SMU Continuing & Professional Education

ETL - COVID Vaccination Data

Project 2 Group #4

Megan Adams

Calvin Cusick

Enoc Serge Kouegbe

Joseph Onwukeme

Pre-Processing

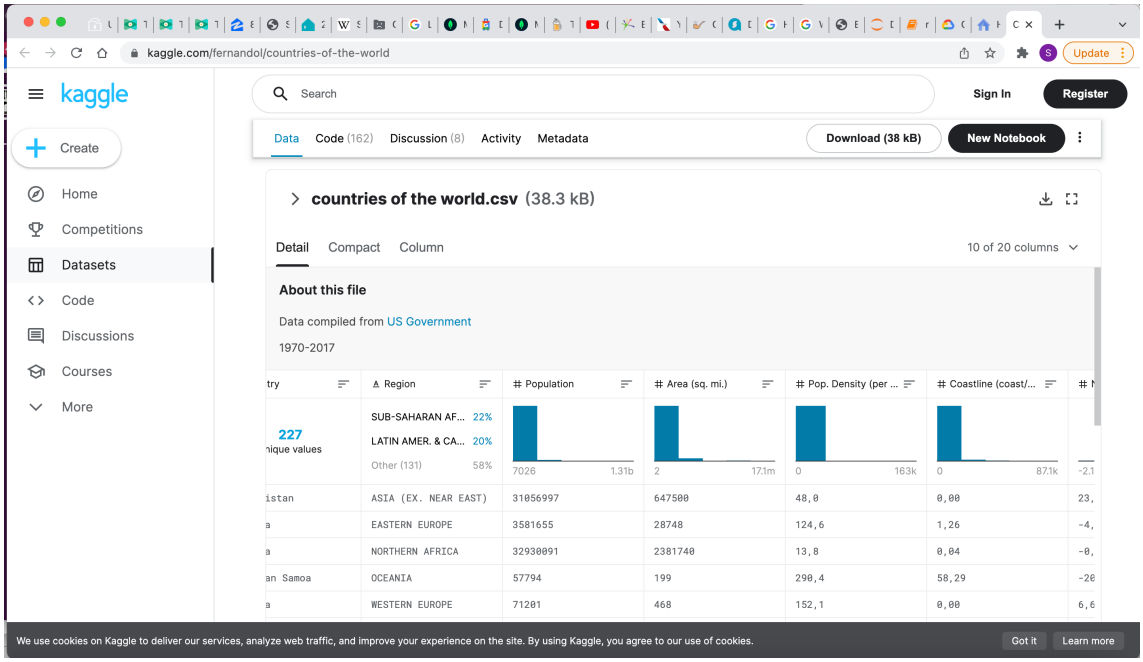
Our project explores vaccination rates for Covid-19 globally. We looked at data from multiple sources: CDC, WHO, University of Oxford's Global Change Data Lab, NY Times, Kaggle, and Google Cloud Platform's Covid-19 public datasets. We focused on the public dataset from GCP because it included information on vaccine provider by country. We also pulled country metadata from Kaggle to provide an interesting comparison.

Once the data was obtained using multiple CSV files, we cleaned and organized our data using Pandas in Jupyter notebook. Once the data was organized in an accessible read format, the step was to transfer our final output to pgAdmin using Google Cloud Platform.

