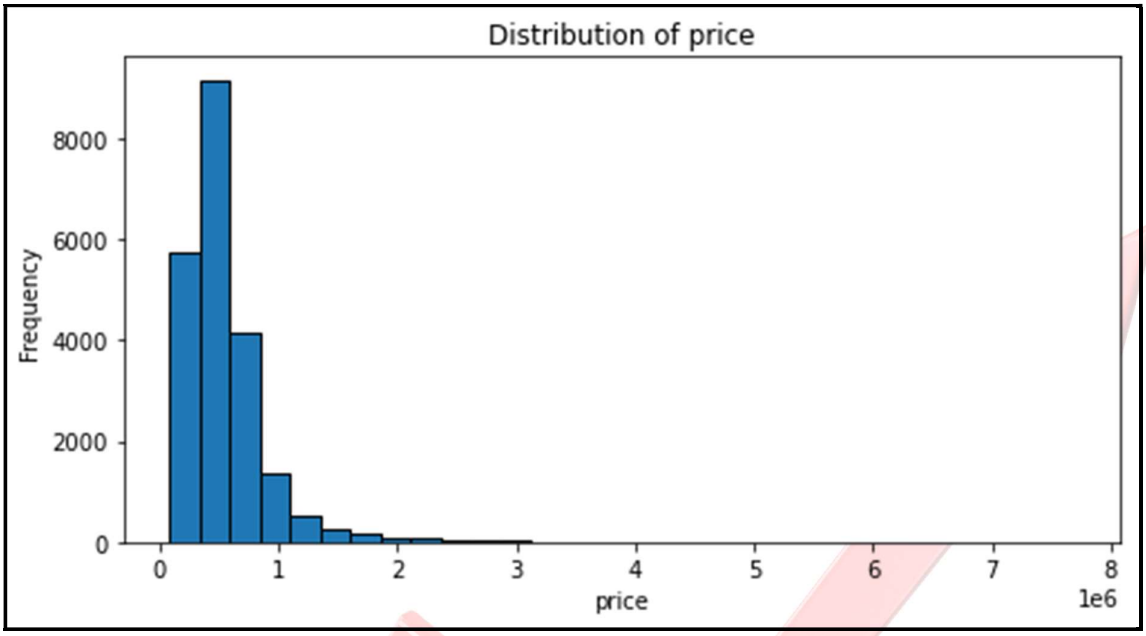


Figure 10 Distribution of price

'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.***

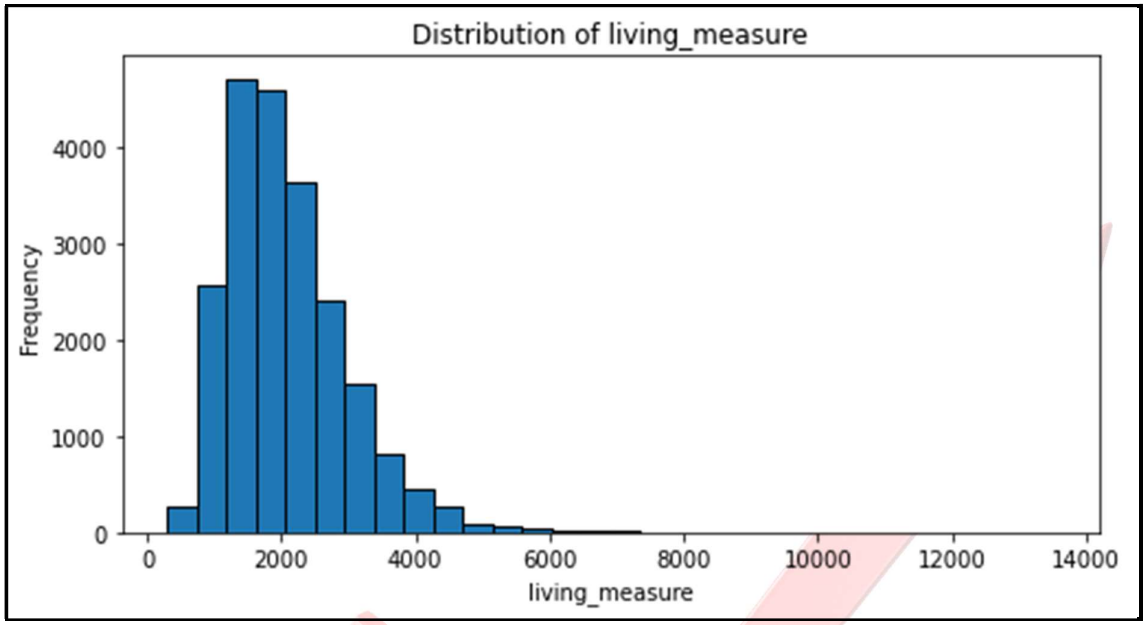


Figure 11 Distribution of Living measure

'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.***

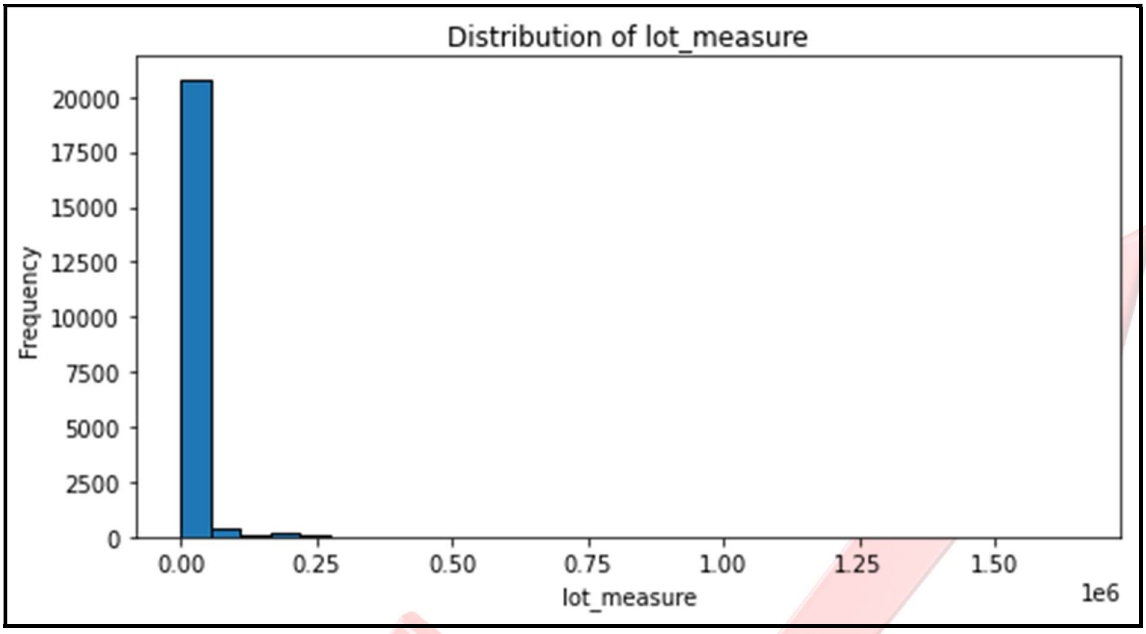


Figure 12 Distribution of lot measure

'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.***

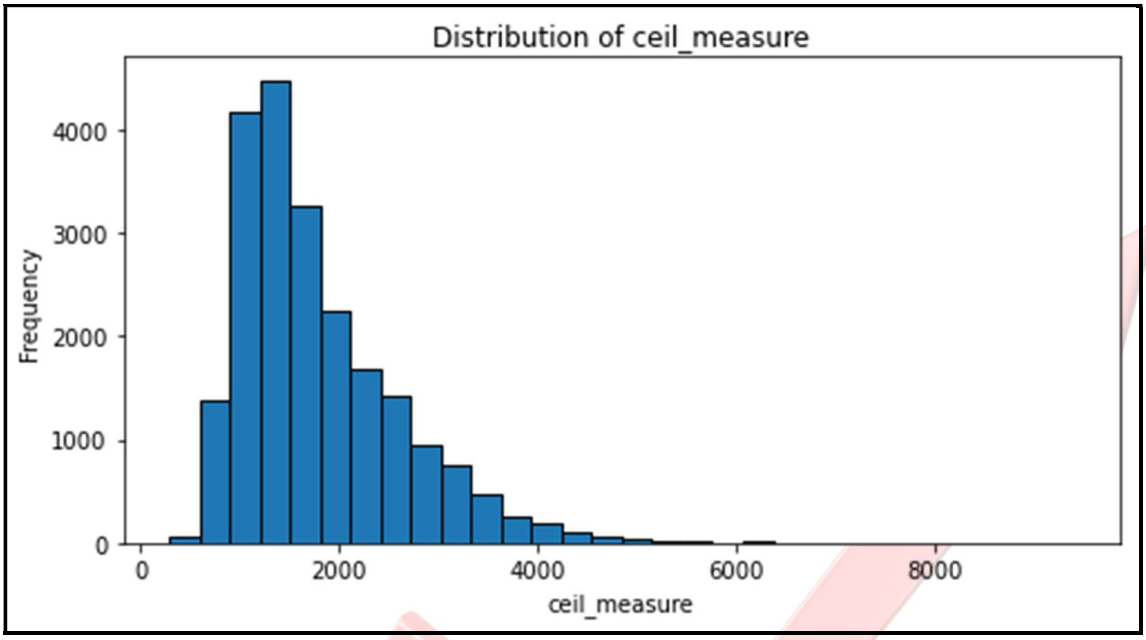


Figure 13 Distribution of ceil measure

'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.***

