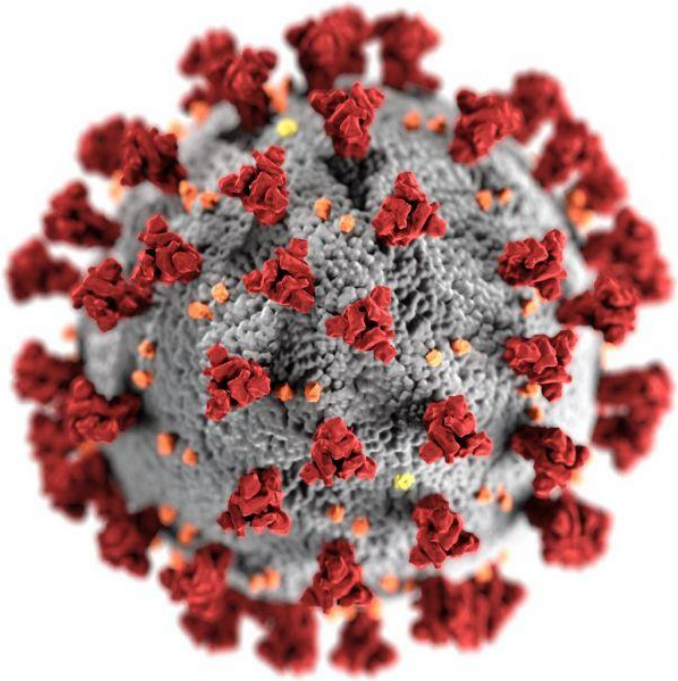
United States COVID-19 Pandemic Cases

Exploratory Data Analysis and Time Series Forecasting



Main Objective of Analysis

The COVID-19 pandemic began in late 2019 and upended the lives of the global population. Its impact lessened with widespread vaccination and other public health measures. To this day, COVID-19 still circulates, but it is more manageable in society and some countries have downgraded it to an endemic, much like the flu. This analysis focuses on the case counts in the United States from the beginning of the pandemic until 2022 by state and county. Part way during this time period, vaccines became publicly available. Also notable was the infection rates during surges, when case counts became especially high. Exploratory data analysis was performed to observe infection rates in various states in the US as well as different counties in California. A forecasting model was created for Los Angeles County to predict infection rates.

Description of Dataset

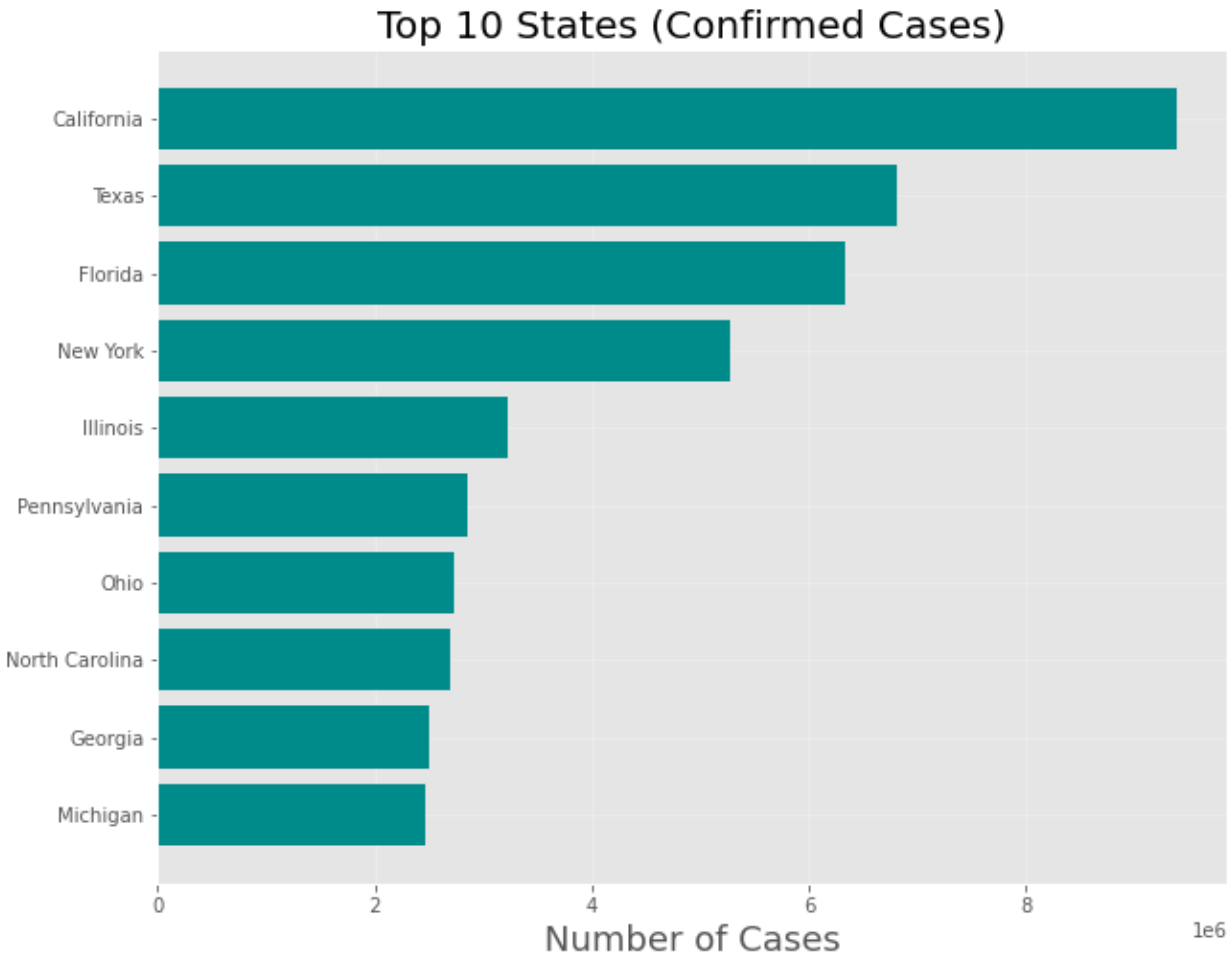
Data on case counts was retrieved online from the New York Times and updated daily from the beginning of the pandemic until May 13, 2022 after which the data was no longer available. The dataset contained the following attributes:

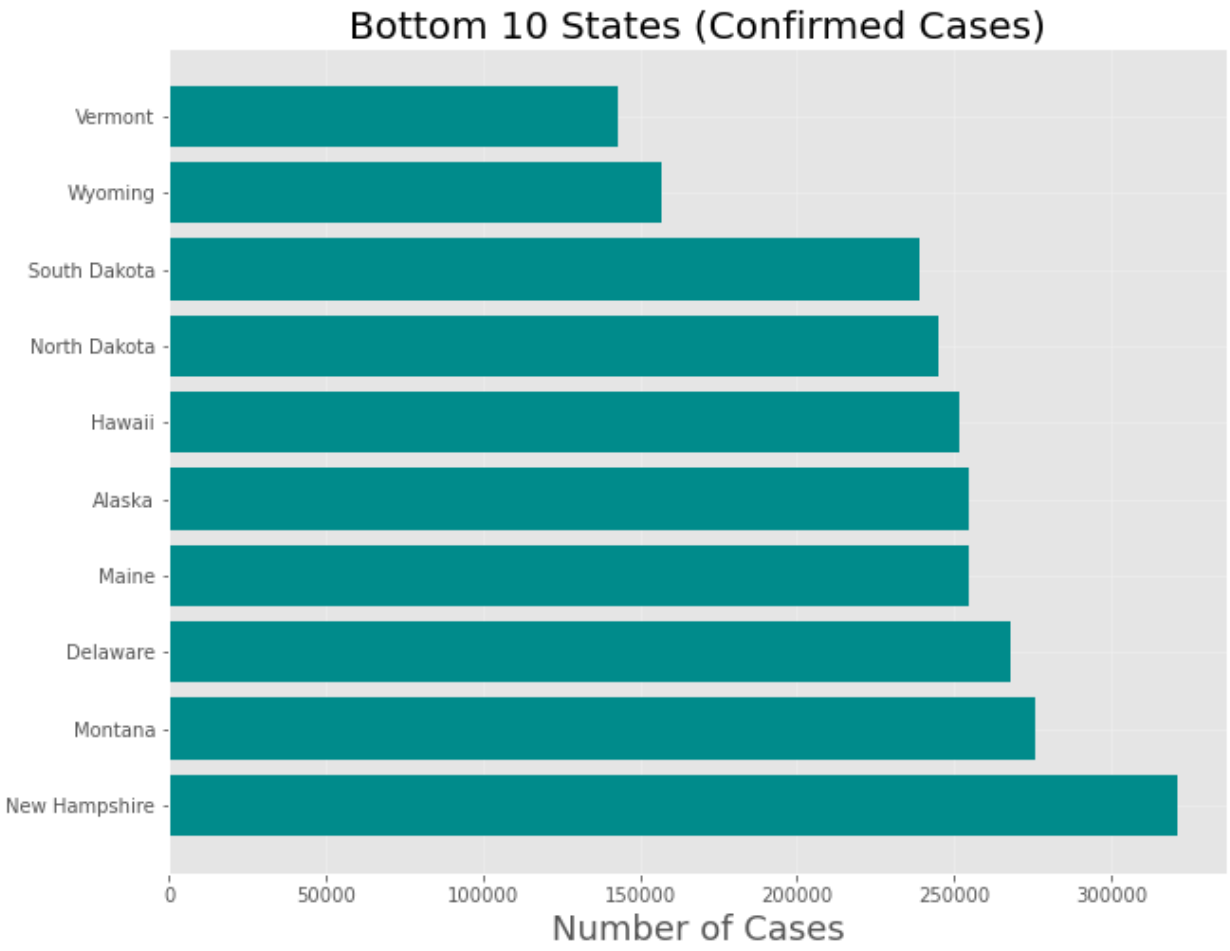
- date (object)
- county (object)
- state (object)
- fips (float64)
- cases (int64)
- deaths (float64)

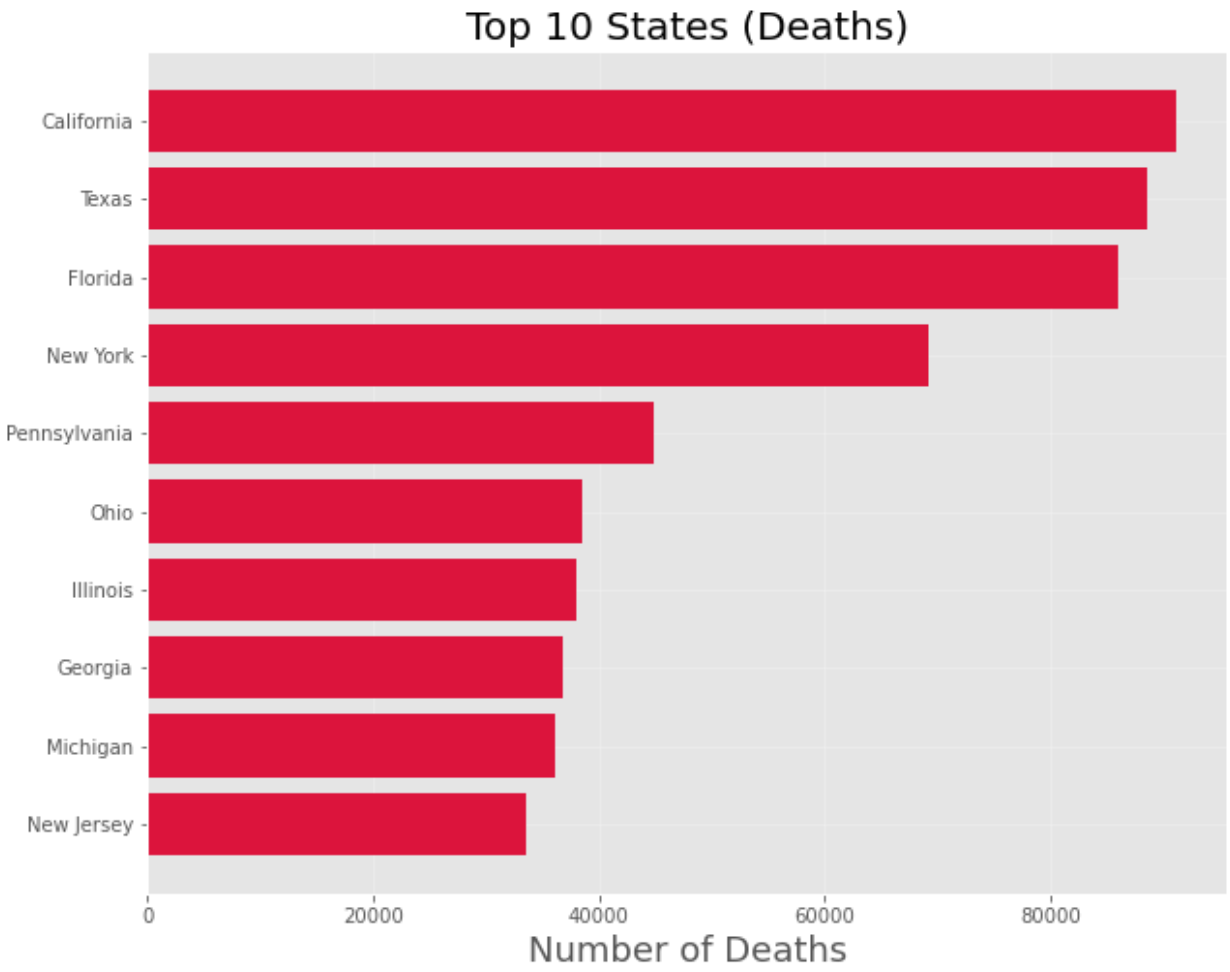
Exploratory Data Analysis, Data Cleaning, and Feature Engineering

