International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Data Analysis for COVID-19 Study Forecast with Small Scale Data Sub-sampling

Shaikh Noman Nabi, Prof. Tagde R.O

Diploma in Computer Engineering. Professor, Dept of CO Sau Sundar Bai Manik Adsul Polytechnic , Chas

Abstract - The COVID-19 pandemic generated a large volume of healthcare data related to patient demographics, symptoms, disease severity, testing, hospitalization, vaccination, recovery, and mortality. Analyzing such data is essential for identifying patterns, understanding healthcare requirements, and presenting complex information in a clear and meaningful manner. This project, "COVID-19 Data Analytics and Interactive Dashboard," focuses on applying data analytics and visualization techniques to a structured dataset containing 1,000 patient records.

The primary objective of this project is to transform raw COVID-19 healthcare data into meaningful insights through a systematic data analytics process. The methodology includes data collection, data preprocessing, data cleaning, data transformation, exploratory data analysis, data modelling, visualization, and dashboard development. During preprocessing, the dataset was examined for missing values, duplicate records, inconsistent categorical values, incorrect data types, and potentially invalid numerical observations. Relevant transformations, including age-group classification and date-based analysis, were performed to make the dataset suitable for further analysis.

Microsoft Power BI was used as the primary visualization and dashboard development tool. Four interactive dashboards were developed: COVID-19 Analytics Dashboard, Patient Analysis Dashboard, COVID-19 Hospitals Dashboard, and COVID-19 Vaccination Dashboard. These dashboards use KPI cards, bar charts, line charts, donut charts, and interactive slicers to present information effectively. Users can filter the analysis using State and City selections, allowing both general and location-specific analysis.

The exploratory analysis identified several patterns within the project dataset. The dashboard contains 340 total cases, 946 recovery records, 54 death records, and 430 vaccinated records in the displayed analysis. The gender distribution is relatively balanced, with 504 female and 496 male records. The 60+ age group has the largest displayed representation, while Mild cases form the largest disease-severity category. Geographical analysis shows differences in patient and vaccination records across states and cities. Hospital analysis provides information related to patients, isolation, ventilator requirements, testing, ICU activity, and recovery. The vaccination dashboard further presents state-wise vaccination, vaccine brands, dose distribution, and booster activity over time.

Keywords : - SARS-CoV-2, COVID-19,Pandemics,Big Data Analytics, Bio Engineering , Predictive Modelling

CHAPTER 1: INTRODUCTION

Coronavirus Disease 2019 (COVID-19) emerged as a major global public health crisis after the identification of a novel coronavirus, later named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). The disease was first identified in Wuhan, China, in late 2019 and subsequently spread rapidly across countries and continents. Due to its high transmission rate and significant impact on healthcare systems, economies, education, employment, and everyday life, COVID-19 became one of the most significant global challenges of the twenty-first century.

The pandemic generated an enormous amount of health-related data. Information contain confirmed cases, deaths, recoveries, hospital admissions, intensive care unit (ICU) admissions, oxygen requirements, ventilator use, vaccination status, symptoms, age, gender, and geographical distribution was continuously collected. This created an opportunity to use data analytics techniques to understand the development and impact of the pandemic.

Problem Statement

The COVID-19 pandemic produced large amount of datasets containing information about patients, testing, hospitalization, vaccination, recovery, and mortality. However, raw datasets are often difficult to interpret directly because they may contain missing values, duplicate records, inconsistent formats, incorrect data types, and outliers. Without appropriate preprocessing and analysis, these problems can affect the accuracy and reliability of analytical results.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

For that reason we took a dummy dataset of 1000 peoples for our analysis for easier of output. In that data we used metro cities to describe our analysis report

A large amount of data sets is uneasier to get understand what happened across the world. To make work easier we get a dummy data of a portion by state, cities, district , hospitals.

Objectives of the Project

The major objectives of this project are:

1. To collect and organize COVID-19-related healthcare data.
2. To clean and preprocess the dataset before analysis.
3. To identify and handle missing, duplicate, inconsistent, and abnormal values.
4. To perform exploratory data analysis to understand the characteristics of the dataset.
5. To analyze COVID-19 cases based on demographic and clinical factors.
6. To examine hospitalization, ICU admission, and ventilator requirements.
7. To analyze recovery and mortality patterns.
8. To examine vaccination status and vaccination trends.
9. To create meaningful charts and interactive dashboards.
10. To present analytical findings in a clear and understandable manner

CHAPTER 2: LITERATURE REVIEW

The pandemic of Coronavirus Disease 2019 (COVID-19) is a timely reminder of the nature and impact of Public Health Emergencies of International Concern. As of 12 January 2022, there were over 314 million cases and over 5.5 million deaths notified since the start of the pandemic. The COVID-19 pandemic takes variable shapes and forms, in terms of cases and deaths, in different regions and countries of the world. The objective of this study is to analyse the variable expression of COVID-19 pandemic so that lessons can be learned towards an effective public health emergency response.

