Impact of COVID-19 Across Indian States: A Data-Driven Analytical Study

1. Project Overview

This project analyzes COVID-19 case and vaccination trends across Indian states using real datasets. Python was used for data cleaning and exploratory analysis, SQL was used for deriving state-wise insights, and Tableau was used to visualize patterns and comparisons. The study highlights the most affected regions, recovery and mortality trends, and vaccination progress. The results help in understanding the pandemic's impact and healthcare response across India.

2. Dataset Summary

The project uses state-wise COVID-19 case data and vaccination records collected during the pandemic period in India. The dataset includes columns such as **Date, State, Confirmed Cases, Recovered Cases, Deaths, and Vaccination Counts**. An additional field **Active Cases** was calculated to analyze ongoing case pressure on healthcare systems. The dataset is clean, structured, and suitable for time-series and comparative state-level analysis.

3. Exploratory Data Analysis (EDA) Using Python

Exploratory Data Analysis was performed in Python to understand the spread of COVID-19 across different states and time periods. The dataset was first cleaned by removing unnecessary columns and converting the **Date** column into a proper datetime format. A new feature **Active Cases** was derived using the formula:

Active Cases = Confirmed – (Recovered + Deaths)

State-wise summaries were generated to identify the **top affected states**, while recovery and mortality rates were calculated to measure health outcomes. Visualization libraries such as **Matplotlib** and **Seaborn** were used to create bar charts and line graphs for trend comparison

Key Python Code Used

```
covid_df['Active_Cases'] = covid_df['Confirmed'] - (covid_df['Cured'] + covid_df['Deaths']
```

```
top_10_active =
covid_df.groupby("State/UnionTerritory").max()[["Active_Cases"]].sort_values("Active_Cases",
ascending=False).head(10)

sns.barplot(data=top_10_active, x=top_10_active.index, y="Active_Cases")
plt.xticks(rotation=45)
plt.title("Top 10 States with Highest Active Cases")
plt.show()
```

Top 10 States with Highest Active Cases

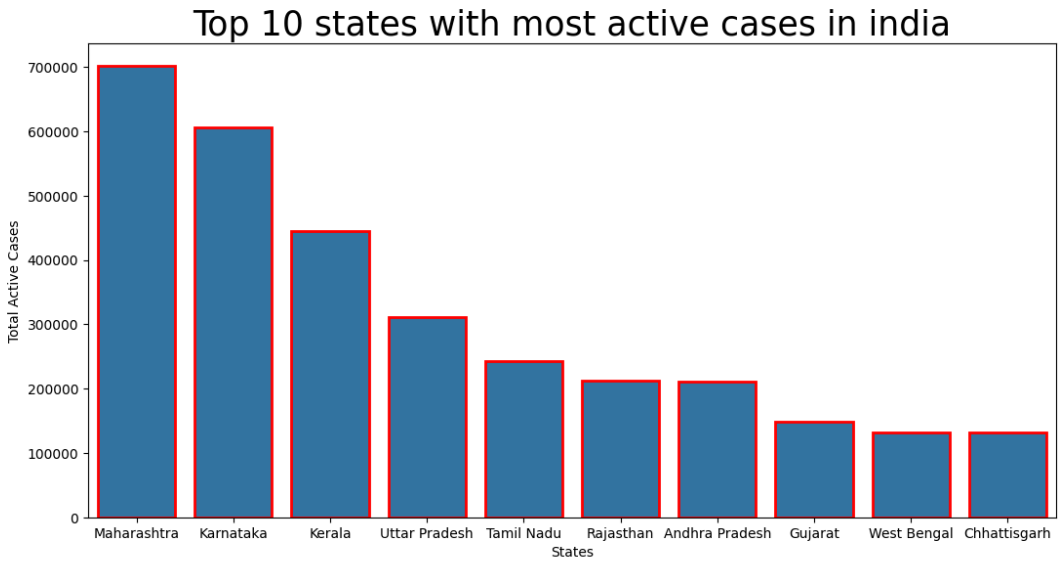

Figure 1: Bar graph showing the states with the maximum number of active COVID-19 cases

Top 10 States with Highest Death Cases

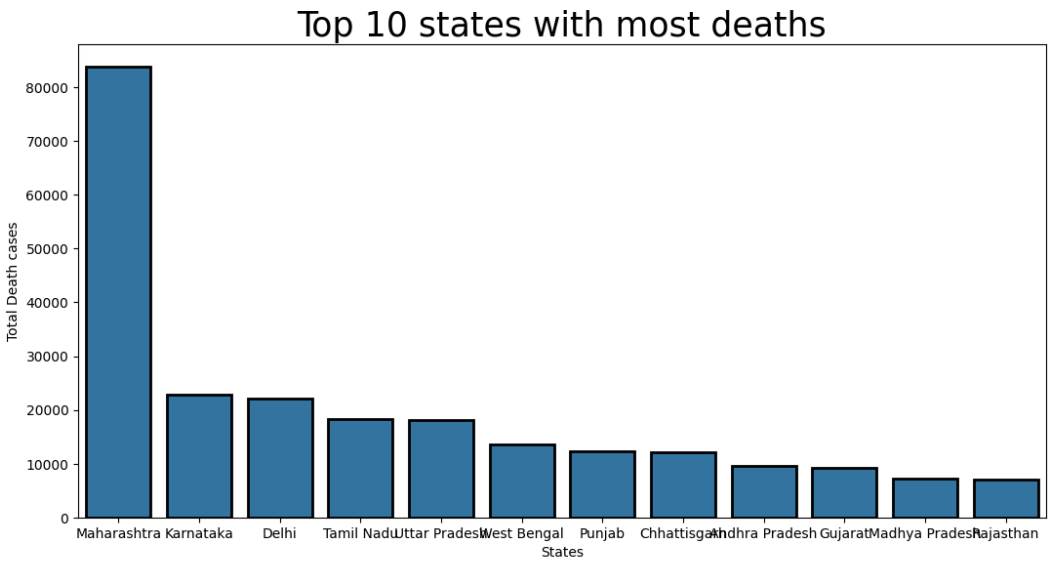

Figure 2: Bar graph showing the states with the maximum number of Death COVID-19 case**s.**

Insights from EDA

- Maharashtra, Kerala, Karnataka, and Tamil Nadu were the most heavily impacted states.
- Recovery rates increased over time due to improved medical response and vaccination.
- Mortality varied by region, influenced by healthcare infrastructure and population density.
- Multiple waves of increasing and decreasing case trends were observed throughout the timeline.

4. Data Analysis Using SQL

SQL was used to perform analytical queries on the cleaned COVID-19 dataset stored in PostgreSQL. The purpose of the SQL analysis was to identify the most affected states, calculate national recovery and mortality rates, and study vaccination distribution across regions. Aggregation functions such as **SUM**, **MAX**, and **GROUP BY** were used to compute trends, while **ORDER BY** helped in ranking states based on case severity. This analysis supported deeper interpretation of COVID-19 spread and population health outcomes.

Key SQL Queries and Insights

1. Total Confirmed, Recovered and Death Cases in India

```
SELECT SUM(confirmed) AS total_confirmed,
       SUM(cured) AS total_recovered,
       SUM(deaths) AS total_deaths
FROM covid_cases;
```

Insight: Shows the overall impact of COVID-19 across the country.

---

2. Top 10 States by Active Cases

```
SELECT state, MAX(active_cases) AS highest_active
FROM covid_cases
GROUP BY state
ORDER BY highest_active DESC
LIMIT 10;
```

Insight: Identified Maharashtra, Kerala, Karnataka, and Tamil Nadu as the most affected states.

---

3. Recovery Rate and Mortality Rate in India

```
SELECT
(SUM(cured)::float / SUM(confirmed)) * 100 AS recovery_rate,
(SUM(deaths)::float / SUM(confirmed)) * 100 AS mortality_rate
FROM covid_cases;
```

Insight: Recovery rate improved over time, while mortality remained comparatively low.

---

4. Trend Analysis (Monthly Confirmed Cases)

```
SELECT DATE_TRUNC('month', date) AS month,
       SUM(confirmed) AS monthly_cases
FROM covid_cases
GROUP BY month
ORDER BY month;
```

Insight: Clearly shows wave patterns and rise/fall cycles during the pandemic periods.

---

5. Vaccination Gender Distribution

```
SELECT SUM(male_vaccinated) AS total_male,
       SUM(female_vaccinated) AS total_female
FROM covid_vaccination;
```

Insight: Vaccination among males was slightly higher, but gap narrowed over time

Summary of SQL Work

- The COVID-19 dataset was stored and analyzed using PostgreSQL.
- SQL queries were used to calculate national statistics and compare state-wise severity.
- Recovery and mortality rates provided insights into healthcare outcomes.
- Vaccination data helped analyze gender distribution and state-level progress.

5. Dashboard in Tableau

Finally, we built an interactive dashboard in **Tableau** to present insights visually

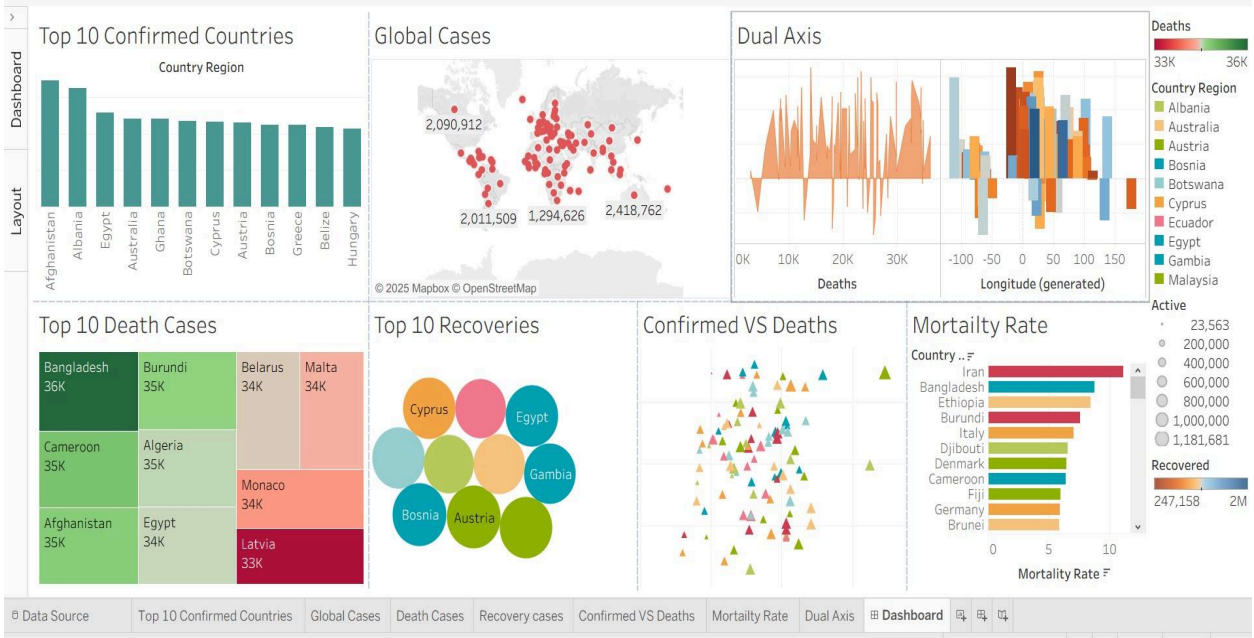

Figure 1: Tableau dashboard displaying overall COVID-19 case statistics.

Dashboard Insights

- Maharashtra, Kerala, Karnataka, and Tamil Nadu were consistently the most affected states.
- Recovery rate improved significantly across several states after vaccination rollout.
- Case trends show multiple waves, indicating periodic outbreak patterns.
- Visual comparison helps understand how different states responded to the pandemic.

6. Key Findings

Finding	Interpretation
Maharashtra, Kerala, Karnataka and Tamil Nadu had the highest number of confirmed and active cases.	These states have high population density and major urban centers, increasing transmission risk.
The recovery rate improved steadily over time across most states.	Indicates better medical support, awareness, and vaccination effect.
Mortality rates varied between states.	Differences in healthcare infrastructure and response strategies influenced outcomes.
Multiple waves of infection were observed during the timeline.	Shows that the virus spread was periodic and influenced by public movement and safety measures.
Vaccination rates increased significantly, reducing active cases in later stages.	Vaccination played a key role in controlling the pandemic impact.

7. Conclusion

This project successfully analyzed the spread and impact of COVID-19 across India using real case and vaccination data. Python enabled efficient data cleaning and visualization, SQL provided deeper analytical insights, and Tableau helped represent the results interactively. The findings highlight the most affected states, improvements in recovery, and the role of vaccination in reducing case severity. Overall, the project provides meaningful understanding of pandemic trends and supports decision-making in public health analysis.

8. Future Scope

- Apply Machine Learning to **forecast** upcoming case trends.
- Automate data updates using live API integration.
- Create a web-based dashboard for real-time monitoring.