# Advanced Data Handling with R (1.5 ECTS)

**Course code: PNG0090**

**Subject: General science**

**Course type:** This course will be given as a distance learning course, where students will take a large degree of responsibility for attaining the learning outcomes. Guided individual learning will be followed by an individual or small-group activity and reflection, before a lecture or workshop (online or in person) 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 handling tasks. This will allow you to save time and handle bigger data-sets.

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

1. Write reproducible code
    1. Write code that can be reused yourself or used and modified by a complete stranger
    2. Standardize code and data structure
    3. Basic use of Github
2. Confidently manipulate data and R-objects (never touch excel ever again)
    1. Understand data manipulation tools such as the apply family and and the plyr package
    2. Index, group and aggregate data using the above functions
    3. Use advanced techniques for string manipulation
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 be split across the three themes of the learning outcomes above. In each theme, the students will be given some learning materials (online video and example code) and a task to complete. There will be a non-compulsory online question and answer session for each theme, before a class-wide workshop where students will present their work.

**1. Write reproducible code**
**Individual study:** Here students will practice how to write reproducible code. Students will consult style guides and then be given a simple exercise using one of two sample data sets. The idea here is that the functions used in the exercise should already be familiar to the students, but that the students will write the code in a reproducible way. Students will create a new (or use an existing) GitHub account to upload their work to a course project site. Students will then be assigned into groups of two and will have to use each other's code to first re-run the original exercise, and then use the same code to

complete the exercise with the other data set. Students will discuss together how their code could be improved to become more understandable and reproducible.
**Workshop:** Each pair will present the results of their exercise, and teachers will lead a discussion based on the students' experiences.

**2. Confidently manipulate data and R-objects**
**Individual study**: Study material will teach the students different ways to manipulate large and heterogeneous datasets in a reproducible way. Students will then be given one of two data sets and be asked to produce a set of specific figures. As in the first theme, students will be assigned to groups of two and check and comment on each other's code for clarity and reproducibility.
**Workshop**: Each pair will present the results of their exercise, and teachers will lead a discussion based on the students' experiences.

**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 datasets. They will then receive a 'buggy' dataset (optional: own data set), and using these skills and the knowledge gained in the rest of the course to clean and restructure the dataset in order to produce a set of specified figures. Again, students will be assigned to groups of two and check and comment on each other's code for clarity and reproducibility.
**Workshop**: Each pair will present the results of their exercise, and teachers will lead a discussion based on the students' experiences.

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

**Time table:**
Scheduled whole day course meetings and workshops on **November 4, November 11 and December 2** Please note that participants will get learning materials in advance and are expected to complete exercises before each scheduled meeting day.

**Contact for application and further information:**
Apply for the course **no later than October 4th** by sending an email to the course organizers: Alistair Auffret: alistair.auffret@slu.se and Lorenzo Menichetti: lorenzo.menichetti@slu.se

**Literature:**
Online learning materials will be distributed at the beginning of the course.

**Additional Information:**
The course will follow the current recommendations concerning the public health and Covid-19 pandemic from the Public Health Agency of Sweden (FHM) and SLU.

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

.