DATA MINING FLIGHT INFORMATION FOR DELAYS AND CANCELLATIONS

DON'T BE LIKE THESE TWO

BUT I LIKE SLEEPING ON THOSE STONE HARD CHAIRS AT DIA

You may want to look away soon then. Through the data given at the website https://www.kaggle.com/usdot/flight-delays , I would like to determine several factors that may contribute to a flight cancellation. Such as weather, airport, airline, etc., and how likely they are to cancel given those condition, and try and find out which airlines are the most trustworthy.

PREVIOUS WORK ON THIS MAGICAL INFO?

The data set given in this presentation is actually pretty recent (released just under 2 weeks ago). Thus, so far not much has been done on the data! Other datasets may have been used but am currently uncertain of their size.

- Datasets:
  - https://www.kaggle.com/usdot/flight-delays (mentioned earlier)
  - https://github.com/j-lopez/DataMiningPlaneDelays (github for any other files)

PLANS MOVING FORWARD

As mentioned before, I would mostly be focused on pattern matching. That is using the data to find out how often planes get cancelled or delayed and what usually causes them to do so.

LIST OF TOOLS

- Data Cleaner
  - Since data is relatively new, there may be a few areas that are messy in the original, massive 5 million line flights.cvs

- Pattern Matcher
  - Qubole looks like a beginner friendly interface to assist in finding patterns in the data.

PRESENTING

- Giving statistics of how one or more factors can contribute to a cancellation or delay of a flight, and compare those results against other big airlines.

QUESTIONS?