



Regression Model to predict the effect of PM2.5 on COVID-19

Aman Rajoria - 2020CH70153PM

Harshit Aggarwal -2020CH10087

Tanishq Yadav - 2020CH10138

Avik Ghosh -2020CH10081

Abstract

In this work, we will attempt to develop a relationship between primary reproduction no (R_0) and pollutant concentrations of PM_{2.5} and PM₁₀. We gather statistics on daily covid instances in 14 different Indian states. Delhi, Maharashtra, Karnataka, West Bengal, Assam, Bihar, Punjab, Madhya Pradesh, Kerala, Tamil Nadu, Mizoram, Odisha, Gujarat, and Chandigarh are among the states included.

We also gathered daily particle concentrations and other air pollutant concentrations, as well as total AQI, for all of the main cities in the previously stated states. We then attempted to establish a link between polluting matter and Covid distribution. We also considered the density of contaminants. The initial work was centred on the states of the United States, and we expanded on it in the Indian setting.

Contents

1	Introduction	3
2	Basic Nomenclature	4
3	Problem Formulation	5
4	Data Description	6
5	Numerical Analysis	7
6	Results and Graphs	11
7	Conclusion	20
8	Path Forward	21
9	Self-Assessment	21
10	Reference	22

Introduction

The first instance of the fatal COVID19 virus was discovered in Wuhan, China, in December 2019. The World Health Organization eventually determined it to be a major health risk (WHO). The coronavirus is an enclosed, single-stranded RNA coronaviridae virus. It swept over the world, wreaking havoc on many countries. Conditions had deteriorated to the point that harsh measures such as lockdowns had to be used. While authoring this report, around 5 million people had already died as a result of the virus. When people are in close physical contact, it spreads through the air via respiratory droplets. Many other disease epidemics have previously occurred, including HIV/AIDS, Dengue fever, Malaria, Influenza, and Ebola. Due to a lack of expertise in dealing with this type of sickness, politicians were forced to impose a rigorous lockdown and restrict travel. India had its first COVID-19 case in January 2020, and the illness has since spread throughout the country, with over 30 million total cases and a death toll that has already surpassed half a million. To combat the virus's rapid spread, India imposed its first shutdown at the end of March. After nearly two years of combating two waves, India is gradually attempting to restore things to pre-covid days. It is critical to stop the virus's transmission, which requires focusing on the components that cause it, two of which are thought to be pollutant matter 2.5 (PM_{2.5}) and pollutant matter 10. (PM₁₀). It was proposed that droplets containing virus particles may bind to Particulate Matter (PM), hence boosting viral droplet dispersion in the air. Scientists all across the world are struggling to uncover patterns in observable differences in epidemic propagation speed and/or intensity. More particular, we want to learn about the consequences of air pollution on COVID-19 transmission. The primary reproduction number is used as a measure of transmissibility (R_0). R_0 assesses SARS-CoV-2 transmissibility in a completely susceptible (non-resistant) population with no social distancing. The original article collected and evaluated data for US

states, and we will do the same for select Indian states. We'll compare R_0 levels to PM2.5 and PM10 concentrations to see if we can find a correlation.

Basic Nomenclature-

P rimary Reproduction Number (R_0) - The basic reproduction number assesses SARS-CoV-2 transmissibility in a completely susceptible population in the absence of intervention (social distancing, quarantine). In a completely susceptible population, it is also defined as the average number of secondary infections induced by a first infected individual.

PM2.5 is a fine solid aerosol with a particle diameter of 2.5 μ m that is suspended in the atmosphere. PM2.5 is mostly obtained from typical outside sources such as motor vehicles, biomass burning, and industrial pollutants. Prolonged exposure to PM2.5 is especially hazardous to human health because this small particulate matter is easily absorbed and can enter deep into the lungs.

S - Susceptible

E - Exposed

I - infected

R - recovered

D - Cumulative Detected Cases

N - Population Size.

β - the rate of virus transmission from an infected to the encountered susceptible individual

σ - the inverse of the average incubation period (~3 days), γ - the inverse of the average period of infectiousness

ε - the detection efficiency (as not every infected individual becomes detected)

δ - the detection rate.

PROBLEM FORMULATION

We extracted R_0 using previously reported techniques, including examination of broad infection growth regimes and extraction of R_0 from the exponential growth phase, which we previously used on a global scale.

The flow between the model compartments in the early stages of epidemics, before social distancing measures are introduced, causes changes in the compartment member abundances S (susceptible), E (exposed), I (infected), R (recovered), and D (cumulative detected cases), which are described by the following system of ordinary differential equations:

$$\frac{ds}{dt} = -\frac{\beta SI}{N} \quad (1)$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E \quad (2)$$

$$\frac{dI}{dt} = \sigma E - \varphi I \quad (3)$$

$$\frac{dR}{dt} = \varphi I \quad (4)$$

$$\frac{dD}{dt} = \varepsilon \delta I \quad (5)$$

The model can then be linearized by invoking $S/N \approx 1$, reducing the model to two linear differential equations (1.2) and (1.3). Solving for the eigenvalues of this system,

$$\lambda_{\pm} = \frac{-(\varphi + \sigma) \pm \sqrt{(\varphi + \sigma)^2 + 4\beta\sigma}}{2}$$

$$R_0 = 1 + \frac{\lambda_+(\varphi + \sigma) + \lambda_+^2}{(\varphi * \sigma)}$$

Now we want to make a regression model to see the correlation between PM2.5 and Covid-19. We have extended this modelling for PM10 too.

So, if we consider Y as our dependent variable which is nothing but the R₀ and X as the concentration of PM2.5 or PM10 then the simplest model can be possible as of the following:

$$Y = a_0 + a_1X$$

Where a₁ is the slope of the line and a₀ is the intercept.

