Syllabus: PPOL 5203 - Data Science I: Foundations

Data Science and Public Policy, McCourt School for Public Polcy, Georgetown University

Course Description

This first course in the core data science sequence teaches Data Science for Public Policy (DSPP) students how to synthesize disparate, possibly unstructured data in order to draw meaningful insights. Topics covered include the fundamentals of object-oriented programming in Python; literate programming; an introduction to algorithms and data types; data wrangling, visualization, and extraction; an introduction to machine learning methods, and text analysis. In addition, students will be exposed to Git and Github for version control and reproducible research. The objective of the course is to teach students how incorporate data into their decision-making and analysis. No prior programming experience is assumed or required.

Class Website: https://tiagoventura.github.io/ppol5203_fall_2023

Learning Goals

After completing this course, the students will be able to:

- General understanding of python's object oriented programming syntax and data structures.
- Competency using version control (Git/Github).
- Learn to manipulate and explore data with Pandas and other tools.
- General understanding of analyzing algorithms and data structures.
- Learn to extract and process data from structured and unstructured sources.
- Get some intuition of modeling text data in Python.
- Learn the basics of machine learning as a modeling approach.

- Learn basics of using SQL to query databases.

Instructors and TAs

Instructor

- Professor: Dr. Tiago Ventura
- Pronouns: He/Him
- Email: tv186@georgetown.edu
- Office hours:
  - Time: Every Thursday, 4pm - 6pm
  - Location: 125E, Office Numbers766

Teaching Assistants:

- Aastha Jha (DSPP Second-Year Student)
  - Email: aj935@georgetown.edu
  - Office Hours:
    - Every Wednesdays, from 1pm to 2pm.
    - Location:
- Shirui Zhou (DSPP Alumni)
  - Email: sz614@georgetown.edu
  - Office Hours:
    - Every Monday, from 1pm to 2pm
    - Location:

Our classes

Classes will take place at the scheduled class time/place and will involve a combination of lectures, coding walkthrough, breakout group sessions, and questions. We will start our classes with a lecture highlighting what I consider to be the broader substantive and programming concepts covered in the class. From that, we will switch to a mix of coding walk through and breakout group sessions.

For every lecture, you will have access to a notebook (`.ipynb`) covering the topics and code discussed in class. I will upload these materials (which I call lecture notes every day before the class starts). In addition, you will also have access (in at least a week in advance), of required

readings (book chapters, articles, blog posts or coding tutorials) for every class. What you will take from this class will be tremendously improved if you work through all these materials.

Note that this class is scheduled to meet weekly for 2.5 hours. I will do my best to make our meetings dynamic and enjoyable for all parts involved. We will take one or two breaks in each of our lecture.

Required Materials

Readings: We will rely primarily on the following text for this course.

- **McKinney, W. , 2022. Python for Data Analysis. O'Reilly Media, Inc.**(Online version: https://wesmckinney.com/book/).
- **Vanderplas, J.T., 2016. "Python data science handbook: tools and techniques for developers."** ***O'Reilly.*** (Online version: https://jakevdp.github.io/PythonDataScienceHand
- **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). "An Introduction to Statistical Learning: with Applications in R".** ***New York: springer.***
- **Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as data: A new framework for machine learning and the social sciences. Princeton University Press.**
- **Salganik, M. 2017. Bit by Bit: Social Research in the Digital Age. Princeton, NJ: Princeton University Press.**

Additional readings will be posted for each class and can be found on the course website. Most reading materials are open source and available via a link on the weekly schedule. Otherwise it can be found on Canvas.

Course Infrastructure

Class Website: A class website https://tiagoventura.github.io/ppol5203 will be used throughout the course and should be checked on a regular basis for lecture materials and required readings.

Class Slack Channel: The class also has a dedicated slack channel (ppol-564-fall-2024.slack.com). The channel serves as an open forum to discuss, collaborate, pose problems/questions, and offer solutions. Students are encouraged to pose any questions they have there as this will provide the professor and TA the means of answering the question so that all can see the response. If you're unfamiliar with, please consult the following start-up tutorial (https://get.slack.help/hc/en-us/articles/218080037-Getting-started-for-new-members). Please follow the ***invite link*** to be added to the Slack channel.

