OWID covid 2020 visual analysis

Lang (Ron) Chen 1181506

The raw dataset of 'owid-covid-data.csv' was downloaded from the database "Our World in Data", a scientific online publication that focuses on global problems. The data set provided had several limitations, most notably in the form of missing data which often occurred in large chunks (e.g. large chunks of total death and new deaths in Afghanistan). This could have been because some locations did not record statistics at the start of the pandemic. Another limitation is the appearance of negative numbers for new deaths, which created ambiguity as to whether they were mis-recorded or were amendments for previous days as no explanation was provided.

In order for the two scatter plots to be produced (Part A Task 2 only), several pre-processing steps had to be undertaken. First, the CSV read from the server was loaded into a Pandas DataFrame, and all rows of data representing dates in 2021 were removed as per the requirement of the assignment. Then, the DataFrame was cut down so that only rows 'total_cases', 'new_cases', 'total_deaths' and 'new_deaths' remained, before groupby() was used to aggregate all the data in each location into a single row (for the columns involving 'total', the value in the last row in each groupby object was taken while for the columns involving 'new', the sum was taken). A new column of case_fatality_rate was then calculated based on new_cases and new_deaths, with special consideration needed for instances where either of these data were missing (nan) or the denominator new_cases had value of 0; for all three aforementioned occurrences the case_fatality_rate column recorded nan as per the important instruction to "not impute missing values". There was now enough information to plot the graph for Part A Task 2.1, using columns 'new cases' and 'case_fatality_rate'. For Part A Task 2.2, further manipulation was done in that a list of log10 values of 'new cases' was used as the x axis for the scatter plot instead of just the original 'new cases' list. Colour was used purely for better visualisation.

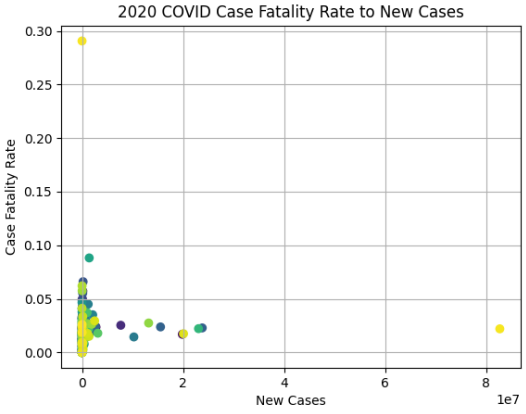

Figure 1 scatter-a.png

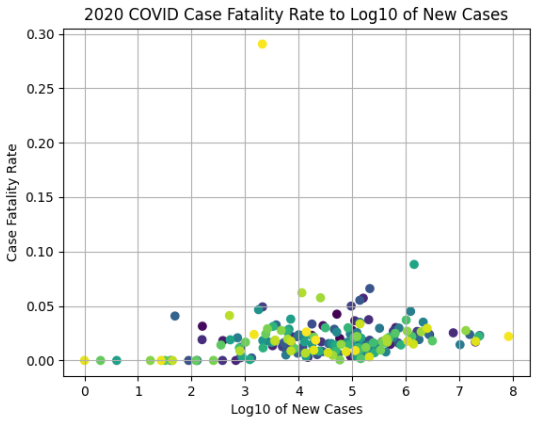

Figure 2 scatter-b.png

Scatter-a.png is a scatterplot of Case Fatality Rates (new death per new case) to New Cases, while scatter-b.png is almost the same except that the x-axis is instead the log10 values of new cases. Patterns observed in scatter-a.png is that the dots were mostly grouped in the lower regions of values for new cases (mostly on/along the y axis). A secondary trend is a weak hyperbolic (negative) correlation, demonstrating a decrease of case fatality rate as the number of cases increased. There are two outliers on the top left and bottom right-hand

corner of the plot. The primary trend in scatter-b.png shows an exponential (positive) correlation, with case fatality rate rising sharply at 7.5 and 12.5. This suggests the fatality rate grows higher with rising confirmed cases. A secondary trend is linear like (constant/horizontal), which arguably exists in all data points between 0 and 0.025 case fatality rate, but however some of these points could be said to be part of the more significant exponential trend. There is one significant outlier at log10 of new deaths = 3.5 at which the cases fatality rate is 0.3.

The two plots showed a contradiction in trends, which could have been due to the fact that the logarithm scale in scatter-b.png freed up the congestion at lower values of new cases in scatter-a.png. The trend in scatter-a.png was also quite weak with few numbers of dots to support the hyperbolic shape, so overall scatter-b.png's (positive) trend is more reliable and better supported than scatter-a.png's negative trend and hence scatter-b.png is more valuable in terms of visualisation than scatter-a.png.

Word count: 598