R Bootcamp

Day 2 Exercises: dplyr

September 14, 2015

Questions in boldface are "bonus" and may require functions or techniques that we haven't covered in class. If you don't want to try them, that's fine; if you do, I encourage you to practice finding the information you need via Google and using it on your problem. This is an invaluable skill.

Fire up a new R Markdown document. Silently load `dplyr`, `ggplot2` and any other packages you will use.

Download the flights.RDS file from the class smartsite. Read the file into R using readRDS. Get a sense for what the size of the data and what it represents using `str`, `summary`, and/or `head`: What spatial and temporal domain do they cover? There is additional information about the data at http://stat-computing.org/dataexpo/2009/the-data.html (http://stat-computing.org/dataexpo/2009/the-data.html). Because the dataset is quite larger, it may be advantagous to build your analyses on a small subset of the data, and once you have them all in or near final form, biuld your document using the full dataset.

1. How many flights are there in the dataset?
2. Which airline has the most flights? The least?
3. How many airlines fly from LAX to SFO?
4. a. To what airports can you fly from Sacramento (SMF)?
   b. Which of these represents the longest distance flight?
   c. Plot the average time to each airport from SMF. Make sure the order of the destination airports makes the plot easy to read.
5. a. Identify the ten busiest airports and five highest-volume airlines.
   b. Among those, compute the number of flights each airline had at each origin airport.
   c. Display this information graphically. (**One possibility is a heatmap using `geom_tile`.**)
6. Answer all of the following questions with a plot.
   a. How does flight volume change over the days of the week?
   b. How do flight delays change over the days of the week?
   c. What is the relationship between departure and arrival delays?
   d. Calculate the difference between arrival delay and departure delay. What is the relationship of that value and flight distance? How do you explain that relationship?
7. For this question, be sure you're handling missing values appropriately.
   a. What proportion of flights departed late?
   b. What proportion of flights arrived late?
   c. **What proportion of flights that departed late arrived late**?
8. a. Calculate the mean, median, and modal flight distance.
   b. Plot the distribution of flight distances and add lines for each of the three summary statistics. **Do this programmatically and ensure that the statistic represeneted by each line is clear to the viewer.**
   c. How many pairs of airports share the modal flight distance (don't worry about origin-dest

vs. dest-origin)? **Display them, with their number of count, in a nicely formatted table. The `xtable` package may be helpful here.**