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 `portfolio` 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 adding R code to clean some data. The third assignment will be to visualize the data you have cleaned up here, so keep that in mind. Your final project will then culminate in an analyses that leverages the data and visualizations you'll produce.

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 (as an `.rds` file) 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 wild data into a format that gets you closer to your final project, and 3) using lecture materials, live coding sessions, and function documentation to help you build your pipeline.

Grading Rubric

- 50%: Your repository contains and R script called `data_processing_script.R` that:
  - 5%: Clearly indicates packages loaded
  - 5%: Uses relative paths to read / write data files
  - 20%: Uses `dplyr()` verbs and `tidyr()` functions. Note that simply renaming columns or changing the order of columns is not enough. You must use at least 3 of the following[1]:
    1. `filter()`
    2. `mutate()`

[1] or any of their advance cousins like `filter_at()`

        3. `group_by()` and `summarize()`
        4. `pivot_longer()` or `pivot_wider()`
        5. A type of `*_join()`
    - 10%: The exported data conform to `tidy` data standards[2].
    - 10%: Contains code documentation using comments `#`.
- 25%: I can run your code and reproduce your clean data.
- 25%: README file:
    - 10% Explains the main objective of your project. 2-3 sentences is enough.
    - 15% Lists the column names of the clean data, and includes information on data type and a description of the column.

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

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 from the WCPFC or IATTC

- 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

Data from the National Center for Environmental Information NCEI

STRAVA data[3]

- 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

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

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

Dive watch data[4]

- 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.

- Jouffray et al., 2025. Identifying and closing gaps in corporate reporting of ocean impacts
- Clark et al., 2023 Global assessment of marine plastic exposure risk for oceanic birds | GitHub repo
- Burguess et al, 2018. Protecting marine mammals, turtles, and birds by rebuilding global fisheries | GitHub repo
- A database of mapped global fishing activity 1950–2017
- Kuczenski et al. 2021Plastic gear loss estimates from remote observation of industrial fishing activity | GitHub repo
- Oyanedel et al., 2025. Improving detectability of illegal fishing activities across supply chains | GitHub repo
- O'Connor et al., 2024 Effects of anthropogenic noise on marine mammal abundances informed by mixed methods
- Any other paper of your liking

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[5] and you can build a script that helps you get your data into an analysis-ready format.

Other data sources:

- Animal tracking data from MoveBank
- Vessel tracking data from Global Fishing Watch

[4] The data may not exported as `.csv` but I can help work around that

[5] or ask chat GPT for it

None of these are working for you?

Come to office hours and we can discuss ideas.