Course overview

Data visualization (DSC 302) Fall 2022

Table of Contents

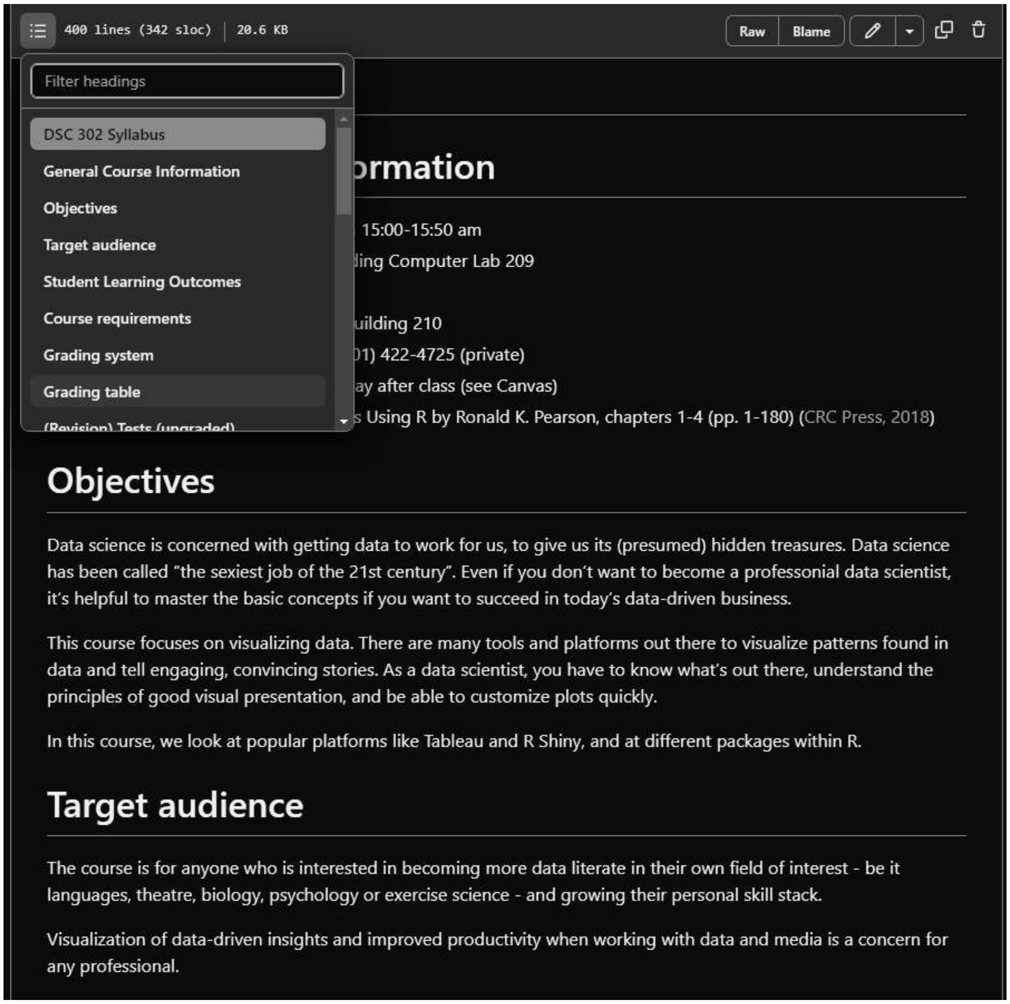
Figure 1: The Red Tower/La tour rouge (Giorgio de Chirico, 1913) Source: Guggenheim

1 MUTUAL INTRODUCTIONS

Figure 2: Marc Chagall, Over the town (2018) Source: Wikiart

1. Why are you here?
2. What would delight you?
3. What would disappoint you?
4. Where are you headed?

2 COURSE SYLLABUS (on GitHub and on Canvas)



ormation

15:00-15:50 am
ding Computer Lab 209

uilding 210
01) 422-4725 (private)
ay after class (see Canvas)
s Using R by Ronald K. Pearson, chapters 1-4 (pp. 1-180) (CRC Press, 2018)

Objectives

Data science is concerned with getting data to work for us, to give us its (presumed) hidden treasures. Data science has been called "the sexiest job of the 21st century". Even if you don't want to become a professional data scientist, it's helpful to master the basic concepts if you want to succeed in today's data-driven business.

This course focuses on visualizing data. There are many tools and platforms out there to visualize patterns found in data and tell engaging, convincing stories. As a data scientist, you have to know what's out there, understand the principles of good visual presentation, and be able to customize plots quickly.

In this course, we look at popular platforms like Tableau and R Shiny, and at different packages within R.

Target audience

The course is for anyone who is interested in becoming more data literate in their own field of interest - be it languages, theatre, biology, psychology or exercise science - and growing their personal skill stack.

Visualization of data-driven insights and improved productivity when working with data and media is a concern for any professional.

