HOUSE PRICE PREDICTION

CONTENT

- Problem Statement and solution Approach
- Data information
- Data pre-processing for EDA
- Exploratory Data Analysis (EDA)
- Model Building and Performance
- Model Validation and Interpretation
- Model performance comparison
- Business insight and Recommendation

PROBLEM STATEMENT AND SOLUTION APPROACH

- Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad.

- 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 neighbourhood and based on gathered data you will try to assess your house price.

- The objective of this problem statement is to predict the housing prices of a town, or a suburb based on the features of the locality provided and identify the most important features to consider while predicting the prices.

- We will develop machine learning model on the data obtained, which will accurately predict house prices and recommend the most important features to be considered while buying or selling a house.

- If successful, the model will help buyers to make more informed decision about buying houses by having reliable estimate of prices. This can help buyers to avoid overpayment or missing out on good deals.

- Sellers can also benefit by setting optimal prices for their houses. Knowing the key factors that influence house prices will help them to competitively position their homes in the market.

DATA INFORMATION

The dataset consists of 21613 houses and each house has 23 different features. The target variable is the price.

FEATURES	DESCRIPTIONS
cid	a notation for a house
dayhours	Date house was sold
price	Price is prediction target (in $)
room_bed	Number of Bedrooms per house
room_bath	Number of bathrooms per bedrooms
living_measure	square footage of the home
lot_measure	square footage of the lot
ceil	Total floors (levels) in house
coast	House which has a view to a waterfront (0 - No, 1 - Yes)
sight	Has been viewed
condition	How good the condition is (Overall out of 5)
quality	grade given to the housing unit, based on grading system
ceil_measure	square footage of house apart from basement
basement	square footage of the basement
yr_built	Built Year
yr_renovated	Year when house was renovated
zipcode	zip code
lat	Latitude coordinate
long	Longitude coordinate
living_measure15	Living room area in 2015 (implies-- some renovations) This might or might not have affected the lot size area
lot_measure15	lotSize area in 2015 (implies-- some renovations)
furnished	Based on the quality of room (0 - No, 1 - Yes)
total_area	Measure of both living and lot

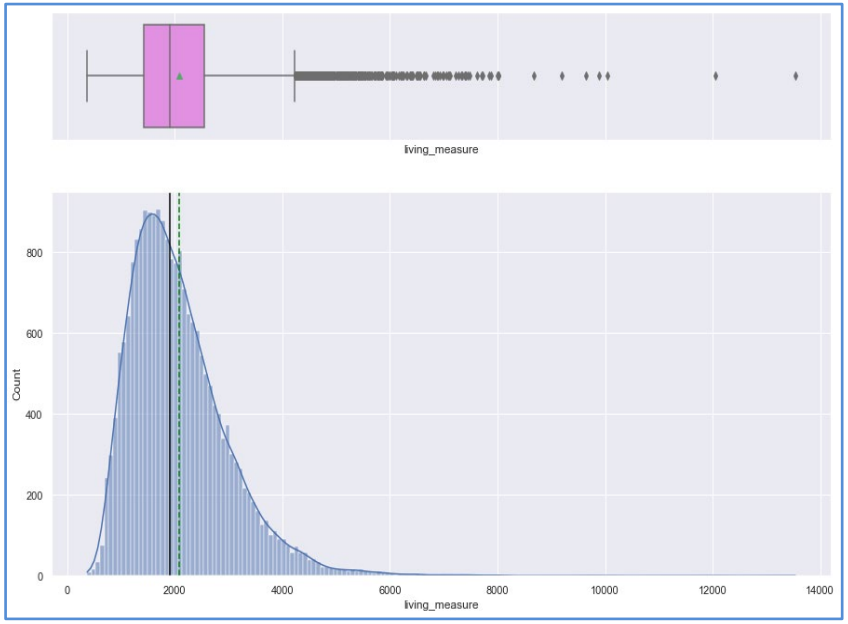
DATA PRE – PROCESSING FOR EDA

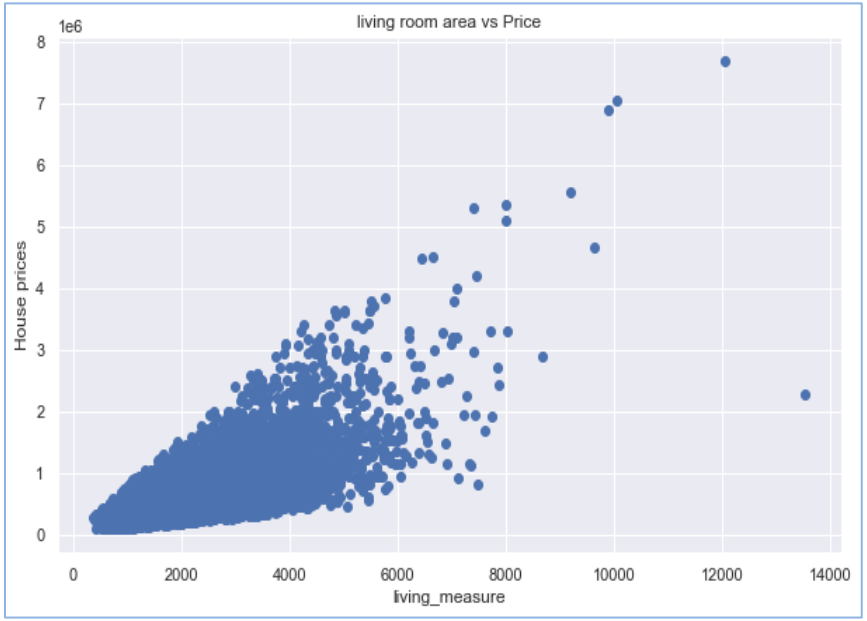
- There were no duplicate values in the dataset.
- Columns; ceil, coast, condition, yr_built, long, and total_area had '$' as part of their values, which distorted their data type. The '$' were removed and replaced with NaN. After this replacement, the affected columns exhibited missing values.
- We imputed all missing values using the median and the mean.
  - Columns with outliers, the impute strategy was median
  - Columns with no outliers, the impute strategy was the mean.
- Houses with no bedrooms and/or no bathrooms were dropped. This reduced the houses in the dataset from 21613 to 21597 for further analysis. 16 houses were dropped. The number of bathrooms were in decimals (float), so they were rounded up for easy use for analysis.
- There were few outliers in the dataset. However, they were not treated because they are all proper values.

