Lab 7: Wrangling at Cruising Altitude

In this lab, you'll continue your work with the flights data and apply a new set of operations that greatly empower your ability to deftly wrangle a data set with remarkably concise syntax.

- Adding columns
- Sorting a data frame
- Finding unique values
- Grouped Operations

Begin, as before, by loading in the data from the `boxofdata` package and then into Python.

```r
# install.packages("remotes")
# remotes::install_github("andrewpbray/boxofdata")
library(boxofdata)
library(reticulate)
library(dplyr)
data(flights)
small_flights <- flights %>%
  sample_frac(.10)
```

```python
import pandas as pd
flights = r.flights
```

These wrangling puzzles are challenging. I recommend starting with pencil and paper, laying out each of the steps, and keeping handy the help file that describes each of the variables.

1. Start by revisiting the questions from the end of the previous lab. Use `.isin`, `.groupby`, and `agg()` to find the mean and median departure delay from the Bay Area to the NYC area in July and November.

2. What proportion of the flights in the full data set are red-eyes? As part of answering this question, add a new boolean column to the data frame indicating whether each flight is a red-eye.

3. On average, are red-eye flights from SFO to JFK more or less likely to arrive (on time or early) or (late)?

4. On showing this result to a friend, they reply:

   Yes, but that's not really due the flights being red-eyes. As the day goes on, flights at an airport get more and more backed-up. Flights that leave in the

evening, like red-eyes, are just more likely to be delayed.

Evaluate this explanation using this data set in two ways: looking at (the proportion of flights that are delayed) and (the average delay) at different points in the day. How do you interpret these results and what does it suggest about red-eyes and delays?

Various and sundry

The remaining questions provide a sense of the diverse questions that can be answered with this data set. Some guidelines and hints:

- For each question, describe any wrinkles/surprises that came up and how you addressed them. Relatedly...
- Some of these questions are not fully answerable given the data in `flights`. If that's the case, be clear about the special case of the question that you're focusing on or the assumptions that you're using (do not bring in other data).
- It can be easiest to work these problems backwards: what does the data frame look like that will answer this question? What is the unit of observation of that data frame? Usually the answer to that question is a hint about any variable you may need to group by.
- If in your work you get a `SettingWithCopyWarning`, read through the explanation of that warning here.

5. Return to the single flight that you studied in the previous lab. Compared to other flights to that destination, how does your arrival delay compare? Answer this by comparing your flight to other flights on the same route taken within two weeks of the day of the flight. Calculate the percentile of your arrival delay in that distribution and then bring the data into R to visualize where your flight is in the context of that distribution.

6. Which routes are served by the greatest number of carriers? List the top five.

7. Which airplane has traveled the greatest distance over the six months covered by this data set? How long in total was it airborne? How many distinct routes did it fly?

8. Certain airports are notorious for the amount of time you have to spend on the tarmac waiting to taxi to your gate. What are the top five airports (the worse offenders) in terms of time spent on the tarmac taxiing upon arrival?