School of Information Date: October 27, 2021

SI 618 Data Manipulation and Analysis

Project 1 Report

Investigation on the Impact of COVID Vaccination from Different Aspects

Name	Junqi Chen
Uniqname	junqich
UMID	03846505
Course	SI 618
Instructor	Ceren Budak

1 Motivation

Due to the COVID-19 epidemic worldwide, a great change has taken place in our daily lives in the past two years. The vaccination is proved to be an efficient method to substantially reduce the risk of severe illness and death. In this project, I am interested in exploring how COVID-19 vaccination makes a difference to the world: did the distribution of vaccines lowers the COVID inflection rate, or ease the epidemic? We are going to look into the relationship between the vaccination rate and the influence of COVID through some indicators like the reproduction rate (it will be explained later in the data source part). Unfortunately, we have to admit that the spread of COVID is not solely affected by vaccination, which involves multiple factors including economic, geographical and cultural ones. As a result, the result may not be conclusive or deterministic like "more vaccination leads to less COVID infection". Instead, the result tends to be an exploration of the overall tendency on how vaccines change the world.

Research Questions:

- Is vaccination effectively prevent or ease the spread of COVID?
- Is vaccination rate affect the COVID policy of government?
- Is the COVID infection cases promote people's willingness to get vaccinated?

2 Data Sources

In this project, no API involves and I used two datasets from *Kaggle*:

- COVID-19 Dataset
- COVID-19 World Vaccination Progress

The detailed information about two datasets is shown below.

2.1 COVID-19 Dataset

Source: https://www.kaggle.com/deepshah16/covid19-dataset?select=covid.csv
This dataset is a collection of the COVID-19 data maintained by *Our World in Data*. It includes the number of confirmed COVID cases, hospital information and various indicators for different countries from 2020 to 2021. I adopt `covid.csv` from this dataset in this project (download on Oct.22, 2021). Detailed documentation of the important variables is listed:

- `location`: Country name
- `date`: Date of observation
- `total_cases`: Total confirmed cases of COVID-19
- `new_cases`: New confirmed cases of COVID-19
- `total_cases_per_million`: Total confirmed cases of COVID-19 per 1,000,000 people
- `new_cases_per_million`: New confirmed cases of COVID-19 per 1,000,000 people
- `reproduction_rate`: Real-time estimate of the effective reproduction rate (R) of COVID-19
- `stringency_index`: Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response)

2.2 COVID-19 World Vaccination Progress

Source: https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations.csv
This dataset is a record for country level vaccination data collected daily from *Our World in Data*. It includes the vaccination records for each country from January to October in 2021. I adopt `country_vaccinations.csv` from this dataset in this project (download on Oct.22, 2021).

- `country`: Country for which the vaccination information is provided
- `date`: Date for the data entry
- `daily_vaccinations`: Number of vaccination for that date/country
- `daily_vaccinations_per_million`: Ratio (in ppm) between vaccination number and total population for the current date in the country

Attention that here `daily_vaccinations_per_million` records only the times of vaccination. If one person get two doses vaccination, it will be counted two times. That is why for the later sum of `daily_vaccinations_per_million`, the total amount of some country many exceed 1 million (even reaching 2 millions).

3 Data Manipulation Methods

In this section, I will break the whole project into serveral steps, which is shown as the work flow diagram below:

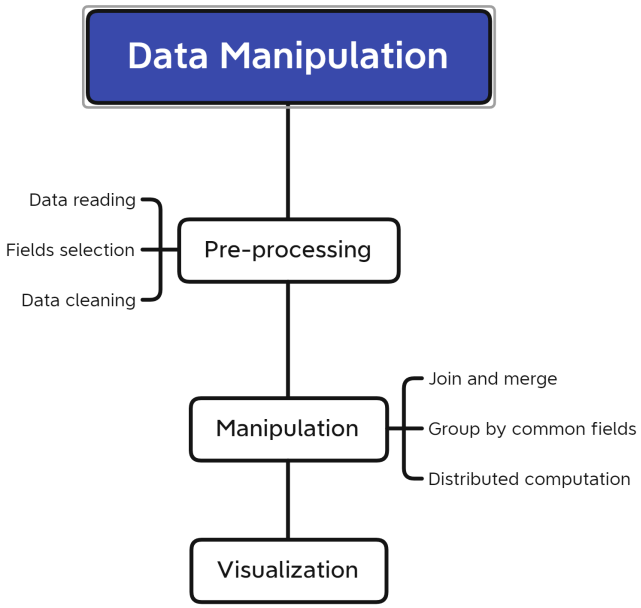


Figure 1: Data manipulation workflow diagram

, where *PySpark* is adopted to deal with the first two steps for a faster computation while the *Python* in *jupyter notebook* is used to plot the diagrams for the result since it is convenient for debugging.

