Problem Statement

Delays prove to be very disruptive and frustrating, especially if we often commute by train. So, we have been trying to solve this pain point and devise a solution to get rid of inconvenience caused to the user, predicting the ETA (Estimated time of Arrival) and ETD (Estimated time of Departure) for upcoming stations for next run dates.
This assignment focuses on the same problem of improving online Train Running Status using AI/ML based approaches.

Data

You will be given two data sets namely train.csv and test.csv(unzip .gz files) which will contain information of 15 trains with their stopping stations for different running dates.

Given below are different columns and their description -
1. runDate - start date for the train
2. stations - railway station code
3. trainCode - train code
4. trainStationId - railway station id
5. scheduledArrival - scheduled arrival timestamp of the train at a station
6. scheduledDeparture - scheduled departure timestamp of the train from a station
7. actualArrival - actual arrival timestamp of the train at a station
8. actualDeparture - scheduled departure timestamp of the train from a station
9. distance - distance covered by the train
10. dayCount - number of days the train takes to complete the journey
11. ArrivalDelay - Derived column (actualArrival - scheduledArrival)
12. DepartureDelay - Derived column (actualDeparture - scheduledDeparture)

Goal

- Based on the retrospective data given in the **train.csv**, predict the ETA (Estimated time of Arrival) / Arrival Delay and ETD (Estimated time of Departure) / Departure Delay using AI / ML models.

- We are looking for a clean and well documented markdown file with snippets and plots (wherever needed).
- It would be nice to see the final approach and the intermediate approaches to derive insights.

Deliverables

- Submitting a file with name **candidateName_solution.csv**
- Separate jupyter notebooks for exploratory data analysis (EDA) and modelling (with proper documentation)
- requirements.txt for package dependencies.
- A list of techniques tried, including: evaluation criteria, improvements, the steps taken to explore different algorithms and options, accuracy and precision.