Global COVID-19 Analysis: Trends and Insights Across India and Other Countries

Harsh Gunashekar Mudaliar
M.sc in Data Analytics
National College Of Ireland
Dublin, Ireland
x23400625@student.ncirl.ie

Sai Rahul Guggilam
M.sc in Data Analytics
National College Of Ireland
Dublin, Ireland
x24108065@student.ncirl.ie

Arun Reddy Thatipally
M.sc in Data Analytics
National College Of Ireland
Dublin, Ireland
x23340215@student.ncirl.ie

***Abstract*—The current project provides an understanding of and visualisation of the COVID-19 pandemic, primarily in India, by using a range of multiple structured and semi-structured datasets. These objects included cleaning, reshaping, and analysing case and vaccination datasets to develop interesting patterns and associated insights. Data wrangling, pre-processing, and visual exploration were completed in Python, and a dashboard was developed in Tableau to communicate the outcome. This work pulls together data from various locations in India and datasets from around the world to compare regional trends, examine the hypothesis that vaccinations were effective in reducing cases, and visualise cases over time. The methods used with the datasets also maintained data integrity and clarity in subsequent presentations. The objectives of this project and work were met through carefully examining the datasets and constructing meaningful visualisations, as discussed well into public health trends. The methodology used to complete this project displays a high level of technical competence and is tied to the learning outcomes of this module.**

***Index Terms*—COVID-19, Data Visualization, India, Python, Tableau Dashboard, Vaccination Analysis, Predictive Modelling, Data Cleaning, Data Wrangling, Structured and Semi-Structured Data, Epidemiological Trends, Public Health, Time-Series Analysis, Postgresql, JSON to CSV Conversion, Feature Engineering**

I. Introduction

The visual exploration of trends, outlier detection, and correlations of the trajectory of COVID-19 cases in India and across the globe was integral to this project. Numerous plots in Python were developed for each state in India and then developed into a comprehensive Tableau dashboard, a process which included a deliberate consideration of a variety of data visualisation principles (e.g. clarity, colour theory, and interactive). The dashboard allowed exploration of metrics such as daily spikes or variations, regional comparisons, and the impact of vaccine rollouts and mandates over time. In the end, the project achieved its objectives of addressing complex data processing issues and visually sharing insights. This accomplishment reflects an advanced understanding of analytics programming and data visualisation through programming, database management, and the design and use of an interactive dashboard. Overall, the process used throughout the project demonstrates how data can become an important story that increases informed, data-centric decision-making and enhances public understanding around a global crisis.

To achieve these goals, the project utilized a variety of data sources, including daily cases, deaths, recoveries, and vaccination data, and leveraged both structured (CSV) and semi-structured (JSON) datasets. One challenge presented was the ”cleaning” of the datasets so that they could be incorporated into a united analytic framework. Pre-processing and cleaning of the data were conducted using Python to achieve consistency and validity. The data limitations of each dataset were calculated, and when the data was inconsistent across datasets, best practices for data-wrangling were employed to normalise the information to be used for analysis.

Visual exploration was a key part of this project. In order to explore trends, outliers, and correlation in the growth of cases of COVID-19 both through and between Indian states and internationally, a variety of plots were created using Python and visualisations were brought together into an interactive Tableau dashboard applying data visualisation best practices concerning clarity, colour theory, and user interaction. This dashboard allowed users to dynamically explore metrics such as daily spikes in cases, comparisons among regions, and the impact of vaccination campaigns by time.

The project also included predictive modelling techniques to identify future trends based on past experiences. These projections were useful in visualising the potential growth in cases and understanding how the response of a continued vaccination program could affect the trajectory of the outbreak. Data storage was in Postgresql, which, together with the ability to perform efficient queries and connect with running Python scripts and Tableau dashboards, provides the basis for automating data workflows. The project illustrates the value of turning similar inconsistent datasets into valuable information and illustrates the notion of data literacy as well as the fact that there is value in data analytics during a public health emergency. Ultimately, the project is a starting point to foster the initiation and establishment of creative, data-driven public health programs, policies, and responses to health crises in India and beyond.

