COVID-19 Project

Ezra Oppenheimer

07/12/2021

Introduction

In this assignment, we have branched off several different hypotheses into one singular R Markdown file. To keep things organized, and to help differentiate between authors, we have compartmentalized our Rmd into the following sections:

Ezra Oppenheimer's Report

Meizhu Wang's Report

Cagri Isilak's Report

Packages

This project utilizes the following packages. Ensure they are initialized in order for it to work.

```
library(tidyverse)
library(stringr)
library(lubridate)
library(maps)
library(gridExtra)
library(ISLR)
```

Datasets

```
# info about world birthrate
birthrate <- read_tsv("BIRTHRATE.csv")

# full dataset for COVID cases worldwide
covid_cases <- read_csv("https://covid.ourworldindata.org/data/owid-covid-data.csv",
                        col_types = cols(
                          .default = col_double(),
                          date = col_date(format =""),
                          location = col_character()
                        ))

# information regarding COVID vaccinations in particular
covid_vaccs <- read_csv("https://raw.githubusercontent.com/owid/covid-19-data/master/public/dat
a/vaccinations/vaccinations.csv")

# authoritative codebook for the Oxford COVID-19 Government Response Tracker
# EZRA!!! THIS GETS OVERWRITTEN
covid_response <- read_csv("https://raw.githubusercontent.com/OxCGRT/covid-policy-tracker/maste
r/data/OxCGRT_latest.csv",
                           col_types = cols(
                             .default = col_character(),
                             Date = col_date(format = "%Y%m%d"))
                           ) %>%
    mutate(
      across(starts_with("Confirmed"), as.integer),
      across(ends_with("Flag"), as.logical),
      across(ends_with("Index"), as.double),
      across(ends_with("ForDisplay"), as.double),
      )

# cosmetic information about every country, such as names, country codes, and continents.
countries <- read_csv("countries with regional codes.csv") %>%
  select(c("name", "alpha-3", "region")) %>%
  rename(location = name, "iso-code" = "alpha-3", continent = region)


# worldwide fertility @ https://www.humanfertility.org/cgi-bin/stff.php
fertility <- read_csv("humanfertility.csv",
                      col_types = cols(
                        .default = col_double(),
                        CountryCode = col_character()
                      ))

# types of regimes for each region
democracyindex <- read_tsv("DEMOCRACYINDEX.csv")

# population of each country
population <- read_tsv("POPULATION.csv")
```

```
# the values in the table are in `% of GDP` of the region.
educationexp <- read_tsv("EDUEXP.csv")

# GDP per capital in the region
gdppp <- read_tsv("GDPPP.csv")

# Public debt in the region
publicdebt <- read_tsv("PUBLICDEBT.csv")

# GINI index in the region
gini <- read_tsv("GINI.csv")
```

Ezra Oppenheimer Report

Return to Introduction

Hypotheses

For my research, I wanted to tackle a number of hypotheses regarding COVID-19. Two of these stem from conspiracies, whereas the other is more original. Regardless, I hope that with this report, I could shed some light onto these problems:

1. As vaccinations increase, the deaths due to COVID will decrease, however the infections will remain the same.
2. Vaccines do not cause infertility.
3. Authoritarian countries have a lower infection rate than democratic countries.

These are three handy functions which will make filtering easier for future datasets.

```
isPeriod0 <- function(date) {
  date < ym("2021-05")
}

isPeriod1 <- function(date) {
  date >= ym("2021-05") & date < ym("2021-09")
}

isPeriod2 <- function(date) {
  date >= ym("2021-09")
}
```

Hypothesis 1

Return to Hypotheses

As vaccinations increase, the deaths due to COVID will decrease,

however the infections will remain the same.

I was wondering if vaccines have affected COVID-19's spread. Conspiracists suggest that vaccines don't work, *period*. However, I've heard elsewhere that vaccines prevent *fatalities*, yet *do not stop infections*.

This is a hard one to pin down scientifically. The raw infection & vaccine data cannot validate the strength of the vaccine, because:

- COVID cases may involve unvaccinated people. (the data doesn't differentiate these)
- Vaccines may not be widespread in certain regions.
- Missing `total_cases` and `total_deaths` entries, which will hinder accuracy for cumulative sums.

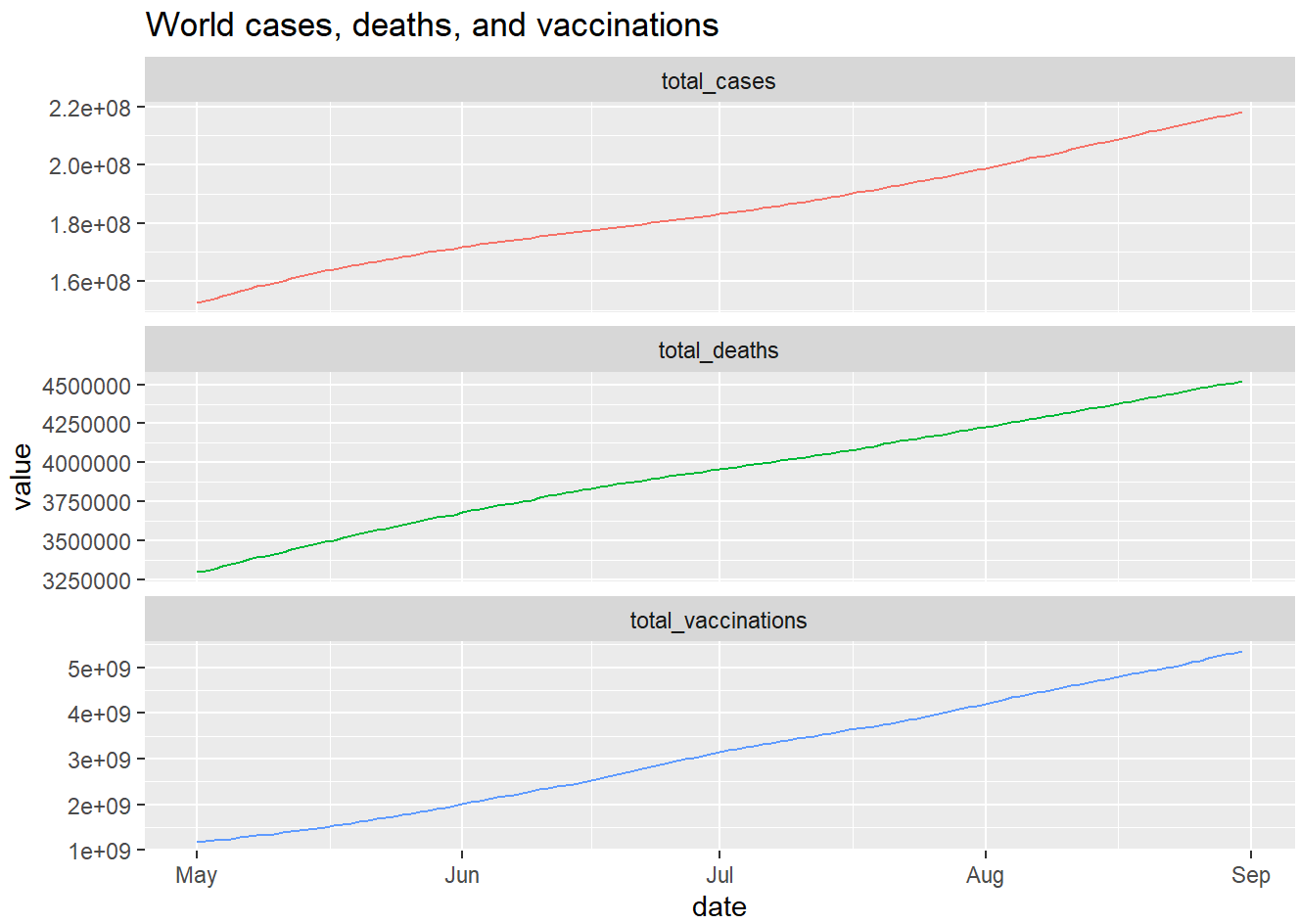
Dates

The interval between May 1st, 2021 and September 1st, 2021 is the most effective for this thesis. The vaccine was entering full rollout during early May (at least in Canada). This is where change should really start to occur.

Anticipated Factors

As stated above, the number of vaccinated infectees & casualties should play a major role in this statistic. I am only interested in COVID cases & deaths involving vaccinated people, which is hard to come by, since this dataset isn't courteous to this statistic.

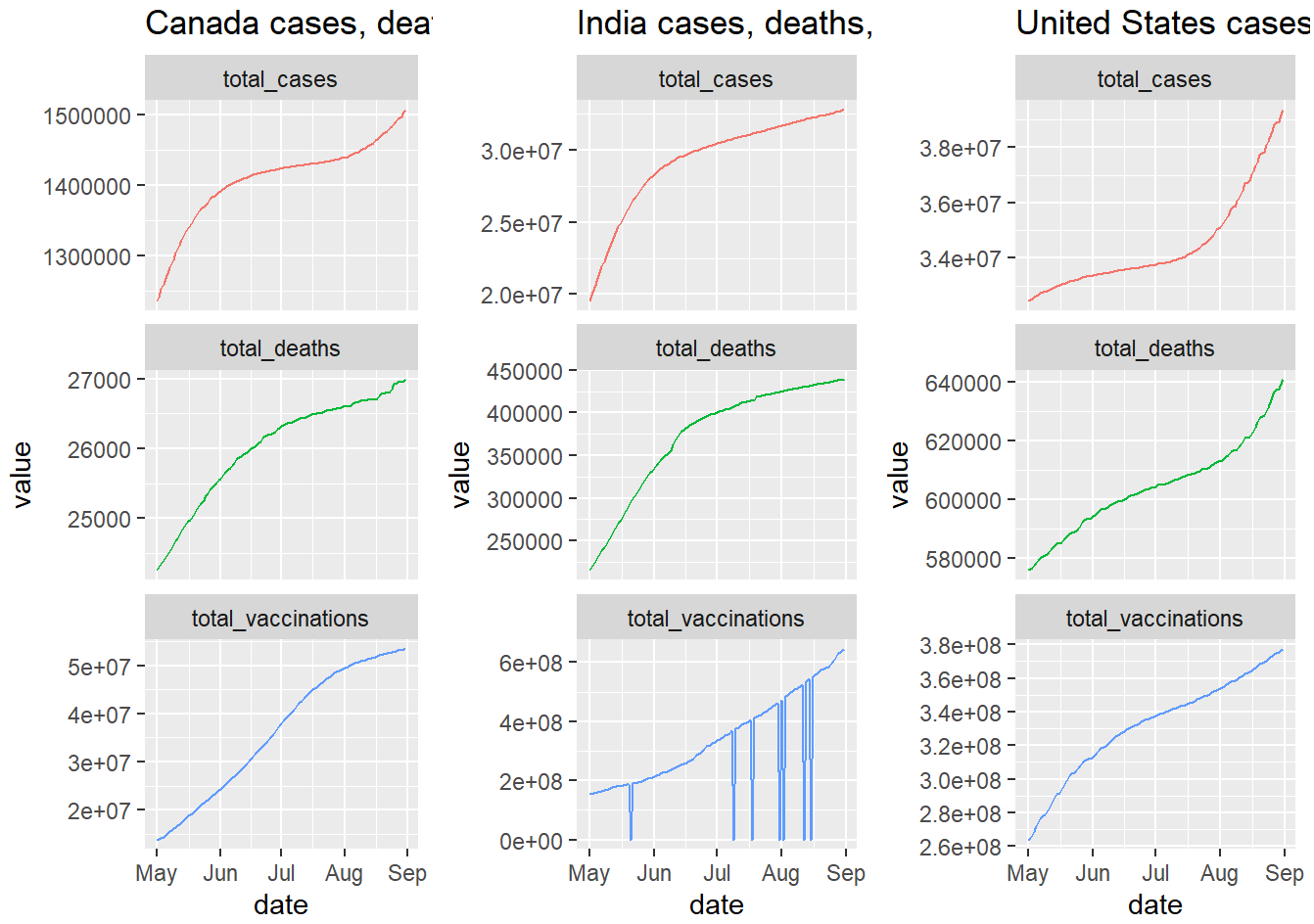
```
# filters the datasets to our desired time interval
covid_cases1 <- covid_cases %>%
  filter(isPeriod1(date)) %>%
  select(c(location, date, total_cases, total_deaths))

covid_vaccs1 <- covid_vaccs %>%
  filter(isPeriod1(date)) %>%
  select(c(location, date, total_vaccinations))

# joins the relevant datasets into one
covid_joined1 <- covid_cases1 %>%
  inner_join(covid_vaccs1, by = c("location", "date")) %>%
  left_join(countries, by = "location")


# given a vector of locations, this will plot each
hypothesis1 <- function(i) {
  covid_joined2 <- covid_joined1 %>%
    filter(location == i) %>%  # this is where you can select which filter you want, be it count
ries or continents
    group_by(date) %>%
    summarise(
      total_cases = sum(total_cases, na.rm = TRUE),
      total_deaths = sum(total_deaths, na.rm = TRUE),
      total_vaccinations = sum(total_vaccinations, na.rm = TRUE)
    ) %>%
    gather(c("total_vaccinations", "total_cases", "total_deaths"), key = "key", value = "value")

  (ggplot(covid_joined2, mapping = aes(x = date, group = key, color = key)) +
    ggtitle(str_c(i, " cases, deaths, and vaccinations")) +
    geom_line(mapping = aes(y = value), show.legend = FALSE) +
    facet_wrap(~key, scales = "free_y", ncol = 1))
}

# plot graph
hypothesis1("World") # this "World" statistic is included in the dataset, which accounts for the
global statistics
```



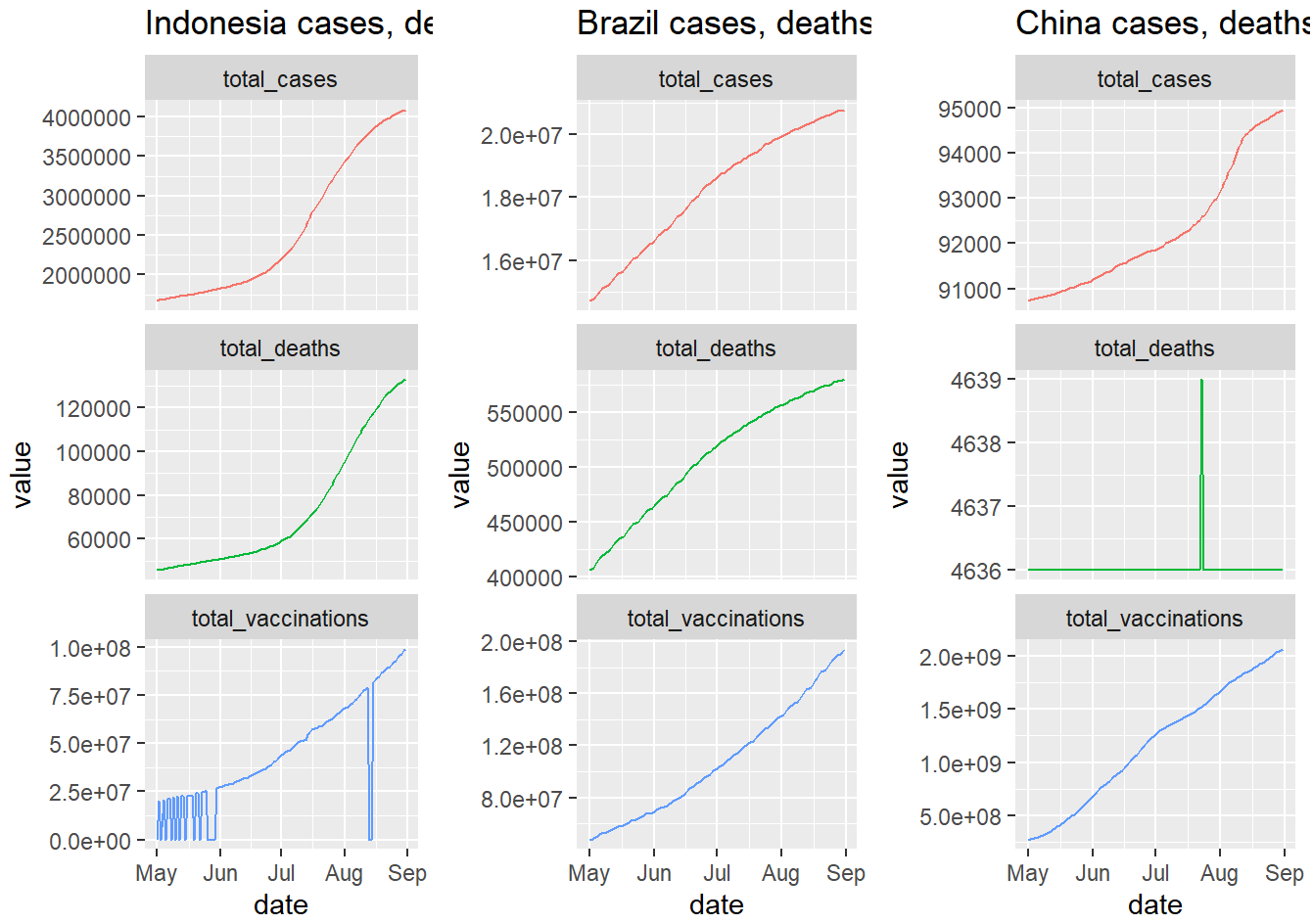
From the data above, it's disheartening to see that vaccines haven't hindered COVID-19 globally. **This does not mean that vaccines malfunction**. On the contrary, the data could suggest that, despite vaccines being in circulation, we haven't seen any big change in infections. Perhaps people aren't being vaccinated, or the vaccine *is* working, but not as effectively as expected.

Let's try a few more countries as a thought experiment.

```
plot1 <- hypothesis1("Canada")
plot2 <- hypothesis1("India")
plot3 <- hypothesis1("United States")
grid.arrange(plot1, plot2, plot3, nrow=1, ncol=3)
```




```
plot4 <- hypothesis1("Indonesia")
plot5 <- hypothesis1("Brazil")
plot6 <- hypothesis1("China")
grid.arrange(plot4, plot5, plot6, nrow=1, ncol=3)
```



HAHAHAhahaha, oh my god… Sorry… I am too fixated by the China statistic. *Despite* COVID originating in China, and *despite* the Chinese government ***literally ghettoizing the Wuhan citizens during COVID's outbreak (https://youtu.be/wEkldGht-S8)***, OF COURSE the Wuhan virus would have only killed a couple of thousand people. Got it.

Besides the CCP's fake data, there seems to be no correlation between the vaccinations and the case & death count. The red and green curves seem to be unaffected by the curvature of the vaccination curve. For these reasons, someone cannot suggest that cases & deaths have gone down due to vaccines.

There are also little graphical glitches in the India and Indonesia graphs. Those would be missing data entries from the `covid_vaccs` dataset. It happens for a lot for this dataset, so I am not particularly surprised. *(Lesson learned, I fix this in Hypothesis 2)*

Overall, the statements that I've taken away from this are:

> Vaccines have not stopped the spread of COVID-19 globally. The death rate is still increasing globally, and the infection rate remains unchanged.

Not what I was expecting. I was expecting more middle-ground, but I guess the doom-and-gloomers are right that COVID isn't getting better.

Hypothesis 2

Return to Hypotheses

Vaccines do not cause infertility.

Hysteria became rampant when vaccines were introduced. Conspiracists believed that taking the vaccine would depopulate the earth to euthanize the population. I mean, what is there to say, except for us to get into it! :)

Dates

So, the dates here are kind of deviating from the DATA-100 assignment. Conception of a child takes approximately 9 months, which means that the data will only be meaningful 9 months after vaccination. As of December 8th, 2021, vaccines have onlt existed for around 10 months, so it is hard to say if I will get any meaningful results using *any* date.

Anticipated Factors

I am not expecting there to be a correlation between these two datasets. It is simply cathartic to *see* that there is no depopulation scheme in the works. The only limitation with this is, this cannot confirm *scientifically* that vaccines affect fertility, because:

- Births may involve unvaccinated people. (the data doesn't differentiate these)
- Vaccines may not be widespread in certain regions.
- Assuming there *is* a decrease in births, it could be chalked up to antinatalist movements, or other factors.

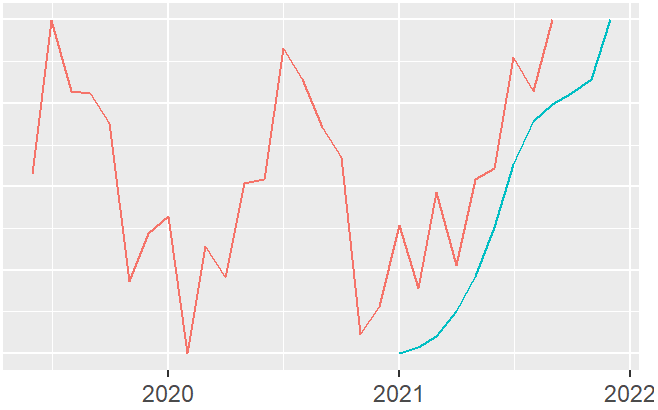
```
# first I need to tidy the fertility dataset
fertility2 <- fertility %>%
  rename("iso-code" = CountryCode) %>%
  gather("January", "February", "March", "April", "May", "June", "July", "August", "September",
"October", "November", "December", key = "month", value = "births") %>%
  mutate(date = ym(str_c(Year, month))) %>%
  inner_join(countries, by = "iso-code") %>%
  select(location, date, births)


# secondly, i was pissed off that `covid_vaccs` contained so many NA entries for total_vaccinati
ons. this was unacceptable, so i interpolated it using a cumulative sum approach. by sampling th
e daily_vaccinations, i was able to eliminate all NA entries, and it snaps back to its ground-tr
uth value when applicable.
covid_vaccs2 <- tibble(
    location = c(),
    date = c(),
    total_vaccinations = c(),
    interpolated_vaccinations = c()
  )

for (country in unique(covid_vaccs $ location)) {
  temp <- covid_vaccs %>%
    filter(location == country)

  cum_sum <- c(temp $ total_vaccinations [1])
  for (j in 2:length(temp $ daily_vaccinations)) {
      value <- temp $ daily_vaccinations[j]
      if (is.na(value)) {
        value <- 0
      }
      value <- cum_sum[length(cum_sum)] + value

      if (!is.na(temp $ total_vaccinations[j])) {
        value <- temp $ total_vaccinations[j]
      }
      cum_sum <- c(cum_sum, value)
  }

  covid_vaccs2 <- rbind(
    covid_vaccs2,
    tibble(
      location = temp $ location,
      date = temp $ date,
      total_vaccinations = temp $ total_vaccinations,
      interpolated_vaccinations = cum_sum
  ))

}
```

Finally with that out of the way, I can accurately process the data.

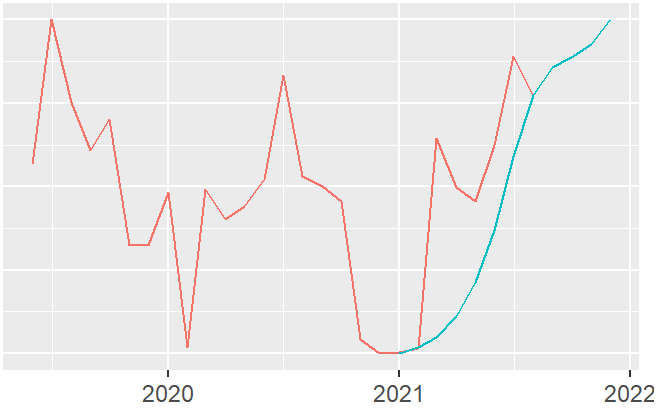
```
covid_joined1 <- left_join(fertility2, covid_vaccs2, by = c("location", "date")) %>%
  arrange(location, date) %>%
  select(-total_vaccinations) %>%
  rename(total_vaccinations = interpolated_vaccinations) %>%
  gather(births, total_vaccinations, key = "key", value = "value")

hypothesis2 <- function(i) {
  ggplot(
    left_join(fertility2, covid_vaccs2, by = c("location", "date")) %>%
      arrange(location, date) %>%
      select(-total_vaccinations) %>%
      rename(total_vaccinations = interpolated_vaccinations) %>%
      filter(location == i & date >= ym("2019-06")) %>%
      mutate(
        births = (births - min(births, na.rm = TRUE)) / (max(births, na.rm = TRUE) - min(births,
na.rm = TRUE)),
        total_vaccinations = total_vaccinations / max(total_vaccinations, na.rm = TRUE),
      ) %>%
      gather(births, total_vaccinations, key = "key", value = "value"),
    mapping = aes(x = date, group = key, color = key)
    ) +
    ggtitle(str_c(i, "'s births and vaccinations")) +
    xlim(ym("2019-06"), ym("2021-12")) +
    theme(
      axis.title.x = element_blank(),
      axis.title.y = element_blank(),
      axis.text.y = element_blank(),
      axis.ticks.y = element_blank()
    ) +
    geom_line(mapping = aes(y = value), show.legend = FALSE)
}

temp <- unique(covid_joined1 $ location)
unique_countries <- c()

for (i in temp) {
  covid_joined2 <- covid_joined1 %>%
    filter(location == i & key == "total_vaccinations")
  covid_joined2
  if (sum(covid_joined2 $ value, na.rm = TRUE)) {
    unique_countries <- c(unique_countries, i)
  }
}

for (i in seq(1, length(unique_countries), by = 4)) {
  country1 <- unique_countries [i]
  country2 <- unique_countries [i + 1]
  country3 <- unique_countries [i + 2]
  country4 <- unique_countries [i + 3]
  plot1 <- hypothesis2(country1)
  plot2 <- hypothesis2(country2)
  plot3 <- hypothesis2(country3)
```

```
  plot4 <- hypothesis2(country4)
  grid.arrange(plot1, plot2, plot3, plot4, nrow=2, ncol=2)
}
```

