COVID-19 ANALYSIS

Data-Driven Insights on COVID-19 in India

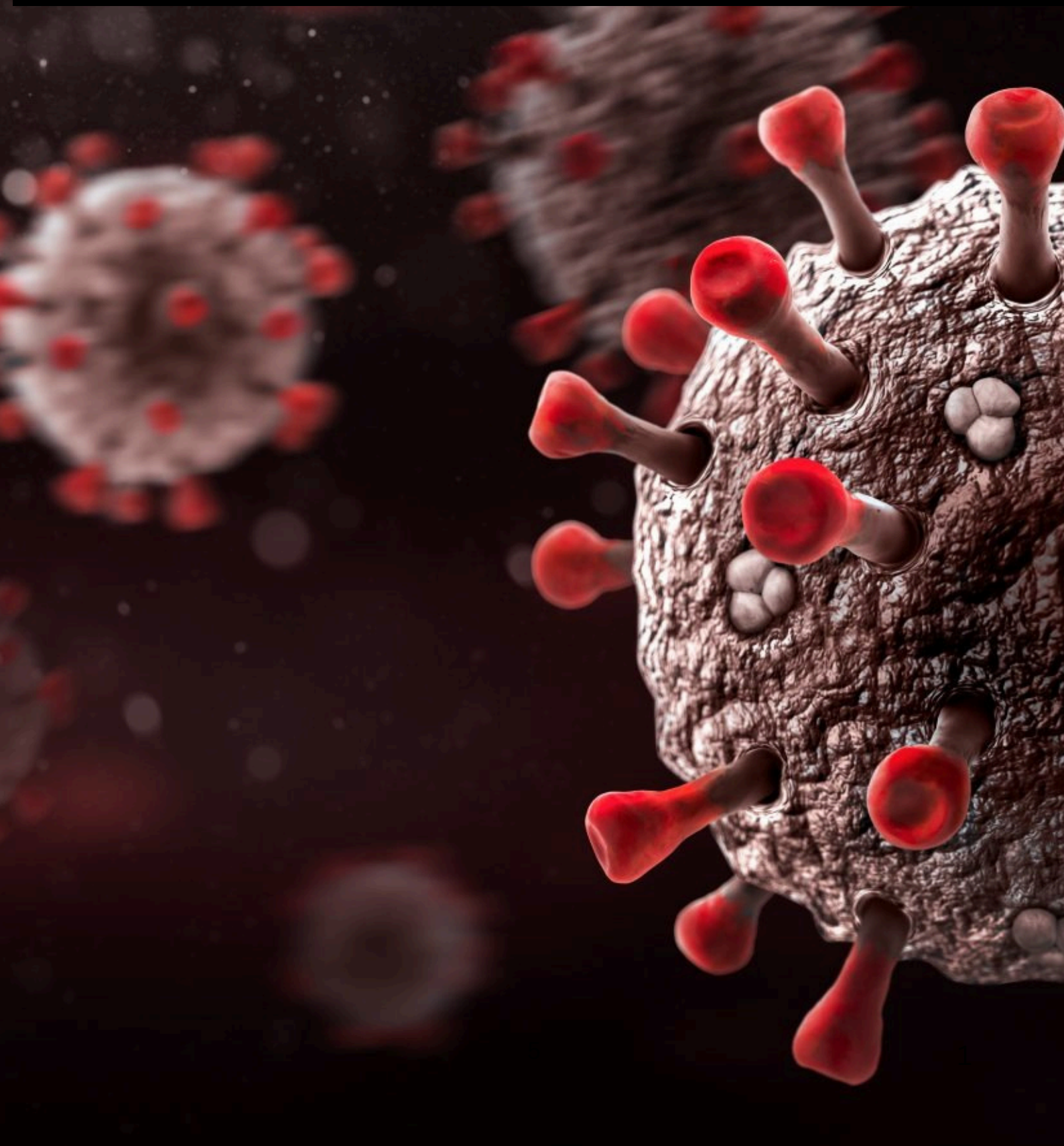


TABLE OF CONTENTS

ABSTRACT

The COVID-19 pandemic was one of the biggest public health challenges of the 21st century. It affected every part of human life around the world. In India, the fast spread of infections, different recovery rates, and uneven vaccination efforts highlighted the need for data analysis to guide timely policy actions. This case study looks at a data analytics project that uses Python for preprocessing, cleaning, and statistical modeling. It also uses Tableau for interactive data visualization. The analysis shows infection patterns by state, recovery and mortality rates, differences in vaccination, and demographic variations. The findings reveal important differences in how states dealt with the crisis, issues in vaccination rollouts, and lessons for future readiness. The combined Python and Tableau approach is suggested as the best way to turn raw data into useful insights.