COVID-19 Data Analysis

Several studies have focused on analyzing COVID-19 datasets to understand the spread and progression of the disease. Researchers have used statistical analysis and visualization techniques to examine confirmed cases, deaths, recoveries, and trends over time. Time-series analysis has been particularly useful for identifying increases and decreases in reported cases and understanding different stages of the pandemic.

COVID-19 datasets generally contain multiple variables, including dates, geographical locations, demographic characteristics, symptoms, testing results, hospitalization information, and patient outcomes. Analysis of these variables allows researchers to identify relationships between different factors. For example, patient-level datasets can be examined to determine whether age, underlying health conditions, or disease severity is associated with hospitalization or mortality.

Healthcare Data Analytics

Healthcare data analytics involves the systematic examination of healthcare information to identify useful patterns and support decision-making. It combines techniques from statistics, computer science, data management, and visualization. Healthcare analytics can be categorized into descriptive, diagnostic, predictive, and prescriptive analytics.

Descriptive analytics focuses on understanding what has happened. During the COVID-19 pandemic, descriptive analytics was widely used to report the number of cases, deaths, recoveries, hospital admissions, ICU admissions, and vaccinations. Diagnostic analytics attempts to explain why particular patterns occurred by examining relationships between variables. Predictive analytics uses historical data and statistical or machine-learning models to estimate possible future outcomes.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

COVID-19 Risk Factors and Patient Characteristics

A significant area of COVID-19 research has involved identifying factors associated with severe disease and poor patient outcomes. Researchers have examined demographic characteristics such as age and gender as well as clinical factors including comorbidities and oxygen requirements.

Older patients and individuals with certain underlying health conditions were frequently identified as higher-risk groups in clinical research. Researchers also examined symptoms and indicators such as oxygen saturation, temperature, and disease severity to understand patient conditions.

Research Gap

Although extensive research has been conducted on COVID-19, many studies focus on individual aspects of the pandemic, such as case trends, mortality, vaccination, clinical outcomes, or transmission. There is an opportunity to develop an integrated analytical approach that combines patient characteristics, clinical information, hospitalization, testing, vaccination, recovery, and mortality within a single analytical framework.

Another limitation is that raw healthcare datasets are often difficult for non-technical users to interpret. A comprehensive interactive dashboard can address this issue by presenting important indicators and relationships in an accessible visual format.

Data analytics is a most important key role play in a future pandemic situation. Using that analysis we can predict what we needs to fight with this pandemic. According to previous disaster we will ready for future challenges. What we can do in that situation how can we manage such a large population over 8 billions present in a earth and other species which can play such a important role in a human life. It is also be a part of our environment and other activities.

CHAPTER 3: DATA PREPROCESSING

Dataset Size and Structure

The primary dataset consists of **1,000 records**, where each record represents information associated with a patient or COVID-19 case. Each row contains multiple attributes describing the patient's demographic characteristics, clinical condition, treatment or hospitalization information, testing details, vaccination status, and final outcome.

The dataset follows a tabular structure in which:

- **Rows** represent individual patient records.
- **Columns** represent attributes associated with each patient.
- **Categorical variables** describe groups such as gender, state, severity, blood group, vaccine type, and test result.
- **Numerical variables** represent values such as age, temperature, oxygen level, and recovery duration.
- **Date variables** represent events such as testing, admission, discharge, vaccination, recovery, and death.

Raw data can help us to create a prototype of that particular disaster to treat them easily without getting any problem when cause in future. Huge data can create confusion to make a planning. We took less than 1% data for our analysis which is better to understand and

Patient Identification and Demographic Information

The patient-related section of the dataset contains information that allows individual records to be distinguished and grouped for analysis. A **Patient ID** is used as a unique identifier for each record. It helps prevent confusion between records and can also be used when creating relationships between different tables in a larger database structure.

