Milestone 1 Report — Data Cleaning and Feature Engineering

Project Title: *AirFly Insights: Data Visualization and Analysis of Airline Operations*
Intern Name: Sarthak Mokal
Organization: *Infosys – Internship Program (Data Analytics & Visualization)*
Milestone 1 Duration: *Week 1 – Week 2*

---

1. Introduction

This project, *AirFly Insights*, aims to analyze large-scale airline flight data to uncover operational trends, delay patterns, and cancellation reasons through data visualization and analytics.
Milestone 1 focused on preparing a clean, consistent, and analysis-ready dataset for further visual exploration and modeling in subsequent milestones.

---

2. Dataset Overview

- **Source:** Kaggle – US Domestic Airlines Flights Data
- **Records Loaded:** 484,551 rows
- **Columns (before cleaning):** 35
- **Columns (after cleaning and feature engineering):** 44
- **Storage Path:** /Volumes/workspace/default/airlines/Flight_delay_cleaned_final.csv

The dataset contains flight-level details such as flight date, airline carrier, origin and destination airports, scheduled departure/arrival times, delays (caused by carrier, weather, NAS, security, late aircraft), and cancellation codes.

---

3. Key Performance Indicators (KPIs)

To guide the later analysis and dashboard creation, the following KPIs were defined:

KPI	Description
Average Departure Delay	Mean delay (minutes) per flight
Average Arrival Delay	Arrival delay across carriers
Total Cancelled Flights	Count and percentage of cancelled flights
On-Time Performance Rate	% of flights with delay ≤ 15 minutes
Top Busiest Routes	Flights per Origin–Destination pair
Peak Hour and Day	Flight volume distribution by hour and day of week

---

4. Objectives and Tasks Completed

Milestone 1 objectives were to achieve full data readiness through systematic cleaning and feature engineering.

Task Category	Description of Work Completed
Data Acquisition	Imported CSV file using pandas; verified schema and memory footprint.
Null Value Treatment	Replaced missing delay values with 0 (while retaining audit flags); handled empty cancellation codes.
Datetime Formatting	Parsed and validated DepDatetime; no invalid timestamps (NaT = 0).
Feature Engineering	Created derived columns — Month, DayOfWeek, DayName, DepHour, DepMinute, DepDate, and Route (ORIGIN–DEST).
Data Type Optimization	Downcast numeric types for memory efficiency.
Duplicate Handling	Removed 2 exact duplicates to ensure unique records.

Task Category	Description of Work Completed
Output Generation	Saved final cleaned dataset as Flight_delay_cleaned_final.csv.

---

5. Methodology

The following workflow was implemented:

1. **Data Loading & Inspection** – Checked datatypes, nulls, and memory usage.
2. **Column Standardization** – Renamed and trimmed columns for consistency.
3. **Datetime Engineering** – Unified DepDatetime field and derived temporal features.
4. **Feature Creation** – Added aggregatable dimensions (Month, DayOfWeek, Route).
5. **Data Validation** – Verified record counts and checked for parsing errors.
6. **Optimization & Saving** – Downcast numeric columns and saved to workspace.

All operations were performed in **Databricks (Python environment)** using pandas and numpy.

---

6. Insights and Observations

- Dataset is fully consistent with no broken timestamps.
- Derived time-based features enable trend and seasonal analysis in later milestones.
- Only 2 duplicate records out of ~485k confirm data integrity.
- Delay columns converted to numeric enable direct aggregation and visual summaries.
- New columns (Month, DayOfWeek, Route) lay the foundation for KPIs and dashboard metrics.

---

7. Data Dictionary

Column Name	Description	Type
FlightDate / DepDatetime	Scheduled departure timestamp	datetime
ORIGIN	Origin airport IATA code	string
DEST	Destination airport IATA code	string
Route	Combined route (ORIGIN–DEST)	string
DepDelay	Departure delay in minutes	float
ArrDelay	Arrival delay in minutes	float
CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay	Categorical delay causes	float
Cancelled	Flight cancellation indicator (True/False)	boolean
Month, DayOfWeek, DayName	Temporal dimensions	int / string
DepHour, DepMinute	Hour and minute of departure	int
_was_missing flags	Boolean flags for audit of filled values	boolean

---

8. Challenges and Resolutions

Challenge	Resolution
Inconsistent datetime formats across files	Unified to single DepDatetime column via robust parser.
Mixed data types in delay columns	Coerced to numeric with pd.to_numeric(errors='coerce').
Missing delay entries	Filled with 0 for aggregation; added audit flags for traceability.
Duplicate records	Identified and removed using df.drop_duplicates().

---

9. Tools and Libraries Used

- **pandas**, **numpy** – data handling and cleaning
- **Databricks Notebook** – execution and data management
- **Python 3.10** – core programming language
- **File Storage:** Databricks Workspace Volumes

---

10. Conclusion

Milestone 1 successfully established a robust data foundation for subsequent visual and statistical analysis.
The dataset is clean, validated, and augmented with key features that enable exploratory data analysis and KPI tracking in Milestone 2 (Univariate and Bivariate Analysis).

---

11. Next Steps (Milestone 2 Preview)

- Perform **Univariate Analysis** to identify top airlines, routes, and seasonal trends.
- Conduct **Bivariate Analysis** on delay causes by airline and airport.
- Generate 8 core visualizations (bar, box, line, and heat plots) for insight presentation.