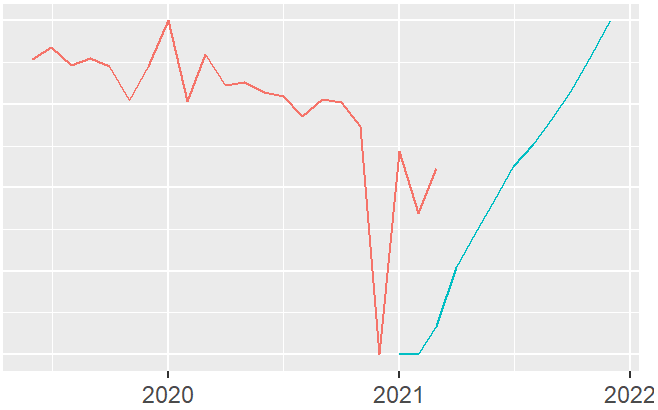
Austria's births and vaccinations



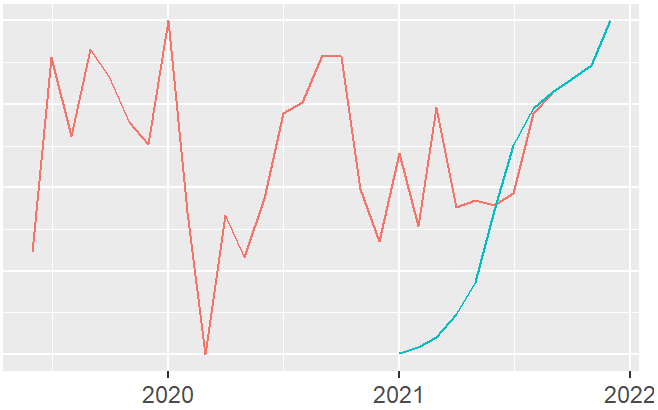
Belgium's births and vaccinations



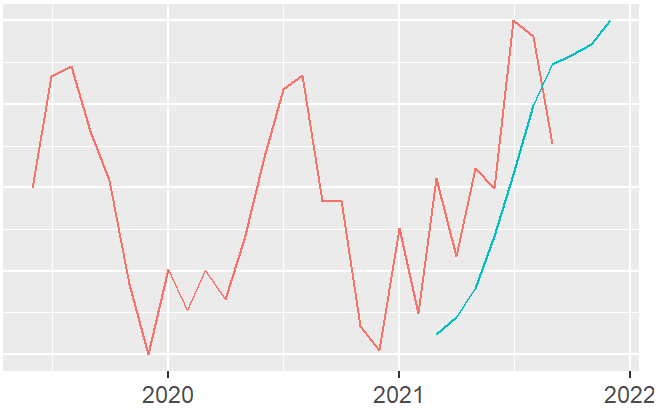
Chile's births and vaccinations



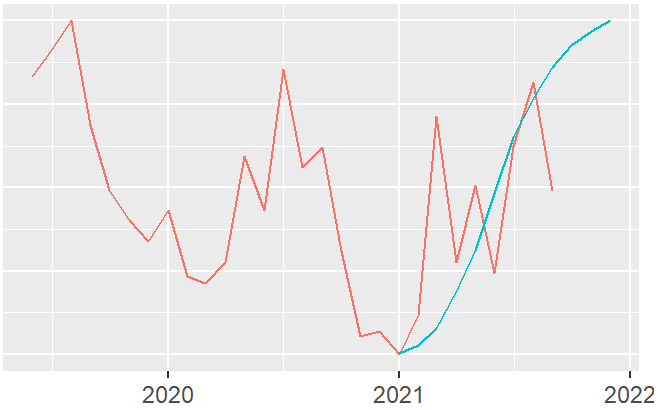
Croatia's births and vaccinations



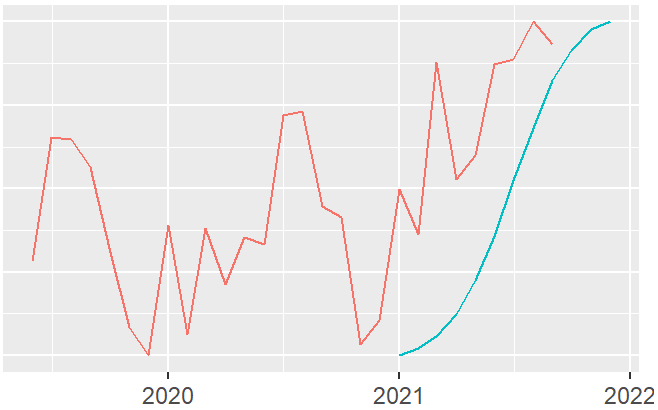
Denmark's births and vaccinations



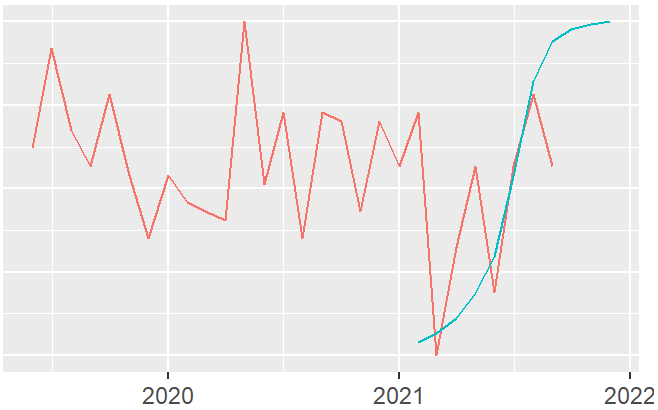
Estonia's births and vaccinations



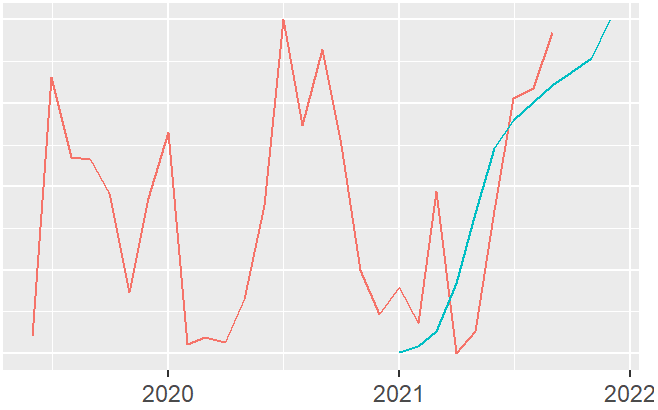
Finland's births and vaccinations



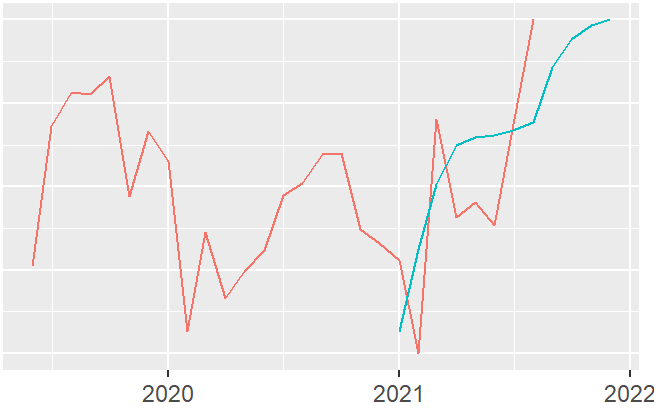
Greenland's births and vaccinations



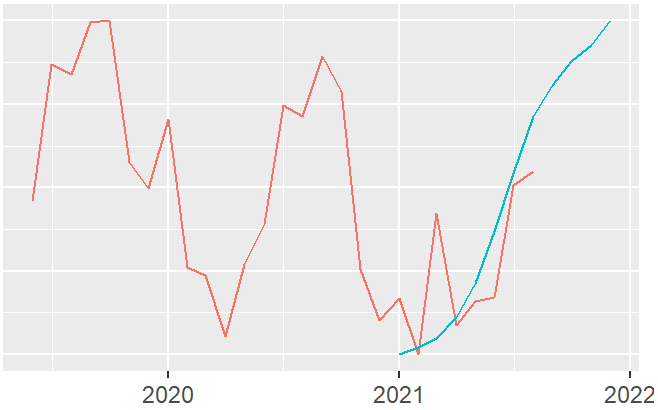
Hungary's births and vaccinations



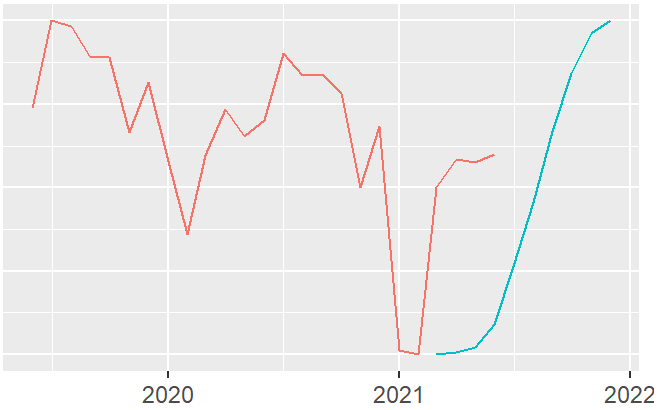
Israel's births and vaccinations



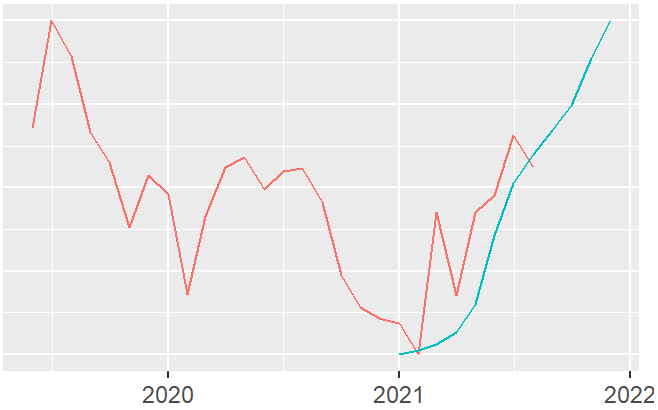
Italy's births and vaccinations



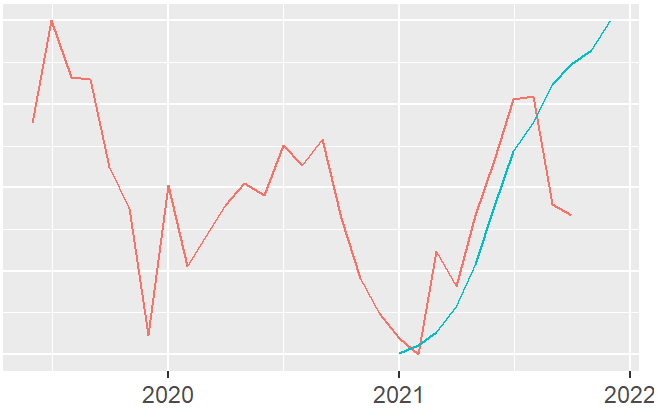
Japan's births and vaccinations



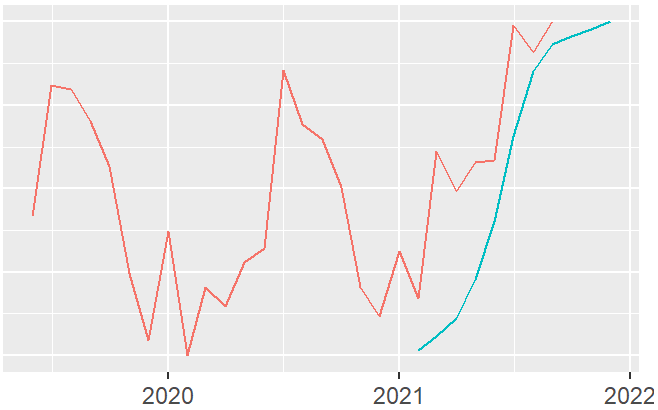
Latvia's births and vaccinations



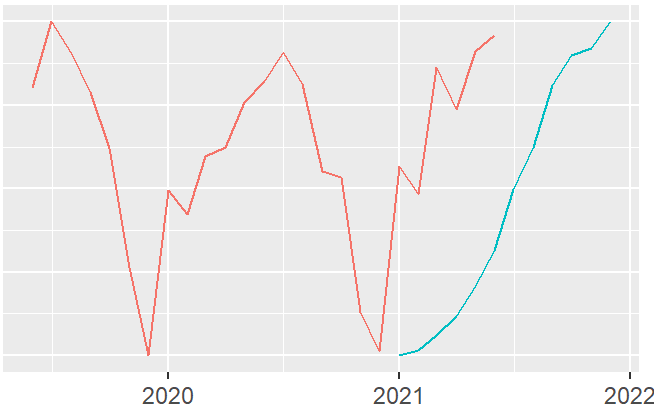
Lithuania's births and vaccinations



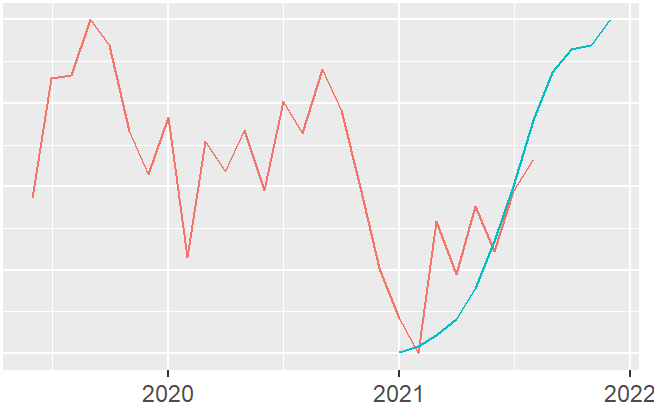
Netherlands's births and vaccinations



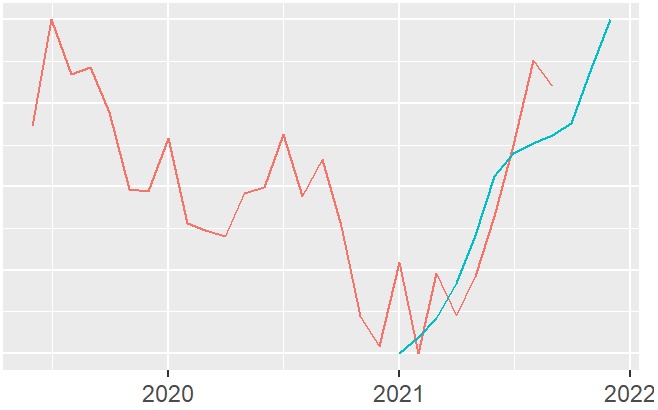
Norway's births and vaccinations



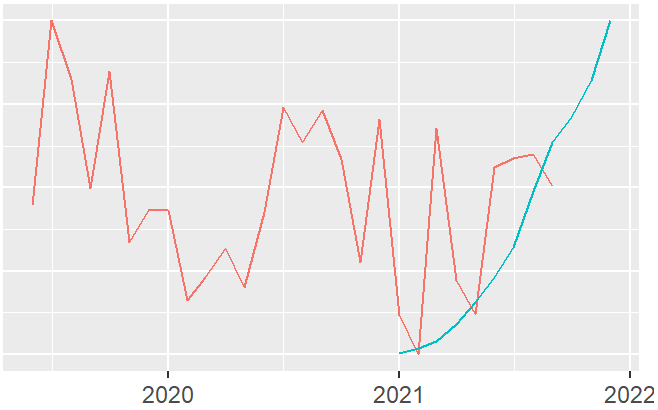
Portugal's births and vaccinations



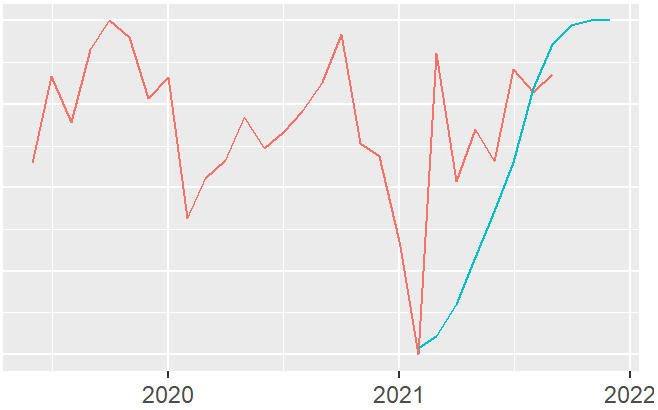
Romania's births and vaccinations



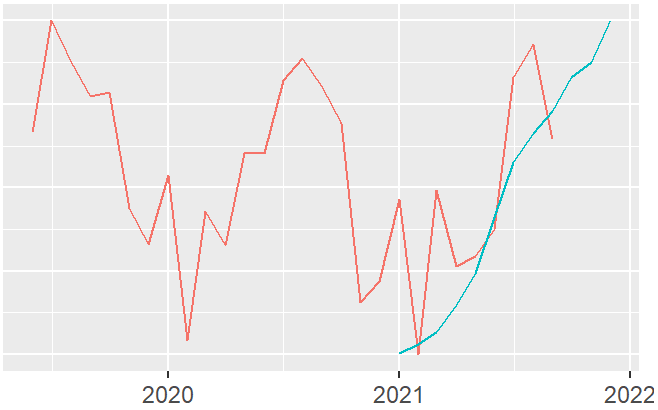
Russia's births and vaccinations



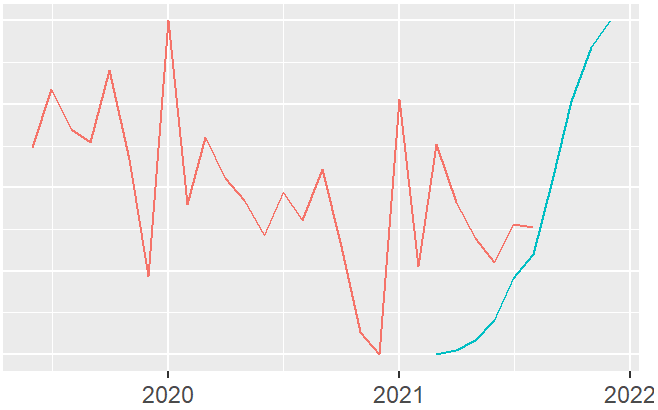
Singapore's births and vaccinations



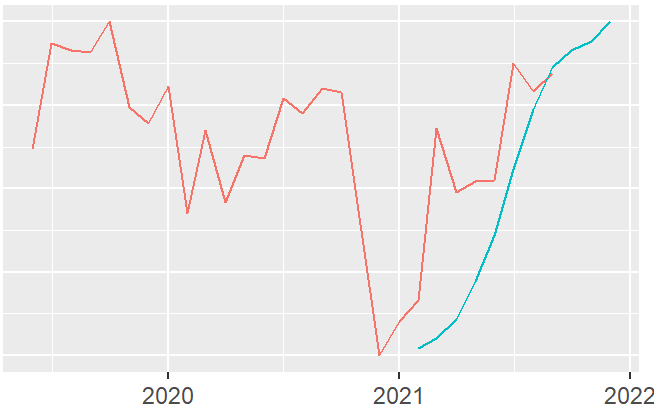
Slovenia's births and vaccinations



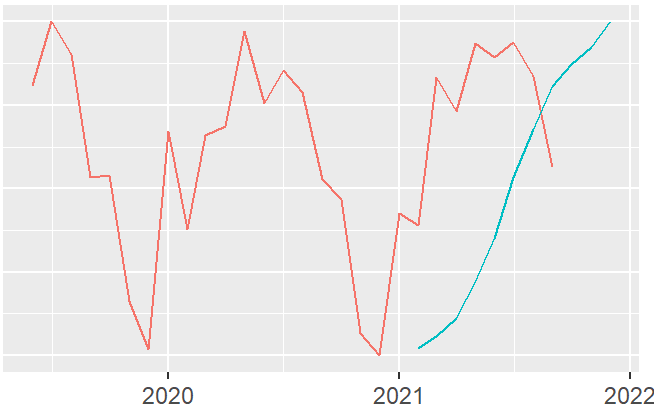
South Korea's births and vaccinations

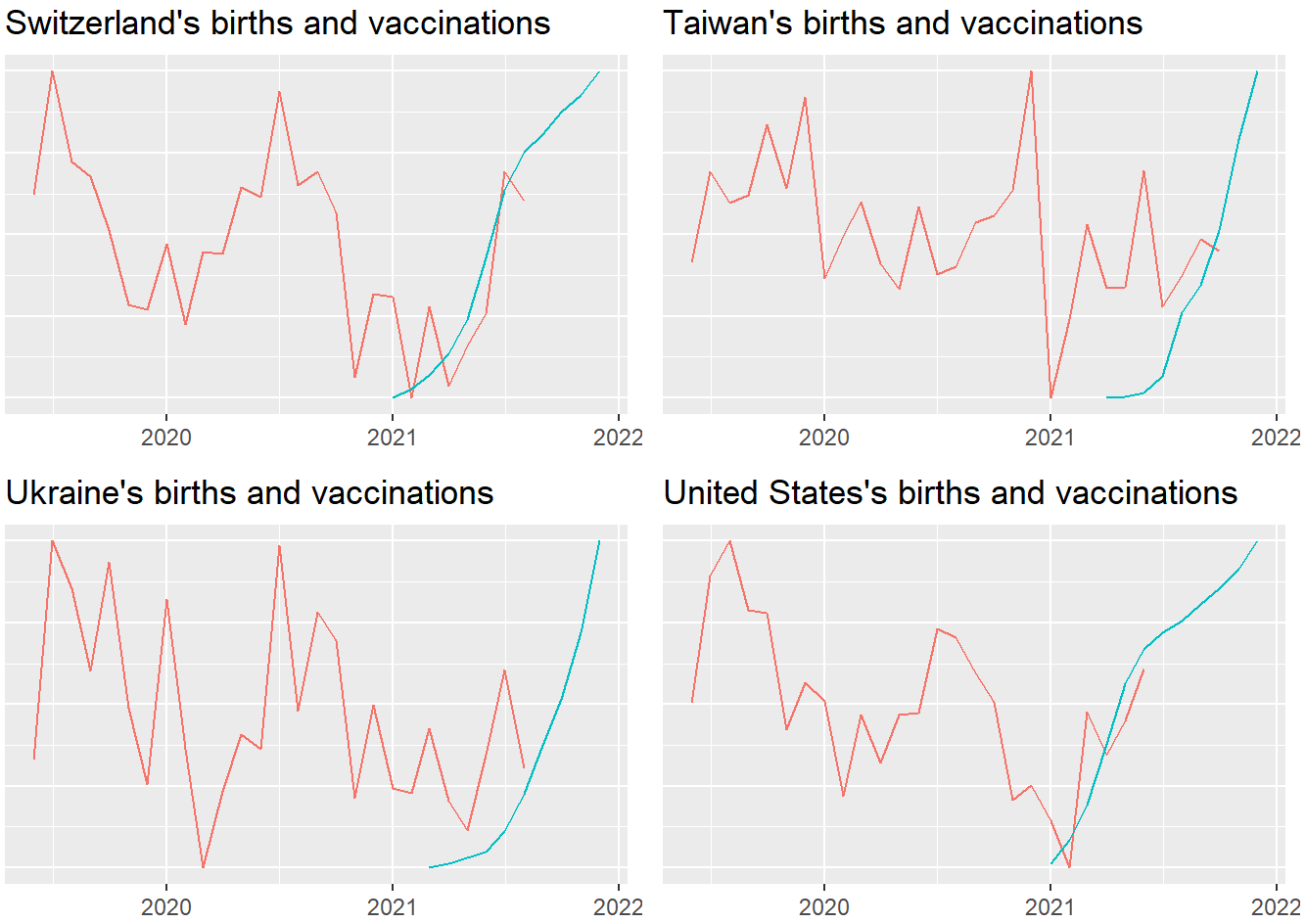


Spain's births and vaccinations



Sweden's births and vaccinations





What did I learn from this? Well, for one, it seems that very little of the `fertility` data overlaps with the `covid_vaccs` data. It is too early to correlate vaccines and infertility to each other. We need a good nine months in order to get conclusive answers, and the dataset that I am using cannot account for the future.

If I had to pick *any* observations, I was initially struck by the unanimous decrease around January 2021. All of the graphs suddenly dip, which would mark precisely the nine-month mark after lockdowns were introduced in March 2020.

Could this mean that, when lockdowns were introduced, did people stop populating? I don't think so. If we look at the earlier trends (before COVID happened), there is a consistent dip at each January mark. This suggests that January's dip is ordinary, which trivializes our suspicion that "COVID stopped repopulation altogether".

Overall, to boil things down into a singular statement:

> As of December 8th, 2021, it is inconclusive that vaccines have lowered births, due to insufficient overlapping data.

Hypothesis 3

Return to Hypotheses

Authoritarian countries have a lower infection rate than democratic

countries.

After seeing the `democracyindex`, I was curious whether totalitarian countries would have a smaller infection rate. After all, totalitarian regimes are more stringent on general policies, therefore the prediction is that since they restrict social gatherings, there would be less infections overall.

Dates

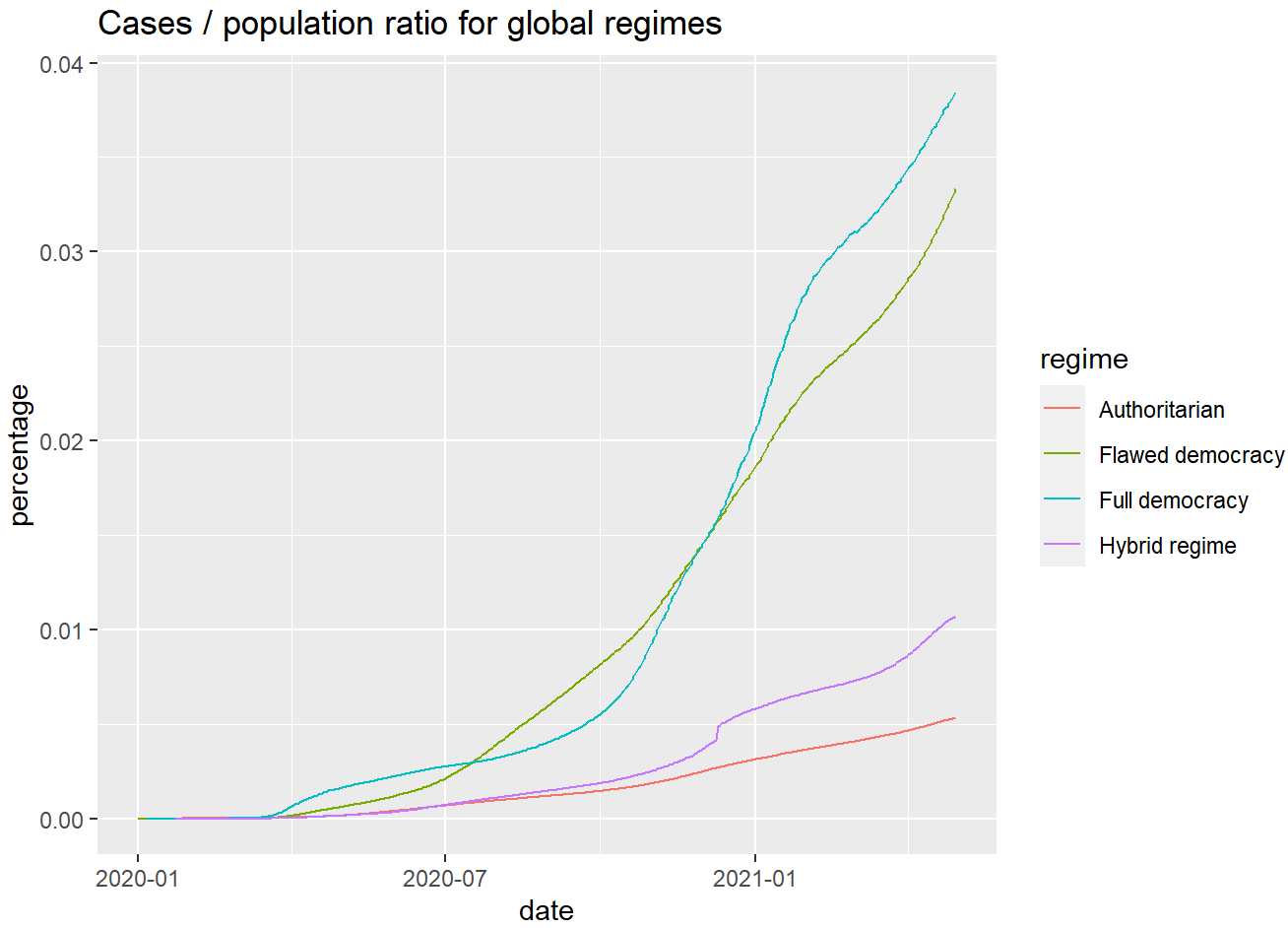
Anytime before May 1st, 2021 is informative for this type of dataset. When COVID broke out, it was crucial for countries to stop it before its spread.

Anticipated Factors

Population definitely plays a huge role in this type of dataset.I am not expecting there to be a correlation between these two datasets. It is simply cathartic to *see* that there is no depopulation scheme in the works. The only limitation with this is, this cannot confirm *scientifically* that vaccines affect fertility, because:

- Births may involve unvaccinated people. (the data doesn't differentiate these)
- Vaccines may not be widespread in certain regions.
- Assuming there *is* a decrease in births, it could be chalked up to antinatalist movements, or other factors.

```
covid_cases1 <- covid_cases %>%
  select(c("location", "date", "total_cases"))

democracyindex1 <- democracyindex %>%
  rename(location = Country, regime = "Regime type") %>%
  select(c("location", "regime"))

population1 <- population %>%
  rename(location = name, population = value) %>%
  select(c("location", "population"))

covid_joined3 <- left_join(covid_cases1, population1, by = "location") %>%
  left_join(democracyindex1, by = "location") %>%
  filter(!is.na(regime)) %>%
  group_by(regime, date) %>%
  summarise(total_cases = sum(total_cases, na.rm = TRUE), population = sum(population, na.rm = T
RUE)) %>%
  mutate(percentage = total_cases / population)

covid_joined3 %>%
  filter(isPeriod0(date)) %>%
  ggplot(aes(x = date, y = percentage, group = regime, color = regime)) +
  ggtitle("Cases / population ratio for global regimes") +
  geom_line()
```



FINALLY a hypothesis that turned out to be correct! As you can see by the results, the countries that are more democratic tend to have a higher rate of infection, whereas the countries that are more stringent have a smaller rate of infection.

This does not necessarily mean that the regime policies created an impact on the spread alone. We must also factor in population *size*. A largely populated area will tend to be prone to infections. It seems that the more populated countries are democratic, which in turn result in more infections. This is all speculation of course. One thing is for sure though:

> Democracies have higher infection rates than authoritarian regimes.

Conclusion

This project has been really fun to tackle. I initially wanted to cram as many conspiracy theories as possible into my hypotheses in order to see which one's were true or not. However, time did not permit me to splurge on ideas, hence I present you with these three humble statements.

This project improved my comfort with R overall. Knowing how to join datasets is an extremely helpful skill, especially when the data is sprawling with different conventions (tidy or untidy).

So what exactly did I learn from this project? I think the most glaring lesson from this project is that China's death numbers are fake. China's national deaths peak at 4636, which is the most bold-faced lie I could ever imagine. They ought to remove those fake numbers from the dataset altogether, because fake numbers can invalidate the weights of other countries.

I also learned that vaccines are not the end-all solution. Despite the world getting vaccinating, there is no visible decline in global cases of COVID. The deaths aren't going down either. This is pretty negative news to all the cynicalists I'm not sure exactly how much intervention we have caused to hinder the spread of COVID, given the underwhelming data, but at least it has altered my view on the sentiment, "Vaccines is stopping COVID-19! Look at the decreases! Pfizer/Moderna is the final solution." Overall, this is very bad news for incentivizing COVID-19 vaccines. Since there is no decrease of cases due to vaccinations, this could cause a pushback against vaccine passports and others.

Another thing that I learned is that, people who claim that vaccines cause infertility were just spouting it on a basis of hysteria. At the time there was no way to prove that the vaccines caused infertility, and there STILL isn't enough time to conclude that vaccines cause infertility. People like to jump the gun. My lesson is to not accept theories at face value when they are conceived very early on and/or are fueled by hysteria.

Lastly, Totalitarian countries are doing *something* right! Just look at their low cases! Indeed, the more stringent countries have a lower increase of COVID cases than democratic countries. This could either be chalked up to smaller populations, or more hopefully, due to more serious COVID laws.

Thank you for reading my report!

