Data Science for Public Policy

PPOL 670-01

Wednesday 6:30 – 9 PM
January 19 – May 6
Spring 2022

Instructor

Aaron R. Williams (he/him)
arw109@georgetown.edu

Teaching Assistant

Nicholas Stabile (he/him)
ncs31@georgetown.edu

Course Description and Learning Goals

This course aims to provide a broad foundation of modern data science skills, while also exposing students to specific applications and challenges of these methods in policy analysis and governance. Aiming to complement training in econometrics and statistical inference, this course will primarily use the programming language R, which we will use for data analysis, visualization, literate programming, collecting data from the web, geospatial analysis, machine learning, and data imputation. Students will also be exposed to Git and GitHub for reproducible research, using the cloud for big data analysis, and SQL for interacting with databases.

This class is highly applied and is meant to prepare students for roles in either computational social science or applied data science in the public sector. Students will have the opportunity to work with real data on relevant policy problems. They will train to analyze data with a broader set of methods, and to tell better and more compelling stories with data. No prior programming experience is required or assumed.

Course Schedule

Topic 0: Syllabus and introductions
Topic 1: Introduction to R
Topic 2: Introduction to the tidyverse
Topic 3: Data visualization
Topic 4: Reproducible research with R Markdown
Topic 5: Reproducible research with Git
Topic 6: More R programming and APIs
Topic 7: Exploratory Data Analysis
Topic 8: Geospatial analysis
Topic 9: Supervised machine learning -- regression
Topic 10: Supervised machine learning -- classification

Topic 11: Unsupervised machine learning -- dimension reduction
Topic 12: Unsupervised machine learning -- cluster analysis
Topic 13: Introduction to the cloud and AWS S3
Topic 14: AWS EC2
Topic 15: Data imputation techniques in R

Flex Topics

- Text analysis
- Web scraping
- SQL
- Advanced Cloud Computing
- Advanced Git

[Week 1] 2022-01-19: Topic 0 and Topic 1
[Week 2] 2022-01-26: Topic 2
[Week 3] 2022-02-02: Topic 3a and Topic 4
[Week 4] 2022-02-09: Topic 3b and Topic 5
[Week 5] 2022-02-16: Topic 6a and Topic 7
[Week 6] 2022-02-23: Topic 6b and Topic 8a
[Week 7] 2022-03-02: Topic 8b and Topic 9a
[Week 8] 2022-03-16: Topic 9b and Topic 10a
[Week 9] 2022-03-23: Topic 10b and Topic 11
[Week 10] 2022-03-30: Topic 12 and Topic 13a
[Week 11] 2022-04-06: Topic 13b and Topic 14a
[Week 12] 2022-04-13: Topic 14b and flex topic
[Week 13] 2022-04-20: No Class During Class Time -- Project Meetings and Asynchronous Flex Topic
[Week 14] 2022-04-27: Topic 15
[Finals Week] 2022-05-06 7:00-9:00 PM: Project Presentations

Office Hours

We will host office hours at least twice per week. We will determine the office hours schedule based on polling during the first week of class. Students are welcome to attend office hours for either 670-01 (this section) or 670-02. The office hours for both sections will be shared after the first week of class.

We will manage a combined Slack channel for both sections of the course. This channel is for asking clarifying questions, sharing materials related to the course, collaborating with students, and asking questions about assignments and projects. Questions must be asked using reproducible examples. Simply copying-and-pasting questions or answers in the Slack channel is not allowed.

Course Materials, Books, and Resources

All the books and readings for this course are available at no cost to students. We will be drawing significantly from the three books listed below, but they are available free online and you should feel no obligation to purchase them. Additionally, we will draw from a wide range of free online resources that will be posted on Canvas. **Students are expected to check the course Canvas site each week to see required readings.**

Wickham, H., & Grolemund, G. (2016). "R for data science: import, tidy, transform, visualize, and model data". O'Reilly Media, Inc. [Link]

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). "An Introduction to Statistical Learning: with Applications in R". New York: springer. [Link]

Boehmke, B. & Greenwell, B, (2020). "Hands-On Machine Learning with R." CRC Press. [Link]

Amazon Web Services

We will use Amazon Web Services (AWS) to learn how to use cloud services, including data storage (AWS S3) and individual cloud servers (AWS EC2). Students should be able to complete all course assignments using free AWS student credits, available here: https://aws.amazon.com/education/awseducate/

I will provide detailed instructions when necessary.

Software

This list covers all the software we will use throughout the course. You will be given explicit instructions for any downloads, installations, or other preparation you will need to do, so there is no need to do anything in advance of those instructions.

- R – A popular statistical programming language
- RStudio - A free and powerful development environment for R
- Git (GitBash for Windows) - a modern and flexible version control system that we will use for disseminating and submitting assignments. You will also need to sign up for a free account on GitHub.com.
- SQL - Structured Query Language, a type of language
- Jupyter Notebooks - a software interface in which we will write code and see results when we are working in the cloud, specifically with Apache Spark.

Course Assignments

Problem Sets There are going to be at least eight problem sets in this course, covering the most critical class topics. *Due dates are subject to change.*

1. Introduction to R, Due Tuesday, January 25th
2. Introduction to the tidyverse, Due Tuesday, February 1st
3. Data Manipulation and Analysis in R, Due Friday, February 11th (6:00 PM)
4. Data Visualization, Literate Programming, & Reproducible Research, Due Friday February 25th (6:00 PM)

5. APIs and Geospatial Visualization, Due Friday, March 18th (6:00 PM)
6. Supervised Machine Learning - Part I, Due Friday, March 25th (6:00 PM)
7. Supervised Machine Learning - Part II, Due Friday, April 1st (6:00 PM)
8. Introduction to Data Analysis in the Cloud, Due Tuesday, April 19th

Assignments 4-8 will contain "stretch" exercises. Students must complete the stretch exercises for at least two of those five assignments. At least one stretch exercise must be from assignments 4, 5, or 8. The lowest stretch exercise grades will be dropped if more than two stretch exercises are submitted (i.e. only the two highest stretch exercises will be considered if four stretch exercises are completed).