#	Column	Non-Null Count	Dtype
0	cid	21613 non-null	int64
1	dayhours	21613 non-null	object
2	price	21613 non-null	int64
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	21541 non-null	float64
8	coast	21582 non-null	float64
9	sight	21556 non-null	float64
10	condition	21528 non-null	float64
11	quality	21612 non-null	float64
12	ceil_measure	21612 non-null	float64
13	basement	21612 non-null	float64
14	yr_built	21598 non-null	float64
15	yr_renovated	21613 non-null	int64
16	zipcode	21613 non-null	int64
17	lat	21613 non-null	float64
18	long	21579 non-null	float64
19	living_measure15	21447 non-null	float64
20	lot_measure15	21584 non-null	float64
21	furnished	21584 non-null	float64
22	total_area	21545 non-null	float64

EXPLORATORY DATA ANALYSIS

EDA – LIVING MEASURE

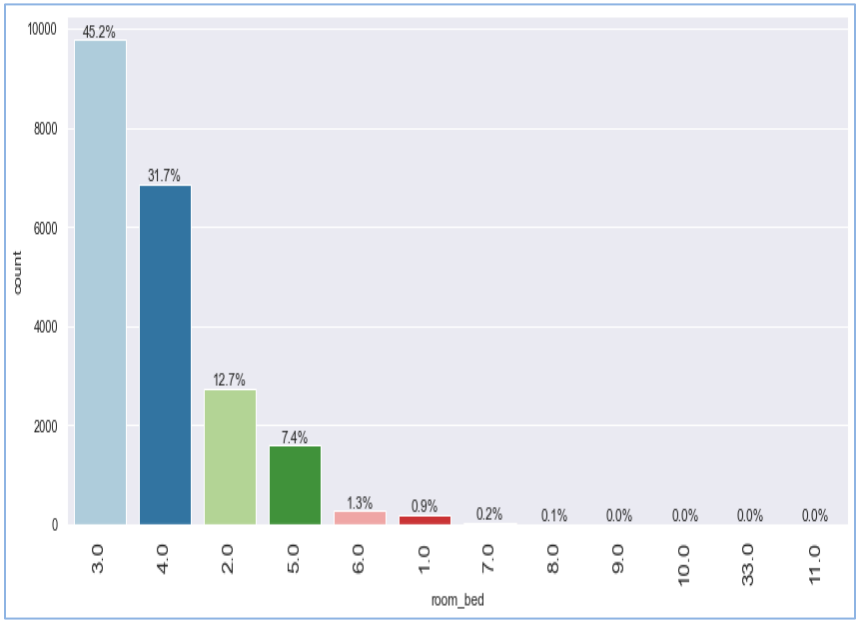


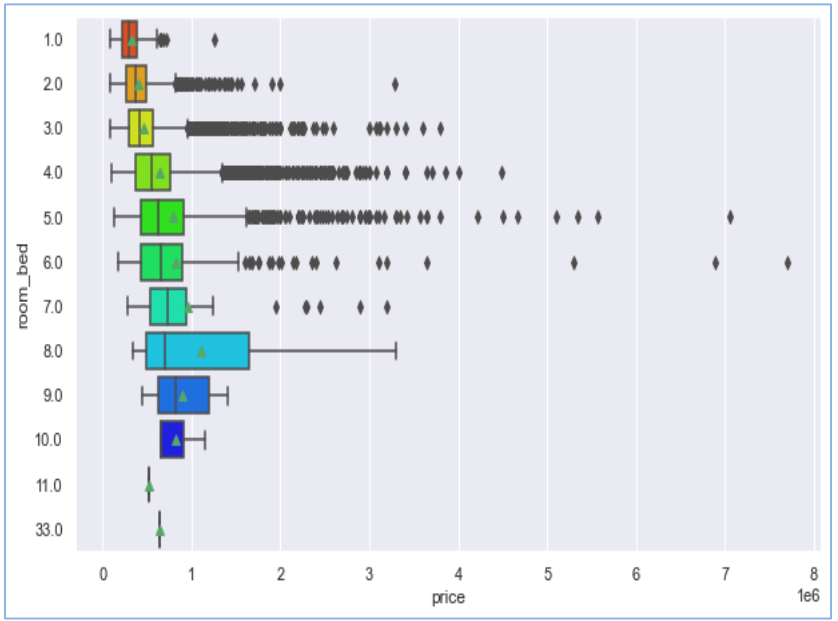


The distribution of the houses living area is skewed to the right with few high outliers. More than 50% houses have lower living area.