3.1 Pre-processing

In this step, I first upload two source datasets `covid.csv` and `country_vaccinations.csv` to hadoop file system and then read them as dataframes in *PySpark*. After that, I select the necessary fields indicated as important variables and build the target dataframes called `covid` and `vacc`.

- Variables in `covid`: `location`, `date`, `reproduction_rate`, `stringency_index`, etc.
- Variables in `vacc`: `country`, `date`, `daily_vaccinations_per_million`, etc.

To deal with the missing data, I simply drop the line with null values (for only columns being used like `reproduction_rate`, `stringency_index`). It makes little difference to the analysis and the result since the later operations (for example, groupby) cover a long period of time like a month or a year. There are also less then 1% of null values in the selected field. As a result, dropping the null values is an acceptable method to deal with the missing values.
For the convenience of later manipulation, I extract the `year`, `month` and `day` from the `date` (string) variable for each dataframe, storing them as additional fields in an easy-groupby manner (int for `year` and `month`, string for `day`).
The schema for both target dataframes looks like:

```
>covid.printSchema()
>root
> |-- location: string (nullable = true)
> |-- iso_code: string (nullable = true)
> |-- date: timestamp (nullable = true)
> |-- total_cases: string (nullable = true)
> |-- new_cases: string (nullable = true)
> |-- total_cases_per_million: string (nullable = true)
> |-- new_cases_per_million: string (nullable = true)
> |-- reproduction_rate: string (nullable = true)
> |-- stringency_index: string (nullable = true)
> |-- year: integer (nullable = true)
> |-- month: integer (nullable = true)
> |-- day: string (nullable = true)

>vacc.printSchema()
>root
> |-- country: string (nullable = true)
> |-- iso_code: string (nullable = true)
> |-- date: timestamp (nullable = true)
> |-- daily_vaccinations: string (nullable = true)
> |-- daily_vaccinations_per_million: string (nullable = true)
> |-- year: integer (nullable = true)
> |-- month: integer (nullable = true)
> |-- day: string (nullable = true)
```

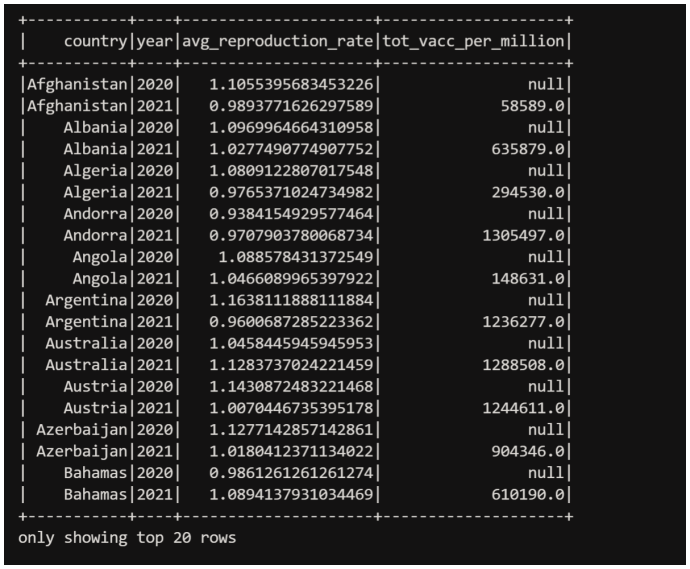
The code for this part is indicated in the "Data Pre-processing" part in the sources file `si618_project1_junqich.py`.

3.2 Manipulation

In this part, I adopt *SparkSQL* to merge the two target dataframe mainly by `Group By` function. I will discuss how and why I merge two target dataframes in detail.

