

# **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

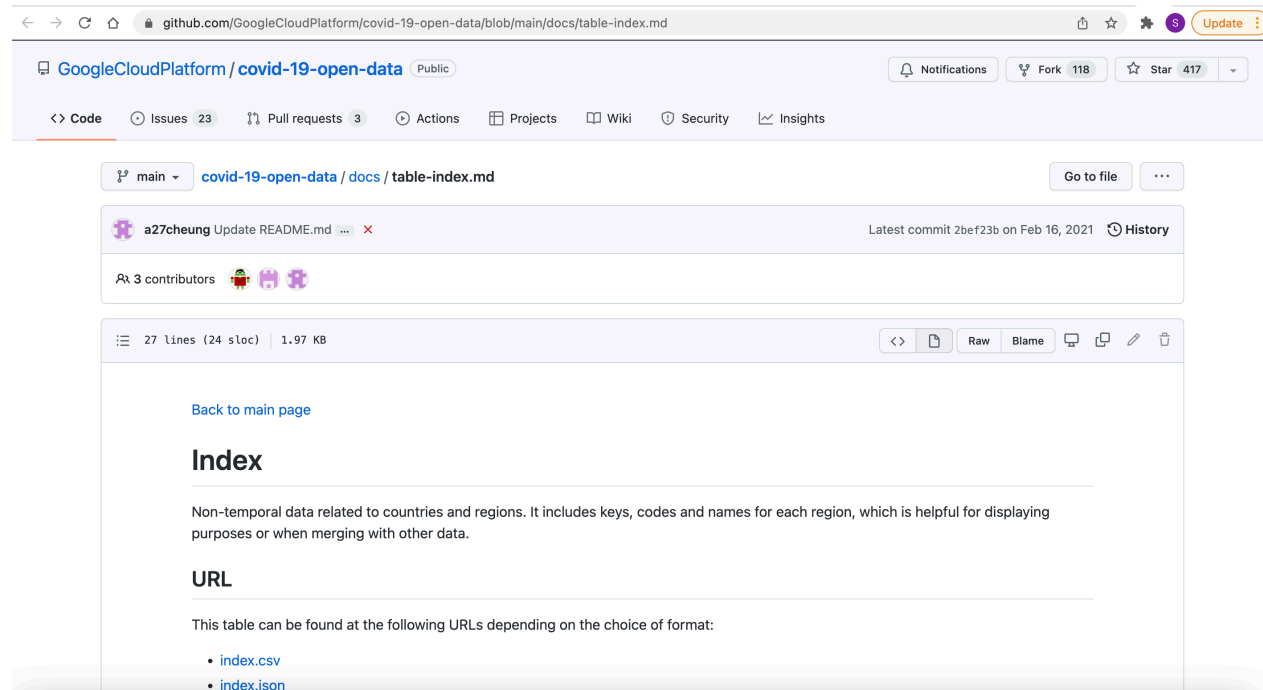
We used two different datasets: the public platform Kaggle which led us to Country names linked to region, population, area size, GDP, mortality, etc. The second CSV file is from Google Cloud Platform containing COVID vaccination data by providers from many countries worldwide.

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

The screenshot shows the Kaggle website interface for the 'countries of the world.csv' dataset. The page includes a sidebar with navigation links, a search bar, and a table of dataset details. The table shows columns for Country, Region, Population, Area, Population Density, and Coastline. A bar chart shows the distribution of countries by region: SUB-SAHARAN AF... (22%), LATIN AMER. & CA... (20%), and Other (131) (58%).

Country	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area)	Population Density (per sq. mi.)
istan	ASIA (EX. NEAR EAST)	31856997	647500	48,0	0,00	23,
a	EASTERN EUROPE	3581655	28748	124,6	1,26	-4,
a	NORTHERN AFRICA	32930091	2381740	13,8	0,04	-0,
an Samoa	OCEANIA	57794	199	290,4	58,29	-20
a	WESTERN EUROPE	71261	468	152,1	0,00	6,6

