Course overview

Data visualization (DSC 302) Fall 2024

Marcus Birkenkrahe

October 21, 2024

Contents

Figure 1: The Red Tower/La tour rouge (Giorgio de Chirico, 1913) Source: Guggenheim

1 Pre-test

I posted this infographic in the Google Chat Space claiming that is was "not a great visualization"? Do you see anything wrong with it?

Figure 2: Source Splashsys.com

- **Distraction:** Colors are unmotivated (except to resemble the colors of the Google logo), but humans tend to look for meaning in the color pattern
- **Lingo:** Seems to be aimed at newbies and not digitally literate people ("bros" not "pros") but uses lingo like "operator".
- **Understanding:** The information can better be presented as a table so that descriptions that belong together (operator | example | explanation) can more easily be compared (and be remembered).
- **Ordering**: There is no ordering principle, and no attempt to order them.
- **Extension:** There is no reference or information on how to find out more even though it seems unlikely (but we're not told) that these are the only search tricks available (I found at least 40).

2 Mutual introductions

1. Why are you here? *(What is it about visualization)*
2. What would delight you? *(In this course)*
3. What would disappoint you? *(In this course)*
4. Where are you headed? *(In life)*

3 Course syllabus (on GitHub and on Canvas)

- General information & standard policies
- Course information (grading, attendance)
- Schedule with dates of tests and assignments
- The GitHub repo contains course material: github.com/birkenkrahe/dviz
- Notes from each session.

Figure 3: Marc Chagall, Over the town (2018) Source: Wikiart

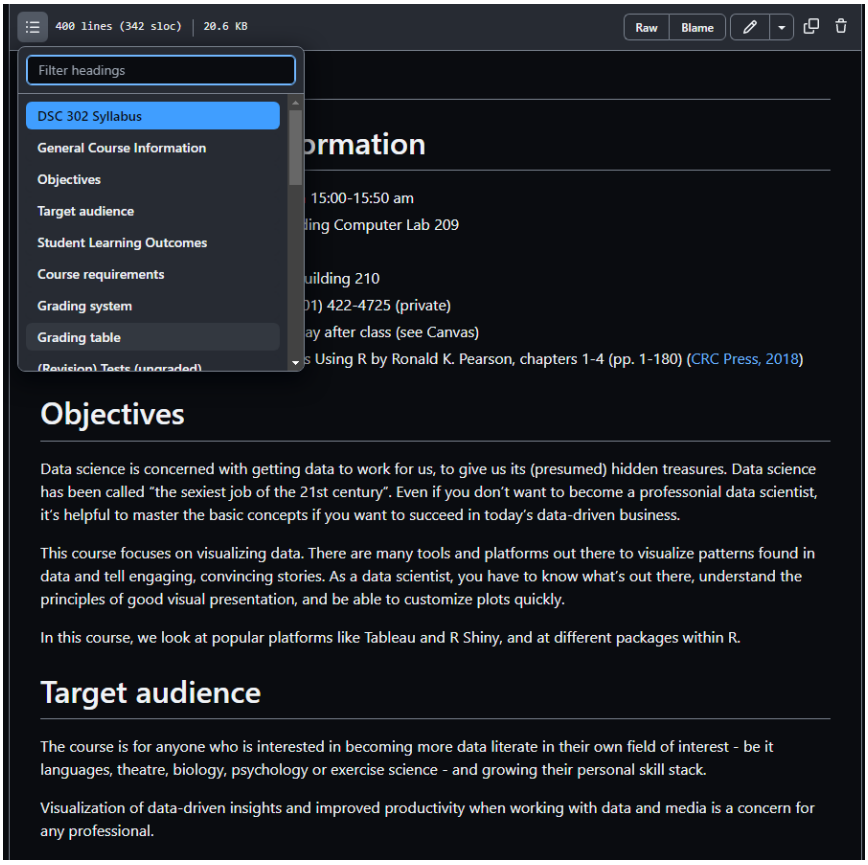


Figure 4: DSC 302 Syllabus on GitHub

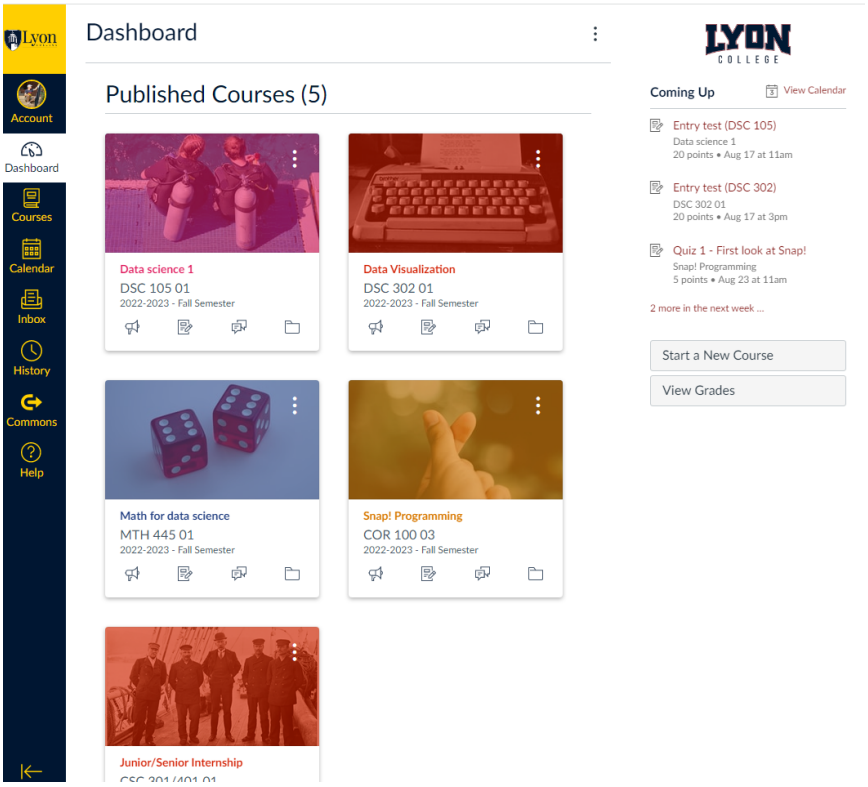


Figure 5: Course topics

4 Course "learning" platform: Canvas

- All grades should be visible at all times
- Control your own notifications
- Course links on a (wiki) page
- New CMS for me & for Lyon: bear with us[1]

5 Course topics (illustrated)

"An illustration of several plots of the same data with curves fitted to the points, paired with conclusions that you might draw about the person who made them. These data, when plotted on an X/Y graph, appear to have a general upward trend, but the data is far too noisy, with too few data points, to clearly suggest any specific growth pattern. In such a case, many different mathematical and statistical models could be presented as roughly fitting the data, but none of them fits well enough to compellingly represent the data." Source: explainxkcd.com 09/2018.

6 Course topics (spelled out)

1. Exploratory Data Analysis (EDA) using R (Python)
2. Graphics in base R with applications
3. Working with external data (critically)

7 Why "data visualization"?

- The purpose of data science is *pattern identification*
- *Visualization* happens in the head of the researcher first
- *Graphing* happens throughout, *storytelling* happens last

[1] CMS = Content Management System; these are the most common systems in business applications - present whenever people create 'content' of any sort (documents e.g.) and need to store it for later. CMS systems rely on database technology. In the case of Canvas, that's MySQL.

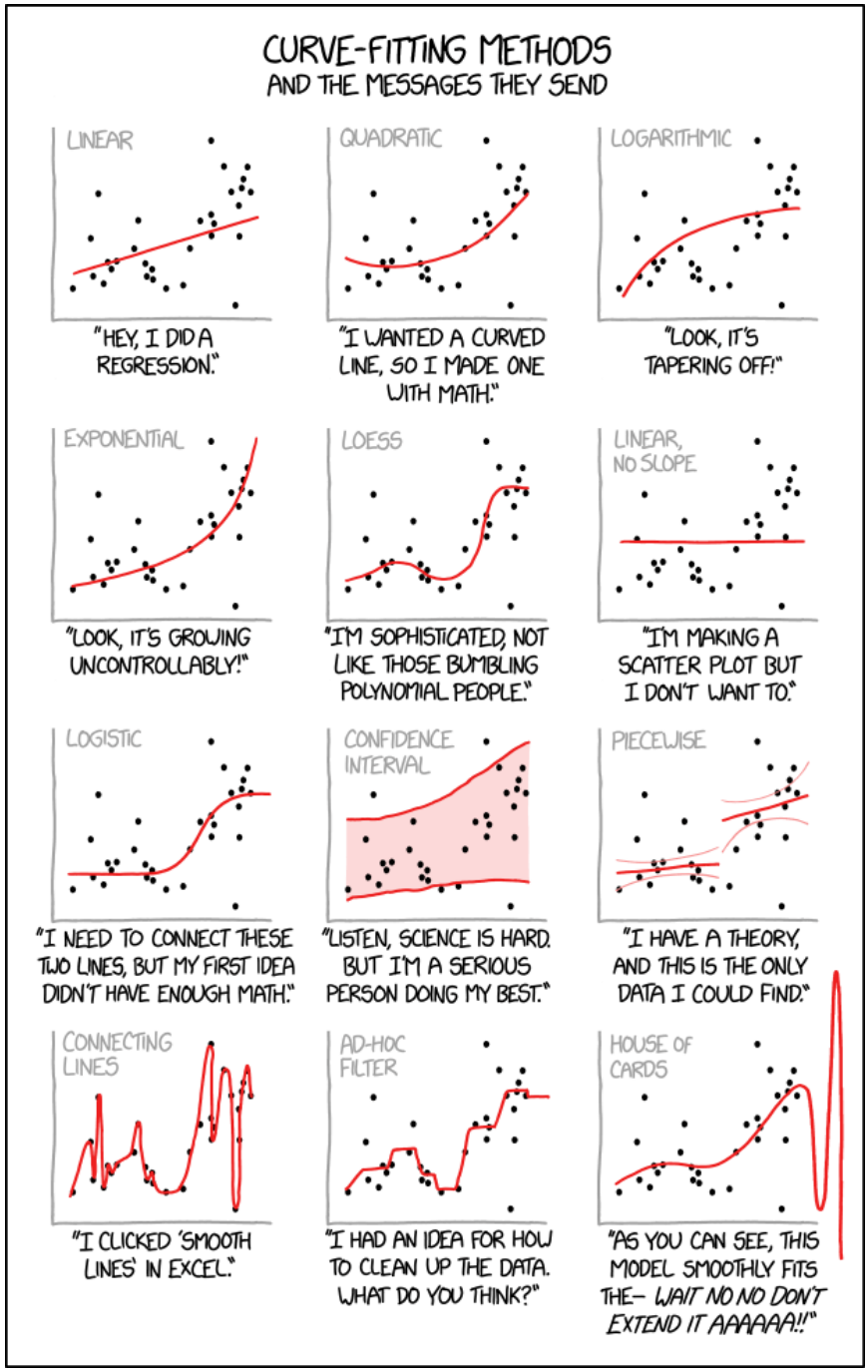
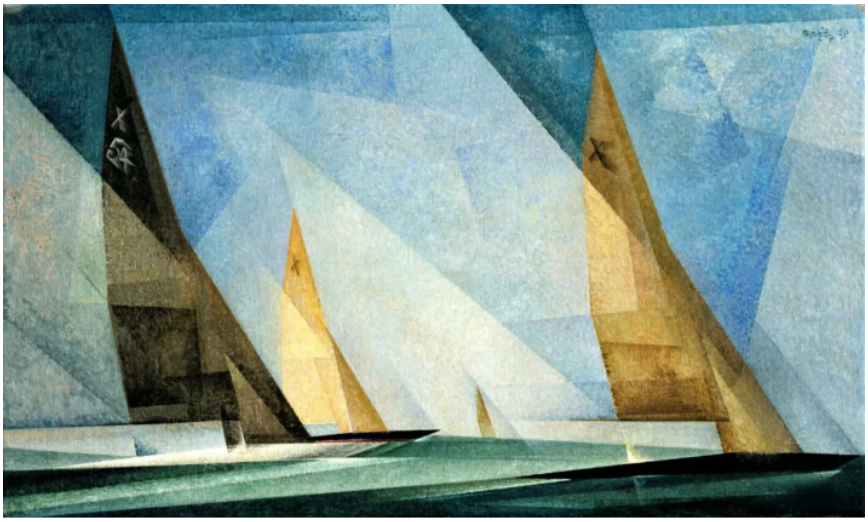


Figure 6: xkcd 2048: curve-fitting methods and the messages they send



Lyonel Feininger, *Sailboats*, 1929, Detroit Institute of Arts, Detroit, MI, USA.

Figure 7: Lyonel Feininger, Sailing Boats (1929)

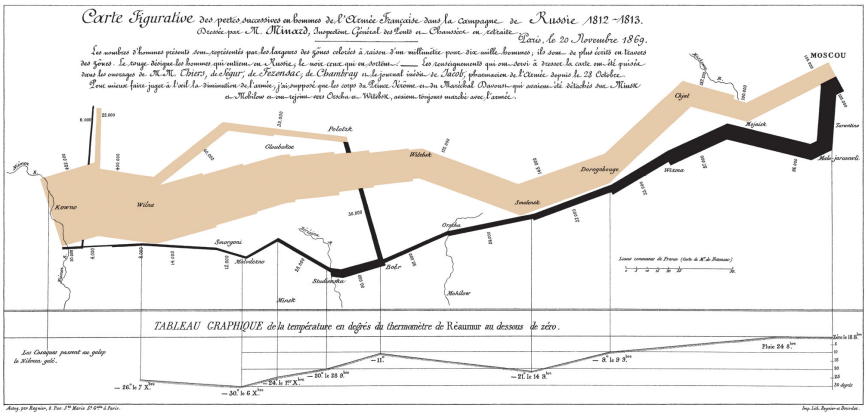


Figure 8: Charles Minard, Napoleon's Russian campaign 1812

- The diagram by Charles Minard (1869) tells the story of Napoleon's disastrous Russian campaign in 1812 (datavizblog.com, 2013)
- Variables: army location, temperature, size over time
- Diagram type: Sankey flow diagram (many examples)
- Data type: time series (an object class, `ts`, in R)
- The story of this campaign is also the backstory for Tolstoy's novel "WAR AND PEACE" (, 1867)

8 Get the story behind the stats

Even *The Fayetteville Observer* is trying to catch readers with data visualization / data story offers:

9 Agile [team] project

The team project makes up 20% of your final grade for this course.

See the GitHub FAQ for answers to these questions:

- What is a team project?
- Do you have examples for data science projects?
- Can you do a project as an absolute beginner?

  Note: the first *sprint review* is on August 31. Use it to present your initial results (see FAQ on what to deliver, and 1st sprint review).

10 IMRaD and Scrum

- Introduction (research question - what you want to find out)
- Method (how you want to do it)
- Results (what you found out)
- Discussion (what it means)

(Video: Research Writing with IMRaD)

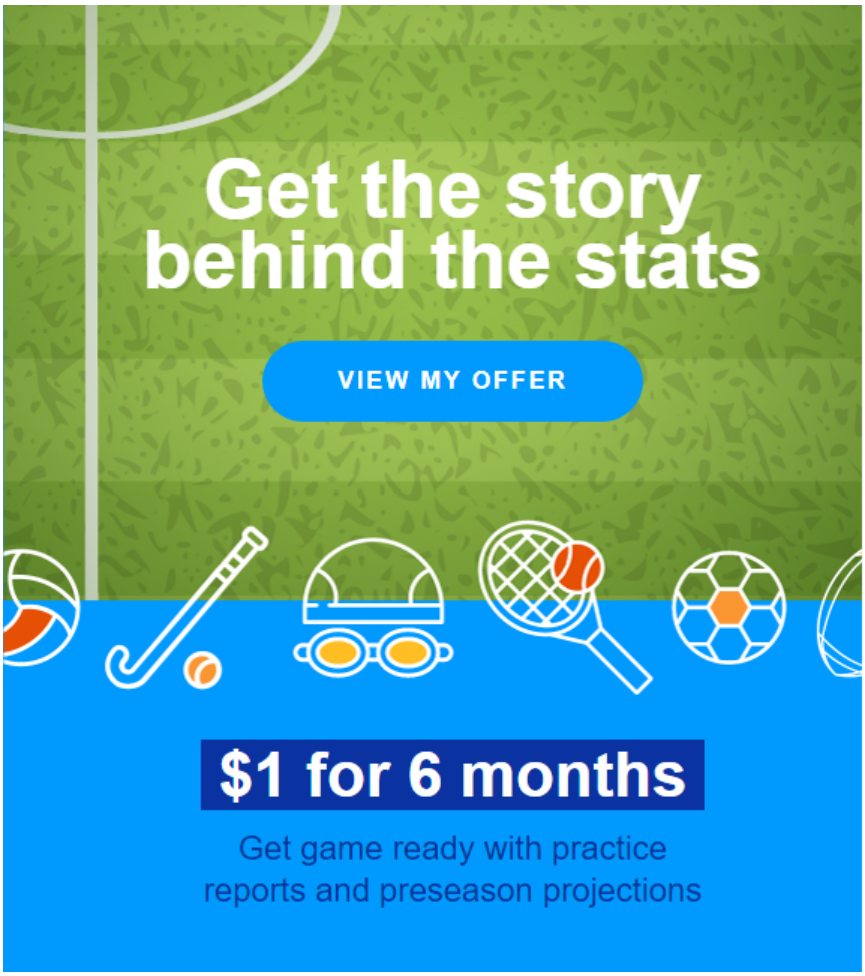


Figure 9: The Fayetteville Observer ad (Aug 5, 2022)

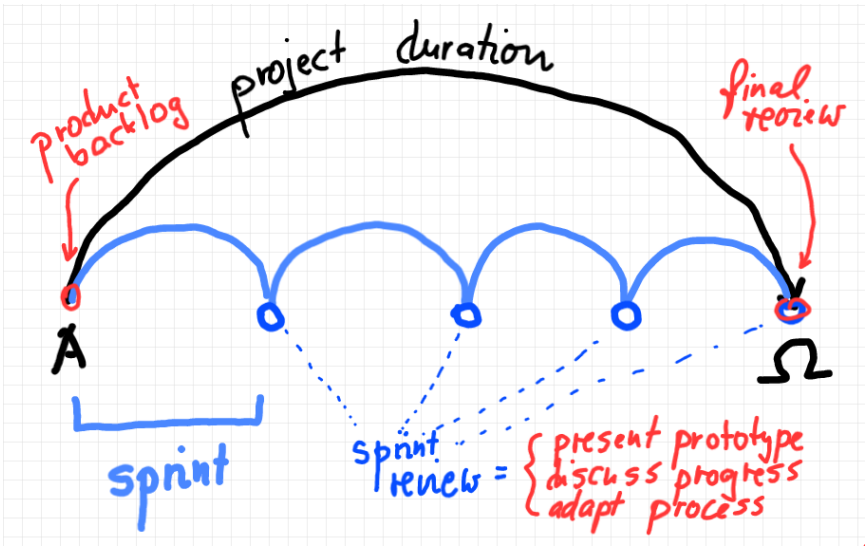


Figure 10: Agile (Scrum) project

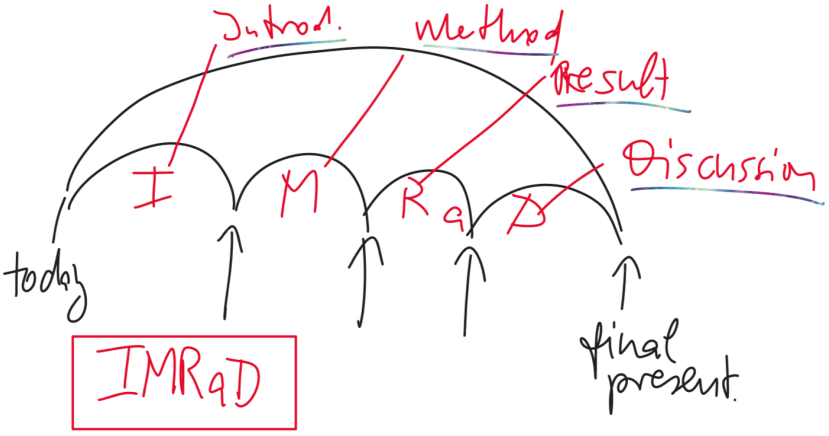
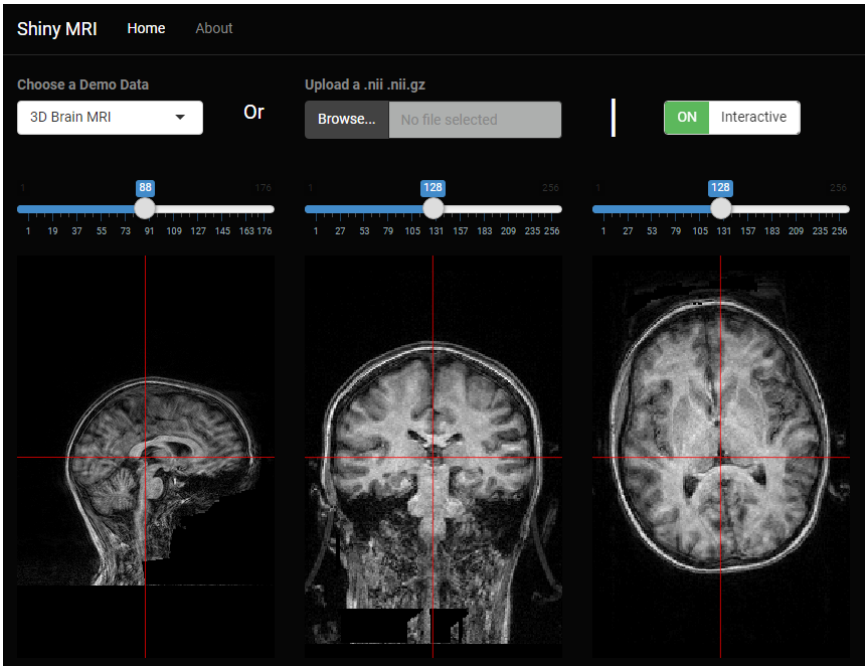


Figure 11: Agile (Scrum) project

11 Many project opportunities

n



Visualize 3D/4D medical imaging data in the browser

- Create an interesting data visualization (examples)
- Explore a graphics or animation package (like here)
- Solve a real-world problem (like here)
- Road scouts! Explain how maps are made nowadays (cp. with 1940)
- Analyse existing visualizations (like here)
- See DataCamp projects for examples, or a DataCamp competition
- Explore a data visualization tool
- Visualize whale song / double up between 2 or 3 courses
- Explore any of these graphics solutions (`base`, `ggplot2` and Shiny are covered in this course already):

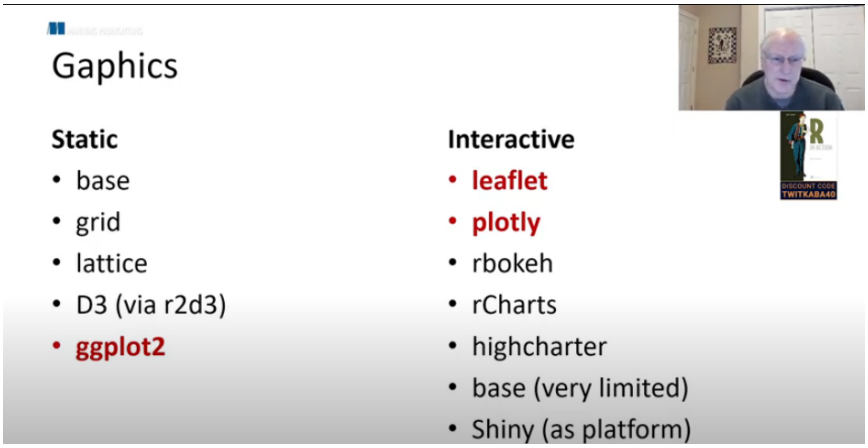


Figure 12: Source: Modern Data Visualization with R (Kabacoff, 2021)

12 Video lectures

- Emacs + Org-mode + R (Tutorial videos Spring '22)
- Introduction to R: installation and shell
- Vectors in R (part 1, part 2, part 3)
- Data frames, matrices, lists, factors in R
- Data frames in R
- Base R plotting
- Plotting with ggplot2
- Data import with R
- RStudio R Notebooks and literate programming

13 Introduction to DataCamp

- DataCamp is a data science learning platform
- Access for you is free (classroom license)
- 9/15 assignments are DataCamp assignments

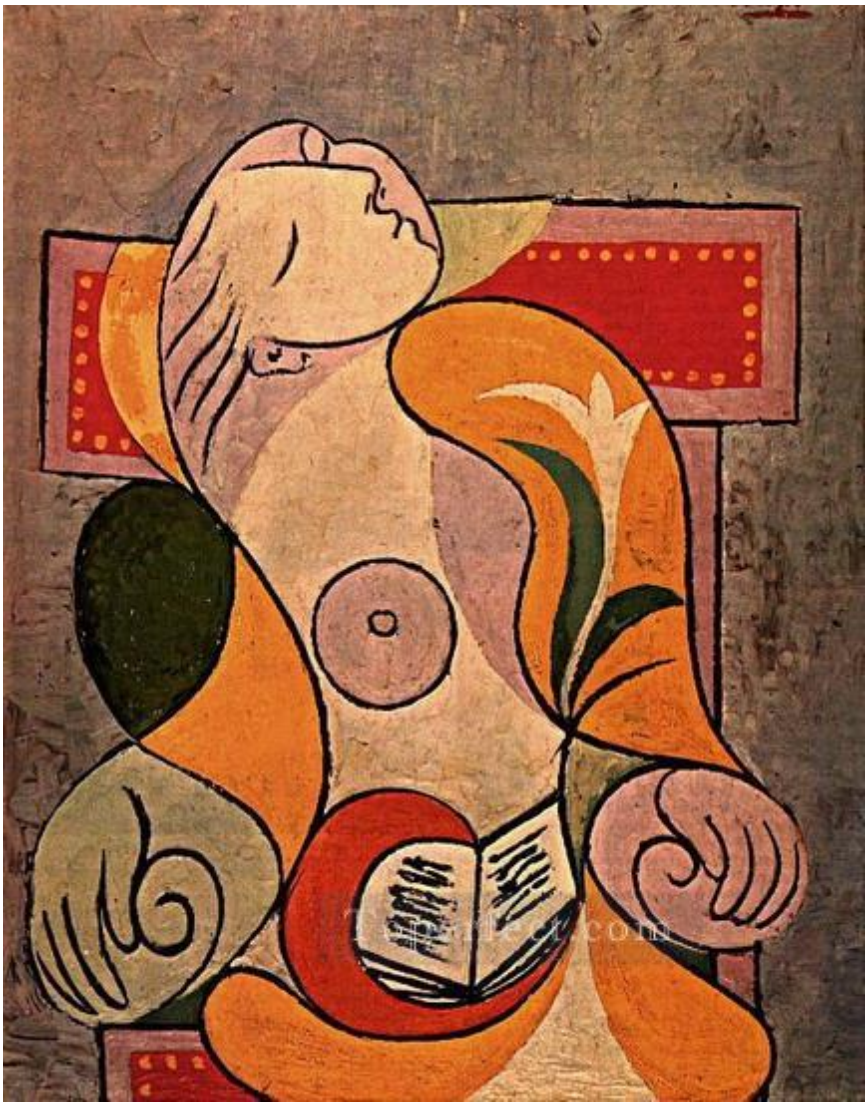


Figure 13: La lecture Marie Therese (Picasso, 1932)

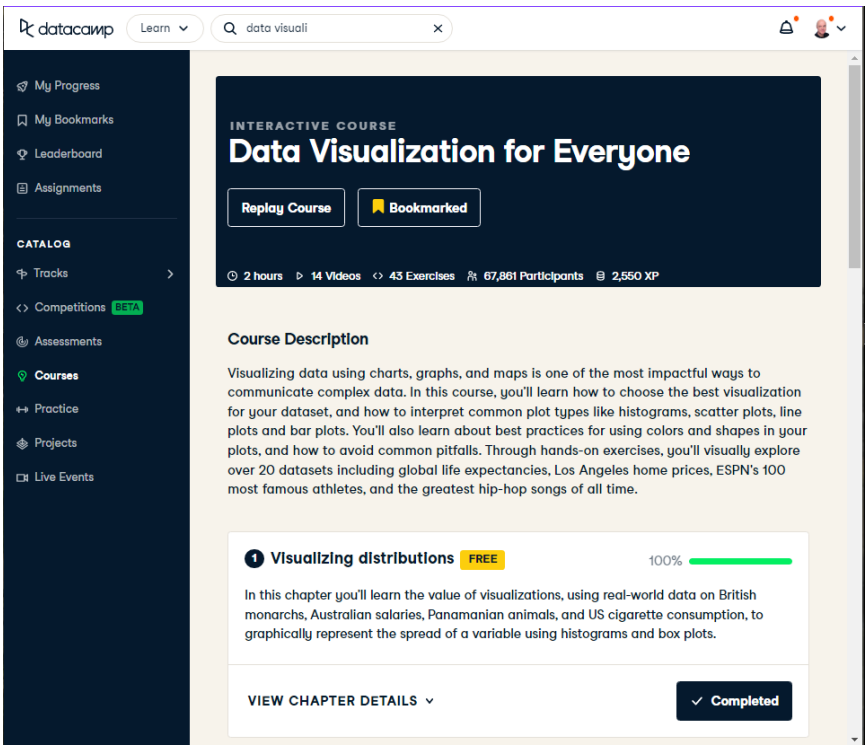


Figure 14: DataCamp course "Data Visualization For Everyone" start page

- Assignments are drawn from 5 courses
    1. Data visualization for everyone
    2. Data visualization with R
    3. Introduction to data visualization with ggplot2
    4. Building web applications with Shiny in R
    5. Introduction to Tableau
- Complete them on time to get full points
- Completed DataCamp courses can support your resume

14 Introduction to the textbook

- R is *FOSS* with focus on stats and graphics
- Pearson's "EDA Using R" is extensive (563 pp.)
- You don't have to read along but it might help

15 Other sources

- Introduction to data visualization: Wilke (2019) - **in library**
- Many other tutorials and textbooks available
- The best (free) short online tutorial: Matloff's "fasteR"
- The best complete textbook: Davies' "Book of R" - **in library**
- Beware of ideologies (cp. Matloff's "TidyverseSceptic")

16 Introduction to GNU Emacs + ESS + Org-mode

- Emacs: self-documenting, extensible *FOSS* text editor
- Process, file and package management (like an OS)
- *Literate programming* environment for 43 languages
- *IDE* for R programming and *REPL* for interactive coding

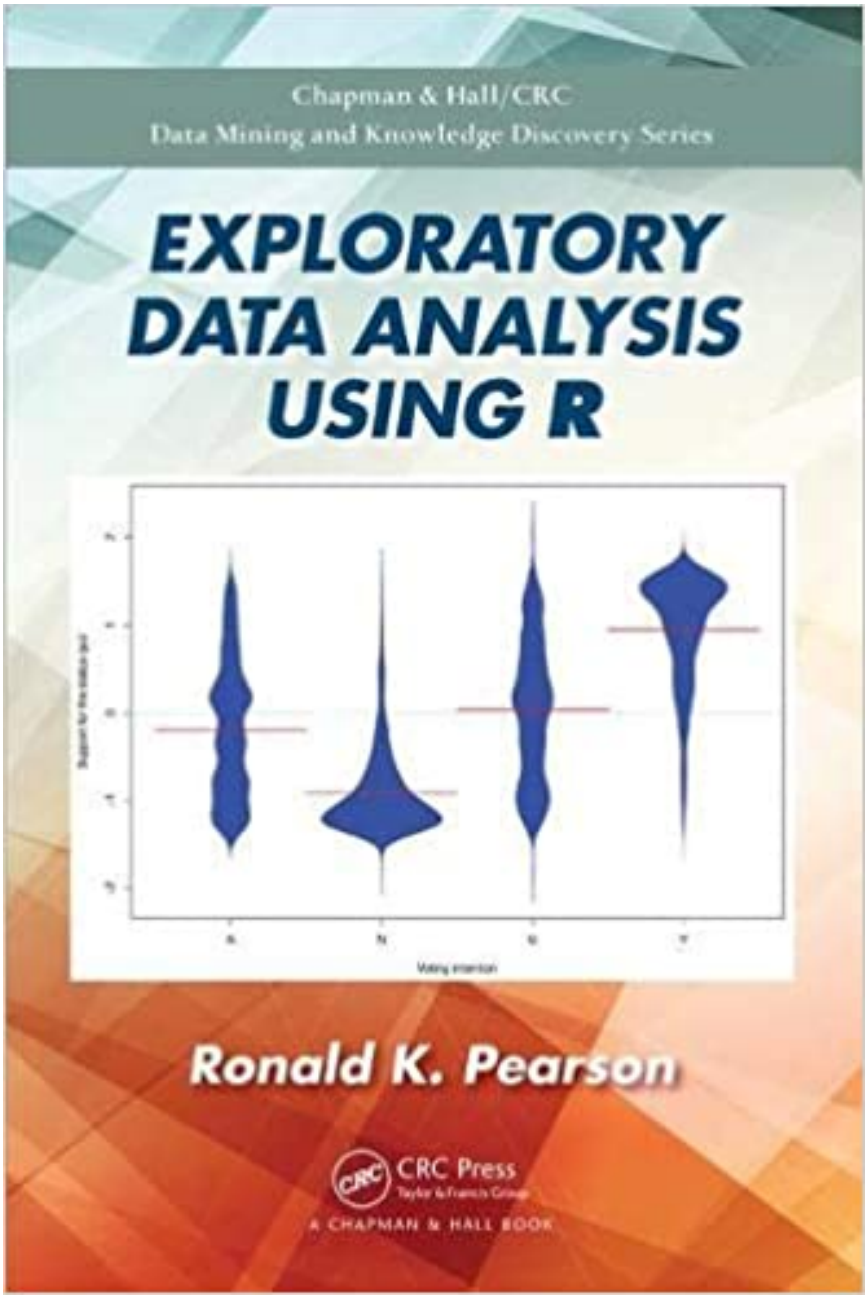


Figure 15: Cover of EDA Using R (Pearson, 2018)

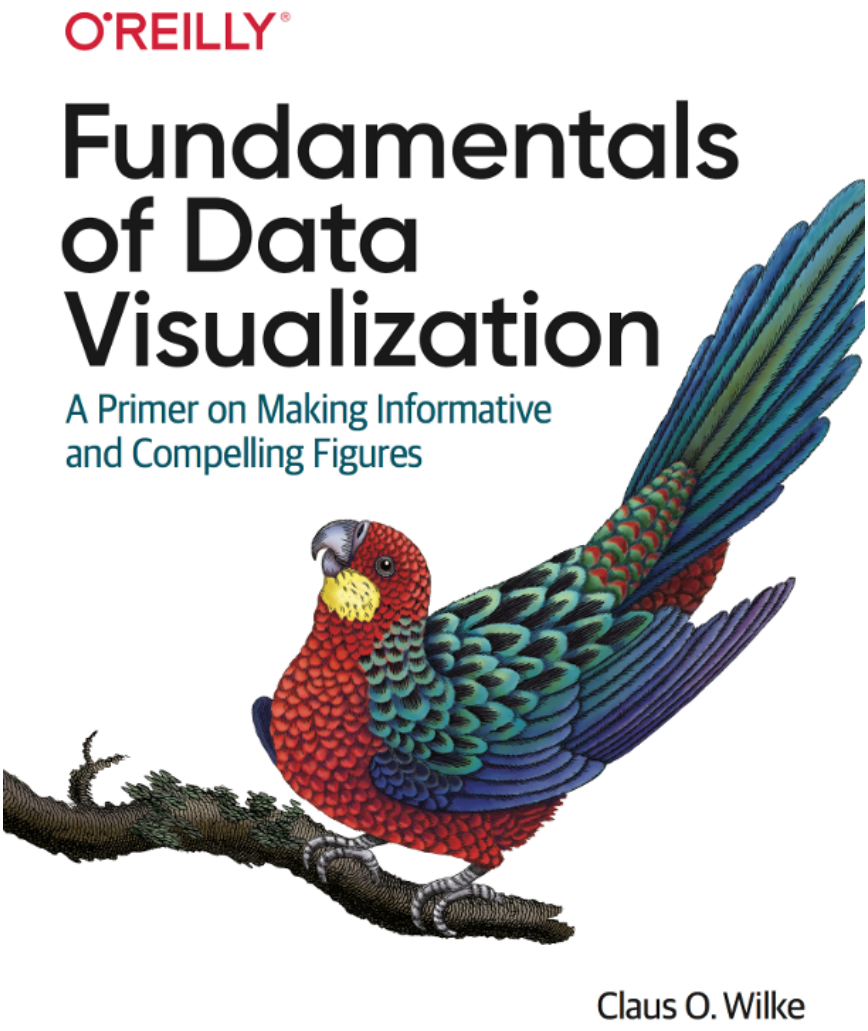


Figure 16: Cover of Fundamentals of Data Visualization (2019) by Claus Wilke

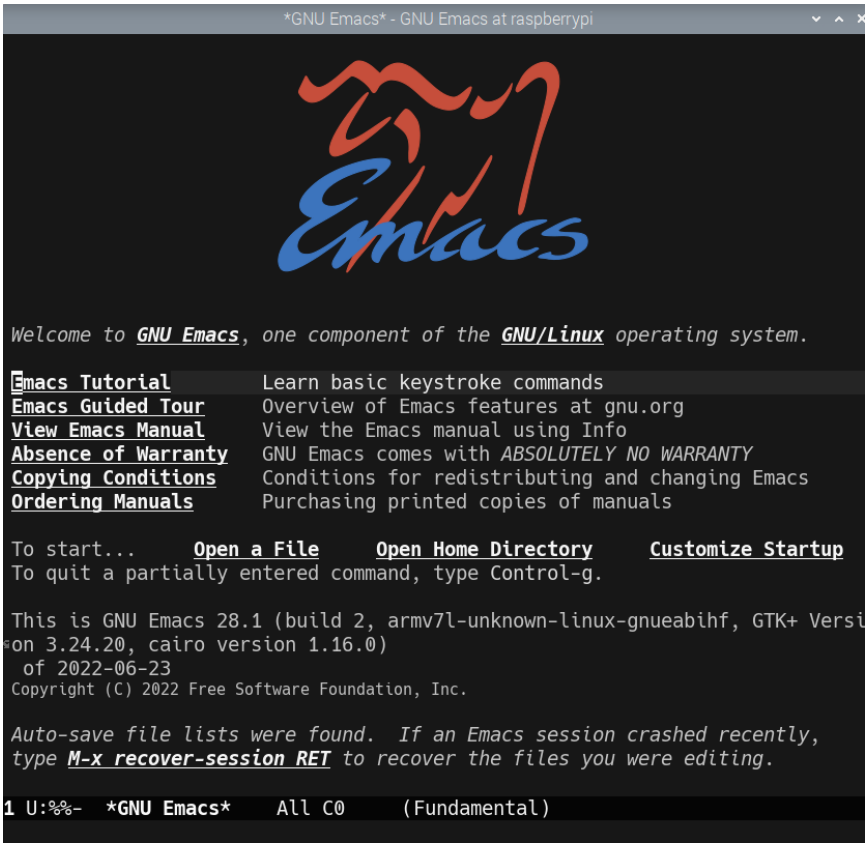


Figure 17: GNU Emacs start page

17 Literate programming

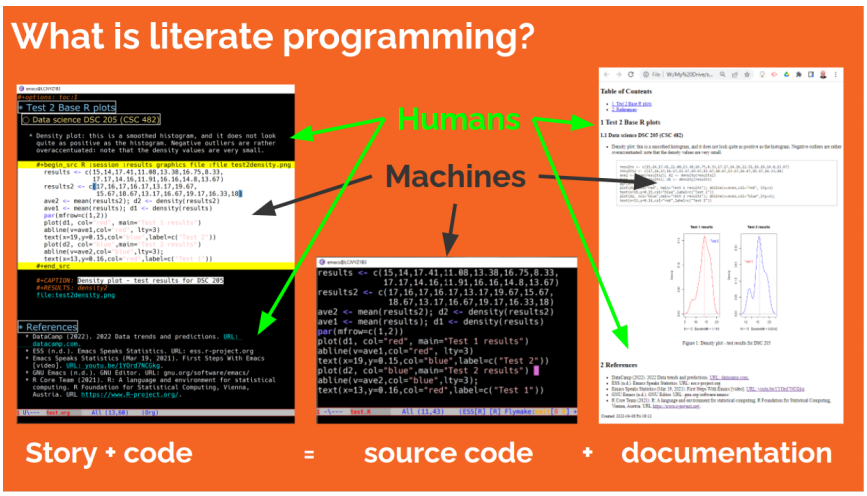


Figure 18: What is literate programming?

Source: "Teaching data science with hacker tools" (2022)

- Common practice among data scientists
- *Paradigm* behind interactive computing notebooks
- Useful when learning any programming language

18 Home assignments

- Register with DataCamp and complete the DataCamp chapter "Visualizing distributions" from the course "Data visualization for everyone".
  - Motivating visualization of data
  - Continous vs. categorical variables
  - Plot types: histograms and box plots
- If you don't know Emacs, complete the Emacs on-board tutorial!
  - Get comfortable with Emacs keyboard bindings
  - Learn how to create, view, edit, save files
  - Learn how to insert a time stamp automatically

19 Tests

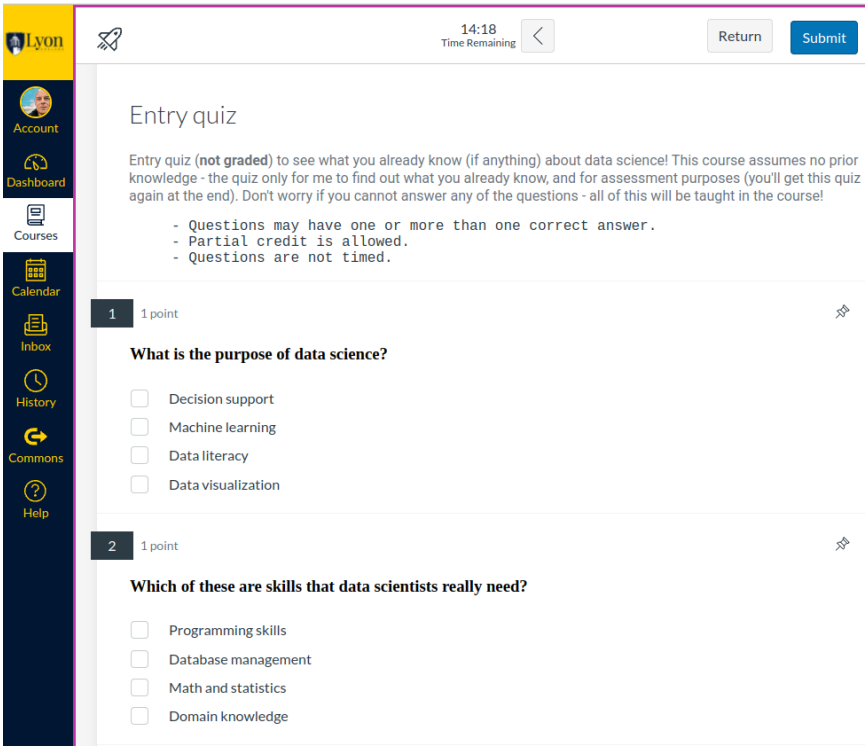


Figure 19: Start page of the entry quiz on Canvas

- Tests have to be completed online, are timed, and have a deadline; after the deadline, you can play them an unlimited number of times
- There will be a revision quiz on Canvas every week, consisting of 5-10 multiple choice, matching and true/false questions.
- A subset of the test questions will form the final exam, which is optional (you don't have to do it if you're happy with your grade).

20 Practice: Course infrastructure

- GitHub
- Linux

- Emacs

21 Glossary

TERM	MEANING
Command line	aka terminal/shell to talk to the OS
Emacs	GNU self-extensible text editor
FOSS	Free and Open Source Software
GitHub	Software development platform
Git	Version control software
GNU	GNU's not Unix
IDE	Integrated Development Environment
"Literate Programming"	Story + code => source code + doc
Paradigm	A standard way of looking at things
R	FOSS statistical programming language
REPL	Read-Eval-Print-Loop
Repo	Code repository
"Tidyverse"	Popular R package bundle
Scrum	Agile project management method
Sprint review	Period to complete a prototype
Prototype	Intermediate (not perfect) solution

22 References

- datavizblog.com (May 26, 2013).DataViz History: Charles Minard's Flow Map of Napoleon's Russian Campaign of 1812. Online: datavizblog.com
- Davies T D (2016). The Book of R. NoStarch Press.
- Pearson R K (2018). Exploratory Data Analysis Using R. CRC Press.
- Wilke C (2019). Fundamentals of Data Visualization. O'Reilly Media. Online: clauswilke.com