Meizhu Wang Report

Return to Introduction

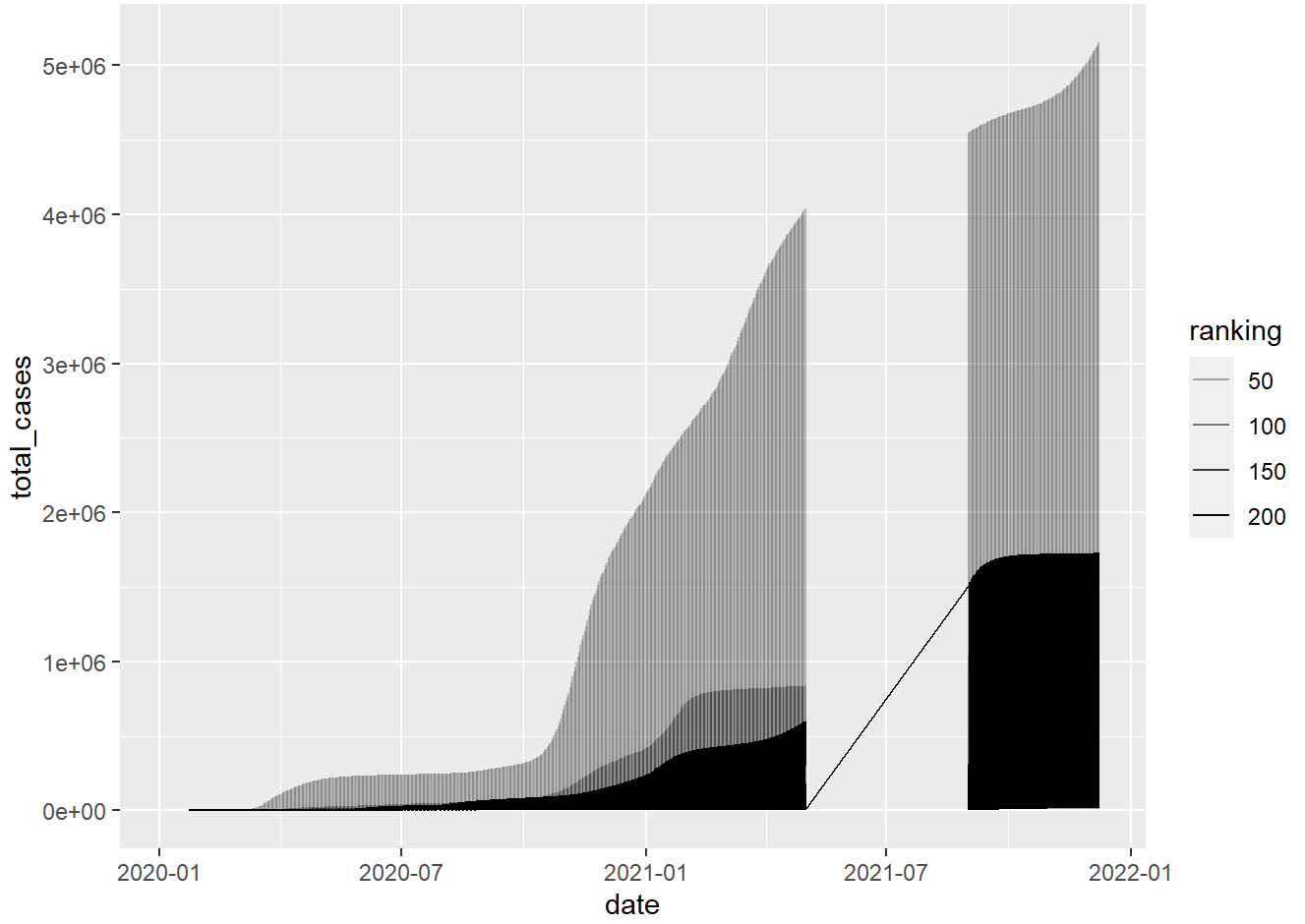
Public debt and COVID-19:

When a country has a large portion of public debt, it means the country owe much to their citizens. In this case, when an unintentional crisis comes such as COVID-19, the ability of handle the crisis and recovery might be suffered. Defining the relationship between the COVID-19 and the public debt would help a country to reconsider the portion of their debt and reduce the risk to experience an out of controlled crisis. Under COVID-19, the economy has recession that many people lose their jobs and must stay at home without earnings. To make the society stable, governments need to spend on subsidies and help economy recovery. However, with large portion of public debt, governments have less money available that can be spent on the important industry such as the health caring or hospital and other essential infrastructures. In this way, we assume that with a larger portion of public debt would decreases the performance against COVID-19(slower vaccination, more cases, more death per million). To clarify the difference between the country with different public debt, before May 01 and between May and Sept. would be better to see the difference in countries or territories ability to deal with the crisis. We pick `total cases` , `new cases` , `total deaths` , `ICU patients` , and `positive rate` to analyze the relationship between the public debt and COVID-19. First, we pick top ten and bottom 9 countries of the public debt and construct the graphs on 5 variables.(Colored based on whole world/Black based on Top 10 &bottom 9 countries/territories)

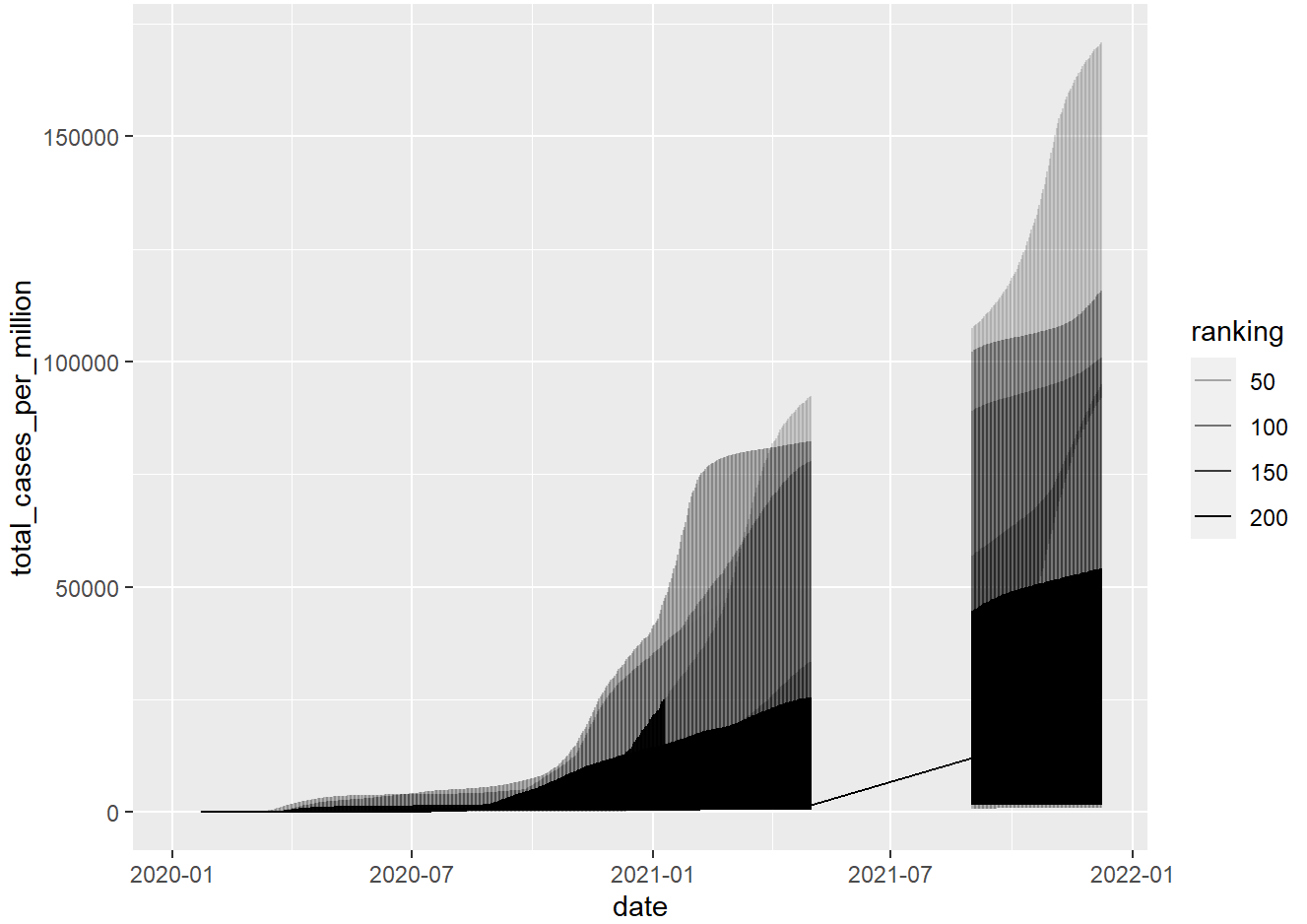
```
covid<-covid_cases%>%filter(date<=ymd("2021-05-01")|date>=ymd("2021-09-01"))
```

```
covid_pubdebt<-left_join(publicdebt, covid, by = c("name" = "location"))
pubdebt_top10<-covid_pubdebt%>%filter(ranking<=10|ranking>=199)
```

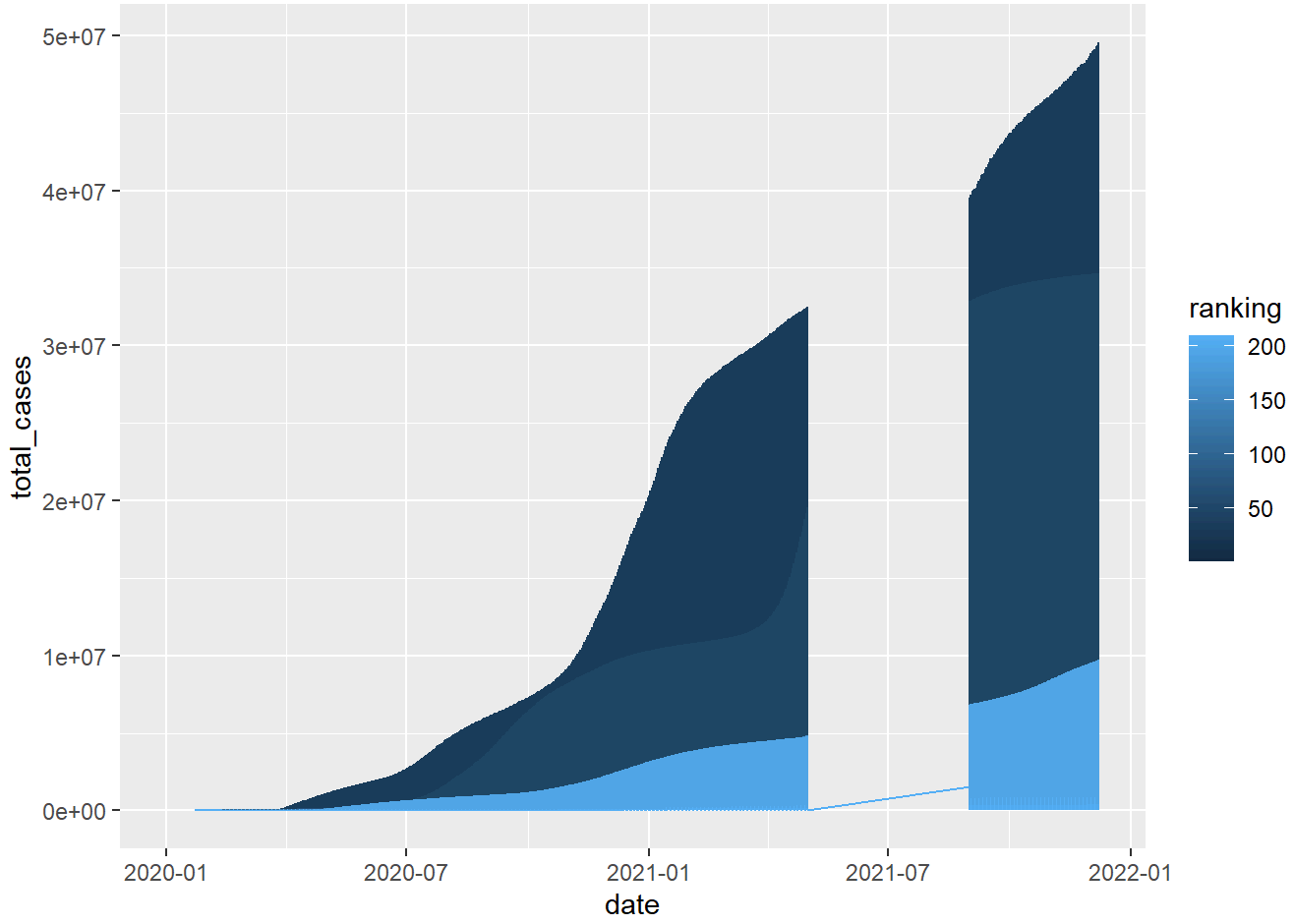
```
ggplot(pubdebt_top10,mapping = aes(x=date,y=total_cases))+geom_line(mapping = aes(alpha=rankin
g))
```



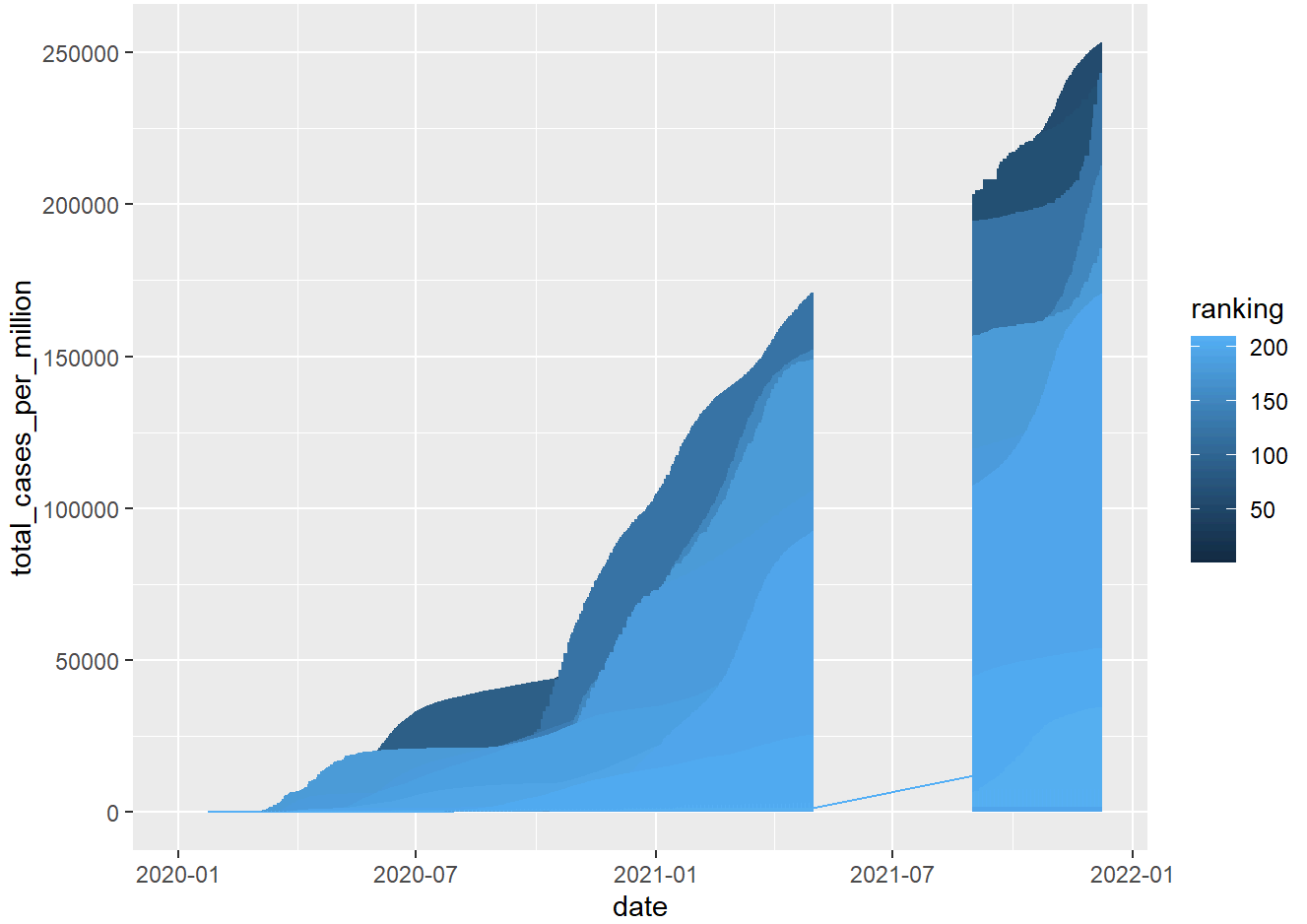

```
ggplot(pubdebt_top10,mapping = aes(x=date,y=total_cases_per_million))+geom_line(mapping = aes(al
pha=ranking))
```




```
ggplot(covid_pubdebt,mapping = aes(x=date,y=total_cases))+geom_line(mapping = aes(color=rankin
g))
```

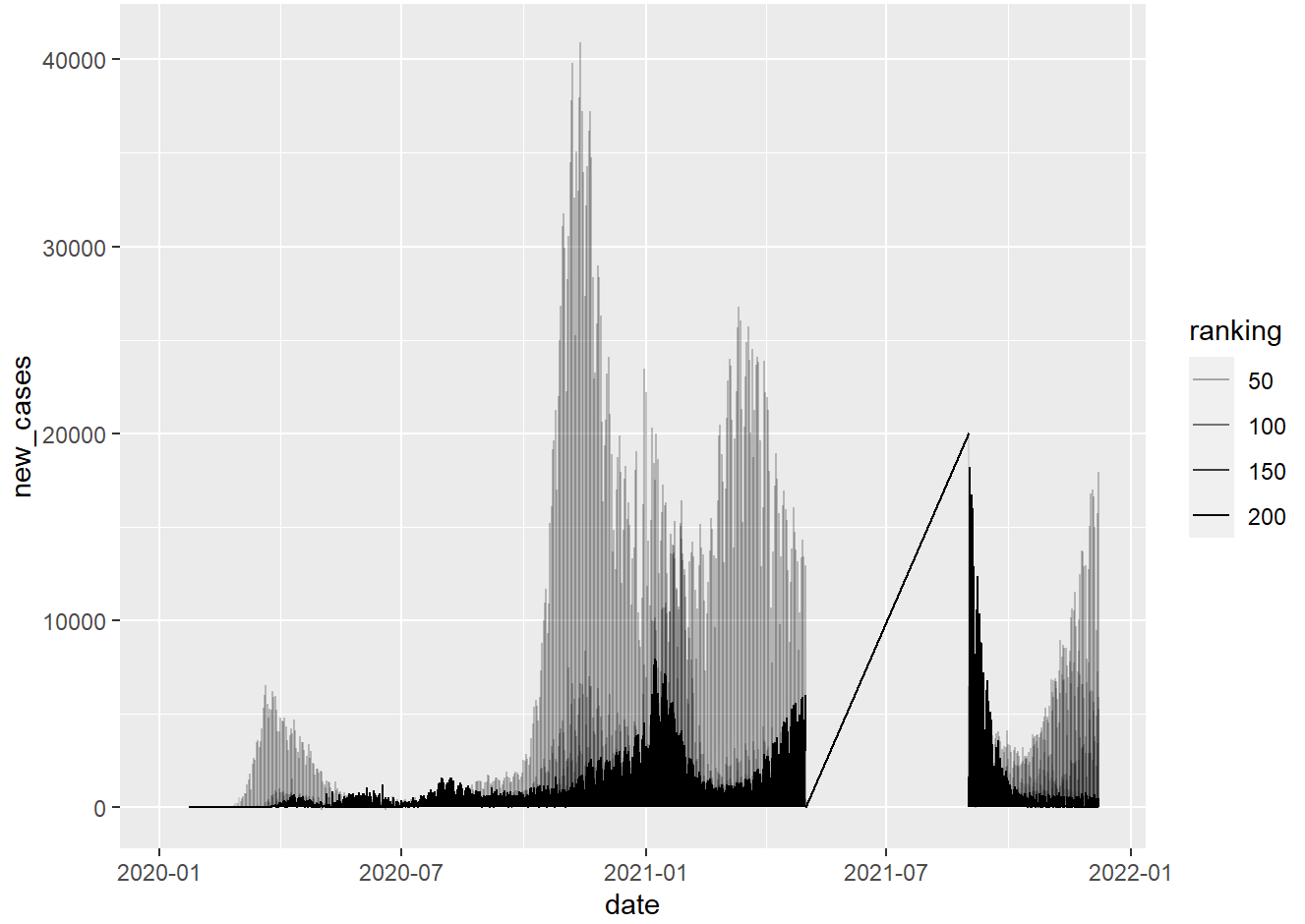



```
ggplot(covid_pubdebt,mapping = aes(x=date,y=total_cases_per_million))+geom_line(mapping = aes(co
lor=ranking))
```

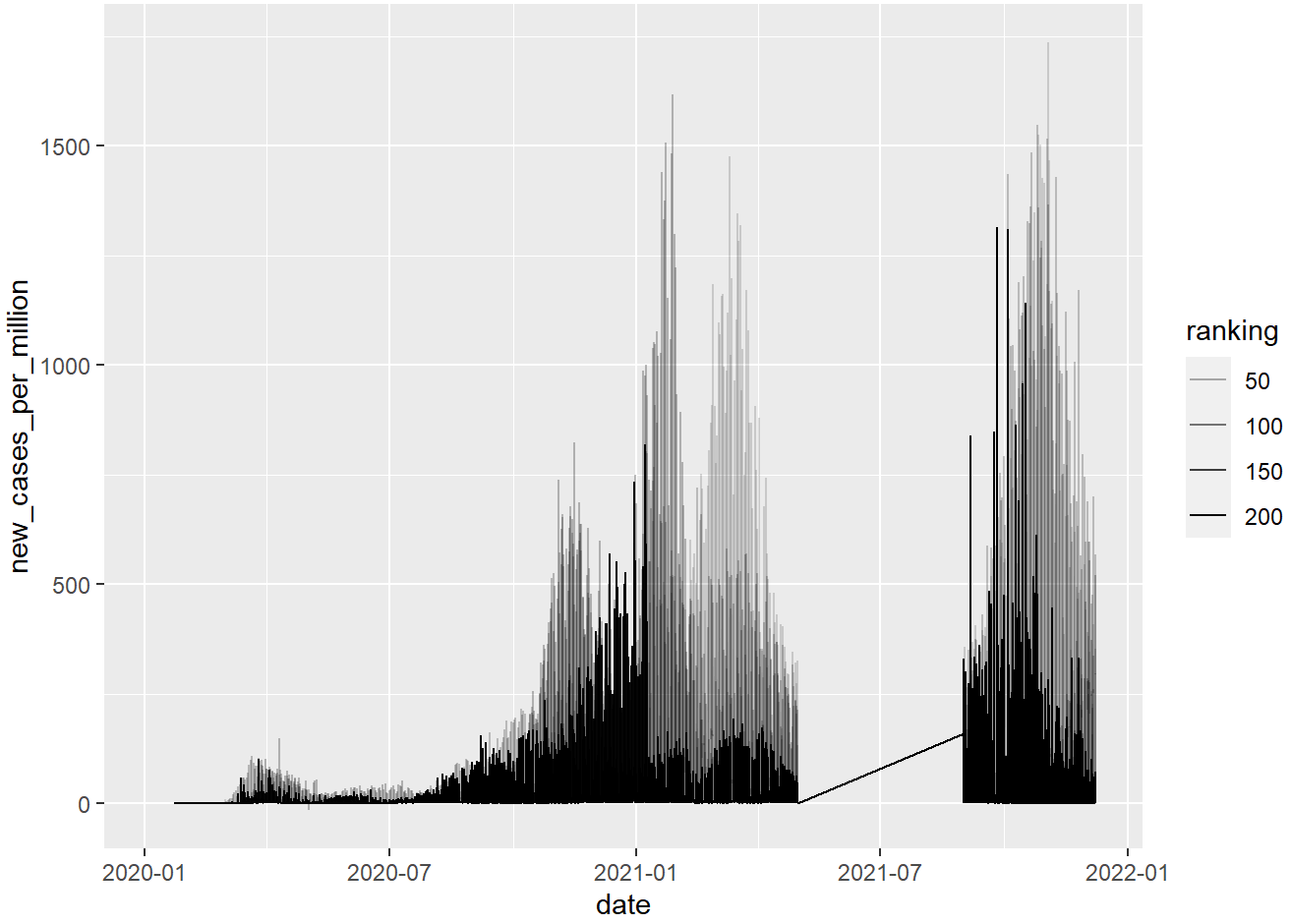


These graphs focus on the total cases in the top 10 and bottom 9 countries.From the graph of total cases, it is obvious that top 10 country has relatively high total cases with the time increases. Moreover, after Sept 01, bottom 9 countries have almost no increasing in total cases, while top 10 countries still illustrate a upward trend. To avoid the population influence on the analysis, the total cases per million is used to construct the graph again. Although the total cases per million is large for bottom 9 countries, the characteristic found in the first graph is still worked. The pattern is worked on all of the country.

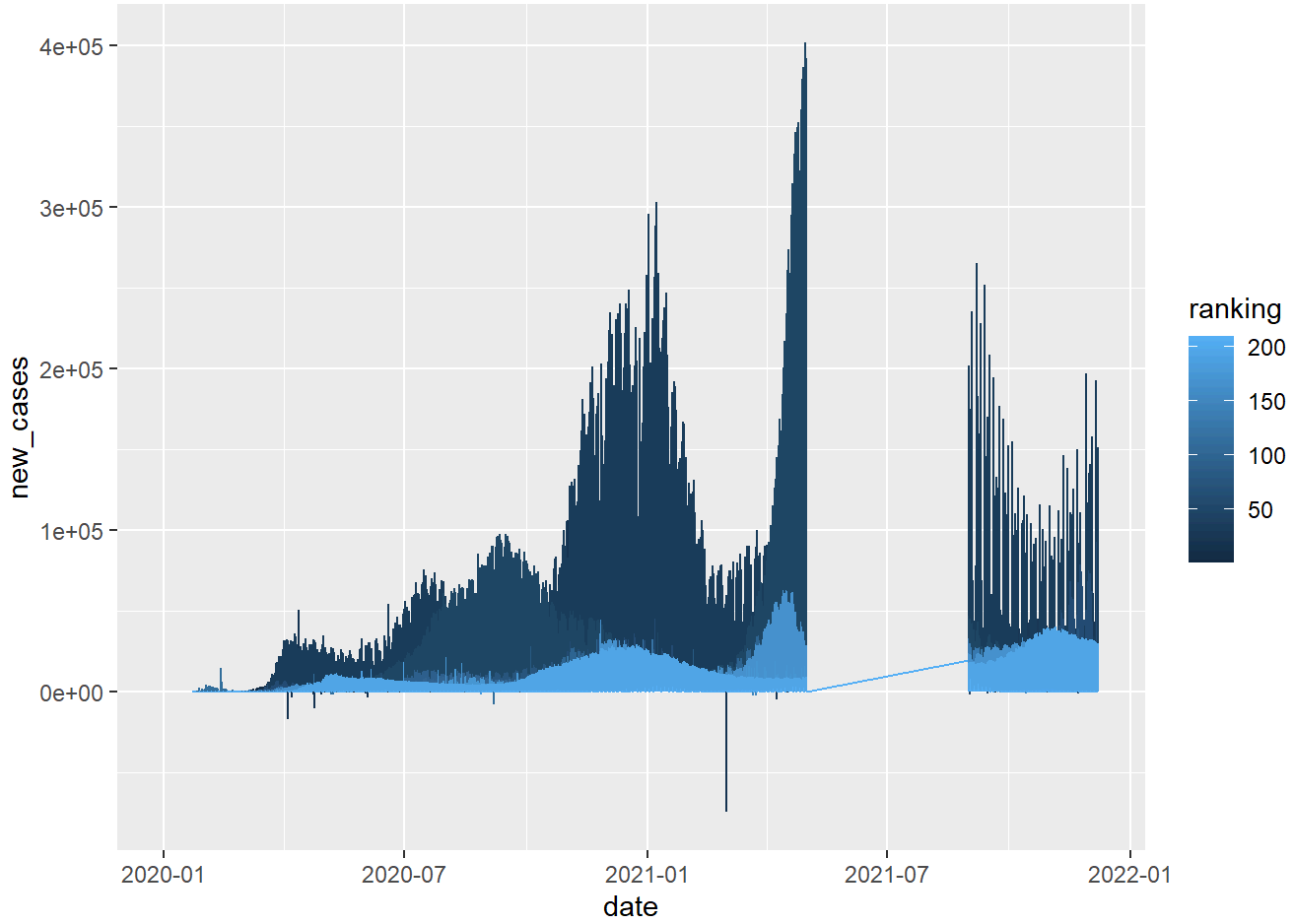
```
ggplot(pubdebt_top10,mapping = aes(x=date,y=new_cases))+geom_line(mapping = aes(alpha=ranking))
```



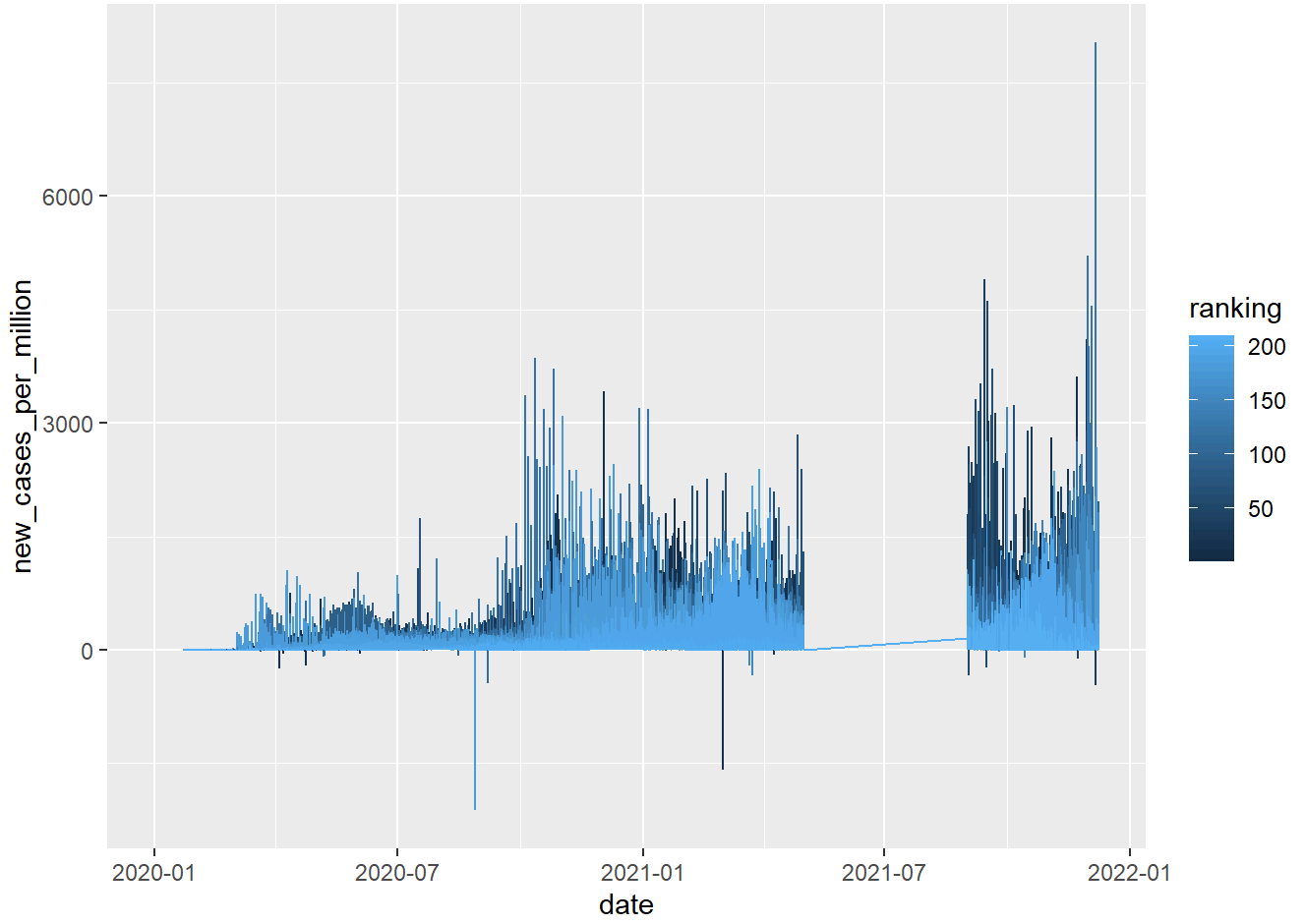

```
ggplot(pubdebt_top10,mapping = aes(x=date,y=new_cases_per_million))+geom_line(mapping = aes(alph
a=ranking))
```




```
ggplot(covid_pubdebt,mapping = aes(x=date,y=new_cases))+geom_line(mapping = aes(color=ranking))
```

