R for Computational Sciences

Part 1. getting started, EDA, data wrangling

John Little

Center for Data & Vizualization Sciences

R for computational sciences

getting started, EDA, data wrangling

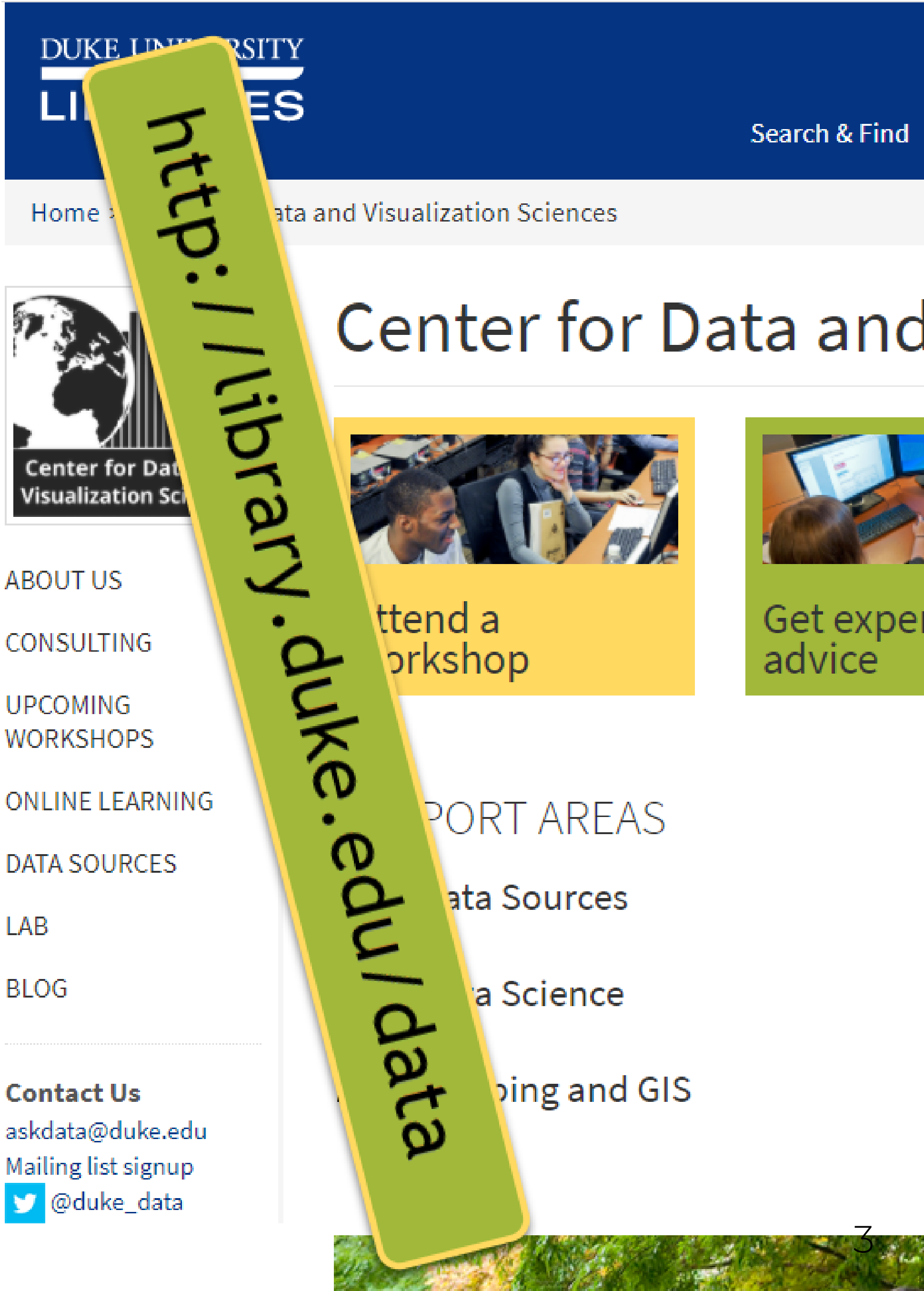
Flipped Workshop

Sept. 15, 1pm to 3pm

15:00

Whoami

John Little
Data Science Librarian
Host of **Rfun.library.duke.edu**
Center for Data & Visualization Sciences

