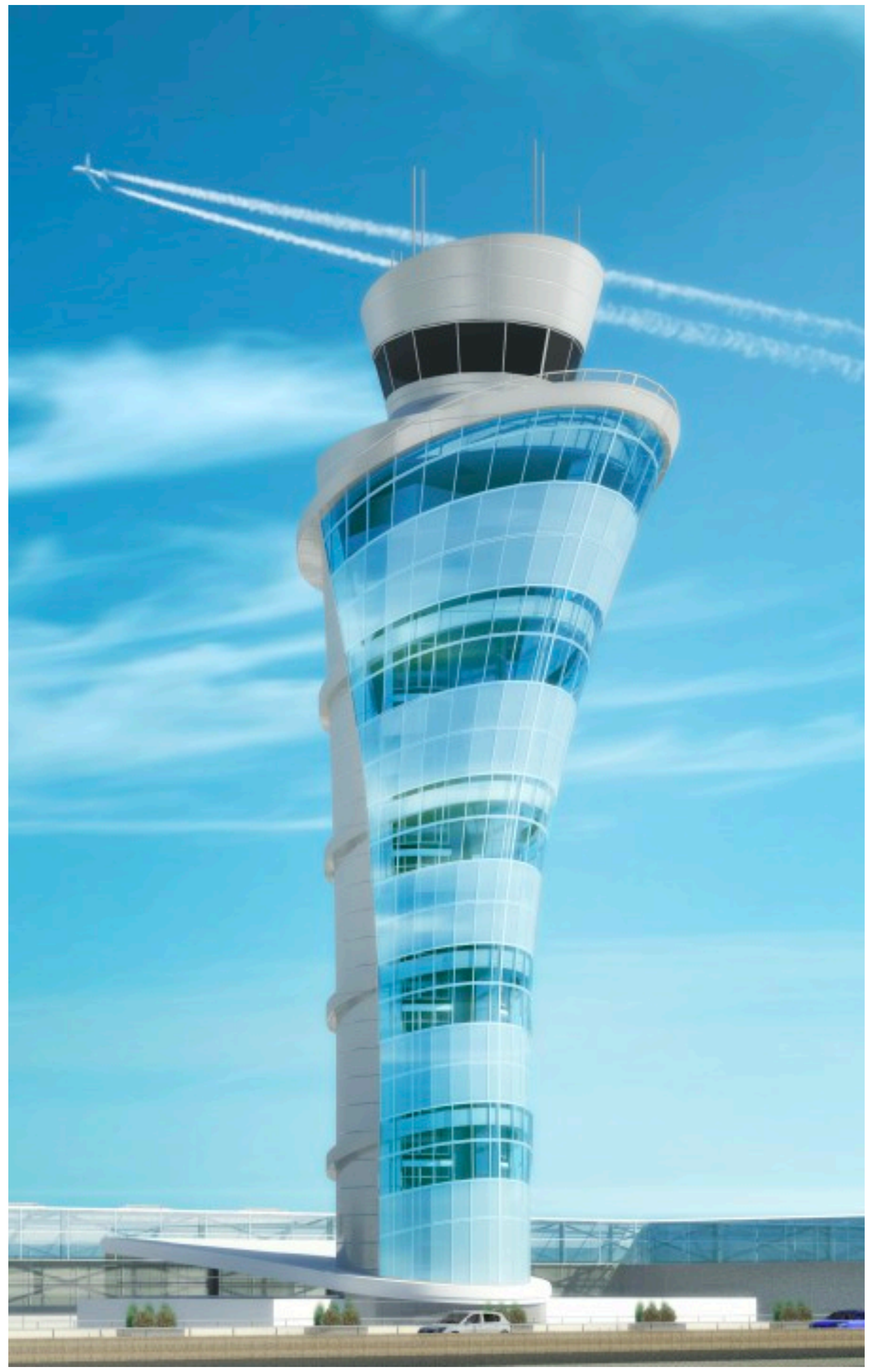
Predictors of Flight Delays

Joe Nelson // Flatiron Mod3 Final

Modeling Flight Delays

- Flight Delays are the worst; they cost businesses money and travelers time from work and family.
- Our app will analyze flight-delay data and determine which features of a given flight are most predictive of a potential delay.
- This will empower travelers to select flights that will minimize their time stuck in an airport, away from work and family.
- At this early stage, we are developing the model which will show these predictive features.



The Data Set

- To build the model, we utilized the 2015 flight data from the FAA, obtained via Kaggle: https://www.kaggle.com/usdot/flight-delays
- We encountered difficulty in building the model due to its sheer size, so we had to reduce the dataset in a few ways, including:
  - Only including data from January - June (this was the biggest cut)
  - Only including primary airports
  - Removing unnecessary columns (such as year, tail number, flight number, etc.)

The Model

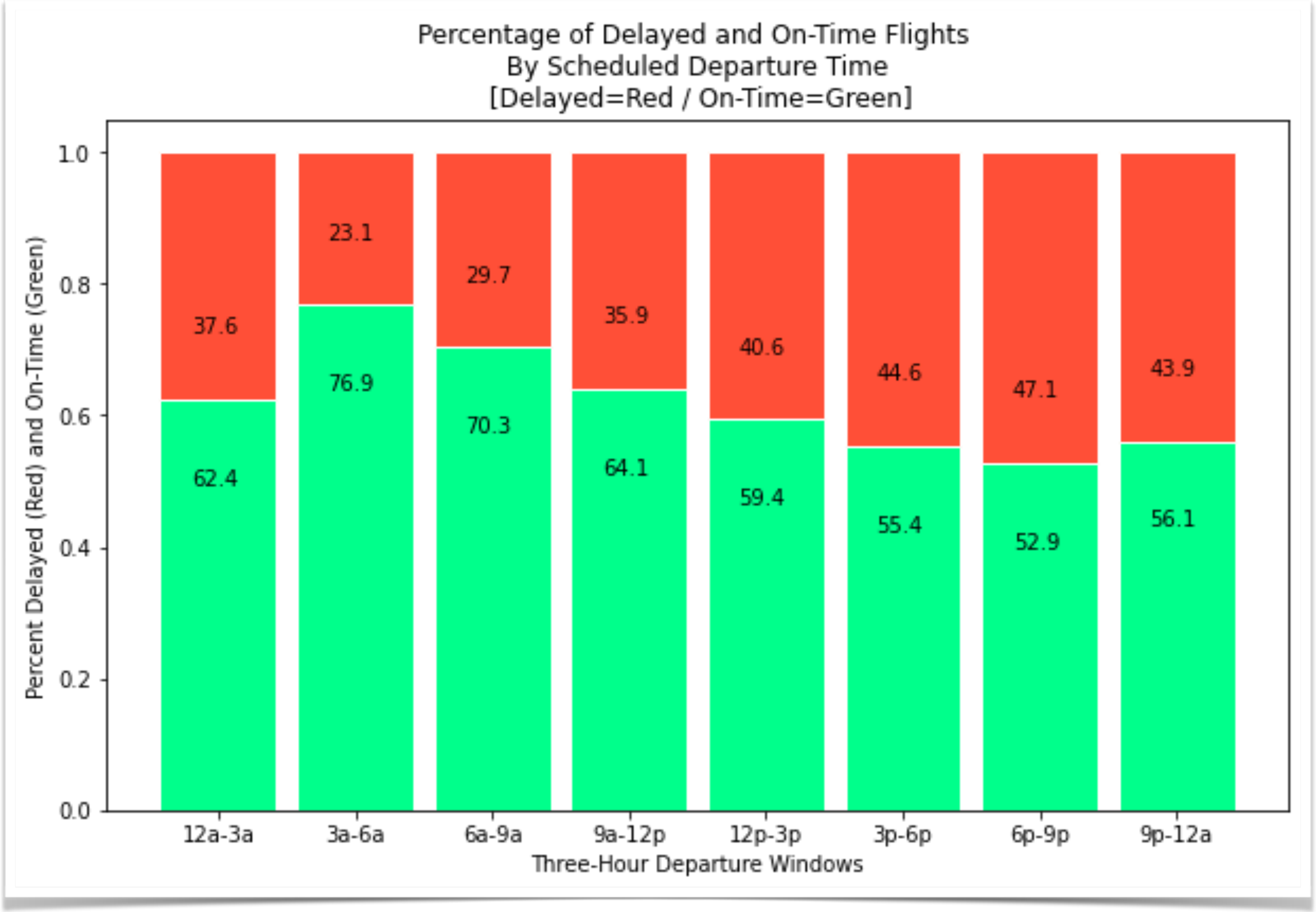
- Considering the goal of identifying predictors of flight delays, we initially used a decision-tree model.
- The delay "threshold" is user-selectable within the model, and will default to zero (i.e., any delay past the scheduled arrival time is considered "late").
- We further developed the model using a few other techniques.

