Introduction to Programming for Applied Political Data Science
Government 496/696-002

Ryan T. Moore[*]

21 February 2018 at 10:29

Course Information

Government GOVT 496/696-002
Introduction to Programming for Applied Political Data Science
Wednesday, 2.30–5.20pm
Anderson Hall B-11

Instructor Information

Ryan T. Moore, Ph.D.
Assistant Professor of Government
Office: Kerwin Hall 226
Telephone: 202.885.6470
Fax: 202.885.2967
Homepage: http://www.ryantmoore.org
Email: rtm (at) american (dot) edu
Office Hours: Thursday, 11:30am-1:30pm or by appointment
(Please use https://calendly.com/ryantmoore to schedule times.)

Course Description

This course introduces concepts and techniques required to engage in modern applied political data analysis. Data science is often described as the intersection of statistics, programming, and substantive knowledge. This course develops skills in the first two, with application to questions in applied politics. Modern political data come from large-scale experiments, text, networks, and other survey and administrative sources. Students develop skills in data wrangling, visualization, collaborative version control, statistical modeling, and scientific presentation.

As designed, this course tends toward the programming side of data science, and less toward the statistical side. For example, I plan to tackle version control and package building rather than machine learning models and natural language processing. We will adapt the plan below to

[*] Department of Government, American University, Kerwin Hall 226, 4400 Massachusetts Avenue NW, Washington DC 20016-8130. tel: 202.885.6470; fax: 202.885.2967; rtm (at) american (dot) edu; http://www.ryantmoore.org.

accommodate the experience and needs of the undergraduate, master's degree, and Ph.D. students in the course.

Learning Objectives

By the end of the course, you should be able to

- Acquire, transform, tidy, explore, analyze, and visualize political data using R,
- Typeset social scientific methods and results legibly using LaTeX and RMarkdown,
- Build an original R package,
- Control code and analysis versions using GitHub,
- Conduct original data analysis that uses techniques from the course to answer a relevant political science question.

Learning Strategies

Readings

Readings should be completed before the course meeting under which they are listed below. The course readings are primarily from two textbooks that are also available online. The textbooks include frequent exercises that I encourage you to complete for practice, even when they are not directly assigned. We may occasionally have short quizzes over the reading.

The primary textbooks for the course are

Hadley Wickham and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly Media, 2017. http://r4ds.had.co.nz/.

Hadley Wickham. *R Packages: Organize, Test, Document, and Share Your Code.* O'Reilly Media, Inc., 2015. http://r-pkgs.had.co.nz.

In the calendar below, these are denoted *DS* and *PKGs*, respectively. The online version of *DS* is denoted *DSO*. The online version of *PKGs* does not include section numbers.

Computers and Notes in Class

For most class meetings, we will focus our attention on computational implementations of social scientific techniques. There will often be time in class to pose your specific questions about code. As such, you may want to bring a laptop to class to try out new code, to update your code files, etc.

Requirements and Evaluation

Students are required to do the weekly reading, attend class, complete all assignments, and contribute significantly to course discussions about the material.

The student's final course assessment includes three components: problem sets (40%), a final project and a roughly 10-minute oral presentation and defense of that project (50%), and engagement in course conversations through attendance, in-class participation, quizzes, and Slack participation (10%). These four components each will be scored 0, 1, 2, 3, or 4, roughly representing the following outcomes:

- 0: absence of a good-faith effort to complete the work; failure to complete the work on time; demonstrates mastery of very little of the material.
- 1: a timely, good-faith effort to complete the work that demonstrates mastery of a small part of the material.
- 2: a timely, good-faith effort to complete the work that demonstrates mastery of more than half of the material.
- 3: a timely submission that demonstrates mastery of most of the material.
- 4: an excellent submission that demonstrates mastery of virtually all of the material.

You should think about these scores as being roughly scaled like grade point averages. A 3 out of 4 represents a B, not a 75% C, for example.

A summary of the course assessments is in Table 1.

Assignment	Weight	Due date
Problem sets	40%	roughly weekly
Final paper and presentation	50%	April 24
Participation (Attendance, Slack, quizzes, paper memo, exercises)	10%	ongoing

Table 1: Course Assessment Summary

No late work will be accepted. If you cannot submit an assignment on time, arrange to submit it early. I encourage you to use office hours to discuss any specific assignments, difficulties, or questions about the course.

Academic integrity is a core value of institutions of higher learning. It is your responsibility to avoid and report plagiarism, cheating, and dishonesty. Please (re-)read the University policy on academic integrity at http://www.american.edu/academics/integrity/code.cfm, particularly Sections I and II.

Problem Sets

The problem set exercises should generally be completed outside of class. You should upload a copy of your solutions (in `.R`, `.pdf`, or other format as appropriate) before the start of the class in which the exercises are due. You should submit your solutions to the course Blackboard site under Content/Assignments/Problem_Sets/[PS number]. You may work with others on the problem sets, but every keystroke of your submission must be your own.

Final Project

For the final project, you will conduct original political science research and submit your work as an R package, a GitHub repository, and/or a data analysis report as appropriate. You may select your own topic and work with at most one other class member.

One possibility is that you may use real data that policymakers want to learn about. In conjunction with The Lab @ DC, a research arm of the Executive Office of the Mayor, we will provide you with a handful of data sets pertaining to policies and programs of Washington, DC. Topics will include campaign finance and expenditures, ANC budgeting, public goods and the 311 request system, transit, and affordable housing. These data are available at http://opendata.dc.gov.

