Introduction

The data source from an organization, Our World in Data, which they collect data from researchers all around the world. They update the database every now and then. The raw data of covid-19 cases downloaded from this organization consists many sections, the most important data are total cases, total deaths, new cases and new deaths, together are some other dataset that can also be used.

The limitations of this data are, first there are lots of columns in the csv file, however, many of the data in those columns are empty for many countries. This could affect the usability of this data, if the researcher needs specific data that is missing. Second, some of the data is negative, which means the source of the data may not be reliable for data processing. When process the data using functions, these numbers might be included, and affect the final result. If the program goes through all data and process it, this problem can be solved, however, the runtime will surge.

Preprocessing steps:

1. Read the file online. Take out the data that is needed in each column. Separate the 6 data into two group. One is total case/death, and the other is new cases/deaths
2. The reason of doing this is because it makes it easier to process data. The data needed for total case/death is the maximum number of each month, since it's already added up. New case/death needs the sum up data for all days in a month
3. Convert date to monthly based, use groupby function to separate each country and month, while finding total case/death(maximum) and new case/death(sum)
4. Combine two group of data using concat function (location and month)
5. Pick the data in total death and total case column, then calculate monthly case fatality rate, add into the dataframe using concat function, rearrange the columns into the required format
6. Convert monthly data into yearly data, then calculate the case fatality rate again (using the same method when calculating monthly data)
7. Use the yearly data to produce two scatter plot of case fatality rate vs. new cases, including axis names. The points in the scatter plot is colored, which makes it easier to observe.

Explanation:

Observing the first graph, the majority of the case fatality rate lays in the left lower corner of the scatter plot. The majority of the case fatality rate lays between 0 to 0.05, with some data lays in between 0.05 to 0.1, and the outlier that has a rate of around 0.29. Observing this graph along, it's hard to find matched pattern, and couldn't have a clear view of the distribution of number of new cases.

Observing the second graph. The case fatality rate is exactly the same and has the same range, however the data is distributed on the bottom of the scatter plot. From which it is easily observed with a null pattern. It is also clear that the distribution of number of new cases follows a normal distribution, where it disperses from $1*10^5$.

Discussion:

Comparing the two graphs, the second graph with the log scale, is better than the first for data analysis.

It not only shows the distribution of the case fatality rate, but also distribution of the number of new cases. Observing from the graph, the null patterns suggests an uncorrelated relationship between case fatality rate and number of new cases. The number of new cases has a normal distribution and spreads from $1*10^5$. The case fatality rate is also a normal distribution with a mean of around 0.025.