 INDIAN INSTITUTE OF REMOTE SENSING,

DEPARTMENT OF SPACE, ISRO,

GOVT. OF INDIA.

**COVID-19 ANALYSIS USING BIGDATA TOOLS/ FRAMEWORKS**

GROUP ASSIGNMENT 2025

MODULE: BIG GEODATA PROCESSING

**SUBMITTED TO:**

**Dr. Kapil Oberai**

Scientist/ Engineer- SF

Geoinformatics Department

Indian Institute of Remote Sensing, ISRO

Govt. of India, Dehradun.

**SUBMITTED BY,**

*Jakka Durga Prasad*

*Shreya Singha*

*J Exousia Josephine*

**CONTENTS:**

|  |  |  |
| --- | --- | --- |
| SL NO. | TOPIC | PAGE NO. |
| 1 | **CHAPTER- 1** |  |
| 1.1 | INTRODUCTION | 1 |
| 1.2 | LITERATURE REVIEW | 1 |
| 1.3 | OBJECTIVE | 1 |
| 1.4 | STUDY AREA | 2 |
| 2 | **CHAPTER- 2** |  |
| 2.1 | DATA REQUIRED | 3 |
| 2.1.1 | DATA USED | 3 |
| 2.2 | WORKFLOW | 3 |
| 2.2.1 | TOOLS USED | 4 |
| 3 | **CHAPTER- 3** |  |
| 3.1 | SCALABILITY | 4 |
| 3.2 | RESULTS AND DISCUSSION | 4 |
| 3.2.1 | OBSERVATIONS | 8 |
| 3.3 | CONCLUSION | 9 |
| 4 | **BIBLIOGRAPHY** | 9 |

**LIST OF FIGURES AND TABLES**

|  |  |  |
| --- | --- | --- |
| FIG NO. | FIGURE NAME | Pg NO. |
| 1 | Boundary Map of Study Area | 2 |
| 2 | Workflow of Covid 19 Analysis | 3 |
| 3 | Scalability Analysis | 4 |
| 4 | Covid-19 Analysis for the month of January | 4 |
| 5 | Covid-19 Analysis for the month of February | 5 |
| 6 | Covid-19 Analysis for the month of March | 5 |
| 7 | Covid-19 Analysis for the month of April | 5 |
| 8 | Covid-19 Analysis for the month of May | 6 |
| 9 | Covid-19 Analysis for the month of June | 6 |
| 10 | Covid-19 Analysis for the month of July | 6 |
| 11 | Covid-19 Heat Map for the month of February | 7 |
| 12 | Covid-19 Heat Map for the month of March | 7 |
| 13 | Covid-19 Heat Map for the month of April | 7 |
| 14 | Covid-19 Heat Map for the month of May | 7 |
| 15 | Covid-19 Heat Map for the month of June | 8 |
| 16 | Covid-19 Heat Map for the month of July | 8 |

**CHAPTER-1**

* 1. **INTRODUCTION**

The COVID-19 pandemic has emphasised the necessity for advanced data analytics to collect, manage, analyse and mitigate its impact. This emerged as a overall global health problem in late 2019 and very Fastly it becomes a public health emergency. Very large data are produced daily, including data regarding the pandemic. Traditional data processing tools like RDBMS often fall short when handling the vast and complex datasets. Here, Big Data technologies, plays a crucial role in handling this large dataset, offer scalable solutions for processing and analysing extensive datasets efficiently (Laham, Shafinah, Hamid, Ismail, & Ainuddin, 2023). This project mainly focuses on the pandemic data analysis using big data tools and other visualization techniques that are available .Hive, framework built on top of Hadoop (GeeksforGeeks, 2024), is used for a state-wise metrics analysis focusing on confirmed cases, number of deaths, recoveries, incident rate, and case fatality rate.

* 1. **LITERATURE REVIEW**

Hadoop and hive are two crucial frameworks for bigdata handling like covid-19 pandemic large scale health datasets. Apache Hive which is a Hadoop based system was used by (Sheng, Khan, & Wang, 2020) to perform predictive analytics on a large covid-19 data, which provided the efficient storage, querying, and pattern discovery.

Similar work was done (Pletcher, 2011) used hive for trend analysis which provided the powerful real-time insights for corona pandemic. Together, these studies highlight the practical application and methodological importance of Hive in COVID-19 data analysis, especially for structured querying and visualization tasks.

(Singh, Kumar, Singh, & Kaur, 2023) demonstrates the abilities of Power BI, which can be used to design the interactive dashboards which can provide the valuable insights into the complex datasets.

* 1. **OBJECTIVE**

1. To analyse COVID-19 data like confirmed cases, number of deaths, recoveries, incident rate, and case fatality rate, using Cloudera Hive framework.
2. To create a dashboard and visualize via BI tools like Power BI.
   1. **STUDY AREA**

