SI618 Project 2

Name: Junqi Chen

Uniqname: junqich

UMID: 03846505

Date: 2021/12/7

I. Motivation

Due to the COVID-19 epidemic worldwide, a great change has taken place in our daily lives in the past two years. I want to investigate how COVID make an influence on the activities of people viewing from the country level. Take a step further, I want to look into how people's behaviors and reactions affect the spread of COVID in reverse.

Research Questions:

- How the condition of a country make influence on the spread of COVID?
- How COVID affect human behavior in the country level?
- How human behaviors make influence on the COVID over time?

II. Data Sources

Basic Info

The dataset I use in this project is **Covid-19 Dataset**, which 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. There are two version for this dataset:

- Version 2 is a previous version the same as I used in project 1. A record of COVID from 2019-12-31 to 2021-10-19.
- Version 3 is an newly updated version. It emitted the past records but only keep the current record for all countries in 2021-11-05. It can be a great start without involving the time series data.

The main file I will focus on is the `covid.csv`, which records the COVID cases in a time series with various useful indicators like reproduction rate and stringency Index. I will possibly use the `vaccinations.csv` if needed.

Documentation

Source: https://www.kaggle.com/deepshah16/covid19-dataset?select=covid.csv

The documentation of the some important variables is listed (for more information, please refer to the source data):

COVID Indicators

Variable	Description
`total_cases`	Total confirmed cases of COVID-19
`new_cases`	New confirmed cases of COVID-19
`new_cases_smoothed`	New confirmed cases of COVID-19 (7-day smoothed)
`total_cases_per_million`	Total confirmed cases of COVID-19 per 1,000,000 people
`reproduction_rate`	Real-time estimate of the effective reproduction rate (R) of COVID-19. See https://github.com/crondonm/TrackingR/tree/main/Estimates-Database

Country & Human Indicators

Variable	Description
`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)
`gdp_per_capita`	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available
`handwashing_facilities`	Share of the population with basic handwashing facilities on premises, most recent year available
`life_expectancy`	Life expectancy at birth in 2019
`human_development_index`	A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506
`total_vaccinations_per_hundred`	Total number of COVID-19 vaccination doses administered per 100 people in the total population
`people_vaccinated_per_hundred`	Total number of people who received at least one vaccine dose per 100 people in the total population

III. Methods

Part 1: Country conditions & COVID

- **Data manipulation**: First, I select the targeted variables (including `iso_code`, `gdp_per_capita`, `total_cases_per_million`) to make a new dataframe. With the function `pd.cut()`, I make new columns called `gdp_level` and `covid_level` for the different level of GDP and COVID cases.
- **Data cleaning**: I simply drop the rows with the missing data. Since the dropping rows are around 10% of all rows, it does not make a difference to the final result.
- **Challenge**: For the function `pd.cut()`, we have to decide the number of the bins. I originally consider 3 bins for each level but finally switch to 2 bins for COVID cases since there will be no country in the {high_gdp, high_covid} set.

Part 2: COVID & Human reactions

- **Data manipulation**: Similarly, I select the targeted variables (including `iso_code`, `reproduction_rate`, `reproduction_rate`) to make a different new dataframe. With the function `pd.cut()`, I make new columns called `rr_level` and `si_level` for the different level of reproduction rate and stringency index.
- **Data cleaning**: I simply drop the rows with the missing data.
- **Challenge**: For the Mosaic plot for the 3 bins size, it does not have a clear relationship between the reproduction rate and stringency index as I expect. However, with the chi-square test, I finally make a conclusion on the these uncorrelated variables.

Part 3: Human reaction & COVID over time

- **Data manipulation**: In this question, I adopt the Version 2 of Covid-19 dataset to take the advantage of the time series data. I first extract the needed variables (including `date`, `"new_cases_per_million"` and `new_vaccinations_smoothed_per_million`) for a new dataframe. Then I perform the conversion from the date time to `year`, `month` and `day`, which is used for later `groupby()` operation. With the data retrieved, we can easily plot a time series graph on both new cases and new vaccinations.
- **Data cleaning**: Similarly, I simply drop the rows with the missing data in the independent variables `new_cases` and `new_vacc`.
- **Challenge**: For the time series data in this question, I originally consider plotting the data in the daily level. However, the large number of data results in a messy plot in the end. So I decide to plot the data in a monthly level, where the change of the variables can be clearly depicted. Besides, it takes me some times to put two variables in the same plot.

IV. Analysis and Results

Part 1: Country conditions & COVID

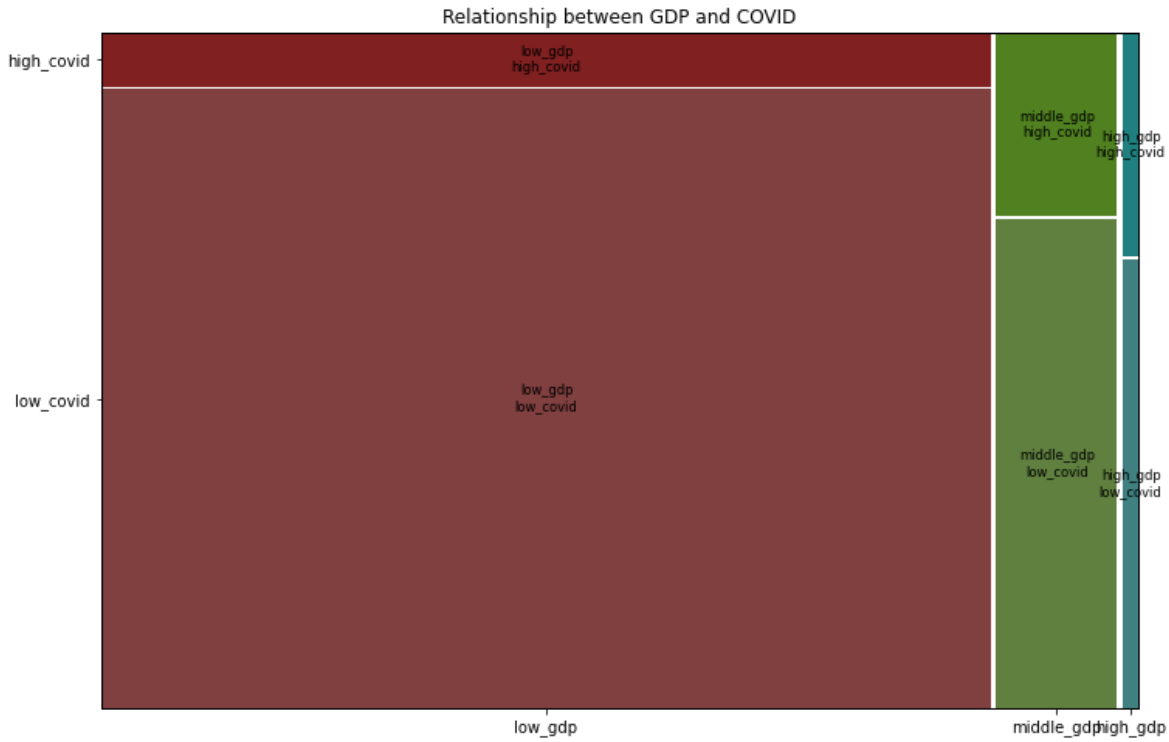
In this question, I want to investigate how the condition of a country (indicated by some indicators like GDP) make influence on the spread of COVID (indicated by total confirmed cases).

Q1-1: The relationship between the GDP and COVID cases

The first subquestion is to explore on whether the GDP of a country `gdp_per_capita` relates to the total count of COVID cases `total_cases_per_million`. Intuitively, we may consider that countries with higher GDP will have better social welfare and healthcare resources due to the highly-developed economic. With the Python, we then explore whether our assumption is ture.

