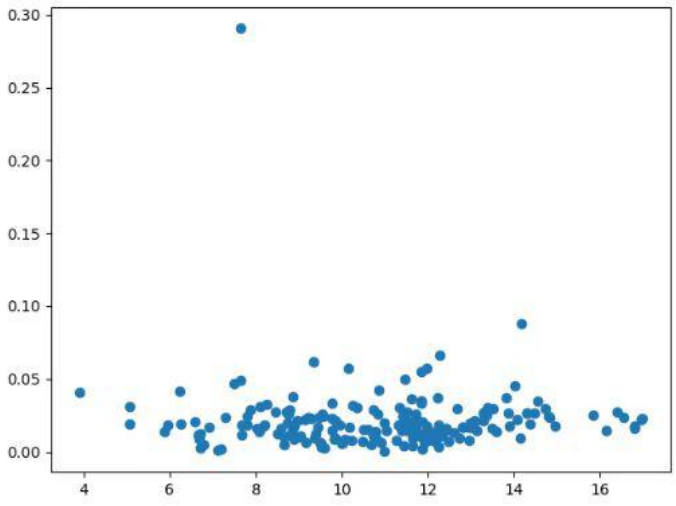
The raw data shows us how the COVID-19 impacted our world. These data recorded the daily situation of the coronavirus epidemic in different countries around the world from year 2020 to year 2021. The variables of this dataframe including the country, date, total confirmed cases, new cases each day, total deaths, new deaths each day and many more . Moreover, these data sorted by the initial letter of alphabetical list from A to Z, and date from earliest to latest, among different countries.

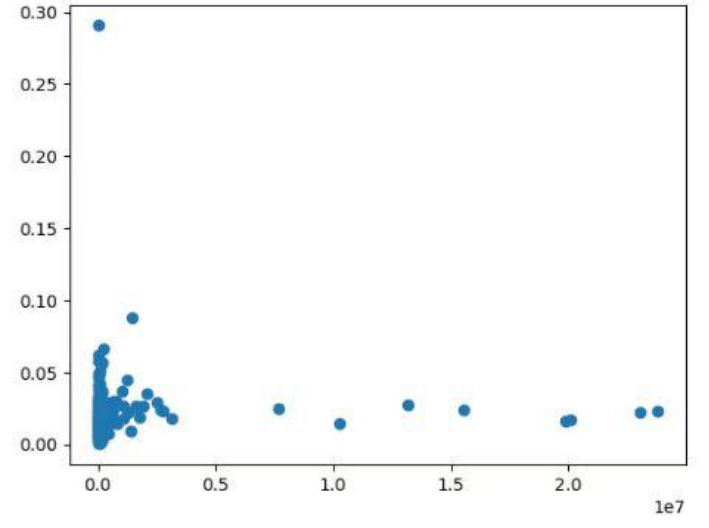
However, we can find some limitations from this dataframe. First of all, for different countries, the start and end point of recording date could be different. For example, Afghanistan recorded it's first confirmed case at 2020/02/24, however, other countries like Cuba and Germany has first confirmed case on 2020/03/12 and 2020/01/27 respectively. The recording date mismatch may bring some influence to our subsequent data analysis. Secondly, we have some redundant data in our dataset, which including population, population density, age group, gdp, handwashing facilities, hospital beds and many more. We can find that, for the same country, these data will not change regardless of the date of data recording, which means we don't need to record them so many times to occupied unnecessary data space. At the last, there are several missing value in the data, and this may cause the NaN value returned when we do the analysis.

To produce the visualisations from our raw data, we need to do the following steps.

1. Joint the data by year 2020 and location
2. Define the maximum value for total_deaths and total_cases variable, the sum value for new_cases and apply to the every countries and continents.
3. Define case_fatality_rate as total_death / total_cases, calculate the value and apply to the every countries and continents.
4. Draw a scatter plot, input case_fatality_rate at the y-axis, new_cases at x-axis as the scatter-a plot
5. Draw a scatter plot, input case_fatality_rate at the y-axis, take the log scale on new_cases value and put it on x-axis as the scatter-b plot



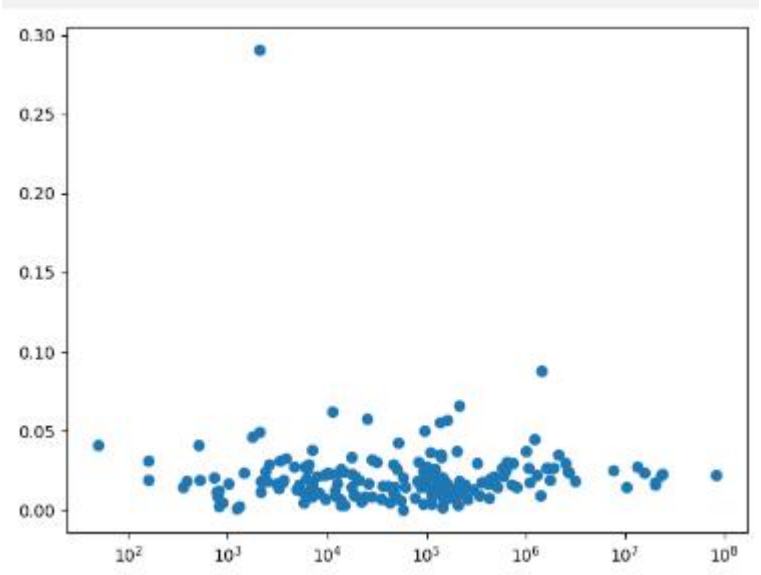
scatter-a plot



scatter-b plot

We can find from the scatterplot, there is a clear outlier with nearly 30% case fatality rate, which is the data represents Yemen. This may be caused by insufficient medical conditions or the government's failure to prevent and control the COVID-19 epidemic. Moreover, we can know that there is no clear pattern between two variables, which means the larger number or smaller number of new cases can't lead a higher or lower case fatality rate, For the most countries, they can keep the case fatality rate below 5%, somehow, still have some economically deprived countries like Sudan, Somalia, or Bolivia, their case fatality rate is slightly above the 5%, but still can keep it below the 10% level. When we take the log scale on the new_cases value, we can find that there are some blue docs with large number of new_cases value, and we can find these represents the different continents or world data, but as we can see, the case fatality rate still under 5%, which means in the worldwide angle, the governments still can control this COVID-19 very well, to prevent people dying from the epidemic.

To compare these two scatter plots, we can find for the y-axis, there is nothing changed. But on the x-axis, scatter-b plot with a very large value, before we adjust the accuracy, which means people can't observe the pattern or results very well. So we can choose log to let the accuracy become a power of 10 or using plt.xscale('symlog') command to optimize it, thus make it easier to visualize.



Optimized scatter-b plot