Covid-19 Data Analysis

Saipavan Tadikonda
Computer Science
UNC Greensboro
Greensboro, USA
s_tadikonda@uncg.edu

Anshu Belkhede
Computer Science
UNC Greensboro
Greensboro, USA
a_belkhede@uncg.edu

Silpa Yerramreddy
Computer Science
UNC Greensboro
Greensboro, USA
s_yerramred@uncg.edu

Srilekha Geda
Computer Science
UNC Greensboro
Greensboro, USA
s_geda@uncg.edu

***Abstract*—The COVID-19 pandemic has resulted in significant human losses and disruption to global economies, societies, and healthcare systems. To effectively control the spread of the epidemic, a thorough understanding of its characteristics and behavior is crucial. This report presents an analysis of COVID-19 in the United States during the second half of 2022. The study compares different states, counties, and the overall situation in the USA with other countries. Data regarding population, deaths, and cases were obtained from the USAFacts platform and other relevant sources. The report employs statistical modeling and forecasting techniques to predict future numbers of COVID-19 cases and deaths. Furthermore, the analysis explores the impact of COVID-19 on the United States, considering factors such as Omicron variant, low vaccination rate, public gathering and other public health measures. These findings and our understanding of the pandemic's progression can further help in enabling policymakers, healthcare professionals, and the general public to make informed decisions and implement effective strategies in combating COVID-19 and other future pandemics.**

I. Introduction

The coronavirus disease 2019 (COVID-19) pandemic resulted in 764,474,387 reported cases and 6,915,286 deaths overall, out of which 104,538,730 reported cases and 1,130,662 deaths in USA till April, 26 2023. The COVID-19 pandemic has become one of the most significant global health crises in recent history, profoundly impacting societies, economies, and healthcare systems worldwide. Originating in late 2019, the novel coronavirus disease quickly spread from its epicenter in Wuhan, China, to become a global pandemic. Elderly individuals, particularly those above the age of 65, individuals with pre-existing medical conditions, along with frontline healthcare workers have been significantly impacted from COVID-19. Additionally, marginalized and disadvantaged populations, including racial and ethnic minorities, low-income communities, and individuals with limited access to healthcare resources, have faced greater challenges during the pandemic. This report focuses on the analysis of COVID-19 deaths and new cases in the United States during the second half of 2022, comparing different states, counties, and the overall situation in the USA with other countries. The objective of this analysis is to gain insights into the progression and impact of the pandemic, as well as to provide predictions for future COVID-19 cases and deaths.

The outbreak of COVID-19 was first reported in December 2019, when clusters of pneumonia cases of unknown origin emerged in Wuhan, Hubei Province, China. The causative agent was identified as a novel coronavirus, later named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). By early 2020, the virus had rapidly spread across borders, leading to an alarming increase in the number of infections and deaths globally.

In the United States, the initial cases of COVID-19 were reported in January 2020. As the virus continued to spread, the number of cases and deaths escalated rapidly. In the early stages of the pandemic, the United States witnessed a surge in cases concentrated in several regions, including New York, California, and Washington State. The healthcare system faced immense pressure as hospitals struggled to accommodate the influx of patients, and the country grappled with the challenges of widespread testing, contact tracing, and implementing containment measures.

To initiate the study, data was collected from various sources, including the USAFacts platform, which provided crucial information on population, deaths, and cases. Additional datasets, such as Census Demographic data, Social, Economic, and Housing datasets, Employment data, Presidential Election Results, and Hospital Beds data, were incorporated to enhance the understanding of the factors influencing the pandemic's impact. To resolve inconsistencies and get meaningful interpretations, we first cleaned them by removing unwanted spaces in the beginning and end of each column data, deleting the necessary data like rows in the County Name column having Statewide Unallocated data and converting each data frame from wide to long format. We merged Covid-19 cases, Covid-19 deaths, Covid-19 population datasets together and further merged it with enrichment datasets.

Following Data Modeling and Distributions were conducted. Statistical techniques such as the Method of Moments (MoM), Maximum Likelihood Estimation (MLE), and Kernel Density Estimation (KDE) were employed to fit appropriate distributions to the number of COVID-19 new cases. To account for population differences, data normalization was performed. Correlation analyses were also conducted to identify potential relationships between COVID-19 variables and socio-economic factors, forming the basis for hypothesis formulation. Here we converted the daily data into weekly data to see the weekly trends like peaks and valleys of new cases and new deaths across the USA and other countries. We observe

there were peaks in the months between June-August 2022, possibly due to the Omicron variant and November-December 2022, probably due to public gathering and transport for holidays like Thanksgiving and Christmas. Also, no peaks in the months between September-November were observed, as the second booster was given to all age groups, particularly for the older age group. Moving forward, basic Machine Learning techniques were applied. Linear and Nonlinear (polynomial) regression models were used to predict future COVID-19 cases and deaths in the United States. The accuracy of these models was assessed using Root Mean Square Error (RMSE). Trend lines were plotted to visualize the data and provided one-week-ahead forecasts. Also, we did hypothesis testing to check if the null hypothesis is correct or an alternate hypothesis.

In addition to the analytical stages, a simple interactive dashboard was developed using Plotly and Dash frameworks. This dashboard serves as a user-friendly interface, enabling us to explore the analyzed COVID-19 data through dynamic visualizations. The dashboard facilitates a deeper understanding of trends, distributions, and predictions related to COVID-19 cases and deaths in the United States during the second half of 2022.

II. DATASETS DESCRIPTION

A. COVID-19 Cases Dataset

This covid confirmed cases dataset contains the confirmed cases for each county across the USA from January 22nd, 2020, to January 16th, 2023.

Name	Datatype	Description
countyFIPS	Integer(int64)	Unique five-digit number for each county
CountyName	Object	Name of the county
State	object	Name of the State.
stateFIPS	Integer((int64)	Unique number for each state.
Cases (on each date)	Integer(int64)	Number of cases on each date from January 22nd , 2020, to January 16th 2023

B. COVID-19 Deaths Dataset

This covid deaths dataset contains the deaths for each county across the USA from January 22nd, 2020, to January 16th, 2023.

Name	Datatype	Description
countyFIPS	Integer(int64)	Unique five-digit number for each county
CountyName	Object	Name of the county
State	object	Name of the State.
StateFIPS	Integer((int64)	Unique number for each state.
Deaths (on each date)	Integer(int64)	Number of Deaths on each date from January 22nd , 2020, to January 16th 2023

C. County Population Dataset:

This county population dataset contains population in each county across the USA.

Name	Datatype	Description
countyFIPS	Integer(int64)	Unique five-digit number for each county
CountyName	Object	Name of the county
State	object	Name of the State.
population	Integer((int64)	Population in each county.

D. Enrichment -Datasets

Further in our work we used four different enrichment datasets namely Demographic,ACS Social, Economic and Housing, Employment Data set and Presidential Election Results (Political leanings) for understanding the covid-19 patterns and their spread.

1) Census Demographic ACS: This enrichment dataset contains all the demographic information of all states across USA by county level for the year 2021. It includes the estimates of population by age (different age groups), races (different races), sex and voting population. Originally this raw census demographic ACS dataset contains 715 columns, Out of all columns we used only required and useful columns for our analysis.

2) ACS Social, Economic, and Housing: For enrichment dataset we have used data from data.census.gov. It covers 100,000+ different geographies like states, counties, places, tribal areas, etc., and for each topics like education, employment, health, and housing are covered.

