Data Handling with R (3 ECTS)

Course code: PNG0100

Subject: General science

Course type: This course will be given as a combined in-person and distance learning course. Much of the course is assignment-based, with students taking a large degree of responsibility for attaining the learning outcomes. Guided individual learning will be followed by a hybrid in-person (at SLU Uppsala) and online workshop where the whole class will meet to discuss the work.

Language: English

Prerequisites: A basic knowledge of the R language, enough to use it as a statistical tool for research. For example, being able to read/write files, simple manipulation and indexing of R objects, basic analyses, plotting data. Admitted to PhD-studies.

Objectives:
The course aims at improving the effectiveness of the scientific code you write by tapping into generally less utilized capabilities of R and its software ecosystem. In particular the course will teach you how to write more ordered code that can be easily reused and incorporated into other projects and to deal with the automation of data and code handling tasks. This will allow you to save time and handle bigger datasets.

Learning outcomes:
After completing the course, students should be able to:

1. Write reproducible code
    1. Basic use of GitHub and Rmarkdown.
    2. Write code that can be reused yourself or used and modified by a complete stranger
    3. Standardize code and data structure
2. Confidently manipulate data and R-objects (never touch excel ever again)
    1. Understand data manipulation tools such as the apply family
    2. Index, group and aggregate data
    3. Write simple loops and functions
3. Check for and fix errors in data and code
    1. Data cleaning and error checking (tests such as look for outliers, NAs, patterns)
    2. Diagnostic plotting
    3. Basic functions for debugging and most common errors

Content:
The course will start with a short introductory workshop. Thereafter the content will be split across the three themes of the learning outcomes above. In each theme, the students will be given some learning materials and a task to complete. Students will have the opportunity to communicate online among themselves and with the teachers using asynchronous tools while working on each theme, before a class-wide hybrid workshop where students will discuss the assignments and work on additional tasks.

Introduction: The course will start with a short online introduction where students and teachers will get to know each other and the structure of the course will be explained. Time will then be given for students to install and connect to the relevant software that will be used during the course.

1. Write reproducible code
Individual study: Here students will acquaint themselves with the very basics of using GitHub for code backup, archiving, sharing and editing, uploading their work to a course project site. Working in small groups, students will practice writing reproducible code using one of two sample data sets. The idea is that, while the functions used in the exercise should already be familiar to the students the students will write the code in a reproducible way that with very little editing allows the same code to be used on an alternative data set by a stranger.
Workshop: Students and teachers will discuss the students' experiences in writing reproducible code and additional tasks will be provided.

2. Confidently manipulate data and R-objects
Individual study: Study material will teach the students different ways to manipulate large and heterogeneous data sets in a reproducible way. Students will then be given one of two data sets and be asked to organize the data in a specific way. As in the first theme, students will work via GitHub in small groups, checking and commenting on each other's code for clarity and reproducibility.
Workshop: Students and teachers will discuss the students' experiences in writing reproducible code and additional tasks will be provided.

3. Check for and fix errors in data and code
Individual study: In this theme, students will learn some ways to identify and deal with errors in code and/or data sets. They will then receive a 'buggy' data set (optional: own data set), and using these skills and the knowledge gained in the rest of the course to clean and restructure the data in order to produce a set of specified figures. Again, students will work via GitHub in small groups, checking and commenting on each other's code for clarity and reproducibility.
Workshop: Students and teachers will discuss the students' experiences in writing reproducible code and additional tasks will be provided.

Examination:
To receive course credits, students are required to have completed all individual exercises and played an active part in all workshops.

Timetable:

14 November 2023	Kick-off workshop: student and teacher introductions. Connecting to course platforms.
30 November 2023	Workshop: Theme 1.
15 December 2023	Workshop: Theme 2.
19 January 2024	Workshop: Theme 3.

Please note that participants will get learning materials in advance and are expected to complete exercises before each scheduled workshop.

Contact for application and further information:
Apply for the course **no later than 1 November 2023** by sending an email to Alistair Auffret: alistair.auffret@slu.se ,who will lead the course together with Lorenzo Menichetti: lorenzo.menichetti@slu.se.

Literature:
Online learning materials will be distributed during the course.

Additional Information:
The course is organized as part of the of the NJ-faculty research schools *Ecology -basics and applications* and *Focus on Soils and Water*