With the data you select, you will pose an appropriate political research question that the data can answer with quantitative methods and analyze the data. As appropriate, you will write a data analysis report, bundle your analysis, data, and original functions into an R package, and develop this package as a GitHub repository. Not all of these products may be appropriate for all projects, and we will discuss individually what makes most sense for your project's goals, your standing as a graduate or undergraduate, etc. You will present your research to the class in the last meeting.

The best projects will be invited to present their work to city officials, particularly representatives of the Office of the City Administrator, including The Lab @ DC, the Performance Team, and others.

Your project should represent original data analysis and code development. It should represent quantitative social science at the highest level you can muster. You may work with one other student on the final project. Working collaboratively is typical in political science research.

Software, Statistics, Data, and Literature Support

The primary software for the course is R. See http://j.mp/2e8zBkC for help getting started with R and RStudio. A brief overview is also available at http://j.mp/2ELPqFO. We will introduce LaTeX and RMarkdown for scientific communication. See http://j.mp/2EO0TEM for an introduction to LaTeX. We will introduce GitHub for version control. See http://j.mp/2ELRKfV for a brief overview.

Support for statistical software is available through CTRL. See http://j.mp/ZrBr2Z for CTRL's workshop schedule.

The Department of Mathematics and Statistics offers statistical consulting services, with extensive hours. For the schedule and contact information, see http://j.mp/1EmVqkY.

The library itself offers support for various software. Our librarian is Olivia Ivey, whom I recommend reaching out to as you formulate a question, search for data, and try to put your question in a larger intellectual or policy context. You can schedule time with her at oliviaivey.youcanbook.me.

Intellectual Property

Course content is the intellectual property of the instructor or student who created it, and may not be recorded or distributed without consent.

Course Evaluation

The course evaluation will take place in class towards the end of the semester. We will take time in class to complete the evaluation.

Further Information for American University Students

For further detailed information on the important issues of academic integrity, emergency preparedness, academic support, discrimination, and use of social media, please see here.

Calendar

The calendar below is one possible instantiation of our general approach to the semester. Expect that it will be amended and updated to accommodate our speed, depth, and interests.

Please note: the calendar below refers to chapters in the **print** copy of *R for Data Science*. These differ from the chapter numbering in the online version. Please ask if you have any questions.

Date	Topic	Reading	To Submit
January 17	Introduction: R overview, scientific communication, version control		(in class: *DS* pp. 6-7, online §3.2.4, exercises 1-5)
January 24	R basics, visualization	This syllabus. *DS* ch. 1; *DSO* §3	Install R and RStudio. Join Slack. Submit a `.R` file of *DS* pp. 12-13, online §3.3.1, exercises 1-6. See also pp. 15-16 1-6, pp. 20-21 1-6, and p. 31 1-4.
January 31	Data transformation	*DS* ch. 2-4; *DSO* §4, 5, 6	*DS* p. 49, online §5.2.4, 1-4.
February 7	EDA, R projects	*DS* ch. 5, 6; *DSO* §7-8	*DS* pp. 90-1, online §7.3.4, 1-3 and p. 93, online §7.4.1, 1-2
February 14	Read, wrangle, and tidy data	*DS* ch. 7, 8, 9; *DSO* §9-12. "Tidy Data" paper: [2]	Create an R project in a directory named `ps-4`. Then create your `.R` file for this PS. *DS* p. 124, online §10.5 **4**; p. 129, online §11.2.2 **5**; p. 137, online §11.3.5 **7**; p. 151, online §12.2.1 **2-3**; p. 156, online §12.3.3 **4**; p. 160, online §12.4.3 **1**. Start thinking about a final project!

Date	Topic	Reading	To Submit
February 21	Scientific communication: LaTeX and RMarkdown	*DS* ch. 21, 23-24; *DSO* §26-27, §29-30. http://j.mp/2EO0TEM	Download MacTeX or MiKTeX. Download an editor. Compile `sample.tex` file.
February 28	Packages 1: Structure and metadata	*PKGs* ch. 2, 3, 4	A LaTeX'ed `.pdf` of a paper of yours. A `.pdf` of your solutions to PS5, created with RMarkdown.
March 7	Packages 2: Documentation, testing, `NAMESPACE`, data, files, demos, and releases	*PKGs* ch. 7, 8, 9, 11, 12, 15	Final project preliminary memo
March 14	(Spring Break)		
March 21	Version control with GitHub	*PKGs* ch. 13. Guide to starting with GitHub.	Package `.tar.gz` file
March 28	(Final project work)		Link to package repository
April 4	Relational data: Joins	*DS* ch. 10; *DSO* §13	
April 11	Strings, factors, and dates	*DS* ch. 11-13; *DSO* §14-16	
April 18	Programming: Pipes, vectors, and functions. Evaluations.	*DS* ch. 14, 15, 16; *DSO* §17-20	
April 25	Conclusions and presentations		Final project
May 2-8	Select presentations to DC city policymakers. John A. Wilson Building, 1350 Pennsylvania Ave NW, Washington DC 20004, near Metro Center station on the Red Line.		

References

[1] Hadley Wickham. *R Packages: Organize, Test, Document, and Share Your Code.* O'Reilly Media, Inc., 2015. http://r-pkgs.had.co.nz.

[2] Hadley Wickham et al. Tidy data. *Journal of Statistical Software*, 59(10):1–23, 2014.

[3] Hadley Wickham and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 2017. http://r4ds.had.co.nz/.