CSC 605/705 01 Group 1 Paper

Kevin Hayes
University of North Carolina at Greensboro.
Spring Garden St, North Carolina, 27412
Email: khhayes2@uncg.edu

Ayodeji Iwayemi
University of North Carolina at Greensboro.
Spring Garden St, North Carolina, 27412
Email: a_iwayemi@uncg.edu

Neetha Ravva
University of North Carolina at Greensboro.
Spring Garden St, North Carolina, 27412
n_ravva@uncg.edu

***Abstract*—Use 9 point Times New Roman Bold for the abstract. Set your line spacing to be 10 points rather than single space. Indent the first line by 0.125 inches and type the word "Abstract" in 9 point Times New Roman Bold Italic. This should be followed by two spaces, a long dash (option / shift / minus), two spaces, and then the first word of your abstract (as shown above). A more professional look will result if all the spaces are set to a font style of regular rather than bold. Times font is an acceptable substitute for Times New Roman font. After the abstract, you should list a few key words from the IEEE approved "Index Terms" (send email to keywords@ieee.org for the latest list) that describe your paper. The index terms are used by automated IEEE search engines to quickly locate your paper. Typically, you should list about 5 to 7 key words, in alphabetical order, using 9 point Times New Roman Bold font. An example is shown next.**

Ceramics, coaxial resonators, delay filters, delay-lines, power amplifiers.

I. Introduction

The COVID-19 pandemic has drastically affected daily life globally, with significant repercussions for countless individuals and economies, particularly in the United States where it has led to numerous cases and fatalities. To analyze its impacts more thoroughly, we will integrate multiple datasets. Primary data will come from usfacts.org, featuring a daily tracker of COVID-19 cases, deaths, and population metrics at the county level. This will be elaborated upon in Stage I of our analysis. Additionally, we will utilize three enrichment datasets related to COVID-19, accessible via the following links:

- **Employment Dataset** (BLS Website: `https://www.bls.gov/cew/downloadable-data-files.html`): Provides detailed employment statistics and earning potentials by county. This dataset helps explore correlations between employment patterns and COVID-19 infection rates, offering insights into socio-economic influences on the pandemic's spread across regions.
- **Presidential Election Results Dataset** (Kaggle: `https://www.kaggle.com/unanimad/us-election-2020`): Contains the results of the 2020 presidential election by county, showing which candidate won and the margin of victory. This dataset allows examination of how political affiliations correlate with responses to the COVID-19 pandemic, potentially revealing regional disparities in healthcare management and policy compliance.
- **Census Demographic ACS Dataset** (Census Bureau: `https://data.census.gov/cedsci/table?q=dp&tid=ACSDP1Y2018.DP05`): Includes extensive demographic information with population breakdowns by age groups for each county. This analysis identifies age-related vulnerabilities to COVID-19, aiding in the development of targeted healthcare interventions and efficient resource distribution.

II. Stage I((Data and Project Understanding)

Stage I focused on establishing a strong foundation for our COVID-19 analysis by thoroughly understanding the datasets available. Our team began by acquiring the primary dataset from usafacts.org, which includes a daily county-level tracker of COVID-19 cases, deaths, population , employement and census metrics. This dataset's detail allows for an in-depth analysis of infection rates per 100,000 people, essential for understanding regional pandemic trends. Initially, comprehending the dataset's structure and variables was crucial to analyzing the pandemic's regional dynamics.

During this phase, we conducted data cleaning operations that included removing duplicate entries, addressing missing values, standardizing formats, and combining the separate datasets on COVID-19 cases, deaths, and population data into a unified and well-organized dataset named SuperCovid19.

To enhance our analysis and expand the COVID-19 dataset, every team member explored different enrichment datasets independently. This included datasets such as the Census Demographic ACS, the Employment Dataset, and the Presidential Election Results dataset.Our team's responsibility extended beyond simply understanding each dataset on its own; we also needed to consider how to integrate these with the main COVID-19 dataset effectively. We focused on identifying shared variables and potential points of connection between the datasets to facilitate this merger. Additionally, our goal was to explain how these auxiliary datasets could enhance our analysis of the COVID-19 spread. We formulated initial

hypotheses, exploring potential links between socio-economic factors, political preferences, and the progression of the pandemic. Below are the data dictionaries for cases and deaths and population.

TABLE I
DESCRIPTION OF COVID19 CASES DATASET

Name	Definition	Data Type	Possible Values	Req.?
CountyFIPS	Unique County ID (unknown counties have an ID of 0)	Integer	1001, 1003, etc.	YES
County	County Name	Text	Crook, Vilas, etc.	YES
State	State Name	Text	TX, PA, WI, etc.	YES
StateFIPS	State FIPS Code	Text	01, 02, etc.	YES
Date	Cases per day from 1/22/2020 to 7/23/2023	Integer	0 to 5	YES

TABLE II
DESCRIPTION OF COVID19 DEATHS DATASET

Name	Definition	Data Type	Possible Values	Req.?
countyFIPS	Unique County ID (unknown counties have an ID of 0)	Integer	0, 1001, 1003, etc.	Yes
County	Name of County	Text	Wake, Orange, etc.	Yes
State	Name of State	Text	TX, PA, WI, etc.	Yes
StateFIPS	State ID	Text	01, 02, 10, etc.	Yes
Deaths	Number of Death per day from 1/22/2020 to 7/23/2023	Integer	0, 26, 1310, etc.	Yes

TABLE III
DESCRIPTION OF POPULATION DATASET

Name	Definition	Data Type	Possible Values	Req.?
County FIPS	The code for the counties	Integer	1200, 1133, etc.	Yes
County Name	Name of the County	Text	Telon County, etc.	Yes
State	Shorthand notation of the US State names	Text	TX, VT, NC, etc.	Yes
Population	Number of people	Integer	67892, 99887, etc.	Yes

A. *State-wise Insights: Analyzing COVID-19 Trends*

We conducted a detailed analysis of last week's COVID-19 trends across different states, with each member focusing on a specific state. This approach allowed us to assess whether cases were rising, falling, or stabilizing, providing crucial insights into regional differences in the pandemic's progression. Additionally, this analysis revealed that the recorded cases exhibit stable peaks, along with potential influencing factors that are distinct for each state. These insights on a state-by-state basis were vital in understanding the localized effects and set the stage for focused investigations in the later phases of our project.

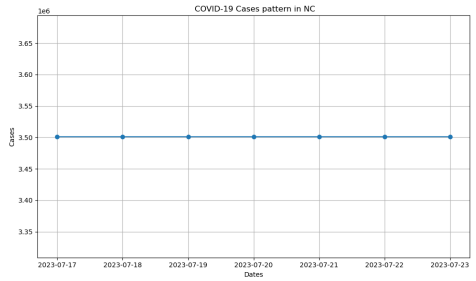

Fig. 1. Covid 19 cases trends for NC state.

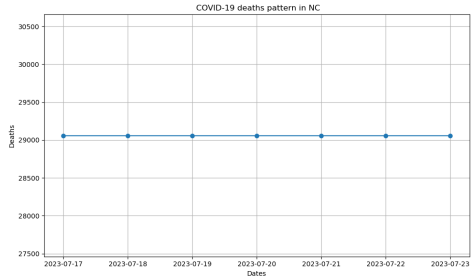

Fig. 2. Covid 19 deaths trends for NC state.

III. STAGE II

During the second phase of our analysis, the team conducted an extensive comparative study of COVID-19 trends within the United States and in selected countries around the world. We used weekly data on case and death counts to gain insights into the global trajectory of the pandemic. By carefully examining the mean, median, and mode values of these counts within the U.S., we highlighted how the pandemic has progressed over time.

The accompanying graphs depict the monthly averages of COVID-19 cases and deaths, demonstrating a direct proportionality between time and increases in both metrics. Notably, peaks in both cases and deaths are observed in December.

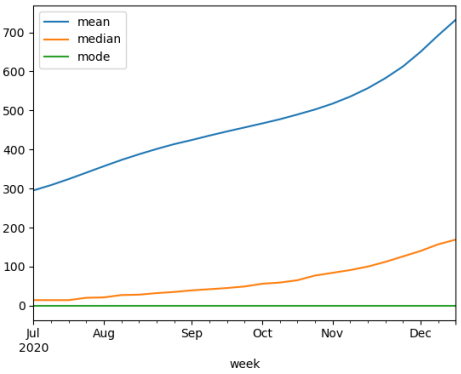

Fig. 3. Weelky statistics of covid 19 cases.

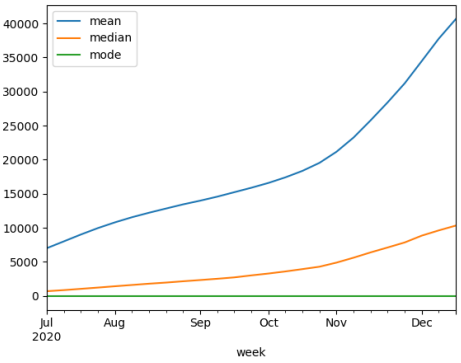

Fig. 4. Weelky statistics of covid 19 deaths.

Expanding our analysis beyond national borders, we embarked on comparing these trends against data from five countries with similar population demographics. Utilizing data from Our World in Data, we plotted and examined weekly trends for cases and deaths, employing various normalization techniques. By aggregating, normalizing by population, calculating differences in cases, and employing log normalization, we aimed to elucidate nuanced patterns and variations in the spread of COVID-19 across different regions.

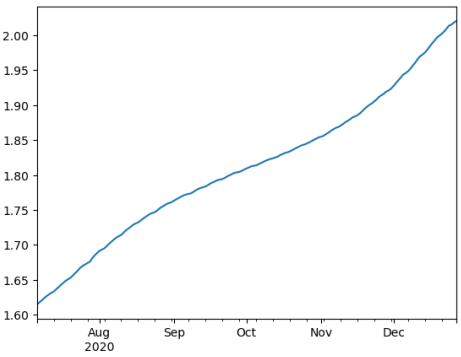

Fig. 5. Cases normalized by population

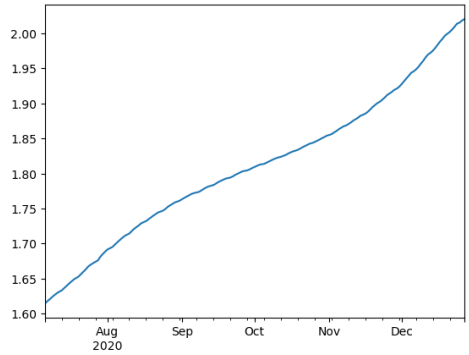

Fig. 6. Deaths normalized by population

The below statistics provide insights into the distribution and central tendency of COVID-19 cases and deaths in several countries: Japan (JPN), Brazil (BRA), Germany (DEU), Canada (CAN), China (CHN), and Russia (RUS).

