R for data science:

custom functions, iteration, and regex

John Little
Center for Data and Vizualization Sciences
Duke University Libraries
October 26, 2022

Links / urls for today

```
15:00
```

I will post the links in chat once we get started
The slides are in the GH repo

- **Preworkshop survey**: https://is.gd/79959b
- Code for today
  - **GH repo**: https://github.com/libjohn/workshop_rfun_iterate OR
  - import code from GH: https://github.com/libjohn/workshop_rfun_iterate.git OR
  - download code as zipped directory: https://github.com/libjohn/workshop_rfun_iterate/archive/refs/heads/main.zi OR
  - launch as a container in RStudio: See the *launch binder* button in the GH readme https://mybinder.org/v2/gh/libjohn/workshop_rfun_iterate/main?urlpath=rstudio
- YT_playlist_this_workshop https://www.youtube.com/watch?v=PrUnbYlC1kY&list=PLIUcX1JrVUNWW7RgPh9ysmJM3mBpIAlYG

Prerequisites

Per the note in the registration:

- familiarity with R/RStudio/Tidyverse

You may refer to earlier **Rfun** workshops (code, video, slides, data).
See: Rfun *quickStart* ★ https://rfun.library.duke.edu/portfolio/r_flipped

Land Acknowledgement

I would like to take a moment to honor the land in Durham, NC. Duke University sits on the ancestral lands of the Shakori, Eno and Catawba people. This institution of higher education is built on land stolen from those peoples. These tribes were here before the colonizers arrived. Additionally this land has borne witness to over 400 years of the enslavement, torture, and systematic mistreatment of African people and their descendants. Recognizing this history is an honest attempt to breakout beyond persistent patterns of colonization and to rewrite the erasure of Indigenous and Black peoples. There is value in acknowledging the history of our occupied spaces and places. I hope we can glimpse an understanding of these histories by recognizing the origins of collective journeys.

Outline

Case study demonstration of iterating with custom functions. The case-study goal is data cleaning that involves web scraping, interaction with the local file system (LFS), and general data wrangling.
In this case study, we will data-scrape and download about five excel workbooks files from a US Census web page (the Census pulse survey). Each workbook consists of approximately 51 worksheets, one for each state and DC.

Outline (continued)

- download data
- import the data into R
- wrangle, clean, and normalize data
- Make one big data frame
- subset the data for two states (i.e. two worksheets) from each workbook
- further subset the data for one category of data from the Census pulse survey
- generate ten faceted bar-graphs using ggplot2, and save those images back to the LFS
- along the way we'll use regex to find string patterns

Motivation

- **data wrangling** is 50-80 percent of any data analysis project
- **R is a functional**, data-first programming language (no FOR loops ; iteration via recursion)
  - “Functional Programming is an approach to replace iterative FOR loops – tidymodels book club
  - Rule of thumb: *Do anything more than three-times: compose a function*
  - The tidyverse approach prefers **tall** data formats and **data frames** (contrast with wide data frames or lists). Purrr iteration leverages this convention.

Motivation (continued)

Recursive Iteration takes some getting used to

- Most people have heard of **For loops** for controlling the flow of the language. Instead of a for loop, we're going to use {purrr} which uses the `map()` functions. (Similar to `lapply`, `mapply`, `sapply`)
- FOR LOOPS ARE FINE ; The Tidyverse way is easier
- **Tidyverse means** tall data in a data frame, looping is done by going 1-row-at-a-time over your **data frame**
- If you prefer base-R and the list data-type, you can still use all those functions. I will be focusing the iconic tidyverse approach (à la *pythonic* approach in python)

4 parts coding + 2 parts case study

1. Iterate with vectorized functions (`read_csv()`): import
    - Along the way we're gong to learn some lesser known *dplyr* functions and techniques that apply to many data wrangling needs.
    - `unite`, `pivot`, `separate`, `separate_rows`
2. Introduce the stringr package to leverage regular expressions, or regex. (finding patterns in strings)
3. Compose Custom functions; Introduce tidy evaluation, indirection, data masking, data variables and environment variables
4. Nesting list columns of data frames. Use `purrr::map` to apply custom functions to each row of a parent data frame
5. Devise a strategy to manipulate a single excel workbook
6. Map the procedures to a set of Census survey data in multiple Excel workbook files.

Learing in this workshop

Aim questions at presented material. Schedule me for one-on-one consultations.
The best way to learn R it to take simple atomic problems with data that you know, and analyses that are familiar to you, then replicate the analysis in R. But, doing these things with familiar data is not always possible in a big diverse group. I invite you to schedule me for consultations.

Post workshop survey

https://forms.gle/MmrzadXkq5TMHBqv5
.bg-washed-blue.b–navy.ba.bw2.br3.shadow-5.ph4.mt5[

Rfun

John R Little

Data Science Librarian
Center for Data & Visualization Sciences
Duke University Libraries
https://johnlittle.info
https://Rfun.library.duke.edu
https://library.duke.edu/data