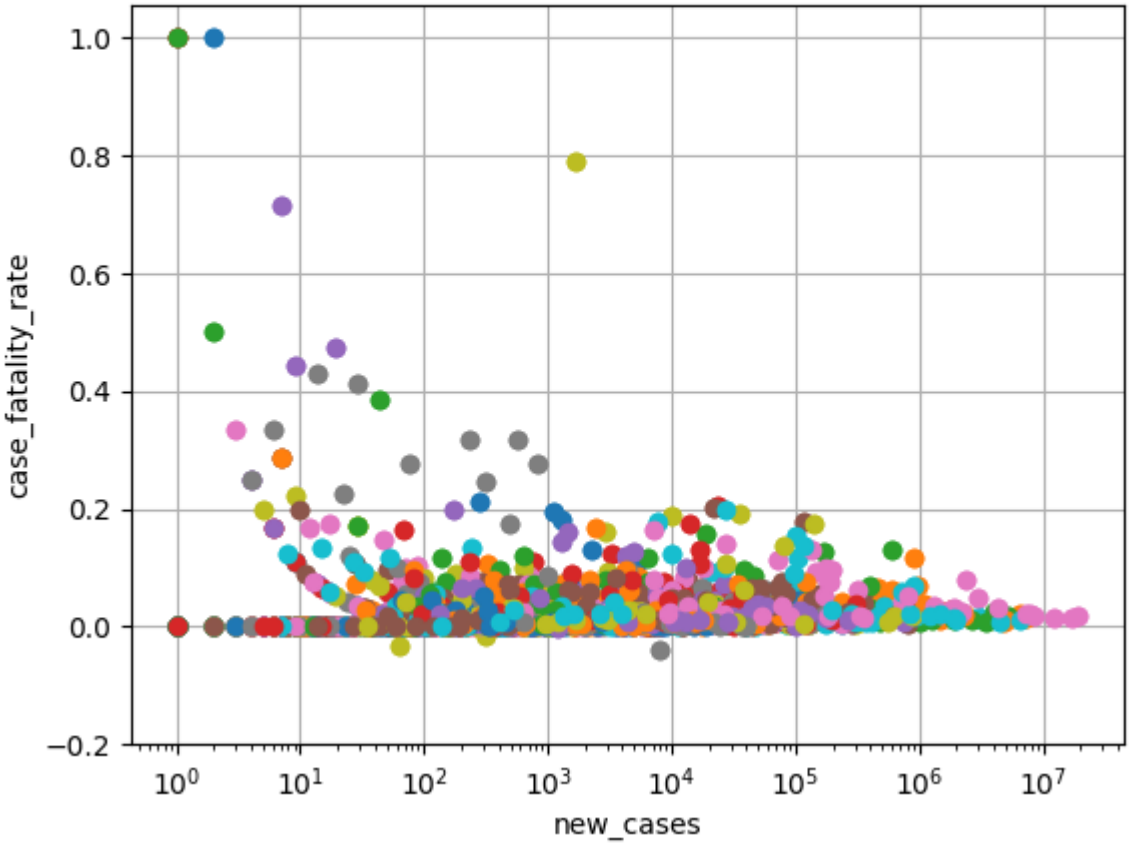
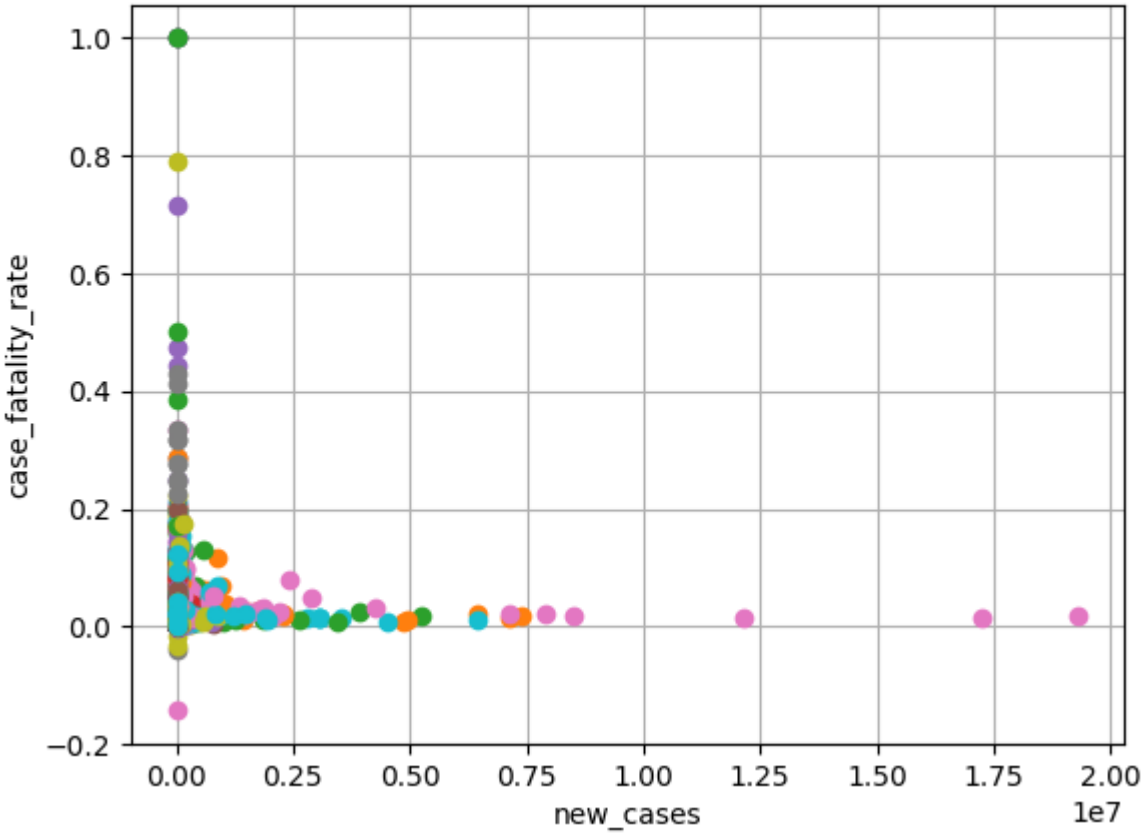
The raw data from the website deliver some features from the different location in the covid-19 period, and the data is retrieved from the GitHub page of 'our World in Data'. The raw data is been processed to a CSV file with a group of different columns, which including the total confirmed cases of covid-19(total_cases), the new confirmed cases of covid-19(new_cases) and so on. There are some missing data in the table, the cases of the missing data always happened in the columns of ICU_patients and the patient in hospital and some columns related to the medical system. Some data might be missing not at random since the difficulties of collecting the data from a location should be considered, but the potential is high that those data in the ICU_patients and hosp_patients are missing at random. The majority of the data from ICU_patients and hosp_patients are missing in almost all the location in the CSV file, whatever the location has a high man_development_index or not. Some missing data seems to impact the visualization. For visualizing the data of confirmed cases and the death cases, the missing data of total confirmed cases and new confirmed cases would impact the result of visualization. Besides, the users might hard to find out the unit of some columns, these might affect the result of some visualization task. Base on the imperfection of the raw data, it could be assumed that a more comprehensive and authoritative source of data might improve the quality of visualization. For preprocessing the data, the daily report is converted to the monthly report. In the process of converting, the monthly new cases are measure by the cumulative sum of each daily new cases in each different month, and the monthly total cases are obtained from the total cases at the end of each month. Besides, the missing data are changed to 0 for avoiding impute the missing data. And also, some NaN value in the case_fatality_rate are changed to 0 for avoiding the case of zero division. In the plotting of scatter-a(case_fatality_rate vs new cases), most of the data points are concentrated on the left-bottom side of the plotting, which means most of the country have similar new confirmed cases and fatality rate. In the scatter-a, some country's case fatality rate is highly concentrated on the y-axis. In contrast, the data from the different location are distributed more evenly on the scatter-b. This phenomenon could be concluded that the range of the x-axis is too large in the scatter-a, it leads most of the data are concentrated on a certain area in scatter-a. After applying the log value on the x-axis, the interval of the x-axis(new_cases) are expanded, so that the visualization is much more readable in scatter-b. For the tendency of the two plotting, some countries have a high case fatality rate even their new cases are at a low level. However, most location's case fatality rate is lower than 0.2 even the new confirmed cases are large. Also, there is some outlier on the graph, for instance, some location's case fatality rate is below zero on the plotting of scatter-b. In general, some missing data might affect the result of the pre-processing stage, and most of the location doesn't have a large proportion of deaths people per new confirmed cases base on the observing on the plotting.


1.0
0.8
0.6
0.4
0.2
0.0
−0.2
case_fatality_rate
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
new_cases


1.0
0.8
0.6
0.4
0.2
0.0
−0.2
case_fatality_rate
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e7
new_cases