India Fights Back: COVID 19 Pandemic

Abbu Bucker Siddique
Department of Computer Science and engineering
PES UNIVERSITY
Bangalore , India
abbubucker124@gmail.com

Aditya S Raj
Department of Computer Science and engineering
PES UNIVERSITY
Bangalore , India
sundarrajaditya@gmail.com

Eknath Reddy
Department of Computer Science and engineering
PES UNIVERSITY
Bangalore , India
ekanathreddydinsi009@gmail.com

Abstract — This paper describes our first research experience in the field of data analytics by using open-source software tools such as R studio and other Python Interpreters. Eventually, the project focuses on Exploratory Data Analysis and Data Visualization for the data-sets in picture. This would help us study the behavior of the attributes and their values thus highlighting points of interest for us to work ahead with.

Keywords— *EDA, Data Visualization,CoronaVirus,R,Python*

I. Introduction

Two years ago, the world was marked by the SARS-CoV-2 virus pandemic causing the infectious COVID-19 disease leading to millions of deaths , tens of millions of illnesses, hundreds of millions remaining quarantined and billions of people having had their lives changed. Although the virus was new, a year of research by scientists around the world led to the development of several safe and effective vaccines in India, the world's second most populous country.
Several factors speak for this country and its performance in the global market.Examining what affected or continues to affect India and how India as a country controlled the pandemic will help draw and announce common conclusions.

II. RELATED WORK

Our data has been obtained from three different open data-sets from Kaggle that together will be used to draw any observations based on our defined statements.

In paper [1] the author specifies the research performed in the field of “corona virology” considering how coronavirus evolved and analyzed this virus function .They have commonly used reverse genetic functions and titration techniques to identify the cellular receptors.

In paper[2] researchers have studied and analyzed how COVID-19 virus spreaded across the world using “Bailey’s Model”. It was interesting to see that high correlation coefficients(91%) was observed using Pearson's correlation method.Also WHO’s daily reports were considered for the analysis of countries across the globe.It also indicated the difficulties in correctly predicting the future spread of the pandemic.

In paper [3] researchers used a Prophet Model to provide understanding of the number of people who were affected by this disease. Prophet is an additive model introduced by FaceBook which is very popular for forecasting time series data.It detects separately the non-linear trends in the time series and then combines them together to obtain the forecast value. This model forecasts 90 days future growth trends and finds the peak time for all 6 countries and 6 states of India.Well this model has achieved around 85% MAPE for all the 6 countries and the 6 states of India.

Paper [4] investigates how the ARIMA model was developed to analyze the spread of outbreak for 21 states of India and the top 6 countries of the world.This model provides an understanding of the number of people affected daily by this disease.The proposed model has achieved around 85% in terms of accuracy for all the 6 countries and 21 states of India. This model consists of 3 parts : (i) an autoregressive part (AR) ,(ii) a contribution from a moving average (MA) and (iii) an integration part and the model is denoted as ARIMA(p,d,q)

We assumed that the selected dataset was reliable and already cleaned. We limit our research to India only as defined in our title.

III. THE DISPROPORTIONAL IMPACT OF COVID19 ON THE INDIAN ECONOMY AND ITS STAKEHOLDERS

A. Dataset

We obtained our datasets from the World Factbook collection that has been put out for public use and the Our World in Data GitHub Repository that actively collects and stores data every single day. The parameters vary based on the targeted solution. We chose the data that felt relevant to our study of the Indian Subcontinent and its ability to tackle the endemic.

B. Exploratory Data Analysis

1)
The dataset with the vaccination details of different countries has 86512 ROWS and 8 COLUMNS.
It has a very sizable number of missing values, here 184790 observations across the data-set.
Data inconsistency prevails as long as missing values are not treated properly.
Duplicates are also looked into and resolved due to the combined uniqueness of two attributes in this particular data-set.
Missing values have been filled with zeroes as no other metric is suitable.
This is done to ensure completeness and help us with our further observations.

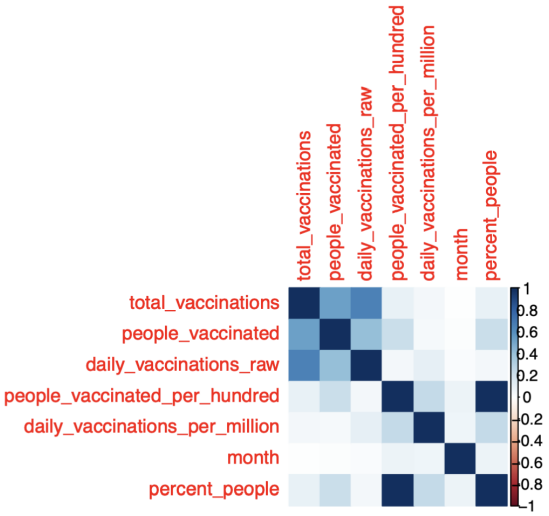

Fig.1. Correlation between attributes in the country vaccines dataset

The correlation plot can be observed to say there is no negative correlation between any of the attributes.
percent_people and people_vaccinated_per_hundred is very strongly correlated.
Most attributes that depend on people or attributes that directly contribute to another attribute (eg: people_vaccinated and total_vaccinations) show high correlation.

Outliers were identified by transforming into time series data but could not be replaced by a suitable metric since this data-set comprises of real time data which is necessary for our study.
Hence we will not be addressing them as outliers thus making the outlier count equal to 0.