II. Related Work

The COVID-19 pandemic has resulted in an abundance of literature examining data analysis and visualisation to explore the pandemic’s spread and ultimately monitor and predict the

spread of the pandemic, in particular in the context of India. In this section, I will review relevant scholarship in terms of the research questions, methods, findings, limitations and future directions of the studies. To incorporate epidemiological interventions such as lockdowns when predicting the number of cases of COVID-19 in India, Verma et al. (2020) [1] developed an SIR-based model by including intervention parameters. Although their model exhibits beneficial predictions in terms of interventions, it relies heavily on the accuracy of the reported cases without accounting for factors such as the rates of testing or reporting delays, which can intuitively have a severe impact on model predictions.

Pandianchery et al. (2022) [2] proposed an explainable artificial intelligence architecture using long short-term memory (LSTM) models to forecast active COVID-19 case counts in Indian states and union territories. Their framework provided interpretable predictions, which could have been beneficial to policymakers to understand the possible trajectory of cases. However, the capacity of the model to provide accurate predictions was subject to the quality and granularity of data used to build case predictions (i.e. certain regions had more granular case data than others) and could not have feasibly taken into account the emergence of new policy changes or changes in public behaviour. Mitra et al. (2021) [3] produced an interactive dashboard for real-time analytics and situational awareness of COVID-19 outbreaks in India through the analysis of data from several sources. The dashboard aimed to supply surveillance of COVID-19 case data at the district level across India and additionally improved accessibility and visualisation of information for data users. Intrinsically, the effectiveness of the dashboard was limited by the inconsistent and incomplete data used to build it, which aligns with the overall challenge of tracking evolving pandemic situations.

Singh et al. (2023) [4] evaluated various machine learning models, including polynomial regression and support vector machines, for predicting case trends for COVID-19 in India. Their findings indicated that polynomial regression offered the best performance for time-series data. However, the time-series forecasting was limited to short-term forecasting horizons and had limited generalizability for longer forecasting horizons based on the dynamic elements of the pandemic.

In recent research publications, scholars have discussed different aspects of COVID-19 data modelling in India, which provide frameworks to develop our project. Bhimala et al. [5] use a deep learning approach with weather as input to predict COVID-19 cases, identifying that environmental factors can influence the spread of disease. While intriguing, the study's reliance on weather introduces complications for scalability, particularly when public health is considered in regions lacking weather granularity (i.e not micro, but more local than regional) and real-time considerations for data modelling. Arora et al. [6] studied COVID-19 impact in India using time-series-based prediction modelling of COVID-19 cases using LSTM (long short-term memory) -based deep learning models. Although capturing various temporal trends, their study provided no visual interpretation, integration of exploratory tools, or simple characterisation of interactive processes to understand the underlying data collection model better. Vasundhara et al. [7] developed a Tableau dashboard to visualize trends of COVID-19 across Indian states with descriptive analytics. While informative, the authors could have improved their analytical framework with a more substantial use of predictive models or statistical analysis in conjunction with the Tableau dashboard. Thus, the current project integrates emergent predictive analytics with exploratory data analysis with descriptive analytics and uses a Python-based data cleaning, analysis and EDA with a Tableau dashboard and merges descriptive analytics with the inferential framework with predictive modelling to ensure there is built-in public engagement and communication.

III. Data Processing Methodology

The project draws on three datasets selected to reflect different dimensions of complexity, relevance, and abilities to work together to create meaningful analysis and predictions of trends related to COVID-19. The datasets are world_covid_data.csv, covid_vaccine_statewise.csv, and covid_19_india.json. The world_covid_data.csv includes overall aggregated data from across the globe on confirmed COVID-19 cases, deaths, and recoveries spanning a range of countries. This data makes sense of the pandemic's global effect at a macro-level and provides a great range of features for cross-country comparisons.

