

# AirFly Insights – Airline Operations and Delay Analysis

## Infosys – Internship Program (Data Analytics & Visualization)

**Intern Name:** Sarthak Mokal

**Duration:** June 2025 – November 2025

---

### Introduction

The *AirFly Insights* project, part of the Infosys Internship Program in Data Analytics and Visualization, focuses on analyzing large-scale airline flight data to uncover operational trends, delay patterns, and cancellation reasons using data visualization techniques.

The goal is to derive actionable insights that help airlines and airports improve scheduling efficiency, reduce delays, and enhance overall reliability. The dataset, sourced from **Kaggle's U.S. Domestic Airlines Flights Data**, contains over **60 million flight records**; for this project, a refined subset of **484,551 flights** was analyzed.

---

# Milestone 1 – Week 1 & 2: Data Cleaning & Feature Engineering

## 1. Introduction

This project, *AirFly Insights*, aims to analyze large-scale airline flight data to uncover operational trends, delay patterns, and cancellation reasons through data visualization and analytics.

Milestone 1 focused on preparing a clean, consistent, and analysis-ready dataset for further visual exploration and modeling in subsequent milestones.

---

## 2. Dataset Overview

- **Source:** Kaggle – U.S. Domestic Airlines Flights Data
- **Records Loaded:** 484,551 rows
- **Columns (before cleaning):** 35
- **Columns (after feature engineering):** 44
- **Storage Path:**  
/Volumes/workspace/default/airlines/Flight\_delay\_cleaned\_final.csv

The dataset contains flight-level details such as flight date, airline carrier, origin and destination airports, scheduled departure/arrival times, delays (carrier, weather, NAS, security, late aircraft), and cancellation codes.

---

## 3. Key Performance Indicators (KPIs)

KPI	Description
Average Departure Delay	Mean delay (minutes) per flight
Average Arrival Delay	Arrival delay across carriers
Total Cancelled Flights	Count and % of cancelled flights
On-Time Performance Rate	% of flights $\leq$ 15 min delay
Top Busiest Routes	Flights per Origin–Destination pair
Peak Hour & Day	Flight volume by hour and weekday

---

## 4. Objectives and Tasks Completed

Milestone 1 objectives were to achieve full data readiness through systematic cleaning and feature engineering.

### Tasks Performed:

- **Data Acquisition:** Imported CSV via pandas and verified schema.
  - **Null Value Treatment:** Replaced missing delay values with 0 while retaining audit flags.
  - **Datetime Formatting:** Parsed and validated timestamps.
  - **Feature Engineering:** Derived columns — Month, DayOfWeek, DayName, DepHour, DepMinute, DepDate, Route.
  - **Data Type Optimization:** Downcast numeric types for efficiency.
  - **Duplicate Handling:** Removed two exact duplicates.
  - **Output Generation:** Saved as Flight\_delay\_cleaned\_final.csv.
- 

## 5. Methodology

1. Data Loading & Inspection – Checked datatypes and nulls.
2. Column Standardization – Renamed for consistency.
3. Datetime Engineering – Unified DepDatetime field.
4. Feature Creation – Added Month, DayOfWeek, Route.
5. Validation – Verified record counts.
6. Optimization – Downcast numeric columns.

All operations were executed in Databricks using Python (pandas, numpy).

---

## 6. Insights and Observations

- Dataset fully consistent; no broken timestamps.
- Time-based features enable trend analysis.
- Only 2 duplicates found in ~485 k rows.
- Delay columns converted to numeric for aggregation.
- Added features lay foundation for KPI visualization.

---

## 7. Data Dictionary

(FlightDate, ORIGIN, DEST, Route, DepDelay, ArrDelay, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay, Cancelled, Month, DayOfWeek, DepHour etc.)

---

## 8. Challenges and Resolutions

Challenge	Resolution
Inconsistent datetime formats	Unified parser for DepDatetime
Mixed data types	Coerced to numeric
Missing delay entries	Filled with 0 and audit flags
Duplicate records	Removed using drop_duplicates()

---

## 9. Tools and Libraries Used

pandas, numpy, Databricks Notebook, Python 3.10

---

## Milestone 2 – Week 3: Univariate & Bivariate Visual Analysis

---

### Objectives

1. Identify top-performing airlines, popular routes, and high-traffic airports.
  2. Explore flight-frequency patterns by day of week, departure hour, and route.
  3. Analyze delay trends across months, airlines, and weekdays.
  4. Visualize correlations among continuous delay factors.
  5. Represent operational metrics using diverse, clear visual formats for storytelling.
- 

### Graph Explanations & Insights

#### 1. Top Airlines by Flight Count (Bar Chart)

- **Explanation:** Bar chart shows the total flights per airline.
- **Insight:** Major carriers like *Delta*, *American*, and *Southwest* dominated operations — indicating a concentrated market share.

