ABDULLAH SAHIN

FLIGHT DELAY PREDICTION USING LOGISTIC REGRESSION

OUTLINE

- Problem
- Data Gathering
- Exploratory Data Analysis
- Feature Engineering
- Logistic Regression
- Conclusion and Future Work
- Recommendation for Clients

PROBLEM

In the last ten years, according to the Bureau of Transportation Statistics (BTS), only 79.63% of all flights have performed on time. Only a few remaining percentage were cancelled or diverted, less than 2%; rest of them were delayed mainly due to late arriving aircraft followed by the cause of the national aviation system and air carrier.

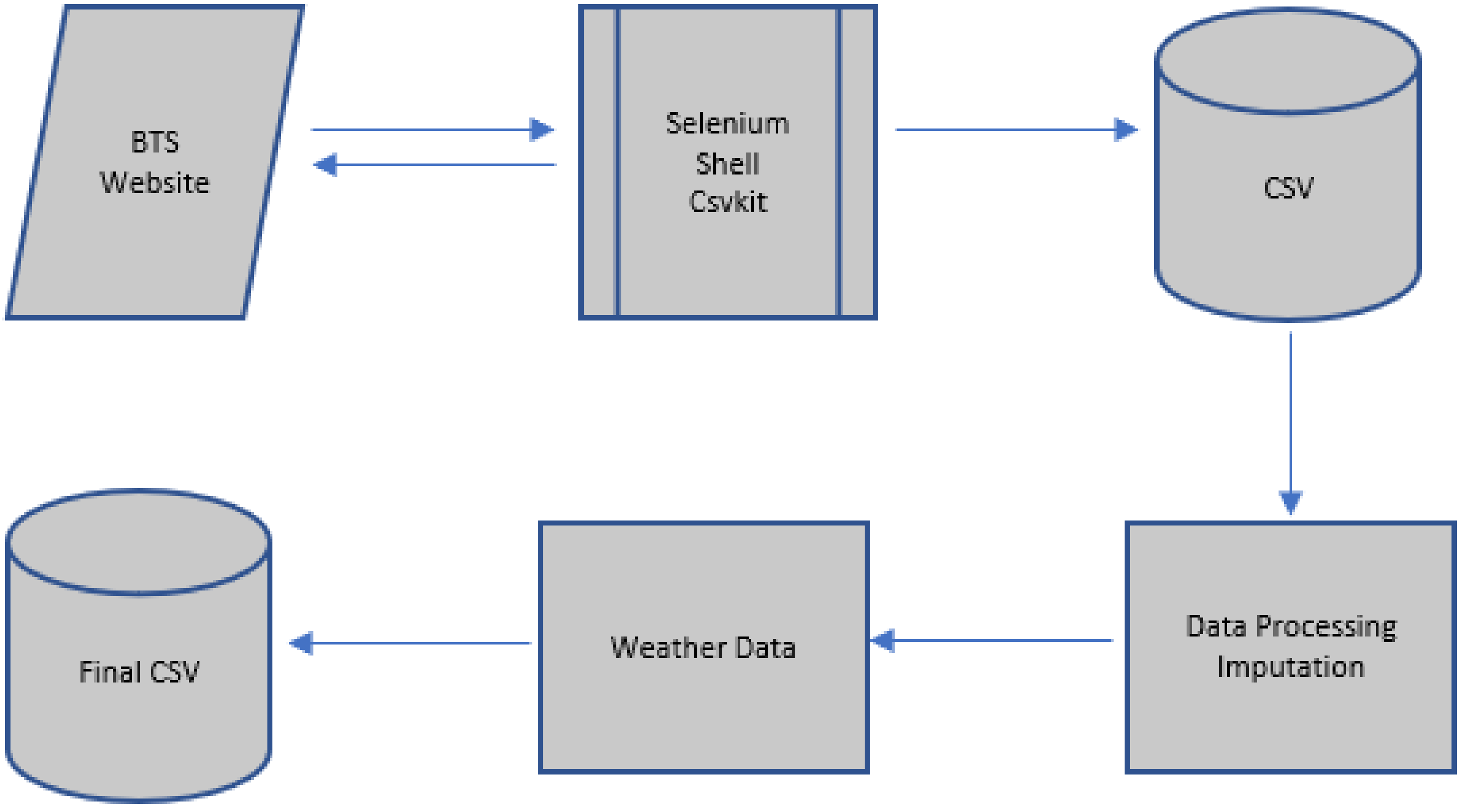
These series of delays cause a serious financial burden on airlines. In 2010, the Federal Aviation Administration commission estimated that flight delays cost the airline companies $8 billion a year, most of which due to increased budget on crews, fuel and maintenance

Overall, the findings of this study can provide a high-profile achievement by addressing aviation delay problems with a robust prediction model and helping people and businesses better on planning their flights.

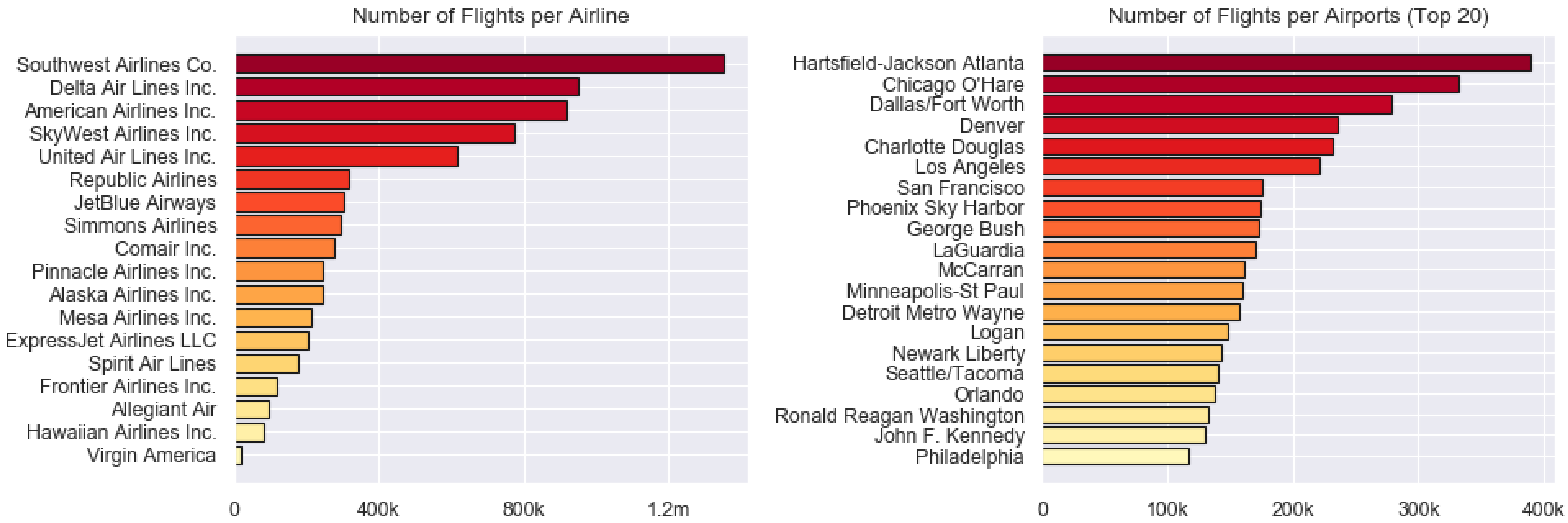
DATA GATHERING

- Airline On-Time Performance Data: is available on The Bureau of Transportation Statistics' website.
- Weather Data: was obtained from Iowa State University's Environmental Mesonet Platform.
- Airport Data: obtained from The Bureau of Transportation Statistics' website manually
- ICAO Codes: acquired from OpenFlight dataset.

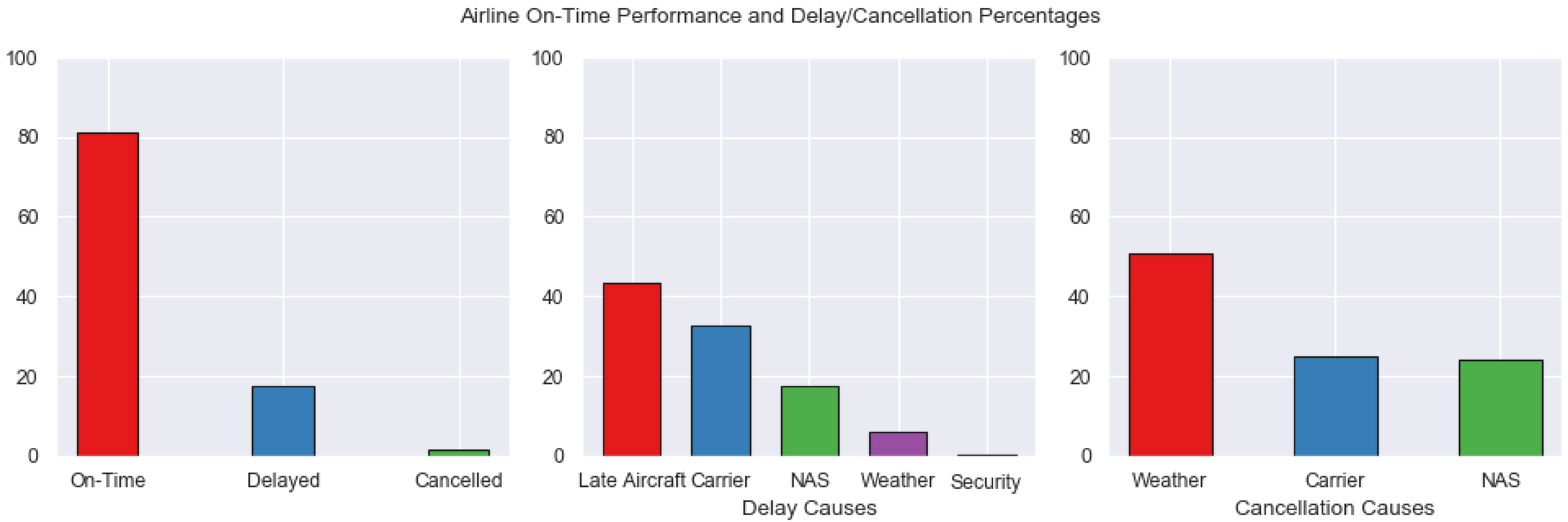
FLOW CHART


BTS Website
Selenium Shell Csvkit
CSV
Data Processing Imputation
Weather Data
Final CSV

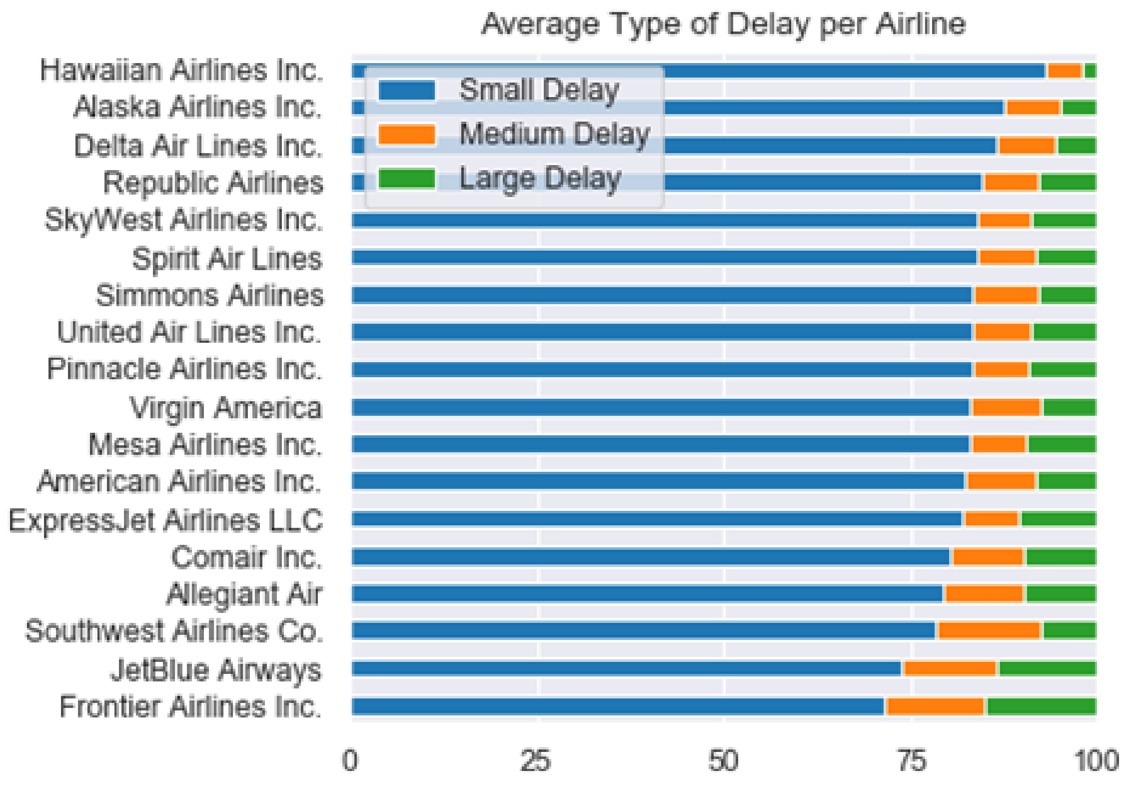
EXPLORATORY DATA ANALYSIS

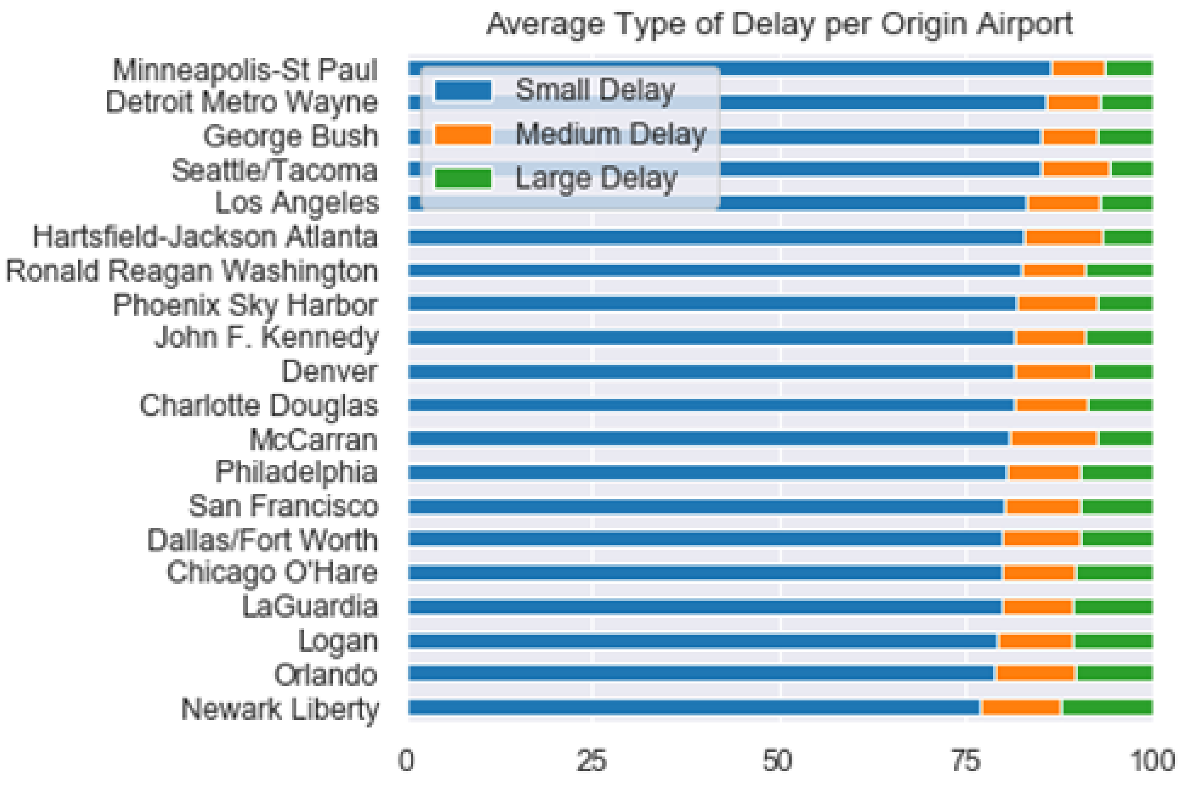


BUSIEST AIRLINES AND AIRPORTS

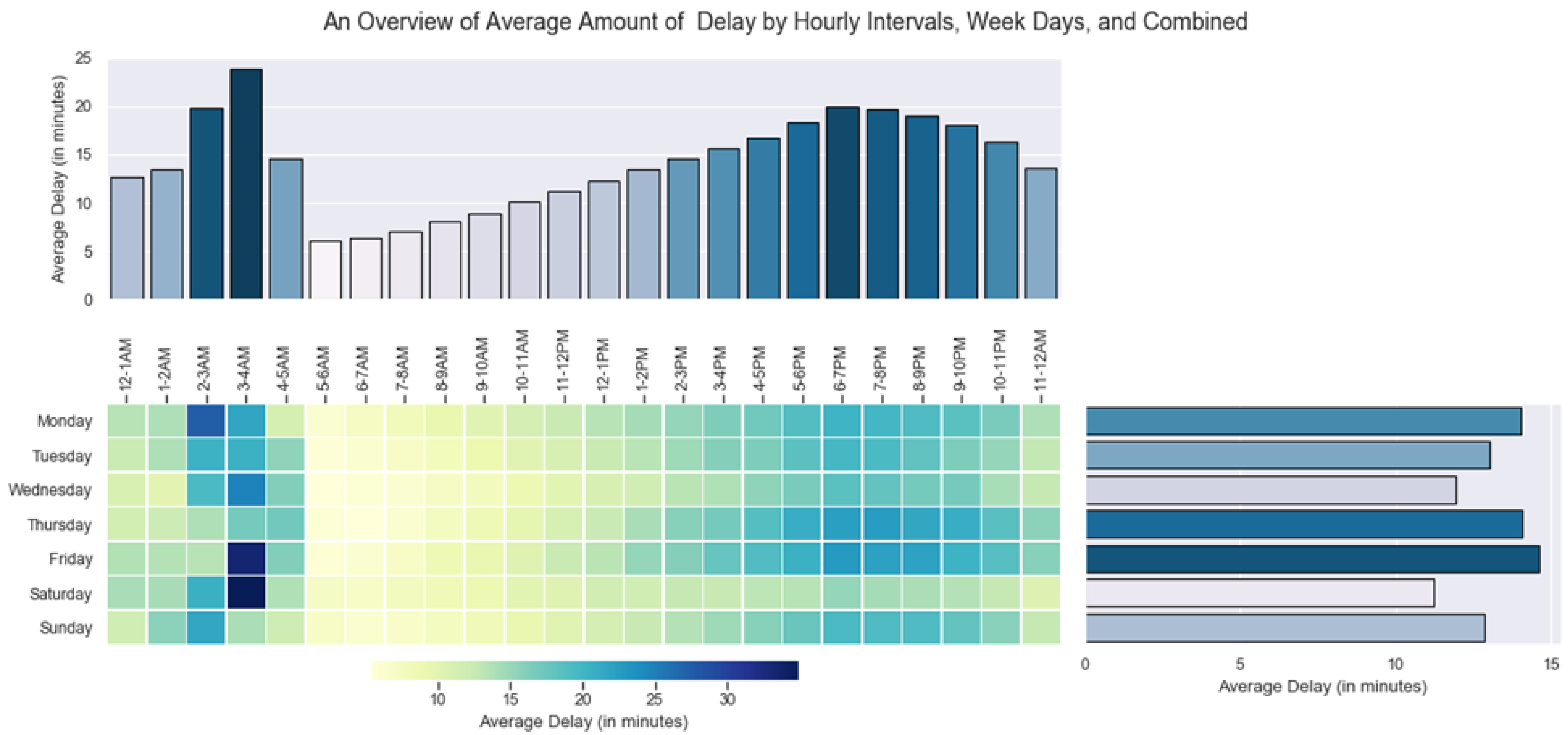


AIRLINE ON-TIME PERFORMANCE

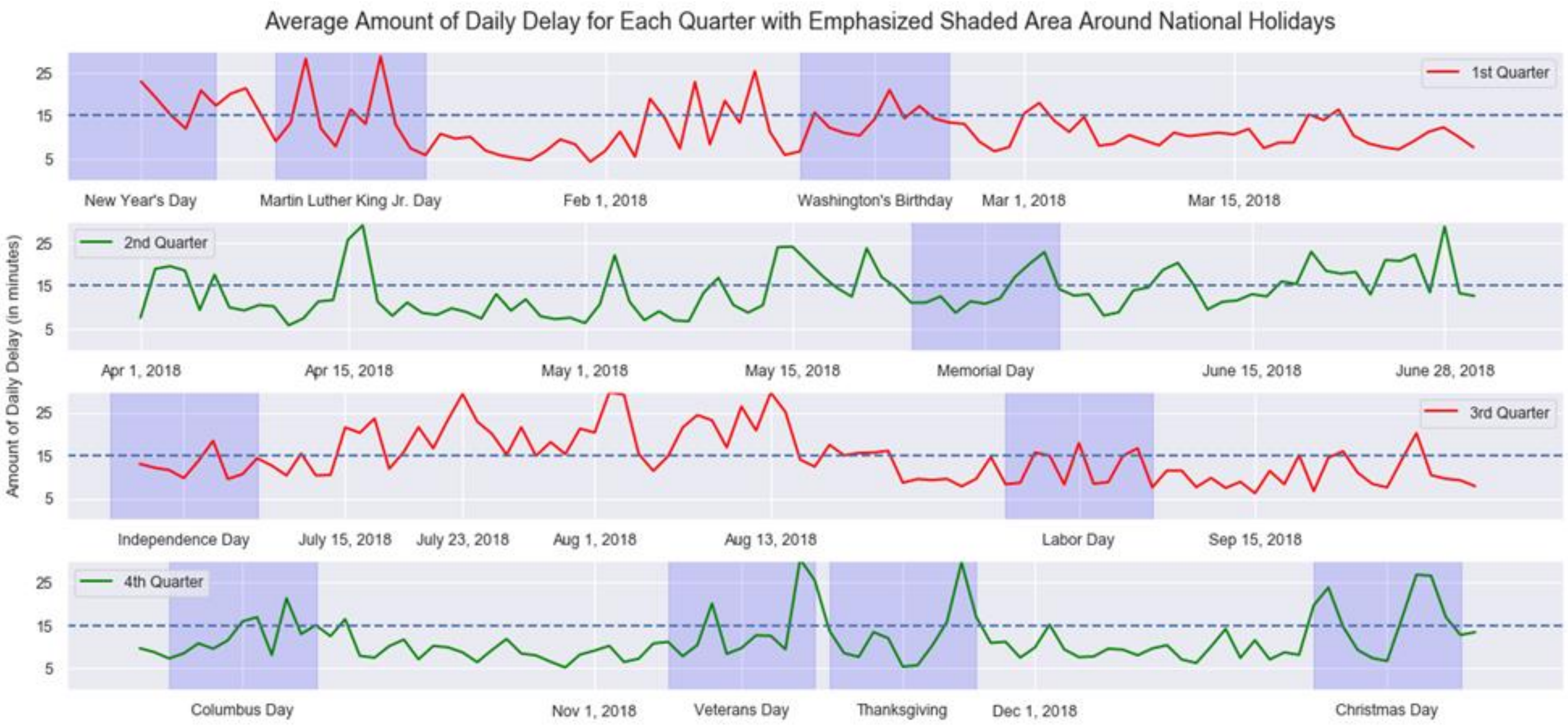




