Visualizing COVID-19 datacamp project

Essraa A.

15/12/2020

1. From epidemic to pandemic¶

In December 2019, COVID-19 coronavirus was first identified in the Wuhan region of China. By March 11, 2020, the World Health Organization (WHO) categorized the COVID-19 outbreak as a pandemic. A lot has happened in the months in between with major outbreaks in Iran, South Korea, and Italy.

We know that COVID-19 spreads through respiratory droplets, such as through coughing, sneezing, or speaking. But, how quickly did the virus spread across the globe? And, can we see any effect from country-wide policies, like shutdowns and quarantines?

Fortunately, organizations around the world have been collecting data so that governments can monitor and learn from this pandemic. Notably, the Johns Hopkins University Center for Systems Science and Engineering created a publicly available data repository to consolidate this data from sources like the WHO, the Centers for Disease Control and Prevention (CDC), and the Ministry of Health from multiple countries.

In this notebook, we'll visualize COVID-19 data from the first several weeks of the outbreak to see at what point this virus became a global pandemic.

Please note that information and data regarding COVID-19 is frequently being updated. The data used in this project was pulled on March 17, 2020, and should not be considered to be the most up to date data available

Load packages

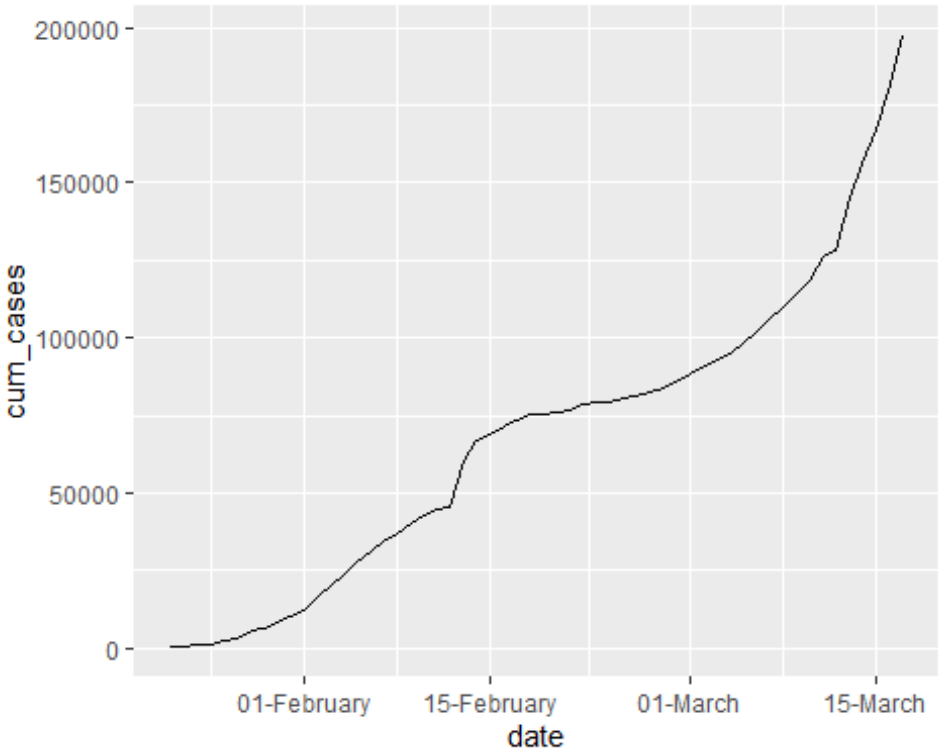
```
library(tidyverse)

library(ggplot2)
library(lubridate)
```

Read confirmed_cases_worldwide.csv into confirmed_cases_worldwide and show the data

```
##         date cum_cases
## 1 2020-01-22       555
## 2 2020-01-23       653
## 3 2020-01-24       941
## 4 2020-01-25      1434
## 5 2020-01-26      2118
```

2. Confirmed cases throughout the world The table above shows the cumulative confirmed cases of COVID-19 worldwide by date. Just reading numbers in a table makes it hard to get a sense of the scale and growth of the outbreak. Let's draw a line plot to visualize the confirmed cases worldwide.

```
plt_confirmed_cases_worldwide
```



3. China compared to the rest of the world¶ The y-axis in that plot is pretty scary, with the total number of confirmed cases around the world approaching 200,000. Beyond that, some weird things are happening: there is an odd jump in mid February, then the rate of new cases slows down for a while, then speeds up again in March. We need to dig deeper to see what is happening.

