HOUSE PRICE PREDICTION

MULTIPLE LINEAR REGRESSION

NAME	PAVITHRA M
TITLE	HOUSE PRICE PREDICTION
DATE	NOVEMBER 2025

TABLE OF CONTENT

1. Introduction

House price prediction is a critical task in the real estate domain, enabling buyers, sellers, and businesses to make informed decisions based on data-driven insights. This project focuses on developing a **Multiple Linear Regression** model to predict house prices using key property features such as size, location-related factors, and other relevant attributes.

The project follows an end-to-end data analytics and machine learning approach, including data preprocessing, exploratory data analysis, model training, and performance evaluation. To enhance usability, the trained model is deployed as an **interactive Streamlit web application**, allowing users to input property details and obtain real-time price predictions. This project demonstrates practical application of regression techniques and model deployment in a real-world scenario.

2. Aim

The aim of this project is to design and implement a house price prediction system using a **Multiple Linear Regression** model. The project focuses on analyzing housing data to identify the key factors affecting house prices and applying statistical and machine learning techniques to build an accurate predictive model.

The project also aims to provide an interactive and user-friendly interface by deploying the trained model as a **Streamlit web application**, enabling users to input property-related information and obtain real-time house price predictions. Through this project, practical knowledge of data preprocessing, regression analysis, model evaluation, and deployment is effectively demonstrated

3. Problem Statement

In the real estate industry, accurately estimating house prices is a challenging task due to the influence of multiple factors such as property size, location, number of rooms, and other amenities. Traditional pricing methods often rely on subjective judgment or limited analysis, which can lead to inaccurate price estimation and inefficient decision-making.

There is a need for a data-driven approach that can analyze historical housing data and predict house prices with greater accuracy and consistency. The challenge lies in selecting relevant features, handling data preprocessing, and building a reliable predictive model that can generalize well to unseen data.

This project addresses the problem by developing a **Multiple Linear Regression–based house price prediction system** and deploying it as an interactive **Streamlit web application**, enabling users to obtain accurate and real-time price predictions based on input property features.

4. Project Workflow

- Data Collection
- Data Understanding
- Data Cleaning
- Data Preprocessing
- Exploratory Data Analysis (EDA)
- Model Training using Multiple Linear Regression
- Model Evaluation
- Streamlit Application Development
- Model Deployment

5. Data Collection

The dataset required for this project was sourced from **Kaggle**, a publicly available data platform. The dataset consists of historical housing data containing various property-related features along with their corresponding house prices. The data was collected in a structured format, making it suitable for analysis and model development.

6. Data Understanding

After collecting the dataset from Kaggle, an initial analysis was performed to understand its structure and characteristics. The dataset was examined to identify the number of records, feature types, and the relationship between independent variables and the target variable (house price).

This step involved checking for missing values, detecting inconsistencies, and analyzing basic statistical summaries. Understanding the data helped in selecting relevant features and planning appropriate preprocessing techniques for effective model building.

The dataset contains :

- Rows - 545
- Columns - 13

7. Data Cleaning

Data cleaning ensures the dataset is **accurate, consistent, and ready for modeling**.

Key steps performed:

- **Handling Missing Values :** There is no missing values.
- **Removing Duplicates :** There is no duplicate rows.
- **Correcting Data Types :** Ensured numeric and categorical columns had appropriate types.

8. Data Preprocessing

Data preprocessing transforms the raw dataset into a format suitable for modeling.

