CovidViz

Elena

2021/8/16

Set up the environmet

Load the readr, ggplot2, and dplyr packages

```
library(readr)
library(ggplot2)
library(dplyr)
library(reshape2)
```

Data preparation

Read datasets/confirmed_cases_worldwide.csv into confirmed_cases_worldwide

```
confirmed_cases_perday <- read_csv("F:/Corona_Viz_R/coronavirus.csv")
```

```
## Rows: 469362 Columns: 7
```

```
## -- Column specification --------------------------------------------------------------------------------------
--------
## Delimiter: ","
## chr  (3): province, country, type
## dbl  (3): lat, long, cases
## date (1): date
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
confirmed_cases_country <- confirmed_cases_perday %>%
  group_by(country, date) %>%
  filter(type == "confirmed",date <= as.Date('2020-03-17')) %>%
  summarise(cases = sum(cases)) %>%
  mutate(cum_cases = cumsum(cases))
```

```
## `summarise()` has grouped output by 'country'. You can override using the `.groups` argument.
```

```
confirmed_cases_worldwide <- confirmed_cases %>%
  group_by(date) %>%
  filter(type == "confirmed") %>%
  summarise(cases = sum(cases)) %>%
  mutate(cum_cases = cumsum(cases)) %>%
  filter(date <= as.Date('2020-03-17'))
```

See the result

```
summary(confirmed_cases_worldwide)
```

```
##       date                cases           cum_cases
##  Min.   :2020-01-22   Min.   :   98.0   Min.   :   557
##  1st Qu.:2020-02-04   1st Qu.:  946.2   1st Qu.: 26707
##  Median :2020-02-18   Median : 2246.5   Median : 75402
##  Mean   :2020-02-18   Mean   : 3571.9   Mean   : 68654
##  3rd Qu.:2020-03-03   3rd Qu.: 3964.5   3rd Qu.: 93556
##  Max.   :2020-03-17   Max.   :15962.0   Max.   :200024
```

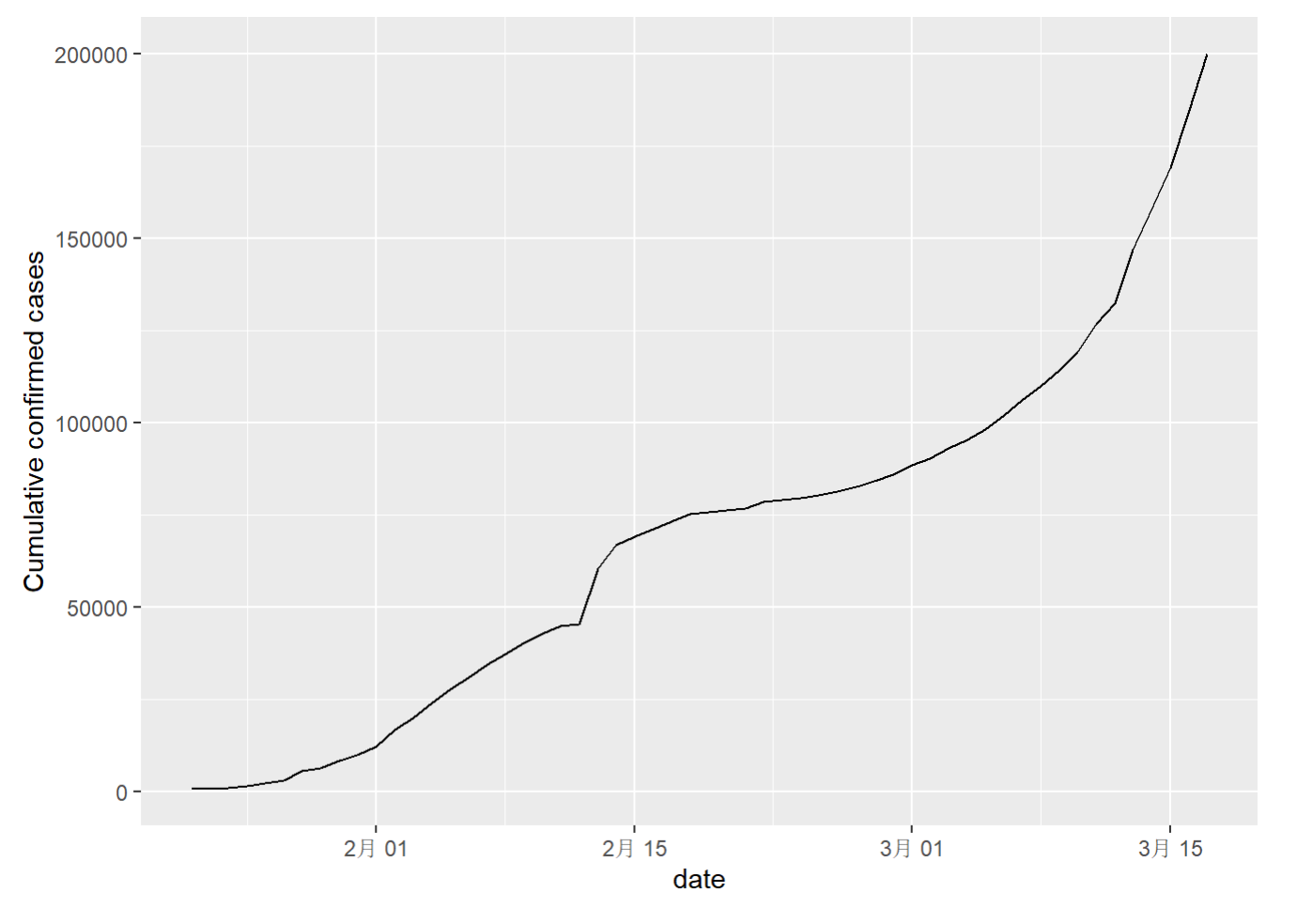
```
str(confirmed_cases_worldwide)
```

```
## tibble [56 x 3] (S3: tbl_df/tbl/data.frame)
##  $ date     : Date[1:56], format: "2020-01-22" "2020-01-23" "2020-01-24" "2020-01-25" ...
##  $ cases    : num [1:56] 557 98 286 492 685 ...
##  $ cum_cases: num [1:56] 557 655 941 1433 2118 ...
```