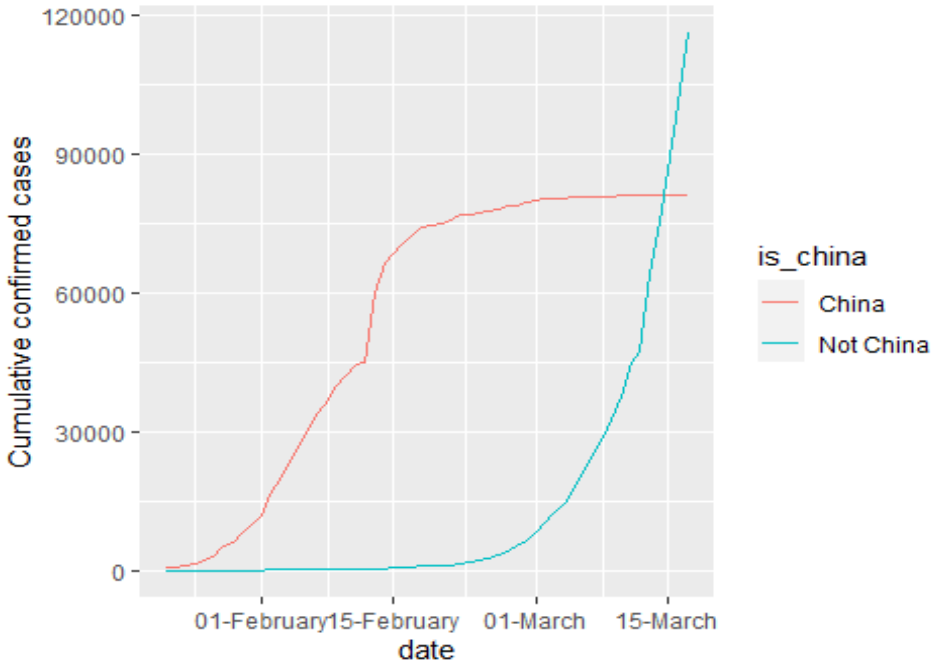
Early on in the outbreak, the COVID-19 cases were primarily centered in China. Let's plot confirmed COVID-19 cases in China and the rest of the world separately to see if it gives us any insight.

We'll build on this plot in future tasks. One thing that will be important for the following tasks is that you add aesthetics within the line geometry of your ggplot, rather than making them global aesthetics.

```
Read confirmed_cases_china_vs_world.csv and show the data

##   is_china       date cases cum_cases
## 1    China 2020-01-22   548       548
## 2    China 2020-01-23    95       643
## 3    China 2020-01-24   277       920
## 4    China 2020-01-25   486      1406
## 5    China 2020-01-26   669      2075
```

Draw a line plot of cumulative cases vs. date, grouped and colored by is_china

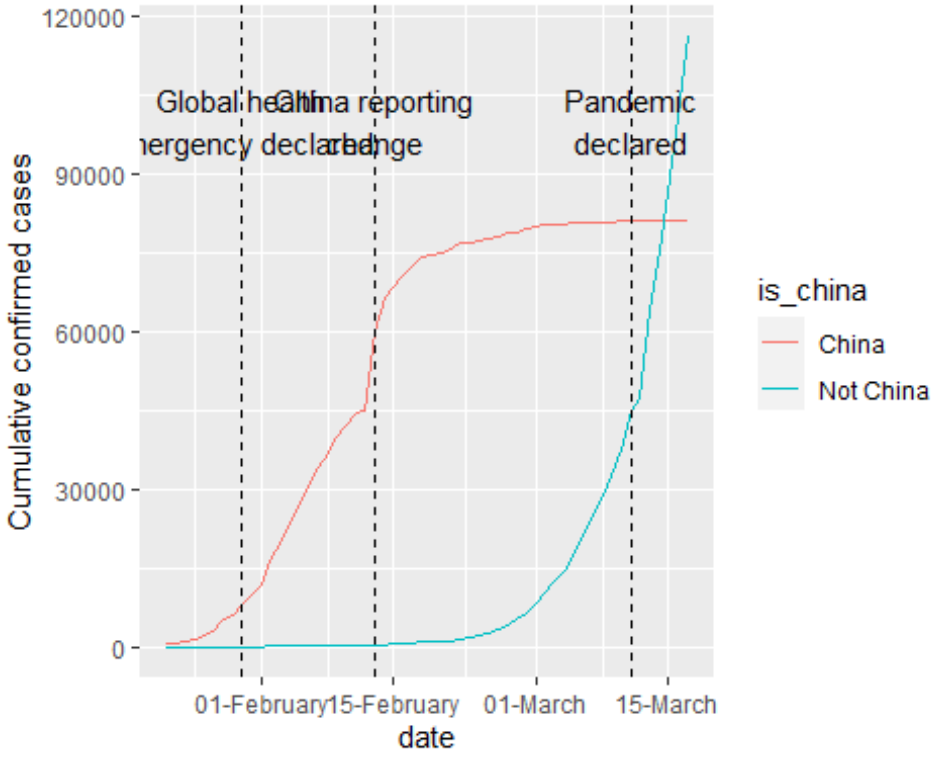
```
plt_cum_confirmed_cases_china_vs_world
```



