Visual analysis of OWID COVID-19 dataset (2020)

Name: Rory Healy (964275)

Word count: 584 words

The COVID-19 dataset maintained by *Our World In Data (OWID)* is a large, public dataset containing information on the confirmed cases, deaths, tests, hospitalisations, and vaccinations regarding the COVID-19 disease. It is updated daily with information compiled from various sources, listed on their public GitHub page. The raw CSV data contains many columns, though for the purpose of this visualisation, the only columns needed were `location, date, total_cases`, and `total_deaths`. The dataset is limited though, as it is not fully complete. It lacks information for some locations for the entirety of 2020 (such as Hong Kong), and is not reflective of the actual true values, but the reported values. As a result, it is common to see negative numbers of new cases or deaths as reports get revised.

Pre-processing steps that were taken involve filtering out all records except those on December 31st, 2020. The last date of the year contains the total number of cases and the total number of deaths for the year, which were then used combined in a DataFrame and a third attribute was created – the case fatality rate for 2020 for each country. This was used for the following visualisations. However, not all locations were selected for visualisation and had to be filtered out. A notable outlier was Yemen, which had the highest case fatality rate of roughly 23%, far higher than all other locations.

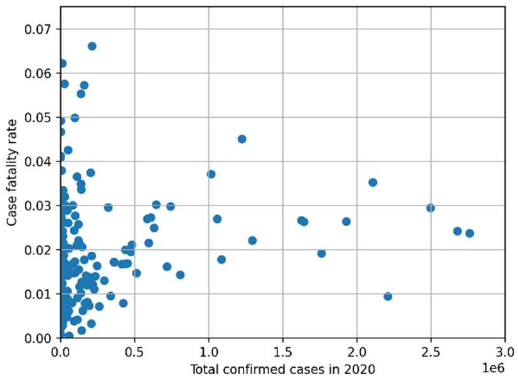

Figure 1 – Case fatality rate plotted against the total confirmed cases in 2020, by country

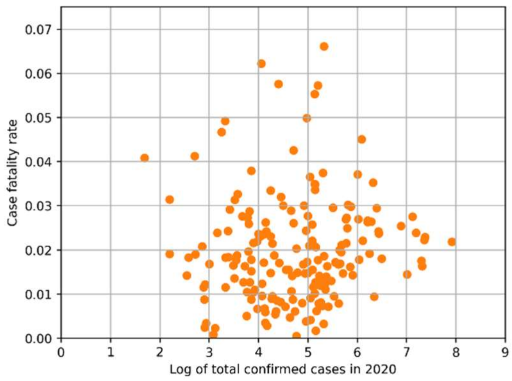

Figure 2 – Case fatality rate plotted against the log (base 10) of the total confirmed cases in 2020, by country

The two scatterplots generated are shown in *Figure 1* and *Figure 2*. *Figure 1* has the case fatality rate plotted against a linear scale of the total confirmed cases in 2020, where each point represents a country. *Figure 2* has the case fatality rate plotted against a logarithmic scale of the total confirmed cases in 2020, with each point representing a country. In both plots, Yemen is excluded as an outlier with a much higher case fatality rate, and in *Figure 1*, the ‘World’ location is also excluded.

Figure 1 shows a large cluster of locations with varying case fatality rates, all seemingly close to zero total confirmed cases. Looking at the underlying dataset, it is clear to see that this figure is not an accurate portrayal of most locations, as the data is mostly blocked view in the plot, and thus very little information about most locations can be extracted. However, for locations with a high number of total cases (such as Asia, North America, European Union), there does appear to be a weak positive linear correlation.

Figure 2 shows a cluster of locations with varying case fatality rates. There does not appear to be enough of a correlation between the two variables to conclude that they are even weakly correlated. Instead, there are a few locations with a very high case fatality rate (> 0.04), several with a very low case fatality rate (<0.01), but most locations fit somewhere in between, with no correlation to the total number of cases. For example, the ‘World’ location (the rightmost point on the plot) has a very similar case fatality rate to Eswatini despite having very different population sizes, and very different numbers of cases.

Contrasting the two figures, *Figure 2* is a much better representation of the data as it makes the visual encoding much more transparent. The plot isn’t hard to interpret, nor is the plot unclear. *Figure 1* hides most locations as most locations did not record over a million cases. However, it does show a weak correlation at much higher orders of magnitude, which is harder to see in *Figure 2*.