SIS-750-007: Data Analysis

Meetings: Wednesdays, 5:30-8:00pm
Classroom: Anderson Computing B14

Professor Austin Hart (ahart@american.edu)
Office: SIS 345
Office hours: Mon/Fri 2:00-4:00pm
Website: https://austin-hart-pols.github.io
Github: https://github.com/austin-hart-pols

Overview and learning outcomes

This course introduces the art and practice of data analysis. It focuses on the preparation of data (data wrangling), professional presentation of analysis, and the process of producing scalable, replicable work (programming). More than just learning the requisite code, I emphasize the professional outputs associated with a career in data analysis. This semester is a chance to get your hands dirty in the world of analysis and explore the new and challenging puzzles you face every day in this profession.

Students who complete this course will be proficient junior analysts. They will be able to (i) build, clean, and manipulate quantitative data, (ii) generate professional data visualizations, (iii) present data-driven research to a professional audience, and (iv) write clear, efficient, reproducible code in the R programming language.

Expectations and workload

This is an advanced graduate course. The material is challenging, and the workload is intense. However, my expectations for the course and of your performance are appropriate to your standing as professionals and aspiring professionals in a top graduate program.

Success in this course will require 10-15 hours of work outside of class each week. Please do not enroll in this course if you cannot commit to these hours.

Course material and software

I will post handouts, data, R scripts, and links to reading materials on our course github site. While there is no textbook to purchase, I will draw regularly on the following: - Jenny Bryan's Stat 545 collection - Claus Wilke's foundational text on visualization, Fundamentals of Data Visualization - Winston Chang's Cookbook for R - A basic introduction to R, my own Survival Guide

You need regular and reliable access to R and RStudio IDE, and I urge you to bring your laptop to class. These programs work across platforms and are available for free. You will also need to connect RStudio with Github, a cloud-based `Git` repository for code sharing and collaboration (i.e., version control). You can find installation instructions in the Getting Started repository.

Becoming proficient with these programs is a challenging, though necessary, part of this course. When you run into trouble (i) Review course materials, (ii) search online (seriously, Google knows the answer), (iii) ask your classmates, and (iv) ASK ME.

Assignments and grading

Grades for the course will be assigned based on performance in the following areas:

- **Problem Sets and Deliverables (80%)** Hands-on experience is the only way to build proficiency. So you will complete nine, equally-weighted exercises that ask you to put into practice what we cover in class. *Problem sets (6)* feature a series of discrete questions of varying degree of difficulty. I automatically drop the lowest problem set score when calculating final grades. *Deliverables (3)* include a recorded slide presentation, a research poster, and a codebook. I will distribute assignments two weeks in advance of the due date. Budget 10-15 hours to complete each exercise. I encourage you to consult with classmates, but your code and your written answers must be yours and yours alone. If someone else wrote or executed your code, it is not individual work. If you sent your code or scripts to someone else, you did not participate earnestly in the process of individual work. You may not consult with anyone not currently enrolled in or teaching this course.

- **Analyst Portfolio (10%)** At the end of the semester, you will create a portfolio that includes an edited collection of your best work from the course and an updated resume tailored to a position in data analytics. You will post the portfolio to your GitHub page.

- **Participation and professionalism (10%)**
  Professionalism is critical to success in data analytics positions. Aside from what one can do on a laptop, a good analyst engages with peers constructively and respectfully, asks questions when expectations are unclear, and creates inclusive environments for co-workers and clients. I will evaluate participation and professionalism in terms of your engagement with peers, with myself, and with the material during class-related activities. Do you come to class prepared? Do you ask and answer questions? Are you contributing to a constructive and inclusive environment?

Please take note of the assignment due dates and plan accordingly. Except in truly unusual and unavoidable circumstances I will not accept late work.

Schedule Overview

We will proceed according to the outline below. Details and required materials for each week are available on our github site.

