University of Colorado
Boulder

COVID-19 Comprehensive Study

CSPB-4502 Data Mining Project

Mohammed Alsailani

Andrew Byrnes

Collin Coakley

Gilberto Zamarron

5/1/2024

Introduction

Our project focuses on analyzing data surrounding COVID-19, including deaths, vaccination rates, and cases viewed on a per-county basis, covering dates from 1/1/2020 to 9/17/2022. Our group decided to use the COVID-19 Open Data by Google[1]. The intriguing questions we aim to answer are:

- How do local variables, such as temperature and mobility rates, affect COVID rates?
- Did high vaccination rates help mitigate the deaths and spread of the virus?
- Understand the impact of detailed weather variables on COVID-19 rates
- How do regional outbreaks occur and do them spread between states/counties?

[1] https://health.google.com/covid-19/open-data

Tools and Methodology

- Overleaf
- Git/GitHub
- Python:
  - Pandas
  - NumPy
  - Geopandas
  - Sklearn
  - Matplotlib

- Agile Method:
  - Weekly check ins
  - Dedicated time to work individually and as a group
  - Shared goals for the week
  - Helped others troubleshoot obstacles

Data Preprocessing

Data Transformation:

- Standardized incidence rate per 100,000
- Min-Max Scaling implemented

Data Repositories:

- Primary GitHub repository
- Storage of all county CSVs for analysis and a subset for testing data mining strategies.

Data Reduction:

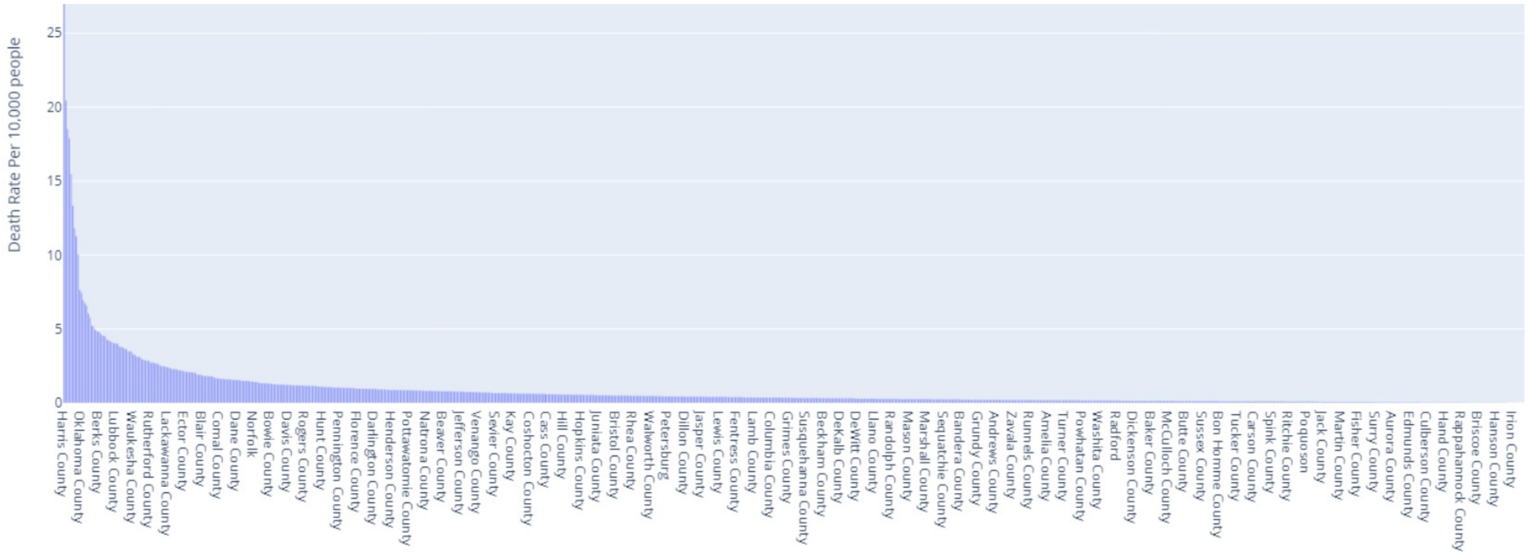
- Focus on the U.S.
- Dimension reduction

Data Cleaning

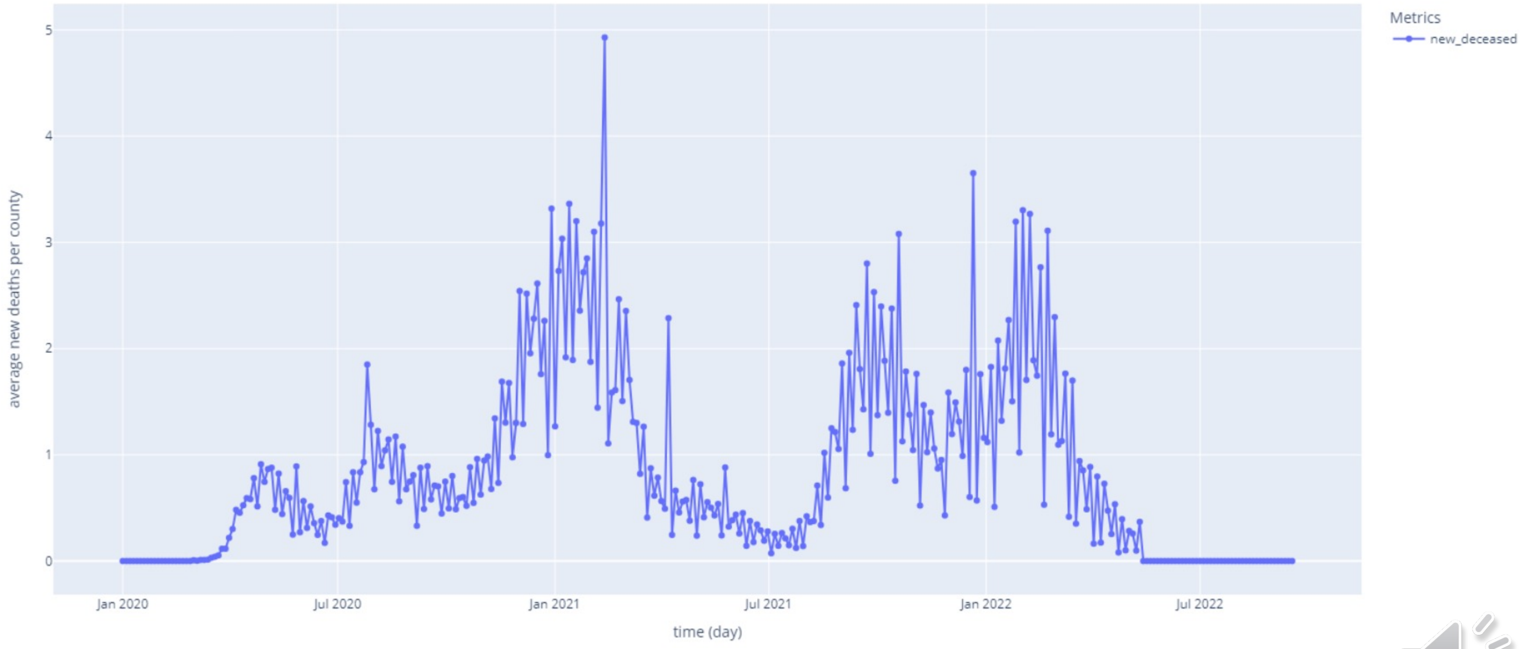
- Made data consistent amongst all files
- Excluded attributes lacking too much data

