AirFly Insights: Data Visualization and Analysis of Airline Operations

---

1: Data Foundation and Cleaning

Project: Flight Delay Analysis
Dataset: airfly_raw_data.csv (484,552 rows × 30 columns)

Project Context

Air travel delay analysis is a crucial use case for understanding operational efficiency, customer experience, and route optimization. This project aims to build a **clean, structured dataset** to support downstream analysis and modeling of **flight delays**, **cancellations**, and **route performance**.

The raw dataset contains U.S. domestic flight operations, including schedule times, actual times, delays, cancellation codes, and airport information.

Goals

- Build a **clean and reliable data foundation** for flight delay analysis.
- Define and extract **key temporal and operational features** to support modeling and visualization.
- Ensure data quality by handling missing values, formatting inconsistencies, and type mismatches.
- Store preprocessed data in a reusable format for faster downstream development.

Key KPIs

KPI	Description
Average Arrival Delay	Mean of ArrDelay by carrier and route
Cancellation Rate	Proportion of flights canceled over total
On-Time Performance	% of flights with ArrDelay <= 0
Route Popularity	Number of flights per Origin–Destination pair
Peak Departure Hour	Hour of day with highest departures

WEEK 1: Project Initialization and Dataset Setup

1. Define Goals, KPIs, and Workflow

- Project scope established for delay and cancellation analysis.
- Workflow defined: Data ingestion → Preprocessing → Feature Engineering → EDA & Modeling.

- KPIs identified to measure airline performance and flight punctuality.

---

2. Load CSVs Using pandas

Raw data was loaded from:

/Volumes/airfly_workspace/default/airfly_insights/airfly_raw_data.csv

using pandas.read_csv().

The dataset contains:

- **484,552 rows**
- **30 columns**

---

3. Explore Schema, Types, Size, and Nulls

Check	Findings
Schema & Dtypes	Mix of int, float, object; time columns stored as int (HHMM); dates as strings
Size	~90 MB CSV file
Nulls	Missing values found in ArrTime, DepTime, Org_Airport, Dest_Airport, and Cancelled
Duplicates	Some repeated flight records by FlightNum, TailNum, Date

---

4. Sampling and Memory Optimizations

- Random sample of 5,000 rows used for quick inspection.
- Columns downcasted to efficient dtypes (e.g., int32, category) to reduce memory footprint.
- Times converted only once during preprocessing to avoid repeated parsing overhead.

WEEK 2: Preprocessing and Feature Engineering

1. Handle Nulls in Delay and Cancellation Columns

- Delay columns (ArrDelay, DepDelay, CarrierDelay, etc.) → filled with **0**.
- Cancelled → filled with **0** where missing (interpreted as not canceled).
- CancellationCode → filled with 'None' where missing.

---

2. Create Derived Features

New columns added for temporal and route-based analysis:

Feature	Description
Month	Extracted from Date
DayOfWeekNum	0–6 representation for Monday–Sunday
DepHour	Extracted from DepTime after conversion
Route	Concatenation of Origin and Dest (e.g., IND-BWI)

---

3. Format Datetime Columns

- Date parsed as datetime with dayfirst=True to handle DD-MM-YYYY format.
- DepTime, ArrTime, and CRSArrTime converted from **HHMM integers** to datetime.time objects for proper time analysis.

---

4. Save Preprocessed Data for Fast Reuse

The cleaned dataset was saved in:

/Volumes/airfly_workspace/default/airfly_insights/flights_cleaned.csv

and also optionally as Parquet for faster downstream reads.

Feature Dictionary

Column	Description
DayOfWeek	Day of week (1=Monday, etc.)
Date	Flight date
DepTime	Actual departure time (HH:MM)
ArrTime	Actual arrival time (HH:MM)
CRSArrTime	Scheduled arrival time
UniqueCarrier	Airline code
Airline	Airline name
FlightNum	Flight number
TailNum	Aircraft tail number
ActualElapsedTime	Actual flight time (minutes)
CRSElapsedTime	Scheduled flight time

Column	Description
AirTime	Airborne time (minutes)
ArrDelay	Arrival delay (minutes)
DepDelay	Departure delay (minutes)
Origin	Origin airport code
Org_Airport	Origin airport name
Dest	Destination airport code
Dest_Airport	Destination airport name
Distance	Distance in miles
TaxiIn	Taxi in time (minutes)
TaxiOut	Taxi out time (minutes)
Cancelled	1 if flight cancelled, else 0
CancellationCode	Reason for cancellation
Diverted	1 if flight diverted
CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay	Delay causes
Month, DayOfWeekNum, DepHour, Route	Derived features for analysis

---

WEEK 3: Univariate and Bivariate Visual Analysis

Explore patterns in flights, delays, cancellations, and routes using the cleaned dataset.

All analysis is performed on the **cleaned dataset**, which includes derived features like Month, DayOfWeekNum, DepHour, and Route.

Univariate Analysis

Univariate analysis focuses on a single variable at a time to understand its distribution and key characteristics.

Tasks performed:

1. **Top Airlines:** Counted the number of flights per airline to identify the busiest carriers.
2. **Top Routes:** Counted flights for each origin–destination pair to find the most frequent routes.
3. **Busiest Months:** Counted flights per month to identify peak travel periods.
4. **Flights by Day of Week:** Analyzed the number of flights per day to see weekly patterns.

5. **Departure Hour Distribution:** Checked the number of flights per hour to understand peak departure times.
6. **Flights by Origin Airport:** Counted flights from each airport to identify the busiest airports.
7. **Arrival Delay Distribution:** Examined the distribution of arrival delays to understand general punctuality.

Visualization methods used:

- Bar charts for categorical counts (Airlines, Routes, Month, Origin)
- Histograms for numeric distributions (DepHour, ArrDelay)
- Boxplots for numeric spread (ArrDelay for detecting outliers)

Bivariate Analysis

Bivariate analysis studies the relationship between two variables to identify patterns, trends, and dependencies.

Tasks performed:

1. **Arrival Delay by Airline:** Compared the distribution of delays across different airlines using boxplots.
2. **Average Arrival Delay by Month:** Calculated the mean arrival delay for each month and visualized it with a line plot to observe seasonal trends.
3. **Optional Analyses:**
    - Delay vs Departure Hour to see if flights at certain hours are more delayed.
    - Delay vs Distance to check if longer flights are more prone to delays.

Visualization methods used:

- Boxplots for numeric vs categorical relationships (ArrDelay vs Airline)
- Line plots for numeric trends over time (ArrDelay vs Month)
- Scatter plots for numeric vs numeric comparisons
