🏠 ML-POWERED HOUSE PRICE PREDICTION

Prepared by:

Hari Krishna Kumar .S

Contents:

1.Problem Statement:

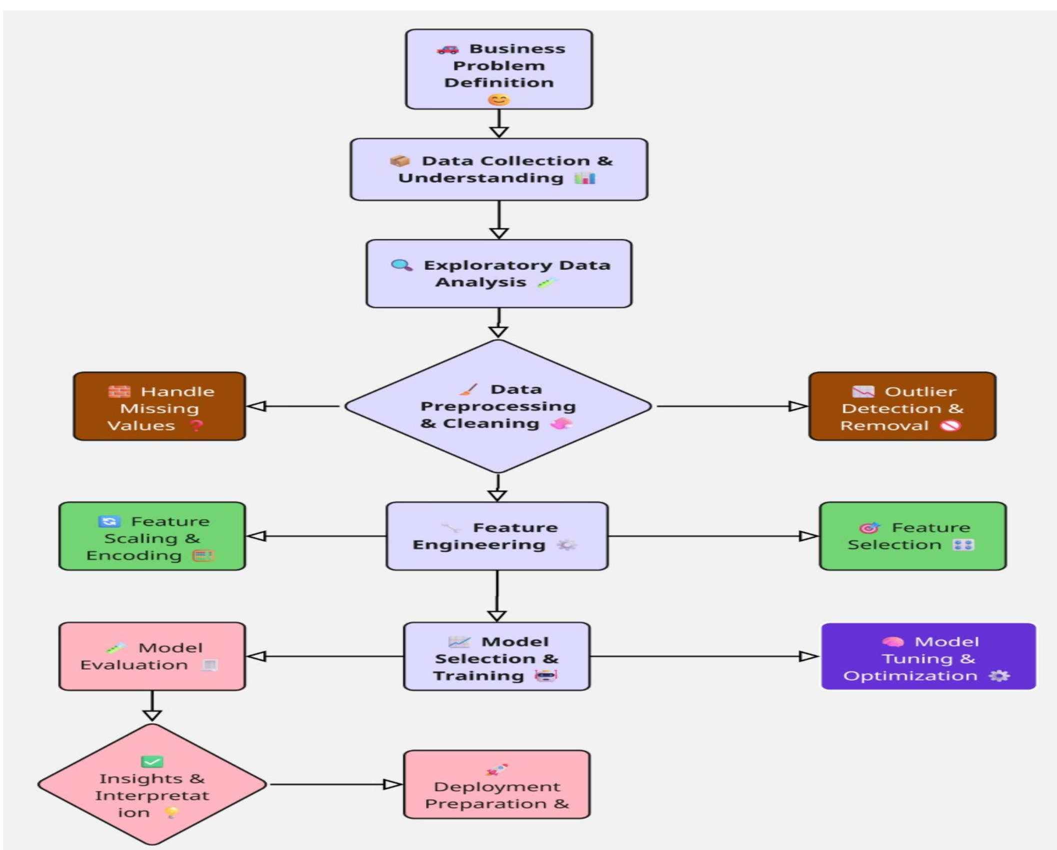
- Inaccurate Property Valuations Leading to Lost Revenue or Over-payment
- Inefficient Investment Decisions for Real estate investors to identify the high-potential properties in competitive markets
- Governments and tax authorities need to value millions of proprieties so manual processes are costly and slow
- Banks and mortgage lenders face risks to identify the real property values to lend their money

2.Dataset:

Dataset Link: 🏠 House Price Prediction

This dataset consists of **13 features and 545 rows**.
The features in the dataset include: price, area,bedrooms,bathrooms,stories,Main road,guestroom,basement,Hot water heating,Airconditioning,parking,Preferred area,Furnishing status

3.Project Workflow:



4.Business Problem & ML Objective:

The Business Challenge:

- Inaccurate Property Valuations Leading to Lost Revenue or Over-payment
- Inefficient Investment Decisions for Real estate investors to identify the high-potential properties in competitive markets
- Governments and tax authorities need to value millions of proprieties so manual processes are costly and slow
- Banks and mortgage lenders face risks to identify the real property values to lend their money

Why This Matters:

- It reduce the time by doing manual processes
- Reduce the risk of estimating the property value in the markets
- Money lenders to identify the real value of the property

Machine Learning Objective

Task: Regression models to predict house price

- Input: Feature with house amenities
- Output: Predict the House price

Success Metrics:

Primary -R2 score to identify the model performance. It range between 0 – 1 (nearer to 1 is predicting good)

Secondary- RMSE score

Business KPI- Reduce the risk of estimation and time

Beneficiaries:

* Banks and Money Lenders
* Governments and tax authorities
* Real Estate Companys

