University of Nottingham
UK | CHINA | MALAYSIA

COMP3021
Fundamentals of Information Visualisation

Data Visualization & Analysis Report

Covid 19 Dataset

March 11, 2022

Individual Coursework

Name	Student ID	Username
Teo Shi Bin	20183717	hcyst2

Contents

List of Figures

Introduction

This report will go through PNG plots, GIF animations and interactive HTML plots. Some graphs are generated as a prove of concept for how data can be presented in different ways and more importantly data analysis via visualization. This project contains a lot of interactive and animated aspect which are hard to present within the report, please do give it a try to get the full grasp of it. All plots are pre-generated and stored in a folder for the convenience.

Dataset

The datasets used here are retrieved from kaggle.

- Covid-19 Global Dataset
- COVID-19 World Vaccination Progress

Implementation

Folder Structure

```
.
└── covid19_analysis/
    ├── Data/
    ├── Function/
    ├── Output/
    │   ├── HTML/
    │   ├── GIF/
    │   ├── PNG/
    │   └── Screenshot/
    ├── covid19_analysis.Rproj
    ├── main.Rmd
    ├── ANALYSIS.md
    ├── README.md
    └── ...
```

`Data` Input dataset
`Function` Auxiliary functions
`Output` Contains different type of generated outputs
`covid19_analysis.Rproj` Main R project file
`main.Rmd` Main markdown script
`ANALYSIS.md` Quick presentation of all generated outputs
`README.md` Read me

Dependencies

A package called "renv" is used within the project to manage package dependencies, run "renv::restore()" to install all relevant packages specified within "renv.lock" file. "pacman" package is used to load and unload packages, packages that have yet to be installed will be installed and loaded automatically during execution.

Data Importing & Cleaning

Simple NA value replacements are applied to the imported dataset.

Data Visualization & Analysis

General Questions

What is the proportion of recovered cases?

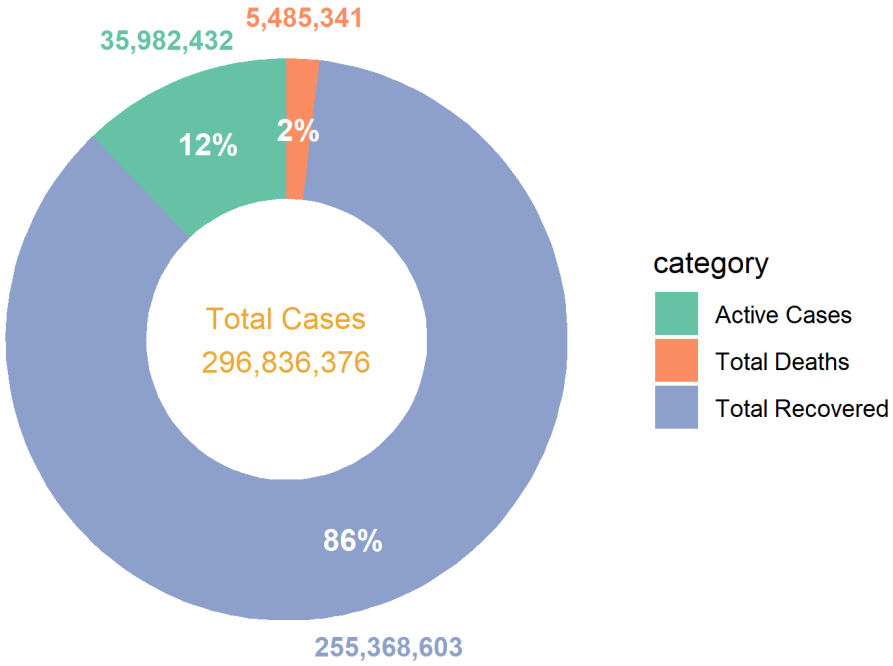


Figure 1: Overview of Covid cases

I selected a simple donut chart to visualize the categories of different Covid cases. Along with some simple math using dplyr to compute the percentage of each categories, the plot is generated easily. We can see that most of them have recovered and a low death rate out of all the cases.

