

R4DS - Unit 1: Tidyverse



Andrea Gilardi
March 17, 2023



Outline and main concepts

- The objective of this course is to introduce a set of effective and modern tools for data science, version-control and R packages' development.
- At the beginning of this class, I will briefly mention a set of (**opinionated**) views that may improve your experience when developing R code.
- The examples are based on the Rstudio IDE (version 2022.7.2.576), but similar considerations hold for other Rstudio versions and different IDEs.



Outline and main concepts (cont)

- Then, we are going to briefly present the tidyverse and some of its most important packages via several examples.
- These practical examples we will based on a series of datasets shared by the Department for Transport:
<https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data>
- We are not going to review the basics of the R language, but if you have any question feel free to ask!

But first, my favourite analogy!



via boredpanda, bbc, reddit

Your taste develops faster than your ability.

Source: https://www.youtube.com/watch?v=7oyiPBjLAWY&t=448s&ab_channel=RConsortium



What They Forgot (WTF!!!)

- Always start R with a blank state!
- Adopt a project-oriented workflow.
- Practice safe “paths”.
- Work and share examples in a reproducible environment, a so-called `reprex` (see Unit 2).

The examples are taken from <https://rstats.wtf/>.
Another course on similar topics (not R-based and slightly more advanced): <https://missing.csail.mit.edu/>.

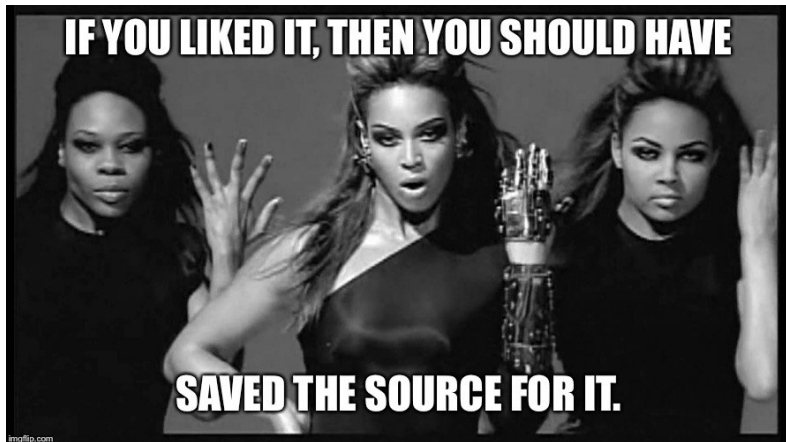


Always start R with a blank state!

- By default, when you terminate an R session, the software asks you if you want to save the current workspace.
- Similarly, the R startup mechanism¹ loads a saved image of the user workspace (i.e. a .Rdata file) if there is one.
- Unfortunately, this behaviour might be really dangerous, especially if you don't remember how the saved objects were generated. Citing the Python's style guide PEP20: *Explicit is better than implicit*.

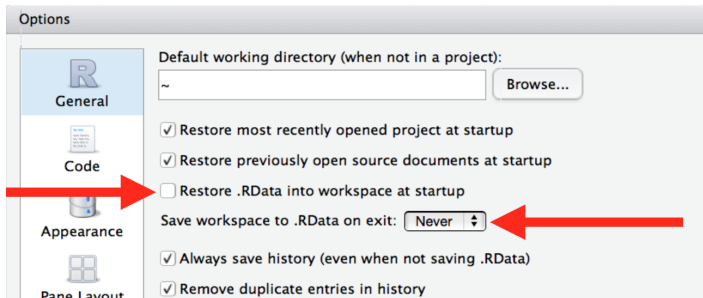
¹See also `?Startup` and `?quit` for more details.

or, as Beyoncé said ...



So what can I do?

- If you run R from the shell, use `R --no-save --no-restore-data`.
- In Rstudio, set the following options via Tools -> Global Options.





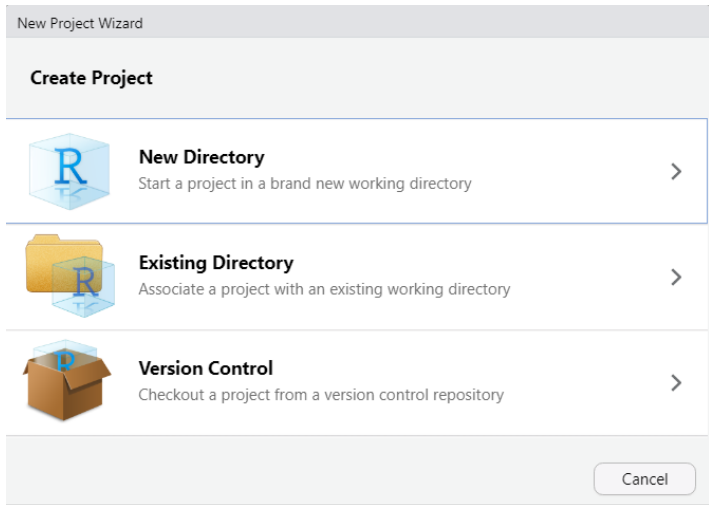
Adopt a project-oriented workflow

- There are many R scripts that begin with the following lines:
 - `rm(list = ls())` # "clean" the environment
 - `setwd("path/that/I/only/have")` # adjust the wd
- What is the problem with the previous code chunks?
 - `rm(list = ls())` is not enough to properly clean your R session. Let's see an example!
 - The previous `setwd(...)` command is (almost surely) going to fail for everyone who is not the original user 😞

⇒ Organise your analyses as independent (Rstudio) projects, each belonging to a separate folder. Let's try!




Adopt a project-oriented workflow





Practice safe “paths”

- Why do we usually run `setwd(...)`? Because we want to specify the “relative” path of a file according to a directory.
- Unfortunately, `setwd(...)` may raise several portability issues as seen before.
- The [here](#)  provides a convenient way to perform the same operation without manual interventions and the aforementioned drawbacks \implies Practice safe “paths”!
- Let's see an example and more details!



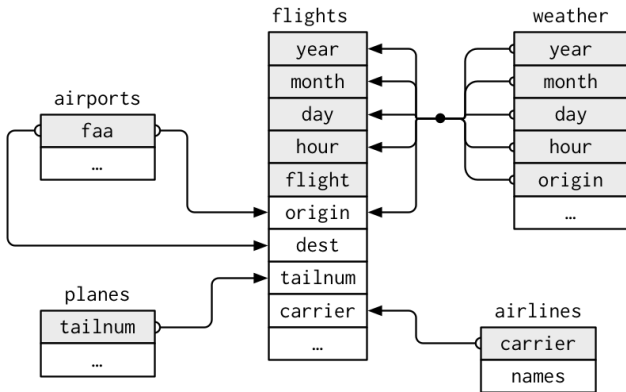
The tidyverse!

The tidyverse is a set of packages that work in harmony because they share common data representations and API design.

<https://tidyverse.tidyverse.org/>

EDA with the Tidyverse

We are going to briefly showcase the tidyverse toolkit using a series of relational dataset obtained from [here](https://r4ds.had.co.nz/relational-data.html).



Source: <https://r4ds.had.co.nz/relational-data.html>.



**Enough theory, let's start
coding 🤖**