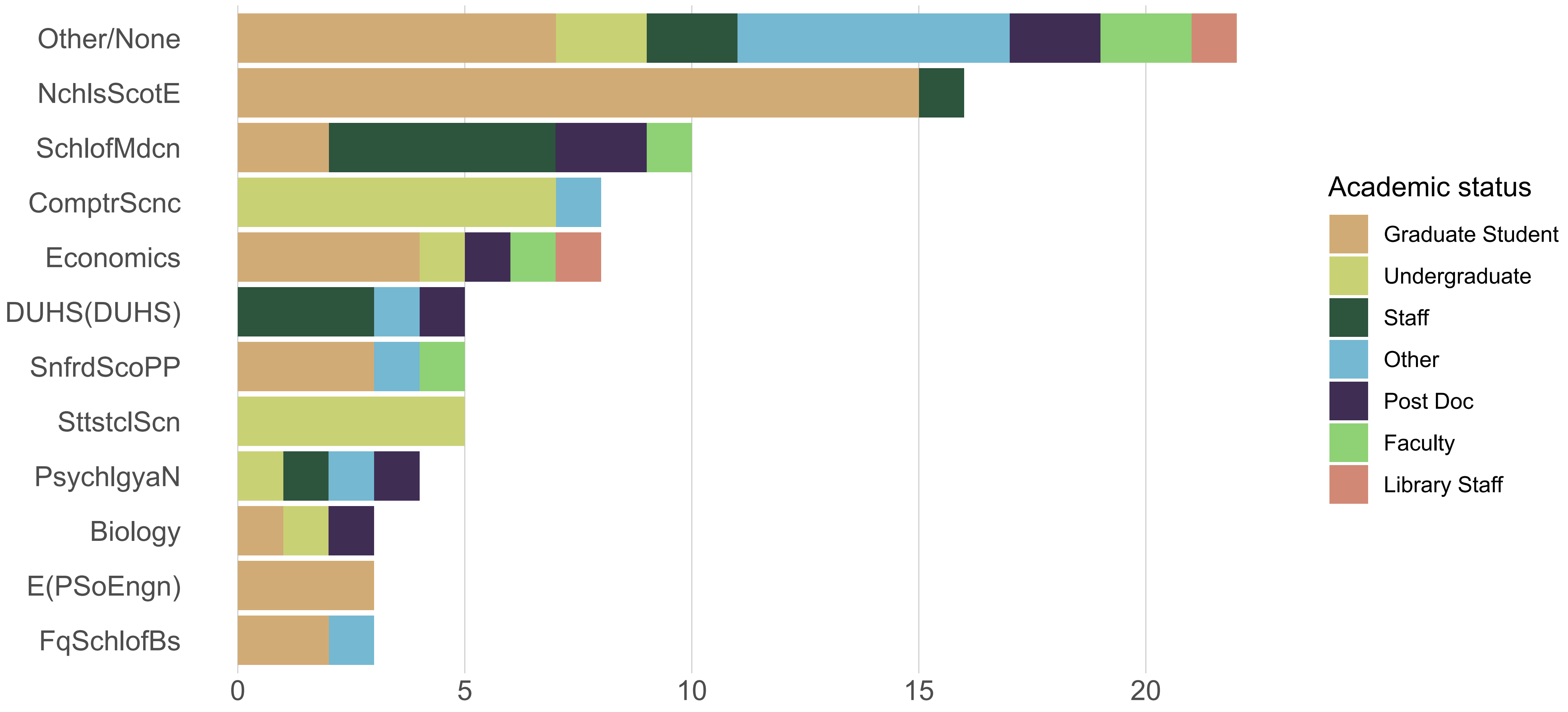

http://library.duke.edu/data

Duke University: 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.

Attendance by Discipline

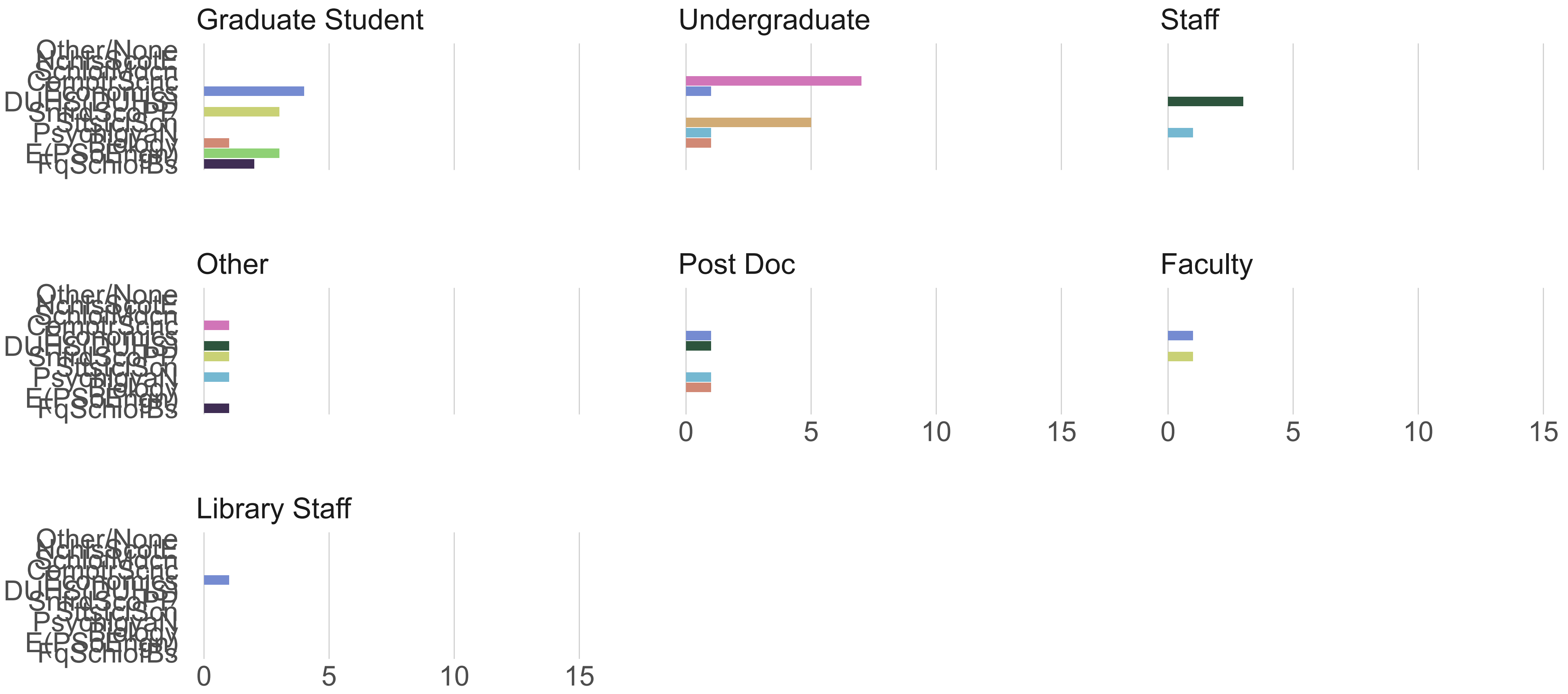
R for data science: getting started, EDA, data wrangling



