CAPSTONE PROJECT

House Price Prediction

SARANG S PANICKER

PGDSBA.O.JULY23.A

INDEX

CONTENT

Sl no.	Particulars	Page no.
1	Introduction to business problem	3
2	Data Report	5
3	Exploratory Data Analysis	8
4	Data cleaning and pre-processing	20
5	Business insights from EDA	21
6	Model building and interpretation	22
7	Model tuning and validation	31
8	Final interpretation/recommendation	32

FIGURES AND GRAPHS

Sl No.	Particulars	Page no.
1	Descriptive statistics of data	5
2	Datatypes	6
3	Dataset info	7
4	Histogram and Boxplot of Price	8
5	Histogram and Boxplot of room_bed	8
6	Histogram and Boxplot of room_bath	9
7	Histogram and Boxplot of living_measure	9
8	Histogram and Boxplot of lot_measure	9
9	Histogram and Boxplot of ceil	10
10	Histogram and Boxplot of coast	10
11	Histogram and Boxplot of sight	10
12	Histogram and Boxplot of Condition	11
13	Histogram and Boxplot of Quality	11
14	Histogram and Boxplot of ceil_measure	11
15	Histogram and Boxplot of basement	12
16	Histogram and Boxplot of yr_built	12
17	Histogram and Boxplot of furnished	12
18	Histogram and Boxplot of living_measure15	13
19	Histogram and Boxplot of lot_measure15	13
20	Correlation	14

21	Price x Living Measure	15
22	Price x Quality	16
23	Price x Furnished	17
24	Price x Room Bath	18
25	Price x Ceil Measure	19
26	RSME of models (test)	25
27	R^2 of models (test)	26
28	Adjusted R^2 of models (test)	26
29	RSME, R^2 and adjusted R^2 of models (test)	27
30	RSME of models (train)	27
31	R^2 of models (train)	28
32	Adjusted R^2 of models (train)	28
33	RSME, R^2 and adjusted R^2 of models (train)	29
34	Performance metrics of models (test and train)	29
35	Feature importance	32

Introduction to business problem

a) Problem Statement

A house value is simply more than location and square footage. Like the features that make up a person, an educated party would want to know all aspects that give a house its value. For example, you want to sell a house and you don't know the price which you may expect — it can't be too low or too high. To find house price you usually try to find similar properties in your neighborhood and based on gathered data you will try to assess your house price.

b) Need of the study/project

Ever wondered why two houses which look identical cost different. This can be due to a lot of factors like – room size, number of rooms, square feet area, number of bathrooms, etc. House prices are decided by the market using various attributes. Therefore, it would be good to build a model which can read the dataset, learn from it, and then predict the price of houses.

This project can be divided into 3 parts:

- Exploratory Data Analysis
- Build models
- Model tuning
- Business recommendations

This project will be helpful for homeowners, real estate agents, and potential buyers to make decisions based on accurate property values. It can help:

- The sellers to price their homes competitively in the market
- Assist buyers in making fair offers
- Aid real estate agents in providing better advice to their clients

c) Understanding business/social opportunities

From a business perspective, developing a reliable model to predict house prices presents significant opportunities:

- Real Estate Companies: Can use predictive models to provide more accurate pricing strategies, improving customer satisfaction and sales efficiency.
- Financial Institutions: Lenders can better assess the value of properties when issuing mortgages.
- Government and Policy Makers: Accurate house price assessments can inform housing policies and economic strategies.

Socially, this project addresses the need for fairness and transparency in real estate transactions. Accurate pricing helps to:

- Ensure buyers and sellers are treated fairly.
- Reduce disputes and dissatisfaction in property transactions.
- Enhance the overall trust in the real estate market.

By using data analytics and machine learning to analyze various factors influencing house prices, the study aims to provide a more accurate and comprehensive valuation method. This approach not only benefits individual stakeholders but also contributes to a more stable and transparent real estate market.

1. Data Report

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

The data is based on an excel sheet with information about housing price of different years.

The steps which will be followed in this project is given below:

- Exploratory Data Analysis
- Build models
- Model tuning
- Business recommendations

b) Visual representation of data

- The number of rows (observations) is 21613
- The number of columns (variables) is 23

	count	mean	std	min	25%	50%	75%	max
cid	21,613.0	4,580,301,520.9	2,876,565,571.3	1,000,102.0	2,123,049,194.0	3,904,930,410.0	7,308,900,445.0	9,900,000,190.0
price	21,613.0	540,182.2	367,362.2	75,000.0	321,950.0	450,000.0	645,000.0	7,700,000.0
room_bed	21,613.0	3.4	0.9	0.0	3.0	3.0	4.0	33.0
room_bath	21,613.0	2.1	0.8	0.0	1.8	2.2	2.5	8.0
living_measure	21,613.0	2,079.9	918.4	290.0	1,427.0	1,910.0	2,550.0	13,540.0
lot_measure	21,613.0	15,107.0	41,420.5	520.0	5,040.0	7,618.0	10,688.0	1,651,359.0
ceil	21,613.0	1.5	0.5	1.0	1.0	1.5	2.0	3.5
coast	21,613.0	0.0	0.1	0.0	0.0	0.0	0.0	1.0
sight	21,613.0	0.2	0.8	0.0	0.0	0.0	0.0	4.0
condition	21,613.0	3.4	0.7	1.0	3.0	3.0	4.0	5.0
quality	21,613.0	7.7	1.2	1.0	7.0	7.0	8.0	13.0
ceil_measure	21,613.0	1,788.4	828.1	290.0	1,190.0	1,560.0	2,210.0	9,410.0
basement	21,613.0	291.5	442.6	0.0	0.0	0.0	560.0	4,820.0
yr_built	21,613.0	1,971.0	29.4	1,900.0	1,951.0	1,975.0	1,997.0	2,015.0
yr_renovated	21,613.0	84.4	401.7	0.0	0.0	0.0	0.0	2,015.0
zipcode	21,613.0	98,077.9	53.5	98,001.0	98,033.0	98,065.0	98,118.0	98,199.0
lat	21,613.0	47.6	0.1	47.2	47.5	47.6	47.7	47.8
long	21,613.0	-122.2	0.1	-122.5	-122.3	-122.2	-122.1	-121.3
living_measure15	21,613.0	1,986.6	685.4	399.0	1,490.0	1,840.0	2,360.0	6,210.0
lot_measure15	21,613.0	12,768.5	27,304.2	651.0	5,100.0	7,620.0	10,083.0	871,200.0
furnished	21,613.0	0.2	0.4	0.0	0.0	0.0	0.0	1.0
total_area	21,613.0	17,186.9	41,589.1	1,423.0	7,035.0	9,575.0	13,000.0	1,652,659.0

Fig 1 – Descriptive statistics of the data

As you can see from the descriptive data, we need to drop some columns like *zipcode* which are not necessary and also need to change the data types like *yr_built* and *yr_renovated*.

c) Understanding of attributes (variable info, renaming if required)

```
price                 int64
room_bed            float64
room_bath           float64
living_measure      float64
lot_measure         float64
ceil                 object
coast                object
sight               float64
condition            object
quality             float64
ceil_measure        float64
basement            float64
yr_built             object
yr_renovated          int64
lat                 float64
long                 object
living_measure15    float64
lot_measure15       float64
furnished           float64
total_area           object
dtype: object
```

Fig 2 - Datatypes

Here, some of the datatypes need to be changed. But before that, it has been noticed that there are some columns with non-numeric character. We need to remove them. It was found that the character was an alpha-numeric one which stands for dollar ($). Therefore, a code was run which replaced these characters with "np.nan".

