**📘 Linear Regression Assignment: Boston Housing Dataset**

**🏠 Dataset Description**

The **Boston Housing dataset** contains information collected by the U.S Census Service about housing in the area of Boston, Massachusetts. It includes **506 rows and 14 columns**.

**🔑 Features (Independent Variables):**

* CRIM: Per capita crime rate by town
* ZN: Proportion of residential land zoned for lots over 25,000 sq. ft.
* 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 (parts per 10 million)
* RM: Average number of rooms per dwelling
* AGE: Proportion of owner-occupied units built prior to 1940
* DIS: Weighted distances to five Boston employment centers
* RAD: Index of accessibility to radial highways
* TAX: Full-value property tax rate per $10,000
* PTRATIO: Pupil-teacher ratio by town
* B: 1000(Bk - 0.63)^2 where Bk is the proportion of Black people by town
* LSTAT: % lower status of the population

**🎯 Target Variable:**

* MEDV: Median value of owner-occupied homes in $1000s

**🧠 Section A: Conceptual Understanding**

**Task A1: Define and Explain Linear Regression**

* Define Linear Regression and explain its types.
* Why is Linear Regression suitable for the Boston Housing dataset?

**Task A2: Assumptions of Linear Regression**

* Linearity
* No multicollinearity
* Homoscedasticity
* Normality of residuals
* Independence of errors

📌 Deliverable: Describe each assumption and how you plan to check it in this dataset.

**Ans:-**

**Definition**:  
Linear Regression is a supervised machine learning algorithm used for predicting a continuous dependent variable (target) based on one or more independent variables (features). It establishes a linear relationship between the variables by fitting a line (in simple linear regression) or a hyperplane (in multiple linear regression) that minimizes the difference between actual and predicted values.

**Types of Linear Regression**:

1. **Simple Linear Regression**:
   * Involves one independent variable and one dependent variable.
   * Example: Predicting house price (MEDV) based on number of rooms (RM).
2. **Multiple Linear Regression**:
   * Involves two or more independent variables.
   * Example: Predicting MEDV using CRIM, RM, AGE, LSTAT, etc.
3. **Ridge and Lasso Regression (Regularized Linear Models)**:
   * These are advanced versions of linear regression that include penalty terms to prevent overfitting and handle multicollinearity.

**Why is Linear Regression suitable for the Boston Housing dataset?**

* The target variable, **MEDV**, is continuous.
* Many features have a potential linear relationship with MEDV (e.g., RM, LSTAT).
* The dataset has a manageable number of features and is widely used for regression model practice.
* It allows interpretation of feature impact through coefficients.

### **Task A2: Assumptions of Linear Regression**

To build a valid and interpretable linear regression model, we must verify these assumptions:

#### 1. **Linearity**

* **Definition**: The relationship between independent variables and the target is linear.
* **How to Check**:
  + Use scatterplots or a correlation matrix to visualize linear relationships.
  + Check residual vs. fitted value plots: residuals should be randomly scattered without a clear pattern.

#### 2. **No Multicollinearity**

* **Definition**: Independent variables should not be highly correlated with each other.
* **How to Check**:
  + Calculate **Variance Inflation Factor (VIF)** for each feature. A VIF > 5 (or 10) indicates multicollinearity.
  + Use a heatmap or correlation matrix to visually inspect pairwise correlations.

#### 3. **Homoscedasticity**

* **Definition**: Constant variance of residuals for all levels of independent variables.
* **How to Check**:
  + Plot residuals vs. predicted values. The spread should be constant (not funnel-shaped).
  + Perform statistical tests like Breusch-Pagan test for confirmation.

#### 4. **Normality of Residuals**

* **Definition**: Residuals (errors) should be approximately normally distributed.
* **How to Check**:
  + Use a **Q-Q plot** (quantile-quantile plot) for visual inspection.
  + Plot histogram of residuals.
  + Perform **Shapiro-Wilk test** or **Kolmogorov-Smirnov test** for normality.

#### 5. **Independence of Errors**

* **Definition**: Residuals should be independent of each other.
* **How to Check**:
  + Use **Durbin-Watson test** to detect autocorrelation.
  + Plot residuals over time (if data is time-based) to ensure no patterns.

**🔬 Section B: Exploratory Data Analysis (EDA)**

**Task B1: Load and Explore the Dataset**

* Display basic stats (.describe(), .info(), missing values).
* Visualize the distribution of MEDV.

**Task B2: Feature Relationships**

* Plot:
  + Scatter plot of RM vs MEDV
  + Scatter plot of LSTAT vs MEDV
  + Correlation heatmap of all features

📌 Deliverable: Identify the top 3 most correlated features with MEDV (both positive and negative).

**🧮 Section C: Model Building**

**Task C1: Preprocessing**

* Normalize/standardize data.
* Handle any missing or invalid values (none in this dataset).
* Optional: Use train\_test\_split() to split data (80/20).

**Task C2: Build a Simple Linear Regression Model**

* Use only RM to predict MEDV.

**Task C3: Build a Multiple Linear Regression Model**

* Use all available features.

📌 Deliverable: Compare the performance of both models.

**🧪 Section D: Model Evaluation**

**Task D1: Evaluation Metrics**

Calculate and interpret using Python code:

* **MSE** (Mean Squared Error)
* **RMSE**
* **R² Score**
* **Adjusted R²** (for multiple regression)

📌 Deliverable: Discuss which model is better and why.

**📈 Section E: Visualization**

* Plot regression line for Simple Linear Regression (e.g., RM vs MEDV)
* Residuals plot
* Actual vs Predicted values plot for multiple regression

📦 **Submission Format**

* Push your completed python code file to your GitHub repository.