DS Agenda

Agenda for CSC482/DSC205 Introduction to Advanced Data Science Spring 2022

Table of Contents

1 README

This file contains the agenda overview (what I had planned), the objectives (what we managed to do) and (much of the) content of each taught session of the course. I want to avoid splitting the content up over many files - so that you have to navigate as little as possible (like a book)!

The companion file to this file, less structured and with the captain's log, is the notes.org file.

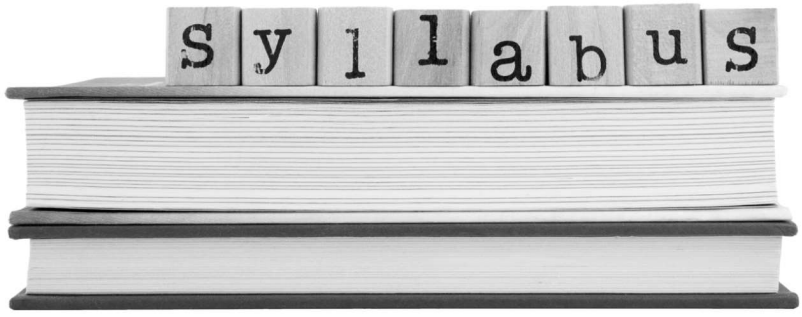
2 Course introduction - w1s1 (12-Jan)

2.1 Welcome

- Aspirations - changes spring 2022
- Ambitions - program 2021-2023
- Antagonization - new data science credo
- Syllabus - this course
- DataCamp assignments
- GNU Emacs Org-mode

  "After a course is launched, we don't consider it to be complete: the launch is just the start of data collection." Richie Cotton, DataCamp

2.2 Syllabus



- Syllabus in Schoology
- Syllabus in GitHub
- Schedule in GitHub

2.3 Aspirations (Changes in Spring 2022)

Cp. Good-bye fall 2021

FALL 2021	SPRINT 2021
Base R (stick shift) instead of "TidyVerse" (automatic)	Adding the "Tidyverse"
Use of interactive notebooks (literate programming!)	Intro to RStudio IDE and Emacs
Use GitHub as a code and materials repository	GitHub repo
Create lots of (ungraded) tests	Graded quizzes and tests
Use of DataCamp assignments	DataCamp assignments
Avoid mathematics as much as possible	No math
Reuse tests for the final exam	Reuse quizzes for final exam
Let students pick their own projects	No projects (only optional)

2.4 Ambitions (DS program 2021-2023)

CLASS	CODE	TERM	Topics
Data Science Tools and Methods	DSC 101	Fall 2021	R, Basic EDA, Base R
Introduction to Advanced Data Science	DSC 205	Spring 2022	R, Advanced EDA, Tidyverse, shell
Database Theory and Applications	CSC 330	Spring 2022	SQL, SQLite
Operating Systems	CSC 420	Spring 2022	Bash, awk, sed, regular expressions
Applied Math for Data Science	DSC 482/MTH 360	Fall 2022	Probability, Statistics + R
Data Visualization	DSC 302	Fall 2022	D3, Processing, Javascript, Bokeh
Machine Learning	DSC 305	Spring 2023	Predictive algorithms, neural nets
Digital Humanities	CSC 105	Spring 2023	Data science applications

2.5 DataCamp

Intermediate R Conditionals and Control Flow Chapter	Team	Active	Jan 31, 15:00 CST	0	0	0%	View
Intermediate R Loops Chapter	Team	Active	Feb 7, 15:00 CST	0	0	0%	View
Intermediate R Functions Chapter	Team	Active	Feb 14, 15:00 CST	0	0	0%	View
Intermediate R The apply family Chapter	Team	Active	Feb 21, 15:00 CST	0	0	0%	View
Intermediate R Utilities Chapter	Team	Active	Feb 28, 15:00 CST	0	0	0%	View
Introduction to the Tidyverse Data wrangling Chapter	Team	Active	Mar 7, 15:00 CST	0	0	0%	View
Introduction to the Tidyverse Data visualization Chapter	Team	Active	Mar 14, 15:00 CDT	0	0	0%	View
Introduction to the Tidyverse Grouping and summarizing Chapter	Team	Active	Mar 28, 15:00 CDT	0	0	0%	View
Introduction to the Tidyverse Types of visualizations Chapter	Team	Active	Apr 4, 15:00 CDT	0	0	0%	View
Exploratory Data Analysis in R Exploring Categorical Data Chapter	Team	Active	Apr 11, 15:00 CDT	0	0	0%	View
Exploratory Data Analysis in R Exploring Numerical Data Chapter	Team	Active	Apr 20, 15:00 CDT	0	0	0%	View
Exploratory Data Analysis in R Numerical Summaries Chapter	Team	Active	Apr 25, 15:00 CDT	0	0	0%	View
Exploratory Data Analysis in R Case Study Chapter	Team	Active	May 2, 15:00 CDT	0	0	0%	View

- Why are we using it?
- How are we using it?
- What will you have to do?

2.6 Antagonization

A new credo.

> “Getting it right is crucial when people’s lives are affected.” -Jonathan Steinhart

Figure 4: Lego fencing (Source: Unsplash)

2.7 What's next?

- See schedule:
  - install R / Emacs IDE - may do this together
  - Entry quiz (by Tue 18 Jan) - you should get > 50%
- Watch online lecture on "Systems" (to be published)
- Online followup notes (`notes.org` in GitHub)
- See you Friday 14-Jan online!
- Hopefully Wednesday 19-Jan in class!

3 Installing R / Windows PATH - w1s2 (14-Jan)

3.1 Overview

HOW	WHAT
Practice	Install R from CRAN
	Set PATH environment variable
	Test R in terminal and GUI
	Install GNU Emacs + ESS (FAQ)
	Set PATH environment variable
	Test R in Emacs
	Set .emacs init file
	Create Org file
	Run R code blocks in an Org file

3.2 Objectives

- [X] Install R
- [X] Set PATH environment
- [X] Test R in terminal and GUI
- [ ] Install GNU Emacs
- [ ] Test R in Emacs

4 Installing and setting up GNU Emacs - w2s3 (19-Jan)

4.1 I'm back

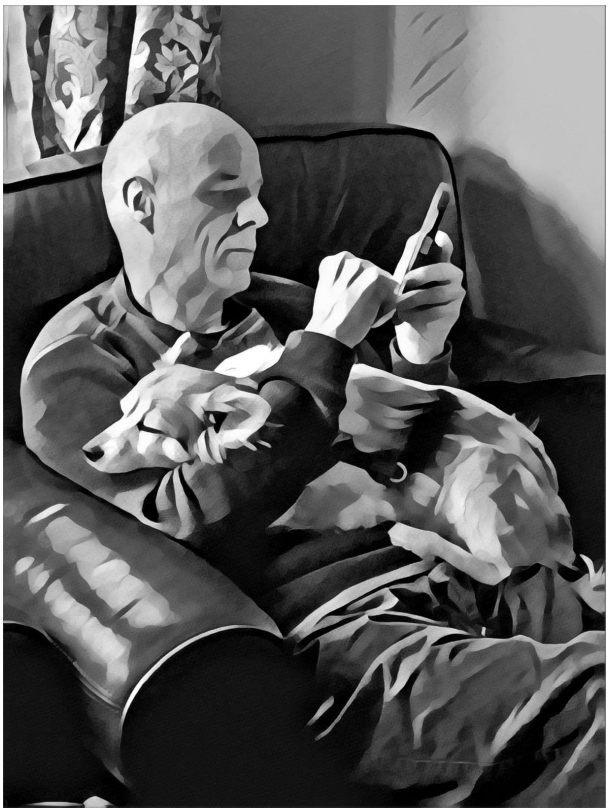


Figure 6: "I'm back, baby."

