Data wrangling with Covid-19

Yelin Shin

5/4/2020

Github repository:

https://github.com/YelinShin/2020-Spring-data-wrangling

Introduction:

Around late December 2019, people got notice about a virus that was speared from China. After the virus got spread quickly, WHO named it as Covid-19 and decided as epidemic. However, at that time, people in non-Asian country underestimated how fast the virus can be spread to their countries. Therefore, most of the countries were not prepared and faced rapid increase of confirmed cases and even death cases from the virus. Also, it is relative to our class because of the virus, every classes hold lectures remotely. Therefore, in current situation, people wonder when the virus have a lull. I decided to show the changes by various data visualization.

Data set 1:

1. The main world time series dataset is from GitHub, "https://github.com/datasets/covid-19/blob/master/data/time-series-19-covid-combined.csv". Since the data contains detail information about the province/state in each country, I summarize each cases number by country and date. So, it will contain only one pair of (country, date).
2. The time series corona case file contains 2018 population of each countries. Therefore, I decided to change the number to 2019 population from Wikipedia, https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations) by web-scrapping the page. Since the population formatted with comma, I erased the comma then converted it to number type
3. Also, I grab the time-series update of the number of tested cases in each country by "https://github.com/owid/covid-19-data/blob/master/public/data/testing/covid-testing-all-observations.csv". This csv file contains too many information and the country name contains '- tests performed', I extract the country name only and the current update of testing number. Moreover, some countries have 2 resources to track the testing number.

4. So, I fix one resource per country. After I got the 3 cleaned data, I use join to get the finalized data that contains all the cases, test, and population. Moreover, I make a table for latest updated case number for each country. Since some countries did not update/share the testing number every day, I grab most recent number of testing into this table, and put active number by mutate. (Active = confirmed – deaths – recovered). When I joined the latest Covid and latest testing data, I based on latest Covid date since sometime the publisher update testing dataset faster than Covid dataset

5. Lastly, I get the world ranking by GDP from Wikipedia ("https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)"). Since there are too many countries in the data, it is better to show subset of countries. Therefore, it is good to show the economically developed countries for data visualization.

For all data source, I edit the countries' name to match with 'country.regions' to use chroplethrMaps.

Raw table looks like . . .

Date	Country/Region	Province/State	Lat	Long	Confirmed	Recovered	Deaths
2020-02-16	Kuwait	NA	29.5000	47.7500	0	0	0
2020-04-26	Bahamas	NA	25.0343	-77.3963	80	22	11

Note: [a] Random 2 rows in world time-series covid 19

Entity	Date	Source URL	Source label	Notes	Cumulative total	Daily change in cumulative total	Cumulative total per thousand	Daily change in cumulative total per thousand	3-day rolling mean daily change	3-day rolling mean daily change per thousand	7-day rolling mean daily change	7-day rolling mean daily change per thousand
Ecuador - samples tested	2020-03-26	https://www.gestionderiesgos.gob.ec/wp-content/uploads/2020/03/INFOGRAFIA-NACIONALCOVI-19-COE-NACIONAL-26032020-17h00-propuestav2.pdf	Government of Ecuador	Sum of confirmados and descartados	3125	527	0.177	0.030	351.000	0.020	350.429	0.020

Entity	Date	Source URL	Source label	Notes	Cumulative total	Daily change in cumulative total	Cumulative total per thousand	Daily change in cumulative total per thousand	3-day rolling mean daily change	3-day rolling mean daily change per thousand	7-day rolling mean daily change	7-day rolling mean daily change per thousand
Nepal - people tested	2020-04-04	https://github.com/raunakms/covid19nepal/blob/master/data/data_total.tsv	Ministry of Health and Population	Made available on GitHub by Raunak Shrestha	1521	213	0.052	0.007	125.333	0.004	92.286	0.003

Note: [a] Random 2 rows in world time-series testing update

	Country or area	UN continental region[4]	UN statistical region[4]	Population(1 July 2018)	Population(1 July 2019)	Change
164	Guyana	Americas	South America	779,006	782,766	+0.48%
58	Sri Lanka	Asia	Southern Asia	21,228,763	21,323,733	+0.45%

Note: [a] Random 2 rows in world 2019 population

	Rank	Country/Territory	GDP(US$million)
88	83	Slovenia	48,455
204	186	São Tomé and Príncipe	337

Note: [a] Random 2 rows in GDP Ranking

After Join and clean up...

