PhD in Chemical Sciences
Department of Chemistry "Ugo Schiff"
University of Florence

PhD course "Introduction to data analysis with R"

SYLLABUS

1   Lecturer information

Name and Surname: Fabio Giardi

National Institute for Nuclear Physics (INFN), Section of Florence

Via Bruno Rossi 1, I-50019 Sesto Fiorentino (Florence), Italy

fabio.giardi@fi.infn.it

Proposed by:

Franco Lucarelli, lucarelli@fi.infn.it

2   Title of the course

Introduction to data analysis with R

3   Course program

The course aims to introduce the basics of R which is a language purely oriented to data analysis. Knowing R means being able to automate workflows that can include data import, data cleaning and manipulation, statistical analysis and visualization. This therefore means having the possibility of reusing the same code for a potentially infinite number of times with the advantage of having accuracy and reproducibility and being able to easily process a large amount of data. This is not a computer theory course but based on the practical use that can be done with R in research, from the exploration of data to the production of images suitable for publication, thanks to the graphic potential offered by some function packages.

4   Course content detailed per lesson of two hours (possibly with dates and room real and virtual)

Lesson 1 – Introduction to the course. Advantages and disadvantages of the R data analysis language. Introducing the RStudio graphical user interface (GUI). Structure and basic commands of the code: mathematical functions, value assignment, logic programming (if and iterative loops), types of objects and data structure. Data import.

Lesson 2 – Data processing and analysis with the main packages: data exploration, modification (data cleaning, identification and removal of outliers and treatment of missing data). Processing of text strings and time data. Statistic analysis. Custom functions.

Lesson 3 – Graphical representation: data preparation, construction of the main types of graphs and discussion of all the levels that compose them. Chart customization.

Lesson 4 – Particular graphical representations: combinations of graphs, maps. Sharing of results with reports in pdf and html. Code versioning with Git. Practical applications on data provided or proposed by students.

5   Suggested reading
None

6   Learning Objectives
The aim of the course is to provide the knowledge to use the basic R language, its most popular packages and the RStudio graphical interface to allow reliable and reproducible data analysis and data representation.

7   Knowledge and Skills to be acquired
Students will acquire the minimum knowledge of the R language to create workflows that allow data analysis and their representation.

8   Prerequisites
To be able to run the code in first person during the lessons on your computer it is recommended to install R (https://cran.r-project.org) and the RStudio (https://posit.co/download/rstudio-desktop/) graphical interface.

9   Teaching Methods

☐ MODE 1 - Pre-recorded lessons uploaded on the moodle platform (a meeting must be organized with PhD students in order to clarify eventual doubts)

X MODE 2 (preferred) - Lessons delivered in-person and in remote with simultaneous recording by the WEBEX platform

(The lessons must be recorded and available to all the students that cannot take part to the lessons in streaming. The Webex platform must be used. All course content should be uploaded to the Moodle platform on the Chemical Sciences PhD page "Courses and Seminars of the PhD in Chemical Sciences 2022-2023")

10 Further information

11 Type of Assessment

Students will be asked to create a workflow which, starting from the import of data on file, leads to the export of processed data and the production of at least one graph.

The final evaluations will have to be validated maximum 1 month after the end of the course

12 Period

From June