Source: CDVS Workshop Registration

Attendance by Discipline

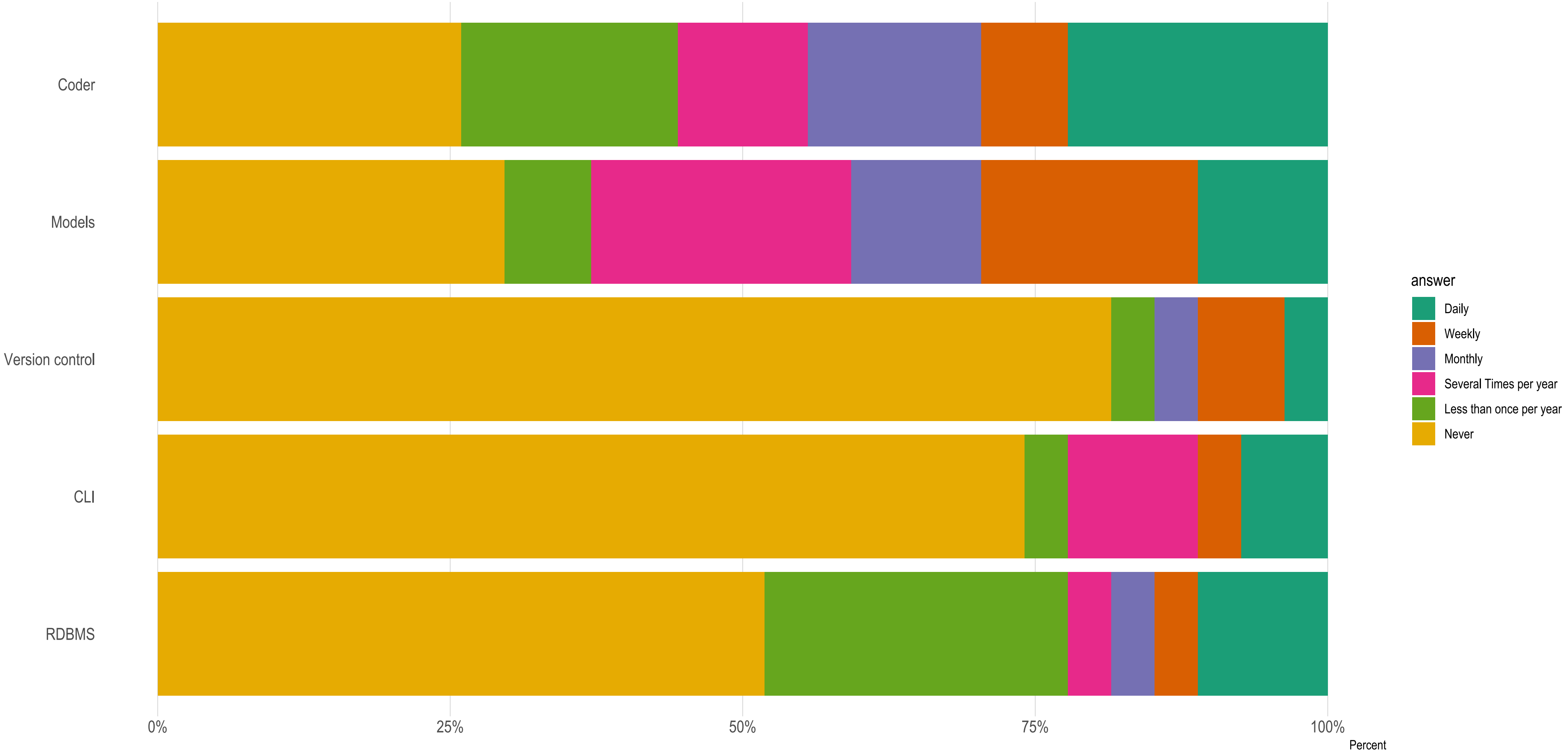
R for data science: getting started, EDA, data wrangling



