COMP4651 Cloud Computing and Big Data Systems: Project Report

Analysis of OWID COVID Data

By Yeung Wai Kit (20608456)

Abstract

The global outbreak of COVID-19 pandemic has lasted for one and a half years and an increasing number of COVID-19-related data are generated and aggregated every day. The analysis of this massive amount of daily updating data requires cloud computing and big data systems. In this project, I will leverage existing cloud computing and big data technology to analyze COVID-19 related data. The goal of this project is to (1) analyze the COVID data to discover trends of the pandemic and (2) discover the relationship between different data in the COVID-19 dataset.

Introduction

The dataset I use is OWID COVID-19 data. It is a dataset maintained by Our World in Data and is updated daily (at the time the project is done the date is 30/04/2021 so the dataset is last updated at 30/04/2021). The dataset includes data on confirmed cases, deaths, hospitalizations, testing, vaccinations, and some demographic data. The dataset gives a macroscopic view on the COVID-19 pandemic and the data structure allows global comparison and time-series analysis. Thus, I choose to use this dataset to analyze COVID-19 trends. The github repository of the data is https://github.com/owid/covid-19-data/tree/master/public/data/ .

The analysis is conducted via Databricks community edition. As the size of data is not super large, I did not use amazon s3 as storage. Databricks is a cloud-based big data platform where users can access to micro-cluster, cluster manager and a python notebook environment. As the data is structured, I mainly use pyspark dataframe module to perform the data analysis for this project.

This report will describe the implementation of the project in detail. In this project, there are several analyses and they all require different preprocessing and implementation methodologies. Thus, the report session is divided according to the type of analysis.

Data Structure

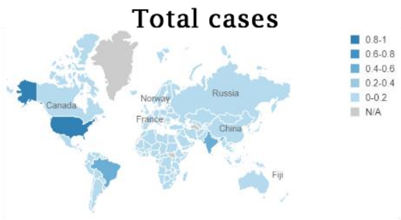
In the first section I explored the structure of the dataset. As mentioned, the data is structured and can be directly loaded as a dataframe. The shape of the dataframe is (85778,59). Among the 59 columns, 3 are categorical data, 1 is timestamp data and the rest are all numerical data. There are 219 unique countries and 483 unique dates in the dataset.

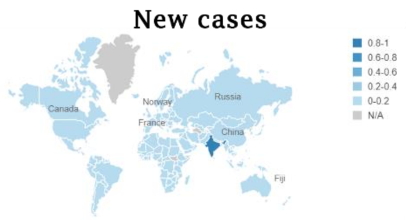
Global Comparison

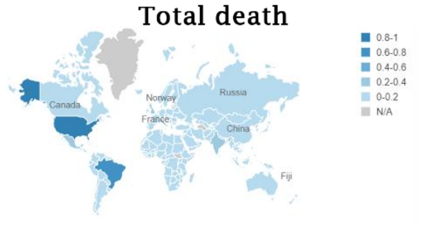
This section aims to find global trends of the dataset at the time the report is done (30/04/2021).

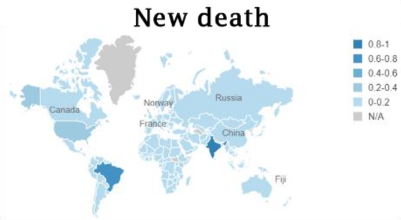
The preprocessing of data includes filtering and normalization. As I want to analyze the most recent data, the data entries with latest timestamp are kept. The normalization is based on min-max normalization algorithm.

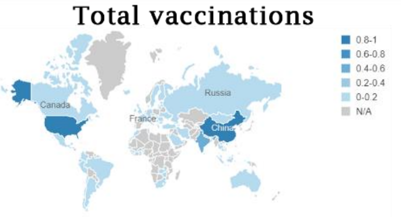
After preprocessing, I plot the world map graph using databrick's built-in display function and gathered the following results.

