# CIS 545: Big Data Analytics

*Fall 2018*

## Homework 1: Data Wrangling

Due September 26, 2018 by 10pm

Big data analytics often requires (1) importing data from multiple sources, possibly *extracting* content from text; then (2) combining data from multiple sources, possibly from multiple different organizations, in heterogeneous formats. For this assignment, our primary goals are to get you comfortable with importing, extracting, saving, and combining data -- largely using Pandas DataFrames as the basic abstraction. You'll read data from files, address issues with missing values, save the data in a database, combine the data, and generate visualizations.

This assignment is the first of four in the course that consists of a **Basic** component, to be done by everyone, and an **Advanced** component, to be done by students who wish to do 3 homeworks and a project. Please see the separate steps for the Advanced component.

## The Basic Goals of the Assignment

Most of you likely were on a plane over the summer, and chances are that at least one of your flights got delayed. Did you ever wonder how well different airlines do? We'll answer those questions in this assignment! *(Caveat: with apologies to international travelers, most of the data is only available for US cities and routes!)*

**Terminology**. We'll generally use **field**, **column**, and **attribute** interchangeably to mean a named column in a DataFrame. We'll also generally assume that **table**, **DataFrame**, and **relation** mean the same thing.

**Notebooks.** In order to give you the best feedback, we are using a system called **nbgrader**, which provides a "skeleton" notebook into which you can add your code. It will also be used to automatically grade your submissions.

We supply some test cases in the notebook to help you figure out if your code is following our expectations. Note that there are many *additional tests* that we'll be using to grade your homework, but that passing the validation tests will be a prerequisite to passing the rest!

## The "Basic" Assignment

We'll be bringing together data from OpenFlights.org with data from the US government Bureau of Transportation at http://www.transtats.bts.gov. To start, go to **Jupyter Notebook** in your web browser (http://localhost:8888/tree with the big token as before). Click on your **work** directory, then New|Terminal. Run:

```
git clone https://bitbucket.org/pennbigdataanalytics/hw1.git
```

to get your initial data sets and skeleton notebooks.

### What to Work on

Homework 1 has two notebooks:

- **Homework 1-1.ipynb**: start with this one, which loads and cleans data, and ultimately writes it to a local database.
- **Homework 1-2.ipynb**: when you are done with Part 1, continue from here.

## The Data You'll be Using

The data files, whose contents are described in the provided notebook Dataset Descriptions, are:

- **airports.dat.txt** - data on airports, in comma-separated values (CSVs) with no header row
- **airlines.dat.txt** - data on airlines, in CSV with no header row
- **routes.dat.txt** - data on flight routes, in CSV with no header row
- **aircraft_incidents.htm** - webpage that lists commercial aircraft incidents by year
- You'll also pull a file from **docs.google.com/uc?export=download&id=1PPtjGx8lr_cDUfVa3qwlk1W8yY6hY91n** - data on actual flights, with performance info, with a header row

## Submitting Homework 1

Once both of your Jupyter notebooks are sanity-checked and pass all tests, go into your **work/hw1** directory on Jupyter Notebook. Run **zip hw1.zip Homework*.ipynb**.

Next, go to the submission site, and if necessary click on the Google icon and log in using your Google@SEAS or GMail account. At this point the system should know you are in the appropriate course. Select CIS 545 **Homework 1** and upload `hw1.zip` from your `Jupyter/hw1` folder, typically found under `/Users/`*{myid}*.

# The "Advanced" Assignment

You should make sure the basic assignment is complete before moving on to this part. To start, go to **Jupyter Notebook** in your web browser (http://localhost:8888/tree with the big token as before). Click on your **work** directory, then New|Terminal. Run:

```
git clone https://bitbucket.org/pennbigdataanalytics/hw1-advanced.git
```

to get your initial data sets and skeleton notebooks. Then `cp hw1/HW1_DB hw1-advanced` to copy the SQLite database.

## What to Work on

In Jupyter's tree browser, go into `hw1-advanced`.

- For this component, you'll be working on **Homework 1-Advanced.ipynb**.
- If you successfully completed the Basic part of the assignment, your HW1_DB database should include the data needed by the Advanced portion. (There will also be a few more files downloaded by the notebook from the Web.)

## Submitting Homework 1-Advanced

Once both of your Jupyter notebooks are sanity-checked and pass all tests, go into your **work/hw1-advanced** directory on Jupyter Notebook. Run **zip hw1-advanced.zip Homework*.ipynb**.

Jupyter Notebook. Run **zip hw1-advanced.zip Homework*.ipynb**.

Next, go to the submission site, and if necessary click on the Google icon and log in using your Google@SEAS or GMail account. At this point the system should know you are in the appropriate course. Select CIS 545 **Homework 1-Advanced** and upload `hw1-advanced.zip` from your `Jupyter/hw1-advanced` folder, typically found under `/Users/`*{myid}*.