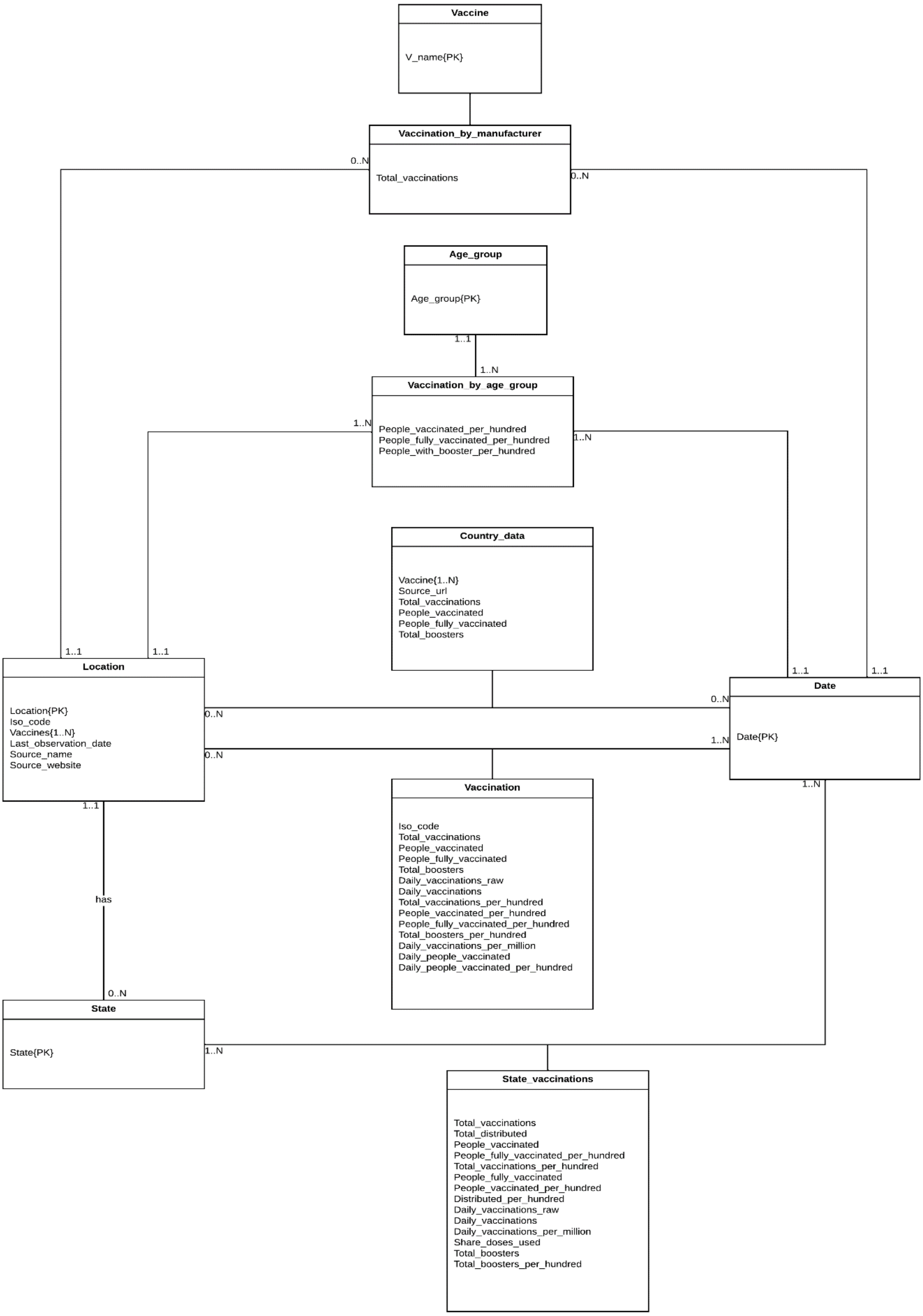
Part B

ER Diagram

Initial Design:



Problems:

1. When removing redundant data, there are several attributes in entites that can be derived from other attributes in vaccination.csv and us_state_vaccinations.csv. For example: Total_vaccinations_per_hundred can be derived from the formula:

   $$Sampling\_population = \frac{\text{Total\_vaccinations\_per\_hundred} \ \times 100}{\text{Total\_vaccinations}}$$

   By using the formula above, we can derive other attributes that end with “per_hundred” or “per_million” with their corresponding attributes. However, I found out that the calculated sampling population value for each attribute is different in vaccination entity which means I cannot derive all other attributes with a single sampling population. Therefore, I decided to keep every attributes that ends with “per_hundred” or “per_million”.

2. In country_data entity, there are several data repetition in the attributes (Total_vaccinations, People_vaccinated People_fully_vaccinated, Total_boosters). They can be found in the vaccination entity which is then removed.

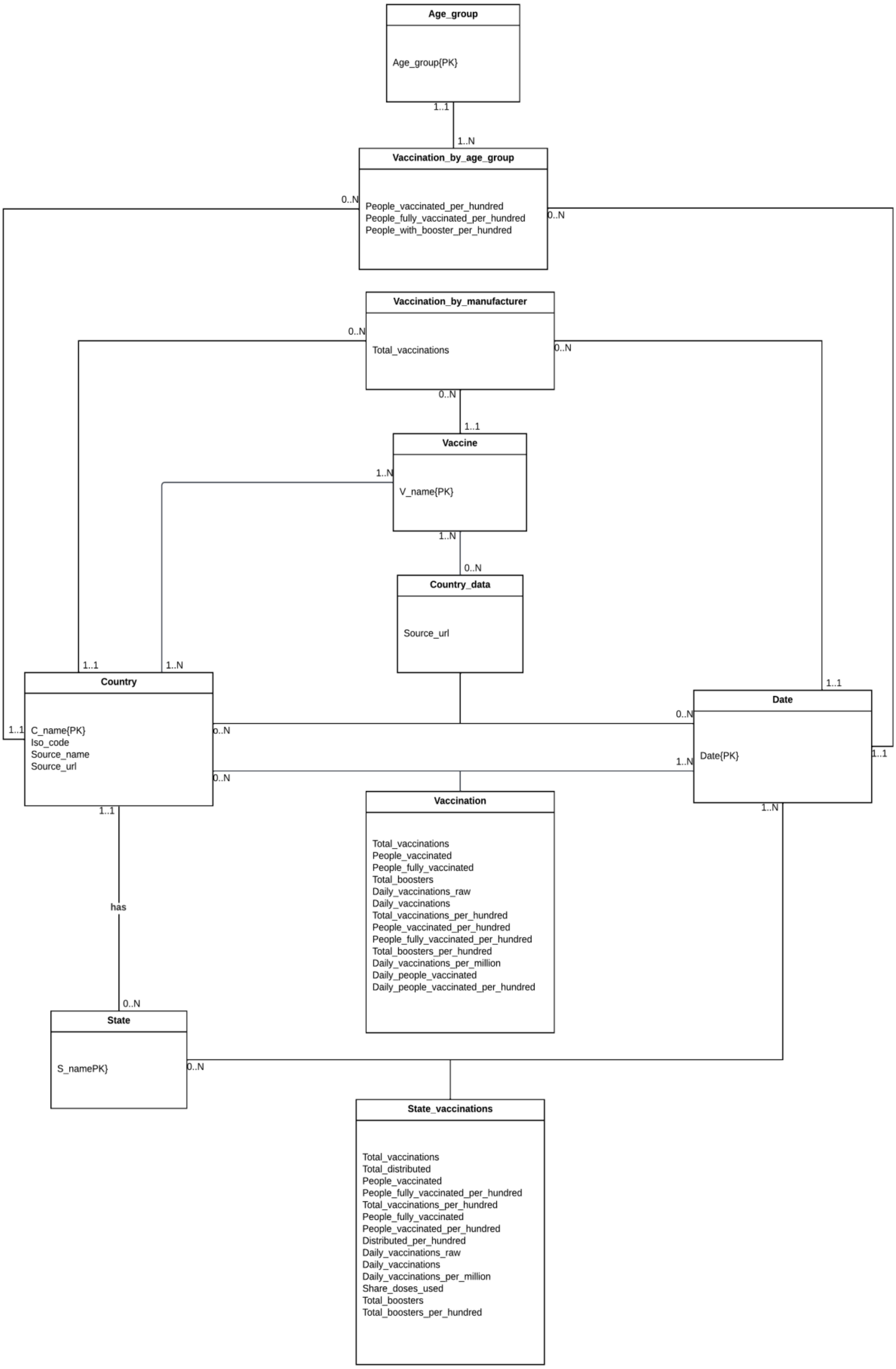
3. In vaccination entity, the Iso_code attribute can be derived directly from the C_name in Country entity. As there exist a FD where C_name -> Iso_code. Therefore, it is removed from the vaccination entity.