- **Question 1: Vaccination rate and reproduction rate**:
  To investigate how vaccination make influence on the spread of the COVID, we first look at the total confirmed cases number. However, since we can hardly figure out the criterion to identify the relationship between two rapidly changing variables (`daily_vaccinations_per_million` and `total_cases`), we turn to look at the the efficient indicators.

  The indicator we adopt in this question is `reproduction_rate`, which refers to the Effective reproduction rate, which is an powerful indicator to measure the ability of viral transmission. I merge two target dataframes by the `country` and `year`, grouping them in an one-year scale. For each country, I count the total number of vaccination rate (ratio, in ppm) for 2021 and the average of reproduction rate for 2020 and 2021 to investigate whether the distribution of vaccines did efficiently restrain the spread of COVID.

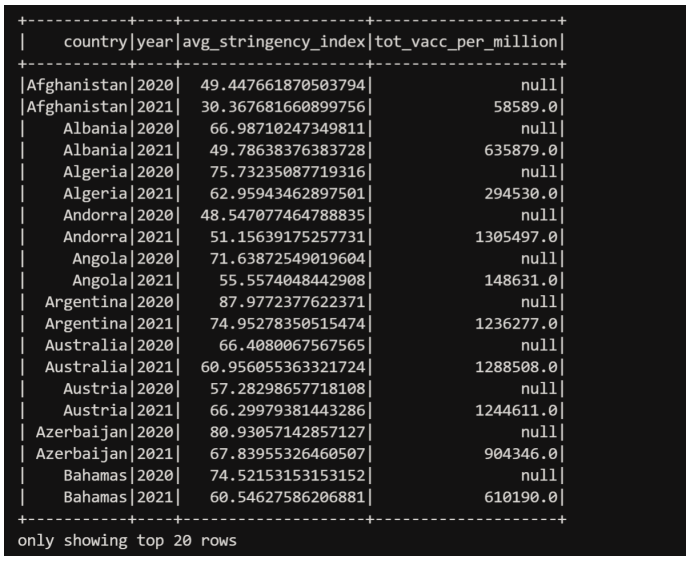


country	year	avg_reproduction_rate	tot_vacc_per_million
Afghanistan	2020	1.1055395683453226	null
Afghanistan	2021	0.9893771626297589	58589.0
Albania	2020	1.0969964664310958	null
Albania	2021	1.0277490774907752	635879.0
Algeria	2020	1.0809122807017548	null
Algeria	2021	0.9765371024734982	294530.0
Andorra	2020	0.9384154929577464	null
Andorra	2021	0.9707903780068734	1305497.0
Angola	2020	1.088578431372549	null
Angola	2021	1.0466089965397922	148631.0
Argentina	2020	1.1638111888111884	null
Argentina	2021	0.9600687285223362	1236277.0
Australia	2020	1.0458445945945953	null
Australia	2021	1.1283737024221459	1288508.0
Austria	2020	1.1430872483221468	null
Austria	2021	1.0070446735395178	1244611.0
Azerbaijan	2020	1.127714285714286	null
Azerbaijan	2021	1.0180412371134022	904346.0
Bahamas	2020	0.9861261261261274	null
Bahamas	2021	1.0894137931034469	610190.0

only showing top 20 rows

Figure 2: Result table after merging for the vaccination rate and reproduction rate

- **Question 2: Vaccination rate and stringency index**:
  The investigation for how vaccination make influence on the COVID policy of government works almost the same. The indicator adopted is `stringency_index`, which is an composite measurement based on the response policies to COVID from the government. I first join two target dataframe `covid` and `vacc` by `country` and `year`. Then I group by the same fields while counting the average of `stringency_index` in a year and the total number of vaccination rate in corresponding timestamps.



country	year	avg_stringency_index	tot_vacc_per_million
Afghanistan	2020	49.447661870503794	null
Afghanistan	2021	30.367681660899756	58589.0
Albania	2020	66.98710247349811	null
Albania	2021	49.78638376383728	635879.0
Algeria	2020	75.73235087719316	null
Algeria	2021	62.95943462897501	294530.0
Andorra	2020	48.547077464788835	null
Andorra	2021	51.15639175257731	1305497.0
Angola	2020	71.63872549019604	null
Angola	2021	55.5574048442908	148631.0
Argentina	2020	87.9772377622371	null
Argentina	2021	74.95278350515474	1236277.0
Australia	2020	66.4080067567565	null
Australia	2021	60.956055363321724	1288508.0
Austria	2020	57.28298657718108	null
Austria	2021	66.29979381443286	1244611.0
Azerbaijan	2020	80.93057142857127	null
Azerbaijan	2021	67.83955326460507	904346.0
Bahamas	2020	74.52153153153152	null
Bahamas	2021	60.54627586206881	610190.0