Date	region	confirmed	recovered	deaths	actives	cumulative_test	population
2020-03-26	bangladesh	44	11	5	28	920	163046161
2020-04-09	slovenia	1124	128	43	953	33047	2078654
2020-04-06	czech republic	4822	121	78	4623	91247	10689209
2020-04-21	russia	52763	3873	456	48434	2142604	145872256
2020-03-29	japan	1866	424	54	1388	58036	126860301

Note: [a] Random 5 row in time-series covid 19 table

Date	region	confirmed	recovered	deaths	actives	population	cumulative_test
2020-05-03	vietnam	271	219	0	52	96462106	261004
2020-05-03	italy	210717	81654	28884	100179	60550075	2153772
2020-05-03	turkey	126045	63151	3397	59497	83429615	1135367
2020-05-03	malaysia	6298	4413	105	1780	31949777	195833
2020-05-03	finland	5254	3000	230	2024	5532156	103445

Note: [a] Random 5 row in most recent covid 19 table

region	rank
united states of america	1
china	2
japan	3
germany	4
united kingdom	5
france	6
india	7
brazil	8
italy	9
canada	10

Note: [a] World ranking top 10 by GDP # Further data visualization

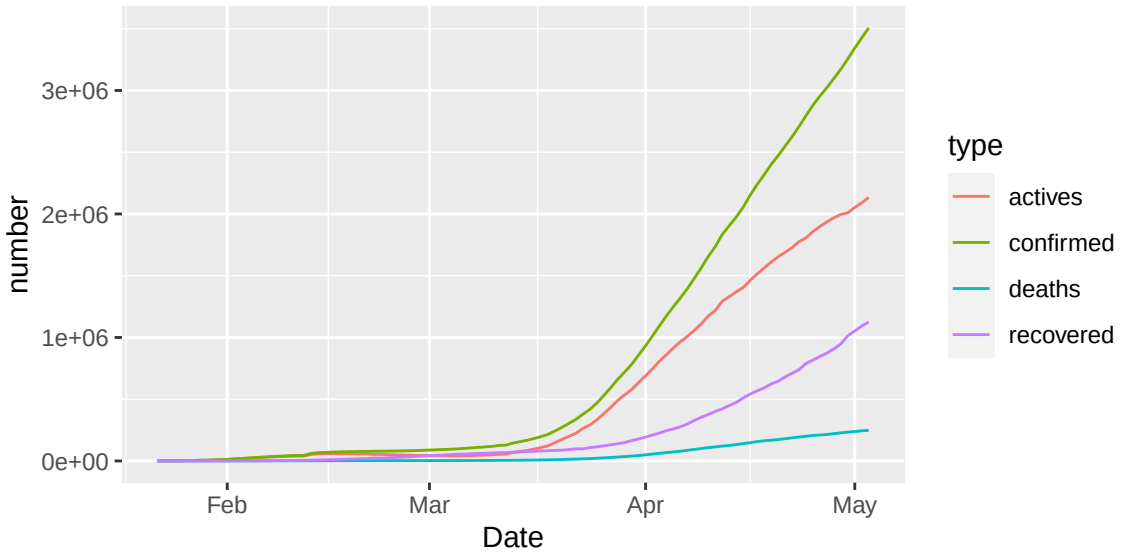
Most updated number of cases & time-series for cases

Now we can get the most updated number of each cases (confirmed, recovered, death, actives) in the world.

By looking at the table, the number of actives in world is still over 2 million cases. And I was quite surprise that the number of recovered is almost one-third of confirmed cases. It indicates that good amount of confirmed people cured by medicine or self-recovered. Since lots of countries' hospitals face frontlines of crisis because of coronavirus, the number of people get recovered is important number to see for checking whether the virus blows over.

The time-line of each case is also helpful to understand overall situation and changes. Even though number of confirmed have high slope and uptrend, active number's slop winces little compare to past month or weeks.

update_date	confirmed	recovered	deaths	actives
2020-05-03	3506729	1125236	247470	2134023



World map by active cases number and rate (by population)

Since we have worldwide active case, it is good to visualize what continent/country have more active case than other. Therefore, I tried to visualize the map in two ways because the actives case is depending on the population of country. In the bottom two tables, the top active countries would not show up in top active rate countries since their number is relatively small in population.

