Utilizing Open Source Resources to Teach Introductory Data Science

useR! 2022

Dr. Tyler George

Cornell College | Dept. of Mathematics and Statistics

June 22nd, 2022

About Cornell College

- Small liberal arts college of about 1000 students
- One-Course-at-A-Time block schedule
  - Each class occurs over 3.5 weeks including 18 days of 4 hours of instructional time and three weekends
- New majors in Applied Statistics and Data Science

Course Objectives

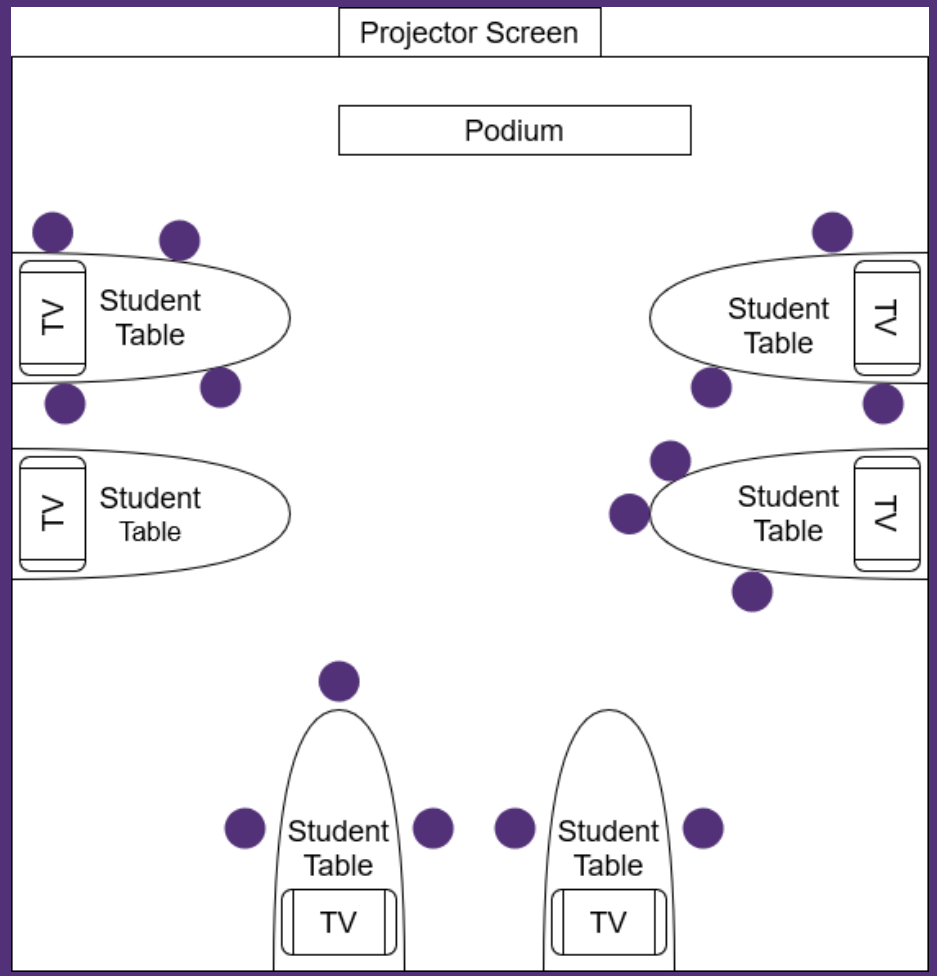
- From the syllabus "respect, explore, understand, and utilize data in a way that is reproducible using version control."
- Data wrangling
- Utilizing version control & good organization to facilitate reproducible work
- Building technical skills in R, and RStudio, namely *tidyverse* functions
- Understanding and learning to consider ethical problems in data science
- Creating understandable and honest data visualizations
- Communicating statistical analysis

Utilized Resources

- Data Science in a Box by Mine Çetinkaya-Rundel (2021)
- Data science ethics assignments from the Quantitative Analysis Institute at Wellesley College (Pattanayak, Gan, Li, Liang, and Wong, 2022)
- Workshops by Julia Silge available at https://juliasilge.github.io/tidytext-tutorial/site/
- Many R packages were use throughout the class but ghclass was particularly useful for teaching/running the course

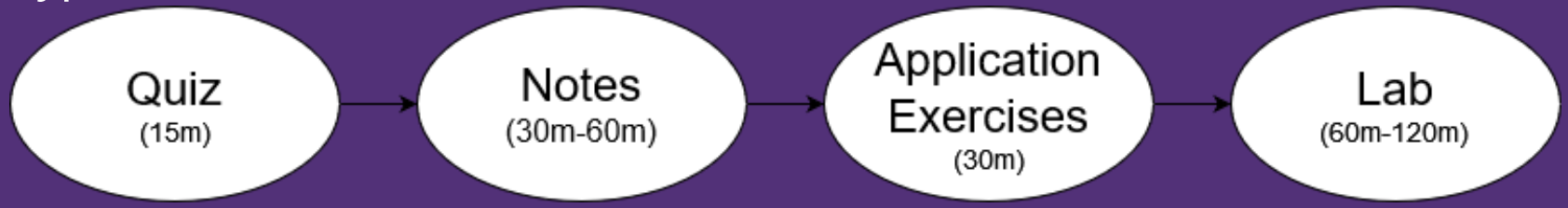
The Class

- This course has a pre-requisite of *either* introductory statistics or introductory computer science (python)
- Total of 15 students



A Day in the Class

- Each day consists of 2, 2 hour sessions.
- Typical class flow:



- After the afternoon class each day students were assigned:
  - One R homework,
  - Read 1-3 chapters of content from sources such as R for Data Science
  - Read ethics related materials

Course Schedule

Day	Topics Covered
Day 1	Meeting R and GitHub (S1-S3,AE1a,L1)
Day 2	Data Visualization (S3-S6,AE2,AE3)
Day 3	Tidy data and Data Wrangling (S7-S11,L2)
Day 4	Data Types and Classes (S12,S13,AE5,L3)
Day 5	Project Meetings
Day 6	Importing and Recording Data (S14,S15,AE6)
Day 7	Effective Visualizations (S16,L4,AE7)
Day 8	Statistical Confounding and Doing Data Science (S17-S19,L5)
Day 9	Web Scraping, Functions, and Iteration (S20-S24,AE8,L8)
Note:	
In reference to Data Science in a Box with S = Slides, AE = Application Exercises, L = Labs	

Course Schedule

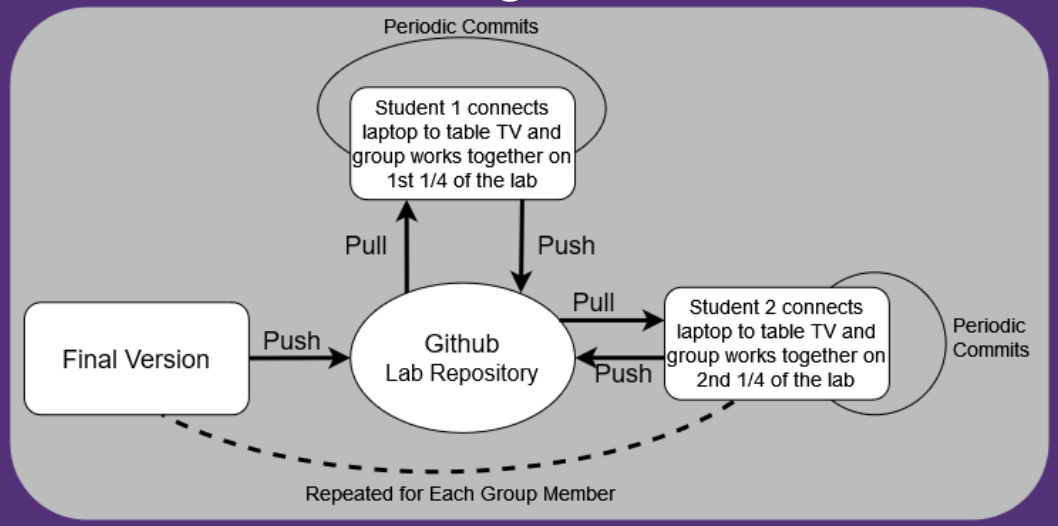
Day	Topics Covered
Day 10	Project Meetings and Midterm
Day 11	Data Ethics (S25-S27,Videos,L9)
Day 12	Linear Modeling (S27,S28,L10)
Day 13	Nonlinear Models and Multiple Regression (S29,S30,L11)
Day 14	Prediction, Feature engineering, Cross Validation (S33-S35,L11)
Day 15	Interactive Web Apps (S42-S45, RStudio Shiny Tutorial (RStudio 2021))
Day 16	Text Analysis (Modification of S40,S41 and tidytext workshop (Silge, 2021)
Day 17	Project Presentations
Day 18	Final Exam
Note:	
In reference to Data Science in a Box with S = Slides, AE = Application Exercises, L = Labs	

Technologies

- Learning Management System (LMS) - Moodle
- RStudio Server (new implementation for this course, see Çetinkaya-Rundel and Rundel (2018) for a discussion on infrastructure)
  - A combination of RMarkdown and R scripts were used
- Git and Github.
  - Students used Git through the RStudio panel within an R project
  - I also used Github desktop application for some convenience such as pulling one students current work to help them
- R packages - primarily within the *tidyverse*
  - *ghclass* for distributing and collecting repositories (Rundel and Çetinkaya-Rundel, 2022).

Teaching Pedagogies

- Collaborative learning (Roseth, Garfield, and Ben-Zvi, 2008)



Note: This was deviated from for projects and examples to simulate merge conflicts

- Using tidyverse (Çetinkaya-Rundel, Hardin, Baumer, McNamara, Horton, and Rundel, 2021)
- Version control to lead to reproducible work (Beckman, Çetinkaya-Rundel, Horton, Rundel, Sullivan, and Tackett, 2021)

Learning the Bare Necessities

- I was new to Git, Github, RStudio Server (Linux), and a handful of the tidyverse packages and functions
- What did I *need* to learn to teach this class?
  - Git and Github: Basic Git functions such as push,pulling, and handling merge conflicts in RStudio
  - *ghclass* R package made much of the Git/Github learning much smoother. See the R script I created here
  - RStudio server implementation and management required learning some basic Linux commands. See here for the commands I ended up needing
  - Learning the Xaringan R package to modify slides (now I'm using it for this!)
  - I converted the notes to pdf's. See my simple script here. I will host these using Github pages next time.

Student Comments

- "I enjoy working with my group on the project, although it did take countless hours outside of the classroom I enjoyed it a lot and felt accomplished after presenting."
- "Working on labs and application exercises were beneficial for me. After completing those I could refer back to them which helped a lot. When working alone I often would get stuck on a exercise but when working in groups we were able to put our heads together and solve the problems. I really enjoyed that."
- "There were far too many notes in this class."
- "Going over the notes are necessary but I thought the length of the slides were way too long. It was hard to stay engaged the whole time when going over the notes because I was not using the code in r I was just seeing the results on the powerpoint. I think the notes could be trimmed down some to free up more space for application exercises."

Future Updates

- Expanding the ethics discussion in the course
  - I will be working with a Cornell College faculty in Philosophy this summer.
  - Using some modules from (Baumer, Garcia, Kim, Kinnaird, and Ott, 2022)
- Trying out gradetools to make grading quick while giving useful feedback to students (Ricci, Medina, and Dogucu, 2022)
- Less statistical inference and more data wrangling
- Higher project expectations
- Github actions
- Rethinking of exams. The estimated time required was way off.

Acknowledgments

Dr. Ajit Chavan, Assistant Professor of Computer Science at Cornell College, for setting up the cluster used to run the RStudio Server and answering many of my questions about its use.

References

Baumer, B. S., R. L. Garcia, A. Y. Kim, et al. (2022). "Integrating Data Science Ethics Into an Undergraduate Major: A Case Study". In: *Journal of Statistics and Data Science Education* 30.1, pp. 15-28. DOI: 10.1080/26939169.2022.2038041. eprint: https://www.tandfonline.com/doi/pdf/10.1080/26939169.2022.2038041 . URL: https://www.tandfonline.com/doi/abs/10.1080/26939169.2022.2038041 .

Beckman, M. D., M. Çetinkaya-Rundel, N. J. Horton, et al. (2021). "Implementing Version Control with Git and GitHub as a Learning Objective in Statistics and Data Science Courses". In: *Journal of Statistics and Data Science Education* 29.sup1, pp. S132-S144. DOI: 10.1080/10691898.2020.1848485. eprint: https://doi.org/10.1080/10691898.2020.1848485 . URL: https://doi.org/10.1080/10691898.2020.1848485 .

Çetinkaya-Rundel, M. (2021). *Data Science in a Box*. available at https://www.datasciencebox.org.

References

Çetinkaya-Rundel, M., J. Hardin, B. S. Baumer, et al. (2021). *An educator's perspective of the tidyverse*. DOI: 10.48550/ARXIV.2108.03510. URL: https://arxiv.org/abs/2108.03510.

Çetinkaya-Rundel, M. and C. Rundel (2018). "Infrastructure and tools for teaching computing throughout the statistical curriculum". In: *The American Statistician* 72.1, pp. 56-65. DOI: 10.1080/00031305.2017.1397549. eprint: https://www.tandfonline.com/doi/abs/10.1080/00031305.2017.1397549. URL: https://www.tandfonline.com/doi/abs/10.1080/00031305.2017.1397549.

Pattanayak, C., D. Gan, T. Li, et al. (2022). *Quantitative Analysis Institute Online Resources*. Wellesley College. https://sites.google.com/wellesley.edu/qai-online-resources/home?authuser=0.

References

Ricci, F. Z., C. Medina, and M. Dogucu (2022). *gradetools: Tools to Assist with Providing Grades and Personalized Feedback to Students*. https://federicazoe.github.io/gradetools/, https://github.com/federicazoe/gradetools/.

Roseth, C. J., J. B. Garfield, and D. Ben-Zvi (2008). "Collaboration in Learning and Teaching Statistics". In: *Journal of Statistics Education* 16.1, p. null. DOI: 10.1080/10691898.2008.11889557. eprint: https://doi.org/10.1080/10691898.2008.11889557 . URL: https://doi.org/10.1080/10691898.2008.11889557.

RStudio (2020). *Learn Shiny*. https://shiny.rstudio.com/tutorial/.

Rundel, C. and M. Çetinkaya-Rundel (2022). *ghclass: Tools for Managing Classes on GitHub*. R package version 0.2.1, available at https://github.com/rundel/ghclass.

References

Silge, J. (2021). *Text Mining Workshop*. Available at https://juliasilge.github.io/tidytext-tutorial/site/.

Thank You!