Purpose and Scope

- The purpose of the Our World in Data COVID-19 vaccination dataset is to provide timely, comparable data across countries to understand the scale and rate of the vaccine rollout. The dataset tracks the total number of COVID-19 vaccinations administered in each country, broken down by first and second doses (where national data is available), and derived daily vaccination rates and population-adjusted figures.
- The data stored in the database includes the total number of COVID-19 vaccinations administered in each country, broken down by first and second doses (where national data is available), and derived daily vaccination rates and population-adjusted figures. The data is compiled from official sources, including health ministries, government reports, and official social media accounts.
- The intended users of the database are journalists, policymakers, researchers, and the public. The dataset has been used by leading media outlets, including the New York Times, the BBC, the Financial Times, and The Economist. The WHO relies on this dataset for its official COVID-19 dashboard, and policymakers use it to benchmark the performance of national vaccination programs across countries.

Analysing and Normalising Data

FILE NAME	COLUMNS	DESCRIPTION
locations.csv	location	Name of the country
	iso_code	Three letter country codes
	vaccines	List of vaccines administered in the country
	last_observation_data	Date of last observation of the data
	source_name	Name of source for data collection
	source_website	Web location of the source

Observation:

- **location** and **iso_code**: Each country has a unique iso_code. The **iso_code** can be the primary key since each country should have a unique ISO code.
- **vaccines**t appears that some countries use the same set of vaccines.
- **last_observation_data**: This seems to be unique for each row, so there's no redundancy here.

FILE NAME	COLUMNS	DESCRIPTION
vaccinations-by-manufactuter.csv	location	Name of country
	date	Date of observation
	vaccine	List of vaccines administered in the country
	total_vaccinations	Total number of doses administered

Observation:

- The **location** field is repeated for each date and vaccine type.

FILE NAME	COLUMNS	DESCRIPTION
vaccinations-by-agegroup.csv	location	Name of country
	date	Date of observation
	age_group	Grouped by ages 3-11, 1217, 18-29, 30-39, …., 90-99 and 100+
	people_vaccinated_per_hundred	People vaccinated per 100 people in the total population within the same age group in the country
	people_fully_vaccinated_per_hundred	People fully vaccinated per 100 people in the total
		population within the same age group in the country
	people_with_booster_per_hundred	People with booster per 100 people in the total population within the same age group in the country

Observation:

- The **location** field is repeated for each date and age group. So is the **age_group.**

FILE NAME	COLUMNS	DESCRIPTION
Germany.csv France.csv England.csv Australia.csv	location	Name of country
	date	Date of observation
	vaccine	List of vaccines administered in the country
	source_url	Web location of the source
	total_vaccinations	Total number of doses administered
	people_vaccinated	Total number of people who received at least one vaccine dose
	people_fully_vaccinated	Total number of people who received all doses prescribed by the initial vaccination protocol.
	total_boosters	Total number of COVID-19 vaccination booster doses administered

Observation:

- **location** column has the same value throughout the table
- **source_url** appears to be repetitive.
- **vaccine** field has a list of vaccines, separated by commas (multi-valued).
- **people_vaccinated, people_fully_vaccinated, total_vaccinations** have empty cells
- **total_vaccinations** can be derived from **people_vaccinated, people_fully_vaccinated, total_boosters**
    - Each person who is not fully vaccinated has received one dose.
    - Each fully vaccinated person, excluding those who have taken booster shots, has received two doses.
    - Each person who has received a booster has taken three doses in total (2 initial + 1 booster).
- **total_vaccinations = people_vaccinated + people_fully_vaccinated+ total_boosters**

FILE NAME	COLUMNS	DESCRIPTION
us_state_vaccinations.csv	location	Name of state or federal entity
	date	Date of observation
	total_vaccinations	Total number of doses administered
	total_vaccinations_per_hundred	total_vaccinations per 100 people in the total population of the state
	daily_vaccinations_raw	Daily change in the total number of doses administered calculated for consecutive days (raw measure provided for data checks and transparency)
	daily_vaccinations	New doses administered per day (7-day smoothed)
	daily_vaccinations_per_million	daily_vaccinations per 1,000,000 people in the total population of the state
	people_vaccinated	Total number of people who received at least one vaccine dose
	people_vaccinated_per_hundred	people_vaccinated per 100 people in the total population of the state
	people_fully_vaccinated	Total number of people who received all doses prescribed by the initial vaccination protocol.
	people_fully_vaccinated_per_hundred	people_fully_vaccinated per 100 people in the total population of the state
	total_distributed	Cumulative counts of COVID-19 vaccine doses recorded as shipped in CDC's Vaccine Tracking System

