

SENTIMENTAL ANALYSIS OF COVID 19 PANDEMIC (USA)

Class: CIS 635

Team : Group -3

Sarika Vemana - vemanas@mail.gvsu.edu

Brenda Ondieki - ondiekib@mail.gvsu.edu

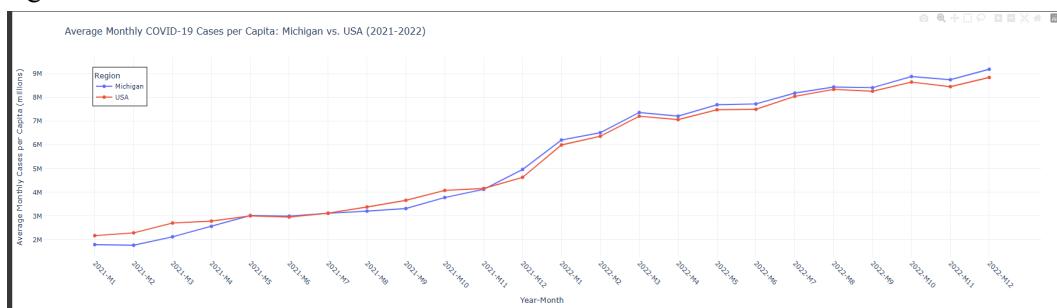
Datasets :

- US state vaccinations:
https://raw.githubusercontent.com/owid/covid-19-data/refs/heads/master/public/data/vaccinations/us_state_vaccinations.csv
- For covid tweets : <https://www.kaggle.com/datasets/arunavakraborty/covid19-twitter-dataset>
- For covid deaths: <https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/>
- For covid cases: <https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/>

AIM: To analyze the case and death trends of COVID-19 in the United States, focusing on Sentiment analysis and perform modelling on predicting cases and deaths for the year 2022.

EDA Part -1

Fig:1



The chart visualizes the average monthly COVID-19 cases per capita for Michigan and the USA, with separate lines for each region and markers for data points.

Trend Analysis: Case rates for Michigan and the USA align closely until late 2021, after which Michigan starts to lag slightly behind the national trend.

Spike in Cases: The late 2021 and early 2022 spikes are associated with the emergence of the Omicron variant, which was more transmissible but less lethal than previous strains.

Drop in Cases: The decrease in mid-2022 is likely due to widespread immunity (vaccination and prior infections) and seasonality factors.

Fig:2



The chart visualizes the average monthly COVID-19 cases per capita for Michigan and the USA, with separate lines for each region and markers for data points.

Trend Analysis: The death rates in Michigan remain consistently higher than the national average throughout 2021–2022. There is a notable increase in deaths around late 2021 (October–December) for both Michigan and the USA.

Spike in Deaths: The late 2021 spike corresponds to the Delta variant wave, which was highly transmissible and caused severe outcomes, particularly among unvaccinated populations.

Drop in Deaths: Following early 2022, death rates begin to stabilize, likely due to increased vaccination rates, booster campaigns, and improved treatments such as antivirals.