Preliminary plots (data familiarization)

- Temporal graphs
- Bar charts



3-Day County Moving Average of Death Rate Over Time"



Geospatial Analysis

- Used TIGER/Line Shapefiles to color each county to get more accurate representation of severity at a particular time interval

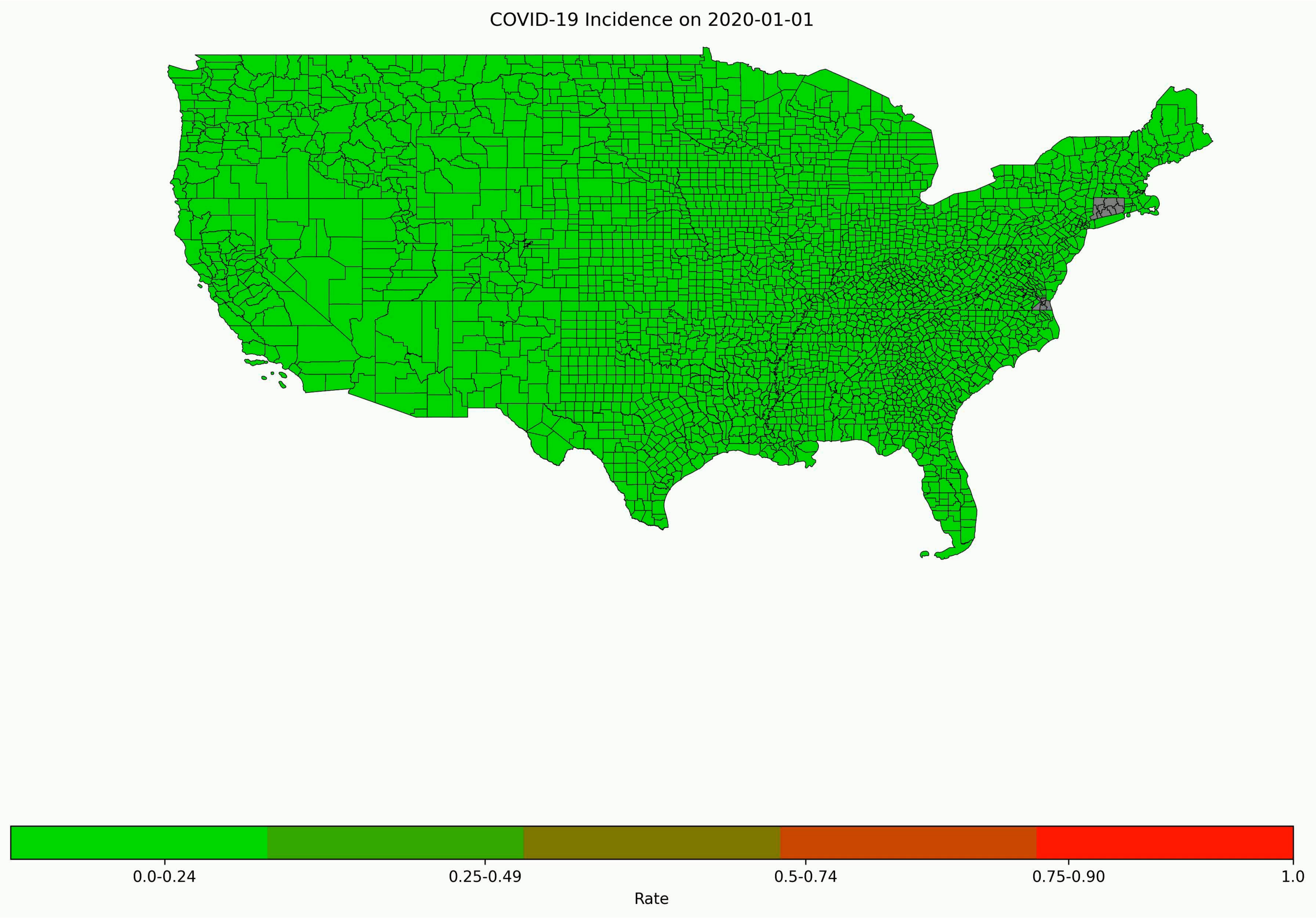
- Aggregated different timeframes ranging from 3-day intervals to 1-year intervals

- Binned to examine different scopes.

- We used the Python module Geopandas to facilitate plotting.