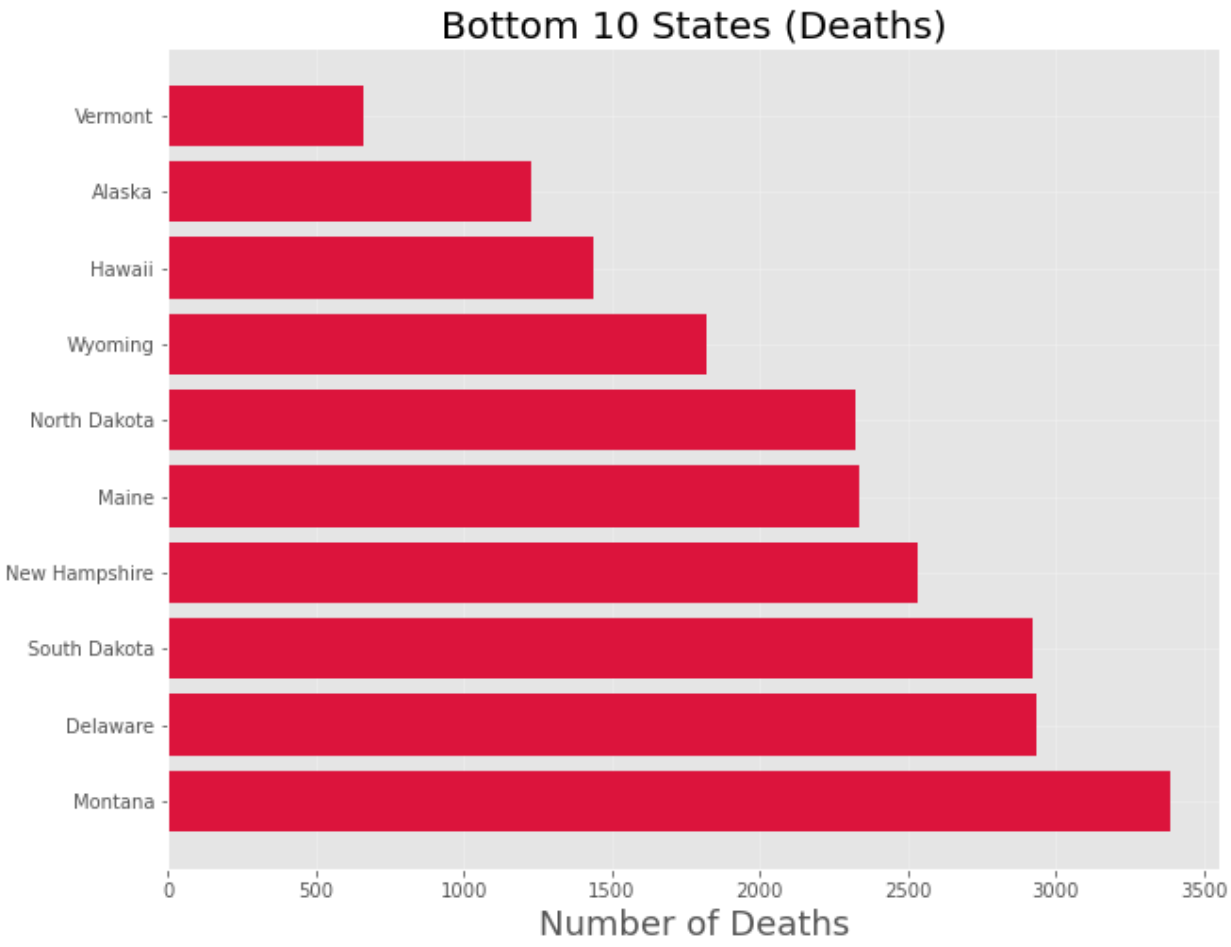
An unnecessary attribute was fips, which is an indexing number, so it was dropped column-wise. Date was changed from a string to a datetime object. The values were sorted by date, county, and state. The analysis would only include the 50 US states, so District of Columbia was excluded, along with the US territories: Guam, Northern Mariana Islands, Puerto Rico, Virgin Islands, American Samoa. A separate data table was created with only the most recent case and death counts for 50 states. To make a forecasting model for case counts using Meta's library Prophet, a data table was created that only included dates and case counts for Los Angeles County.

Case counts as well as deaths of the top 10 most numerous and bottom 10 least numerous states as of May 13, 2022 were plotted on bar charts. Case counts align with the population of states and more populous states had bigger case counts. Public health measures to counteract the spread of the virus varied by state, but what is most reflected is the raw number of people residing in each state. Deaths after infection followed a similar trend.

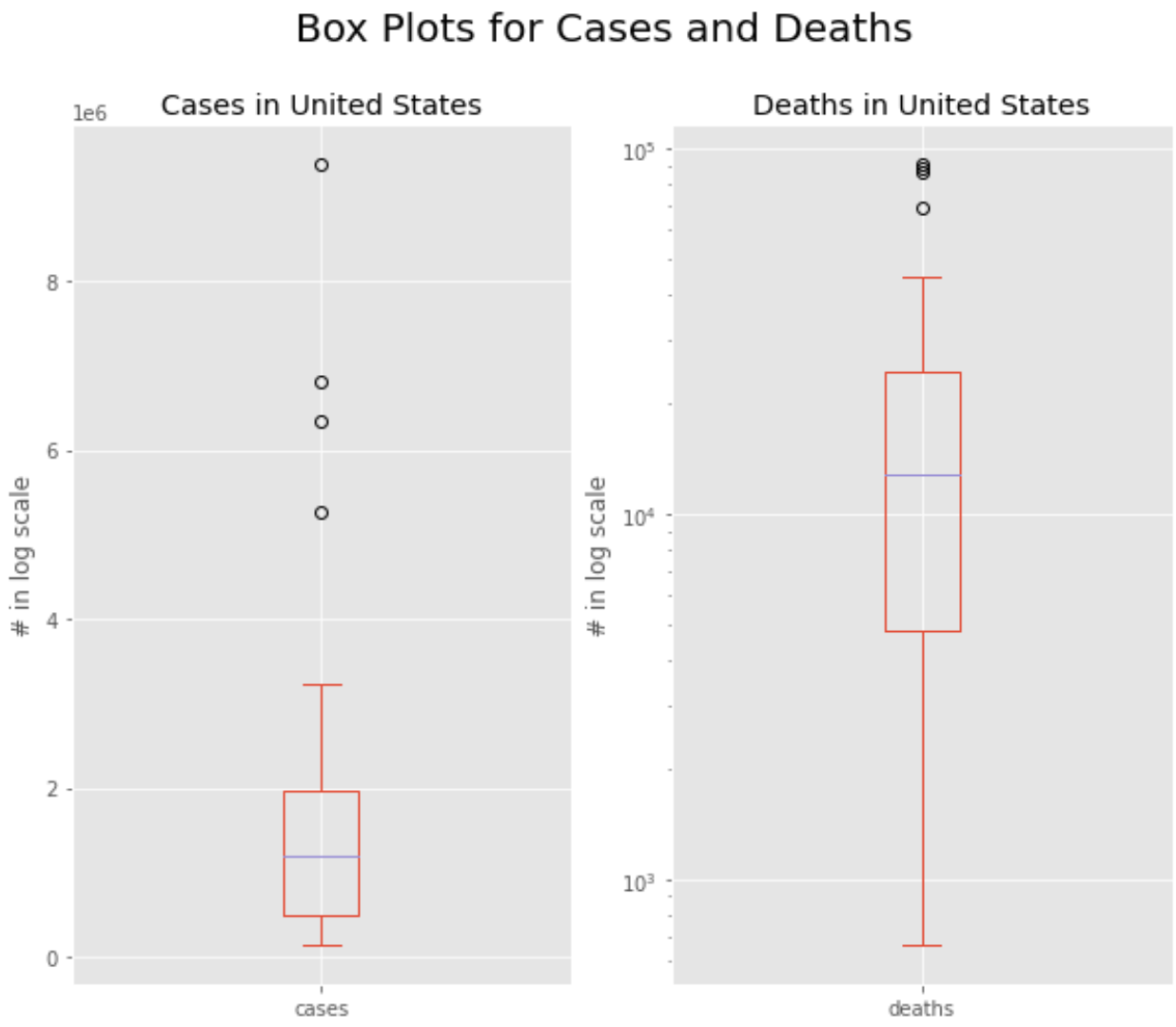



Bottom 10 States (Confirmed Cases)
Vermont
Wyoming
South Dakota
North Dakota
Hawaii
Alaska
Maine
Delaware
Montana
New Hampshire
0
50000
100000
150000
200000
250000
300000
Number of Cases

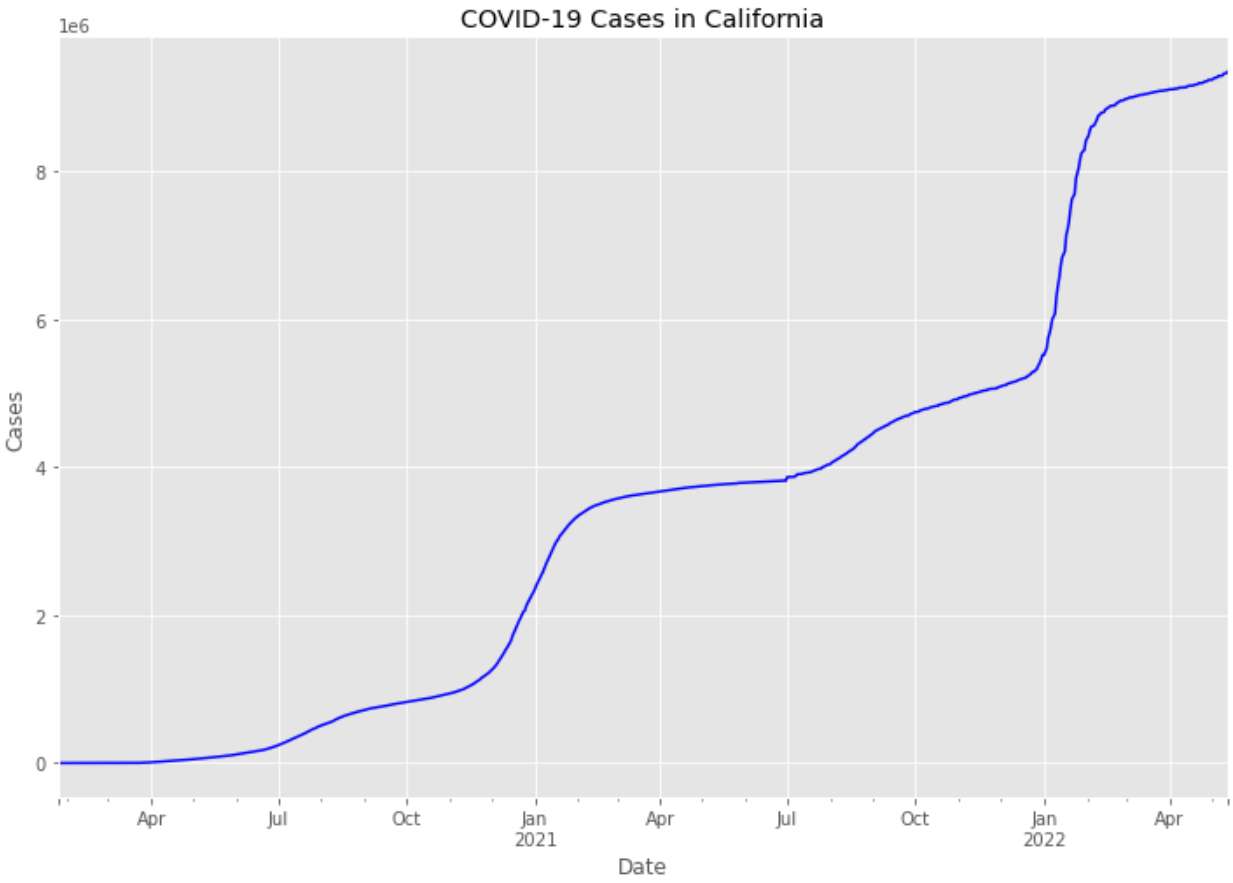

Top 10 States (Deaths)
California
Texas
Florida
New York
Pennsylvania
Ohio
Illinois
Georgia
Michigan
New Jersey
0
20000
40000
60000
80000
Number of Deaths

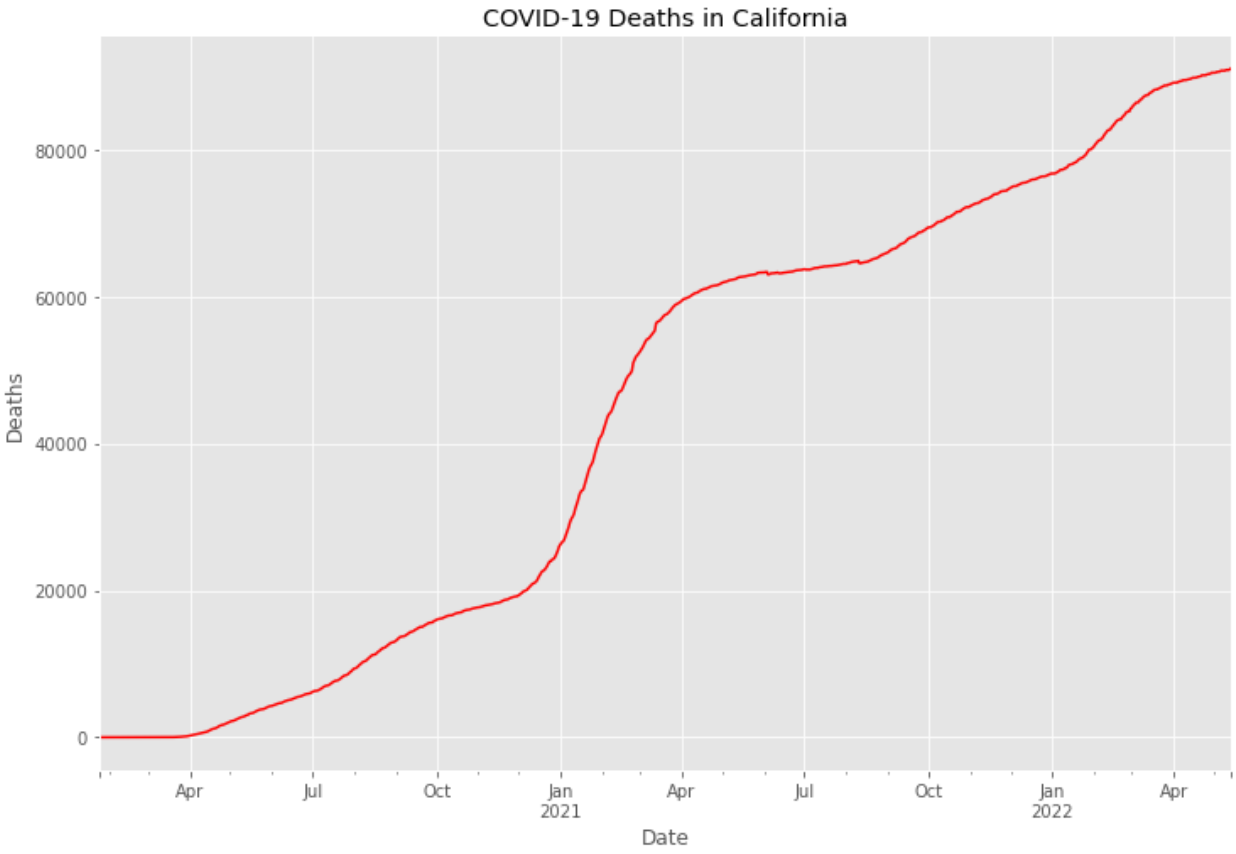

Bottom 10 States (Deaths)
Vermont
Alaska
Hawaii
Wyoming
North Dakota
Maine
New Hampshire
South Dakota
Delaware
Montana
0
500
1000
1500
2000
2500
3000
3500
Number of Deaths

