**A**

**Project Report**

**on**

**Big Data Analytics**

**Breaking Down Covid-19**

**Submitted by-**

**Yarala Hruthik Reddy**

**1700289C203**

**CSE-1**

**Nikita Agarwala**

**1700247C203**

**CSE-1**

**Under Guidance of-**

**Dr. Yogesh Gupta**

Assistant Professor

******

**Department of Computer Science and Engineering**

**SCHOOL OF ENGINEERING AND TECHNOLOGY**

**BML MUNJAL UNIVERSITY GURGAON-122413, INDIA**

***DECEMBER 2020***

**Acknowledgement**

We are extremely grateful for a course that allows us to build projects and let us attain the inherent experience. We would like to thank our mentor Dr. Yogesh Gupta for enabling us to choose a project that solves real-world issues.

Thanking You

Nikita Agarwala

Yarala Hruthik Reddy

**INDEX**

[Abstract 4](#_Toc58163008)

[Motivation 5](#_Toc58163009)

[1. INTRODUCTION 6](#_Toc58163010)

[2. PROBLEM STATEMENT 7](#_Toc58163011)

[3. LITERATURE REVIEW 7](#_Toc58163012)

[3.1 Existing State of Art 7](#_Toc58163013)

[3.2 Details of the Existing State of Art 8](#_Toc58163014)

[4. METHODOLOGY 8](#_Toc58163015)

[4.1. Datasets 8](#_Toc58163016)

[4.2. Global Analysis 8](#_Toc58163017)

[4.2.1. Exploratory Data Analysis 8](#_Toc58163018)

[4.2.2. Efficiency of Lockdowns 12](#_Toc58163019)

[4.2.3. Diabetes Effect on Death Rate 13](#_Toc58163020)

[4.2.4. Obesity Effect on Death Rate 14](#_Toc58163021)

[4.2.5. Testing Effect on Death Rate 17](#_Toc58163022)

[4.3. Forecast 18](#_Toc58163023)

[4.4. Indian Analysis 19](#_Toc58163024)

[4.5. Technologies Used 21](#_Toc58163025)

[5. RESULTS and DISCUSSION 22](#_Toc58163026)

[6. CONCLUSIONS and FUTURE WORK 22](#_Toc58163027)

[7. REFERENCES 22](#_Toc58163028)

# Abstract

The global outbreak of Covid-19 has affected and is continuing to affect all of us since December 2019. Governments across the world have taken various measures to contain the global outbreak of COVID-19 but it seems somewhere we are going wrong as currently the US, and Europe are again facing a new load of viral infections. Therefore, this study intends to understand how Covid-19 has spread over the past few months and who are the people who are at greater risk. This study indicates that the pandemic situation might get even worse if we do not maintain social distancing from others. This statement can be affirmed from the fact that cities with high population density indeed have a high number of people affected with COVID-19. In addition, we were able to observe aged people and people with previous health ailments such as diabetes and obesity are the most vulnerable to this disease. It may seem to some that the pandemic is coming to an end, but the forecast does not say the same. Hence, this study helps us understand various factors that may make one more vulnerable to COVID-19 because it will take some more time before the world returns to normalcy.

# Motivation

Our motivation was simple. We wanted to understand and analyze some of the factors that contributed to the spread of Covid-19 across the globe. Also, we wanted to see for ourselves how data scientists and data analysts around the world handle this type of raw data which grows by volumes daily.

# INTRODUCTION

In the family of Coronaviruses that have caused epidemics such as SARS (2002-2004), MERS (2012-Present) and COVID-19 (2019-Present), only COVID-19 has been declared a pandemic by the WHO. The number of confirmed cases crossing 60 million and reported deaths crossing 1.4 million as of 29th November 2020 as reported in the weekly epidemiological update of WHO [1]. Though not deadly as its counterparts SARS and MERS but is more contagious than both. Covid-19 and SARS have similar base R0,but the symptoms of SARS were more visual than Covid-19 resulting in less spread i.e., there was decent contrast between the healthy and the infected. Covid-19 was also noted to spread through asymptomatic patients [2].

