Predicting US flight delays and their causes

aviation industry use case project

Sebastian T. Gomez

19.5.2022

The problem

- Flight delays are expensive for airports and airlines
- Increasing traffic volumes makes schedules more sensitive to delays and disruptions
- Delays increase environmental impact
- No one likes to hang out in parked planes and airport gates

The data

Monthly flight departures, delays, and cancellations per carrier and airport in the US during 2003-2020

30 airports, 28 carriers, 73k data points (monthly figures per carrier per airport)

https://www.transtats.bts.gov/OT_Delay/Homepage.asp

The data

Year, month, carrier, airport

number of delayed flights and breakdown by cause

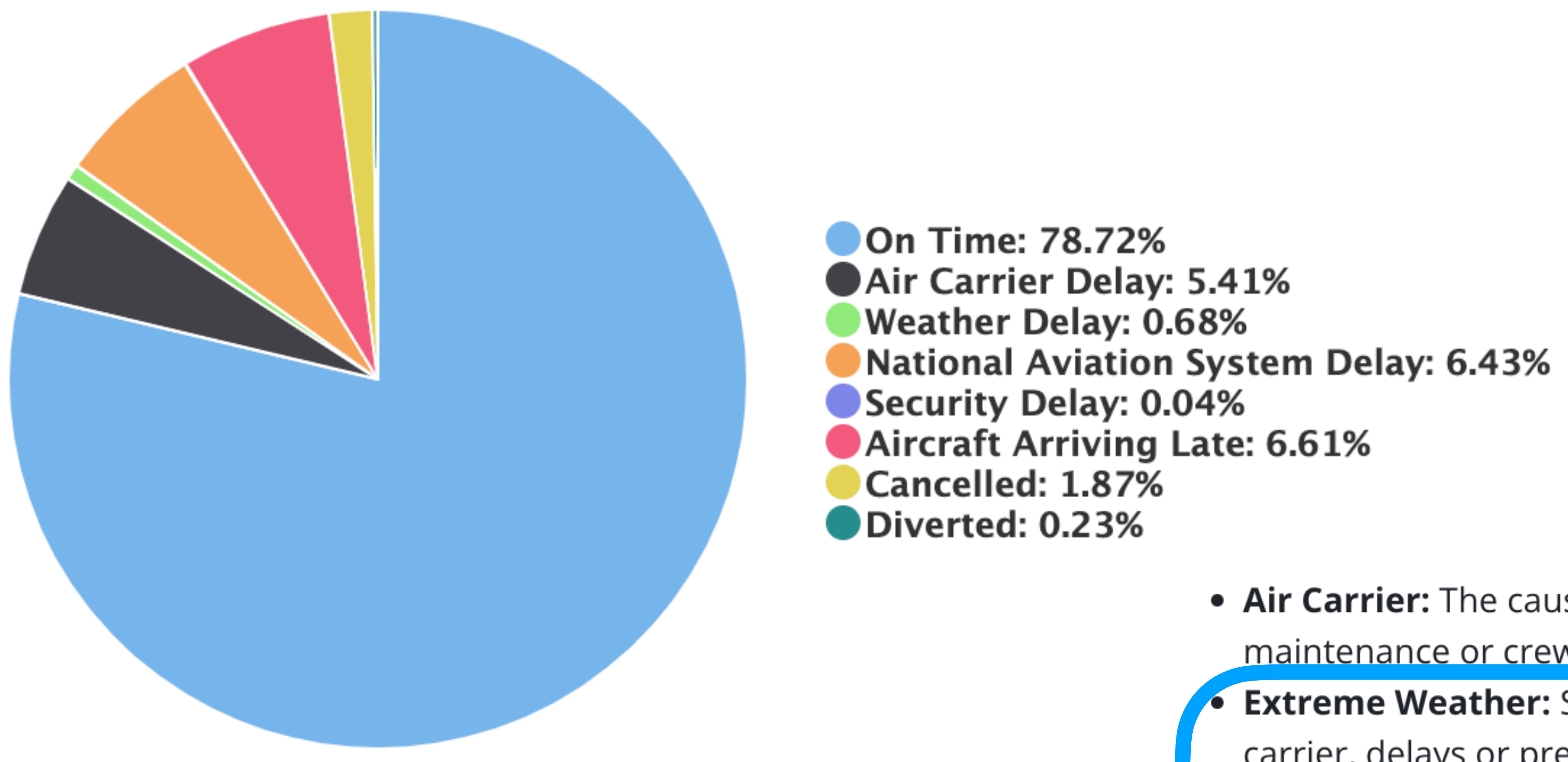
delay duration and breakdown by cause

Column Name	Description
date	Year and month, in the format YYYY-M (e.g., 2018-1)
carrier	The two character designator for the carrier/airline.
carrier_name	The full name of the carrier/airline.
airport	The three character designator for the arrival airport.
airport_name	The full name of the arrival airport.
arr_flights	The total number of arriving flights for the carrier-airport pair for the month specified.
arr_del15	The number of arriving flights that were delayed. Delayed is when a flight arrives more than 15 minutes later than the scheduled arrival time.
carrier_ct	The number of arriving flights delayed due to a carrier issue.
weather_ct	The number of arriving flights delayed due to a weather issue.
nas_ct	The number of arriving flights delayed due to a national air system issue.
security_ct	The number of arriving flights delayed due to a security issue.
late_aircraft_ct	The number of arriving flights delayed due to an earlier late arrival of an aircraft.
arr_cancelled	The number of cancelled flights.
arr_diverted	The number of diverted flights.
arr_delay	The total number of delayed minutes due to delays.
carrier_delay	The total number of delayed minutes due to carrier issues.
weather_delay	The total number of delayed minutes due to weather issues.
nas_delay	The total number of delayed minutes due to national air system issues.
security_delay	The total number of delayed minutes due to security issues.
late_aircraft_delay	The total number of delayed minutes due to earlier later arrival of aircraft.

source: https://towardsdatascience.com/learn-python-data-analytics-by-example-airline-arrival-delays-e26356e8ae6b

The data

Delay causes



- **Air Carrier:** The cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
- **Extreme Weather:** Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane.
- **National Aviation System (NAS):** Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
- **Late-arriving aircraft:** A previous flight with same aircraft arrived late, causing the present flight to depart late.
- **Security:** Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.

The questions

1. Is the delay probability and its duration predictable from time, airport, and carrier information?
2. Are the leading causes of delays predictable?
3. What are the main factors that help predict flight delays?

Data exploration

… but first, some boring details

- **clean the data: removing the few missing values by dropping rows**

```
RangeIndex: 73282 entries, 0 to 73281
Data columns (total 21 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   year                 73282 non-null  int64
 1   month                73282 non-null  int64
 2   carrier              73282 non-null  object
 3   carrier_name         73282 non-null  object
 4   airport              73282 non-null  object
 5   airport_name         73282 non-null  object
 6   arr_flights          73240 non-null  float64
 7   arr_del15            73211 non-null  float64
 8   carrier_ct           73240 non-null  float64
 9   weather_ct           73240 non-null  float64
 10  nas_ct               73240 non-null  float64
 11  security_ct          73240 non-null  float64
 12  late_aircraft_ct     73240 non-null  float64
 13  arr_cancelled        73240 non-null  float64
 14  arr_diverted         73240 non-null  float64
 15  arr_delay            73240 non-null  float64
 16  carrier_delay        73240 non-null  float64
 17  weather_delay        73240 non-null  float64
 18  nas_delay            73240 non-null  float64
 19  security_delay       73240 non-null  float64
 20  late_aircraft_delay  73240 non-null  float64
dtypes: float64(15), int64(2), object(4)
memory usage: 11.7+ MB
```

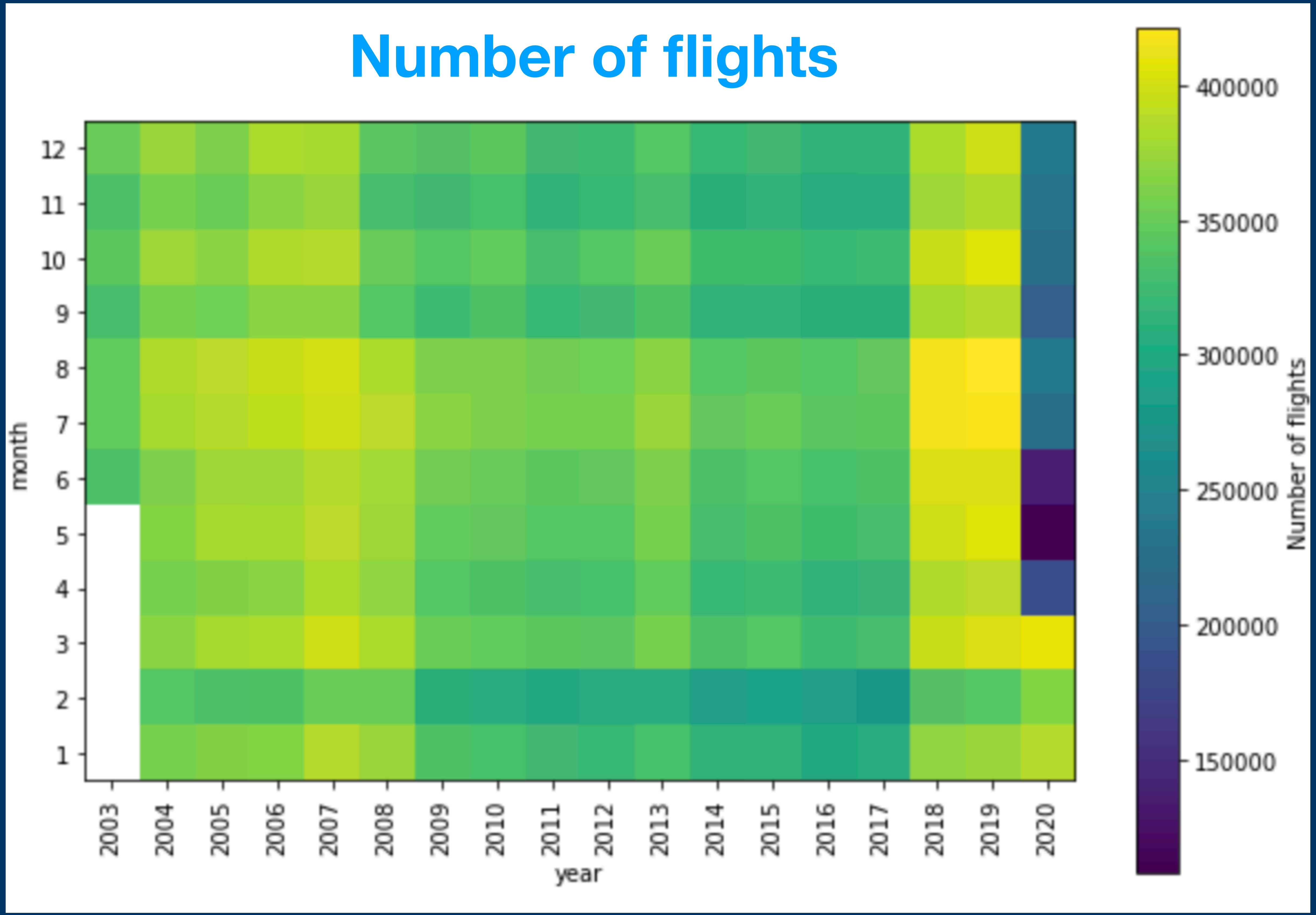
Data exploration

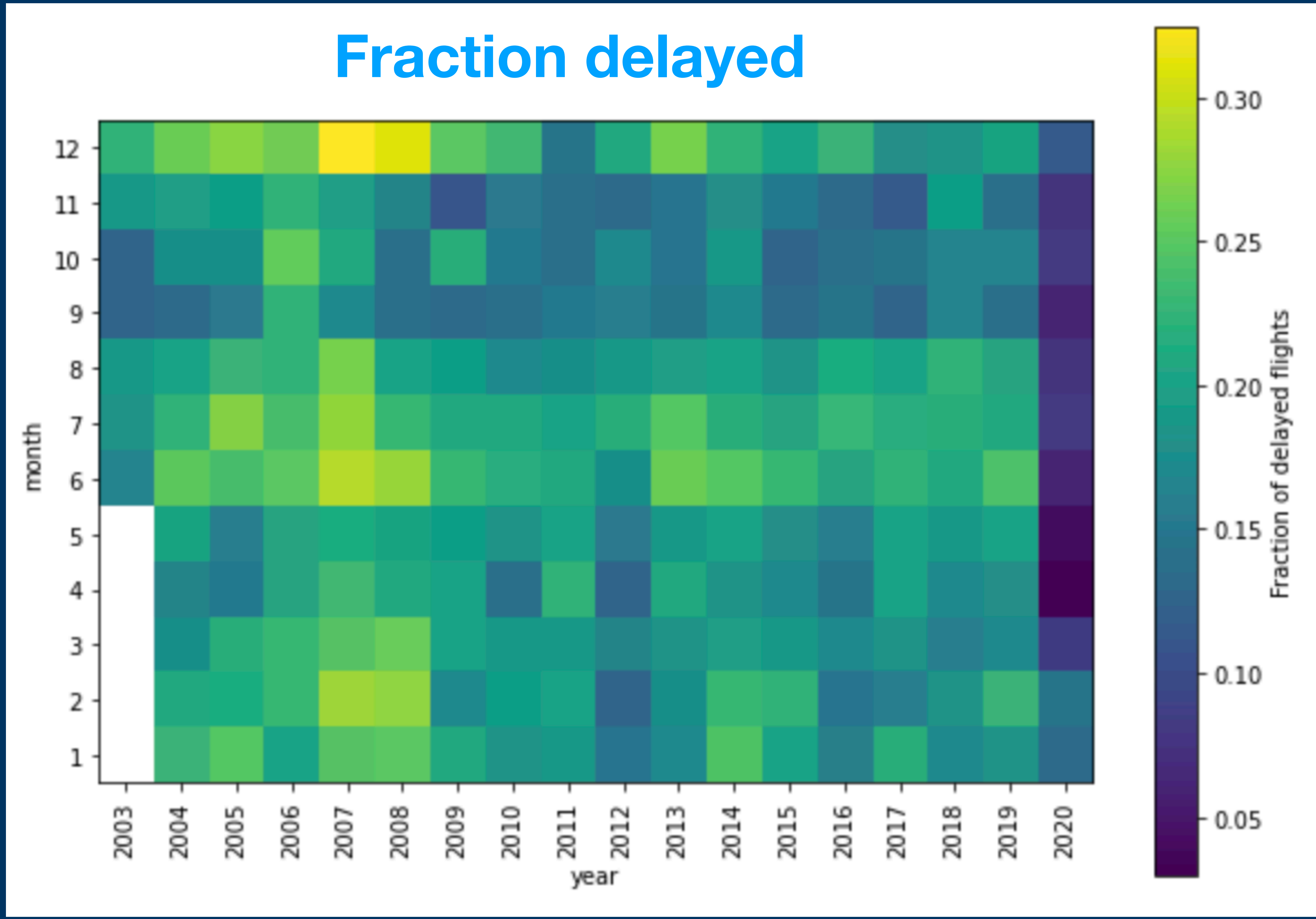
Look for general trends:

- **Collapse data along carrier and airport dimensions to look for seasonal and yearly trends**
- **Collapse all delay data into three summary variables:**
    1. **total flights per month and year**
    2. **fraction delayed**
    3. **main delay cause**

