Here is your **comprehensive and fully updated analytical report** based on the latest version of data\_cleaning.py.

# ✈️ Flight Data Cleaning & Preprocessing Report

**Project:** Business Intelligence Project **Contributors:** Zehra Ahmed, Farah Inayat, Kisa Fatima, Zuha Aqib **Date:** 18-May-2025 **Dataset:** flights\_sample\_3m.csv **Objective:** Transform raw flight data into a clean, analysis-ready format with zero ambiguities, leveraging domain logic, Python functions, and statistical verification.

## 🧩 Section 1: Introduction

This report documents every step in the data cleaning process performed on a 3M+ row U.S. flight dataset. The cleaning pipeline ensures:

* All columns are in the correct format.
* Missing data is logically imputed.
* Redundant or misleading fields are removed.
* Time-based delays are accurately calculated.
* New BI-friendly features are created.

## 📥 Section 2: Dataset Loading & Basic Metadata

* Loaded with pd.read\_csv('../data/flights\_sample\_3m.csv').
* Set pd.set\_option('display.max\_columns', None) for full visibility.
* Basic checks: df.shape, df.dtypes, and df.describe(include='all').

## 🔄 Section 3: Data Type Conversions

Applied custom functions for type coercion:

| Column | From | To | Reason |
| --- | --- | --- | --- |
| FL\_DATE | object | datetime | For extracting year/month/day |
| AIRLINE, ORIGIN, DEST, etc. | object | string | For clarity in BI tools |
| CANCELLED, DIVERTED | float64 | int64 | To enable binary logic |

## 🧹 Section 4: Redundant Column Removal

* **Dropped:** AIRLINE\_DOT
  + Justification: Derived by combining AIRLINE + AIRLINE\_CODE. Redundant.

## 🔍 Section 5: Text Columns Audited for Validity

Checked for:

* Misspellings
* Redundant categories
* Inconsistent casing

### Checked columns:

* AIRLINE, AIRLINE\_CODE, ORIGIN, DEST, CANCELLATION\_CODE

✅ All values were found consistent and valid.

## ❌ Section 6: Missing Value Treatment

### 6.1 Cancelled Flights (Full Null Rows)

* 77,615 rows had nulls in nearly all time and delay columns.
* Verified: All were CANCELLED == 1.
* ✅ Filled with 0:
  + DEP\_TIME, WHEELS\_OFF, ARR\_TIME, ELAPSED\_TIME, AIR\_TIME, all delay fields
* ⚠️ CRS\_ELAPSED\_TIME was deferred for later calculation from scheduled times.

### 6.2 DEP\_DELAY (29 missing)

* Calculated using: DEP\_DELAY = ACTUAL\_DEP\_TIME - SCHEDULED\_DEP\_TIME (both in hhmm)
* Converted hhmm → minutes using:

def hhmm\_to\_minutes(hhmm):  
 return (hhmm // 100) \* 60 + (hhmm % 100)

### 6.3 TAXI\_OUT (1,191 missing)

* All 1,191 rows were cancelled flights.
* ⚠️ Some cancelled flights still had TAXI\_OUT, meaning taxiing began before cancel.
* If TAXI\_OUT was null and CANCELLED == 1: filled with 0.

### 6.4 WHEELS\_ON (802 nulls)

* Mostly diverted flights (DIVERTED == 1)
* 2 flights were neither diverted nor cancelled → incomplete → dropped
* For diverted flights:
  + Filled with 0: WHEELS\_ON, TAXI\_IN, ARR\_TIME, AIR\_TIME, ARR\_DELAY, etc.

### 6.5 ARR\_DELAY Calculation

* If ARR\_DELAY was null but both times were present: ARR\_DELAY = ACTUAL\_ARR\_TIME - SCHEDULED\_ARR\_TIME
* Handled wrap-around using +1440 if needed.

### 6.6 CRS\_ELAPSED\_TIME

* Derived using:

CRS\_ELAPSED\_TIME = SCHEDULED\_ARR\_TIME - SCHEDULED\_DEP\_TIME

* Wrap-around for overnight flights.

### 6.7 ACTUAL\_ELAPSED\_TIME

* Calculated from:

ACTUAL\_ELAPSED\_TIME = ACTUAL\_ARR\_TIME - ACTUAL\_DEP\_TIME

### 6.8 AIR\_TIME

* Based on:

AIR\_TIME = WHEELS\_REACHED\_DESTINATION - WHEELS\_LEFT\_ORIGIN

### 6.9 Delay Reason Columns

| Column | Imputation |
| --- | --- |
| DELAY\_DUE\_CARRIER | 0.0 |
| DELAY\_DUE\_WEATHER | 0.0 |
| DELAY\_DUE\_NAS | 0.0 |
| DELAY\_DUE\_SECURITY | 0.0 |
| DELAY\_DUE\_LATE\_AIRCRAFT | 0.0 |

Justification: Null implies no delay, and 0 makes statistical sense.

## 🧠 Section 7: Feature Engineering

### 7.1 Date Components from FL\_DATE

| New Column | Purpose |
| --- | --- |
| year | Yearly trends |
| quarter | Quarterly analysis |
| month | Seasonality |
| day | Daily patterns |

### 7.2 Binary Categoricals

| Original Column | New Categorical |
| --- | --- |
| CANCELLED | cancelled\_c (‘Yes’/‘No’) |
| DIVERTED | diverted\_c (‘Yes’/‘No’) |

## ✏️ Section 8: Label Enhancement

Mapped CANCELLATION\_CODE:

* 'A' → 'Carrier'
* 'B' → 'Weather'
* 'C' → 'Airspace System (NAS)'
* 'D' → 'Security'
* NaN → 'Not Cancelled'

Improves interpretability for stakeholders.

## 🏷️ Section 9: Column Renaming

| Old Column | New Column |
| --- | --- |
| DEP\_TIME | ACTUAL\_DEP\_TIME |
| CRS\_DEP\_TIME | SCHEDULED\_DEP\_TIME |
| ARR\_TIME | ACTUAL\_ARR\_TIME |
| CRS\_ARR\_TIME | SCHEDULED\_ARR\_TIME |
| ELAPSED\_TIME | ACTUAL\_ELAPSED\_TIME |
| CRS\_ELAPSED\_TIME | SCHEDULED\_ELAPSED\_TIME |
| FL\_DATE | FLIGHT\_DATE |
| WHEELS\_OFF | WHEELS\_LEFT\_ORIGIN |
| WHEELS\_ON | WHEELS\_REACHED\_DESTINATION |
| TAXI\_IN | TAXI\_TIME\_DESTINATION |
| TAXI\_OUT | TAXI\_TIME\_ORIGIN |
| AIR\_TIME | IN\_AIR\_DURATION |

## ✅ Section 10: Final Check & Output

* ✅ All missing values were resolved.
* ✅ All time delays and durations are computed or filled.
* ✅ Dataset saved as:
  + flights\_sample\_3m\_cleaned.csv
  + flights\_sample\_1k\_cleaned.csv (sample)

## 📌 Summary

**Total Nulls Fixed:** ✅ All relevant **New Features Created:** ✅ 8+ **Redundant Fields Removed:** ✅ 1 **Final Status:** Clean, robust, and ready for BI dashboards or analytics pipelines.

Let me know if you’d like me to package this into a downloadable DOCX or PDF as well — or generate visual timelines or infographics to explain the flight timeline and delays!