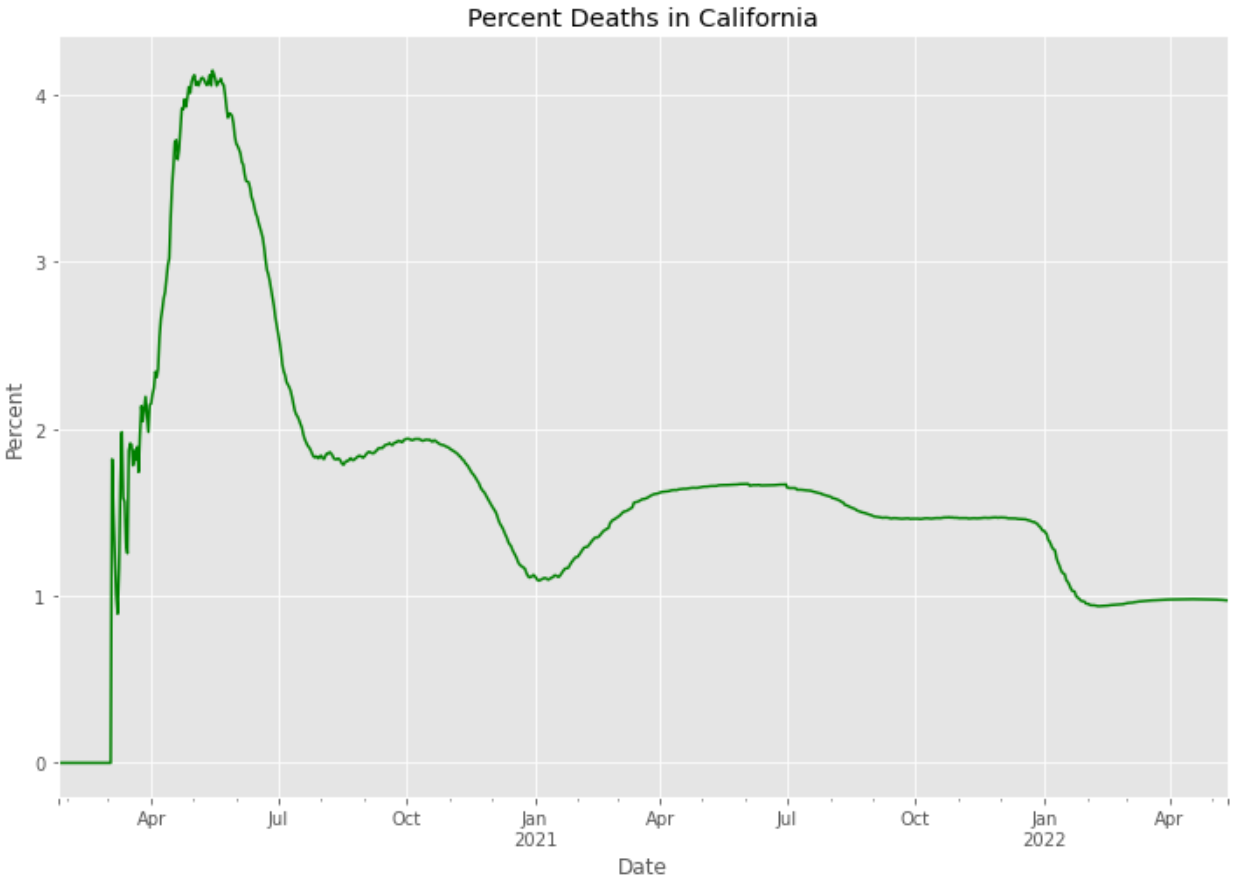
The maximum, minimum, quartiles and median of case counts and deaths in states as of May 13, 2022 were visualized with box plots. The spread of case counts was relatively narrow with outliers which included the most populous states. Deaths had a wider margin between the first and third quartiles. Perhaps this is due to differences in medical resources between various states in which urban areas have more access while rural areas have less.



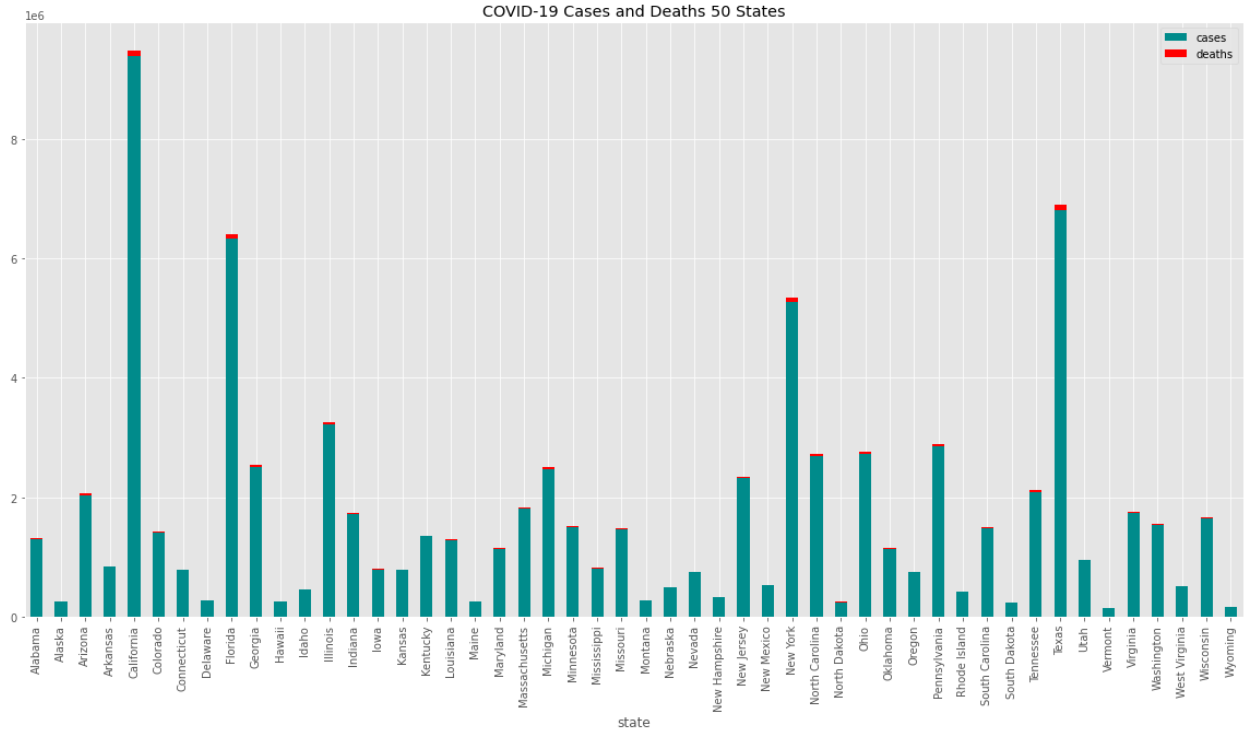
Cumulative case counts, deaths, and the percent of cases resulting in death over time from the beginning of the pandemic were plotted for California. The steep rises followed by plateaus reflect the periodic surges during the pandemic including the emergence of the delta and omicron variants. With the introduction of vaccines, the curve for the rise in deaths is shallower and the percent of cases leading to death stabilized to a lower number.




COVID-19 Deaths in California
Deaths
80000
60000
40000
20000
0
Apr
Jul
Oct
Jan
2021
Apr
Jul
Oct
Jan
2022
Apr
Date

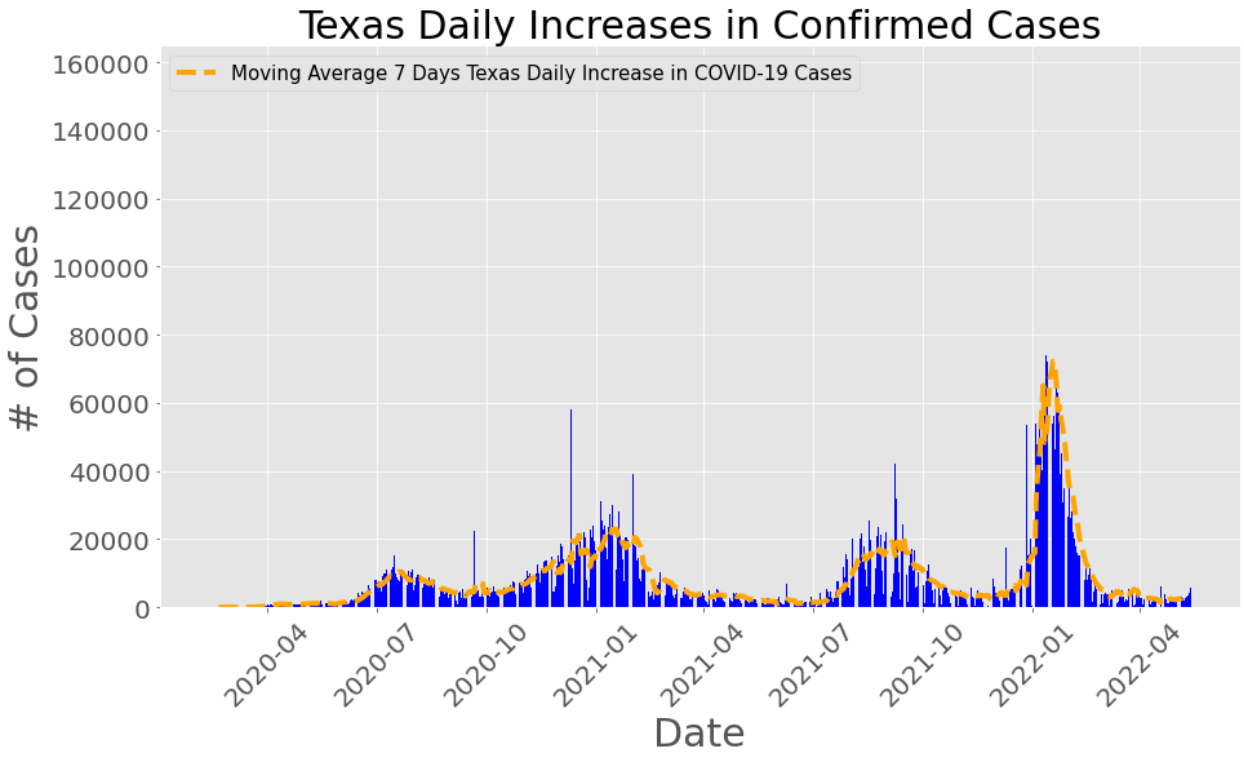

Percent Deaths in California
Percent
0
1
2
3
4
Apr
Jul
Oct
Jan
2021
Apr
Jul
Oct
Jan
2022
Apr
Date

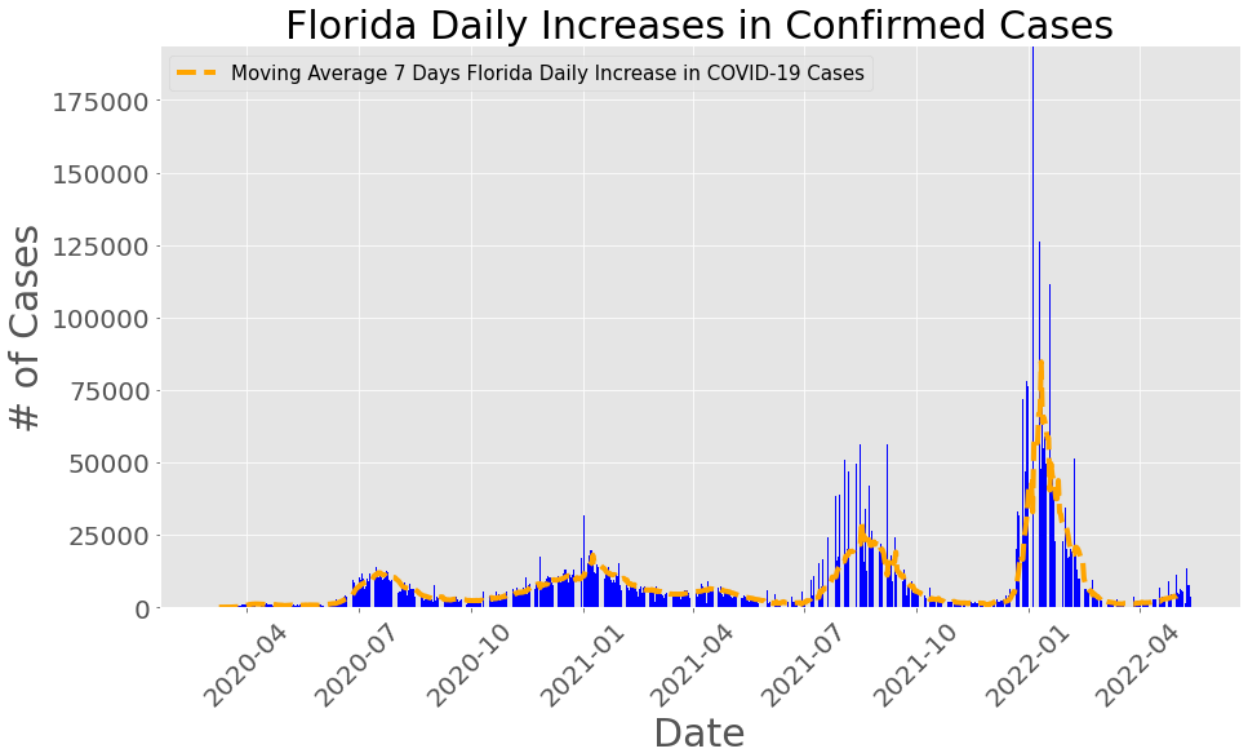
The case counts and deaths as of May 13, 2022 for all 50 states were plotted for comparison. As indicated, more rural states that are not densely populated and have smaller populations had the least cumulative case counts and deaths while larger states such as California and Texas far surpassed smaller states in the sheer numbers of cases.