What is the scale of infected population in different continent or country?

Figure 2: Scale of Covid cases in different continents interactive plot

(a) Layer 2 Countries

(b) Layer 3 Categories

Figure 3: Sub-layers encoding

Interactive Tree plot is selected to visualize the proportion of Covid cases in different continent and countries. There are 3 layers of encoding within the plot, we can zoom in 3 layers by clicking on it. 3 layers include continents, countries and Covid categories (categories as in Figure 1). It turns out that most of the cases happen in Europe, Asia and North America and this is not really a surprise.

Notice that groups with smaller values will converge to the bottom right corner of the area and when that happens it will turn into a black ball of text that is not readable. My Solution to this is to filter values that are too small out and the way I'm filtering them is using percentile. However, an extra step is applied so that percentile filtering is done within each continent, therefore preserving smaller continent groups such as the tiny blue continent in the first layer. A reasonable filtering percent value is chosen for the best possible aesthetic and readability while preserving most of the groups and data.

How many people suffered from covid?

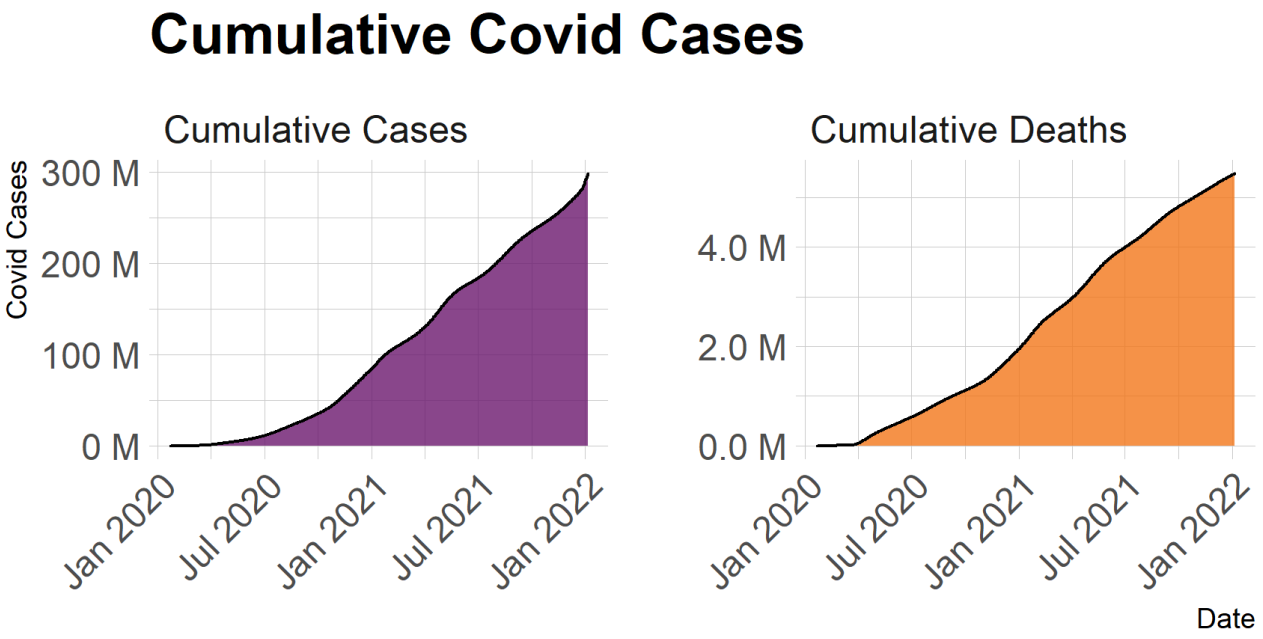


Figure 4: Cumulative cases & Cumulative death cases interactive plot

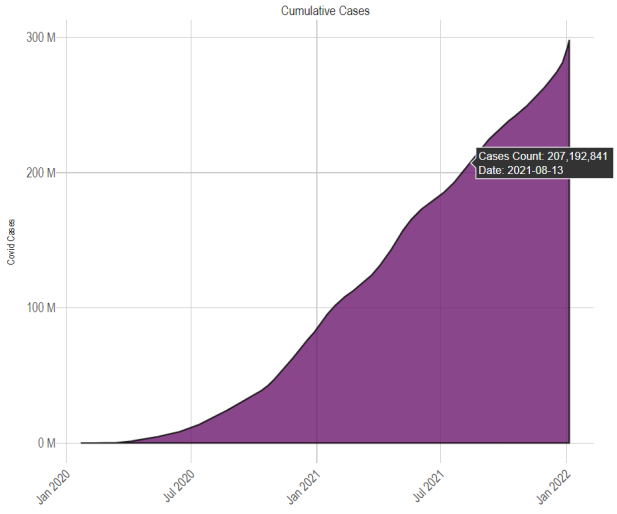


