Analysis of Covid-19 Virus Spread Hotspots in India

Siddharth Kekre

April 2020

Abstract

At the time of writing this report Covid-19 had already affected 14,84,811 people worldwide with 88,538 death toll and 5,734 confirmed cases in India. With the increase in Covid-19 Virus cases all around the world and in our Country, it is necessary to identify Hotspots of the spread so as to implement necessary constraints and regulations to avoid any more casualties. The project aims at identifying cities in each state of India where the number of cases is maximum and list those cities as Hotspots of the State.

Identification of such hotspots is done through Two Approaches, including Mathematical Approach and Visual Representation Approach so as to cross verify the obtained results and avoid any discrepancy in results.

The project on completion will be made Open Source along with all the Codes, Data Set and Documentation in hopes that other interested and educated minds would bring more usefulness from it than this project alone can.

Table of Contents

List of Figures

List of Tables

Abbreviations

Sr. No.	Abbreviation	Meaning
1	Covid-19	Novel Corona Virus
2	I/O	Input - Output
3	MH	Maharashtra
4	TN	Tamil Nadu
5	DL	Delhi
6	MP	Madhya Pradesh
7	CSV	Comma Separated Value File Format
8	GB	Giga Byte (Unit of Memory in Computer System)

Chapter-1

INTRODUCTION

1.1 Introduction

According to the World Health Organization (WHO), viral diseases continue to emerge and represent a serious issue to public health. On February 11, 2020, the WHO Director-General, announced that the disease caused by this new CoV was a "COVID-19," which is the acronym of "coronavirus disease 2019". This new virus seems to be very contagious and has quickly spread globally. In a meeting on January 30, 2020, per the International Health Regulations (IHR, 2005), the outbreak was declared by the WHO a Public Health Emergency of International Concern (PHEIC) as it had spread to 18 countries. World governments are at work to establish countermeasures to stem possible devastating effects.

The project aims at identifying cities in each state of India where the number of cases is maximum and list those cities as Hotspots of the State. Identification of such hotspots is done through Two Approaches, including Mathematical Approach and Visual Representation Approach so as to cross verify the obtained results and avoid any discrepancy in results.

The underlying distributed architecture design and implementation is discussed in this report, together with the strategy followed to get an accurate prediction matching with the real data retrieved by reliable sources.

1.2 Objective

The value of the Life is much higher than that of International Trades, Currencies, Markets etc. However, it is still controversial how communities and governments have failed to stop the spread of this highly dangerous virus. Therefore, so as to help our Indian government and local authorities this project will aim at providing names of cities where more focus is required so that all available resources could be deployed more efficiently.

The project describes the implementation of an approach that addresses this latter challenge. The proposal lies in two pillars: First, it is focused on mathematical calculations through various algorithms and known patterns. And the Second is focused on Visual representation of the data for better and easy understanding.

As a proof of concept, in the presented work Three majorly affected States of India namely Maharashtra, Tamil Nadu and Delhi have been chosen to showcase the effectiveness of the concept. Due to personal attachment to the state, Madhya Pradesh has also been taken in this cluster.

Chapter-2

SYSTEM REQUIREMENT ANALYSIS

2.1 Information Gathering

Two major sources of Data have been used in this project :

1. Official State wise data from Ministry of Health, Govt. of India

   The Ministry of Health, Government of India has been very kind to share data on Daily Basis that contains Number of Registered Patients, Number of Cures/Migrated/Discharged Patients and Number of Unfortunate Deaths in India due to Covid-19 Virus.

2. Reliable Open-Source data that has date-wise city's data

   This open source data from Covid-19 Dashboard, provides a live data updated every 4 hours and contains details about Covid-19 Spread bifurcated in 18 Column Heads. This live database obtains it's data from reliable sources such as India Today, Indian Express, Live Mint, NDTV, Hindustan Times, The Hindu etc.

Through Web-Scrapping methodology this data is obtained directly from the source, without creating any soft copy that can be tampered with. Hence, the data obtained is reliable, secure and fair to use.

2.2 System Feasibility

2.2.1 System Feasibility (Technical)

Using Data Analytics enables us to create flexible, data-driven models of Covid-19 Spread without fear of overfitting. The central insight of our approach is that Data Analytic methods, designed to produce insights that can be beneficial while deploying available relief resources making the process more efficient, reliable and robust. Data obtained in 4 hours interval helps archiving the goal. All of this data is collected and is processed in 3 stages. In the first stage, the data in obtained and Top 3 most affected States are identified. In the second stage, separate dedicated data clusters of these States are formed for detailed analytics. The third stage, Top 4 most affected Cities in each state are identified and among these most affected cities, the City with Maximum number of cases is declared as a Hotspot of that State.

2.2.2 System Feasibility (Behavioural)

There is little existing work on representing Data in various graphical formats; which are not operated by profit-maximizing agents, but none of them identify hotspots in the regions clearly. Using the new proposed methodology, the whole sole goal of this project is to identify the hotspots are report them as soon as data is obtained. Through such a mechanism the aim of optimizing relief resource distribution can be achieved.

2.3 Platform Specification

2.3.1. Minimum Hardware Specification

Processor	Intel Core i3 or equivalent
RAM	4 GB
Storage Capacity	1GB
Input Devices	Internet, Basic I/O like Keyboard
Output Devices	Image File, Basic I/O like Monitor

Table 2.1 : Minimum Hardware Specification

2.3.2 Software Platform Specification

Front End	Image File
Back End	CSV File, Python 3, Jupyter Notebook/Visual Studio
Database	Pandas Data Frame, Google Spreadsheet
Web Browser	Google Chrome (recommended)

Table 2.2 : Software Platform Specification

Chapter-3

MATHEMATICS REQUIRED

3.1 Introdiction to Terminologies

- Boolean Matrix

  A logical/ binary/ relational matrix is a matrix with entries from the Boolean domain B = {0, 1}. Such a matrix can be used to represent a binary relation between a pair of finite terms.

- Cartetian Plane

  A Cartesian plane is defined by two perpendicular number lines: the x-axis, which is horizontal, and the y-axis, which is vertical. Using these axes, we can describe any point in the plane using an ordered pair of numbers.

- First Quartile

  The first quartile (Q1) is defined as the middle number between the smallest number and the median of the data set.

- Third Quartile

  The third quartile (Q3) is the middle value between the median and the highest value of the data set.

- Inter Quartile Range

  Difference Between Third Quartile and First Quartile.

- Outliers

  The Data Points that exist outside Inter Quartile Range are termed Ouliers. These data points usually represents out of the general pattern of data; these could be Extreme Values in the dataset.

- Plot / Graph

  A way of visually representing Inter-Dependency of Numeric Data with respect to each other through shapes, lines, colors etc.

3.2 Graphs for Visualization

3.2.1. Box Plot

In statistics, the box plot is a standardized way of displaying the distribution of data based on the five number summary: minimum, first quartile, median, third quartile, and maximum.

