Airfly Insights Report

# 1. Dataset Overview

The airline delay dataset was analyzed to understand flight patterns, delays, and overall performance. The raw dataset contained 484,551 rows and 29 columns. After cleaning and feature engineering, the final dataset has 484,549 rows and 33 columns.

# 2. Data Cleaning Steps (Using Pandas)

The following steps were performed to clean and preprocess the dataset:  
1. Converted 'Date' column into datetime format with explicit format (DD-MM-YYYY).  
2. Extracted additional features: Month, DayNumber, Hour, and Route.  
3. Handled missing values – filled/removed nulls in 'Org\_Airport' and 'Dest\_Airport'.  
4. Checked and confirmed zero null values in final dataset.  
5. Removed duplicate rows (2 duplicates found, resulting in 484,549 rows).  
6. Verified schema and feature consistency.

# 3. Metrics and Insights

- Raw Dataset Shape: (484,551, 29)  
- Cleaned Dataset Shape: (484,549, 33)  
- Null Values (RAW): Org\_Airport (1177), Dest\_Airport (1479)  
- Null Values (CLEANED): 0  
- Duplicate Rows (RAW): 2  
- Duplicate Rows (FINAL): 0  
- Extra Derived Columns: Month, DayNumber, Hour, Route

Insights:  
• The dataset now contains enriched time-based features for better trend analysis.  
• Null values have been fully resolved, ensuring consistency.  
• Duplicate records were removed, preventing bias in downstream analysis.