Plotting — worldwide

Let's draw a line plot to visualize the confirmed cases worldwide.

```
ggplot(confirmed_cases_worldwide) +
  geom_line(mapping = aes(x = date, y = cum_cases)) +
  ylab("Cumulative confirmed cases")
```



Early on in the outbreak, the COVID-19 cases were primarily centered in China. Let's plot confirmed COVID-19 cases in China and the rest of the world separately to see if it gives us any insight.

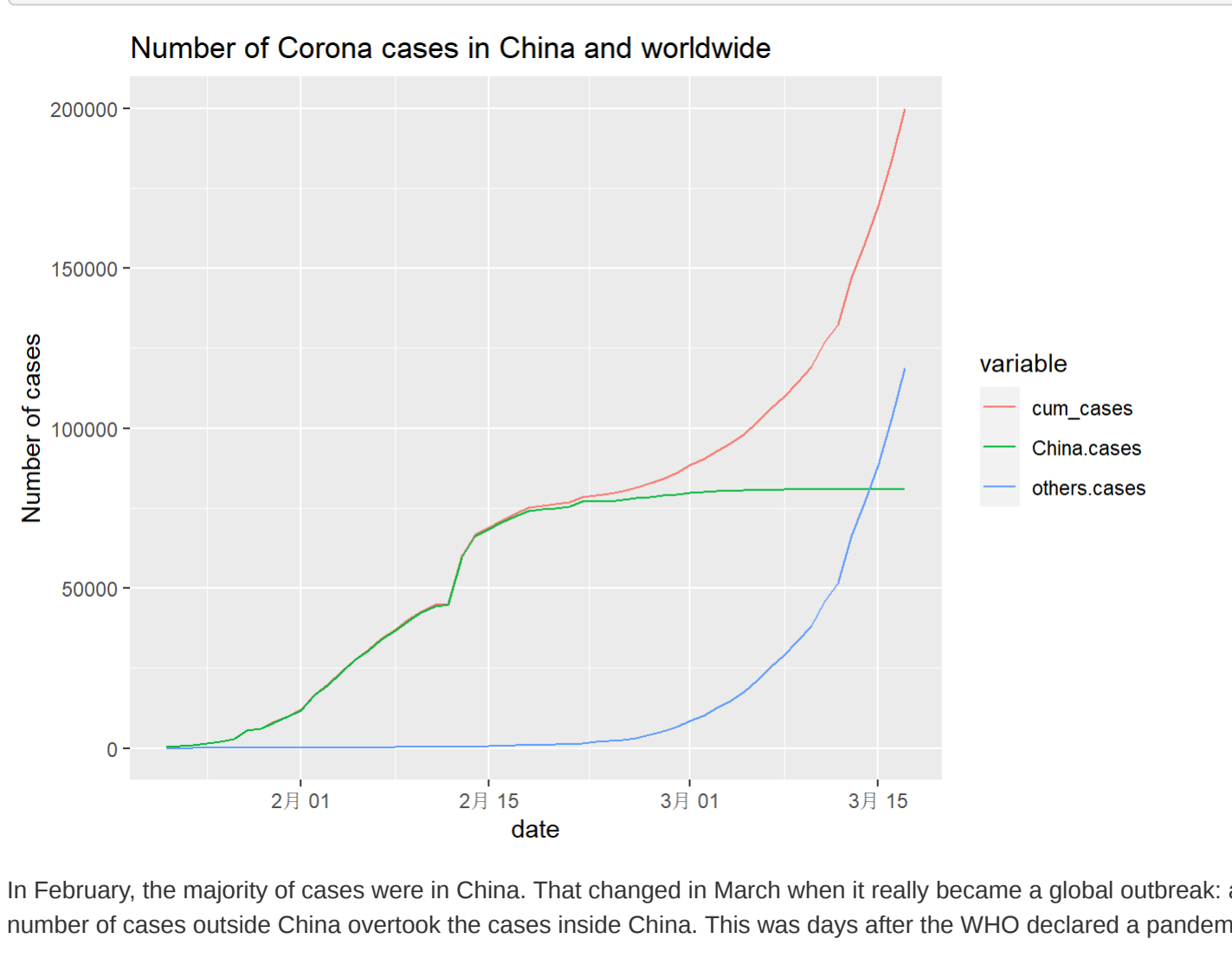
Plotting - China

```
confirmed_cases_china <-
  confirmed_cases_country %>%
  filter(country == "China") %>%
  group_by(date)

confirmed_cases_china_vs_world <-
  confirmed_cases_worldwide %>%
  mutate(China.cases = confirmed_cases_china$cum_cases) %>%
  mutate(others.cases = cum_cases - confirmed_cases_china$cum_cases)

confirmed_cases_china_vs_world <- confirmed_cases_china_vs_world[-c(2)]

china_vs_world_melt <-
  melt(confirmed_cases_china_vs_world, id = "date")

china_vs_world_melt
```

date	variable	value
<date>	<fct>	<dbl>
2020-01-22	cum_cases	557
2020-01-23	cum_cases	655
2020-01-24	cum_cases	941
2020-01-25	cum_cases	1433
2020-01-26	cum_cases	2118
2020-01-27	cum_cases	2927
2020-01-28	cum_cases	5578
2020-01-29	cum_cases	6167
2020-01-30	cum_cases	8235
2020-01-31	cum_cases	9927

1-10 of 168 rows Previous 1 2 3 4 5 6 ... 17 Next

```
plt_cum_confirmed_cases_china_vs_world<- ggplot(china_vs_world_melt)+
  geom_line(mapping = aes(x= date, y = value, color = variable, group = variable))+
  labs(title = "Number of Corona cases in China and worldwide", y = "Number of cases")

plt_cum_confirmed_cases_china_vs_world
```

