European University Institute
DEPARTMENT OF POLITICAL AND SOCIAL SCIENCES
Co-funded by the Erasmus+ Programme of the European Union

Bottom-up Workshop

3rd term 2018-2019

Data management in R

Tricks to save time when preparing your data for analysis

Organized by: Elias Dinas, Juho Härkönen, and Adrián Del Río Rodríguez

Instructor for introductory part: Shpend Kursani

Instructor for data management part: Adrián Del Río Rodríguez

Register online: here

Description

Typically, researchers need to assemble, shape and create different datasets and variables into a useful form for a variety of analyses. In fact, these tasks are often more time-consuming than the analysis itself; and prone to a wide range of errors and inconsistencies. This workshop provides a set of routines to enter, cleanse and manage your data in the most efficient manner in R. Specifically, participants will learn tricks to solve automatically common problems when dealing with panel data, event data and text as data in a transparent and reproducible manner. Moreover, participants will learn about some add on packages that make R powerful in Data wrangling. The course's length is 3 days. Participants are free to attend either all the days or any combination of days. The first day is devoted to an introduction to R programming, data structure and Rmarkdown. The second day provides a set of tools to solve the most common challenges in data manipulation and data store in a clean and consistent way. The last day shows a handful number of techniques to join multiple data sets together, creation of data joins and manipulate individual characters (e.g. words) within character vectors.

Requirements

The process of getting datasets into R in a useful form for data visualization and modelling is a crucial step of research projects. Therefore, this course is intended for all quantitative oriented researchers. Basic knowledge of R is not required since the first session is devoted to an introduction.

Before the workshop

You must have R and RStudio installed prior to the beginning of the Workshop. Follow the below steps:

- Download and install **R** from here. Click on the "Download R 3.5.2 for Windows"
- Download and install **R Studio** from here. Click on the "RStudio Desktop" FREE version.

Schedule

1st Day. Introduction to R and in-class exercises.

April 9th: 9:00 – 13:00, Seminar Room 2 at Badia Fiesolana

2nd Day. Data management I: Shape and create different datasets and variables

April 11th: 14:00 – 18:00, Seminar Room at Villa Paola

3th Day. Data management II: Relational Data and regular expressions

April 12th: 14:00 – 18:00, Seminar Room at Villa Paola

Organization and expected goals

1st Day: Introduction to R and in-class exercises.

The first part of the seminar covers the introductory steps to R programming. R is an open source software that enables users to conduct statistical analysis, other mathematical operations, varieties of qualitative analysis, web-scraping, creation of texts and graphs of "publication quality" with particular ease. This introduction shows the fundamental logic of how R operates and the functions that will enable participants to carry out more complex tasks (even by themselves). **After the workshop, participants will learn:**

- The logic of the R platform, environment, and its general capabilities
- The creation of objects that can be used for different analytical purposes
- The logic of RMarkdown, environment and its general capabilities
- Import and export datasets in a variety of formats
- Participants will be acquainted to basic, yet necessary, functions to carry out various quantitative and qualitative analyses (e.g ifelse() command and function())
- If time allows, basic descriptive analysis in R will be taught

2nd Day. Data management I: Shape and create different datasets and variables

The second part of the workshop introduces participants to a handful number of routines and functions in data management. Special attention is devoted to the process of manipulating, sorting, summarizing, and reshaping datasets into a tidy form (e.g. expand datasets into a wide or long format conditioned on a column). Moreover, practical advices will be shared in order to exploit the capabilities of R functions; and encourage participants' creative thinking in order to deal with common challenges in data management. **After the workshop, participants will learn:**

- How to manipulate and re-shape datasets in a consistent, complete and informative manner
- How to handle the time dimension of datasets: introduction to time data-types and its use to expand datasets or create variables (e.g. age)
- Introduction to R studio connect: save and share scripts online
- A handful number of functions from the tidyverse package
- How to identify and correct errors in datasets: missing values, creation of ids and the problem of duplicated observations

3th Day. Data management II: Relational Data and regular expressions

The third part of the workshop deals with common challenges when joining multiple datasets together and the manipulation of characters (e.g. words) within character vectors. Specifically, it introduces a number of functions to merge datasets and show their capabilities to deal with a number of data errors. In addition to this, the workshop introduces participants to the functions and language, known as regular expressions, to manipulate characters within character vectors. **After the workshop, participants will learn:**

- methods of data joins in R and work with relational data
- functions and strategies to manipulate strings (characters within character vectors)
- regular expressions: a concise language for describing patterns in strings

Useful materials

- An Introduction to R
- R Installation and Administration

Assistance

For further questions regarding the topic that instructors deal or suggestions, write to:

Adrian Del Río Rodríguez: Adrian.DelRioRodriguez@eui.eu

Shpend Kursani: Shpend.Kursani@eui.eu