PSC203A

Data Management and Cleaning
Winter 2026
Monday 2:10-5 PM, Young Hall 166

Instructor:
Dr. Emorie D. Beck (she/her/hers)
E-mail: edbeck@ucdavis.edu
Office: Young Hall 268J / BROB 157
Office Hours: TBD as winter schedules finalize

- Drop-in hour:
- One-on-one hour: [sign-ups through calendly]
- By other appointment: edbeck@ucdavis.edu

This syllabus is a living document and will be updated throughout the course with readings, shifts in topics, and more as makes sense given student interests tensions between depth and breadth, and more. Please check Canvas for most updated version.
Draft date: 01/02/2026

Course Site: https://emoriebeck.github.io/psc203a-data-WQ26/

Table of Contents

Course Description

In graduate education, training on research (and statistical) methods and conceptual frameworks far outpaces training on key technical skills that underpin all research, empirical or otherwise. On average, researchers spend about 80% of their (analytic) time on data cleaning, but we spend comparatively little teaching those skills. This course aims to fill that gap by helping researchers to (1) build their reproducible research workflow and (2) improve their data cleaning and general statistical programming skills. To that end, each session will be split to address each of these goals, with the beginning of class focused on conceptual ideas about best practices in building a workflow and the latter half focused on technical training on programming and cleaning data in R. This course will be set up as a “bring your own data” course to allow students to anticipate specific challenges that face different types of research.

This course is not a “pure” data science (i.e. we won’t be working with databases, etc.) because it focuses on the skills and tools most common within the social sciences. Science is a collaborative enterprise, and these tools are widely used among many social scientists, which promotes an open, equitable workflow by using tools available and most commonly used by the majority of our peers.

Prerequisites

There are no prerequisites, but students are encouraged to take this course with at least a basic familiarity of the R programming language and data cleaning in R. If the course fills, priority will be given to upper-level graduate students.

You will spend about 50% of this course using R. Therefore, to get the most out of this class, I highly recommend having some basic experience with the R programming language.

This course is a “bring your own data” course. Although having your own data is not a formal requirement, your learning experience will be better if you are able to work with your own data or data specifically related to your field of research.

Learning Outcomes

After successful completion of this course, you will be able to:

1. Build your own research workflow that can be ported to future projects.
2. Learn new programming skills that will help you efficiently, accurately, and deliberately clean and manage your data.
3. Create a bank of code and tools that can be used for a variety of types of research.

Course Materials

There is no official textbook for this course (but if there was, it'd be Wickham, Cetinkaya-Rundel, & Grolemund's R for Data Science [2nd edition]). However, many of you are coming in with different levels of knowledge and different types of questions, so I am providing some suggested readings below.

I have arranged for students in this course to receive free access to Data Camp, a library of R (other programming languages) tutorials. Sign up using your UC Davis email here:
https://www.datacamp.com/groups/shared_links/47a82efe9c4fee59e3ba4592743caa2326d5c091badc782aec969e244e98ab5e

We will pull from the following two (freely available) books:

Hadley Wickham & Garret Grolemund: R for Data Science
Hadley Wickham: Advanced R

All course materials comply with copyright/fair use policies.

Technology Requirements

The lecture presentations, links to articles, assignments, and rubrics are located on the Canvas site for the course. To participate in learning activities and complete assignments, you will need:

- Access to a working computer that has a current operating system with updates installed;
- Reliable Internet access and a UCD email account;
- A current Internet browser that is compatible with Canvas;
- R and R Studio (see below)
- Reliable data storage for your work, such as Box, Office 365, or a USB drive.

We will do all of our data cleaning work in this class using the R programming language. We will use RStudio to interface with R console for a more user-friendly experience. If you are a Python user, you may consider their newer offering, Positron.

Please install both R and RStudio before the first day of class. Here's how:

1. Get the most recent version of R (free). Download the version of R compatible with your operating system (Mac, Linux, or Windows). If you are running Windows or MacOS, you should choose one of the *precompiled binary* distributions (i.e., ready-to-run applications; .exe for windows or .pkg for Mac) linked at the top of the R Project's webpage.
2. Once R is installed, download and install R Studio or Positron. R Studio and Positron are "Integrated Development Environments", or IDEs. This means they are front-ends for R that makes it much easier to work with. Both are also free, and available for Windows, Mac, and Linux platforms.
3. Install the tidyverse library and several other add-on packages for R. These are sets of tolls or functions that will aid us in cleaning and wrangling data, and more. This is a non-exhaustive list that will get us started.

```
my_packages <- c(
      "plyr", "tidyverse", "furrr", "broom",
      "MASS", "quantreg", "rlang", "scales",
      "survey", "srvyr", "devtools", "future"
      )

install.packages(my_packages, repos = "http://cran.rstudio.com")
```

Minimum Technical Skills Needed

Minimum technical skills are needed in this course. All work in this course must be completed and submitted online through Canvas and all assignments will be completed in R / Rmarkdown / Quarto. Therefore, you must have consistent and reliable access to a computer and the Internet.

The basic technical skills you have include the ability to:

- Organize and save electronic files;
- Use UCD email and attached files;
- Check email and Canvas a few times / week;
- Download and upload documents;
- Locate information with a browser; and
- Use Canvas.

However, you will spend about 50% of this course using R. Therefore, to get the most out of this class, I highly recommend having a better-than-beginner understanding or and experience with the R programming language. R is a skill, just like understanding the components of quality data and workflows, and for the purposes of this course, both are equally necessary and important. If you have any concerns about whether your R skills are strong enough for the course, please talk to the instructor or consider taking the course in a future year.

Course Assignments and Grading

General Assignment Information

- All coursework (assignments) is secured in Canvas with a username and password.
- All assignments are due on the day indicated on the course schedule.
- Complete rubrics (final project presentations and paper only) will be provided in Canvas.

Bi-Weekly Assignments

The goal of this course is not simply to teach you *how* to clean hypothetical or convenient data. Rather, the goal is to teach you principles of good, accurate,

reproducible, and efficient data cleaning and management, how to identify features of high quality data, and how to produce results of analyses efficiently.

Bi-weekly homework (40%) in this class will focus on programming concepts from the previous two weeks. Each week, you will complete one problem set, applying the skills you learned that week ***to your own data***. Submit each of these via Canvas by **the start of the next class session**.

These will be graded for completion (you turned it in), relevance (it should be clear that you actually tried to do what you asked), and effort (please show your work). You will not receive feedback on them unless there is an ongoing problem (e.g., lack of depth or effort).

This is good opportunity to:

1. Better understand challenges with your own data (relative to others)
2. Reflect on features of your current workflow you like or dislike
3. Critique your own work and note ideas to improve (I will probably do this a lot in class!).
4. Create a repository of ideas and code for future research.

Final Exam

There is **no** final exam for this course; rather, there is a final project, due at the day and time of the scheduled final exam. The last day of the course will (likely) be used for presentations on the final project in order to receive feedback from the class and instructor.

Additional information on the project will be provided as a separate document on Canvas, announced in week 4 or. The project will not be long and the goal will be for you create a document outlining your workflow.

To ensure that your workflows are as effective as possible, this will proceed in five parts:

1. Initial proposal of an idea submitted via Canvas (10% of your grade).
2. Updated proposal submitted via Canvas (if needed).
3. 5-10 minute presentation to the class on the last day of the course (10% of your grade).
4. Final Project (20%).

Extra Credit

- Participate in a https://www.tidytuesday.com.
- 2 pt extra credit for each one you participate in (max 6 pt total).
- Can post on Twitter or just create a document with the code and output
- Submit on Canvas
    - If posting, link the post in the Canvas submission
    - If not posting, attach the knitted file on Canvas

Evaluation and Grading Scale

All grades will be posted on Canvas. You are strongly encouraged to check your scores in Canvas regularly. A final letter grade will be assigned based on percentages.

Assignment Weights	Percent
Class Participation	20%
Problem Sets	40%
Final Project Proposal	10%*
Class Presentation	10%*
Final Project	20%*
Total	100%

* If presentations are omitted, proposals will be worth 15% and Final Projects 25%.

