2A: Data Cleaning Review

Today we're going to be learning how to ask a question and figure out how to set up our data to answer it. To do so we're going to have to quality control and wrangle historical CTD and discretely measured data from DaRTS.

Make a CTD Data CSV

First, let's load in a data table. This time, instead of using cruise data from this year, we are going to process and load a spreadsheet that has data from multiple years of DARTS cruises from the excel table in the R google drive called: DARTS_cruise_data_2012-2020 - 20201103.xlsx

Please open up this excel table in either excel or google drive, and create a new table for R that reformats the data within the tab "CTD all years" into a csv file that can be read by R.

In your spreadsheet program (Excel or Google Sheets) change the headers to the following column names:

Date
Station
Depth_m
Conductivity
Temperature_C
Salinity_PSU
Density
PAR
Fluorescence
TurbidityNTU
BeamC
O2Conc
O2Saturation

Save the table to you project directory as "DaRTS_CTD_data.csv" in your R project directory.

Make discrete data CSV file

We've previously created a CSV file with all the CTD data from `DaRTS_cruise_data_2012-2020 - 20201103.xlsx`. Before combining our CTD data with the discrete data, we need to create a CSV file with the discrete data.

1. Open `DaRTS_cruise_data_2012-2020 - 20201103.xlsx` (the full data set file) in Excel
2. Go to the "Depth Discrete Data" sheet
3. Go to "File" -> "Save As" and select to save the file as a "CSV (Comma delimited)" file named DaRTS_discrete_data (selecting "yes"" to the windows that pop up)
4. Exit the Excel file (selecting "no" you don't want to save changes if asked)
5. Open the csv file you just created in Excel (the easiest way to do this is open Excel, then do "File" -> "Open" -> "Browse", then navigate to the file and select "Open")

Once you've opened the CSV file

1. Delete the second PicoPlankton Conc column
2. Combine the information in the first three rows into the one row by replacing the spaces, parentheses and forward slashes with underscores
3. Delete the extra rows at the top of the sheet
4. Change the dates `10/1/2013 and 10/4/2013` to `10/1/2013`
5. Press CTRL-F on Windows, or the Command-F on a Mac, to bring up the "Find and Replace" window.
6. Go to the Replace tab, enter "Cruise" in the "Find What" box and nothing in the "Replace with" box, hit "Replace all"
7. Repeat the previous step for the word "NA"
8. Repeat the previous step for the word "N/A"
9. Repeat the previous step for the word "BLOS" and close the window
10. Delete the rows with "Pier" as a Station value (only occur in 2012)
11. Rename the blank column (i.e. the 2nd column) to "Cruise"
12. Save as DaRTS_discrete_data.csv and close the CSV file.