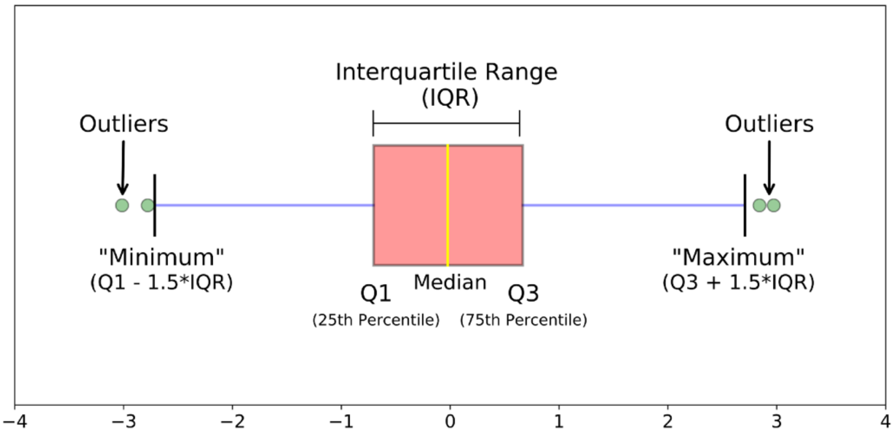

Fig 3.1 :Demo Box Plot

3.2.2. Bar Plot

A bar plot or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

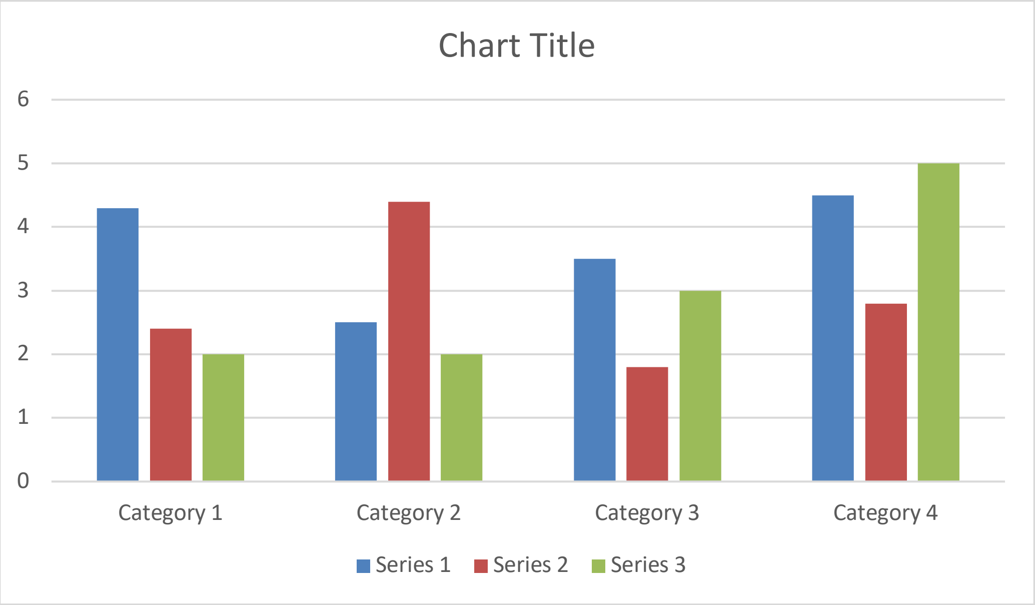

Fig 3.2 : Demo Bar Plot

Chapter-4

LIBRARIES AND PACKAGES

4.1 Libraries and packages Required

4.1.1 Pandas

In particular, it offers data structures and operations for manipulating numerical tables and time series. The most used application of 'Pandas' Library is the functionality of Pandas Data Frames which are life 2-D Matrix or a simple Spreadsheet for easy understanding.

4.1.2 Requests

'Requests' is a Python HTTP library, released under the Apache License 2.0. The goal of the project is to make HTTP requests simpler and more human-friendly.

4.1.3 Beautiful Soup

Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

4.1.4 Matplotlib

'Matplotlib' is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications

4.1.5 Seaborn

'Seaborn' is a library for making statistical graphics in Python. It is built on top of matplotlib and closely integrated with pandas data structures.

Chapter 5

ANALYSIS OF DATA

5.1 Getting Data in Required Format

5.1.1 Data Sourcing

Among other things, it was crucial to get the data in required format in order to understand and operate upon it so as to visualize the proposed system and make it ready for the real world. Also, in order to make it more efficient and trustworthy for future users we had to test it on certain parameters.

Data was Obtained by using 2 techniques :

1. Web Scrapping HTML Table from Ministry of Health, Govt. of India website through Beautiful Soup library and storing that data in a Pandas Data Frame.
2. Web Scrapping a Google Spreadsheet CSV file through Pandas 'read_csv' function and storing that data in Pandas Data Frame.

5.1.2 Data Pre-Processing

All the data rows with missing data are to be removed or else they will cause irregularity while performing mathematical operations. This is done by using 'dropna()' function of the Pandas Library.

Now, only data from required states has to be stored and rest all has to be discarded so as to save memory and reduce mathematical complexity.

5.2 Analyzing the Data

5.2.1 Analysing State wise data

Through a conditional check a Boolean Matrix of Entire Data Set is created and passed into the Data Frame to obtain data only of a particular state. Since the dataset is unordered, grouping by the name of city is done for easy analysis and visualization.

5.2.2 Identifying Hotspot in Each State

Boxplots of cities of a particular state is plotted to identify extreme values and Bar Plots of Top 4 most affected Cities is plotted to identify the Hotspots of the particular state.

