Dear Professor YY,

It's been two months since I started working on your COVID-19 data visualization project. I've updated data, adjusted the scales, and made improvements based on your earlier suggestions. In this email, I'll briefly explain what changes I made, how I manipulated the data, some issues I am facing, and ask for some advice for future directions.

The email is quite long. I am so sorry for that.

The most updated visualization is here. The three updated & improved Observable notebooks are here: Spreading trends, Confirmed vs. New cases, and Case fatality rate.

Here is the updated covid-19 data repo, and here is the covid-19 dash-board repo.

Data side

I manipulated the original data directly without going through the covid-19 data repo, as you suggested. I almost did it. Here is the notebook detailing how I manipulated the data on ObservableHQ. The most updated dataset from Our World in Data (OWID) already contains country codes and population size, so I am directly using them.

However, OWID dataset does not have the information on country's region classification as you (and the data from World Bank) did. It only specifies in which continent each country resides. I think world regions makes more sense than simply continents. Therefore, I assigned a region to each country using Wrold Bank's data. There are ten countries / areas whose data were not found in the World Bank data. For those, I used data from ISO-3166-Countries-with-Regional-Codes, as the README file has detailed.

The only problem is that in your original visualizations, "ERI" (Eritrea) was labelled as "East & North Africa" in your data, but it was classified as in "Sub-Sahara Africa" in World Bank data. Also, "ERI" is the only country with the label of "East & North Africa". So I changed its label to "Sub-Sahara Africa" and deleted the label of "East & North Africa" in all visualizations.

"BES" was the only one labeled as "America & Caribbean" in your Data. I changed it to "Latin America & Caribbean" according to ISO-3166-Countries-with-Regional-Codes. I then deleted the label of "America & Caribbean" in all visualizations.

I changed the starting date for COVID-19 Confirmed vs. New cases to be Dec. 31th, 2019, because it presented a whole picture of the pandemic from the earliest start. To do this, I modified the `scripts/nyt_state_data.py` such that data for all US states also started from 2019-12-31; otherwise, there will be errors in the visualization.

I didn't know to use Snakemake. Since right now, there are only two Python scripts, I managed to automate the data updating process using a bash script. It's not as elegant but it worked fine.

Visualization side

As you can see in the updated dashboard, I've created a slider for date range specification, and also a slider for play speed. As the number of days gets larger, people might want to speed up, or slow down.

I've also created a delay parameter for the Case Fatality Rate visualizations. It was more complicated than I thought since many things need to be tweaked to accommodate the change. Fortunately, it worked well, at least for now. I set the range to be 0 to 100 days because according to this study, "The time from symptom onset to death ranges from 2 to 8 weeks for COVID-19 (p. 15)". I will shorten it if you think it's too broad.

I didn't implement "more mouse-over annotations". In fact, I didn't know how to decide which countries should have the annotation of "The actual case number is likely much bigger" and which ones have "Likely more deaths in the future". Also, ten months into the pandemic, these annotations might not be necessary for now.

As you can see on the visualization dashboard, I put the side notes within the body. This is because on laptops, the screen is not wide enough to show both the visualization and the side notes.

My questions

1). Four countries/areas, i.e., BEL, ESP, SWE and HKG do not have the most updated data. The first three are constantly 1 to 2 days behind and Hong Kong is more than a week behind. In COVID-19 Spreading trends' last_index, I had to accommodate this through "last*index - 9". Another solution is to simply remove these countries/areas' data, especially that from Hong Kong. I might also try to use a different parameter to replace "last*index".

2). You did not include the visualization of COVID-19 Confirmed vs. New cases. I guess it was because this visualization was a recreation of Aatish Bhatia's. If copy right was the concern, I understand it and will remove it from the dashboard. I do feel that this visualization is better than the original one, though.

3). Case Fatality Rate (CFR) tends to be much larger than 1 in the beginning, if the delay parameter is set. I am thinking of using an if statement to set the CFR to be zero when the number of confirmed cases is 0.01. Is it necessary? Or I can leave it as it is for now.

4). Should I put all the visualizations in one singe page or put them in seperate pages? You can see what's the result of have three separate pages here, and here is the repo of this webpage. I thought about having three separate pages because 1) right now, if we put them in one page, the loading needs time. I thought separate pages will make the loading faster but it didn't. 2) It's seems too packed with three visualizations in each page. I think it's better to stick to one page because three separate pages makes the loading much slower, rather than faster.

5). How should I (re)organize the covid19-data repo? I created a folder named "depreciated" and put all files not in use in it. Is this acceptable? I find it easier than creating folders of "depreciated" within each sub-folders, i.e., "data_sources", "scripts", "output", etc.

Future directions

1). Using Python to manipulate the data. I manipulated the data on ObservableHQ because I thought I can import the output directly within the platform but I didn't find a way to do it. So, right now, I need to run the data manipulation notebook each day and use the bash script in the Repo to update the data. I might redo the manipulation using a Python script so that everything can be automated using one single bash script.

2). Creating a racing bar chart. I am thinking of creating a bar chart race for confirmed cases and deaths, like this visualization.

3). Creating a time series global map. I am thinking of creating a temporal map showing the growth in confirmed cases and deaths in each country, like this exmaple. Each country will have a circle whose area represents represent the number of cases / deaths. Ideally, I'll have a slider for time where people can see time progression and adjust the date for visualizations. I guess other people might have done it but I only saw that one by Bing, which later disappeared. OWID has a similar one, in which they are using categorical colors to represent the size.

4). Improving the website. I am thinking of creating an "About" page in front, which might shift the audience's attention from the initial data loading. Also, I am thinking of designing a foldable sidebar, like the one on OWID website. Third, I'll try making it responsive. Right now, the webpage looks ugly on mobile phones. Fourth, I'll try enabling side notes when the screen is wide enough.

Could I know what other future directions you have in mind?

I had been learning D3.js in the first month and I logged my process of learning. A by-product of this is D3 tutorial I created. I also logged my work every day in this spreadsheet.

The deadline of PhD application is quickly approaching, could I know whether I can focus more on my application in next two months? I will still work on this project but will devote less time than before.

How to remove the signs in Confirmed & New cases on top?

Internationalization, CJK

In Confirmed & New Cases, country names won't show even when being pointed on.

Should I later submit a PR to the original Observable notebook? I don't want to screw up your codes.