iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** II **Month of publication:** February 2023

DOI: https://doi.org/10.22214/ijraset.2023.48991

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue II Feb 2023- Available at www.ijraset.com

Covid-19 for India: A Data Analytics Approach

Riddhi Chatterjee[1], Ritwik Mukherjee[2], Soumik Podder[3]
Department of Electronics and Communication Engineering, Guru Nanak Institute of Technology

Abstract: ***The COVID-19 epidemic has caused a large number of human losses and havoc in the economic, social, societal, and health system around the world. India also severely impacted throughout 34 states. First case reported on: 30 January 2020. As of 10th June Ministry of Health and Family Welfare reported a total of 276,804 cases, 134,843 recoveries and 7,751 deaths. In this project ,we aim to analyse COVID-19 data in India and find the key insights on patient demographics, patient clusters, state and district level spread. To be able to foresee the local transmission rate, top affected districts and predict the saturation point of the disease spread. In this work some datasets for each requirement were collected and organized those as per our requirements. All redundant data were erdaicated and made it useful for each requirement. The these data were grouped using Hadoop. After this query was written using Hive. To improve this planning Power BI was used for enhancing the visions on COVID-19 as per the said requirements .The collected data were used for state purpose to understand how COVID-19 spreads, the severity of disease it causes, how to treat it, and how to stop it and update the strategy based on evidence.***
Keywords: COVID-19 data set, Hive Tools, BI generator, Big data, Data Analytics, Hadoop, Distributed File System

I. INTRODUCTION

In the current pandemic situation where COVID-19 has spread across the world impacting 150+ countries, India too is severely impacted throughout 34 states. So, now analysing COVID -19 data for India using Big Data is become the most popular and important topic for understanding patient clusters, state and district level spread and find key insights on patient demographics and be able to foresee the local transmission rate, top affected districts and predict the saturation point of the disease spread. In this project the patient demographics were analysed to understand the spread and trend in daily cases treatment .As India is a vast country with a geographic area of 3,287,240 square of about 1.3 billion so, we consider state wise to Identifies the top 6 affected districts in India. Big Data is not only a Broad term but also a latest approach to analyze a complex and huge amount of data; there is no single accepted definition for Big Data. The challenge of Big Data is how to use it to create something that is value to the user. How to gather it, store it, process it and analyze it to turn the raw data information to support decision making.

Hadoop allows to store and process Big Data in a distributed environment across group of computers using simple programming models. It is intended to scale up starting with solitary machines and will be scaled to many machines. In this paper Hive tool is used. The primary goal of Hive is to provide answers about business functions, system performance, and user activity. To meet these needs strongly dumping the data into MYSQL data set, but now since huge amount of data in Terabytes which is injected into Hadoop Distributed File System files and processed by Hive Tool.

Power BI is a Data Visualization and Business Intelligence tool that converts data from different data sources to interactive dashboards and BI reports. In this project we also use power BI for data visualization. In the present work we are identifying the top 6 affected districts in India in order to improve the planning and strategy .Then we are identifying the patient demographics to understand the spread and trend in daily cases treatment. After that we will analyse the key insights for new policy formation like lockdown extension and also predict the saturation point for the spread. Customer can even get information about average time taken by patients to recover and average time a patient stays in hospital.Customer can also track load on healthcare facility and get the predicted number of ventilators required in future.

II. SIGNIFICANCE

Hadoop is a highly s calable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel so that thousands of terabytes of data and information regarding COVID-19 could easily be accessed. The motivation behind the use of Big data analytics derives innovative solutions and it helps in understanding and targeting customers. It helps in optimizing business processes and also helps in improving science and research. Here we Identifies patient demographics to understand the spread and trend in daily cases treatment so It improves healthcare and public health with availability of record of patients. Here the number of ventilators are predicted in order to satisfy the requirement in future, so using this information many hospitals can understand the number of mechanical ventilators they needed to have available.

It is a cost effective storage solution for exploring data sets. This provides reliable prediction of cases and performs even better when modified to include more states and parameters. Straight forward and simple to use.Every second additions are made. One platform carry unlimited information. Anyone can access vast information via surveys and deliver answer of any query.

III.METHODOLOGY

Hadoop is used for this data analytics. Hive is mainly used for structured data assuming all the Hadoop tools have been installed and having semi structured information on COVID-19 data. Power BI is used for COVID-19 data visualization.

Methodology used are as follows:

1) Create tables with required attributes
2) Extract semi structured data into table using the load a command
3) Analyse data for the following queries
a) List of top 6 affected districts in India Improves planning and strategy .
b) List of patient numbers to understand the spread and trend in daily cases treatment.
c) Average time taken by patients to recover.
d) The saturation point for the spread .

In the current pandemic situation where COVID-19 has spread across the world impacting 150+ countries, India too is severely impacted: 34 states / UT affected. First Case Reported on: 30 January 2020 as of 10th June, Ministry of Health and Family Welfare reported a total of 276,804 cases, 134,843 recoveries and 7,751 deaths the infection rate of COVID-19 in India is: 1.7 (significantly lower than in the worst affected countries.

TABLE I

Req_ID	Requirements
1	Data preparation: Loading into Hadoop Quality data preparation Identifying tool for BI
2	Identifies the top 6 affected districts in India Improves planning and strategy to provide
3	Identifies patient demographics to understand the spread and trend in daily cases treatment
4	Predicts the saturation point for the spread
5	Average time taken by patients to recover, average time a patient stays in hospital – hospital occupancy

IV.RESULT AND DISCUSSION

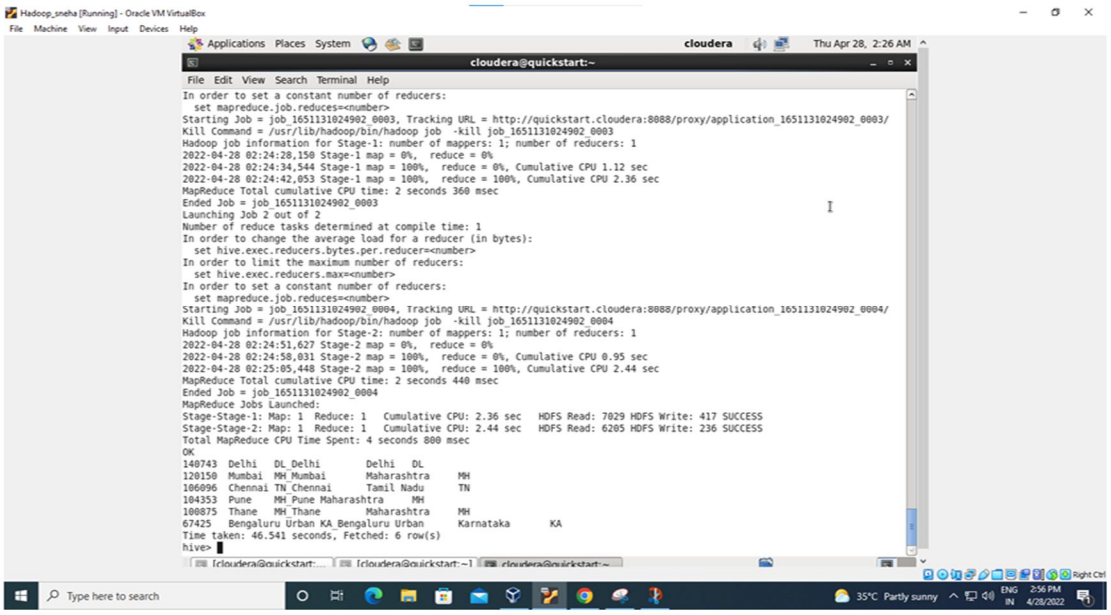


Fig 1: Identification of the top 6 affected districts in India Improves planning and strategy

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue II Feb 2023- Available at www.ijraset.com

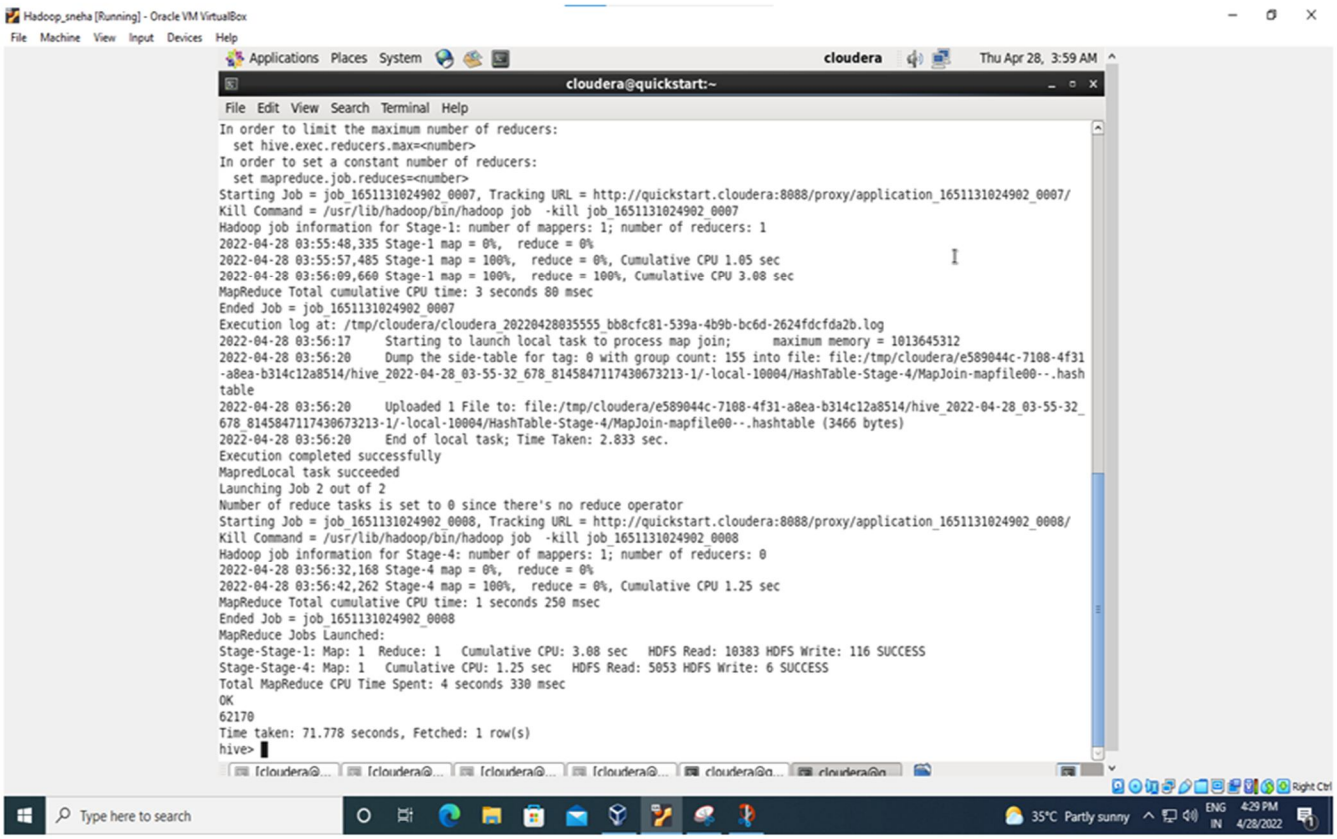


Fig 2: Identification of patient demographics to understand the spread and trend in daily cases treatment

