**Predict Delivery Time Based on Sorting Time**

Project Overview

The project’s ambition is to predict the ‘Delivery Time’ based on ‘Sorting Time’.

Project Steps

1. Importing Libraries and Loading the Data  
Essential Python libraries such as pandas, seaborn, and matplotlib were imported in the Jupyter environment. The dataset was then loaded, and initial data cleaning was performed, including checking for and handling missing values.

2. Exploratory Data Analysis (EDA)  
A detailed EDA was conducted to understand the data distribution and relationships:

* Correlation Heatmap:  
  A heatmap was used to visualize the correlation between 'Delivery Time' and 'Sorting Time', revealing a strong positive correlation of 0.83, suggesting a linear relationship.
* Histogram Analysis:  
  A histogram of 'Delivery Time' was plotted, indicating a slight right skew with a skewness value of approximately 0.33.
* Box Plot Analysis:  
  A box plot was generated to detect outliers. No significant outliers were identified, so no imputation or removal was necessary.

Model Building

The model was trained using both the **full dataset** and different **train-test split ratios**, such as **70:30**, to evaluate its performance. The splitting and training were done using scikit-learn's linear regression implementation.

Results

Among the three linear regression models developed, **Model 3** demonstrated the best overall performance. It achieved the **lowest RMSE of 1.51** and the **highest R² score of 0.5318**, indicating a moderate fit to the data and better predictive accuracy compared to the others.

In contrast, **Model 1** and **Model 2** both had significantly lower R² scores of **-2.0548**, suggesting that these models performed worse than simply predicting the mean of the target variable. Although Model 1 had a slightly lower RMSE (1.6135) than Model 2 (1.9198), the negative R² scores for both indicate poor model performance.

Therefore, **Model 3 is the most reliable and accurate model among the three**, based on both its error metrics and explanatory power.