# Internship Report – Week 1 & 2

**Intern Name:** Abhinand P K
**Internship Duration:** Week 1 & 2
**Domain:** Data Visualization & Data Analysis
**Organization/Platform:** [Infosys Springboard]
**Mentor/Supervisor:** [Swaraj Aalla]

---

## Week 1: Project Initialization and Dataset Setup

### Objectives:

- Define the project goals, key performance indicators (KPIs), and workflow.
- Load and explore the dataset to understand its structure and quality.

### Tasks Completed:

- Defined the overall objectives of the internship project and established KPIs for measuring progress.
- Loaded the Flight_delay.csv dataset using Pandas.
- Explored the dataset thoroughly by:
    - Examining the first and last rows (head(), tail()).
    - Checking data types, column names, and dataset shape.
    - Generating summary statistics using describe() and inspecting detailed information with info().
- Identified and removed duplicate rows to ensure data integrity.

- Performed random sampling for initial analysis:
    - 1% random sample and first 100,000 rows of the dataset.
- Optimized memory usage for large dataset processing by:
    - Downcasting numeric columns to appropriate smaller data types.
    - Converting object-type columns to categorical type.
- Verified memory optimization by comparing memory usage before and after changes.

### Skills Acquired:

- Dataset exploration and schema understanding.
- Memory-efficient handling of large datasets.
- Practical use of Pandas for initial data exploration and sampling.

---

## Week 2: Preprocessing and Feature Engineering

### Objectives:

- Clean and preprocess the dataset to prepare it for analysis and visualization.
- Derive meaningful features to enable advanced data analysis.

### Tasks Completed:

- Handling Missing Values:

- Filled missing values in categorical columns (Date, UniqueCarrier, Airline, TailNum, Origin, Org_Airport, Dest, Dest_Airport,CancellationCode.) with "Unknown".
- Replaced missing numeric delay values (ArrDelay, DepDelay, etc.) with 0.
- Filled missing values in elapsed time and airtime columns with the median.
- Updated CancellationCode to "NotCancelled" for flights that were not cancelled.
- Removed any remaining rows with missing values to ensure a clean dataset.

- Feature Engineering:
  - Converted the Date column to datetime format.
  - Extracted departure hour and minute from DepTime and created a unified DepDatetime column.
  - Derived new features for enhanced analysis:
    - Month – month of departure (1–12).
    - DayOfWeek – name of the weekday (Monday, Tuesday, etc.).
    - Hour – hour of departure.
    - Route – combination of Origin and Dest airports (e.g., ATL-LAX).
- Preprocessed Data Export:

- Saved the cleaned and feature-enriched dataset as Flight_delay_cleaned.csv for efficient reuse in subsequent analysis.

## Skills Acquired:

- Advanced data cleaning and preprocessing using Pandas.
- Datetime manipulation and feature derivation for data analysis.
- Preparing datasets for efficient visualization and analysis workflows.