Covid-19 data analysis

2024-09-11

Data source

The data is found on this page: https://github.com/CSSEGISandData/COVID-19

Quoted from the page:

```
This is the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by
the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also,
Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab
(JHU APL).
```

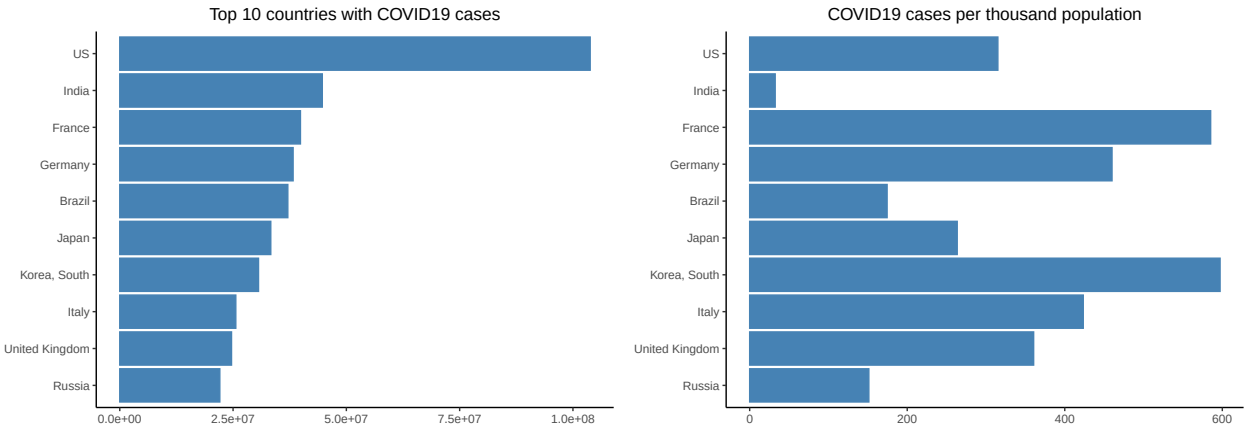
Question of interest

COVID19 has affected most of the people in the world, and different countries take different actions to deal with it. To some degree, the "cases per population" and "deaths per population" are indicators to tell how well it is controlled.

In this report, I am going to compare the top cases/deaths countries, and look especially into the data in China, since it is a country with a big population and it has a very strict policy on controlling cases.

Compare cases between countries

First, I will select the top 10 countries with the most cases, and plot cases per thousand population.

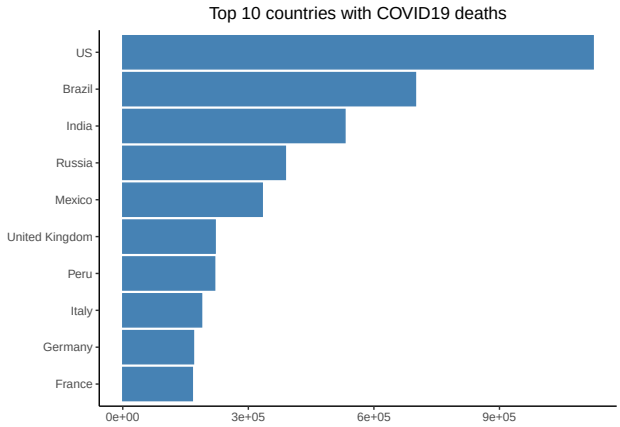


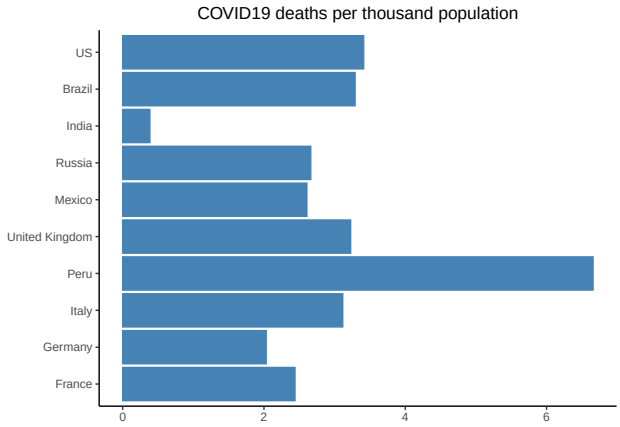
From the plot we can see that in terms of absolute number of cases, the US is the top 1, and it has more than twice as many cases as the second place, which is India. And other countries have relatively similar cases.

When we check the "cases per thousand" plot, we can see the US is similar to other countries. So from this point of view, the US is not doing that badly. And India has a very low value, further analysis would be needed to understand why.

Compare deaths between countries

Second, I will select the top 10 countries with the most deaths, and plot the deaths per thousand population.