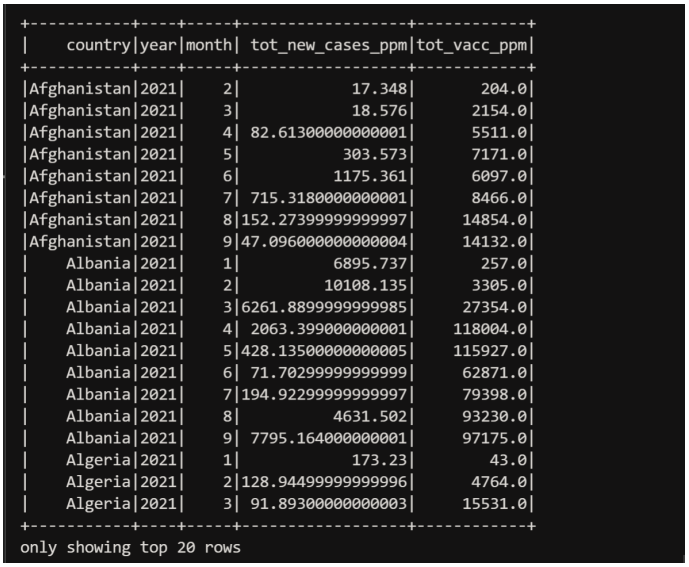
only showing top 20 rows

Figure 3: Result table after merging for the vaccination rate and stringency index

The reason for null values in the last column is that vaccination recorded only in the year of 2021. I am going to compare the results for 2020 and 2021 to see how vaccination make a difference on these indicators.

- **Question 3: New cases rate and vaccination rate**:
  To see if the severity of COVID affect people's willingness to get vaccinated, I look into the relationship between the number of daily new cases and vaccination rate (both in ratio), i.e. `new_cases_per_million` and `daily_vaccinations_per_million`. In this part, I group them by

`month` additionally to the `country` and `year` to detect minor changes in a month for each country. The summation for new COVID cases (in ppm) is calculated for each month in each country in comparison to the total vaccination rate (in ppm) in the same condition.



country	year	month	tot_new_cases_ppm	tot_vacc_ppm
Afghanistan	2021	2	17.348	204.0
Afghanistan	2021	3	18.576	2154.0
Afghanistan	2021	4	82.61300000000001	5511.0
Afghanistan	2021	5	303.573	7171.0
Afghanistan	2021	6	1175.361	6097.0
Afghanistan	2021	7	715.3180000000001	8466.0
Afghanistan	2021	8	152.27399999999997	14854.0
Afghanistan	2021	9	47.096000000000004	14132.0
Albania	2021	1	6895.737	257.0
Albania	2021	2	10108.135	3305.0
Albania	2021	3	6261.8899999999985	27354.0
Albania	2021	4	2063.399000000001	118004.0
Albania	2021	5	428.13500000000005	115927.0
Albania	2021	6	71.70299999999999	62871.0
Albania	2021	7	194.92299999999997	79398.0
Albania	2021	8	4631.502	93230.0
Albania	2021	9	7795.164000000001	97175.0
Algeria	2021	1	173.23	43.0
Algeria	2021	2	128.94499999999996	4764.0
Algeria	2021	3	91.89300000000003	15531.0

only showing top 20 rows

Figure 4: Result table after merging for the new cases rate and vaccination rate

`ATTENTION`:
In data manipulation part, I mainly adopt the daily data for vaccination rate as well as new cases rate instead of using pre-calculated total number of cases is to avoid the conflicting in between. Also, a large number of missing data in the column of total cases (like `total_cases_per_million`) is hard to deal with. So I adopt the daily data to compute the summation and average on my own. It appears to be more reliable and accurate.
Besides, variables except indicators in this part are measured in ratio, which means that they can be a common measurement for different countries regardless of their population.

