TASK 6.1: SOURCING OPEN DATA

Marilize de Villiers

Data Source

Dataset: COVID-19 Vaccinations

Data Sourcing:

This dataset is an external open dataset. The data is collected, aggregated, and cumulated by Our World in Data. They collect the data from official reports from governmental institutes in each country as well as the World Health Organization. They have been collecting and updating the dataset since vaccine rollouts started (end of 2020/beginning of 2021). Since Our World in Data only rely on official reports, governments and the WHO, the data can be judged as reliable and trustworthy.

Data Collection:

The data is administrative. Since COVID-19 is a Global pandemic, each country is obligated to keep track of their statistics related to COVID-19 and to make these statistics available to anyone. Organisations such as WHO cumulate these statistics that they either receive directly from member states or obtain from 3rd party organisations, such as Our World in Data. A major caveat of this data is that not all countries publish their statistics daily and some organisations who collect and share the data may have differing inclusion criteria or cut-off times (e.g., WHO might publish a day's vaccination counts in the evening and another institution may publish their statistics for that day only after WHO has already published the day's statistics). Time differences and different estimation tools may also cause inaccuracies in the data.

Data Content:

The dataset contains a daily count of vaccinations by country since the vaccine rollout started in December 2020. It contains information on how many people have been fully vaccinated, had their first shot, as well as total vaccinations to date per country.

Why did I choose this data (data relevance):

I chose this data since I am very interested in health statistics/data, coming from a health industry background. Since COVID-19 is a global pandemic, I thought it would be interesting to see how vaccination initiatives fares in each country, how the vaccination rates and totals differ for each country, as well as how vaccinations have progressed over time – globally and for each country. I am hoping to merge data on COVID-19 deaths to try and compare death rates before vaccine rollouts started to current death rates. My ultimate goal is to determine the success rate of the COVID-19 vaccines.

Data Profile

Dataset Details

The original dataset had 81976 rows and 15 columns. These are the column descriptions:

- **country**: name of the country (or region within a country).
- **iso_code**: ISO 3166-1 alpha-3 – three-letter country codes.
- **date**: date of the observation.
- **total_vaccinations**: total number of doses administered. For vaccines that require multiple doses, each individual dose is counted. If a person receives one dose of the vaccine, this metric goes up by 1. If they receive a second dose, it goes up by 1 again. If they receive a third/booster dose, it goes up by 1 again.
- **people_vaccinated**: total number of people who received at least one vaccine dose. If a person receives the first dose of a 2-dose vaccine, this metric goes up by 1. If they receive the second dose, the metric stays the same.
- **people_fully_vaccinated:** total number of people who received all doses prescribed by the initial vaccination protocol. If a person receives the first dose of a 2-dose vaccine, this metric stays the same. If they receive the second dose, the metric goes up by 1.
- **daily_vaccinations_raw**: daily change in the total number of doses administered. It is only calculated for consecutive days. This is a raw measure provided for data checks and transparency, but we strongly recommend that any analysis on daily vaccination rates be conducted using daily_vaccinations instead.
- **daily_vaccinations:** new doses administered per day (7-day smoothed). For countries that don't report data on a daily basis, we assume that doses changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window. An example of how we perform this calculation can be found here.
- **total_vaccinations_per_hundred**: total_vaccinations per 100 people in the total population of the country.
- **people_vaccinated_per_hundred:** people_vaccinated per 100 people in the total population of the country.
- **people_fully_vaccinated_per_hundred**: people_fully_vaccinated per 100 people in the total population of the country.
- **daily_vaccinations_per_million:** daily_vaccinations per 1,000,000 people in the total population of the country.
- **vaccines** – names of vaccines used in the country (up to date)
- **source_name** - source of the information (national authority, international organization, local organization etc.)
- **source_website** - website of the source of information.

Consistency Checks & Cleaning

Data Types:

There were no mixed data types.

I changed "date" from *float64* to *datetime64[ns].*

Missing Values:

I changed NaN values to 0 since they weren't really missing. Some countries only report their statistics once per week and since the dates are daily, some days have no statistics reported. I might have to aggregate the daily dates to monthly summaries for future analysis.

Duplicates:
There were no duplicates.

Dropped Columns:

I dropped iso_code, source_name, and source_website. I might drop a few of the other columns later but since I am only starting out with basic questions, I am not 100% sure which variables I would want to explore further.

Consistency:

I checked that all country names have a consistent format and are correct. I also checked the counts for each date. It is clear that not all countries have an entry for each day, thus it would make sense to aggregate these into monthly sum per country (I am not yet sure how to do that – would have to research a bit more).

Basic Descriptive Statistics:

After checking and cleaning procedures, the data set now has 81976 rows and 12 columns.

See next page for a summary table of descriptive statistics.

	total_ vacci-nations	people_vacci-nated	people_fully_ vaccinated	daily_vaccina tions _raw	daily_vaccina tions	total_vacci-nations_per_ hundred	people_vacci nated_per_ hundred	people_fully_ vacci-nated_per_h undred	daily_vaccina tions_per_mil lion
count	81976	81976	81976	81976	81976	81976	81976	81976	81976
mean	21791285	8045972	5914203	113150.1	133933.8	39.13193	19.21886	15.48715	3327.514
std	1.53E+08	46989618	35901413	802359.6	782905.8	60.60931	28.31254	25.39302	3968.625
min	0	0	0	0	0	0	0	0	0
25%	0	0	0	0	934	0	0	0	674
50%	7423.5	0	0	0	7557.5	0.13	0	0	2146
75%	3628498	1830708	1102393	14016.5	44892	65.9	37.09	24	4788
max	3.17E+09	1.27E+09	1.23E+09	24741000	22424286	336.16	124.65	121.53	117497

Data Limitations & Ethics

The biggest possible limitation of this dataset is that not all countries have entries for every single day which would make a time series a bit difficult. I would have to find a way to aggregate or group the daily dates by months for ach year. This way I will be able to make comparisons between countries based on their monthly sum of vaccinations. Some of the data may also be a bit inaccurate since some countries may have possible have bias in the way they report their number. It is possible that some countries may not report accurate vaccination rates due to the global pressure to vaccinate as quickly as possible.

Since there is no PII in this dataset, ethical issues such as data privacy and security is not a concern for me.

Initial Questions

1. Which countries are using which vaccines?
2. How has the vaccine rollouts been progressing in each country?
3. Which countries are more advanced and why?
4. Which countries had the first batch of vaccines?
5. Eventually, I would also want to know what the effect of the vaccines have been on the COVID-19 situation worldwide? This might be explored when a second dataset about daily covid-19 deaths can be introduced and merged.