Syllabus Module 215: "Introduction to R software for data sciences"

N° : 215	Introduction to R software for data sciences
Coordinator	Nolwenn LE MEUR, Eng, PhD Professeur of Biostatistics, EHESP Nolwenn.LeMeur@ehesp.fr
Teachers	Nolwenn Le Meur, EHESP Amin Gharbi, Teaching asssistant, Mondor Institute for Biomedical Research / Ecole normale supérieure
Dates	Monday 16h – Friday 20th October 2023
ECTS	3ECTS
Duration	30 hours (5 days of 6h)
Location	EHESP 20 Avenue George Sand 93210 LA PLAINE ST DENIS
Description	The course will teach students how to use the R statistical software in data sciences to start analyzing and presenting data. Data management, statistical analysis, visualizations, and generating reports will be the main achievements of this week course. R is not only a free statistical software but also a language and environment for statistical computing and graphics. R is highly extensible and runs on a wide variety of UNIX platforms, Windows and MacOS. During this module, the emphasis will be on learning by doing. Practice material and data will be grounded on actual research questions and are intended to illustrate the kinds of issues that often arise when practicing quantitative data analysis, whatever the public health topic (statistics, epidemiology, environmental, policy …). Each session will be a combination of didactic lecture and hands-on practice. Students will be encouraged to apply the material to their own research interests.
Prerequisites	Students are assumed to be familiar with elementary statistical methodology (such as descriptive analysis, hypothesis testing, regression analysis)
Course learning objectives	The course will familiarize students with the programming skills necessary to: • Use the R statistical computing environment to enter, read, clean, organize, index and manipulate data in R. • Apply R functions and packages to describe and analyze data (statistical tests and models). • Apply R functions to visualize, present data. • Understand the R transition from classic (S3) to grammar-oriented programming (with dplyr, tidyr, ggplot2 libraries) • Understand how R can be extended with functions. • Save analysis and create report for reproductible research
Competences	**Competences:** 1) Knows how to retrieve, analyze and appraise evidence from all data sources to support decision-making 2) Uses vital statistics and health indicators effectively to increase knowledge and generate evidence about population health, including within at-risk and vulnerable groups 3) Applies data sciences methods, (digital) technologies and good practices for managing, analyzing, and storing data and health information. **Teaching activities:** 1)The students will choose a public health issue of their interest and retrieve (open access) data on the matter to be analyze. A short list of datasets will be available if none identified by the student. 2 & 3) The R statistical software will be used to compute statistical parameters to summarize and model on vital statistics, health measurements and health indicators to generate evidence about population health **Evaluation:** A study report on the student's public health question(s) will be written using the statistical software R. The report will take the form of a research paper (including Introduction, Material and Methods, Results and Conclusion). The input document will have to be well commented and the final document (output) should be reproducible.

Structure
(details of sessions title/speaker/date/duration)
	Class meets Monday from thru Friday from 9:00 am – 4:00 pm (30 hours total). Classes encompass lectures, practical, quiz and final project assignment. Dr Nolwenn Le Meur delivers all sessions, details of each session are provided below.
	• R environment - Day 1 – 9:00 am to 4:00pm
	  ○ R and RStudio
	  ○ Data management: Importing, exporting, data types and structure, recoding variables, handling missing values.
	• Exploratory data analysis – Day 2 - 9:00 am – 4:00 pm
	  ○ Summary, univariate statistical test, intro to grammar-oriented programming
	  ○ Basic plot and gglopt2
	• Machine Learning #1 Day 3 -9:00 am – 4:00 pm
	  ○ Regression statistical models (lecture and practical) – Day 4 - 9:00 am – 4:00 pm
	• Machine Learning #2 Day 3 -9:00 am – 4:00 pm
	  ○ Tree based models.
	  ○ Classification methods
	• Project: Day 5. 9:00 am – 4:00 pm

