# Airfly Insights Documentation

## Milestone 1: Data Foundation and Cleaning

### Week 1: Project Initialization and Dataset Setup

**Objectives**

- Define goals, KPIs, and workflow
- Load CSVs using pandas
- Explore schema, types, size, and nulls
- Perform sampling and memory optimizations

**Tasks Completed**

- Defined project goals, KPIs, Workflow and identified performance metrics (e.g., average delay, cancellation rate, flight volume).
- Imported the dataset Flight_delay.csv using **Pandas**.
- Conducted exploratory data analysis:
    - Viewed dataset structure using head(), tail(), info(), and describe().
    - Checked column names, data types, and missing values.
- Removed **10 duplicate rows** to maintain data integrity.
- Performed **random sampling** for preliminary inspection (1% sample & first 100,000 rows).
- Optimized memory usage by:
    - Downcasting numeric columns.
    - Converting categorical features to category type.
- Verified memory reduction post-optimization.

**Skills Acquired**

- Efficient handling of large datasets in Pandas.
- Data quality assessment and optimization techniques.
- Understanding dataset schema and initial data profiling.

## Week 2: Preprocessing and Feature Engineering

### Objectives

- Handle nulls in delay and cancellation columns
- Create derived features: Month, Day of Week, Hour, Route
- Format datetime columns
- Save preprocessed data for fast reuse

### Tasks Completed

- **Missing Value Treatment:**
  - Filled missing values in categorical columns (e.g., Airline, TailNum, Origin, Dest) with *"Unknown"*.
  - Replaced missing numeric delay fields with 0.
  - Updated CancellationCode to "NotCancelled" where applicable.
  - Removed remaining incomplete rows.
- **Feature Engineering:**
  - Converted Date to datetime format.
  - Created new fields:
    - Month, DayOfWeek, Hour, and combined Route (Origin-Dest).
  - Added OnTime binary flag to mark punctual flights.
- Formatted Date column.
- Exported the processed dataset as **Flight_delay_cleaned.csv**.

### Skills Acquired

- Advanced preprocessing with Pandas.
- Time-series feature creation for delay trend analysis.
- Preparing clean, reusable datasets for further analytics.

# Milestone 2: Visual Exploration and Delay Trends

## Week 3: Univariate and Bivariate Visual Analysis

Objectives

- Top airlines, routes, and busiest months
- Flight distribution by day, time, and airport
- Plot bar charts, histograms, boxplots, and line plots

**Tasks Completed**

- Analyzed flight volumes by:
  - **Day, Hour, and Month** — Thursday and March emerged as the busiest.
- Identified **top airlines and routes**:
  - **Southwest Airlines** dominated flight count (~26.5% share).
  - **ORD–LGA** (Chicago–LaGuardia) was the most frequent route.
- Examined performance metrics:
  - **JetBlue Airways** had the highest average arrival delay (~70 min).
  - **Frontier Airlines** showed the lowest (~42 min).
  - **Southwest Airlines** had high volume but moderate delay.
- Verified correlation between **Distance** and **AirTime** (strong positive).
- Discovered that flight durations and delays were consistent across weekdays.

**Skills Acquired**

- Visual analysis interpretation (bar charts, scatterplots, heatmaps).
- Identifying operational inefficiencies and correlations.
- Airline and route-level performance comparison.

## Week 4: Delay Cause and Pattern Analysis

### Objectives

- Compare delay causes by airline
- Explore carrier delays, weather delays, NAS delays
- Visualize delays by time of day and airport

### Tasks Completed

- Identified **Carrier** and **Late Aircraft Delays** as the dominant contributors.
- Analyzed delay distribution:
  - Majority of delays near 0 minutes with significant extreme outliers.
- Studied **time-of-day impacts**:
  - Delays peak early morning (3–5 AM), dip during mid-day, rise again late night.
- Examined **airline and airport-level performance**:
  - **American Eagle Airlines** and **Frontier Airlines** faced severe "Late Aircraft" issues.
  - **O'Hare (ORD)** and **LAX** had peak delays in early morning.
- Found **weak correlation** among delay causes — each operates independently.