Date	Topics	Due in class
Pre	Getting started	
Jan 17	Intro and Monty Hall	
Jan 24	R programming crash course	PS 1
Jan 31	Wrangling 1: frames, tibbles, lists	PS 2
Feb 7	Wrangling 2: reshape, merge, append	Draft resume
Feb 14	Wrangling 3: all about variables	PS 3
Feb 21	Wrangling round-up	
Feb 28	Visualization 1: grammar of graphics	Codebook
Mar 6	Visualization 2: facets and patchwork	PS 4
Mar 20	Markdown 2: slide decks	PS 5
Mar 27	Careers in analytics	Presentation 1
Apr 3	Analysis 1: linear associations	
Apr 10	Analysis 2: advanced topics	PS 6
Apr 17	Analysis round-up	Poster
Apr 24	Scaling up	
May 1	Submit Analyst Portfolio	Portfolio

Making the most of our time

I strongly encourage you to drop in during office hours or set up meetings on a semi-regular basis. I love working with students in this course, and I will do my best to help you think through challenges both big and small. I'm happy to discuss problem sets, the field of data analytics, your intellectual and professional goals, and whatever else is on your mind. I regularly work with students (current and former) to write letters of recommendation, organize independent study, supervise theses, hire TA/RAs, etc. The sooner you come in to consult, the better.

Academic Integrity, Group Work, and AI

Assignments in this course are individual exercises. The written work you submit must be substantially your own work. You are welcome to make use of appropriate research tools and materials to help learn how to answer a question or otherwise fulfill the assignment. Two in particular require special attention:

- *Collaborating with classmates:* I encourage you to study with your classmates and to talk through challenges and questions in written work. You may not co-write, edit, exchange, or otherwise share your answers or code. Posting your answers in a group chat or copying code posted by a classmate, for example, constitutes a violation of the Academic Integrity Code.
- *Using generative AI (ChatGPT, Bard, etc):* Use of AI is strongly discouraged. These tools may be able to answer some/all of questions on our assignments, and they can be incredibly useful for folks with strong foundational skills. For those just starting out, however, I believe that reliance on this technology inhibits the development of those core skills. If you choose to use AI tools in this course, you must submit an appendix with your work including verbatim text of your prompt and the output produced by the tool.

In each case, you must start and finish your written work independently. Regardless of what resources you rely on in your work, you bear sole responsibility for the integrity of what you submit.

Standards of academic conduct are set forth in the University's Academic Integrity Code. By registering for classes, you have acknowledged your awareness of the Academic Integrity Code, and you are obliged to become familiar with your rights and responsibilities as defined by the Code. Violations of the Academic Integrity Code will not be treated lightly, and disciplinary actions will be taken should such violations occur. Please see me if you have any questions about the academic violations described in the Code in general or as they relate to particular requirements for this course.

Background knowledge

A background in introductory statistics—-equivalent to SIS-600-—is necessary for this course. So I will assume familiarity with descriptive statistics (e.g., means, standard deviations), hypothesis testing (through linear regression), and basic statistical computing (e.g., R or Stata).

Snow days, COVID, and random apocalypses

If we're unable to meet for our regular class time due to a University closure (e.g., snow day) or something on my end (e.g., I get COVID), we will hold class virtually on Zoom. If you are sick (e.g., you get COVID), do not come to our cramped lab. Join virtually instead. Let me know that you'll be virtual and contact a classmate to dial you in for class (I can set you up with someone if you're not sure who to ask).

Support services

I will make every effort to enable full participation in this course by all students. Students needing ongoing special accommodations should contact me as early as possible (within the first two weeks of the semester, or as soon as those needs arise) to discuss how they can be met within the structure of the course. Every effort will be made to maintain the confidentiality of personal information. You may also wish to contact Disability Support Services (Mary Graydon Center, Room 206) or the Counseling Center (Mary Graydon Center, Room 214) if you have questions about campus policies and services or wish to register for specific accommodations or assistance with physical, medical, or psychological disabilities.

Requests for special accommodations

Students anticipating any difficulty in completing assigned work on time should consult with the professor well in advance of course deadlines. Any student seeking exceptions to course policies or requesting special accommodations due to medical or familial issues must first consult with the Office of the Dean of Students (Butler Pavilion, Room 408). Should the Office of the Dean of Students determine that accommodations are appropriate, a letter will be issued from the Office of the Dean of Students to the student's professors. Students must then follow up with their professors either in person or via email as soon as circumstances permit in order to discuss new arrangements for required work and deadlines. Simply providing a letter from the Office of the Dean of Students does not constitute a waiver for course requirements or deadlines, nor does such a letter excuse work missed prior to the provision of documentation unless the Office of the Dean of Students specifically indicates that the accommodations are retroactive. Late penalties and other course policies will apply to any revised assignment arrangements or deadlines.