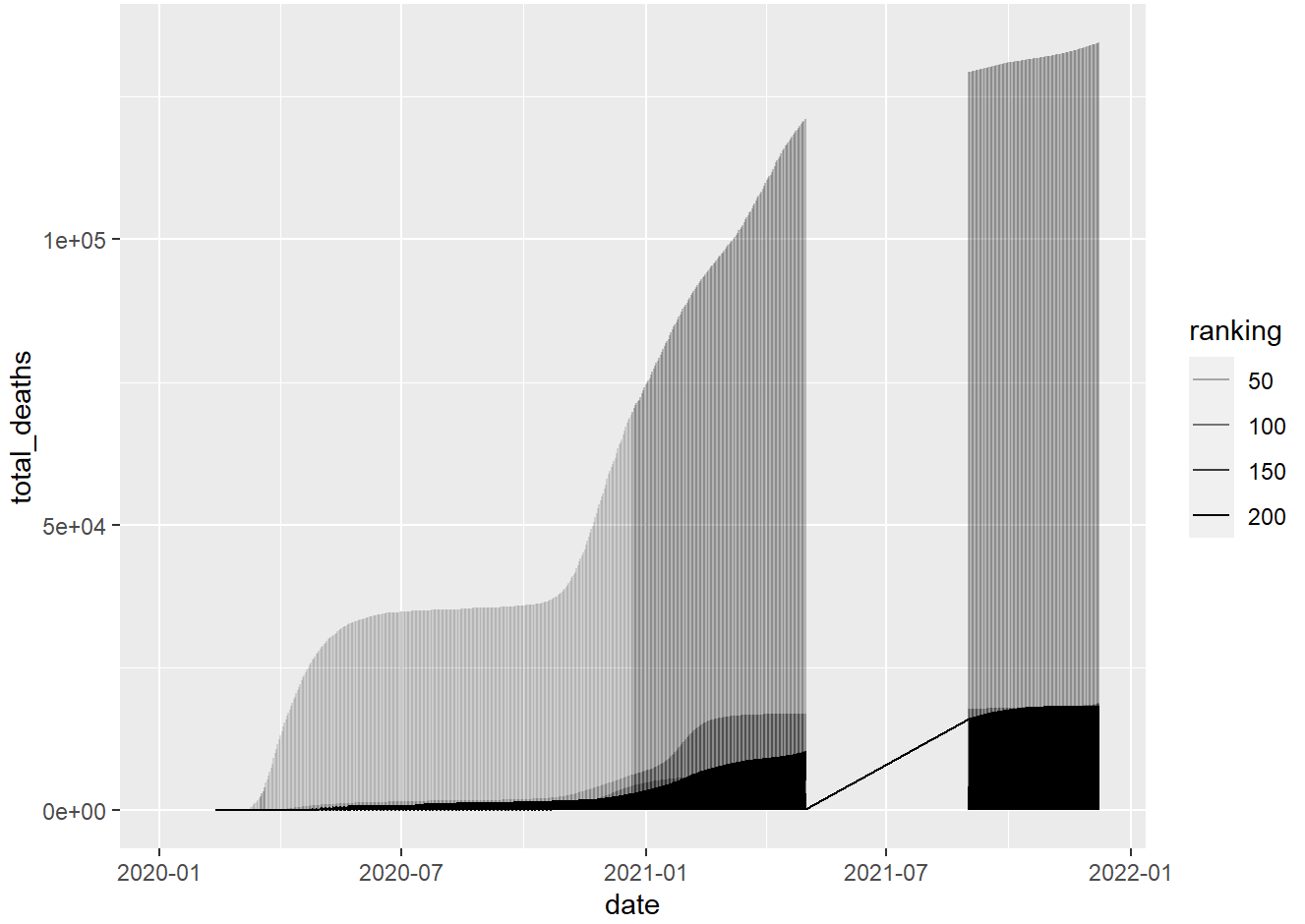



```
ggplot(covid_pubdebt,mapping = aes(x=date,y=new_cases_per_million))+geom_line(mapping = aes(colo
r=ranking))
```

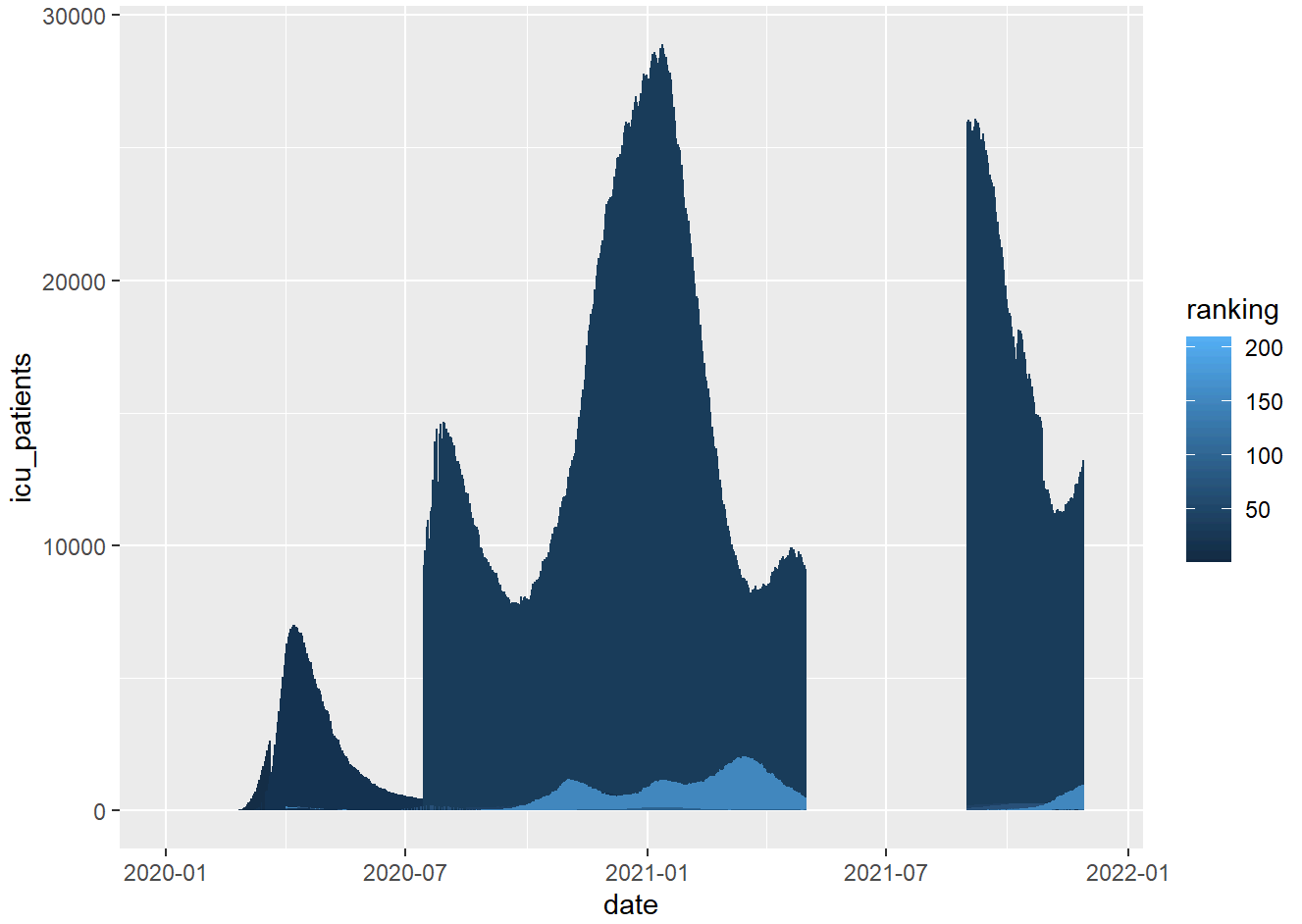


The graphs related to the new cases are more violated than the total cases. Although, top 10 countries' new cases are more than bottom 9 countries. Furthermore, top 10 countries has more peaks than bottom 9 countries before May 01, and after Sept 01, the new cases in bottom 9 countries has dramatic decrease compare to an increase in to 10 countries. Besides, before May 01, the distribution of new cases per million for the bottom 9 countries is more right skewed that the peak of the bottom 9 countries' new cases per million is earlier than top 10 countries'. It also works after Sept 01.

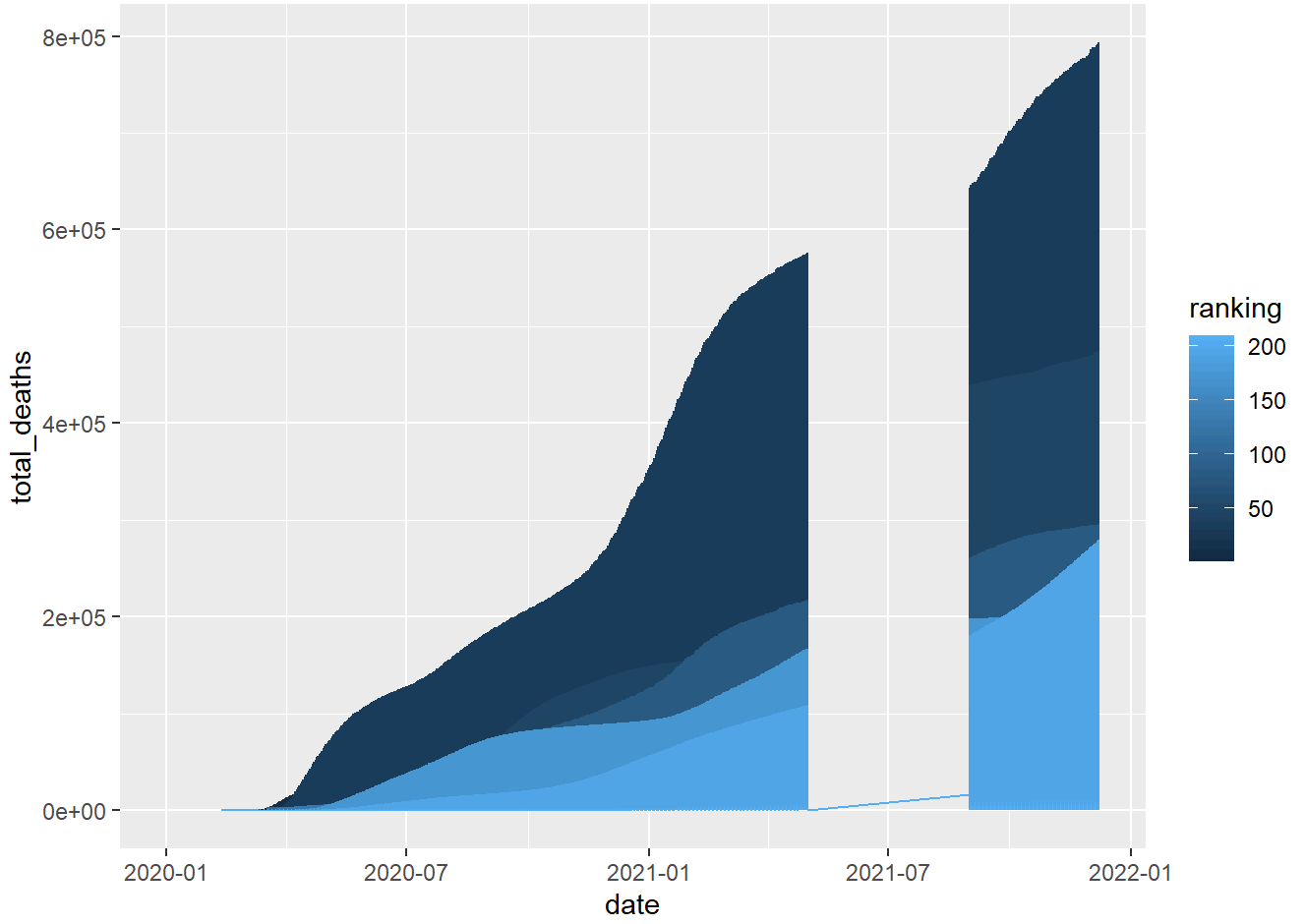
```
ggplot(pubdebt_top10,mapping = aes(x=date,y=total_deaths))+geom_line(mapping = aes(alpha=rankin
g))
```



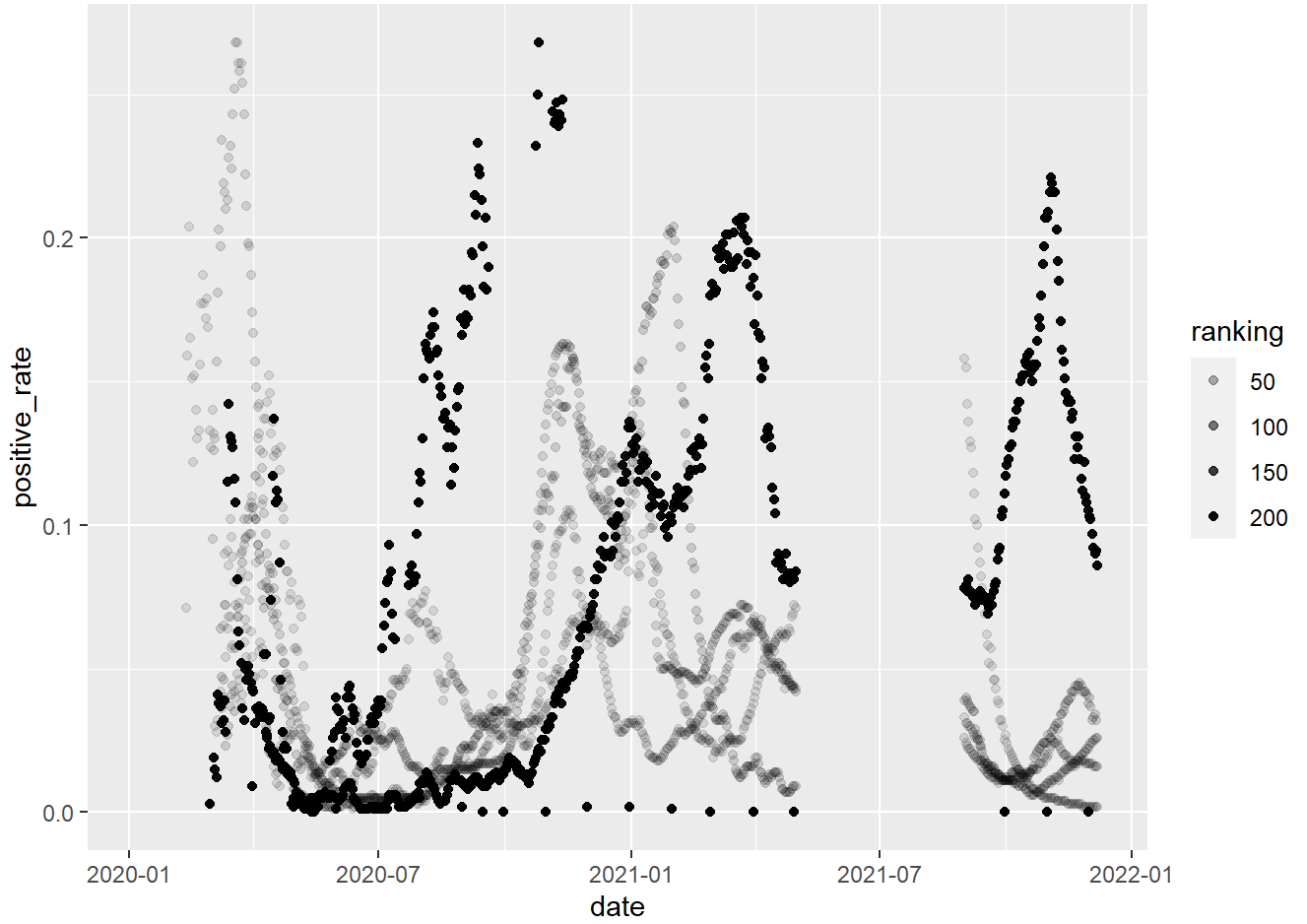

```
ggplot(covid_pubdebt,mapping = aes(x=date,y=icu_patients))+geom_line(mapping = aes(color=rankin
g))
```




```
ggplot(covid_pubdebt,mapping = aes(x=date,y=total_deaths))+geom_line(mapping = aes(color=rankin
g))
```

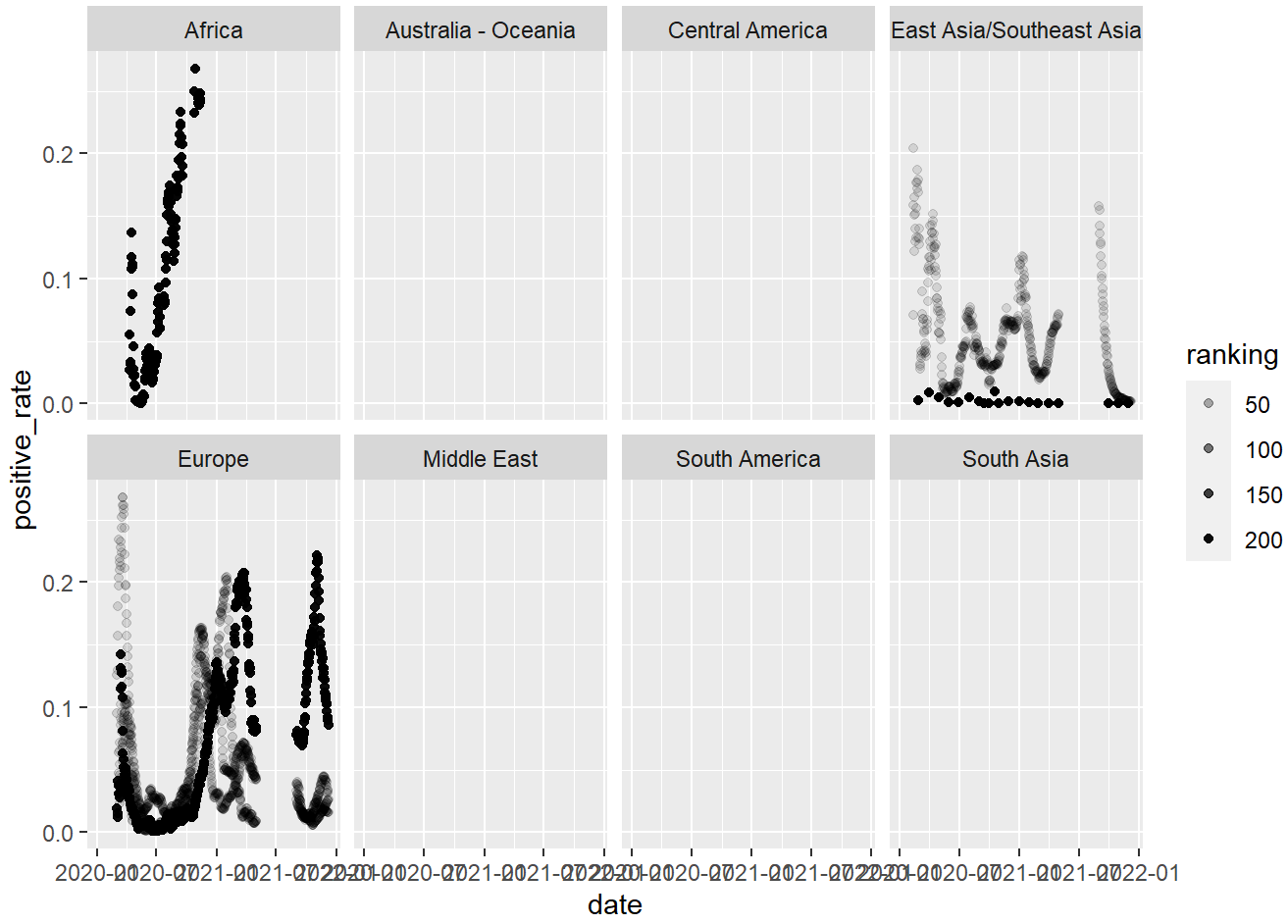



```
ggplot(pubdebt_top10,mapping = aes(x=date,y=positive_rate))+geom_point(mapping = aes(alpha=ranki
ng))
```

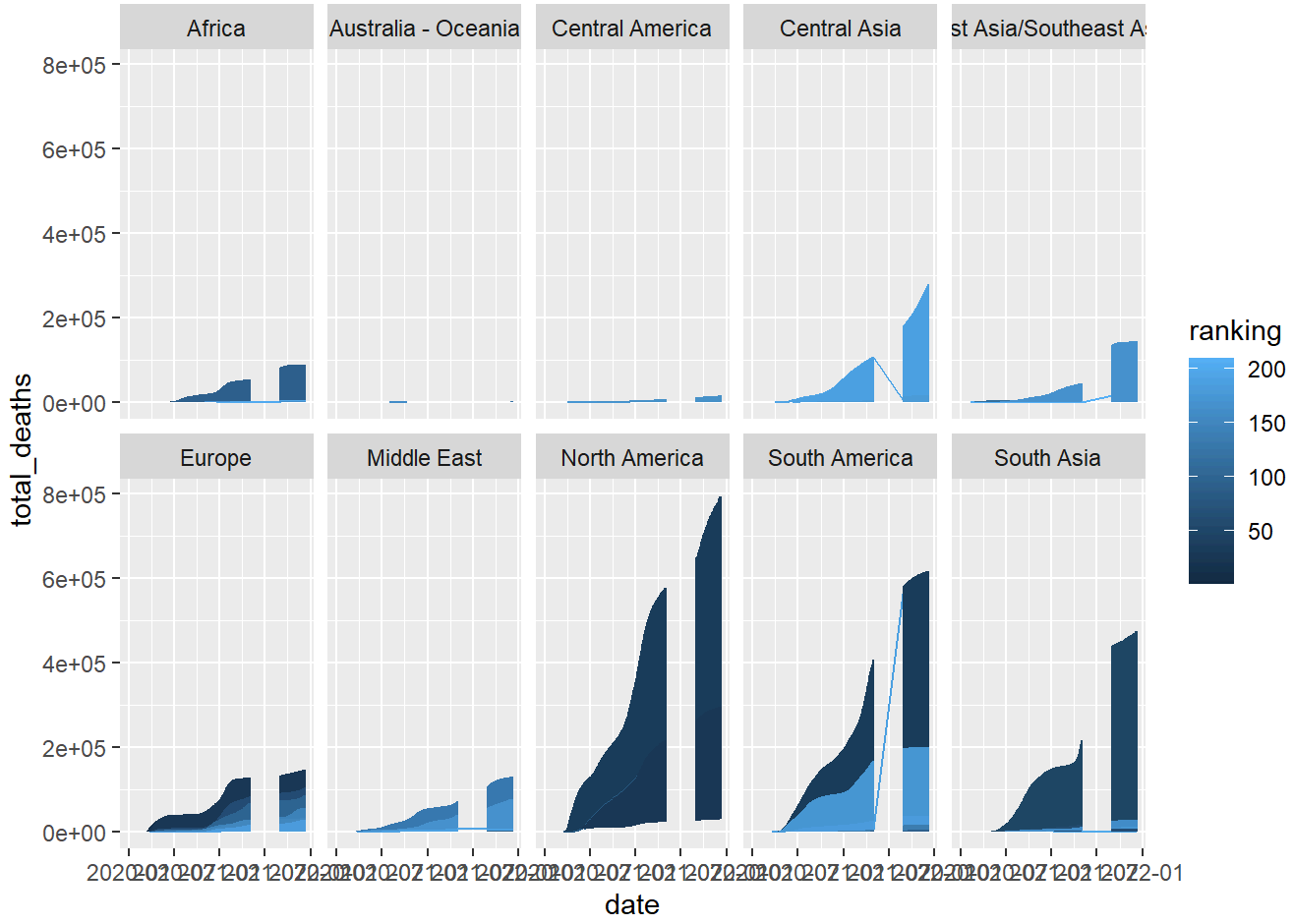


Looking at the total deaths, top 10 countries have far more deaths than bottom 9 countries. Same for the ICU patients, higher ranking countries have more peaks and larger number than lower ranking countries. However, the positive rate seems having different result. It might be caused by the region.

