

AirFly-Insights

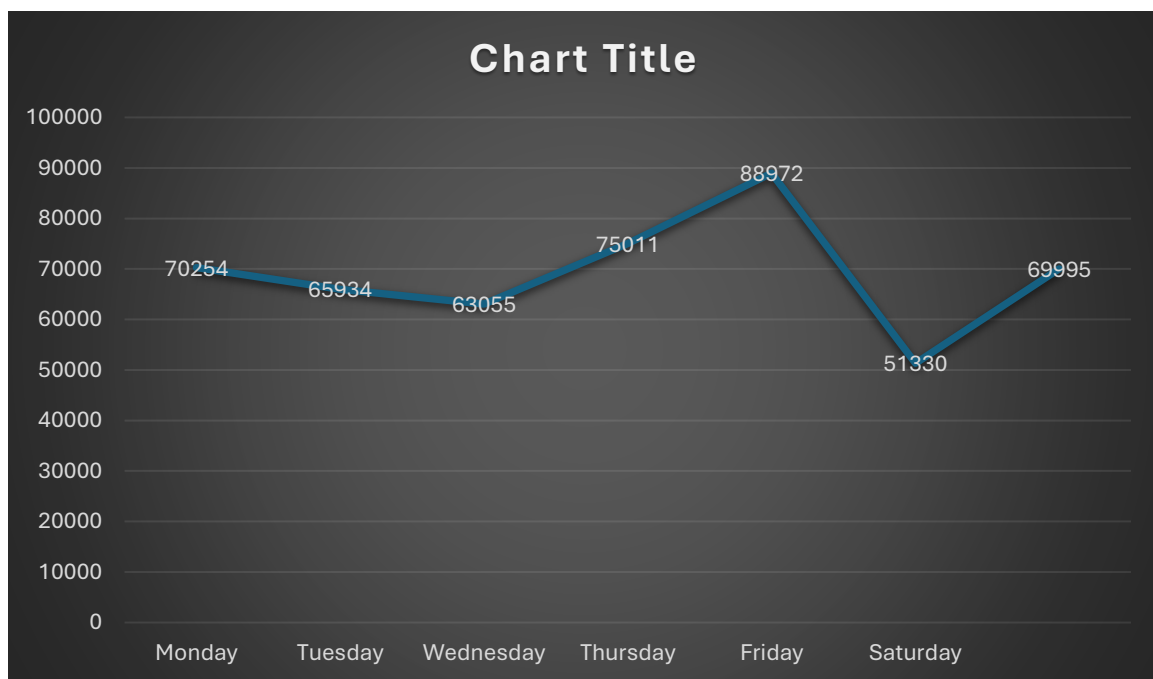
Infosys SpringBoard Internship

Dataset:

- **Kaggle:** Flight Delay and Causes
- <https://www.kaggle.com/datasets/underscore01/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.

Key Performance Metrics:

Flights by Day:

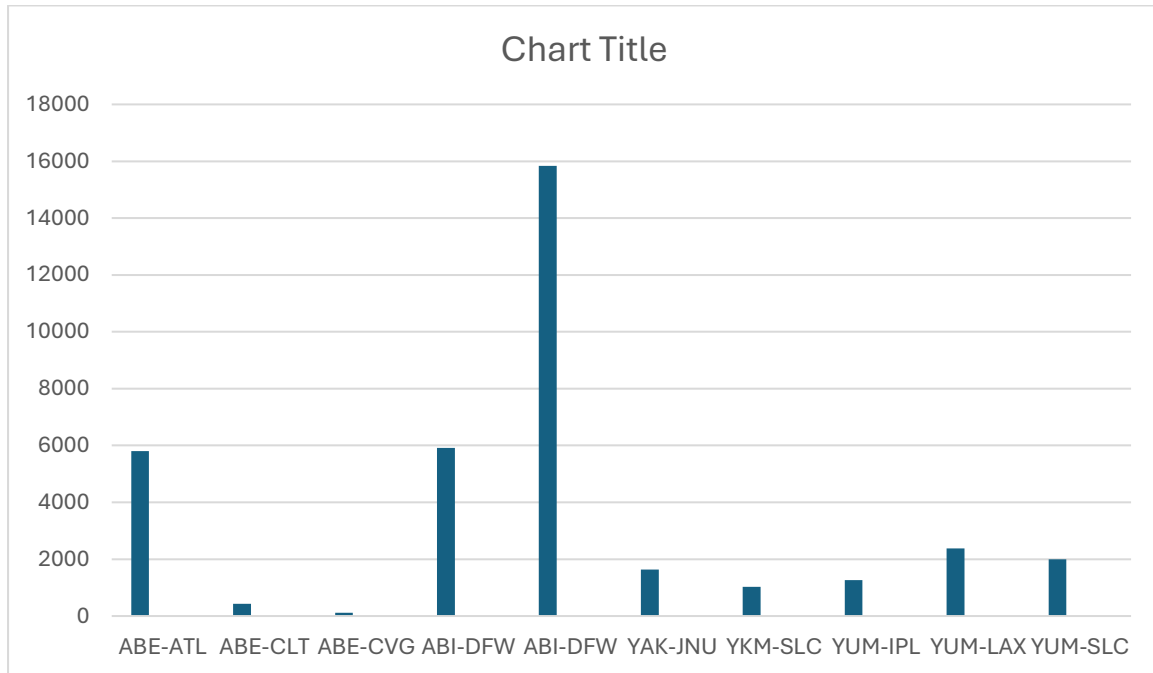


Distance:

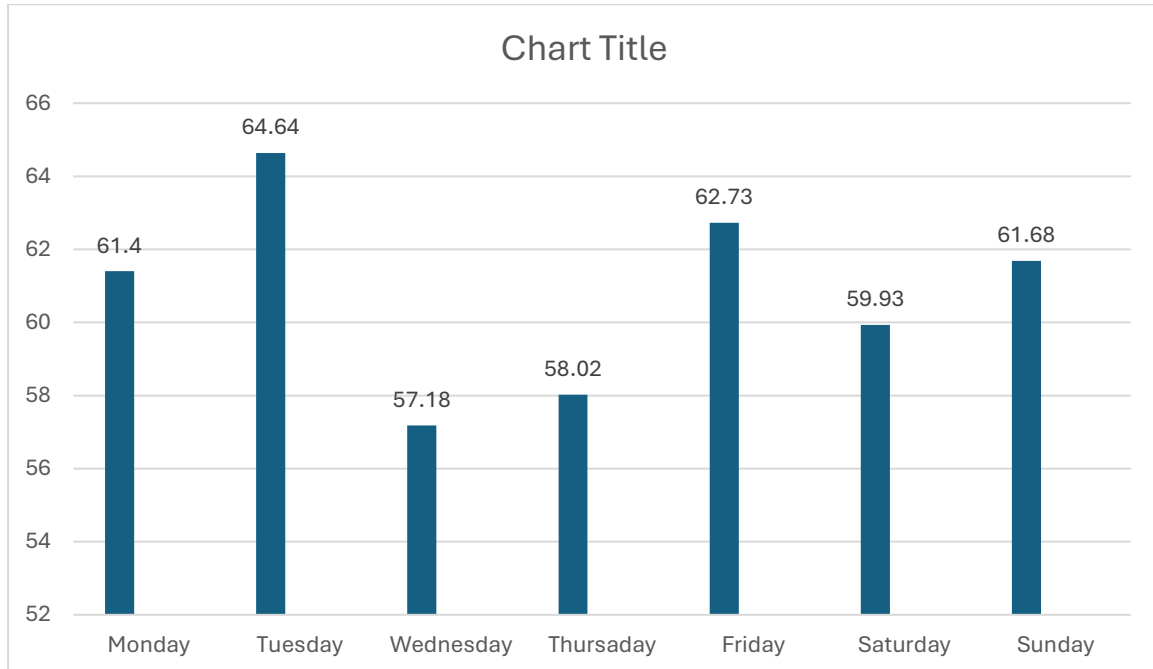
- a. Minimum Flight Distance: 31 miles
- b. Maximum Flight Distance: 4,502 miles
- c. Average Flight Distance: 752.3844 miles