4. In location.csv, last_observation_date column of each country can be derived from the date column last row of each country in the vaccination.csv which means it can be removed.

5. In country_data entity and country entity, there is a multi-valued attributes named vaccine. One country in location.csv can have multiple vaccines while one vaccine can be used in different countries. Therefore, there is a M:N relationship between country entity and vaccine entity. The same case goes for country_data entity.

6. Notice that country entity has source_website attribute that can be found at the source_url column in the last row of each country in country_data entity. However, there are only four countries in the country_data entity which does not cover other countries in the country entity. Therefore, they are kept in their original state

Final Design:



Assumptions:

1. Not all dates stored in the date entity exist in the state_vaccinations relationship
2. Not all dates stored in the date entity exist in the vaccination relationship
3. Not all dates stored in the date entity exist in the country_data relationship
4. Not all dates stored in the date entity exist in the vaccination_by_manufacturer relationship
5. Not all dates stored in the date entity exist in the vaccination_by_age_group relationship
6. Some countries may not have states (i.e. we only keep states information in United States)
7. Some countries may not have country_data relationship with date (i.e. we only keep four countries data)
8. Not all countries are included in the vaccination_by_manufacturer relationship
9. Not all countries are included in the vaccination_by_age_group relationship
10. Not all countries stored in the vaccine entity exist in the country_data relationship
11. Not all vaccines stored in the vaccine entity exist in the vaccination_by_manufacturer relationship

Mapping ER Model to Relational Model

1. Map Strong Entities

   Country (C_name, Iso_code, Source_name, Source_url)
   State (S_name)
   Date (Date)
   Vaccine (V_name)
   Age_group (Age_group)

2. Map Weak Entities

   No weak entities found

3. Map 1:1 Relationships

   No 1:1 Relationships found

4. Map 1:N Relationships

   State (S_name, L_name*)

5. Map M:N Relationships

   State_vaccinations (S_name*, Date*, Total_vaccinations, Total_distributed People_vaccinated, People_fully_vaccinated_per_hundred, Total_vaccinations_per_hundred, People_fully_vaccinated, People_vaccinated_per_hundred, Distributed_per_hundred, Daily_vaccinations_raw, Daily_vaccinations, Daily_vaccinations_per_million, Share_doses_used, Total_boosters, Total_boosters_per_hundred)

Vaccination (C_name*, Date*, Total_vaccinations, People_vaccinated, People_fully_vaccinated, Total_boosters, Daily_vaccinations_raw, Daily_vaccinations, Total_vaccinations_per_hundred, People_vaccinated_per_hundred, People_fully_vaccinated_per_hundred, Total_boosters_per_hundred, Daily_vaccinations_per_million, Daily_people_vaccinated, Daily_people_vaccinated_per_hundred)
Country_data (C_name*, Date*, Source_url)
Country_data_vaccine (C_name*, Date*, V_name*)
Country_vaccine (C_name*, V_name*)

6. Multi-valued Attributes

   No multi-valued attributes found

7. Map higher-degree relationships
   Vaccination_by_manufacturer (C_name*, Date*, V_name*, Total_vaccinations)
   Vaccination_by_age_group (C_name*, Date*, Age_group*, People_vaccinated_per_hundred, People_fully_vaccinated_per_hundred, People_with_booster_per_hundred)

Complete Relational Model

Country (C_name, Iso_code, Source_name, Source_url)

State (S_name, L_name*)

Date (Date)

Vaccine (V_name)

Age_group (Age_group)

State_vaccinations (S_name*, Date*, Total_vaccinations, Total_distributed, People_vaccinated, People_fully_vaccinated_per_hundred, Total_vaccinations_per_hundred, People_fully_vaccinated, People_vaccinated_per_hundred, Distributed_per_hundred, Daily_vaccinations_raw, Daily_vaccinations, Daily_vaccinations_per_million, Share_doses_used, Total_boosters, Total_boosters_per_hundred)

Vaccination (C_name*, Date*, Total_vaccinations, People_vaccinated, People_fully_vaccinated, Total_boosters, Daily_vaccinations_raw, Daily_vaccinations, Total_vaccinations_per_hundred, People_vaccinated_per_hundred, People_fully_vaccinated_per_hundred, Total_boosters_per_hundred, Daily_vaccinations_per_million, Daily_people_vaccinated, Daily_people_vaccinated_per_hundred)

Country_data (C_name*, Date*, Source_url)

Country_data_vaccine (C_name*, Date*, V_name*)

Country_vaccine (C_name*, V_name*)

Vaccination_by_manufacturer (C_name*, Date*, V_name*, Total_vaccinations)

Vaccination_by_age_group (C_name*, Date*, Age_group*, People_vaccinated_per_hundred, People_fully_vaccinated_per_hundred, People_with_booster_per_hundred)

Normalization

Constructing FDs

1. Country

   FD1: C_name -> Iso_code, Source_name, Source_url

   Note that it is found that one source_name can have different source_url and one source_url can have different name

2. State

   FD1: S_Name -> L_name

3. Date

   No FDs

4. Vaccine

   No FDs

5. Age_group

   No FDs

6. State_vaccinations

   FD1: S_name, Date -> Total_vaccinations, Total_distributed, People_vaccinated, People_fully_vaccinated_per_hundred, Total_vaccinations_per_hundred, People_fully_vaccinated, People_vaccinated_per_hundred, Distributed_per_hundred, Daily_vaccinations_raw, Daily_vaccinations, Daily_vaccinations_per_million, Share_doses_used, Total_boosters, Total_boosters_per_hundred

7. Vaccination

   FD1: C_name, Date -> Total_vaccinations, People_vaccinated, People_fully_vaccinated, Total_boosters, Daily_vaccinations_raw, Daily_vaccinations, Total_vaccinations_per_hundred, People_vaccinated_per_hundred, People_fully_vaccinated_per_hundred, Total_boosters_per_hundred, Daily_vaccinations_per_million, Daily_people_vaccinated, Daily_people_vaccinated_per_hundred

8. Country_data

   FD1: C_name, Date -> Source_url

9. Country_data_vaccine

   No FDs

10. Country_vaccine
    No FDs

11. Vaccination_by_manufacturer
    FD1: Vaccine_name, C_name, Date -> Total_vaccinations_per_vaccine

12. Vaccination_by_age_group
    FD1: Age_group, C_name, Date -> People_vaccinated_per_hundred, People_fully_vaccinated_per_hundred, People_with_booster_per_hundred

Primary Keys for relations

1. Country
   PK: C_name

2. State
   PK: S_name

3. Date
   PK: Date

4. Vaccine
   PK: V_name

5. Age_group
   PK: Age_group

6. State_vaccinations
   PK: S_name, Date

7. Vaccination
   PK: L_name, Date

8. Country_data
   PK: L_name, Date

9. Country_data_vaccine
   PK: L_name, Date, V_name

10. Country_vaccine
    PK: L_name, V_name

11. Vaccination_by_manufacturer
    PK: V_name, L_name, Date

12. Vaccination_by_age_group
    PK: Age_group, L-name, Date

Decomposition

1. Country
   Relation already in 3NF

2. State
   Relation already in 3NF

3. Date
   Relation already in 3NF

4. Vaccine
   Relation already in 3NF

5. Age_group
   Relation already in 3NF

6. State_vaccinations
   Relation already in 3NF

7. Vaccination
   Relation already in 3NF

8. Country_data
   PK: L_name, Date

9. Country_data_vaccine
   Relation already in 3NF

10. Country_vaccine
    Relation already in 3NF

11. Vaccination_by_manufacturer
    Relation already in 3NF

12. Vaccination_by_age_group
    Relation already in 3NF

Combine Relations

R1: Country (C_name, Iso_code, Source_name, Source_url)

R2: State (S_name, L_name*)

R3: Date (Date)

R4: Vaccine (V_name)

R5: Age_group (Age_group)

R6: State_vaccinations (S_name*, Date*, Total_vaccinations, Total_distributed, People_vaccinated, People_fully_vaccinated_per_hundred, Total_vaccinations_per_hundred, People_fully_vaccinated, People_vaccinated_per_hundred, Distributed_per_hundred, Daily_vaccinations_raw, Daily_vaccinations, Daily_vaccinations_per_million, Share_doses_used, Total_boosters, Total_boosters_per_hundred)

R7: Vaccination (C_name*, Date*, Total_vaccinations, People_vaccinated, People_fully_vaccinated, Total_boosters, Daily_vaccinations_raw, Daily_vaccinations, Total_vaccinations_per_hundred, People_vaccinated_per_hundred, People_fully_vaccinated_per_hundred, Total_boosters_per_hundred, Daily_vaccinations_per_million, Daily_people_vaccinated, Daily_people_vaccinated_per_hundred)

R8: Country_data (C_name*, Date*, Source_url)

R9: Country_data_vaccine (C_name*, Date*, V_name*)

R10: Country_vaccine (C_name*, V_name*)

R11: Vaccination_by_manufacturer (C_name*, Date*, V_name*, Total_vaccinations)

R12: Vaccination_by_age_group (C_name*, Date*, Age_group*, People_vaccinated_per_hundred, People_fully_vaccinated_per_hundred, People_with_booster_per_hundred)

Final relational database schema

R1: Country (C_name, Iso_code, Source_name, Source_url)

R2: State (S_name, L_name*)

R3: Date (Date)

R4: Vaccine (V_name)

R5: Age_group (Age_group)

R6: State_vaccinations (S_name*, Date*, Total_vaccinations, Total_distributed, People_vaccinated, People_fully_vaccinated_per_hundred, Total_vaccinations_per_hundred, People_fully_vaccinated, People_vaccinated_per_hundred, Distributed_per_hundred, Daily_vaccinations_raw, Daily_vaccinations, Daily_vaccinations_per_million, Share_doses_used, Total_boosters, Total_boosters_per_hundred)

R7: Vaccination (C_name*, Date*, Total_vaccinations, People_vaccinated, People_fully_vaccinated, Total_boosters, Daily_vaccinations_raw, Daily_vaccinations, Total_vaccinations_per_hundred, People_vaccinated_per_hundred, People_fully_vaccinated_per_hundred, Total_boosters_per_hundred, Daily_vaccinations_per_million, Daily_people_vaccinated, Daily_people_vaccinated_per_hundred)

R8: Country_data (C_name*, Date*, Source_url)

R9: Country_data_vaccine (C_name*, Date*, V_name*)

R10: Country_vaccine (C_name*, V_name*)

R11: Vaccination_by_manufacturer (C_name*, Date*, V_name*, Total_vaccinations)

R12: Vaccination_by_age_group (C_name*, Date*, Age_group*, People_vaccinated_per_hundred, People_fully_vaccinated_per_hundred, People_with_booster_per_hundred)