Owid COVID Visual Analysis

Truong Giang Hoang: 1166323

I. Introduction & Description:

The given dataset provides information about statistics of different countries during COVID on a daily basis. It also summarizes these statistics with regard to each continent in particular and the world in general.

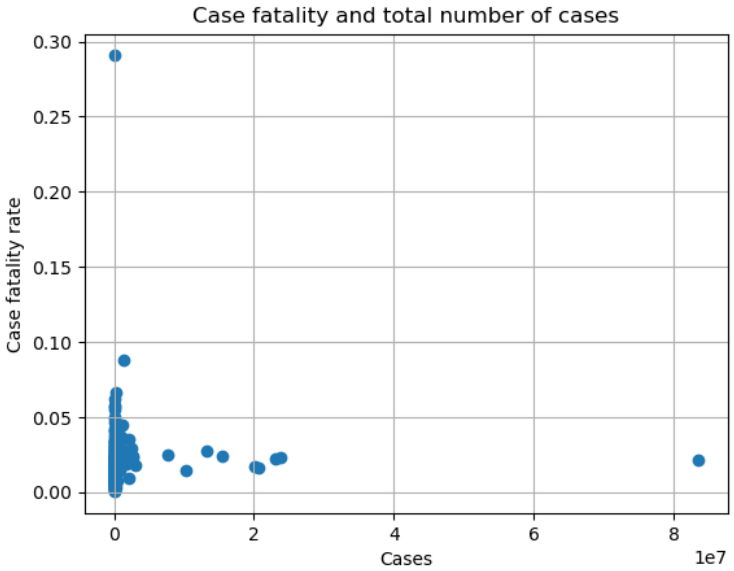
Dataset is retrieved from https://covid.ourworldindata.org/data/owid-covid-data.csv, with important data fields being the number of new cases, new deaths, total cases and total deaths. The dataset covers a period of February 2020 to present day. As only the year 2020 needs to be taken into consideration, any data from 1st January, 2021 should be removed. As of 14th April, 2021, the data set has a total of 81,223 entries, and as of 31st December, 2020, there were 83558756 total cases, with 1824715 fatalities.

A limitation when preprocessing is a considerable lack of data, represented by multiple Nan values throughout this dataset, which makes it more challenging considering case fatality rate must be calculated. In addition, there are a number of redundant data fields in which there are no values. Therefore, when preprocessing the dataset, only data in 2020 can remain, which are then aggregated by location and month. Unnecessary columns are also oonly non nan values are visited. When calculating case fatality rate, integer division of new deaths by new cases is required, so places where there is 0 new case in a given month are untouched.

II. Plotting:

A revised case fatality rate must be calculated for both, which is equal to total deaths divided by total cases as of 31st December 2020.

1. Plot 1:

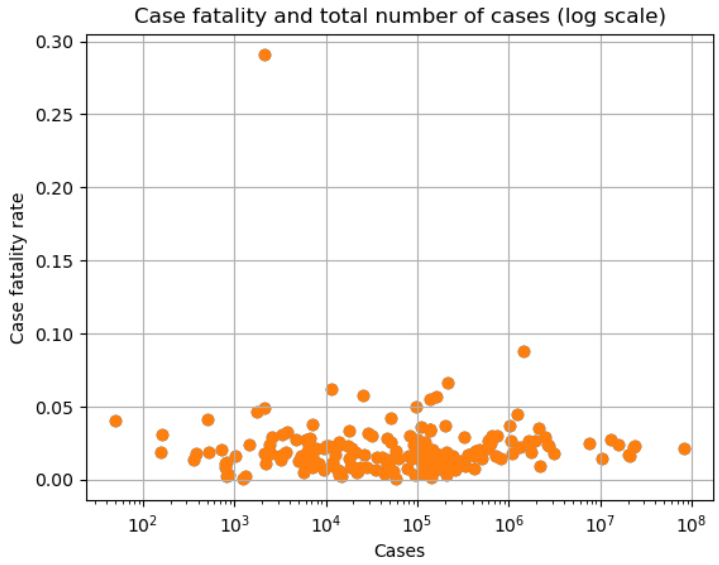


The 'Case fatality and total number of cases' plot has its x-axis being the number of cases, y-axis being the yearly fatality rate. The plot includes not only data from all countries, but also a summary of each continent and the world.

Most data points are clustered below the 0.10 mark for case fatality and 0.5*10^7 for cases. However, data are too clustered, so no further visible trends can be observed.

The data point corresponding to over 80 million cases is a summary of the whole world, with fatality rate of 0.022. A few other data points around the 2*10^7 mark, each belongs to a continent and 3 countries: United States, Brazil and India. Yemen has the highest case fatality of 0.29, with 2099 cases and 610 deaths.

2. Plot 2:



The 'Case fatality and total number of cases' plot has its x-axis being the number of cases in log scale, y-axis being the yearly fatality rate. The plot includes not only data from all countries, but also a summary of each continent and the world.

From this plot, it is observed that the majority countries have between 10^3 and 10^7 cases. The 10^5 mark is where data are the densest.

No further trends are observed in case fatality rate, as rescaling is not applied. The data point at the top of the plot is of Yemen.

A linear behaviour is observed, suggesting that the case fatality rates are largely the same throughout the world, and are relatively low, most of which are lower than 0.05.

III. Discussion:

Both plots provide the same information about number of cases and fatality rate. Therefore, the distribution of data along the y-axis is the same. However, plot 2 is more interpretable, thanks to its evenly distributed data points.

Rescaling has a considerable effect on how data are presented to users. An appropriate scale helps identify trends faster and more effectively, while a dense distribution of data will lead to failure to investigate certain properties of the dataset.