Visual analysis report for assignment 1 part A

This report focuses on the visual analysis for COMP20008 assignment 1 part A. it contains three main ideas, an introduction and description of the raw data, an explanation of the scatter plots and patterns observed, and contrasting the two scatter plots.

The raw data has come from https://covid.ourworldindata.org/data/owid-covid-data.csv. It is about the COVID-19 statistics around the world collected by Our World in Data. It contains many columns such as location, total_cases, new_cases, total_deaths, new_deaths. After reading through the raw data, we can discover that this raw data has a significant problem regarding data missing. For example, some locations' new_cases are null. Also, in the raw data, some countries have negative values. For example, location Angola, date 8/10/2020, new_deaths is -3. This is not possible. New_dealths value can not be a negative number. As a result, this raw data is not reliable. There are also some columns in the raw data that are not filled in, such as icu_patients, so those columns are unnecessary.

To process raw data, there are few steps required for processing data to produce the final plots. First, we need to get valuable columns from raw data to create a data frame. Second, extract yearly(2020) data from the data frame. Third, aggregate total_deaths, new_deaths, total_cases, new_cases from the data frame, and then use this data frame to plot to scatter plots.

From both Figure 1 and Figure 2, we can observe the x-axis and y-axis are increasing trends. For Figure 1, the range for the x-axis is from 0 to 8*107, and for the y-axis is from 0.00 to 0.30. For Figure 2, the range for the x-axis is from 0 to 107, and for the y-axis is from 0.00 to 0.30.

In Figure 1, we can observe two dots are far from the other dots(location). The first dot(location) is around case_fatality_rate 0.30, and the second dot(location) is around new_case 8*10 7. These two dots(locations) are referred to as the world. The world((location)) is sum up all the data across all the locations, so it is far from the other dots(locations). Same as Figure 2, there is a dot(world) around case_fatality_rate 0.30.

From Figure 1, it is not hard to find out that most dots(locations) are concentrated in the bottom left of the plot. This is because, for most countries, case_fatality_rate and new_cases will not be significantly higher than other countries due to the

government's need to try its best to minimize the deaths and new cases. However, there are still some dots a little bit far from the concentrated part. This might because some least developed counties (LDC) it has a relatively immature medical system, less infrastructure. Consequently, it leads to a relatively higher new_case value and case_fatality_rate value. The same principle is applied in Figure 2. There is some dots' case_fatalit_rate value that is higher than 0.05.

The difference between these two plots is in Figure 2 x-axis uses a log-scale rather than the normal number in Figure 1 x-axis. Unlike in Figure 1, most dots are concentrate in the bottom left of the plot, and the log scale makes the dots more scattered. It makes us much easier to observe how case_fatality_rate is distributed.

Overall, as this report showed, this raw data is not reliable. Plot the scatter plot after processing the raw data, and we can find out that Figure 2 is much clearer than Figure 1 to observe the pattern.

Figures:

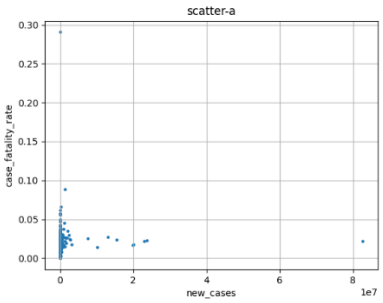

Figure 1

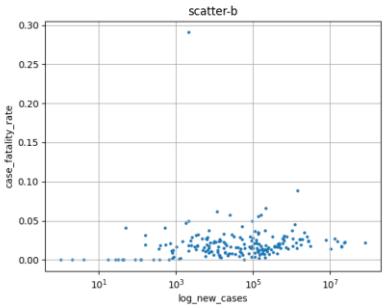

Figure 2