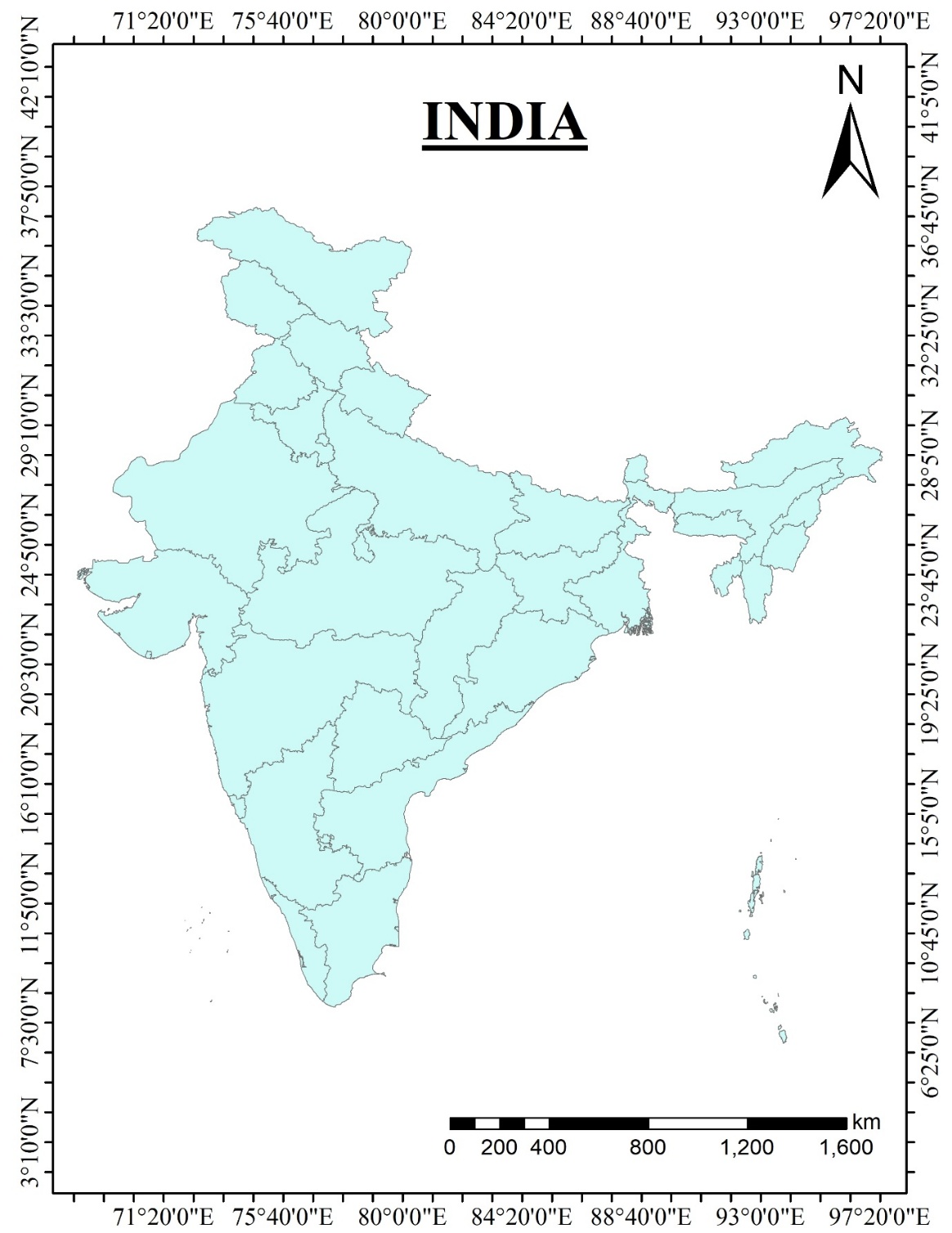
India is the most populous country in the entire world. Because of its high population density, it offers a complex environment for epidemiological studies. Indian subcontinent is situated north of the equator, between 8°4' N to 37°6' N latitude and 68°7' E to 97°25' E longitude. It is the seventh largest country in the world. It is the subcontinent with Indian Ocean in the south, Bay of Bengal in the east and Arabian Sea in the west. It has Ganga-Brahmaputra River system which feeds population of this land. The annual mean temperature lies between 24°C to 27°C and shows a tropical, hot, and humid weather. The average elevation of the area is about 1,000 meters above the sea level (Pletcher, 2011). It has about 28 states and 8 union territories.

Fig 1 : Boundary map of Study Area

**CHAPTER-2**

**2.1 DATA REQUIRED**

**COVID DATA:** CSSE COVID-19 daily reports (John Hopkins University)

* + 1. **DATA USED**
* Confirmed cases
* No. of deaths
* Recovered cases
* Incident rate
* Case Fatality rate
* No. of Doctors

**2.2 WORKFLOW**

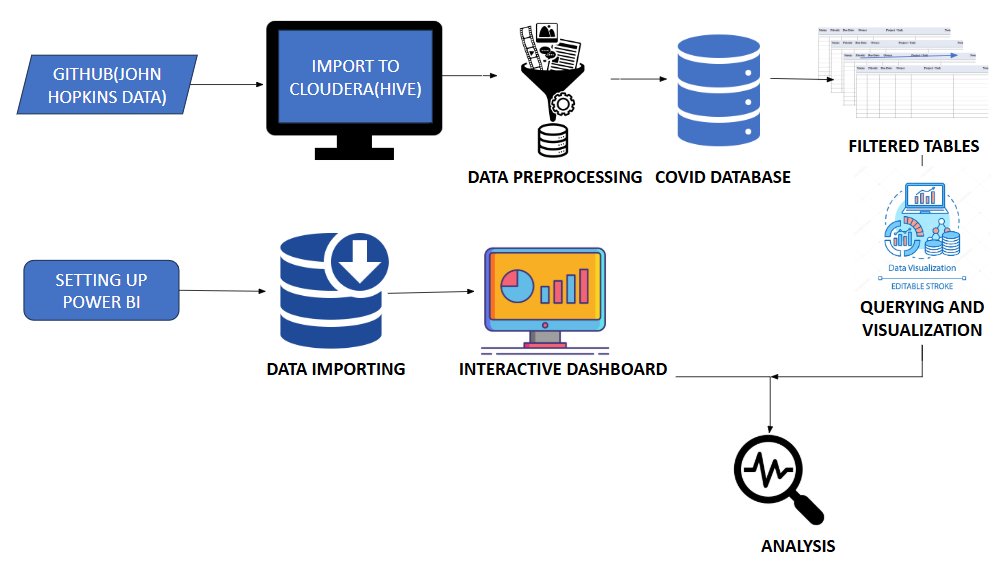
****

Fig 2 : Workflow of Covid 19 Analysis

**2.2.1 Tools Used**

* **Cloudera QuickStart VM**: Used for data storage, preprocessing, and querying with Hive.
* **Power BI**: Used for dashboard creation and data visualization.

**CHAPTER-3**

**3.1 SCALABILITY**

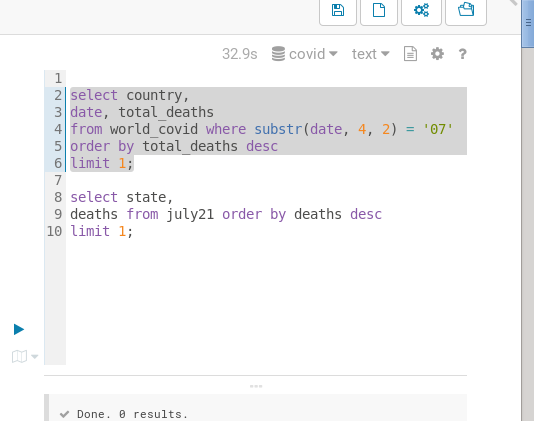
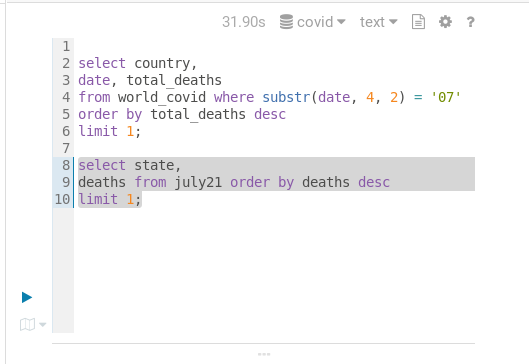
****