Unlike MERS (Except for the ROK outbreak) and SARS, Covid-19 outbreak epicenters kept shifting. Initial epicenter was China, then shift was towards Europe (Italy and neighboring countries) and currently the US is the new epicenter of Covid-19 [3].

A picture containing text, screenshot, indoor

Description automatically generated

Figure 1. Almost every country has reported cases.

Containment measures included social distancing, contact tracing, quarantine of suspected cases and isolation of all contracted cases. The other coronaviruses outbreaks did not call for social distancing. But SARS-CoV-2 spreads faster, and it affects the old-aged people disproportionately. Social distancing helps keep further infected people from spreading the virus while keeping National Healthcare Systems from overwhelming.

India is no stranger to this pandemic and in fact it is one of the countries that has been affected the most by this pandemic. As of 6th December 2020, more than 9.6 million people have been tested positive for Covid-19 in India and around 1.5 lakh people have succumbed to this disease. India is home to around 1.3 billion people and this staggering number has just added to the woes of the authorities. Covid-19 has not only resulted in loss to human lives but we can also observe adverse socio-economic effects therefore it is essential for us to act fast. Although strategies such as lockdown, contact tracing and social distancing have proved to be highly effective in controlling the spread of Covid-19 but now it is essential to unleash the power of data. Various agencies have been able to collect humongous amount of data regarding Covid-19 and now we can use of different computation models, statistical tools, and quantitative analyses to study and control the further spread of this highly contagious disease.

# PROBLEM STATEMENT

Any patterns identified in the dynamics of Covid-19 can be crucial. These patterns could help devise strategies in better handling the pandemic. We are analyzing two datasets: a global dataset and an Indian dataset of cases. This way we can learn from other countries and analyze how Covid-19 progressed in those places. This analysis will help us in gathering insights that will be crucial in understanding the spread of Covid-19 and finding people and communities that are more vulnerable to it. A forecast of the cases is also done for select countries using the Holt-Winter’s Model [4].

# LITERATURE REVIEW

Virus being an infectious agent multiplies inside of a living organism and hence pose a great threat to mankind. Since they spread through interaction of humans, represent an expensive consequence. A loss of $60 billion per year is estimated in the case of a pandemic disease event. This loss is generally due to disruption of daily activities, meaning most of the economy comes to a stall in case of a huge outbreak. Though the transmission dynamics usually define the type of containment measures, respiratory pathogens cause the most damage. SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2), the virus that causes the Covid-19 disease, is in fact a respiratory pathogen, a virus typically attacking the lungs.

## Existing State of Art

There are extensive analysis and analytics of this virus being done, every analysis is trying to find patterns which we also tried finding. The dataset is the problem for any of these studies, as data is either estimated, incorrect, or simply is not authentic enough to provide consistent results across all the models.

## Details of the Existing State of Art

|  |  |  |  |
| --- | --- | --- | --- |
| S. No. | Existing state of art | Drawbacks in existing state of art | Overcome |
| 1 | Many analytical models | Data is ambiguous in many cases | Standard datasets needed |

# METHODOLOGY

## Datasets

We had two datasets relating to Covid-19. A global dataset which had daily data of cases from every country that reported Covid-19 and an Indian dataset which had district wise data. The global dataset was obtained from Our World in Data (OWID) git repository [5] and the Indian dataset was obtained from a volunteer driven and crowd sourced database [6].

The OWID dataset had 52873 rows with 41 columns. It contains data from 212 countries (including provinces of China and other miscellaneous states).

In the Covid-19 Indian Dataset, data is being constantly updated on different Google Sheets and volunteers are responsible for collecting this data from trusted sources and updating the sheet. This is a live dataset, and it is updated after every few minutes.

We retrieved district-wise data regarding India  using API provided by www.covid19india.org. On any given day, the dataset comprises around 801 entries for some 16 parameters. The 801 entries correspond to different districts in India. Latitude and Longitude information was missing from the above-mentioned dataset, so we thought of appending these columns to gauge the overall situation in our country.