4. Let's annotate! Wow! The two lines have very different shapes. 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.

There were a couple of other landmark events that happened during the outbreak. For example, the huge jump in the China line on February 13, 2020 wasn't just a bad day regarding the outbreak; China changed the way it reported figures on that day (CT scans were accepted as evidence for COVID-19, rather than only lab tests).

By annotating events like this, we can better interpret changes in the plot

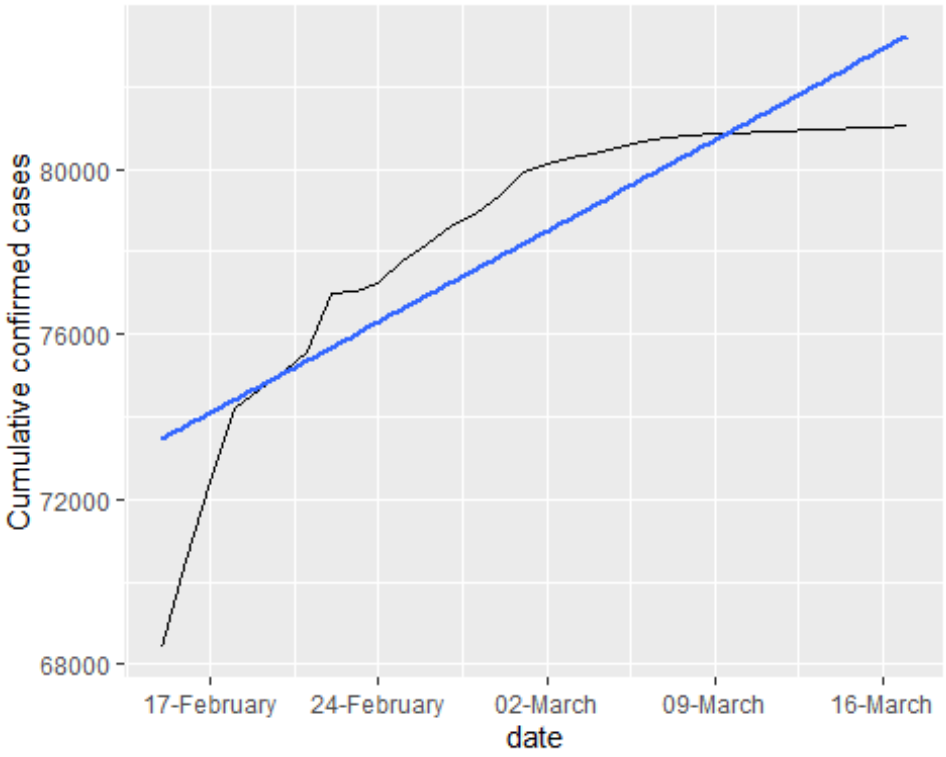


5. Adding a trend line to China When trying to assess how big future problems are going to be, we need a measure of how fast the number of cases is growing. A good starting point is to see if the cases are growing faster or slower than linearly.

There is a clear surge of cases around February 13, 2020, with the reporting change in China. However, a couple of days after, the growth of cases in China slows down. How can we describe COVID-19's growth in China after February 15, 2020?

```
##   is_china       date cases cum_cases
## 1    China 2020-02-15  2055     68413
## 2    China 2020-02-16  2100     70513
## 3    China 2020-02-17  1921     72434
## 4    China 2020-02-18  1777     74211
## 5    China 2020-02-19   408     74619
```

```
plt_china_after_feb15
```

