The raw data used was the COVID-19 data sourced from 'Our World in Data'. It is a collection of COVID-19 data from all over the world which is updated daily. This data includes confirmed cases, deaths, dates, along many others. Part A Task 2 is focused on the locations, total cases, total deaths, and new cases in 2020. The raw data was preprocessed first by selecting data from year 2020. The data included the location, total cases, total deaths and new cases. Then the data was further preprocessed by grouping the data by location, finding the total cases, total deaths and total new cases in 2020 per location. I added the case fatality rate for each location inside the data. The case fatality rate is defined as the number of deaths per confirmed case in a given period, which in our case, is 2020. As such, my equation is (total deaths / total cases). I avoided using total new cases (or even new deaths which is in the raw data) as the deaths in total deaths or new deaths could have been from cases before 2020, which would not be accurate to the case fatality rate in 2020. Thus, my definition for the case fatality rate would be the case fatality rate as of 2020.

There were two major limitations I observed from the data. Firstly, the data did not start from the beginning of 2020, thus the data that we want, which is in 2020, would not be as accurate as compared to if we had the full data for 2020. Secondly, as mentioned regarding the case fatality rate, we could have found the case fatality rate just for year 2020 alone if we had the full data on the new cases in 2020 and the full data on the new deaths in 2020 that were only from the new cases in 2020.

Now that I have data on the case fatality rate and total new cases in 2020 per location, I created two scatterplots. Both scatterplots were of case fatality rate (y-axis) and confirmed new cases (x-axis) by locations in year 2020. The only difference was that the x-axis was changed to a log scale in the second scatter plot (Figure 2). Scatterplots for 2 numeric variables are used to see how change in one variable relates to change in the second variable. So, with the scatterplots, what I would expect to find is if the change in confirmed new cases affected the case fatality rate.

Prior to the analysis and patterns observed in scatterplots, I analyzed difference between the two scatterplots first. The scatterplot in Figure 1 was not an ideal scatterplot to observe as I had an overplotting problem due to the outliers. So, Figure 2 would be a better scatterplot to use for analysis as it remedies the problem by changing the x-axis to be log scaled. This helped reduce the overplotting and made it easier to observe a pattern or relation.

The pattern observed would be null. As seen in Figure 2, the case fatality rate did not change with the increase in confirmed new cases (ignoring the factor of locations). Although there was a strong correlation in the scatterplot, I concluded that the pattern is null as most of the case fatality rates did not increase nor decrease as confirmed new cases increased. This analysis disregards the location as a factor and only looked at the pattern of confirmed new cases to case fatality rate in 2020.

Figure 1: scatter-a.png

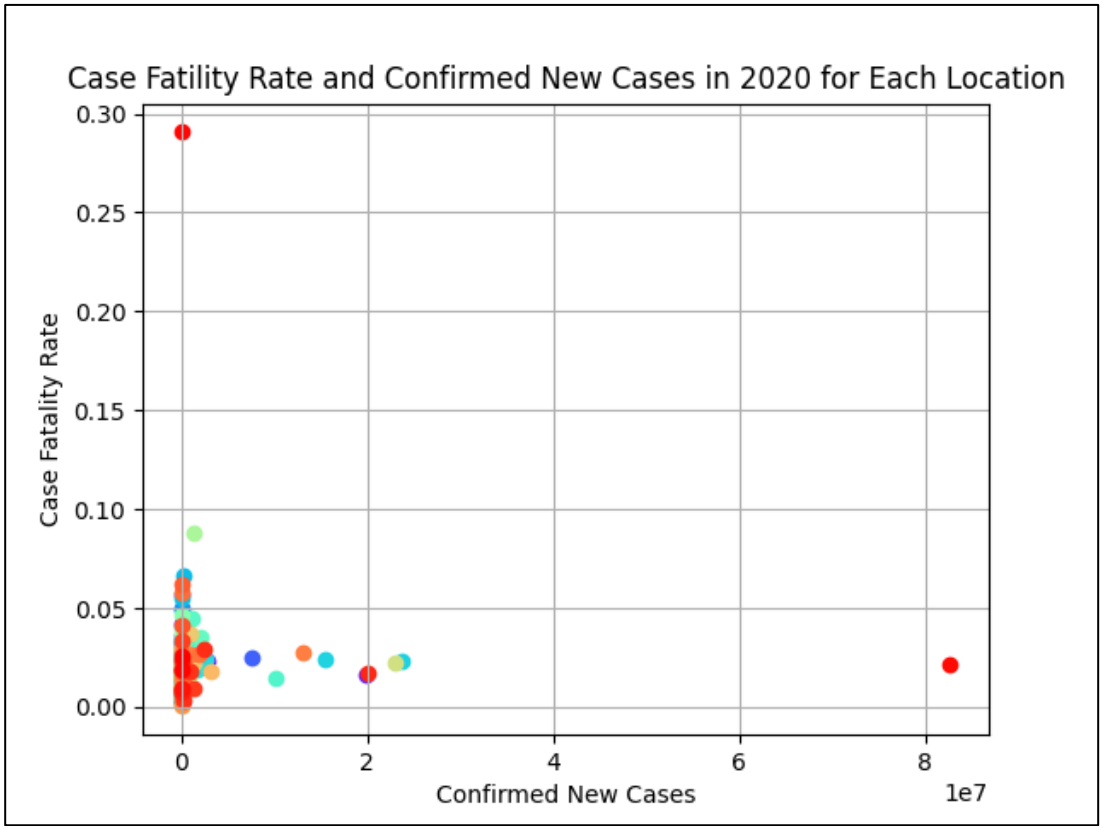

Figure 2: scatter-b.png