EVR 628- Intro to Environmental Data Science

Assignment 2: Data wrangling

Juan Carlos Villaseñor-Derbez (JC)

The big picture

Remember that the final goal is to have a GitHub repository where you can showcase your work. Assignment 1 was to create the repository, which should be mostly empty. Assignment two is to start developed one R script to clean some data. Your choice of data will lock you in a path: The third assignment will be to visualize the data you have cleaned up here. Your final project will leverage the data and visualizations you'll produce to wrap it all together.

This assignment

Task: Develop one data wrangling script that reads raw data (*e.g.* in `.xlsx` or `.csv`), performs data cleaning, transformation, or wrangling, and exports a processed version of the data (`.rds`) to be used in later assignments.

This assignment will require quite a bit of work on three fronts:

1. Coming up with an idea for a final project
2. Thinking about how you will get the raw data into a format that gets you closer to your final project, and
3. Using lecture and live coding session materials –and function documentation– to help you build your pipeline.

I have included a list of potential data sources and ideas for final projects. Feel free to come up with your own ideas, I recommend choosing something you may be genuinely interested in.

Grading Rubric

- 25%: `README.md` file:
  - 10% Explains the main objective of your project. 2-3 sentences is enough.
  - 5% Lists the contents of the project repository.
  - 10% Lists the column names of the clean data, and includes information on data type and a description of the column.

- 50%: Your repository contains and R script called `data_processing.R`, saved to your `scripts/01_processing` folder that:
  - 10%: Contains code documentation using comments `#`
  - 5%: Clearly indicates packages loaded at the top of the script
  - 5%: Uses **relative** paths to read / write data files
  - **20%: Uses `dplyr` verbs and / or `tidyr` functions to achieve the data cleaning task.** Note that simply renaming columns or changing their order is not enough. If your raw data happen to be in the perfect format, consider starting your analysis instead. If you decide to do this, your script should be named `analysis.R` and saved in the corresponding folder.
  - 10%: The exported data conform to `tidy` data standards[1].
- 25%: I can clone your repo and reproduce your data without having to modify your code

Some examples of repositories that would get a 100%:

- Analyzing mid-term survey data
- Building a geospatial dataset of Mexican ports, by McGill undergraduate student Emma Zgonena

Turning in your assignment

- Please share the link to your github repo via Canvas
- The deadline for this assignment is October 19, 2025 by 11:59

Project ideas with different data sets

Below are examples of data sources and final project ideas. For the purpose of this assignment you are not expected to provide an answer. You will simply prepare the data necessary to eventually provide the answer.

Tuna Data

Access tuna data from the WCPFC or IATTC and then:

- Build a time series of catch-per-unit-effort for a species of your liking
- Find the year in which total tuna catch peaked
- Find the month where tuna catch is, on average, maximum
- For every year, find the coordinates where tuna catch was the highest. Do we see any changes in it's location?

[1] If you are using this assignment as part of your own research and you have a legitimate reason not to use a `tidy` data format, that is ok. Just make sure you include a justification in your code documentation and `README.md` file

Oceanic Indices

Go to NOAA's Physical Sciences Laboratory and find an interesting data set

- Use monthly NINO3 to produce a figure of NINO / NINA years
- Compare indices among them
- Relate indices here to tuna catch data above

STRAVA data

Export your data from strava[2], and then:

- Find your longest activity
- Find your total traveled distance since you started logging activities
- Find the average length of your runs
- Analyze your trends in heart rate vs speed
- Analyze the pace of runs performed with the Rosenstiel Running Club

Dive watch data

Your dive watch should allow you to export your data[3], although you might need a special cable / software.

- Calculate how much time you've spent underwater per week, month, or year
- If you have air pressure info associated (or you keep a dive log with it), calculate your Surface Air Consumption Rate
- What's your longest dive?
- What's your deepest dive?
- What is the average depth of all your dives?
- What is the maximum number of dives you have done in a single day? a week?

Data from papers

The following papers have data available. You could download their data and recreate or augment their figures.

- Burguess et al, 2018. Protecting marine mammals, turtles, and birds by rebuilding global fisheries | GitHub repo
- Kuczenski et al. 2021Plastic gear loss estimates from remote observation of industrial fishing activity | GitHub repo
- Clark et al., 2023 Global assessment of marine plastic exposure risk for oceanic birds | GitHub repo
- A database of mapped global fishing activity 1950–2017

[2]Data will not exported as `.csv` but I can help work around that
[3]The data may not exported as `.csv` but I can help work around that

- O'Connor et al., 2024 Effects of anthropogenic noise on marine mammal abundances informed by mixed methods
- Jouffray et al., 2025. Identifying and closing gaps in corporate reporting of ocean impacts
- Oyanedel et al., 2025. Improving detectability of illegal fishing activities across supply chains | GitHub repo

Other sources of data

- The National Center for Environmental Information NCEI contains loads of interesting data sets.
- Animal tracking data from MoveBank
- Vessel tracking data from Global Fishing Watch

Synthetic data

- Maybe you are in the midst of collecting data for your project. You don't yet have the raw data, but you have a sense of what they will look like (number of columns, number of observations…). I can help you simulate the data and you can build a script that helps you get your data into an analysis-ready format.

None of these are working for you?

Come to office hours and we can discuss ideas.