Figure 5: Hover Tool-tip

One way of presenting groups of cumulative values is to use stacked area plot. The interactive version of this plot will display the details on hover (Figure 5). We can clearly see that up until January of 2022, we have accumulated over 300 Millions of cases and nearly 6 Millions of death globally.

One decision that I have made was to utilize small multiples (Implemented using ggplot facet wrap) in order to avoid the unbalance scaling of Y axis. The main reason to this is because that death count is far smaller than accumulated cases, thus, splitting them into smaller sub plots will be more visually effective to analyse. The creation of this plot requires a lot more knowledge in data manipulation to achieve this, some advance functions such as “gather”, “ group by” and “summarize” are used.

Which country has the most active cases?

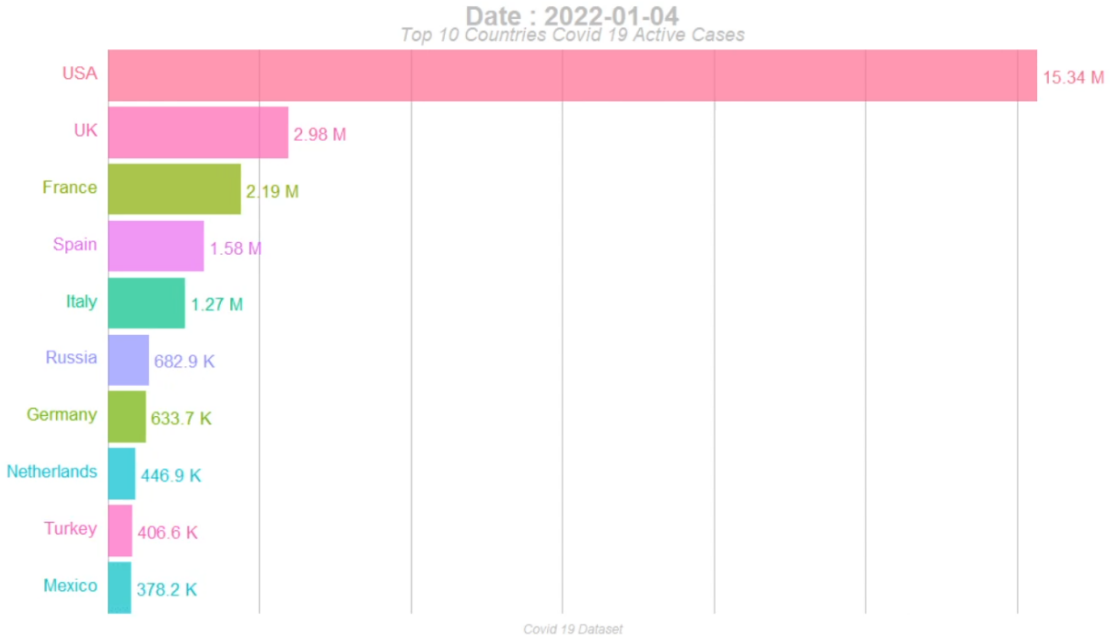


Figure 6: Active Cases Country Ranking GIF

This is an animated GIF that animates the ranking of active cases starting from Jan 2020 to Jan 2022. The reason that I have selected bar chart is to allow us to compare between countries easily. This is done using ggplot along with gganimate. Active cases visualized in a more interesting way then standard line plot. The answer for the question above is fairly obvious. I also notice quite a few spikes in different countries and that will be something to be discussed in further section.