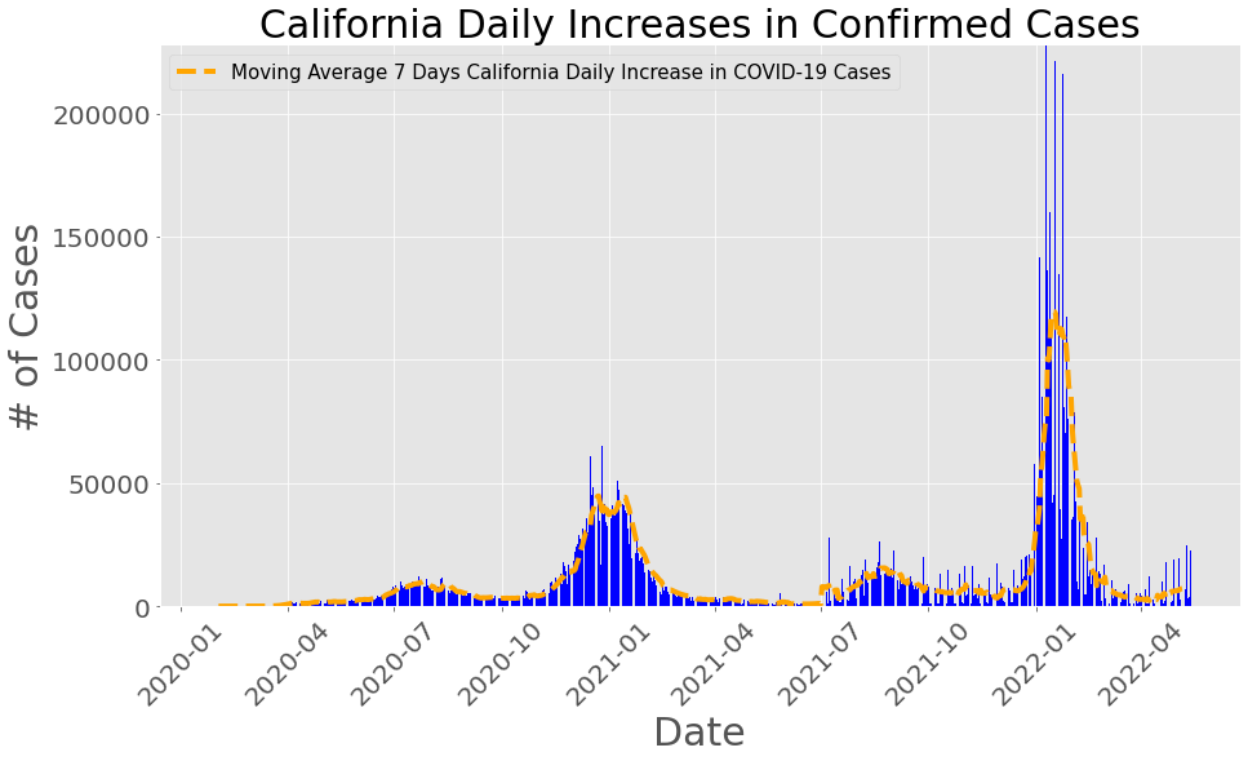


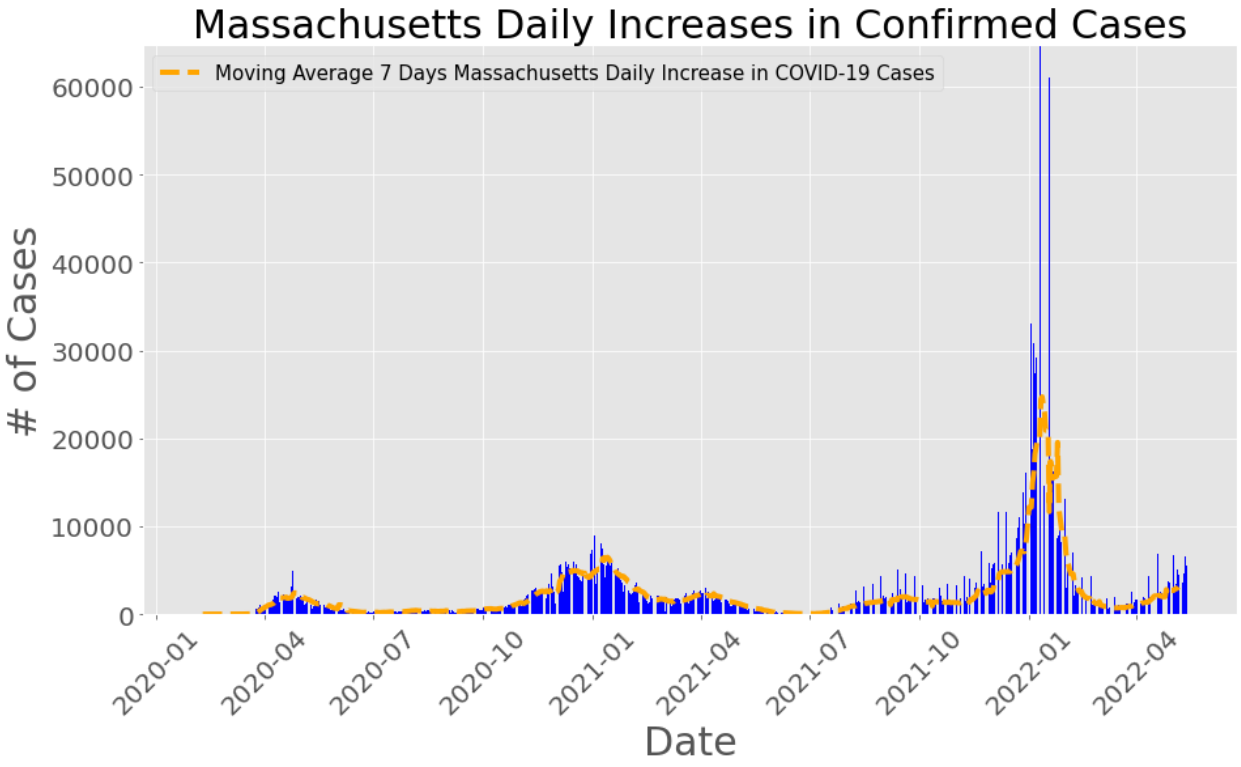
Plots were constructed for infection rates using functions that hone in on state and county levels. Moving averages were included for case counts and deaths per day. While the total number of case counts and deaths per day varied by geography, all plots reflected the initial summer surge, the 2020-2021 winter surge, as well as the emergence of the delta and omicron variants.

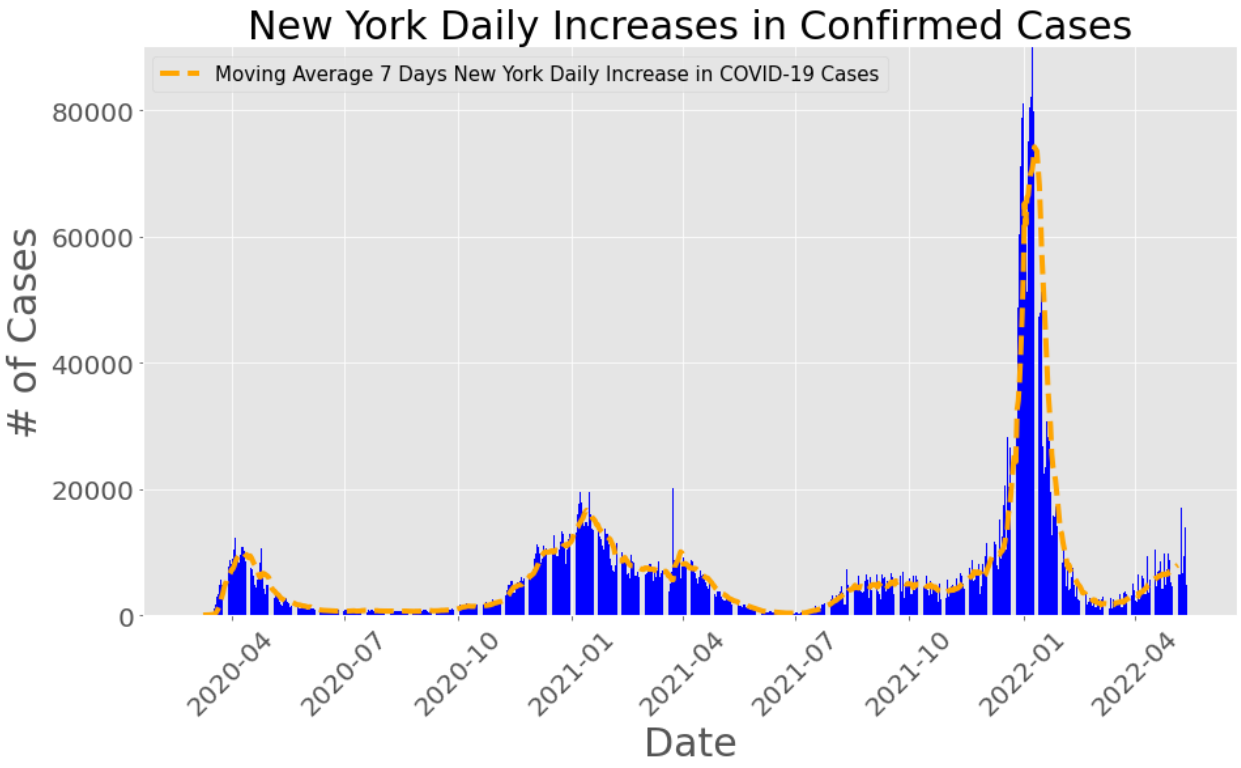
State Level Case Counts



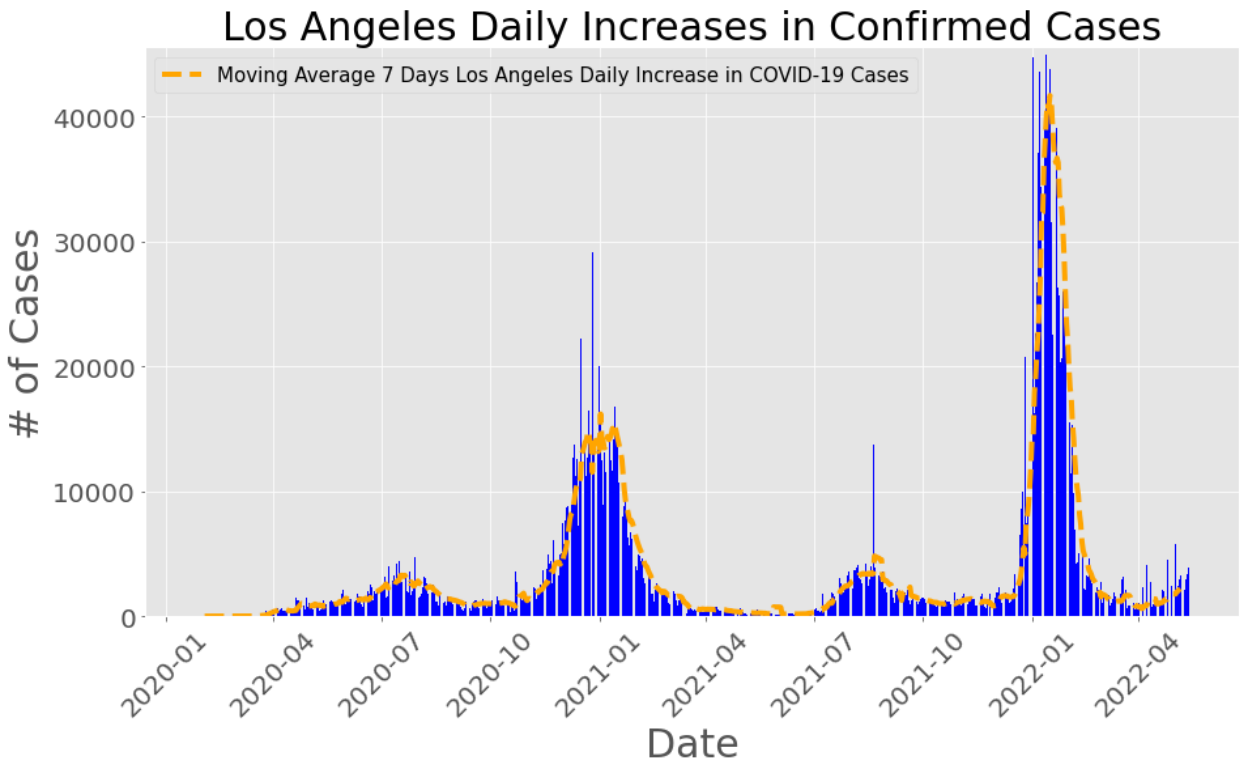

Florida Daily Increases in Confirmed Cases
Moving Average 7 Days Florida Daily Increase in COVID-19 Cases
# of Cases
175000
150000
125000
100000
75000
50000
25000
0
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date