Insights from the data

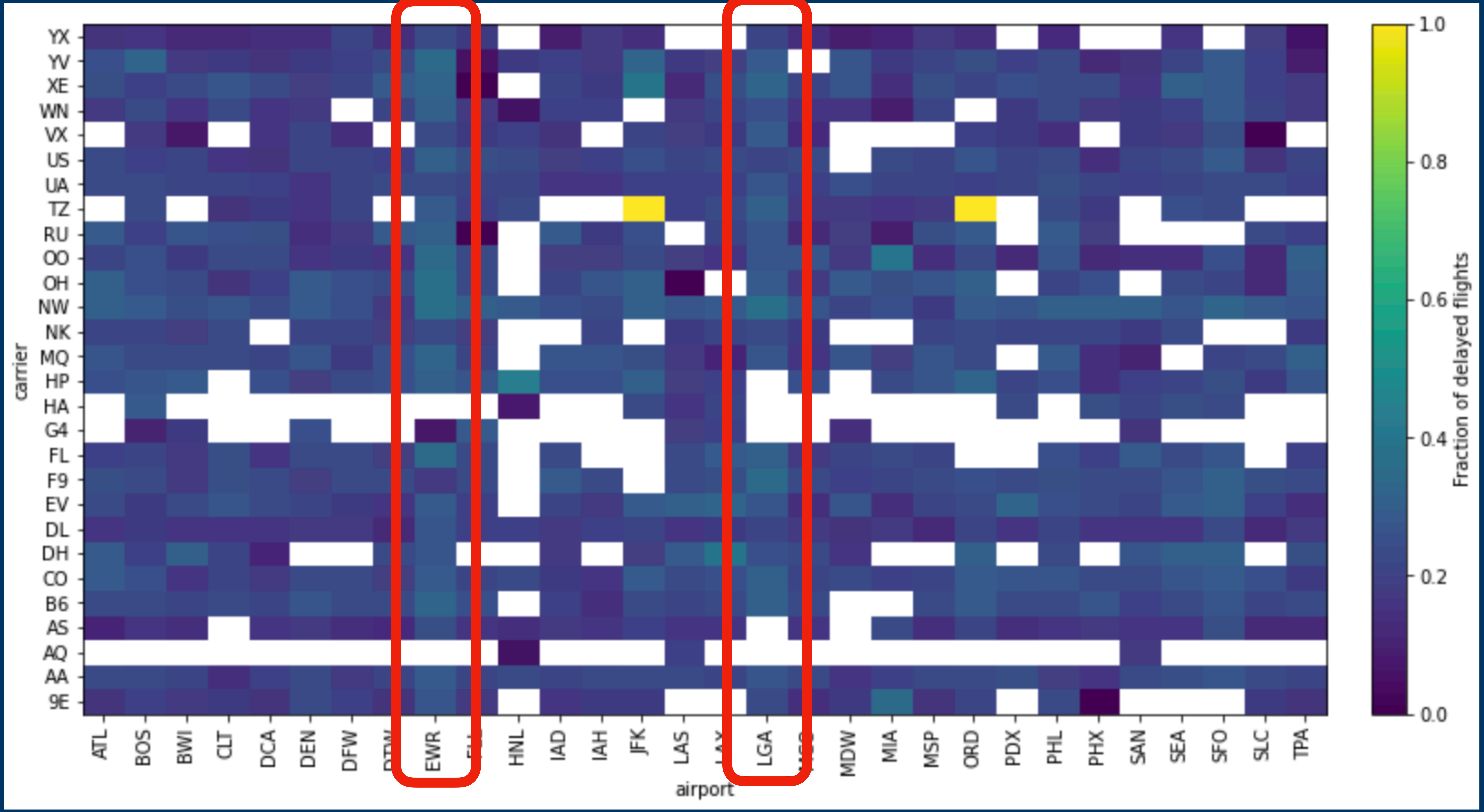
Seasonal and yearly variation





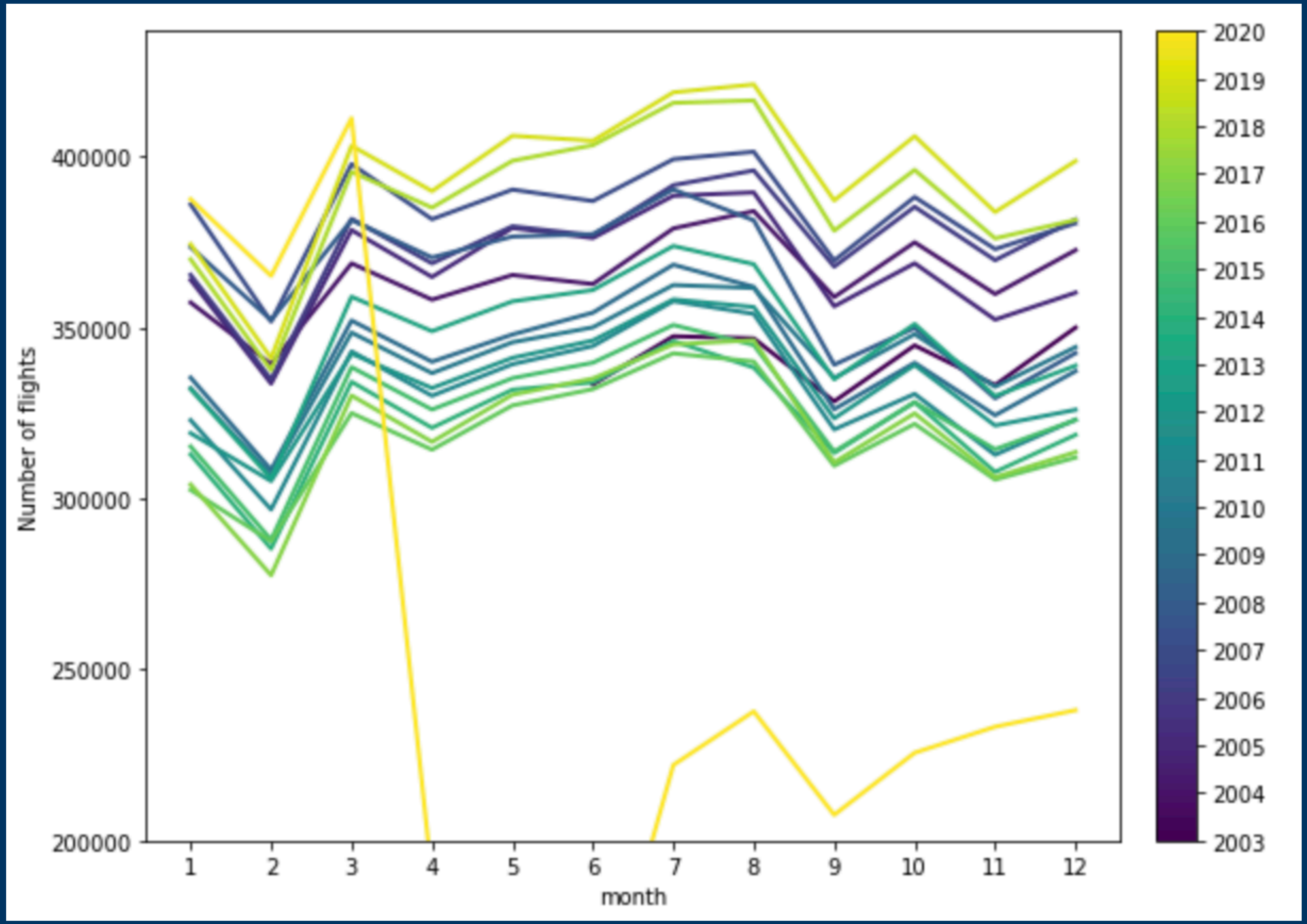
Insights from the data

Carrier and airport trends



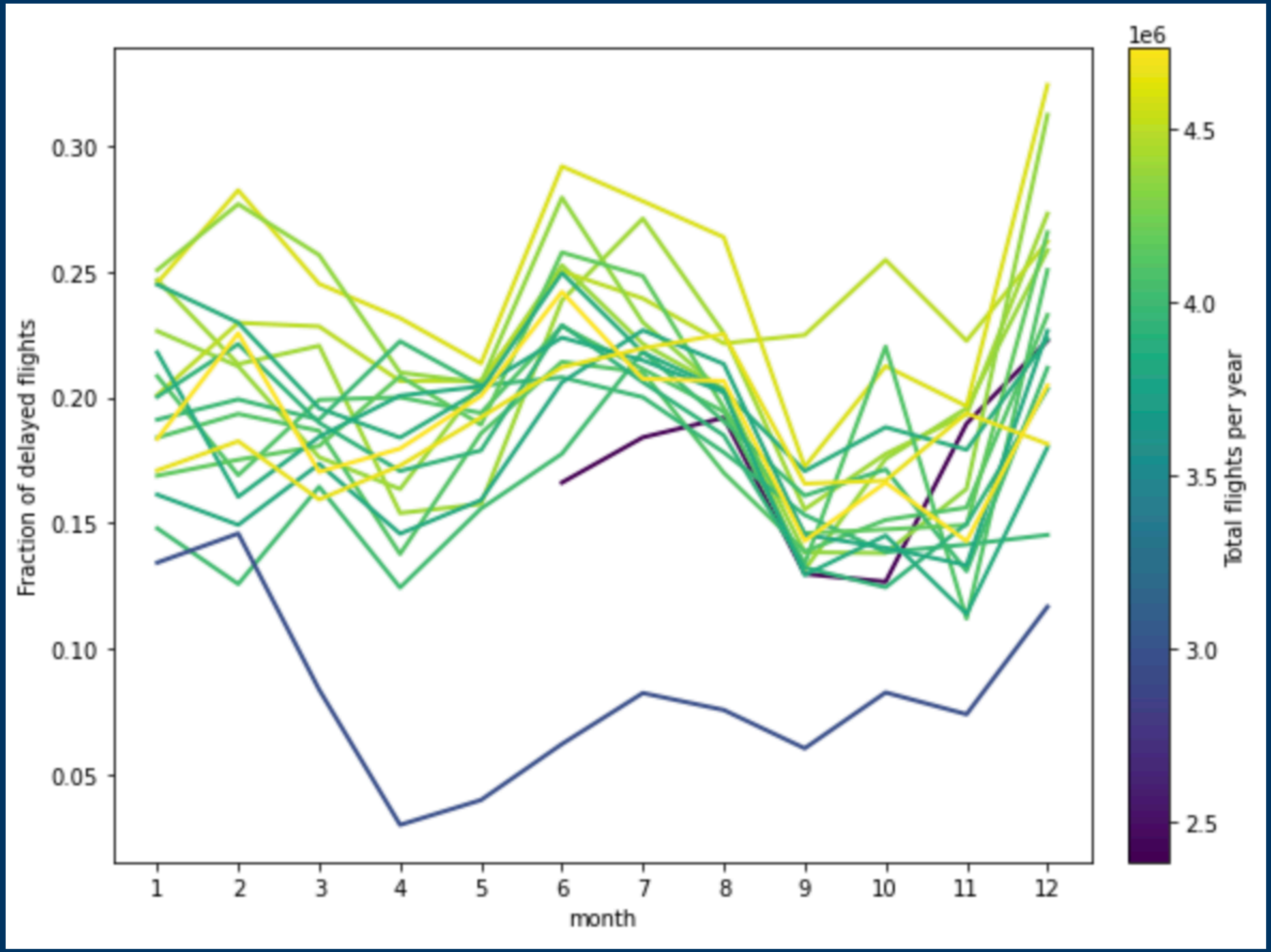
Zoom-in on seasonal trends

- **Seasonal trend strong across all years**
- **2020 flows trends until March and becomes outlier due to pandemic**



