ASSIGNMENT A1

Title: Linear Regression

Problem Statement: The following table shows the results of a recently conducted study on the correlation of the number of hours spent driving with the risk of developing acute backache. Find the equation of the best fit line

Hours spent driving	10	9	2	15	10	16	11	16
Risk score (0-100)	95	80	10	50	45	98	38	93

Objective: To implement linear regression and analyze the given data

Outcomes: Students will be able to
- Understand how to find the correlation between two variables
- Calculate accuracy of the linear model and plot graph using Matplotlib

Requirements: python3, pandas, numpy, jupyter

Theory:

Simple Linear Regression: Linear regression assumes a linear or a straight line relationship between the input variables (x) and output variable (y).
- The relationship between the variables and coefficients can be calculated using statistics, and these coefficients can then be used to predict the dependent variable, when given the independent variable as input.
- This can be represented by the equation $y = b_0 + b_1 x$, where b_0 and b_1 are the coefficients.

- The statistical properties of the given distribution are calculated (mean, variance and covariance).

$$b_1 = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{n}(x_i - \bar{x})^2} \qquad b_0 = \bar{y} - b_1\bar{x}$$

where n is the number of datapoints

- Above can be re-written as a representation of covariance and variance as follows:

$$b_1 = \frac{\text{covariance}(x, y)}{\text{variance}(x)}$$

- The error metric used is the RMSE (root mean square error) It is the standard deviation of the residuals (prediction errors.) Residuals are a measure of how far from the regression line the data points are.

Regression analysis is a form of predictive modelling. The given data ~~is~~ could be best used to predict the dependent variable (risk score of backache) using simple linear regression, or a best-fit straight line.

Types of regression:
Other forms of regression analysis are logistic (probability of event occurrence / not, mainly used in classification problems,), polynomial (best-fit line is a curve), ridge (independent variables are highly correlated)

Algorithm
1. Calculate mean of features
2. Calculate covariance and variance of training data

3. Calculate coefficients using values obtained in steps 1 and 2
4. Predict values in test set using the calculated coefficients
5. Calculate error

Test Cases

Test case (Driving hours)	Expected risk score	Actual risk score
11	63.05	63.05
3	26.35	26.35
8	49.29	49.29
15	81.40	81.40
2	21.76	21.76

Result

Coefficients : $b_0 = 12.585$ $b_1 = 4.588$ RMSE : 25.54

Conclusion : Successfully performed linear regression on the given data and predicted values based on the best-fit line found.