Data Collection and Preprocessing Phase

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

Data Quality 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 Quality Report:

Data Source	Data Quality Issue	Severity	Resolution Plan
Kaggle Dataset	Missing values in ArrTime, ActualDepTime columns	Moderate	Impute missing values using mean or remove affected rows.
Kaggle Dataset	Outliers in DepDelay and ArrDelay values (extreme delays > 500 min)	High	Cap extreme outliers or remove unrealistic values.
Kaggle Dataset	Categorical columns: Carrier, Origin, Destination not encoded	Moderate	Apply Label Encoding or One-Hot Encoding.
Kaggle Dataset	Unused or redundant columns like FlightNum	Low	Drop irrelevant columns during preprocessing.

Kaggle Dataset	Inconsistent case formatting in categorical fields (msp, MSP)	Low	Standardize all categorical values to uppercase.