Lab 8: Wrangling Arrival

This is the third and final installment in the data wrangling series. You'll wrap things up by practicing three last skills common in data science.

- Date-time methods
- String methods
- Joins

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)
```

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

1. You saw in lecture that if you have a column that is a Series of strings, you can append the *string accessor* `.str` to have access to a host of useful element-wise operations for strings. You can access analogous functionality for a Series of date-time data using accessor `.dt`. Take a look at the available methods either in the Pandas docs online or by calling `dir(flights["time_hour"].dt)` (note that when you call `flights.dtypes`, `time_hour` is of type `datetime64`).

   Use date-time methods to determine which day of the week is best to fly from the Bay Area to Seattle if you want to minimize delays. Does the answer differ between Oakland and San Francisco Airports?

2. Return to a question from the second data wrangling lab:

   > 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?

   The main challenge was the calculation of total time (total time minus air time gives you what can be interpreted as tarmac time). Arrival and departure times are always

listed in local time, so it doesn't work to simply find the difference between these two if the plane changed time zones. You can now account for this programmatically because the timezone of all of the airports is available in the data set in `boxofdata` called `airports`.

Use your knowledge of how to join multiple tables to re-answer this question in a more satisfactory way.

3. Return to a question from the first data wrangling lab:

   > Create a second data set called `nyc_jul_flights` that is similar, but contains data from July. Calculate the mean and median departure delay. Use any knowledge you have about air travel and weather to speculate why you see this difference between July and November.

   You found that, perhaps surprisingly, delays were more common in July than they are in November. One of the explanations offered by several students is that during the summer, San Francisco Bay often gets fog that can delay flights at both airports. Evaluate this explanation using the data found in the `weather` data set in `boxofdata`. Explain clearly how you are using the weather data to infer the presence of fog.

4. The two biggest manufacturers of commercial aircraft are Boeing and Airbus, followed by Embraer and Bombardier. For each of these manufacturers, which airline schedules the greatest proportion of their flights out of SFO and OAK using planes made by that manufacturer? The `planes` data frame in `boxofdata` is useful here. Note that the manufacturers sometimes use slight variants on their names; you will want to collapse those variants into a single name.