# DATA 620-Assignment 2b (Group 4)

The dataset group 4 has selected to use is the highly popular flights database. The origin of this data can be found here:
https://openflights.org/data.html

We can extend our data to include elements such as ferry, terminal, and station if needed from the master flights data found here:
https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports-extended.dat

In order to meet the "one categorical variable" requirement, project member Saayed transformed a numerical variable into a categorical variable which will be touched upon in more detail later on.

## Data preparation:

In order to prepare the data for downstream analysis, we need to perform some operations on our data such that the data resembles a pairwise structure necessary for turning the data into a graphical network. We envision the nodes as being airport destinations where each pair represents flights to and from destinations, hence edges. We derive our own categorical variable from the arrival and departure columns. Our new classification says that if the value is 1, then the flight was early or on time. If the value is 0, then the flight was delayed. The data is prepared with a simple r script that will be included in the project. We then export the data into a csv and upload to a GitHub location.

## Data Loading:

The data will be read into the jupyter notebook environment directly from the GitHub url where the data is being kept. This is mainly to foster an environment where our work can be reproducible.

## Hypothetical Outcome:

By definition, degree centrality is a measure the number of neighbors a node has. This measure is a good indicator if a node is important or not. In our case, if a particular airport is a central hub that has flights going to more destinations than other smaller airports. We expect airports such as LAX to have a high number of neighbors but we also expect such a big airport to have the most delays as shown with our derived categorical value.