**COVID Data Analysis**

Keerthi Devireddy, Sony Annem, Supriya Kottam

***Abstract – This report introduces a structured approach to examining how COVID-19 has impacted the United States. The framework is divided into five phases, addressing different project aspects, such as collecting the data, data modelling, analyzing distributions, testing hypothesis, implementing fundamental machine learning, and crafting a dashboard. Our method relies on data-driven techniques, employing advanced statistical and machine leaning methods to explore pandemic patterns. The objective is to provide valuable insights that can guide healthcare experts and policymakers in making informed choices to control COVID-19 transmission.***

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

I. INTRODUCTION

he, global impact of the COVID-19 pandemic has halted regular life, significantly impacting countless lives and economies worldwide. In the United States the pandemic has caused a substantial number of cases and fatalities. To comprehend its effects better, merging various datasets can offer valuable insights into the pandemic’s spread and consequences.

T

We plan to use data from usfacts.org, which includes a daily county-level tracker of COVID-19 cases, deaths, Population that can be discussed in more detail in Stage I. And we have three enrichment datasets for COVID-19 which can be gathered from the below links:

* **Employment Dataset**

<https://www.bls.gov/cew/downloadable-data-files.html>

The Employment Dataset provides comprehensive information on employment levels and earning potential across various counties. Integrating this dataset allows us to investigate potential links between employment demographics and COVID-19 spread. Examining the impact of economic activities and workforce distribution on infection rates could shed light on socio-economic factors influencing the pandemic's trajectory within different regions.

* **Presidential Dataset**

<https://www.>[kaggle.com/unanimad/us-election-2020](https://www.kaggle.com/unanimad/us-election-2020)

The Presidential Election Results dataset offers invaluable insights into the political landscape of each county, detailing the winning candidate and their margin of victory. By incorporating this dataset, our analysis can explore correlations between political leanings and responses to the COVID-19 pandemic. Understanding how political affiliations intersect with pandemic responses may unveil regional disparities in healthcare management, policy adherence, and resource allocation.

* **Census Demographic**

<https://data.census.gov/cedsci/table?q=dp&tid=ACSDP1Y2018.DP05>

The Census Demographic ACS dataset encompasses detailed demographic information for each county, including population estimates categorized by age groups. Incorporating this dataset enables an exploration of how population demographics intersect with COVID-19 trends. Analyzing age-specific infection rates and population distributions could help discern vulnerabilities among different age cohorts and aid in formulating targeted intervention strategies and healthcare resource allocation.

II. PROJECT STAGES

*A. Stage I* (Data and Project Understanding)

The primary focus of Stage I was to lay the foundational groundwork for our COVID-19 analysis by comprehensively understanding the datasets at our disposal. Our team embarked on this phase by sourcing the primary COVID-19 dataset from usafacts.org. This dataset offers a daily county-level tracker of COVID-19 cases, deaths, and population data. Its granularity enables a detailed examination of infection rates per 100,000 individuals, providing a nuanced view crucial for our analysis. Understanding this dataset's structure and variables was the initial step toward unravelling the pandemic's regional dynamics.   
  
The below are the data dictionaries of the population data from usfacts.

