COVID Data Analysis (April 2023)

A. Chavan, A. Dasari, D. Dasaroju, L.M. Bobbili, V. Pulibandla

***Abstract*— The ongoing COVID-19 pandemic has affected the lives of millions of people worldwide. This paper presents an analytical framework that aims to study the patterns of COVID-19 spread and its effects in the United States. Our framework consists of five stages, each focusing on a specific aspect of the project, namely data and project understanding, data modeling, distributions, and hypothesis testing, basic Machine Learning, and dashboard development. Our approach is data-driven and utilizes advanced statistical and machine-learning techniques to analyze the data and derive insights into the pandemic's impact. Our findings can help policymakers and healthcare professionals make informed decisions to mitigate the spread of COVID-19.**

***Index Terms*—Data modeling, Hypothesis testing, Machine Learning, Pandemic analysis, COVID-19, Statistical analysis, Data-driven insights, Dashboard Development, Public health, United States.**

I. Introduction

THE, the COVID-19 pandemic has bought the world to a standstill, affecting millions of lives and economies worldwide. The United States has been hit hard by the pandemic, with a high number of cases and deaths reported across the country. To gain a better understanding of the pandemic's impact, several datasets can be combined to provide insights into its spread and effects.

One such dataset is the census Demographics ACS, which contains the demographic information for each county in the US. This data can be combined with the COVID-19 data to determine the level of infection in each age group, providing valuable insights for healthcare professionals and policymakers. Additionally, the ACS Social, Economic, and Housing dataset can provide insights into the type of population living in a county/state, providing a better understanding of the pandemic's impact on different communities.

Another important dataset is the Employment Dataset, which provides employment and earning potential by county. This data can help identify regions that may be more vulnerable to economic disruption caused by the pandemic.

The Presidential Election Results dataset can also provide insights into the political learnings of different counties, which may affect the response to the pandemic. Additionally, the Hospital Beds Dataset can provide valuable information on the available Hospital Beds Dataset can provide valuable information on the available hospital beds and ICU units by county/state, showing the decreasing capacity of beds due to COVID-19.

By combining and enriching these datasets, a more comprehensive and nuanced understanding of the COVID-19 pandemic's impact on the UNITED States can be gained. This can help healthcare professionals and policymakers make informed decisions to mitigate the spread of the virus and protect vulnerable populations.

II. Project Stages

A. Stage I

We were given three sets of covid namely Confirmed Cases, Deaths, and Population. The attributes in each table and their description are mentioned in the tables below.

The below table is the data dictionary for the population data set, that we have taken from.

https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

Name	Definition	Data type	Possible Values	Req?
County FIPS	Specific code for the county	Integer	1001,1003	Yes
County Name	Name of the County	String	Barber County	Yes
State	Name of the State	String	AL, AK	Yes
Population	Number of People.	Integer	65432	Yes

The below two tables are the data dictionary for the Number of Cases and the Number of Deaths respectively. These were from the website

https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

Name	Definition	Data type	Possible Values	Req?
County FIPS	Specific code for the county	Integer	1001,1003	Yes
County Name	Name of the County	String	Barber County	Yes
State	Name of the State	String	AL, AK	Yes
State FIPS	Code for State	Integer	1234	Yes
Date	The deaths on that date.	Integer	2,4	Yes

Along with these, we have also answered the question posed as "*Describe how your enrichment data can help in the analysis of COVID-19 spread. Pose initial hypothesis questions.",* and each of our answers was as follows.

From "*Census Demographics ACS*" it was answered as "*We know that covid-19 has a big impact on the world population. This demographic enrichment dataset helps in analyzing which race, age group, sex(male/female), county, and state got most affected"*

From "*ACS Social, Economic, and Housing dataset*" it was answered as "*When we look at the data present, we can see that the main categories are Housing Occupancy, Number of housing units, Total number of rooms and bedrooms, Vehicles available, the years people moved into the homes, and the number of people occupying this housing unit. By analyzing this data, we can see that the number of occupants per house defines the density of the population in that area for that county. So, with this pattern, we can predict the number of COVID cases, as the higher the density of people, the more likely the increase in cases of COVID. There is another column about Vehicles available, with this, we can determine the spread of COVID and the point of origin of the cases from that county if traveled to the other*."

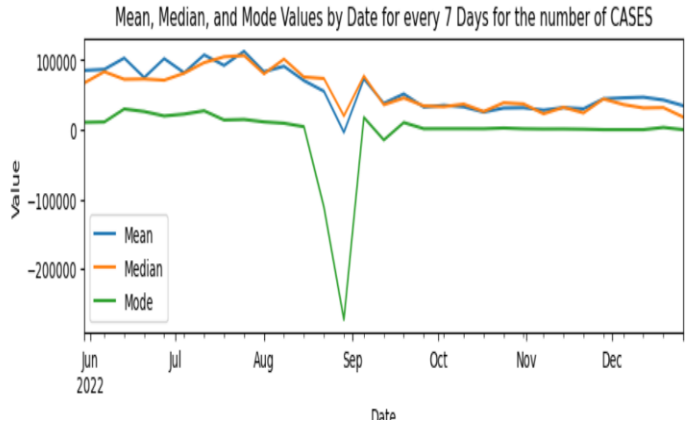
From "*Employment Dataset*" it was answered as *"In Employment Dataset, we have columns such as Area, Ownership, Industry, Area Type, etc. In each county, how many COVID cases are spreading in Industries owned by the Federal, how many COVID cases are spreading in Industries owned by State etc. can be analyzed and Employment Dataset also has Quarterly Wages and Each month employment columns. Based on these columns, we can analyze the COVID spread where the employment count is high or low (i.e., if more people work in an Industry, how the COVID spread is or we can also analyze the covid spread where the wage rate is too high)."*

From the "*Presidential Election Results dataset*" it was answered as *"However, the data does not show which candidate/party was previously elected for the respective position. Therefore, we will collect the previously elected candidate details of the particular state or find evidence of the political inclination for each county or state and combine it with the current dataset to draw proper conclusions."*

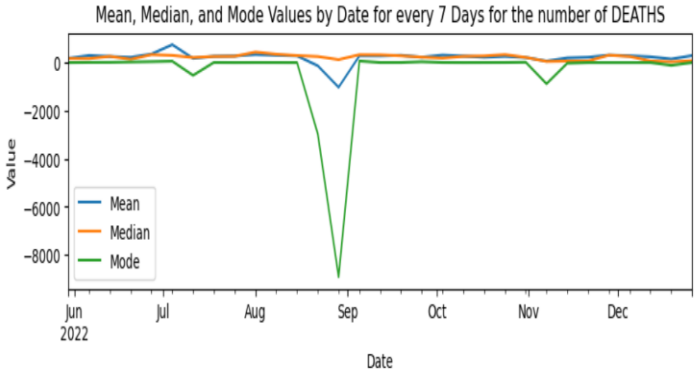
From "*Hospital Beds Dataset*" it was answered as "*In hospital datasets, we have columns such as total beds used, total ICU beds, and many. In each county how many beds are being available and how many beds are used. we have different categories like adult, pediatric ICU, and soon. we have weekly averages and the sum of the beds that are used and the available ones. we can analyze in which county or state the cases are high or low. Does the increasing number of cases in the state have an increasing number of total beds in the state."*

B. Stage II

In Stage-2 we have filtered the covid data from June to December of 2022 across the US. Here, we have considered weekly data from Monday to Sunday, calculated the mean, median, and mode weekly, and plotted their trends.



The above graph shows the weekly mean, median, and mode for the cases across the US.

