Data Collection and Preprocessing Phase

Date	28 July 2025
Project Title	Flight Delays Prediction Using Machine Learning
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Report:

The Data Quality Report summarizes issues identified in the flightdata.csv dataset from Kaggle. It includes severity levels and proposed resolutions for each issue.

Data Collection Plan:

Section	**Description**
Project Overview	The machine learning project aims to predict flight delays based on flight attributes such as departure and arrival times, distance, and day of travel. Using a dataset sourced from Kaggle, the goal is to develop a model that accurately classifies flights as delayed or on-time, enabling airlines and passengers to better anticipate disruptions and improve operational planning.
Data Collection Plan	• Search for publicly available datasets related to flight schedules and delays. • Prioritize datasets containing essential features such as scheduled and actual times, day-of-week, and delay indicators.
Raw Data Sources Identified	The raw data source for this project is the

	Flight Delay Dataset obtained from Kaggle. It contains flight-level records including details such as airline, scheduled and actual departure/arrival times, delay information, and distance traveled.

Raw Data Sources Report:

Source Name	Description	Location/URL	Format	Size	Access Permissions
Kaggle Dataset	Features include airline, month, day of month, day of week, origin airport, destination airport, scheduled and actual departure times, arrival time, delay indicator, and distance.	https://drive.google.com/file/d/1HNYx6fX5hvRDX43egcAAUsrQ9sccv4AR/view?usp=sharing	CSV	~2.3 MB	Public