Assignment #1: Exploring and Visualizing Data

1. Introduction

COVID-19 (Coronavirus disease 2019) is a pandemic caused by the SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) virus, that engulfed most of the planet in 2020. It is an ongoing crisis. In this report, we present the data exploration results on the daily worldwide COVID-19 new reported cases data, sourced from ECOC (European Centre for Disease Prevention and Control). The data set has counts of daily cases and deaths from countries and territories worldwide from Dec 31, 2020, to date. Countries not listed in the data have no reported cases or deaths. The data set also contains the countries/territories' population in 2019 and a cumulative case count per 100,000 people over the previous 14 days.

2. Data preparation

The data set has 12 attributes. Among them, we are interested in dateRep, cases, deaths, popData2019, Cumulative_number_for_14_days_of_COVID-19_cases_per_100000, and countriesAndTerritories. As part of the data preparation, we transformed dateRep from string format to date-time format (dd/mm/yy). We also removed any records with missing values or records with cases or deaths or the 14-day cumulative counts per 100,000 people less than 0. This resulted in 3018 fewer records in the data. We also computed six new variables per country/territory:

Total cases = sum of all cases to date

Total deaths = sum of all deaths to date

*Case fatality rate = (number of deaths to date/ number of diagnosed cases to date) * 100*

*Mortality rate = (number of deaths to date / total population) * 100*

*Cases Per Million = (number of cases / total population)*1000000*

*Deaths per Million = (number of deaths / total population)*1000000*

3. Data exploration

Based on the mean and median values, we determined that the values for cases, deaths, the cumulative 14 days cases count (per 100K), and the countries/territories' population in the data are severely positively skewed with long tails. For these attributes, the mean is significantly higher than the median. The presence of long tails is indicative of outliers in the data. As expected, there is a strong positive correlation between cases and deaths, regardless of the country or territory. Also, in our analysis, we determined it is imperative to analyze the data individually for each country. This is because of the variations in the availability and quality of medical care, COVID-19 guidelines with each country.

4. Data Visualization

Fig 1 shows the case totals and deaths to date for the top 20 most affected countries. The USA, India, and Brazil are the top 3 countries with the most cases and the most deaths. Among the continents, North and South America have the most cases and deaths, followed by Asia and Europe. Oceania has the least number of cases and deaths.

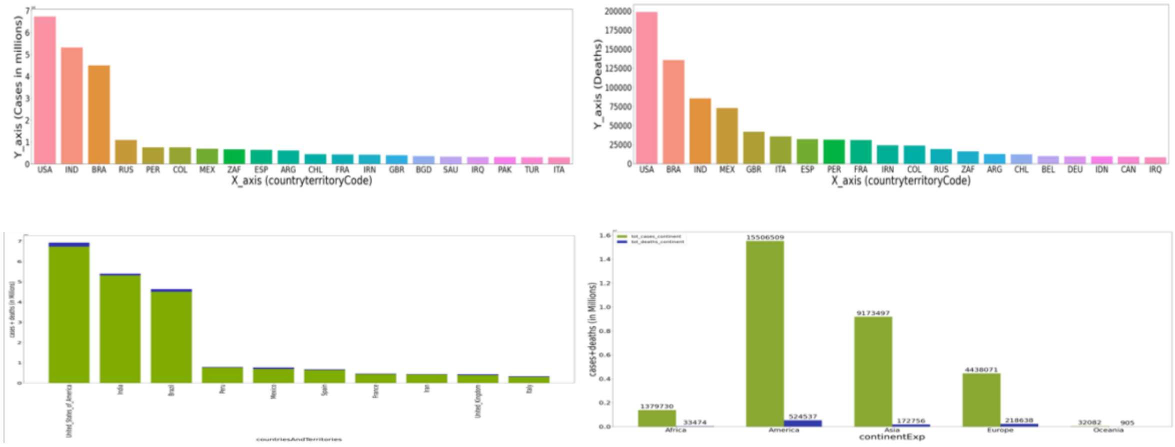
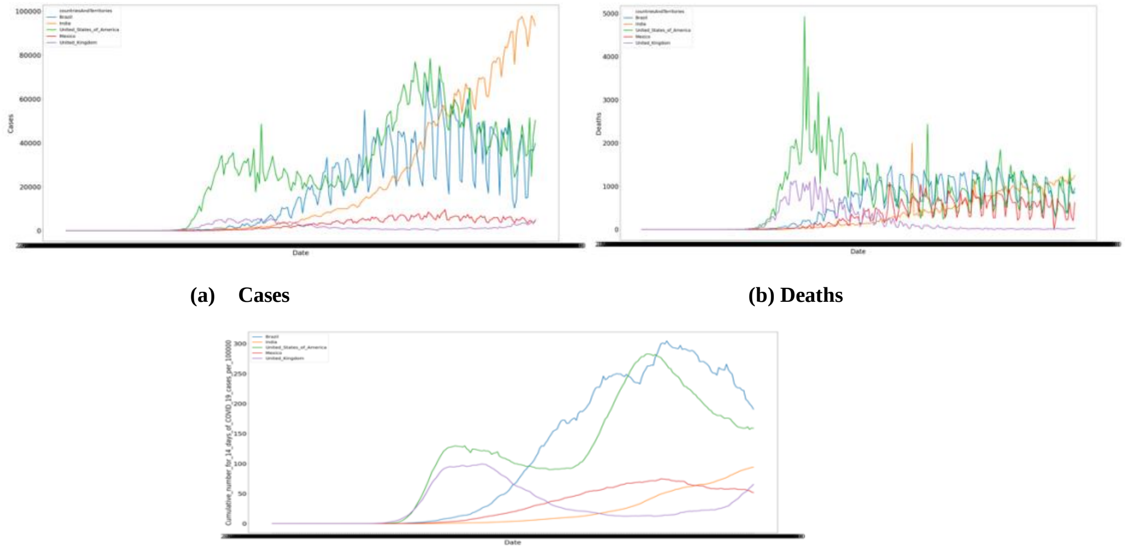