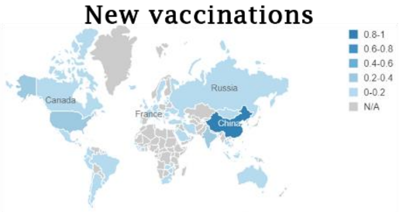

Total cases

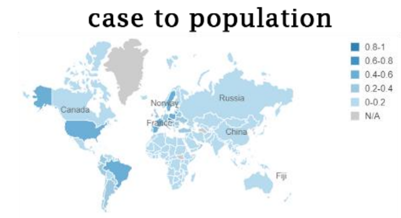

New cases

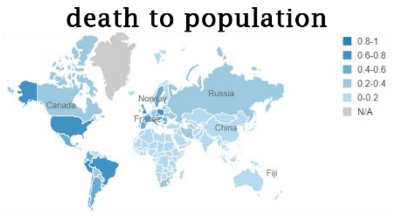

Total death

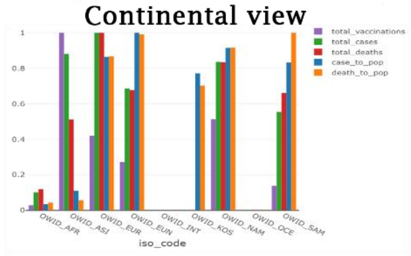

New death


Total vaccinations


New vaccinations


case to population


death to population


Continental view

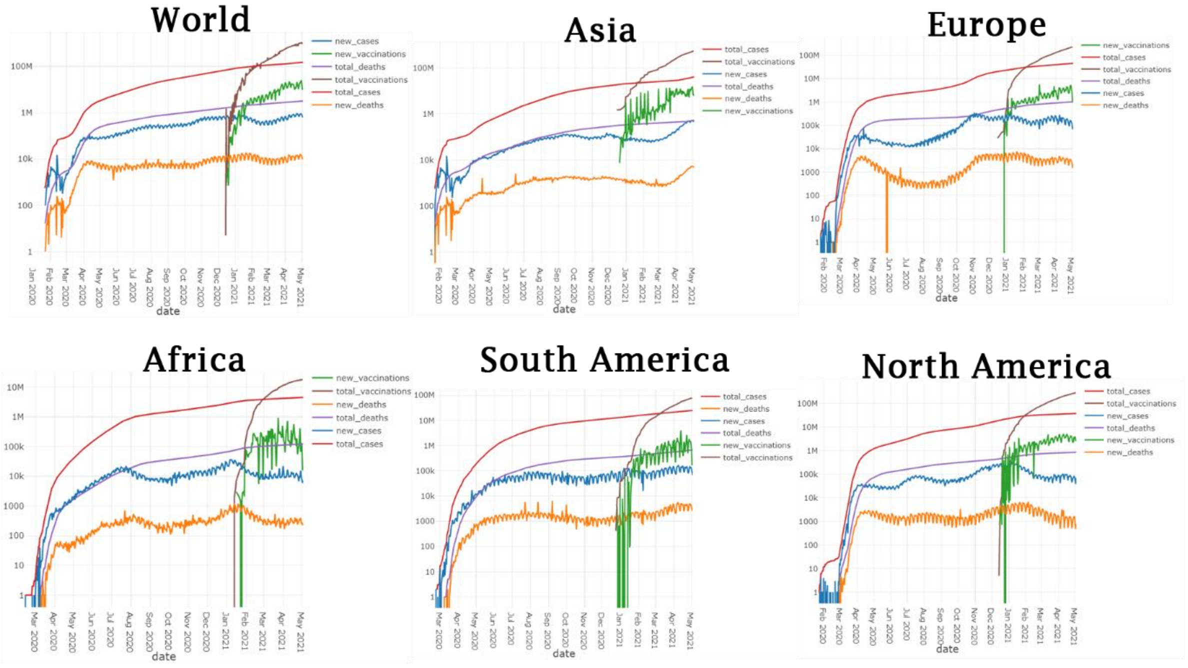
From the above graphs, we can draw the following conclusions.

(1) COVID-19 does a significant damage to USA and Brazil. They have the most confirmed cases and deaths numerically. When compared to their population, the confirm cases and deaths still own a large portion comparing to the rest of the world.
(2) The recent soar in confirmed cases and deaths in India made India the third most affected country in COVID-19 pandemic.
(3) The three most vaccinated countries are USA, China and India. Among them USA and India are significantly affected by the pandemic. Thus, the effect of vaccination against COVID-19 may not be significant.
(4) Comparing the figures intercontinentally, Europe is severely affected by COVID-19. It has the highest number of confirmed cases and deaths. Meanwhile, countries in the European Union have the highest confirmed cases to population ratio and deaths to population ratio. Thus, countries in Western Europe suffer the most in this pandemic.

Time-Series Analysis

This section aims to find trends of the pandemic across time. The preprocessing only requires filtering of data based on locations.

After preprocessing, the results are plotted using databrick's built-in line graph function. The data on the graph are log scaled to enable comparison. The graphs are shown below.



World

Asia

Europe

Africa

South America

North America

There are several patterns concerning the above graph.

Across the world, COVID-19 has similar patterns where the number of total cases, new cases, total deaths, and new deaths are all increasing exponentially since the start of the pandemic. Since December 2020, vaccines are massively produced and effectively distributed across countries and the number of vaccinations skyrocketed starting from December 2020.

Globally speaking, the log-scaled graph of total cases and deaths start to flatten recently. The lines indicating new cases and new deaths also dropped recently. Thus, the pandemic is more stably controlled recently.