Fig 3 : Scalability Analysis

**3.2 RESULTS AND DISCUSSION**

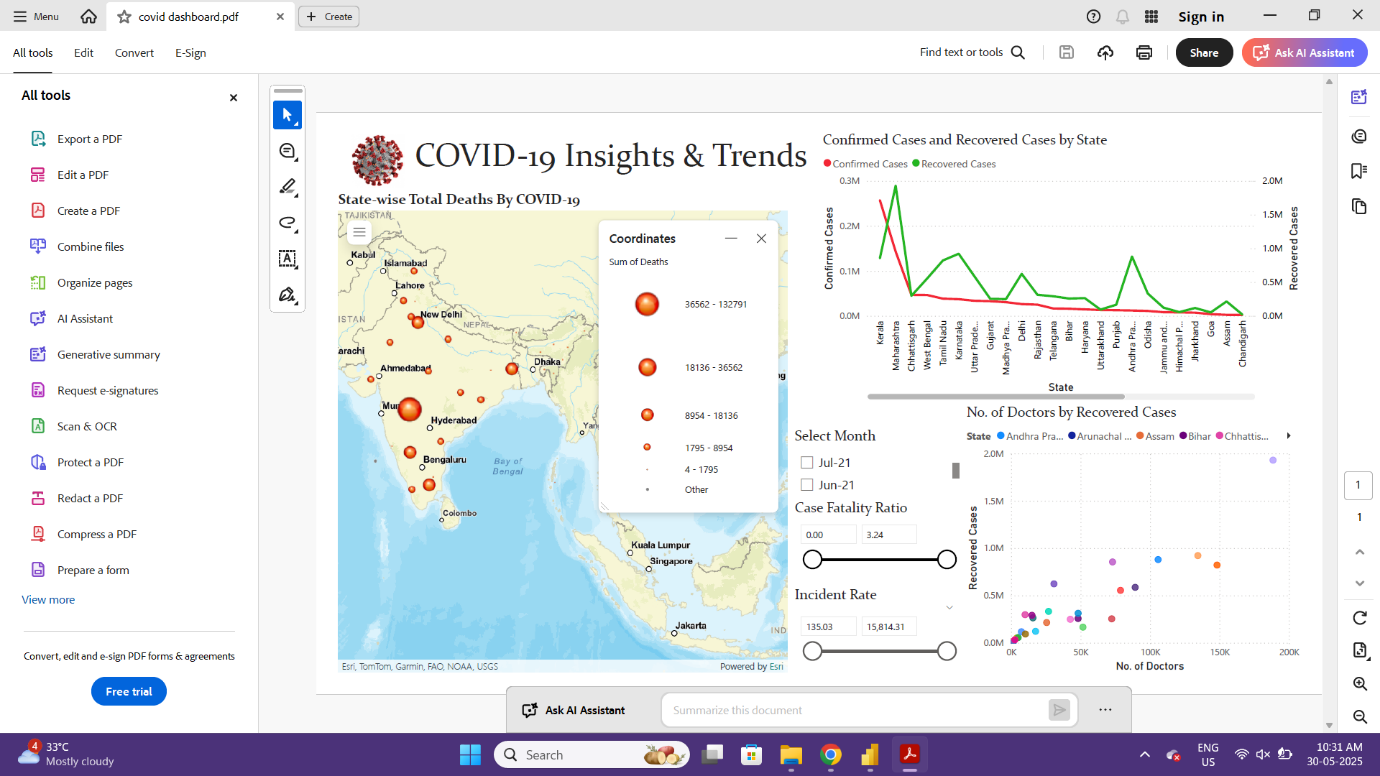
****The Key Feature of Power BI that we have used here is an interactive map visualization, which displays the overall geographical spread of the covid- 19 across India. In this map we did a location-based data to highlight the hotspots that are in a region. This will help the users to quickly identify areas with very high infected or active people**.**

