Data Science in R
25th – 26th April 2016
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 used 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 covered:

- 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

Attendance

The course was attended by 18 delegates and feedback was received from 14 of them. The statistics quoted in this document are based on the 14 who completed the evaluation.

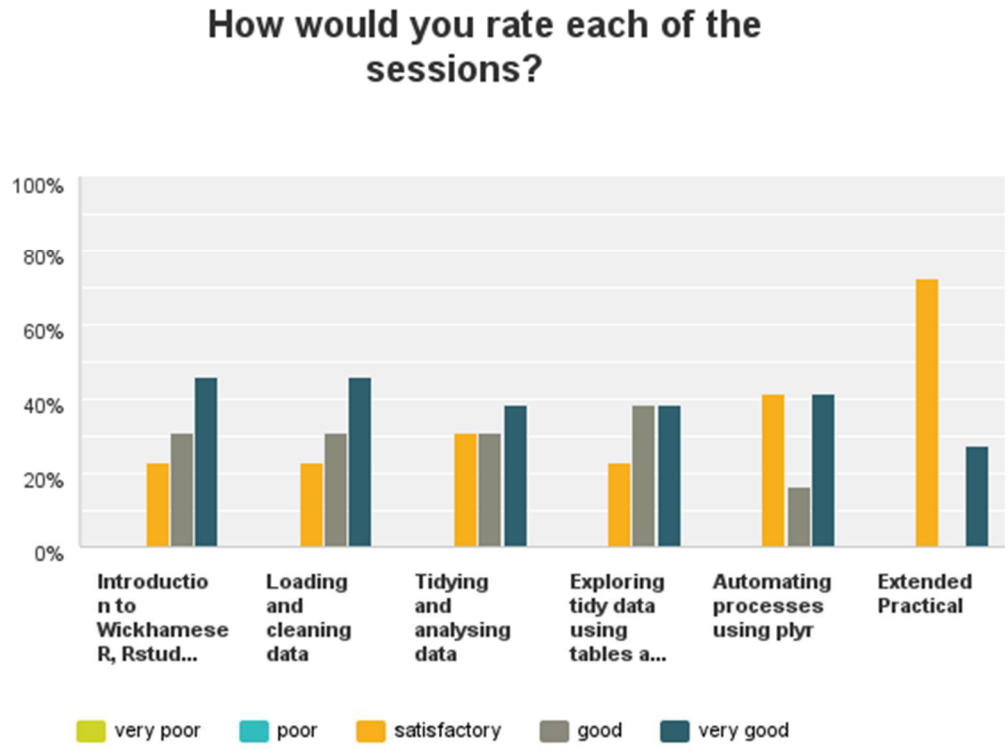
The delegates rated themselves as 'complete beginner' 38%, 'very little experience' 23% and 'quite experienced' 38% in relation to using R.

71% of delegates were made aware of the course via an AQMeN newsletter, and the remainder were informed of the course from an email bulletin (21%) or colleague/supervisor (7%).

Course Structure and Content

The course was positively received with 29% of delegates noting it exceeded their expectations and the remainder that it met expectations, bar one delegate who stated it did not meet expectations, but failed to leave a comment as to why this was.

As detailed below, sessions were well received with the majority rating sessions as very good or good and the rest as satisfactory.



The most popular session was ‘Introduction to Wickhamese R, Rstudio and Research Project Management’ and ‘Loading and cleaning data’, which 46% of delegates, rated as ‘very good’.

Delegates were asked if they had any general comments on the theory and practical session’s overall, and some noted the amount of material to get through in the two days was rather large and the fact they would have benefitted from more direction from the course leader-

- In my opinion, practical sessions should have had more guidance, perhaps an overview given by the instructor to explain what the practical was about.
- There was a challenging amount of material to cover in 2 days. I'm not sure how many people got onto the extended practical - I certainly didn't. The final session on automating processes using plyr seemed a jump up in terms of complexity from the earlier material and as this came at the end of a long 2 days, I'm not sure I got all I could out of this part - probably either omit or provide written material for advanced study after the course.
- I liked being able to work through booklet at my own pace. Class was a good friendly and supportive learning environment. Only comment would be a bit more warning about the pre-course materials.
- Very few people got on to the practical, so I would suggest reducing some of the material.
- More teacher-led training would have been beneficial, but not necessary.
- I think the course may have benefited from more teacher-led content. After a brief introduction, attendees were basically left to work through a workbook. This approach has some advantages but as I had never used R before, it was challenging. That said, the support from the course leaders if encountering problems was great.
- I would have preferred a balance between "proceed at your own pace" and guidance on the basics of each session, like some 5-10 minutes just to briefly explain how for instance 'loading data' works and the main purpose of the task.
- Although there was not much theory part, the support during the practical part was more than enough to make concepts very clear
- The theory and practical sessions were clearly laid out in the handbook and it was helpful to be able to work through the course content at my own pace. However, I would have found it helpful to have a tutor led introduction and example for each of the main sessions.
- They were excellent! And very applicable to my own work. Notes: I did not complete the extended practical, choosing instead to work on my own data with the techniques outlined

earlier in the course. I missed the first session at the course, but worked on the material in my own time.
- The manual was a bit unclear on what was hypothetical exercises

Knowledge and Understanding

On a scale of 1-5 where 1 is not at all and 5 is a great deal, all delegates felt their knowledge of R and RStudio had improved by attending this course by scoring 3 or above. The average rating was 3.54.

All but one of the attendees felt more confident in using the techniques taught following the course and all delegates bar one felt they were likely to use R and/or RStudio in the future in various ways-

- For my PhD project
- I think as a young researcher, there is an expectation for us to use R. It is a highly flexible system however with the steep learning curve a large time commitment is required initially which will be difficult.
- Unfortunately, I still don't feel using R would be more helpful than using excel for data cleaning. Perhaps because my data set isn't as large.
- I hope to but will need to go on an introductory course to do so.
- I would like to get to grips with RStudio as I can see it has advantages over other packages like SPSS. It will be a case of getting over the steep learning curve at the beginning.
- I use it already, but now can do so *much* more efficiently

The one delegate who stated they did not feel more confident in using the techniques taught and did not feel likely that they would use it in the future left the following comment:

- I am more familiar with R and its language but I don't feel confident in using independently

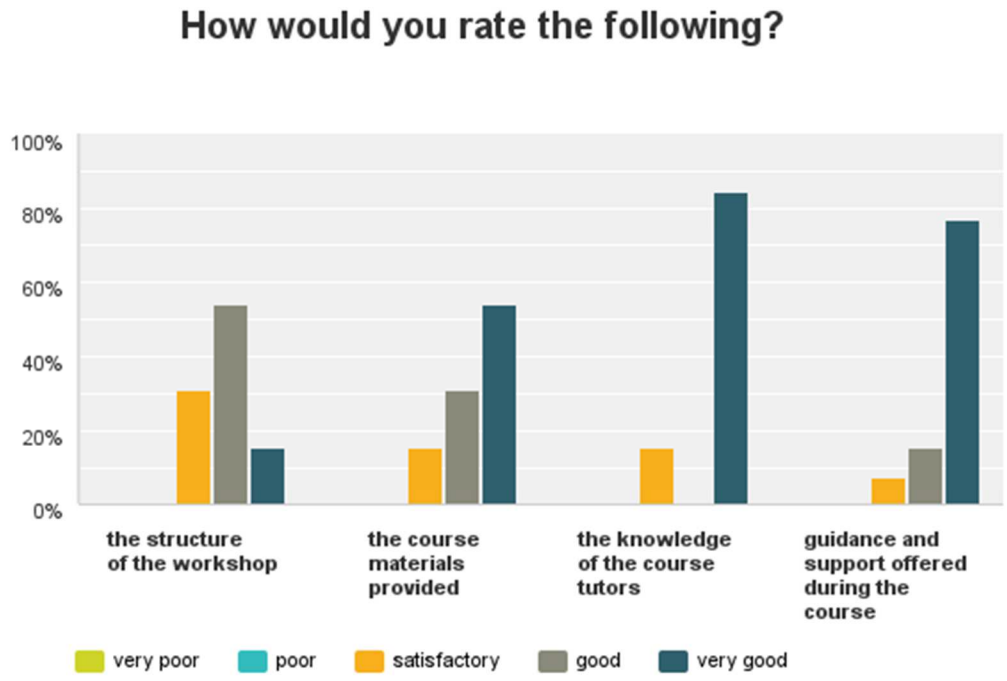
On a scale of 1-5 where 1 is not at all and 5 is a great deal, the majority of delegates (77%) felt the course was useful in terms of expanding their knowledge of Quantitative Methods. The average rating was 3.31.

Event Organisation

Delegates were also asked to rate the different aspects of the course including the structure of the workshop, course materials provided, knowledge of the tutor and guidance and support offered through the course.

All areas were rated by the majority as ‘very good’ or ‘good and in particular, delegates noted that the course leader’s knowledge was ‘very good’ (85%) -

- For a training like this, which is mainly hands on I would recommend having one or two more assistants to help students when they get stuck with the coding.
- The knowledge of the course leader, Jon Minton, was exemplary.
- Materials were great. Tutors were excellent and always at hand to provide help. 90% of the time was spent going through the handbook and I feel there could have been more time spent teaching.
- The course materials would maybe need to be updated on the basis of the different problems emerged from different configurations of the software. "compulsory" exercises shall maybe be more clearly highlighted in contrast to optional ones.
- As above - I think more tutor-led introductions to each session might have been helpful.
- As everyone was working on the materials at their own pace the tasks which called for discussion in groups were somewhat redundant, as we all reached these exercises at different points.



Conclusion

In conclusion 54% of attendees rated the workshop overall as ‘very good’, and the remaining as ‘good’ (15%) or ‘satisfactory’ (31%).

When asked to describe what they most enjoyed about the course, a lot of the feedback focused on the materials and the knowledge of the course leader–

- Enjoyed all sessions, especially practical session.
- Practical’s
- Quality of materials and help available
- Good tutors and learning environment
- Handbook was very high quality and the staff were always on hand to provide help.
- Material
- I enjoyed being introduced to R and R Studio having never used it before.
- Knowledge and help from the tutors
- The support of the other colleagues in the course. The help of instructors
- The piping technique
- Getting the code to work through many iterations. Working with other course attendees to problem solve.
- Tremendously useful and applicable to my work
- The easy way to find applications

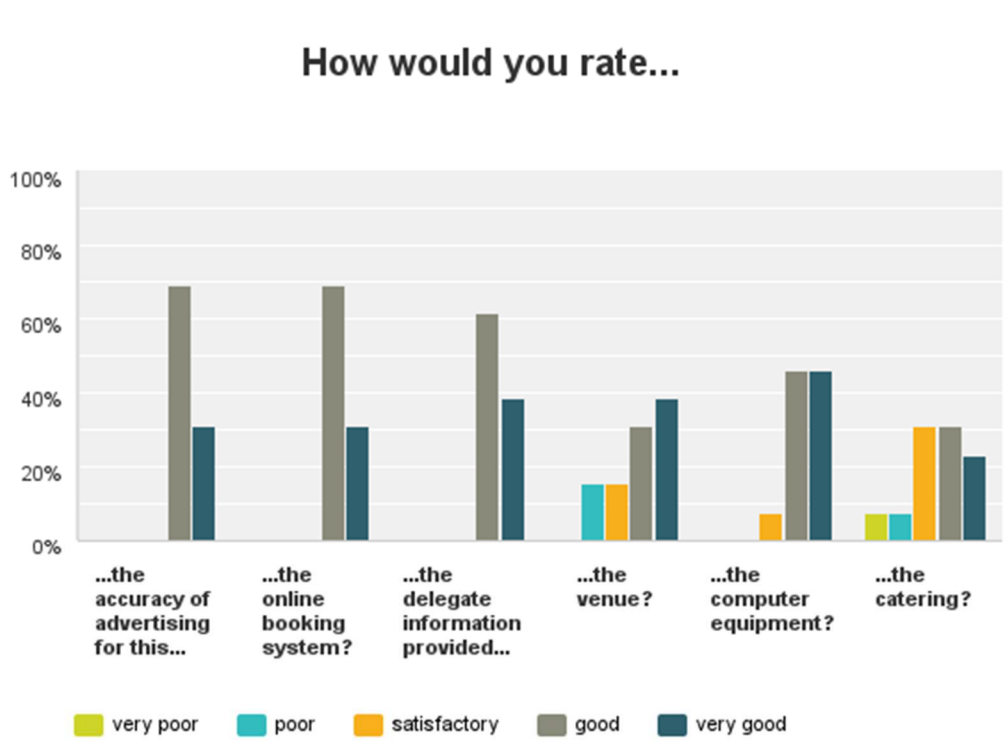
When asked what they enjoyed least about the course, some noted the time constraint of trying to work through so much material in the space of two days and others felt they would have benefitted from more teacher-led sessions–

- Nothing. I enjoyed all sessions.
- Lack of time to cover all the material
- It was a challenge to complete in the time available. Catering and break out environment were not great.
- as said earlier a little more warning about pre-course material

- There was too much material to cover, which increased the pressure to rush though material.
- Less teacher-led than expected.
- The course format did not work for me (though I'm sure it worked well for others) so it's not a criticism of the course (or course leaders).
- I found it a little tedious to working through the handbook and would have found some more tutor-led discussions/explanations of the exercises helpful.
- The venue
- I think this course have to be given in more than two sessions or in different modules as covering and understanding all the material in two sessions was impossible for me.

Future Courses

Delegates were also asked to rate overall the different aspects of the course including venue, catering, application process and information provided prior to the event. Overall the majority of the areas were rated ‘very good’, ‘good’, or at least ‘satisfactory’ however two delegates did rate the venue as poor and one rated the catering at poor and another as very poor–



Delegates were asked to provide comments on how this event could be improved in the future –

- Two days training period is short time to learn. My suggestion is that if you extend this training period it would be better.
- I think second afternoon session was quite long and with the room being stuffy and having no windows it contributed to the afternoon slump
- Make the manual a bit clearer

Lastly, there were several suggestions for follow up courses –

- Further training event should be on R but the topic should be different. Its better.
- I strongly suggest organizing a course on computer programming for social sciences as it is becoming a skill demanded by employers. Currently, the Consumer Data Research Center offers a Summer School on this topic. Why not reproducing this training!!

https://www.cdrc.ac.uk/training-session/cdrc-summer-school-computer-programming-for-social-scientists/

- I think a refresher in a years' time would be really useful.
- I think any R event will be well attended. I would like to attend an event on Machine Learning/Neural Networks. Also a big data session using R would be very useful and popular.
- It would be good to link this course to an introductory R course. Another general comments is that although my expectations for the course were not met, this was because my expectations were not realistic in light of what the course was trying to achieve. It is not a criticism of the course or course leaders.
- I think an introductory course to R would be a helpful per-requisite for absolute beginners like myself.
- A course for modifying datasets e.g. tranposing, taking certain info to add columns, using if arguments and loops.