Resources	**The following texts are recommended:** • An Introduction to R. Free at http://cran.r-project.org/doc/manuals/R-intro.pdf • Aragon TJ, et al. Applied Epidemiology Using R. Free at https://bookdown.org/medepi/phds/ • *Bradley Boehmke & Brandon Greenwell Hands-On Machine Learning with R 2020-02-01* *https://bradleyboehmke.github.io/HOML/index.html* • *Hadley Wickham R for Data Science https://r4ds.had.co.nz/*
Course requirement	Students have to come to class with a computer with the software R and the software RStudio installed (Instructions for installation will be send one week before the course)
Grading and assessment	**In-class Exercises: 20%** Starting the second day, the morning sessions will begin with a quiz related to the previous day. The quiz will be proposed via the EHESP REAL platform. **Final Project: 80%** At the beginning of the week the students will choose a dataset on a public health issue of their interest and they will analyze it using R. A short list of datasets will be available if none identified by the student. Student will be able to start working on the data at home and will have the last day of the week to finalize the analysis. Specific questions will be asked to assess the student's competencies. The objectives of the final project are to : • Describe the data (present summary statistics and graphs) • Investigate the factors associated with public health issue Students are required to submit a final paper (**7 pages max**) and the associated R code presenting their analysis. The assignment is intended to allow the student the opportunity to demonstrate his or her ability to conduct preliminary analyses of a public health problem using R. • Cover Page: Title, student's name, date • Introduction (short paragraph): What is known about the public health topic, what is not known, why it is important. • Materials and Methods: Analytic design or method to be presented, including variables and their roles in the analysis (explanatory, confounding, interaction, etc...). The main R packages used with version. • Results: bullet points walkthrough of the analysis with graphics, tables, and short text of interpretation of the principal results. • Discussion: Including limitations, conclusions, possible next steps. • References: if appropriated and limited to 15 citations (not counted toward pagelimit).

Course policy	**Attendance & punctuality** **Regular and punctual class attendance is a prerequisite for receiving credit in a course.** Students are expected to attend each class. Attendance will be taken at each class. The obligations of attendance and punctuality cover every aspect of the course: - lectures, conferences, group projects, assessments, examinations, as described in EHESP Academic Regulations http://mph.ehesp.fr EHESP Academic Regulation Article. 3). If students are not able to make it to class, they are required to send an email to the instructor and to the MPH program coordinating team explaining their absence prior to the scheduled class date. All supporting documents are provided to the end-of-year panel. . Students who miss class are responsible for content. Any student who misses a class has the responsibility for obtaining copies of notes, handouts and assignments. If additional assistance is still necessary, an appointment should be scheduled with the instructor. Class time is not to be used to go over material with students who have missed class. **Lateness**: Students who are more than 10 minutes late may be denied access to a class. Repeated late arrivals may be counted as absences (See http://mph.ehesp.fr EHESP Academic Regulation Article. 3 Attendance & Punctuality) **Maximum absences authorized & penalty otherwise** Above 20% of absences will be designated a fail for a given class. The students will be entitled to be reassessed in any failed component(s). If they undertake a reassessment or they retake a module this means that they cannot normally obtain more than the minimum pass mark (i.e. 10 out of 20) **Exceptional circumstances** Absence from any examination or test, or late submission of assignments due to illness, psychological problems, or exceptional personal reasons must be justified; otherwise, students will be penalized, as above mentioned. Students must directly notify their professor or the MPH academic secretariat before the exam or before the assignment deadline. Before accepting the student's justification, the professor or the MPH academic secretariat has the right to request either a certificate from the attending physician or from a psychologist, or from any other relevant person (See http://mph.ehesp.fr EHESP Academic Regulation Article 4 Examinations). **Courtesy**: All cell phones/pages MUST be turned off during class time. Students are required to conduct themselves according to professional standards, eating during class time is not permitted during class time, such as course or group work.
Valuing diversity	Diversity enriches learning. It requires an atmosphere of inclusion and tolerance, which oftentimes challenges our own closely-held ideas, as well as our personal comfort zones. The results, however, create a sense of community and promote excellence in the learning environment. This class will follow principles of inclusion, respect, tolerance, and acceptance that support the values of diversity. Diversity includes consideration of: (1) life experiences, including type, variety, uniqueness, duration, personal values, political viewpoints, and intensity; and (2) factors related to "diversity of presence," including, among others, age, economic circumstances, ethnic identification, family educational attainment, disability, gender, geographic origin, maturity, race, religion, sexual orientation and social position.
Course evaluation	EHESP requests that you complete a course evaluation at the end of the school year. Your responses will be anonymous, with feedback provided in the aggregate. Open-ended comments will be shared with instructors, but not identified with individual students. Your participation in course evaluation is an expectation, since providing constructive feedback is a professional obligation. Feedback is critical, moreover, to improving the quality of our courses, as well as for instructor assessment.