AirFly-Insights

Infosys SpringBoard Internship

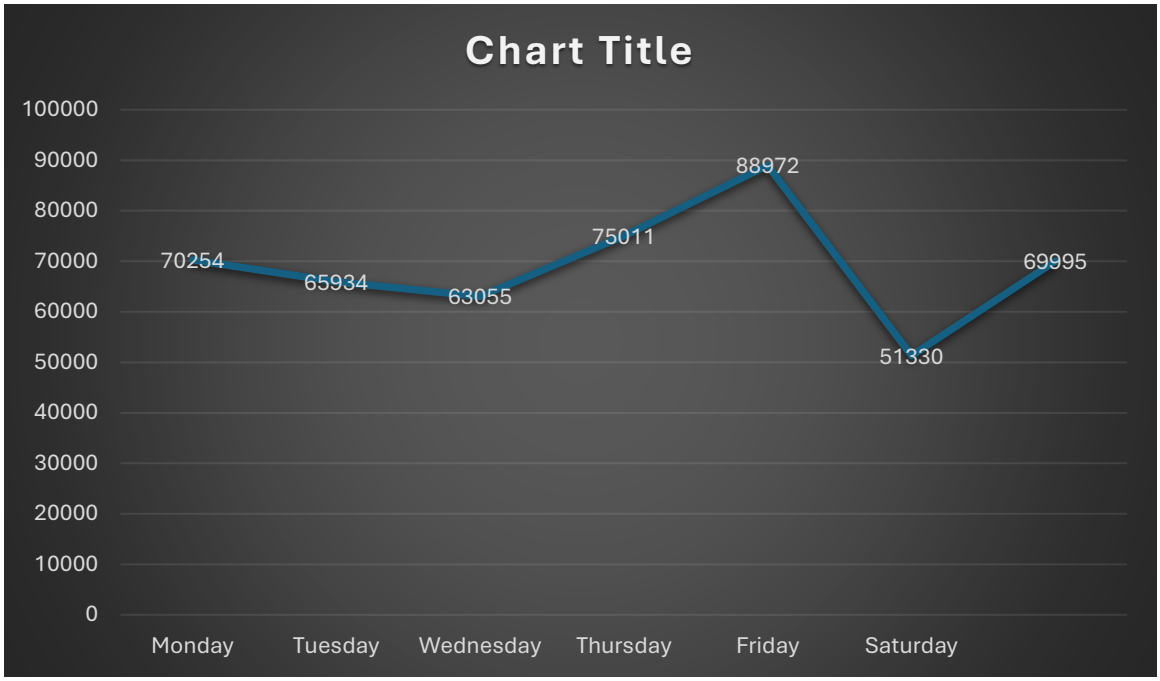
Dataset:

- **Kaggle:** Flight Delay and Causes
- https://www.kaggle.com/datasets/undersc0re/flight-delay-and-causes
- This dataset contains airline flight records including flight details,delays,airport and operations.
- It contains 484,559 rows and 29 columns initially.
- After preprocessing 484,310 rows and 33 columns.

Milestone 1: Data Foundation and Cleaning

Key Performance Metrics:

Flights by Day:

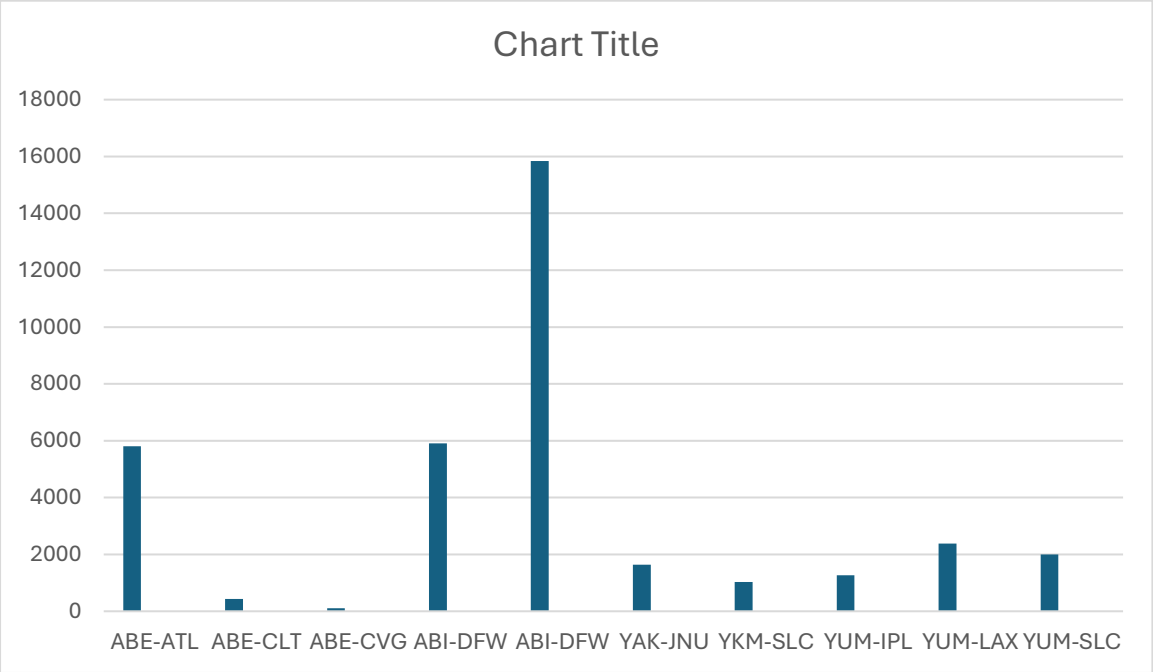


Distance:

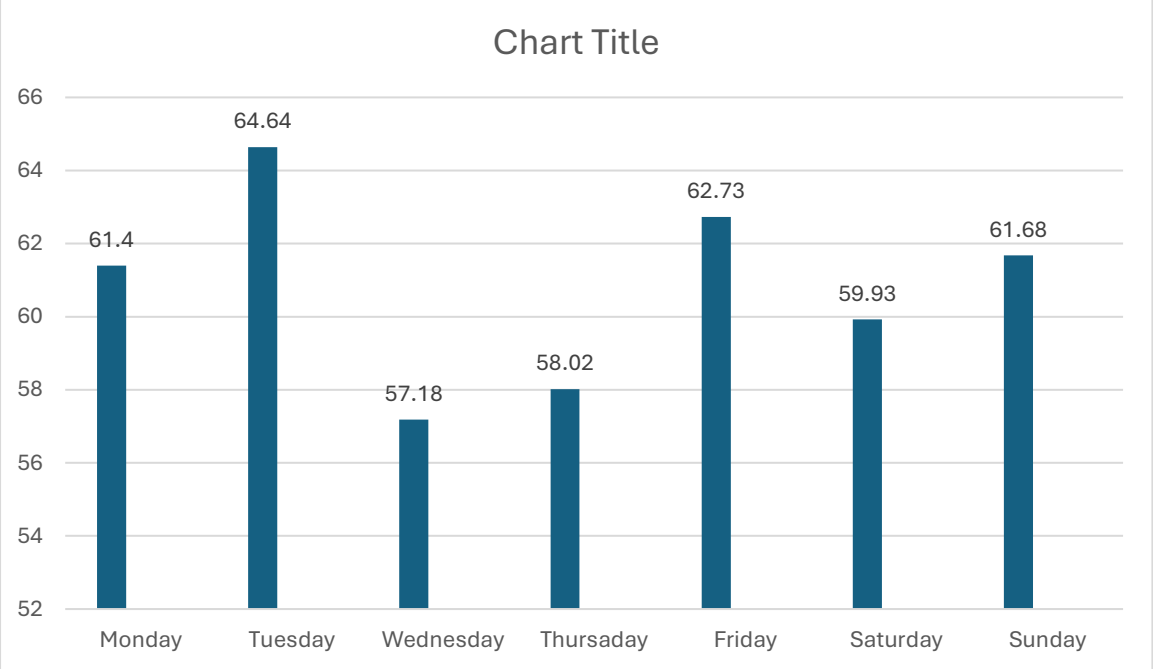
a. Minimum Flight Distance: 31 miles
b. Maximum Flight Distance: 4,502 miles
c. Average Flight Distance: 752.3844 miles

Delay:

TOTAL ROUTE DELAY



BASED ON DAY OF WEEK



Insights

Travel Patterns:

- Friday is the busiest travel day (88,972 flights), while Saturday has the lowest traffic (51,330 flights).

Route-Level :

- The **ORD → LGA** route is the most delay-prone, reflecting congestion challenges at both a major Midwest hub (Chicago O'Hare) and a ighly slot-constrained East Coast airport (LaGuardia).
- The **ALB → CVG** route records the lowest average delays, highlighting the benefits of regional connectivity between less congested airports with smoother operations.

Week based delay:

- Wednesday has less delay.
- Tuesday has more delay.

Preprocessing Technique:

Data Cleaning Steps

a. Duplicate Removal

- All duplicate rows were detected and removed using drop_duplicates().

b. Handling Missing Values

- **Numeric Columns (Delays, Distance, Time):**
  - Columns: DayOfWeek, ActualElapsedTime, CRSElapsedTime, ArrDelay, DepDelay, Distance, AirTime, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay, Diverted, TaxiIn, TaxiOut.
  - Missing values imputed with **column-wise mean**.
- **Categorical/Time Columns (DepTime, ArrTime, Org_Airport, Dest_Airport):**
  - Missing values filled using the **mode within each flight group** (groupby(FlightNum) and filled with most frequent value).
- **Remaining Missing Values:**
  - Any leftover null entries were dropped to ensure a clean dataset.

Data Type Conversions

- **Date column:** Converted from string to datetime format.
- **Time columns (DepTime, ArrTime, CRSArrTime):**

- Converted from integer (hhmm) to proper **HH:MM string format** using a custom parsing function.

Feature Engineering

- **Month:** Extracted from Date column.
- **Hour:** Extracted from Date (departure hour).
- **Route:** Created by concatenating Origin and Dest (e.g., *ORD → LGA*).
- **DayName:** Mapped DayOfWeek values (1–7) into actual weekday names (Monday–Sunday).
- **TotalDelay:** Computed as the sum of individual delay causes:
  - WeatherDelay + CarrierDelay + NASDelay + SecurityDelay + LateAircraftDelay.

Final cleaned dataset saved as:
Flight_delay_cleaned.csv

Milestone 2: Visual Exploration and Delay Trends:

Week 3: Univariate and Bivariate Visual Analysis

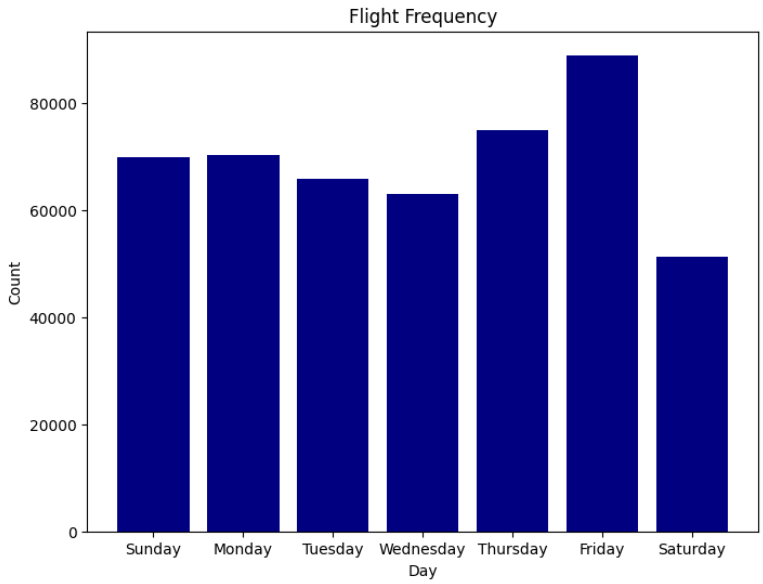
• Top airlines, routes, and busiest months

• Flight distribution by day, time, and airport

• Plot bar charts, histograms, boxplots, and line plots

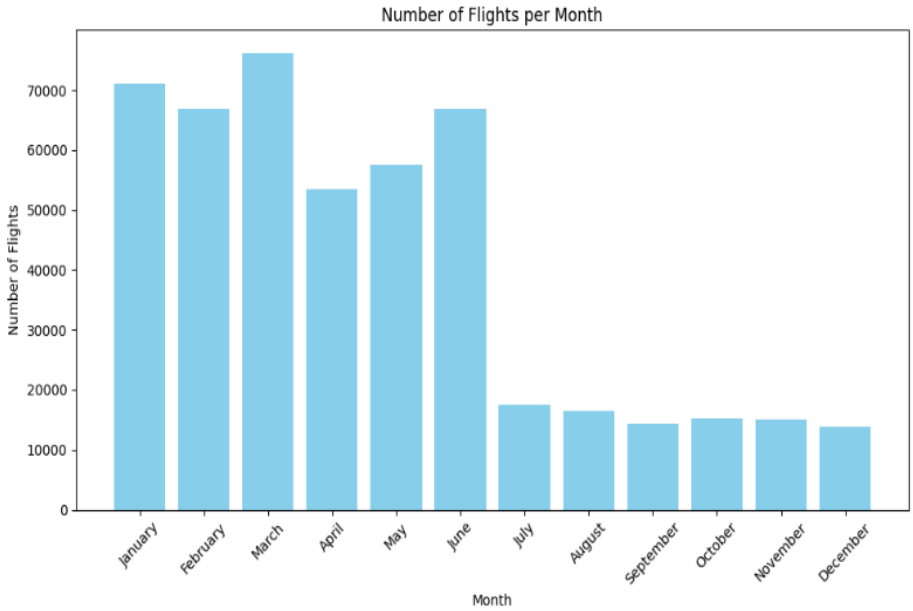
EDA(Exploratory Data Analysis):

