POL 478H1F: Data management and visualization for political science*

Michael J. Donnelly†

Updated September 6, 2023

This course is designed to expose political science majors and specialists, as well as others with an interest in the topic, to the basics of data visualization. When you are done, you will be able to

1. Critically evaluate data visualizations in the media, understanding not just what the figure says but what decisions went into making it say that and not something else.
2. Download and clean data of a variety of types commonly found in political science and related fields.
3. Use that data to create many different types of figures.
4. Choose the right figures to include in a data-oriented research presentation, taking into account a variety of possible audiences and goals.
5. Understand the ethical dimensions of data collection, analysis, and visualization.

This course relies on R, a free, open-source software that is widely used both in the social sciences and industry. Facility with a statistical software is essential to the practice of quantitative political science and to many careers into which our graduates move.

*A note on spelling: I am originally from the US, so I will usually use a "z" (rhymes with 3, not bed) for the fourth word in this title. You can use what you please.

†Note the "J." There is another (retired) UT political scientist named Michael Donnelly. If you email him, I will not respond.

1 Contact Information

Professor: Michael J. Donnelly (he/him)
Office: 315 Bloor St (Munk Observatory), Room 213
Office Phone: 416-946-8936
Email: mj.donnelly@utoronto.ca
URL: http://www.MichaelJDonnelly.net

2 Logistics

- Lectures: Wednesdays 11-1. Sid Smith 561 (basement, not fifth floor).
- Michael's Office hours: Thursday 10-12. Appointments by email are also available.
- Email policy: I endeavor to respond to all emails within **two working days**. If I have not gotten back to you by then, feel free to send a reminder. If you email me about an assignment fewer than two working days before it is due, I cannot guarantee that I will respond in time for you to use my comments.

3 Course Requirements

Your final grade is based on participation, in-class exercises, two types of assignments, and a final project.

- **Participation** (total of 10%): This is a lab-based course, mixing lecture, discussion, and group activities. Participation is therefore an essential part of the learning process. Attendance is mandatory. If you expect to miss class, or if you miss class unexpectedly, I expect you to let me know quickly. You can also participate by discussing your classmates' visualization critiques on the discussion board or in class.
- **Visualization critiques** (15%): In each of weeks 2-11, you will write a one paragraph critique of some data visualization you see online, providing a link to the figure.[1] You should point out strengths and weaknesses of the figure and speculate about decisions that the creator made when producing the figure. You will post these on a discussion board by 9pm the night before class to share with your classmates. Each critique is worth 1.5 points. You should be prepared to discuss your critique in class.
- **Problem sets** (25%): There will be five problem sets of varying length. These will use real data and produce real, potentially interesting figures. I will grade these generously.[2]

[1] If it is paywalled, save a copy of the figure and upload it too.

[2] That is, you can get full credit without getting everything right. You can get partial credit if you try and fail but you explain in text what you are trying to do in the code. As in "I want to make a scatterplot here with the points scaled to represent the population of the cities, but the scaling doesn't seem to be working. I'm using aes(size = sqrt(population)), but that isn't doing what I need and I can't figure out why." That would get you a fair amount of credit.

- **Timed exercises** (15%): There will be three timed in-class coding exercises where you will be asked to complete basic tasks without outside aid. Think of these as tests of your fluency in R.

- **Final project** (35%): The final project will be a report and presentation on a data set or data sets of your choice. The topic must be relevant to politics,[3] and you must have a practical way of getting access to the relevant data. We will discuss many potential sources in class. The report should have a minimum of 10 figures (though this number could get much larger depending on the data), with an introduction describing the topic and analytic importance of the data, a discussion of the data source, an analysis of each figure, and a discussion of the overall findings. The discussion of the source should include a description of the data source(s), including the measurement, methodological, and ethical issues involved. Details on measurement and methods, recoding decisions, etc. can be included in an appendix, but ethical considerations should be included in the main text. **The report is due on December 13.**

  - The report should not have any evidence that it is a class paper. Ideally, this is something you would put up on your LinkedIn page or attach to a grad school application. The report should be html or pdf,[4] with the decision based on your comfort and audience. If you want to think of this as something you'd send as a writing sample in an academic context, then you should produce a pdf. If you want to send this to an employer as evidence of your tech skills, an html makes more sense.
  - You will give a summary in an oral presentation in the final two weeks of class (we will schedule these sometime early in the semester). The presentation **will account for $\frac{10}{35}$ points** from the final project.
  - This project may be cooperative. That is, you may work with one other person on the same data, producing a report and presentation. If you do this, I expect twice as much work. You should clearly document for yourselves and for me who contributed which parts of the report and the presentation and I will give separate grades.
  - In Week 4, you will submit a 1-2 paragraph memo describing the data you plan to use and confirming that you have obtained the data. Completion of this memo is worth 1/35 points for the final project.

4 General rules for problem sets

Submission All assignments are due by 9am on the day of class. They should be submitted as a pdf or html document.

Plagiarism & cheating Cooperation on the problem sets is encouraged, but all students should understand the answers they submit (i.e., be able to explain them) and write up the answers separately (i.e., don't copy and paste). If you copy someone's code, even if you make changes, it will be obvious to a more experienced coder (me).