Source: CDVS Workshop Registration

Self-reported tool usage

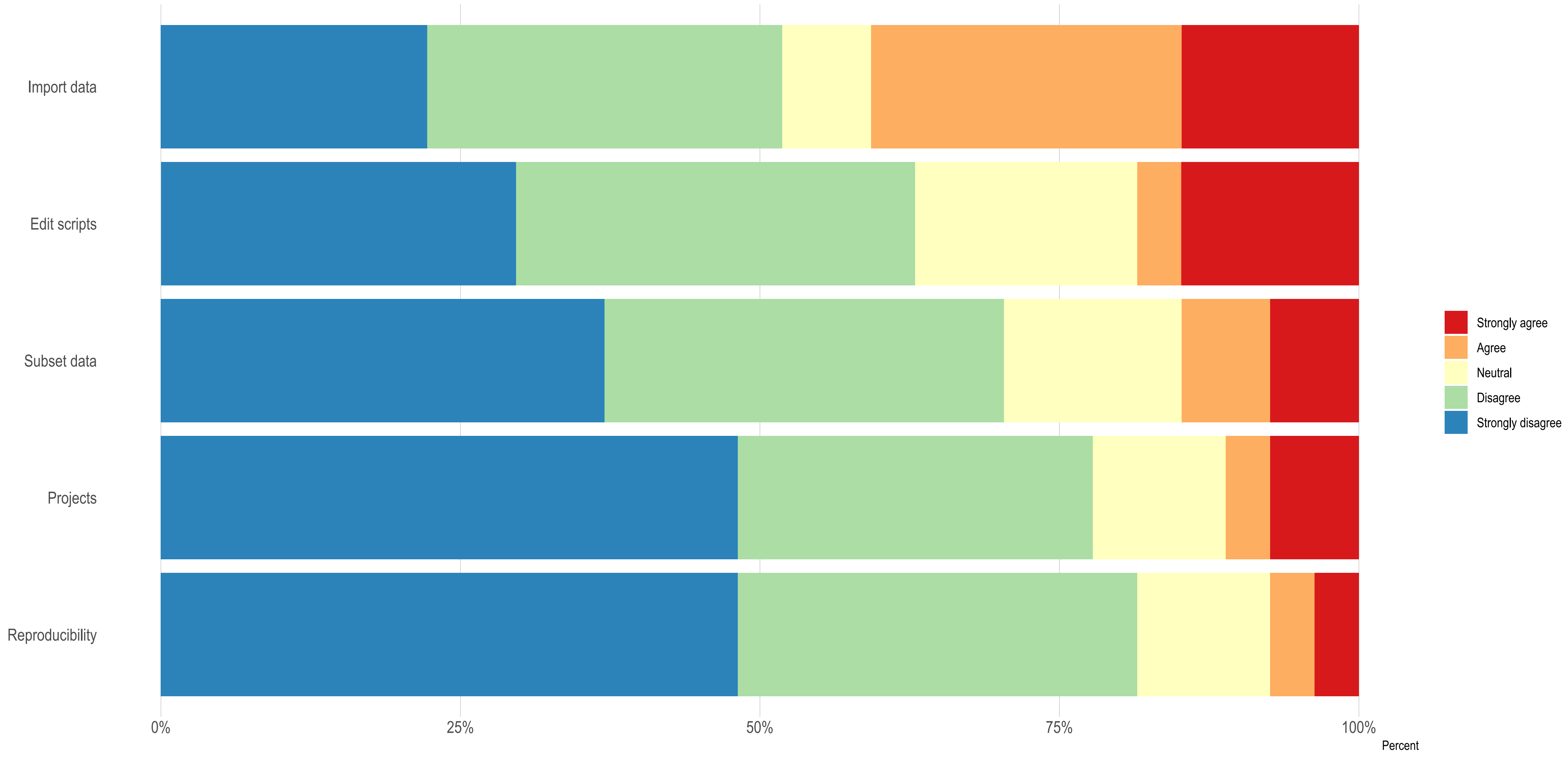
Respondent's use of a tool / technology / technique



Source: CDVS Pre-workshop survey

Self-reported R skills

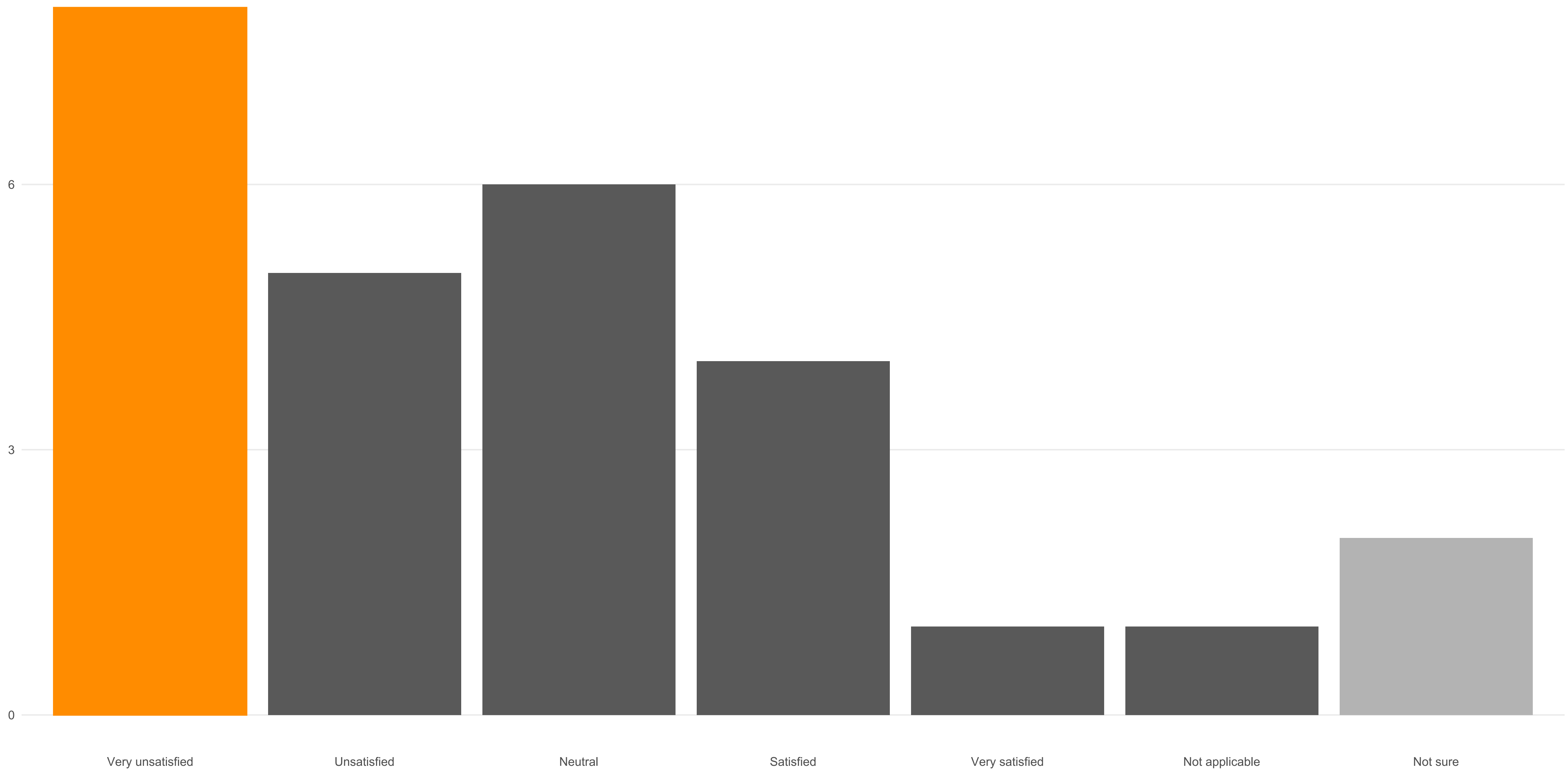
Respondents feel capable of completing a Tidyverse task



