



DATABASE DESIGN PROJECT

Database Concepts, Due: 29 Oct

Samuel Mainwood
S3939120

Part A Reflection:

Data was observed and downloaded via the Github. The project walkthrough was followed to put all data files in 1NF.

For some clarity, the Readme was followed as well as the journal article linked. It was challenging to be certain of repeating/redundant information with the vaccination rates. I also was fortunate to pick up that Australia's Source URL changed on the final observation which affected my design.

There is redundant data in the
Calculated population by dividing:
 $X = \text{pop_value_by_hundred}$
 $Y = \text{total_value}$

$\text{Population} = y / (x/100).$

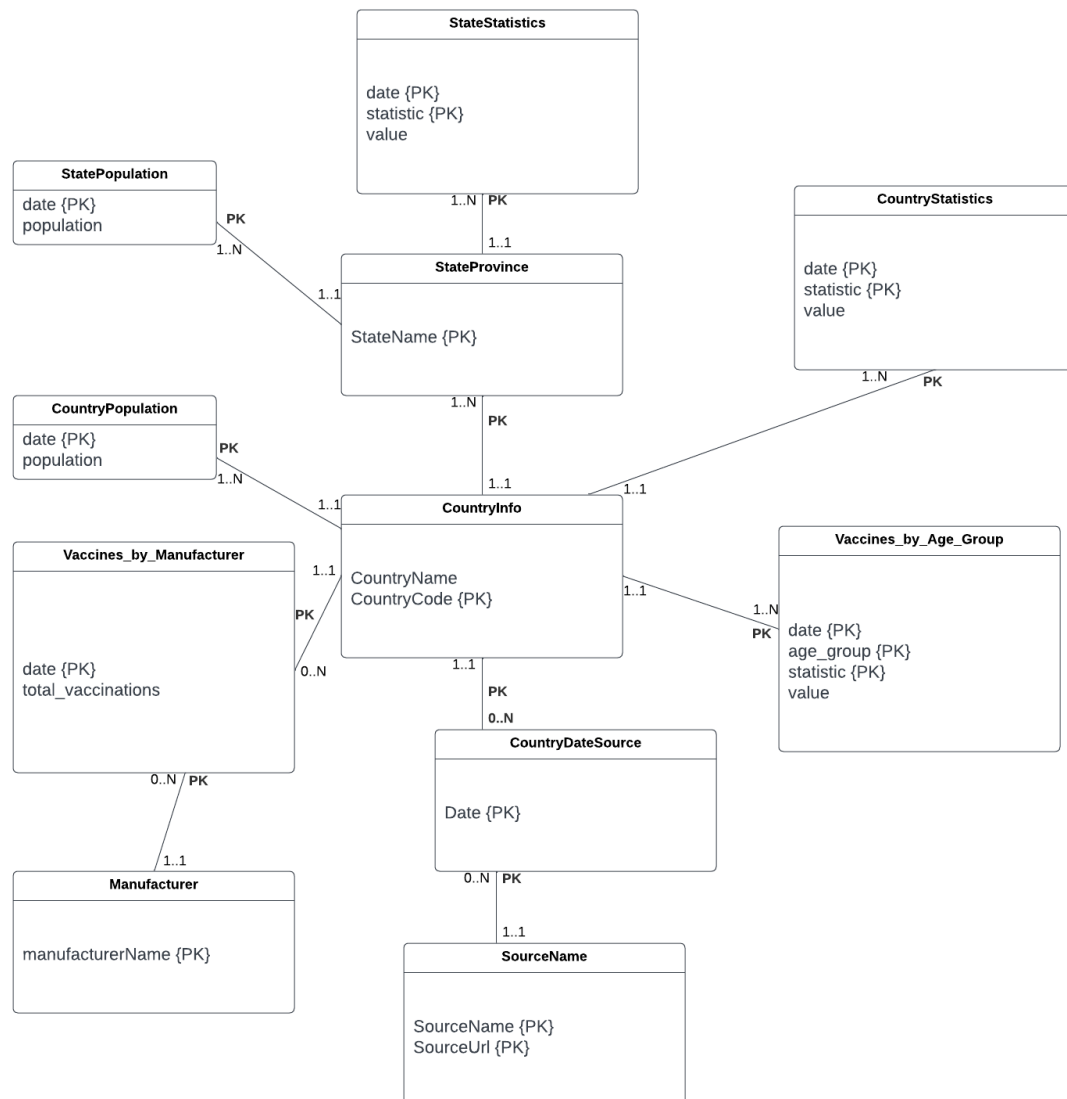
Daily Vaccinations and Daily Vaccinations Per Hundred were the best candidates for X and Y as they appeared to have the most complete information.

Aggregate data (income-levels, continents) was removed from files.