Enrichment data DP03 is selected for Economic characteristics. In this dataset stats of all the counties in US States along with their geographic ID. Count of population is given based on their employment status, occupation, commuting to work, industry they work, class of worker, income health insurance coverage.

3) Employment Dataset: The employment dataset provides the level of employment and the earning potential by Geographical impact. The Reason we have chosen employee data set to show the super Covid-19 data set to show the impact of covid_19 with respect to the employee enrichment data within each ”Area” and ”State”.

4) Presidential Election Results (Political leanings): The Presidential Elections data set was taken from the below mentioned site. This dataset contains the election information such as number of parties or candidates and who won of each county which was held on November 8th 2022.

III. DATA MODELING

After cleaning the data set, we merged the COVID cases and COVID deaths data sets into a single data frame to obtain all the details of cases and deaths in the USA by county level. We then merged the county population data frame with the above merged data frame to include the population of each county in our analysis.

Next, we converted the date column data type from object to datetime. This allowed us to perform analysis on the second half of the data, from June 1st, 2022, to December 31st, 2022. We separated this data from the original merged data frame and calculated the sum of cases and deaths grouping on date using the groupby function. This allowed us to see the total number of cases and deaths across the USA on each date.

To understand the daily trends of new cases and new deaths, we calculated the difference between consecutive dates using the diff() function. We removed the data related to May 29th, 2022, as it contained NaN.

We then converted the daily data into weekly data to see the weekly trends of new cases and new deaths across the USA. This allowed us to identify any patterns or trends in the data over a larger time frame.

With obtained weekly data of USA for the second half of the 2022 if we plot the weekly trends of the new cases across the USA using plotly express library it looked as follows:

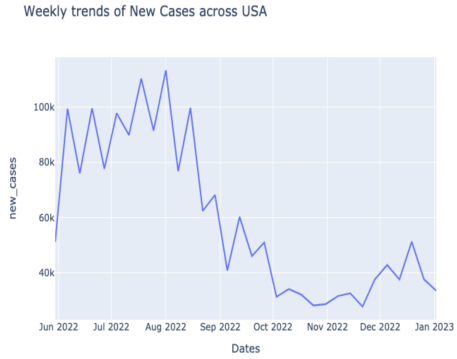

Fig. 1. Weekly trends of New Cases across USA

From the Fig 1 we can observe that there were peaks in the months between June to August. Thereafter, there is a downfall in between September-November. Again by the end of the November and in the month of December there are again small peaks in the new cases. Our team analysis on this peaks in new cases will be explained in the later section.

Now the weekly trends of the new deaths across the USA using plotly express library is as follows:

From the Fig 2 we can observe that there was a huge spike for new deaths in the month of July 2022. And there were minimal peaks in the remaining months in the second half of the year 2022.

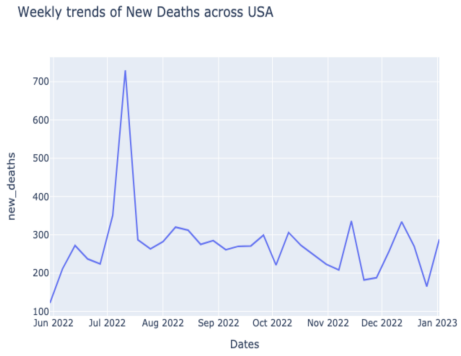

Fig. 2. Weekly trends of New Deaths across USA

Furthermore, we compared the USA data against three other countries of the world: Indonesia, Pakistan, and Nigeria. We selected these countries from the world data frame as their population is somewhat similar to that of the USA. We performed the same data cleaning steps and converted the daily data into weekly data for comparison purposes. This allowed us to understand how the USA's COVID-19 cases and deaths compare to other countries with similar population sizes.

Weekly trends plot of the new cases across the USA along with the selected three countries Indonesia, Nigeria and Pakistan is as follows:

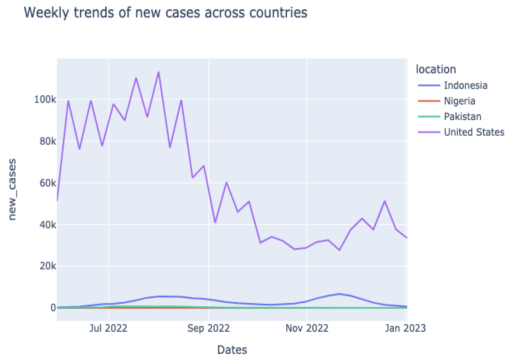

Fig. 3. Weekly trends of New Cases across selected countries

Weekly trends plot of the new deaths across the USA along with the selected three countries Indonesia, Nigeria and Pakistan is as follows:

From the Fig 3 and Fig 4 we can observe that when compared to other countries the new cases and new deaths registered across the USA are higher in the second half of the 2022 year. But if you observe the minimal peak months of new cases across the selected three countries are similar to the peak months in the USA. Coming to the deaths there were very less new deaths in all the selected three countries compared to the USA.

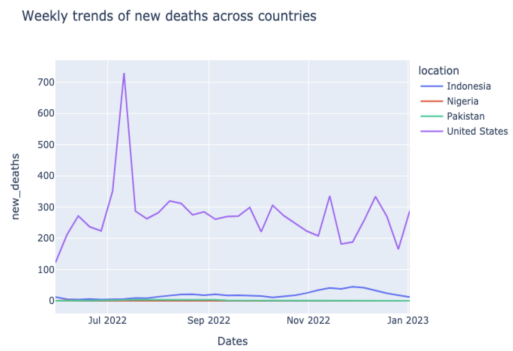

Fig. 4. Weekly trends of New Deaths across selected countries

Till now we have seen weekly trends of new cases and new deaths across the USA. In this project we also did same weekly trends analysis for individual states across USA. Repeated the same steps i.e., considering the state data from June 1st 2022 to December 31st 2022, converting the daily data into weekly data which are followed for obtaining the weekly trends across USA.

Weekly trends plot of the new cases across the California State obtained is as follows:

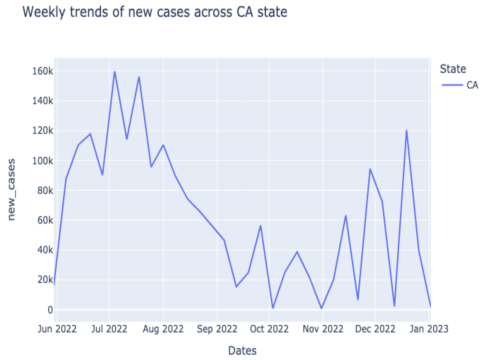

Fig. 5. Weekly trends of New Cases across California State

From the Fig 5 we can observe that the peak months of the new cases across the California State are same as the peak months of the new cases across the USA i.e., from June to August and again in the end of November and also in the month of December. So, from this we can understand that all the States in the USA are following the same pattern as the USA pattern.

Weekly trends plot of the new deaths across the California State obtained is as follows:

From the Fig 6 we can observe that the peak month of the new deaths across the California State are same as the peak month of the new deaths across USA i.e., in the month of July.

