**Preprocessing Logic for Flight Dataset(FOR CANCELLATION ANALYSIS)**

**Objective**

Clean and transform the raw flight dataset (~1M rows) to prepare it for **seasonal and cancellation analysis**, ensuring:

* Correct time formats
* Consistent delay/taxi columns
* Standardized cancellation codes
* No critical null values
* Additional features for analysis (day\_of\_week, month, route)

**1. Load and Initial Cleanup**

* Load CSV file using pandas.read\_csv.
* Strip extra spaces from column names.
* Ensure FL\_DATE is a valid datetime (MM-DD-YYYY) format.
* Remove any rows with invalid dates.

**Rationale:** Prevent errors when extracting time-based features.

**2. Convert Time Columns (HHMM → HH:MM)**

**Columns:**  
CRS\_DEP\_TIME, DEP\_TIME, WHEELS\_OFF, CRS\_ARR\_TIME, ARR\_TIME, WHEELS\_ON

**Action:**

* Converted integer HHMM format (e.g., 1 → 00:01, 1136 → 11:36) to string format HH:MM.
* Created new columns with suffix \_CLEAN for all time columns.

**Rationale:** Standardized time format for readability and further calculations.

**3. Remove Noisy Rows**

* Identified rows where **actuals are missing but scheduled times exist**:
  + DEP\_DELAY missing while DEP\_TIME exists
  + TAXI\_OUT missing while DEP\_TIME exists
  + ARR\_DELAY missing while ARR\_TIME exists
  + TAXI\_IN missing while ARR\_TIME exists
* Dropped these rows as they represent noise.

**Rationale:** Ensure only meaningful rows remain; noisy rows could distort analysis.

**4. Impute Remaining Values**

* **Departure/arrival delays, taxi times** (DEP\_DELAY, ARR\_DELAY, TAXI\_OUT, TAXI\_IN) → fill missing values with 0.
* **Missing actual times** (DEP\_TIME\_CLEAN, ARR\_TIME\_CLEAN) → filled with scheduled times (CRS\_DEP\_TIME\_CLEAN, CRS\_ARR\_TIME\_CLEAN) when missing.
* **Cancellation code (CANCELLATION\_CODE)** → replace blank/NaN with 'N' (Not Cancelled).
* **Delay reason columns** (DELAY\_DUE\_CARRIER, DELAY\_DUE\_WEATHER, DELAY\_DUE\_NAS, DELAY\_DUE\_SECURITY, DELAY\_DUE\_LATE\_AIRCRAFT) → fill missing values with 0.

**Rationale:**

* Ensures **numeric columns have no nulls** for analysis.
* Standardizes cancellation codes for consistent categorization.

**5. Feature Engineering**

* **Month (MONTH)** → extracted from FL\_DATE (1 = January, 12 = December).
* **Day of Week (DAY\_OF\_WEEK)** → extracted from FL\_DATE (0 = Monday, 6 = Sunday).
* **Route (Route)** → concatenation of ORIGIN and DEST (e.g., ATL-DAB).

**Rationale:** These features are essential for:

* Seasonal analysis (monthly/weekday trends)
* Route-level cancellation patterns

**6. Drop Original Raw Columns**

**Dropped Columns:**

* Original HHMM/delay columns:  
  CRS\_DEP\_TIME, DEP\_TIME, DEP\_DELAY, TAXI\_OUT, WHEELS\_OFF, WHEELS\_ON, TAXI\_IN, CRS\_ARR\_TIME, ARR\_TIME, ARR\_DELAY

**Rationale:** Cleaned columns with \_CLEAN suffix replace originals; avoids duplication and confusion.

**7. Final Checks**

* Verified **no nulls** remain in the critical columns for analysis.
* Confirmed all time columns are standardized (HH:MM).
* Dataset ready for **cancellation, seasonal, and route-level analysis**.