# Exploratory Data Analysis Of COVID-19 Dataset by Our World in Data

Mohammad Rouintan

Computer Science, Shahid Beheshti University

**Abstract**

*The outbreak of the novel coronavirus (COVID-19) poses an unprecedented global challenge, and a comprehensive understanding of its development is essential to effectively combat its impact. In this research, we present an exploratory data analysis of COVID-19 using a diverse range of data attributes including testing, vaccination, hospitalizations, and more. Drawing on the extensive datasets compiled by Our World in Data, we provide valuable insights into the progress of the epidemic and illuminate key trends and patterns observed over time. By combining research findings and statistical analysis, we aim to provide a timely and reliable assessment of the current state of the epidemic.*

## Introduction

Our task for this project is to perform exploratory data analysis to find important data behaviors and characteristics. In carrying out this project, we have used data science libraries such as NumPy and Pandas for exploratory data analysis, as well as Matplotlib and Seaborn for data visualization.

This dataset contains approximately 350,000 samples collected from different days in different countries of the world, each of which has the same features (67 features) that can help us find different reasons for the number of new cases each day, the number of deaths and The number of deaths per day, the number of people who have been vaccinated, etc. helps.

Through our comprehensive data analysis, we are uncovering important insights into the COVID-19 pandemic, enabling policymakers, researchers and public health officials to make informed decisions and devise effective strategies. The availability of reliable and timely data plays an essential role in combating this outbreak, and our work in Our World in Data strives to facilitate a better understanding of the progress of the pandemic.

As the world continues to grapple with the challenges posed by COVID-19, we emphasize the importance of ongoing data collection and analysis and the need for a collaborative, evidence-based approach. Using the power of data-driven insights, we can work collectively to reduce the impact of the pandemic and protect global health.

## Exploratory Data Analysis (EDA)

### Data Claening and Handling Missing Values

There are a lot of missing values in this dataset, so there are a lot of features, more than half of them are missing, which causes a lot of problems. At first, we dropped features that were more than 80% missing, and this caused our dataset to decrease from 67 features to 52 features. However, there are still a lot of missing data values, but for various reasons, we have left them unchanged for this project, which we will explain below.

Our data is time series data, so that each row of information represents one day (as we can see, the date of each day is included as a feature), and if we fill the missing data in the usual ways, with a high probability is that the information we receive from the data is wrong (for example, if we fill in the missing values of the New Death feature and then try to show the trend of death by date, we may see a trend that is far from reality, so reporting these results will not be useful). So with these interpretations, in order to properly fill such missing values, we need more complex neural networks and deep learning models in the field of time series, so that the trend of features in this way can be close to reality (for example We used iterative imputer from the Scikit-Learn library to populate a series of features, and what happened was that the numbers inserted for null values were outside the original domain defined for that feature. Finally, we left these unchanged so that we can discover the correct information from the same amount of data we have. In addition, for the outliers, the data contained a series of features that had many outliers, for example (New Cases, New Deaths, New Tests), which we have removed using

the percentile method with a ratio of (0.0 and 0.98).

## Analysis

In the Analysis section of this data, we review and analyze a series of numerical and categorical data. We will examine this using different types of plots.

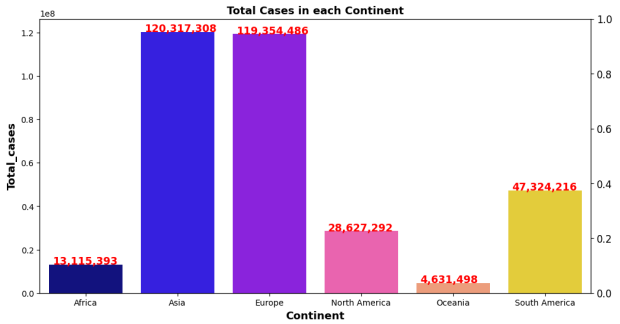


**Figure 1:** *This BarPlot shows the total number of cases in each continent, and as you can see, Asia, Europe, and South America are ranked first to third, which seems correct considering the size of the continent.*

