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, the work of scientists from around the world over the course of one whole year led to the development of several vaccines that remain safe with high effectiveness.We have looked at the vaccination process in relation to countries, particular to the country which houses the second highest population in the world-India.
Several factors speak for the country and its performance in the global market. Finding what influenced or continues to influence and how India as a nation was able to control the pandemic will help us draw common conclusions and herald the work that has been done by the authorities.

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)

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.

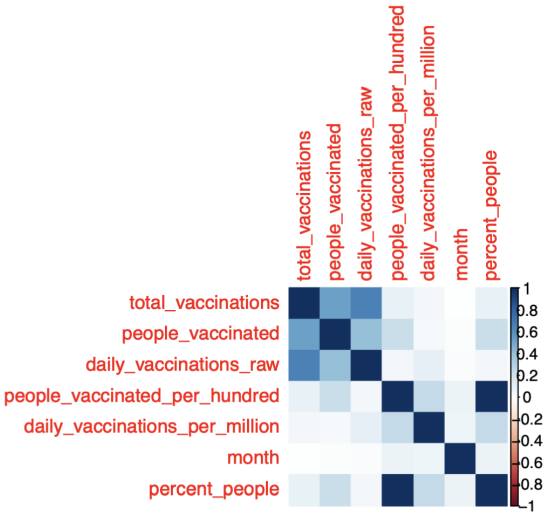


Fig.1. Correlation between attributes in the country vaccines dataset

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.

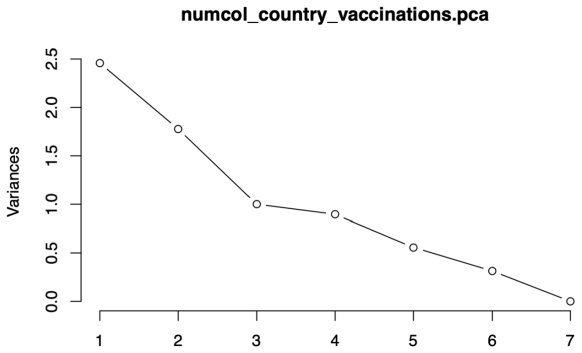


Fig.3. Arm bend plot to study cumulative contribution of numerical attributes in the data-set

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.

