Study of Delays Prediction in the US Airline Network

Ahmad Latiffi
Nicholas Dubois
Gavin Frings
Nurul Sapari
Mariya Siddiqui

Airline Delay Data

July 2023 US Flight Records
Source: DOT Bureau of Transportation Statistics

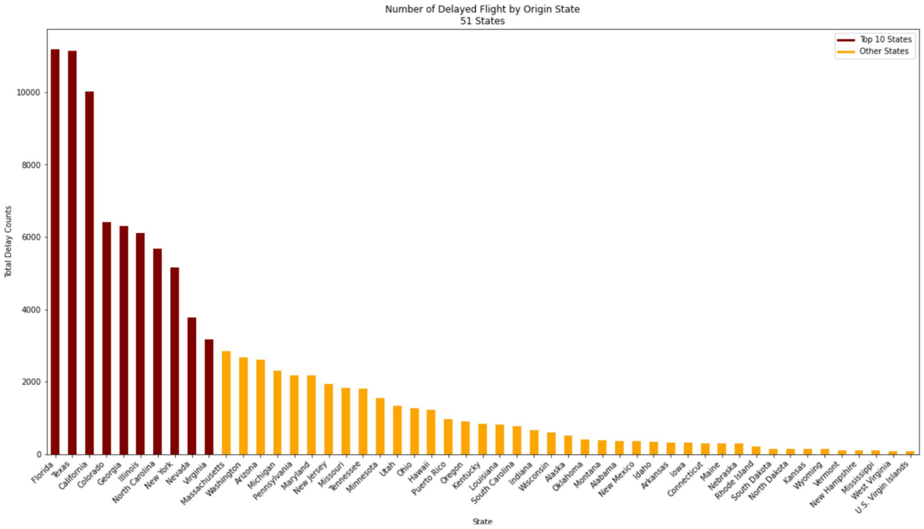
Metrics

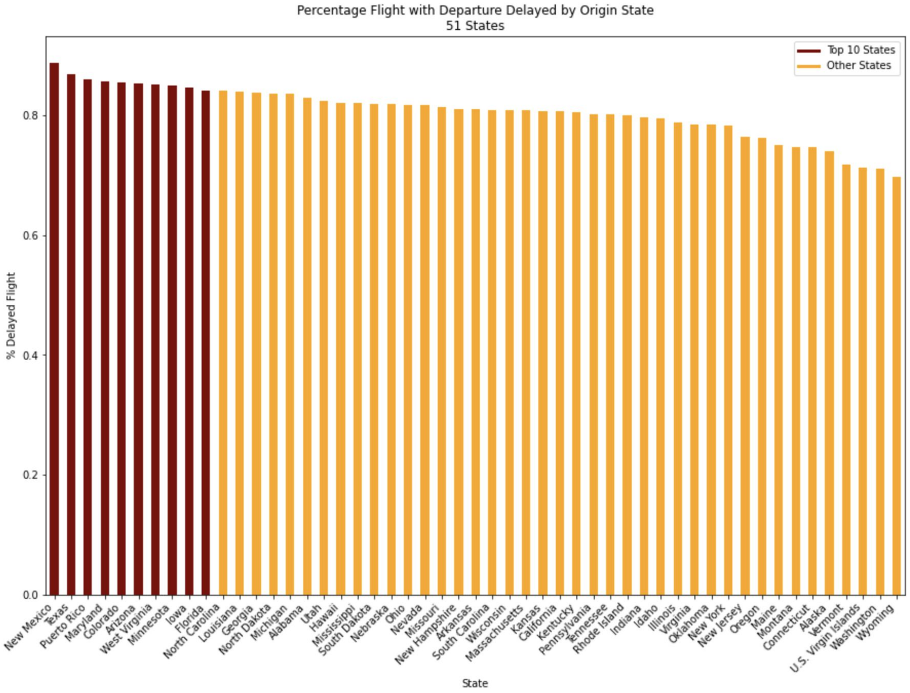
- Flight Delayed 15+ minutes (True/False)
- Length of Delay (minutes)

