**ASSIGNMENT: - 06**

**Problem Statement: -**

Download the temperatures dataset from following link:

https://www.kaggle.com/datasets/venky73/temperatures-of-india.

This data consists of temperatures of INDIA averaging the temperatures of all places month-wise.

Temperature values are recorded in CELSIUS.

a) Apply Linear Regression using a suitable library function and predict the Month-wise temperature.

b) Assess the performance of regression models using MSE, MAE and R-Square metrics

c) Visualize a simple regression model.**S/W, Library and Package:**

1. Software: Python
2. Library: scikit-learn (sklearn) - for linear regression model and metrics calculation
3. Package: pandas - for data manipulation and preprocessing

**Theory:**

Methodology:

Linear regression is a statistical technique used to model the relationship between a dependent variable (target) and one or more independent variables (features). The methodology involves finding the best-fit linear equation that represents the relationship between variables by minimizing the sum of squared differences between observed and predicted values.

Advantages:

Interpretability: Provides interpretable results, as coefficients represent the impact of features on the target variable.

Simplicity: Easy to implement and understand, making it suitable for initial data analysis and baseline modeling.

Efficiency: Computationally efficient, allowing quick predictions even with large datasets.

Versatility: Can be used for both regression (predicting continuous values) and classification (logistic regression for binary classification).

Applications:

Predictive Modeling: Used for predicting sales, stock prices, housing prices, etc.

Risk Assessment: Predicting credit risk, insurance claims, and medical outcomes.

Marketing Analysis: Analyzing the impact of marketing campaigns on sales.

Resource Planning: Forecasting demand for resources like electricity, water, etc.

Limitations:

Assumption of Linearity: Linear regression assumes a linear relationship between variables, which may not always hold true.

Sensitivity to Outliers: Outliers can significantly affect the model's performance.

Overfitting/Underfitting: Can suffer from overfitting (too complex model) or underfitting (too simple model) issues.

Multicollinearity: High correlation among independent variables can lead to unstable coefficients.

Example:

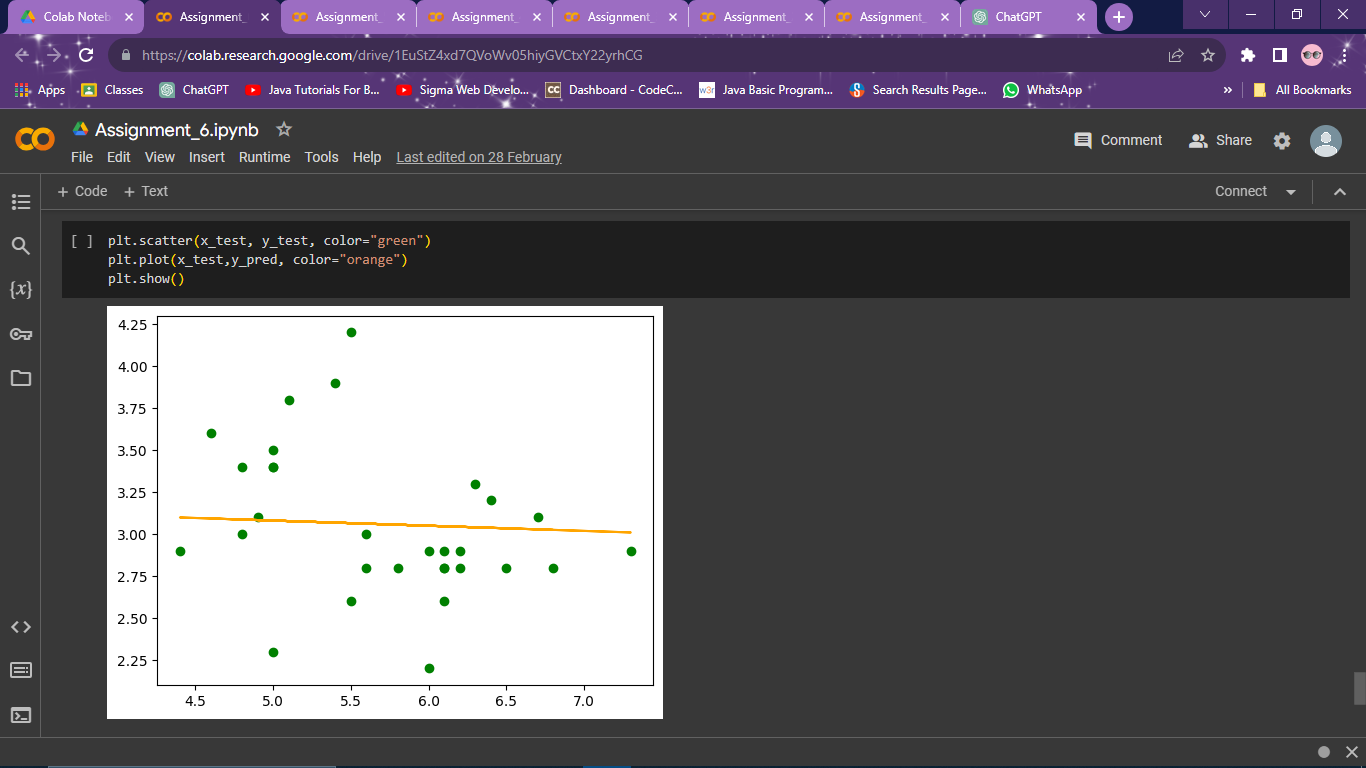
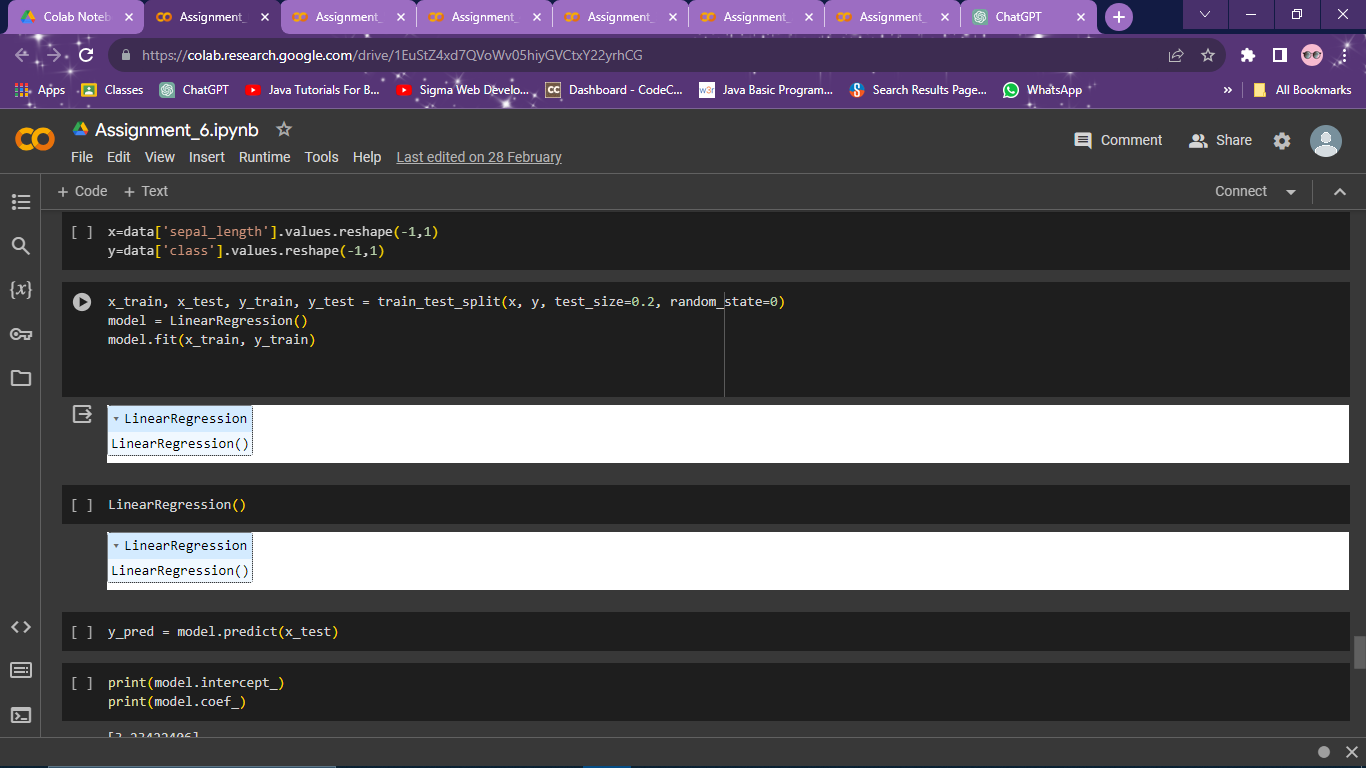
In the provided example, we applied linear regression to predict month-wise temperatures based on the given dataset. The methodology involved splitting the data into training and testing sets, fitting a linear regression model, predicting temperatures, and assessing model performance using metrics such as MSE, MAE, and R-Square. Visualizing the regression model helped understand the relationship between the month and temperature, showcasing the simplicity and interpretability of linear regression.

**Working/ Algorithm:**

Linear regression is a fundamental technique in statistics and machine learning used to model the relationship between a dependent variable and one or more independent variables. The goal is to create a linear model that predicts the dependent variable based on the independent variables.

Here's how a linear regression model works for prediction:

* Importing necessary libraries and modules: They provide pre-built functionalities and extend your program's capabilities
* Data Collection: Collect data on the variables of interest. For example, in a simple linear regression, you would have one independent variable (for eg. Year here) and one dependent variable (for eg. Temperature here).
* Data Preprocessing and EDA: This step involves cleaning the data and analysing it intricately.
* Splitting the Data: Split the dataset into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance.
* Model Training: Use the training data to fit a linear regression model. The model tries to find the best-fitting linear relationship between the independent and dependent variables. In simple linear regression, this relationship is represented by a line (y = mx + b), where m is the slope and b is the intercept.
* Making Predictions: Once the model is trained, use it to make predictions on the testing data. The model calculates the predicted values of the dependent variable based on the values of the independent variable(s).
* Evaluating the Model: Evaluate the model's performance using metrics such as mean squared error (MSE) or R-squared. These metrics measure how well the model's predictions match the actual values in the testing data.

****

**Conclusion:**

Linear regression is a versatile and widely used statistical technique with several advantages, such as interpretability, simplicity, and efficiency. It finds applications in predictive modeling, risk assessment, marketing analysis, and resource planning. However, it has limitations like the assumption of linearity, sensitivity to outliers, and potential overfitting/underfitting issues. Overall, linear regression provides valuable insights and predictive capabilities but should be used judiciously considering its assumptions and limitations.