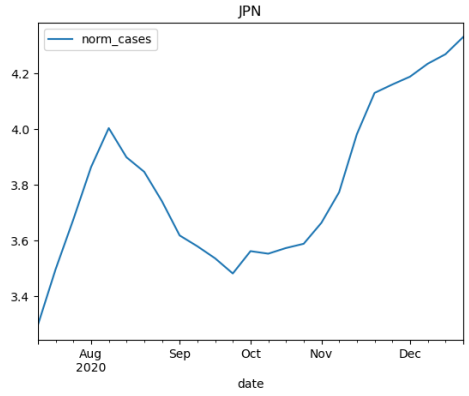

Fig. 7. Covid 19 cases in Japan

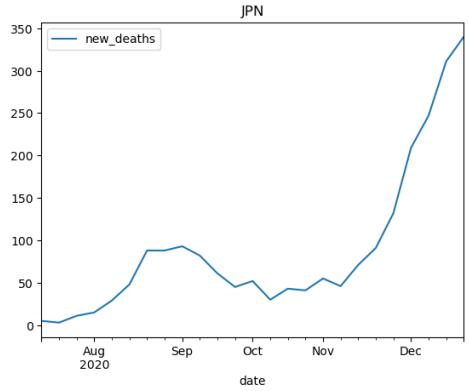

Fig. 8. Covid 19 cases in Japan

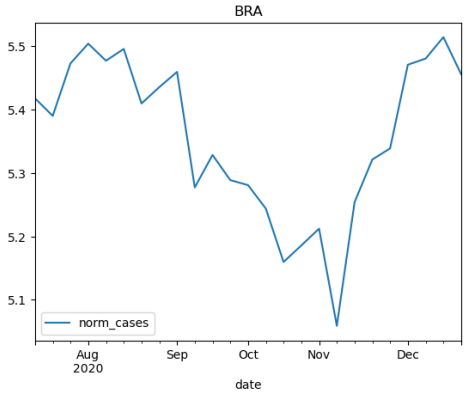

Fig. 9. Covid 19 cases in Brazil

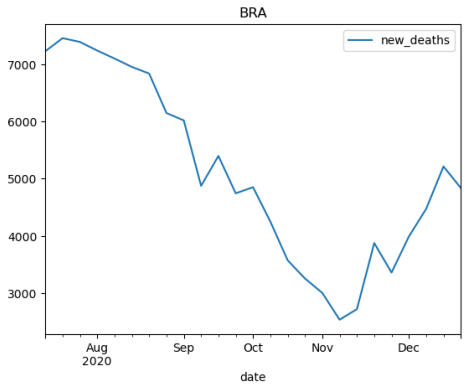

Fig. 10. Covid 19 deaths in Brazil

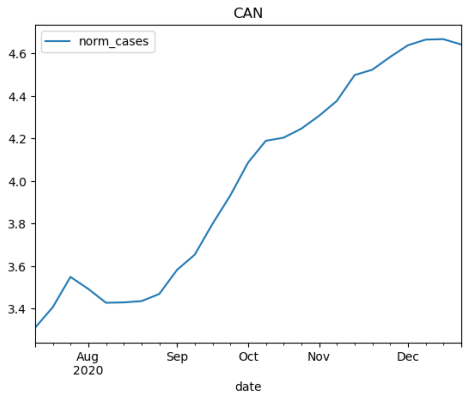

Fig. 11. Covid 19 cases in canada

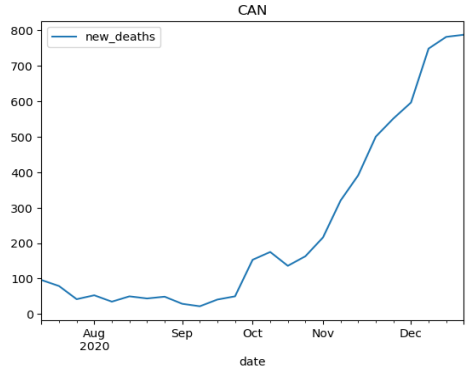

Fig. 12. Covid 19 cases in canada

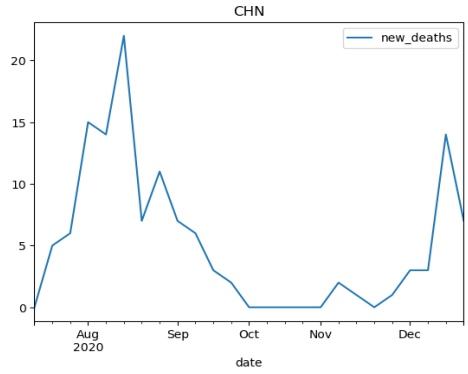

Fig. 13. Covid 19 cases in China

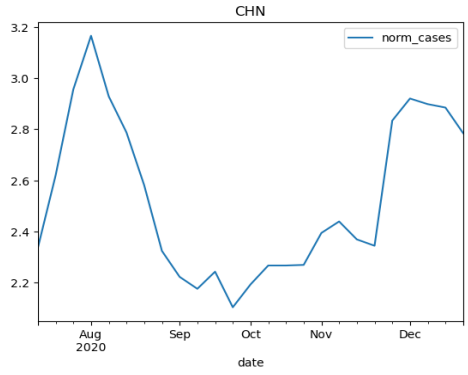

Fig. 14. Covid 19 deaths in China

The mean represents the average number of new cases or deaths reported daily, illustrating the typical magnitude or severity of the outbreak within each country. The median and mode being 0.0 for both cases and deaths indicate that there were many days with no new reported cases or deaths, highlighting days of low transmission or effective control measures. This comprehensive analysis provides critical insights into the pandemic's progression across different regions.

Peak COVID-19 Related Deaths Based on the data, the peak week for COVID-19 related deaths in the USA is identified:

Peak Week: December 21, 2020, with a total of 2,338,334 deaths.

Potential Reasons for the Peak in Deaths

- **Holiday Season:** The convergence of Christmas and New Year's Eve increases social gatherings and travel, potentially elevating virus transmission rates.
- **Winter Weather:** Colder weather prompts more indoor gatherings, heightening the risk of virus spread.

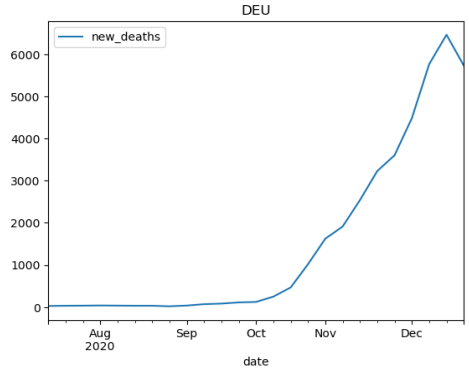

Fig. 15. Covid 19 cases in germany

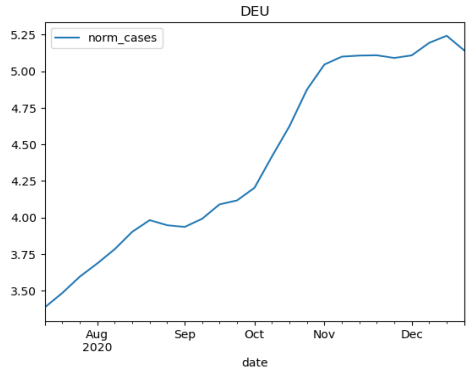

Fig. 16. Covid 19 deaths in germany

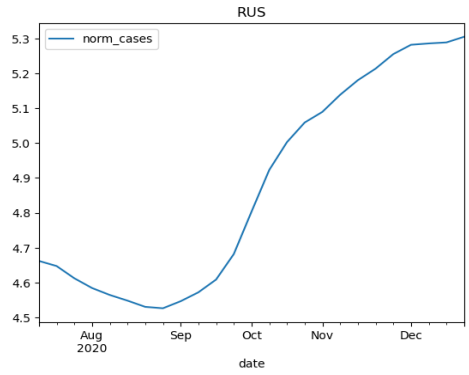

Fig. 17. Covid 19 cases in Russia

- **Thanksgiving Effects:** Activities and gatherings during Thanksgiving likely increased virus transmission.
- **Delayed Reporting:** Potential delays in data reporting could artificially create peaks by accumulating death reports.

Peak COVID-19 Related Cases Analysis shows that the peak week for new COVID-19 cases in the USA is the same as that for deaths:

Peak Week: December 21, 2020, with a total of 129,737,094 new cases.

Potential Reasons for the Peak in Cases

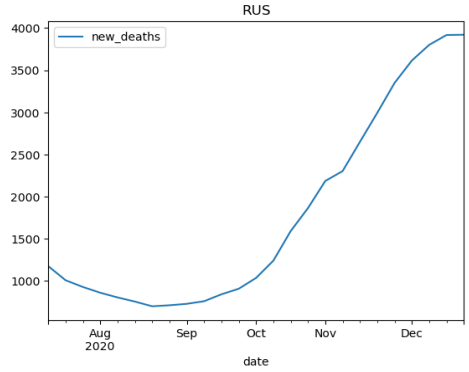

Fig. 18. Covid 19 deaths in Russia

- **Holiday Season:** Increased travel and gatherings during the holiday season may lead to a surge in new cases.
- **Winter Weather:** The cold weather increases indoor activities where the virus spreads more readily.
- **Thanksgiving Effects:** The Thanksgiving holiday period may have contributed to a rise in cases, with effects persisting into December.
- **Delayed Reporting:** Delays in the reporting process can result in peaks appearing more pronounced due to accumulated case reports.

Conclusion The peak in COVID-19 cases and deaths around December 21, 2020, appears to be influenced by seasonal activities, weather conditions, and reporting delays. Detailed analysis and further research are necessary to fully understand these dynamics and to better prepare for future outbreaks.

For the Stage II member task, each team member chose a particular state and produced weekly statistics for that state. The following plot represents data from the state of North Carolina. In North Carolina, the progression of COVID-19 showed a consistent increase in the number of cases and deaths throughout the period under review. From June 30 to July 6, 2020, the reported cases averaged 699.67, alongside 321 deaths. This trend of rising figures continued in the following weeks. By the week of December 29, 2020, to January 4, 2021, the average cases had surged to 5248.91, with 2722 deaths recorded. This data indicates a steady escalation in COVID-19 cases and fatalities, highlighting the virus's unrelenting spread across the state during this time. These increasing figures emphasize the need for continued vigilance, robust public health strategies, and effective interventions to mitigate the virus's spread and impact in North Carolina.