The animation of this is relatively easy due to the fact that the package has abstracted a lot of technical stuff from us. The only thing we need to deal with is to rank active cases by countries and it can all be done using rank function. Some details that you can find is that I have round up different values in different scales and in different decimal points The rest of the work is just tweaking themes and transitions. Here's an image of the bar performing it's transition (Figure 7).

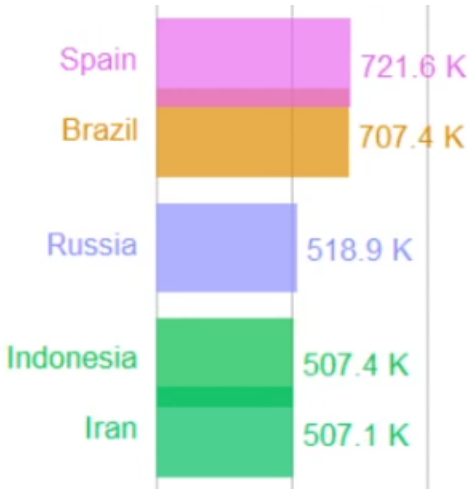


Figure 7: Transitions between frames

How active cases evolve around the world?

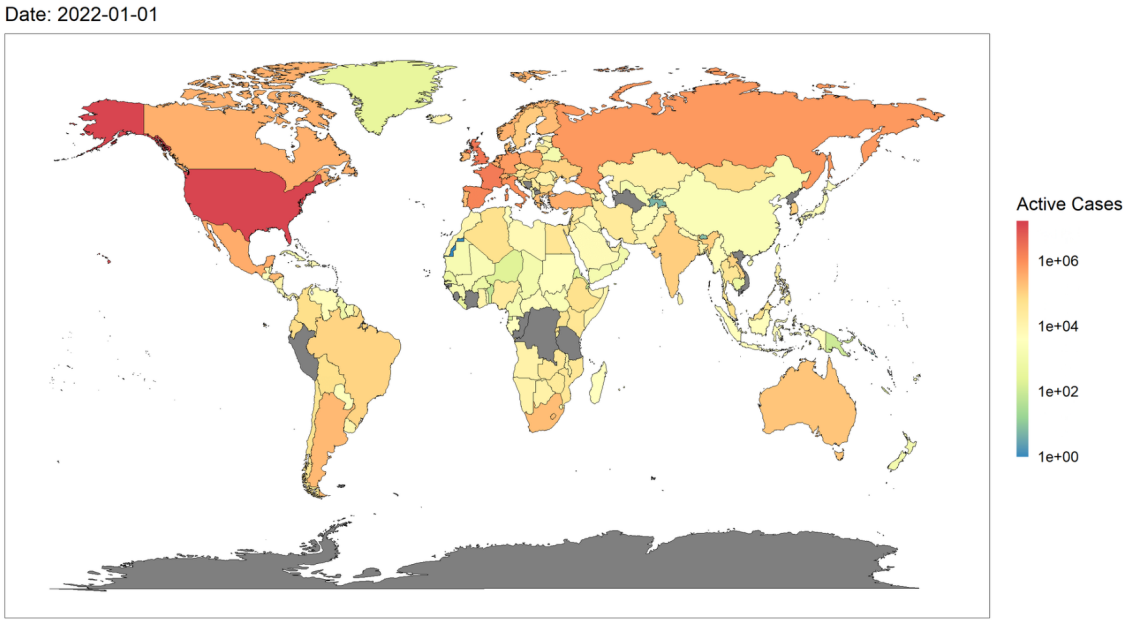
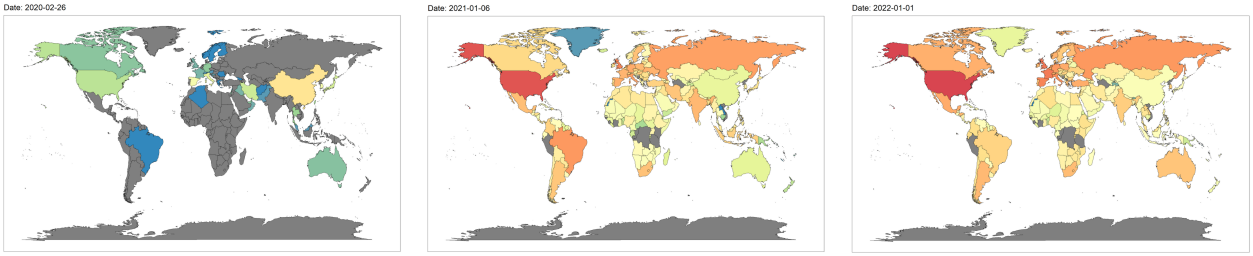