There were 864 cells with null values, and these values were dropped.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 21288 entries, 0 to 21612
Data columns (total 20 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   price             21288 non-null  int64
 1   room_bed          21288 non-null  float64
 2   room_bath         21288 non-null  float64
 3   living_measure    21288 non-null  float64
 4   lot_measure       21288 non-null  float64
 5   ceil              21288 non-null  float64
 6   coast             21288 non-null  float64
 7   sight             21288 non-null  float64
 8   condition         21288 non-null  float64
 9   quality           21288 non-null  float64
 10  ceil_measure      21288 non-null  float64
 11  basement          21288 non-null  float64
 12  yr_built          21288 non-null  float64
 13  yr_renovated      21288 non-null  int64
 14  lat               21288 non-null  float64
 15  long              21288 non-null  float64
 16  living_measure15  21288 non-null  float64
 17  lot_measure15     21288 non-null  float64
 18  furnished         21288 non-null  float64
 19  total_area        21288 non-null  float64
dtypes: float64(18), int64(2)
memory usage: 3.4 MB
```

Fig 3 – Dataset info

3) Exploratory Data Analysis

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

Price

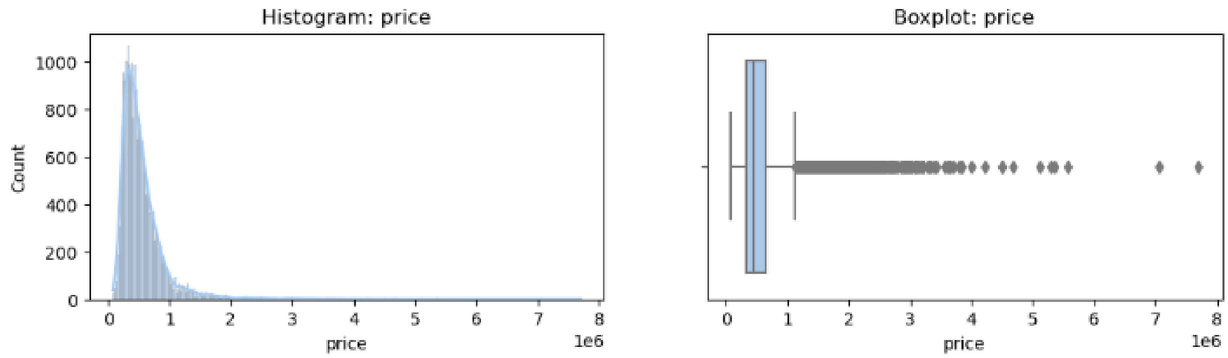

Fig 4 – Histogram and Boxplot of Price

Number of Bedrooms (room_bed)

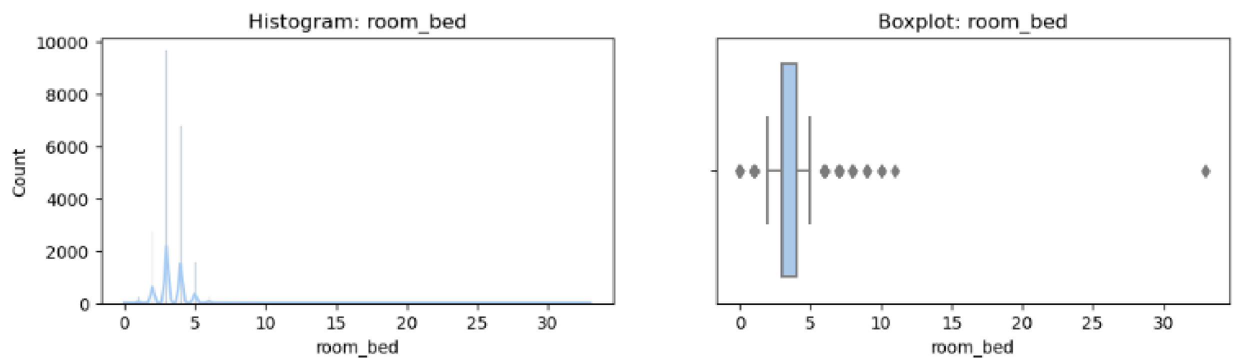

Fig 5 – Histogram and Boxplot of room_bed

Number of bathrooms (room_bath)

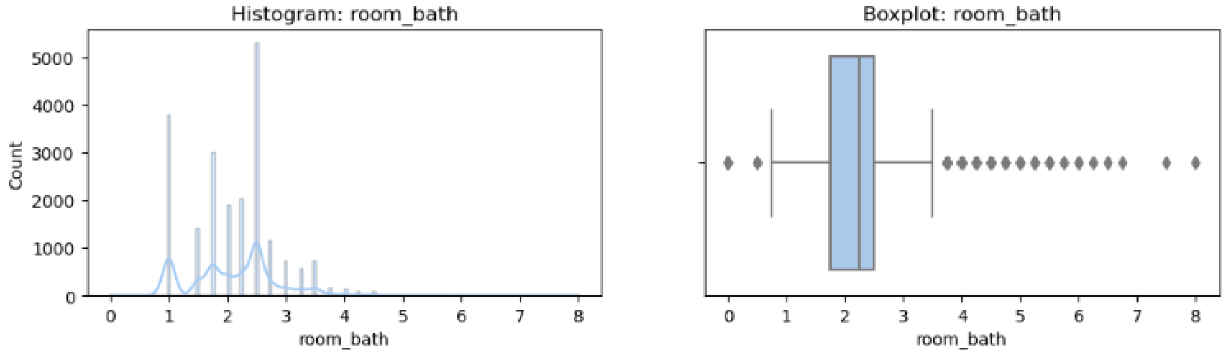

Fig 6 – Histogram and Boxplot of room_bath

living_measure

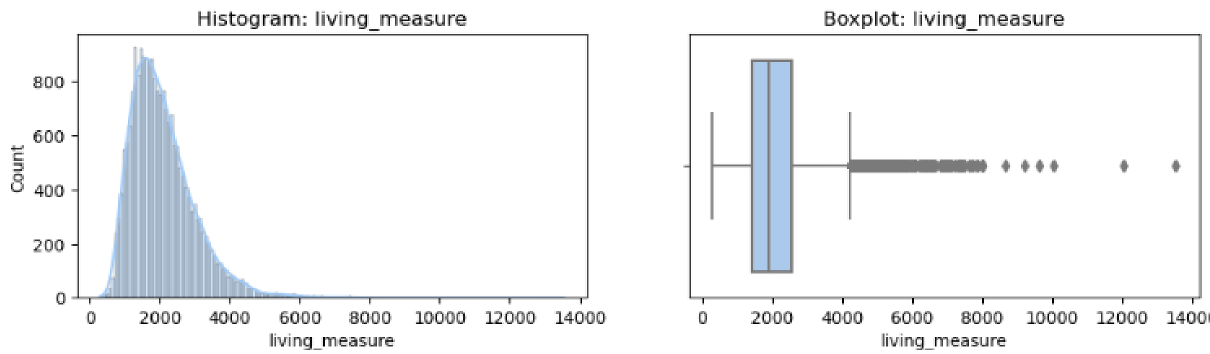

Fig 7 – Histogram and boxplot of living_measure

lot_measure

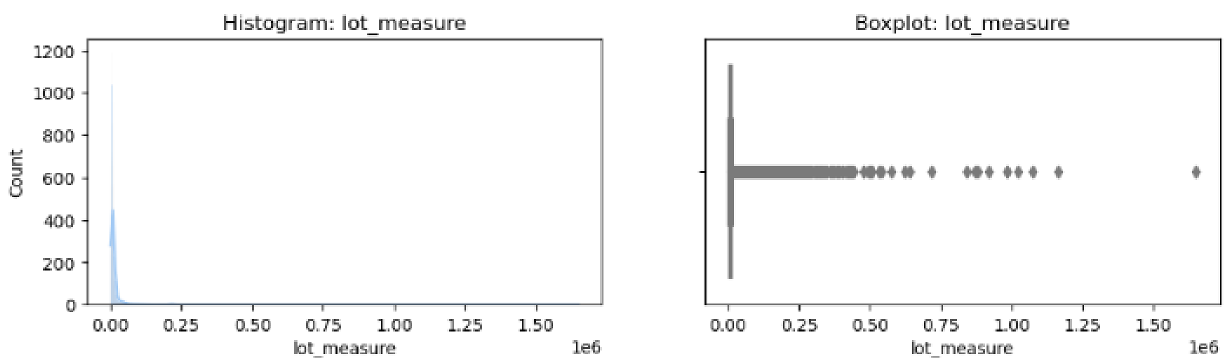

Fig 8 – Histogram and boxplot of lot_measure

Ceil

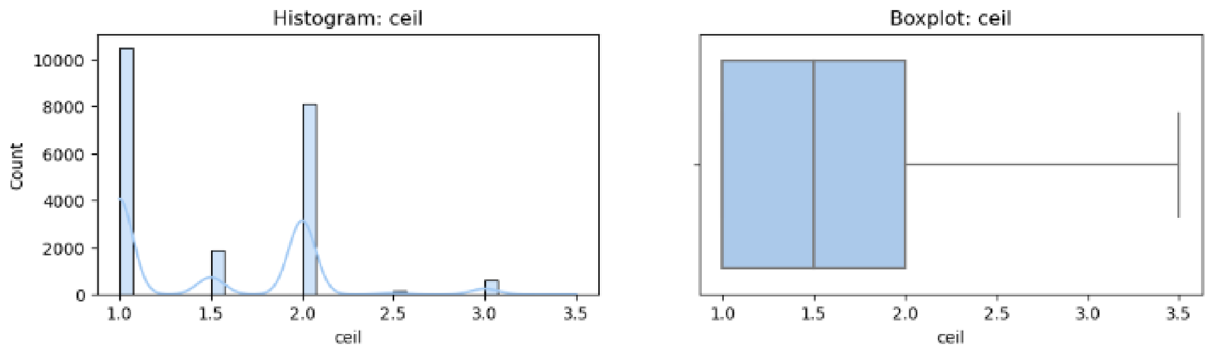

Fig 9 – Histogram and boxplot of ceil

Coast

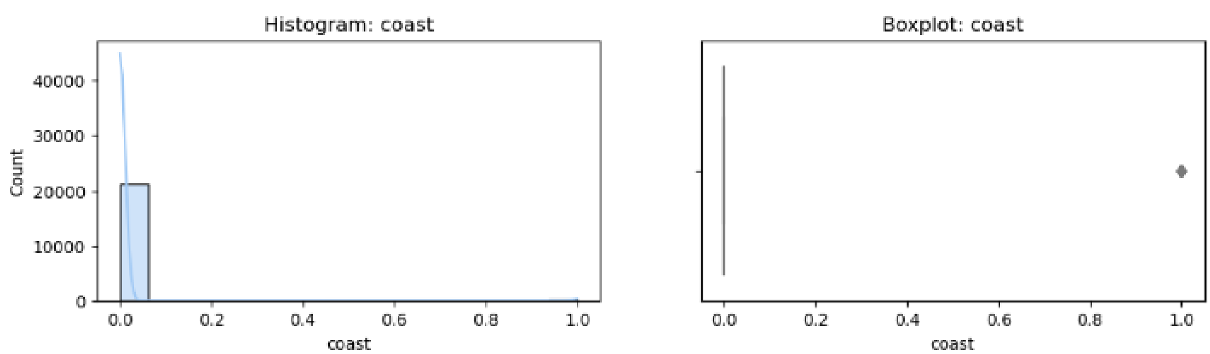

Fig 10 – Histogram and boxplot of coast

Sight

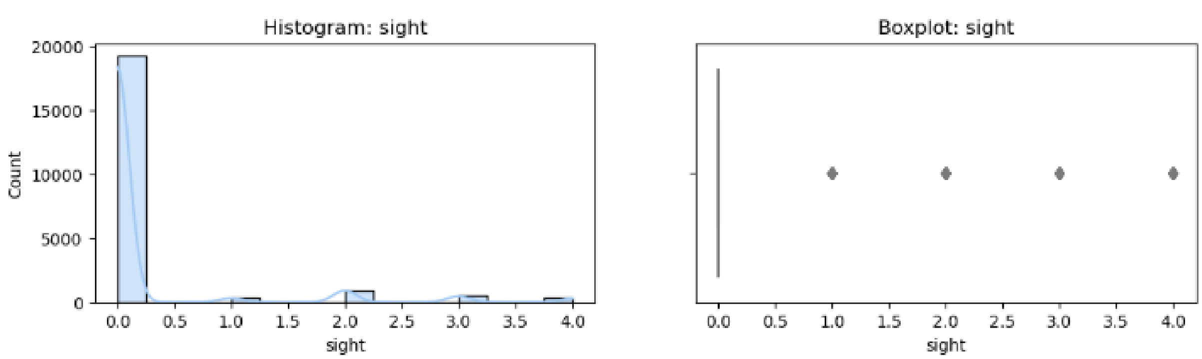

Fig 11 – Histogram and boxplot of sight

Condition

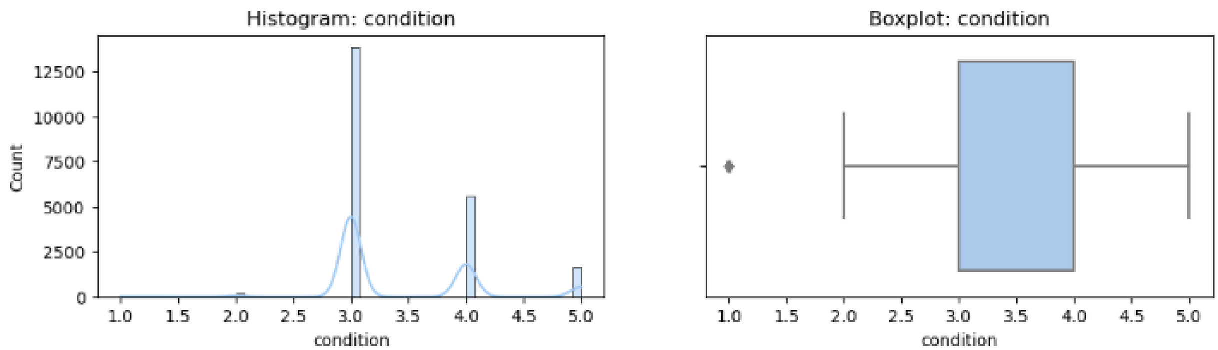

Fig 12 – Histogram and boxplot of Condition

Quality

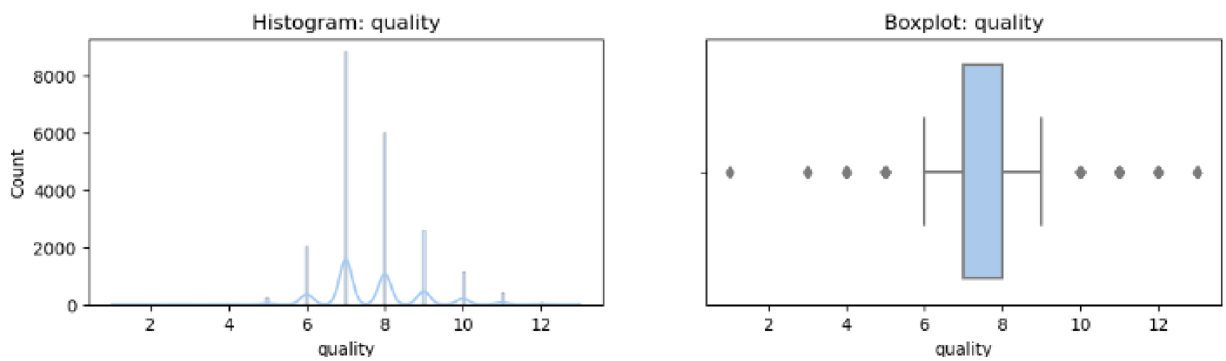

Fig 13 – Histogram and boxplot of quality

ceil_measure

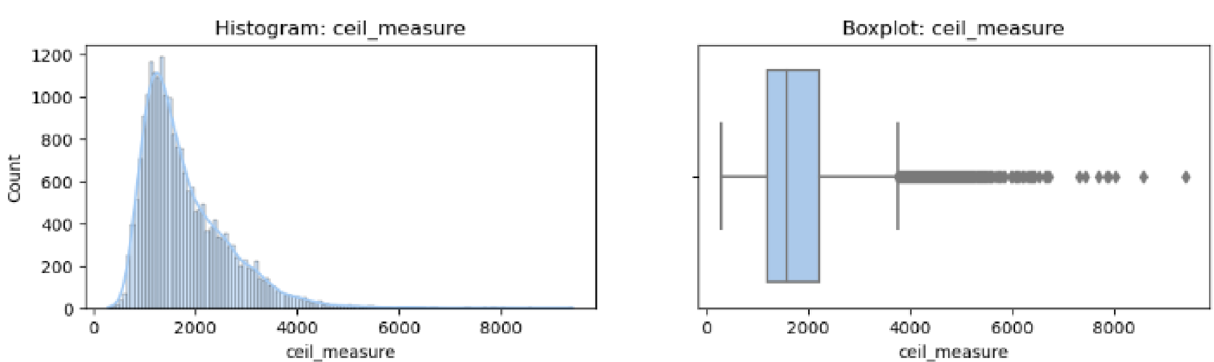

Fig 14 – Histogram and boxplot of ceil_measure

Basement

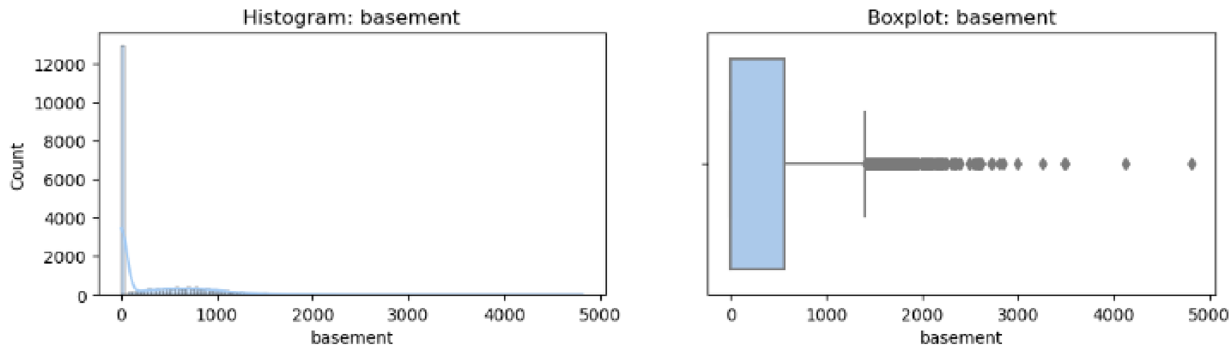

Fig 15 – Histogram and boxplot of basement

yr_built

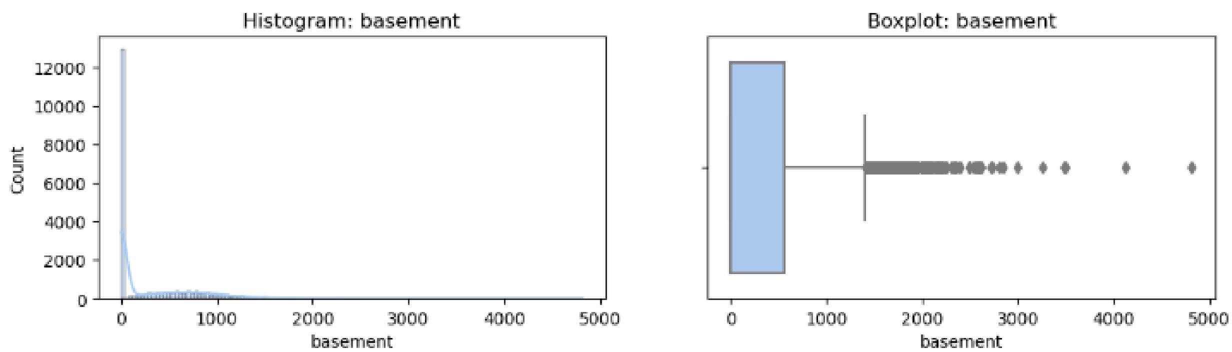

Fig 16 – Histogram and boxplot of yr_built

furnished

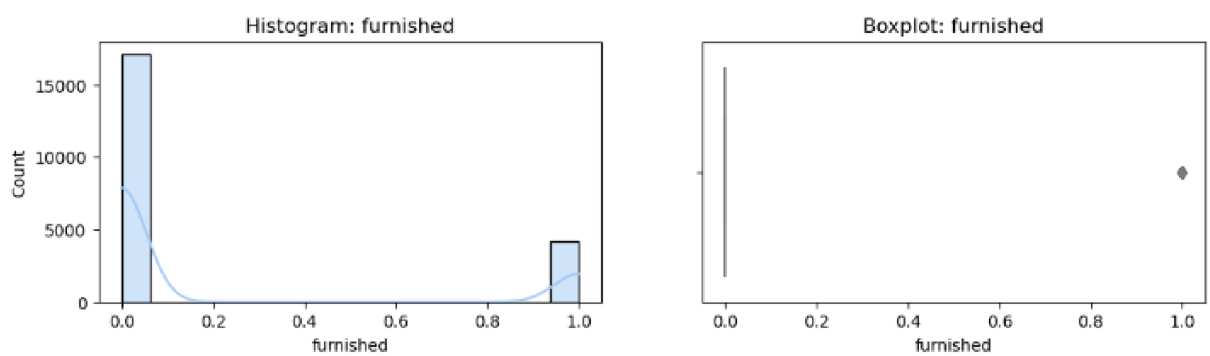

Fig 17 – Histogram and boxplot of furnished

living_measure15

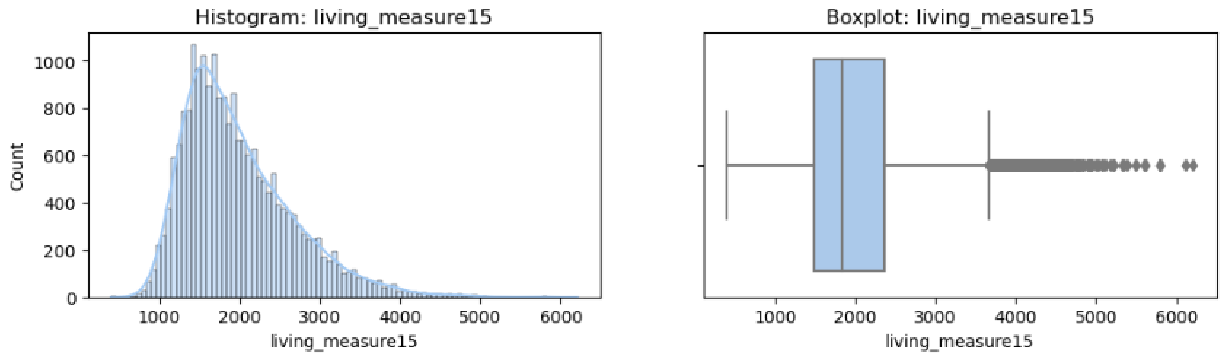

Fig 18 – Histogram and boxplot of living_measure15

lot_measure15

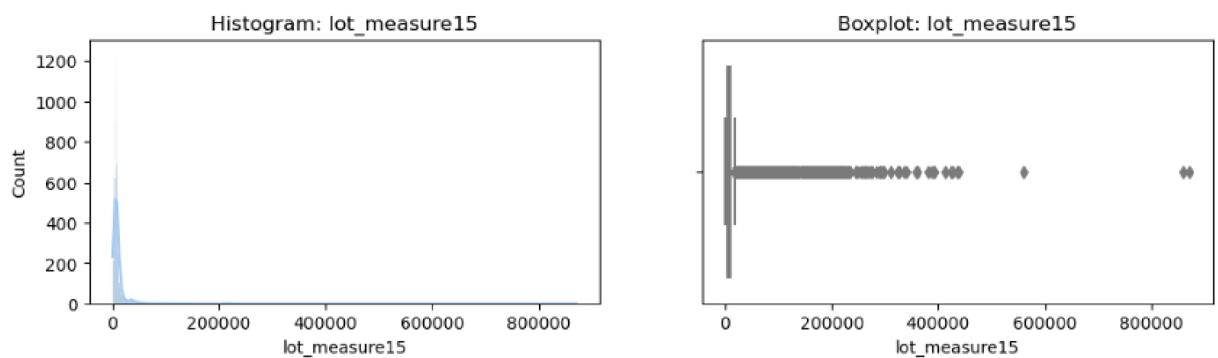

Fig 19 – Histogram and boxplot of lot_measure15