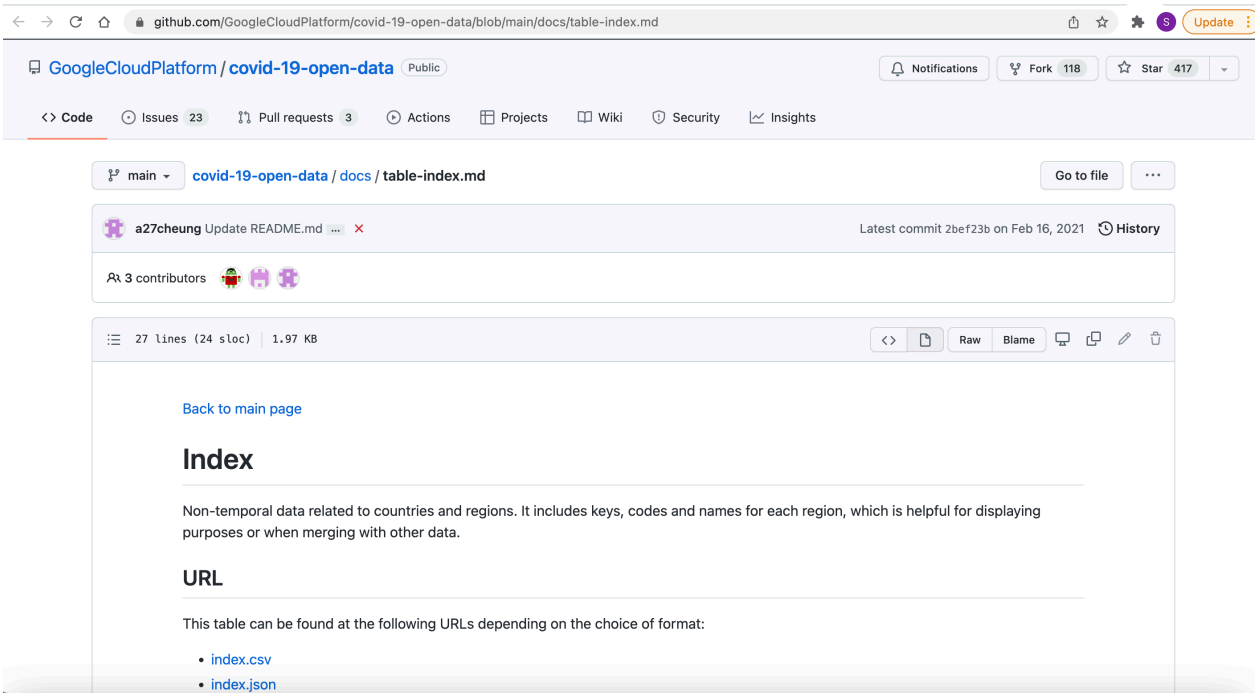
Extraction

We used two different datasets: the public platform Kaggle which led us to Country names linked to region, population, area size, GDP, mortality, etc. The second CSV file is from Google Cloud Platform containing COVID vaccination data by providers from many countries worldwide.

https://www.kaggle.com/fernandol/countries-of-the-world



https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-index.md



Transformation

To transform the public data and use it in our study, we performed the following:

- Used Pandas functions in Jupyter Notebook to load both CSV files.
- Reviewed the files and transformed them into data frames
- We also streamlined the data set to decrease repetition using the vaccination ID.
- We then shifted the six columns for each vaccine into rows sorted by date, country, and vaccine ID.
- Pulled data vaccine by providers
- Used a mask to pull some specifics countries relevant for the focus of our study.
- Limited our second dataset only to eleven countries (Population, infant mortality, GDP, birthrate, and death rate)
- Created vaccination id CSV

Countries Selected

In [25]:
```
df3 = pd.read_csv(filepath2)
df3
```

Out[25]:

	country_code	country	population	infant_mortality	gdp	birthrate	deathrate
0	AU	Australia	20264082	4.69	29000	12.14	7.51
1	BR	Brazil	188078227	29.61	7600	16.56	6.17
2	CA	Canada	33098932	4.75	29800	10.78	7.80
3	CN	China	1313973713	24.18	5000	13.25	6.97
4	CI	Cote d'Ivoire	17654843	90.83	1400	35.11	14.84
5	EG	Egypt	78887007	32.59	4000	22.94	5.23
6	IN	India	1095351995	56.29	2900	22.01	8.18
7	MX	Mexico	107449525	20.91	9000	20.69	4.74
8	NG	Nigeria	131859731	98.80	900	40.43	16.94
9	US	United States	298444215	6.50	37800	14.14	8.26
10	ZW	Zimbabwe	12236805	67.69	1900	28.01	21.84

