PSYC 259: Principles of Data Science
Winter 2025

Instructor:
John Franchak

Substitute Instructors:
Olivia Atherton (olivia.atherton@ucr.edu)
Stephen Antonoplis (stephen.antonoplis@ucr.edu)
Tabea Springstein (tabea.springstein@ucr.edu)

Office Hours: by appointment only (email)

** If you have a question about the class, please email all three substitute Instructors and Teaching Assistant.

Teaching Assistant:
Madison Montemayor-Dominguez (madison.montemayordominguez@email.ucr.edu)

In-Person Office Hours: Tuesdays, 2pm-4pm: Olmsted 2107
Zoom Office Hours: Fridays, 10am-12pm: https://ucr.zoom.us/my/macmonted

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 Quarto 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 Github page will have all of the code repositories, slides, and assignments.

- Readings from *R for Data Science* are available online.
- PDFs of other course readings are linked in the schedule below and lectures
- Other helpful resources are compiled for your future reference.

Course Policies
Keep an eye on Canvas for updates to the syllabus and materials, as well as 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	30%
Workflow critique	20%
Final project	30%

Attendance and Participation
You are expected to attend each class and participate in class. Readings should be completed prior to class so that you are prepared. In the event of absences due to illness or family obligations, please contact the instructors immediately so that we can make arrangements. Everyone is in a different situation, so we will always work with you to figure out a way to handle disruptions to learning. We want everyone in this class to succeed!

Weekly Assignments
We 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 2-4 students to complete these exercises if you would like, but each student should turn in their own assignment. Please indicate the names of the students you worked with at the top of your assignment. Some class time will be given so that you can start work on the assignment with the help of the instructors/TA. Office hours are also available for help on assignments. To submit your assignments, fork the corresponding homework repository, complete the homework, and when you are finished, add Madison as a collaborator. This will send Madison an email about the repository. Madison's username is macmonted.

Workflow Critique
The goal of this assignment is to take an existing workflow and critically analyze it with an eye for 1) efficiency, 2) fidelity, and 3) sharing/reproducibility. The project will take the form of an in-class presentation (5 minutes). You should choose a current or past data analysis project that you have worked with (or one from your lab or a public repository if you are a new student or don't have data to work with). You will make changes to this workflow for your final course project, so choose something that will be useful to you! Each student should work on their own assignment (unlike homework, this is not a collaborative project). You will give your presentation in class during class on Feb 10, Feb 24, or Mar 3. Regardless of your presentation date, your slides are **due on Sunday Feb 9 by 5 pm.** Export your slides as a PDF with file name "Lastname_critique.pdf" and upload to Canvas.

In your presentation, you will have two sections: workflow and critique. The workflow section should be briefer (about 1-2 minutes) so that you can focus more on the critique. Avoid spending time describing the project aims/research goals; please keep your focus on the

workflow and implementation of the project. Here are more details about what these sections should entail:

1. **Workflow**

   In the workflow section, describe the end-to-end workflow of your data analysis project. Provide only enough detail here so that we can understand the critique section. What are sources of raw data, and how are they stored? What steps are needed to combine and/or process the raw data? What research personnel, computing resources, software, and hardware devices are involved in your workflow? What is/are the end product(s) needed for statistical analysis? You don't have to answer every question, just what helps you explain the critique.

2. **Critique**

   Next, you will take a critical eye to your workflow and discuss:

   Efficiency: Discuss what time-consuming steps in the workflow could be improved with coding (or coding with more automation). What parts of the workflow are iterative and need to be repeated when making changes or correcting errors?

   Fidelity: Discuss where errors in the workflow are most likely to occur and what procedures you might implement to improve the fidelity of data, such as eliminating copy/paste or other manual procedures or implementing formal data checks. How do you track what data are cleaned, and what participants/sessions should be included in final analyses?

   Sharing/reproducibility: Examine how an outsider might view the project. How well are the files organized and how well documented are the procedures? Would someone be able to reproduce the analyses? If not, what steps could be taken to make this possible (e.g., coding the analyses rather than using a drop-down menu)?