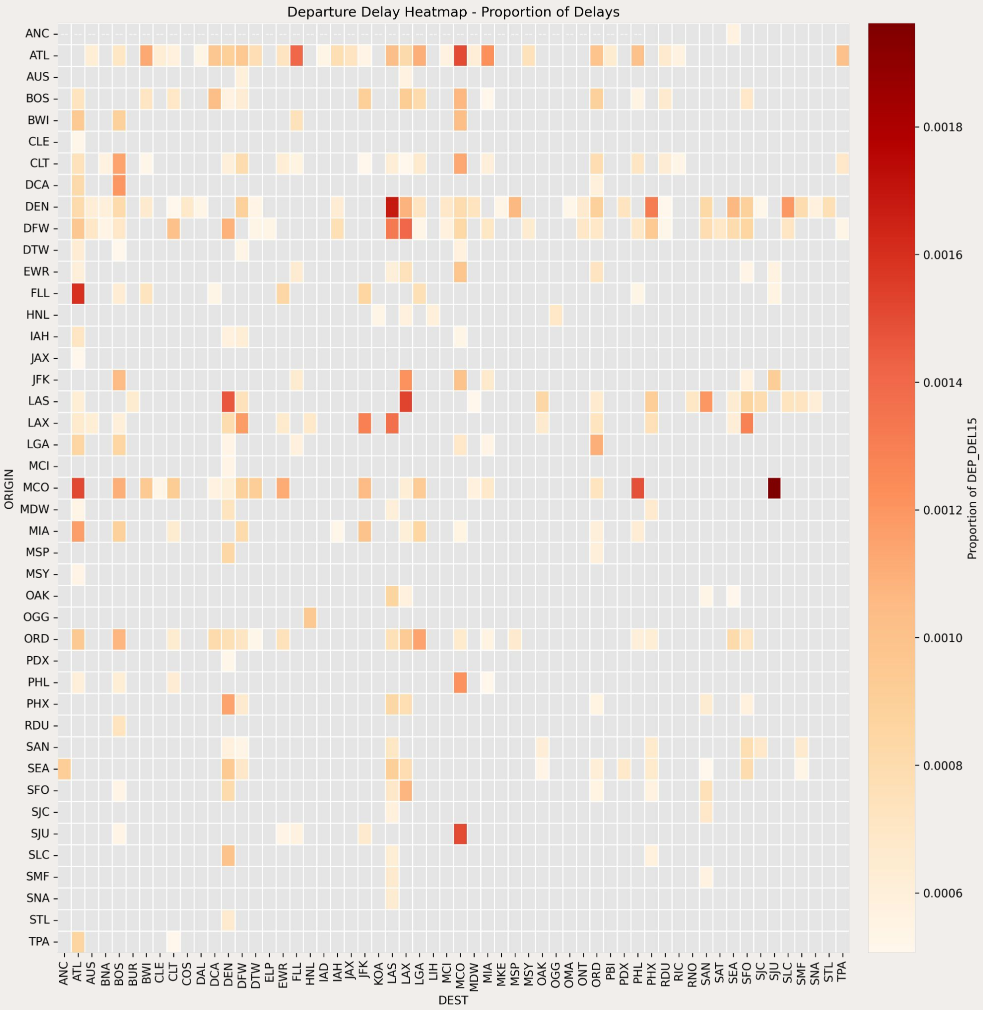
Factors

- Carrier
- Origin State
- Departure Time

Distribution of Delay Count by Origin State


Number of Delayed Flight by Origin State
51 States

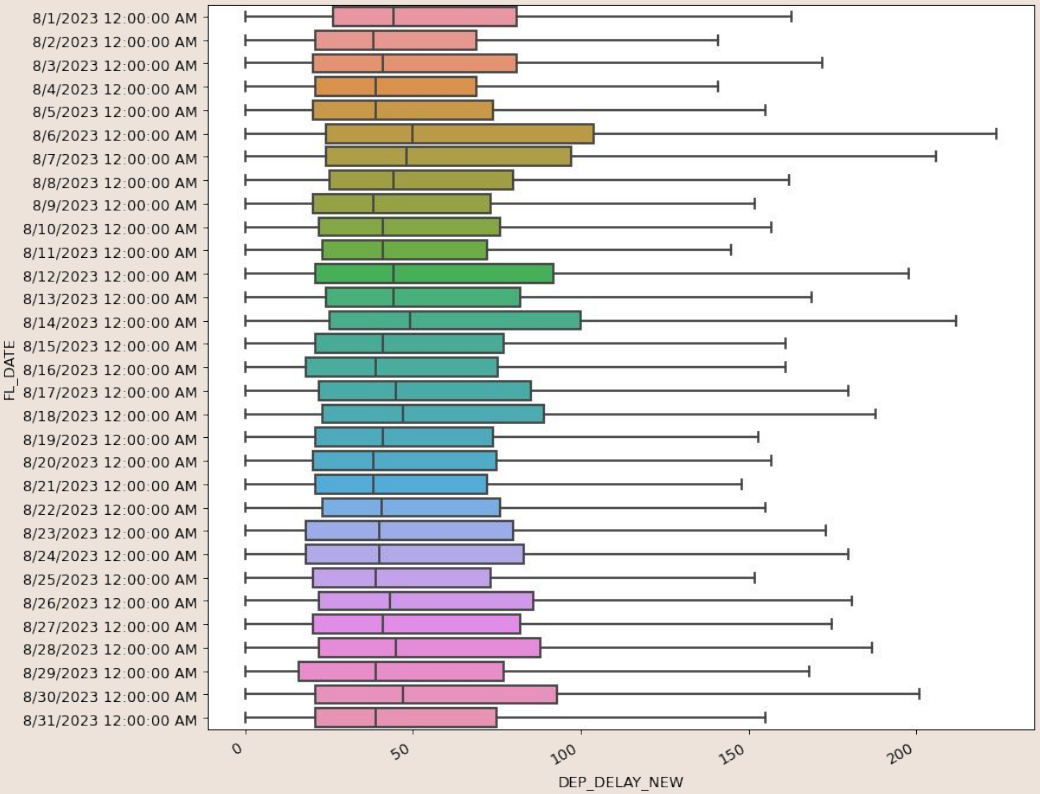

Percentage Flight with Departure Delayed by Origin State
51 States



Heatmap: Proportion of Flight Delays across the entire set of flights

- Gray boxes - cases where there are no flights between the origin (row) and destination (column) states (NaN)
- Light yellowish boxes - state pairs where delays occur
- Red boxes - hotspots where the delay occur the most across the origin-state pair
  - (MCO to SJU)
  - (DEN to LAS)
  - (FLL to ATL)