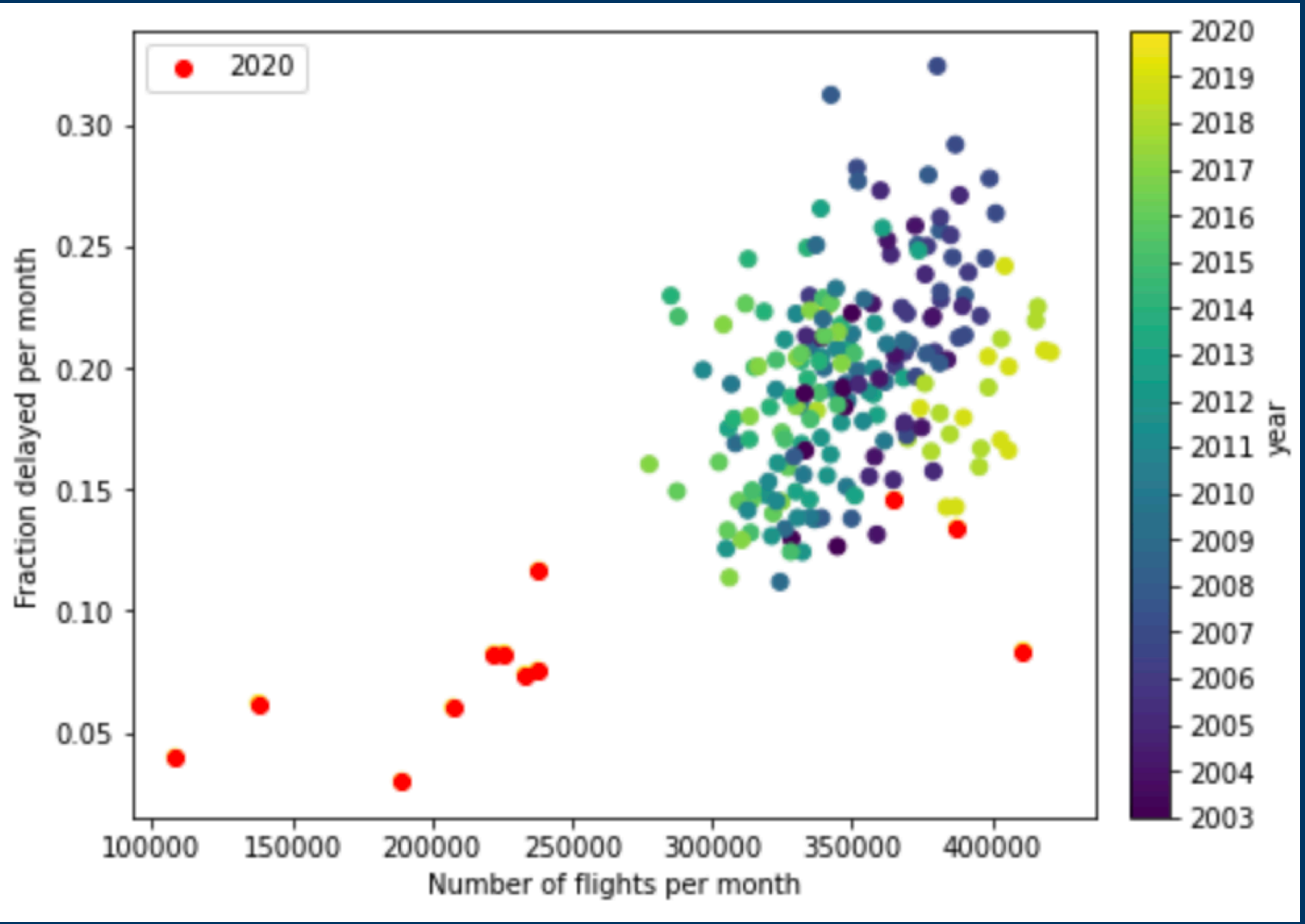
Zoom-in on seasonal trends

- **Delayed fraction increases with yearly flight volume**



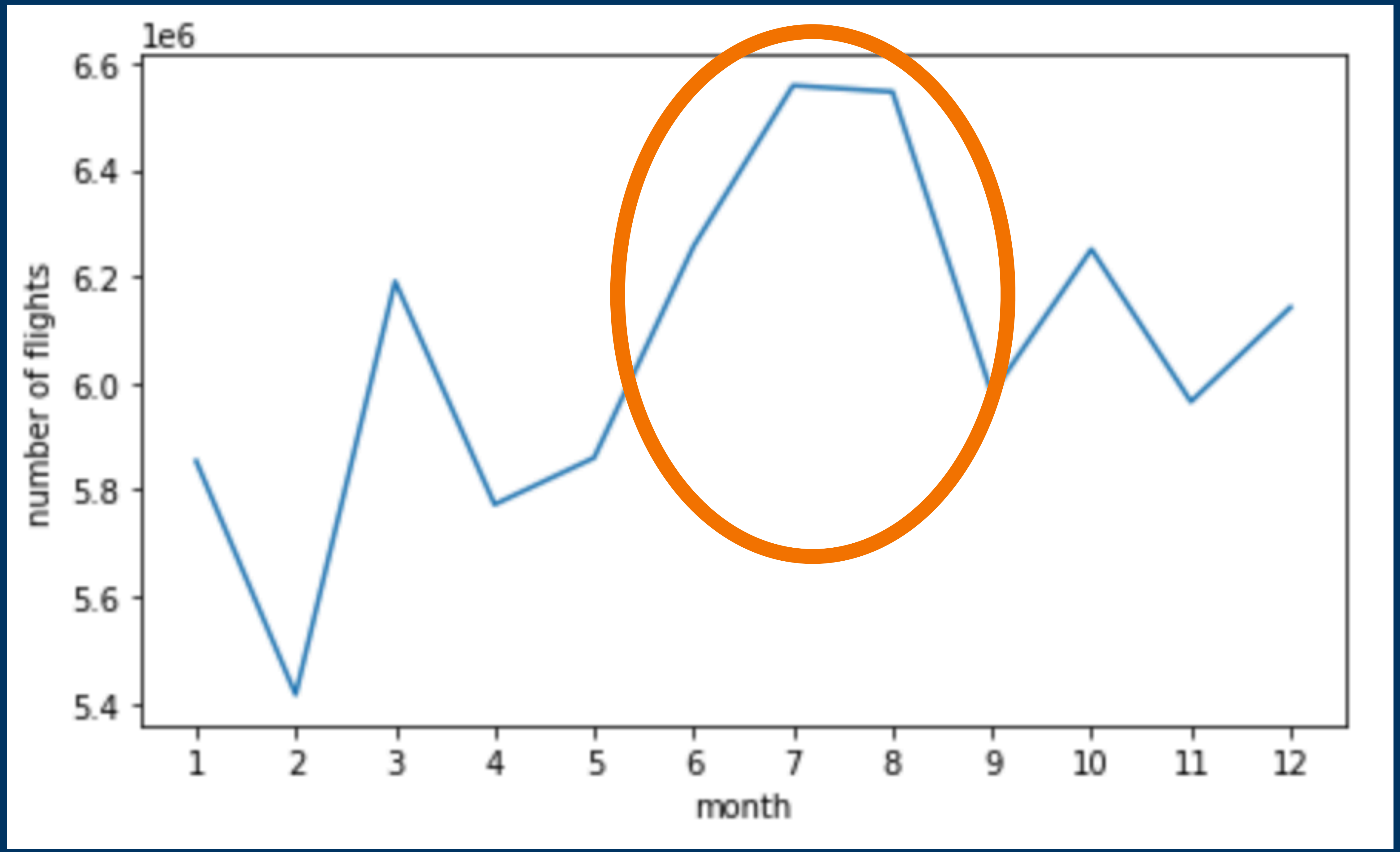
Zoom-in on seasonal trends

- Seasonal and yearly correlation between delayed fraction and volume
- Pre-pandemic 2020 had fewest delays despite high volume