Fig:3

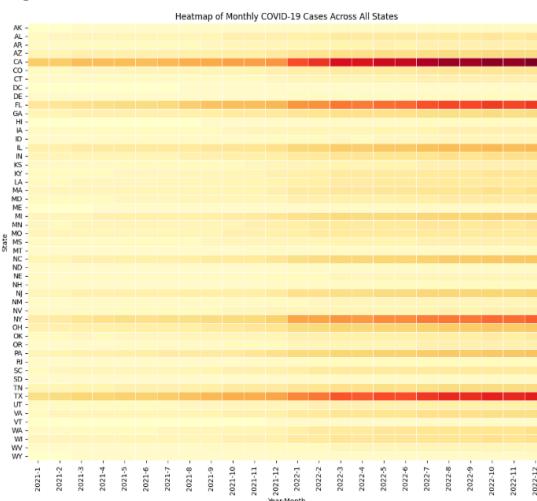
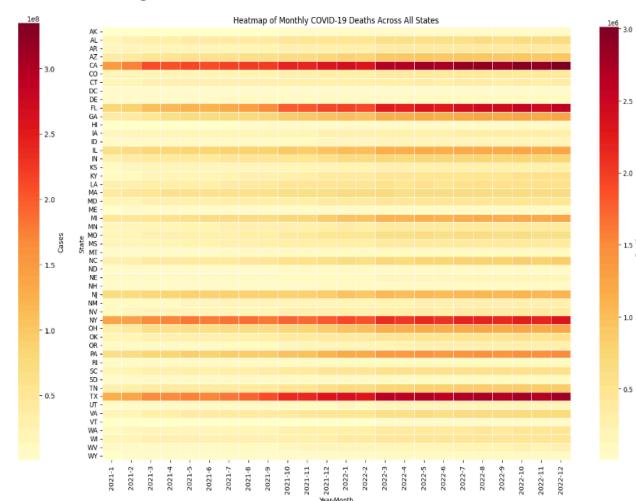
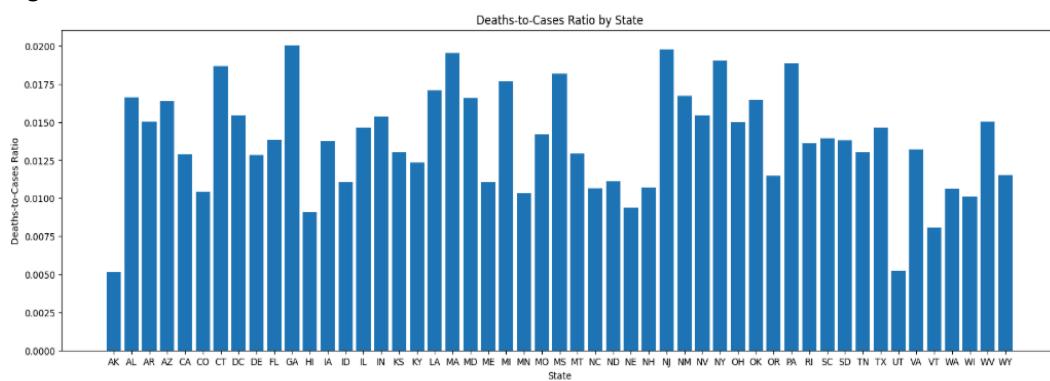


Fig:4



A heatmap to visualize the monthly total COVID-19 cases & deaths across all U.S. states to identify the condition of the USA.

Fig:5



A bar chart is plotted to visualize this ratio for each state, showing variations in mortality relative to reported cases. States like Florida and Georgia exhibit higher ratios, potentially due to older populations and healthcare system strain during surges. States like Vermont and Hawaii have lower ratios, which may be attributed to better public health measures, higher vaccination rates, and younger populations.

Fig :6

Fig:7

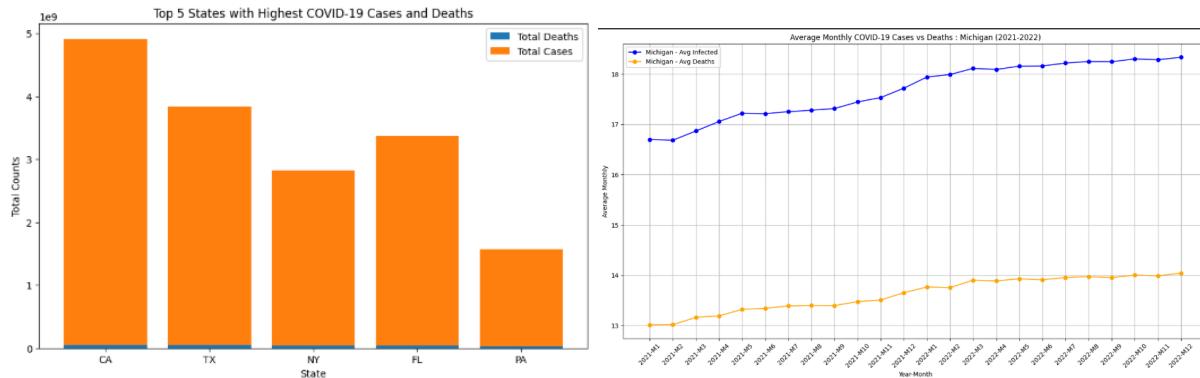


Fig:6 A stacked bar chart to compare the total deaths and cases in these states, with deaths forming the lower part of each bar and cases stacked on top, providing a clear visual comparison.