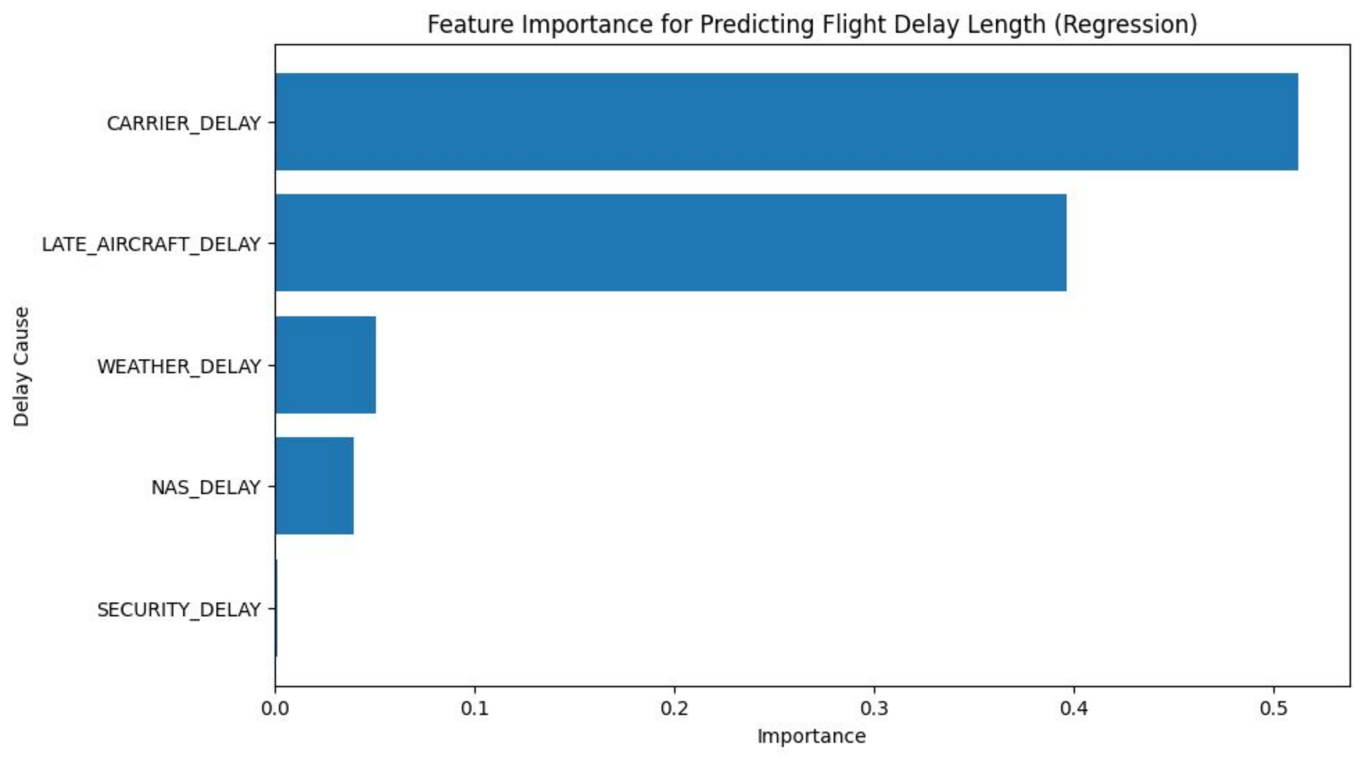
Delay Distribution by Date



Difference in minutes between scheduled and actual departure time:

- Median departure delay time is 50 minutes
- Greater skew toward longer delays over five days
  8/5/2023 - 8/9/2023
