HARVARD T.H. CHAN | SCHOOL OF PUBLIC HEALTH

ID529 Data Management and Analytic Workflows in R
Winter Session 2023

Course Listing: ID529
Course Hours: 1:30-5:30 PM
Course Dates: Monday January 9th - Friday January 13th, Tuesday January 17th - Friday January 20th, 2023
Limit 40 students, priority for Population Health Science (PHS) students

Instructional Team:

Christian Testa
Department of Social and Behavioral Sciences
ctesta@hsph.harvard.edu
https://scholar.harvard.edu/ctesta

Dean Marengi
Department of Environmental Health
dean_marengi@g.harvard.edu

Jarvis Chen
Department of Social and Behavioral Sciences
jarvis@hsph.harvard.edu
https://www.hsph.harvard.edu/profile/jarvis-chen/

Credits
2.5 Credits

Course Description
Data Management and Analytic Workflows in R will introduce students to R programming and modern data management and analysis workflows applied to examples from population health science. Throughout, we will emphasize reproducibility, open science, data visualization, and dynamic document generation. Specific skills learned will include the use of the RStudio integrated development environment, tidy data management practices/workflows, how to get help in programming, and how to use GitHub to track changes in code, disseminate professional work, and integrate feedback. Coursework will consist of lectures, in-class group work, homework, peer assessment, and time for discussion. This course complements graduate-level courses in statistics and quantitative research methods by helping students develop practical skills for conducting independent research incorporating modern data science principles. Students completing this course will have a solid foundation enabling them to handle complex data management tasks and data communication skills for research and professional work.

Prerequisites
Prerequisites include an interest in learning R programming and an interest in population health science. Suggested prior coursework includes at least one of PHS2000A, BST201 or 210, ID201 or ID207.

Key Learning Objectives
Upon successful completion of this course, students will be able to:
- Articulate and use best practices for data cleaning, management, and project organization in the context of R programming-based analyses focused on population health science using example datasets from sources commonly used in the public health field such as the CDC National Health and Nutrition Examination Survey (NHANES), CDC WONDER (Wide-ranging ONline Data for Epidemiologic Research), US Census American Community Survey (ACS), and longitudinal datasets such as the National Longitudinal Survey of Youth (NLSY).
- Build and implement reproducible workflows and describe the merits of reproducibility in data analysis and scientific software.
- Use R Markdown to create dynamic reports for their findings.

- Use Git and GitHub version control software to disseminate code and engage in collaborative work and peer review which enables students to showcase their work to public audiences.
- Perform exploratory data analyses (e.g., data visualization, working with regression models to present output in professional quality tables, etc.).
- Leverage online resources to learn more and get help with any R programming challenges.
- Paint a better picture of the variety of spectacular kinds of data analysis they could go on to do!

Suggested References

- *R for Data Science* by Hadley Wickham and Garrett Grolemund, available here: https://r4ds.had.co.nz/
- *What They Forgot to Teach You About R* by Jenny Bryan and Jim Hester, available here: https://rstats.wtf
- *ggplot2 Elegant Graphics for Data Analysis* by Hadley Wickham, Danielle Navarro, and Thomas Lin Pedersen, available here: https://ggplot2-book.org/
- *Happy Git and GitHub for the 2ser* by Jenny Bryan, available here: https://happygitwithr.com/
- *The tidyverse Style Guide*, available here: https://style.tidyverse.org/
- *R Markdown Cookbook* by Yihui Xie, Chirstophe Dervieux, Emily Riederer, available here: https://bookdown.org/yihui/rmarkdown-cookbook/
- *R Packages* by Hadley Wickham and Jenny Bryan, available here: https://r-pkgs.org/
- *Fundamentals of Data Visualization* by Claus O. Wilke, available here: https://clauswilke.com/dataviz/
- *The Epidemiologist R Handbook*, available here: https://epirhandbook.com
- *Regression and Other Stories* by Andrew Gelman, Jennifer Hill, and Aki Vehtari, pdf download available here: https://avehtari.github.io/ROS-Examples/index.html
- *Introduction to Data Science: Data Analysis and Prediction Algorithms with R* by Rafael Irizarry, available here: https://rafalab.github.io/dsbook/
- *Public Health Disparities Geocoding Project 2.0 Training Manual*, available here: https://phdgp.github.io/PHDGP2.0/

