Milestone 3 – Week 5 Report

Route and Airport-Level Analysis

Project Title: AirFly Insights: Data Visualization and Analysis of Airline Operations
Intern Name: *Sarthak Mokal*
Organization: Infosys – Internship Program (Data Analytics & Visualization)
Milestone: 3 – Week 5

1. Introduction

The focus of **Week 5** was to extend the delay analysis from airline-level trends (covered in Week 4) to a **route and airport-level perspective**.
While previous milestones emphasized delay causes such as weather or carrier inefficiencies, this phase aimed to uncover **where** and **along which routes** these inefficiencies were most prominent.

The objective was to identify **busiest routes**, analyze **average delays by origin and destination airports**, and visualize **flight network congestion** geographically.
This analysis provided a spatial understanding of delay distribution and operational efficiency across the U.S. air transport network.

2. Objectives

The main objectives for Week 5 were:

- To identify the **Top 10 origin–destination pairs** by number of flights.
- To visualize **average departure delays** by route and airport using heatmaps.
- To map **busiest airports** and analyze their average delay intensity.
- To compare **flight volumes** across top origin and destination airports.
- To assess the **correlation between distance and arrival delay**.
- To analyze the **relationship between departure and arrival delays** across airports.

3. Tasks Completed

Task	Description
Top 10 Routes	Identified top 10 origin–destination pairs by number of flights using bar chart visualization.
Delay Heatmap by Route	Created ranked heatmap of average departure delays by each origin–destination route.
Average Delay by Airport	Visualized average departure delay per origin airport in a single-axis heatmap.
Geographic Visualization	Mapped busiest U.S. airports by flight count and average departure delay using Plotly.
Flight Volume by Airport	Compared top 10 origin and destination airports by total flights handled.
Distance vs Arrival Delay	Examined correlation between flight distance and arrival delay using scatter plot.
Departure vs Arrival Delay	Analyzed airport-level comparison between average departure and arrival delays.

4. Methodology

1. **Data Preparation:**
   Loaded the cleaned dataset (Flight_delay_cleaned_final.csv) and verified key columns such as Org_Airport, Dest_Airport, DepDelay, ArrDelay, and Distance.

2. **Feature Engineering:**
   Created a combined Route column (Org_Airport → Dest_Airport) and added latitude/longitude values for mapping.

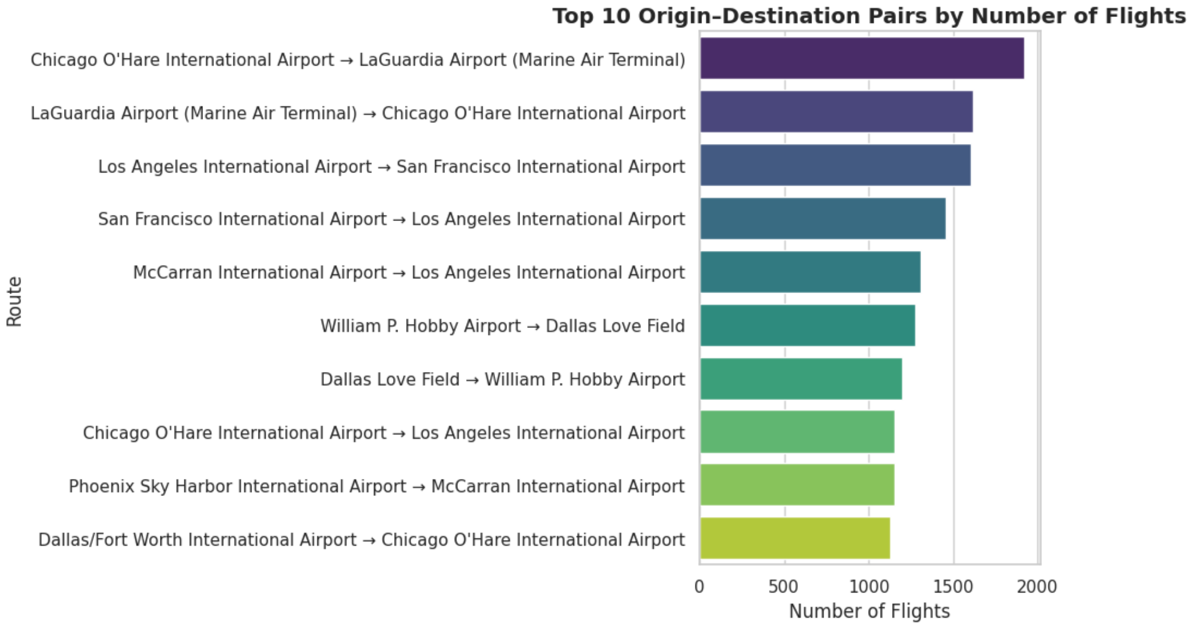
3. **Visualization Techniques:**
   Used bar charts, single-axis heatmaps, scatter plots, and geo-maps to analyze route- and airport-level delay patterns.

