Week 1 Report – Project Initialization & Dataset Setup

Dataset Overview

The dataset `flights_sample_100k.csv` was successfully loaded.
It contains 100,000 records with 32 columns, representing various flight operations including flight dates, airlines, origins, destinations, times, delays, and cancellations.
Memory usage was checked and optimization steps were applied where possible (e.g., converting object columns with low cardinality into categorical types).

Initial Exploration

The dataset includes key columns such as:
- Identifiers: FL_DATE, AIRLINE, AIRLINE_CODE, FL_NUMBER
- Airports & Routes: ORIGIN, DEST, Route (created as ORIGIN–DEST)
- Times: CRS_DEP_TIME, DEP_TIME, ARR_TIME, CRS_ARR_TIME
- Performance Metrics: DEP_DELAY, ARR_DELAY, ELAPSED_TIME, DISTANCE
- Delay Causes: DELAY_DUE_CARRIER, DELAY_DUE_WEATHER, DELAY_DUE_NAS, DELAY_DUE_SECURITY, DELAY_DUE_LATE_AIRCRAFT
- Flags: CANCELLED, DIVERTED

Missing values were identified in several delay-related columns (DEP_DELAY, ARR_DELAY, weather-related delays, etc.), as well as some time fields.

Feature Engineering

1. Date & Time Features
- Converted FL_DATE into proper datetime format.
- Extracted Year, Month, Day, and DayOfWeek.
- Extracted DepHour and ArrHour from scheduled departure and arrival times.

2. Route Information
- Created a new feature Route = ORIGIN + '-' + DEST.

3. Delay Metrics
- Created TotalDelay = DEP_DELAY + ARR_DELAY.
- Created IsDelayed flag = 1 if TotalDelay > 15 minutes, else 0.

Descriptive Analysis

- Airlines & Airports
  * Number of unique airlines was recorded.
  * Top 5 airlines by flight count were identified.
  * Top origin and destination airports were extracted.

- Flights Distribution

  * Flights grouped by Year, Month, DayOfWeek show seasonal and weekly variations.
  * Total flights in the dataset: 100,000.

- Delays
  * Average departure and arrival delays were calculated.
  * Summary statistics were generated for all delay-related columns.
  * Top routes were listed based on frequency.
  * Longest and shortest flights (by distance) were identified.

- Cancellations & Diversions
  * Cancelled and diverted flights were counted.

- Data Quality Checks
  * A check for invalid times (arrival before departure) was performed.
  * A missing value report function was created.

Outputs Generated

• Optimized dataset saved in both Pickle (flights_week1.pkl) and CSV (flights_week1.csv) formats.
• Delay summaries and grouped statistics were generated.
• Final reporting steps (schema, documentation) are reserved for later weeks.

Summary

By the end of Week 1, the dataset has been successfully loaded, inspected, cleaned at a basic level, and enriched with new features (Year, Month, DayOfWeek, DepHour, ArrHour, Route, TotalDelay, IsDelayed). The structure and quality issues (missing values, invalid times) have been identified. The dataset is now ready for Week 2 preprocessing and feature engineering.