R Working Group

Data Management in R using `dplyr`

Jason Thomas
thomas.3912@osu.edu

Friday, February 17th, 2023
11am - 12:30pm

Motivation

- R – past 5 years (or so) has included great (user-friendly) strides in data management
  - `dplyr package` – "grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges"
  - a lot of functions coded in C++ (for speed)
  - many other packages make use of `dplyr`
- Goal for today is to establish basic fluency with this grammar

Background

- `dplyr` is part of `tidyverse`
    - `ggplot2`, `forcats`, `tibble`, `readr`, `stringr`, `tidyr`, `purrr`
    - may also want to check out `tidycensus`[https://walker-data.com/tidycensus/articles/basic-usage.html]
- `dplyr` logic: "By constraining your options, it helps you think about your data manipulation challenges."
    - 5 commands will take you a long way
    - readability and simplifying code (with pipes)
- **R for Data Science**
    - https://r4ds.had.co.nz/

Grammar

- **Rows** – selecting and organizing observations/cases
    - **Groups of Rows**
- **Columns** – cleaning and creating new variables
- **Merging Data** (not going to go into this today)
- Additional features of the language
    - tibble data structure (similar to data frames, but crankier)
    - "pipe" %>% for stringing multiple commands together
    - useful for keeping number of objects to a minimum and for plotting (e.g., adding separate symbols for subgroups)

Grammar for Rows

- `filter()`
- `slice()`
- `arrange()`
- *Groups of rows*: `summarize()`
  - useful companion: group_by()
  - collapse across rows with `summarize()`

Grammar for Columns

- `select()`
- `rename()`
- `mutate()`