International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 04, April-2020

An Exploratory Data Analysis of COVID-19 in India

Sarvam Mittal
Data Science
IIIT-B
Banglore, India

Abstract - The number of COVID-19 cases in India is increasing at a rapid pace. The National and local authorities are having a hard time to create a pattern, analyze and forecast the spread of COVID-19 in India. The main aim of this paper is to draw a statistical model for better understanding of COVID-19 spread in India by thoroughly studying the reported cases in the country till 22 April 2020. An Exploratory Data Analysis (EDA) technique is being implemented to study and analyze the reported COVID-19 cases in India. The result of the analysis divulge the impact of COVID-19 in India on daily and weekly manner, analogize India with abutting countries as well as with the countries who are badly affected and arrangement of India's Healthcare sector for such epidemic.

Keywords— COVID-19, exploratory data analysis technique, India's analysis, abutting countries analysis, healthcare sector analysis

I. INTRODUCTION

COVID-19 is a contagion belongs to the “Nidovirus family”, or “Nidovirales” which includes “Coronaviridae”, “Artieviridae” and “Roiniviridae” family, responsible for respiratory illness in humans which may cause common cold to more austere diseases such as “Middle East Respiratory Syndrome(MERS)” and “Severe Acute Respiratory Syndrome(SARS)”. The most common symptoms or traits of COVID-19 are fever, tiredness, dry cough, aches and pain, nasal congestion, runny nose or sore throat. The main thing to note here is that some people get infected and don't get these symptoms or traits and doesn't feel unwell. All age group people who has a medical history of blood pressure, cardiovascular disease or diabetes are more prone to get infected and if anyone with fever, cough and breathing difficulties should immediately seek for medical attention. COVID-19 is a “communicable” disease and can be passes through the droplets from nose or mouth when an infected person coughs or exhales and this is the main reason to maintain 1m (3 feet) distance from the sick person. Studies till date indicate that COVID-19 is mainly spread through contact rather than transmitted through air. As many people only experienced mild symptoms so it is a high probability to catch COVID-19 from the person who has mild cough or doesn't feel ill.

Protection from and prevention of spreading COVID -19 can be minimized by including some of the simple and easy to adopt precautions in daily habits which include thoroughly cleaning hands with alcohol based hand rub or washing them with soap and water, avoid touching eyes, nose and mouth as hands touches several surfaces which might be contaminated and hands could act as a carrier for COVID- 19 and virus can enter our body, stay home if you feel unwell and most importantly avoid traveling as much as possible. Follow National and local authorities only as they will have the most up to date information about the situation.

On 30 January 2020, India reported its first coronavirus case in Kerala when a student returned from Wuhan (epicenter of coronavirus) and till then the number of cases has been increasing exponentially. In recent times there is no vaccine or medicine available particularly for treatment of COVID-19 and currently are under investigation. This paper analyzes the current trend of COVID-19 based on certain criterion using “Exploratory Data Analysis”. Exploratory Data Analysis (EDA) is the way to explore the data with the aim of extracting useful and actionable information from it. EDA is the revelatory step in any kind of analysis.

II. LITERATURE SURVEY

In [1] the researchers analyzed the transmission trend of COVID-19 from China to other countries, confirmed cases on daily basis, surveillance strategy of China, South Korea, Japan, Italy , Iran and Spain from the first day of outbreak along with the effect of government policies of the above countries in controlling the COVID-19 outbreak by finding the linear relation between outbreak condition and “case fatality rate(CFR)” by taking global daily statistics such as confirmed, death and recovered cases and making prediction with respect to China using “Linear Regression”.

In [2] authors describes the research performed in the field of “coronavirology” with the overview of coronavirus replication and pathogenesis along with the evolution of coronavirus, the organ cultures and cells preparation, as well as techniques for analyzing the virus function, commonly used reverse genetic techniques of coronavirus and virus cell fusion as well as titration techniques, identification of cellular receptors and virus cell fusion along with visualization of virus replication complexes and covers the “coronavirus life cycle” in great detail.

In paper [3] researchers study and analyze the COVID-19 virus spreading statistics from the cases of different countries using “Bailey's model”. High correlation coefficients (91.4%) were resulted using Pearson correlation method and determinants (83.98) were also considered for the correctness of the model. “World Health Organization's” daily report were considered were considered for analyses of 204 countries and also indicates the difficulties in correctly predicted the future spread-reduction variable of the pandemic.

