Beginners Sports Analytics NFL Dataset | Kaggle - great notebook… spatial yes but geographic? Not really
NFL Combine - Performance Data (2009 - 2019) | Kaggle

Seems most likely
https://covid19.who.int/data Covid dash data source

Data Source Description

Data Source: https://covid19.who.int/data
I chose this dataset because I wanted to work with something "real" and relevant. There are already dashboards (including the one I sourced this data from) using it to provide useful information about Covid-19, and I believe there are many possible other pieces of information that could be utilized with this data. For example comparing GDP or Gini Index with country level Covid data.
It appears to have a good breadth of information that should permit multiple potential analytical routes, be flexible in terms of what else I can integrate with it, and be a relevant topic with data that potentially has real patterns vs. a sample dataset.

These are all **external** sources, since none of it is my data, nor the WHO's self-collected data. As far as I can tell, this is all **survey** data since various organizations are reporting it to the WHO.

- Primarily World Health Organization (WHO) data for Case and Death counts
    - "From the 31 December 2019 to the 21 March 2020, WHO collected the numbers of confirmed COVID-19 cases and deaths through official communications under the International Health Regulations (IHR, 2005), complemented by monitoring the official ministries of health websites and social media accounts. Since 22 March 2020, global data are compiled through WHO region-specific dashboards (see links below), and/or aggregate count data reported to WHO headquarters daily."
    - " Case detection, definitions, testing strategies, reporting practice, and lag times (e.g. time to case notification, and time to reporting of deaths) differ between countries, territories and areas. These factors, amongst others, influence the counts presented with variable under or overestimation of true case and death counts, and variable delays to reflecting these data at a global level."
    -
- Vaccination data is pooled from numerous sources including WHO review of public data, direct reports from Member States, or third party sites such as Our World in Data.
    - Third party data has not been validated by WHO.
    - Some disagreement in counts is expected due to different inclusion criteria and cutoffs

- Population data are drawn the following sources (copied from WHO Covid-19 dashboard data description):
    - “United Nations, Department of Economic and Social Affairs, Population Division. World Population Prospects 2019, Online Edition. Rev. 1 (2020 projections). Available Online.
    - Eurostat. “Demo_pjan” (population as of 1 January by country, year, age and gender; last updated by Eurostat on 2021-02-12; last year of data: 2020). Available online.
    - National Statistics Office Malta. Available online.
    - The Government of the Pitcairn Islands. Available online.
    - Statistics Netherlands (CBS). Estimates of population for Bonaire, Sint Eustatius and Saba. Available online.”
- Public Health and Social Measures (PHSM) data\
    - “PHSM data is collected from publicly available sources from governments, international, national, and regional authorities and media.”

Data Understanding

- Case Table
    - This is summarized data. It has been consolidated into more aggregations than the underlying Cases data. There is one entry per country, and a variety of time-specific measures (new cases in the last 7 days, for example). I could download updated versions of the dataset to get more datapoints for some of these numbers, should they be actively updating the dataset.
    - This should prove pretty useful for spatial analysis–essentially some of the aggregation work has just been done for me.
- Cases
    - This is the raw data feeding into the Case Table dataset. Here is where the bulk of my data volume lies, and where most of my time series options lie, like cumulative cases over time.
- Vaccine data
    - Much like the Case Table dataset, this presents aggregated figures for vaccinations per country. Unlike the covid case data, I do not have the underlying reports (a Cases dataset equivalent).
    - I can potentially use this to try to establish a relationship between aggregate Case Table data and figures from this dataset. It might even be most useful to merge the two datasets.
- Vaccine metadata
    - This seems like the least useful dataset. It is really just a list of countries and which vaccines from which companies they use. There is start/end date information for vaccine deployment, but it is incomplete. End Date makes sense, since there is no reason to end our usage of Covid vaccines yet. But it ends up with mostly tangential or auxiliary relevance.

Data Limitations & Ethics

- There are no substantial PID concerns with this dataset. No personal information is provided whatsoever.
- Bias could exist in the collection methods. If a country is not reporting accurate numbers, or lacks the infrastructure to collect accurate numbers, that could affect my analysis. The WHO states in their dataset description that they cannot validate the vaccination data, and I believe the same to be true for Case data as well.
- It is possible that lack of information about medical infrastructure and normal background or even illness-related death rates for every country could bias my analysis. If more people die of illness in poorer countries every year for example, then I would expect more covid-related deaths than other countries even if we considered it identical to any other illness (like a common flu). Or if my analysis cannot explain a higher death rate using data from this set such as vaccinations and case volume, there may be important factors external to the dataset such as medical infrastructure, education, budget, and personnel.

Investigation Questions

- Does an increased rate of covid cases predict the same increase in death rates worldwide?
- As vaccinations increase, do covid deaths and cases decrease? Is the rate of decrease the same for each?
- In which countries has covid had the highest rate of cases? Does this change year to year?
    - The same for deaths–and are these identical maps?
    - Is this the same heatmap as a population map (I call this 'contingency X…KCD')?