4. **Tools and Libraries:**

   - Python: pandas, numpy
   - Visualization: matplotlib, seaborn, plotly.express
   - Environment: Databricks Notebook

5. Visual Analysis and Insights

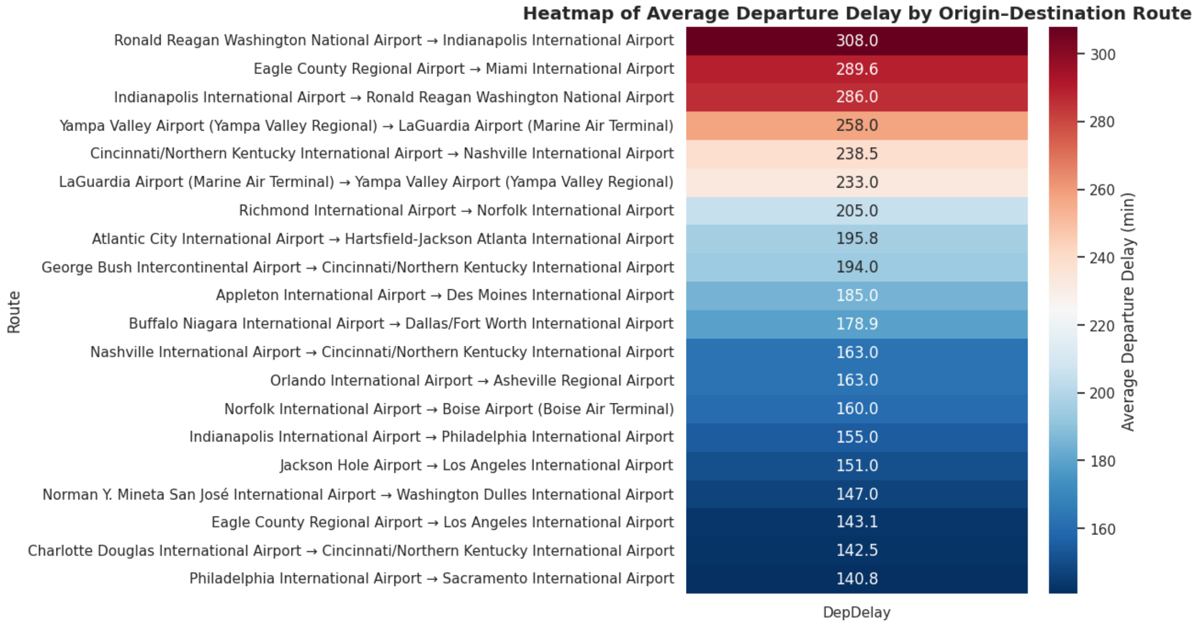
5.1 Top 10 Origin–Destination Pairs by Number of Flights



Insights:

- The **Chicago O'Hare → LaGuardia (New York)** route recorded the highest number of flights (~1900).
- **LAX–SFO** and **San Francisco–LAX** routes were also among the most active, indicating high bi-directional business travel.
- Major hubs such as **Dallas/Fort Worth**, **McCarran**, and **Phoenix** frequently appear in the busiest routes list.