Fig 4 : Covid-19 Analysis for the month of January

Fig 4 : Covid-19 Analysis for the month of January

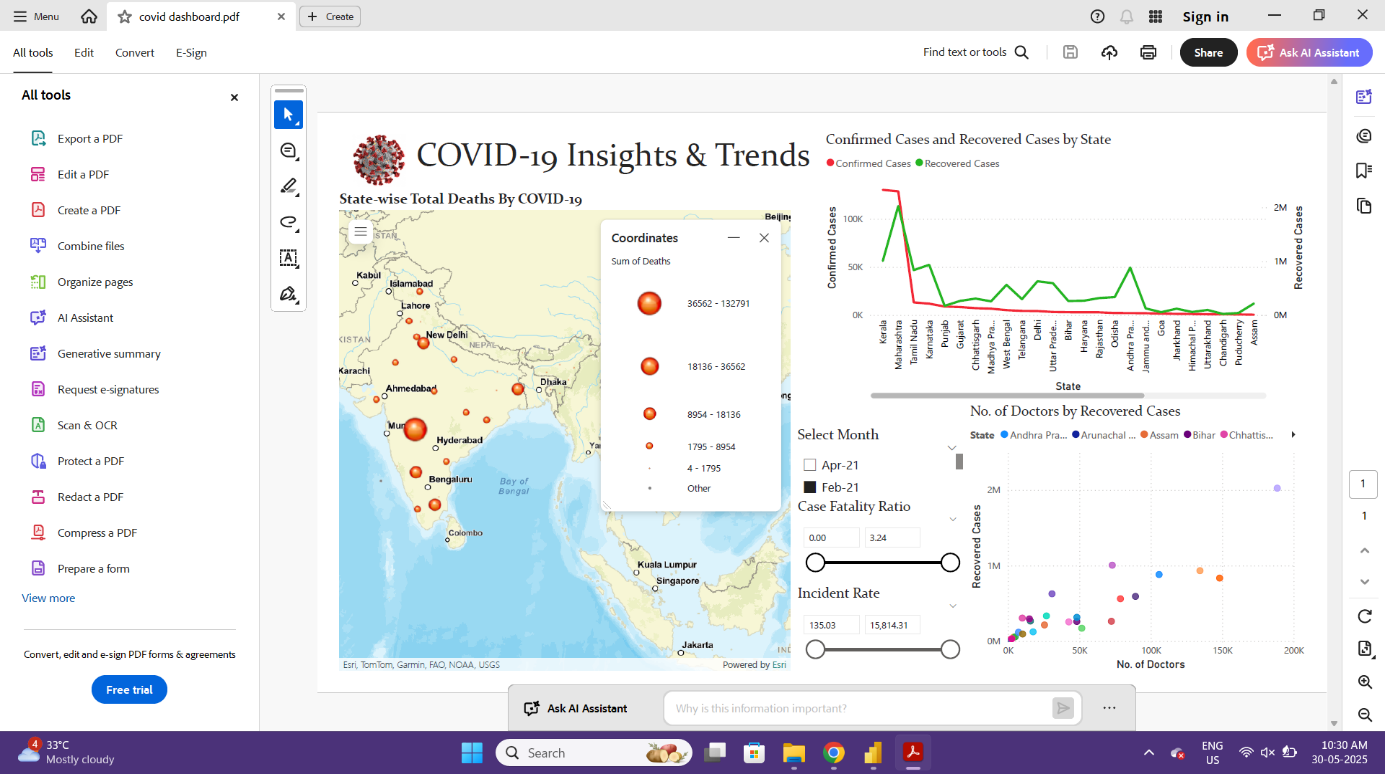
****

Fig 5 : Covid-19 Analysis for the month of February

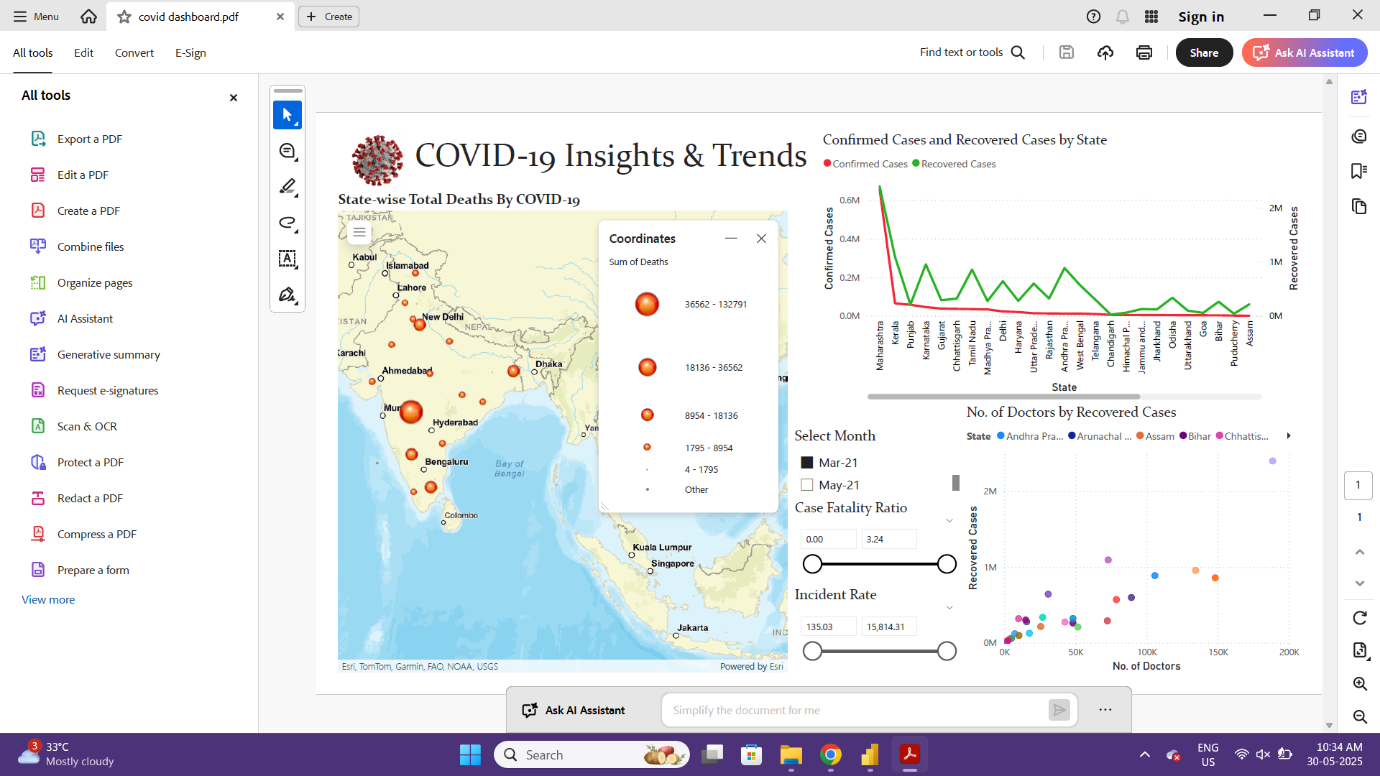
****

Fig 6 : Covid-19 Analysis for the month of March

