Part a task 3

The data was an excel spreadsheet describing the daily covid data for a series of worldwide locations taken from the Our World in Data online database.

In order to preprocess the data relevant to my task, I had to extract the location, date, new case, and new deaths columns from the full dataset to a new data frame. I then changed the date format from a string representing the full date, to an integer representing the month and removed any data that was not from 2020, as only we were interested in the monthly data of each location in 2020. I experienced limitations in the string format of the date as the length of the date string was not consistent throughout the data. In addition, I originally tried to remove every entry that wasn't from 2020 directly from the data frame which made my program very slow, before I then recorded each index representing irrelevant data and removed all of them at once.

I then iterated through each location creating a dictionary of the monthly data for each location and appending the data to a list of all the monthly data for each location. Once I had the monthly data I created a new data frame with the monthly new case and new death data for each location, and iterated through each location to retrieve the total death and total case data using the cumulative sum function. This was difficult as there were differing lengths of the sections representing each location as a result of the vast discrepancies in the amount of data collected between each location due to the greatly differing size of location in the data. I had to use a dictionary with the indexes corresponding to each location to navigate this limitation.

Once I had the full monthly data for each location, I extracted the total confirmed cases by finding the maximum total cases for each location, and calculated the average case fatality rate by dividing the total deaths by the total cases. I then plotted this data using a scatter plot.

The main limitation I experienced was seeded in the vastness of the data. The number of locations was so great it was not possible to identify different locations in the scatter plots. In addition, as some locations were so much larger than others so it is difficult to recognise trends in the scatter plots as the range of values has to be so large to accommodate the world and continent data as well as the individual country data. This resulted in the non-log scale plot (figure 1) being impossible to analyse as it shows a cluster around the origin of most of the data, then the outliers at the extremes of the axes.

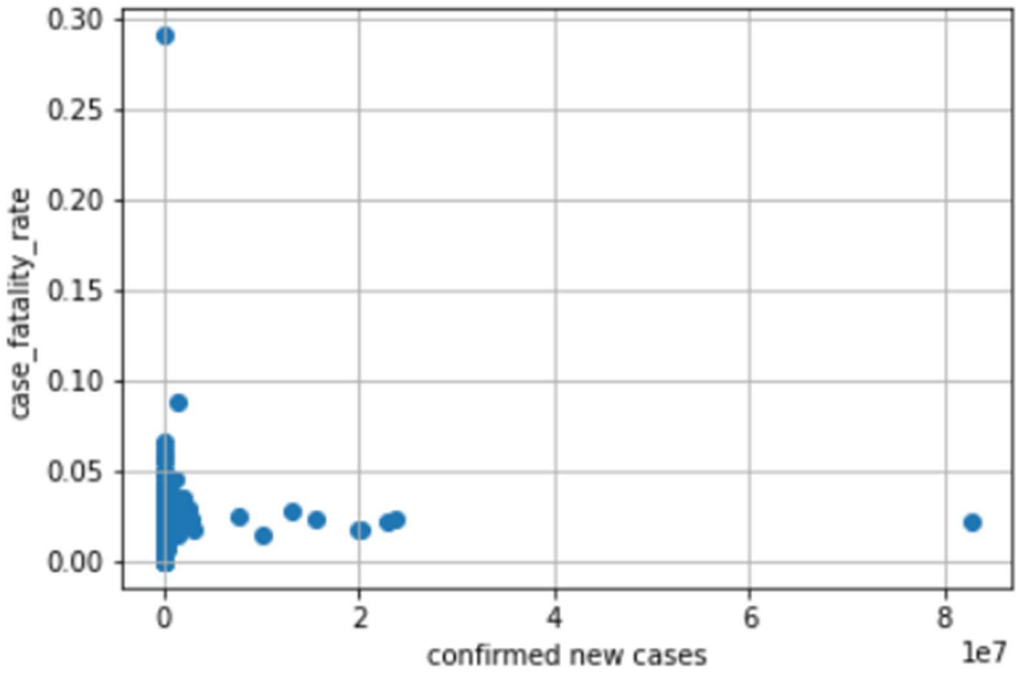

Figure 1

The log scale plot (figure 2) reveals an overall trend in increasing number of confirmed cases resulting in an increase in case fatality rate, as it no longer has the limitation of most of the relevant data being clustered around the start of the x axis.

