**Assignment – 6**

**Problem Statement:** Assignment on Regression technique.

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 Packages and H/W apparatus used**:

Software used:

1. Python 3.x
2. Google Colab

Libraries and packages used:NumPy, Matplotlib, scikit-learn

**Theory:**

To understand regression techniques in the context of temperature prediction using linear regression, let's delve into some key concepts:

**1. Linear Regression:** - Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). - It assumes a linear relationship between the independent and dependent variables, represented by the equation: \( y = mx + c \), where \( y \) is the dependent variable, \( x \) is the independent variable, \( m \) is the slope (coefficient), and \( c \) is the intercept. - Linear regression aims to find the best-fitting line that minimizes the sum of squared differences between the observed and predicted values.

**2. Mean Squared Error (MSE):** - MSE is a common metric used to assess the accuracy of regression models. - It calculates the average squared difference between the actual and predicted values.

- Lower MSE values indicate better model performance.

**3. Mean Absolute Error (MAE):** - MAE measures the average absolute difference between the actual and predicted values. - It provides a more interpretable estimate of error compared to MSE. - Like MSE, lower MAE values indicate better model performance.

**4. R-Squared (Coefficient of Determination):** - R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables. - It ranges from 0 to 1, with higher values indicating a better fit of the model to the data. - An R-squared value of 1 indicates a perfect fit, while a value of 0 indicates that the model does not explain any of the variance in the dependent variable.

In this assignment:

- We will use the temperatures dataset to predict month-wise temperatures using linear regression.

- We'll assess the performance of the regression model using MSE, MAE, and R-squared metrics to evaluate its accuracy.

- Finally, we'll visualize the simple regression model to understand the relationship between the independent and dependent variables graphically.

By applying linear regression and evaluating its performance metrics, we can gain insights into the effectiveness of the model in predicting month-wise temperatures and its overall accuracy.

Linear Regression Types:

Basic Linear Regression:   
A linear regression procedure is referred to as simple linear regression if it uses one independent variable to predict the value of a number of dependent variables.

Multiple Linear Regression: This type of linear regression method is employed when multiple independent variables are combined to predict the value of a numerical dependent variable.

the steps to apply linear regression and assess its performance using MSE, MAE, and R-squared metrics:

**Data Preparation:** Download the temperatures dataset from the provided link.

Load the dataset into a dataframe using a suitable library like Pandas.

Explore the dataset to understand its structure, features, and any missing values.

Feature Selection: Identify the independent variable (feature) and dependent variable (target) for regression.

In this case, the independent variable could be the month, and the dependent variable could be the temperature.

**Data Splitting:**

Split the dataset into training and testing sets.

Typically, around 70-80% of the data is used for training and the remaining for testing.

**Model Training:** Import the Linear Regression model from a suitable library like Scikit-learn.Fit the model to the training data to learn the relationship between the independent and dependent variables.

**Model Prediction:**

Use the trained model to make predictions on the testing data.

**Model Evaluation:**

Calculate the Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared metrics to assess the performance of the regression model.

MSE and MAE can be calculated using functions provided by libraries like Scikit-learn or NumPy.

R-squared can be calculated using the r2\_score function from Scikit-learn.

Visualize the Model:

Plot the actual vs. predicted values to visualize how well the model fits the data.

Additionally, you can plot the regression line to see the relationship between the independent and dependent variables.

**Applications:**

1. Student grades determined by the number of hours studied (ideally):In this case, exam scores are dependent on the number of hours studied, but the number of hours studied is independent.
2. Estimating agricultural yields using rainfall data: The measure of precipitation is an independent variable, and yield is a dependent variable.
3. Estimating an individual's salary based on years of experience: Experience is now the independent variable, and salary is the dependent variable.

**Limitations:**

1. Assumes linearity: Regression models, particularly linear regression, assume a straight-line relationship between the independent and dependent variables. If the underlying relationship is more complex (curved, exponential, etc.), the model may not accurately capture the true association.
2. Sensitive to outliers: Outliers (data points significantly different from the majority) can disproportionately influence the regression line, leading to misleading results.
3. Doesn't establish causation: Even if a strong correlation is found between variables, regression models cannot determine causality. There might be a third, unseen variable influencing both the independent and dependent variables, creating a false association.