Figure 8: Active Cases World Map GIF



(a) Feb 2020 Map (b) Jan 2021 Map (c) Jan 2022 Map

Figure 9: Maps in Different Dates

This is another cool way of visualizing active cases using GIF. The main reason that I selected map as the plot is to visualize all the countries including those with less Covid cases, this way we get the benefit of plotting everything, unlike top ranking bar chart. We can clearly see that countries with high population are having a hard time fighting this pandemic, but with some exception such as China, they are doing pretty well. Notice that the color scale here represents the log10 scale of the actual value, meaning it is 100 times harder to change color from one to another, this truly emphasise that some of the countries did really well in defending the virus.

The Covid dataset is first joined with the map dataset which defines the boundaries of each countries. Right join (depending on the order of the arguments) is being used here in order to preserve all the country boundaries from the map. When the data is plotted, countries with na daily cases will be colored as grey, otherwise color gradient.

Finally, animating this GIF is not as straight forward as previous figure (Figure 6), the reason to this is that I can't get gganimate to work for this massive joined map data. Therefore, the alternative solution to this is to create all the necessary images on my own using for loop and stitch them together using gifski package to form a GIF at the end.

In depth Questions

Is there any relation between population and Covid cases?

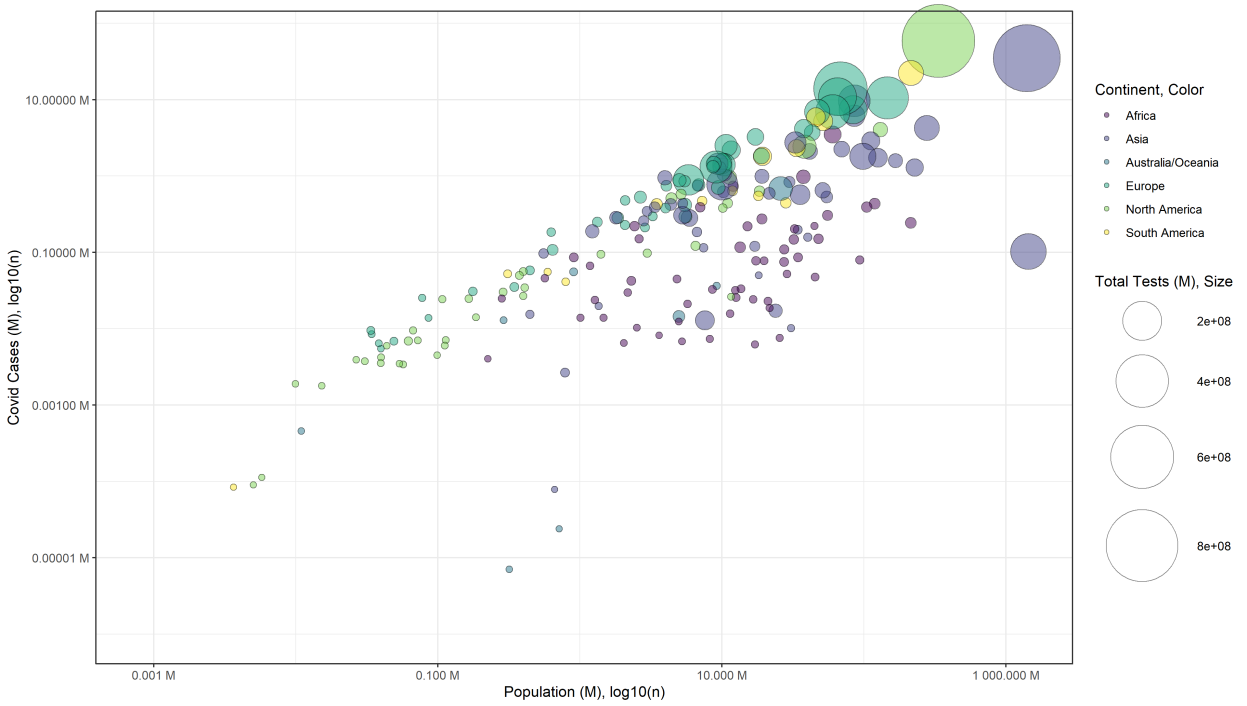


Figure 10: Covid Cases vs Population plot

After going through bar ranking and map GIFs, I have noticed that most of the cases are spreading within high population countries and that leads me to question the relation between cases and population. I have selected bubble plot to visualize the relation for population, total confirmed Covid cases, PCR tests and 1 extra continent color encoding. We can clearly see that, there is a positive correlation between these 2 variables as the values grow larger, it does the same for the other variable. Apart from that, the bubble size denotes the number of PCR tests done in that country and there is a pattern where it is getting larger towards the top right corner. In summary, we can say that all these 3 variables, cases, tests and population are all positively correlated. However, we do not know the dependencies between variables, we do not know which causes which to grow higher, and since the growth of population does not depend on Covid cases and in fact higher cases might cause the population to go down instead of up, thus we know that high population can definitely cause high Covid cases or high tests, alternatively we can also say that high tests is returning high Covid cases or vice versa.

The implementation of this plot is quite simple, in fact it is just a scatter plot with 2 extra encoding and some interactive components. The only changes that I've made to the data is to arrange it by descending order of total tests, so that larger bubble will stay behind smaller bubbles.