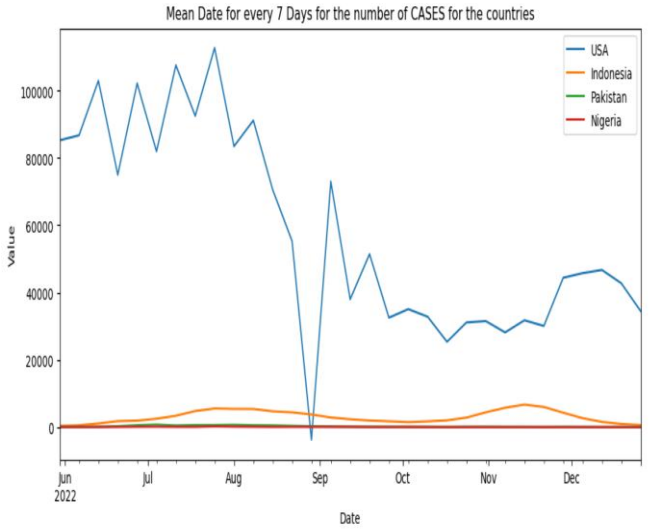


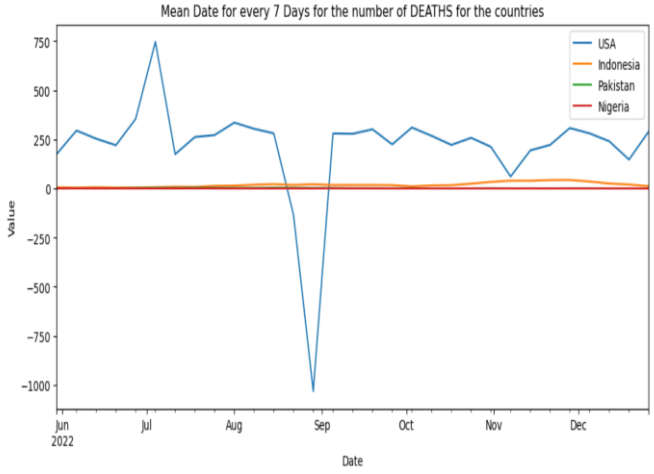
The above graph shows the weekly plot of mean, median, and mode for the deaths across the US. Here in the above cases mean, median, and mode values are not rounded to an integer value.

In the second task, the mean value is rounded to the integer value, and we again do the mean, median, and mode for weekly data and just display the values.

In the next task, we had to compare the data against other countries of the world, so we have chosen the following countries: Indonesia, Pakistan, and Nigeria.

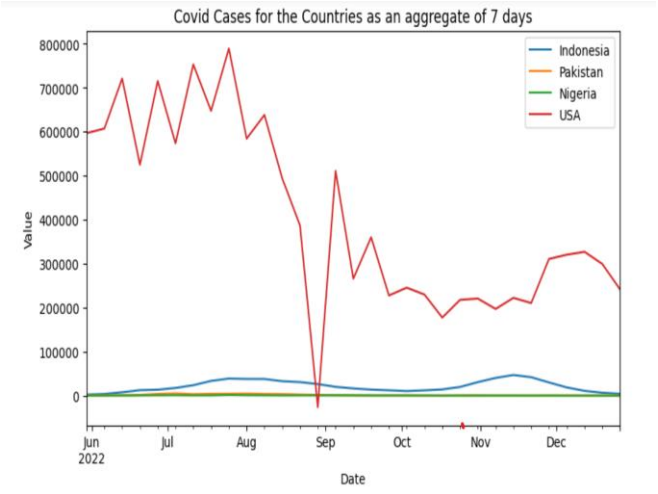
The reason we chose these 3 countries is that according to the Sensex conducted (https://www.worldometers.info/world-population/) we can see that the 3 countries with a similar population base are these 3 even though they are approx. a hundred million people lesser than the USA.

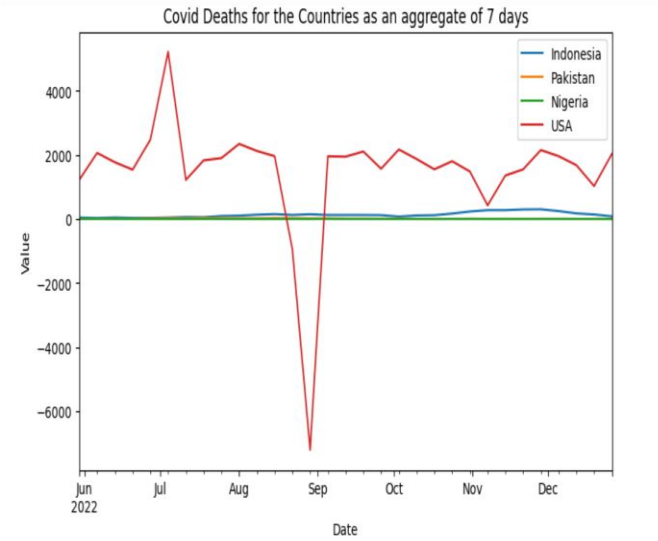




The above two trends show the weekly mean of the cases and deaths in the countries. The trend is going into negative as we have negative data in our data and we cannot clean them as we are finding out the daily cases.

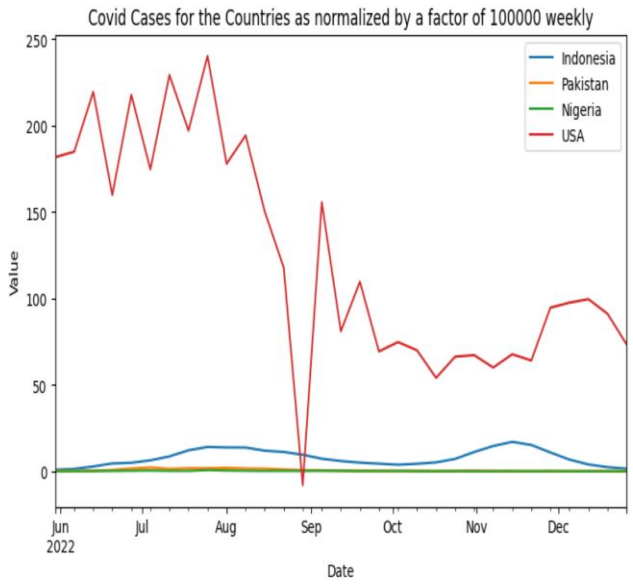
In the next task, we aggregated the cases and deaths for the countries. The trends are shown below:

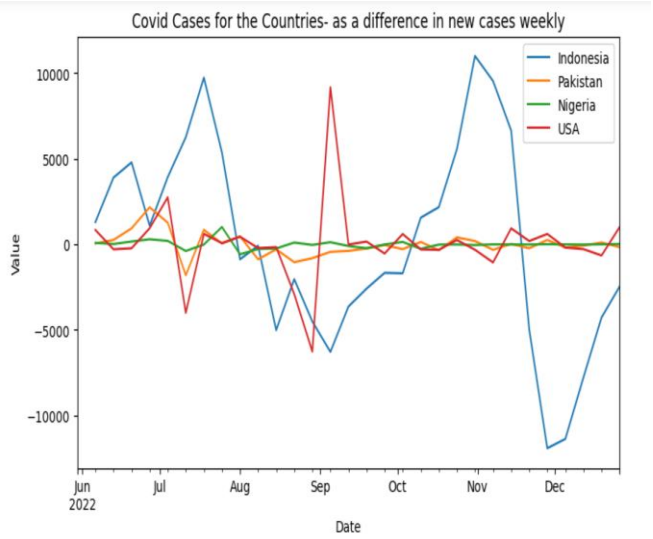




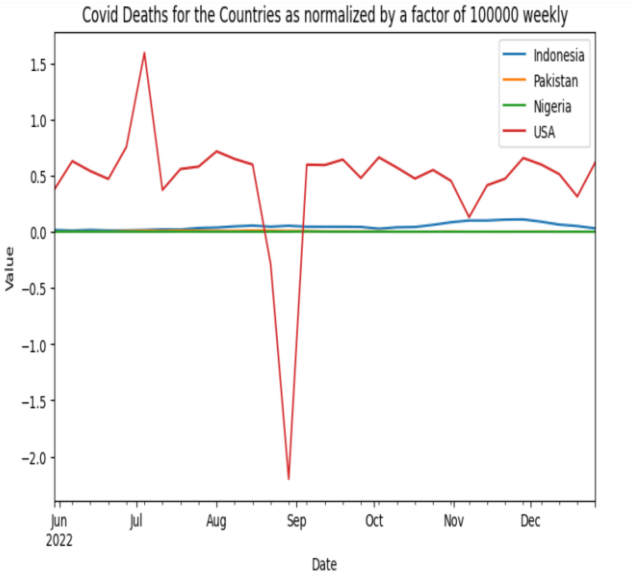
The trend in the above graph is negative because we have negative values in our data, and we cannot clean them as we are finding out the new cases and deaths.

