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 33 columns.

Key Statistics:

- Total flights: 484,549
- Original features: 29
- New features after engineering: 33
- Time period: Multiple years of flight data
- Coverage: Various airlines, airports, and routes across the United States

2. Column Descriptions

Flight Identification

- **Airline**: Carrier code (e.g., AA, DL, UA)
- **FlightNum**: Unique flight number
- **Origin**: Departure airport code
- **Dest**: Destination airport code
- **Route**: Origin-Destination combination

Time & Date

- **Date**: Flight date (datetime format)
- **Month**: Month extracted from date
- **DayNumber**: Day of week (1-7)
- **Hour**: Hour of departure
- **CRSDepTime**: Scheduled departure time
- **CRSArrTime**: Scheduled arrival time

Duration Metrics

- **ActualElapsedTime**: Real flight time in minutes
- **CRSElapsedTime**: Scheduled flight time in minutes
- **AirTime**: Actual time in air
- **TaxiIn**: Time from landing to gate (minutes)
- **TaxiOut**: Time from gate to takeoff (minutes)

Delay Information

- **ArrDelay**: Arrival delay in minutes
- **DepDelay**: Departure delay in minutes
- **CarrierDelay**: Delay caused by airline
- **WeatherDelay**: Delay due to weather
- **NASDelay**: National Air System delay
- **SecurityDelay**: Security-related delay
- **LateAircraftDelay**: Delay from previous aircraft

Flight Status

- **Cancelled**: 1 if flight cancelled, 0 otherwise
- **Diverted**: 1 if flight diverted, 0 otherwise
- **CancellationCode**: Reason for cancellation (A=carrier, B=weather, C=NAS, D=security, N=Not cancelled)

Distance & Location

- **Distance**: Flight distance in miles
- **Org_Airport**: Origin airport details
- **Dest_Airport**: Destination airport details

3. Data Cleaning Steps (Using Pandas)

Handling Missing Values

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

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 codes

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
- Top 10 longest flights identified for analysis

Data Quality Achieved

- Zero null values in final dataset
- Zero duplicate records
- All dates properly formatted
- Enhanced with time-based features
- Ready for machine learning and analysis

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