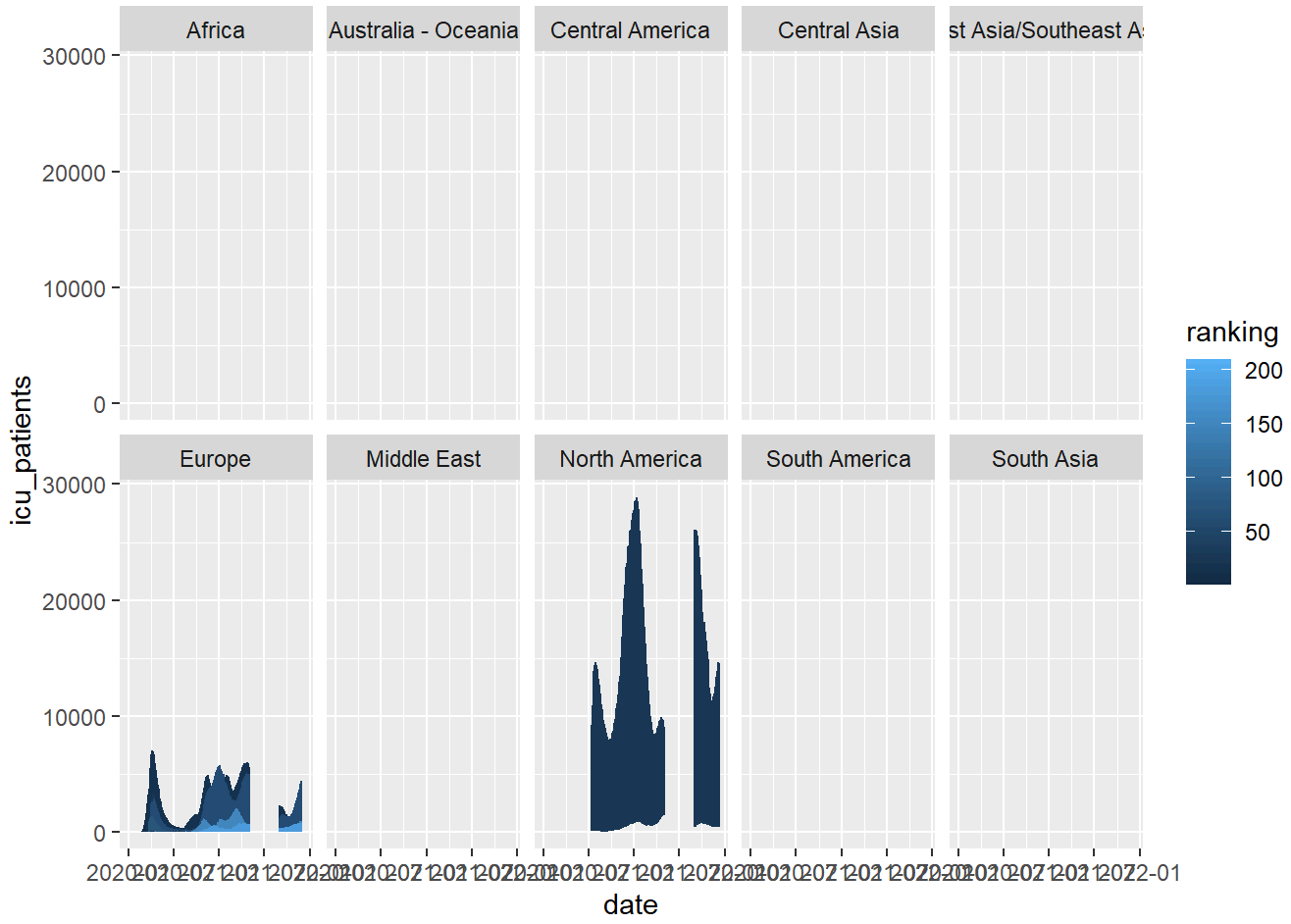
```
ggplot(pubdebt_top10,mapping = aes(x=date,y=positive_rate))+geom_point(mapping = aes(alpha=ranki
ng))+
  facet_wrap(~region, nrow = 2)
```



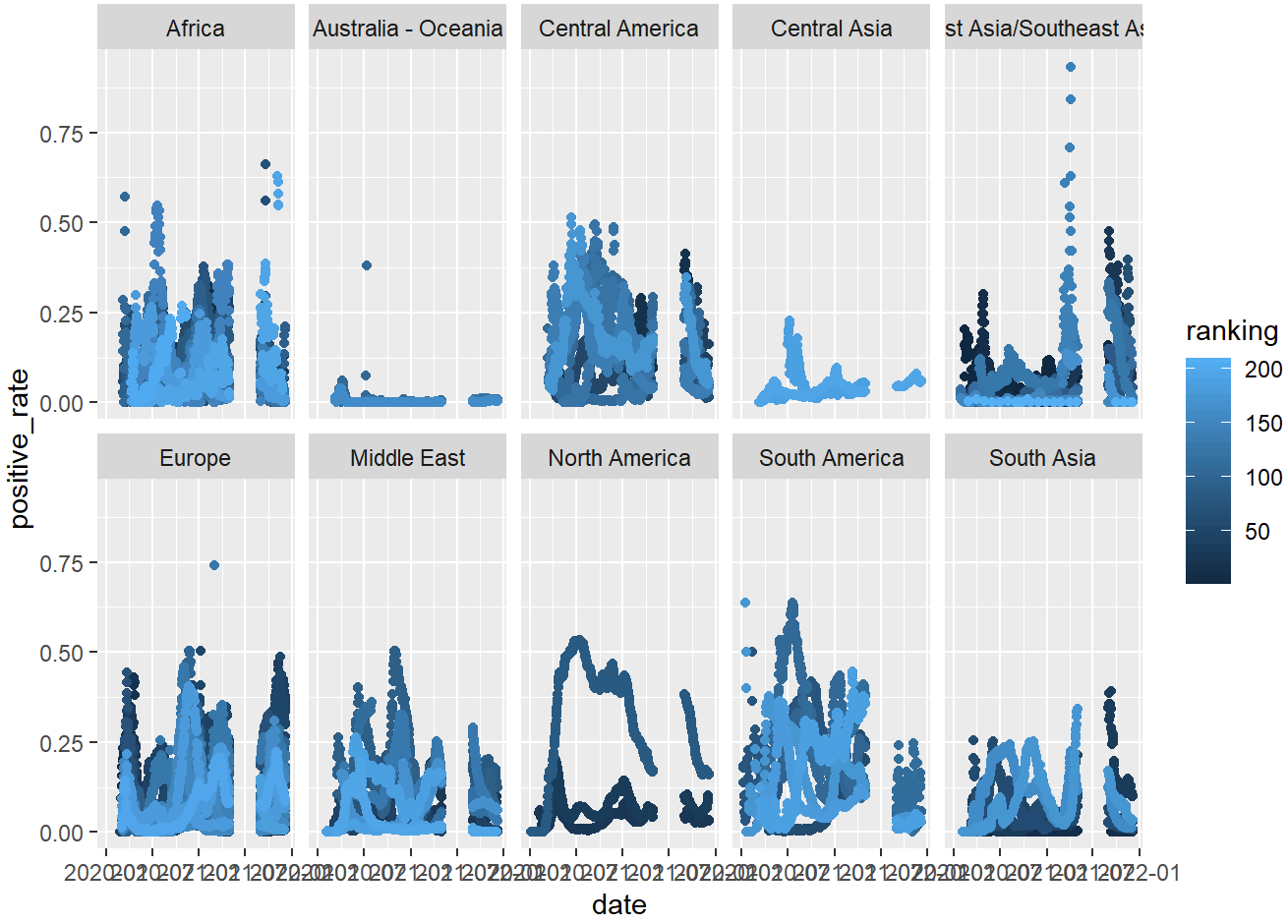

```
ggplot(covid_pubdebt,mapping = aes(x=date,y=total_deaths))+geom_line(mapping = aes(color=rankin
g))+
  facet_wrap(~region, nrow = 2)
```



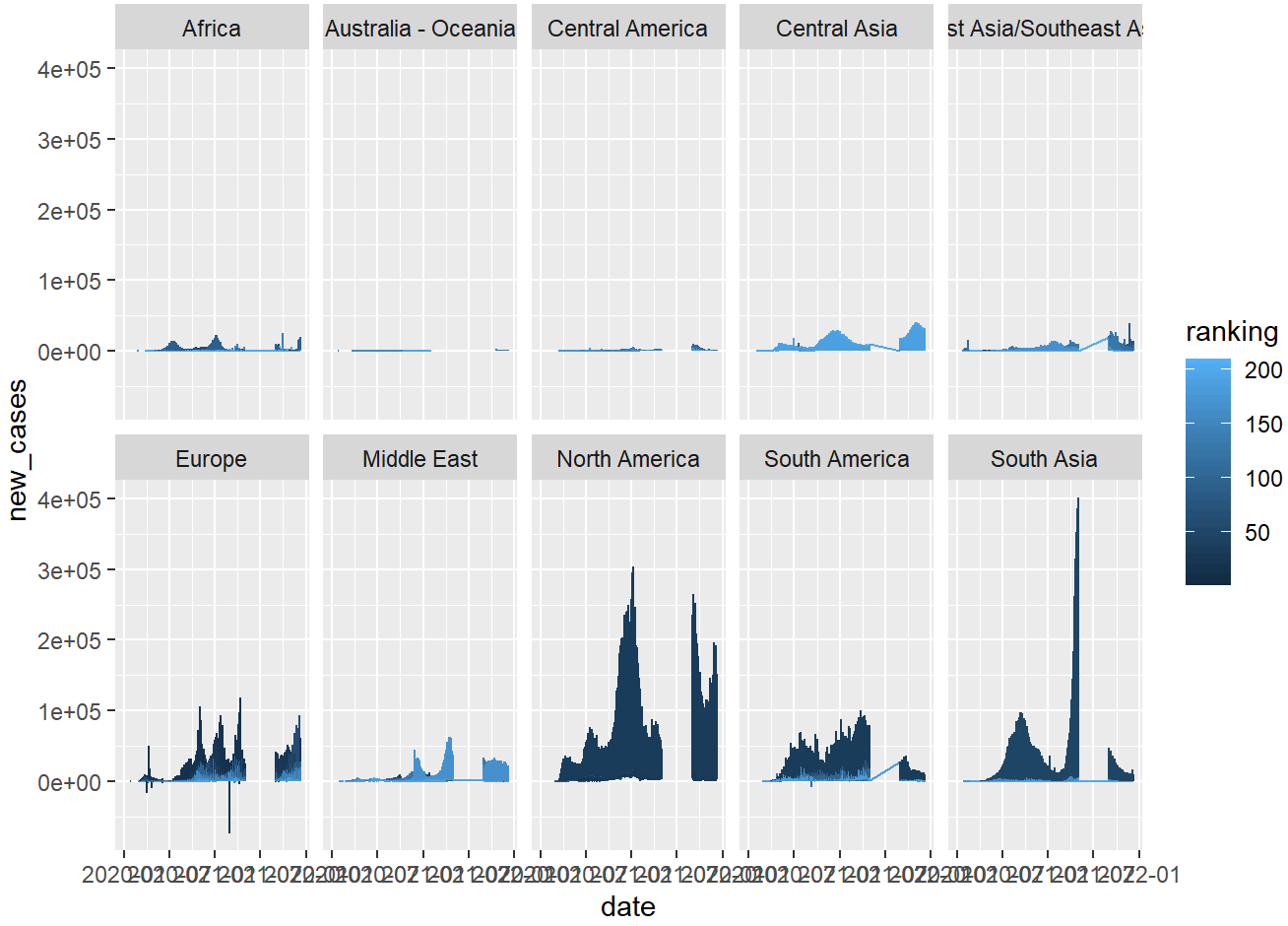

```
ggplot(covid_pubdebt,mapping = aes(x=date,y=icu_patients))+geom_line(mapping = aes(color=rankin
g))+
  facet_wrap(~region, nrow = 2)
```



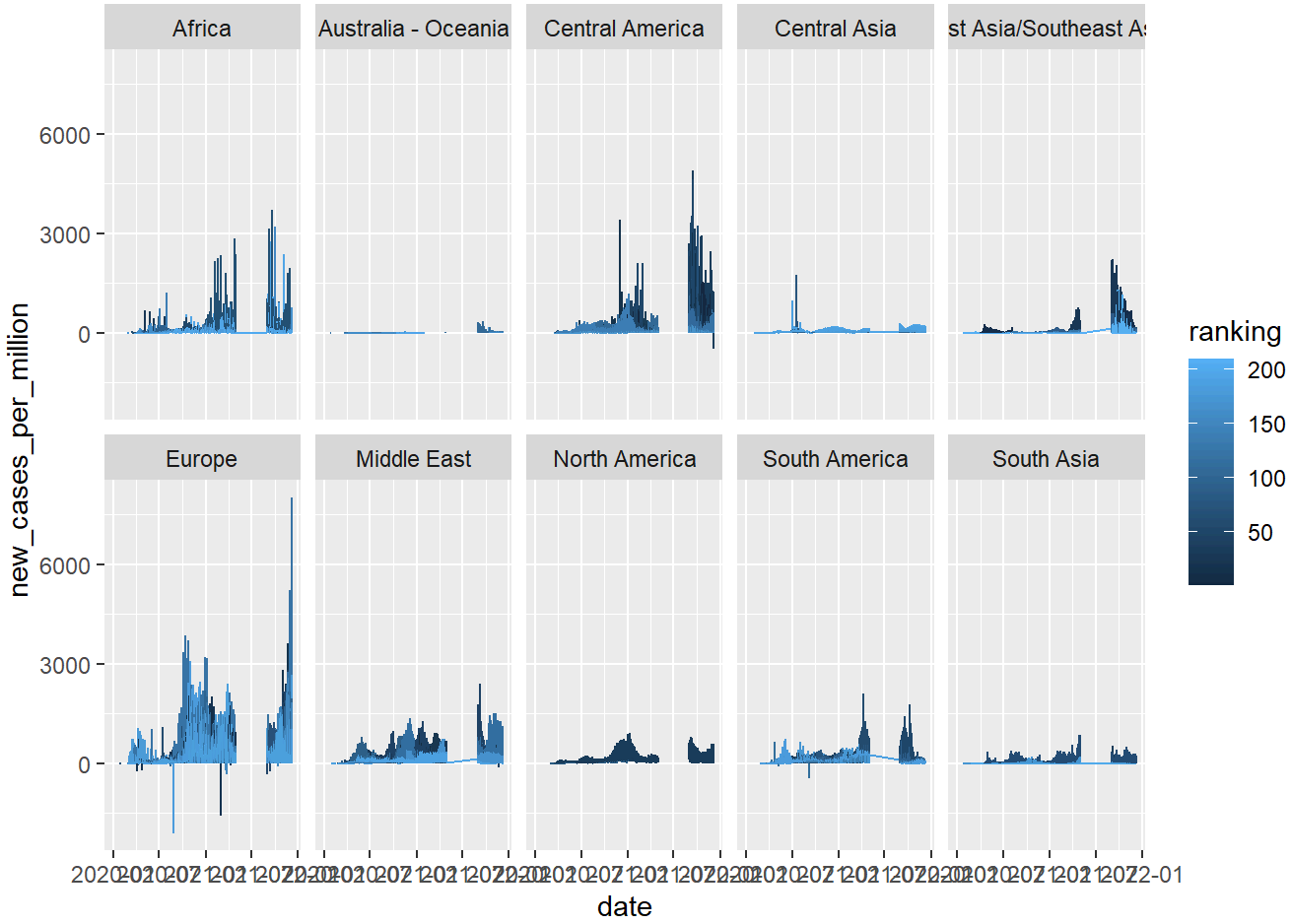

```
ggplot(covid_pubdebt,mapping = aes(x=date,y=positive_rate))+geom_point(mapping = aes(color=ranki
ng))+
  facet_wrap(~region, nrow = 2)
```




```
ggplot(covid_pubdebt,mapping = aes(x=date,y=new_cases))+geom_line(mapping = aes(color=ranking))+
  facet_wrap(~region, nrow = 2)
```




```
ggplot(covid_pubdebt,mapping = aes(x=date,y=new_cases_per_million))+geom_line(mapping = aes(colo
r=ranking))+
  facet_wrap(~region, nrow = 2)
```



From the first graph, Africa might have low public debt and high positive rate, while East Asia and Europe have high positive rate with high ranking. To analyze the influence from the region, the next five graphs are based on region. European seems been affect by the public debt. North America seems get weak affect by the public debt. When comparing the response variable, the new cases per million is the most clear factor that higher public debt would gets more new cases per million. From the previous analysis, the public debt has impact on the COVID 19 performance of a country. However, the impact might interact or influenced by other factors such as the wealthy of the country, the democracy of the country, or the region of the country.

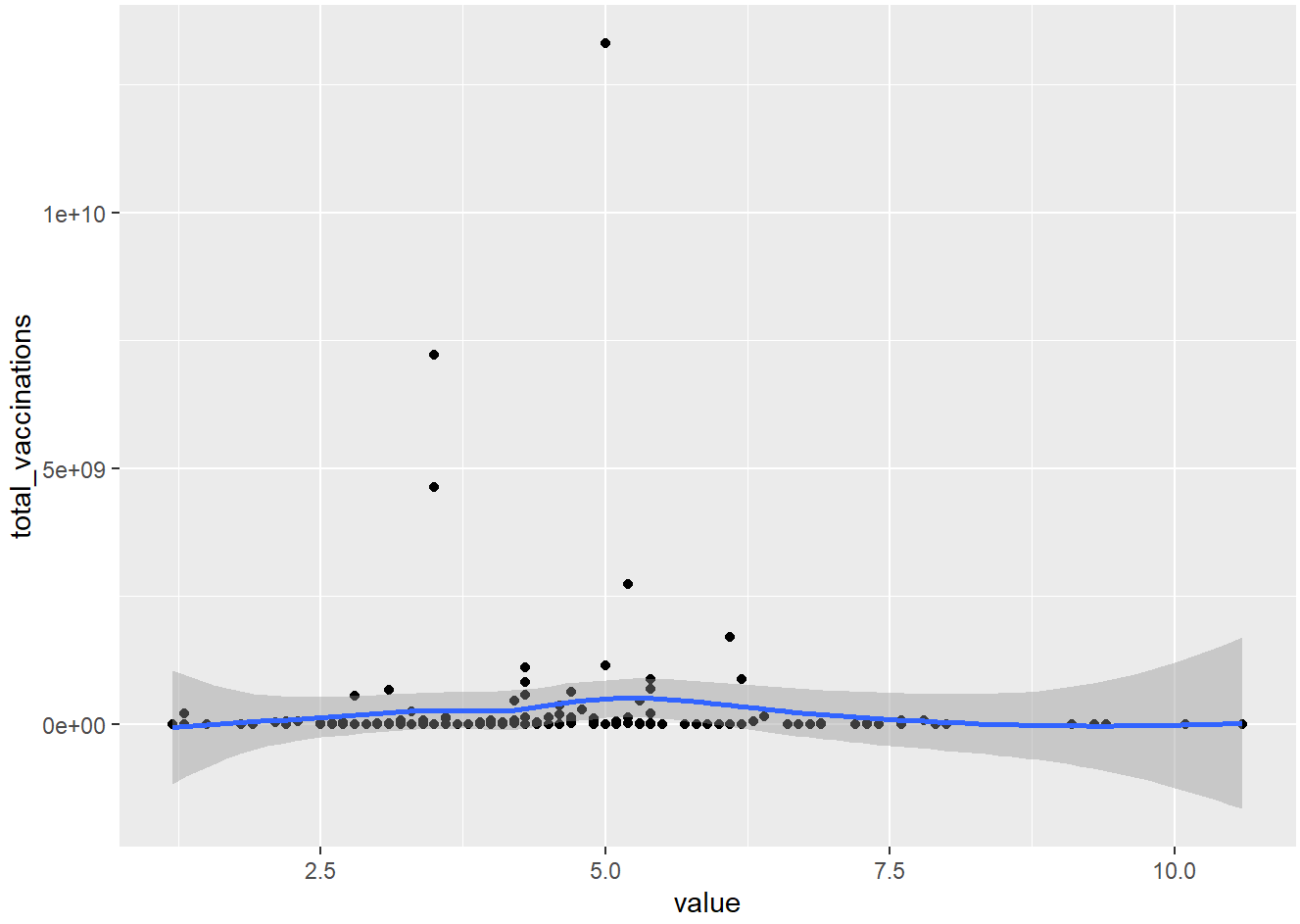
Education expenditure and vaccination

Higher education expenditure always means more attention from government to the education industry. It might address more high-educated people and higher average education level. Furthermore, better educated people may have more knowledge and acceptance about vaccination. In this way, the higher education expenditure might causes the vaccination rate higher. The clarification of the relationship between education expenditure and COVID-19 vaccination could describe a new way to prohibit the spread of pandemic or encourage people to take regular vaccine in normal life. As the COVID vaccine been created, many people criticize the safety of the vaccination. Some of them fear the vaccination. It is interesting that some of those people who fear vaccination may not provide convinced information to prove their points. We assume that they might be lack of the knowledge about the vaccination that caused them refused vaccination. With high education expenditure, it would be considered as high educated country or territories that would have high vaccination rate. The time period are still before May 01 and After Sept 01. The choice based on the random select.

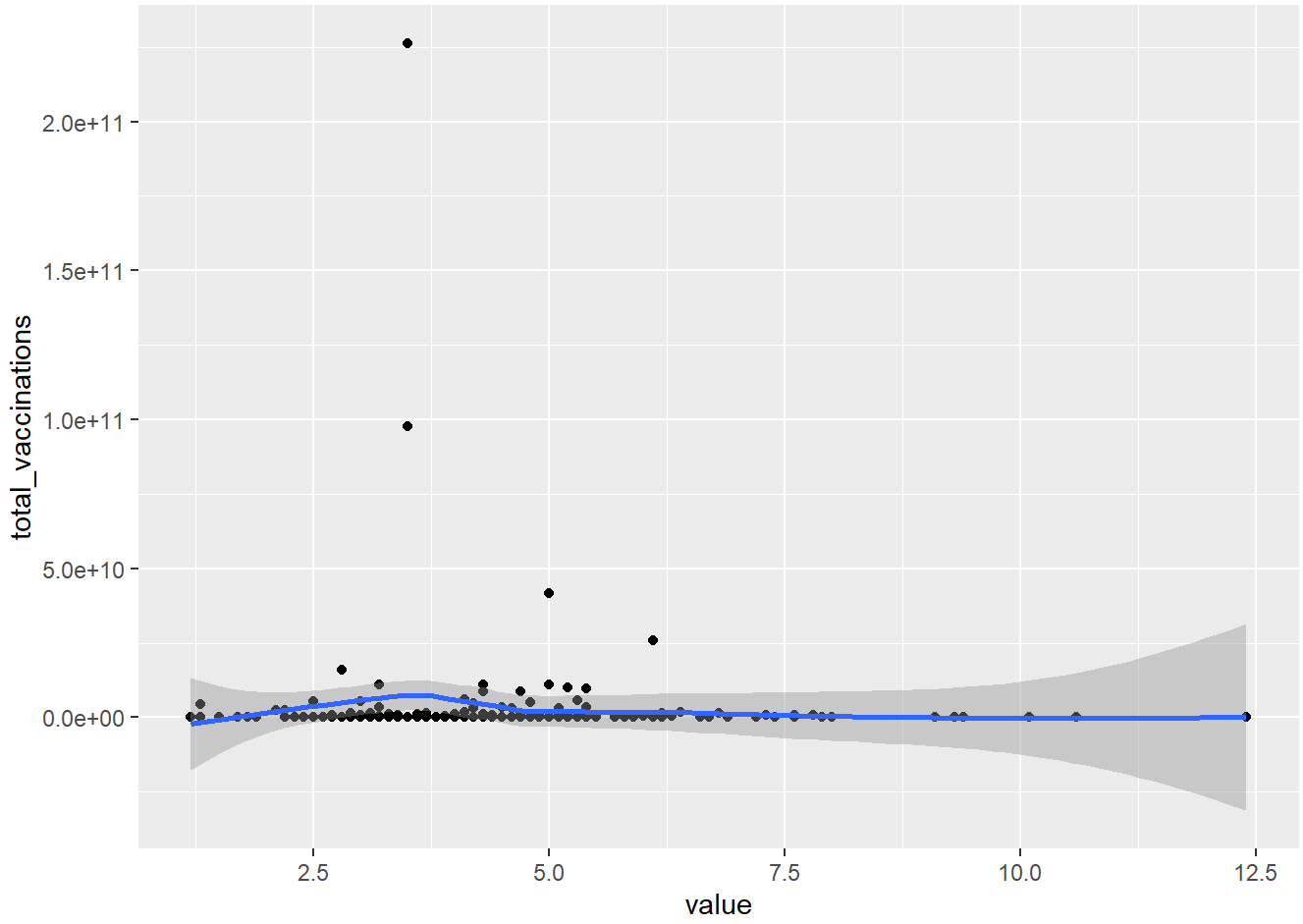
```
covid_edu<-left_join(educationexp, covid_vaccs, by = c("name" = "location"))%>%filter(date<=ymd
("2021-05-01")|date>=ymd("2021-09-01"))
covid_edu_before<-covid_edu%>%filter(date<=ymd("2021-05-01"))%>%select(name,value,region,total_v
accinations,people_fully_vaccinated,people_vaccinated,total_vaccinations_per_hundred,people_full
y_vaccinated_per_hundred)%>%group_by(name,region)%>%summarise(total_vaccinations=sum(total_vacci
nations,na.rm = TRUE), people_fully_vaccinated=sum(people_fully_vaccinated,na.rm = TRUE),people_
vaccinated=sum(people_vaccinated,na.rm = TRUE), total_vaccinations_per_hundred=sum(total_vaccina
tions_per_hundred,na.rm = TRUE), people_fully_vaccinated_per_hundred=sum(people_fully_vaccinated
_per_hundred,na.rm = TRUE),value=mean(value))%>%arrange(value)
covid_edu_after<-covid_edu%>%filter(date>=ymd("2021-09-01"))%>%select(name,value,region,total_va
ccinations,people_fully_vaccinated,people_vaccinated,total_vaccinations_per_hundred,people_fully
_vaccinated_per_hundred)%>%group_by(name,region)%>%summarise(total_vaccinations=sum(total_vaccin
ations,na.rm = TRUE), people_fully_vaccinated=sum(people_fully_vaccinated,na.rm = TRUE),people_v
accinated=sum(people_vaccinated,na.rm = TRUE), total_vaccinations_per_hundred=sum(total_vaccinat
ions_per_hundred,na.rm = TRUE), people_fully_vaccinated_per_hundred=sum(people_fully_vaccinated_
per_hundred,na.rm = TRUE),value=mean(value))%>%arrange(value)
```

To analyze the relationship, we have pick: total_vaccinations, people_fully_vaccinated, people_vaccinated, total_vaccinations_per_hundred, and people_fully_vaccinated_per_hundred. The data includes all of the five variables' total before May 01 and after Sept 01.

```
ggplot(covid_edu_before,mapping = aes(x=value,y=total_vaccinations))+geom_point(mapping = aes())
+geom_smooth()
```

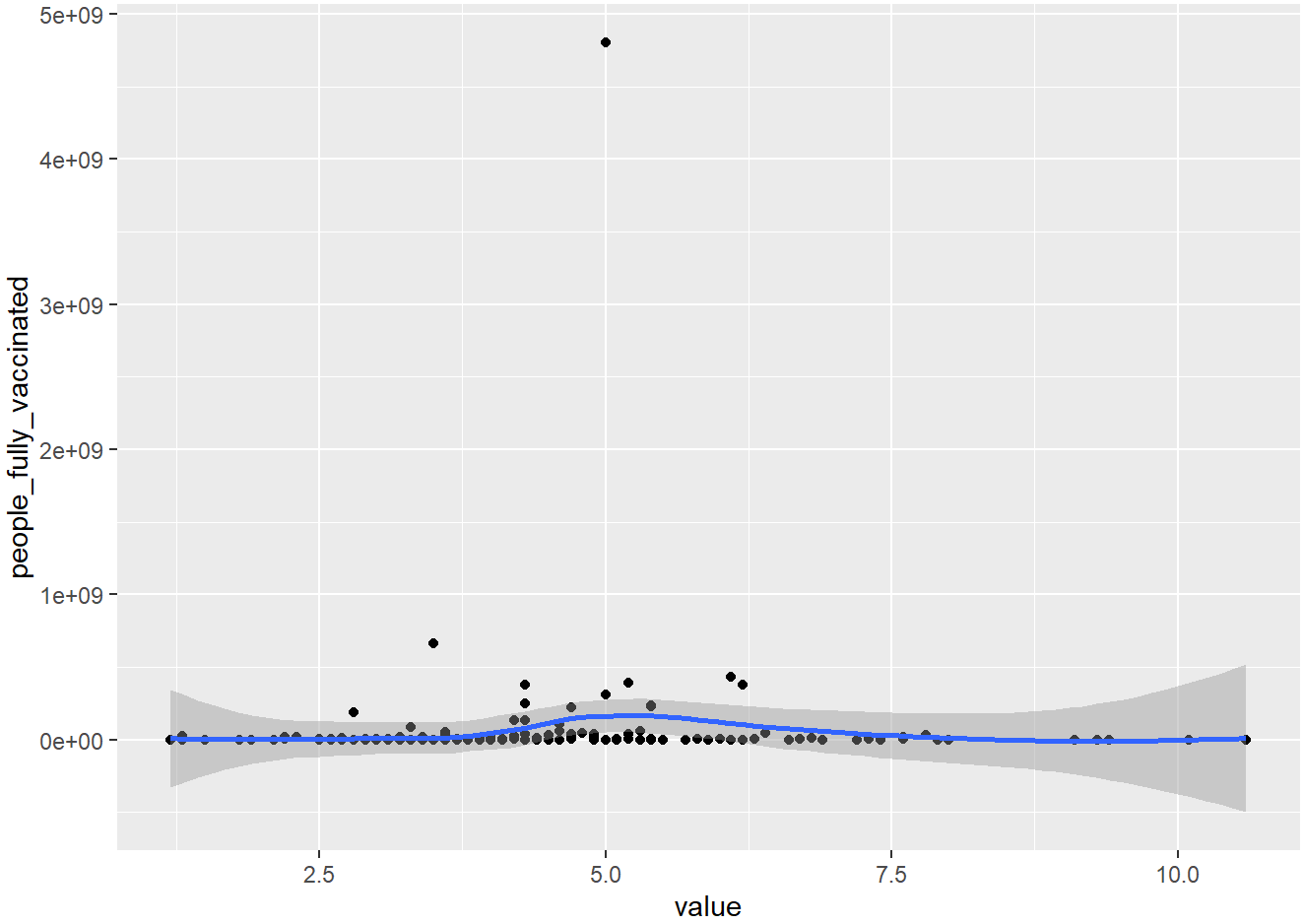


```
ggplot(covid_edu_after,mapping = aes(x=value,y=total_vaccinations))+geom_point()+geom_smooth()
```

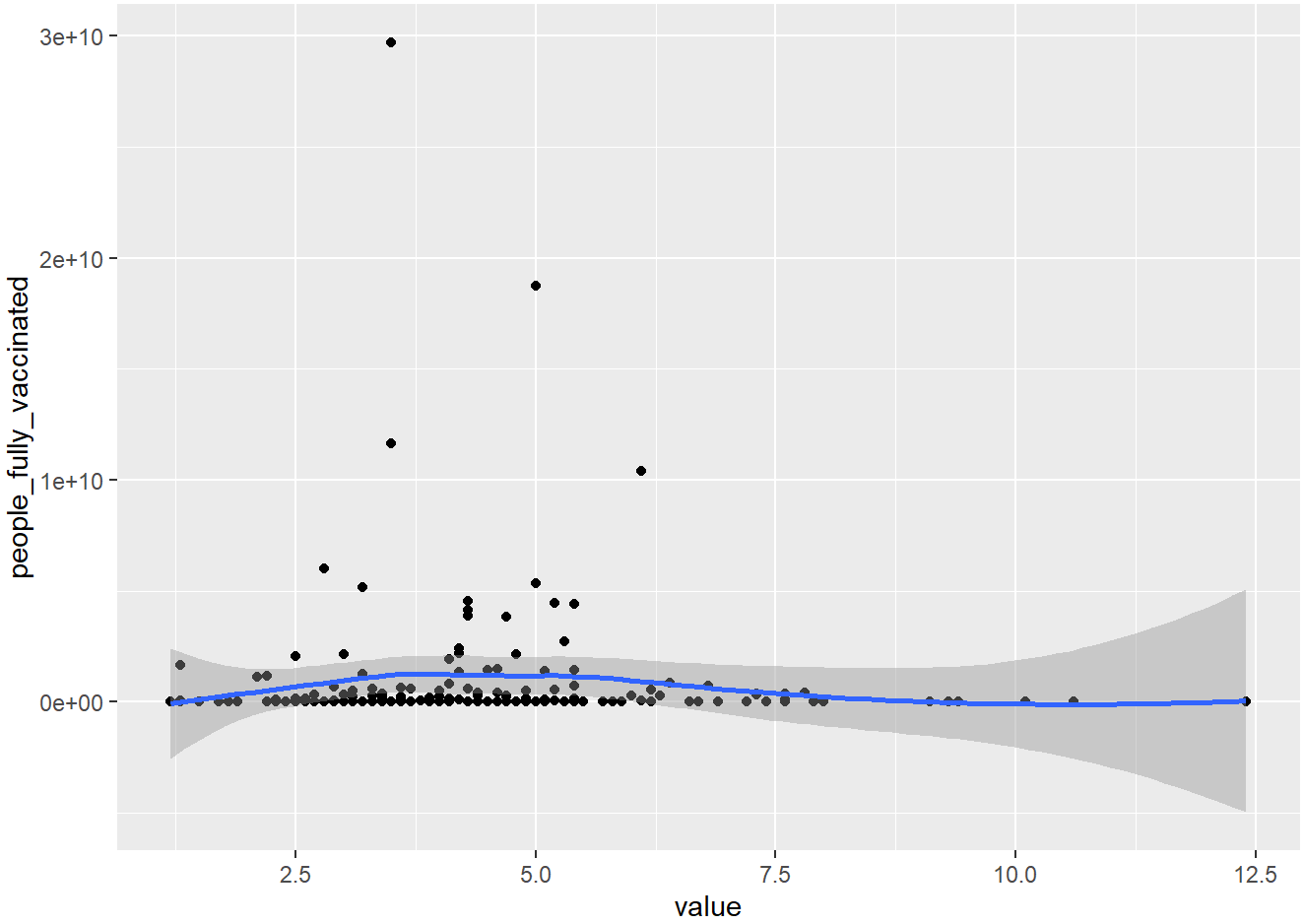


These graphs describe the relationship between total vaccination and the value of education expenditures. Before May 01, we can see small portion of country began vaccination that might cause the immature proceed of producing vaccine. The countries that have positive vaccine are close to the lower part of the education expenditures. Although the increase in the period after Sept 01 is still slight and at lower education expenditures, the increased country before May 01 has more education expenditure comparing to period after Sept 01.

```
ggplot(covid_edu_before,mapping = aes(x=value,y=people_fully_vaccinated))+geom_point()+geom_smoo
th()
```