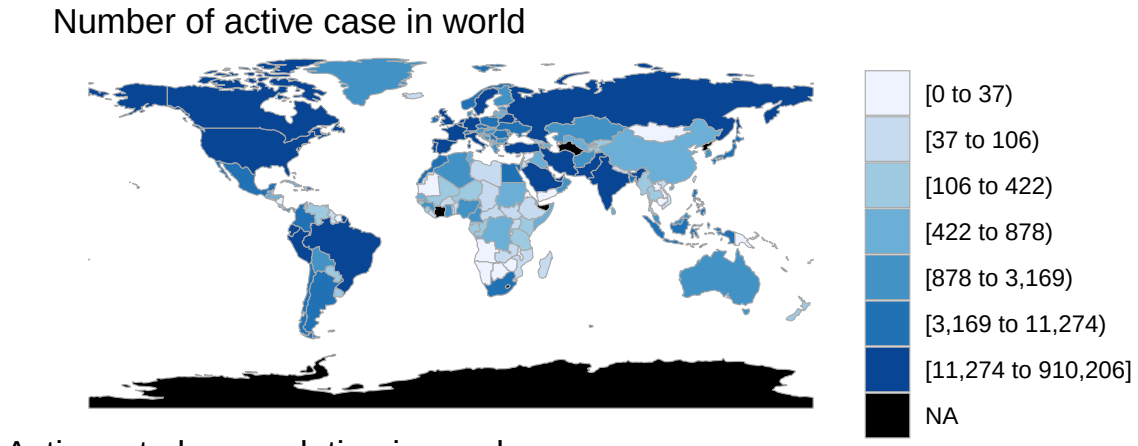
Top 3 country by active cases

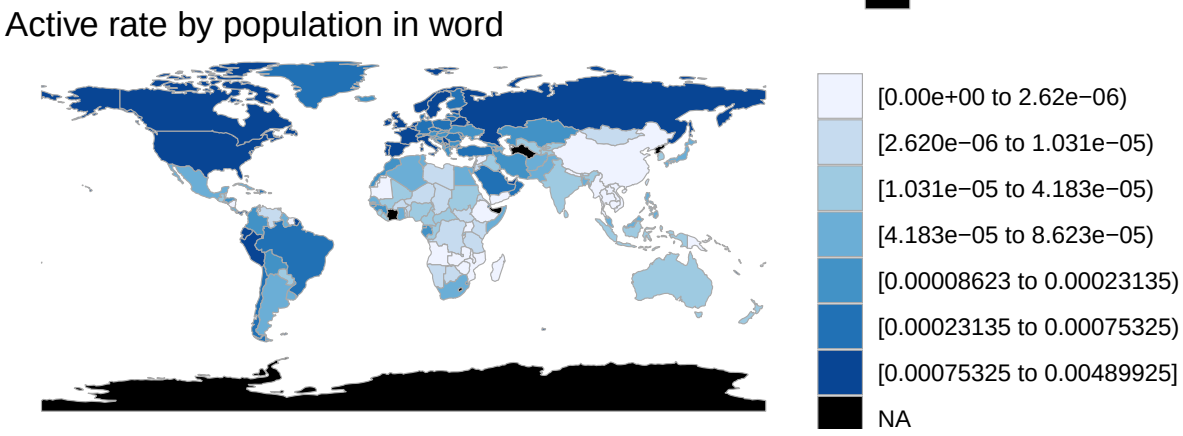
region	actives	ratio_active	population
united states of america	910206	0.2766038	329064917
united kingdom	158421	0.2345929	67530172
russia	116768	0.0800481	145872256

Top 3 country by active rate

region	actives	ratio_active	population
san marino	455	1.3437685	33860
qatar	13875	0.4899248	2832067
singapore	16779	0.2890769	5804337

The bottom two graphs shows some countries have lighter or darker color in active rate map compare to active number itself. Norway, Ireland, Gabon, and Chile have darker color in rate map.