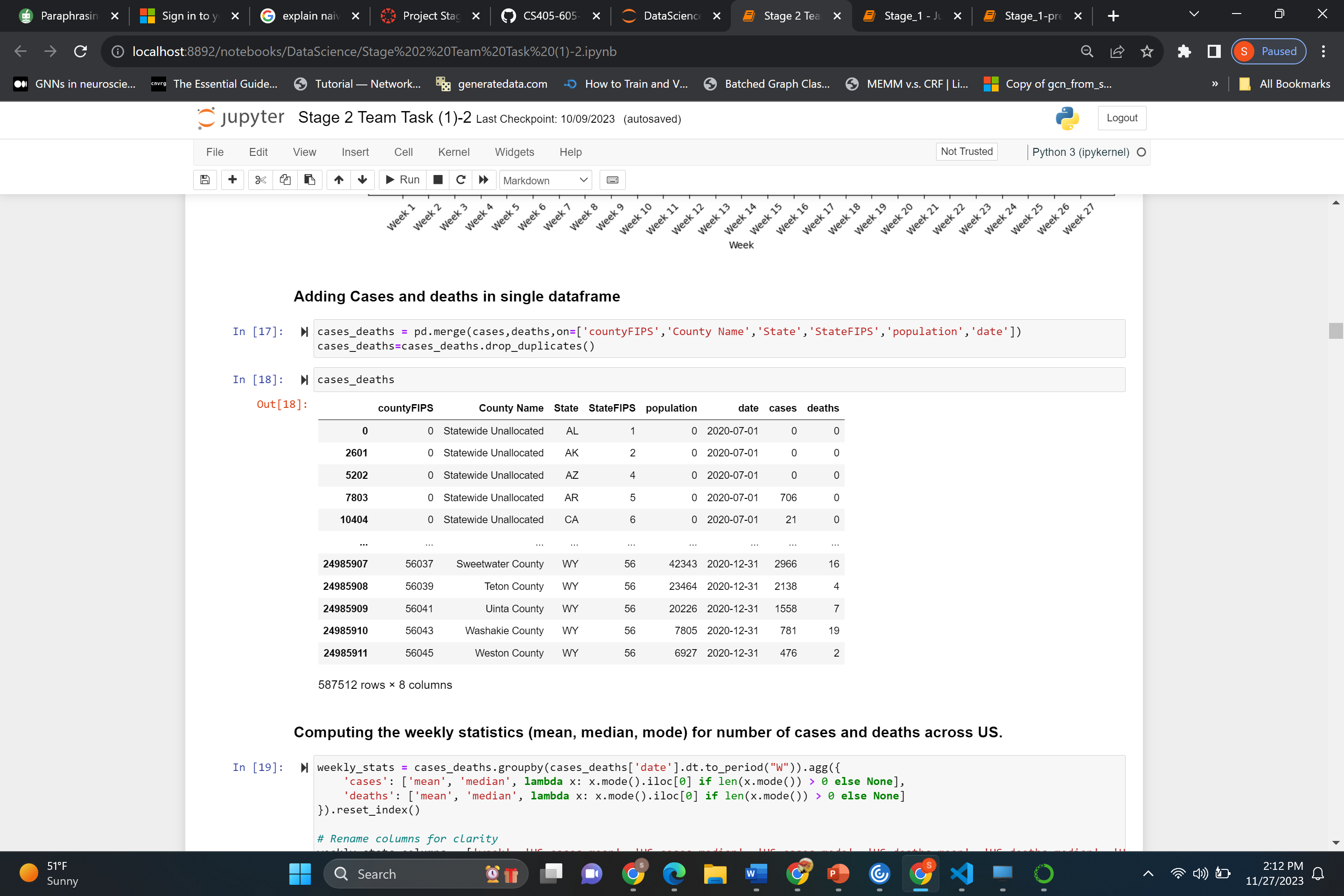
<https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Definition | Date type | Possible values | Req? |
| County FIPS | The code for the counties | Integer | 1200, 1133…. | Yes |
| County Name | Name of the County | String | Telon County… | Yes |
| State | Shorthand notation of the US Sate names | String | TX, VT, NC… | Yes |
| Population | Number of people | Integer | 67892, 99887… | Yes |

The below table is the data dictionary for the Number of Cases and the Number of Deaths from usfacts.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Definition | Date type | Possible values | Req? |
| County FIPS | The code for the counties | Integer | 1200, 1133…. | Yes |
| County Name | Name of the County | String | Telon County… | Yes |
| State | Shorthand notation of the US Sate names | String | TX, VT, NC… | Yes |
| Date | The count of number of cases for that specific county in that particular date (number of cases dataset). The count of number of deaths (number of deaths dataset). | Integers | Positive numbers | Yes |
| State FIPS | The code for the US States | Integer | 1120, 1986… | Yes |

The data cleaning operations are also performed in this stage, we have removed duplicate values, handled missing values, standardised formats and merged the covid datasets (cases, deaths, and population) to form a single well-structured SuperCovid19 dataset. The below image shows the transformed data frame:



To enrich our analysis and augment the COVID-19 dataset, each team member delved into separate enrichment datasets. These included the Census Demographic ACS dataset, the Employment Dataset, and the Presidential Election Results dataset. The Census Demographic ACS offers comprehensive demographic insights into each county, including age-group-based population estimates. Simultaneously, the Employment Dataset provides information on employment levels and earning potential across counties. Additionally, the Presidential Election Results dataset sheds light on the political landscape, detailing winning candidates and their margins of victory within each county.