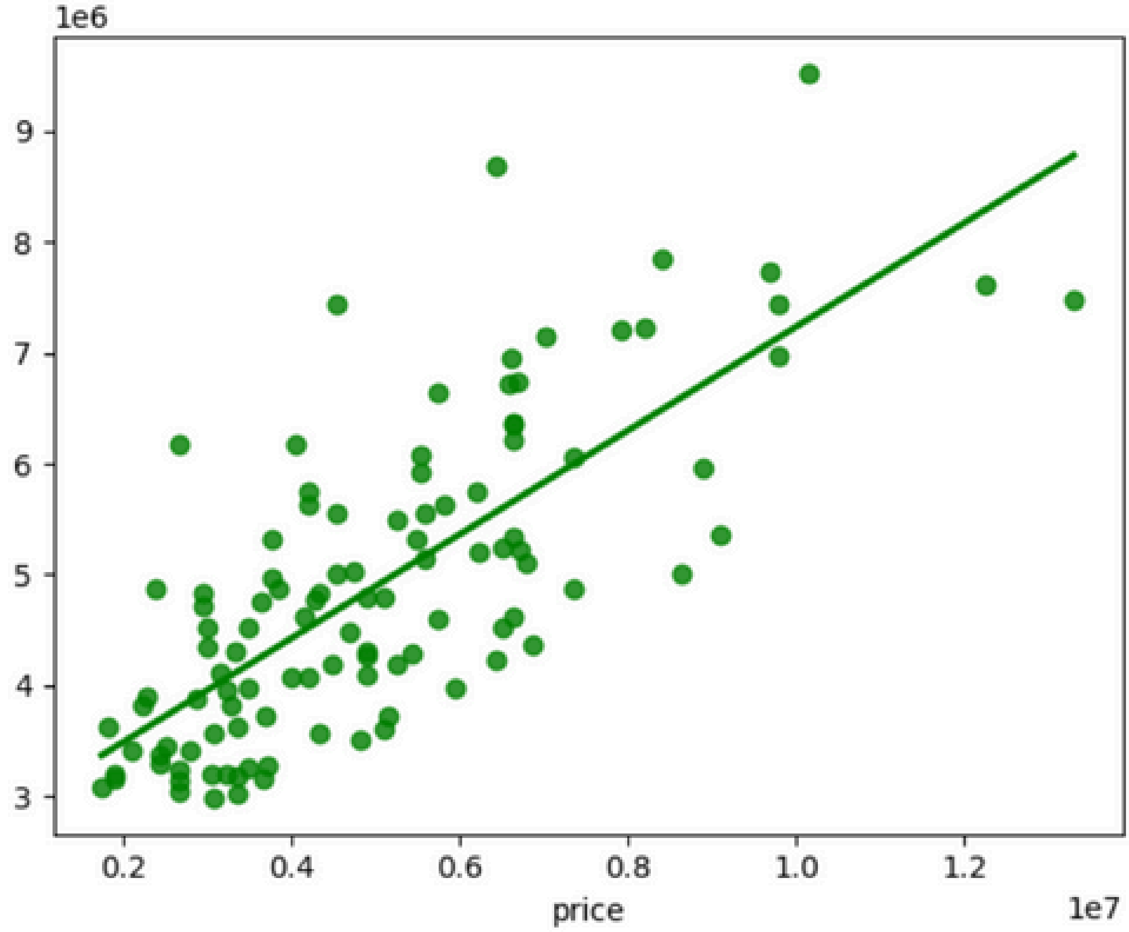
Categorical variables were encoded into numeric formats. The data was split into training and testing sets. After these steps, the dataset was fully preprocessed and ready for model building and evaluation.

9. Exploratory Data Analysis

In this project, EDA was primarily focused on analyzing the performance of the Multiple Linear Regression model through visualizations. Since the dataset had been cleaned and preprocessed, the key goal of EDA was to compare the actual house prices with the predicted values to evaluate how well the model captures underlying patterns.

Regression Plot

- A regression plot was created to visualize the relationship between the actual Price values (y_test) and the predicted Price values (y_predict).
- The plot allows a visual inspection of how closely the predictions align with the true values.
- A line along the diagonal represents perfect predictions, and the scatter points show how actual predictions deviate from this ideal.
- The plot indicated that most predicted values were close to the actual values, confirming that the model was performing well.

