ETL project proposal

Wills and Andrew

The transformations preformed are intended to prepare the data to test the correlation between weather in a given city and the rates of light delays from that city.

* **E**xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).

From Kaggle: Historical Hourly Weather Data 2012-2017 -> weather.csv

From Kaggle: 2015 Flight Delays and Cancellations -> airport.csv and flight.csv

* **T**ransform: what data cleaning or transformation was required.

Removing irrelevant columns.

Filter data so that the weather and flight data have the same date range

Convert day and time columns in fight.csv to a timestamp format

Change SCHEDULED\_DEPARTURE column from an integer to a string for further analysis

Round datetime from flight data to nearest hour

Merged data frame from flight.csv with airport.csv

Changed column names to match schema

* **L**oad: the final database, tables/collections, and why this was chosen.

Create schema in postgres because this was what we were most conferrable with.