COVID-19 Vaccination Visualization and Analysis

Judy Jin, Yaoyuan Luo

Introduction

In this project, we analyze and visualize the dataset for COVID-19 cases and vaccination. In order to do the visualization, we use 2 datasets (vaccination dataset and cases dataset) and a supplement country dataset for plotting geographic maps. In the project, we perform 4 visualization tasks to understand the COVID-19 vaccination from different perspectives. First, we do a visualization overview for global and country vaccination. Second, we perform a detailed visualization for top 10 countries. Third, we plot a choropleth to understand the country vaccination pace. Last, we combine the cases data and the vaccination data to understand if there is a relationship between the two using a scatter plot.

Motivation

COVID-19 has spread rapidly worldwide since the end of 2019, which has caused a severe economic loss and population death. Therefore, we want to choose data that is related to COVID-19 as our analysis target for this project. However, most institutions have created plots for COVID-19 confirmed cases and deaths. We decide not to focus our project on analyzing confirmed cases. We notice that most countries have begun their vaccination process since this could be the solution to end the pandemic. Thus, we want to have an understanding of the current vaccination process and we also want to look at both COVID-19 cases data and the vaccination data to not only understand the relationship between them but also may suggest potential reallocation of the vaccination to help countries that have a shortage.

Data Pipeline

Data Source

In our project, we use 2 datasets and 1 supplement datasets.

The first dataset (vaccination) comes from Kaggle, which contains information of COVID-19 vaccination for each country and each day. This dataset has 15 attributes (country, iso_code, date, total_vaccinations, people_vaccinated, people_fully_vaccinated, daily_vaccinations_raw, daily_vaccinations, total_vaccinations_per_hundred, people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred, daily_vaccinations_per_million, vaccines, source_name, source_website) and a total of

4043 entries.

vaccination

	country	iso_code	date	total_vaccinations	people_vaccinated
0	Albania	ALB	2021-01-10	0.0	0.0
1	Albania	ALB	2021-01-11	NaN	NaN
2	Albania	ALB	2021-01-12	128.0	128.0
3	Albania	ALB	2021-01-13	188.0	188.0
4	Albania	ALB	2021-01-14	266.0	266.0
...	...	...	...	...	...
4038	Zimbabwe	ZWE	2021-02-19	NaN	NaN
4039	Zimbabwe	ZWE	2021-02-20	NaN	NaN
4040	Zimbabwe	ZWE	2021-02-21	NaN	NaN
4041	Zimbabwe	ZWE	2021-02-22	1314.0	1314.0
4042	Zimbabwe	ZWE	2021-02-23	4041.0	4041.0

4043 rows × 15 columns

Fig 1. Data “vaccination”

The second dataset (cases) comes from GitHub, which contains information of COVID-19 for each country and each day. This dataset has 5 attributes (Date, Country, Confirmed, Recovered, Deaths) and a total of 76800 entries.

cases

	Date	Country	Confirmed	Recovered	Deaths
0	2020-01-22	Afghanistan	0	0	0
1	2020-01-23	Afghanistan	0	0	0
2	2020-01-24	Afghanistan	0	0	0
3	2020-01-25	Afghanistan	0	0	0
4	2020-01-26	Afghanistan	0	0	0
...	...	...	...	...	...
76795	2021-02-20	Zimbabwe	35768	32096	1432
76796	2021-02-21	Zimbabwe	35796	32125	1436
76797	2021-02-22	Zimbabwe	35862	32216	1441
76798	2021-02-23	Zimbabwe	35910	32288	1448
76799	2021-02-24	Zimbabwe	35960	32410	1456

76800 rows × 5 columns

Fig 2. Data “cases”

The third dataset (country) is a supplement dataset which is only used for plotting geographic visualizations. This country dataset comes from the data provided for DSC106 HW4.

country

	id	name	alpha2	alpha3
0	4	Afghanistan	af	afg
1	8	Albania	al	alb
2	12	Algeria	dz	dza
3	20	Andorra	ad	and
4	24	Angola	ao	ago
...	...	...	...	...
188	862	Venezuela (Bolivarian Republic of)	ve	ven
189	704	Viet Nam	vn	vnm
190	887	Yemen	ye	yem
191	894	Zambia	zm	zmb
192	716	Zimbabwe	zw	zwe

193 rows × 4 columns

Fig 3. Data “country”

Data Filtering and Selection

For the vaccination dataset, we only care about the process of vaccination. We delete columns regarding vaccination source, while keeping the columns regarding vaccination number and people vaccinated.

