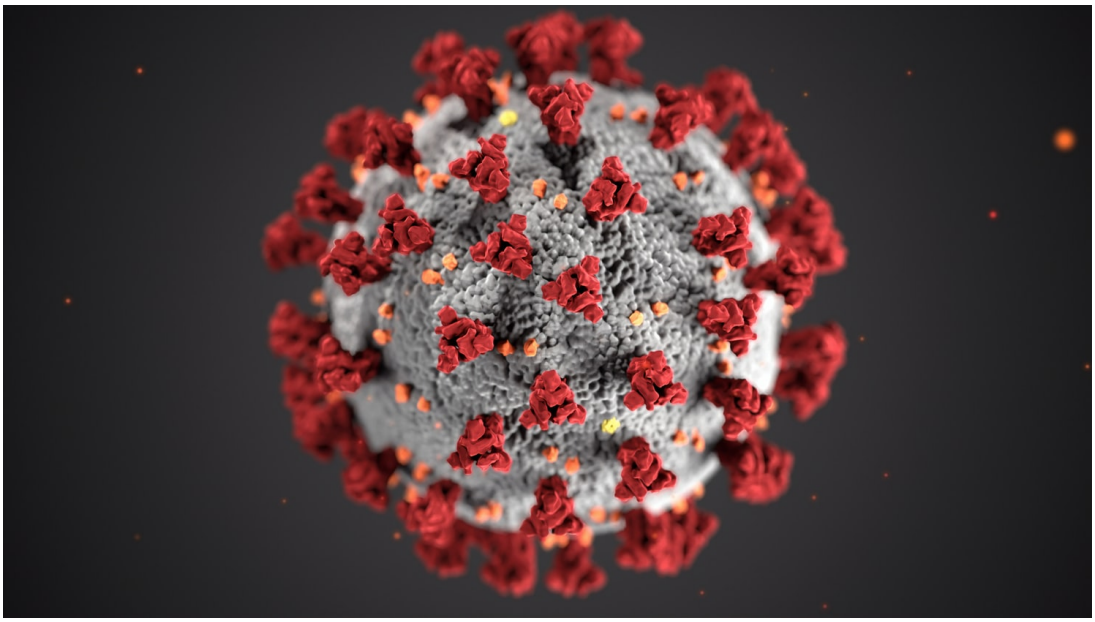
Data Analysis on COVID-19 in India



Contents

1. Introduction

1.1 Background

On 31 December 2019 China first reported it's Coronavirus in Wuhan, Hubei Province. A novel coronavirus was eventually identified, On 5 January 2020 WHO(World Health Organisation) announced an outbreak of coronavirus named it "COVID-19", On 13 January 2020 Thailand reported the first COVID-19 case outside of china late WHO declared a worldwide health emergency.

In India, The country reported its first three cases in Kerala, all of whom were students who had returned from Wuhan, China. The transmission escalated in March after several cases were reported all over the country, most of which were linked to people with a travel history to affected countries. On 10 March, the total cases reached 50. On 12 March, a 76-year-old man who had returned from Saudi Arabia became the first victim of the virus in the country. The total cases reached 100 on 15 March and 250 on 20 March and by April 30 the country reported 35365 total, 9065 cured, and 1152 death cases.

1.2 Problem

Even though many websites give a clear insight into a pandemic in India. I want to create a simple insight into the outbreak using python scripting language from pandas, matplotlib, folium libraries.

2. Exploring Datasets

The data set is available from the "Ministry of Health and Family Welfare" website(https://mohfw.gov.in), where we have to explore which data set is suitable for our requirement.

3.Downloading and Prepping

Data downloaded or scraped from "Ministry of Health and Family Welfare" website(https://mohfw.gov.in), and I have collected Indian states latitude and longitude from google maps as I have to plot the geospatial visualisation of corona cases throughout the country, after scrapping the data into DataFrame using pandas library, I've observed that data should be prepared for further stages as a part I have removed last three unwanted rows from the data frame and there only three columns which indicate cured, death,& total however for a better solution I have added another column "Active" for better understanding I have changed names the columns from {Name of State / UT': 'State/UT', 'Total Confirmed cases (Including 111 foreign Nationals)':'Total', 'Cured/Discharged/Migrated':'Cured'}.

4.The Total States Affected by COVID -19

As of 30 April 2020 26 states & 6 union territory have been affected with corona virus, in pandas using groupby attribute I have sorted in descending ordered from the data frame and plot as below using matplotlib.