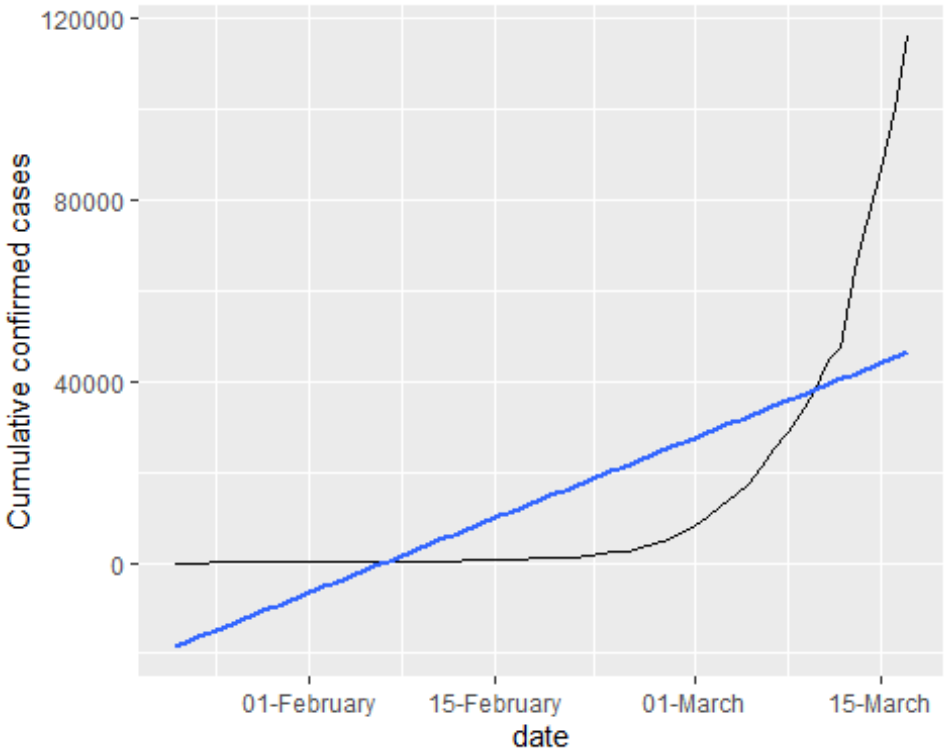


6. And the rest of the world? From the plot above, the growth rate in China is slower than linear. That's great news because it indicates China has at least somewhat contained the virus in late February and early March.

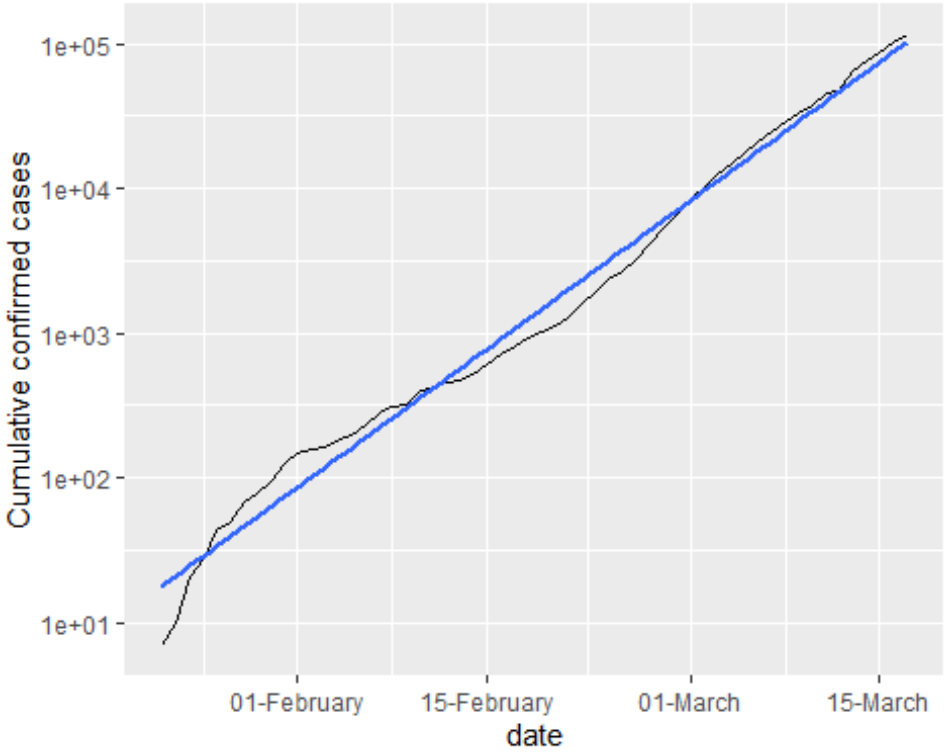
How does the rest of the world compare to linear growth?

```
##    is_china       date cases cum_cases
## 1 Not China 2020-01-22     7         7
## 2 Not China 2020-01-23     3        10
## 3 Not China 2020-01-24    11        21
## 4 Not China 2020-01-25     7        28
## 5 Not China 2020-01-26    15        43
```

```
plt_not_china_trend_lin
```



