PPOL 670-01
Introduction to Data Science for Public Policy

Spring 2019

Instructor

Professor: Eric Dunford

- **Office**: 404 Old North
- **Office Hours**: Tuesdays 3pm to 5pm
- **Email**: eric.dunford@georgetown.edu

Teaching Assistant: Nathan Lovin

- **Office Hours**: by appointment online (https://nathanlovin.youcanbook.me/)
- **Email**: nhl8@georgetown.edu
- **Skype**: nathan.lovin

Course Description

The Introduction to Data Science for Public Policy is a survey course of fundamental concepts and techniques used in data science. This course teaches students how to synthesize disparate, possibly unstructured data to better understand and characterize the world around us and to draw meaningful inferences from data. Topics covered include fundamentals of functional programming in R, data wrangling and probing, data extraction (via web scraping and APIs), data visualization, and data ethics. The course surveys commonly used data science approaches, such as text analysis, machine learning, network, and geospatial analysis.

The course is focused on telling stories with data in order to make compelling, fact-based arguments. The value of this course is in the mortar, not the bricks. The course aims at offering students a practical toolkit for data analytics and exposure to different analytical approaches and tools. The objective of the course is to equip students with the skills to incorporate data into their decision-making and analysis. No prior programming experience is assumed or required.

Time and location

Classes will be held on **Mondays** from ***6:30pm to 9:00pm*** in **Reiss 282**:

- January 9 (Wednesday), 28
- February 4, 19 (Tuesday), 11, 25
- March 11, 18, 25
- April 1, 8, 15, 29

Holidays/Breaks (No class):

- January 14 (no class due to inclement weather)
- January 21
- March 4
- April 22

Course Objectives

This course focuses on providing students with an applied knowledge of the `R` programming environment while placing emphasis on developing a practical data science toolkit that students can implement quickly and efficiently. To this end, the course takes a 'Tidyverse' approach to `R` programming, which provides users an intuitive grammar for data manipulation and visualization. The goal is to establish a practical toolkit for analysis in `R` without getting too bogged down in the nuts and bolts of functional programming.

1. Receive exposure to the main concepts associated with data science.
2. Learn how to wrangle (prepare and clean) different types of data.
3. Learn to analyze data for descriptive and predictive purposes.
4. Understand the basics of programming in `R` with emphasis on the "tidy" ecosystem of packages.
5. Learn to identify and visualize important trends and findings.
6. Debate and articulate ideas related to data.
7. Tell a coherent, meaningful data science story.

Required Materials

Textbook: There is no required textbook for this class. Required class readings will be posted on Canvas.

- *Suggested Textbook*: We'll be relying on heavily on Garrett Grolemund and Hadley Wickham's phenomenal book, **R for Data Science**, throughout the course. In an effort to keep costs as low as possible, we'll resort to the online presentation of these materials. That said, many students find it useful to have a hard copy of the materials. If that is the case, I strongly encourage students to purchase this book. It will serve as a valuable reference both during the semester and into the future.

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

Computing: Programming task for in-class activities and assignments will be conducted using `R`. Students are strongly encourages to utilize Rstudio, which offers an accessible and widely-utilized graphical user interface for programming in `R`.

NOTE: In-class activities will include programming in `R`. If you do not have access to a laptop on which you can install `R` and `Rstudio`, please contact the professor and/or TA for assistance.

Course Requirements

Assignment	Percentage of Grade
Participation	15%
Problem sets	30%
Project	55%

Preparation and Participation (15%): Data science techniques are used to solve problems; therefore, a significant portion of class time will be spent on interactive problem-solving activities. It is imperative that you arrive to class prepared for these hands-on activities. As a result, 15% of each student's grade will be based on in-class activities and class participation.

Problem Sets (30%): Students will be assigned six problem sets. 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 problem sets will account for 30% of the final grade. Problem sets are due on the date and time posted on Canvas and must be submitted on Canvas. Late assignments will be penalized 20% for every day they are overdue.

Project Memos (20%) and **Showcase** (35%): 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. Over the course of the semester, you will work in a small group of students on a project that applies what you have learned in the course. The project, which will be completed in stages, will comprise 55% of your course grade.

Students will use Git/Github for version control and will use the text-based practices covered the first week of class. Details regarding each aspect of the project will be posted on CANVAS leading up to the first due date. Each team will be required to create a public Github repository in which all aspects of their project will be housed.

Requirement	**Due**	**Percentage**
Teams Assigned	January 28	
Establish Team Repository	February 4	
Memo 1: Topic and Data	February 11	10%
Memo 2: Analysis Plan	March 11	10%
Project Report	April 29	20%
Presentation	April 29	10%
Peer Evaluation	April 29	5%

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%

Managing the Workload: How to Succeed in this Course

- **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.
  - As this class is quite hands-on, it is expected that students bring their computers to class to partake in computational activities. 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.**
    - Work in groups, but do so wisely. Collaboration is the greatest source of creativity and innovation. Better yet, working with classmates is a great way to learn from each other. Often, classmates will have some way of explaining things that clicks for you, and, more often than not, the act of explaining something to someone else will make things click for you. This only works, though, if you prepare by yourself first. If you show up and wait for classmates to do the work, you can probably muddle through the homeworks, but you'll have trouble participating in classes and may fall behind as the material we cover cumulates and needs to be understood at each step.
    - collaboration should not result in verbatim submissions (e.g. no copy cats). As everyone writes code following their own unique logic, the chance of identical submissions is unlikely and easily detectable. Non-unique code will be penalized.
    - Finally, 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 homeworks 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 before the homework is due. Also, the more you are doing homeworks, 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

Communication

- 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.***

- The class also has a dedicated slack channel (ppol670introt-5qu2508.slack.com). The channel serves as an open forumn 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).

Electronic Devices

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.

Consult the Syllabus

The professor and TA reserve the right not to respond to emails where the answer is in the syllabus.

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 noted 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 25% per day, and late project deliverables will be penalized 50% 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.

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, 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 record 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.

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/

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:

Jen Schweer, MA, LPC
Associate Director
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
1	9-Jan	Reproducibility and Version Control	
Canceled	14-Jan		
2	28-Jan	Introduction to Programming in R	
3	4-Feb	Data wrangling in R	
4	11-Feb	Data Visualization	Problem Set 1 Due; Memo 1 Due
5	19-Feb	Web Scrapping and APIs	
6	25-Feb	Exploratory Data Analysis and Descriptive Analytics	Problem Set 2 Due
7	11-Mar	Inference and Modeling	Memo 2 Due
8	18-Mar	Predictive Analytics	Problem Set 3 Due
9	25-Mar	Text as Data	Problem Set 4 Due
10	1-Apr	Networks	Problem Set 5 Due
11	8-Apr	Geographic Information Systems (GIS)	
12	15-Apr	Ethics and Project Management	Problem Set 6 Due
13	29-Apr	Final Presentations	

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 on Canvas.