Milestone 2 – Week 4 Report

Delay Analysis – Airline and Weather

Project Title: AirFly Insights: Data Visualization and Analysis of Airline Operations
Intern Name: Sarthak Mokal
Organization: Infosys – Internship Program (Data Analytics & Visualization)
Milestone: 2 – Week 4

---

1. Introduction

The objective of Week 4 was to conduct *delay cause–specific and temporal visual analysis* to uncover deeper insights into airline operations.
While Week 3 focused on univariate and bivariate patterns such as flight counts, route popularity, and delay trends, this week expanded the analysis toward understanding **why** delays occur — their causes, severity, and temporal behavior.

This milestone aimed to explore delay contributions by airline, month, hour, and airport to identify operational inefficiencies, weather influences, and systemic delay propagation.

---

2. Objectives

The main objectives of Week 4 were:

- To analyze the **distribution and contribution of delay causes** (Carrier, Weather, NAS, Security, Late Aircraft).
- To evaluate **airline-level delay composition** through stacked and normalized comparisons.
- To identify **monthly and hourly delay trends** driven by carrier and weather conditions.
- To visualize **delay distributions and outliers** to understand spread and operational consistency.
- To extend analysis to **top-performing airlines and airports** across different time dimensions.

---

3. Tasks Completed

Task	Description
Average Delay by Cause per Airline	Stacked bar chart showing mean delay minutes by cause for each airline.
Normalized Cause Proportion by Airline	100% stacked chart displaying the relative share of delay causes per airline.
Mean Delay per Delay Type	Horizontal bar graph comparing average delay minutes for all causes.
Boxplot of Delay Types	Visualized spread and outliers for each delay cause to detect variability.
Departure & Arrival Delay Distributions	Histogram with KDE overlays to reveal distribution shape and skewness.
Monthly Trend: Carrier vs Weather vs NAS Delay	Multi-line comparison showing how major delay causes fluctuate across months.
Hourly Delay Heatmap (Cause vs Hour)	Illustrated mean delay of each cause across hours of the day (time-based variability).
Average Arrival Delay by Hour (Top 5 Airports)	Line chart comparing arrival delay trends by hour for busiest origin airports.
Average Departure Delay by Hour (Top 5 Airlines)	Line chart comparing departure delays by hour for top 5 airlines.
Airport-Level Cause Proportion	Stacked bar chart showing how delay causes vary among top 10 origin airports.

4. Methodology

1. **Data Preparation:**
   Loaded the cleaned dataset (Flight_delay_cleaned_final.csv) containing time, airline, and delay features.
2. **Feature Selection:**
   Focused on five delay types – *CarrierDelay, WeatherDelay, NASDelay, SecurityDelay,* and *LateAircraftDelay* – as primary metrics.
3. **Visualization Techniques:**
   - **Stacked and normalized bar plots** for cause-wise airline comparisons.
   - **Line charts** for time-based delay patterns (hourly, monthly).
   - **Histograms and boxplots** for distribution and outlier detection.
   - **Heatmaps** for multivariate hourly delay analysis.
4. **Tools Used:**
   - Python (pandas, matplotlib, seaborn)
   - Databricks Notebook for visualization and execution.

---

5. Visual Analysis and Insights

5.1 Average Delay by Cause per Airline

- Each airline showed distinct delay cause profiles.
- **Insight:** Carrier and Late Aircraft delays were dominant across most airlines, indicating internal operational and turnaround inefficiencies.

5.2 Normalized Cause Proportion per Airline

- Visualized proportional contribution (0–100%) of each delay cause.
- **Insight:** Some airlines had higher Weather and NAS delays, suggesting regional dependencies and routing factors.

5.3 Mean Delay of Each Delay Type

- Compared average delay minutes per cause using a bar graph.
- **Insight:** Late Aircraft and Carrier delays had the highest mean durations, reflecting cascading delay effects.

5.4 Delay Outliers by Cause (Boxplot)

- Boxplots revealed variability and extreme values for each cause.
- **Insight:** Late Aircraft delays showed wide spread and extreme outliers, emphasizing operational propagation issues.

5.5 Delay Distributions (Arrival & Departure)

- Histograms with KDE overlays illustrated distribution shape.
- **Insight:** Both Arrival and Departure delays were right-skewed, with a majority of flights under 30-minute delays but occasional extreme cases.

5.6 Monthly Delay Trend (Carrier vs Weather vs NAS)

- Multi-line graph tracked monthly averages of major delay causes.
- **Insight:** Weather delays spiked during winter and monsoon months, while Carrier and NAS delays remained relatively stable.

5.7 Hourly Delay Heatmap (Cause × Hour)

- Heatmap displayed hourly delay intensity by cause.
- **Insight:** Delay peaks were visible during evening hours (17:00–21:00), mainly driven by Carrier and NAS delays.

5.8 Average Arrival Delay by Hour (Top 5 Airports)

- Line chart compared hourly delay patterns across top origin airports.
- **Insight:** Major airports such as ATL, DFW, and ORD showed higher evening arrival delays due to high air traffic density.

5.9 Average Departure Delay by Hour (Top 5 Airlines)

- Analyzed hourly departure delay patterns by airline.
- **Insight:** Southwest and American Airlines exhibited higher departure delays between 16:00–20:00, aligning with peak operational load.

5.10 Airport-Level Delay Cause Proportion

- Stacked bar chart summarized delay cause share across top 10 airports.
- **Insight:** Regional weather conditions and airport congestion significantly influenced the proportion of Weather and NAS delays.

---

6. Summary of Key Insights

Category	Summary of Findings
Delay Causes	Carrier and Late Aircraft delays remained dominant across airlines.
Seasonal Impact	Weather delays showed clear seasonal trends (winter and monsoon peaks).
Temporal Variation	Peak delays occurred between 17:00–21:00 hours due to evening congestion.
Airport Influence	Busiest airports recorded higher cumulative delays and varied cause composition.
Operational Efficiency	Some airlines consistently showed higher internal (Carrier) delays, highlighting potential scheduling inefficiencies.

---

7. Tools and Libraries Used

- **Python:** pandas, numpy
- **Visualization:** matplotlib, seaborn
- **Environment:** Databricks Notebook
- **Dataset:** Kaggle Airlines Flights Data

---

8. Conclusion

Week 4 successfully expanded upon Week 3's findings by focusing on *delay causes, time-based variation, and operational insights*.
Through stacked, normalized, and temporal visualizations, this milestone identified the root contributors to delays and highlighted seasonal and hourly behavioral trends.

The outcomes of this analysis provide valuable direction for **predictive modeling (Milestone 3)** and **dashboard development (Milestone 4)**, enabling data-driven performance optimization for airlines and airports alike.