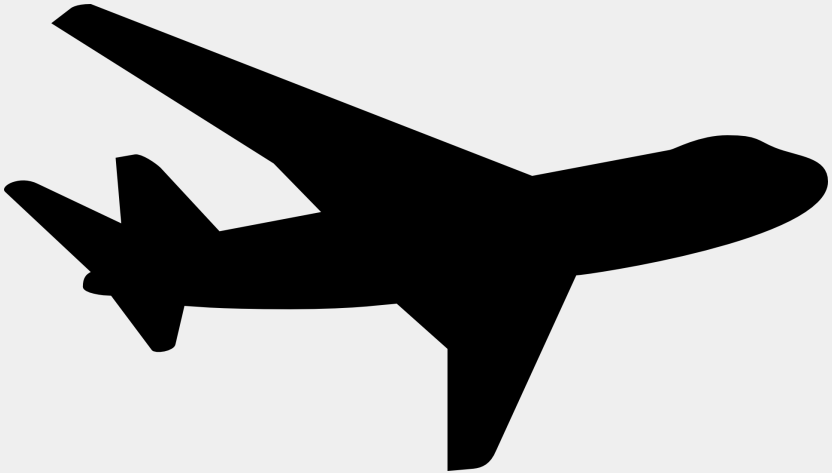
Flight Insights from AviationStack Data

ETL Pipeline Development with AviationStack API

Melissa Lau

Table of Contents



Project Goals

- ❏ Design a Python-based ETL pipeline that automates flight data collection and is adaptable for both personal and business travel insights.
- ❏ Design an interactive dashboard for exploring flight data across time, location, and airline
- ❏ Showcase pipeline’s utility through practical analytical use cases

README

Project Overview

This is a Python-based ETL pipeline that extracts real-time and historical flight data from the AviationStack API (Paid Plan), stores it in a SQL database, and visualizes key travel insights in Streamlit dashboard. The system is designed to support both personal travel planning and business-level operational insights through analysis of flight delays, airline performance, and airport activity.

Tools Used

Python
SQL
Streamlit
AviationStack API

Key Features

- **Query Flights by Date(s) and Airport(s)**
  Supports querying by date/date range, departure, and/or arrival airports, and airlines.

- **Historical Data Collection**
  Note: AviationStack only provides access to historical flight data from the past 3 months. For long-term insights, the pipeline must run periodically (e.g., daily or weekly) to build a historical dataset.

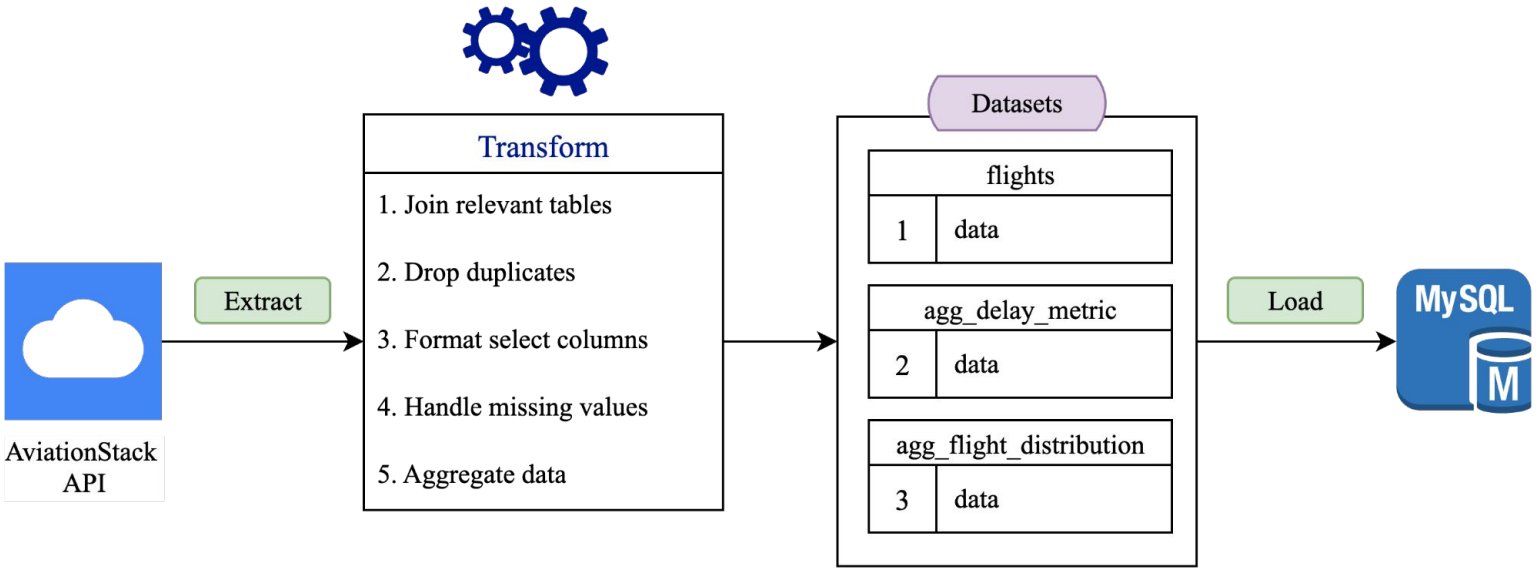
- **Visual Insights with Streamlit Dashboards**

  - View flight delays and on-time performance
  - Explore trends by airline, airport, and time

Want a Deeper Dive?

Check out my [Github repo]

Schematic of ETL Pipeline



Customizable Parameters

Data Retrieval

Parameters	Definition
`access_key`	AviationStack API Key
`start_date`	Start date in 'YYYY-MM-DD' format
`end_date`	End date in 'YYYY-MM-DD' format
`dep_code_list`	IATA code(s) of departure airport (e.g., SFO) (*optional*)
`arr_code_list`	IATA code(s) of arrival airport (*optional*)
`airline_code_list`	List of airline IATA codes (*optional*)

While `dep_iata_code` and `arr_iata_code` are optional parameters, at least one of them **must be provided** to successfully retrieve data.

MySQL Database Connection

Parameters	Definition
`username`	MySQL username
`database_password`	Database password
`hostname`	Hostname or IP address
`port`	Port number
`database_name`	Name of database to store the data

Optional Parameters

Parameters	Definition
`create_database`	Set to True (*default*) to create a new database or False to use an existing one
`if_exists`	Choose "append" to add new data to existing table, or "replace" to overwrite the table with new data.
`version_tag`	Optional tag or suffix to distinguish tables

USE CASE 1: Delta Airline Monthly Performance

1. Objective
2. Dashboard Overview
3. June Report Results
4. Conclusions and Recommendations

Objective

Provide data-driven insights into Delta's flight performance using an interactive dashboard and recommend strategies to improve its performance.

Approach

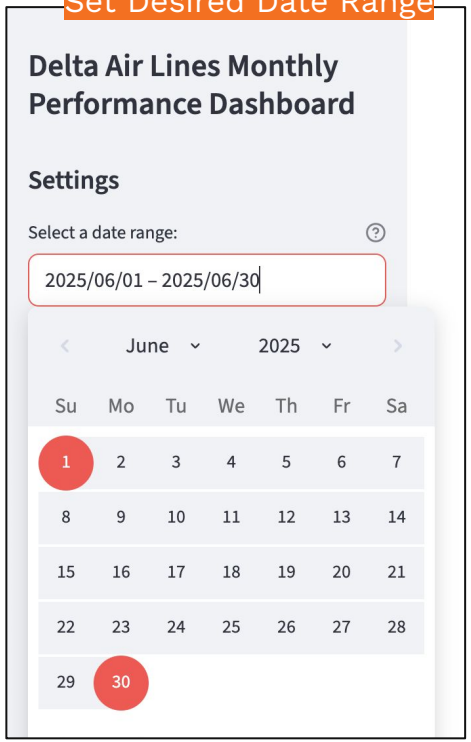
- **Data Sources:** Historical flight performance data including delay metrics, airports, routes, airlines, and time-of-day patterns. For this section, I will focus on Delta flights departing from 12 of U.S. busiest airports for the month of June (size of data: **194,237**). United Airlines and American Airlines will be included as benchmark performance comparison.