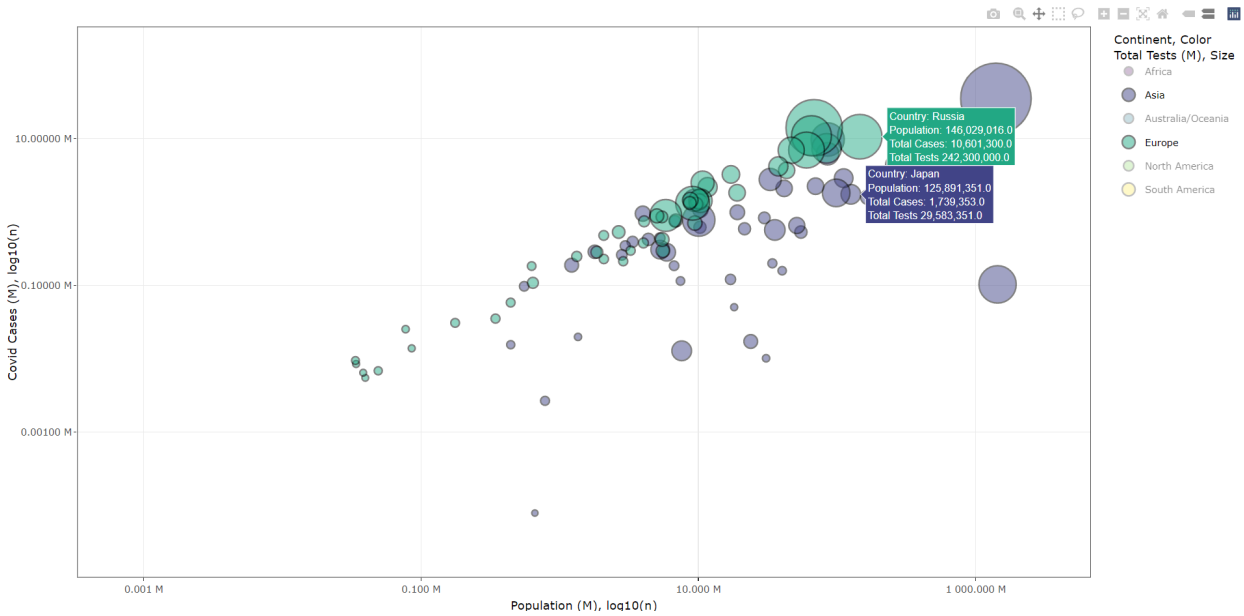


Figure 11: Covid Cases vs Population interactive plot

This is the interactive version of the same plot as you see me hovering my mouse on top these 2 selected continents (the legend) using comparison mode. These tool-tips are comparing to nearby x-axis and we can see that Japan has similar population as Russia but the number of Covid cases are incredibly different. Notice that it is a log scaled plot and this is due to the high number difference in population and also Covid cases which makes it hard to read when plotted without using it. If I were to plot it without using log scale, all the bubbles will concentrate to the left bottom corner while a few stay at the top right corner.