We have normalized the cases and deaths with a normalization factor of 100000 and divided it with the population and plotted for the countries.



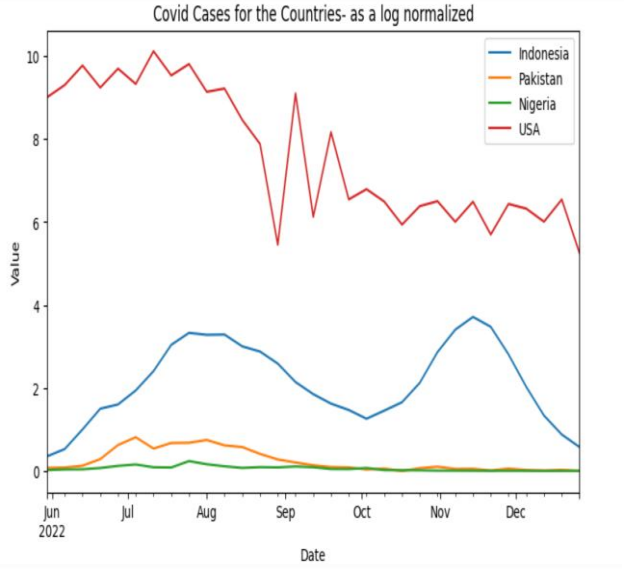


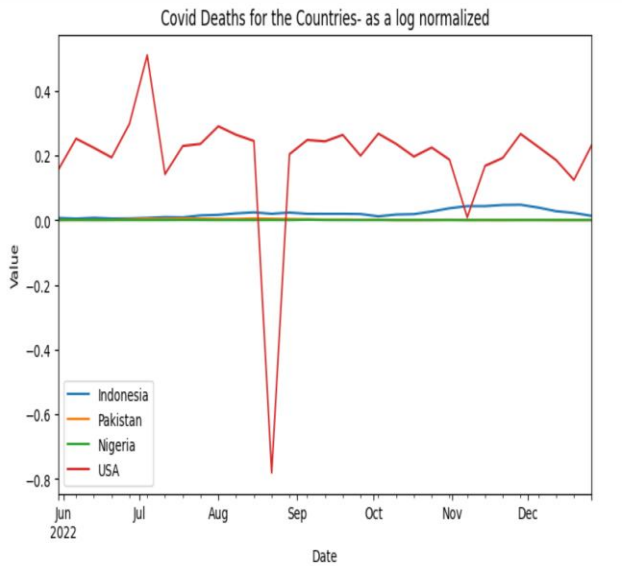
We have also plotted the log normalized plots for the cases and deaths for all the countries.



The trend in the above graph is negative because we have negative values in our data, and we cannot clean them as we are finding out the new cases and deaths.

Below are the plots of the new cases and deaths weekly:





Here we are adding the 1 before taking the log to avoid logging a zero/ negative value. It's a very common practice done by mathematicians.

The purpose of taking log values is to reduce the scale of the data and make it easier to compare between countries. The actual values of the data are less important in this context, as the focus is on the trends and patterns in the data. Hence, there may be log values plotted that is not in between the range of 0 and 1.

The graph has negative values because we have negative values in our data, and we cannot clean them as we are finding out the new cases and deaths.

We have also identified the peak weeks for the case and deaths in the US and other countries:

	Country	Date	deaths
0	USA	2022-07-04	5225
1	Indonesia	2022-11-28	302
2	Pakistan	2022-08-15	27
3	Nigeria	2022-09-05	6

	Country	Date	cases
0	USA	2022-07-25	789033
1	Indonesia	2022-11-14	46863
2	Pakistan	2022-07-04	5080
3	Nigeria	2022-07-25	1492

Here we are considering the highest peak for the number of cases and deaths in the respective countries and showing them from the date of starting of that week.

According to the research that we did, we go the following findings as to why there was a sudden rise in COVID-19cases/deaths:

For "The *USA",* There were many cases reported in the months of June, July, and August. During these months there were a lot of reasons for people to gather which could explain the reason for the increase:

Roe vs Wade rally: in the months of May and June, there were many rallies and protests conducted which made a huge number of people gather, which could have caused an increase in COVID.

4th of July celebration (Independence Day): it's one of the most celebrated holidays in the USA, where many people (friends and family gather to celebrate).

Oak Fire incident: In the month of July and August, there were loads of forest fires that occurred on the west coast, due to which lots of people had to relocate and migrate to other areas, if even a few people had any symptoms of COVID when they migrated to another place, they could have transmitted them to others.

For "*Indonesia*", There were cases reported in the months of August and November. The reasons for the increase in COVID cases are:

The outbreak of Monkey Pox: There was a sudden outbreak of Monkeypox in the country which cause panic among the masses. Because of this people started migrating from the infected areas to other places. During this migration, COVID cases increased.

Earthquake: In the month of November there was an earthquake that struck the island of Cianjur Regency, in West Java, which caused many people to lose their homes and people had to relocate to different places which may have caused an increase in COVID cases.

For "*Pakistan",* There were cases reported in the months of July and August. The reason for this increase is that during these months there were few festivals, namely Moharram and Eid - ul - Azah which are considered national holidays. During these festivals, the whole family gathers to celebrate which may be a reason for the increase in COVID cases.

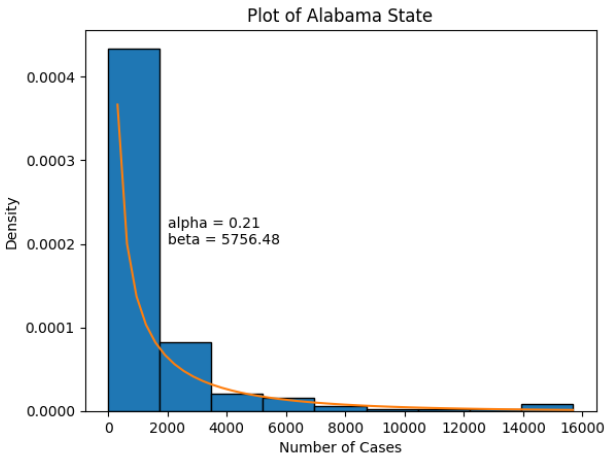
For "*Nigeria*", There was a rise in the number of cases from the month following July, the reason is that the government up until then had strict regulations regarding COVID and gathering of people but in July they relaxed the regulations on COVID, and because of this there was a rise in COVID.