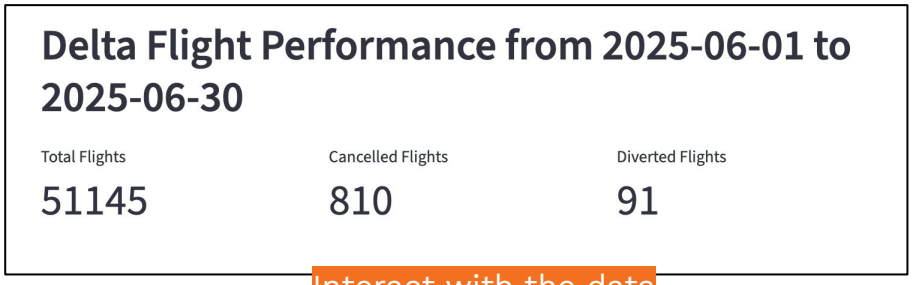
- **Pipeline Parameters:**
  - `dep_iata_code = ['ATL', 'DFW', 'DEN', 'ORD', 'LAX', 'CLT', 'MCO', 'LAS', 'MIA', 'PHX', "JFK", "SFO"]`
  - `start_date = "2025-06-01"`
  - `end_date = "2025-06-30"`
  - `airline_code_list = ["DL", "UA", "AA"]`

- **Analysis:**
  1. Departure punctuality (% on-time)
  2. Delay trends
  3. Cancellations
  4. Flight volume patterns

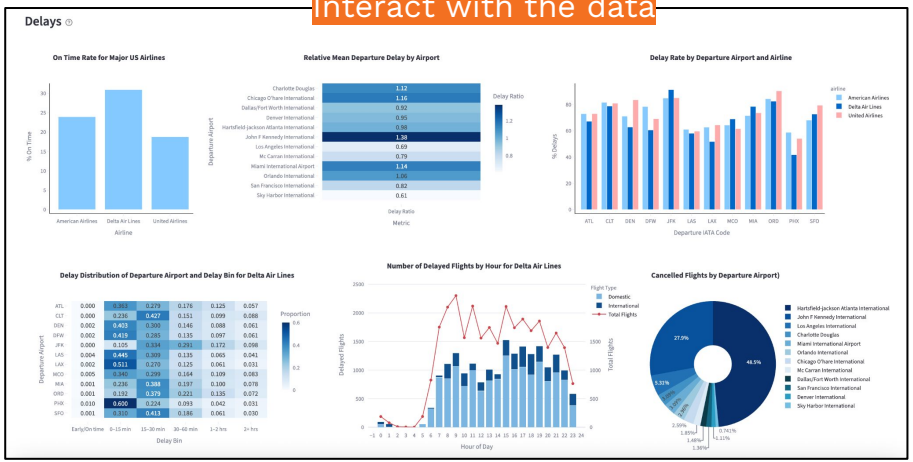
Dashboard Preview

Set Desired Date Range



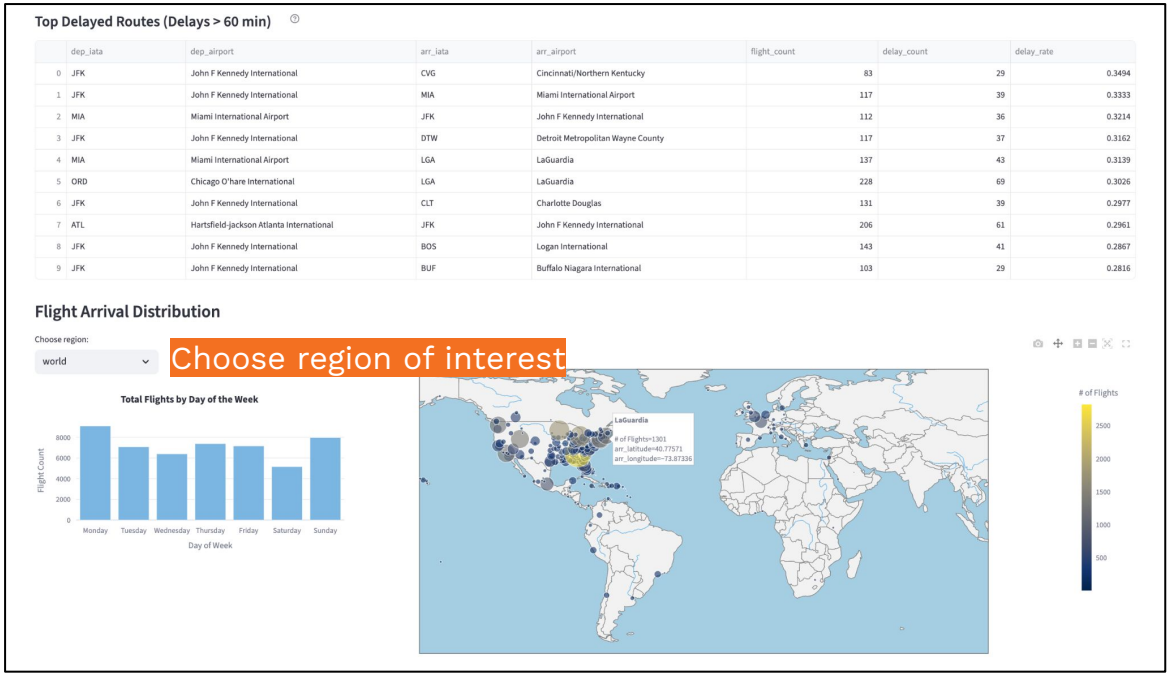


Interact with the data



Dashboard Preview

2 of 2



Delta’s Key Metrics Dropped on Week 3

Overview of Key Metrics

Drop in Week 3 performance likely due to Juneteenth holiday, and spike in cancellation flights
Week 4 is likely due to Atlanta storms.

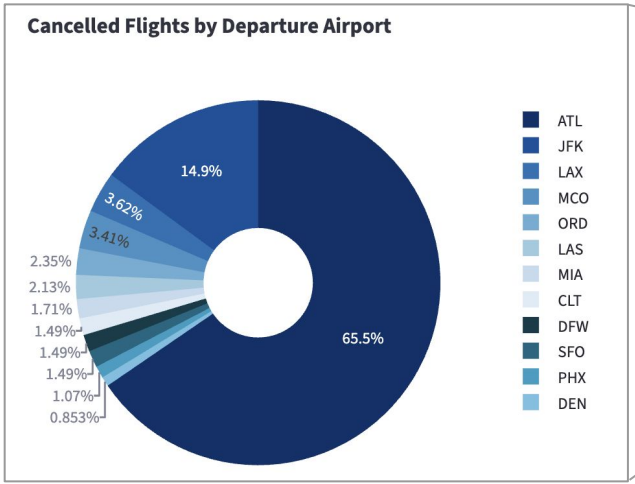
	Total Flights	Cancelled Flights	On Time Rate
Week 1 (6/2 - 6/8)	12691	110	38.49%
Week 2 (6/9 - 6/15)	12736 ▲ 0.35%	61 ▼ 44.55%	27.25% ▼ 29.20%
Week 3 (6/16 - 6/22)	9518 ▼ 25.27%	134 ▲ 119.67%	23.13% ▼ 15.12%
Week 4 (6/23 - 6/29)	12660 ▲ 33.01%	469 ▲ 250.0%	30.20% ▲ 30.57%

