PPOL 564-01
Data Science I: Foundations
Fall 2021

Instructor

Professor: Eric Dunford

- **Office**: 404 Old North
- **Office Hours**: Mondays/Wednesdays 9:00am to 10:00am (EST) or by appointment
- **Email**: eric.dunford@georgetown.edu
- **Pronouns**: He/Him

Teaching Assistant: Madeline ("Maddie") Pickens

- **Office Hours**: by appointment
- **Email**: mp1595@georgetown.edu
- **Pronouns**: She/Her

Teaching Assistant: Chandler Dawson

- **Office Hours**: by appointment
- **Email**: cad162@georgetown.edu
- **Pronouns**: He/Him

Class Website: www.ericdunford.com/ppol564

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; and an introduction to machine learning methods. 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.

Time and location

Classes will be held on **Wednesdays** from ***6:30pm to 9:00pm*** in **Car Barn 204**:

- September 1, 8, 15, 22, 29
- October 6, 13, 20, 27
- November 3, 10, 17, 24
- December 1

Asynchronous & Synchronous Lectures

The lecture will be broken up into *synchronous* and *asynchronous* components.

- The ***asynchronous components*** will cover the main concepts of the lecture. These materials will take the form of embedded videos in class lecture notes on the course website. Students are required to review this content along with the lecture notes and readings prior to the start of class. ***Asynchronous materials will be made available a week prior to the scheduled lecture date.***
- The ***synchronous component*** will take place at the scheduled class time/place and will involve active coding walkthrough, breakout group sessions, and questions. The aim of the synchronous class time is to reinforce the concepts covered in the asynchronous lecture materials. Thus, it is imperative that students complete the asynchronous material *prior to the start of the synchronous lecture.*

Note that this class is scheduled to meet weekly for 2.5 hours. I will do my best to ensure that the asynchronous and synchronous material in combination does not exceed 2.5 hours weekly. Put differently, students will not be required to commit more than 2.5 hours to lecture (on average). This does not include readings, homework and/or coding discussions; rather, bifurcating lecture materials into synchronous and asynchronous components is necessary when learning technical material. Lectures that exceed an 1.5 hours are not effective. Moreover, to best absorb technical concepts, it helps to be able to return to them. The recorded asynchronous lecture material allows students this opportunity to pause, ponder, and repeat.

Five minute breaks will be taken approximately every 40 minutes during the synchronous lecture.

All synchronous lecture material will be recorded and stored on the class Canvas site. Students who are unable to attend the synchronous lecture will be able to review the materials covered in class at a future date. *It is the students responsibility to review all lecture materials and to keep pace with the course.*

Virtual Classroom

In case we must switch to a virtual classroom due to a COVID-19 outbreak, we will use **Zoom** (a web-conferencing platform) to hold class each week. Class will meet at its regularly scheduled time each week for synchronous lectures. If you do not have Zoom, you can download it **here** prior to the start of class.

A link for the synchronous component of the weekly lecture along with a link for virtual office hours is posted on the course website and Canvas. Students will use this link to access the live Zoom call for lecture.

If the link breaks or does not function properly, please check the `#general` channel on Slack for information regarding the new link. If there is no message regarding a new link, please contact the professor and/or TA via Slack. All synchronous lecture material will be recorded.

Course Objectives

The course aims to provide students with the following competencies:

- 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.
- Learn to use statistical learning approaches to effectively explore and ask questions from data.

Required Materials

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

- **Vanderplas, J.T., 2016. "Python data science handbook: tools and techniques for developers."** ***O'Reilly.*** (Online version: https://jakevdp.github.io/PythonDataScienceHandbook/)
- **Miller, B. and Ranum, D., 2013. "Problem Solving with Algorithms and Data Structures using Python."** (Online version: https://runestone.academy/runestone/books/published/pythonds/index.html)

- **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). "An Introduction to Statistical Learning: with Applications in R". *New York: springer*.**
- ***Additional readings will be posted for each class and can be found on the course website***. Most reading material is open source and available via a link on the reading list, otherwise it can be found on Canvas.

Class Website: A class website (www.ericdunford.com/ppol564) 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-2021.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, readings, and assignments. All readings and assignments 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.

Course Requirements

Assignment	Percentage of Grade
Participation/Attendance	5%
Coding Discussion	15%
Problem sets	40%
Final Project	40%

Preparation, Participation and Attendance (5%): It is imperative that students keep up with the asynchronous materials and attend synchronous lectures on time. See the "Participation" section in the Course Policies section for more details.

Coding Discussions (15%): Some weeks there will be a coding problem/prompt/dataset to explore pushed to the class Github repository. Students will be required to submit an original response to the prompt. A point will be awarded for (1) submitting on time and

(2) the quality of the submission. "Quality" is defined as an *original*, well-commented, and clear solution/analysis (i.e. the solution follows the guidelines required for assignments - see below).[1]

All submission must be posted Sunday 11:59 PM (EST). The schedule for the coding assignments are listed below.

No.	Coding Discussion Week	Date Assigned
1	Week 2	September 8
2	Week 3	September 15
3	Week 6	October 6
4	Week 10	November 3
5	Week 11	November 10
6	Week 12	November 17

The goal of the coding discussions is to apply a concept learned during the week in a way that helps build a greater level of programming fluency. Programming skills are honed through active usage and repetition. Learning to read other people's code and detecting issues is vital to successful collaborations in applied work. The point is not to be "right" necessarily but rather to try, learn, and collaborate.

Problem Sets (40%): Students will be assigned four 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. All assignments will be posted Wednesday evening by 9:00PM (i.e., by the scheduled end of class) on the class Canvas site for the weeks marked on the syllabus. Problem sets are due on the date and time posted on Canvas and must be submitted on Canvas. Generally, a week will be allotted to complete each assignment. ***Late assignments will be penalized a letter grade for every day they are overdue***.

The assignments can be in the form of a Jupyter Notebook (`.ipynb`) or R Markdown (`.Rmd`). Student's must submit completed assignments as a rendered `.html` file to Canvas on the assigned due date. All assignment submissions must adhere to the following guidelines:

- (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),

[1]Students are encouraged to generate a response before looking at the responses of their peers. If a student appears to have copied an answer of another student, we'll examine the time stamp and only award a quality point to the first entry of the timeline of responses that appear duplicative.

  - 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 should contain no grammatical or spelling errors;
- (iv) All mathematical formulas should be written in LaTex;
- (v) All solutions must be completed in Python 3.

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

Assignment	Date Assigned	Date Due
No. 1	September 22	September 29
No. 2	October 13	October 20
No. 3	October 20	October 27
No. 4	November 24	December 1

Final Project (40%): Data science is an applied field and therefore, it is important that you understand how to conduct a complete analysis from collecting data, to cleaning and analyzing it, to presenting your findings. Toward the end of the semester, you will complete an independent data science project, *applying concepts learned throughout the course*. The project is composed of three parts: a 2 page project proposal, an in-class presentation, and a 12-page project report. Due dates and breakdowns for the project are as follows:

Requirement	**Due**	**Length**	**Percentage**
Project Proposal	November 3	2 pages (500 words)	5%
Presentation	December 1	7 minutes	10%
Project Report	December 16	12 pages (3000 words)	25%

Students will use Git/Github for version control to track progress made on their analysis. Each student will be required to create a public Github repository and use it to track progress made on the project. Failure to version control one's work on the project could result in a deduction in points on all components of the project.

Details regarding each aspect of the project will be posted on the course website leading up to the first due date (i.e. the Project Proposal). Until then, we will not discuss the project in class. The reason for this is that students need to reach a basic level of data competency before thinking through a project idea. Thus, discussion of the final project and the development of a project proposal will align with the final portion of the class.

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	$< 70\%$

How to Succeed

- **Come Prepared.**
  - Do the readings. Think about the readings on their own terms, but also in terms of how the concepts apply to things you're interested in.
  - It is expected that students bring their computers to class to partake in computational activities or play with coding being discussed in class. Moreover, students should have all relevant software up and running on their machines.
- **Ask Questions.**
  - Formulating a question helps you engage with the material much more deeply. If you have a question, it's almost certain that others do too; asking a question will not only help yourself, but you will help others. Most importantly, asking questions helps keep the class on track. If there are lots of questions, we'll slow down and get things figured out. If there are few questions, we'll charge ahead.
- **Collaborate.**
  - Utilize **the class slack channel** to pose any questions, insights, coding problems and concerns. The channel will offer an open forum to communicate, collaborate, and collectively problem solve.
- **Start homework early.**
  - Sometimes the data doesn't cooperate, or there is an error in your code that will take you awhile to figure out and debug. You don't want to find this out at 11pm the night the homework is due. Also, the more you are doing homeworks on time, the more you will be able to follow the lectures.

- **Try doing it the hard way.**
  - A core factor in the success of a data scientist is being able to explain how an algorithm or analysis was constructed, not just use software. In this class, where possible, build from scratch rather than an overly convenient library. This will allow you to become more creative down the line.

Course Policies

Participation

Participation is required in this course. I define participation as:

- Attending synchronous lecture components over Zoom (if we must switch to a virtual format).
- Completing the readings and asynchronous materials prior to the synchronous (in-person) lecture.
- Asking questions and participating in class.
- During synchronous lectures, cameras are active at all times.
- Paying attention to the professor during lecture
- Engage in break-out group discussions when assigned.
- Responding to questions asked during synchronous sessions.
- When ***in-person***
  - Not looking at your computer screen for extended periods of time;
  - Never looking at your phone during class;
  - No side conversations during lecture.

I reserve the right to deduct participation points from students who are not participating as expected.

Attendance

- ***When in-person***, a sheet of paper will be made available at the front of the room at the start of every class. Students must write their name on the paper. The ***paper will be removed 5 minutes after the start of class***. Students who walk in late after that point will not have an opportunity to write their name and will be considered absent. This log will be used, in part, to calculate the attendance grade.
- ***When virtual***, attendance will be drawn from the Zoom attendance log. The instructor will take attendance at the start of class. Students who join the Zoom call after that point will not be counted.

If absent, each student is responsible to make up the materials missed during a lecture on their own. All lecture notes/lecture materials will be posted on the class website. Thus, students who missed a lecture should reach out to their peers in the class for lecture notes. It is not the responsibility of the Professor/TA to fill absentee students in on any missed content.

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. Please address the professor and TA by their last name unless stated otherwise.

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.

Assignments and Late Work

Assignments should be clear, legible, and submitted in the required format. Writing assignments will be graded on the basis of content, logic, analysis, mechanics, organization, and research. Due dates for all assignments will be posted on Canvas and are non-negotiable. Exceptions to this policy will be made only under extremely unusual circumstances and will require valid documentation from the student. ***Late problem sets will be penalized a letter grade per day.***

Proof of Diligent Debugging

When reaching out to the professor or teaching assistant regarding a technical question, error, or issue you ***must*** demonstrate that you made a good faith effort to debugging/isolate your

problem prior to reaching out. In as concise a way as possible, send a record of what you tried to do. ***The professor/TA is a resource of last resort***. As software is continually being refined in data science and new approaches continually emerge and changing, learning how to frame your question and find a similar solution online is a key tool for success in this domain. If you make a diligent effort beforehand to solve your problem, we will do the same in trying to help you figure out a solution.

You should always consult with the TAs regarding your problem before reaching out to the professor.

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 my syllabus, lectures, videos, handouts, problem sets, and problem set answers to be my intellectual property. I respectfully request that you refrain from sharing my materials in any electronic (or paper) format. You are welcome to save my lectures for your own use, but they should not be posted anywhere. Sharing notes, on an occasional basis, with others in the class is fine as long as they are not posted. Students found in breach of this policy will fail the course.

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 (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/.

Plagiarism

Plagiarism is the intentional or unintentional presentation of another person's idea or product as one's own. Plagiarism includes, but is not limited to the following: copying verbatim

all or part of someone else's written work; using phrases, charts, figures, illustrations, code, or mathematical / scientific solutions without citing the source; paraphrasing ideas, conclusions, or research without citing the source; and using all or part of a literary plot, poem, film, musical score, or other artistic product without attributing the work to its creator. In technology, plagiarism is the verbatim use of a code chunk from a peer or third party website to complete an assignment questions. Students can avoid unintentional plagiarism by following carefully accepted scholarly practices. Students who plagiarize will receive a 0 on the plagiarized assignment and may fail the course, if deemed necessary.

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 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.

Statement on Sexual Misconduct

Please know that as a faculty member I am committed to supporting survivors of sexual misconduct, including relationship violence, sexual harassment and sexual assault. However, university policy also requires me to report any disclosures about sexual misconduct to the Title IX Coordinator, whose role is to coordinate the University's response to sexual misconduct.

Georgetown has a number of fully confidential professional resources who can provide support and assistance to survivors of sexual assault and other forms of sexual misconduct. These resources include:

```
Associate Director
Jen Schweer, MA, LPC
Health Education Services for Sexual Assault Response and Prevention
(202) 687-0323
jls242@georgetown.edu

Erica Shirley
```

```
Trauma Specialist
Counseling and Psychiatric Services (CAPS)
(202) 687-6985
els54@georgetown.edu
```

More information about campus resources and reporting sexual misconduct can be found at http://sexualassault.georgetown.edu.

Course Calendar

Week	Date	Topic	Assignment	Coding Discussion
1	1-Sep	Introductions, Installations, and IDEs		
2	8-Sep	Version Control, Workflow, and Reproducibility		X
3	15-Sep	Object-Oriented Programming in Python		X
4	22-Sep	Introduction to Algorithms	Assignment 1 Assigned	
5	29-Sep	From Nested Lists to Data Frames	Assignment 1 Due	
6	6-Oct	Approaches to Data Manipulation in Python		X
7	13-Oct	Data Visualization and Exploration	Assignment 2 Assigned	
8	20-Oct	Drawing from (Un-)Structured Data Sources	Assignment 2 Due; Assignment 3 Assigned	
9	27-Oct	Introduction to Statistical Learning	Assignment 3 Due	
10	3-Nov	Continuous Outcomes and Linear Regression	Project Proposals Due	X
11	10-Nov	Probability, Bayes Theorem, and Classification		X
12	17-Nov	Algorithmic Approaches to Supervised Learning		X
13	24-Nov	Interpretable Machine Learning	Assignment 4 Assigned	
14	1-Dec	Project Presentations	Assignment 4 Due	

IMPORTANT: This syllabus is subject to change and may be amended throughout the course to reflect any changes deemed necessary by the professor. Any changes will be announced in class or over Slack.