## Global Analysis

## Exploratory Data Analysis

We did an exploratory data analysis on our dataset to narrow down the selection to few countries that were at the risk of Covid-19. The countries we chose based on this analysis are:

Table 1. Countries chosen for analysis.

|  |  |
| --- | --- |
| No | Country |
| 1 | United States |
| 2 | Sweden |
| 3 | Bolivia |
| 4 | Serbia |
| 5 | Japan |
| 6 | Canada |
| 7 | India |
| 8 | Ireland |
| 9 | Denmark |
| 10 | Mexico |
| 11 | Chile |
| 12 | Belgium |
| 13 | United Kingdom |
| 14 | France |
| 15 | Italy |

Figure 2 shows the scatter plot between total deaths per million and median age as of 27-10-2020. This figure gives insights into the effect of Covid-19 on countries with respect to median age.

Chart, scatter chart

Description automatically generated

Figure 2.

Countries with median age greater than 35 and high deaths:

Table 2. Countries with median age greater than 35 and high deaths.

|  |  |
| --- | --- |
| No | Country |
| 1 | Chile |
| 2 | United States |
| 3 | Belgium |
| 4 | United Kingdom |
| 5 | Sweden |
| 6 | France |
| 7 | Italy |

But there are a lot of other countries with similar and even greater median age but less deaths. The reason being the population greater than 65 was higher for the countries in table 2. This was noticed in figure 3 when a similar plot was made between total deaths per million and population aged more than 65 years.

Chart, scatter chart

Description automatically generated

Figure 3.

Countries with greater population of 65 aged people high deaths:

Table 3. Countries with greater population of 65 aged people high deaths.

|  |  |
| --- | --- |
| No | Country |
| 1 | Belgium |
| 2 | United States |
| 3 | United Kingdom |
| 4 | Sweden |
| 5 | France |
| 6 | Italy |

These countries in the table 3 were also the worst hit by Covid-19 when the death rates are compared to other countries. This means Covid-19 targets an inherently older population.

Yet Japan with the highest median age and the oldest population had significantly less deaths. This shows that Japan had an edge over the virus though being highly susceptible to it.

Figure 4 and figure 5 show the general situation of the USA and Japan, respectively. It is clear that the USA did not attempt enough to contain the virus while Japan came close to containing the virus before a new wave of infections started once again.

Chart, line chart, histogram

Description automatically generated

Figure 4. USA. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

A picture containing graphical user interface

Description automatically generated

Figure 5. Japan. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

## Efficiency of Lockdowns

We plotted new cases per million of Sweden and India who implemented the minimum and the maximum restrictions respectively to measure the efficiency of lockdowns had on the spread of the virus. Figure 6 and figure 7 show the respective plots.

Chart, line chart, histogram

Description automatically generated

Figure 6. Sweden. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

A picture containing chart

Description automatically generated

Figure 7. India. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

Early restrictions in India helped in containing the virus to a large extent even though being the second most populous country. Unlike Sweden, which saw frequent rises and falls in its new cases. Though no mandatory restrictions were imposed, the public were well aware of the virus and tool voluntary measures to curb the spread.

## Diabetes Effect on Death Rate

The effect of diabetes on Covid-19 was not much pronounced implicating it had negligible effect on the death rate. The comparison between the USA and Ireland gave this insight. The USA has the highest prevalence rate for diabetes while Ireland has the least. Figure 8 and figure 9 show the respective plots.