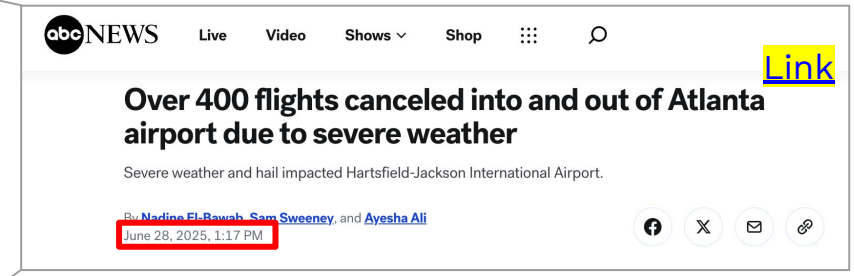
Table 1. Summary of Key Metrics: Key metrics for evaluating Delta’s June 2025 performance include total flights, cancellations, and on-time rates.

Severe storms in Atlanta were Likely the Primary Driver of Delta's Spike in Flight Cancellations During Week 4

Week 4 flight cancellations from Atlanta **exceeded** the total Atlanta-related cancellations from all prior weeks combined.

June 23 - June 29





Link

*"**Delta Air Lines**, which has a **major hub at Atlanta**, is experiencing the most significant impact from the severe weather, with **542 cancellations and 684 delays on Saturday** in total across the country. Delta said it expects additional delays and cancellations as it works to recover from **Friday's storms in Atlanta**, with teams resetting aircraft and flight crews completing required rest." (El-Bawab, Sweeney, and Ali).*

Delta Has Better Relative Performance but Experiences More Delays at JFK

On Time: < 15 min departure time

Delta experiences the highest departure delays from John F Kennedy Airport.

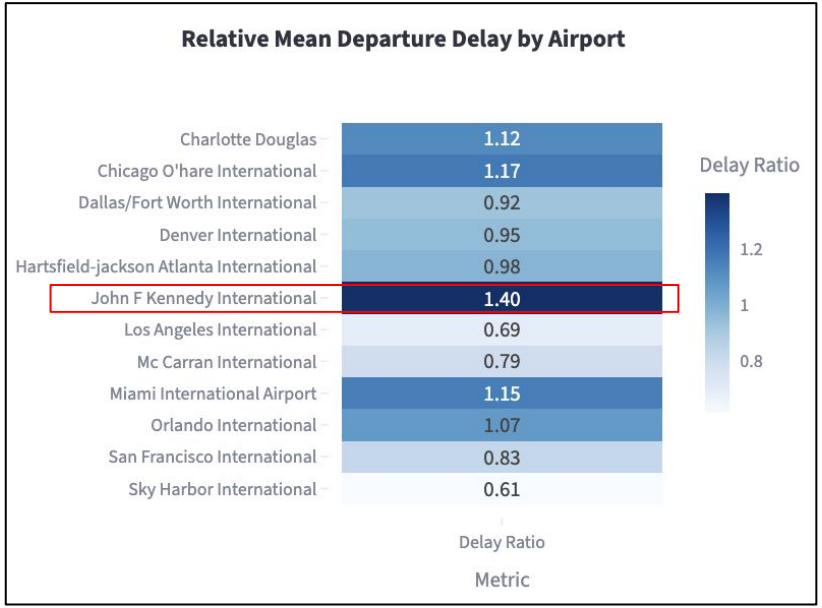


Figure 1. Relative Mean Departure Delay by Airport: The relative mean departure delay is calculated as the mean delay at each airport divided by the overall mean delay across all airports. A value greater than 1 indicates that the airport experiences above-average delays, while a value less than 1 indicates below-average delays.

Compared to American and United, Delta has the fewest delays at 8 of the 12 busiest U.S. airports.

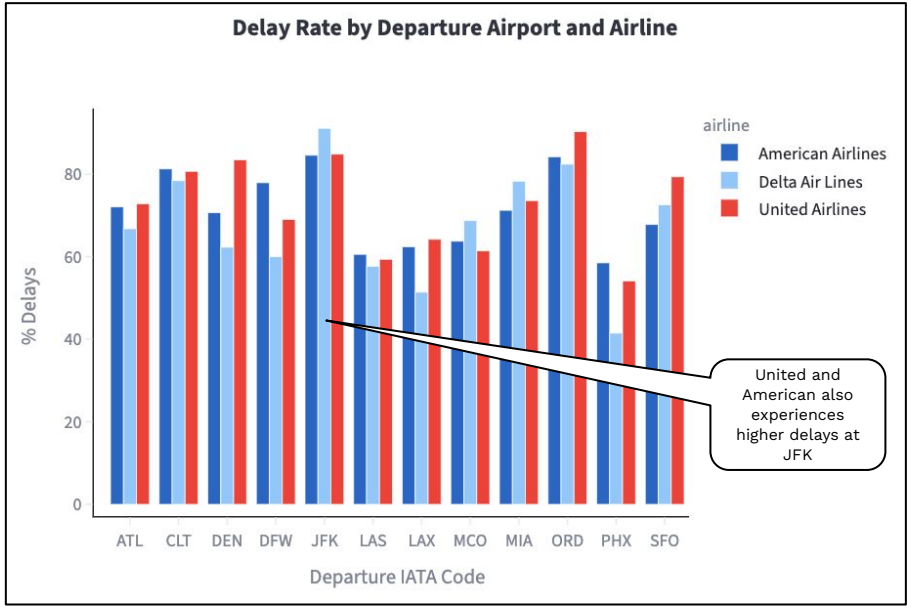


Figure 2. Delay Rate by Airport and Airline: Percent delay is calculated as the number of flights with departure delays of over 15 minutes divided by the total flights departing from that airport.