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5    PC6        PC7
## Standard deviation     1.5678 1.3329 1.0001 0.9466 0.74534 0.5600 1.016e-12
## Proportion of Variance 0.3511 0.2538 0.1429 0.1280 0.07936 0.0448 0.000e+00
## Cumulative Proportion  0.3511 0.6049 0.7478 0.8758 0.95520 1.0000 1.000e+00
```

Fig.2. Principal Component Analysis followed by summary done on the dataset

Proportion of variance for all 7 numeric principal components is low and PCA would not be the best option.
Other transformations also do not seem fit due to the nature of this data-set.

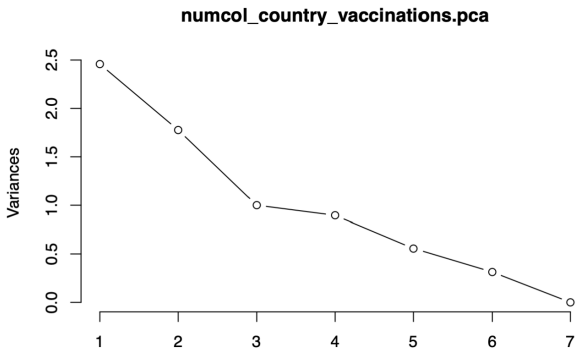

Fig.3. Arm bend plot to study cumulative contribution of numerical attributes in the data-set

In the screen-plot above, the 'arm-bend' represents a decrease in cumulative contribution.
The above plot shows the bend at the third principal component.

Outliers were identified and capped to fall within a suitable range but that would not benefit our study.
Hence, we have not treated them as of now.

For pure observation, the relevant numerical columns have been plotted as a function of time.The following figures show the same.

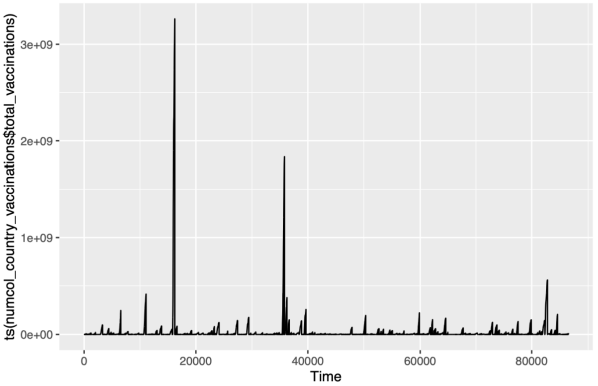

Fig.4.1. Plot of total_vaccines as a function of time

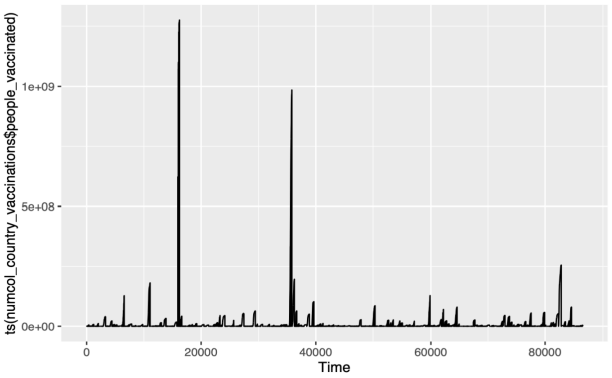

Fig.4.2. Plot of people_vaccinated as a function of time

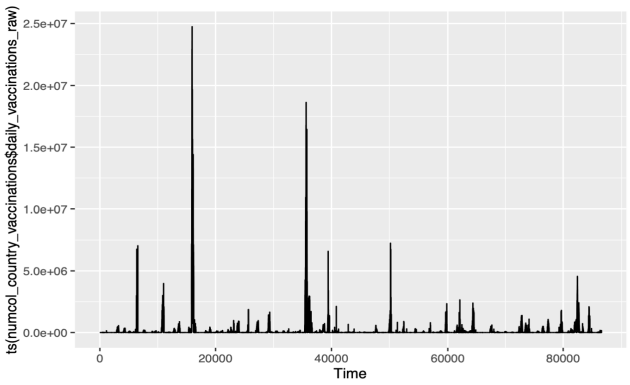

Fig.4.3. Plot of daily_vaccinations_raw as a function of time

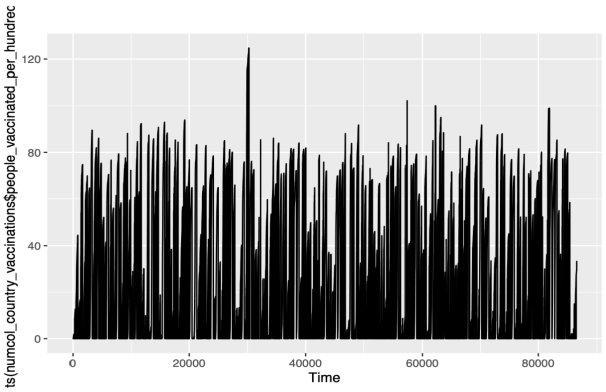

Fig.4.4. Plot of people_vaccinated_per_hundred as a function of time

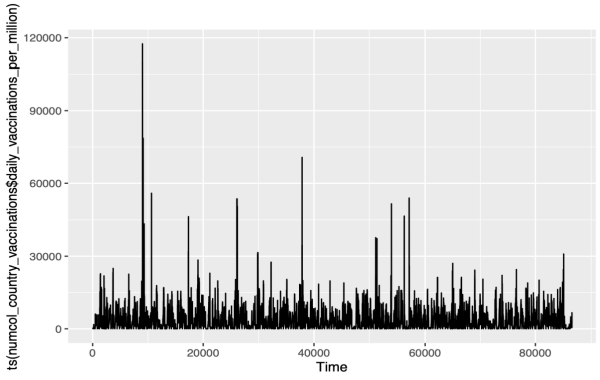

Fig.4.5. Plot of people_vaccinated_per_million as a function of time

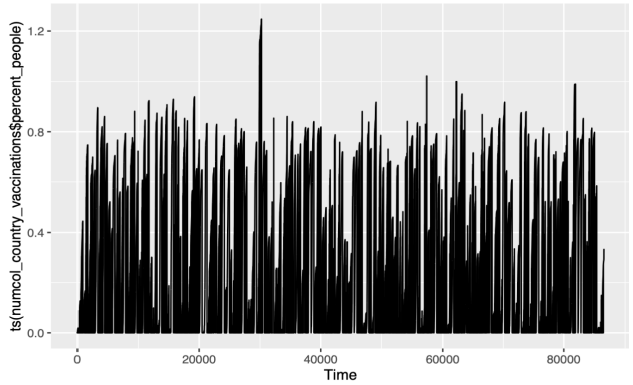

Fig.4.5. Plot of percent_people as a function of time

The plots suggest that there are outliers but we will not be treating them for our analysis.
This variation in values is what makes the base of our study.

2)

The dataset with the details about the manufacturer's of vaccines has 9895 ROWS and 4 COLUMNS.
One good thing about this dataset is that it has no missing values.
We had to change the date attribute from type object to type datetime64.
We found the total number of vaccines provided by each company and observed the below table:

vaccine	total_vaccinations
Pfizer/BioNTech	3.801997e+08
Moderna	1.786214e+08
Oxford/AstraZeneca	3.141898e+07
Sinovac	2.197000e+07
Johnson&Johnson	1.475870e+07