Geospatial Progressive Timelapse

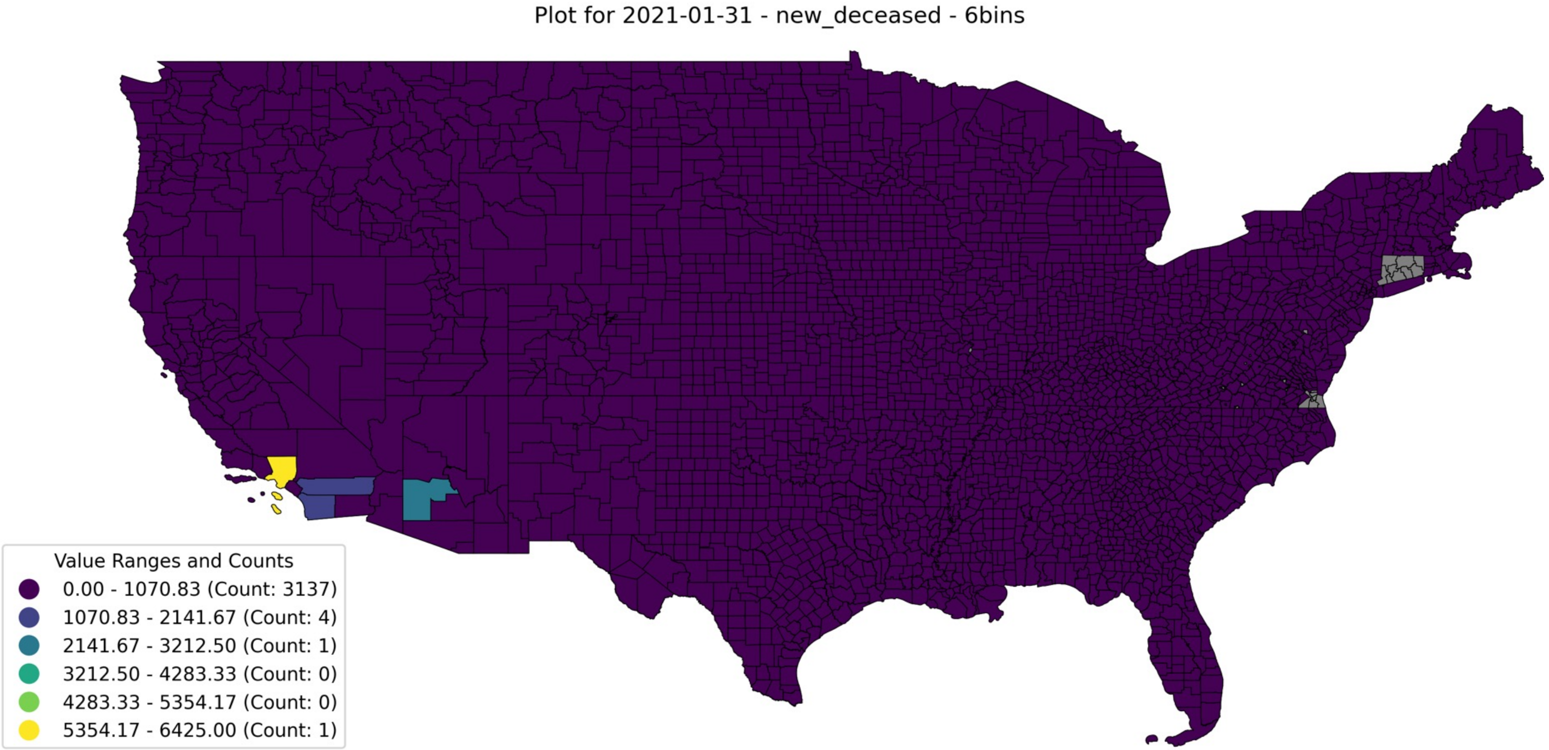
- Seem to have two waves of covid.
- Regions with outbreak seem to expand to others.
- Aggregation based on 17 days, per normalized per 100,000



Geospatial on Worst County Outbreaks

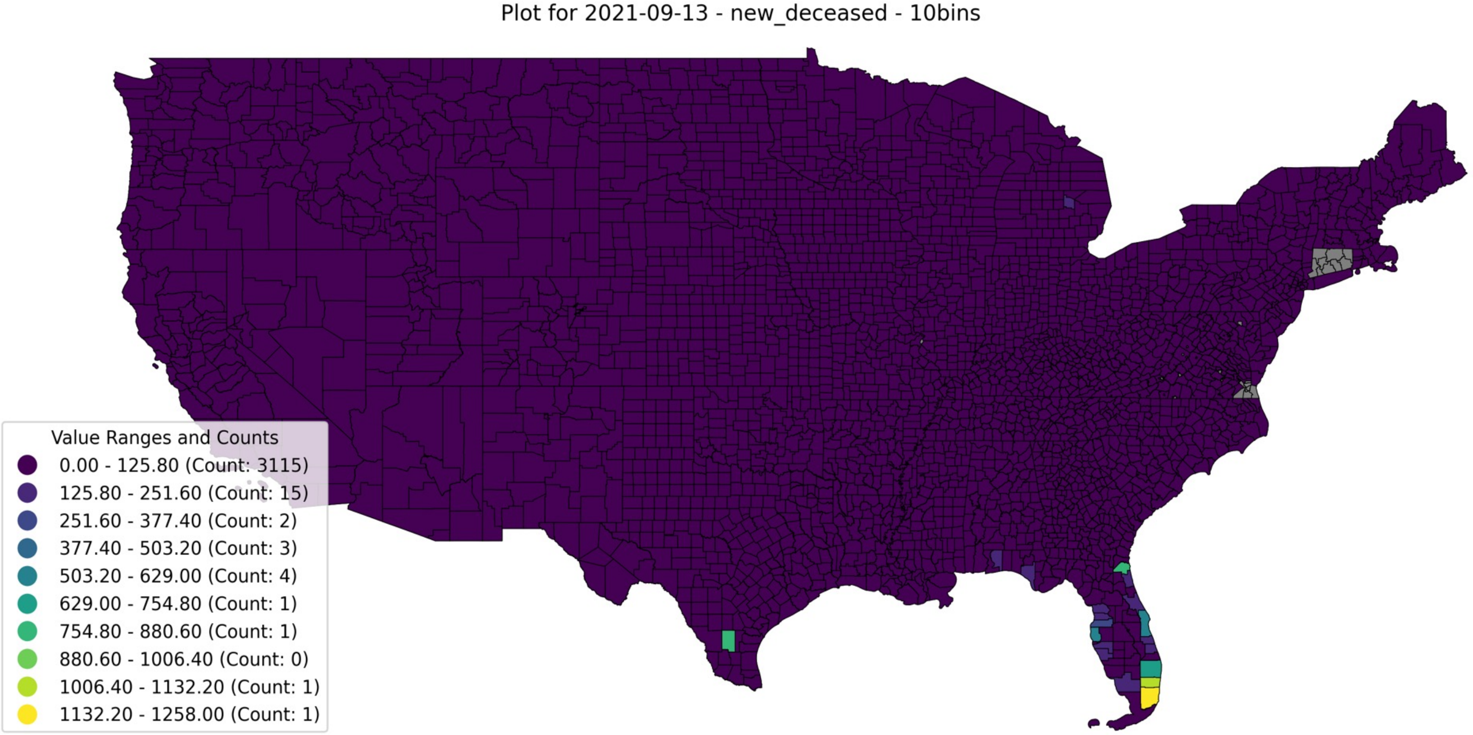
- During December 2020 through February 2021, the West Coast had a significant breakout.

