Name: Brandon Widjaja
Student ID: 1187107

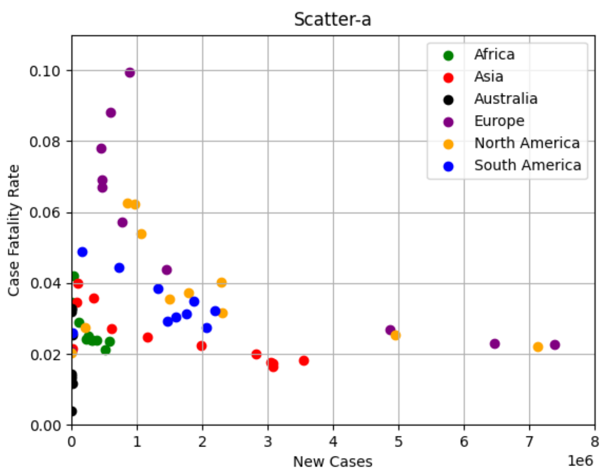
Visual Analysis of Covid-19 Cases in 2020

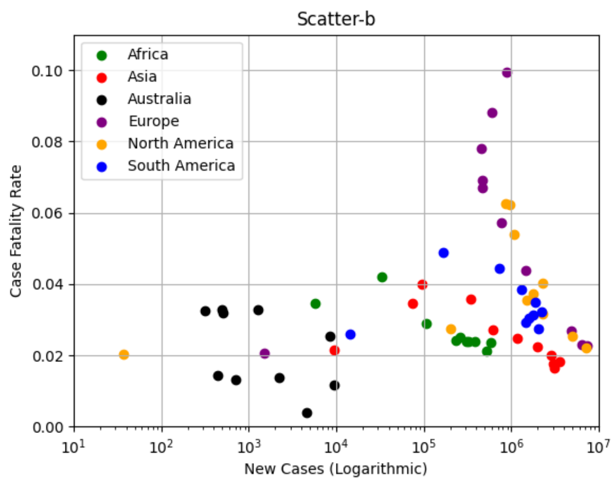
Introduction

This visual analysis is for the relationship between new confirmed covid-19 cases and the case fatality rate for each habitable continent in 2020. The data was sourced from the "Our World in Data COVID-19 dataset" (from https://covid.ourworldindata.org/data/owid-covid-data.csv).

A limitation of the visual analysis includes that to avoid overplotting, I opted to only input data from each continent rather than every country. E.g., instead of plotting every country in Europe, the values were aggregated to represent Europe as a whole. Due to this, the data cannot be analysed on a country-by-country basis and displays only the "big picture".

Some pre-processing was required to represent the desired data. This included converting the dates from a dd/mm/yyyy format to represent only months, making it easier to then aggregate the data by months and locations. The case fatality rates also had to be calculated by dividing the new deaths by the new cases for each month. I chose to represent each continent by a different colour to make the points distinguishable on the plot and a legend was added for identification of each point.





Explanation

Both scatter-a and scatter-b represent the case fatality rates against new cases for each month on each continent in 2020 with the only difference being that the new cases on scatter-b are plotted on a logarithmic scale.

It can be observed from scatter-a that there is a strong negative exponential correlation between the number of new cases and the case fatality rate. The case fatality rate exponentially decreases to around 0.025 as the number of new cases per month is higher. This may suggest that the actual case fatality rate is close to 0.025 for most continents. We can also see that near 0 new cases the case fatality rates are scattered (suggesting no correlation) along the y-axis and that the correlation is much more visible at around 10,000 new cases, however due to the scale of the x-axis, the correlation is unclear.

For scatter-b we can again observe a negative correlation between the number of new cases and the case fatality rate however although it still appears to be a negative exponential correlation, it is much closer to a negative linear correlation compared to scatter-a. In scatter-b, it is much more apparent that there is no correlation between the case fatality rate and the new cases per month for when the new cases are less than 10,000.

Name: Brandon Widjaja
Student ID: 1187107

The scatter plots both suggest that when there are over around 10,000 new cases in each month there is a strong negative exponential correlation between case fatality rates and new cases and that under 10,000 new cases there is no correlation.

Discussion

Although both scatter plots represent the same data, they each have their benefits for viewing and analysing the data.

For scatter-a, the strong negative exponential correlation is much more obvious for the larger values (greater than 10,000), however the fatality rates for the new cases below 10,000 are much less visible and it is not possible to make a conclusion on the correlation of those points.

On the other hand, scatter-b makes it much easier to see that there is no correlation between points under 10,000 new cases and the negative exponential correlation for cases above 10,000 cases is still visible, but it is much less obvious and looks more linear than in scatter-a. In conclusion, scatter-a is much better for viewing the correlation for the larger number of new cases (over 10,000) while scatter-b is much better for viewing the correlation between the lower new cases.