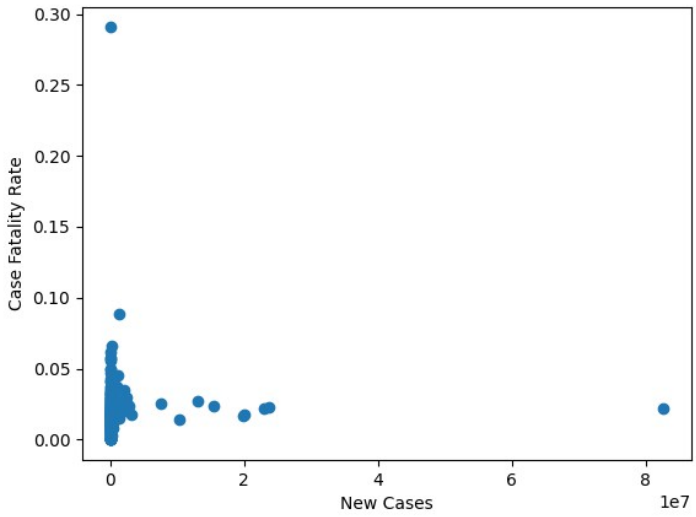
## Pre-processing Report

The raw data we are given is owid-covid-data.csv, which is a collection of covid data provided by Our World in Data. The dataset given contains data on all countries, continents, and the world, outlining their data collected on the Covid-19 pandemic and other generic information about their health and hospital capacity.

The dataset contains 73385 columns and 59 columns although we only use 6 of the columns for our data analysis, which include location, date, total_cases, new_cases, total_deaths and new_deaths. In which we aggregated by month and location for the year of 2020, e.g. aggregating all of the data for Australia in January gets grouped and aggregated into a new Australia – 1 – Data… row.

The data does have some limitations, with some countries starting their covid-19 data collection later then others, the earlier months can be a bit hit or miss regarding wether or not they have data. Alongside the fact that there was no worldwide consensus on how the data should be collected, their may also be slight discrepancies between countries.

## Visual Analysis

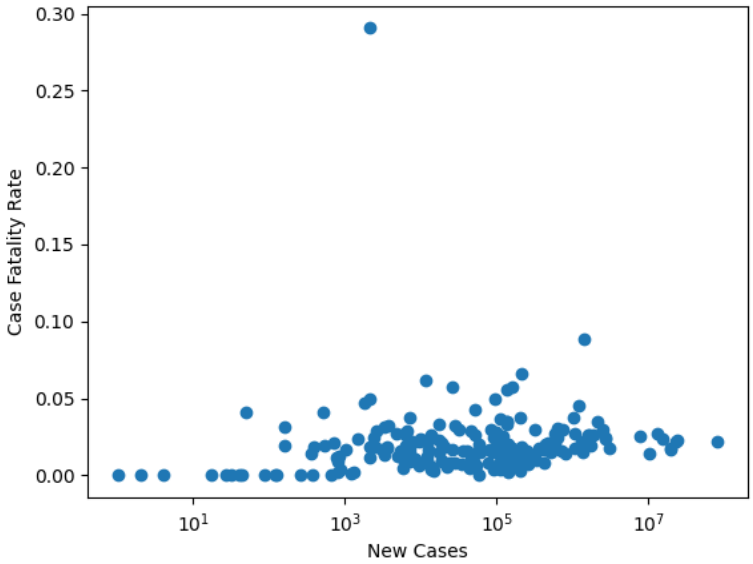



Scatter A compares the Case Fatality Rate on the Y-axis with a linear scale to the number of New Cases on the X-axis also with a linear scale. With each dot representing one region which can be either a country, continent, or the globe.

The 3 groupings of data are independent countries on the far left which are hard to read, the 7 continents slightly spaced apart in the middle, and the world total to the far right.

One of the outliers on our graph our Yemen with a near 30% fatality rate which could be attributed to either poor testing, or more worryingly a bad grip on combating the virus. This datapoint skews our y-axis and makes the graph a lot harder to compare the countries with a lower fatality rate.

The second outlier on our graph is the data point representing the World, which as much as being a good data point to show the world average severely skews the rest of our data to the left making it much harder to read.




Scatter b shows the same data as scatter a, but the scale on the x-axis is now logarithmic, which helps to negate the effects of having the larger world total of new-cases.

We can still see the 3 groupings of the world, the continents and then the countries. But the log scale allows us to see disparities between countries easier.

The large fatality rate of Yemen still skews our graph to contain a much larger y-scale then needed, which makes it harder to compare the other data points.

Making the y-axis logarithmic as well could help to solve our issues of skew on the y-axis, but overall, it would probably be a better option to just remove the data point of Yemen, to allows our axis to show our data better.

It could also work well to change the colours of our continent and world data, so that viewers observers don’t get confused by the large disparity between the rest of data points and those of continents and the world.