Figure 3: DSC 302 Syllabus on GitHub

- General information & standard policies
- Course information (grading, attendance)
- Schedule with dates of tests and assignments
- The GitHub repo contains course material

3 COURSE TOPICS (ILLUSTRATED)

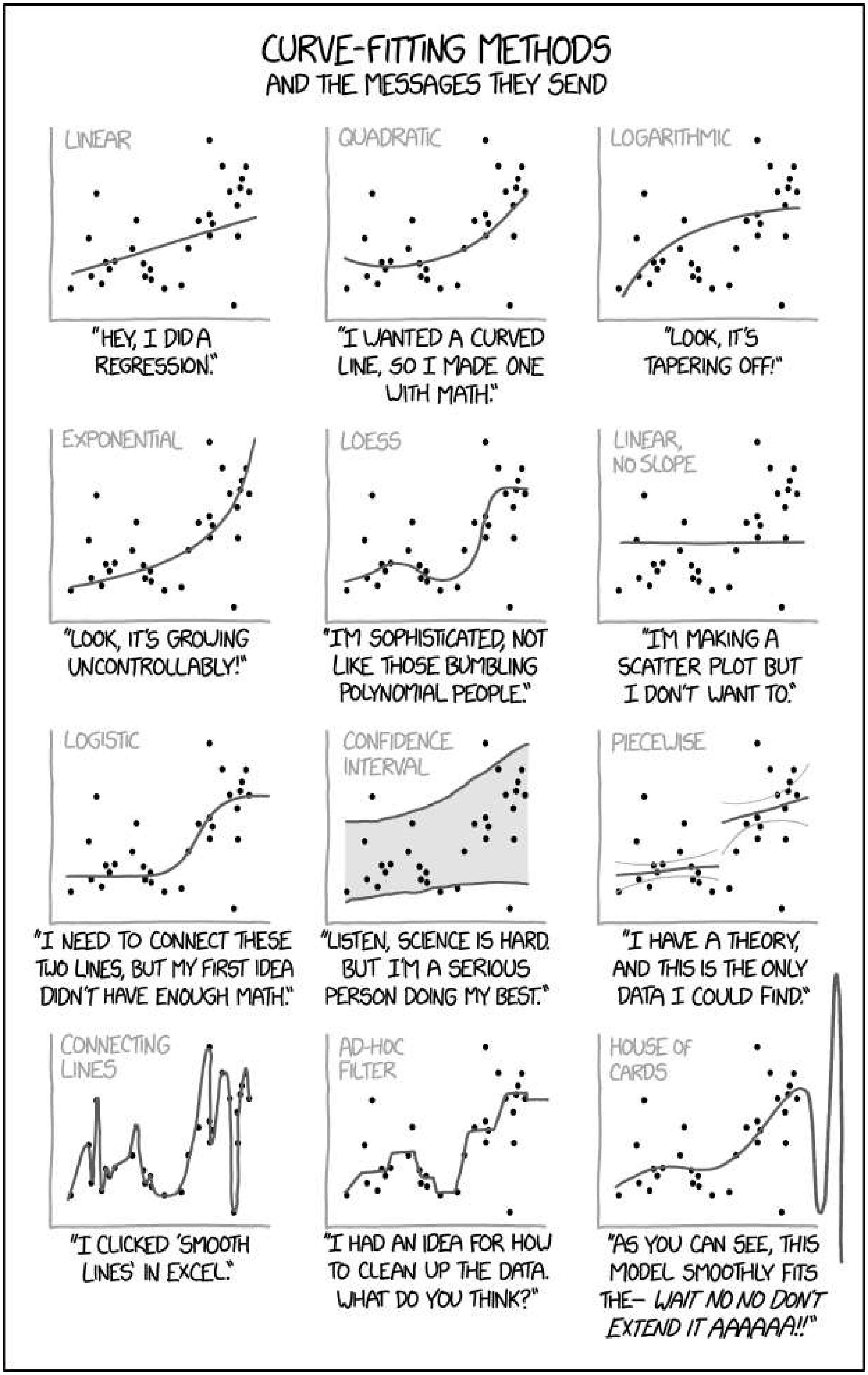
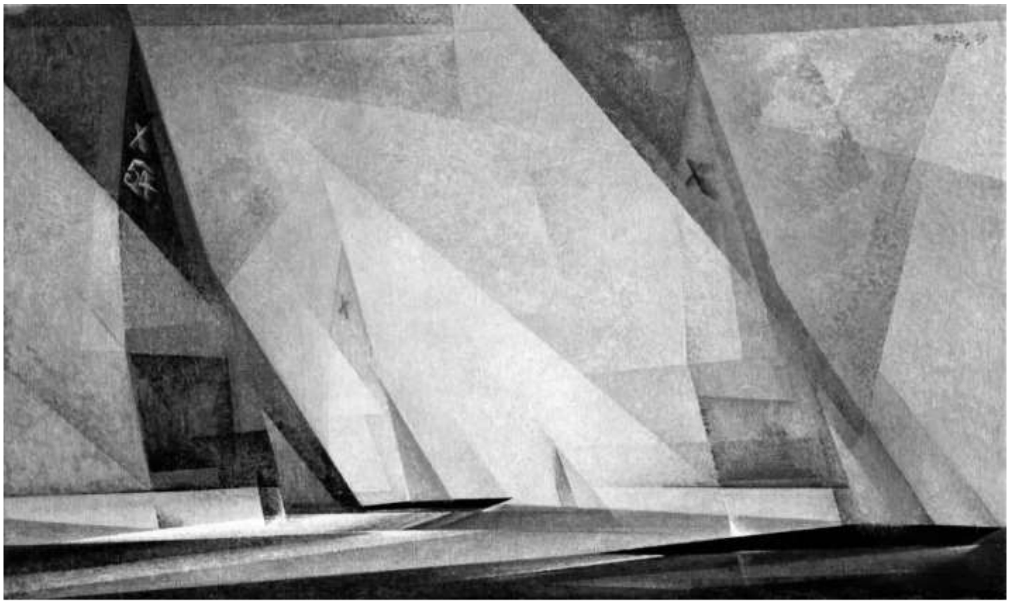

CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!!"

Figure 4: Course topics

4 COURSE TOPICS (SPELLED OUT)



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

Figure 5: Lyonel Feininger, Sailing Boats (1929)

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

5 WHY "DATA VISUALIZATION"?

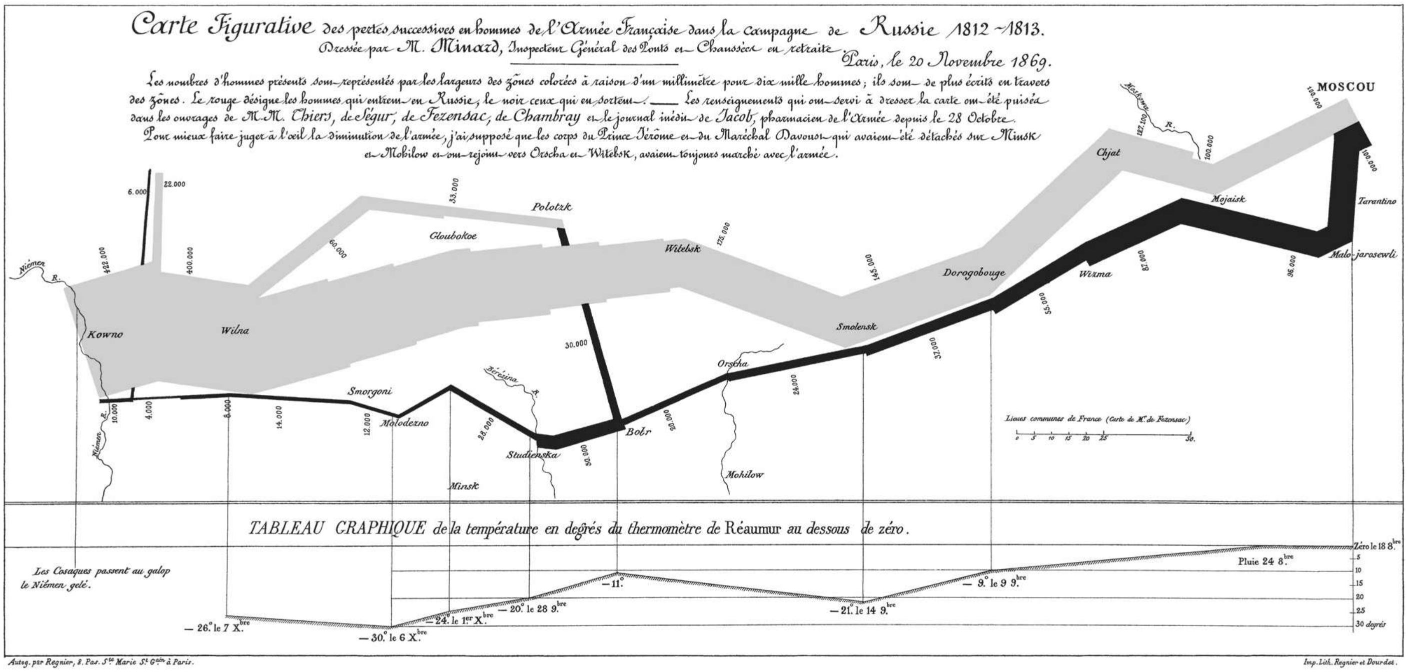

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de MM. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
Niémen R.
422.000
6.000
22.000
400.000
60.000
33.000
Polotzk
Gloubokoe
Witebsk
175.000
145.000
Dorogobouge
Chjat
127.100
Moskowa R.
100.000
MOSCOU
100.000
Mojaisk
Tarantino
Malo-jarosewli
96.000
87.000
Wizma
55.000
Smolensk
37.000
Orscha
24.000
20.000
Bobr
30.000
Studienska
50.000
Bérézina R.
28.000
Minsk
Molodexno
12.000
Smorgoni
14.000
8.000
4.000
10.000
Kowno
Wilna
Niémen R.
Mohilow
Lieues communes de France (Carte de M. de Fezensac)
0 5 10 15 20 25 50
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
— 26° le 7 X.bre
— 30° le 6 X.bre
— 24° le 1er X.bre
— 20° le 28 9.bre
— 11°
— 21° le 14 9.bre
— 9° le 9 9.bre
Pluie 24 8.bre
Zéro le 18 8.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.
Imp. Lith. Regnier et Dourdet.

Figure 6: Charles Minard, Napoleon's Russian campaign 1812

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

- 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" (Война и мир, 1867)

6 GET THE STORY BEHIND THE STATS

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

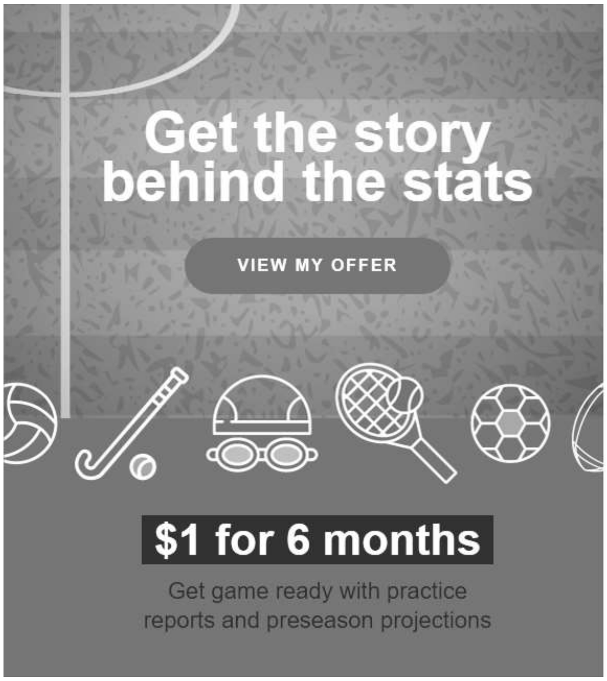

Get the story behind the stats
VIEW MY OFFER
$1 for 6 months
Get game ready with practice reports and preseason projections

Figure 7: The Fayetteville Observer ad (Aug 5, 2022)

7 AGILE TEAM PROJECT

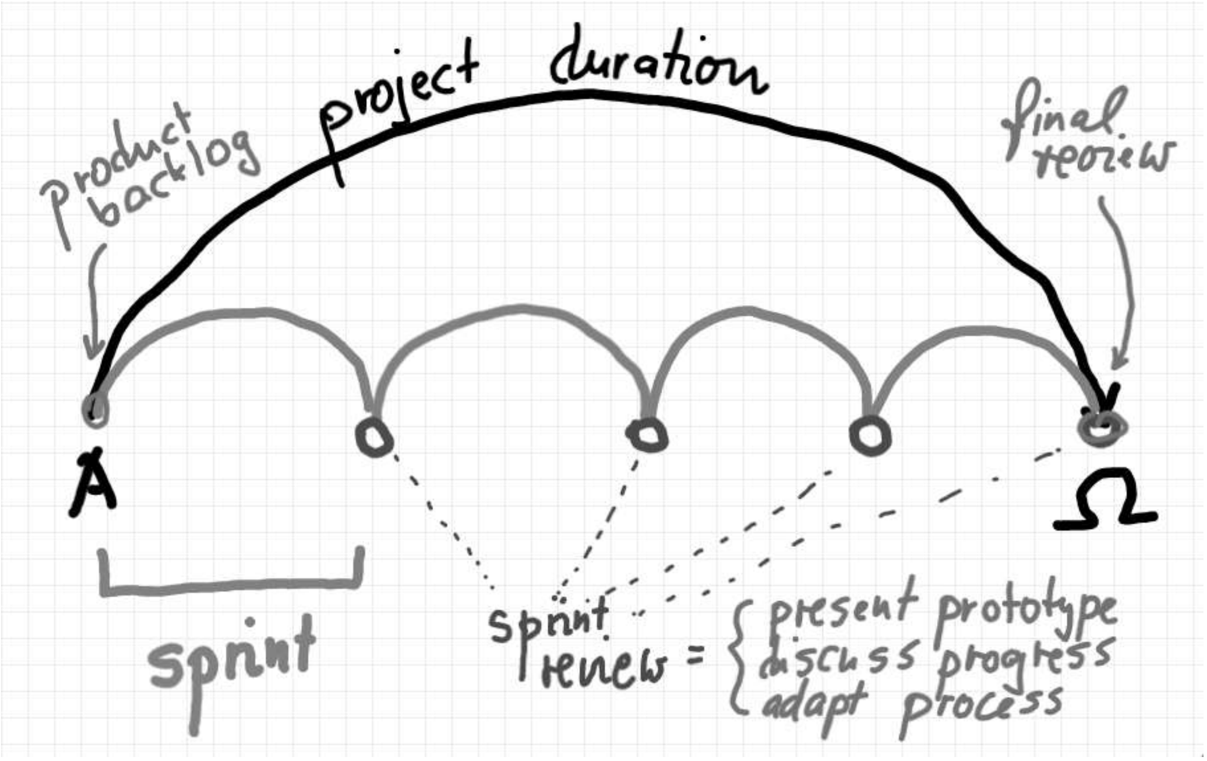


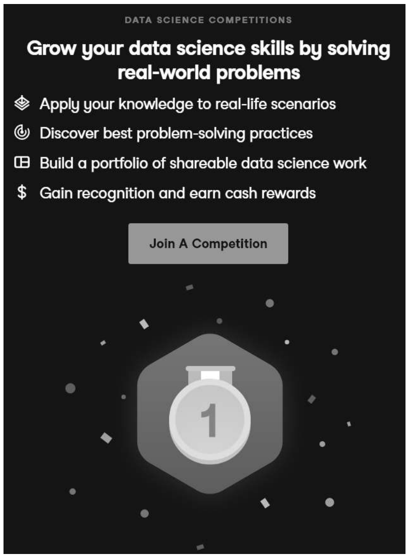
Figure 8: Agile (Scrum) project

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

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

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).

8 MANY PROJECT OPPORTUNITIES


DATA SCIENCE COMPETITIONS

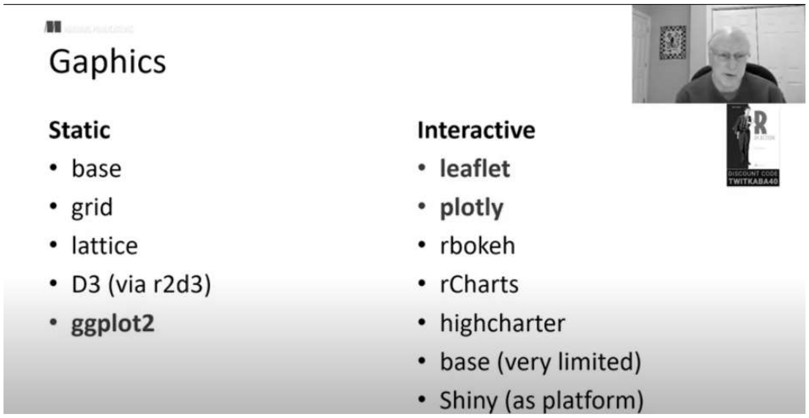
Grow your data science skills by solving real-world problems

- Apply your knowledge to real-life scenarios
- Discover best problem-solving practices
- Build a portfolio of shareable data science work
- Gain recognition and earn cash rewards

Figure 9: DataCamp competition announcement

- Create an interesting data visualization
- Explore a graphics or animation package
- Solve a real-world problem
- Analyse existing visualizations
- See DataCamp projects for examples
- 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):

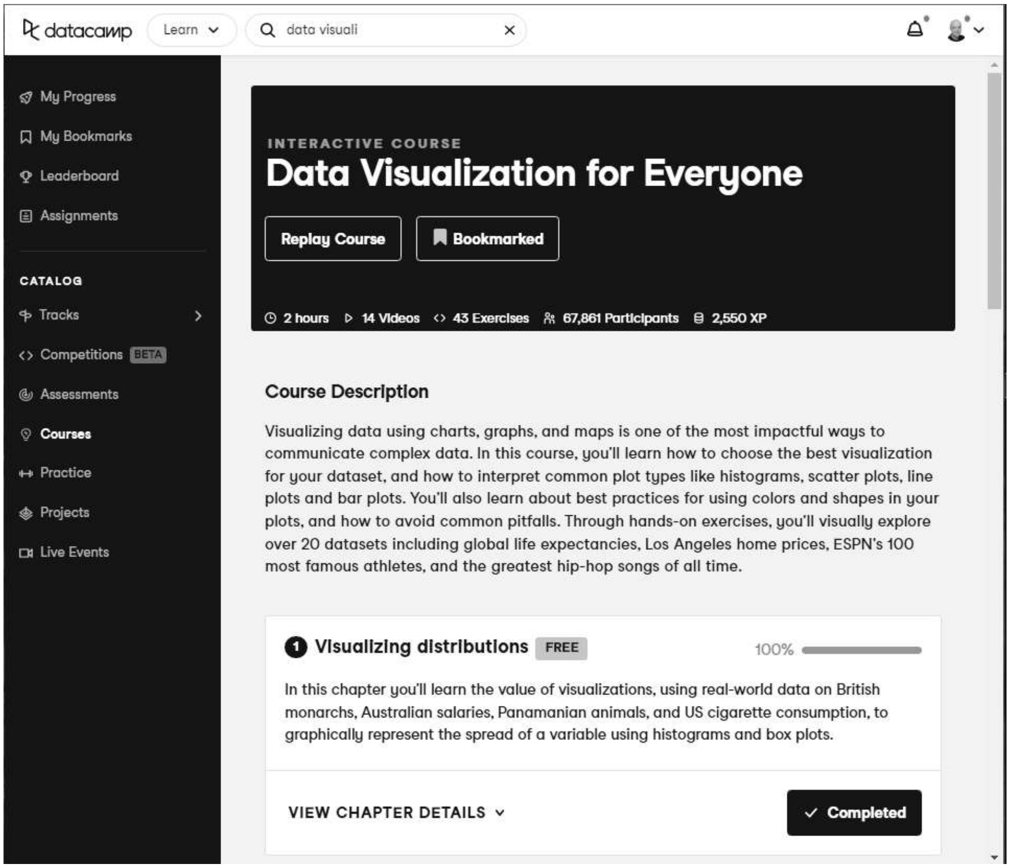

Gaphics

Static
- base
- grid
- lattice
- D3 (via r2d3)
- **ggplot2**

Interactive
- **leaflet**
- **plotly**
- rbokeh
- rCharts
- highcharter
- base (very limited)
- Shiny (as platform)

Figure 10: Source: Modern Data Visualization with R (Kabacoff, 2021)

9 INTRODUCTION TO DataCamp



INTERACTIVE COURSE

Data Visualization for Everyone

2 hours 14 Videos 43 Exercises 67,861 Participants 2,550 XP

Course Description

Visualizing data using charts, graphs, and maps is one of the most impactful ways to communicate complex data. In this course, you'll learn how to choose the best visualization for your dataset, and how to interpret common plot types like histograms, scatter plots, line plots and bar plots. You'll also learn about best practices for using colors and shapes in your plots, and how to avoid common pitfalls. Through hands-on exercises, you'll visually explore over 20 datasets including global life expectancies, Los Angeles home prices, ESPN's 100 most famous athletes, and the greatest hip-hop songs of all time.

1 Visualizing distributions FREE 100%

In this chapter you'll learn the value of visualizations, using real-world data on British monarchs, Australian salaries, Panamanian animals, and US cigarette consumption, to graphically represent the spread of a variable using histograms and box plots.

Figure 11: DataCamp course "Data Visualization For Everyone" start page

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

10 INTRODUCTION TO THE TEXTBOOK

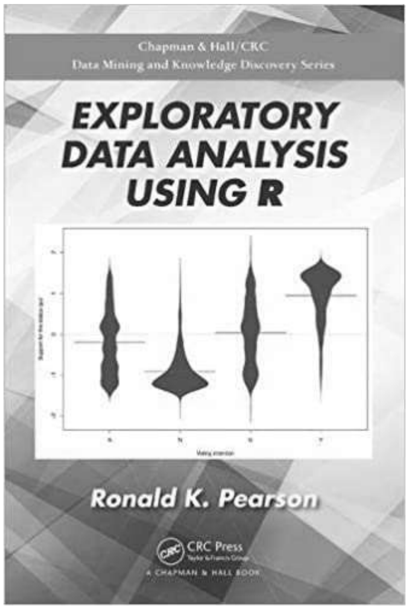

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series
EXPLORATORY DATA ANALYSIS USING R
Ronald K. Pearson
CRC Press
A CHAPMAN & HALL BOOK

Figure 12: Cover of EDA Using R (Pearson, 2018)

- 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

11 OTHER SOURCES

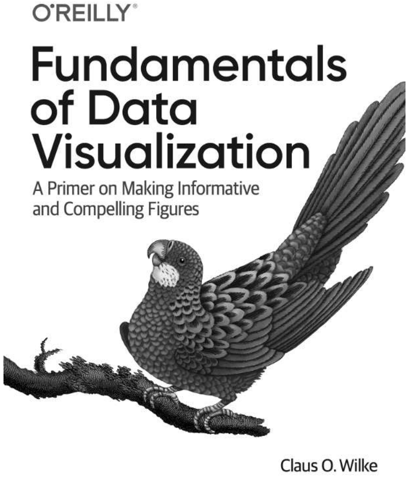
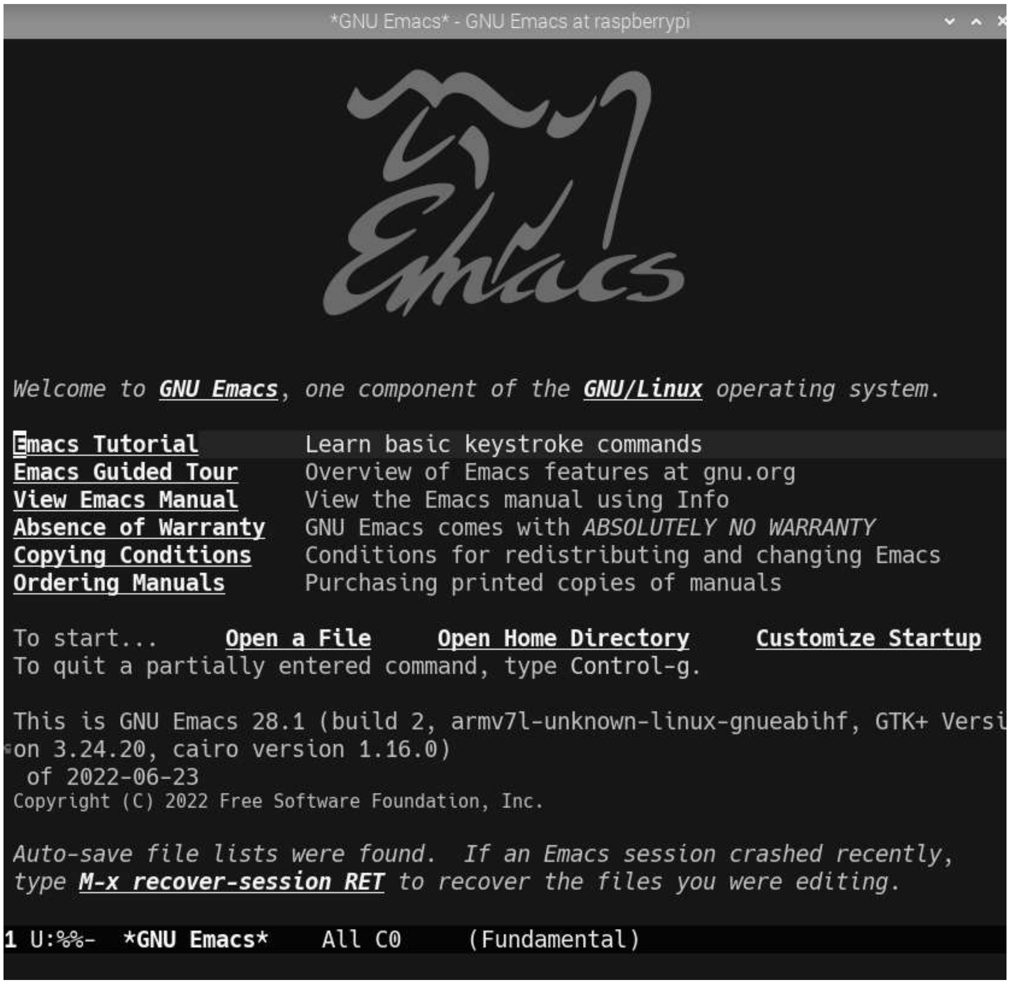

O'REILLY
Fundamentals of Data Visualization
A Primer on Making Informative and Compelling Figures
Claus O. Wilke

Figure 13: Cover of Fundamentals of Data Visualization (2019) by Claus Wilke

- 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")

12 INTRODUCTION to GNU Emacs + ESS + Org-mode

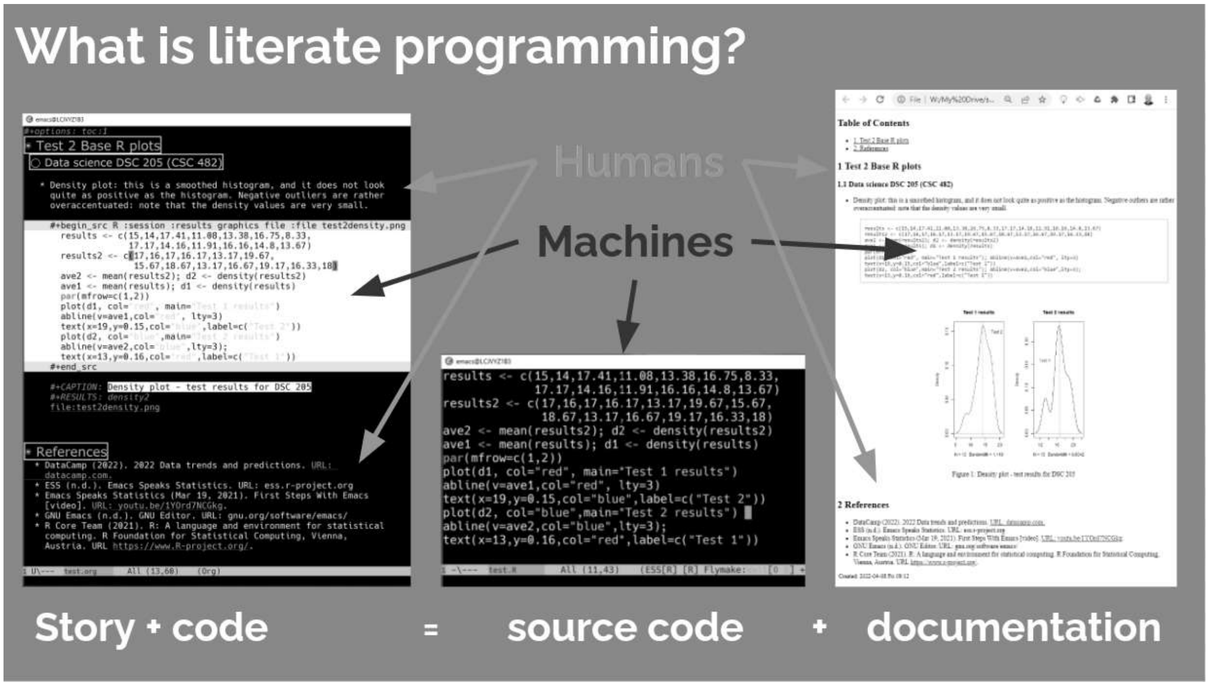


```
*GNU Emacs* - GNU Emacs at raspberrypi

Welcome to GNU Emacs, one component of the GNU/Linux operating system.

Emacs Tutorial          Learn basic keystroke commands
Emacs Guided Tour       Overview of Emacs features at gnu.org
View Emacs Manual       View the Emacs manual using Info
Absence of Warranty     GNU Emacs comes with ABSOLUTELY NO WARRANTY
Copying Conditions      Conditions for redistributing and changing Emacs
Ordering Manuals        Purchasing printed copies of manuals

To start...     Open a File     Open Home Directory     Customize Startup
To quit a partially entered command, type Control-g.

This is GNU Emacs 28.1 (build 2, armv7l-unknown-linux-gnueabihf, GTK+ Versi
on 3.24.20, cairo version 1.16.0)
 of 2022-06-23
Copyright (C) 2022 Free Software Foundation, Inc.

Auto-save file lists were found.  If an Emacs session crashed recently,
type M-x recover-session RET to recover the files you were editing.

1 U:%%-  *GNU Emacs*    All C0     (Fundamental)
```

Figure 14: GNU Emacs start page

- 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

13 LITERATE PROGRAMMING



What is literate programming?

Humans

Machines

```
#+options: toc:1
* Test 2 Base R plots
○ Data science DSC 205 (CSC 482)

  * Density plot: this is a smoothed histogram, and it does not look
    quite as positive as the histogram. Negative outliers are rather
    overaccentuated: note that the density values are very small.

    #+begin_src R :session :results graphics file :file test2density.png
      results <- c(15,14,17.41,11.08,13.38,16.75,8.33,
                   17.17,14.16,11.91,16.16,14.8,13.67)
      results2 <- c(17,16,17,16.17,13.17,19.67,
                    15.67,18.67,13.17,16.67,19.17,16.33,18)
      ave2 <- mean(results2); d2 <- density(results2)
      ave1 <- mean(results); d1 <- density(results)
      par(mfrow=c(1,2))
      plot(d1, col="red", main="Test 1 results")
      abline(v=ave1,col="red", lty=3)
      text(x=19,y=0.15,col="blue",label=c("Test 2"))
      plot(d2, col="blue",main="Test 2 results")
      abline(v=ave2,col="blue",lty=3);
      text(x=13,y=0.16,col="red",label=c("Test 1"))
    #+end_src

    #+CAPTION: Density plot - test results for DSC 205
    #+RESULTS: density2
    file:test2density.png


* References
 * DataCamp (2022). 2022 Data trends and predictions. URL:
   datacamp.com.
 * ESS (n.d.). Emacs Speaks Statistics. URL: ess.r-project.org
 * Emacs Speaks Statistics (Mar 19, 2021). First Steps With Emacs
   [video]. URL: youtu.be/1YOrd7NCGkg.
 * GNU Emacs (n.d.). GNU Editor. URL: gnu.org/software/emacs/
 * R Core Team (2021). R: A language and environment for statistical
   computing. R Foundation for Statistical Computing, Vienna,
   Austria. URL https://www.R-project.org/.

1 U\---  test.org     All (13,60)    (Org)
```

```
results <- c(15,14,17.41,11.08,13.38,16.75,8.33,
             17.17,14.16,11.91,16.16,14.8,13.67)
results2 <- c(17,16,17,16.17,13.17,19.67,15.67,
              18.67,13.17,16.67,19.17,16.33,18)
ave2 <- mean(results2); d2 <- density(results2)
ave1 <- mean(results); d1 <- density(results)
par(mfrow=c(1,2))
plot(d1, col="red", main="Test 1 results")
abline(v=ave1,col="red", lty=3)
text(x=19,y=0.15,col="blue",label=c("Test 2"))
plot(d2, col="blue",main="Test 2 results")
abline(v=ave2,col="blue",lty=3);
text(x=13,y=0.16,col="red",label=c("Test 1"))

1 -\---  test.R     All (11,43)    (ESS[R] [R] Flymake: [0 ] +
```

Table of Contents

- 1. Test 2 Base R plots
- 2. References

1 Test 2 Base R plots

1.1 Data science DSC 205 (CSC 482)

- Density plot: this is a smoothed histogram, and it does not look quite as positive as the histogram. Negative outliers are rather overaccentuated: note that the density values are very small.

```
[illegible]
```

Figure 1: Density plot - test results for DSC 205

2 References

- DataCamp (2022). 2022 Data trends and predictions. URL: datacamp.com.
- ESS (n.d.). Emacs Speaks Statistics. URL: ess.r-project.org
- Emacs Speaks Statistics (Mar 19, 2021). First Steps With Emacs [video]. URL: youtu.be/1YOrd7NCGkg
- GNU Emacs (n.d.). GNU Editor. URL: gnu.org/software/emacs/
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.r-project.org/.

Created: 2022-04-08 Fri 09:12

Story + code = source code + documentation

Figure 15: What is literate programming?

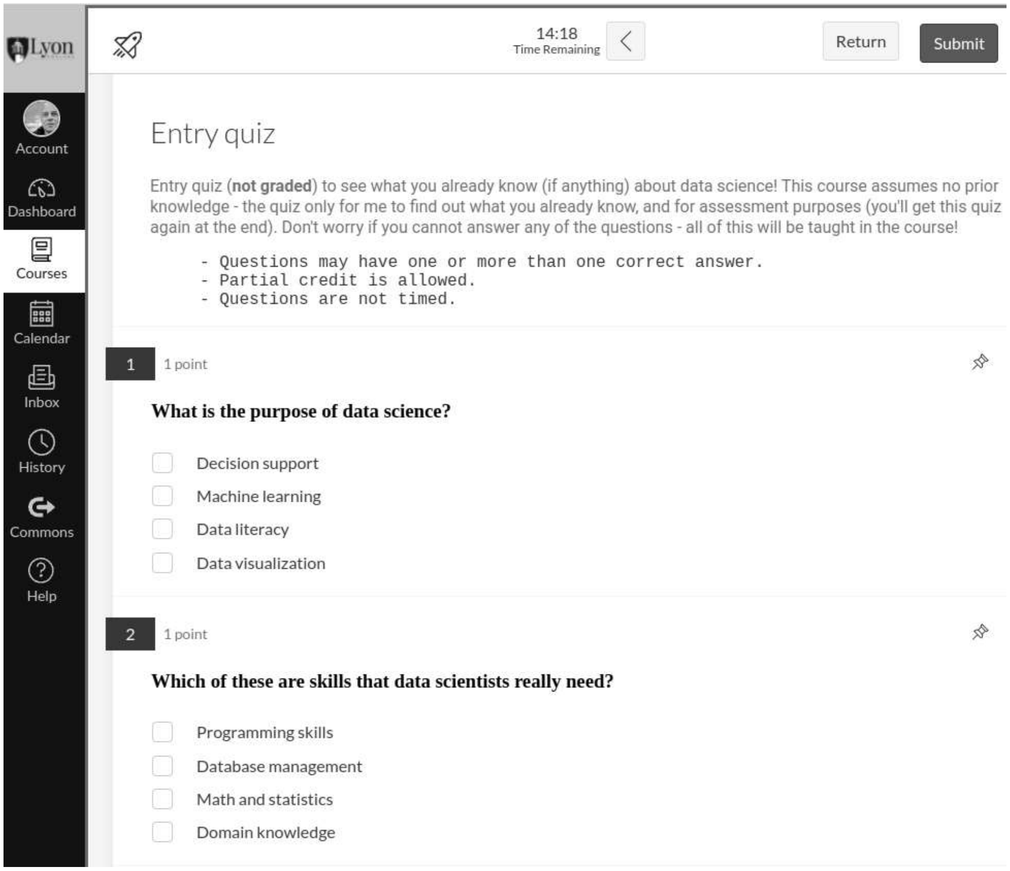
Source: "Teaching data science with hacker tools" (2022)

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

14 HOME ASSIGNMENTS

- There are 15 programming assignments altogether = 10 points each, or 30% of your final grade.
- Register with DataCamp and complete the DataCamp chapter "Visualizing distributions" from the course "Data visualization for everyone" by Monday, 22 August at 3 pm (ca. 20 min).
  - Motivating visualization of data
  - Continous vs. categorical variables
  - Plot types: histograms and box plots
- Complete the Emacs on-board tutorial and upload an edited copy to Canvas by Friday, 26 August at 11 am (ca. 60 min).
  - Get comfortable with Emacs keyboard bindings
  - Learn how to create, view, edit, save files
  - Learn how to insert a time stamp automatically

15 TESTS (NOT GRADED)



Entry quiz

Entry quiz (**not graded**) to see what you already know (if anything) about data science! This course assumes no prior knowledge - the quiz only for me to find out what you already know, and for assessment purposes (you'll get this quiz again at the end). Don't worry if you cannot answer any of the questions - all of this will be taught in the course!

```
- Questions may have one or more than one correct answer.
- Partial credit is allowed.
- Questions are not timed.
```

1 1 point

What is the purpose of data science?

□ Decision support
□ Machine learning
□ Data literacy
□ Data visualization

2 1 point

Which of these are skills that data scientists really need?

□ Programming skills
□ Database management
□ Math and statistics
□ Domain knowledge

Figure 16: 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 (20% of your final grade) - we will practice in the last week before the exam.

16 PRACTICE - COURSE INFRASTRUCTURE

Useful: take notes! Practice leads to mastery and the practice exercises will often come back to haunt you in the tests.

1. Open a browser
2. Find the GitHub repos (birkenkrahe/ds1 and /org)
3. Open the command line terminal
4. Open/close R
5. Open Emacs
6. Find the Emacs tutorial
7. Open/close R inside Emacs
8. Run R in an Org-mode file
9. Close Emacs
10. Close the command line terminal

Note: Class room practice completion = 10 points each for active participation.

17 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

18 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

Author: Marcus Birkenkrahe

Created: 2022-08-07 Sun 16:45