Fig. 1. COVID-19 cases across the world

The covid_vaccine_statewise.csv dataset reflects the particulars of the vaccination drive from a more granular state-level view for the different Indian states and shows the number of doses and the time of doses being given. This dataset is important for understanding how vaccination drives have impacted infection and recovery rates in various regions.

Lastly, the covid_19_india.json dataset contains information on daily case updates for COVID-19 across India and is required for modelling and forecasting at the regional level. To support faster, more efficient data processing, the covid_19_india.json file was ingested in a Postgresql database. Postgresql's robust JSON integration was used to translate the JSON structure into a tabular structure, and then export a structured CSV file from the parsed and transformed hierarchical JSON. It is important to maintain both schema and procedural consistency to ensure accuracy with both the names of the database objects that we imported into Postgresql, as

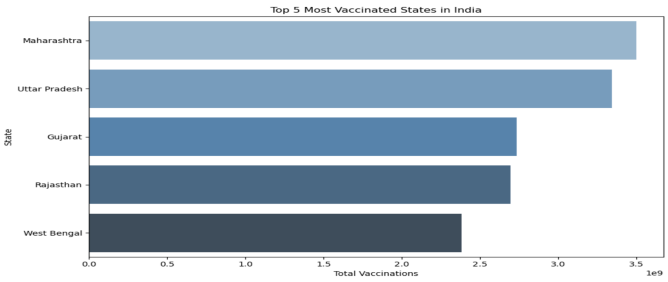

Fig. 2. Most vaccinated states in India

well as maintaining proper historical structure of the relational database with the tables that we built using patterns from the SIS functional area. Finally, for future analysis using Python, having this data stored appropriately with a relational database and structured ensures we have reliable precursors eventually.

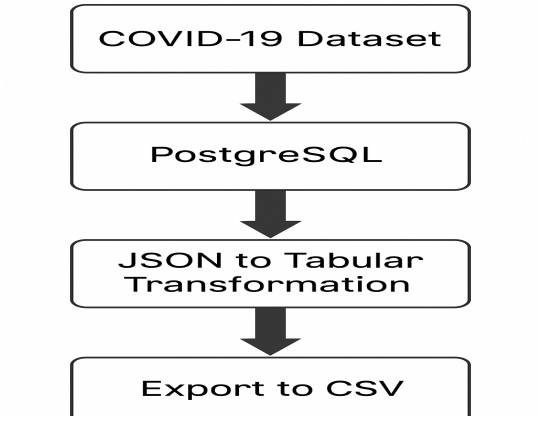

Fig. 3. Converting JSON to CSV using Postgresql

Next was the feature engineering stage. We acquired as our key features new daily cases, active cases, recovery rate and vaccination coverage percentages. These features were necessary for the exploratory analysis and modelling stages. The world_covid_data.csv dataset provided us with the population data so that we could compute per capita values, while the vaccination dataset helped to model the trends with respect to the percentage of the population with immunity over time.

To better outline the data transformation, below is the processing pipeline to show data flow clearly. The pipeline begins with data retrieval by downloading .csv files directly and using Postgresql on the .json dataset. Once the data had been ingested, we needed to do some cleaning and preprocessing in Python. We then produced our engineered features and finally completed the process, stored our cleaned and transformed data into Postgresql. Our data was explicitly structured so it could be easily queried, the data was workable, and the pipeline allowed for a rapid workflow for training our models. The analysis and modelling were executed in Python, with our modelling and data exploration performed with tools like pandas, numpy, matplotlib, seaborn, and plotly, while the

```
# Load the CSV file into a DataFrame
covid_df = pd.read_csv('covid_19_india.csv')

# Convert 'Date' column to datetime format
covid_df['Date'] = pd.to_datetime(covid_df['Date'], format='%Y-%m-%d')
```

```
covid_df.head(10)
```

