Milestone 2 – Week 3 Report

Univariate and Bivariate Visual Analysis

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 3

---

1. Introduction

The purpose of this milestone was to perform **univariate and bivariate visual analysis** on the cleaned airline dataset to explore operational trends, route popularity, airline performance, and delay behavior.

This stage focuses on **visual storytelling** — transforming raw numbers into insights through exploratory data visualization. Using Python libraries like **pandas**, **matplotlib**, and **seaborn**, multiple patterns were identified, such as flight distributions, delay trends, and on-time performance variations.

---

2. Objectives

The key objectives of this week's analysis were:

- To identify **top-performing airlines**, **popular routes**, and **frequent airports**
- To explore **flight frequency patterns** by day of week, departure hour, and route
- To analyze **delay patterns** across months, airlines, and days
- To visualize **correlations** among continuous delay factors
- To represent operational metrics using clear and diverse visual formats

---

3. Tasks Completed

Task	Description
Top Airlines Analysis	Identified the top 10 airlines by total flight count using bar plots.
Top Routes Identification	Found the busiest routes based on flight frequency.
Airport Analysis	Determined the top 10 origin airports handling the highest number of flights.
Time-Based Analysis	Examined flight distribution by day of the week and by hour of departure.
Delay Analysis	Compared average departure delay by month, airline, and day of week.
On-Time Performance	Segmented flights as "On-Time" vs "Delayed" for performance insights.
Delay Cause Visualization	Plotted a pie chart showing the percentage contribution of each delay type (Carrier, Weather, NAS, etc.).
Correlation Heatmap	Explored relationships among continuous delay variables.

---

4. Methodology

1. **Data Loading and Validation** – Imported the cleaned CSV using pandas and verified column structure, nulls, and datatypes.
2. **Univariate Analysis** – Used count plots and bar charts to show distribution of flights across airlines, routes, days, and airports.
3. **Bivariate Analysis** – Plotted line and box plots to compare average delays across time and categories.
4. **Pie Chart Visualization** – Added percentage-based insights for delay causes.
5. **Correlation Study** – Generated a heatmap to examine interdependence among delay factors.

Each visualization was created using **matplotlib** and **seaborn** for clarity, consistency, and analytical readability.

---

5. Visual Analysis and Insights

5.1 Top Airlines by Flight Count

- Bar chart displayed top airlines with maximum flight operations.
- *Insight:* Major carriers such as Delta, American, and Southwest dominated total flights, showing market concentration.

5.2 Top Routes by Flight Count

- Visualized the 10 most frequent origin–destination pairs.
- *Insight:* Routes connecting major hubs (like ATL–DFW, ORD–JFK) recorded the highest frequency, indicating heavy domestic demand.

5.3 Flight Distribution by Day and Hour

- Count plots were created for both day-wise and hourly flight distributions.
- *Insight:*
    - **Days:** Midweek days (Tuesday–Thursday) showed consistent flight activity, while weekends were comparatively lighter.
    - **Hours:** Two peaks were observed — early morning (6–9 AM) and evening (4–8 PM).

5.4 Average Departure Delay by Month

- Line plot depicted average departure delay variation over months.
- *Insight:* Delays tended to increase slightly during summer and holiday months, suggesting higher congestion and weather sensitivity.

5.5 Average Departure Delay by Airline

- Compared mean delay across airlines using horizontal bar plots.
- *Insight:* A few airlines exhibited higher average delays, indicating potential operational inefficiencies or congested routes.

5.6 On-Time vs Delayed Flights by Day of Week

- Grouped bar plot showed the balance between on-time and delayed flights for each weekday.
- *Insight:* Weekends showed a better on-time percentage compared to weekdays, possibly due to reduced traffic.

5.7 Delay Cause Contribution (Pie Chart)

- Pie chart illustrated total delay minutes distributed across Carrier, Weather, NAS, Security, and Late Aircraft delays.
- *Insight:*
    - **Carrier Delays** and **Late Aircraft Delays** formed the largest portions, showing internal airline and turnaround dependencies.
    - **Weather Delays** contributed moderately, reflecting predictable seasonal effects.

5.8 Correlation Heatmap of Delay Factors

- Heatmap represented correlation coefficients among numeric delay columns.
- *Insight:*
    - **Departure and Arrival Delays** were strongly correlated.
    - **Carrier and Late Aircraft Delays** showed moderate correlation, suggesting chained delay effects.

---

6. Summary of Key Insights

Category	Insight Summary
Airline Operations	Top 3 airlines handled the majority of flights, indicating market dominance.
Flight Timing	Morning and evening peaks confirm strategic scheduling for high-demand slots.
Delay Behavior	Carrier and Late Aircraft delays were primary contributors to total delay minutes.
On-Time Performance	Around 80–85% of flights were on-time, showing good operational reliability.
Data Completeness	Dataset represents high-quality, clean data suitable for dashboarding and forecasting.

---

7. Tools and Libraries Used

- **Python** (pandas, numpy) – Data processing and feature extraction
- **matplotlib & seaborn** – Visualization and trend analysis
- **Databricks Notebook** – Environment for execution and data management
- **CSV Dataset (Kaggle)** – Primary data source

---

8. Conclusion

This milestone successfully demonstrated **exploratory visual analysis** on the airline dataset using univariate and bivariate techniques.
The insights derived provide a clear understanding of airline performance, operational bottlenecks, and delay dynamics.

These findings form a strong analytical base for **Week 4 (Delay Cause & Seasonal Analysis)**, where deeper cause-specific trends and predictive modeling can be developed.