5.Data Cleaning :

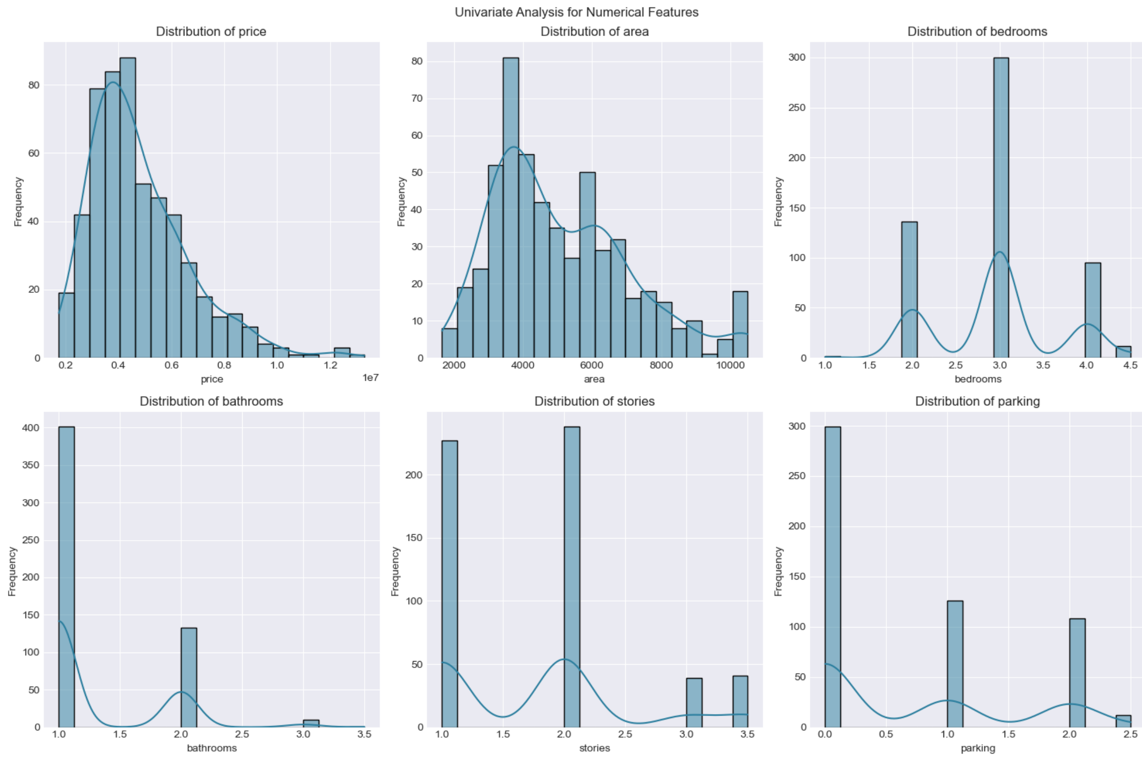
No null values ,Missing values and duplicate values present in the dataset

6.Exploratory Data Analysis:

Exploratory Data Analysis helps us understand how the features of the dataset vary for the different variables of the House Price Prediction data. We first start by understanding the features alone (Uni-variate Analysis) and then perform bi-variate and multivariate analysis to understand the data and relations between the features better.

Uni-variate Analysis:

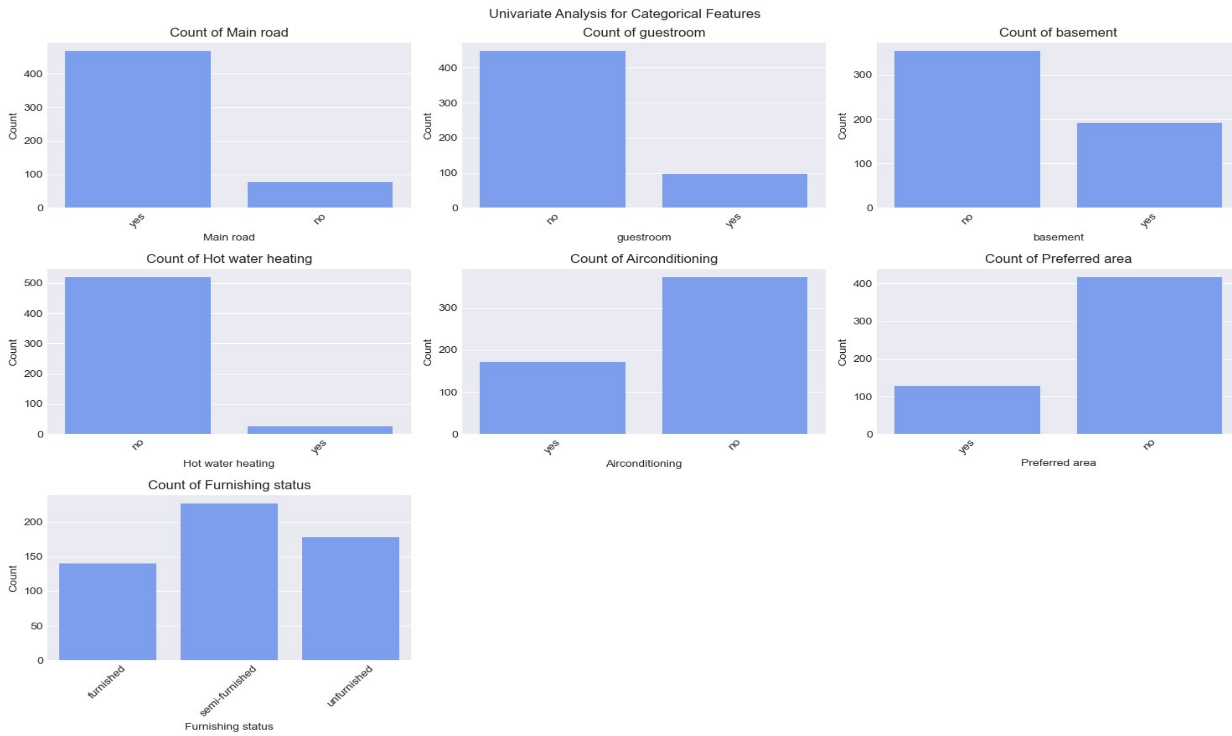
Distributions of Numerical Features



💡 Key Observation

- Price and areas are positively skew
- 3 bedroom houses are present higher
- 1 bathroom houses are present higher
- 1 and 2 stories houses are present higher
- without parking houses are present higher

Distributions of Categorical Features

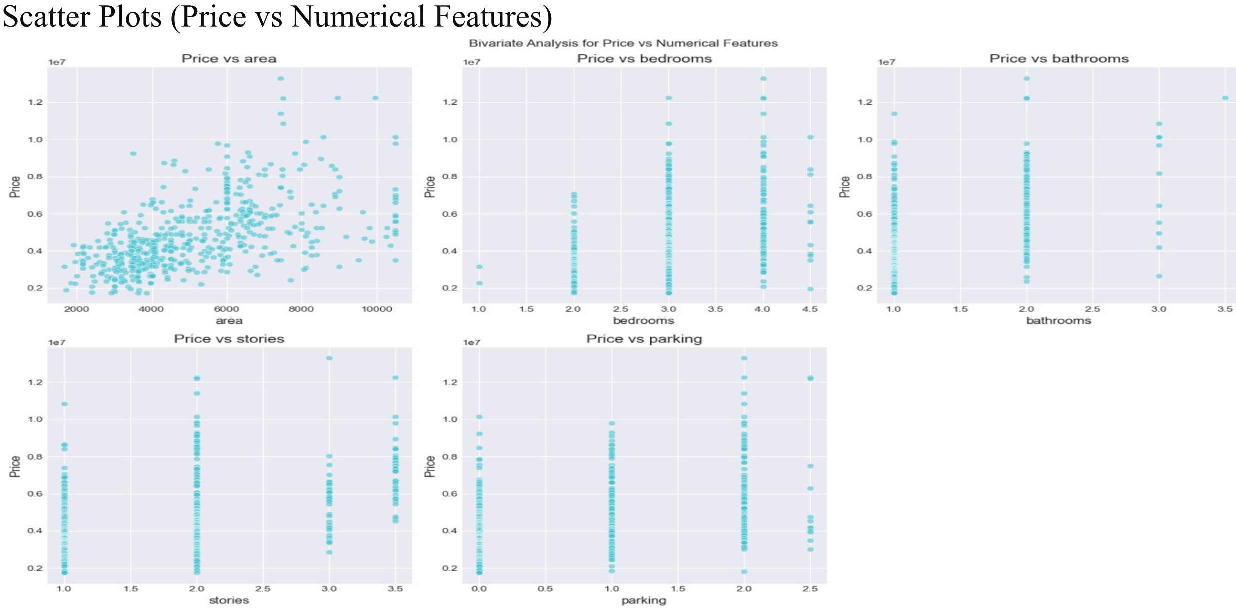


💡 Key Observation

- Most of the houses located near main road
- Most of the houses are semi-furnished

Bi-variate Analysis for numerical and categorical features:

