College of William & Mary
Government 391 / Data Science 340
Fall 2023

Data Visualization for the Social Sciences

Prof. Maurits van der Veen

Class
TuTh 12:30-13:50
Location: Washington Hall 301
Office hours: Tuesdays 2-4pm, or by appointment

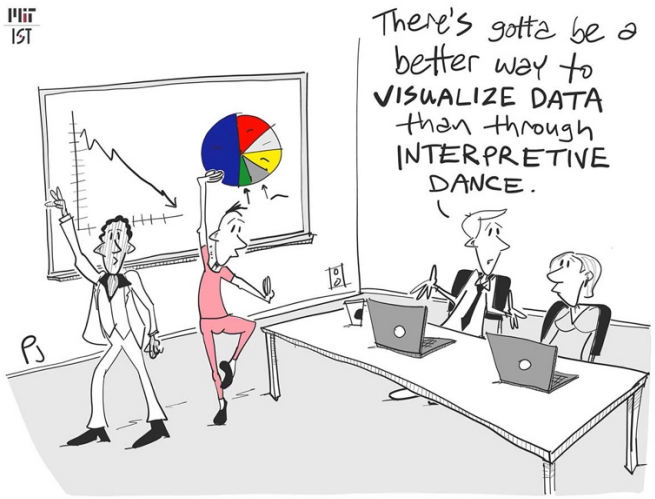
Office
355 Chancellor Hall
maurits@wm.edu

Course description

Visualizations are a key form of communication. Data visualizations let data "speak" to an audience. The story they tell and the message they convey may or may not be interesting, important, and successful; the design of the visualizations themselves plays an essential role in their impact.

In this course, we'll learn basic design principles, the building blocks of successful charts, as well as the coding necessary to produce such charts. By the end of the course you will have become both a critical consumer of data visualization and a producer of attractive, informative, and honest (as in not misleading) charts, in a wide range of forms.

Producing good visualizations can be immensely satisfying, but also quite frustrating. We'll work together to keep the amount of frustration to a minimum. Mastering the course material should be rewarding and even fun!



To become a confident producer of successful visualizations, you need to master a programming vocabulary. Two of the most common languages for data visualization are R and python. R is more commonly used by social scientists; python by those with a computational background. This course is agnostic about which you prefer. All in-class examples, assignments, etc. can be done equally well in R and python. You can do all your work in one language or go back and forth between languages — the choice is yours.

We begin the semester by becoming familiar with the basic grammar of visualizations as represented in the visualization package ggplot (R), or its python version, plotnine. These packages are both elegant and powerful, and are used by many professional data scientists as their primary visualization tools. Accordingly, both packages also offer a wide range of fine tuning options, which we will cover towards the latter part of the semester.

No one package can do everything. In particular, certain specialized visualizations, such as word clouds, rely on dedicated software tools, and we cover some of those as well. Finally, charts produced in ggplot can be used as the basis for dynamic charts as well as for dashboard-style interactive visualizations. We will take a closer look at each of these possibilities towards the end of the semester.

Reading material

There is one required book for the course:

- Healy, Kieran. 2018. *Data visualization*. Princeton, NJ: Princeton University Press. (https://socviz.co/)

While the book is available online at the link above, I do recommend you purchase it. It will be valuable to have alongside as you're working on your laptop, as well as to use as a reference.

All other readings are, or will be, available online: on Blackboard, at a given URL, or through the Electronic Journals feature of the W&M library website. This includes two additional books from which we use several chapters. (Each is excellent, and well worth purchasing.)

- Dougherty, Jack, and Ilya Ilyankou. 2022. *Hands-on data visualization* O'Reilly publishing. (https://handsondataviz.org/)
- Wilke, Claus. 2019. *Fundamentals of Data visualization.* O'Reilly Publishing. (https://clauswilke.com/dataviz/)

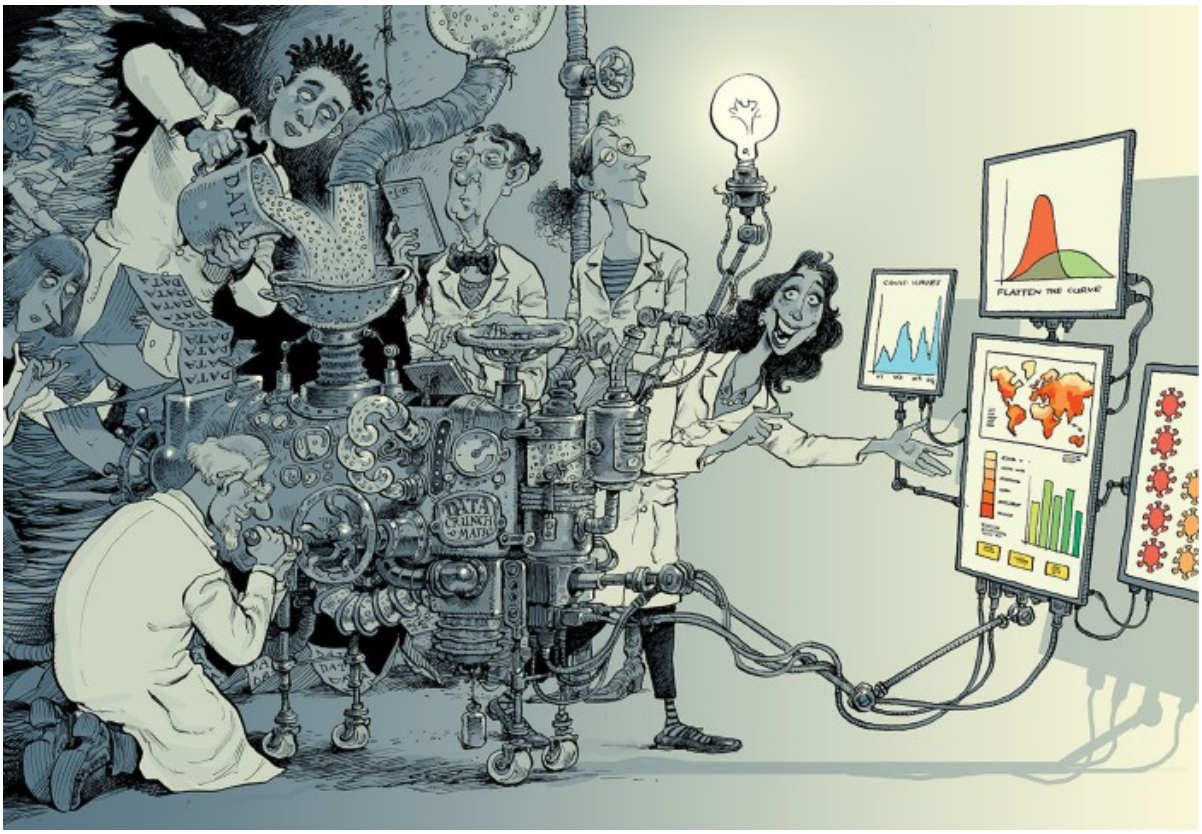


Illustration in Nature by David Parkins (https://www.nature.com/articles/d41586-022-00792-2)

Evaluation

You are responsible for your own learning. My role is to engage and assist you in learning more about data visualization. You will get the most out of the class if you complete all the readings on the assigned dates, and regularly practice what you learn.

Your grade will be composed of:

- Participation — 25%
- Programming assignments — 40% (5% each)
- Final project — 35% (10% presentation; 25% final product)

Grade scale

A	94-100	C+	77-79	D-	60-63
A-	90-93	C	73-76	F	0-59
B+	87-89	C-	70-72		
B	83-86	D+	67-69		
B-	80-82	D	63-66		

Programming assignments
The programming assignments are intended to give you hands-on experience with the principles and particulars of data visualization. Most weeks have such an assignment. Usually I will introduce that week's topic as well as demonstrate the goal product (or something like it) on Tuesday, and make the assignment instructions available the same day. On Thursday we'll discuss more advanced work using the topic/tools of the week; you'll get much more out of class on Thursday if you have begun that week's programming assignment already.

Final project
You will design your own final project and present your project to the class. Projects can be completed individually or in groups of two. They require acquiring your own dataset, exploring it, and producing a thoughtful, informative story about it, using at least three graphs/charts.

Course Policies

Attendance and class participation
Success in this course is predicated on regular class attendance. When you miss class, you miss an opportunity to learn something new and to gain a deeper understanding of the course material, as well as the chance to ask questions, learn from your peers, and show me where I may need to slow down, speed up, or retrace particular steps. Moreover, your risk missing out on important details, changes to the syllabus, and discussion of assignments.

If you are unable to attend class, whether due to COVID exposure or any other reason, please let me know as soon as possible, and we'll find the best way for you to catch up. Note: if you miss a class, do not ask me if you missed anything, because the answer will always be 'yes'.

Readings
Many of the readings include code to try out. We will generally go through the readings and the code together during class. You will get the most out of the material if you glance over it prior to class, follow along actively during class, and make sure to go back over anything that didn't completely make sense afterwards.

Assignments
Assignments are to be submitted via Blackboard. You should complete them individually, but discussion with your classmates is allowed. If you discussed the work with another student, please mention that at the top of the assignment. Assignments are due at the start of class on the due date. Assignments that are late will lose 0.5 points (out of 5) for each day they are late. If your homework assignment is more than a week late without explanation, you will receive a 0.

Programming environment
The recommended environment for R is Rstudio; for python it is Jupyter. If you have no experience with either of these and are unsure where or how to start, let me know, and we'll make sure you get a working set-up for your preferred language.

Class communication
I have set up a class Slack workspace, DataViz, where you can post questions and receive answers from me or from your fellow students. I will send an invitation link after the first class. To make sure everyone has successfully added themselves to the Slack workspace, I would like you to post a message in the #course channel with your current idea of what you think you might want to do for your final project. This can be very broad and I will not hold you to it, but it will help me check that everyone is on Slack, while at the same time giving me some information about your interests (which will help me tailor my class presentations).

Office hours (student hours) and email
In addition to regular office hours, I am available for both Zoom and live meetings by appointment. Please do get in touch if you have any questions regarding the reading material, the assignments, or the course in general. You are also welcome to email me if you have minor questions, but better still is to post a question on our class Slack, where others can see it too.

Accommodations
Any student who needs accommodation is requested to consult with Student Accessibility Services (SAS) as early as possible. I will follow the recommendations from SAS, and all discussions, issues, concerns, and accommodations will remain confidential.

Academic Integrity and the Honor Code
I adhere to the College's official policies regarding academic honesty. More information about the College's Honor Code can be found here. You are expected to be familiar with the College's policies on academic honesty. There will be **zero tolerance** for any cases of plagiarism, fabrication, cheating and facilitation of other forms of academic dishonesty.

The course calendar starts on the next page. It is a general plan for the course; deviations (always announced to the class in advance by the instructor) may be necessary.

Calendar

I. Introduction

Week 1: Introduction to the course

Thu. Aug. 31 ***Introduction to the course***

- Brief overview of the course
- Setting up: Rstudio (R), Jupyter (Python), Slack

Recommended

- Podcast: "Florence Nightingale: Data Viz Pioneer." (*99% invisible podcast*, presented by Tim Harford and including readings by Nightingale's distant relative Helena Bonham Carter: https://99percentinvisible.org/episode/florence-nightingale-data-viz-pioneer/)
- Andrews, R.J. 2022. "How Florence Nightingale changed data visualization forever." *Scientific American.* (https://www.scientificamerican.com/article/how-florence-nightingale-changed-data-visualization-forever/)

Week 2: Design principles

Tue. Sep. 5 ***Principles of data visualization***

- Healy, chapter 1: "Look at data"
- Discussion of Edward Tufte's *The visual display of quantitative information* by Guy Pursey (https://guypursey.com/blog/202001041530-tufte-principles-visual-display-quantitative-information)

Recommended

- Tufte, Edward. 1982. *The visual display of quantitative information.*
- Mansky, Jackie. 2018. "W.E.B. DuBois' visionary infographics come together for the first time in color." *Smithsonian Magazine* (https://www.smithsonianmag.com/history/first-time-together-and-color-book-displays-web-du-bois-visionary-infographics-180970826/).

Thu. Sep. 7 ***Pitfalls***

- Ingraham, Stephen. 2018. "Study: Charts change hearts and minds better than words do." *Washington Post*, 15 June (https://www.washingtonpost.com/news/wonk/wp/2018/06/15/study-charts-change-hearts-and-minds-better-than-words-do/)

- Few, Stephen. 2011. "The chartjunk debate. (https://perceptualedge.com/articles/visual_business_intelligence/the_chartjunk_debate.pdf)

Recommended

- "WTF Visualizations" (https://viz.wtf/)

Notes / assignments:

- Become comfortable with your chosen environment (R or Python)
- Think about DuBois' data visualizations: Which do you find most evocative, informative, striking, etc.? How important is text in the visualizations?
- Read Healy, chapter 2: "Get started"

II. Nuts & bolts of visualization

Week 3: The components of a chart

Tuesday, September 12: **Getting started**

- Healy, chapter 3: "Make a plot"
- Robbins, Naomi. 2012. "Are grid lines useful or chartjunk?" (https://www.forbes.com/sites/naomirobbins/2012/02/22/are-grid-lines-useful-or-chartjunk/?sh=57a163e04283)

Thursday, September 14: **Showing the right data**

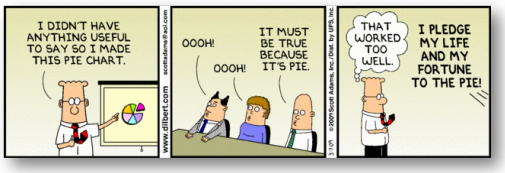
- Healy, chapter 4: "Show the right numbers"
- Cisneros, Mike. "Colors and emotions in data visualization" (https://www.storytellingwithdata.com/blog/2021/6/8/colors-and-emotions-in-data-visualization)

Recommended

- Wilke, chapters 4 ("Colour scales") & 5 ("Directory of visualizations)

Notes / assignments:

- Do not wait until the day before it is due to start an assignment!
- Assignment 1: Reproduce a DuBois chart (due Sep. 19)



Week 4: Doing charts right

Tuesday, September 19: **Amounts and proportions**

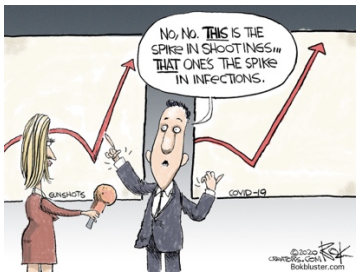
- Wilke, chapters 6 ("Visualizing amounts") & 10 ("Visualizing proportions")
- Kosara, Robert. 2016. "A reanalyisis of a study about (square) pie charts from 2009." (https://eagereyes.org/blog/2016/a-reanalysis-of-a-study-about-square-pie-charts-from-2009)
- Kosara, Robert. 2008. "Engaging readers with square pie / waffle charts." (https://eagereyes.org/blog/2008/engaging-readers-with-square-pie-waffle-charts)

Thursday, September 21: **Distributions**

- Healy, chapter 5: "Graph tables, add labels, make notes"
- Wilke, chapter 9: "Visualizing many distributions at once"

Notes / assignments:

- Assignment 2: Trends in the distribution of freedom in the world (due Sep. 26)



Week 5: Visualizing uncertainty

Tuesday, September 26: **Visualizing statistical analyses**

- Healy, chapter 6: "Work with models"

Thursday, September 28: **Conveying uncertainty**

- Wilke, chapter 16: "Visualizing uncertainty"

Recommended

- Correll, Michael, and Michael Gleicher. 2014. "Error bars considered harmful: Exploring alternate encodings for mean and error."
- Kosara, Robert. 2017. "Communicating uncertainty when lives are on the line." (https://eagereyes.org/blog/2017/communicating-uncertainty-when-lives-are-on-the-line)

Notes / assignments:

- Assignment 3: Produce 2 different charts (with indications of uncertainty) based on a statistical analysis of the Varieties of Democracy (V-Dem) data (due Oct. 3)

Week 6: Visualizing time

Tuesday, October 3: **Time within a figure** (dumbbell charts, slope charts, etc.)

- Wilke, chapter 14: "Visualizing trends"

Thursday, October 5: **Time across figures** (facets, dynamic charts)

- Fisher, Danyel. 2010. "Animation for visualization: Opportunities and drawbacks." Chapter 19 in Julie Steele and Noah Iliinsky, eds. *Beautiful visualization.* O'Reilly Media, Inc.

Notes / assignments:

- Assignment 4: Dynamic chart of freedom over time (due Oct. 10)
- Start planning your final project: general topic, data sources, etc.
  (Highly recommended: Dougherty & Ilyankou, ch. 3: "Find and question your data")

Week 7: Interlude: How to lie with charts

Tuesday, October 10: **Deceptive charts**

- Dougherty & Ilyankou, chapter 14: "Detect lies and reduce bias"

Recommended

- Lin, Chujun, and Mark Thornton. 2021. "Fooled by beautiful data" (https://psyarxiv.com/dnr9s/)

Thursday, October 12: *Fall Break (no class)*

Notes / assignments:

- Assignment 5: Use data on school diversity in the US to produce 3 charts: one that makes things look better than they are, one that makes things look worse than they are, and one that does neither (due Oct 17)

Week 8: Visualizing geography: Maps

Tuesday, October 17: **Data reflected in geographic units**

- Healy, chapter 7: "Draw maps"

- Yau, Nathan. 2016. "Firearms dealers vs. burgers, pizza, and coffee." (https://flowingdata.com/2016/06/14/firearms-dealers-vs-burgers-pizza-and-coffee/)

Thursday, October 19: **Data overlaid on maps**

- Kahle, David, and Hadley Wickham. 2013. "ggmap: Spatial visualization with ggplot2." *R Journal* 5(1): 144-161.
- Aisch, Gregor, and Josh Keller. 2015. "How gun traffickers get around state gun laws." *New York Times* (https://www.nytimes.com/interactive/2015/11/12/us/gun-traffickers-smuggling-state-gun-laws.html)

Notes / assignments:

- Assignment 6: Map global mortality data (due Oct. 24)

Week 9: Visual summaries (dimensionality reduction)

Tuesday, October 24: **Principal component analysis and alternatives (t-sne, umap)**

- Conlen, Matthew, and Fred Hohman. 2018. "The beginner's guide to dimensionality reduction." (https://dimensionality-reduction-293e465c2a3443e8941b016d.vercel.app/)

Recommended

- Chari, Tara, and Lior Pachter. 2022. "The specious art of single-cell genomics." *BioRxiv,* https://doi.org/10.1101/2021.08.25.457696

Thursday, October 26: **Topic modeling**

- Kaufman, Micki. "Quantifying Kissinger." (https://blog.quantifyingkissinger.com/) (Look in particular at https://blog.quantifyingkissinger.com/category/visualizations/force-directed-graphs/ and the two interactive graphs linked at the bottom of that page.)

Notes / assignments:

- Identify final project topic, data, and proposed visualizations (due Oct. 31)

Week 10: Finishing touches & customatization

Tuesday, October 31: **Refining charts & arranging plots**

- Healy, chapter 8: "Refine your plots"
- Wang, Henry. "ggplot2 theme elements demonstration." (https://henrywang.nl/ggplot2-theme-elements-demonstration/)
- Ggplot2 book, chapter 9: Arranging plots (https://ggplot2-book.org/arranging-plots.html)
- Auguié, Baptiste. 2019. "Laying out multiple plots on a page." (https://cran.r-project.org/web/packages/egg/vignettes/Ecosystem.html)

Recommended

- Yau, Nathan. 2022. "When there were more deaths than births in the U.S.." (https://flowingdata.com/2022/02/03/when-there-were-more-deaths-than-births-in-the-u-s/)

Thursday, November 2: **Selections & tooltips**

- Agrawal, Pulkit. 2021. "Why tooltips are terrible and how to better design them." (https://www.chameleon.io/blog/why-tooltips-are-terrible-and-why-you-should-use-them)

Notes / assignments:

- Assignment 7: Refine & add tooltips to a previous assignment (due Nov 7)

Week 11: Dashboards

Tuesday, November 7: *Election day. No class*

- If you haven't voted early, vote today! (if you're eligible to vote!)
- Work on the final assignment

Thursday, November 9: **Dashboards — an overview**

- What is a dashboard? A complete overview (https://www.tableau.com/learn/articles/dashboards/what-is)
- (video) Kruchten, Nicolas. 2022. "Why *interactive* data visualization matters for data science in Python." (https://www.youtube.com/watch?v=tlcMlOVbEpw)
- (video) "Python Dashboarding Shootout and Showdown" (*https://www.youtube.com/watch?v=4a-Db1zhTEw*)

Recommended

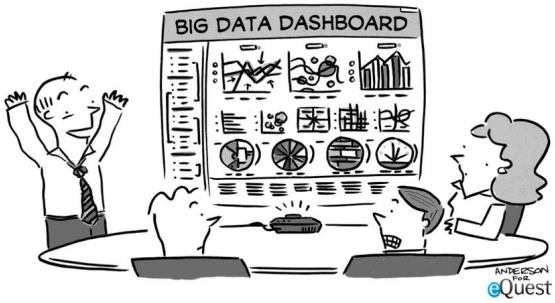
- Sievert, Carson. Chapter 2: "Overview" in *Interactive, web-based data visualization with R, plotly, and shiny*. (https://plotly-r.com/overview.html)
- Play around with sample dashboards using (for R) flexdashboard (https://rstudio.github.io/flexdashboard/articles/examples.html) and (for both R and python) Dash (https://github.com/plotly/dash-sample-apps/)

Tuesday, November 14: **Dashboards — implementing**

- (video) "Tableau how to create dashboard" (https://www.youtube.com/watch?v=Nr31rv9tsJ8)

Notes / assignments:

- Assignment 8: Make a simple dashboard (using Tableau, Dash, or flesdashboard), expanding on a previous assignment (due Nov 21)



"After careful consideration of all 437 charts, graphs, and metrics, I've decided to throw up my hands, hit the liquor store, and get snockered. Who's with me?!"

IV. Highlights

Thursday, November 16: **Favorite charts of W&M faculty** Presentation sign-up

Tuesday, November 21: **Favorite visualizations of/by non-W&M people**

(Thanksgiving week, no in-person class; lecture recorded, readings TBD)

Thursday, November 22: *Thanksgiving Break (no class)*

V. Presentations

Tuesday, November 28: Student presentations
Thursday, November 30: Student presentations

Tuesday, December 5: Student presentations
Thursday, December 7: Student presentations

Final project is due by midnight on Dec. 19 (in lieu of the final exam scheduled for that date)