- **Mosaic Plot**:

  To figure out the relationship between these two variables, I first cut the COVID indicator `total_cases_per_million` into 2 equal sized bins and GDP `total_cases_per_million` into 3 bins with clear labels. Then, the masaic plot is shown below:



  According to the masaic plot, we can find out that there is an obvious decrease on the COVID cases as the GDP level increase, indicating that countries with higher GDP are likely to fewer COVID cases. However, most countries are classified into the `low_gdp` bins since I use the default bin with equal length. It may lead to some bias for the different bin size (different numbers of countries in different bins). Hence the conclusion still need further inspection.

- **Chi-square Test**:

  To make further examination on the previous conclusion, we consider the chi-square test as an applicable statistic method in this situation. With the chi-square test, we get the following statistics:

```
chi2 =  9.149203444961929
p-val =  0.010310404762794585
degree of freedom =  2
```

  Since the p-value of the test is smaller than the pre-defined $\alpha = 0.05$, hence we have no enough evidence to reject the null hypothesis statistically. It indicates that a higher GDP relate to a lower total COVID cases for a country.

Q1-2: OLS between COVID cases and country indicators

To further investigate how the condition and situation of a country make influence on the COVID cases, I adapt the OLS method to fit several country indicators. The model I used is:

$$y = \beta_1 \mathrm{CI}_1 + \beta_2 \mathrm{CI}_2 + \beta_3 \mathrm{CI}_3 + c$$

, where y is a COVID indicator (`total_cases_per_million` or `reproduction_rate`) and CI_i are selected country indicators (`gdp_per_capita` , `handwashing_facilities` and `hospital_beds_per_thousand`). The specific models are shown in the sample code:

```
model1 = smf.ols('total_cases_per_million ~ gdp_per_capita + handwashing_facilities +
hospital_beds_per_thousand', data=df1).fit()
model2 = smf.ols('reproduction_rate ~ gdp_per_capita + handwashing_facilities +
hospital_beds_per_thousand', data=df1).fit()
```

i.e. we have two models like:

$$\text{model1: total\_cases\_per\_million} = \beta_1 \text{gdp\_per\_capita} + \beta_2 \text{handwashing\_facilities} + \beta_3 \text{hospital\_beds\_per\_thousand} + c$$

$$\text{model2: reproduction\_rate} = \beta_1 \text{gdp\_per\_capita} + \beta_2 \text{handwashing\_facilities} + \beta_3 \text{hospital\_beds\_per\_thousand} + c$$

The fitting results are shown below:

> **Model1: total_cases_per_million ~ gdp_per_capita + handwashing_facilities + hospital_beds_per_thousand**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     total_cases_per_million   R-squared:                       0.477
Model:                                 OLS   Adj. R-squared:                  0.456
Method:                      Least Squares   F-statistic:                     22.22
Date:                     Tue, 07 Dec 2021   Prob (F-statistic):           2.51e-10
Time:                             15:25:57   Log-Likelihood:                -887.59
No. Observations:                       77   AIC:                             1783.
Df Residuals:                           73   BIC:                             1793.
Df Model:                                3
Covariance Type:                 nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                   -1.138e+04   5997.132     -1.897      0.062   -2.33e+04     575.950
gdp_per_capita                  0.9827      0.527      1.865      0.066      -0.068       2.033
handwashing_facilities         335.4258    122.277      2.743      0.008      91.728     579.123
hospital_beds_per_thousand    7293.2487   2014.612      3.620      0.001    3278.133    1.13e+04
==============================================================================
Omnibus:                       20.067   Durbin-Watson:                   2.204
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               31.066
Skew:                           1.031   Prob(JB):                     1.80e-07
Kurtosis:                       5.330   Cond. No.                     2.18e+04
==============================================================================
```

> **Model2: reproduction_rate ~ gdp_per_capita + handwashing_facilities + hospital_beds_per_thousand**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      reproduction_rate   R-squared:                       0.116
Model:                            OLS   Adj. R-squared:                  0.079
Method:                 Least Squares   F-statistic:                     3.115
Date:                Tue, 07 Dec 2021   Prob (F-statistic):             0.0315
Time:                        15:25:57   Log-Likelihood:                -1.3847
No. Observations:                  75   AIC:                             10.77
Df Residuals:                      71   BIC:                             20.04
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                      0.7016      0.061     11.556      0.000       0.581       0.823
gdp_per_capita              1.129e-05   5.38e-06      2.098      0.039    5.61e-07     2.2e-05
handwashing_facilities         0.0008      0.001      0.628      0.532      -0.002       0.003
hospital_beds_per_thousand    -0.0121      0.020     -0.597      0.552      -0.053       0.028
==============================================================================
Omnibus:                       10.539   Durbin-Watson:                   1.655
Prob(Omnibus):                  0.005   Jarque-Bera (JB):               10.804
Skew:                          -0.768   Prob(JB):                      0.00451
Kurtosis:                       4.048   Cond. No.                     2.20e+04
==============================================================================
```