4.2 Overview

HOW	WHAT
Review	Entry quiz
	Quiz 1 + feedback + discussion
Practice	Install GNU Emacs + ESS (FAQ)
	Set PATH environment variable

HOW	WHAT
	Test R in Emacs
	Set .emacs init file

4.3 Objectives

- [X] Install GNU Emacs + ESS
- [X] Set PATH environment to run R in Emacs
- [X] Test R in Emacs (however, see course FAQ)
- [ ] Configure Emacs

4.4 Next

- Create Emacs Org file
- Run R code blocks in an Org file
- DataCamp assignments beginning soon!

5 Understand Emacs Org-mode - w2s4 (21-Jan)

5.1 Overview

HOW	WHAT
Lecture/Demo	GNU Emacs Org-mode
Practice	GNU Emacs Tutorial (gh)
Homework	Set `emacs` init file
	Create `.org` file
	Run code in an `.org` file

5.2 Objectives

- [X] Understand what Org-mode is and what it's for
- [ ] Create an `.emacs` init file for GNU Emacs
- [ ] Create an Org file
- [ ] Run a code block in your Org file

5.3 Next

- Create Emacs Org file
- Run R code blocks in an Org file
- DataCamp assignments beginning soon

6 Customizing Emacs (init file) - w3s5 (24-Jan)

6.1 Overview

HOW	WHAT
Review	Quiz 2
Lecture/Demo	GNU Emacs Org-mode (Part 2)
	New: video playlist
Practice	GNU Emacs Tutorial cont'd (gh)
- Package manager	`M-x package-list-packages RET`
- Start R shell in Emacs	`M-x R` (R must be installed & in the `PATH`)
- Add init file	.emacs sample file (GitHub)
Assignment[1]	Set `emacs` init file
Assignment	Read 2022 Data trends and predictions
	Put your summary thoughts in an `.org` file
	Check the FAQ "How should you read?"

6.2 Objectives

- `[X]` Create an `.emacs` init file for GNU Emacs
- `[ ]` Create an Org file
- `[ ]` Run am R code block in your Org file

6.3 Reading assignment

- Read "2022 Data trends and predictions" (DataCamp, 2022).
- Prepare for discussion in class:
  - Which quantitative and which qualitative predictions were made?
  - What do you think how valid these predictions are?
  - Put your thoughts in an Org-mode file (filename = `YourName.org`)
  - Upload your submission to assignment/2022_predictions on GitHub

To identify yourself, use the `#+AUTHOR:` option. You can see how

this works from the options in the header of this `README.org` file.

There is no upper or lower limit on the number of words. The main point is to create a proper Org-mode file.

6.4 Next

- Create Org-mode file with R code in it and run it
- Org-mode assignment
- DataCamp assignments beginning soon (due Jan 31)

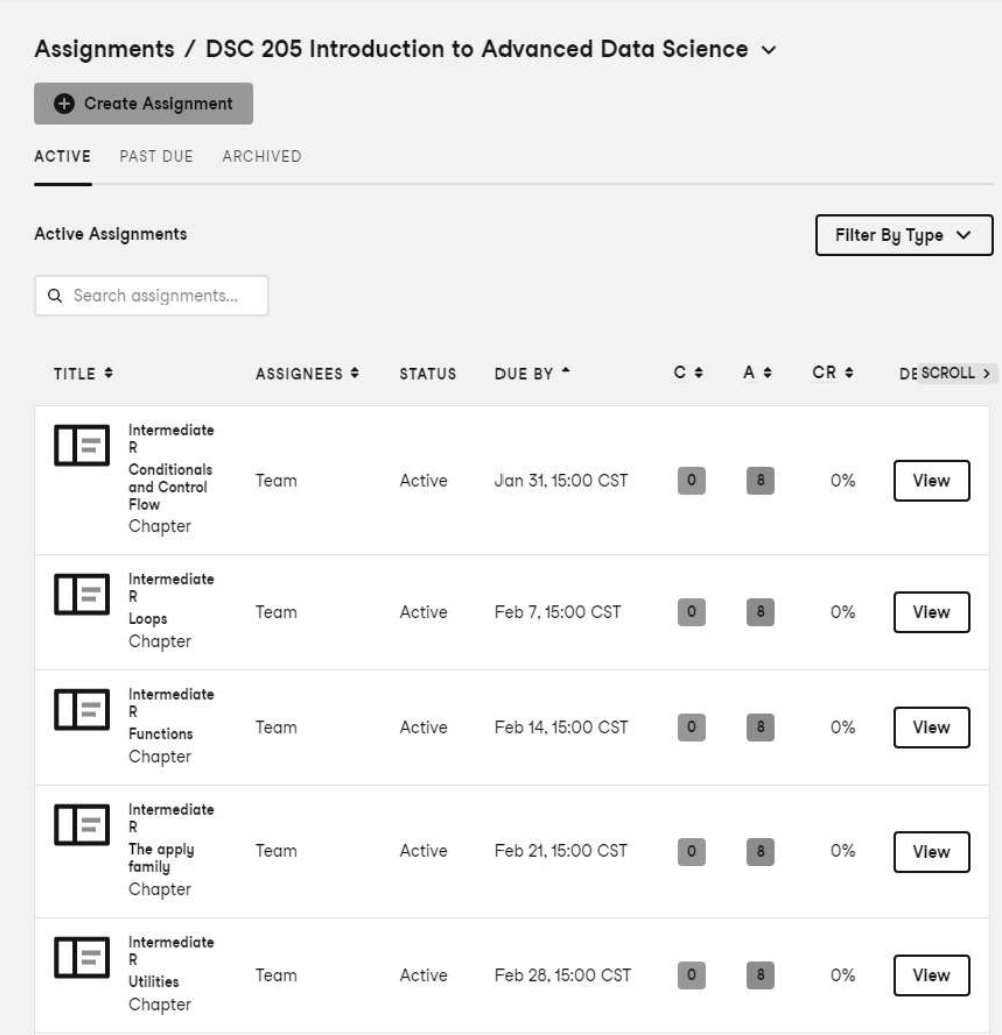


Figure 7: DataCamp assignments

7 Running code in Org-mode 1 - w3s6 (26-Jan)

7.1 Overview

HOW	WHAT	Link
Preview	DataCamp course "Intermediate R"	datacamp.com
Demo	Creating an Emacs Org-mode file with code and run it	README.org
Practice	Create Org-mode file with an R code block	