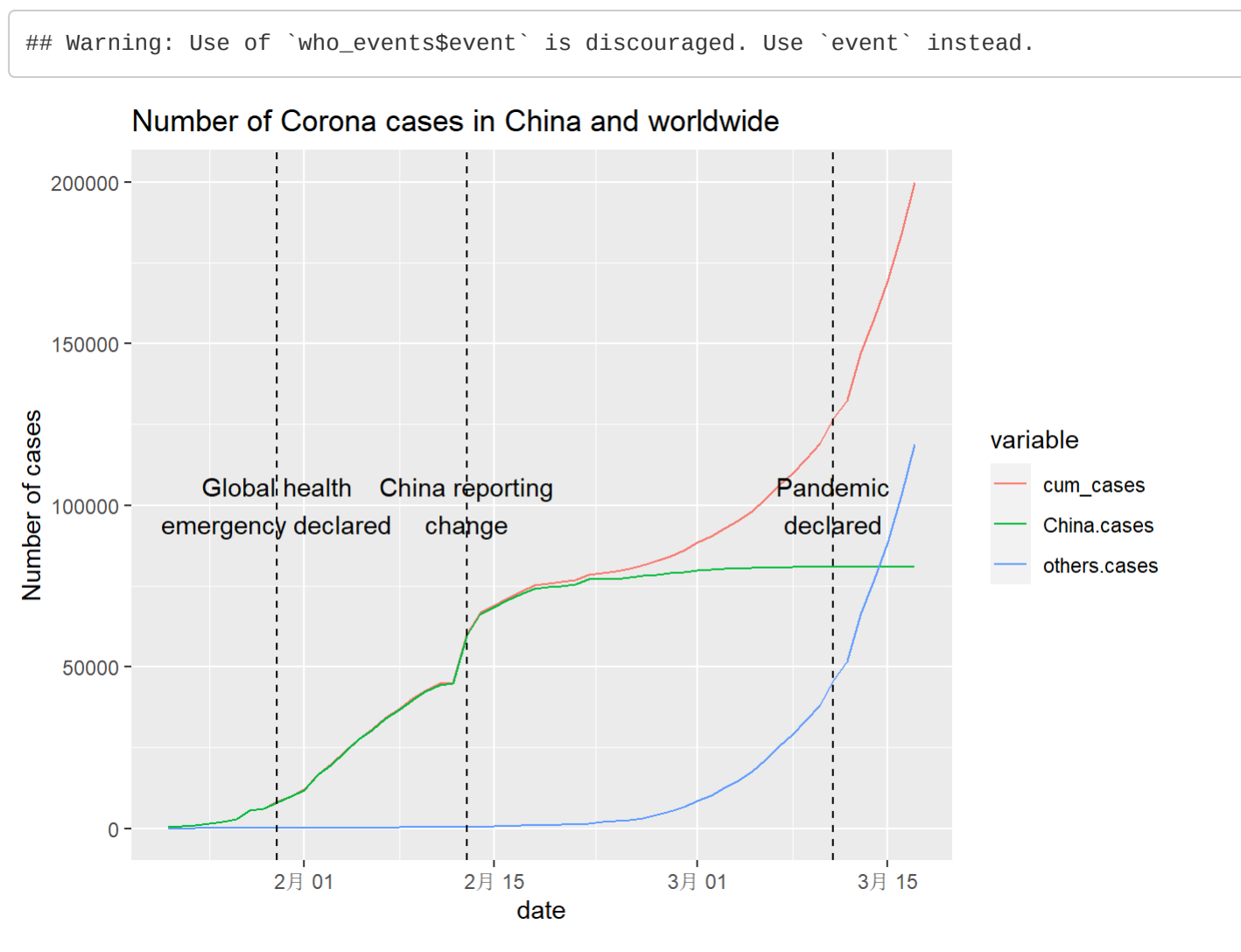


In February, the majority of cases were in China. That changed in March when it really became a global outbreak: around March 14, the total number of cases outside China overtook the cases inside China. This was days after the WHO declared a pandemic.

