PP 5397 Special Topics: Data Analytics with R for Policy & Management

Spring 2024

School of Public Policy

University of Connecticut

Class hours and location: Tuesday, 6:30 – 9 pm, HTB 220

Office hours: By appointment, see https://calendly.com/dmitre/officehours

Instructor: David Mitre Becerril (pronunciation: Me·tre Beh·se·reel)

Email: david.mitre@uconn.edu

Overview

This course will provide you with the computational tools and techniques necessary to acquire, organize, and visualize complex data to answer social science research questions. The course will cover topics on basic programming, data cleaning, and more advanced data processing techniques such as text mining, regular expressions, geocoding, mapping, web scraping, and interactive visualizations. The course will use R, an open-source, object-oriented scripting language. It is assumed that students have no previous R programming experience.

Most datasets are messy, unstructured, and require substantial processing. Surveys suggest that analysts can spend nearly 80% of their time cleaning data.[1] Accordingly, learning how to perform these tasks in an efficient, structured, and replicable way can be a good investment for those aiming to conduct social science research. In short, think of this course as an opportunity to learn the skills needed to gather raw data from multiple sources, process it, and present it in an understandable format ready for statistical analysis.

Course objectives

This course will help you develop the skills in R programming needed to process and visualize complex data used in social science research. At the end of the semester, students should gain:

- An understanding of the data analytics pipeline.
- An ability to acquire, organize, and transform data using R.
- An ability to effectively visualize and present quantitative data.

Class format

We will meet regularly in person and, on exceptional occasions, online, as described in the class plan. This course follows a modified **flipped classroom**: the students are expected to watch the weekly R programming online lectures and solved a short quiz before class. Class time is used to work on hands-on exercises and in-depth examples.

Use of technology and required software

We will use **RStudio** for all the statistical analysis and data visualization. Students are responsible for having **RStudio installed** on their computers or electronic devices by **the first week of classes**. RStudio is an open software, available for free. Students should refer to the UConn Software

[1] See for example https://www.emerald.com/insight/content/doi/10.1108/JD-08-2021-0167/full/html, https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=2f36ae5c6f63

Catalog website (Free/Open Source >> RStudio) for the instructions on how to install it on their computers https://software.uconn.edu/software/?licenseSelect=Free+%26+Open-Source

Students are expected to bring fully charged laptops to class to solve the in-class exercises and engage in the class. Students should avoid using their electronic devices for non-class topics, as it can distract their classmates and affect everyone's grades.[2]

Course communication

All course announcements and materials will be posted on the course website on HuskyCT. For any class questions, please **email** me and **include "PP 5397" on the subject** for a timely response. Expect responses within 24 hours on weekdays and 48 hours on weekends. You are also welcome to come to office hours. Students are encouraged to contact me for advice on any class question.

Textbooks

There are **no required textbooks**. Students looking for additional material beyond class content may find the following useful:

- Wickham, H., Çetinkaya-Rundel, M. & Grolemund, G., (2023) *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly. 2nd Edition. Authors' free online version: https://r4ds.hadley.nz/

Evaluation

Grades will be based on problem sets, quizzes, a group project, and participation.

- **Problem sets**. There will be **two take-home problem sets**. Students can work individually or in pairs (submit only one problem set per group). The problem sets will consist of data tasks solvable using RStudio. They will be an application of the knowledge learned in class.
- **Quizzes**. There will be **11 take-home, open-book quizzes**. Students will solve these quizzes individually after watching the weekly recorded lecture but before class. Students can take the quizzes twice until the due date; the highest grade will be counted. The lowest student's quiz score will be dropped from the final grade. The quizzes will assess your understanding of R programming.
- **Group project**. The students will present a group project at the end of the semester. Students will work in groups of two or three people. The project will allow students to apply the skills learned in class to acquire, process, and visualize of a dataset chosen by the group. It will include two components:
  - **Proposal**. The group will submit a one-page document stating their proposed topic based on a dataset conducive to data cleaning, processing, and visualization and explain their motivation, data cleaning tasks to perform, and data sources.
  - **Presentation**. Students will communicate and present the data project to the class, summarizing their main data cleaning and processing tasks and the most relevant findings.
- **Participation**. Students will present their solutions to the in-class exercises during the regularly scheduled class time. **Students should make at least 10 correct participations during the semester for a full participation grade**. Absences will impede your ability to participate.

