Part A Task 3 Discussion and visual analysis (4 marks)

1. The raw data consists of a csv file concerning information about COVID-19 during the timespans of 2020 to 2021. The data involves confirmed cases and deaths, hospitalization and ICU admissions, testing for COVID-19, vaccinations and other variables about each case such as the median age, gender, smoking status, etc. This data is also separated by location, date and each countries' ISO code.

   Some limitations within the raw data were that not all columns were consistently filled, such as the 'total vaccinations', or 'new vaccinations' columns. This was due to the vaccine only being recently developed but the lack of previous data in 2020 may skew the data model. Outliers within countries that had an overwhelming amount of 'total_deaths' or countries that had very little 'total_deaths' may also be a data limitation within the raw data.

   The pre-processing steps used on the data consisted of importing the csv file to a data frame using the pandas library within python. Once in a data frame format, the data was able to be easily aggregated using an aggregate function as well as be sorted for month and location within 2020 using "year_2020 = df.loc[year_filter].sort_values('location')". Finally, the fatality rate was easily calculated by dividing the total deaths by the total cases within the year 2020.

   This was then plotted using mathplotlib, and the log scale was developed for scatter-b.png via "plt.xscale("log")".

2. Scatter-a.png demonstrated the fatality case rate vs confirmed new cases by location by the year 2020 with a normal numerical staggered scale. This data showed a higher fatality rate in the beginning with a few outliers at 1.0, 0.5 and 0.4 case fatality rate. The case fatality rate seems to taper down to approx. 0.1-0.2 case fatality rate the higher the number of cases.

   Scatter-b.png shows a similar trend with the same data plotted being the fatality case rate vs new confirmed cases by location within the year 2020. Alike to scatter-a.png the same outliers exist at 1.0, 0.5 and 0.4 fatality case rate.

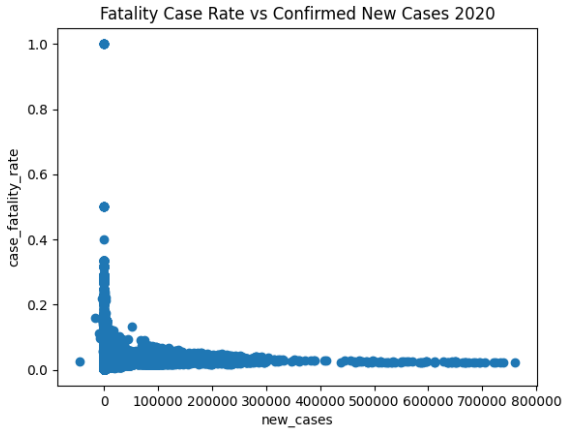

Figure 1:Scatter-a.png

   Overall both scatterplots imply that with an increasing number of new cases, the fatality case rate slowly decreases and stays at a constant fatality case rate average.

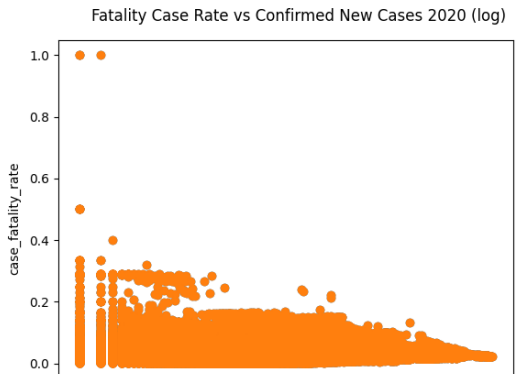

Figure 2:Scatter-b.png

3. Whilst both scatterplots demonstrate the same data, the overall trend is easier to read in scatter-a.png whilst scatter-b.png provides a more detailed spread of the data points. There seems to be a sharp negative relationship between the x and y values in scatter-a.png whilst in scatter-b.png there is only a slight negative relationship.