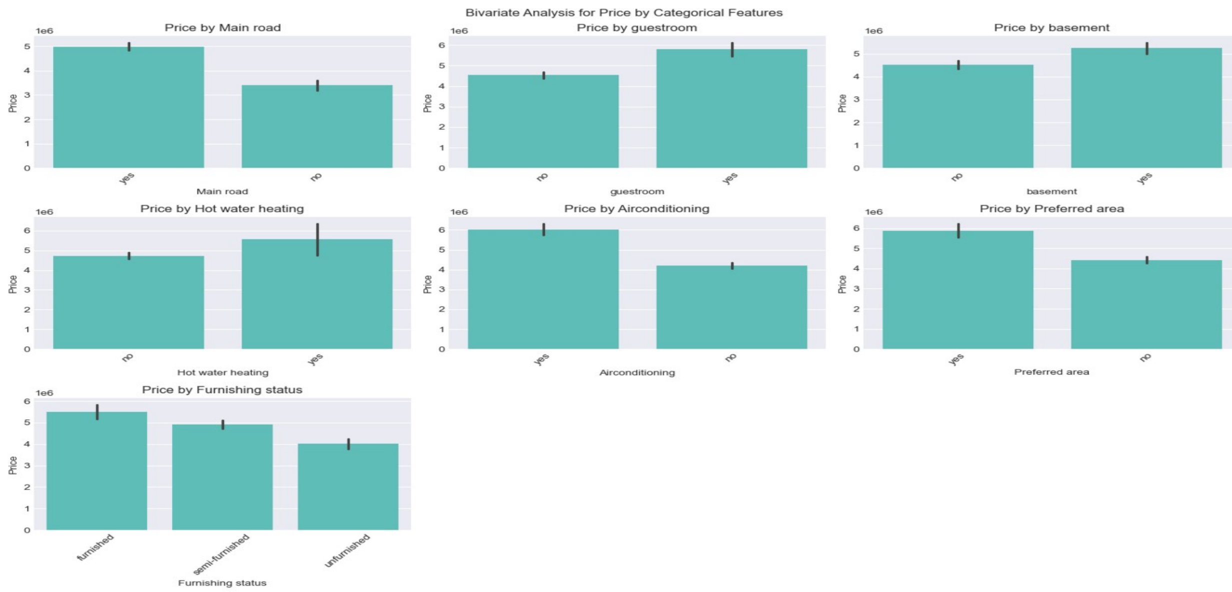
Scatter Plots (Price vs Numerical Features)



💡 Key Observation :

- House amenities increase price also increases

Box Plots (Price by Categorical Features)

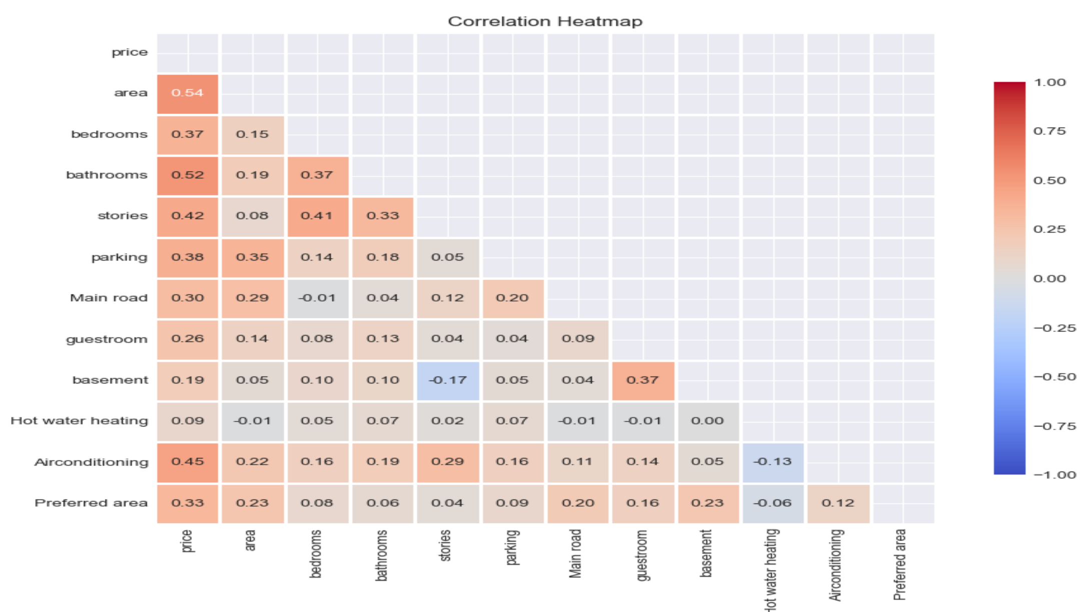


💡 Key Observation :

- House amenities increase price also increases

Numerical Analysis:

Correlation between features

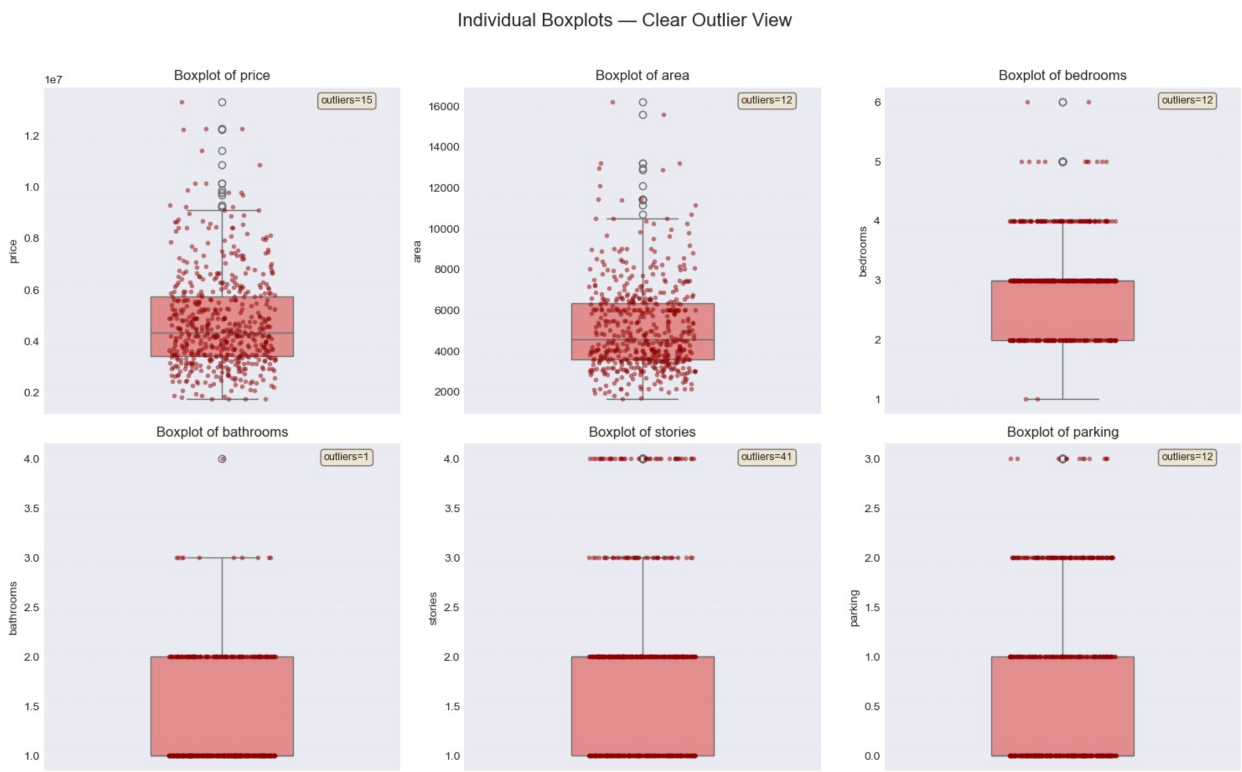


💡 Key Observation :

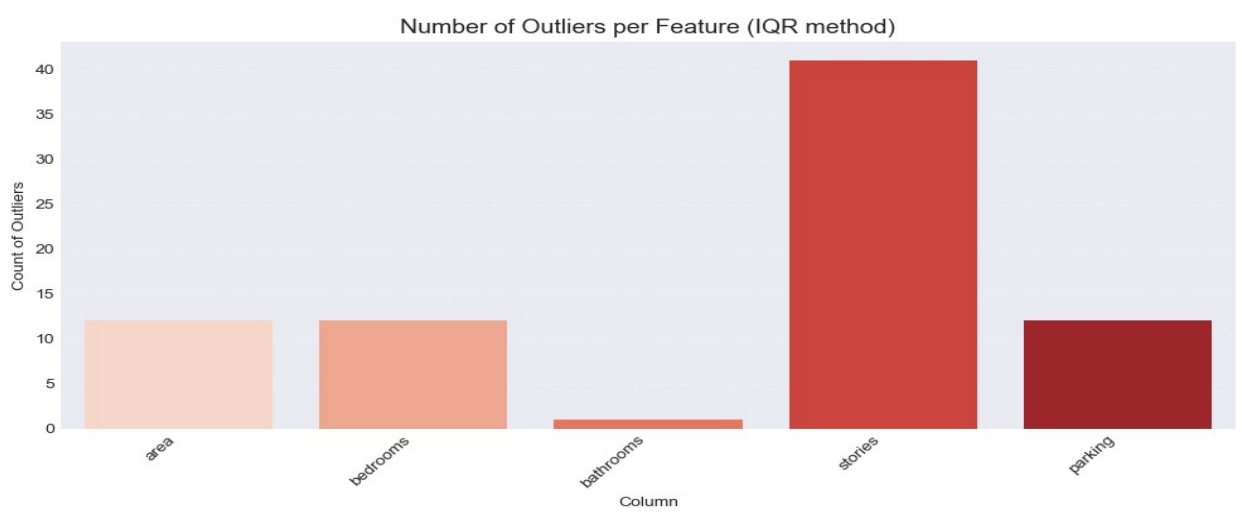
- compare to other features Area and bedroom highly correlated with price
- compare to other features Bathroom and stories highly correlated with bedroom