The Results

- Overall model accuracy across all iterations offered only marginal efficacy in predicting a late flight (~64%, vs. 50% of plain-old guessing).
- However, the models did identify several features of flights which were important in determining whether a flight would be delayed.
- These include the departure time frame, the length of the flight, the airline, and the month of the flight (within our January-June time frame).

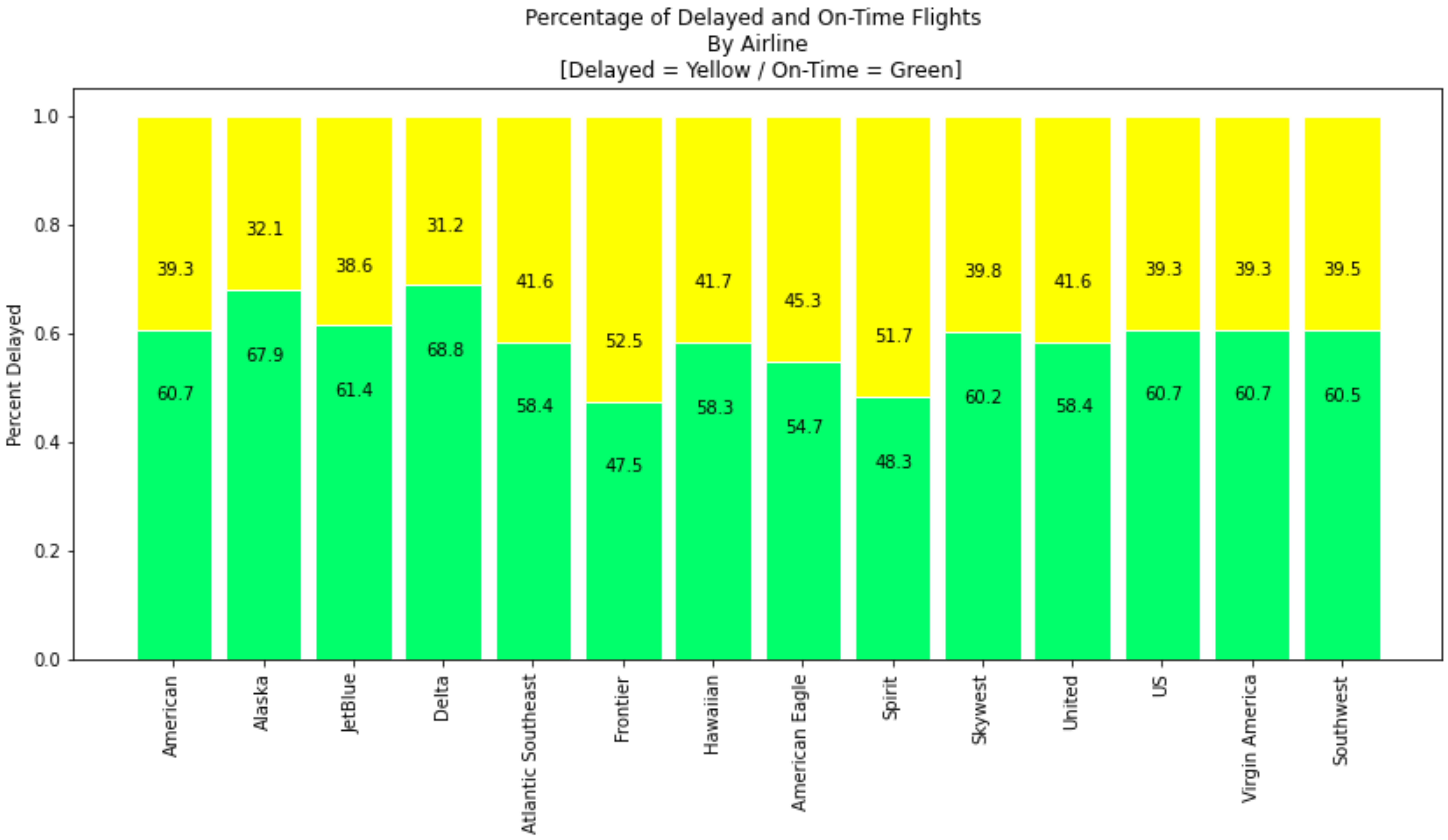
Time of Departure as Predictor of Delay

- In the initial model, departure times were the most important feature.
- Departure times were grouped into three-hour windows.
- Departures between 3am and 6am had the best on-time performance.