Add annotation

```
who_events <- tribble(
  ~ date, ~ event,
  "2020-01-30", "Global health\nemergency declared",
  "2020-02-13", "China reporting\nchange",
  "2020-03-11", "Pandemic\ndeclared"
) %>%
  mutate(date = as.Date(date))

# Using who_events, add vertical dashed lines with an xintercept at date
# and text at date, labeled by event, and at 10000 on the y-axis

plt_cum_confirmed_cases_china_vs_world +
  geom_vline(aes(xintercept = date), data = who_events, linetype = "dashed") +
  geom_text(aes(date, label = who_events$event), data = who_events , y = 1e5)
```

```
## Warning: Use of `who_events$event` is discouraged. Use `event` instead.
```

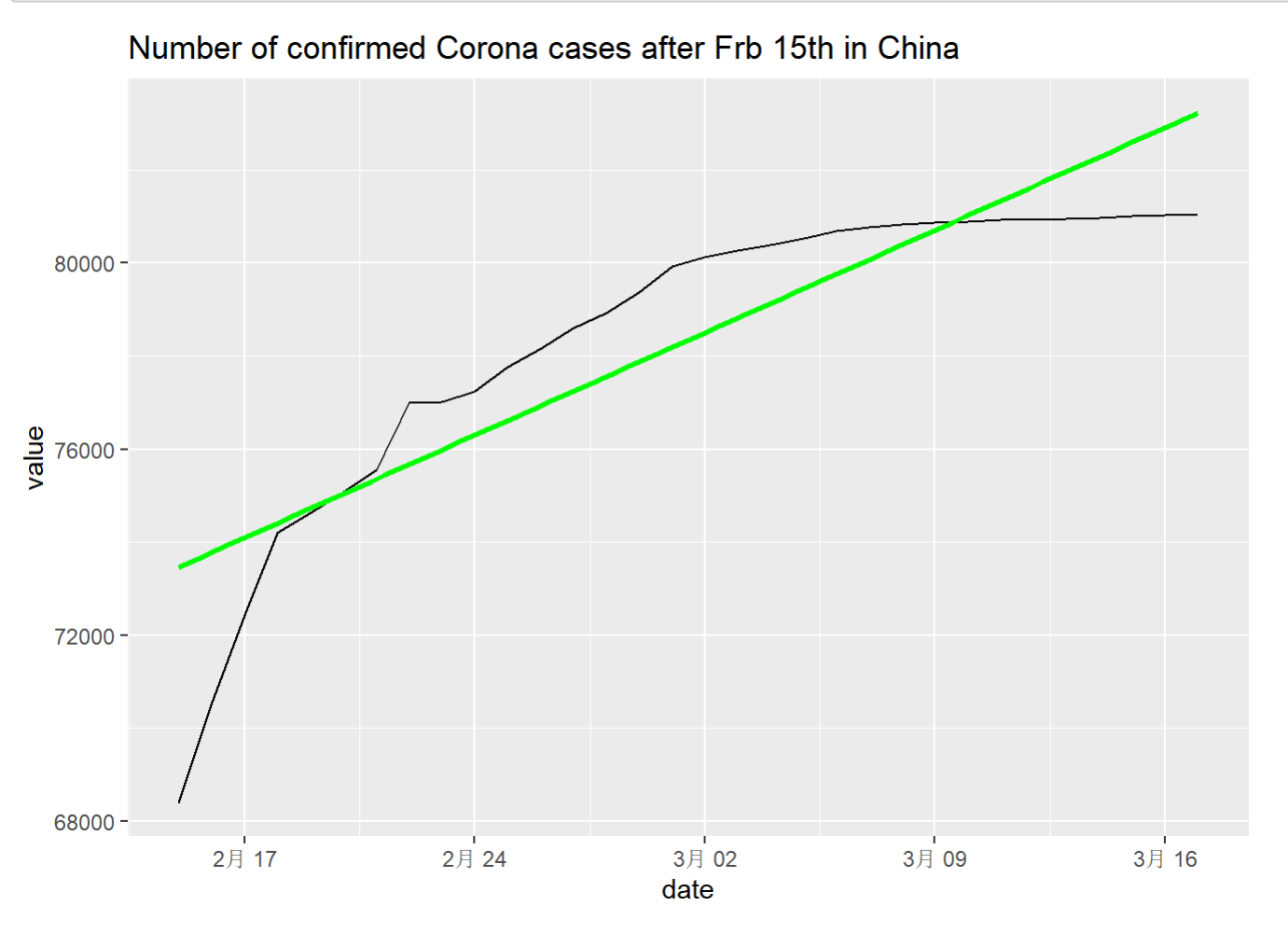


After Feb 15th (China)