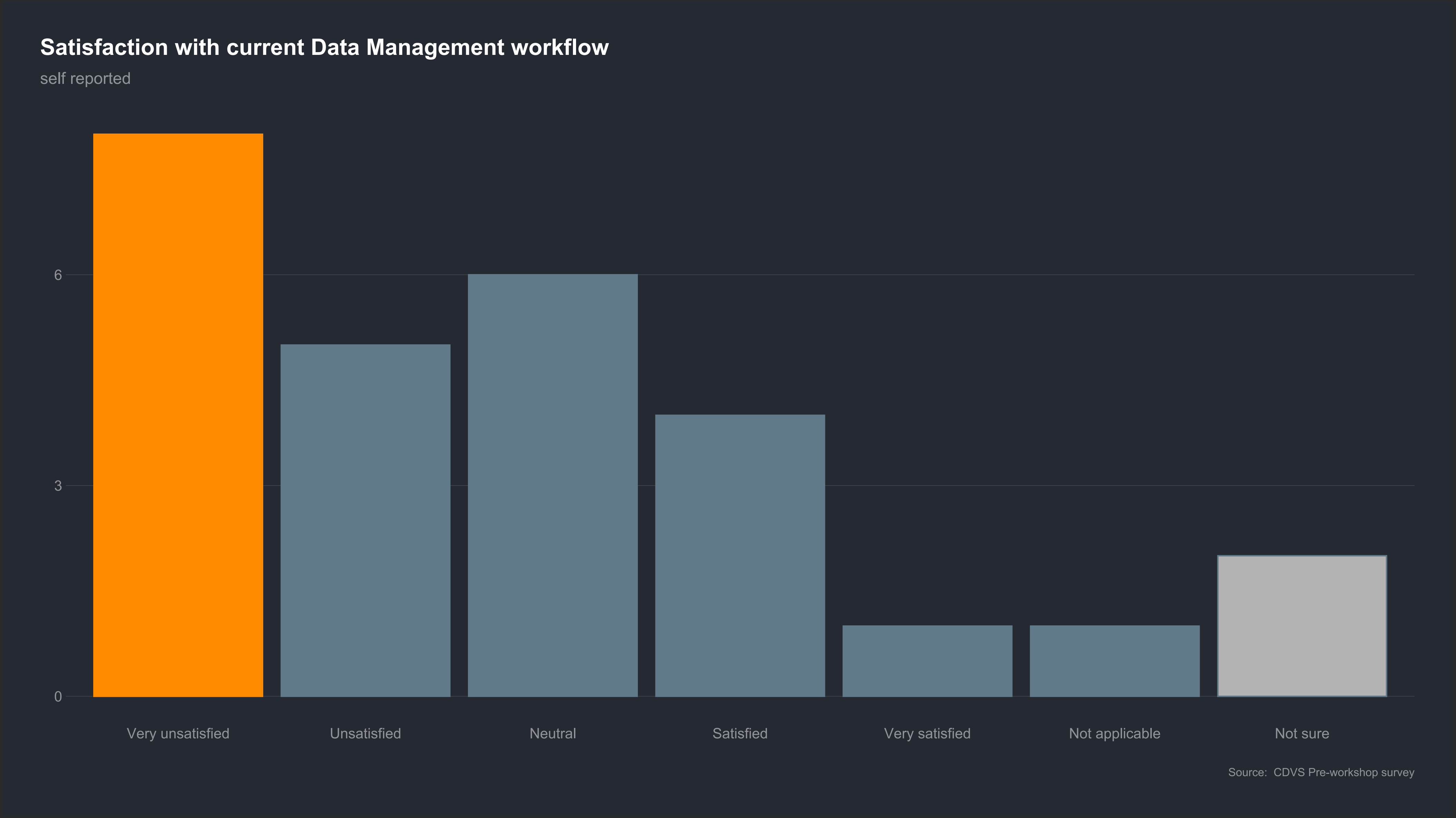
Source: CDVS Pre-workshop survey

Satisfaction with current Data Management workflow

self reported



Source: CDVS Pre-workshop survey



Satisfaction with current Data Management workflow

self reported

Source: CDVS Pre-workshop survey

Consulting and Assistance

Title	URL
Schedule me for consultations	https://is.gd/littleconsult
Consulting & AskData@Duke.edu	https://library.duke.edu/data/consulting

We're happy to consult with you

In a consultation, let us help make the details of R relevant to your project

Resources

It's all online

Title	URL
Code for this workshop	https://github.com/libjohn/rfun_flipped
exercises	https://github.com/libjohn/intro2r_exercises
Rfun	https://rfun.library.duke.edu
Center for data & Viz	https://library.duke.edu/data