How vaccination affect the spread of covid?

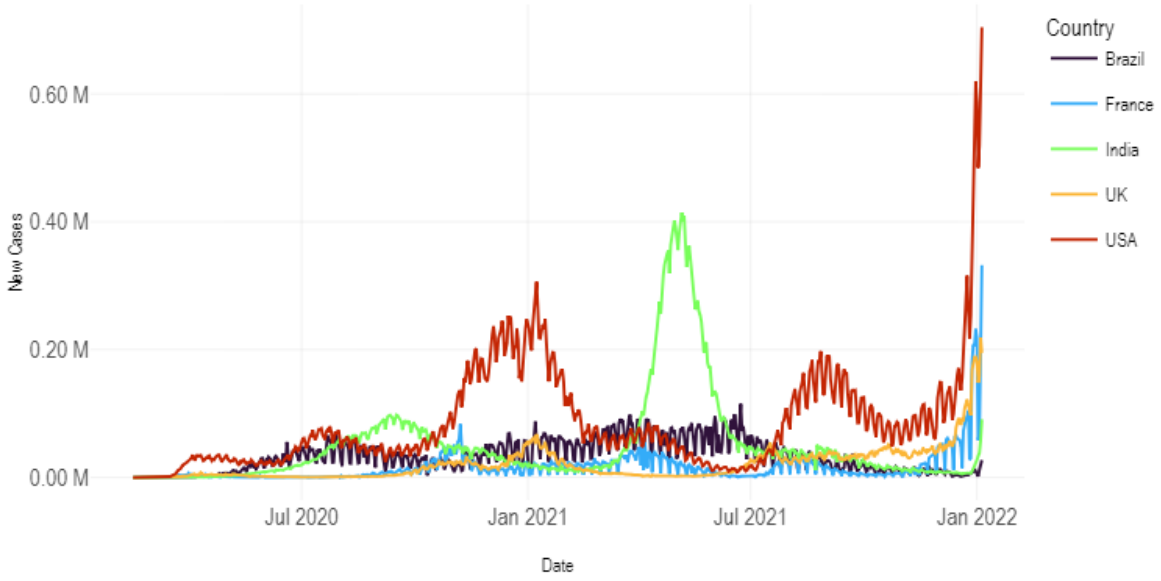


Figure 12: Top 5 variance of group by country daily new case interactive plot

This question is derived from the previous bar ranking chart, where there are some huge spike incline and decline in some countries. I figure that something is preventing it from happening, is it city lockdown or vaccination count? However, I can't find reliable dataset of lockdown dates for all countries, so I'll go towards the vaccination path. Now, let's find a country that we want to focus on using daily new cases (Fig12).