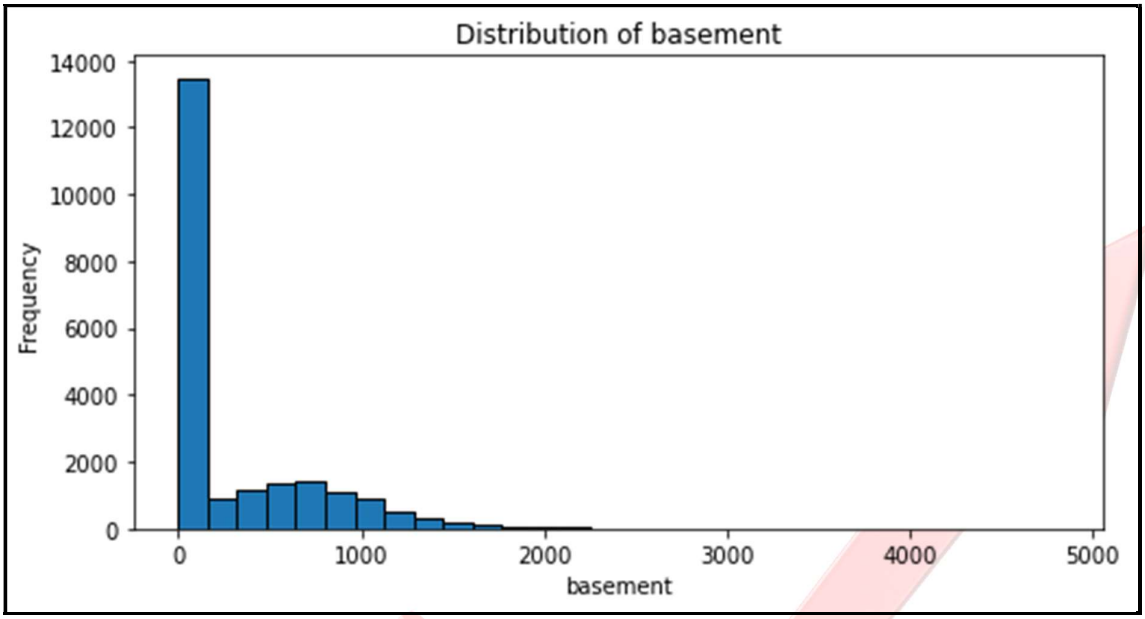


Figure 14 Distribution of basement

'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.***

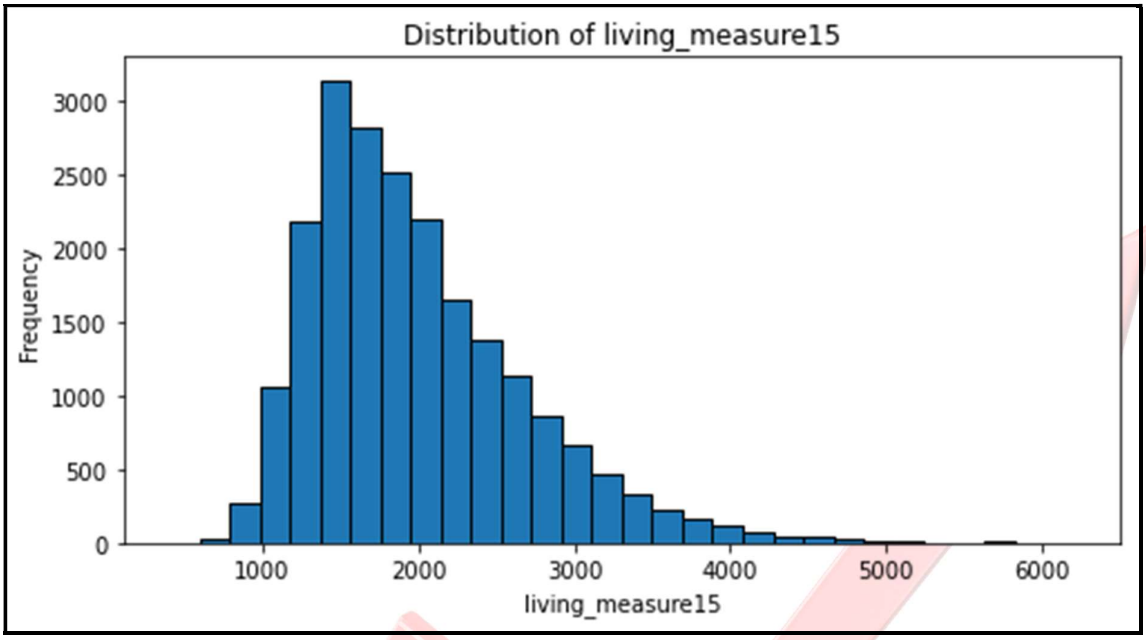


Figure 15 Distribution of Living measure15

'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.***

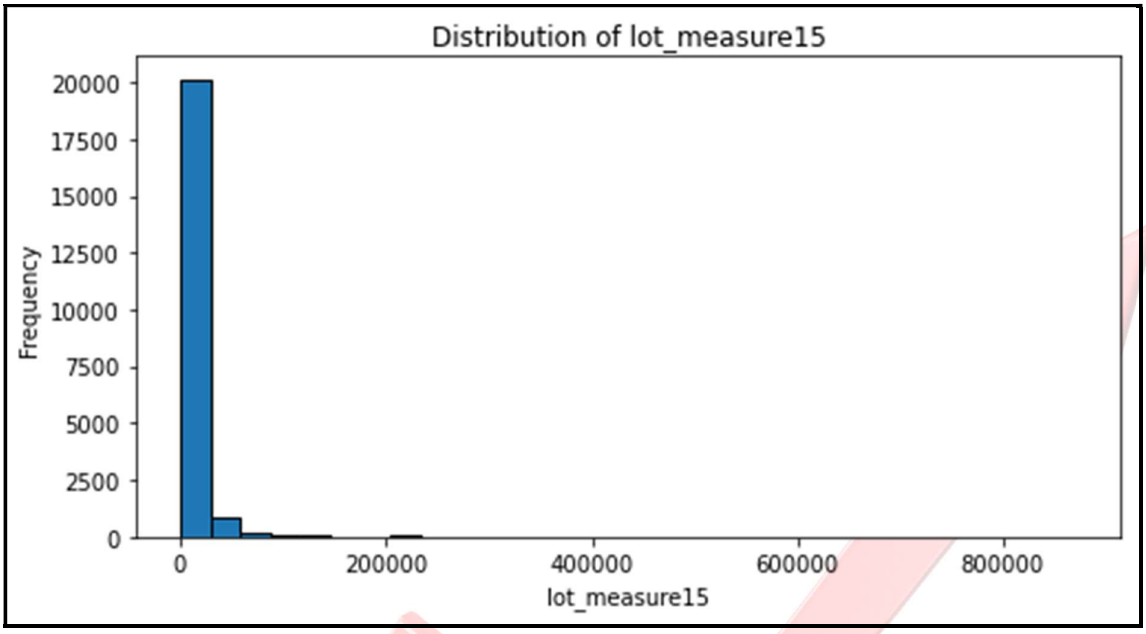


Figure 16 Distribution of lot measure15

'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.***

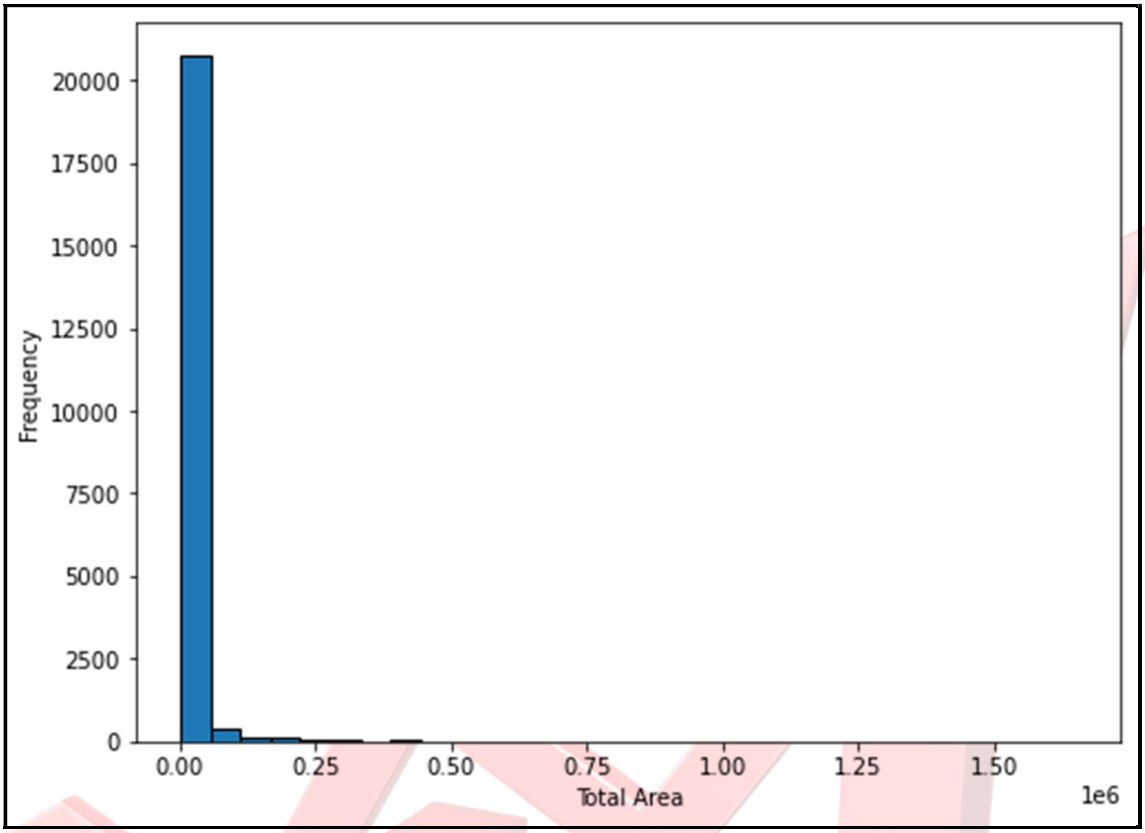


Figure 17 Distribution of Total Area

'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).