The dataset includes **gender information**, which is used extensively in the Patient Analysis dashboard. According to the dashboard, the dataset contains **504 female records and 496 male records**, giving a total of 1,000 patient records. This allows gender-based comparisons to be performed.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Age is another important demographic variable. Rather than analyzing every individual age separately, the data can be grouped into categories such as:

- 17–30 years
- 31–45 years
- 46–60 years
- 60+ years

The Age Group Distribution chart in the main COVID-19 Analytics Dashboard uses these categories. Age-group analysis is important because it allows the distribution of patients across different age categories to be compared visually.

4 GEOGRAPHICAL INFORMATION

Geographical information is an important component of the dataset because COVID-19 cases can vary significantly between different regions. The dataset contains **State** and **City** fields, allowing the records to be analyzed geographically.

The **Patient by City** visualization displays patient distribution across cities such as Hyderabad, New Delhi, Jaipur, Chennai, Bengaluru, Kolkata, Ahmedabad, Surat, Lucknow, and Mumbai. Similarly, the **Cases by State** and **Vaccination by State** charts provide comparisons between different states.

This geographical structure makes it possible to identify areas with relatively higher or lower numbers of patients and vaccinations.

Clinical and Patient Health Information

The clinical portion of the dataset contains information related to the patient's health condition. Important variables include **symptoms, temperature, disease severity, oxygen level, and comorbidities**.

The **Severity** field is particularly important for the Patient Analysis dashboard. The dashboard categorizes patients into:

- Mild
- Moderate
- Severe
- Critical

The Patient Severity by Disease visualization allows the number of patients in each severity category to be compared. In the displayed dashboard, mild and moderate cases represent large portions of the dataset, while severe and critical cases account for smaller groups.

COVID-19 Testing Information

Testing information forms another important part of the dataset. The testing section contains variables such as:

- Test date
- Test type
- Test result
- Test laboratory

These fields allow the testing activity to be analyzed according to date, test type, result, or laboratory.

The dataset contains **340 records classified as positive** in the dashboard's displayed KPI. This value is useful for demonstrating how a raw categorical field can be transformed into a meaningful dashboard indicator.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Recovery and Mortality Information

Patient outcomes are represented through recovery and mortality-related fields. These include **recovery status, recovery date, death status, death date, and recovery time in days**.

The recovery information allows analysts to calculate or display the number of recovered patients and examine the duration required for recovery. The hospital dashboard contains an **Average Recovery Days** visualization, which presents recovery-duration values.

The dataset also contains death-related information that can be used to analyze mortality trends. The hospital dashboard combines death information with ICU counts to provide a daily comparison.

Vaccination Information

Vaccination is one of the major analytical components of the project. The vaccination section of the dataset contains fields such as **vaccination status, vaccine name, Dose 1, Dose 2, and Booster Dose**.

The Vaccination Dashboard provides several visualizations based on these variables. The **Vaccination by State** chart compares vaccination records across states. The displayed dashboard includes states such as Tamil Nadu, Telangana, Delhi, Gujarat, Kerala, Rajasthan, West Bengal, Uttar Pradesh, Maharashtra, and Karnataka.

The **Top Vaccine Brands** visualization provides a graphical view of vaccine-related records. This enables vaccine brands to be compared based on their occurrence in the dataset.

CHAPTER 4 : EXPLORATIRY DATA ANALYSIS

Overall COVID-19 Case Analysis

The first level of analysis focuses on obtaining an overall understanding of the dataset. The main COVID-19 Analytics Dashboard contains KPI cards for **Total Cases, Recovered, Death, and Vaccinated**. These KPIs provide a quick summary of important indicators and allow users to understand the overall status of the dataset.

The dashboard displays **340 total cases** in the Total Cases KPI, **946 recovered records**, **54 deaths**, and **430 vaccinated records**. These values provide a high-level summary of the selected data and can be used as starting points for more detailed analysis

Cases Trend Over Time

Time-based analysis is important for understanding how COVID-19-related records change across different dates. The **Cases Trend Over Time** line chart was developed to represent changes in case counts across the available period

Age Group Distribution

Age is an important demographic variable in the dataset. To make the analysis easier to interpret, individual ages were grouped into four categories:

- **17–30**
- **31–45**
- **46–60**
- **60+**

The Age Group Distribution chart indicates that the **60+ age group has the highest number of records** among the displayed categories. The 46–60 group follows, while the 17–30 and 31–45 groups have comparatively fewer records.

State-Wise Case Analysis

Geographical analysis was performed using the State and City attributes. The **Cases by State** bar chart provides a comparison of COVID-19-related records among different states.

The displayed visualization includes states such as **Tamil Nadu, Telangana, Delhi, Gujarat, Kerala, Rajasthan, West Bengal, Uttar Pradesh, Maharashtra, and Karnataka**. Tamil Nadu appears at the top of the displayed ranking, followed by Telangana and Delhi

Gender Distribution

Gender-based analysis was conducted to understand the composition of the dataset. The Patient Analysis Dashboard displays separate KPI cards for **Female** and **Male** patients.

The dataset contains **504 female records and 496 male records**, indicating a relatively balanced distribution between the two categories. The gender distribution can also be examined through the visualizations available on the dashboard

Patient Distribution by City

City-level analysis provides a more detailed geographical perspective than state-level analysis. The **Patient by City** visualization displays the number of patient records across selected cities.

The displayed dashboard includes **Hyderabad, New Delhi, Jaipur, Chennai, Bengaluru, Kolkata, Ahmedabad, Surat, Lucknow, and Mumbai**. Hyderabad and New Delhi appear among the cities with the highest displayed patient counts.

Patient Severity Analysis

Disease severity is one of the important clinical variables analyzed in the project. The dataset categorizes patients into four severity levels:

- Mild
- Moderate
- Severe
- Critical

The **Patient Severity by Disease** chart shows a larger number of patients in the Mild category, followed by Moderate and Severe categories. Critical cases form the smallest category among the displayed groups

Blood Group Distribution

The Patient Analysis Dashboard contains a **Blood Group Distribution** chart. It displays the frequency of different blood groups among the patient records.

The dashboard includes O−, B+, B−, AB−, O+, A−, A+, and AB+. The values are relatively close to one another, although some groups appear more frequently than others.

This analysis demonstrates how categorical variables can be summarized using bar charts. It provides descriptive information about the dataset but should not be interpreted as a medical relationship between blood group and COVID-19 outcomes.

ICU and Death Analysis

The **Death and ICU Count Over Day by Day** chart combines time-based information about deaths and ICU-related records. The visualization shows fluctuations in the values across different dates.

The presence of peaks indicates days with relatively higher recorded activity. Comparing ICU and death information can provide useful descriptive insights into healthcare outcomes and critical cases.

Recovery Analysis

Recovery is another major outcome variable. The dashboard contains a **Recovered** KPI and an **Average Recovery Days** visualization.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

The recovery-duration chart displays different recovery-day values, ranging from shorter to longer recovery periods. This allows the distribution of recovery duration to be examined.

Vaccination Analysis

Vaccination data was explored through a dedicated Vaccination Dashboard. The dashboard includes **Vaccination by State, Top Vaccine Brands, Vaccination Rate, and Booster Dose Day by Day**.

The Vaccination by State chart shows differences in vaccination records between states. Tamil Nadu, Telangana, Delhi, and Gujarat are among the higher-ranked states displayed in the visualization.

The Vaccination Rate donut chart compares **Dose 1, Dose 2, and Booster Dose** records. The dashboard also includes a line chart showing booster-dose activity across different days.

CHAPTER 5 : METHODOLOGY

Research and Analytical Approach

This project follows a **descriptive and exploratory data analytics approach**. Descriptive analytics is used to summarize the information contained within the COVID-19 dataset, while exploratory analysis is used to identify patterns, variations, and relationships between different variables.

The project does not attempt to provide medical diagnosis or establish clinical causation. Instead, it focuses on understanding the available data and presenting it through meaningful analytical visualizations.

The overall methodology can be represented as:

Data Collection → Data Understanding → Data Cleaning → Data Transformation → Exploratory Analysis → Data Modelling → Visualization → Dashboard Development → Insight Generation

Data Collection

The first stage of the methodology was the preparation of the COVID-19 dataset. The dataset contains **1,000 patient records** and includes information related to demographic characteristics, clinical conditions, geographical location, hospital treatment, testing, vaccination, recovery, and mortality.

The major categories of information include:

- Patient identification
- Age and gender
- State and city
- Symptoms
- Temperature
- Disease severity
- Oxygen level
- Comorbidities
- Hospital information
- Admission and discharge dates
- ICU admission
- Ventilator requirement

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

- Vaccination status
- Vaccine name
- Dose information
- Testing information
- Recovery information

Death information

Data Preprocessing

Data preprocessing was performed to improve the quality and reliability of the dataset. Raw data can contain missing values, duplicates, inconsistent formats, and invalid observations, which can affect the accuracy of analytical results.

The preprocessing stage included:

Missing-Value Handling

The dataset was checked for blank and null values. Missing information was reviewed according to the type and importance of the variable. Appropriate treatment was applied where necessary.

Duplicate Removal

Duplicate records were checked to prevent the same patient from being counted multiple times. Patient identification information was used to help identify repeated records.

Data-Type Correction

Numerical fields were converted into appropriate numerical formats, while date fields were

converted into proper date formats. This allowed calculations and time-based visualizations.

Data Standardization

Categorical values such as State, City, Gender, Severity, and Vaccination Status were checked for inconsistent spelling, capitalization, and unnecessary spaces.

Data Validation

Numerical variables such as age, temperature, oxygen level, and recovery duration were checked for invalid or unrealistic values.

Geographical Findings

Geographical analysis revealed differences in the number of patient records among states and cities.

The **Cases by State** visualization shows that Tamil Nadu has the highest displayed number of cases among the selected states, followed by Telangana and Delhi. Other states represented in the visualization include Gujarat, Kerala, Rajasthan, West Bengal, Uttar Pradesh, Maharashtra, and Karnataka.

At the city level, the dashboard displays patient records for cities including Hyderabad, New Delhi, Jaipur, Chennai, Bengaluru, Kolkata, Ahmedabad, Surat, Lucknow, and Mumbai.

Disease Severity Findings

Disease severity analysis categorizes patients into **Mild, Moderate, Severe, and Critical** groups.

The findings indicate that **Mild cases form the largest severity category**, followed by Moderate and Severe cases. Critical cases represent the smallest category among the displayed groups.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Hospital and Healthcare Findings

The Hospital Dashboard provides information about healthcare utilization through KPIs and visualizations. The dashboard includes **Isolation, Ventilators, Tests, and Recovered** indicators.

The Patients by Hospitals chart shows differences in patient records among hospitals such as **Bombay Hospital, Care Hospitals, Ruby Hall Clinic, AIIMS Delhi, and Apollo Hospitals**.

Recovery Findings

Recovery is one of the major outcomes analyzed in the project. The dataset contains recovery status, recovery date, and recovery time in days.

The dashboard shows **946 recovered records** in the displayed KPI. The Average Recovery Days visualization also demonstrates that recovery durations vary between patients.

CHAPTER 6 : CONCLUSION

In conclusion, the **COVID-19 Data Analytics project successfully demonstrates how data analytics can be applied to healthcare information to produce meaningful and interactive insights**. Starting with a dataset containing 1,000 patient records, the project followed a structured process of data preparation, preprocessing, exploratory analysis, visualization, and dashboard development.

The analysis learned multiple dimensions of COVID-19 information, which contain patient demographics, age, gender, geographical distribution, disease severity, blood groups, hospital activity, ICU records, ventilator requirements, recovery, mortality, testing, and vaccination

The project demonstrates that raw data becomes significantly more useful when it is properly cleaned, structured, analyzed, and visualized. KPI cards provide quick summaries, charts reveal patterns, and interactive slicers allow users to investigate specific geographical areas. Together, these features create a comprehensive analytical environment.

Most importantly, the project establishes a strong foundation for future healthcare analytics. While the current work focuses primarily on descriptive and exploratory analysis, the same dataset and analytical framework can be extended toward predictive modelling, machine learning, statistical research, and advanced decision-support systems.

Overall, the project achieves its primary goal of demonstrating the practical application of **data analytics and Power BI in understanding COVID-19-related healthcare data**. It shows how a structured analytical methodology can transform complex patient and healthcare information into clear, organized, and interactive insights, while also highlighting the importance of data quality, appropriate visualization, and responsible interpretation of analytical results.

REFERENCES

1. World Health Organization. (2023). *WHO Coronavirus (COVID-19) dashboard*. World Health Organization. WHO COVID-19 Dashboard
2. World Health Organization. (2024). *COVID-19 epidemiological update – 24 December 2024*. World Health Organization. WHO Epidemiological Update
3. World Health Organization. (2024). *COVID-19 epidemiological update – 15 July 2024*. World Health Organization. WHO Epidemiological Update – July 2024
4. Ministry of Health and Family Welfare, Government of India. (2026). *COVID-19 India dashboard*. Government of India. COVID-19 India Dashboard
5. Ministry of Health and Family Welfare, Government of India. (2023). *COVID-19 guidelines and notifications*. Government of India. COVID-19 Notifications
6. World Health Organization. (2023). *WHO COVID-19 dashboard: About the data*. World Health Organization. WHO COVID-19 Dashboard – About
7. World Health Organization. (2023). *WHO COVID-19 dashboard: COVID-19 deaths*. World Health Organization. COVID-19 Deaths Data