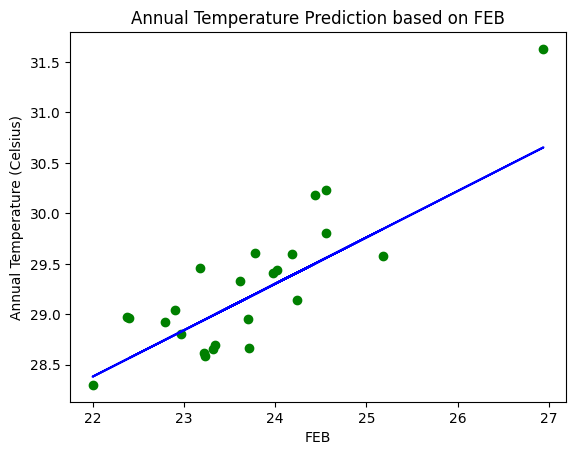
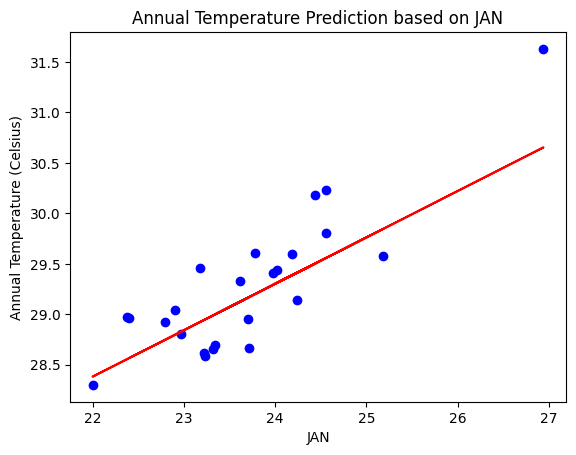
**Working:**

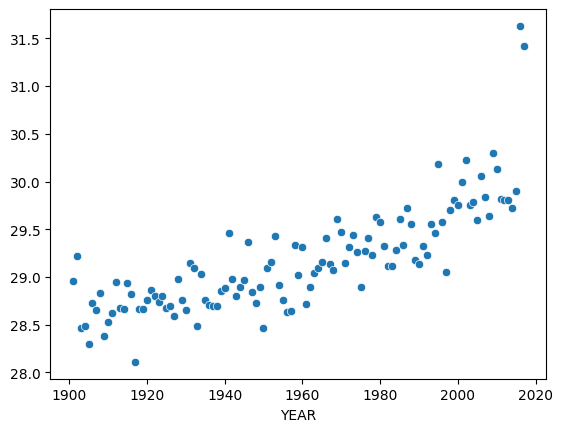
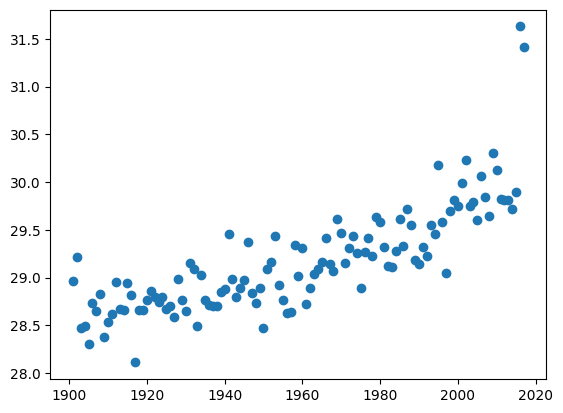
Linear regression is a fundamental technique in statistics and machine learning used to model the relationship between a dependent variable and one or more independent variables. The goal is to create a linear model that predicts the dependent variable based on the independent variables.

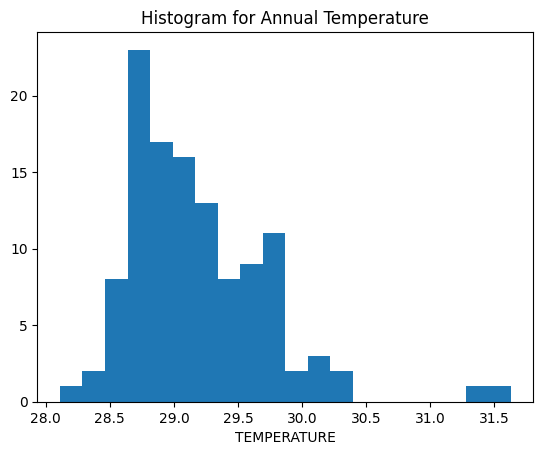
Here's how a linear regression model works for prediction:

* 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.

**Diagram:**







**Conclusion:** In summary, a simple linear regression model uses a straight line to determine the association between one independent variable and one dependent variable.