Data Processing & Visualization
QMSS G4063
Columbia University

Spring 2017
Lecture: Mondays 6.10 - 8pm (but see weekly schedule)
Location: Kent Hall 413

Instructor: Thomas Brambor tb2729@columbia.edu IAB 270C Mon 11-12pm

TA1: Luronne Vaval lv2301@tc.columbia.edu IAB 509 Fri 11-12pm

TA2: Pablo Vicente Juan pv2288@columbia.edu IAB 509 Tue 6-7pm

---

Quick Links

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.

Course Website

We will be using a Q&A forum on piazza: https://piazza.com/columbia/spring2017/qmssg4063

I have added all enrolled students at the beginning of the term; for late-joiners please add yourself. The forum will be used to exchange questions about lectures, assignments, software etc. Students are encouraged to help each other!

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

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, 1st ed. – Detailed tutorial style book for using graphs in R using the base package plots and ggplot2. Some parts of the book's content are free.
- Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer, 2nd ed. – Intro to ggplot2 by Hadley Wickham, the creator of the package. 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, 1st ed. – Beginners guide to interactive data visualization for the web using D3. Beginners guide to interactive data visualization for the web using D3. The book is available for free.
- Alberto Cairo. *The Functional Art: An Introduction to Information Graphics and Visu- alization.* 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 Knaflic. *Storytelling with Data: A Data Visualization Guide for Business Professionals.* Wiley, 1 ed. – The books 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.
- Edward R. Tufte. The Visual Display of Quantitative Information. Graphics Press, 2 ed. – Edward Tufte is one of the pioneers in the field of data visualization, and this book is one of his most famous 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

- IDRE at UCLA has lots of tutorials, code examples, for R and other statistical packages.
- 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.
- Awsome-R A curated list of great R packages and tools.
- Wickham, H., & Grolemund, 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.

Git and GitHub

- Git
  - Official git command line and GUI clients, official documentation

- Clients
  - Desktop Client for Mac and Windows
  - Sourcetree - another free visual Git
- Tutorials
  - Setting up git
  - Try.github
  - 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

Coding Help Sites

- http://stackoverflow.com/ Programming Q&A site. Excellent first stop if you have questions on coding. Searching 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/. Some familiarity with the software, in particular with regards to importing 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.

Assessments

1. Final report (40%): A final report in the form of a conference paper. 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 OR network visualizations and (4) 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 (50%): 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. In addition, course participants should aim to do the circulated readings before class and take part in the discussion during the session and online.

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

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.

Other

Turn off or silence your cell phones prior to the beginning of class. I reserve the right to answer all calls (your's, not mine) received during class time and let your friends know what you are learning that day.

Feel free to use laptops in class - in fact, I encourage it. Respecting your classmates and myself, please refrain from using Facebook, shopping sites or other random distractions during class.

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

I am most easily reached via email under tb2729@columbia.edu. I will try to respond within 24h of the email, however, may take up to 48h. Please consider these response times when asking about assignments etc.

Lecture Topics

Part 1 - Introduction and Plotting with ggplot2

Week 1: (Jan 23) 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).
- **Group work:** Exchange contact details with your group.

- **Exercise:** Go through the babynames exercise. Try the suggested extensions
- **Recommended Reading**:
  - Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (2nd ed.). New York, NY: Springer. - The book is available for free.

Week 2: (Jan 30) Continuing with ggplot2. Addtional chart types. Grammar of Graphics. Understanding additional layers - coordinates, facets. Additional data visualization and perception recommendations.

- **On your own:**
  - Set up your own GitHub account. Familiarize yourself, if necessary with the suggested tutorials.
  - Continue learning ggplot2 with the data exercise on "Guns and Deaths in America"
- **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
  - 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
- **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)

Week 3: (Feb 6) Continuing with ggplot2: Refining plots, themes, publication-ready

- **Exercise:** Assignment 1: Plotting with ggplot using WorldBank Datset ID4D (Due Tuesday, Feb 14)
- **Recommended Reading:**
  - Tufte, E. R. (2001) [first published 1983]. *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, Connecticut: Graphics Press.

Part 2 - Working with Spatial Data

Week 4: (Feb 13) 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.

`Feb 20 - President's Day. No Class`

Week 5: (Feb 27) tmap, sp and raster packages. Projections, polishing maps

- **Exercise:** Assignment 2: Mapping AirBnBs in NYC.
- **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.

  - 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.

Part 3 - Text Mining and Visualization

Week 6: (Mar 6) 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 from http://nlp.stanford.edu/IR-book/ - This books 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

`SPRING RECESS March 13 - March 17. No class.`

Week 7: (Mar 20) Word clustering, sentiment analysis, topic detection and more.

- **Recommended Reading:**
  - Quanteda package vignette on text analysis
  - tm package vignette
- **Exercise:** Assignment 3: Analyzing NY Op-Eds.

Part 4 - Getting Dynamic and Interactive

Week 8: (Mar 27) Making ggplot graphs dynamic and interactive using plot.ly

- **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

Week 9: (Apr 3) Interactive Maps using Leaflet

- **Recommended Reading:**
  - R Studio provides a great tutorial for the basics of using the `Leaflet` package.
  - A recent presentation of the current 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

Week 10: (Apr 10) 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
- **Exercise:** Assignment 4: Interactive visualizations.

Week 11: (Apr 17) R Shiny and Data Driven Documents (D3). Hosting your own interactive visualization.

- **Recommended Reading:**
  - R Shiny tutorial: https://shiny.rstudio.com/tutorial/. I recommmended completeting 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 24) 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 codeacademy

Week 13: (May 1) Final Student Presentations

- Final Project Websites and Comments Assignment