#### 2. Top Routes by Flight Count (Bar Chart)

- **Explanation:** Displays the ten most frequent origin-destination pairs.
- **Insight:** Routes between major hubs (e.g., *ATL–DFW*, *ORD–JFK*) had the highest frequencies, showing heavy domestic air traffic.

#### 3. Flight Distribution by Day & Hour (Count Plots)

- **Explanation:** Count plots show how flights are distributed across weekdays and hours.
- **Insights:**
  - *Days:* Mid-week (Tue–Thu) had steady operations; weekends slightly lower.
  - *Hours:* Two clear peaks — *early morning (6–9 AM)* and *evening (4–8 PM)* — matching business-travel demand.

#### 4. Average Departure Delay by Month (Line Chart)

- **Explanation:** Line graph compares mean delay times across months.

- **Insight:** Delays rise during *summer and holiday* months due to congestion and weather impacts.

5. Average Departure Delay by Airline (Horizontal Bar Chart)

- **Explanation:** Compares average departure delay among airlines.
- **Insight:** Some carriers show consistently higher delays, suggesting operational inefficiencies or congested routes.

6. On-Time vs Delayed Flights by Day (Grouped Bar Chart)

- **Explanation:** Compares proportions of on-time vs delayed flights for each weekday.
- **Insight:** *Weekends* have better punctuality, likely due to lighter traffic loads.

7. Delay Cause Contribution (Pie Chart)

- **Explanation:** Pie chart shows percentage of total delay minutes by cause.
- **Insights:**
  - *Carrier* and *Late Aircraft* delays dominate, showing internal scheduling issues.
  - *Weather* delays moderate, indicating predictable seasonal effects.

8. Correlation Heatmap of Delay Factors (Heatmap)

- **Explanation:** Displays correlation coefficients among delay variables.
- **Insights:**
  - *Departure* and *Arrival* delays highly correlated.
  - *Carrier* and *Late Aircraft* delays moderately correlated, hinting at cascading delay effects.

---

 **Overall Insights Summary**

Category	Key Takeaways
Airline Operations	Top 3 airlines handle majority of flights → market dominance.
Flight Timing	Morning & evening peaks reflect demand-based scheduling.
Delay Behavior	Carrier and late-aircraft delays are main contributors.

Category	Key Takeaways
On-Time Performance	~80–85 % flights on-time → good reliability.
Data Quality	Dataset clean and suitable for dashboards & forecasting.

---

## Conclusion

This milestone successfully applied **univariate and bivariate visual analysis** to uncover operational trends, delay patterns, and airline performance. The results provide a solid analytical foundation for future stages — particularly **Week 4's Delay Cause & Seasonal Analysis**, where deeper trend modeling and predictive analytics will refine operational insights.

---

## Milestone 2 – Week 4: Delay Cause and Seasonal Analysis

### Objectives

1. Analyze the distribution and contribution of delay causes (Carrier, Weather, NAS, Security, Late Aircraft).
  2. Compare airline-level delay composition using stacked and normalized bar charts.
  3. Identify delay trends by **month**, **hour**, and **airport** influenced by carrier and weather conditions.
  4. Visualize delay variability and outliers to assess operational consistency.
  5. Extend delay analysis to **top airlines and airports** across different time periods.
- 

### Graph Explanations & Insights

#### 1. Average Delay by Cause per Airline (Stacked Bar Chart)

- **Explanation:** Displays mean delay minutes of each cause (Carrier, Weather, NAS, etc.) per airline.
- **Insight:** *Carrier* and *Late Aircraft* delays dominated across most airlines, revealing internal operational inefficiencies and turnaround issues.

#### 2. Normalized Cause Proportion per Airline (100% Stacked Chart)

- **Explanation:** Shows relative percentage of each delay type per airline.
- **Insight:** Airlines with higher *Weather* and *NAS* delay proportions are likely impacted by specific regional and routing conditions.

#### 3. Mean Delay of Each Delay Type (Horizontal Bar Chart)

- **Explanation:** Compares average delay minutes among all delay categories.
- **Insight:** *Late Aircraft* and *Carrier* delays had the highest mean durations, reflecting compounding effects of prior delays.