For the other two dataset, we keep all the information for our analysis.

Data Preprocessing

For the vaccination dataset, we notice that there is missing data for some date. We preprocess the data to fill the NaN field with the previous number for columns “total_vaccinations", "people_vaccinated", "people_fully_vaccinated", "total_vaccinations_per_hundred",

"people_vaccinated_per_hundred", and "people_fully_vaccinated_per_hundred". We also notice that some countries do not have up-to-date information. In this case, we do not add entries for the latest data to make sure we don't misinterpret the data.

For the other two dataset, we do not preprocess the data to keep all the data clean. However, in each visualization task, we will perform specific data processing to get the ideal visualization.

Visualization Methods and Analysis

Vaccination Overview

For the first visualization, we want to have an overview on the vaccination process. Thus, we use the vaccination dataset and plot the number of vaccinations over time using a line chart. To incorporate the country information, we encode the country information as hue so that users can compare different countries by different colors. Moreover, we provide interactive features that allow users to zoom in and out for more detailed information. We also add a selection feature that allows users to select specific countries that they want to learn more about. Users can also learn more information by pointing to the specific point to read the tooltip.

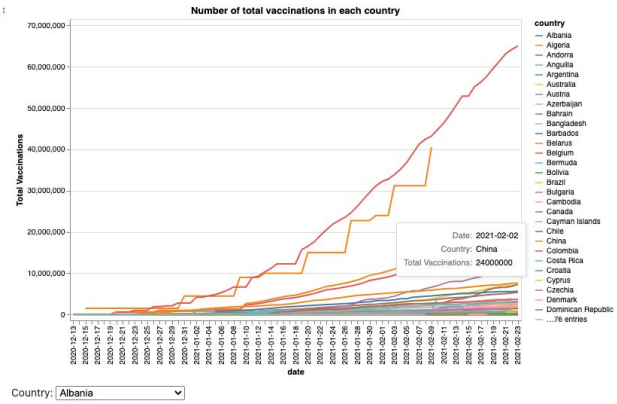

Fig 4. Global vaccination overview

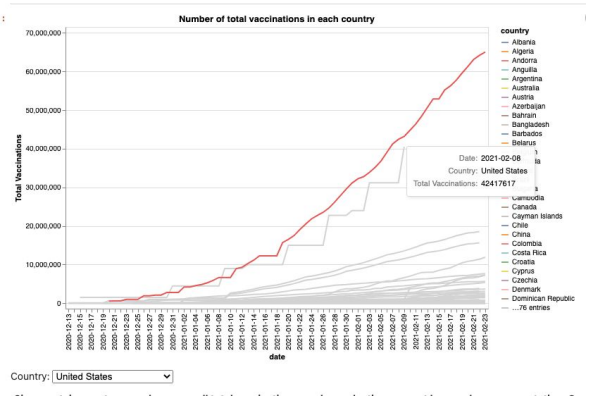

Fig 5. Vaccination overview selected US

For some countries that have very little vaccinations, selection may not reveal the best visualization. We also allow users to only plot the country that they care about.

```
plot.plot1_line_graph(vaccination,country = "United States"
```

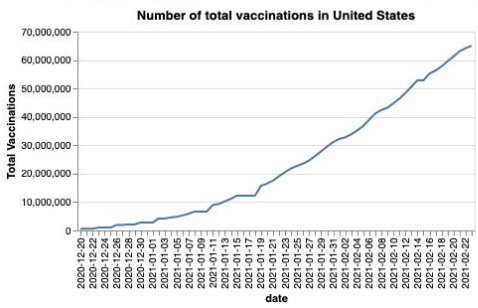

Fig 6. Vaccination overview only US

```
plot.plot1_line_graph(vaccination,country = "Zimbabwe"
```

