Part A Task 3: Discussion and Visual Analysis

Simon Keng-Jeng Wong, Student ID: 995097

The raw data, “Our World in Data COVID-19 dataset” obtained from https://covid.ourworldindata.org/data/owid-covid-data.csv, is a csv file that contains the figures of COVID-19 within each country and continent, as well as the entire world. It ranges from as early as January 2020 through to March 2021 and includes statistics such as the daily number of cases and deaths, as well as smaller quantities such as median age, life expectancy and GDP per capita. Although Our World in Data is a credible source for data, there may be some possible limitations associated with the information. As scientists have backtracked COVID-19 to have first emerged in December 2019, we see that since our data begins in January 2020 there would be several cases not documented. Additionally, as the statistics are derived from many different sources, they may vary in accuracy as well as precision. Considering we are given data of developing countries from continents such as Africa and South America, this information may be delayed significantly and/or inaccurate.

Initially, the data is trimmed down to only contain the statistics for the end of each month of 2020, “2020-01-31” representing January, “2020-02-29” representing February and so on. The column, ‘date’ is relabelled to ‘month’, and the index is then relabelled to ‘location’. The number of ‘new cases’ is redefined to be the difference between the number of ‘total cases’ of the current month and the month prior, leaving it null if no data exists for the month prior. Finally, the ‘case fatality rate’ is computed by dividing the number of ‘new deaths’ by the number of ‘total cases’ for that month.

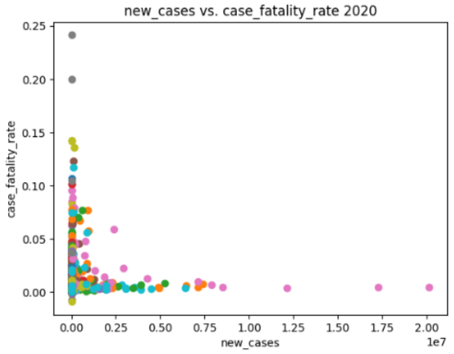

Figure 1

Both scatterplots, Figure 1 and Figure 2 represent the number of new cases compared to the case fatality rate for every month in the year of 2020, with every plot representing a month, and each different colour representing a different location. As there are roughly 200 locations, a legend that lists all locations with corresponding colours would not provide any further clarification into the data and is thus left out. Looking at Figure 1, it is hard to decipher many trends within the data as it is heavily skewed to the right, having locations with larger populations, such as the entire globe gravitating more towards the right, while locations with smaller populations grouping towards the left. Figure 2 however, rescales the x-axis using the natural log, and highlights that there is very little if no correlation between the number of new cases and case fatality rate per month. This draws further into the scatter plot beginning to take resemblance of a normal distribution, which is to be expected if we are given data from a large sample space that are independent and thus have no correlation. From this overall trend, we expect any new data from 2021 to lie somewhere within the confines of the Figure 2, and can predict that while conditions remain consistent, case fatality rate should not increase drastically.

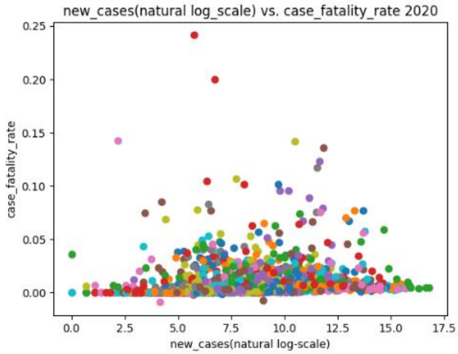

Figure 2

The biggest difference between Figures 1 and 2 lie in their readability. Although they portray the same information, it is very difficult to decipher any trends within Figure 1, which Figure 2 brings light to. This lies solely in the rescaling of the number of ‘new cases’ for each location using the natural log, which shrinks down the data. Whilst no trend or correlation can be deciphered from Figure 1, Figure 2 hints at the absence of correlation and thus independence of the two variables which, realistically make sense.