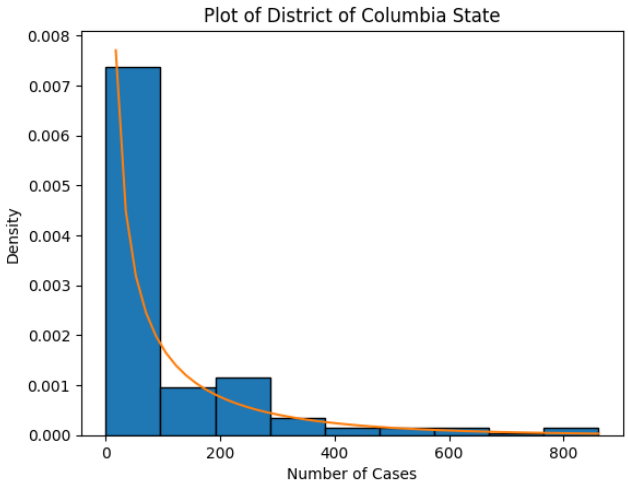
C. Stage III

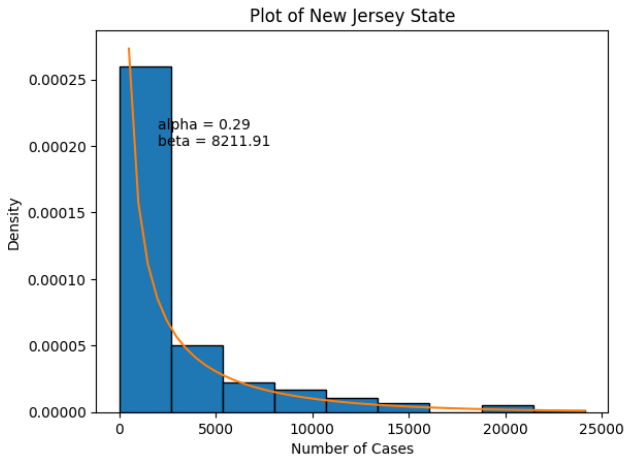
The goal of this Stage is to develop distributions and formal hypothesis tests for the intuitions. Here we use the state data generated in Stage II to fit a distribution to the number of COVID-19 new cases using any of the MoM, MLE, and KDE methods.

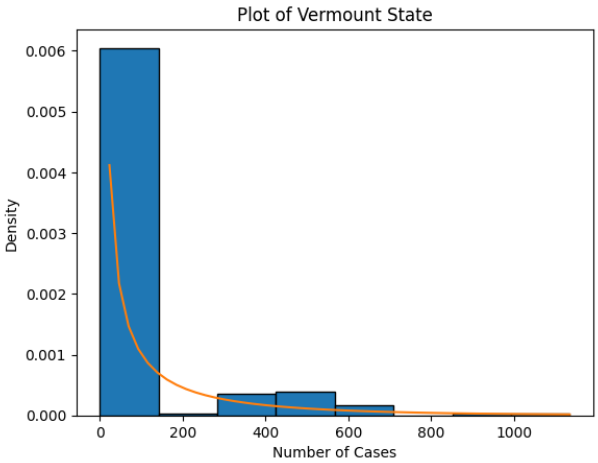
Here each member of our team has done different approaches/ methods to fit the distribution.

In statistics, the "*Methods of Moments (MoM)*" is a method of estimation of population parameters. The same principle is used to derive higher moments like skewness and kurtosis. It starts by expressing the population moments as functions of the parameters of interest. Those expressions are then set equal to the sample moments.

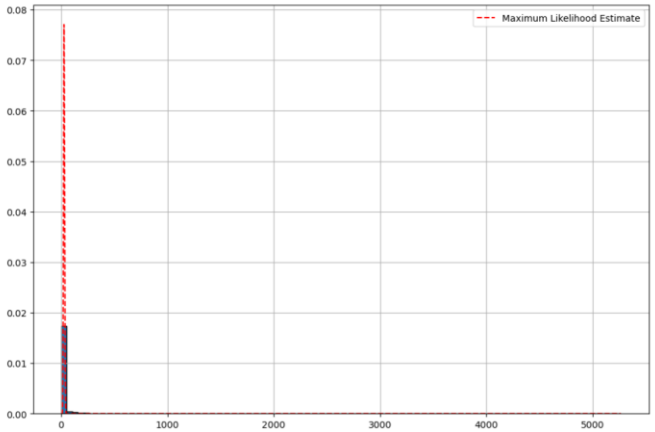


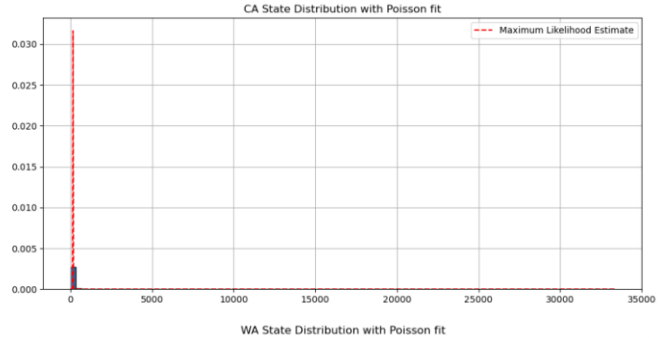


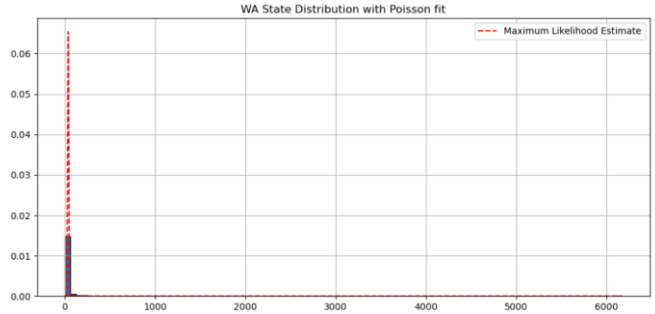


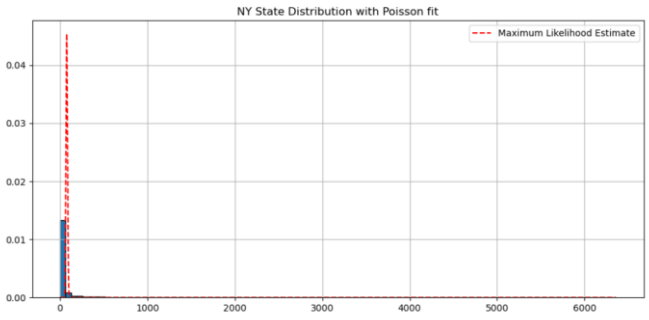


In statistics, "M*aximum Likelihood Estimation (MLE)*" is a method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable.