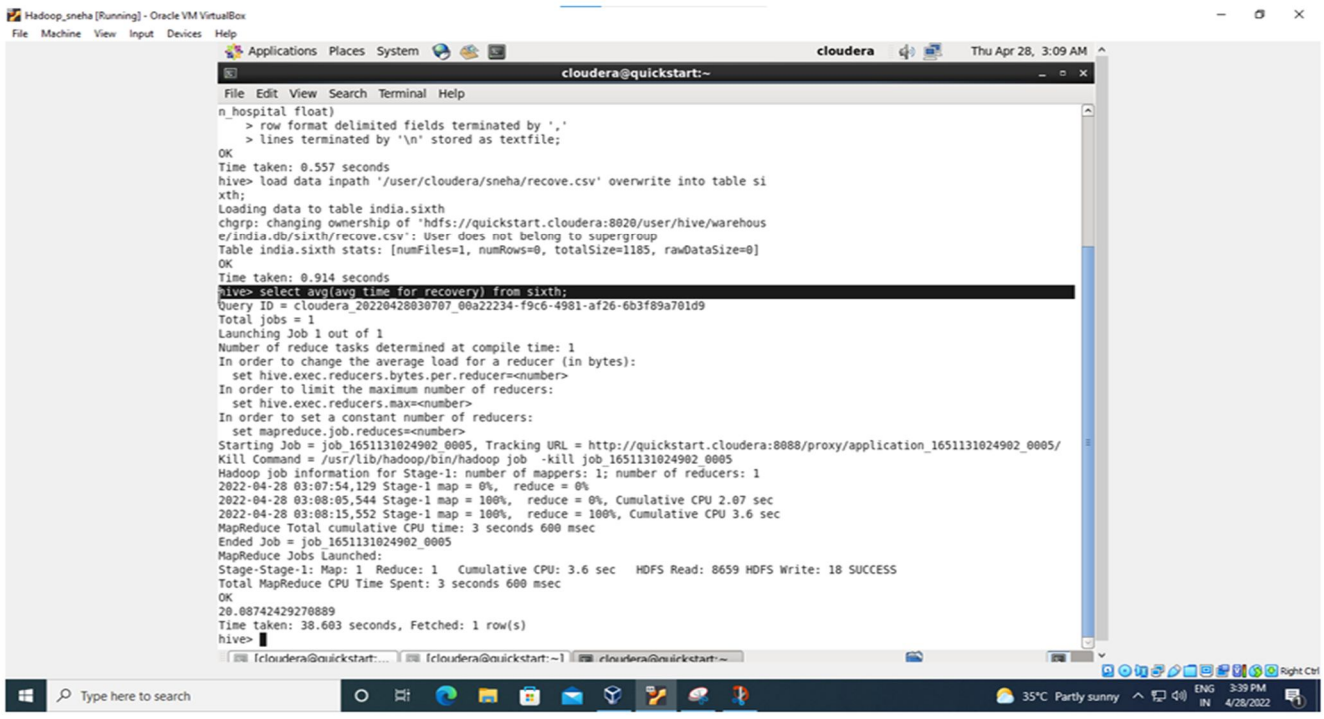


Fig 3: Identification of the Average time taken by patients to recover

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue II Feb 2023- Available at www.ijraset.com

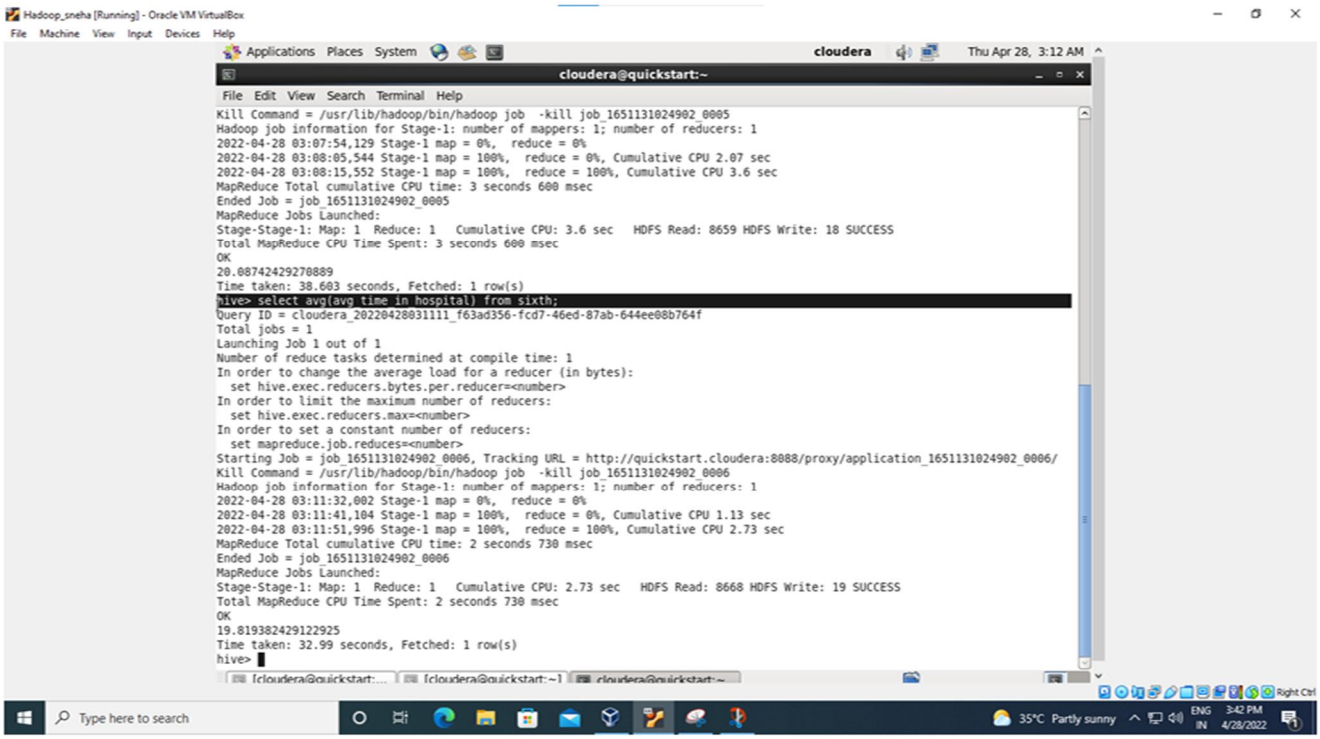
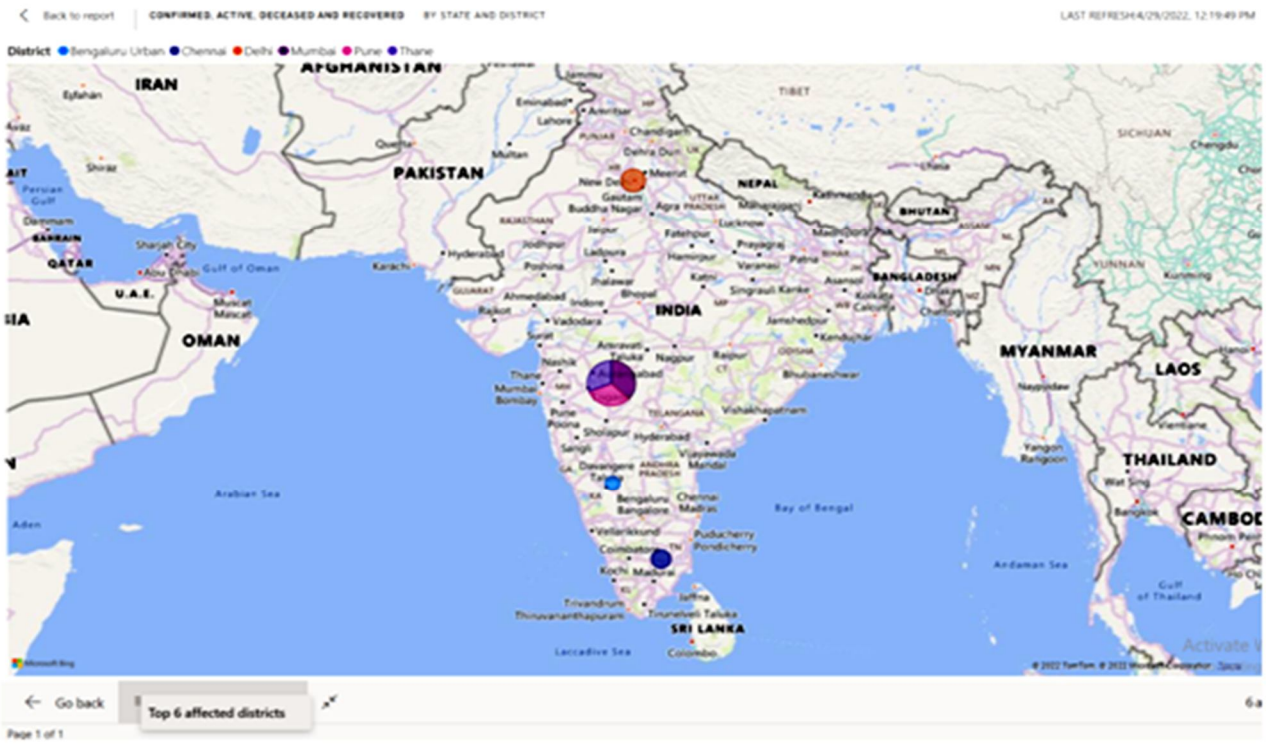


Fig 4: Identification of the average time a patient stays in hospital

A. Visualization of COVID-19 in India

1) Requirement - 2

Identification of patient demographics to understand the spread and trend in daily cases treatment

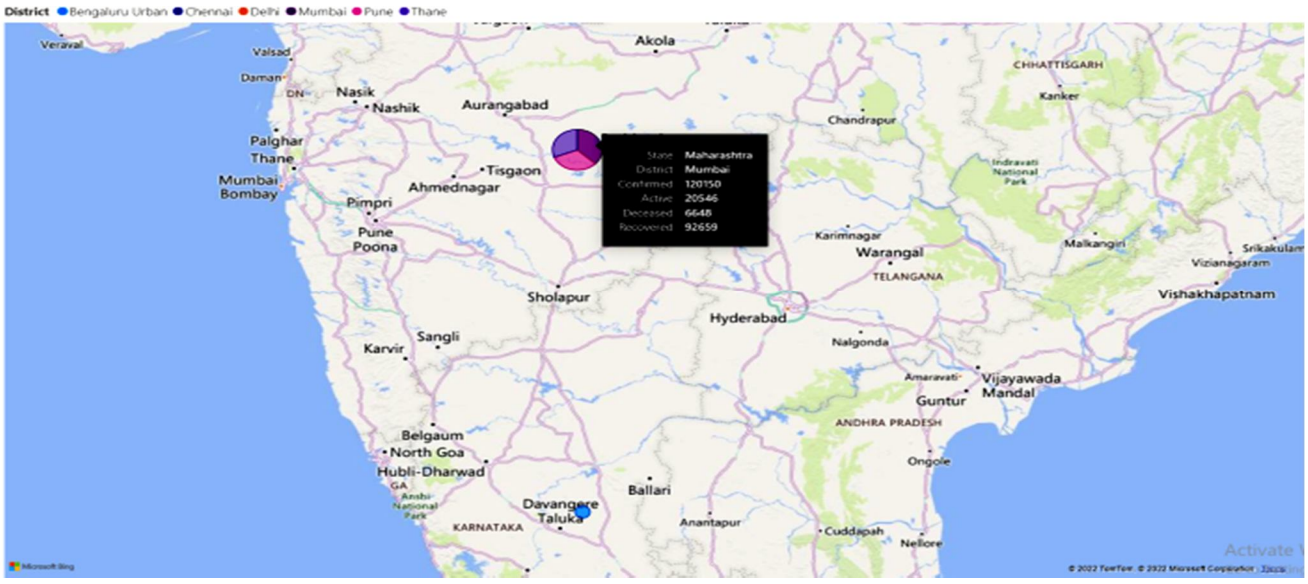


A Power BI Dataset can work as a collection of data for use in Power BI reports, and can either be connected to or imported into a Power BI Report. A Dataset can be connected to and get their source data through one or more Data flows.
The tool we used to develop this visualisation is power bi. To identify the top 6 districts we used map view as a visual from the entire state data we got this total 6 districts as most increased cases.

