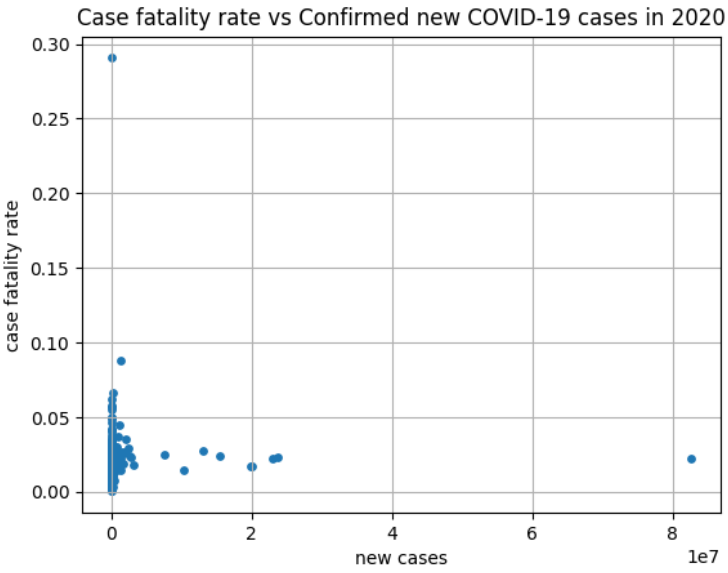
OWID COVID-19 2020 Visual Analysis Report

The data used to produce the plots contained within this report was sourced from Our World in Data at https://covid.ourworldindata.org/data/owid-covid-data.csv.

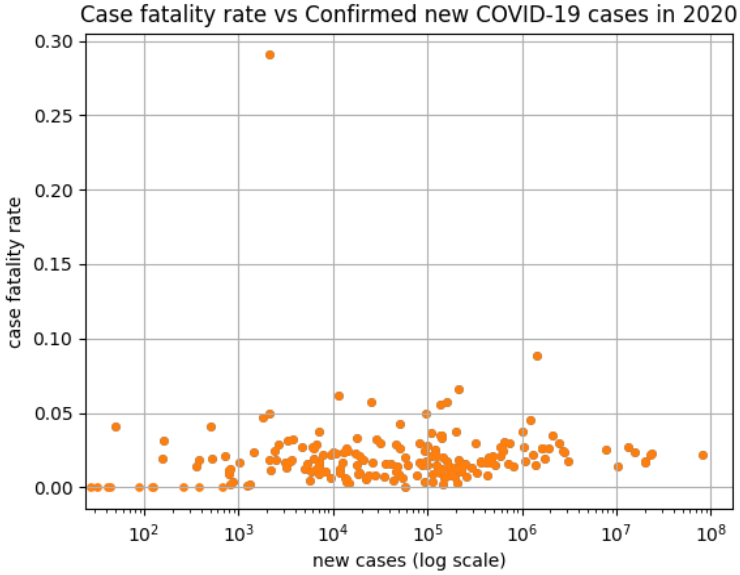
The raw data includes a vast number of values relating to COVID-19 cases and deaths over the past couple of years in many locations throughout the world. To produce the plots, the data was filtered down to 2020 total cases, total deaths, new cases, and new deaths, and was grouped together by location so that there was a single set of 2020 data points per location. There were negative values present in the filtered data which have been interpreted as corrections in death / case counts and were therefore left unaltered. However, these negative values could potentially be errors in data entry have therefore possibly led to slightly inaccurate data. It is worth noting that these errors would be insignificant to the visualisations due to the scale of the plots in comparison to the small magnitude of these potential errors. There were data values for new cases / deaths that have been assumed to be missing not at random and these have been imputed by replacement with the value 0. It has been assumed that missing new cases / deaths could potentially signify that no new cases / deaths occurred. However, it again is possible that these missing values were the result of error in data entry or inaccurate reporting of data in the relevant location (missing completely at random).

The plots compare the number of confirmed new cases in 2020 to the 2020 case fatality rate (calculated by dividing the number of deaths in each period by the number of new cases in each period) for each location. Upon inspection of the plots, similar case fatality rates can be observed in both locations with a lower number of new cases and locations with a higher number of new cases. A line of best fit between the two variables would be approximately horizontal and therefore the correlation coefficient between these variables is approximately 0. This suggests that there essentially isn't any correlation whatsoever between the case fatality rate and the number of confirmed new cases in a given period. The outlier for the new cases is the number of new cases which occurred across the whole world, which naturally is going to be significantly larger than all the other samples as it is the aggregation of all other data points. The outlier for the case fatality rate (Yemen) has a rate more than 3 times higher than the next highest case fatality rate which could either be due to an extremely poor quality healthcare system in Yemen or inaccurate reporting of their new cases and deaths. The mean case fatality rate appears to lie around 1-3% indicating that, on average, approximately 1-3% of new cases in 2020 resulted in deaths. This can be confirmed by the world data point which appears to have a rate of about 2.5%

Plot A



Plot B



It is extremely difficult to analyse any patterns between most of the samples in Plot A due to the inclusion of the number of new cases in the world which stretches the x-axis by a large amount. It is considerably easier to study the distribution of the samples and identify patterns in Plot B due to the logarithmic scale on the x-axis; more specifically, the outlier does not stretch the x-axis in Plot B due to the logarithmic scale. However, the magnitude of the spread is somewhat lost and can much more clearly be observed in Plot A where, for example, the difference between the maximum number of new cases (the world) and the rest of the locations is more accurately represented. Plot B was created using pre-processed data in which missing new case counts and case fatality rates were assumed to be 0 while Plot A was created without any imputation being performed and is therefore missing some data points compared to Plot B.