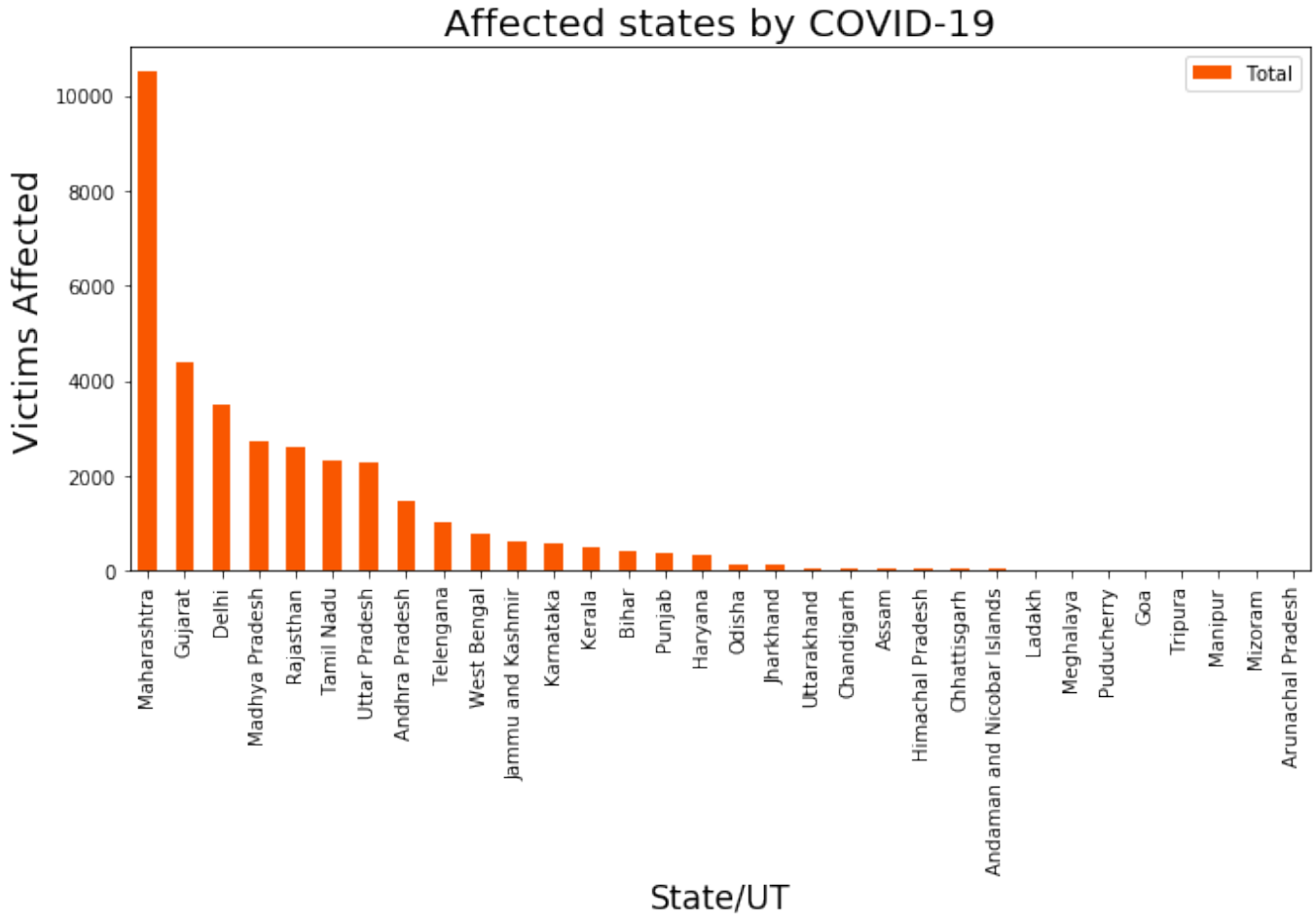


Figure 1. Bar graph of affected states by corona

4.1 Top 5 states affected by COVID-19

As we know from the data Maharashtra, Gujarat, Delhi, Madhya Pradesh, & Rajasthan are being top 5, as we know the incident(Tablighi Jamaat) happen in Delhi on before lockdown as there is more crowd in the incident as they have traveled throughout the country and government recognised it as coronavirus super-spreader event, with more than 4,000 confirmed cases and at least 27 deaths linked to the event reported across the country.

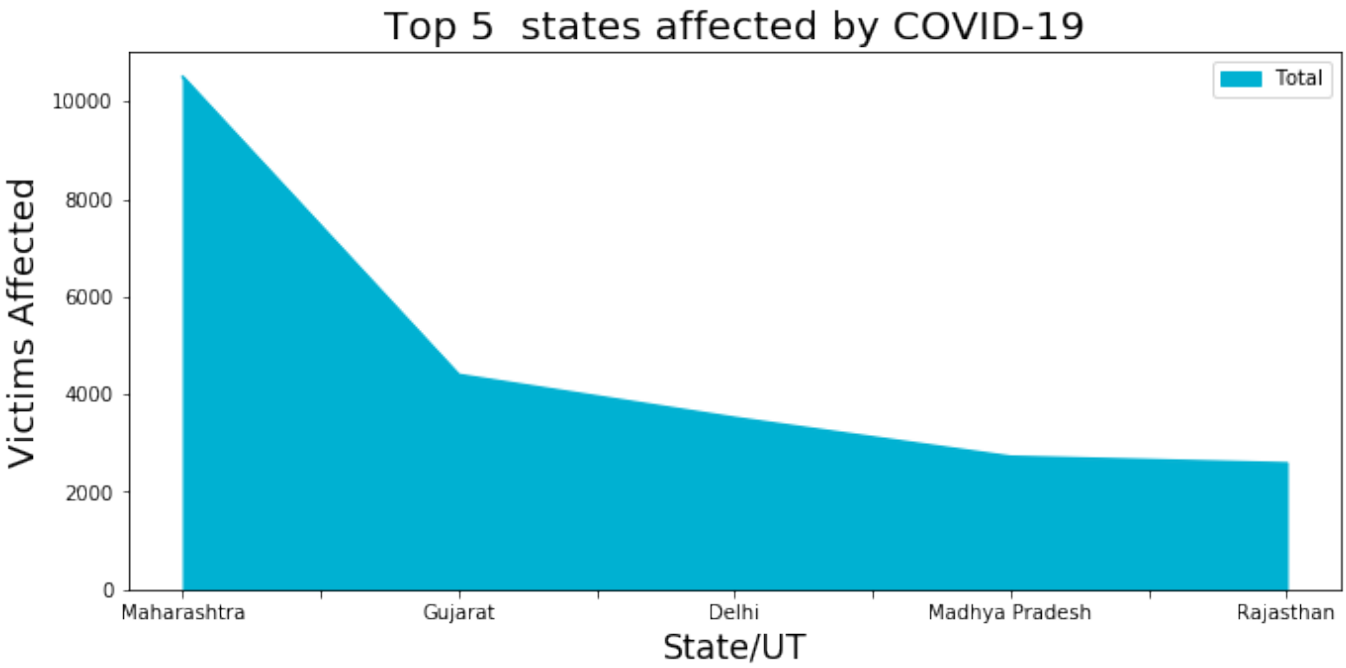


Figure 2. Area plot of Top 5 states affected by corona virus

4.2 Least affected states

Goa, Tripura, Manipur, Mizoram, Arunachal Pradesh are the least affected states by the coronavirus in the country. 7 corona cases in Goa, 2 corona cases in Tripura & Manipur, 1 case in Arunachal Pradesh.

