

Data Visualization

QMSS GR5063 - Columbia University

Spring 2021

Lecture: Mondays 8.10 - 10pm (but see weekly schedule)

Location: Zoom

Instructor: Thomas Brambor

thomas.brambor@columbia.edu

Office Hours Mon 8.30am-9.30am EST

TA1: Nicolas Rojas nr2670@tc.columbia.edu

Office Hours Thu 9.00am-10.00am / Sections Wed 20.00pm-21.00pm EST

TA2: Yifei Dong yd2564@tc.columbia.edu

Office Hours Fri 8.30am-9.30am / Sections Wed 20.00pm-21.00pm EST

Quick Links

- [Course Description](#)
- [References and Resources](#)
- [Requirements and Assessments](#)
- [Policies](#)
- [Lecture Topics](#)
 - [Part 1 - Introduction and Plotting with ggplot2](#)
 - [Part 2 - Working with Spatial Data](#)
 - [Part 3 - Text Mining and Visualization](#)
 - [Part 4 - Getting Dynamic and Interactive](#)

Course Description

This course will provide a hands-on introduction to visualizing a wide variety of different types data. It is aimed at graduate students for the Master of Arts Degree in Quantitative Methods in the Social Sciences (QMSS). The course combines tutorial style introductions to different software tools and visualization packages centered around the R language, practical tips on analyzing and presenting real data, and some readings and discussion of the principles of data visualization. In the course, we progress from a set of basic static graphs to mapping geographic data, text, social networks, and other forms of data in dynamic and interactive displays. Examples will be drawn from a variety of disciplines in and beyond the social sciences, and you will be encouraged to work with your own data to create custom graphics. For the capstone project, students incorporate several of the techniques learned in class to visualize data of their own choice and present a website as the final product.

Course Website

All lecture materials, exercises, and (links to) readings will be made available in the GitHub course repository.

The materials and topics indicated below are a provisional road map that will be adjusted to the needs of the students. I will let you know well ahead of time of any changes.

Communications

For all questions to the members of the teaching team, we will be using a discussion forum on Campuswire. The forum will be used to exchange questions about lectures, assignments, software etc. Students are encouraged to help each other!

Find our class page at: <https://campuswire.com/p/G0F020D17>

If you have not joined yet, please use the code 9309 to sign up for the course.

Students are asked to customize their Campuswire notification preferences to receive immediate (ASAP) notifications of messages and announcements through the third-party provider of choice (e.g. email, SMS/text). Students are also asked to log into the course regularly (more than twice a week) and check announcements and the Campuswire inbox immediately upon logging in to stay on top of developments in the course as they occur.

Please send all emails and messages to the instructor and teaching assistants through Campuswire. Messages sent through the Campuswire Inbox (Send a Message) feature will be answered within 24 hours during the week and within 48 hours on weekends. Please consider these response times when asking about assignments etc.

References and Resources

Books

There are no required books for the course. All required readings will be provided as PDFs or links. However, here are some books that you may find useful in addition to the lectures and course readings.

- Winston Chang. *R Graphics Cookbook: Practical Recipes for Visualizing Data*. O'Reilly Media, 2n ed. – Detailed tutorial style book for using graphs in R using the base package plots and ggplot2. The book is available for free.
- Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, 3rd ed. – Intro to ggplot2 by Hadley Wickham, the creator of the package. The book is available for free.

- Scott Murray. *Interactive Data Visualization for the Web: An Introduction to Designing with D3*. O'Reilly Media, 2nd edition – Beginners guide to interactive data visualization for the web using D3.] The first edition is available for free.
- Alberto Cairo. *The Functional Art: An Introduction to Information Graphics and Visualization*. New Riders, 1st ed.: Graphics journalist Alberto Cairo provides a very readable intro to understanding data visualization, how to use it to communicate with your audience, and helpful how-to guides to improve your own data graphics.
- Alberto Cairo. *The Truthful Art: Data, Charts, and Maps for Communication*. New Riders, 1 edition – Following up on his earlier book, Cairo provides a set of principles for data visualization and puts them to practice by showing loads of examples of good (and bad) graphical displays.
- Cole Nussbaumer Knafl. *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley, 1 ed. – The book provides many excellent examples of how to improve visual displays to get your point across and communicate effectively with data. All plots are done in Excel, showing that it can be done (though I won't recommend).
- Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2nd ed. – Edward Tufte is one of the pioneers in the field of data visualization, and this book is one of his most important works. The book is on the theory and design of data graphics. Originally printed in 1983 it won't help you directly with your computer exercises, but can give you some insight into what makes a data graphic good or not.