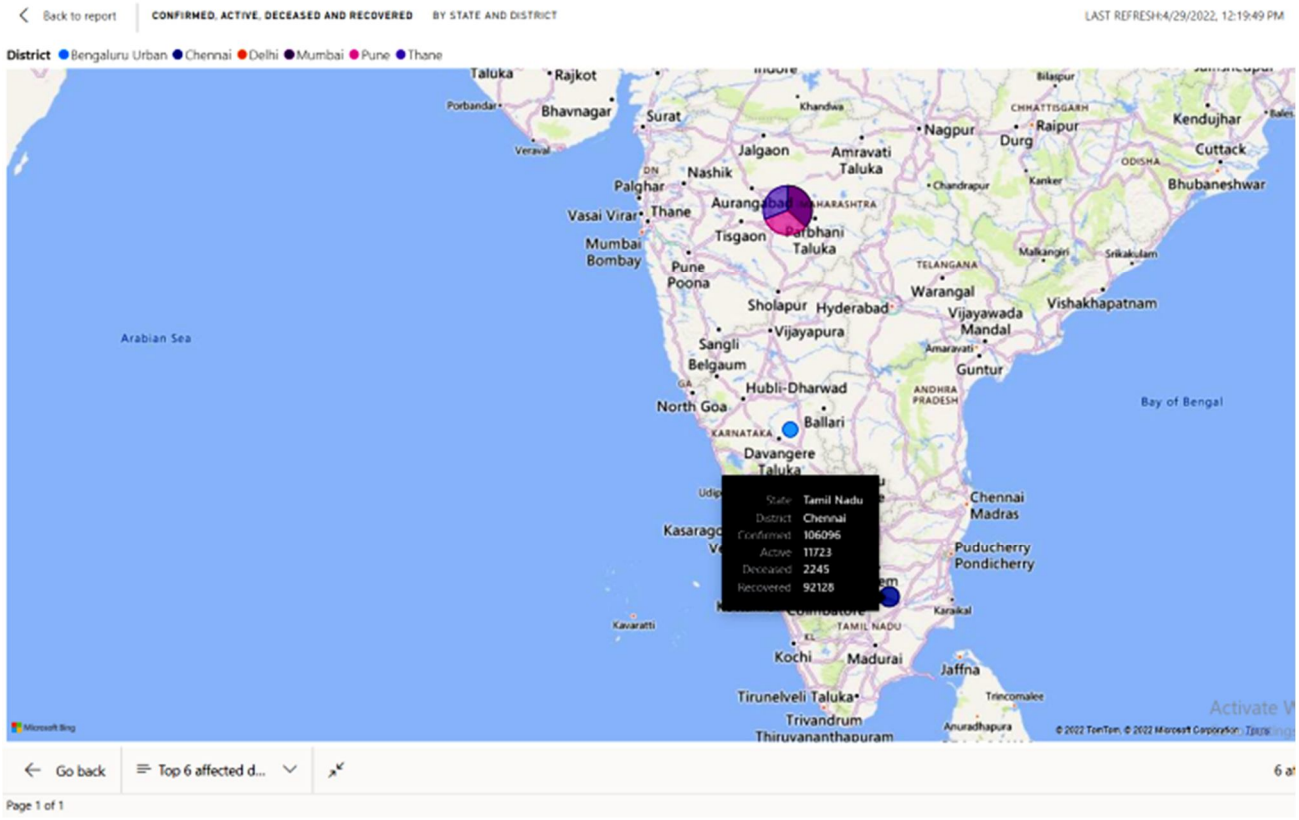
From the top 6 lists

Delhi has cases more cases compare to the other districts
Total cases : 140743
Present active cases are 9561
Decreased cases are 4058
Total recovered cases are 127124
Delhi in district made up 22.00% of confirmed cases out of 6 districts.

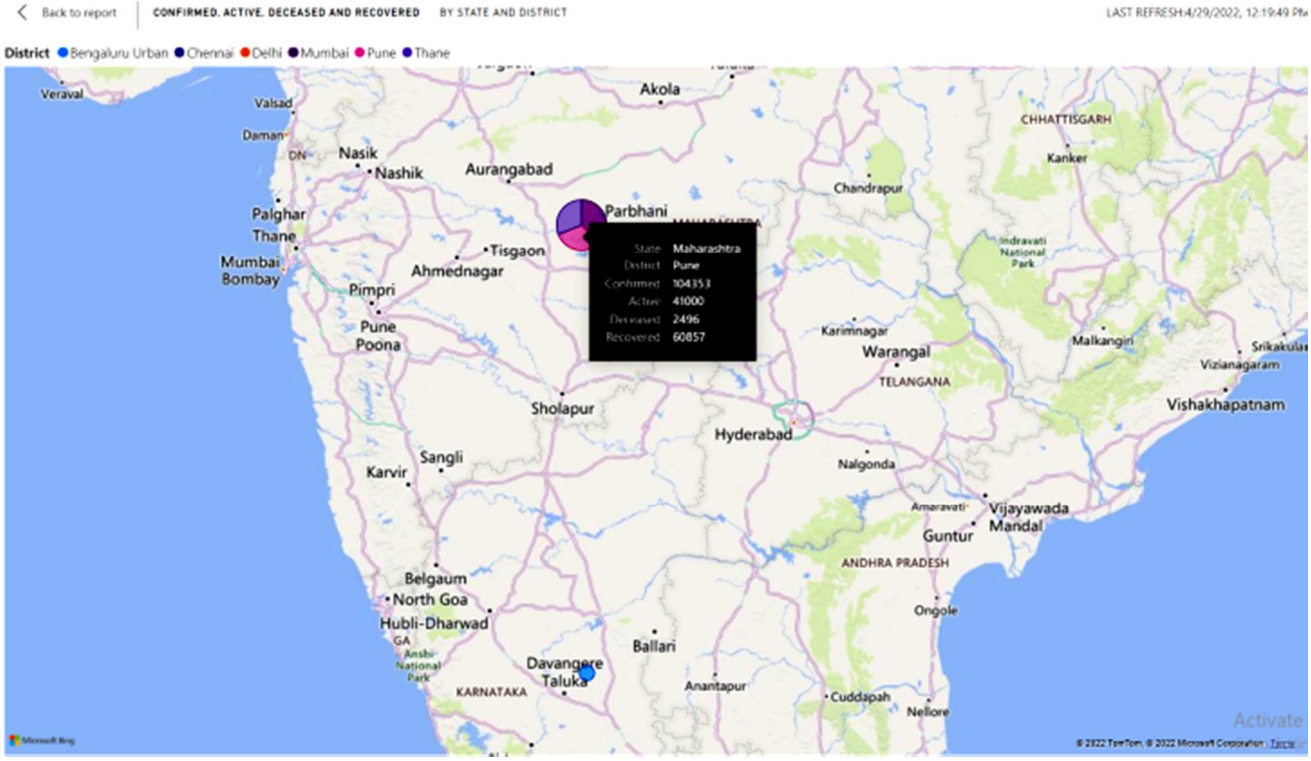
2nd highest cases are in Maharashtra state in Mumbai district with the confirmed cases of 120150. Total active are 20546, decreased 6648, recovered 92659 Mumbai in district made up 18.78% to display this visual we used map .



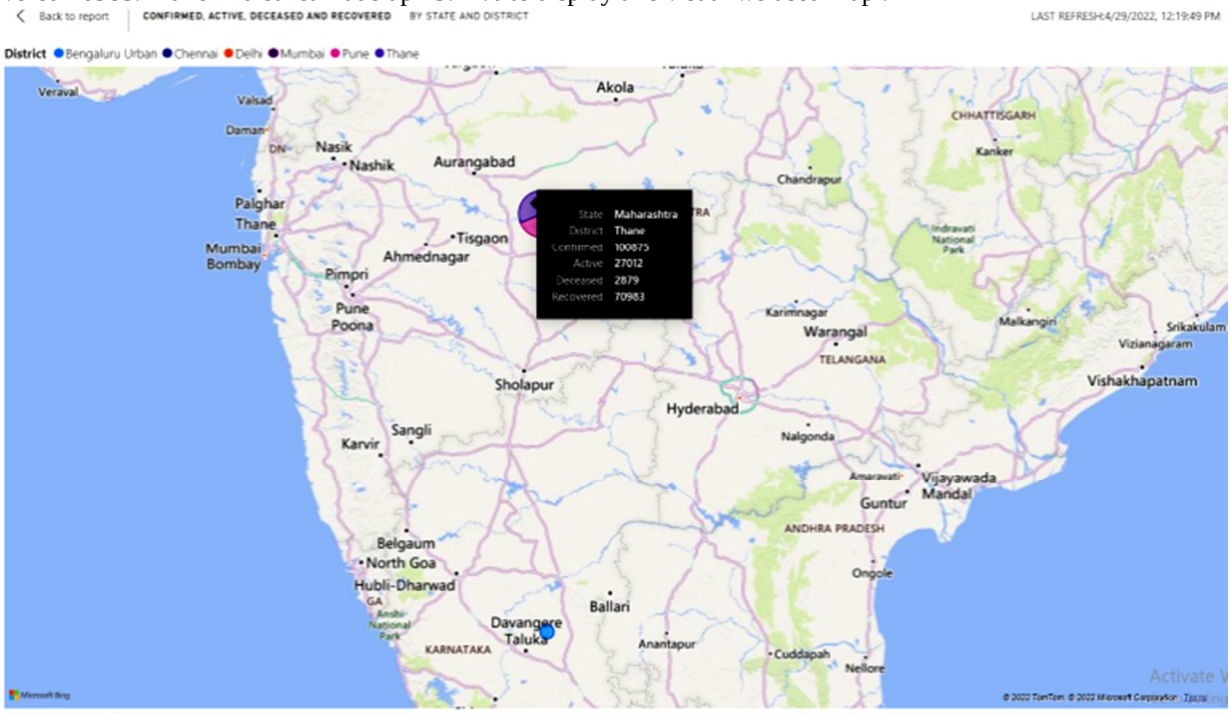
3rd highest cases are in Tamil nadu state in Chennai district with the confirmed cases of 106096. Total active are 11723 , decreased 2245, recovered 92128 .Chennai in district made up 16.59% to display this visual we used map.



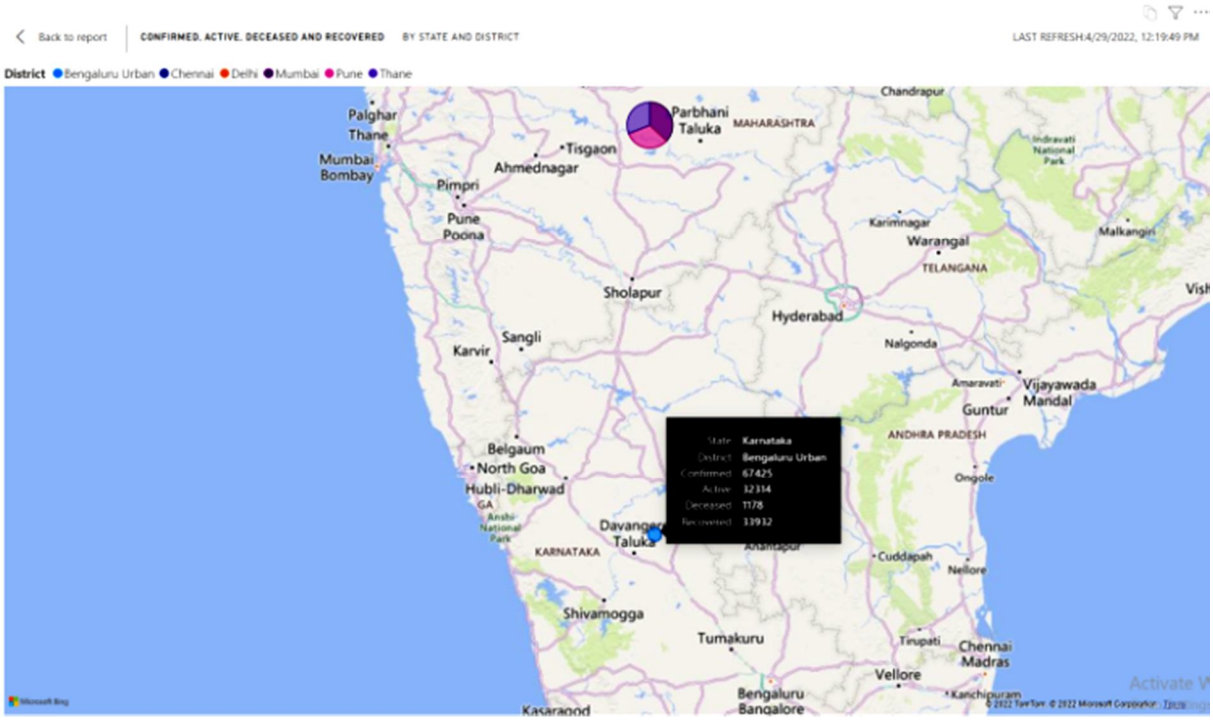
4th highest cases are in Maharashtra state in Pune district with the confirmed cases of 104353.total active are 41000 , decreased 2496, recovered 60857 .Pune in district made up 16.31% to display this visual we used map .