Fig 1: Cases (in million) and Deaths to date– most affected 20 countries

Fig 2 shows the time-plot of the cases and deaths from Dec 31, 2020, to date for the top 5 most affected countries (The USA, Brazil, India, Mexico, and the UK). In Fig 2, for cases, we can note that the USA has two peaks (in March and again in July). Brazil too has two peaks - the first peak in May was much smaller than the second peak in July. On the other hand, India is showing an increasing trend and does not appear to have peaked. For the deaths, to date, the USA had the largest number of deaths per day reported in March and since then, has more or less a stable downward trend, with some upward spikes in July and August. India shows an upward trend in deaths with a small slope. Brazil appears to have a stable trend in deaths, from the start of May to now.



(a) Cases **(b) Deaths**

(c) Cumulative number for 14 days of COVID-19 cases per 100000

Fig 2: Time-plot of cases, deaths, and cumulative number for 14 days of COVID-19 cases per 100,000 population for 5 countries (USA, Brazil, India, Mexico, and UK) from Dec 31, 2020 to date

Fig 3 shows the case fatality rates, mortality rates, cases per million, and deaths per million for the top 10 most affected countries. The countries' case fatality rates and the mortality rates are computed over the period of the pandemic (from Dec 31, 2019, to date). An interesting point to note is that even though the case fatality rate to date is higher for Mexico, the UK, Italy, and France, these countries do not have the most fatalities. Several factors that affect the case fatality rates are the quality of medical care available in a country, guidelines in the country regarding who should get tested, precautions to prevent infection that are followed, availability of COVID-19 testing, etc. Also, in Fig 3, the mortality rates are the same for the USA, Brazil, Mexico, the UK, Italy, though the number of deaths is actually more in the USA and Brazil. Among the countries shown, Peru has the highest mortality rate, and India has the lowest, though India is the second-most populous country. A 0.01 mortality rate in India would result in more fatalities than in another country with the same mortality rate but with a smaller population, due to India’s very large population. Therefore, caution has to be exercised in interpreting these metrics.

