Make DaRTS data R friendly

Yesterday, we started with a nice CSV file from one DaRTS cruise that we didn't need to format and could just load directly into R. In practice, the CTD data David shares with you from the cruises will NOT be in that format. You'll need to reformat it a bit before it can be loaded into R. The main purpose of this lesson is to make R friendly DaRTS data.

We will learn:

1. Best practices for organizing data in a spreadsheet
2. Create spreadsheets that are R-readable
3. Export data from spreadsheets into R

Formatting Data

Download the spreadsheet from 2024 Cruise 1 to your working directory and open it in Excel, Numbers or another spreadsheet viewing software. This is the type of spreadsheet David will share with you after each cruise.

- What are the strengths of this Excel file?
- What do you see as some weaknesses of it?
- Could you load this sheet into R as-is?

Generally, let's discuss:

- What makes a good spreadsheet?
- What are some common pitfalls of spreadsheets?

Spreadsheets for R

General tips for creating an R-readable spreadsheet:

- Keep it simple
- One header row
- Keep a separate tab or text file describing your data in detail.
- Keep your original files!!

Things to consider:

- Split date columns by year, month and day to avoid confusion.
- Keep data type consistent within each column (e.g. all numbers)
- Include columns for all variables, even if they are split up by sheets (e.g. by station/year etc).
- Keep columns consistent between sheets (to the best of your ability!).

- Avoid complicated headers (minimize keystrokes during analysis!).
- Don't carry out any cell operations within the spreadsheet... save that for R!

- Same applies for plotting!

Final tip: Make a plan about data collection at the *beginning* of a project!

Cleaning up our DaRTS data

Note that:

- Each tab represents a variable (station)
- File ID also provides important information (Cruise date, Project ID, Cruise Number)

Cleaning steps

1. Open a new, empty spreadsheet file, name it appropriately (e.g., I named mine `DaRTS_SingleCruise_CTDdata.xlsx`) and save to your sea change R lab directory
2. Identify columns and column names (hint: for this data I recommend 15 columns)
3. Rename columns with simple headers
4. Copy and paste data from the 2024 Cruise 1 spreadsheet to your new spreadsheet, noting date and station number as data is transferred
5. Add a new tab and add some notes on what you just did, including listing the old column headers and what you replaced those with. This is your "README" tab, for when you come back to this file and want to remember where the data came from, and if any processing had been done to it.
6. Export data as a CSV (details below)

Saving as a CSV

If you opened the file in Excel:

1. Go to "File" -> "Save As" and select to save the file as a "CSV (Comma delimited)" file named DaRTS_CTD_data (selecting "yes" " to the windows that pop up)
2. Exit the Excel file (selecting "no" you don't want to save changes if asked.)

If you opened the file in Numbers:

1. Go to "File" -> "Export To" -> CSV
2. Check the box that says "Create a file for every table". You are going to create a CSV file for every sheet (table) in the file. You will be asked to name the folder all the CSV files will be saved in - pick something sensible!
3. Once saved, exit the numbers file.

But what about when we do our next cruise (and the one after that. . .)?

1. Open the R friendly Excel file we made above (it contains the README tab)
2. Open the new cruise CTD Excel file from David
3. Copy and paste the new cruise data onto the bottom of the Cruise 1 data, noting data, cruise and station number as before.
4. Edit the README tab so it states what files / cruises this file now contains.
5. Export data as a CSV.

Final Comments

What do you think about that process? Was it foolproof?

- We added a README to help us track everything back to the original data.
- We did a lot of copy and pasting: there's room for error here!
- We did some manual typing: there's room for error here!

Is there a better approach?

We could have created our csv / data table directly in R, by reading all the different sheets, adding in columns for dates (taken from the file name), station, etc. This would be the most reproducible approach

(and generally what I would recommend), but would have involved a lot of data wrangling skills that we aren't familiar with.

But - is there an even better approach?

We could set ourselves up for success from the start: we could create our original spreadsheets in an already R-friendly way, so we can open and read them easily within R. Or we could even work directly with the raw data in R, and create a whole data processing pipeline or workflow that can be implemented on every CTD file.

Think about how you want to save your data for your independent research - make it easy for yourself when you come to do your data analysis!