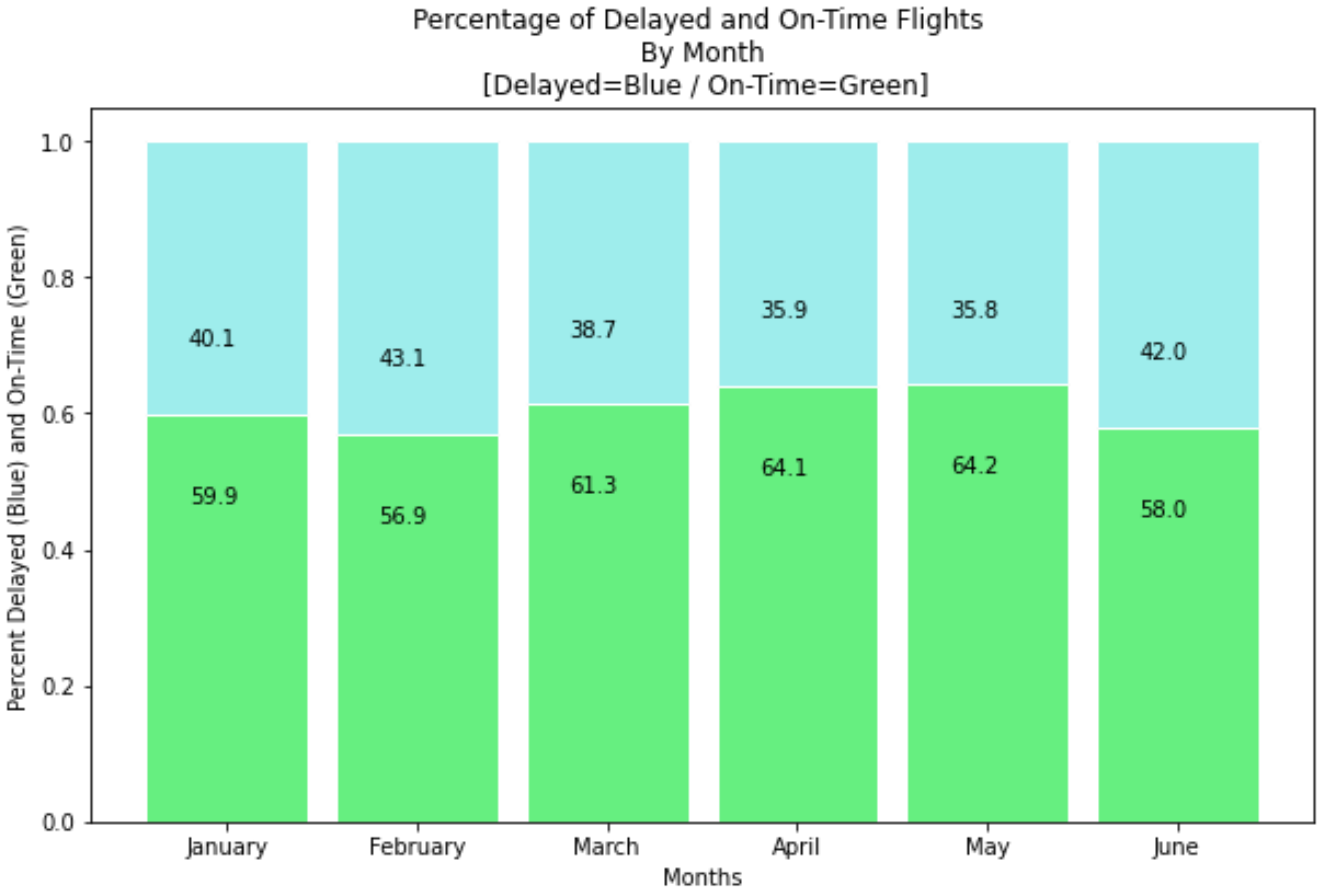
Airline as Predictor of Delay

- Airlines themselves were also an important feature in the model
- Southwest had the most flights (22%), while Virgin American had the fewest (1%).
- On-time performance of these airlines is shown to the right; Alaska and Delta did best, while Frontier and Spirit faired the worst.