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

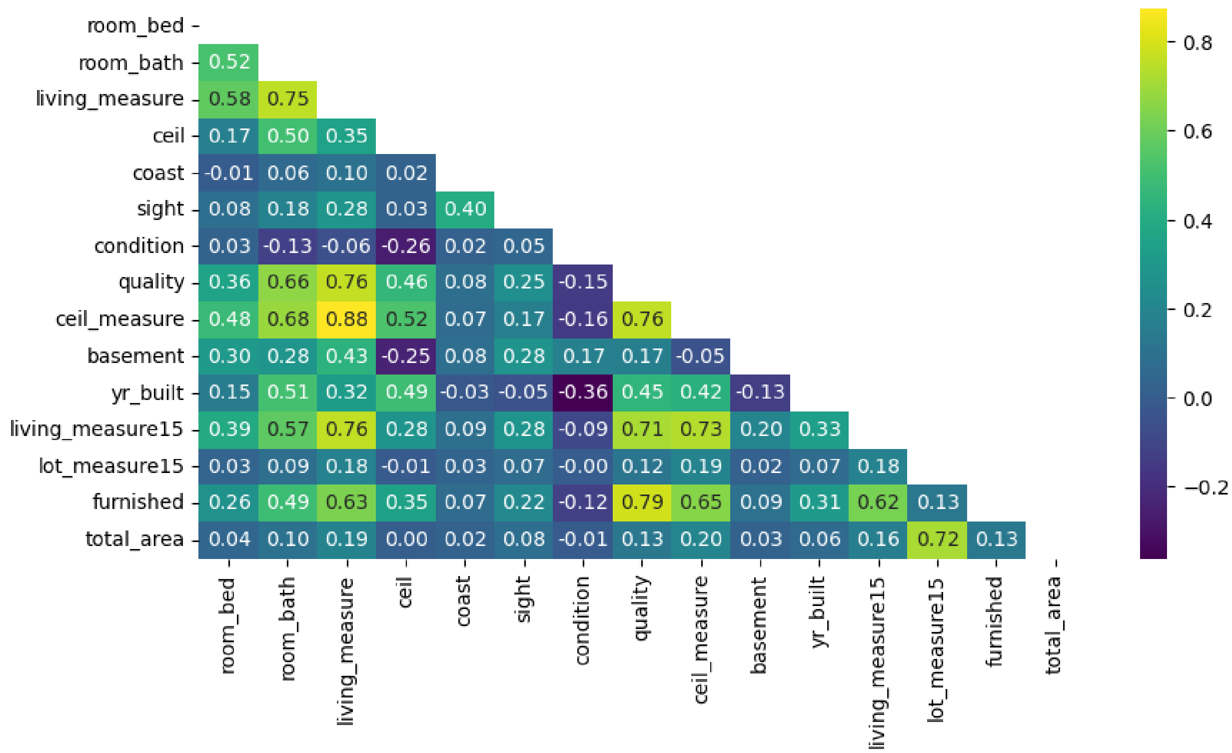

Fig 20 – Correlation

Here, we are taking price as the dependent variable.

Correlation of other variables with price is as follows:

price	1.0
living_measure	0.7
quality	0.7
ceil_measure	0.6
furnished	0.6
room_bath	0.5
sight	0.4
basement	0.3
room_bed	0.3
coast	0.3
ceil	0.3
total_area	0.1
lot_measure15	0.1
yr_built	0.1

condition 0.0

According to this, living measure, quality, ceil measure, furnished, room bath has highest correlation. Therefore, we will be comparing those in our bivariate analysis.

Price vs Living Measure

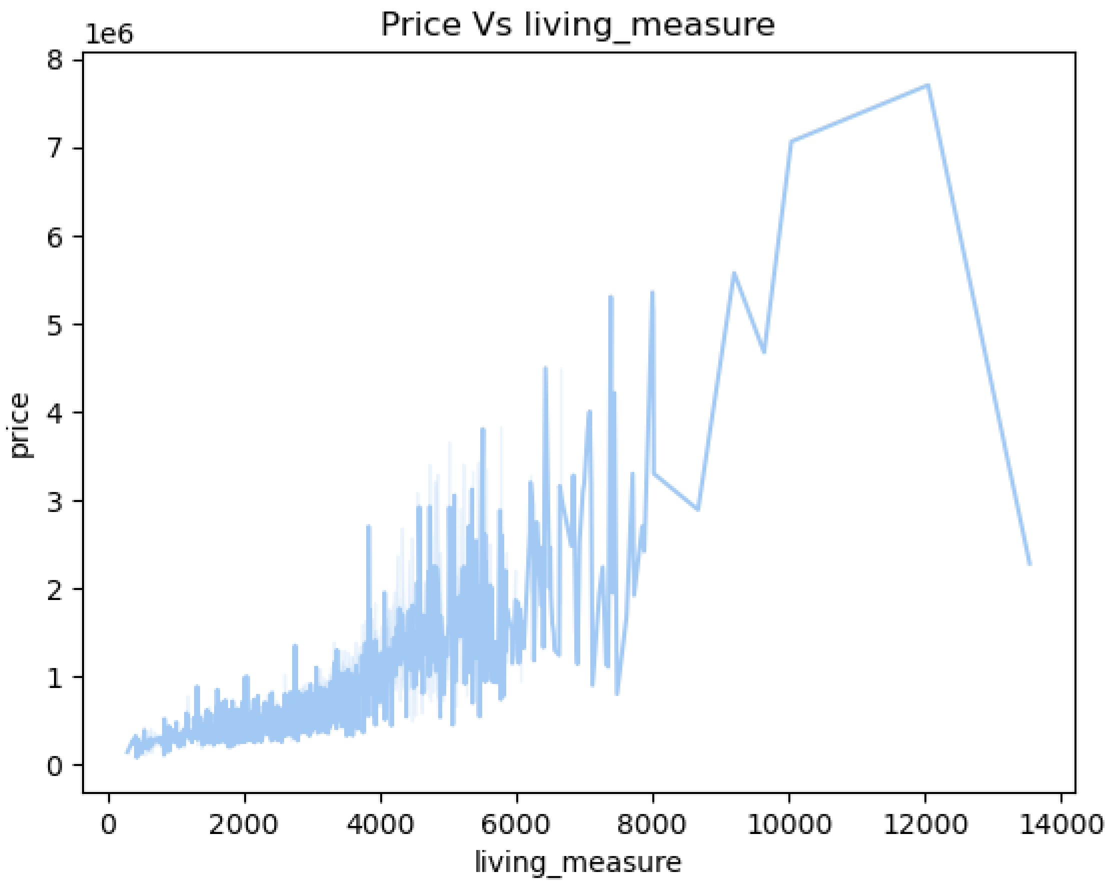

Fig 21 – Price x Living Measure

- Price increases as the square footage of the house increases.
- The price peaked at 12,000 sq. feet.

Price vs Quality

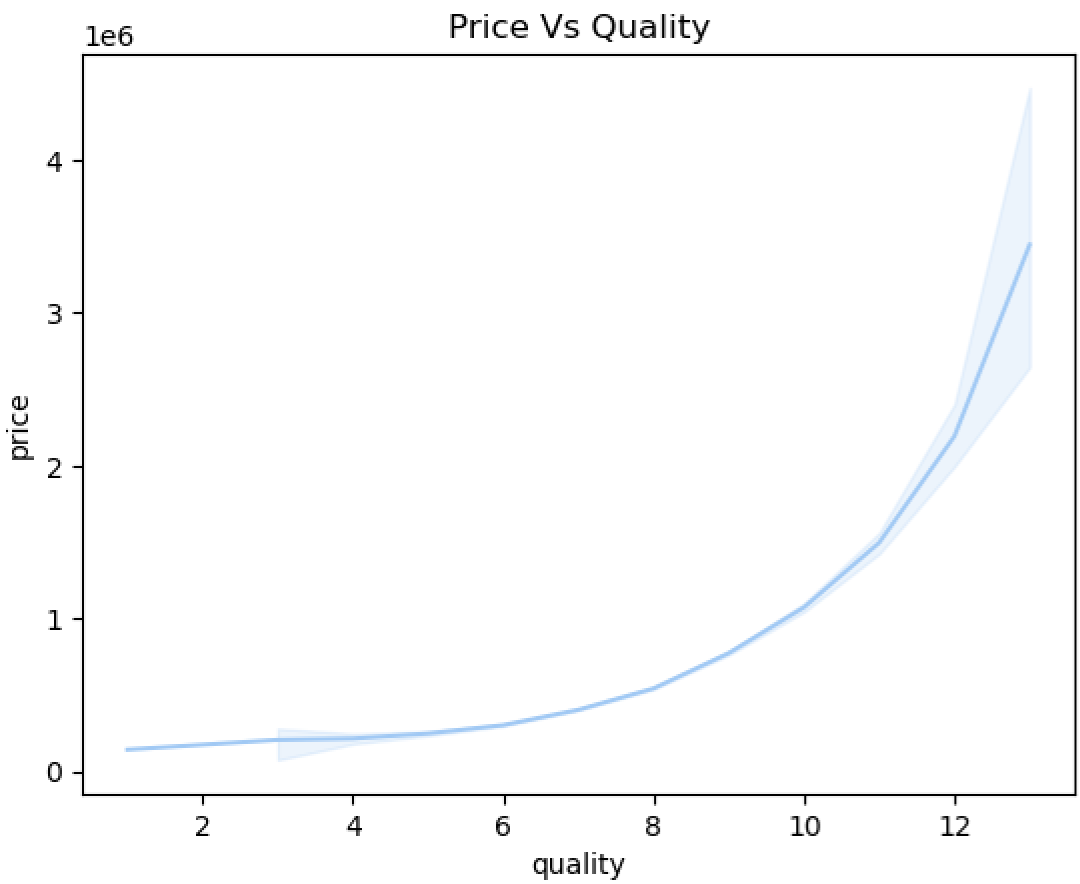

Fig 22 – Price x Quality

- Price and quality are directly proportional to each other.
- The price increased as the quality increased.
- Quality is the score given to the house

Price vs Furnished

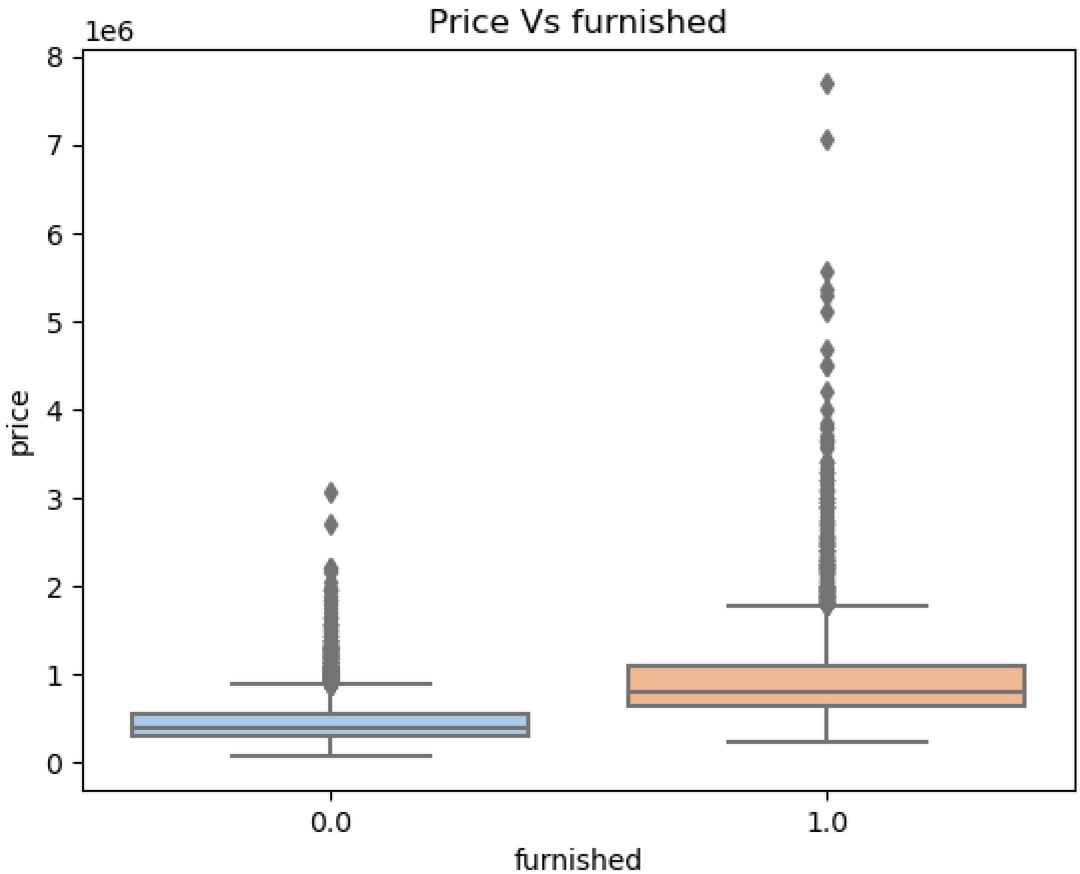

Fig 23 – Price x Furnished

- Furnished houses tend to have higher prices and more variability in prices compared to unfurnished houses.
- Outliers in furnished houses indicate the presence of some extremely high-priced properties.
- This visualization helps in understanding the impact of furnishing on house prices, showing that furnished houses generally command higher prices.

Price vs Room Bath

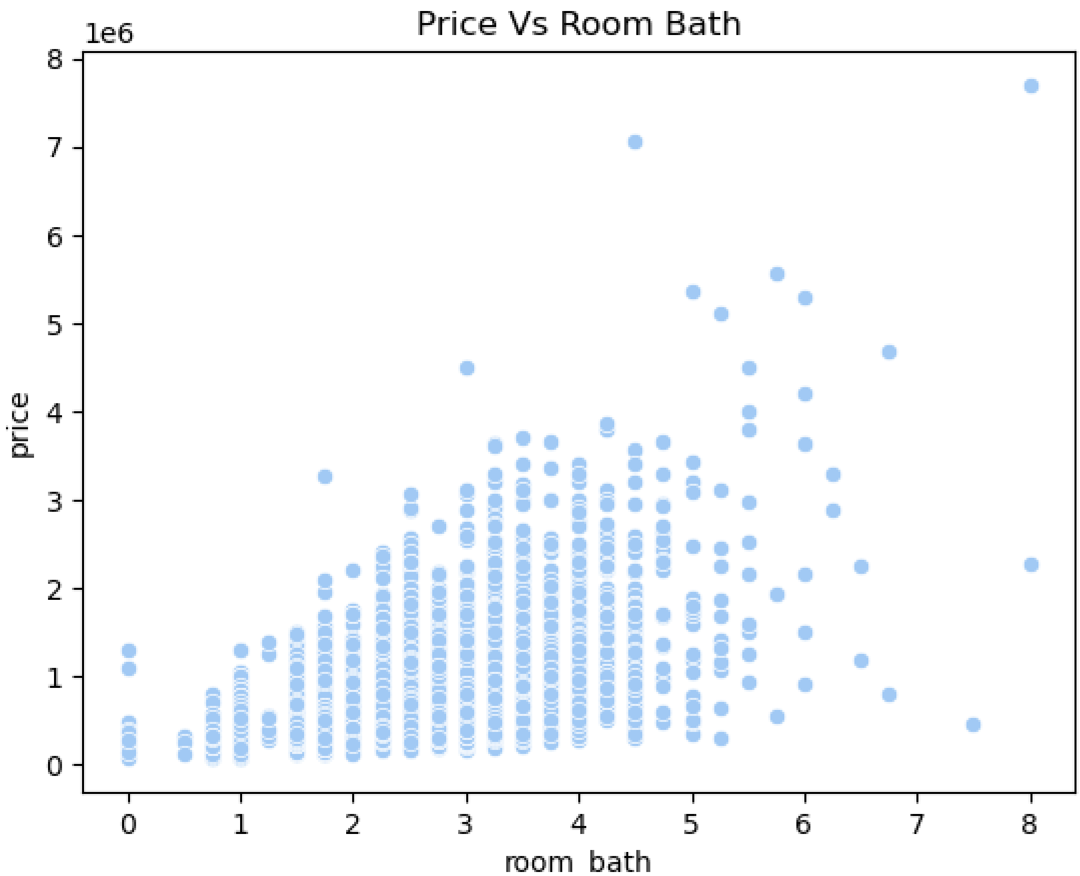

Fig 24 – Price x Room Bath

- **Positive Correlation:** There is a general positive correlation between the number of bathrooms and house prices.
- **Price Clusters:** Lower-priced houses tend to have fewer bathrooms, while higher-priced houses tend to have more bathrooms.
- **Outliers and Variance:** High-priced outliers are more frequent in houses with more bathrooms, indicating that luxury properties often have more bathrooms.
- **Saturation Point:** The price increase becomes less significant after about 5-6 bathrooms, indicating a potential saturation point in the market.

This visualization helps in understanding how the number of bathrooms affects the house price, and having more bathrooms generally adds value to a property, though with diminishing returns at the higher end.

Price vs Ceil Measure

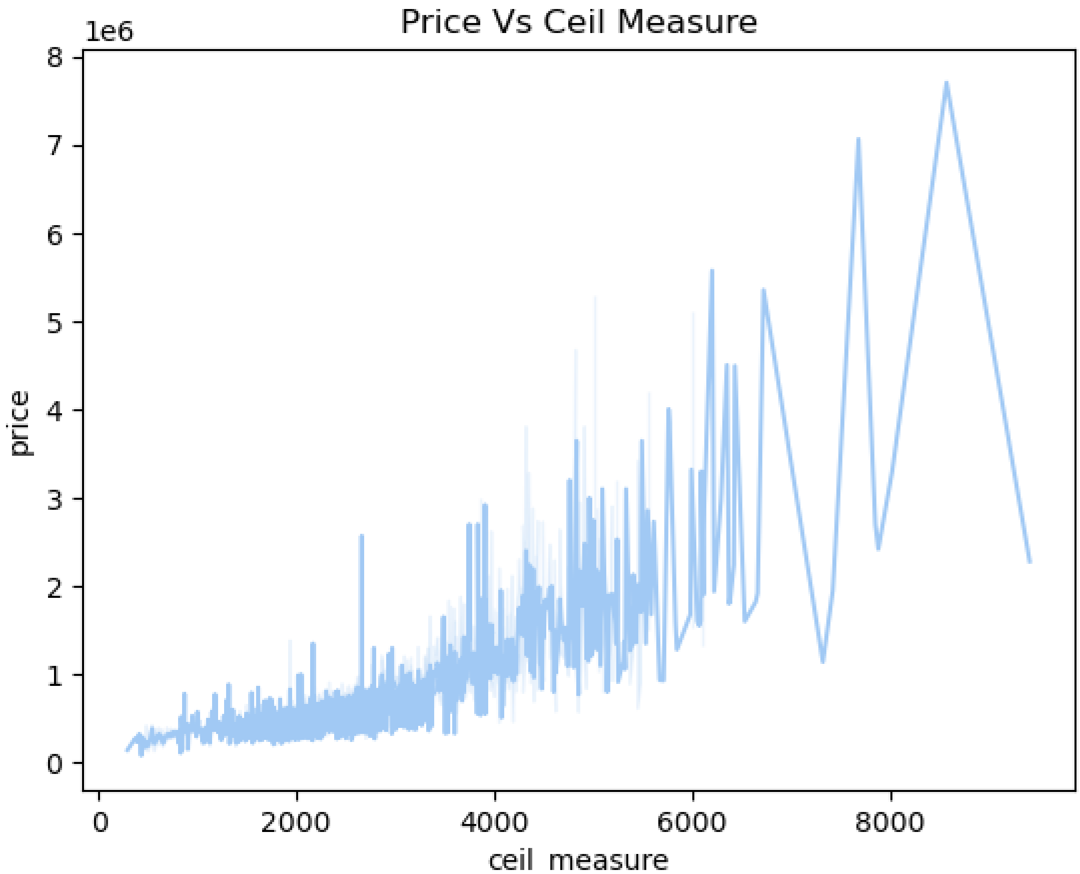

Fig 25 – Price x Ceil Measure

- **Positive Correlation:** There is a general positive correlation between ceiling measure and house prices.
- **Steady Increase:** House prices increase steadily with ceiling measure up to a certain point (around 4000 square feet).
- **High Variability in Larger Sizes:** Beyond 4000 square feet, the prices show greater variability and include some extremely high values.
- **Luxury Properties:** The large fluctuations and spikes in prices for higher ceiling measures indicate the presence of luxury properties.