Weekly trends plot of the new cases across the top3 infected counties of California State obtained is as follows:

From the Fig 7 we can observe that the peaks of the new cases across the top3 infected counties across the California

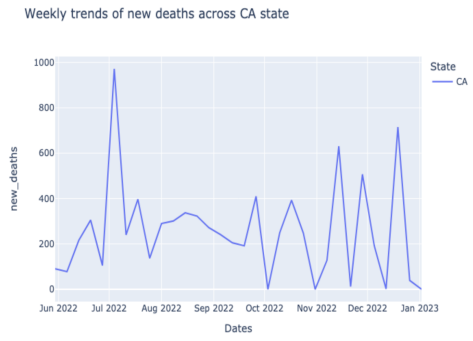

Fig. 6. Weekly trends of New Deaths across California State

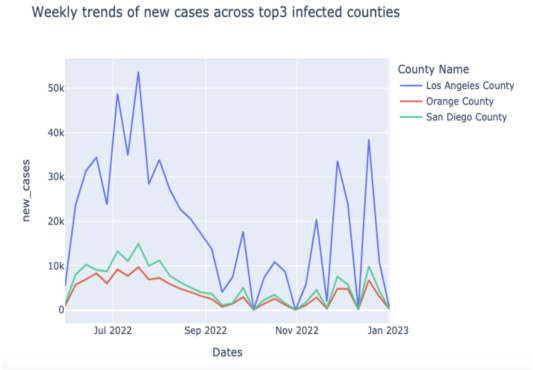

Fig. 7. Weekly trends of New Cases across top3 infected counties in California State

state are in the same months as the peaks months in the CA State. So, from this we can understand that all the counties in the California State are following the State pattern.

Overall, these steps were crucial in preparing the data for analysis and providing a better understanding of the COVID-19 situation in the USA.

Background Research:

Our research revealed that compared to the three selected countries, the United States of America had a higher number of new cases and new deaths in the second phase of 2022. We observed peaks in the months between June-August 2022, which we believe was due to the highly transmissible Omicron variant. Other factors that may have contributed to this increase include relatively low booster uptake compared to primary vaccination. We also noticed that there were no peaks in the months between September-November 2022, as the second booster was given to all age groups, particularly for the elder age group. As the cases and deaths in the elder age group were more in this second phase, this could explain the reduction in peaks during this period. However, there were small peaks in the months of November and December due to public gatherings that occurred during Thanksgiving and Christmas holidays. In terms of analyzing the data at a more granular level, we observed that California state followed the same pattern as the USA as a whole, with peaks in the months

between June-August 2022 and smaller peaks in November and December. Additionally, the top three counties in the USA also followed a similar pattern to California, with peaks during the summer months and smaller peaks during the holiday season. Overall, our research provides important insights into the COVID-19 situation in the USA and highlights the impact of the Omicron variant on the number of new cases and new deaths. By analyzing the data at different levels of granularity, we were able to identify patterns and trends that can inform public health policy and interventions to mitigate the spread of the virus.

IV. Distributions and Hypothesis

Distribution of a dataset is an important aspect of data analysis because it provides insight into the underlying patterns and characteristics of the data. By analyzing the distribution of a dataset, you can gain a better understanding of the shape, central tendency, variability, and outliers in the data. The distribution provides a summary of the data that can be used to compute descriptive statistics such as the mean, median, mode, variance, and standard deviation and can also be helpful in selecting an appropriate model for the data. Different types of distributions have different properties and can be used to model different types of data. Here we performed multiple probability distributions on our data and analysed which distribution is best fit by using distribution estimators such as Method of Moments, Maximum Likelihood and Kernel Density Estimators(KDE).

A. Gamma Distribution

Gamma distribution is one of the continuous probability distributions that is often used in data analysis to model positive, skewed data. It is a flexible distribution that can take on a wide range of shapes, from exponential to normal-like, depending on the values of its parameters. For the implementation of this distribution we used the ”scipy.stats” module. We then create a Gamma distribution object using the gamma function from the ”scipy.stats” module, passing in the parameters. We can then use the pdf/cdf method of the distribution object to calculate the probability density function or a cumulative distribution function at a given value which are new cases in our cases.

Now applying this gamma distribution to see if this distribution is a good fit on the California State data which was generated in the before step. For implementing gamma distribution using mom method we need to find out the mean and variance as MoM depends on the mean and variance. Using the mean and variance values calculated the alpha and beta parameters as they are essential parameters for the gamma distribution. Now plotting the distribution over the histogram of new cases across California state Using the obtained alpha and beta parameters

From the Fig 8 we can observe that gamma distribution using mom did not fit correctly for some part of the histogram on the new cases.

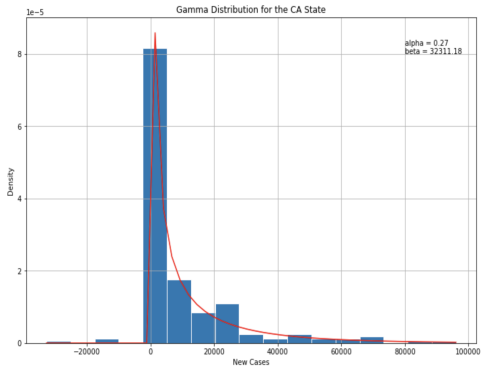

Fig. 8. Gamma Distribution for new cases across CA State

B. Poisson Distribution

The Poisson distribution is a discrete probability distribution that is often used to model the number of occurrences of an event in a fixed interval of time or space. With this we can assume that COVID-19 data is discrete event which occurs in a interval of time. So we have used the same module ”scipy.stats” for implementation of Poisson distribution. Here Poisson has a single parameter, lambda(mu), which represents the expected number of occurrences in the interval, which we have taken as the mean of the cases.

Now applying the Poisson distribution to see if this distribution is a good fit on the California State data. In Poisson distribution mean is equal to variance. Plotting the Poisson distribution on the new cases histogram using the mean value is as follows:

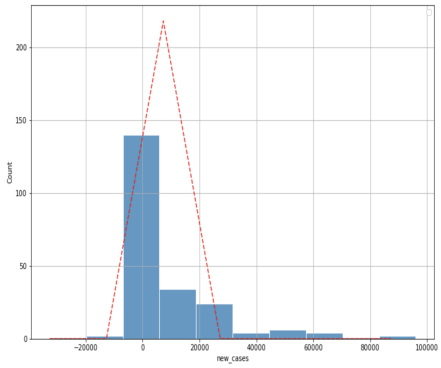

Fig. 9. Poisson Distribution for new cases across CA State

From the above Fig 9 we can observe that Poisson distribution is not so good fit for the California State data as the data is widely spread i.e., data is over-dispersed. If the data is widely spread then that Poisson distribution is not good.

C. Negative Binomial Distribution

The negative binomial distribution can be used in data analysis to model count data that have over dispersion, which means that the variance is greater than the mean. The negative binomial distribution models the probabilities for the rth

success occurring on the Nth trial when you know the event probability. Which is basically a discrete probability distribution that models the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified number of failures occurs. To use the negative binomial distribution in Python, we can use the ”scipy.stats” module.

Plotting the negative binomial distribution on the new cases histogram of the California State is as follows:

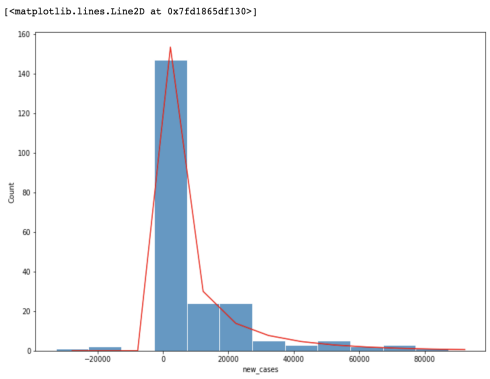

Fig. 10. Negative Distribution for new cases across CA State

From the above Fig 10 we can observe that negative binomial distribution using MoM fitted good for the California state new cases data compared to the poisson and gamma using MoM.