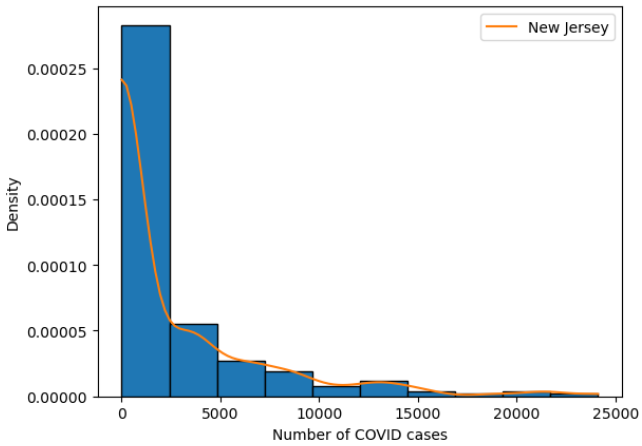




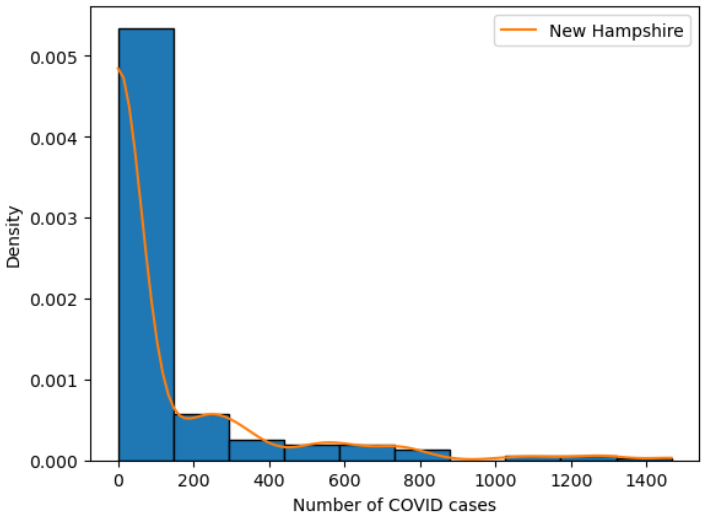


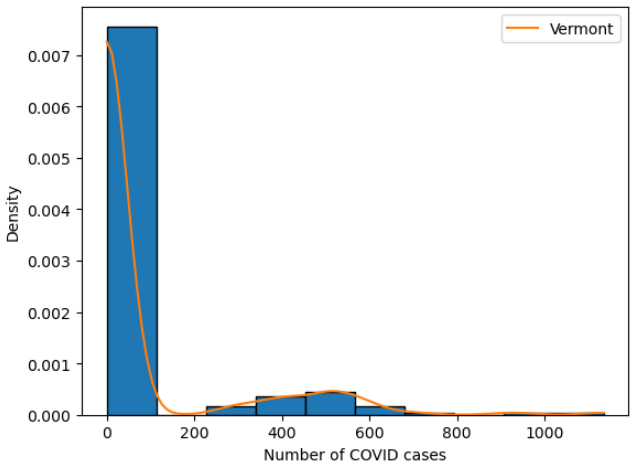


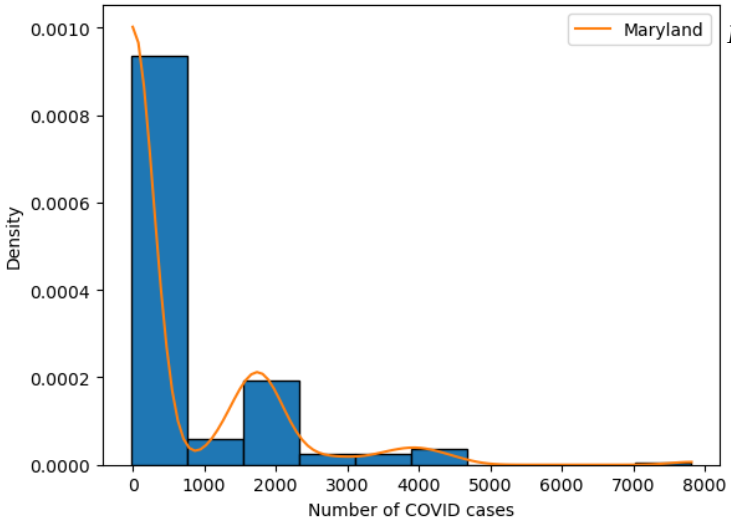
In statistics, "*Kernel Density Estimation (KDE)*" is the application of kernel smoothing for probability density estimation, i.e., a non-parametric method to estimate the probability density function of a random variable based on kernels as weights.



Below are the graphs that show a comparison of other states for a KDE Method.



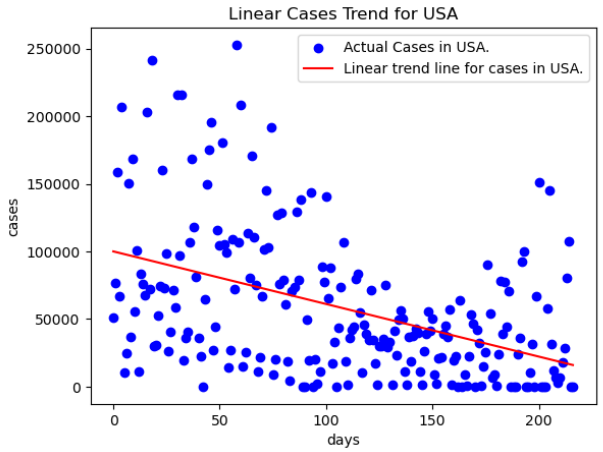


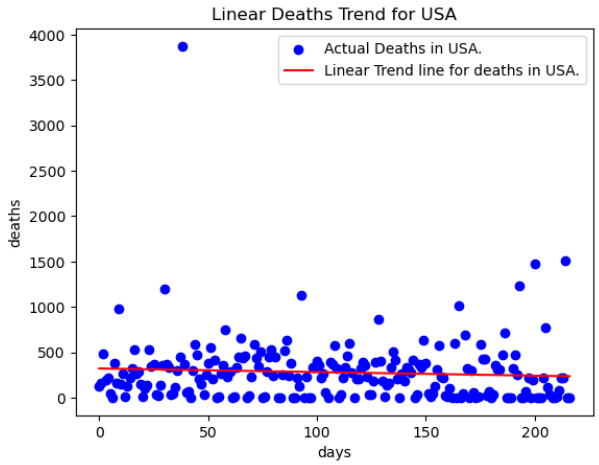


D. Stage IV

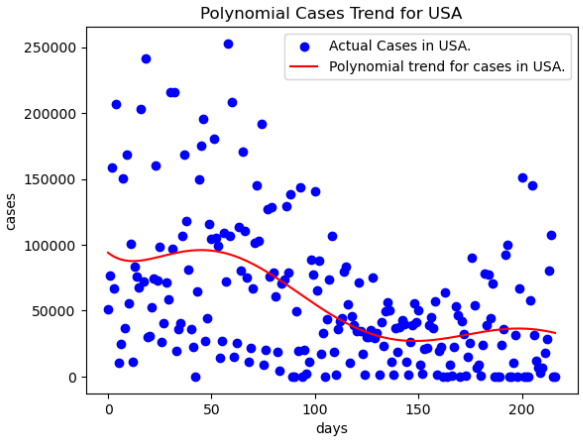
The goal of this stage is to utilize machine learning and statistical models to predict the trend of COVID-19 cases/deaths. Here we are developing the Linear/ Non-Linear regression models for predicting cases and deaths for not only the USA, but we would be comparing these models with 3 other countries with similar population density.

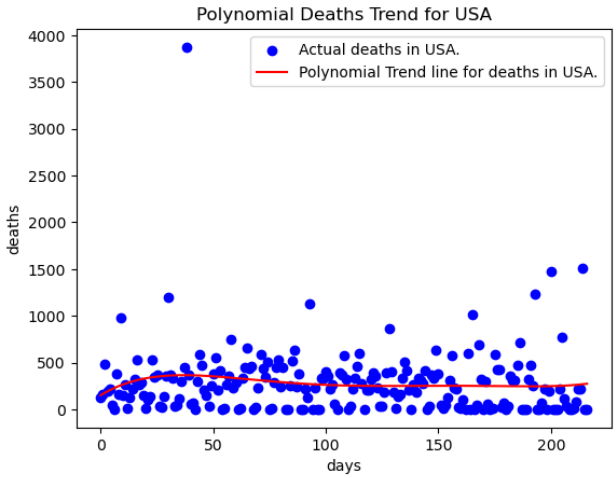
The following are the regression model graphs for the USA.