Weekly Analysis across states: We carried out an in-depth analysis that compared weekly data on COVID-19 cases and deaths across a selected state and five additional states. To ensure fair comparisons, we normalized these statistics by the population, calculating rates per 10,000 or 100,000 individuals. The line graph plotting these normalized values week by week for each state provided fascinating insights into their different COVID-19 trajectories. These variations in infection rates can be attributed to a range of factors, including different public

health policies, population densities, healthcare systems, and compliance with preventive measures. Identifying peak periods allowed us to align these with the broader national trends, revealing any similarities or differences in the timing and severity of outbreaks.

Over time, cases escalated, reaching a climax in November and December due to factors such as holiday gatherings, colder weather, and the emergence of new variants. The direct impact of population size on the case and death numbers in each state demonstrated a significant correlation.

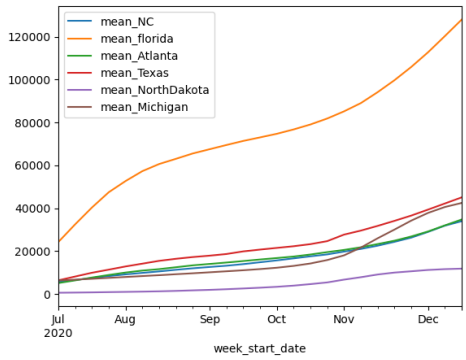

Fig. 19. Weekly statistics across the countries.

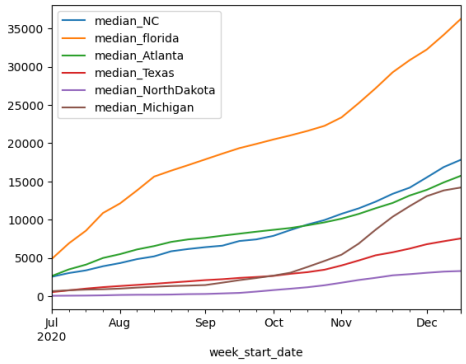

Fig. 20. Weekly statistics across the countries.

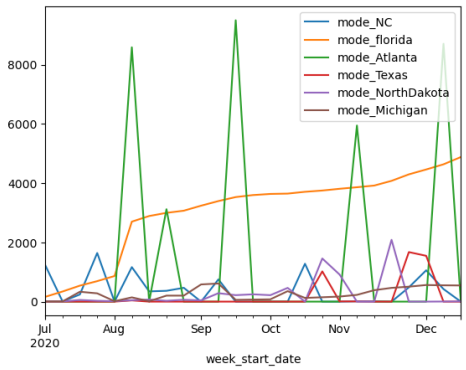

Fig. 21. Weekly statistics across the countries.

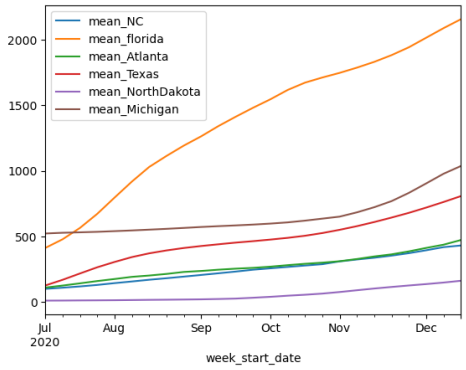

Fig. 22. Weekly statistics across the countries.

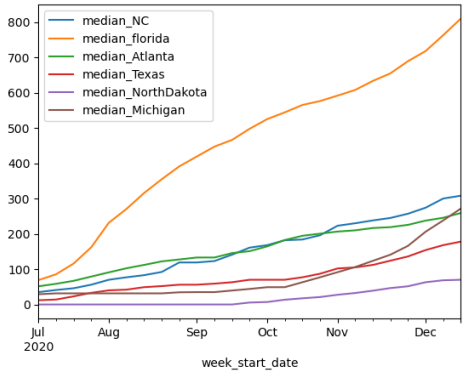

Fig. 23. Weekly statistics across the countries.

State-Specific Findings A key observation regarding differing rates across states is detailed below:

- Low deaths were recorded for North Dakota, whereas the state of Florida recorded the highest number of deaths in the last analyzed week.
- The highest and lowest cases and deaths are directly proportional to the population sizes of the states, indicating that Florida (being the most populated among the studied states) recorded the highest numbers, while North Dakota (with the smallest population) recorded the lowest.

County-Level Analysis As our analysis progressed, we concentrated on five counties within a designated state that showed particularly high COVID-19 case and death rates. We mapped out weekly trends for these most impacted counties, enabling us to pinpoint distinct patterns and characteristics specific to these areas. The visual representations included both raw and log-normalized data, which were instrumental in clarifying the progression of infections

I can observe that the counties pretty much follow the same pattern as their respective state does. The cases and deaths are higher in count in the week (end of december) for all the counties, the same is observed in the case of the North Carolina state as well.

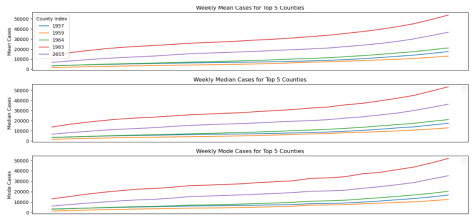

Fig. 24. County wise analysis.

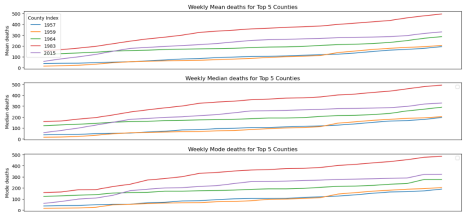

Fig. 25. County wise analysis.

IV. Stage III

A. Ayodeji Iwayemi

1) Distribution Analysis of New York State COVID-19 Cases: A graphical plot of the distribution of COVID-19 cases in New York State is provided in Figure 26. Descriptive statistics for New York's weekly COVID-19 cases are presented:

- Mean: Approximately $523,370$ cases
- Median: $460,987.5$ cases
- Standard Deviation: Approximately $144,267.094$
- Variance: Approximately $20,812,994,547.344677$
- Skewness: Approximately 1.527 (positively skewed)
- Kurtosis: Approximately 1.267
- The distribution was identified as unimodal based on graphical inspection and the presence of a single peak.

2) Comparison of COVID-19 Case Distributions Across States:

- New York exhibits the highest positive skewness, followed by Ohio, Kansas, North Carolina, Texas, and Florida.
- Higher skewness values indicate more pronounced skewness towards higher case numbers.

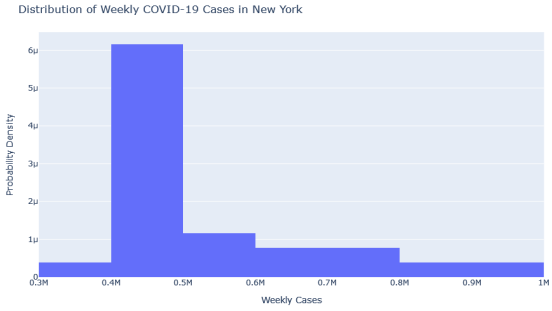

Fig. 26. Distribution of COVID-19 cases in New York State

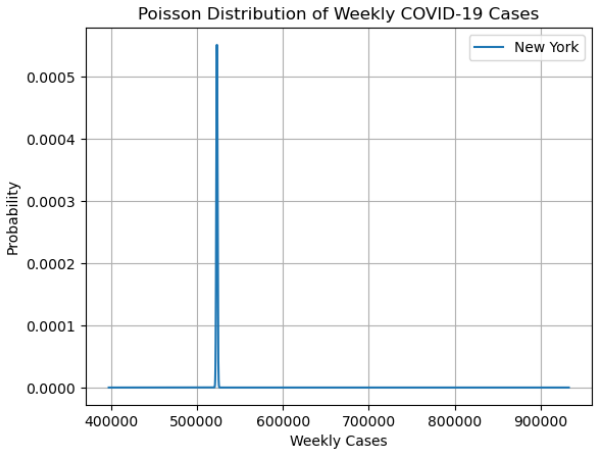

Fig. 27. Poisson distribution of COVID-19 cases in New York State

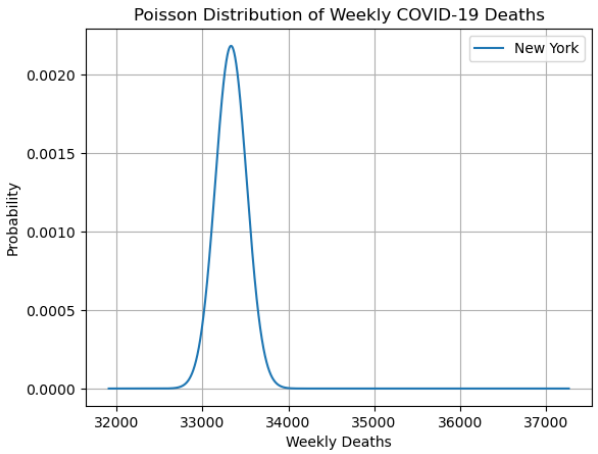

Fig. 28. Poisson distribution of COVID-19 deaths in New York State

3) Poisson Distribution Modeling:

- Figure 27 shows the Poisson distribution of COVID-19 cases in New York State, while Figure 28 depicts the distribution of deaths.

4) Correlation Analysis between Enrichment Data and COVID-19 Cases:

- Correlation was performed within New York State, comparing county-level COVID-19 cases with voter turnout. A moderate positive correlation coefficient of 0.560 was observed.
- Possible explanations for this correlation include community engagement, health awareness, population density, and socioeconomic factors.

5) Formulation of Three Hypotheses:

1) First Hypothesis Statement: The total number of votes cast in New York counties is correlated with the number of COVID-19 cases reported in those counties.

   Null Hypothesis (H0): There is no significant correlation between the total number of votes cast in the 2020

presidential election and the number of COVID-19 cases in New York counties between 2020-06-01 and 2021-01-03.

Alternative Hypothesis (H1): There is a significant correlation between the total number of votes cast in the 2020 presidential election and the number of COVID-19 cases in New York counties.

Results of Hypothesis Test 1:

Pearson Correlation Coefficient: The Pearson correlation coefficient measures the strength and direction of the linear relationship between two variables. In this case, the coefficient of -0.004790540566556201 suggests a very weak negative correlation between the total number of votes cast and the number of COVID-19 cases in New York counties. However, the correlation is close to zero, indicating that there is almost no linear relationship between the variables.

p-value: The p-value associated with the Pearson correlation coefficient test is 0.991867419065854. This p-value represents the probability of observing the given correlation coefficient (or more extreme) if the null hypothesis were true (i.e., if there were no correlation between the variables). Since the p-value is very high (close to 1), we fail to reject the null hypothesis. This suggests that there is no significant correlation between the total number of votes cast and the number of COVID-19 cases in New York counties.