- **Flights by Day:**



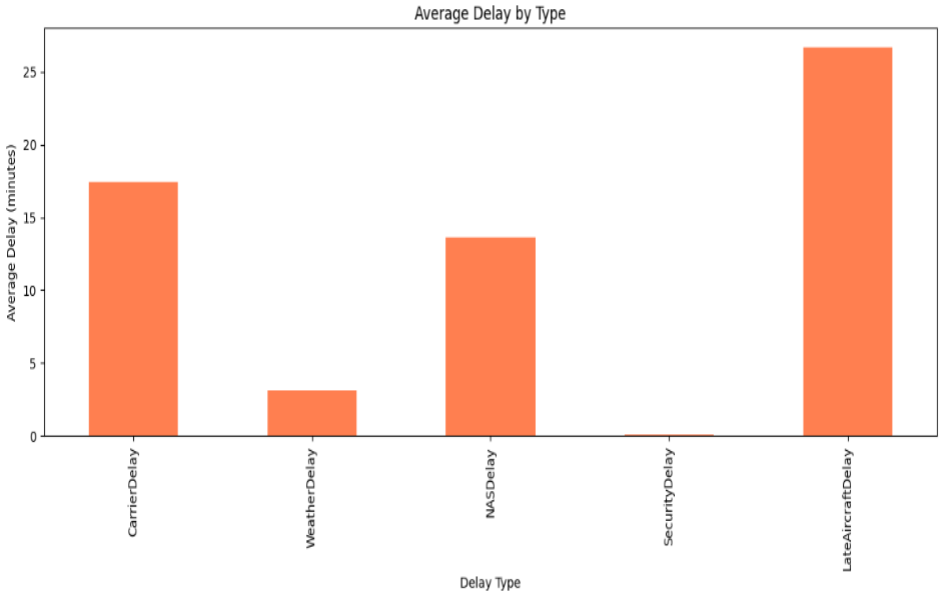
- **Flights by Day:** Friday has the most flights, while Saturday has the least.

➢ **Flights by Month:**



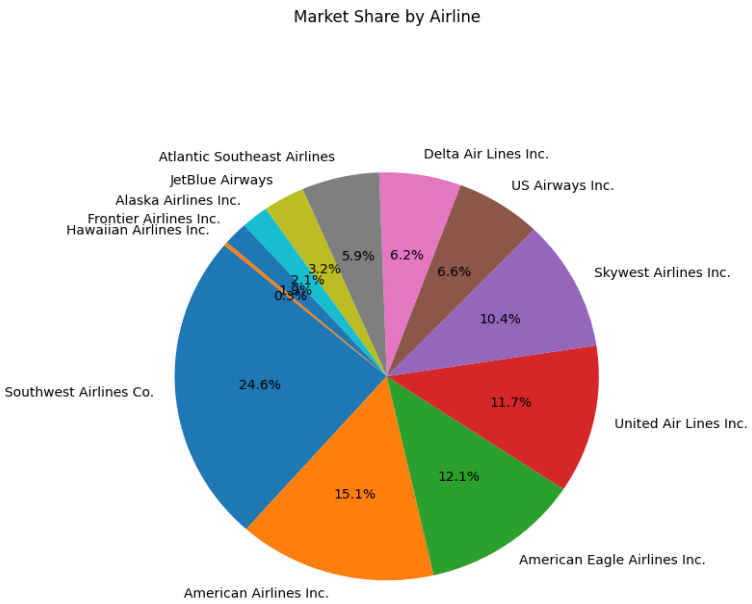
♦ **Flights by Month:** Jan-Jun has highest number of flights compared to Jul-Dec. March has highest number of flights.

➢ **Average Delay by delay type:**



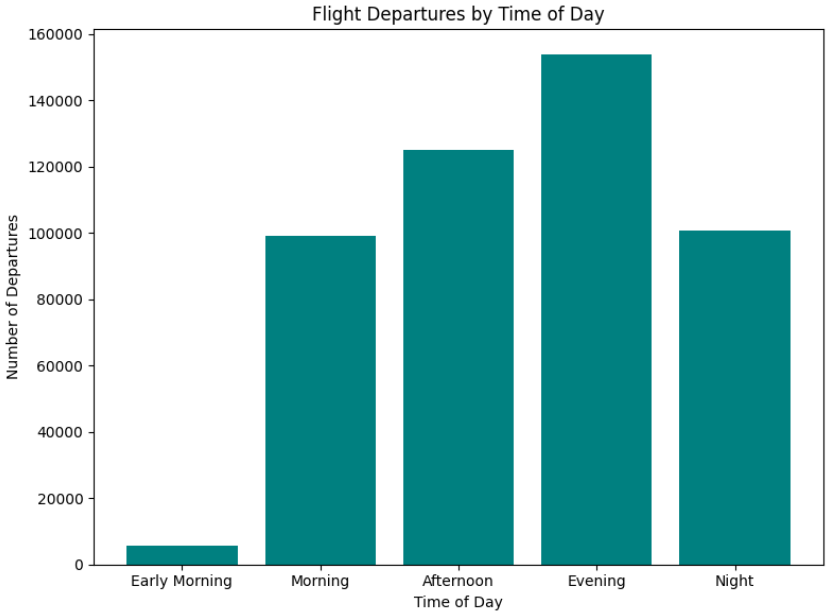
♦ **Delay type:** Late Aircraft delay has highest delay while Security delay has lowest

➢ Market Share by Airline:



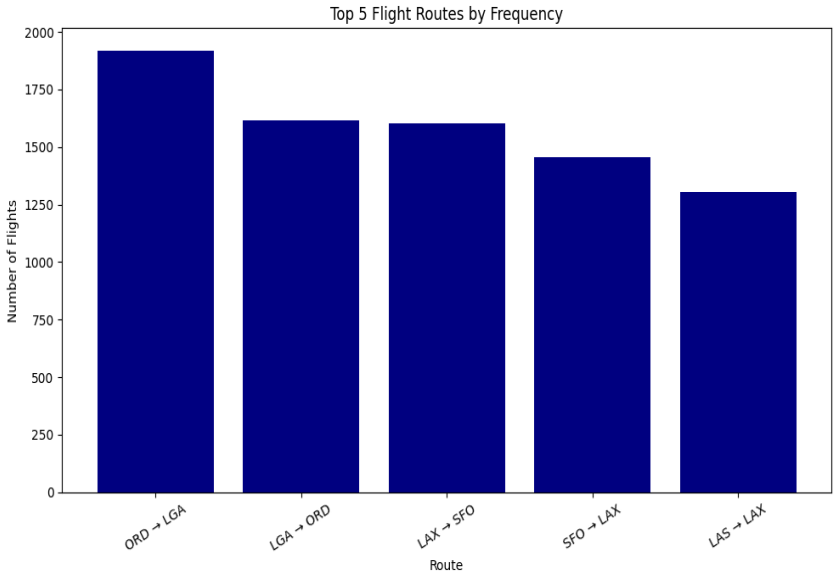
- **Market share:** Southwest Airlines Co has highest share while Hawaiian Airlines Inc. has lowest share.

➢ Flight departures by Time of Day:



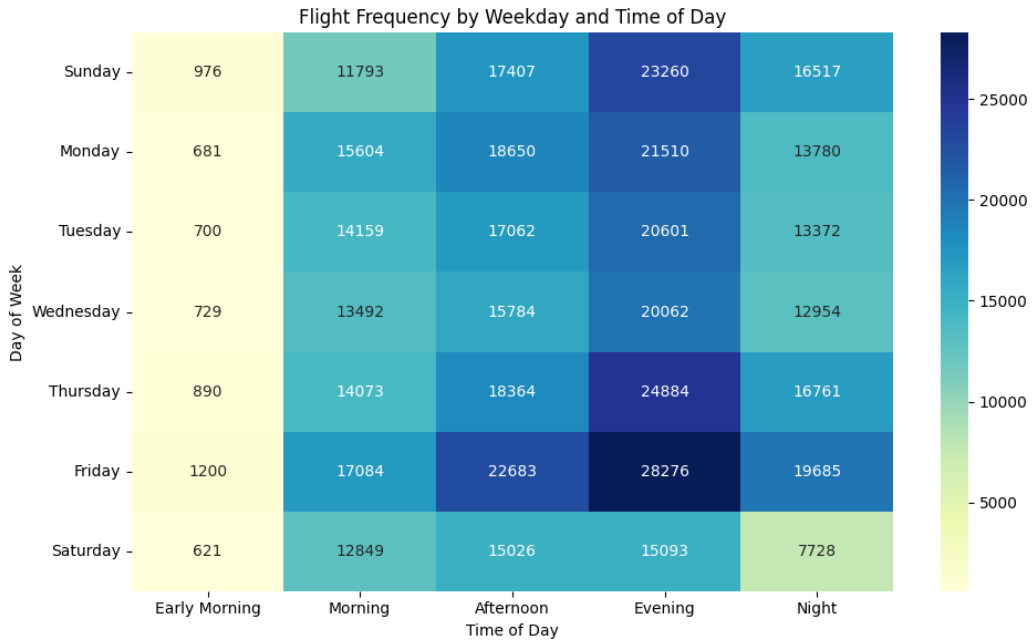
- **Flight departures by Time:** Evening has highest flight while early morning has lowest flight departures.

➢ Top 5 Flight Routes by Frequency:



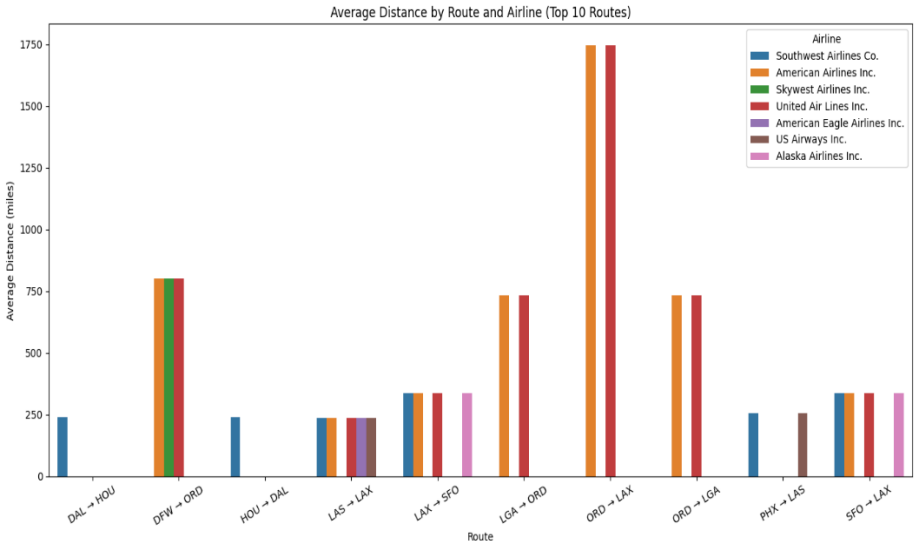
- **Top 5 Flight routes:** ORD -> LGA has highest frequency of flights.

➢ Flight frequency by Time and Day of Week:



- **Flight frequency:** Friday Evening has higher frequency while Saturday morning has lower frequency.

➢ Average distance by route and airline:



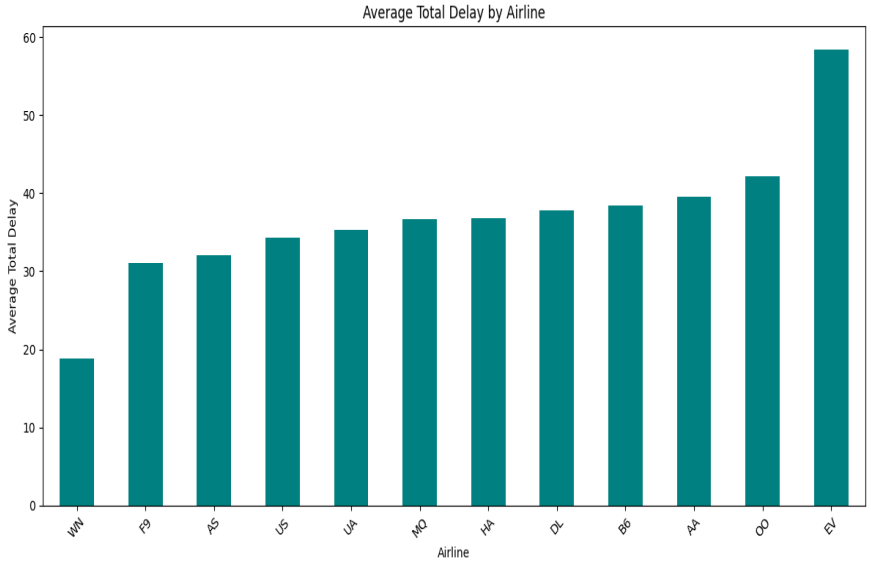
Week 4: Delay Analysis – Airline and Weather

• Compare delay causes by airline

• Explore carrier delays, weather delays, NAS delays

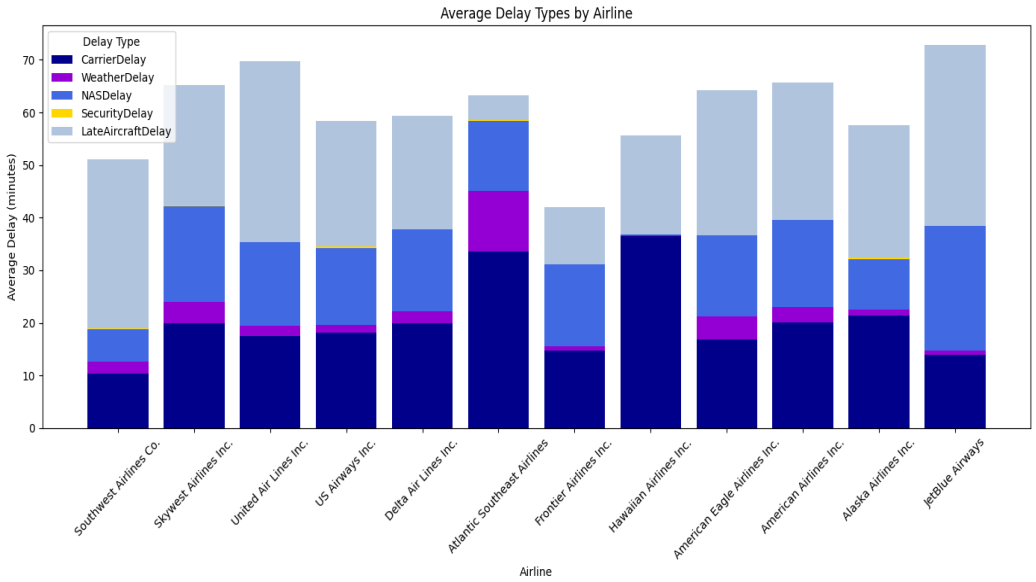
• Visualize delays by time of day and airport

➢ Average Total Delay by Airline:



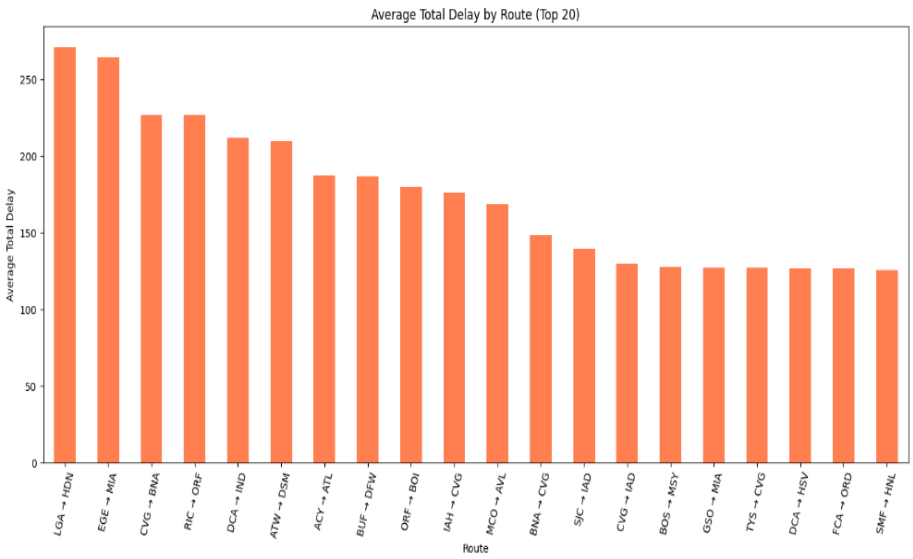
- **Average Total Delay:** EV Airline has highest flight delay while WN has lowest flight delay

➢ Average Different Delay by Airline



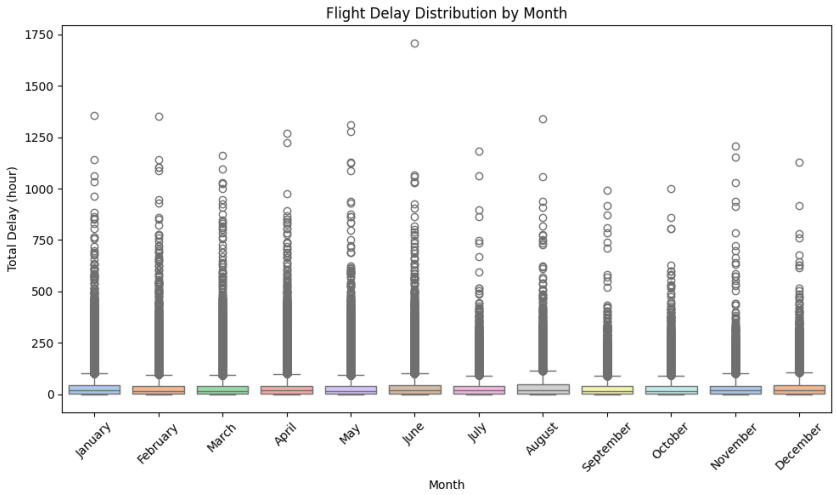
- Southwest Airlines has highest late aircraft delay. Atlantic Southeast Airline has highest weather delay. Hawaiian Airlines has highest carrier delay and JetBlue Airways has highest NAS delay

➢ Average Total delay by route:



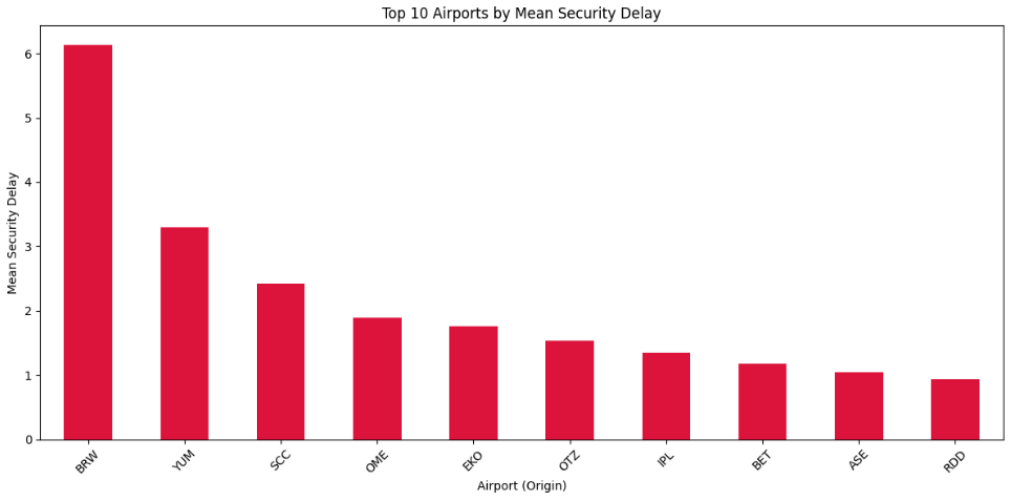
- LGA-HDN has highest delayed route.

➢ **Flight Delay Distribution by month:**



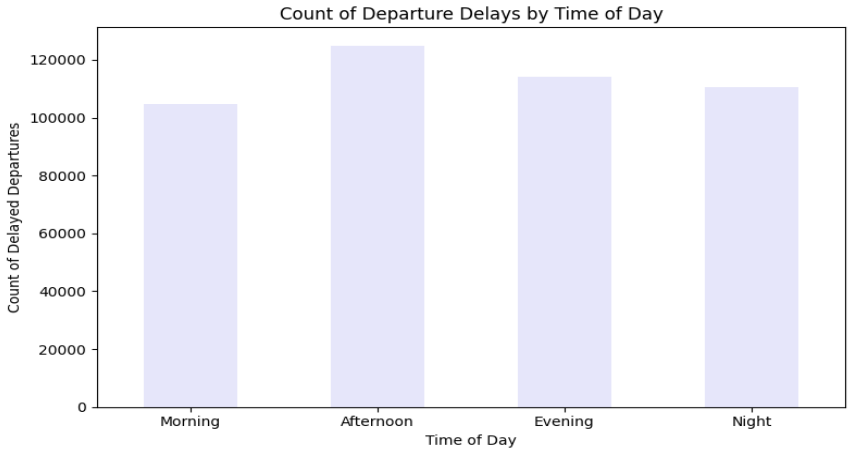
- March has highest delay

➢ **Top 10 Airports by Security Delay:**

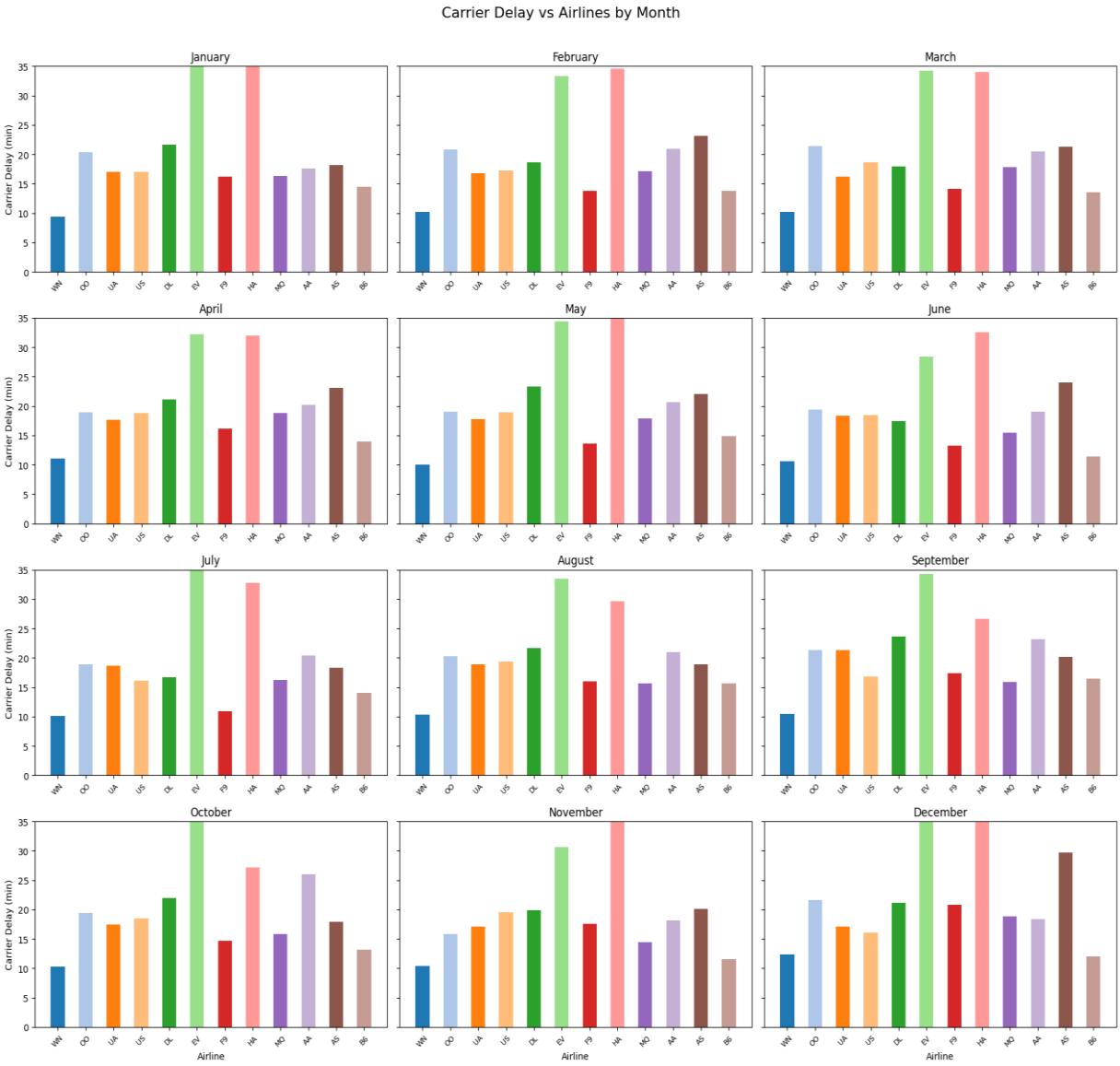


- Wiley Post-Will Rogers Memorial Airport has highest Security Delay.

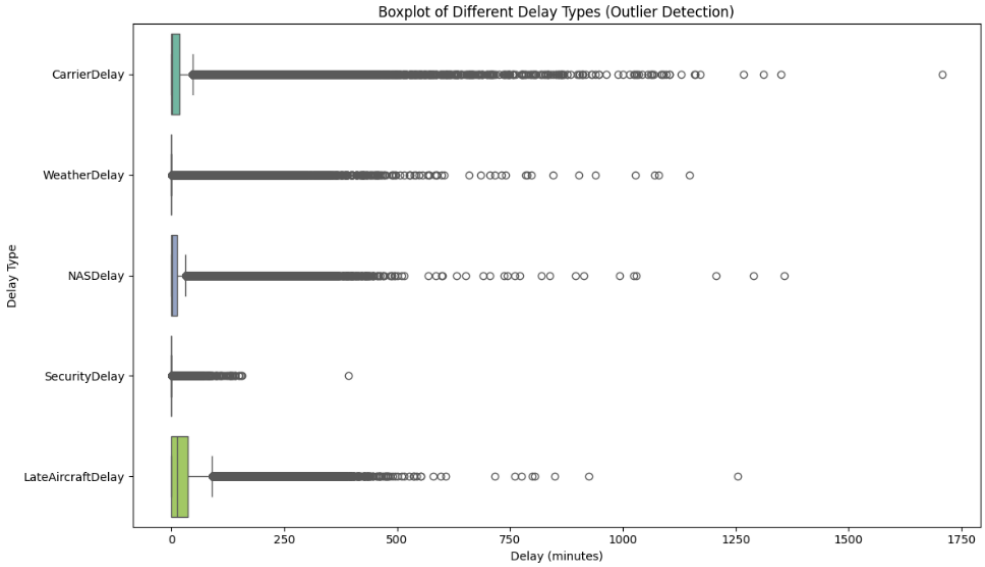
➢ **Count departure delays by time of day:**



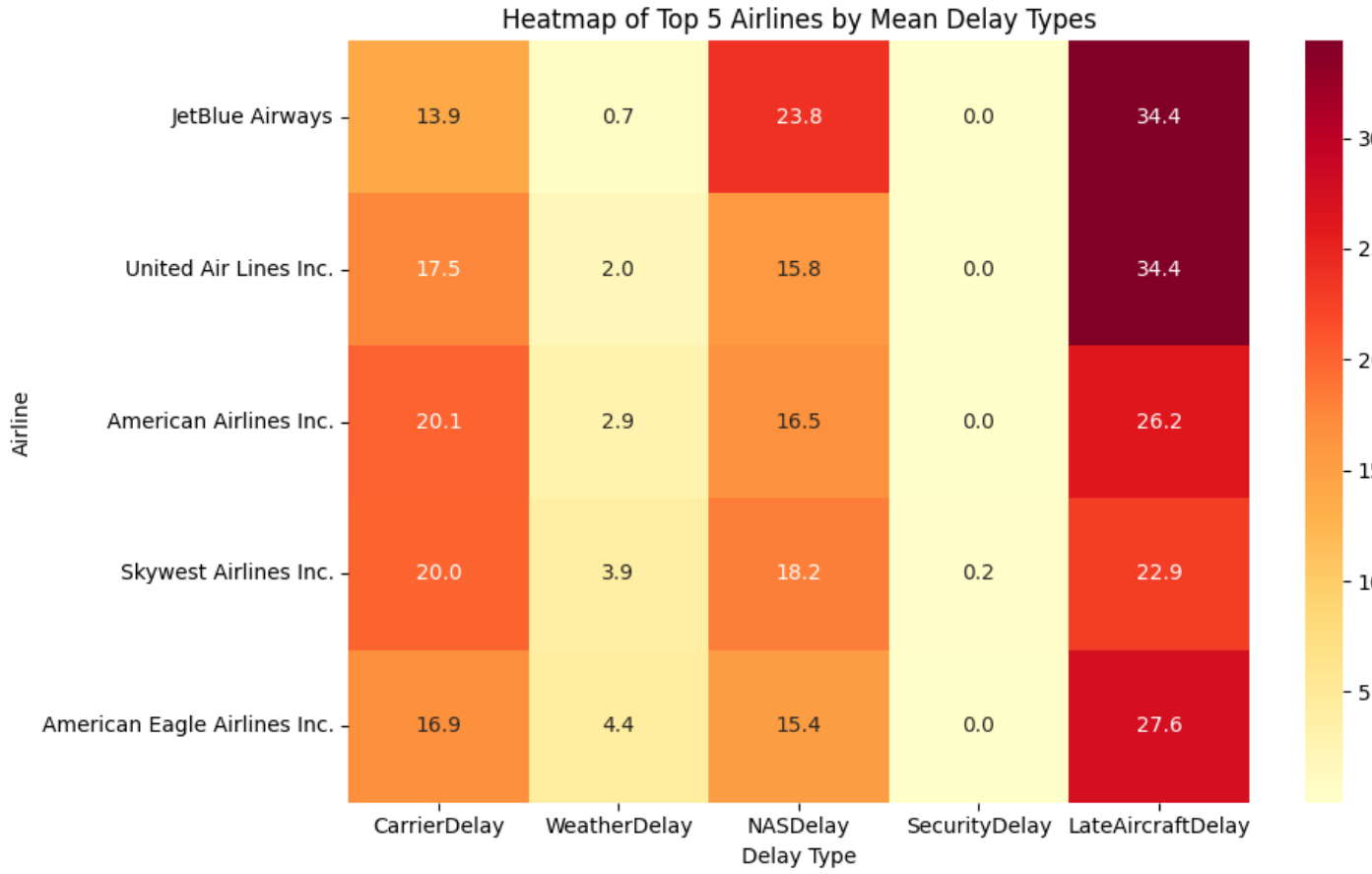
➢ Carrier Delay vs Airlines by Month:



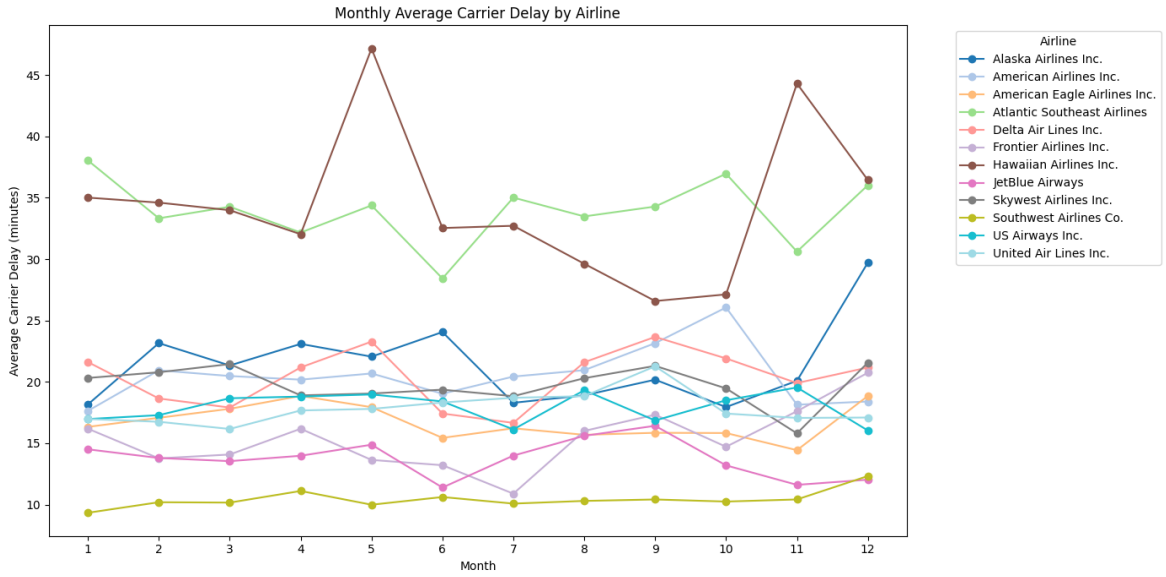
- **Boxplot of different delay types:**



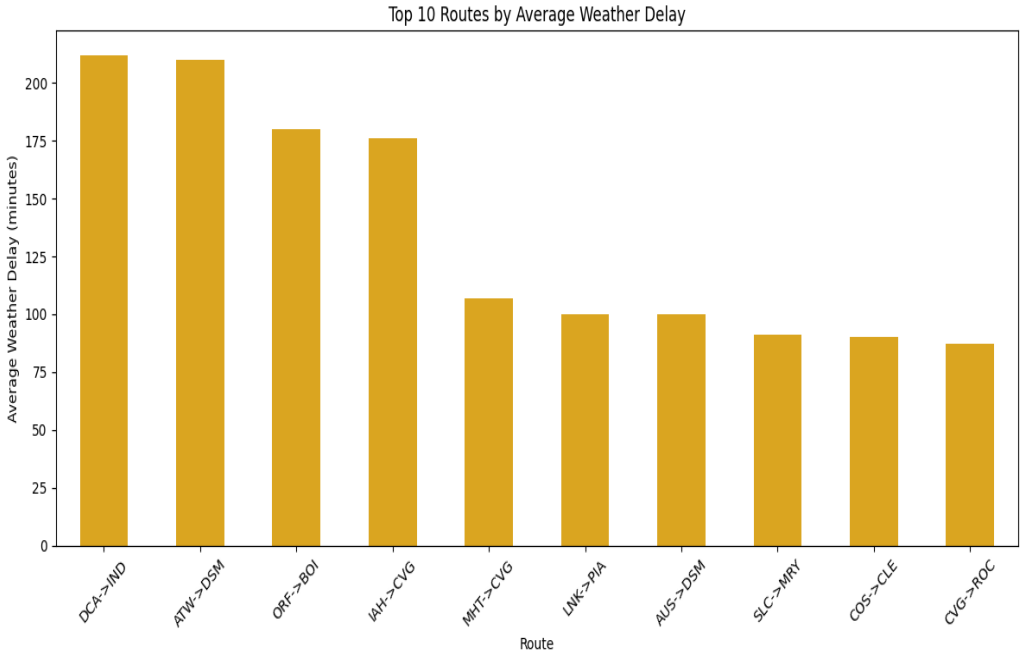
- **Heatmap of top 5 Airline and delay types:**



➢ Monthly Average Carrier Delay by Airline:

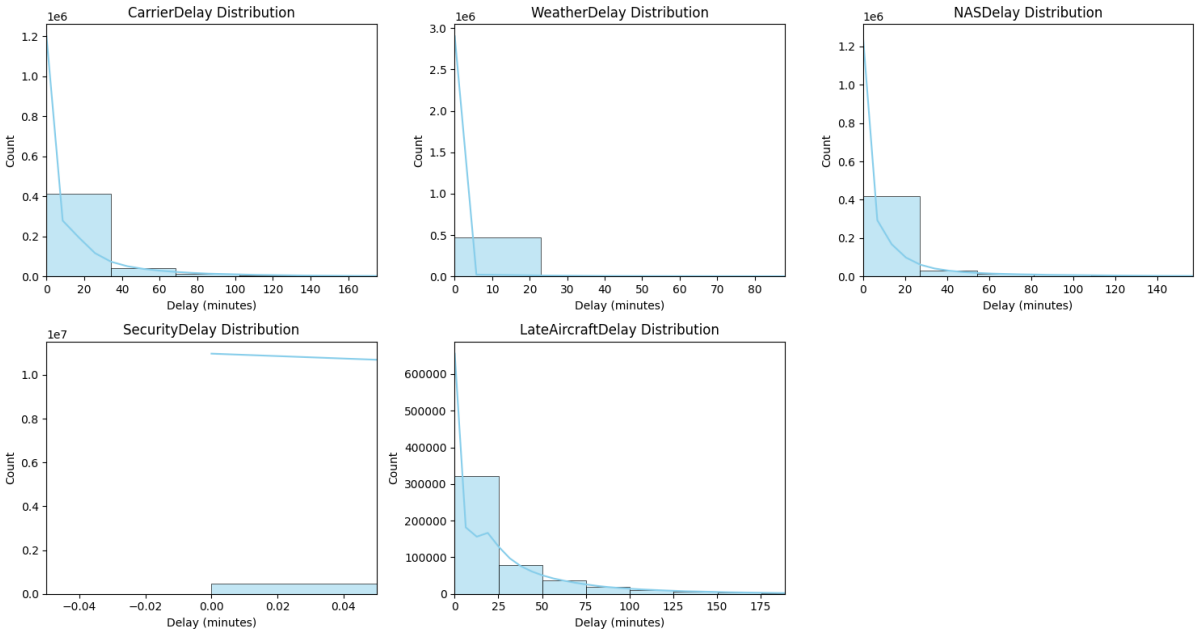


➢ Routes by Average Weather Delay:



- ♦ DCA-IND and ATW-DSM has highest weather delay

➢ Right Skewed delay:



➢ Airports by Security Delay: