Teaching Data Cleaning and Wrangling with R's `data.table` Package

Erin Franke Sara Colando
Carnegie Mellon University, Department of Statistics & Data Science

Background

`data.table` is a powerful R package for **computationally efficient** data wrangling with **concise syntax** and **minimal dependencies**.

We devised materials for **teaching introductory** data cleaning and wrangling with **`data.table`**, which include:

1. Lecture slides with visualizations and poll questions on the **`data.table`** syntax for the six main verbs of data wrangling (*select, filter, mutate, arrange, group by, and summarize*)
2. A lab with try-it-yourself activities

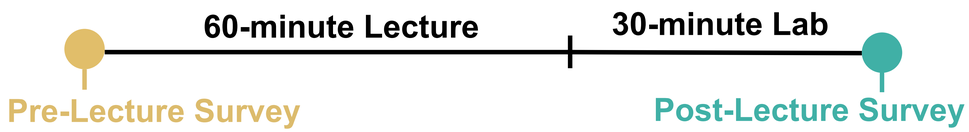
Methods and Assessing Efficacy

We led a 90-minute session to students in Carnegie Mellon's Summer Undergraduate Research Experience (SURE) 2025 program.

Of the 25 students who completed both surveys:

- **72%** are rising seniors, **28%** are rising juniors
- **44%** attend large institutions, **24%** attend mid-size institutions, **32%** attend small institutions
- **52%** are majoring or minoring in statistics or data science

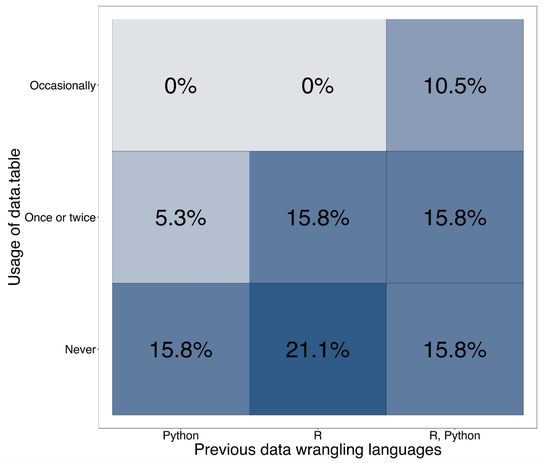
Session Timeline:



Most Students Had Not Used `data.table`

- **20%** of students had never used R before SURE
- **40%** of students had never heard of **`data.table`** before our session
- **76%** of students had wrangled data before

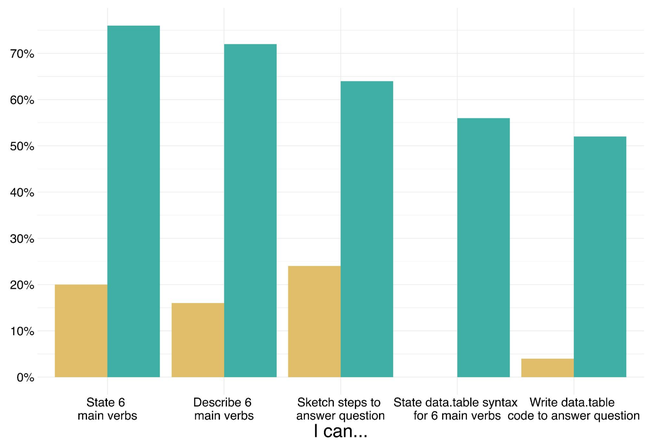
Of those that had wrangled data before:



The Learning Objectives

Level 1	- **State** and **describe** the six main verbs of data wrangling - **State** the general syntax for the six main verbs with **`data.table`**
Level 2	- **Implement** the **`data.table`** syntax for the six main verbs - **Sketch** the steps needed to answer a substantive question with the six main verbs
Level 3	- **Devise** a data wrangling pipeline using **`data.table`** syntax based on the previously sketched steps

Students' Self-Efficacy Improved Across All 5 Learning Objectives



The proportion of students who felt **at least fairly confident** increased from pre-lecture to post-lecture across all 5 learning objectives.

Post-lecture, students indicated they'd use **`data.table`** for the following tasks:

- **87%**: big data analysis
- **65.2%**: everyday data wrangling
- **60.9%**: special features (e.g. grouped models, visualizations, **`dtplyr`**)
- **30.4%**: **`fread`**, **`fwrite`** functions

Key Takeaways

- Students from diverse coding backgrounds improved across all 5 learning objectives after the lecture and lab session.
- Different features of data.table resonated with different students, with **96%** saying they would use **`data.table`** again in some capacity.
- Providing students with more practice would likely help them achieve higher self-efficacy in level 2 and 3 learning objectives.

Acknowledgements: this work was funded by a travel ambassador grant as part of data.table's NSF POSE award (Award Number 2303612) and conducted under the Eberly Center's Umbrella IRB. We thank Allison Connell Pensky and Alex Reinhart for their guidance in preparing the teaching materials, as well as Quang Nguyen, Ron Yurko, and the SURE 2025 students for allowing us to assess the efficacy of our teaching materials.