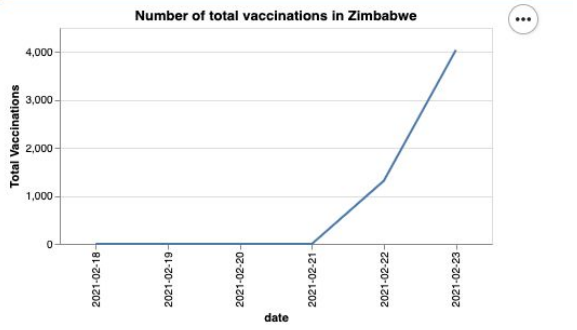

Fig 7. Vaccination overview only Zimbabwe

Top 10 Countries Analysis

After understanding the global vaccination process, we want to take a closer look at the countries with the top 10 number of vaccinations. We preprocess the data by grouping the vaccination by country and select the maximum number of vaccination to get total vaccination for each country and then we select the top 10 countries.

```
vaccination_top10
```

	country	total_vaccinations
102	United States	1.571491e+09
21	China	6.662880e+08
101	United Kingdom	4.736484e+08
32	England	4.049339e+08
51	Israel	2.363890e+08
46	India	2.064034e+08
100	United Arab Emirates	1.611197e+08
38	Germany	1.301326e+08
15	Brazil	1.201132e+08
98	Turkey	1.142023e+08

Fig 8. Top 10 vaccination countries

To visualize this information, we choose a barchart and encode countries information as hue. We use the barchart because we think it allows users to make comparisons. We also sort the bars to allow easier comparisons.

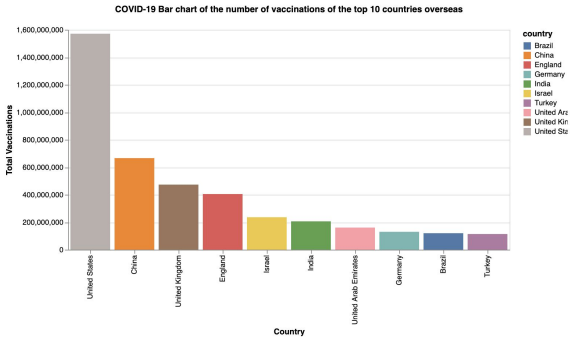

Fig 9. Top 10 countries vaccination barchart

We also want to draw the development trend about Covid19 vaccination, such as the top 10 countries in vaccination coverage and the number of vaccinations each day of the top 10 countries. Thus, we process our data by selecting only the countries from top 10 countries for our vaccination dataset.

We have two plots for visualizing this idea. The first plot is a trending line graph, with each point representing the number of total vaccinations for that day. We encode hue as country and position as the time.

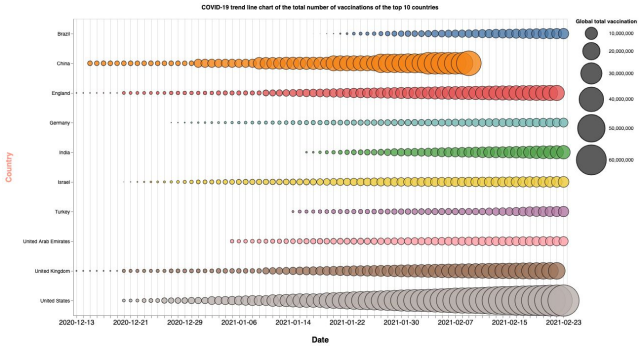

Fig 10. Top 10 countries trending line

We also visualize this using another technique, a stacked barchart. We encode the position as the number of vaccinations and hue as the time. In order to have a clear distinction, we use a color scheme for qualitative data as our encoding.

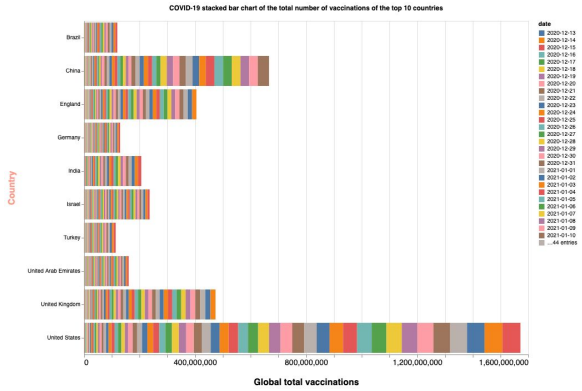

Fig 11. Top 10 countries stacked barchart

Country Vaccination Pace

We also want to understand the progress of vaccination for each country. We want to know how many people have been vaccinated for each country and compare them.

For this task, we choose choropleth for our visualization. We process the data by combining the country data and vaccination data in order to get country_id for geographic plot. However, we notice that countries that have more population tend to have more vaccination. To eliminate this factor, we choose to plot a choropleth for total vaccination per hundred and a choropleth for people fully vaccinated per hundred. We use a sequential color scheme for the encoding and we also allow tooltips for additional information.

```
plot.plot3_geo_per100(country_vac,"total_vaccinations_per_hundred")
```

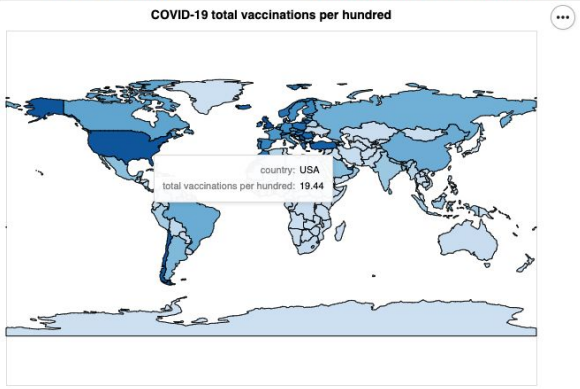

Fig 12. Choropleth for total vaccination per hundred

```
plot.plot3_geo_per100(country_vac,"people_fully_vaccinated_per_hundred")
```

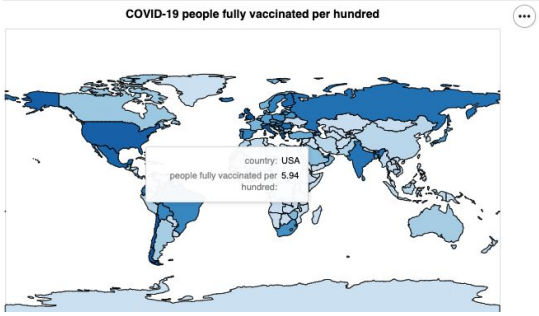

Fig 13. Choropleth for people fully vaccinated per hundred

Relationship Between COVID Cases and Vaccination

Last, we hypothesize that the countries that have more confirmed cases should care more about vaccinations since they are eager to stop the spreading of viruses.

For this task, we choose to plot a scatter plot to see whether there is an association between number of confirmed cases and number of vaccinations. We group the cases data by country and find the maximum confirmed cases, then we join the cases data with the vaccination data. For plotting our

graph, we plot the number of confirmed cases on the x axis, the number of vaccinations on the y axis, and we encode the number of fully vaccinated features by size. We also enable the interactive feature and the tooltips for more information.

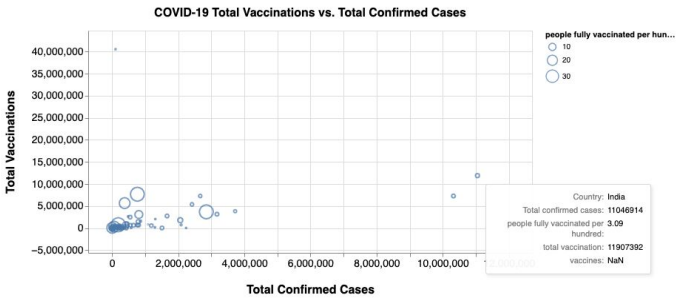

Fig 14. COVID-19 Vaccinations vs. Cases

By looking at this plot, we see a roughly positive correlation.

Conclusion

In our project, we visualize the COVID-19 vaccination and cases data. We found out most countries have started the vaccination process and the overall vaccination increased steadily, especially for the top 10 countries. We also realize that countries which have a fast pace of vaccination (larger number of people vaccinated per hundred) tend to be developed countries (North America, Europe, etc…). Moreover, we notice a slightly positive relationship between the number of confirmed COVID-19 cases and number of vaccinations, suggesting that countries which have a larger number of confirmed cases care more about vaccinations.