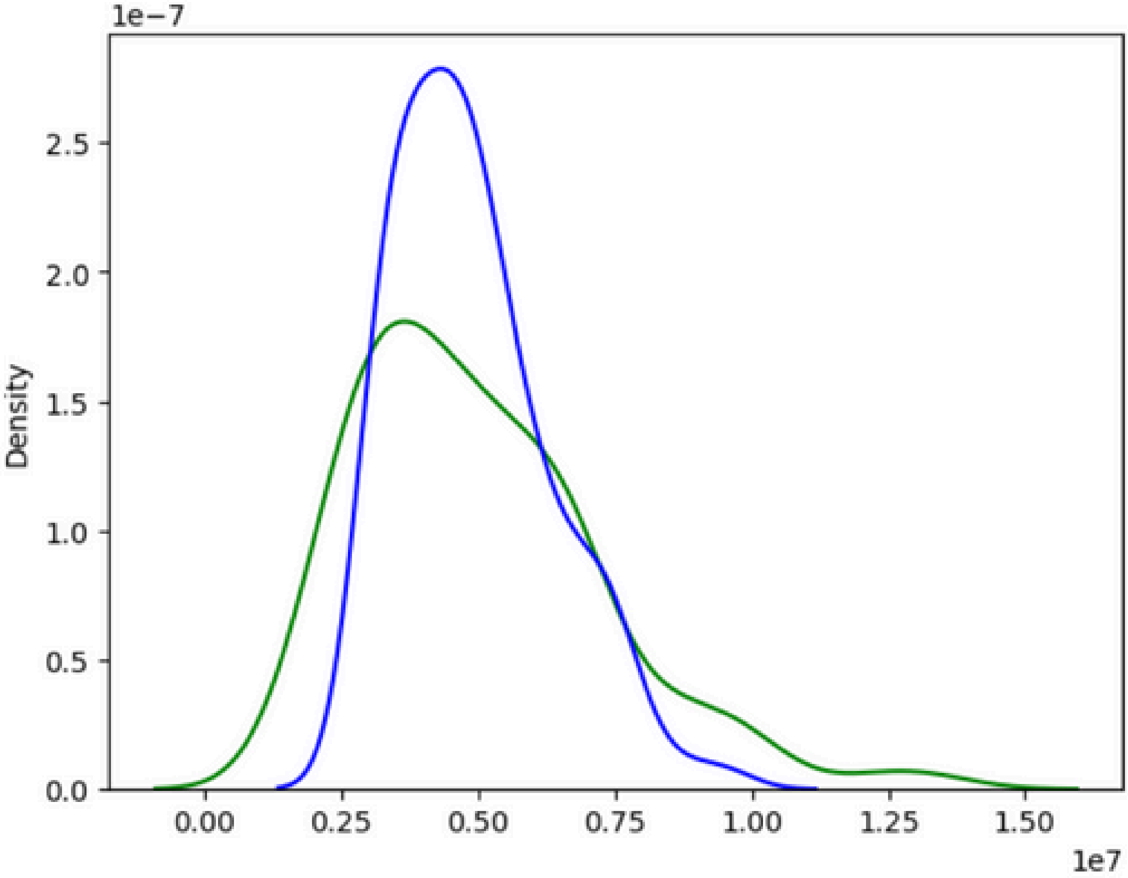


Insights :

- **Positive correlation:** The plot shows a clear upward trend, indicating a strong positive correlation between actual and predicted house prices.
- **Model performance:** The Multiple Linear Regression model captures the overall trend well.
- **Prediction accuracy:** Most points are close to the regression line, showing predictions are generally accurate.
- **Interpretation:** The model performs reasonably well but may benefit from additional features or outlier handling to improve accuracy.

Actual vs Fitted Value Plot

- Another key visualization plotted the fitted values versus the actual values to further assess model performance.
- This plot helps in identifying any patterns or systematic errors in the predictions, such as underestimating or overestimating prices for specific ranges.
- The results showed a strong alignment between actual and fitted values, suggesting that the Multiple Linear Regression model captures the main trends in house pricing effectively.



Insights :

- **Distribution comparison:** The green line (actual values) and blue line (predicted values) show similar overall distribution patterns.
- **Model fit:** Peaks of predicted values align closely with actual values, indicating the model captures central tendencies well.
- **Deviation:** Some differences in the tails suggest the model underestimates or overestimates extreme house prices.
- **Insight:** The model performs reasonably well for most prices but may require additional features or transformation to better capture outliers.