[3] I will be very generous in interpreting "relevant to politics." If you are unsure if a topic is political, ask.
[4] I will show you how to do both within R

Plagiarism and other violations of academic integrity will not be tolerated. See the university policies[5] for more details.

Late assignments Late assignments will receive deductions of 20% per day. When submitting to Quercus, recognize that it can be slow, and that can sometimes push your submission past the deadline. Similarly, it is sometimes down for maintenance. I will not grant extensions for normal Quercus delays, so make sure to leave yourself a time cushion.

Grade appeals must be made within two weeks of receiving the grade. They must include a 100-200 word written statement of why the assignment deserves to be re-graded. The grade will change only in cases where the second grading is more than 10 points different from the first (i.e. a 60 will not be changed unless the second grading produces a score of 70+ or 50-). Grades can go up or down on the second grading.

Lost assignments Keep a backup of everything. Have you backed up your computer to the cloud or to an external hard drive this week? No? Go do that now and then come back and finish reading this.

5 Using R

Learning R is required for this course. I will conduct classes on the assumption that you work using RStudio, a front-end interface for R. It is freely available for any of the major operating systems. **Before** the first class, you should download

- R from https://cran.r-project.org/
- RStudio from https://www.rstudio.com/products/RStudio/ (note that the free desktop version is the one you're looking for).
- The textbook's package, for which instructions can be found at https://socviz.co/

If you have any trouble downloading and installing these, please let me know in the first class. In the final paper, you should cite R, RStudio, and any packages you use.[6]

Use of generative AI You may use ChatGPT or other generative artificial intelligence tools to help you with your code. I suggest not using it for a first draft of the code, but when you are stuck or when you have bugs, it is quite good at suggesting solutions. If you do so, document it. Each time you use it, put a comment in your code of the following form:

```
## Typo solved by pasting code into ChatGPT.
```

or

```
## Two level for loop structure corrected with ChatGPT.
```

[5] http://www.utoronto.ca/academicintegrity

[6] There are ways of automating this. Ask me about them if you want to save time.

In the final project, you should cite any generative AI tools that you use in the works cited section.

Norms about citation of such tools are still developing, but for now, this is how we will do it for take-home work. In-class assessments will **not** allow the use of generative AI.

6 Readings

The main textbooks for this course are

- Kieran Healy. *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press, 2018. URL: https://socviz.co/,
- Rohan Alexander. *Telling Stories with Data*. Toronto: Bookdown, 2023. URL: https://www.tellingstorieswithdata.com/index.html.

Both books are available in print (from Princeton UP and Chapman, respectively) or as (free) web sites.

Healy covers both basic visualization and an introduction to R. It does not assume any substantial statistical or programming background. Though the author is a sociologist, the examples come from political science, sociology, economics, and related fields. He is Irish, but has lived in the US for a while and the book was published in the US. Unsurprisingly, then, the book uses many American examples, but it also includes examples relevant to comparative politics and international relations. Where possible, I have inserted additional Canadian examples.

Alexander does less introduction to R, and covers a somewhat different spectrum of topics. He is Australian and lives in Toronto, so his examples are a bit less American-centric (though the US has a lot of data, so...).

Other textbooks that I find useful (available for free online):

- Hadley Wickham and Garret Grolemund. *R for Data Science: Visualize, Model, Transform, Tidy, and Import Data*. Sebastopol, CA: O'Reilly, 2017. URL: https://r4ds.had.co.nz/
- Alex Douglas et al. *An Introduction to R*. Online: Bookdown, 2022. URL: https://intro2r.com

Here is a link to a list of more than 200 free books about R: https://www.r-bloggers.com/2021/05/big-book-of-r-has-over-200-books/. Some are good. Some are not.

Reading Healy Each week, you will be assigned to work through a section of Healy. This should be done with your computer open and RStudio running. You do not have to run every command as you read, but do NOT skip over them. Unlike reading for many purposes, reading to understand these concepts requires reading computer code. When you do run them, you will be tempted to copy and paste. Try to resist that temptation and write the code in an R or RMarkdown script and then execute it. It is part of the process.

Supplementary reading I have noted additional readings (denoted with a star) that might help you figure out what Healy, Alexander, or I am saying or that introduce related concepts or approaches. You also can and should regularly Google the basic concept you are reading about and find a wealth of good (and lots of bad too) resources.

Web resources that might be useful Below, I've listed a variety of blogs and resources where you should poke around if you are looking for visualizations or interesting data.

- City Hall Watcher https://cityhallwatcher.com/#free[7]
- Healy's blog: https://kieranhealy.org/blog/.
- Andy Gelman's blog: https://statmodeling.stat.columbia.edu/
- The data journalism site https://fivethirtyeight.com/
- Statcan's infographic feed https://www150.statcan.gc.ca/n1/en/catalogue/11-627-M
- Eric Grenier's Canadian election website https://www.thewrit.ca/
- Patrick Fournier's Canadian election website https://338canada.com/
- R-Ladies https://rladies.org/
- A nice blog with occasional useful examples https://rforpoliticalscience.com/
- *The Economist*'s Instagram: https://www.instagram.com/theeconomist/?hl=en

Writing up the problem sets Most of the questions in the problem sets require both data manipulation **and** written responses or interpretations. This may seem pointless in some cases, but the questions are designed in part to get you used to writing based on figures. Part of writing such work is saying things in multiple ways. To see this, pick up a newspaper article accompanied by a data visualization, and examine how the author expands on the figure, interprets it, or highlights the key messages.