- Used binning to emphasize contrast for worst counties


Plot for 2021-01-31 - new_deceased - 6bins
Value Ranges and Counts
● 0.00 - 1070.83 (Count: 3137)
● 1070.83 - 2141.67 (Count: 4)
● 2141.67 - 3212.50 (Count: 1)
● 3212.50 - 4283.33 (Count: 0)
● 4283.33 - 5354.17 (Count: 0)
● 5354.17 - 6425.00 (Count: 1)

Geospatial on Worst County Outbreaks (Cont.)

- Florida experienced a similar outbreak that was significantly more severe than other counties.
- Different aggregation timeframes led to identifying different outbreaks.

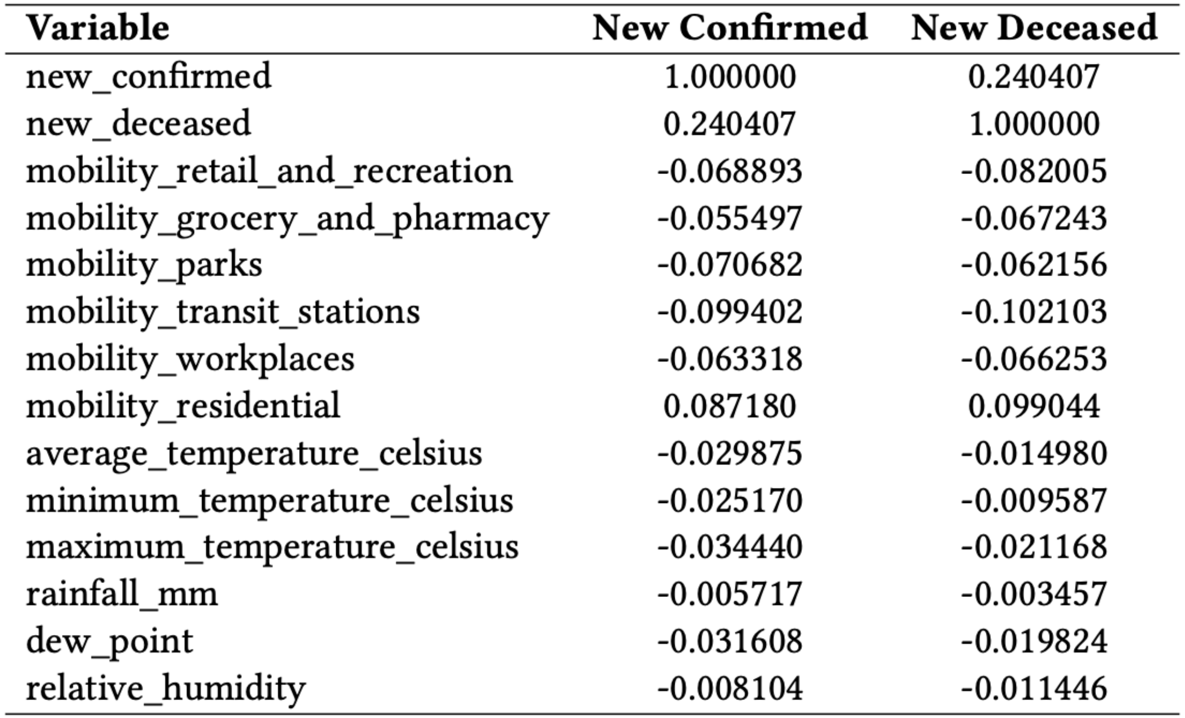


Plot for 2021-09-13 - new_deceased - 10bins

Value Ranges and Counts
● 0.00 - 125.80 (Count: 3115)
● 125.80 - 251.60 (Count: 15)
● 251.60 - 377.40 (Count: 2)
● 377.40 - 503.20 (Count: 3)
● 503.20 - 629.00 (Count: 4)
● 629.00 - 754.80 (Count: 1)
● 754.80 - 880.60 (Count: 1)
● 880.60 - 1006.40 (Count: 0)
● 1006.40 - 1132.20 (Count: 1)
● 1132.20 - 1258.00 (Count: 1)