7.DATA Preprocessing :

Box plot for identifying outliers with counts:



Barplot of outlier counts:



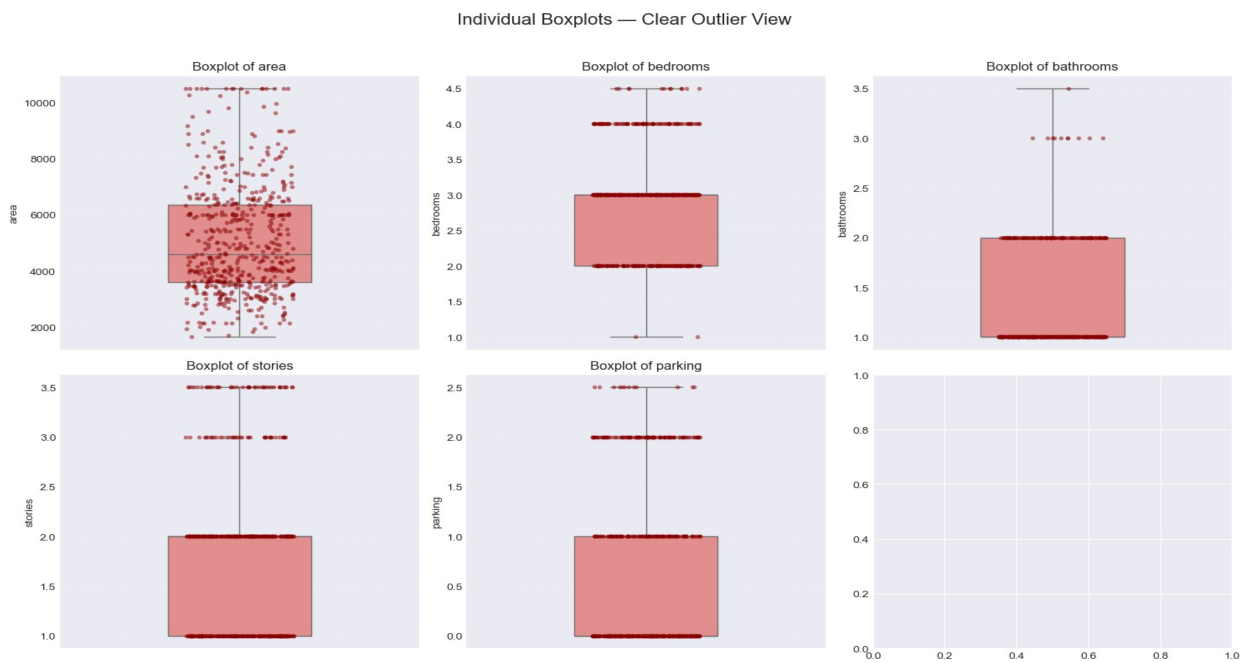
💡 Key Observation :

- Outliers present in all the features
- Stories has high outliers compare to others

Handling outliers:

Replace outlier with suitable values by Quantile method and Inter quartile range

Boxplot and Stripplot after Handling outliers:



Feature Engineering:

Feature engineering is the process of transforming raw data into meaningful features (variables) that enhance the performance, accuracy, and interpretability of machine learning models. It improved Model Performance and Reduced Over-fitting and Increased Generalization

New engineered features are 'price_log', 'area_per_room', 'area_per_bedroom', 'area_per_bathroom', 'small_house', 'large_house', 'total_rooms', 'bed_bath_ratio', 'bath_per_bedroom', 'amenity_score', 'luxury_score', 'stories_parking', 'has_parking', 'high_parking', 'furnishing_score', 'area_x_aircon', 'area_x_prefarea', 'area_x_luxury', 'bedrooms_x_aircon', 'log_area', 'log_parking', 'area_bin', 'price_bin'

Splitting of Data

Splitting the dataset to x as independent variable and y as target variable

8.Train and test split:

Train and test split in the data to train the models. Stratified split to preserve class balance.

Train Test split values and distributions:

- Training set: 436 samples (80%)
- Testing set: 109 samples (20%)

9. Feature Scaling:

RobustScaler reduces the impact of outliers by scaling data using median and interquartile range (IQR) which makes it fit to extreme values. We use it when our data contains many outliers and we need to maintain relative distances between non-outlier data points or we're working with algorithms which are sensitive to extreme values.

