Introduction

The raw data set shows the spread of COVID-19 worldwide. It contains 59 variables, including total cases, new cases, new deaths, total deaths and so on. Each variable is categorized by location and date.

To evaluate the relationship between mortality and new cases, a new variable 'case_fatality_rate' was created. It is defined as the number of deaths per confirmed case in a given period. Before calculating the new variable, the raw dataset was grouped by location and months in 2020. Then the calculation of the new variable was finished in the grouped table, and thus we could get facility rate which is in given period and location.

Explanation of Plots

1. Figure 1.1 shows the relationship between facility rate and new cases in 2020, the data was grouped by location (each point illustrates the facility rate and new cases of a location).

   As shown in Figure 1.1, the points are concentrated in the lower left corner of the scatter plot, which means the new cases of most of locations are less than 500,000 in 2020, just in a few locations, the new cases in 2020 are more than 500,000. For facility rate, in most of the locations the facility rate is lower than 10%, only in one location the facility rate is nearly 30%.

   The relationship between the two is hard to see. It doesn't look like a positive or a negative correlation. Therefore, I tend to believe there is no relationship between the two from this plot.

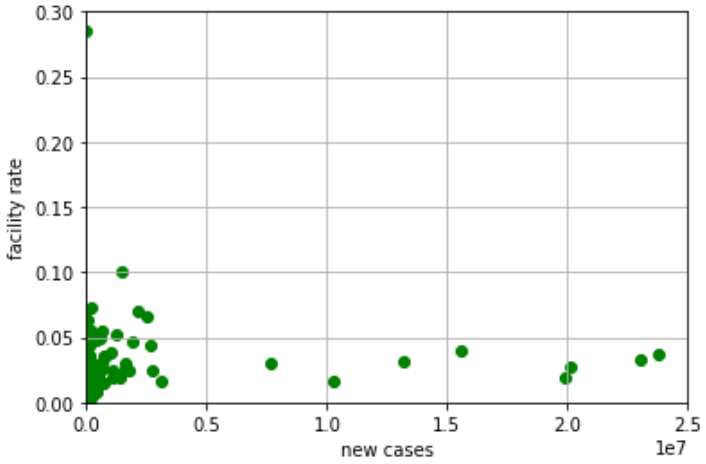

Figure 1.1 Facility rate vs New cases

2. Figure 1.2 shows the same relationship as Figure 1.1, the only change is the x-axis is changed to a log-scale.

   From this figure, the pattern is much easier to be observed. The scatter points are distributed in a fan shape, which means the relationship between the two variables may be positive (the more new cases, the higher facility rate).

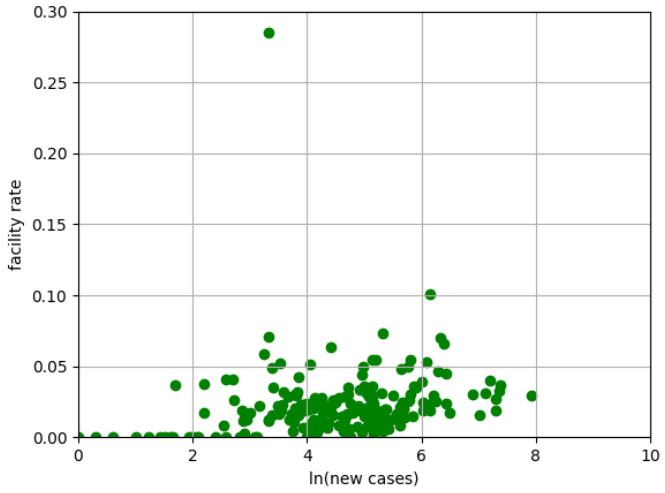

Figure 1.2 Facility rate vs ln (New cases)

Contrasting the Two Scatter Plots

1. Scatter-a (Figure 1.1)
   Merit:

   I . Shows the exact data range of x-axis.

   Demerit:

   I . The span between span is too large, the pattern is hard to observe.

2. Scatter-b (Figure 1.2)
   Merit:

   I . Reduce the absolute value of data and narrow the span between data, which makes the data to be more concentrated on a scatter plot.

   II . The relationship between the two variables can be easier to observed.

   Demerit:

   I . Can only show scaled data.