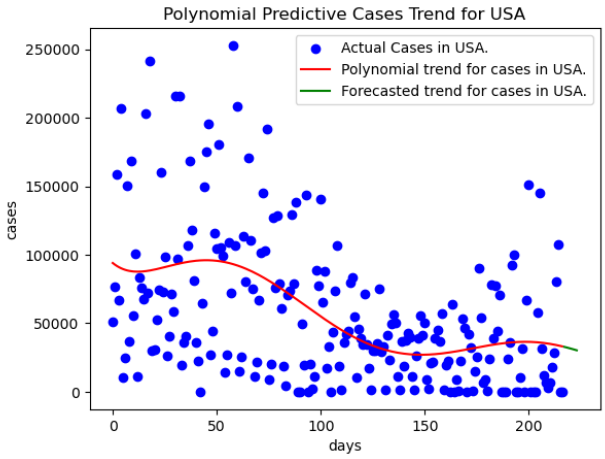


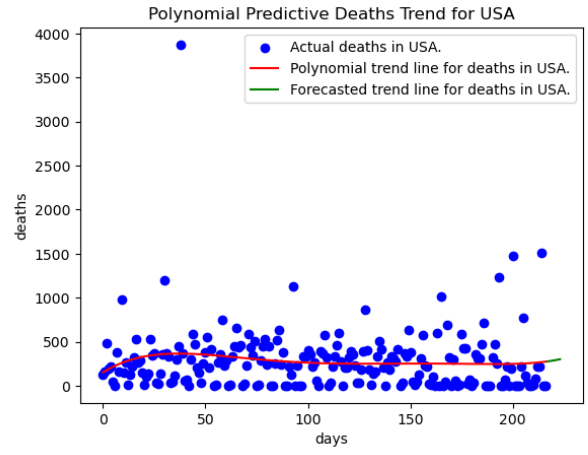
After seeing this graph we can say that the Linear regression model doesn't answer our prediction hence we have to conclude that we need to use a non-linear model for predicting the cases and deaths





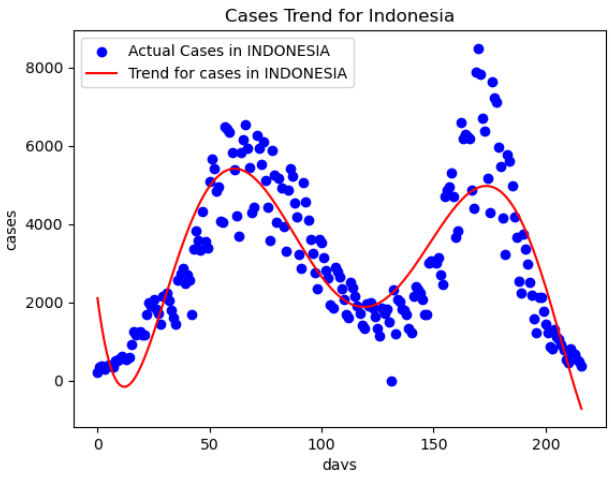
After finding these we can find the prediction for the next week using this model. Since we have trained this model, we can predict next week. (Here this prediction line is denoted by the green line at the end of the red trend line).

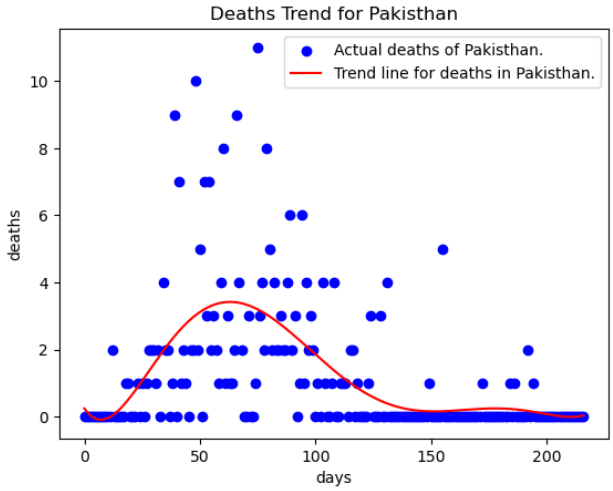


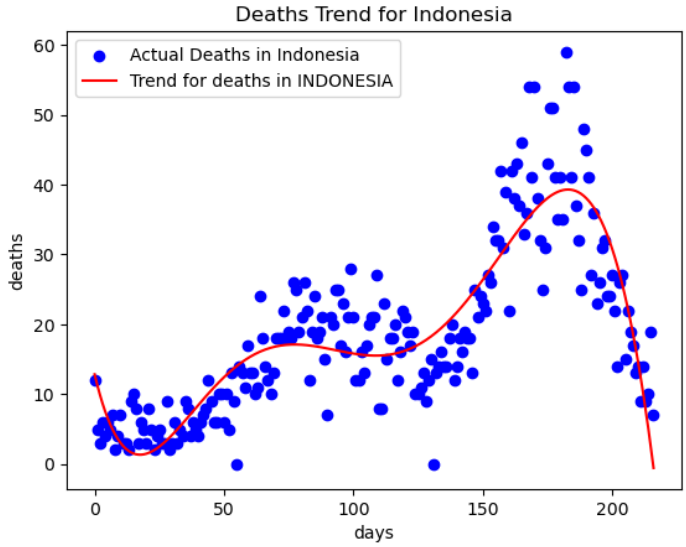


Now we will even plot the graphs for the other countries.

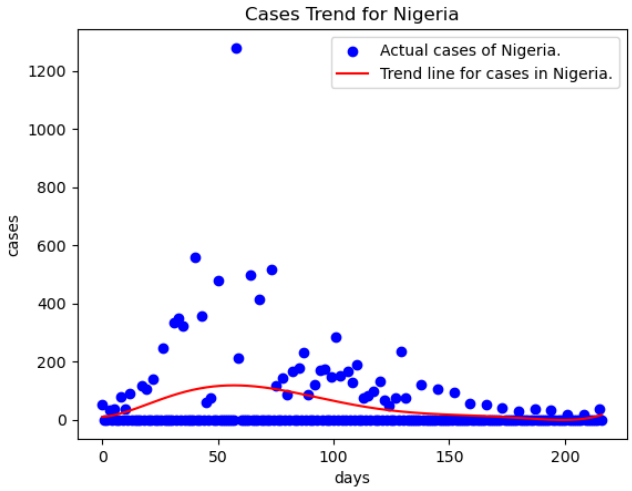
"*Indonesia"*



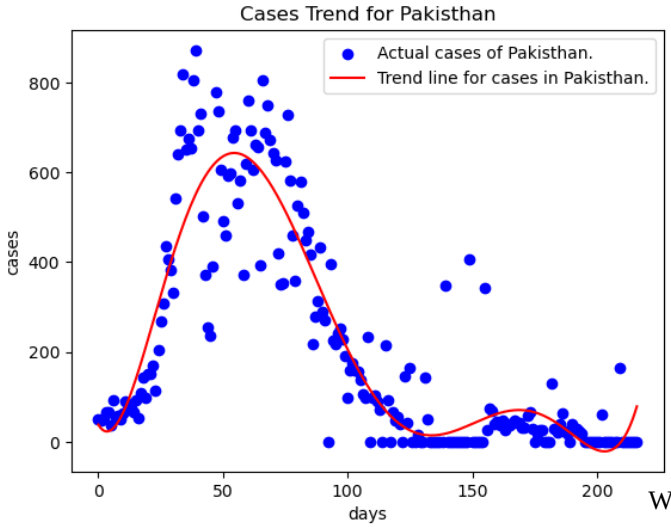


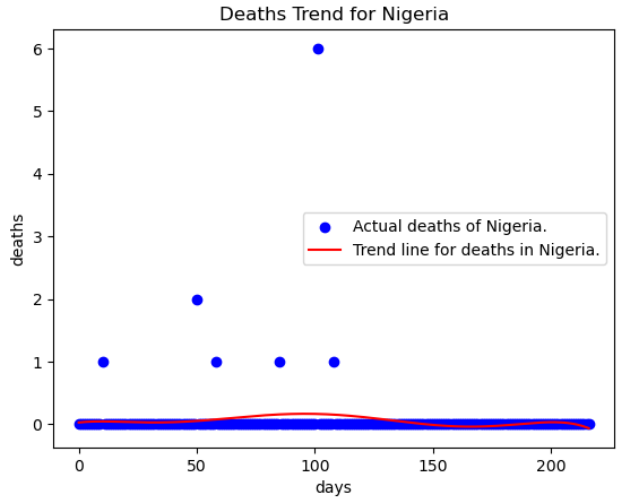


"Nigeria"



"Pakistan"





When compared to other nations the USA has the highest number of cases per day where as Nigeria has the lowest number of cases each day. At the end of the last six months of 2022, Nigeria's covid cases are almost equal to 0, whereas the USA has thousands of covid cases even at the end. Pakisthan covid cases have declined after 100 days and reported less than 100 cases for the next hundred days. Cases in Indonesia are very fluctuating cases sharply declined between 100 to 150

days and then rapidly increased after 150 days and reported almost 0 cases after the 200th day.

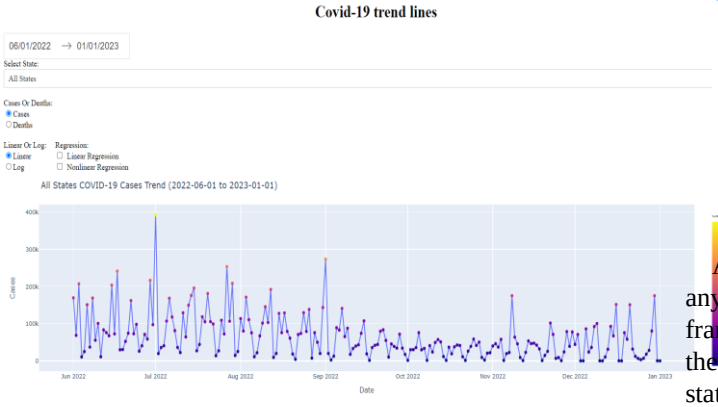
Nigeria has the lowest number of deaths on each day whereas the USA has the highest number of deaths on each day. Pakistan has its maximum number of deaths between 50 to 100 days and Indonesia has its highest number of deaths on each day between 150 to 200 days. Overall USA has witnessed a maximum number of deaths each day for the last 6 months of 2022.

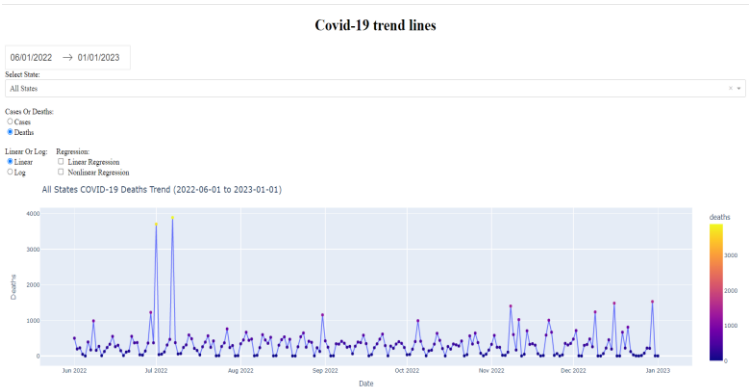
E. Stage V

In this stage, we have created an interactive dashboard to show the graphs and plots that we have created until now. To achieve this, we have used Plotly and Dash to create this interactive graph dashboard. Plotly library is an interactive open-source plotting library that supports over 40 unique charts covering a wide range of statistical, financial, geographical, scientific, and 3-dimensional use cases.

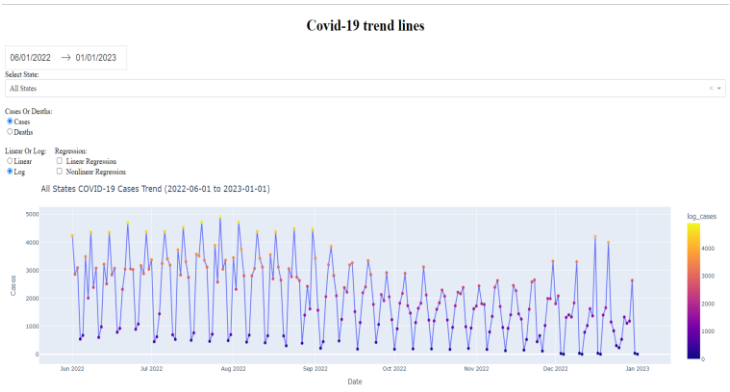
Along with Plotly, there's Dash. It is the best way to build analytical apps in Python using Plotly figures. Many specialized open-source Dash libraries exist that are tailored for building domain-specific Dash components and applications. Some examples are Dash DAQ, for building data acquisition GUIs to use with scientific instruments, and Dash Bio, which enables users to build custom chart types, sequence analysis tools, and 3D rendering tools for bioinformatics applications.

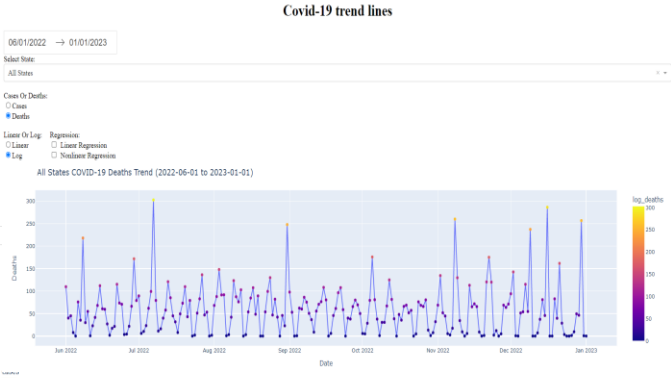
The following show the plots from the dashboard that we have created:



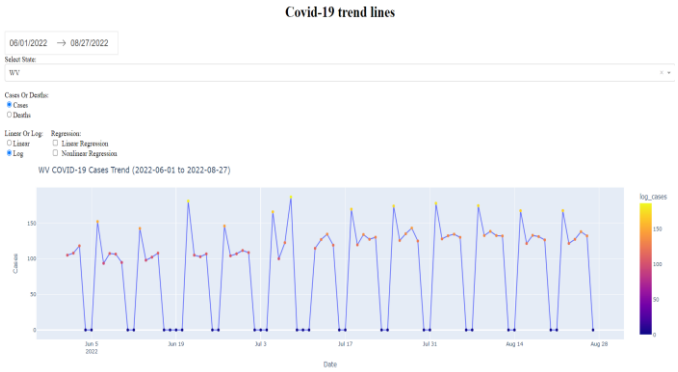


The above two graphs show the linear graph plot of the cases for all the states combined.