This visualization helps in understanding how ceiling measure affects house prices, with larger ceiling areas generally corresponding to higher house prices, and significant variability indicating the impact of other factors on high-end properties.

Data cleaning and pre-processing

Removal of unwanted variables

We have removed zipcode, latitude, longitude, cid, dayhours, etc.

Missing value treatment

There were null values and some alphanumerical values. So we replaced them with nan numbers.

Outlier treatment

There are many outliers in the following columns: room_bed, room_bath, living_measure, ceil_measure, basement, living_measure15, ceil_measure15, total_area.

The steps for outlier treatment:

1. Calculate percentiles
2. Calculate Inter-quartile range
3. Determine bounds of outliers
4. Cap outliers

Business insights from EDA

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

Some data is unbalanced due to different skewness.

We can use stratified sampling and feature engineering to combat this. It might even help in making a new feature/variable which will be useful for realtors, sellers and buyers.

In the context of real estate, understanding and addressing data imbalance is crucial for accurate and fair price prediction models. Here's why:

- **Fair Pricing:** Accurate models ensure that houses are priced fairly, benefiting both sellers and buyers.
- **Market Insights:** Helps identify trends and patterns in underrepresented segments (e.g., luxury houses or low-cost housing).
- **Resource Allocation:** Real estate companies can allocate resources better by understanding market demand and supply dynamics.

Some business insights:

1. Accurate pricing strategies can be used by developing this model further.
2. This will help all three parties involved humongously.
3. This will help the builders in creating houses according to their demand in the market.
4. Resource allocation will have a revamp. For eg, builders will be purchasing only necessary wood from the lumberjacks.
5. This will have an environmental impact.

Model Building and Interpretation

Modeling

We will be using 7 different classification models to find out which is the optimum model.

a) **Build Various models**

The following models will be tried out:

- 1. Linear Regression
- 2. Decision tree
- 3. Random forest
- 4. KNN (K nearest neighbour)
- 5. Bagging
- 6. XG Boost
- 7. OLS Statsmodel

Linear Regression

Linear regression is a fundamental statistical and machine learning technique used to model the relationship between one or more independent variables (features) and a dependent variable (target). It seeks to find the best-fitting straight line through the data points to make predictions. In classification tasks, logistic regression is often used instead, which models the probability of the default class.

Decision Tree

Decision tree is a tree-like model where each internal node represents a "decision" based on a feature, each branch represents the outcome of the decision, and each leaf node represents a class label or regression value. Decision trees are easy to interpret and can handle both numerical and categorical data.

Random Forest

Random forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or average prediction (regression) of the individual trees. It improves accuracy and reduces overfitting compared to a single decision tree.

Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly.

KNN – K Nearest Neighbours

KNN is a simple and effective algorithm for classification. It classifies new data points based on the majority class among their k nearest neighbors in the feature space. KNN does not explicitly learn a model but instead stores instances of the training data.

Bagging

Bagging (Bootstrap Aggregating) is an ensemble technique where multiple instances of a base learning algorithm (e.g., decision trees) are trained on different subsets of the training data, sampled with replacement. Bagging helps reduce variance and improve the stability and accuracy of the model.

XG Boost

XGBoost (Extreme Gradient Boosting) is a scalable and efficient implementation of gradient boosting machines. It builds an ensemble of weak learners (typically decision trees) sequentially, where each new tree corrects errors made by the previous ones. XGBoost is known for its performance and often outperforms other gradient boosting implementations.

OLS (Ordinary Least Squares)

OLS is a method for estimating the unknown parameters in a linear regression model. It minimizes the sum of squared residuals between the observed responses in the dataset and the responses predicted by the linear approximation.

- **Decision Tree**, **Random Forest**, **KNN**, **Bagging**, and **XGBoost** are classification algorithms that aim to predict categorical outcomes.
-Each model has its strengths and weaknesses depending on the dataset and problem at hand. Choosing the optimal model often involves evaluating performance metrics such as accuracy, precision, recall, and F1-score, as well as considering interpretability, scalability, and computational efficiency

b)Test your predictive model against the test set using various appropriate performance metrics

We will be using the following performance metrics:

1. RMSE (Root Mean Squared Error)
2. R square
3. Adjusted R square

RMSE (Root Mean Squared Error)

***Definition*:** RMSE is a measure of the differences between predicted values by the model and the observed values. It is the square root of the average of the squared differences between predicted and actual values.

Interpretation:

- RMSE is measured in the same units as the target variable.
- A lower RMSE indicates that the model is better at predicting the target variable because it suggests that the predictions are closer to the actual values.