Sinopharm/Beijing	3.041437e+05
Sputnik V	2.594869e+05
CanSino	1.794493e+05

Fig.5. Table listing total vaccinations by each vaccine manufacturer

To try and understand it graphically , we plotted a bar graph.

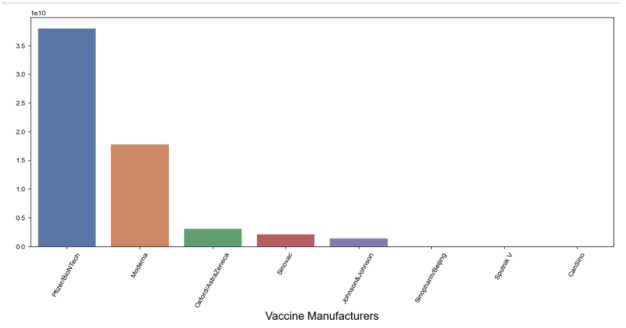

Fig.6. Plot of total vaccinations against various vaccine manufacturers

It was observed that Pfizer/BioNTech was way ahead than the other manufacturers.

3)

The dataset with the details of countries and various other factors like GDP, literacy rate, Birth Rate etc has 227 ROWS and 20 COLUMNS.
There are 110 null values present in this data set.

It was wise to replace the null values with their respective mean since the data set was not too large and all of the numerical attributes followed a near normal distribution.

The following figures show the distribution of observations of the relevant numerical attributes in the data set:

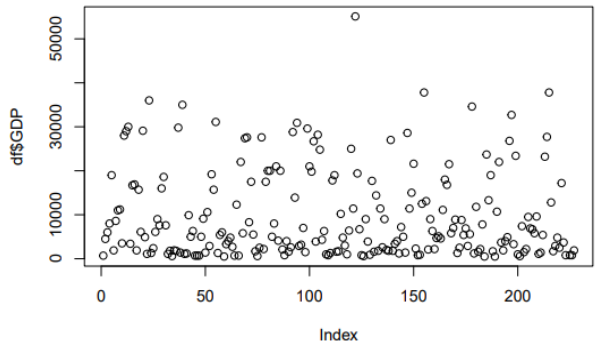

Fig.7.1. Plot of attribute GDP

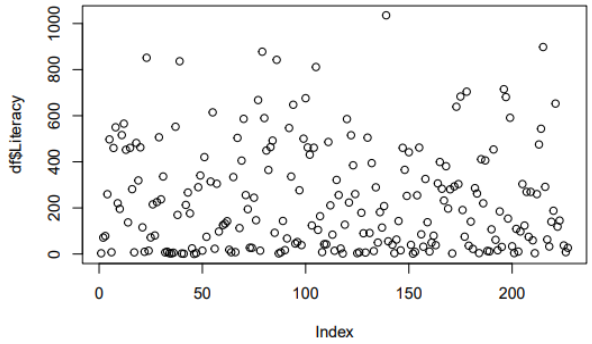

Fig.7.2. Plot of attribute Literacy

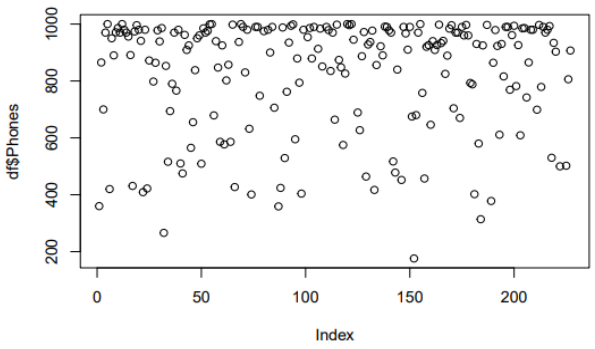

Fig.7.3. Plot of attribute Phones

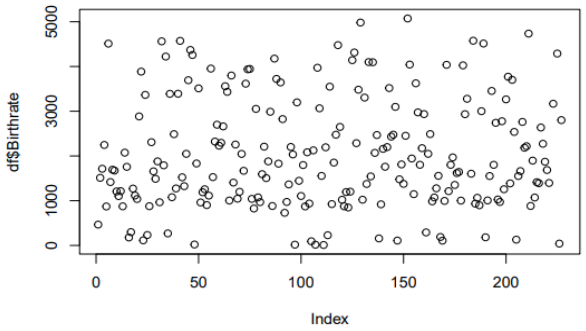

Fig.7.4. Plot of attribute Birth Rate

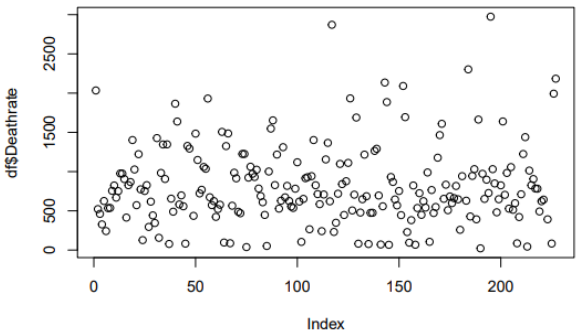

Fig.7.5. Plot of attribute Death Rate

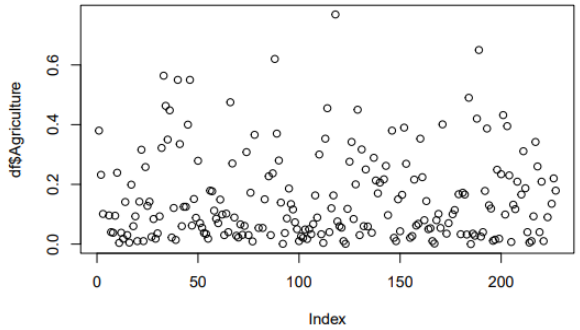

Fig.7.6. Plot of attribute Agriculture

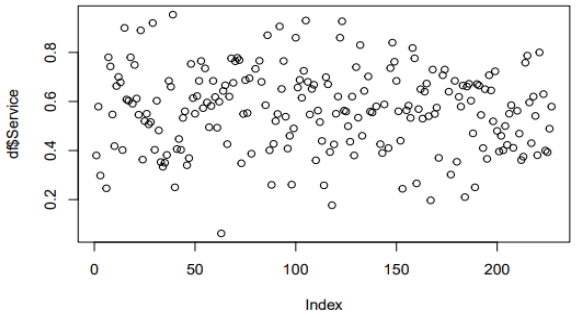

Fig.7.7. Plot of attribute Service

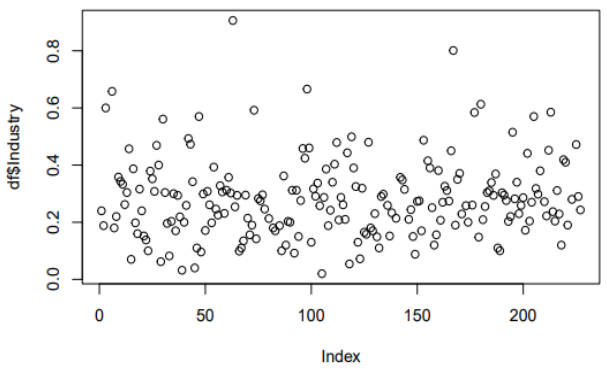

Fig.7.8. Plot of attribute Industry

We plotted the Spearman's Correlation matrix to find out the correlation between various attributes present in the data set.

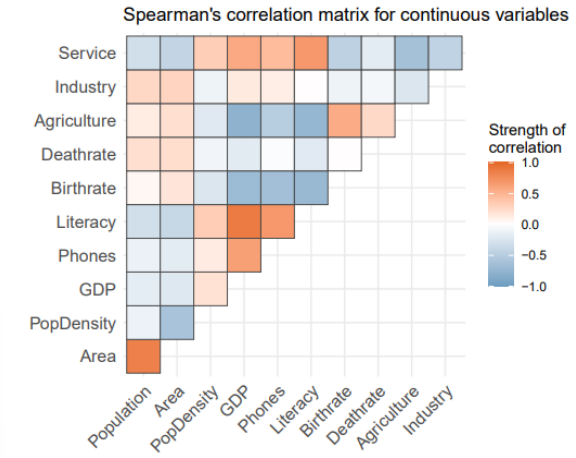

Fig.9. Correlation between attributes in the countries of the world dataset

It can be observed that metrics that depend on the population measure of people are positively correlated. The rest are not significant for our study as of now.

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.023 1.2896 1.1350 0.98996 0.94461 0.77946 0.72161
## Proportion of Variance 0.372 0.1512 0.1171 0.08909 0.08112 0.05523 0.04734
## Cumulative Proportion  0.372 0.5232 0.6403 0.72938 0.81050 0.86573 0.91307
##                           PC8     PC9    PC10    PC11
## Standard deviation     0.69830 0.56276 0.38653 0.05030
## Proportion of Variance 0.04433 0.02879 0.01358 0.00023
## Cumulative Proportion  0.95740 0.98619 0.99977 1.00000
```