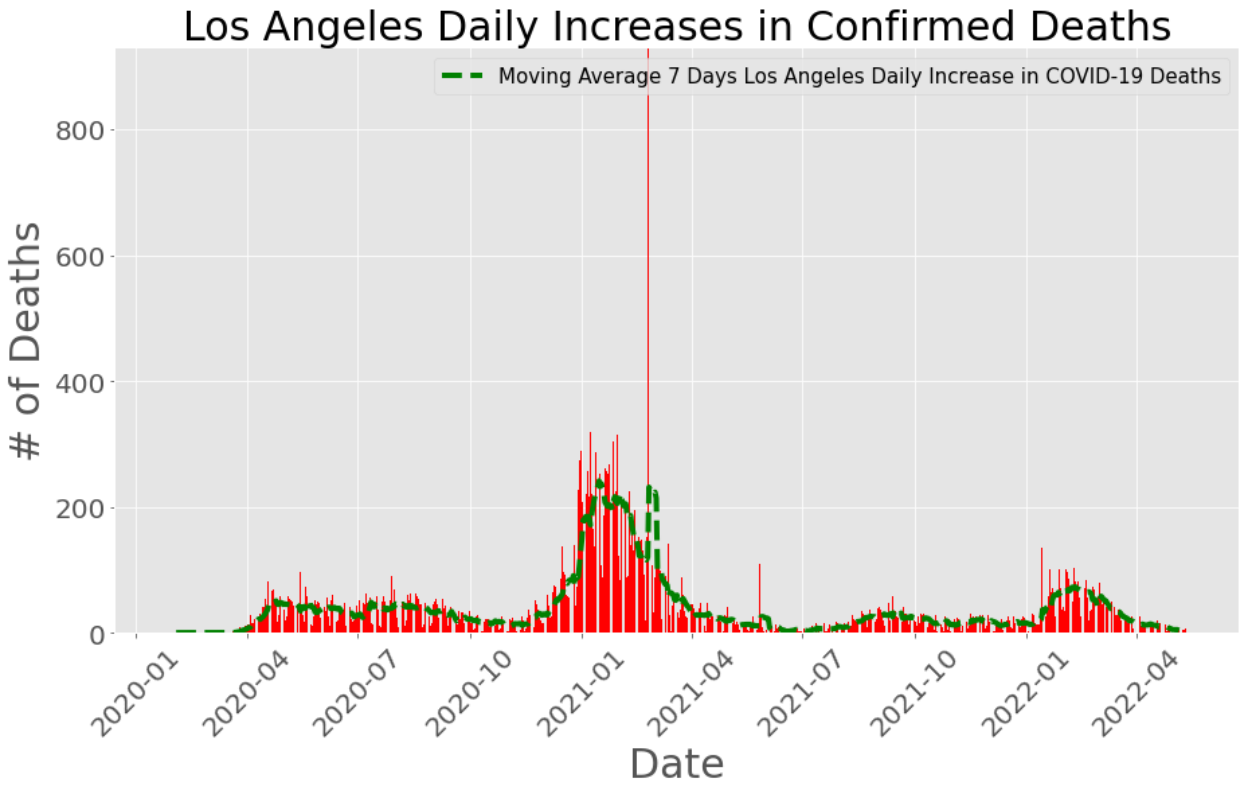

California Daily Increases in Confirmed Cases
Moving Average 7 Days California Daily Increase in COVID-19 Cases
# of Cases
200000
150000
100000
50000
0
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date

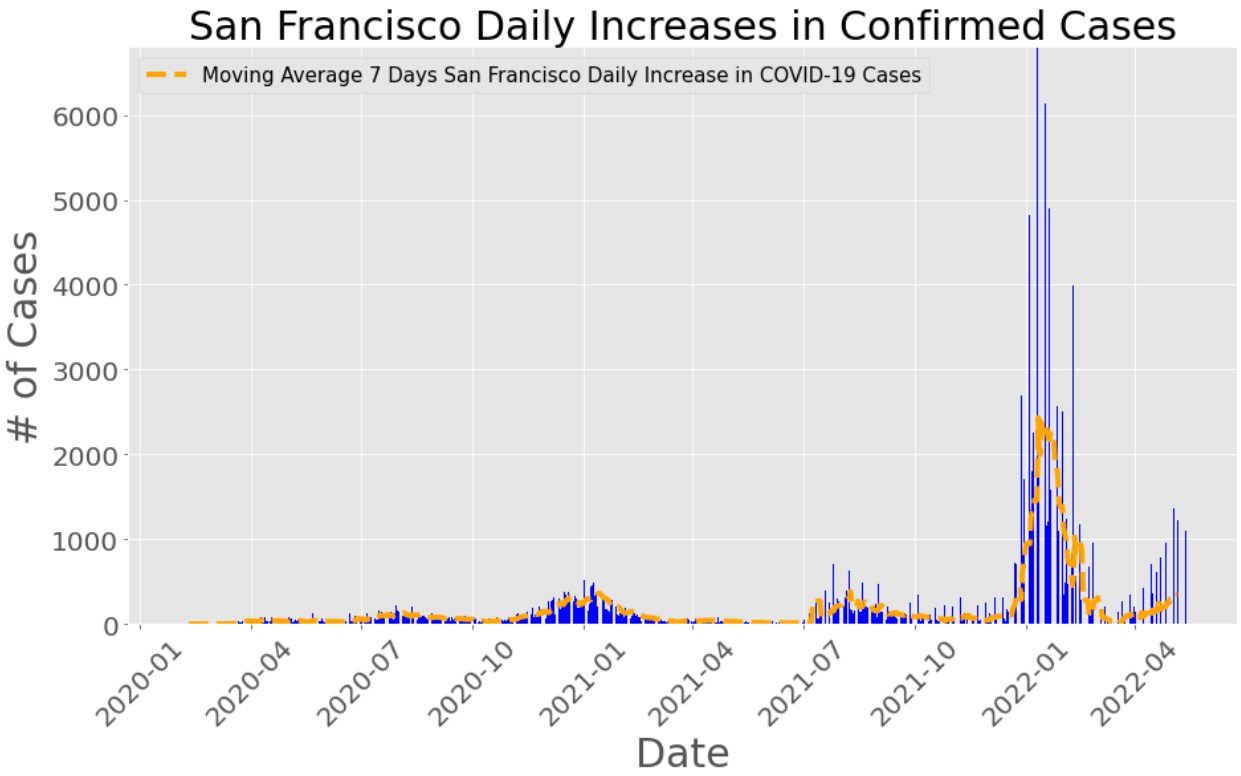

Massachusetts Daily Increases in Confirmed Cases
Moving Average 7 Days Massachusetts Daily Increase in COVID-19 Cases
# of Cases
0
10000
20000
30000
40000
50000
60000
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date

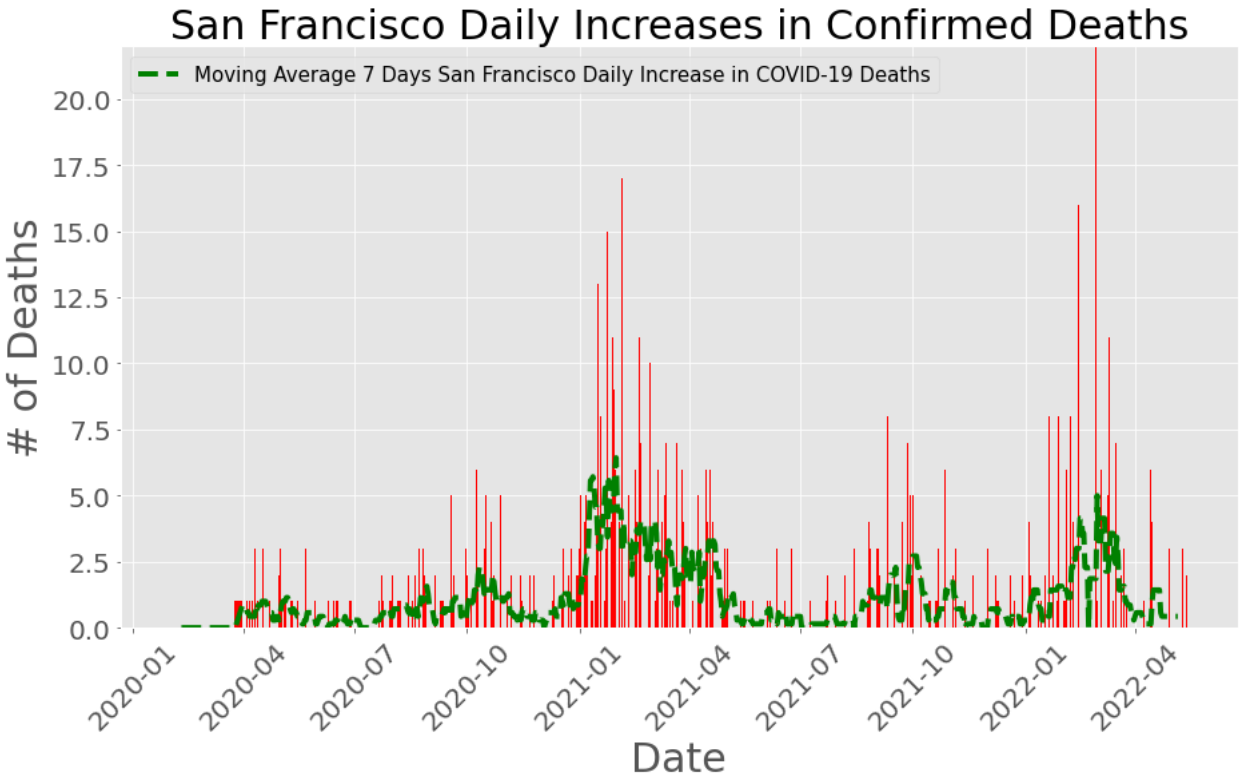

New York Daily Increases in Confirmed Cases
Moving Average 7 Days New York Daily Increase in COVID-19 Cases
# of Cases
0
20000
40000
60000
80000
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date

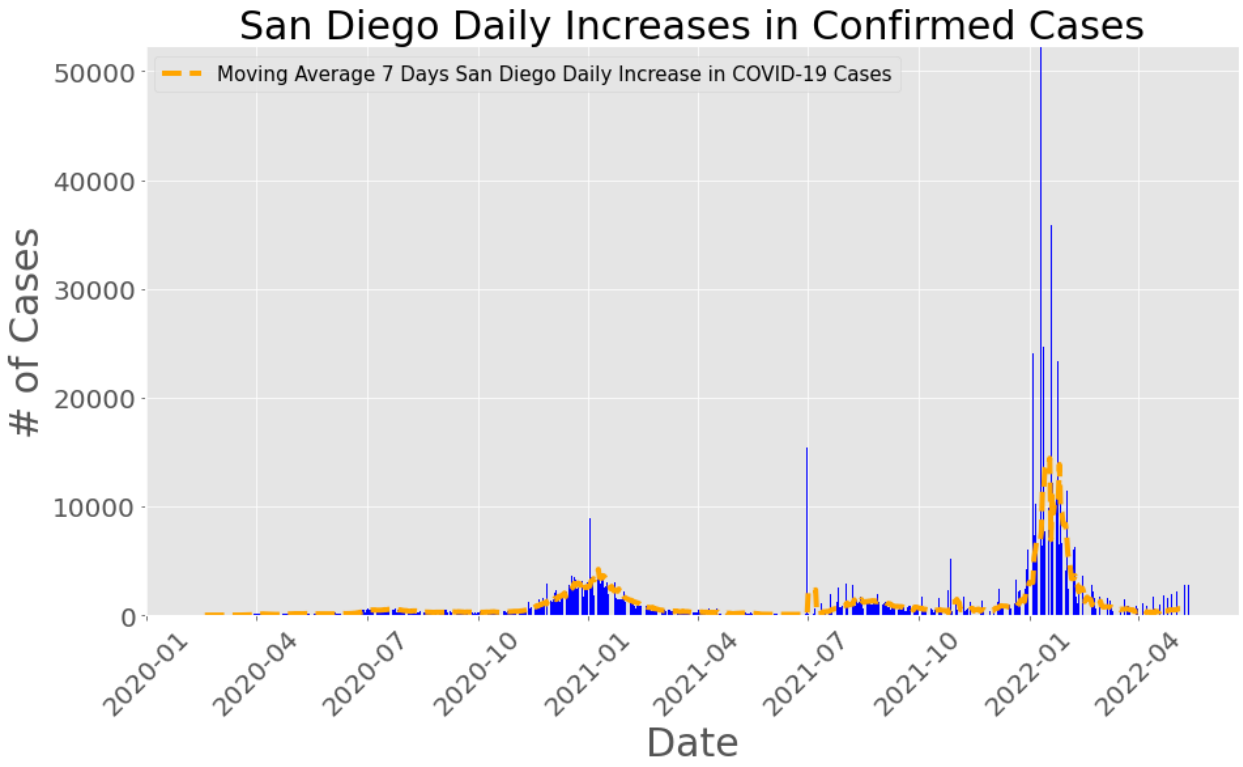
California Counties Case Counts and Deaths