Bivariate Analysis:

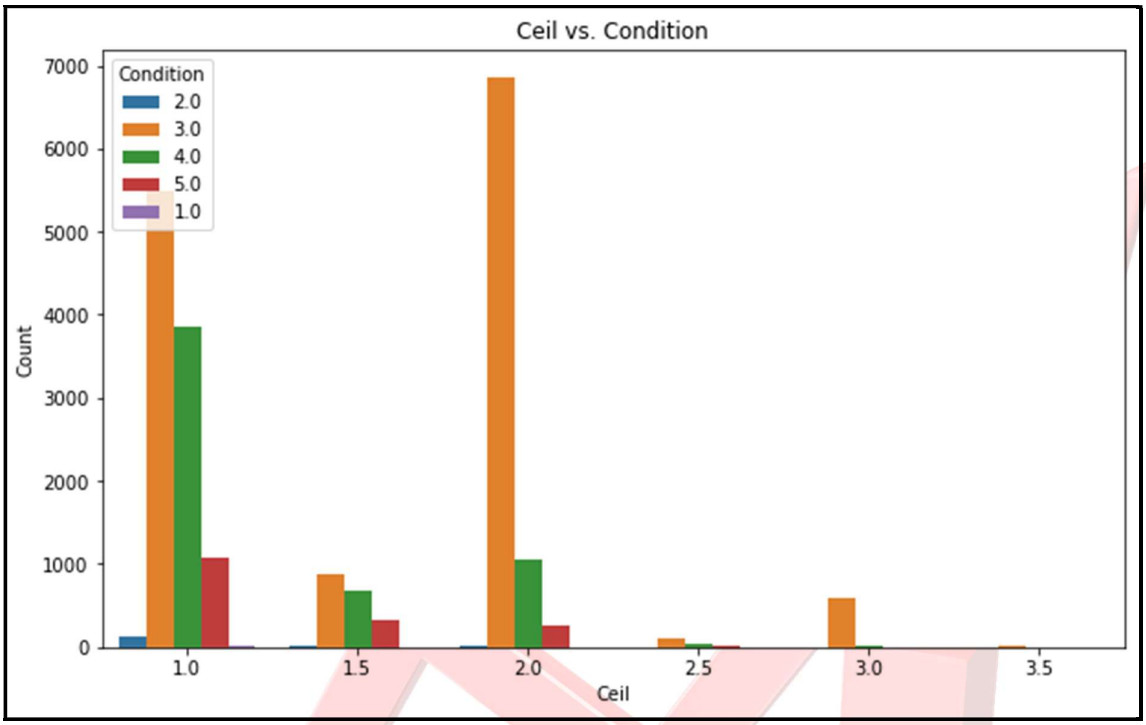


Figure 18 Ceil Vs Condition

- ***The cross-tabulation provides a summary of the relationship between the ceiling type ('ceil') and the house condition ('condition').***
- ***The table shows the count of houses with different combinations of ceiling type and condition.***
- ***For example, there are 23 houses with a condition of 1.0 and a ceiling type of 1.0.***
- ***It appears that houses with different ceiling types (1.0, 1.5, 2.0, etc.) are distributed across various condition ratings (1.0, 2.0, 3.0, 4.0, and 5.0).***

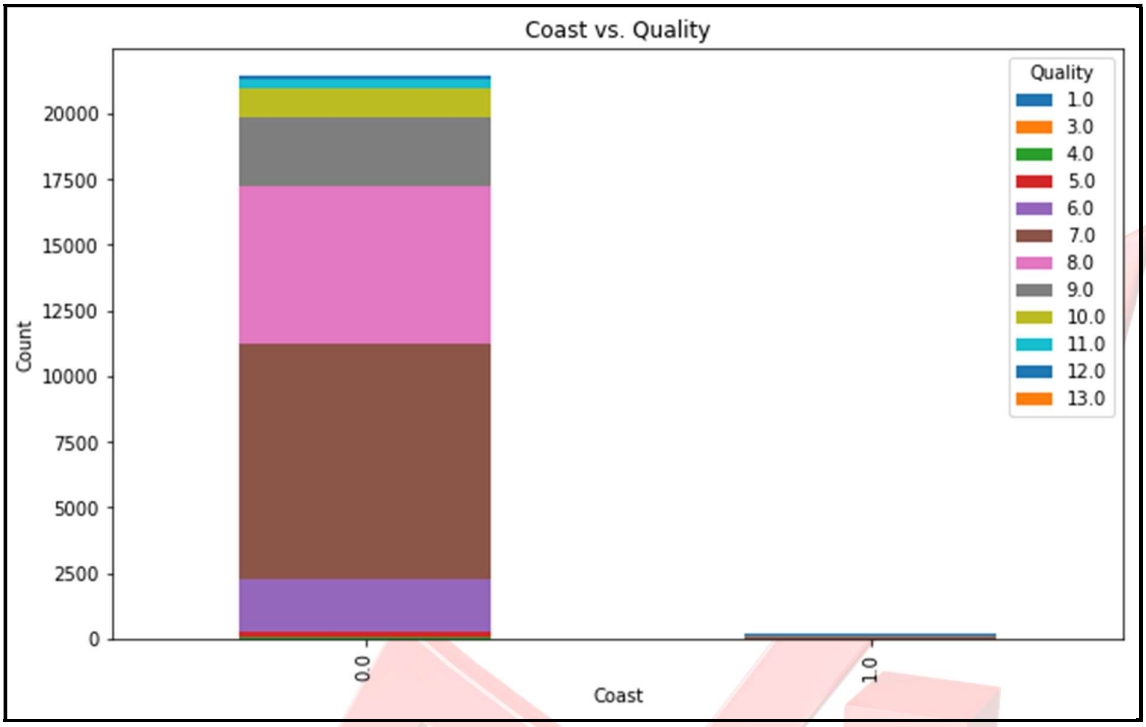


Figure 19 Coast Vs Quality

- ***The cross-tabulation provides a summary of the relationship between the proximity to the coast ('coast') and the quality of houses ('quality').***
- ***The table shows the count of houses with different combinations of coast proximity and quality.***
- ***Most houses with quality ratings are not located near the coast (coast value 0.0).***
- ***Few houses with quality ratings are located near the coast (coast value 1.0).***
- ***This suggests that higher-quality houses are not necessarily clustered near the coast.***

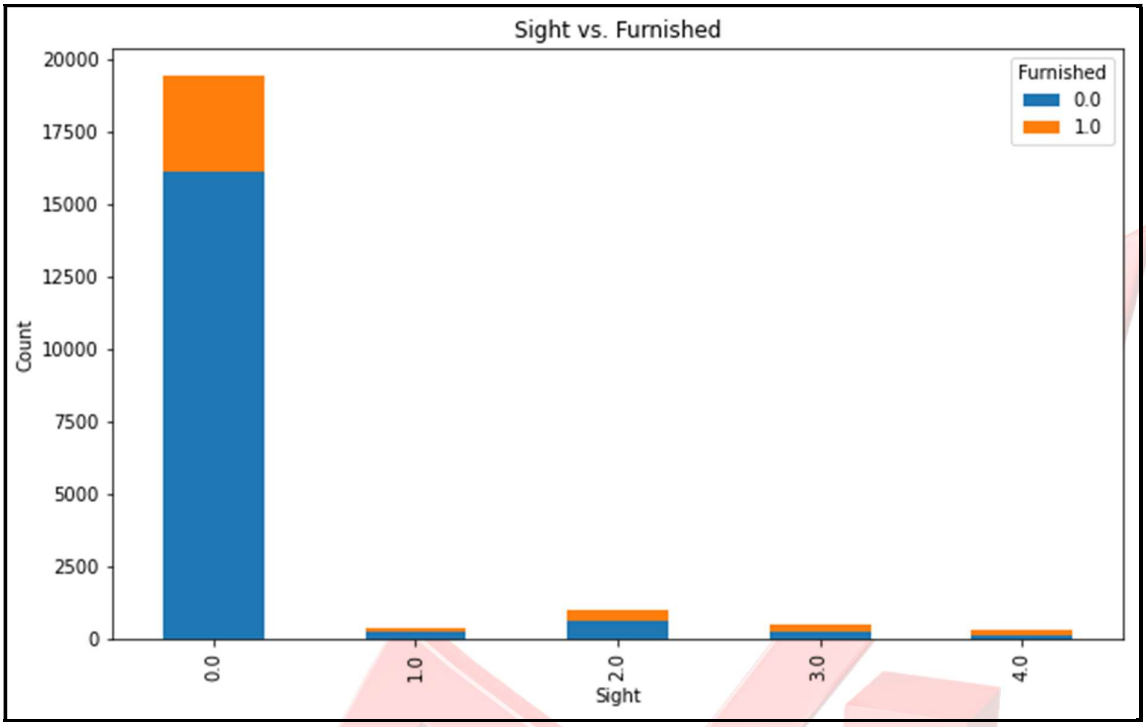


Figure 20 Sight Vs Furnished

- ***The cross-tabulation summarizes the relationship between the sight rating ('sight') and whether the house is furnished ('furnished').***
- ***The table shows the count of houses with different combinations of sight rating and furnished status.***
- ***Most houses with various sight ratings are not furnished (furnished value 0.0).***
- ***A smaller number of houses with different sight ratings are furnished (furnished value 1.0).***
- ***This provides insights into the distribution of furnished and unfurnished houses based on their sight ratings.***

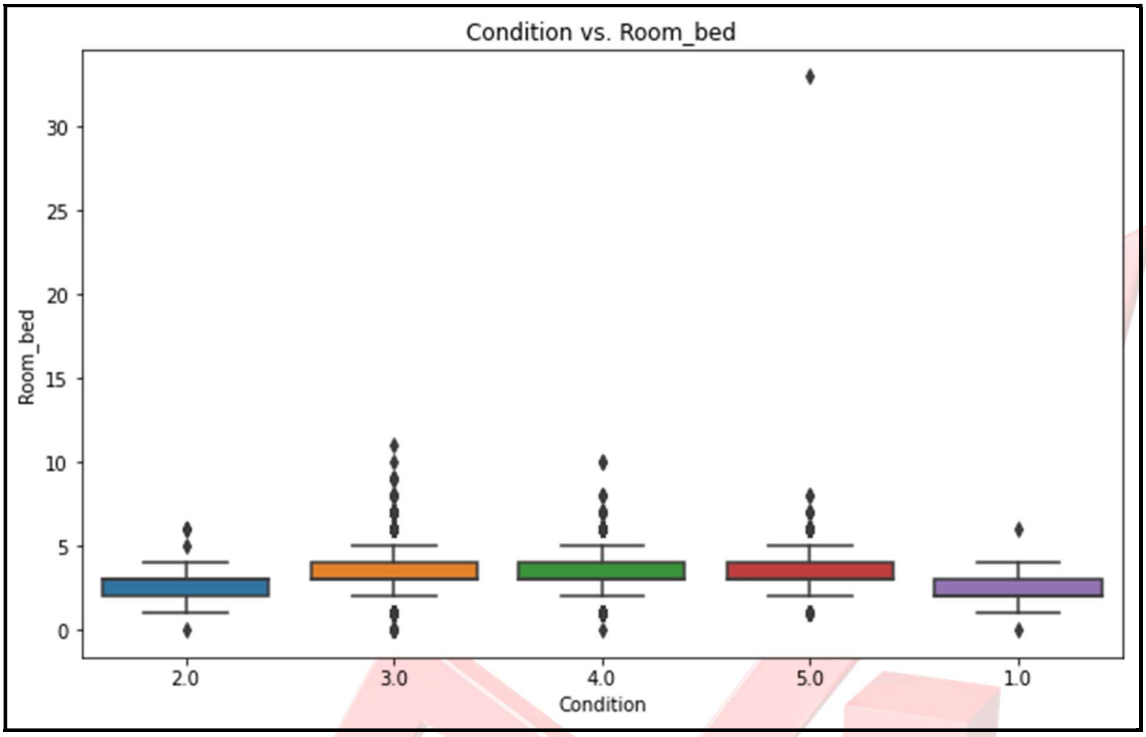


Figure 21 Condition Vs Room bed

- ***The summary statistics provide information about the relationship between the house condition ('condition') and the number of bedrooms ('room_bed').***
- ***It includes count, unique values, the most frequent (top) value, and its frequency.***
- ***For example, houses with a condition of 3.0 most frequently have 3 bedrooms.***
- ***This information helps understand the typical number of bedrooms in houses with different condition ratings.***

