PROJECT REPORT
SUBMITTED BY : GOVIND LAKHOTIA
ROLL NO : 045014
GITHUB LINK : https://github.com/govindlakhotia/DEV1_Project

Project : Design a Web Scraper to Scrape Data from a Website. Analyze the Data and Make a Report on the Analysis

Topic : Analyzing data of Covid Impact on different states taken from https://api.covid19india.org/csv/latest/state_wise.csv (Data until August 2021)

PROJECT OBJECTIVES

The primary objectives of this analysis are as follows:

- To assess the regional variations in COVID-19 impact, including confirmed cases, deaths, and active cases.
- To identify high-risk areas that require targeted resource allocation.
- To emphasize data-driven decision-making in pandemic response.
- To explore communication strategies that consider regional differences.
- To encourage collaboration among regions and flexibility in response strategies.
- To underscore the importance of continuous monitoring and surveillance.

DESCRIPTION OF DATA

❖ Data Sources

The COVID-19 data used in this analysis was sourced from reputable and authoritative organizations, including government health agencies and established data providers. This data forms the basis of our analysis and provides a comprehensive view of the pandemic's impact at the regional level.

❖ Data Overview

```
[6] # Read the CSV file using pandas
    df = pd.read_csv("covid_data.csv")
    df
```

	State	Confirmed	Recovered	Deaths	Active	Last_Updated_Time	Migrated_Other	State_code	Delta_Confirmed	Delta_Recovered	Delta_Deaths	State_Notes
0	Total	34285612	33661339	458470	152606	13/08/2021 23:27:22	13197	TT	0	0	0	NaN
1	Andaman and Nicobar Islands	7651	7518	129	4	13/08/2021 23:27:22	0	AN	0	0	0	NaN
2	Andhra Pradesh	2066450	2047722	14373	4355	13/08/2021 23:27:22	0	AP	0	0	0	NaN
3	Arunachal Pradesh	55155	54774	280	101	13/08/2021 23:27:22	0	AR	0	0	0	[July 25]: All numbers corresponding to Papum ...
4	Assam	610645	600974	5997	2327	13/08/2021 23:27:22	1347	AS	0	0	0	[Jan 1]: 1347 cases i.e Covid +'tive patients ...
5	Bihar	726098	716390	9661	46	13/08/2021 23:27:22	1	BR	0	0	0	[June 9] : 3951 deceased cases have been repor...
6	Chandigarh	65351	64495	820	36	13/08/2021 23:27:22	0	CH	0	0	0	NaN
7	Chhattisgarh	1006052	992159	13577	316	13/08/2021 23:27:22	0	CT	0	0	0	NaN
8	Dadra and Nagar Haveli and Daman and Diu	10681	10644	4	2	13/08/2021 23:27:22	31	DN	0	0	0	NaN
9	Delhi	1439870	1414431	25091	348	13/08/2021 23:27:22	0	DL	0	0	0	[July 14]: Value for the total tests conducted...
10	Goa	178108	174392	3364	352	13/08/2021 23:27:22	0	GA	0	0	0	NaN
11	Gujarat	826577	816283	10089	205	13/08/2021 23:27:22	0	GJ	0	0	0	NaN
12	Haryana	771252	761068	10049	135	13/08/2021 23:27:22	0	HR	0	0	0	NaN
13	Himachal Pradesh	224106	218410	3738	1942	13/08/2021 23:27:22	16	HP	0	0	0	NaN
14	Jammu and Kashmir	332249	326915	4432	902	13/08/2021 23:27:22	0	JK	0	0	0	NaN
15	Jharkhand	348764	343518	5138	108	13/08/2021 23:27:22	0	JH	0	0	0	NaN
16	Karnataka	2988333	2941578	38082	8644	13/08/2021 23:27:22	29	KA	0	0	0	NaN
17	Kerala	4968657	4857181	31681	79266	13/08/2021 23:27:22	529	KL	0	0	0	NaN
18	Ladakh	20962	20687	208	67	13/08/2021 23:27:22	0	LA	0	0	0	NaN
19	Lakshadweep	10365	10270	51	0	13/08/2021 23:27:22	44	LD	0	0	0	NaN
20	Madhya Pradesh	792854	782215	10524	115	13/08/2021 23:27:22	0	MP	0	0	0	[14 Oct'20]: 4469 confirmed cases and

The given screenshot gives us an idea of the data that has been collected from the website. The dataset consist of COVID-19 data of the different states and union territories of India. The provided snapshot offers an overview of the dataset, outlining the columns and variables available for analysis. Key variables include "State," "Confirmed Cases," "Deaths," "Recoveries," and "Active Cases." This snapshot serves as the foundation for our analysis, allowing us to delve into the details of COVID-19 trends and impacts across regions.

ANALYSIS OF DATA

- **Data Preparation**

  Data Cleaning

To ensure the integrity and accuracy of our analysis, a meticulous data cleaning process was executed. This involved addressing missing values, rectifying data inconsistencies, and identifying and handling any potential outliers. The goal was to transform the data into a reliable and consistent format for further analysis.

```
# Filter necessary columns
df = df[df['State'] != 'State Unassigned']
df = df[df['State'] != 'Total']
df = df[["State", "Confirmed", "Deaths", "Recovered", "Active"]]

# Print the data
df
```

	State	Confirmed	Deaths	Recovered	Active
1	Andaman and Nicobar Islands	7651	129	7518	4
2	Andhra Pradesh	2066450	14373	2047722	4355
3	Arunachal Pradesh	55155	280	54774	101
4	Assam	610645	5997	600974	2327
5	Bihar	726098	9661	716390	46
6	Chandigarh	65351	820	64495	36
7	Chhattisgarh	1006052	13577	992159	316
8	Dadra and Nagar Haveli and Daman and Diu	10681	4	10644	2
9	Delhi	1439870	25091	1414431	348
10	Goa	178108	3364	174392	352
11	Gujarat	826577	10089	816283	205
12	Haryana	771252	10049	761068	135

- **Exploratory Data Analysis (EDA)**

Data Overview

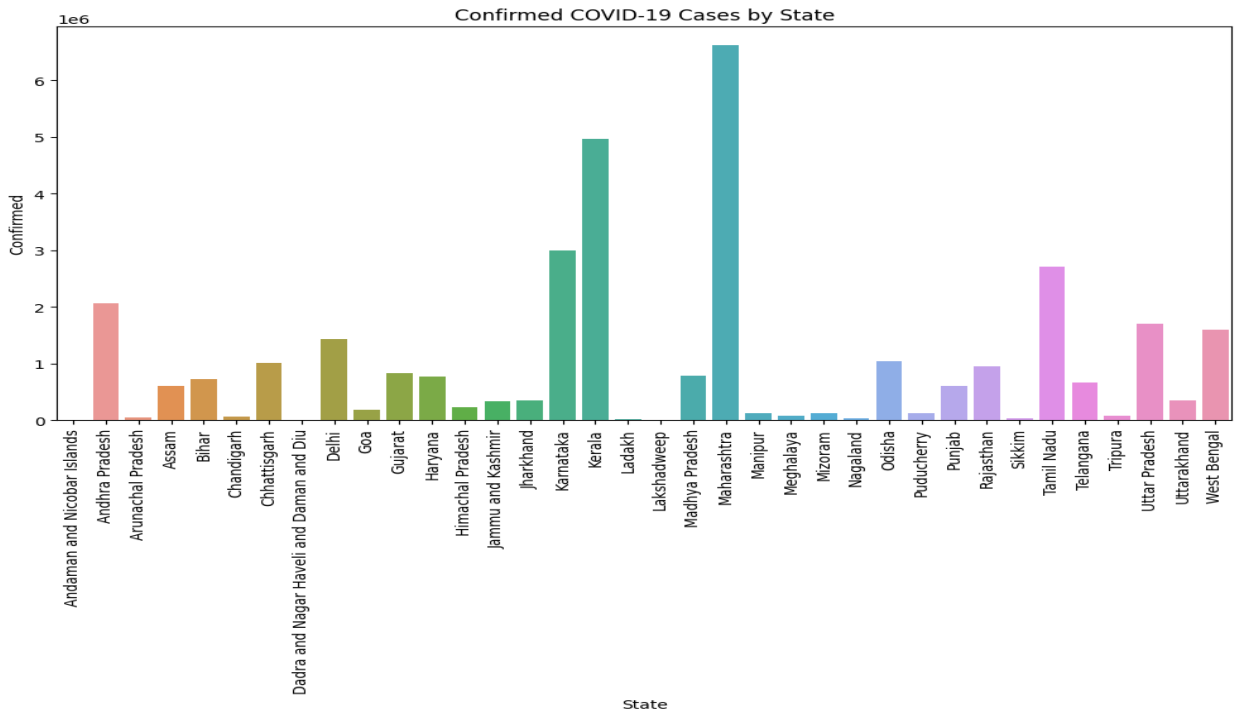
Exploratory Data Analysis (EDA) played a pivotal role in understanding the dataset. We conducted an initial assessment of key statistics, such as means, medians, and standard deviations, to gain insights into the central tendencies and variabilities of COVID-19 metrics.

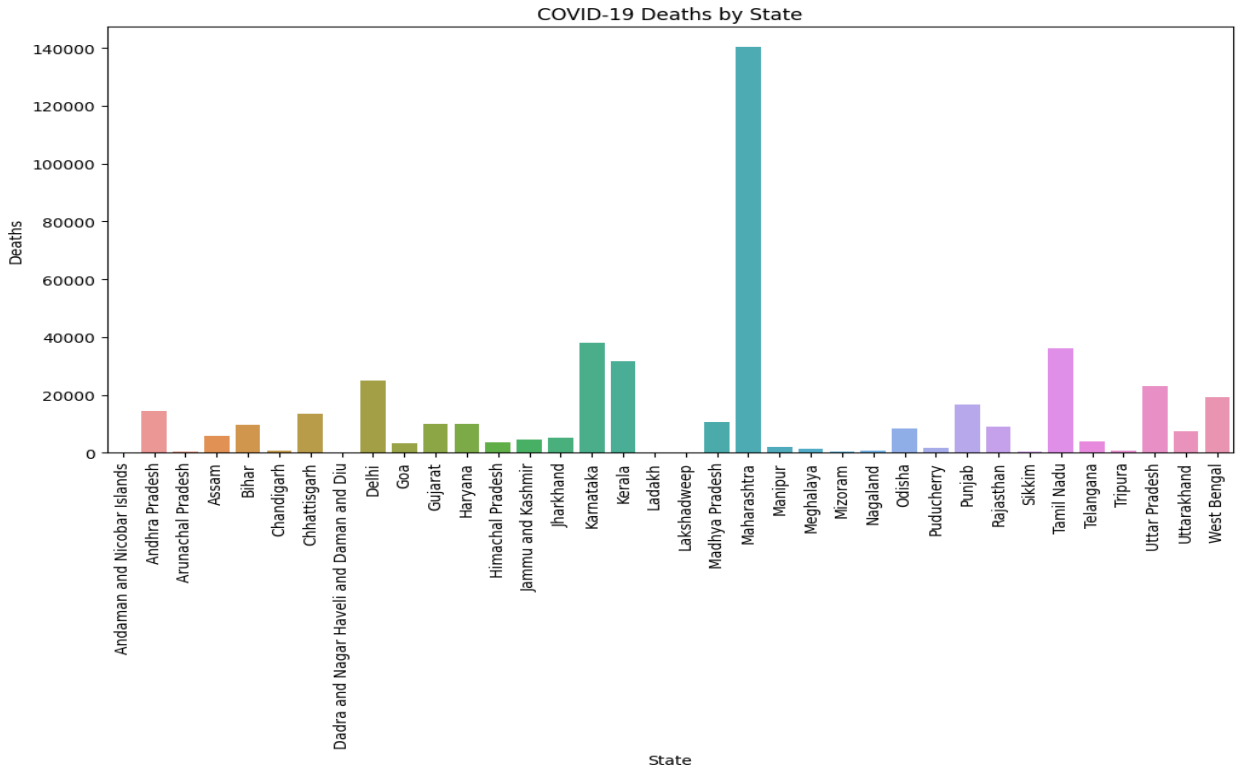
```
# Basic data exploration
print(df.head())
print('\n')
print(df.describe())
```