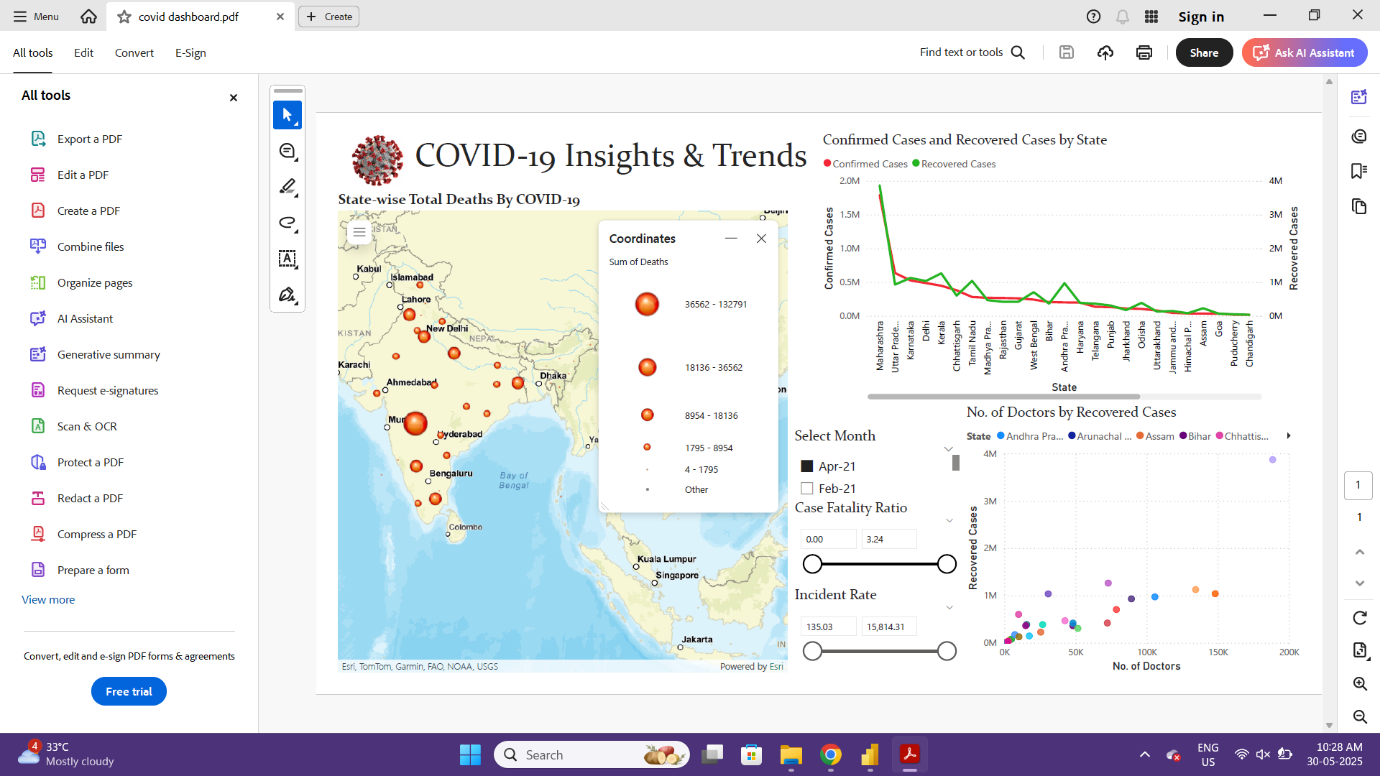
****

Fig 7 : Covid-19 Analysis for the month of April

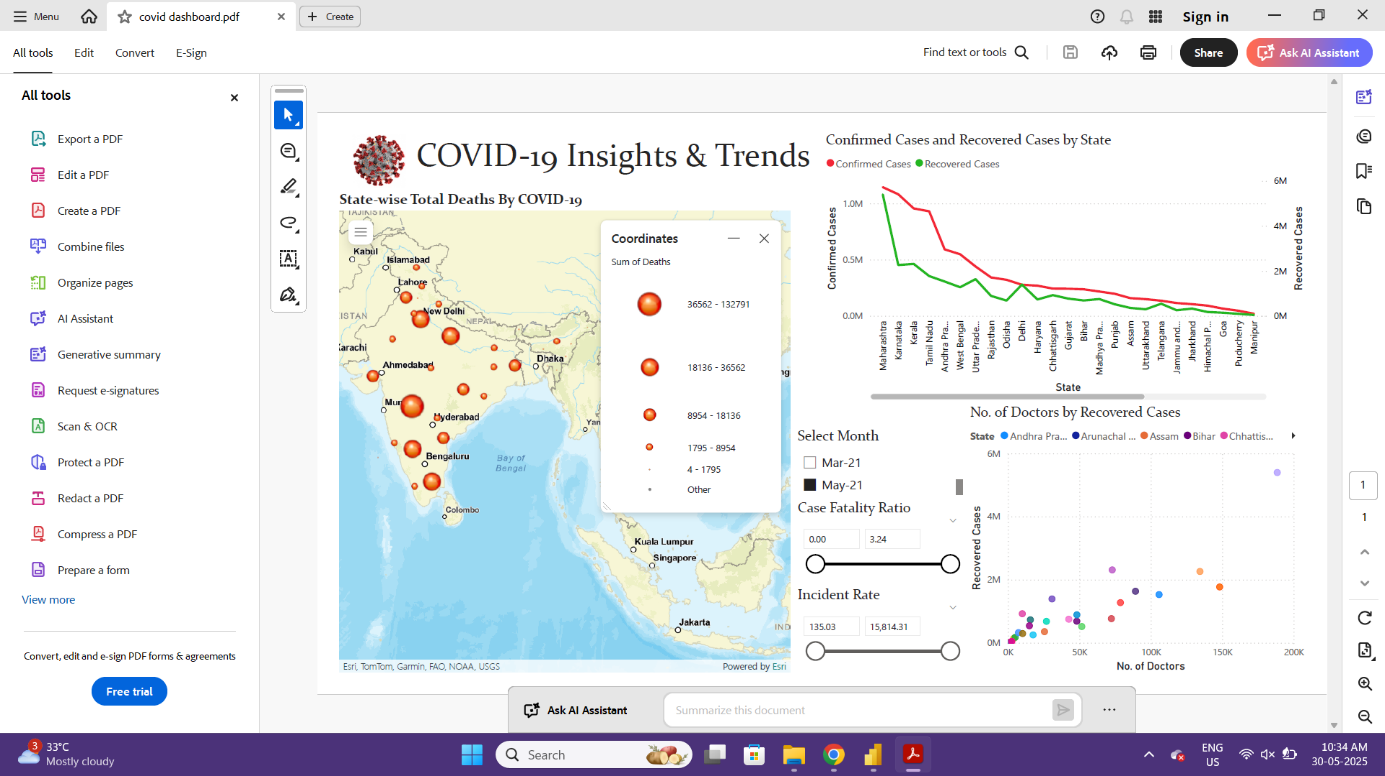
****

Fig 8 : Covid-19 Analysis for the month of May

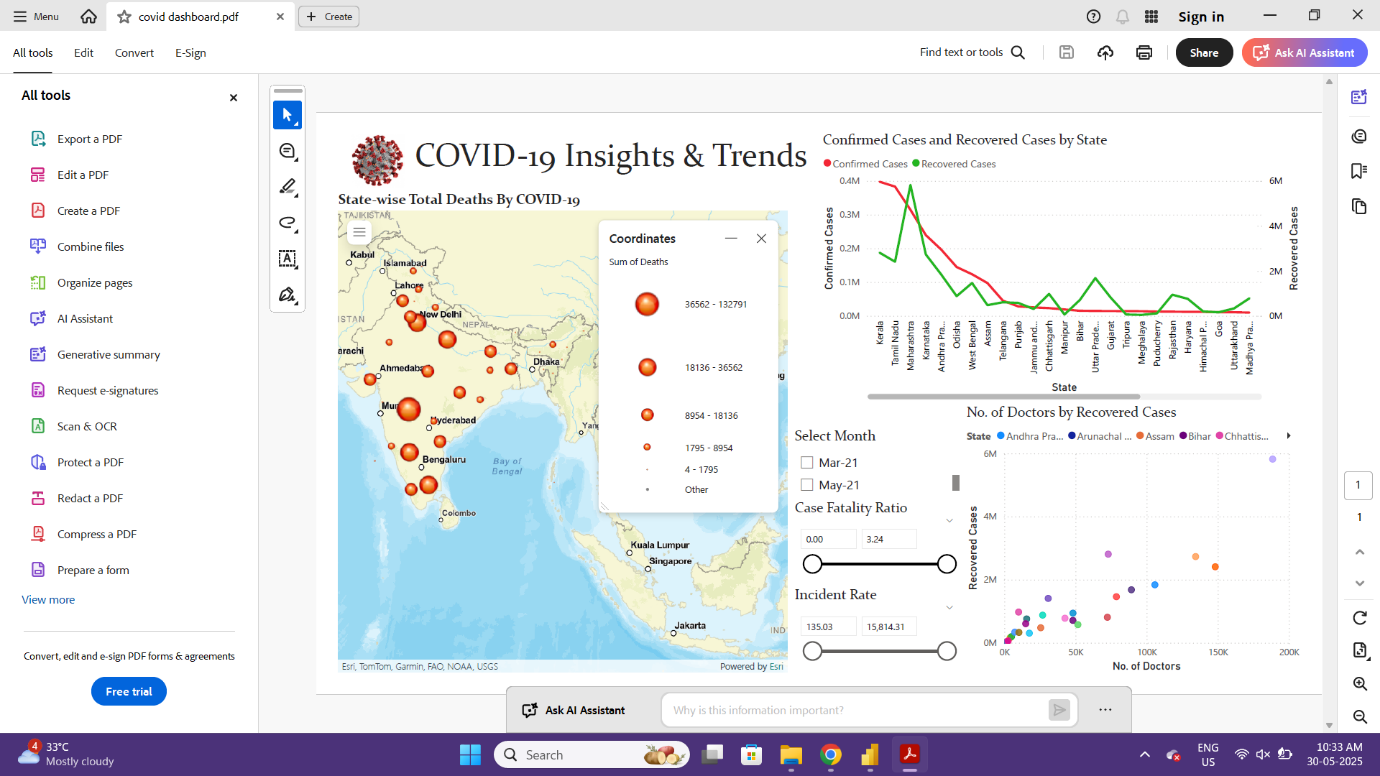
****

Fig 9 : Covid-19 Analysis for the month of June

