Data Analysis for Policy Research Using R

Columbia | SIPA

Fall 2025

Instructor: Harold Stolper (he/they)
Office: IAB 906A
E-mail: hbs2103@columbia.edu
OH: Walk-in Wed 2-4 | or sign up with Calendly
Website: https://hreplots.github.io/U6614/

Class: Tues 1:10-3p | 3:10-5p (IAB 510a)
Recitation: Thurs 10-11:50a (IAB 510a) (Seung Min)
Thurs 3:10-5:00p (IAB 510a) (Nico)
TAs: Seung Min Kim (sk5316)
Nico Rojas (nr2670)

Course Description

This course will develop the skills to prepare, analyze, and present data for policy analysis and program evaluation using R. In Quant I and II, students are introduced to probability and statistics, regression analysis and causal inference. In this course we focus on practical microeconometric applications of these skills to explore data and policy questions (note that we do not cover macroeconomic topics or forecasting methods). The goal is to help students become effective analysts and policy researchers: given the available data, what sort of analysis would best inform our policy questions? How do we think about research design, prepare data and implement statistical methods using R? How can we begin to draw conclusions about the causal effects of policies, not just correlation? What should we keep in mind to make sure we're using "data for good", especially when the focus is on marginalized communities using data on personal identity?

We'll learn these skills by exploring data on a range of policy topics: Caste-based expenditure gaps in India; racial disparities in NYPD subway fare evasion enforcement; water shutoffs in the city of Detroit; Village Fund grants in Indonesia; public health insurance and child mortality; Stand Your Ground laws and gun homicides; and student projects on topics of your choosing.

Course Learning Goals

- **Research design:** understanding how data structure impacts analysis and causal inference
- **Data wrangling:** cleaning and structuring data for analysis
- **Exploratory analysis:** identifying and analyzing key factors in your analysis
- **Econometric analysis:** estimating (causal?) relationships between variables to inform policy
- **Data visualization and presentation:** conveying findings to your target audience
- **Policy writing and interpretation:** communicating statistical analysis in accessible terms
- **Data advocacy**: thinking critically about using *data for good*
- **R programming and troubleshooting skills** (these skills support all above the areas)

Prerequisite Requirements

1. Students should have some basic exposure to R, or a demonstrated aptitude for object-oriented programming languages. For students who are brand new to R, to satisfy this prerequisite you can complete two short DataCamp Courses in advance: Introduction to R, and Introduction to the Tidyverse. These are 4-hour courses that can be accessed for free via our own dedicated DataCamp Classroom. **If you have no prior R experience, you are required to complete these short courses before the first class meeting.** If you completed *U6593 R for Public Policy* or you are already familiar with the tidyverse package for R, then you are exempt from both DataCamp prerequisite courses.
2. Students should have completed both U6500 and U6501 (Quant I and II) or equivalent.

Please note that completion of the two DataCamp courses and Quant I and II does not guarantee enrollment in the class, as a waitlist is likely.

The prerequisites aim to ensure students can keep up with the pace of the course, and have the necessary econometric foundation to inform their work with R. Learning a new programming language requires a steady investment of time and energy: investing in this foundation from the outset allows us to explore interesting and realistic data exercises in (relatively) short order.

Required Software

The course will be taught using R, a free, open-source programming language. R has become the most popular language for statistical analysis in many circles. One advantage to using R is the thousands of open-source "libraries" created by R users. By learning R you'll be able to carry out practically any statistical method and access powerful capabilities for data collection, manipulation, and visualization. It is necessarily more complex than Stata, but far more flexible.

We'll be working with R using RStudio. Instructions on installing R + RStudio can be found at https://stat545.com/install.html. ***Please install both R and RStudio on your laptop prior to our first class session.***

Class will be taught in the 510a computer lab, but we'll all be working on our own laptops because the lab computers don't currently support the necessary RStudio functionality.