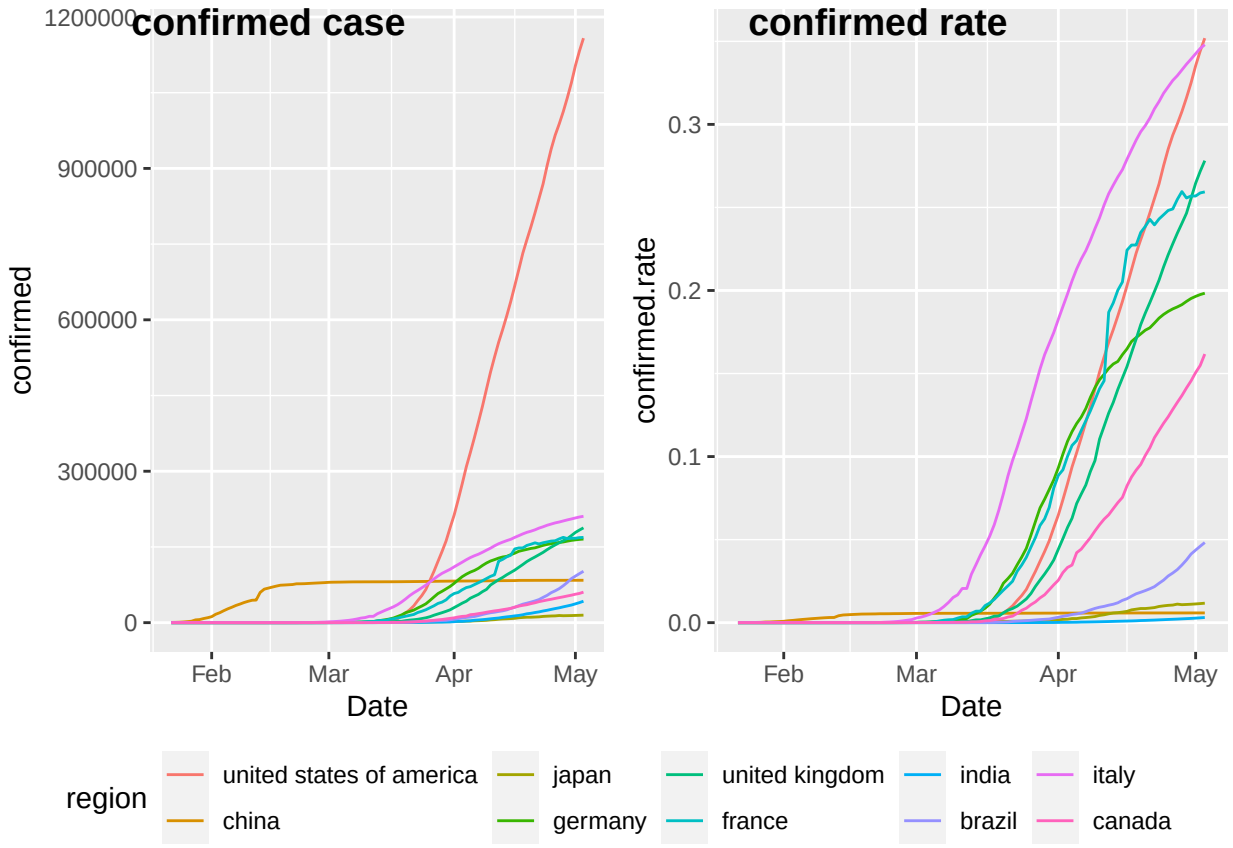


Confirmed cases and Confirmed ratio by population

To look the number deeply, I first plotted the number of confirmed cases and confirmed rate by time in specific countries. I used rate by population because it makes easy to compare various countries in a one plot. Also, I chose only top 10 GDP ranked countries' cases since most of people have interested to see the number in developed countries, and how they deal with this situation.

Before March, most countries have very low confirmed cases, except China. So, I checked up the first date of testing case in countries. Even though some countries started testing in January and February, they did not get confirmed case that much. However, after mid-March the graph starts to have sharp increasement especially in United States.

However, rapid increase in U.S. confirmed cases does not mean that U.S. people tend to have positive result in testing than other countries. Therefore, if I look the right plot (confirmed rate), actually Italy has higher confirmed rate than U.S. Also, in rate plot shows better to understand that actually most of countries have more rapid change around mid-March.