*Our team's task was not only to comprehend these datasets individually but also to envision their integration with the primary COVID-19 dataset. By identifying common variables and potential mapping points between datasets, we aimed to merge these datasets effectively. Furthermore, our objective was to articulate how these enrichment datasets could bolster our analysis of COVID-19 spread. Initial hypothesis questions were posed, anticipating correlations between socio-economic factors, political inclinations, and the trajectory of the pandemic.*

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

As part of our team's collaborative efforts, we delved into a comprehensive analysis of COVID-19 trends across three states. Focusing on the trends of the last week, our analysis encompassed an in-depth examination of state-level COVID-19 data. Each team member chose a specific state for analysis, enabling us to construct a detailed picture of the pandemic's trajectory within these regions. By calculating COVID-19 data trends and assessing whether cases were increasing, decreasing, or remaining stable, we gained invaluable insights into regional variations. Furthermore, this analysis allowed us to identify that the cases recorded are stable peaks, and potential influencing factors unique to each state. These state-wise insights formed a crucial foundation for understanding localized impacts and paved the way for targeted investigations in subsequent stages of our project.

A screenshot of a computer

Description automatically generated

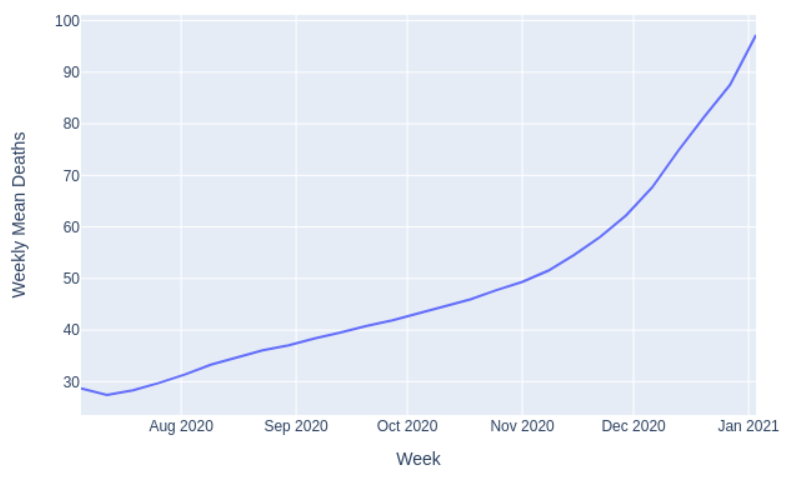
A graph with a line going up

Description automatically generated

*B. Stage II (Data Modeling)*

The team's Stage II analysis focused on a comprehensive comparative study of COVID-19 trends across the United States and selected countries worldwide. Utilizing weekly statistics for the number of cases and deaths, our analysis aimed to provide insights into the pandemic's trajectory on a global scale. We conducted a meticulous comparison of mean, median, and mode values for cases and deaths within the US, shedding light on the pandemic's progression over time.

The below graphs show the monthly mean cases and deaths. We can observe that the rise of cases and deaths is directly proportional to the time and the peaks are observed in the months of November, December, and January.

A graph with a line going up

Description automatically generated

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.

A graph of the covid-19 cases

Description automatically generated

A graph of the covid-19 cases

Description automatically generated

A graph of a number of people

Description automatically generated

A graph of a number of cases

Description automatically generated

The generated plots and comparative trends present an invaluable resource for understanding the global dynamics of the pandemic, showcasing fluctuations, identifying critical periods, and providing a nuanced view of the impact of diverse socio-environmental factors on disease spread.