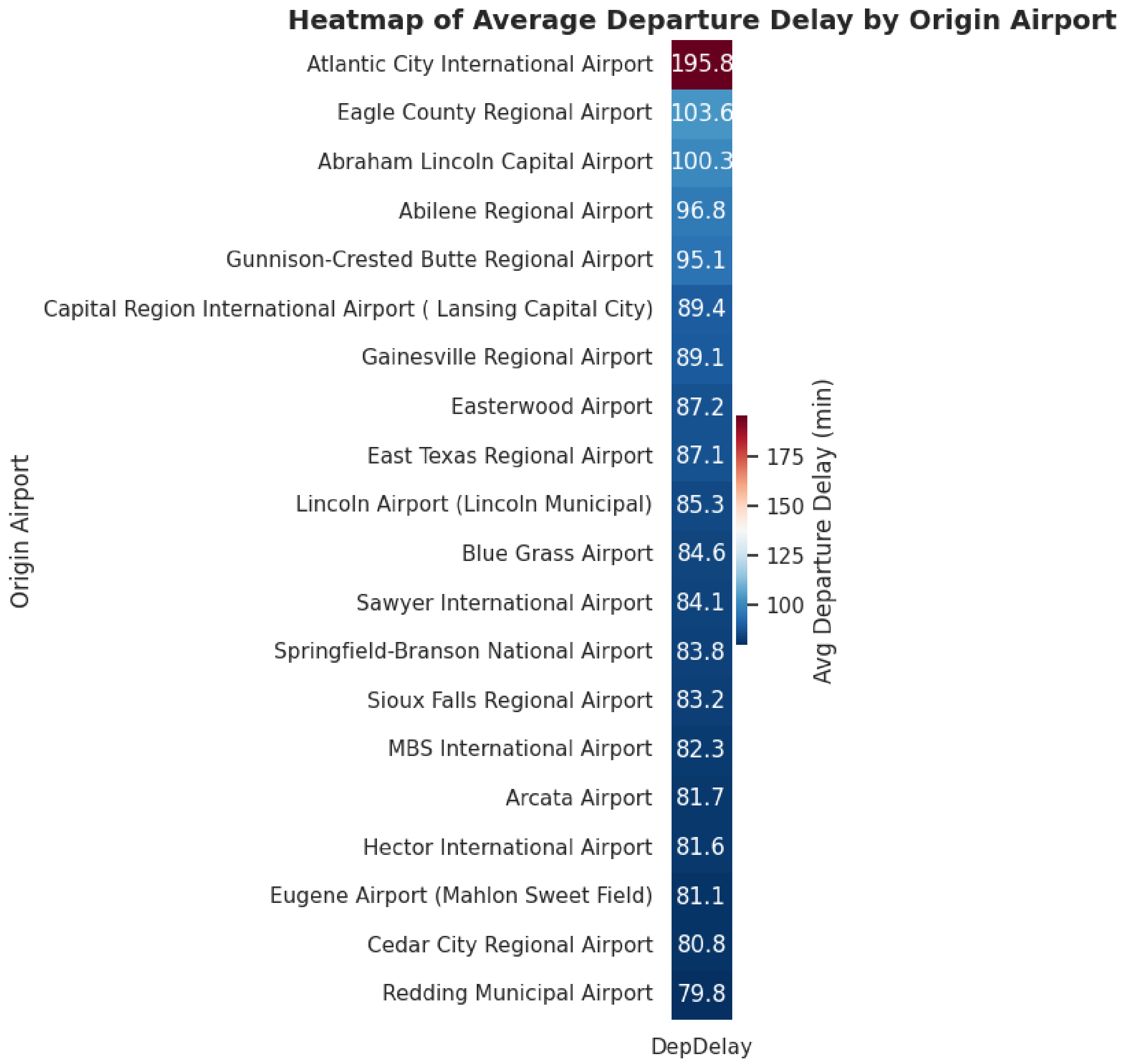
5.2 Heatmap of Average Departure Delay by Route



Insights:

- **Ronald Reagan–Indianapolis** and **Eagle County–Miami** routes showed the **highest average delays (280–300 min)**.
- Several regional routes such as **LaGuardia–Yampa Valley** and **Richmond–Norfolk** also reported high delays, indicating regional operational challenges.
- Routes with high delay intensity were concentrated around **East Coast and Midwest airports**.