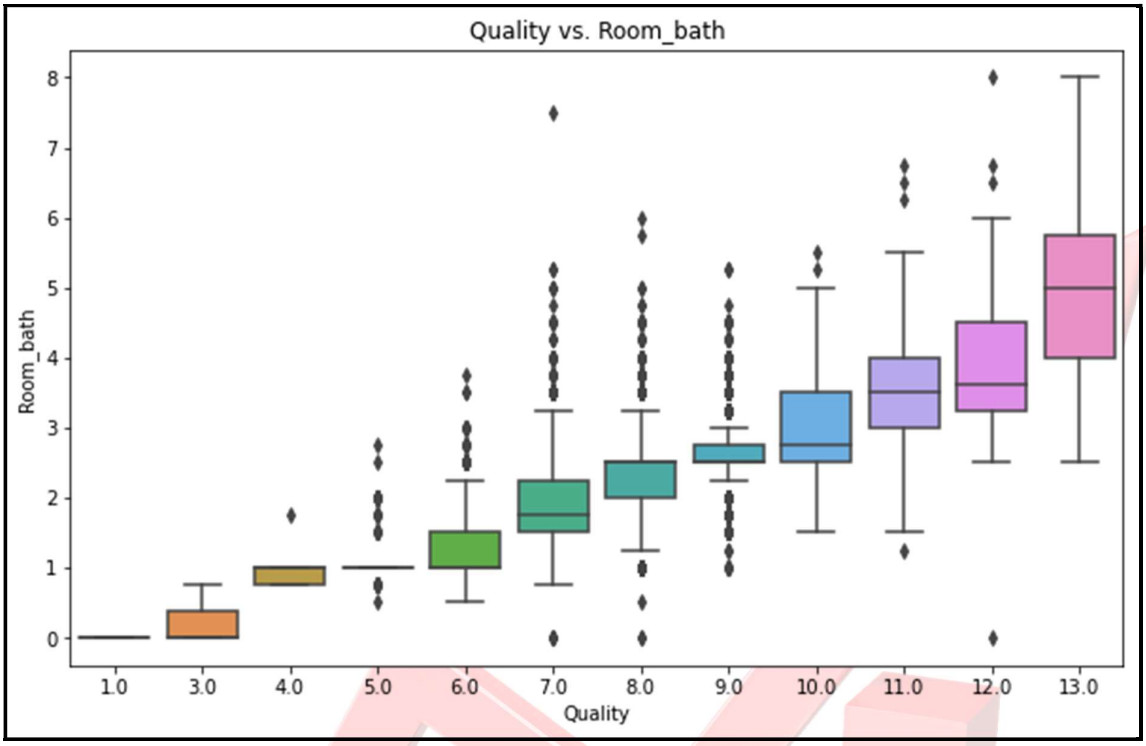


Figure 22 Quality Vs Room bath

- ***The summary statistics provide information about the relationship between the house quality ('quality') and the number of bathrooms ('room_bath').***
- ***It includes count, unique values, the most frequent (top) value, and its frequency.***
- ***For example, houses with a quality rating of 7.0 most frequently have 2.50 bathrooms.***

This information helps understand the typical number of bathrooms in houses with different quality ratings.

These cross-tabulations and summary statistics offer insights into the relationships between various attributes in our dataset. They can be valuable for understanding patterns and associations between different features, which can inform further analysis and decision-making in our project.

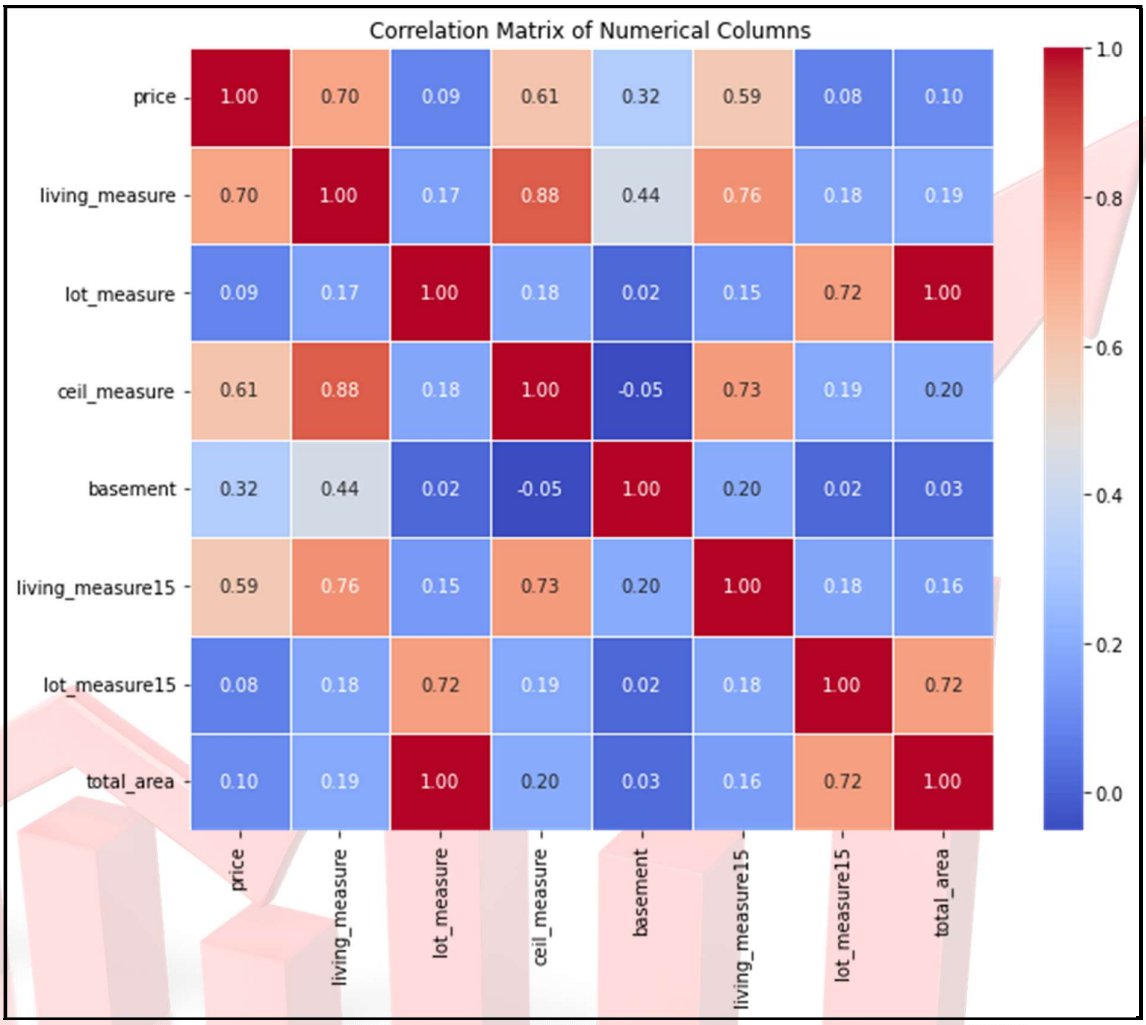


Figure 23 Heat Graph

Price vs. Living Measure:

- ***Correlation: 0.702149***
- ***Inference: There is a strong positive correlation (0.70) between the price of the house and its living area size. This suggests that as the living area size increases, the price of the house tends to increase as well.***

Price vs. Lot Measure:

- ***Correlation: 0.089733***
- ***Inference: There is a relatively weak positive correlation (0.09) between the price of the house and its lot size. This indicates that lot size has a weaker influence on house price compared to living area size.***

Price vs. Ceiling Measure

- ***Correlation: 0.605593***
- ***Inference: There is a strong positive correlation (0.61) between the price of the house and its ceiling area size. This suggests that houses with larger ceiling areas tend to have higher prices.***

Price vs. Basement:

- ***Correlation: 0.323825***
- ***Inference: There is a moderate positive correlation (0.32) between the price of the house and the size of the basement. This implies that houses with larger basements tend to have higher prices.***