In [4] authors review the “pathogen”, “clinical features”, “diagnosis” and “treatment” of COVID-19 and also explain the “epidemiology” and “pathology” based on current evidence and recommend that symptoms, exposure history and manifestation on chest CT imaging could be consider as a clinical diagnosis in the COVID-19 affected areas where there is shortage of “RT-PCR” testing kits. Further the crucial role of “S-protein” is also depicted for COVID-19 as S-protein mediates receptor binding and membrane fusion which is crucial for transmission is also explained along with the suggestion that transmission mode is human to human and majorly it gets transmitted through droplets and close contact. The detailed analysis of 14 days “quarantine period” is also clearly explained as 95% of people experience symptoms within 12.5 days of contact.

The paper [5] investigates the influence of air temperature and relative humidity on the transmission of COVID-19 by calculating the “effective reproductive number”(R) and under “Linear Regression” framework they found out that one-degree Celsius rise in temperature and one percent increase in the relative humidity lower R by 0.0225 and 0.0158 respectively and provides an indication that arrival of summer and rainy season in the northern hemisphere can effectively reduce the transmission of COVID-19.

III. COVID-19 INDIA’s DATA ANALYSIS

Current COVID-19 outbreak motivates to do an EDA on the datasets, scraped from different sources such as “Ministry of Health and family Welfare” [6], “COVID-19 India website” [7], “John Hopkins GitHub repository” [8], “Worldometer” [9] and “Wikipedia” [10] using “Python” and thus analyzing the spread and trend of the COVID-19 in India and comparison with the neighboring and worst affected countries of the world. The dataset that uses EDA undergoes the process of normalization, choosing of essential columns using filtering, deriving new columns, and visualizing the data in the graphical format. This paper used “Python” for “data processing” and “web scrapping”, “pandas” library to process and extract information from the available dataset. Appropriate graphs were created for better visualization are the results of “Matplotlib” and “Seaborn” library of the Python.

A. COVID-19 Spread in India over time

In the Figure 1, the X axis represents the Dates on an interval of 15 Days from 22 Jan 2020 till 22 April 2020 and Y axis represents the number of cases (in thousands).Orange line shows “Confirmed cases” (positive cases), Red line shows “Deaths” representing the number of cases who had lost their lives, Green line shows “Recovered” which depicts the count of people who has recovered and Blue line shows “Active” cases, the difference of Deaths and Recovered from Confirmed cases.

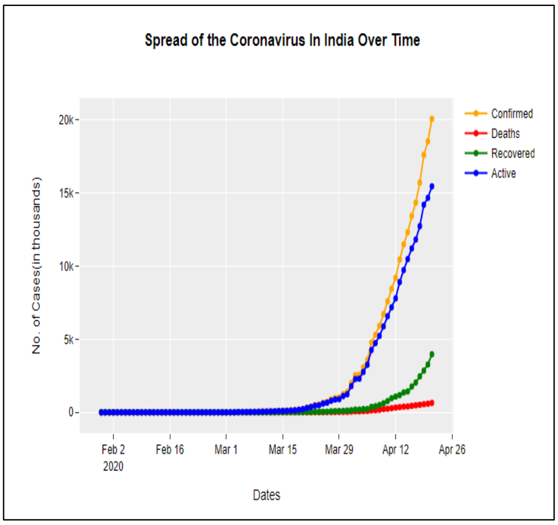


Fig. 1. Spread of COVID-19 in India over time

Inference from the Figure 1 is as follows:

- From last week of January, COVID-19 cases started to pop up in India, till March 15 the number of cases had been increasing on a constant level but after that it started increasing significantly and today on April 19 it have risen exponentially.
- Approximately 17 thousand “confirmed” cases has been registered till April 19 out of which 14 thousand are currently active.
- Out of 17 thousand confirmed cases round about 2800 cases have been recovered and 550 resulted in death.
- As on 22 April 2020, India has registered around 21 thousand cases out of which 700 has been resulted in death and around 4300 has been recovered leaving around 16 thousand as active and these numbers are increasing on daily basis.

B. COVID-19 Spread in India vs Other Countries

In the Figure 2, the X axis represents the Dates on 15 days interval and Y axis represents the number of cases (in lakhs). “Orange” line represents the cases in US, “Blue” in China, “Green” in Italy, “Red” in Spain, “Purple” in France and “Brown” in India.

Inference from the Figure 2 is as follows:

- All the countries started to detect COVID-19 cases from the January end except China which tells us that the epicenter of the virus is China.