Course Structure

This is a two-week winter session course. It will be composed of daily lectures, in-class activities including discussions and Q&A, individual homework assignments, peer-review, and a final project. Each class session will consist of approximately three hours of lecture-time broken up by discussion, programming and group activities, and Q&A (totaling four hours per session).

The final grade for this course will be based on:

- Attendance (10%)
- Homework (35%)
- Peer Review (25%)
- Final Presentation (30%)

Homework

There will be eight individual homework assignments. Assignments will be brief and are intended to help students gain hands-on experience implementing the data management concepts covered during class sessions. Each homework assignment is due by the next class after it's been assigned.

You are encouraged to collaborate on homework assignments with your peers; however, you must submit your own work.
Homework will be graded according to the following rubric:

Objective/Principle	Percent of Grade
Does it accomplish the stated goal? Is it complete?	25%
Is it well documented and commented?	25%
Is it transparent and clearly motivated? Is it elegant (i.e., not kludgey)?	25%
Does it incorporate what's been taught? Does it reflect growth?	25%

Each of these items will be scored on a 1–5-point scale.

Participation and Peer Review
Students are encouraged to participate in class discussion and must participate in peer review.

Peer review will be graded according to the following rubric:

Objective/Principle	Percent of Grade
Does it include constructive criticism?	50%
Does it include positive feedback (i.e., things you liked about their approach)?	50%

Constructive (excellent) feedback examples should:	Fair (mediocre) feedback examples of peer review are:	Not-helpful (bad) feedback examples of peer review are:
- Relate criticisms and feedback to the objectives of the assignment - Point out specific aspects of the code that can be improved upon - Connects the feedback to principles of why the code falls short or is high quality	- Less detailed than constructive - Less explicit in its connection to good coding principles and practices	- Overly terse - Not specific - Harsh, overly critical, or not in the spirit of constructive collaboration
Example: I liked how your code comments went into the details of how you came up with your approach and referenced some of the sources you learned from. It was interesting to see that you were using base R methods to create your figure — though I wonder if that means you missed out on some of the data visualization features that are automatically included when using ggplot2. If I had one aspect I'd most like to see you improve on, it would be to use more informative variable names, like instead of just "x" or "var" you could have used "df" to indicate that the data is a data-frame and "facet_var" to indicate that the variable will be used for faceting.	Example: I thought your comments were good. I think you should use ggplot2 next time though. Your variable names could be improved.	Example: Your comments were verbose. Your figure doesn't look very good though. Your code was hard to read.

Final Presentation

The final project will be a presentation of best practices related to a specific aspect of data analysis and management (see Course Schedule, last day for the list of specific topics). The presentations will be assessed according to the following objectives/principles:

Objective/Principle	Percent of Grade
Does the presentation describe principles and clear guidance on methods to achieve those principles that can be applied in the context of their chosen topic or subject matter?	75%
Does the presentation communicate how students have learned and grown through completing the coursework?	25%

Attendance

Because this course is so short (as it is a winter session course), attendance is mandatory. We will provide accommodations on a case-by-case basis for students with health (including mental health) emergencies. Students requesting accommodations should also contact the appropriate Accessible Education Office for their school (see below). Lectures will be recorded and available via the Zoom tab on Canvas.

Students may be excused from one or more in-person sessions so long as they provide advanced notice to the teaching team (i.e., at least 24-hours prior to a session). Students with unexcused absences will receive a 1% deduction in their attendance grade per session missed (e.g., 1 session missed = 10 - 1 % = 9 %; 2 sessions missed = 10 - 2% = 8%; etc.)

