Data Skills for Public Policy - 30531

Professor: Peter Ganong
TAs: TBD
Course email: uchicagodatasci@gmail.com
Lab (Required) to Work on Problem Sets: Monday and Wednesday, 1:30 PM - 2:50 PM
Lecture: Monday and Wednesday 9:00 AM - 10:20 AM

Course Description

This course is the second of a three-quarter sequence for the Harris Data Science certificate. This certificate is designed to train you to work in the rapidly-expanding field of data analytics in the public sector after graduation. Although the course is designed for MPPs, undergraduates are welcome to enroll as well.

The goal of this course is to teach you to *quickly engage an important policy question with a data visualization.* Doing this requires two new skills.

First, we will teach you to be proficient in R. We will closely follow Hadley Wickham and Garret Grolemund's R for Data Science. The online textbook is free.

Second, we will teach you to use data to improve the performance of public sector organizations. The course material draws on Peter's experience helping to start the Citywide Analytics Team in Boston. The certificate description contains more examples of how teams like this are transforming government. During the course, you will complete nine problem sets. Through repeated analysis, you will gain an in-depth knowledge of three public sector datasets:

- Chicago and Boston 311 service requests
- traffic data for Chicago captured at 5-minute intervals from Waze
- shift-level human resources data from a large transit agency.

The last two datasets are proprietary. To use these two datasets, you will need to agree to abide by the confidentiality rules from the data providers.

This course will differ in three ways from the typical Harris course. Learning R, just like learning a foreign language, is hard and requires lots of repetition.

1. Different students have different styles for learning how to code. As a result, we will use a flipped classroom. In addition to Monday and Wednesday morning lecture, the course will have a *mandatory* time to meet where you will work on your problem sets on Tuesday and Thursday afternoon.
2. The best way to learn to write good code is to write lots of code. As a result, this course will not have any exams and will have one problem set per week.
3. You are encouraged to discuss challenges with your classmates, but what you submit must be your own work.

Prerequisite: As a part of the Harris Data Science certificate, you must have taken 30550 “Introduction to Programming for Public Policy”. The course is also open to students with significant prior programming experience. If you have not taken 30550 and would like to enroll in this course, you need to either have taken a programming course in Python or you can send examples of code you have written to the course email address and we will let you know if it is OK for you to enroll.

Lectures

About two-thirds of each lecture will consist of material from the *R for Data Science* textbook. One-third of each lecture will be about applying these data skills to improving public sector performance.

Part I. Explore -- Make Simple Plots and Tables Quickly

1. Ch 1 & 2 (Intro), Ch 3 (Data visualization)
2. Ch 4 (Workflow: basics), Ch 6 (Workflow: scripts), Ch 27 (Intro to Rmarkdown)
3. Ch 5 (Data transformation). *PS1 due Friday Jan 12*
4. Ch 7 (Exploratory Data Analysis), Ch 8 (Workflow: projects)

Part II. Wrangle -- Cleaning and Structuring Data

5. Ch 9 (Intro), Ch 10 (Tibbles), Ch 11 (Data import) *PS2 due Friday Jan 19*
6. Ch 12 (Tidy data)
7. Ch 13 (Relational data) *PS3 due Friday Jan 26*
8. Ch 14 (Strings), Ch 15 (Factors), Ch 16 (Dates and times)

Part III. Program -- Handle Larger, More Complicated Data

9. Ch 17 (Intro), Ch 18 (Pipes), Ch 19 (Functions) *PS4 due Friday Feb 2*
10. Ch 20 (Vectors)
11. Ch 21 (Iteration) *PS5 due Friday Feb 9*

Part IV. Model

12. Ch 22 (Intro), Ch 23 (Model basics)

Part V. Communicate -- Share Your Findings with Decision-makers

13. Ch 26 (Intro), Ch 28 (Graphics for communication) *PS6 due Friday Feb 16*
14. Dashboards I
15. Dashboards II *PS7 due Friday Feb 23*
16. Dashboards III

Part VI. Maps

17. Maps I *PS8 due Friday Mar 2*
18. Maps II
19. Maps III *PS9 due Friday Mar 9*

Problem Sets, Grading, Honesty

Problem sets (80% of grade) will be submitted using github. Register here. There are 9 problem sets and they are due each Friday by 5PM. Problem sets will have two components:

- highly structured exercises from the textbook as well as
- less structured prompts where you will create data products for decision-makers using the three public sector datasets (311, Waze, human resources data)

Group work will be allowed on all but one solo problem set. I will assign you to a groups, which will rotate every week. Late problem sets will not be accepted. However, I will drop your lowest problem set grade. (The only problem set which cannot be dropped is the solo problem set.)

Attendance (10% of grade) at office hours Tuesday and Thursday afternoon to work on problem sets. Students are allowed one unexcused absence during the quarter without penalty. You must provide written documentation of your absence (e.g., send an email to the course address!). Failure to document absences (or frequent late arrivals) will result in the loss of points from your attendance grade. Full points for the quarter will be earned for full attendance.

Discussion Board Answers (10% of grade). We expect everyone to ask and answer questions on the class discussion board. Your grade will be based on the quality of the answers that you give your classmates.

Academic Honesty Writing code is substantially different from writing essays: it is standard practice to find individual functions or google things that don't work, and copy a line or two from the manual or stackoverflow. I encourage you to discuss general strategies for solving problems with your group members as well as other classmates and friends. Questions and answers on the discussion board will naturally include code. However – you should never ask to see another's solutions, and you absolutely should not copy code from your classmates. No one but you should type your code. If you find more than a single line/method, you should attribute the source in your comments.

Answers to Student Questions

Relation to Other Courses

How does this course compare to…

- MACS 30500 -- It uses the same textbook and is quite similar. 30531 will focus more on data cleaning and on public policy subject matter. There will be no machine learning or text analysis (Guillaume will cover those in the spring).
- Chicago Booth 41201: Big Data -- This would be a good course to take after completing the data science sequence at Harris. The focus of that course is statistical concepts for high-dimensional data. It assumes that you know R and that you know basic statistics.
- Chicago Booth 41205 Data Analysis with R -- 41205 is a 5-week intro to statistics and R. 30531 is a 10-week course, so it is more in-depth. It emphasizes cleaning data, visualization and using R. It does not emphasize statistics

Will the course be graded on a curve?

Yes. The curve will be similar to, or slightly more generous than, the harriswide curve.

A	A-	B+	B	B- and below
1/8	1/4	1/4	1/4	1/8