	total_distributed_per_hundred	Cumulative counts of COVID-19 vaccine doses recorded as shipped in CDC's Vaccine Tracking System per 100 people in the total population of the state
	share_doses_used	Share of vaccination doses administered among those recorded as shipped in CDC's Vaccine Tracking System
	total_boosters	Total number of COVID-19 vaccination booster doses administered
	total_boosters_per_hundred	total_boosters per 100 people in the total population

Observation:

- **loaction** appears to be repetitive.
- **total_vaccinations_per_hundred**: This can be derived from **total_vaccinations** by dividing it by the total population of the state and multiplying by 100.
- **daily_vaccinations_per_million**: This can be derived from **daily_vaccinations** by dividing it by the total population of the state and multiplying by 1,000,000.
- **people_vaccinated_per_hundred**: This can be derived from **people_vaccinated** by dividing it by the total population of the state and multiplying by 100.
- **people_fully_vaccinated_per_hundred**: This can be derived from **people_fully_vaccinated** by dividing it by the total population of the state and multiplying by 100.
- **total_distributed_per_hundred**: This can be derived from **total_distributed** by dividing it by the total population of the state and multiplying by 100.
- **share_doses_used**: This can be derived by dividing **total_vaccinations** by **total_distributed**.
- **total_boosters_per_hundred**: This can be derived from **total_boosters** by dividing it by the total population of the state and multiplying by 100.
- **total_vaccinations = people_vaccinated + people_fully_vaccinated + total_boosters**

FILE NAME	COLUMNS	DESCRIPTION
vaccinations.csv	location	Name of country
	iso_code	Three letter country codes
	date	Date of observation
	total_vaccinations	Total number of doses administered
	total_vaccinations_per_hundred	total_vaccinations per 100 people in the total population of the country
	daily_vaccinations_raw	Daily change in the total number of doses administered calculated for consecutive days (raw measure provided for data checks and transparency)

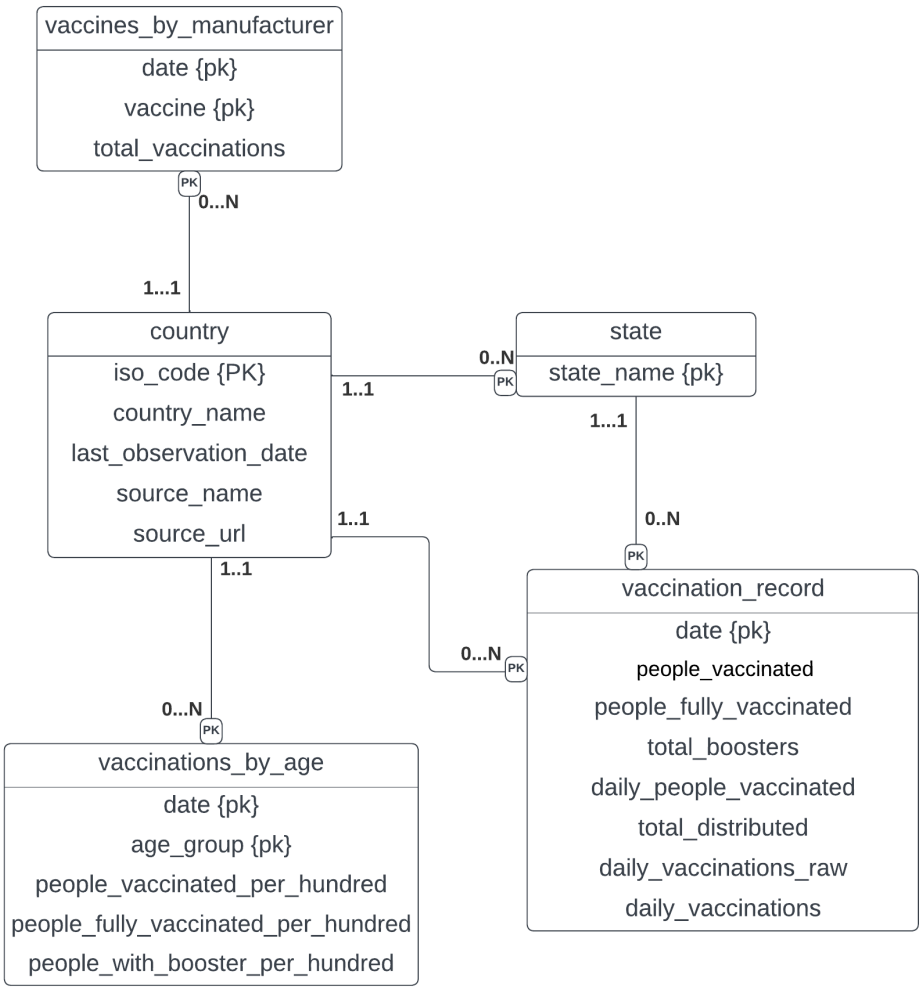
	daily_vaccinations	New doses administered per day (7-day smoothed)
	daily_vaccinations_per_million	daily_vaccinations per 1,000,000 people in the total population of the country
	people_vaccinated	Total number of people who received at least one vaccine dose
	people_vaccinated_per_hundred	people_vaccinated per 100 people in the total population of the country
	people_fully_vaccinated	Total number of people who received all doses prescribed by the initial vaccination protocol.
	people_fully_vaccinated_per_hundred	people_fully_vaccinated per 100 people in the total population of the country
	total_boosters	Total number of COVID-19 vaccination booster doses administered
	total_boosters_per_hundred	total_boosters per 100 people in the total population
	daily_people_vaccinated	Daily number of people receiving a first COVID-19 vaccine dose (7-day smoothed)
	daily_people_vaccinated_per_hundred	daily_people_vaccinated per 100 people in the total population of the country

Observation:

- **total_vaccinations_per_hundred**: This can be derived from **total_vaccinations** and the total population of the country.
- **daily_vaccinations_per_million**: This can be derived from **daily_vaccinations** and the total population.
- **people_vaccinated_per_hundred**: Can be derived from **people_vaccinated** and the total population.
- **people_fully_vaccinated_per_hundred**: Can be derived from **people_fully_vaccinated** and the total population.
- **total_boosters_per_hundred**: Can be derived from **total_boosters** and the total population.
- **daily_people_vaccinated_per_hundred**: Can be derived from **daily_people_vaccinated** and the total population.
- **total_vaccinations = people_vaccinated + people_fully_vaccinated + total_boosters**

ER Diagram:

- The country specific records overlap with the vaccinations data, hence we can merge the two tables.
- We are assuming that states have unique names in a country.



Now we have the following tables:
Country table:

- iso_code (Primary Key)
- location
- last_observation_date
- source_name
- source_website

State table:

- iso_code (Primary Key, Foreign Key)
- state_name (Primary Key)

Vaccination Record Table (For country):

- date (Primary Key)
- iso_code (Primary Key, Foreign Key)
- total_distributed
- people_vaccinated

- people_fully_vaccinated
- total_boosters
- daily_people_vaccinated
- daily_vaccinations_raw
- daily_vaccinations

Since we do not have information about other states it is a safer choice to make a separate table for US states. Although in a combined table the primary key would be (date, iso_code, state_name).
Vaccination Record Table (For state):

- date (Primary Key)
- state_name (Primary Key, Foreign Key)
- total_distributed
- people_vaccinated
- people_fully_vaccinated
- total_boosters
- daily_people_vaccinated
- daily_vaccinations_raw
- daily_vaccinations

Vaccinations-By-Manufacturer Table:

- iso_code (Primary Key, Foreign Key)
- vaccine (Primary Key)
- date (Primary Key)
- total_vaccinations

Vaccinations-By-Age Table:

- iso_code (Primary Key, Foreign Key)
- date (Primary Key)
- age_group (Primary Key)
- people_vaccinated_per_hundred
- people_fully_vaccinated_per_hundred
- people_with_booster_per_hundred

Normalization
Country Table:
1NF – First Normal Form:
Each column contains only atomic (indivisible) values, and there's a primary key.
2NF – Second Normal Form:
source_name and **source_website** are not directly dependent on the **iso_code** alone, so we separate the dependent attributes into another table
Country table:

- iso_code (Primary Key)
- location
- last_observation_date
- source_name (Foreign Key)

Source table:

- source_id (Primary Key)

- source_name
- source_website

3NF – Third Normal Form:

location is transitively dependent on **iso_code** since **iso_code** can uniquely determine **location**. Split the "Country" table to remove the transitive dependency.