Correlation Analysis

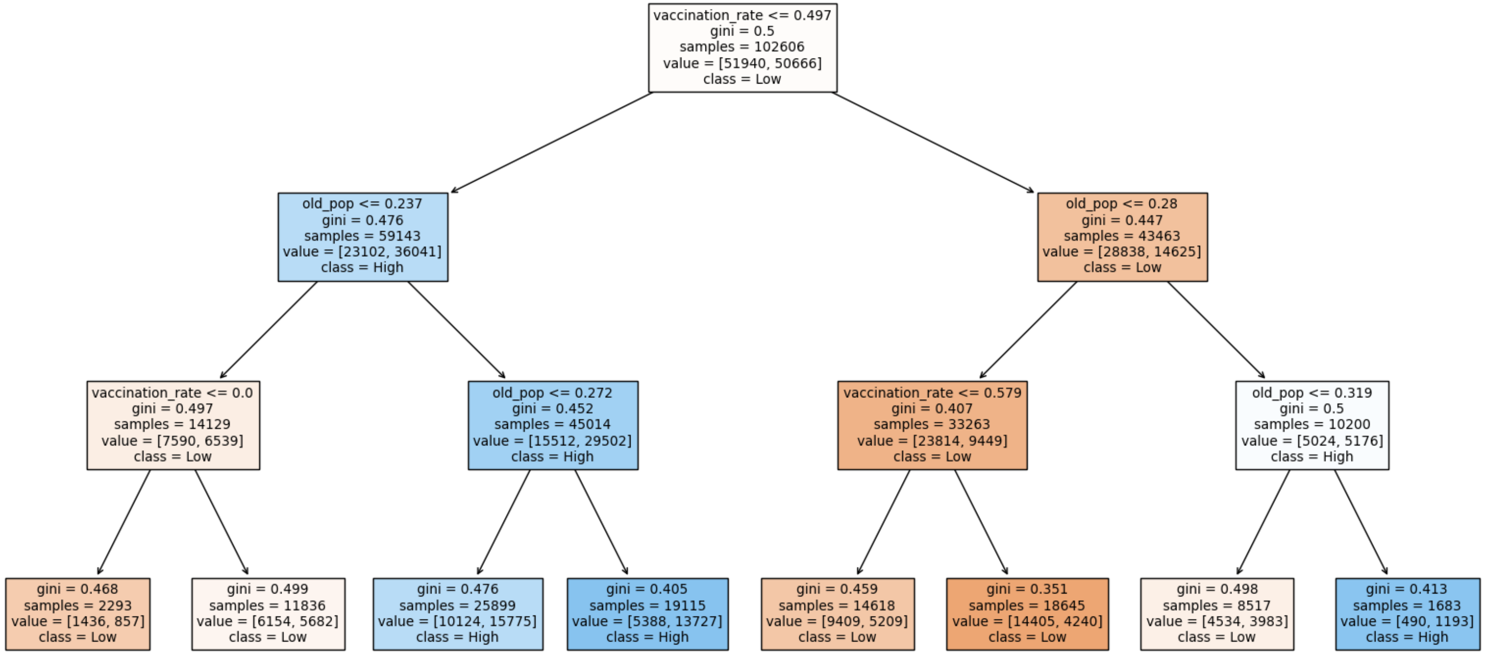
- Pearson Correlation Coefficients: Linear Relation.
- Moderate positive correlation between the new confirmed cases and new deceased cases.
- Mobility data show weak negative correlations with new confirmed cases
- Residential mobility shows a weak positive correlation with new confirmed cases.
- Temperature shows weak negative correlations with newly confirmed cases.



Variable	New Confirmed	New Deceased
new_confirmed	1.000000	0.240407
new_deceased	0.240407	1.000000
mobility_retail_and_recreation	-0.068893	-0.082005
mobility_grocery_and_pharmacy	-0.055497	-0.067243
mobility_parks	-0.070682	-0.062156
mobility_transit_stations	-0.099402	-0.102103
mobility_workplaces	-0.063318	-0.066253
mobility_residential	0.087180	0.099044
average_temperature_celsius	-0.029875	-0.014980
minimum_temperature_celsius	-0.025170	-0.009587
maximum_temperature_celsius	-0.034440	-0.021168
rainfall_mm	-0.005717	-0.003457
dew_point	-0.031608	-0.019824
relative_humidity	-0.008104	-0.011446

Impact of Vaccination on COVID-19 Death Rates

- Decision Tree Classification
  - Feature variables: Vaccination Rate and Old Population Rate
  - Target variable: Death Rate
- Increased vaccination rate in counties with higher older population rates decreases the mortality rate in the county.
- This shows a correlation between vaccines and lower mortality rates for older populations.



Comparative Regression Analysis with and without Lagged Variables - Ordinary Least Squares (OLS)

