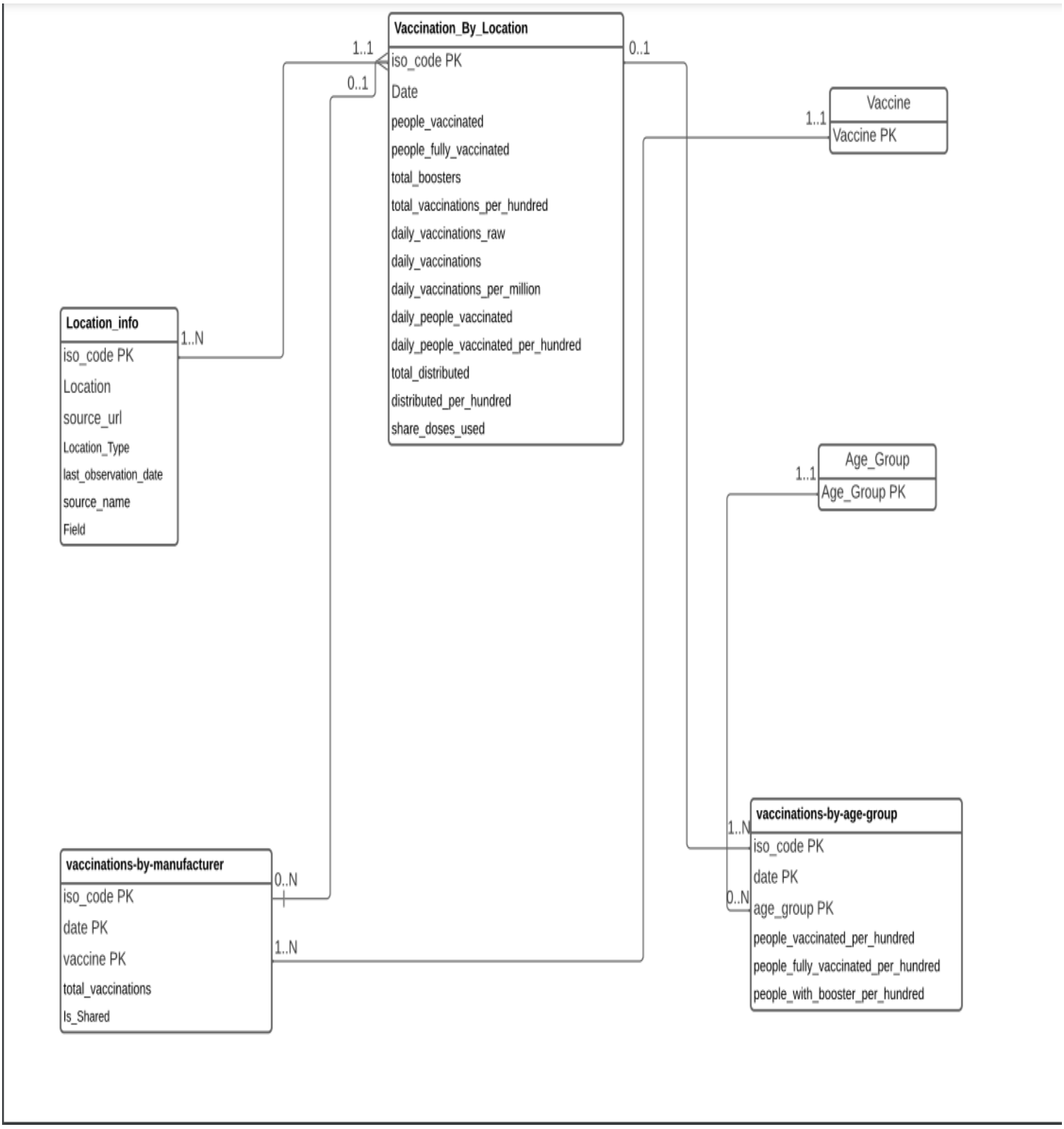
Contents

Some normalisation challenges were… 3

Assumptions 3

Database Schema 4

Design


Vaccination_By_Location
iso_code PK
Date
people_vaccinated
people_fully_vaccinated
total_boosters
total_vaccinations_per_hundred
daily_vaccinations_raw
daily_vaccinations
daily_vaccinations_per_million
daily_people_vaccinated
daily_people_vaccinated_per_hundred
total_distributed
distributed_per_hundred
share_doses_used
1..1
0..1
0..1
Vaccine
Vaccine PK
1..1
Location_info
iso_code PK
Location
source_url
Location_Type
last_observation_date
source_name
Field
1..N
Age_Group
Age_Group PK
1..1
vaccinations-by-manufacturer
iso_code PK
date PK
vaccine PK
total_vaccinations
Is_Shared
0..N
1..N
vaccinations-by-age-group
iso_code PK
date PK
age_group PK
people_vaccinated_per_hundred
people_fully_vaccinated_per_hundred
people_with_booster_per_hundred
1..N
0..N

Some normalisation challenges were…

Including US states in the normal locations, this requiered the location name to be used as an ISO_CODE.

Another challenge when it came to normalisation, was that the vaccine type in the files of the individual countries (Australia, Canada etc.) were all included in one field, what had to be done was if there was for example, Australia, Pfizer and Moderna Vaccine, and 1000 vaccinations. There were two rows created with Australia, Pfizer Vaccine, and 1000 vaccinations, and then Australia, Moderna Vaccine, and 1000 vaccinations. And the colunm IS_SHARED had to be created to ensure that there is not a confusion when it comes to the duplicated individual rows. As IS_SHARED reperesents if this number of total vaccines are shared with others.

Another challenge was row duplication, as there were the same named rows in different tables that had to be excluded. The only problem was that in some cases, even though the rows seemed auto-generated, they were either inconsistent, or not auto-generated.

Another issue, as said, was relying on the fact that the data was consistent between tables, this was more of an assumption, and an assumption that was relied on heavily throughout this process.

Furthermore, another challenge was that there were fields that are aleady in multiple tables but had inconsistent data.

Field names were consistantly ambiguos, and hard to dicipher their meanings. GitHub README was reffered to in order to assist in this challenge

Assumptions

1. In general data was integral
2. Removed Vaccine from the Master table assuming there is data integrity
3. people_with_booster_per_hundred = total_boosters_per_hundred
4. source_url = source_website

Database Schema

Location_info(iso_code, Location, source_url, Location_Type, last_observation_date, source_name)

Vaccination_By_Location(iso_code*, Date, people_vaccinated, people_fully_vaccinated, total_boosters, total_vaccinations_per_hundred, daily_vaccinations_raw, daily_vaccinations, daily_vaccinations_per_million, daily_people_vaccinated, daily_people_vaccinated_per_hundred, total_distributed, distributed_per_hundred, share_doses_used)

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

vaccinations_by_age_group(iso_code*, date, age_group*, people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred, people_with_booster_per_hundred)

Vaccine(vaccine)

age_group(age_group)