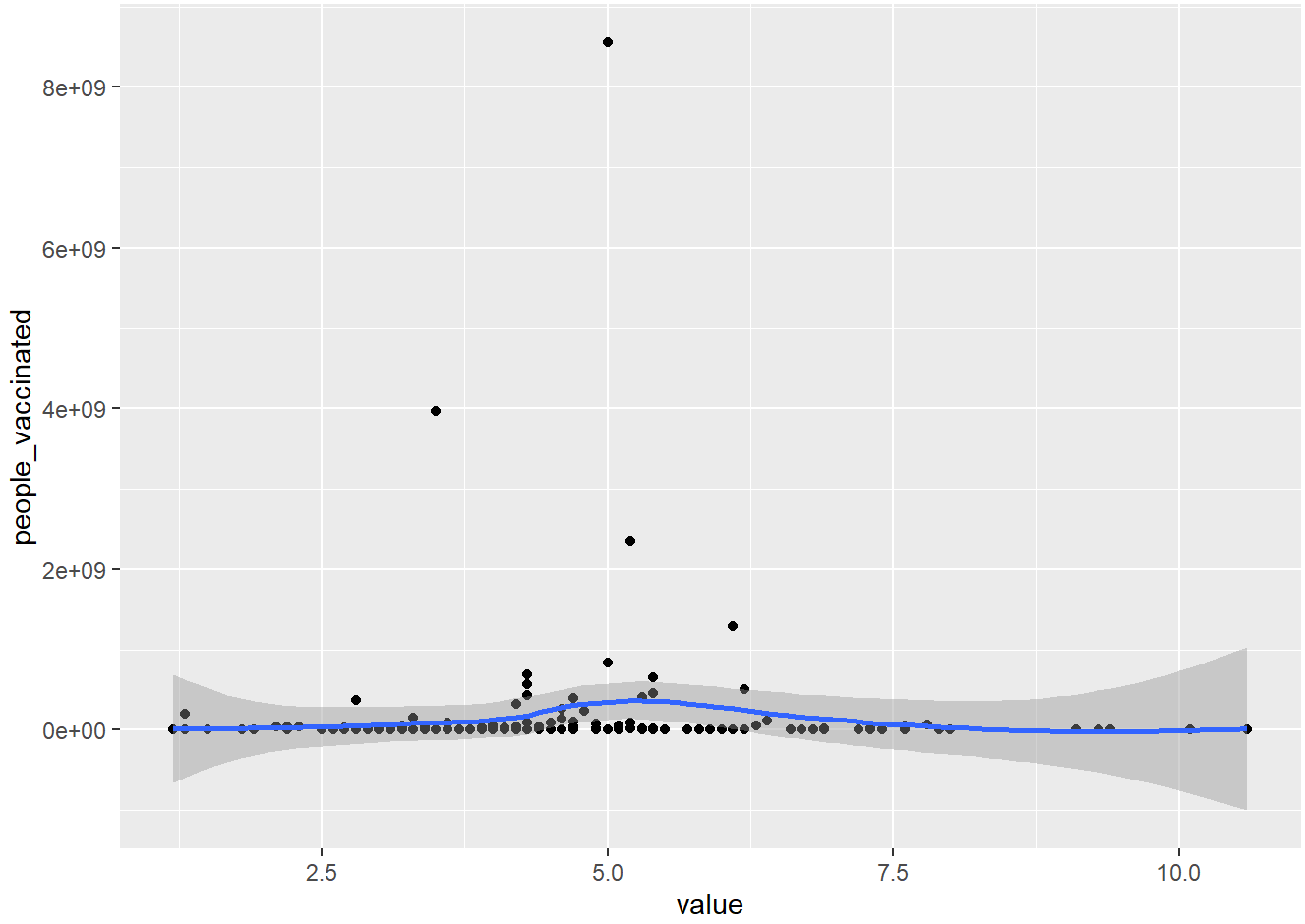



```
ggplot(covid_edu_after,mapping = aes(x=value,y=people_fully_vaccinated))+geom_point()+geom_smooth()
```

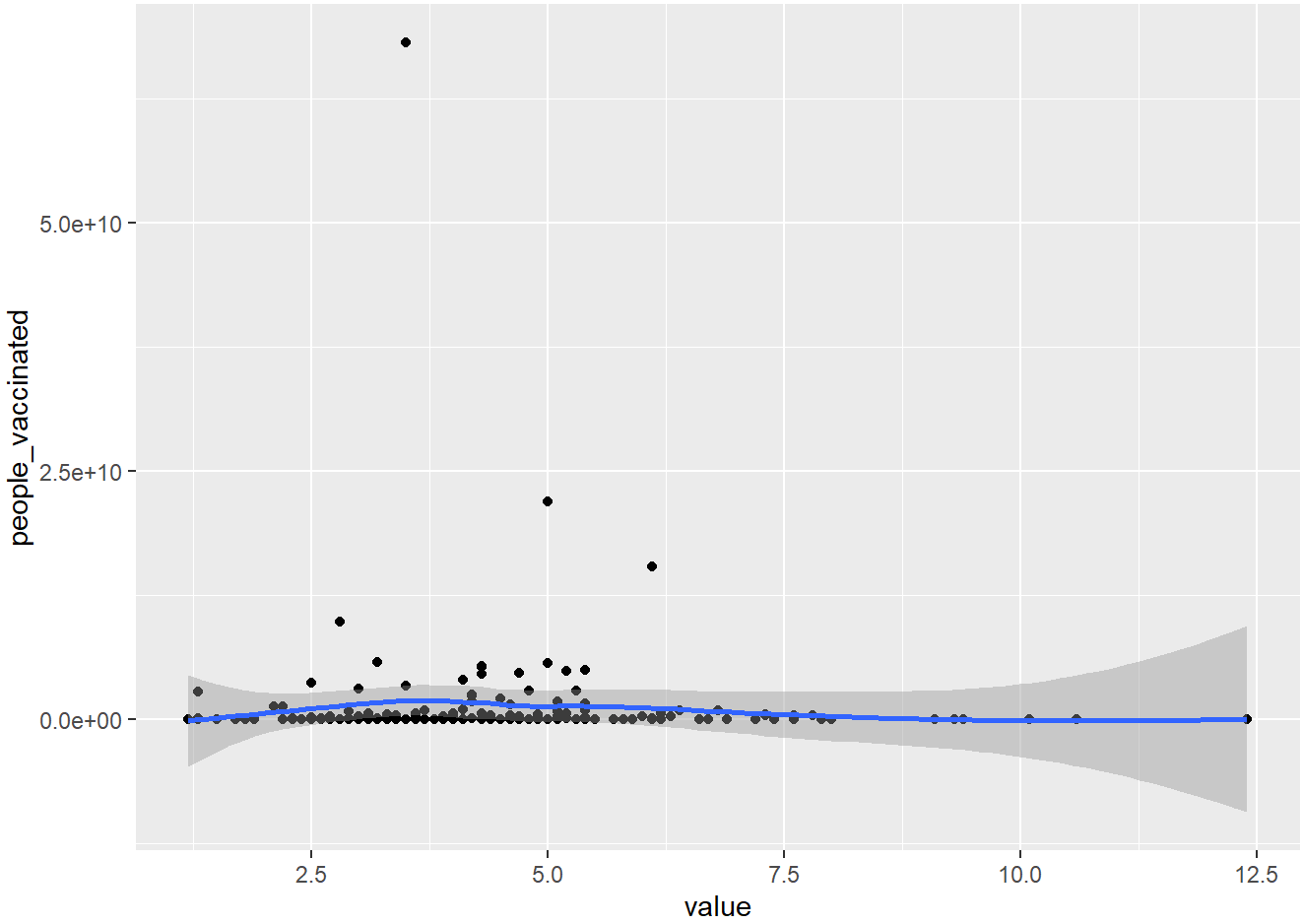


Comparing the people gets fully vaccinated, it is more clear that the countries which has relatively higher education expenditure would gets vaccination earlier than those who gets lower education expenditures. This illustrates that the education somehow influence the people who gets fully vaccinated.

```
ggplot(covid_edu_before,mapping = aes(x=value,y=people_vaccinated))+geom_point()+geom_smooth()
```

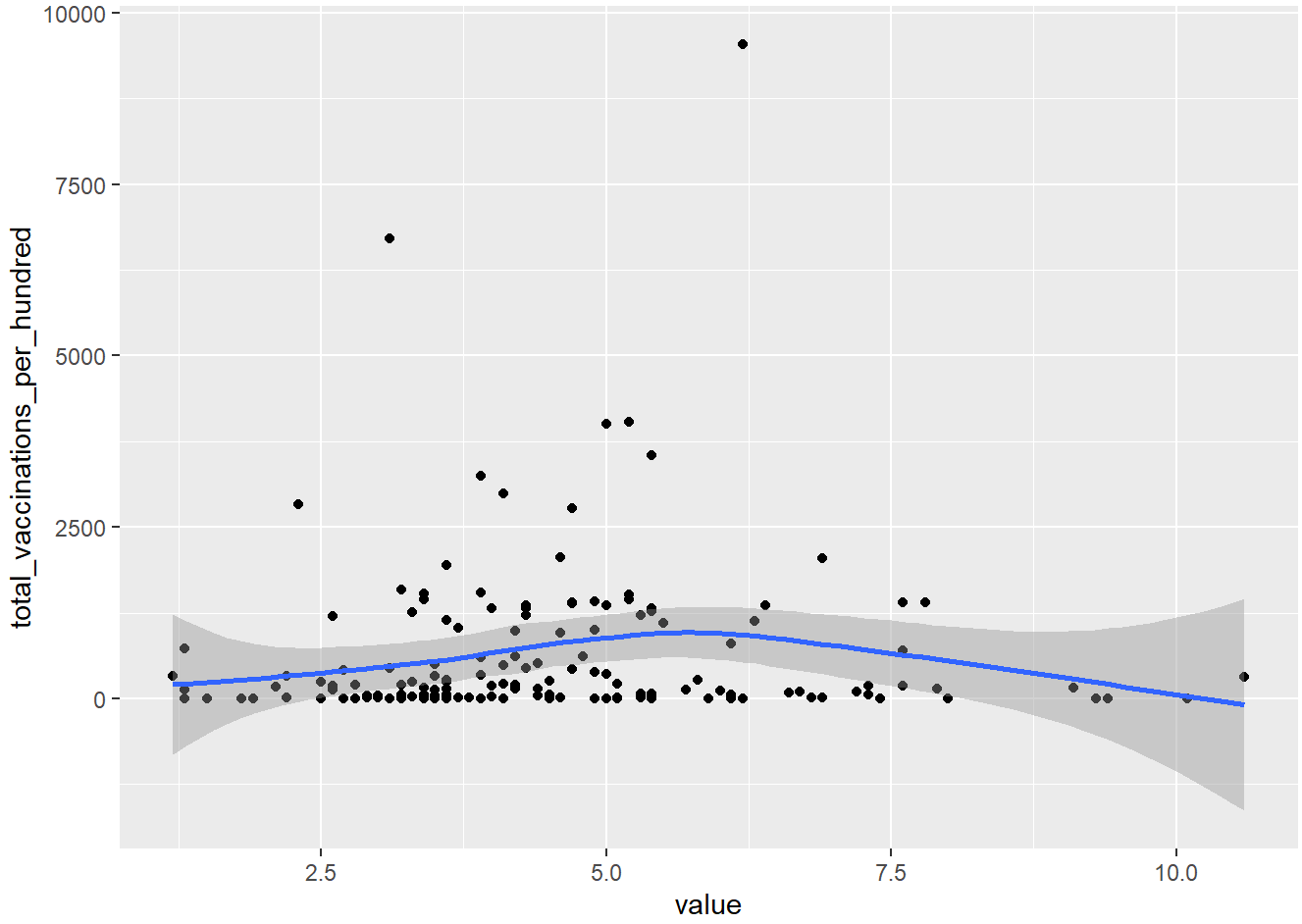



```
ggplot(covid_edu_after,mapping = aes(x=value,y=people_vaccinated))+geom_point()+geom_smooth()
```

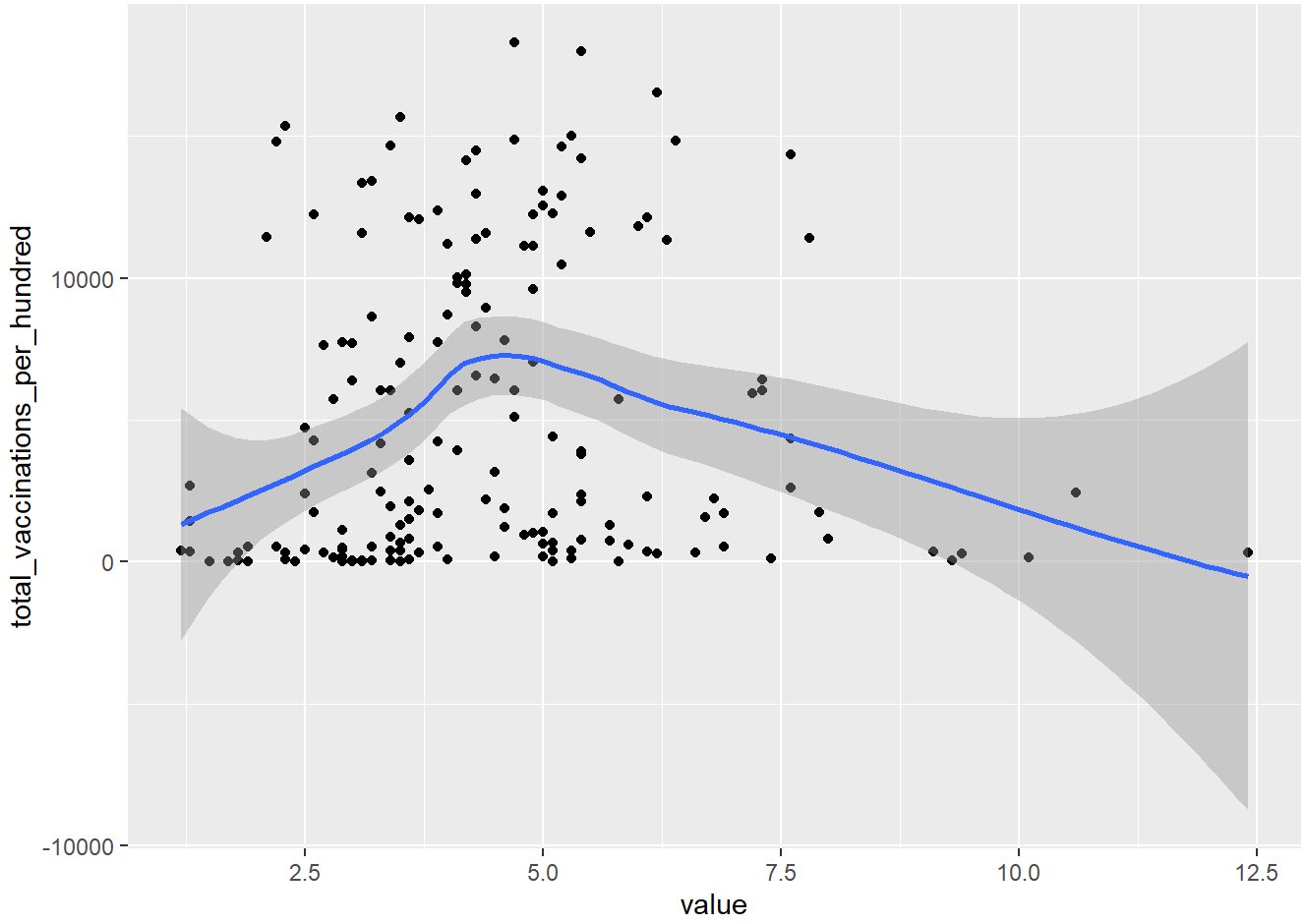


People vaccinated graphs shows the people living in relative higher education expenditures would get vaccine earlier than those who live in lower education expenditures.

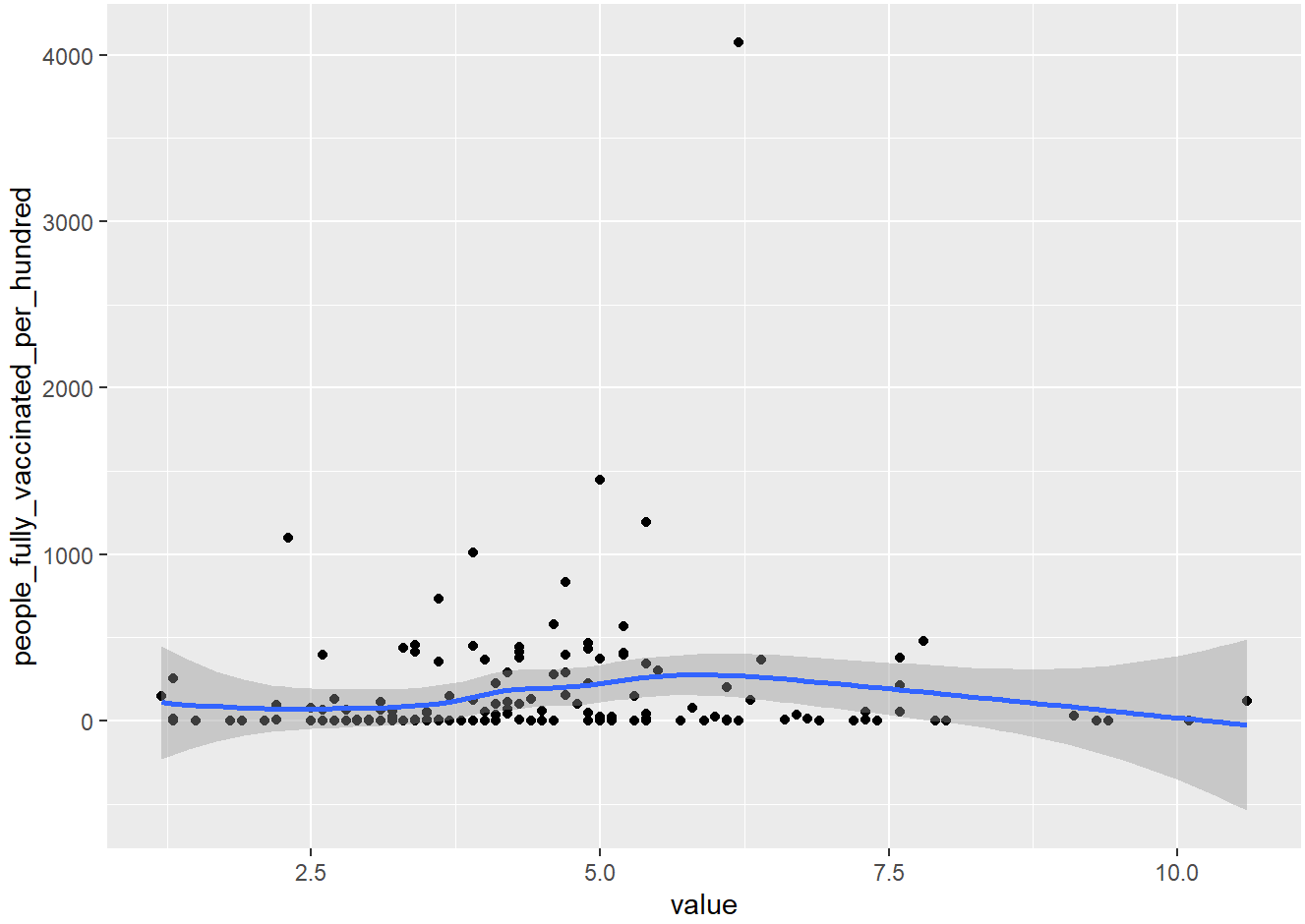
```
ggplot(covid_edu_before,mapping = aes(x=value,y=total_vaccinations_per_hundred))+geom_point()+geom_smooth()
```



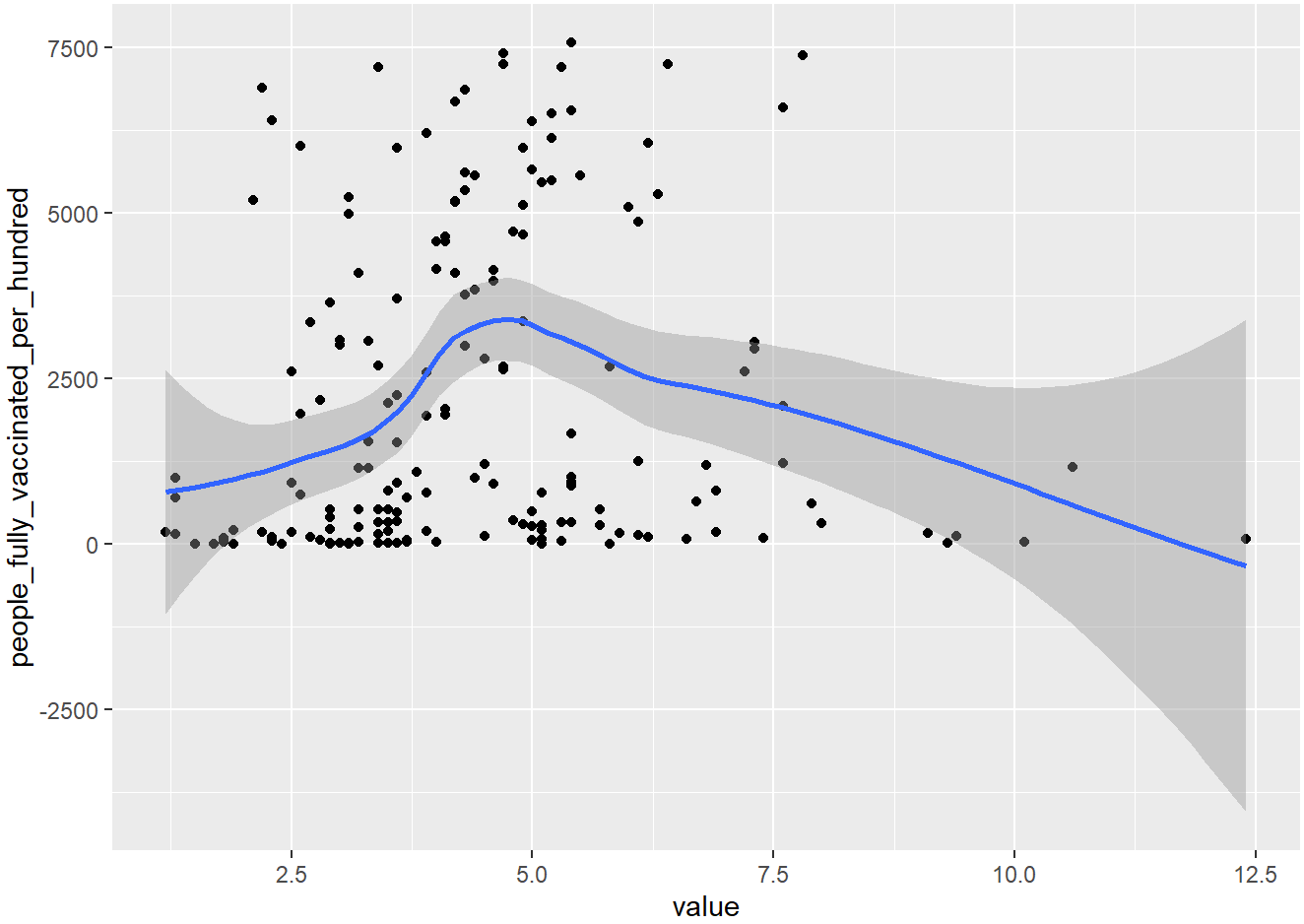

```
ggplot(covid_edu_after,mapping = aes(x=value,y=total_vaccinations_per_hundred))+geom_point()+geo
m_smooth()
```




```
ggplot(covid_edu_before,mapping = aes(x=value,y=people_fully_vaccinated_per_hundred))+geom_point
()+geom_smooth()
```

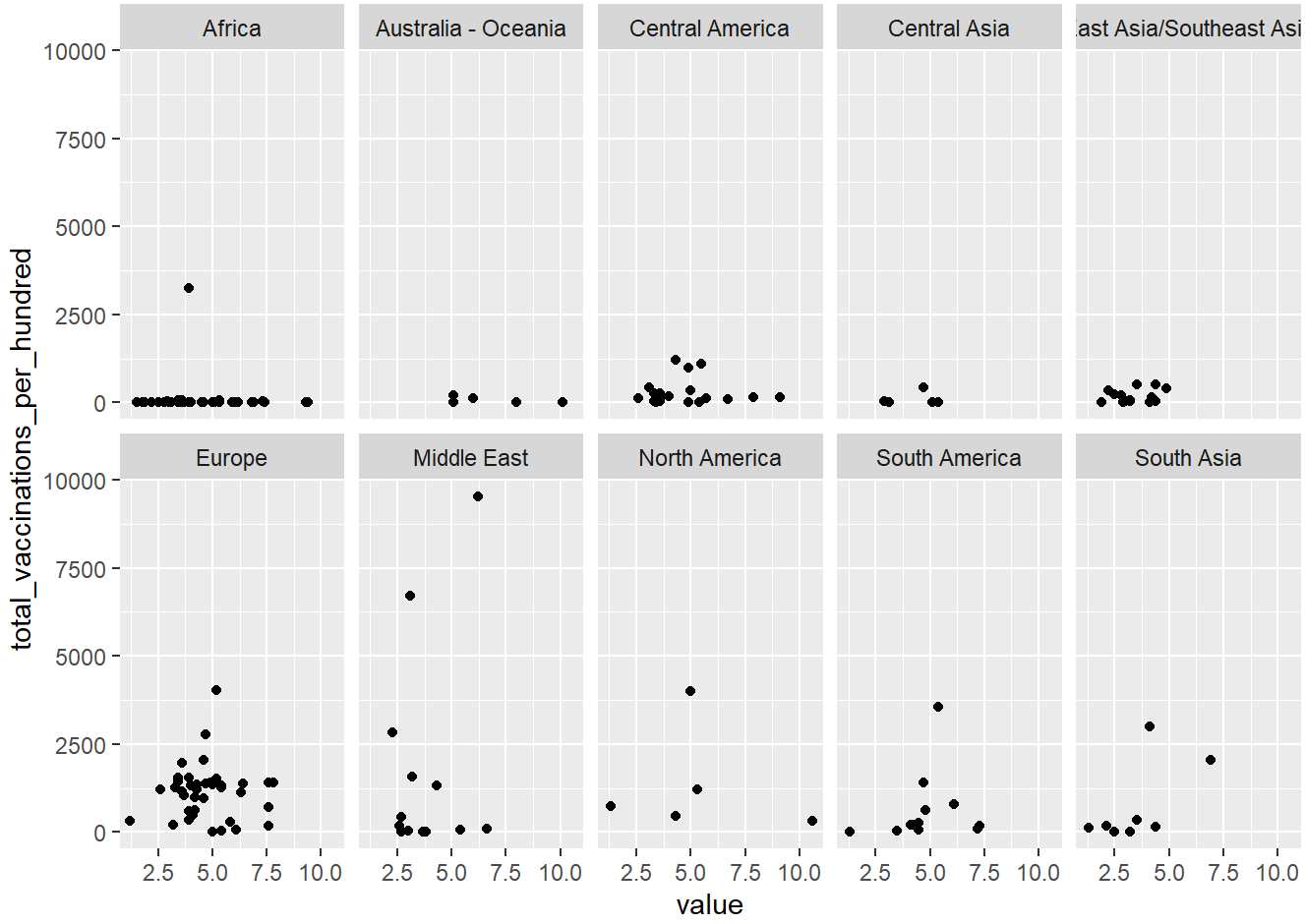



```
ggplot(covid_edu_after,mapping = aes(x=value,y=people_fully_vaccinated_per_hundred))+geom_point
()+geom_smooth()
```

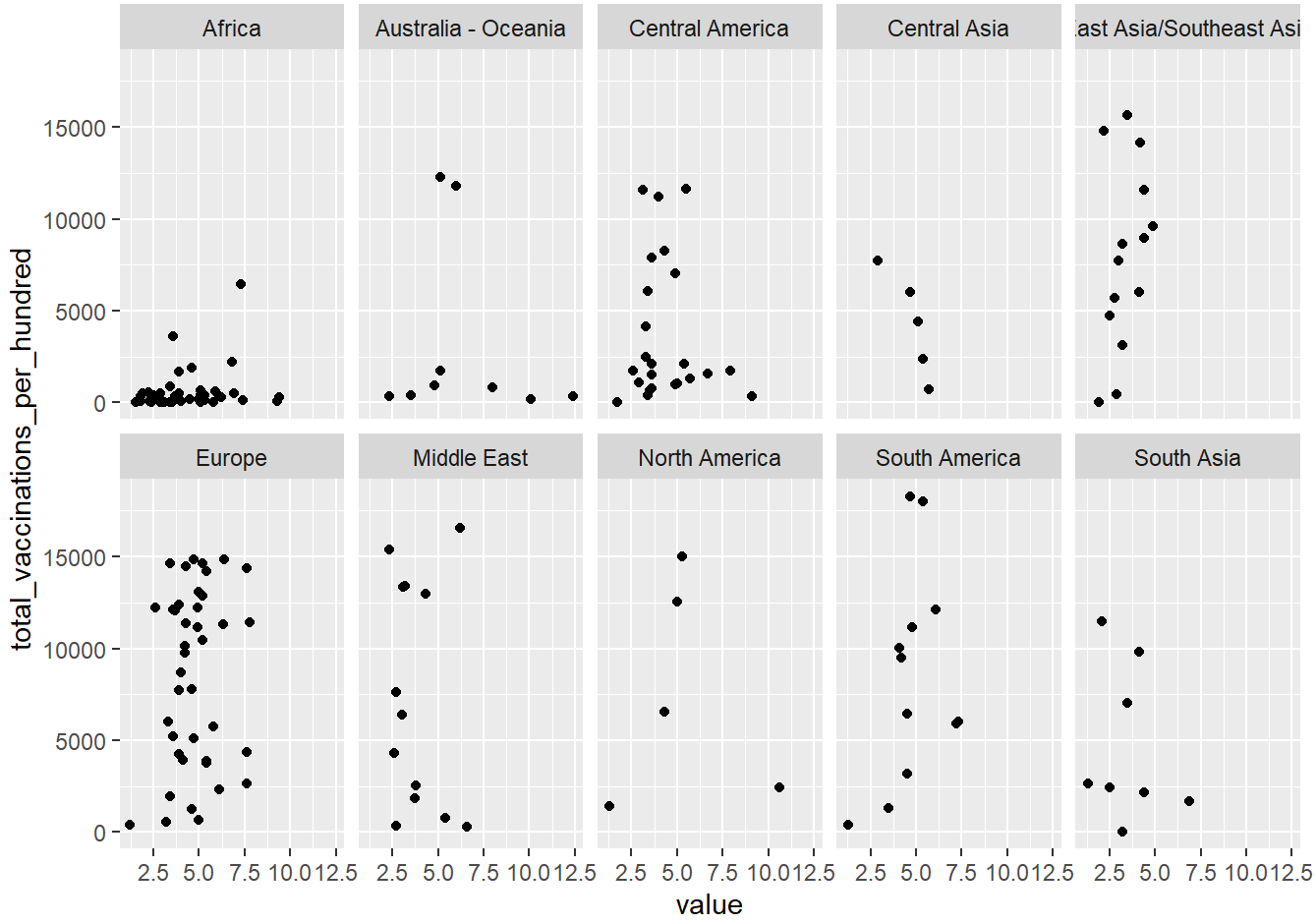


From total vaccinated per hundred and people fully vaccinated per hundred, we can see a more spread graph in period after Sept 01 than the period before May 01. It illustrates the country have relative higher education expenditure would have better performance in vaccination. However, we can see there are a few countries which have high education expenditure would have limited vaccination. Does that means the region influence the vaccinations? We pick total_vaccinations_per_hundred and people_fully_vaccinated_per_hundred to construct the new graph because it is more clear to analyze the difference between the different time spot.