7. Adding a logarithmic scale 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?



8. Which countries outside of China have been hit hardest? 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.

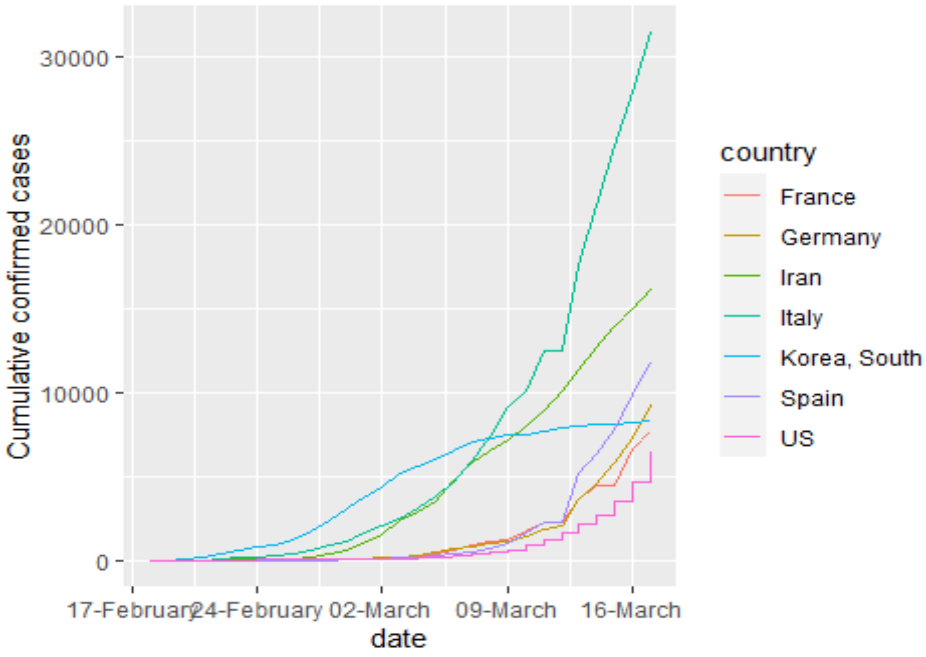
Not all countries are being affected by COVID-19 equally, and it would be helpful to know where in the world the problems are greatest. Let's find the countries outside of China with the most confirmed cases in our dataset.

```
Read confirmed_cases_by_country.csv and show data

##                country province       date cases cum_cases
## 1          Afghanistan          2020-01-22     0         0
## 2              Albania          2020-01-22     0         0
## 3              Algeria          2020-01-22     0         0
## 4              Andorra          2020-01-22     0         0
## 5 Antigua and Barbuda          2020-01-22     0         0

top_countries_by_total_cases

## # A tibble: 5 x 2
##   country      total_cases
##   <fct>              <int>
## 1 France              7699
## 2 Germany             9257
## 3 Iran               16169
## 4 Italy              31506
## 5 Korea, South        8320
```

9. Plotting hardest hit countries as of Mid-March 2020 Even though the outbreak was first identified in China, there is only one country from East Asia (South Korea) in the above table. Four of the listed countries (France, Germany, Italy, and Spain) are in Europe and share borders. To get more context, we can plot these countries' confirmed cases over time.

Finally, congratulations on getting to the last step! If you would like to continue making visualizations or find the hardest hit countries as of today, you can do your own analyses with the latest data available here.

```
Read confirmed_cases_top7_outside_china.csv and show data

##         country       date cum_cases
## 1       Germany 2020-02-18        16
## 2          Iran 2020-02-18         0
## 3         Italy 2020-02-18         3
## 4 Korea, South 2020-02-18        31
## 5         Spain 2020-02-18         2



plt_confirmed_cases_top7_outside_china
```


30000
20000
10000
0
Cumulative confirmed cases
17-February
24-February
02-March
09-March
16-March
date
country
France
Germany
Iran
Italy
Korea, South
Spain
US