In summary, the result of hypothesis 1 indicates that there is no significant linear relationship between the total number of votes cast and the number of COVID-19 cases in New York counties based on the given data.

2) Second Hypothesis Statement: Counties where a Democratic candidate won the election led to a higher number of COVID-19 cases compared to counties where a Republican candidate won.

   Null Hypothesis (H0): There is no significant difference in the average number of COVID-19 cases between counties where a Democratic candidate won and counties where a Republican candidate won.

   Alternative Hypothesis (H1): Counties where a Democratic candidate won the election have a higher average number of COVID-19 cases compared to counties where a Republican candidate won.

   Results of Hypothesis Test 2:
   The mean computations indicated that the Average number of COVID-19 cases in Democratic won counties was 13,074,044.17 and the Average number of COVID-19 cases in Republican won counties was 12,985,408.4 Democratic-won counties seemed to have a slightly higher average number of COVID-19 cases compared to Republican-won counties. Hence, further statistical analysis, such as hypothesis testing, was needed to determine if this difference was statistically significant.

   The output ”Fail to Reject Null Hypothesis: There is no significant difference in the average number of COVID-19 cases between Democratic-won and Republican-won counties” indicated that based on the statistical analysis conducted, there was insufficient evidence to conclude that the victory of a political party influenced the rise in COVID-19 cases in those counties.

   In other words, the data did not provide enough evidence to suggest that there was a significant difference in the average number of COVID-19 cases between counties won by Democratic and Republican parties. Therefore, we could not attribute the rise in COVID-19 cases solely to the victory of a particular political party in those counties.

   This result suggests that other factors, such as population density, adherence to public health guidelines, healthcare infrastructure, and socioeconomic factors, might have had a more significant impact on the rise in COVID-19 cases in the analyzed counties, rather than the political affiliation of the county’s residents.

3) Third Hypothesis Statement: The conduct of the election in 2020 and the announcement of the victory of Joe Biden in New York State took place on November 3, 2020.
   Null Hypothesis (H0): The victory of Joe Biden in the 2020 presidential election in New York has no significant influence on the increase in COVID-19 cases in the state.

   Alternative Hypothesis (H1): The victory of Joe Biden in the 2020 presidential election in New York has a significant influence on the increase in COVID-19 cases in the state.

   Results of Hypothesis Test 3:
   The output indicates the results of a two-sample t-test conducted to analyze the influence of Joe Biden’s victory in the 2020 presidential election on the increase in COVID-19 cases in New York.

   t-statistic: The t-statistic measured the difference between the means of the two samples (cases before and after the election) relative to the variability in the data. In this case, the t-statistic value was -17.2545. The negative t-statistic value reflects the magnitude of

the difference in COVID-19 cases before and after the election, indicating a significant change over time.

p-value: The p-value represents the probability of observing a test statistic as extreme as the one calculated under the null hypothesis. A lower p-value suggests stronger evidence against the null hypothesis. Here, the p-value is approximately 1.258e-50, which is extremely close to zero.

Interpretation: With such a small p-value (close to zero), we reject the null hypothesis that there is no significant influence of Joe Biden's victory on the increase in COVID-19 cases in New York. Instead, we conclude that there is a significant influence of Joe Biden's victory on the increase in COVID-19 cases in New York.
Overall, based on the results of this hypothesis test, we can infer that Joe Biden's victory in the 2020 presidential election had a significant impact on the increase in COVID-19 cases in New York.

Overall, the analysis provides insights into the distribution, correlation, and potential relationships between COVID-19 cases and various factors, including political engagement and socioeconomic indicators. Further exploration based on the formulated hypotheses could deepen understanding of these relationships.

B. Neetha Ravva

At this stage, I conducted an in-depth analysis of the distribution of COVID-19 cases across the state of North Carolina. Graphical analyses indicated that a log-normalized distribution most accurately fits the data, as confirmed by goodness-of-fit tests comparing normal, exponential, and various other distributions. The characteristics of these distributions, such as modality, skewness, and kurtosis, were also examined. Statistical measures of the distributions, including central tendency and variance, were analyzed, providing insights into the spread and central location of COVID-19 cases within the state.

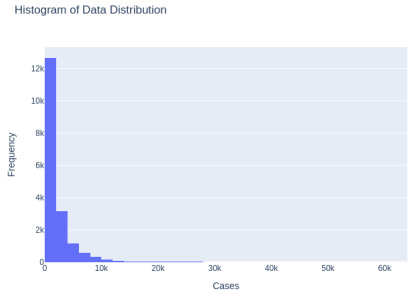

Fig. 29. Normal distribution.

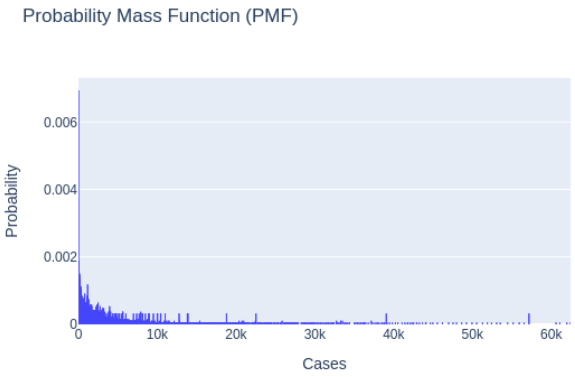

Fig. 30. Probability mass function.

In this phase, we carried out a detailed examination of the COVID-19 case distribution in North Carolina. Graphical analysis showed that a log-normal distribution was the most suitable model for the data, as verified through goodness-of-fit tests against normal, exponential, and several other distributions. We also assessed distribution characteristics including modality, skewness, and kurtosis. Additionally, we looked at statistical metrics such as central tendency and variance, which shed light on the distribution and central concentration of COVID-19 cases across the state.

Poisson distribution for covid 19 cases Modeling the COVID-19 cases and deaths using a Poisson distribution offered unique insights distinct from the initial distribution modeling. The Poisson distribution, focused on new cases and deaths per 100,000 population, relied on the mean value as its parameter. Probability mass functions helped visualize the probability at different case levels, facilitating comparisons between states. The differences between the initial distribution and the Poisson model unveiled that modeling COVID-19 cases and deaths using a Poisson distribution is an oversimplification because the actual distribution of COVID-19 cases and deaths doesn't strictly follow a Poisson distribution. We have proved the same by computing the AIC values in the beginning.

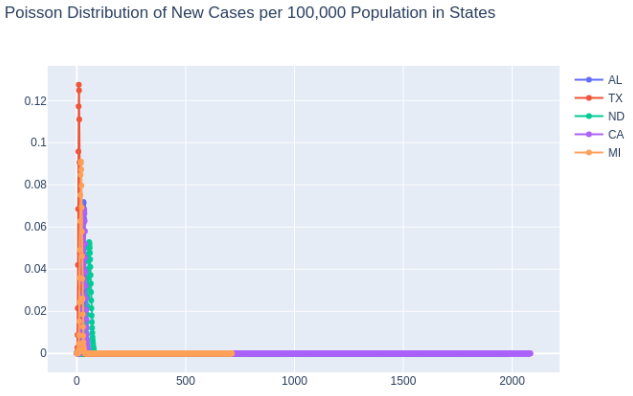

Fig. 31. Poisson distribution for cases across the countries

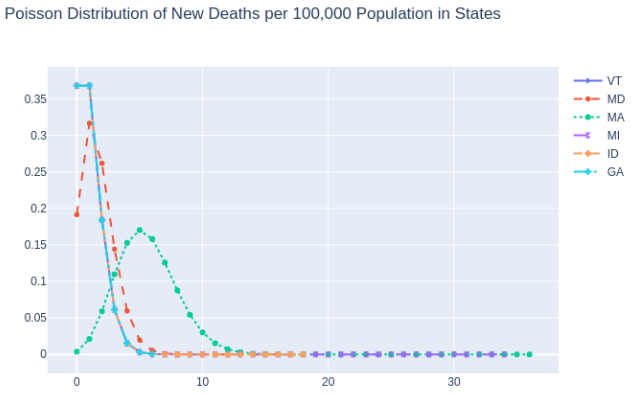

Fig. 32. Poisson distribution for deaths across the countries

Correlation Analysis between Enrichment Data and COVID-19 Cases. A correlation analysis was performed between COVID-19 cases and respective enrichment variables of each team members.

Hypothesis 1: Political Party Affiliation and COVID-19 Case Counts Null Hypothesis (H0): There is no significant correlation between the political party of the winning candidate in a state and the total number of COVID-19 cases reported in that state.

Alternative Hypothesis (H1): There is a significant correlation between the political party of the winning candidate in a state and the total number of COVID-19 cases reported in that state.

Hypothesis 2: Voter Turnout and COVID-19 Case Counts Null Hypothesis (H0): There is no significant relationship between voter turnout in a state and the total number of COVID-19 cases reported in that state.

Fig. 33. Correlation Analysis between Enrichment Data and COVID-19 Cases.

Alternative Hypothesis (H1): There is a significant relationship between voter turnout in a state and the total number of COVID-19 cases reported in that state.

Hypothesis 3: Population Size and COVID-19 Case Counts Null Hypothesis (H0): There is no significant correlation between the population size of a state and the total number of COVID-19 cases reported in that state.

Alternative Hypothesis (H1): There is a significant correlation between the population size of a state and the total number of COVID-19 cases reported in that state.

C. Kevin Hayes

Kevin Hayes based his data on the state of Florida.

He began by creating a histogram of the state of Florida's cases as an image using matplotlib, as shown in figure 34

he then created a function to get the mean, median, mode, variance, skewness, kurtosis, maximum value, minimum value, and standard deviation of a given dataset. This is then used in a function that creates a gamma distribution of a given state's cases and then plots that on top of a histogram of that states cases. How this looks for Florida is shown in figure 35.

Then this procedure was repeated for the states of North Carolina, Alabama, California, Maine, and Missouri. The resulting diagrams can be seen in figures 36,37,38,39, and 40.

The notable features of these graphs are that

1) The distributions look similar to each other, but NC and MO are bimodal.
2) For most of them, the gamma distribution fits well.

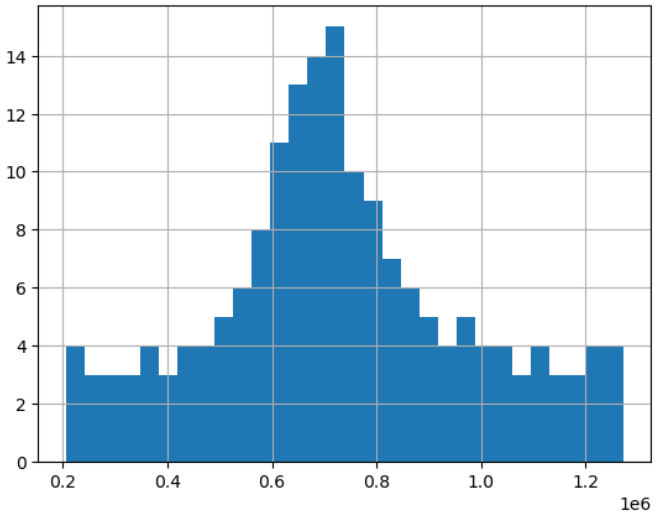

Fig. 34. Histogram of cases in Florida.