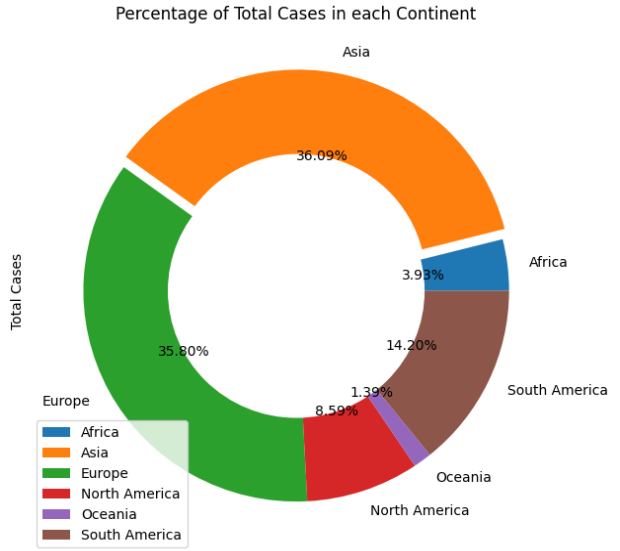


**Figure 2:** *This DonutPlot shows the percentage of total number of cases in each continent, and as you can see, Asia, Europe, and South America are ranked first to third, which seems correct considering the size of the continent.*

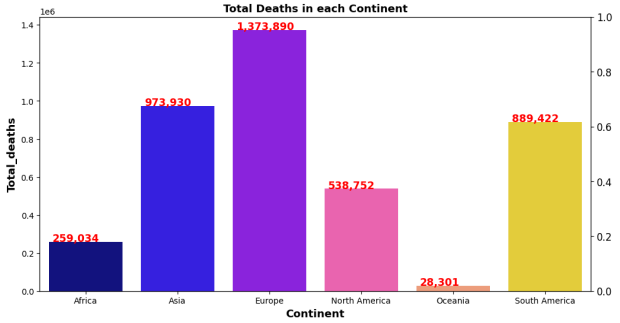


**Figure 3:** *This BarPlot shows the total number of deaths in each continent, and as you can see, Europe, Asia, and South America are ranked first to third.*

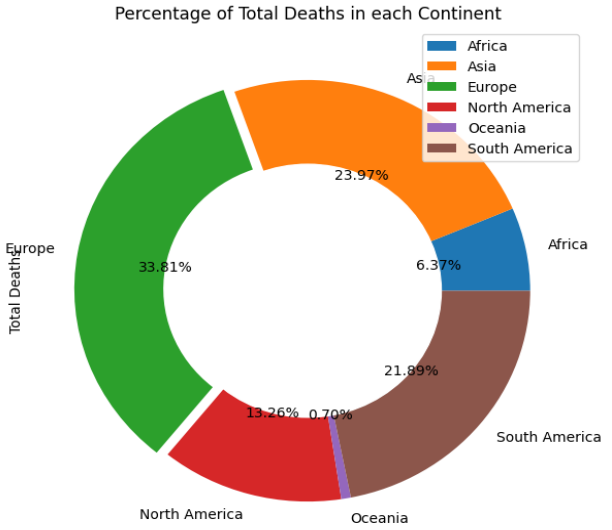


**Figure 4:** *This DonutPlot shows the percentage of total number of deaths in each continent, and as you can see, Europe, Asia, and South America are ranked first to third.*