Communication with the Instructional Team and Peers

If you have questions about material from class, we encourage you to post your questions to the course Slack workspace (accessible through the Canvas page for the course). Teaching fellows will respond to questions as a thread. By posting on Slack, the whole class can benefit from the questions posed and others can ask follow-up questions. This is a great way to keep up with the material asynchronously, as it allows you to ask questions that you otherwise may have asked during lecture

- For questions about homework, or any other course-related questions, please email the full instructional team.
- For questions about disability accommodations, please email Jarvis (jarvis@hsph.harvard.edu). Students will be advised to contact their school-specific Accessibility Education Office (for HSPH students: Colleen Cronin ccronin@hsph.harvard.edu; for GSAS students: Robyn Bahr dao@fas.harvard.edu)

In general, you can expect a response within 24 hours, with the exception of weekends.

Feedback

We invite feedback about your learning experience in this course. Please email the instructional team if you have comments, concerns, or suggestions for how we can improve your learning experience.

Harvard Chan Policies and Expectations

Inclusivity Statement

Diversity and inclusiveness are fundamental to public health education and practice. Students are encouraged to have an open mind and respect differences of all kinds. I share responsibility with you for creating a learning climate that is hospitable to all perspectives and cultures; please contact the faculty instructor (jarvis@hsph.harvard.edu) if you have any concerns or suggestions.

Bias Related Incident Reporting

The Harvard Chan School believes all members of our community should be able to study and work in an environment where they feel safe and respected. As a mechanism to promote an inclusive community, we have created an anonymous bias-related incident reporting system. If you have experienced bias, please submit a report here so that the administration can track and address concerns as they arise and to better support members of the Harvard Chan community.

Title IX

For information on Harvard University policies and procedures and Title IX Resource Coordinators at Harvard Chan, please see:

- Harvard University Interim Title IX Sexual Harassment and Interim Other Sexual Misconduct policies and procedures: https://titleix.harvard.edu/policies-procedures
- Title IX Resource Coordinators: https://titleix.harvard.edu/coordinators
- Title IX Sexual Harassment and Other Sexual Misconduct resource guide: https://titleix.harvard.edu/resource-guide

Academic Integrity

Each student in this course is expected to abide by the Harvard University and the Harvard T.H. Chan School of Public Health School's standards of Academic Integrity. All work submitted to meet course requirements is expected to be a student's own work. In the preparation of work submitted to meet course requirements, students should always take great care to distinguish their own ideas and knowledge from information derived from sources.

Students must assume that collaboration in the completion of assignments is prohibited unless explicitly specified. Students must acknowledge any collaboration and its extent in all submitted work. This requirement applies to collaboration on editing as well as collaboration on substance.

Should academic misconduct occur, the student(s) may be subject to disciplinary action as outlined in the Student Handbook. See the Student Handbook for additional policies related to academic integrity and disciplinary actions.

Accommodations for Students with Disabilities

Harvard University provides academic accommodations to students with disabilities. Any requests for academic accommodations should ideally be made before the first week of the semester, except for unusual circumstances, so arrangements can be made. Students must register with the Local Disability Coordinator in the Office for Student Affairs to verify their eligibility for appropriate accommodations. Contact Colleen Cronin ccronin@hsph.harvard.edu in all cases, including temporary disabilities.

Religious Holidays, Absence Due to

According to Chapter 151c, Section 2B, of the General Laws of Massachusetts, any student in an educational or vocational training institution, other than a religious or denominational training institution, who is unable, because of his or her religious beliefs, to attend classes or to participate in any examination, study, or work requirement on a particular day shall be excused from any such examination or requirement which he or she may have missed because of such absence on any particular day, provided that such makeup examination or work shall not create an unreasonable burden upon the School. See the student handbook for more information.

Grade of Absence from Examination
A student who cannot attend a regularly scheduled examination must request permission for an alternate examination from the instructor in advance of the examination. See the student handbook for more information.

Final Examination Policy
No student should be required to take more than two examinations during any one day of finals week. Students who have more than two examinations scheduled during a particular day during the final examination period may take their class schedules to the director for student affairs for assistance in arranging for an alternate time for all exams in excess of two. Please refer to the student handbook for the policy.