Flight Delays Occur More Often Past 2:00 PM

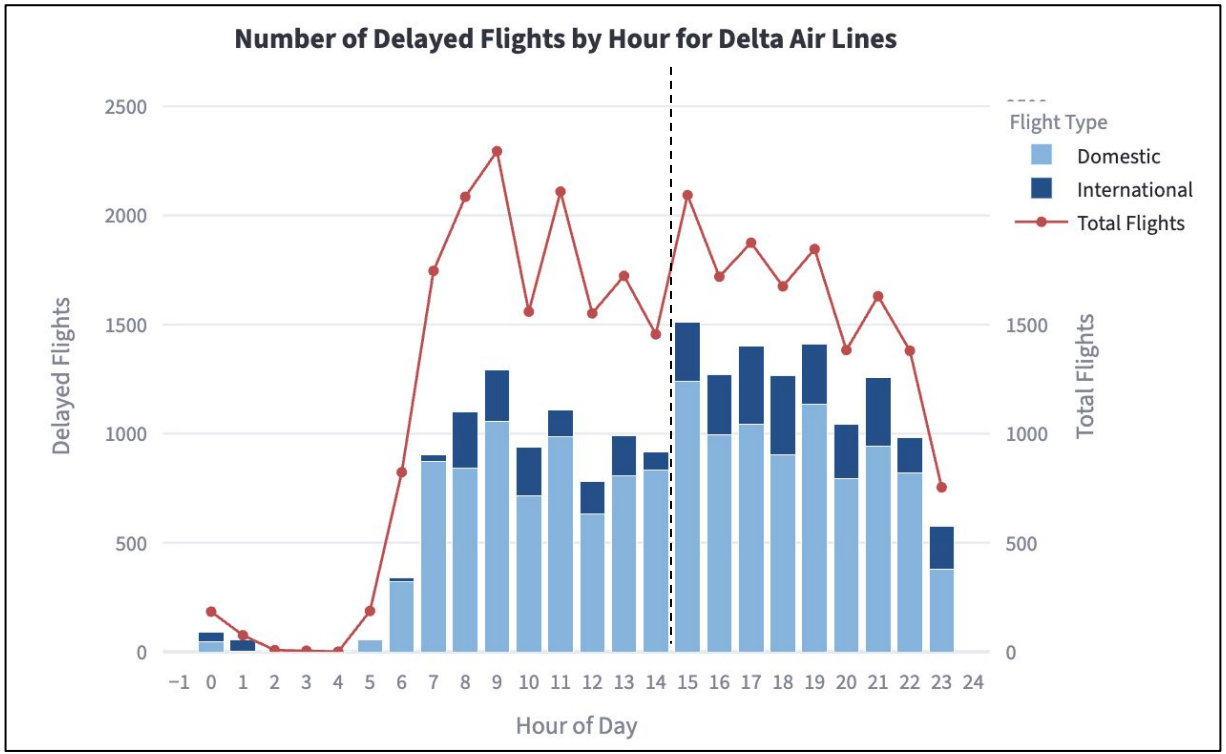


Figure 3. Number of Delayed Flights by Hour: Delayed flights were totaled for each hour and overlaid by the total flight count for comparison.

Majority of Delta's Domestic Flights from Busier Airports Fly to Eastern Regions

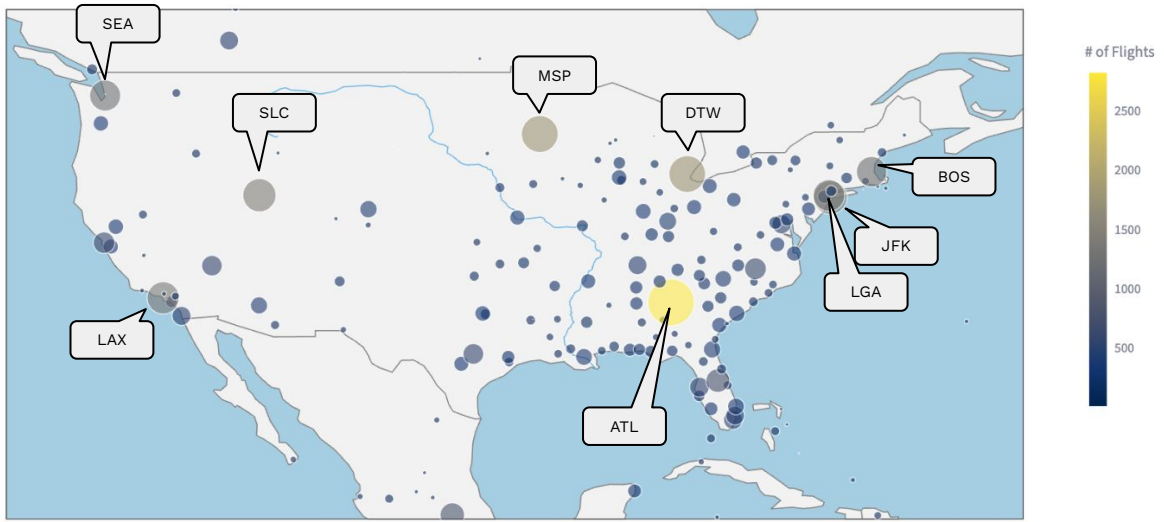


Figure 4. Flight Arrival Distribution: Size and color of the circles represent the total flight volumes for the month of June 2025 at each arrival location from the 12 selected U.S. airports.

Departure Airports

East Coast hubs: JFK, MIA, CLT, ATL, MCO
West Coast hubs: LAX, SFO, PHX, LAS
Central hubs: ORD, DEN, DFW

Delta May Need to Focus Efforts on Reducing NY Delays

NOTE:

To avoid over-inflating delay proportions, I filtered out destinations with very few flights. As a result, many international routes were excluded unless they had both high flight volume and prevalent delays. This filtering may explain why domestic flights are predominantly listed in the table.

Rank	Departure Airport	Arrival Airport
1	JFK	CVG
2	JFK	MIA
3	MIA	JFK
4	JFK	DTW
5	MIA	LGA
6	ORD	LGA
7	JFK	CLT
8	ATL	JFK
9	JFK	BOS
10	JFK	BUF

Table 2. Top 10 Delayed Routes (Delays >60 min): Top delayed routes are defined as those with delay times exceeding 60 minutes. The proportion of flights delayed on each route, delay rate, is used to rank routes with the highest delays. To reduce bias from routes with very few flights, only routes with a flight count above the overall median (82.0) are included.

Every route in the top 10 for delays involved flights to or from New York.

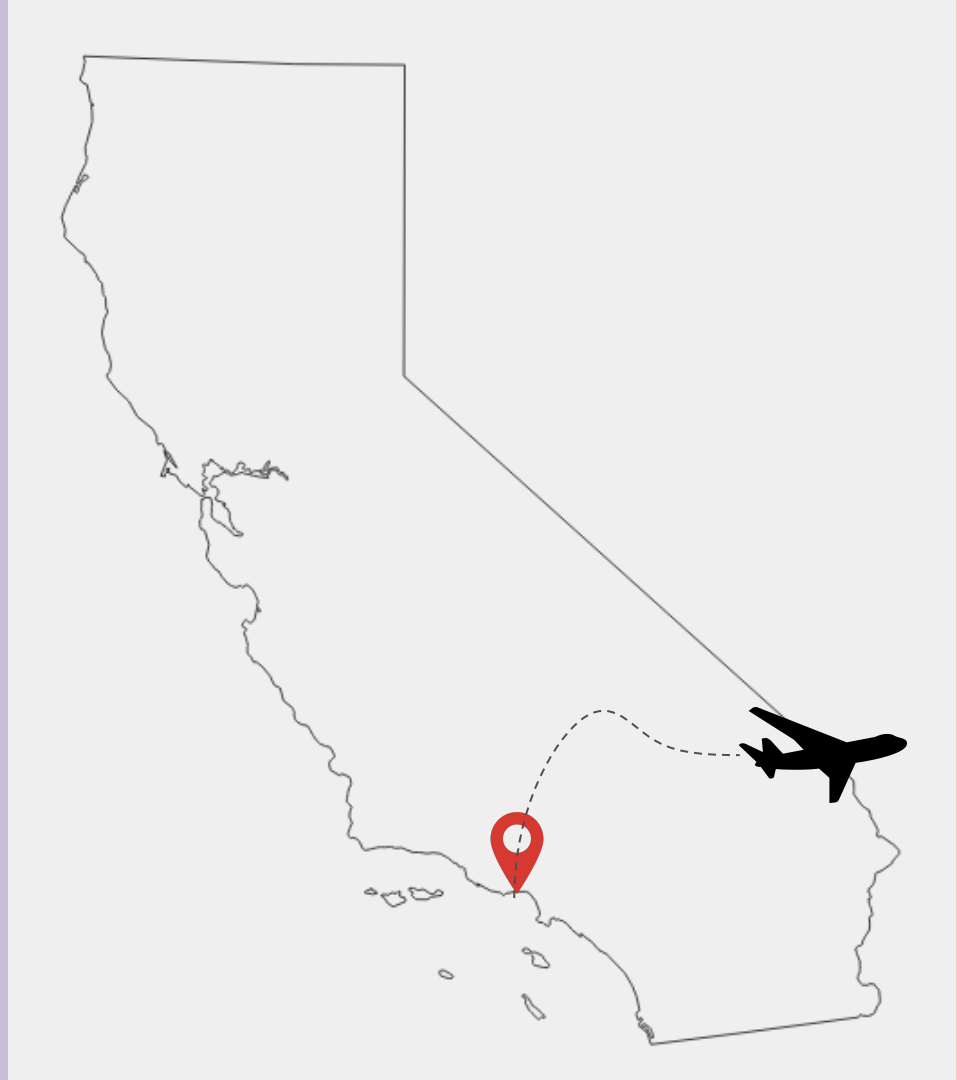
Conclusions

- Based on the weekly flight performance data for Delta, Week 1 showed the strongest on-time rates and average cancellations. Subsequent weeks experienced higher disruptions, likely influenced by external factors such as severe weather on the East Coast and holiday travel peaks.