3.3 Visualization

After computing the result with *MySQL*, I save the results in each part to several csv files for later visualization and analysis. Please refer to next section for more details.

4 Analysis and Visualization

In this section, I adopt *Seaborn* from *Python* to plot different figures for visualization. Based on the results data and visualization, I make practical analysis to discuss how vaccination make influence on the COVID from different aspects.

4.1 Vaccination and Spread of COVID

This part mainly answer the first question: *Is vaccination effectively prevent or ease the spread of COVID.* I investigate the relationship between the **vaccination rate** and the spread of COVID, indicated by **reproduction rate**. Here are the detail explanation on the variables:

- `vaccination_rate` is counted by the following formula:

$$\text{Vaccination rate (for a country)} = \frac{\text{Sum(Daily vaccination count)}}{\text{Population}} = \text{Sum(daily vaccination rate)}$$

, where the count of daily vaccination can be counted repeated for one person since there are second dose even the third dose. In simply word, the vaccination rate represents the coverage of vaccination in a country.

- `reproduction_rate` refers to the effective reproduction rate (aka effective reproduction number, R_t), which is an indicator for the spreading ability of a virus taking every anti-epidemic measures into account. Practically, an `reproduction_rate` less than 1 means that COVID is effectively controlled while an `reproduction_rate` larger than 1 means that the virus is spreading. I use the `reproduction_rate` as a reliable reference for the spread of virus.

I look into the first question from two different aspects:

- **Is the distribution of vaccination in 2021 play a part in preventing the spread of COVID comparing to 2020?**

  Since the distribution of the vaccine started approximately from the early 2021, we plot the average of reproduction rate for each country in two years, 2020 and 2021, to get overview on how the spread of COVID change between years. According to Figure 5, we can find out that for most countries, the reproduction rate in 2021 is lower than the one in 2020, indicating that the spread of the COVID has been effectively suppressed comparing to the previous year.

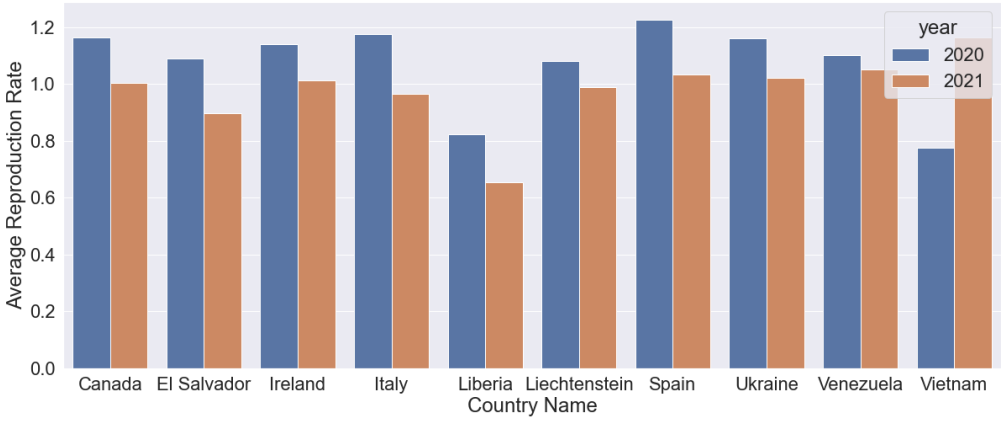


Figure 5: Average reproduction rates of 10 random countries in 2020 and 2021

However, it does not necessary mean that vaccination do good to the epidemic prevention. We can also see that there are also countries like Vietnam suffer from a more severe epidemic in 2021. As a result, we still need more evidence to investigate the relationship between vaccination and the spread of COVID.

- **Is a higher vaccination rate relates to a decrease on the COVID spread?**

  To make further investigation on how vaccination relate to the spread of the COVID (here using reproduction rate), we plot the scatter plot of the vaccination rate (in percent) verse reproduction rate for each country.