Fig.10. Principal Component Analysis followed by summary done on the dataset

Proportion of variance for all 11 numeric principal components is low and PCA would not be the best option.
Other transformations will be applied as and when required.

IV. PROPOSED SOLUTION

After some discussion, we arrived at a suitable solution that could very well answer all the questions regarding the problem at hand.
We planned to create a graph and compare total cases, new cases, and active cases. We also planned to determine India's ranking in terms of total deaths due to COVID-19. We tried to present the solution in a very unique way. We thought of using a choropleth map to help easily visualize how a variable changes over a region, or to show the degree of variation within a region.
We plan to analyze the selected dataset and create some nice charts to help visualize the data.
Clearly determine which factors rank countries and determine global rankings based on available datasets.
Hopefully, the plan itself will allow India to defend and prove its greatness in fighting COVID-19.
This could possibly be the best solution to prove how India as a "Developing Nation" has dealt with COVID-19.

V. EXPERIMENTAL OBSERVATIONS

Since EDA is already in place we start visualizing the requirements for coming up with suitable and valid conclusions.
We start with the most important aspect of successful vaccinations i.e the number of vaccinated people per day per country. This comparison will help us track India while also acknowledging the other well to do countries.
In many countries there is an ascending trend, but also large fluctuations.
Since, this analysis is done over a certain period, it can be largely varied for minor changes in time

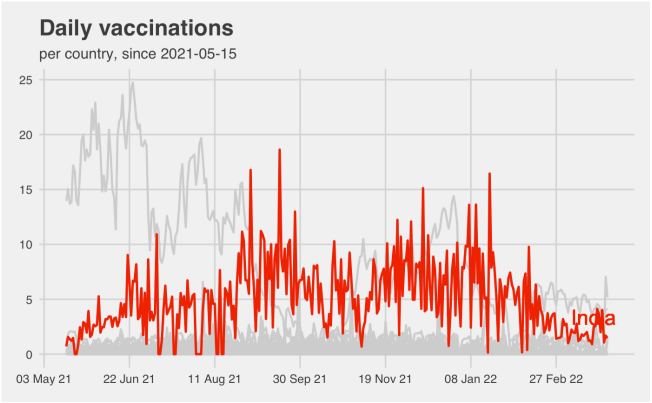

Fig.11. Daily Vaccinations Per Country over certain time period

India was the leader in vaccinations for a certain time mainly due to its massive population. This linear relation between the number of successful vaccinations and the population at harm was the first step in ensuring global dominance.

However, the situation over any number of days is cumulative and needs to be analyzed separately. This means that in the absence of attached people, the value doesn't change, while with each newly vaccinated person the value increases.