Since, data is live and changes every 4 hours, the Plots also keep changing to identify hotspots.

Chapter 6

DATA VISUALIZATION

6.1 Analyzing Data Visually

6.1.1 State Wise Analysis

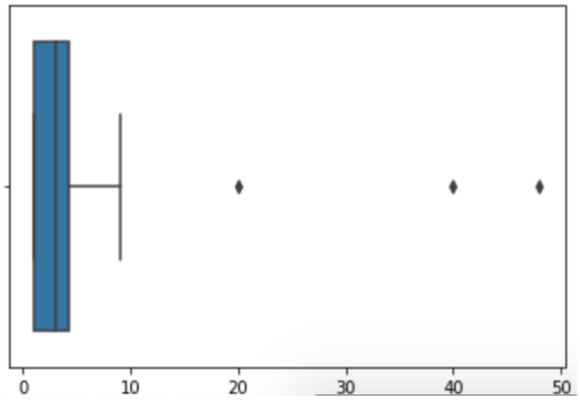

Fig 6.1 : Box Plot of Maharashtra

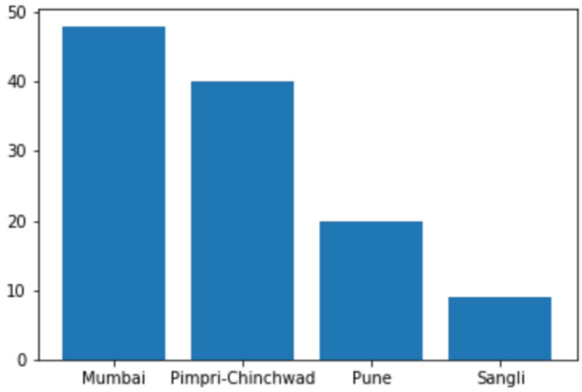

Fig 6.2 : Bar Plot of Maharashtra

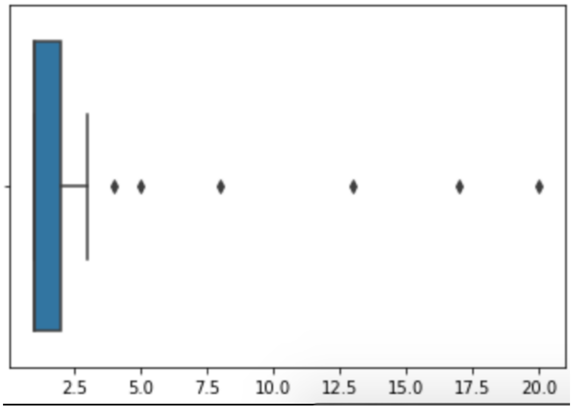

Fig 6.3 : Box Plot of Tamil Nadu

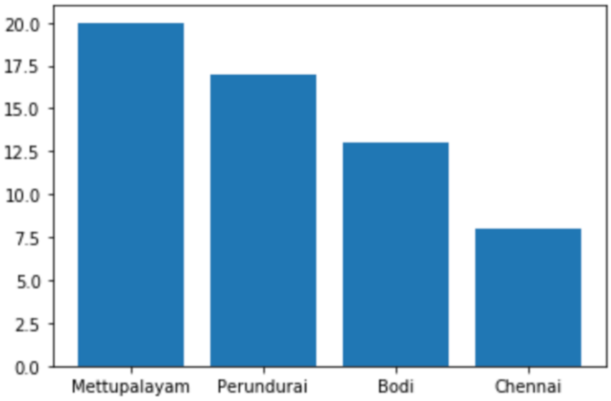

Fig 6.4 : Bar Plot of Tamil Nadu

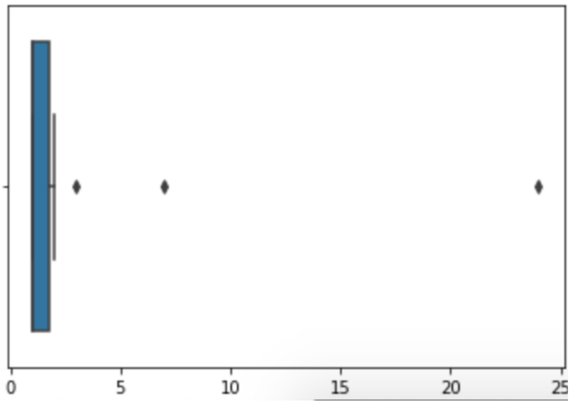

Fig 6.5 : Box Plot of Delhi

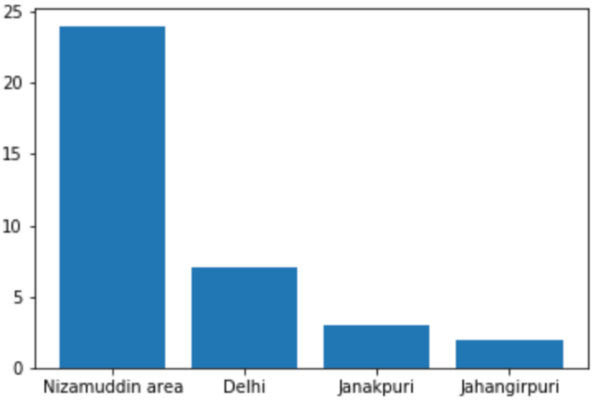

Fig 6.6 : Bar Plot of Delhi

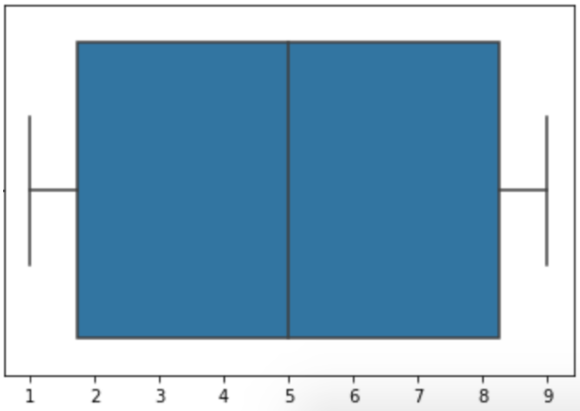

Fig 6.7 : Box Plot of Madhya Pradesh

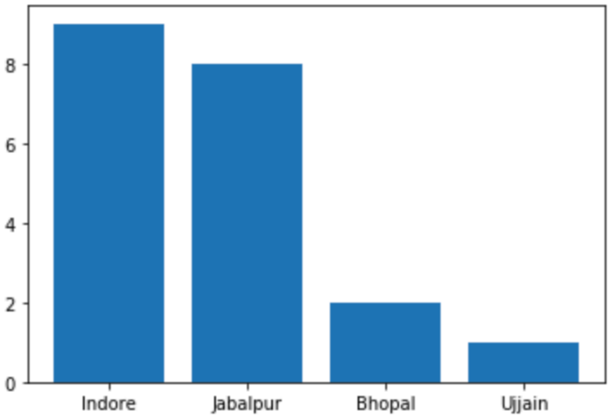

Fig 6.8 : Bar Plot of Pradesh

6.1.2 Hotspot Identification in each state Visually

From the Box Plots and Bar Plots shown above, Hotspots of Some of the most affected States can be stated as follows:

State	Hotspot
Maharashtra	Mumbai
Tamil Nadu	Mettupalayam
Delhi	Nizamuddin Area
Madhya Pradesh	Indore

Table 6.1 : State Wise Hotspot

Chapter 7

LIMITATIONS

7. Limitations

- The data could be incomplete, missing values, even the lack of a section or a substantial part of the data, could limit its usability.
- Data collected from different sources can vary in quality and format.
- The pool of training and test data might not be large enough for apt results.
- Data Integration from multiple sources could generate ambiguity.

Chapter 8

BIBLIOGRAPHY AND REFERECES

8. Bibliography and References

- State wise Data from Ministry of Health, Government of India
  - https://covid19dashboard.social
  - https://www.mohfw.gov.in
- Open Source Live Data Set for Covid-19 City Wise
  - https://www.covid19india.org
- ‘Pandas’ Library
  - https://pandas.pydata.org/
- ‘Seaborn Library
  - https://seaborn.pydata.org/
- ‘Matplotlib’ Library
  - https://matplotlib.org/