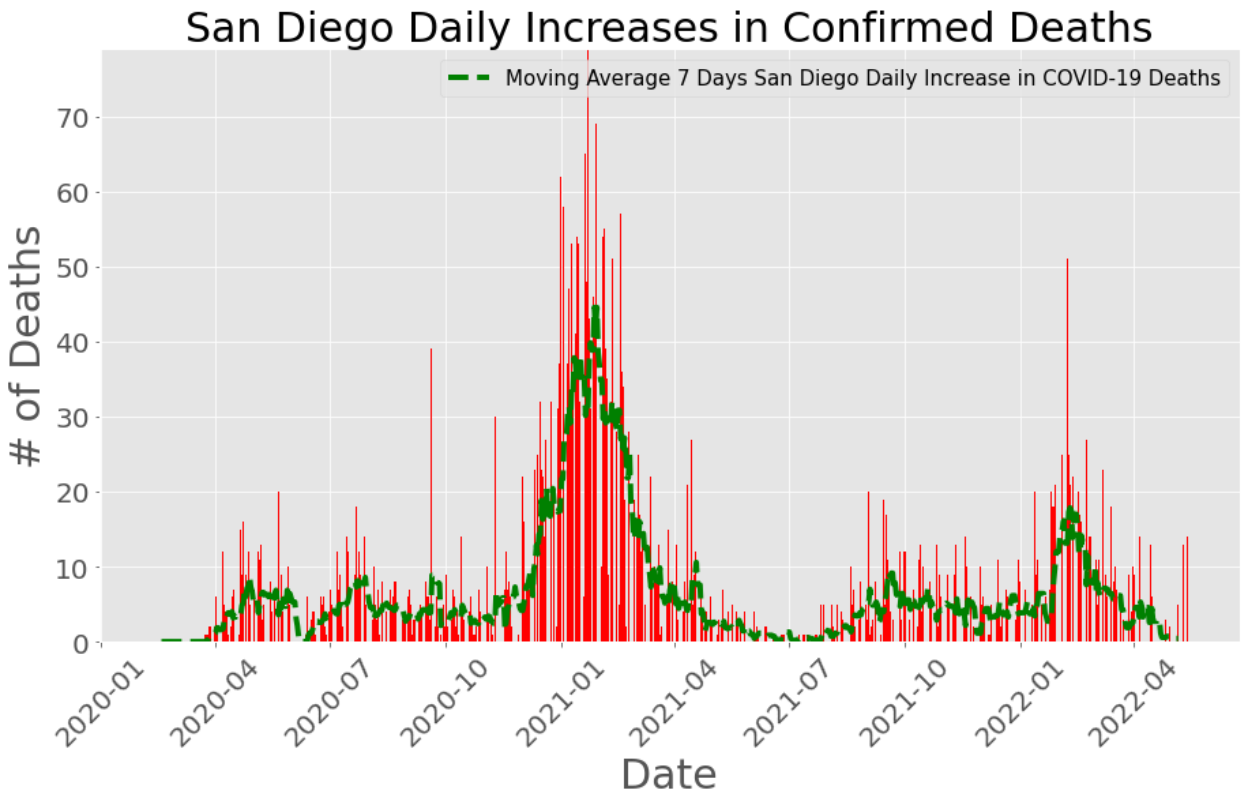


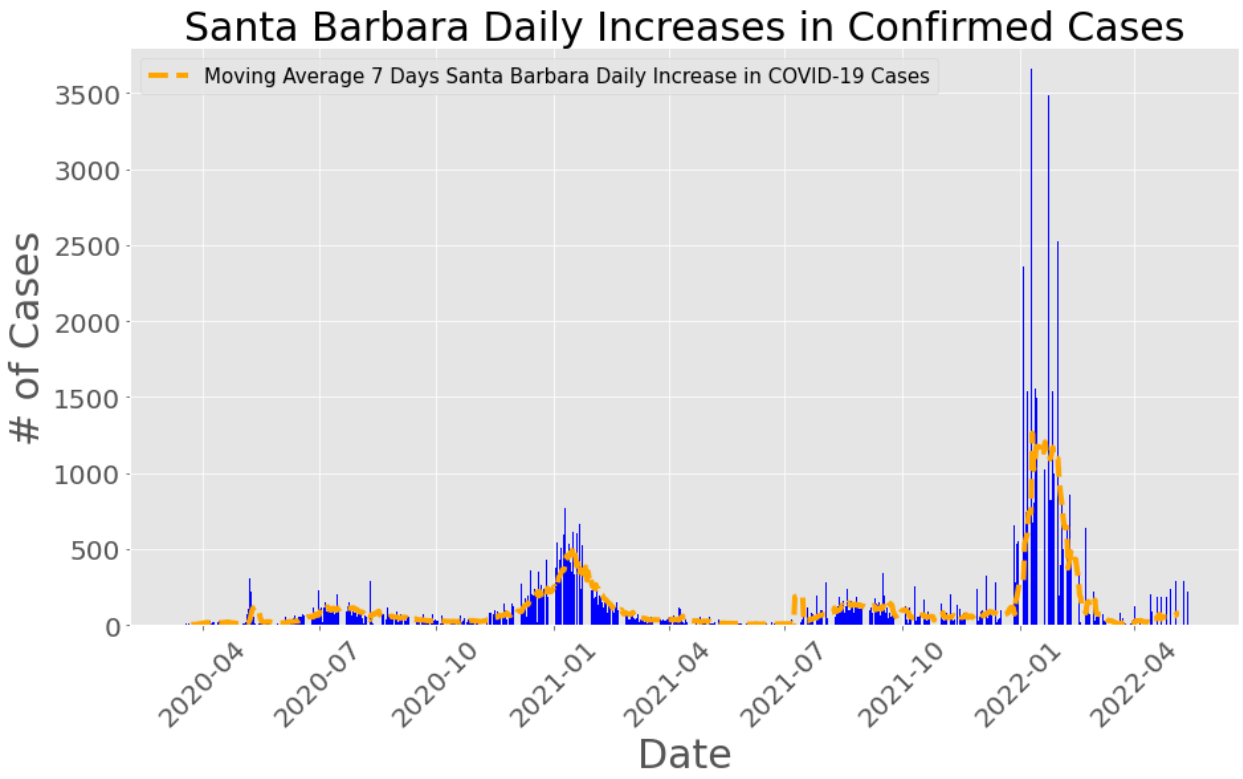


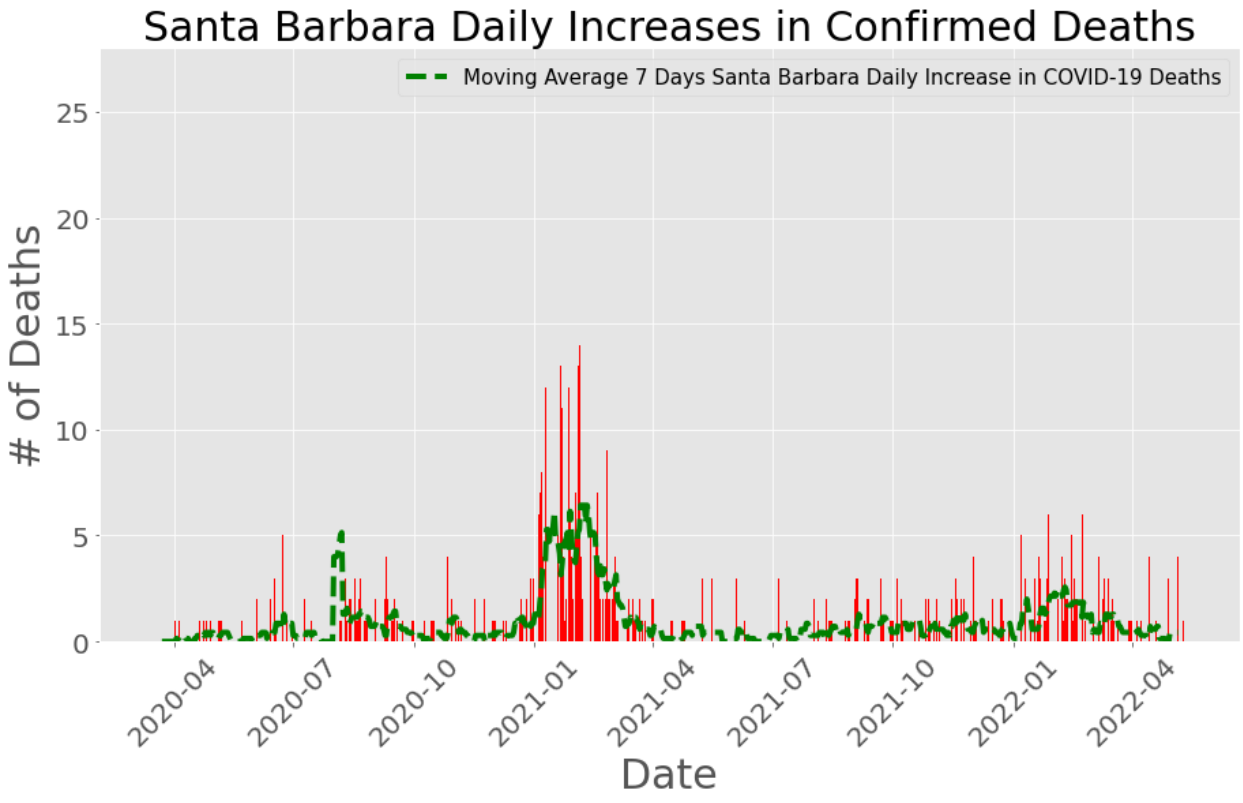

San Francisco Daily Increases in Confirmed Cases
Moving Average 7 Days San Francisco Daily Increase in COVID-19 Cases
# of Cases
0
1000
2000
3000
4000
5000
6000
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date

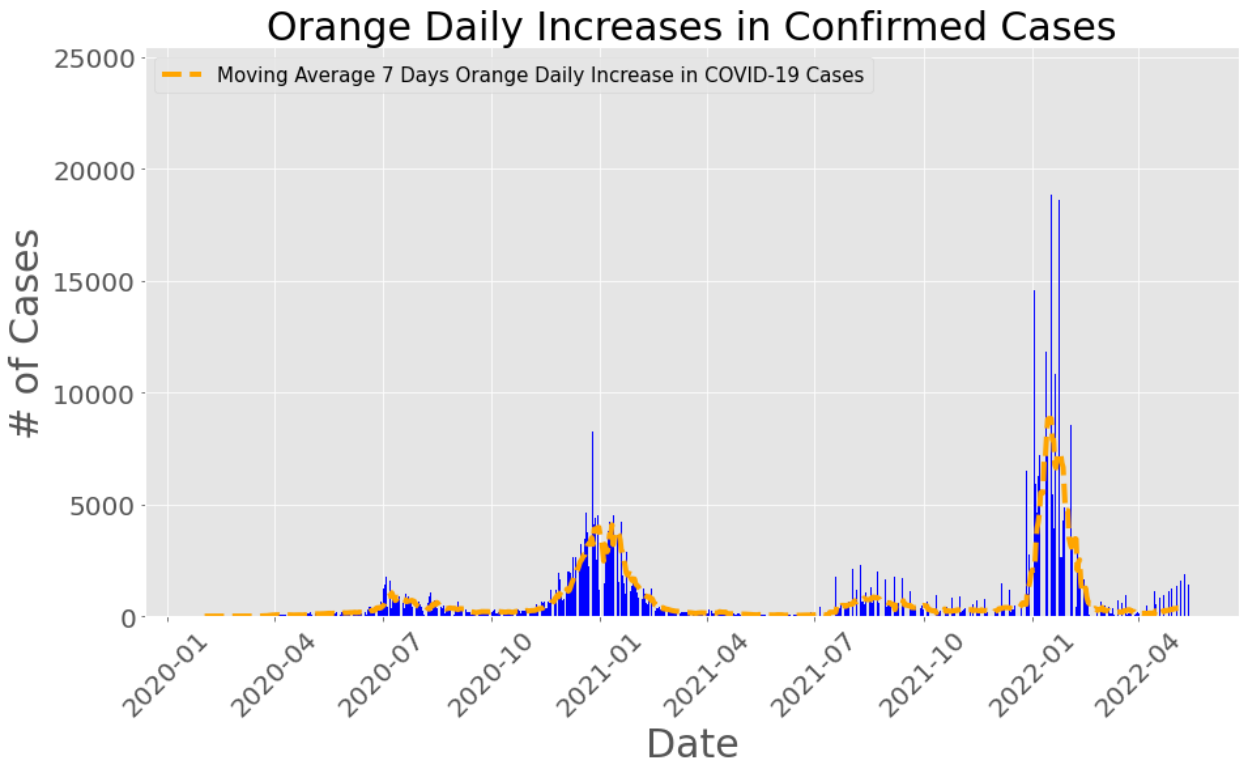

San Francisco Daily Increases in Confirmed Deaths
Moving Average 7 Days San Francisco Daily Increase in COVID-19 Deaths
# of Deaths
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date