California, Texas, New York, Florida, and Pennsylvania have had a high number of COVID-19 cases and deaths due to their huge populations, urban density, and early virus epicenters, such as New York.

State-specific rules, such as Florida's early reopening and Texas' uneven enforcement of prohibitions, contributed to the increase. Furthermore, these states experienced severe healthcare system strain, particularly during waves caused by highly transmissible variations. Population statistics, particularly among elderly people, increased sensitivity to severe outcomes.

Fig:7 The blue line shows a steady increase in average monthly infections, peaking around late 2021 and early 2022, likely due to the Omicron variant, which caused widespread but generally less severe illness than Delta. The orange line also rises, but at a slower rate, reflecting improved treatments, widespread vaccination, and the less severe nature of Omicron. The widening gap between cases and deaths highlights the effectiveness of vaccines and treatments in reducing mortality despite the rise in infections.

Fig:8

Fig:9

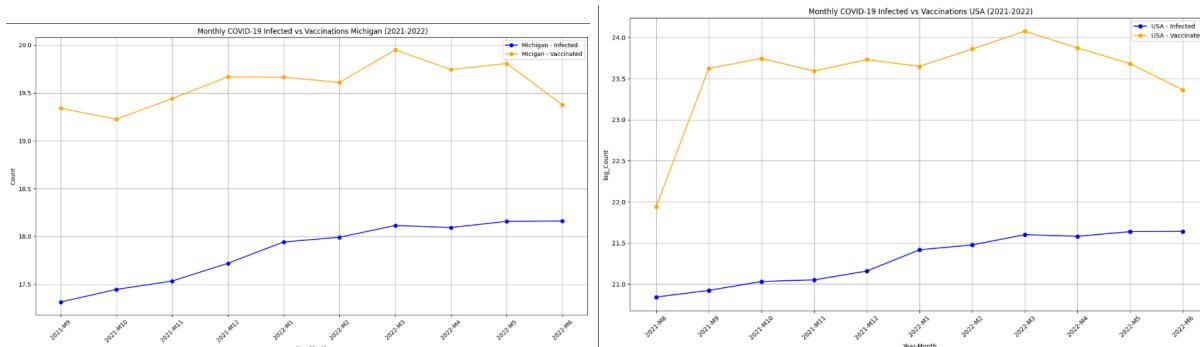
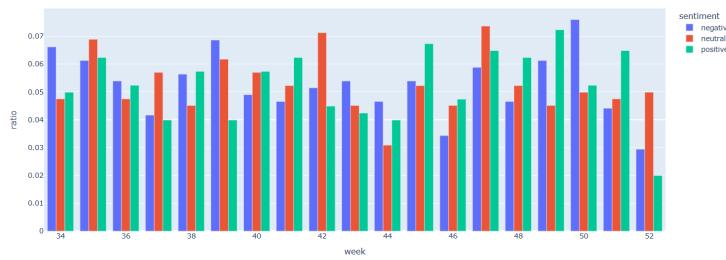


Fig:8 The vaccination curve (orange) starts high, dips slightly in late 2021, and stabilizes until mid-2022 before declining, reflecting initial vaccination efforts followed by a drop in demand. Infections (blue) gradually rise, with notable spikes in late 2021 and early 2022, corresponding to new variant waves. The relatively stable vaccination rate in Michigan helped reduce the severity of cases, but could not fully prevent infection surges.

Fig:9 The orange line depicts the monthly vaccination count, which peaks in August 2021, continues high, and then drops dramatically after May 2022, potentially due to vaccine saturation or decreased demand. Despite vaccine efforts, infection rates (blue) are constantly increasing, most likely due to variations like as Delta and Omicron. The disparity between rising infections and declining vaccination rates shows that the virus continued to propagate in vaccinated populations, possibly due to fading protection or immunological escape variations.

EDA Part-2 (Sentimental Analysis)

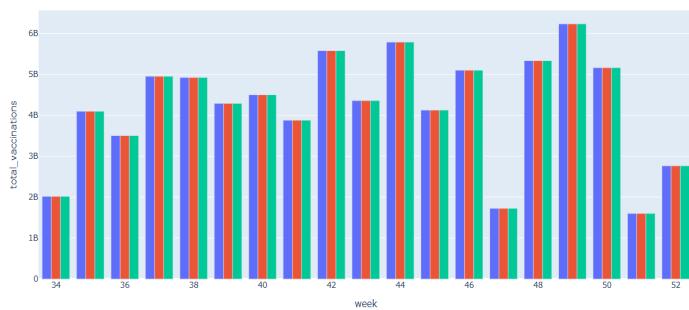
Fig 10



In fig 10, There was no significant positive or negative shifting for the weeks as most of the conversations followed the typical public discussion surrounding COVID-19 updates and policies, and the rollout of the vaccines. Positive sentiment rose towards year end by accomplishments in vaccination campaigns, holiday season, although it declined in week 52 due to Omicron variant, disruptions during the holiday travel season, and renewed restrictions. Mid-year content tone was mixed both by Delta variant fears and vaccine hesitancy; despite the vaccination, early fall saw slightly negative content sentiment and then positive sentiment due to booster campaigns and New year celebrations.

In fig 11, we used rolling average to smoothen the weekly sentiment trends in fig 10. The sentiments softened in the vicinity of weeks 45–46 on uncertainty over Delta variant as well as winter months. The positive driver of sentiment was booster campaigns and optimistic holiday expectations, which were evident in weeks 47-50, but negativity due to Omicron, disruption in travel, and renewed restrictions surfaced in week 52.

Fig 12



In fig 12, the number of vaccinations tended to rise from week 34 through 44, as the fight against the Delta variant unfolded, while the number of tweets remained balanced due to the informative nature of the topics being factual. Positive text sentiment indeed reached its highest level during weeks 44 and 48–50, with the marked slowdown in vaccination growth in week 52 due to the holidays, the spread of Omicron, and the slow pace.

In fig 13, slightly more of the tweets were neutral, which would suggest most of the tweets were strictly and purely informative regarding COVID-19 without much opinion or emotion. The positive and negative sentiments were almost balanced, the positivity was around the vaccination roll outs while negativity was observed around new variants of COVID and new lock downs restrictions.

Fig 11

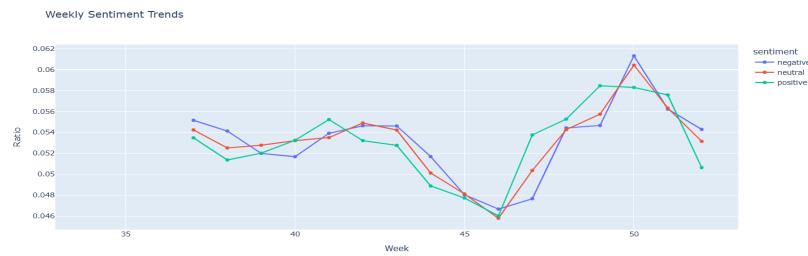
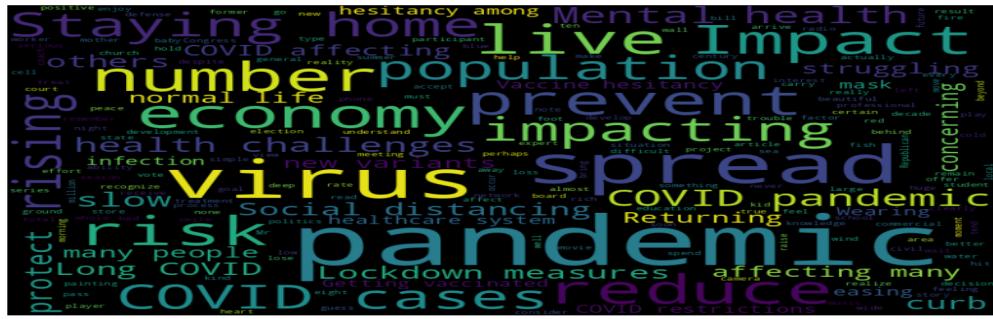


