# AIRFLY INSIGHTS REPORT

## 1. Dataset Overview

The airline delay dataset was analyzed to uncover trends in flight delays, operational performance, and temporal patterns. The original dataset contained **484,559 rows** and **29 columns**. After thorough data cleaning and feature engineering, the final processed dataset consists of **484,549 rows** and **33 columns**, ensuring data integrity and enhanced analytical capability.

**Initial Dataset:**

- Rows: 484,559
- Columns: 29

**Final Dataset:**

- Rows: 484,549
- Columns: 33
- Total Data Points: 14,051,979

## 2. Data Cleaning Steps (Using Pandas)

The following data cleaning and preprocessing steps were performed using **Pandas** to ensure data quality and usability:

### 2.1. Handling Missing Values

- **Org_Airport**: 1,177 null values
- **Dest_Airport**: 1,479 null values
- **Canceled**: 13 null values
- These missing values were imputed to ensure a complete dataset.

### 2.2. Removing Duplicate Rows

- **Duplicate Rows (Raw)**: 10
- **Duplicate Rows (Final)**: 0
- Duplicates were identified and removed to avoid skewed analysis.

### 2.3. Data Type Conversion

- **Date Column**: Converted to datetime format for time-series analysis.

### 2.4. Feature Engineering

- **Month**: Extracted from the date column.
- **DayNumber**: Day of the week (1 = Monday, 7 = Sunday).

- **Hour**: Extracted from departure time.
- **Route**: Created by combining Origin and Destination airports.
- **Duration**: Created by converting the AirTime minutes into readable Time.

### 2.5. Handling Delay Columns

- Delay-related columns (e.g., CarrierDelay, WeatherDelay, etc.) were checked for consistency and missing values.
- Zero delays were retained as valid entries.

## 3. Metrics and Insights

### 3.1. Dataset Metrics

- **Raw Dataset Shape**: (484,559, 29)
- **Cleaned Dataset Shape**: (484,549, 33)
- **Null Values (Raw)**: Org_Airport (1,177), Dest_Airport (1,479), Cancelled(13)
- **Null Values (Cleaned)**: 0
- **Duplicate Rows (Raw)**: 10
- **Duplicate Rows (Final)**: 0

### 3.2. Distance Analysis

- **Minimum Distance**: 31 miles
- **Maximum Distance**: 4,502 miles
- **Average Distance**: 752.14 miles
- **Flights with Distance > 1,000 miles**: Extracted for further analysis (includes FlightNum, Origin, Dest, Distance)

### 3.3. Temporal Insights

- **Flights by Day of Week**:
    - Monday (1): 70,254
    - Tuesday (2): 65,934
    - Wednesday (3): 63,055
    - Thursday (4): 75,011
    - Friday (5): 88,972
    - Saturday (6): 51,330
    - Sunday (7): 69,995

- **Peak Day**: Friday (Day 5) with 88,972 flights.

### 3.4. Operational Metrics

- **Average Taxi In Time**: 6.78 minutes
- **Average Taxi Out Time**: 19.15 minutes

### 3.5. Additional Insights

- **Top N number of Longest Flights**: Extracted with details (Airline, FlightNum, Origin, Dest, Distance).
- **Top N number of Shortest Flights**: Extracted with details (Airline, FlightNum, Origin, Dest, Distance).
- **Route Feature**: Added to enable route-based analysis.
- **Date Column**: Successfully parsed and ready for time-series modeling.

## Insights Summary

- The dataset is now fully cleaned, with no missing or duplicate values.
- New time-based features (Month, DayNumber, Hour, Route) enhance analytical depth.
- Friday is the busiest day for flights.
- Long-haul flights (>1,000 miles) have been isolated for targeted analysis.
- Taxi times indicate longer delays during departure than arrival.
- The dataset is ready for advanced analysis, including delay prediction and seasonal trend modeling.