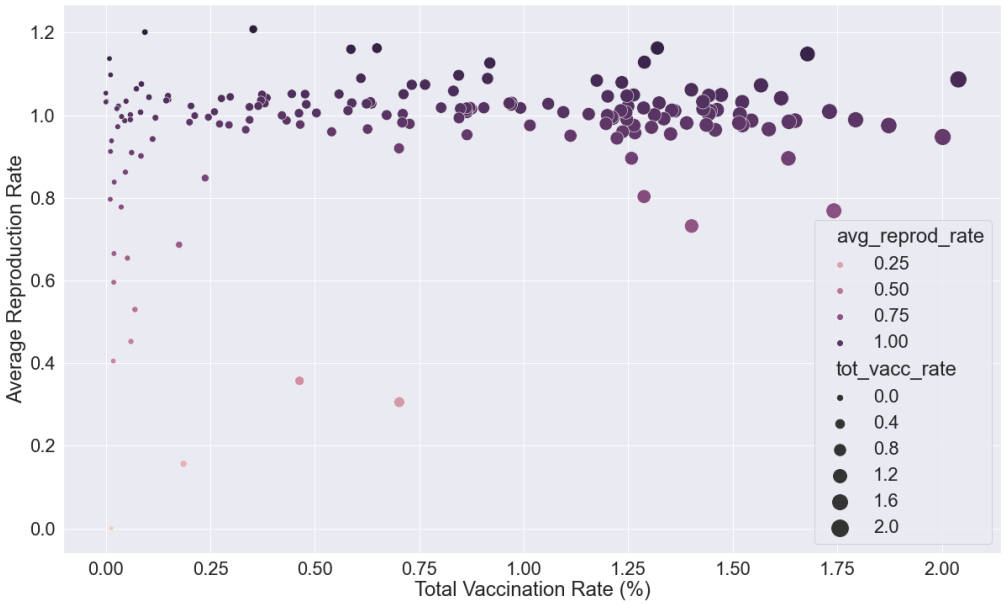


Figure 6: Average reproduction rates v.s. Total Vaccination Rate for all countries in 2021

According to Figure 6, we can hardly find out the correlation between the total vaccination rate and the reproduction rate since points are distributed horizontally. Even some countries with little vaccination rate can still maintain a lower reproduction rate, indicating COVID can hardly spread in these countries. But for most countries, the COVID has been successfully controlled since they have an average reproduction rate lower than 1.

Conclusively, we can **hardly figure out any direct connection between the vaccination rate and reproduction rate** since the spread of COVID is not solely depend on a single factor like vaccination. The investigation above at least tells us vaccination may affect and prevent the epidemic to some extent.

4.2 Vaccination and Policy

Then I am going to look into the second question: *Is vaccination rate affect the COVID policy of government.* To figure out the relationship between the vaccination and the government's policy, two variables vaccination rate and the stringency index are examined in this part. The explanation for the varibales:

- `stringency_index` is a composite indicator to measure the government's COVID policies based on 9 response indicators including school closures, workplace closures, and travel bans. It ranges from 0 to 100, where 0 means the most relaxing policy while the 100 means the strictest response.

Similarly, I also look into the first question from two different aspects:

- **Is the distribution of vaccination in 2021 make an influence on the COVID policies by government comparing to 2020?**

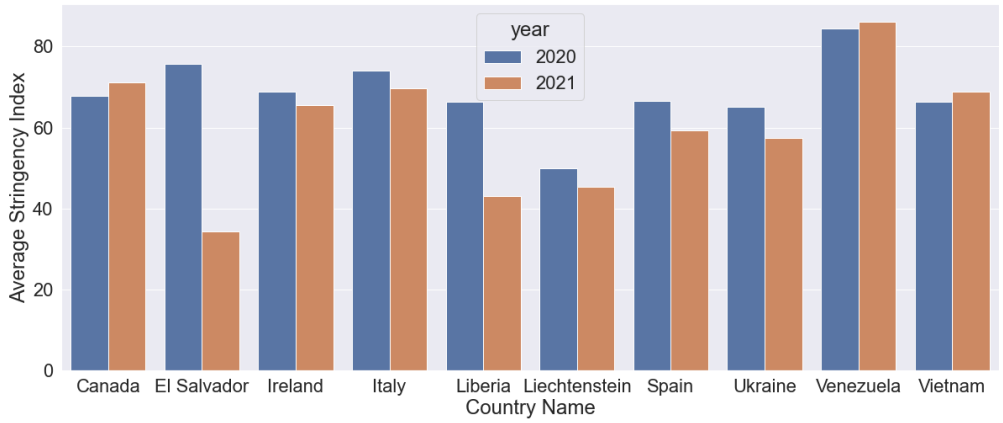


Figure 7: Average stringency index of 10 random countries in 2020 and 2021

The figure for the average stringency index is plotted follow the same construction as the previous part for reproduction rate. According to Figure 7, we can figure out that a more relaxed policy against COVID has been adopted for most countries in 2021. It can also be resulted from the fact that epidemic has been well controlled in 2021.

