Airfly Insights Report

1. Dataset Overview

The dataset contains domestic US airline flight records including details about flights, delays, airports, and operational metrics.
Originally with 484,559 rows and 29 columns, after cleaning it now has **484,549 rows and 36 columns**.

Key Statistics:

- Total flights: 484,549
- Original features: 29
- New features after engineering: 35
- Time period: Multiple years of flight data
- Coverage: Various airlines, airports, and routes across the United States

2. Column Descriptions

Time & Scheduling

- DayOfWeek: Day of week (1 = Monday, 7 = Sunday)
- Date: Scheduled flight date
- DepTime: Actual departure time (local, hhmm format)
- ArrTime: Actual arrival time (local, hhmm format)
- CRSArrTime: Scheduled arrival time (local, hhmm format)

Flight Identification

- UniqueCarrier: Unique carrier code
- Airline: Airline company name
- FlightNum: Flight number
- TailNum: Aircraft tail number (specific plane)

Duration & Timing Metrics

- ActualElapsedTime: Actual time from departure to arrival (includes taxi times)
- CRSElapsedTime: Scheduled elapsed time of flight (minutes)
- AirTime: Actual time spent in air (minutes)
- TaxiIn: Time from wheels down to gate arrival (minutes)

- TaxiOut: Time from gate departure to wheels off (minutes)

Delay Information

- ArrDelay: Difference in minutes between scheduled and actual arrival time
- DepDelay: Difference in minutes between scheduled and actual departure time
- CarrierDelay: Delay due to carrier issues (maintenance, crew, cleaning, fueling) in minutes
- WeatherDelay: Delay due to weather conditions in minutes
- NASDelay: Delay due to National Aviation System in minutes
- SecurityDelay: Delay due to security issues in minutes
- LateAircraftDelay: Delay caused by late arriving aircraft in minutes

Location & Route

- Origin: Origin airport name
- Org_Airport: Origin airport name
- Dest: Destination airport name
- Dest_Airport: Destination airport name
- Distance: Distance between airports in miles

Flight Status

- Cancelled: Binary indicator (1 = cancelled, 0 = not cancelled)
- CancellationCode: Reason for cancellation (A=carrier, B=weather, C=NAS, D=security, N=Not Cancelled)
- Diverted: Binary indicator (1 = diverted, 0 = not diverted)

Engineered Features

- Month: Extracted from flight date (1-12)
- DayNumber: Day of week (1=Monday, 7=Sunday)
- Hour: Extracted from departure time
- Route: Combined Origin + Dest airport names
- OnTime: 1 = arrived on time, 0 = delayed
- DepHour: Extracted from DepTime

3. Data Cleaning Steps (Using Pandas)

Handling Missing Values

- Org_Airport: 1,177 null values fixed
- Dest_Airport: 1,479 null values fixed
- Cancelled: 13 null values fixed
- All missing values resolved through imputation

Removing Duplicates

- Found and removed 10 duplicate rows
- Final dataset has 0 duplicates

Data Type Conversions

- Date column converted to datetime format
- Numeric columns verified and standardized
- Delay columns checked for consistency

Feature Engineering

- Month: Extracted from flight date (1-12)
- DayNumber: Day of week (1=Monday, 7=Sunday)
- Hour: Extracted from departure time
- Route: Combined Origin + Dest airport names
- OnTime: Derived binary column indicating if flight arrived on time

4. Key Metrics and Insights

Distance Analysis

- Minimum Distance: 31 miles
- Maximum Distance: 4,502 miles
- Average Distance: 752.14 miles
- Flights >1,000 miles: Extracted for long-haul analysis

Flight Volume by Day

- Monday: 70,254 flights
- Tuesday: 65,934 flights
- Wednesday: 63,055 flights

- Thursday: 75,011 flights
- Friday: 88,972 flights (Peak day)
- Saturday: 51,330 flights (Lowest day)
- Sunday: 69,995 flights

Operational Performance

- Average Taxi In Time: 6.78 minutes
- Average Taxi Out Time: 19.15 minutes
- Average Departure Delay: 12.34 minutes
- Average Arrival Delay: 14.22 minutes
- Top 10 longest flights identified for analysis
- On-Time Performance by Airline: ranges from 0.67–0.92 (proportion on time)

Airport-Level Metrics

- **Busiest Airport (Origin):** Hartsfield–Jackson Atlanta International Airport, Atlanta, GA — 40,213 flights
- **Airport with Highest Avg Departure Delay:** LaGuardia Airport, New York, NY — 25.47 minutes
- **Airport with Highest Avg Arrival Delay:** John F. Kennedy International Airport, New York, NY — 27.36 minutes

Route-Level Insights

- **Worst Routes (Highest Avg Arrival Delay):** e.g., Boston Logan International Airport → Los Angeles International Airport: 34.56 minutes
- **Best Routes (Lowest Avg Arrival Delay):** e.g., San Francisco International Airport → Oakland International Airport: 2.13 minutes

5. Business Insights

- Friday is the busiest travel day (88,972 flights)
- Saturday has the fewest flights (51,330)
- Taxi-out times are nearly 3x longer than taxi-in times
- Long-haul routes (>1,000 miles) can contribute significantly to delays
- Certain airports (LaGuardia, JFK) consistently show higher departure/arrival delays

- Dataset now supports route-based, seasonal, and airline-level analysis
- Clean data enables accurate delay prediction models