D. Kernel Density Estimation

Kernel Density Estimation is a non-parametric method of estimating the PDF function. It is non-parametric method because it does not assume any underlying distribution. In Kernel Density estimation we use the gaussian kernel. The advantage of Gaussian kernel is that it automatically determines the bandwidth which is the important parameter of the kernel density estimator. Now plotting the kernel density estimation using the scipy stats kde on the new cases histogram across the CA State.

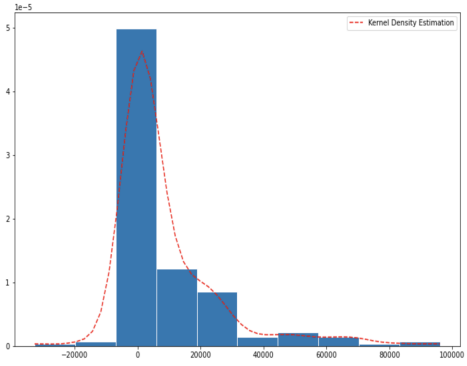

Fig. 11. Kernel Density Estimation for new cases across CA State

From the above Fig 11 we can observe that the Kernel Density Estimation method is a very good fit compared to the MoM method. Also if we try to adjust the bandwidth between 0.1-0.5 then we could get optimally smoothed kernel density estimate for the new cases data of CA State.

When the bandwidth is given as 0.3 the kernel density estimation obtained is as follows:

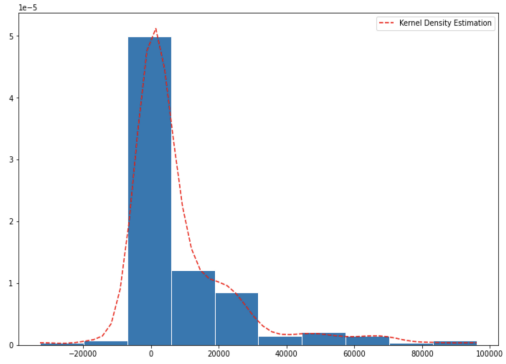

Fig. 12. Kernel Density Estimation for new cases across CA State with bandwidth 0.3

Apart from the California State we have tried to fit the above mentioned distributions to other selected states North Carolina, New Jersey and New york. The moment of distributions obtained for these selected states along with the California State is as follows:

	State	Mean	Variance	Skewness	Kurtosis
0	CA	8861.899083	285024960.7	2.374971	6.803399
1	NC	2623.298165	53952872.0	2.799712	6.647819
2	NJ	2380.316514	19482974.7	2.367865	6.013621
3	NY	4831.096330	45247698.0	2.720358	10.184112

Fig. 13. Moments of Distribution for selected states

From the above Fig 13 we can observe the moments of distribution for each state and we can see each state follows the same trends.Coming to the Skewness the values are almost same for each state and are positive that means each state data are right skewed. Right tails are more compared to the left tails.

Coming to the kurtosis the NY state has the highest kurtosis remaining states are almost similar. So the NY state data is highly tailed.

Distributions of each state looks different because each state registers different number of cases. Histogram spread depends on the number of new cases so the spread would be different from one state to another. So the distributions differs. Negative binomial distribution using MoM is more suitable for state whose data is more widely spread followed by the gamma distribution. If state data is not widely spread then the Poisson distribution will be a very good fit. Also if we observe the state data which we have with us it is a discrete data not continuous. So discrete distribution fits good for our data.

From the observations we can say that from these distributions, Kernel density estimation method works very good

compared to the method of moments method. We can get best kernel density estimate by giving the bandwidth parameter. If we select between 0.1-0.5 then we can see optimally smoothed kernel density estimate.

E. Correlation

In this step we took five-seven columns from each enrichment dataset and grouped them by state as per our comparison and then normalized per state population to understand correlation on normalised cases/deaths per state population. As we are four in number, each member used the columns from their respective enrichment dataset to obtain the correlation between the normalized enrichment variables per state population and the Normalized cases/deaths per state population.

F. Hypothesis

In this step we formulated the hypothesis between Enrichment datasets and number of cases and deaths to be compared against states. As we are four in number we used four enrichment datasets. The hypothesis for the Census Demographic ACS is mentioned below as an example:

1) Does the increase in white population increases in the spread of covid-19 cases?

2) Covid cases are more in the age group of 65 years and above

3) Covid cases spread is low in the Male population

4) Covid cases are less among the Asian population

V. Machine Learning Models

Machine Learning Algorithms are one of the widely used frameworks for the future analysis. Now in this section we will use Regression based models for predictions by fitting our COVID-19 data and training the model. Here we built basic linear and polynomial regression models and fit with our data and made the predictions for one week ahead.

Regression is a statistic tool which is used to model the relation between variables and make predictions to the dependant variable. So using our daily covid data as dependent data to train the model and predict the cases in future. We also fit our model to other selected countries COVID-19 data. We tested this model for US and other countries and also within US by considering few states and counties in that particular states respectively.

A. Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. Linear regression model will be highly biased if the input data is correlated. It works to identify the best-fit line or regression line that can predict the output variable given the input features under the assumption that there is a linear relationship between the input features and the output variable. So here the independent input variable will be the days from the start day of the infection/deaths and the covid-19 cases/deaths will be the dependent values on the dates. By deafult, the linear regression model has high bias and it has low variance, and this low variance will help if we have less input data. RMSE metric is used to evaluate our model. Linear Regression model equation is given as follows:

$y = b + ax$

Here y is the response

x is the feature

b is the intercept

a is the coefficient of x

Here a, b together are called model coefficients. Before creating the model we must learn the values of these model coefficients. Once after learning these coefficients are used to predict the best line of the data points.

In order to implement this linear regression model we imported LinearRegression from sklearn.linear_model. For calculating the root mean square error we imported the mean_squared_error from the sklearn.metrics and sqrt from the math library.

On implementing the linear regression model, the model predicted trend line and model predicted one week ahead forecast for new cases across the USA is as follows:

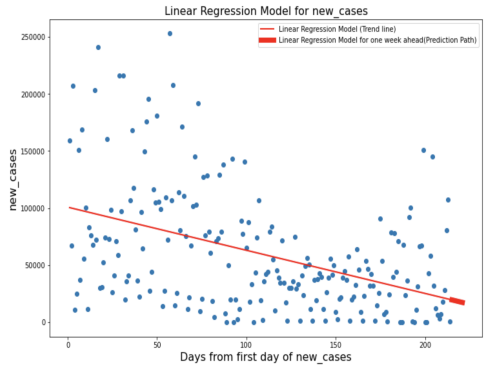

Fig. 14. Linear Regression Model for new cases across USA

Root Mean Squared error of Linear Regression Model for new cases across USA is 48442.23155959759.

B. Polynomial Regression

Polynomial regression model is the special case of the linear regression model which is often used to make predictions using polynomial powers of independent variables.It posits a non-linear relationship between the input features and the output variable and is an extension of linear regression. By using higher order components in the regression equation, the procedure in polynomial regression attempts to fit a polynomial curve rather than a straight line to the data.The data points are fitted to a polynomial equation of degree n in polynomial regression, where n is an integer greater than or equal to 2. The polynomial regression model equation is given as follows:

$y = b + a_1x + a_2x^2 + ... + a_nx^n$

where n is the degree of the polynomial.

In this project we implemented the polynomial regression model from degree 2 to the degree 5. The respective equations until degree 5 are as follows:

For,

$n = 2, y = b + a_1x + a_2x^2$

$n = 3, y = b + a_1x + a_2x^2 + a_3x^3$

$n = 4, y = b + a_1x + a_2x^2 + a_3x^3 + a_4x^4$

$n = 5, y = b + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5$

In order to implement this Polynomial regression model we imported PolynomialFeatures from sklearn.preprocessing.