**Insights into Sudden Surges in COVID-19 Cases and Deaths**

Our investigation into the sudden spikes in COVID-19 cases and deaths across diverse countries revealed compelling patterns and influential factors contributing to these surges. In the global context, the month of December 2020 witnessed a substantial increase in COVID-19 cases worldwide, stemming from multifaceted factors that triggered resurgence after initial containment efforts.

United States

The surge in COVID-19 cases during December correlated with several significant events. Post-Thanksgiving gatherings prompted a surge in infections, compounded by extensive New Year celebrations. Additionally, the winter season exacerbated transmission, and scepticism surrounding the COVID-19 vaccine hindered its widespread adoption, further contributing to the surge.

Indonesia

The sudden increase in cases aligned with outbreaks of Monkey Pox and seismic activities, particularly an earthquake in November. These events induced population movements, fostering conditions conducive to COVID-19 spread.

Pakistan

High case counts in November and December were associated with festive celebrations, where families congregated, potentially facilitating virus transmission. Moreover, the onset of winter exacerbated respiratory conditions. Plans to procure vaccines signalled hope, albeit the surge persisted amid hesitancy regarding vaccine uptake.

Nigeria

A rise in cases following December correlated with relaxed lockdown measures and reduced adherence to preventive protocols. The relaxation of mask usage in public spaces coincided with New Year celebrations, fostering an environment conducive to virus transmission.

India

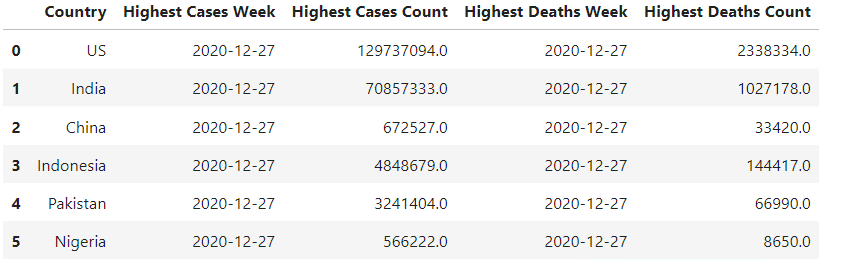
The introduction of the vaccination program in January 2021 marked a significant milestone. Despite this, the surge can be attributed to various factors, including New Year festivities, winter conditions, and widespread travel during festive seasons, especially in public transportation, augmenting transmission risks.

China

Notably, China experienced a relatively lower caseload due to stringent lockdown measures and an elimination strategy until early 2022. However, a surge emerged in December 2022 despite high vaccination coverage. This anomaly underscored the challenges in sustaining containment efforts over extended periods.

These findings underscore the intricate interplay between social behaviours, seasonal influences, vaccination efforts, and public health policies in influencing COVID-19 transmission. While each country faced unique challenges, a convergence of factors, including societal gatherings, seasonal variations, and vaccine hesitancy, contributed to the sudden spikes observed globally, highlighting the need for sustained vigilance and adaptive strategies in combating the pandemic.

As a part of identifying peak weeks of the cases and deaths in United States and other countries. We have generated the below table and observed that last week of December is the week where highest number of cases and deaths are reported for all the countries.



As a part of Stage II member task, each member of our team has selected a specific state and generated weekly statistics for the state. The below is the plot for the state of North Carolina.

A graph of a number of cases

Description automatically generated

A graph with numbers and a line

Description automatically generated

In North Carolina, the COVID-19 trajectory showed a steady increase in both cases and deaths throughout the monitored period. From June 30 to July 6, 2020, the recorded cases stood at 699.67, with 321 deaths reported. Over subsequent weeks, a consistent upward trend was observed in both cases and deaths. Notably, by December 29, 2020, to January 4, 2021, the cases had escalated significantly to 5248.91, accompanied by 2722 deaths. The data signifies a persistent rise in COVID-19 prevalence, suggesting a continual and concerning spread of the virus within the state over the observed period. These escalating numbers underscore the importance of ongoing vigilance, public health measures, and strategic interventions to curb the transmission and impact of the virus within North Carolina.