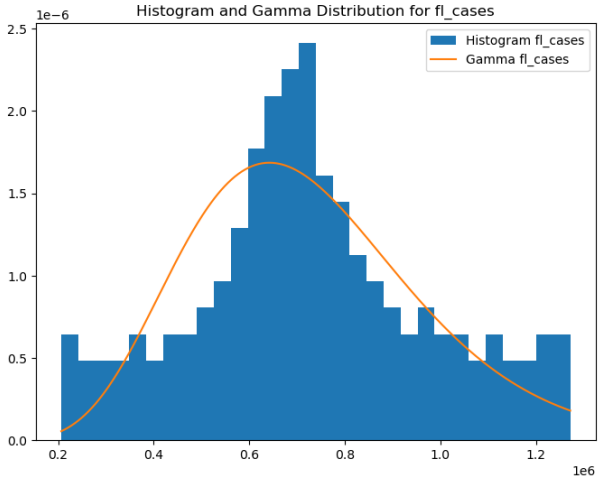

Fig. 35. Histogram and Gamma Distribution for Florida cases

3) Florida has more neutral skew while the others are more right skewed. This means that Florida had more values that are higher than the other states.

Kevin then made a function that automatically creates a Poisson distribution of a state given the state ID and the requested dataset.

He then made Poisson functions to graph the cases and deaths.

Kevin then created a correlation between his enrichment data and total number of COVID-19 cases by state.

The immediate best corralations found were the following:

1) It looks like the best correlations available are:
   a) DP05_0011E: Estimate!!SEX AND AGE!!Total population!!35 to 44 years
   b) DP05_0002E: Estimate!!SEX AND AGE!!Total population!!Male

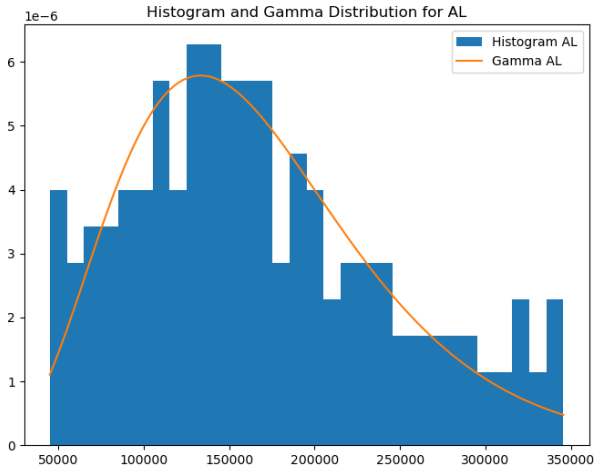

Fig. 36. Alabama Gamma Distribution

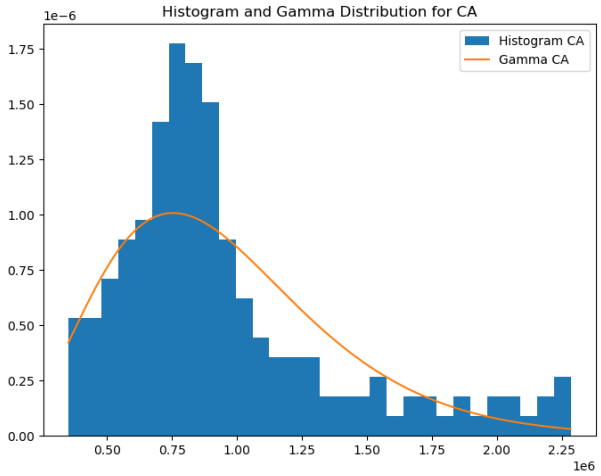

Fig. 37. California Gamma Distribution

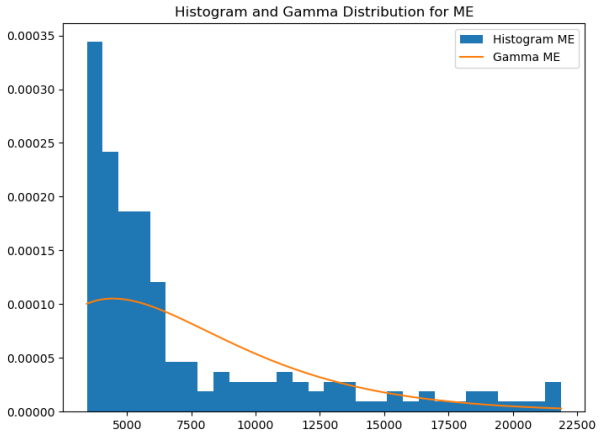

Fig. 38. Maine Gamma Distribution

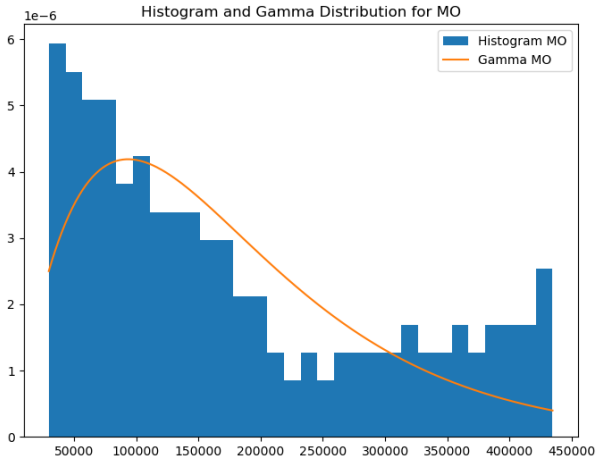

Fig. 39. Missouri Gamma Distribution

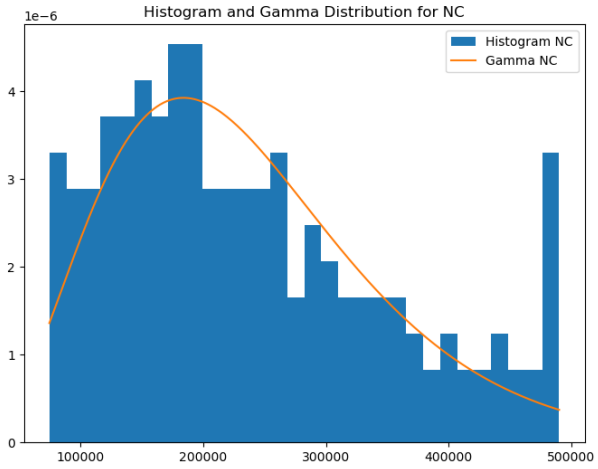

Fig. 40. North Carolina Gamma Distribution

   c) DP05_0026E: Estimate!!SEX AND AGE!!Total population!!18 years and over!!Male
   d) DP05_0090E: Estimate!!CITIZEN, VOTING AGE POPULATION!!Citizen, 18 and over population!!Male
   e) DP05_0010E: Estimate!!SEX AND AGE!!Total population!!25 to 34 years
2) Note that these are all negative corralations. The best positive correlations we have (Aside from mean value) are
   a) DP05_0013E: Estimate!!SEX AND AGE!!Total population!!55 to 59 years
   b) DP05_0014E: Estimate!!SEX AND AGE!!Total population!!60 to 64 years
3) From this it looks like more young people $\implies$ less COVID and more old people $\implies$ more COVID.

All of the best correlation data came from age, so Kevin decided to filter by race data to see what the best correlations there were.

1) Our best negative correlation involving race is: DP05_0078E: Estimate!!HISPANIC OR LATINO AND RACE!!Total population!!Not Hispanic or Latino
2) Our best positive correlation involving race is: DP05_0073E: Estimate!!HISPANIC OR LATINO AND RACE!!Total population!!Hispanic or Latino (of any race)

Then, to form the hypothesis questions, Kevin chose the following

1) Does having a higher total population of Hispanic or Latino people of any race in a given state lead to more COVID-19 infections in that state?
   a) Null Hypothesis: having a higher total population of Hispanic or Latino people of any race in a given state has no effect on number of COVID 19 infections in that state
   b) Alternate Hypothesis: having a higher total population of Hispanic or Latino people of any race in a given state increases the number of COVID 19 infections in that state
2) Does having more people of 35 to 44 years of age in a given state leads to fewer COVID-19 infections in that state?
   a) Null Hypothesis: having more people of 35 to 44 years of age in a given state has no effect on COVID-19 infections in that state
   b) Alternate Hypothesis: having more people of 35 to 44 years of age in a given state leads to fewer COVID-19 infections in that state.
3) does having more people aged 60 to 64 in a given state lead to more COVID-19 cases in that state?
   a) Null Hypothesis: having more people aged 60 to 64 in a given state has no effect on COVID-19 cases in that state
   b) Alternate Hypothesis: having more people aged 60 to 64 in a given state leads to more COVID-19 cases in that state

V. Stage IV: Prediction

Dealing with a pandemic requires one to be able to figure out how events will likely work out before they happen. We attempt to do this through the use of machine learning models. We are using fairly basic regression models. These models are linear regression models and polynomial regression models.

Linear regression and polynomial regression both have their pros and cons.

1) Linear Regression
   a) Pros
      i) Simplicity. Linear regression models are very fast and resource efficient to train
      ii) Low variance. This model does not tend to be over-fitted.
   b) Cons
      i) High bias. Tends to be very under-fitted.

ii) Can not properly handle the prediction of complex trends

2) Polynomial Regression
   a) Pros
      i) Low bias. Can be fitted to handle complex trends.
      ii) Can be well trained to avoid under-fitting.
   b) Cons
      i) High variance. A small change in the training data can lead to large changes in the model. The model is also prone to over-fitting.
      ii) Cost of training. Compared to a linear model, it is more expencive to train the polynomial model.

We then use linear and polynomial models to predict the number of cases and deaths for COVID-19 during the time between may 25Th, 2020 and January 10Th, 2021. We do this by creating a function that makes a plot of a regression function with an arbitrary number of degrees, and then plotting all of these data with degree = 1 and degree = 2 for the polynomial plot. We find that the polynomial regression functions tend to be better at predicting the next week of data as ,within the bounds that have been set by our model, the linear model tends to predict answers close to the mean number of cases.

