COVID-19 DATA ANALYSIS AND DATA VISUALIZATION

Harinishri.S
Electronics and Communication
Rajalakshmi Engineering College
Chennai, Tamil Nadu, India

Abstract— COVID-19 (coronavirus disease 2019) is a viral infectious disease and presently, World Health Organization (WHO) has declared it as a global pandemic. By viewing the reports of August 2020, nearly 18 million people have been affected on the whole with over 700,000 deaths. As the cases are getting increased day by day, there is a huge demand for data manipulations and data storage. Thereby, data analytics play a vital role in identifying, arranging the data and visualizing in various forms. This can largely help in controlling the diseases in the regions and isolating those regions. Thus, data analytics is a new dawn in the medical field.

Keywords: Data Analysis, plots, Graphs, Covid-19, Global pandemic, Deaths.

I. INTRODUCTION

Data mining has already established as a novel field for exploring knowledge from hidden lattices in the big datasets. Data Mining can be defined as a non-trivial extraction of implicit, formerly not known and significantly valuable from given data. Briefly, it is a process to analyze the data from different perspective and gather the knowledge from it. This meticulously gathered knowledge can be used in different domains, majorly in healthcare industry. So we will use data mining techniques along with big data and IOT to assist doctors and other people in making decision of virus control in early stages.

Big data is quite an old method which is becoming the most primitive method to harness data as many health care organizations are keen in discovering new opportunities for better understanding and predict customer behaviors. It is the kind of data that outflows the processing ability of conventional or the traditional Database Management Systems(DBMS). A patient who gets admitted in a hospital, his/her activities are strictly monitored and is taken into account. This complete set of data is stored under the patients name and is referred when there is a need for a check-up. It is now very crucial to investigate patient care and reducing both mortality and morbidity concerned with covid attacks that can be used to improve or to set up an alert program.

II. METHODS

A) DATA ANALYSIS

The data of Covid-19 affected patients is taken from one of the most reliable sources. The individual data is taken from https://www.covid19india.org/ and the state level data is taken from https://www.mohfw.gov.in/. This data taken from the Indian states and used for the complete analysis of Covid-19 Pandemic. The cumulative analysis is done using Jupyter Notebook. The Covid-19 data taken from 01/03/2020 to 06/08/2020. In dataset1, we have the state-wise Covid attacked patients details and in dataset2, we have the state-wise-population and the area of that state.

![Fig 1:First 5 data of dataset1](image1)

![Fig 2:First 5 data of Dataset2](image2)

III. ACCESSING OF DATA

Data mining is known as the process of extracting valuable data from an enormous amount of any raw data. It also does analysis of data patterns in numerous number batches of data.
with the help of one or more software. Data mining has a wide range of applications in multiple fields, like technology and science. Data mining deals with efficient data collection, storing it as well as processing it. Data mining can also be called as Knowledge Discovery in Data (KDD). In order to view the widespread of virus in a India for a particular day, the below code is used for analysis.

```python
data_latest = data[data['Date']=='5/8/2020']
data_latest
```

This gives the data of people who got affected in all the indian states, people who got cured and also the number of deaths which took place.

To get a clear view of the collected data, we can using various types of graphs in Python Jupyter notebook. Some of the graphs available are

- Scatter Plot
- Line Graph
- Pie Chart
- Area Chart
- Legend Graph
- Relplot
- Bubble Chart
- Heat Map
- Funnel Chart
- Waterfall Graph
- Stacked Bar Graph

### A) GRAPHS IN PYTHON

#### SCATTER PLOT

To visualize the data, several graphs are used. The most reliable and efficient graph is scatter plot. A scatterplot is one kind of data representation that gives us the relationship between two independent/dependent variables. Each and every value present in the dataset gets plotted, like a dot whose $(x, y)$ coordinates corresponds to the values for the two variables.

If the $y$-variable increase when the $x$ variable increases, then it is called as a positive correlation between the two variables else the vice versa. A scatter plot must be used either if one continuous variable is under influence of the examiner and the other is dependent on it or when both these continuous variables are independent. The control parameter or the independent variable is plotted in the $x$-axis if there is a linear increasing/decreasing relationship exist.