Created vaccination id CSV

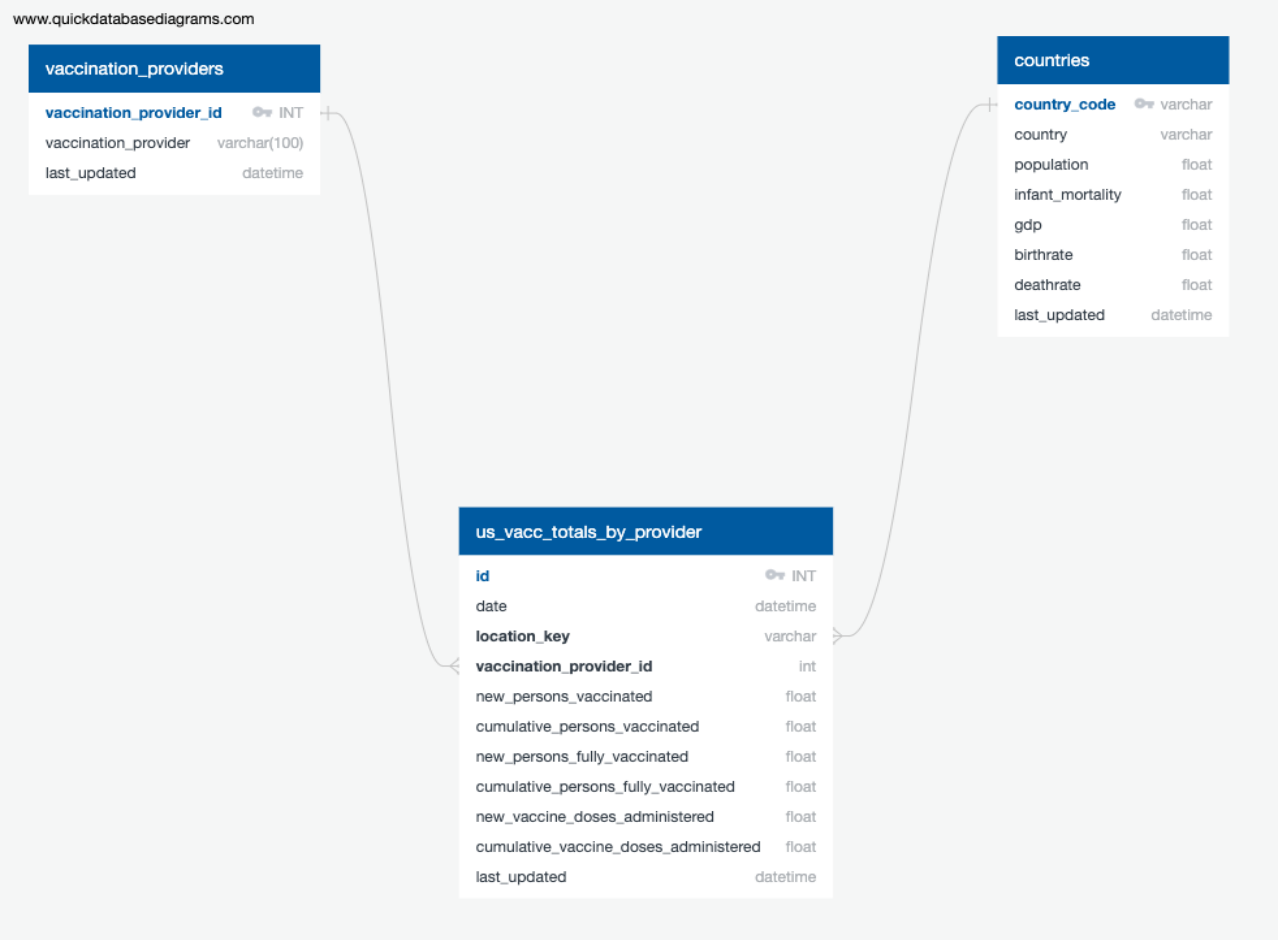
In [3]:
```
#create vaccination id csv
data = [[1, 'vaccination_totals'], [2, 'pfizer'], [3, 'moderna'], [4, 'jassen'], [5, 'sinovac']]
vacc_id_df = pd.DataFrame(data, columns = ['vaccination_provider_id', 'vaccination_provider'])
vacc_id_df
```

Out[3]:

	vaccination_provider_id	vaccination_provider
0	1	vaccination_totals
1	2	pfizer
2	3	moderna
3	4	jassen
4	5	sinovac

ERD Diagram

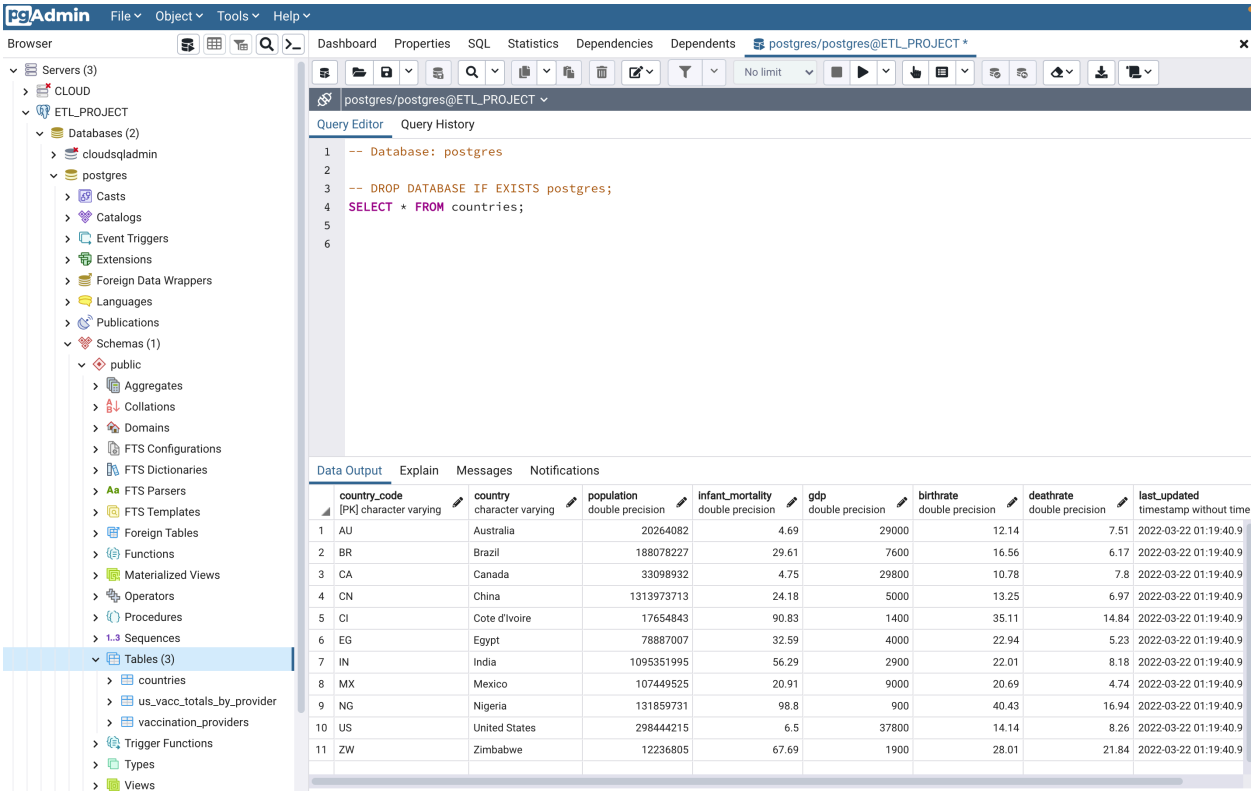
After transforming the data, we conducted a data engineering and Entity-Relationship Diagram (ERD) using an open-source toolkit called Quickdatabasediagrams. The model looks as follows:



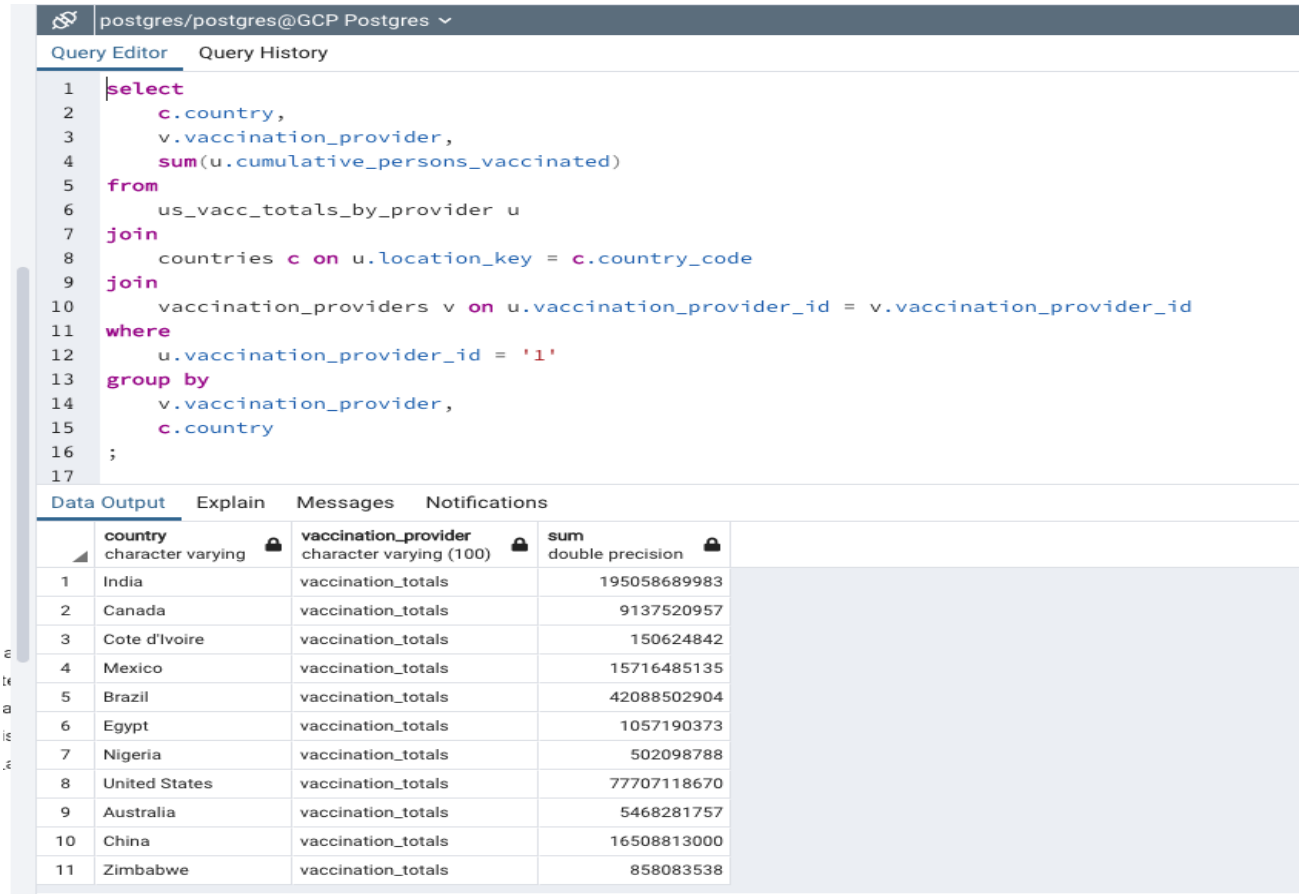
Load

We then used pandas in a jupyter notebook to load the tables. We did an initial connection to the Postgres database using PG admin to store our original clean data sets.

Postgres Database:



Queries



postgres/postgres@ETLProject

Query Editor | Query History

```
SELECT
*
from
us_vacc_totals_by_provider


Select
us.date,
co.country_code,
us.cumulative_persons_vaccinated,
us.cumulative_persons_fully_vaccinated,
us.cumulative_vaccine_doses_administered
FROM
us_vacc_totals_by_provider as us
inner join countries as co
    on co.country_code=us.location_key
WHERE
co.country_code = 'US'
ORDER BY
    us.date asc;
```

Data Output | Explain | Messages | Notifications

	date date	country_code character varying	cumulative_persons_vaccinated double precision	cumulative_persons_fully_vaccinated double precision	cumulative_vaccine_doses_administered double precision
1	2020-12-13	US	[null]	[null]	[null]
2	2020-12-13	US	[null]	[null]	[null]
3	2020-12-13	US	[null]	[null]	[null]
4	2020-12-13	US	[null]	[null]	[null]
5	2020-12-13	US	24614	5706	29586
6	2020-12-14	US	29025	5825	34157
7	2020-12-14	US	[null]	[null]	[null]
8	2020-12-14	US	[null]	[null]	[null]
9	2020-12-14	US	[null]	[null]	[null]
10	2020-12-14	US	[null]	[null]	[null]
11	2020-12-15	US	[null]	[null]	[null]
12	2020-12-15	US	76421	6085	83882
13	2020-12-15	US	[null]	[null]	[null]

Summary

There were some limitations to our project due to the data available. With more time and better resources, we could conceivably answer more questions on Covid vaccination worldwide.