We then discuss bias vs variance in relation to our model. Our discussion is as follows

> Bias is assumptions that the model makes in order to simplify the target function. Too much of it, and you'll end up with a straight line, regardless of how straight or curved the data is. This is where variance comes in; it acknowledges the actual data trends and starts to curve the target function to fit the model closer to it. In order to get a more accurate regression function, variance is important. However, too much of it, and the model begins overfitting the data, where instead of predicting a function that closely fits the data, the line follows each datapoint to the next. Because it gets so specific in the training set of the data, it becomes more difficult for the model to make predictions on any new dataset. You need a balance between the two in order to get the most efficiently accurate model.

We then compare the United States infection and death data to Japan, Brazil, Germany, Greenland, China, and Russia. Our findings are as follows:

1) **Japan** The regression line was an upward trend. The new cases were trending upward, and the forecasted data predicted increasing new cases over the coming week. The polynomial model predicted an increase like the linear model. The same pattern was seen for the prediction of new deaths.
2) **Brazil** The pattern for Brazil was slightly different in that it was unstable, that is, fluctuating in the cases' linear model. However, in the deaths slope, it was a downward trend indicating a decreasing death rate. This pattern was consistent for both the linear and polynomial models in both cases and deaths in Brazil. The predicted values for cases and deaths did not forecast any sharp increase or sharp decline.
3) **Germany** The trend of the linear model was an upward view for both cases and deaths. The polynomial model also indicated an upward trend in Germany for both deaths and cases. The polynomial model however predicted fewer cases and fewer deaths in the next week in slight contrast to the prediction of the linear model for the deaths in the coming week.
4) **Greenland** The linear model gives a slightly upward trend for the new cases. The deaths however remained stable without increase or decrease for the linear model. The polynomial model was also like the linear model for both cases and deaths.
5) **China** The polynomial model produced a slightly upward trend for both cases and deaths with the cases more skewed upwards. However, the prediction for the next week was a slight increase.
6) **Russia** The trend was an upward pattern indicating rises for both the linear and polynomial models in Russia. The models predicted an increase in the coming week.
7) **USA** The trend was an upward line indicating rises for both linear and polynomial models for the USA. Both models predict an increase in the coming week. This pattern is the same for both new cases and deaths in the USA.

For our team work we plotted for individual states and counties rather than for the US as a whole. Then we did our hypothesis testing for the questions we made in stage 3.

VI. Stage V: Dashboard

The COVID-19 pandemic has had a significant impact on the economy, the health of people all around the world, and everyday life. Cases and fatalities caused by COVID-19 need to be monitored and analyzed in order to have a better knowledge of the virus's transmission and to make decisions regarding public health actions. The data on COVID-19 cases and fatalities for a number of states in the United States were shown through the use of a dashboard that was developed as part of this research. At the same time, a callback function was built so that the dashboard could be dynamically updated based on the input provided by the user.

A. Dashboard Design

The dashboard was created using Dash, a Python framework for building analytical web applications. It consists of the following components:

- **Title:** The dashboard is titled "COVID-19 Dashboard" to convey its purpose effectively.
- **Date Pickers:** Users can select the start and end dates to specify the time range for data visualization.
- **Mode Selector:** Users can choose between linear and logarithmic scales for the y-axis.

- **Performance Options:** Checklists allow users to toggle between displaying actual values, trendlines, and 7-day moving averages for both cases and deaths data.
- **State Selectors:** Dropdown menus enable users to select one or more states for analysis.
- **Graphs:** Two graphs are displayed side by side. The first graph visualizes COVID-19 cases and deaths for three selected states (AL, AR, and CA) in linear scale. The second graph shows the same data but in logarithmic scale.

B. Callback Function

A callback function was defined to update the graphs dynamically based on user input. The function takes inputs such as selected dates, mode (linear or logarithmic), performance options, and selected states. It then filters the data accordingly and generates traces for the graphs. Specifically:

- For each selected state, actual values, trendlines, and 7-day moving averages are plotted based on the user's performance options.
- Trendlines are calculated using polynomial regression.
- One-week predictions are generated for selected states, such as NC, and appended to the graphs.
- The callback function ensures that the dashboard reflects the latest data and provides insights into COVID-19 trends.

C. Screenshots

Twelve screenshots (Figures 41 - 52) were captured to showcase different aspects of the dashboard and callback function.

VII. Conclusion

The project began with a data understanding in stage one, data modelling in stage two, distributions development in stage 3, and application of machine learning and statistical model in stage four. The activities of stages one to four were presented in a capsular summary by developing a simple interactive dashboard based on the analysis. This was the stage five. An intuitive user interface was designed by the dashboard design and callback function, which allowed for the analysis of data pertaining to COVID-19 cases and deaths. By providing users with the ability to examine the data in an interactive manner, compare the data from other states, and display patterns, the dashboard makes it easier for researchers, users and decision makers to make educated decisions and effectively implement public health initiatives. In general, the analysis of the COVID19 data has helped us to achieve the goal and objectives of the course (Data Science) offered at the Department of Computer Science, UNCG, through hands-on activities implemented in five broad stages. Additional upgrades might include the presentation of additional data, such as geographical maps, in order to give other insights into the COVID-19 epidemic. In general, the project made a contribution to the ongoing efforts that are being made to both prevent the spread of COVID-19 and lessen the damage that it has on society.

Acknowledgment

For the Summary paper submission only, no acknowledgements are allowed.

References

Note: For the Summary paper submission only, references to the authors own work should be cited as if done by others to enable a double-blind review. **Citations must be complete and not redacted, allowing the reviewers to confirm that prior art has been properly identified and acknowledged.**

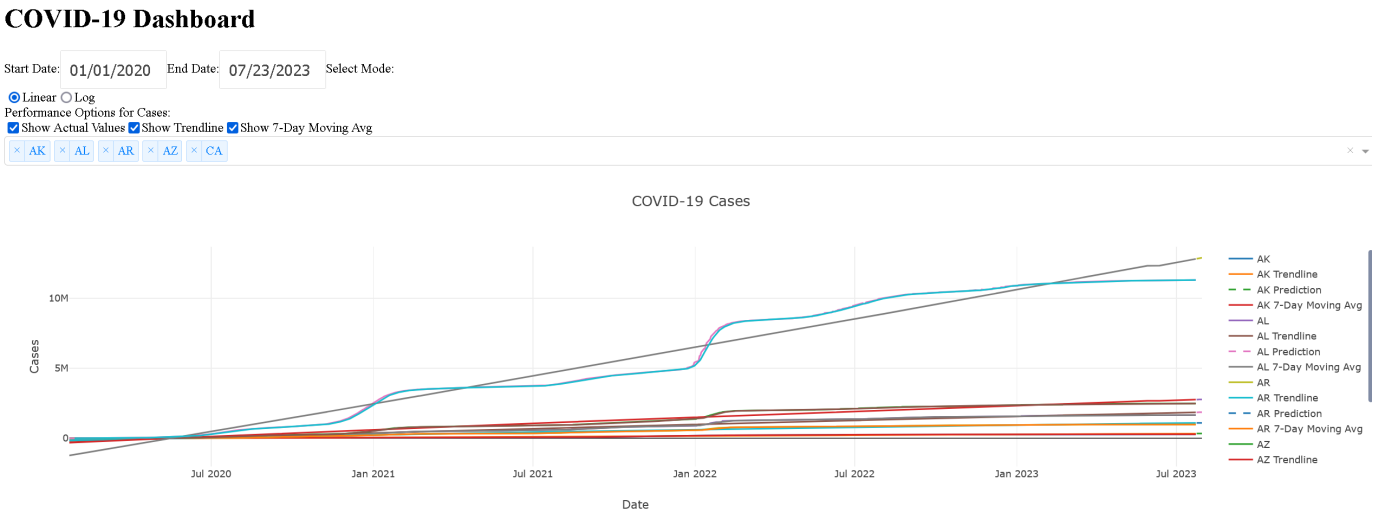

Fig. 41. Linear plot of Five states (AK, AL, AR, AZ, and CA states -LINEAR) selected for the dates from January 1, 2020, to July 23, 2023 for cases on the dashboard.

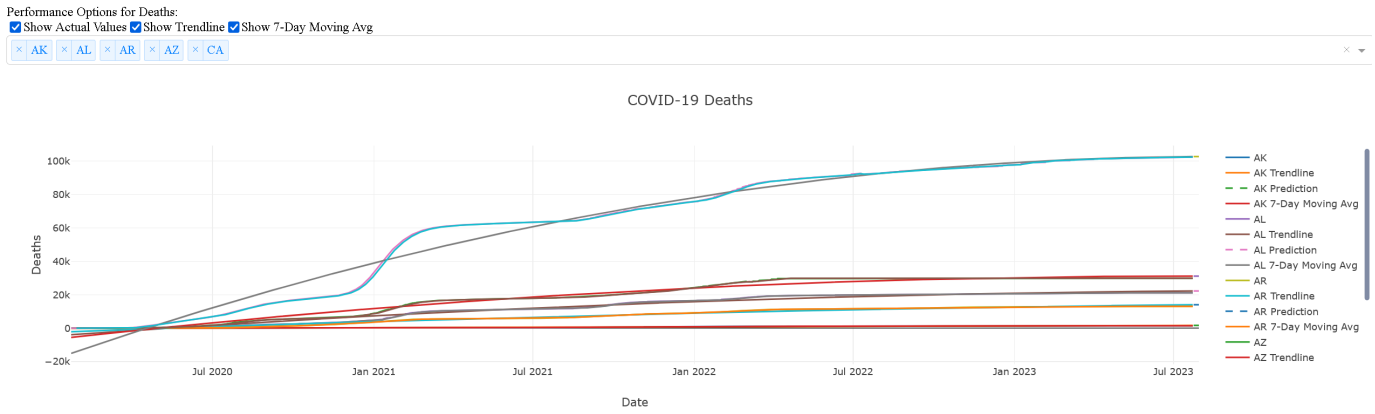

Fig. 42. Linear plot of Five states (AK, AL, AR, AZ, and CA states -LINEAR) selected for the dates from January 1, 2020, to July 23, 2023 for deaths on the dashboard.