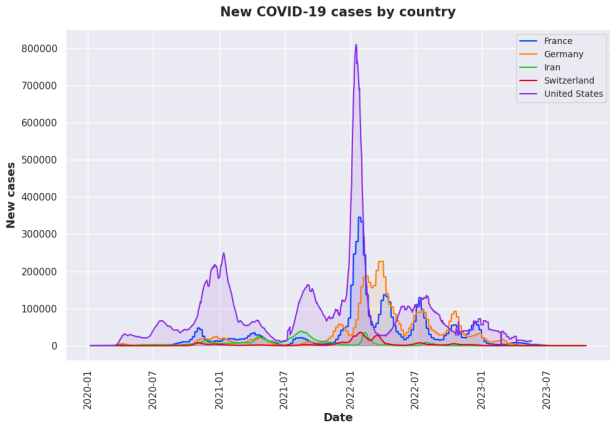


**Figure 5:** *This LinePlot shows the trend of new cases from 2020 to 2023 for 5 countries (Iran, United States, France, Germany, Switzerland). According to this graph, we can understand in which period of time the number of cases reached the peak in each of these 5 countries or in which period of time the trend was normal. As it is clear from the graph, the best trend belongs to Switzerland.*

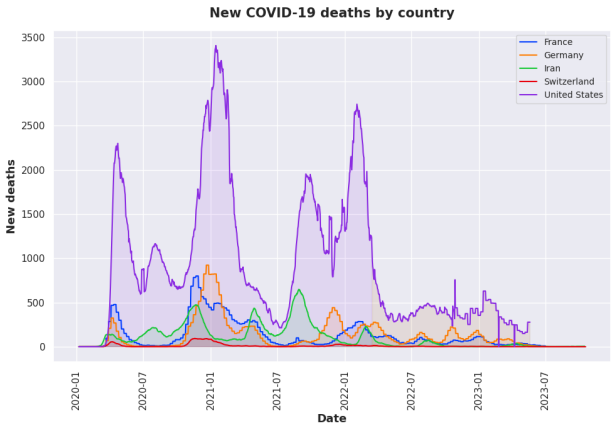


**Figure 6:** *This LinePlot shows the trend of new deaths from 2020 to 2023 for 5 countries (Iran, United States, France, Germany, Switzerland). According to this graph, we can understand in which period of time the number of deaths reached the peak in each of these 5 countries or in which period of time the trend was normal. As it is clear from the graph, the best trend belongs to Switzerland.*

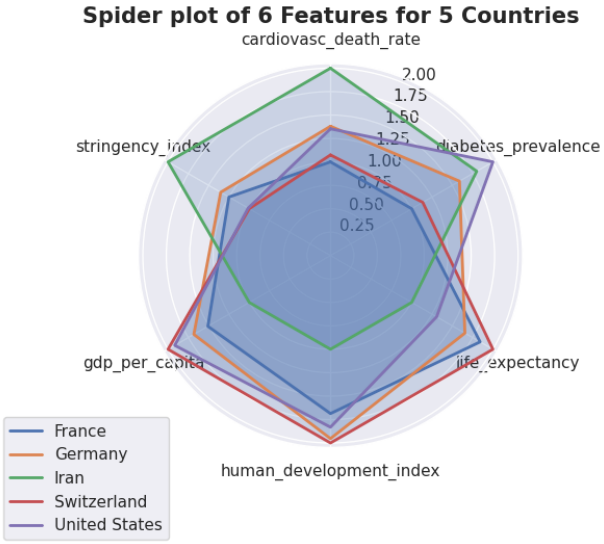


**Figure 7:** *First of all, in order to be able to draw the spider plot, we need to scale the features we want to use using appropriate scaling. Here we've used min-max scaling and then plus-one each value (ie we've actually shifted it by one unit) because if a value is zero it sticks to the center of the graph and doesn't display anything . This spider plot shows the intensity and weakness of each of the mentioned features in 5 different countries (Iran, United States, France, Germany, Switzerland). For example, the prevalence of diabetes is high in the United States, or the rate of cardiovascular deaths is high in Iran. In general, this chart wants to compare 5 countries according to these features.*

