EODP report for Assignment 1 Part A

Student Name: Jie Xie

Student Number: 1174437

For COMP20008 Assignment 1 part A, the relative data of COVID-19 over the world was crawled and various elements were analysed.

A brief description of the raw data:

From the link given in the assignment 1 specification, I found the source of the raw data comes from the COVID-19 Data Repository by the Centre for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU).

However the raw data is not quite reliable. The number of cases/deaths reported by any institution on a given day is probably inaccurate, because of the long reporting chain between a new case/death and the update in the dataset. There are some significant flaws in the source. Many columns are empty which makes the data inconsistent. This issue also causes trouble to the calculation of the case fatality rate. Meanwhile decreasing values in cases/deaths can appear if a country had previously overestimated the number of cases/deaths and wants to correct the data. Even major changes could happen to a country's entire time series if the institute decides to correct all those previous values.

All pre-processing steps:

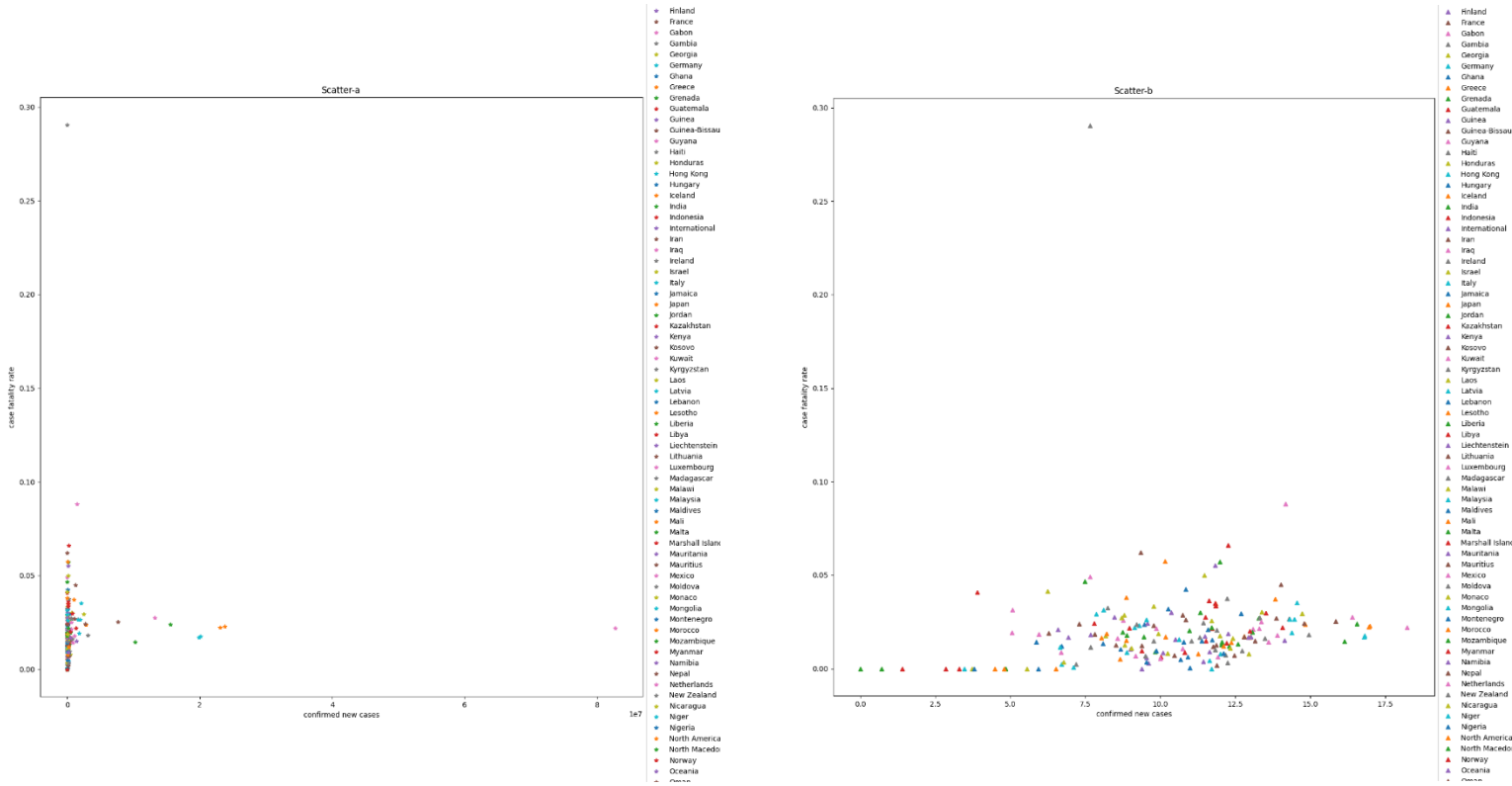
In order to visualise the raw data I accessed it as a csv file, then locates the month and year from the 'date' column. After selecting the data in 2020 from the whole file, I created an empty data frame, followed by filling the data in the required order. Then I made a copy of 'location' and 'month' from the original data and created a double for loop to fill the data frame step by step. I initialised the value of both total cases and total deaths as 0 between the two for loops, as the situation varies in different countries. I built an empty list to temporarily store different columns and a temp data to transfer each specific row of data. Then by following the required order the empty list is completely filled, however a list cannot be directly appended on a data frame. Therefore I tried to transfer the empty list (now filled with data) to a series, and used the column name as the index. In the next step the data frame is firstly formed with six columns.

For the new column 'case_fatality_rate' I simply calculated it and inserted the results after 'month' by using the existed function in pandas. I then started the steps for visualisation with the first step to split the data frame into groups. Since the graph displays the data in 2020, I found the sum of the new cases in each month as the variable on x-axis. Similarly on the y-axis the case fatality rate was calculated again in the annual level. Then the regular plotting steps were applied and the second graph was done in a very similar way, except changing the unit on the x-axis to log.

Explanation of the plots observed:

Scatter-a: The x-axis trend reveals that most countries do not have massive new cases, a few countries have many new cases, and an outlier has a huge amount of new cases. The y-axis trend demonstrates that in most countries the case fatality rate is below 0.05, some countries has this value from 0.05 to 0.10, however the outlier almost reaches 0.30. From the overall trend we can conclude that in most countries the situation is under control as they do not have much new cases nor high case fatality rate. Nevertheless, some countries are in dangerous situation and there are two outliers on the graph.

Scatter-b: On the x-axis most countries are in the range from 5 to 15, and there is no obvious outlier. The y-axis trend shows the case fatality rate in most countries is below 0.05, some countries' value vary from 0.05 to 0.10, however the outlier still exists and is close to 0.30. Overall the new cases are uniformly distributed and the case fatality rate is generally controlled in a low level, with a rather high exception.



(Scatter-a on the left and Scatter-b on the right)

Comparison between the two scatter plots:

Scatter-b just changes the unit of the x-axis but evenly expands the x-axis instead leaving it stick together like Scatter-a does. In the second plot by importing log the difference between x values become smaller. This brings convenience to our analysis as each value can now be clearly viewed.