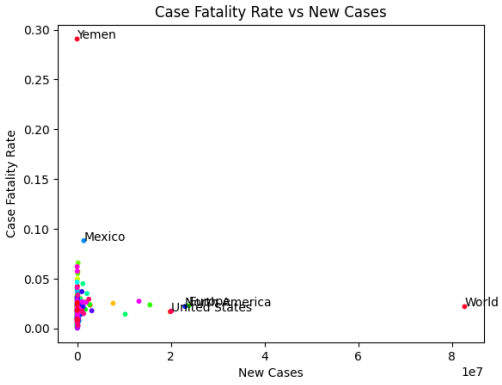
Visual Analysis of Covid-19 Dataset

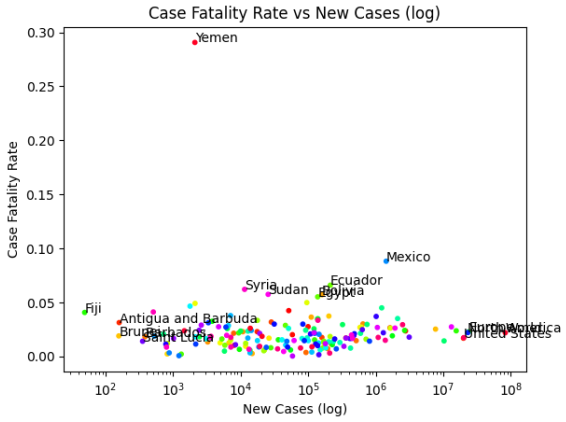
The raw data of Covid-19 comes from Our World in Data, an organisation which does research and data collection of the world's largest problems. The dataset contains a lot of missing data in different variables, where an accurate data cleaning would be required to remove any missing data from the preprocessing steps.

The preprocessing steps include separating and extracting values of year from date to slice the data so that only data in year 2020 are aggregated. The following steps include summing up values of new cases and new deaths per location in 2020 which results in the total amount of each variable and calculating case fatality rate. To prevent from overcrowding the scatter plot with data points, each location will only have one point in the scatter plot, representing aggregated data for the whole year. A few extra steps are taken into providing a way to indicate the location for each plot point to provide better visibility of the points in the scatter plot, although it comes with a cost of not being able to identify every single location in each plot.



Scatter plot 1

The first scatter plot with linear scale in both axes is extremely dense in the bottom left portion of the plot due to nearly every country having less than 10 million cases in 2020. There are few outliers visible from this plot which includes Yemen on the top left, having significantly higher case fatality rate of nearly 0.3, the World on the bottom right, which is the sum of cases in all countries having more than 80 million cases, and to a certain extent Mexico, which is noticeable in this plot just above the highly dense area of the plot.



Scatter plot 2

The second scatter plot is extremely dense in the bottom area after switching the x-axis scale from linear scale to logarithmic scale. As visible from the above scatter plot, the total number of new cases for each location varies from less than 100 up to above 1 million cases. Similar to scatter plot 1, the outliers from scatter plot 1 are also noticeable in scatter plot 2, excluding the World total. An interesting case is the number of new cases for United States being very close to the number of new cases for North America, meaning United States contributes a huge number to the number of new cases for North America.

Both scatter plots have similar density in the bottom area of each scatter plot, although the points in scatter plot 2 is more spread out and provides more information on the number of new cases for each location than scatter plot 1. Lower number of new cases are also much more noticeable in scatter plot 2 as opposed to scatter plot 1. Furthermore, outliers are much easier to indicate in scatter plot 2, in contrast to scatter plot 1. Finally, due to better visibility of plots and information, scatter plot 2 provide a much better explanation of the situation of Covid-19 around the world.