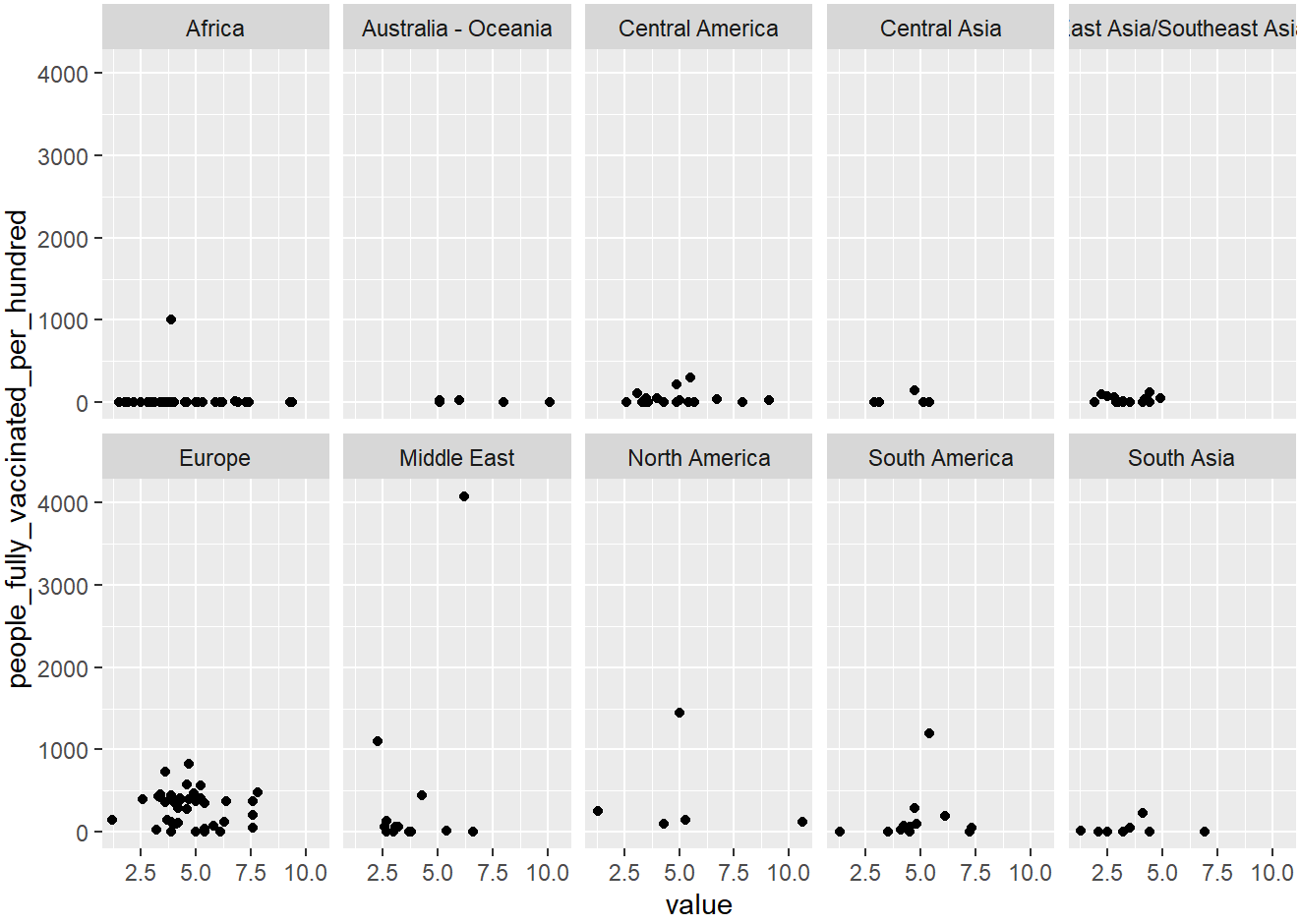
```
ggplot(covid_edu_before,mapping = aes(x=value,y=total_vaccinations_per_hundred))+geom_point()+
  facet_wrap(~region, nrow = 2)
```



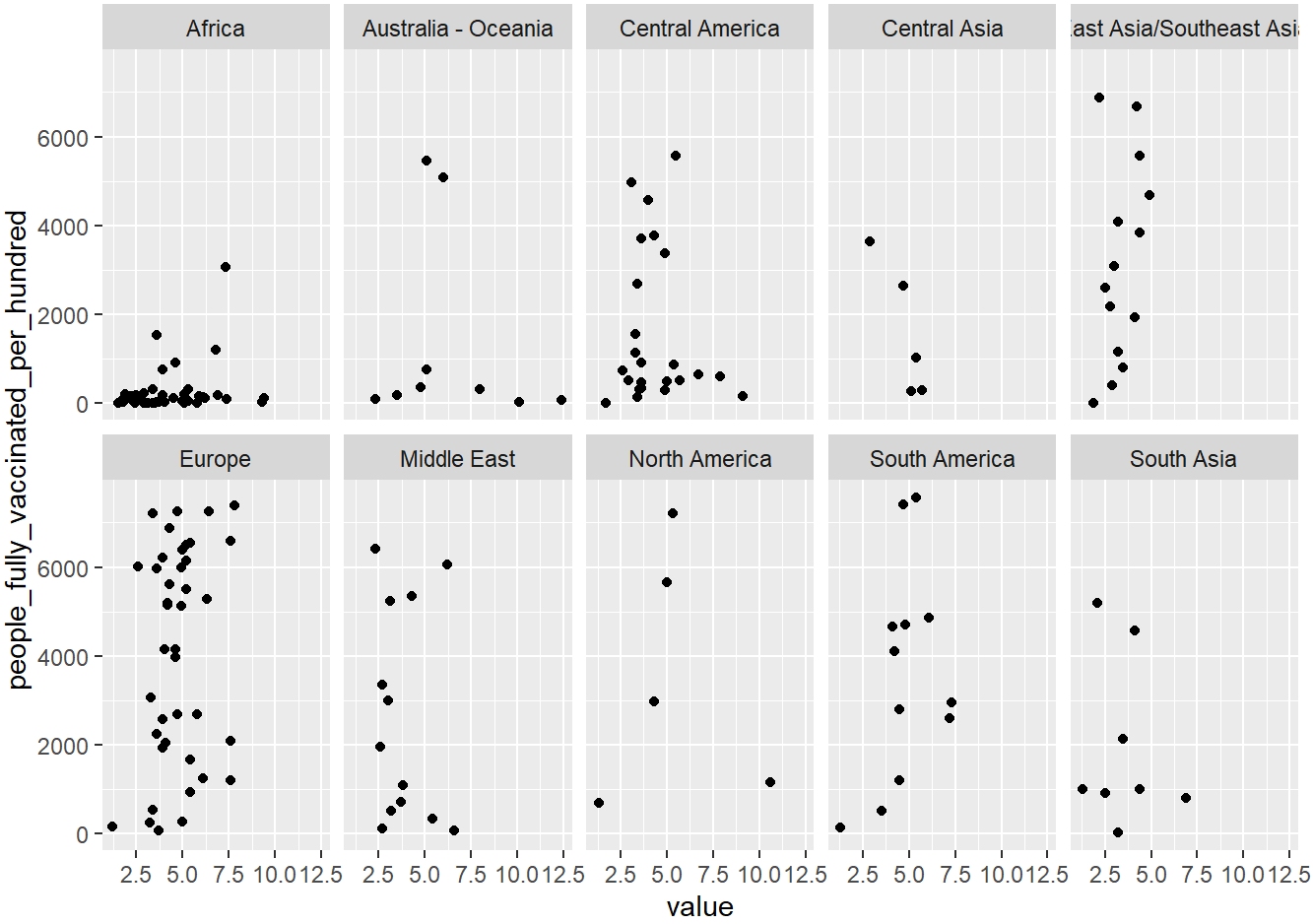

```
ggplot(covid_edu_after,mapping = aes(x=value,y=total_vaccinations_per_hundred))+geom_point()+
  facet_wrap(~region, nrow = 2)
```




```
ggplot(covid_edu_before,mapping = aes(x=value,y=people_fully_vaccinated_per_hundred))+geom_point
()+
  facet_wrap(~region, nrow = 2)
```




```
ggplot(covid_edu_after,mapping = aes(x=value,y=people_fully_vaccinated_per_hundred))+geom_point
()+
  facet_wrap(~region, nrow = 2)
```



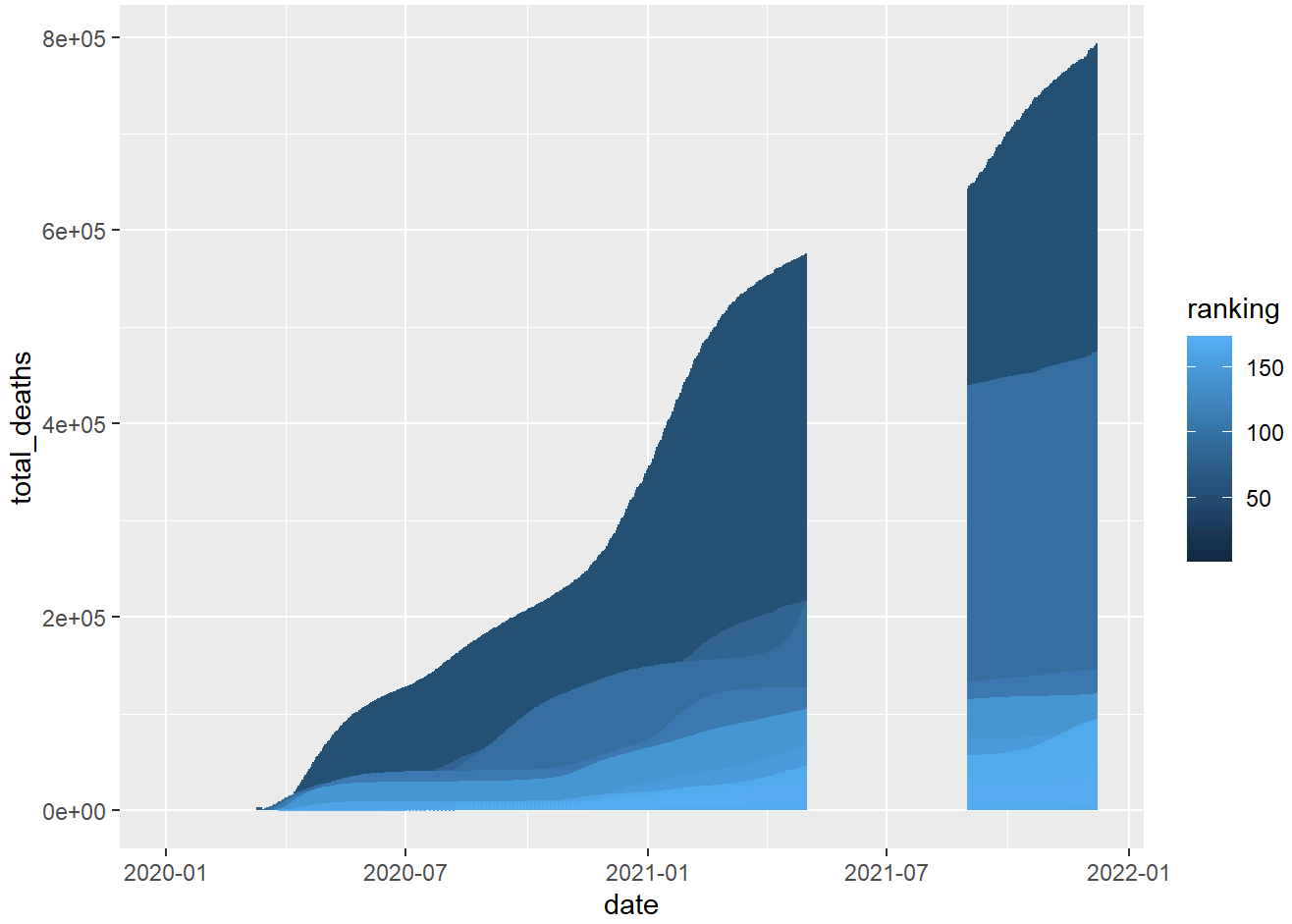
Before May 01, East Asia/Southeast Asia, Europe, and South America has a larger portion of total vaccinations per hundred and people fully vaccinated per hundred. Comparing different periods, these regions also has a positive trend between total vaccination/hundred or people fully vaccinated/ hundred and the value of education expenditure. Moreover, after Sept 01, in Africa, Middle East, North America, and south Asia, there is a positive relationship if the outliers (value>7.5) are excluded. In this case, in some region, the education expenditure seems influence the speed of the vaccination in a country. However, there might be any other factors that would influence the relationship between education expenditure and vaccination. For example, the large education expenditure in some country might not means a higher educated population. With higher population, government need to spend more on fundamental education structure such as primary school.

GINI coefficient and COVID-19

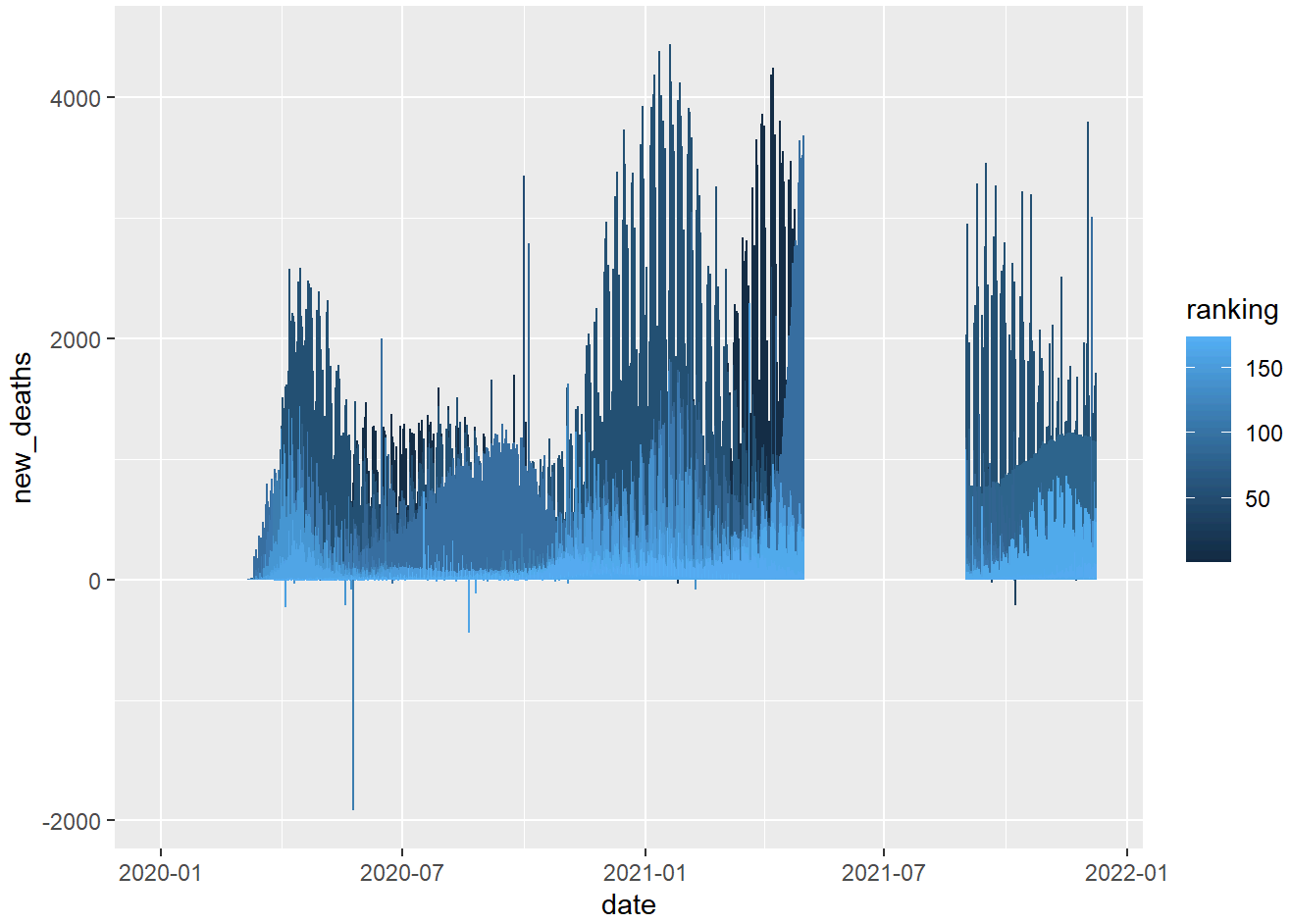
Gini coefficient is a popular measurement on the inequality of a country. Inequality always influences many fields such as the stability of a society, the happiness of the citizens, and soon. It is interesting to define the impact from Gini coefficient on COVID-19 that could help us to understand more deeply with how inequality would affect a country. A higher Gini coefficient illustrates a high inequality in a country that more poor people would be influenced when they are forced to stay at home. They are lack of money and other supports that have higher risk to get infect or die.In this way, we assume a country that have a high gini coefficient would have more deaths. To defined the impact from Gini coefficient on deaths, we randomly pick before May 01 and After Sept 01 and the variables:total_deaths, new_deaths, and total_deaths_per_million.

```
covid_gini<-left_join(gini, covid_cases, by = c("name" = "location"))%>%filter(date<=ymd("2021-05-01")|date>=ymd("2021-09-01"))
```

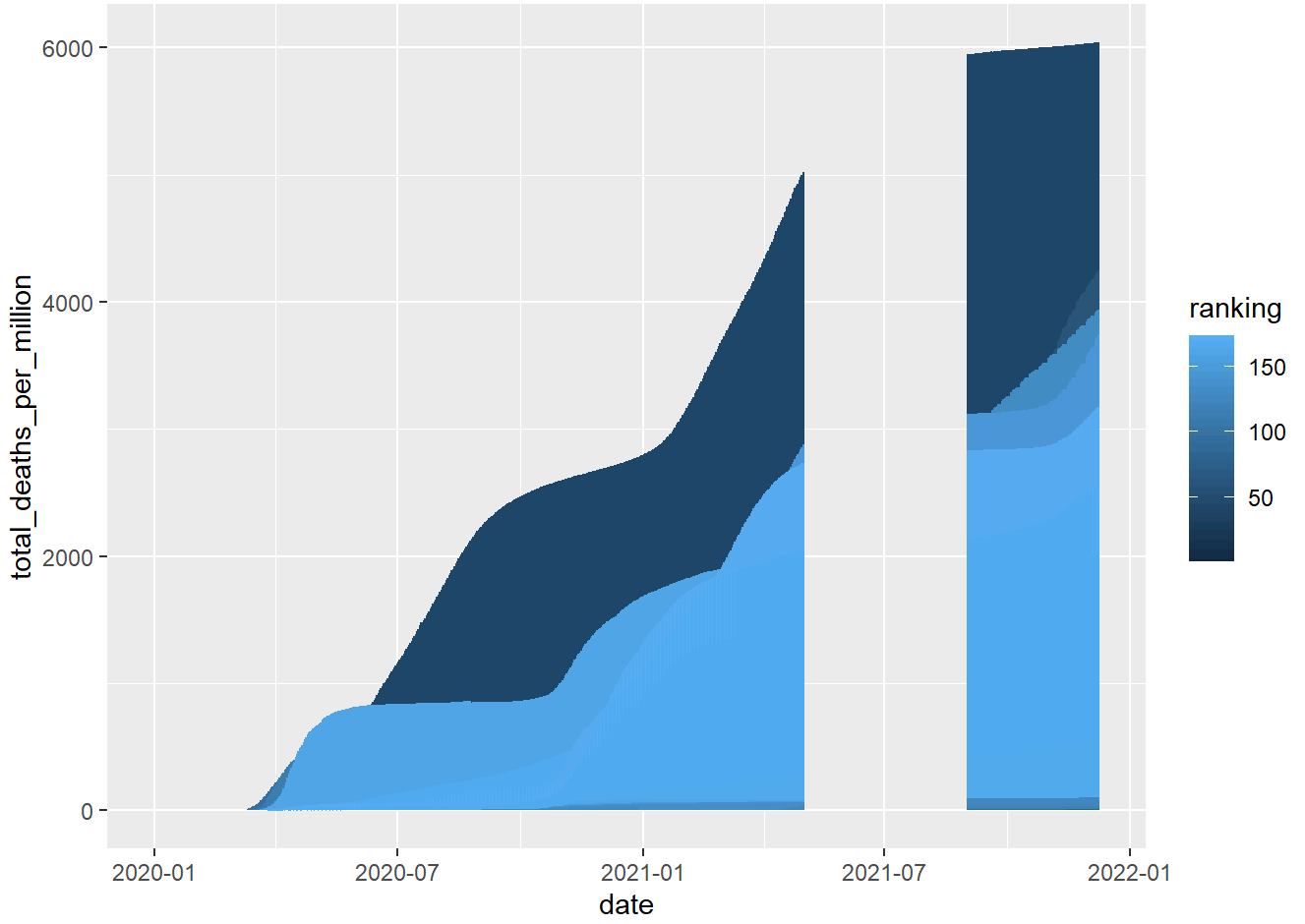
```
ggplot(covid_gini,mapping = aes(x=date,y=total_deaths))+geom_line(mapping = aes(color=ranking))
```




```
ggplot(covid_gini,mapping = aes(x=date,y=new_deaths))+geom_line(mapping = aes(color=ranking))
```

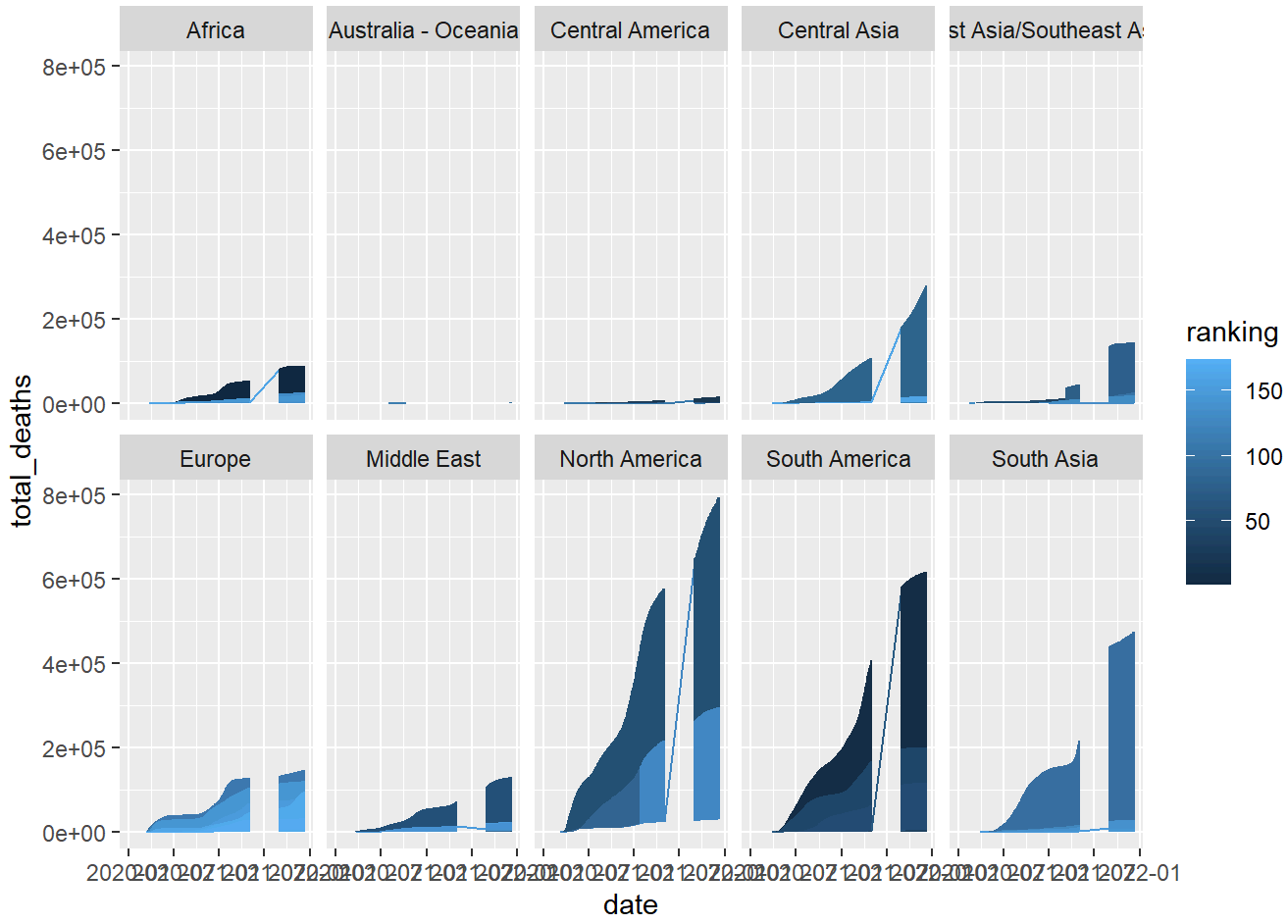



```
ggplot(covid_gini,mapping = aes(x=date,y=total_deaths_per_million))+geom_line(mapping = aes(colo
r=ranking))
```

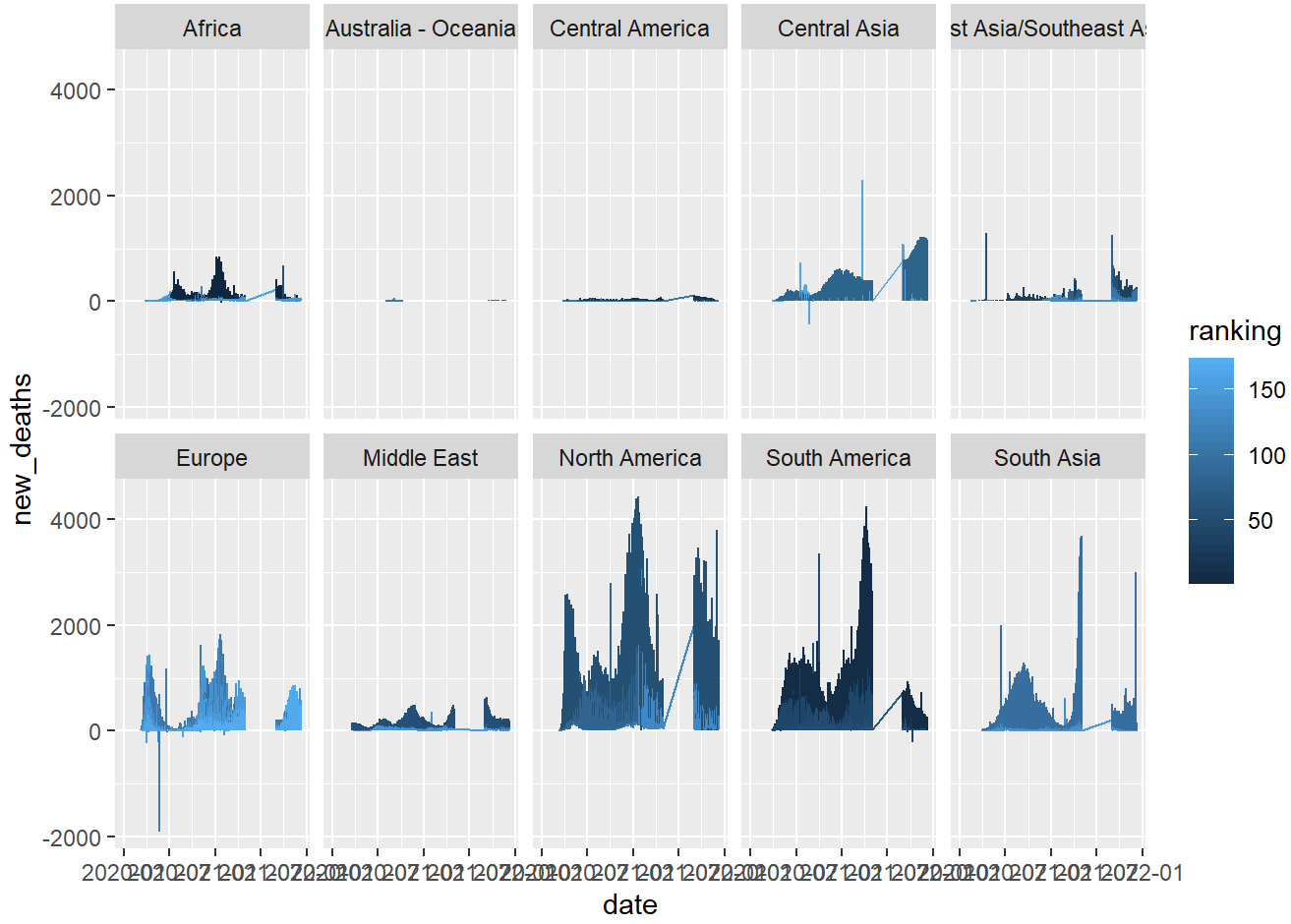


From the graphs above, with higher ranking in the Gini coefficient, the country would have more total deaths, new deaths, and total deaths per million than other country.For those which have a low Gini coefficient country, the total deaths are quit small and have a constant number after Sept 01 that means those countries have controlled the COVID-19. While the country which has high Gini coefficient have a rising number on total deaths, even after Sept 01. The interpretation of other two variable is same as total deaths that Gini coefficient has impact on deaths.