Price vs. Living Measure 15

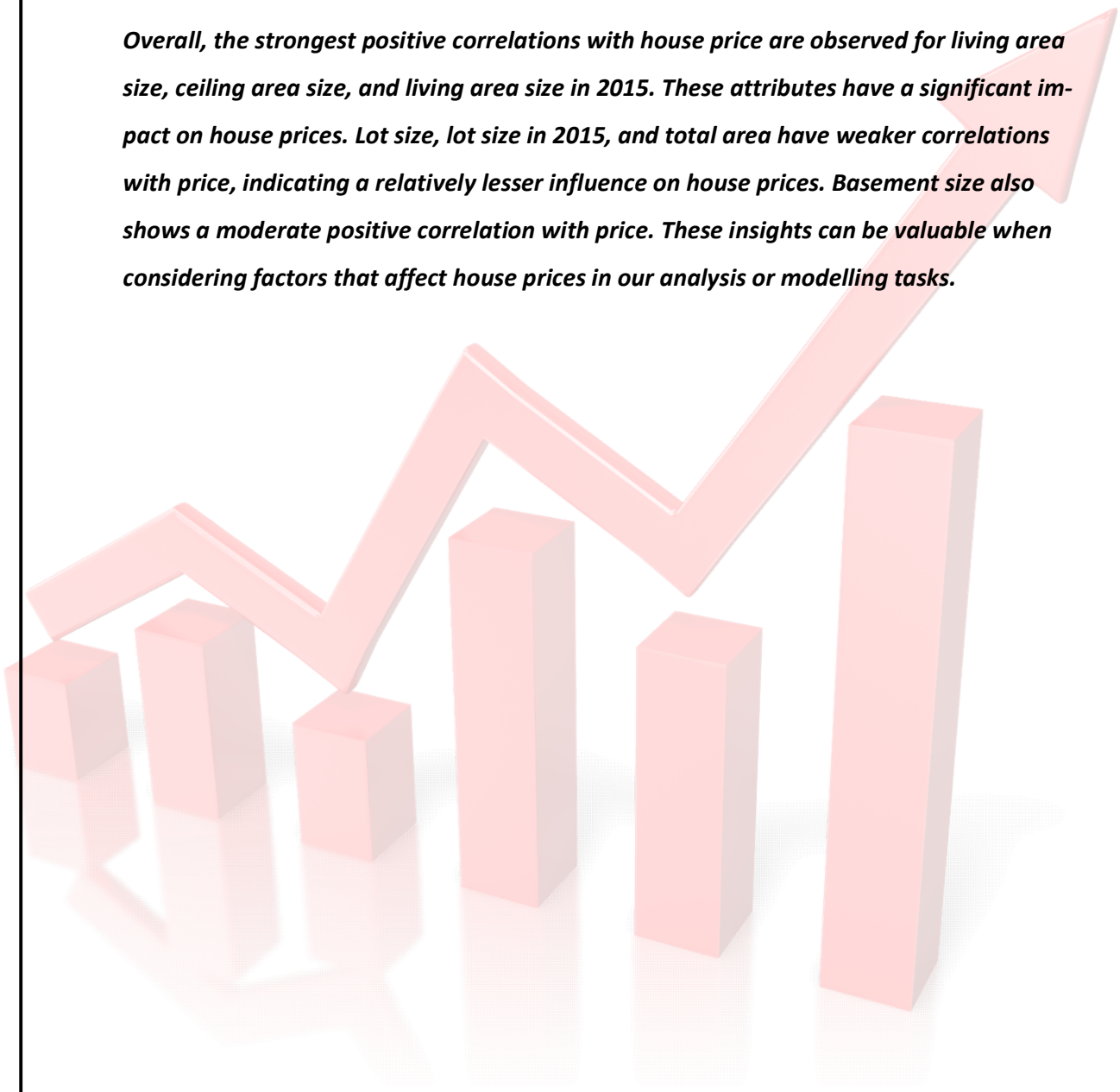
- ***Correlation: 0.585186***
- ***Inference: There is a strong positive correlation (0.59) between the price of the house and the living area size in 2015. This suggests that the size of the living area in 2015 is also a strong predictor of house price.***

Price vs. Lot Measure 15

- ***Correlation: 0.082603***
- ***Inference: There is a relatively weak positive correlation (0.08) between the price of the house and the lot size in 2015. Similar to the correlation with lot measure, this indicates that lot size in 2015 has a weaker influence on house price compared to living area size.***
- ***Price vs. Total Area:***
- ***Correlation: 0.104929***

- ***Inference: There is a relatively weak positive correlation (0.10) between the price of the house and the total area (sum of living area and lot size). This suggests that the total area has a moderate but weaker influence on house price compared to living area size.***

Overall, the strongest positive correlations with house price are observed for living area size, ceiling area size, and living area size in 2015. These attributes have a significant impact on house prices. Lot size, lot size in 2015, and total area have weaker correlations with price, indicating a relatively lesser influence on house prices. Basement size also shows a moderate positive correlation with price. These insights can be valuable when considering factors that affect house prices in our analysis or modelling tasks.



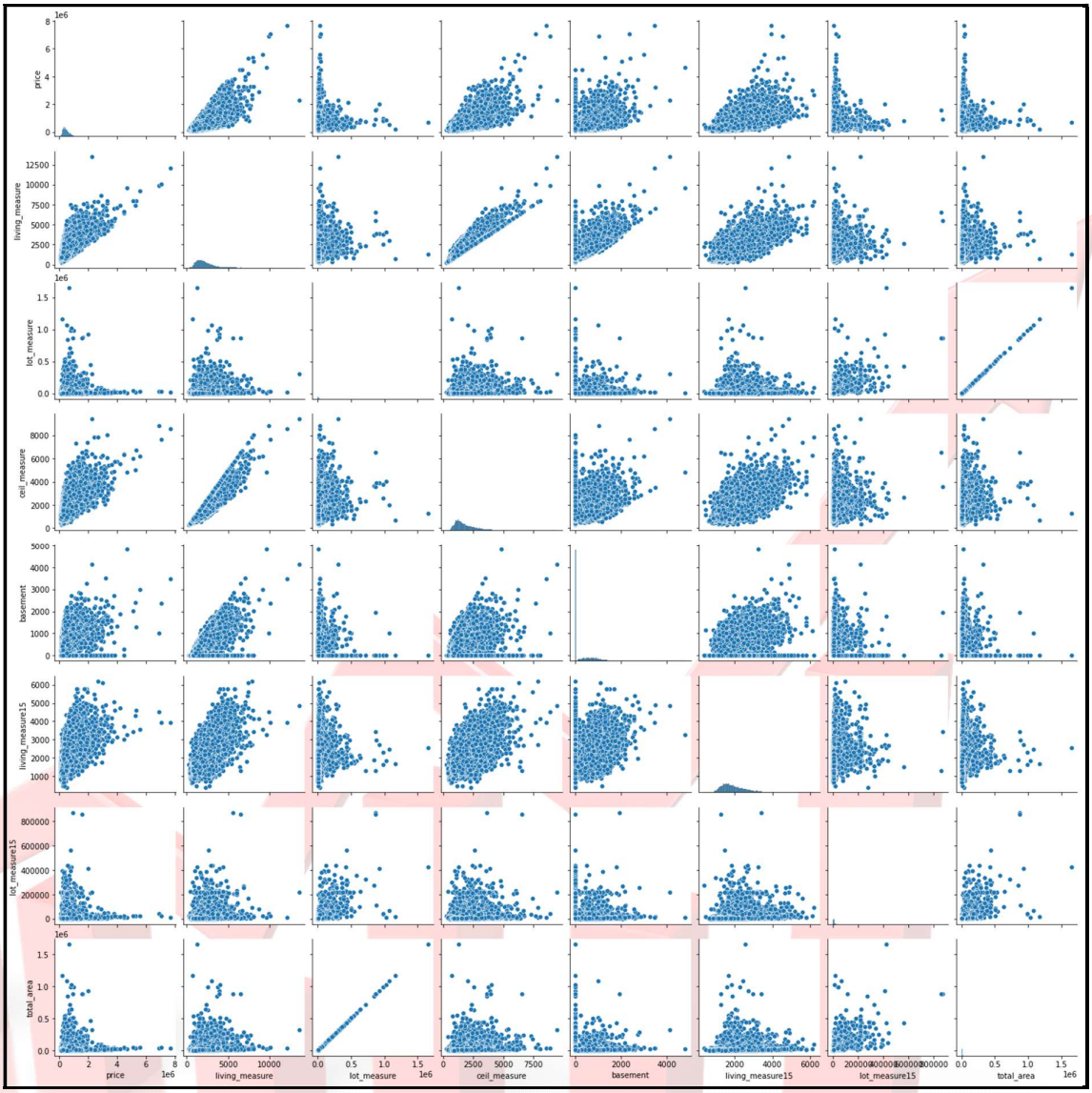


Figure 24 Pair Plot

Price vs. Living Measure:

- ***Correlation: 0.702149***
- ***Inference: There is a strong positive correlation (0.70) between the price of the house and its living area size. As the living area size increases, the price of the house tends to increase as well.***

Price vs. Ceiling Measure

- ***Correlation: 0.605593***
- ***Inference: There is a strong positive correlation (0.61) between the price of the house and its ceiling area size. This suggests that houses with larger ceiling areas tend to have higher prices.***

Living Measure vs. Ceiling Measure:

- ***Correlation: 0.876653***
- ***Inference: There is a very strong positive correlation (0.88) between the living area size and ceiling area size. This indicates that as the living area size increases, the ceiling area size tends to increase as well.***

Price vs. Living Measure 15:

- ***Correlation: 0.585186***
- ***Inference: There is a strong positive correlation (0.59) between the price of the house and the living area size in 2015. This suggests that the size of the living area in 2015 is also a strong predictor of house price.***

Lot Measure vs. Lot Measure 15:

- ***Correlation: 0.717727***
- ***Inference: There is a strong positive correlation (0.72) between the lot size and the lot size in 2015. This indicates that the lot size tends to be consistent over time.***

Total Area vs. Lot Measure and Total Area vs. Lot Measure 15

- ***Correlation: 0.999764 (with lot measure) and 0.719328 (with lot measure 15)***
- ***Inference: There is an extremely strong positive correlation (0.9998) between the total area (sum of living area and lot size) and the lot size. This is expected since the total area is the sum of its components.***

These correlations provide valuable insights into the relationships between different attributes in our dataset. Notably, living area size, ceiling area size, living area size in 2015, and lot size appear to be positively correlated with house price, indicating their importance in influencing house prices. Additionally, the high correlation between living area size and ceiling area size suggests that these attributes are closely related.

c) Removal of unwanted variables (if applicable).

Based on this analysis, it's evident that the "dayhours" and "yr_renovated" columns have the highest percentage of missing values (100% and 95.77%, respectively). These columns contain a substantial amount of missing data and may not provide meaningful information for analysis.

Therefore, it is reasonable to remove these two columns from our dataset to eliminate the high proportion of missing values and ensure that our analysis and modelling are not significantly affected by these missing values. Removing these columns will help our work with a cleaner and more informative dataset.

d) Missing Value treatment (if applicable) d) Outlier treatment (if required).

I, proceed to treat outliers in the selected numerical columns using the IQR method.

The outliers are identified and then capped to the lower and upper bounds determined by the IQR.

This helps in mitigating the impact of outliers on subsequent analyses or modelling tasks.

