FNCE 237:
PREDICTING FLIGHT DELAYS

Sophia Africk & Kaitlynn Soo

Questions & Context

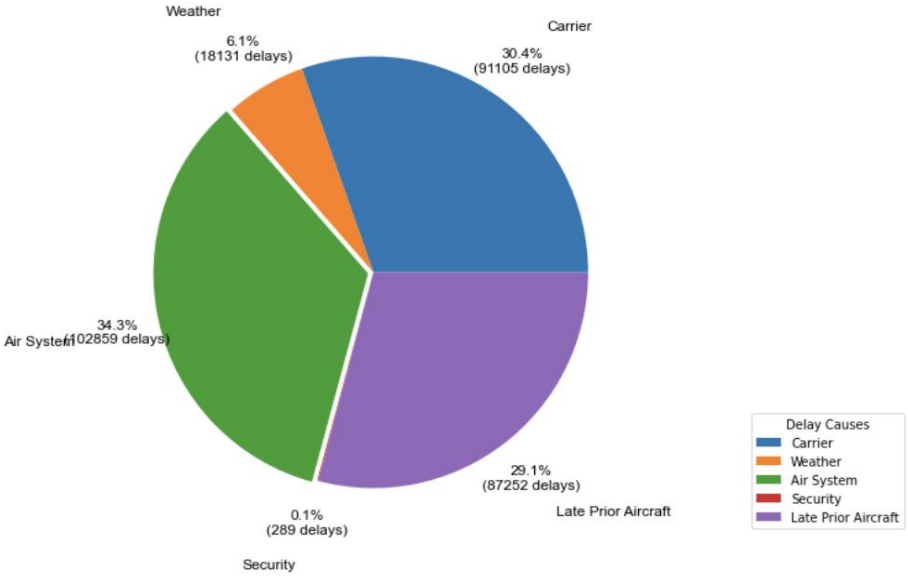
1. ***Cause:*** What is the primary cause of delay for flights in the United States? How does the cause of delay impact the length of the delay?
2. ***Correlated Factors:*** What other factors (airport size, airline, season, day of the week, time of day, etc.) are correlated with delays and the length of delays?

Defining Delayed

The Department of Transportation's Bureau of Transportation Statistics defines a flight as delayed if it ***arrives at or departs from the gate 15 minutes or more after the scheduled time***

The Bureau of Transportation Statistics assigns each minute of a delay to one of five categories:

Delay Count and Percentage by Delay Cause



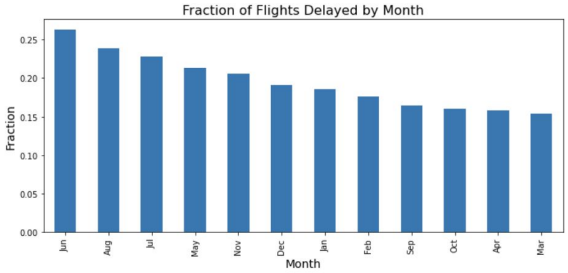
Data

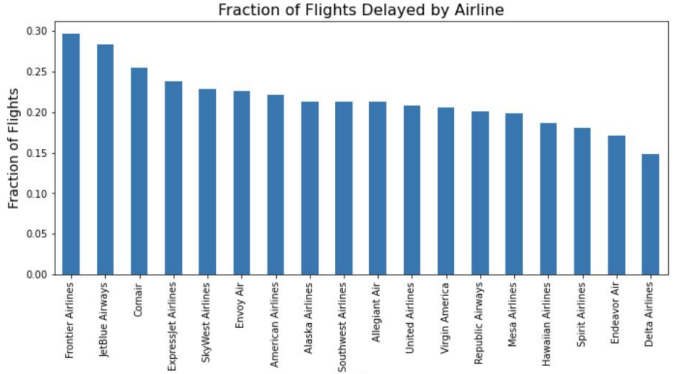
Original data set 2009-2018 with ***60 million+ observations***

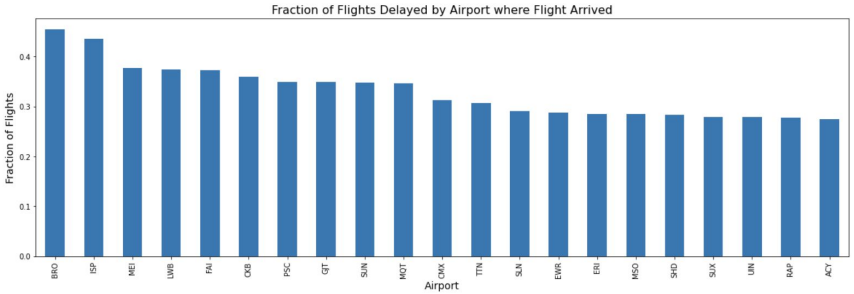
- Narrowed scope to 2018 and top three origin airports → ~7 million flight observations and ~900,000 delayed flights

Notable findings (conditioned on delay):