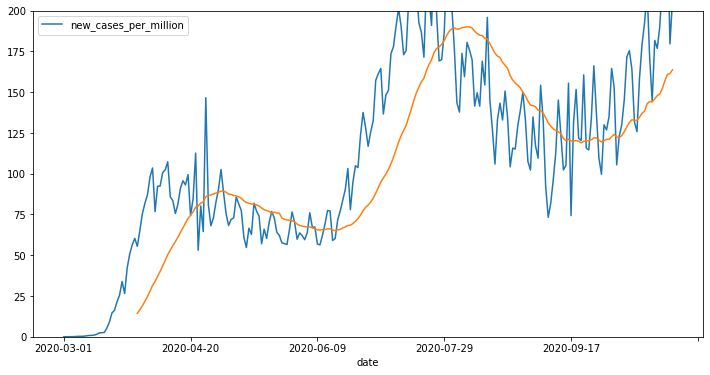


Figure 8. USA. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

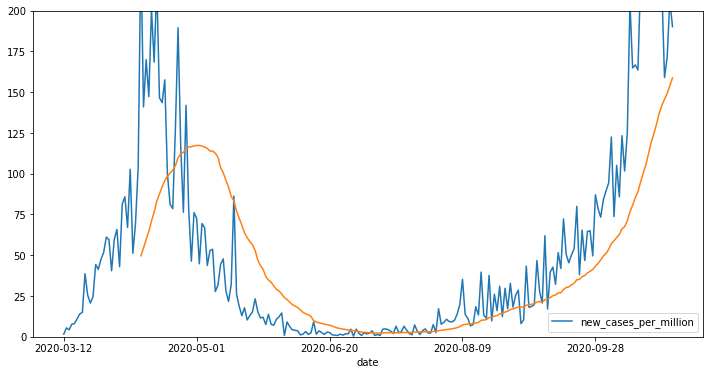


Figure 9. Ireland. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

## Obesity Effect on Death Rate

Unlike diabetes, obesity had a more pronounced effect on the death rate due to the virus. The comparisons between the USA, Canada, India, and Japan gave this insight. Figures 10, 11, 12 and 13 show the respective plots.

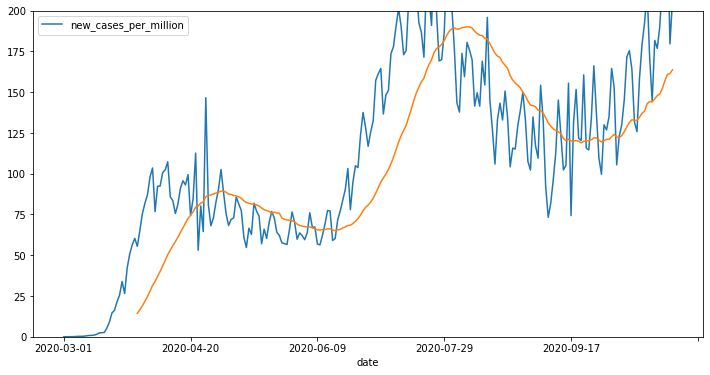


Figure 10. USA. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

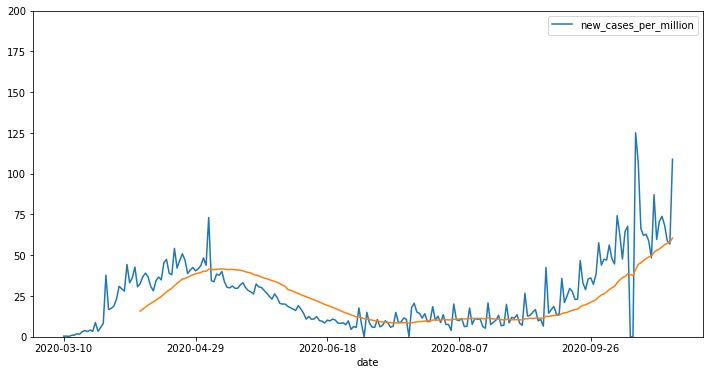


Figure 11. Canada. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

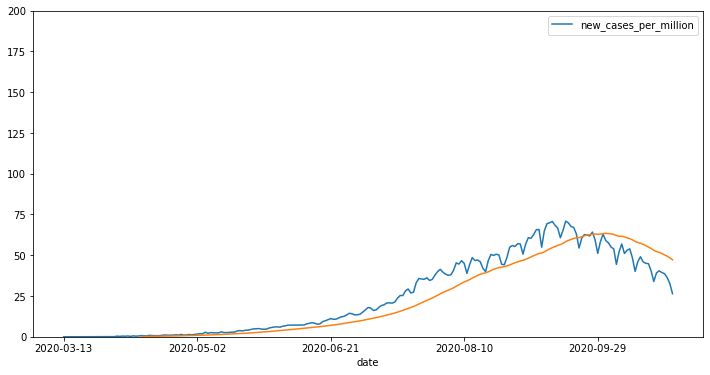


Figure 12. India. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

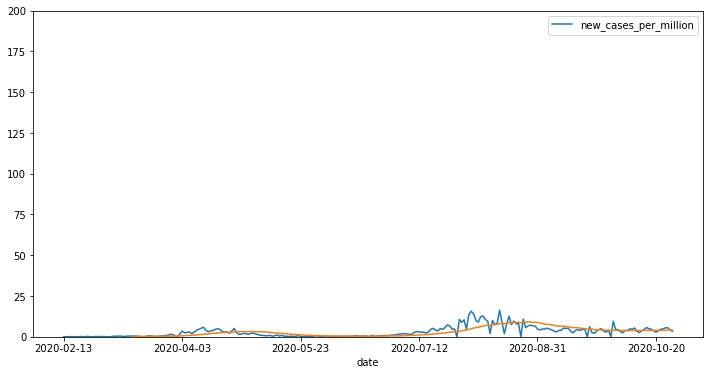


Figure 13. Japan. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

Another reason for Japan having less death rate could be due to its healthy population relative to other countries. Also, the reason for India even with a greater population having less death for a contagious virus.

## Testing Effect on Death Rate

The effect of testing was pronounced for certain countries such as Denmark and Mexico who have had the highest and the least testing capacity respectively. Figure 14 and figure 15 show the respective plots.

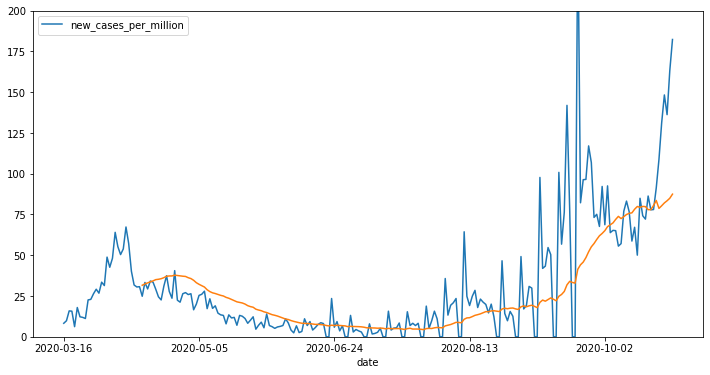


Figure 14. Denmark. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

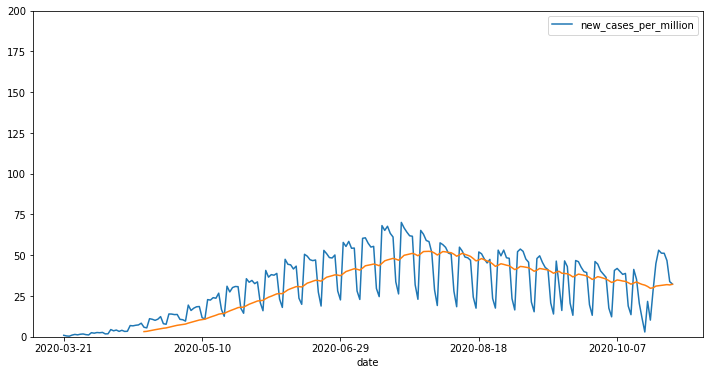


Figure 15. Mexico. The orange line indicates average i.e., if blue line is above the orange line than the situation is bad.

## Forecast

We used the Holt-Winter’s Model which models three aspects of time series being the average, trend, and seasonality. In doing so it tries to predict the future provided the series is seasonal and repetitive over time.

The model was run on three countries: United States, Brazil, and Japan. There were different training and testing sets for each of these as the number of entries were different. The forecast was done 100 days into the future with additive model and a seasonal period of 14.

In order to test the forecast, we used data collected up to the month of October though we had data available till November. Our forecast showed that November month matched closely to the actual scenario of the month.

The accuracy of the model ranged from 90% to 95% depending on the country and the amount of data available. The accuracies for USA, Brazil, and Japan were 94%, 92%, and 93% respectively.

Chart, line chart

Description automatically generated

Figure 16. USA. The orange line indicates the forecast.

A picture containing chart

Description automatically generated

Figure 17. Brazil. The orange line indicates the forecast.

Chart

Description automatically generated

Figure 18. Japan. The orange line indicates the forecast.

## Indian Analysis

After retrieving all the data associated with India, we wanted to analyze the same. So, we made use of Folium to understand the spread of coronavirus in India. Folium is a Python library which is used for visualizing geospatial data on a leaflet map. It is a Python wrapper for Leaflet.js which is an open-source JavaScript library used for plotting interactive maps.

We used Folium to visualize various factors such as the total number of active cases and total number of deceased to understand the spread. We did this in three phases:

1. Global Analysis
2. State Wise Analysis
3. District Wise Analysis

In this phase wise analysis regions were highlighted based on:

1. Number of confirmed and active cases
2. Number of deceased cases

We see that Figures 19 and 20 reinstate the fact that no one is safe from Covid-19 and over the course of 9 months this pandemic has affected all of us. Metropolitan cities such as Mumbai, New Delhi and Pune are worst affected due to the high population density in these cities. These cities report the highest number of active live cases and deaths each day.

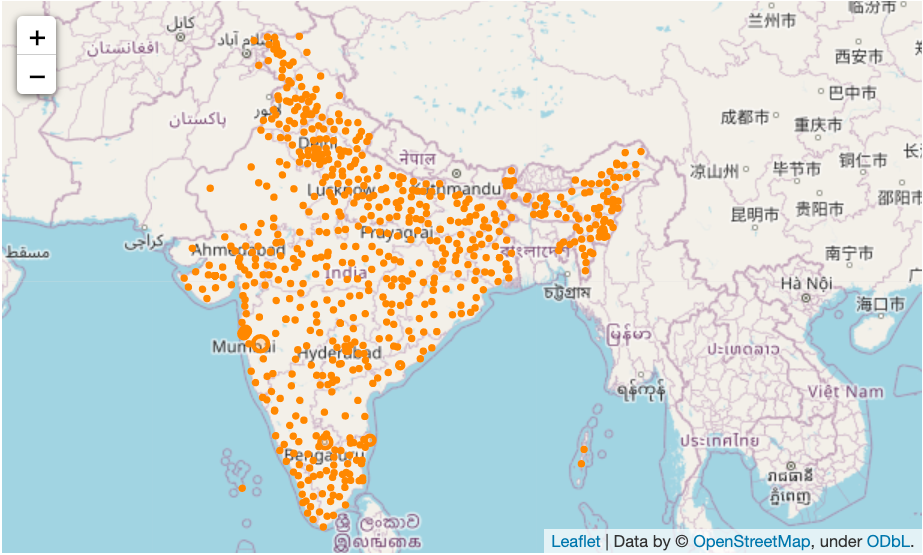


Figure 19. India. Spread of Covid-19.

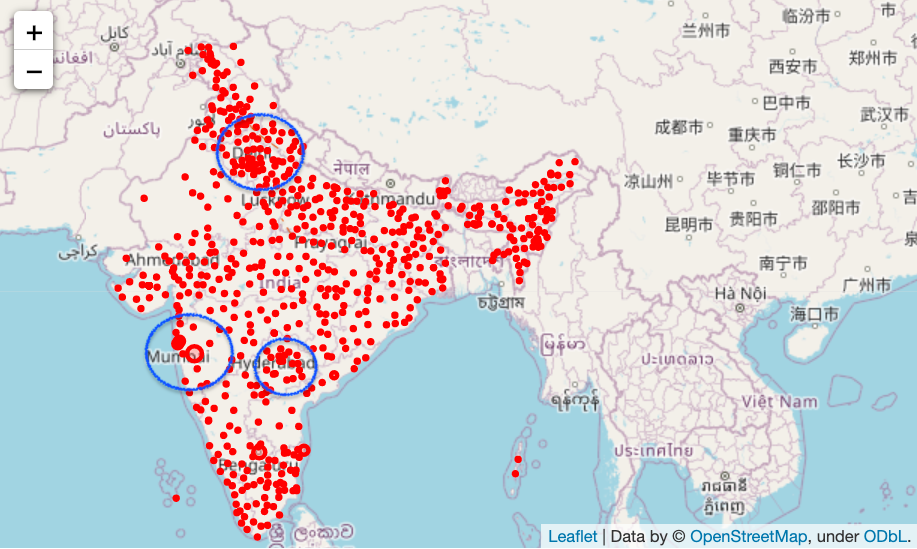


Figure 20. India. This map shows the total deceased in each district till date due to Covid-19.