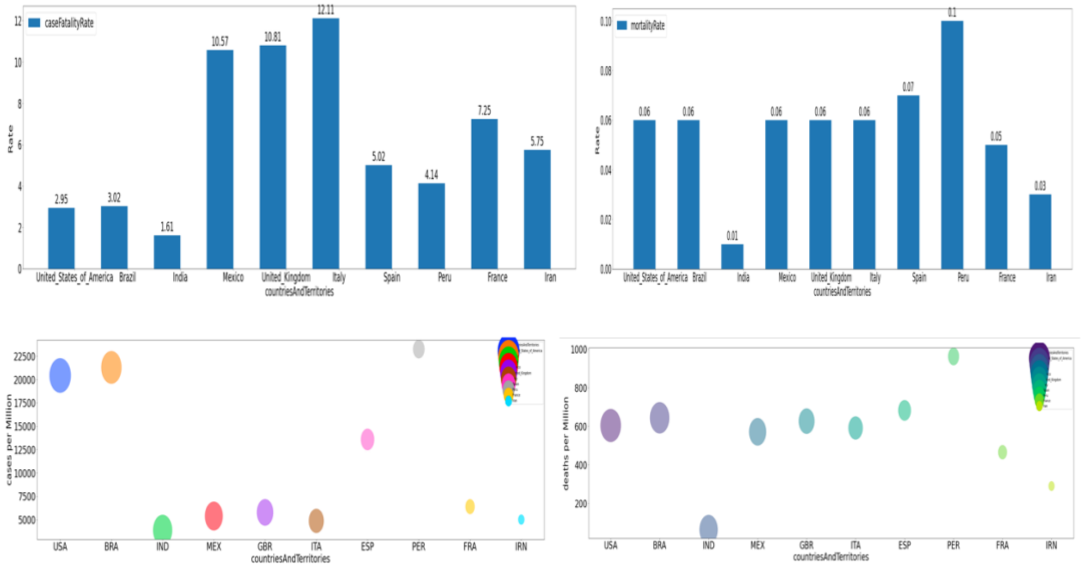
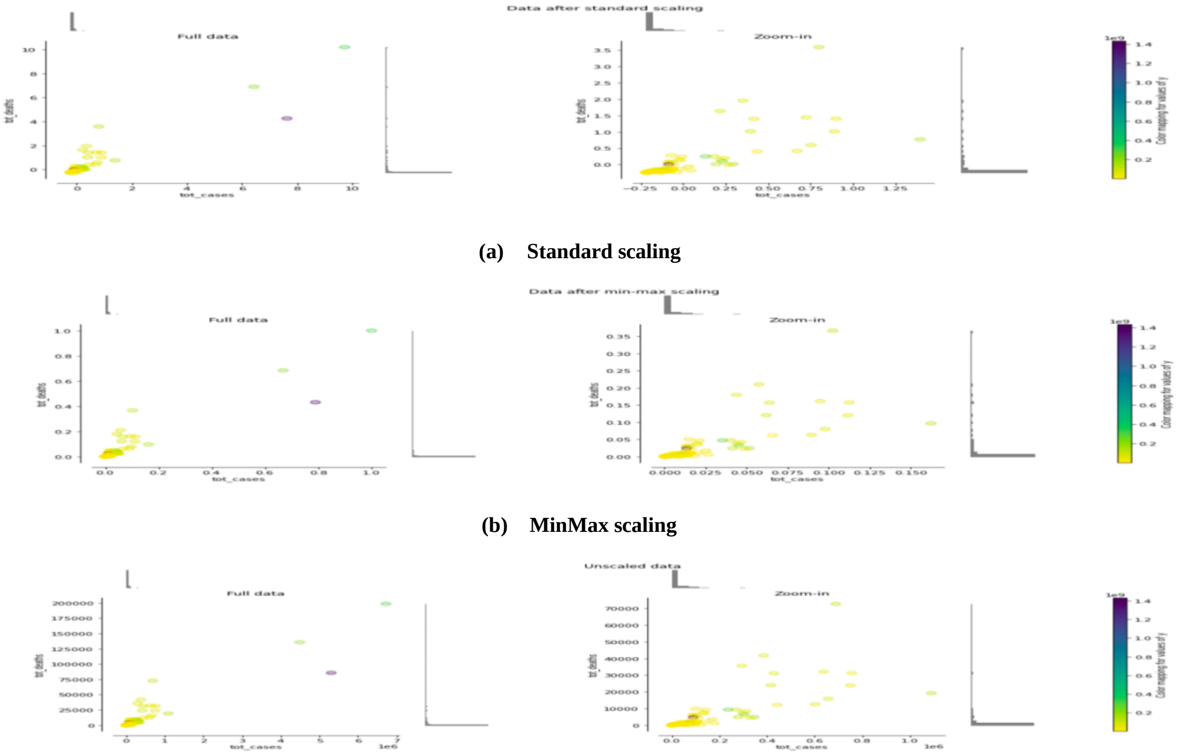


