Statistical Visualization II —

Stat 5810, Section 005 & Stat 6910, Section 004

Spring 2018

Instructor: Dr. Jürgen Symanzik
Office: AnSc 313
Phone: 435–797–0696
FAX: 435–797–1822
e–mail: symanzik@math.usu.edu
Web: `http://www.math.usu.edu/~symanzik/`
`http://www.math.usu.edu/~symanzik/teaching/2018_stat5810_005/stat5810_005.html`

Office Hours: MWF 4:30pm – 5:30pm and by appointment.

Classes & Rooms:
MWF 2:30am – 3:20pm, We 1/17 – Fr 4/20, 2018 (tentatively): AnSc 320.
Please visit the course Web page listed above and/or Canvas frequently for lecture notes, data sets, graphical examples, R code, etc. — in particular if you miss class for any reason.

Detailed Class Schedule:
For a 2–credit course, we need 29 lectures/lecture days (in contrast to 43 or 44 lectures/lecture days for a 3–credit course). Those days are marked as "Lecture 01" to "Lecture 29" in the overview below:

Week	Monday	Wednesday	Friday
1	1/8: No class	1/10: No class	1/12: No class
2	1/15: No class	1/17: Lecture 01	1/19: Lecture 02
3	1/22: Lecture 03	1/24: Lecture 04	1/26: Lecture 05
4	1/29: Lecture 06	1/31: Lecture 07	2/2: Lecture 08
5	2/5: Lecture 09	2/7: Lecture 10	2/9: No class
6	2/12: No class	2/14: No class	2/16: No class
7	2/19: No class	2/21: No class	2/23: No class
8	2/26: Lecture 11	2/28: Lecture 12	3/2: Lecture 13
9	3/5: Spring Break	3/7: Spring Break	3/9: Spring Break
10	3/12: Lecture 14	3/14: Lecture 15	3/16: Lecture 16
11	3/19: Lecture 17	3/21: Lecture 18	3/23: Lecture 19
12	3/26: Lecture 20	3/28: Lecture 21	3/30: Lecture 22
13	4/2: Lecture 23	4/4: Lecture 24	4/6: Lecture 25
14	4/9: Lecture 26	4/11: Lecture 27	4/13: Lecture 28
15	4/16: Lecture 29	4/18: Possibly	4/20: Possibly
16	4/23: Backup	4/25: Backup	4/27: Backup

Note: "No class" means guaranteed no class that day. I anticipate that I may have to reschedule 1 or 2 lectures, but this has not been finalized yet. If this is confirmed, we will use the dates marked as "Possibly". I have marked the last week as "Backup", e.g., in case we miss lectures because I am sick. But hopefully, this won't happen. If nothing goes wrong, our tentative last lecture date will be on Fr 4/20/2018.

Course Objectives:

Statistical graphics and data visualization are critical elements of modern data analysis and presentation. From initial exploration of a data set to the final presentation of results to the end user, statistical graphics play a vital role in shaping our understanding of our data. Through proper use of graphics, we can make critical discoveries, and communicate them clearly. Conversely, poor use or misuse of graphics can seriously mislead (by accident or design).

The course will address three main questions:

1. Why statistical graphics (and which ones to draw)?
2. How to construct statistical graphics in R?
3. How to distinguish between **good** and **bad** statistical graphics?

This course is **not** an introduction into a single R graphics package. Rather, a variety of R graphics packages will be used, such as baseR, `ggplot2`, `lattice`, etc.

Even more than most aspects of statistics, graphics and visualization involve art as well as science. In most cases, there are many reasonable approaches. Only an understanding of the options available and the underlying principles will lead to a successful analysis and presentation.

Prerequisites:

I expect basic knowledge of R as taught in the "Introduction to R" course. More importantly, I expect that you previously took my "Statistical Visualization I" course or have equivalent knowledge. Moreover, you should be familiar with a tool such as R Markdown, knitr, or sweave that allows you to combine text, R code, graphics, and numerical results in high-quality documents. LaTeX is a plus but is not formally required at the 5000 level, but it will be required at the 6000 level of this course.

Moreover, I expect basic "operational" knowledge from an introductory stats course such as Stat 2000, Stat 3000, or higher. "Operational" means that you still recall sufficient details from regression, ANOVA, hypothesis tests, etc. (it is not sufficient that you have taken such a course several years ago and have forgotten almost all details).

IDEA Center Learning Objectives:

Objective 1) Gaining factual knowledge (terminology, classifications, methods, trends).

Objective 2) Learning fundamental principles, generalizations, or theories.

Objective 3) Learning to apply course material (to improve thinking, problem solving, and decisions).

Objective 11) Learning to analyze and critically evaluate ideas, arguments, and points of view.

Topics: (subject to change)

This course will continue where "Statistical Visualization I" ended:

1. Statistical Maps.
2. Color and Cognition.
3. Graphs for Trivariate Data.
4. Graphs for "Hypervariate" (High–Dimensional) Data.

5. History of Graphics.
6. Interactive and Dynamic Graphics.
7. Web–Based Graphics.
8. Others (as time permits).

We will work with some data sets suitable for particular concepts introduced in class. However, our primary data sets for homeworks and projects will be data sets related to the "Data Expo 2018" from the ASA Statistical Computing and Statistical Graphics sections, accessible at `http://community.amstat.org/stat-computing/data-expo/data-expo-2018`. These data sets will contain surprises — for you and for me. Do not expect that someone is going to give you the final answer or model. We jointly will have to work towards such an answer or model.

For MS and PhD students majoring in Statistics, it is important to learn LaTeX — from basic document preparation, over the inclusion of R graphics into your LaTeX documents to advanced topics such as Sweave (`https://leisch.userweb.mwn.de/Sweave/`) and the LaTeX bibliography BibTeX (`http://www.bibtex.org/`). LaTeX is essential for graduate work (at the MS and PhD level) and will be used for many theses, dissertations, and scientific publications. Therefore, LaTeX will have to be used for all homeworks, projects, presentations, etc. at the 6000 level of this course.

Homework Assignments:

There will be a variety of assignments throughout the semester. Each assignment will include a value (typically 20–100 points) that it will be scored out of. Your final grade will be determined by the sum of your points in all assignments. Some assignments will include combinations of computer work in R (or others) and short oral presentations. The value of each assignment will be roughly proportional to its importance and the amount of work involved.

Regular homework assignments will be done individually or in groups of 2 or 3 students. For individual assignments, you will be allowed to discuss general approaches to questions on the assignments with other students, but each student must write and submit their own code and comments. Any students caught sharing code will fail the class.

The following deductions will be applied to late homework submissions: 1 min – 24 hours late: 10% off; > 24 hours – 48 hours late: 25% off; > 48 hours – 72 hours late: 50% off. Homeworks won't be accepted later than 72 hours (i.e., 3 days) after the submission deadline.

There will be no (in–class or take–home) quizzes, midterm exams, or final exams. Nevertheless, this will be a very challenging course that requires a lot of individual time to work on the assignments (and projects). Just attending classes will not be enough to pass this course! In addition, you will have to do a lot of individual reading of textbooks, online documentation, and help pages, and search for available information on the web.

Projects (Stat 6910 only):

There will be one or two projects during the semester. This could be the presentation of an R package, a summary of a journal paper (related to graphics), an extended open–ended analysis of a data set with a focus on graphics, etc. Projects will require the preparation of a final project report and possibly a short presentation of your work for the other students in this course. The projects will account for about 30% of your final grade.

Textbooks:

Carr, Daniel B., and Pickle, Linda W. (2010) *Visualizing Data Patterns with Micromaps*, Boca Raton, Florida: Chapman & Hall/CRC Press, `https://www.crcpress.com/Visualizing-Data-Patterns-with-Micromaps/Carr-Pickle/p/book/9781420075731` & `http://mason.gmu.edu/~dcarr/Micromaps/`.

Tufte, Edward R. (1983) *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press.

Unwin, Antony (2015) *Graphical Data Analysis with R*, Boca Raton, FL: CRC Press/Taylor & Francis.

Wickham, Hadley (2009) *ggplot2 — Elegant Graphics for Data Analysis*, New York, NY: Springer, `http://www.springer.com/us/book/9783319242750` & `http://ggplot2.org/book/`.

Every student should have access to each of these books, but it is not necessary that every student buys all of these books. Perhaps you can make arrangements with some of the other students in class who purchases which book(s). If you plan to work in the area of statistical visualization for your MS or PhD degree, you should consider to purchase these books for an ongoing use beyond this course.

Software:

We will primarily be using R (`http://cran.r-project.org/`), a free software environment for statistical computing and graphics. Please install the current version of R, i.e., 3.4.3, on your own computer so we can exchange code. Also install RStudio (`https://www.rstudio.com/`) as a front end to R.

Courtesy:

Please turn off cell phones and similar devices before class, and please keep conversations to a minimum during lectures. Please do not read/reply to your e–mails or browse other web pages than the ones discussed during class.

I will not keep track if you come to class or not. However, I would highly recommend to attend all lectures. In case you miss a lecture, please see the recording in Canvas/Panopto. I plan to record all lectures, but technology (web access, wireless connections, etc.) may fail once in a while so I can't give a 100% guarantee that all lectures will be available as recordings.

Americans with Disabilities Act:

If a student has a disability that will likely require some accommodation by the instructor, the student must contact the instructor and document the disability through the Disability Resource Center (DRC), during the first week of the course. Any requests for special considerations relating to attendance, pedagogy, taking of examination, etc. must be discussed with and approved by the instructor. In cooperation with the Disability Resource Center, course materials can be provided in alternative formats — large print, audio, or Braille.

Note:

The above schedule and procedures in this course are subject to change in the event of extenuating circumstances.