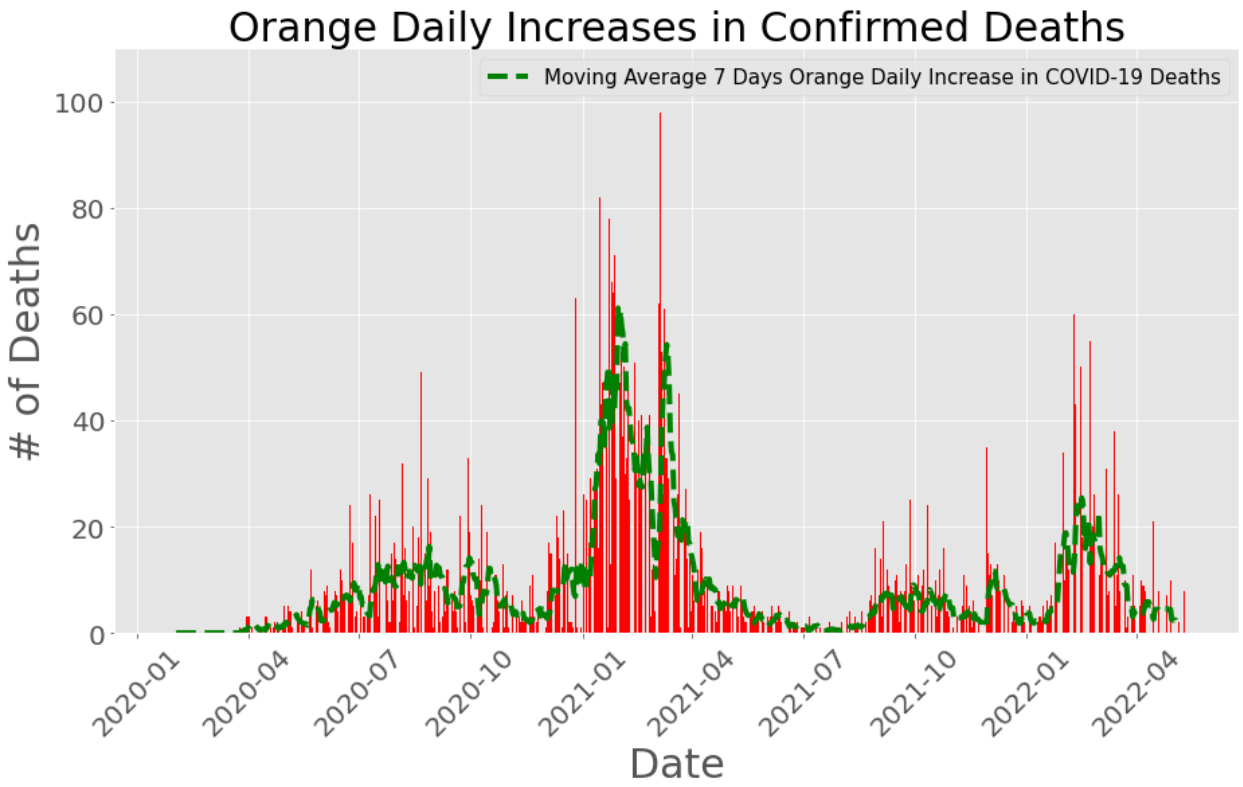

San Diego Daily Increases in Confirmed Cases
Moving Average 7 Days San Diego Daily Increase in COVID-19 Cases
# of Cases
50000
40000
30000
20000
10000
0
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date

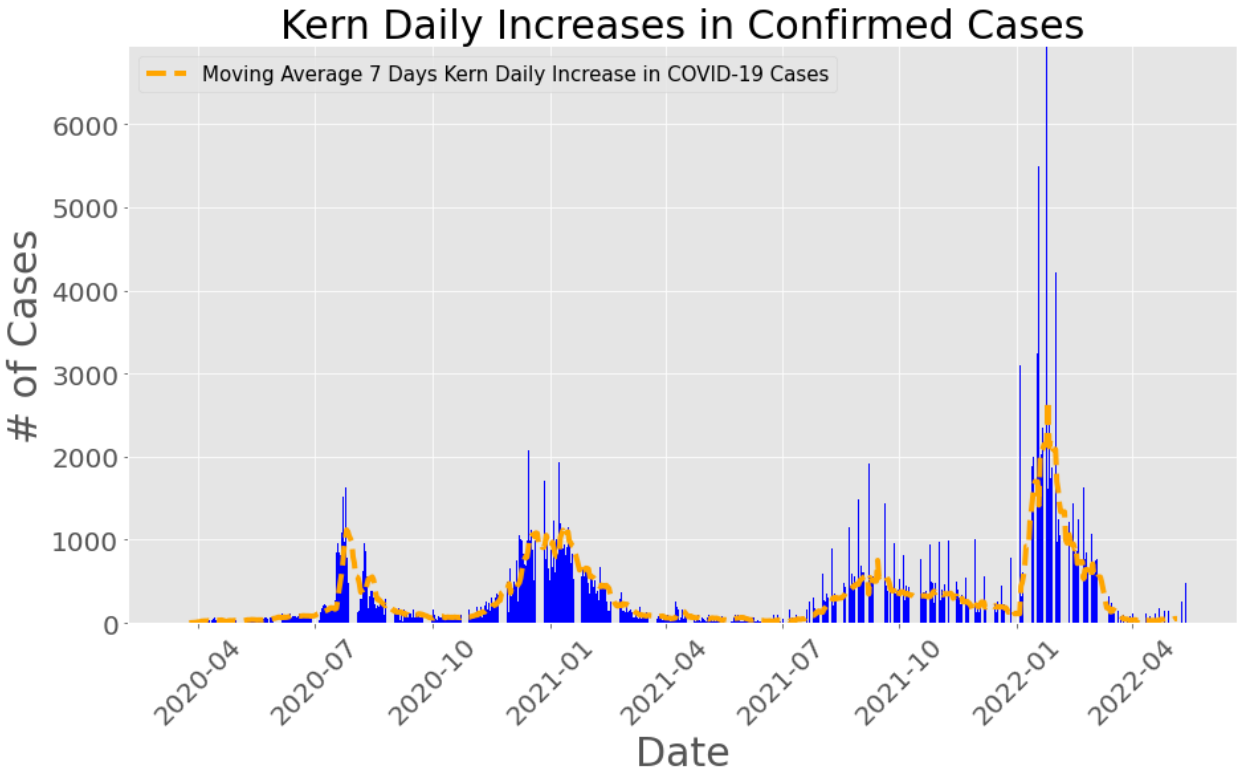

San Diego Daily Increases in Confirmed Deaths
Moving Average 7 Days San Diego Daily Increase in COVID-19 Deaths
# of Deaths
70
60
50
40
30
20
10
0
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date

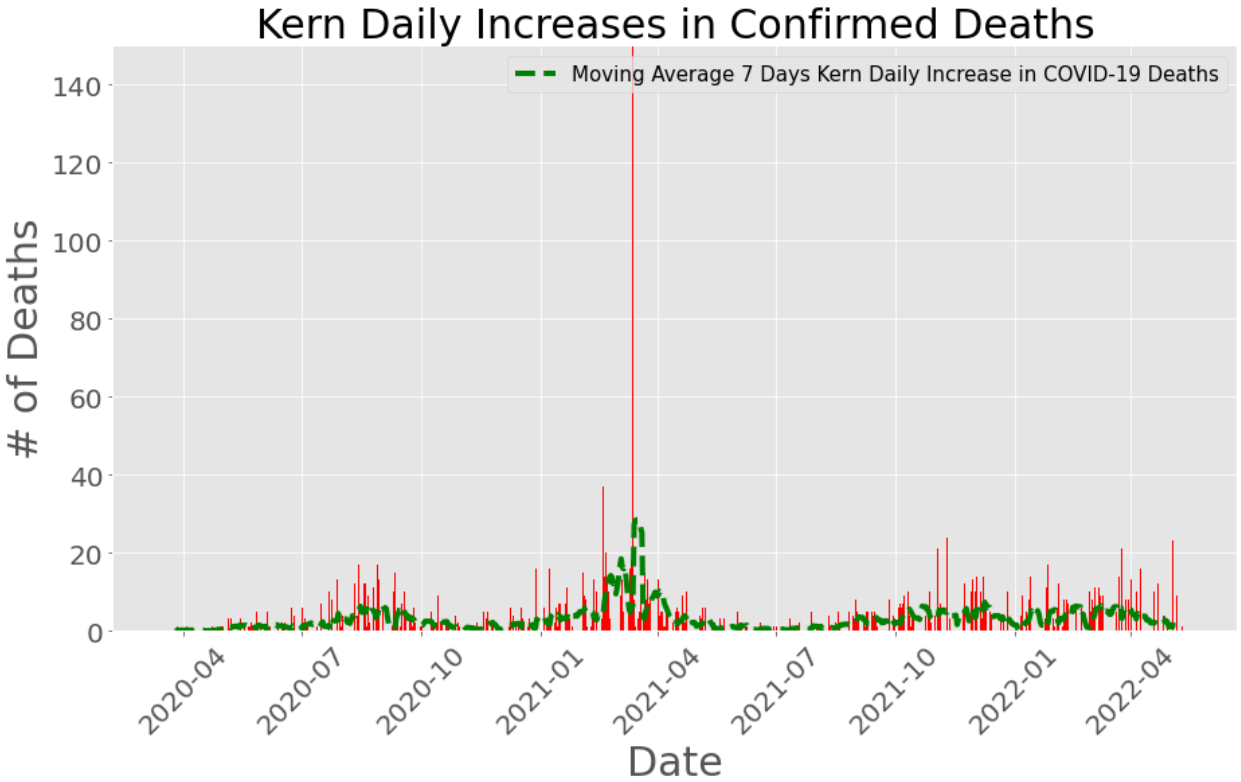

Santa Barbara Daily Increases in Confirmed Cases
Moving Average 7 Days Santa Barbara Daily Increase in COVID-19 Cases
# of Cases
0
500
1000
1500
2000
2500
3000
3500
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date


Santa Barbara Daily Increases in Confirmed Deaths
Moving Average 7 Days Santa Barbara Daily Increase in COVID-19 Deaths
# of Deaths
0
5
10
15
20
25
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date

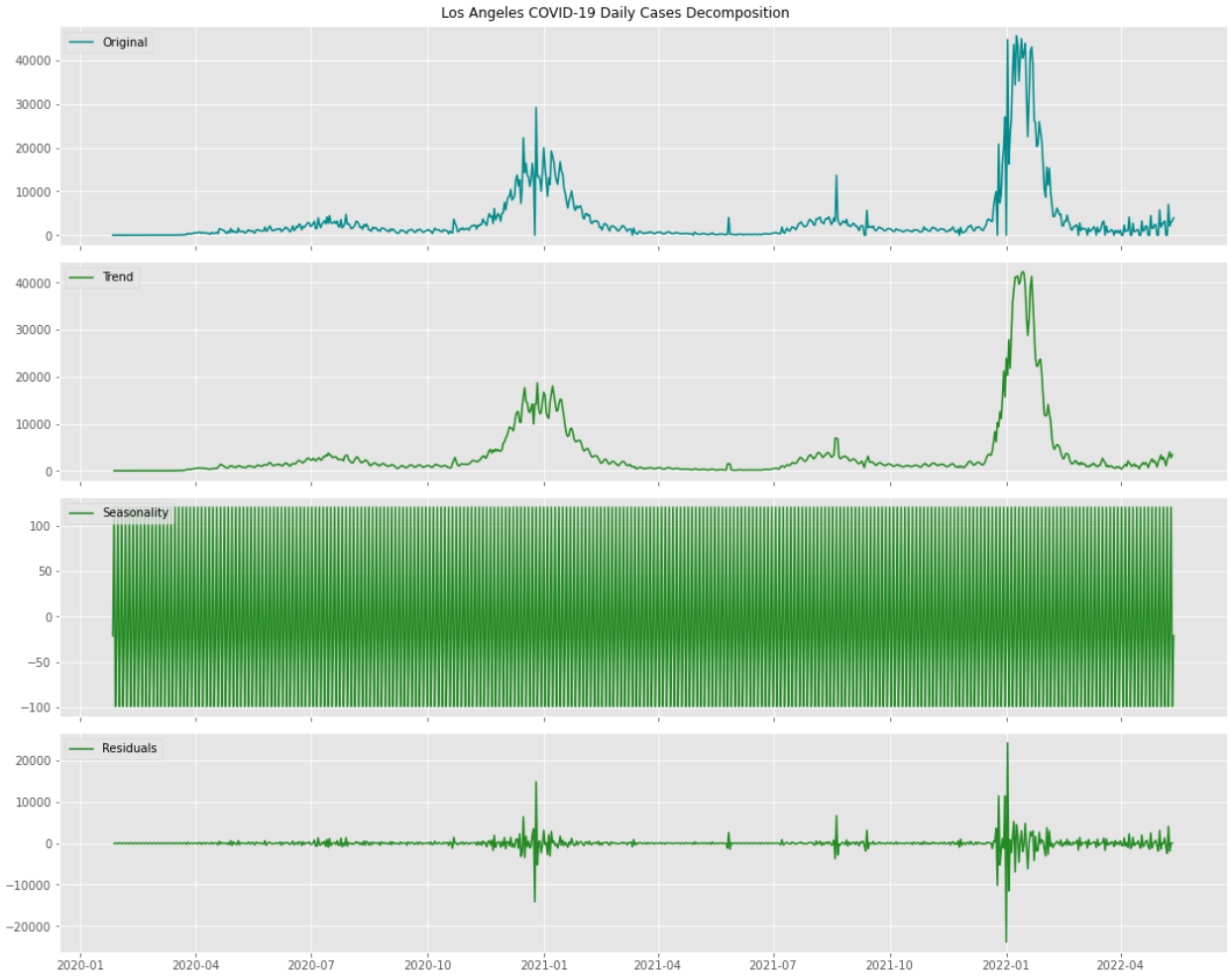

Orange Daily Increases in Confirmed Cases
Moving Average 7 Days Orange Daily Increase in COVID-19 Cases
# of Cases
25000
20000
15000
10000
5000
0
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date


Orange Daily Increases in Confirmed Deaths
Moving Average 7 Days Orange Daily Increase in COVID-19 Deaths
# of Deaths
100
80
60
40
20
0
2020-01
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date