Reprex

The most efficient way to get help

REPRoducible EXample and Code

https://reprex.tidyverse.org

Use the smallest, simplest, most built-in data possible

Include commands on a strict “need to run” basis

Pipes and Assignment

A couple things to remember…

Assignment

Give an object name particular value

```
<-
```

"gets value from"

```
answer <- 5 * 5

mutate(answer2 = answer * 2)
```

Keyboard shortcut for `<-` is *alt-dash*

Pipe

Chain functions together (a tidyverse or magrittr conjunction)

`%>%` or `|>`

"and then"

```
answer %>% sqrt()
```

Keyboard shortcut: *Ctrl/Cmd-Shift-M*

Definitions

R is a **data-first programming language** with mature sense of the *data life-cycle* and reproducibility

R - programming language / language interpreter
RStudio - an IDE or Integrated Development Environment
Tidyverse - a coherent and opinionated system of packages for data manipulation, exploration, and visualization

Definitions

Tidy data - a foundational concept governing the shape of your data. https://vita.had.co.nz

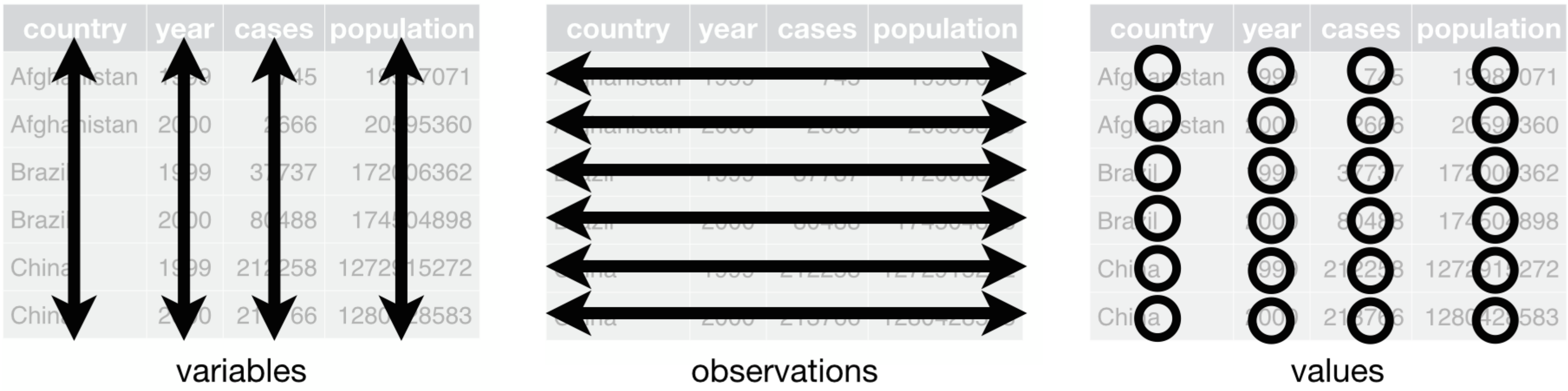


Image Credit: *R for Data Science*

Outline

Reproducibility

RStudio projects

Literate coding / Coding Notebooks

5 dplyr verbs

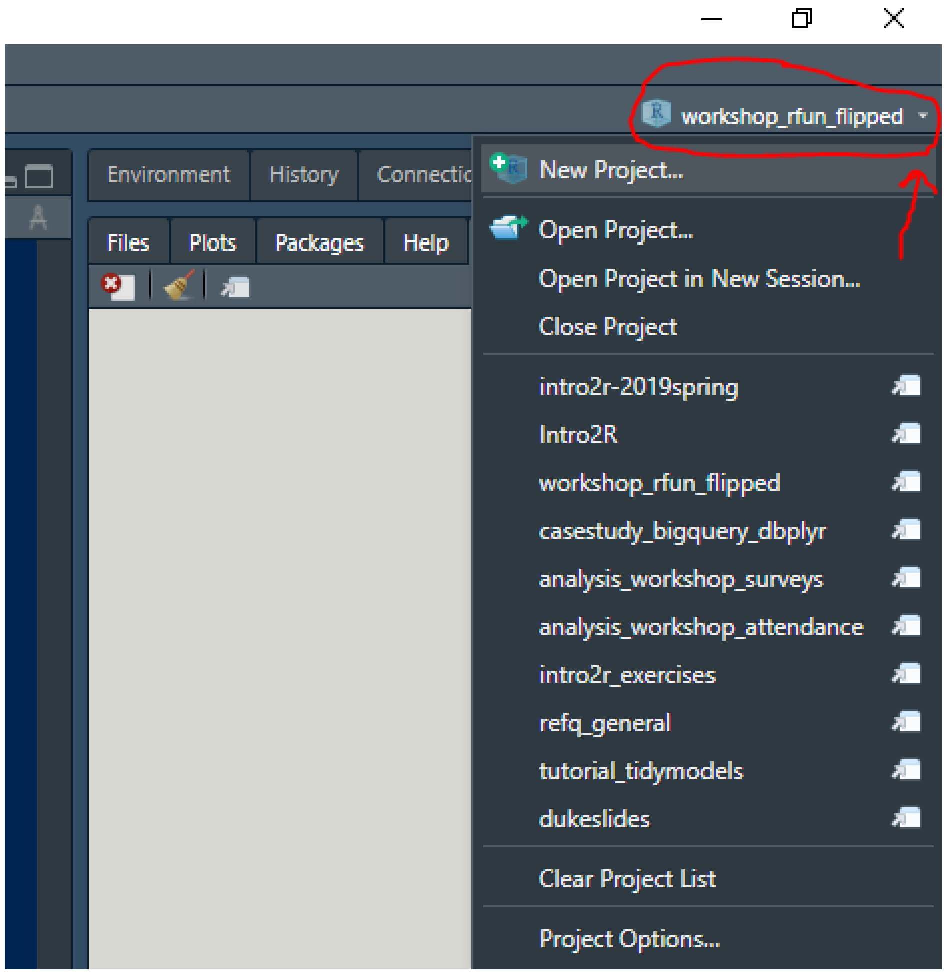
Reproducibility

Obtaining computational results using the same input data, computational steps, methods, code, and conditions of analysis

RStudio projects

Managing each project in a discrete directory that can be easily shared with others. That is, your projects can work on other computers without rewriting the code.