![Visualization to display the variation in COVID-19 figures in different Indian states](image1.png)

**FIG 3: DETAILED REVIEW OF COVID-19 CASES IN INDIA**

![scatter plot of number of covid-19 cases per 10 million people](image2.png)

**Fig 4: scatter plot of number of covid-19 cases per 10 million people**
From the graph, we can visualize that when we take the state population into account, Maharashtra, Delhi, Tamil Nadu and Punjab have the highest number of covid-19 infected people. Followed by these states, other states like Nagaland, Goa, Chandigarh and also union territory Puducherry have considerable less covid-19 affected cases. From the statistics of Ladakh, 14 out 20 confirmed cases have been recovered.

- **BAR GRAPH**

A bar graph is a graph that represents data in a categorical fashion with rectangular bars having lengths tantamount to the values that contain. These bars can be plotted either in vertical or in horizontal direction. Bar graphs are mainly used in comparison of values between different categories or to make a note of series of changes happening over time. Bar graphs are exceptional when measuring changes over long intervals of time.

Matplotlib API in Python has the `bar()` function that must be used while plotting bar plots. The State with maximum number of cases has been plotted with bar graph.

Matplotlib API in Python has the `bar()` function that must be used while plotting bar plots. The State with maximum number of cases has been plotted with bar graph.

From the above bar graph visualization, we can infer that Maharashtra has the maximum number of confirmed COVID-19 cases as of now. It’s peak value is nearly 7000 cases per day. It is an unusual case that no other state in India has even marked up to half of its total cases

Alternatively, the states with maximum number of deaths has been described in the below mentioned bar graph.

- **RELPLOT**

Relplot is similar to scatterplot, where relplot lets us to create subplots in a single feature. Analyzing, deaths occurred in India using relplot. This relplot is available under the Seaborn library.

A legend is a line denoting the values of the variables in the graph. `Legend()` function is available in the matplotlib library.
By separating the 5 states individually, i.e Maharashtra, Kerala, Delhi, Rajasthan and Gujarat. The data is firstly segregated and then applied to the legend().

Fig 8: LEGEND OF 5 SELECTED STATES IN INDIA
From the above graph, it is visible that these curves gradually keeps increasing i.e as the severity of the virus increasing day by day, the number of confirmed cases also increases in the same fashion. But as an exceptional case, the southern state, Kerala has contained the virus spread and the amount of cases has become tremendously low and had set them up as an example to other states in India
Whereas the situation in Maharashtra has gone out of hands. The curve keeps increasing drastically and there seems to be no hope or possibility to slow down. Thus, it would be helpful for Maharashtra to consider Kerala as a forerunner in containing the covid-19 outbreak and follow their steps taken for protection. Gujarat’s curve was under control for the first 40 days and after that, unfortunately there was a steep increase in the number of confirmed cases.

B) ANALYZING USING THE POPULATION IN INDIA

The data mentioned below has the population count of the number of people in an individual state. The preliminary details were taken from the Aadhar (Identity proof in India) card from the official website controlled by the government of India. By using these figures, it is very efficient to calculate the estimate of the cases accordingly. By using this, Cases per 10 million is found out.

Fig 9: FIRST 5 COVID-19 CONFIRMED CASES/10 MILLION IN INDIA
Decreasing order of cases/10million is found out using the below mentioned code.

Fig 10: FULL DESCRIPTION OF COVID-19 CONFIRMED CASES/10 MILLION WITH THEIR POPULATION IN INDIA

IV. CONCLUSION
The coronavirus seems to increase in an unpredictable manner and is highly uncontrollable. The ILO’s four pillar policy framework has passed the circular in regard to the virus outbreak, about the precautionary measures to be followed and possible ways to cure the affected patients. In order to manage and predict the data, we must arrange them appropriately, thus its visualization becomes efficient and error-free. As data analytics plays a vital role in controlling the spread of the virus by intimating the areas as red zones and green zones.
respectively. Analyzing its death rate and recovery rate. As a whole it is recommended to conduct these analysis to get a clear picture of day-to-day statistics. It is believed these encoded information from the raw sources play a remedial role amongst the Pandemic and helps us to lead a better life.

V. REFERENCE


