# Indian Institute of Technology Gandhinagar


# CS 328
# Data Science

Authors

20110130 Dishant Patel

20110079 Hritik Ladia

20110165 Rishabh Patidar

18110006 Abhinav Singh

Under the Supervision of
**Prof. Anirban Dasgupta**

## Application of Clustering On State Wise Time Series Data Of COVID-19 In India

30th April, 2023

Repository Link

# Index

# Introduction

The COVID-19 pandemic has been one of the most significant global health crises in recent history. It has affected millions of people worldwide and caused unprecedented social and economic disruptions. The pandemic started in late 2019 in China and quickly spread to other countries through international travel and trade. The World Health Organization (WHO) declared it a public health emergency of international concern on January 30, 2020, and a pandemic on March 11, 2020.

India was one of the countries that was severely hit by the pandemic. The first case of COVID-19 in India was reported on January 30, 2020, in Kerala. Since then, the virus has spread to all states and union territories of India, with varying degrees of intensity and impact. As of April 2023, India has reported over 30 million confirmed cases and over 400,000 deaths due to COVID-19, making it the second worst-hit country after the United States. However, there is a lot of variation and uncertainty in the data reported by different sources and regions in India. Some of the challenges faced by India in collecting and reporting accurate data include:

1. Lack of adequate testing and tracing facilities and resources
2. Inconsistencies and delays in reporting by different authorities and agencies
3. Underreporting or misreporting of cases and deaths due to stigma, fear, or political pressure
4. Differences in definitions and criteria for testing, confirmation, and classification of cases and deaths
5. Lack of disaggregated data by age, gender, location, and other relevant factors
6. Therefore, it is important to analyze the data carefully and critically to understand the dynamics and impact of the pandemic in India. Data science can play a vital role in this process by providing tools and techniques to collect, process, analyze, visualize, and communicate data effectively and efficiently.

# Problem Definition

The COVID-19 pandemic has had a devastating impact on the world, affecting millions of people, causing widespread illness, and leading to significant social and economic disruptions. Data analysis can play a critical role in understanding the spread of the virus and developing effective strategies to mitigate its impact. As part of this project, we aim to explore various datasets related to the COVID-19 pandemic and identify important trends and insights that might have helped policymakers and healthcare professionals make informed decisions. Some possible questions are:

1. How can we identify anomalies in the reported data and what are their possible causes?
2. How can we find correlations between case growth and other indicators like mortality rate?
3. How can we identify factors that contributed to certain regions doing better or worse than others in terms of controlling the pandemic?

To answer these questions, we will use data-driven approaches such as causal modeling and machine learning. The methodology for this project will involve data cleaning, exploratory data analysis, feature engineering, model building and evaluation, and geospatial analysis.

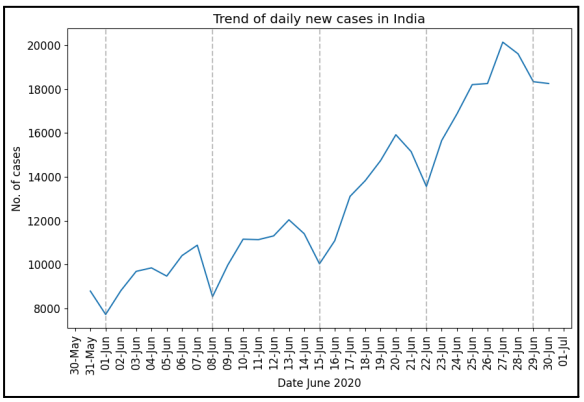
# Broad Overview of Our Methodology

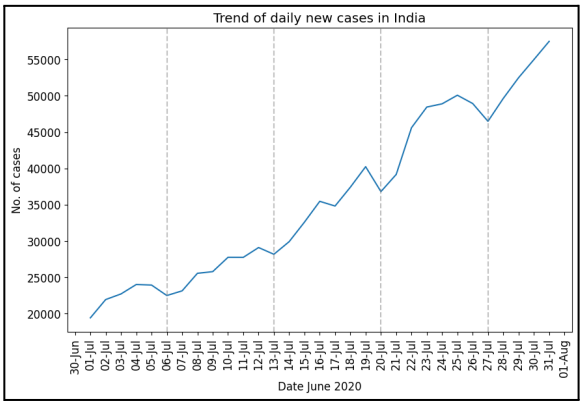
1. **Data Cleaning**: We collected data from various sources on different aspects of the pandemic. The data was not consistent and reliable across different sources and regions. We performed data cleaning to remove inconsistencies, missing values, and outliers.
2. **Exploratory Data Analysis (EDA)**: We conducted EDA to gain insights into the data and identify patterns and correlations. We used statistical methods, such as visualization and descriptive statistics, to explore the data. Through EDA, we were able to answer some basic questions about the data and generate some hypotheses for further analysis.
3. **Feature Engineering**: We created new features from existing ones to improve the performance of our models. For instance, we created a new feature that represents the percentage of the population vaccinated in a region. We also created a new feature that represents the case fatality rate (CFR).
4. **Model Building and Evaluation**: We built machine learning models such as clustering to identify factors that contributed to certain regions doing better than others.

# Application of Clustering:

## Data Collection and Preparation:

The primary data source utilized for this study is the COVID19-India API, which provides daily data on confirmed cases, deaths, and recoveries at the state and district level in India. The data for daily new cases and deaths were collected from July 2020 to December 2020 for each state in India for clustering and from May 2020 to April 2021 for each state in India for heatmaps.




In the data preparation phase, we identified an anomaly in the reporting of cases, specifically a consistent drop in the number of daily new cases reported on Mondays compared to other weekdays. Upon further investigation, we determined that this was due to a lower number of tests being performed on Sundays for administrative reasons, as evidenced by the figures presented above.

To address this anomaly and ensure the robustness of our clustering analysis, we employed the Sakoe-Chiba radius parameter with a value of r = 7 when computing the DTW score for every pair of states. By setting the Sakoe-Chiba radius parameter to a value of r = 7, we limit the warping between the two time series being compared to a maximum of 7 time steps. This effectively makes the clustering analysis more robust to Sunday anomalies and any other delays in the reporting of cases that may occur.

It has been observed that many studies and reports on COVID-19 data have used states as a dimension of reporting, despite the fact that each state in India has a varying area and population. This can lead to bias towards larger states, as they are likely to have a higher number of COVID-19 patients. To mitigate this bias, we have chosen to normalize the COVID-19 patient data by dividing it by the total population of each state. This normalization allows us to make meaningful comparisons between regions with varying population sizes, including smaller states like those in the north-eastern region of India and union territories.

## Model and Techniques:

### Case rate and Mortality rate:

We define case rate for region r on day t as

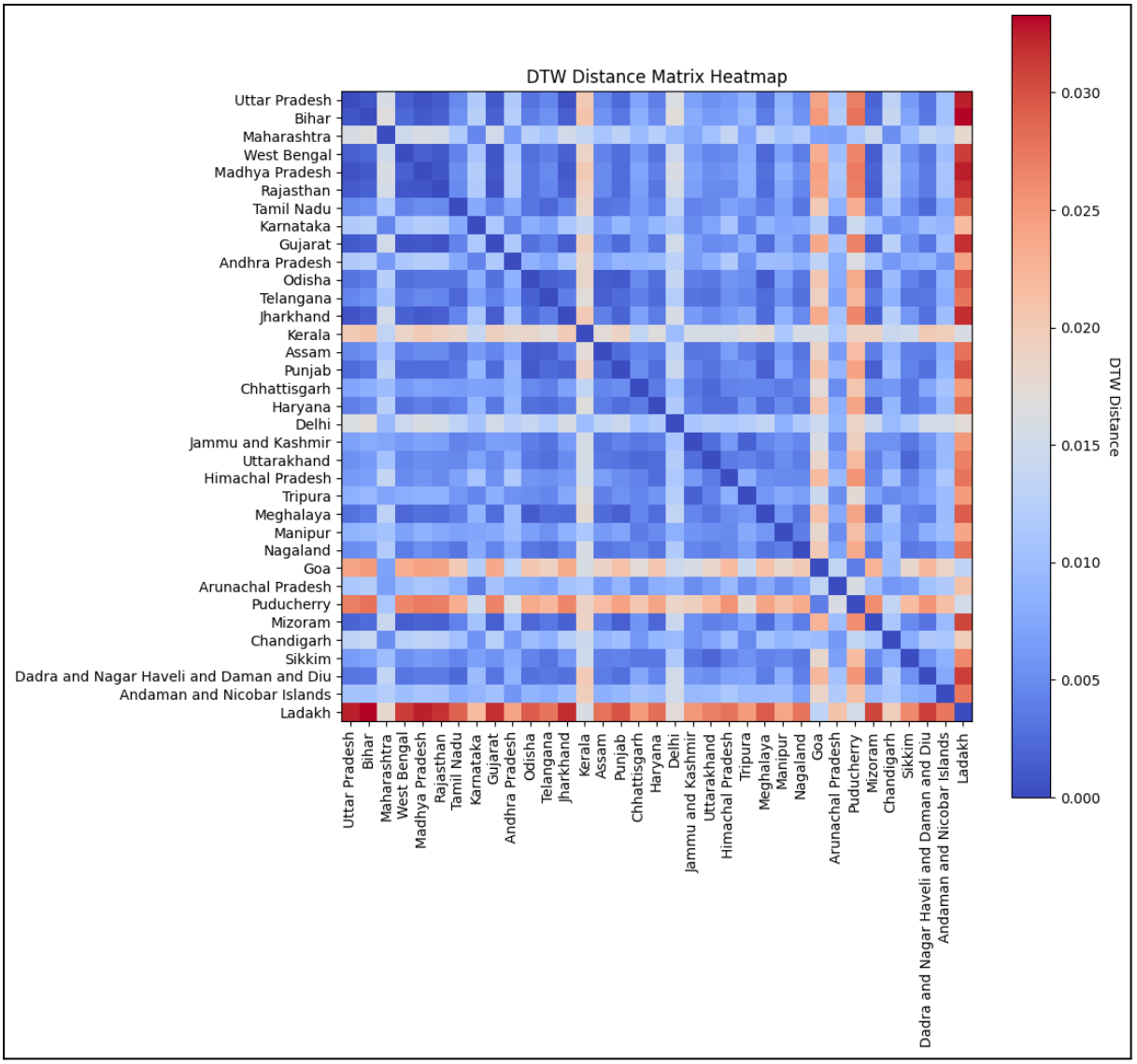
$$\textbf{Case rate} = \frac{Total\ number\ of\ new\ cases\ in\ region\ r\ on\ day\ t}{Population\ of\ region\ r}$$

We define mortality rate for region r on day t as

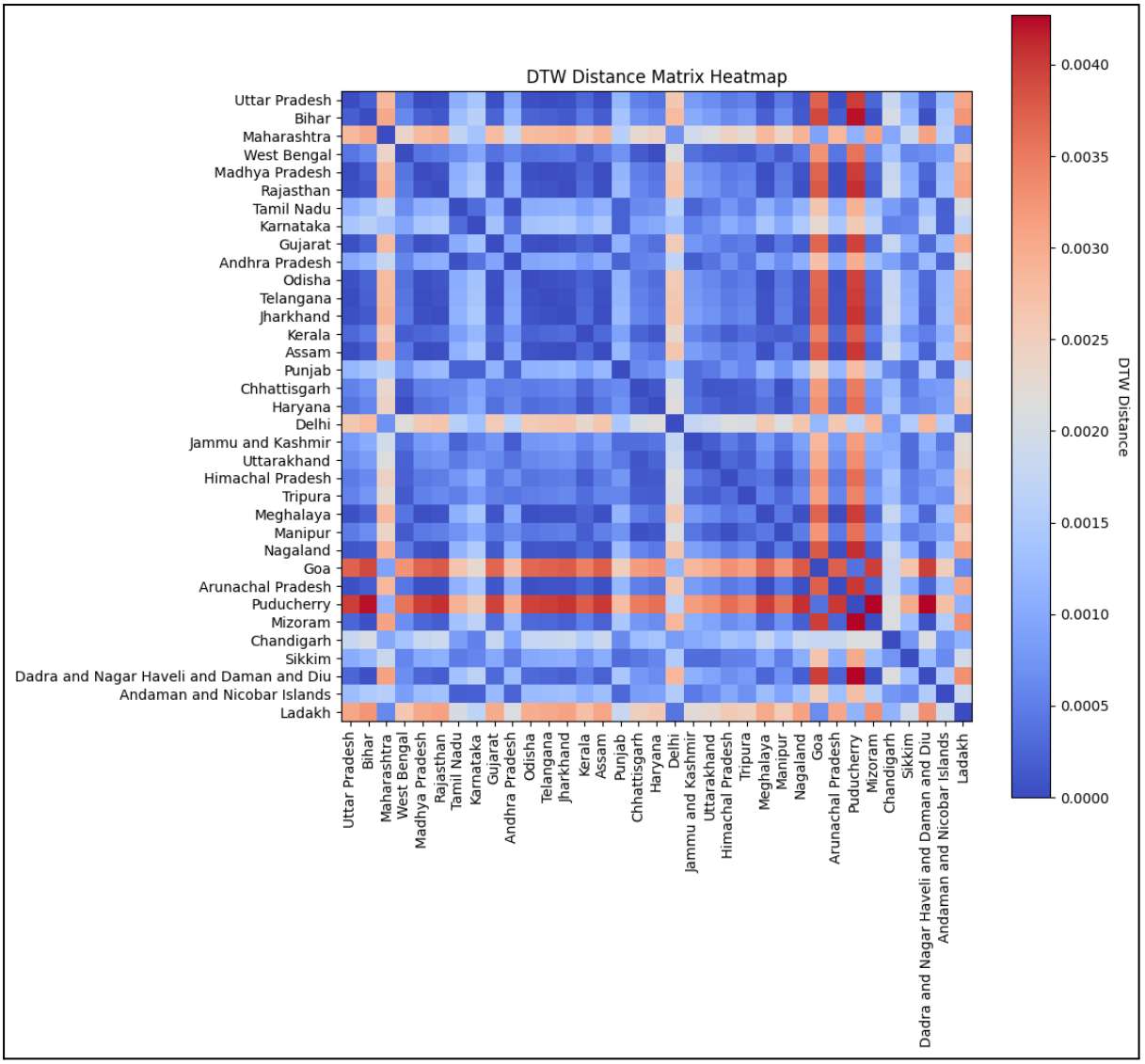
$$\textbf{Mortality rate} = \frac{Total\ number\ of\ deaths\ in\ region\ r\ on\ day\ t}{Populatio\ ,of\ region\ r}$$

### Distance Matrix:

Our main goal is to compare the patterns of COVID-19 rise and decline across various states and districts in India. To achieve this, we utilized a shape-based approach for time series clustering, as some regions experienced multiple waves of infections. Dynamic Time Warping (DTW) score, which measures the similarity between two time series based on their shapes, was employed for this analysis. DTW (Dynamic Time Warping) is preferred over the traditional Euclidean distance metric for time-series clustering applications due to its ability to handle time series that are of different lengths and may be warped in time. Unlike Euclidean distance, which is sensitive to shifts and distortions in the time series, DTW aligns the two series in a way that minimizes the difference between them. This alignment process allows DTW to capture the shape similarity between two time series, even when they have different lengths or when there are nonlinear time warps. As a result, DTW has been shown to have higher accuracy for time-series clustering applications. The lower the DTW score $DTW(x,y)$ of a pair of time series $x$ and $y$, the similar they are, with $DTW(x,x) = 0$. DTW is a symmetric function, with $DTW(x,y) = DTW(y,x)$. The distance matrix for the above clustering was prepared based on this score.




DTW Heatmap for Case rate




DTW Heatmap for Mortality rate

## Agglomerative clustering:

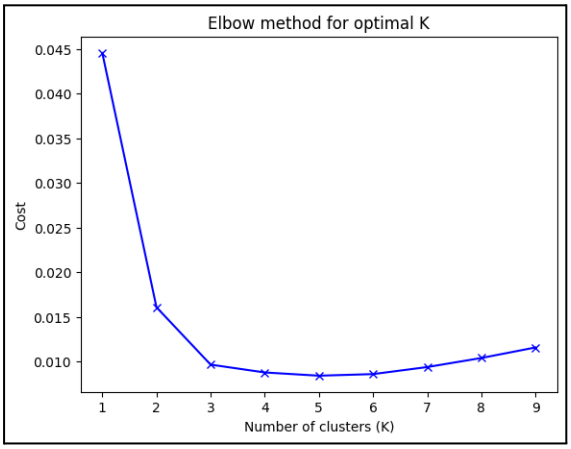
The resulting distance matrix was then used to perform hierarchical clustering using agglomerative clustering. This is a widely used technique in data mining and statistics to build a hierarchy of clusters, where each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy.

## Elbow Method:

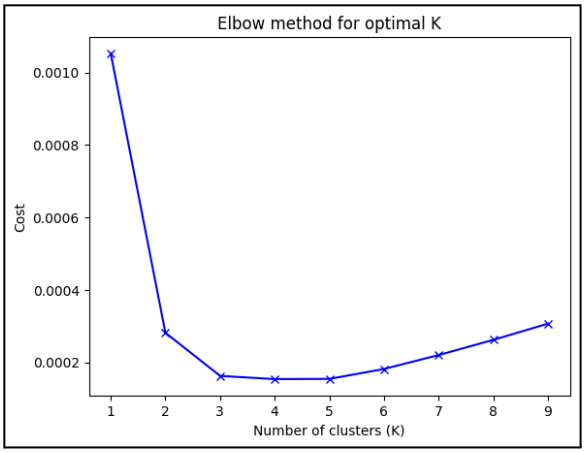
The elbow method is a commonly used heuristic in time series K-means clustering to determine the optimal number of clusters based on a cost function. Our cost function combines the Residual Sum of Squares (RSS) with a penalty term that accounts for the number of clusters, represented by "k". Specifically, we define the cost as the sum of RSS and α times the logarithm of k.

$$\textbf{Cost} = \textbf{RSS} + \alpha \times log(k)$$

After applying the elbow method we have figured out the optimal number of clusters for Case rate analysis is 5 and Mortality rate analysis is 4.



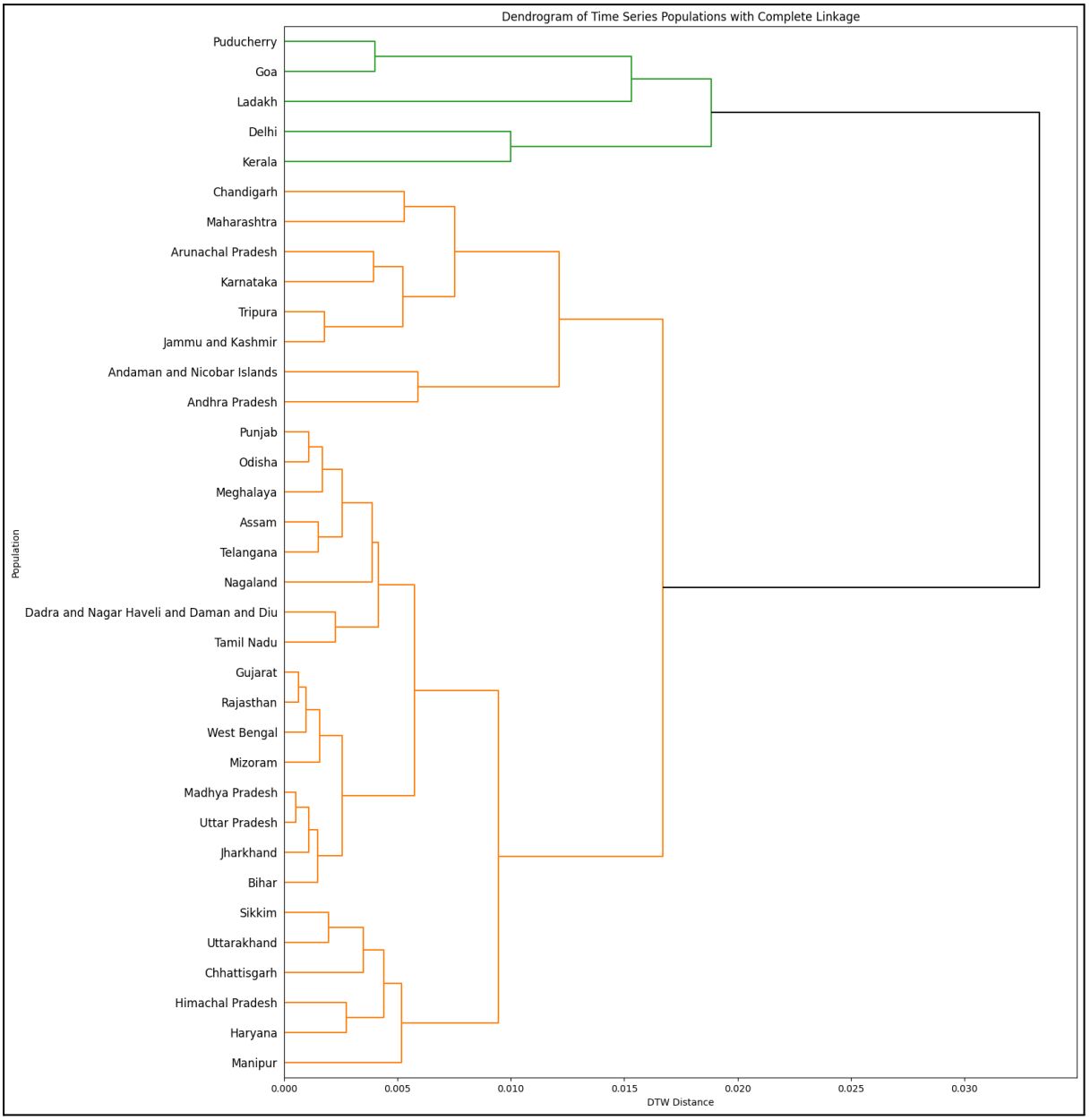

Cost function vs no. clusters for case rate




Cost function vs no. of clusters for mortality rate

## Dendrogram:

The results of a hierarchical clustering is commonly represented by a dendrogram, a tree-like diagram which describes the series of steps taken by the clustering technique from $n$ distinct singleton clusters to a single cluster containing all $n$ individuals.

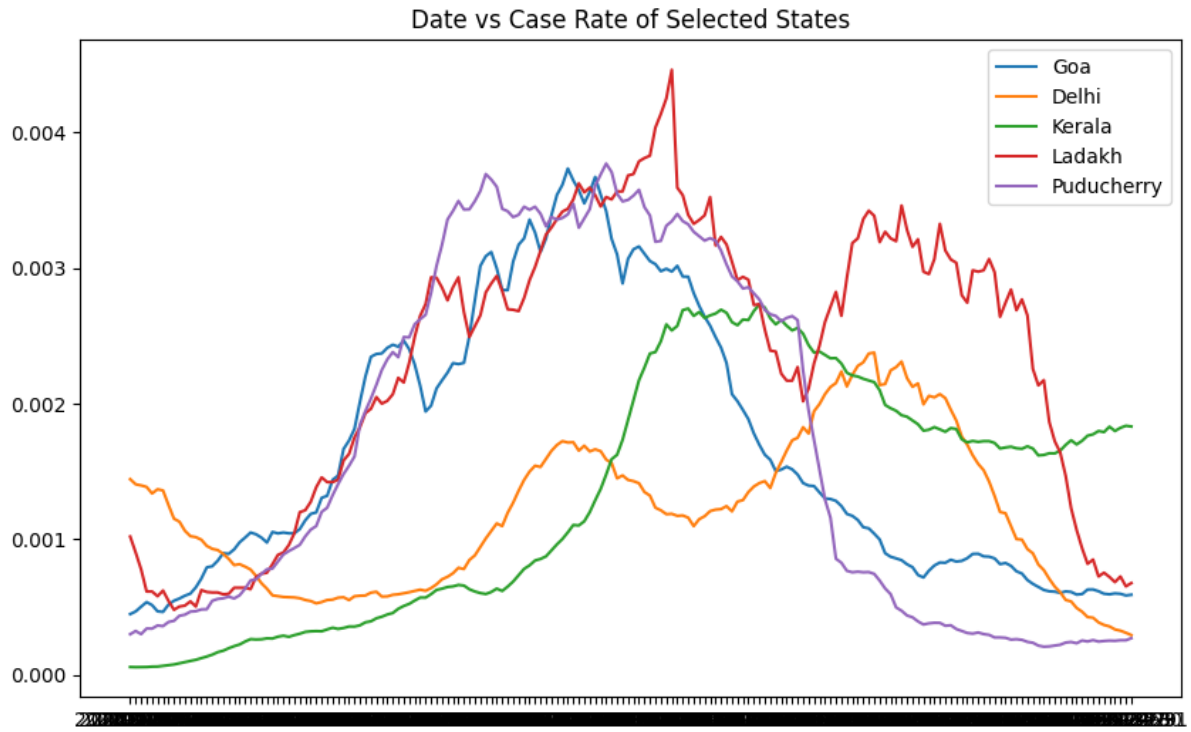



Dendrogram for states using Complete Linkage in *Case Rate* analysis

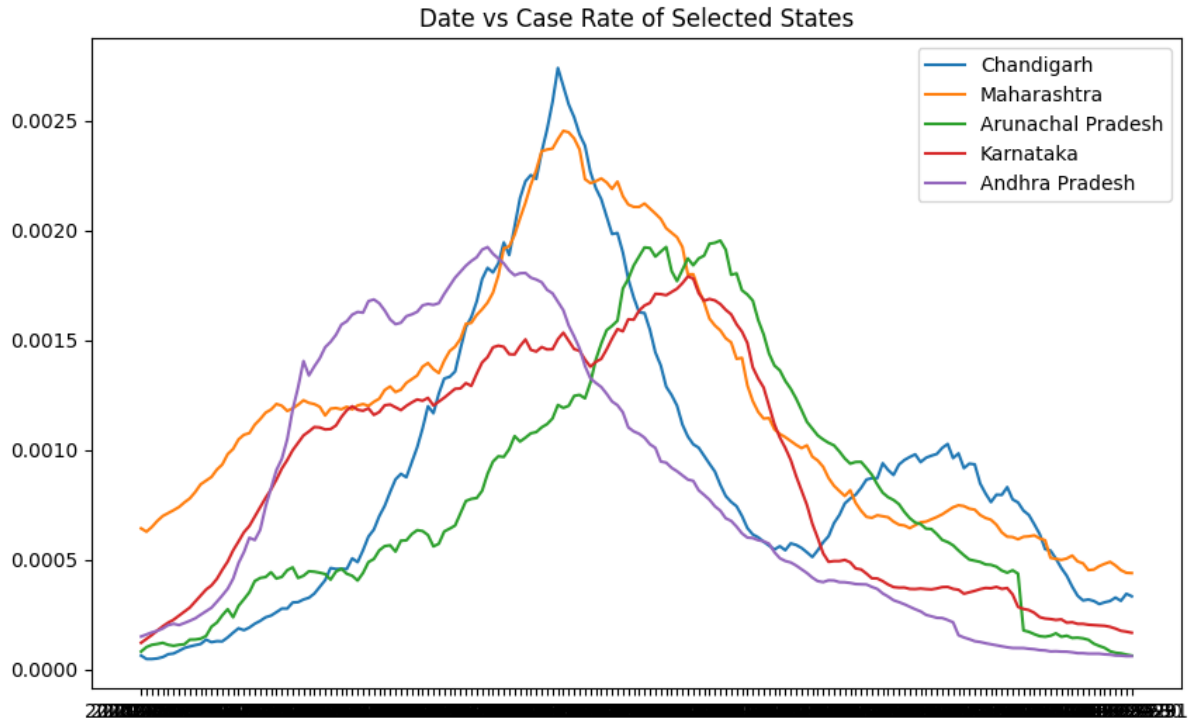
We now provide a brief explanation of the time series clusters obtained and characterize the same.

## 1. Case rate analysis:

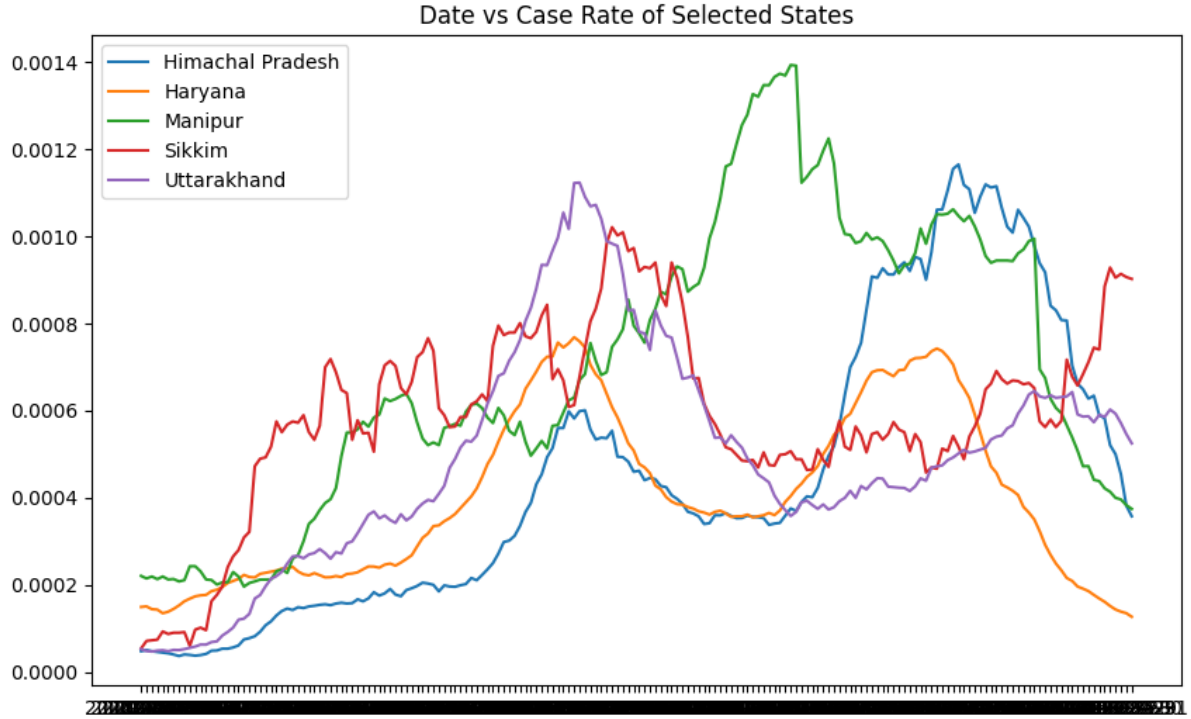
***Cluster 1:*** This cluster has the highest peak of *Case rate*. The states Goa, **Delhi,** Kerala, Ladakh, Puducherry of this cluster may well be labeled as the "Critical states" since they experienced the highest *Case rate* with multiple peaks, shown in figure below,




***Cluster 2:*** This cluster (Hotspot states) may be regarded as the set of states which have experienced significantly higher *Case rate* than the national average but have still performed better than the Critical states. As evident in Figure below there is a consistent rise and decline in the states with the peak being just relatively lower than the critical states. States belonging to this cluster are Chandigarh, Maharashtra, Arunachal Pradesh, Karnataka, Andhra pradesh, **Tripura, J&K**, Andaman and Nicobar.

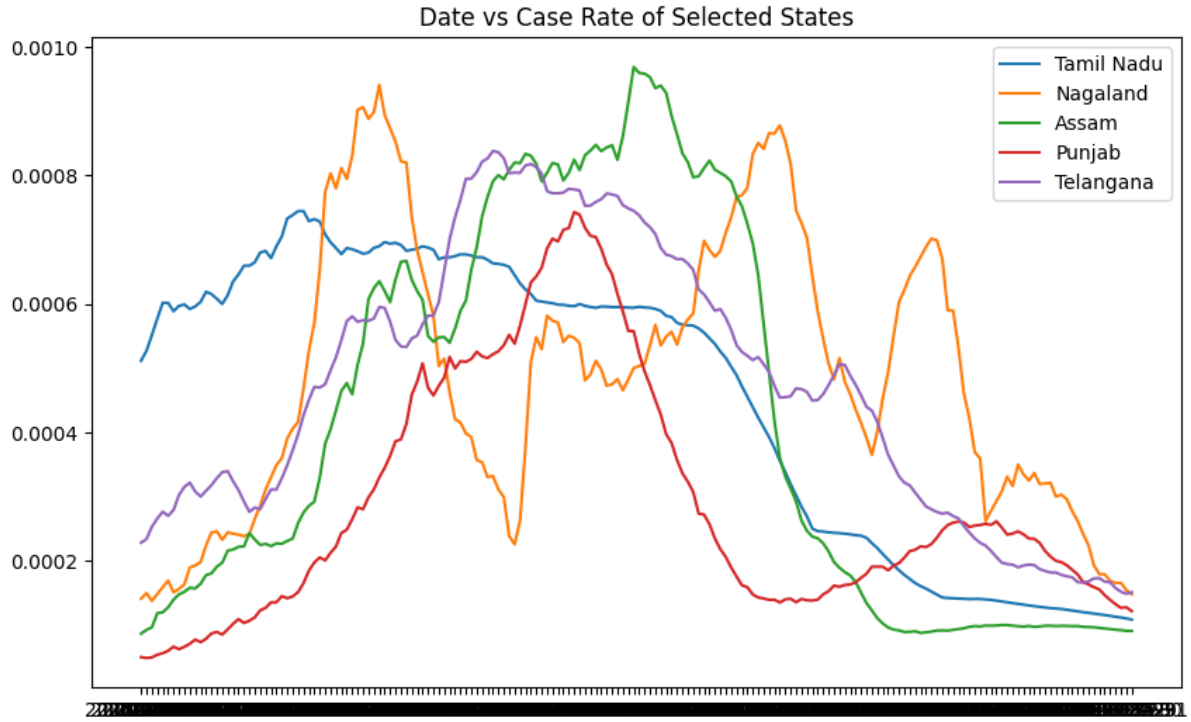


***Cluster 3:*** This cluster (Severe Zone states) consists of states that have a lower case rate compared to the Critical and Hotspot states, but still high enough to pose a threat. States belonging to this cluster are Himachal Pradesh, **Haryana**, Manipur, Sikkim, Uttarakhand, and Chattisgarh.

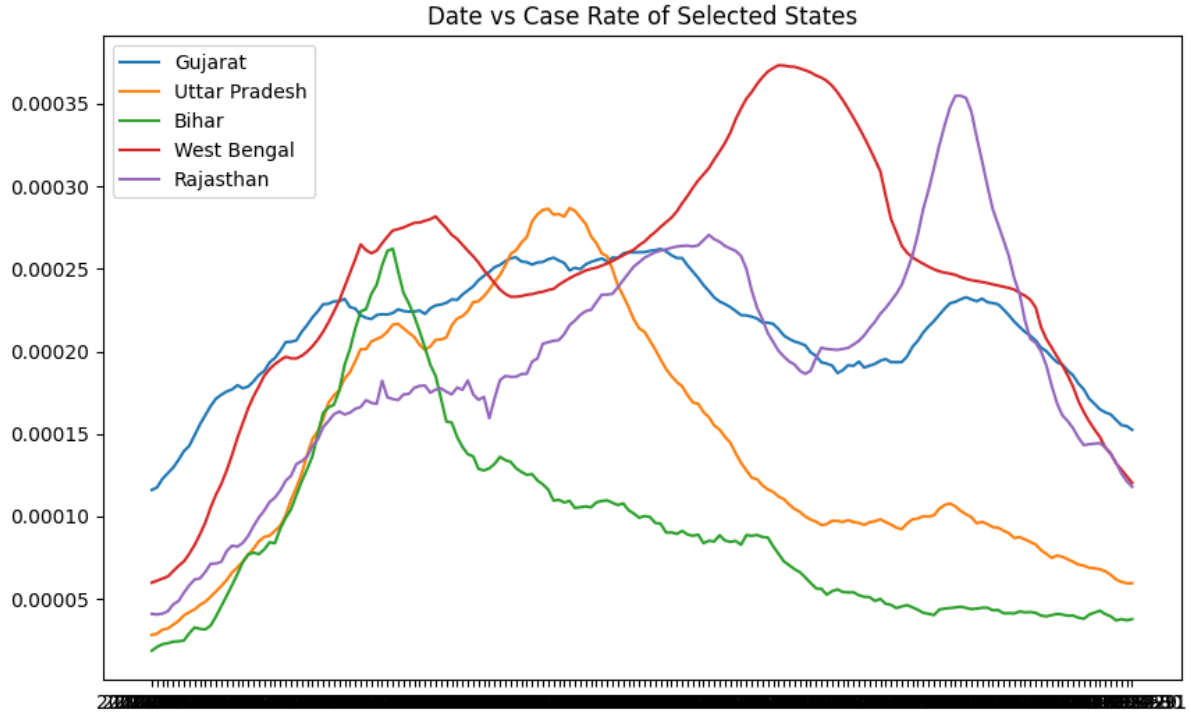



***Cluster 4:*** This cluster (Moderate states) can be considered as the set of states which have had a moderate case rate during the COVID-19. As seen in the Figure below, the states belonging to this cluster have had a relatively lower case rate compared to the hotspot, critical states and severe states. This cluster includes Tamil nadu, Nagaland,

Assam, Punjab, Telangana, Meghalaya, Odisha and Dadra nagar haveli.

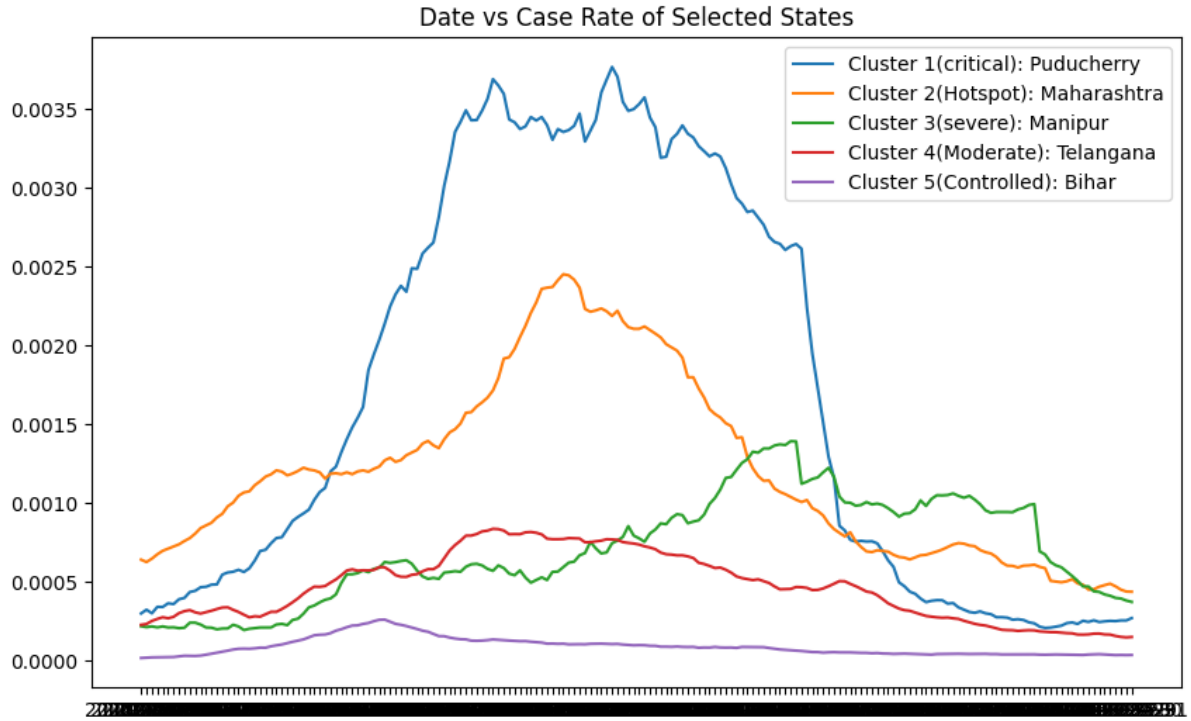



***Cluster 5:*** The plateau-like nature with not many sharp peaks depicted by the curve of the states belonging to this cluster, indicates that these states performed very well in comparison to other states. Thus, these states have a controlled and subdued outbreak of the pandemic. Interestingly, these states represent a compact geographical region in the northern plains of India as shown in Figure. This cluster includes Gujarat, Uttar pradesh, Bihar, West Bengal, Rajasthan, Mizoram, Madhya pradesh and Jharkhand.




We can now try to summarize the entire picture using the below inter-cluster plot of

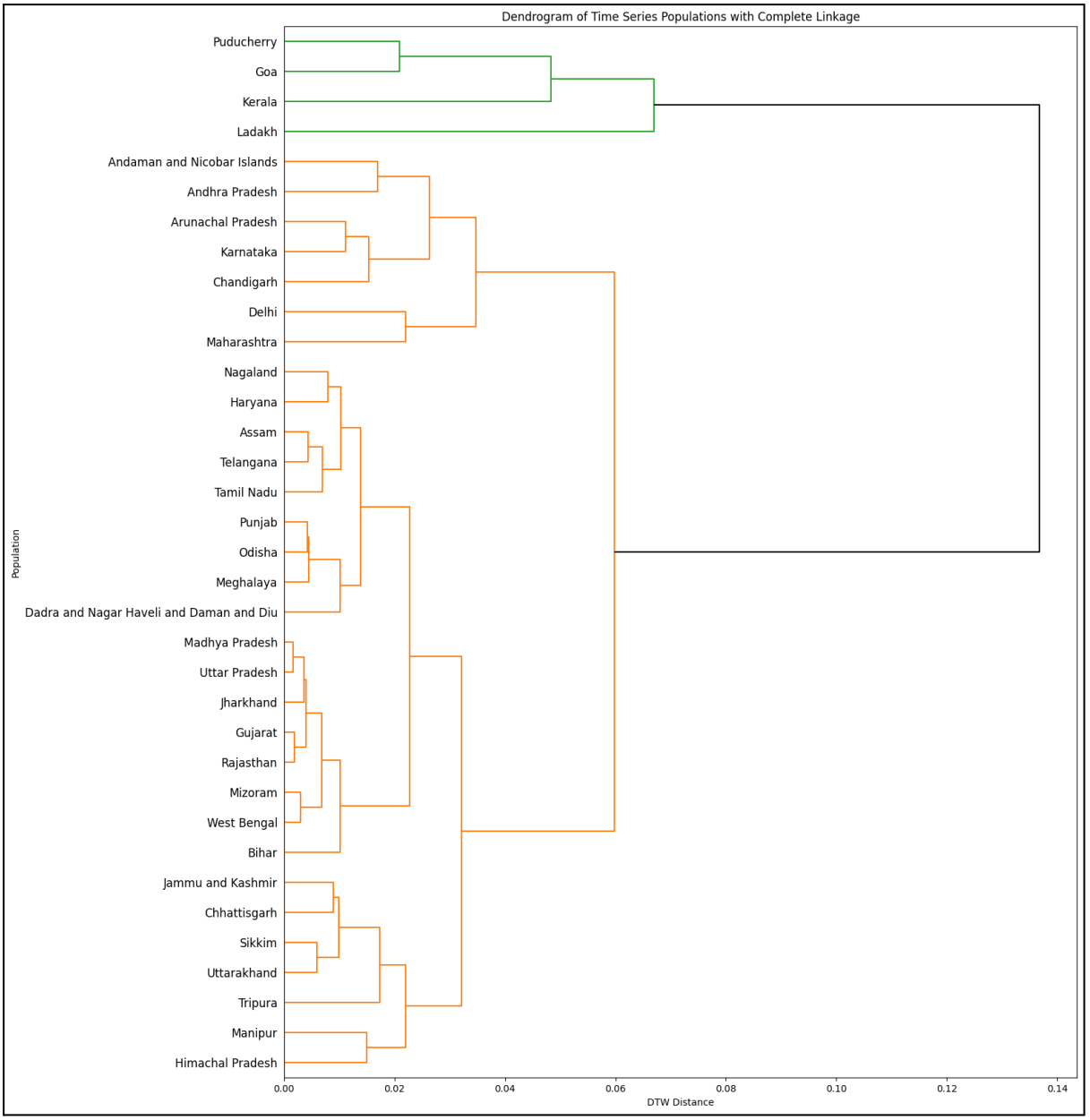
states shown in Figure below. For this figure, a representative state has been picked up from each of the above four clusters.




To compare the current model for clustering, we implement the standard Euclidean distance instead of DTW score between two time-series. On comparing the clusters produced by DTW model with that of the clusters produced by Euclidean model, we observe the following differences:

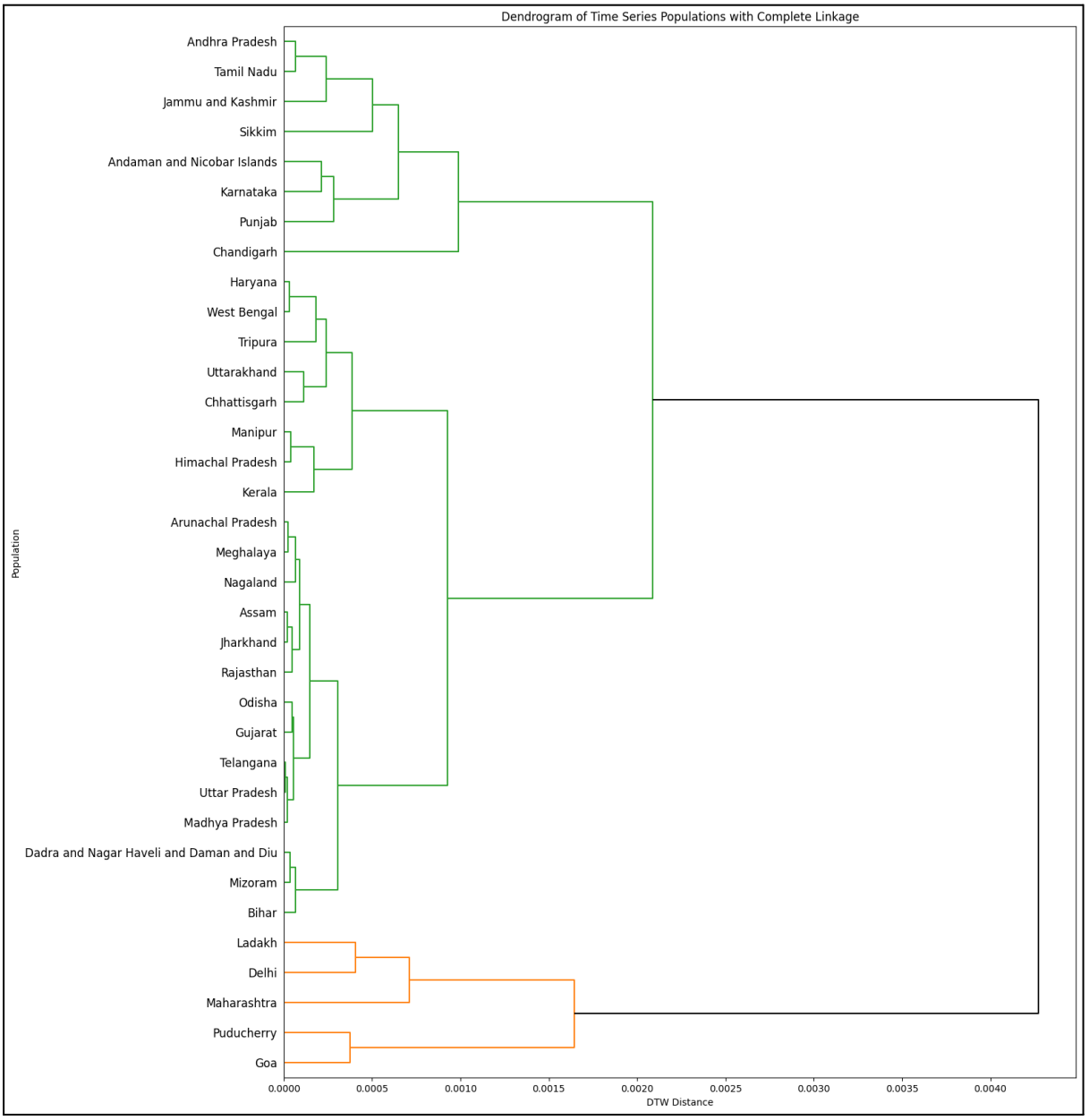
1. Delhi shifts from Critical to Hotspot states.
2. Tripura and J&K shifts from Hotspot to Severe states.
3. Haryana shifts from severe to moderate states.

This is mainly due to Sakoe-Chiba radius which is explained above, which is used in calculation of the DTW score, being absent in this Euclidean case. This makes the Euclidean distance between Delhi and other time series of *Critical Cluster* larger than that of Delhi and other time series of *Hotspot Cluster*. So, Kerala shifts to Cluster 2 from Cluster 1 in this case. The same argument remains valid for Tripura, J&K and Haryana.



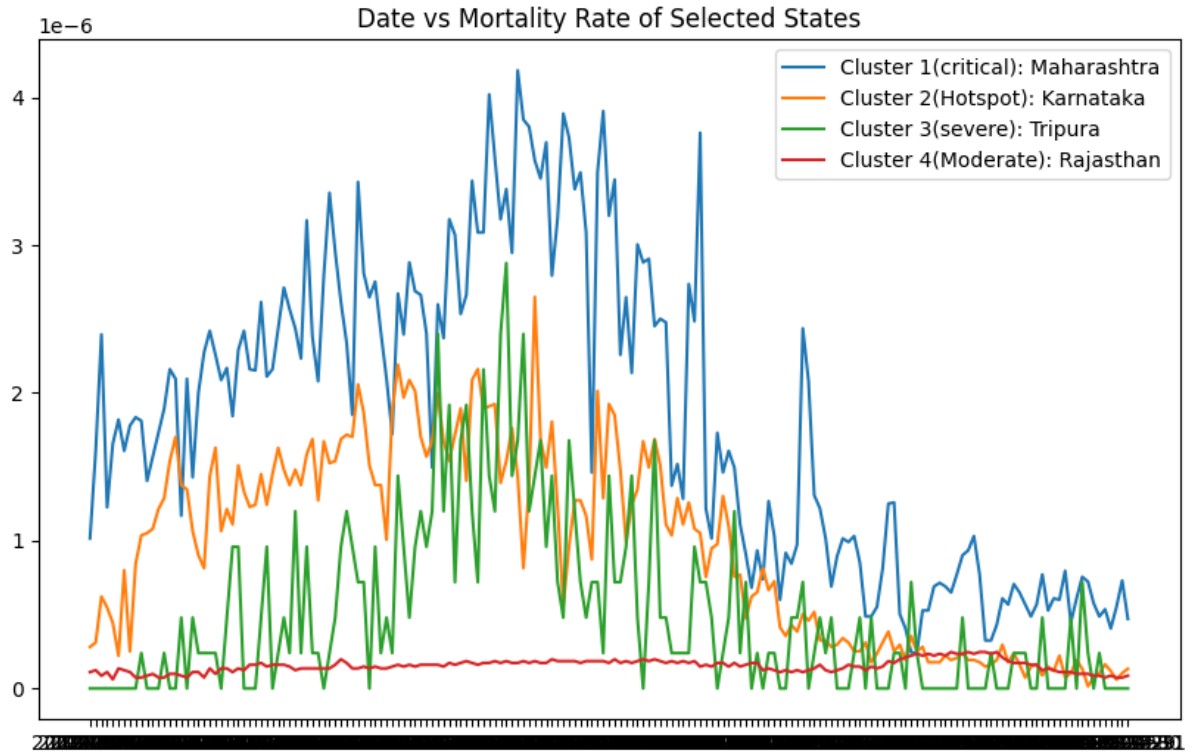

Dendrogram for states using Complete Linkage in *Case Rate* analysis using Euclidean distance matrix

## 2. Mortality Rate Analysis:




Dendrogram for states using Complete Linkage in *Mortality Rate* analysis

We apply the above methods for clustering and analysis of *mortality rate* time series at the state level. We obtain four distinct clusters of the states. The *mortality rate* time series for a representative state from each of the clusters is plotted in Figure below,




| | Cluster A | Cluster B | Cluster C | Cluster D |
|---|---|---|---|---|
| **Cluster 1** | LA, DL, MH, PD, GA | | KL | |
| **Cluster 2** | | CH, AR, KA, AN | | AP, TR |
| **Cluster 3** | | | HR, CG, UK, | |
| **Cluster 4** | | PB | | NL, AS, TG, ML, OD, DN |
| **Cluster 5** | | | WB | GJ, UP, BH, RJ, MZ, MP, JH |

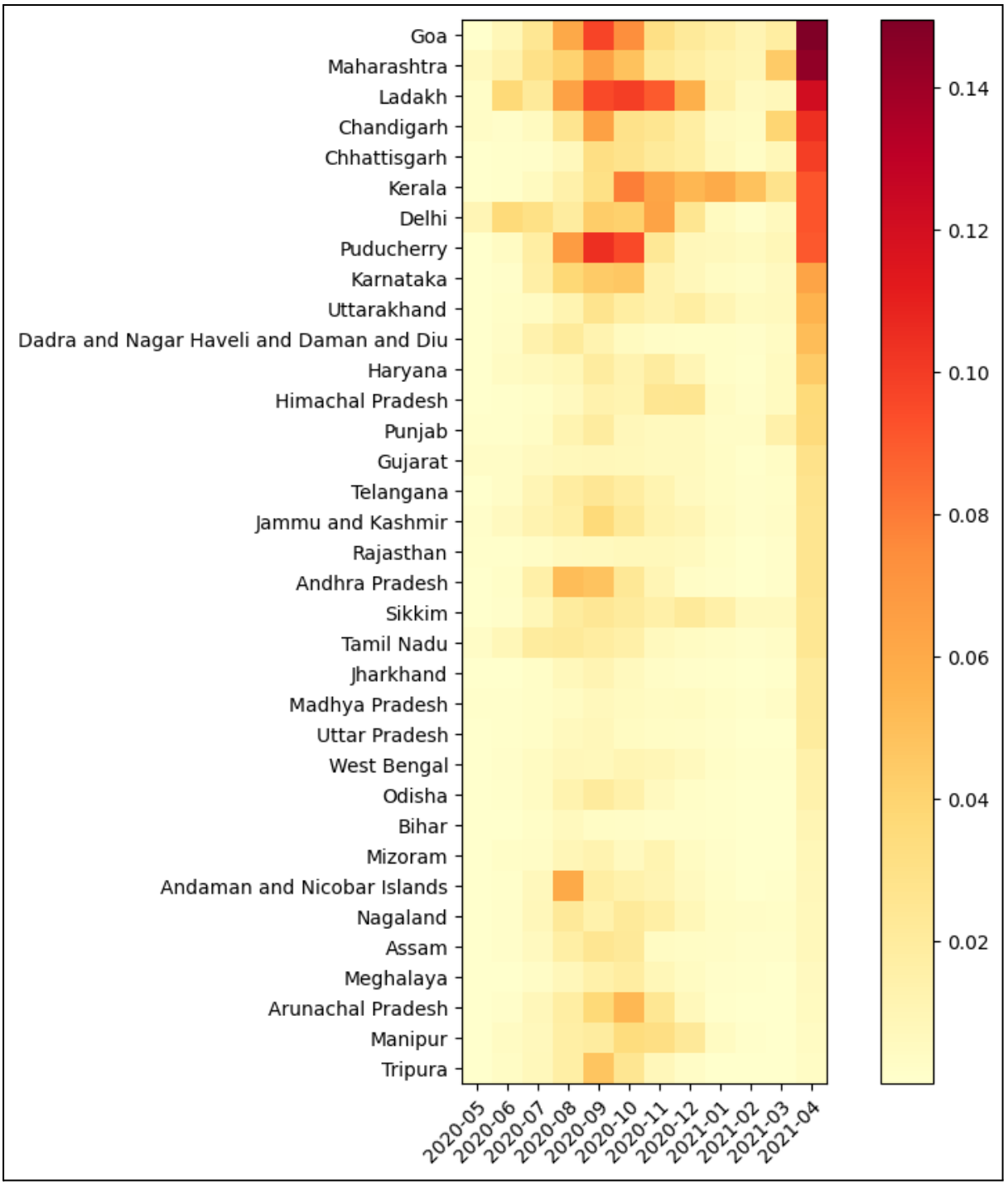
The mortality rate clusters are ordered according to their mortality rate from highest to lowest: A > B > C > D. Suppose a cluster of ith row has a one-to-one correspondence with a cluster of jth column, then any state in cell (i,p) has performed relatively badly if the cluster of pth column is higher in the ordering than the cluster of the jth column and vice versa.

For instance, Kerala (KL) in cell (1, C) has performed better in terms of mortality rate as it is associated with Cluster 1, which is correlated with Cluster A. However, Punjab(PB) in cell (4, B) has performed worse in terms of mortality rate as it is associated with Cluster 4, which is correlated with Cluster D, but it is placed in Cluster B.

The clustering analysis conducted on the data revealed that the clusters tended to form a compact geographical region, which suggests that the spread of COVID-19 is influenced by distance contiguity. It was observed that hotspot clusters included districts in major cities such as Delhi, Pune, and Bengaluru, indicating that regions with good connectivity and a significant economy were more susceptible to the virus. Additionally, the analysis of mortality rates revealed that the number of cases in a region was not necessarily an accurate indicator of its ability to handle the pandemic. Certain regions were found to be more vulnerable in terms of fatalities, highlighting the need for targeted public health interventions in different regions to effectively combat the pandemic.

## 3. Some other analysis from Heatmaps:

### 1. Month-wise Heatmap of different states:

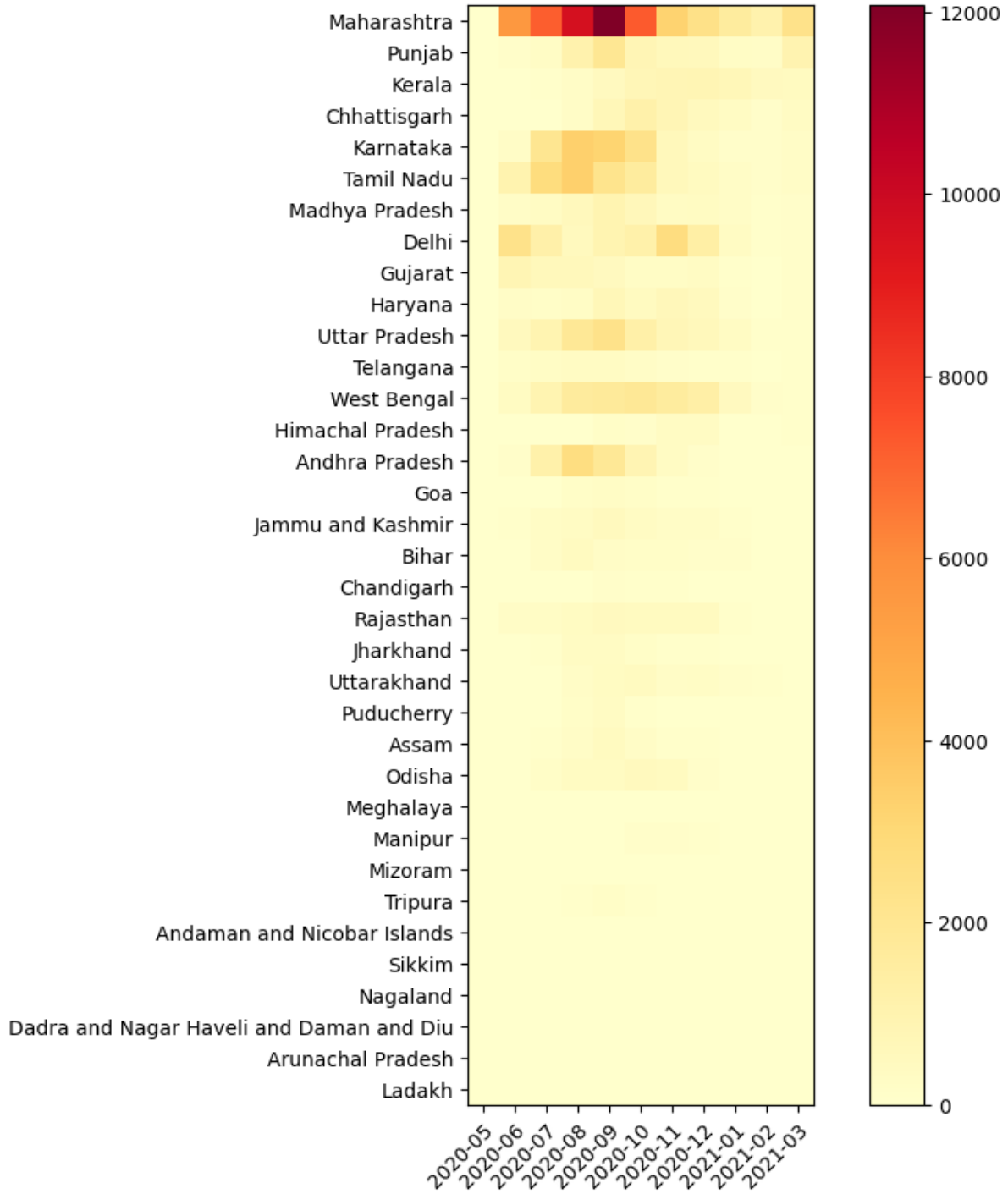



Heatmap for ranking of states based on *Case Rate*

During the period between May 2020 and April 2021, Goa had the highest monthly case rate among all the states in India. On the other hand, Kerala had moderate case rates until September 2020, but then experienced a rapid increase in cases, which led to its ranking as the 6th highest state in the heatmap. Tamil Nadu had a high case rate until

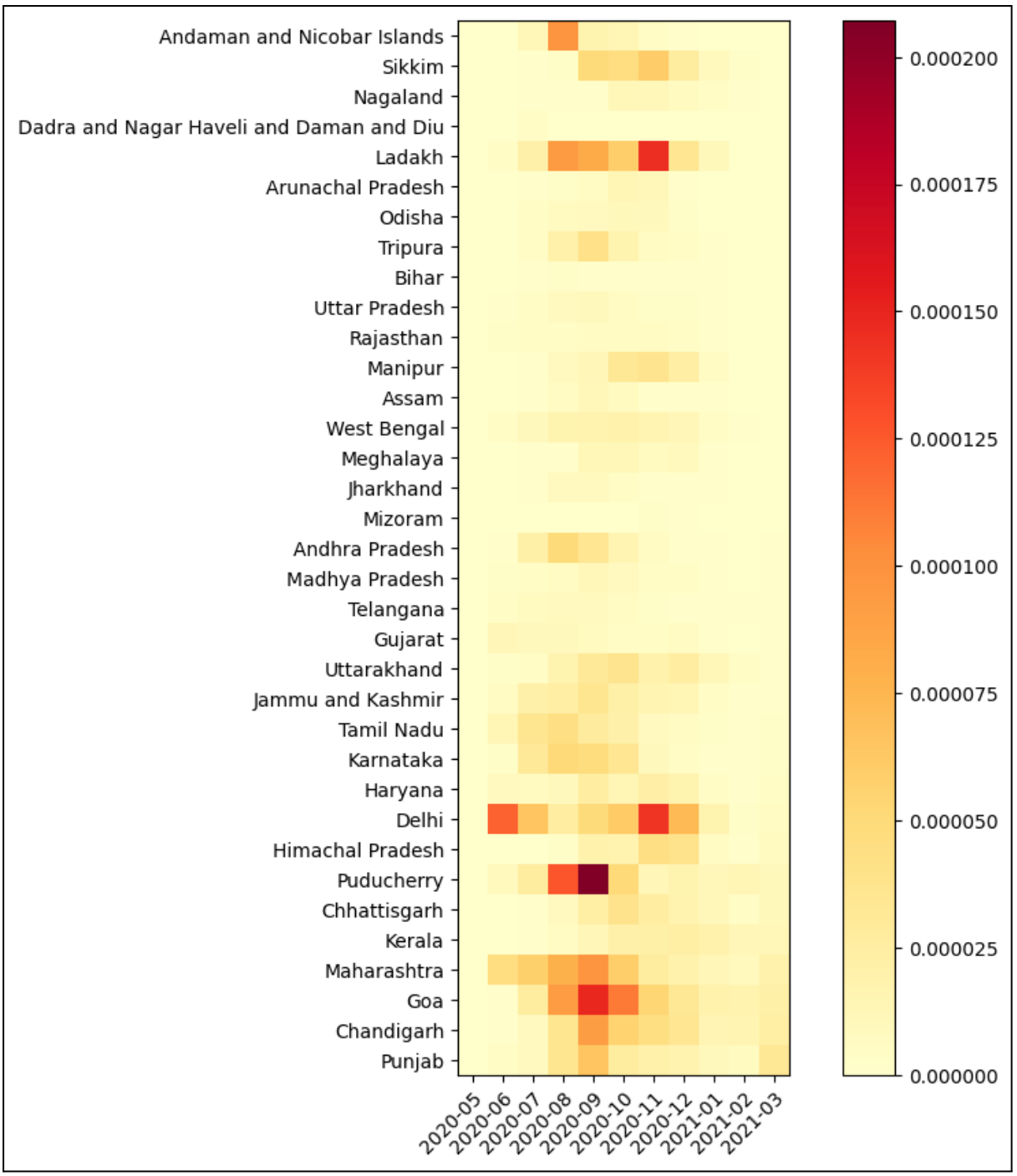
October 2020, but the situation improved afterward. It is important to note that these findings highlight the need for continuous monitoring and adaptive strategies in response to changing trends and patterns of the pandemic in different regions. As such, it is imperative for health officials and policymakers to remain vigilant and agile in their decision-making processes.

**2. Heatmap of absolute daily new deaths:**




Heatmap for ranking of states based on absolute daily new deaths

In the above Figure, we have analyzed the daily death count clustered over a period of 30 days from May 1, 2020 to March 30, 2021. The figure highlights that highly populated states like Maharashtra, West Bengal, Uttar Pradesh have recorded the highest number of deaths. Despite being less populous states (Delhi nineteenth and Punjab sixteenth) in terms of total population, Delhi and Punjab have recorded high daily death counts, which can be attributed to their high population density. Delhi, for example, has recorded more than two distinct waves of infections, unlike most other states. Gujarat, the ninth most populous state in India, has improved its situation after being one of the worst affected states in the first few months. Bihar, despite its high population, has a relatively low daily death count, which may be attributed to under-reporting. To overcome such anomalous results, it is essential to analyze the data using normalized metrics like death rates.




The above figure illustrates the cluster dynamics of the mortality rate, indicating that smaller population states such as Chandigarh, Goa, Ladakh and Puducherry were significantly affected by COVID-19. Each state has a unique pattern of Mortality rate over time. Gujarat has managed to keep its mortality rate relatively low since June 2020, possibly due to its lower case rate during that period as shown in Heatmap of Case rate. However, Kerala, which had lower mortality rate from April to October 2020, experienced an increase in mortality rate from November 2020, when its case rates were higher than the rest of the country (From the heatmap of case rate).

# Conclusion:

- The study uses case rate and mortality rate to analyze COVID-19 data across various states and districts in India. Dynamic Time Warping (DTW) score is used to measure the similarity between two time series based on their shapes for time series clustering afterward using the Euclidean distance matrix, we compared the two to figure out changes in clusters like 4 states (Delhi, Tripura, J&K and Haryana) were in different clusters for both the method. Agglomerative clustering is used to perform hierarchical clustering, and the elbow method is used to determine the optimal number of clusters based on a cost function. A dendrogram is used to represent the results of hierarchical clustering.
- We also established a useful correspondence between clusters of *case rate* and *mortality rate* at the state level and identified that some of the states like Kerala did well in reducing the mortality risks, whereas Punjab is the only state which had relatively higher mortality risks than expected.
- The study provides a brief explanation of the time series clusters obtained for the case rate analysis and characterizes each cluster. It identifies five clusters, including the "Critical states", “Hotspot states", “Severe states”, “Moderate states” and "Controlled states". Similarly, the study identifies four clusters for the mortality rate analysis, including the "High mortality states", "Medium mortality states," "Low mortality states," and "Negligible mortality states." dendrograms are used to visualize the results.
- From the analysis conducted on the data, it can be concluded that effective handling of the COVID-19 pandemic requires a targeted approach that takes into account the unique characteristics of different regions. This includes not only the number of cases but also the mortality rates, as certain regions may be more vulnerable in terms of fatalities like Punjab. Furthermore, the findings suggest that distance contiguity and economic connectivity play a role in the spread of the virus, highlighting the need for measures such as social distancing and targeted economic interventions.
- The analysis of the data on COVID-19 cases and deaths in India highlights the need for continuous monitoring and adaptive strategies in response to changing trends and patterns of the pandemic in different regions. The heatmap of case rates reveals that highly populated states like Maharashtra, West Bengal, and Uttar Pradesh have recorded the highest number of cases. However, smaller population states such as Chandigarh, Goa, Ladakh, and Puducherry were also significantly affected by COVID-19. Each state has a unique pattern of different rates over time. Delhi, for example, has recorded more than two distinct waves of infections, unlike most other states.The heatmap of absolute daily new deaths shows that highly populated states have recorded the highest number of deaths, but Delhi and Punjab also had high daily death counts due to their high population density. These findings suggest that health officials and policymakers must remain vigilant and agile in their decision-making processes and adopt targeted public health interventions in different regions to effectively combat the pandemic. It is also important to note that the analysis of the data using normalized metrics like death rates can help to overcome anomalous results such as under-reporting in certain regions.

# References


1. https://data.covid19india.org/
2. Aleta, A., Blas-Laína, J.L., Tirado Anglés, G. *et al.* Unraveling the COVID-19 hospitalization dynamics in Spain using Bayesian inference. *BMC Med Res Methodol* **23**, 24 (2023). https://doi.org/10.1186/s12874-023-01842-7
3. Roy S, Ghosh P (2020) Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking. PLoS ONE 15(10): e0241165. https://doi.org/10.1371/journal.pone.0241165
4. Miralles‐Pechuán, Luis & Kumar, Ankit & Suárez-Cetrulo, Andrés. (2023). Forecasting COVID‐19 cases using dynamic time warping and incremental machine learning methods. Expert Systems. 10.1111/exsy.13237.
5. Sadeghi B, Cheung RCY, Hanbury M. Using hierarchical clustering analysis to evaluate COVID-19 pandemic preparedness and performance in 180 countries in 2020. BMJ Open. 2021 Nov 9;11(11):e049844. doi: 10.1136/bmjopen-2021-049844. PMID: 34753756; PMCID: PMC8578186.
6. Abdullah D, Susilo S, Ahmar AS, Rusli R, Hidayat R. The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. Qual Quant. 2022;56(3):1283-1291. doi: 10.1007/s11135-021-01176-w. Epub 2021 Jun 3. PMID: 34103768; PMCID: PMC8173859.
7. Rojas F, Valenzuela O, Rojas I. Estimation of covid-19 dynamics in the different states of the United States using time-series clustering, medRxiv 2020. https://doi.org/10.1101/2020.06.29.20142364.
8. Raj, A., Bhattacharyya, P. & Gupta, G.R. Clusters of COVID-19 Indicators in India: Characterization, Correspondence and Change Analysis. *SN COMPUT. SCI.* **3**, 210 (2022). https://doi.org/10.1007/s42979-022-01083-3
9. Shastri S, Singh K, Kumar S, Kour P, Mansotra V. Time series forecasting of covid19 using deep learning models: India-USA comparative case study. Chaos Solit Fract.