# COMP20008 Elements of Data Processing
# Assignment 1

## Analysis of Case Fatality Rate vs Confirmed New Cases by Countries in 2020

Hasne Hossain: 1102602

The analysis was done using “Our World in Data COVID-19 dataset" (\owid-coviddata") from https://covid.ourworldindata.org/data/owid-covid-data.csv. The dataset includes daily statistics related to COVID of various locations from 2020 and going upto 2021. We used this data to calculate the “total confirmed new cases” by location in 2020 and the “case fatality rate” for the year 2020 and then plotted Case Fatality Rate vs Confirmed New Cases by Countries in 2020 to observe the pattern.

To derive these data, we first constrained the dataset to have data from only 2020. Furthermore, we also removed incomplete or null rows for ease of calculation. We grouped the data by month and location which was followed by grouping the resulting data by location to get the total confirmed cases for each location in 2020. Since the dataset had a column called “total cases” which cumulatively updated the total cases for each location, we had to just take the last row of data for each location for 2020 to get their total cases. Similar technique was followed to obtain total deaths of each location in 2020 while to get the sum of new cases and new deaths for each location in 2020, we summed up all their rows of 2020 as their values were not cumulative. It is important to note that the value of the column heading “total cases” were not necessarily the total cases for just 2020 as counting may have begun from before. Same goes for the column “total deaths”. So, to analyse based on new cases and new deaths that occurred strictly in 2020 only, we used the values of column “new cases” and “new deaths” to derive case fatality rate (new cases in 2020/new deaths in 2020).

The graphs obtained are given below:

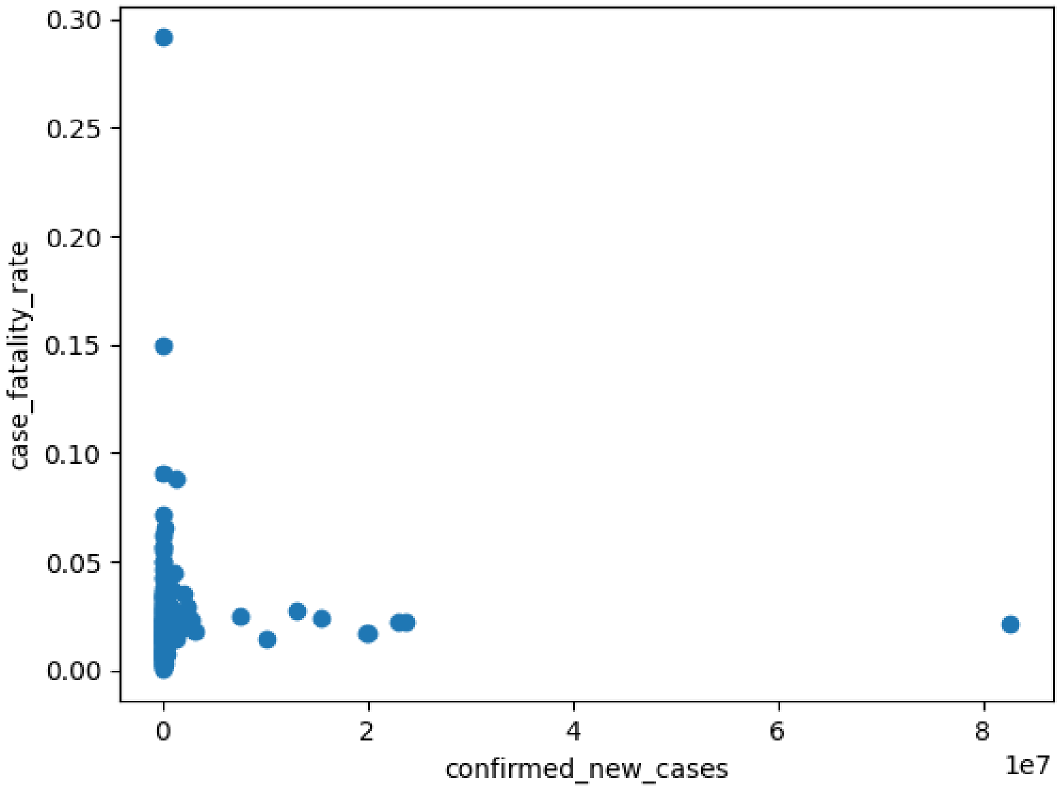

Figure: scatter-a

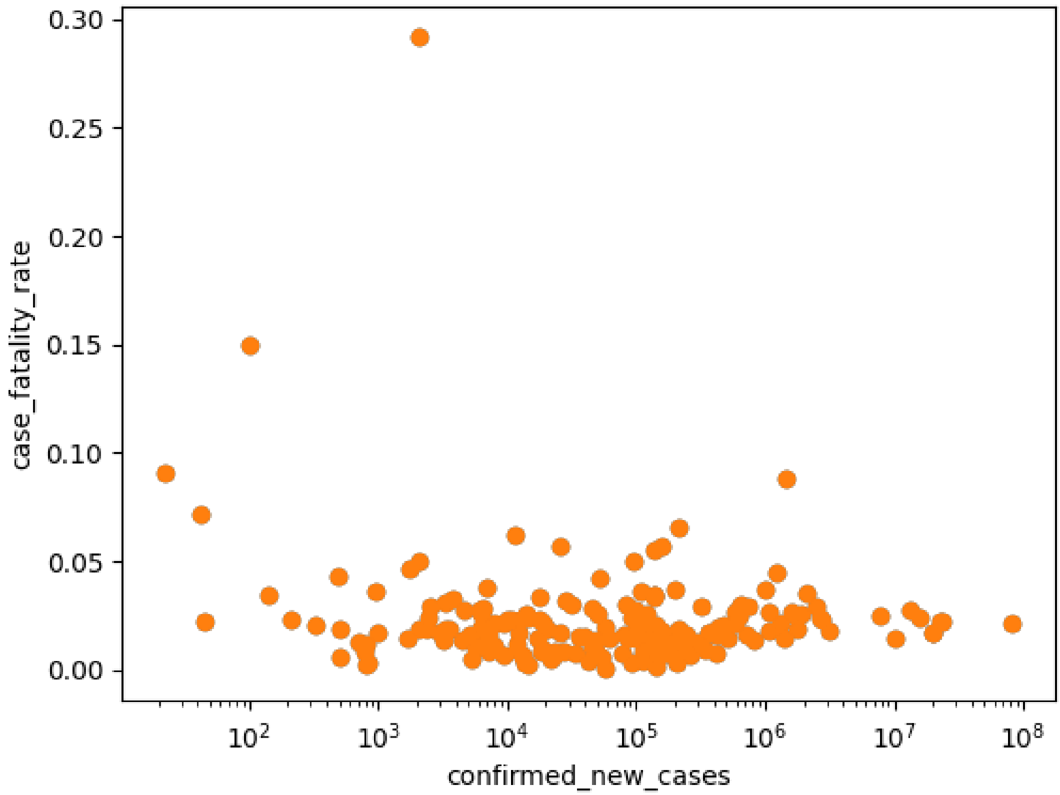

Figure: scatter-b

Both graphs represent the same set of data with just 1 change in scatter-b: the x-axis was changed to a log scale. Therefore, the pattern in the data is much easier to see in scatter-b than in scatter-a as they are more spreaded and less overlapping. Each dot represents a location. The results indicate that the majority of the locations had a fatality rate between 0 and 0.5. As can be visually seen from the graph, the case fatality rate is fairly constant irrespective of confirmed new cases. This shows that the relationship between case fatality rate and confirmed new cases is fairly weak.