**Weekly Statistics Comparison Across States:**

Conducted a comprehensive analysis comparing weekly statistics for COVID-19 cases and deaths across a specific state and five other states. To ensure accurate comparisons, we normalized these statistics by population, allowing for a fair assessment based on per capita rates (e.g., per 10,000 or 100,000 individuals). The plotted line graph juxtaposing the values across the weeks for these states revealed intriguing insights into the varying trajectories. These divergent patterns in COVID-19 rates among states stemmed from multifaceted factors, including differing public health policies, population densities, healthcare infrastructure, and adherence to preventive measures. The identification of peaks enabled a comparison with the broader US pattern, elucidating similarities or disparities in the timing and intensity of outbreaks.

Cases surged over time, peaking in Nov-Dec due to factors like holidays, winter, and variant impact. Population directly influences state-wise case and death numbers, highlighting a clear correlation.

A graph with different colored lines

Description automatically generatedA graph of a number of people

Description automatically generated On observing the above plots and comparing the COVID-19 cases and deaths rates across five states—Alaska (AK), Alabama (AL), Arizona (AZ), Texas (TX), and Virginia (VA)—reveals distinct trajectories in the pandemic's impact. Alaska initially exhibited the lowest rates among the states, gradually rising but maintaining a comparatively lower trajectory throughout the observed period. Alabama and Arizona portrayed escalating rates early on, experiencing intermittent fluctuations, while Texas showcased a steadier climb in both cases and deaths rates. Virginia demonstrated a gradual but more controlled increase, consistently maintaining a moderate rate compared to the other states. Peaks and troughs in these rates varied, reflecting diverse responses to the pandemic, including varying public health measures, population density, and community adherence to safety protocols. The fluctuations indicate the evolving nature of the pandemic's impact on these regions, emphasizing the importance of localized strategies to curb the virus's spread."

**County-Level Analysis:**

In furthering our analysis, our team identified and delved into five counties within a chosen state that exhibited notably high rates of COVID-19 cases and deaths. Plotting weekly trends for these top five infected counties allowed us to discern patterns and peculiarities unique to these localized areas. The plotted graphs showcased raw values and log-normalized values, aiding in understanding the trajectory of infections and highlighting significant peaks.

A graph of different colored lines

Description automatically generated

A graph of different colored lines

Description automatically generated Examining the data and plotted graphs revealed initial fluctuations in cases and deaths, indicating sporadic rises rather than a consistent trend. The peak in cases was notably observed during the week of December 22 to December 28 across all counties. Higher case counts and deaths were consistently reported during November and December, except for Duplin County. Factors contributing to these peaks include the festive season, winter conditions, delayed public health measures, limited healthcare infrastructure, global connectivity, non-adherence to safety protocols, and the emergence of more transmissible variants like Delta and Gamma.

A key determinant for varying rates among counties is closely tied to population size. Counties with higher populations, such as Robeson County, consistently reported the highest cases and deaths. Conversely, counties with smaller populations, like Greene County, recorded fewer cases and deaths. This consistent pattern across all five counties underscores a direct correlation between case numbers, deaths, and the county's population size.

I can also observe that the counties pretty much follow the same pattern as their respective state does and the cases and deaths are higher in count in the week of 2020-12-22 to 2020-12-28 for all the counties, the same is observed in the case of the North Carolina state as well.

*C. Stage III (Distributions and Hypothesis Testing)*

**Distribution of COVID-19 Cases**

In this stage, a thorough analysis of COVID-19 case distributions was conducted for North Carolina State. Graphical representation revealed a log normalized distribution would be the best fit for the data, when the goodness of fit test is run on fitting normal, exponential and various others distributions. The distributions are also characterized by modality, skewness, kurtosis. The distribution statistics highlighted the center and variance and other moments, showcasing the spread and central tendency of COVID-19 cases within the state.

A graph of a line

Description automatically generated with medium confidence

Comparison with five other states exhibited distinct distribution patterns. Notably, we observe that the graphical plot of the distribution of all the 5 states pretty much looks the same, same when compared with NC as well. The frequency, that is count of cases are different the major cause for the difference in frequency may be the population of the stage (we have already analyzed this in stage II, a state with higher population has more number of cases registered when compared with the state with least number of cases), but the distribution plots are alike, that is the peak is observed at the higher range of cases. The modality of all the six states (5 states compared and NC state) is also the same, Multimodal which implies that the data has multiple modes, several values in the data occur with relatively higher frequency.