5th highest cases are in Maharashtra state in Thane district with the confirmed cases of 100875.total active are 27012, decreased 2879, recovered 70983.Thane in district made up 15.77% to display this visual we used map .



6th highest cases are in Karnataka state in bengaluru urban district with the confirmed cases of 67425.total active are 32314, decreased 1178, recovered 70983.bengaluru urban in district made up 15.77% to display this visual we used map .



The total cases in these districts are 639642, active 142156, decreased 19504 , recovered 477683.

To improve this planning the visions on COVID-19 were improved as per the said requirements. The collected data for state is employed to understand the top 6 affected districts. One of the main things we've learned is that the faster all cases are found, tested

and isolated, the harder we make it for this virus to spread. This principle will save lives and mitigate the economic impact of the pandemic. The strategy requires a whole-of-government and whole-of-society response. The resolve and sacrifice of frontline health workers must be matched by every individual and every political leader to put in place the measures to end the pandemic.

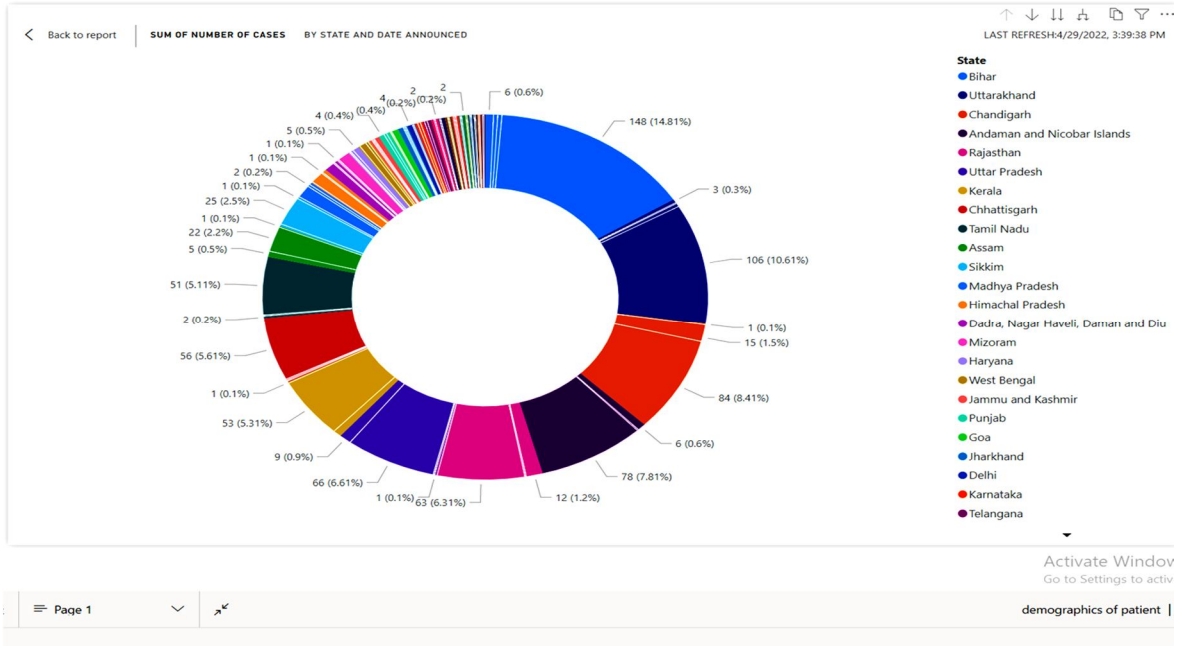
2) Requirement - 3

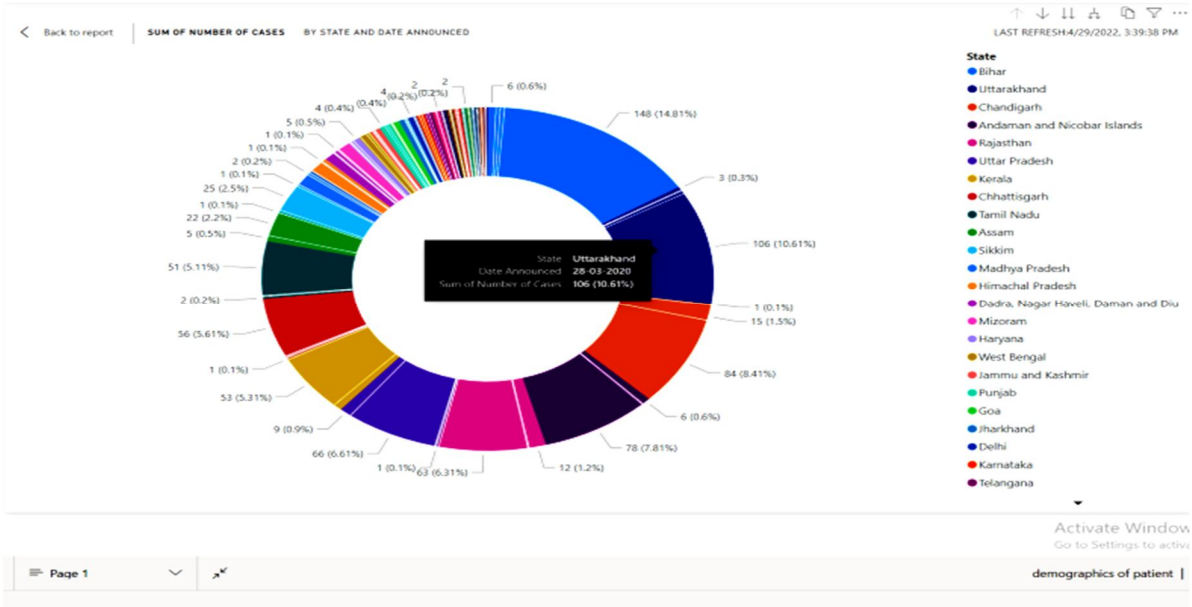
Identification of patient demographics to understand the spread and trend in daily cases treatment

This visualization was developed by using Donut chart for clear visual of states. To prevent the spread and trend in the daily cases treatment, we need to minimize social and economic impact through multisectoral partnerships. communicate critical risk and event information to all communities and counter misinformation.

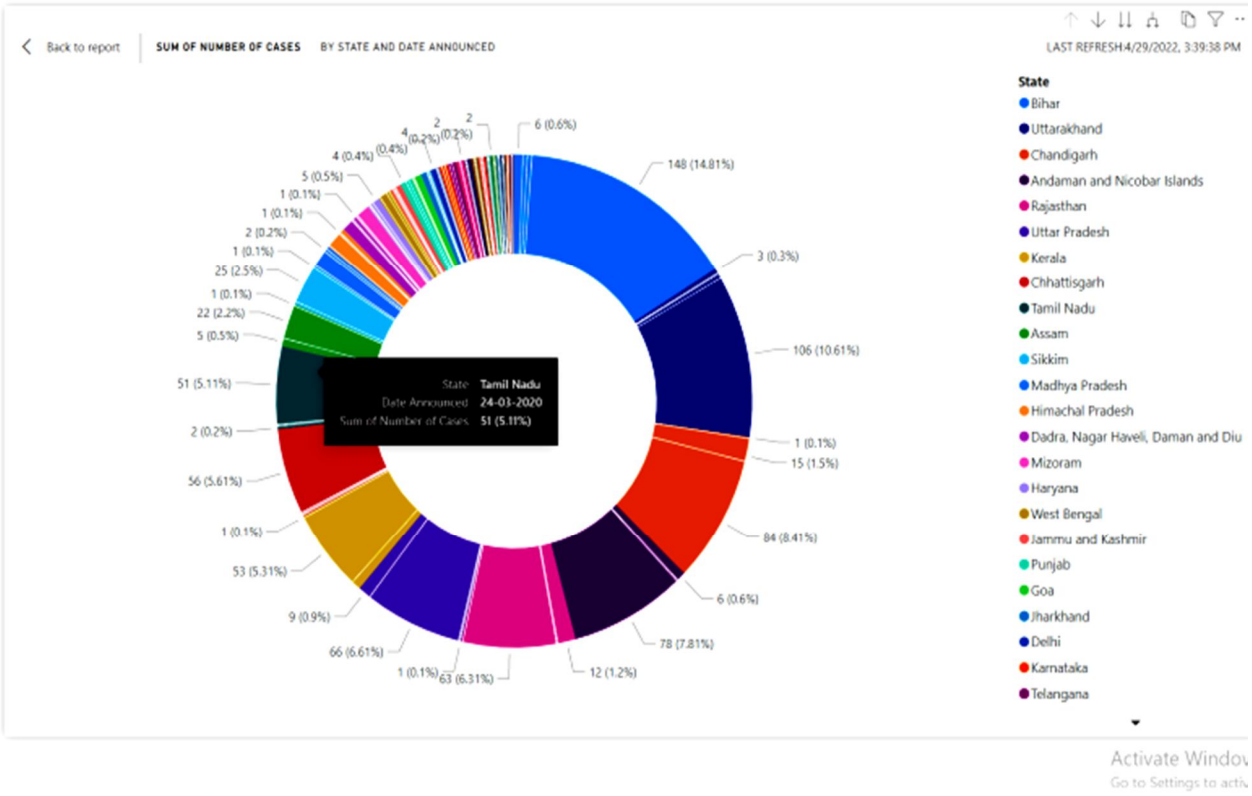
This can be achieved through a combination of public health measures, such as rapid identification, diagnosis and management of the cases, identification and follow up of the contacts, infection prevention and control in health care settings, implementation of health measures for travellers, awareness-raising in the population and risk communication.