In the Asia graph, the log-scaled graph of total cases and deaths soared then started to flatten from February 2020 to March 2021. This pattern matches with the worldwide pattern of the pandemic. Yet, starting from March 2021, the number of confirmed cases and deaths started to rise again. A reason for this is the recent outbreak in India.

In the Europe graph, we can see the number of cases and deaths soared at the beginning of the global outbreak. The graph of total cases and deaths were flattened and the number of new cases and deaths showed a decrease from April 2020 to November 2020. An explanation for this is during this period, countries in Europe implemented strict social restriction measures.

From December 2020 to recent, these countries loosened the restriction measures and this results in a significant increase in cases and deaths.

The Africa graph has similar patterns comparing to the globe but the number on the y-axis is much lower than that of other continents. This may indicate Africa is less affected by COVID-19 comparing to other continents. The total number of cases and deaths starts to flatten, and the number of new cases and deaths have started to drop since February 2021.

The South America graph shows a continuous increase in number of cases and deaths since the start of the pandemic.

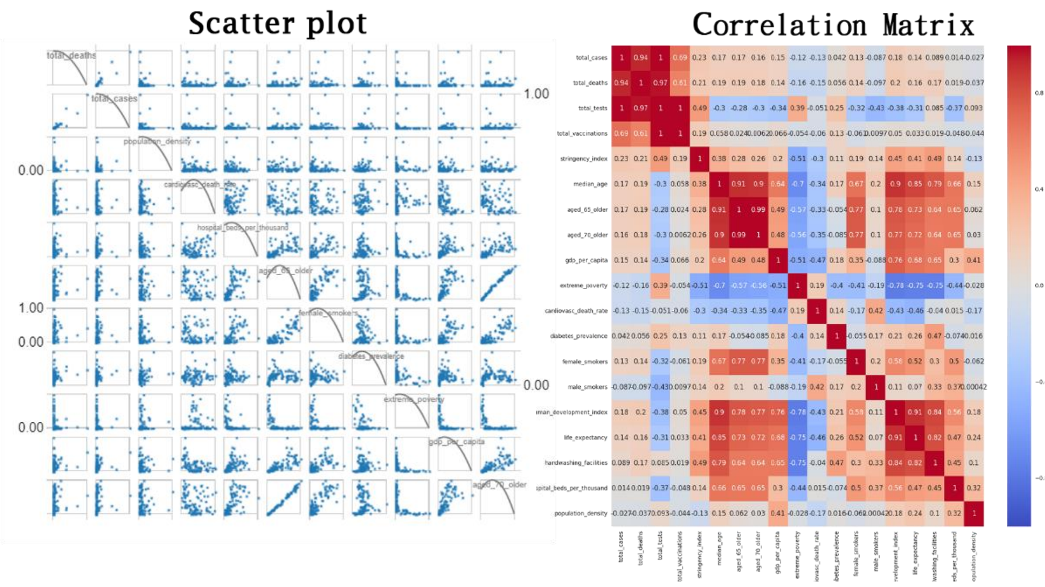
The North America graph shows a similar pattern comparing to the rest of the world from March 2020 to January 2021. The total number of cases and deaths starts to flatten, and the number of new cases and deaths have started to drop since January 2021.

Among the Five main continents, Europe, Africa and North America show a drop in the number of new cases and deaths since start of vaccinations. Yet, in Asia and South America, increase in number of vaccinations does not reduce the number of new cases and deaths. Several factors may lead to this result including (1) difference in types of vaccines, (2) Massive outbreak variant virus of in India, and (3) stringency of social restriction measures.

Correlation analysis

This section aims to discover the correlation between demographic data and the pandemic.

The preprocessing of data includes filtering and normalization. I filtered the data entries with latest timestamp to do the correlation analysis. The normalization is based on min-max normalization algorithm. The findings are shown below.



From the above graph, we can discover some interesting findings.

First, COVID-19 pandemic does greater damage in more developed countries. The correlation matrix shows the number of cases and deaths are slightly positively correlated to human development index, life expectancy, and stringency index. Meanwhile the number of cases and deaths are slightly negatively correlated to extreme poverty. Combining with the global analysis above, we can conclude that more developed countries suffer more severely in this pandemic.

Second, better environment and hygiene facilities does not lead to better prevention of the pandemic. In the correlation matrix, number of hospital beds, number of handwashing facilities, and population density are all uncorrelated to the total number of cases and deaths.

Conclusion

The above analyses show the trends of the pandemic around the world. The main insights are (1) Even after one and a half years, the pandemic figures are still growing constantly, (2) The outbreak of the pandemic is not related to demographic data, and (3) the effect of vaccines are limited for now. To conclude, there is still a long way to the end of this pandemic. Despite vaccinations and the increase of knowledge about the virus, we should maintain good hygiene awareness and obey the social restriction measures to fight against the pandemic.

Links

Github link of this project:

https://github.com/hkust-comp4651-21S/project-comp_4651_owid_data_analysis