On implementing the Polynomial Regression model with degree 2, the model predicted trend line and model predicted one week ahead forecast for new cases across the USA is as follows:

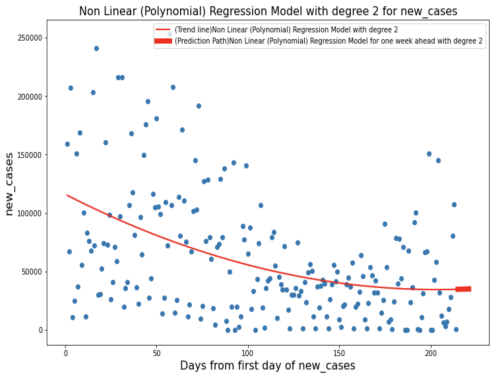

Fig. 15. Polynomial Regression Model with degree 2 for new cases across USA

Root Mean Squared error of Polynomial Regression Model with degree 2 for new cases across USA is 47975.931372924955.

The Polynomial Regression Model with degree 3 predicted trend line and the predicted path for one week ahead forecast for new cases across the USA is as follows:

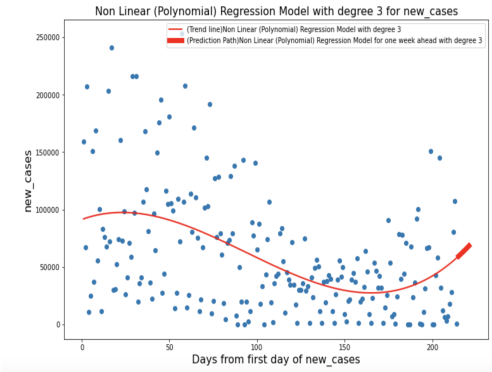

Fig. 16. Polynomial Regression Model with degree 3 for new cases across USA

Root Mean Squared error of Polynomial Regression Model with degree 3 for new cases across USA is 47114.51151060665.

The Polynomial Regression Model with degree 4 predicted trend line and the predicted path for one week ahead forecast for new cases across the USA is as follows:

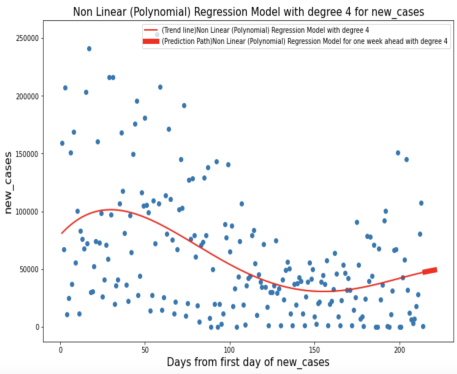

Fig. 17. Polynomial Regression Model with degree 4 for new cases across USA

Root Mean Squared error of Polynomial Regression Model with degree 4 for new cases across USA is 46962.83088865819.

The Polynomial Regression Model with degree 5 predicted trend line and the predicted path for one week ahead forecast for new cases across the USA is as follows:

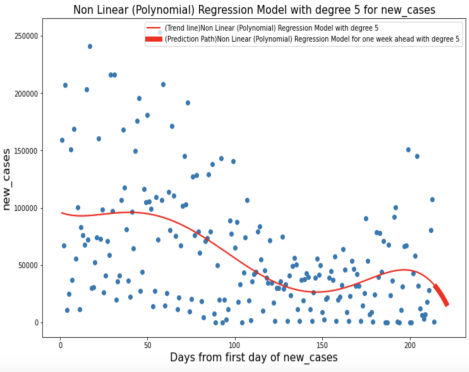

Fig. 18. Polynomial Regression Model with degree 5 for new cases across USA

Root Mean Squared error of Polynomial Regression Model with degree 5 for new cases across USA is 46730.42323778381.

From the above figures of linear regression model and non linear regression models for new cases across USA, we can observe that the linear regression model predicted trend line is highly biased. Coming to the polynomial regression model predicted trend lines as the degree increased the bias is getting reduced but the variance started increasing. The trend lines from polynomial regression model with degree 4 and 5 are some what over fitting to the data. We can say a model is a best fit for the data when it has moderate bias and moderate variance. So, Polynomial regression with degree 3 is the best fit for the new cases data across the USA.

Also, if we observe the root mean squared error for both linear regression model and polynomial regression the RMSE is high for the linear regression compared to the polynomial regression. Also for polynomial regression as the degree gets increased the root mean squared error is getting less.

The linear regression model predicted trend line and model predicted one week ahead forecast for new deaths across the USA is as follows:

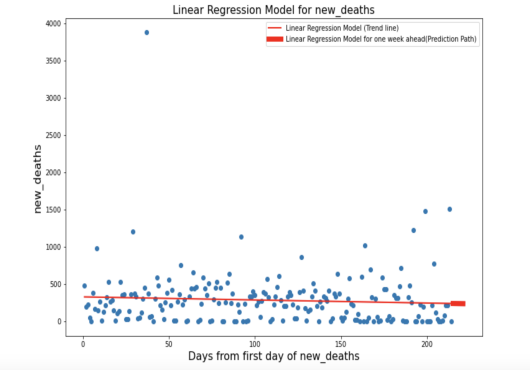

Fig. 19. Linear Regression Model for new deaths across USA

Root Mean Squared error of Linear Regression Model for new deaths across USA is 361.0230192808004.

The Polynomial Regression model with degree 2 predicted trend line and model predicted one week ahead forecast for new deaths across the USA is as follows:

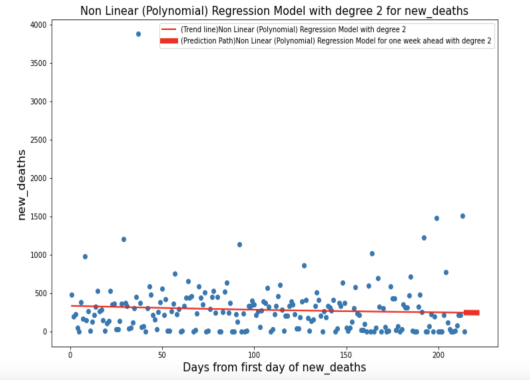

Fig. 20. Polynomial Regression Model with degree 2 for new deaths across USA

Root Mean Squared error of Polynomial Regression Model with degree 2 for new deaths across USA is 361.01114385674805 .

The Polynomial Regression Model with degree 3 predicted trend line and the predicted path for one week ahead forecast for new deaths across the USA is as follows:

Root Mean Squared error of Polynomial Regression Model with degree 3 for new deaths across USA is 360.10151947743867 .

The Polynomial Regression Model with degree 4 predicted trend line and the predicted path for one week ahead forecast for new deaths across the USA is as follows:

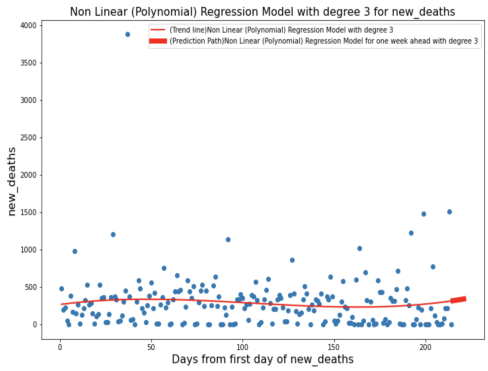

Fig. 21. Polynomial Regression Model with degree 3 for new deaths across USA

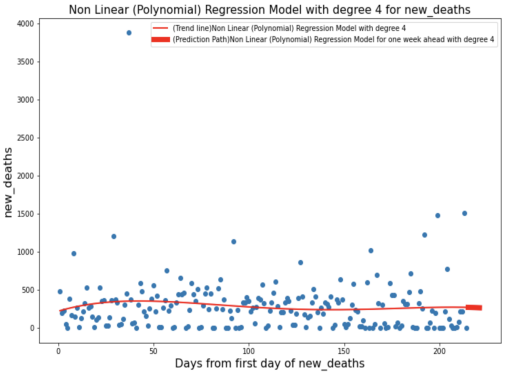

Fig. 22. Polynomial Regression Model with degree 4 for new deaths across USA

Root Mean Squared error of Polynomial Regression Model with degree 4 for new deaths across USA is 359.7999092590437.

