PPOL 670-02
Introduction to Data Science

Spring 2020

Instructor

Professor: Eric Dunford, Ph.D.

- **Office**: 404 Old North
- **Office Hours**: Tuesdays 1pm to 3pm
- **Email**: eric.dunford@georgetown.edu
- **Pronouns**: he/him

Teaching Assistant: Trey Billing, Ph.D.

- **Office Hours**: by appointment online (https://calendly.com/billingtt/ppol670)
- **Email**: tb990@georgetown.edu
- **Skype**: billingt1
- **Pronouns**: he/him

Course Description

This course teaches students how to synthesize disparate, possibly unstructured data in order to draw meaningful insights from data. Topics covered include fundamentals of functional programming in R, literate programming, data wrangling, data visualization, data extraction (via web scraping and APIs), text analysis, and machine learning methods. In addition, students will be exposed to Git and Github for reproducible research. The course aims to offer students a practical toolkit for data exploration. 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, but prior statistics training is required.

Time and location

Classes will be held on **Wednesdays** from ***3:30pm to 6:00pm*** in **Car Barn 203**:

- January 15, 22, 29
- February 5, 12, 19, 26
- March 4, 18, 25

- April 1, 8, 15, 22

Holidays/Breaks (No class):

- January 8 (Monday classes on Wednesday)
- March 11 (Spring Break)

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. Understand the basics of programming in `R` with emphasis on the "tidy" ecosystem of packages.
2. Learn how to wrangle (prepare and clean) different types of data.
3. Learn to identify and visualize important trends and findings.
4. Learn to extract and process data from unstructured sources, such as the web and/or text.
5. Learn to use statistical learning approaches to effectively explore and ask questions from data.

Pre-Requisites

- ***Required***: PPOL501/531 - Statistical Methods for Policy Analysis (or an equivalent course)
- ***Preferred***: PPOL502/532 - Regression Methods for Policy Analysis (or an equivalent course)

Required Materials

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

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

  - 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 book materials. I strongly encourage students to purchase this book. It will serve as a valuable reference both during the semester and into the future.

- **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/ppol670) 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 (ppol670_Intro-to-DS_Spring2020). 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 Slack, 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 periodically throughout the course and should be checked on a regular basis. All 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 encouraged 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	10%
Problem sets	45%
Final Project	45%

Note that the grades on Canvas are not weighted, and thus, may not accurately reflect a student's final grade.

Preparation and Participation (10%): It is imperative that you arrive to class prepared for lecture and any hands-on activities. As a result, 10% of each student's grade will be based on class participation. See "Attendance/Participation" in the Course Policies section for more details.

Problem Sets (45%): Students will be assigned five 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**. Problem sets are due on the date and time posted on Canvas and must be submitted on Canvas. Late assignments will be penalized a letter grade for every day they are overdue.

All problem sets must be submitted as PDFs with clean, readable code chunks using `RMarkdown`. Along with the `.pdf`, student's must submit a `.zip` file containing the `.rmd` file they used to knit the `.pdf` and the data used to complete the assignment. The `.rmd` file should be completely reproducible and contain no machine specific information (e.g. a file path). 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 code's purpose by reading the comment),
  - Coding solutions should be broken up into individual code chunks, not clumped together into one large code chunk (See examples in class or reach out to the TA/Professor if this is unclear),
- (iii) Non-coding responses should all be written in Markdown and should contain no grammatical or spelling errors;
- (iv) All programming solutions should employ concepts learned during the course. Specifically, students must use `tidyverse` solutions learned in class, over base `R` solutions pulled from the internet.

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

Assignment	Date Assigned	Date Due
No. 1	February 5	February 12
No. 2	February 19	February 26
No. 3	February 26	March 4
No. 4	March 18	April 1
No. 5	April 15	April 22

Final Project: 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	March 25	2 pages	5%
Presentation	April 22	7 minutes	10%
Project Report	May 7	12 pages	30%

Students will use Git/Github for version control and will use the reproducibility practices covered the first week of class. Each student will be required to create a public Github repository in which all aspects of the project will be housed. 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; once we've broadly covered most of the fundamental data topics covered in this course.

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 post 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

Attendance/Participation

Participation is required in this course. Participation can be decomposed into class attendance, engagement, and completing the class assignments. Specifically, I define "engagement" as:

- Asking questions and participating in class (no zombies)
- Paying attention to the professor during lecture
  - no messaging during class
  - never looking at your phone during class
- No side conversations during lecture

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

Attendance will be taken daily. 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.

If absent, each student is responsible to make up the materials missed during a lecture on their own. All lecture notes 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

- 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.
- For general, class-relevant questions, `Slack` is the preferred method of communication. Please use the general or the relevant channel for these questions.
- I will 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 5PM) or during the weekend until Monday (9AM). 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

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 along with a reproducible example emulating the error. (See the materials for Week 3 on how to generate a reproducible example (`reprex`)). 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. Note that the ***professor/TA is a resource of last resort***: only come to them after you've exhausted all other options.

Class Seating

Students should (and must) sit beside a different student each time the class meets. The aim is to facilitate diverse interactions. If all or some students fail to follow this procedure, then the professor reserves the right to generate a random seating assignment each class. Failure to comply with the class seating policy will result in a deduction in participation points.

Instructional Continuity

In the event the university is closed for a scheduled class, lecture materials will be provided via a video link on Canvas. Students will be expected to watch the recording and answer a brief quiz on the materials cover therein before the next scheduled class period. Students will be notified via Slack when the video lecture has been posted to Canvas. Assignment due dates will not change due to the closure.

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 elsewhere online. 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/

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
1	15-Jan	Reproducibility and Version Control	
2	22-Jan	Introduction to Programming in R	
3	29-Jan	Reproducibility in Practice	
4	5-Feb	Data Wrangling in R	
5	12-Feb	Data Visualization	Problem Set 1 Due
6	19-Feb	Web Scrapping and APIs	
7	26-Feb	Text as Data	Problem Set 2 Due
8	4-Mar	Introduction to Statistical Learning	
-	11-Mar	Spring Break	Problem Set 3 Due
9	18-Mar	Applications in Supervised Learning (Regression)	
10	25-Mar	Applications in Supervised Learning (Classification)	Project Proposal Due
11	1-Apr	Applications in Unsupervised Learning	Problem Set 4 Due
12	8-Apr	Exploratory Data Analysis	
13	15-Apr	Work Day + Using Spatial Data	Problem Set 5 Due
14	22-Apr	Project Presentations	
Final	7-May	Final Project Due (6:00 PM)	

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 Slack.