SIMPLE LINEAR AND MULTIPLE LINEAR REGRESSION ASSIGNMENT

Problem Statement:

This assignment focuses on understanding and applying **Simple Linear Regression** and **Multiple Linear Regression** to predict outcomes based on given datasets. By exploring the datasets, students will learn how to build regression models, evaluate their performance, and interpret their coefficients to derive meaningful insights.

Guidelines:

1. Foundational Knowledge:
- Understand the concepts of Simple and Multiple Linear Regression.
- Learn how to estimate the relationship between independent and dependent variables using linear models.
- Recognize the assumptions and limitations of Linear Regression.
- Differentiate between Simple and Multiple Linear Regression.

2. Data Exploration:
- Analyze the dataset's structure and characteristics using techniques such as scatter plots, box plots, pair plots, and correlation matrices.
- Identify relationships between variables and assess their suitability for regression modeling.

3. Preprocessing and Feature Engineering:
- Handle missing values, outliers, and categorical variables appropriately.
- Scale or normalize features if necessary.
- Split the dataset into training and testing sets.

4. Model Construction:
- Choose an appropriate implementation for Simple and Multiple Linear Regression.
- Train a Simple Linear Regression model using one independent variable.
- Train a Multiple Linear Regression model using multiple independent variables.
- Interpret the coefficients of the regression models.

5. Model Evaluation:
- Evaluate the trained models using metrics such as **Mean Squared Error (MSE)**, **Mean Absolute Error (MAE)**, **R-squared (R^2)**, and **Adjusted R-squared (Adjusted R^2)**.
- Assess how well the models fit the data and explain any discrepancies.

6. Model Optimization and Insights:
- Analyze the residuals to ensure the assumptions of Linear Regression are met.
- Compare the performance of Simple Linear Regression and Multiple Linear Regression models.
- Draw conclusions based on the coefficients, p-values, and R-squared values.

Step-by-Step Approach to Linear Regression Modeling:

1. Setup and Data Preparation:
- Import necessary libraries such as `pandas`, `matplotlib`, `seaborn`, and `scikit-learn`.
- Load the dataset for regression analysis.
- Conduct exploratory data analysis (EDA) to understand the dataset.
- Preprocess the data by handling missing values, encoding categorical variables, and normalizing/standardizing features if needed.

2. Simple Linear Regression:
- Select one independent variable (predictor) and one dependent variable (response).
- Train a Simple Linear Regression model using the selected variables.
- Visualize the regression line on a scatter plot to show the relationship between the predictor and response variable.

3. Multiple Linear Regression:
- Select multiple independent variables (predictors) and one dependent variable (response).
- Train a Multiple Linear Regression model using the selected features.
- Analyze the regression coefficients to understand the impact of each predictor on the response variable.

4. Model Evaluation:
- Calculate evaluation metrics for both Simple and Multiple Linear Regression models:
  - Mean Squared Error (MSE)
  - Mean Absolute Error (MAE)
  - R-squared (R^2)
  - Adjusted R-squared (Adjusted R^2)
- Compare the performance of both models and justify the results.

5. Residual Analysis:
- Perform residual analysis to validate the assumptions of Linear Regression:
  - Linearity
  - Homoscedasticity (constant variance)
  - Independence of errors
  - Normality of residuals

6. Model Optimization:
- Identify and remove irrelevant or highly correlated predictors to improve the model's performance.
- Assess the impact of feature selection on the R-squared and Adjusted R-squared values.

Links to Datasets for the Assignment:

1. Medical Cost Personal Datasets
[https://www.kaggle.com/datasets/mirichoi0218/insurance/data]
Use the dataset to predict medical insurance costs based on individual attributes like age, BMI, smoking status, region, etc..

2. Startup Dataset
[https://www.kaggle.com/datasets/karthickveerakumar/startup-logistic-regression/data]
Predict profit based on R&D Spend, Administration, Marketing Spend, and State.

3. Calculate Concrete Strength Dataset
[https://www.kaggle.com/datasets/prathamtripathi/regression-with-neural-networking/data]
Analyze the relationship between concrete features (e.g., water, cement, age) and the resulting concrete strength.