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

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

All course materials will be made available on `https://courseworks.columbia.edu/`.

In addition, 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, 1 edition : Detailed tutorial style book for using graphs in R using the base package plots and ggplot2.
- Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, 2 edition – Intro to ggplot2 by Hadley Wickham, the creator of the package.
- Scott Murray. *Interactive Data Visualization for the Web: An Introduction to Designing with D3*. O'Reilly Media, 1 edition – Beginners guide to interactive data visualization for the web using D3.
- Alberto Cairo. *The Functional Art: An Introduction to Information Graphics and Visualization*. New Riders, 1 edition edition – 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 edition – 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 edition – 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 and R Studio

- `http://www.ats.ucla.edu/stat/r/` – Lots of tutorials, code examples, for R and other statistical packages.

- http://tryr.codeschool.com/ – Try R. In-browser, interactive online tutorial. Particularly useful if you have not used R (much) before.
- https://www.rstudio.com/resources/cheatsheets/ Cheat sheets for data wrangling, data visualization, general use of R, R Studio, R Markdown etc.
- http://rmarkdown.rstudio.com/ Lots of R Studio resources for R Markdown. Get started here.

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 and Reading Assignments

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.

Group work: Exchange contact details with your group.

Exercise: Go through the babynames exercise. Try the suggested extensions.

Week 2: (Jan 30) Continuing with ggplot2. Addtional chart types. Grammar of Graphics. Understanding additional layers – coordinates, facets.

On your own: Set up your own GitHub account. Familiarize yourself, if necessary with the suggested tutorial.

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

Exercise: Assignment 1: Plotting with ggplot

Part 2: Working with Spatial Data

Week 4: (Feb 13) Working with Spatial Data in standard data frames and ggplot using the ggmap package

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

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

Exercise: Assignment 2: Mapping NYC's Income Differences

Part 3: Text Mining and Visualization

Week 6: (Mar 6) Bag of Words. Cleaning and preprocessing, data structures. Frequency visualization, word clouds.

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

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

Exercise: Assignment 3: Visualizing Tweets.

Part 4: Getting Dynamic and Interactive

Week 8: (Mar 27) Making ggplot graphs dynamic and interactive

Week 9: (Apr 3) Network visualizations. R Shiny and Data Driven Documents (D3). Hosting your own interactive visualization.

Week 10: (Apr 10) Refining interactive plots, maps, and text analyses

Exercise: Assignment 4: Interactive app with your own data.

Week 11: (Apr 17) Other Topics – depending on interest

Week 12: (Apr 24) Wrap-up and conclusion, showcase of techniques, preview of further topics

Week 13: (May 1) Final Student Presentations

--- END OF SYLLABUS ---