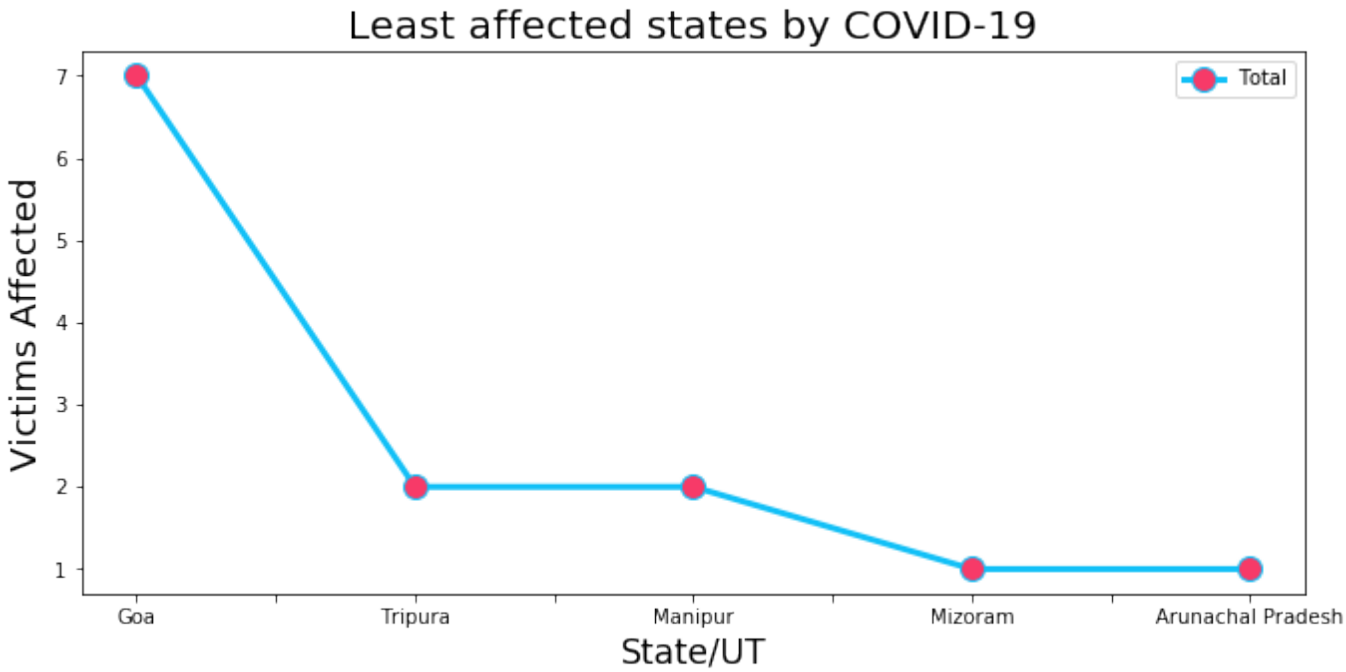


Figure 3. Least affected states

4.3 Active cases in India

As of 30 April 2020, there are 19 states with 25,148 active cases whereas remaining states 13 states have no active out of 32 states that have been reported corona cases in the country.

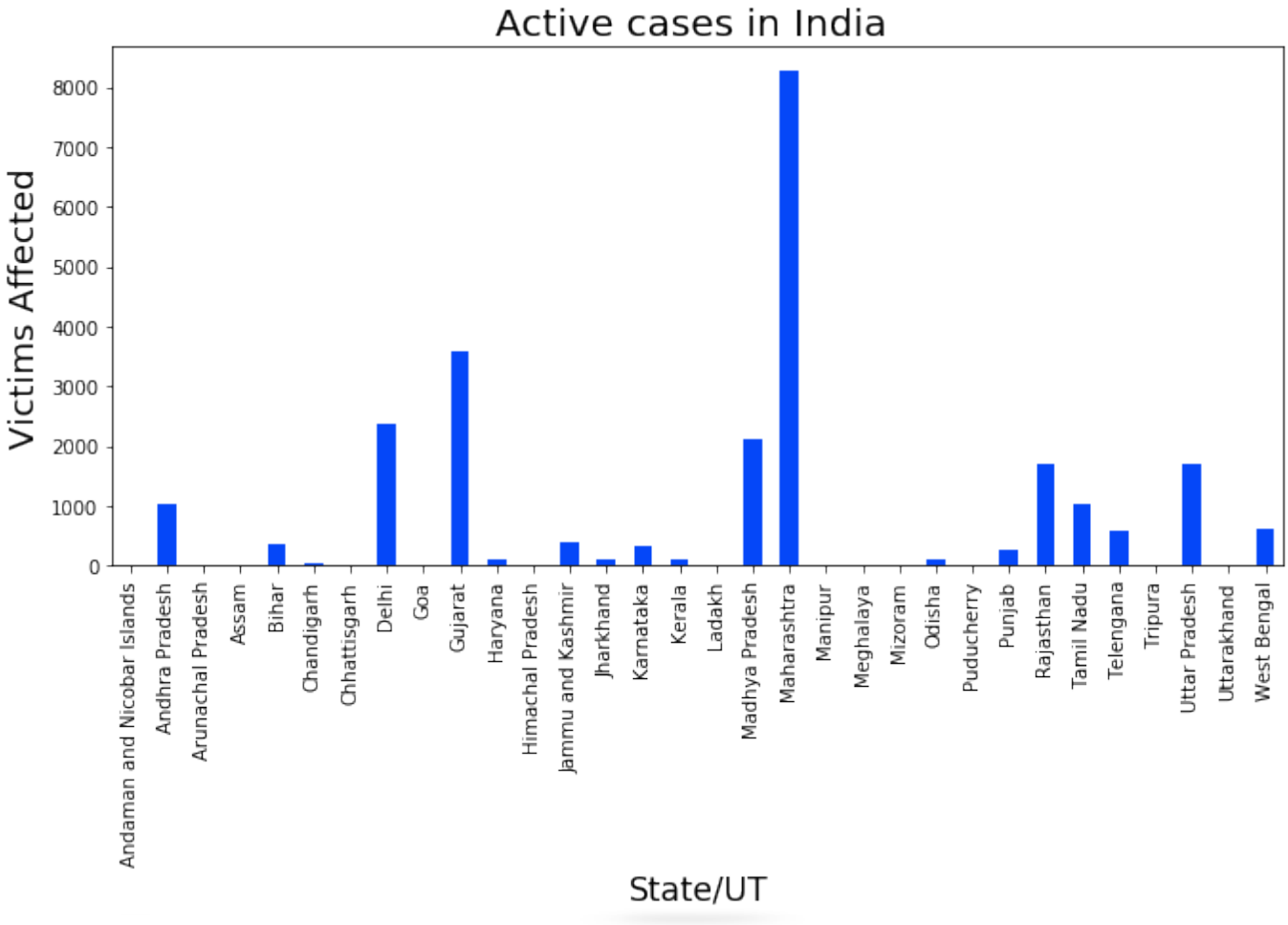


Figure 4. Active Cases in India

4.4 Top 5 Active States

As we know from the data set Maharashtra, Gujarat, Delhi, Madhya Pradesh, & Rajasthan are being top 5 as they have high corona cases and it will take almost 2 weeks to recover from coronavirus, as there is a recent spike in all these states which might lead to being a high

chance of spreading coronavirus in these states.

