In the given assignment, students were set the task of wrangling raw data through Python to categorise them by location(country) and by date(month). The dataset in question is the Covid-19 dataset compiled by Our World in Data (OWID). OWID is a publication with a goal pertaining to research and data to “make progress against the world’s largest problems”. Within the Covid-19 dataset, are records related to Covid-19 updated on a daily-basis. Components of these are among the likes of location, date, new and total cases and deaths, and so on. These six aforementioned components were the target focus of the Assignment 1 for Elements of Data Processing.

While the dataset is packed with records from every country it could get data from, unfortunately, not all data is equal. It was observed throughout the process of making Part A that there were differences in the amount of data that each country could report. For example, different countries started reporting Covid-related statistics at different times, and to name one, Bhutan doesn’t have records on the number of deaths due to covid. The latter later on became somewhat of an obstacle regarding one of the tasks given to the students, namely, computing for Case Fatality Rate. Due to avoiding imputation, no Case Fatality Rate could be computed overall for some countries.

A number of pre-processing steps were taken to organise the data in a way that would fit the needs of the assignment. The functionality of choosing specific components from the dataset was programmed into the function, ‘datasetFilter()’. Another function was coded into the program, named ‘datasetAdjust()’. This function would sum each value of every new case and death record into one-month intervals and perform a cumulative sum for total cases and deaths. The function would then group these in relation to the country where the data originated from. In order to avoid imputation, a minimum count was set to 1, as to avoid the pandas groupby() function from imputing blank cells to 0. Finally, as was tasked in the spec, we had to compute for a Case Fatality Rate. This was done through the ‘addCaseFatalityRate()’ function, which divided total deaths by total cases . As was stated in the previous paragraph, some countries could not be computed a Case Fatality Rate as a result of avoiding imputation.

Figure 1: scatter-a.png

A scatter plot was used to represent the data from a slightly modified version of parta1. The modification in question was, rather than making the data monthly per country, it was making it yearly. Essentially, this made each country/location have one point on the scatter plot. On the y-axis is Case Fatality Rate, and on the x-axis is the number of Confirmed New Cases. The difference between scatter-a.png and scatter-b.png is that the latter uses a log-scale on the x-axis. Aside from a few outliers, it is noticeable in both plots that the data-points tend to be around a similar value in the y-axis. This is particularly notable in scatter-b.png.

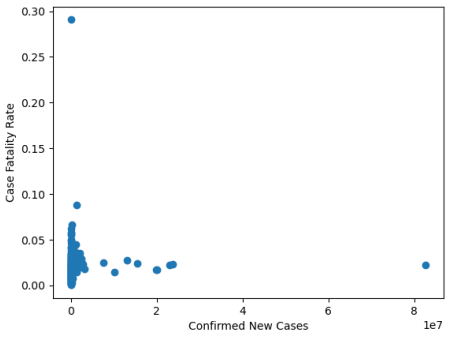

Figure 2: scatter-b.png

Despite both graphs being composed of the same values, it is notable in scatter-a.png that a majority of the data points are compressed to the left-hand side of the graph, and there are some datapoints a large degree away from them. This seems to be tackled through the usage of the log-scale in the x-axis, which further reveals the trend that the data generally lies around a close-range of Case Fatality Rate values. However, while there is a trend in the data, there is no correlation.