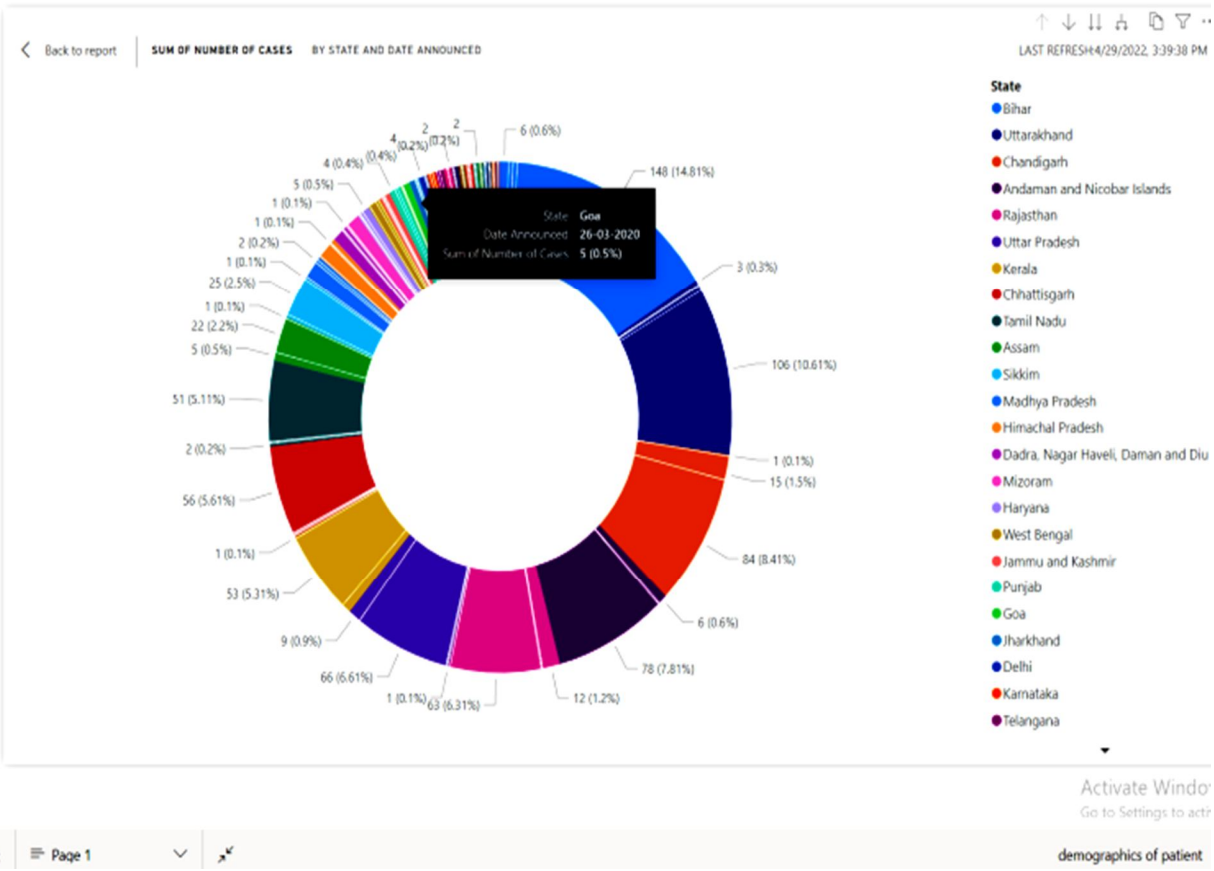
Here in the above picture it shows the data visualization which is made.

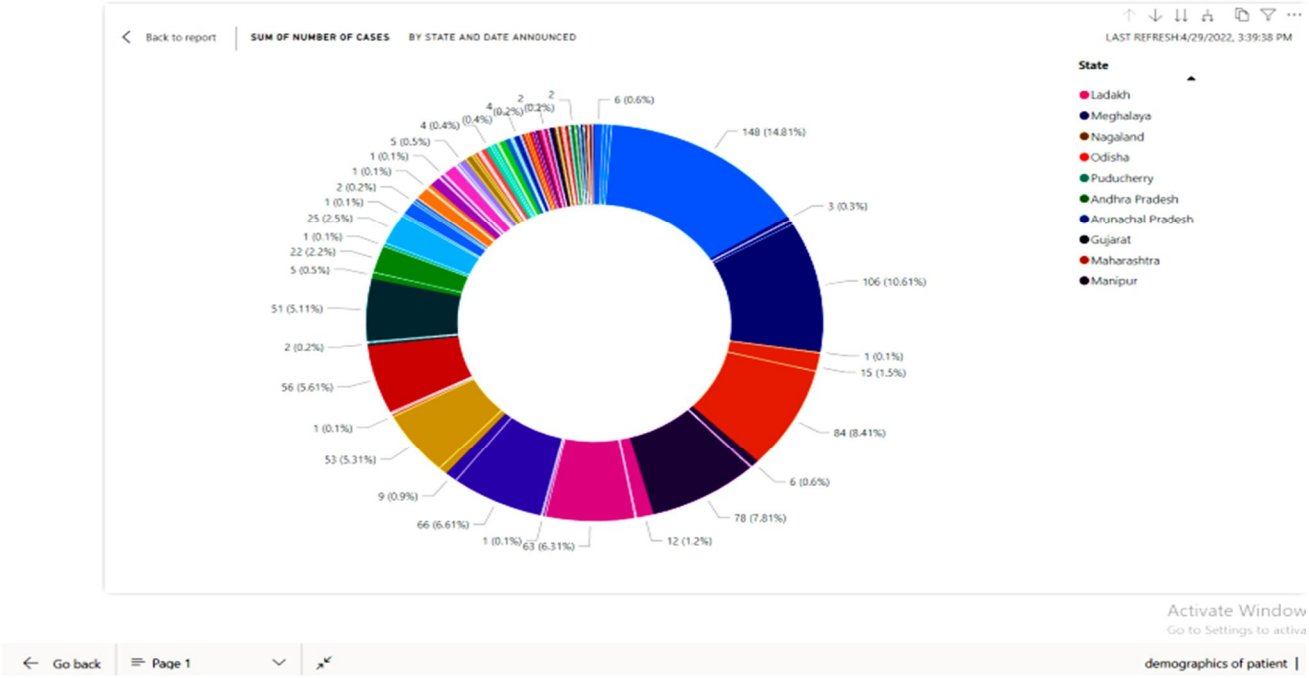




International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue II Feb 2023- Available at www.ijraset.com