Part B: Database Design

1. Entity – Relationship Diagram

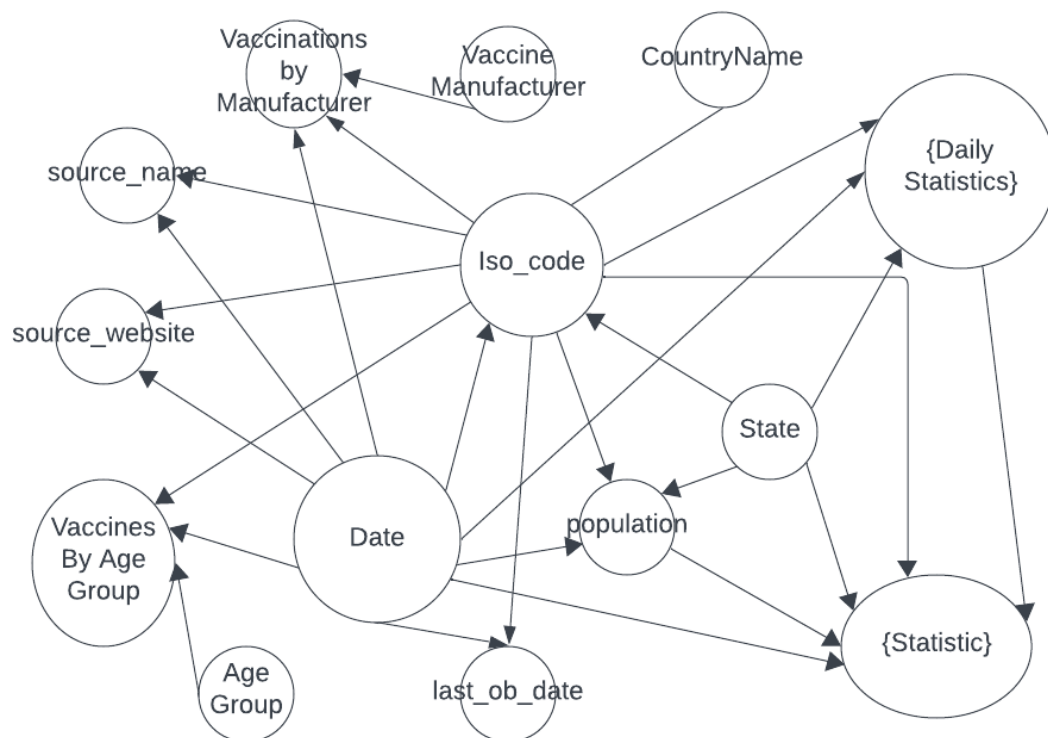
Figure 1: ER Diagram



2. Explanation of Normalisation Changes

A simplified FD diagram is below. All stats were seen to be functionally dependent on States, Countries, Manufactures and Sources were identified as strong entities. A simplified diagram is below in figure 2. The daily statistics of “daily people vaccinated” etc., and “total people vaccinated” etc. have been aggregated into a {Statistic} nodes.

Figure 2: FDs



This results in the Following Dependencies:

{IsoCode} → {Country Name}

{IsoCode, Date} → {Statistic}

{IsoCode/State, Date} → Population

{State} → IsoCode

Note, this allows scalability if the data was to incorporate more than just US states

{State, IsoCode, date, {Statistic} → StatisticValue

{State/IsoCode, Date, Population, {Statistic}} → Relative Daily Stat (i.e Vaccines_per_hundred)

Note, this does not need to be stored as it is a calculation based on other available data

{Country, LastObsDate} → Source Website

*Note, Last Obs Date is used to differentiate Australia having 2 different URLs as sources at different dates.

{Country, Date} → {SourceWebsite, SourceName}

{VaccineManufacturer, Country, Date} → VaccinesByManufacturer

Normalisation for Statistic Attributes:

Recognising that there are multiple overlapping primary keys that can be aggregated, the statistics columns were pivoted to conform with 3NF restrictions.

1NF Table

<u>Country</u>	<u>Date</u>	Stat1	Stat2	Stat3
Australia	03/03/2022	X	Y	Z

3NF

<u>Country</u>	<u>Date</u>	<u>Statistic</u>	Value
Australia	03/03/2022	Stat1	X
Australia	03/03/2022	Stat2	Y
Australia	03/03/2022	Stat3	Z

This maintained all values in all statistic columns whilst also minimising data storage of nulls whilst achieving 3NF.

3. Database Schema

1. Map Strong Entities

Country (CountryCode, CountryName)

Source (SourceUrl, SourceName)

Manufacturer (ManufacturerName)

2. Map Weak Entities:

State (CountryCode*, StateName)

StatePopulation (CountryCode*, StateName*, date, population)

CountryPopulation (CountryCode *, date, population)

3. 1:1 Relationships:

None

4. 1:N Relationships:

StateStatistics (date, StateName*, CountryCode *, statistic, value)

CountryStatistics (date, CountryCode*, statistic, value)

VaccinesByAgeGroup (date, CountryCode*, ageGroup, statistic, value)

5. M:N Relationships

VaccinesByManufacturer (date, CountryCode *, ManufacturerName*, total_vaccinations)

CountryDataSource (LastObsDate, CountryCode*, SourceUrl*, SourceName)

6. Multi-valued-attributes:

None

7. Higher-Degree Relationships:

None

Final Schema

Country (CountryCode, CountryName)

Source (SourceUrl, SourceName)

Manufacturer (ManufacturerName)

State (CountryCode*, StateName)

StatePopulation (CountryCode*, StateName*, date, population)

CountryPopulation (CountryCode*, date, population)

StateStatistics (date, StateName*, CountryCode *, statistic, value)

CountryStatistics (date, CountryCode*, statistic, value)

VaccinesByAgeGroup (date, CountryCode*, ageGroup, statistic, value)

VaccinesByManufacturer (date, CountryCode*, ManufacturerName*,
total_vaccinations)

CountryDataSource (LastObsDate, CountryCode*, SourceUrl*, SourceName*)