Also, let us associate this data with another factor, temperature. To understand this let us consider New Delhi, the capital of India. The minimum temperature in New Delhi has been below 10°C for the last few days and a maximum temperature of 46.4°C was recorded in the month of June. Initially there were various reports that suggested that higher temperatures stopped the spread of coronavirus cases but from all the observations so far, we can observe that the virus has transmitted in all areas, including areas with hot and humid weather such as New Delhi.

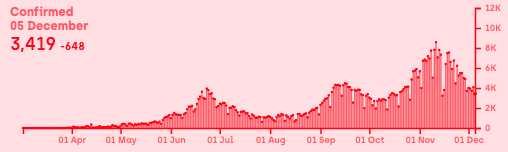


Figure 21. New Delhi, India. Number of Confirmed Cases.

## Technologies Used

* Google Script for appending the latitude and longitude data to the existing database.
* NumPy
* Pandas
* Seaborn
* Plotly
* StatsModels
* Github for collaboration and version control
* Jupyter Notebooks

# RESULTS and DISCUSSION

The results of the analysis and analytics gave clear insights into the dynamics of this virus. Disproportionately, affecting the old and co-morbid population. Still there were instances where a country like Japan were able to contain the virus to a large extent though being a high-risk country. Other countries can infer through these kinds of analytics and figure out better and efficient ways of handling viral diseases such as Covid-19 in the foreseeable future.

Lockdowns played crucial part in containing the virus in many populous countries such as India. The effect of obesity was huge while as the prevalence rate of diabetes was not much on the death rate of the virus. Testing provided countries with important data that enabled them to better manage the Covid-19.

Our forecast model was also able to provide the analytics of this virus 100 days into the future with more than 90% of accuracy. With more incoming data every day, this model will improve linearly.

# CONCLUSIONS and FUTURE WORK

It is clear from this study that Covid-19 is in fact a virus that affects the old and co-morbid population. Handling this population should be high priority for every country. In fact, this analysis and analytics can be performed on any virus that is similar to the dynamics of Covid-19 with minor tweaks.

The forecast model could also be used in similar epidemics with adjustments. More improvements can come from handling more parameters that the datasets provide. Adding weather and climatic conditions to the dataset would have achieved more detailed analysis.

# REFERENCES

|  |  |
| --- | --- |
| [1] | World Health Organisation, "Weekly epidemiological update - 1 December 2020," 1 December 2020. [Online]. Available: https://www.who.int/publications/m/item/weekly-epidemiological-update---1-december-2020. [Accessed 2 December 2020]. |
| [2] | Centers for Disease Control and Prevention, "Covid-19 FAQs," 6 May 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/hcp/faq.html. |
| [3] | J. Dubey, "Covid-19 Factoid: Shifting of epicentre to the US, and 3 other data trends," 1 April 2020. [Online]. Available: https://www.business-standard.com/article/current-affairs/covid-19-factoid-shifting-of-epicentre-to-the-us-and-3-other-data-trends-120040100168\_1.html. |
| [4] | R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice," 2018. [Online]. Available: OTexts.com/fpp2. [Accessed 2 December 2020]. |
| [5] | D. B. D. G. C. G. J. H. B. M. E. M. E. O.-O. H. R. M. R. Cameron Appel, "Data on COVID-19 (coronavirus) by Our World in Data," [Online]. Available: https://github.com/owid/covid-19-data/tree/master/public/data. [Accessed 23 October 2020]. |
| [6] | COVID-19 India Org Data Operations Group, "{Dataset for tracking COVID-19 spread in India," [Online]. Available: https://github.com/covid19india/api. [Accessed 23 October 2020]. |