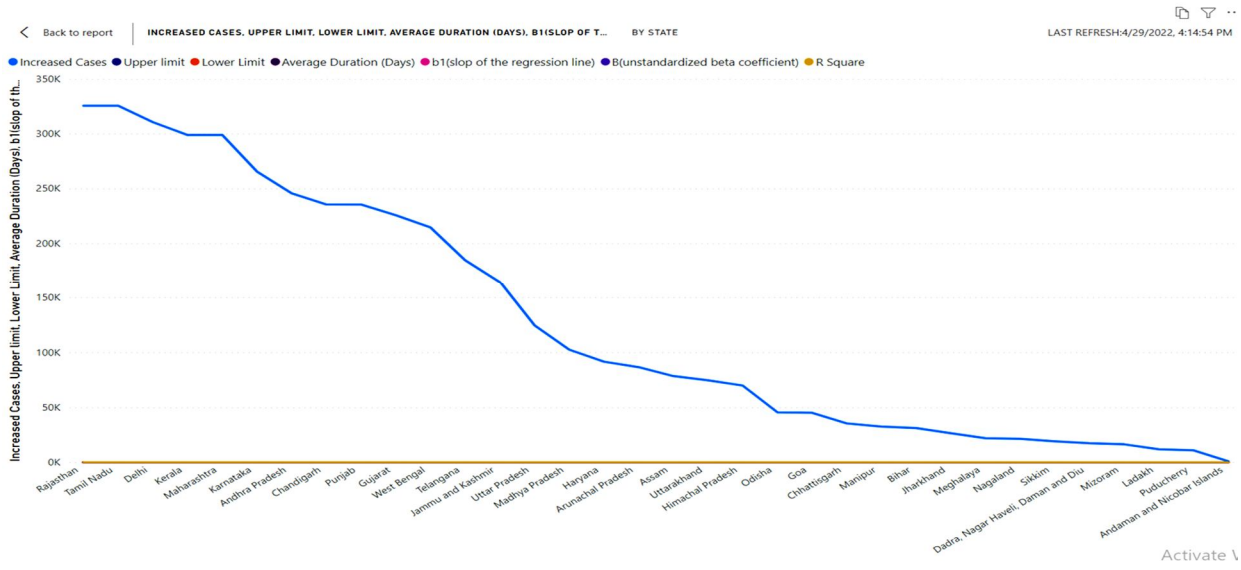




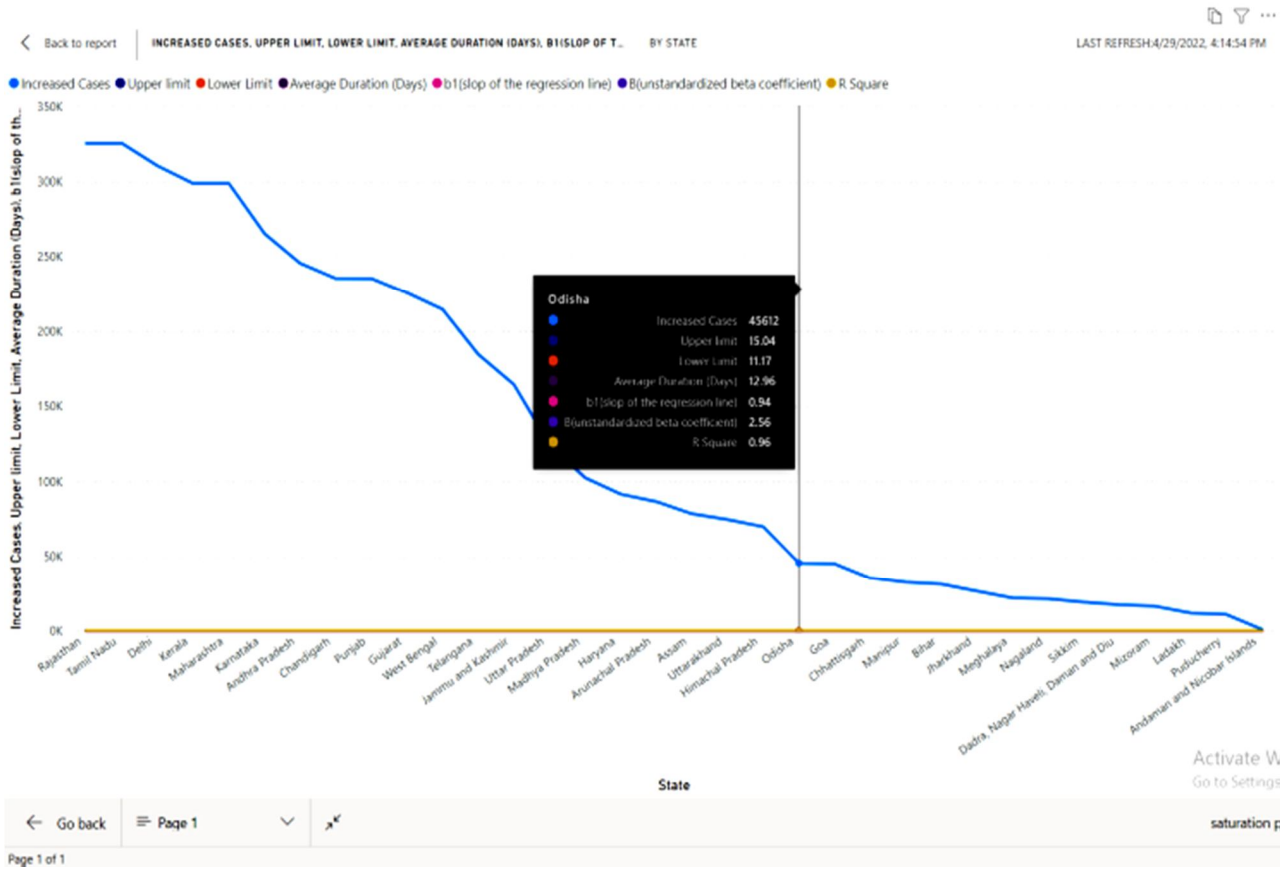
3) Requirement 4

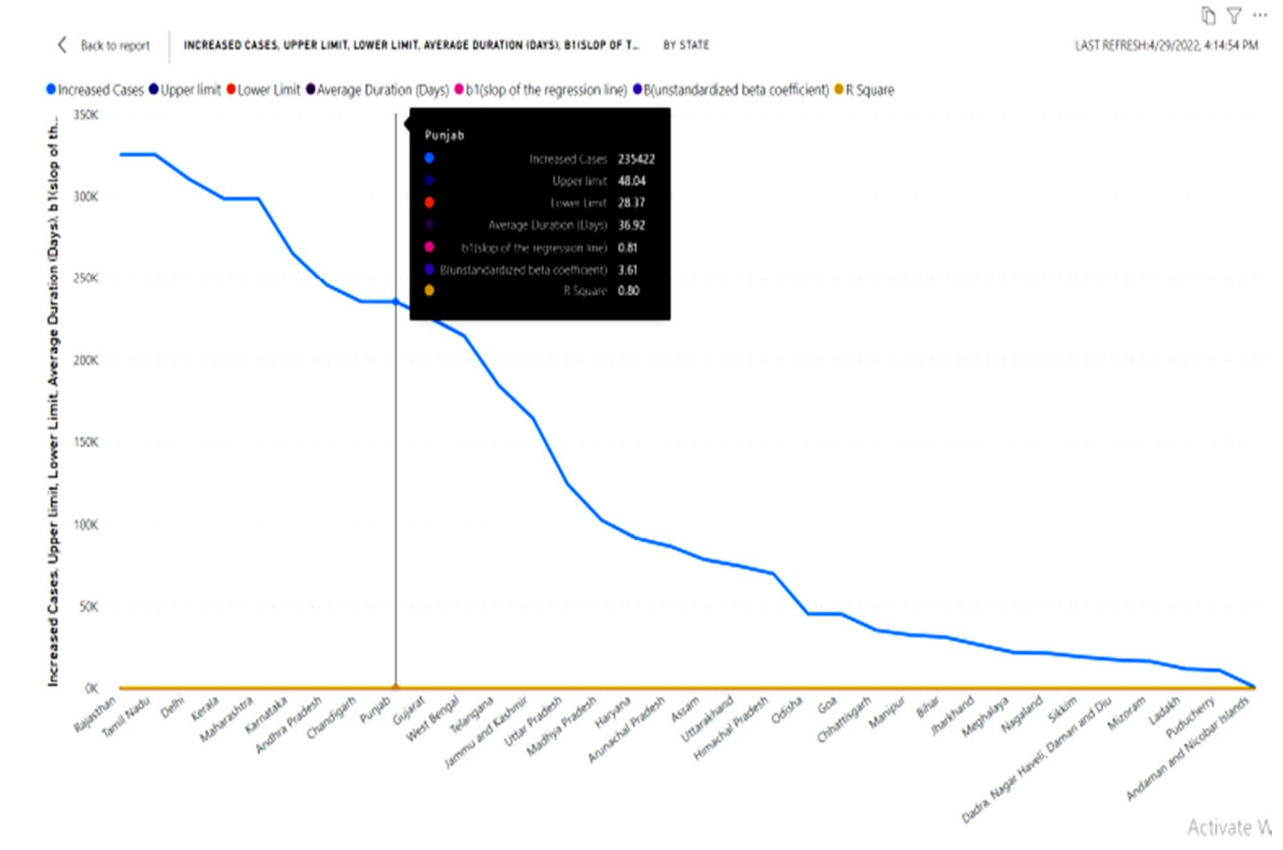
Predicts the saturation point for the spread

A line representing the predicted moduli resulting from the equations spread in the data points at particular degree of saturation .in this the regression analysis yields a predicted value for the criterion resulting from a linear combination of the predictors. To predict the saturation point we used line chart to display the visual of increased cases in the states, average duration of per patient to find out the saturation where can the cases are more by using the limits as a prediction . upper limit and lower limit is taken as a key to know the saturation point of spread in states as per cases and it shows for every state . The appearance and fast spreading of Covid-19 took the international community by surprise. Collaboration between researchers, public health workers, and politicians has been established to deal with the epidemic. One important contribution from researchers in epidemiology is the analysis of trends so that both the current state and short-term future trends can be carefully evaluated. This can be used for other researchers collaborating with and advising health institutions around the world during the Covid-19 outbreak or any other epidemic that follows the same pattern. We hope it may help facilitate policy decisions, the review of in-place confinement measures, and the development of new protocols.

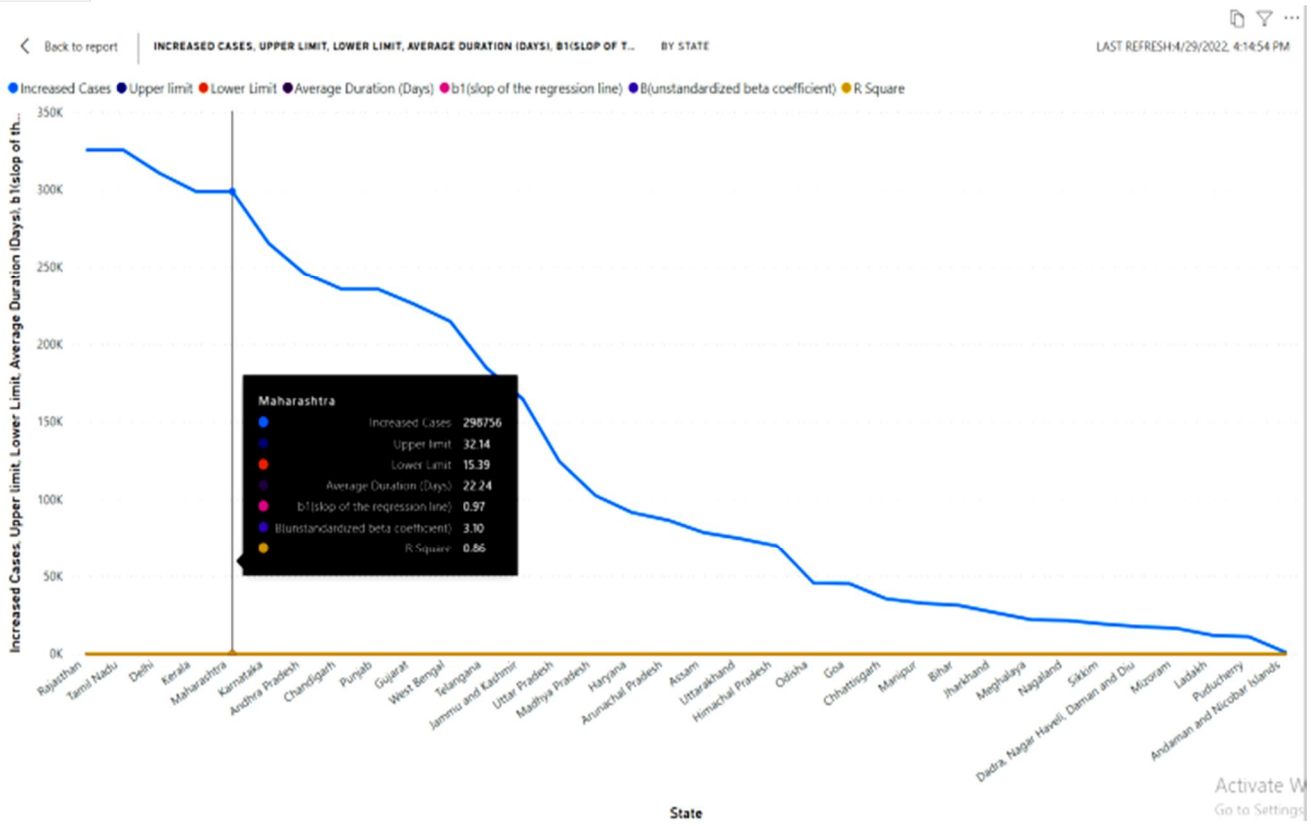


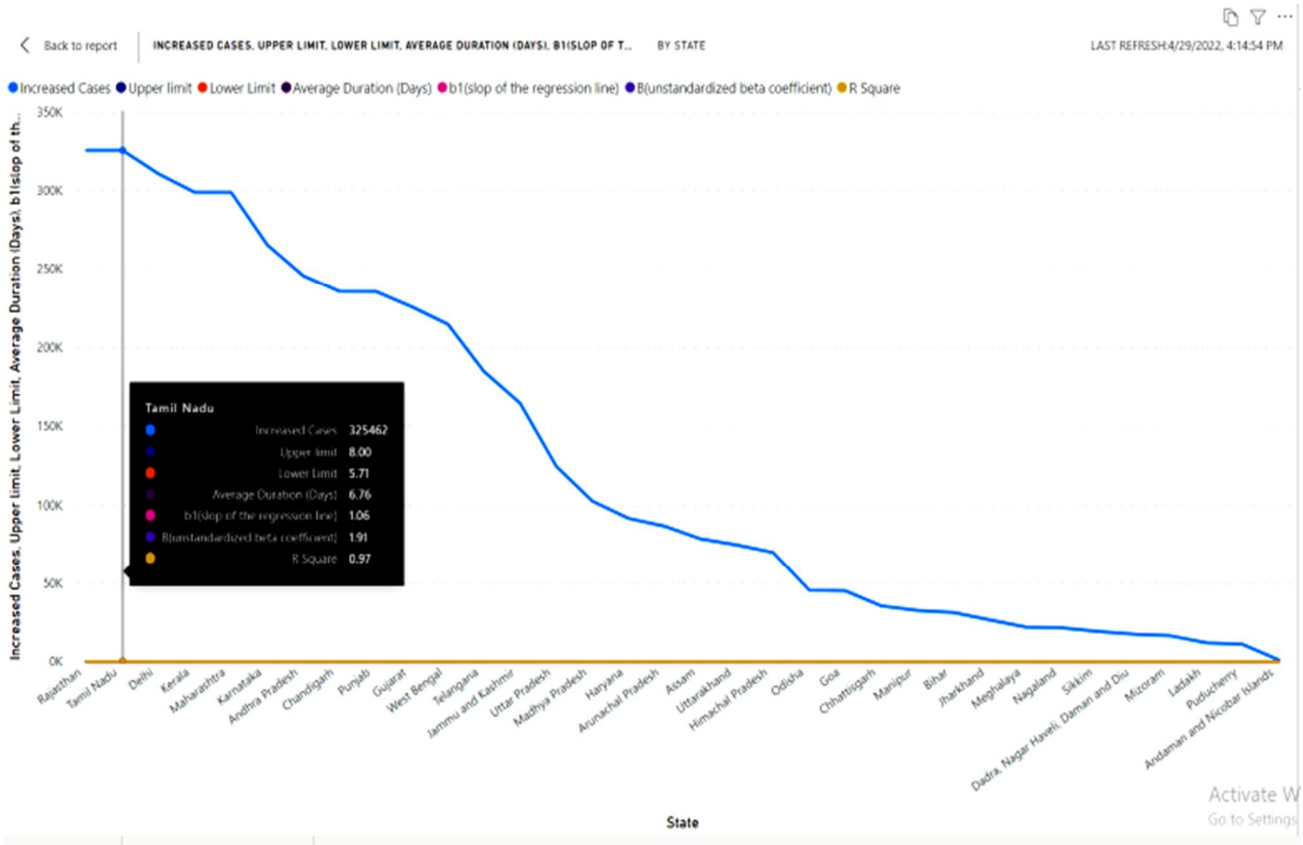
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue II Feb 2023- Available at www.ijraset.com





ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue II Feb 2023- Available at www.ijraset.com



