Advanced Programming Tools in Data Science

Instructors: P.BESBEAS, P.PAPASTAMOULIS

Course Code: **.....**
Course Type: **Elective**
Course Level: **Graduate (MSc)**
Year of Study: **B'**
Semester: **4rth**
ECTS: **5**
Language: **Greek or English**

Course Description
Data Science refers to the scientific study of data. The module brings together two eminent programming tools for analysing data: `Python` and `R`.
The first component of the module will introduce students to `Python`, a programming language that has become the leading choice for scientific computing. Students will learn how to use `Python` in order to perform a range of statistical analyses for modern Data Science. You'll learn the latest versions of pandas, `NumPy` and `Jupyter` in the process.
The second component of the module consists of advanced data manipulation techniques and efficient coding in `R`. We will combine data engineering libraries with web scraping applications. Then we are going to speed up `R` by integrating with `C++`, something that can be extremely advantageous when writing source code for computationally demanding problems. Finally, we will extend `R` by creating our own package.

Prerequisites
Familiarity with the `R` programming language.

Target Learning Outcomes
- Develop skills for choosing the right tool between `Python` and `R` for your data
- Use the Jupyter notebook and I`Python` shell for exploratory computing
- Take advantage of basic and advanced features in `NumPy`
- Get started with data analysis tools in the `pandas` library
- Use flexible tools to load and manipulate data
- Create informative visualizations with `matplotlib`
- Utilise `Python` for advanced scientific computing
- Data engineering: `dplyr`, `tidy`
- Web scraping `html`, `css`, `rvest`, selector gadget, `rselenium`
- Integrating `R` and `C++`
  - basics of bridging `R` with other languages
  - `Rcpp` (and related) packages
- Creating extensions with `R`.

Indicative Reading

Randall L. Eubank, Ana Kupresanin, “Statistical Computing in C++ and R”, Chapman & Hall/CRC, 2023

H. Wickham, “Advanced R”, Second Edition (Chapman & Hall/CRC The R Series), 2019

M. Dawson, “Python Programming for the Absolute Beginner”, 3rd edition, 2011, Cengage, ISBN 9781435455009

A.B. Downey, Think Python, 2nd edition, 2015, O'Reilly, ISBN: 9781491939369

M. Lutz, Learning Python, 5th edition, 2013, ISBN: 9781449355715

W. McKinney, Python for data analysis, 2013, O'Reilly, ISBN: 9781449323622

Friedrich Leisch (2009). Creating R packages: a tutorial. R-developer core team.

Eddelbuettel D, François R (2011). “Rcpp: Seamless R and C++ Integration.” *Journal of Statistical Software*, **40**(8), 1–18. doi:10.18637/jss.v040.i08.

Eddelbuettel D (2013). *Seamless R and C++ Integration with Rcpp*. Springer, New York. doi:10.1007/978-1-4614-6868-4, ISBN 978-1-4614-6867-7.

Eddelbuettel D, Balamuta J (2018). “Extending R with C++: A Brief Introduction to Rcpp.” *The American Statistician*, **72**(1), 28-36. doi:10.1080/00031305.2017.1375990.

Hadley Wickham, Mine Çetinkaya-Rundel, Garrett Grolemund (2023). R for Data Science, 2nd Edition. ISBN: 9781492097402

Teaching and Learning Activities
Classroom teaching and assignments.
Contact hours

- Total contact hours: 24
- Private study hours: 76
- Total study hours: 100

Assessment and Grading Methods
Combination of (i) weekly assignments, (ii) assessment, and (iii) final exam.