<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

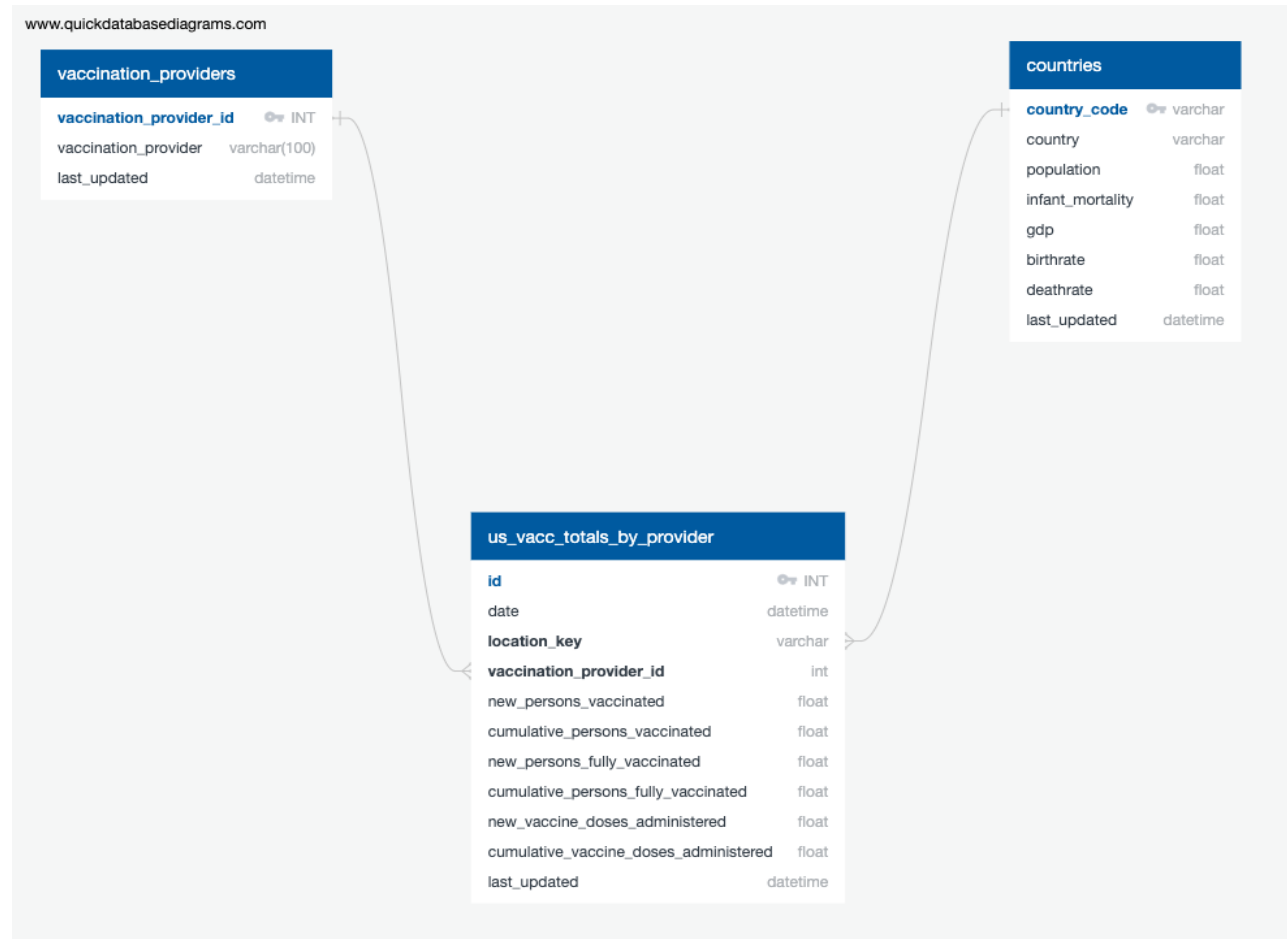
```
In [3]: #create vaccination id csv
data = [[1, 'vaccination_totals'], [2, 'pfizer'], [3, 'moderna'], [4, 'jassen'], [5, 'sinovac']]
vacc_id_df = pd.DataFrame(data, columns = ['vaccination_provider_id', 'vaccination_provider'])
vacc_id_df
```

```
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:

**pgAdmin** File Object Tools Help

Browser: postgres/postgres@ETL\_PROJECT \*

Servers (3)  
CLOUD  
ETL\_PROJECT  
Databases (2)  
cloudsqladmin  
postgres  
Casts  
Catalogs  
Event Triggers  
Extensions  
Foreign Data Wrappers  
Languages  
Publications  
Schemas (1)  
public  
Aggregates  
Collations  
Domains  
FTS Configurations  
FTS Dictionaries  
FTS Parsers  
FTS Templates  
Foreign Tables  
Functions  
Materialized Views  
Operators  
Procedures  
Sequences  
Tables (3)  
countries  
us\_vacc\_totals\_by\_provider  
vaccination\_providers  
Trigger Functions  
Types  
Views

Query Editor Query History

```
-- Database: postgres
-- DROP DATABASE IF EXISTS postgres;
SELECT * FROM countries;
```

Data Output Explain Messages Notifications

	country_code [PK] character varying	country character varying	population double precision	infant_mortality double precision	gdp double precision	birthrate double precision	deathrate double precision	last_updated timestamp without time
1	AU	Australia	20264082	4.69	29000	12.14	7.51	2022-03-22 01:19:40.9
2	BR	Brazil	188078227	29.61	7600	16.56	6.17	2022-03-22 01:19:40.9
3	CA	Canada	33098932	4.75	29800	10.78	7.8	2022-03-22 01:19:40.9
4	CN	China	1313973713	24.18	5000	13.25	6.97	2022-03-22 01:19:40.9
5	CI	Cote d'Ivoire	17654843	90.83	1400	35.11	14.84	2022-03-22 01:19:40.9
6	EG	Egypt	78887007	32.59	4000	22.94	5.23	2022-03-22 01:19:40.9
7	IN	India	1095351995	56.29	2900	22.01	8.18	2022-03-22 01:19:40.9
8	MX	Mexico	107449525	20.91	9000	20.69	4.74	2022-03-22 01:19:40.9
9	NG	Nigeria	131859731	98.8	900	40.43	16.94	2022-03-22 01:19:40.9
10	US	United States	298444215	6.5	37800	14.14	8.26	2022-03-22 01:19:40.9
11	ZW	Zimbabwe	12236805	67.69	1900	28.01	21.84	2022-03-22 01:19:40.9

## Queries

postgres/postgres@GCP Postgres

Query Editor Query History

```
select
  c.country,
  v.vaccination_provider,
  sum(u.cumulative_persons_vaccinated)
from
  us_vacc_totals_by_provider u
join
  countries c on u.location_key = c.country_code
join
  vaccination_providers v on u.vaccination_provider_id = v.vaccination_provider_id
where
  u.vaccination_provider_id = '1'
group by
  v.vaccination_provider,
  c.country
;
```

Data Output Explain Messages Notifications

	country character varying	vaccination_provider character varying (100)	sum double precision
1	India	vaccination_totals	195058689983
2	Canada	vaccination_totals	9137520957
3	Cote d'Ivoire	vaccination_totals	150624842
4	Mexico	vaccination_totals	15716485135
5	Brazil	vaccination_totals	42088502904
6	Egypt	vaccination_totals	1057190373
7	Nigeria	vaccination_totals	502098788
8	United States	vaccination_totals	77707118670
9	Australia	vaccination_totals	5468281757
10	China	vaccination_totals	16508813000
11	Zimbabwe	vaccination_totals	858083538

postgres/postgres@ETLProject

Query EditorQuery History

1SELECT

2\*

3from

4us\_vacc\_totals\_by\_provider

5

6

7Select

8us.date,

9co.country\_code,

10us.cumulative\_persons\_vaccinated,

11us.cumulative\_persons\_fully\_vaccinated,

12us.cumulative\_vaccine\_doses\_administered

13FROM

14us\_vacc\_totals\_by\_provider as us

15inner join countries as co

16on co.country\_code=us.location\_key

17WHERE

18co.country\_code = 'US'

19ORDER BY

20us.date asc;

21

Data OutputExplainMessagesNotifications

	<div>date</div> <div>date</div>	<div>country_code</div> <div>character varying</div>	<div>cumulative_persons_vaccinated</div> <div>double precision</div>	<div>cumulative_persons_fully_vaccinated</div> <div>double precision</div>	<div>cumulative_vaccine_doses_administered</div> <div>double precision</div>	
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.