Airfly_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.

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)

2.1 Duplicate Rows

- **Raw Dataset:** 2 duplicate rows
- **After Cleaning:** 0 duplicates

2.2 Missing Values

- **Raw Missing Values:**
  - Org_Airport: 1,177 nulls
  - Dest_Airport: 1,479 nulls
- **After Cleaning:** 0 nulls

- **Date Column:** Converted to datetime format; missing values filled using forward/backward fill
- **ArrTime:** Missing values replaced with CRSArrTime

2.3 Derived Columns Added

- Month
- DayOfWeekNum / DayName
- Hour
- Route

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:** 2 → 0

- **Extra Derived Columns:** Month, DayNumber, Hour, Route
- **Flights by Day of Week:**
    - Day 1 → 70,254
    - Day 2 → 65,934
    - Day 3 → 63,055
    - Day 4 → 75,011
    - Day 5 → 88,972 (highest)
    - Day 6 → 51,330
    - Day 7 → 69,995

6. Delay Analysis

6.1 Average Arrival & Departure Delays by Airline

- Identify airlines with highest delays:
    - 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

1. **Flight Count & Basic Stats**
    - Count of flights by each airline using `value_counts()`.
    - Top airlines: Southwest Airlines Co., American Airlines Inc., American Eagle Airlines Inc., etc.
2. **Date and Departure Time Processing**
    - Convert `Date` column to datetime format (`dayfirst=True`).
    - Convert `DepTime` from HHMM integers to separate hours (`DepHour`) and minutes (`DepMinute`).
    - Combine `Date`, `DepHour`, and `DepMinute` into a single datetime column `DepDatetime`.
3. **Derived Features**
    - Extract month from `DepDatetime` → `Month`.
    - Extract day of the week name → `DayOfWeek`.
    - Extract hour of departure → `Hour`.
    - Create route string by combining `Origin` and `Dest` → `Route`.
    - Optional numeric day of week → `DayOfWeekNum`.
    - Day name derived from date → `DayName`.
4. **Time Conversion**
    - Convert HHMM integers for `DepTime`, `ArrTime`, `CRSArrTime`, and `ActualElapsedTime` to HH:MM:SS string format.
    - Handles missing or invalid values safely.
5. **Data Quality Checks**
    - Null values checked → `df.isnull().sum()` → all columns filled.
    - Final dataset contains 484,549 rows and 37 columns.
6. **Export Cleaned Data**

- Save processed dataset to CSV → `Airfly_Data_Analysis.csv`.

12. Insights Summary

- Dataset enriched with time-based features for trend analysis
- Null values resolved, duplicates removed
- Distance statistics: Min = 31 mi, Max = 4,502 mi, Avg = 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
- Final dataset is clean and ready for analysis