

Introduction to R

Bo Markussen
Department of Mathematical Sciences

Data Science Lab
<https://datalab.science.ku.dk>

What is Data Science Lab?

- ▶ Collaboration between Math and Computer Science.
- ▶ Funding from SCIENCE.
- ▶ Overall aim: Enhance quality of scientific data analyses in research carried out at SCIENCE.
- ▶ Activities: Consultancy meetings for students and researchers, research collaborations, workshops.
- ▶ More info and contact details: <https://datalab.science.ku.dk/english/>

Who are you? And what's the purpose of this event?

We hope that you

- ▶ are interested in tools for data exploration, data analysis, and data visualization,
- ▶ are open-minded and ready to play with R in a moment.

We don't expect that you have any experience with R.

Who are you? And what's the purpose of this event?

We hope that you

- ▶ are interested in tools for data exploration, data analysis, and data visualization,
- ▶ are open-minded and ready to play with R in a moment.

We don't expect that you have any experience with R.

By this course, we hope to

- ▶ get you started with R, so you can start using R “at home”,
- ▶ give you an impression about the possibilities with R.

What is R and RStudio?

What is R and RStudio?

- ▶ R is a software environment for statistics and graphics.
- ▶ RStudio is a user interface to R (but you can use R without it).

A lot of things has happened with R over the last decade, in particular regarding data handling and graphics. Buzzword: *Data Science* (but not so much AI).

Why use R?

- ▶ Easy to manipulate data (in the good sense) and explore data.
- ▶ All “standard statistical tools/methods” are available in R — and many, many non-standard ones, too.
- ▶ Graphics: Quite easy to make high quality figures as you want them.
- ▶ Programming language: All steps in your analysis may be saved in a script (reproducible research) — and you can implement your own methods.
- ▶ Simulation (making artificial data to examine your methods).
- ▶ Open-source, free, runs on all platforms.
- ▶ Large community, R-packages made available by other users.

How to access R

Two possibilities: (1) Locally on your own laptop, (2) in the “cloud”.

1. To use R on your own laptop:

- ▶ Install R from <https://www.r-project.org>
- ▶ Install RStudio from <https://posit.co/products/open-source/rstudio/>
- ▶ Use data and script files placed on your own laptop.
- ▶ For further info on how to install R see the *Welcome email* we sent to you.

2. Cloud solutions, for example

- ▶ The ERDA platform at UCPH, see <http://www.erda.dk>.
- ▶ UCloud from SDU eScience Center that support HPC computing in Denmark, see <https://docs.cloud.sdu.dk>. and

What is going to happen?

Mix of lecture-type presentations and hands-on exercises.

Tentative plan:

- ▶ 9.30–10.30: I. Working with R and RStudio.
- ▶ 10.30–12.00: II. Working with data in R.
- ▶ 12.00–12.30: Lunch break.
- ▶ 12.30–14.00: III. Graphics in R.
- ▶ 14.00–15.30: IV. Quarto and R markdown.
- ▶ Not done: V. Statistical analysis in R.

Course material

Available in a ZIP-file at

`https://datalab.science.ku.dk/english/course/r-intro/`

- ▶ These slides.
- ▶ R presentations as R scripts, and as Rmd, pdf and html.
- ▶ Exercise sheets as Rmd, pdf and html.
- ▶ Data for presentations and exercises (xlsx).
- ▶ Solutions for exercises as Rmd, pdf and html.

The exercises

- ▶ There are presumable far too many questions in the exercises! Not important that you get through everything.
- ▶ Do whatever you feel like, according to interests and experience.
- ▶ The exercises do not explain much statistics — this is a course on R rather than statistics.
- ▶ Often: “Try the following commands, and explain what happens”. You are encouraged to play!

Best: Work together in groups and discuss whatever comes up. Ask for help, clarifications, and discussions.

Material about R

- ▶ Many statistics textbooks use R (also online guides and videos):
 - ▶ Wickhan and Grolemund: *R for Data Science* (2nd ed). Available online at <https://r4ds.hadley.nz>.
 - ▶ Wickham: <https://ggplot2-book.org/>
 - ▶ Venables and Ripley: *Modern Applied Statistics with S*.
- ▶ Help pages inside R and RStudio: `?'function name'`, `??'topic'`
- ▶ RStudio: Links to cheat sheets in the Help-menu.
- ▶ google: Search R `whatever-topic-you-need`.
- ▶ Chatbots.
- ▶ RStudio now comes with an internal AI: **Posit Assistant**.
 - ▶ There is a user fee, and I haven't yet tried this.

I. Working with R and RStudio

- ▶ Make folders on your computer for files associated to R work.
- ▶ Work with R scripts (or Quarto scripts); then your work is saved for later use.
- ▶ Save your files often. It happens rarely, but R may crash if you ask it to do too weird things.
- ▶ Write comments to yourself about the code.
- ▶ Default RStudio set-up has four windows:

Upper-left: Editor for R scripts (and other programs).

Lower-left: Console for executing R code.

Upper-right: Show data objects in R session (also history of executed code).

Lower-right: Files, graphical output, package manager, help pages, ...

II. Working with data in R

- ▶ Many (newer) tools for working with data: selection of sub-datasets, transformation, sorting, ...
- ▶ It is important to distinguish between data files (often xlsx), R datasets, variables, observations.
- ▶ Datasets (either 'data frame' or 'tibble' inside R):
 - ▶ Rows are called observations (or datalines).
 - ▶ Columns are called variables.
- ▶ Variables can be extracted from datasets, but it is better to work on them **in** the dataset.
- ▶ Data types (categorical/numerical/ordinal/...) are important.

III. Graphical systems available in R

- ▶ Base R:
 - ▶ Pen on paper system.
 - ▶ Functions: `plot()`, `hist()`, `boxplot()`, ...
 - ▶ Produces nice graphics usable for scientific publications.
- ▶ `ggplot2`:
 - ▶ Add-on package.
 - ▶ Syntactical description of plots (*Grammar of Graphics*, Wilkinson 1999).
 - ▶ Extended and implemented by Wickham (2010).
 - ▶ Produces even nicer graphics — probably the best graphics in town.
- ▶ Other add-on packages:
 - ▶ Previous to `ggplot2` the `lattice` package was a good choice, and it is still popular (but not recommendable any more).
 - ▶ Extensions to `ggplot2`: `ggfortify`, `gridExtra`, `patchwork`, ...

Basic structure

- ▶ An **empty plot**

```
ggplot()
```

- ▶ with **data** and **aesthetics**

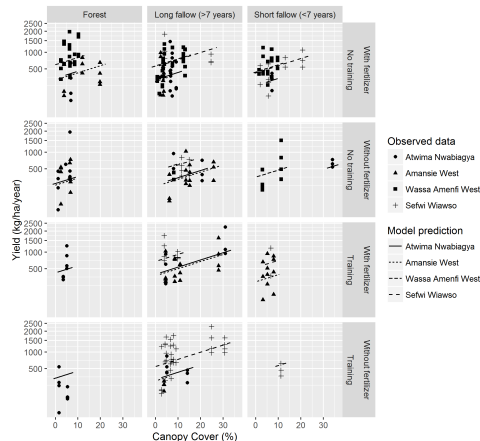
```
data=..., mapping=aes(...)
```

- ▶ is extended by **adding**

+

- ▶ **graphical objects**, e.g. points

```
geom_point()
```



Example gallery: <https://r-graph-gallery.com/>

IV. Quarto and R Markdown

Quarto is a file format for making dynamic documents.

- ▶ Input: Text and R code (in so-called chunks)
- ▶ Output, as html/docx/pdf: Formatted text, R code, R output, graphs

Benefits:

- ▶ Higher degree of reproducibility of research because “everything” is collected in one place.
- ▶ Nice for communication, also with people who don't use R.
- ▶ Easy to control the content of the final product, fx. should code and/or output be included?
- ▶ Possibly using a “visual” editor (WYSIWYG).

V. Statistical analyses in R

- ▶ All standard analyses are built into R. Workhorses for this

`lm()`, `glm()`, `lmer()`, `glmer()`, `nls()`, ...

- ▶ A model is fitted and saved to an object. Information from this object may be extracted or processed further, e.g. by

`print()`, `summary()`, `anova()`, `confint()`, `predict()`,
`plot()`, `drop1()`, `update()`, `step()`, ...

- ▶ `emmeans`-package provides estimated marginal means (a powerful method of presenting the results of a model) via

`emmeans()`, `pairs()`, ...