Free Online Resources

R, RStudio, and R Markdown

- Try R. In-browser, interactive online tutorial. Particularly useful if you have not used R (much) before.
- Cheat sheets for data wrangling, data visualization, general use of R, R Studio, R Markdown etc.
- R Studio resources for R Markdown. Get started here with markdown.
- Datacamp courses on Intro to R Lots of courses at all skill levels from basic intro to R to leveling up your skills as an R programmer.

Getting and transforming data in R

For this course on visualization, I won't have time to cover in detail all the steps to access, clean, "munge" and organize data.

For an excellent overview of the topic, I recommend the following book:

- Wickham, H., & Golemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (1 edition). O'Reilly Media. – Great as introduction on how to use R. From the creator of many R packages that we use in the course, this will help with the usual tasks of data import and management, modeling, and some visualization. Book is available for free online.

For a more advanced treatment of coding in R, consider the following book:

- Wickham, H. (2014). *Advanced R* (1 edition). Boca Raton, FL: Chapman and Hall/CRC. Book is available for free online

Lastly, Datacamp is a great option to complement your skills on importing, cleaning, and manipulating data.

Git and GitHub

- Git
 - Official git command line and GUI clients, official documentation
- Clients
 - Desktop Client for Mac and Windows. Comes standard with github and is easy to use, but limited.
 - Gitkraken. Professional git client. Students get the full-version for free. Recommended!
- Tutorials
 - Learning resources for git and github
 - Setting up commandline git
 - Interactive tutorial to learn branching
 - Hello World - GitHub for the non-programming beginner.
 - Guides at GitHub
 - Git and Github guide from plot.ly Extensive screen-shot guided intro to Git, Github, Git in RStudio and GitHub pages.
 - Pro Git - a full book with lots of details
 - Datacamp - Introduction to Git for Data Science

Coding Help Sites

- <http://stackoverflow.com/> Programming Q&A site. Excellent first stop if you have questions on coding. Search for keywords and restrict your queries by adding tags about the coding language or package in square brackets, e.g. [R], [ggplot], or [shiny].
- <http://stats.stackexchange.com/> A stackoverflow off-shoot with a bit more focus on conceptual questions in statistics.
- <http://rseek.org/> Search engine for R-related stuff, including tutorials and code.

Data Visualization

- <http://www.thefunctionalart.com/> Alberto Cairo, an infographics and data visualization journalist, also has a great blog.
- <http://andrewgelman.com/> Andrew Gelman's insightful comments on social science analyses go beyond data visualization, but are well worth your time.
- <https://flowingdata.com/> Nathan Yau, a statistician with a knack for visualization. Author of a few data visualization books. Blog has great examples and some tutorials.
- <http://www.r-graph-gallery.com/> Need inspiration? R graph gallery you can just scroll through graphs - good ones, and not so good ones - and get the code behind them.
- <http://www.informationisbeautiful.net/> Another site for inspiration, with less code but nicer visuals.
- <http://chartporn.org/> Slightly corny name, but good links to beautiful charts, graphs, maps, and interactive data visualization tools around the web.

Requirements and Assessments

Requirements

Basic knowledge of statistics on the level of an introductory level graduate course in statistics or econometrics is assumed. The focus is on data visualization, but some statistical concepts appear here and there.

The course uses the software package R for most exercises. The program R itself can be downloaded for free at <http://cran.r-project.org/>. Intermediate familiarity with the software, in particular with regards to importing, cleaning, and reshaping data is assumed. Knowledge of specific packages and other software tools will be built throughout the course.

You will need to have access to your own computer to install software and packages, do your assignments etc. I highly recommend bringing your laptop to class to follow along the coding tutorials and examples.

In addition, I will provide some information on how to visualize data in Python in parallel.

Assessments

1. Final group project (30%): A final group project (3-4 students) presented in the form of a website and accompanying project book. You will analyze data of your own choosing and report the results using (1) static images based on `ggplot2`, (2) maps using geospatial data, (3) visualizations of

text analyses, and/or (4) network visualizations AND (5) prepare a hosted, interactive display of some of your visualizations. There will be in-class presentations of the final projects (if class size allows).

2. Assignments (60%): There will be 4 individual assignments. These assignments will be due 7 days after being handed out and returned graded by the teaching assistants. The assignments will ask you to use the specific visualization techniques we cover in the individual subparts of the course.
3. Class participation & Commentary on other student project (10%): You will be asked to comment on another student's visualization techniques at one point in the course. Campuswire participation will also be a part of the grade. Both live and asynchronous participation in lectures are possible and not participating live is not in any way detrimental to the participation grade.

