A gentle introduction to data visualization in R for TAP Striking Statistics

March 2020

1 Goals of this tutorial

It is often easier and quicker to communicate coding concepts in an interactive in-person workshop. We are unlikely to have that opportunity soon. The idea of this tutorial is to share a full R worlflow of a striking stat we have already started discussing in Teams. I have also shared online guides that I believe give a good introduction to key topics. Hopefully, you would be able to go back and forth between the guides, which give a conceptual overview of R, and the example code of our own striking stat, which provides an opportunity to directly apply the new concepts. That self-guided process will be enhanced by a Webex Q&A whenever convenient.

Figure 1 shows the current version of the striking stat (using Tanou's export data) and a possible publication-ready version.

I have selected resources and examples with two goals in mind:

1. You can explore examples with enough background information to get a fair feel for the possibilities and challenges of data visualization in R
2. We have enough of a shared understanding to jointly edit the occasional visualation, even if we use mostly use different tools

> For example, a few weeks ago we needed to change a couple of maps to represent Western Sahara. Alexis had the correct shapefiles; Hannah and I had two different maps produced with two different workflows. If we had been a bit more coordinated, changing the underlying maps would have involved one line of reproducible code. (See example `5_plot_wb_world_map.R`). As it was, Hannah and I painted over our existing maps in Microsoft Paint.

This is not

- A comprehensive introduction to R, to data visualization or even to data visualization in R
- (Yet :-)) An evangelistic mission to convert everyone to R

2 Why R for data visualization?

Advantages	Disadvantages
Scripting language for better reproducibility	Steep learning curve
Open source, freely available everywhere	Overkill for many purposes
Significant complementarities to other analytical skills (statistical analysis, data wrangling incl. web scraping etc., machine learning...)	High barriers to entry make collaboration more difficult
Lots of active development for new data viz tools	

3 Basic setup

- Download and install R from the R-project website `https://www.r-project.org/about.html`
- Download and install the free version of RStudio `https://rstudio.com/products/rstudio/download/`
- Once in R, install the tidyverse package

  ```
  install.packages(''tidyverse'')
  ```

  Please follow the defaults unless you have a good reason not to.
- All code and material for this tutorial can be downloaded here `https://github.com/ammapanin/striking-statistics-tap/tree/master/R-for-TAP`

4 Outline of this tutorial

1. Open `1_plot_simple_export_bars.R` Run the code and produce your first plot! Don't worry about understanding each command.
2. Read through the tutorial on 'Data Types and Structures' at the following link `https://swcarpentry.github.io/r-novice-inflammation/13-supp-data-structures/index.html`
3. Go back to `1_plot_simple_export_bars.R` Try to understand the different data types that can already be found in the code. Get more familiar with the dataset and the plotting options.
4. Read through this chapter introducing ggplot `https://r4ds.had.co.nz/data-visualisation.html`
5. Read about dply rbasics `https://r4ds.had.co.nz/transform.html`
6. Open `2_plot_intermediate_export_bars.R`

   Get familiar with some of the customizations that are offered in this script. Again, don't worry if some pieces of code don't make complete sense.
7. Then read more about 'Understanding Factors'. Factors are a specific data type for storing categorical variables. You would eventually need to be comfortable with them as they are important for plotting (e.g. ordering country labels)

   `https://swcarpentry.github.io/r-novice-inflammation/12-supp-factors/index.html`
8. Then attempt to create the final graphic with `3_plot_advanced_export_bars.R` Look for your output in the main project folder. It will be called something like `E4T_export_compliance.png`
9. Go through `4_tidy_downloaded_data.R` to look under the hood at the data cleaning and transformation process to go from the World Bank download to the tidy dataset we used for analysis

Figure 1: A striking stat to be customized in R