**Flight Data Transformations**

**Data Quality Operations**

These are cleaning steps applied to improve the dataset:

1. **Column Removal**
   * Dropped unnecessary or redundant columns like IDs, elapsed time, cancellation codes, and delay breakdowns to reduce noise.
2. **Date Handling**
   * Converted FL\_DATE (originally with time) into proper Date type.
   * Reconstructed missing FL\_DATE values using YEAR, MONTH, and DAY\_OF\_MONTH.
3. **Missing Values Handling**
   * Checked missing values for all columns and summarized them.
   * Visualized missingness per column using a bar chart.
4. **Duplicate Handling**
   * Counted duplicates.
   * Removed exact duplicate rows with distinct().
   * Verified duplicates were successfully removed (pie chart for duplicate vs unique).
5. **Outlier Detection**
   * Checked numeric columns for negative/impossible values.
   * Created a boxplot for DISTANCE to highlight extreme values.
6. **Date Range Coverage**
   * Computed and reported date range of flights.
   * Histogram for flight coverage across time (2015–2024).
7. **Flight Trends**
   * Plotted flights per year for completeness check (ensures data exists across the years).
8. **Record Identifier Creation**
   * Added a unique FLIGHT\_ID column (combination of FL\_DATE and ORIGIN).

**Data Validation Rules**

These ensure data integrity and correctness:

1. **Year Range Validation**
   * Confirmed all YEAR values are between 2015 and 2024.
2. **Month Validation**
   * Checked that MONTH values fall between 1 and 12.
3. **Distance Validation**
   * Ensured all DISTANCE values are strictly positive.
4. **Delay Range Checks**
   * Extracted ranges for ARR\_DELAY and DEP\_DELAY to inspect extreme values (like -107 and 2000+ mins).
5. **Duplicate Row Validation**
   * Verified number of duplicate rows before and after cleaning.
6. **Missing Value Validation**
   * Reported missing values per column in detail (missing\_summary).

**Plots for Validation & Quality**

* **Missing values bar chart** → checks completeness.
* **Duplicate vs unique pie chart** → ensures deduplication.
* **Histogram of flight dates** → checks temporal coverage.
* **Flights per year bar chart** → validates consistency across years.
* **Boxplot of distance** → checks outliers/implausible records.