Volag

Data-Driven Analysis of Flight Delays and their Causes

Travis Cox*, Alex Hahn*, Cory Sabol, and Josh Moore

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

Motivation

- 8M+ domestic flights per year
- 600M+ passengers per year
- According to our analysis, passengers spent a combined 25 ***years*** of delay time in 2015
- What are the most significant causes of these delays?
- Can we predict future delays?

THE UNIVERSITY *of* NORTH CAROLINA GREENSBORO

Goals

- Identify causes of flight delays
  - Focus on weather observations on the day of a particular delayed flight
- Find other correlations in data
- Predict flight delays
  - Whether or not a flight *will* be delayed
  - How *long* a flight will be delayed

Data

- 2015 Flight Delays and Cancellation Data Set
  - Department of Transportation’s Bureau of Transportation Statistics
  - 5.8M records
    - Origin and destination airport, scheduled departure and arrival times, origin and arrival delays, delay type, cancellations, etc.
- 2015 Global Surface Summary of the Day (GSOD) Data Set
  - National Oceanic and and Atmospheric Administration
  - 4.2M records
    - Station, temperature, wind speed, precipitation, gust, etc.
- Merged data based on airport and weather station proximity

THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

Approach

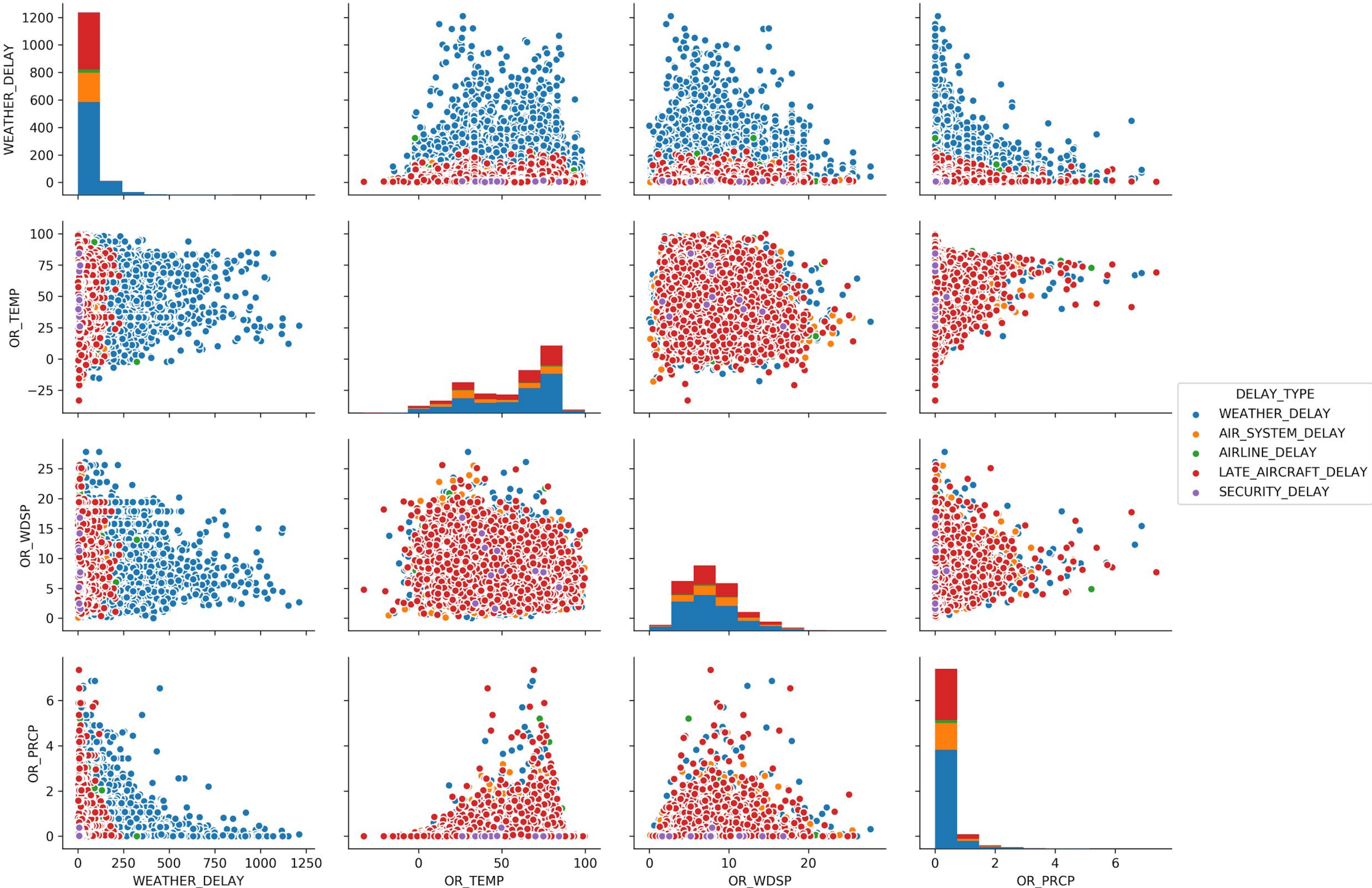
- Performed analysis in Python using Jupyter notebooks
  - Merged, cleaned, and normalized data sets
  - Correlation analysis
  - Statistical analysis
  - Machine learning
- Open science using reproducible results - GitHub and Jupyter notebooks