- Among the most delayed routes, flights to and from New York were more frequently delayed.

- Targeting New York airports, especially JFK, for operational improvements could allow Delta to outperform competitors in managing delays at this critical hub.

- To do that Delta can:
  - Add schedule buffers to prevent delays from cascading.
  - Standby crews & planes during stormy periods in key hubs like NYC and Atlanta.
    - Utilize weather analytics to plan ahead
  - Improve coordination between JFK, LGA, and EWR to reduce air New York traffic congestions.



USE CASE 2: LAX Flights Analysis

1. Problem Statement/ Objective
2. Approach
3. Key Learning/Results
4. Conclusion

Problem Statement and Objective

Problem Statement:

- Flight delays disrupt travel plans, cause missed connections, and create uncertainty for travelers.
- Travelers often lack accessible data to anticipate and avoid high-delay risk routes, airlines, or travel times.

Objective:

- Help travelers plan smarter trips by highlighting delay-prone routes, days, and airlines, while also recommending frequently traveled destinations based on flight volume, along with strategies to avoid or minimize delays.

Approach

- **Data Sources:** Historical flight performance data including delay metrics, airports, routes, airlines, and time-of-day patterns. For this project, I'm focusing specifically on flight patterns from LAX during Memorial Day week, 2025. (size of data: **21,534**)

- **Pipeline Parameters:**
  - `dep_iata_code = ["LAX"]`
  - `start_date = "2025-05-19"`
  - `end_date = "2025-05-25"`

- **Analysis:**
  1. Spot the most delay-prone routes
  2. Rank airlines by delay rates
  3. Reveal delay trends by day and time
  4. Highlight frequently traveled routes

Important Notes

- This dataset is used for illustrative purposes only. To draw stronger conclusions, we'd need flight data from multiple years around the same time.
- These trends are based on flights departing from LAX. Patterns may differ when considering other departure airports.
- This analysis focuses on a holiday week to assess airline and airport performance. Trends observed here may differ from those on typical travel days.

Hub-Connected Flights *May* Suggest Popular Destinations

LAX connects travelers to numerous international destinations.

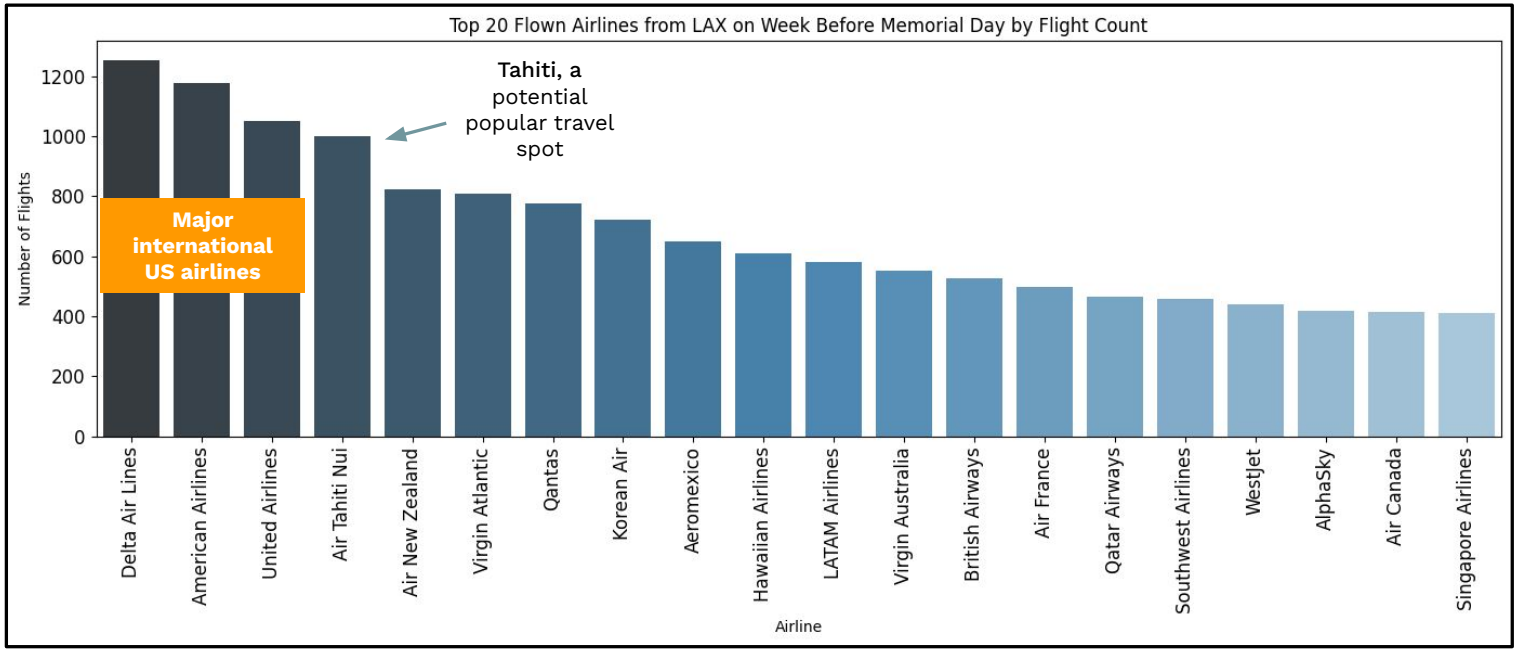


Fig 2. Top 20 Flown Airlines from LAX by Flight Count: Totaled flights for each airline for every day of the week of 2025-05-19 to 2025-05-25.

