Principles of Data Science

PSYC 259

Winter 2021

Prof. John Franchak
E-mail: franchak@ucr.edu

Office Hours: By Appointment
Office Hour Zoom Link

TA: Jacob Elder
E-mail: jacob.elder@email.ucr.edu

Office Hours: T 11am-12pm
Office Hour Zoom Link

Class Time: W 1:30-4:20 PM
Class Zoom Link
Course Materials

Course Description

Most quantitative courses (importantly) focus on the final steps of data analysis—conducting and understanding statistical tests. However, much of the work in data science is taking raw data, often from multiple, incompatible sources, and processing those data into a usable form. This course will emphasize the importance of robust, documented, and automated workflows for processing data to save time, reduce errors, improve reproducibility, and facilitate collaboration among multiple researchers. We will also spend time on data visualization and communication—an important part of creating, checking, and collaborating on data workflows. We will use the R programming language, Github, and Rmarkdown to work through examples, but the focus is on concepts/best practices that can be applied to any software or programming language. The course is open to students who have little programming experience or experience with R. The goal is for students at all levels of programming experience to set goals to improve their data science skills.

Course Objectives

The goals of this course are for you to critically analyze and improve your data analysis workflows. Implementing robust, automated procedures for handling data will allow you to:

- Foster open science through increased transparency, reproducibility, and easier data sharing
- Increase the fidelity of your data and analyses by detecting and preventing errors
- Better understand and communicate about your data
- Save time by preventing errors, automating tasks, and reusing code
- Facilitate collaboration with organized and documented workflows

Course Materials

- The course Github page has links to project files, readings, and the syllabus
- We will work through examples and complete assignments through a course workspace on RStudio Online. To sign up, follow this link. You will need to create a free user account. We will go over how to access materials on the first day of class.

- Readings from *R for Data Science* (Wickham & Grolemund, 2017) are available online.
- PDFs of other course readings referenced below are available on Github.

Other helpful resources

- *R Markdown: The Definitive Guide*
- papaja: Reproducible APA manuscripts with R Markdown
- RStudio cheatsheets for base R, data import, ggplot, R Markdown, and other packages.
- STAT 545, another good, free textbook on data wrangling and exploration
- Happy Git with R, for all of your Github-Git-R-RStudio needs

Course Policies

Please follow the course Github page for updates to the syllabus and materials. I will communicate changes through email and/or Canvas announcements.

Assignments and Grading

Grading for the class is S/NC. Your grade will be based on the following:

Component	Weight
Participation	20%
Weekly assignments	20%
Workflow critique	20%
Final project	40%

Participation

You are expected to attend each class and participate in class discussions. Readings should be completed prior to class so that you can contribute to discussions.

Weekly Assignments

I will assign short practical assignments after most weeks of class. These exercises will develop your skills in R programming and other concepts that we discuss. You are allowed to work in groups of 2-3 students to complete these exercises if you would like, but each student should turn in their own assignment. Please indicate in the top of your assignment the names of the students you worked with. Some class time will be given so that you can start work on the assignment with the help of the instructor/TA. Office hours are also available for help on assignments.

Workflow Self-Critique

You should choose a current or past data analysis project that you have worked with (or one from your lab if you are a newer student). In a short paper (3-4 pages), you will first describe the end-to-end workflow of your data. What are sources of raw data? How are those sources combined and/or processed? What research personnel, computing resources, software, and hardware devices are involved? What is the end product needed for statistical analysis? Next, you will take a critical eye to your workflow and identify 1) Where are errors most likely to occur, 2) What time-consuming steps could be automated, and 3) how your workflow could be made more transparent and reproducible. You will briefly present your workflow critique in class during week 4 or 5. Each student will work individually on this assignment.

Final Project

In your final project, you will improve the data workflow that you chose uing skills learned in this class. Your final project should be shared with the instructor and TA through an online repository (such as RStudio Online, Github, OSF, or Code Ocean) and allow your end-to-end data workflow to be reproduced (e.g., include raw data files, functions that implement processing steps, etc.). You can use whatever programming languages are necessary (it doesn't need to just be in R), but you should consult with the instructor if R will not be used to ensure that the instructor can run your code (or alternatively, that you demonstrate your workflow to the instructor).

Your project should be a report, either to demonstrate exploratory analyses or to communicate the results, that is written in R Markdown and contains visualizations written in R. In a brief presentation in the last week of class, you will include summarize the changes you implemented in the workflow, describe the problems you believe they have solved, and reflect on what future changes may need to be implemented.

