2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT) | 978-1-6654-4137-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISCIPT53667.2021.00107

Exploratory Data Analysis on the Usage of COVID-19 Vaccine

Liangqi Chen
School of Information Science and Engineering
Hangzhou Normal University
Hangzhou, China
813806051@qq.com

Ben Wang*
School of Information Science and Engineering
Hangzhou Normal University
Hangzhou, China
* Corresponding author: 20170056@hznu.edu.cn

Yinggui Wang
School of Information Science and Engineering
Hangzhou Normal University
Hangzhou, China
227659811@qq.com

Xiya Wang
School of Information Science and Engineering
Hangzhou Normal University
Hangzhou, China
1647101212@qq.com

***Abstract*—The COVID-19 epidemic has swept the world for more than a year. Besides the efforts of the medical staff to fight the outbreak, a number of researchers at home and abroad have conducted a visual analysis of the data of COVID-19 epidemic. They have made contributions to the fight against the epidemic mainly in two aspects: situation display and epidemic prediction. With the advance of research, more and more countries have developed effective vaccines. In this paper, we conducted exploratory data analysis on the existing data of COVID-19 vaccine. It reveals the types and quantities of vaccines currently in use, shows the comparison of vaccination data from different countries and global vaccination trends, made the comparison between China and India, and has a more in-depth and clear understanding of the world's fight against the epidemic situation.**

Keywords- COVID-19; vaccine; exploratory data analysis; data visualization

I. INTRODUCTION

It has been more than a year since the COVID-19 outbreaks in late 2019. In history, data visualization analysis of major events can achieve direct expression, exchange of ideas, and enhance persuasion. And it is also helpful to mine the hidden, deep-seated rules and trends in the event. For instance, John Snow's Cholera Map of London, Charles Minard's Russo-French War (Napoleonic Expeditions Map), etc. [1].

A lot of work has been done on the visualization of COVID-19 epidemic data, mainly focusing on epidemic situation display and epidemic prediction. Zhao Shanlu et al. analyzed 195 COVID-19 cluster cases in Hunan province of China through spatiotemporal visualization, and summarized the transmission mode of the virus. Li Chechen et al. analyzed the basic information, gender and age distribution, patient behavior, and epidemiological characteristics of confirmed COVID-19 cases in the high-prevalence area of Henan Province, as well as the local prevention and control measures through visualization of statistical charts, so as to provide reference for further epidemic prevention and control work [1]. Pronoy Roy et al. explained how data visualization can help us minimize the damage caused by COVID-19 [2].

Fortunately, the world's top scientists are working on the vaccine. As research advances, more and more countries have developed effective vaccines, while few studies have been conducted on the use of the vaccine. Therefore, in order to show the quantity and usage of the existing vaccines more intuitively, this paper decided to use exploratory data analysis method to analyze the data of current vaccine. For exploratory data analysis, Rahul Pradhan et al. used it to analyze and display various factors of the Olympic Games evolving over time in the form of graphs, and conduct comparative research among various factors [3].

II. EXPLORATORY DATA ANALYSIS

A. Definition

Exploratory data analysis (EDA) is to analyze data in a specific way without definite problems. In the process of data analysis, it finds new problems and solves the problem of circular exploration process [4]. The main work is to clean and describe data, visualize the distribution, compare relationship between them, cultivate the intuition from data, summarize the results, and so on.

B. Comparison with traditional data analysis methods

Based on the difference of steps, data analysis can be divided into descriptive data analysis, exploratory data analysis, and confirmatory data analysis. Descriptive data analysis should be the simplest kind of data analysis. It adopts methods such as calculating statistical values and drawing charts to find the rules of data surface. After the emergence of exploratory data analysis, the process of data analysis is divided into two stages, the exploration stage and the verification stage. In the exploration stage, we focus on discovering patterns or patterns hidden behind the data; In the validation stage, we focus on verifying whether the new model from the data exploration stage is correct. Exploratory data analysis can help us investigate extra features hidden behind the data, so as to discover better effective model [5].

978-1-6654-4137-7/21/$31.00 ©2021 IEEE
DOI 10.1109/ISCIPT53667.2021.00107

III. EDA OF COVID-19 VACCINE

A. Data acquisition

Kaggle is a popular data science competition platform that offers a wide variety of data sets. This paper used the COVID-19 Vaccine Progress data set provided by Kaggle, and use Python to read CSV files. The data, 11175 records on April 14, 2021, have been obtained. Some of its data fields are shown in TABLE I.

TABLE I. PARTIAL DATA FIELDS OF THE DATASET

Field Name	Field Description
Country	Country name
Iso_code	ISO Code for country
Date	Calendar date
Total_vaccinations	Total vaccinations per date and country
People_vaccinated	Number of people vaccinated
People_fully_vaccinated	Number of people fully vaccinated
Daily_vaccinations	Daily vaccination
People_fully_vaccinated_per_hundred	People fully vaccinated percent
Vaccines	Vaccines scheme (the combination of vaccines used by a country)

B. Data cleaning

Data cleaning is the process of re-detecting data files, identifying errors and standardizing operations, such as deleting duplicate data, processing invalid values and missing values [6]. Data cleaning is the starting point of data research and one of the important means to improve data quality. The purpose of data cleaning is to conduct standardized inspection on the massive data with high redundancy, and to mine out the key data information necessary for subsequent data analysis and visualization [7].

Figure 1 and Figure 2 show that this data set contains a large number of missing values NaN, which have an impact on subsequent data analysis. To solve this problem, this article uses Python to fill in the missing values with 0.

	country	iso_code	date	total_vaccinations	people_vaccinated
0	Afghanistan	AFG	2021-02-22	0.0	0.0
1	Afghanistan	AFG	2021-02-23	NaN	NaN
2	Afghanistan	AFG	2021-02-24	NaN	NaN
3	Afghanistan	AFG	2021-02-25	NaN	NaN
4	Afghanistan	AFG	2021-02-26	NaN	NaN

Figure 1. The dataset contains missing NaN values.

```
data_df.isna().sum()
```

country	0
iso_code	0
date	0
total_vaccinations	4510
people_vaccinated	5169
people_fully_vaccinated	6872
daily_vaccinations_raw	5590
daily_vaccinations	196

Figure 2. Statistics of the number of missing values in the dataset.

C. Analysis of Vaccine Combination

1) Type and quantity of vaccine

Effective vaccines have been developed in various countries. In the provided dataset, ten vaccines have been put into production, used and inoculated in the world. As a result, there are 29 vaccine combinations from these 10 vaccines. The most popular inoculations of vaccine combinations in the world at present are described in Figure 3 and Figure 4. Among them, Johnson & Johnson, Moderna, Pfizer / BioNTech combination had the largest number of vaccinations, reaching 175 million times.

	total_vaccinations
Johnson&Johnson, Moderna, Pfizer/BioNTech	175773904.0
Sinopharm/Beijing, Sinopharm/Wuhan, Sinovac	154187928.0
Covaxin, Oxford/AstraZeneca	94605157.0
Moderna, Oxford/AstraZeneca, Pfizer/BioNTech	90096893.0
Oxford/AstraZeneca, Pfizer/BioNTech	47545213.0

Figure 3. Total vaccinations of vaccine combination.

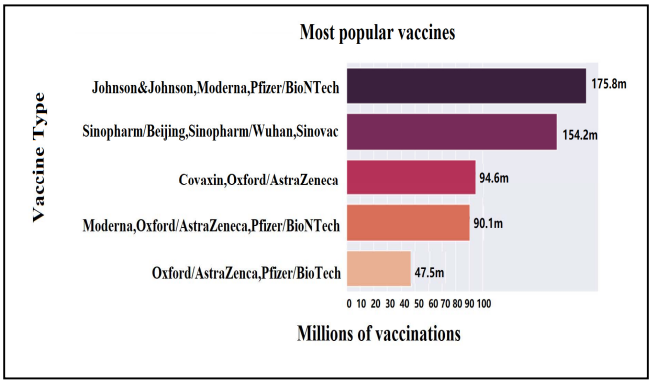

Figure 4. Histogram of total vaccinations of vaccine combination.

Figure 5 is a map of the world that more intuitively shows the geographical location of global vaccine usage. Different colors represent different vaccine combinations. As can be seen from the figure, Europe mostly uses Pfizer/ BioNTech vaccines.

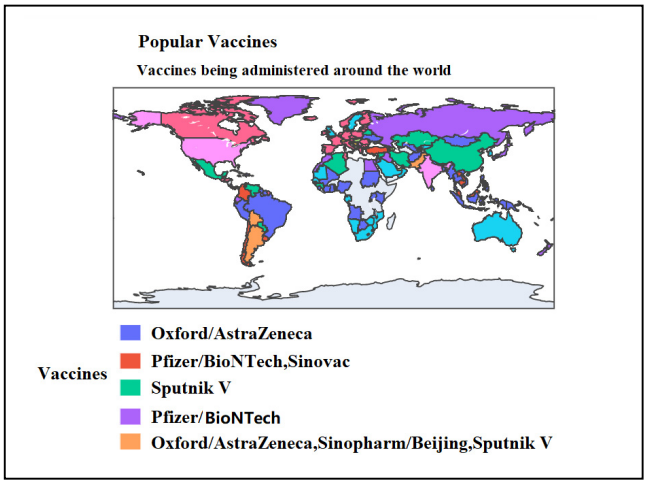

Figure 5. Map of global vaccine usage.

2) Comparison of vaccination data from different countries

Table II shows the data of the top 20 countries in total vaccinations. The United States ranks first, with a total of 187,047,131 times and 119,242,902 people. The second is China, with a total number of 167,343,000 inoculations. However, the people vaccinated number, not provided by the dataset, has been replaced with 0 in the data cleaning stage. India ranks third, with more than 100 million vaccinations and more than 91 million people vaccinated. In order to more intuitively show the comparison of the total number of vaccinations in different countries, this paper uses histogram and line chart as visual tools to show.

TABLE II. PARTIAL DATA OF THE TOP 20 COUNTRIES RANKED BY TOTAL NUMBER OF VACCINATIONS

Country	*Total vaccinations*	*People vaccinated*	*Date*
United States	187047131.0	119242902.0	2021-04-11
China	167343000.0	0.0	2021-04-11
India	104528565.0	91587400.0	2021-04-11
United Kingdom	39587893.0	32121353.0	2021-04-10
Brazil	26741261.0	20654434.0	2021-04-11
Turkey	18494796.0	10907432.0	2021-04-11
Germany	18231747.0	13196552.0	2021-04-11
Indonesia	15081949.0	10002901.0	2021-04-10
France	14444958.0	10757542.0	2021-04-10
Russia	14108341.0	8692848.0	2021-04-11
Italy	13032996.0	9108332.0	2021-04-11
Chile	12031595.0	7369321.0	2021-04-10
Mexico	11395137.0	9325316.0	2021-04-11
Israel	10256698.0	5320075.0	2021-04-11
Spain	10231825.0	7159716.0	2021-04-09
United Arab Emirates	9005444.0	3480415.0	2021-04-11
Morocco	8606571.0	4471831.0	2021-04-10
Canada	7996122.0	7198857.0	2021-04-11
Poland	7687617.0	5581068.0	2021-04-11
Saudi Arabia	6281357.0	0.0	2021-04-11

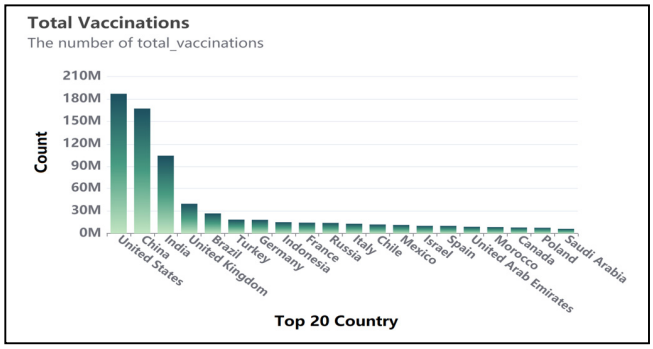

Figure 6. Top 20 countries in total number of vaccinations.