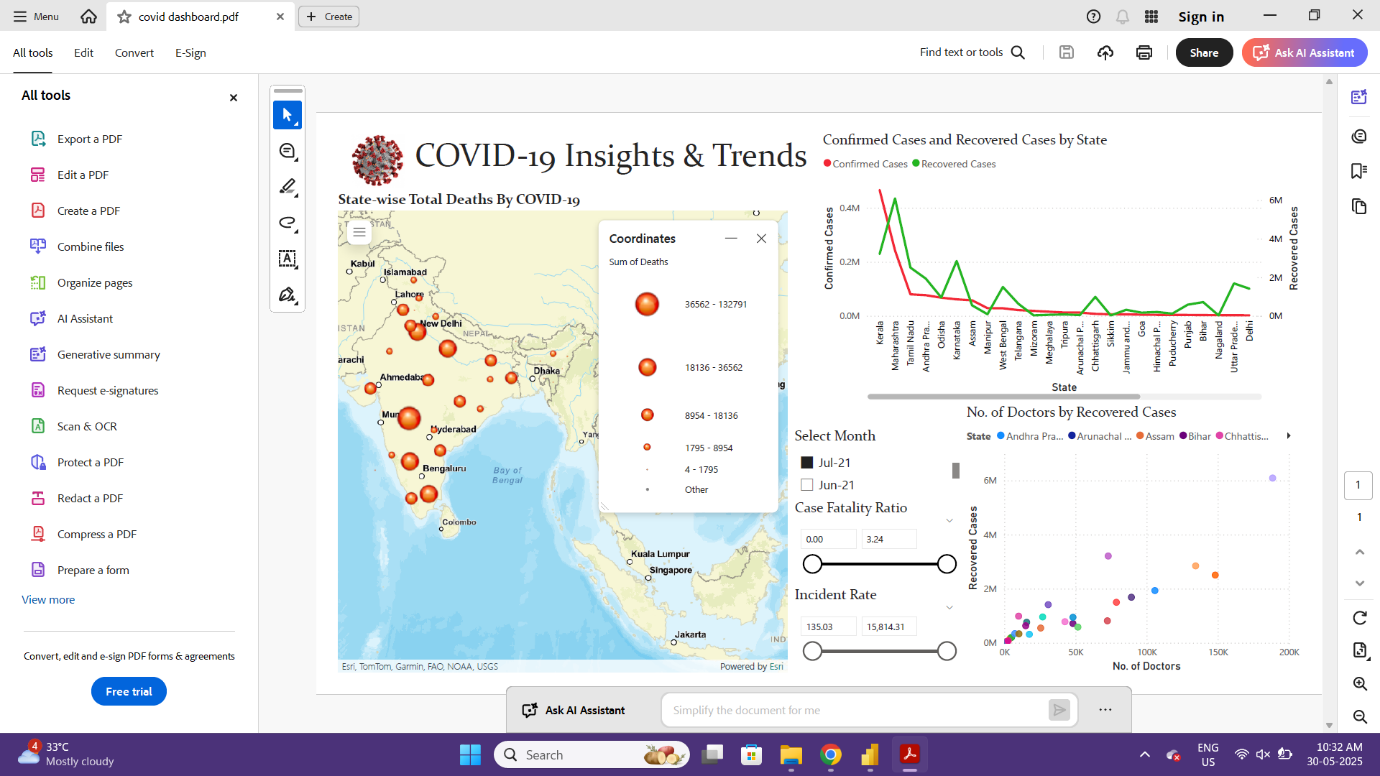
****

Fig 10 : Covid-19 Analysis for the month of July

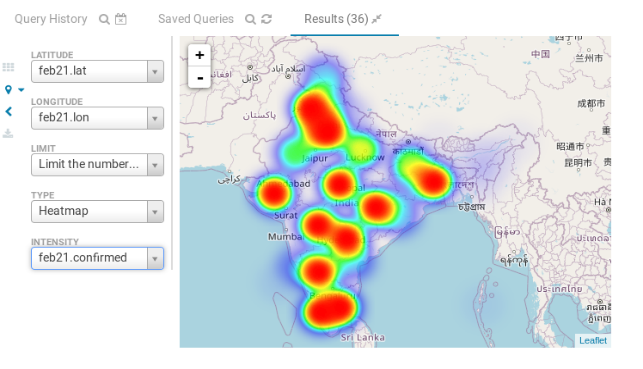
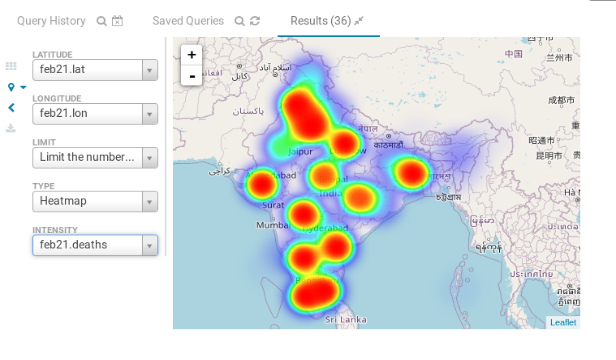
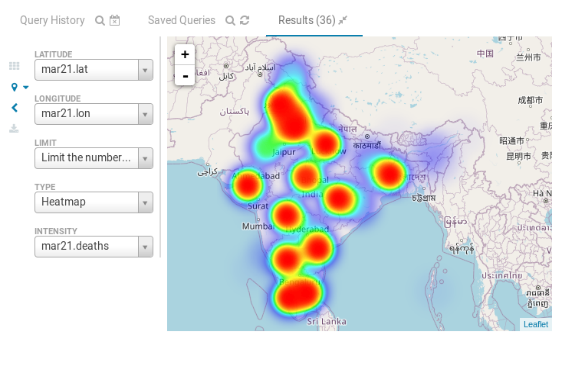
****

Fig 11 : Covid-19 Heat Map for the month of February

****

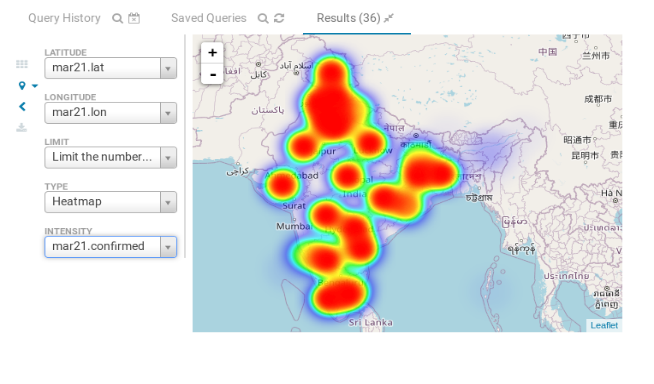