10. Feature Extraction(Dimensionality Reduction):

Create new features by combining or deriving information from existing ones to provide more meaningful input to the model.

PCA:

PCA (Principal Component Analysis) is a dimensionality reduction technique and helps us to reduce the number of features in a dataset while keeping the most important information. It changes complex datasets by transforming correlated features into a smaller set of uncorrelated components.

11. Model Training:

1.Linear Regression:

- Train the models with PCA and without PCA.
- Model without PCA Score is good.
- Model without PCA Score-MAE: 600340.02| RMSE: 970342.81| R2Score: 0.813
- Model without PCA Score-MAE: 556604.62| RMSE: 948141.11| R2Score: 0.822

2.Random Forest Regression:

- Train the models with PCA and without PCA.
- Model without PCA Score is good.

- Model without PCA Score-MAE: 482886.39| RMSE: 873841.18| R2Score: 0.848
- Model without PCA Score-MAE: 739039.97| RMSE: 1179273.01| R2Score: 0.724

3.XGBoost Regression:

- Train the models with PCA and without PCA.
- Model without PCA Score is good.
- Model without PCA Metrics=MAE: 472300.71| RMSE: 888729.65| R2Score:0.843
- Model with PCA Metrics= MAE: 626229.18| RMSE: 1037736.66| R2Score: 0.786

4.LightGBM Regression:

- Train the models with PCA and without PCA.
- Model without PCA Score is good.
- Model without PCA Metrics=MAE: 485159.25| RMSE: 834056.82| R2Score: 0.862
- Model with PCA Metrics= MAE: 671764.20| RMSE: 1086534.19| R2Score: 0.766

5.Ridge Regression:

- Train the models with PCA and without PCA.
- Model without PCA Score is good.
- Model without PCA Metrics=MAE: 574467.57| RMSE: 956358.98| R2Score: 0.819
- Model with PCA Metrics= MAE: 555918.64| RMSE: 948964.66| R2Score: 0.821

6.Lasso Regression:

- Train the models with PCA and without PCA.
- Model without PCA Score is good.
- Model without PCA Metrics=MAE: 595687.85| RMSE: 967260.07| R2Score: 0.814
- Model with PCA Metrics= MAE: 556604.20| RMSE: 948146.83| R2Score: 0.822

7.ElasticNet Regression:

- Train the models with PCA and without PCA.
- Model without PCA Score is good.
- Model without PCA Metrics=MAE: 567270.31| RMSE: 954954.73| R2Score: 0.819

- Model with PCA Metrics= MAE: 555613.46| RMSE: 950067.96| R2Score: 0.821

LightGBM Regression without PCA is consider as a best model when compared to others with the error metrics of MAE: 485159.25| RMSE: 834056.82| R2Score: 0.862.so It is selected as a final model

12.Hyper Parameter Optimization:

Hyperparameters are the parameters that determine the behavior and performance of a machine-learning model. These parameters are not learned during training but are instead set prior to training. The process of finding the optimal values for these hyperparameters is known as hyperparameter optimization

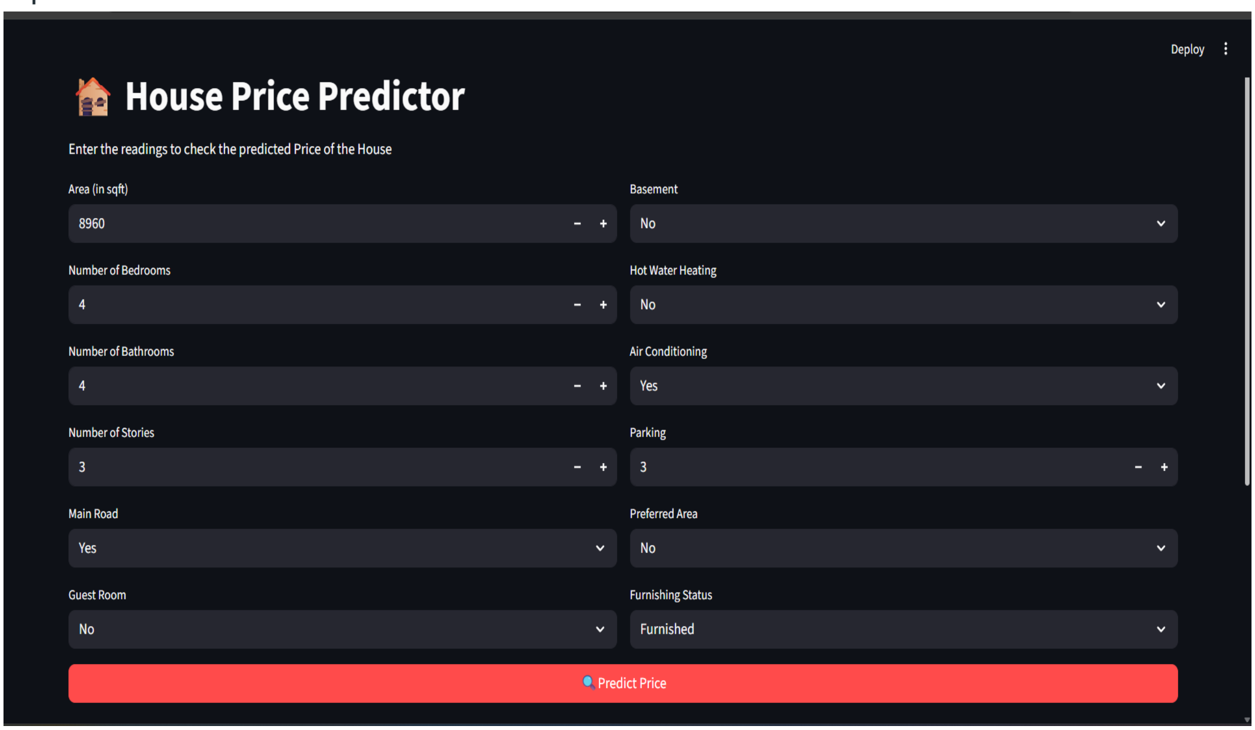
RandomSearchCV:

