Chair of Computational Molecular Medicine
Technical University of Munich

Data Analysis and Visualization in R (IN2339)

A practical introduction to Data Science

Contents

List of Figures

List of Tables

Preface

This is the lecture script of the module Data Analysis and Visualization in R (IN2339).

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Acknowledgments

This script has been first put together in the winter semester 2020/2021 by Felix Brechtmann, Alexander Karollus, Daniela Klaproth-Andrade, Pedro Silva, and Julien Gagneur with help from Xueqi Cao, Laura Martens, Ines Scheller, Vangelis Theodorakis, and Vicente Yépez.

We leveraged work from colleagues who helped creating lecture slides since 2017: Žiga Avsec, Ines Assum, Daniel Bader, Jun Cheng, Bašak Eraslan, Mathias Heinig, Jan Krumsieck, Christian Mertes, and Georg Stricker.

Prerequisites

Basics in probabilities are required. Chapters 13-15 ("Introduction to Statistics with R", "Probability" and "Random variables") of the Book "Introduction to Data Science" https://rafalab.github.io/dsbook/ make a good refresher. Make sure all concepts are familiar to you. Check your knowledge by trying the exercises.

Feedback

For improvement suggestions, reporting errors and typos, please use the online document here.

Introduction

Data Science: What and why?

Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data. The goals of Data Science include discovering new phenomena or trends from data, enabling decisions based on facts derived from data, and communicating findings from data. It is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics.

Data Science is at the heart of the scientific method, which starts with making data-driven observations to formulate testable hypotheses. It furthermore comes into play to visualize and assess experimental results. Data science skills are therefore necessary to any field of scientific research. Data science is the main tools of epidemiology, the study of health and disease in populations, which largely relies on observational data. Moreover data science is important in the industry, to understand operational process, and in business analytics, to understand a particular market. Hence, with the rise of big data in all areas of society, data science skills are some of the most demanded skills on the job market. Last, but not least, in an era of fake news, data science skills are important for citizens of modern societies.

What you will learn and not learn

The goal of this course is to provide you with general analytic techniques to extract knowledge, patterns, and connections between samples and variables, and how to communicate these in an intuitive and clear way.

This course focuses on front-end data science. This means, it teaches practical skills to analyse data. We will focus on tidy data, visualizations, and data manipulation in R. To only then dive into the math required to understand and interpret analysis results.

This course does not teach back-end data science, i.e. it does not teach how to develop your own statistical or machine learning models, nor how to develop scalable data processing software.

Other courses offered by the faculty of Informatics cover data science back-end skills.

The R language

R is a statistical programming language designed for data analytics. It is a great language for front-end data science, i.e. to rapidly manipulate, visualize and come to raising interesting hypotheses.

Seen from a software developer point of view (i.e. from a back-end data science point of view), R can be seen cumbersome and not using memory and computing resources efficiently. The purpose of the R language is to reduce time spent in coding to maximize user's brain time on looking at the data and thinking about it, rather than reducing computer's running time. Of course, there are ways to develop efficient R software, notably by relying on implementation in lower languages such as C. This course does not cover such R developer skills.

Another advantage of R is that it offers a very large set of libraries from many application areas.

Course overview

The lecture is structured into three main parts covering the major steps of data analysis:

1. **Get the data**: After basic introduction to R, learn how to fetch and manipulate real-world datasets. How to structure them to most conveniently work with them (tidy data).
2. **Look at the data**: Basic and advanced visualization techniques allows navigating large and complex datasets, identifying interesting signal, and formulating hypotheses. Typical sources of confounding are discussed. Recommendation to present an analysis in compelling fashion are also given.
3. **Conclude**: Concepts of hypothesis testing will allow concluding about the statistical robustness of discovered associations. Also, methods from supervised learning will allow to model data and build accurate predictors.

The chapters of this script corresponds to individual lectures. Appendices provide further technical details as well as R tricks and tips.

Complementary reading

These books offer complementary information to this script:

- Introduction to Data Science, Rafael A. Irizarry [https://rafalab.github.io/dsbook/]
- R for Data Science, Garrett Grolemund and Hadley Wickham [https://r4ds.had.co.nz/]
- Statistical Inference via Data Science, Chester Ismay and Albert Y. Kim [https://moderndive.com/]
- Fundamentals of Data Visualization, Claus O. Wilke [https://clauswilke.com/dataviz/]
- Advanced R, Hadley Wickham [https://adv-r.hadley.nz/]