There is strong correlation between living measure and price of the houses

EDA – BEDROOM PER HOUSE

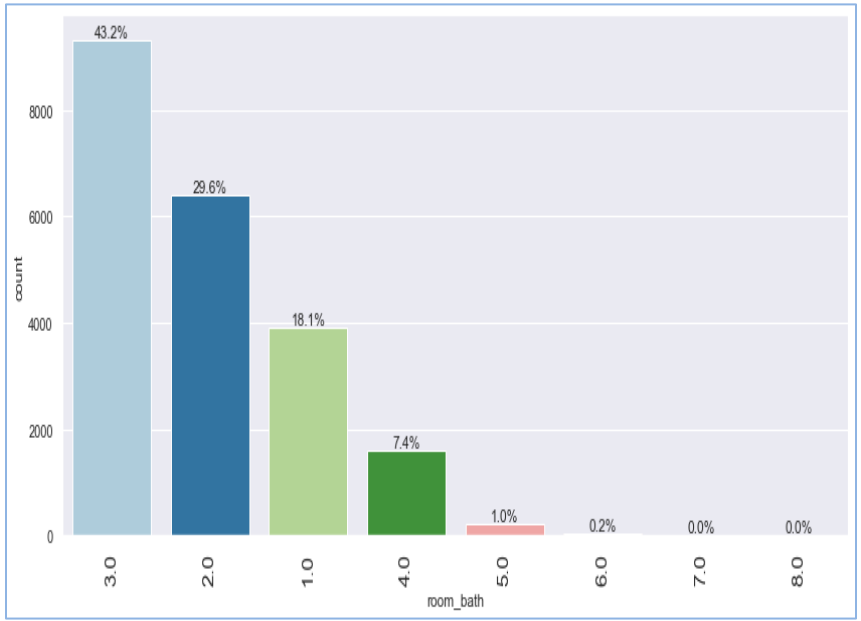


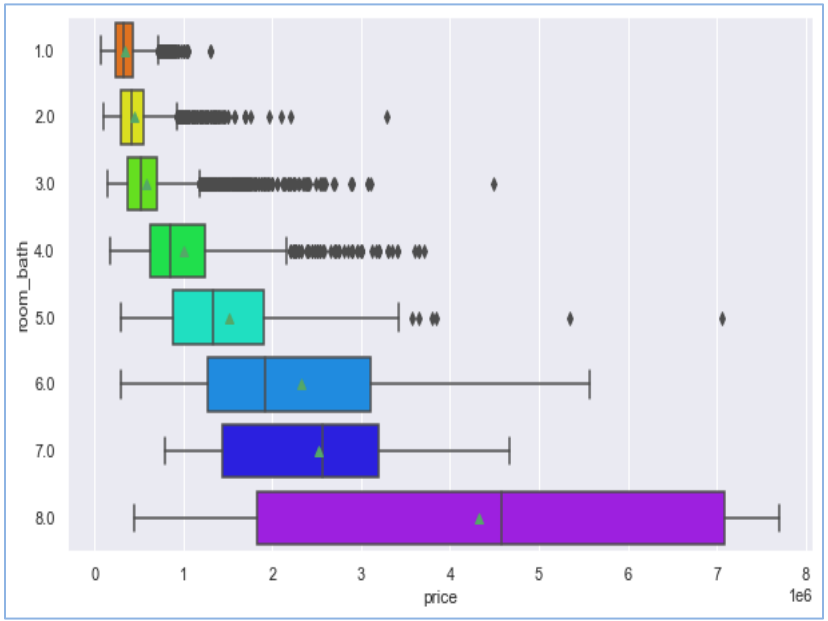


3-bedroom houses have the highest frequency in the dataset follow by 4-bedroom houses. 11, 33, 10, 9, 8, and 7-bedroom houses appeared to be few in the dataset.

Generally, Houses with more bedrooms have higher prices that houses with few bedrooms

EDA – BATH ROOM PER HOUSE

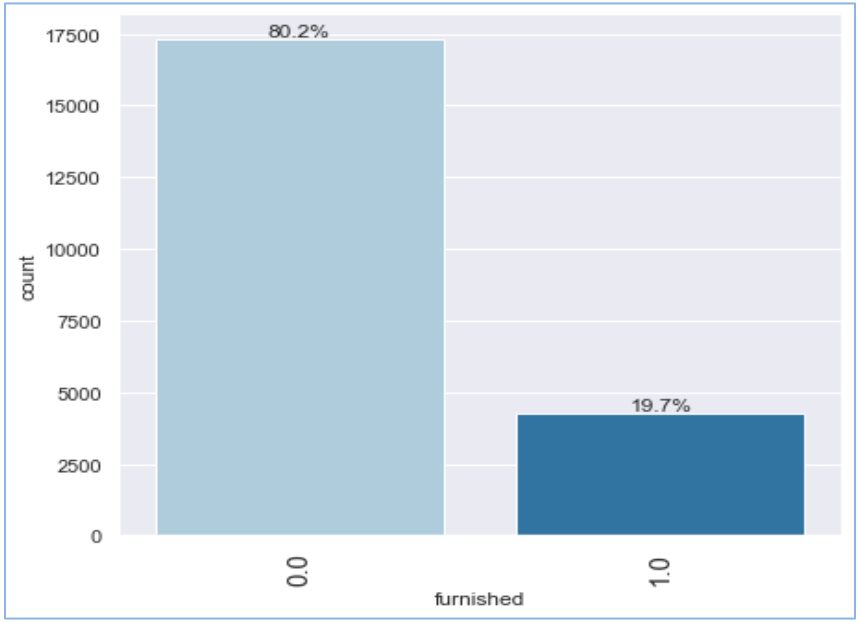


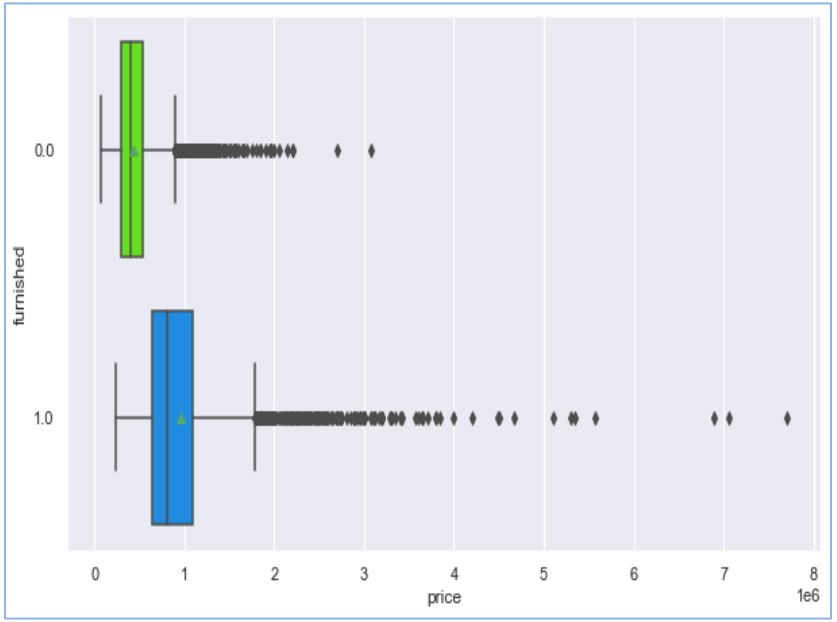


Houses with 3 bathrooms have the highest frequency in the dataset followed by houses with 2 bathrooms. Houses with 8, 7, 6, and 5 bathrooms appear to be few in the dataset.

Generally, houses with more bathrooms appeared to have higher prices than houses with few bathrooms.

EDA – FURNISHED

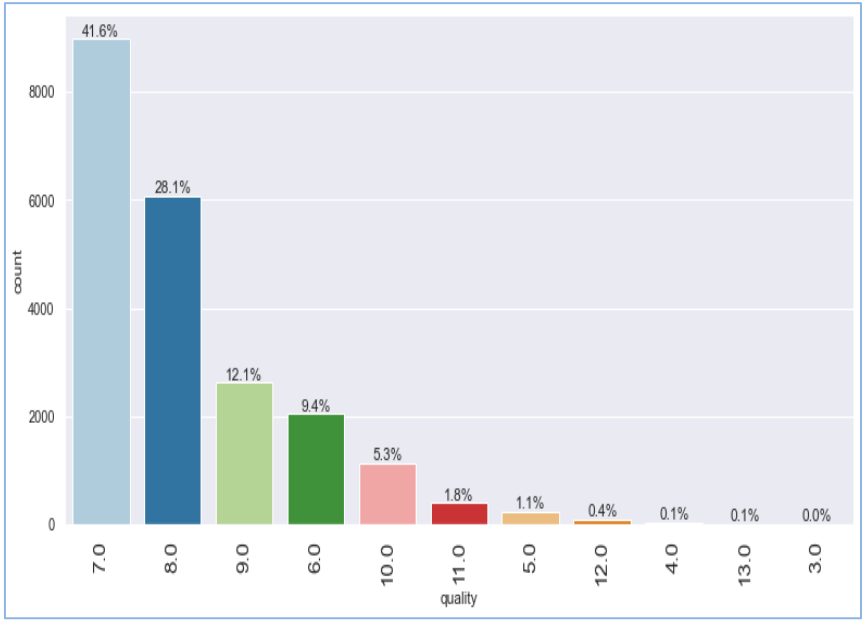


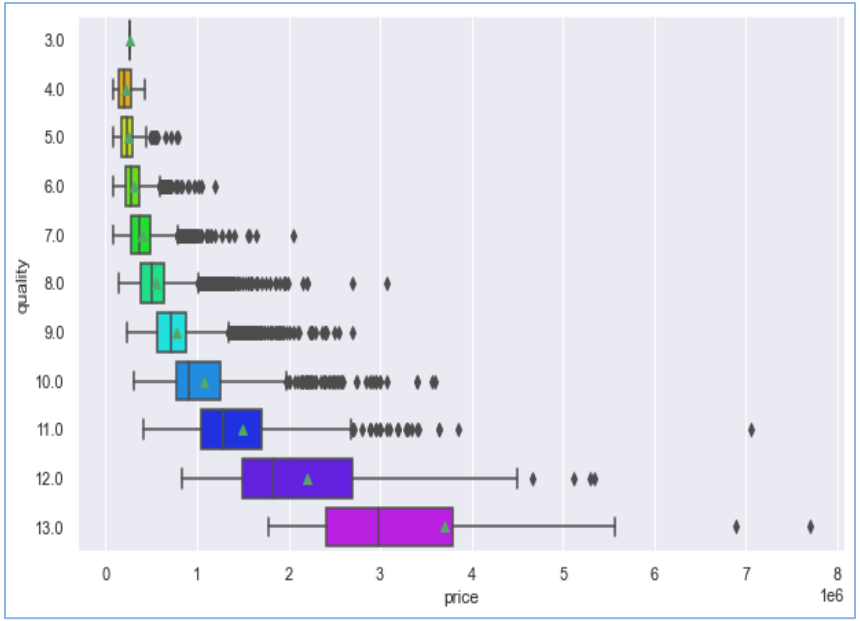


About 80% of the houses are not furnished

Generally, furnished houses have higher prices than unfurnished houses

EDA – QUALITY





About 42% of the house were graded 7 and 28% were graded 8 moderate quality.

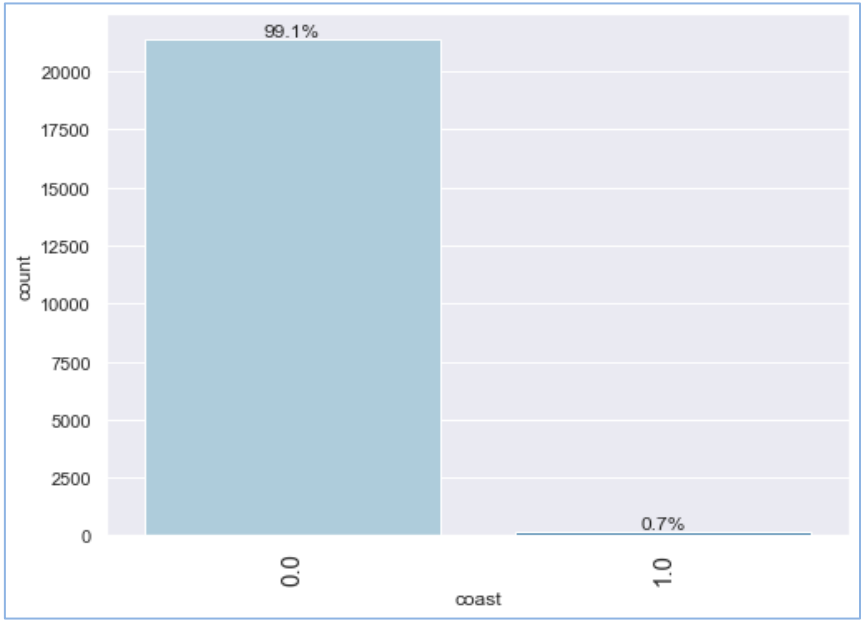
Houses graded 3, 4, 5, and 6 could be suggested a low-quality houses and house graded 9, 10, 11, 12 and 13 could also be suggested as high-quality houses.

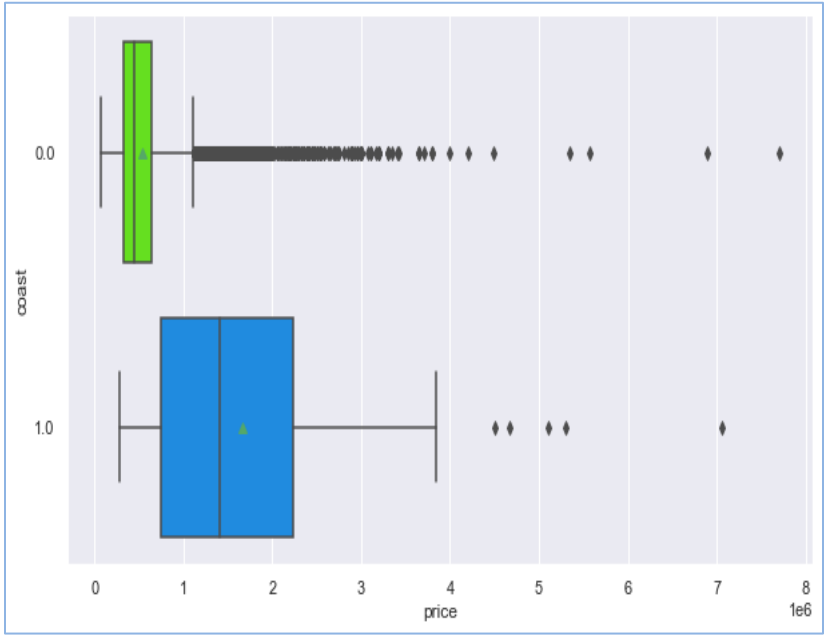
High-quality houses and low-quality houses are very few in the dataset

Generally high-quality houses have higher prices than lower quality houses.

comparatively, moderate quality houses have moderate prices.

EDA – HOUSES WITH WATERFRONT

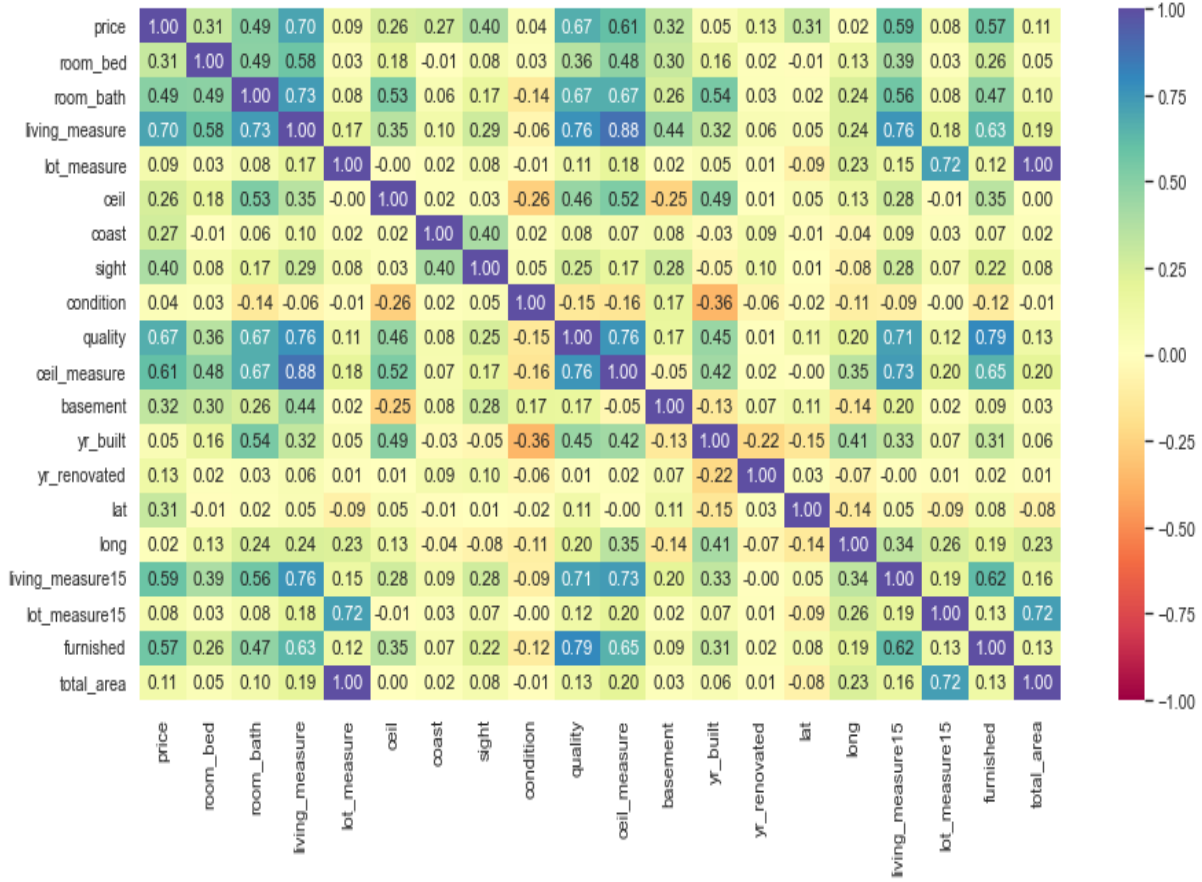




About 99% of the houses have waterfront

Generally, houses with waterfront attract higher prices.

EDA – CORRELATION/HEATMAP



- Numerous variables display minimal correlation or no correlation with each other
- The target variable (price) is correlated to the following variables: living measure, quality, ceil_measure, living measure15 and furnished.
- The highest correlation exist lot_measure and total_area in the plot. Possibly, total_area will be dropped when building our linear regression model as it will affect the performance of the model – collinearity.
- Lat and long variables have no correlation with other different variables.
- Furnished is highly correlated to quality. This indicates that the quality of a house determines its level of furnishing.

MODEL BUIDING AND PERFORMANCE

OVERVIEW – DATA FOR MODEL BUILDING

- The dataset for the models building contains **21957 houses** and 26 features

- Before we proceed to build the models, we will drop 'yr_built', 'cid', 'yr_renovated', 'yr_sold', 'total_area' and 'dayhours'

- Why drop these variables:
  - ✓ cid - Not significant in the model building - they are just codes assigned to the houses.
  - ✓ renovation_status (new variable) has been created from yr_renovated.
  - ✓ yr_sold, yr_built & dayhour - yr_before_sold (new variable) has been created to replace these variables
  - ✓ total_area comprises; living_measure & lot_measure. Once were are adding these two variables to the model, total_area, which is their sum can be dropped.

- We have **20 features** for the models building:
  - ✓ Price – dependent variable.
  - ✓ 19 others – independent variables.

- We will split the dataset into train and test to be able to evaluate the model that we will build on the train data.
  - ✓ Test data will cover 30% of the dataset
  - ✓ Train data will also cover 70%

MODEL OPTIONS

- We want to build a model to predict the price of the house using train dataset and validate their performances on the test data.

- There are several options for model building. Each option comes with its strengths and considerations, depending on the nature of the data and the specific requirements of the problem at hand.

- Here are some common approaches for model building in house price prediction:
  - ✓ Linear Regression Model (OLS)
  - ✓ Decision Tree Regressor Model
  - ✓ Random Forest Regressor Model
  - ✓ XGBoost Regressor Model.

Linear Regression (OLS)

- It is widely used statistical method for predicting continuous outcomes. It is used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting linear equation that describes the relationship between the variables. The term "ordinary least squares (OLS)" refers to the method used to minimize the sum of the squared differences between the observed and predicted values.

- Linear regression makes these assumptions; No multicollinearity in independent, linear relationship between dependent and independent variables , independence of errors, homoscedasticity (constant variance of errors), and normality of errors. Violations of these assumptions can affect the reliability of the model.

MODEL OPTIONS

Decision Tree:

- Decision trees are simple and interpretable models that can handle both numeric and categorical features. They are well suited for capturing non-linear relationships in the data. However, they can be prone to overfitting, so using techniques like pruning can help improve their performance.

Random Forest:

- Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions to make more accurate and robust predictions. It works well with a mixture of feature types and helps reduce overfitting compared to decision trees.

XGBoost:

- XGBoost is an ensemble learning technique based on the gradient boosting framework. It builds a strong predictive model by combining the predictions of multiple weak learners (usually decision trees) sequentially. It can handle missing values, categorical variables, and can handle large datasets efficiently

MODEL EVALUATION METRICS

We will evaluate the performance of the models using the following metrics.

- ✓ RMSE (Root Mean Squared Error)
- ✓ MAE (Mean Absolute Error)
- ✓ R-squared (Coefficient of Determination)
- ✓ Adjusted R-squared
- ✓ MAPE (Mean Absolute Percentage Error)

MODEL VALIDATION AND INTERPRETATION

LINER REGRESSION MODEL (OLS)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	181785.73	106521.90	0.75	0.75	21.60

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	185359.29	108567.58	0.75	0.75	21.63

Inference

- The evaluation metrices for the test and train data are comparable. This indicates that there is no overfitting or underfitting
- The model has performed and generalized well on the test dataset.
- R-squared and adjusted R-squared (after considering the variables without multicollinearity) are 0.75. This is a good score.
- This indicates that 75% of the variability in the price of the houses are explained by independent variables in our model.
- RMSE – On average, the model's predicted prices on the test data differ from the actual prices by 185,359.29
- The MAE value is 108,567.58. This represents the average absolute difference between the model's predicted prices and the actual prices on the test data.
- MAPE - on average, the model's predictions deviate by 21.63% from the actual Prices.

DECISION TREE REGRESSOR MODEL(DEFAULT)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	3041.06	126.72	1.00	1.00	0.04

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	181846.99	102022.04	0.76	0.76	18.62

- Minimal errors on the training set, each price has been predicted correctly.
- The model has performed very well on the training set.

- The decision tree model is overfitting the data as expected and not able to generalize well on the test set.
- We will have to prune the decision tree.

DECISION TREE REGRESSOR MODEL(TUNNED)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	119688.59	63209.73	0.89	0.89	11.60

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	164115.60	88583.43	0.81	0.80	15.88

- The performance of the model after hyperparameter tuning has become generalized.
- The model can explain 80% of the variability in the prices with the independent variables in the test data within 15.88% error margin

RANDOM FOREST REGRESSOR MODEL(DEFAULT)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	48483.71	25932.75	0.98	0.98	4.90

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	131914.36	70290.35	0.87	0.87	12.94

- The metrics are not comparable. Adj. R-squared for Train and Test data are 0.98 and 0.87, respectively.
- The model is overfitting on the train dataset. Let's try to reduce overfitting and improve the performance by hyperparameter tuning.

RANDOM FOREST REGRESSOR MODEL(TUNNED)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	135397.34	81995.00	0.86	0.86	16.12

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	156704.65	88947.29	0.82	0.82	16.62

- After tunning the model, the model performance has generalized.
- We have Adj. R-squared 0.86 for train and 0.82 for test.
- On the average, the model's predicted prices deviated from actual prices by 16.62% on the test and 16.12% on train data

XGBOOST REGRESSOR MODEL(DEFAULT)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	56998.02	40549.90	0.98	0.98	9.03

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	126688.00	70253.64	0.88	0.88	12.90

- The model is overfitting on the train data – Adj. R-square is 0.98 and that of Test data is 0.88
- Let's try to reduce the overfitting and improve the performance of the model

XGBOOST REGRESSOR MODEL(TUNNED)

Training Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	79639.41	54700.92	0.95	0.95	11.47

Test Performance

	RMSE	MAE	R-squared	Adj. R-squared	MAPE
0	124680.15	70297.02	0.89	0.89	13.07

- After tunning the model, the model performance has generalized.
- We have Adj. R-squared 0.95 for train and 0.89 for test.
- There is big difference between RMSE for train data and RMSE for Test data.

MODELS PERFORMANCE COMPARISON

Model/Metrics	RMSE	MAE	R-squared	Adj. R-squared	MAPE
Linear Regression_Train	181,785.73	106,521.90	0.75	0.75	21.60
Linear Regression_Test	185,359.29	108,567.58	0.75	0.75	21.63
Decision Tree_Train	3,041.06	126.72	1.00	1.00	0.04
Decision Tree_Test	181,846.99	102,022.04	0.76	0.76	18.62
Tunned Deccion Tree_Train	119,688.59	63,209.73	0.89	0.89	11.60
Tunned Deccion Tree_Test	164,115.60	88,583.43	0.81	0.80	15.88
Random Forest_Train	48,678.37	25,945.45	0.98	0.98	4.92
Random Forest_Test	132,558.54	70,525.86	0.87	0.87	12.98
Tunned Random Forest_Train	135,397.34	81,995.00	0.86	0.86	16.12
Tunned Random Forest_Test	156,704.65	88,947.29	0.82	0.82	16.62
XGBoost_Train	56,998.02	40,549.90	0.98	0.98	9.03
XGBoost_Test	126,688.00	70,253.64	0.88	0.88	12.90
Tunned XGBoost_Train	79,639.41	54,700.92	0.95	0.95	11.47
Tunned XGBoost_Test	124,680.15	70,297.02	0.89	0.89	13.07

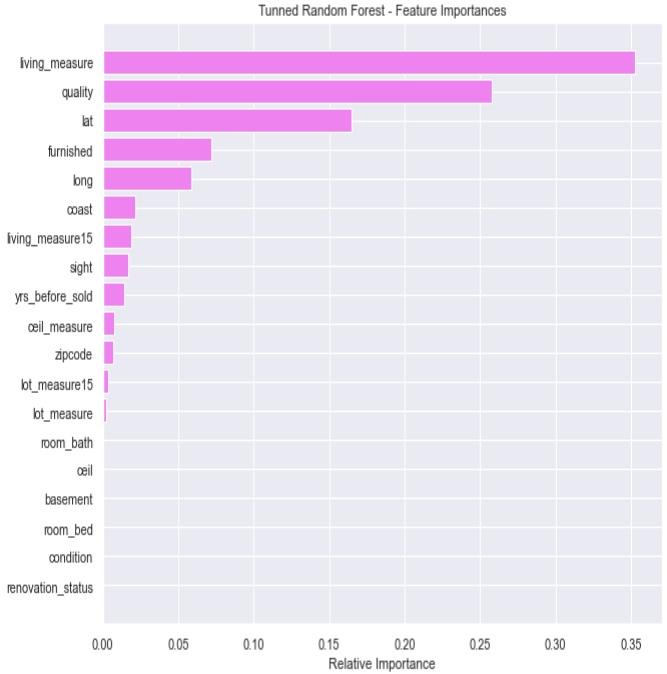
Tunned Random Forest is the best performing model.

THE BEST MODEL PERFORMANCE

Tunned Random Forest Regression Model:

- The evaluation metrices for the test and train data are comparable. This indicates that there is no overfitting or underfitting
- The model has performed and generalized well on the test dataset.
- R-squared and adjusted R-squared for the train data is 0.86 and that of the test is 0.82. This is a good score.
- This indicates that 82% of the variability in the price of the houses are explained by independent variables in our model on the test data.
- RMSE – On average, the model's predicted prices on the test data differ from the actual prices by 156,704.65
- The MAE value is 88,947.29. This represents the average absolute difference between the model's predicted prices and the actual prices on the test data.
- MAPE - on average, the model's predictions deviate by 16.62% from the actual Prices on the test data.

Fig - Important Features in the model



The top five variables in the model that contributed largely to the prediction of the house prices:

- **Living_measure**
- **Quality**
- **Furnished**
- **Lat & long (location of the house)**

BUSINESS INSIGHT AND RECOMMENDATION

living_measure:

- Insight: Living_measure is an important predictor. Large square footage of a home might indicate high price and small square footage will attract low price
- Recommendation: Buyers should look for homes the maximize living space with smart designs and other needs to meet their lifestyles.

Quality & Furnished:

- Insight: Quality and furnished are crucial factors. Furnished is highly correlated to quality. This indicate Quality of a house determines whether the house is furnished or not.
- Recommendation: Buyers should focus on the quality of the property during purchase. A well-built and maintained house often correlates with being furnished. Sellers can also invest in properties that are well-maintained and of high-quality to attract high prices and potential buyers.

Lat & long (location of the house):

- Insight: Location also plays a crucial in the house pricing. Good or advantageous location might attract high price. Proximity to good schools, parks, public transportation, and other excellent amenities can influence the price of a house.
- Investing in houses or buying houses with a strategic and advantageous location can offer both a comfortable living environment and potential financial gains over time.

- Additional information of the houses can be collected to gain better insights. Information such as: location crime rates, school quality, proximity to public transportation/train station, house in gated community, presence of a security system, etc. These may appeal to certain buyers and can significantly influence house prices.

- Regular Model Updates: Continuously update and refine the house price prediction model using new data. As economic and financial conditions. As economic and financial conditions change, the model should adapt to identify emerging trends in house price patterns.

END OF REPORT