Lab Assignment 1: Simple Linear Regression using the Boston Housing dataset

Objective:

- Understand the basic concept of simple linear regression and implement it using Python.
- Use the Boston Housing dataset to train and test the simple linear regression model.

Prerequisites:

- Basic knowledge of Python programming.
- Familiarity with NumPy and Pandas library.
- Familiarity with Matplotlib library for data visualization.

Steps:

1. Import the necessary libraries (NumPy, Pandas, Matplotlib).
2. Load the Boston Housing dataset using Pandas.
3. Explore the dataset by printing the first few rows and checking the statistics of the dataset.
4. Create a scatter plot to visualize the relationship between the dependent variable (MEDV) and the independent variable (RM).
5. Split the dataset into training and testing sets.
6. Train a simple linear regression model using the training dataset.
7. Evaluate the model using the testing dataset.
8. Print the model coefficients (intercept and slope) and the R-squared value of the model.
9. Create a line plot to visualize the relationship between the dependent variable. (MEDV) and the independent variable (RM) along with the predicted values from the model.
10. Repeat steps 6-9 with different independent variables and compare the results.

Assignment Questions:

1. What is the relationship between the independent variable RM and the dependent variable MEDV in the Boston Housing dataset?
2. How does the R-squared value of the model change when you use different independent variables?
3. What are the pros and cons of using a simple linear regression model?

Submission Guidelines:

- Submit a Jupyter notebook with the complete code and the answers to the assignment questions.
- The Jupyter notebook should be well-documented, with clear explanations of the code and the steps taken.
- The code should be clean, readable, and well-organized.
- The visualizations should be labeled and clearly visible in the notebook.

Note:

- You can use the scikit-learn library to perform simple linear regression.
- The Boston Housing dataset is available in the scikit-learn library, and you can load it using the load_boston() function.
- The MEDV variable represents the median value of owner-occupied homes in $1000s.
- The RM variable represents the average number of rooms per dwelling.

The code will:

- import the necessary libraries, including pandas, sklearn's LinearRegression, mean_squared_error, r2_score, and train_test_split.
- load the Boston Housing dataset using sklearn's load_boston function and convert it into a Pandas dataframe.
- divide the data into input and output variables (X and y)
- Split the data into training and test sets using sklearn's train_test_split function
- fit the linear regression model on the training data using the LinearRegression().fit() method
- predict the values for the test set
- calculate the mean squared error and R-squared score using sklearn's mean_squared_error and r2_score functions