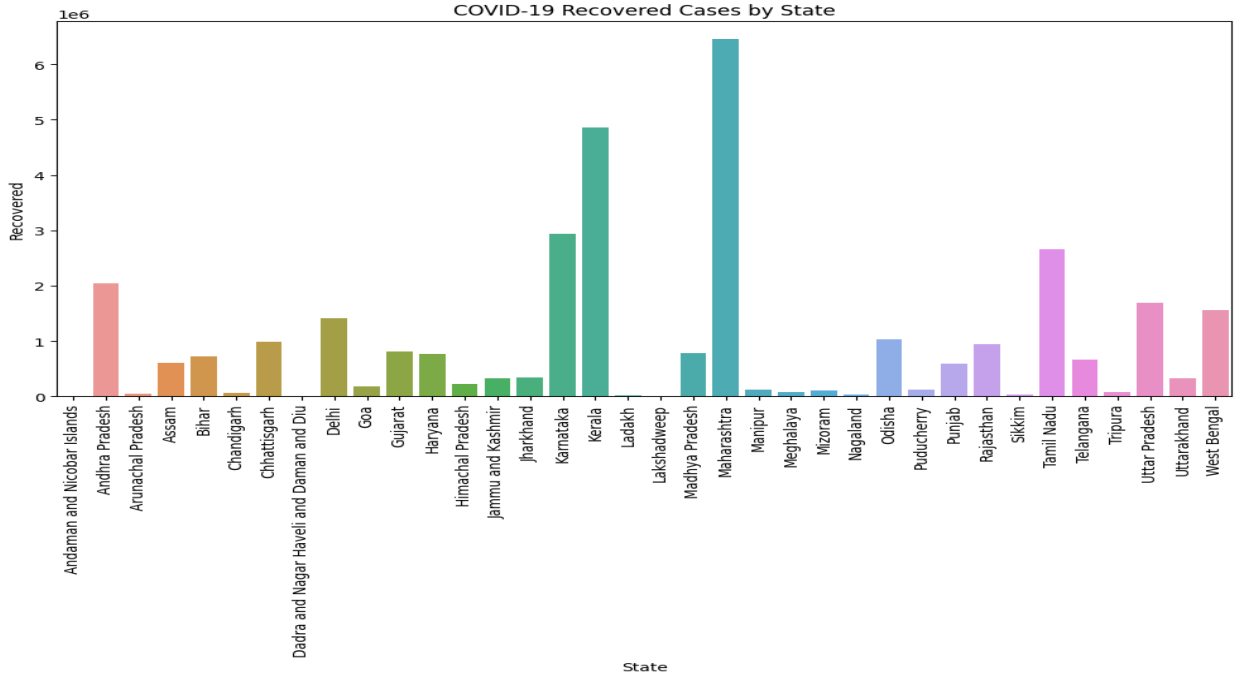
```
                          State  Confirmed  Deaths  Recovered  Active
1  Andaman and Nicobar Islands       7651     129       7518       4
2               Andhra Pradesh    2066450   14373    2047722    4355
3            Arunachal Pradesh      55155     280      54774     101
4                        Assam     610645    5997     600974    2327
5                        Bihar     726098    9661     716390      46


          Confirmed         Deaths      Recovered        Active
count  3.600000e+01      36.000000   3.600000e+01     36.000000
mean   9.523781e+05   12735.277778   9.350372e+05   4239.055556
std    1.423395e+06   24208.694026   1.392346e+06  13406.951453
min    7.651000e+03       4.000000   7.518000e+03      0.000000
25%    8.425775e+04     818.250000   8.303600e+04    107.750000
50%    4.755825e+05    5567.500000   4.645545e+05    283.500000
75%    1.014903e+06   13776.000000   1.001406e+06   2726.250000
max    6.611078e+06  140216.000000   6.450585e+06  79266.000000
```

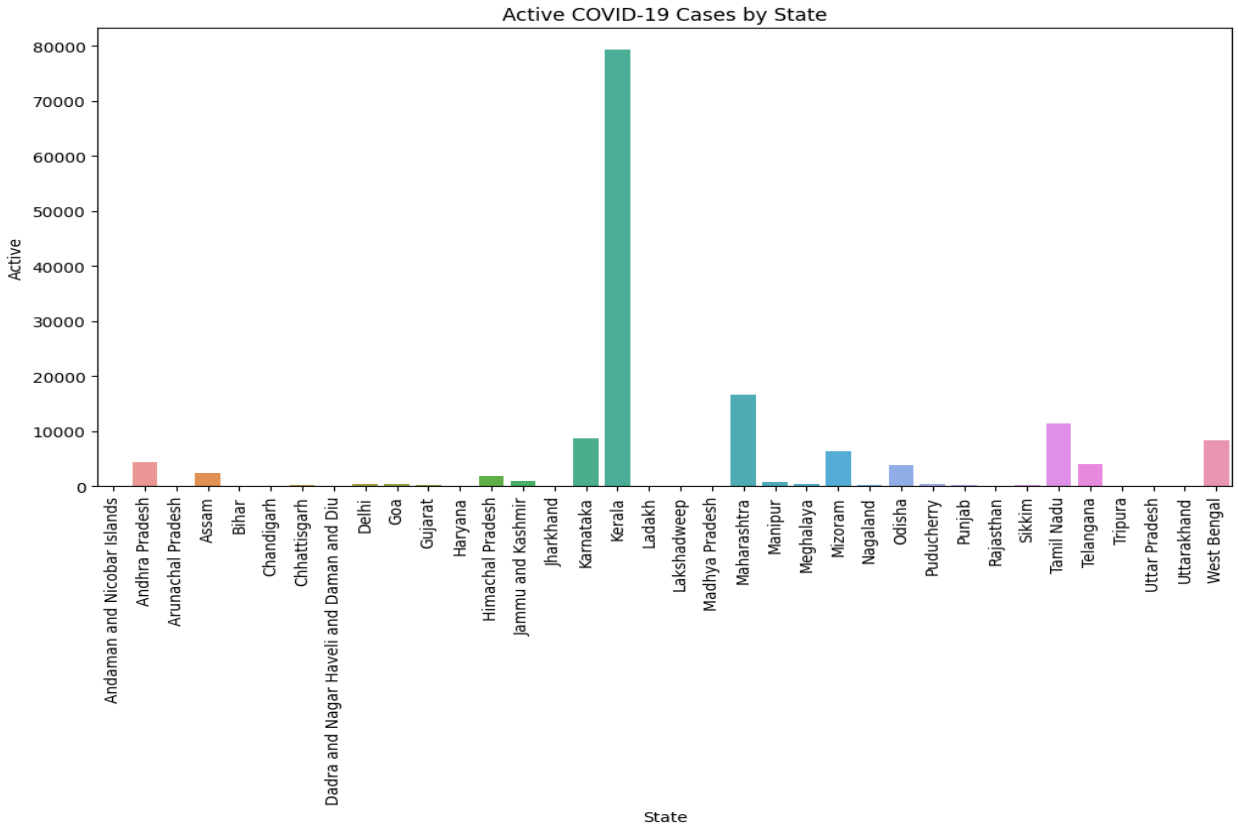
Descriptive Analysis

The descriptive analysis yielded several key insights. For example, we observed that certain states exhibited higher average daily case counts, suggesting different rates of transmission. This information is vital for resource allocation and response planning.




COVID-19 Deaths by State
Deaths
0
20000
40000
60000
80000
100000
120000
140000
Andaman and Nicobar Islands
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Dadra and Nagar Haveli and Daman and Diu
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu and Kashmir
Jharkhand
Karnataka
Kerala
Ladakh
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
State

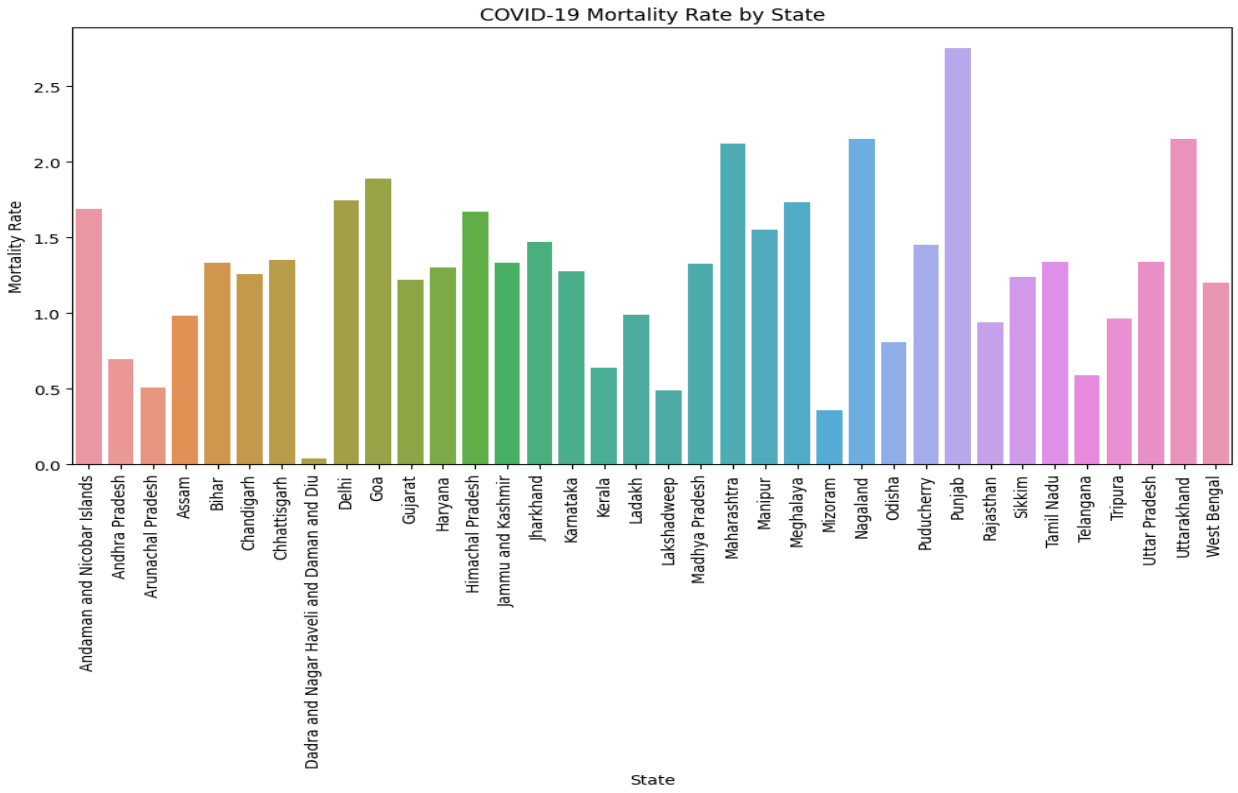

COVID-19 Recovered Cases by State
1e6
Recovered
0
1
2
3
4
5
6
Andaman and Nicobar Islands
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Dadra and Nagar Haveli and Daman and Diu
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu and Kashmir
Jharkhand
Karnataka
Kerala
Ladakh
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
State

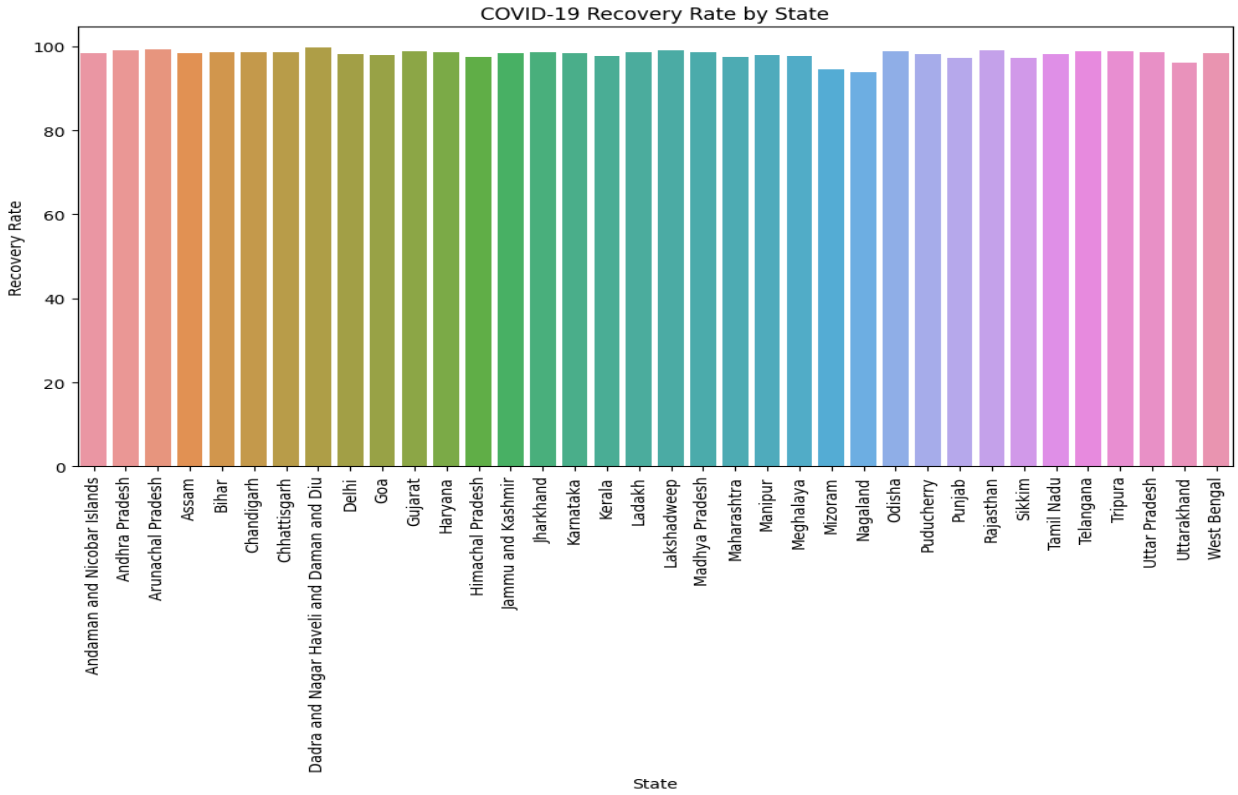


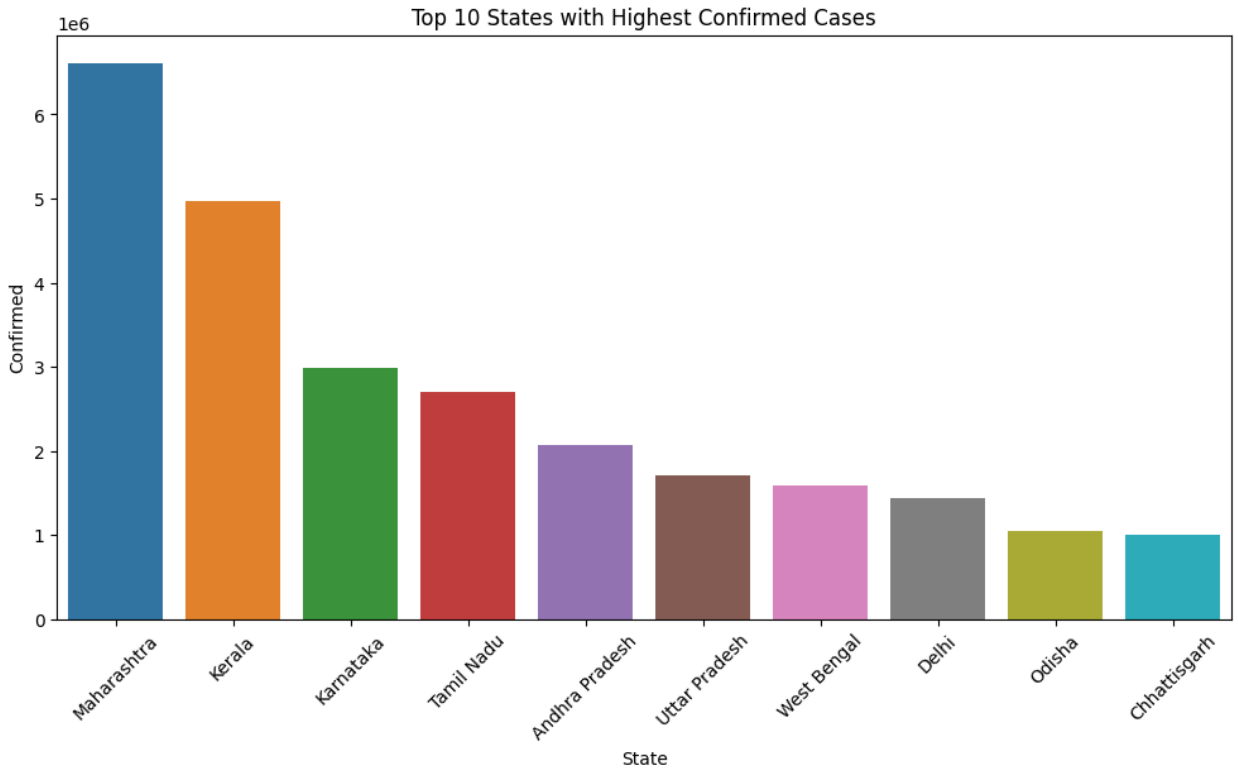
FINDINGS AND INFERENCES

Mortality Rate and Recovery Rate by State

From the graphs, we can see that the recovery rate is very high in each and every state. This concludes that most of the patients who got COVID were recovered successfully. On the other hand, the mortality rate is showing a lot of variations, which takes us to the conclusion that the deaths/confirmed ratio is different for each and every state.

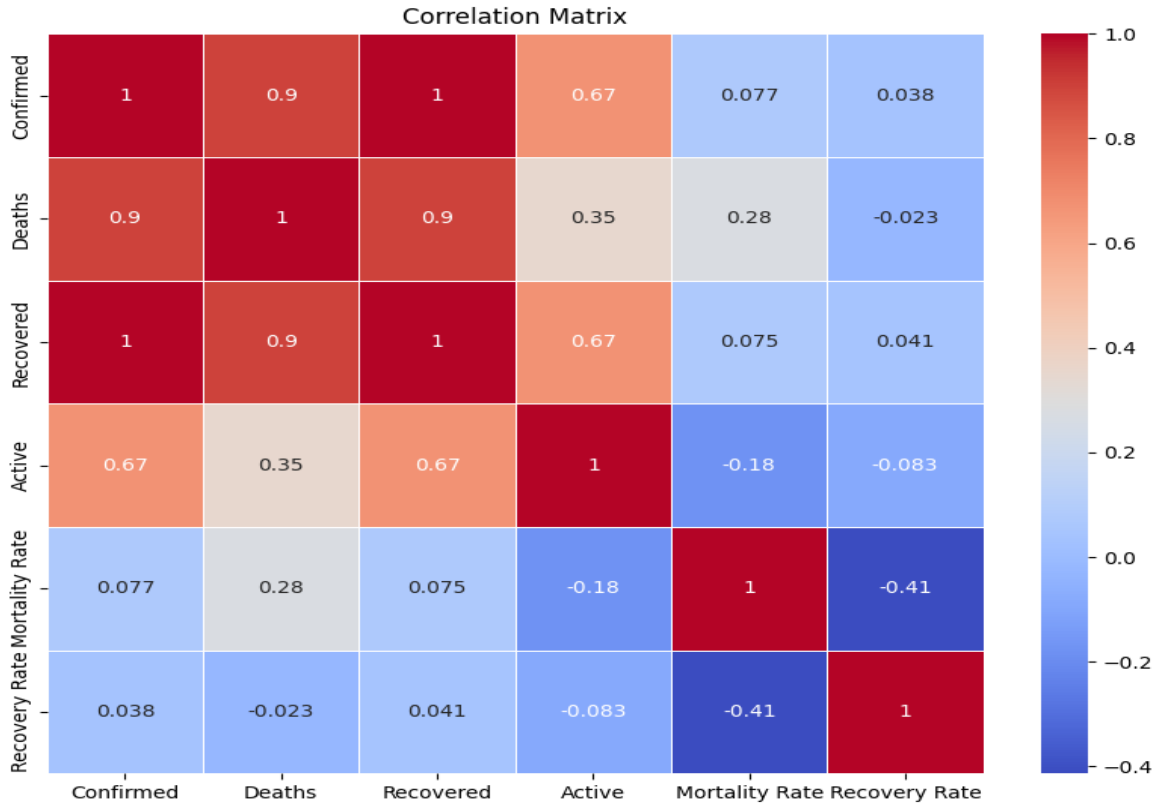

COVID-19 Mortality Rate by State
Mortality Rate
0.0
0.5
1.0
1.5
2.0
2.5
Andaman and Nicobar Islands
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Dadra and Nagar Haveli and Daman and Diu
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu and Kashmir
Jharkhand
Karnataka
Kerala
Ladakh
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
State


COVID-19 Recovery Rate by State
Recovery Rate
0
20
40
60
80
100
Andaman and Nicobar Islands
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Dadra and Nagar Haveli and Daman and Diu
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu and Kashmir
Jharkhand
Karnataka
Kerala
Ladakh
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
State

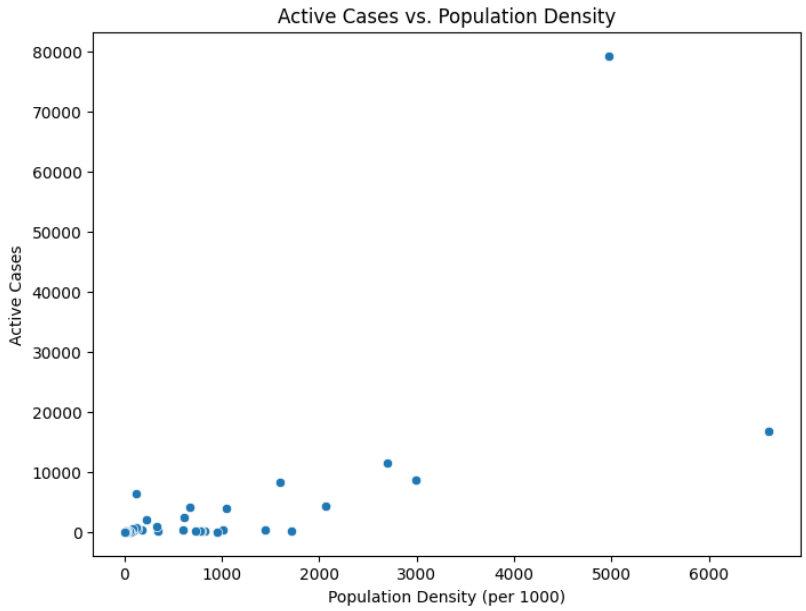



```
Top 5 States by Confirmed Cases:
            State  Confirmed
21    Maharashtra    6611078
17         Kerala    4968657
16      Karnataka    2988333
32     Tamil Nadu    2702623
2   Andhra Pradesh    2066450

Bottom 5 States by Confirmed Cases:
                                      State  Confirmed
25                                Nagaland      31842
18                                  Ladakh      20962
8   Dadra and Nagar Haveli and Daman and Diu      10681
19                              Lakshadweep      10365
1                Andaman and Nicobar Islands       7651
```

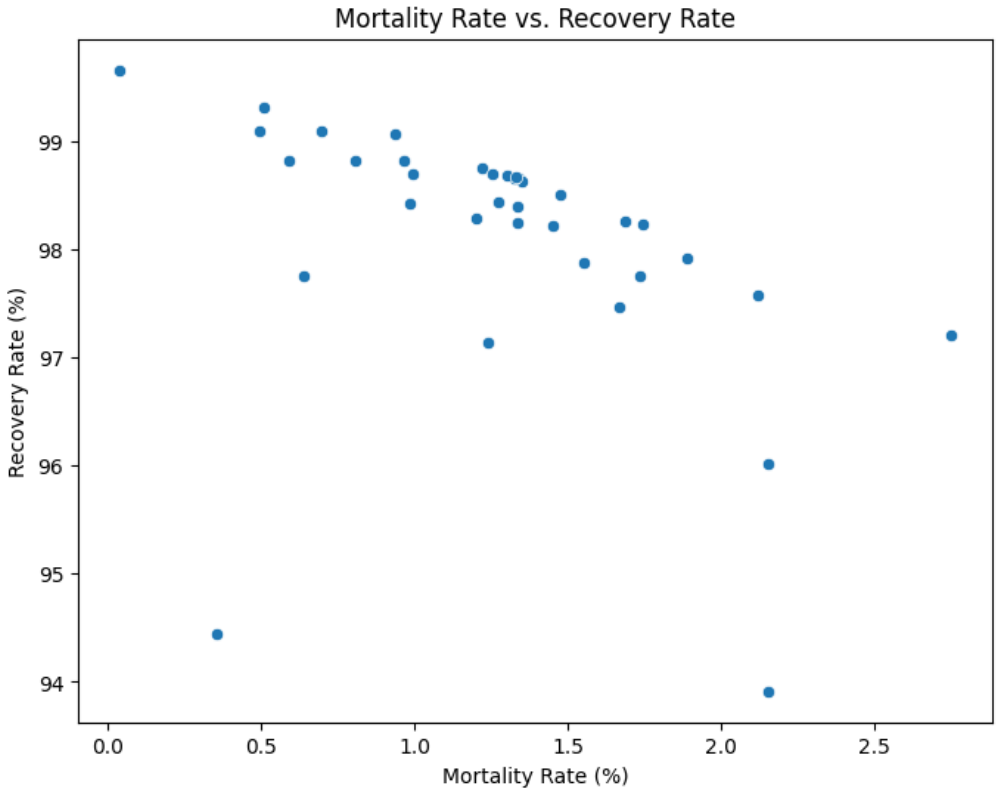
Correlation between the columns of the Data



This matrix is showing us the relation between the different variables which shows us the dependency of a certain variable on another. The correlation is given in the form of probabilities so, closer it is to 1 more dependency we can identify.



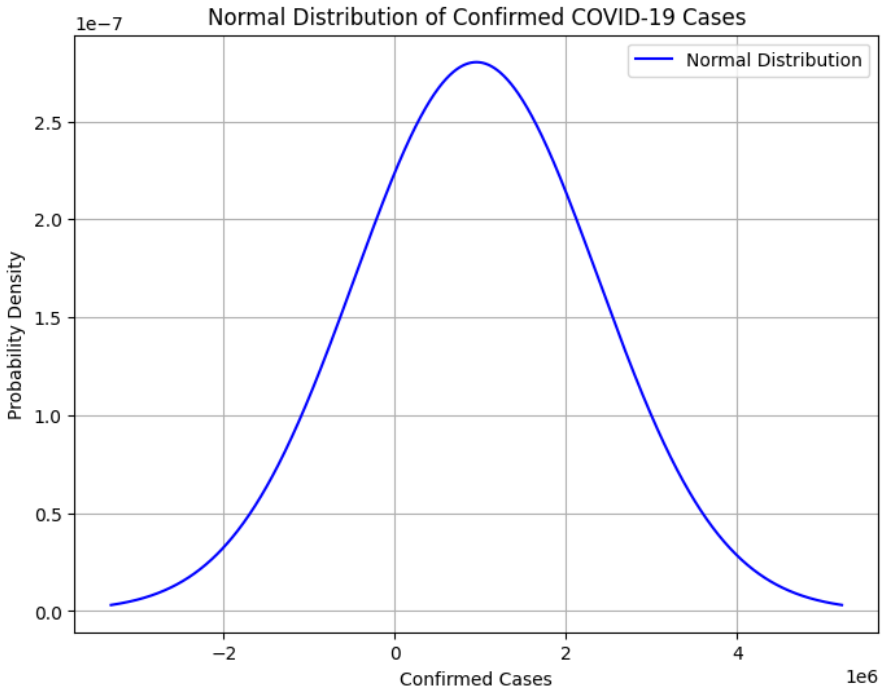
Here we can see that the scatterplot shows us the relationship between the active cases and the population density which we have assumed to be 1,000,000.



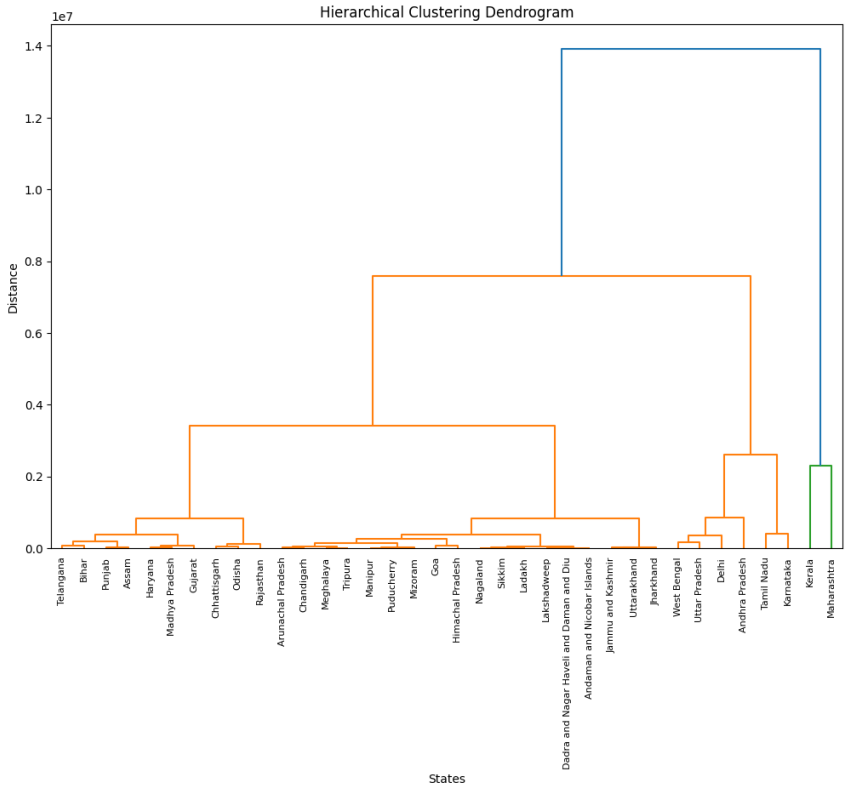
Here we can see the relationship between the mortality rate and the recovery rate.

```
State with the highest mortality rate: Punjab
State with the lowest mortality rate: Dadra and Nagar Haveli and Daman and Diu

Average Values:
Confirmed    952378.111111
Deaths        12735.277778
Recovered    935037.194444
Active         4239.055556
dtype: float64

Median Values:
Confirmed    475582.5
Deaths         5567.5
Recovered    464554.5
Active          283.5
dtype: float64

Variance Values:
Confirmed    2.026054e+12
Deaths       5.860609e+08
Recovered    1.938626e+12
Active       1.797463e+08
dtype: float64

Standard Deviation Values:
Confirmed    1.423395e+06
Deaths       2.420869e+04
Recovered    1.392346e+06
Active       1.340695e+04
dtype: float64
```

Normal Distribution



Cluster Analysis



States and Their Clusters:

	State	Cluster
21	Maharashtra	1
17	Kerala	1
16	Karnataka	2
32	Tamil Nadu	2
2	Andhra Pradesh	2
35	Uttar Pradesh	2
37	West Bengal	2
9	Delhi	2
26	Odisha	4
7	Chhattisgarh	4
29	Rajasthan	4
11	Gujarat	4
20	Madhya Pradesh	4
12	Haryana	4
5	Bihar	4
33	Telangana	4
4	Assam	4
28	Punjab	4
15	Jharkhand	3
36	Uttarakhand	3
14	Jammu and Kashmir	3
13	Himachal Pradesh	3
10	Goa	3
27	Puducherry	3
22	Manipur	3
24	Mizoram	3
34	Tripura	3
23	Meghalaya	3
6	Chandigarh	3

4	Assam	4
28	Punjab	4
15	Jharkhand	3
36	Uttarakhand	3
14	Jammu and Kashmir	3
13	Himachal Pradesh	3
10	Goa	3
27	Puducherry	3
22	Manipur	3
24	Mizoram	3
34	Tripura	3
23	Meghalaya	3
6	Chandigarh	3
3	Arunachal Pradesh	3
30	Sikkim	3
25	Nagaland	3
18	Ladakh	3
8	Dadra and Nagar Haveli and Daman and Diu	3
19	Lakshadweep	3
1	Andaman and Nicobar Islands	3

Correlation and Covariance Matrix

```
# Create a DataFrame with the selected columns
selected_data = df_sorted_confirmed[selected_columns]

# Covariance Matrix
covariance_matrix = selected_data.cov()

# Correlation Matrix (Pearson correlation)
correlation_matrix = selected_data.corr()

# Display the covariance matrix
print("Covariance Matrix:")
print(covariance_matrix)

# Display the correlation matrix
print("\nCorrelation Matrix (Pearson Correlation):")
print(correlation_matrix)
```

```
Covariance Matrix:
                Confirmed        Deaths     Recovered        Active
Confirmed  2.026054e+12  3.089712e+10  1.981829e+12  1.283956e+10
Deaths     3.089712e+10  5.860609e+08  3.018567e+10  1.136233e+08
Recovered  1.981829e+12  3.018567e+10  1.938626e+12  1.254474e+10
Active     1.283956e+10  1.136233e+08  1.254474e+10  1.797463e+08

Correlation Matrix (Pearson Correlation):
           Confirmed    Deaths  Recovered    Active
Confirmed   1.000000  0.896646   0.999986  0.672814
Deaths      0.896646  1.000000   0.895535  0.350079
Recovered   0.999986  0.895535   1.000000  0.672024
Active      0.672814  0.350079   0.672024  1.000000
```

MANAGERIAL INSIGHTS AND IMPLICATIONS

1. **Regional Variations in COVID-19 Impact:**
    - The analysis reveals significant regional variations in the impact of COVID-19, with some states experiencing higher confirmed cases, deaths, and active cases compared to others.
    - Managers and policymakers should recognize that a one-size-fits-all approach to pandemic response may not be effective. Tailored strategies are needed to address the specific challenges faced by different regions.

2. **Identifying High-Risk Areas:**
    - Understanding the regional distribution of COVID-19 cases can help managers and policymakers identify high-risk areas.
    - Resource allocation, including healthcare resources, testing facilities, and vaccination campaigns, can be more effectively targeted to regions with the highest need.
3. **Resource Allocation for Healthcare Facilities:**
    - Managers can use the regional distribution of COVID-19 cases to inform decisions regarding the allocation of hospital beds, medical equipment, and healthcare personnel.
    - Hospital administrators can adjust their capacity planning to respond to surges in cases effectively.
4. **Data-Driven Decision-Making:**
    - The project underscores the importance of data-driven decision-making during a public health crisis. Accurate and up-to-date data is crucial for identifying trends, making informed decisions, and monitoring the effectiveness of interventions.
    - Managers should invest in data collection, analysis, and reporting infrastructure to support ongoing decision-making.
5. **Public Communication Strategies:**
    - Managers should recognize that the public's perception of risk and willingness to comply with public health measures can vary by region.
    - Tailored communication strategies that consider regional differences can be more effective in conveying the urgency of preventive measures and vaccination.
6. **Collaboration and Information Sharing:**
    - Collaboration among states or regions with similar COVID-19 trends can be beneficial for sharing best practices, resources, and strategies.
    - Managers should encourage inter-state cooperation and information sharing to improve pandemic response.

7. **Flexibility and Adaptation:**
    - The project highlights the need for flexibility in response strategies. As the situation evolves, managers must be prepared to adapt policies and resource allocation based on changing trends and new data.
8. **Monitoring and Surveillance:**
    - Continuous monitoring and surveillance of COVID-19 data are essential for early detection of outbreaks and timely response.
    - Managers should establish robust surveillance systems that can provide real-time data to support decision-making.