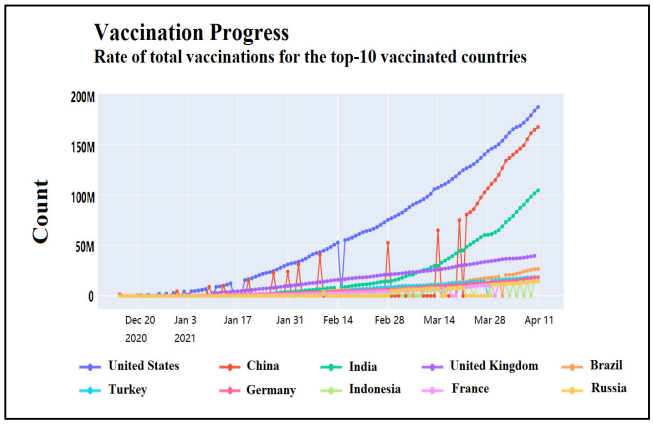

Figure 7. Trends in total number of vaccinations by country.

According to Figure 6, there is a large gap in the total number of vaccinations in various countries around the world. The United States, China and India are in the lead. In Figure 7, the total number of vaccinations in the United States has been in the leading position almost all the time. In China, the vaccination rate has increased gradually in the past month. India ranked fourth before March 2021, and then the vaccination rate also increased, but its acceleration is still less than that of China. According to the picture as a result, we can simply predict that Chinese total number of vaccinations will overtake the United States in the near future.

How about ranking people fully vaccinated rates? Fully vaccinated people, refers to the people who are vaccinated fully dose of the vaccine, and in China's case, two doses count as a full vaccination. Table III and Figure 8 reveal this result.

TABLE III. PARTIAL DATA OF THE TOP 20 COUNTRIES RANKED BY PEOPLE FULLY VACCINATED PER HUNDRED

Country	*People fully vaccinated per hundred*	*Date*
Gibraltar	88.13	2021-04-10
Israel	57.03	2021-04-11
Seychelles	40.13	2021-04-06
Bermuda	27.88	2021-04-04
Chile	24.39	2021-04-10
Bahrain	23.00	2021-04-11
San Marino	22.69	2021-04-10
Jersey	22.54	2021-04-04
Monaco	22.41	2021-04-02
United Arab Emirates	22.12	2021-04-11
United States	21.72	2021-04-11
Serbia	17.14	2021-04-10
Malta	15.70	2021-04-10
Hungary	12.60	2021-04-11
Northern Cyprus	11.53	2021-03-25
Morocco	11.20	2021-04-10
United Kingdom	11.00	2021-04-10
Singapore	9.16	2021-04-06
Turkey	9.00	2021-04-11
Iceland	8.15	2021-04-09

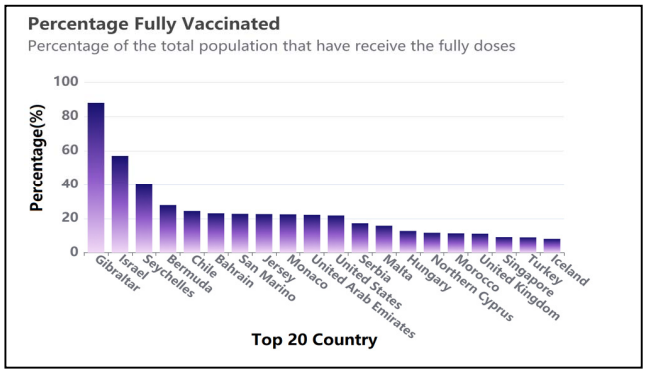

Figure 8. Top 20 countries in fully vaccinated percentage.

Among the countries and regions in the world, Gibraltar ranks first, with a full vaccination rate of 88.13% of its total population. That means nearly 90 percent of the country's population has received all doses of COVID-19 vaccine. Further analysis of the data set revealed that the vaccine used in Gibraltar was Pfizer/ BioNTech. As of 10 April 2021, its total number of vaccinations is 63,671, the number of people who have received at least one dose was 33979, and fully vaccinated is 29692. According to the data, Gibraltar has total population of 34733 in 2020. After calculation, the vaccination rate of at least one dose will reach about 98%.

The vaccination rate also depends on the population base. Of the 10 countries with the highest first-dose vaccination rates, only the United States, the United Kingdom and Chile have populations of more than 10 million. At the current rate of vaccine production and vaccination, most of the countries in the world that can achieve high vaccination rates are small countries with a population of one million.

