Data science with 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 1.0 (11.03.2022)

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. 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 using both traditional statistics such as regression analysis, as well as modern tools from the field of machine learning, and (iv) communicating the results via visually appealing and reproducible reports. These technical insights will be complemented by some fundamental aspects from the philosophy of science such that you are able to reflect your applied work in an adequate manner.

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.

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.

Structure and basic philosophy

The course comprises aspects that resemble a classical lecture and that introduce the theoretical foundations for the various concepts covered, as well as aspects that resemble a lab session in which we will implement these concepts in R hands on. In practice, both parts will blend into each other regularly, which is why it is essential that all students bring their own laptops to all sessions: the course focuses of *applied data science*, which is why we will apply new concepts together as quickly as possible.

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.

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 in the second lecture week 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.

There will be practical exercises provided through Moodle during the semester. While not being mandatory, I strongly encourage you to work through all exercises. 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): https://datascience-euf-spring22.netlify.app/

Note that all communication as well as the grading of course assignments will take place only via Moodle (course number: 9652; password will be provided upon request).

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

> Wickham, H. & Grolemund, G. (2017): R for Data Science. Online: https://r4ds.had.co.nz/index.html

In several instances, we will refer to chapters in the following two books:

> *Hanck, C., Arnold, M., Gerber, A. & Schmelzer, M. (2021): Introduction to Econometrics with R. Online:* https://www.econometrics-with-r.org/

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/

Throughout the philosophical sessions we regularly refer to the following book, which is not free but is available in the library:

Shrader-Frechette, K. (2014): *Tainted. How Philosophy of Science Can Expose Bad Science*, New York, NY: Oxford University Press.

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

Tentative outline

There are two dates for the course: the one on Thursday 10am-12am will take place every week, the one on Wednesday 12am-2pm every two weeks. Please refer to the tentative schedule below for more details and 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	17.03.22	Thu	General introduction and remarks on how to install the relevant software
2	23.03.22	Wed	Philosophy of science I and resolving installation problems
3	24.03.22	Thu	Introducing the basics of R and R Studio
4	31.03.22	Thu	Data types in R
5	06.04.22	Wed	Data visualization I
6	07.04.22	Thu	Project Management and data import
7	20.04.22	Wed	Data wrangling I
	21.04.22	Thu	No Session
8	28.04.22	Thu	Introducing R Markdown
9	04.05.22	Wed	Advanced programming tools
10	05.05.22	Thu	Review of statistics: probability and inference
11	12.05.22	Thu	Mid-term exam
12	18.05.22	Wed	Models of data I: introduction & philosophical foundation
13	19.05.22	Thu	Models of data II: simple linear regression
14	01.06.22	Wed	Models of data III: multivariate regression
15	02.06.22	Thu	Models of data IV: regression diagnostics
16	09.06.22	Thu	Philosophy of science III: paradigms and the theory-laddenness of observation
17	15.06.22	Wed	Data wrangling II
18	16.06.22	Thu	Data visualization II
19	23.06.22	Thu	General linear models
20	29.06.22	Wed	Philosophy of science IV: models, verification, and validation Supervised machine learning I
21	30.06.22	Thu	Supervised machine learning II
22	06.07.22	Wed	Unsupervised machine learning
23	07.07.22	Thu	Summary, open questions, and outlook