Observe the number of cases in China after Feb 15th.

```
china_after_feb15 <- china_vs_world_melt %>%
  filter(variable == "China.cases", date >= as.Date("2020-02-15"))

# glimpse(china_after_feb15)
#
# View(china_vs_world_melt)

plt_china_after_feb15 <- ggplot(china_after_feb15, aes(x = date, y = value)) +
  geom_line()+
  geom_smooth(method = "lm", se = FALSE, color = "green") +
  labs(title = "Number of confirmed Corona cases after Frb 15th in China")

plt_china_after_feb15
```

```
## `geom_smooth()` using formula 'y ~ x'
```



From the plot above, the growth rate in China is slower than linear. That's great news because it indicates China has at least somewhat contained the virus in late February and early March.

How does the rest of the world compare to linear growth?

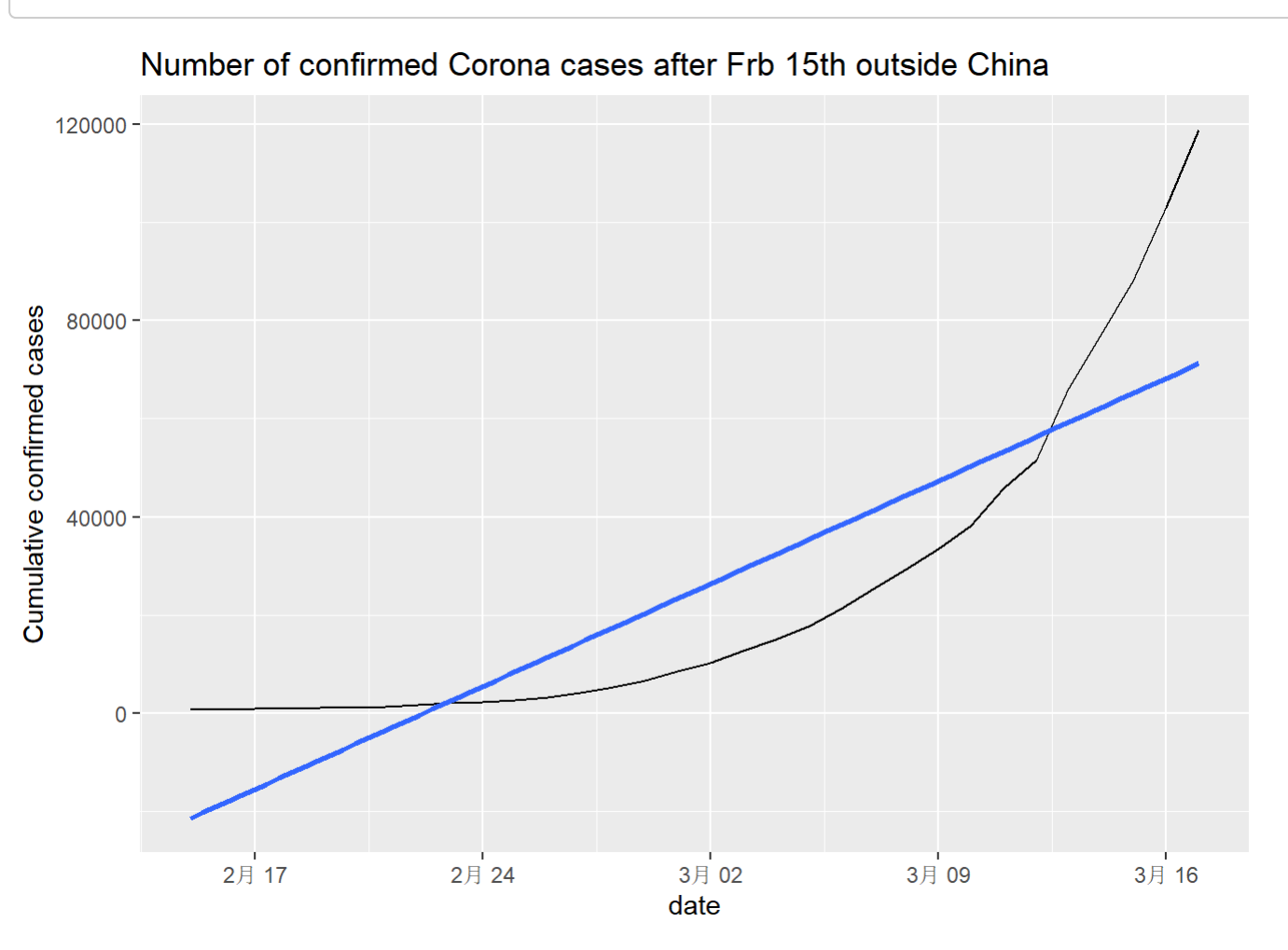
After Feb 15th (Other countries outside China)

```
non_china_after_feb15 <- china_vs_world_melt %>%
  filter(date >= "2020-02-15", variable == "others.cases")


glimpse(non_china_after_feb15)
```

```
## Rows: 32
## Columns: 3
## $ date     <date> 2020-02-15, 2020-02-16, 2020-02-17, 2020-02-18, 2020-02-19, 2020-02-20, 2020-02-21, 2020-02-
22, 20~
## $ variable <fct> others.cases, others.cases, others.cases, others.cases, others.cases, others.cases, others.ca
ses, o~
## $ value    <dbl> 639, 722, 836, 941, 1033, 1135, 1291, 1601, 1960, 2305, 2646, 3211, 4137, 5195, 6658, 8464, 1
0246, ~
```