Course Evaluations
Constructive feedback from students is a valuable resource for improving the teaching and learning experience. The feedback should be specific, focused, and respectful. It should address aspects of the course and teaching that are positive, as well as those which need improvement.

For registered students, submission of course evaluations is considered to be a School requirement because of their importance. The course evaluation system will open during the last week of the term and remain open for a three week period. You will gain access to your grades for the term after you have completed your course evaluations, and the course evaluation system has closed.

Course Schedule

In-Class Work/Activities	Learning Objectives Upon successful completion of this coursework, students will be able to:	Assignments
Week 1: Monday January 9th - Friday January 13th		
Monday January 9th • Lecture: Demonstration of Modern Data Science Practices in an R Project Based Workflow • Activity: Group reflection on principles shown in the demonstration • Lecture: Introduce the **Final Presentation** Assignment • Lecture: Intro to RStudio and R • Activity: Discussion and Self-Introductions • Lecture: Intro to Git and GitHub • Activity: Setup GitHub accounts + SSH authentication + join our course organization	• Identify the components of a reproducible data management workflow • Setup their own GitHub profile and push files to it • Demonstrate the ability to use basic commands to create and manipulate objects in R • Describe the basic object types in R and how to use them • Reference online documentation and help pages to get help while working on R programming tasks	Homework 1: Write a bio for yourself in day1/README.md file and include a picture
Tuesday January 10th • Lecture: Intro to R Programming (including conditionals, control-flow, etc.) • Lecture: Data Dictionaries and Documentation • Activity: Discussion • Lecture: Reading in data from diverse data sources • Lecture: Intro to ggplot2 (common types of figures, faceting, legends, patchwork, and saving figures) • Discussion: What makes an effective data visualization?	• Understand and demonstrate how and when to use for-loops, while-loops, and conditionals. • Read various file types into R (Excel, CSV, fixed-width, SAS, SPSS, STATA) • Construct useful data dictionaries • Create common statistical plots and figures using ggplot2	Homework 2: Read in one of the suggested datasets and create a figure using ggplot2. We want to see students include titles, subtitles, captions, data sources, legends, etc.
Wednesday January 11th • Lecture: Starting R Projects – Project Based Workflow in R • Lecture: Intro to dplyr – Comparing base R vs. tidyverse • Lecture: Cleaning Text Data • Lecture: Writing Functions • Activity: Write functions together • Activity: Discussion Q&A from ggplot2 homework	• Set up R projects and reproducible data management and analysis workflows • Use dplyr to manipulate and clean data • Use the stringr package to clean text data • Write functions to automate repetitive tasks	Homework 3: Choose one of the prespecified datasets and, using an R project, implement concepts covered in lecture to prepare data for analysis and visualization. Peer Review for Homework 2

In-Class Work/Activities	Learning Objectives Upon successful completion of this coursework, students will be able to:	Assignments
Thursday January 12th • Lecture: Diverse Data Sources (APIs [tidycensus, WHO, World Bank, qualtRics, twitter], scraping web data, datapasta) • Discussion: What kind of sources are students interested in using in their dissertation work? • Lecture: How to handle factors and date-times • Lecture: Regression Modeling – How to fit regression models, extracting model summaries from fit model objects, visualizing model estimates • Lecture: Creating maps in R • Activity: Get setup with software dependencies to make maps in R	• Read data into R from online data sources including publicly available databases and by scraping web pages • Clean categorical data using the factor data type in R • Clean date-times based data and perform common operations with them (e.g., binning, calculating durations) • Extract and report on regression model estimates (coefficient estimates, confidence intervals, p-values) • Create maps in R	Homework 4: • Choose one of the prespecified geospatial datasets and create a map in R • Fit and report on a regression model including categorical (factor) variables Peer Review for Homework 3
Friday January 13th • Lecture: Why reproducibility and robustness are important principles in science and data analysis acknowledging the pressures in academia that push people away from reproducible science • Lecture: Regression Modeling (Poisson, logistic, smoothing splines) • Lecture: Data Linkage Methods (different kinds of joins, probabilistic join methods) • Lecture: Introduction of the Onikye et al reproduction article • Activity: Course evaluation (note that this course evaluation is in addition to the HSPH course evaluation)	• Articulate and defend their choices to use reproducible and robust programming practices • Be able to use and communicate results from more advanced regression techniques such as smoothing splines, GLMs, etc. • Understand the limitations and features of conventional, probabilistic, and deterministic record linkage approaches • Identify the dangers of scientific practices that are difficult to reproduce (copy/paste, lack of version-control, lack of open-source code, inadequate documentation)	Homework 5: • Fill out activity template for Onikye et al reproduction article • Combine multiple data sources using joins • Finding Online Resources Exercise Peer Review for Homework 4

