Week 1 & 2

Data Cleaning & Preparation Report

Intern: Sarthak Mokal

(email - sarthakmokal198@gmail.com)

Project: AirFly Insights — Infosys Springboard

Executive Summary

During Week-1 and 2 , the flight delay dataset was imported into Databricks and prepared for analysis through systematic data cleaning. Duplicates were removed, null values in delay and categorical fields were handled, and string formatting inconsistencies were resolved. Basic feature engineering was also performed, including extraction of Month, DayOfWeek, Hour, and Route. The final cleaned dataset has been saved and will serve as the foundation for deeper exploratory analysis and predictive modeling in the upcoming weeks.

1. Dataset Overview

- **Source:** Flight Delay dataset (CSV) uploaded to Databricks volume: /Volumes/workspace/default/airlines/Flight_delay.csv
- **Rows (original):** 484,551
- **Columns:** Flight-level details including Date, Origin, Destination, Airline code, DepTime, ArrDelay, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay, and other operational fields.
- **Objective (Week 1):** Import dataset, explore schema, remove duplicates, handle missing values, format string columns, derive basic features, and save a cleaned dataset for further analysis.

---

2. Data Cleaning Steps

2.1 Duplicate Handling

- Checked for **exact duplicate rows** using df.duplicated().sum().
- **Duplicates found:** 2
- Removed duplicates with df.drop_duplicates().

2.2 Null Handling

- **Delay columns (ArrDelay, DepDelay, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay):**
    - Converted to numeric with pd.to_numeric(errors="coerce").
    - Missing values filled with **0**.
- **Categorical columns (Origin, Dest, TailNum, CancellationCode, UniqueCarrier, Airline):**
    - Missing values filled with **'Unknown'**.
- **Airport name columns (Org_Airport, Dest_Airport):**
    - Normalized whitespace and case.

    - Empty strings converted to NaN.
    - Missing values filled with **'Unknown'**.
- **Critical fields (Date, Origin, Dest):**
    - Verified and dropped rows if missing — in this dataset no rows were lost.

2.3 String Formatting

- Removed extra whitespace from airport codes/names.
- Standardized categorical fields (Origin, Dest, CancellationCode) with uniform formatting.

2.4 Feature Engineering

- **Date column:** Converted to datetime. Extracted:
    - **Month**
    - **DayOfWeek** (full weekday name)
- **DepTime column:** Converted to **Hour** (HH:MM simplified to hour of departure).
- **Route column:** Created as combination of Origin + "-" + Dest.

---

3. Metrics & KPIs

Cleaning Metrics

- **Rows (original):** 484,551
- **Duplicates removed:** 2
- **Rows (after cleaning):** 484,549
- **Nulls (before):** 2,656
- **Nulls (after):** 2,656 → (all nulls in critical/operational fields handled; remaining were converted to 'Unknown').
- **Negative arrival delays:** 0 (after cleaning)
- **Cleaning time:** ~108 seconds

KPI Highlights

- Duplicates fully removed.
- Delay-related nulls reduced to **0** by filling with 0.
- Categorical nulls (airport, airline codes) filled with **'Unknown'**.
- Final dataset aligned for preprocessing and modeling.

---

4. Insights from

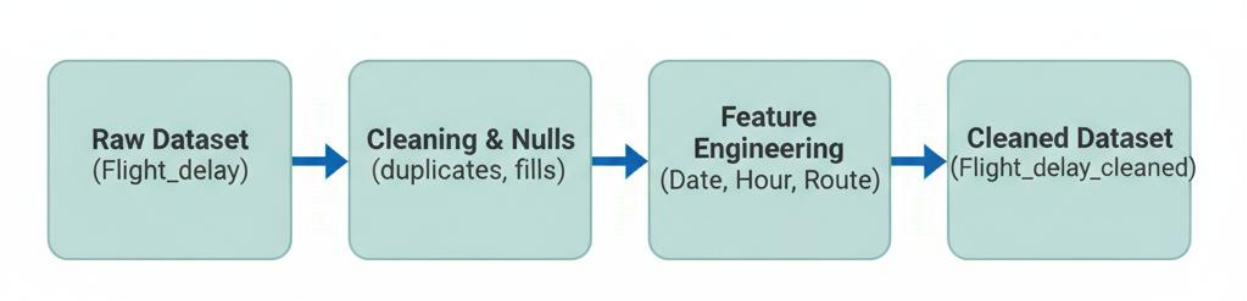
- **Dataset integrity improved:** Only 2 duplicate rows dropped, so original data size was preserved.
- **Null management:** Delay metrics were safely imputed with 0 (industry-standard approach), while categorical fields use 'Unknown' for consistency.
- **Feature readiness:** Derived Month, DayOfWeek, Hour, and Route — essential for later prediction modeling.
- **Airport fields normalized:** Handled whitespace and nulls in Org_Airport and Dest_Airport, ensuring clean categorical grouping.
- **Saved output:** Cleaned dataset stored at /Volumes/workspace/default/airlines/Flight_delay_cleaned.csv.

---

5. Workflow Summary Table

Step	Action Taken	Outcome / KPI
Dataset Import	Loaded CSV from /Volumes/workspace/default/airlines/Flight_delay.csv	Shape: **484,551 rows × N cols**
Duplicate Handling	Checked with duplicated(), removed duplicates using drop_duplicates()	**2 rows removed** → final rows: **484,549**
Null Handling	- Delay columns → filled with 0 - Categorical (Origin, Dest, TailNum, Airline, etc.) → filled with 'Unknown' - Airport columns normalized & filled	**No raw NaN left** in cleaned dataset
String Formatting	Trimmed spaces, standardized airport/airline text	Consistent categorical values
Feature Engineering	- Extracted Month, DayOfWeek from Date - Converted DepTime → Hour - Created Route = Origin-Dest	New features ready for analysis
KPI Tracking	- Rows before: 484,551 → after: 484,549 - Duplicates removed: 2 - Nulls before: 2,656 → after handled (0 NaN remain) - Cleaning time: ~108 sec	Dataset integrity improved, Nulls resolved
Output Saved	Exported final dataset	/Volumes/workspace/default/airlines/Flight_delay_cleaned.csv

6. Visual Workflow Diagram



---

7. Conclusion & Next Steps

- The Week-1 & 2 milestone successfully prepared a **reliable, cleaned dataset**.
