fall_2021_meetings

Ari Anisfeld

9/1/2021

Class 1: Reading files and 'dplyr'

Do now

Do now:

- Complete the intro poll at `bit.ly/acc_intro_poll`
- Let us know if you do not have R and RStudio installed!

After the poll

- Download `fall_lab_1` from the course webpage: `harris-coding-lab.github.io`.

Expectations

From you:

- do the work
- engage in course!
- have R and RStudio installed!

From us:

- prepare engaging lesson materials to teach you R basics
- address your questions about R
- help you gain self-confidence and self-sufficiency

From everyone:

- be nice to each other and create a growth-focused environment

Do the work

- Step 1. Videos
- Step 1a. Basics
- Step 2. Live sessions
- Step 3. Finish Labs

R basics

We cover:

- how to read in and manipulate data
- how to work with basic data structures
  - tibbles
  - vectors
- programmer logic
  - if statements
  - loops
  - functions

R we don't cover

Your homework! That's your professors' gift to you!

We won't cover **in depth**:

- most statistical tools
- advanced tools like webscrapping, package development and so forth
- some fundemental data skills
  - how to join data together
  - how to convert data from long to wide (pivoting)
  - how to deal with very messy data
  - how to work with specific data types (e.g. dates, advanced strings)

Address your questions about R

- Ed discussions (linked on canvas)
- Office hours (. . . ..)
- Here!

Tip: If you're stuck on your homework because of R, reframe your question as an R question.

And help you build self-sufficency

And help you build self-sufficency

- use ?
- test code in console. try to break it.
- ask teammates / try googling
- ask us! (Ed / Office hours / live now)

If it's not "mission critical", you can safely move on without full understanding. (Imagine learning a language and trying to figure out all the grammar and vocabulary at the same time!)

In this sessions

Walk away confident in reading in data and manipulating data with `dplyr` verbs.

- Warm-up
- Guided practice – setting up
- Independent or group practice
- comprehension check

Warm up

1. How would you install the `haven` package?

Warm up

2. In the videos, you learned about `head()`. What if you wanted to get the tail end of your data instead?

Warm up

3. Recall our `dplyr` verbs.

What is the purpose of each function?

- mutate()
- filter()
- select()
- arrange()
- summarize()

Soon we'll add:

- group_by()

Warm up

4. Imagine you have a data set, `df` with 4 variables, `county`, `year`, `income`, and `employment`. You only need the year and employment status of people whose income is below $5000. Which two `dplyr` commands do you need to do this? Can you write the code for this?
5. Remember the `mean()` function? What `dplyr` commands would we need if we want the average income? How many rows will the resulting dataset be?

Question: Can you explain pipes?

- Pipes %>% take the left hand side and put them into the first position on the right hand side.

```
storms %>% filter(year > 2010) %>% glimpse()

recent_storms <- filter(storms, year > 2010)
glimpse(recent_storms)
```

Notice

- `filter()` takes data in the first position and then an arbitrary number of filtering expressions.
- `glimpse()` takes data in the first position

Quick look at group_by()

Compare:

```
txhousing %>%
  summarize(total_volume = sum(volume, na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##   total_volume
##          <dbl>
## 1 858502159353
```

```
txhousing %>%
  group_by(year) %>%
  summarize(total_volume = sum(volume, na.rm=TRUE))
```

```
## # A tibble: 16 x 2
##     year total_volume
##    <int>        <dbl>
##  1  2000  33342410971
##  2  2001  35804815138
```

Lesson 0: Intro to R, RStudio and the `tidyverse`

- navigate and use Rstudio's features
  - particularly, the console, the text editor and help
- assign objects to names with <-
- use functions by providing inputs and learn more with ?
- `install.packages()` (once) and then load them with `library()` (each time you restart R)

Lesson 1: Key points: Reading files

- Tabular data is stored in a lot of different formats.
  - e.g. `.csv`, `.xlsx`, `.dta`
- Read tabular data of a given type with the proper function.
  - e.g. for `csvs` we have `read_csv()`
  - If you get a new type, Google "How to read xxx files into R tidyverse".
- We need to be aware of the file path and can `setwd()`.
- We know there are useful tools built into the `read_xxx()` functions.
  - Though we just scratched the surface.

Lesson 1: Manipulating data with `dplyr()`

- Choose columns with `select()`.
- Choose rows based on a match criteria with `filter()`.
  - We were introduced to comparison operators like == and `%in%`.
- Make new columns with `mutate()`.
- Sort data with `arrange()` and `arrange(desc())` or `arrange(-x)`.
- Create summary statistics with `summarize()`.