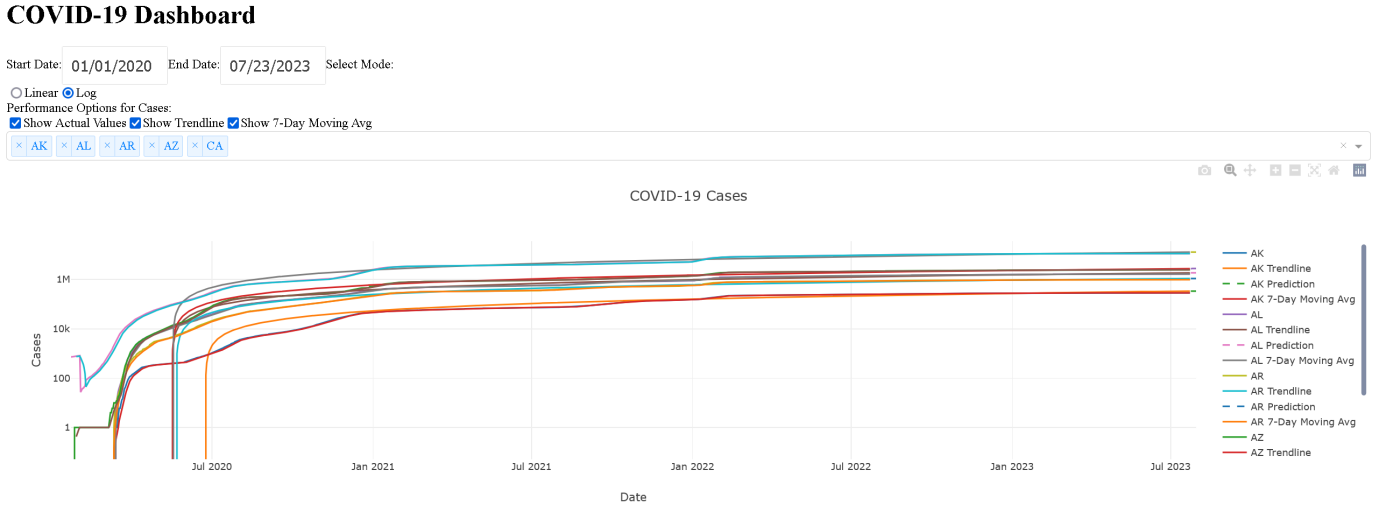

Fig. 43. AK, AL, AR, AZ, and CA states -LOG from 1-1-2020 to 07-23-2023 cases plot.

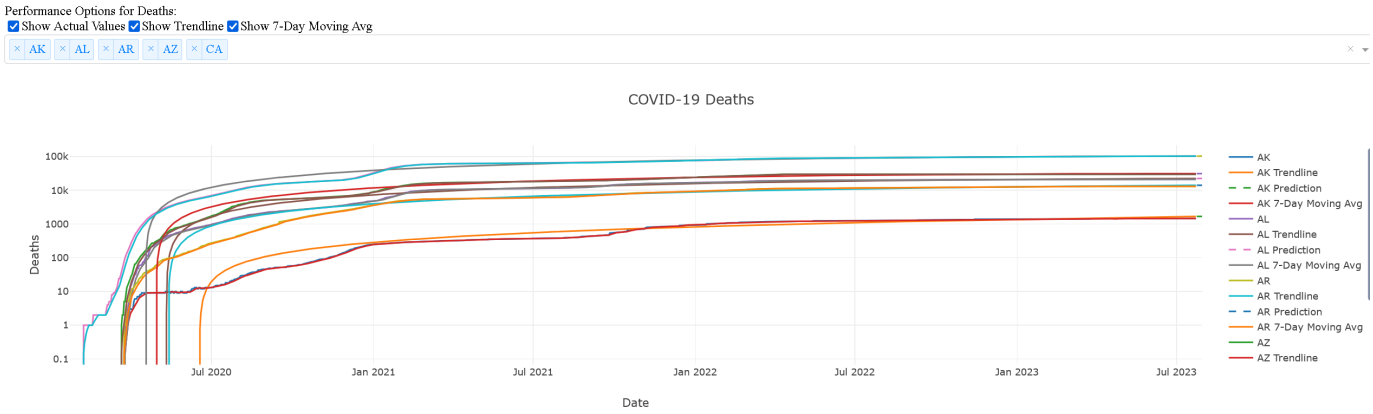

Fig. 44. AK, AL, AR, AZ, and CA states -LOG from 1-1-2020 to 07-23-2023 deaths.

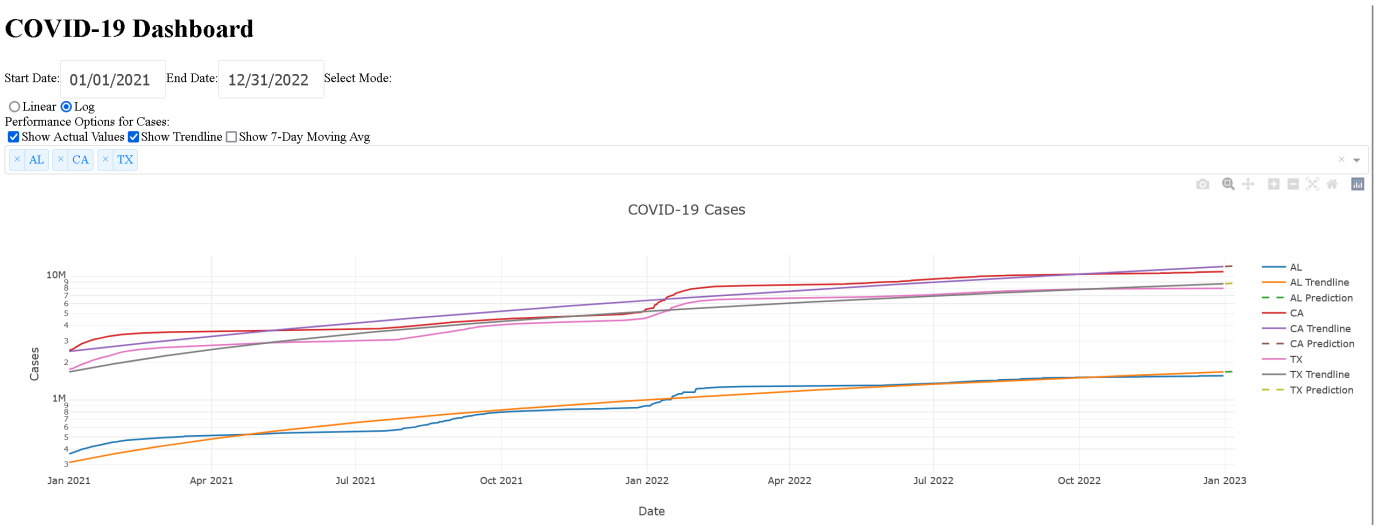

Fig. 45. Visualization when AL, CA, and TX states -LOG from 1-1-2021 to 12-31-2022 cases was selected.

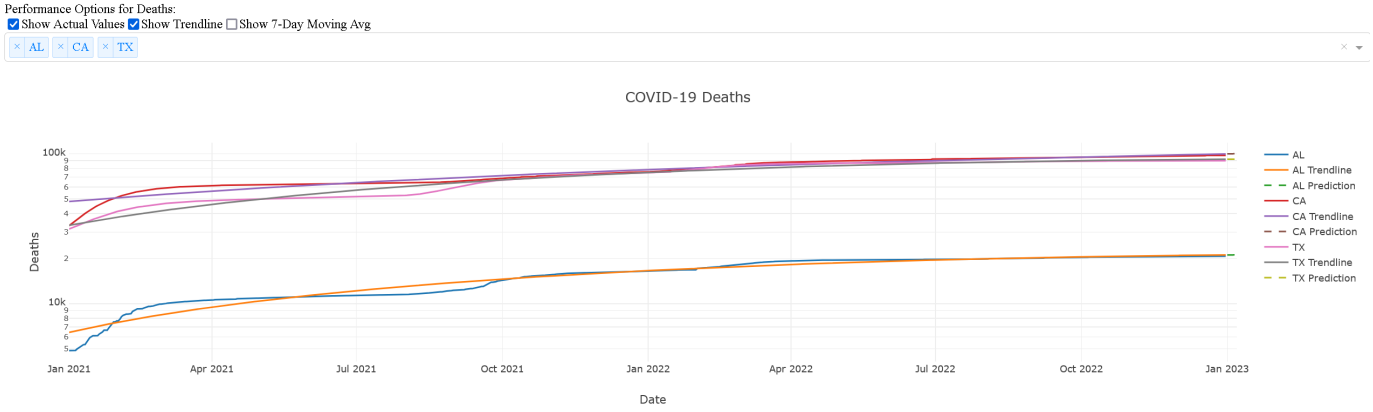

Fig. 46. Visualization when AL, CA, and TX states - LOG from 1-1-2021 to 12-31-2022 deaths was selected.

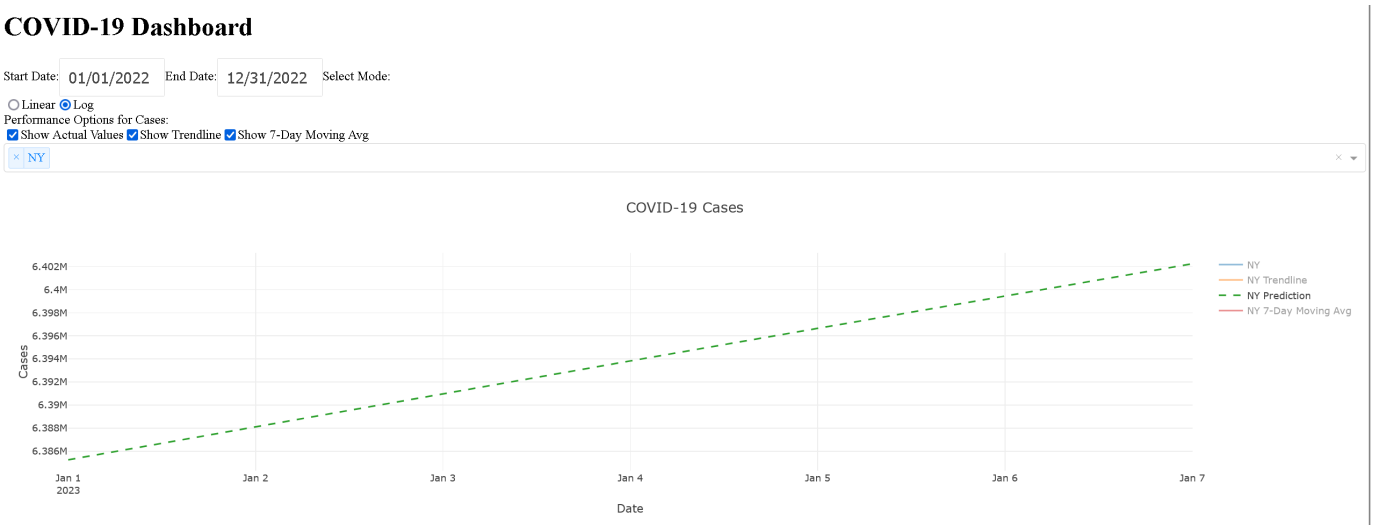

Fig. 47. One-week prediction for NY based on LOG from 1-1-2022 to 12-31-2022 cases

