STA 141A – Fundamentals of Statistical Data Science

Department of Statistics; University of California, Davis

Instructor: Dr. Akira Horiguchi (ahoriguchi@ucdavis.edu)
A01 TA: Zhentao Li (ztlli@ucdavis.edu)
A02 TA: Zijie Tian (zijtian@ucdavis.edu)
A03 TA: Lingyou Pang (lyopang@ucdavis.edu)

Section 1: About this course and R

Spring 2025 (Mar 31 – Jun 05), MWF, 01:10 PM – 02:00 PM, Young 198

Overview

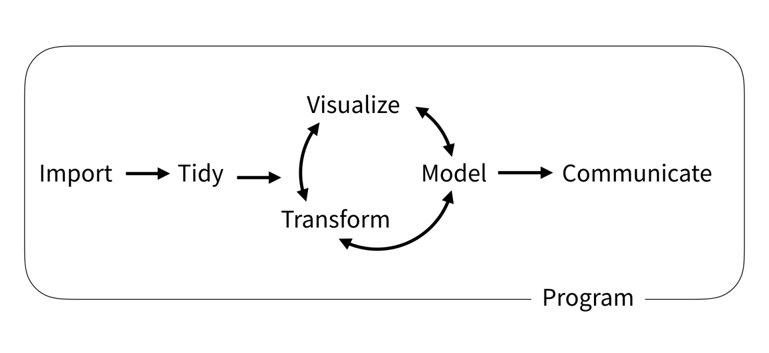
Section 1: About this course and R

- **What is the course about? – The model of data science**
- **Programming**
- **About R**

About this course and R

What is the course about? – The model of data science

What is the course about? – The model of data science



Wickham and Grolemund (2017)

⇒ We want to do this with R !!!

About this course and R

Programming

General comments

- Surrounding all the mentioned tools is programming.
- Being an expert programmer or data scientist is not needed. However, learning more about programming pays off since it allows to automate and simplify common tasks

In this class we are going to learn...

- ... general programming concepts;
- ... visualize our results;
- ... statistical programming, computation techniques for data analysis/statistics purposes.

Key high-level programming concepts

- Data Objects (vectors, arrays, matrices, lists, data frames, etc.)
- Operations (vector arithmetic, selecting and modifying, element-wise operations, matrix multiplication, matrix decompositions, etc.)
- Control statements (conditional execution, repetitive execution, etc.)
- Functions (built-in functions, writing own functions)
- Data manipulation (how to manipulate/transform data frame objects)
- Data visualization

About this course and R

About R

R and RStudio

- R: is a programming language and software for statistical computing and graphics. It provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques.
- RStudio is an Integrated Development Environment (IDE), i.e. an application that enables programmers to consolidate the different aspects of writing a computer program by combining common activities of writing software into a single application: editing source code by syntax highlighting and autocomplete, debugging.

By the end of the week, please make sure you can run R/RStudio from a machine you have access to.

Why R?

- Easy to learn and to use.
- R can be used to generate graphics based on complex data sets very quickly.
- Very popular and one of the standard languages for statistics, data science, computational biology, finance, industry, etc.
- New technology and ideas often appear first in R.
- Supported by a vast community that maintains and updates R.
- A lot of high quality packages.
- Free and open-source.
- Runs on basically any platform.

What is the effect of learning R?

- By programming in R, you will learn general concepts of high-level programming and languages.
- Since R is a complete programming language, learning it allows you to transfer the concepts to other languages.
- Syntax and available libraries may differ between languages, but how you approach a computational task and reason about the computations is similar.
- It enables you to learn another programming language much easier.

R libraries/packages

R is divided into:

- 1. The **base R system**
  - This contains, among other things, the base package which is required to run R, and the most fundamental functions.
  - The 'base' system contains also some other packages.
- 2. In about **20,000 libraries** (or packages) that you can install and use:
  - CRAN [1] 'contributed' packages (or sometimes in BioConductor project or in Github repositories).
  - These already do pretty much anything you have in mind (data manipulation, advanced visualizations, machine learning models, etc.).

[1] The Comprehensive R Archive Network

RStudio

- RStudio is an Integrated Development Environment (IDE) for R.
- It makes it easier to interface yourself with R, and comes with extra functionalities.
- While R is a community open-source project, RStudio is a business (for profit) which offers some of its products for free.

Homework due in two days

Now you have enough to do homework zero (due this Wednesday at 9pm!)

- Purpose: ensure you know how to use R Markdown or Quarto
- Check Canvas for assignment
- If you already have used RMD/Quarto, should take you $<$ 10 minutes
- Otherwise, use this week's discussion period