Lab Assignment 2: Comparing Simple and Multiple Linear Regression Models

Objective:

- To understand the difference between simple linear regression and multiple linear regression.
- To use the California Housing dataset to train and test simple and multiple linear regression models.
- To evaluate the performance of the models and compare them.

Instructions:

6. Start by importing the necessary libraries: pandas, numpy, fetch_california_housing, LinearRegression, r2_score,mean_squared_error, train_test_split and matplotlib.pyplot.
1. Load the California Housing dataset using fetch_california_housing and store the feature variables in a dataframe named X and the target variable in a dataframe named y.
2. Split the data into training and testing sets using train_test_split with test_size = 0.2 and random_state = 42.
3. Simple Linear Regression:

- Choose 'MedInc' as the independent variable.
- Create new dataframes X_simple and X_test_simple that contain only the 'MedInc' column from X_train and X_test respectively.
- Train the simple linear regression model using X_simple and y_train.
- Predict using the testing set and store the predictions in a dataframe named y_pred_simple.
- Print the coefficients and R-squared value of the model.
- Create a scatter plot to visualize the relationship between 'MedInc' and 'MedianHouseValue'.

5. Multiple Linear Regression:

- Choose the independent variables 'MedInc', 'HouseAge', 'AveRooms', 'AveBedrms', 'Population', 'AveOccup', 'Latitude' and 'Longitude'.
- Create new dataframes X_multi and X_test_multi that contain only the chosen independent variables from X_train and X_test respectively.
- Train the multiple linear regression model using X_multi and y_train.
- Predict using the testing set and store the predictions in a dataframe named y_pred_multi.

- Print the coefficients and R-squared value of the model.
- Create a scatter plot to visualize the relationship between the independent variables and 'MedianHouseValue'.

7. Compare the performance of the two models by evaluating the R-squared value and mean squared error of each model and discuss the effect of including multiple independent variables on the model.
8. Write a conclusion summarizing the results of the lab and discussing the difference between simple and multiple linear regression models.
9. What are the pros and cons of using a simple linear regression model?

Submission Guidelines:

- Submit a Jupyter notebook with the complete code.
- 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:

- Make sure to comment your code and add appropriate titles and labels to your plots.
- Remember to use the appropriate metrics for regression problems such as R-squared and mean squared error.
- For the multiple regression, you can use the corr() function to check the correlation between the independent variables and the target variable.
- Make sure to interpret the results of the model and explain what they mean.
- calculate the mean squared error and R-squared score using sklearn's mean_squared_error and r2_score functions