Data Management in R

Dr Jon Minton, AQMEN Research Fellow

25-26 April, University of Glasgow

Processes of data management - including finding data, cleaning data; renaming, joining, filtering, deriving formatting, classifying and restructuring data tables and variables – often take up the bulk of the time involved in doing quantitative data analysis, with one widely quoted estimate suggesting these ‘janitorial’ tasks take up around 80% of the time involved in doing statistical work. However, whereas there are many courses which describe how to do statistical analyses with datasets that have been ‘prepared earlier’, there are far fewer courses that describe the methods and processes needed to do such data preparation yourself. Instead, such skills tend to be learned ‘on the job’, through trial and error, often when project deadlines are looming and stress is mounting.

This course fills the gap in data management teaching, discussing how R and RStudio can be used to quickly and efficiently handle the complete process involved in turning poorly structured and formatted datasets into ‘tidy data’ tables which can be analysed and explored quickly and nimbly. The course will use a new suite of interlocking R packages, in particular tidyr and dplyr, which have been developed to make the many small tasks and challenges involved in preparing and management much more straightforward. If these tasks take up 80% of research project time, then it follows that becoming even slightly more efficient at performing them can easily double or even triple the amount of time available to produce the analyses and results that make great papers and help make better decisions.

This course will cover:

* The data-to-value chain, and the differences and overlaps between statisticians and data scientists
* How to user R with Rstudio
* How and why to set up RStudio projects
* How to make R code easier to build, test and read using ‘code piping’
* Ways of adapting existing code and functions to use ‘code piping’ conventions
* How to read and write data files in a range of formats
* Automated data cleaning using stringr and related packages
* The tidy data philosophy and the benefits of tidy data structures for nimble analysis
* Data Wrangling using tidyr and dplyr
* Ways of performing rapid data analysis using tidyr, dplyr and other packages
* An introduction to ggplot2 as an extension of code piping
* An introduction of plyr for automating the reading and writing of files, including image files

**Course outcomes:**

* An understanding of important data project management concepts, such as the data-to-value chain and tidy data structures
* A familiarity with packages and processes useful for achieving these aims, and how to apply them in practice
* Knowledge about how to use R for all stages of data analysis.
* Hints and tips about how to develop further as efficient and effective data producers and users.

**Course requirements:**

* Some familiarity with script-based approaches to quantitative analysis and basic programming concepts
* Some knowledge of how computers store and encode data, and the implications of this for data management
* A willingness to be challenged and learn, to persist and keep learning