Airline Recommendations Based on Delays

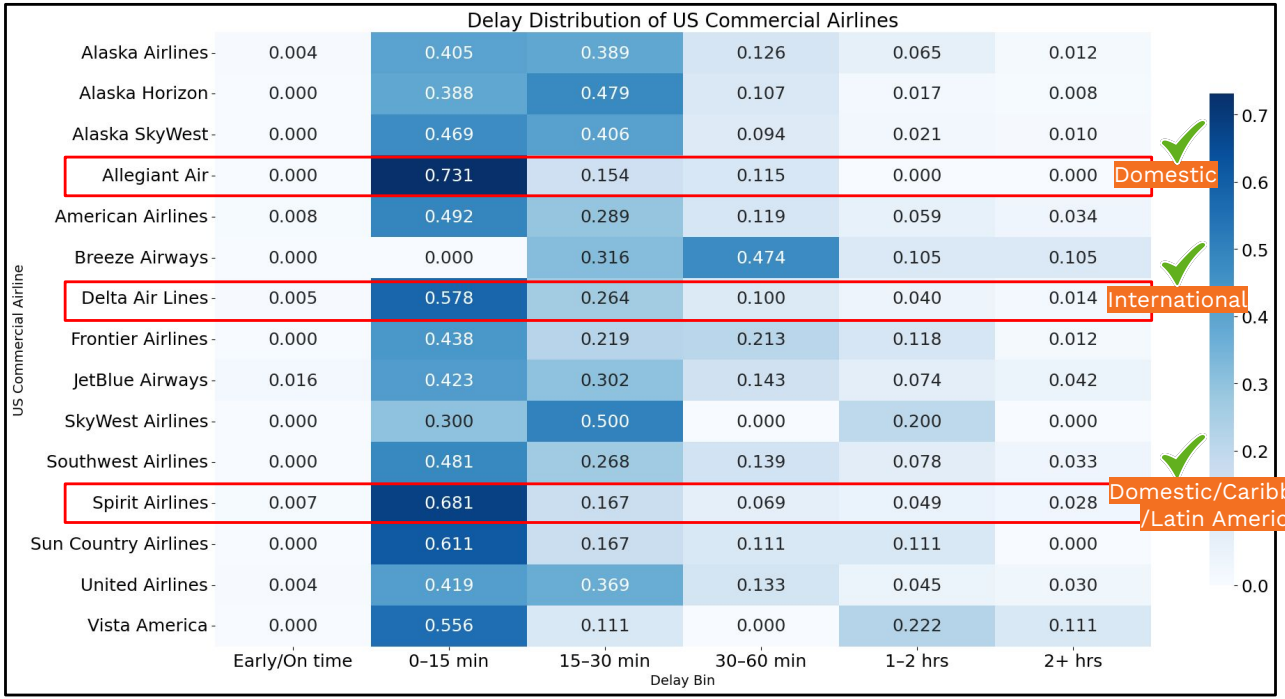


Fig 1. Delay Distribution of US Commercial Airlines: Proportion in the heatmap reflects the normalized counts across each delay bins per airline. Delays from *both* domestic and international routes were used to calculate the proportions.

Delta Airlines
Fewest delays among U.S. international carriers; most delays fall within 0–15 minutes. Tends to be more expensive.

Allegiant
Budget-friendly option for domestic flights with minimal delays.

Spirit Airlines
Affordable choice for domestic travels and international destinations like the Caribbean or Latin America.

Evening Flights Around 8:00PM Have Fewer Delays and Congestions

Peak travel times are associated with higher delays.

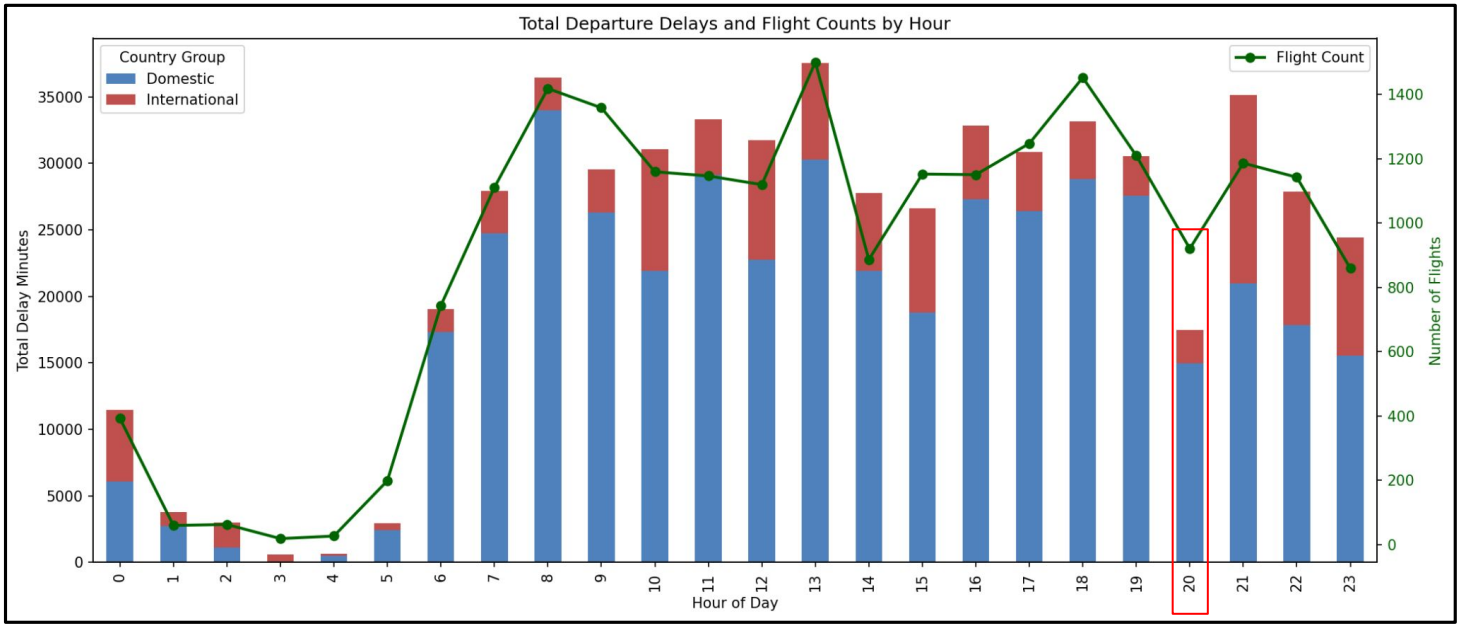


Fig 3. Total Departure Delays and Flight Counts by Hour: Departure flight delays are summed across both domestic and international flights to give the total delay in minutes. Flight counts are included to provide context, as lower delay times may be influenced by reduced flight activity during those hours, and vice versa.

Monday, Tuesday, and Sunday Tend to Have Lower Delays

Over 40% of daily flights experience delays under 15 minutes.

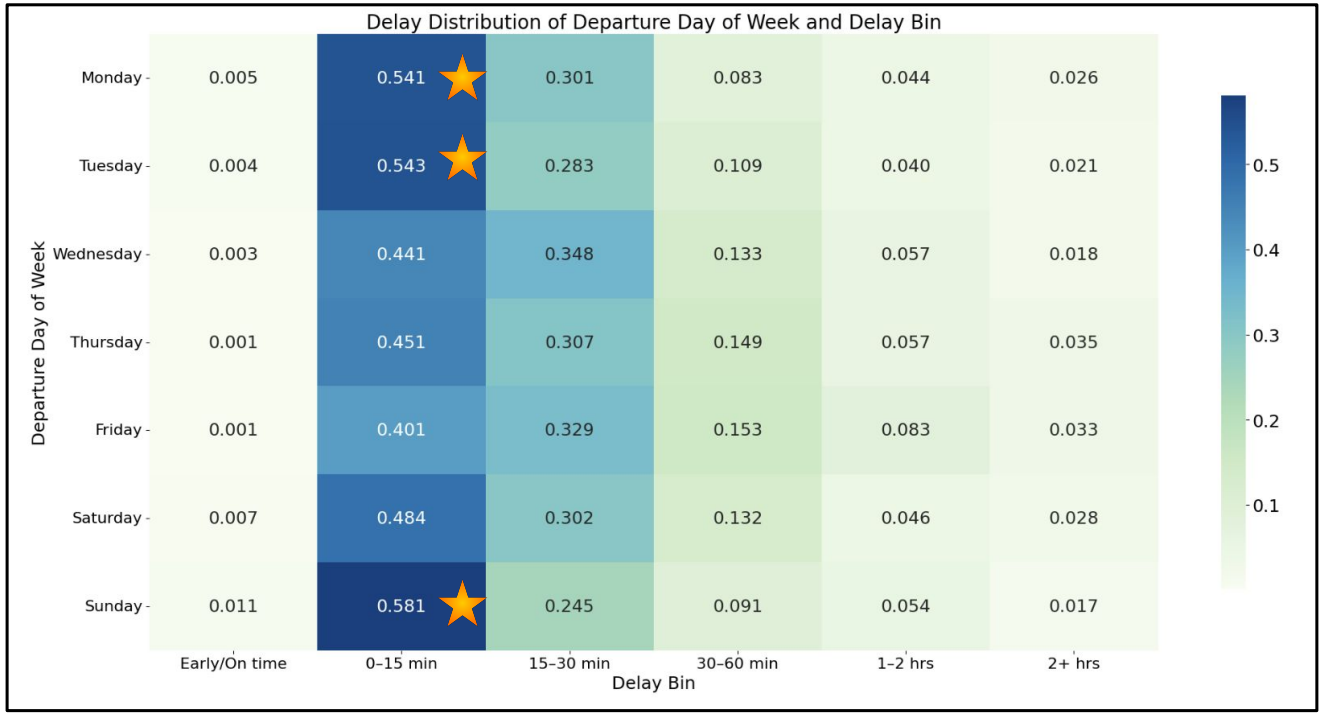


Fig 4. Delay Distribution of Departure Day of Week: Proportion in the heatmap reflects the normalized counts across each delay bins per day of week. Delays from ***both*** domestic and international routes were used to calculate the proportions.

