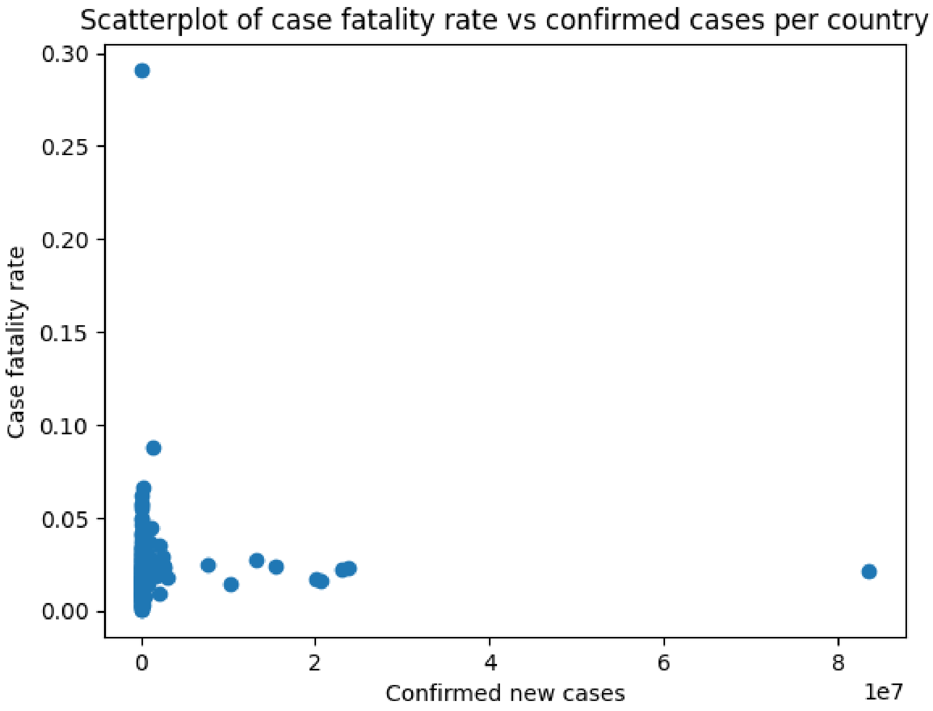
Comp20008 assignment 1

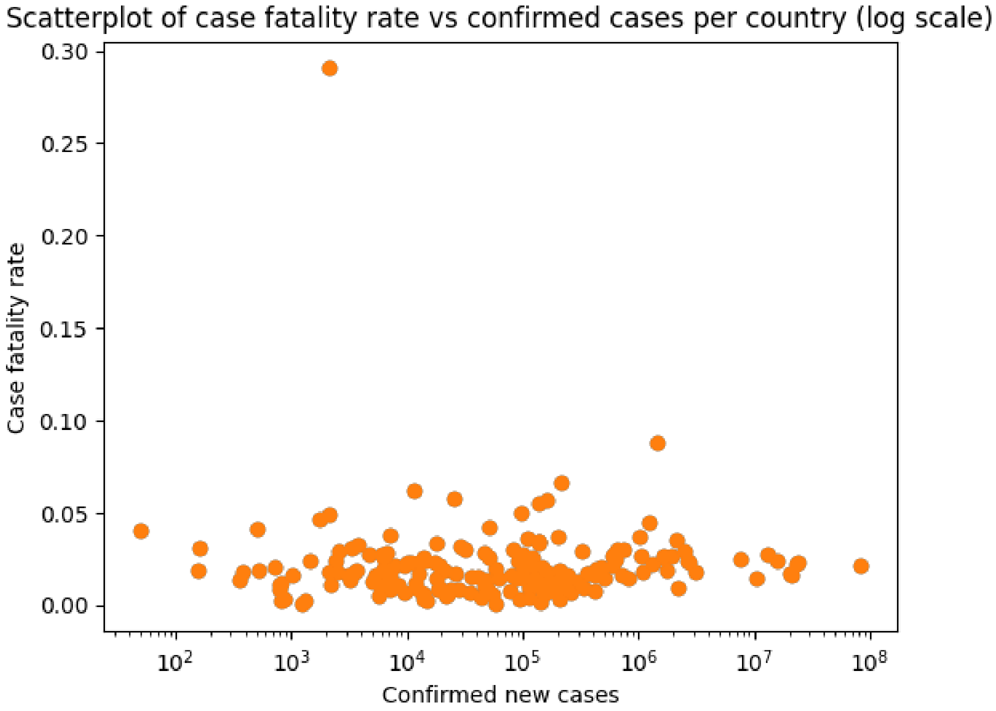
Brief introduction:

In this assignment, the raw data is taken from the Our World in Data website (https://covid.ourworldindata.org/data/owid-covid-data.csv). This file contains all the relevant information of the world's regions with regard of Covid-19. A number of preprocessing procedures were applied to the data. Firstly, since the objective of the assignment is to examine the total number of new cases and deaths recorded per country/region without considering the causes and how each country/region responded to the pandemic, only 4 variables were used in the report: total cases, new cases, total deaths and new deaths. Secondly, only the data in the year 2020 were retained. Thirdly, the data were aggregated on a monthly basis. And lastly, to better understand the data, a new variable is introduced: case fatality rate, a measure of the proportion between number of deaths and number of new cases each period. Entries with no values are left as is. The processed data were then plotted against 2 figures, 1 standard scatter plot, and 1 similar figure but have a log-scaled x-axis. Any NaN values are excluded from the plot.

Below are the 2 plots for the data:



scatter plot a



scatter plot b

Explanation of the plots and patterns:

It can be derived from both of the plots that the majority of countries and regions recorded a case fatality rate between 0 and 0.05. The only outlier with a case fatality rate of 0.03 is Yemen, recorded a total of about 600 deaths and 2100 cases.

The farthest point on the plot to the right is, undoubtedly, the world data, recorded a total of 80 million total cases in 2020 and a case fatality rate of 0.022. World data is closely followed by the data of the entire continents and highly populated countries such as the United States, Brazil and India, thus such large value in total cases.

When examining scatter plot a, the only sensible conclusion about the number of total cases is that the majority of countries/regions recorded a number less than 5 million. This is not useful.

In scatter plot b, the x-axis are represented on a logarithmic scale, making the data easier to interpret. One can look at the graph and actually see the trend of total cases per country/region, which ranges from 10000 to 1 million cases.

For those who have only access to the plot and not to the data behind it, scatter b provides better insights as data points are more spread out. Such information can be the relevant number of points in the graph, how many points lie between a certain range, etc.

One important takeaway after examining the 2 plots is that the case fatality rate of Covid-19 is relatively unaffected by the total number of people that have the disease.

The contrast between 2 plots, even though they derive from the same dataset, demonstrate that visual representation is an important part in our understanding of the information and the ability to reach reasonable conclusions.