# R-Bootcamp: Assignment

Matteo Tanadini and Claude Renaux

31 January - 3 February 2022

## Admin

In order to obtain the credits for the course "**R**-Bootcamp" students must provide evidence of their successful participation. To do that, students must hand in a document, hereafter *the assignment*, where the tasks listed below are carried out.

## Find a use case that comes with some data

Roughly speaking the assignment represents a complete data analysis where you use a wide variety of **R** functionalities. The first task is to find an interesting use case that comes with data. The choice of the case is completely up to you. Nevertheless, makes sure the following criteria are fulfilled:

- the use case comes with some data
- the dataset must contain at least a few hundred observations and a dozen variables
- among the variables there must be numeric and categorical ones
- the dataset should also contain variables that are dates or geographic locations (both is even better)
- if the dataset itself does not contain dates or geographic locations, you must find another additional dataset with this information (see example below)
- the data can be publicly available or come from e.g. your employer. If the data is not publicly available make sure that you can use the data and discuss the results with the course instructors.

Example of a use case I:

We want to model ice cream sales in Switzerland. So we get the data about ice cream sales in each Swiss city during the past 5 years (say weekly sales provided by Friscolino AG), data about climate (i.e. day temperatures and precipitations e.g. from MeteoSuisse) and a data about city populations (e.g. from the Swiss Federal Office of Statistics). We will then merge all these datasets into a single dataset.

Example of a use case II:

We want to model train arrival times in Switzerland. So we get the departure and arrival times of all SBB trains (e.g. from Puntlichkeit.ch), we the get the stations coordinates and we also get the local vacations days (cantons homepage).

Ideally, your datasets should come into different formats (e.g. xlsx, zip, ...) and from different sources (directly from websites via url, locally downloaded files, ...).

## Prepare the data for the analysis

As mentioned above, datasets need to be merged and cleaned before starting the analysis.

## Visualise the data appropriately

The first step is to inspect the data once it is ready. Use summary statistics and even more imporant, use graphs to inspect the data. The latter is a requirement for your assignment.

## Fit model(s)

Note that the focus of this course is not modelling. Therefore, do not invest time in finding fancy models. A very simple model will do.

You may want to produce graphs (e.g. predictions or residual diagnostics) for your model fits.

If you wish, you can also compare several models via CV. However, remember that the focus of this course in not the modelling part. But rather the coding part.

## A chapter of your choice

In this chapter we want you to use a package that was not mentioned in the course and perform a task that was not directly discussed in the course. Be creative! Note: Here we don't want you to use a new statistical/machine learning method. We rather want you to use a method to prepare or display data. It could even be a package that enables you to create prettier documents. What we don't want to see is you fitting e.g. a regression to your data. Please insert a separate section in your report and call it "Chapter of choice" such that we can easily recognize it!

If you need input: feel free to ask us.

## Dynamic documents and reproducibility

We want you to create the pdf/html document with Rmarkdown. Make sure that your analysis is fully reproducible and comprehensible for anyone reading it.

## Comments

The analysis needs to be commented. Keep in mind that you should make up a story to tell to a client. So, we want you to comment what you are doing and why you are doing things. We also want you to comment on the results. Putting 20 uncommented plots of the model fits is not something a client would like to see. Think about a "conclusions" chapter that the client may want to see.

Note, however, that you should not add a very long text section... just to have some text. Always keep in mind that a potential client will read your work, he/she wants to understand what you did and why and what the results of the analysis are. However, he/she does not want to get bored.

Finally, note also that very often Rmarkdown documents become very long. Two hints to shorten them is to use the chunk options *message* and *results* such that only really needed message and results are shown. For example, in the chunk where you load packages, you may want to set *message* to *FALSE*.

## Sell the story

Your analysis must be a story that you would like to sell to a client. It is completely up to you how the story goes and flows. Nevertheless, keep in mind that you need to make clear:

- where the data comes from
- what the aims of the analysis are
- what the conclusions of your analysis are

## Formal requirements

Here a few formal points about the assignment:

- students must work in pairs
- deadline to hand in the assignment is on Friday 3 weeks after the end of the course at 5 pm (hard deadline! No later submission will be accepted)
- the assignment must be uploaded on Ilias on the delivery folder
- you must hand a zip file named after your family names (e.g. Tanadini_Mueller.zip)
- the zip file must contain:
  - the dataset analysed
  - the Rmarkdown file (.Rmd)
  - the pdf or html output file
- the zip file should contain:
  - structured folders (e.g. "Data", "Scripts", ...)
  - a readMe.txt file
- the length of the pdf/html document:
  - should not exceed 25 pages
  - anything longer than 30 pages will not be considered
  - you can export your list of functions as a pdf and append it to the analysis document
- your work will be graded as *passed* or *not passed*
- the evaluation criteria are:
  - completeness (are all/most topics addressed in class there?)
  - sophistication (adapted to your level)
  - quality of the end product (e.g. commented plots and summaries)
- where to find data:
  - anywhere!
  - note that kaggle is one of many sources of datasets (too often used by students)
  - You may want to consider websites like opendata.swiss, https://opendata.swisscom.com, https://datasetsearch.research.google.com/, http://puenktlichkeit.ch/, www.mldata.io, https://www.pxweb.bfs.admin.ch/, https://www.ostluft.ch/, http://www.agrometeo.ch/de/meteorology/datas, https://data.stadt-zuerich.ch/, https://data.world/uci/, https://databank.worldbank.org/source/world-development-indicators, https://data.gov.sg/dataset, (just to mention a few examples...)
  - **R** itself comes with a few hundreds datasets (type "data()" to see the list), the vast majority of add-on packages also comes with datasets
  - note also that if you found a cool dataset that seems to be too small/simple... you may complement it with another dataset (e.g. add coordinates, meteo or similar things).