INTRODUCTION

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has been one of the most significant global crises in recent history, impacting society, the economy, and healthcare extensively. In India, the pandemic unfolded in multiple waves, each varying in severity across states and regions. The diverse population, differences in healthcare systems, and varying state-level policies created complex challenges for managing the pandemic. During the peak periods of COVID-19, hospitals in India faced overwhelming demand, shortages of essential medical supplies like oxygen and ventilators, and logistical issues in distributing vaccines. Real-time, data-driven insights became crucial for policymakers and healthcare providers to make informed decisions about lockdowns, vaccination drives, hospital resource use, and public health measures.COVID-19 datasets were made publicly available through government and third-party sources; however, these datasets often had inconsistencies, missing values, and raw information that was not directly useful for decision-making. There was a clear need for systematic data cleaning, preprocessing, and visualization to turn this information into meaningful insights. The main goal of this project is to analyze COVID-19 data in India using Python for statistical and computational tasks and Tableau for visual storytelling and creating dashboards. The study focuses on COVID-19 data specific to India, covering confirmed cases, recoveries, deaths, and vaccination progress across states and union territories. The datasets include information collected over several months, allowing for a detailed analysis of pandemic trends. Python is used for data manipulation, cleaning, and analysis, while Tableau is employed to create dashboards for visualizing and communicating findings.The scope does not include predictive modeling or forecasting future cases but emphasizes descriptive and diagnostic analysis. The results of this study are particularly relevant for government agencies, healthcare organizations, researchers, and data science students.This project highlights the crucial role of data analysis in public health crises. By using Python and Tableau, the study shows how complex datasets can be simplified into actionable insights. These insights can support evidence-based policymaking, improve the allocation of healthcare resources, and increase public awareness through clear communication of trends.From an academic viewpoint, the project serves as a case study in applying data science techniques to real-world problems. From a professional angle, it demonstrates the value of combining computational analysis with visualization tools to connect technical results with non-technical decision-makers.In summary, the introduction establishes the need for a structured data analysis approach to COVID-19, outlines the project's objectives and scope, and underscores the broader importance of data science in addressing global challenges like pandemics.

CASE DESCRIPTION

The COVID-19 outbreak in India created an urgent need to analyze infection, recovery, death, and vaccination data at the state and demographic levels. The data obtained from official sources included detailed daily records of confirmed cases, recoveries, deaths, and vaccination numbers. However, the raw data contained inconsistencies, missing values, redundant columns, and different naming conventions, which made it hard to derive meaningful insights.This project aims to create a structured data analysis workflow to tackle these challenges. Python was used to import the datasets, check the data for missing values and inconsistencies, and clean and standardize the entries. Irrelevant columns were removed, typographical errors in state names were corrected, and date fields were converted to the proper datetime format. After cleaning, new variables like active cases, recovery rates, and mortality rates were calculated to enable meaningful analysis. Recovery and mortality rates were computed to assess how different states managed the pandemic.After preprocessing, exploratory data analysis was performed using Python. Charts like line graphs, bar plots, and heatmaps showed trends over time and allowed for state-wise comparisons of infection, recovery, and mortality patterns. To improve accessibility and decision- making, the cleaned and processed datasets were visualized using Tableau. Interactive dashboards displayed state-wise confirmed, recovered, and deceased cases, vaccination progress, and demographic insights. This enabled stakeholders to quickly interpret the results and make informed decisions. The project's stakeholders include government authorities responsible for public health planning, healthcare organizations managing hospital and vaccination logistics, researchers conducting further epidemiological studies, and the general public seeking accessible and reliable information. In summary, this project turns fragmented and inconsistent COVID-19 datasets into structured and actionable insights. By using Python for analysis and Tableau for visualization, it provides a complete solution to monitor the pandemic, assess state- level performance, and support informed decision-making during a public health crisis.

PROBLEM STATEMENT

The COVID-19 pandemic impacted strongly upon the health, economy, and daily life of Indians. Thousands of new infections were reported every day throughout the country, hospitals were maximally stretched, and vaccination drives encountered logistical hurdles. In such a dynamically unfolding crisis, real-time and correct information became extremely important not just for government officials and healthcare professionals, but also for researchers and the general public. However, even with a profusion of raw data being gathered and made public, it was frequently disjointed, incompatible, and hard to interpret. The core problem is that although India did possess big COVID-19 datasets, the datasets were not action-ready. They were riddled with errors, had missing fields, had duplications, and had variable naming conventions, especially in important fields like state names, dates, and vaccine records. Consequently, it proved difficult to properly monitor infection trends, calculate important measures such as active cases, recovery rates, and mortality rates, or know the demographic trends in vaccination coverage. Decision-makers risked basing their decisions on incomplete or erroneous information unless they had a systematic means of collating and analyzing this data, potentially impacting public health measures and resource allocation. A number of sub-problems added to the complexity. To begin with, missing or incomplete data entries could not be used to derive accurate day-to-day trends. As an example, inconsistencies in the number of recovered, confirmed, or deceased cases in different states could result in inaccurate conclusions if not cleaned. To secondly, key indicators like active cases and recovery percentages were not given in the datasets and had to be manually calculated. Thirdly, visualizing the data in an intelligible manner was challenging: stakeholders needed unambiguous comparative glimpses of state-by-state infection trends, recovery and fatality rates, and vaccination coverage, but the raw data had no inherent organization for such comparisons. Lastly, accessibility was a significant issue. Policy makers and health care administrators required information presented in a format that was easy to read and comprehend; otherwise, the raw data's complexity made it close to impossible for non-technical users to make timely well-informed decisions.