- Enables the use of relative paths instead of `setwd()`
- Using R Markdown to Restart R and run all chunks instead of `rm(lists = ls())`
- Integrates with version control (e.g. *Git*)



Literate coding

Integrate and intersperse prose with code. Explain your analysis with natural language. Ideally, render various outputs from the same code-prose document

R Markdown & Jupyter notebooks are an example of literate coding

Why

Using *reproducibility* and *literate code* techniques within *RStudio projects* + *version control* enables better workflows; workflows that are not dependent on cut & paste mousing

Today, we'll use .Rmd files to render R Markdown notebooks

`dplyr`

A grammar of data manipulation

- consistent verbs to solve common data transforms
- https://dplyr.tidyverse.org/

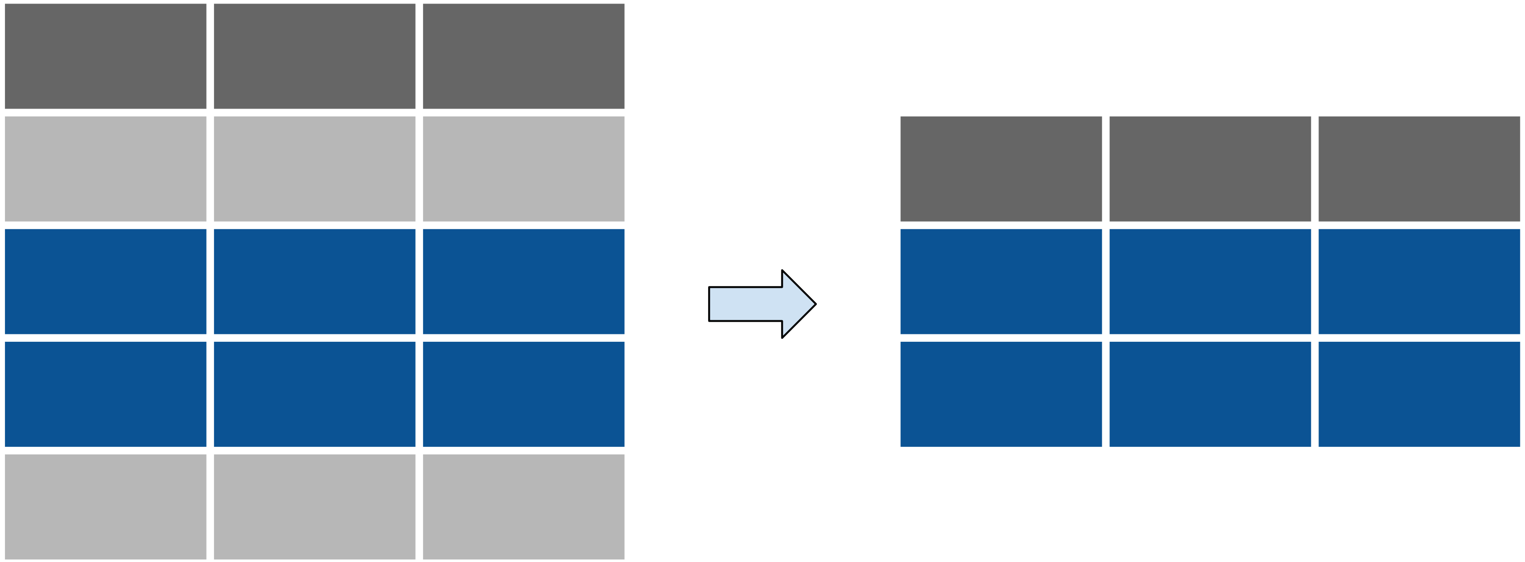
Five `dplyr` Verbs

Function	Usage
`filter`	subset rows
`select`	subset columns
`arrange`	sort rows by variables
`mutate`	change cell or variable values
`count`	
`summarize`	powerful when used with `group_by()`

There are many more dplyr functions

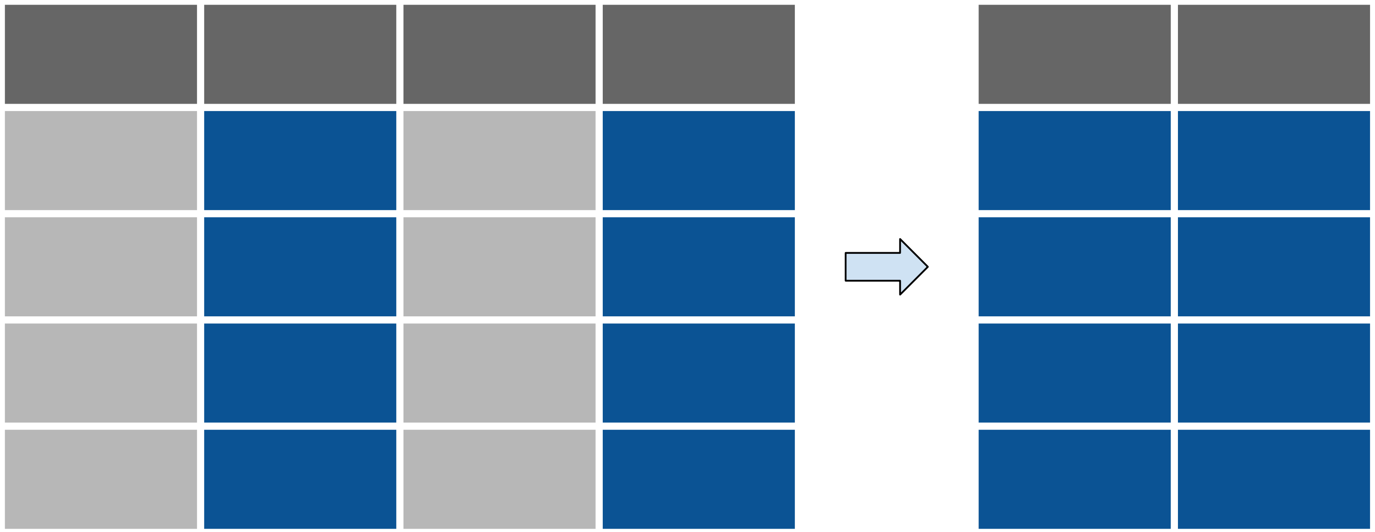
`filter` Subset **Rows** by variables

```
starwars %>% filter(eye_color == "orange")
```



