Description of the Raw Data:

The project made use of a COVID-19 dataset with information from around the world. The dataset contained attributes including location, total cases of COVID-19, new cases by time period, new deaths, GDP per capita among other things. For this project, only confirmed cases and deaths were used which was sourced from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). While the dataset is quite large, there are many missing values. Since the data is reported daily, there is a possibility of inaccuracy. As mentioned on the GitHub repository, data might be inaccurate due to the long reporting chain. There is also inconsistency in usage of the source. Until November 2020, data for confirmed cases and deaths were sourced from European Centre for Disease Prevention and Control.

Preprocessing:

The data was provided in a csv file from a url. The preprocessing started by turning the csv file into a data frame for easy readability. Next, I proceeded to make a new data frame from our original one with only the columns that were required in our analysis – location, date, total cases, new cases, total deaths and new deaths. Since, we only need to work with the year 2020, I made a subset data frame consisting of data from 2020. I added a new column to the data frame called 'month' that showed the month data from the 'date' attribute. The 'date' column was deleted, and the remaining columns were grouped and aggregated based on 'location' followed by 'month'. 'Case_fatality_rate' was calculated based on total deaths divided by total cases. For the purpose of visualization, the data was annualized. In the first graph, confirmed new cases was put on the x axis and case fatality rate on the y axis. The second graph was similar with the one exception of applying log on the x axis.

Scatter a:

Scatter plot a show that the case fatality rate for most countries is within 5% with some between 5% and 10%. However, there is an outlier whose case fatality rate is above 20%. In terms of new cases, most countries are doing well as they are hovering around 0. There are some countries where the number of confirmed cases is concerning with one outlier where number of confirmed cases is very high. In conclusion, it can be said that most countries are doing a good job of managing the crisis.

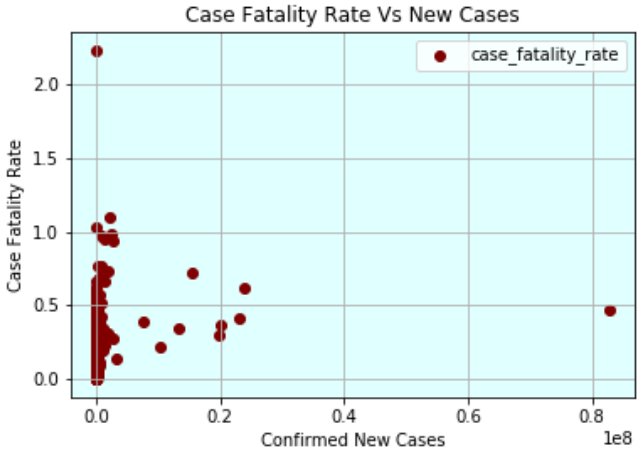

Figure: Scatter a

Scatter b:

There are no outliers for confirmed new cases. Most of the countries are situated between 5 and 15 confirmed cases. The outlier for case fatality rate can still be observed here. Case fatality rate is still below 5% for the majority of countries present in the dataset. New cases are more uniformly distributed for this graph.

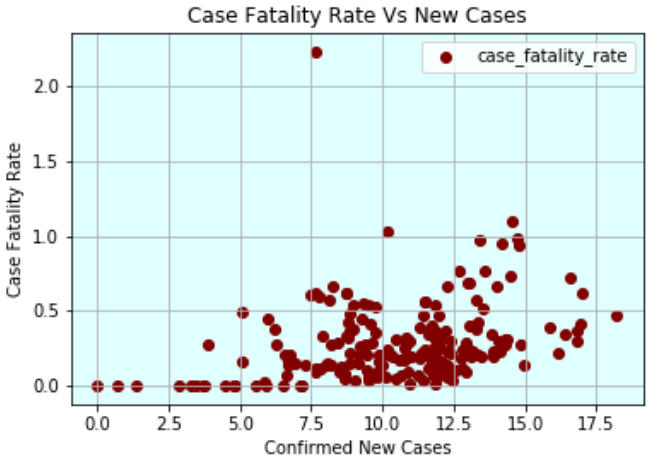

Figure: Scatter b

Comparison:

The usage of log in the second plot brings more clarity to the confirmed new cases. In the first graph, majority of the points are situated on the left corner. In the second one, the data is more spread out allowing us to read the graph properly