#### 4. Delay Outliers by Cause (Boxplot)

- **Explanation:** Highlights variability and extreme values in delay times across causes.

- **Insight:** *Late Aircraft* delays had the widest spread and extreme outliers, signaling cascading and unpredictable delays.

## 5. Delay Distributions (Arrival & Departure) (Histogram + KDE)

- **Explanation:** Shows overall distribution and skewness of delay times.
- **Insight:** Both distributions were **right-skewed**, meaning most delays were short (<30 min), but a few extreme long delays occurred.

## 6. Monthly Delay Trend – Carrier vs Weather vs NAS (Multi-Line Chart)

- **Explanation:** Plots monthly average delays for the three major causes.
- **Insight:** *Weather delays* peaked during **winter and monsoon** months, while *Carrier* and *NAS* delays remained fairly stable.

## 7. Hourly Delay Heatmap (Cause × Hour)

- **Explanation:** Heatmap showing mean delay per cause across 24-hour time intervals.
- **Insight:** Peak delay hours were **17:00–21:00**, mainly driven by *Carrier* and *NAS* delays due to evening congestion.

## 8. Average Arrival Delay by Hour (Top 5 Airports) (Line Chart)

- **Explanation:** Compares hourly arrival delay trends for busiest airports.
- **Insight:** Airports like *ATL*, *DFW*, and *ORD* showed increased evening arrival delays due to high air traffic density.

## 9. Average Departure Delay by Hour (Top 5 Airlines) (Line Chart)

- **Explanation:** Shows departure delay variation across hours for major airlines.
- **Insight:** *Southwest* and *American Airlines* recorded higher departure delays between **4–8 PM**, aligning with peak operations.

## 10. Airport-Level Delay Cause Proportion (Stacked Bar Chart)

- **Explanation:** Shows how delay causes vary across top 10 origin airports.
  - **Insight:** Delay composition was influenced by **regional weather and congestion**, with weather-heavy airports facing more *NAS* and *Weather* delays.
-

### 💡 Overall Insights Summary

Category	Key Findings
Delay Causes	Carrier and Late Aircraft delays dominated overall delay time.
Seasonal Trends	Weather delays peaked in <b>winter and monsoon</b> seasons.
Time-of-Day Impact	Peak delays between <b>17:00–21:00</b> hours due to evening congestion.
Airport Influence	High-traffic airports experienced greater delay accumulation.
Operational Efficiency	Certain airlines consistently showed internal (Carrier) delays — potential for scheduling optimization.

---

### 🔍 Conclusion

Week 4 built upon Week 3’s exploratory analysis by diving into **delay causes, timing patterns, and airport-level dynamics**.

Through stacked, normalized, and time-based visualizations, it revealed that **Carrier and Late Aircraft delays** are the most influential factors, while **weather conditions** drive strong **seasonal variations**.

Evening peaks indicate **system congestion**, especially for major hubs and airlines.

These findings set the foundation for **predictive modeling (Milestone 3)** and **dashboard development (Milestone 4)** — enabling **data-driven strategies** to improve flight scheduling, delay management, and airline efficiency.

---

## Milestone 3 – Week 5: Route and Airport-Level Analysis

---

### Objectives

1. Identify the **Top 10 busiest origin–destination routes** by number of flights.
  2. Visualize **average departure delays** by route and airport using heatmaps.
  3. Map **busiest airports** and their **delay intensities** geographically.
  4. Compare **flight volumes** between top origin and destination airports.
  5. Assess the **correlation between distance and arrival delay**.
  6. Analyze **departure vs arrival delays** across different airports.
- 

### Graph Explanations & Insights

#### 1. Top 10 Origin–Destination Pairs by Number of Flights (Bar Chart)

- **Explanation:** Bar chart visualizing the most frequent routes in the dataset.
  - **Insights:**
    - *Chicago O'Hare → LaGuardia (NY)* had the highest flight count (~1900).
    - *LAX–SFO* and *SFO–LAX* were also major routes due to heavy business travel.
    - Hubs like *Dallas/Fort Worth*, *McCarran*, and *Phoenix* frequently appeared, indicating dense domestic connectivity.
- 

#### 2. Average Departure Delay by Route (Heatmap)

- **Explanation:** Heatmap displaying ranked average departure delays across routes.
- **Insights:**
  - *Ronald Reagan–Indianapolis* and *Eagle County–Miami* routes had the **highest delays (280–300 min)**.

