DAV 5400 Fall 2019 Project 1 (100 Points)

This project will allow you to demonstrate your ability to: (1) make use of Python's Pandas library; (2) add, remove and transform data within a data frame; and (3) generate basic summary statistics and graphics as part of your exploratory data analysis work. ****You may work in small groups of no more than <u>three</u> (3) people for this project.*** **

The data set we'll be using is a 20,000 row subset of the **hflights** package provided within the **R** programming language and sourced originally from the US Bureau of Transportation Statistics:

https://www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=120&Link=0

The data set contains information and metrics for commercial domestic (USA) airline flights that departed from 2 separate airports located in Houston, TX (IAH and HOU) during the 2011 calendar year.

A summary / description of the data set can be found here:
https://cran.r-project.org/web/packages/hflights/hflights.pdf

The flight attributes provided within the data set are as follows:

- **Year, Month, DayofMonth**: date of departure
- **DayOfWeek**: day of week of departure (useful for removing weekend effects)
- **DepTime**: departure time (in local time, hhmm)
- **ArrTime**: arrival time (in local time, hhmm)
- **UniqueCarrier**: unique abbreviation for a carrier
- **FlightNum**: flight number
- **TailNum**: airplane tail number
- **ActualElapsedTime**: elapsed time of flight, in minutes
- **AirTime**: flight time, in minutes
- **ArrDelay**: arrival delay, in minutes,
- **DepDelay**: departure delay, in minutes
- **Origin**: origin airport code
- **Dest**: destination airport code
- **Distance**: distance of flight, in miles
- **TaxiIn**: taxi in time in minutes
- **TaxiOut**: taxi out time in minutes
- **Cancelled**: cancelled indicator: 1 = Yes, 0 = No
- **CancellationCode**: reason for cancellation: A = carrier, B = weather, C = national air system, D = security
- **Diverted**: diverted indicator: 1 = Yes, 0 = No

To load the data set into your personal Python environment, use the following Pandas function call (**NOTE**: be sure to change the name of the data frame to something more appropriate - don't use "YourDataFrameName" within your own code):

***** NOTE: The Project Description Continues on the Following Page *****

```
-------
filename = “https://raw.githubusercontent.com/jtopor/DAV-5400/master/Project1/hflights.csv”

YourDataFrameName  = pd.read_csv(filename)
--------
```

After reviewing the data, you should define at least three (3) (or more if you are motivated!) interesting research / analytical questions you will seek to answer using the data. When formulating your questions, be sure to think about who would care about the answer to the questions and how they might make use of the results of your analysis. Provide a short written narrative that explains your justification for your questions **using formatted Markdown cells in your Jupyter notebook .** Use the research questions you formulate to guide your work throughout this Project.

Once you’ve loaded the data set, you should perform some basic transformations on the data frame. For example, you might decide to make use of only a particular subset of the provided attributes as you work to answer the analytical questions you’ve defined. If so, you might choose to simplify the data frame by subsetting only those attributes you plan to make use of, or rename any column names you find confusing.

You might also choose to create one or more new columns within the data frame by creating new attributes you derive via calculations based on the original data (e.g., ‘flight date’, etc.), or perhaps filter out certain rows of the data frame based on the requirements of the research questions you’ve chosen to answer. **Be sure to provide a written narrative using formatted Markdown cells in your Jupyter notebook explaining any transformations you have chosen to make to the data set**.

You should then perform some exploratory data analysis, including the generation of basic summary statistics and basic Pandas graphics (e.g., histograms or box plots or scatter plots as appropriate) for the attributes you have elected to work with for purposes of answering your research questions and **provide a written narrative using formatted Markdown cells in your Jupyter notebook that explains your exploratory data analysis** to a reader of your research, e.g., “the summary statistics for the ___variable indicate right skew in the variable’s distribution, as evidenced by .. ”, etc.

Once you’ve completed your exploratory data analysis, answer the research questions you’ve defined, making use of descriptive and summary statistics as well as basic Pandas DataFrame graphics ***where appropriate***. Be sure to include a narrative **using formatted Markdown cells in your Jupyter notebook** describing your approach to answering each of your research questions.

Save all of your work for this project within **a single Jupyter Notebook** and upload it to your online DAV5400 GitHub directory. Be sure to save your Notebook using the following nomenclature : **first initial_last name_Project1**" (e.g., J_Smith_Project1). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team’s work within Canvas.***

As a reminder, Project 1 is due no later than 11.59pm on Sunday Sep 29.