Merging and Cleaning Data

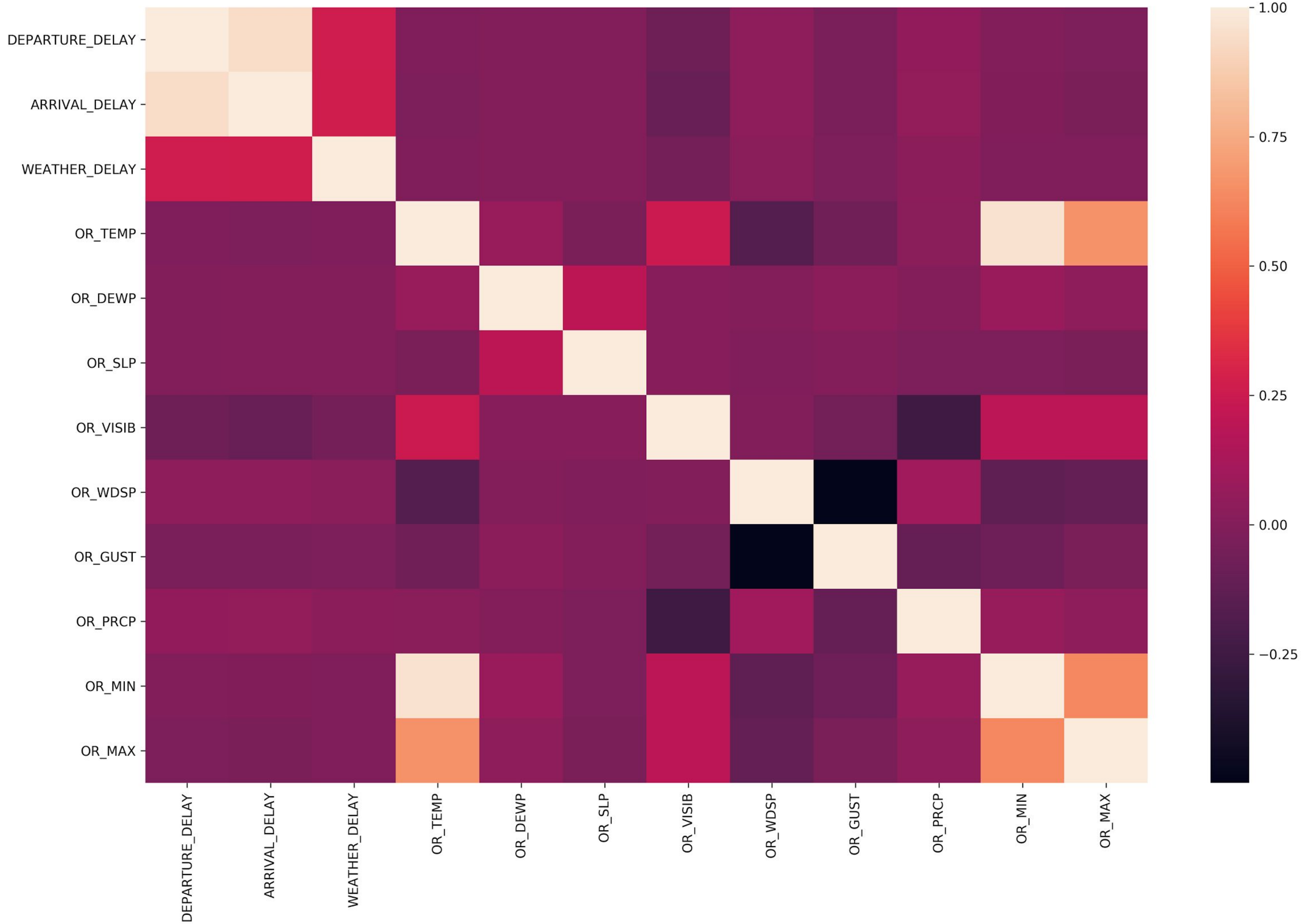
- Connected airports to weather stations based on proximity (many weather stations are actually *at* airports)
- Used weather stations to connect daily weather data to flights
- Removed and replaced missing data
- Added delay type classifications

Correlation Analysis

Does weather correlate with flight delay?