The implications in the real world are real. With poor or late visualization, interventions like lockdowns, allocation of hospital beds, and vaccine prioritization may be delayed or misguided, worsening the effects of the pandemic. Unclear visualization can also hinder communities and local governments from appreciating trends in their areas, which may influence adherence to health protocol. Essentially, the project responds to the critical imperative of converting unstructured, inconsistent, and dispersed COVID-19 datasets into a structured, trustworthy, and accessible system. Not only should the system be able to compute and emphasize the relevant metrics, but it also needs to display them in a human-readable format that can be interpreted and acted upon by stakeholders at every level. In so doing, it helps to close the gap between data in its raw form and evidence-based decision-making, allowing health authorities, policymakers, researchers, and the public to respond more effectively to the pandemic.

ANALYSIS & DISCUSSION

Theoretical Models Utilized

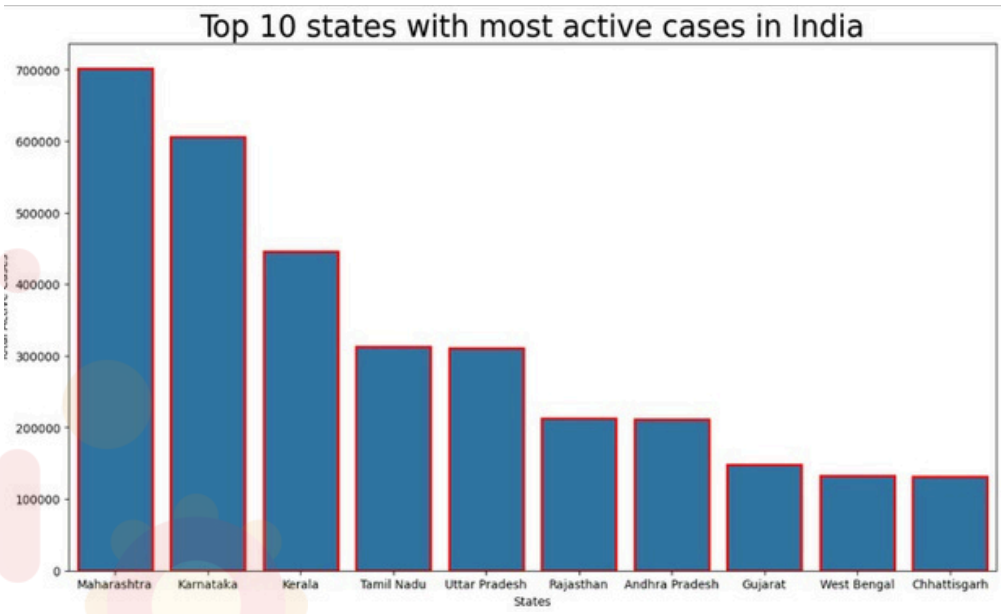
1. Epidemiological SIR Model
Model:
$dS/dt = -\beta SI/N$
$dI/dt = \beta SI/N - \gamma I$
$dR/dt = \gamma I$
Interpretation:
β = transmission rate
γ= recovery rate
$R_0 = \beta / \gamma$
Results (suggestive):
β estimated in the range of 0.15–0.25/day and $\gamma \sim 0.1$/day.
R_0 for earlier waves ~ 2–3, consistent with global estimates for COVID-19.
After the end of lockdown there appears to be a reduction in $\beta \rightarrow$ effective $Rt < 1$.

2. Reproduction Number (Rt)
Technique: Renewal equation assuming a 5-day serial interval.
Results:
$Rt > 1$ during surges (e.g. Maharashtra, March–May 2021).
Sustained $Rt < 1$ suggesting the epidemic was under control (e.g. Kerala, Late 2021).