A screenshot of a graph

Description automatically generated

The probability mass function plotted for North Carolina in the above image shows that, the probability of each possible outcome occurring within a discrete set of values, illustrating the likelihood of specific values in the covid dataset.

**Modeling COVID-19 Cases and Deaths with Poisson Distribution**

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.

A screenshot of a graph

Description automatically generated A screenshot of a graph

Description automatically generated

**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.

A table of numbers with black text

Description automatically generated

Hypothesis Formulation:

Hypothesis 1:

Null Hypothesis (H0): No significant correlation exists between COVID-19 cases and Democratic (DEM\_votes) votes.

Alternative Hypothesis (H1): A significant correlation exists between COVID-19 cases and Democratic votes.

Reasoning: The correlation coefficient of -0.042656 implies a weak negative relationship. The value's proximity to zero suggests an almost negligible correlation between COVID-19 cases and Democratic votes. The weak negative correlation doesn't support a conclusive relationship between these variables.

Hypothesis 2:

Null Hypothesis (H0): There is no significant correlation between COVID-19 cases and Republican (REP\_votes) votes.

Alternative Hypothesis (H1): A significant correlation exists between COVID-19 cases and Republican votes.

Reasoning: With a correlation coefficient of -0.378338, there's a negative correlation, albeit not strong. The moderate negative correlation suggests a tendency for Republican votes to decrease with an increase in COVID-19 cases, though the relationship isn't highly pronounced.

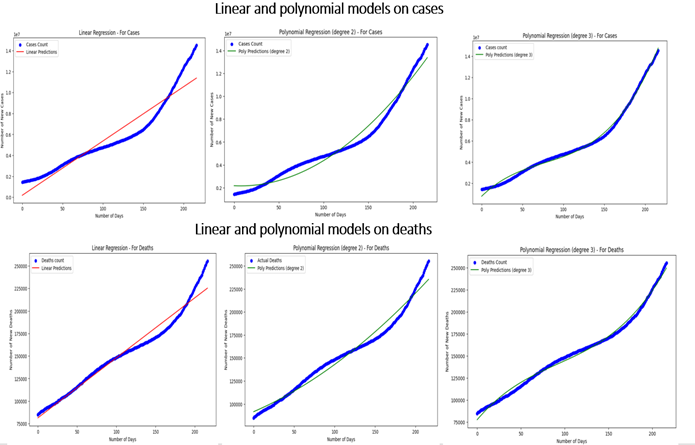
Hypothesis 3:

Null Hypothesis (H0): No significant correlation exists between COVID-19 cases and Libertarian (LIB\_votes) votes.

Alternative Hypothesis (H1): A significant correlation exists between COVID-19 cases and Libertarian votes.

Reasoning: A correlation coefficient of -0.649093 indicates a moderate to strong negative relationship. The robust negative correlation signifies a more notable tendency for Libertarian votes to decrease as COVID-19 cases increase, exhibiting a comparatively stronger association than the previous hypotheses.

*D. Stage IV (Basic Machine Learning)*



Regression models to compare trends of North Carolina

A group of graphs showing different types of data

Description automatically generated with medium confidence

**Linear and Non-Linear Regression Models for US Cases and Deaths:**

Utilizing data from June 1, 2020, to January 3, 2021, for US infections, Linear and Polynomial Regression models were constructed. The X-axis represents the number of days, while the Y-axis indicates the count of new cases and deaths.

**Root Mean Square Error (RMSE) Analysis:**

RMSE for both linear and non-linear models was computed to evaluate predictive accuracy. RMSE quantifies the differences between predicted and observed values.

