Statistical Visualization II —

Stat 5560, Section 001 & Stat 6560, Section 001

Spring 2024 (2 Credits)

Instructor: Dr. Jürgen Symanzik
Office: AnSc 313
Phone: 435–797–0696
e–mail: symanzik@math.usu.edu
Web: `http://www.math.usu.edu/~symanzik/`
`http://www.math.usu.edu/~symanzik/teaching/2024_stat5560/stat5560.html`

Office Hours: Tuesday (T) & Thursday (H) 10:00am – 11:00am in AnSc 313,
Friday (F) 9:00am – 10:00am (via Zoom), and by appointment.

Teaching Assistant & Grader / Office Hours: Kelvyn Bladen (kelvyn.bladen@usu.edu) /
Wednesday (W) 9:00am – 10:20am in AnSc 101 & Thursday (H) 9:00am – 10:00am in AnSc 101.

Classes & Rooms:
TH 12:00noon – 1:15pm, T 1/16 – H 3/28, 2024 (tentatively): AnSc 320 (face–to–face).

Please visit the course Web page listed above for emergency announcements, e.g., when Canvas is unavailable. Otherwise, visit Canvas frequently for lecture notes, data sets, R code, etc. — in particular if you miss our face–to–face lectures for any reason. All (additional and updated) materials, announcements, discussions, recordings, etc. from Canvas are part of the course materials. Not seeing one of these in time does not serve as an excuse for not getting point deductions for the course. Deadlines may change or unexpected new *Coronavirus/Covid–19* regulations and requirements may occur. It is your responsibility to make sure to receive all announcements in time.

Detailed Class Schedule:
For a 2–credit course, we need 20 lectures/lecture days (in contrast to 29 or 30 lectures/lecture days for a 3–credit course). Those days are marked as "Lecture 01" to "Lecture 20" in the overview below:

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

Note: "No class" means guaranteed no class that day. I have marked the dates in the four last weeks of the semester as "Backup", e.g., in case we miss lectures because of weather conditions, sickness, or any other reason. If nothing goes wrong, our tentative last lecture date will be on H 3/28/2024. Backup lectures (if any) will be held via Zoom (and not face–to–face). There is no need to stay in Logan after H 3/28/2024 for a backup lecture in this course.

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:

STAT 5050 ("Introduction to R") with a C− or better. STAT 5550 ("Statistical Visualization I") with a C− or better. 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 that you still have "operational" knowledge from your STAT 2000, STAT 3000, or STAT 5100 (or higher) course. "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 a basic understanding of the subject (e.g., factual knowledge, methods, principles, generalizations, theories).

Objective 4) Developing specific skills, competencies, and points of view needed by professionals in the field most closely related to this course.

Objective 13) Learning appropriate methods for collecting, analyzing, and interpreting numerical information.

Topics: (subject to change)

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

1. Graphs for Bivariate Data.
2. Graphs for Trivariate Data.
3. Graphs for "Hypervariate" (High–Dimensional) Data.
4. Color and Cognition.
5. Statistical Maps.
6. Interactive and Dynamic Graphics.
7. Web–Based Graphics.
8. History of Graphics.
9. Others (as time permits).

We will work with some data sets suitable for particular concepts introduced in class. 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://stat.ethz.ch/R-manual/R-devel/library/utils/doc/Sweave.pdf`), knitr (`https://yihui.org/knitr/`), 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 3 HW assignments for this course, roughly one every three weeks. Each HW assignment will include a value (typically 20–100 points) that it will be scored out of. HW assignments will contribute to 100% (for Stat 5560), repectivley 70% (for Stat 6560), of your course grade. The value of each HW assignment will be roughly proportional to its importance and the amount of work involved.

You will be allowed to discuss general approaches to questions on the HW assignments with other students, but each student must write and submit their own R code and comments. Any students caught sharing R code or other parts of their homework submissions will fail the class.

Unless otherwise stated on the HW assignment sheet, all homework assignments have to be submitted electronically via Canvas. **You will have 2 or 3 weeks after the last lecture to finalize and submit the last HW assignment.**

There will be no (in–class or take–home) quizzes, midterm exams, or final exams. We will have a few worksheets for training purposes only. Nevertheless, this will be a very challenging course that requires a lot of individual time to work on the assignments. 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.**

No Excuse Needed Late Homework Submission Policy:

Each student has **3 tokens** for late homework submissions. These tokens can be used in multiple ways, e.g., one token at a time for each of the 3 HWs or all 3 tokens can be used for a single late HW submission. You will need 1 token if your HW is 1 min – 24 hours late; you will need 2 tokens if your HW is $>$ 24 hours – 48 hours late; and you will need all 3 tokens if your HW is $>$ 48 hours – 72 hours late. Once you have used up all 3 tokens, late HWs will count as 0 points, even if such a HW is only one additional minute late. As an example, if you used 2 tokens early on, you will have only 1 token left and can submit your next HW only up to 24 hours late. If you submit that HW more then 24 hours late, it will count as 0 points. All times are based on the Canvas time stamp that is created when a HW gets submitted — and not on your local computer clock.

It is your responsibility to keep track of the number of tokens you have used. Record that information on your phone or check how you can extract the submission dates and times in Canvas on your side. I will not be able to tell you shortly before the next submission deadline how many tokens you have left. In fact, I only plan to use that information when I determine final course grades. You may therefore get assigned a score $>$ 0 points for a certain HW submission initially, but that score may get adjusted to 0 points if it turns out that this was for a late submission and you had already used up all your tokens.

You will not get any credit for any unused tokens, so you could use the remaining ones for proofreading, checking that all figures and R code are included, etc., in particular for the later HWs — and if you still have tokens left at that time!

As the name indicates, there is no excuse needed when you use a token, be it for a minor sickness, personal travel due to family events, or just needing more time as you couldn't finish that HW by the deadline. If you are officially absent from USU when a HW is due, e.g., due to travel with a USU sports team or band, attending a conference, doing fieldwork, etc., I need an official statement from a coach or supervisor — and no token will be used. However, we then need to determine on a case–by–case basis when that particular HW will be due for you — and when you would have to use additional token(s) to further extend your extended deadline.

You can resubmit your HW as often as you want in Canvas prior to the deadline. Apparently, if you resubmit it after the deadline, one (or more) tokens will be used. In that case, also let me know via e–mail at the time of the deadline that you are still working on

your HW beyond the original deadline and plan to resubmit a new version later on. Once a HW has been graded on my side, you will no longer be allowed to resubmit a revision of that HW (or in case you did, you still would get the points based on the last submission prior to the original deadline).

Project (Stat 6560 only):

There will be one major project towards the end of the semester. This will include the preparation of a final project report and a short presentation of your work for the other students in this course. The project will be done in a small group of students. The project will account for 30% of your course 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`, `https://ggplot2.tidyverse.org/`, and `https://ggplot2.tidyverse.org/reference/`.

Every student should have access to each of these books, but it is not necessary that every student buys all of these books. The USU library holds several of these books or provides online access. 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 a recent version of R, i.e., 4.3.2 or later, on your own computer so we can exchange code. Working with version 4.3.0 or 4.3.1 should still be OK. I would discourage the use of version 4.2.3 or older. Also install RStudio (`https://www.rstudio.com/`) as a front end to R and a version of LaTeX that can be used in connection with RStudio. My personal recommendation is TeX Live (`https://www.tug.org/texlive/`). For Windows, also install Rtools4.3 (`https://cran.r-project.org/bin/windows/Rtools/`).

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. If you have to miss a lecture, there will be a Kaltura recording of the lecture available in Canvas later in the day (if the technology doesn't fail).

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 – `https://www.usu.edu/drc/`), preferably 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.