3) Global vaccination trends

With the continuous marketing and vaccination of the COVID-19 vaccine, the global epidemic has indeed shown a certain turnaround. Countries with high vaccination rates, such as the United States and the United Kingdom, have seen significant improvements in epidemic. The central variable in whether and when the global epidemic can be effectively controlled remains vaccination. Currently, the number of people who have received at least one dose of the vaccine has reached 327,040,631, and the number of people who have been fully vaccinated is 130,613,907. As of January 2021, the world's population is estimated at 7.585 billion, which means that only 1.7% of the world's population has currently completed the entire vaccination phase.

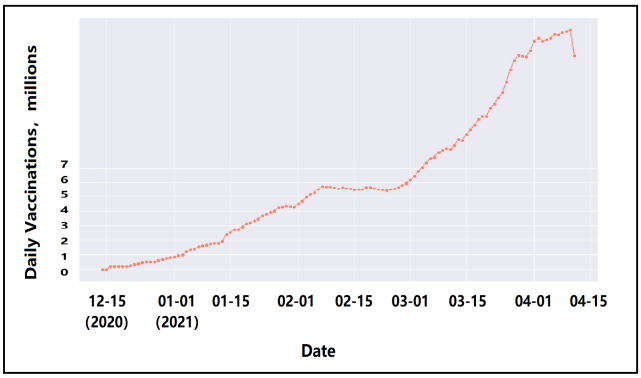

Figure 9. Global trend of daily vaccinations.

Figure 9 shows the global trend of daily vaccinations (million) of COVID-19 vaccine. In terms of daily vaccinations, the global average is now about 6 million doses a day, it is expected that the global daily vaccinations of COVID-19 vaccine will continue to increase as the number of licensed vaccines increases and vaccine capacity expands.

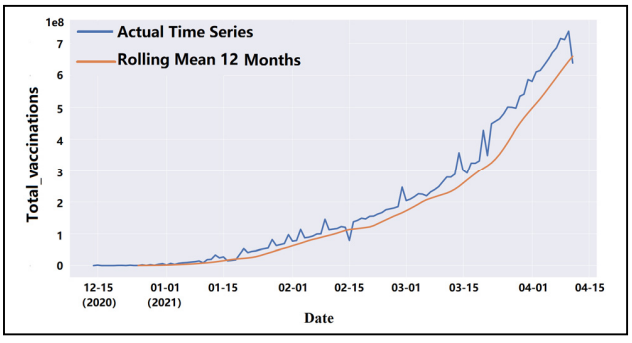

Figure 10. Global trends in the number of total vaccinations.

Figure 10 shows a global time series plot of the number of total vaccinations. Actual time series represent the situation of real data. The curve shows an upward trend on the whole, but there are still fluctuations in some parts. In order to eliminate accidental variations in the real data and find out the trend of the total number of vaccinations, this figure introduces the Rolling mean, which is commonly used to flatten out short-term fluctuations in time series data and to highlight long-term trends. Its essence is to use historical records to create data that can replace the original data. From the Rolling mean curve, it is very clear that the growth rate of the total number of vaccinations in the world is growing.

4) Comparison between China and India

There are two countries in the world with a population of more than 1 billion. They are China and India. By the end of 2019, Chinese total population has reached 1.443 billion, and Indian has 1.39 billion. Due to the limited information available in China, the analysis will focus on the total and daily vaccinations.

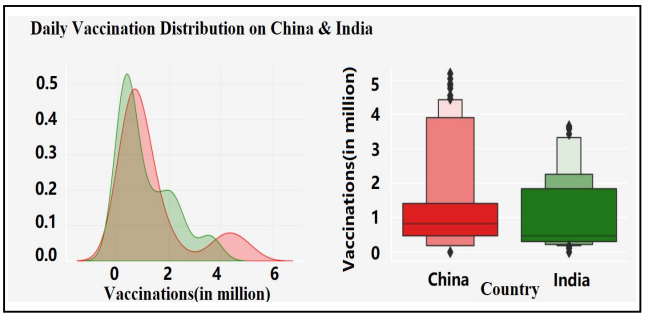

Figure 11. The distribution and boxplot of vaccinations.

