MSDS 16:954:597

Data Wrangling and Husbandry

Spring 2020

Instructor:	Prof. Jason M. Klusowski	**Time:**	Monday 6:40–9:30 PM
Email:	jason.klusowski@rutgers.edu	**Place:**	ARC-105

Course Page: All course materials, including class notes, homework, projects, and a copy of this syllabus (which may be periodically updated), will be posted on Canvas.

- https://rutgers.instructure.com/courses/46291/

Office Hours: Hill Center, Room 492. Monday 5:00–6:00 PM, or by appointment.

Teaching Assistants: Ziyue (Jeff) Wang, Email: ziyue.wang@rutgers.edu & Penghui Fu, Email: penghui.fu@rutgers.edu.

Teaching Assistant Office Hours: Thursday 5:00–6:00 PM.

Main References: Below are the required textbooks we will use throughout the course. You will need to consult them frequently. Note that the first text is available for free online, while the second is available online to Rutgers students (you need a valid netID to gain access to the electronic copy).

- Garrett Grolemund & Hadley Wickham, *`R` for Data Science*, O'Reilly.
  - Available online: http://r4ds.had.co.nz/
- Bradley C. Boehmke, *Data Wrangling with `R`*, Springer.
  - Available online: https://catalog-libraries-rutgers-edu.proxy.libraries.rutgers.edu/vufind/Record/5725290/

Objectives: This course provides an introduction to the principles and tools to retrieve, "tidy", clean, and visualize data in preparation for statistical analysis. Principles of reproducibility and reusability are emphasized. It teaches *husbandry* techniques for data *wrangling* and exploration. The emphasis is on preparation of data to ease the analysis rather than sophisticated analyses. Topics include methods to convert data from diverse sources into suitable form for data visualization and analysis; methods to scrape data from websites; data visualization; elementary database operations; principles of reproducibility and reuseability.

- *Wrangle*. To round up, herd, or take charge of (livestock): the horses were wrangled early.

- *Husbandry.* 1. the care, cultivation, and breeding of crops and animals: crop husbandry. 2. management and conservation of resources.

Prerequisites: A basic understanding of statistics, e.g., correlation analysis and linear regression. Some programming experience will be helpful but is not necessary.

Course Outline (Tentative):

1. Introduction to the course, `R`, and `RStudio`
2. Report writing using `R Markdown`
3. Basic data management in `R`
4. Introductory data visualization
5. Data manipulation and the split-apply-combine paradigm
6. String manipulation
7. Text analysis
8. Writing `R` functions
9. Project organization, including the use of `github`
10. Getting data off the web
11. Basics of writing `R` packages
12. Reproducible results, task automation, and automated analytical pipelines

Important Dates:

- There will be no class on Monday, March 16 for Spring Recess.
- Take-home Midterm Exam (Tentative) ..... March 9-10, 2020.

Course Policy:

- Regular attendance is essential and expected.
- All homework and projects must be done individually and not in groups. Discussion with others is permitted (and even encouraged!), but you must write your own work in your own words.

Grading Policy:

- Homework (40%), Take-home Midterm (30%), and Final Project (30%).

Academic Integrity:

- Please consult Rutger's policies on academic conduct: `http://academicintegrity.rutgers.edu/`