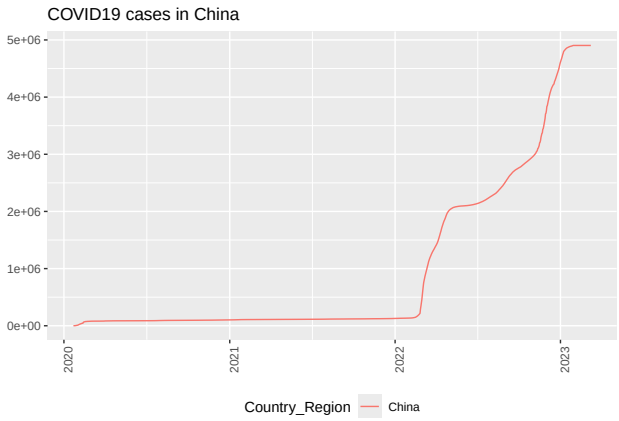


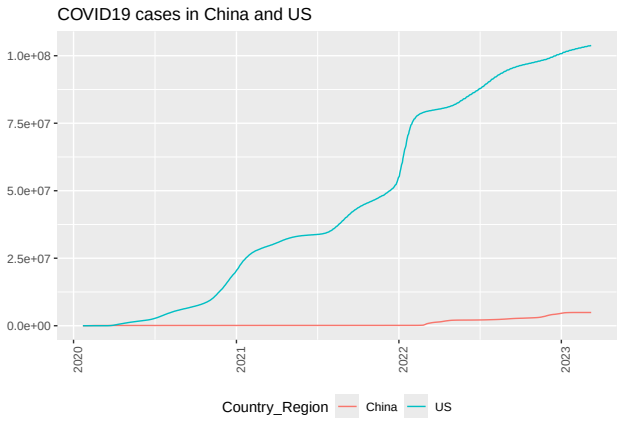
In terms of the death value, the US is still the top 1 in absolute value, but checking the "deaths per Thousand" plot, it is among the average countries. Peru has many deaths about twice as high as the other countries, and India again has a very low value. Further analysis is required to understand the reason.

How is China doing?

As we all know, China has a very large population, and has suffered from COVID19 since the very beginning. However, it is not listed in the top 10 countries in terms of total cases. Let's take a look into more details by comparing it with the US data.

First let's plot the cases in the time series.





From the first plot we can see the cases in China are kept very low till March 2022. The reason is that China has imposed a strict "Dynamic Clearing" policy to keep cases down. There are two outbreaks in the plot:

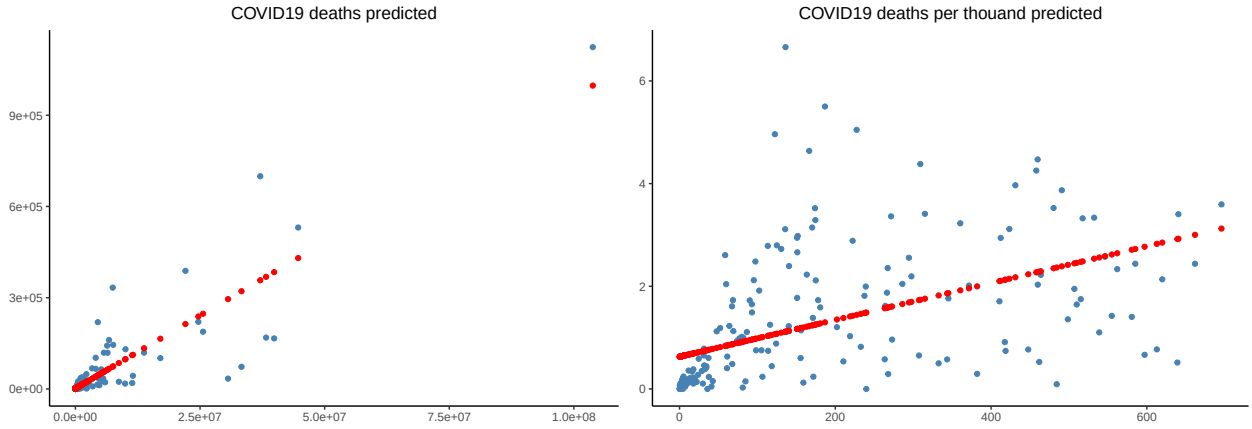
1. The outbreak around March 2022 was due to the new "Omicron variant".

2. The outbreak around December 2023 was because China stopped the "Dynamic Clearing" policy.

Comparing with the US cases in the second plot, we can see the difference in cases is quite obvious, the strict policy in China keeps the total cases much lower. And from this point of view, China is doing better at controlling the cases.

Data model

In this section, I will use a model to test the hypothesis that, globally, the deaths are linear to cases, and deaths_per_thou are linear to cases_per_thou.



As we can see, both models are somewhat fit with the data but with big residuals. Which means if we predict the deaths only from cases, it is not quite accurate. I think we need to consider more factors to the deaths, e.g. number of hospitals, economic conditions, age, etc.

Summary and conclusion

In this report I utilized the COVID19 data from Johns Hopkins University to do some analysis. Compared the cases and deaths of the top countries, took a close look at the China data and finally created two models to analyze the relationship between cases and deaths.

The analysis raised some uncertain questions which would require further analysis to answer.

Some bias might exist in the data source in where the data come from, and how they are collected. Some results might also be biased that, not every aspect was considered. For example, cases are well controlled in China, which is good, but the side effect of that is not take in to account.