IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# "Global Covid 19 Data Analysis And Visualization"

**[1]Prof. S. P. Gunjal, [2]Tanaya Sandbhor, [3]Sadichha Talekar, [4]Nisha Tekade, [5]Tejaswini Pawar**

[1] Assistant Prof SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra

[2,3,4,5] UG students SKN Sinhgad Institute of Technology & Science, Lonavala, Department of Computer Science

***ABSTRACT***: *Effective data analysis is essential for driving strategic planning and optimizing operational efficiency, particularly within dynamic fields such as global public health. This project presents a structured, end-to-end data science framework for the analysis, processing, and visualization of global COVID-19 statistics to derive actionable insights for stakeholders. The methodology encompasses acquiring data from diverse sources, including CSV/Excel files, SQL/NoSQL databases, and web APIs/scraping. Rigorous preprocessing was implemented, including Mean/Median/Mode imputation for missing values, Z-score/IQR outlier detection, and Min-max scaling for normalization The core analysis leverages Python libraries (Pandas for manipulation, NumPy for numerical computation, Matplotlib/Seaborn for EDA) and SQL for efficient data management and targeted retrieval. Finally, interactive dashboards were developed using Microsoft Power BI and Tableau to visualize key pandemic metrics such as confirmed cases, recovery rates, age distribution of cases, and vaccination progress. The resulting framework successfully demonstrates how data analysis transforms raw health data into crucial information for understanding global health patterns and guiding strategic interventions.*

**Keywords**

COVID-19, Data Analysis, Data Visualization, Python, Pandas, SQL, Microsoft Power BI, Tableau, Business Intelligence.

## I. INTRODUCTION :

Data analysis is a fundamental skill that drives decision-making across all industries, including finance, healthcare, and technology. In the digital age, organizations generate vast amounts of data, and analyzing this data is crucial for uncovering patterns, trends, and insights that inform strategic planning and business success. Effective data analysis is vital for enhancing operational efficiency, improving decision-making accuracy, and supporting innovation by predicting future trends.

This project focused on applying these principles to the Global COVID-19 pandemic data, providing hands-on experience in handling real-world datasets. The key objectives of this study were to:

Develop a pipeline for acquiring and preprocessing data from multiple sources, including databases, APIs, and web scraping.

Implement data cleaning techniques to ensure data integrity, focusing on handling missing values, removing duplicates, and standardizing data formats.

Utilize Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn for data manipulation, statistical analysis, and visualization.

Develop interactive dashboards using tools like Power BI and Tableau to enable stakeholders to explore and interact with the data.

Demonstrate how data analysis guides strategic decision-making in areas like resource allocation and performance tracking.

This report covers the complete workflow of data analysis, from raw data collection to visual storytelling through dashboards, detailing the methodologies and tools applied.

## II. LITERATURE SURVEY

The vast literature on COVID-19 underscores the multifaceted nature of pandemic research, spanning epidemiology, data science, big data analytics, visualization, forecasting and health-policy intervention analysis. Below is a summary of major strands relevant to our project. 1. Data Analytics & Visualization of COVID-19 Researchers have used data analytics and visualization to map the spread of COVID-19 across countries, regions and time periods. For example, Data analytics for novel coronavirus disease (COVID-19) examined different aspects of COVID-19 and presented visualisation of the infection spread globally. Another study, Data analysis and modeling of COVID-19, analysed trends of the spread both globally and specifically for India, using graphs and modelling. These studies demonstrate the value of combining descriptive statistics (cases, deaths, recoveries) with dynamic visual elements (maps, time-series, heat-maps) to offer insight beyond raw numbers. 2. Forecasting & Predictive Modelling Forecasting future waves or case counts has been a major focus. For example, the paper COVID-19 prediction models: a systematic literature review reviewed multiple forecasting approaches (statistical, machine learning, time-series) for COVID-19. Another work, Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data- driven analysis, presented short-term forecasts for multiple countries and assessed fatality-rate risk drivers. These papers highlight methods such as ARIMA, wavelet transforms, regression trees and hybrid models to forecast case counts; however they also emphasise limitations: short forecast horizon, data quality issues, variability across countries. 3. Big Data, Machine Learning & Healthcare Systems The rise of big-data analytics in the healthcare sector has been applied to COVID-19 in multiple ways: detection, prediction, pattern recognition. The review A Systematic Literature Review and Future Perspectives for Big Data in COVID-19 highlighted the role of big data technologies to detect outbreaks and monitor community health.Additionally, AI/ML-based approaches for diagnosis (e.g., chest x-ray/CT images) have been surveyed in the review Diagnosis of COVID-19 Using Machine Learning and Deep Learning: A review. 4. Data Quality, Reporting & Limitations High-quality data is foundational: many studies emphasise issues of missing values, under reporting, differences in country reporting standards, delayed updates and inconsistent variable definitions. For example, in the review Coronavirus Disease 2019 (COVID-19): A Literature Review, authors summarise challenges in diagnosis, treatment and data completeness. Such limitations have implications for any global dataset: cross-country comparisons must account for reporting bias, testing capacity divergences and variations in definitions (e.g., death “with” vs “from” COVID-19). 5. Interactive Dashboards & Decision-making Tools Early in the pandemic, the dashboard created by Johns Hopkins COVID-19 Dashboard became a widely-cited real-time tracking tool. Similarly, the literature acknowledges that visualisation platforms help non-technical stakeholders interpret complex data quickly. This project builds on that by creating interactive dashboards (using Power BI/Tableau) that support filtering, drill-down, region/time slicing and dynamic insights.

## III. PROBLEM STATEMENT

The outbreak of the COVID-19 pandemic generated massive volumes of heterogeneous data worldwide, including infection rates, mortality statistics, vaccination progress, demographic impacts, and regional transmission patterns. However, this data was dispersed across multiple formats such as CSV files, government dashboards, APIs, databases, and online sources, making it difficult for researchers, policymakers, and health organizations to derive meaningful insights in real time.

A major challenge lies in *collecting, cleaning, integrating, and analyzing* this multi-source data to extract reliable trends and support evidence-based decision-making. Due to missing values, inconsistent reporting formats, outliers, and rapid data fluctuations, the absence of a structured analytical framework often leads to inaccurate interpretation and delayed responses.

Therefore, there is a need to design a *systematic data analysis and visualization framework* that can preprocess raw global COVID-19 data, perform exploratory and statistical analysis, and present results through interactive dashboards. Such a framework would help stakeholders monitor the pandemic’s progression, identify high-risk regions, track vaccination effectiveness, and assess public health strategies, thus enabling timely and informed decision-making.

## IV. PROPOSED METHODOLGY

The methodology adopted in this study follows a sequential yet iterative workflow consisting of dataset acquisition, preprocessing, exploratory analysis, database integration, dashboard development, insight generation, and documentation. COVID-19 data were collected from reliable global repositories such as WHO, Our World in Data, and national public-health portals to ensure accuracy and completeness. The collected datasets underwent extensive preprocessing involving removal of duplicates, handling of missing values, normalization of key indicators (such as cases per million and vaccination percentages), alignment of date formats, and standardization of region names to maintain continuity in time-series analysis. Exploratory Data Analysis (EDA) was conducted using Python libraries including Pandas, NumPy, Matplotlib, Seaborn, and Plotly to generate descriptive statistics, visualize weekly and cumulative trends, construct choropleth maps, and compute correlations between variables such as vaccination rates and case reduction. A relational

SQL database (MySQL/PostgreSQL) was designed to store cleaned data using structured tables for country information, time-series records, and metadata, followed by optimized SQL queries to extract key metrics like top affected countries, vaccination thresholds, and weekly case variations. Interactive dashboards were developed in Power BI and Tableau by importing SQL/CSV data and creating dynamic visualizations such as time-series graphs, heatmaps, and geographical maps with user-controlled filters and tooltips. Insights were derived by examining dashboard outputs to identify patterns such as regions with high vaccination yet rising cases, low vaccination with high mortality, and time-lag effects between vaccination rollout and decline in cases, leading to evidence-based recommendations for resource allocation and monitoring strategies. Finally, comprehensive documentation—including user guides, code explanations, dataset metadata, and optional web-deployment links—was prepared to ensure reproducibility, usability, and effective dissemination of the findings

## V. CONCLUSION

The "Global Covid-19 Data Analysis and Visualization" project successfully demonstrated the effectiveness of a structured data science workflow, integrating Python for processing, SQL for robust data management, and Business Intelligence tools (Power BI, Tableau) for visual storytelling. The analysis provided meaningful insights by identifying the most affected countries, regions with the highest recovery counts, and overall pandemic trends through filtering, grouping, and aggregation queries. This framework is invaluable for transforming raw data into actionable information, supporting informed decision-making in healthcare planning, resource allocation, and guiding future health strategies.

## VI REFERENCES

[1]Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases, 20(5), 533–534.

[2]Kucharski, A. J., et al. (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. The Lancet Infectious Diseases, 20(5), 553–558.

[3]Zhou, F., et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet, 395(10229), 1054–1062.

[4]Wu, J. T., Leung, K., & Leung, G. M. (2020). Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak. The Lancet, 395(10225), 689–697.

[5]Li, Q., et al. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New England Journal of Medicine, 382(13), 1199–1207. Add all this contain in wor