Majority of LAX Flights Serve U.S. Cities

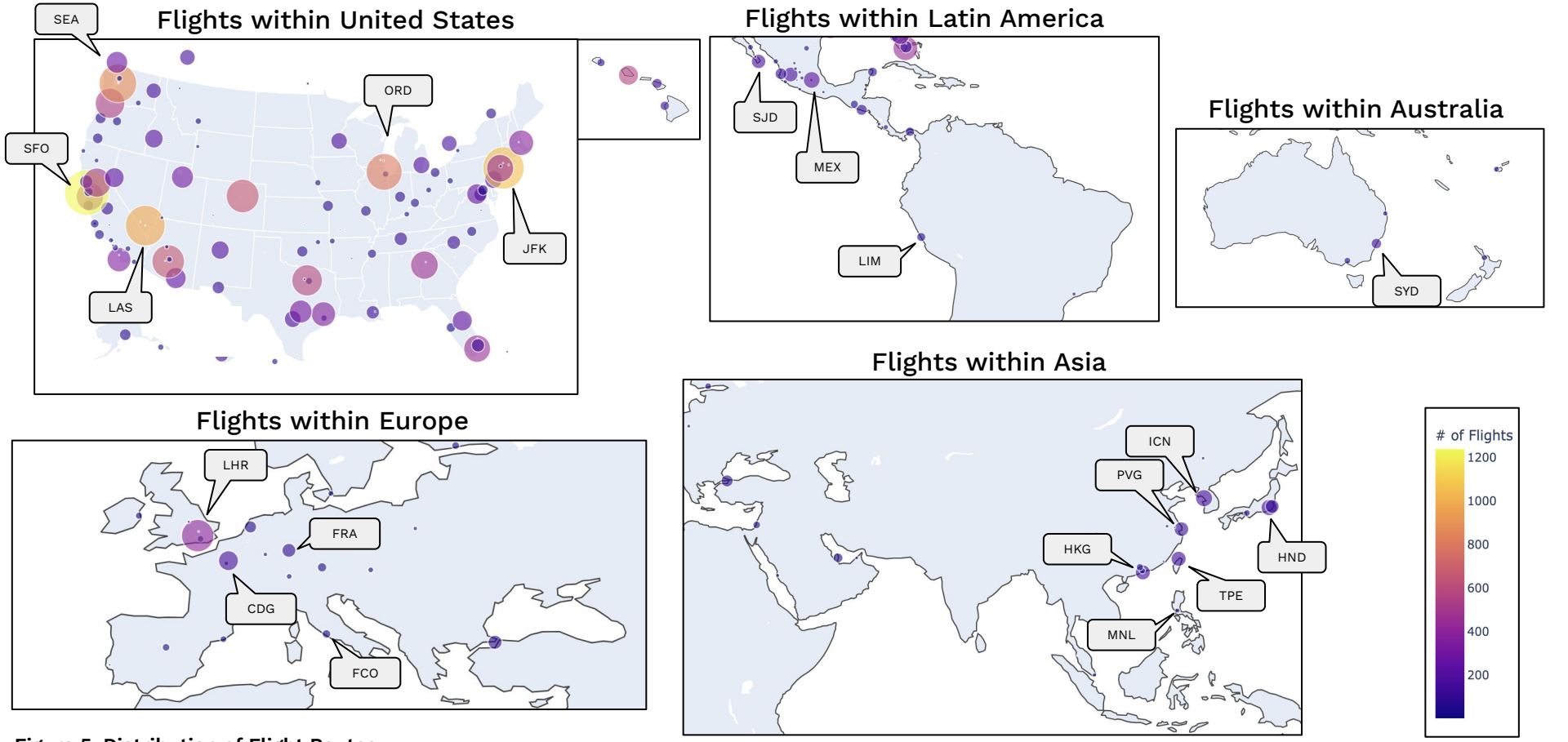
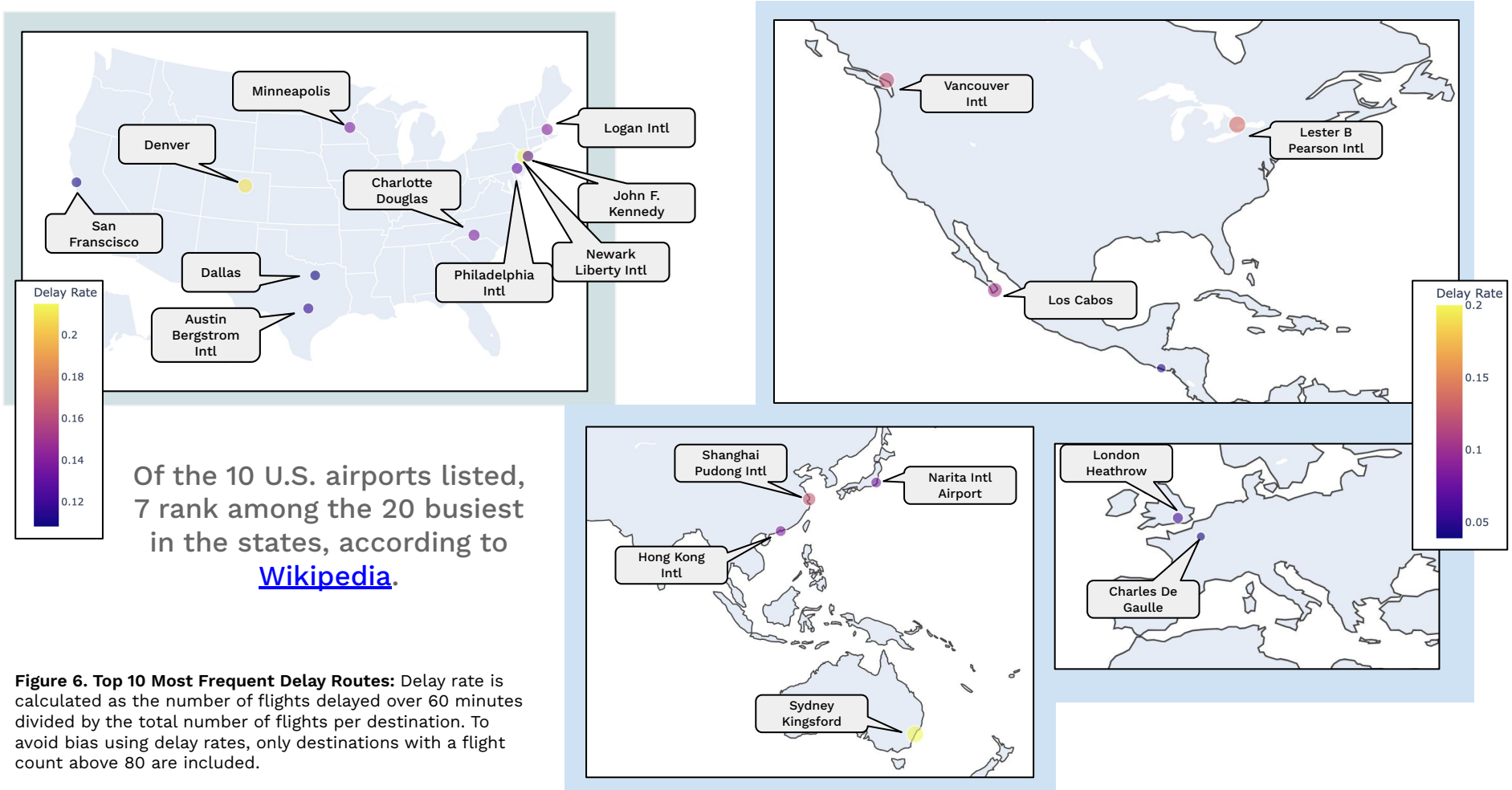