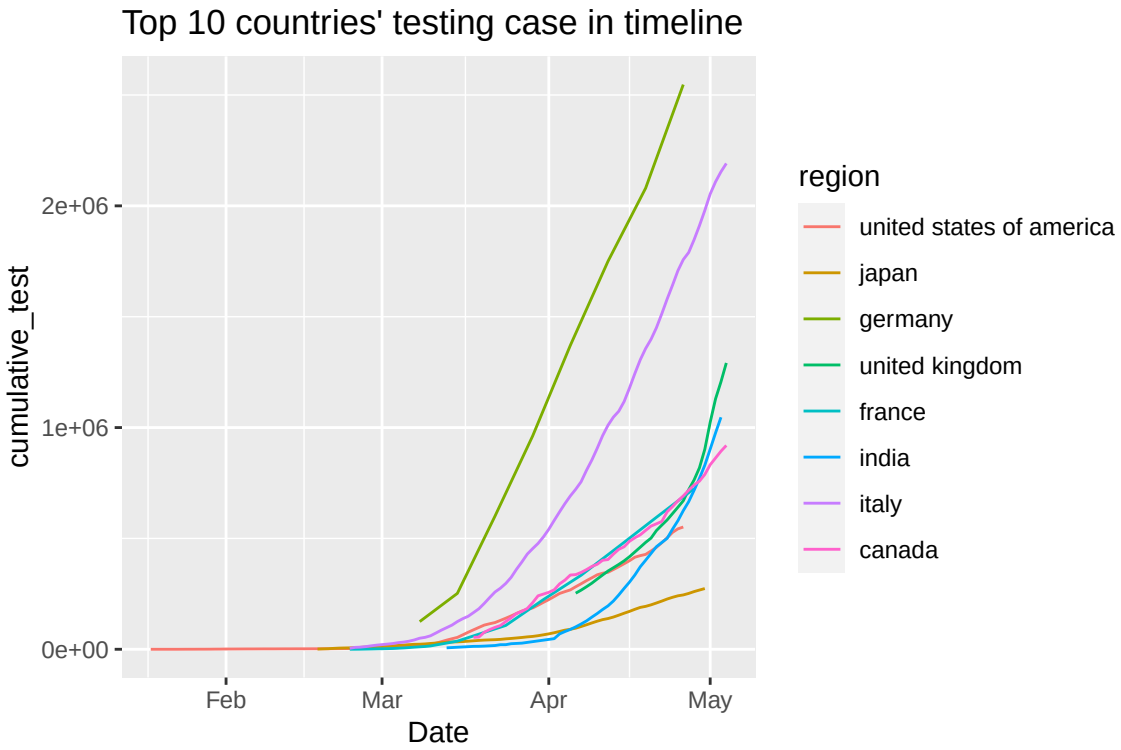


region	first_testing
united states of america	2020-01-18
japan	2020-02-18
germany	2020-03-08
united kingdom	2020-04-06
france	2020-02-24
india	2020-03-13
italy	2020-02-24
canada	2020-03-18

Testing number changes

After I figured out the confirmed case is depending on population, I decided to compare cumulative testing number change also. Since lots of countries do not have enough number of covid-19 testing kit, there might be some countries have lower confirmed cases because they did not test enough. Also, some people said the number of testing in developed countries busted their bubble. They thought developed countries can deal with the virus, but in fact, they also facing difficulty with supplying sanitized product such as mask, sanitizer, gloves, and etc.

By looking at the change in testing in time, most of countries have rapid increase stat at March, but interestingly in India and Japan has the increase point at April even though they are Asian countries which located close to China.