7.2 Objectives

- [X] Understand DataCamp assignment 1
- [X] Create an Org file
- [X] Run an R code block in your Org file

7.3 Next

- Submit Org-mode assignment in Schoology
- DataCamp assignments due Jan 31

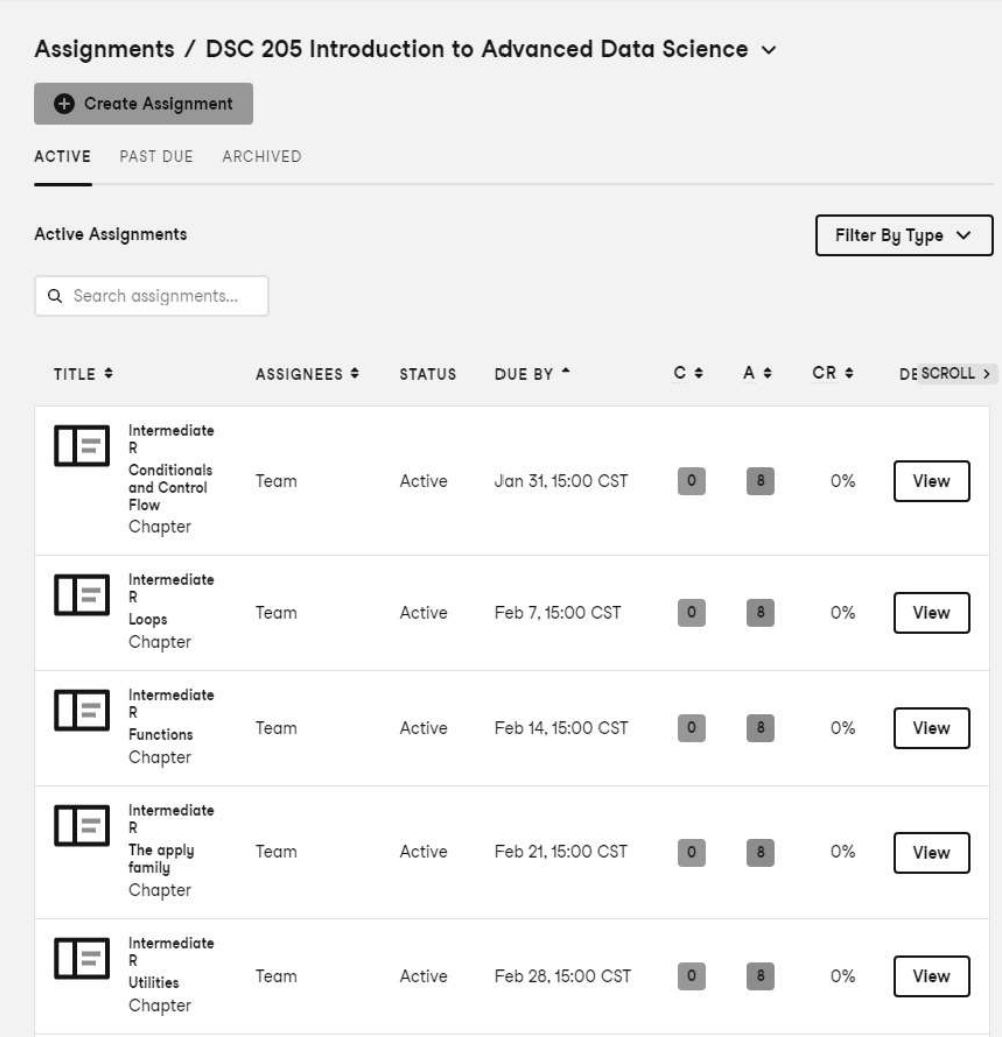


Figure 8: DataCamp assignments

8 Running code in Org-mode 2 - w3s7 (28-Jan)

1. We continue where we left it last Wednesday
2. Fixing the .emacs problem on Windows lab computers

3. Change of some deadlines - to finish basic Emacs training

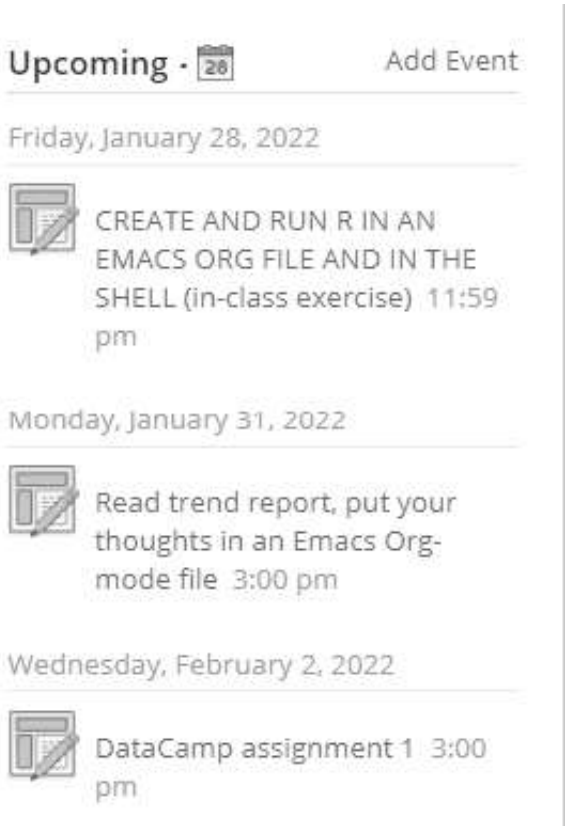


Figure 9: deadline changes in Schoology

4. Finish (expanded) Org-mode assignment
5. Submit results to Schoology.

9 Org-mode lab session - w4s8 (31-Jan)

Figure 10: Teaching Emacs on Dagobah

We will hold a special lab session tomorrow, Monday 31 January 3-3.50 PM, to sort out any issues related to Emacs and R. Bring your own PC to the session, or work on a lab desktop. I will spend the time going round to make sure that you can

- Install/ open / use the Emacs editor
- Create, run and tangle Org-mode files with R code
- Install / use the R programming language

- Understand the recent program assignments

  The necessary steps are also demonstrated in this tutorial video playlist.

