Ian Hochstead

August 5, 2021

(DAT201) Python Final (Capstone) Project Presentation

[www.kaggle.com/usdot/flight-delays](http://www.kaggle.com/usdot/flight-delays)

Why?

“general public”

data collection bias

data dictionary

2015 (Entire year)

Domestic (US)

14 (major) airlines / 322 airports

Initial Data Analysis

6.3 million flights (rows), 31 data points (columns)

5 different types of delays

air system, security, airline, late aircraft, weather

44 Megabytes memory allocation – how?

Why? hardware, int64, all flights

Sample(s)

July 15th, July 16th, July 17th

15th: (Wed) 4924 of 17296 28%

16th: (Thu) 3027 of 17419 17%

17th: (Fri) 3169 of 17469 18%

Worst? (in samples)

easy to determine using python \*mode\* function

15th: Southwest, ATL departures, LGA arrivals

16th: Southwest, ORD (O’Hare) departures, ORD arrivals

17th: Southwest, ORD (O’Hare) departures, ORD arrivals

Conclusions:

Avoid Southwest?

Use MDW (Midway)?

Bad weather? Hub?

Further/Future work: More data / Better samples

Specialized information (“city”, hubs)

Beyond single most (mode), top 3/10

Visualizations

Thank you!

Any questions?