Country table:

- iso_code (Primary Key)
- location

Country Vaccination Table:

- iso_code (Primary Key, Foreign Key)
- last_observation_date
- source_name (Foreign Key)

Source table:

- source_id (Primary Key)
- source_name
- source_website

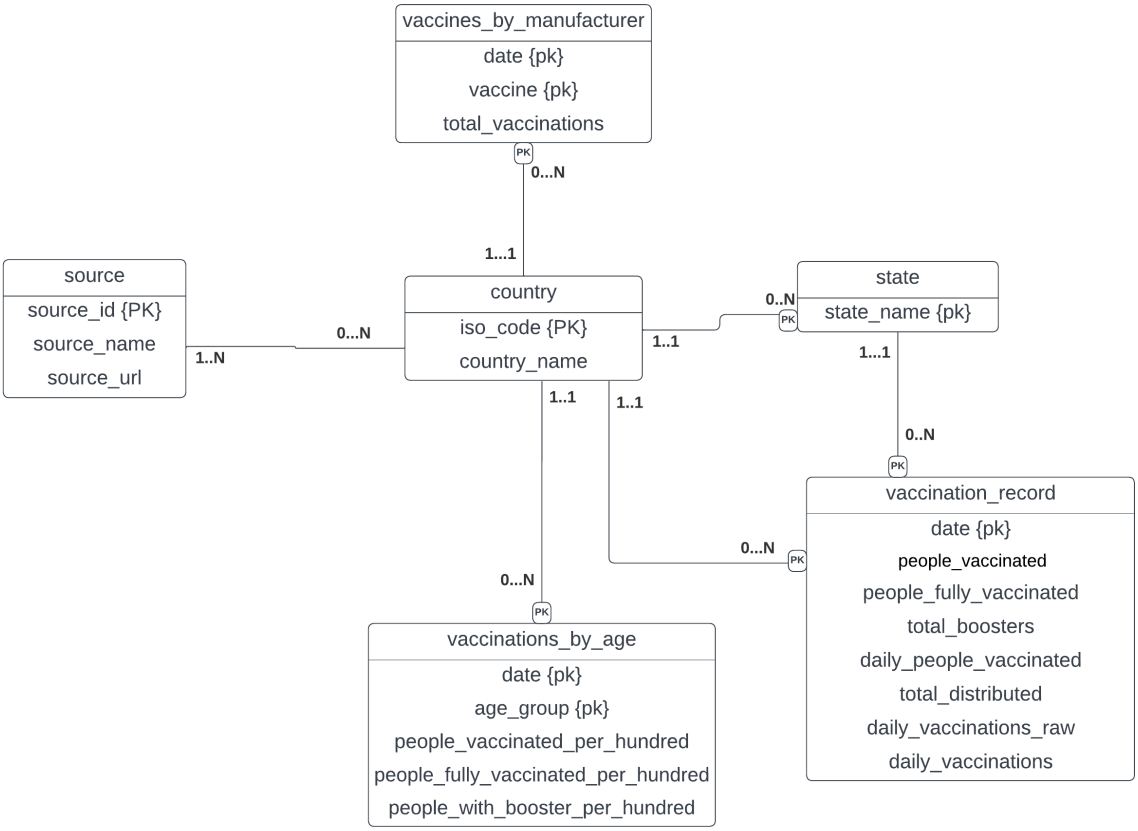
The last observation date can also be observed from the vaccination records of the country. Instead of Country Vaccination table we can have a country source junction table.

Country_Source Table:

- iso_code (Primary Key, Foreign Key)
- source_id (Foreign Key)

All the other tables are in 3NF.

Updated ER Diagram:



Database Schema:

source (source_id, source_name, source_url)

country (iso_code, country_name)

country_source (iso_code*, source_id*)

state (state_name, iso_code*)

world_vaccination_record (date, iso_code*, people_vaccinated, people_fully_vaccinated, total_boosters, daily_people_vaccinated, total_distributed, daily_vaccinations_raw, daily_vaccinations)

us_state_vaccination_record (date, state_name*, people_vaccinated, people_fully_vaccinated, total_boosters, daily_people_vaccinated, total_distributed, daily_vaccinations_raw, daily_vaccinations)

vaccinations-by-maufaturer (date, vaccine, iso_code*, total_vaccinations)

vaccinations-by-age (date, age_group, iso_code*, people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred, people_with_booster_per_hundred)