Lab 03

Data wrangling

Shawn Santo

Announcements

- Complete the Qualtrics survey (available in Slack in channel `#general`). The information gathered will be used to form lab groups. Answer honestly.
- Continue working on Homework 02, ask questions, attend office hours
- Carefully review feedback given on past assignments

Goals

- Use data wrangling to extract meaning from data
- Practice using the **seven helpful verbs (functions)**

A grammar of data manipulation

Package `dplyr` is based on the concepts of functions as verbs that manipulate data frames (tibbles).

Common single data frame functions / verbs:

Function	Description	Operates on
`filter()`	pick rows matching criteria	rows
`slice()`	pick rows using indices	rows
`arrange()`	reorder rows	rows
`select()`	pick columns by name	columns
`mutate()`	add new variables	columns
`summarise()`	reduce variables to values	groups of rows

... many more.

`dplyr` rules

1. First argument is *always* a data frame
2. Subsequent arguments say what to do with that data frame
3. Almost always returns a data frame
4. Doesn't modify in place

Based on rules 1 and 3, it is natural to apply `%>%` in a sequence of `dplyr` functions for data wrangling purposes.

Pipes in R

The `%>%` is a forward-pipe operator. It allows you to pipe an object forward into a function.

You can think about the following sequence of actions - find keys, unlock car, start car, drive to school, park.

Expressed as a set of nested functions in R pseudo code this would look like:

```
park(drive(start_car(unlock_car(find("keys"))), to = "campus"))
```

Writing it out using pipes give it a more natural (and easier to read) structure:

```
find("keys") %>%
  unlock_car() %>%
  start_car() %>%
  drive(to = "campus") %>%
  park()
```

Lab 03

- Accept and create your private repository of the assignment at https://classroom.github.com/a/j6tXRehp
- The directions are available on the course website at http://www2.stat.duke.edu/courses/Spring21/sta199.003/labs/lab_03.html.
- This is an individual lab assignment. As you work on Lab 03
    - work with individuals in your breakout room,
    - ask questions,
    - don't be afraid to experiment in R, you can't break anything.
- Pay special attention to the submission procedure detailed at the end of the instructions (it is the same as Lab 02 and Homework 02).

Less commonly used `dplyr` functions

These are all single data frame functions.

Function	Description
`pull()`	grab a column as a vector
`transmute()`	create new data frame with variables
`distinct()`	filter for unique rows
`sample_n()` / `sample_frac()`	randomly sample rows

Additional `dplyr` resources

- `dplyr` cheat sheet
- dplyr vignette
- Chapter 5, R for Data Science