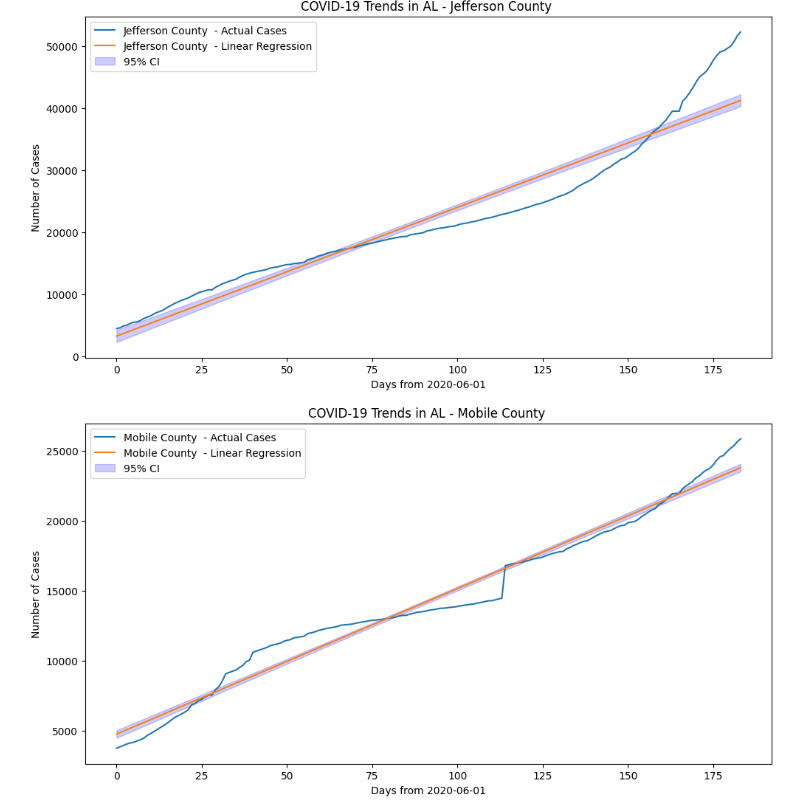
Bias-Variance Tradeoff Discussion: Assessing the models' performance, consideration was given to

**Trend Line and 1-Week Forecast:**

Trend lines were plotted based on historical data, demonstrating trends in cases and deaths.

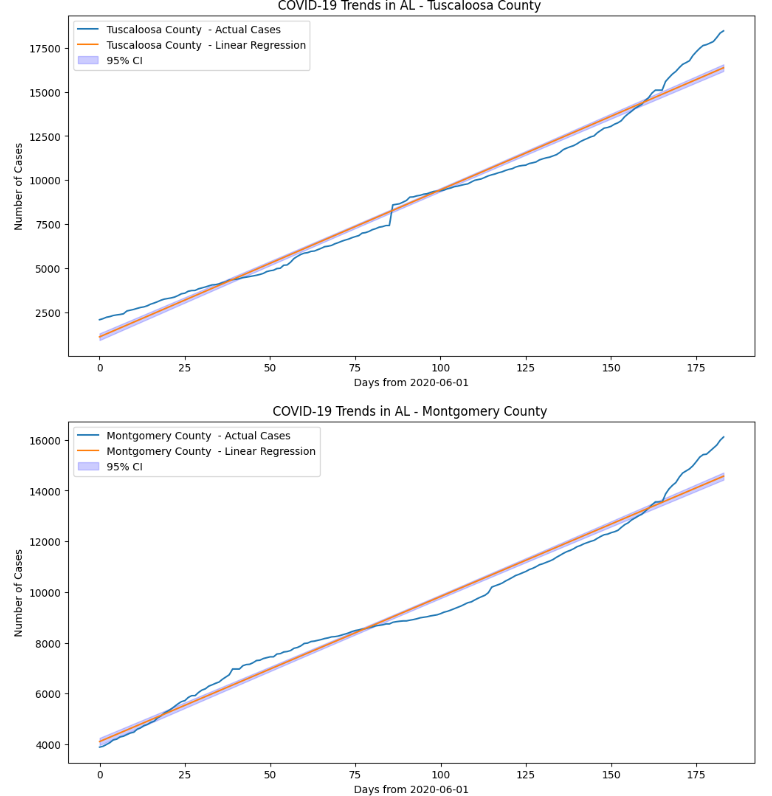
Forecasts for the following week (January 4 - January 10, 2021) were generated and overlaid on the existing trend lines to predict the trajectory of COVID-19 cases and deaths.

