Model

Dataset Manipulation:

- The data on total vaccines, people vaccinated, and total booster given was deleted on all the specific country datasets because this same data can be found on the vaccinations dataset.
- A new spreadsheet of vaccines used per country was created from the locations csv because there was list of multiple vaccines in each cell. The vaccines column was then deleted from the locations csv
- All the specific country csvs are combined into one called countries_with_source. The list of vaccines in the vaccines column will be made into a new spreadsheet called countries_vaccines_by_day and will be deleted from the countries_with_source csv. This new csv will contain the location date and vaccine used on that date with multiple rows for each location and date if there are multiple vaccines used on each day.
- All the per_capita columns except the one without missing data was deleted from the vaccinations and us_state_vaccinations as per capita information for other information can be calculated using the kept per capita and the information and the other columns
- Iso_code is removed from all csv except the locations csv because location and iso_code is not needed as they both describe the country or location.
- For us_state_vaccinations, changed existing location column name to state
- Added missing locations in locations because some are in vaccinations that are not in locations
- Add location US in state vaccinations csv to be used as foreign key

Initial Schema (after initial dataset manipulation):

Vaccinations(location*, date, total_vaccination, people_vaccinated, people_fully_vaccinated, total_boosters, daily_vaccinations_raw, daily_vaccinations, daily_vaccinations_per_million)

Vaccinations-by-age-group(location*, date, age_group, people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred, people_with_booster_per_hundred)

Vaccinations-by-manufacturer(location*, date, vaccine, total_vaccinations)

Locations(location*, iso_code, last_observation_date, source_name, source_website)

Locations_vaccines(location*, vaccine)

Countries_vaccines_by_day(location*, date, vaccine)

countries_source_by_day(location*, date, source_url)

Us_state_vaccinations(date*, state, total_vaccinations, total_distributed, people_vaccinated, people_fully_vaccinated, daily_vaccinations_raw, daily_vaccinations, share_doses_used, total_boosters, location*)

Normalisation:

Vaccinations is in third normal form

Vaccinations_by_age_group is in third normal form

Locations:

Location → iso_code, last_observation, source_url
source_url→ source_name

Locations_vaccines is in third normal form

Countries_vaccines_by_day is in third normal form

countries_source_by_day is in third normal form

Us_state_vaccinations is in third normal form

Vaccinations_by_manufacturer is in third normal form

Final Schema:

Vaccinations(location*, date*, total_vaccination, people_vaccinated, people_fully_vaccinated, total_boosters, daily_vaccinations_raw, daily_vaccinations, daily_vaccinations_per_million)

Vaccinations-by-age-group(location*, date*, age_group, people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred, people_with_booster_per_hundred)

Vaccinations-by-manufacturer(location*, date*, vaccine, total_vaccinations)

Locations(location*, iso_code, last_observation_date, source_url*)

Source(location_url*, source_name)

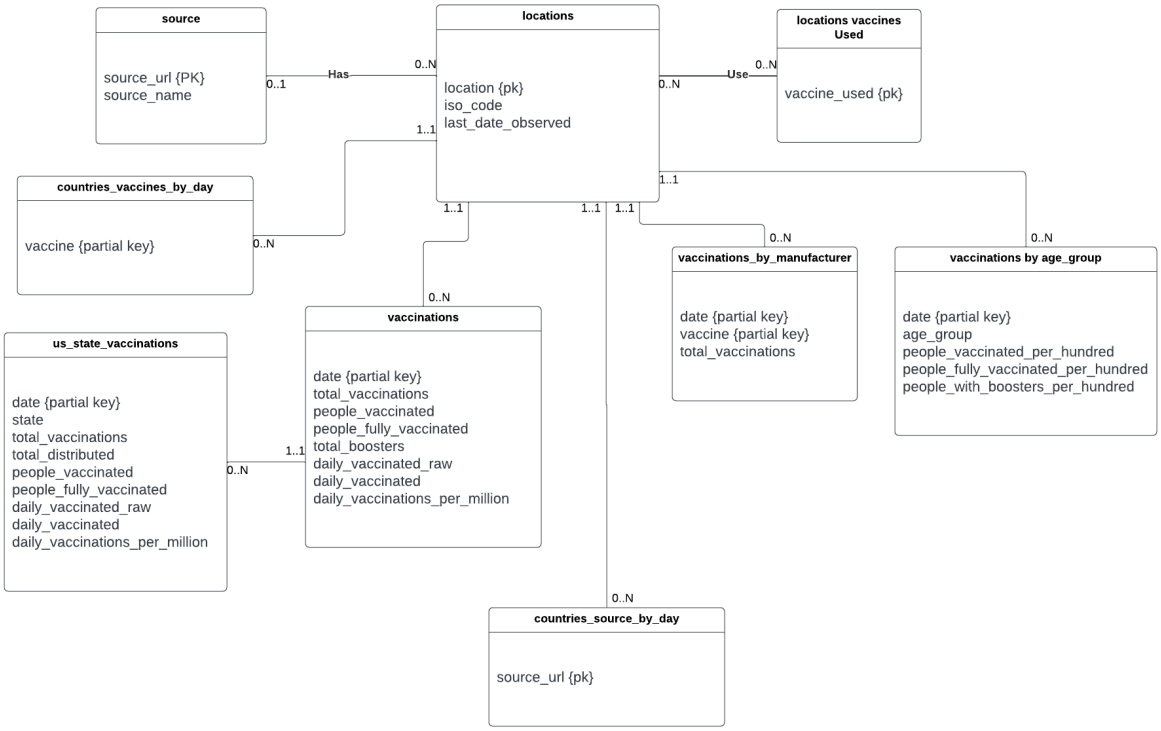
Locations_vaccines_used(location*, vaccine)

countries_vaccines_by_day(location*, date*, vaccine)

countries_source_by_day(location*, date*, source_url*)

Us_state_vaccinations(date, state, total_vaccinations, total_distributed, people_vaccinated, people_fully_vaccinated, daily_vaccinations_raw, daily_vaccinations, share_doses_used, total_boosters, location*)

ER Model:


source
source_url {PK}
source_name
Has
0..1
0..N
locations
location {pk}
iso_code
last_date_observed
Use
0..N
0..N
locations vaccines Used
vaccine_used {pk}
1..1
countries_vaccines_by_day
vaccine {partial key}
0..N
1..1
1..1
1..1
1..1
0..N
0..N
vaccinations_by_manufacturer
date {partial key}
vaccine {partial key}
total_vaccinations
vaccinations by age_group
date {partial key}
age_group
people_vaccinated_per_hundred
people_fully_vaccinated_per_hundred
people_with_boosters_per_hundred
0..N
vaccinations
date {partial key}
total_vaccinations
people_vaccinated
people_fully_vaccinated
total_boosters
daily_vaccinated_raw
daily_vaccinated
daily_vaccinations_per_million
us_state_vaccinations
date {partial key}
state
total_vaccinations
total_distributed
people_vaccinated
people_fully_vaccinated
daily_vaccinated_raw
daily_vaccinated
daily_vaccinations_per_million
1..1
0..N
0..N
countries_source_by_day
source_url {pk}