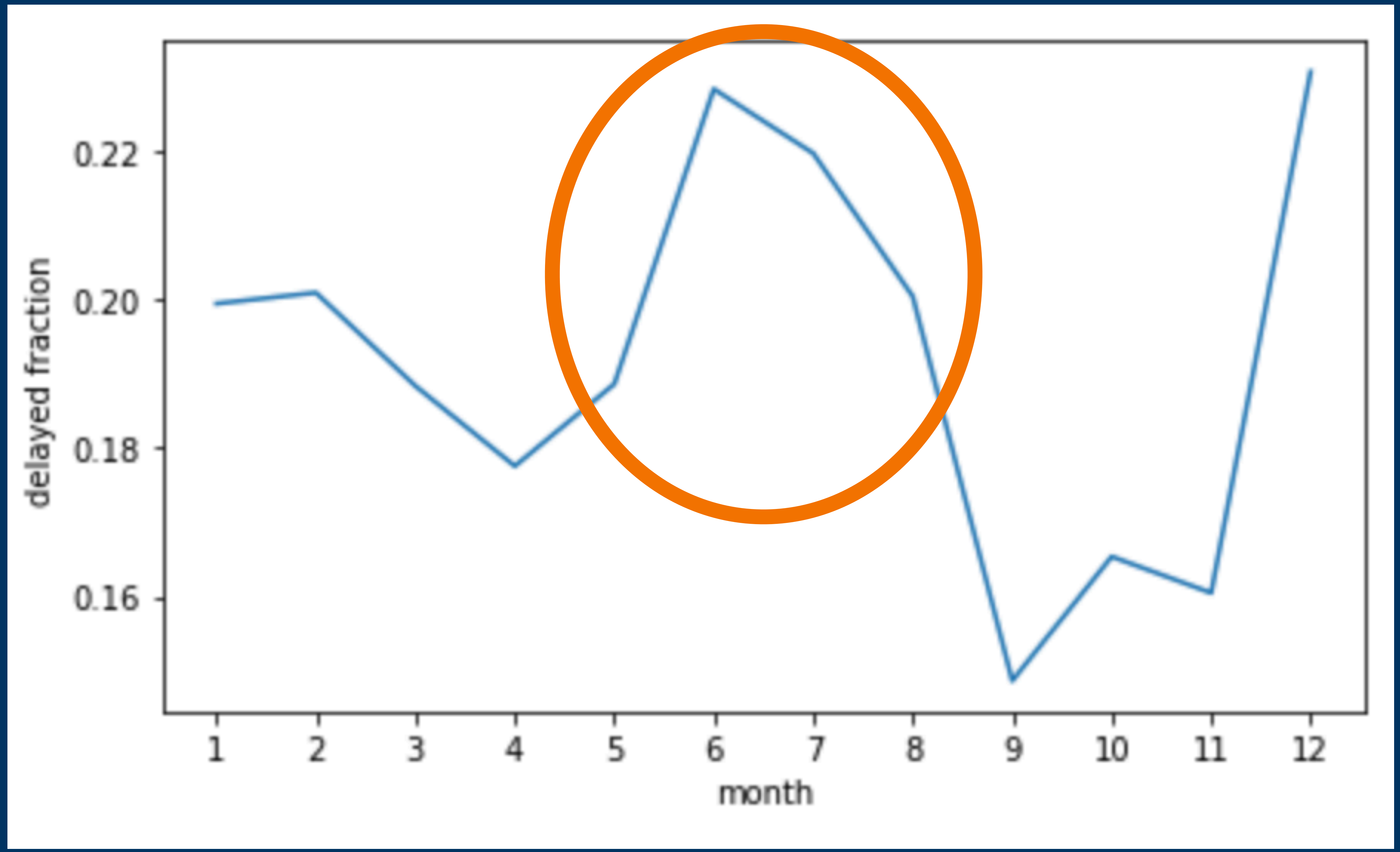


Zoom-in on seasonal trends

Number of flights

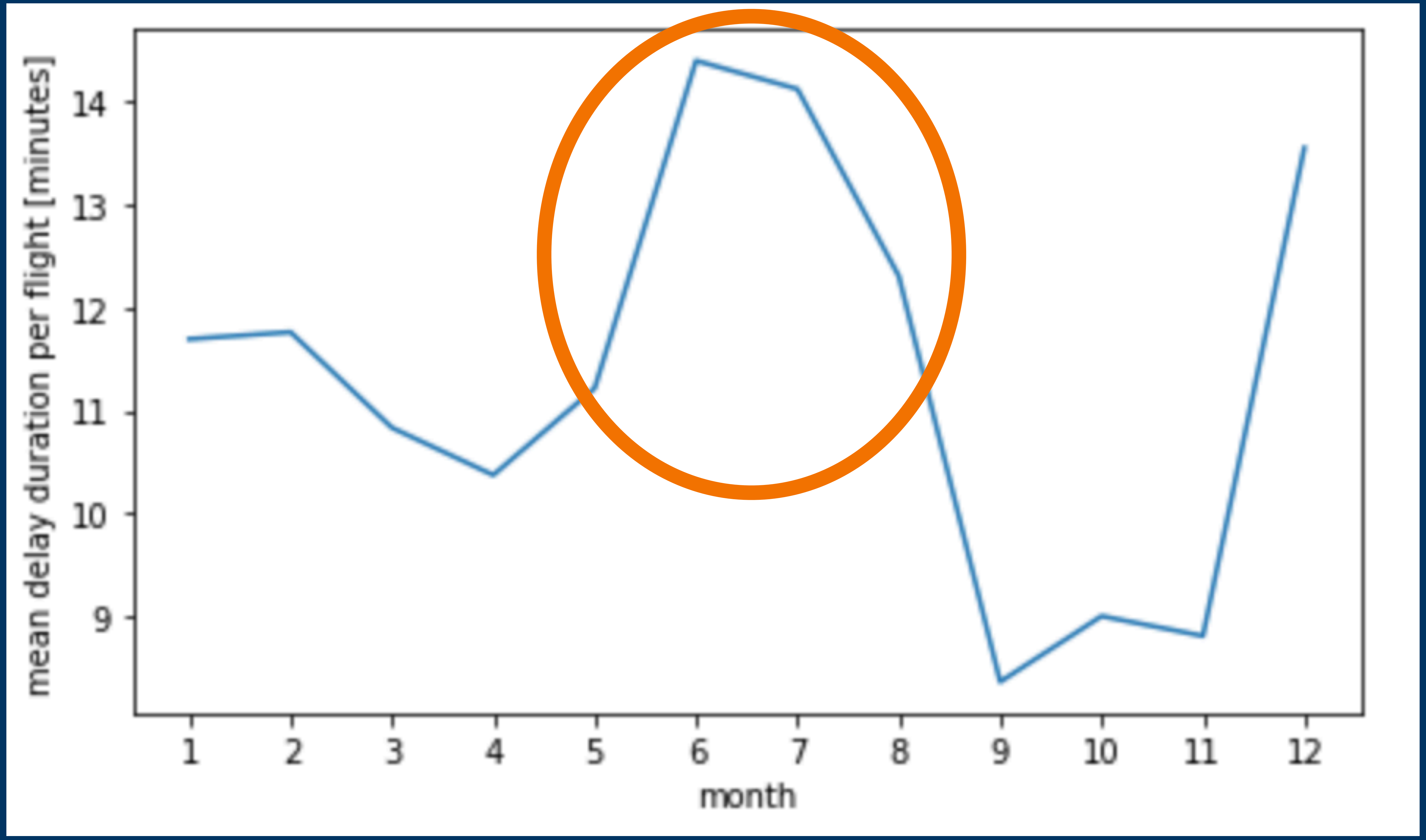


Delayed fraction

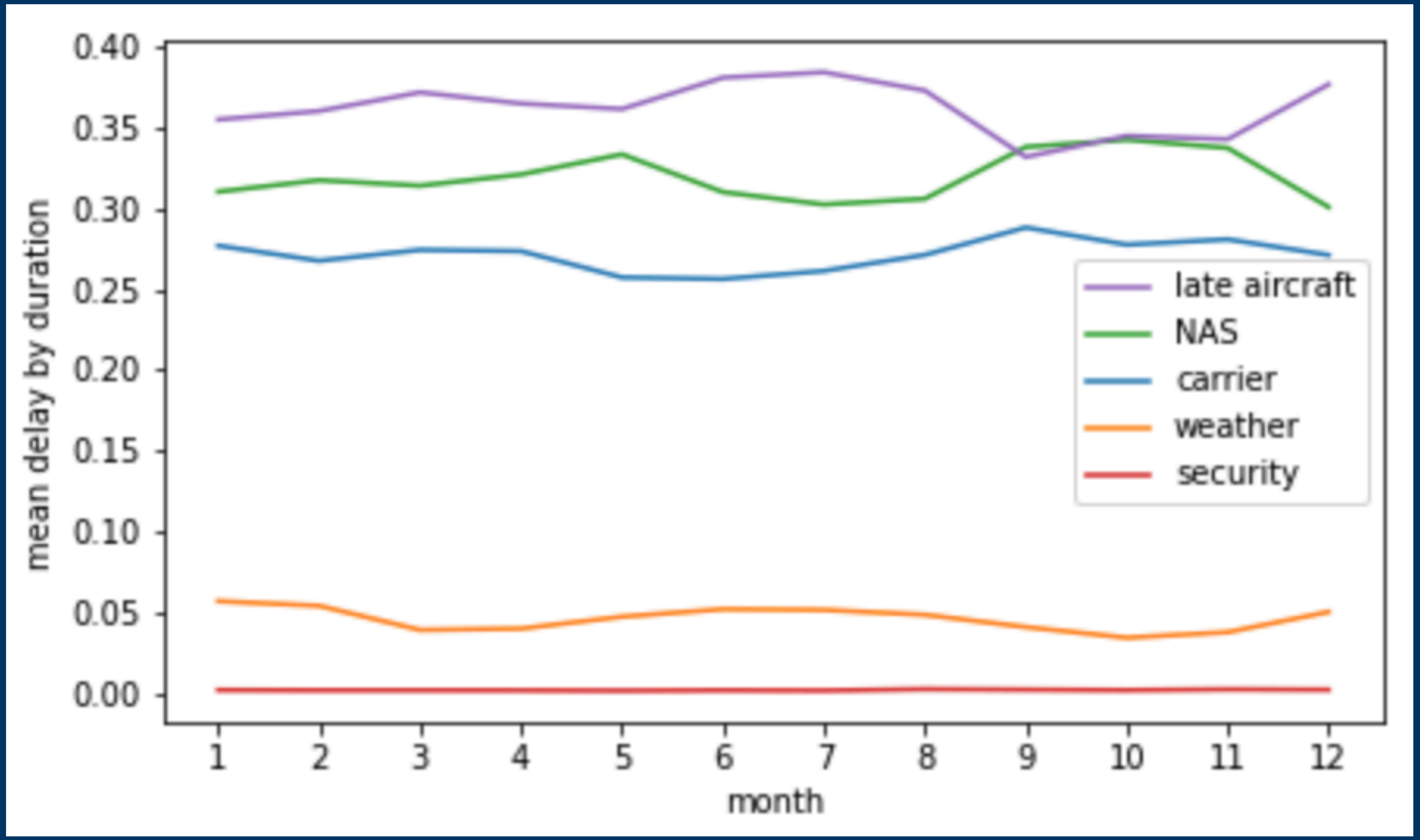


Zoom-in on seasonal trends

Mean delay duration



Mean delay cause



Use ML to predict delay statistics

1. Assume data is *representative of all flights* in the US
2. Select algorithm
   - Random Forest (Breiman 2001):
     an ensemble of decision trees optimized to find the best rules for predicting values or categories
   - good for both regression and classification
   - *efficient, accurate, interpretable & good out of the box*

Use ML to predict delay statistics

3. Prepare data
   - Design feature variables:
     - year, month, airport, carrier, *flight volume*
   - Design target variables:
     - delay *probability* = delayed/total flights
     - delay *duration* [min.]
     - delay *cause* (carrier, weather, NAS, late aircraft, security)
4. Encode categorical variable (C) using integers
5. Set aside random 20% of data for testing performance