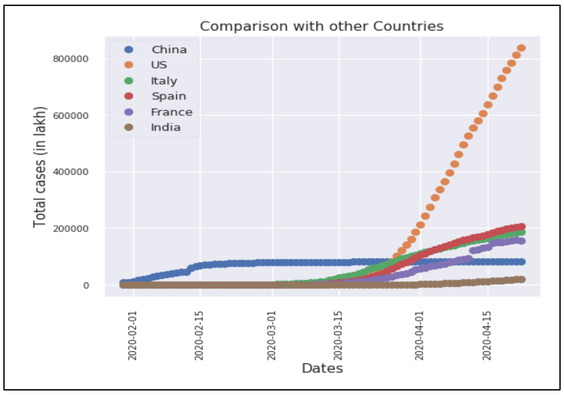


Fig. 2. COVID-19 Spread in India vs Other Countries

- China shows a growth in number of cases from January till mid Feb but after that the number of cases has been constant in the China.
- US turns out to be the worst affected country in world and till April 19 the number of cases has been touched 7 lakhs and it has increased to more than 8 lakh on 22 April 2020.
- Italy, Spain, France almost follow the same pattern and the number of cases has touched there approximately 2 lakhs till 22 April 2020.
- In India, the condition is better compared to other countries as the number of cases has touched only 22 thousand which is very less as compared to other countries till April 22 2020.

C. *Age-wise spread of COVID-19 in India*

Inference from the Figure 3 is as follows:

- The pie char analyzes the spread of COVID-19 in India to understand which age group is affected most.
- Age group 21-40 which is considered to be young age group is affected most i.e.42% as compared to elderly who are more prone to the diseases, only account 17% of the total cases in the country.

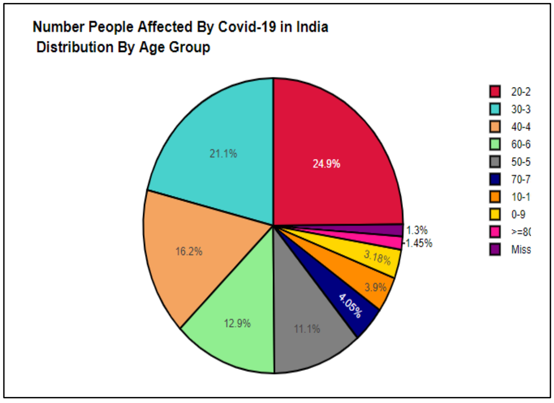
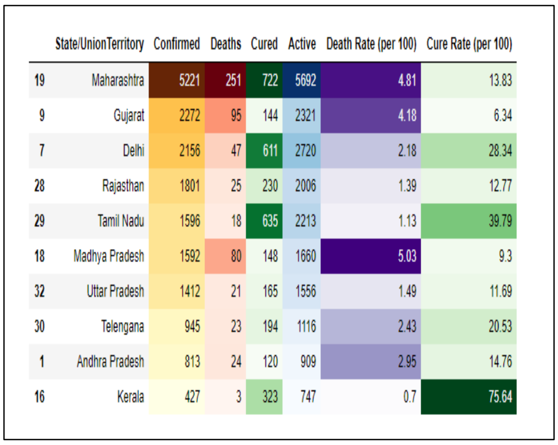


Fig. 3. Age-wise spread of COVID-19 in India

- The 9% are in between 0-20 years, 33% in 41-60 years and 17% in the people above 60 years while 1.3% are still unknown

D. *State-wise Analysis of COVID-19*

Kerala reported the first coronavirus case in India on January 30 when a student who had returned from Wuhan. Till February 3, two more students were tested positive after their return from Wuhan. Till then the spread of the COVID-19 in India has been on rampage. Fig. 4 depicts the top 10 states of India along with the statistics such as confirmed, deaths, cured, active, death rate(per 100) and cure rate(per 100). Inference from Fig. 4 can be drawn that Maharashtra tops the list by having 5221 confirmed cases, 251 deaths having 4.81 death rate (per 100) and 13.83 cure rate (per 100). After that Gujarat having 1893 confirmed cases followed by Delhi, Madhya Pradesh, Tamil Nadu, Rajasthan, Uttar Pradesh, Telengana, Andhra Pradesh and Kerala respectively.