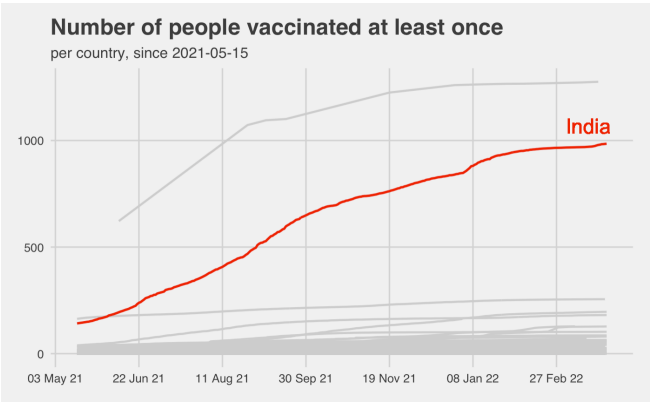

Fig.12. Number of people vaccinated at least once

We can see that in India the increase is quite rapid compared to other countries - this is due to the fact that the number of vaccinations daily increased over time due to major changes in government orders and people's personal safety.

So India currently has a large number of vaccinated people, and it also vaccinates a large number of people everyday- so is that enough to call it a leader? No, there are a few independent indicators that provide deeper insight for the same.

The answer is the percentage of vaccinated inhabitants - the most independent indicator. Let us explore it for a better understanding of our country's situation in the past.

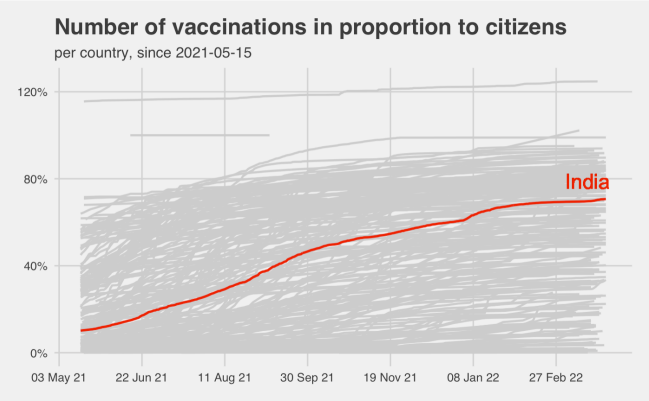

Fig.13. Number of vaccinations in proportion to population

At the very beginning of the vaccination process, India did not come close to the other superpowers of the world in terms of percentage of population vaccinated. This might be due to the vast population but it should be no excuse for the same.
Currently, India matches its progress to that of similar countries and this has been achieved rather quickly. This ratio may exceed 100% as most of the available vaccines are two-dose, i.e. two doses must be taken at a time interval to achieve immunity.
So what is the secret of this country? What made us get so many vaccines so quickly, while many countries (even of the same size or similar GDP per capita) have a serious problem to vaccinate at least 10% of the population at the same time? So let's look at the process in a broader perspective, looking for the differences in data between India and the rest of the world.

We start with the geographical location - we look at the vaccination process in neighboring countries, comparing the most objective measure, which is the number of vaccinated per 100 inhabitants of the country. As we can see, neighboring countries in SouthEast Asia are very far reaching and perform much worse in the proportion of vaccinated people. Therefore, we can observe on these examples that vaccination is also carried out more efficiently in other small territorial countries. It's worth adding that for some countries these data are not updated , while we only have information about the number of vaccinations.

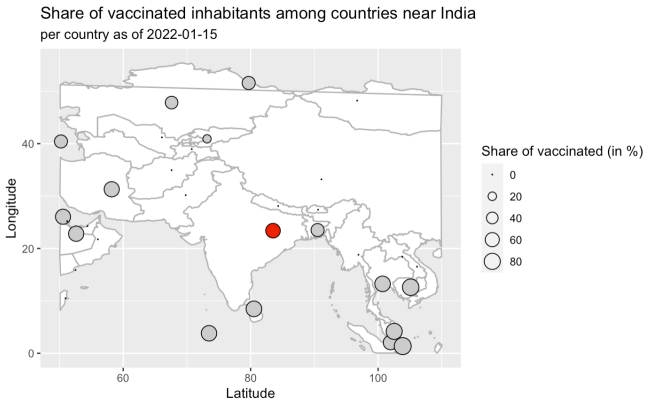

Fig.14. Share of Vaccinated inhabitants among nearby countries

Another idea that might make vaccinations different is spreading them over the course of the week. Perhaps the differences may be due to stopping the vaccination on weekends or a different strategy. We can see that both in India and in the rest of the world there are no big differences in vaccination schedules by day of the week. Comparing the averages, we can see that the differences are so small between the two analyzed groups that we can undoubtedly say that the weekly schedule is not the reason for the success of this country.

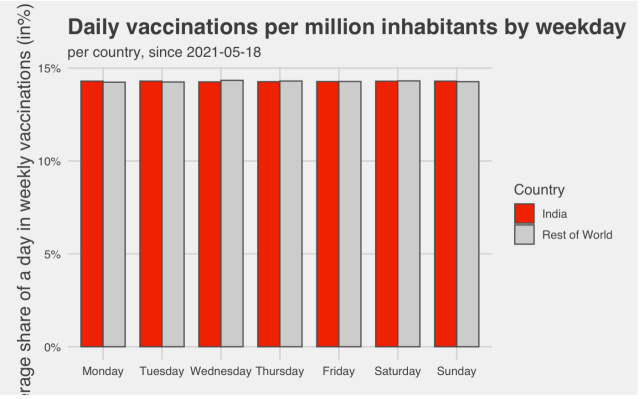

Fig.15. Daily vaccinations per million inhabitants by weekday

As we recalled at the beginning, not one, but several vaccines have been approved for use (and their number may increase). So let's take a look at India's performance in terms of the number of different vaccine suppliers compared to other countries. The record holder here is the United Arab Emirates, where as many as 6 different vaccines are used. Some European countries already use the services of 3 different suppliers, while some of them only use 2

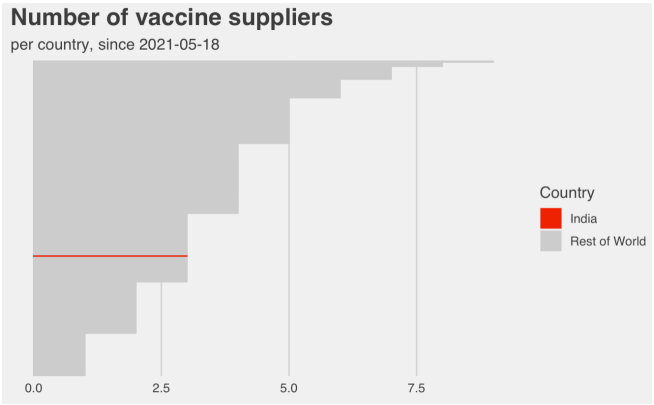

Fig.16. Number of Vaccine Suppliers per country

The vaccination process in India doesn't seem to be crucial for their rapid pace compared to other countries. The decisive factor was the rapid start of mass vaccination, while other countries were just thinking about organizing the process. But why was it possible to organize it so quickly in India? Perhaps a quick purchase of vaccines, perhaps public pressure after the epidemic escalated and the lockdown was introduced?

