Point wise explanation of process for developing a multiple linear regression model using the "Car Purchasing Data" dataset:

**Dataset Overview:**

- The dataset, called "Car Purchasing Data," consists of 9 variables:

- Customer Name

- Customer e-mail

- Country

- Gender

- Age

- Annual Salary

- Credit Card Debt

- Net Worth

- Car Purchase Amount

**Data Cleaning and Feature Selection:**

**Removed Variables:**

The first three variables—Customer Name, Customer e-mail, and Country—were removed from the dataset.

**Reason**: These variables do not have a direct or significant relationship with the dependent variable (Car Purchase Amount). They do not contribute to predicting the outcome and are considered irrelevant for the model.

**Selected Variables**:

The independent variables selected for the model are:

- Gender: A categorical variable that represents the gender of the customer.

- Age: A numerical variable that indicates the age of the customer.

- Annual Salary: A numerical variable that represents the customer’s yearly income.

- Credit Card Debt: A numerical variable that shows the total credit card debt of the customer.

- Net Worth: A numerical variable representing the customer’s overall financial worth.

The dependent variable is:

- Car Purchase Amount: The amount spent by the customer on purchasing a car.

**Model Development Using OOP:**

The model was developed using Object-Oriented Programming (OOP) concepts, specifically using classes and objects.

Class and Objects:

A class was created to encapsulate the entire regression model, including data processing, training, and prediction.

Objects of the class were used to execute the model operations such as training the model on the dataset and making predictions.

Purpose of the Model:

The purpose of this multiple linear regression model is to predict the car purchase amount based on the selected independent variables (Gender, Age, Annual Salary, Credit Card Debt, and Net Worth).

The model aims to establish the relationship between these predictors and the dependent variable, allowing accurate predictions of car purchase amounts for new customers.

**Model Overview**

Developed a multiple linear regression model to predict the car purchase amount.

The dataset was split into **80%** training data and **20%** testing data, which is a standard practice to evaluate the model's performance on unseen data.

**Training Metrics Explanation**

**Accuracy (R-squared) of the Training Data**: `0.9999999799259697`

This value represents the proportion of the variance in the car purchase amount that is explained by the model's independent variables during training.

A value very close to 1 indicates an extremely high level of accuracy, meaning the model explains nearly all of the variance in the data. However, this high accuracy may suggest overfitting, especially given how close it is to 1.

**Training Mean Squared Error (MSE):** `2.22`

MSE measures the average of the squared differences between the actual and predicted values.

A low MSE indicates that the model’s predictions are close to the actual values during training.

**Training Mean Absolute Error (MAE):** `1.17`

MAE measures the average of the absolute differences between the actual and predicted values.

A lower MAE value shows that on average, the model’s predictions are very close to the true car purchase amounts.

**Training Root Mean Squared Error (RMSE)**: `1.49`

RMSE is the square root of MSE and provides an estimate of the average prediction error in the same units as the dependent variable (car purchase amount).

A lower RMSE suggests good performance, indicating that the model’s predictions are quite accurate during training.

**Testing Metrics Explanation**

**Accuracy (R-squared) of the Test Data:** `0.9999999799259697`

This accuracy score on the test data is exactly the same as on the training data, suggesting that the model performs equally well on unseen data.

Such high test accuracy further indicates potential overfitting, as it is unusual for a model to perform this perfectly on both training and testing data.

**Testing Mean Squared Error (MSE)**: `1.83`

This is slightly lower than the training MSE, indicating the model's error on the test set is marginally smaller than on the training set.

**Testing Mean Absolute Error (MAE)**: `1.1092`

The MAE for the test data is comparable to that of the training set, suggesting that the average prediction error is consistently low for unseen data.

**Testing Root Mean Squared Error (RMSE)**: `1.35`

RMSE on the test data is slightly lower than the training RMSE, further indicating the model's strong predictive capability even on unseen data.