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

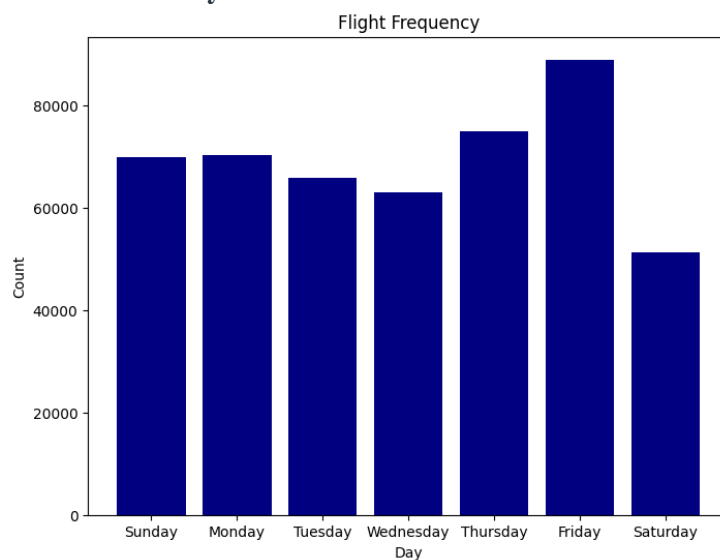
- **Month:** Extracted from Date column.
- **Hour:** Extracted from Date (departure hour).
- **Route:** Created by concatenating Origin and Dest (e.g., *ORD* → *LGA*).
- **DayName:** Mapped DayOfWeek values (1–7) into actual weekday names (Monday–Sunday).
- **TotalDelay:** Computed as the sum of individual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

Final cleaned dataset saved as:

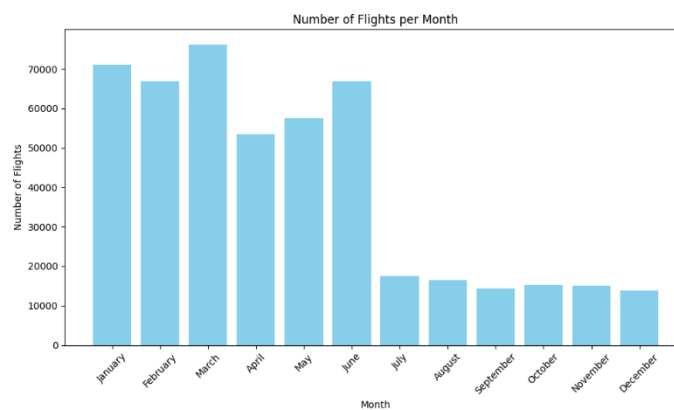
Flight_delay_cleaned.csv

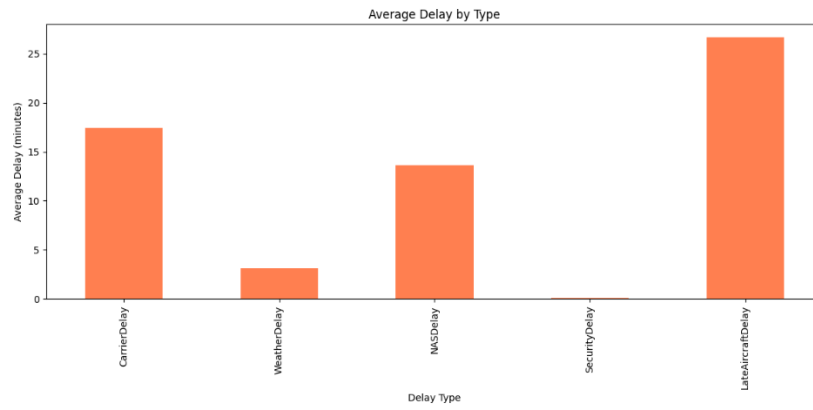
EDA(Exploratory Data Analysis):

- **Univariate Analysis**

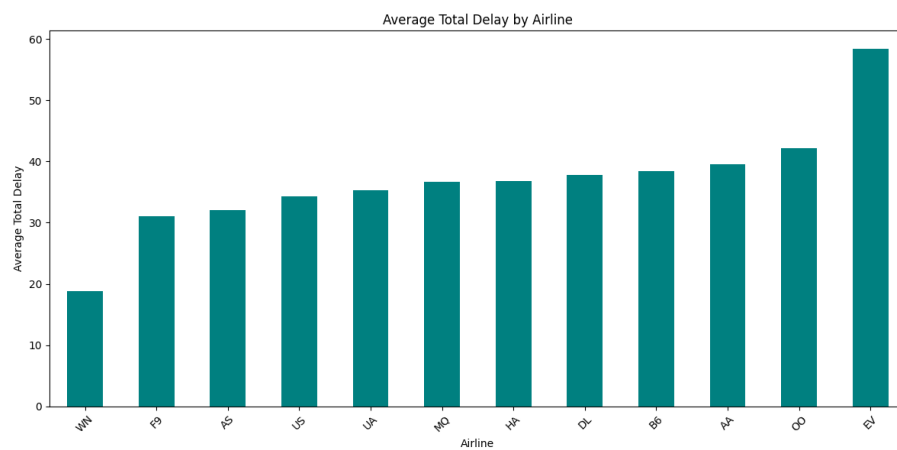
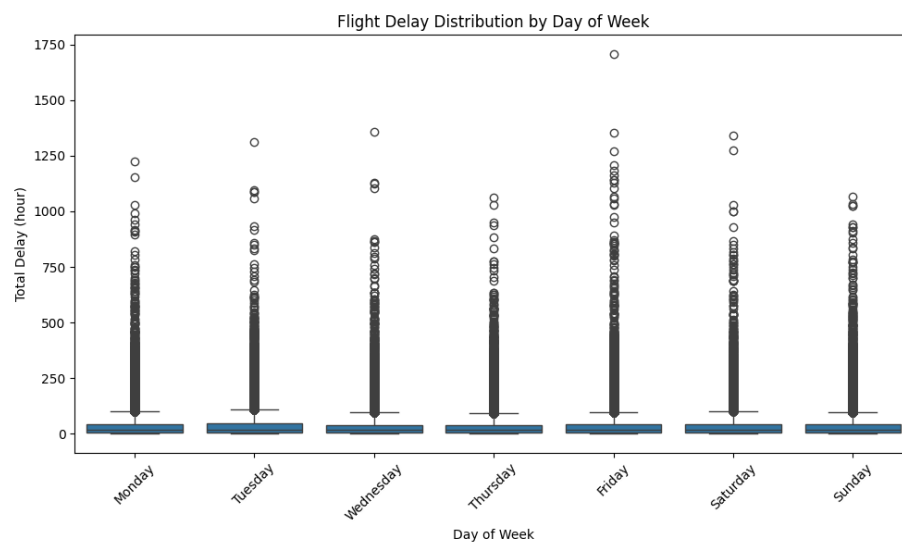


