Introduction to R and the tidyverse

Paolo Crosetto
Lecture 0: getting started
The course

This course will be a 12 hour brief introduction to R, from the point of view of the tidyverse, a set of packages developed by Hadley Wickham.

We will be dealing with the basics of data handling:

- data input & output
- data wrangling, selecting; filtering, mutating (with dplyr)
- data reshaping and merging (with tidyr)
- data plotting (with ggplot)

Slides will be in English for a simple reason: all R documentation is in English. Out there you will be needing help, and it is good to be taught in the language you’re going to get helped with.
Lecture plan

- **Lecture 1**: introduction to R, plotting
- **Lecture 2**: data input & output, data wrangling
- **Lecture 3**: reshaping, merging and tidying
- **Lecture 4**: applying it all: each student a dataset to explore & a question to answer. Dive in, find out.
Contacts

Paolo Crosetto

- paolo.crosetto@gmail.com
- Office 413, 4th floor, BATEG
- do not hesitate to pop by or send an email if needed
Tools of the trade

- the course is (loosely) based on the *R for Data Science* book. Free reading and code here
- we will use the (free and wonderful) editor Rstudio
- within Rstudio, we will use the (free and even more wonderful) packages from the tidyverse
- we will use freely downloadable data from several sources
Your R background

How many of you have already used

- Excel
- SAS, SPSS
- Stata
- Base R
- Tidyverse R
- Python, Jupyter, Julia, C, C++...
Why R?

Why bother with R, of all tools?

- it’s free (as in free beer and also as in free speech)
- it has a huge code base and an ever increasing user base
- it is a programming language built for doing statistics
- it forces you to understand what you are doing (hard at first, but...)
- it forces you to use scripts -> reproducible research
- and if all of this does not convince you, it is always better than Excel
Why not Excel? [a short rant]

Excel is really bad for serious work

- It has a click-based interface that is a source of horrible errors
- Reinhardt & Rogoff see here for more
- You can run macros, but then it gets awkward anyway
Whenever you do serious analysis, consider this

- in any analysis, even the basics, you make *hundreds* of operations on your data
- each of them can be a mistake, can contain one, can be forgotten, or done twice
- if you get back to the data after 1 month, will you remember the choices made?
- NO
- think how hard it can then be for *other people* to read your work
- take care of your future self: do reproducible work
Reproducible work

- You need to keep track of what you do. So:
- always use a script-based language
- experiment interactively, try and re-try, but then lock what you did in a script
- comment extensively
- or, if too lazy, use a code that is readable
Further resources

*Where to find more about *R*, the tidyverse, get help, etc.*…*

- The tidyverse website contains all the package documentation and galleries of examples
- The *R* for Data Science book covers all of the course and more
- There is a nice cheatsheet about ggplot here
- Another nice cheatsheet about data wrangling here
If you are lost and need to answer a question...

- try to input ? or ?? followed by the command/topic of interest in R
- try StackExchange: there is an R channel here
- (learn how to ask meaningful questions first, though...)}