Airfly Insights Report

1. Dataset Overview

The Airfly dataset serves as a comprehensive record of domestic airline operations across the United States. It provides detailed information on scheduled and actual flight timings, delays, cancellations, and airport performance metrics. The raw data initially consisted of 484,559 flights with 29 columns, and after applying cleaning and feature engineering, the final dataset contains 484,549 flights and 33 columns.
This transformation ensured data integrity, improved usability, and prepared the dataset for advanced analytics such as trend detection, delay prediction, and operational benchmarking.

- Key Highlights
  - Total flight records analyzed: 484,549
  - Original attributes: 29
  - Features after enrichment: 33
  - Time span: Multiple years of airline activity
  - Coverage: Wide range of airlines, airports, and U.S. routes

The dataset is now in a structured format that can directly support both descriptive statistics and predictive modeling

2. Dataset Features Explained

- To better understand the scope of the dataset, features are grouped into logical categories:
- Flight Identification
  - Airline – Two-character code representing the airline (e.g., AA = American Airlines).
  - FlightNum – A unique identifier for each scheduled flight.
  - Origin / Destination – Departure and arrival airports coded using IATA standards.
  - Route – A derived feature combining Origin and Destination to represent city pairs.
- Time and Scheduling
  - Date – Calendar date of the flight, stored in proper datetime format.
  - Month – Extracted month number for seasonal analysis.
  - DayNumber – Day of the week (1 = Monday through 7 = Sunday).
  - Hour – Departure time converted into 24-hour clock format.
  - CRSDepTime / CRSArrTime – Scheduled departure and arrival times from airline timetables.
- Duration Metrics
  - ActualElapsedTime – Total real travel time (gate-to-gate).

- CRSE lapsed Time – Planned flight time from the schedule.
- AirTime – Minutes spent in the air excluding taxiing.
- TaxiIn / TaxiOut – Time spent after landing and before takeoff respectively.

- Delay Information
  - ArrDelay / DepDelay – Minutes of delay at arrival and departure.
  - CarrierDelay – Delays caused by the airline itself.
  - WeatherDelay – Minutes delayed due to adverse weather conditions.
  - NASDelay – Delays caused by the National Aviation System.
  - SecurityDelay – Delays caused by security-related incidents.
  - LateAircraftDelay – Cascading delays due to late arrivals of previous flights.
- Flight Status
  - Cancelled – Indicates if the flight was canceled.
  - Diverted – Shows if the flight was diverted mid-route.
  - CancellationCode – Provides the specific reason for cancellation (e.g., Carrier, Weather, NAS, or Security).
- Geography & Distance
  - Distance – The flight distance in miles.
  - Org_Airport / Dest_Airport – Extended airport details after enrichment.

3. Data Cleaning & Preprocessing (Using Pandas)

- The dataset required multiple preparation steps to ensure reliability:
  1. Handling Missing Values
     - *Org_Airport*: 1,177 missing entries replaced with “Unknown”.
     - *Dest_Airport*: 1,479 missing entries replaced with “Unknown”.
     - All other missing values resolved via imputation or removal.
  2. Duplicate Removal
     - Detected 10 duplicate records, which were removed.
     - Final dataset contains zero duplicates.
  3. Data Type Standardization
     - Converted Date to proper datetime type.
     - Verified that numeric columns such as Distance and Delays are consistent.
  4. Feature Engineering
     - Extracted Month, DayNumber, Hour from the Date/Time fields.
     - Created a Route field to capture Origin–Destination pairs.
  5. Quality Validation
     - Final dataset has no nulls, no duplicates, and is enriched with time-based features for improved analysis.

4. Metrics and Analytical Insights

- Distance Analysis
  - Minimum route length: 31 miles
  - Maximum route length: 4,502 miles
  - Average distance: 752.14 miles
  - Long-haul flights (>1,000 miles) extracted for specialized trend analysis.
- Flight Volume by Day of Week
  - Monday: 70,254 flights
  - Tuesday: 65,934 flights
  - Wednesday: 63,055 flights
  - Thursday: 75,011 flights
  - Friday: 88,972 flights *(busiest travel day)*
  - Saturday: 51,330 flights *(lowest travel day)*
  - Sunday: 69,995 flights
- Operational Performance
  - Average Taxi-In time: 6.78 minutes
  - Average Taxi-Out time: 19.15 minutes
  - Taxi-out is roughly 3 times longer than taxi-in, suggesting runway congestion and departure bottlenecks.
  - Top 10 longest flights were identified for further performance evaluation.
- Data Reliability Achieved
  - 0 nulls in the final dataset.
  - 0 duplicate records.
  - All temporal features properly formatted.
  - Dataset is now ready for machine learning and advanced analytics.

5. Business Insights & Recommendations

- Peak Demand: Fridays represent the busiest travel day with 88,972 flights, indicating strong pre-weekend demand for both leisure and business travelers.
- Low Demand: Saturdays record the fewest flights, suggesting this day may be ideal for airlines to conduct planned maintenance or resource optimization.
- Operational Bottlenecks: Taxi-out delays are substantially higher than taxi-in times, pointing towards potential issues with ground operations and runway allocation.
- Strategic Use: With enhanced features like Route, Month, and DayNumber, airlines can conduct seasonal demand forecasting and route profitability analysis.
- Predictive Modeling Potential: The cleaned dataset now provides a solid foundation for building delay prediction models, which can help reduce passenger inconvenience and improve airline efficiency.