However, we'll look for answers in other data, looking more broadly for the economic and social aspects in which India stands out from other countries. For this we will use data from the "World Factbook" describing the various country-specific variables.

The data isn't the latest and comes from 2017 or earlier years, but it should be enough to find the current dependencies. We have selected several variables that may have a substantive impact on the current situation - from demographic variables to access to information.

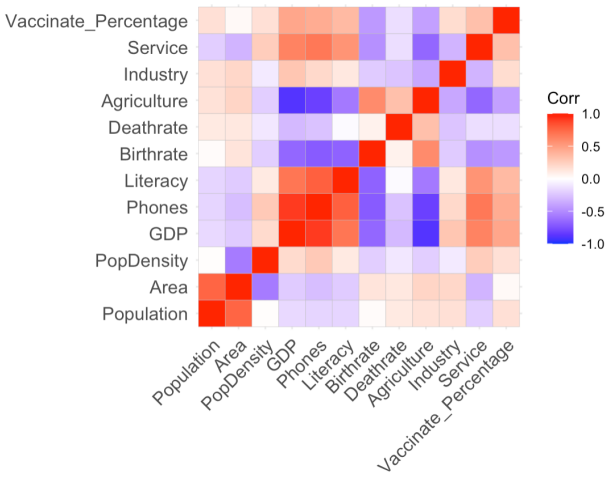

Fig.17. Correlation matrix of contributing features

The share of vaccinated people in a country is not strongly correlated with any other variable we have in the second base. The strongest negative correlations, which are only around -0.5, are the size of the GDP and the share of agriculture in the economy. This means that the increase in the value of these variables more often coincides with low values of the percentage of vaccinated persons in the society. Moderate positive correlations can be seen with the following variables: GDP per capita, number of phones per 1000 inhabitants, share of literacy and share of people working in service sectors. The lack of strong correlations also makes it difficult to find in this information any significant reasons for the rapid progress of coronavirus vaccination in some countries.

We come to the position of India among other countries according to the given variables. We're looking for a system where India can stand out from other countries, just as it excels in vaccination. We start with the country area, population, and a combination of these two variables, i.e. population density . Logarithmic scales have been used for the distribution of the variables on the further axis to improve readability and better show the differences between countries. In all three characteristics, India is in the middle of the middle without any distinction. Therefore, this country doesn't significantly influence the negative correlations of vaccines with these variables and isn't a decisive factor for such a large percentage of the vaccinated society. We keep looking.

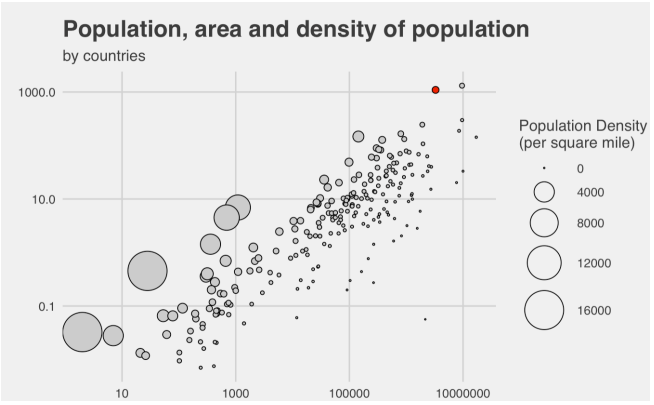

Fig.18. Population and Area density

Moving on, we look at the demographic variables: death rate and birth rate, that is, respectively, the number of deaths and the number of births per 10,000 inhabitants per year. When the birthrate value is lower than the death rate value in that country, the population (not including migration) decreases. India is a country with a low number of deaths in relation to the population and a moderate number of births (the number of inhabitants increases every year). It doesn't stand out from other countries, although the demographic situation in this country is much better than in many wealthier countries. This doesn't seem to be a significant reason - in other countries with similar demographics and numbers the number of vaccinated people is low

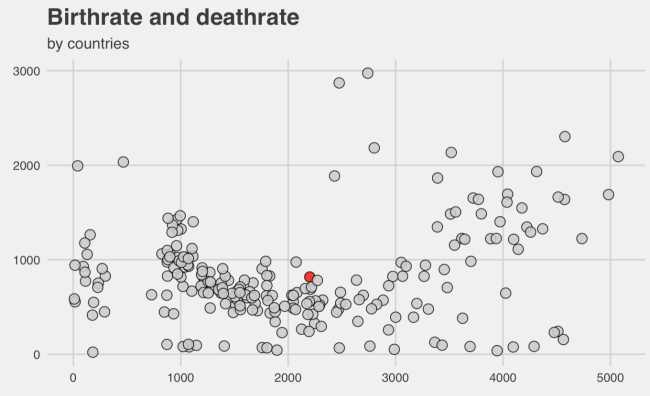

Fig.19. Birthrate and Death Rate

If vaccination isn't affected by geographic and demographic variables, access to information may have an impact. Better information flow in society means more people will get vaccinated, more people will get truthful information about vaccines and will be more likely to get vaccinated. In the database, we have two pieces of information related to this topic: the rate of the society that can write and read and the number of phone calls per 1,000 inhabitants.

Most countries have over 90% of the population that can read and communicate information other than by speech and access telephones over 500 devices per 1,000 inhabitants. This group does not include India, where the number of telephones is particularly low for the population. However, this doesn't distinguish this country, because in today's world such values are present in many countries on all continents, so access to information is equally easy in many countries, which have not yet embraced even 5% of the society.

