UNIVERSITY OF LONDON

# Course information 2023-24

# ST2195 Programming for data science

## General information

**MODULE LEVEL:** 5

**CREDIT:** 30

**NOTIONAL STUDY TIME:** 300 hours

## Summary

In the last decade the demand for programming skills related to managing and visualizing data has grown remarkably. Python, R and SQL feature consistently in the top skills listed in data science and data analyst jobs. Knowing how to write efficient software code to handle and visualise data is an essential skill for any modern data scientist. This course will cover the main principles of computer programming with a focus on data science applications by following the entire pathway from raw data to databases, data wrangling and visualisation, machine learning frameworks up to software development.

## Conditions

None

## Aims and objectives

- Gain knowledge on the main principles of programming in the Data science context
- Develop ability to handle and visualise data
- Apply computational thinking in various applications domains
- Provide training in state-of-the-art tools, e.g. SQL, Python, R and Git
- Communicate the data analysis results to stakeholders and share work with people in the Data Science industry

## Learning outcomes

At the end of the course and having completed the essential reading and activities students should be able to:

- Convert raw data to relational databases such as SQL
- Import data to Python and R, apply data manipulation and visualisation
- Program in Python and R
- Develop software using version control via Git

Please consult the current EMFSS Programme Regulations for further information on the availability of a course, where it can be placed on your programme's structure, and other important details.

## Essential reading

McKinney W. *Python for Data Analysis*, 2nd edition O'Reilly (2017)

Gutagg J.V. *Introduction to Computation and Programming using Python*, MIT Press, 2nd edition (2017)

Wickham H. and Grolemund G. *R for Data Science*, 1st edition O'Reilly (2017)

Wickham H. *Advanced R.*, 1st edition Chapman & Hall (2015)

Rammakrishnan R. and Gehrke J. *Database Management Systems*, 3rd edition, McGraw Hill (2002)

## Assessment

This course is assessed by an individual case study piece of coursework (50%) and a three-hour and fifteen-minute closed-book written examination (50%).

## Syllabus

**From raw data to databases**

Real examples of raw data, relational databases models, structured query languages (SQL), data extraction, processing of various human-readable data formats (e.g. JSON, XML, CSV), importing to Python and R, data types and data structures.

Relevant programming concepts, such as IDEs, control flow structures, variables, functions, loops, errors and exception handling, and data input-output operations.

**Key steps of a data-analytic pipeline**

Starting with formulation of a data science problem, going through manipulation and visualisation of data, and, finally, creating actionable insights.

**Data wrangling**

Data cleaning and transformation, representation of data using tabular data structures and their manipulation. Programming and handling data types in R and Python such as scalars, factors, vectors, matrices, arrays, lists and data frames. Introduction to NumPy and Pandas in Python, and the data wrangling utilities in base R and the tidyverse collection of R packages.

**Data visualisation**

Methods for explanatory data analysis, using various statistical plots such as histograms and boxplots, data visualisation plots for time series data, multivariate data, dimensionality reduction methods for visualisation of high-dimensional data, graph data visualisation methods. Hands on experience with Python (matplotlib and seaborn) and R (base R graphics, ggplot2).

**Interacting with machine learning frameworks**

Introduction to Machine Learning via standard frameworks in Python (SciPy, Scikit Learn) and R (glm methods, mlr, caret).

Relevant programming concepts such as modularisation and aspects of parallel computing.

**Data Analytics software development**

Use of version control via git to share work and collaborate with others in the Data Science industry. Software testing methods and test-driven development (using unit testing). Developing an R package.

Please consult the current EMFSS Programme Regulations for further information on the availability of a course, where it can be placed on your programme's structure, and other important details.

Please consult the current EMFSS Programme Regulations for further information on the availability of a course, where it can be placed on your programme's structure, and other important details.