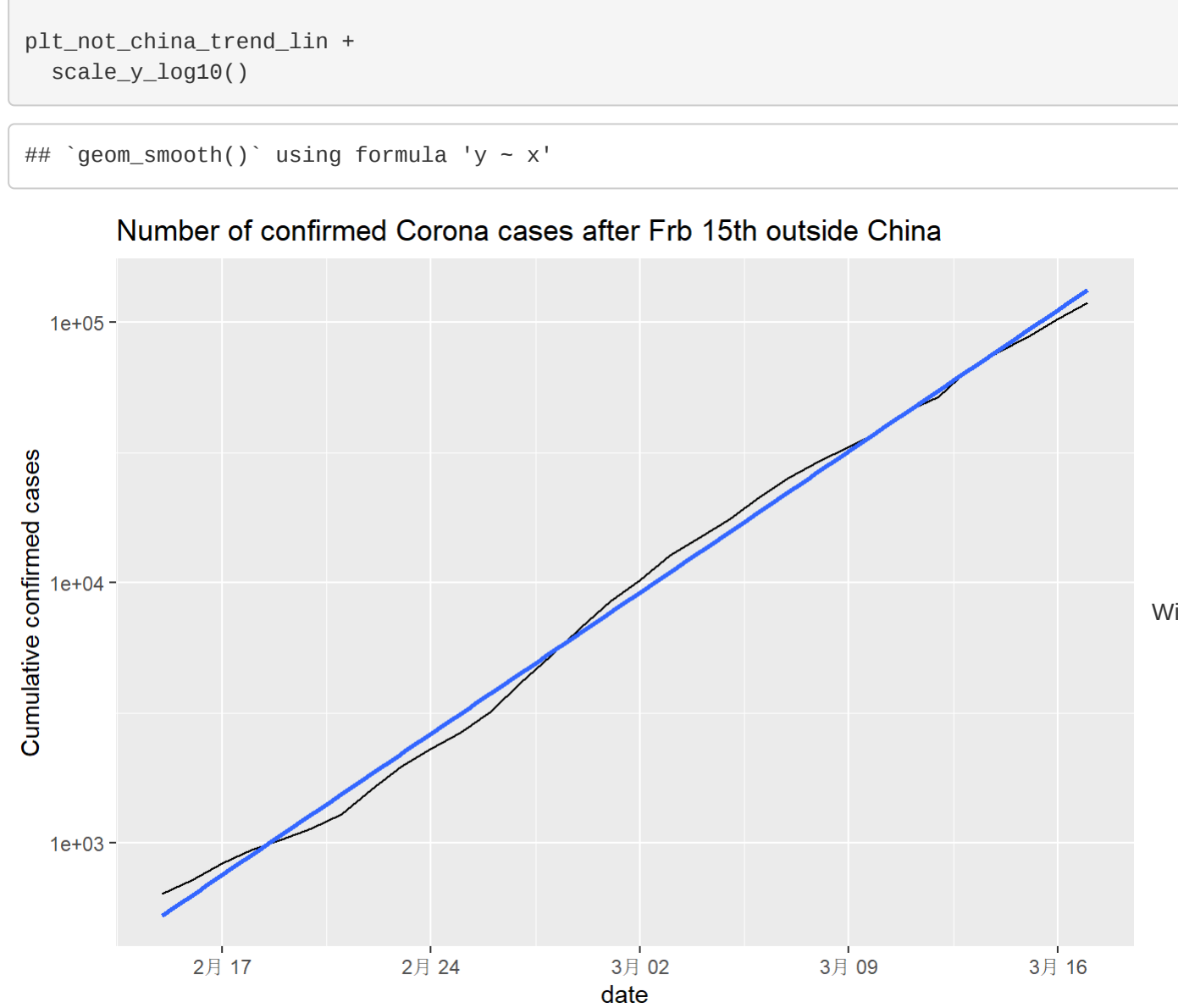
```
plt_not_china_trend_lin <- ggplot(data = non_china_after_feb15, aes(x= date, y = value)) +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Number of confirmed Corona cases after Frb 15th outside China", y = "Cumulative confirmed cases")

plt_not_china_trend_lin
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# From the plot above, we can see a straight line does not fit well at all, and the rest of the world is growing
 much faster than linearly.
# What if we added a logarithmic scale to the y-axis?

plt_not_china_trend_lin +
  scale_y_log10()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



With the logarithmic scale, we get a much closer fit to the data. From a data science point of view, a good fit is great news. Unfortunately, from a public health point of view, that means that cases of COVID-19 in the rest of the world are growing at an exponential rate, which is terrible news.

Countries hit the most by Covid by Mid March

Plot which country was hit the most by Covid outside China by the Mid March