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

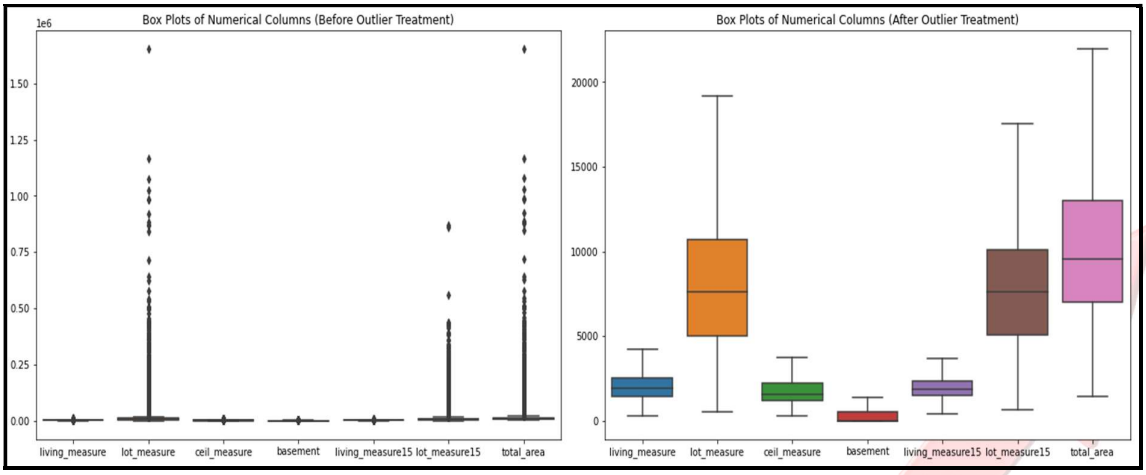


Figure 25 Before and after treatment of outliers

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.***

Therefore, it is reasonable to remove these two columns from our dataset to eliminate the high proportion of missing values and ensure that our analysis and modelling are not significantly affected by these missing values. Removing these columns will help our work with a cleaner and more informative dataset.

e) Variable transformation (if applicable).

For the 'dayhours' column, I acknowledge the suggestion to perform variable transformation. The 'dayhours' column indeed contains date and time information, but it is currently in the form of an object data type. To make the data more suitable for time-series analysis, I plan to convert the 'dayhours' column into a datetime data type. Additionally, I have observed that there are several derived columns such as 'date,' 'year,' 'month,' 'day,' 'hour,' 'minute,' and 'second' that either have already been extracted from the 'dayhours' column or can be derived for further time-based analysis. These transformations will help in handling and utilizing the time-related aspects of the data effectively.

f) Addition of new variables (if required).

It's a good approach to assess the need for new variables as our progress with data analysis and modelling. Adding new variables should be driven by specific insights or hypotheses we want to test, and it's perfectly fine to make that decision during the modelling phase based on the requirements and objectives of our analysis. If I, identify any specific

variables that could enhance the predictive power of our model or provide valuable insights, I, always consider adding them at that stage.

4) Business insights from EDA

a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business.

To check if the data is unbalanced, we can examine the distribution of the target variable in a classification problem or the distribution of the response variable in a regression problem. Here's how we can check for class imbalance in a classification problem:

For Classification Problems (e.g., Predicting Binary Outcomes):

- ***If we're dealing with a binary classification problem (two classes), our can calculate the distribution of the target variable. For example, if we predicting whether an email is spam (1) or not spam (0), our can count the number of samples for each class.***
- ***If we're dealing with a multi-class classification problem (more than two classes), our can calculate the distribution of each class in the target variable.***
- ***Visualize the distribution using a bar plot or a pie chart to see if there's a significant class imbalance. If one class dominates the dataset (e.g., 90% of the samples belong to one class), it indicates class imbalance.***
- ***Calculate and compare the proportions of each class. If there's a substantial imbalance (e.g., one class has much fewer samples than the others), it can affect the performance of our machine learning model.***

For Regression Problems (e.g., Predicting Continuous Values):

Capstone Project notes-I-4

- *In regression problems, we can plot the distribution of the response variable (the variable we are trying to predict) to see if it is skewed or has outliers.*
- *Check for extreme values or outliers in the response variable. Outliers can sometimes indicate data imbalance, especially if they are far from the majority of data points.*

In our specific case of predicting house prices, it's a regression problem. We can visualize the distribution of house prices (the target variable) and look for any skewness or outliers. If we have extreme outliers, it's essential to decide how to handle them, as they can impact the performance of our regression model.

Our can use tools like histograms, box plots, and summary statistics to assess the distribution of the target variable (house prices) and identify any potential issues related to imbalance or outliers.

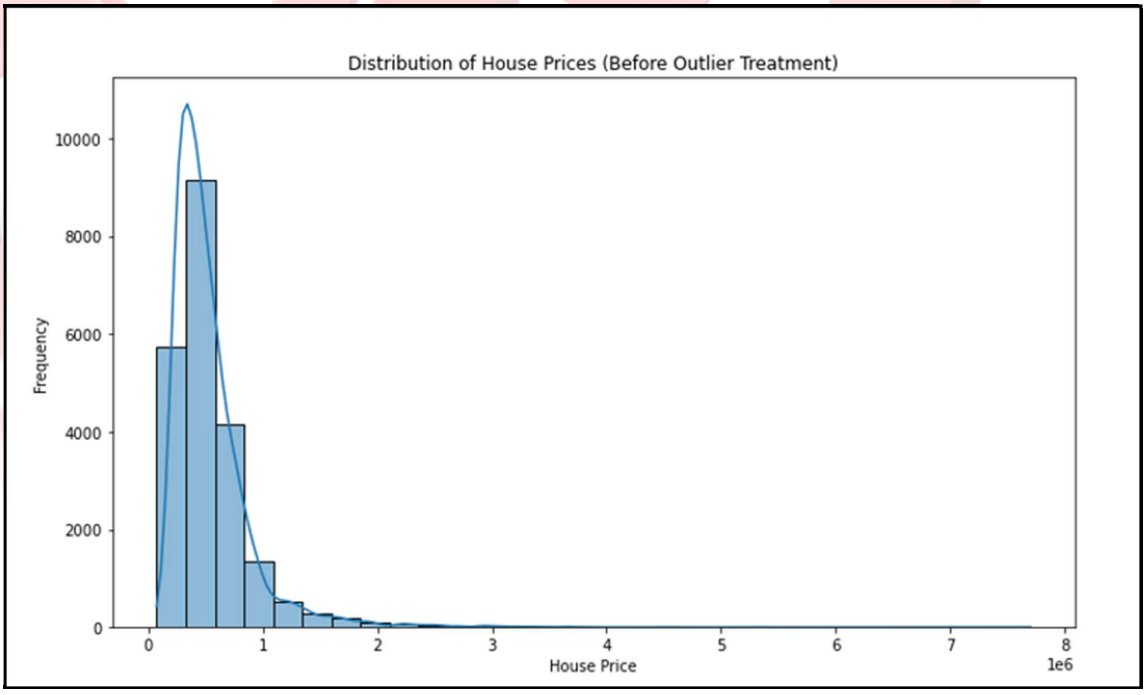


Figure 26 Distribution of Price Before Outlier Treatment

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

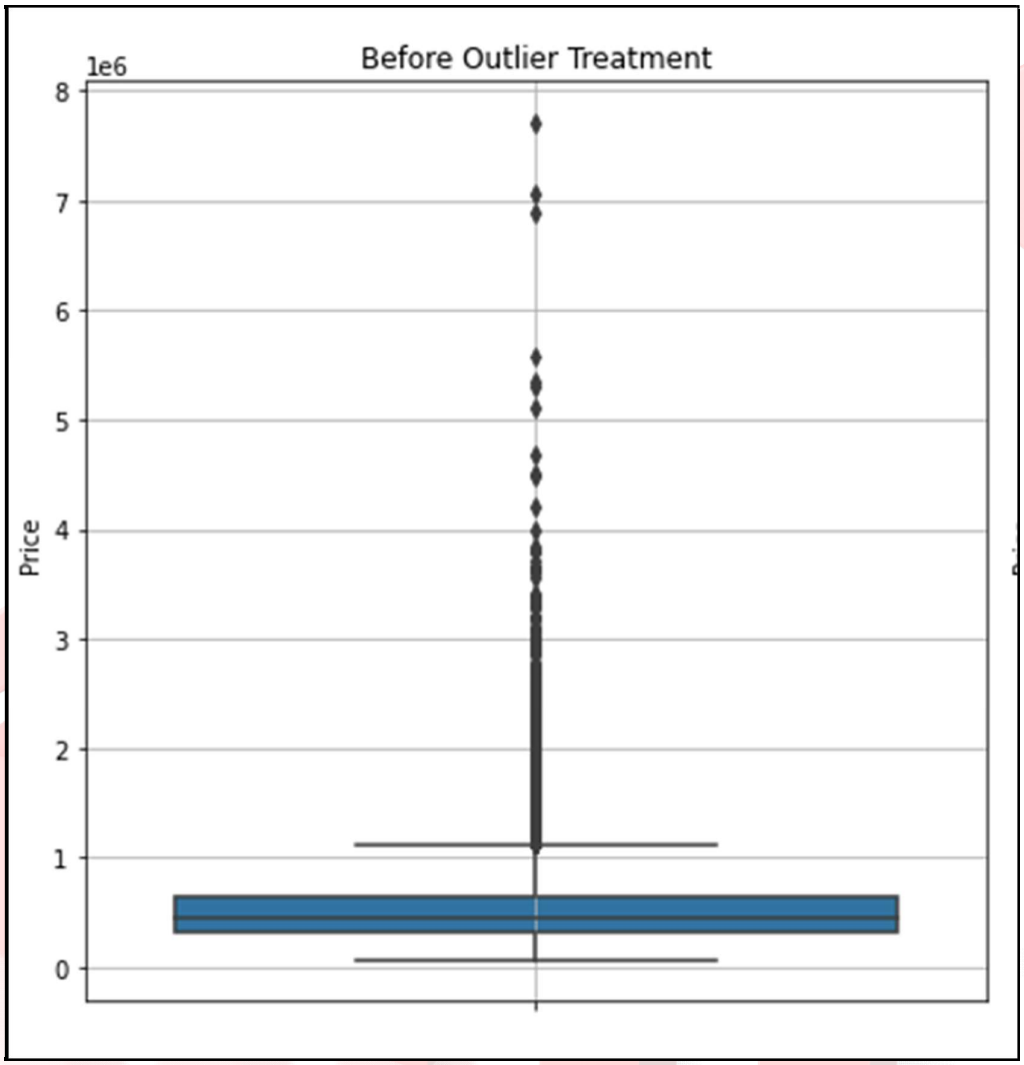


Figure 27 Box plot for Price Attribute

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.