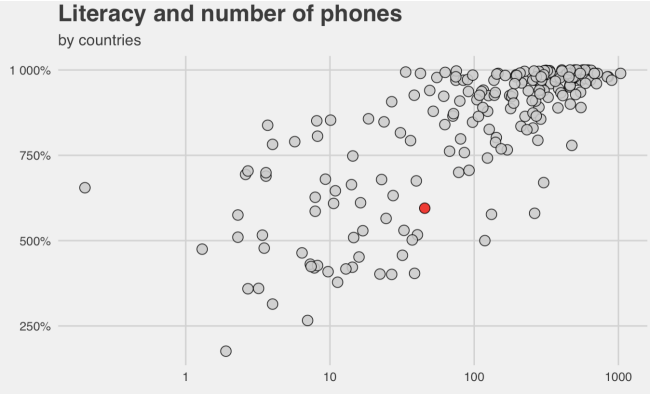

Fig.20. Literacy and Number of Phones comparison

We now look at the most popular economic indicator which is Gross Domestic Product per capita in US dollars. The value of this ratio for India is well below the average and median for over 200 analyzed countries. This country stands out from the rest, but it cannot be denied that richer countries find it easier to negotiate the purchase of vaccines - if the economic situation of the country was much worse, India would probably not become a leader. Much richer countries do not vaccinate such a large percentage of the population, so

the economic situation doesn't always make the country cope well with the mass vaccination process.

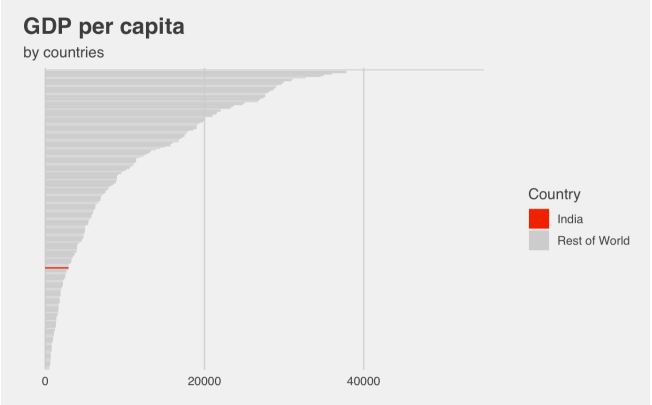

Fig.21. GDP per capita

Next we look at countries in the structure of the economy, i.e. the share of agriculture, industry and services in employment. Due to the large number of countries, the chart is limited to a dozen or so countries with the closest GDP, excluding very small countries (mainly island countries) from this group. All of the represented countries (except two) are more than half employed in services, including India. All countries have a very small share of agriculture in their employment. India doesn't differ completely from the structure of employment of the population from the average profile of the country, the labor market is certainly not atypical in this country. The poor correlation of variables with the number of vaccinated people per population also indicates that this is not a feature that influences the situation with vaccinations.

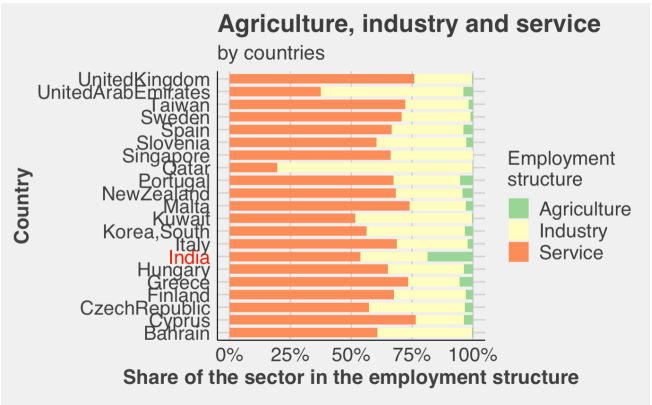

Fig .22. Comparison of GDP factors for similar countries

VI. CONCLUSIONS

The aim of the analysis was to find in the selected data the reasons why India is so good at vaccination. We approached the subject from two sides: analyzing vaccination data between countries and looking for other variables (including economic, demographic, etc.) in which India stands out and looking for variables that may affect the vaccination system in this country..

It turns out that it's extremely difficult to find features that contribute to becoming a leader, even if only to a moderate extent. Certainly India was in such a high position in late 2021 as it began mass vaccination the fastest and continues to maintain a good pace. Such a quick action could have been helped by a not very large area and number of inhabitants as well as low GDP values, but these are not key factors, as other countries with similar parameters are doing much worse. Perhaps there are other variables that distinguish India to a much greater extent, and perhaps it isn't a question of the economic and social profile of the country but of the reasons given by the media, such as political goals of the current government, a broad information campaign or mobilization resulting from the very difficult epidemic situation in the fall. New data comes in every day and India has lost its leadership position and was not the first country to start mass vaccination, but will forever remain the first country to vaccinate a huge fraction of its population in a very short time.

In this era, nothing can be declared without proof.

We have tried to include a few plots which could very well conclude our discussion.

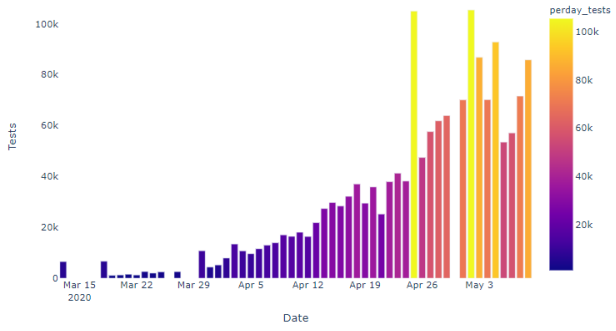

Fig.23.Plot of number of COVID-19 tests conducted

From this plot it could be very well observed that India left no stone unturned in testing its citizens. The graph indicates that the per day tests were increasing as time passed.

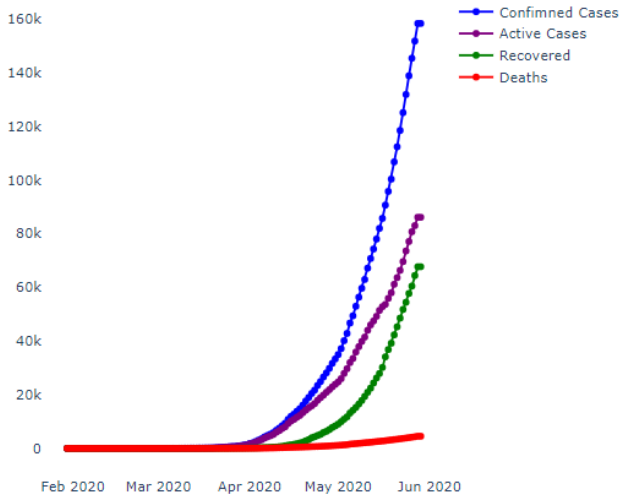

Fig.24. Plot indicating the confirmed,active,recovered and deaths.

