AIRLINE FLIGHT DATA

MALLA REDDY ENGINEERING COLLEGE

FOR WOMEN

(TELANGANA- HYDERABAD)

1. **NAVANEETHA**

**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

INTRODUCTION

The **AirFly Insights** project focuses on analyzing large-scale airline flight data to uncover operational trends, delay patterns, and cancellation reasons. By leveraging data visualization and exploratory data analysis (EDA) techniques, the project aims to provide actionable insights that can help airlines, airports, and analysts improve decision-making, optimize flight schedules, and enhance passenger experience.

The dataset used for this project contains over **60 million flight records**, covering details such as flight dates, times, delays, cancellations, routes, and carriers. Using **pandas** for preprocessing and **visualization libraries** such as Matplotlib, Seaborn, Plotly, and Folium, we performed a thorough analysis of flight operations.

This document summarizes the data cleaning steps, metrics generated, and insights discovered during the project. It also highlights deliverables that can serve as a foundation for building interactive dashboards and future predictive models.

# 1. Project Goals

• Understand and explore the Kaggle airline dataset.

• Perform data cleaning by handling missing values and correcting data types.

• Create derived features such as Month, Day of Week, Hour, and Route to enable deeper analysis.

• Save the cleaned dataset for fast and reusable access in further milestones.

• Document the data preprocessing logic clearly for reproducibility.

# 2. Key Performance Indicators (KPIs)

• 100% of missing values in delay and cancellation columns handled.

• Optimized dataset size by converting columns to appropriate data types.

• Created at least 4 derived features (Month, Day of Week, Hour, Route).

• Generated a feature dictionary describing all columns and their data types.

• Cleaned dataset saved as Parquet file and verified for accuracy.

# 3. Workflow

• Load the raw dataset into Databricks using pandas.

• Inspect data shape, schema, column types, and null values.

• Perform memory optimization by converting data types.

• Handle missing values in delay and cancellation columns.

• Feature engineering: extract Month, Day of Week, Hour, and Route columns.

• Remove duplicates and invalid records if any.

• Save cleaned data as a parquet/CSV file for reuse in later analysis.

• Document all steps in Jupyter Notebook/Word file for tracking.

DATA CLEANING AND PREPROCESSING

During **Milestone 1**, the following steps were performed to clean and prepare the Kaggle airline dataset for analysis:

1. **Loading the Dataset**

Used pandas.read\_csv() to load multiple CSV files containing flight data into a single DataFrame.

Combined datasets for different years (if applicable) using pd.concat().

1. **Exploratory Checks**

Examined dataset shape, schema, and column data types using:

df.shape, df.info(), df.describe()

Checked for missing values using df.isnull().sum() to identify columns needing cleaning.

1. **Handling Missing Values**

**Delay columns** (CarrierDelay, WeatherDelay, etc.) were filled with 0 since a missing value implies no delay.

**Cancellation codes** were replaced with "Not Cancelled" to avoid null entries.

Dropped rows with critical missing values (if Origin or Dest was missing).

1. **Data Type Conversion & Memory Optimization**

Converted numeric columns from int64 → int32 and float64 → float32 to reduce memory usage.

Converted string/categorical columns (e.g., UniqueCarrier, Origin, Dest) to category dtype.

1. **Datetime Parsing**

Converted FlightDate into a proper datetime format:

df['FlightDate'] = pd.to\_datetime(df['FlightDate'])

1. **Feature Engineering**

Created derived columns for better analysis:

**Month:** df['Month'] = df['FlightDate'].dt.month

**Day of Week:** df['DayOfWeek'] = df['FlightDate'].dt.dayofweek

**Hour:** Extracted hour from DepTime.

**Route:** Concatenated Origin and Dest into a new column Route.

1. **Duplicate Removal**

Checked and removed duplicate rows using df.drop\_duplicates().

1. **Saving Cleaned Data**

Saved the cleaned dataset into **Parquet format** for faster reads in subsequent analysis:

df.to\_parquet('cleaned\_airline\_data.parquet', index=False)

## 1. Dataset Overview

The airline dataset was explored to study flight schedules, delays, and cancellation behavior.

* **Raw Dataset:** 484,551 rows × 29 columns
* **After Cleaning & Feature Engineering:** 484,549 rows × 33 columns
* **Total Records Processed:** 14+ million data points

## 2. Data Cleaning with Pandas

### 2.1 Duplicate Handling

* Identified **2 duplicate rows** in the raw dataset.
* After cleaning, **all duplicates were removed**, leaving 484,549 valid records.

### 2.2 Missing Values

* **Origin Airport:** 1,177 nulls
* **Destination Airport:** 1,479 nulls

**Fixes Applied:**

Missing airport codes filled logically.

Missing ArrTime values replaced using CRSArrTime.

Date column converted to datetime and missing values handled via forward/backward fill

### 2.3 New Features Created

* **Month**
* **Day of Week (both number & name)**
* **Hour of Departure**
* **Route (Origin–Destination combination)**

### 2.4 Data Types Optimization

* Converted numeric columns (ArrDelay, DepDelay, Distance, etc.) into integer/float types.
* Converted text-based columns (Origin, Dest, CancellationCode) into categorical types to save memory.

## 3. Distance Statistics

* **Minimum Distance:** 31 miles
* **Maximum Distance:** 4,502 miles
* **Average Distance:** 752.14 miles
* **Segmented long-haul flights** (>1,000 miles) for separate analysis

## 4. Metrics & Insights

**Dataset Shape:** Before = (484,551 × 29), After = (484,549 × 33)

**Day of Week Flight Count:**

Mon: 70,254

Tue: 65,934

Wed: 63,055

Thu: 75,011

Fri: 88,972 (highest)

Sat: 51,330

Sun: 69,995

**Delays:**

* Southwest: Avg ArrDelay = 8.5 min, DepDelay = 7.2 min
* Delta: Avg ArrDelay = 10.3 min, DepDelay = 9.0 min
* Carrier & NAS delays most frequent
* Weather delays less frequent but more severe
* Late Aircraft delays are the major cause of cascading delays

**Flight Patterns:**

* Busy hours: 06:00–10:00 & 16:00–20:00
* Lowest activity: 01:00–05:00
* Popular routes: IND–BWI, ISP–BWI, IND–LAS

**Taxi Times:**

* Avg Taxi In = 6.78 min
* Avg Taxi Out = 19.15 min → suggests congestion at major airports