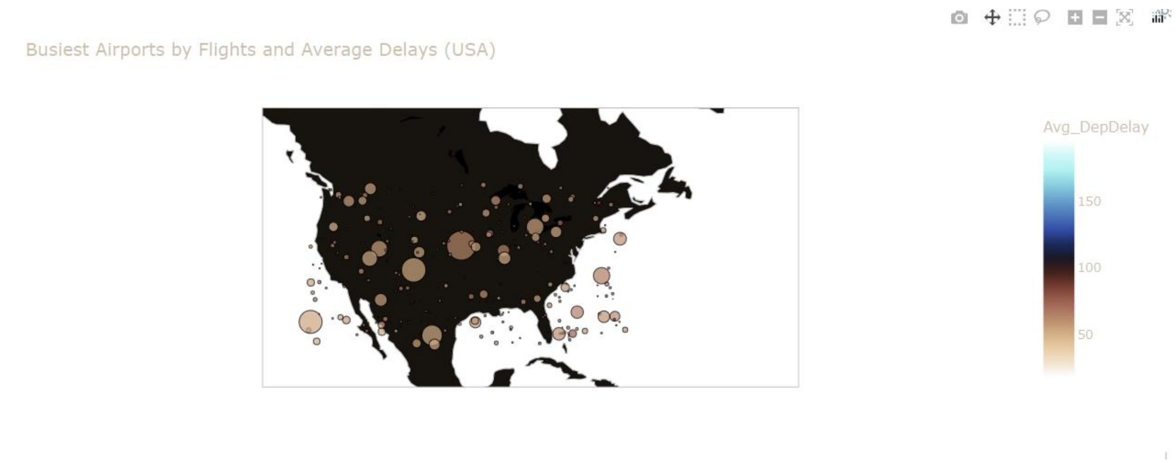
5.3 Heatmap of Average Departure Delay by Origin Airport



Insights:

- **Atlantic City International Airport** recorded the **highest delay (195.8 min)**, far exceeding the average.
- Other high-delay airports included **Eagle County (103.6 min)** and **Abraham Lincoln Capital (100.3 min)**.
- Many smaller regional airports displayed larger average delays than major hubs, suggesting **resource and scheduling inefficiencies**.

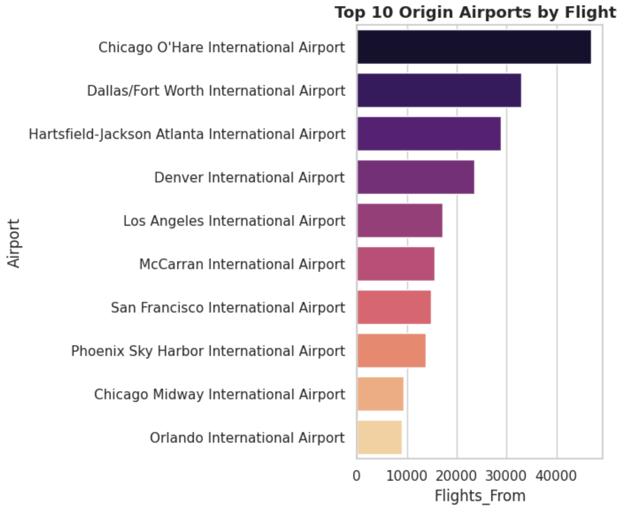
5.4 Geographic Visualization – Busiest Airports by Flights and Average Delays

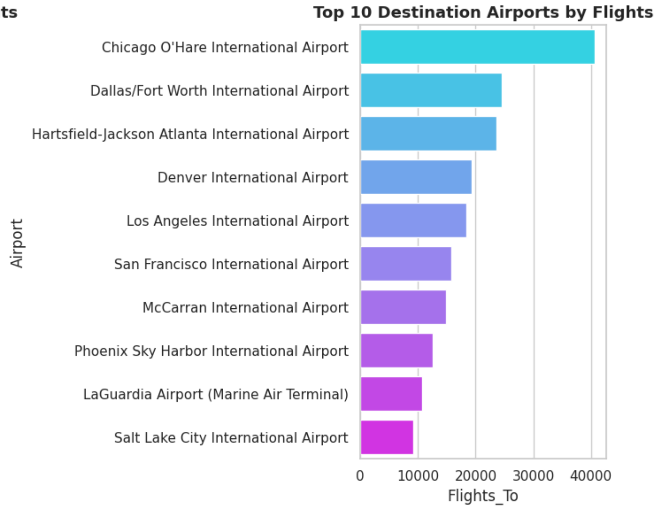


Insights:

- Clusters of **high-traffic airports** were visible in the **Midwest, East Coast, and California**.
- Despite heavy traffic, major hubs like **ATL**, **ORD**, and **DFW** maintained moderate delay averages, showing better operational control.
- Outlier airports with large average delays were smaller regional airports with limited handling capacity.