- RMSE penalizes large errors more severely compared to Mean Absolute Error (MAE).

R square

***Definition*:** R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

Interpretation:

- R-squared ranges from 0 to 1. Higher values indicate that a larger proportion of the variance in the dependent variable is explained by the independent variable(s).
- A value of 1 indicates a perfect fit where all variations in the dependent variable are explained by the model.
- A value of 0 indicates that the model does not explain any variability in the dependent variable.

Adjusted R square

Definition: Adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a regression model. It penalizes the addition of unnecessary predictors that do not improve the model significantly.

Interpretation:

- Adjusted R-squared is always lower than R-squared or equal if all predictors are useful.
- It increases only if the new term improves the model more than would be expected by chance.

c)Interpretation of models

All the 7 classification and regression models were evaluated on the basis of performance metrics.

The RSME, R square and Adjusted R square values of the models (test data) are given below:

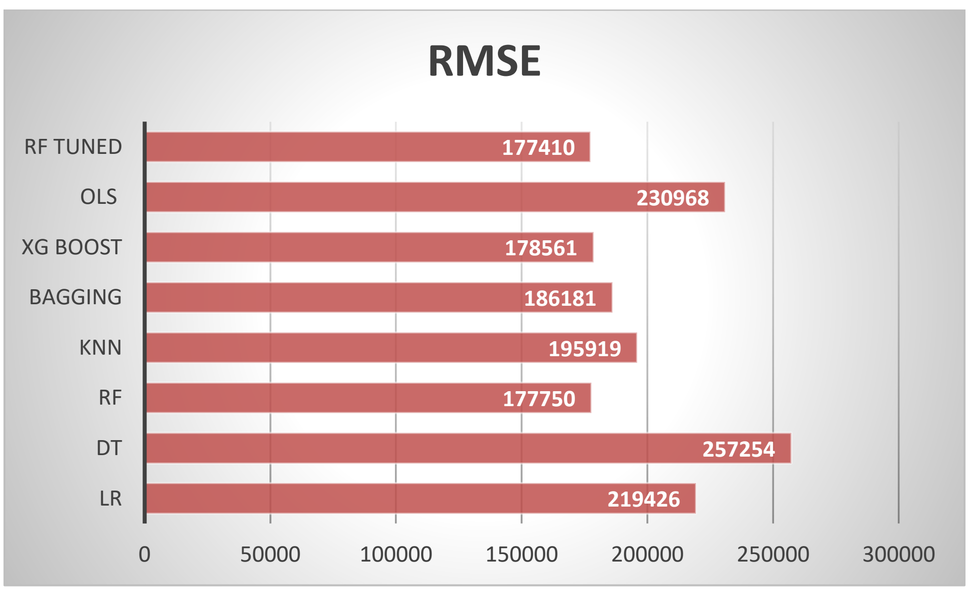

Fig 26 – RSME of models (test)

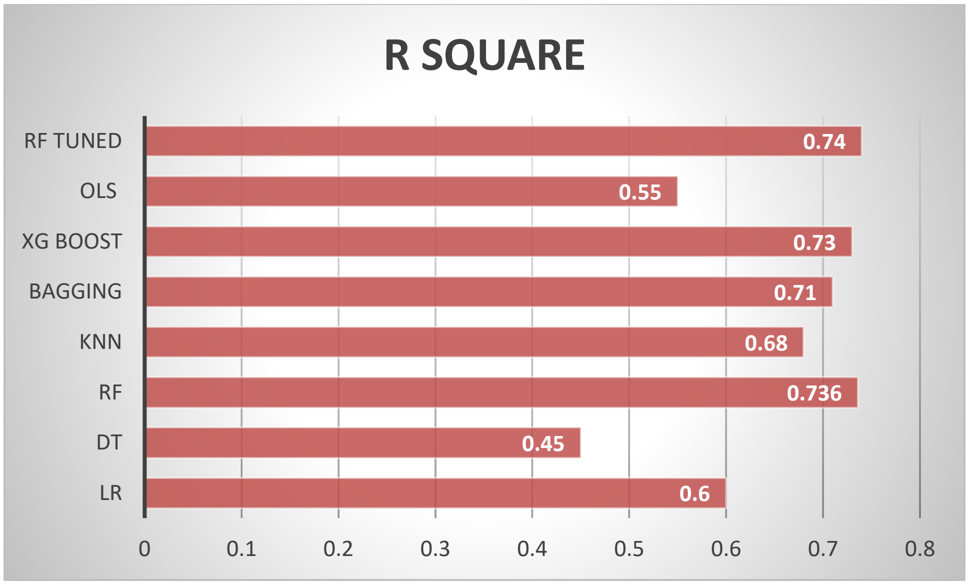

Fig 27 – R square of models(test)

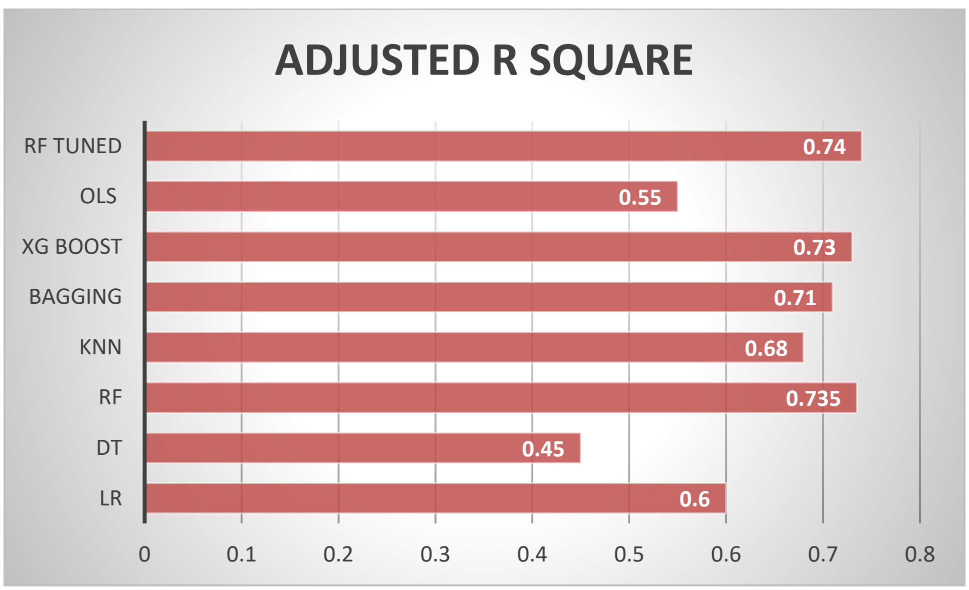

Fig 28 – Adjusted R square of models(test)

Models	RMSE	R SQUARE	ADJUSTED R SQUARE
LR	219426	0.6	0.6
DT	257254	0.45	0.45
RF	177750	0.736	0.735
KNN	195919	0.68	0.68
BAGGING	186181	0.71	0.71
XG BOOST	178561	0.73	0.73
OLS	230968	0.55	0.55
RF TUNED	177410	0.74	0.74

Fig 29 – RSME, R square and Adjusted R square values of models(test)

The RSME, R square and Adjusted R square values of models (train data) are given below.

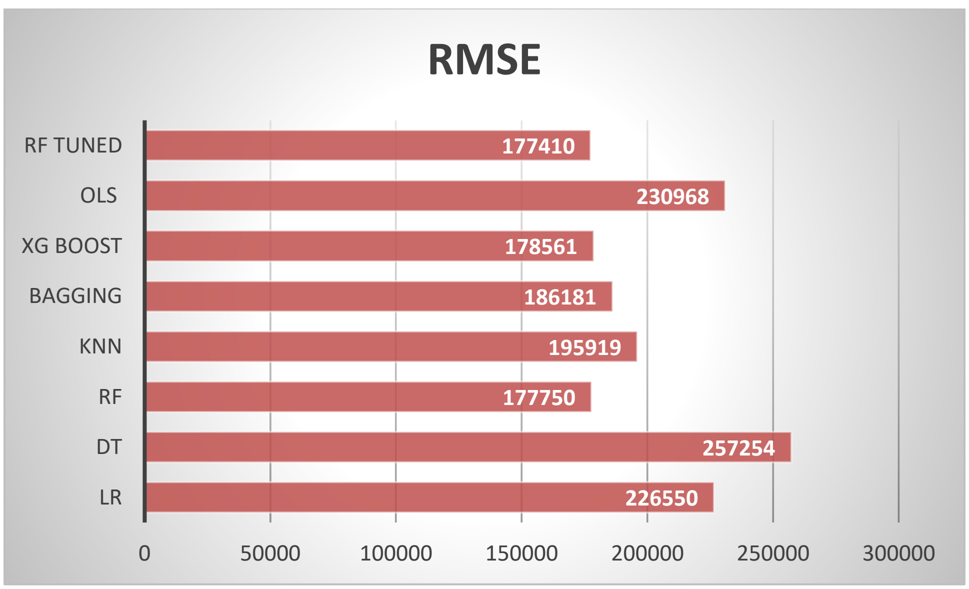

Fig 30 – RMSE of models (train)

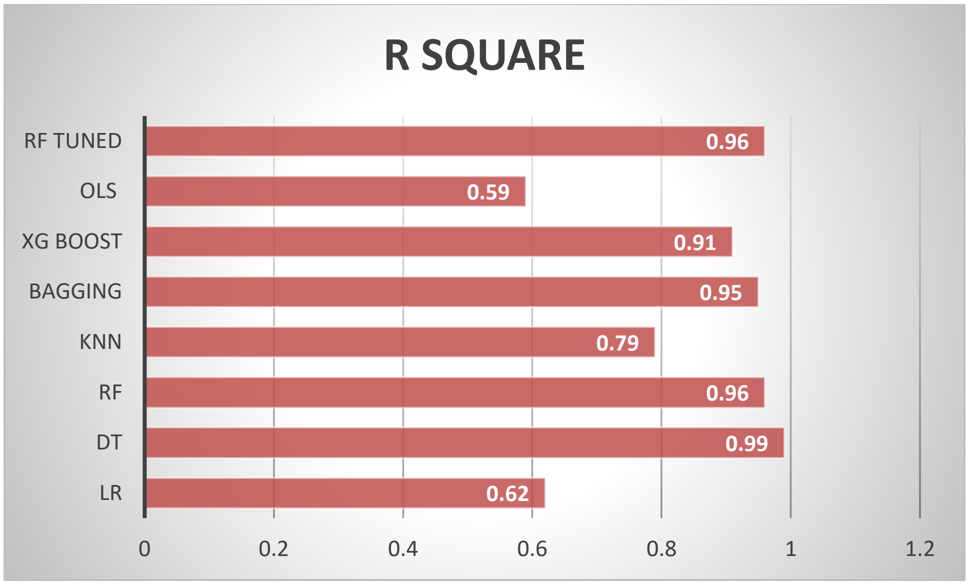

Fig 31 - R square of models (train data)

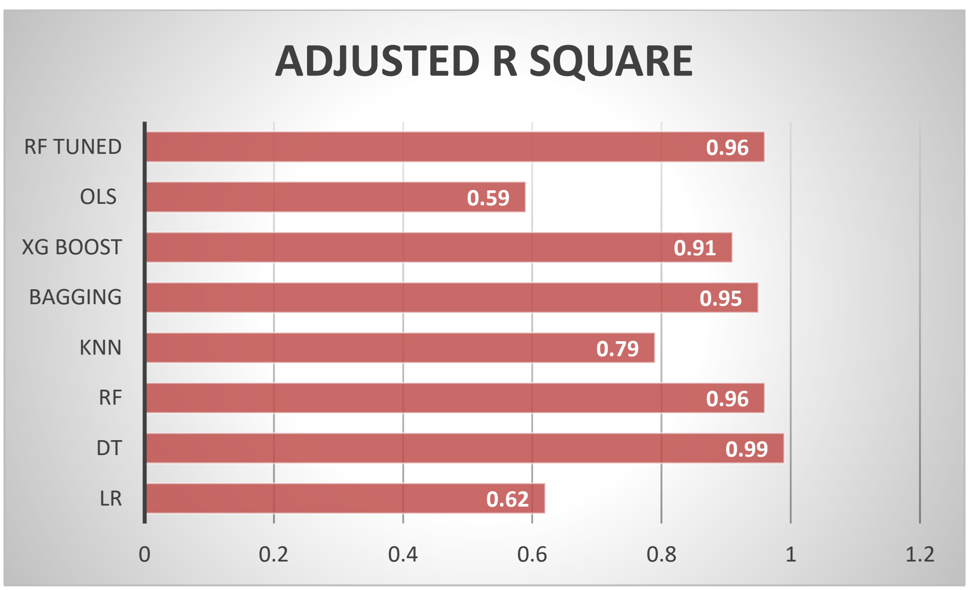

Fig 32 - Adjusted R square of models (train data)

Models	RMSE	R SQUARE	ADJUSTED R SQUARE
LR	226550	0.62	0.62
DT	257254	0.99	0.99
RF	177750	0.96	0.96
KNN	195919	0.79	0.79
BAGGING	186181	0.95	0.95
XG BOOST	178561	0.91	0.91
OLS	230968	0.59	0.59
RF TUNED	177410	0.96	0.96

Fig 33 – RSME, R square and Adjusted R square values of models(train)

To evaluate the models, we need to compare the test data with train data.

	Test			Train		
Models	RMSE	R SQUARE	ADJUSTED R SQUARE	RMSE	R SQUARE	ADJUSTED R SQUARE
LR	219426	0.6	0.6	226550	0.62	0.62
DT	257254	0.45	0.45	257254	0.99	0.99
RF	177750	0.736	0.735	177750	0.96	0.96
KNN	195919	0.68	0.68	195919	0.79	0.79
BAGGING	186181	0.71	0.71	186181	0.95	0.95
XG BOOST	178561	0.73	0.73	178561	0.91	0.91
OLS	230968	0.55	0.55	230968	0.59	0.59
RF TUNED	177410	0.74	0.74	177410	0.96	0.96