****

Fig 12 : Covid-19 Heat Map for the month of March

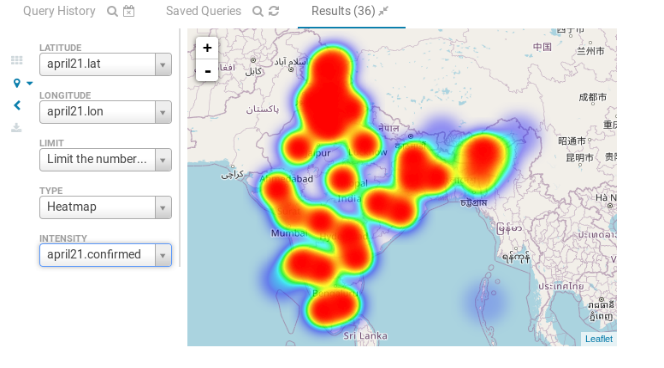
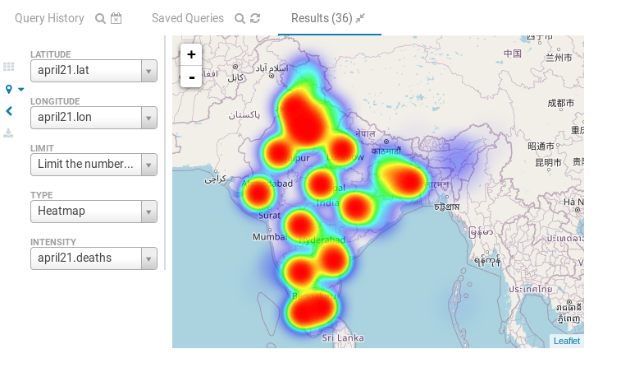
****

Fig 13 : Covid-19 Heat Map for the month of April

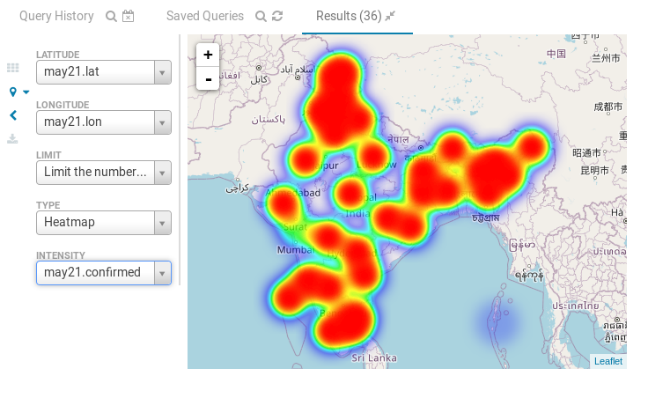
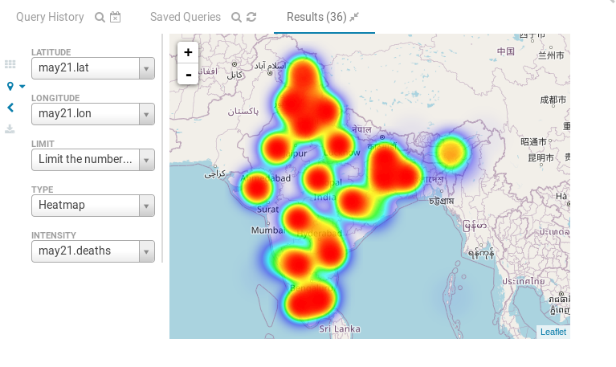
****

Fig 14 : Covid-19 Heat Map for the month of May

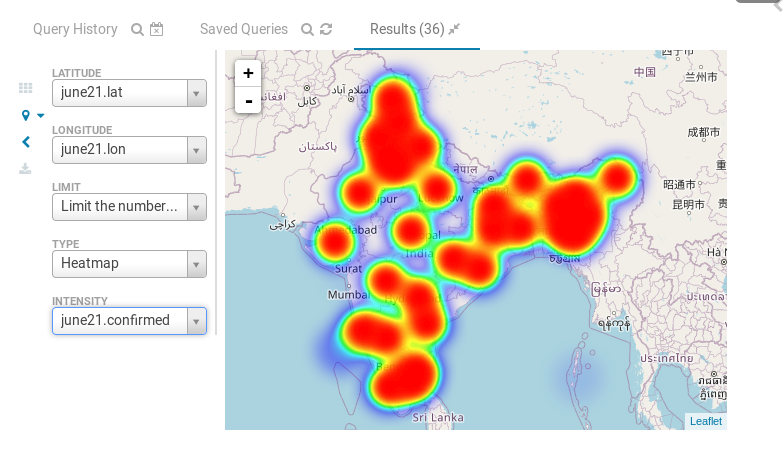
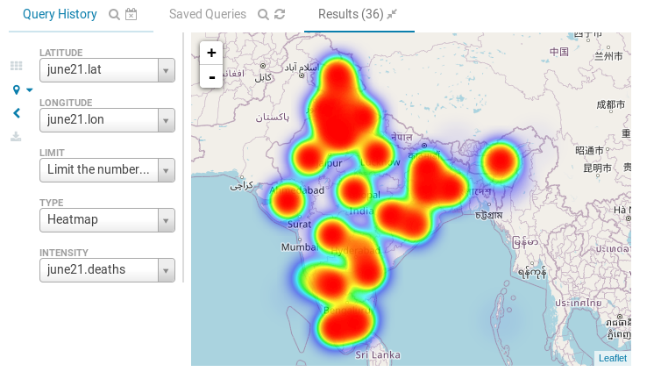
****

Fig 15 : Covid-19 Heat Map for the month of June

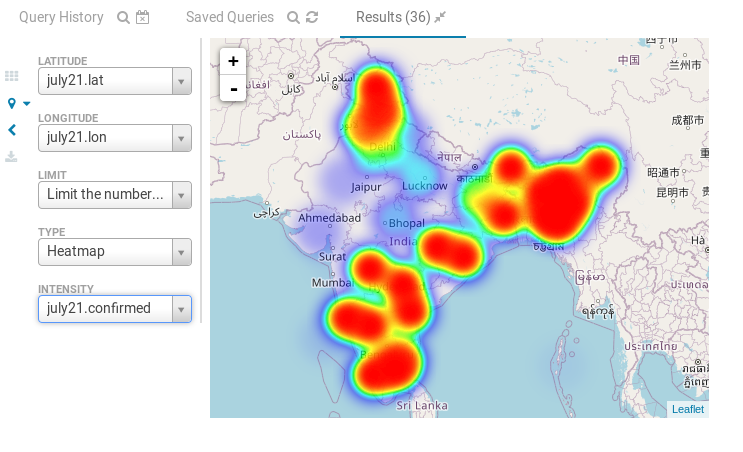
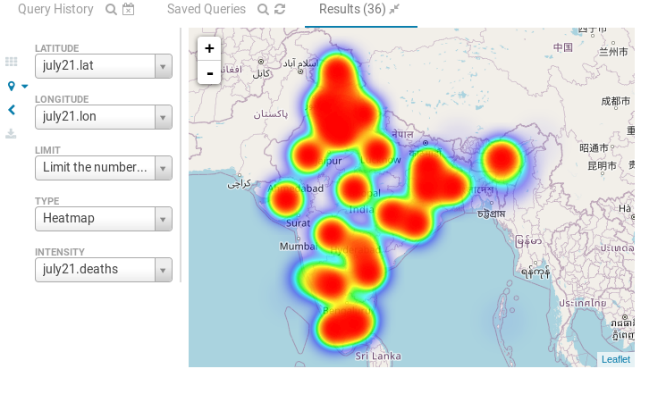
****

Fig 16 : Covid-19 Heat Map for the month of July

**3.2.1 OBSERVATIONS**

* Kerala is showing the highest number of confirmed cases and Daman & Diu showing the lowest cases.
* Top five states with highest confirmed cases are: Kerala, Maharashtra, Chhattisgarh, West Bengal and Tamil Nadu
* Maharashtra has the highest death cases and Lakshadweep has the lowest.
* The top five states with highest death cases are: Maharashtra, Tamil Nadu, Karnataka, Delhi and West Bengal.
* Maharashtra shows the highest recovered cases and Lakshadweep with lowest recovered cases.
* The top five states in recovered cases are: Maharashtra, Karnataka, Andhra Pradesh, Kerala and Tamil Nadu
* Punjab is showing highest cases fatality ratio (3.24) and Daman & Diu with lowest ratio (0.06)
* Ladhak has highest incident rate (3543.71) and Lakshadweep with lowest rate (135.03)
* The scatterplots between the number of doctors and recovered cases showing us the strong positive correlation, where number of doctors is the independent variable and number of cases is a dependent variable. This indicates a higher number of doctors leads to higher recovered cases.

**3.3 CONCLUSION**

This project demonstrated the powerful combination of big data processing and modern visualization tools in analysing public health data. Cloudera QuickStart VM enabled efficient data management, while Power BI provided intuitive and insightful visualizations. As COVID-19 continues to affect lives and spread in new waves, the importance of such analytical platforms remains vital. This study not only informs current strategies but also lays the groundwork for responding to future pandemics. In conclusion, as covid-19 awareness is on increasing trend one side, especially in the developing regions of India, successful strategies can be made for future public health planning.

**BIBLIOGRAPHY**

GeeksforGeeks. (2024). Apache Hive | GeeksforGeeks.

Kareem, A., & Mahdi, L. (2023). Predictive analytics on COVID-19 data using Hive based on Hadoop cluster. *Indonesian Journal of Electrical Engineering and Computer Science*, *31*, 945. https://doi.org/10.11591/ijeecs.v31.i2.pp945-956

Laham, M., Shafinah, K., Hamid, N., Ismail, Z., & Ainuddin, S. (2023). Performance Analysis of Apache Hadoop Using Hive on COVID19 Datasets. *Journal of Computer Science & Computational Mathematics*, *13*, 83–87. https://doi.org/10.20967/jcscm.2023.03.003

Pletcher, K. (2011). *The geography of India*. 376.

Sheng, J., Khan, Z., & Wang, X. (2020). COVID‐19 Pandemic in the New Era of Big Data Analytics: Methodological Innovations and Future Research Directions. *British Journal of Management*, *32*, 1164–1183. https://doi.org/10.1111/1467-8551.12441

Singh, G., Kumar, A., Singh, J., & Kaur, J. (2023). Data Visualization for Developing Effective Performance Dashboard with Power BI. *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, 968–973. https://doi.org/10.1109/ICIDCA56705.2023.10100169