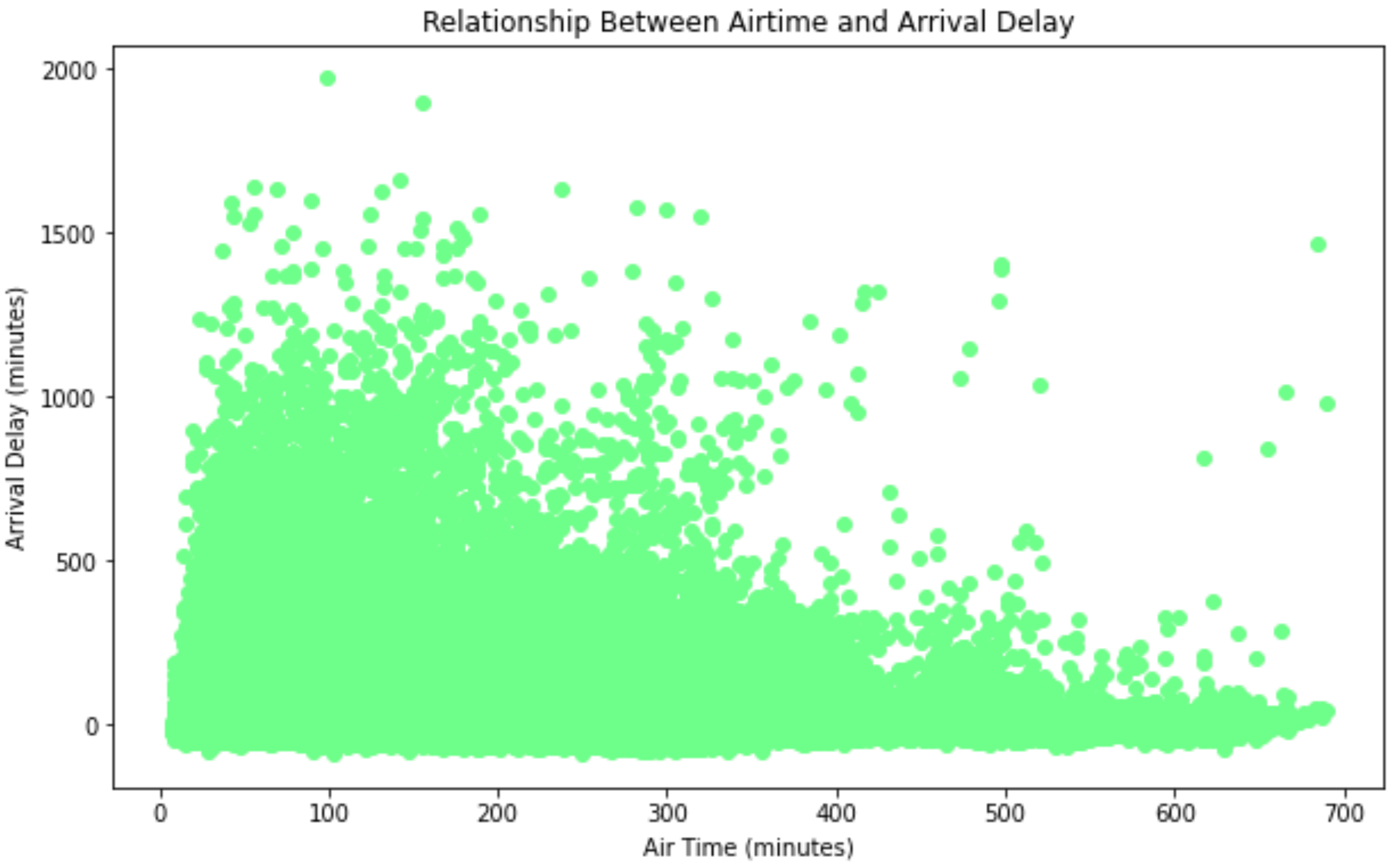
Month of Departure as Predictor of Delay

- The month of departure also played a significant role in determining on-time performance.
- This was curious, since there is not a overt difference in on-time performance over the different months, as shown on the right.
- Overall, flights in April and May were on time more than in the other months.



In-Air Time as Predictor of Delay

- As we developed the model further, the flight's time in the air also became important in predicting any delay.
- To explore this, air time is plotted against the delay in minutes (early flights are included, and sit below zero in the chart).
- The relationship shows an increase in delays with shorter flights (both in substance and in frequency).



Conclusion

- Paying attention to each of the four categories shown will help travelers avoid flight delays.
- However, this model cannot predict which flights actually *will be* delayed.
- Future development includes further exploration of the entire calendar years, and further tuning of the existing model (perhaps by further restricting airports based on the number of flights in/out, etc.).
- Any questions?

Thank You!