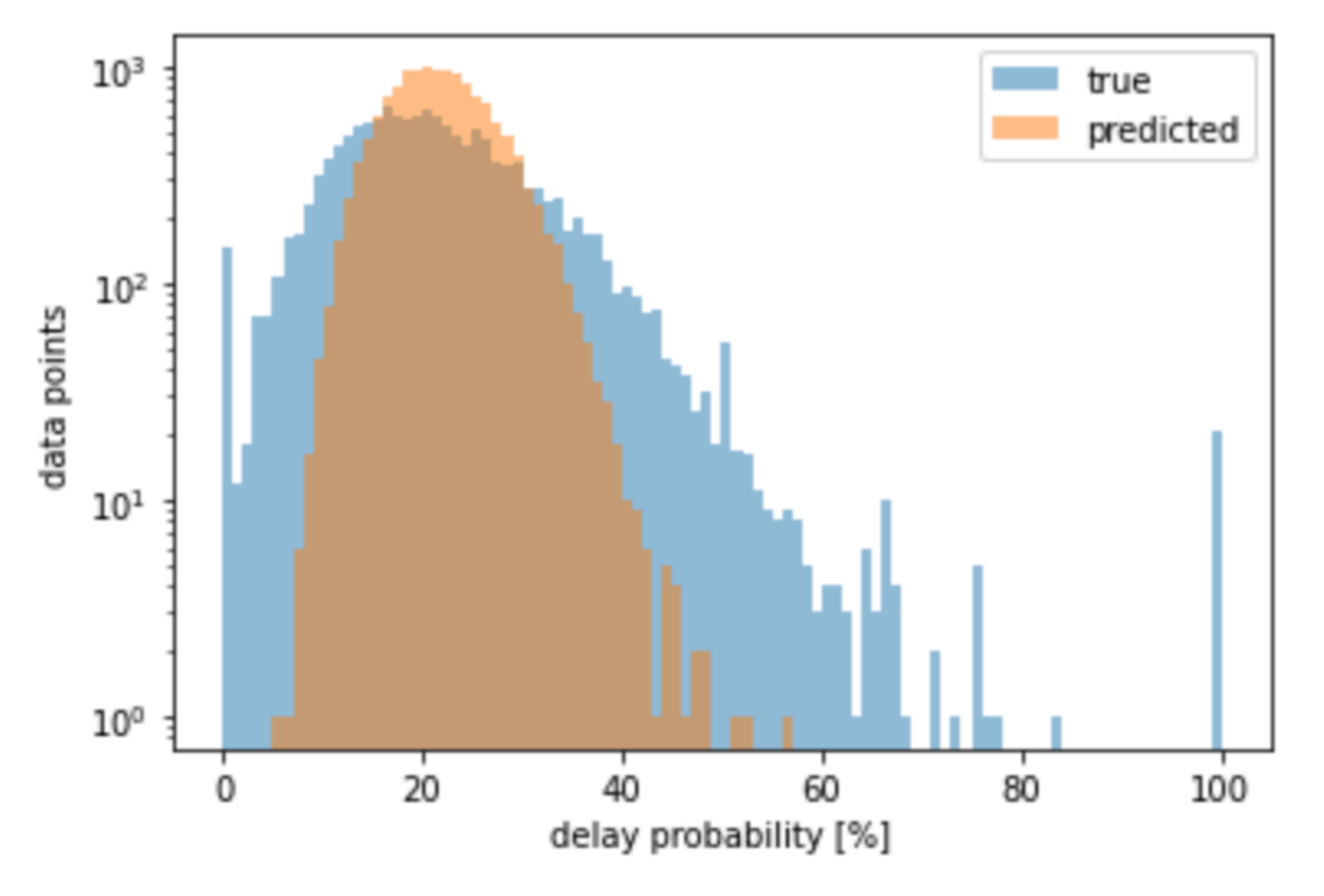
Use ML to predict delay statistics

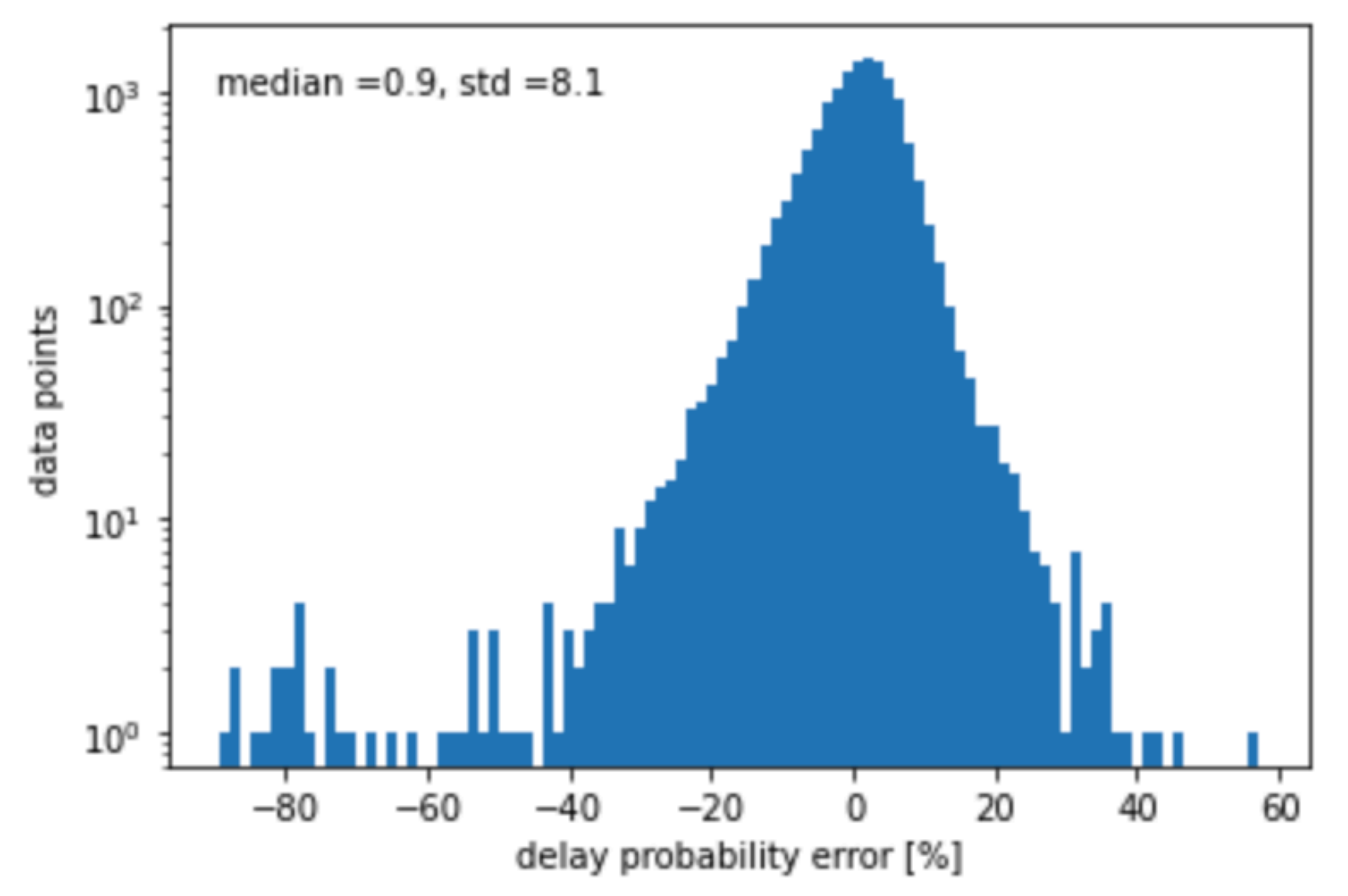
6. Select performance metric
   - Delay *probability* and *duration* (regression): R^2
   - Delay *cause* (classification): balanced accuracy
7. Tune model hyperparameters using grid search cross-validation
8. Python libraries: `numpy`, `matplotlib`, `pandas`, `scikit-learn`

Delay probability prediction

Delay probabilities:

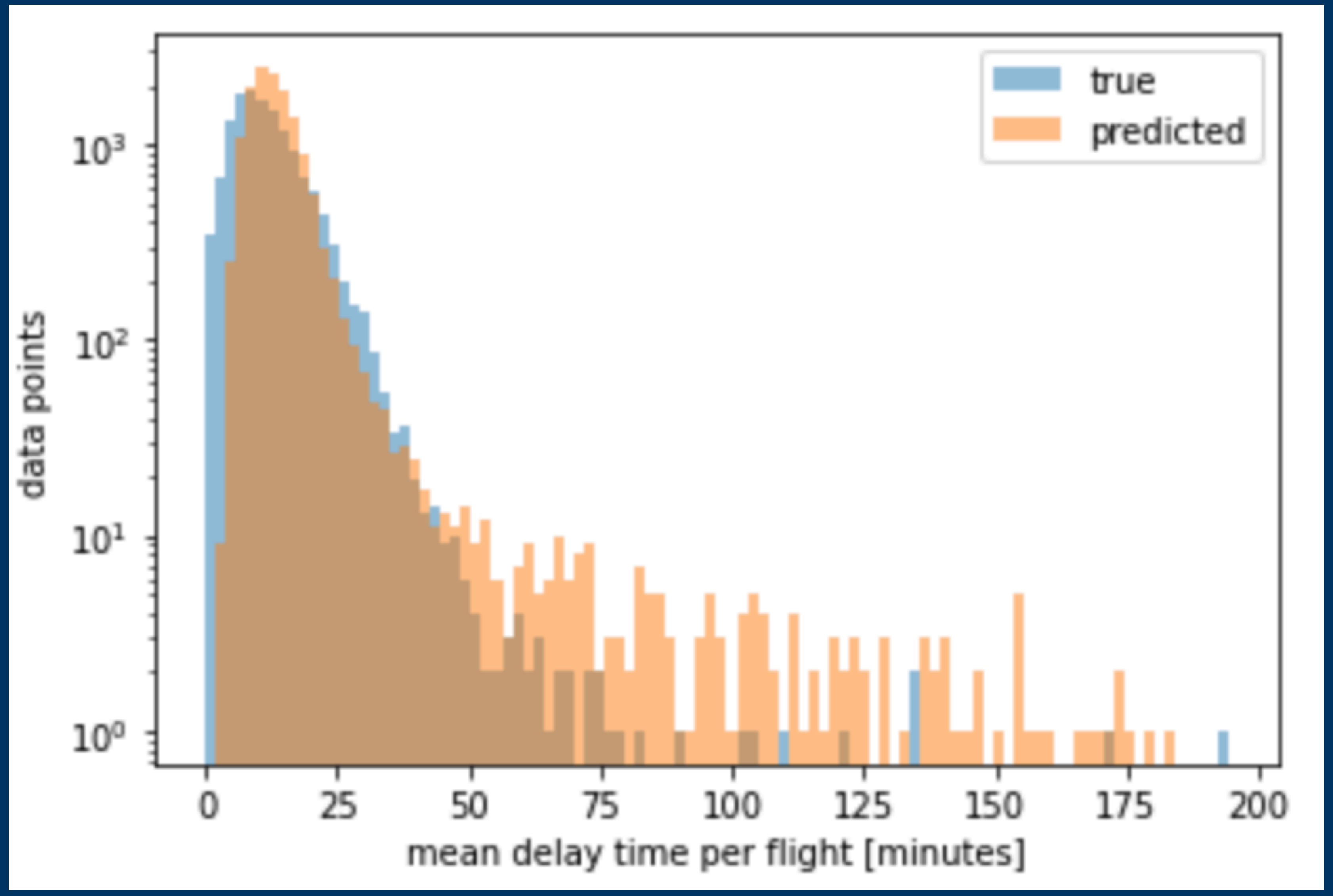
- $R^2 = 0.42$ (model captures about half of the variance in the data
- ~10% systematic bias
- ~8% precision

