# Boston Housing Price Prediction using Linear Regression

## ■ Abstract

This project predicts median housing prices in the Boston area using **Linear Regression**. The model analyzes various socio-economic and environmental factors that influence housing prices. By applying regression techniques, the project demonstrates how statistical learning can be used for price forecasting and real-estate analysis.

## ■ Dataset Information

- **Source:** [Boston Housing Dataset on Kaggle](https://www.kaggle.com/datasets/vikrishnan/boston-house-prices)
- **Attributes:**
- `CRIM`: Per capita crime rate by town
- `ZN`: Proportion of residential land zoned for large lots
- `INDUS`: Proportion of non-retail business acres per town
- `CHAS`: Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- `NOX`: Nitric oxide concentration (ppm)
- `RM`: Average number of rooms per dwelling
- `AGE`: Proportion of owner-occupied units built before 1940
- `DIS`: Weighted distances to employment centers
- `RAD`: Accessibility index to radial highways
- `TAX`: Property tax rate per $10,000
- `PTRATIO`: Pupil-teacher ratio
- `B`: Proportion of Black population
- `LSTAT`: % lower status of the population
- `MEDV`: Median value of owner-occupied homes (target variable)

## ■■ Project Workflow

1. **Data Loading and Exploration**
- Import dataset, inspect missing values, and examine feature distribution.
2. **Data Preprocessing**
- Handle missing data, feature scaling, and correlation analysis using a heatmap.
3. **Feature Selection**
- Select important features influencing `MEDV` based on correlation values.
4. **Model Training**
- Train a **Linear Regression** model using Scikit-learn.
5. **Model Evaluation**
- Evaluate using **R^2 Score**, **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **Root Mean Squared Error (RMSE)**.
6. **Visualization**
- Compare actual vs. predicted prices and plot residual errors.

## ■ Mathematical Explanation

Linear Regression is modeled as:

$$ y = \beta_0 + \beta_1x_1 + \beta_2x_2 + ... + \beta_nx_n + \varepsilon $$

Cost Function (MSE):

$$ J(\beta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\beta(x^{(i)}) - y^{(i)})^2 $$

Gradient Descent:

$$ \beta_j := \beta_j - \alpha \frac{\partial}{\partial\beta_j} J(\beta) $$

Where:
- `α` = learning rate
- `m` = number of training examples

The model minimizes the cost function iteratively to find the best-fit line.

## ■ Results and Analysis
- **R^2 Score:** ~0.73 (model explains 73% of variance)
- **MAE:** Low average error between predicted and actual prices
- **Observation:** Features like `RM`, `LSTAT`, and `PTRATIO` have the strongest influence on house prices.

The predicted vs. actual plot demonstrates a strong linear relationship, validating model performance.

## ■ Conclusion
The Linear Regression model effectively predicts Boston housing prices, providing insights into key factors affecting home values. Further improvements could include polynomial regression or ensemble methods for enhanced accuracy.

## ■■■ Author
Elavarasi Chinnadurai
Department of Agriculture Engineering
Passionate about Data Science, Machine Learning, and Predictive Modeling.