AIRLINE ON-TIME PERFORMANCE PER AIRLINE COMPANIES AND AIRPORTS

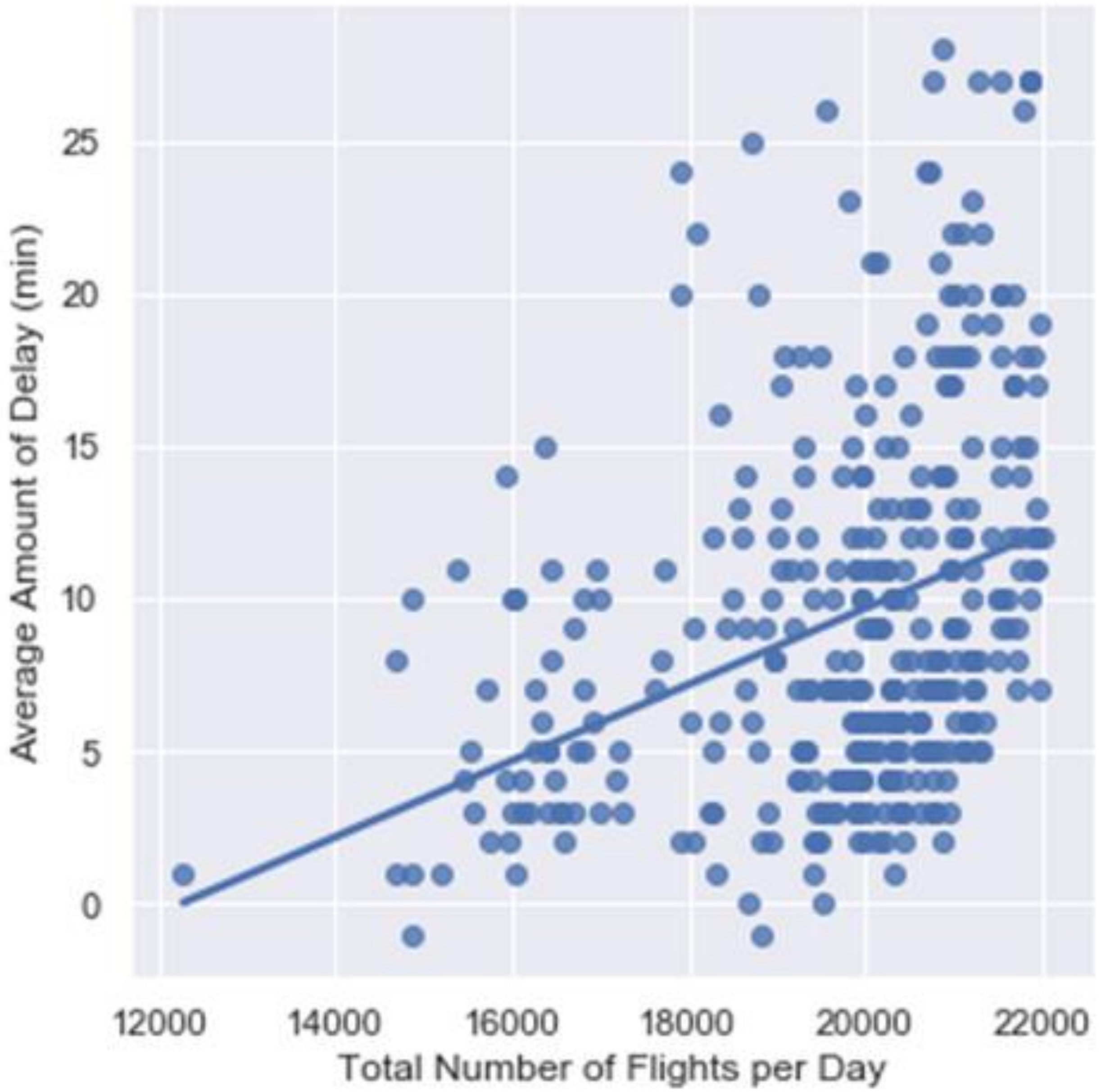


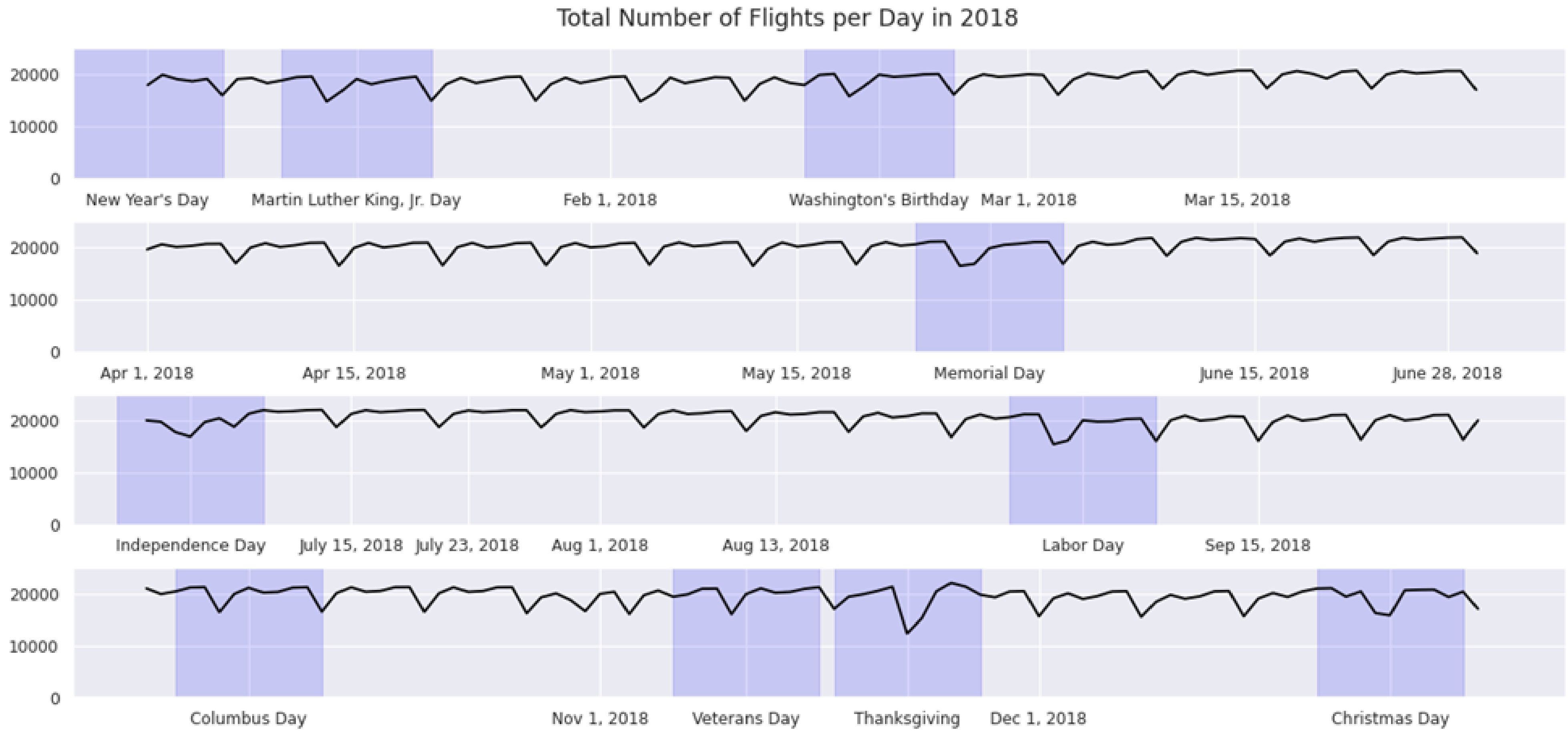
EFFECT OF DEPARTURE TIME AND WEEKDAYS OF A FLIGHT ON DEPARTURE DELAY



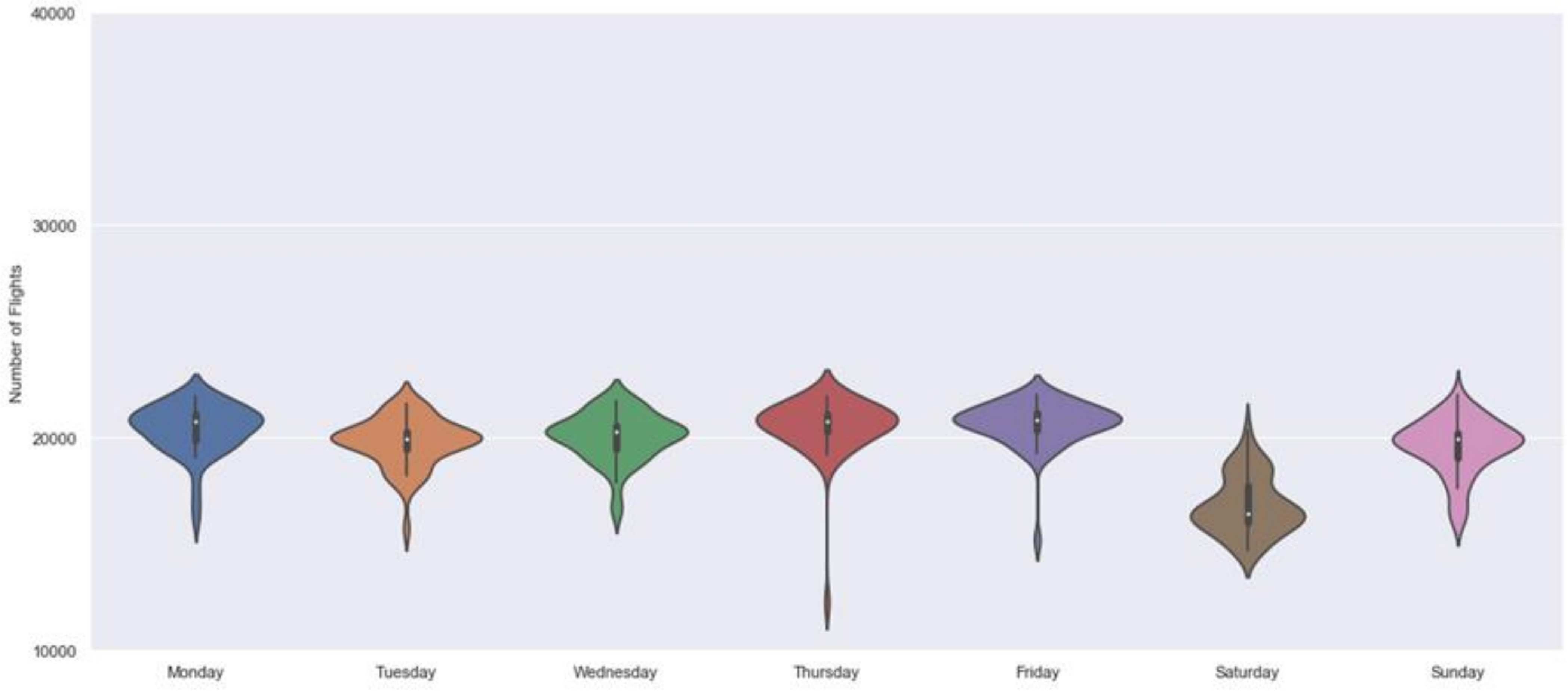
TRENDS BY QUARTERS

EFFECT OF TOTAL NUMBER OF FLIGHTS ON AVERAGE AMOUNT OF DELAY

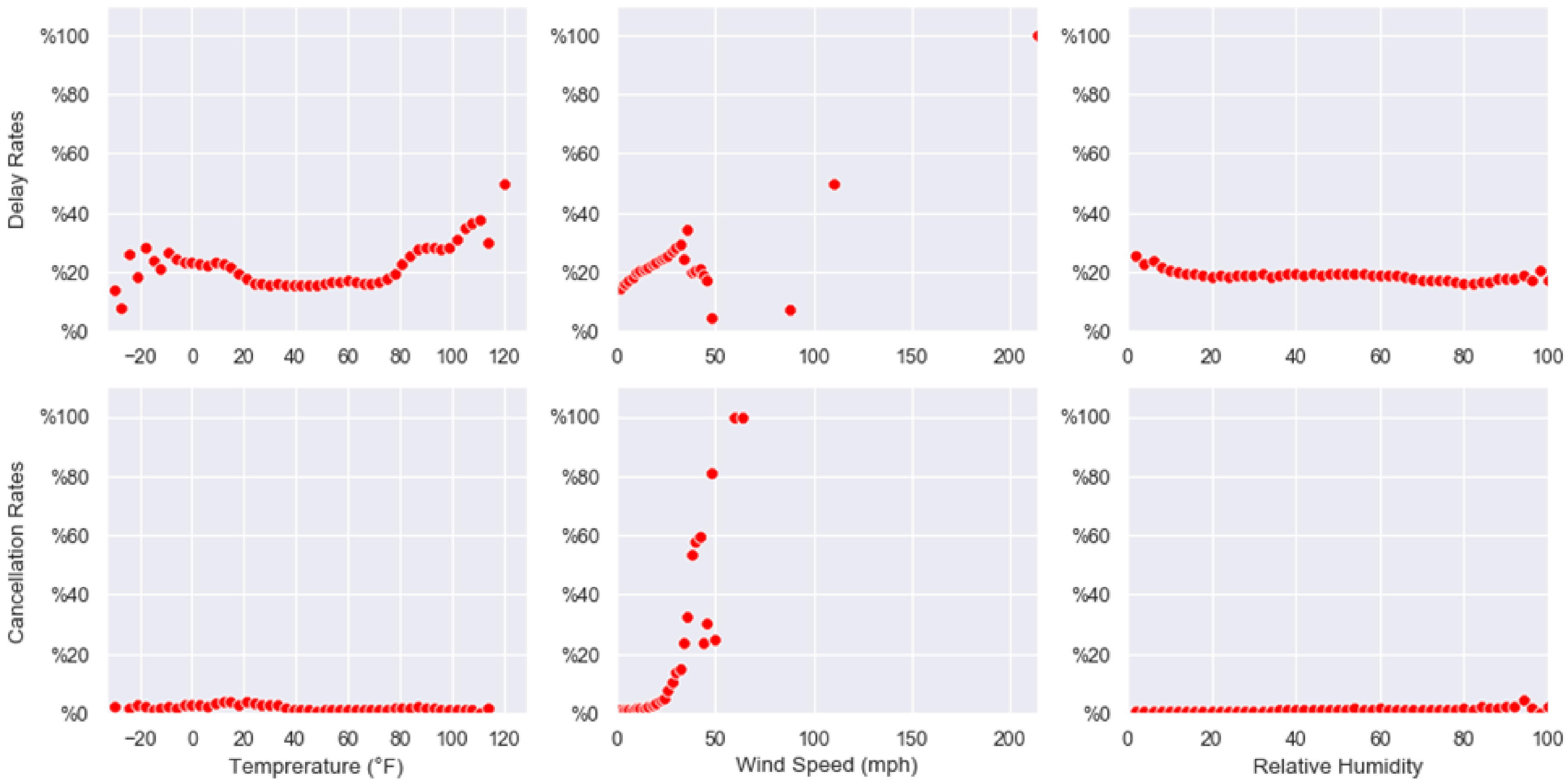




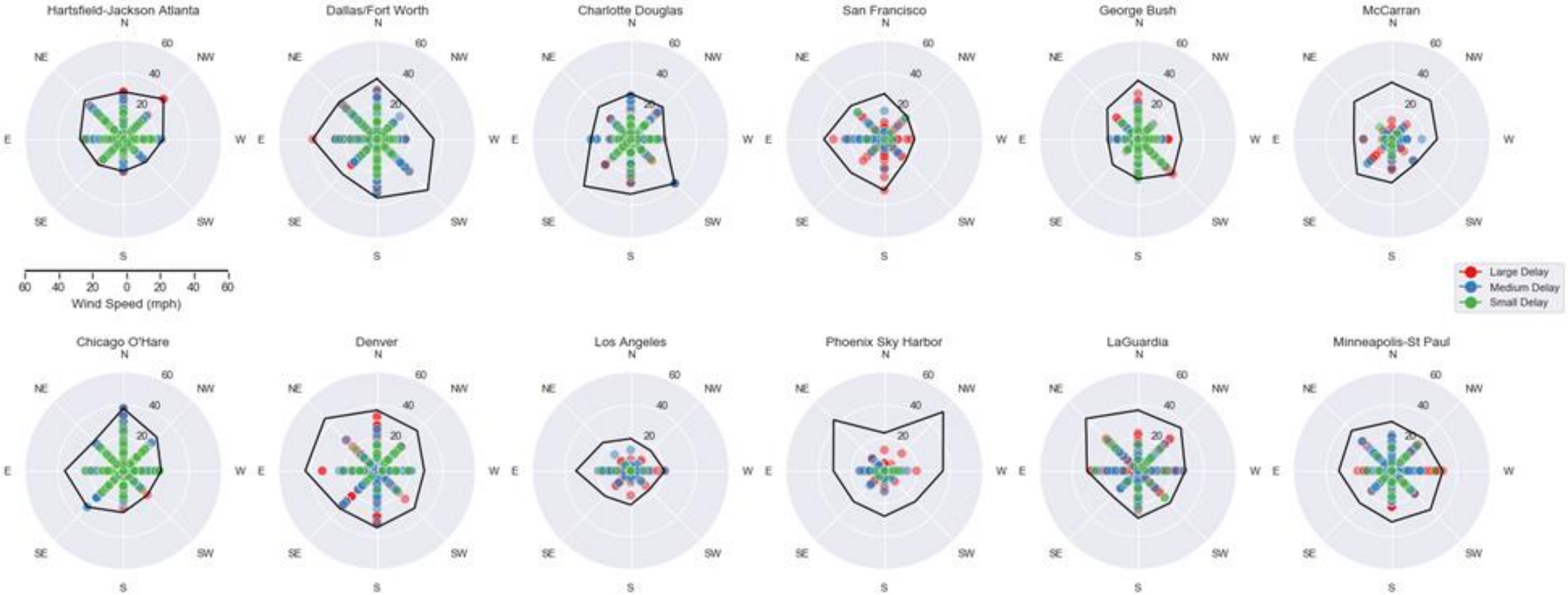
NATIONAL HOLIDAY'S EFFECT ON TOTAL NUMBER FLIGHTS



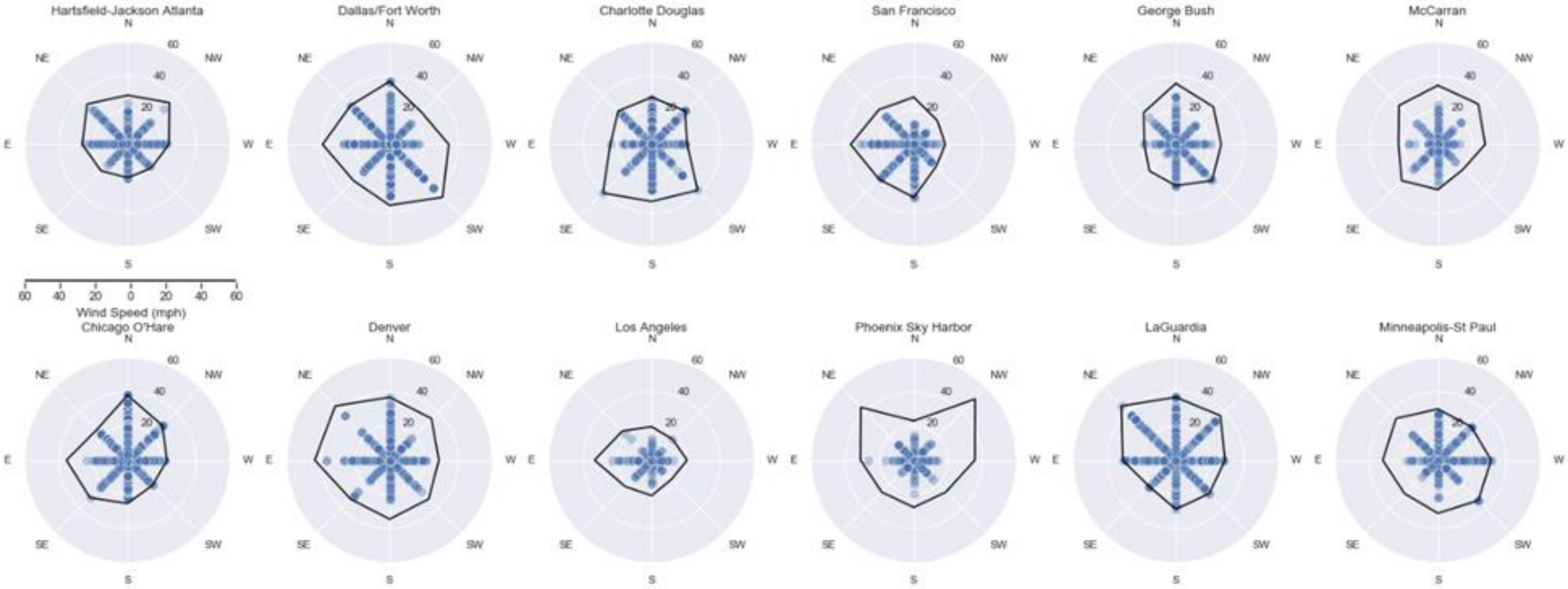
TOTAL NUMBER OF FLIGHTS (DAILY) DISTRIBUTION PER WEEKDAY



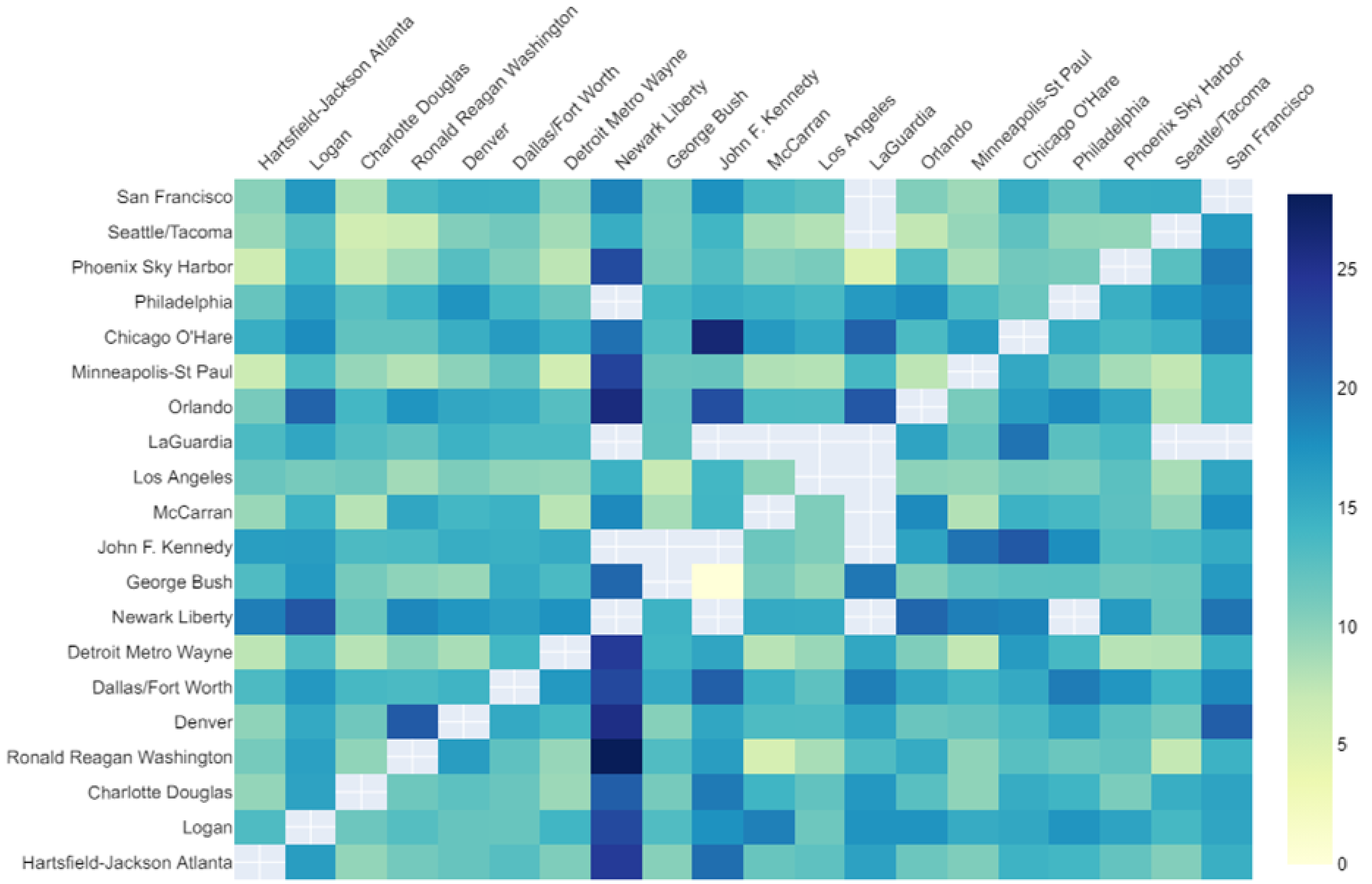
EFFECT OF WEATHER CONDITIONS ON DELAYS AND CANCELLATIONS



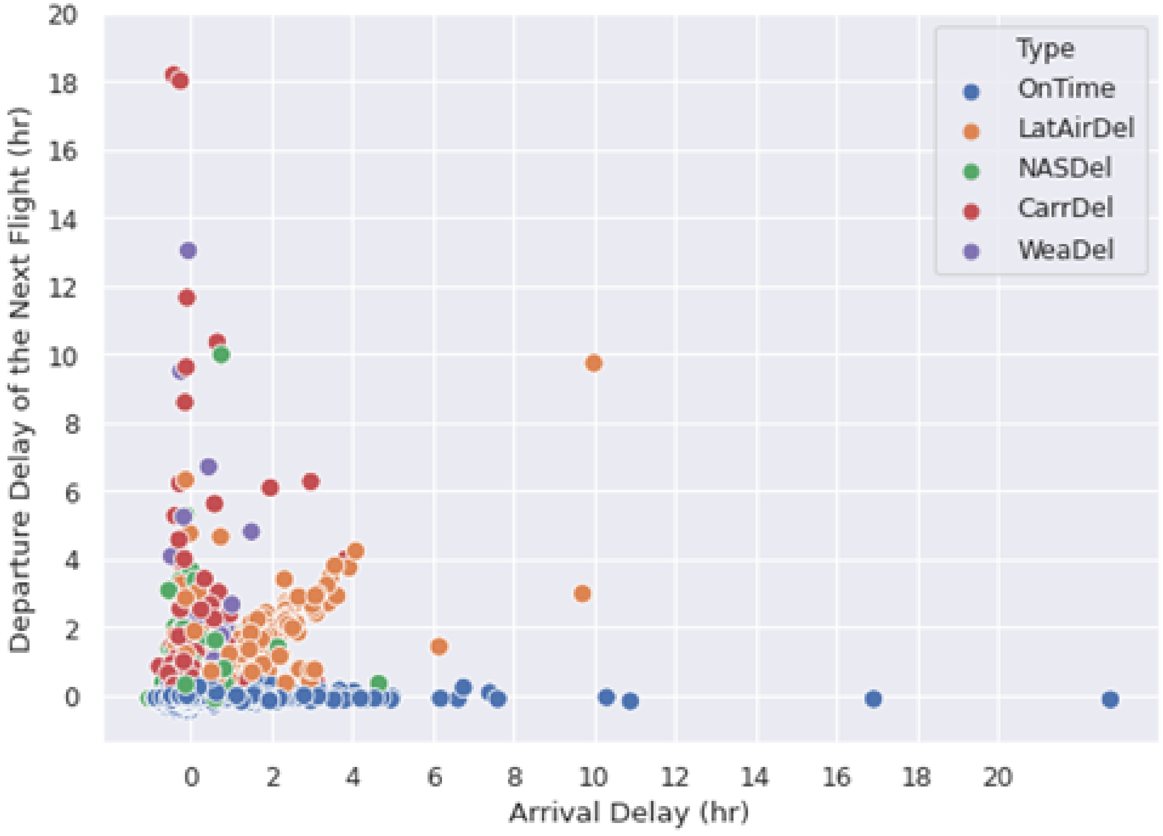
EFFECT OF WIND SPEED AND DIRECTION ON WEATHER CAUSED DELAYS

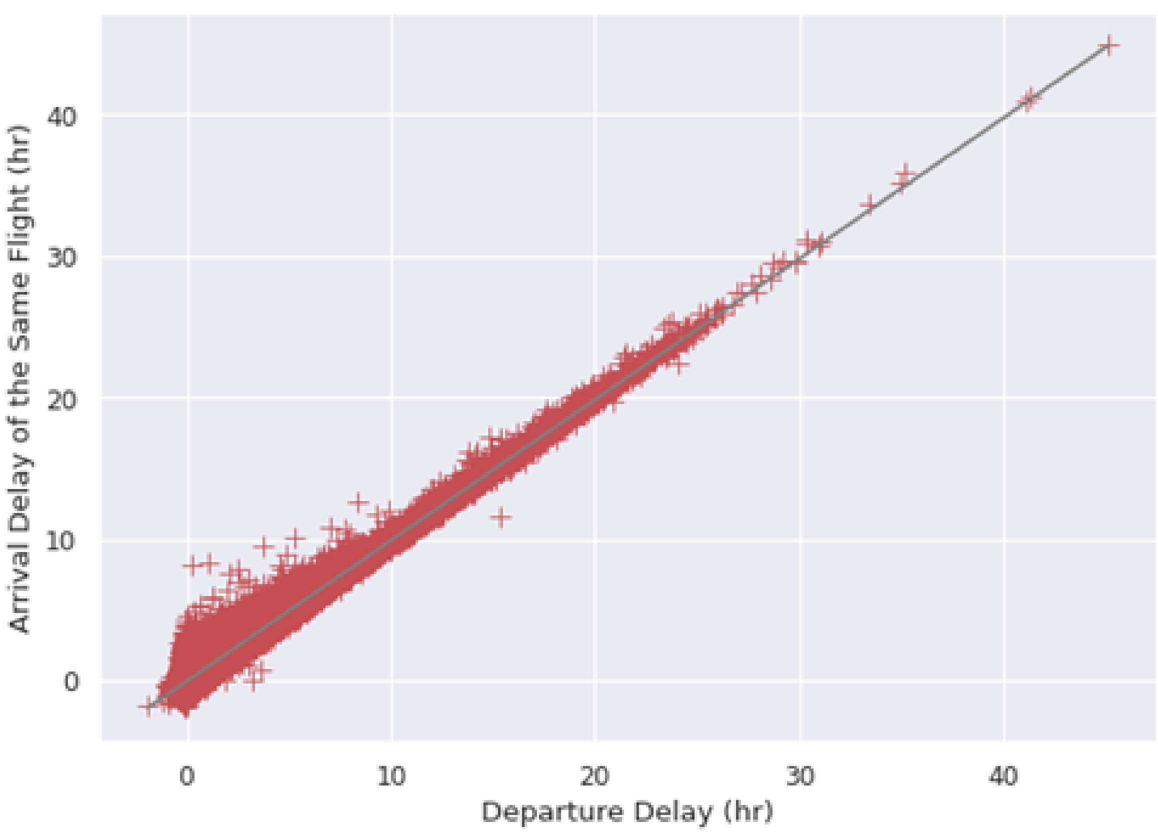


EFFECT OF WIND SPEED AND DIRECTION ON WEATHER CAUSED CANCELLATIONS



ORIGIN-DESTINATION PAIR AVERAGE AMOUNT OF DELAY





LATE AIRCRAFT DELAYS

FEATURE ENGINEERING

FEATURE ENGINEERING

Time: We start with first and foremost feature for a delayed flight: time-related data. These features will be dummy coded in the modeling part.

Holidays: For particular holidays, such as New Year's Day and Thanksgiving Day, average amount of delay is increased the set threshold, 15 minutes.

Number of Flights: Number of flights is not a function of how close a flight to the federal holidays. Number of flights can cause average amount of delay.

Airline and Airports: Certain airlines and/or airports have better performance than others.

Weather: Extreme weather conditions have an adverse effect on flight delays and cancellation.

Late Aircraft Delay: Almost half of the flight delays caused by late aircraft. It is crucial to take into account this effect.

LOGISTIC REGRESSION

LOGISTIC REGRESSION (BASELINE)

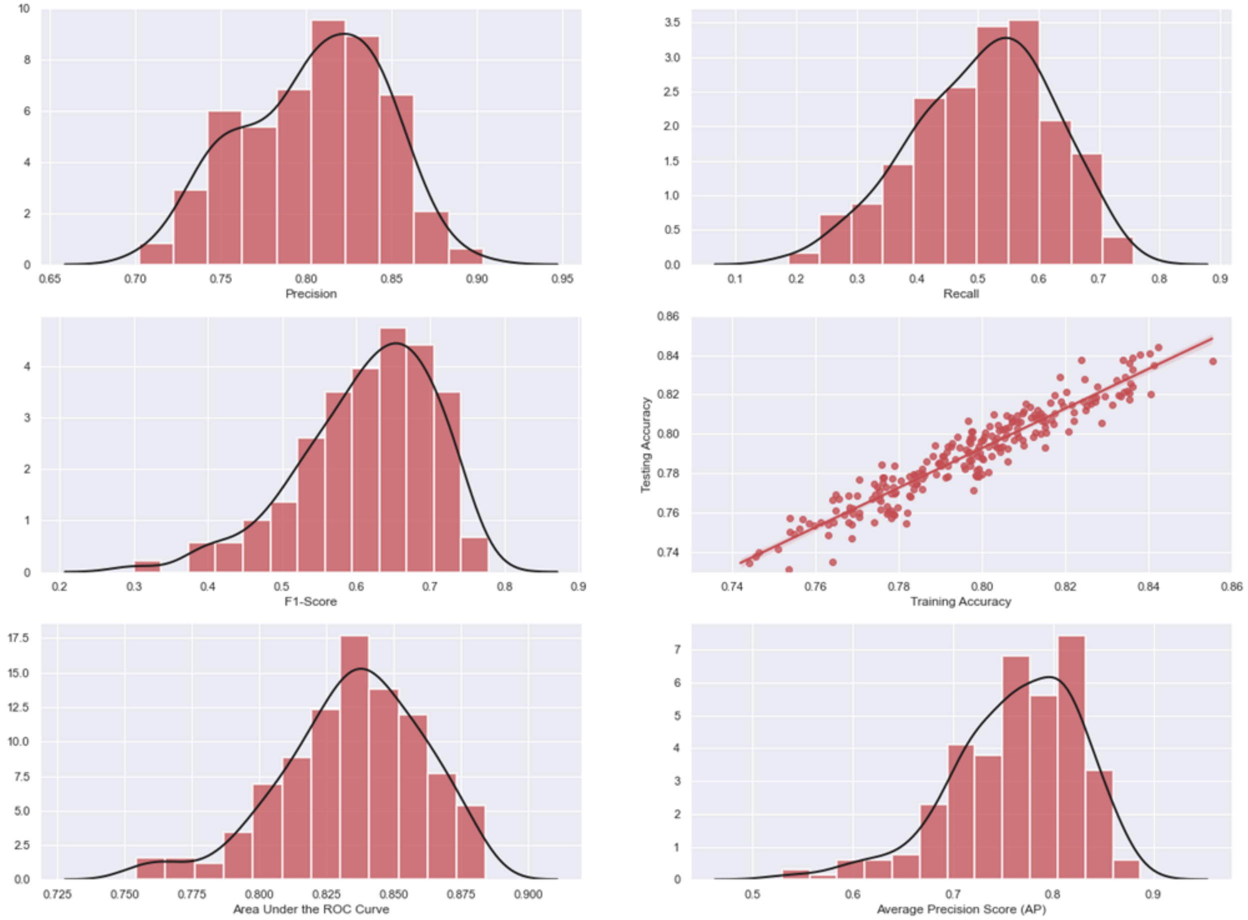
Few steps before modeling:

- One Hot Encoding: On categorical columns
- Data Splitting: 75-25
- Resampling: Random Over Sampling
- Hyperparameter Tuning with Cross Validation: C and class_weight
- Scaling: MaxAbsScaler

Origin-Destination Slicing

LOGISTIC REGRESSION (BASELINE)

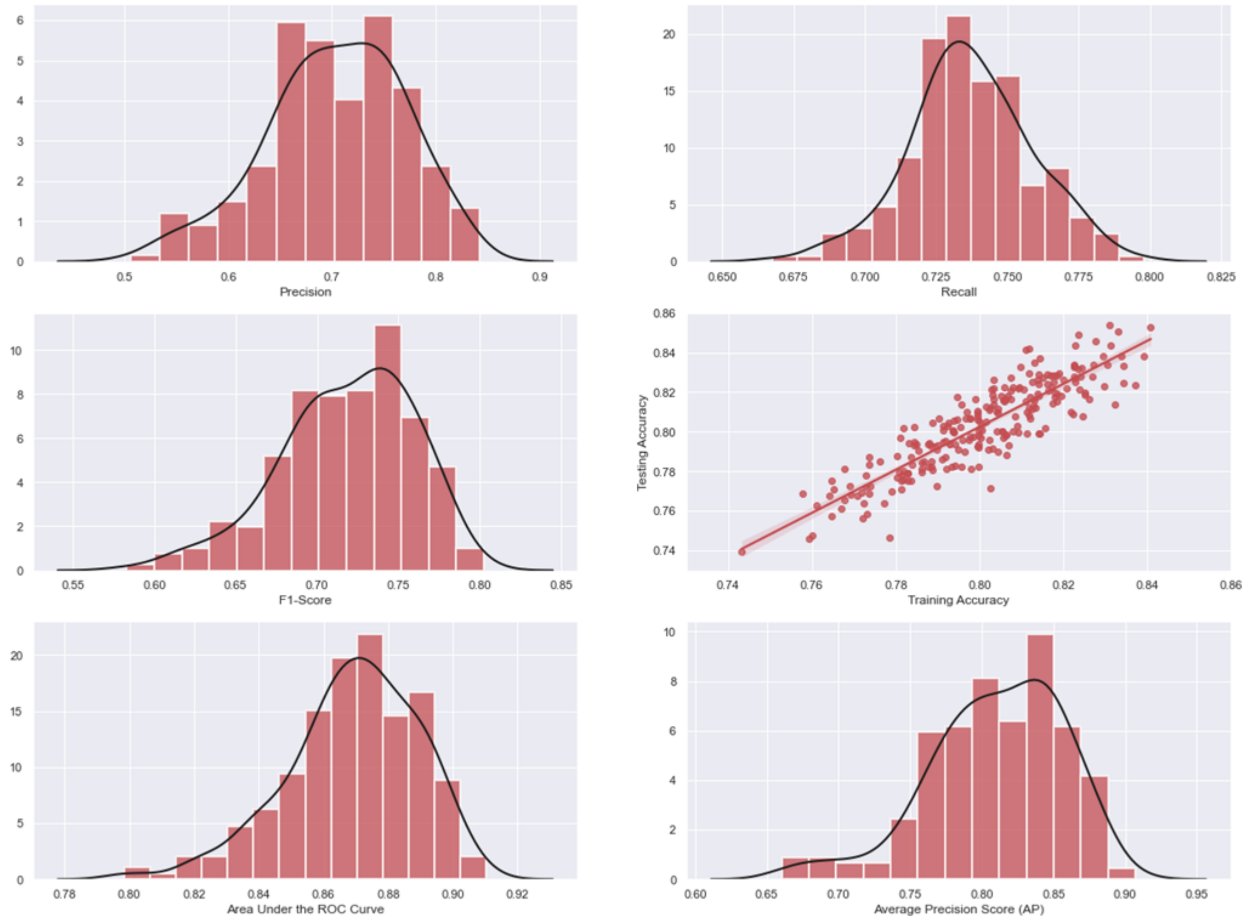
Logistic Regression (Baseline)	Average Performance Metrics
Precision (Delayed Class)	0.80
Recall	0.50
F1-Score	0.61
AUC	0.83
AP	0.76
Training Accuracy	0.79
Testing Accuracy	0.79



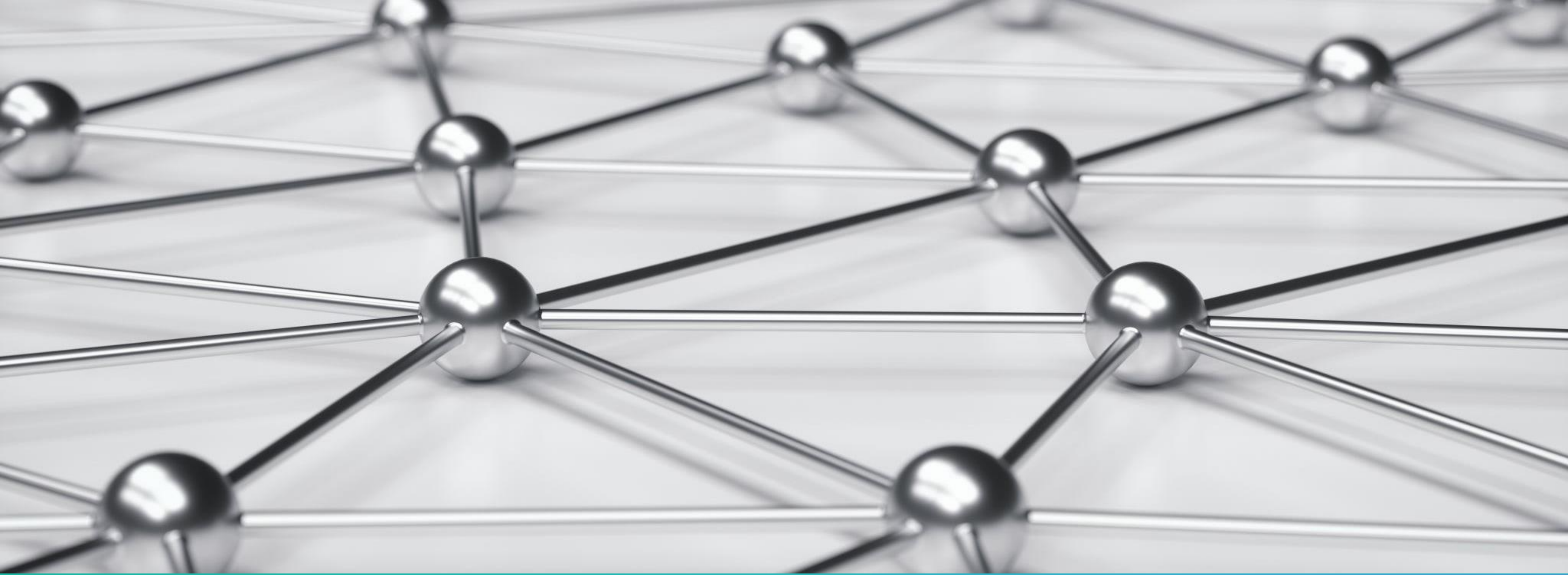
LOGISTIC REGRESSION (BASELINE)

LOGISTIC REGRESSION

Performance Metrics	Baseline	Logistic Regression
Precision (Delayed Class)	0.80	0.70
Recall	0.50	0.74
F1-Score	0.61	0.72
AUC	0.83	0.87
AP	0.76	0.81
Training Accuracy	0.79	0.80
Testing Accuracy	0.79	0.80



LOGISTIC REGRESSION



CONCLUSIONS AND FUTURE WORK

CONCLUSION AND FUTURE WORK

- Over sampling and cross validation improved prediction significantly
- SMOTE is computationally expensive
- Achieved over-fitting
- Late aircraft feature is the most important one
- One hot encoding helped in terms of precision and recall metrics
- Focus on second, third order time dependencies
- Additional feature might help in increasing model performance
- More of available data

RECOMMENDATION FOR CLIENTS

- Our model can help airline companies to pinpoint underlying causes of flight delays so that they can improve their service

- Companies and booking agencies can provide probability of a flight being delayed early to their customers at the time of booking or they can build early alert system in order to avoid wait times at the airport

- Passengers can take advantage of flight delay prediction models to schedule their flights to minimize their losses in time, business, and money

Thank you!

For your questions, please email to sahin.csci@gmail.com