Applied Data science using R

Prof. Dr. Claudius Gräbner-Radkowitsch

International Institute of Management and Economic Education, Europa University Flensburg
Institute for the Comprehensive Analysis of the Economy, Johannes Kepler University, Linz
ZOE. Institute for future-fit economies, Cologne
Web: https://claudius-graebner.com | Email: claudius.graebner@uni-flensburg.de

Version 1.3 (24.04.2023)

Overall goal of the course

You will be introduced to the statistical programming language R and acquire practical knowledge about the fundamental tools of data science and machine learning. At the end of the course, you will be able to perform all essential steps of a quantitative data analysis in R yourself. This includes (i) data acquisition and preparation, (ii) visualization of the data on a publication-ready level, (iii) analysis of the data (with a focus on regression analysis), and (iv) communicating the results via visually appealing and reproducible reports.

The course does not require you to have any prior knowledge in R or any other programming language. Depending on your prior knowledge or affinity to programming, the course will be quite demanding, but equip you with computational skills that are most valuable both within academia and the business world. Moreover, please note that the course stresses collaborative and cooperative work, so we will work in teams a lot and support each other when tackling the challenge of learning a new programming language to the best extent possible.

Learning goals

At the end of the course, students will have acquired the following competences:

- Use R together with the integrated development environment R Studio
- Understand the use of R packages to perform specific data analytic tasks
- Write reproducible data analysis reports using Quarto/R Markdown
- Transform raw data as typically obtained into tidy data, which is suitable for further analysis
- Choose and justify the correct visualization approach, and create appealing visualizations using the R package ggplot2
- Implement and interpret linear regression models with numerical and categorial variables

Why R?

R is – together with Python and Julia – the *lingua franca* of data scientists all over the world. It is open source and free to use and runs on all operating systems. The community of R users is large, growing, and extremely amicable. Despite being a language specialized on data science, R is among the most widely used programming languages, and jobs for R programmers abound (and are comparatively well paid). In a nutshell: R is an indispensable part of the growing field of data science, and it is among the most widely used and influential tools for data preparation, visualization, and analysis.

Europa-Universität Flensburg
International Institute of Management and Economic Education
Department of Pluralist Economics

Structure and basic philosophy

The course prioritizes computational implementation over mathematical derivation, which is why I will focus on intuition and implementation, and often put mathematical derivations and proofs into optional further readings that are not part of the core course.

The course material is mainly hosted via a course homepage. Here you find links to the relevant material. Moreover, there is a Moodle room, which is important because it is where you can submit tasks, where the relevant announcements are made, and where you can find a Moodle forum, in which I invite you to pose your questions. This way, everybody can benefit from the answers. Note that I do not answer questions via email for this reason.

I will provide you with online exercises that you can do at home on your own. These exercises feature automated feedback. Below you see a tentative schedule where I added the activities associated with each session. Of course, you are quite free in how to handle the asymmetric learning activities (you should only have watched the videos until the date they are mentioned).

Tentative outline

Please keep in mind that this schedule will be subject to regular adjustments during the course. The respective announcements will be made via Moodle. The mandatory and optional readings for each session are provided on the course homepage.

#	Date	Day	Topic
1	15.03.23	Wed	General introduction and installation
2	16.03.23	Thu	Introducing the basics of R and R Studio
3	23.03.23	Thu	Basic object types in R (learning video)
4	29.03.23	Wed	Advanced object types in R
5	30.03.23	Thu	Recap and practice
6	06.04.23	Thu	Data visualization
			11.04.23 – 13.03.23: Reading period
7	20.04.23	Thu	Project Management and data import
8	26.04.23	Wed	Video lecture on data wrangling I
9	27.04.23	Thu	Video lecture on data wrangling II
10	04.05.23	Thu	Explorative data analysis in practice
11	10.05.23	Wed	Introducing Quarto and R Markdown
12	11.05.23	Thu	Recap and practice
NN	18.05.23	Thu	No lecture due to Christi Himmelfahrt
NN	**24.05.23**	**Wed**	**Mid-term exam (60 minutes)**
13	25.05.23	Thu	Introduction to data analytics
14	01.06.23	Thu	Sampling theory
15	07.06.23	Wed	Simple linear regression
16	08.06.23	Thus	Multiple linear regression
17	15.06.23	Thu	Selected machine learning tools I
18	21.06.23	Wed	Selected machine learning tools II
19	22.06.23	Thu	Recap and practice
NN	**28.06.23**	**Mon**	**Final exam; 10:00 – 11:30, MAD 130**

Timing overview

In the following I outline how the expected workload is allocated among the different course activities:

Activity	**Frequency · on-site hours**
On-site lectures	$19 \cdot 1{,}5 = 28{,}5$ hours
Expected self-study and group work time (including videos)	121,5 hours
Total workload	**150 hours**

Note that I strongly recommend to allocate self-study time rather equally over the semester. It will save you time and effort if you learn and practice constantly, not only shortly before the exams.

Expected contributions of the students and software used

You need to install R, R Studio and Git on your personal laptop. We will reserve one session to work on problems that may occur during the installation process, but it is vital that you give your best to install these programs yourself as soon as possible. You will also need to sign up and create an account at GitHub and use the service of Netlify (which you can use via your GitHub account). During each session we will use a collaborative online pad/chat platform that allows you to summarize the key messages for yourself and pose questions that will be answered by myself after the session. To this end we will most likely use Jitter. Please note that using these tools is mandatory.

I will provide you with practical exercises. While not being mandatory, I strongly encourage you to work through all exercises, and also to attend the exercise sessions. The same is true for the online tutorials you are asked to complete between some of the sessions: while I do not test the completion of the tutorials immediately, you will run into trouble later if you do not complete them in due course.

I encourage you to complete the exercises and tutorials in teams since teamwork is (a) more fun, (b) a more realistic preparation for your later work, and (c) more insightful since you learn from each other. I also expect that we help each other in our learning processes as much as possible: learning a programming language is a community effort. So please post your problems and questions in the Moodle forum and try to help others wherever you can.

Evaluation

The overall grade for the lecture will depend on:

- A mid-term exam on your computer, to be written in the University (50%)
- A final exam on your computer, to be written in the University (50%)

Literature and course material

All course materials will be provided via the course homepage (which has been developed completely in R). The link will be published in due course.

Note that all communication as well as the grading of course assignments will take place only via Moodle (course number: 11973; password: DataAnalysis23):

https://elearning.uni-flensburg.de/moodle/course/view.php?id=11973

During the course we will refer to a number of textbooks, all of which are available online for free. Our main references will be:

Wickham, H. & Grolemund, G. (2023): R for Data Science. Online: https://r4ds.hadley.nz/

Ismai, C. & Kim, A. (2021): Statistical Inference via Data Science. Online: https://moderndive.com/index.html

For more advanced details on the fundamentals of programming in R, I recommend the following book, which is also available online:

Wickham, H. (2019): Advanced R. Online: https://adv-r.hadley.nz/

For the model-related parts of the lecture I recommend the following book:

James, G., Witten, D., Hastie, T., Tibshirani, R. (2021): An Introduction to Statistical Learning with Applications in R, 2nd ed.
Book page: https://www.statlearning.com/
Online PDF: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Further (optional) references will be provided in due course.