Figure 5. Distribution of Flight Routes

Ten Domestic and International Routes with the Most Frequent Delays (>60min)



Of the 10 U.S. airports listed, 7 rank among the 20 busiest in the states, according to Wikipedia.

Figure 6. Top 10 Most Frequent Delay Routes: Delay rate is calculated as the number of flights delayed over 60 minutes divided by the total number of flights per destination. To avoid bias using delay rates, only destinations with a flight count above 80 are included.

Flight Distribution of U.S. Airlines on Most Delayed Routes

Which U.S. airline contributes the most to delays on the most delayed routes?

Figure 7. Proportion of Airline-Attributed Delays on High-Delay Routes: Proportions are calculated as the number of delayed flights (>60 min) per airline per city divided by the total number of delayed flights for that destination.

What proportion of any U.S airline to the top delayed destinations were delayed >60 min?

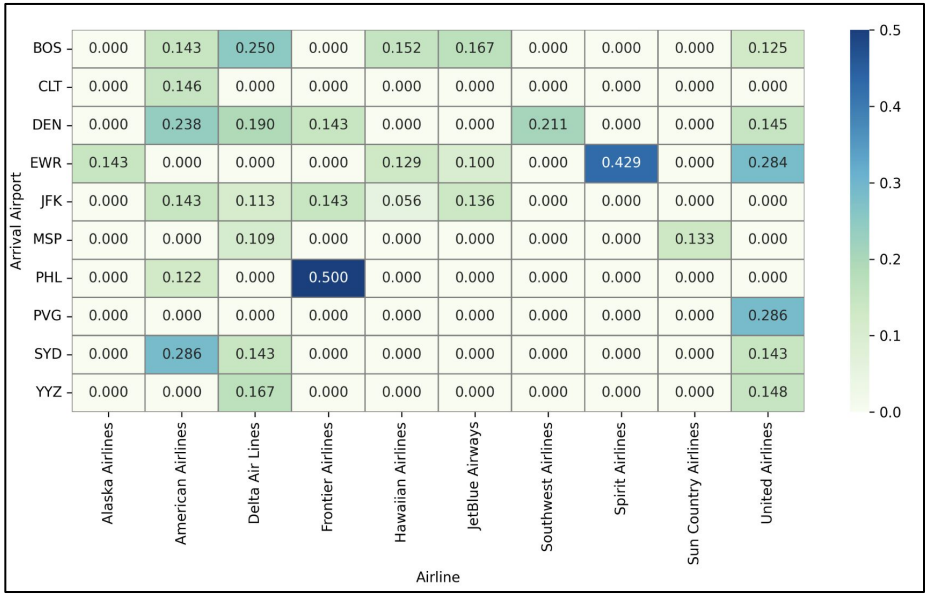


Figure 8. Proportion of Delayed Flights (>60 min) by Airline per Destination: Proportions are calculated as the number of flights delayed over 60 minutes for each airline and city, divided by the total number of flights (delayed and non-delayed) for that airline and city.

Airline X contributes XX% of all delayed flights to City Y.

XX% of Airline X flights to City Y were delayed

Airport	Top Delay Contributor	Highest Delay Rate	Airline With Most Impact on Delays	Tie-Breaker Rationale
BOS	jetBlue	Delta Air Lines	Delta Air Lines	
CLT	American Airlines	American Airlines	American Airlines	
DEN	United	American Airlines	American Airlines	
EWR	United	Spirit	United	The majority of delayed flights to EWR (Newark) were from United Airlines. While Spirit shows a higher proportion of its flights being delayed to EWR, this is likely inflated due to the small number of flights it operates to that airport.
JFK	American Airlines	American Airlines, Frontier Airlines	American Airlines	
MSP	Delta Air Lines	Sun Country Airlines	Delta Air Lines	Sun Country Airlines has hub advantage, meaning there are high flight frequencies to/from MSP operated through Sun Country airlines.
PHL	American Airlines	Frontier Airlines	American Airlines	Since Frontier only operated a small number of flights to PHL, its delay rate may be skewed upward.
PVG	United	United	United	DELTA IS WINNER!! *(AMONG THE BIG 3)*
SYD	American Airlines	American Airlines	American Airlines	
YYZ	United	Delta Air Lines	United	Proportion of delays for both plots were high for United.

Delta Air Lines
American Airlines
Frontier Airlines
Sun Country Airlines

Table 1. Delay by Airline Summary: Summary of airlines with the most impact on delays for top delayed routes using Figure 7 and Figure 8. “Top Contributor” and “Highest Delay Rate” refers to the highest proportion observed for airports in Figure 7 and Figure 8, respectively.

Conclusion

- **Memorial Day week** was chosen for analysis due to higher travel volume, making it ideal for assessing airline reliability and optimal flight times.

- During this period, **Delta** stood out as most reliable among the major international airlines, making it a solid option for travelers prioritizing **on-time performance for their international flights**

- Delta can also be a great option for domestic flights, but it tends to be more expensive. **Allegiant** and **Spirit** both showed relatively low delay rates, making it a more **budget-friendly option for domestic flights** during this period.

- If you're considering a trip during Memorial Day week, this dataset points to a few U.S. destinations that were especially popular if flying from LAX:

  - **Las Vegas** for fun nightlife, food, and 24/7 energy
  - **San Francisco** and **Seattle** for cool weather, laid-back vibes, nature, and a touch of tech culture

- As for international trips, the **United Kingdom** emerged as a popular destination from LAX during this time.

- When booking your flights for this period, the data suggests to avoid flights connecting to BOS, CLT, DEN, EWR, JFK, MSP, PHL, PVG, SYD, and YYZ and to book departure days for Monday, Tuesday, or Sunday for a lower chance of delay.

Project Challenges & Limitations

Project Challenge & Limitations

- Certain data from AviationStack API can be sparse. Pipeline fills in missing delay values via imputations. If data is sparse, too much data imputations may impact the analysis.

- AviationStack API limits historical data collection to only 3 months. For more extensive data, this will require running the pipeline periodically

- This pipeline requires a paid subscription to run

Next Steps

Next Steps

- Current analysis does not account for factors like **weather**, **social**, or **political** events, which may impact flight delay patterns
- As next steps, I can **integrate weather API** to provide additional insights on factors contributing to aviation delays
- Then to enable greater insights, I can work to **automate the pipeline to collect year-round data** for identifying seasonal trends.
- I can then build a **predictive model** to identify which routes are most likely to experience delays.

THANK YOU FOR TAKING THE TIME TO REVIEW MY PROJECT!