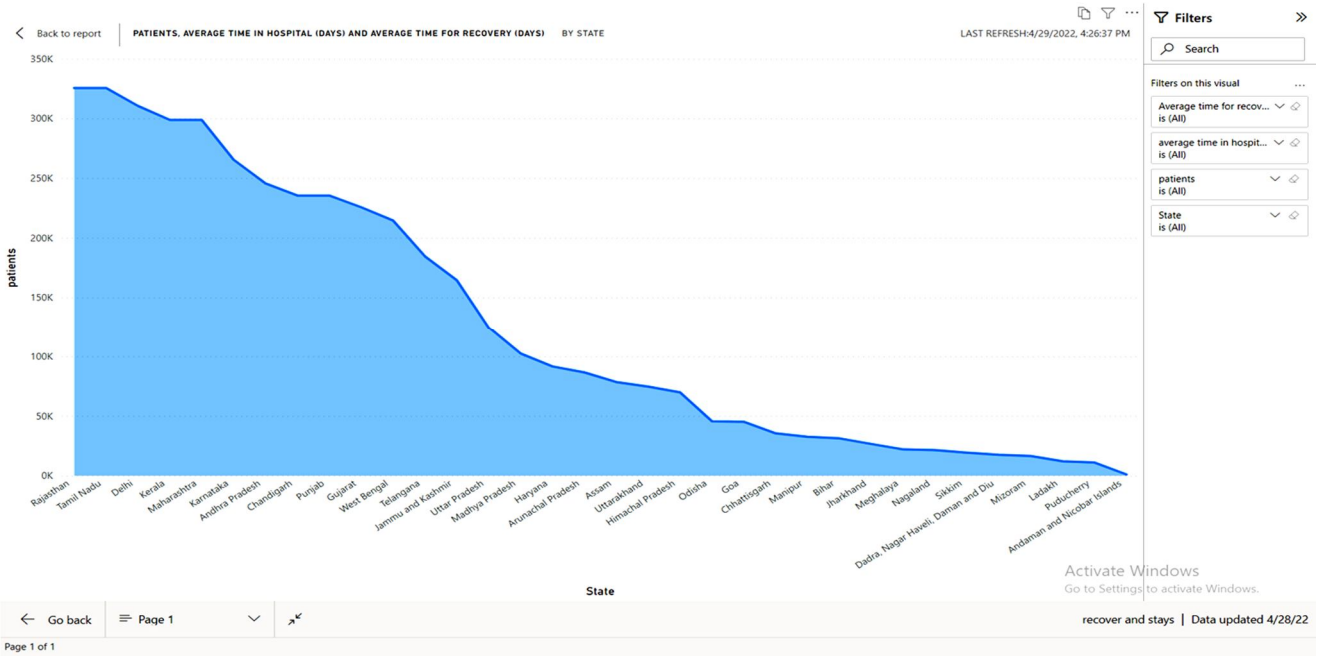


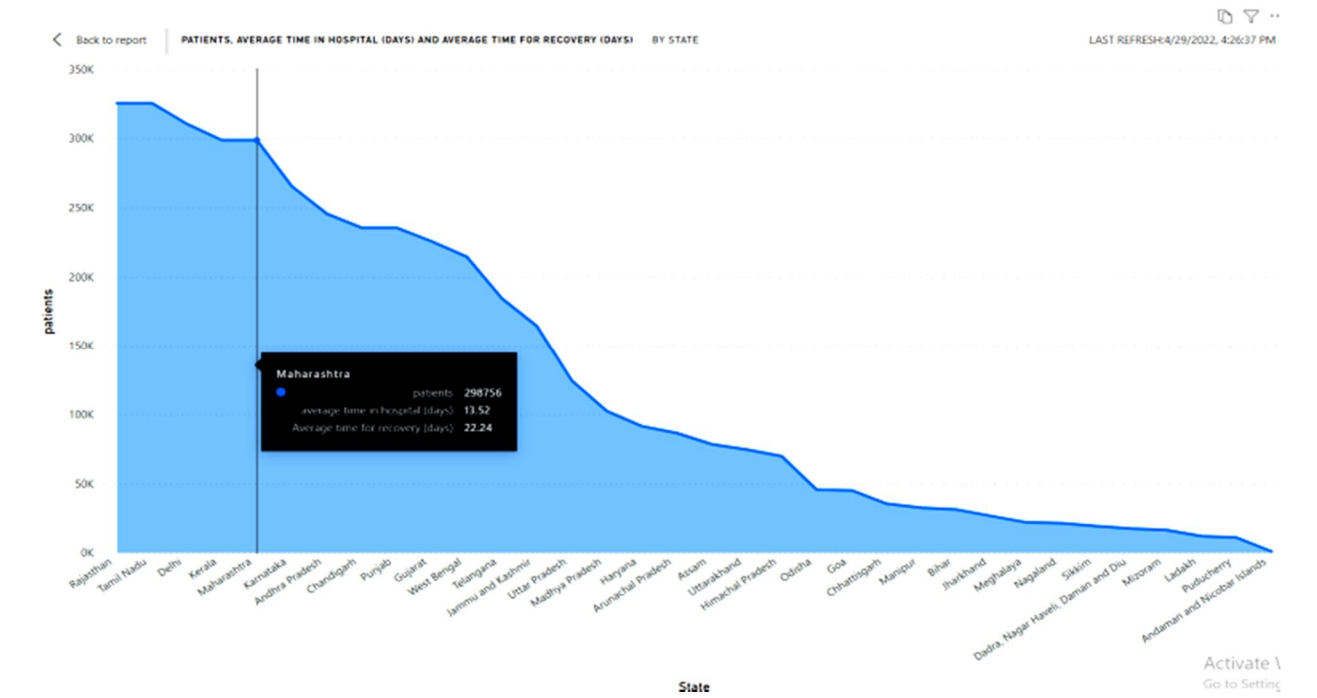
4) Requirement 5

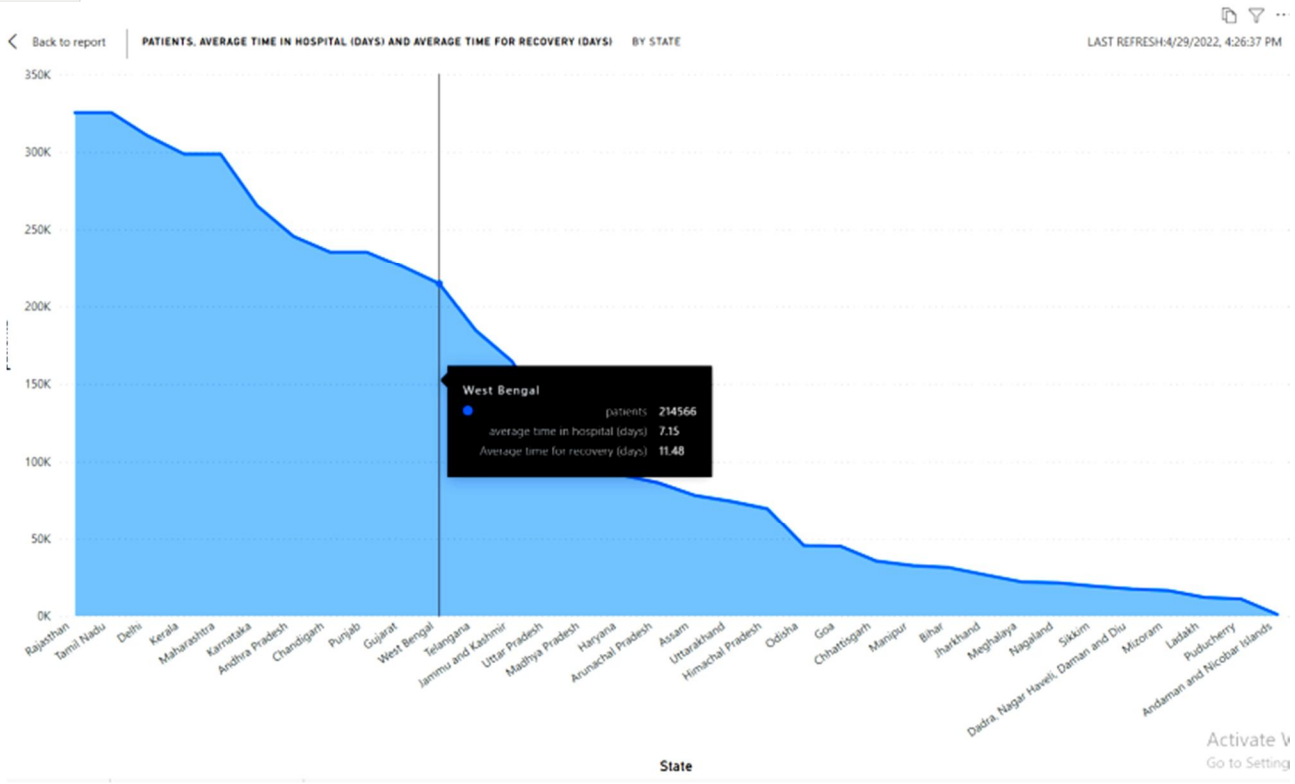
Average time taken by patients to recover, average time a patient stays in hospital-hospital occupancy.

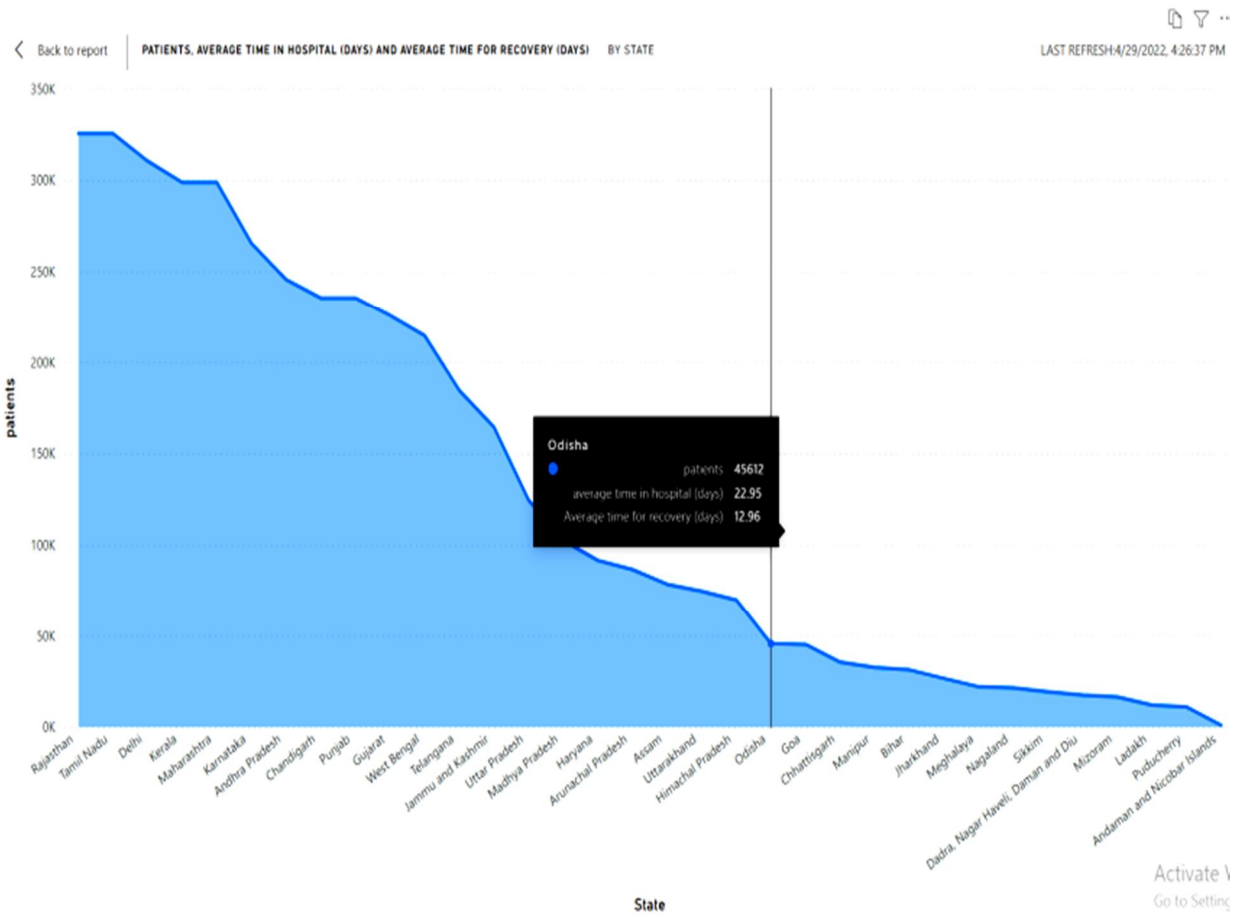
The COVID-19 pandemic has placed an unprecedented strain on health systems, with rapidly increasing demand for healthcare in hospitals and intensive care units (ICUs) worldwide. As the pandemic escalates, determining the resulting needs for healthcare resources (beds, staff, equipment) has become a key priority for many countries. Projecting future demand requires estimates of how long patients with COVID-19 need different levels of hospital care.