Final Project

The goal of this assignment is to take an existing workflow and improve its 1) efficiency, 2) fidelity, and 3) sharing/reproducibility. There are two parts of the project. The first part is a Github repository that tracks changes you made to your project and shows how your new workflow is set up. The second part is a written report that describes the changes you made, why you made them, and what benefits each one should create. This report should use Quarto to illustrate changes you made to code, the examples of error checking and EDA, and figures made from your data. All of the files, including data files, scripts, figures, and the Quarto report, should be stored on a **private** Github repository. The final project is **due on Sunday March 18 at 5 pm**. You should share your **private** Github repository with the three Instructors and the TA by adding them as collaborators.

Part 1: Github repository

Before you start to make changes to your project, set up a **private** Github repository with your current project files so that we can see the "before". Because the focus of the course has been on data wrangling and exploratory analysis, your project should contain raw data files, script(s) that import and clean the data, and scripts that explore the data and/or check for errors. If you have raw data that contains sensitive information, please de-identify the data or make a simulated data file that has the correct format/structure so

that we can understand the project. Please share your **private** Github repository with the three Instructors and the TA by adding them as collaborators.

As you work on your project and make changes, commit and push the changes to Github so that we can see what you changed. If you've already started making changes and they're not tracked in Github, let us know and we can try to figure out a way to see the "before" version of the project. If your project uses a coding language other than R for data import/cleaning/checking, that's OK. Let us know in advance so that we can make sure we have a way to see/understand what your project does. No matter what language(s) you use in the project, everyone should use Quarto and R to make the report.

What changes should you make?

Based on your Workflow Critique assignment and our feedback, you should make changes to improve the efficiency, fidelity, and sharing/reproducibility of the project. You do not need to cover these three categories equally. If improving the efficiency is most important to your project, focus on that. Of course, some changes might fall into multiple categories (e.g., automating reading data from a file is more efficient and is less likely to lead to mistakes). Please describe the reasoning behind the changes you chose to make. Your final project should contain at least one example of the following items:

- Wrote a custom function to reduce code repetition and/or split long scripts into more manageable files
- Used some form of automation (map, for loop, across, read_csv) to replace repetitive code or manual data entry/copy/paste
- Plotted graphs or created data checks to explore the data
- Improved documentation of the project, file organization, and/or readability of the code

We do not expect the final project to be perfect (whatever that means!). It would be impossible to make every change you might want to make in a few weeks. You can, instead, demonstrate in one section of the code how you used automation to improve the workflow, and describe how you might apply a similar technique in other sections of your project. Prioritize making changes that will help your project the most and help you learn skills you want to learn.

Part 2: Quarto report

Use Quarto to prepare a report that summarizes the work you did to improve your project. Because everyone's changes will be different, you should structure the report in a way that makes sense to you. For each change that you made, please be sure to explain how you did it and why you did it. Make use of Quarto code chunks to illustrate the before and after of a change you made to code (not every change, but changes that are representative of the type of change). You are encouraged to embed figures to illustrate the results of your data checking and exploration. Render your Quarto file as a PDF or HTML so that we can view it easily.

Schedule

Date	Lead	Topic	Reading (before class)	Assignment (all due at 5 PM)
Jan 27	Tabea	File Organization and Workflow	Workflow Basics Scripts and projects Data import	Data Importing HW due Sun Feb 2
Feb 3	Olivia	Data Transformations *[Overview of Workflow Critique Assignment]*	Data transformation Numbers Strings	Data Transformations HW due Fri Feb 7 **Workflow Critique due Sun Feb 9**
Feb 10	Tabea	Data Structure *[Workflow Critique Presentations Group #1]*	Tidy data Factors Dates	Data Structure HW due Fri Feb 14
Feb 17	No class (Presidents Day)			--
Feb 24	Stephen	Automation: Functions and Iteration *[Overview of Final Project Assignment + Workflow Critique Presentations Group #2]*	Functions Vectors Iteration	*Optional* Functions HW (extra credit; due Fri Feb 28)
Mar 3	Olivia	Exploratory Data Analysis *[Workflow Critique Presentations Group #3]*	Data visualization Layers Exploratory data analysis	Integrating Skills HW due Fri Mar 7
Mar 10	Stephen	Data Sharing and Reproducibility	Quarto Quarto formats	Rmarkdown HW due Fri Mar 14
Mar 17	Tabea	Visualization	Chartjunk-Tufte 1990; 2001; 2006 Graphics for communication	**Final Project due Wed Mar 18**