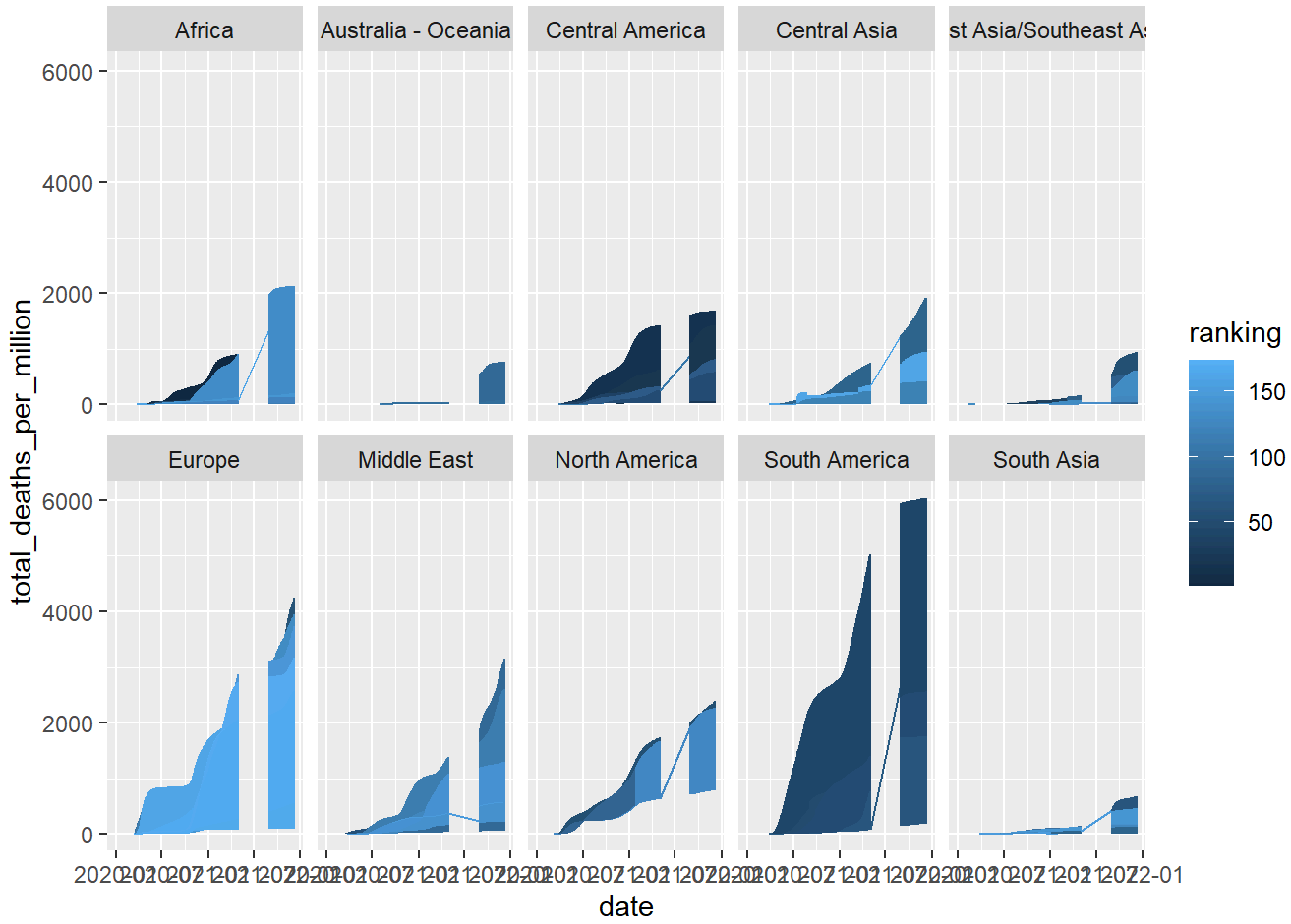
```
ggplot(covid_gini,mapping = aes(x=date,y=total_deaths))+geom_line(mapping = aes(color=ranking))+
  facet_wrap(~region, nrow = 2)
```




```
ggplot(covid_gini,mapping = aes(x=date,y=new_deaths))+geom_line(mapping = aes(color=ranking))+
  facet_wrap(~region, nrow = 2)
```




```
ggplot(covid_gini,mapping = aes(x=date,y=total_deaths_per_million))+geom_line(mapping = aes(colo
r=ranking))+
  facet_wrap(~region, nrow = 2)
```



From the graphs above, Africa, Middle East, South America, and East Asia/Southeast Asia show the impact from gini coefficient on deaths more clear. The reason might be these regions have relatively higher Gini coefficient than other regions. Europe, North America and South Asia have low Gini coefficient with a relatively unclear pattern on the relationship between Gini coefficient and deaths. However, comparing across regions, with lower Gini coefficient, Europe would have a lower peak than South America which have high Gini coefficient. To conclude, Gini coefficient seems have obvious impact on COVID-19 deaths number. However, in some regions, the patterns are not clear of the deaths are relatively small such as East Asia/Southeast Asia. It might be explained by other factors. For example, the political concentrated might affect the control of the COVID-19 that might help to decrease the deaths.

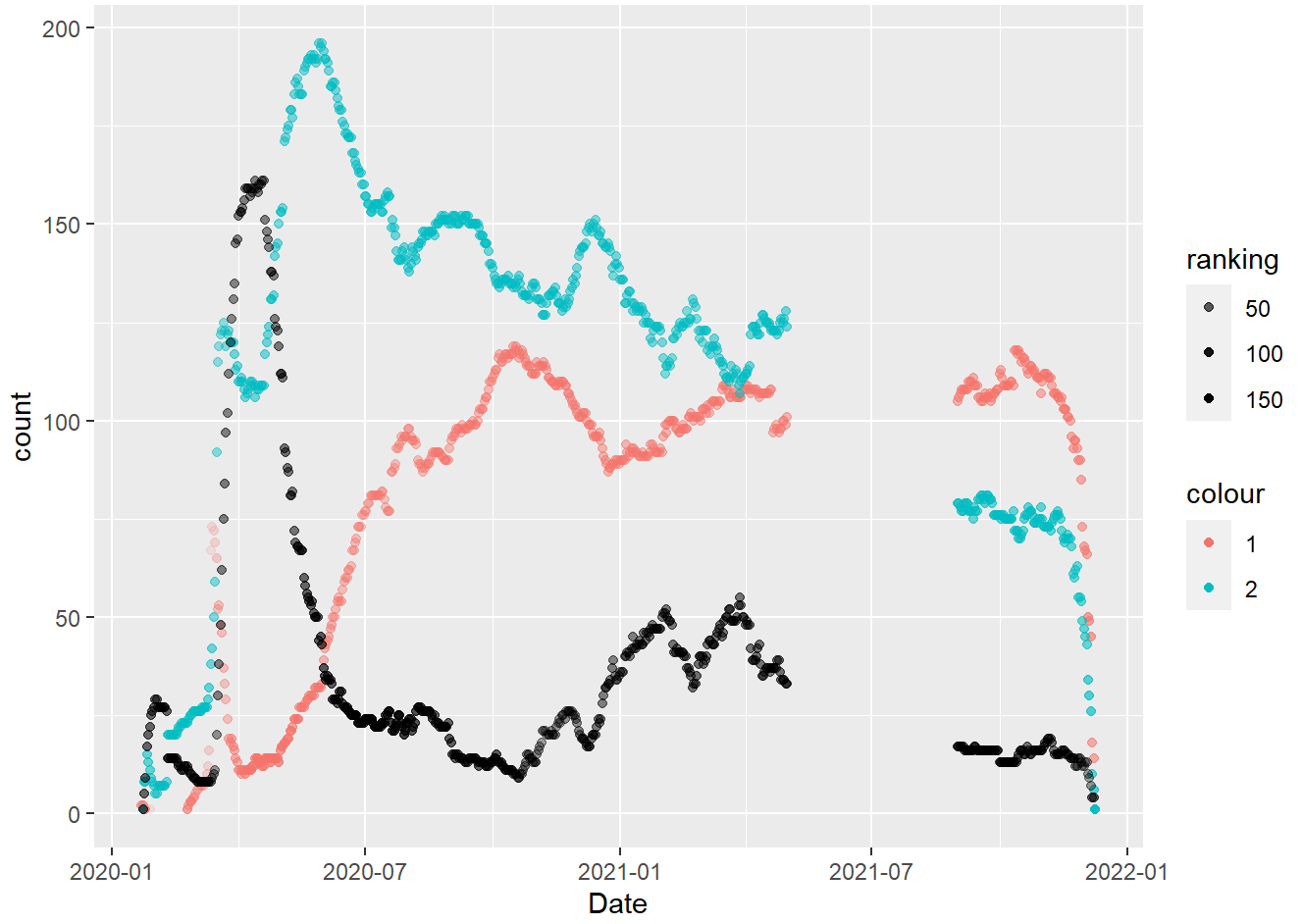
COVID-19 and GDP

GDP is a main part to represents the performance of a country. It counts the total final goods and services produced within a country during a year. That means a higher GDP illustrates a countries high ability in productivity. In this way, when facing a natural crisis, it would be more likely to actively deal with the problem. Under COVID-19, we assume a high GDP country would have ability to handle and afford the cases in closing the work place that might decrease the production of a country. The period is randomly select that are before May 01 and after Sept 01.

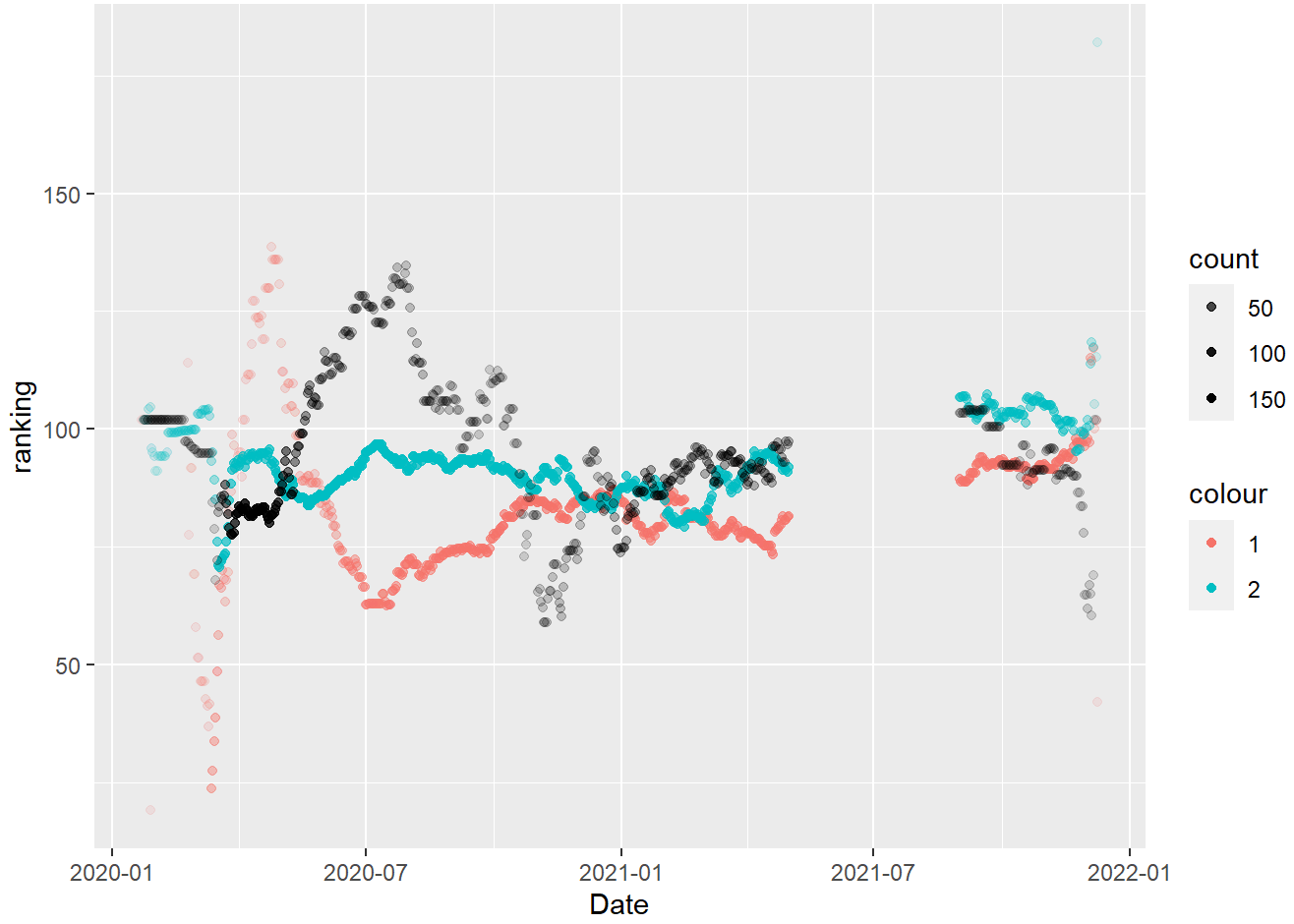
```
library(dplyr)
library(magrittr)
covid_reponse<-covid_response%>%mutate(Date=as.Date(Date,format ="%Y-%m-%d"))
covid_gdp<-left_join(gdppp, covid_response, by = c("name" ="CountryName"))%>%filter(Date<=ymd("2
021-05-01")|Date>=ymd("2021-09-01"))#%>%mutate(value=as.integer(as.character(value)))
```

To format the data, we calculate the total number of the countries that choose 1-recommend closing (or recommend work from home) or all businesses open with alterations resulting in significant differences compared to non-Covid-19 operation, 2-require closing (or work from home) for some sectors or categories of workers, and 3-require closing (or work from home) for all-but-essential workplaces (eg grocery stores, doctors). Then we take the mean of the ranking that illustrates the countries' position in GDP. In the following graphs, points illustrate the number of countries and date, and the line illustrates the average ranking and date.

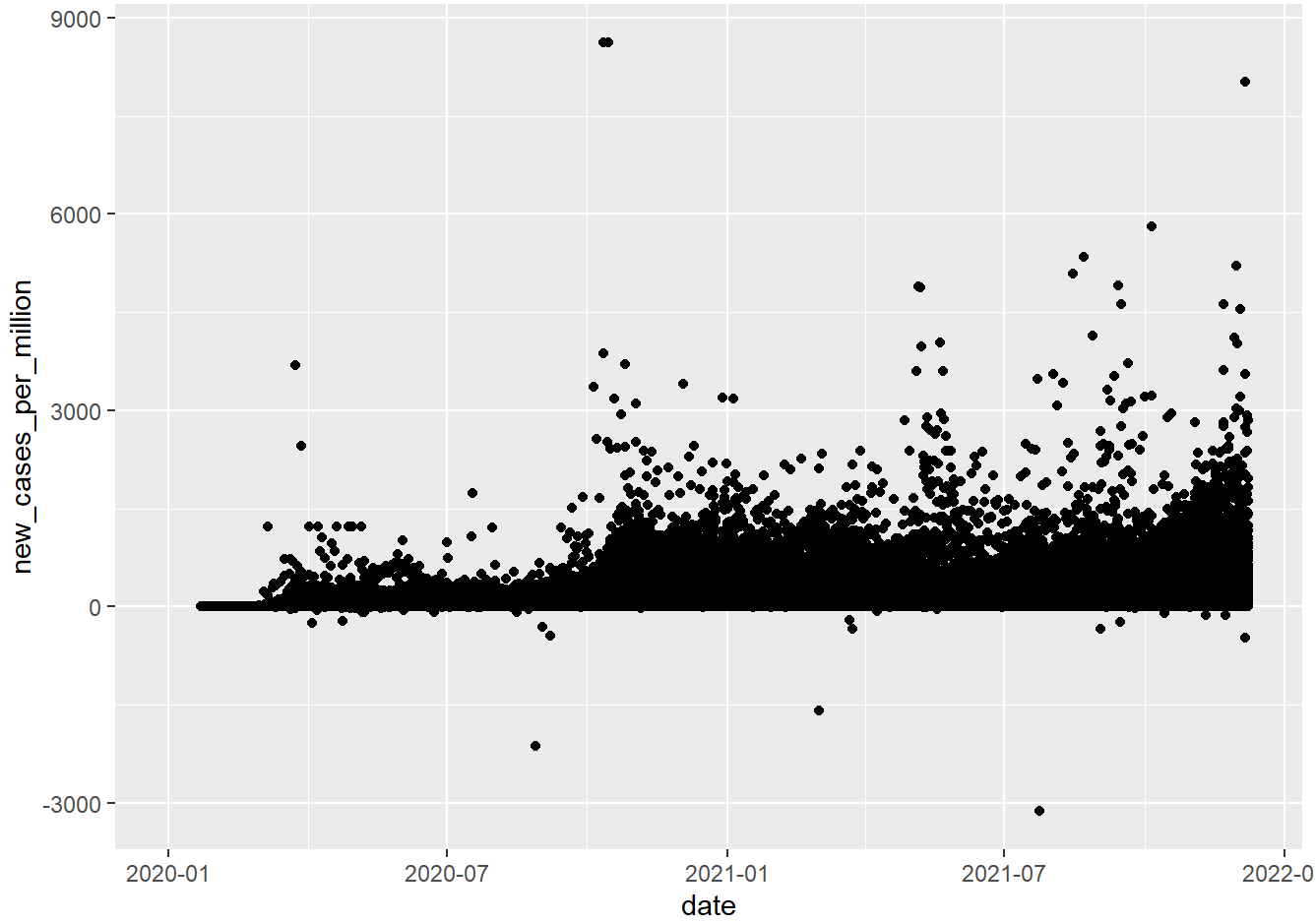
```
covid_gdp_work_1<-covid_gdp%>%filter(`C2_Workplace closing`=="1"|`C2_Workplace closing`=="1.0
0")%>%group_by(Date)%>%select(name,Date,ranking)%>%summarise(count=n(),ranking=mean(ranking))
covid_gdp_work_2<-covid_gdp%>%filter(`C2_Workplace closing`=="2"|`C2_Workplace closing`=="2.0
0")%>%group_by(Date)%>%select(name,Date,ranking)%>%summarise(count=n(),ranking=mean(ranking))
covid_gdp_work_3<-covid_gdp%>%filter(`C2_Workplace closing`=="3"|`C2_Workplace closing`=="3.0
0")%>%group_by(Date)%>%select(name,Date,ranking)%>%summarise(count=n(),ranking=mean(ranking))

ggplot(covid_gdp_work_1)+geom_point(covid_gdp_work_1,mapping=aes(x=Date,y=count,color="1",alpha=
ranking))+geom_point(covid_gdp_work_2,mapping=aes(x=Date,y=count,color="2",alpha=ranking))+geom_
point(covid_gdp_work_3,mapping=aes(x=Date,y=count,alpha=ranking))
```




```
ggplot(covid_gdp_work_1)+geom_point(mapping = aes(x=Date,y=ranking,color="1",alpha=count))+geom_
point(covid_gdp_work_2,mapping = aes(x=Date,y=ranking,color="2",alpha=count))+geom_point(covid_g
dp_work_3,mapping = aes(x=Date,y=ranking,alpha=count))
```




```
ggplot(covid_gdp_work_1)+geom_point(covid_cases,mapping=aes(x=date,y=new_cases_per_million))
```



From the graphs above, at the beginning of the COVID-19, most countries choose option 3 that close all of the work places to control the spread of the COVID-19. After the COVID-19 gets better, countries tend to release the restrictions on work places to option 2 and option 1. However, countries that choose option 3 is the smallest number compare to other two options that means some country would not like to close all of the work place even it is a better way to control the COVID-19 spread. Compare to option 1 and 3, countries prefer to choose option 2 that is not too restricted or too wide. The second graph illustrates the average ranking vs date that base on three options and the count of three options. As the count increase, the mean of the rank decrease that is obvious. However, when the count is lower, the mean of ranking is still lower that might mean the country would have higher GDP. We can see that, at the end of 2020, both the average ranking and the count of the country are both small that means these country have high GDP that would like to take option 3. Besides, at the beginning of the 2020, most countries took option 3 to deal with COVID-19, while few countries chose option 1 that has large mean of ranking that means the countries with low GDP would prefer option 1 to option 3. After this period, the mean of ranking for option goes up with small number of country might be seen as the lag of the decision on going into the most restrict option for those countries with lower GDP. When we see the graph 3, on 2020-07, we can see there is a gap that means most country has controlled the COVID-19, while in graph 3 some of lower GDP country began to close all-but-essential workplaces (eg grocery stores, doctors). To conclude, we have discovered a likelihood of higher GDP countries would prefer actively dealing with the pandemic on a better option to prohibit the spread.

Cagri Isilak Report

Return to Introduction

Hypothesis: Covid cases went down after vaccines came out/during lockdown

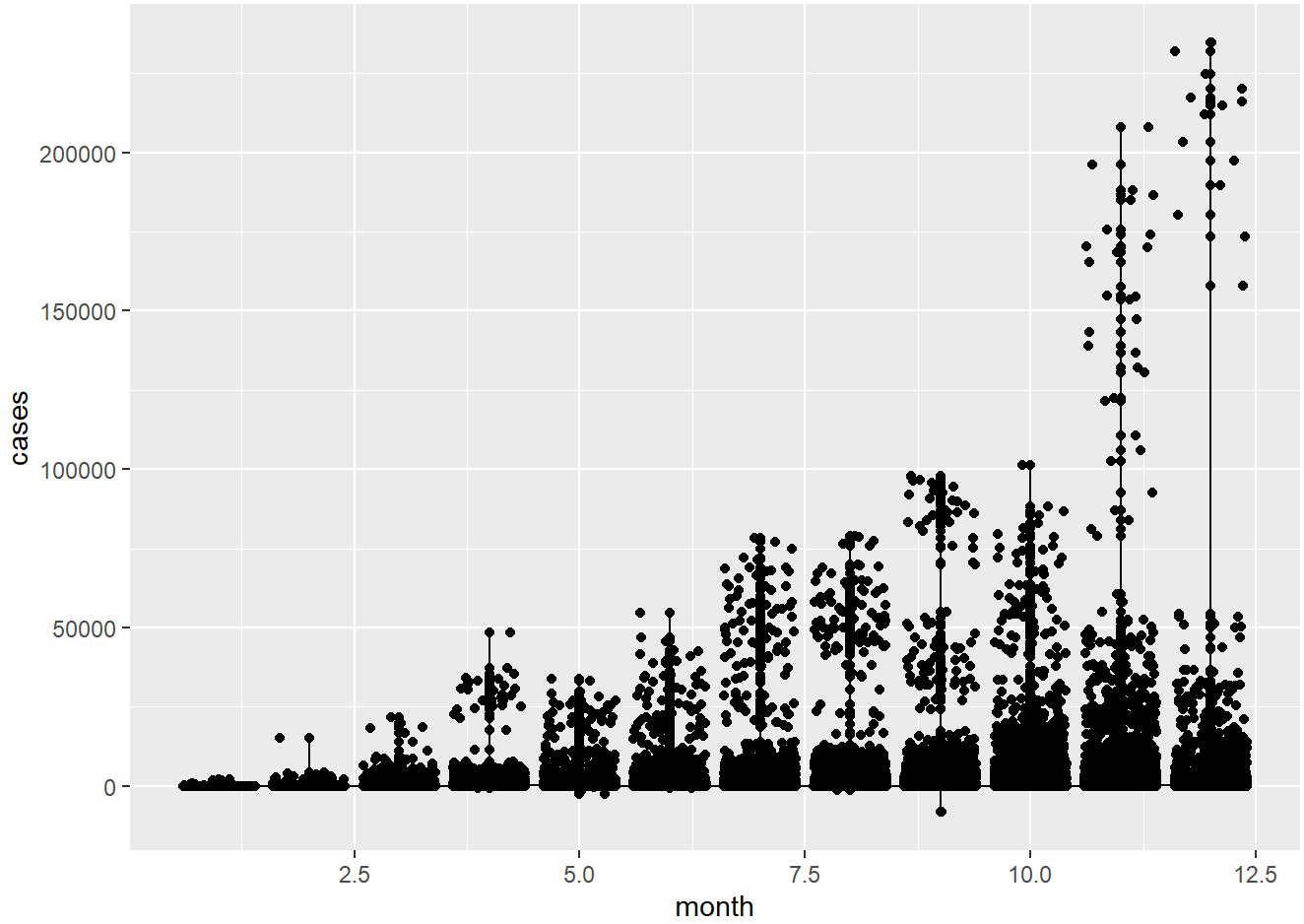
Vaccines began to come out in Canada around May 2020. Covid lockdowns began in March 2020.

During lockdown, there is less people interacting and much more social distancing, meaning the virus will have a much harder time spreading. The dataset below is the covid 19 case-count per day for 2020 (all countries). I will be basing my analysis of the lockdown and vaccination dates of Canada, which were not too different from the rest of the world.

```
# retrieved @ https://data.europa.eu/data/datasets/covid-19-coronavirus-data-daily-up-to-14-december-2020?locale=en
Daily_Cases <- read_csv("DailyCases.csv")

Daily_Cases %>%
  head()
```

```
## # A tibble: 6 x 12
##   dateRep   day month  year cases deaths countriesAndTer~ geoId countryterritor~
##   <chr>   <dbl> <dbl> <dbl> <dbl>  <dbl> <chr>            <chr> <chr>
## 1 14/12/~    14    12  2020   746      6 Afghanistan      AF    AFG
## 2 13/12/~    13    12  2020   298      9 Afghanistan      AF    AFG
## 3 12/12/~    12    12  2020   113     11 Afghanistan      AF    AFG
## 4 11/12/~    11    12  2020    63     10 Afghanistan      AF    AFG
## 5 10/12/~    10    12  2020   202     16 Afghanistan      AF    AFG
## 6 09/12/~     9    12  2020   135     13 Afghanistan      AF    AFG
## # ... with 3 more variables: popData2019 <dbl>, continentExp <chr>,
## #   Cumulative_number_for_14_days_of_COVID-19_cases_per_100000 <dbl>
```

```
ggplot(Daily_Cases, aes(x=month,y=cases))+
  geom_line()+
  geom_point()+
  geom_jitter()
```



In the graph above, we can see that the increase of cases slowed down a bit in March when lockdowns started, and also dipped slightly down in May which is when vaccines began coming out. From there on we can see that although cases continued increasing in most countries, more and more countries had lower cases, and we can see this the most clearly in December. In addition, I believe the branching off of two visible sections, countries with ever increasing covid cases and countries with decreasing covid cases, can be explained since by then society was getting more used to life in lockdown and more people were getting vacccinated, however at the same time many people had enough of the lockdown and wanted restrictions to lift, which led to many large scale protests that were broadcasted to television around that time. I believe that these protests along with the rising number of people who believe in conspiracy theories combined with those who are not yet vaccinated and anti-vaxxers (people who refuse to be vaccinated because they do not trust the vaccine or for other reasons) are the reason why Covid cases are still increasing. Another very big reason for a continuous increase of Covid cases is the introduction of mutations that can nullify the protection of some vaccines. Mutations, such as a Delta Variant are most likely why there is a large spike in cases in many countries in December, as that is when the Delta variant was first identified, and it had probably begun spreading a lot more by then.