Course Structure and Approach to Learning R

Course Structure

1. **Pre-class lessons** will be shared via the course website with the expectation that students work through them independently in advance of class. Each class meeting will begin with a short (closed-note) CW quiz on this content. The idea is to introduce key concepts and syntax in R and some statistical methodology to prepare for in-class discussion and coding exercises. These pre-class lessons will take the form of web-based lessons (html files) including sample code and output that students can replicate on their own as they go. Pre-class lessons for Tuesday's class session will be posted by the previous Thursday.
2. **In-class workshop-style instruction using R** will take up the majority of our in-class time together. We'll be working through R code together using RStudio to prepare and explore data for analysis. This will include a mix of reviewing pre-filled code line by line, and short exercises for students to arrive at their own coding solutions. We will also spend considerable time for in-class discussion about how we can use R code to craft and implement a

research design with appropriate econometric methods–"why" and "when", not just "how" to work with data using R.

3. **Six weekly data assignments including a write-up of findings and policy conclusions.** The data assignments will require you to expand on the work we do together in class and write up your work using clear, accessible language. We will introduce R Markdown as a tool for you to write up your work and present code and findings in a single document. Data assignments will be due before midnight on Mondays, in advance of Tuesday's class session.

4. **A group data project** chosen by students (with instructor approval) to be conducted in consultation with the teaching team and presented and submitted towards the end of the semester. Students are required to work in ***groups of three*** (not two or four). The project will require you to use R to explore a policy-relevant research question with readily available data. It must focus on analyzing the effect of at least one independent variable of interest on some relevant outcome variable, though the majority of work you do will involve data cleaning, manipulation, and exploratory data analysis to inform the specification of appropriate statistical models. In the latter half of the semester, student groups are ***required*** to sign-up for three individual meetings with the instructor and TA to discuss their progress.

5. **Recitations.** During recitation time, the TAs will review code to reinforce the material for the week, typically building on the same data exercises from "lecture" to help prepare students for the assignment. Later in the semester, recitation will transition to walk-in OH focused on support for student projects. ***You are required to be able to attend one of the recitations each week.***

6. **Teaching team office hours.** The instructor and TAs will all hold weekly office hours, scheduling and sign-up details will be posted at the start of the semester.

7. **A course discussion board** where students can post reactions to pre-class lessons and ask questions/comments to share with classmates and the teaching team. If you're stuck or experiencing problems with R more generally, odds are others are too. Posting questions and concerns allows us all to benefit from each others knowledge. We'll be using **Ed** via Courseworks for our online discussion. When asking R questions via Ed, please include as many details to replicate the "error" (if applicable), including code snippets, output and screenshots where appropriate. The teaching team will do our best to reply within 48-72 hours, but we encourage students to reply to other posts with their own insights. *Thoughtful contributions will count towards your overall class participation grade*.

Approach to Learning R

Our approach will emphasize "learning by doing" by working through R code together in class to explore data and inform policy questions. The pre-class lessons will introduce key concepts to prepare us for the class workshop exercise. Assignments will task you with refining and expanding the code from in-class workshop exercises, putting your new knowledge to work.

It will take us some time to build up the skills to effectively explore messy, real-world data. Learning a new programming language can feel a bit overwhelming at times, and this class is only the beginning. The goal of this course is not to become proficient in the sense of memorizing all the commands you think you will need, but rather to understand the basics of R syntax and develop the comfort level to explore new functionality and troubleshoot on your own.

Online resources and coding "cheat sheets" will be shared periodically, but learning how to find

and employ answers from both within RStudio and using Google will be some of your most valuable resources.

Virtually all course materials will be generated using RStudio, posted to the course website along with source code for you to consult as examples.

The following open-source textbooks are good supplementary learning resources that we'll rely on throughout the course:

- Bryan, J. (2018). *STAT 545: Data wrangling, exploration, and analysis with R*. Retrieved from https://stat545.com.
- Grolemund, G., & Wickham, H. (2018). *R for Data Science*. Retrieved from http://r4ds.had.co.nz.
- Hanck et al (2020). *Introduction to Econometrics with R*. Retrieved from https://www.econometrics-with-r.org/index.html.
- Xie, Y., Allaire, J. j., & Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Retrieved from https://bookdown.org/yihui/rmarkdown.