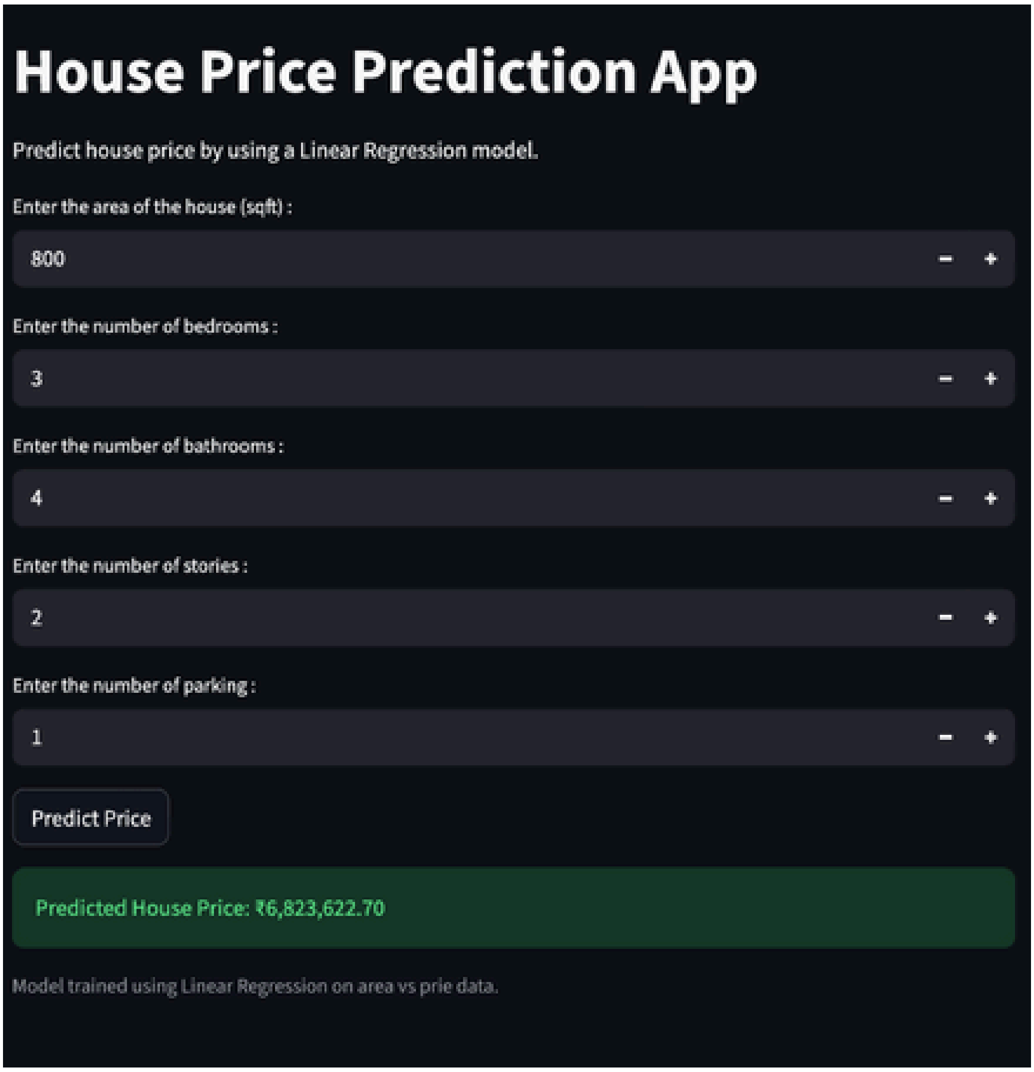
10. Model Building & Evaluation

For predicting house prices, **Multiple Linear Regression (MLR)** was selected as the modeling approach because the target variable (Price) is continuous.

The model was evaluated using metrics such as **Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R² score** to ensure accuracy and reliability of predictions.

11. Model Deployment

The trained Multiple Linear Regression model was saved for future use and loaded whenever predictions were needed. Using Streamlit, a simple web application was created to allow users to input house features and get real-time price predictions. The app can be run in PyCharm or deployed online, making the project interactive and user-friendly.



House Price Prediction App

Predict house price by using a Linear Regression model.

Enter the area of the house (sqft) : 800

Enter the number of bedrooms : 3

Enter the number of bathrooms : 4

Enter the number of stories : 2

Enter the number of parking : 1

Predicted House Price: ₹6,823,622.70

Model trained using Linear Regression on area vs prie data.

12. Overall Insights

- The Multiple Linear Regression model demonstrates a strong positive relationship between actual and predicted house prices.
- Regression plots indicate that the model fits the data well, with most predictions closely aligned to actual values.
- Distribution analysis shows that predicted values follow a similar pattern to actual prices, capturing the central trend effectively.
- Minor deviations and spread in higher price ranges suggest the presence of outliers and feature limitations.

- The model performs well for average-priced houses but shows reduced accuracy for extreme values.

13. Conclusion

The House Price Prediction model developed using Multiple Linear Regression demonstrates strong predictive capability, with a clear positive relationship between actual and predicted prices. Exploratory and regression analyses confirm that the model effectively captures the underlying trend in the data and provides reliable predictions for most cases. While minor deviations are observed due to outliers and feature limitations, the overall performance is satisfactory. With further feature enhancement and advanced modeling techniques, the model's accuracy and robustness can be improved, making it a valuable baseline for house price estimation.