Assignment 8

This is a group assignment. I have taken the following files for this assignment from the nycflights13 library in R. Additional information about these datasets is available at this link and at this one. You will need to use Python and Pandas to complete this assignment using Jupyter Notebook. You may not have to use all these files to complete the assignment. Assume all files are in the same folder as the notebook files. Provide your answers in the notebook itself by running those cells are markdown cells.

Part A

Read the following files into Pandas data frames

```
airlines.csv
airports.csv
flights.csv
planes.csv
weather.csv
```

Provide the following using Python:

i. Print out the number of rows in each of the five files.
ii. Show histograms of all airlines for departure delay. Which airlines is the best in terms of on time departure. Do this using
    a. ggplot, and
    b. matplotlib

    in separate Jupyter notebook cells
iii. Create a dataframe containing the means and standard deviations of inter-arrival times for the top 10 busiest airports (in terms of highest average number of flights arrived per day) in the list provided. The data frame should contain the full name of the airport and the average number of flights per day, means of inter-arrival times, and standard deviations of inter-arrival times.
iv. Provide the histogram and density plot of the inter-arrival times for the top five busiest airports in the list provided. All plots should be in the same figure.

Part B

i. Use the **flights** and **weather** date and left join flights and weather on "year", "month", "day", "hour", and "origin"
ii. Select all flights that departed from JFK that were delayed in departure by more than 10 hours and store them in a separate table called **`mydelay`**
iii. What were the average weather conditions for these delayed flights at the time of scheduled departure in terms of temperature, humidity, wind speed and precipitation?
iv. Provide plots of distributions for all these weather variables.
v. Save data from mydelay into a csv file called **`mydelay.csv`**.
vi. Is there a difference, across airlines, how arrival delay is distributed? Please analyze this using histograms **faceted** by airlines.