Data Community

In-class exercises and discussion are designed to foster a data community where students can interact among themselves and with the teaching team to share ideas. Data and coding obstacles generally feel less overwhelming when you can exchange ideas with others. The Ed discussion board will also help us learn from each other and help troubleshoot data and coding issues.

Accessing Course Materials

Courseworks will used for weekly quizzes and submitting assignments, but all other course materials will be shared via the website and the Ed discussion board. Bookmark these external sites, as you'll be visiting them regularly.

Towards an Anti-Racist Learning Experience

Every course should be an anti-racist course, quant courses included. In this course we'll cover examples that reflect systemic gaps based on race, ethnicity, immigration status, and gender identity, among other aspects of personal identity. It is critical to acknowledge that the social and economic marginalization reflected in the data is rooted in systemic oppression that upholds opportunity for some at the expense of others. We should all be thinking about our own role in upholding these systems. Over the course of the semester, we'll engage in discussions to help challenge our own notions about how to use "data for good" – both during class time and via the Ed discussion board.

Assignments, Grading and Course Requirements

Six Weekly Assignments (Data Memos)

[25% of your overall grade]

Weekly assignments are due by midnight on Monday night. Assignments will be graded on a check plus/minus scale. Late submissions will not receive a grade as we will be discussing solutions during class.

Courseworks Quizzes on Pre-class Material

[10% of your overall grade]

CW quizzes at the beginning of scheduled class meetings will account for 10% of your total grade. These *closed-note* quizzes will consist of multiple choice questions, and are designed to encourage you to engage with the pre-class lessons in advance of class so we can put new R functionality to work in class and focus on application and discussion. Note that CW keeps a log of your quiz activity and allows the instructor to see if/when you leave your quiz browser tab; yet another reason to follow the academic integrity policies described below. Also note that your lowest quiz score will be automatically dropped when calculating your course grade.

Project Work

[50% of your overall grade]

Your project grade will include an-class presentation of your work to-date near the end of the semester (15% of your total grade), a report or policy brief (25% of your total grade), and a short peer evaluation from your fellow group members to ensure everybody is making both coding and non-coding contributions (10%). The data project will also involve three required meetings with the teaching team for project advising, and include several intermediate deliverables: (1) submitting research ideas; and (2) and a proposal with summary statistics. Intermediate deliverables will not receive their own grade, but late intermediate submissions will result in a one grade deduction from your overall project grade for every day late (e.g. from an A to A-).

Participation, discussion and experimental input

[15% = 5% in-class participation + 3% Categorizing Identity response + 7% experimental input]

Students are required to attend weekly class sessions, participate in the discussion both in-class and through Ed, as well as reply to one required discussion board prompt ("Categorizing Identity"). This component can make the difference between an A and B for your overall course grade, for example, so please come to class prepared and ready to participate. Multiple *unexcused* absences may result in additional deductions to your overall course grade beyond any deductions for forgone participation.

An experiment on student belonging. This year I was named as a Provost's Senior Faculty Teaching Scholar, for which I am tasked with leading a project to advance teaching and learning at the university. My project will explore student belonging by trying to identify moments over the course of the semester when students feel more or less of a sense of belonging in the course and engagement with the material. This information will inform improvements to course design intended to ensure that more students feel like they truly belong in the course and can engage more deeply and consistently. It will also serve as a model for gauging student belonging in other courses.

I will be providing a link for you to periodically check-in by answering one or two brief questions on your own sense of belonging in the course. Students will be encouraged to complete a check-in as often as they would like, with a minimum of 3 check-ins required to receive all 7 points for this component of your overall grade (7% of the total grade). Responses will be periodically downloaded by the instructor to award credit for completing each check-in, and then anonymized before reading any substantive information.