The polynomial regression with degree 3 best fits the data when compared with other regression techniques, came to this conlusion by observing the plots and computing the RMSE (root mean square error) values.



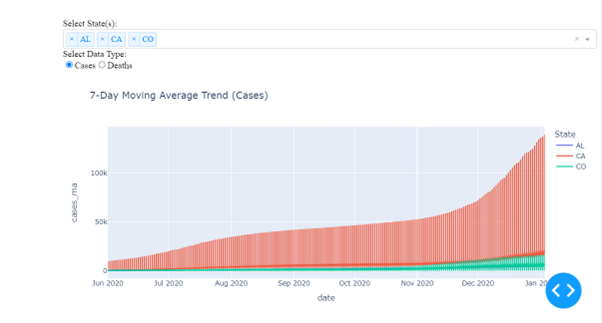
A graph of a covid-19 virus

Description automatically generated



*E. Stage V (Dashboard)*

**Covid 19 Data Analysis Dashboard**



Cases trend using moving average for Single State

A graph showing the average rate

Description automatically generated with medium confidence

Deaths trend using moving average for Single State

A graph showing the average trend

Description automatically generated

Deaths trend using moving average for multiple states

A graph showing the average trend

Description automatically generated with medium confidence

Cases trend using moving average for multiple states.

A graph showing a number of different colored lines

Description automatically generated with medium confidence

*F. Conclusion*

The culmination of our project, spanning multiple stages, involved a holistic exploration of COVID-19 trends, employing diverse methodologies encompassing statistical analyses, machine learning models, and hypothesis testing. This rigorous approach facilitated a deeper understanding of the pandemic's progression and its correlation with various socio-political factors.

Insights from Data Analysis:

Stage I & II:

Initial stages focused on data exploration, revealing patterns and correlations between COVID-19 cases/deaths and factors like time, geography, and socio-economic variables. This unearthed key insights into the outbreak's dynamics.

Stage III:

Statistical models, such as correlation and hypothesis testing, were leveraged to discern relationships between COVID-19 metrics and enrichment data, shedding light on potential correlations between variables like political affiliations and infection rates.

Utilizing Machine Learning & Predictive Models:

Stage IV:

The application of regression models, both linear and non-linear, enabled the prediction of COVID-19 trends in the US and across individual states. These models not only forecasted future infection and mortality rates but also facilitated comparisons with other countries' trends.

Impact and Significance:

The analysis conducted throughout the project provides actionable insights crucial for public health policies, resource allocation, and risk assessment. Understanding correlations between various factors and the pandemic's trajectory aids in informed decision-making and strategy formulation.

Key Learnings:

We learned that COVID-19's spread and impact are multifaceted, influenced by a multitude of factors including social behaviors, healthcare infrastructure, political environments, environmet condition, and geographic variables.

The bias-variance tradeoff in modeling highlighted the need for balanced model complexity to ensure accurate predictions without overfitting or underfitting.

Acknowledgement

Our deepest gratitude to all team members for their unwavering dedication and collaboration throughout this project. Additionally, heartfelt thanks to all sources providing data, guidance, and support, contributing significantly to the success of our endeavor.

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

REFERENCES

[1] How to write a good README for your GitHub project? (https://bulldogjob.com/news/449-how-to-write-a-good-readmefor-your-github-project)

[2] What is a data dictionary? A simple and thorough overview. (https://analystanswers.com/what-is-a-data-dictionary-a-simplethorough-overview/)

[3] Our World Data

(https://analystanswers.com/what-is-a-data-dictionary-a-simplethorough-overview/)

[4] How to plot a Confidence Interval in Python.

(https://www.statology.org/plot-confidence-interval-python/)

[5] Plotly

(https://plotly.com/)

[6] Plotly- Dash (https://plotly.com/dash/) [13] Getting started with Dash.

(https://www.youtube.com/watch?v=hSPmj7mK6ng)

[7] Introducing JupyterDash

(https://medium.com/plotly/introducing-jupyterdash-

811f1f57c02e)