We will continue with our regular program on Wednesday, 2nd February at 3 PM - a short quiz will be available before.

For those who know or can do all of this already: here's a second challenge (with solution) to practice while I sort others out.

9.1 What's next

- Deadline for 1st DataCamp assignment is looming (Wed 2 Feb 3pm)
- Scenario building for "Data Trends and Predictions 2022" report (assignment) - think about the 2 most important dimensions & watch this video about scenario planning
- Complete **quiz 3** including a **poll** on the prediction report before class
- Check out the webinar recording with DataCamp luminaries (panel)
- Use the breathing space to complete the Emacs tutorial (`C-h t`)

10 2022 Data Trends - w4s9 (2-Feb)

We meet today at 3-3.5- PM in the seminar room Lyon 106 - this room is directly adjacent to 104, our usual lab. We'll discuss the DataCamp 2022 trend report. The quiz will be available before end of the week. The planned first test (in class) will take place next Wednesday instead. (Schoology Update)

10.1 Overview

HOW	WHAT
Discussion	DataCamp 2022 report on Data Trends
Groupwork	Data science scenario planning (video)

10.2 Objectives

- [X] Understand the implications of the 2022 DataCamp trend report
- [X] Understand and apply the scenario planning technique

10.3 Next

- Quiz 3 - Conditionals and Control Workflow (DataCamp review)
- Test 1 (Friday 11 Feb 3 PM)
- Interactive R notebook - Writing functions

11 Studying with DataCamp - w5s10 (7-Feb)

11.1 Overview

HOW	WHAT
Review	Quiz 3 - Relational and logical operators
	How to study R with DataCamp
Preview	While and For Loops
Lecture	Writing functions in R
Test info	Test 1 on Friday 11 Feb 3.05-3.50 pm

11.2 Objectives

- [X] Review quiz 3 & how to study with DataCamp
- [X] Understand test conditions (Friday 11 Feb)
- [ ] Understand how to write functions in R (lecture)

11.3 Test 1 info

- Online in Schoology
- Entry quiz and Quiz 1-3 are not visible during the test
- The 10 hardest questions of entry quiy + quiz 1-3 (< 50%)
- 10 new questions
- Maximum time = 45 min

11.4 Next

- Interactive R notebook - loop problems
- Test 1 (Friday 11 Feb 3 PM)

12 Installing packages, using index vectors - w5s11 (9-Feb)

12.1 Overview

HOW	WHAT
Review	While and For loops
Lecture	Writing functions in R (part 1)

12.2 Objectives

- [X] Org-mode `PROPERTY` "shebang" stuff (meta data)
- [X] Review: install packages and loading datasets
- [X] Understanding and using index vectors

12.3 Next

- Test 1 (Friday 11 Feb 3-3.50 PM)
- Matthew Stewart, Stone Ward (Friday 18 Feb 3-3.50 PM)

13 Writing functions 1- w6s13 - (14-Feb)

13.1 News

- 2022 Data analytics competition (accounting data)
- Matthew Stewart, Stone Ward (Fri 18 Feb 3-3.50 PM) in Derby 209

13.2 Overview

HOW	WHAT
Class assignments	How do they work?
Practice Class assignments	Write a hello world function Installing loading packages .Rprofile configuration file
Review	Writing functions (DataCamp)
Interactive Lecture	Writing functions in R (part 2) Statistical functions in R

13.3 Objectives

- [X] Mark guest talk in your calendar (Fri 18-Feb) Derby 209
- [X] Understand how "class assignments" work
- [X] Complete a couple of class assignments
- [ ] Practice: install packages and loading datasets
- [ ] Review DataCamp chapter on writing functions

13.4 How do class assignments work?

- In-class assignments are **10%** of your total grade
- They are labeled **class assignments** in the Schoology gradebook
- You get the points if you attend and participate **actively**
- If you check your phone instead, you're **not** active
- If you could not attend (with a good excuse), submit **late**
- Submit an **Org-mode file**, not a screenshot

13.5 Next

- Wednesday: Review of test 1
- See some fun plotting techniques

14 Reviewing test 1, xkcd, plots - w6s14 (16-Feb)

14.1 News

- Eliminated some DataCamp assignments
- Remaining assignments mostly bi-weekly
- Emacs package of the week: `xkcd`

14.2 xkcd - life is too serious sometimes

- Package is pre-installed (list: `M-x package-list-packages`)
- `M-x xkcd` opens current comic
- `o` in `xkcd` mode opens browser with current topic
- `C-h ? m` opens full mode description

14.3 Overview

HOW	WHAT
Review	Hello function Test 1 - first month of class
How to make up for bad test results	Complete a mini-project

14.4 Objectives

- [ ] Review: Hello function
- [ ] Review: results of test 1
- [ ] Learn how to plot a density distribution and the mean
- [ ] Understand factor vectors
- [ ] Master Vector element extraction
- [ ] Understand the difference: Emacs Org-mode, ESS, and Base R
- [ ] Understand R comments
- [ ] Understand `NA`
- [ ] Understand the difference: object, storage class, data type
- [ ] Understand the help available in and outside of R
- [ ] Understand `print` and `paste`
- [ ] Understand vectorization
- [ ] Understand purpose and properties of interactive notebooks

14.5 CHALLENGE: Write a hello function with your name as an argument

- You already learnt how to write a `hello()` function without arguments. Write a function that takes your name as an argument and prints `"Hello, [your name]"`. Write and test the function in the same code block.

```
hello <- function(name) {
  print(paste("Hello, ", name))
}
hello(name="Marcus")
```

```
[1] "Hello,  Marcus"
```

- Another solution, this time with two arguments.

```
hello2 <- function(fname,lname) {
  print(paste("Hello, ", fname, lname,"!"))
}
hello2(fname="Marcus", lname="Birkenkrahe")
```

```
[1] "Hello,  Marcus Birkenkrahe !"
```

14.6 Lab 104 Emacs check

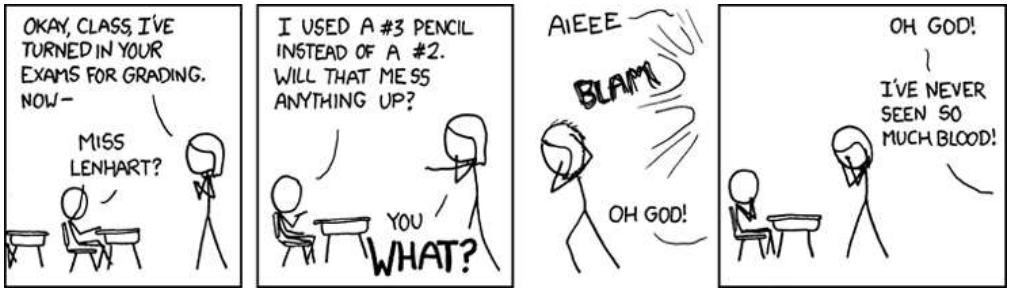
- First thing, when you sit down at your desktop in the computer lab, open Emacs, write a code block in an Org-mode file (`test.org`), and try to run it:

```
str(mtcars)
```

- If it does not work but instead complains about missing `org-babel` whatever, you need to install a `.emacs` file in the `$HOME` directory.
- Download the file or its content from `https://tinyurl.com/lyonemacs`. Make sure the file has the right name, then restart Emacs and run the code block again.
- You unfortunately need to do this any time you sit at a computer in the lab you have not sat at before.
- To make things easier, you could also put a `.emacs` file in your GDrive and download it in one go.

14.7 Test review

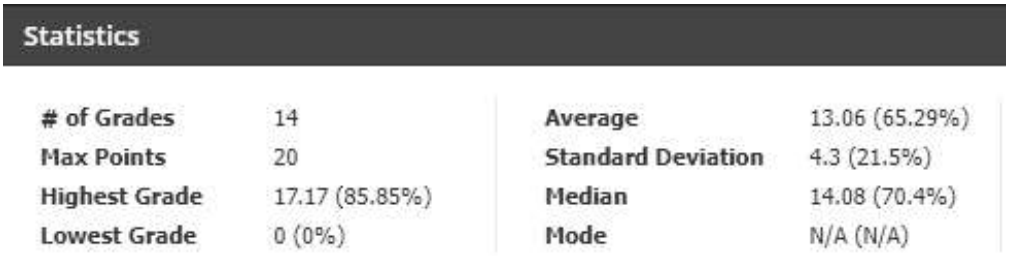
14.7.1 Paper vs Screen



Never again! Preparing such a test on paper and grading it while allowing for partial credit is a nightmare: future tests will be online in Schoology!

14.7.2 Test 1 results

- The test results are OK (average 70%). Better next time!



Statistics

# of Grades	14	**Average**	13.06 (65.29%)
Max Points	20	**Standard Deviation**	4.3 (21.5%)
Highest Grade	17.17 (85.85%)	**Median**	14.08 (70.4%)
Lowest Grade	0 (0%)	**Mode**	N/A (N/A)

Figure 12: Test 1 results (Schoology)

```
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)
```

- When checking the stats with R, I find different results. Why?[2]

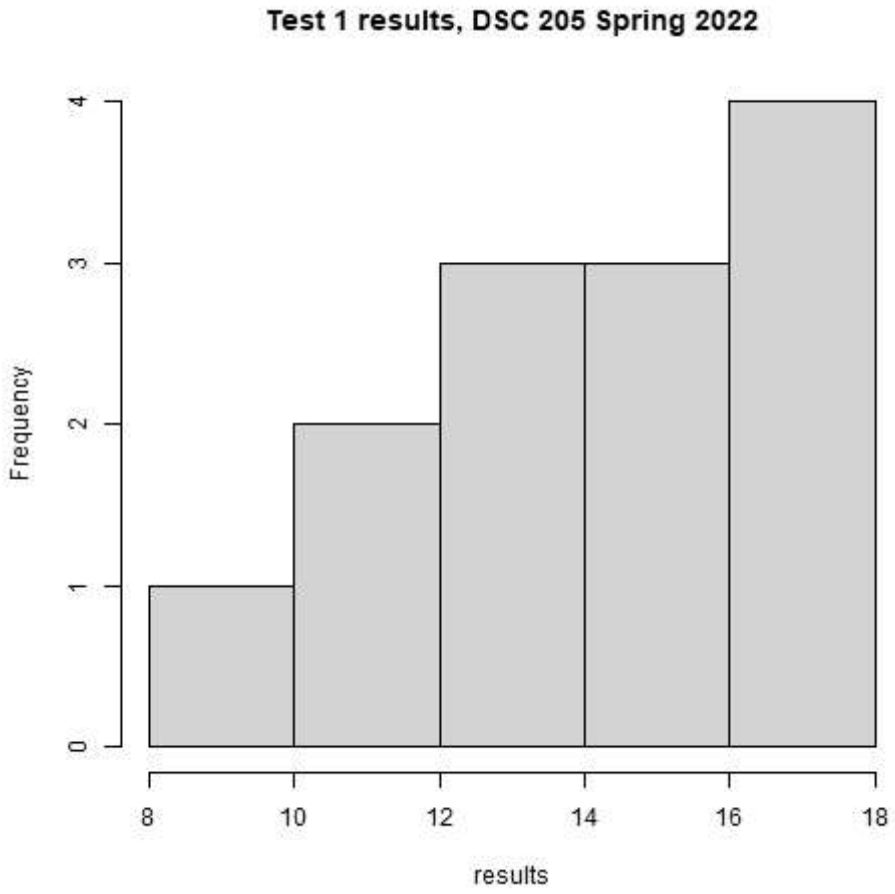
```
paste("Sample:",length(results))
paste("Standard deviation:", sd(results))
paste("Average:", 100*mean(results)/20)
summary(results)
```

```
[1] "Sample: 13"
[1] "Standard deviation: 2.59571120632991"
[1] "Average: 70.7"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.33   13.38   14.16   14.14   16.16   17.41
```

- Let's make some plots: histogram, boxplot and density plot.
- Fetch the vector from GitHub and run the code in Emacs.

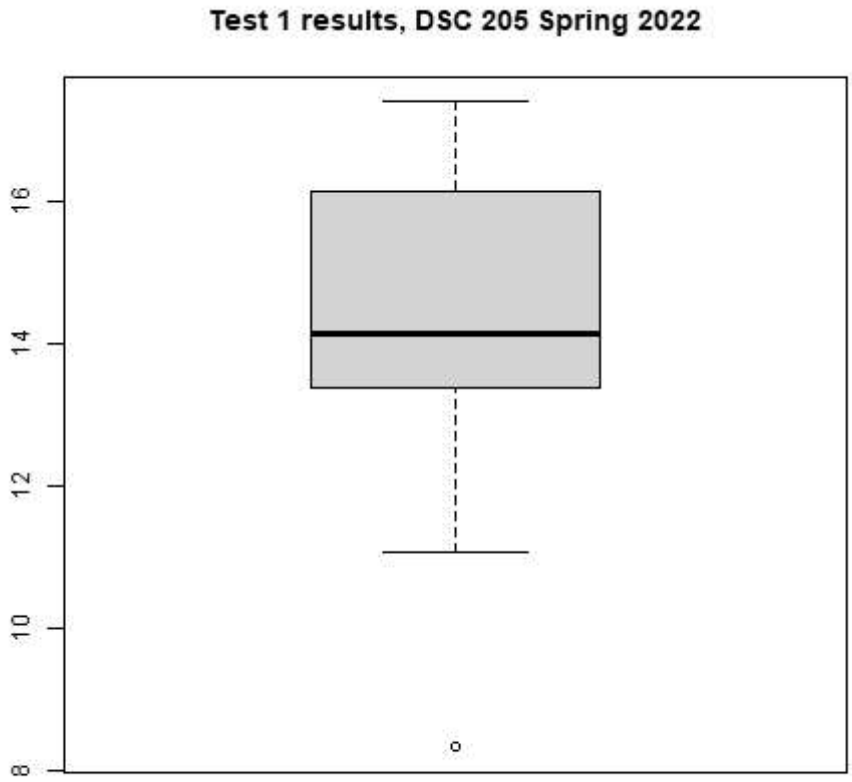
- Histogram. Demonstrates the fact that almost the entire course but one is above 50% (= pass). Looks more positive than the whole truth, because the x-axis ends with the maximum result achieved, and not with the maximum points available (20).

```
hist(results, main="Test 1 results, DSC 205 Spring 2022")
```



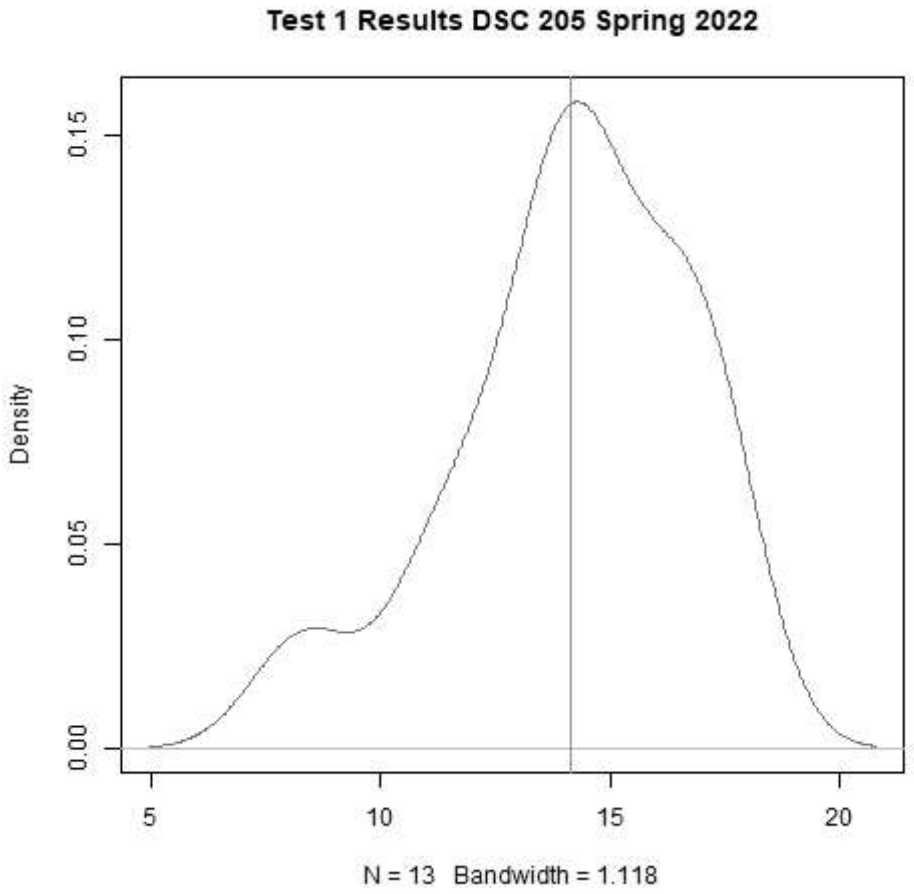
- Boxplot: this graph is deceptively positive, because it doesn't show the maximum points (20) but only the maximum achieved points. The "whiskers" correspond to the outliers, and the thick black line is the median (the middle value).

```
boxplot(results, main="Test 1 results, DSC 205 Spring 2022")
```



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

```
ave <- mean(results)
med <- median(results)
d <- density(results)
plot(d, col="steelblue",main="Test 1 Results DSC 205 Spring 2022")
abline(v=ave,col="red")
abline(v=med,col="green")
```



14.7.3 Analysis - feedback and action points

- Test 1 can now be played an unlimited number of times. I will add feedback to all new questions by the end of today.
- If you didn't play the other quizzes until you reached 100%, you had it coming. (My question: why wouldn't you do that?)
- What surprised me most was that many of you did not use the available time. However, I have not (yet) been able to correlate test time and test success (it's a project).
- Plots: I'd like the histogram and the density plot (a smoothened histogram) to peak more to the right, and for the boxplot to be smaller and higher up.
- See also: "I can teach it to you but I cannot learn it for you"
- Questions:
  - How did you study for this test?
  - If you didn't perform well, what will you change?
  - What can I do to help you help yourself?
- Changes to be applied in future quizzes/tests:
  - Fewer multiple choices (max. 4)
  - Announce if a question has > 1 answer (and/or how many)
  - Try to avoid having > 1 test on the same day

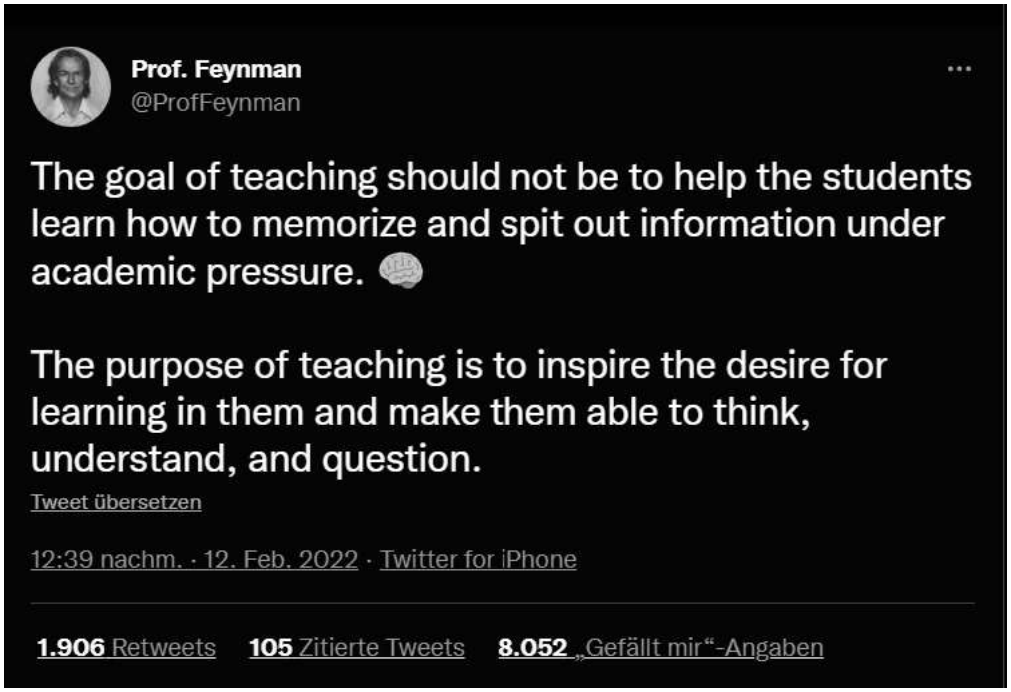


Figure 16: Feynman (via Twitter)

14.8 Next (topical)

- Writing R system functions
- Statistical functions
- Reading tables with `read.table`

15 Guest talk - Stone Ward - w6s15 (18-Feb)

15.1 Potential questions:

These are my questions informed e.g. by the 2022 data trends report.

1. What do your clients typically expect from you with regard to data science?
2. In the 2022 data trends report, we read that "upskilling [with data literacy skills] becomes a mandate". What is the level of data literacy (with examples) at Stone Ward? Where would you like it to be?
3. How well did your studies prepare you for what you're doing now as a data scientist?
4. What should undergraduates at Lyon know before they decide to embark on a potential career as data scientists or data analysts?
5. How important is machine learning in 2022 - and where is it going?
6. If you compare data science from an industry perspective 5 years ago, now, and 5 years from now - what's different?
7. What should students know before they approach you/Stone Ward for internships? What if they approach Stone Ward for a job?
8. What about a data science minor/major: important? Useful? Relevant?
9. Which projects would you like students to have attempted or completed? Is project experience important at all?
10. Which soft skills are most relevant at Stone Ward?

15.2 Presentation questions:

These are some of my questions after leafing through a pre-view of Matthew's presentation "Data in Business":

1. Why do clients want analysis? What do they do with the results? (Example)
2. Are clients typically more interested in descriptive (historic), prescriptive (normative) or predictive (future) analyses?
3. How much time do you still spend coding? Reading about R, new packages etc. How important do you think this is?
4. Tidyverse or base R?
5. How important is Excel to your work? How important is it to your clients still? (Compared to R or Python, or platforms like Tableau or Power BI)
6. What's with Plato's cave!?
7. Clients only remember "1-3 numbers" - which numbers are these (example)? How would I know what's important to them?
8. What if I screw up as a data analyst (example)?
9. How did you learn to talk about data and data science?
10. Do clients ever ask you for helicopter presentations like these, or only data analysis presentations (close to the result)?
11. What is a "non-data minded person"? (What are they missing?)
12. Who is on the analytics team?
13. Have you had interns or employees from Lyon College yet?
14. Can you tell us more about the scope of the problem or problems to be tackled in a mini-internship? How much does a student have to know?
15. How large are the data sets that you encounter at clients?[3]
16.

16 Guest talk - Post mortem - w7s16 (21-Feb)

16.1 News

- If you answered TRUE for question 18 on vectorization, contact me and you'll get an extra point for your test. My question was too confusing because the comparison could be seen as vectorization: check with `is.vector("hello")` - scalars and characters are internally represented as vectors, hence the simultaneous application of an operator (<) to all its elements could be called vectorization!

16.2 Objectives

Link to the slides

- [ ] Summarize talk (small group discussion + presentation)
- [ ] Identify what is most relevant to you
- [ ] Critically review claims and recommendations
- [ ] Apply the presentation to your own learning and career
- [ ] Learn more about the Google Data Analytics certificate
- [ ] Understand the problems with Excel
- [ ] Understand the difference between the "Tidyverse" and Base R
- [ ] Learn more about `ggplot2`

16.3 Overview

WHAT	HOW
Discussion in small groups	What did you think? Summarize main messages
Overview	Google Data Analytics Certificate Excel Errors Tidyverse vs. Base R The `ggplot2` package

16.4 What did you think of the talk?

- Summarize the main messages of the presentation?
- What were your personal takeaways?
- Is there anything you'd like to know from me?
- Do you want to have more presentations like this one?
- Are you interested in the mini-internships at all?

16.5 Next

- Writing functions
- Reading tables
- `apply` family of functions

17 Emacs recent files, Writing functions 2 - w8s17 (28-Feb)

17.1 News

- Updated schedule: `Rcpp`, `bash`, `Excel` and `SQLite`
- Interesting GitHub issues for "forward studying"
- The quizzes are hard(er): I uploaded PDF versions to GitHub
- Emacs package(s) of the week: `recentf` (and `ace-window`)

17.2 Objectives

- `[X]` Get started with interactive Emacs notebooks
- `[X]` Save and load user-defined functions
- `[X]` Practice writing functions (system, ~~stats~~)
- `[ ]` Understand stats functions in R
- `[X]` Learn a new Emacs package (or two)

17.3 Emacs package of the week: how to display recent files

- Package is actually built in. You call it with `M-x recentf-open-files`, which, on my Windows Emacs, leads to this buffer right now (the buffer below shows the buffer list - `C-x C-b`).

- It is nice that this also works when you kill Emacs by mistake. It is very handy to just be able to continue your work after you've opened and worked in 5-10 files!

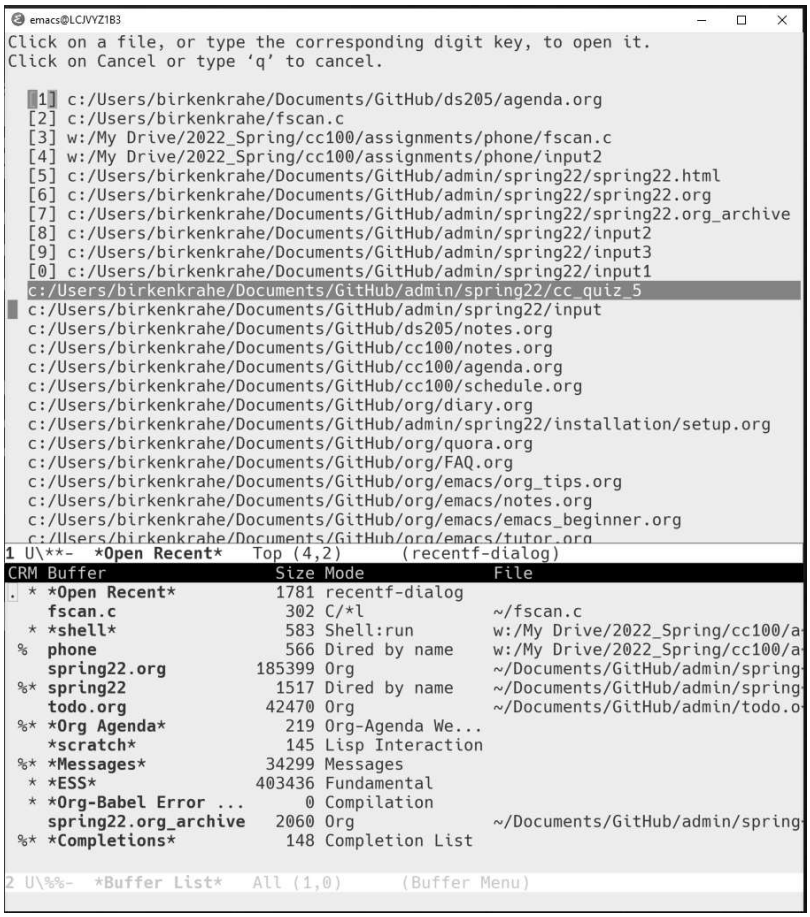


Figure 17: recent files on Emacs

- You can put the following code into your `.emacs` file to bind the command to `C-x r e` (or any other available combination you choose), and to enable it.

```
;; enable recentf mode and bind it to
(recentf-mode 1)
(global-set-key (kbd "C-x rf") 'recentf-open-files)
```

- More information:
  - Details on Emacs keybindings (Petersen, 2019)
  - Customizing Key Bindings: GNU Emacs manual (also available inside Emacs: `C-h i` opens the Emacs info reader)

17.4 Interactive notebook practice

- Download save_nb.org from GDrive and work through it in class.

17.5 Next

- Statistical functions (lecture)
- Tour of `apply` (notebook)
- `ggplot2` (lecture + notebook)
- Visualizing COVID-19 (project)

18 Change R download repo - w8s18 + 19 (4-Mar)

18.1 News

- [X] Mid-term grades - Improve your grade with a project (FAQ)
- [X] Waking up in `AppData/Roaming`? Change your Emacs HOME now.

18.2 Writing functions: set default R download repo

- Download `repos_nb.org` from GDrive and get cracking!
- You can find the solution in `repos_solution_nb.org` in GDrive.

18.3 Next

- Statistical functions
- Tour of `apply`
- `ggplot2` graphics
- Visualizing COVID-19 (DataCapmp project)

19 Graphical devices dev.list, .Library - w9s19 (7-Mar)

19.1 News

- [X] Waking up in `AppData/Roaming`? Change your Emacs HOME now.

19.2 Writing functions: set default R download repo

- Download `repos_nb.org` from GDrive and get cracking!
- You can find the solution in `repos_solution_nb.org` in GDrive.

19.3 Next

- Tour of `apply`
- `ggplot2` graphics
- Visualizing COVID-19 (DataCapmp project)

20 Gapminder, Pomodoro timer - Tour of ggplot - w9s20 (9-Mar)

20.1 Preparations for test 2 (Mon 14-Mar)

- Test 2 will only cover questions from quiz 4-6 + new questions.
- You can find quiz 4-6 with solutions + feedback as PDF (in /quiz)
- I will create an update of content Org files (in pdf/)

20.2 Emacs package of the week: Pomorodo timer

- Auto-complete `org-timer-`
- `org-set-timer`
- `org-timer-start`
- `org-timer-stop`

- Lines for `.emacs`:

```
;; pomodoro timer
(require 'org)
(setq org-clock-sound "c:/Users/birkenkrahe/Documents/Sounds/ding.mp3")
```

20.3 Multiple device control

Device control is quite important in all graphical program. In R, there are multiple functions that achieve this - look up `help(dev.list)` for the documentation, especially `example(dev.list)` (examples at the end of the help file).

Even though the 1 is supposedly Unix-specific, it runs under Windows, too. `x11` is the windows manager. In the example, a series of devices is opened, used, and finally closed again.

```
x11()
plot(1:10)
x11()
plot(rnorm(10))
dev.set(dev.prev())
abline(0, 1) # through the 1:10 points
dev.set(dev.next())
abline(h = 0, col = "gray") # for the residual plot
dev.set(dev.prev())
dev.off(); dev.off() #- close the two X devices
```

```
windows
      3
windows
      4
windows
      3
windows
      4
windows
      2
```

```
windows
      2
windows
      3
windows
      2
windows
      3
null device
          1
```

- Tour of `apply`
- Tour of statistical functions

20.4 ggplot2 and dplyr

- [ ] In winter 2020 (DSC101) I gave an introductory lecture to `ggplot2` using this Google Colaboratory workbook. You can copy the workbook to your GDrive and work through it if you like.
- [ ] This workbook came from an Org-mode file that I have uploaded in the practice directory in GDrive as `ggplot2_2021.zip` with all plots and images required to view and run it if you so wish. There is a also a 24-page PDF copy in the pdf directory.
- [ ] In this advanced introductory class, we will instead dive right into an interesting EDA example, the gapminder dataset, to explore the possibilities of ggplot2 and dplyr.
- [ ] I say it once at the outset (and hopefully not too many times). I am not a friend of the "Tidyverse" ideology. In practice, I prefer `data.table` over `dplyr`, and base R graphics over ggplot2. The reasons are manifold - see here for details.
- [ ] See also my notes on the "Tidyverse" and ggplot2 from the post mortem session of our guest talk by Matthew Stewart.

20.5 Next

- Quiz 4-6 check and test preparation
- `ggplot2` (lecture + notebook)
- Visualizing COVID-19 (project)

21 Test preparation / quiz 4-6 - w9s21 (11-Mar)

- [ ] You find the materials of the past month in GitHub (pdf/)
- [ ] Review of quiz 4-6
- [ ] Review of DataCamp chapters "apply" and and "utilities"
  - `lapply`, `sapply`, `vapply` and their uses
  - Useful mathematical functions: `abs`, `sum`, `mean`, `round`
  - Pattern matching: `grep`, `grepl`, `sub`, `gsub`
  - Time and data manipulation and calculation
- [ ] Best way to prepare: practice "Intermediate R" in the DataCamp app - my questions are likely to come from that source!

22 References

- Birkenkrahe (Jan 11, 2022). Interactive shell vs. interactive notebook (literate programming demo). URL: youtu.be/8HJGz3IYoHI.
- Cotton (Oct 25, 2018). How DataCamp Handles Course Quality [blog]. URL: www.datacamp.com.
- 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/
- Petersen (2019). Mastering Key Bindings in Emacs [website]. URL: masteringemacs.org.
- 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/.
- System Crafters (Aug 1, 2021). Emacs Has a Built-in Pomodoro Timer?? [video]. URL: youtu.be/JbHE819kVGQ.

Footnotes:

[1] Submission of the assignment by Monday 24 January 3pm gives 10 extra credit points.

[2] Somehow Schoology counts 14, not 13 participants.

[3] This was answered in the talk later. MS said that he had been asked to analyze time series data sets containing no more than 3 months of data. Depending on the number of observations, this could mean that the data set consists of 90 lines only, which is very small indeed.

Author: Marcus Birkenkrahe
Created: 2022-03-11 Fri 13:38
Validate