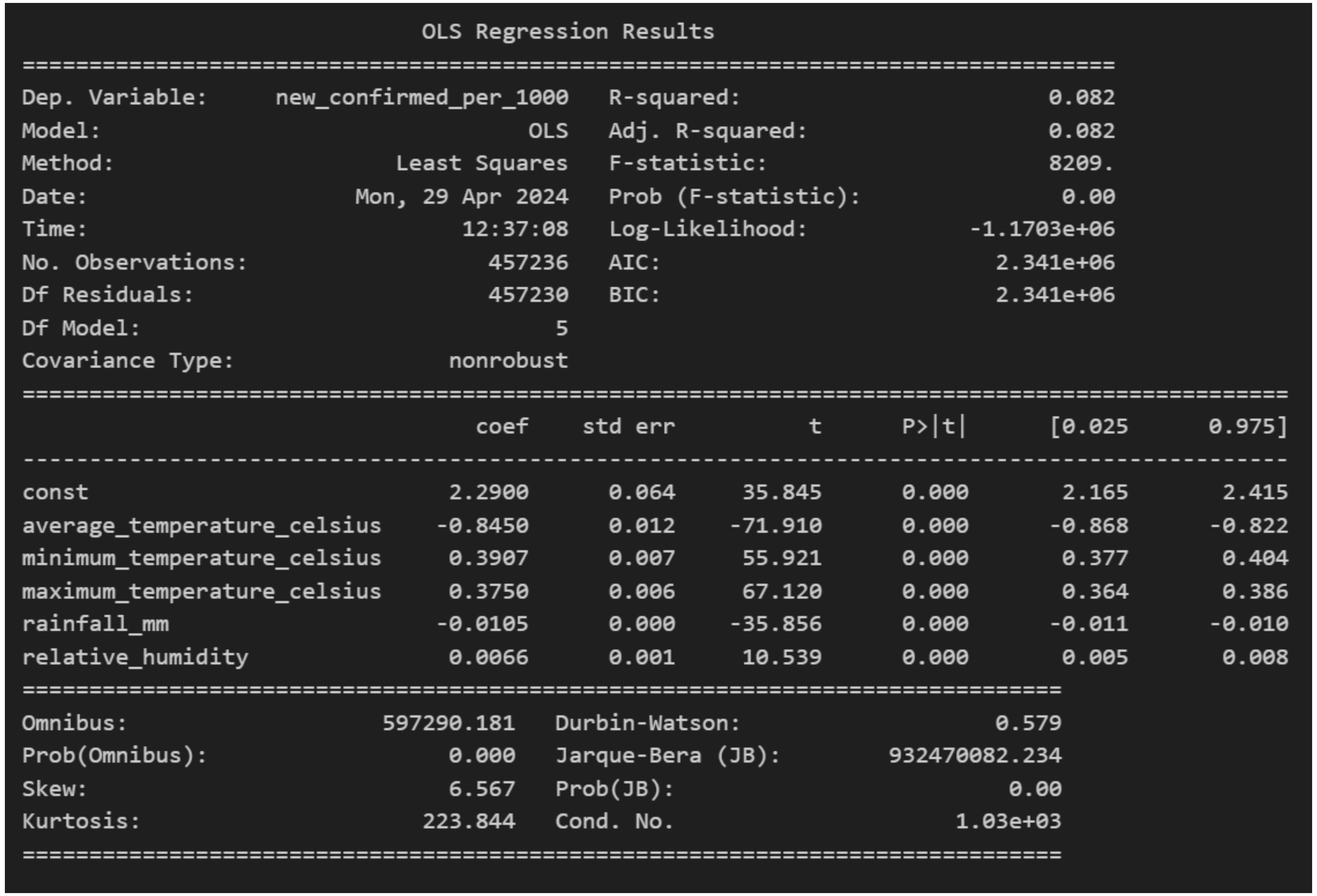
Objective: Examine the effectiveness of incorporating lagged variables into regression models for predicting COVID-19 cases.

Methodology: Two OLS regression models were developed; one with current environmental data and another enhanced with lagged environmental data from the previous week.

Rationale: To capture the delayed effects of environmental factors on COVID-19 transmission rates.