The Polynomial Regression Model with degree 5 predicted trend line and the predicted path for one week ahead forecast for new deaths across the USA is as follows:

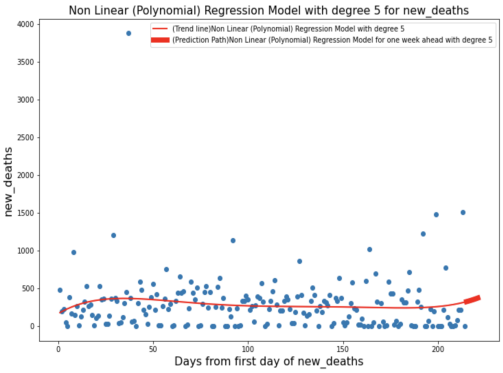

Fig. 23. Polynomial Regression Model with degree 5 for new deaths across USA

Root Mean Squared error of Polynomial Regression Model with degree 5 for new deaths across USA is 359.47356777308124.

For the new deaths data across the USA the polynomial regression model with degree 3 looks like a best fit.

In this we project we also build the both linear and polynomial regression models for three selected countries Indonesia, Nigeria and Pakistan. We build models for both new cases and new deaths across these selected countries. In this report we are keeping the information regarding the new cases as the new deaths were very less in these countries.

The linear regression model predicted trend line and model predicted one week ahead forecast for new cases across the Indonesia is as follows:

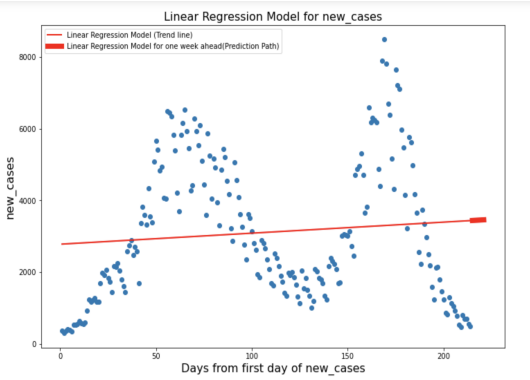

Fig. 24. Linear Regression Model for new cases across Indonesia

Root Mean Squared error of Linear Regression Model for new cases across Indonesia is 1901.240100901625.

The Polynomial Regression model with degree 2 predicted trend line and model predicted one week ahead forecast for new cases across the Indonesia is as follows:

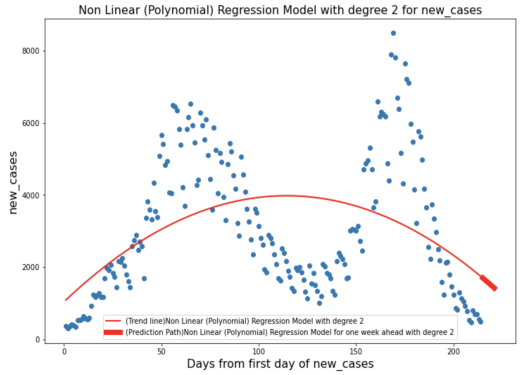

Fig. 25. Polynomial Regression Model with degree 2 for new cases across Indonesia

Root Mean Squared error of Polynomial Regression Model with degree 2 for new cases across Indonesia is 1739.0115716545006.

The Polynomial Regression model with degree 3 predicted trend line and model predicted one week ahead forecast for new cases across the Indonesia is as follows:

Root Mean Squared error of Polynomial Regression Model with degree 3 for new cases across Indonesia is 1722.6965267555518.

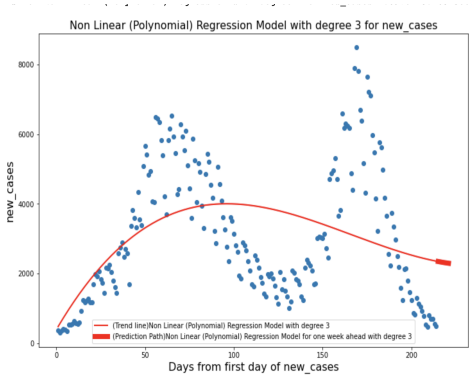

Fig. 26. Polynomial Regression Model with degree 3 for new cases across Indonesia

The Polynomial Regression model with degree 4 predicted trend line and model predicted one week ahead forecast for new cases across the Indonesia is as follows:

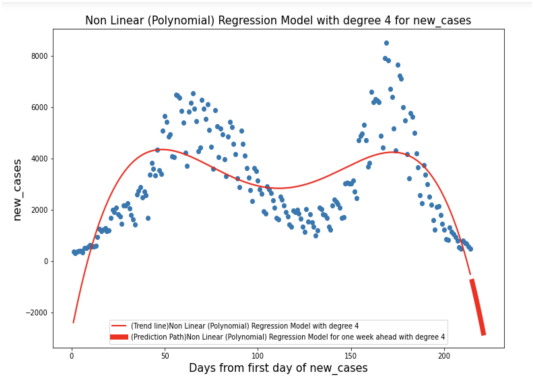

Fig. 27. Polynomial Regression Model with degree 4 for new cases across Indonesia

Root Mean Squared error of Polynomial Regression Model with degree 4 for new cases across Indonesia is 1401.9981528637072.

The Polynomial Regression model with degree 5 predicted trend line and model predicted one week ahead forecast for new cases across the Indonesia is as follows:

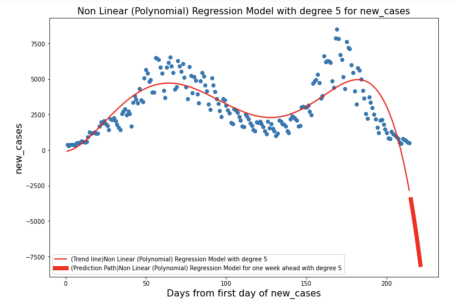

Fig. 28. Polynomial Regression Model with degree 5 for new cases across Indonesia

Root Mean Squared error of Polynomial Regression

Model with degree 5 for new cases across Indonesia is 1185.1229292368532.

The linear regression model predicted trend line and model predicted one week ahead forecast for new cases across the Nigeria is as follows:

Fig. 29. Linear Regression Model for new cases across Nigeria

Root Mean Squared error of Linear Regression Model for new cases across Nigeria is 113.53730794274986.

The Polynomial Regression model with degree 2 predicted trend line and model predicted one week ahead forecast for new cases across the Nigeria is as follows:

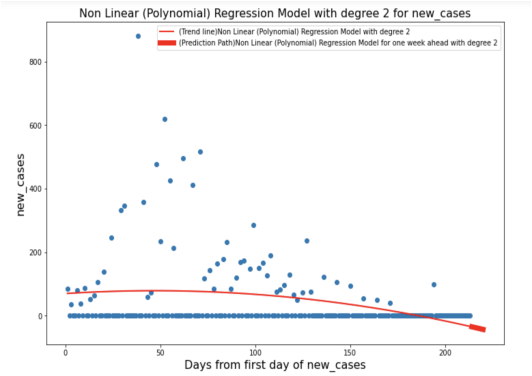

Fig. 30. Polynomial Regression Model with degree 2 for new cases across Nigeria

Root Mean Squared error of Polynomial Regression Model with degree 2 for new cases across Nigeria is 112.69803020347341.

The Polynomial Regression model with degree 3 predicted trend line and model predicted one week ahead forecast for new cases across the Nigeria is as follows:

Root Mean Squared error of Polynomial Regression Model with degree 3 for new cases across Nigeria is 110.46174893631874.

The Polynomial Regression model with degree 4 predicted trend line and model predicted one week ahead forecast for new cases across the Nigeria is as follows:

Root Mean Squared error of Polynomial Regression Model with degree 4 for new cases across Nigeria is 110.16415139460786.

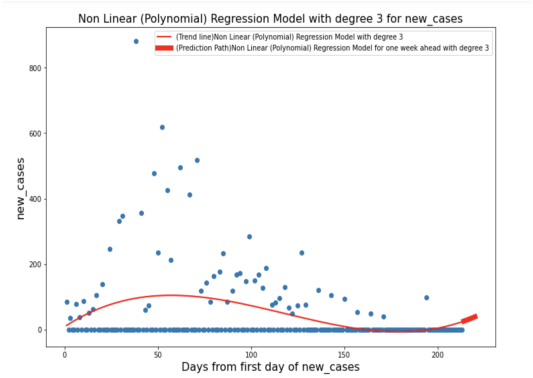

Fig. 31. Polynomial Regression Model with degree 3 for new cases across Nigeria

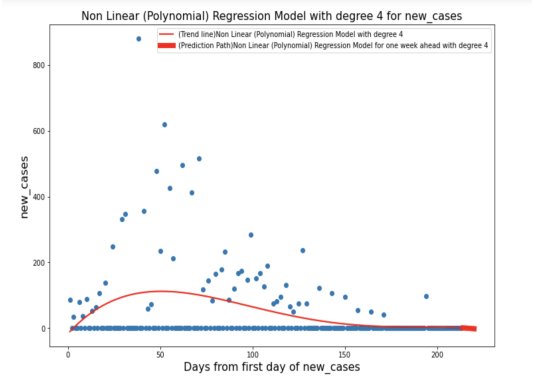

Fig. 32. Polynomial Regression Model with degree 4 for new cases across Nigeria

The Polynomial Regression model with degree 5 predicted trend line and model predicted one week ahead forecast for new cases across the Nigeria is as follows:

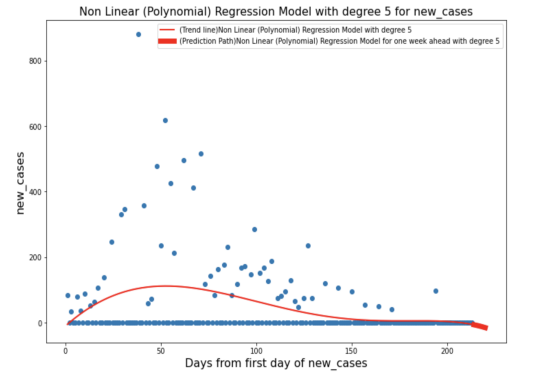

Fig. 33. Polynomial Regression Model with degree 5 for new cases across Nigeria

Root Mean Squared error of Polynomial Regression Model with degree 5 for new cases across Nigeria is 110.14889636025761.

The linear regression model predicted trend line and model predicted one week ahead forecast for new cases across the

Pakistan is as follows:

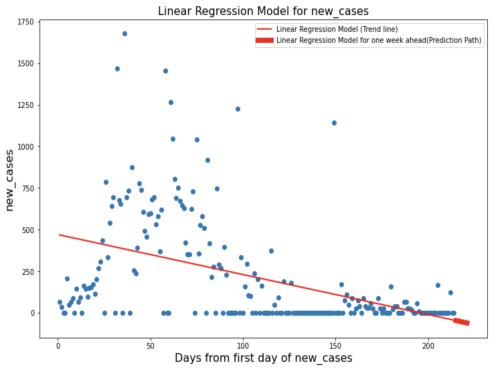

Fig. 34. Linear Regression Model for new cases across Pakistan

Root Mean Squared error of Linear Regression Model for new cases across Pakistan is 287.843090167427.

The Polynomial Regression model with degree 2 predicted trend line and model predicted one week ahead forecast for new cases across the Pakistan is as follows:

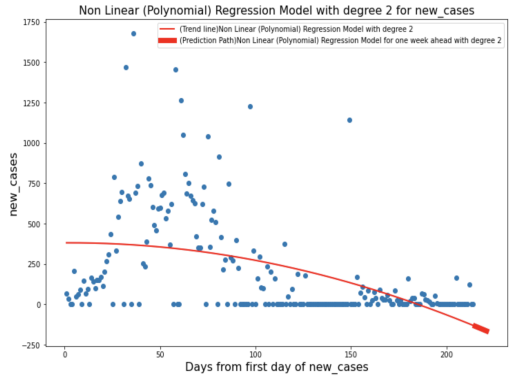

Fig. 35. Polynomial Regression Model with degree 2 for new cases across Pakistan

Root Mean Squared error of Polynomial Regression Model with degree 2 for new cases across Pakistan is 285.17491553272765.

The Polynomial Regression model with degree 3 predicted trend line and model predicted one week ahead forecast for new cases across the Pakistan is as follows:

Root Mean Squared error of Polynomial Regression Model with degree 3 for new cases across Pakistan is 253.17193539485368.

The Polynomial Regression model with degree 4 predicted trend line and model predicted one week ahead forecast for new cases across the Pakistan is as follows:

Root Mean Squared error of Polynomial Regression Model with degree 4 for new cases across Pakistan is 238.0998248072289.

The Polynomial Regression model with degree 5 predicted trend line and model predicted one week ahead forecast for new cases across the Pakistan is as follows:

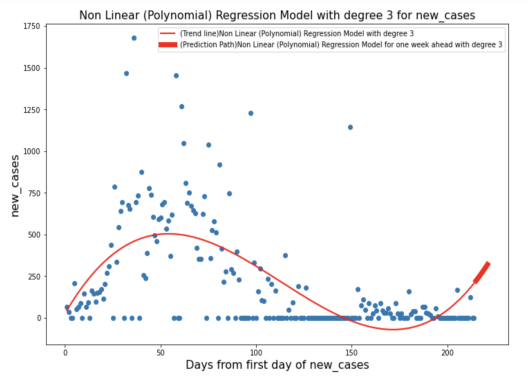

Fig. 36. Polynomial Regression Model with degree 3 for new cases across Pakistan

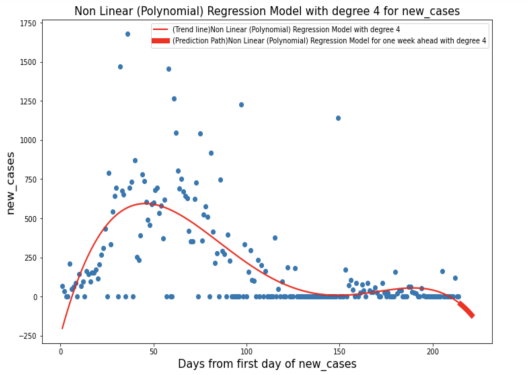

Fig. 37. Polynomial Regression Model with degree 4 for new cases across Pakistan

Root Mean Squared error of Polynomial Regression Model with degree 5 for new cases across Pakistan is 237.84179724725666.

VI. VISUALIZATION USING DASHBOARD

In this project, we developed an interactive dashboard using the JupyterDash and plotly as the visualization is the best way to express the results so that it makes our findings more approachable and more understandable by the users.

Generally the Jupyter Dashboard contains two main things one is app layout and other is app callback. In app layout we define the structure of the dashboard like what all input components should be available for the users to select and also the output components where the users can see the results for their inputs.

In order to achieve this dash board we are using dash core components of plotly. We added few options to plot trend and that can be changed by user like selecting the specific timeline to see the trend (by default it is limited from June 1st 2022 to Dec 31st 2022 as we performed analysis on the second half of the year 2022), dropdown where cases/deaths input parameter can be selected (by default cases will be selected), dropdown for selection of states (By default we can see the new cases trend for all USA data), also added linear, log-normalized option to entire data, we can also project a linear/non-linear

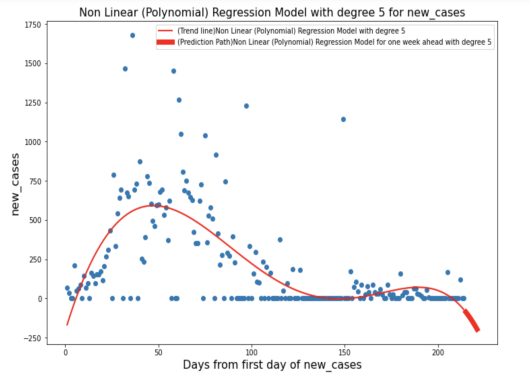

Fig. 38. Polynomial Regression Model with degree 5 for new cases across Pakistan

regression model trend lines to the data and also a 7-day moving average on the data. Below are some of the insights of the developed dashboard.

Fig. 39. Default appearance of the dashboard

If we select the date selection from 15th June 2022 to 15th September 2022 and if cases is selected then the plot is as follows:

Fig. 40. Plot of new cases across USA for selected date range

Now if we select log normalized radio button for above date range and cases the plot is as follows:

Fig. 41. Plot of log normalized new cases across USA for selected date range

Again if we select the non-linear regression model with degree 3 from the dropdown of regression along with the above selections then the obtained plot is as follows:

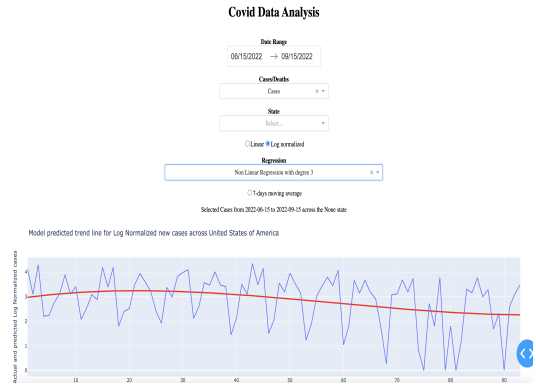

Fig. 42. Non-linear Regression Model with degree 3 for USA

Now instead of selecting regression if we select 7 day moving average on the above selected data then the obtained plot is as follows:

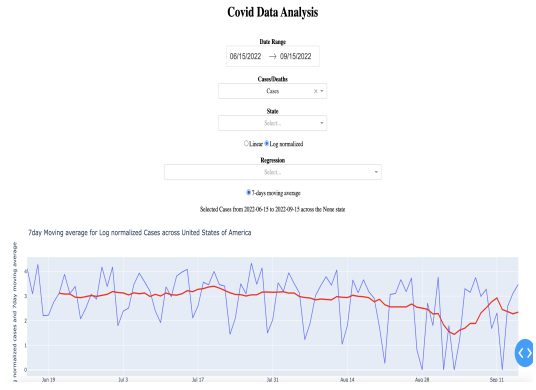

Fig. 43. 7day moving average for the above selected data

Now on this selection if we choose select multiple states i.e., CA and NY then the 7 day moving average plot for these states looks as follows:

Fig. 44. 7day moving average for the selected states

Selecting Non-linear regression model with degree 3 instead of the 7 day moving average for the selected states then the obtained plot is as follows:

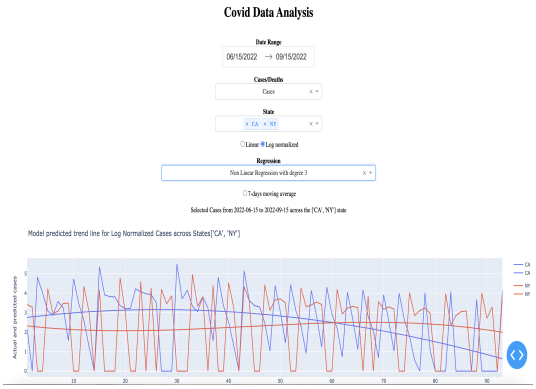

Fig. 45. Non-linear Regression Model with degree 3 for selected states

VII. Conclusion

In this project we have combined the data of covid cases, covid deaths and population of the USA into one to understand how USA got affected especially in the second half of the year 2022 i.e., from June 1st 2022 to December 31st 2022. With the combined data we tried to plot the weekly trends using the plolty express library. Along with the USA we also plotted the weekly trends with other countries whose population is similar to the USA. On comparing with those selected countries the cases and deaths in the USA are more compared to those countries in the second phase of the year 2022. Also, as we are team of four each member took four different states to see the weekly trends in those states and checked whether the patterns of states in these phase are matching with the countries trend. Also each member took different enrichment dataset to understand the patterns of spread of covid-19. Each member formed there own hypothesis with the help of the enrichment datasets. Also, we build the Linear and Non-Linear regression models for the USA data to predict the cases and deaths in the future and also to see the trends of cases and deaths across the states as well as the entire USA.

In terms of our observations, the cases and deaths are more in the months from June-August as there was effect of transmissible omicron variant and also immunity of the people from the covid-19 got lesser in this period. Especially in the USA the deaths in the age group of 65 years and above are more due to lack of immunity from the virus. Later after the second booster was given to people there was decrease in cases and deaths across USA. Again at the end of November and in the month of December there were again some peaks due to the public gatherings and holidays. States like California, New York, New Jersey, Florida and Texas registered lot of cases and deaths in the second phase of the year 2022.

Some inferences that can be made from this analysis of COVID data, such as people of all ages have been impacted by COVID-19, although the elderly and those with underlying medical issues are more at risk of developing serious sickness or passing away. It has been shown that a variety of measures, including social isolation, mask use, and immunization, are successful in slowing the development of COVID-19. Due to decreased demand and operational limitations brought on by COVID-19, many enterprises are finding it difficult to continue in business.

References

[1] https://covid19.who.int/
[2] https://covid.cdc.gov/covid-data-tracker/datatracker-home
[3] https://www.askpython.com/python/examples/polynomial-regression-in-python
[4] https://www.analyticsvidhya.com/blog/2020/03/polynomial-regression-python/
[5] https://dash.plotly.com/dash-core-components/graph
[6] https://plotly.com/python/
[7] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.moment.html
[8] https://github.com/c-zhu3/CSC405-605-Spring2023-Project/tree/master/Lectures/
[9] https://github.com/q-tong/CS405-605-Data-Science/tree/main/lecture
[10] https://www.statology.org/plot-confidence-interval-python/
[11] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rolling.html