  8/25/2023 - 8/31/2023
  Factors ?
  - Weather patterns
  - Increase in number of passengers
  - The end of summer break

Feature Importance



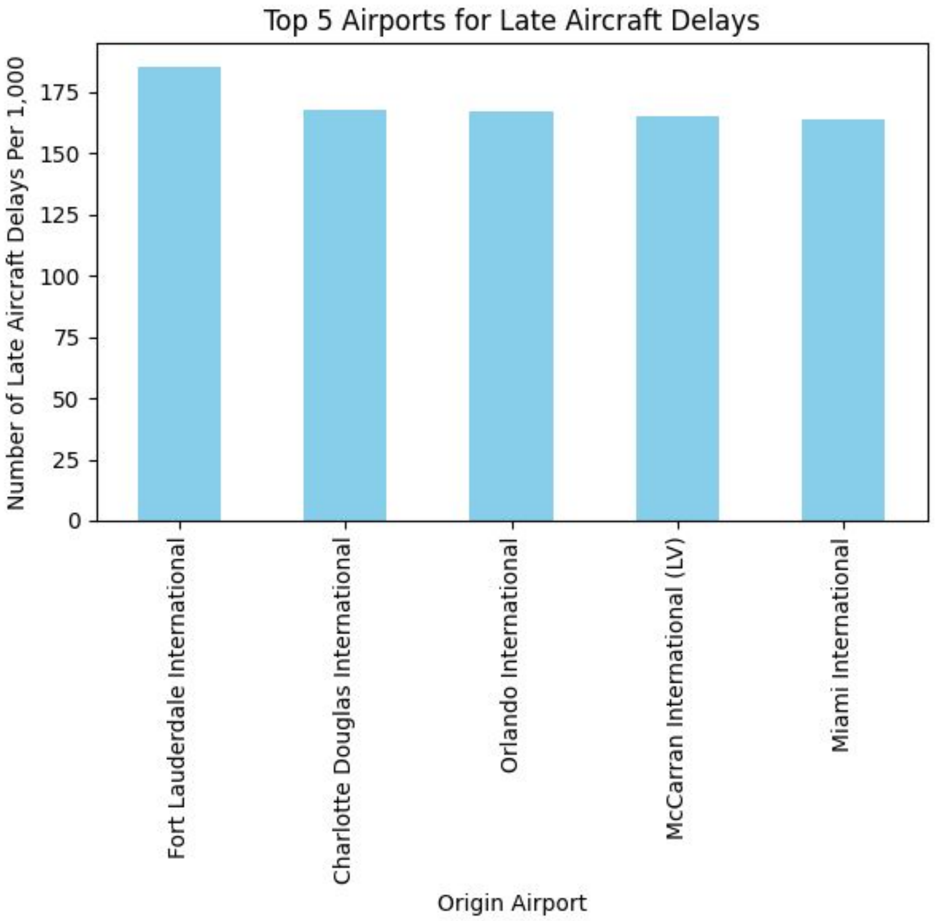
Carrier - In control of air carrier, i.e. baggage, cleaning/damage, fueling, etc.