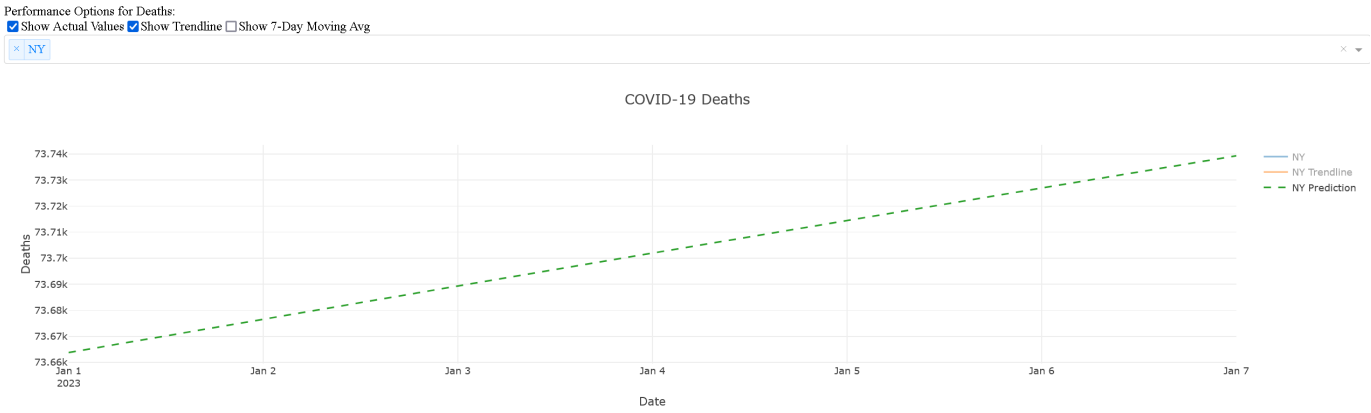

Fig. 48. One-week prediction for NY based on LOG from 1-1-2022 to 12-31-2022 deaths

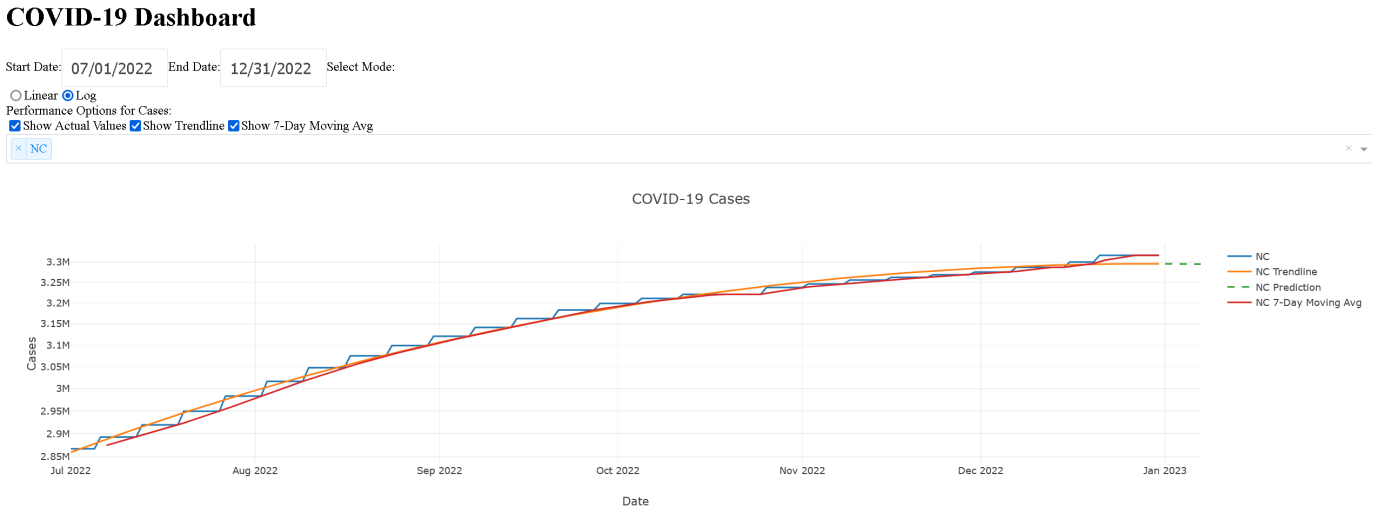

Fig. 49. The entire plot including the trendline and prediction for NC state - LOG from 7-1-2022 to 12-31-2022 cases.

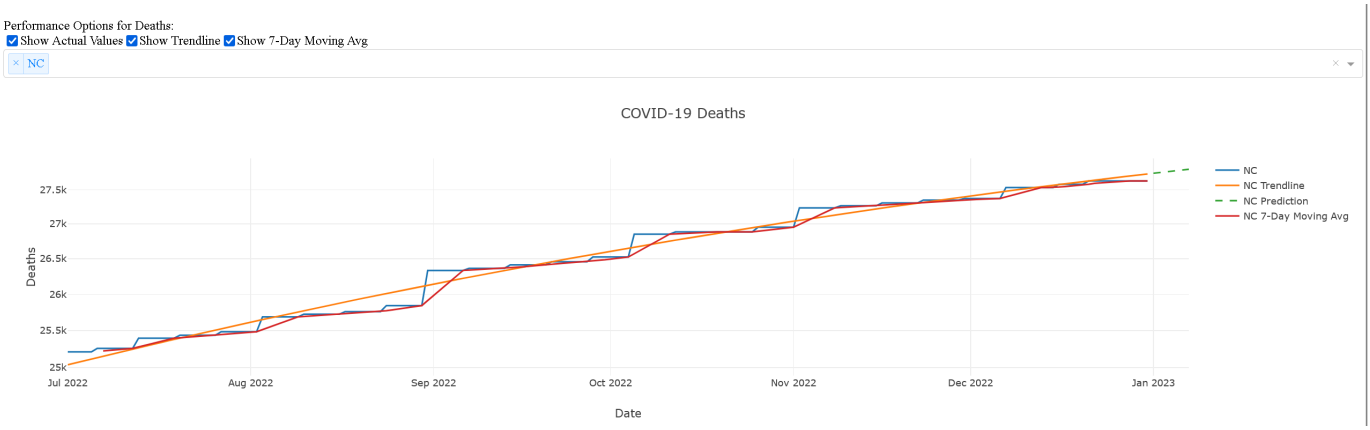

Fig. 50. The entire plot including the trendline and prediction for NC state -LOG from 7-1-2022 to 12-31-2022 deaths.

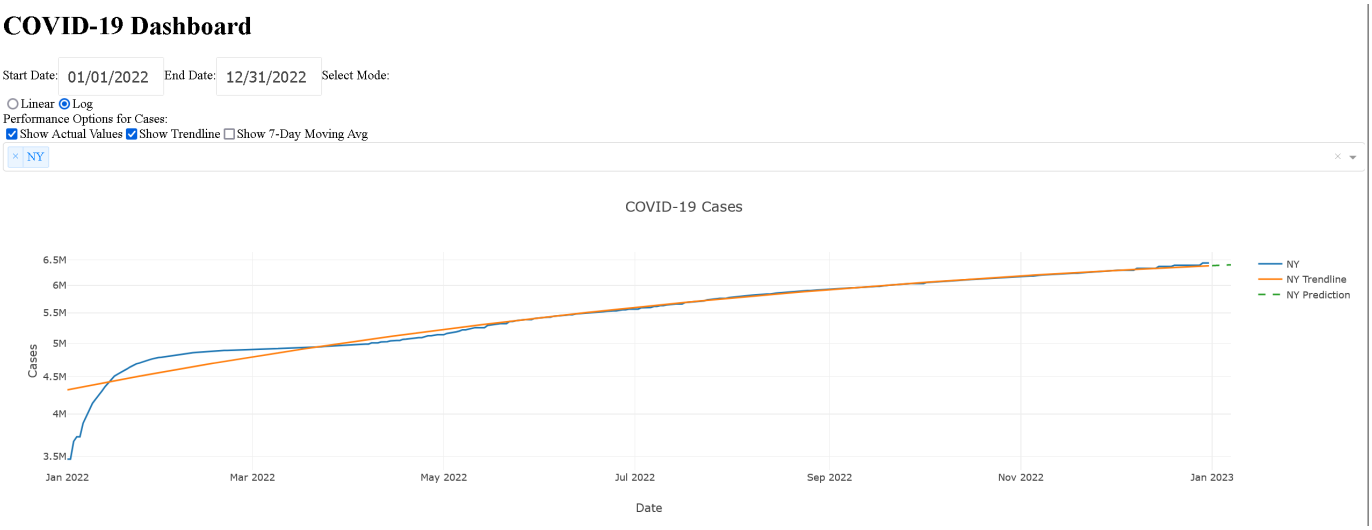

Fig. 51. Visualization when NY state - LOG from 1-1-2022 to 12-31-2022 cases was selected.

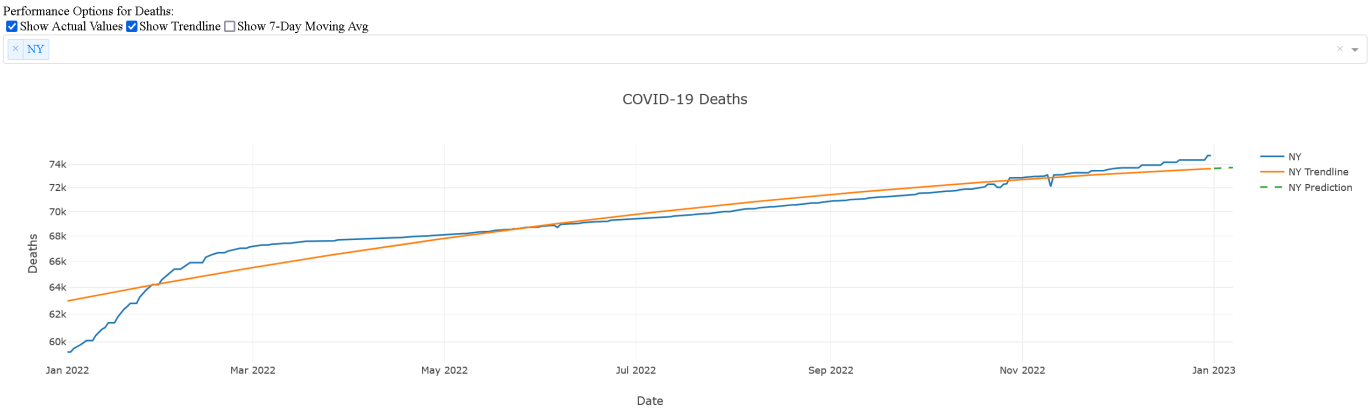

Fig. 52. Visualization when NY state -LOG from 1-1-2022 to 12-31-2022 deaths was selected.