5.5 Airport-Wise Flight Volume (Top Origin & Destination Airports)

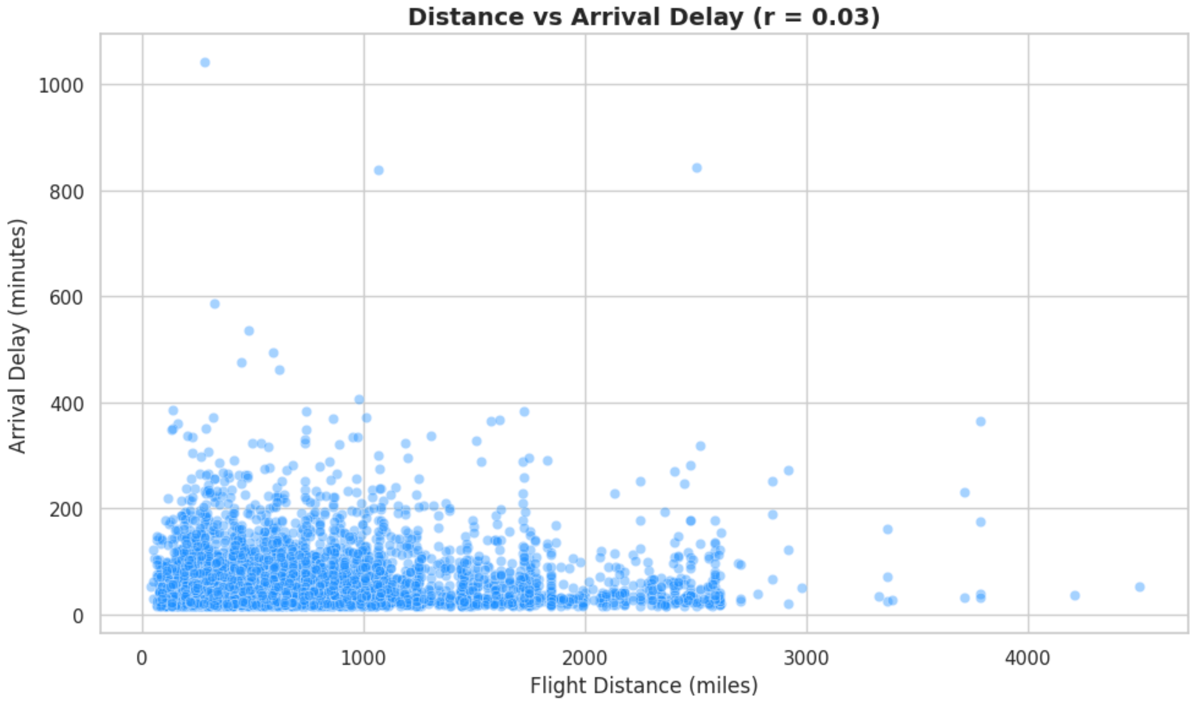




Insights:

- **Chicago O’Hare**, **Dallas/Fort Worth**, and **Atlanta** were the **top origin airports** by flight count.
- The same airports also dominated as **top destinations**, reflecting their central hub roles.
- **LaGuardia** and **Phoenix** featured strongly among top destinations due to dense domestic connectivity.