Data Description:

We gather statistics on daily covid instances in 14 different Indian states. Delhi, Maharashtra, Karnataka, West Bengal, Assam, Bihar, Punjab, Madhya Pradesh, Kerala, Tamil Nadu, Mizoram, Odisha, Gujarat, and Chandigarh are among the states included.

We also gathered data on daily particle concentrations and other air pollutants, as well as total AQI, for all of the main cities in the aforementioned states.

These statistics cover the period from 2020-01-30 to 2020-07-1, when the Covid-19 was in its first stages.

Data Source:

Air Pollutants: <https://www.kaggle.com/rohanrao/air-quality-data-in-india>

Covid-19: https://www.kaggle.com/sudalairajkumar/covid19-in-india?select=covid_19_india.csv

Population: <https://www.findeasy.in/top-indian-states-by-population/>

Data Processing:

We selected 14 states and union territories from a total of 28 in the Covid-19 data since we have a separate dataset of contaminants. In addition, we eliminated several of the cities included in the air pollution dataset.


We replaced all NAN and missing values of air pollutant concentrations with the average value across a specific city/state and eliminated -- irrelevant data. Furthermore, for convenience of coding, we replaced the city name with the state in which it is located.

All of the graphs and data processing tasks were created using Python libraries. All of the code for this purpose is included in the notebook that was supplied, along with additional codes.

Numerical analysis:

- Data: R_0 = Birth rate of Covid-19, X = Conc. Of PM2.5
 Y (Birth rate of Covid-19) is the dependent variable in our case.
- Modelling: $Y_i = a_0 + a_1X_i$
- Target: Finding the value of a_1 and a_0 to linearise the relation between PM2.5 concentration and Birth rate of Covid-19.

If we have N data points as the following, we can model it like this.

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$


$$y = a_0 + a_1x$$

$$x = b_0 + b_1y$$

$$\frac{x}{b_1} = \frac{b_0}{b_1} + y$$

$$y = \left(\frac{1}{b_1}\right)x + \left(-\frac{b_0}{b_1}\right)$$

If we find the value of slope and intercept, we can find that from the following formula.

$$y = a_0 + a_1x$$

$$a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

$$x = b_0 + b_1y$$

$$b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum y_i^2 - (\sum y_i)^2}$$

$$b_0 = \bar{x} - a_1 \bar{y}$$

Our focus here to follow the first modelling and find the value of intercept and the slope which minimize the error or fit the curve.

We want to do the same for multivariable regression where we treat $R_0(Y)$ as our dependent variable and conc. Of different pollutant like PM2.5 and PM10 as the independent variable.

- Data: R_0 = Birth rate of Covid-19, X = Conc. Of PM2.5
 U = Conc. Of PM10
- Modelling: $Y_i = a_0 + a_1X_i + a_2U_i$

Target: Finding the value of a_2 , a_1 and a_0 to linearise the relation between PM2.5 concentration, PM10 concentration and Birth rate of Covid-19.

Multilinear Regression

Data: $(x_i, u_i, w_i, y_i) \quad i = 1 \text{ to } N$

Model: $\hat{y} = a_o + a_1x + a_2u + a_3w$

Error: $e_i = y_i - \hat{y}_i = y_i - [a_o + a_1x_i + a_2u_i + a_3w_i]$

Objective: To find a_o, a_1, a_2, a_3 such that

$$S_r = \sum_{i=1}^N e_i^2$$

$$\min \sum_{i=1}^N (y_i - [a_o + a_1x_i + a_2u_i + a_3w_i])^2$$

$$\frac{\partial S_r}{\partial a_o} = \sum_{i=1}^N -2(y_i - a_o - a_1x_i - a_2u_i - a_3w_i) = 0$$

$$\frac{\partial S_r}{\partial a_1} = \sum_{i=1}^N -2x_i(y_i - a_o - a_1x_i - a_2u_i - a_3w_i) = 0$$

$$\frac{\partial S_r}{\partial a_2} = \sum_{i=1}^N -2u_i e_i = 0$$

$$\frac{\partial S_r}{\partial a_3} = \sum_{i=1}^N -2w_i e_i = 0$$

$$\sum y_i = a_o N + a_1 \sum x_i + a_2 \sum u_i + a_3 \sum w_i = 0$$

$$\sum y_i x_i = a_o \sum x_i + a_1 \sum x_i x_i + a_2 \sum u_i x_i + a_3 \sum w_i x_i = 0$$

$$\sum y_i u_i = a_o \sum u_i + a_1 \sum x_i u_i + a_2 \sum u_i u_i + a_3 \sum w_i u_i = 0$$

$$\sum y_i w_i = a_o \sum w_i + a_1 \sum x_i w_i + a_2 \sum u_i w_i + a_3 \sum w_i w_i = 0$$

Solve by Gauss Elimination/Gauss Seidel Method

$$\begin{bmatrix} N & \sum x_i & \sum u_i & \sum w_i \\ \sum x_i & \sum x_i x_i & \sum u_i x_i & \sum w_i x_i \\ \sum u_i & \sum x_i u_i & \sum u_i u_i & \sum w_i u_i \\ \sum w_i & \sum x_i w_i & \sum u_i w_i & \sum w_i w_i \end{bmatrix} \begin{bmatrix} a_o \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum y_i x_i \\ \sum y_i u_i \\ \sum y_i w_i \end{bmatrix}$$