This plot in itself explains how India fought COVID-19 even when having very high confirmed cases.The doctors, our life saviors fought very hard to increase the number of recovered cases and decrease the active cases and deaths.

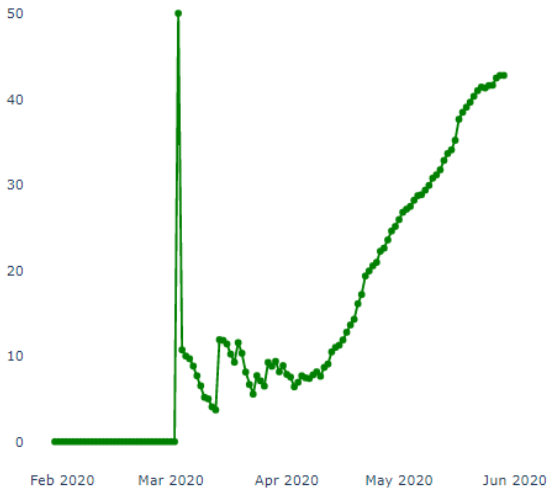

Fig.25. Trend recovery rate of India

This graph shows why India deserves to be known as the best COVID warrior across countries. The increase in plot gives us an idea on how the recovery rate increased even when there was an alarming rate of increase in total infections per day.

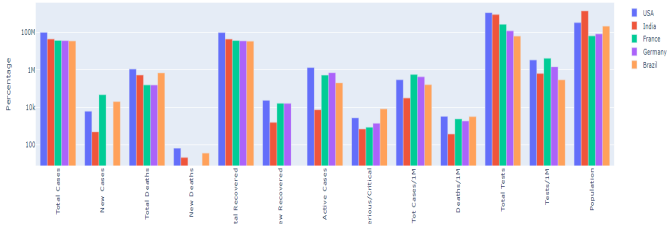

Fig.26. 5 countries with most COVID cases

Yet another plot indicating how India even after having a high number of daily cases, has the best total number of recovered counts.

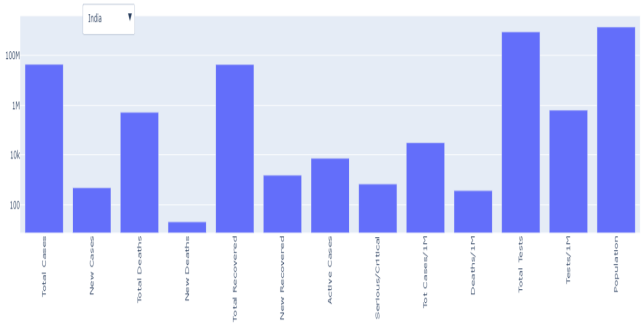

Fig.27. India and it's COVID-19 cases

The previous plot compared India with the other 4 countries that had a very high number of total cases.
This plot clearly gives insight into India's statistics, again proving India's role in COVID-19 was worth mentioning.

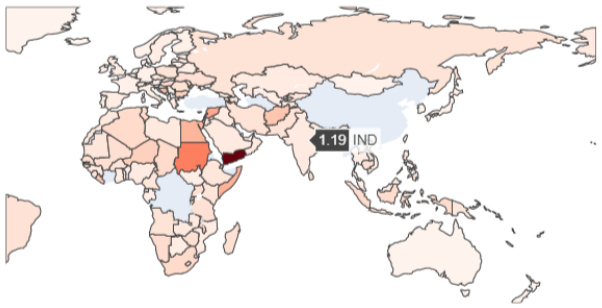

Fig.28. Death rate of countries
A very beautiful Choropleth plot indicates that India has a very low death rate when compared to the rest of the countries.

Disparity in contribution was never an issue due to the ample number of datasets, attributes and corresponding features.
Every section of this report has been dealt with as a team and the implementation is also based on common understandings of each team member involved.
Some general findings were included as a part of the report which were not the principal work of any of the members.
All relevant documentation at each step has been equally divided due to the numerosity of the datasets.

It becomes very difficult to summarize what we have learnt through this project
We have got a decent touch on this subject, learnt how to clean data, pre-process it and perform EDA.
There was never an end to learning new things because of the immense growth in the field of Analytics.
We came across many different amazing tools,packages etc that helped us play with the dataset.
Very well understood that Data Analytics is neither a "hard" nor a "soft" skill but is instead a process that involves a combination of both.This journey has given us the idea of thinking differently across situations.
Common take-aways are the vast application and scalability of analytical measures that would narrow down the largest pile of data into the most accurate and insightful facts.

Let's hope that we will soon be able to say that the virus has gone down in history in all corners of the world - data trends give it a lot of hope!

REFERENCES

[1] A. R. Fehr,S.Pealman," Coronavirus:An Overview of Their Replication and Pathogenesis," Springer, Methods Mol Biol., 2015;1282:1-23,doi: 10.1007/978-1-4939-2438-7_1

[2] D. Gondauri, E.Mikautadze and M. Batiashvili, "Research on covid-19 virus spreading statistics based on the examples of the cases from different countries," Electron J Gen Med.2020;17(4),em(209)

[3] Sujata Dash , Chinmay Chakraborty, Sourav K.Giri, Subhendu Kumar Pani "Intelligent computing on time-series data analysis and prediction of COVID-19 pandemics "August 2021 Pattern Recognition Letters 151(5) DOI:10.1016/j.patrec.2021.07.027

[4] Sujata Dash , Chinmay Chakraborty, Sourav K.Giri, Subhendu Kumar Pani, Jaroslav Frnda "BIFM: Big-Data Driven Intelligent Forecasting Model for COVID-19 " , Electronic ISSN: 2169-3536,INSPEC Accession Number: 21050274,DOI: 10.1109/ACCESS.2021.3094658