Fig 13



Fig 14

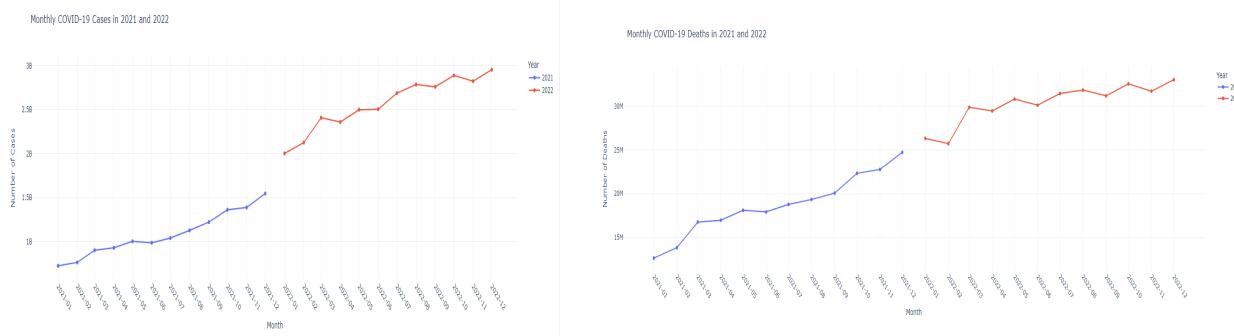


When it comes to the word cloud of COVID-19, key areas of discussions involving health and economics can be noted in fig 14 with the help of such leading terms as virus, pandemic, and economy that points at the main concerns and challenges of public health. Vaccination hesitancy and other new variants, social and lifestyle changes continue to persist and stake evidence of the ever-encompassing impact of the pandemic.

Predictive Modeling:

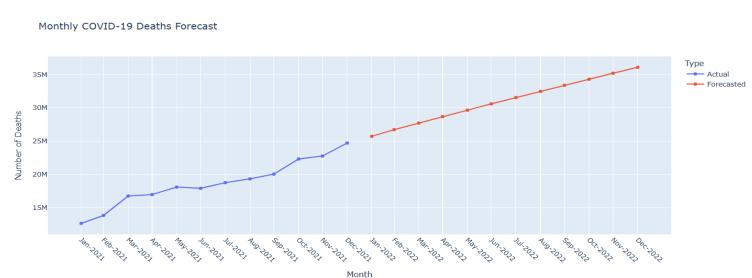
We used the Time Series ARIMA model and the Random Forest Regressor model to predict cases and deaths based on a dataset containing monthly data for 2021 and 2022 across all states. Both models were trained on 2021 data and tested on 2022 data to evaluate how accurately they could predict cases and deaths for 2022.

First, we started by visualizing the actual data for 2021 and 2022 for both cases and deaths as shown below. The figure on the left captures cases data in blue line and on the right is deaths data captured on red line.



ARIMA Forecast For Cases and Deaths:

The model uses 5 lag observations ($p=5$), 1 differencing ($d=1$) to make the data stationary, and 3 lag forecast errors ($q=3$) to bring additional accuracy to the model. Below are the charts we got after the modeling. The figure on the left, the red line shows the 2022 number of cases predictions, similarly the chart on the right shows the 2022 number of deaths predictions using the red line.

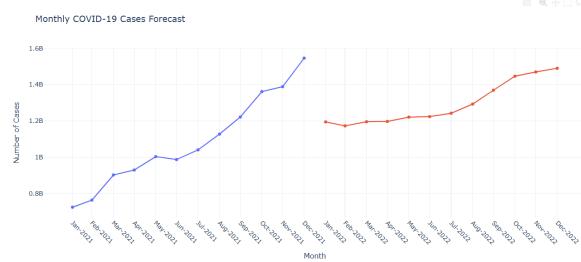


We used the walk-forward validation technique to analyze the performance of the ARIMA model, and we used root mean square error (RMSE) to assess it. The Test RMSE for cases was 15,467,997.31 and deaths was 181,367.09. The higher RMSE values are due to the large scale of data in our dataset but overall the result shows ARIMA performed quite well in time series forecasting.

Random forest Cases & Deaths:

We used the lag features to help the model to understand and predict 2022.

Cases:



Deaths:



The evaluation metrics for the Random Forest model indicate poor performance for predicting both "Cases" and "Deaths." For "Cases," the RMSE (Root Mean Squared Error) is extremely high at 1,289,054,221.84, indicating significant deviations between predicted and actual values. The R2 (R-squared) value of -18.7658 suggests the model is performing much worse than a simple mean-based prediction, explaining none of the variance in the data. Similarly, for "Deaths," the RMSE is 7,706,208.60, again showing large prediction errors, while the R2 value of -11.5299 reinforces the model's inability to capture meaningful patterns.