Late aircraft - plane arrives late, causing delay for next flight

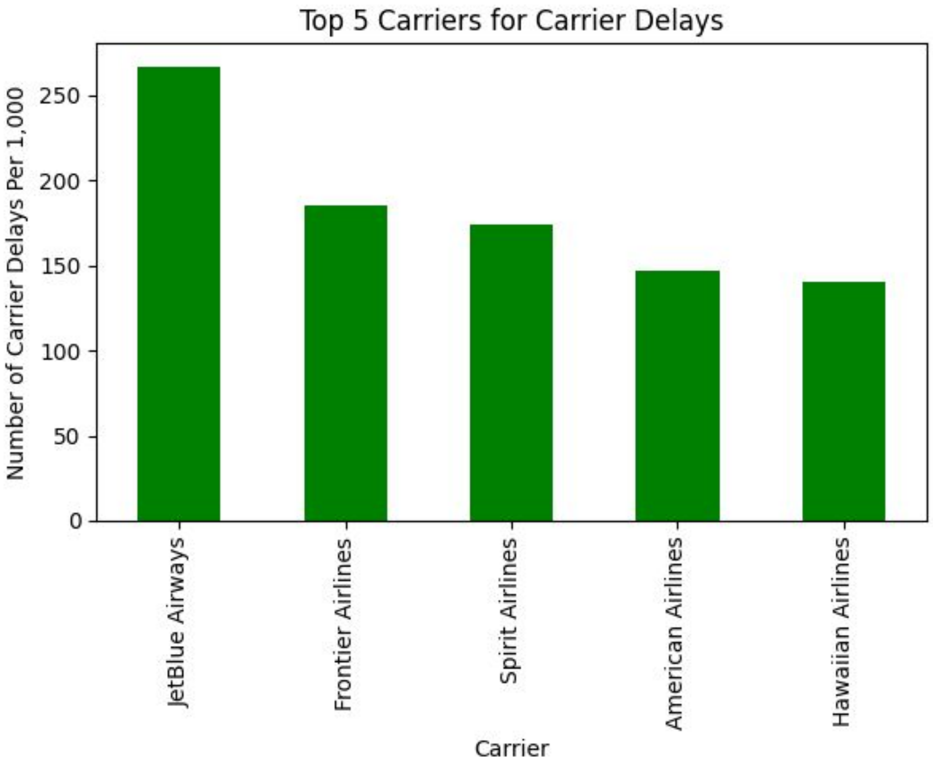
Weather - extreme conditions causing inability to take off

NAS - National Airspace Security, includes air traffic/runway control

Security - Evacuation of terminal or reboarding due to security breach



Top 5 Airports for Late Aircraft Delays



Top 5 Carriers for Carrier Delays

Difficulties

Data Size

- 600,00+ flights in one month
- Large number of qualitative variables

Limited factors to base prediction on

- Weekday, origin state, departure time, carrier

Missing Data

- Some rows have NaNs that do not match existing patterns.
- May not reflect reality

Results of Classification

Proportion of Flights Delayed 15+ Minutes: **0.288 (1 - 0.712)**

Highest Classification Accuracy: **0.733**

- Gradient Boosting and Logistic Regression

Results of Regression

Variance of Delay Time: **4800**

MSE of Ridge Regression: **5000**

- R^2: **0.04**

Conclusions

We are unable to effectively predict airline delays

Limitations

- Insufficient Data
  - Other factors such as weather could benefit the analysis
- Methods
  - Current techniques may not be advanced enough