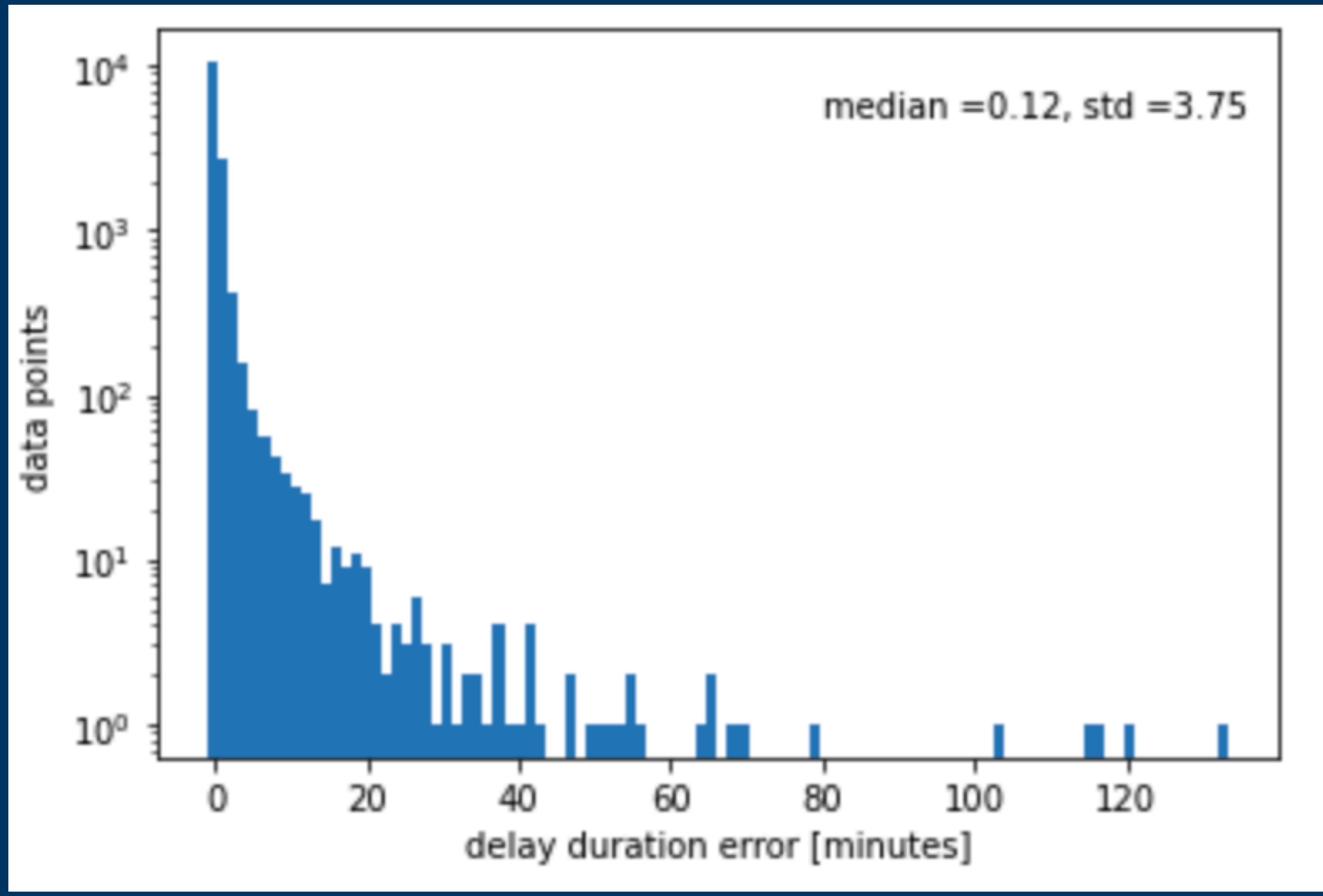




Delay duration prediction

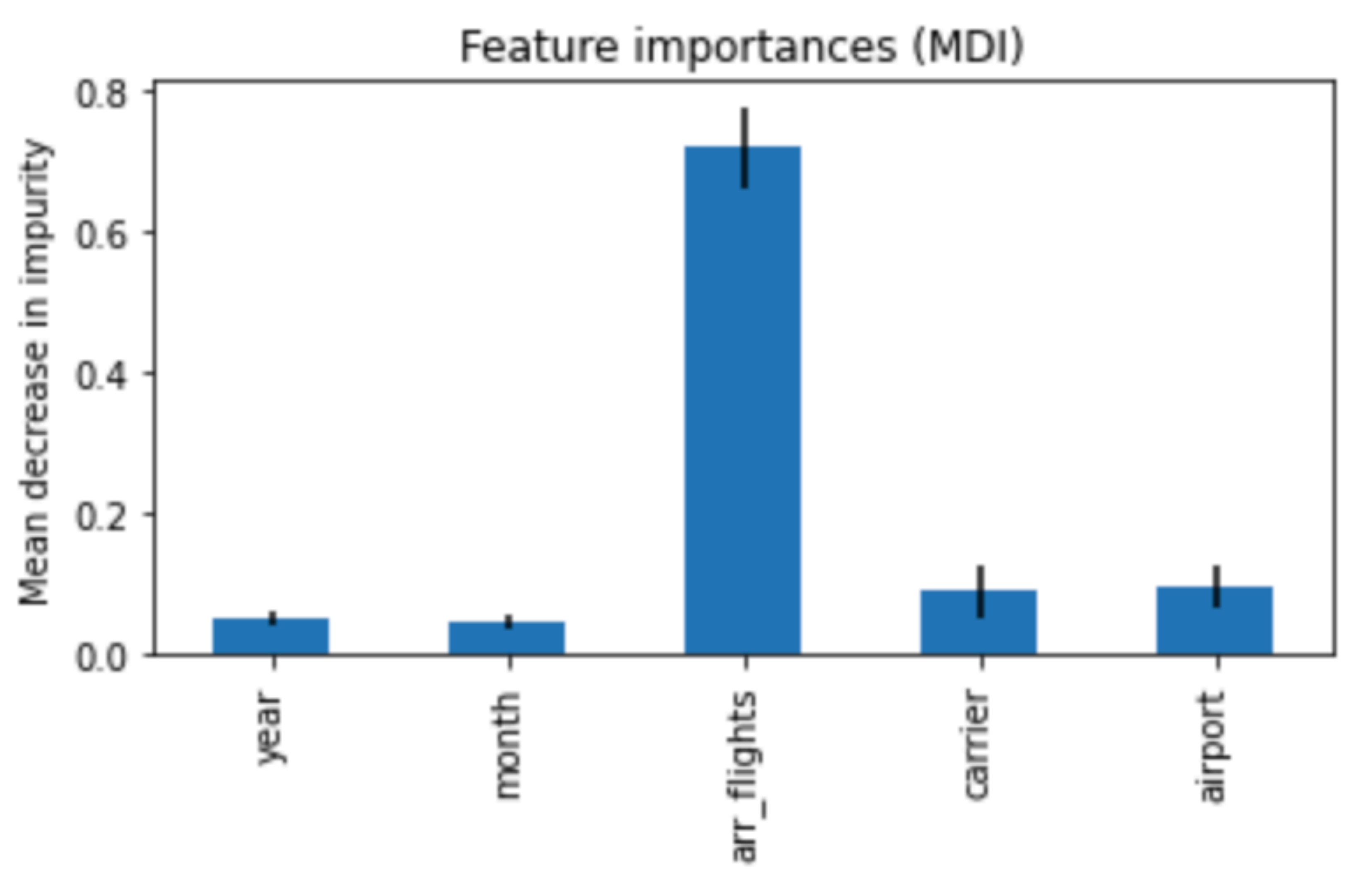
- R^2 = 0.89 (model captures most of the variance in the data)
- negligible (~10 sec.) systematic bias
- ~4 minute precision!

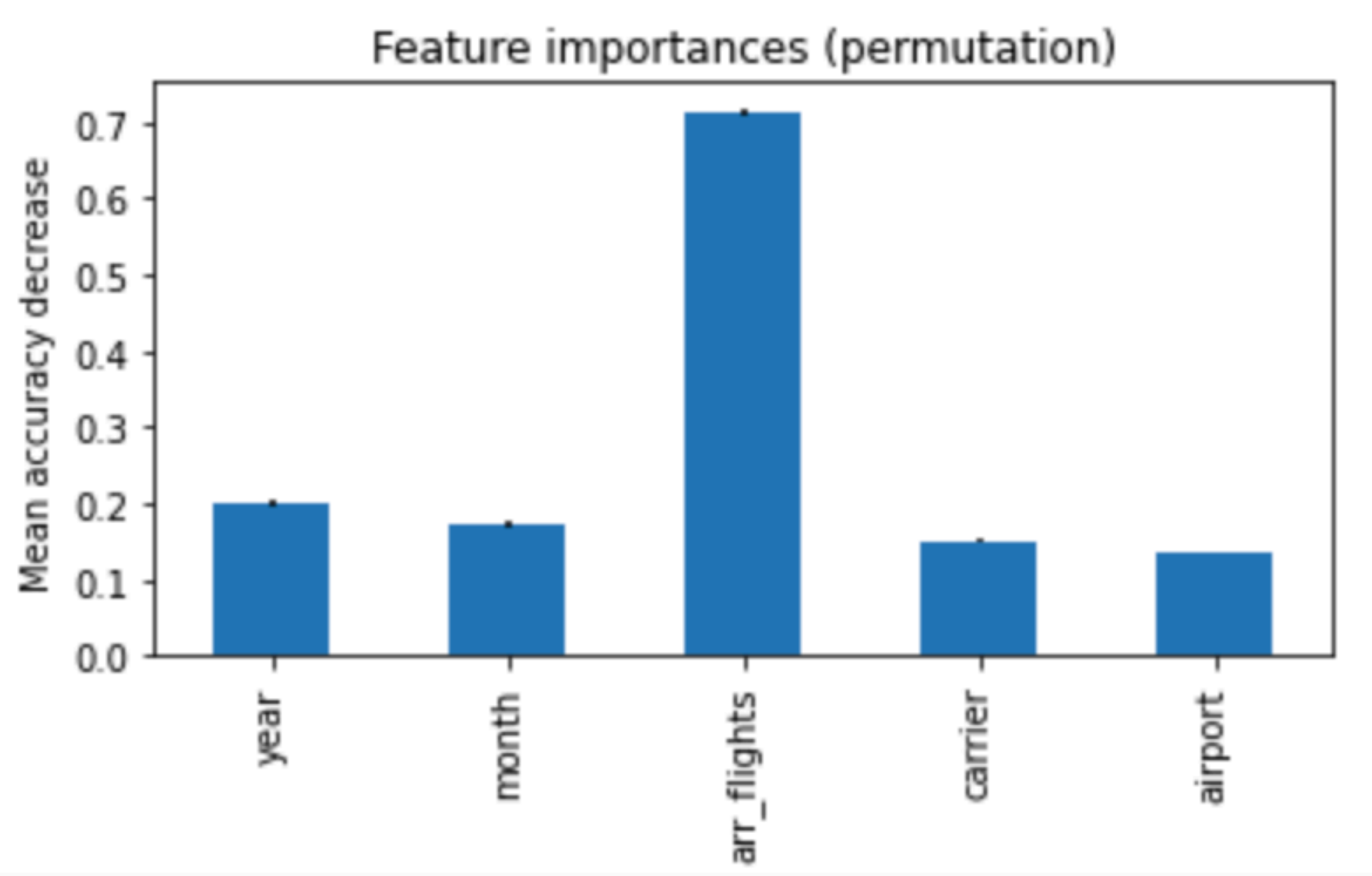




Importance of predictive features

- RF has built-in interpretability
- MDI measures feature importance
- flight volume dominates predictions
- caveat: affected by cardinality
    - use permutation importances
    - year and month have larger impact than airport and carrier
- may need to account for feature covariance