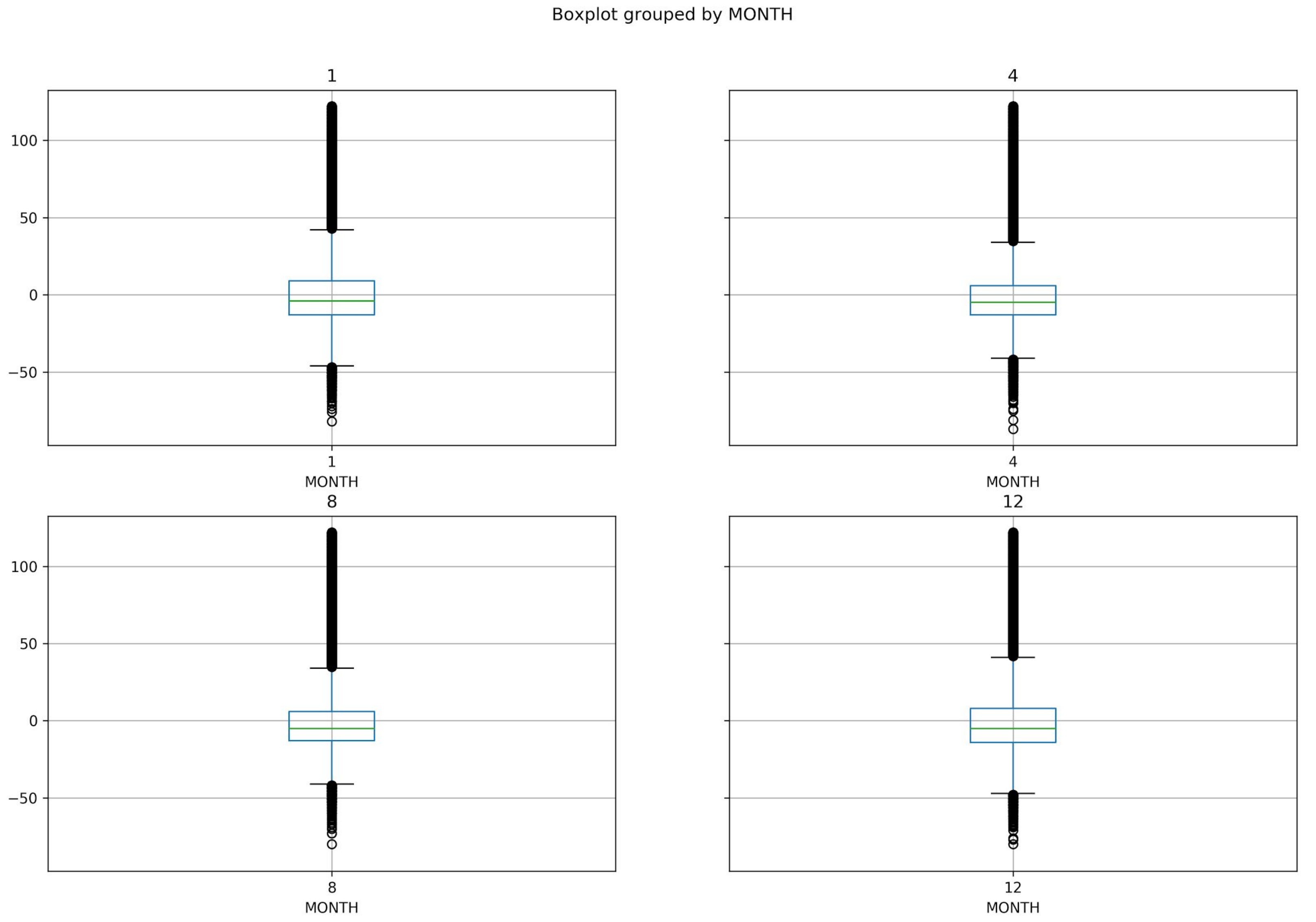
Pair plot of weather features vs. weather-delayed flights