### Measures Suggested

- Add buffer times between early flights.
- Strengthen maintenance and ground operations.
- Create airline-specific delay management strategies.

### Skills Acquired

- Multi-factor delay analysis and visualization.
- Identifying temporal and operational delay trends.
- Strategy formulation for delay mitigation.

# Milestone 3: Route, Cancellation, and Seasonal Insights

## Week 5: Route and Airport-Level Analysis

### Objectives

- Top 10 origin-destination pairs
- Delay heatmaps by airport and route
- Maps showing busiest airports and average delays

### Tasks Completed

- **Busiest Routes:**
  - ORD–LGA and LGA–ORD had the most flights (~2,000 each).
- **Longest Delays:**
  - LGA–ORD (155 mins) and ORD–LGA (145 mins).
- **Airport-Level Findings:**
  - **ORD**, **DFW**, and **ATL** were the busiest origins.
  - **ACY (Atlantic City)** had the highest average delay (~380 mins).
- **Time-Based Delays:**
  - Early hours (4–6 AM) had highest delays across major airports.
- **Route Patterns:**
  - Longest delays seen on MDW–ATL and ATL–ORD routes.
  - Shortest on ORD–MDW and PHX–SFO.

### Measures Suggested

- Redistribute flights during early hours.
- Add operational buffers on delay-prone routes.
- Prioritize ground operations at high-traffic airports.

### Skills Acquired

- Geospatial and temporal route analysis.
- Multi-dimensional visualization interpretation.
- Data-driven operational optimization.

## Week 6: Seasonal and Cancellation Analysis

### Objectives

- Monthly cancellation trends
- Cancellation types: carrier, weather, security, NAS
- Analyze impact of holidays or winter months

### Tasks Completed

- **Monthly Trends:**
  - January (~20K) and December (~16K) had highest cancellations due to winter weather.
  - September had lowest cancellations.
- **Cancellation Causes:**
  - Carrier delays (40K+) were top reason, followed by NAS (~18–20K).
  - Weather and security had minimal effect.
- **Seasonal Comparison:**
  - Winter cancellation rate ≈0.23 vs. non-winter ≈0.10.
- **Airline Patterns:**
  - All airlines showed higher winter cancellations, especially JetBlue and Hawaiian.

### Measures Suggested

- Enhance winter readiness (de-icing, flexible scheduling).
- Improve communication during disruptions.
- Predictive analytics for early risk detection.

### Skills Acquired

- Seasonal trend analysis and visualization.
- Root-cause analysis of operational disruptions.
- Developing mitigation strategies for cancellations.

# Milestone 4: Report and Presentation

## Week 7: Dashboard Design and Visualization (Power BI)

### Objectives

- Combine plots into a coherent storyline
- Use markdown, presentation slides, or Streamlit
- Ensure plots include labels, titles, legends, and axis clarity

### Tasks Completed

- Built multi-page Power BI dashboards covering:
    1. **Dataset Overview** – total flights (483K), avg distance (753 miles), total delay (29M mins).
    2. **Delay Components** – Late Aircraft (26.7 mins) and Carrier (17.4 mins) dominate.
    3. **Airline Comparison** – JetBlue highest avg delay (72.9 mins); Frontier lowest (42 mins).
    4. **Airport & Route Map** – Major hubs (ORD, ATL, DFW) and route congestion.
    5. **Trend Dashboard** – Delay peaks in February and June.
    6. **Cancellation Analysis** – 0.09% overall cancellation rate; Hawaiian highest (0.10%).
    7. **Summary Dashboard** – Frontier best, JetBlue worst, Delta balanced.

### Skills Acquired

- Power BI dashboard design and KPI storytelling.
- Data integration and interactive visualization.
- Insight communication through data storytelling.

## Final Outcome

Over the 7-week period, the **Airfly Insights project** evolved from raw data exploration to a fully interactive analytical dashboard.
Key learnings include:

- Mastery of **data preprocessing, visualization, and interpretation**.
- Development of **insightful airline performance metrics**.
- Creation of a **comprehensive, data-driven reporting system** to guide operational decisions.