## In-Class Exercise 4 - 1

*Note: Work on this with your team. Draft your responses in the space below – this will let us learn from each other’s analysis. Also make sure to copy it into* [*Canvas*](https://canvas.wisc.edu/courses/450230/assignments/2555530) *so that you get credit.*

### Prompt:

The goal of this exercise is to help you learn to critique some hypothetical projects, to help get a sense of both the common traps and the potential for statistical creativity.

For **two** of the project descriptions below, discuss:

* Measurements: What types of measurements are available, and are there any important features that aren't available? Do you think the measurement process was reliable or that the provided features are meaningful? Do you recommend that the team use this data or search for an alternative?
* Relevance: Does the approach seem to answer the questions in which the authors are actually interested? Further, how broadly would the project's conclusions apply (e.g., to other populations)? Should they modify any questions?
* Depth: What types of background will the authors need to develop so that they can study their problem in depth? How will you know that the project has succeeded and not simply given a superficial overview?

Note that some of these questions are framed as Yes/No, but you can also argue that it lies in a grey area in between.

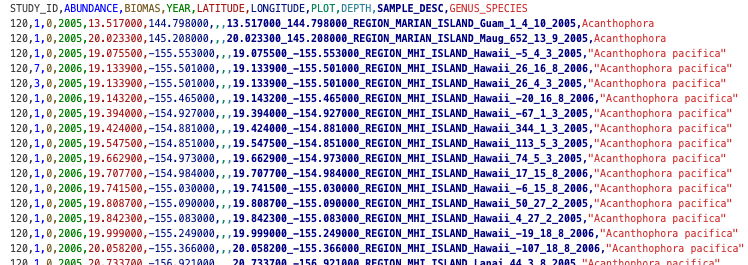
**Project 1: Coral Ecosystems**

We found the [BioTIME](https://biotime.st-andrews.ac.uk/) data in the Data is Plural [newsletter](https://www.data-is-plural.com/archive/2022-09-21-edition/). We have heard about biodiversity loss due to climate change and thought that this dataset would help us study this in more detail. Since the entire dataset seems too much to work with, we want to focus on just the samples related to [tropical coral](https://biotime.st-andrews.ac.uk/selectSpecies.php?biome%5B%5D=Tropical+coral&syear=&eyear=&getForm=Search). In this subset, there are 350K records giving time series of species abundances across 5K species. Each record comes with a time/location ID, and we have included a screenshot below.

We would like to use this dataset to answer:

* Which species seem to become more vs. less abundant? What are the major trends?
* How do these species trends depend on geography? Are there species that are thriving in one ecosystem but not another?
* Which coral reefs seem to be the healthiest/have maintained the most biodiversity over time?

We would like to prepare a report studying these data in depth. We plan on answering the questions above. We will also use visualization to check for any potential anomalies in the data, like whether two studies at nearby reefs gave very different measurements.



**Project 2: QuickDraw**

We all like to doodle, and we want to analyze the drawings in the QuickDraw [dataset](https://huggingface.co/datasets/google/quickdraw). This is a dataset of computer drawings people made within 20 seconds to describe some category (e.g., "backpack"). After those 20 seconds, an AI algorithm is asked to classify the doodles -- the model’s success/failure is also recorded. We would like to use these data to understand:

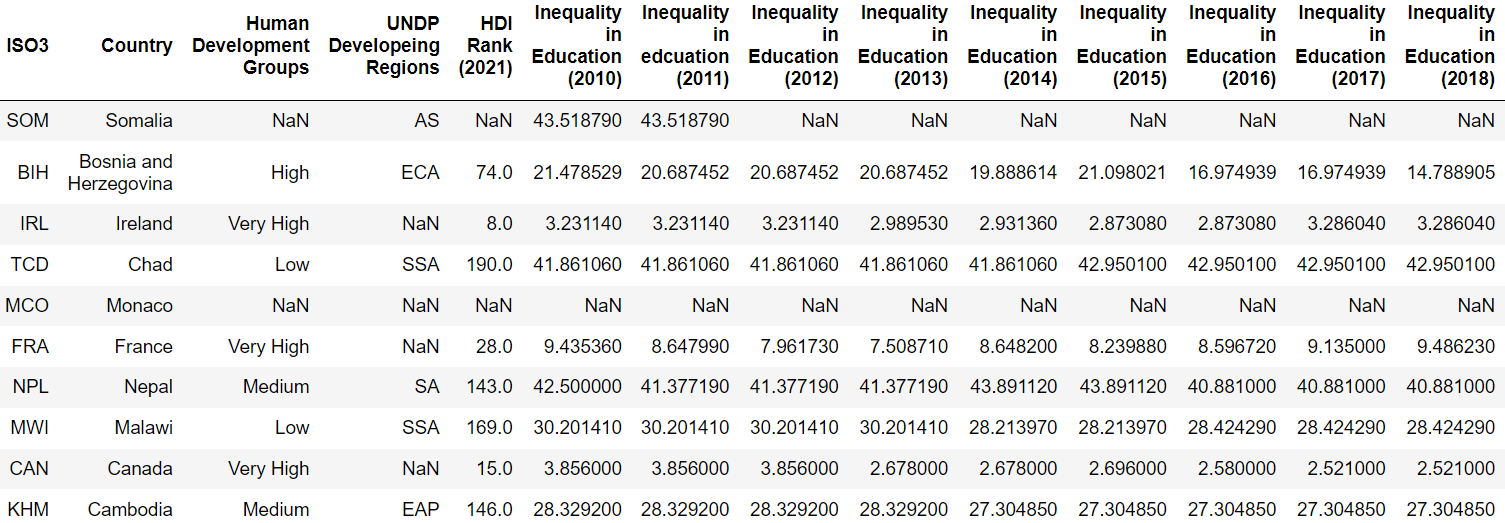
* Who are the best and worst artists in this dataset, and what do their doodles look like?
* Are there major cultural differences in how people represent objects? For example, I heard that snowmen are drawn differently in Japan vs. the US.

Since there are 50 million drawings, we plan on taking a subsample of only those drawings created by a random sample of 10K users. Our interface will allow users to navigate differences in doodles by country, doodle category, and whether the doodle was successfully recognized by the AI. We would also like to have a way for people using the interface to upload their own doodle and see which ones from the dataset are the most similar.

**Project 3: Education Inequality**

Everyone on our team has always been interested in education, and we want to use this project to understand the factors that make it easier for people to become more highly educated. A few of us have read about campaigns to improve literacy rates and wanted to learn more about this type of work.

For this reason, we'd like to create a visualization interface to understand education inequality around the world. We plan on using the UN Human Development Reports [dataset](https://www.kaggle.com/datasets/iamsouravbanerjee/inequality-in-education-around-the-world), and we included a screenshot below. As mentioned on that webpage, the data come from the UN Human Development Reports, and it gives a time series of education inequality from 2010 to 2018. The data are at the country level. We think that this interface will help our readers identify countries where education inequality is increasing or decreasing. They may be able to discover potential explanations for these trends once they see them.



### Responses: