Checkpoint 2

For this assignment the following needs to be in place for your team to get started:

* An agreed upon research topic
* An agreed upon (first) research question

The purpose of this assignment is to collect the first datasets that will form the basis of your project work.

Your mission for this assignment is the following:

### Finding Data

1. Each team member searches for 2 datasets relevant to your project work; for a total of 4 **different** datasets in a team of 2 (6 in a team of 3). These datasets may be accessible in any form (API, download, data you have collected yourself, ...). You can also reach out to data providers and ask them for data (we have signed data release forms for this class before). The only requirement is that the data should already be available - or at least available by the assignment deadline.

**Do not download this dataset yet** but keep track of where you found it.

1. Each team member looks at documentation about their datasets and the dimensions/variables and number of datapoints that are contained - as well as other relevant factors such as when the data was collected.
2. You meet as a team and decide which of the found datasets might best provide enough interesting exploratory content to support your research questions. Your result may be that a single dataset is interesting enough or that you might want to combine some of the datasets you found.  
     
   It is possible that in this process you might want to narrow down or modify your research question. Keep track of decisions that you make in this regard.
3. As a team download or scrape this/these dataset/s and share them among the team

### **Advice for Choosing Datasets**

We advise against using unstructured data from social media streams such as Twitter/X unless you are very experienced with it. If you use them you will have to do a lot of data transformation and cleaning and this work will not be part of your grade for this project. If you are doing the work as part of another class, however, you can combine it.

### **Responsible Scraping**

Before you scrape data from the web, please make sure that you read the website’s ToS (Terms of Service). Some websites don’t allow web scraping of their content. Many (like Wikipedia) allow scraping, provided it is not disruptive to other users. For example, the Wikimedia Foundation<https://foundation.wikimedia.org/wiki/Terms_of_Use/en>, specifies that “automated uses” like web scraping are not allowed if they are abusive or disruptive of the services. However, scraping a small amount of information for academic purposes hardly meets this criteria.  
 **Basic Rules**

* Check if there is an API. If there is, use it. It will make your life easier.
* Try not to scrape too much in a short time. This can bog down the servers, and may result in you being banned from a website.
* Never scrape anything that is not public.

### Data Cleaning

1. Your next task is to transform your data into something ready for analysis - that is something that looks like one or multiple tidy data tables. This task might involve the following steps:
   1. If your data is huge (>a few hundred MB) then reduce it in size (e.g. filter out some years, remove unnecessary columns, empty rows, etc.). OpenRefine might not be able to help you with this task and you might have to do this in Python or another language.
   2. If you already have a data table then make sure it's in a tidy (see the lecture on cleaning) format.
2. Your second task is to do some data cleaning. Take your transformed data and load it into OpenRefine. Here, inspect the data as we learned in the tutorial and correct data errors. Keep track of the types of possible data errors you found and what changes you made to the dataset. Also save your operations in a .json file so you can reuse them in case you need to change your dataset again

### Submission

Fill out the assignment submission form with the details about the datasets you collected.

Grading scheme

This assignment is a checkpoint. It has a simplified grading scheme:

| 0 points | No form submitted |
| --- | --- |
| 1 point | Form submitted but only 1 dataset |
| 2 points | Two appropriate datasets submitted but insufficient detail provided |
| 3 points | Two appropriate datasets submitted and all necessary detail provided |