Canvas: A Canvas site (http://canvas.georgetown.edu) will be used throughout the course and should be checked on a regular basis for announcements. ll announcements for the assignments and classes will be posted on Canvas; they will not be distributed in class or by e-mail. Support for Canvas is available at (202) 687-4949

NOTE: Students are encouraged to run lecture code on their own machines. If you do not have access to a laptop on which you can install `python3`, please contact the professor and/or TA for assistance. Only `python3` will be used in this course.

Weekly Schedule

Weeks	Topic	Date
Week 1	Introductions, Installations, IDEs, Command line	29-Aug-23
Week 2	Version Control, Workflow and Reproducibility: Or just Git & GitHub	12-Sep-23
Week 3	Intro to Python - OOP, Data Types, Control Statements and Functions	19-Sep-23
Week 4	From Nested Lists to Data Frames	26-Sep-23
Week 5	Pandas I: Data Manipulation	3-Oct-23
Week 6	Pandas II: Advanced Data Manipulation and Visualization	10-Oct-23
Week 7	Scrapping: Drawing from (Un-)Structured Data Sources	17-Oct-23
Week 8	Text as data I: Data Mining	24-Oct-23
Week 9	Introduction to Statistical Learning	31-Oct-23
Week 10	Text as Data II: Topics + Supervised Models	7-Nov-23
Week 11	Invited Speaker: Dr. Eric Dunford (META) - Introduction to Algorithms + Coding Interview	14-Nov-23
Week 12	Training Machines and collecting data with Selenium	21-Nov-23
Week 13	SQL + Spark	28-Nov-23
Week 14	Final Presentations	5-Dec-23

Course Requirements

Assignment	**Percentage of Grade**
Participation/Attendance	5%
Coding Discussion	5%
Problem sets	50%
Final Project	40%

Participation and Attendance (5%):

Data science is an cooperative endeavor, and it's crucial to view your fellow classmates as valuable assets rather than rivals. Your performance in the following aspects will be considered when assessing this part of your grade:

- Active involvement during class sessions, fostering a dynamic learning environment.
- Contributions made to your group's ultimate project.
- Assisting classmates by addressing problem set queries through GitHub issues. Supporting your peers will enhance your evaluation in terms of teamwork and engagement
- Assisting classmates with slack questions, sharing interesting materials on slack, asking question, and anything that provides healthy contributtions to the course.

Coding Discussion(5%)

Every class will involve some lecture time, and some coding time. The coding time will be divided between me showing you things, and you working on small problem sets. These problem sets are purposefully constructed to help you understand the concepts we go through in class. You participation and involvement in these group exercises will also be part of your grade.

Problem Sets (50%)

Students will be assigned five problem sets over the course of the semesters. While you are encouraged to discuss the problem sets with your peers and/or consult online resources, **the finished product must be your own work**. The goal of the assignment is to reinforce the student's comprehension of the materials covered in each section.

The problems sets will assess your ability to apply the concepts to data that is substantially messier, and problems that are substantially more difficult, than the ones in the coding discussion in class.

I will distribute the assignment through a mix of canvas and github. The assignments can be in the form of a Jupyter Notebook (`.ipynb`). Students must submit completed assignments as a rendered `.html` file and the corresponding source code (`.ipynb`).

The assignments will be graded in accuracy and quality of the programming style. For instance, our grading team will be looking at:

- (i) all code must run;
- (ii) solutions should be readable
  - Code should be thoroughly commented (the Professor/TA should be able to understand the codes purpose by reading the comment),
  - Coding solutions should be broken up into individual code chunks in Jupyter/R Markdown notebooks, not clumped together into one large code chunk (See examples in class or reach out to the TA/Professor if this is unclear),

  - Each student defined function must contain a doc string explaining what the function does, each input argument, and what the function returns;

- (iii) Commentary, responses, and/or solutions should all be written in Markdown and explain sufficiently the outpus.
- (v) All solutions must be completed in Python.

The follow schedule lays out when each assignment will be assigned.

Assignment	Date Assigned	Date Due
No. 1	Week 2	Before EOD of Friday of Week 3
No. 2	Week 4	Before EOD of Friday of Week 5
No. 3	Week 6	Before EOD of Friday of Week 7
No. 4	Week 8	Before EOD of Friday of Week 10
No. 5	Week 10	Before EOD of Friday of Week 11

Final Project (40%): Data science is an applied field and the DSPP is particularly geared towards providing students the tools to make policy and substantive contributtions using data and recent computational developments. In this sense, it is fundamental that you understand how to conduct a complete analysis from collecting data, to cleaning and analyzing it, to presenting your findings. For this reason, a considerable part of your grade will come from a an independent data science project, *applying concepts learned throughout the course.*

The project is composed of three parts:

- a 2 page project proposal: (which should be discussed and approved by me)
- an in-class presentation,
- A 10-page project report.

Due dates and breakdowns for the project are as follows:

Requirement	**Due**	**Length**	**Percentage**
Project Proposal	October 31	2 pages	5%
Presentation	December 10	10-15 minutes	10%
Project Report	December 17	10 pages	25%

Important notes about the final project

- For the project proposal, you need to schedule a 30min with me at least a week before the due date. For this meeting, I expect you to send me a draft of your ideas. We will do the group assignment and start scheduling meetings by week 4, I will share with you a calendar invite to organize our meetings.

- For the presentation, You will have 10-15 minutes in our last class of the semester to present you project.
- **Take the final project seriously**. After you finish your Masters, in any path you take, you will need to show concrete examples of your portfolio. This is a good opportunity to start building it.
- Your groups will be randomly assigned.

Submission of the Final Project

The end product should be a github repository that contains:

- The raw source data you used for the project. If the data is too large for GitHub, talk with me, and we will find a solution
- Your proposal
- A README for the repository that, for each file, describes in detail:
  - Inputs to the file: e.g., raw data; a file containing credentials needed to access an API
  - What the file does: describe major transformations.
  - Output: if the file produces any outputs (e.g., a cleaned dataset; a figure or graph).
  - A set of code files that transform that data into a form usable to answer the question you have posed in your descriptive research proposal.
  - Your final 10 pages report (I will share a template later in the semester)

Of course, no commits after the due date will be considered in the assessment.

Grading

Course grades will be determined according to the following scale:

Letter	Range
A	95% – 100%
A-	91% – 94%
B+	87% – 90%
B	84% – 86%
B-	80% – 83%
C	70% – 79%
F	$< 70\%$

Grades may be curved if there are no students receiving A's on the non-curved grading scale.

Late problem sets will be penalized a letter grade per day.

Communication

- Class-relevant and/or coding-related questions, **`Slack` is the preferred method of communication**. Please use the general or the relevant channel for these questions.
- For private questions concerning the class, email is the preferred method of communication. All email messages must originate from your Georgetown University email account(s). Please use a professional salutation, proper spelling and grammar, and patience in waiting for a response. The professor reserves the right to not respond to emails that are drafted inappropriately. ***Please email the professor and the TA directly rather than through the Canvas messaging system.*** Emails sent through CANVAS will be ignored.
- I will try my best to respond to all emails/slack questions ***within 24 hours*** of being sent during a weekday. ***I will not respond to emails/slack sent late Friday (after 5:00 pm) or during the weekend until Monday (9:00 am)***. Please plan accordingly if you have questions regarding current or upcoming assignments.
- Only reach out to the professor or teaching assistant regarding a technical question, error, or issue after you made a good faith effort to debugging/isolate your problem prior to reaching out. Learning how to search for help online is a important skill for data scientists.

ChatGPT and others

In the last year, the internet was inundated with popularization of Large Language Models, particularly the easy use of ChatGPT. As a Data Scientist, LLMs will be part of your daily work. I see ChatGPT as Google on steroids, so I assume ChatGPT will be part of your daily work in this course, and it is part of my work as a researcher.

That being said, ChatGPT does not replace your training as a data scientist. If you are using ChatGPT instead of learning, I consider you are cheating in the course. And most importantly, you are wasting your time and resources. So that's our policy for using LLMs models in class:

- Do not copy the responses from chatgpt – a lot of them are wrong or will just not run on your computer.
- Use chatgpt as a auxiliary source.
- If your entire homework comes straight from chatgpt, I will consider it plagiarism.

If you use chatgpt, I ask you to mention on your code how chatgpt worked for you.

Electronic Devices

When meeting in-person: the use of laptops, tablets, or other mobile devices is permitted *only for class-related work*. Audio and video recording is not allowed unless prior approval is given by the professor. Please mute all electronic devices during class.

Georgetown Policies

Disability

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 (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/

Applied to this course, while I encourage collaboration on assignments and use of resources like StackOverflow, the problem sets will ask you to list who you worked on the problem set with and cite StackOverflow if it is the direct source of a code snippet.

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[at]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.

Provost's Policy on 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.