Course Project - Due May 6th
In the final weeks of the class, each student will execute a group project, working through a real world data science analysis of public policy relevant data. More details will be discussed in class and will be available on Canvas. There will be additional due dates for proposals and check-ins during the semester.

Course Expectations

Attendance & Participation
Students are expected to attend every class session. More than one unexcused absence may lead to grade penalties. Students are expected to turn on cameras if joining class virtually.

Late Policy
Please let me know, as early as possible, if you are unable to complete an assignment on time. I will be reasonably accommodating for late assignment submissions; however, without advance notice, a late penalty of -0.5 points per day will be assessed.

Attendance Policy

Students are expected to attend every class session if at all possible. Please inform me if you know in advance that you will not be able to make it to class. This course contains difficult material that builds quickly upon the prior weeks – missing classes will add significant burden to students and should be avoided whenever possible.

I will do my best to adhere to the guidance and best practices put forward by Georgetown University for instructional continuity. The conditions for this semester are less than ideal because of COVID-19. Please let me know if the instructional continuity is inadequate for your needs and **please stay home if you are experiencing COVID-related symptoms or if you have a close contact with someone who tests positive for COVID-19.**

Homework Collaboration Policy

Learning and data science are both collaborative practices. I encourage you to discuss class topics and homework topics with each other. However, the work you submit must be your own. A student should never see another student's code or receive explicit coding instructions for a homework problem (see exception below). Please attend office hours or contact one of the

instructors if you need help or clarification. Plagiarism on homework or projects will be dealt with to the full extent allowed by Georgetown policy (http://honorcouncil.georgetown.edu).

Assignments 4-8 can be done with a lab partner from your section. Each partnership will only need to submit one assignment and partners can share code and directly discuss the exercises.

Grading Schema

93.0 — 100	A
90.0 — 92.9	A-
87.0 — 89.9	B+
83.0 — 86.9	B
80.0 — 82.9	B-
70.0 — 79.9	C
Below 70	F

Use of Class Materials

Increasingly, with the proliferation of certain websites, questions about the ownership of course materials have arisen (and Georgetown is actively working on policies to address these concerns). I consider the syllabus, lectures, handouts, problem sets, and problem set answers to be my intellectual property. I respectfully request that you refrain from sharing the materials in any electronic (or paper) format without my permission. Sharing notes, on an occasional basis, with others in the class is fine as long as they are not posted.

Academic Resource Center/Disability Support

If you believe you have a disability, then you should contact the Academic Resource Center (arc@georgetown.edu) for further information. The Center is located in the Leavey Center, Suite 335 (202.687.8354). The Academic Resource Center is the campus office responsible for reviewing documentation provided by students with disabilities and for determining reasonable accommodations in accordance with the Americans with Disabilities Act (ADA) and University policies.

Important Academic Policies and Academic Integrity

McCourt School students are expected to uphold the academic policies set forth by Georgetown University and the Graduate School of Arts and Sciences. Students should therefore familiarize themselves with all the rules, regulations, and procedures relevant to their pursuit of a Graduate School degree. The policies are located at: http://grad.georgetown.edu/academics/policies/.

Provosts Policy Accommodating Students Religious Observances

Georgetown University promotes respect for all religions. Any student who is unable to attend classes or to participate in any examination, presentation, or assignment on a given day because of the observance of a major religious holiday (see below) or related travel shall be excused and provided with the opportunity to make up, without unreasonable burden, any work that has been missed for this reason and shall not in

any other way be penalized for the absence or rescheduled work. Students will remain responsible for all assigned work. Students should notify professors in writing at the beginning of the semester of religious observances that conflict with their classes. The Office of the Provost, in consultation with Campus Ministry and the Registrar, will publish, before classes begin for a given term, a list of major religious holidays likely to affect Georgetown students. The Provost and the Main Campus Executive Faculty encourage faculty to accommodate students whose bonafide religious observances in other ways impede normal participation in a course. Students who cannot be accommodated should discuss the matter with an advising dean.

Statement on Sexual Misconduct

Georgetown University and its faculty are committed to supporting survivors and those impacted by sexual misconduct, which includes sexual assault, sexual harassment, relationship violence, and stalking. Georgetown requires faculty members, unless otherwise designated as confidential, to report all disclosures of sexual misconduct to the University Title IX Coordinator or a Deputy Title IX Coordinator. If you disclose an incident of sexual misconduct to a professor in or outside of the classroom (with the exception of disclosures in papers), that faculty member must report the incident to the Title IX Coordinator, or Deputy Title IX Coordinator. The coordinator will, in turn, reach out to the student to provide support, resources, and the option to meet. [Please note that the student is not required to meet with the Title IX coordinator.]. More information about reporting options and resources can be found on the Sexual Misconduct Website: https://sexualassault.georgetown.edu/resourcecenter.

If you would prefer to speak to someone confidentially, Georgetown has a number of fully confidential professional resources that can provide support and assistance. These resources include:

Health Education Services for Sexual Assault Response and Prevention: confidential email sarp@georgetown.edu

Counseling and Psychiatric Services (CAPS): 202.687.6985 or after hours, call (833) 960-3006 to reach Fonemed, a telehealth service; individuals may ask for the on-call CAPS clinician

More information about reporting options and resources can be found on the Sexual Misconduct Website.