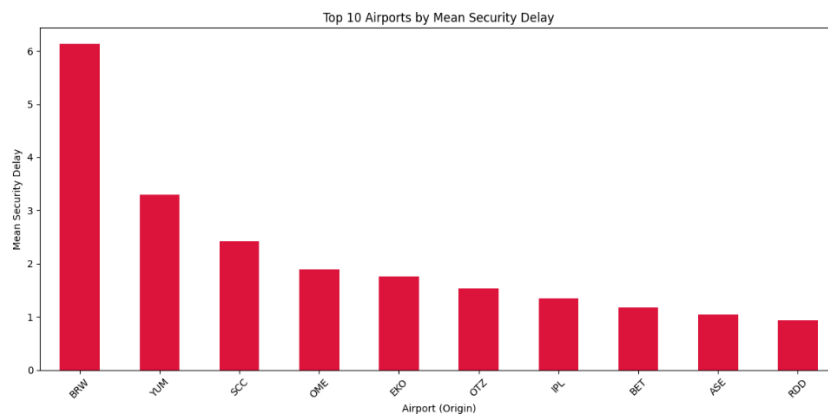
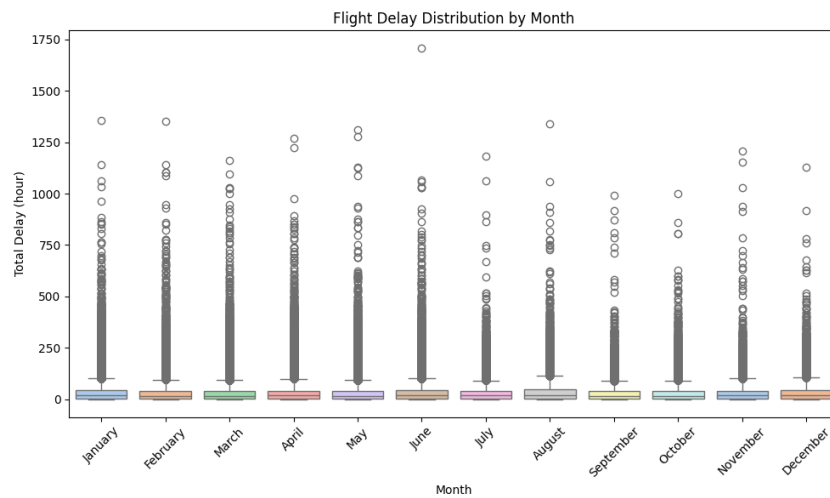
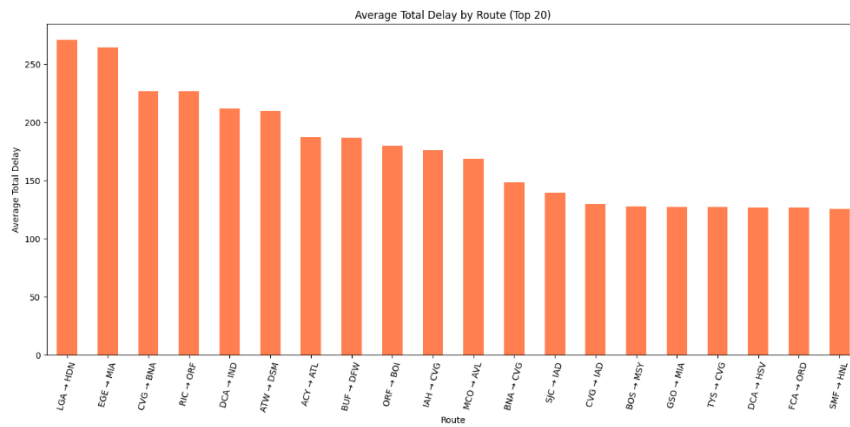
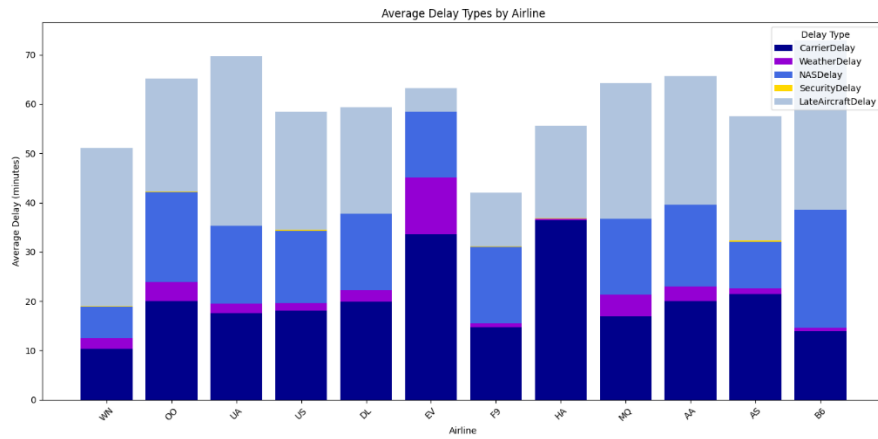
Flight frequency