Remote Instruction

The course will be conducted remotely via Zoom using the course link at the top of the page. This class will be synchronous — instruction will take place at the scheduled time. If you are having a technical difficulty (wifi/power outage/etc.) and lose access to the class (or can't get in), please let the TA and instructor know. You are responsible for all class material, but we will do our best to get you caught up.

In the event of absences due to illness or family obligations, please contact the instructor immediately so that we can make arrangements. Everyone is in a different situation, so I will always work with you to figure out a way to handle disruptions to learning. I want everyone in this class to succeed!

Class Schedule

Readings should be completed prior to each class.

Week 01, 01/06: Course Goals

- Goals of the class
- Logistics (RStudio/Github)
- Readings and assignments

R4DS: Introduction and Workflow Basics

OPTIONAL: Intro R lessons using swirl()

Week 02, 01/13: Data Workflow

- File organization
- Version control
- Workflow example
- *SKILLS:* Basic I/O functions, getting help, using Github

R4DS: Data import, Tibbles, and Projects

OTHER READINGS: Minimizing mistakes article (Rouder, Haaf, & Snyder, 2019)

SOFTWARE SETUP: Follow the directions here to setup R, RStudio, git, and Github. Afterwards, read and follow along with the Github beginner guide.

Week 03, 01/20: Data Structure

- Data types
- Data organization (variables/observations)
- *SKILLS:* Factors/strings, reshaping/aggregating data

R4DS: Data transformation, Tidy data, and Factors

OTHER READINGS: Tidy Data (Wickham, 2014)

Week 04, 01/27: Automation

- Copy/paste is evil (automating data pipeline)
- Drop-down menus are evil (automating your analyses)
- Machine-readable formats
- *SKILLS:* Writing functions, iteration, vectors
- **WORKFLOW PRESENTATIONS GROUP 1**

R4DS: Functions, Vectors and Iteration

Week 05, 02/03: Error Checking

- Data validation
- Automating visualizations
- Exploratory data analysis
- *SKILLS:* Basic visualizations, handling missing data
- **WORKFLOW PRESENTATIONS GROUP 2**

R4DS: Data visualization and Exporatory data analysis

Week 06, 02/10: Reuse Part 1: Encapsulation and Writing Functions

- Best option: Use existing APIs and packages
- Next best option: Writing your own general-purpose functions
- Not great option: Writing overly-specific functions that you can never use again
- *SKILLS:* Sourcing functions, parameterization

OTHER READING: Technical debt (Suryanarayana, Samarthyam, & Sharma, 2014)

Week 07, 02/17: Reuse Part 2: Data Sharing, Archiving, and Documentation

- Code as documentation
- Reuse-minded project management
- *SKILLS:* Comments, specifications, licenses

OTHER READING: Care and feeding of data (Goodman et al., 2014)

Week 08, 02/24: Communication Part 1: Reproducible Reports

- Reproducible reports
- Preserving programming environment and analyses
- *SKILLS:* R Markdown, papaja, package control

R4DS: R Markdown

OTHER READING: Transparency in psychological science (Klein et al., 2018)

Week 09, 03/03: Communication Part 2: Visualization

- Communicating through graphical styles
- Interactive/animated plots for data exploration
- *SKILLS:* ggplot and extensions

R4DS: Graphics for communication

OTHER READING: Designing graphs for decision-makers (Zacks & Franconeri, 2020), Chartjunk from (Tufte, 1990, 2001, 2006), and (optional) Graph construction (Witt, 2019).

Week 10, 03/10: Project Presentations

- How have you changed your workflow?
- What have you learned about your data?
- What problems are still unresolved?

References

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., … Slavkovic, A. (2014). Ten simple rules for the care and feeding of scientific data. *PLoS Computational Biology*, *10*(4), e1003542.

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Hofelich Mohr, A., … Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*(1), 1–15.

Rouder, J. N., Haaf, J. M., & Snyder, H. K. (2019). Minimizing mistakes in psychological science. *Advances in Methods and Practices in Psychological Science*, *2*(1), 3–11.

Suryanarayana, G., Samarthyam, G., & Sharma, T. (2014). *Refactoring for software design smells: Managing technical debt*. Elsevier Science.

Tufte, E. R. (1990). *Envisioning information.*

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.).

Tufte, E. R. (2006). *Beautiful evidence* (Vol. 1). Graphics Press Cheshire, CT.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23.

Wickham, H., & Grolemund, G. (2017). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc.

Witt, J. K. (2019). Graph construction. *Meta-Psychology*, *3*.

Zacks, J. M., & Franconeri, S. L. (2020). Designing graphs for decision-makers. *Policy Insights from the Behavioral and Brain Sciences*, *7*(1), 52–63.