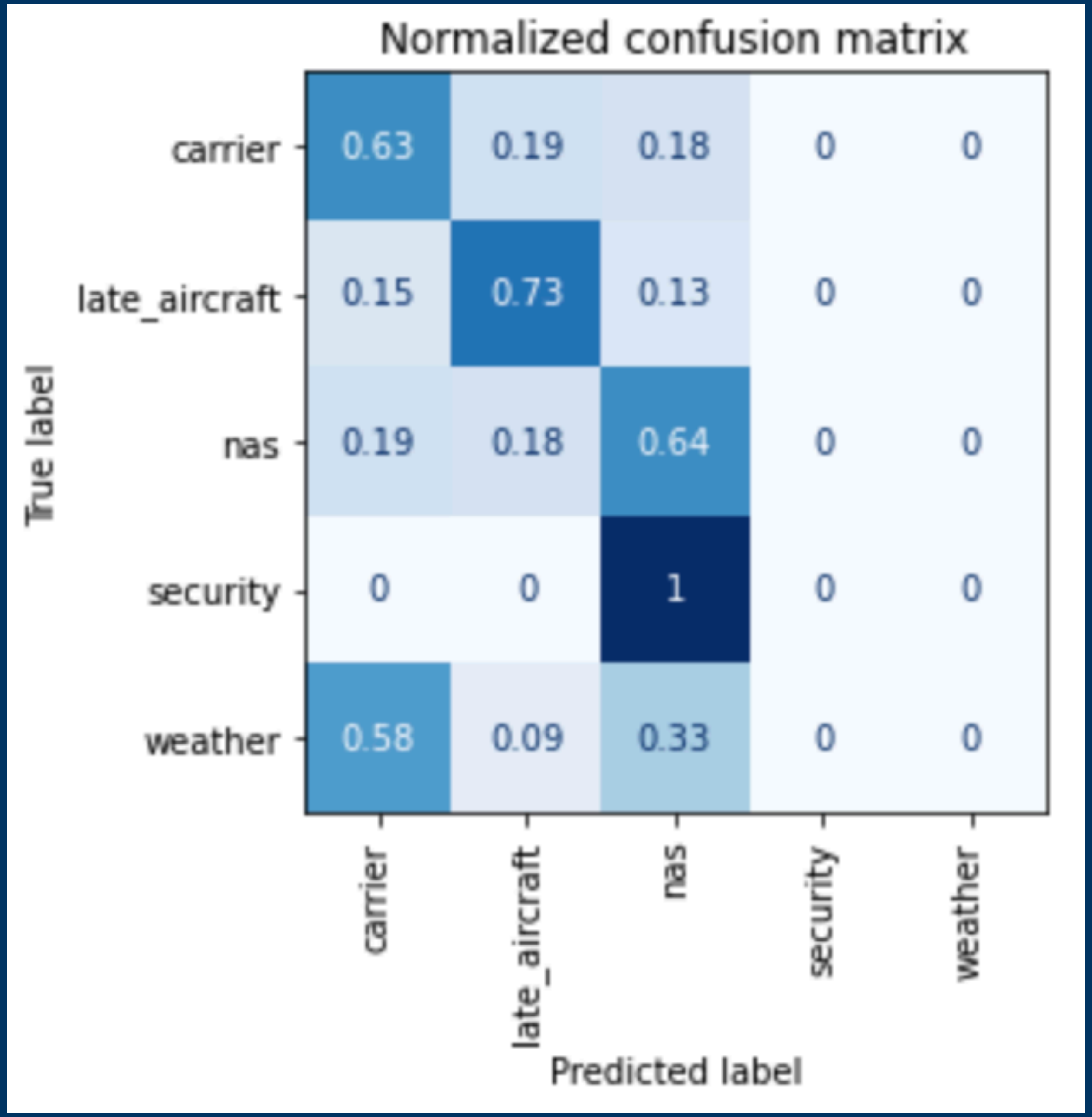




Delay cause prediction

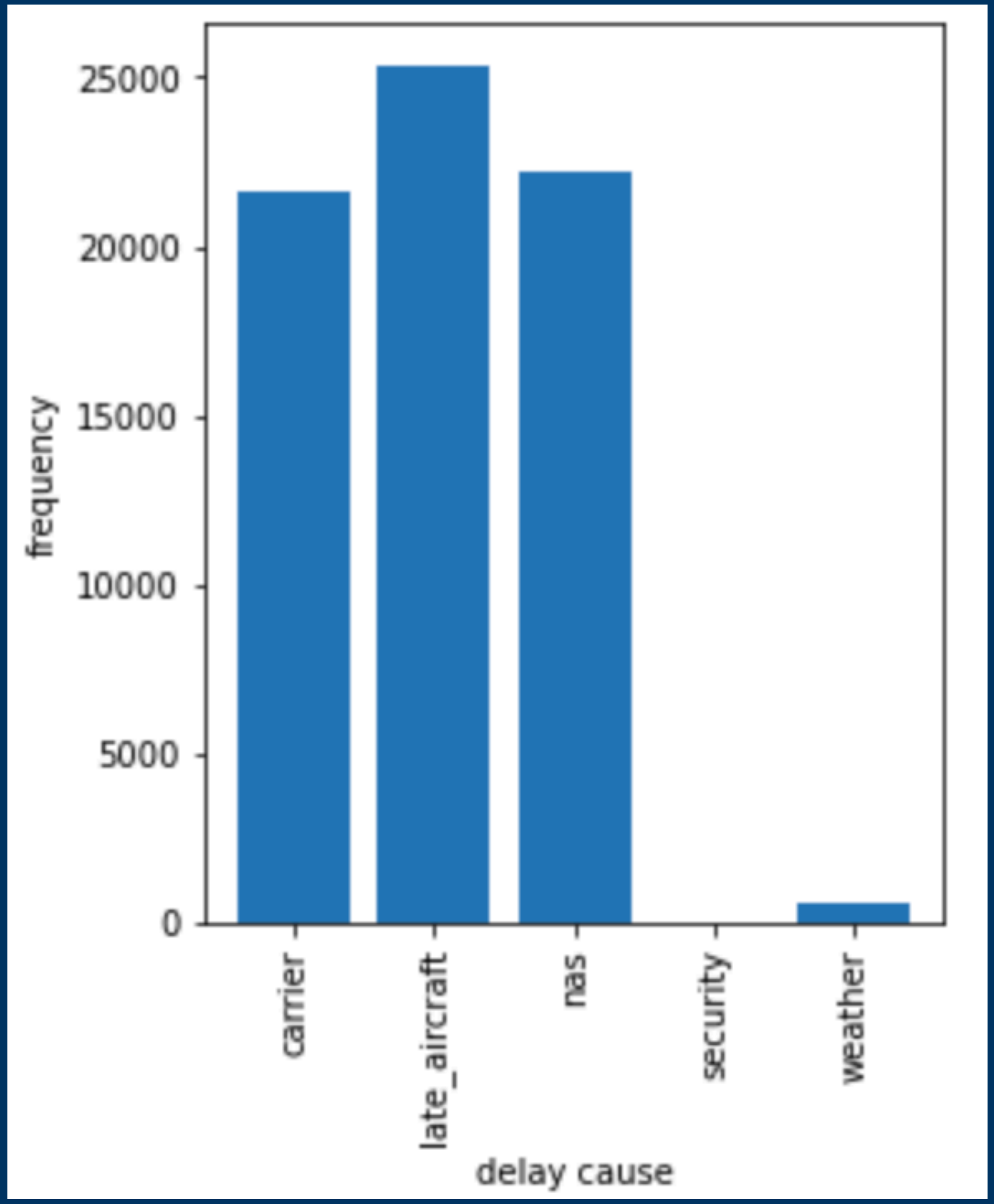
Performance

- **balanced accuracy = 0.38**
- **recall (fraction of correctly predicted causes)= 0.63-0.73 for 3 main causes**
- **recall = 0 for ‘weather’ and ‘security’**
- **weather and security delays are too rare**



Delay cause prediction

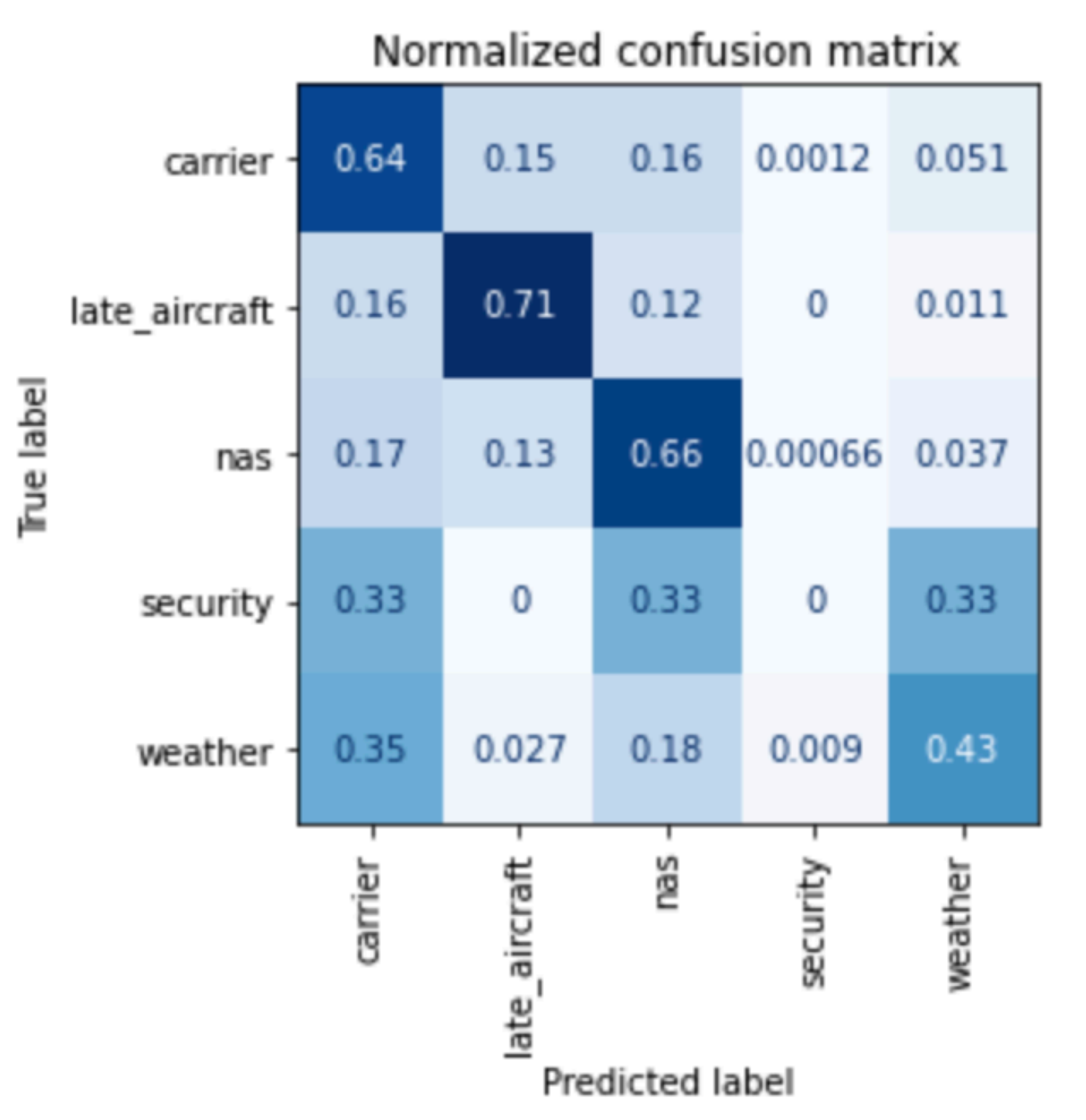
- weather and security delays are too rare and under-represented in training data
- need to account for class imbalance
- use built-in class weights



Delay cause prediction

Performance using balanced weights

- **accuracy = 0.67**
- **balanced accuracy = 0.54**
- **'weather' recall increases to 0.43**



Conclusions

- Flight volume and season have dominant impact on delays
  - airports and carriers less important

- Using only year, month, airport, carrier, flight volume:
  - predicts delay probability and duration
  - predicts main cause of delay in >64% of cases (for 3 leading causes)
  - Delays are driven by previous late arrivals due to high flight volumes

- Predictive performance could be improved by:
  - using tailored algorithms (i.e. deep learning)
  - better handling of class imbalance (e.g. weighting) and tuning decision threshold
  - including more info from discarded features
  - more granular data
  - adding extra features (e.g. location, weather)