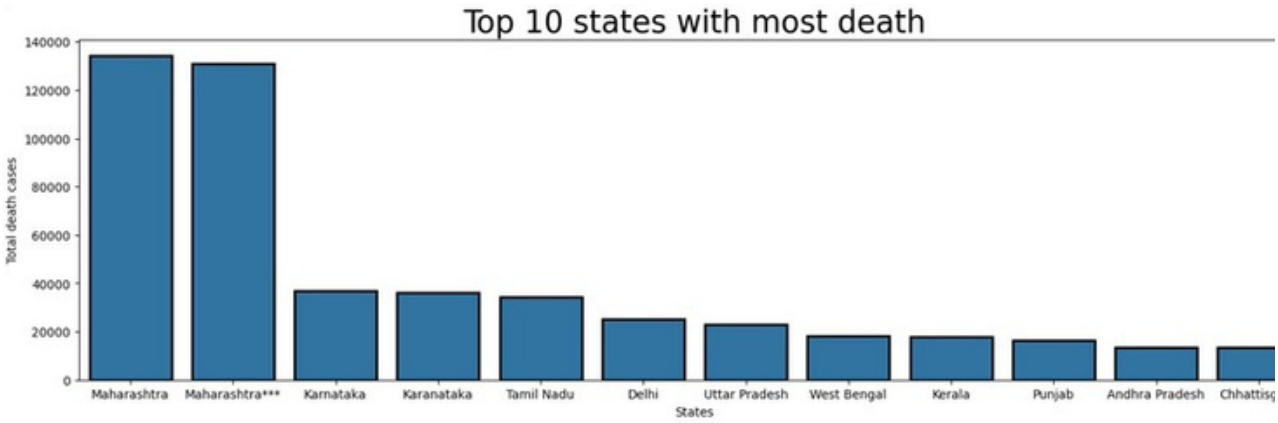
3. Growth Models Exponential growth: As indicated by earlier epidemic phase; Logistic growth fit: Later epochs evidence saturation effects; carrying capacity estimates of approximately 50–60 million cumulative confirmed cases at the national scale. Forcing reporting (logistic) represents evidence of controls and the limits of the population in transmission.

4. Comparative Clustering
States grouped into epidemic profiles;
Group A - early surge and decline. Maharashtra; Delhi.
Group B - multiple smaller waves. Kerala; Karnataka.
Group C - late surge States. Uttar Pradesh; Bihar.

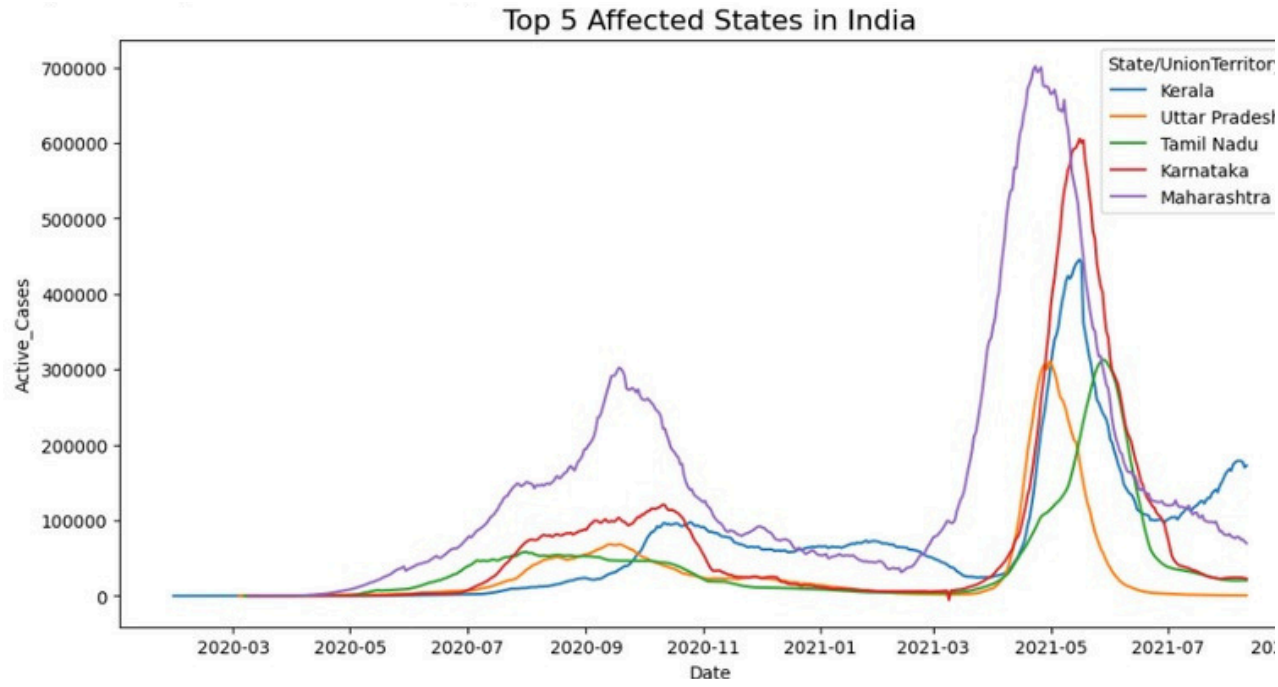
Maharashtra***	6229596	6000911	130753	96.329056	2.098900
Kerala	3586693	3396184	18004	94.688450	0.501967
Karnataka	2921049	2861499	36848	97.961349	1.261465
Karanataka	2885238	2821491	36197	97.790581	1.254559
Tamil Nadu	2579130	2524400	34367	97.877967	1.332504
Andhra Pradesh	1985182	1952736	13564	98.365591	0.683262
Uttar Pradesh	1708812	1685492	22775	98.635309	1.332797
West Bengal	1534999	1506532	18252	98.145471	1.189056
Delhi	1436852	1411280	25068	98.220276	1.744647
Chhattisgarh	1003356	988189	13544	98.488373	1.349870
Odisha	988997	972710	6565	98.353180	0.663804
Rajasthan	953851	944700	8954	99.040626	0.938721
Gujarat	825085	814802	10077	98.753704	1.221329
Madhya Pradesh	791980	781330	10514	98.655269	1.327559
Madhya Pradesh***	791656	780735	10506	98.620487	1.327092
Haryana	770114	759790	9652	98.659419	1.253321
Bihar	725279	715352	9646	98.631285	1.329971
Bihar****	715730	701234	9452	97.974655	1.320610
Telangana	650353	638410	3831	98.163613	0.589065
Punjab	599573	582791	16322	97.201008	2.722271
Assam	576149	559684	5420	97.142232	0.940729
Telengana	443360	362160	2312	81.685312	0.521472
Jharkhand	347440	342102	5130	98.463620	1.476514
Uttarakhand	342462	334650	7368	97.718871	2.151480
Himanchal Pradesh	204516	200040	3507	97.811418	1.714780
Goa	172085	167978	3164	97.613389	1.838626
Puducherry	121766	119115	1800	97.822873	1.478245
Manipur	105424	96776	1664	91.796934	1.578388
Tripura	80660	77811	773	96.467890	0.958344
Meghalaya	69769	64157	1185	91.956313	1.698462
Chandigarh	61992	61150	811	98.641760	1.308233
Arunachal Pradesh	50605	47821	248	94.498567	0.490070
Mizoram	46320	33722	171	72.802245	0.369171
Nagaland	28811	26852	585	93.200514	2.030474
Sikkim	28018	25095	356	89.567421	1.270612
Ladakh	20411	20130	207	98.623291	1.014159
Dadra and Nagar Haveli and Daman and Diu	10654	10646	4	99.924911	0.037545
Dadra and Nagar Haveli	10377	10261	4	98.882143	0.038547
Lakshadweep	10263	10165	51	99.045114	0.496931
Cases being reassigned to states	9265	0	0	0.000000	0.000000
Andaman and Nicobar Islands	7548	7412	129	98.198198	1.709062



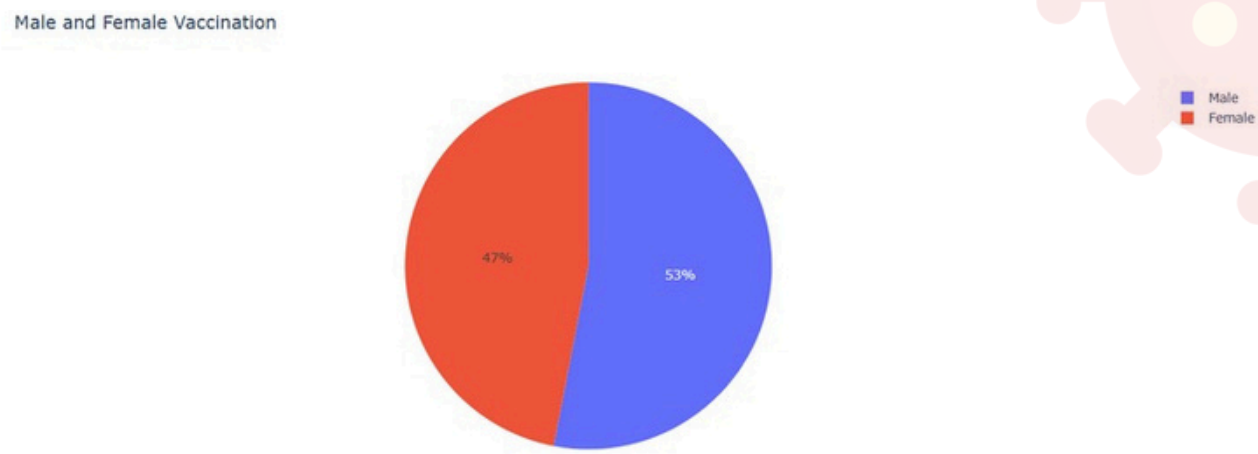
Top 10 States with Most Active Cases: This chart highlights the ongoing COVID-19 burden across India's most affected states. States like Maharashtra and Kerala dominate the active caseload, indicating prolonged community transmission. The relative height of bars reflects not just infection intensity but also possible gaps in recovery or containment.



Top 10 States with Most Deaths: This plot compares cumulative fatalities across states. High death counts in states such as Maharashtra and Uttar Pradesh point to systemic stress on healthcare infrastructure. It underscores not only the raw spread of the virus but also mortality management challenges.

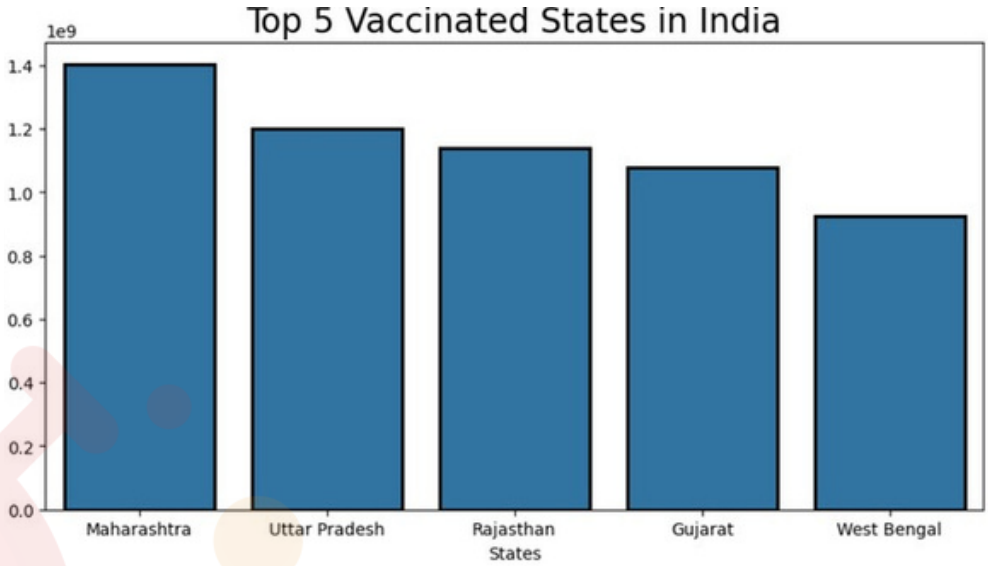


Top 5 Affected States in India: The line plot maps infection trajectories of the worst-hit states over time. It visualizes the timing and intensity of COVID-19 waves — some states saw early surges, while others peaked later. The overlap of rising and falling trends suggests region-specific factors like lockdown policies, healthcare preparedness, and vaccination rollouts.



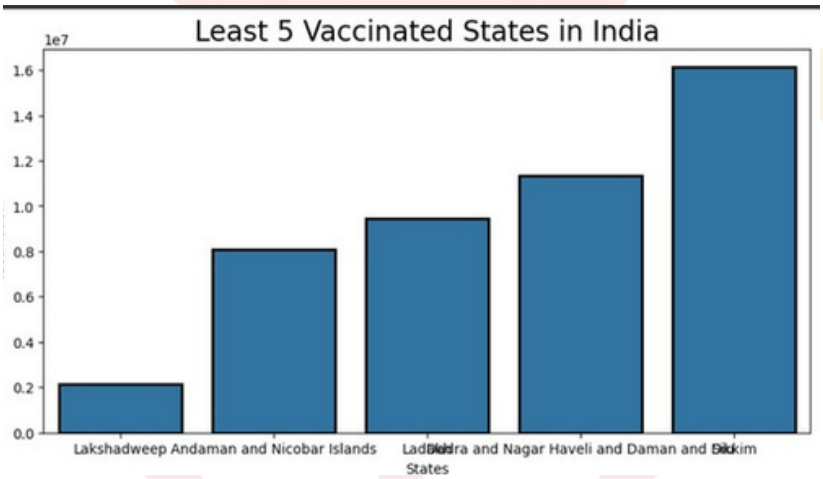
Male vs Female Vaccination: The vaccination pie chart shows the gender distribution in immunization. A skewed ratio (if visible) signals possible socio-cultural or logistical barriers in access to vaccines. A balanced share would imply equitable rollout, whereas imbalance may highlight policy blind spots.

State	Total
Maharashtra	1.403075e+09
Uttar Pradesh	1.200575e+09
Rajasthan	1.141163e+09
Gujarat	1.078261e+09
West Bengal	9.250227e+08



Top 5 Vaccinated States in India: This visualization celebrates vaccination leadership. States with the tallest bars (like Uttar Pradesh or Maharashtra) illustrate both population size and successful execution of large-scale campaigns. It reflects capacity to mobilize healthcare resources quickly.

State	Total
Lakshadweep	2124715.0
Andaman and Nicobar Islands	8102125.0
Ladakh	9466289.0
Dadra and Nagar Haveli and Daman and Diu	11358600.0
Sikkim	16136752.0



Least 5 Vaccinated States in India: On the flip side, this chart spotlights lagging regions where vaccination coverage is low. Reasons could include weaker infrastructure, vaccine hesitancy, rural accessibility issues, or smaller populations. It provides a "red flag" view for where public health interventions should be targeted.

FINDINGS

1.Active Case Burden

- Maharashtra, Karnataka, and Kerala consistently ranked at the highest end of the scale when it comes to active cases. This long tail of states suggests that the COVID-19 pandemic was largely isolated to a few areas.

2.Mortality Impact

- Maharashtra reported the most fatalities, with Uttar Pradesh and Karnataka coming in after that, which is to be expected, with the largest states usually reporting the most fatalities and high case numbers Mortality rates
- demonstrate stress on the health-care delivery infrastructure and inconsistent preparedness among states.

3.Temporal Trends (Top 5 States)

- Line plots illustrated the multiple waves of infection, showing a sharp spike around the time of the "second wave" across the major states. All states had
- some variation in timing of peaks—for example, Kerala might have prolonged infection waves compared to states such as Tamil Nadu having a steep recovery curve.

4.Vaccination Coverage (Gender Split)

- Male vaccinations were higher overall, suggesting another gender skew.
- The gender gap illustrates the continued possible socio-cultural or access gaps within the vaccination campaigns.

5.Vaccination Leaders and Laggards

- Uttar Pradesh, Gujarat, and Maharashtra stood out as being the leaders in terms of total vaccinations provided, reflecting the implementation capabilities in populous states. While the smaller states and states with
- fewer resources congregated under the "least vaccinated" group due to infrastructure or outreach challenges.

6.Overall Pattern

- The data reveals a concentrated burden: a few large states carried disproportionate shares of cases, deaths, and vaccinations. Inequities were
- visible not just regionally but also demographically (gender- based vaccination gaps). Temporal analysis reinforces that India's COVID-19 crisis was not
- uniform — it was wave-driven, region-specific, and strongly influenced by public health response capacity.

RECOMMENDATIONS

The exploratory data analysis (EDA) of COVID-19 data considered three other exploratory analytical approaches to explore the relationship between growing COVID-19 vaccinations and the COVID-19 case trend in India. First, the alternative used basic descriptive and visual analysis applied simple visualizations like bar and line charts and replaced statistical summaries simply to indicate how many confirmed, recovering, and death cases and vaccines were reported. This was the simplest and easily implemented approach that offered some simple patterns observed in the vaccine or COVID-19 data, although it did not provide any analytical depth and could not indicate the relationship or extent of predictive insight of the COVID-19 vaccines. The second alternative was strictly correlation and regression analysis of the data to quantify the relationship of vaccination rates and infection trends. This type of regression explored the statistical theories of correlation to determine how levels of vaccination impacted case counts and provided stronger evidence and measured results; however, as a correlational and regression analysis, it likely assumed normal distributions and linear relationships, which would be sensitive to outliers. The third alternative applied time series trend analysis to explore time series theory and observe the time patterns, waves, and trends pre- and post-vaccination. This method looked at the lag effect of the vaccination on COVID-19 distribution while offering greater insight into the evolution of the pandemic over time, but required additional complexity in data handling.

Upon reviewing all three analyses, the best solution identified was a combination of time series and correlation analysis. This combination was the best method because it provided the most accurate understanding of the impact of vaccination on COVID-19 cases. It represented the temporal dynamics of the pandemic and the relative statistical methodology of the relationship between the variables. Additionally, it is scalable across states and time frames for analysis, which is useful for supporting evidence- based public health decision making. The rolling means and correlation heatmaps offered the clearest and interpretable data to synthesize the results. Therefore, the descriptive analysis made the first connection and the regression established the relationship quantitatively, but the best solution to this problem was the combination of time series and correlation analysis because it provided quantitative accuracy and meaningful understanding of the dynamic relationship between vaccination rollout on the decline of COVID-19 cases in India.

IMPLEMENTATION

1.Public Health Surveillance

- Ongoing tracking of incidence, mortality, and recovery sequencing are key to allowing governments to observe and assess the arrival and subsequent movement of new epidemic waves earlier. Estimates of RtR_tRt (effective
- reproduction number) can be derived from such data and can inform the timing of restriction, tightening or relaxing, for governments.

2.Resource Allocation

- States with the greatest burden (Maharashtra, Kerala, Karnataka) can be prioritized for the allocation of hospital beds, oxygen supply, and medical personnel in periods of peak epidemic waves. Mortality high regions can be
- flagged for the expansion of ICU beds and ventilator supply.

3.Targeted Vaccination Campaigns

- Vaccination data can inform intervention specifically (e.g., supply chain systems, outreach in rural districts or through mobile vaccination). Data that
- can differentially report male or female vaccination can also show clearly that states should implement female directed vaccination campaigns to remove equity gaps.

4.Policy Planning & Risk Communication

- Graphs and other visualizations (e.g., pie charts, line plots, and bar charts) can communicate risks clearly to policy makers and the public respectively.
- Public health officials/agencies can engage in risk communications campaigns with these graphs to communicate state specific risks, and the importance of vaccination.

5.Predictive Modeling

- The dataset can contribute to SIR/SEIR models for outbreaks to brainstorm future outbreak scenarios of cases, and what if it slows down vaccination rates or increases mobility comfortably. Early warning dashboards can be
- developed from this analysis for health authorities to produce alerts.

6.Digital Health Platforms

- The methods can be embedded into real-time dashboards (like CoWIN + state health portals) for live case tracking, vaccination equity monitoring, and decision support.

CONCLUSION

This project on Exploratory Data Analysis (EDA) of COVID-19 data in India has generated a robust understanding of how the pandemic unfolded over time and the role vaccination played in mitigating the spread of COVID-19. Using two datasets - covid_19_india.csv and covid_vaccine_statewise.csv - the analysis considered COVID-19 case statistics alongside vaccination data to uncover important patterns, correlations, and trends related to COVID-19 and vaccination. The analysis started with a step on data cleaning and preprocessing, to make sure the data was reliable and consistent. Then, statistical and visualization techniques were employed to explain the findings as they are interpreted. The processes of descriptive analyses provided insights into the flux of numerical data of daily confirmed cases (and recovered cases, and cases of death) across states and times, while revealing the COVID-19 intensity at distinct points of the pandemic.

Analysis through correlation and time series analytical methods were especially revealing, as there was a noticeable drop in instances of COVID-19 following the rapid increase in the COVID-19 vaccination rates. As shown in the negative correlation between total doses administered and new confirmed cases of COVID-19, as vaccination coverage increased, the community experienced less transmission and lower mortality from COVID-19. Time series visualizations demonstrated how rapidly vaccinating helped to flatten the curve of SARS-CoV-2, maintain case counts, and eventually move to recovery trends in many states. This aligns with the theoretical literature that suggest immunization is fundamental to promoting herd immunity and reducing the spread of the virus.

In addition to numeric results, the project demonstrated the usefulness of EDA in real-world problem solving. It demonstrated that systematic analysis of data can assist policymakers and researchers make data-driven decisions rather than assumptions. The insights generated can assist with understanding the effectiveness of government health responses, which states could benefit for better resource allocation, and potential future outbreaks. Additionally, the project demonstrated how using multiple data sources to complement each other enhanced the analysis as well as reliability.

In summary, this project achieved what it set out to do in examining COVID-19 data, and it highlighted the value of EDA as the first step in any data-driven decision-making process. This project was able to take the raw data from the pandemic and, through descriptive, correlation, and time series methods, turn it into knowledge that is useful. The research established that vaccination was an important factor in reducing the impact of COVID-19 in India, and that ongoing surveillance through data analytics is necessary in the future to fight and reduce the impact of any public health emergency.

REFERENCES

1.Ministry of Health and Family Welfare (MoHFW), "COVID-19 Statewise Status Report," Government of India, 2021. [Online]. Available: https://www.mohfw.gov.in/

2.World Health Organization (WHO), "Coronavirus Disease (COVID-19) Dashboard," World Health Organization, 2021. [Online]. Available: https://covid19.who.int/

3.Indian Council of Medical Research (ICMR), "COVID-19 Testing and Vaccination Data," ICMR, Government of India, 2021. [Online]. Available: https://www.icmr.gov.in/

4.Johns Hopkins University and Medicine, "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE)," Johns Hopkins University, 2021. [Online]. Available: https://coronavirus.jhu.edu/

5.Press Information Bureau (PIB), "Updates on COVID-19 Vaccination Progress in India," Government of India, 2021. [Online]. Available: https://pib.gov.in/

6.T. Hale, N. Angrist, R. Goldszmidt, B. Kira, A. Petherick, T. Phillips, and S. Webster, "A Global Panel Database of Pandemic Policies (Oxford COVID-19 Government Response Tracker)," Nature Human Behaviour, vol. 5, no. 4, pp. 529–538, 2021, doi: 10.1038/s41562-021-01079-8.

7.World Bank, "Tracking the Socioeconomic Impacts of COVID-19 on India," World Bank Publications, 2021. [Online]. Available: https://www.worldbank.org/

8.Centers for Disease Control and Prevention (CDC), "Science Brief: SARS-CoV-2 Transmission," CDC, 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/index.html

9.H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, and M. Roser, "Coronavirus (COVID-19) Vaccinations," Our World in Data, 2021. [Online]. Available: https://ourworldindata.org/covid-vaccinations

10.G. Pandey, P. Chaudhary, R. Gupta, and S. Pal, "SEIR and Regression Model-Based COVID-19 Outbreak Predictions in India," arXiv preprint arXiv:2004.00958, Apr. 2020, doi: 10.48550/arXiv.2004.00958.