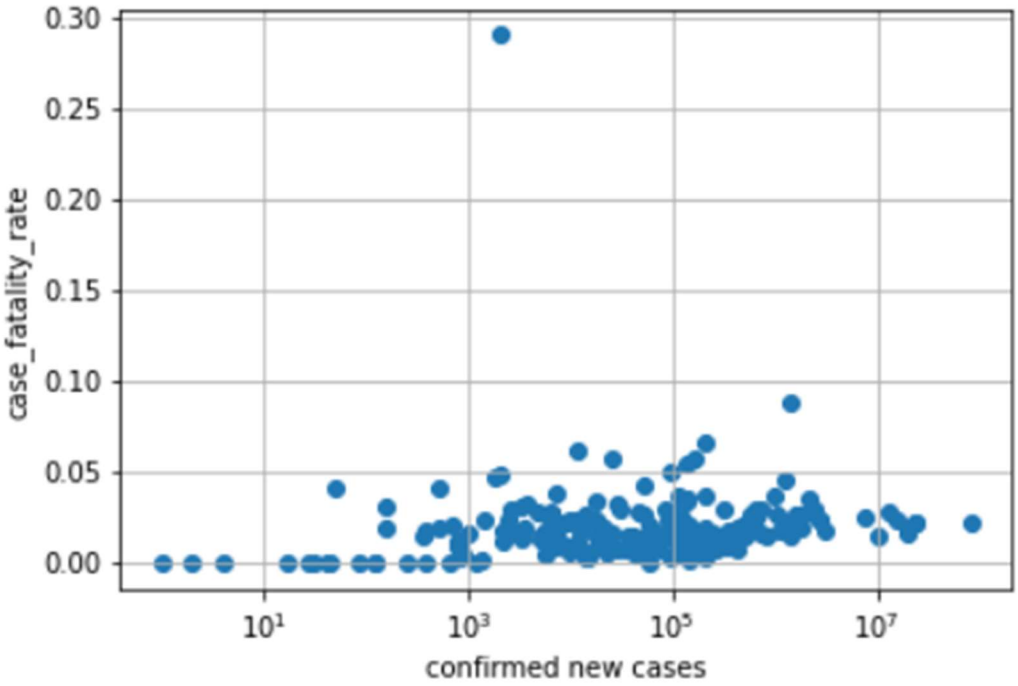

Figure 2

However, the plot still has clusters of data in the cases fatality rate region of [0, 0.10] as it still accommodates the outlier of Yemen, with a cases fatality rate of 0.29 despite only having around

2000. The extreme case fatality rate was largely due to the civil conflict occurring during 2020 which largely prevented appropriate medical assistance in Yemen. If we remove this outlier from the plot the trends are more easily identified, as evident in figure 3, a plot with the y-axis restricted to [-0.01, 0.10].

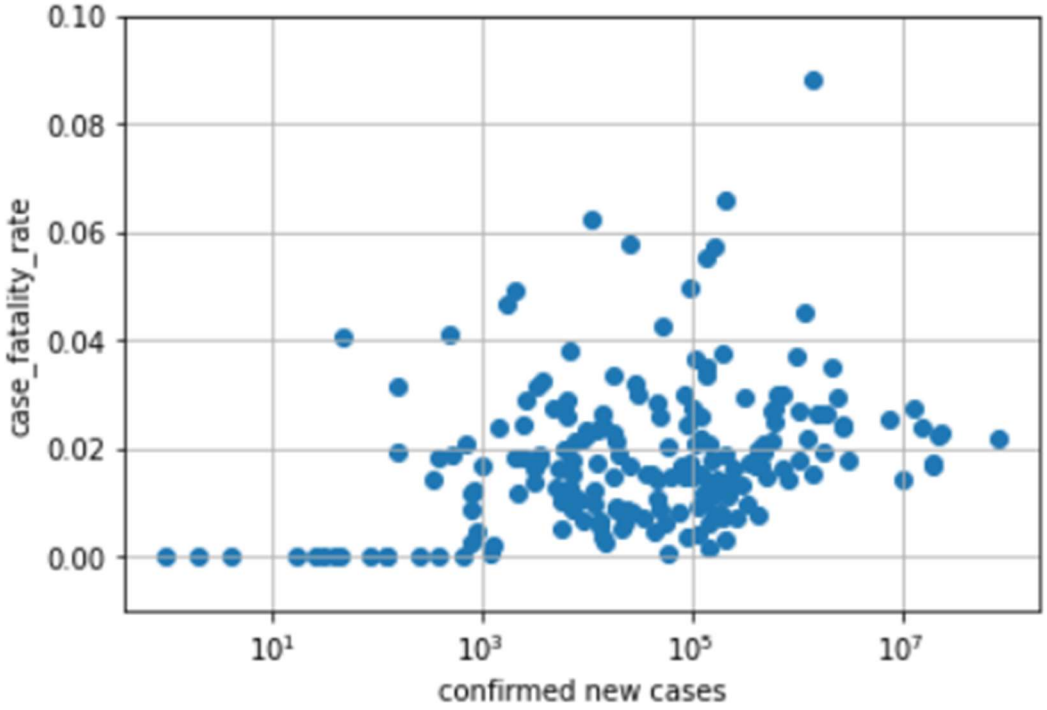

Figure 3