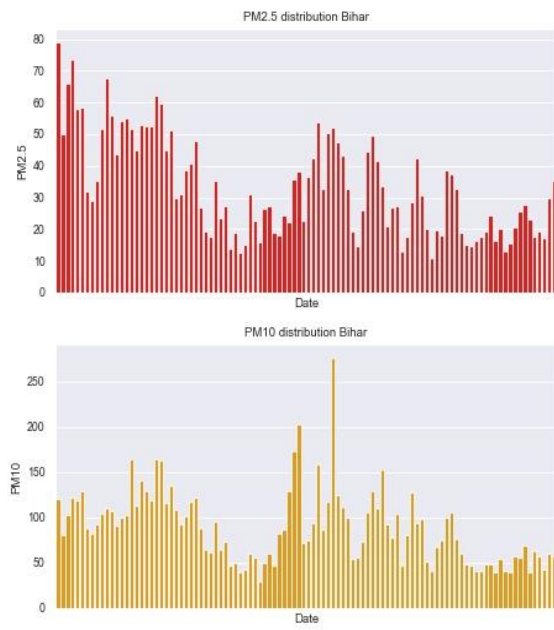
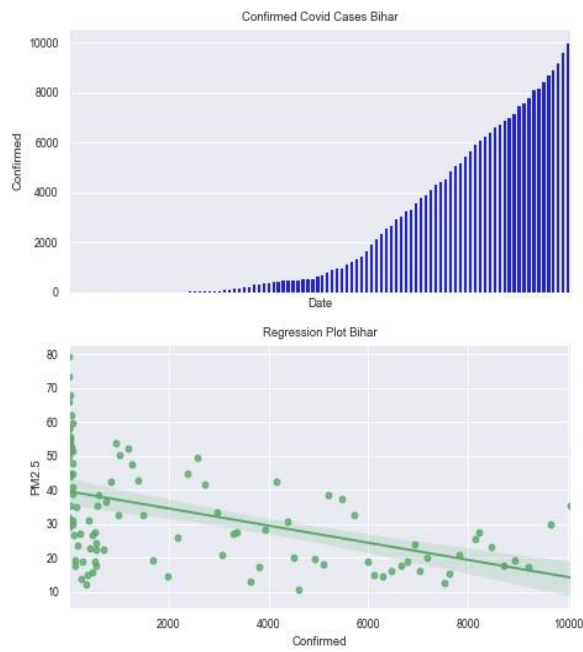
Results and Graphs:

These are the graphs we created using data from public websites. We tried charting confirmed cases vs. time as well as PM2.5 and PM10 concentrations vs. time. Using multivariable regression, we also designed and showed a relationship between COVID cases and PM2.5 concentrations. We have included the CSV files for all of the plotted graphs as well as the programmes used to retrieve the data. We've also included the raw data that we utilised in our computations. In these plots, the dates of covid data and pollutant concentration data are synchronised.

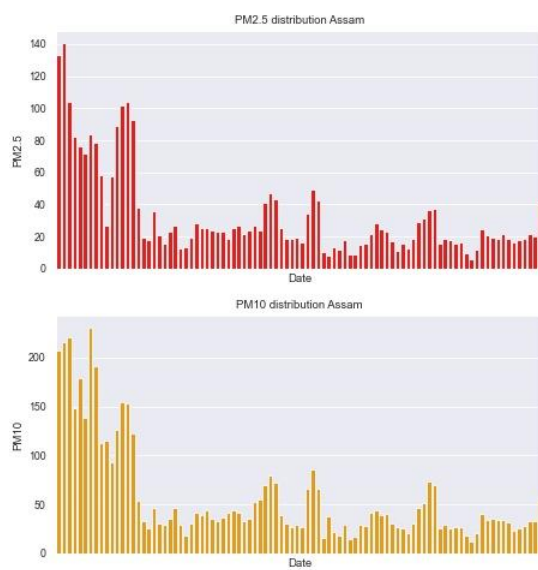
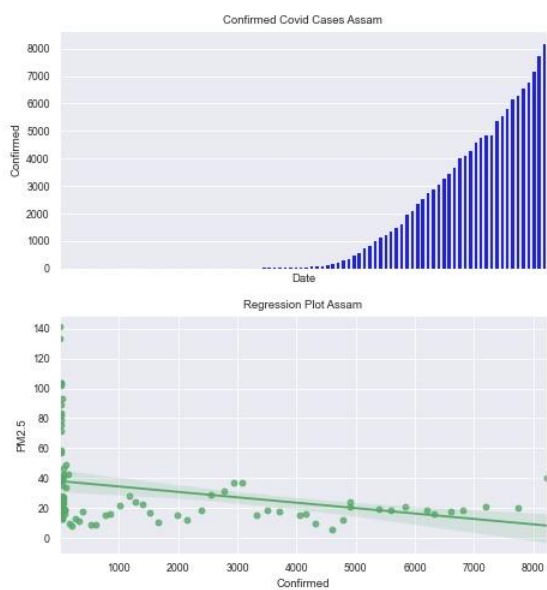
Description: For each fig-plot, it's a combination of 4 plots.

1. Confirmed covid cases VS Date
2. PM2.5 conc. VS Date
3. Regression plot between PM2.5 and Confirmed-covid cases
4. PM10 conc. VS Date

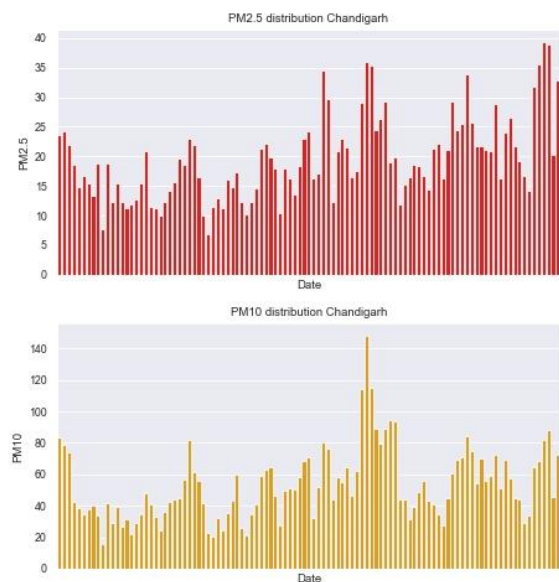
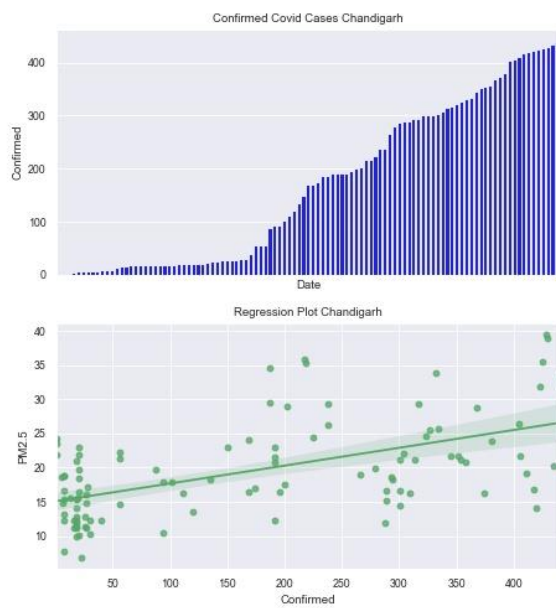
We have also mentioned the name of the states



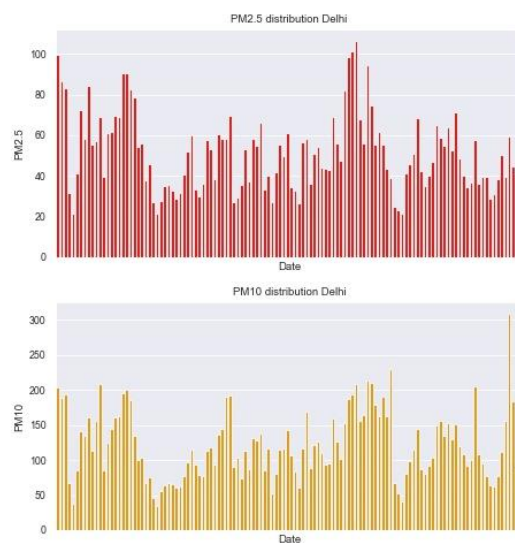
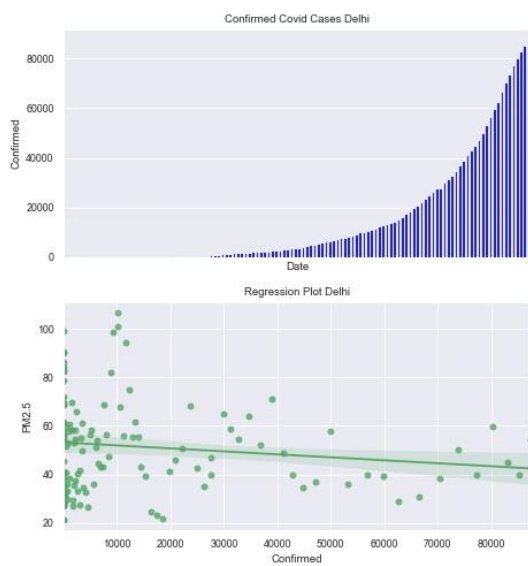
BIHAR