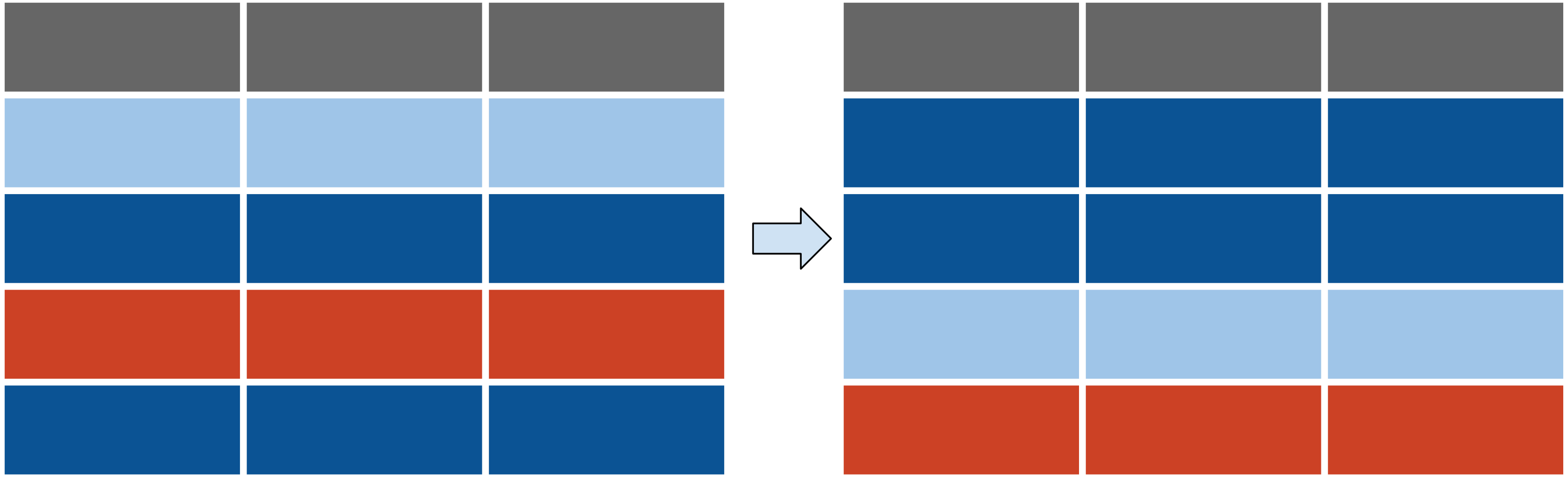
`select` subset by columns (variables)

```
starwars %>% select(hair_color, eye_color)
starwars %>% select(2:4)
starwars %>% select(name:mass, 10, 7, 4:6)
```



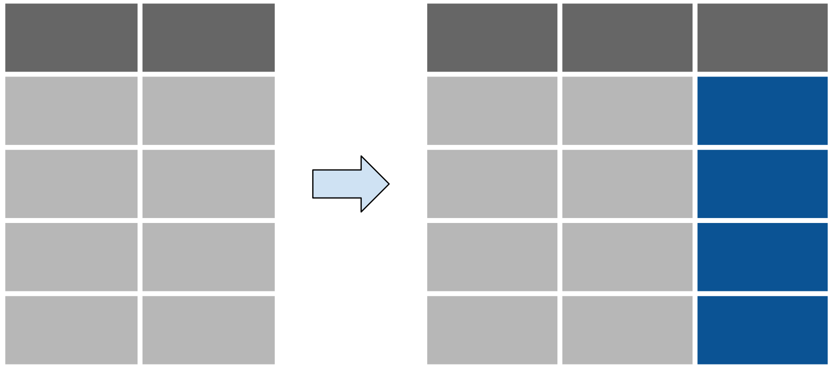
`arrange` Sort **Rows** by variables

```
starwars %>% arrange(eye_color)
starwars %>% arrange(desc(eye_color))
starwars %>% arrange(desc(eye_color), hair_color)
```



`mutate` Change cell values

```
starwars %>% mutate(big_mass = mass * 100)
starwars %>% mutate(BMI = (mass / (height/100)^2))
starwars %>% mutate(
       nickname = str_c("Big", str_to_upper(hair_color),
                        sep = " "))
```



`count` *Count observations by group*

```
starwars %>% count(gender)
```

`summarize` *Reduce multiple values down to a single line*

```
starwars %>%
  drop_na(height) %>%
  summarise(n(), n_distinct(height), min(height), max(height))

starwars %>%
  drop_na(height) %>%
  group_by(gender) %>%
  summarise(Total = n(), n_distinct(height), min(height))
```

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