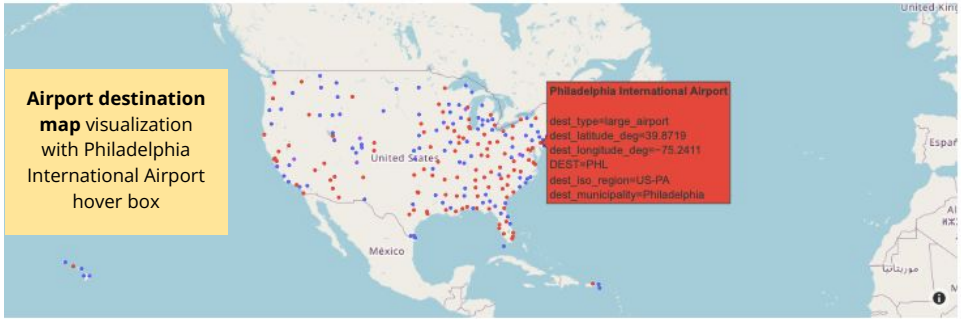
- **Average delay time:** 61.87 minutes
- Delay times and fraction of flights delayed **highest in June** by month and **Summer** by season
- **Airlines with highest average delays:** Frontier, Spirit, ExpressJet
- **Airlines with highest average delay times:** Frontier, Jetblue, Comair
- **Destination airports with highest average delay times:** San Francisco International Airport (SFO), LaGuardia Airport (LGA), and Dallas/Fort Worth International Airport (DFW)
- **Destination airports with highest fraction of delayed flights:** Brownsville South Padre Island International Airport (BRO), Long Island MacArthur Airport (ISP), and Meridian Regional Airport (MEI)
  - EWR only large airport on the list



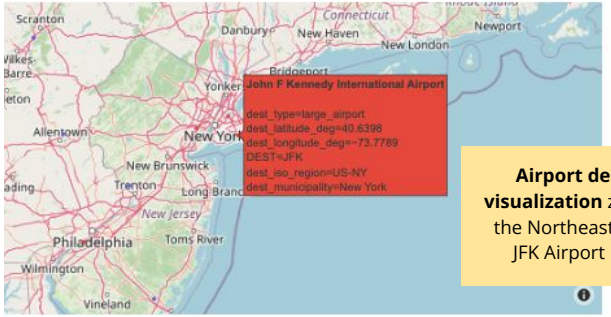




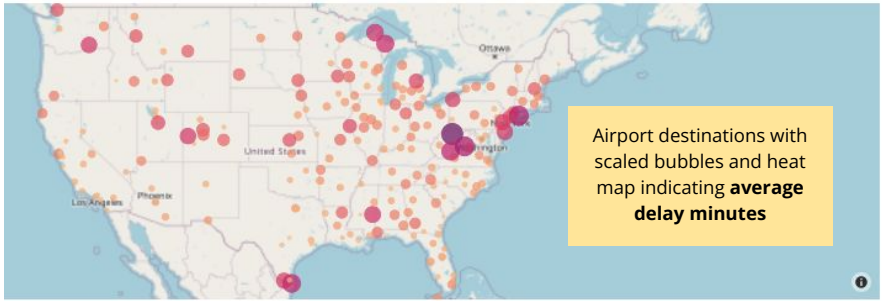
Visualizations with Mapbox



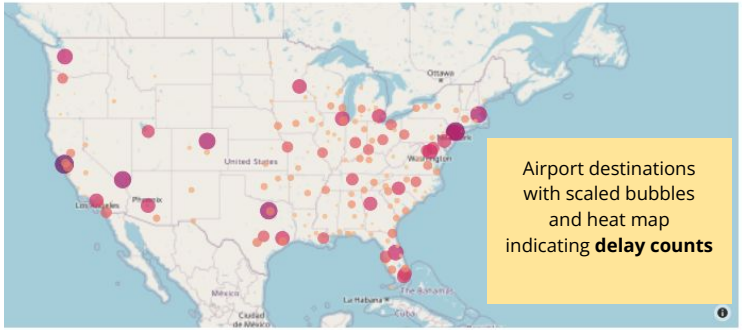
Airport destination map visualization with Philadelphia International Airport hover box



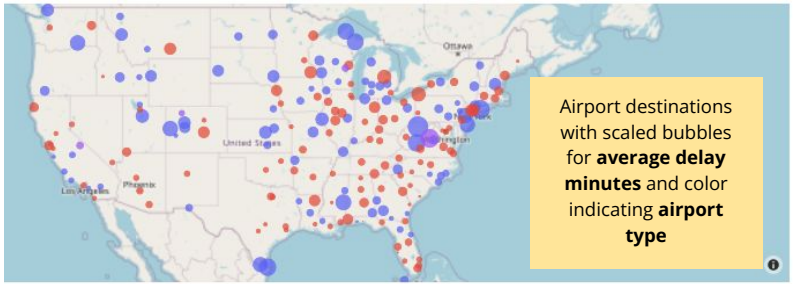
Airport destination visualization zoomed in on the Northeast region with JFK Airport hover box



Airport destinations with scaled bubbles and heat map indicating **average delay minutes**



Airport destinations with scaled bubbles and heat map indicating **delay counts**



Airport destinations with scaled bubbles for **average delay minutes** and color indicating **airport type**

Modeling

Categorical: Will My Flight Be Delayed?

Naive Model

	0	1	pact
0	80.43%	0.00%	80.43%
1	19.57%	0.00%	19.57%
ppred	100.00%	0.00%	100.00%

Cost = c*TP + 0*TN + 0.5c*FP + 2.5c*FN

	0	1	pact
0	66.83%	13.54%	80.37%
1	13.96%	5.67%	19.63%
ppred	80.79%	19.21%	100.00%

Continuous: How Long Will My Delay Be?

Random Forest gives best RMSE, but OLS is more interpretable without sacrificing significant accuracy

```
OLS: -41.830565
Lasso: -41.830478
Random Forest: -40.570830
Gradient Boosting: -40.708335
```

OLS Coefficient Highlights

- Distance: -0.0007
- Reasons: Weather (23.70) → NAS (2.44)
- Airlines: United (3.68), JetBlue (4.95)
- Late Flight: -3.70