Fig 34 – Performance metrics of models (test and train)

Linear Regression

High RMSE indicates higher average error or deviation from the actual values leading to lower models' accuracy. But the model shows low variance as the test and train scores are close.

Decision Tree

High RMSE indicates higher average error or deviation from the actual values leading to lower model's accuracy. R^2 score is also low indicating that model does not fit the data

Random Forest

Lower RMSE and good R^2 score indicates a good accuracy of the model.

KNN (K Nearest Neighbours)

RMSE and R^2 score indicates average model. It has low variation as well.

Bagging

RMSE and R^2 score indicate a relatively average model.

XG Boost

Low RMSE and good R^2 score indicates a good accuracy of the model.

OLS Statsmodel

High RMSE and low R^2 score indicates a higher deviation from actual values and low R^2 score indicates poor model.

Random Forest (Tuned)

Lowest RMSE and good R^2 score indicates a good accuracy of the model.

Model Tuning and Business implication

a) Ensemble modelling, wherever applicable

We have used Random Forest, Bagging and Gradient Boosting (XGBoost).

b) Hyperparameter Tuning

In machine learning models, hyperparameters are parameters that are set before the learning process begins. They control aspects of the learning process and have a significant impact on the performance of the model. Examples of hyperparameters for a Random Forest model include the number of trees (n_estimators), maximum depth of the trees (max_depth), and minimum number of samples required to split an internal node (min_samples_split).

Grid Search Cross-Validation (GridSearchCV)

GridSearchCV is a technique used to tune the hyperparameters of a model by searching over a grid of hyperparameter values and evaluating the model's performance using cross-validation.

Benefits of GridSearchCV

- **Systematic Exploration:** GridSearchCV systematically searches for the best combination of hyperparameters from the specified grid.
- **Cross-Validation:** Uses cross-validation to ensure the evaluation of model performance is robust and not sensitive to how the data is split into training and validation sets.
- **Optimization:** Helps optimize model performance by finding hyperparameters that lead to the best performance metrics on validation data.

Final Interpretation/Recommendations

Interpretation of the most optimum model and its implication in the business

Business Recommendation Insights:
Top important features (from Random Forest (Tuned) model):
- quality
- living_measure
- yr_built
- furnished
- living_measure15
- coast
- lot_measure15
- total_area
- sight
- ceil_measure
- basement
- room_bath
- room_bed
- condition
- ceil

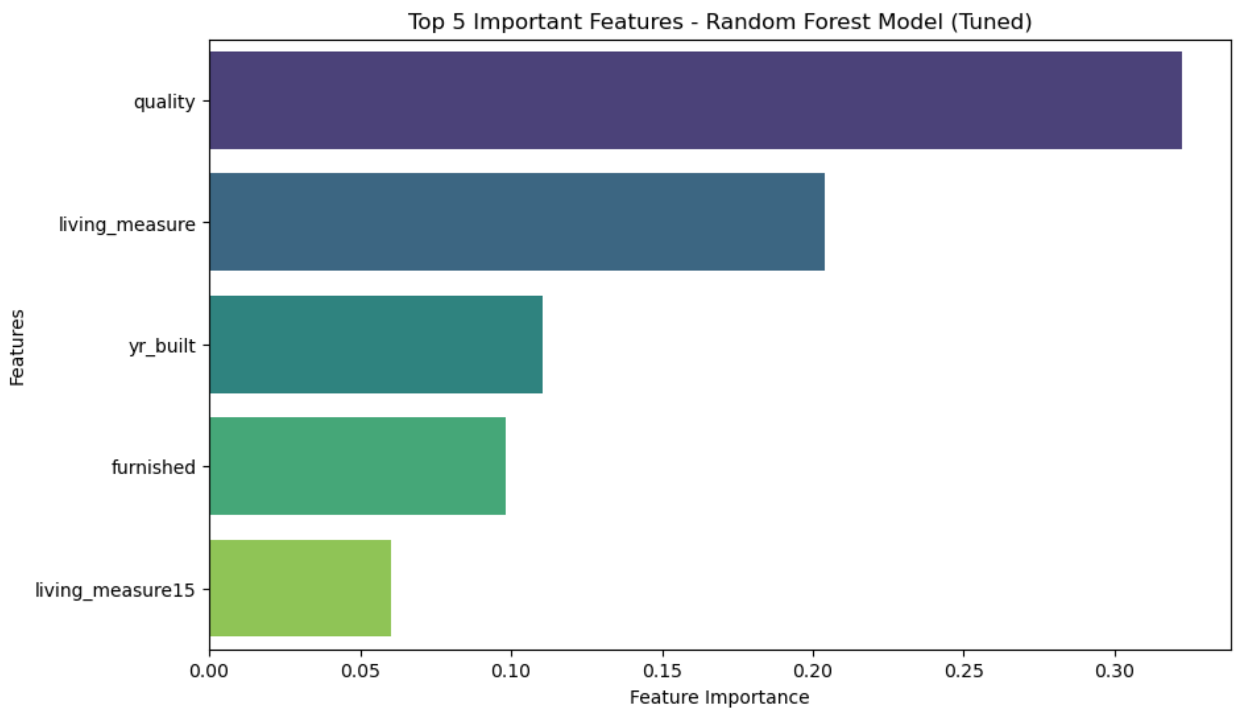

Fig 35 – Feature importance

- The primary factor influencing housing prices is the quality rating of the property. This suggests that buyers place the highest value on the current quality of the house when determining its worth.