Initial Model without Lagged Variables

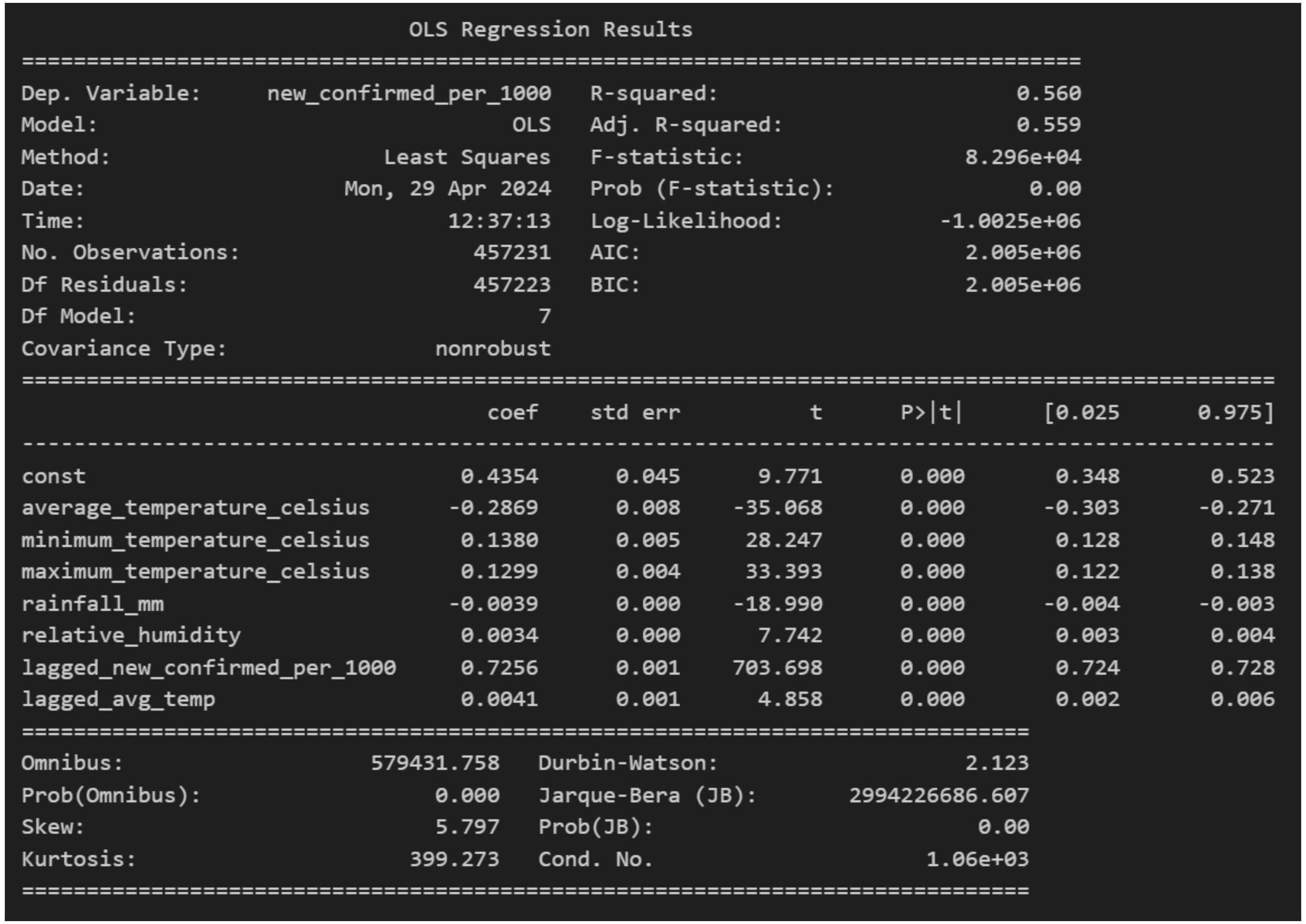
- **Variables Used**: Average temperature, minimum temperature, maximum temperature, rainfall, and relative humidity.
- Results Summary:
    - R-squared: 0.082, indicating that about 8.2% of the variability in new confirmed COVID-19 cases per 1000 people is explained by the model.
    - Significant predictors: All initial environmental factors had a noticeable impact on COVID-19 case predictions.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     new_confirmed_per_1000   R-squared:                       0.082
Model:                                OLS   Adj. R-squared:                  0.082
Method:                     Least Squares   F-statistic:                     8209.
Date:                    Mon, 29 Apr 2024   Prob (F-statistic):               0.00
Time:                            12:37:08   Log-Likelihood:            -1.1703e+06
No. Observations:                  457236   AIC:                         2.341e+06
Df Residuals:                      457230   BIC:                         2.341e+06
Df Model:                               5
Covariance Type:                nonrobust
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                          2.2900      0.064     35.845      0.000       2.165       2.415
average_temperature_celsius   -0.8450      0.012    -71.910      0.000      -0.868      -0.822
minimum_temperature_celsius    0.3907      0.007     55.921      0.000       0.377       0.404
maximum_temperature_celsius    0.3750      0.006     67.120      0.000       0.364       0.386
rainfall_mm                   -0.0105      0.000    -35.856      0.000      -0.011      -0.010
relative_humidity              0.0066      0.001     10.539      0.000       0.005       0.008
==============================================================================
Omnibus:                   597290.181   Durbin-Watson:                   0.579
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        932470082.234
Skew:                           6.567   Prob(JB):                         0.00
Kurtosis:                     223.844   Cond. No.                     1.03e+03
==============================================================================
```

Enhanced Model with Lagged Variables

- **New Variables:** Lagged confirmed cases and lagged average temperature were added.
- Results Summary:
  - R-squared Improved to 0.560, showing that 56% of the variability is now explained by the model, significantly enhancing predictive accuracy.
  - F-statistic: Increased to approximately 82,960, underscoring the model’s robustness.

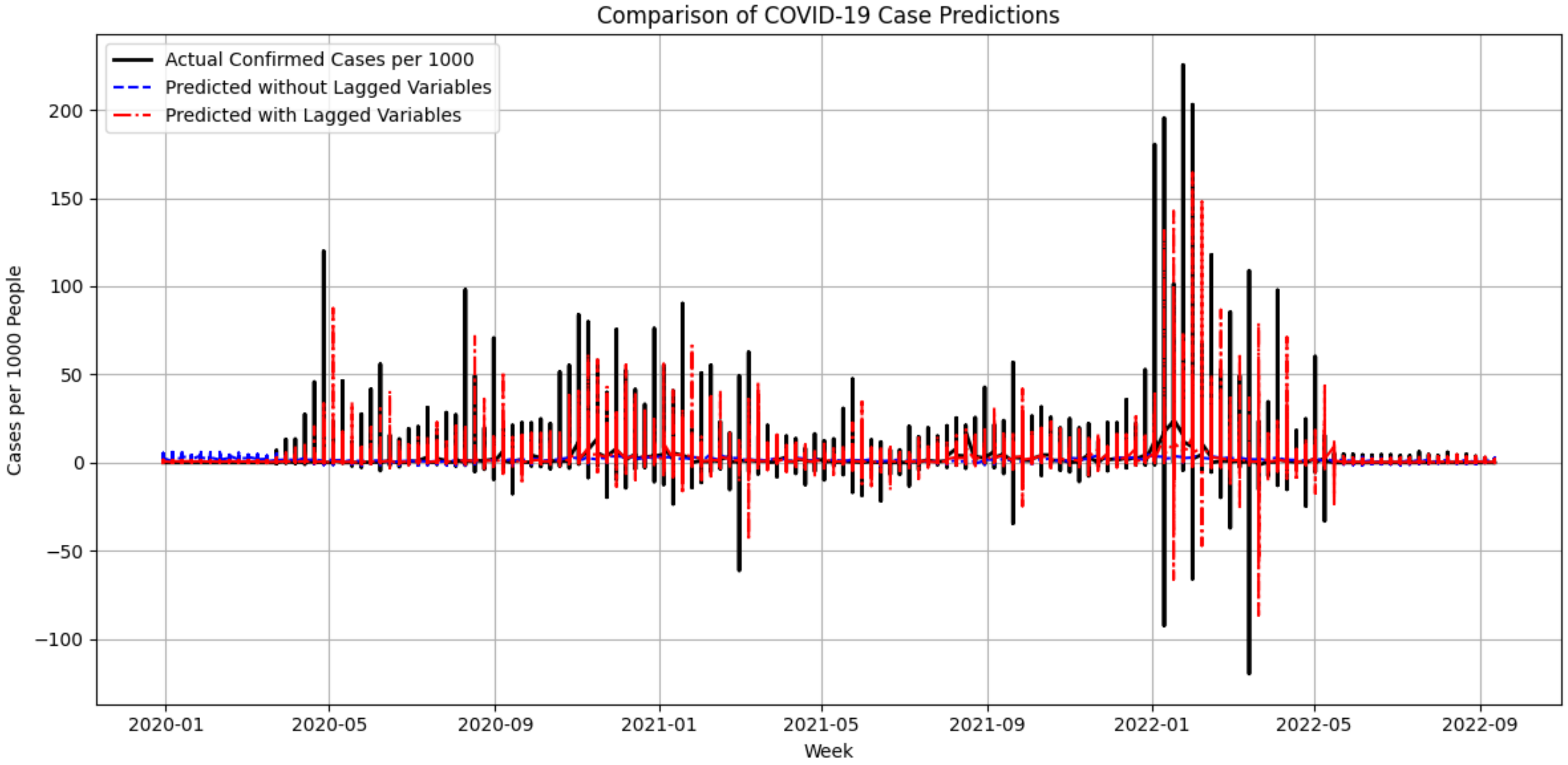