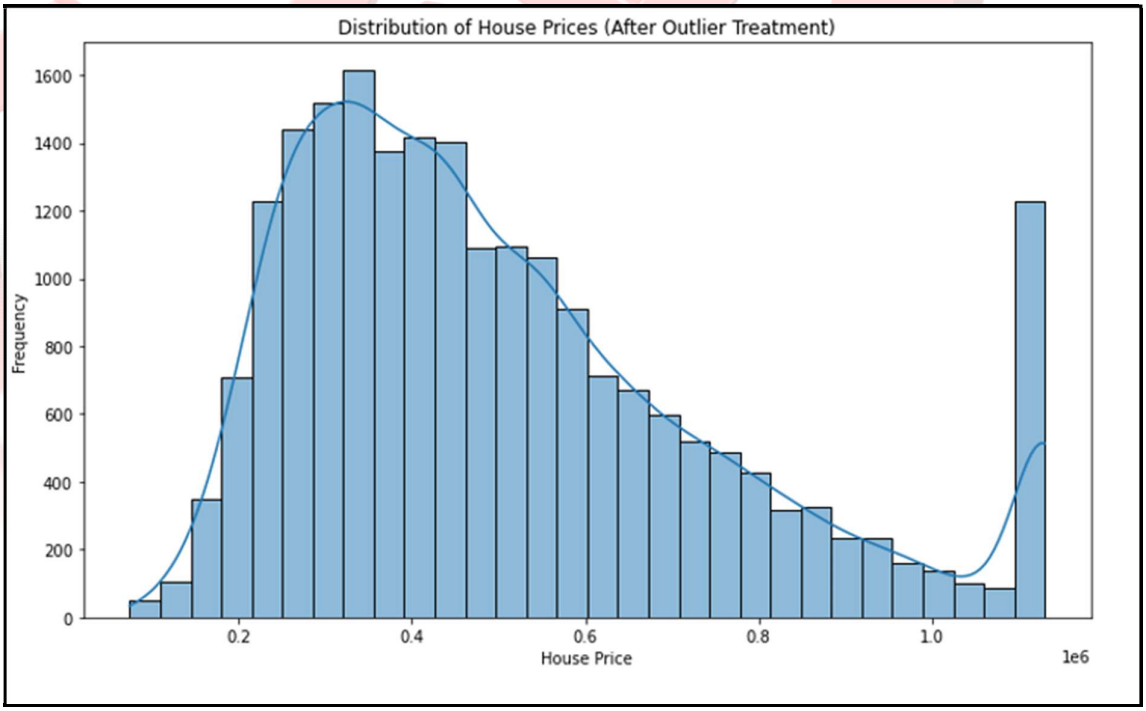


Figure 28 Distribution of Price After Outlier Treatment

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.

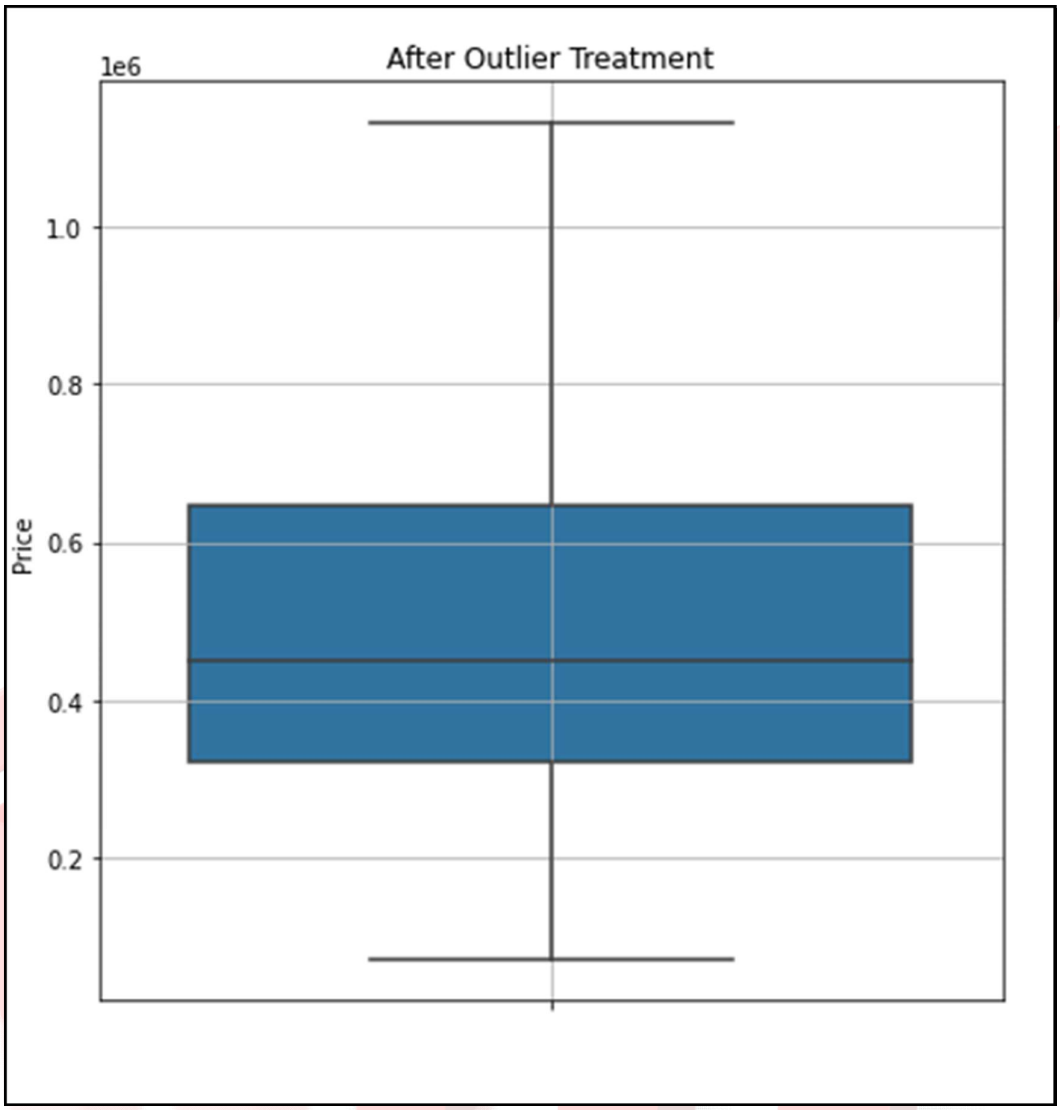


Figure 29 Box plot after outlier treatment

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

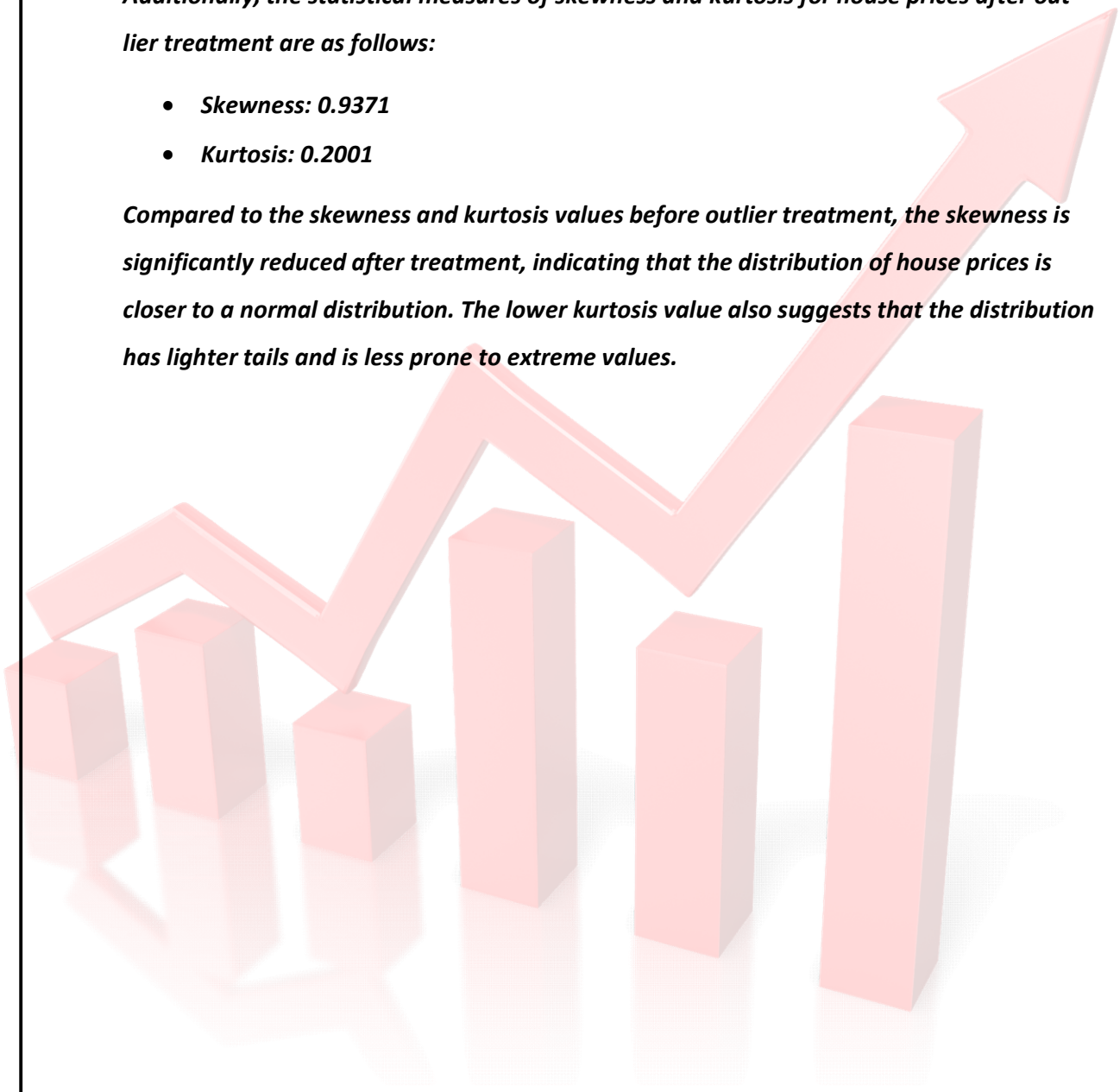
- ***Count: 21,613***
- ***Mean: $511,607.60***
- ***Standard Deviation: $250,047.90***
- ***Minimum Price: $75,000.00***
- ***25th Percentile (Q1): $321,950.00***

- ***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.



b) Any business insights using clustering (if applicable).