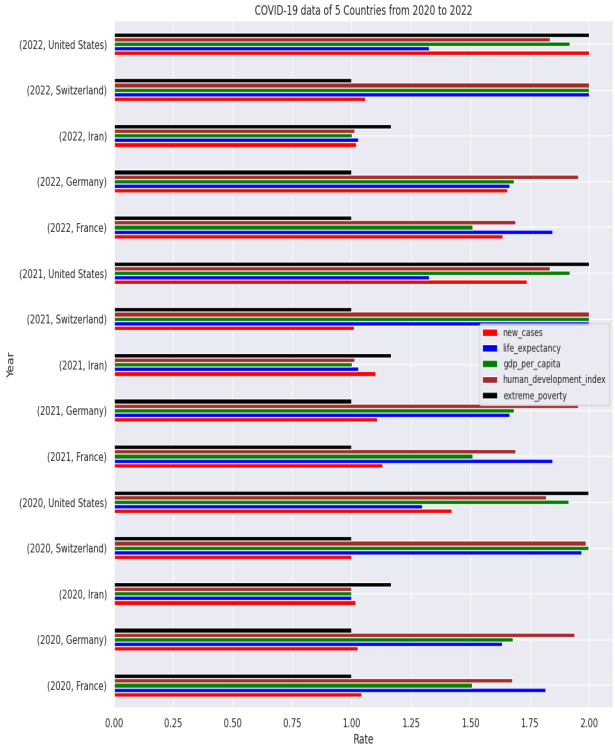


**Figure 8:** *This Barhplot shows the rate of each of the mentioned features in different countries from 2020 to 2022. For example, in France in 2021, life expectancy is high, extreme poverty is low, human development is high, GDP is high, and New Cases is in a relatively normal state.*

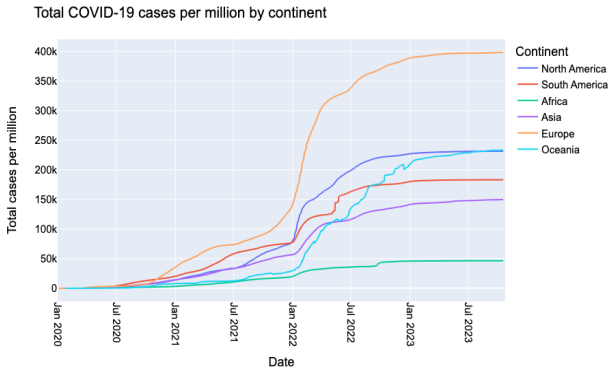


**Figure 9:** *This LinePlot shows the trend of total cases in different continents. As you can see, this trend has made the most progress between January 2022 and July 2022.*

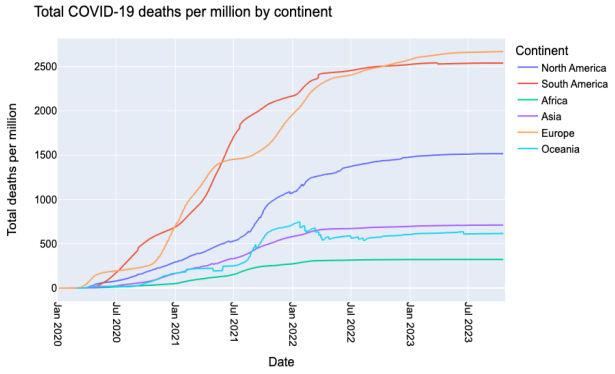


**Figure 10:** *This LinePlot shows the trend of total deaths in different continents. As you can see, this trend has made the most progress between January 2021 and July 2022.*

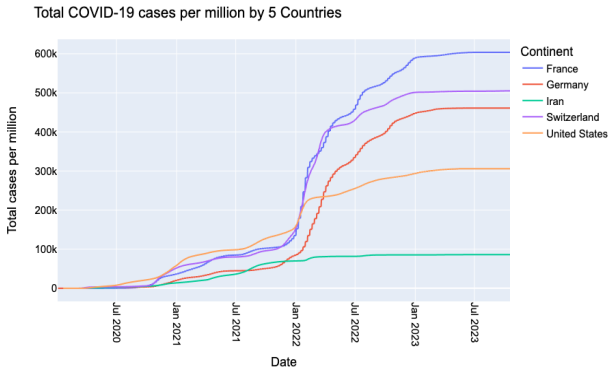


**Figure 11:** *This LinePlot shows the trend of total cases in different 5 countries. As you can see, this trend has made the most progress between January 2021 and January 2023.*