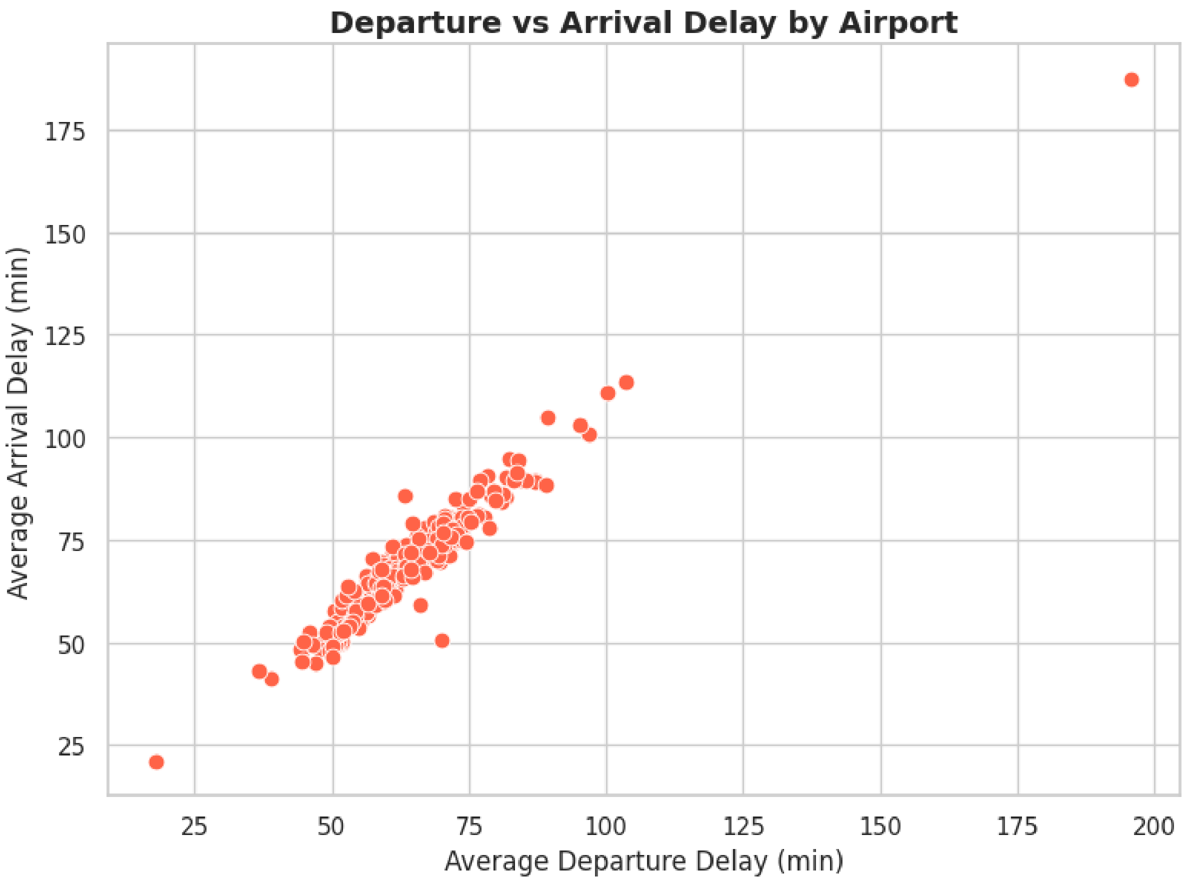
5.6 Correlation Between Distance and Arrival Delay



Insights:

- The correlation coefficient (**r = 0.03**) indicated **no significant relationship** between distance and delay.
- Long-haul flights did not necessarily face longer delays; instead, delays were **primarily influenced by airport congestion and scheduling factors**.
- Most flights experienced moderate delays below 200 minutes.

5.7 Departure vs Arrival Delay by Airport



Insights:

- A **strong positive correlation** between departure and arrival delays was observed.
- Airports near the diagonal line showed balanced delay propagation (arrival delay ≈ departure delay).
- Airports above the line exhibited **compounding mid-air or landing delays**, suggesting airspace congestion or inefficient arrival sequencing.

6. Summary of Key Insights

Category	Summary of Findings
Busiest Routes	Chicago–LaGuardia and LAX–SFO are the most active routes in the dataset.
Delay Hotspots	Atlantic City, Eagle County, and Abraham Lincoln Capital airports recorded the highest average delays.
Geographic Pattern	Delay intensity is concentrated in the Midwest and East Coast corridors.
Efficiency of Major Hubs	Major airports like ATL, ORD, and DFW handle high flight volumes with relatively low delay averages.
Correlation Trends	Distance has minimal correlation with delay; departure delays directly impact arrival delays.
Operational Insight	Smaller regional airports face greater operational inefficiencies and variability in delay times.

7. Tools and Libraries Used

- **Python Libraries:** pandas, numpy, matplotlib, seaborn, plotly
- **Environment:** Databricks Free Edition
- **Dataset:** Flight_delay_cleaned_final.csv (Kaggle Airlines Dataset)
- **Visualization Techniques:** Bar charts, Heatmaps, Scatter plots, Geo visualization

8. Conclusion

Week 5 successfully transitioned the analysis from delay cause identification to **route and airport-level performance evaluation**.
Through visualizations such as bar charts, heatmaps, and geo-maps, the analysis revealed key **traffic patterns, delay hotspots, and efficiency disparities** among U.S. airports.

The findings indicate that **major hubs maintain better delay control** despite large traffic volumes, while **smaller regional airports** contribute disproportionately to average delay times.