GridSearchCV is a technique for hyperparameter tuning that performs an exhaustive search over a predefined set of parameter values for a machine learning model, evaluating each combination using cross-validation to find the optimal settings.

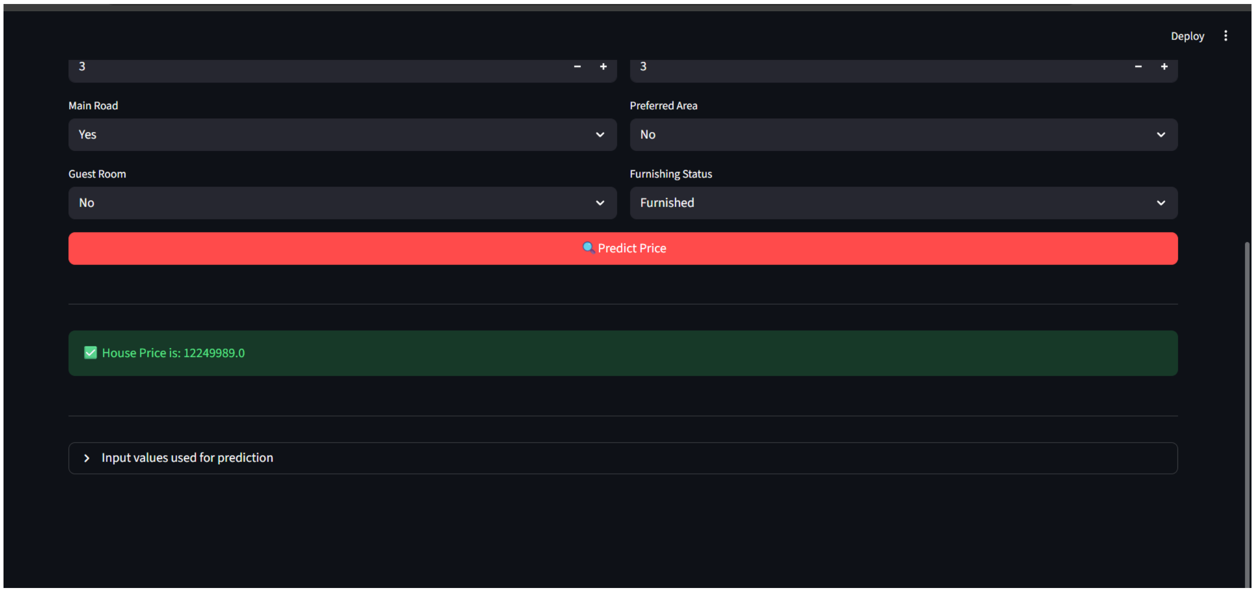
13.Webframe work:

Streamlit is an open-source Python library that makes it easy to create and share custom web apps for machine learning.By using Streamlit you can quickly build and deploy powerful data applications.

Input Data:



Result:



14.Business Insights & Recommendations

- LightGBM Regression achieved the highest R2 score is 0.862

- In future people may use more facilities and amenities in living life. we can use those amenities as a feature to train the model it give the better results

- Currently deployed in streamlit cloud only if client needed we can deploy web-frame work in the cloud(AWS,AZURE,GCP)for the production