**Preprocessing Steps Summary**

**1. Data Loading**

* The dataset is loaded from a CSV file using pandas.
* File path: "C:\\Users\\dandu\\OneDrive\\Desktop\\INF\_FINAL\_PROJ\\MILESTONE 1\\FINAL\_FLIGHT\_DELAYS\_DATA.csv"

**2. Initial Data Exploration**

* Displayed dataset info (df.info()), first 5 rows (df.head()), and shape.
* Identified numerical and categorical columns.
* Generated descriptive statistics for both numerical and categorical columns.

**3. Handling Duplicates and Trimming**

* Dropped duplicate rows (no duplicates found in this case).
* Stripped whitespace from all string columns to ensure consistency.

**4. Missing Value Imputation**

* Identified columns with missing values.
* **Dropped rows with nulls** in key departure-related columns:
  + DEP\_TIME, DEP\_DELAY, TAXI\_OUT, WHEELS\_OFF
* **Dropped rows with nulls** in key arrival-related columns:
  + ARR\_TIME, ARR\_DELAY, TAXI\_IN, WHEELS\_ON
* **Filled missing values** in:
  + CANCELLATION\_CODE → filled with '0' (string)
  + Delay cause columns (e.g., DELAY\_DUE\_CARRIER, DELAY\_DUE\_WEATHER, etc.) → filled with 0

**5. Time Format Conversion**

* Converted time columns from integer HHMM format to string HH:MM format:
  + CRS\_DEP\_TIME, DEP\_TIME, WHEELS\_OFF, WHEELS\_ON, CRS\_ARR\_TIME, ARR\_TIME
* Created new columns with \_HHMM suffix and dropped the original columns.
* Renamed the new columns back to the original names.

**6. Feature Engineering**

* **Date Features**:
  + Converted FL\_DATE to datetime.
  + Extracted MONTH and DAY\_OF\_WEEK.
* **Route Feature**:
  + Created Route as ORIGIN-DEST.
* **Hour Features**:
  + Extracted hour from time columns (e.g., CRS\_DEP\_TIME\_HOUR, DEP\_TIME\_HOUR, etc.).

**7. Final Cleanup**

* Dropped the original time columns (after extracting hours).
* Saved the preprocessed dataset as Preprocessed\_data.csv.

**8. Feature Dictionary**

* Created a CSV file (Feature\_Dictionary.csv) documenting:
  + Feature names
  + Descriptions
  + Data types
  + Additional notes

**Key Preprocessing Decisions**

* **Missing Values**: Rows with missing critical timing data (departure/arrival) were dropped. Other missing values were filled with zeros or a placeholder.
* **Time Conversion**: Converted integer-based time formats to human-readable HH:MM strings and extracted hour components for temporal analysis.
* **Feature Engineering**: Added temporal, categorical, and derived features (e.g., Route, MONTH, HOUR features) to support deeper analysis.
* **Data Quality**: Ensured consistency by trimming strings and removing duplicates.

**Final Output**

* Cleaned and enriched dataset with **35 columns**.
* Saved as Preprocessed\_data.csv.
* Feature dictionary saved as Feature\_Dictionary.csv.