**Name: Sumit Avinash Solanke**

**Roll No: TYITB121**

**Assignment No. 2**

**AIM:** Assignment on Linear Regression on Employee\_Salary\_Dataset.csv

**PREREQUISITE:** Python programming

**THEORY:**

Linear Regression is a fundamental statistical method used in machine learning for predicting an output variable based on one or more input variables. In its simplest form, when only one independent variable is used, it is known as Simple Linear Regression. When multiple independent variables are involved, it becomes Multiple Linear Regression.

**Objective**

The objective of this assignment is to apply linear regression techniques to the *Employee\_Salary\_Dataset.csv* in order to develop a predictive model for estimating employee salaries based on relevant features such as experience, education level, job role, and department. This analysis aims to understand the relationship between these variables and salary, enabling data-driven salary predictions and insights into key salary influencers.

**Simple Linear Regression**

Simple Linear Regression aims to establish a linear relationship between an independent variable (input) and a dependent variable (output). The relationship can be represented by a straight-line equation where one variable influences the other.

In this model:

* The dependent variable is what we want to predict.
* The independent variable is what we use to make predictions.
* The intercept is the constant term that determines the starting point of the line.
* The slope defines the degree of impact the independent variable has on the dependent variable.

The objective of linear regression is to find the best-fitting line that minimizes the difference between the actual values and the predicted values.

**Estimating Coefficients**

The coefficients in the linear regression model are determined using statistical calculations based on the dataset. The slope is estimated by assessing how much the dependent variable changes with respect to the independent variable. Once the slope is obtained, the intercept can be computed easily using the average values of the dataset.

**Making Predictions**

After determining the coefficients, we can use the linear regression equation to predict values for new input data. By substituting different values into the equation, we can estimate the expected output values. A graphical representation helps visualize how well the model fits the data points.

**Assessing Model Performance**

The accuracy of a linear regression model is often evaluated using the Root Mean Squared Error (RMSE). RMSE measures the average magnitude of prediction errors and provides insight into how well the model generalizes to new data. A lower RMSE indicates better performance.

**Importance of Linear Regression**

Linear Regression is widely used in various fields such as economics, finance, healthcare, and engineering. It helps in forecasting trends, identifying relationships, and making data-driven decisions. The simplicity of the model makes it a great starting point for understanding machine learning concepts.

**Dataset Description:**

**Dataset Name :**Employee\_Salary\_Dataset.csv  
**Total Records:** *(Example: 100 entries — you can update this based on your actual data)*  
**Purpose:** To analyze the relationship between employee attributes like experience, age, gender, and their corresponding salary. This dataset is suitable for salary prediction and trend analysis.

**Columns and Their Descriptions:**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| ID | Integer | Unique identifier for each employee |
| Experience\_Years | Integer | Total number of years of professional experience |
| Age | Integer | Current age of the employee in years |
| Gender | Categorical | Gender of the employee (Male, Female, or other values if present) |
| Salary | Float | Current annual salary of the employee in INR or USD (based on your dataset) |

**Sample Records (Preview):**

| **ID** | **Experience\_Years** | **Age** | **Gender** | **Salary** |
| --- | --- | --- | --- | --- |
| 1 | 3 | 25 | Male | 450000.00 |
| 2 | 7 | 32 | Female | 850000.00 |
| 3 | 1 | 22 | Male | 300000.00 |
| 4 | 10 | 40 | Female | 1200000.00 |
| 5 | 5 | 29 | Male | 650000.00 |

**Possible Exploratory Questions:**

* What is the average salary across the company?
* Is there a strong correlation between Experience\_Years and Salary?
* Do older employees tend to have higher salaries?
* What is the salary distribution by Gender?
* Are there any outliers in salary based on experience?
* What is the most common age group in the dataset?

**Preprocessing Checks During EDA:**

* Checking for **missing values** (e.g., null Salary or Age)
* Validating **data types** (ensuring no string values in Experience\_Years)
* Ensuring **unique IDs** for all entries
* Checking for **outliers** in Salary and Age columns
* Verifying **logical consistency** (e.g., Age should be greater than or equal to Experience\_Years)

**CONCLUSION:**

Linear Regression is a fundamental technique in predictive modeling that establishes a relationship between variables. By understanding and applying this method, we can make informed predictions and analyze patterns in data. This method serves as a building block for more advanced machine learning algorithms and real-world applications.