# COMP20008
# Elements of Data Processing

## Assignment #1 - COVID-19 2020 Report

Xing Yang Goh
# 1001969
April 14, 2021

# Introduction

The data is acquired from a collection of up-to-date Covid-19 data around the world from reputable publishers such as JHU, WHO, ECDC, etc, maintained by Our World in Data and presented in a csv data format: Covid-19 Data-set. This data-set contains an extensive volume of data, with some notable key columns on location, date, total cases, new cases, total deaths, new deaths, etc.

**Potential issues that can be encountered in the data-set:**

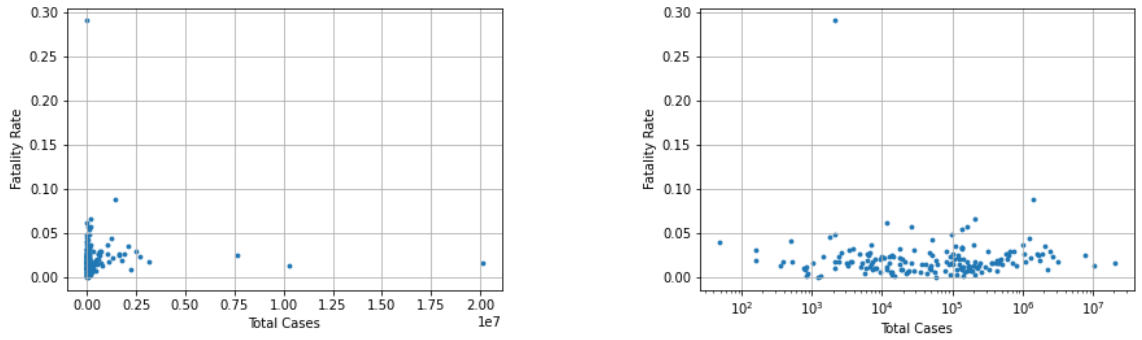
- Inclusion of non-countries Covid-19 data such as aggregated data for different continents, the entire world and the European Union.
- Rows that are missing data for certain columns.
- Non-uniform starting date since the data only begins after a country acquires it's first case and no month column.
- Data-set collects data beyond 2020 and includes columns that are not relevant to this study, which is not required for the purpose of this analysis.

# Pre-Processing and Visualisation

To combat these issues, the data-set is first passed into a pandas data frame that allows for easy pre-processing. The date column is converted into a format that pandas will be able to interpret, then the data points that are not in 2020 will be removed. After this, a numerical month column will be added and only the required columns of the data frame will be kept as per the specifications. Calculation of the case fatality rate follows, taking the fraction of new deaths to new cases for each month and adding it to the data frame. Lastly, the data frame is sorted to location then month in ascending order and saved to a new csv file.

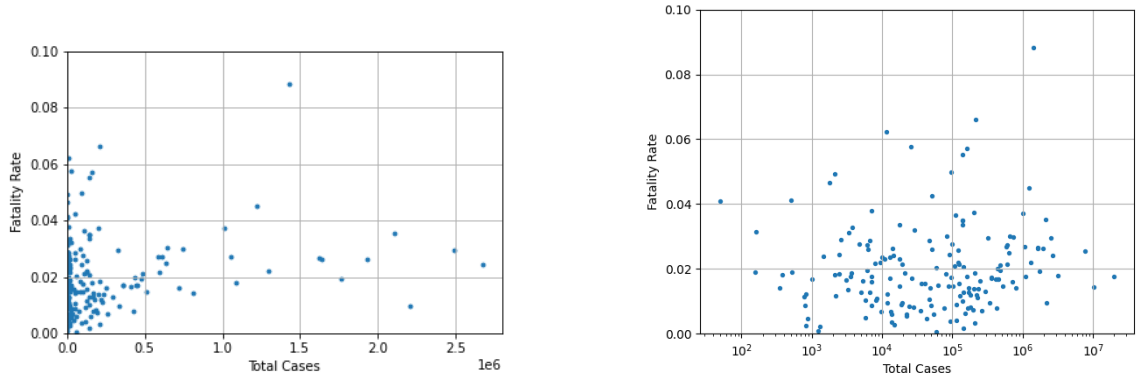
To visualise the data, matplotlib is used to create the scatter plots. Due to the high number of data points making the visualisation of data difficult, instead of plotting each month's case fatality rate to the number of new cases, the case fatality rate is calculated as the fraction of the total number of deaths in a year to the total cases for each country. This improves the clarity of the scatter plot with each data point representing a single country. Further clarity is obtained by adding limits to the scatter plot for far outlier data points and removing the data points of non-country regions such as aggregated data for the world and continents.

# Scatter Plots




(a) Linear Scale (b) Logarithmic scale

Figure 1: Total cases to fatality rate no limits




(a) Linear Scale (b) Logarithmic scale

Figure 2: Total cases to fatality rate with limits

# Discussion

These scatter plots presents the relationship between the total Covid-19 cases in 2020 vs the fatality rate for all the cases in the year. Each dot on the scatter plot represents a single country, as regions such as continents have been removed. Outliers in the data are present from Figure 1, with Yemen having a case fatality rate at almost 0.3, which is much higher than the world average of 0.0218. Also, in terms of the total number of cases, the United States accrued a total of 20 million cases in 2020, a figure much higher than of other countries. Other than outlier data, most countries reside within the 0 to 3 million cases figure with fatality rates from 0 to 0.06 and there is no correlation between the number of cases and fatality rate of countries. This is especially clear in Figure 2b, where the data points are randomly scattered with no distinct relationships. Contrasting the scatter plots with linear scales and logarithmic scales, the logarithmic scale presents a clearer visualisation of the relationship between the two axes, while the linear scale presents a clearer visualisation of where the majority of countries reside. This is because the logarithmic scaling removes the factor of country populations as countries with large populations will naturally have higher number of cases, and this is not the relationship investigated in this visualisation.