- **Is a higher vaccination rate related to relaxation on the COVID policies in a country?**

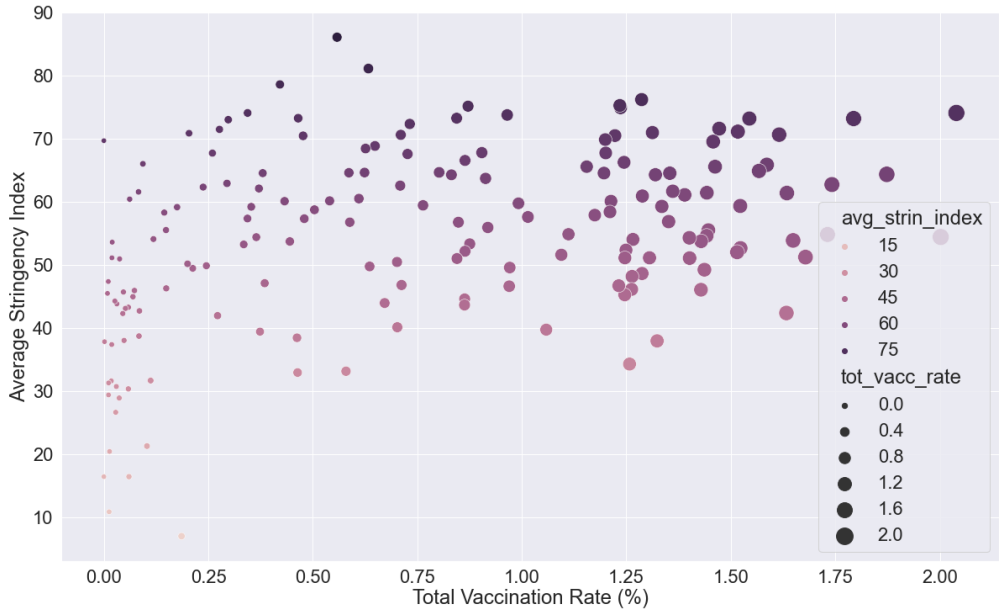


Figure 8: Average Stringency Index v.s. Total Vaccination Rate for all countries in 2021

From the figure above, we can also figure out that the vaccination rate has no obvious correlation to the stringency index, indicating that **a higher vaccine coverage does not necessary mean a more relaxed policy for a country**. From my point of view, the policies may largely depend

on the dynamic change on the epidemic. Governments determine their policies against COVID according to the current condition involving different considerations.

4.3 COVID cases and Vaccination

For the third question: *Is the COVID infection cases promote people's willingness to get vaccinated*, our intuition may be that people actively get vaccinated once seeing such a great epidemic. I am going to answer this question in this part. The variables I adopt are daily new cases and daily vaccination (both in ratio). The explanations are as follow:

- `new_cases_per_million` refers to the new cases rate, which is a daily measurement for the newly confirmed COVID cases for each day. It can be counted by the formula:

$$\text{New cases rate (ppm, for a country)} = \frac{\text{Sum(Daily confirmed COVID cases)}}{\text{Population}} = \text{Sum(daily cases rate)}$$

- `daily_vaccinations_per_million` is the same as previous one but here is counted in per million (ppm). Note that both variables are counted in ratio for the unification for all countries.

Similarly, two sub-questions are put forward to better investigate the main question:

- **Is a higher COVID inflection rate results in more people to get vaccinated in a month?**

  To find out how COVID cases make influence on public willingness to get vaccinated. I randomly choose 6 countries and visualize the change on the total COVID cases and total vaccination rate monthly. From Figure 9, we may think that there is no clear relationship in between for the first glance. But we can find that two variables follows a similar trend even through there may be a lag in between. For example, in Vietnam, the total COVID cases increase from month 6 to 8 and reaching the peak at 9 while the total vaccination rate begins to rise at month 7 (lag approximately 1 month) and keep increasing. Also, two dependent variables in Spain and Suriname appear to have peak at the same timestamps. But it still lack of evidence to verify that that are related.

- **Is a higher COVID inflection rate results in more people to get vaccinated for each country within a year?**

Figure 10: Average reproduction rates of 10 random countries in 2020 and 2021

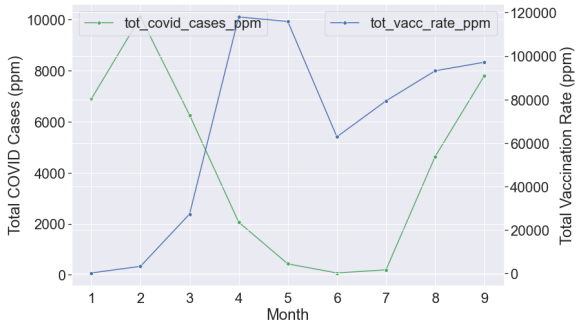
Then we are going to look at the overall tendency between the average vaccination rate and total cases rate. The scatter plot for the average vaccination rate and total cases rate in 2021 is shown. From Figure 9, there is a positive correlation between two variables. Typically, the country of the rightmost point with the highest total COVID cases does have a higher vaccination coverage (around 155%). We can figure out that for each country, higher total cases rate relates to a higher vaccination rate (both in ratio).

In conclusion, the result plots indicates that **the more severe the epidemic in a country, the more people want to get vaccinated to protect themselves**.

5 Challenges

In this project, I did encounter various challenges and difficulties, which did stuck me a lot. Some of the typical problems are discussed below with my ways out:

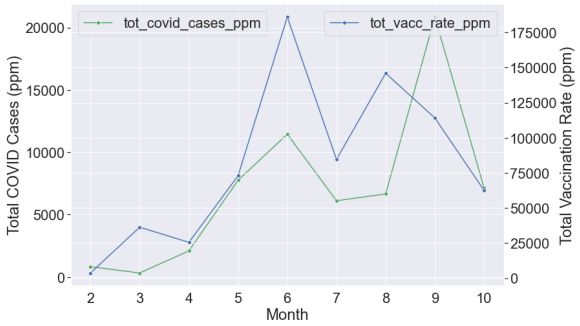
- When I first started to process the source data, I had a hard time to deal with tons of missing data in different columns like `new_cases` and `total_cases` since they were recorded about once a week in the early 2021. My solution out was to adopt the reliable columns like daily COVID cases rate `new_cases_per_million`, which has less missing rows so that I can handle easily, and compute the secondary data like total COVID cases rate on my own. The lesson I learnt from this is that there may be all kinds of exceptions beyond simple missing date in real-world dataset, we have to decide what is needed and what is not.
- Another biggest problem besides the missing data is the decision on the target variables. Since COVID cases are affected by various factors as I discussed in the paper, I can hardly draw a deterministic conclusion simply between two variables. Investigating multi-relationship among 3 or more variables is also impractical. That is why I send the email to the professor and state this problem. Finally, I just investigate on vaccination and it influence regardless of whether a conclusion can be drawn: a non-conclusive result is also acceptable, though.



(a) Albania

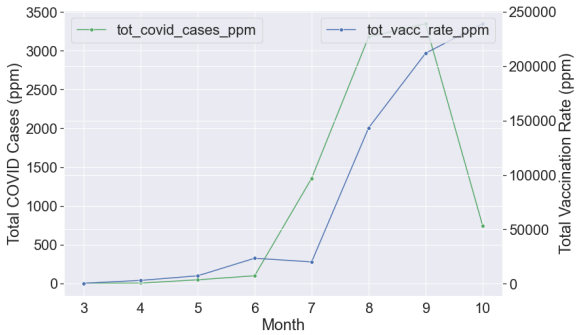
(b) Costa Rica

(c) Spain



(d) Suriname

(e) Turkey



(f) Vietnam

Figure 9: Line plots for the Total COVID cases and Tatal vaccination rate v.s. month for 6 random countries in 2021