```
# glimpse(confirmed_cases)
# confirmed_cases_country

top_countries_by_total_cases_perday <- confirmed_cases_country %>%
  filter(country != "China",date <= as.Date('2020-03-17'), date > as.Date('2020-02-15'))

# head(top_countries_by_total_cases_perday)

top_countries_by_total_cases <- confirmed_cases_country %>%
  filter(country != "China",date <= as.Date('2020-03-17'), date > as.Date('2020-02-15')) %>%
  group_by(country) %>%
  summarise(sum = sum(cases))%>%
  arrange(-sum)%>%
  top_n(7)
```

```
## Selecting by sum
```

```
top_countries_by_total_cases
```

country	sum
<chr>	<dbl>
Italy	31503
Iran	16169
Spain	11746
Germany	9241
Korea, South	8292
France	7703
US	6498

7 rows

```
# Using confirmed_cases_top7_outside_china, draw a line plot of
# cum_cases vs. date, colored by country

top_seven <- top_countries_by_total_cases_perday %>% filter(country %in% pull(top_countries_by_total_cases,countr
y))

glimpse(top_seven)
```

```
## Rows: 217
## Columns: 4
## Groups: country [7]
## $ country   <chr> "France", "France", "France", "France", "France", "France", "France", "France", "France", "F
rance"~
## $ date      <date> 2020-02-16, 2020-02-17, 2020-02-18, 2020-02-19, 2020-02-20, 2020-02-21, 2020-02-22, 2020-02
-23, 2~
## $ cases     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 4, 20, 19, 43, 30, 61, 21, 76, 138, 190, 332, 177, 286, 372, 5
10, 0,~
## $ cum_cases <dbl> 12, 12, 12, 12, 12, 12, 12, 12, 12, 14, 18, 38, 57, 100, 130, 191, 212, 288, 426, 616, 948,
1125, ~
```

```
plt_top_countries <- ggplot(top_seven) +
  geom_line(aes(x= date, y = cum_cases, color = country)) + labs(title = "Top seven countries hit the most by Cov
id by Mid March", y = "Confirmed cases per day") +
  geom_text(data = top_seven %>% filter(date == max(date)), aes(label = country,x = date + 1,y = cum_cases,color
 = country),size = 2)



plt_top_countries
```