This interactive plot is generated using top 5 variance of group by country daily new cases. I have used variance to filter it so that the plot will display countries that have large ranging daily new cases. Now, countries with huge cases spikes show up. India has a really odd looking spike there, focusing into this part along with some vaccination dataset. (Next Page)

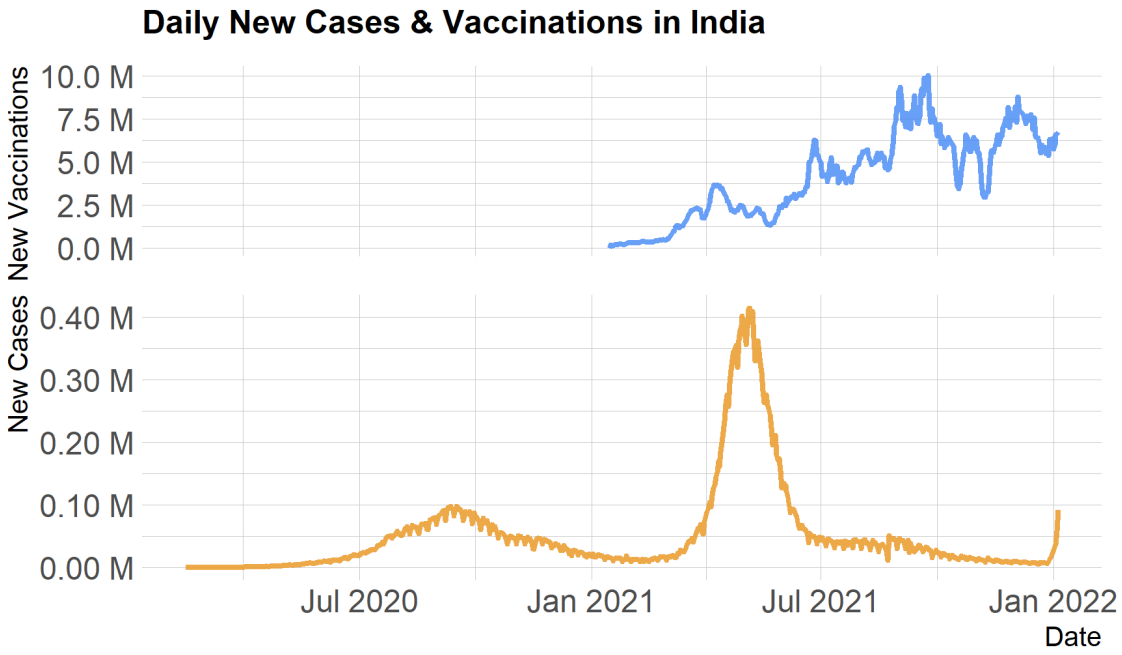


Figure 13: India vaccination comparison plot

Now, we start to see some possible reason of why daily new cases went all the way down. This won't explain everything as lockdown might impact the spread of Covid even more but this is still a good evidence that as India sky rocketed their daily new vaccinations the daily new cases have went down.

The idea behind this plot is that the y-axis suffers from 2 variable that have numbers that are too far apart, making it hard to scale. In the other hand, I did not use log scale for this as we need to see the true scale of the image in order to compare them properly without the interference of log scaling. Therefore, I have stack 2 plots with different Y-axis scale vertically for better comparison by aligning X-axis. One tiny detail here is that I have removed the X-axis of the first plot as they share the same x variable.

Conclusion

In summary, we have gone through the proportion of active cases, recovered cases and death cases as an overview then we take a look at the scale of Covid cases within different continent. After that, we go through the global cumulative Covid cases to see how many people have suffered from the pandemic. In addition, I have introduced 2 GIF plots to visualize active cases in a more interesting manner, one by bar ranking and the other by map. Next, I questioned more about the pattern that we saw in those charts and try to seek for answers or explanations. We have observed the positive correlation between Covid cases over population and then the weird daily cases spike in India against vaccination counts. There are still a lot more to be analysed within this dataset, but as the report grows longer I will stop here.

Extra Information

Demo Video

Quick demo is demonstrated here.
https://youtu.be/3a8Gk9zPDjM

Repository

Further detail is available in README of this repository.
https://github.com/teoshibin/COMP3021_FIV_covid19_analysis