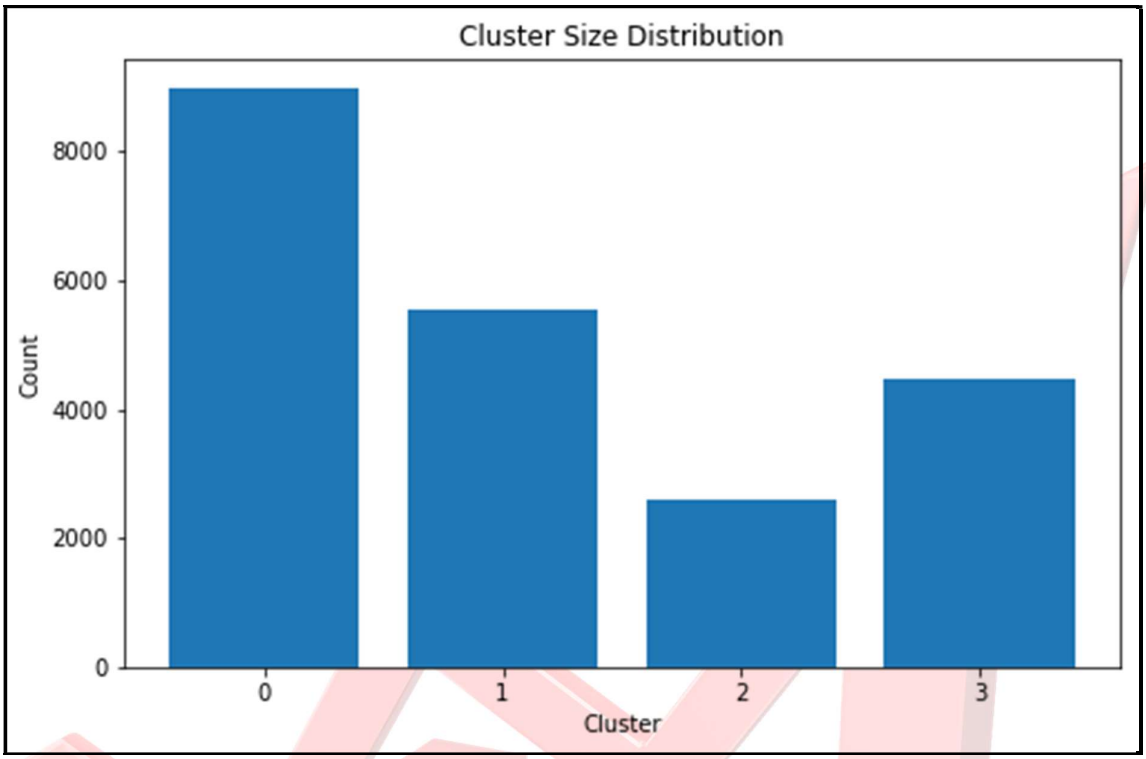


Figure 30 Cluster Size Distribution

Here is the cluster size distribution:

- ***Cluster 0: 8981 data points***
- ***Cluster 1: 5559 data points***
- ***Cluster 2: 2608 data points***
- ***Cluster 3: 4465 data points***

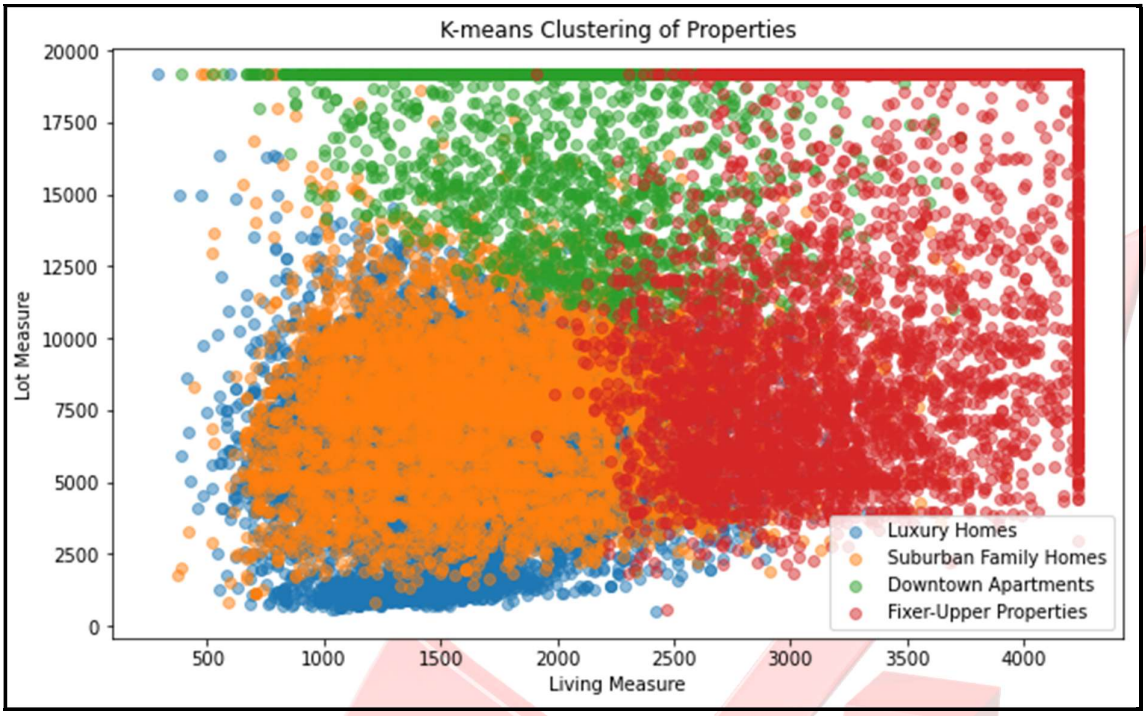


Figure 31 Scatter Plot for Clustering

Cluster 0 - Luxury Homes:

- ***Cluster 0 represents a group of properties that can be categorized as "Luxury Homes."***
- ***These properties have a moderate living measure on average (mean living measure: 1648.82 square feet) with relatively low variance (standard deviation: 497.37 square feet).***
- ***The minimum living measure in this cluster is 290.0 square feet, and the 25th percentile is 1280.0 square feet.***
- ***The median (50th percentile) living measure is 1620.0 square feet.***
- ***The maximum living measure in this cluster is 3330.0 square feet.***

Cluster 1 - Suburban Family Homes:

- *Cluster 1 represents a group of properties classified as "Suburban Family Homes."*
- *These properties also have a moderate living measure on average (mean living measure: 1707.66 square feet) with moderate variance (standard deviation: 562.86 square feet).*
- *The minimum living measure in this cluster is 370.0 square feet, and the 25th percentile is 1270.0 square feet.*
- *The median (50th percentile) living measure is 1660.0 square feet.*
- *The maximum living measure in this cluster is 3920.0 square feet.*

Cluster 2 - Downtown Apartments:

- *Cluster 2 represents properties categorized as "Downtown Apartments."*
- *These properties have a higher living measure on average (mean living measure: 2129.84 square feet) compared to the previous clusters, with moderate variance (standard deviation: 585.04 square feet).*
- *The minimum living measure in this cluster is 384.0 square feet, and the 25th percentile is 1740.0 square feet.*
- *The median (50th percentile) living measure is 2140.0 square feet.*
- *The maximum living measure in this cluster is 4231.13 square feet.*

Cluster 3 - Fixer-Upper Properties:

- *Cluster 3 represents properties identified as "Fixer-Upper Properties."*
- *These properties have the highest living measure on average (mean living measure: 3274.37 square feet) among all clusters, with moderate variance (standard deviation: 584.27 square feet).*
- *The minimum living measure in this cluster is 1910.0 square feet, and the 25th percentile is 2800.0 square feet.*
- *The median (50th percentile) living measure is 3190.0 square feet.*
- *The maximum living measure in this cluster is 4231.13 square feet.*

Insights for House Price Prediction:

- ***The "living_measure" (size of the property) is a crucial predictor across all clusters, with different cluster-specific trends.***
- ***"Quality" and "condition" of the property may significantly impact house prices, especially in luxury homes and suburban family homes.***
- ***Location-related features, such as "zipcode," "coast," and "sight," can be important predictors, especially for downtown apartments.***
- ***"Year built" or "yr_built" might also influence house prices, as older properties may require more maintenance or renovations.***

To build an effective house price prediction model, we should consider using these factors as features and improve the accuracy of price predictions.

Note: For more detailed information on the clustering analysis, including the code and insights derived from the clusters, please refer to the "Clustering.ipynb" Jupyter Notebook provided alongside this summary. This notebook contains step-by-step instructions, code implementation, and business insights extracted from the clustering analysis, offering a comprehensive view of the process and its results.

c) Any other business insights.

1. ***In conclusion, the comprehensive analysis of this dataset spanning the years 2014 and 2015 has provided us with crucial insights that lay the foundation for further exploration and modelling. We have meticulously examined the dataset's characteristics, including data collection time frames, variable types, and the presence of missing values. The data consists of both numerical and categorical attributes, totalling 23 columns, and encompasses 21,613 data points.***
2. ***Descriptive statistics have unveiled essential details about attribute distributions, with a special focus on the 'price' attribute. This attribute exhibited a right-skewed distribution with pronounced kurtosis, indicating the existence of outliers. To***

enhance modelling accuracy, outlier treatment was conducted, resulting in a more normally distributed 'price' attribute.

3. *Moreover, correlation analysis has shed light on potential relationships between numerical attributes and house prices. This exploration has set the stage for more advanced modelling techniques and feature selection. While we have gained valuable insights into the categorical attributes' distribution and relationships.*
4. *Based on the clustering analysis results, we can derive valuable business insights. The silhouette score of 0.3125 suggests a moderate level of cluster separation, indicating that there are discernible patterns within the dataset. Cluster 3, with a significantly higher mean living measure of 3274 square feet and the highest maximum living measure, seems to represent larger and potentially more luxurious properties. This cluster may cater to a higher-end market segment, and businesses targeting this cluster should focus on showcasing the spaciousness and luxury features of these properties.*
5. *Clusters 0, 1, and 2, with relatively lower mean living measures, may represent properties in different market segments. Businesses can tailor their marketing and pricing strategies for these clusters based on the preferences and needs of potential buyers. For example, Cluster 2 (mean living measure of 2129 square feet) might appeal to a mid-range market, while Cluster 0 (mean living measure of 1648 square feet) may attract buyers looking for more compact homes.*
6. *Armed with the knowledge and insights garnered from this preliminary analysis, we are poised to embark on the modelling phase, where we will leverage this information to develop predictive models and extract actionable business insights. The journey ahead promises to uncover the intricate dynamics of house prices and contribute to more informed decision-making in the real estate domain.*