	State/UnionTerritory	Confirmed	Deaths	Cured	Active	Death Rate (per 100)	Cure Rate (per 100)
19	Maharashtra	5221	251	722	5692	4.81	13.83
9	Gujarat	2272	95	144	2321	4.18	6.34
7	Delhi	2156	47	611	2720	2.18	28.34
28	Rajasthan	1801	25	230	2006	1.39	12.77
29	Tamil Nadu	1596	18	635	2213	1.13	39.79
18	Madhya Pradesh	1592	80	148	1660	5.03	9.3
32	Uttar Pradesh	1412	21	165	1556	1.49	11.69
30	Telengana	945	23	194	1116	2.43	20.53
1	Andhra Pradesh	813	24	120	909	2.95	14.76
16	Kerala	427	3	323	747	0.7	75.64

Fig.4 Sate-wise analysis of COVID-19 in India

E. *Symptoms observed for COVID-19*

The bar graph in Fig. 5 X-axis represents the percentages and Y-axis represents the name of symptoms which has been analyzed from the people who has been tested till now in India. This is the observation to keep the common symptoms checklist that has been created to keep a track if some new patients comes and can be helpful in classifying them as positive and negative

Inference from the Figure 5 is as follows:

- The most common symptoms of the COVID-19 is Fever as it has been reported by around 80% of the people ,is represented in black who has been tested positive followed by Dry cough by dark blue i.e. 70%,Fatigue near about 40%.

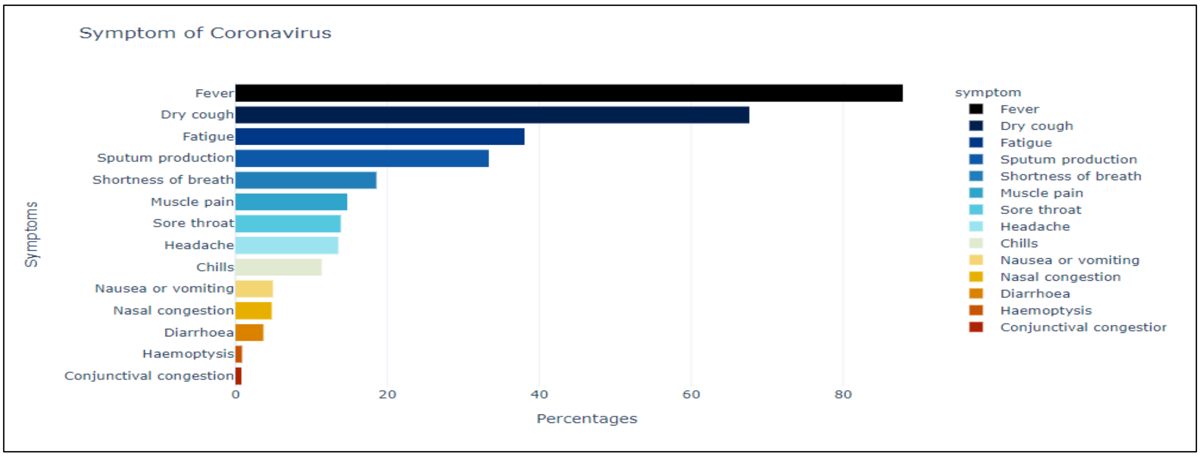


Fig.5. Sympotoms for COVID-19

- Sputum production has been reported by around 30% people, followed by shortness of breath which is reported by less than 20%
- Muscle pain, Sore throat, Headache and Chills has been observed in less than 15% of the people and Nausea, Diarrhea, Hemoptysis and Conjunctival congestion has been reported by less than 5% people.

F. *When Healthcare sector will exhaust in India?*

From Fig 6. It has been observed that if the infectious rate is maintained as it is in the India i.e.2.24 then by the end of August 2020 all the beds in India will be occupied by the COVID-19 patients and if we compare the infectious rate of India with countries like Spain, Italy and US then the convergence rate can be expected sooner than end of Aug 2020 also which is an alarming condition and government has to plan for the worst case scenario like this because as per the government records India has around 7 Lakh beds which can accommodate the patients but keeping is current scenario in place the situation can be much worse than our imagination