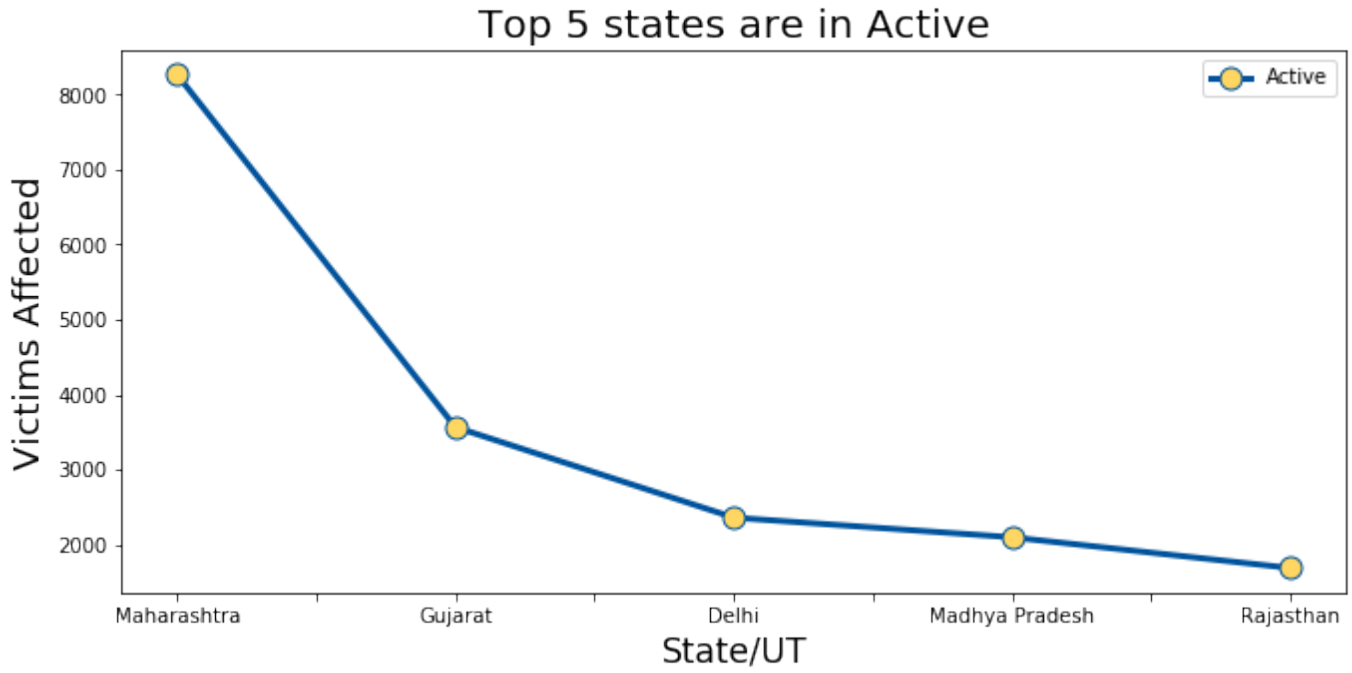


Figure 5. Top 5 Active States

4.5 Have low impact by Covid-19

As these states have a low chance of spreading coronavirus due to knee-high corona cases

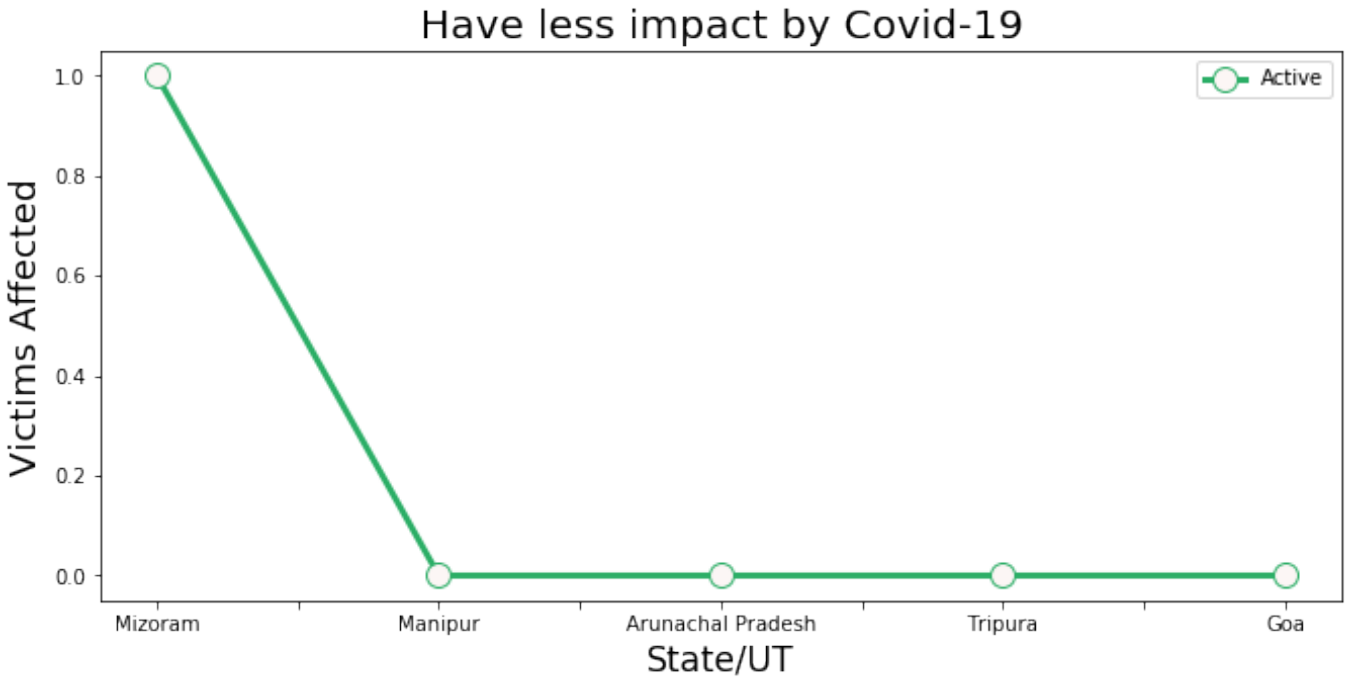


Figure 6. Low Impact states

4.6 Cured patients from Covid-19

As of 30 April 2020, there are 27 states with 9065 cured cases whereas remaining states 5 states have already rehabilitated out of 32 states that have been reported corona cases so far in the country.

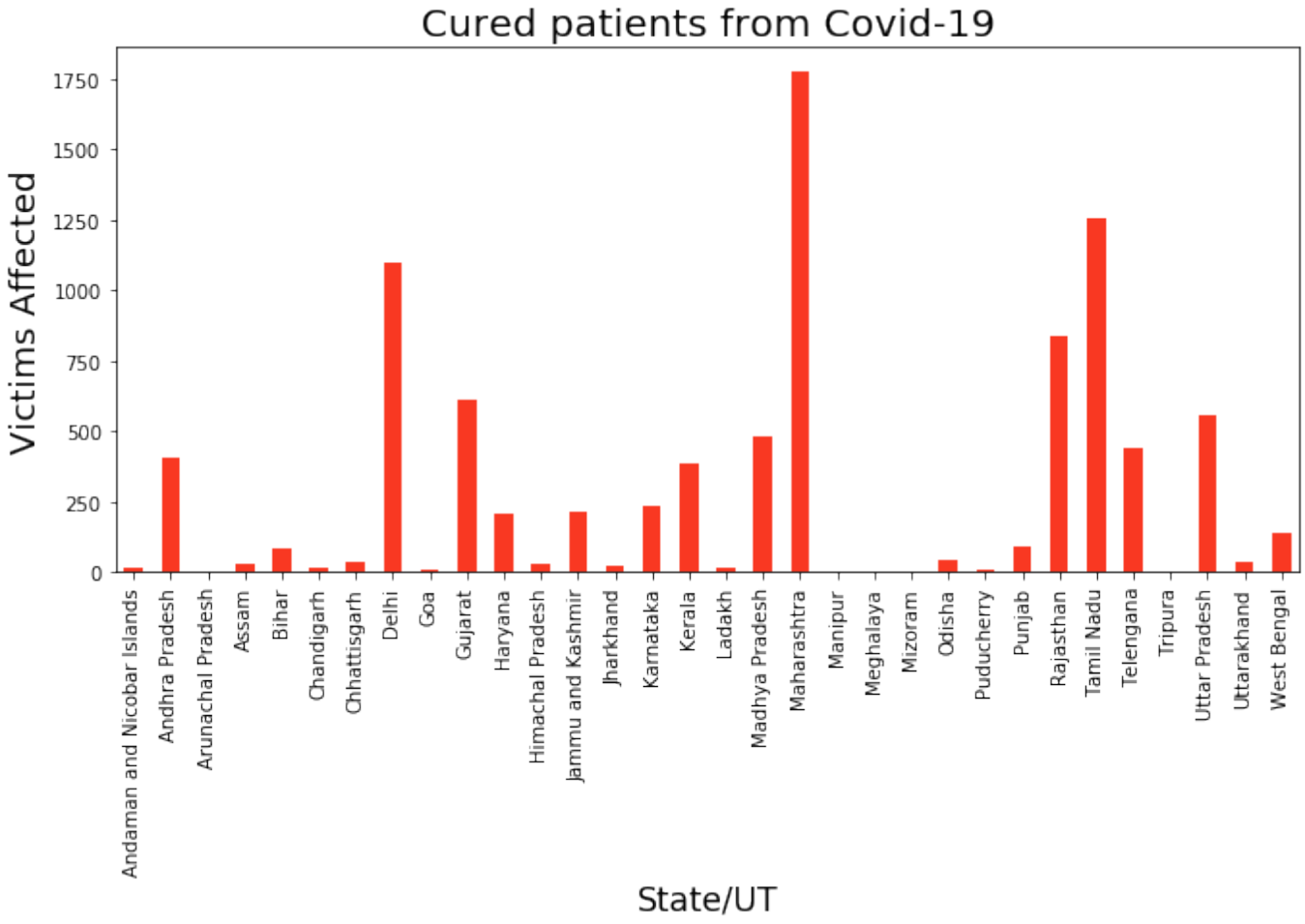


Figure 7. Rehabilitating States

4.7 Fast Rehabilitating States

In some states, the rate rehabilitating is very high because of the affected patients mostly under 60 years age group, as per WHO it is said that under 60 years age the patients who are attacked by coronavirus should be kept in isolation treatment for 2 weeks if they have no

other disease there is a high chance of getting rehabilitated quickly, as per the data set Maharashtra, Tamil Nadu, Delhi, Rajasthan, & Gujarat have a high recovery rate than compared to other states in the country.

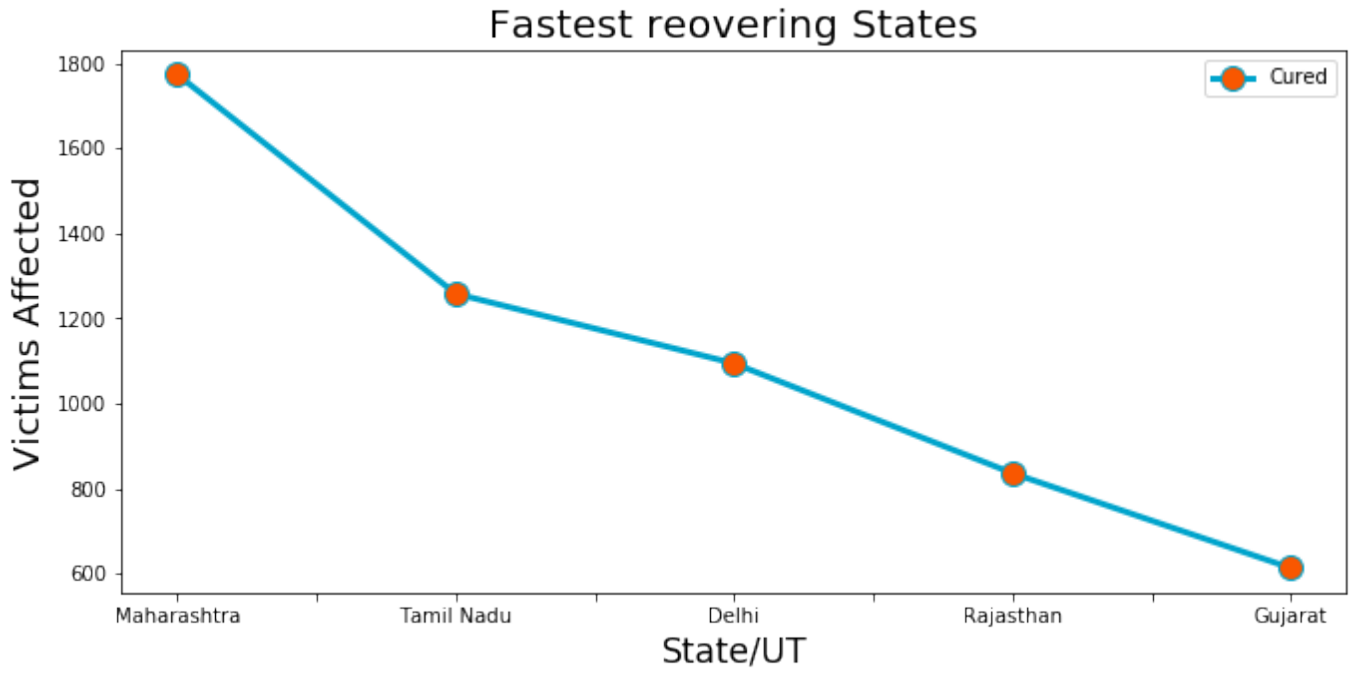


Figure 8. Fastest recovering States

4.8 Deaths by COVID-19

The country has 1152 deaths as of April 2020, as we know under 60 years of age group or who have low immunity or suffering from various diseases have a high chance to die by the coronavirus, as the country has a more young population than the old country is seeing low death rate by a coronavirus.

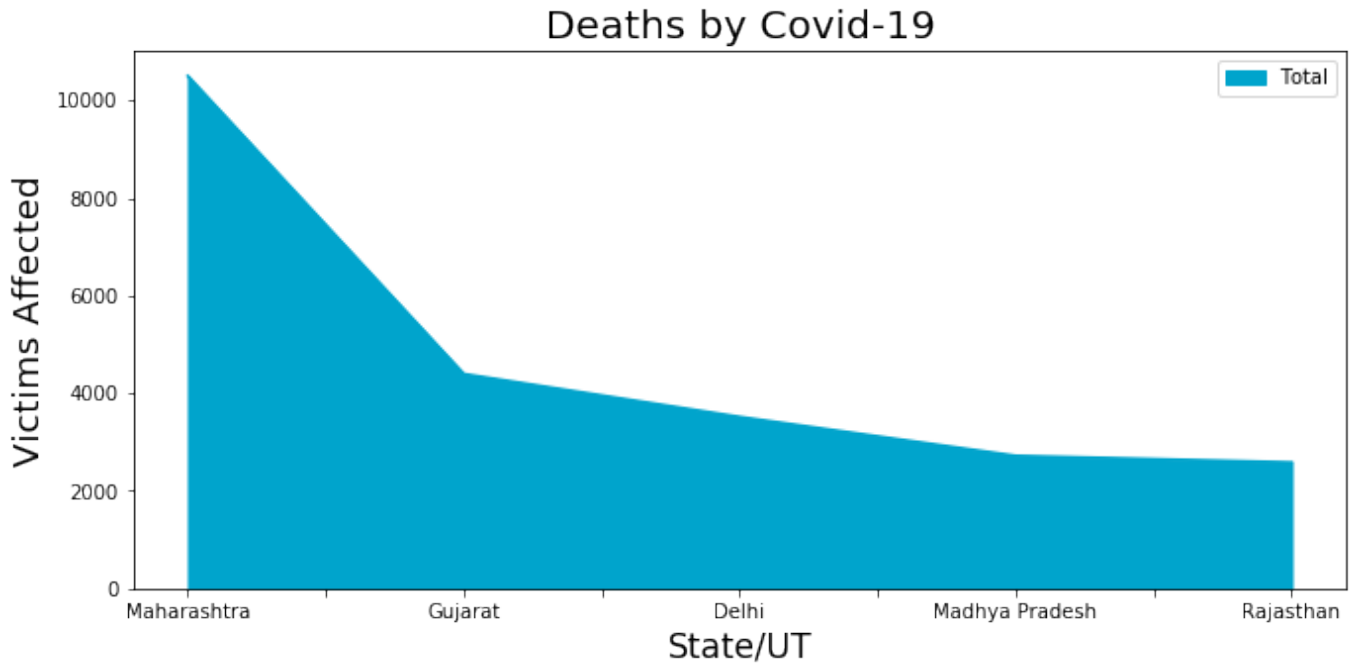


Figure 9. Area Plot of Deaths

4.9 High Death Impact states

Maharashtra share 45% death percentage followed by Gujarat, Madhya Pradesh, Delhi, & Rajasthan.

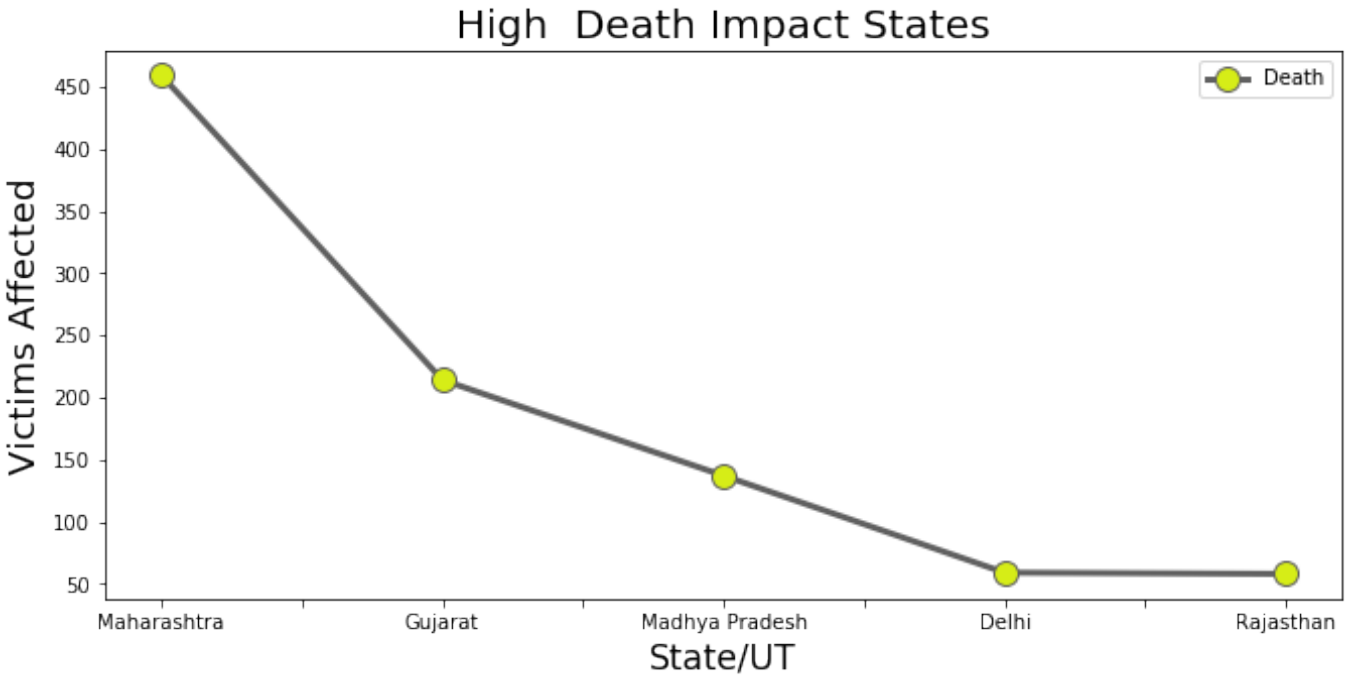


Figure 10. High Death Impact states

4.10 No Death Impact

Puducherry, Ladakh, Mizoram, Uttarakhand, & Andaman and Nicobar Islands have not reported any death by coronavirus so far.

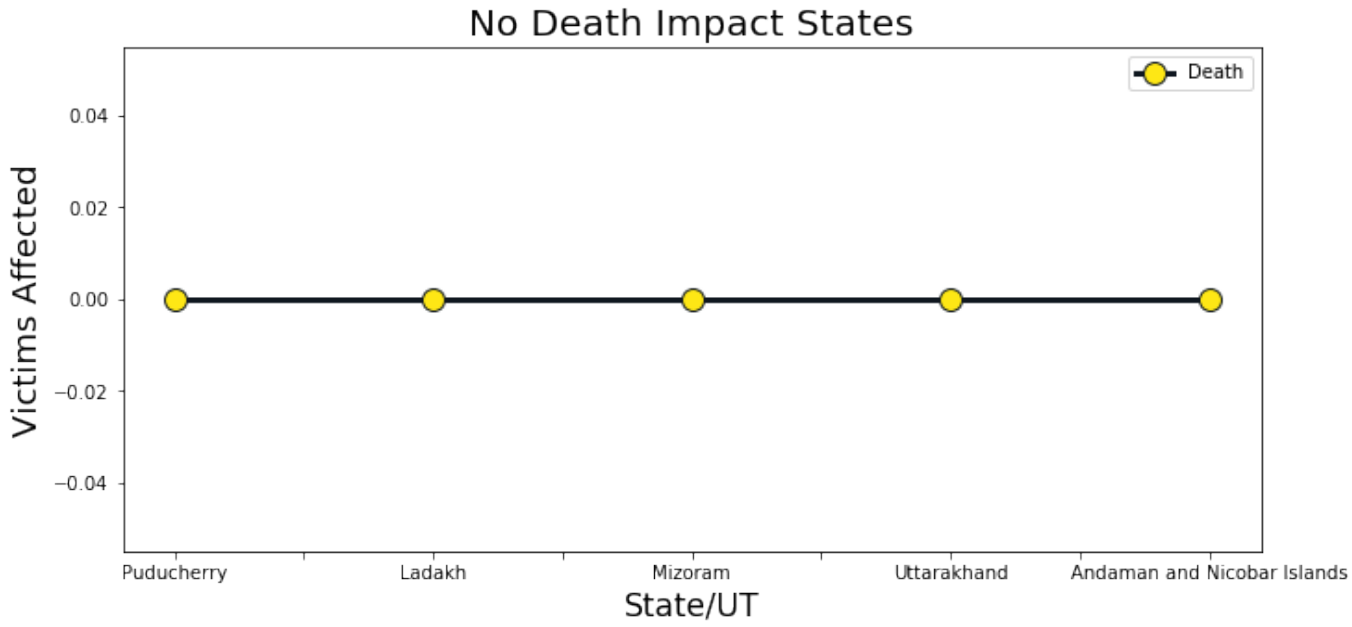


Figure 11. No Death Impact States

5. Regression Plot between Active and Total

Even though we know that, if total cases increase the active cases will also increase, to better understand the data from data set between active cases and total corona cases I have used the regression plot.

In the regression plot, we can see a straight line passes through the active cases and total cases in India.

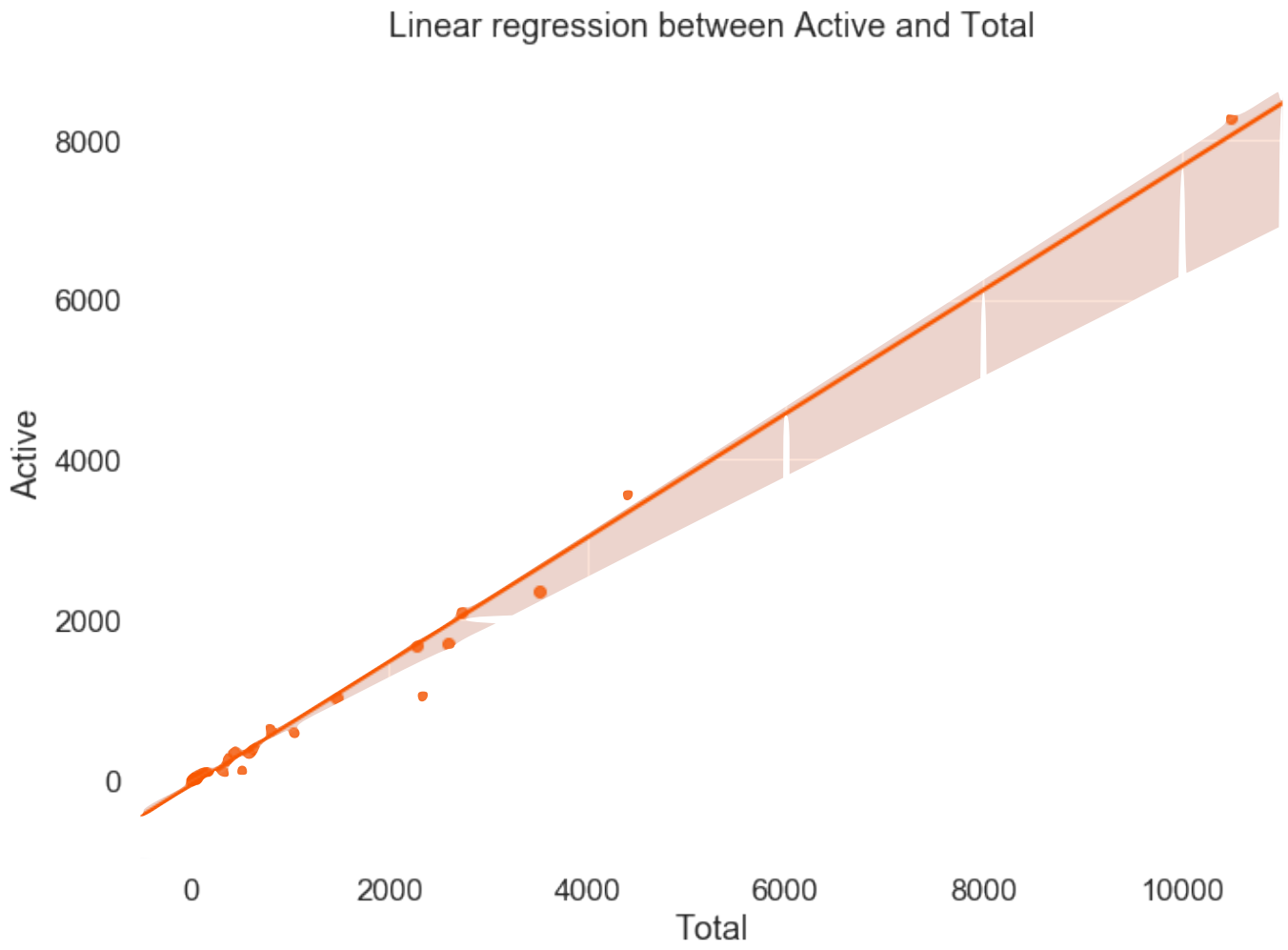


Figure 12. Regression Plot

6. Area Plot

	Active	Cured	Death	Total
State/UT				
Telengana	572	441	26	1039
Maharashtra	8266	1773	459	10498
Tamil Nadu	1038	1258	27	2323
Gujarat	3568	613	214	4395

As my native place is Telangana I want to compare Telangana data with other top 3 states in the country. From the data we can conclude that Telangana has fewer corona cases compared to the other three states as 30 April 2020.

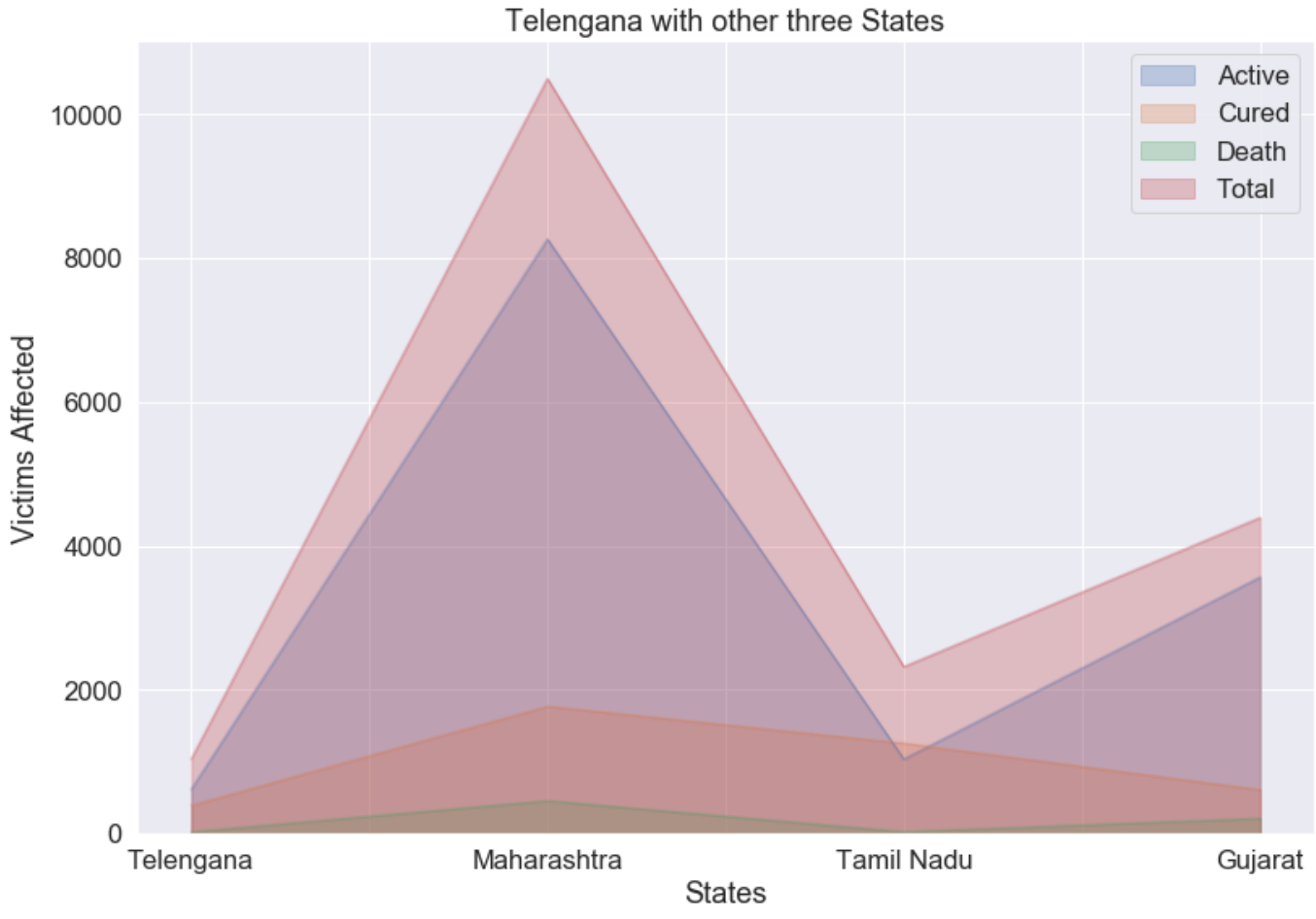


Figure 13. Area Plot

7. Geo-Spatial Visualisation

I have collected latitude and longitude numbers of Indian states scraped from google maps and created into pandas data frame later using merge attribute I have merged data frame from the MOHFW website into a new data frame, using folium library I have plotted on maps.

It will give a glance at every state.

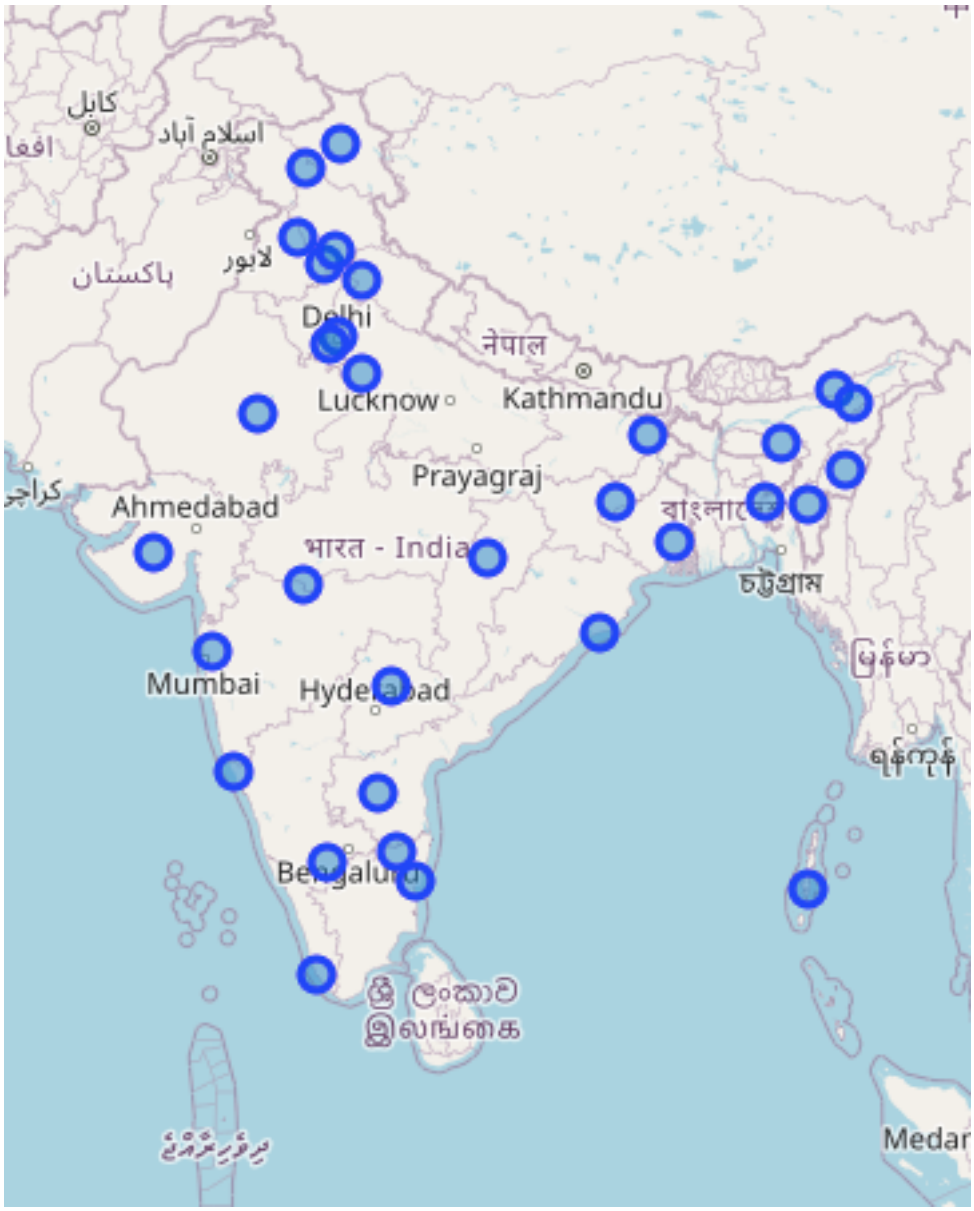


Figure 14. GeoSpatial Visualisation of COVID

8.Conclusion

In this study,I have analyzed coronavirus cases in India in various visualization like a bar graph, area plot, line plot, regression plot, & geospatial in every column from the data frame, even though the data is easily available in the public domain as recently I have learned Data Analysis I would like to apply it on real-time data. From this data set, I came to Know that for better analysis time series should be used.