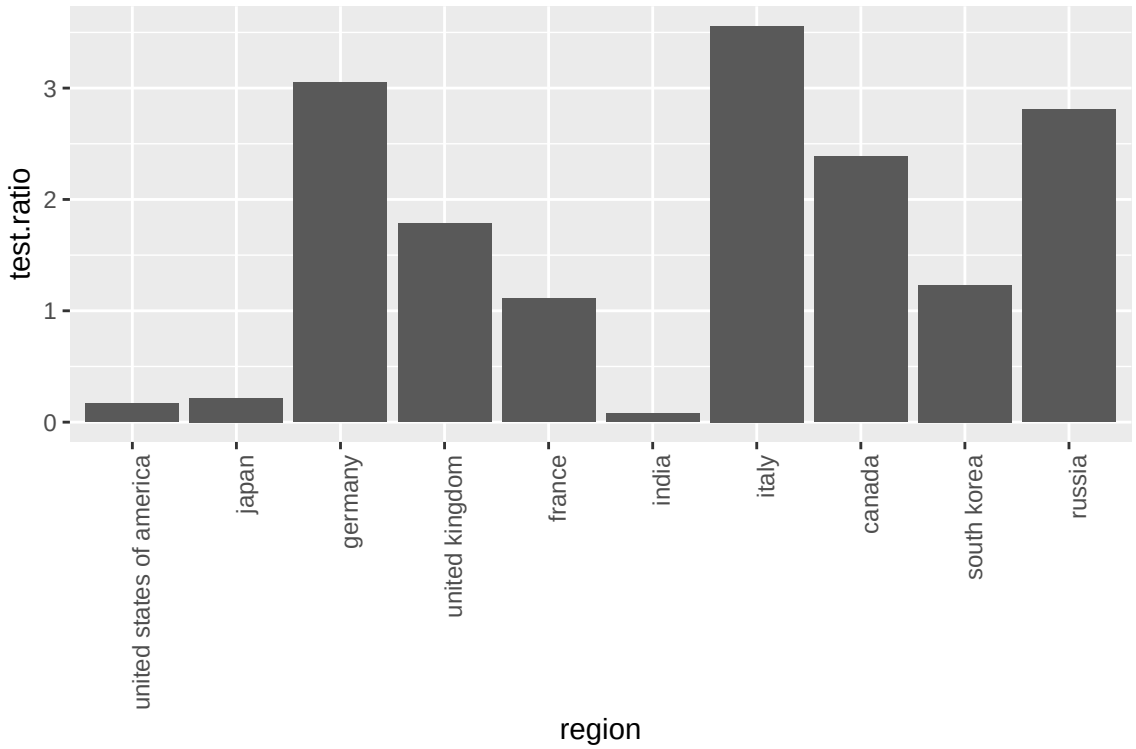
Note: Since China and Brazil does not provide their number of testing, the testing cases change does not contain those two countries.

By looking at the upper plot, it shows that top rank countries do not guarantee they test covid-19. Japan is rank #2 country, but they are the least testing country in May.

To investigate this graph further, I plot the bar plot of tested rate by their populations. If the rate is way small, then it may indicate that country only test a person who has serious symptom because they have lack of physician or doctor, or their population is comparatively larger than others.

U.S., Japan, India have relatively small test rate compare to other top countries. Since I order the bar plot by the world rank, it clearly shows that world ranking is not following the trend of tested ratio.

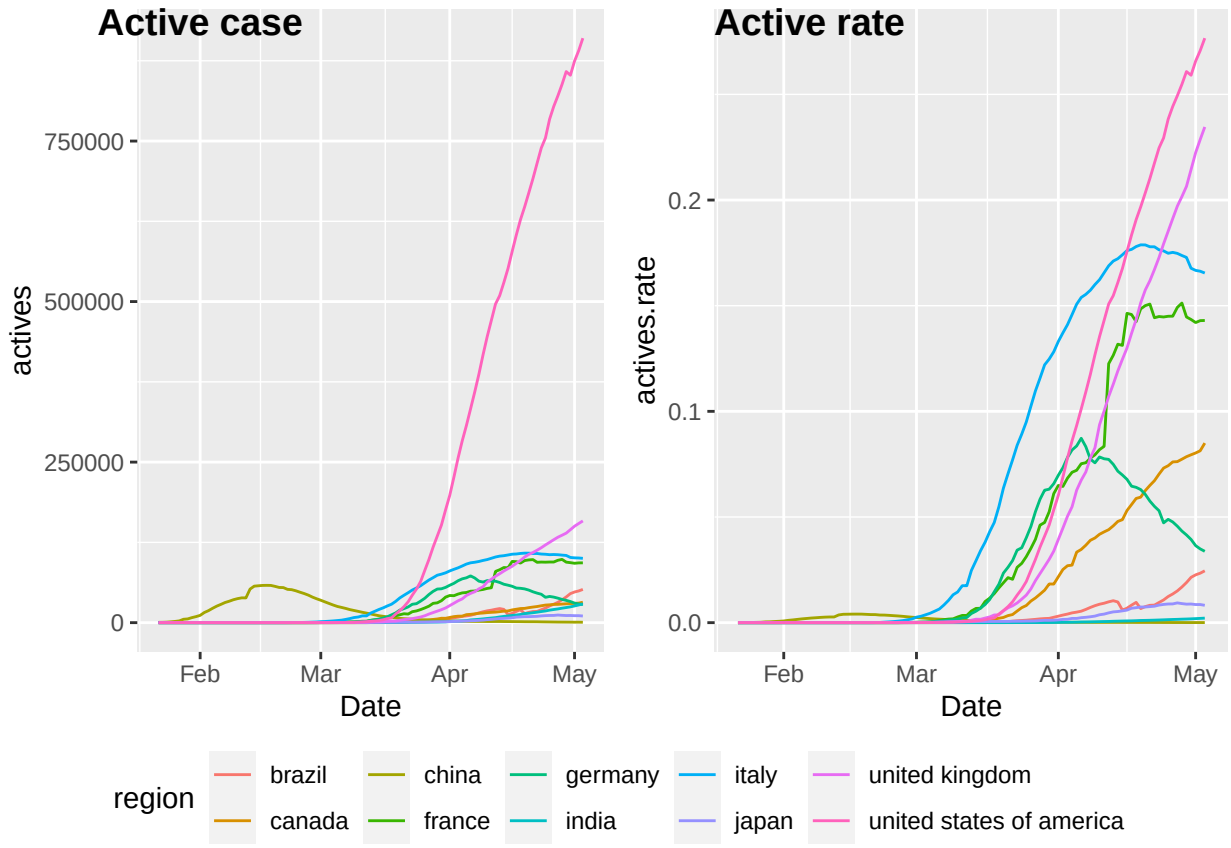
Note: Since some of the countries do not share the cumulative testing number, I plotted top 10 countries who share it.



Active cases and Active ratio by population

After investigating confirmed and tested number, finally I plot the time-line of active case in the world.

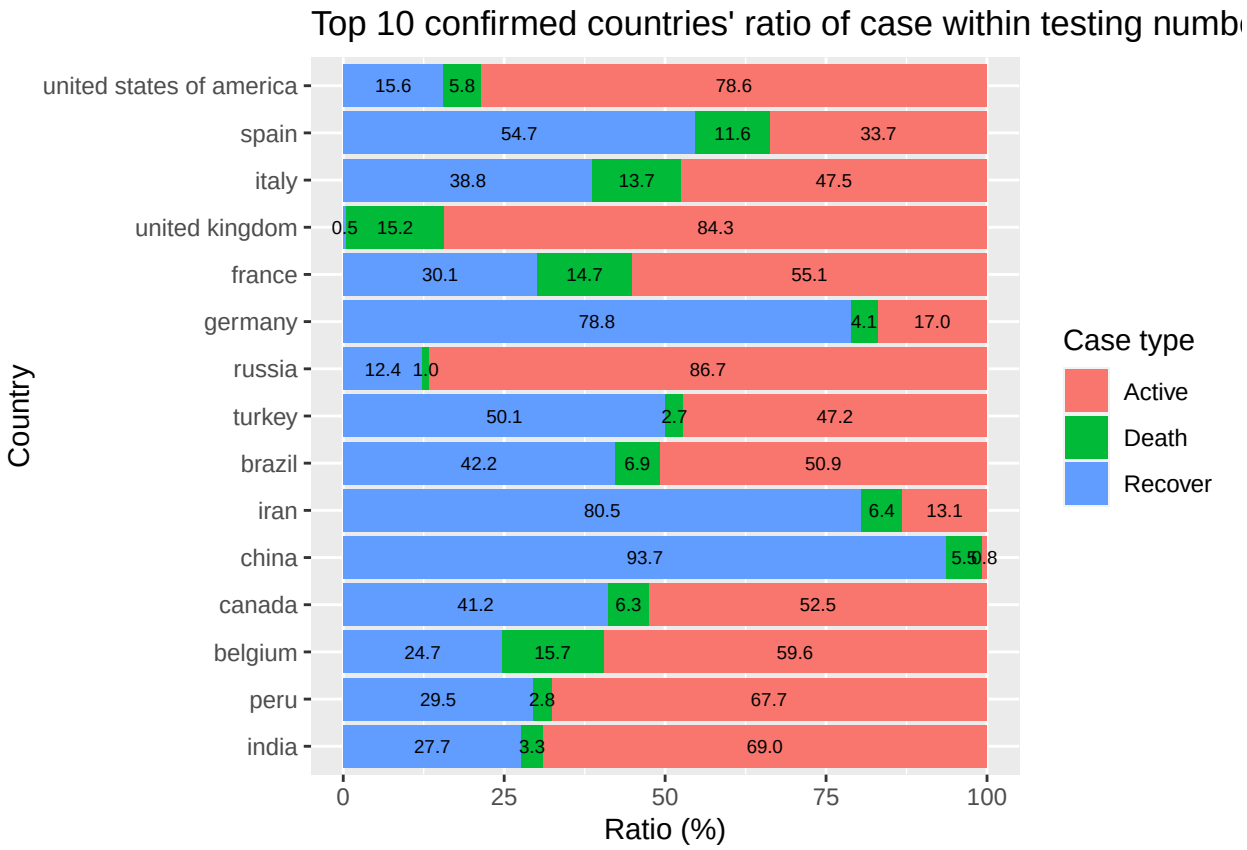
By looking at the bottom graph, some of countries having downtrend after mid-April. However, U.K., U.S., Canada, and Brazil still have uptrend for active cases.



Distribution of active, recover, and death in confirmed cases

After the ingestions data by time series, I wondered how much the case occupied within confirmed cases. To see the distribution well, I grab the top 15 countries who have most confirmed cases.

The proportion is inconsistent between countries. Therefore, it is hard to conclude the trending of each case's distribution. However, the interesting part is death proportion. Except Russia, Peru, Germany, it is quite noticeable. In the next step, I looked into this death portion.



Data set 2:

Distribution of death cases by races in United States

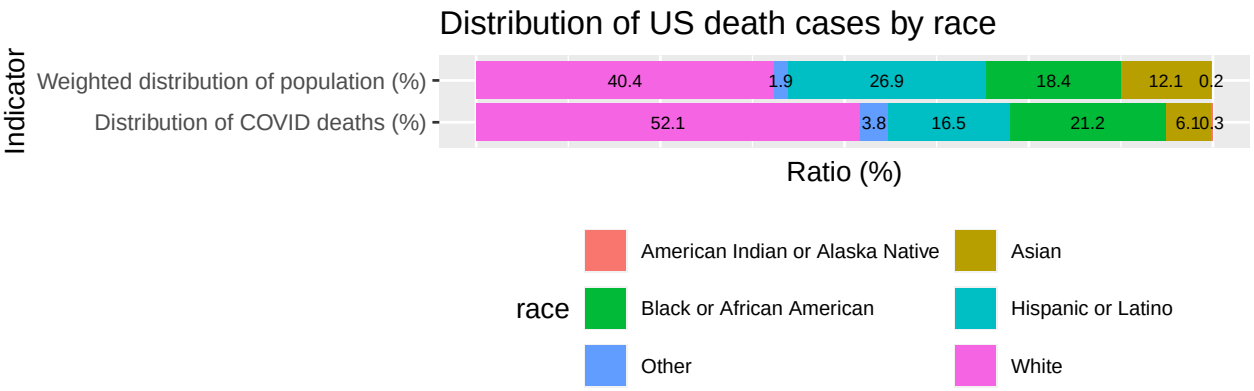
Since the world death rate (deaths / confirmed) is around 7%, I tried to look into more detail information like the distribution of race. Since U.S. is most diverse country and the confirmed number is significantly high, I got U.S. death distribution by race information by race from Centers for Disease Control and Prevention (CDC), https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm.

From the data, I extract the cumulative United State death distribution with 2 cases – Distribution of COVID deaths and Weighted distribution of population.

```
## [1] "Update date: 04/28/2020"
```

Indicator	White	Black or African American	American Indian or Alaska Native	Asian	Hispanic or Latino	Other
Distribution of COVID deaths (%)	52.1	21.2	0.3	6.1	16.5	3.8

Indicator	White	Black or African American	American Indian or Alaska Native	Asian	Hispanic or Latino	Other
Weighted distribution of population (%)	40.4	18.4	0.2	12.1	26.9	1.9



After looking at the distribution of death in U.S., there were significantly large percentage of death report from White than other races even in weighted distribution. By looking at this, obviously we cannot conclude a specific race tend to dead more/less from coronavirus. But good resource to see the trend in diverse country.

Conclusion:

I faced a difficulty to grab right data for testing number. Since some of the countries does not share their testing number, it was hard to show the relationship between testing and confirm or other cases. Even some of them count the number from 2 different sources and they do not match. Also, some of the countries does not update the testing from their source frequently.

However, after visualize the results, I found out higher ranked countries does not guarantee they can handle the situation better than other countries. Some of lower rank countries did more testing and got more recovered cases. We can do further research about relapse cases if we have the data.

Also, by the active graph, even though there were some countries have down trend after April, we should keep eyes on the trend because there is a possibility of having second wave of coronavirus.