AIRLINE FLIGHT DATA

MALLA REDDY ENGINEERING COLLEGE
FOR WOMEN
(TELANGANA- HYDERABAD)

P. THRISHUJUNA
B-TECH - IV YR

AIRLINES

Milestone 1: Data Foundation and Cleaning

Week 1: Project Initialization and Dataset Setup

- Define goals, KPIs, and workflow
- Load CSVs using pandas
- Explore schema, types, size, and nulls
- Perform sampling and memory optimizations

Week 2: Preprocessing and Feature Engineering

- Handle nulls in delay and cancellation columns
- Create derived features: Month, Day of Week, Hour, Route
- Format datetime columns
- Save preprocessed data for fast reuse

Deliverables:

- Cleaned dataset
- Summary of preprocessing logic
- Feature dictionary

- Loads a dataset: **`Flight_delay.csv`** from an airline dataset.
- Uses **Pandas** to explore the data (`df = pd.read_csv(...)`).
- Takes a **10% sample** of the data for testing/analysis (`sampled_df = df.sample(frac=0.1, random_state=42)`).
- Performs **data cleaning**:
  - Fills `NaN` values in delay-related columns (`ArrDelay, DepDelay, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay`) with `0`.
  - Fills `Cancelled` column nulls with `0`.
  - Fills `CancellationCode` nulls with empty string.

Since the dataset is about **flight delays**, here are the main metrics you should calculate:

1. Dataset Overview

- Total rows & columns.
- Missing values count (before/after cleaning).
- Data types & memory usage.

2. Delay Analysis

- **Average Arrival Delay (`ArrDelay`)**.
- **Average Departure Delay (`DepDelay`)**.
- Breakdown of delay causes:
  - Carrier Delay
  - Weather Delay
  - NAS Delay (National Airspace System)
  - Security Delay
  - Late Aircraft Delay

➡️ These can be expressed as **mean delay minutes per flight** and as **percent contribution** to total delays.

3. Cancellation Analysis

- Total number of cancelled flights.
- Percentage of cancelled flights = `(Cancelled flights / Total flights) * 100`.
- Breakdown by **CancellationCode** (e.g., carrier-related, weather, security, NAS).

4. Sampling Impact

- Compare stats (mean delay, cancellation rate) between **full dataset** vs **10% sample**.
- Ensure the sample is representative.

5. KPIs (Key Performance Indicators)

- **On-Time Performance Rate** = % of flights with `DepDelay <= 15 minutes`.
- **Average Delay per Flight** (total delay minutes / total flights).
- **Most common cause of delay**.

Right now, I only have access to the **notebook code**, not the actual dataset (`Flight_delay.csv`). From your notebook, I can see what columns were handled:

- **Delay columns:** `ArrDelay, DepDelay, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay`
- **Cancellation columns:** `Cancelled, CancellationCode`

The notebook fills nulls in those columns, which means they **had missing values** originally.

Here's what you would typically get (once the CSV is loaded):

🔎 Dataset Structure & Cleaning Insights

1. **Shape of the dataset**
   - Number of **rows (flights)** = total records in `Flight_delay.csv`
   - Number of **columns** = flight attributes (delays, cancellations, etc.)
2. **Null Values (Before Cleaning)**
   - `ArrDelay` → had missing values → filled with `0`
   - `DepDelay` → had missing values → filled with `0`
   - `CarrierDelay` → had missing values → filled with `0`
   - `WeatherDelay` → had missing values → filled with `0`
   - `NASDelay` → had missing values → filled with `0`
   - `SecurityDelay` → had missing values → filled with `0`
   - `LateAircraftDelay` → had missing values → filled with `0`
   - `Cancelled` → had missing values → filled with `0`
   - `CancellationCode` → had missing values → filled with `""` (empty string)
3. **Other Columns**
   - Likely include things like flight date, airline, origin, destination, etc. (no cleaning steps shown in the notebook, so probably had few or no nulls).