Elements of Data Processing
Leoni Angela 1179015

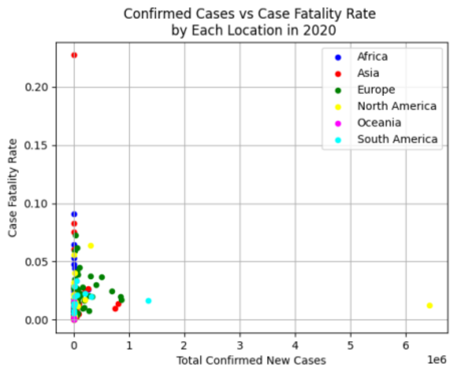
Scatter Plot Visual Analysis

The raw data used to produce these visualisations are the "Our World in Data COVID-19 dataset" from the website: https://covid.ourworldindata.org/data/owid-covid-data.csv. This raw data comes from Our World in Data, a scientific online publication based at the University of Oxford. Our World in Data collects data that revolves around global problems such as poverty, disease, climate change, and, in this case, a pandemic. The raw data taken provides a collection of data related to the COVID-19 pandemic. It provides confirmed cases, deaths, hospitalisations, testing, and vaccination data.

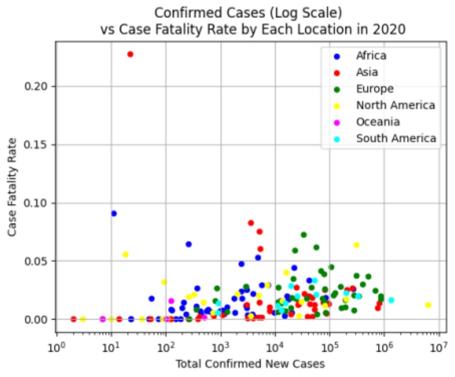
For this specific analysis, the raw data is processed to take the total cases, new cases, total deaths, and new deaths grouped by each location and month in 2020. The raw data is first pre-processed by filtering and grouping the needed variables. The "date" column is broken down into year and month, and then the data can be grouped by locations.

New cases and new deaths column is aggregated using the sum function, and max function finds total cases and total deaths for every month in that location. Another column, the case fatality rate, needs to be derived from dividing total deaths by total cases. To ensure that only valid data is processed, all location and month entries that are empty are removed.

To produce the visualisations, plot the case fatality rate by total cases. The data points are also grouped by continents for further analysis. For scatter plot 'a', the data points are too close together and sometimes overlap. However, there are some interesting observations from scatter plot 'a', there is a location in Asia where its case fatality rate skyrockets above other locations. This location might not have that much COVID-19 case in comparison to other locations, but the fatality is much more prominent. A location in North America records a large number of confirmed cases with a relatively average fatality rate. Indicating that this place might have a large population and a moderately effective healthcare system. The way that this scatter plot is made exaggerates and highlights some outliers within the data.



scatter-a.png



scatter-b.png

Elements of Data Processing
Leoni Angela 1179015

For further analysis, scatter plot 'a' is processed again by scaling the x-axis by log. The produced visualisation provides more insight, as the data points are more visible. Some patterns that can be observed is that there is a common fatality rate ranging from 0.00 to 0.05. Once the x-axis is adjusted the pattern produced from the data points indicate that both Asian and African locations tend to have a higher case fatality rate than other continents in comparison to the number of new cases recorded. This could signify that in some Asian and African locations, handling of the COVID-19 pandemic is less effective. European locations are clustered more to the right section, indicating that they have similar amounts of new cases and a similar fatality rate. This data could be supported by the geographical fact that most European locations are near each other (within one land). This might indicate that the proximity between European locations could closely affect the way the pandemic spreads there.

To better understand the data, the interquartile range of the data set could be taken and made into a scatter plot. Observations and analysis from that plot could be even more specific then.