

PPOL 670-01: Introduction to Data Science



Instructor: Alexander Podkul, Ph.D.

Meeting Location: McCarthy Hall MCSHNSML

Meeting Times: Mon. 6:30p - 9:00p

Email: arp52@georgetown.edu

Course site: <http://ppol670.alexanderpodkul.com/>

Office Hours: By appointment/announcement
(in-person/virtual)

Course Description and Objectives

This course aims at introducing students to data science applications of public policy research. After introducing students to some of the key foundations of data science, the semester will take students through a survey of various topics including: data wrangling, visualization techniques, data collection, statistical learning, and machine learning. This course emphasizes applied data science and students will leave the course with the ability to produce their own public policy research findings developed using data science. Students will also leave the class as thoughtful consumers of data science methodologies used within their research areas. While the course will be taught predominately in the R statistical programming language, no prior experience is required or expected of students entering the class. The primary objective of this course is to give public policy students a data science toolkit that will allow them to pursue advanced topics both in- and outside the classroom.

Prerequisite Courses

Students are required to have completed either PPOL 502 or PPOL 532. Students who have not taken either of these courses are required to get permission from the instructor.

Required Texts

There are two required texts for the class:

1. Magallanes Reyes, Jose Manuel. 2017. *Introduction to Data Science for Social and Policy Research: Collecting and Organizing Data with R and Python*. Cambridge University Press.
2. Wickham, Hadley and Garrett Grolemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly. (Note: While students are welcome to buy the hard copy of this text, it is also available on the web for free at: <https://r4ds.had.co.nz>.)

All other readings will be available for download on the course website. Additional suggested text resources to supplement scheduled course readings can be found under the “Resources” section of the course website.

Assignments

The major assignments for this course are:

- **Problem Sets** – Throughout the course of the semester, there will be five problem set assignments, which include deep dives into producing your own data science work. These assignments will range from replication to producing original work. Each problem set will be assigned one week before it is due. Problem sets are required to be submitted in R Markdown format rendered as an .html file. The full submission should include a zip file containing: the code (.rmd file), the output (.html file), and any data sets used.
- **Final Project** – The largest project of the semester is the final project, which includes a proposal, a presentation, and report. The full project will be formally assigned during the class session on 9/20/2021 (including due dates and other details) after the course has introduced the basics of data science for public policy.
- **Participation** – Participation includes attending class meetings, participating in in-class discussions, asking thoughtful questions, and providing thoughtful suggestions to classmates’ final projects.

Assignment	Share of Final Grade
Problem Sets	50%
Final Project	40%
Participation	10%

Grading

Final course grades will follow:

Grade	Range
A	> 95%
A-	91% - 94%
B+	87% - 90%
B	84% - 86%
B-	80% - 83%
C	70% - 79%
F	< 70%

Course Policies

Notes on Interruptions and Instructional Continuity

In the event that our in-person semester gets interrupted for any reason, we will pivot to a virtual setting (Zoom) that will continue to meet during regularly scheduled course times. If any individual course meeting is affected by a school closure or other emergency, we will pivot to a virtual session as necessary. If for some reason it is not possible to pivot to a virtual session, course content for that session will be delivered via asynchronous format later in the semester. Any changes to regularly scheduled course meetings will be communicated by Canvas Announcement.

Late Policy

It is expected that students will submit assignments according to the course schedule. Students can request extensions under reasonable circumstances. For problem sets, students should submit extensions at least five days before the due date, if possible. Although late work will be accepted, any tardy work without an approved extension will lose a letter grade for each day late. (e.g. Between 0 to 1 days loses one letter grade, between 1 to 2 days loses two letter grades, etc.)

Office Hours

The course professor will be available for office hour appointments by appointment and by announcement throughout the semester. Office hours are available both in-person and virtually. To request an appointment, send an e-mail to the instructor at arp52@georgetown.edu including a few possible meeting times.

Software

The primary software program used in this course will be R. Students will be encouraged to use the R Studio IDE for producing course outputs such as R markdown files (.html/.pdf) for problem sets and final projects. No prior experience with R is necessary for this course. R is available to download for free at <https://cran.r-project.org>. R Studio is also available for free at: <https://rstudio.com/products/rstudio/download>. Installation details will be covered during the first course session.

Technology Policy

The use of technology is encouraged for this course. However, students are expected to use only course-related materials (Zoom, R, etc.) during active course sessions. It is expected that students will refrain from cell phone use during course meetings.

Academic Resource Center/Disability Support

If you believe you have a disability, 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 (ASA) and University policies. For more information, go to <http://academicsupport.georgetown.edu/disability/>.

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/>

Provost's 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 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 bona fide religious observances in other ways impede normal participation in a course. Students who cannot be accommodated should discuss the matter with an advising dean.

Title IX/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.

Class Materials Use

Considering the course syllabus, lectures, handouts, and problem sets as intellectual property, it is requested that students refrain from sharing course materials in any electronic or paper format without permission. Though sharing materials with others in class is acceptable, posting them online is unacceptable. If students have any questions, do not hesitate to reach out to the course instructor.

Weekly Topics and Assignments Schedule

Weekly readings will be posted at least one week before each class session on the course website.

Week 1 - Aug. 25, 2021 - Introduction of Data Science, Getting to Know R

Week 2 - Aug. 30, 2021 - Data Storage and Data Types

NO CLASS ON SEP. 6, 2021

Week 3 - Sep. 13, 2021 - Data Wrangling

Week 4 - Sep. 20, 2021 - Data Visualization Problem Set #1 Due

Week 5 - Sep. 27, 2021 - Data Collection Techniques (API's + scraping) Problem Set #2 Due

Week 6 - Oct. 04, 2021 - Working with Geospatial Data

NO CLASS ON Oct. 11, 2021

Week 7 - Oct. 18, 2021 - Statistical Thinking

Week 8 - Oct. 25, 2021 - Working with Regression

Week 9 - Nov. 01, 2021 - Classification Techniques Problem Set #3 Due

Week 10 - Nov. 08, 2021 - Machine Learning Methods Problem Set #4 Due

Week 11 - Nov. 15, 2021 - Text-as-Data

Week 12 - Nov. 22, 2021 - Missing Data/Working with Relational Databases (Primer to SQL)

Week 13 - Nov. 29, 2021 - *To be determined – Class interest* Problem Set #5 Due

Week 14 - Dec. 6, 2021 - Final Presentations

Final Project Due: Dec. 10, 2021

Any changes to the above schedule will be communicated to students during class sessions.