Fig 3: caseFatalityRate, mortalityRate, casesPerMil, deathsPerMil – for the top 10 most affected countries

In contrast, the cases per million, and the deaths per million seem to be more appropriate in reporting the COVID-19 situation in a country. The size of the bubbles indicates the magnitude of the cases/deaths relative to other countries.

5. Data scaling and comparisons

Our analysis noted that the number of cases and deaths (including their totals) for the various countries vary in magnitude or range. For any ML algorithm, this could pose an issue where higher values might be given higher weights, and smaller values might get lower weights, regardless of the unit of the values. To bring all the values to the same standing, we applied scaling. We leveraged two techniques, standard scaling and MinMax scaling. Standard scaling removes the mean and scales the data to unit variance. MinMax scaling rescales the variables to values in the range [0,1]. We selected two variables – total cases and total deaths – for all the countries in the data set and applied the two scaling methods on them. Fig 4 shows the distribution of the two variables after applying the scaling methods.



(a) Standard scaling

(b) MinMax scaling

(c) Unscaled data

Fig 4: Distribution plots on 2 variables – tot_cases and tot_deaths using a) standard scaling, b) MinMax scaling, c) Unscaled data (for comparison)

In Fig 4, the left figure shows a scatter plot of the full data set, while the right figure will exclude the extreme values (only 99 % of the data set is shown, excluding marginal outliers). Utilizing the unscaled plot, we can note the marginal distribution shows a positively skewed distribution for total number of cases and total number of deaths. Applying either the Standard or the MinMax scaler did not alter that. Most of the data show the points are between 0 and 0.2, with a few points between 0.2 and 0.4. This is true of all three graphs. Based on this comparison, we could apply either the Standard scaling or the MinMax scaling for our analysis.

6. Insights from analysis

The daily COVID-19 worldwide new cases and deaths data from ECOC can help answer the following questions:

a) What is the severity of COVID-19 infection in a country?

This question can be better answered using the cases per million people and deaths per million people metrics. Additionally, an increasing trend for these variables informs us of rising infection rates in a country.

b) Which countries are managing COVID-19 infection better?

In this case too, the cases per million people and the deaths per million people are better metrics to answer the question. Countries with a decreasing trend or even a stable trend might have implemented measures that are helping to combat disease transmission.

c) For what purposes can case fatality rate and mortality rate metrics be used?

The case fatality rate and mortality rates are computed over a period of time. These metrics should typically be computed at the end of the infection. At the start of any novel infection, these rates tend to be very high. But with experience in handling the infection (even without a preventive or a curative solution), the number of cases and the number of fatalities can be reduced. Since these metrics are heavily dependent on the policies and guidelines enforced in a country, testing availability, quality of health care, comparing these metrics among countries is not advisable. However, at the end of an infection, the case fatality rate and mortality rates from COVID-19 infection can be compared to other known diseases such as influenza, cardiovascular disease, etc., for mortality analysis within a country.

d) When is the second-wave of COVID-19 infection going to occur?

An examination and analysis of the increasing values of cases per million and deaths per million metrics can inform public health officials of an upcoming second wave of infection.

In summary, the data to date shows that the USA, Brazil, and India are the three most affected countries with COVID-19 infection. The USA has two peaks - a big peak in March and a smaller peak in July-August. The USA has a somewhat stable trend on deaths after March, though there have been some upward spikes in July and August. Brazil too has two peaks - the first peak in May was much smaller than the second peak in July. Brazil has increased deaths in May, but since then has stabilized deaths from infection. India, on the other hand, is showing an increasing trend in cases and in deaths. Countries, such as Italy, Spain, and France, though initially heavily impacted by the infection, have since stabilized/lowered new cases. However, without a safe and effective preventative solution, such as a COVID-19 vaccine, the data needs to be continuously monitored to identify additional waves of peak infections and deaths.