Part A Task 3 Discussion and visual analysis

All data being used in Part A were extracted from owid-covid-data.csv file. This csv file was provided as part of information source of Assignment 1 in LMS, and been used as raw data in Part A. It contains information of daily records of covid-19 of all countries and regions from start of 2020 to March of 2021 including important data of locations, date, total cases, new cases, new death and total death, etc. However, the limitation of this raw data is obvious. containing daily records of covid-19 of all countries in period of one year makes data frame extremely prolix and interminable. Daily records were stochastic, it can hardly reveal the tendency of change of data in certain period of time. In addition, finding the daily case fatality rate was meaningless, since the number of death cases and confirmed cases could be various from day to day and it leads to a random distribute of case fatality rate. Therefore, it was very difficult to find anything valuable of plot patterns through this raw data. In order to have a better visualization, this report is going to only focus on the data in year of 2020, and daily record were added up and reorganized into monthly record. Only useful data were extracted from raw data including locations, moth, case fatality rate, total cases, new cases total deaths and new deaths. Furthermore, case fatality rates were also calculated based on new generated data and added into the data frame as a new column.

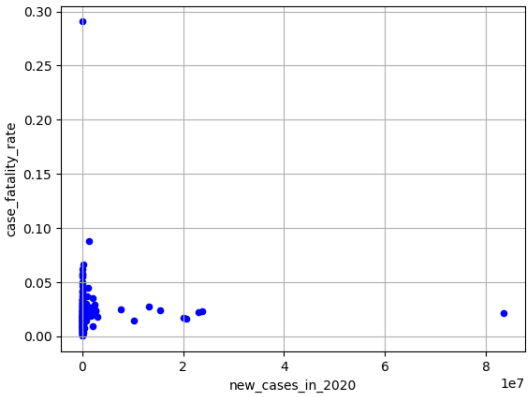

Figure 1 scatter a

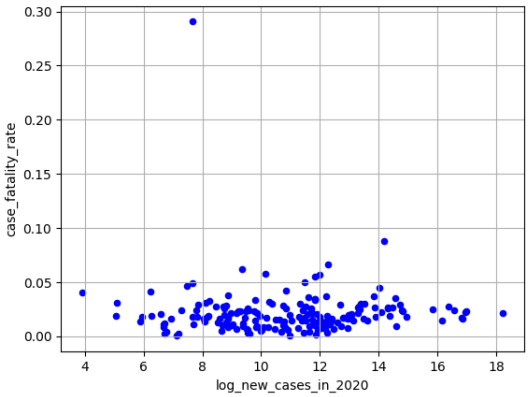

Figure 2 scatter b

Then a scatter plot (figure 1) was generated with new cases in 2020 (on the x axis) and case fatality rate (on the y axis). Each dot on the figure represent a different country. However, most of new cases in 2020 in each country was centralized around one million, so it is very hard to observe the data patterns. Readers can still find that case fatality rate was in the range of 0 to 0.075. However, there are a few outliers exist. The fatality rate of a country is up to 0.3. Another country has more than 80 millions of new cases in 2020 in total. This outlier also directly influenced the scale of x in figure 1. From the figure 2, most of x values were in range of 8 to 14 and most of y values were in range of 0 to 0.05. Only a few countries have case fatality rate above 0.05 and log value of new cases smaller than 6 or larger than 16. There was only one country is outlier, which has x value of 8 and a relatively high value of case fatality rate of 0.3. Overall in figure 2, two variables has relatively strong correlation strength and has null patterns. The line of best fit was a flat line with y value around 0.02. This means that in most countries, case fatality rate was stabilized around the value of 0.02 regardless of the value of new case.

Comparing two figures, they both share the same case fatality rate for all countries. However, readers can hardly observe any useful patterns from figure 1, since all dots were squeezed together due to the scale of x axis. With helping of log scale value, the data of new cases could be dispersed on x axis which make the figure 2 much more readable and easily to discover the relation of two variables. To be noticed, the gap between outliers and main part of x values were also decreased by log calculation in figure 2 comparing with figure 1.