- The living measure, or square footage, is a critical determinant of property value. Larger homes typically command higher prices, underscoring the importance of space in property valuation.
- The year a house was built significantly affects its price. Newer constructions are perceived as higher quality, resulting in a higher valuation. Thus, 'yr_built' is one of the top five most important features.
- Whether a house is furnished also impacts its price. A well-furnished home is valued higher, reflecting the added appeal and convenience of ready-to-live-in properties.
- The living room area as of 2015, implying some renovations were made around that time, also plays a significant role in determining house prices.
- Homes located in coastal areas with waterfront views are generally priced higher, as indicated by the 'coast' feature.

Interestingly, the number of bedrooms, bathrooms, and floors has minimal impact on house prices. This suggests that the overall size of the home is more important than its internal layout.

With an accurate and well-fitting prediction model, we can expect price predictions to deviate minimally from actual prices. This can help stakeholders get a reliable approximation of a house's expected price based on different features.

The feature importance derived from the model can guide buyers and sellers to focus on key factors that significantly impact property prices, while also identifying features that have little effect on value. This knowledge can lead to better decision-making in real estate transactions and improve the experience for customers seeking homes within specific price ranges and with features.

This understanding will enable more informed business decisions and enhance the experience for buyers seeking homes in their desired price range with the most valued features.

Recommendations

- **Focus on quality -** Interestingly, the number of bedrooms, bathrooms, and floors has minimal impact on house prices. This suggests that the overall size of the home is more important than its internal layout.

- **Prediction model -** With an accurate and well-fitting prediction model, we can expect price predictions to deviate minimally from actual prices. This can help stakeholders get a reliable approximation of a house's expected price based on different features.
- **Decision-making -** The feature importance derived from the model can guide buyers and sellers to focus on key factors that significantly impact property prices, while also identifying features that have little effect on value. This knowledge can lead to better decision-making in real estate.
- **Accessibility -** This understanding will enable more informed business decisions and enhance the experience for buyers seeking homes in their desired price range with the most valued features.