



Moving from Excel to Scripting

Nick Rochlin, University of Victoria
RDM Jumpstart
May 13, 2025

Optimist: The glass is $\frac{1}{2}$ full.
Pessimist: The glass is $\frac{1}{2}$ empty.
Excel: The glass is January 2nd.

Have you ever used Excel?

- What did you use it for?
- Did it do a good job? Why/why not?

Diving Into Excel: What it does well



- Intuitive point-and-click graphical user interface (GUI)
- Easily group and order datasets based on specific values
- Easily calculate things like, sum, mean, and other basic statistical methods
- Create graphs and charts with a few clicks
- TLDR: **it's easy and convenient.**

Diving Into Excel: What it doesn't do so well



- Any manual changes made to a document are subject to human error, and tracking these changes and identifying errors can be challenging to reproduce.
- Formulaic changes are often hidden in cells and are subject to “breaking”, making any manipulations and analyses difficult to track and reproduce.
- Manipulations require manual processes for every file.

Diving Into Excel: What it doesn't do so well



- The auto-formatting of cells can be very annoying and cause issues.
- Rows/columns can be hidden, which can cause issues with collaboration.
- Changing variables will overwrite original data, causing issues with reproducibility and provenance.
- TLDR: **it's not very reproducible.**

What's in a Spreadsheet?

.xlsx vs .csv files

What's in a spreadsheet? .xlsx files



- Excel files (.xlsx) are a proprietary format, meaning that the ability to use them relies on a subscription to the Microsoft suite.

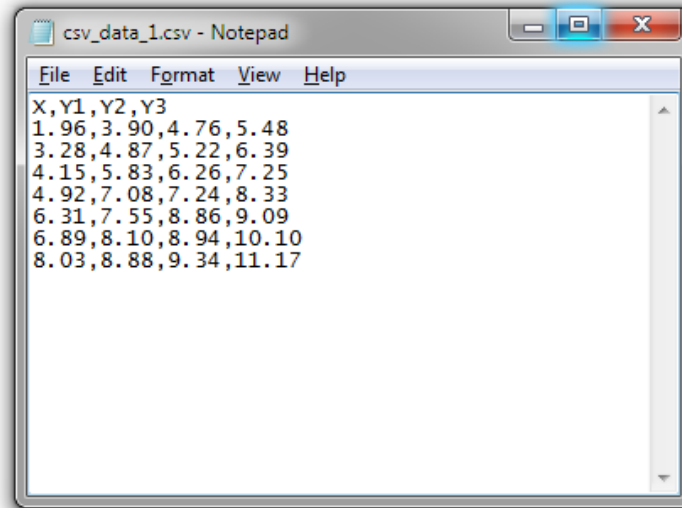
What's in a spreadsheet? .xlsx files



- While it's natural to think of an Excel file as a single file, they are actually zipped archives containing multiple underlying files:
 - Worksheet data
 - Worksheet formatting
 - Charts and graphs
 - Pivot tables
 - Embedded objects
 - External links
 - Metadata

What's in a spreadsheet? .csv files

- .csv file (Comma Separated Values), are tabular (spreadsheet) files, with each row representing a record and values within each row separated by commas (even though they can be presented in cells).



What's in a spreadsheet? .csv files



- **They are plain text files**
 - They don't have any underlying structures like Excel files.
- **They are non-proprietary**
 - Can be opened using any simple and free text editor like Notepad, TextEdit, etc., and work easily with coding software.

What's in a spreadsheet? .csv files



- Because of their simplicity, they are an efficient way to store and transfer tabular data because they are smaller in size, but with this they have limited formatting options.
- While it is possible to handle Excel files in coding languages like R and Python, .csv files are generally the preferred format due to their open and simple nature.



Original value/format	What can happen	Example of change	Why it happens
01234	Leading zero lost	1234	Treated as a whole number
Résumé	Characters corrupted	Rsum / RÃ©sumÃ©	Encoding issues
31/12/2024	Date misread or changed	12/31/2024	Different regional settings for dates
=SUM(A1:A10)	Formula lost	55 (value only)	Plain text only keeps values
Hello	Formatting lost	Hello	Plain text doesn't store formatting
123456789012345	Rounded or scientific notation	1.23E+19 or rounded value	Precision limits

Example: Genetic Data



“In a survey looking at over 11,000 papers with Excel gene lists published between 2014 and 2020, more than 30% contained at least one gene name error caused by Excel’s auto-formatting.

This was due to gene names being converted to standard dates, internal data numbers (5 digits), and floating-point numbers.”

<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008984>


Moving to Scripting

Moving to Scripting: What is it?



- Writing lines of code, or “scripts”, is creating a set of instructions that a coding language can perform.
- For the purpose of the Jumpstart, these instructions are performed on a dataset (noting there can be broader usage).

Moving to Scripting: What is it?

- 
- Much like instructions written in human language, instructions written in coding languages have characters and strings of letters with semantic meaning, and a specific grammatical pattern that needs to be followed.
 - The instructions you create can be quite simple (ex: a cookie recipe), and can scale up to the very complicated (ex: building a bridge).

Literate Programming



- Scripting was just referenced as instructions for a coding language/computer to perform, which is true, but a big focus of this series is making our work reproducible, which needs to include a second element of **instructions for humans to interpret.**

Literate Programming



- Literate programming is a framework that provides a human-language explanation of how a script works in combination with the script itself, so people can accurately interpret and reuse the script.
- We'll be discussing various forms of documentation throughout the week, and part of this will include documentation of our code and literate programming principles.

Moving to Scripting: Why you wouldn't use it



- Learning a coding language is like learning a human language, and can be a painful process.
- Because the things we do with tabular data are often isolated and time-sensitive activities (“I need this now!”), the effort needed to learn how to perform a new task can take much longer than doing the task in Excel.

Moving to Scripting: Why you wouldn't use it

- Even when you are familiar with a coding language, it can still be a struggle to figure out the exact way to do something.
- TLDR: it can be a pain in the a**.

Moving to Scripting: Why you would use it



- Other than the learning curve, scripting overcomes all the downfalls of Excel:
 - Every action can be tracked and checked.
 - There is a clear and ordered list of everything that was done.
 - Can be performed across large amounts of similar files.
 - Can be altered to perform similar tasks on different datasets.
 - No hidden rows or columns.

Moving to Scripting: Why you would use it



- The actions you can perform with scripts are far more precise and powerful than what Excel is capable of.
- TLDR: it's **powerful and reproducible**.

Scripting in the RDM Jumpstart

Scripting in the RDM Jumpstart



- The goal is to learn the foundations of scripting in R, including key semantic functions, general grammar, and documentation, to promote the transparency and reproducibility of your research.
- While we have a set curriculum and activities for you to perform, we encourage you to be curious and explore!

Scripting in the RDM Jumpstart



- Much like learning a human language, if you don't regularly use it, your ability to use it can diminish.
- You won't leave this week being an expert in R, but you will have all the tools to engage with coding and to continue building your skills.

Questions?