[2] About the effect of in-class use of electronics on school performance, see https://doi.org/10.1016/j.econedurev.2017.02.004, https://doi.org/10.1080/01443410.2018.1489046, https://doi.org/10.1177/1469787417721382

Evaluation method
Your grade will be determined as follows:

Assignment	Percentage	Due date
Problem set 1	20	3/10
Problem set 2	20	3/31
Concept quizzes	20	1/22, 1/29 2/5, 2/12, 2/19, 2/26 3/4, 3/18, 3/25, 4/1, 4/8
Group project		
Proposal	5	3/23
Presentation	15	4/23
Participation	20	4/16
Total	**100**	

Late Assignments
All assignments are due at 11:59 pm EST on the due date. Late assignments will result in a 10% cumulative grade deduction per late day (e.g., 10% for one day late, 20% for two days late, 30% for three late days, etc.) up to a 100% deduction. Contact me ahead of time if there are any extenuating circumstances.

Re-grade policy
If you wish to have an assignment regraded, let me know within one week after you receive it. Re-grading an assignment can increase or decrease your grade. If there was an arithmetic error in adding points to your assignment, let me know immediately, and I will correct it.

Grading Scale

Grade	Letter Grade	GPA
93-100	A	4.0
90-92	A-	3.7
87-89	B+	3.3
83-86	B	3.0
80-82	B-	2.7
77-79	C+	2.3
73-76	C	2.0
70-72	C-	1.7
67-69	D+	1.3
63-66	D	1.0
60-62	D-	0.7
<60	F	0.0

Students with disabilities
Please contact me to discuss academic accommodations needed during the semester due to a documented disability. The University of Connecticut is committed to protecting the rights of individuals with disabilities and assuring that the learning environment is accessible. If you anticipate or experience physical or academic barriers based on disability or pregnancy, please let me know immediately to discuss options. Students who require accommodations should contact the Center for Students with Disabilities, Wilbur Cross Building Room 204, (860) 486-2020 or http://csd.uconn.edu/.

Academic integrity
Plagiarism, cheating, and other forms of academic misconduct will not be tolerated. All work that you submit for credit during this course must represent your own work and no one else's. Students should be especially careful in their writing to properly cite material and ideas taken from other sources. A link to the policy on scholarly integrity for graduate students may be found at https://provost.uconn.edu/faculty-and-staff-resources/syllabi-references/.

You can use AI writing tools such as ChatGPT on assignments (I'll alert you when you cannot). Whenever you use them, you must include a brief acknowledgment stating that it and how you used them. Note that all large language models still tend to make up incorrect facts and fake citations. You will be responsible for any inaccurate, biased, offensive, or otherwise unethical content you submit, regardless of whether it originally comes from you or an AI tool.

Disclaimer
Syllabus information may be subject to change, except for materials for purchase. The most up-to-date syllabus is located on the course website on HuskyCT.

Weekly course plan

Week **Date**	**Topic**
Week 1 1/16	**Introduction** ▪ Syllabus review & expectations **Data analytics** ▪ Data analytics pipeline ▪ Introduction to R and RStudio ▪ Getting help in R ▪ Errors vs Warnings ▪ Reproducible examples ▪ OpenAI and coding
Week 2 1/23	**Operations in R** ▪ Arithmetic operations ▪ Logical operations ▪ Missing values ▪ Code syntax ▪ Loading and installing libraries **Data types and structures** ▪ Data frames (import/export data) ▪ Variables and data types ▪ Vectors ▪ Lists
Week 3 1/30	**R Programming Basics** ▪ Conditional statements ▪ Loops ▪ While ▪ Functions
Week 4 2/06	**Data wrangling** ▪ Rearranging rows and columns ▪ Rename and create new variables ▪ Summarize data
Week 5 2/13	**Data wrangling** ▪ Conditional data changes ▪ Merging datasets ▪ Lengthening and widening datasets
Week 6 2/20	**Date and times** ▪ Date-time components

	▪ Time spans
Week 7 2/27	**Text mining** ▪ Working with text ▪ Regular expressions ▪ Detecting patterns
Week 8 3/5	**Complex data manipulations** ▪ Custom functions ▪ Evaluating text as code ▪ Automatizing tasks
Week 9 3/12	**No class - Spring recess**
Week 10 3/19	**Web scraping** ▪ Terms of service ▪ Web-scraping vs APIs ▪ Parsing HTML content
Week 11 3/26	**Spatial data** ▪ Geometry operations ▪ Mapping ▪ Geocoding
Week 12 4/2	**Data visualizations** ▪ Static visualizations ▪ Interactive visualizations ▪ Exploratory data analysis
Week 13 4/9	**Interactive reporting** ▪ Rmarkdown basics ▪ Text formatting ▪ Reproducible reports
Week 14 4/16	**TBD** ▪ Review session** ▪ Advanced data analytics topics** ▪ Group project work time** **Depending on class progress
Week 15 4/23	**Group presentations**