- Several smaller regional routes (e.g., *LaGuardia–Yampa Valley*, *Richmond–Norfolk*) showed high delays.
  - Delay hotspots were mainly concentrated in **East Coast and Midwest regions**.
- 

### 3. Average Departure Delay by Origin Airport (Single-Axis Heatmap)

- **Explanation:** Visual representation of average delay per origin airport.
  - **Insights:**
    - *Atlantic City International Airport* had the **highest average delay (195.8 min)**.
    - *Eagle County* and *Abraham Lincoln Capital* followed with 103–100 minutes.
    - Smaller regional airports showed higher delays than large hubs, implying **resource and scheduling inefficiencies**.
- 

### 4. Geographic Visualization – Busiest Airports (Geo Map using Plotly)

- **Explanation:** Geo-mapped visualization showing airport locations by flight count and delay average.
  - **Insights:**
    - High-traffic clusters in **Midwest, East Coast, and California**.
    - Major hubs (*ATL*, *ORD*, *DFW*) maintained **moderate delay averages**, showing effective management.
    - Smaller airports had higher delays due to limited ground capacity.
- 

### 5. Flight Volume by Airport (Top Origin & Destination Airports – Bar Chart)

- **Explanation:** Comparison of total flights handled by top airports.
- **Insights:**
  - *Chicago O’Hare (ORD)*, *Dallas/Fort Worth (DFW)*, and *Atlanta (ATL)* led as both top origin and destination airports.
  - *LaGuardia (LGA)* and *Phoenix (PHX)* were also key destinations.

- Indicates **central hub dominance** in U.S. airline networks.

---

## 6. Correlation Between Distance and Arrival Delay (Scatter Plot)

- **Explanation:** Scatter plot testing relationship between flight distance and delay.
- **Insights:**
  - Correlation coefficient  **$r = 0.03$**  → negligible relationship.
  - Distance does **not affect delay**; factors like **airport congestion and scheduling** are more impactful.
  - Most flights experienced **moderate delays below 200 minutes**.

---

## 7. Departure vs Arrival Delay by Airport (Scatter Plot)

- **Explanation:** Compares average departure vs arrival delays for each airport.
- **Insights:**
  - Strong **positive correlation** between departure and arrival delays.
  - Airports close to the diagonal line showed **balanced propagation** (arrival  $\approx$  departure).
  - Airports above the line had **compounding delays**, caused by **airspace congestion** or **landing inefficiencies**.

---

### Overall Insights Summary

Category	Key Findings
Busiest Routes	Chicago–LaGuardia and LAX–SFO are the busiest flight corridors.
Delay Hotspots	Atlantic City, Eagle County, and Abraham Lincoln Capital airports recorded maximum delays.
Geographic Trends	Delay concentration strongest in <b>Midwest and East Coast</b> regions.
Major Hub Efficiency	ATL, ORD, and DFW manage large flight volumes with controlled delays.

Category	Key Findings
Correlation Patterns	Distance has minimal effect; departure delays directly influence arrival delays.
Operational Insight	Regional airports face higher delay variability due to limited capacity and scheduling inefficiencies.

---

### Conclusion

Week 5 extended analysis from airline-level delay patterns to **route and airport-level operational insights**.

Through **heatmaps, scatter plots, and geo-mapping**, this milestone highlighted key delay hotspots, busiest flight corridors, and performance contrasts between large hubs and small regional airports.

The findings confirm that **major hubs maintain operational control** despite heavy traffic, while **smaller regional airports** contribute disproportionately to total delay times due to limited infrastructure and turnaround inefficiencies.

This analysis provides a **spatial and operational foundation** for Week 6, focusing on **predictive modeling, geographic forecasting, and performance optimization** in the next milestone.

---

## Milestone 4 – Week 6: Flight Cancellation and Seasonal Delay Analysis

---

### Objectives

1. Identify **monthly and seasonal cancellation trends** across U.S. airlines.
  2. Examine **airline-wise and cause-wise cancellation patterns**.
  3. Evaluate the effect of **winter and holiday seasons** on delay and cancellation rates.
  4. Use **latitude-longitude mapping** to visualize geographical distribution of cancellations.
  5. Analyze relationships between **delay causes** and **cancellation frequency**.
- 

### Graph Explanations & Insights

#### 1. Monthly Flight Cancellation Trends (Bar Chart)

- **Explanation:** Bar chart showing total flight cancellations per month.
  - **Insights:**
    - **March** had the **highest cancellations (~4,700)**.
    - **January** and **February** also recorded high values due to **winter weather** and **traffic congestion**.
    - **April-May** had the fewest cancellations.
  - **Conclusion:** Early-year months (Jan-Mar) are the most operationally challenging.
- 

#### 2. Airline vs Month Heatmap

- **Explanation:** Heatmap comparing monthly cancellations across different airlines.

- **Insights:**
    - **Southwest** and **American Airlines** reported the most cancellations due to high flight volumes.
    - **Frontier** and **Hawaiian Airlines** had minimal cancellations.
  - **Conclusion:** Larger network size increases exposure to cancellations, not necessarily poor reliability.
- 

### 3. Cancellation Causes by Type (Bar Chart / Pie Chart)

- **Explanation:** Visualization showing percentage share of cancellation causes.
  - **Insights:**
    - **Carrier delays (50%)** dominate as the main cause, followed by **NAS (25%)** and **Weather (20%)**.
    - **Security delays (5%)** are negligible.
  - **Conclusion:** Internal airline inefficiencies cause more cancellations than external factors.
- 

### 4. Seasonal Comparison (Winter vs Non-Winter) (Table / Bar Chart)

- **Explanation:** Comparison of average departure delays and cancellation rates between winter and non-winter seasons.
  - **Insights:**
    - **Winter:** Avg Delay = 59 mins, Cancellation Rate = 5.1%.
    - **Non-Winter:** Avg Delay = 57 mins, Cancellation Rate = 4.9%.
  - **Conclusion:** Slight rise in delays and cancellations during **Dec–Feb**, consistent with weather and congestion effects.
- 

### 5. Holiday vs Non-Holiday Comparison (Bar Chart)

- **Explanation:** Comparison of cancellations during the holiday period (Dec 20–Jan 5) vs non-holiday days.
- **Insights:**
  - **Holiday period:** 5.1% cancellations vs **5.0% non-holiday**.

- Minor difference indicates **effective airline scheduling and planning**.
  - **Conclusion:** Airlines are well-prepared to handle holiday rush efficiently.
- 

## 6. Geo Map – Top 10 Busiest Airports (Plotly Map)

- **Explanation:** Geo-scatter map showing top airports by flight count and delay intensity.
  - **Insights:**
    - **ATL, ORD, and DFW** are busiest airports with **moderate delays**.
    - **JFK and LAX** show higher delay intensity despite strong performance.
  - **Conclusion:** Major hubs handle large volumes efficiently; smaller or coastal airports face sporadic high delays.
- 

## 7. Pie Chart – Cancelled Flights in Winter Months

- **Explanation:** Pie chart visualizing reasons for cancellations during winter.
  - **Insights:**
    - **Carrier** and **Weather-related** issues account for **~70%** of winter cancellations.
  - **Conclusion:** Weather amplifies carrier inefficiencies during peak winter months.
- 

## 8. Heatmap – Cancelled Flights by Month and Delay Types

- **Explanation:** Multi-dimensional heatmap comparing delay causes (Carrier, NAS, Weather, etc.) across months.
  - **Insights:**
    - **CarrierDelay** and **NASDelay** dominate throughout the year.
    - **Jan, Feb, Mar** show maximum intensity, aligning with high cancellation months.
  - **Conclusion:** High-delay months directly correspond to peak cancellation periods, confirming weather-traffic correlation.
-

## 💡 Overall Key Insights

Category	Insight
Seasonal Impact	Winter months have slightly higher delay and cancellation rates due to weather and congestion.
Major Cause	Carrier-related issues dominate cancellations, indicating operational inefficiencies.
Geographic Pattern	Large hubs (ATL, ORD, DFW) manage traffic efficiently, while smaller airports face recurrent issues.
Holiday Resilience	Airlines maintain strong performance during holiday seasons through advanced scheduling.
Delay Patterns	Carrier and NAS delays remain persistent throughout the year and correlate with high cancellation months.

---

## 🔍 Conclusion

Week 6 integrated **cause-based, time-based, and geographic analysis** to provide a comprehensive understanding of flight cancellations and delay behavior. Through visualizations such as **bar charts, heatmaps, and geo-maps**, it revealed that while **weather and traffic** influence performance, **carrier-level operational inefficiencies** remain the **primary drivers** of cancellations — especially during **winter and high-demand months**.

This milestone marks the final step of exploratory analysis, laying the groundwork for **predictive modeling and optimization** in future stages.

Airlines can leverage these findings to improve **scheduling resilience, delay mitigation**, and **seasonal performance forecasting**.

---