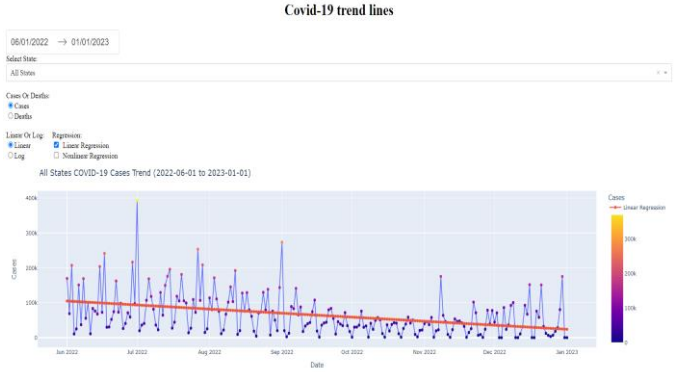


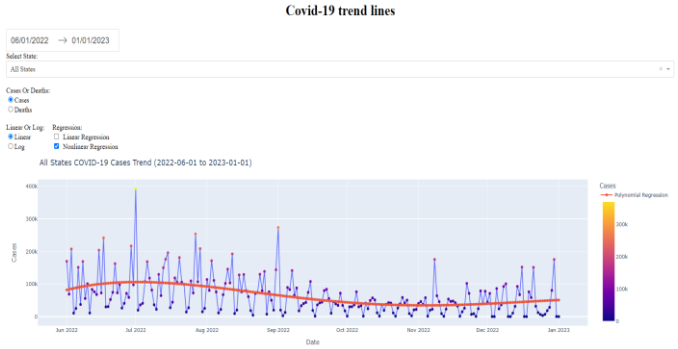


Along with this, we have also given the provision to select any of the chosen states of the US within the desired time frame. As shown below, we have shown the graph plotted for the time frame from June 1st,2022 to August 27th, 2022 for the state of West Virginia.



Along with this, we have also provided the option to check the Linear and non-Linear regression. Below graph show the same for all the state's trend.

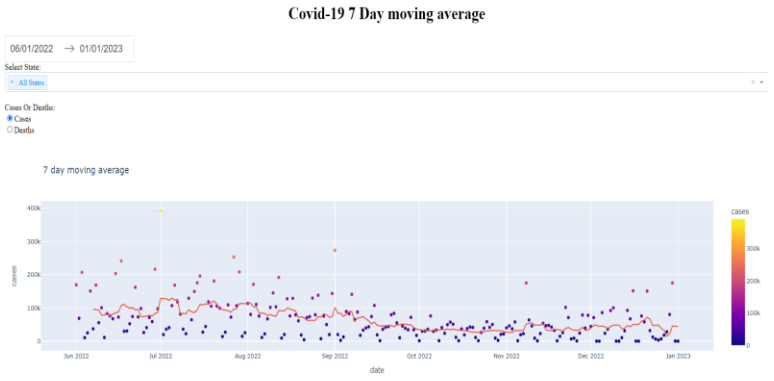


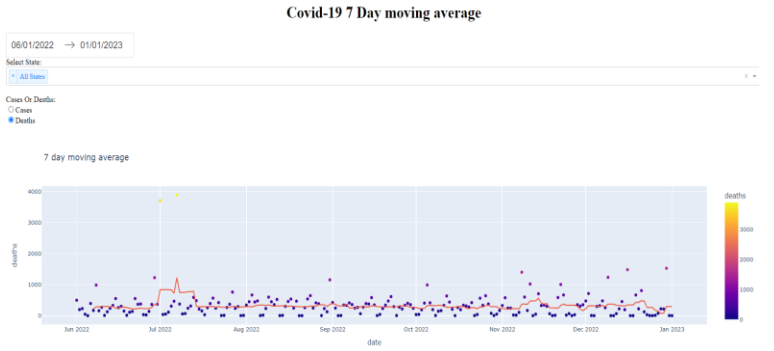


Not only can we see the graphs one at a time, but we can also see the graphs between linear and non-linear regression at the same time just to have an idea of how the data points are getting fitted

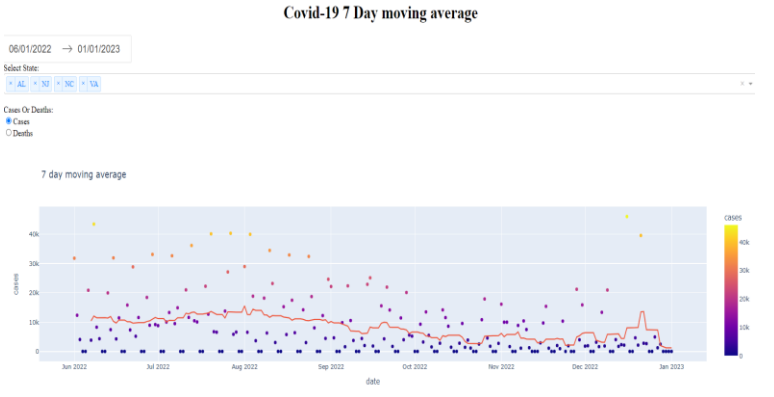
These are the graphs related to regression. But we have also provided along with this to see the rolling moving weekly average for the given timeframe of 6 months. In statistics, a moving average (rolling average or running average) is a calculation to analyze data points by creating a series of averages of different selections of the full data set. It is also called a moving mean (MM) or rolling mean and is a type of finite impulse response filter.

Below is the graph related to it.





Along with this, we have given the option to select multiple states to compare the moving average.



III. Conclusion

The COVID-19 pandemic has created unprecedented challenges across the United States, affecting millions of lives and disrupting economies. To understand the pandemic's

impact and develop effective strategies to mitigate its spread, it is essential to gather and analyze accurate data from multiple sources.

Enrichment datasets, such as the Census Demographic ACS, provide valuable demographic information that can be combined with COVID-19 data to gain a better understanding of the pandemic's impact on different age groups and communities. Additionally, datasets such as the Employment Dataset, Presidential Election Results, and Hospital Beds Dataset provide insights into the economic, political, and healthcare contexts in which the pandemic is unfolding.

By combining and enriching these datasets, it is possible to develop a more comprehensive and nuanced understanding of the COVID-19 pandemic's impact on the United States. This information can help inform decision-making at the local, state, and national levels and ensure a more effective response to the ongoing crisis.

As the pandemic continues to evolve, it is essential to continue gathering and analyzing data from multiple sources to stay informed and adapt to changing circumstances. By working together and using the insights gained from these datasets, we can mitigate the impact of the pandemic and move towards a more sustainable and resilient future.

Appendix

Appendixes, if needed, appear before the acknowledgment.

Acknowledgment

We would like to express our sincere gratitude to all individuals who have contributed to the success of this research project. We would also like to acknowledge the invaluable technical support provided by our teachers, without which this project would not have been possible.

We are grateful to our colleagues and collaborators who provided their support, advice, and expertise throughout the project. We would like to thank the participants who contributed to this study and made this research possible.

Finally, we would like to thank our families and friends for their encouragement and support throughout the research project. We are truly grateful for their understanding and patience during this time.

References

[1] US COVID-19 Cases and Deaths by State (https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/)
[2] Census Demographic ACS (https://data.census.gov/cedsci/table?q=dp&tid=ACSDP1Y2018.DP05)
[3] ACS Social, Economic, and Housing (https://data.census.gov/cedsci/table?q=dp&tid=ACSDP1Y2018.DP05)
[4] Employment Dataset (https://www.bls.gov/cew/downloadable-data-files.htm)
[5] Presidential Election Results (Political Leanings) (https://www.kaggle.com/unanimad/us-election-2020)
[6] Hospital Beds Dataset (https://protect-public.hhs.gov/pages/hospital-utilization)
[7] How to write a good README for your GitHub project? (https://bulldogjob.com/news/449-how-to-write-a-good-readme-for-your-github-project)
[8] What is a data dictionary? A simple and thorough overview. (https://analystanswers.com/what-is-a-data-dictionary-a-simple-thorough-overview/)
[9] Our World Data (https://analystanswers.com/what-is-a-data-dictionary-a-simple-thorough-overview/)
[10] How to plot a Confidence Interval in Python. (https://www.statology.org/plot-confidence-interval-python/)
[11] Plotly (https://plotly.com/)
[12] Plotly- Dash (https://plotly.com/dash/)
[13] Getting started with Dash. (https://www.youtube.com/watch?v=hSPmj7mK6ng)
[14] Introducing JupyterDash (https://medium.com/plotly/introducing-jupyterdash-811f1f57c02e)