Visualization report for Part A3

1. The data set was obtained from Our World in Data and was pre-consolidated as a structed data type file known as a .csv file. The dataset contains daily nation-specific statistics relating to covid-19 for 207 nations and covering 59 variables that are pertinent to the transmission rates and death rates from the virus. Although the dataset appears comprehensive, there are several limitations. It should be acknowledged that the statistics are only accurate to the accuracy of report by each individual country in the dataset. Furthermore, it appears that certain fields in the dataset are missing and it is likely due to external factors such as nations reporting a range of statistical variables which differ from other countries. For example, economically developed nations like the United States and Japan have commenced mass vaccination campaigns and the data is available in the set, while poorer nations still lack the infrastructure, therefore, fields relating to such data would be missing and is a comprehensive analysis were to be caried out, the resulting data may be inaccurate and counter-representative of the world situation. In addition to that, the dataset is dynamic. Hence, in order for the data aggregation and visualization to be up to date, the dataset has to be extracted, transformed and loaded daily.

   In Part A task 1 and 2, the Pandas library was utilized to aggregate the data. We filtered the dataset to only include variables relevant to the study. This included the location, the number of total cases, new cases, total deaths and new deaths for each country. The months had to be extracted manually from the given date. The variables were then grouped by location and the number of new cases and new deaths at the end of the month were recorded and the total cumulative death count and case count was also recorded for each month for the respective countries. The case-fatality rate was calculated via the formula:

$$case\ fatality\ rate = \frac{total\ deaths}{total\ cases} \times 100$$

The statistic was then concatenated onto the DataFrame for each month for every country.

After the data was transformed, the dataset was saved as a .csv file and was loaded again to produce data visualizations.

2.

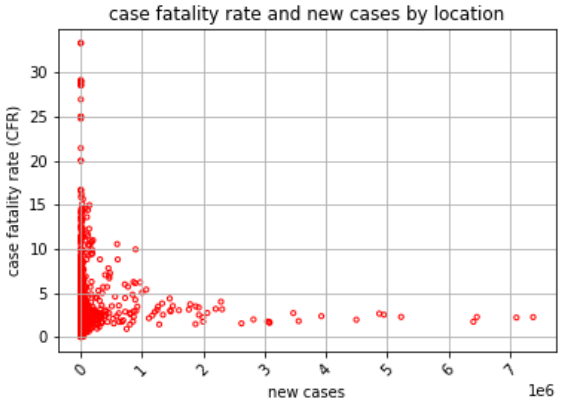

Figure 1 trend of case fatality rate per new cases

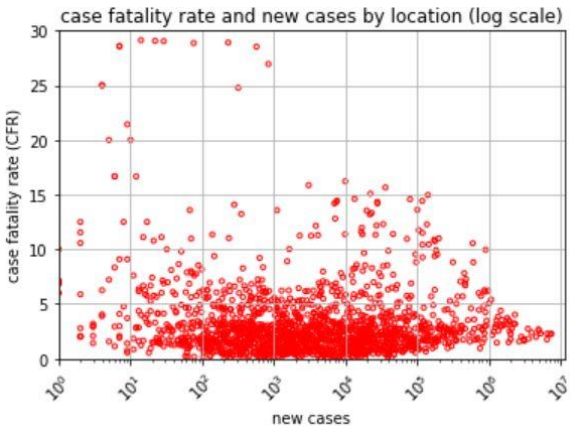

Figure 2 case fatility rate and new cases on a log scale

The figures above are scatter plots that visualise the relationship between case-fatality rates (y-axis) and new cases for every calendar month (x-axis). The figures show that the range of case-fatality rates tends to be wide for low case counts and tends to narrow as more new cases are recorded. There is a clear non-linear negative correlation between the new cases and the case fatality rate as seen in figure 1. Figure 1 also resembles an exponential decay plot. It can also be interpreted in figure 2 that a majority of records have a fairly stable case-fatality rate as new cases grow as shown by the cluster of points in figure 2.

3. Figure 1 describes the relationship with a linear scale while figure 2 uses a logarithmic scale. The logarithmic scale would be a more ideal scale to use for aiding readability as discrete points can be better identified and the values on the x-axis can also be more easily read; especially considering the case-fatality rates of countries with below 2.5×10^6 cases. Via figure 2, we can also interpret that most of the case fatility rates are between 0% and 5% with little regard of the number of cases as shown by the high density of data points in that area. Although figure 2 helps discern ambiguous data points, figure 1 is more viable for looking at the trend by which case fatality is decreasing.