Grading Scale

Range	Letter Grade
92.5% - 100%	A
89.5% - 92.4%	A-
87.5% - 89.4%	B+
82.5% - 87.4%	B
79.5% - 82.4%	B-
77.5% - 79.4%	C+
72.5% - 77.4%	C
69.5% - 72.4%	C-
67.5% - 69.4%	D+
62.5% - 67.4%	D
59.5% - 62.4%	D-
0% - 59.4%	F

Course Policies and Procedures

Many of the below are also outlined in the UC Davis Code of Academic Conduct.

Attendance Policy

When you miss class, you miss important information, not all of which will be available in the zoom recordings. This course is only 10 class meetings, so each meeting comprises 10% of your in-class time. If you need to miss more than one class, I suggest considering whether taking this course in a future term. I will teach this course either annually or biennially, so there will be future opportunities to take this course in many cases (e.g., for example, if you are a second year student who will miss two meetings, taking the course in your fourth year may be more effective).

Late Work/Make-up Policy

Late work will be allowed per instructor discretion. Please try to proactively communicate these needs. Assignments due at midnight will have a 9 hour “grace period” with no penalty. Each day late is subject to a 20% drop in course grade (e.g., a 10-point response is worth 8 points on day 1 late, 6 points on day 2 late, etc.).

Each student will receive one, no-questions-asked extension. To use this, please email me noting which assignment **before** the deadline.

Academic Integrity

You are expected to practice the highest possible standards of academic integrity. Any deviation from this expectation will result in a minimum academic penalty of your failing the assignment, and will result in additional disciplinary measures. This includes improper citation of sources, using another student's work, and any other form of academic misrepresentation.

Plagiarism

Using the words or ideas of another as if they were one’s own is a serious form of academic dishonesty. If another person’s complete sentence, syntax, key words, or the specific or unique ideas and information are used, one must give that person credit through proper citation.

Incomplete Grades

You may assigned an 'I' (Incomplete) grade if you are unable to complete some portion of the assigned course work because of an unanticipated illness, accident, work-related responsibility, family hardship, or verified learning disability. An Incomplete grade is not intended to give you additional time to complete course assignments or extra credit unless there is indication that the specified circumstances prevented you from completing course assignments on time.

Instructional Methods

The course will be taught using multiple instructional methods. I will typically briefly (45-50 minutes) lecture at the beginning of the class on conceptual topics related to data cleaning and management. We will then have a 75 minute workshop, which will be a mix of going through code and examples together and working in small groups (if preferred) on short exercises. The remainder of the class will be available to receive support on Problem Sets for that week and other general questions (optional). The proportion of these will vary by week and portions of the course will be shortened or dropped as needed.

Diversity and Inclusion

The university is committed to a campus environment that is inclusive, safe, and respectful for all persons. To that end, all course activities will be conducted in an atmosphere of friendly participation and interaction among colleagues, recognizing and

appreciating the unique experiences, background, and point of view each student brings. You are expected at all times to apply the highest academic standards to this course and to treat others with dignity and respect.

Accessibility, Disability, and Triggers *[credit to Dr. David Moscowitz]*

I am committed to ensuring course **accessibility** for all students. If you have a documented **disability** and expect reasonable accommodation to complete course requirements, *please notify me at least one week before accommodation is needed*. Please also provide SDRC (https://sc.edu/about/offices_and_divisions/student_disability_resource_center/) documentation to me before requesting accommodation. Likewise, if you are aware of cognitive or emotional **triggers** that could disrupt your intellectual or mental health, please let me know so that I can be aware in terms of course content.

Absences for Personal or Religious Holidays

I am committed to allowing students to exercise their rights to religious freedom. Accommodations on assignment due dates and absences will be allowed for students observing religious holidays that fall on course days. Please email me to let me know ahead of time to allow for accommodations to be made.