According to the result, we can find out that:

- For the first model whose dependent variable is COVID cases, the indicators `handwashing_facilities` and `hospital_beds_per_thousand` are significant since their p-value larger than 0.05. Hence we can conclude that the healthcare condition including the handwashing facilities and available beds in hospital make a great influence to the number of COVID cases.
- For the second model whose dependent variable is reproduction rate (an indicator for the spread of the COVID), only indicator `gdp_per_capita` is significant. Hence we can conclude that GDP makes influence on the spread of COVID in a country.

Part 2: COVID & Human reactions

In this part, I want to investigate how COVID (indicated by total confirmed cases and reproduction rate) affect human behavior (indicated by some indicators like stringency index and vaccination rate) in the country level.

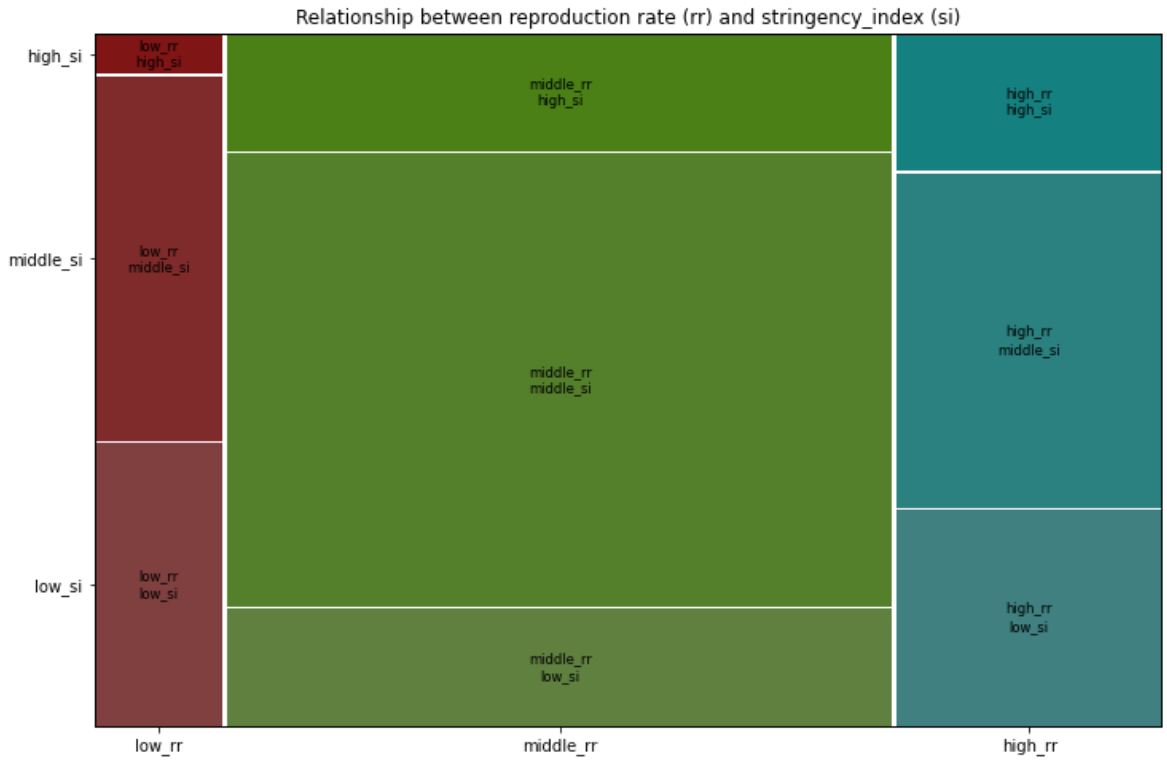
Q2-1: The relationship between the reproduction rate and stringency index

The only question I want to investigate in this part is that how reproduction rate relates to the stringency index. The reproduction rate is an indicator for the spread speed of COVID virus and the stringency index refers to the government's policies against the spread of COVID. We may think that countries with higher reproduction rate would have a more strict policy.

To exam the assumption, we perform the similar operation as the Q1-1:

- **Mosaic Plot**:

  To figure out the relationship between two variables, I first cut the reproduction rate `reproduction_rate` into 3 equal sized bins and indicator for the country policies `stringency_index` into 3 bins with clear labels. Then, the masaic plot is shown below:



  According to the masaic plot, we can find out that there is not obvious tendency between the reproduction rate and stringency index. We can figure out that most countries are of middle-level on both reproduction rate and stringency index since this sell occupies the most area. Hence we need further examination to check whether they are uncorrelated.

- **Chi-square Test**:

  Similarly, I consider the chi-square test to perform further inspection. I perform the chi-square on the following cross tab:

rr_level/si_level	low_si	middle_si	high_si
low_rr	7	9	1
middle_rr	15	58	15
high_rr	11	17	7

The result of chi-square test is shown below:

```
chi2 =  7.6001770453060855
p-val =  0.10737217993872845
degree of freedom =  4
```

Since the p-value is smaller than the pre-defined $\alpha = 0.05$, we have enough evidence to reject the null hypothesis statistically. It indicates that there is not relationship between the reproduction rate and stringency index. Hence we can conclude that the spread of COVID does not necessarily relates to or determine the policies of government.

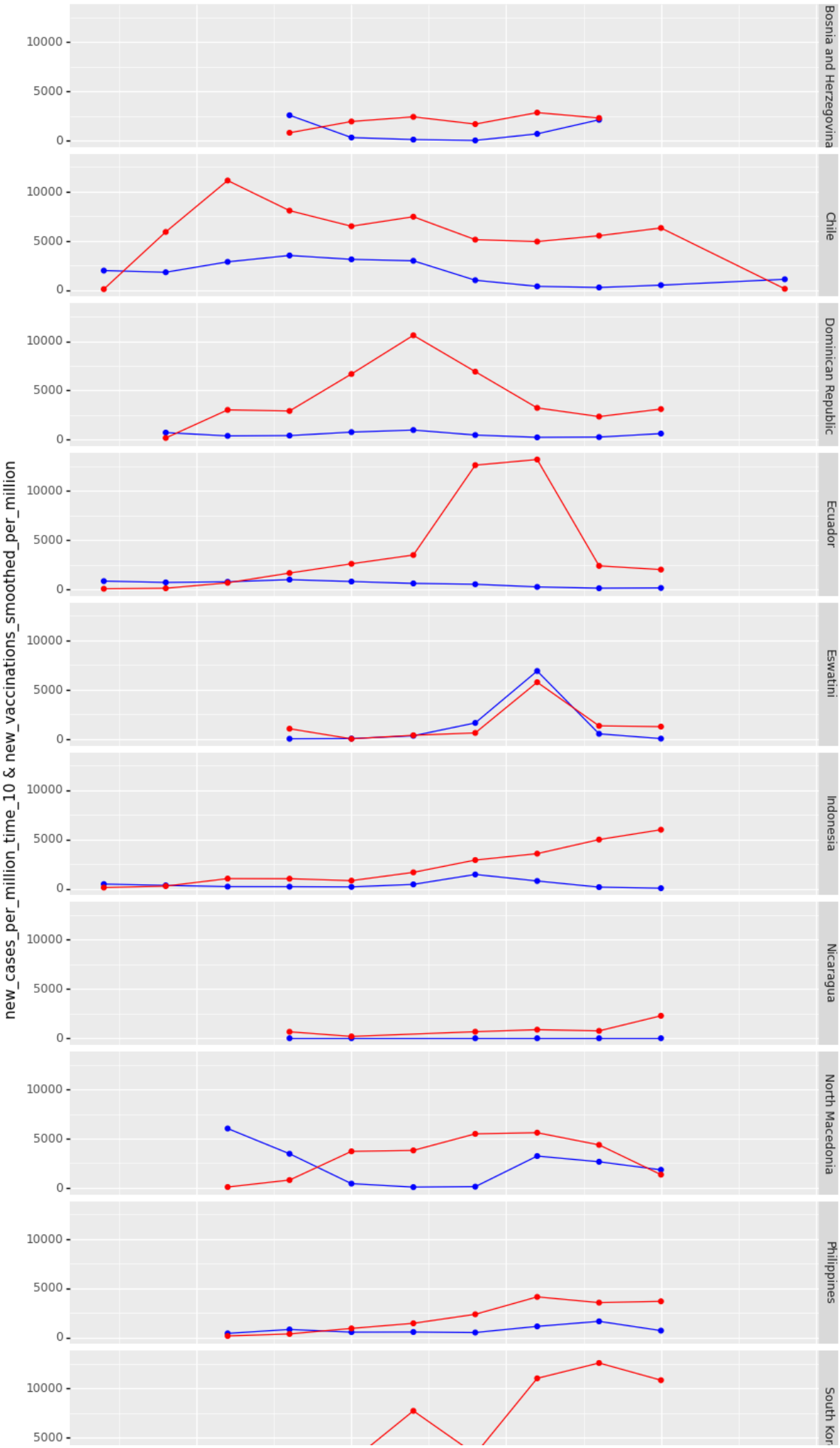
Part 3: Human reaction & COVID over time

