# Introduction to Applied Political Data Science
# Government 470/670

Ryan T. Moore*

2022-08-29 at 14:08

## Course Information

Government GOVT 470/670
Introduction to Applied Political Data Science
Tuesdays, 14:30–17:20 EDT
Kerwin Hall 201

## Instructor Information

Ryan T. Moore, Ph.D.
Associate Professor of Government
Office: Kerwin Hall 228
Telephone: 202.885.6470
Homepage: http://www.ryantmoore.org
Email: rtm (at) american (dot) edu
Office Hours: Monday and Thursday 11:00-12:00 or by appointment
(Please use https://calendly.com/ryantmoore to schedule times.)

## Teaching Assistant Information

Ali Amini
Ph.D. Student
Email: aa6718a@american.edu
Office Hours: Wednesday 3:00pm - 5:00pm and by appointment
Location: Café on ground floor, Kerwin Hall

*Department of Government, American University, Kerwin Hall 228, 4400 Massachusetts Avenue NW, Washington DC 20016-8130. tel: 202.885.6470; rtm (at) american (dot) edu; http://www.ryantmoore.org.

## 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 communication.

As designed, this course tends toward the programming side of data science, and less toward the statistical side. For example, we tackle version control and package building rather than machine learning models and natural language processing. We will adapt the plan below to 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 or build an original tool that uses techniques from the course to answer a relevant political data 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 four components: problem sets (50%), a final project and a roughly 10-minute oral presentation and defense of that project (30%), weekly reading quizzes (10%), and engagement in course conversations through attendance, in-class participation, exercises, and Slack participation (10%).

A summary of the course assessments is in Table 1.

| Assignment | Weight | Due date |
|---|---|---|
| Problem sets | 50% | roughly weekly |
| Reading quizzes | 10% | roughly weekly |
| Final paper and presentation | 30% | December 1 |
| Participation (Attendance, Slack, quizzes, paper memo, exercises) | 10% | ongoing |

Table 1: Course Assessment Summary

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 may submit PDF solutions to the course Canvas site as well. You will submit code to GitHub (thus, you will need a [free] GitHub account). You may work with others on the problem sets, but every keystroke of your submission must be your own. You may not copy code or answers from others, but you may develop your code with classmates. This includes all support from resources outside of class. You are responsible for understanding and being able to explain every line of code you submit.

Please name files following the structure `ps00-Moore.R` (lowercase `ps`, two-digit number, hyphen, capitalized surname, no spaces).

### Reading Quizzes

Reading quizzes will take five minutes at the beginning of class, will have roughly five questions, and will be scored 0-5. The student's best 10 reading quiz scores will be counted, and there are no make-up opportunities. Each counted reading quiz counts 1% toward the final mark. You will receive 0.5% simply for completing the quiz in good faith at the appropriate time. The other 0.5% will reflect your score. E.g., if you score 4/5 on a reading quiz, you will earn

$$\underbrace{0.5\%}_{\text{Completion}} + \underbrace{\frac{4}{5} \cdot 0.5\%}_{\text{Score}} = 0.9\%$$

toward the final mark.

### 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.

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

Graduate students are required to submit a GitHub repository as part of the final project.

## Software, Statistics, Data, and Literature Support

The primary software for the course is R. We will use the RStudio IDE to help us manage our work in R. See http://j.mp/2swvN0p 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/2LWQfQF for an introduction to using LaTeX through R (via `tinytex`). For an introduction to the fuller version of LaTeX, see http://j.mp/2EO0TEM. We will utilize GitHub for version control. See http://j.mp/2ELRKfV for a brief overview.

Support for statistical software is available through CTRL. See https://bit.ly/3ABF1w4 for tutorials and https://bit.ly/3ADGIsV for CTRL's calendar.

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. 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 book an appointment with Olivia at https://aulib.info/oivey.

## 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.

Students are not permitted to make visual or audio recordings, including live streaming, of classroom lectures or any class related content, using any type of recording devices (e.g., smart phone, computer, digital recorder, etc.) unless prior permission from the instructor is obtained, and there are no objections from any of the students in the class.

## Course Evaluation

The course evaluation will take place in class towards the end of the semester. Please take a few minutes to provide this valuable feedback.

## Replication Policy

Students must retain copies of all `.R` and `.Rmd` files that include their data processing and analysis for problem sets, labs, and the final project. In keeping with standard practice in the discipline, these files should be able to be run by others, and should reproduce all results the student submits.

## 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 |
|---|---|---|---|
| 31 August | Introduction: Data science, R overview, scientific communication, version control | | (in class: *DS* pp. 6-7, online §3.2.4, exercises **1-5**) |
| 07 Sept | R tidyverse basics, visualization | This syllabus. *DS* ch. 1; *DSO* §3 | Install R and RStudio. Join Slack. Join GitHub repo. Submit a `.R` file of PS1: *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. |

| Date | Topic | Reading | To Submit |
|---|---|---|---|
| 14 Sept | Data transformation | *DS* ch. 2-4; *DSO* §4, 5, 6 | PS2: *DS* p. 49, online §5.2.4, **1-4** and p. 73, online §5.6.7, **5-6** and p. 75, online §5.7.1, **2-3**. |
| 21 Sept | EDA, R projects | *DS* ch. 5, 6; *DSO* §7-8. Bryan on workflows (j.mp/2QmKx0D), and `here` (j.mp/34J2ysP) | PS3: *DS* pp. 90-1, online §7.3.4, **1-3** and p. 93, online §7.4.1, **1-2** |
| 28 Sept | Read, wrangle, and tidy data | *DS* ch. 7, 8, 9; *DSO* §9-12. "Tidy Data" paper: [3] | Create an R project in a directory named `ps4`. 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! |
| 05 Oct | Scientific communication: LaTeX and RMarkdown | *DS* ch. 21, 23-24; *DSO* §26-27, §29-30. http://j.mp/2EO0TEM | Compile `sample.tex`; add sentence "We obtain $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i^2$". Either install `tinytex`, or use Overleaf.com, or download MacTeX/MiKTeX and an editor. Schedule a conversation about your final project. |
| 12 Oct | Packages 1: Structure and metadata | *PKGs* ch. 2, 3, 4. | All: A `.pdf` of your solutions to PS5, created with RMarkdown. Graduates only: A LaTeX'ed `.pdf` of a paper of yours. Start to set up final project. |

| Date | Topic | Reading | To Submit |
|---|---|---|---|
| 19 Oct | Packages 2: Documentation, testing, `NAMESPACE`, data, files, demos, and releases | *PKGs* ch. 5, 7-10, 12-14, 16, 20. (documentation, testing, NAMESPACE, data, installed files, releasing) | Slack package feedback; Final project preliminary memo (`.pdf` to GitHub) |
| 26 Oct | Version control with GitHub | *Happy Git and GitHub for the useR* (Bryan), parts 2 and 3, Installation and Connect Git, GitHub, RStudio. Guide to starting with GitHub. | Install git tools; introduce yourself; set up HTTPS PAT or SSH key. PS6: `uniquer` package `.tar.gz` file; Slack package feedback. |
| 02 Nov | Relational data: Joins | *DS* ch. 10; *DSO* §13 | Link to repo (see PS07 for assignment) |
| 09 Nov | | *DS* ch. 11-13; *DSO* §14-16 | PS8: *DS* p. 177 online §13.3.1 **1**; p. 186-7 online §13.4.6 **1-3**; p. 191 online §13.5.1 **3**, **5** |
| 16 Nov | Strings, factors, and dates. Programming: Pipes, vectors, and functions. Evaluations. Conversation with course alum Tyler Sanders, Red Oak Strategic. | *DS* ch. 14, 15, 16; *DSO* §17-20 | PS9: online §14.3.2.1 **1**, **2**; online §14.3.3.1 **1**, **5**; online §14.4.3.1 **2** |
| 23 Nov | No class meeting (Thanksgiving holiday). | | |
| 30 Nov | Iteration and modeling | *DSO* §21-23; "Two cultures" paper [1] | |
| 07 Dec | Conclusions and presentations | | Final project |
| 14 Dec | Submit by 11:59pm EST | | Final project |

## References

[1] Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199 – 231, 2001.

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

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

[4] 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/.