Title IX and Gendered Pronouns *[credit to Dr. David Moscowitz]*

This course affirms equality and respect for all gendered identities and expressions. Please don't hesitate to correct me regarding your preferred gender pronoun and/or name if different from what is indicated on the official class roster. Likewise, I am committed to nurturing an environment free from discrimination and harassment. Consistent with Title IX policy, please be aware that I as a responsible employee am obligated to report information that you provide to me about a situation involving sexual harassment or assault.

Values *[credit to Dr. David Moscowitz]*

Two core values, inquiry and civility, govern our class. **Inquiry** demands that we all cultivate an open forum for exchange and substantiation of ideas. Strive to be creative, to take risks, and to challenge our conventional wisdom when you see the opportunity. **Civility** supports our inquiry by demanding ultimate respect for the voice, rights, and safety of others. Threatening or disruptive conduct may result in course and/or university dismissal. Civility also presumes basic *courtesy*: please be well rested, on time, and prepared for class (class time also includes a break to use the restroom, etc.), which includes silencing all personal devices.

My perspective is that we never cease being students of this world, so I believe that attentive, reflective people always have something to learn from others. Good

discussions can be energetic and passionate but are neither abusive nor offensive. Vibrant, vigorous inquiry derives from discussions that:

- challenge, defend, and apply different ideas, theories, perspectives, and skills,
- extend a body of knowledge into different arenas and applications, and
- result in a synergy that compels us to seek resolution to these discussions.

Copyright/Fair Use

I will cite and/or reference any materials that I use in this course that I do not create. You, as students, are expected to not distribute any of these materials, resources, homework assignments, etc. (whether graded or ungraded) without permission from the instructor.

Course Schedule

Day	Date	Topic	Due Today
First Day of Classes January 5			
1	01/05/2026	*Lecture*: Basics of Workflow *Workshop*: Introduction to R & Workflow Basics; Quarto *Readings*: - r4ds: Ch. 2, 3, 4, 6, 28	
2	01/12/2026	*Lecture*: Reproducibility and Workflow Values *Workshop*: Data Transformation: Introduction to **`dplyr`** *Readings*: - r4ds: Ch. 3	
3	01/19/2026	***MARTIN LUTHER KING, JR. DAY, NO CLASS*** ***RECORDED** Lecture* (watch by 01/26/26): Understanding and Assessing Data Quality *Workshop*: Reshaping and Joining: Introduction to **`tidyr`** *Readings*: - r4ds: Ch. 5	Problem Set 1 Due (Grace period until 01/20/2026 at 2 PM)
4	01/26/2026	*Lecture*: Documenting Data and Procedures *Workshop*: Using Codebooks to Aid Data Import *Readings*: - r4ds: Ch. 7, 20, and 23	
5	02/02/2026	*Lecture*: Functions *Workshop*: Iteration: Introduction to **`purrr`** *Readings*: - r4ds: Ch. 25, 26	Problem Set 2 Due
6	02/09/2026	*Lecture*: Review – Putting the Pieces of Your Workflow Together *Workshop*: Review – tidyverse: Using codebooks, functions, and iteration within a tidyverse framework (a series of in-class activities) *Readings*: - None	
7	02/16/2026	***PRESIDENT'S DAY, NO CLASS*** ***RECORDED** Lecture* (Watch by 02/23/26): Data Structures in R	Problem Set 3 Due **PROPOSALS DUE**

Day	Date	Topic	Due Today
		Workshop: Data Transformation: Dates, Strings, regex, and Other Tricky Classes *Readings*: - r4ds: Ch. 12-18	(Grace period until 02/17/2026 at 2 PM)
8	02/23/2026	*Lecture*: N/A (Workshop week) *Workshop*: (Functional) Tables & Figures *Readings*: - (Probably none ☺)	
9	03/02/2026	*Lecture*: Git & GitHub *Workshop*: Parallelization: Introduction to `future` and `furrr` *Readings*: - https://dcgerard.github.io/advancedr/09_future.html	Problem Set 4 Due
10	03/09/2026	In-Class Presentations	(Mini) Problem Set 5 Due
	03/18/2026	Final Project Due	

Other topics:

-

Sources of Inspiration used in the creation of this course:

- R for Data Science
- My own blood, sweat, and tears (mostly caused by own mistakes and inability to ask for help)