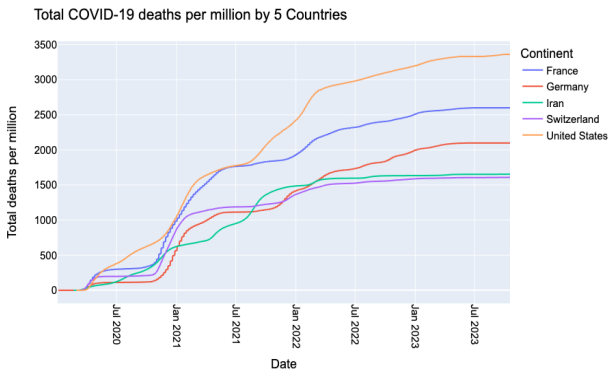


**Figure 12:** *This LinePlot shows the trend of total deaths in different 5 countries. As you can see, this trend has made the most progress between January 2021 and July 2022.*

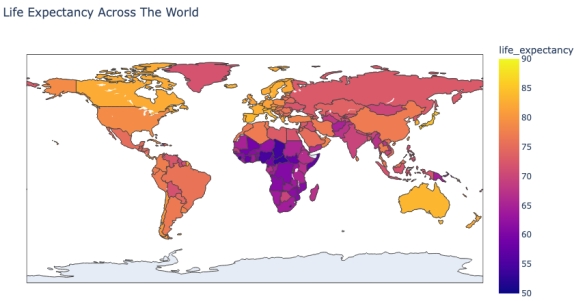


**Figure 13:** *This plot shows the life expectancy in different countries on the world map. You can understand the intensity and weakness of life expectancy in different countries according to the map guide.*

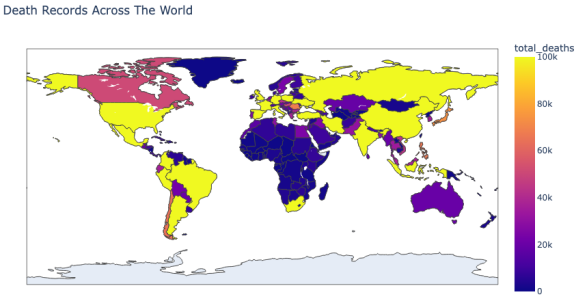


**Figure 14:** *This plot shows the total deaths in different countries on the world map.*

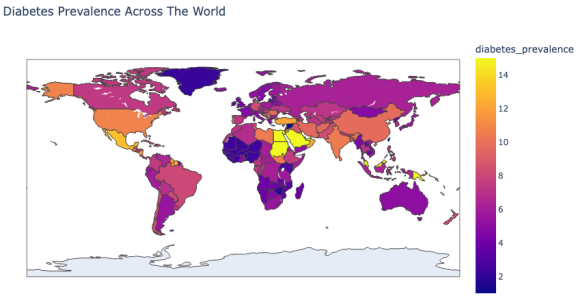


**Figure 15:** *This plot shows the diabetes prevalence in different countries on the world map.*

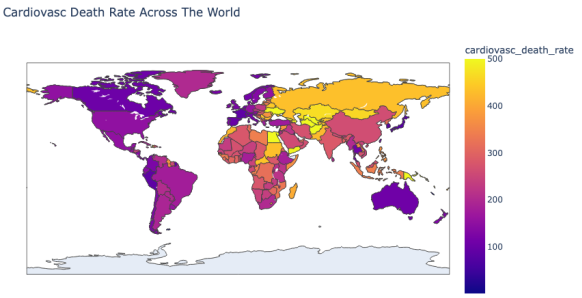


**Figure 16:** *This plot shows the cardiovasc death in different countries on the world map.*

# Results

The main goal of this project was to perform an exploratory data analysis on the Covid-19 data, from which we were able to obtain information to some extent, but due to the problems in the data (missing values are extremely high), we need stronger work to find strong information from the data. For example, we can get help from machine learning and deep learning models to better understand the data, because the problems of data cannot be solved with simple methods.