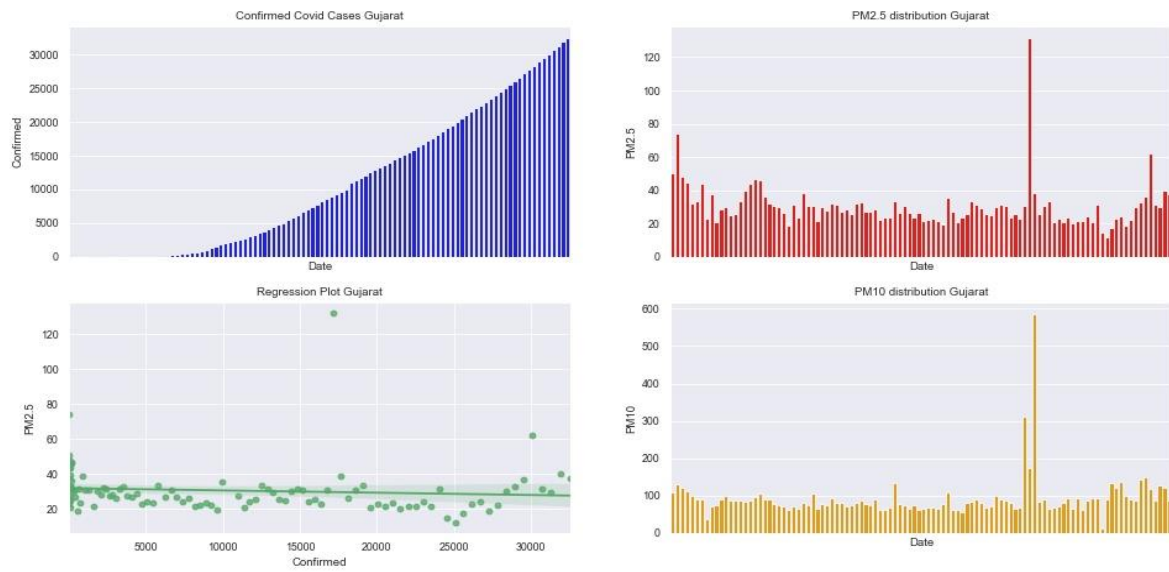
ASSAM



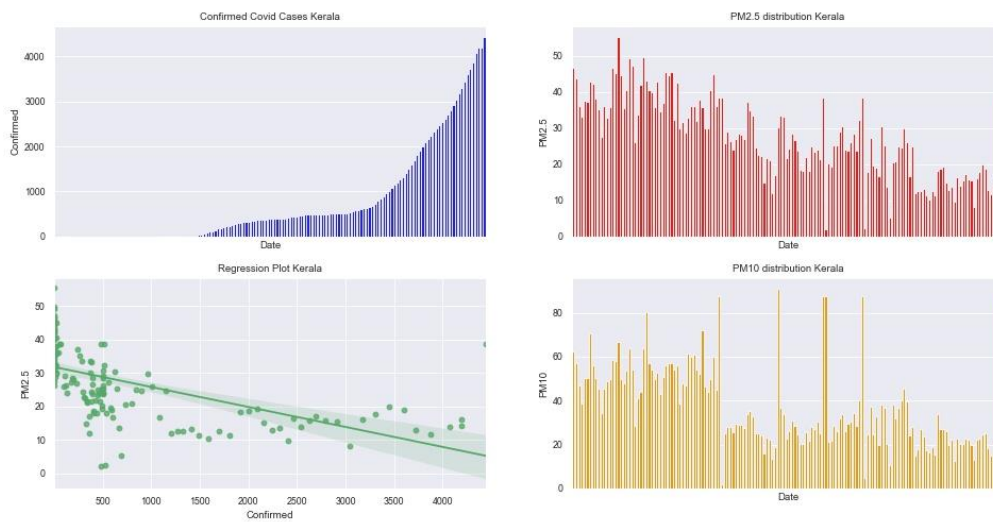
CHANDIGARH



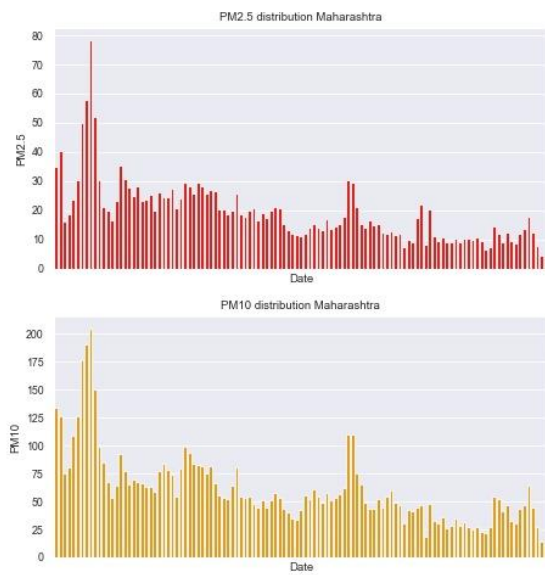
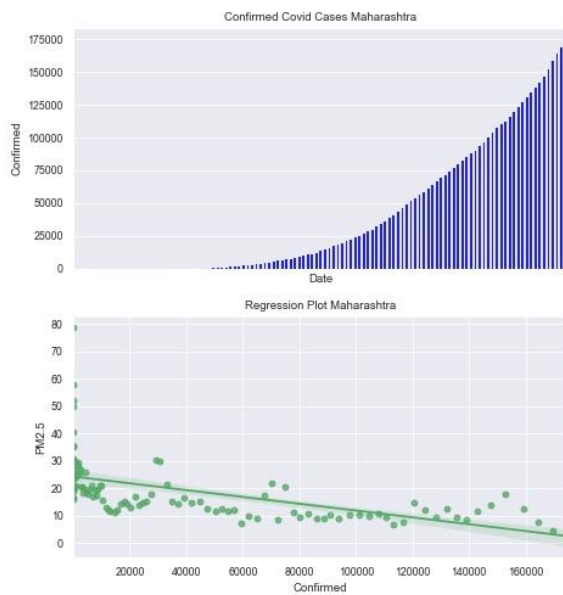
DELHI



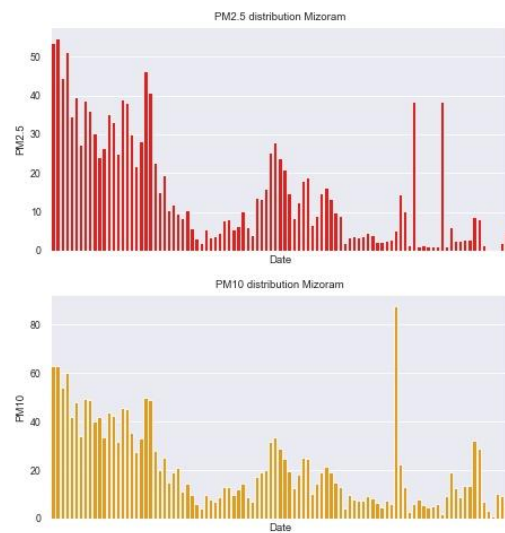
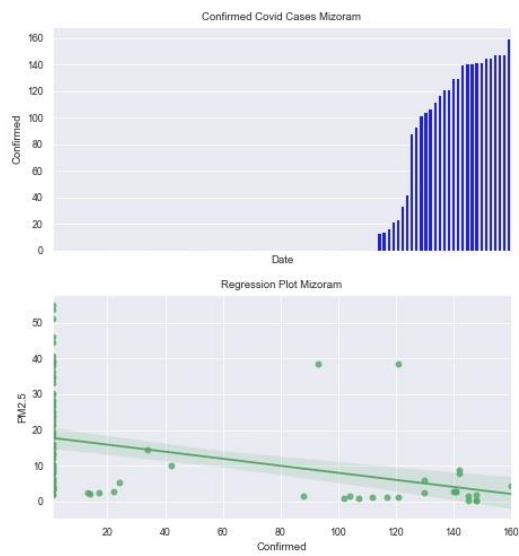
GUJRAT



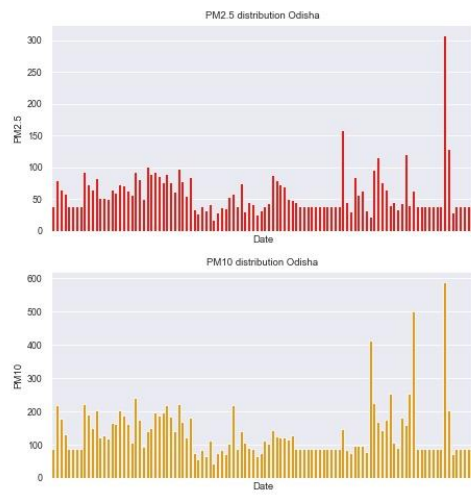
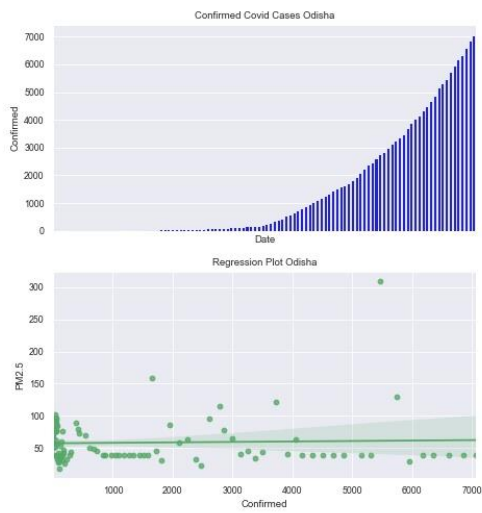
KERALA