In-Class Work/Activities	Learning Objectives Upon successful completion of this coursework, students will be able to:	Assignments
Week 2: Tuesday January 17th - Friday January 20th		
Monday January 16th, 2023: Martin Luther King Jr. Day (Observed Holiday)		
Tuesday January 17th • Activity: Discussion of Onikye et al reproduction article • Lecture: How to create R packages that standardize data loading and cleaning processes • Lecture: How to use R Markdown to produce reproducible reports including tables, visualizations, and inline-quantitative statements. • Activity: Discussion of online resource exercise	• Create R packages to standardize routine processes, such as interfacing with a particular dataset • Use R Markdown to prototype and develop professional (journal-quality) tables and data visualizations with analytic results	Homework 6: • Write an R package that contains a) some data chosen from our example datasets, and b) functions that load that data. • Use R Markdown in your package to document some exploratory data analysis looking at those datasets. Peer Review for Homework 5
Wednesday January 18th • Discussion: When should you choose a Project vs. a Package based workflow? • Lecture: Best practices for reporting on missing data • Lecture: Intro to accessible machine learning methods for exploratory data analysis – k-means clustering, singular value decomposition, principal components analysis, random forests • Discussion: How can you leverage machine learning methods to further understand your data? What are the risks involved?	• Report on missing data including complex patterns of missingness • Use easy-to-implement machine learning methods to assist in exploratory data analysis • Articulate the risks and pitfalls of using mathematical modeling or machine learning methods uncritically	Homework 7: • Add functions to your R package that clean the data • Create an R Markdown document that presents the patterns of missingness in their chosen dataset Peer Review for Homework 6

In-Class Work/Activities	Learning Objectives Upon successful completion of this coursework, students will be able to:	Assignments
Thursday January 19th • Lecture: Using purrr to automate creating many models and summarize them • Discussion: What are the kinds of projects in which you might use many models?	• Use functional programming and the purrr package to automate creating many models, quickly in R • Build simple interactive applications that enable data exploration • Analyze text-based data to produce quantitative findings	Homework 7: • Create many models using one of the example datasets • Report on the fit models and summarize findings • Work with classmates to prepare final presentations
Friday January 20th • **Final Presentation**: Students present in small groups best practices they've identified for data management and analysis related to a topic chosen from the following list: – Clean Code and Code Hygiene – Data Documentation and Dictionaries – Code Commenting and Documentation – Presentation of Mathematical, Statistical, and Machine Learning Models – Data Visualization – Exploratory Data Analysis – Reporting on Missing Data – Sharing Code and Open-Source Science – Use of Interactive Web Applications in Data Analysis – R Project Workflows – R Package Workflows • Lecture: Recap of principles for reproducibility and robustness in open-source science and R programming • Activity: Course evaluation (note that this course evaluation is in addition to the HSPH course evaluation)	• Refer to a shared, common set of principles and advice their peers have identified as the most important in different aspects of data management and analysis workflows	Enjoy having finished the course!

If time permits, we may be able to cover additional topics based on student interest including: using shiny to create interactive data analysis tools and text analysis.