China and India are the two most populous countries in the world, with the highest daily vaccination dose ranking first and second in the world. In Figure 11, although the daily vaccination rate in China is high, the distribution pattern in China and India is basically the same. It can be seen that the median daily vaccination in China is 660000 / day, higher than 360000 / day in India from the enhanced box chart .

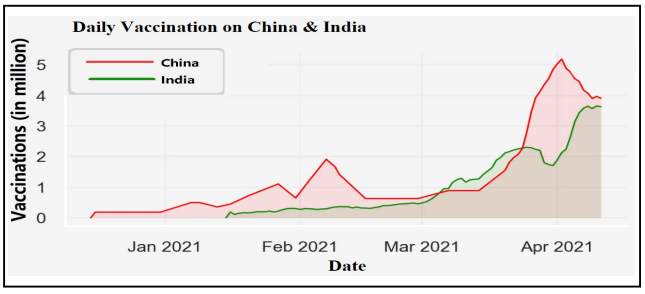

Figure 12. Top 5 countries with the highest daily vaccination.

In Figure 12, China has started their vaccination program earlier than India. In February 2021, China has done 1.5 million vaccinations per day before it drops below 1 million in the same month until the end of February 2021. Contrary with China, India has a more consistent increment since the start of the vaccination program. Rapid increased in daily vaccinations can be seen in the start of March 2021 for China and India. Both countries have more than 2 million daily vaccinations and are expected to continuously rise up. In April 2021, the daily vaccinations in China and India reach an unprecedented level, with a maximum of more than 5 million doses in China, and then dropped rapidly to about 4 million. So far, the highest number in India is only about 3.6 million, which has remained flat since then.

After vaccinating high-risk groups last year, China changed its focus and now targets the working population aged 18 to 59. The over-60s and those with potential health problems such as high blood pressure and diabetes will follow.

IV. Conclusion

The analysis of this data set indicates that the global vaccination situation is improving. There are 10 kinds of vaccines have been put into markets. Among them, Johnson &Johnson, Moderna, Pfizer/BioNTech combination of vaccination, the total amount is about 175 million times. In terms of total vaccinations, the United States, China, and India lead the way. In terms of the fully vaccinated rate, countries and regions are quite unbalanced. Most of the countries that can achieve high coverage rates are small countries with a population of one million. Gibraltar ranks first in the world at 88.13 percent. However, the global vaccination rate is still low, with only 1.7 percent. Under the background of continuous promotion of global COVID-19 vaccination, we expect that the epidemic situation, especially in developed countries in Europe and the United States, will take the lead in showing clear signs of control in 2021. It is still a long time to observe when the global epidemic will be truly controlled. Continue to explore the information hidden in the vaccine data, will have a more in-depth and clear understanding of the whole situation of the fight against the epidemic.

Acknowledgment

This paper was funded by projects: Zhejiang Province (lgf19f020011、Y202044936);Hangzhou(20191203b14).

References

[1] J.X. Liu, H.Y. Liu, X.H. Chen, J. Li, L. Kang, Q.B. Zhao. (2020) Journal of Computer-Aided Design & Computer Graphics, 32(10):1617-1627.

[2] R. Pronoy, D. Ankit, M. Indranil, Maity Dr. Saikat. (2021) Data Visualization to solve COVID-19[J]. Journal of Physics: Conference Series ,1797(1).

[3] P. Rahul, A. Kartik, N. Anubhav. (2021) Analyzing Evolution of the Olympics by Exploratory Data Analysis using R[J]. IOP Conference Series: Materials Science and Engineering ,1099(1).

[4] He Xueying. (2017) Design and Implementation of Exploratory Data Visualization Analysis System [D]. Southwest Jiaotong University.

[5] Sun Lijun. (2005) Exploratory Data Analysis Method and Application [D]. Dongbei University of Finance and Economics.

[6] Wang Yingui, Wang Ben, Huang Yuxing. (2020) Comprehensive Analysis and Mining Big Data on Smart E-commerce User Behavior, Journal of Physics Conference Series, 1616:1-6.

[7] Tan Renchun, Jiang Wei, Ma Yiwen, Chen Junwei. (2021) Epidemic data visualization in Python and FME [J]. Geospatial Information, 19(01):1-4.