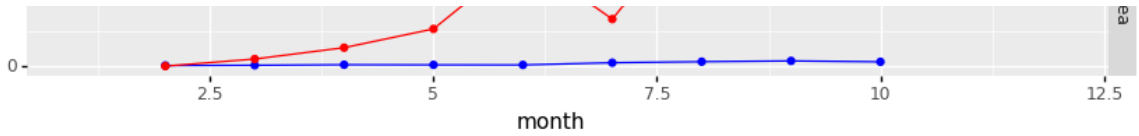
In this part, I want to investigate how human behaviors (indicated by vaccination) make influence on the COVID (indicated by newly confirmed cases and reproduction rate) over time how COVID in a country level.

Q3-1: The relationship between COVID cases and vaccination over time

Intuitively, we may consider that a serious COVID condition will prompt more people to get vaccinated. To figure out whether our assumption is true, i.e. the relationship between the COVID cases and vaccination over time, we plot the time series graph of two variables `new_cases_per_million` and `new_vaccinations_smoothed_per_millio` in a monthly level. I take the average values from both variables in a month as the representatives.

The line plot is shown below (from 10 random sampled countries):


new_cases_per_million_time_10 & new_vaccinations_smoothed_per_million
Bosnia and Herzegovina
Chile
Dominican Republic
Ecuador
Eswatini
Indonesia
Nicaragua
North Macedonia
Philippines
South Kor
10000
5000
0



, where the blue lines stand for the `new_cases_per_million` times 10 and the red lines stand for the `new_vaccinations_smoothed_per_million`.

According to the graph above, we can figure out that there exist some similarity between the change on new COVID cases and new vaccination counts. For example, we can see that both COVID cases and vaccinations reach a peak in August (month 8) in Eswatini. However, most of the curves have a irregular change over months, indicating that we cannot draw a conclusion that there are close relationship between the COVID cases and vaccinations.

Q3-2: The relationship between reproduction rate and vaccination over time

Due to the limitation of page, the analysis for the relationship between reproduction rate and vaccination over time is emitted in this report. Please refer to the source code for more detail information.