Stefan Müller, PhD
Associate Professor
School of Politics and International Relations
University College Dublin
Belfield, Dublin 4, Ireland
✉ stefan.mueller@ucd.ie
https://muellerstefan.net

Level 4 Module; Autumn Trimester 2024

Applied Data Wrangling and Visualisation (POL42540)

Version: September 11, 2024

Latest version at: https://muellerstefan.net/teaching/2024-autumn-adwv.pdf

---

Time: Thursday, 09:00–10:50
Room:
– Weeks 1, 3–12: B002-CSI (→ location)
– Week 2: 114-VET (→ location)
Credits: 5.0
Format: Lecture and computer labs

Module Coordinator: Stefan Müller, PhD
stefan.mueller@ucd.ie | https://muellerstefan.net
Office: Newman Building, G312
Office hours: Wed, 13:00–13:45 (sign up here)
Teaching Assistant: Sarah King
sarah.king4@ucdconnect.ie | https://sarahaking.net

Course Content

Welcome to Applied Data Wrangling and Visualisation! The module offers a comprehensive introduction to the essential techniques and tools required for effective data management and visualisation in R. Students will also learn how to use AI tools as coding assistants, manage projects and handle data in various file formats, ensuring a robust understanding of data cleaning, wrangling, and merging.

The course emphasises the fundamentals of data visualisation, moving from principles to applied data visualisation strategies for compelling data storytelling. Additionally, it delves into the use of relational databases with SQL and data collection through web scraping, enabling students to manage and analyse large datasets efficiently. Applied Data Wrangling and Visualisation prepares module participants for a range of data-intensive roles, equipping them with the knowledge to leverage data visualisation and management tools effectively in their future careers.

Module Instructors

This module is taught by different instructors, each an expert in the topic covered during their assigned weeks.

- Stefan Müller (Fundamentals, R, Applied Data Visualisation)
- James P Cross (AI Tools)
- Jos Dornschneider-Elkink (SQL)

- Sarah King (Web Scraping)
- Mafalda Zúquete (Fundamentals of Data Visualisation)

Sarah King will serve as a Teaching Assistant for this module. She is the first point of contact for you. The communication in this module will take place on Slack, usually in channels that everyone can contribute to. More details on Slack are provided below.

Course Structure

Learning Outcomes

Upon successful completion of the course, students will be able to:

1. Manage and visualise data effectively in R, using VSCode and GitHub, enhancing your proficiency in handling data in diverse file formats.
2. Use relational databases (SQL) and conduct data collection through web scraping, enabling you to analyse large datasets for a range of data-intensive roles.
3. Develop skills in using AI tools as coding assistants, fostering your ability to manage projects efficiently and improve replicability in their work.
4. Robust understanding of data cleaning, wrangling, and merging techniques, preparing you for the challenges of managing large and complex datasets.
5. Solid foundation in the principles of data visualisation, learning to apply these strategies to create compelling data stories that are both intuitive and accessible.

General Readings

The seminar does not build on a single textbook, but relies on papers and book chapters. All readings used in this module are freely available online or accessible through the UCD Library.

- H. Wickham, M. Çetinkaya-Runde, and G. Grolemund (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd edition. Sebastopol: O'Reilly. URL: https://r4ds.hadley.nz
- N. B. Weidmann (2023). *Data Management for Social Sciences: From Files to Databases*. Cambridge: Cambridge University Press. URL: https://cambridge.org/9781108845670
- K. Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press. URL: https://socviz.co
- R. Alexander (2023). *Telling Stories with Data: With Applications in R*. New York: CRC Press. URL: https://tellingstorieswithdata.com
- C. O. Wilke (2019). *Fundamentals of Data Visualization: A Primer On Making Informative and Compelling Figures*. Sebastopol: O'Reilly. URL: https://clauswilke.com/dataviz/

Plagiarism

Although this should be obvious, plagiarism – copying someone else's text without acknowledgement or beyond 'fair use' quantities – is not allowed. Plagiarism is an issue we take very serious here in UCD. Please familiarize yourself with the definition of plagiarism on UCD's website* and make sure not to engage in it.