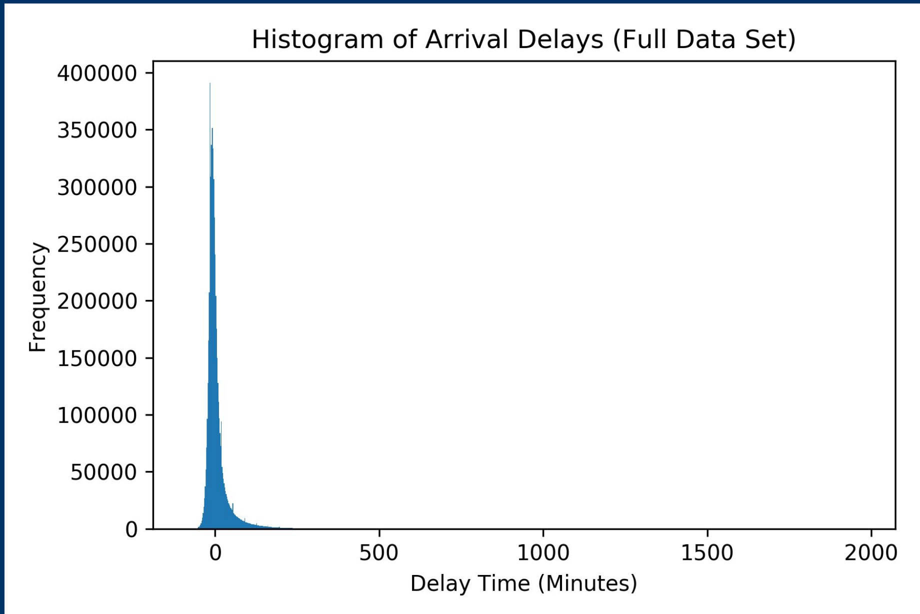
Correlation heatmap of flight delay data vs. weather data