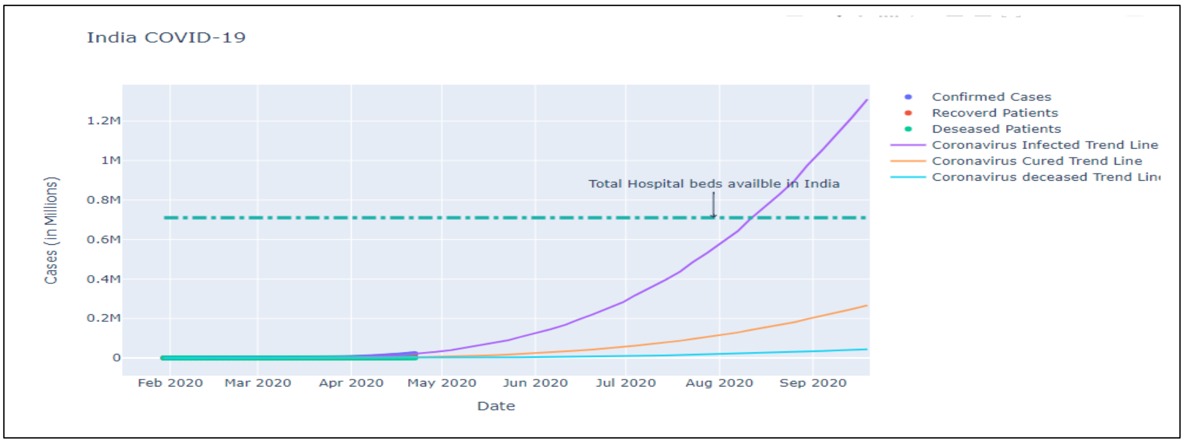


Fig.6. Sympotoms for COVID-19

G. *Future Prediction*

The below graph Fig. [7], has been obtained as output by passing the dataset we scraped which is a time series dataset through ARIMA model and it has been predicted that spread of COVID-19 and increase in the total number of cases in India will follow an exponential curve as obtain in Fig.[7] . The dataset we have used has records up to 22 April 2020, so the prediction graph is from 23 April 2020. The X-axis shows the dates on two days interval and Y-axis shows the total cases in thousand. It has been predicted that till 23 April 2020 approximately 25 thousand cases will be reported in India. If we compared it with the reported cases till 22 April 2020 i.e. 21 thousand then the predicted value is somewhat correct and in span of every 2 days till 7 May 2020 the number of reported cases will reach around 200 thousand in India which is a nightmare not only for India but for any county in the world.

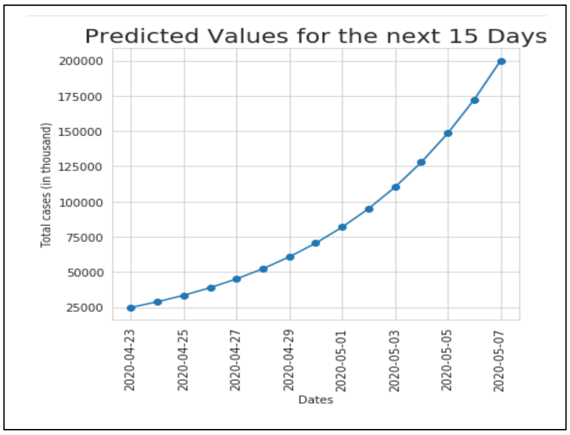


Fig.7. Future Prediction of COVID-19

IV. CONCLUSION and FUTURE WORK

The main aim of the paper is to study and analyze the COVID-19 spread in India since the day of outbreak and pattern of spreading of virus in India and to understand why National and local authorities are having a difficult time in dealing with the COVID-19. Moreover to study about the common symptoms of COVID-19 that are observed till now, age wise spread of COVID -19 to observe which age group is affected most, the spread of disease in India compared to other countries, the state wise trend of the epidemic to get detail understanding of how this is spreading and also to analyze the Healthcare sector of India and lastly to predict the future of epidemic in India.

This paper work can be extended to higher level in future, Predictive model for lasting of COVID-19 that uses Machine Learning algorithms, where the results from each graph of the paper can be taken as independent criteria for the machine learning algorithm. Moreover the Future Prediction analysis can be extended ad resulted in more accurate prediction as to predict more accurate number of total cases in India.

V. REFRENCES

[1] A. Hoseinpour Dehkordi, M. Alizadeh, P. Derakhshan, P. Babazadeh and A. Jahandideh,“Understanding epidemic data and statistics:A Case Study of COVID-19,” in press.

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

[3] 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)

[4] Y. Deng, F. He, W. Li, “Coronavirus disease 2019: What we know?,” J Med Virol,March 2020; doi:10.1002/jmv.25766

[5] W. Jingyuan, T. Wang, F.Kai and L. Weifeng, “High temperature and high humidity reduce the transmission of covid-19,” in press

[6] https://www.mohfw.gov.in/

[7] https://www.covid19india.org/

[8] https://github.com/CSSEGISandData/COVID-19

[9] https://www.worldometers.info/coronavirus/country/india/

[10] https://en.wikipedia.org/wiki/List_of_states_and_union_territories_of_India_by_population

[11] https://pib.gov.in/PressReleasePage.aspx?PRID1539877