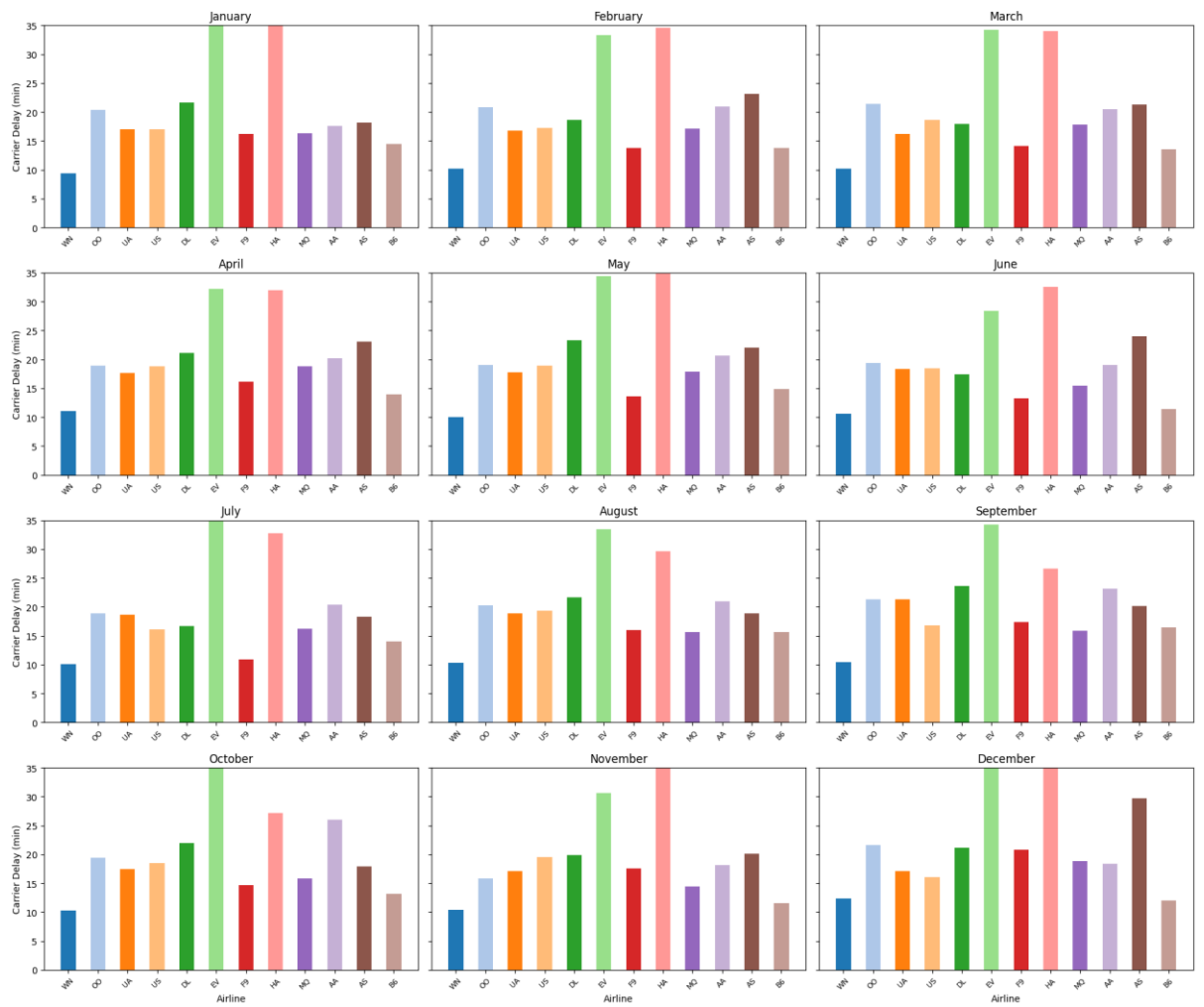


- Bivariate Analysis





Carrier Delay vs Airlines by Month



Count of Departure Delays by Time of Day

