EDS 220 - Homework 1

Task 1: Microsoft Planetary Computer Access Request

Later in the course we will work on Microsoft's Planetary Computer (MPC). The MPC is both a data catalog and a development environment to work with large scale environmental data. You will need to request access to it to use it for this course (I already checked with the MPC team the class can sign up for it).

To complete this activity:

1. Go to the MPC's data catalog and make a note of which datasets interest you the most. You will need this info to request access. If you're interested in previewing a specific dataset you can try loading it on the MPC's Explore panel.
2. Go to: https://planetarycomputer.microsoft.com/account/request.
3. Scroll down and use the following options to request access:

- **Email:** use your @bren email address
- **Name:** your name, of course
- **Affiliated Organization:** UC Santa Barbara - Bren School of Environmental Science and Management
- **Sector:** Academia
- **Primary programming languages:** R, Python (plus any others that apply to you)
- **Country:** United States
- **What datasets are you intersted in?:** Mention some datasets in the MPC data catalog you are interested in and any broad type of data that interests you (ex: remotely sensed, climate, biodiversity, demographics, etc.)
- **What is your area of study?:** Environmental Data Science - Currently enrolled in UCSB's Masters in Environmental Data Science and will use the MPC for data analysis in my courses. *(Or something similar to this)*

4. Agree to terms of use and click submit. That's it.

Task 2: Datasheets for Datasets Paper Reading

So much goes into creating a dataset, and data is more than numbers and words in a file. Without a proper understanding of the whole context where data was created, biases, omissions, and inacuracies can go undetected. The Datasheets for Datasets paper by T. Gebru et al. advocates for transparency about the purpose and contents of datasets.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

Read the paper and write a one-paragraph (between 100 and 150 words) open reflection about it. You may use the following questions as prompts for your reflection, but feel free to discuss any topic in the article that caught your interest.

- Do you know of an example where understanding the context in which data was developed was crucial for appropriately analyzing it? Maybe you've been in such a situation.
- How does a framework like the propsed "datasheets for datasets" can help transparency and accountability in data science? Is it feasable to implement it?
- Do you think there is a particular kind of data that may benefit the most from increased transparency about collection methods and processing methods? Why?

Task 3: Coral Diversity Data

Repository Setup

Follow the next steps to setup for tasks 3 and 4:

1. Fork this repository: https://github.com/carmengg/eds220-hwk-1
2. In the Taylor server, start a new JupyterLab session or access an active one.
3. Using the terminal, clone your `eds220-hwk-1` repository to a new directory under your `eds-220` directory.
4. In the terminal, use `cd` to navigate into the `eds-220-hwk-1` directory. Use `pwd` to verify `eds-220-hwk-1` is your current working directory.

In this task you will practice:

- preliminary data exploration
- accessing data using a URL from a data archive
- selecting data from a data frame

- git
- commenting your code

Follow the instructions in the notebook `hwk1-task3.ipynb` to complete this task.

Task 4:

In this task you will practice:

- accessing data from your directory
- selecting data from a data frame
- creating exploratory graphs
- git
- commenting your code

Follow the instructions in the notebook `hwk1-task4.ipynb` to complete this task.