Please reach out to the instructor if you have any suggestions or concerns about this experiment.

Course Policies

Illness and Personal Challenges

If you are feeling unwell or dealing with hardships outside of class, please email me *in advance* of class and/or any looming deadlines. We will do our best to help you catch up on any material you miss and come up with a plan for submitting any upcoming assignments.

I encourage all students to reach out to me as early as possible when dealing with personal challenges, mental or physical health issues. Being a grad student in today's world can be a demanding and stressful experience, I'm happy to chat about possible accommodations and teaching team support to help you manage course-related anxiety.

Academic Integrity and Generative AI

Generative AI

SIPA does not tolerate cheating or plagiarism in any form. Students who violate the Code of Academic & Professional Conduct will be subject to the Dean's Disciplinary Procedures.

Students who violate this code or the AI and Collaboration guidance below will be referred to Student Affairs and are subject to the Dean's Discipline Policy and Procedures.

*AI-generated code is **NOT PERMITTED** for weekly assignments.* We provide most of the code you need and review all of the functions you need in class and/or the pre-class lessons; learning to modify and augment the code we provide on your own, developing coding logic and debugging intuition are all central goals of these assignments. One can't reliably use AI tools to avoid syntax errors without first developing a foundational understanding of the syntax.

AI tools are PERMITTED for your project work in certain instances as long as this use is properly documented. Permissible and constructive uses include: (1) help with project brainstorming, i.e. using AI tools for enhanced 'Googling' to help identify sources of policy variation, data sources, and relevant research to date; or (2) help with R packages, functions, and more complex coding logic that have *not* been covered in the class materials.

*Using AI tools to interpret regression results and assess internal or external validity is **NOT PERMITTED** nor is it constructive.* The risks of doing so outweigh the potential benefits: recent project submissions that have used AI tools for regression interpretation were readily identifiable by the teaching team for missing the mark and focusing on the wrong information. Using AI to interpret and critique your work is an impediment to developing your own econometric intuition and becoming more comfortable discussing econometrics in both technical and policy-relevant language.

Collaboration

Students must write up Assignments independently, but you are encouraged to discuss with others along the way. While grading your assignments, if we come across answers to any parts of any assignments that are clearly not your own words, all involved parties will receive a zero for those parts and may be referred to Academic Affairs if appropriate.

Disability Accomodations & Neurodiversity

While there are no exams for this course, I am happy to provide all relevant disability accommodations outlined by Student Affairs for students registered with Columbia University's Disability Services (DS). Students who are registered with DS are encouraged to make an appointment with me at the start of the semester to make sure we have a plan for supporting your needs that you feel good about.

Students who are not registered with DS are also encouraged to make an appointment to chat with me about any personal learning challenges. Our neurological diversity and different learning styles is an asset, and our collective learning experience is enhanced when we all feel supported individually.

Academic Freedom in the Classroom

Knowledge flourishes when inquiry is free and respectful. This class, which has been approved as part of the Columbia curriculum by appropriate faculty bodies, aims to advance knowledge through discussion, debate, and carefully selected readings and assignments. In accordance with principles of academic freedom promulgated by the American Association of University Professors and affirmed by many universities, including Columbia, the instructor has the authority to set the class syllabus, which may include controversial material relevant to topics being studied. While all participants and their views will be treated respectfully, no one should expect to be shielded from challenging or even upsetting ideas, since thoughtfully engaging such ideas is crucial to free inquiry and intellectual growth.

Student Support and Teaching Team Communication

Given the large of number of student inquiries, we ask that you rely on scheduled office hours and the Ed discussion board as much as possible. We'll do our best as a teaching team to respond to inquiries within 72 hours, typically sooner.

Course Schedule

Please note this syllabus and due dates are subject to change with appropriate notice.

