Assignment 6

Problem statement

Assignment on Regression technique. Download temperature data from the below link. https://www.kaggle.com/venky73/temperaturesof-india?select=temperatures.csv 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:

OS: **Ubuntu/Windows,**
Tool : Google Colab
Packages: **Numpy, Pandas, Matplotlib and SkLearn**

Theory :

1. Methodology:

1.   **Linear Regression :**
     a. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.
     b. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.
2.   **Types of Linear Regression:**

a. Simple Linear Regression:

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

b. Multiple Linear regression:

If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

3. Assumptions of Linear Regression:

To conduct a simple linear regression, one has to make certain assumptions about the data. This is because it is a parametric test. The assumptions used while performing a simple linear regression are as follows:

- **Homogeneity of variance (homoscedasticity)-** One of the main predictions in a simple linear regression method is that the size of the error stays constant. This simply means that in the value of the independent variable, the error size never changes significantly.
- **Independence of observations-** All the relationships between the observations are transparent, which means that nothing is hidden, and only valid sampling methods are used during the collection of data.
- **Normality-** There is a normal rate of flow in the data. These three are the assumptions of regression methods.
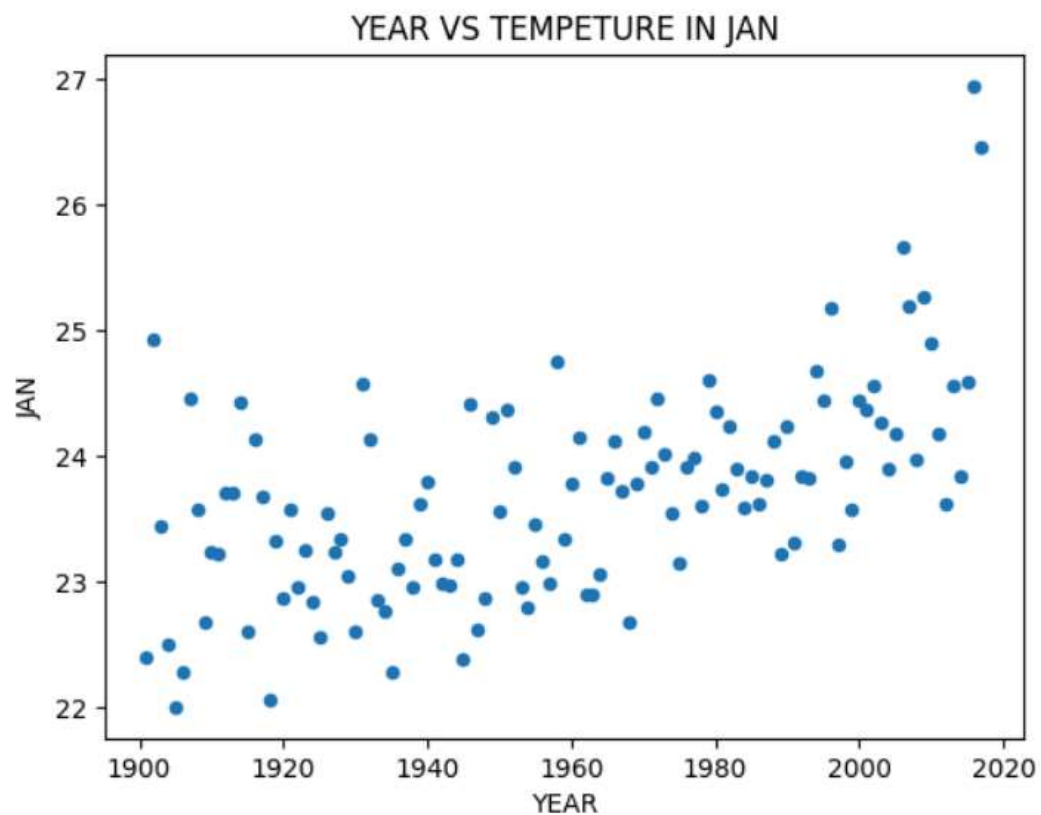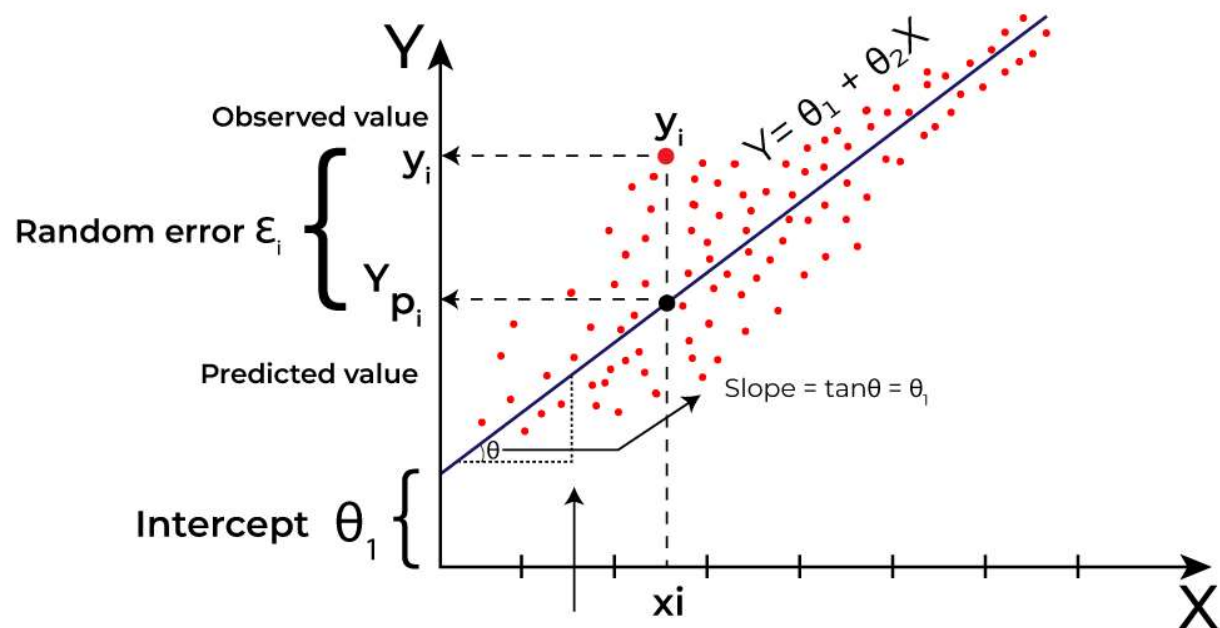
2. Applications of Simple Linear Regression :

1. Marks scored by students based on the number of hours studied (ideally)- Here marks scored in exams are dependent and the number of hours studied is independent.

2. Predicting crop yields based on the amount of rainfall- Yield is a dependent variable while the measure of precipitation is an independent variable.

3. Predicting the Salary of a person based on years of experience- Therefore, Experience becomes the independent variable while Salary turns into the dependent variable.

Working :

1. **Data Collection and Preprocessing:** Gather a dataset containing both independent (features) and dependent (target) variables. Preprocess the data to handle missing values, outliers, and scale features if necessary.
2. **Model Selection**: Choose a regression model suitable for the problem, such as linear regression, polynomial regression, or tree-based regressors.
3. **Model Training:** Split the dataset into training and testing sets. Fit the regression model to the training data, learning the relationship between features and the target variable.
4. **Model Evaluation**: Evaluate the model's performance using metrics like mean squared error (MSE), root mean squared error (RMSE), or R-squared. Assess how well the model generalizes to unseen data using the testing set.
5. **Prediction**: Use the trained model to make predictions on new data by inputting feature values.

Diagram :

Conclusion:

Simple linear regression is a regression model that figures out the relationship between one independent variable and one dependent variable using a straight line.