Individual Accountability and Group Contribution Assessment

Group work is a core part of this module because it cultivates essential skills such as collaboration, problem-solving, and communication, which are vital for professional success. The policy on individual accountability within groups is designed to prevent issues related to uneven contributions in group assignments. Each group member is required to submit a document along with their group project that clearly specifies their individual contributions. This document must transparently detail each student's roles, responsibilities, and actual input. While the aim is to award the same grade to all members of the group, divergent contributions will result in varied grades to ensure fairness. This policy not only promotes equitable grading but also deters disparities in the distribution of workload among team members. Students are encouraged to report any issues with group dynamics *promptly* to facilitate swift intervention and support from the course instructor.

Module policy on the Use of Artificial Intelligence (AI) Tools

I encourage the use of AI tools when completing the assignments for this module, but users of this technology must be aware of what it can and more importantly, what it cannot do well. It is crucial for you to exercise judgement when evaluating the quality and reliability of content generated through AI platforms. AI is not a panacea for all writing challenges; it will not automatically generate a flawless, logically coherent assignment. Instead, use AI as a tool to tackle specific issues such as brainstorming and idea formation, literature discovery, and text drafting issues.

*https://libguides.ucd.ie/academicintegrity.

View your preferred AI platform(s) as useful but imperfect tools that can offer inspiration, new perspectives, and supplementary areas for research for your own work. In-depth research on your part remains essential to ensure coherent, factual, and scientifically informed perspectives in your assignment. Always cross-reference the information AI offers against other independent and reliable sources.

Late Submission Policy

If a student or group submits an assignment late, the following penalties will be applied:

- Coursework received at any time within two weeks of the due date will be graded, but a penalty will apply.
  - Coursework submitted at any time up to one week after the due date will have the grade awarded reduced by two grade points (for example, from $B-$ to C).
  - Coursework submitted more than one week but up to two weeks after the due date will have the grade reduced by four grade points (for example, from $B-$ to $D+$). Where a student finds they have missed a deadline for submission, they should be advised that they may use the remainder of the week to improve their submission without additional penalty.
- Coursework received more than two weeks after the due date will not be accepted. Regulations regarding extenuating circumstances apply.

Questions and Problems

In this module, we will discuss concepts, methods, and software you might not have heard of before. I am aware that parts of this module could be challenging, and I will assist you as best as I can.

We will use Slack for in this module.[†] Make sure to create a Slack account before the first seminar and join the Slack workspace. If you have a question that involve code or concepts, please share your question in #coding, #homework, or #research-paper.

If you struggle to solve problems relating to R, Python or SQL, please follow the steps outlined below before contacting your peers or me. It is very likely that at least one other person faced the same problem before or received the same error message.

1. You are welcome and encourage to use AI tools as coding assistants. Recall that GitHub Copilot is a powerful coding assistant.
2. Try to summarise the problem in your own words and then google this summary or use an AI tool, for example ChatGPT. If the problem relates to R, add `rstats` to your search query. For example: `how to import csv file in rstats`. I am almost certain that you will find a solution to most of your questions.
3. If your R code returns an error, I would advise you to google the text of the error message. For example, you can google the error message "`Error: Can't subset columns that don't exist.`"

$\longrightarrow$ If steps 1–3 still do not solve your problem or question, please ask your question in the Slack channel devoted to this module. Your peers and we will help you.

[†]I have had very positive experiences with Slack in my modules. Müller (2023) discusses both the advantages and shortcomings of Slack for teaching and learning.

Contact and Office Hours

Sarah King, our Teaching Assistant devoted to this module, is your first point of contact. If you have questions about software, code, or assignments, please use Slack first (see more details below) or check if one of your peers already asked this question. We monitor Slack regularly and will respond to your question if other module participants cannot help you.

In addition, Stefan Müller also offers office hours on Wednesdays from 13:00–13:45, either in in person (Room G312, Newman Building) or online. Please sign up for a meeting at https://calendly.com/mueller-ucd/office-hours.

Proper communication practices are crucial for formal email exchanges. Please familiarise yourself with this summary[‡] on professional, efficient, and respectful communication.

Software

In this module, we will run R and SQL code in the code editor VSCode. We will also use GitHub Copilot (free to use for students).

On Brightspace, we have uploaded a detailed guidelines on installing the relevant software and extensions (see the `Installation Instructions` page under `My Learning`).

Important: Make sure to install and set up R, VSCode, and GitHub Copilot *before* our first lecture in Week 1. If you have any question related to (installing) the software, please ask this Question in our Slack workspace.

Syllabus Modification Rights

I reserve the right to reasonably alter the elements of the syllabus at any time by adjusting the reading list to keep pace with the course schedule. Moreover, I may change the content of specific sessions, depending on the participants' prior knowledge and research interests. If I make adjustments, I will email all seminar participants and upload the revised syllabus to Brightspace.

Dignity and Respect

UCD is committed to the promotion of an environment for work and study which upholds the dignity and respect of all members of the UCD community and which supports your right to study and/or work in an environment which is free of any form of bullying, harassment or sexual misconduct (including sexual harassment and sexual violence).

There are a number of supports in place if you are experiencing bullying, harassment or sexual misconduct and you are strongly encouraged to come forward to seek confidential support and guidance on the range of informal options and formal options for resolving issues as appropriate. Reports of bullying, harassment or sexual misconduct can also be made anonymously through UCD's Report and Support tool.

UCD is actively promoting a culture where bullying, harassment and sexual misconduct is not tolerated, where everyone is respected and feels valued, included and that they belong in UCD.

You can find more details on UCD's Dignity and Respect Website at: https://www.ucd.ie/equality/support/dignityrespect/.

[‡] https://www.beyondberea.org/blog/2022/05/03/email-etiquette-for-students/.

Expectations and Grading

Table 1: Overview of assignments and deadlines

Date	Assignment
Week 7 (24 Oct)	Multiple Choice Questionnaire #1: 25%
Week 11 (21 Nov)	Multiple Choice Questionnaire #2: 25%
Week 12	Data Report: Descriptive analysis and visual representation of insights derived from a large dataset: 50%

Week 1: Introduction to R, VSCode, and GitHub (12 September)

Instructor: Stefan Müller

Mandatory Readings:

- H. Wickham, M. Çetinkaya-Runde, and G. Grolemund (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd edition. Sebastopol: O'Reilly: chapters 2 and 27.
- C. Ismay and A. Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press: chapter 1.

Important: Please follow the installation guide (see the `Installation Instructions` page under `My Learning`), available on Brightspace, in order to set up R, VSCode, and GitHub Copilot. Make sure to install the software *before* our first lecture. If you have any question related to (installing) the software, please ask this question through our Slack workspace.

Week 2: Software, Project Management, and Replicability (19 September)

Instructor: Stefan Müller

Location (this week only): 114-VET, UCD Veterinary Sciences Centre (→ location)

Mandatory Readings:

- H. Wickham, M. Çetinkaya-Runde, and G. Grolemund (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd edition. Sebastopol: O'Reilly: chapters 3 and 5.

Optional Readings:

- J. Bryan (2020). *Happy Git and GitHub for the useR*. URL: https://happygitwithr.com (skim if you are interested in version control).
- W. G. Jacoby and R. N. Lupton (2016). *American Journal of Political Science: Guidelines for Preparing Replication Files. Version 2.1*.

Week 3: Using AI Tools as Coding Assistants (26 September)

Instructor: James Cross

Mandatory Readings:

- **TBA**

Optional Readings:

- GitHub Docs (2024). *Asking GitHub Copilot Questions in your IDE*. URL: https://docs.github.com/en/copilot/using-github-copilot/asking-github-copilot-questions-in-your-ide.
- S. Verdi (2024). *How to Use AI Coding Tools to Learn a New Programming Language*. URL: https://github.blog/developer-skills/programming-languages-and-frameworks/how-to-use-ai-coding-tools-to-learn-a-new-programming-language/.

Week 4: Importing, Summarising, Transforming, and Merging Data (3 October)

Instructor: Stefan Müller

Mandatory Readings:

- H. Wickham, M. Çetinkaya-Runde, and G. Grolemund (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd edition. Sebastopol: O'Reilly: chapters 7 and 8.

Week 5: Principles of Data Visualisation (10 October)

Instructor: Mafalda Zúquete

Mandatory Readings:

- C. O. Wilke (2019). *Fundamentals of Data Visualization: A Primer On Making Informative and Compelling Figures*. Sebastopol: O'Reilly: chapters 2, 17, 19, and 29.

Optional Readings:

- C. O. Wilke (2019). *Fundamentals of Data Visualization: A Primer On Making Informative and Compelling Figures*. Sebastopol: O'Reilly: chapters 1, 4, 18, 20, 26.

Week 6: Explorative Data Visualisation (17 October)

Instructor: Stefan Müller

Mandatory Readings:

- K. Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press: chapters 1, 3, and 4.

Week 7: Advanced, Intuitive, and Accessible Data Visualisation (24 October)

Instructor: Stefan Müller

Mandatory Readings:

- K. Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press: chapters 5 and 8.
- C. O. Wilke (2019). *Fundamentals of Data Visualization: A Primer On Making Informative and Compelling Figures*. Sebastopol: O'Reilly: chapters 15, 17, 19, 20, and 22.
- J. P. Kastellec (2023). *Practical Advice for Producing Better Graphs*. URL: https://jkastellec.scholar.princeton.edu/sites/g/files/toruqf3871/files/documents/kastellec_graphs_practical_tips_0.pdf.

Optional Readings:

- J. L. Steenwyk and A. Rokas (2021). "ggpubfigs: Colorblind-Friendly Color Palettes and ggplot2 Graphic System Extensions for Publication-Quality Scientific Figures". *Microbiology Resource Announcements* 10 (44): e00871–21

Week 8: Relational and Non-Relational Databases (31 October)

Instructor: Jos Dornschneider-Elkink

Mandatory Readings:

- **TBA**

Optional Readings:

- **TBA**

Week 9: SQL: Organising and Managing Data (7 November)

Instructor: Jos Dornschneider-Elkink

Mandatory Readings:

- **TBA**

Optional Readings:

- **TBA**

Week 10: Introduction to Web Scraping (14 November)

Instructor: Sarah King

Mandatory Readings:

- R. Alexander (2023). *Telling Stories with Data: With Applications in R*. New York: CRC Press: chapter 7.

Optional Readings:

- **TBA**

Week 11: SQL: Retrieving, Joining, and Summarising Data (21 November)

Instructor: Jos Dornschneider-Elkink

Mandatory Readings:

- **TBA**

Optional Readings:

- **TBA**

Week 12: Advanced Web Scraping (28 November)

Instructor: Sarah King

Mandatory Readings:

- H. Wickham (2022). *rvest: Selector Gadget*. URL: https://rvest.tidyverse.org/articles/selectorgadget.html.

Optional Readings:

- **TBA**

References

Alexander, R. (2023). *Telling Stories with Data: With Applications in R*. New York: CRC Press.

Bryan, J. (2020). *Happy Git and GitHub for the useR*. URL: https://happygitwithr.com.

GitHub Docs (2024). *Asking GitHub Copilot Questions in your IDE*. URL: https://docs.github.com/en/copilot/using-github-copilot/asking-github-copilot-questions-in-your-ide.

Healy, K. (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press.

Ismay, C. and A. Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press.

Jacoby, W. G. and R. N. Lupton (2016). *American Journal of Political Science: Guidelines for Preparing Replication Files. Version 2.1*.

Kastellec, J. P. (2023). *Practical Advice for Producing Better Graphs*. URL: https://jkastellec.scholar.princeton.edu/sites/g/files/toruqf3871/files/documents/kastellec_graphs_practical_tips_0.pdf.

Müller, S. (2023). "How Slack Facilitates Communication and Collaboration in Seminars and Project-Based Courses". *Journal of Educational Technology Systems* 51 (3): 303–316.

Steenwyk, J. L. and A. Rokas (2021). "ggpubfigs: Colorblind-Friendly Color Palettes and ggplot2 Graphic System Extensions for Publication-Quality Scientific Figures". *Microbiology Resource Announcements* 10 (44): e00871–21.

Verdi, S. (2024). *How to Use AI Coding Tools to Learn a New Programming Language*. URL: https://github.blog/developer-skills/programming-languages-and-frameworks/how-to-use-ai-coding-tools-to-learn-a-new-programming-language/.

Weidmann, N. B. (2023). *Data Management for Social Sciences: From Files to Databases*. Cambridge: Cambridge University Press.

Wickham, H. (2022). *rvest: Selector Gadget*. URL: https://rvest.tidyverse.org/articles/selectorgadget.html.

Wickham, H., M. Çetinkaya-Runde, and G. Grolemund (2023). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd edition. Sebastopol: O'Reilly.

Wilke, C. O. (2019). *Fundamentals of Data Visualization: A Primer On Making Informative and Compelling Figures*. Sebastopol: O'Reilly.