Week 1, 2 Sep 2025: Introduction, R Basics and Workflow

- In-class data: gapminder country-level data on life expectancy
- *Assignment 1 posted after class: due before 11:59pm on Monday, 8 Sep*

Week 2, 9 Sep 2025: Data Types & Structures, R Markdown, Intro to the Tidyverse

- In-class data: Caste-based expenditure gaps in India
- *Assignment 2 posted after class: due before 11:59pm on Monday, 15 Sep*

Week 3, 16 Sep 2025: Data Management I

- In-class data: Brooklyn subway fare evasion arrest data part 1: cleaning and analyzing the microdata
- *Assignment 3 posted after class: due before 11:59pm on Monday, 22 Sep*

Week 4, 23 Sep 2025: Data Mgmt. II, Inference, Working with Data on Personal Identity

- In-class data: Brooklyn subway fare evasion arrest data part 2: analyzing enforcement patterns across subway stations
- *Assignment 4 posted after class: due before 11:59pm on Monday, 29 Sep*

Week 5, 30 Sep 2025: Research Design, Regression Analysis, Weighting

- In-class data: Brooklyn subway fare evasion arrest data part 2: analyzing enforcement patterns across subway stations
- *Assignment: post reaction to Categorizing Identity reading to Ed by Monday, 6 Oct*
- *Project deliverable: Find 2 partners for your data project by Tuesday, 7 Oct*

Week 6, 7 Oct 2025: Data Visualization with ggplot

- In-class data: World Bank Gender Statistics: Maternal leave policy and women in STEM
- *Project deliverable: Submit 2 possible research questions by Monday, 13 Oct*
- *Sign up for required project meeting #1 w/instructor during Week 7 (meet by Mon, 20 Oct)*

Week 7, 14 Oct 2025: Working with Panel Data (Part 1: Descriptive Analysis)

- In-class data: Detroit water shutoffs
- *Assignment 5 posted after class: due before 11:59pm on Monday, 20 Oct*

Week 8, 21 Oct 2025: Working with Panel Data (Part 2: Two Way Fixed Effects)

- In-class data: Detroit water shutoffs and public health impacts
- *Project deliverable #2: Proposal with summary statistics due by 11:59pm on Mon, 27 Oct*
- *Sign up for required project meeting #2 with instructor (meet by Fri, 7 Nov)*

Week 9, 28 Oct 2025: String & Date Variables, Mapping with R

- In-class data: Indonesian Village Fund allocations
- *Assignment 6 posted after class: due before 11:59pm on Monday, 10 Nov*

Week 10, 4 Nov 2025: ELECTION DAY BREAK

Week 11, 11 Nov 2025: Difference-in-Differences with Staggered Treatments

- In-class data: Gun-related deaths and Stand Your Ground laws (Cheng & Hoekstra 2013)
- *Sign up for required project meeting #3 with TA (meet by Fri, 21 Nov)*

Week 12, 18 Nov 2025: Formatted Output and R Markdown Presentations

Week 13, 25 Nov 2025: Group Project Presentations

[The lunch break will also be used for presentations, please plan accordingly.]

Week 14, Tues 2 Dec 2025: Web Scraping and Dynamic Visualizations

- In-class data: Police killings in Jamaica

Final papers due Fri, 12 Dec 2025

Summary of due dates

Deliverable	Posted	Due
Assignment 1	Tues, 2 Sep	Mon, 8 Sep
Assignment 2	Tues, 9 Sep	Mon, 15 Sep
Assignment 3	Tues, 16 Sep	Mon, 22 Sep
Assignment 4	Tues, 23 Sep	Mon, 29 Sep
Reaction to categorizing identity	Tues, 29 Sep	Mon, 6 Oct
Project: find 2 partners	–	Tues 7 Oct
Project: submit 2 research Q's	–	Mon, 13 Oct
Assignment 5	Tues, 14 Oct	Mon, 20 Oct
Project: proposals due	–	Mon, 27 Oct
Assignment 6	Tues, 28 Oct	Mon, 10 Nov
Project: presentations	–	Tues, 25 Nov
Project: reports due	–	Fri, 12 Dec