```
                              OLS Regression Results
=================================================================================
Dep. Variable:     new_confirmed_per_1000   R-squared:                      0.560
Model:                                OLS   Adj. R-squared:                 0.559
Method:                     Least Squares   F-statistic:                8.296e+04
Date:                    Mon, 29 Apr 2024   Prob (F-statistic):              0.00
Time:                            12:37:13   Log-Likelihood:           -1.0025e+06
No. Observations:                  457231   AIC:                        2.005e+06
Df Residuals:                      457223   BIC:                        2.005e+06
Df Model:                               7
Covariance Type:                nonrobust
==============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                           0.4354      0.045      9.771      0.000       0.348       0.523
average_temperature_celsius    -0.2869      0.008    -35.068      0.000      -0.303      -0.271
minimum_temperature_celsius     0.1380      0.005     28.247      0.000       0.128       0.148
maximum_temperature_celsius     0.1299      0.004     33.393      0.000       0.122       0.138
rainfall_mm                    -0.0039      0.000    -18.990      0.000      -0.004      -0.003
relative_humidity               0.0034      0.000      7.742      0.000       0.003       0.004
lagged_new_confirmed_per_1000   0.7256      0.001    703.698      0.000       0.724       0.728
lagged_avg_temp                 0.0041      0.001      4.858      0.000       0.002       0.006
==============================================================================
Omnibus:                   579431.758   Durbin-Watson:                   2.123
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       2994226686.607
Skew:                           5.797   Prob(JB):                         0.00
Kurtosis:                     399.273   Cond. No.                     1.06e+03
==============================================================================
```

Comparative Analysis and Model Improvement

- **Improvements Noted:**
  - Adjusted R-squared increased dramatically from 0.082 in the initial model to 0.560 in the enhanced model.
  - The Durbin-Watson statistic improved, indicating reduced autocorrelation among residuals.
- **Diagnostics Issues:**
  - Omnibus and Jarque-Bera tests still indicate non-normal distribution of residuals, suggesting the presence of outliers or model misspecification.
  - Next Steps: Consideration of more sophisticated time-series models or transformations of the dependent variable to further refine model accuracy.

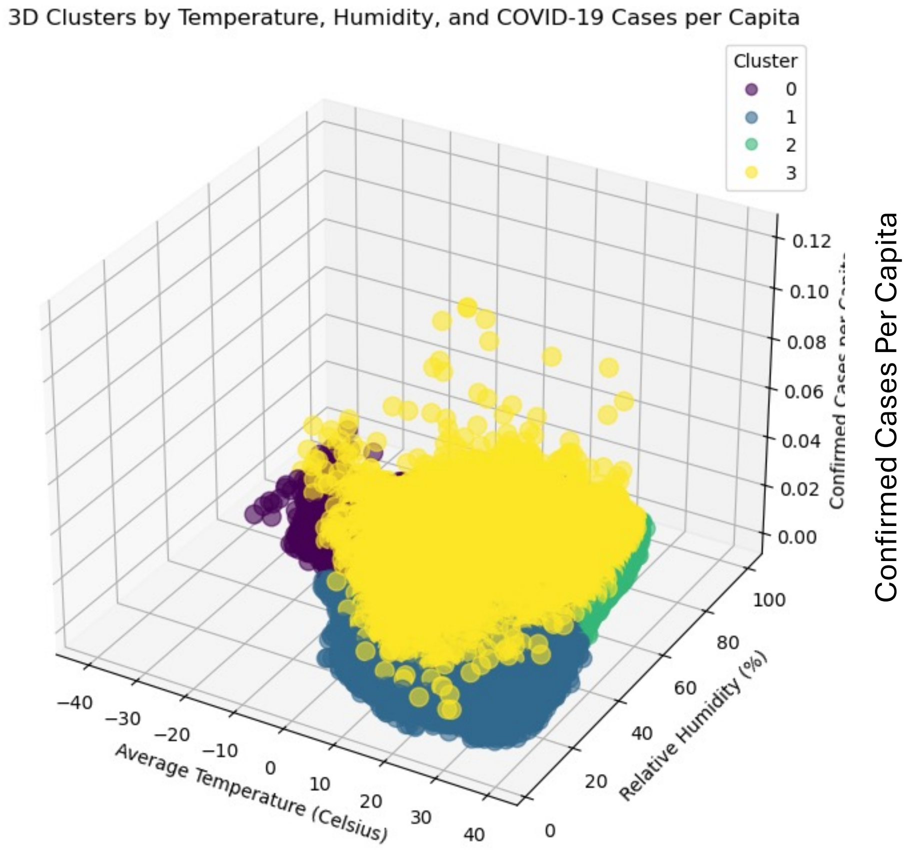
Visual Comparison and Conclusion



- **Conclusion:** The inclusion of lagged variables significantly improves the model's ability to predict new COVID-19 cases, highlighting the importance of considering temporal dynamics in epidemiological modeling

Impact of Weather on COVID-19 Rates

- k-means clustering
  - 4 clusters based in inertia
- Decrease in the transmissibility of COVID-19 at temperatures exceeding 30°C, under 30°C, or under a 20% relative humidity.



Knowledge Gained and Application

- Infections seem to expand from one region to others.
- There appeared to be at least two waves significant outbreaks.
- Higher vaccination rates are related to lower COVID-19 death rates.
- COVID-19 cases decrease in extreme climate.