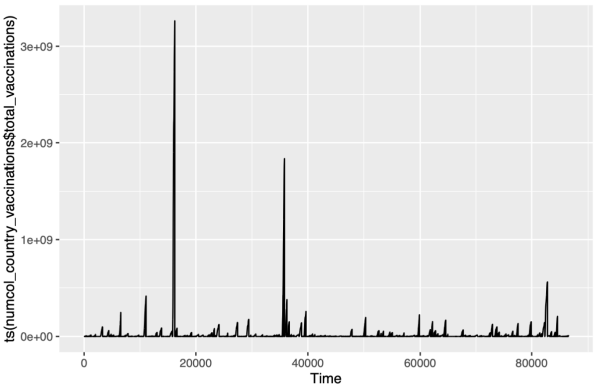


Fig.4.1. Plot of total_vaccines as a function of time

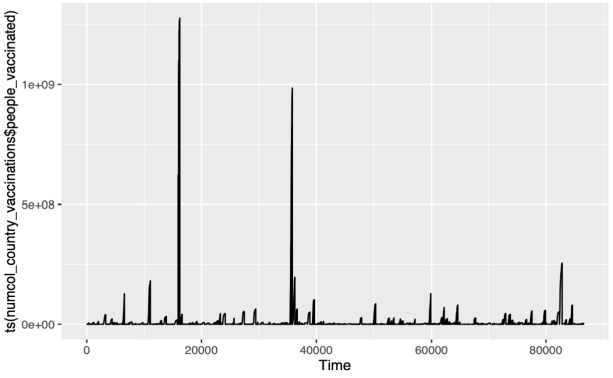


Fig.4.2. Plot of people_vaccinated as a function of time

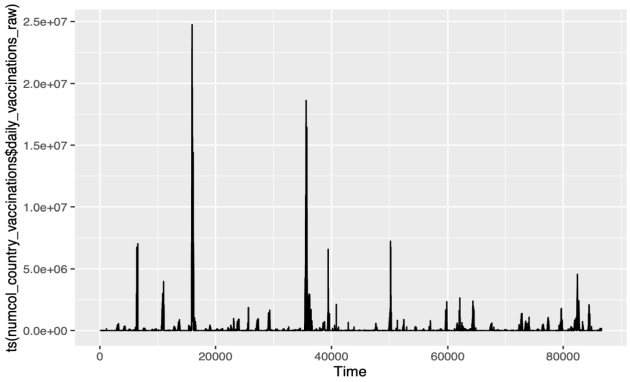


Fig.4.3. Plot of daily_vaccinations_raw as a function of time

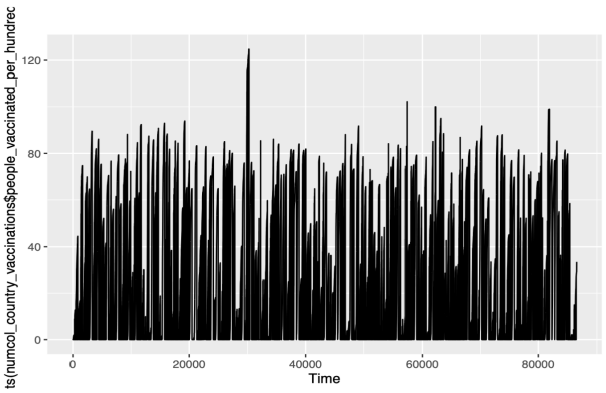


Fig.4.4. Plot of people_vaccinated_per_hundred as a function of time

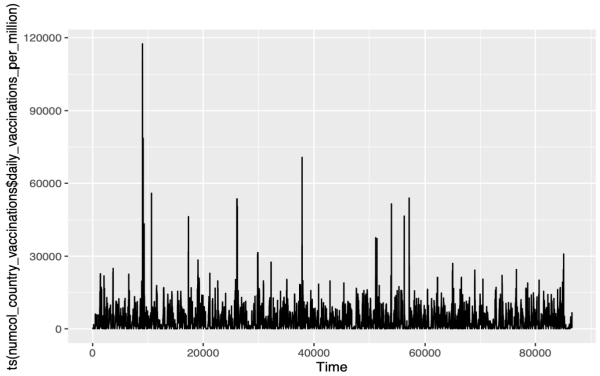


Fig.4.5. Plot of people_vaccinated_per_million as a function of time

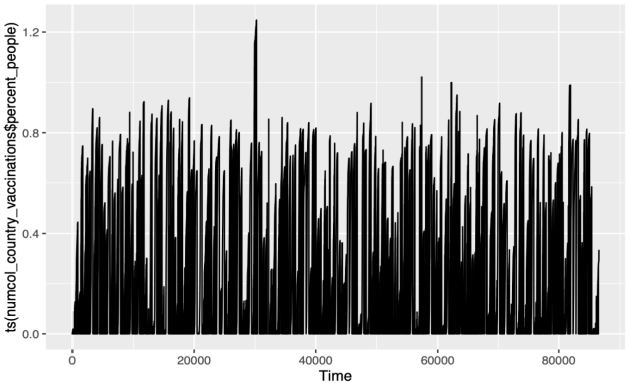


Fig.4.5. Plot of percent_people as a function of time

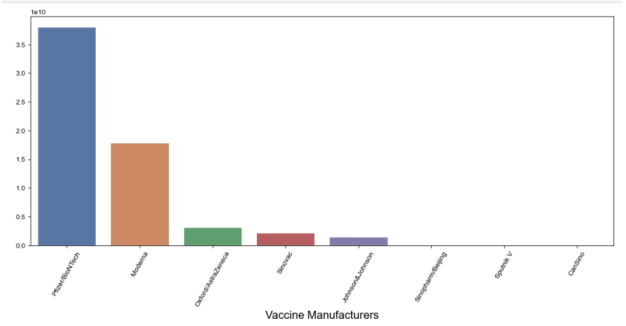
The plots suggest that there are outliers but we will not be treating them for our analysis.
This variation in values is what makes the base of our study.

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

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



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:

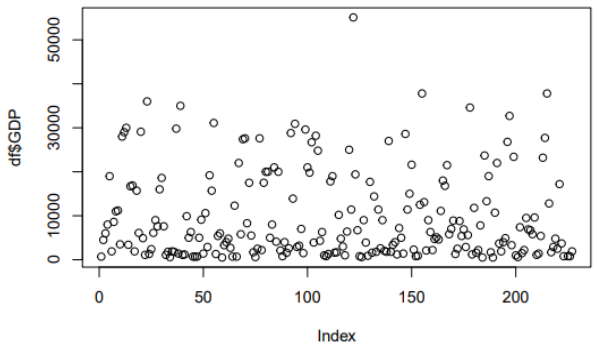


Fig.1.1. Plot of attribute GDP

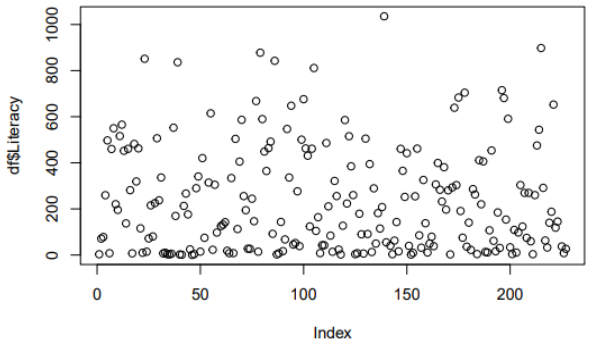


Fig.1.2. Plot of attribute Literacy

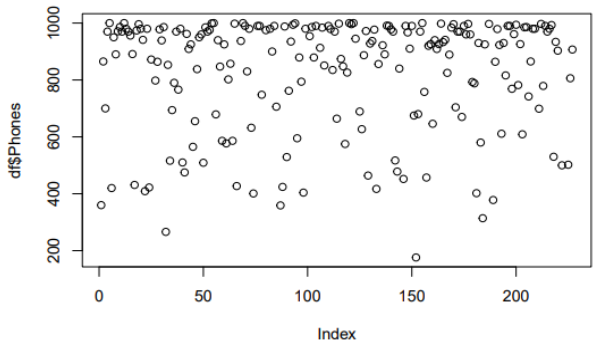


Fig.1.3. Plot of attribute Phones

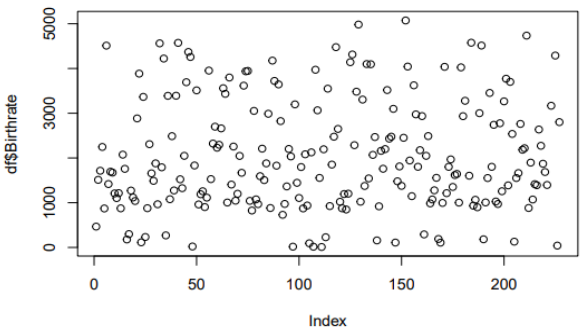


Fig.1.4. Plot of attribute Birth Rate

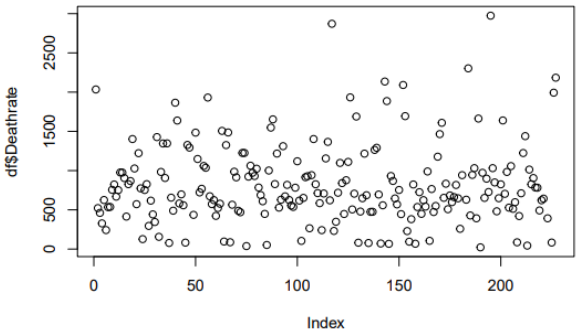


Fig.1.5. Plot of attribute Death Rate

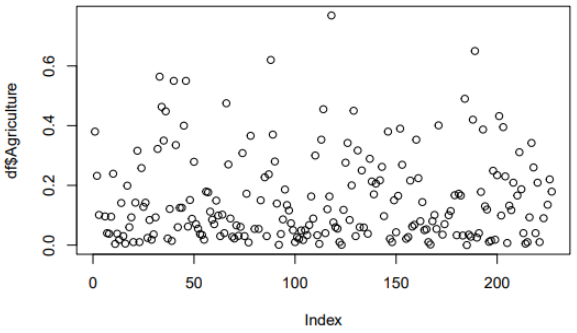


Fig.1.6. Plot of attribute Agriculture

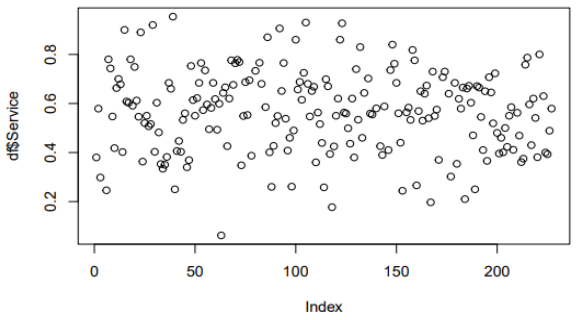


Fig.1.2. Plot of attribute Service

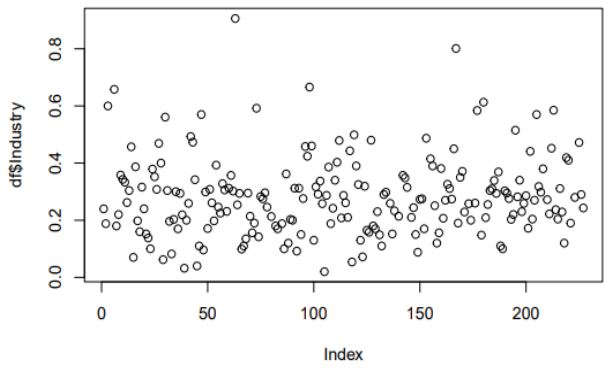


Fig.1.2. Plot of attribute Industry

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

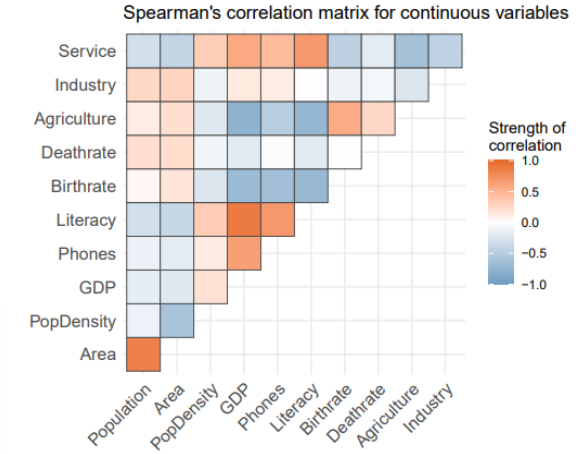


Fig.2. Correlation between attributes in the countries of the world dataset

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.3. 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.

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