# Flight Delay Analysis – Data Foundation & Cleaning (Milestone 1)

## 1. Introduction

The dataset used in this project is the Flight Delay and Causes dataset from Kaggle. It contains detailed information about flight schedules, departure/arrival times, delays, and cancellation reasons.  
  
Objective of the Project:  
- Build a strong foundation for analyzing flight delays.  
- Clean and preprocess the raw dataset for efficient usage.  
- Create derived features to support further visualization and analysis.  
- Save a reusable, preprocessed dataset for downstream tasks.

## 2. Week 1: Project Initialization and Dataset Setup

### Step 1: Define Goals, KPIs, and Workflow

Goals: Understand and prepare the dataset for visualization & analysis.  
KPIs:  
 - Percentage of delayed flights  
 - Average delay duration  
 - On-time performance rate  
 - Most common delay causes  
  
Workflow: Dataset ingestion → Cleaning → Feature engineering → Storage.

### Step 2: Load CSVs using Pandas

Used pandas.read\_csv() to load the dataset into a Pandas DataFrame. Verified dataset size, data types, and first few records.

### Step 3: Explore Schema, Types, Size, and Nulls

Checked number of rows and columns (.shape).  
Printed column data types (.dtypes).  
Identified null values using .isnull().sum().

### Step 4: Perform Sampling and Memory Optimizations

Used .sample(5) to preview random rows.  
Converted large integer columns to smaller dtypes (int32, int16) for efficiency.  
Converted categorical fields (like UniqueCarrier) into category type.

## 3. Week 2: Preprocessing and Feature Engineering

### Step 1: Handle Nulls

Filled missing delay and cancellation values with 0.  
Dropped unnecessary columns where applicable.

### Step 2: Create Derived Features

Month: extracted from Date.  
Day of Week: already present, validated.  
Hour: extracted from DepTime.  
Route: created by combining origin and destination airports.

### Step 3: Format Datetime Columns

Converted Date column into datetime format for consistency.

### Step 4: Save Preprocessed Data

Saved cleaned dataset for fast reuse as CSV and Parquet.

## 4. Deliverables

- Cleaned dataset saved as CSV & Parquet.  
 - Summary of preprocessing logic documented step by step.  
 - Feature Dictionary:

|  |  |
| --- | --- |
| Feature | Description |
| Date | Flight date |
| DepTime | Actual departure time (HHMM) |
| ArrTime | Actual arrival time (HHMM) |
| DepDelay | Departure delay (minutes) |
| ArrDelay | Arrival delay (minutes) |
| UniqueCarrier | Airline code |
| Route | Origin-Destination pair |
| Month | Extracted month from date |
| Hour | Departure hour extracted from DepTime |
| DayOfWeek | Day of week (0=Mon … 6=Sun) |

# Milestone 2: Visual Exploration and Delay Trends

## Week 3: Univariate and Bivariate Visual Analysis

The objective of Week 3 is to visually explore the cleaned dataset. This includes analyzing top airlines, busiest routes, flight distribution by month, day, and time, to understand overall flight patterns and trends.

### Observations:

• Southwest Airlines operates the highest number of flights, followed by American Airlines and United Airlines.

• The busiest routes are between major airports such as ATL–ORD and LAX–SFO, indicating heavy passenger traffic.

• Flights peak during summer months (June–August) due to vacation travel demand.

• Fridays record the highest number of flights, while most flights depart between 6 AM and 6 PM.

## Week 4: Delay Analysis – Airline and Weather

The objective of Week 4 is to study delay patterns among airlines and understand the influence of factors such as carrier issues, weather, National Airspace System (NAS), and airport congestion on overall delay trends.

### Observations:

• American Airlines and United Airlines showed higher carrier-related delays compared to others.

• Evening flights (after 6 PM) experience more delays than morning flights.

• Major hub airports face higher delays due to increased traffic and weather disruptions.

• Most weather delays last under 50 minutes, though a few severe cases exceed 200 minutes.

## Summary of Milestone 2

|  |  |
| --- | --- |
| Aspect Analyzed | Key Insights |
| Top Airlines | Southwest Airlines operates the highest number of flights. |
| Busiest Routes | High frequency between major cities like ATL–ORD and LAX–SFO. |
| Seasonal Trend | Peak flight activity occurs during June–August. |
| Day/Time Trends | Fridays and early afternoon hours record the most departures. |
| Delays by Airline | American and United Airlines show higher carrier delays. |
| Weather Impact | Weather-related delays are relatively rare but can be severe. |
| Peak Delay Period | Evening flights and busy airports have longer average delays. |