Thip Rattanavilay

DSC540-T302 Data Preparation

Project: Milestone 1

**Covid-19 & Vaccine Analysis**

* 3 data sources, along with a description of each one (links to each are fine, no need to submit the actual data)

<https://www.ecdc.europa.eu/en/covid-19/data>

At the request of Member States, data on the daily number of new reported COVID-19 cases and deaths by EU/EEA country will be available to download from 11 March 2021. ECDC will continue to publish weekly updates on the number of cases and deaths reported in the EU/EEA and worldwide every Thursday. The daily and weekly data are available as downloadable files in the following formats: XLSX, CSV, JSON and XML.

<https://ourworldindata.org/coronavirus-source-data>

Our complete COVID-19 dataset is a collection of the COVID-19 data maintained by Our World in Data. It is updated daily and includes data on confirmed cases, deaths, and testing.

<https://covid19api.com/#>

Access data on COVID19 through an easy API for free. Build dashboards, mobile apps or integrate in to other applications. Data is sourced from Johns Hopkins CSSE

*Additional Vaccine dataset –*

<https://www.kaggle.com/imdevskp/corona-virus-report>

<https://datascience.nih.gov/covid-19-open-access-resources>

<https://www.tn.gov/health/cedep/ncov/data/downloadable-datasets.html>

* The relationships between them, or the relationship you will make between them

The relationship between the three plus the additional data sets will be the country names and vaccine type. The base or main table will be the API/CSV/JSON/XML. I will use the countries provided there to pull information from the other data sets as well as vaccine type. This way I can pull resources among the three and get the most up to date tracking and where the world is currently at.

* What you believe you will have to do to the data to accomplish all 5 milestones and what your interpretation is of what the data means (you could provide a data dictionary or a summary of what the data is) – should be at least 250 words

In the following few milestones, I will need to decide if I will go to the website to pull the CSV files, API calls, or if I am going just to download them as a flat file. I don't think that there will be a ton of work involved needing to transform the data too much. I might have multiple flat files to load from the additional resource above.

I had a hard time coming up with this one because I am not as familiar with the web scrape. I think this one will require the most work on my side. I want to scrape CDC/Google/John Hopkins or other news sites and pull data on what is happening. I want to find information around symptoms, severity, cause, death, recovered, and other items related to individuals within each country and provide data on what country has been vaccinated. I want to see if the information shared about the virus helps with slowing down the spread. My current plan for the web scrape is to create a sort of class that I can build a web scrape foundation.

I will probably need to build the API script before the web scrape to look up the countries for the web scrape search. I know how to build this off of JavaScript; I would have to teach myself how to pull API with python; once I figured that out, I don't think it would be too difficult to work with after I learn Python's API.