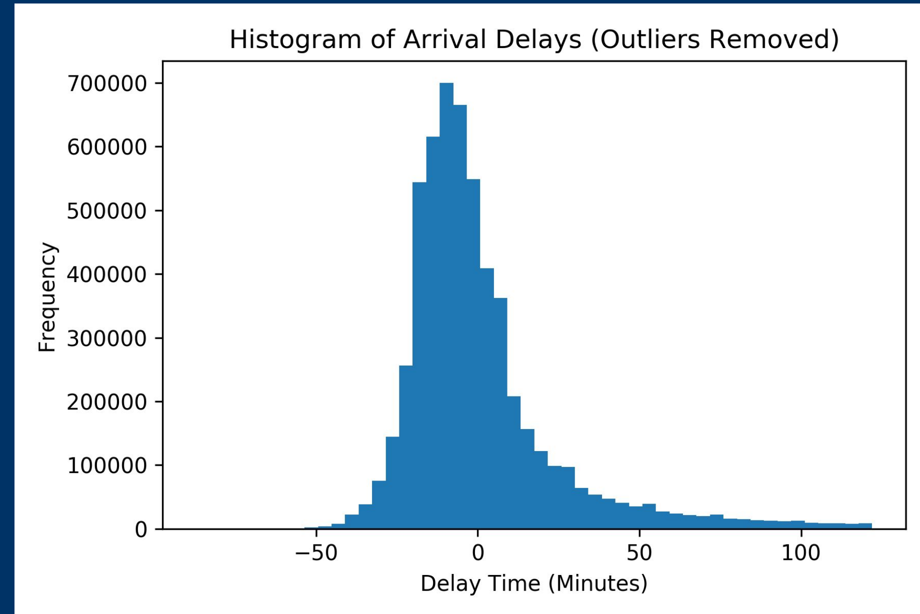
Statistical Analysis

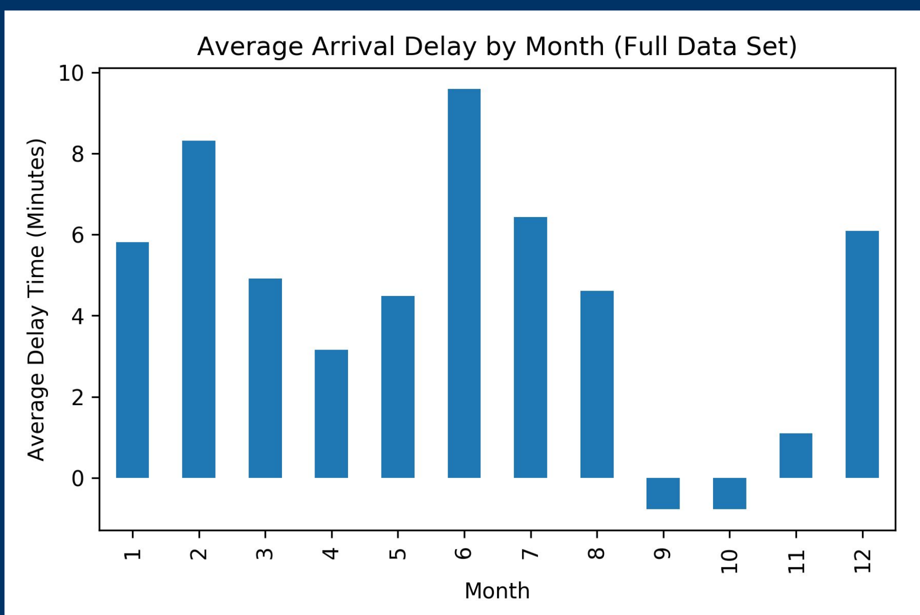
- Analyzed mean and median delay times, standard deviations, distribution shapes, etc.
- Performed hypothesis testing
  - Hypothesis (Alternative): *the weather is significantly different during delayed flights than it is during non-delayed flights*

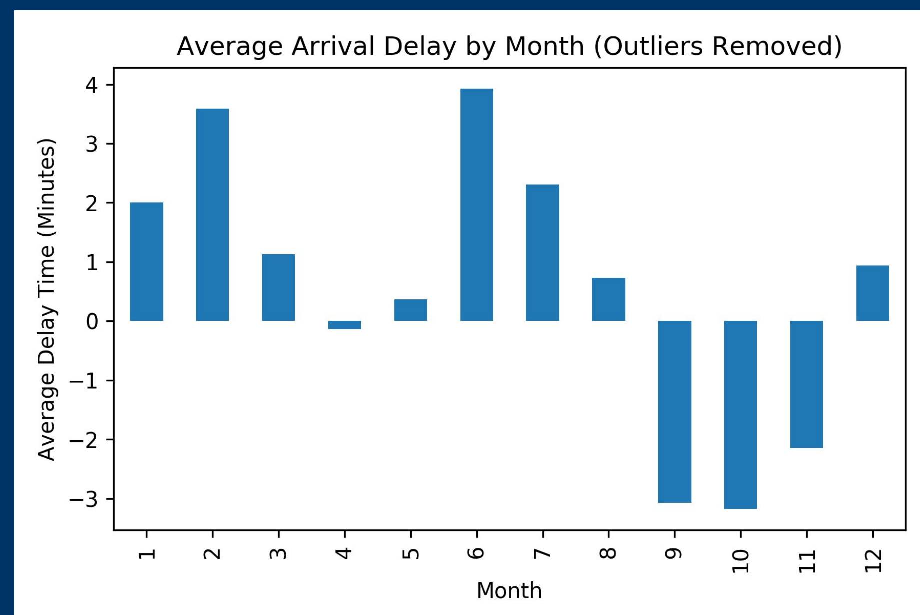


Box Plot Summaries for January, April, August, and December









THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

Statistical Analysis - Results

- **Null Hypothesis:** weather feature *[x]* **is not** significantly different during delayed flights vs. non-delayed flights

- **Alternative Hypothesis:** weather feature *[x]* **is** significantly different during delayed flights vs. non-delayed flights

We were able to reject the null hypothesis (*p-value* < 0.05) for the following *[x]* features:

- Precipitation
- Temperature
- Visibility
- Wind Speed
- Hail
- Tornado/Funnel Cloud
- … and others

THE UNIVERSITY *of* NORTH CAROLINA GREENSBORO

Machine Learning

Classification

- Decision Trees
  - Score = 0.73880
    - 73% accurate classification
- Support Vector Machine
  - On-Time vs. Delayed vs. Cancelled
    - 98% accurate classification
  - Delayed vs. Cancelled
    - 63% accurate classification

Regression

Regression Techniques Used

- Ridge Regressor
- Elastic Net
- Stochastic Gradient Descent
- Support Vector Regression

All performed almost exactly the same as predicting the mean, except SVR which was worse.

Challenges

- Heavily skewed data, especially for delays and precipitation
- Daily weather data instead of hourly
- Large data sets were very slow and resource intensive to process

Conclusion

- Found relationships between several weather features and delayed flights
- However, unable to build predictive models
  - Possibly due to daily weather data (opposed to more frequent observations)
- Flight delays are very low (< 5 minutes on average)

Conclusion

- Multiple causes of flight delays (security, late aircraft, air system, airline)
- In our analysis, weather delays actually contributed a very small fraction of the total delays (~64,000 of 5.8M)

Thanks!