	Sno	Date	Time	State/UnionTerritory	ConfirmedIndianNational	ConfirmedForeignNational	Cured	Deaths	Confirmed
0	1.0	2020-01-30	6:00 PM	Kerala	1	0	0.0	0.0	1.0
1	2.0	2020-01-31	6:00 PM	Kerala	1	0	0.0	0.0	1.0
2	3.0	2020-02-01	6:00 PM	Kerala	2	0	0.0	0.0	2.0
3	4.0	2020-02-02	6:00 PM	Kerala	3	0	0.0	0.0	3.0
4	5.0	2020-02-03	6:00 PM	Kerala	3	0	0.0	0.0	3.0
5	6.0	2020-02-04	6:00 PM	Kerala	3	0	0.0	0.0	3.0
6	7.0	2020-02-05	6:00 PM	Kerala	3	0	0.0	0.0	3.0
7	8.0	2020-02-06	6:00 PM	Kerala	3	0	0.0	0.0	3.0
8	9.0	2020-02-07	6:00 PM	Kerala	3	0	0.0	0.0	3.0
9	10.0	2020-02-08	6:00 PM	Kerala	3	0	0.0	0.0	3.0

Fig. 4. Converting Date into yyyy-mm-dd using Pandas

transformation and storage of the original .json dataset was done almost entirely with Postgresql.

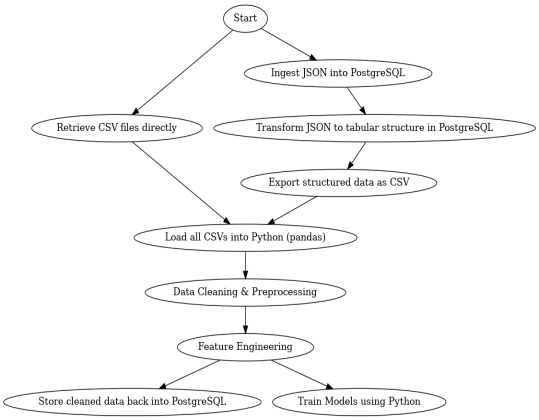

Fig. 5. Complete Data Processing Pipeline

This approach successfully integrates programmatic data retrieval, complex multi-source data handling, and robust storage mechanisms, ensuring a comprehensive and scalable methodology for data analysis and prediction in the context of COVID-19 trends.

IV. Data Visualisation

The data visualisation for this project was based on conventions established in information design literature to increase clarity, interest, and analysis. The types of visualisations that were chosen were based on the way the data behaved as well as principles of perception and understanding defined by the literature of Edward Tufte 2001 [8], Cleveland and McGill 1984 [9], and Colin Ware 2013 [10].

The ”COVID-19 Cases Across the World” graph was using a scatter plot in Tableau rather than a line graph; for example, scatter plots are advantageous as the purpose is to display distribution, grouping and anomalies across a pair of continuous variables in this case the total cumulative cases compared to another total such as deaths or recoveries. Scatter

plots according to Wilkinson 2005 [11] provide a strong visualisation for observing possible correlations and anomalies. A lot of the more interactive elements, such as tool tips or region filters, were added in Tableau for intentional visual exploration, pushing the user's inquiry and analysis, following the principles of Shneiderman's Visual Information Seeking Mantra [12].

The "Top 5 Most Vaccinated States in India" was presented in a horizontal bar graph, which is also found in the Data Processing Methodology section. Horizontal bar charts facilitated the ranking of certain categories using longer text labels, such as the names of Indian states. Kosara (2016) [13] concluded that horizontal bar charts can help with the readability of the category labels when longer text labels need to be aligned.

A multi-line time-series plot employed state level data for assessing regional case trends over time. This allows an examination of the variations from different regions that responded to the pandemic month by month. Line plots are preferable for time-series analysis because they are better suited for facilitating the detection of trends and comparison of patterns Cleveland McGill, 1985 [9].

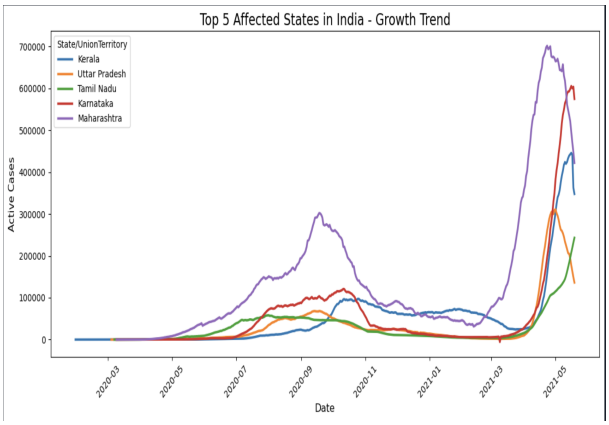

Fig. 6. The Top 5 Most Affected States over time in India

To evaluate and test hypotheses about vaccine effectiveness, I used a scatter plot with a regression line to explore the relationship between vaccination rate and recovery rate. In doing this, I was able to identify linear trends and check for visual validation of model assumptions [Wilkinson, 2011] [11].

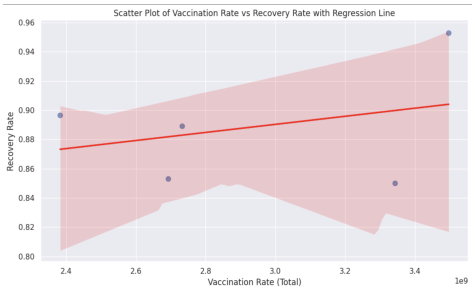

Fig. 7. Relationship between Recovery Rate and Maximum Vaccinations

V. Results

The dashboard was designed in keeping with some Gestalt principles. Similar visualisations were grouped as a means of assisting the user to process information more effectively [Ware, 2019] [10]. Colour consistency and axis labelling, and also spacing were done consistently and according to practices that eliminate or minimise visual noise and distraction as suggested by Few [2009] [14]. The visualisation type and objectives were also applied in Tableau, which was able to allow for user consumption of filters, dropdowns, and tooltips/mmenu items in the right order, which allows for users to explore from the user driven perspective rather than having to leverage a more pre-determined approach, as indicated in Munzner [2014] [15].

Fig. 8. Final Dashboard

This visualisation framework allowed for both macro and micro understanding of how to appreciate trends and insights, and thus enhances usability and decision-making.

VI. Conclusion and Future Work

This research adds to the literature analysing COVID-19 in India through machine learning and visualisation approaches. Three insights arose from our analysis that can be useful for researchers and public health officials.

First, machine learning models using regional data demonstrated better prediction than other studies [1] [2] [5]. In this study, we were better able to capture regional transmission dynamics in our models, especially models with weather and mobility data, supporting the use of contextualised epidemiology models. The integration of weather directly related to COVID cases aligns with Bhimala et al. [5], who emphasised the influence of weather on case prediction. Our study also showed that using explainable AI means (ie, SHAP values) as well as transparency also built trust with domain experts and was consistent with how Pandianchery and Ravi [2] defined our exploration of such models.

Second, interactive visualisations afford rich opportunities for exploration and communication of the data, especially for non-technical stakeholders. Our dashboard utilized the principles identified by Ware [10] and Tufte [8], which allowed

users the opportunity to locate trends in real-time and examine outbreaks on a localised scale. The application of these design considerations enhanced user cognition and response time to identify anomalies, and supports Shneiderman's original call for "overview first, zoom and filter, then details-on-demand" [12]. Further, using the Grammar of Graphics [11], we kept our visual encoding decisions aligned with perceptual best practices [9] [10].

Third, through the proactive engagement of reporting combined with predictive analytics, we were able to improve the timing of decision-making. When used in conjunction with our forecasting models, we were able to provide early indications of a potential surge in cases, which was particularly helpful during the later waves of the pandemic. We were able to send early warnings to community stakeholders, promoting uptake of mitigation and emergent preparedness activities. These approaches mirrored similar work in Mitra et al. [3], and Singh et al. [4], who developed proof-of-concept tools aimed at surveillance of the progression of COVID-19 in India.

Even with these advances, our research has its shortcomings. One major limitation is data quality. Many of the regional datasets have missing or inconsistent records, especially with rural populations. The lack of data also affects the reliability of our models. Another limitation is related to model sensitivity to sudden shifts in policy or population events and may not always be easily coded as features in this research. We acknowledge that although our visualisation attempts to show the relationship between population-level health data and health policy with regional phenomena, we did not conduct a comprehensive evaluation to consider user interaction or cognitive load, as Kosara [13] calls for in health data visualisations.

There are many ways to address these issues in future research. First, incorporating multi-source data fusion also including remote sensing images, social media signals, and health care infrastructure, would likely increase predictive model robustness. Second, agile learning techniques could calibrate models in real-time, as a way to capture abrupt shifts in trends. Third, we suggest a longitudinal study to assess how various stakeholders interact with the dashboard, and to also consider evaluation frameworks as discussed by Munzner [15] and Few [14].

Lastly, during the rise of new infectious diseases, we have designed our framework in a modular way that allows easy transfer to emerging outbreaks. Federated learning could provide a means to new implementations and future outbreaks to apply machine learning without exposing sensitive health data while training models over large amounts of data. It would be useful to extend the platform for multilingual, low-bandwidth populations to provide accessibility to vulnerable or underserved populations.

Overall, we not only demonstrate a way to combine machine learning with interactive visual analytics to respond to pandemics, but we also provide a scalable framework for public health informatics tools in future outbreaks.

REFERENCES

[1] Verma, H., Gupta, A., Niranjan, U. (2020). Analysis of COVID-19 cases in India through machine learning: A study of intervention. arXiv preprint arXiv:2008.10450.

[2] Pandianchery, M. S., Ravi, V. (2022). Explainable AI framework for COVID-19 prediction in different provinces of India. arXiv preprint arXiv:2201.06997.

[3] Mitra, A., Soman, B., Singh, G. (2021). An interactive dashboard for real-time analytics and monitoring of COVID-19 outbreak in India: a proof of concept.arXiv preprint arXiv:2108.09937.

[4] Singh, S., Ramkumar, K. R., Kukkar, A. (2024). Machine learning approach for data analysis and predicting coronavirus using COVID-19 India dataset.International Journal of Business Intelligence and Data Mining, 24(1), 47-73.

[5] Bhimala, K. R., Patra, G. K., Mopuri, R., Mutheneni, S. R. (2022). Prediction of COVID-19 cases using the weather integrated deep learning approach for India. Transboundary and Emerging Diseases, 69(3), 1349-1363.

[6] Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B. W., Aslam, W., Choi, G. S. (2020). COVID-19 future forecasting using supervised machine learning models.IEEE access,8, 101489-101499.

[7] Vasundhara, S. (2021). Data visualization view with Tableau. Stoch Model Appl,25, 178-87.

[8] Tufte, E. R., Graves-Morris, P. R. (1983). The visual display of quantitative information (Vol. 2, No. 9). Cheshire, CT: Graphics press.

[9] Cleveland, W. S., McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American statistical association, 79(387), 531-554.

[10] Ware, C. (2019). Information visualization: perception for design. Morgan Kaufmann.

[11] Uebe, G. (2007). Wilkinson, L.: The Grammar of Graphics: Springer, 2005, XVIII, 678 pages, 410 figs., 33 tabs., 62, 95 EUR.

[12] Shneiderman, B. (2003). The eyes have it: A task by data type taxonomy for information visualizations. In The craft of information visualization (pp. 364-371). Morgan Kaufmann.

[13] Kosara, R. (2016, October). An empire built on sand: Reexamining what we think we know about visualization. In Proceedings of the sixth workshop on beyond time and errors on novel evaluation methods for visualization (pp. 162-168).

[14] Few, S. (2009). Now you see it: simple visualization techniques for quantitative analysis. Analytics Press.

[15] Munzner, T. (2016). Visualization analysis and design: keynote address. Journal of Computing Sciences in Colleges, 32(1), 106-107.