7 Course Outline

Required readings/tasks marked with a •, suggested readings with ∗.

Week 1: Introductions, logistics, introduction to R September 13

- This syllabus. **No, really, read the whole thing! I put it all here for a reason.**
- Healy Preface and Ch. 1: Make sure to follow the instructionsat the end of the Preface for downloading and installing R, RStudio, and the relevant packages.
- ∗ https://www.r-bloggers.com/2020/07/getting-started-with-r-markdown-guide-and-cheatsheet/
- ∗ https://www.dataquest.io/blog/rstudio-tips-tricks-shortcuts/

Week 2: Getting comfortable in R September 20

[7] A local journalist maintains a substack that covers Toronto City Hall politics. You should go to that web site and click on "Request a Free Subscription." He often graphs interesting data from the city. You can use this as a source for your visualization critique assignments and for inspiration for the final project. Some assignments may replicate and extend his visualizations.

- Healy Ch. 2
- Healy Appendix A.1 and A.3
* Alexander Ch. 3
* https://intro2r.com/install-rm.html

Week 3: Data cleaning September 27

- Healy Ch. 3.1-3.2
- Alexander Ch. 11
- Hadley Wickham. "Tidy Data". In: *Journal of Statistical Software* 59.10 (2014), pp. 1–23. URL: http://www.jstatsoft.org/
* https://www.dataquest.io/blog/load-clean-data-r-tidyverse/
* https://towardsdatascience.com/what-is-tidy-data-d58bb9ad2458
* https://www.garrickadenbuie.com/project/tidyexplain/#tidy-data
* Chapter 3 of Carrie Wright et al. *Tidyverse Skills for Data Science*. Online: Bookdown, 2021. URL: https://jhudatascience.org/tidyversecourse/
* Christopher Ingraham. *An Alarming Number of Scientific Papers Contain Excel Errors*. Washington, DC, Aug. 2016. URL: https://www.washingtonpost.com/news/wonk/wp/2016/08/26/an-alarming-number-of-scientific-papers-contain-excel-errors/
- **ASSIGNMENT 1 DUE**

Week 4: Visualization theory October 4

- Healy Ch. 1
- Steven L. Franconeri et al. "The Science of Visual Data Communication: What Works". In: *Psychological Science in the Public Interest* 22.3 (2021), pp. 110–161. ISSN: 21600031. DOI: 10.1177/15291006211051956
- Parts 1 and 3 of Jason Forrest. "W. E. B. Du Bois' Staggering Data Visualizations Are as Powerful Today as They Were in 1900". In: *Nightingale* (July 2018). URL: W.E.B.DuBois'staggeringDataVisualizationsareaspowerfultodayastheywerein1900(Part1)
- Andrew Gelman. *A Checklist for Data Graphics*. 2022. URL: https://statmodeling.stat.columbia.edu/2022/03/15/a-checklist-for-data-graphics/ (visited on 03/15/2022)
* Masataka Okabe and Kei Ito. *Color Universal Design (CUD): How to Make Figures and Presentations That Are Friendly to Colorblind People*. 2008. URL: https://jfly.uni-koeln.de/color/
- **TOPIC SELECTION MEMO DUE**
- **In-class exercise 1**

Week 5: Univariate description October 11

- Healy Ch 3-4
- Alexander Ch 13
- https://kieranhealy.org/blog/archives/2021/12/19/comparing-distributions/ [make sure to click on "more" beneath the code blocks.]

Week 6: Bivariate description — October 18

- Healy Ch 5
* Andrew Gelman and Antony Unwin. "Infovis and Statistical Graphics: Different Goals, Different Looks". In: *Journal of Computational and Graphical Statistics* 22.1 (2013), pp. 2–28
- **ASSIGNMENT 2 DUE**

Week 7: Maps and gifs — October 25

- Healy Ch 7
- https://intro2r.com/loops.html

Week 8: Lab class — November 1

- Healy Ch 8
- **ASSIGNMENT 3 DUE**
- **In-class exercise 2**

No class: Reading Week. Work on Assignment 4, projects — November 8

Week 9: More maps — November 15

- Healy Ch 6
* https://mhallwor.github.io/_pages/Tidyverse_intro
- **ASSIGNMENT 4 DUE**

Week 10: Lab class — November 22

- Alexander Ch 5
- https://kieranhealy.org/blog/archives/2018/03/24/making-slides/
- https://statmodeling.stat.columbia.edu/2014/12/01/quick-tips-giving-research-presentations/
- **In-class exercise 3**

Week 11: Presentations — November 29

- **ASSIGNMENT 5 DUE**

Week 12: Presentations — December 6

No class: **FINAL REPORT DUE.** — December 13