**Data sources:**

* Flat file:

This is a 2015 flight delay data package, with flight info from all US major airport and delay info for the first 3 months from year 2015. The link is below. The original package is about 600 mb. To reduce size and improve speed to load the data, I filtered out the airport I am interested and created a new smaller file, in excel format.

The datafile has 31 columns and >7500 rows.

A screenshot of a computer

Description automatically generated

<https://www.kaggle.com/datasets/usdot/flight-delays>

* API

I am using the openweathermap API, which I can grab the weather info by using zip code or city name. I have obtained API key.

<https://openweathermap.org/api>

* Website:

This is a wiki website with a table of Texas airport related info, like airport code.

<https://en.wikipedia.org/wiki/List_of_airports_in_Texas>

A screenshot of a computer

Description automatically generated

**Dataset relationship:**

The 3 datasets are linked by city code.

* The flat flight info file has an airport code which is related to the interested city.
* The weather API can be called using city name.
* The website table provides a look up between city name and airport code.

**Project subject:**

Simulate the situation that using data preparation and wrangling skills to provide answer to a traveler want to find out which flight is operating from the city he is in, what’s the weather and if it will be delayed due to weather conditions.

**Project plan:**

From the dataset chosen, there are flights and weather information (available airlines in and out of the chosen city, the historical delay information and weather info on current date). This is critical information for a traveler.

I will first use some basic web scraping skills to get airport code from the website using city name. then the city name and airport code will be the keyword to link other data sources.

The flat file, which contains the information on available flights from that departure city to possible destination. It also has historical data of delays on these flights. I will need to create a data frame, prepare, and clean the data, use airport name as key to create subset of data, using wrangling functions from proper python packages to filter out the needed information.

I will then be using API to extract weather information for the departure airport, which will give travelers some ideas to prepare for the travel.

I would like to output a travel summary, like flight available, the weather info and if there is a historical delay on a particular flight/airline.

**Challenges:**

There are some challenges I need to adjust to.

First, there are time mismatches of the information. From the source I could find, the flight delay information is always historical data, while weather API will only provide current weather forecasts.

Second, the dataset I am using only contains information for the first 3 months in the year 2015 so I cannot get information on any date of the year.

Fortunately, the term of this class falls into the first 3 months of year. As a practice project, I will ignore the time difference and pretend at the same year (i.e: flight in Jan 1st 2015 is the same as flight in Jan 1st 2024).