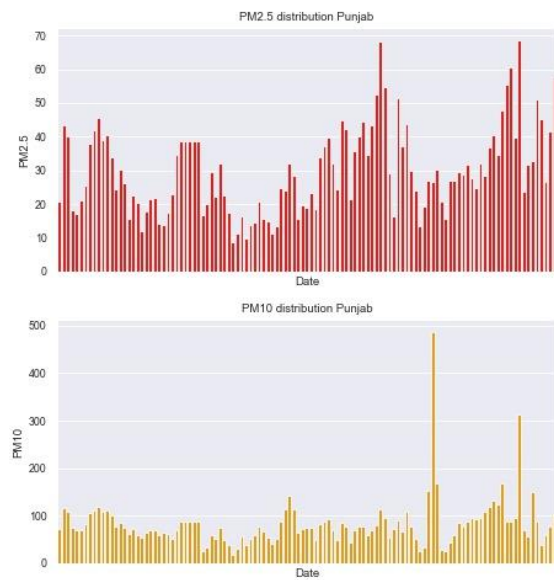
MAHARASHTRA



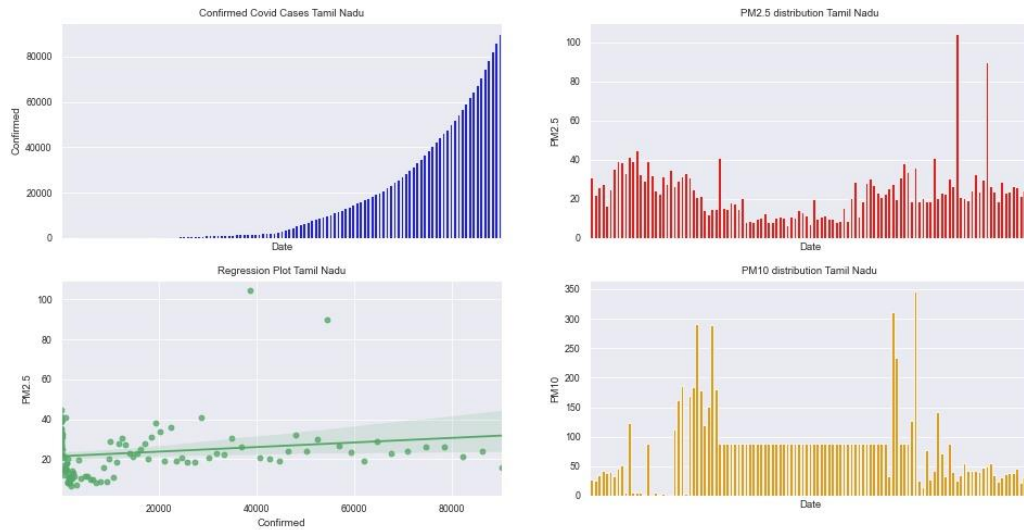
MIZORAM



ODISHA

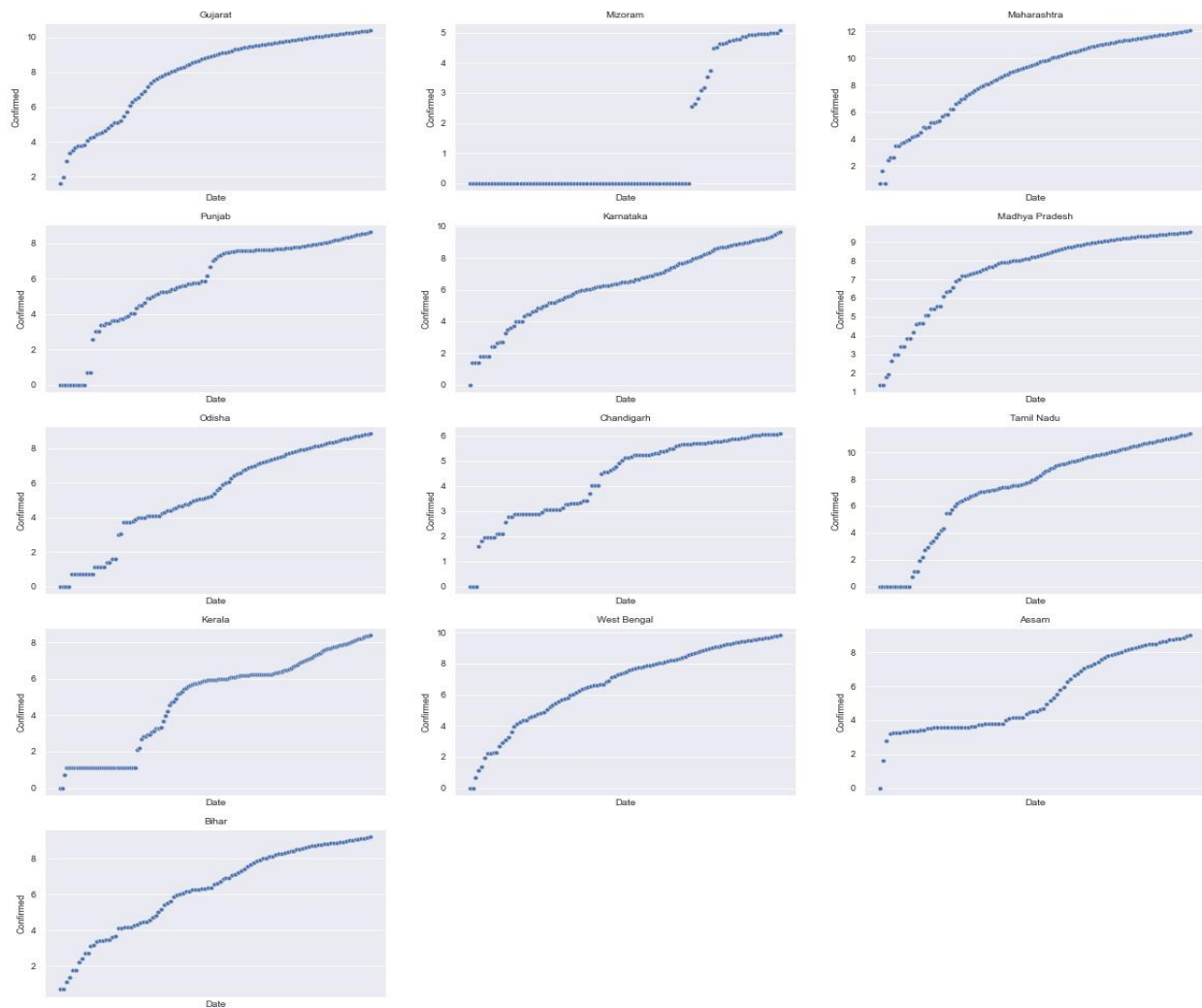


PUNJAB



TAMILNADU

We observed that the logarithm of the confirmed covid cases for each state has a linear relationship with the dates. If we plot all of them we can find the following plots.

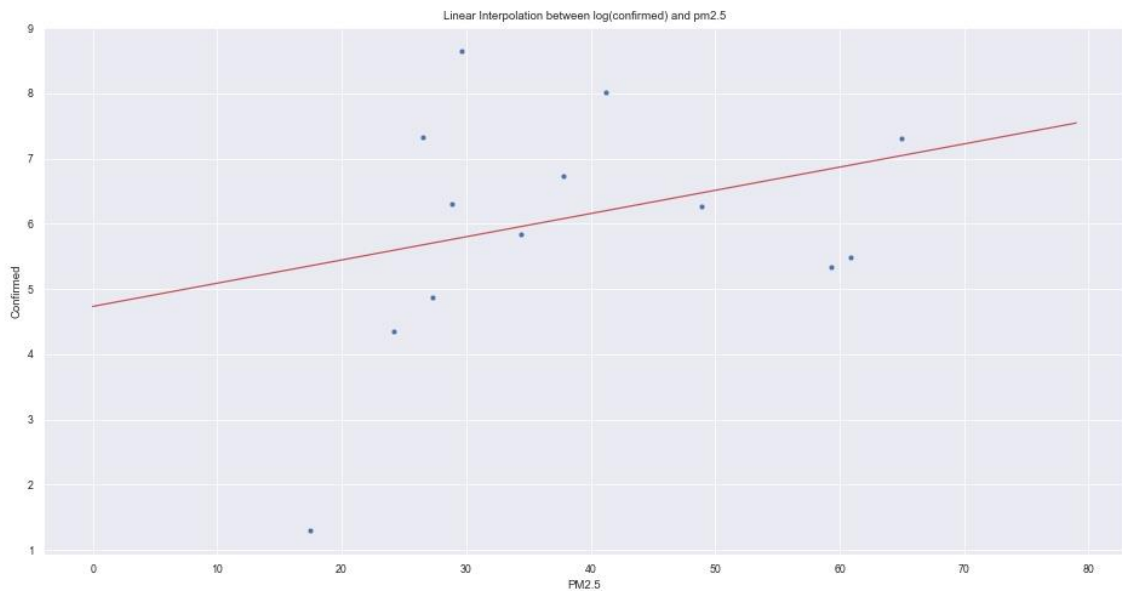


Log(covid cases) VS Dates

Now, if we consider mean value of log (confirmed covid cases) and PM2.5 concentration and follow our regression model, we can get the equation:

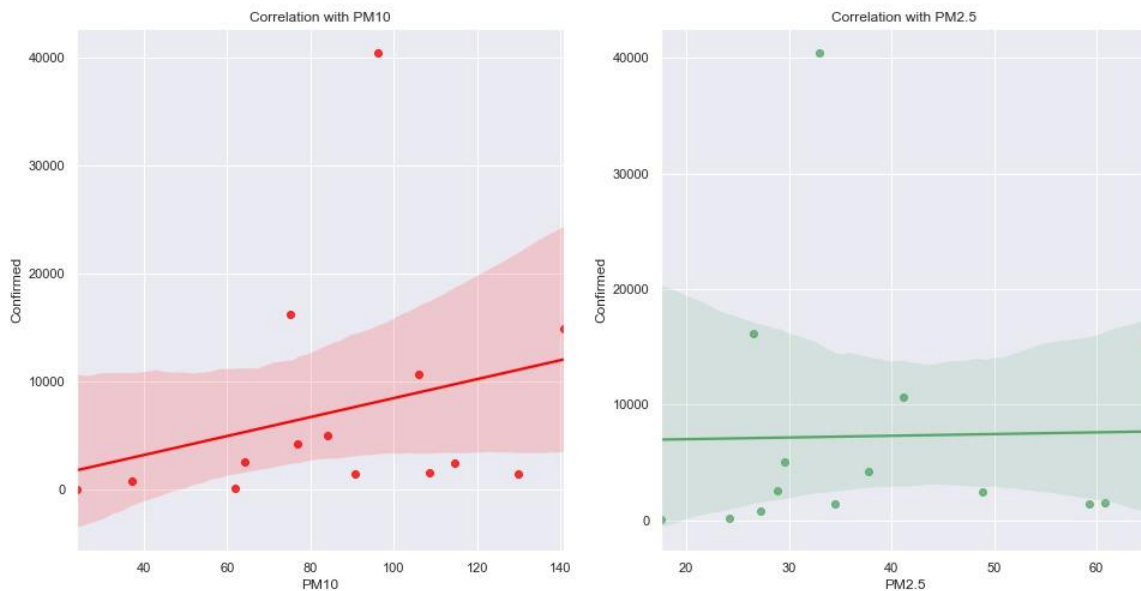
$$Y = 0.0356051 X + 4.7324$$

We can observe the slope is **positive**. If we draw the line and fit to the scatter plot between log(confirmed covid cases) and PM2.5 concentration, we can get the following.



Regression Graph

If we draw regression plot for all of the states and the concentration of PM10 and PM2.5, we can get the following reg plot.



Here, even we can see a high **positive correlation** with pm10 and pm2.5. Though the slope values are different.

Conclusion:

In this paper, we discussed the potential of particulate matter 2.5 (PM2.5) and particulate matter 10 (PM10) to be a potential carrier of the SARS-COV virus, and we attempted to formulate a relationship between the concentration of these pollutant matter and the spread of the Corona Virus using data from 14 different states in India. Using multivariable regression, we attempted to build a linear curve between PM2.5 and total number of cases, and we were able to draw the following conclusions:

1. When we drew the PM2.5 and total case graph, it was linear with a positive slope. As a result, we may infer that PM2.5 is most likely a possible viral carrier.
2. This implies that the virus can be spread via solid aerosols. PM2.5 is defined as fine solids with a particle diameter of 2.5 m that are suspended in ambient air aerosols. Correlations between PM2.5 and other respiratory viruses, such as the influenza virus, have

previously been reported, underlining the likelihood of particulate matter serving as a carrier for SARS-CoV-2.

3. When we attempted the same procedure with PM10, we got identical results. The rise in PM10 concentration resulted in an increase in covid cases. This also implies that PM10 is a possible viral carrier. Furthermore, it implies that there is no relationship between viral concentration and particle size.

Path forward

We would have to neglect many additional aspects that are important in the transmission of the coronavirus due to our lack of mathematical ability and data available in the public domain. Existing research have significant methodological flaws. Also, the premise that the entire population being Susceptible is not optimal. We omit numerous factors that might have had a substantial impact on the results for mathematical reasons. We also ignored the role of any changes in state policy. Someone who is motivated to work on this and has sufficient mathematical skills can evaluate elements such as geography, healthcare, genetics, immunity, and many more. A more better modelling can be possible if we could use various machine learning models like **stochastic gradient descent** or **Lasso regression**. Furthermore, they will require a large amount of additional data, as the data we obtained was mostly from metropolitan regions where AQI calculation apparatus was accessible, but it did not reflect its concentration in the entire state.

Self-Assessment

We recommend our paper at level 2 for the following reasons:

1. We developed our model and solved it, yielding real-world outcomes.

2. We used Indian data rather than the USA states data used in previous studies. We modified the data with little bit of feature engineering, data cleaning and processing.
3. To solve the models, we applied additional numerical approaches along with C++ code of linear regression.
4. For our research, we created an interactive prologue to make it easier and more enjoyable for everyone to grasp the dynamics of an infectious disease like COVID-19.

To summarise, working on this model allowed us to experience the COVID pandemic through the perspective of an epidemiologist, focusing on the technical elements as we uncovered and learnt about the relationships between contaminants and COVID transmission. It also provided us with a greater understanding of what it meant to function as a team. The project also required us to apply the mathematical abilities we gained in CLL113 to real-world challenges.

References:

1. *Ognjen Milicevic, Igor Salom, Andjela Rodic, Sofija Markovic, Marko Tumbas, Dusan Zigic, Magdalena Djordjevic, Marko Djordjevic, PM2.5 as a major predictor of COVID-19 basic reproduction number in the USA, Environmental Research, Volume 201, 2021, 111526, ISSN 0013-9351, <https://doi.org/10.1016/j.envres.2021.111526>, <https://www.sciencedirect.com/science/article/pii/S0013935121008203>*
2. Zhao R, Gu X, Xue B, Zhang J, Ren W. Short period PM2.5 prediction based on multivariate linear regression model. PLoS One. 2018 Jul 26;13(7):e0201011. doi: 10.1371/journal.pone.0201011. PMID: 30048475; PMCID: PMC6062037. <https://pubmed.ncbi.nlm.nih.gov/30048475/>
3. Nor, N.S.M., Yip, C.W., Ibrahim, N. et al. Particulate matter (PM_{2.5}) as a potential SARS-CoV-2 carrier. Sci Rep **11**, 2508 (2021). <https://doi.org/10.1038/s41598-021-81935-9>
4. <https://www.sciencedirect.com/science/article/pii/S0013935121008203>
5. Lecture Notes, CLL113, Prof. Jayati Sarkar

