# Data Cleaning and Preprocessing Notebook

Project Name: UK Train Rides

Author: (CAI2_DAT2_S4)

Date: 25/02/2025

## 1. Introduction

In this phase, data was cleaned and preprocessed using Power BI to ensure its quality before analysis. The cleaning process included handling missing values, correcting data types, and ensuring data consistency.

## 2. Data Loading

The dataset was imported into Power BI using Power Query, and its contents were verified in the Data View.

Data Source: CSV

Number of Records: 31,653 rows

Number of Columns: 18 C

## 3. Data Cleaning Steps

### 3.1 Handling Missing Values

- Missing values were identified using Power Query.
- The following actions were taken:
- Rows (Actual Arrival Time) & (Reason for delay) with missing values in key columns were checked for the cause, as cancellation and on time journeys.

### 3.2 Removing Duplicates

- The **Remove Duplicates** function in Power Query was used to eliminate duplicate records and found that there were no duplicate records.

### 3.3 Data Type Corrections

- Converted (Time of purchase, Departure time, Arrival time, Actual arrival time) columns from text to (Short Time) format also dates columns from text to (Short date) format.

### 3.4 Text Cleaning

- Removed extra spaces from station names using (Trim).

## 4. Data Validation

- Verified that all modifications were applied correctly.
- Exported the cleaned dataset to Power BI Model for further analysis.

## 5. Summary

The data has been successfully cleaned and prepared for analysis, ensuring accuracy and data quality. The next phase involves (Analyzing the dataset and making predictions).