Project Milestone 1

DSC 540

Abed Tabbalat

This project is to collect the most common Covid-19 data and have it stored as one dataset. The dataset will contain positive cases, deaths, recoveries, and vaccine by country.

Data Sources:

API:

- Contains data for positive cases, deaths, and recoveries by country
- The dataset contains the following variables:
    - Country_Region
    - Last_Update
    - Lat
    - Confirmed
    - Deaths
    - Recovered
    - Active
    - Incident_Rate
    - People_Tested
    - People_Hospitalized
    - Mortality_Rate
    - UID
    - ISO3
- https://covid-19-data.unstatshub.org/datasets/cases-country/api

Website:

- Contains data for population by county
- The dataset contains the following variables:
    - Country (or dependency)
    - Population
    - Yearly Change
    - Net Change
    - Density
    - Land Area

    - Migrants
    - Fert. Rate
    - Med. Age
    - Urban Pop
    - World Share
- https://www.worldometers.info/world-population/population-by-country/

Flat File (CSV):

- Contains daily vaccinations by country
- The dataset contains the following variables:
    - Country
    - ISO Code
    - Data
    - Total Vaccinations
    - People Vaccinated
    - People fully vaccinated
    - Daily vaccinations raw
    - Daily vaccinations
    - Total vaccinations per hundred
    - People vaccinated per hundred
    - People fully vaccinated per hundred
    - Daily vaccinations per million
    - Vaccines
    - Source name
    - Source website
- https://www.kaggle.com/gpreda/covid-world-vaccination-progress

Relationships:

The three data sources common relationship is Country name which will be used as the ID. Each dataset contains a dimension that has country name mapped to different segments. API link shows what each country positive test results, deaths, and recoveries are. The website shows the country population. The CSV file includes daily vaccinations by country.

Action Plan:

Covid has changed our lives significantly and everyone has invested their time trying to understand it. Most of the world has invested time to understand the impact that it has done with our lives. However, in my case, I wasn't invested much in it as I try to avoid the news as much as I can. I believe it will be interesting to see how each country in the world is reacting to this virus.

Understanding what the data is doing is key. This project will enhance my skillset in pulling data from websites and APIs as it is the area that I am not familiar with as much as using a flat file. Cleaning the data is the most important. The reason behind that is we want to have the most accurate information to prevent passing on falsified information.

Simplifying headers will help in knowing what our dimensions are and would make the coding process easier. That said, follows to see whether there are certain areas that aren't readable that could be reformatted. Once this step is complete, checking for NAs, outliers and duplicates and figuring out what the reason behind those three components are to prepare our data to be analyzed.

The end goal is to have all the information needed into one dataset so that the user can have a quick access to the most common variables needed.

Covid has impacted the world and changed our lives. I believe working together as a planet to be on top of it would result in over coming this pandemic quicker than it needs to be.

My biggest challenge in this project will be handling the website and API portion of the project as those are the two areas, I am least familiar with in bringing the data into my Python code. Also, in most cases in list by countries, the US usually have things broken down by state which I believe may be a challenge into compiling up the data and present as one country. Some may challenge that each state needs to be treated as its own country and this is something I will have to think about to see what is the best possible way to address it.