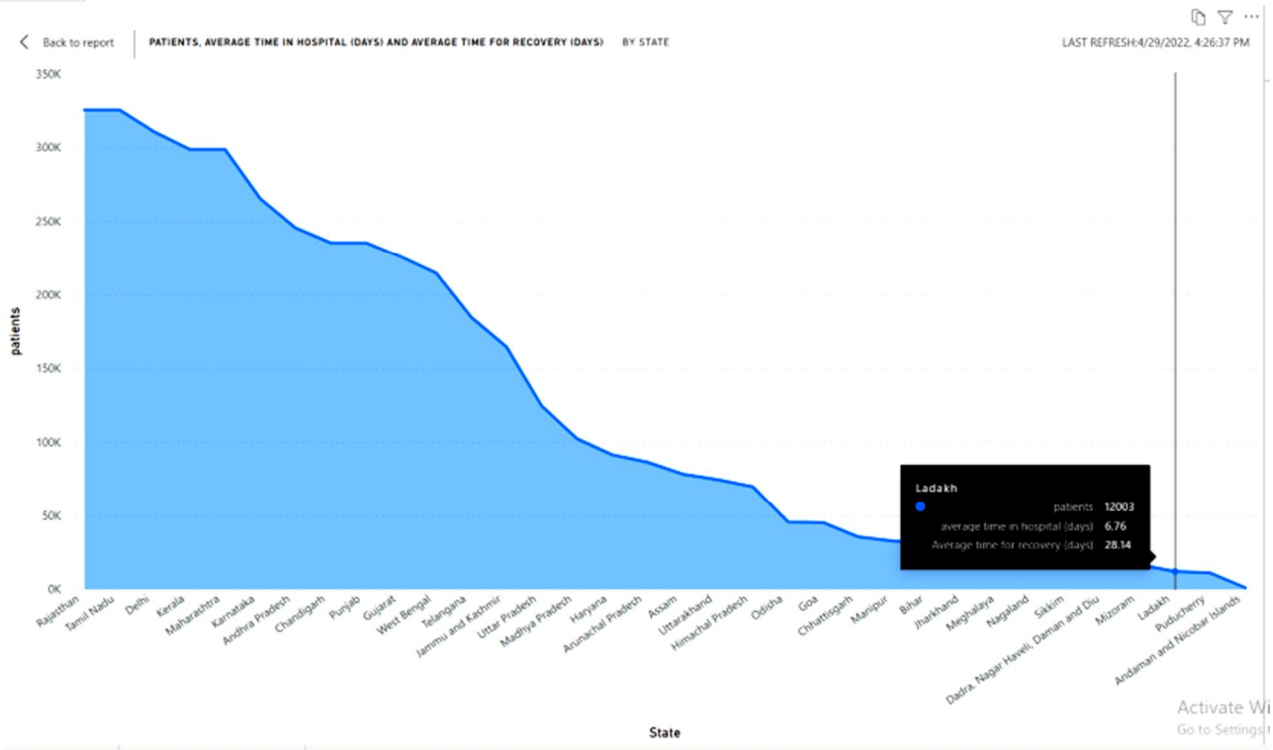
We made a visual report to understand how much time patients took their time to recover and who stays in the hospital .average time of a patient is different in every state because it depends on the patients majority. The area chart shows the patients data that how many patients are staying in hospital and how many are recovered all these are calculated by the patients average rate in per state.






Back to report
PATIENTS, AVERAGE TIME IN HOSPITAL (DAYS) AND AVERAGE TIME FOR RECOVERY (DAYS) BY STATE
LAST REFRESH:4/29/2022, 4:26:37 PM
350K
300K
250K
200K
150K
100K
50K
0K
West Bengal
patients 214566
average time in hospital (days) 7.15
Average time for recovery (days) 11.48
Rajasthan
Tamil Nadu
Delhi
Kerala
Maharashtra
Karnataka
Andhra Pradesh
Chandigarh
Punjab
Gujarat
West Bengal
Telangana
Jammu and Kashmir
Uttar Pradesh
Madhya Pradesh
Haryana
Arunachal Pradesh
Assam
Uttarakhand
Himachal Pradesh
Odisha
Goa
Chhattisgarh
Manipur
Bihar
Jharkhand
Meghalaya
Nagaland
Sikkim
Dadra, Nagar Haveli, Daman and Diu
Mizoram
Ladakh
Puducherry
Andaman and Nicobar Islands
State


Back to report
PATIENTS, AVERAGE TIME IN HOSPITAL (DAYS) AND AVERAGE TIME FOR RECOVERY (DAYS) BY STATE
LAST REFRESH:4/29/2022, 4:26:37 PM
patients
350K
300K
250K
200K
150K
100K
50K
0K
Odisha
patients 45612
average time in hospital (days) 22.95
Average time for recovery (days) 12.96
Rajasthan
Tamil Nadu
Delhi
Kerala
Maharashtra
Karnataka
Andhra Pradesh
Chandigarh
Punjab
Gujarat
West Bengal
Telangana
Jammu and Kashmir
Uttar Pradesh
Madhya Pradesh
Haryana
Arunachal Pradesh
Assam
Uttarakhand
Himachal Pradesh
Odisha
Goa
Chhattisgarh
Manipur
Bihar
Jharkhand
Meghalaya
Nagaland
Sikkim
Dadra, Nagar Haveli, Daman and Diu
Mizoram
Ladakh
Puducherry
Andaman and Nicobar Islands
State

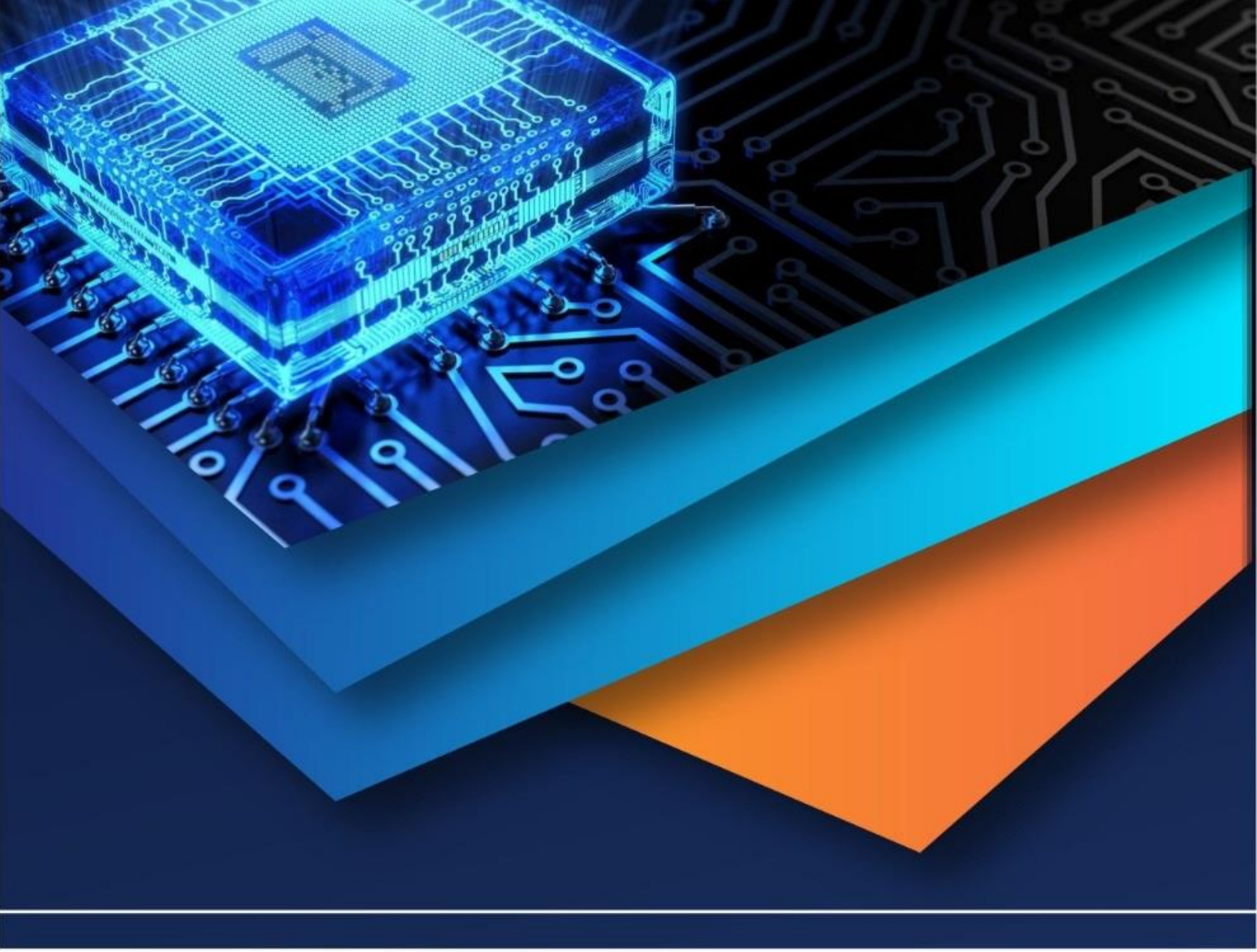


V. CONCLUSIONS

The volume of data increases dramatically over time, especially data generated on the global pandemic caused by COVID-19. Such volume of data requires utilizing big data analytics tools along with BI tool to make sense of the pandemic and its spread in a timely manner. In this study, we presented a review of several data analysis applications for COVID-19 to find key insights on patient demographics, patient clusters, state and district level spread and also to find transmission rate, top affected districts and predict the saturation point of the disease spread. Finally, we highlighted and discussed about track load on healthcare facility and predict number of ventilators required in future. The pandemic has exposed India's cherished inadequate medical infrastructure. India is a country of 1.3 billion people but around 75 precent of the healthcare infrastructure is focused in urban areas and making basic facilities inaccessible to rural areas. In this regards the data analytics can play a crucial role. It can bridge the gap between healthcare access and affordability across the country. So using this information, people of India can be prepared with a complete reformation within the healthcare sector led by digital technologies and the pandemic became a perfect time to effect this change.

REFERENCES

[1] https://www.kaggle.com/datasets
[2] https://www.kaggle.com/datasets/imdevskp/covid19-corona-virus-india-dataset?select=district_level_latest.csv
[3] P. Ghosh, R. Ghosh, , B. Chakraborty, "COVID-19 in India: Statewise Analysis and Prediction", *JMIR Public Health Surveill*, vol. 6(3), pp. e20341, 2020
[4] A. Haleem, M. Javaid, I. Haleem Khan, and R. Vaishya,, "Significant Applications of Big Data in COVID-19 Pandemic", *Indian J Orthop*. Vol. 54(4), pp. 526–528, 2020 .
[5] S. J. Alsunaidi, A.M. Almuhaideb, N. M. Ibrahim, F. S. Shaikh, K. S. Alqudaihi, F. A. Alhaidari, I. Ullah Khan, N. Aslam, and M.d S. Alshahrani, "Applications of Big Data Analytics to Control COVID-19 Pandemic", 21(7): 2282, 2021.
[6] A R Pradana, S R Madjid, H J Prayitno, R D Utami and Y Dharmawan , "Potential Applications of Big Data for Managing the COVID-19 Pandemic", **DOI** 10.1088/1742-6596/1720/1/012002.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY