Air fly Insights Report

1. Dataset Overview

The airline delay dataset was explored to gain insights into flight trends, delay behavior, and performance metrics. Initially, the dataset consisted of 484,551 rows and 29 columns. Following data cleaning and feature engineering, the processed dataset now contains 484,549 rows and 33 columns.

2. Data Cleaning Steps (Using Pandas)

The following steps were performed to clean and preprocess the dataset: **1. Dataset Overview**

Initial Dataset: 484,551 rows × 29 columns

Final Dataset (after cleaning & feature engineering): 484,549 rows × 33 columns

Total Data Points: 14,051,979

2. Data Types (Selected Columns)

• Integer Columns-> ActualElapsedTime, CRSElapsedTime, AirTime, ArrDelay, DepDelay, Distance, TaxiIn, TaxiOut, Cancelled, Diverted, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay

Object Columns-> Origin, Org_Airport, Dest, Dest_Airport, CancellationCode

3. Missing Values (Raw Data)

• Org_Airport → 1,177 null values

Dest_Airport → 1,479 null values

5. Distance Statistics

• Minimum Distance: 31 miles

• Maximum Distance: 4,502 miles

Average Distance: 752.14 miles

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

Minimum Distance: 31 miles

Maximum Distance: 4,502 miles

Average Distance: 752.14 miles

Flights with Distance > 1,000 miles: Extracted for separate analysis (FlightNum, Origin, Dest, Distance)

Flights by Day of Week: 1 → 70,254, 2 → 65,934, 3 → 63,055, 4 → 75,011, 5 → 88,972, 6 → 51,330, 7 → 69,995

Top 10 Longest Flights: Extracted with columns Airline, FlightNum, Origin, Dest, Distance

Average Taxi In Time: 6.78 minutes

Average Taxi Out Time: 19.15 minutes

Date Column: Successfully converted to datetime format

Insights

Dataset now has extra time-based features (Month, DayNumber, Hour, Route).

All null values are fixed.

Duplicate rows removed.

Minimum Distance: 31 miles, Maximum Distance: 4,502 miles, Average Distance: 752.14 miles.

Flights counted by day of week – highest on Day 5 (88,972 flights).

Long flights (Distance > 1,000 miles) extracted for analysis.

Average Taxi In: 6.78 minutes, Average Taxi Out: 19.15 minutes.

Date column converted to datetime format.

Top 10 longest flights identified.

Final dataset is clean and ready for analysis.

WEEK 2

1. Dataset Overview

The airline delay dataset was explored to gain insights into flight trends, delay behavior, and performance metrics. Initially, the dataset consisted of 484,551 rows and 29 columns. Following data cleaning and feature engineering, the processed dataset now contains 484,549 rows and 33 columns.

3. Metrics and Insights

1. Dataset Overview

The airline delay dataset was explored to gain insights into flight trends, delay behavior, and performance metrics.

Initial Dataset: 484,551 rows × 29 columns

Processed Dataset (after cleaning & feature engineering): 484,549 rows × 33 columns

Total Data Points: 14,051,979

2. Data Cleaning Steps (Using Pandas)

Duplicate Rows Removed: 2 duplicates dropped → 0 remaining

Missing Values (Raw Data):

Org_Airport: 1,177 null values

Dest_Airport: 1,479 null values

Missing Values (After Cleaning): 0

Date Column: Converted to datetime format and missing values filled using forward/backward fill

Derived Columns Added: Month, DayOfWeekNum, DayName, Route

Filled Missing ArrTime Values: Replaced nulls with CRSArrTime

3. Data Types (Selected Columns)

Integer Columns: ActualElapsedTime, CRSElapsedTime, AirTime, ArrDelay, DepDelay, Distance, TaxiIn, TaxiOut, Cancelled, Diverted, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay

Object Columns: Origin, Org_Airport, Dest, Dest_Airport, CancellationCode

4. Distance Statistics

Minimum Distance: 31 miles

Maximum Distance: 4,502 miles

Average Distance: 752.14 miles

Flights > 1,000 miles: Extracted separately for long-haul analysis

5. Metrics and Insights

Raw Dataset Shape: (484,551, 29)

Cleaned Dataset Shape: (484,549, 33)

Duplicate Rows (RAW): 2 → **Final:** 0

Extra Derived Columns: Month, DayNumber, Hour, Route

Average Taxi In Time: 6.78 minutes

Average Taxi Out Time: 19.15 minutes

6. Delay Analysis

6.1 Average Arrival & Departure Delays by Airline

Identify airlines with highest delays

Example:

Southwest Airlines: Avg ArrDelay = 8.5 min, Avg DepDelay = 7.2 min

Delta Airlines: Avg ArrDelay = 10.3 min, Avg DepDelay = 9.0 min

6.2 Delay Causes

CarrierDelay: Most frequent minor delays

WeatherDelay: Rare but severe

LateAircraftDelay: Major contributor to cascading delays

7. Flight Patterns

Peak Flight Hours: 06:00–10:00 and 16:00–20:00

Low Activity: 01:00–05:00

Popular Routes: IND-BWI, ISP-BWI, IND-LAS

Longest Routes (>4,000 miles): Few intercontinental or coast-to-coast flights

8. Time-based Feature Analysis

Month-wise trends: Peak months correlate with higher delays

DayOfWeek trends: Friday (Day 5) has highest flights → higher congestion

Hour-of-Day trends: Morning/evening peaks align with business travel

9. Taxi Times Analysis

Average Taxi In: 6.78 minutes

Average Taxi Out: 19.15 minutes

Longer taxi out times indicate busy airports

10. Data Quality and Preparation Notes

Null values filled logically (forward/backward fill or replacement)

Duplicate rows removed

Derived features (Month, DayName, Route) created

Dataset ready for predictive modeling, visualization, or dashboard reporting

11. Recommendations

Airlines with higher delays should investigate operational bottlenecks

Monitor long-haul flights (>1,000 miles) for performance

Airports should optimize peak-hour scheduling

Build an interactive dashboard for daily/weekly flight monitoring

12. Insights Summary & Detailed Insights

Insights Summary:

Dataset enriched with time-based features for trend analysis

Null values resolved, duplicates removed

Distance statistics: Minimum 31 mi, Maximum 4,502 mi, Average 752.14 mi

Flights counted by day of week – highest on Day 5 (88,972 flights)

Long flights (>1,000 miles) isolated for analysis

Average Taxi In: 6.78 min, Taxi Out: 19.15 min

Date column converted to datetime format

Top 10 longest flights identified

Dataset clean and ready for analysis

Detailed Insights:

Dataset now has extra time-based features: Month, DayNumber, Hour, Route

All null values are fixed

Duplicate rows removed

Minimum Distance: 31 miles, Maximum Distance: 4,502 miles, Average Distance: 752.14 miles

Flights counted by day of week – highest on Day 5 (88,972 flights)

Long flights (Distance > 1,000 miles) extracted for analysis

Average Taxi In: 6.78 minutes, Average Taxi Out: 19.15 minutes

Date column converted to datetime format

Final dataset is clean and ready for analysis