Auditing and Pass/Fail Grades

Auditing students (R grade) will need to complete *one* of the homework assignments 2 to 4 and should participate regularly throughout the semester (by viewing lectures and being active in Campuswire). There is no expectation to complete a final project.

Pass/Fail students (P/F grade) will need to complete all requirements as all other for credit students. A P grade will be assigned to any student who received a passing letter grade.

Policies

Academic Integrity

This course is based on the principles of academic integrity established by Columbia University and agreed to by each student. The same rules hold in this course. Academic dishonesty will not be tolerated. All submitted work must be your own work and properly cited.

The full guidelines on academic integrity as well as a review of how or what to cite, can be found here: <http://gsas.columbia.edu/academic-integrity>

Just to be extra clear, here are some pointers that apply to this course:

- No assignment, except the final project, is to be collaborative.
- You may consult with others, but any work you hand in must be your own.
- Do not copy another individual's work, answers or ideas.
- Do not allow another individual to copy your work, answers or ideas.
- Disciplinary action follows for those that choose to disobey these instructions.

Late Assignment Policy

Late assignments are detrimental to progress in the course, because earlier assignments build on later ones. For each day late, the assignment will be reduced 1/3 of a grade. That is, one day late, it goes from A to A-, two days late A- to B+ and so on. Once solutions are posted, submission of the assignment is no longer possible.

Other

Lecture slides will be made available on the course website. However, I believe that learning and understanding is well served when you need to aggregate and structure your notes yourself, so I suggest you do so as well.

Lecture Topics

Part 1 - Principles of Data Visualization and Plotting with ggplot2

Week 1: (Jan 11) Getting Started. Introduction to ggplot2 and R Markdown

- **On your own:**
 - Install R and R Studio on your own computer. Try out R Markdown (use the tutorial to get familiar).
 - Set up your own GitHub account. Familiarize yourself, if necessary with the suggested tutorials.
- **Group work:** Exchange contact details with your group.
- **Exercise:** Go through the babynames exercise. Try the suggested extensions.
- **Recommended Reading:**
 - Wickham, H. (2019). ggplot2: Elegant Graphics for Data Analysis (2nd ed.). New York, NY: Springer. The book is available for free.. There is no need to read the entire book. Just use the book as a reference and read the parts that complement the lecture.

Martin Luther King Jr. Day. NO CLASS ON MONDAY, JANUARY 18.

Week 2: (Jan 25) Continuing with ggplot2. Additional chart types. Grammar of Graphics. Understanding additional layers - coordinates, facets. Additional data visualization and perception recommendations.

- **On your own:**
 - Continue learning ggplot2.
- **Exercise:** Go through the “Guns and Deaths in America” exercise.
- **Recommended Reading:**
 - Wong, B. (2010a). Points of view: Design of data figures. Nature Methods, 7(9), 665-665. <https://doi.org/10.1038/nmeth0910-665>

- Wong, B. (2010b). Points of View: Gestalt principles (Part 1). *Nature Methods*, 7(11), 863-863. <https://doi.org/10.1038/nmeth1110-863>
- Wong, B. (2010c). Points of View: Gestalt principles (Part 2). *Nature Methods*, 7(12), 941-941. <https://doi.org/10.1038/nmeth1210-941>
- Continue using/reading Wickham (2016) as a reference to ggplot2.
- **Further Reading:**
 - Ware, C. (2012). *Information Visualization, Third Edition: Perception for Design* (3 edition). Waltham, MA: Morgan Kaufmann. (especially chapter 5 and the section on preattentive processing)
 - Cleveland, W. S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387), 531-554. <https://doi.org/10.2307/2288400>

Week 3: (Feb 1) Continuing with ggplot2: Refining plots, themes, publication-ready

- **Recommended Reading:**
 - Tufte, E. R. (2001) [first published 1983]. *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, Connecticut: Graphics Press. Chapters 1 - 3.
- **Further Reading:**
 - Tufte, E. R. (2001) [first published 1983]. *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, Connecticut: Graphics Press. Remainder of the book.

Week 4: (Feb 8) Making ggplot graphs dynamic and interactive

- **Recommended Reading:**
 - Hans Rosling’s TED Talk
 - Carson Sievert - plotly for R - A full reference website for using plotly in R.
 - Yi, J. S., ah Kang, Y., & Stasko, J. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6), 1224–1231. <https://doi.org/10.1109/TVCG.2007.70515>

Assignment 1: Plotting with ggplot.

Part 2 - Working with Spatial Data

Week 5: (Feb 15) Working with Spatial Data in standard data frames and ggplot using the ggmap package

- **Recommended Reading:**
 - Meirelles, I. (2013). *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*. Rockport Publishers. (chapter 4 - Maps)

- Kahle, D., & Wickham, H. (2013). *ggmap: Spatial Visualization with ggplot2*. *R Journal*, 5(1), 144–161.

Week 6: (Feb 22) tmap, sp and raster packages. Projections, polishing maps

- **Recommended Reading:**
 - 7 Deadly Sins of (Academic) Data Visualisation - What not do in your maps. The spatial.ly blog is also a good resource for interesting maps and code.
- **Further Reading:**
 - Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R* (2nd ed. 2013 edition). New York: Springer. - A comprehensive and hands-on treatment of how to use spatial data in R for analysis and visualization. This goes well beyond what we cover, but may be helpful as a reference to move further on topics of your interest related to spatial analysis.

SPRING RECESS March 1 to March 5. NO CLASS ON MONDAY, MARCH 1.

Week 7: (Mar 8) Interactive Maps using Leaflet

- **Recommended Reading:**
 - R Studio provides a great tutorial for the basics of using the *Leaflet* package.
 - A presentation of the features of Leaflet at the RStudio 2017 Conference
- **Further Reading:**
 - Roth, R. E. (2013). Interactive maps: What we know and what we need to know. *Journal of Spatial Information Science*, 2013(6), 59–115. <https://doi.org/10.5311/JOSIS.2013.6.105>

Assignment 2: Visualizing geospatial data.

Part 3 - Text Mining and Visualization

Week 8: (Mar 15) Bag of Words. Cleaning and preprocessing, data structures. Frequency visualization, word clouds.

- **Recommended Reading:**
 - Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267–297. <https://doi.org/10.1093/pan/mps028>
 - Silge, J., PhD, & Robinson, D., PhD. (2017). *Text Mining with R: A tidy approach* (1 edition). O'Reilly Media. <http://tidytextmining.com/>
- **Further Reading:**

- Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval (1 edition). Cambridge University Press. Available for free at <http://nlp.stanford.edu/IR-book/> - This book is concerned with information retrieval (think search engines) and includes some treatments of the fundamental theoretical ideas behind text mining.
- Rule, A., Cointet, J.-P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. Proceedings of the National Academy of Sciences, 112(35), 10837–10844. <https://doi.org/10.1073/pnas.1512221112>

Week 9: (Mar 22) Word clustering, sentiment analysis, topic detection and more.

- **Recommended Reading:**
 - Quanteda package vignette on text analysis
 - tm package vignette

Assignment 3: Analyzing text as data..

Part 4 - Dynamic and Interactive Displays using D3 and Shiny

Week 10: (Apr 29) Network visualizations. Twitter API.

- **Recommended Reading:**
 - Meirelles, I. (2013). Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations. Rockport Publishers. (chapter 2 - Networks)
 - Network visualization with R by Katherine Ognyanova (Rutgers University) - is a full workshop on how to work with and visualize networks in R, including GitHub code.
 - ggnet2: network visualization with ggplot2 - a visualization function to plot network objects as ggplot2 objects.
 - James Curley’s slides on Interactive and Dynamic Network Visualization in R
- **Further Reading:**
 - A comprehensive(!) curated list of network visualization info
 - Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US election: divided they blog. In Proceedings of the 3rd international workshop on Link discovery (pp. 36–43). ACM. <http://dl.acm.org/citation.cfm?id=1134277>

Assignment 4: Interactive visualizations of networks.

Week 11: (Apr 5) R Shiny and Data Driven Documents (D3). Hosting your own interactive visualization.

- **Recommended Reading:**

- R Shiny tutorial: <https://shiny.rstudio.com/tutorial/>. I recommend completing the three part introductory video tutorial. The lecture will dovetail closely so you can go back and fill in the blanks.
- R Studio provides an excellent cheat sheet to get started with R Shiny
- **Further Reading::**
 - Interactivity in RMarkdown with D3.js (by James Curley)

Week 12: (Apr 12) R Shiny continued. Refining interactive apps.

- **Recommended Reading:**
 - R Shiny tutorial: <https://shiny.rstudio.com/tutorial/>. The third part of the video tutorial is closest to the lecture.
 - How to use CSS to style your shiny app: <https://shiny.rstudio.com/articles/css.html>
- **Further Reading::**
 - CSS and HTML tutorial at codecademy

Week 13: (Apr 19) Final Student Presentations

- Final Project Websites and Comments Assignment

Project presentation websites due on Apr 19 at noon. Student comments on other projects due on Apr 20 at 8pm. Final student projects due on April 23.