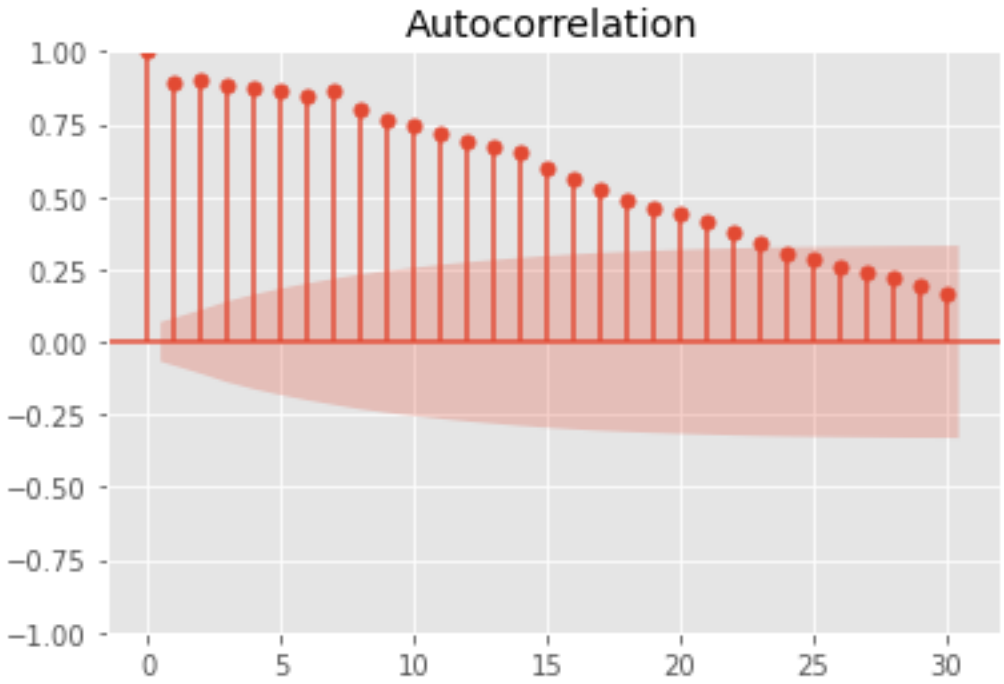

Kern Daily Increases in Confirmed Cases
Moving Average 7 Days Kern Daily Increase in COVID-19 Cases
# of Cases
0
1000
2000
3000
4000
5000
6000
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date

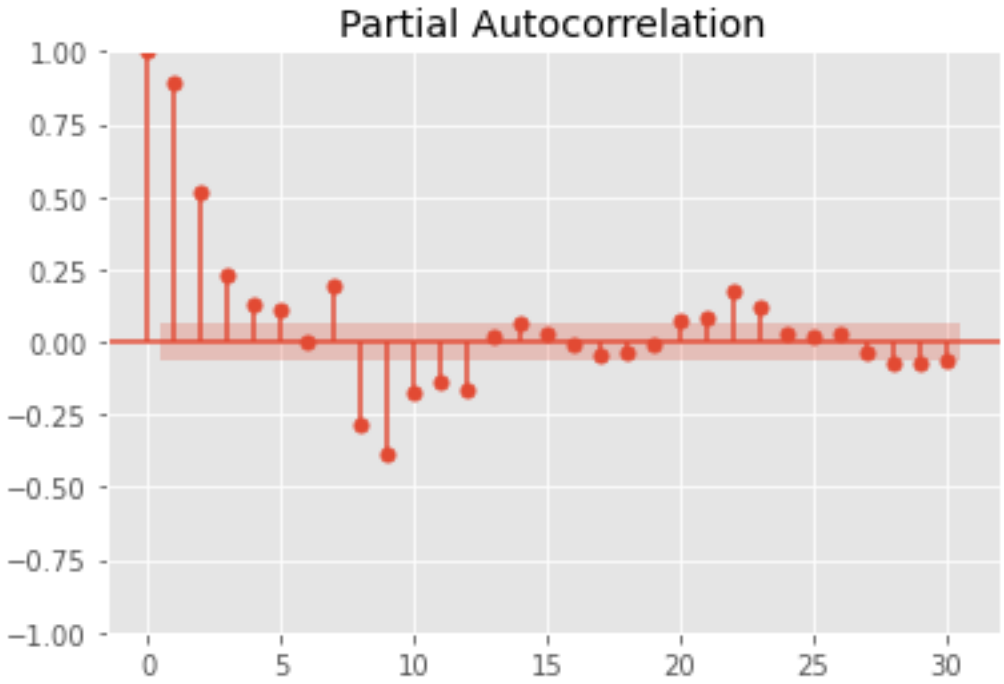

Kern Daily Increases in Confirmed Deaths
Moving Average 7 Days Kern Daily Increase in COVID-19 Deaths
# of Deaths
0
20
40
60
80
100
120
140
2020-04
2020-07
2020-10
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
Date

Data of daily case counts was prepared using numbers from Los Angeles County in California for time series analysis as well as forecasting on Prophet. To visualize the trend, noise was removed from the original time series data. The trend reflects the four surges recorded during the duration of the pandemic for this dataset.

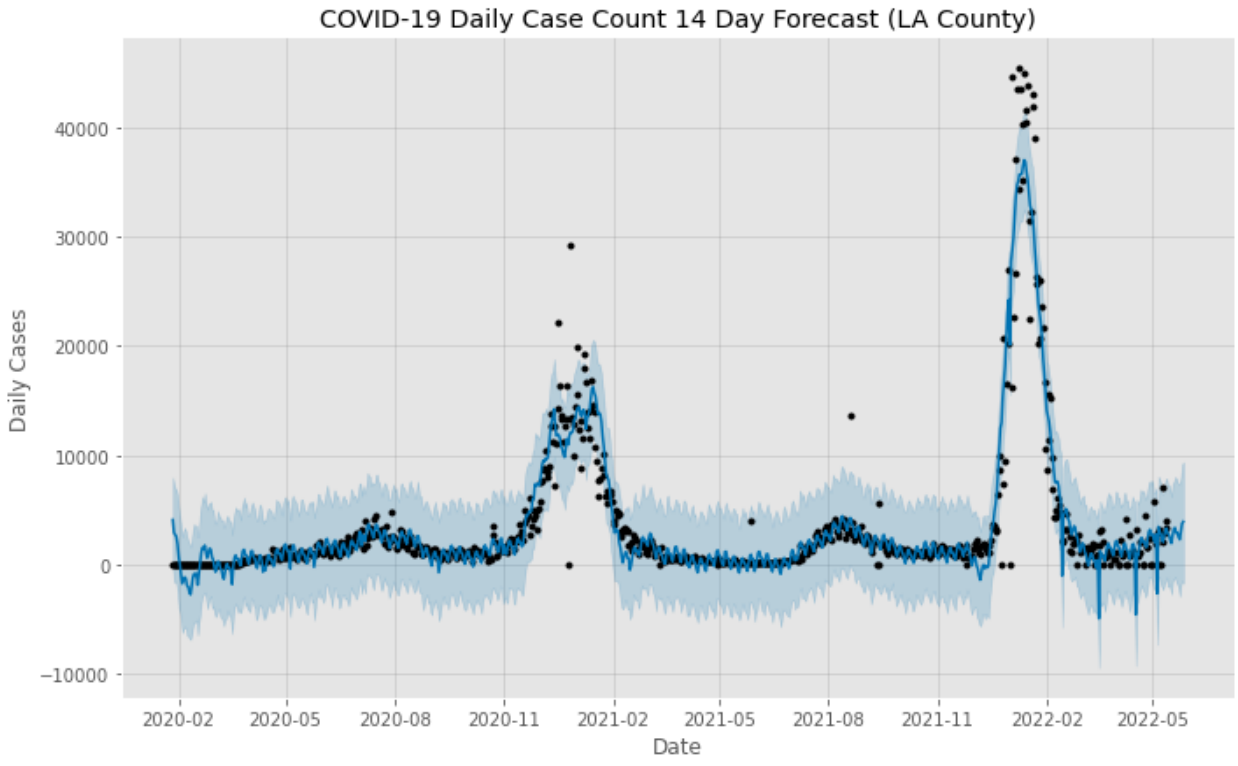


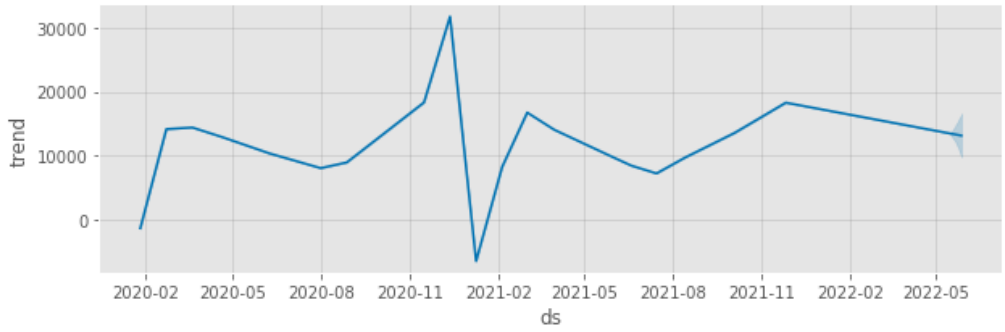
Autocorrelation and partial autocorrelation plots were made to determine if seasonal patterns or trend existed in the data, making it more suitable for forecasting. Points in the autocorrelation plot clearly lie outside of the confidence interval, indicating potential trend. The partial autocorrelation plot shows that after 7 lags, the value becomes insignificantly different from zero so the autoregressive order would be 6.

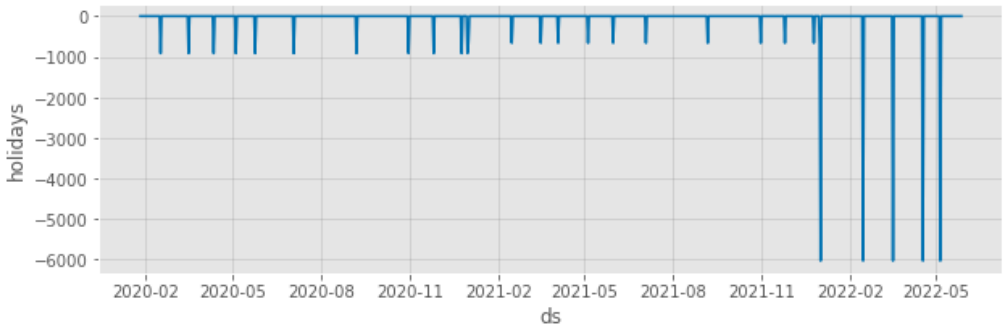


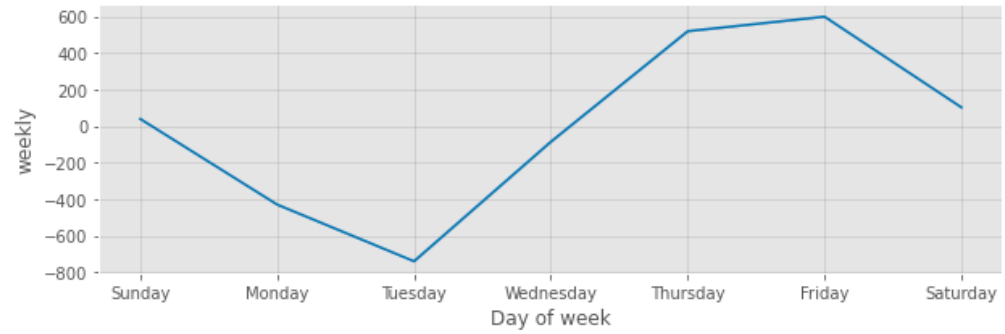


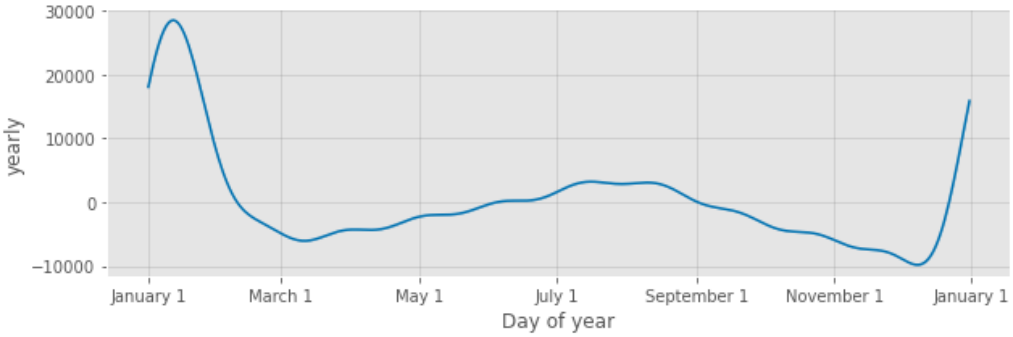
A 14-day forecast was made with dates of US holidays for each year put into consideration for modeling. The project was ongoing during the pandemic and the forecasting model correctly predicted the peak of the delta surge. Summaries of the model's trend, holidays, weekly, and yearly were visualized.

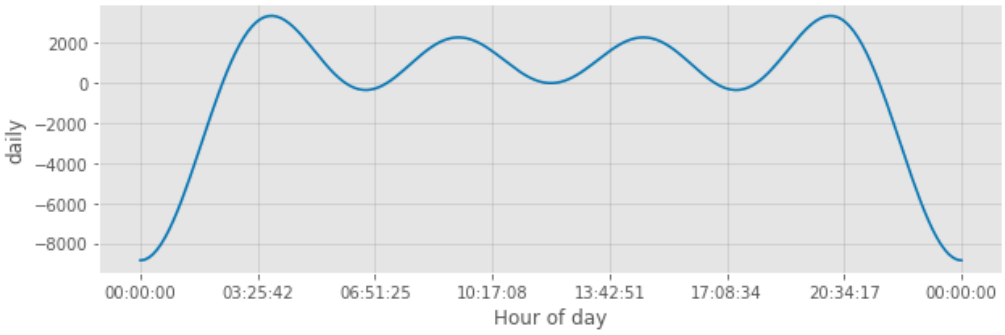



trend
30000
20000
10000
0
2020-02
2020-05
2020-08
2020-11
2021-02
2021-05
2021-08
2021-11
2022-02
2022-05
ds


holidays
0
−1000
−2000
−3000
−4000
−5000
−6000
2020-02
2020-05
2020-08
2020-11
2021-02
2021-05
2021-08
2021-11
2022-02
2022-05
ds


weekly
600
400
200
0
−200
−400
−600
−800
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week


yearly
30000
20000
10000
0
−10000
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year


daily
2000
0
−2000
−4000
−6000
−8000
00:00:00
03:25:42
06:51:25
10:17:08
13:42:51
17:08:34
20:34:17
00:00:00
Hour of day

Key Findings and Insights

Cumulative case counts and daily infection rates varied by state and county due to factors such as population size, population density, medical resources, and public health precautions taken. While the total numbers varied, the states and counties in this analysis all followed the same pattern for daily infections and the four surges included in the dataset were clearly apparent. Vaccination made an obvious difference in curtailing the rate of infection in the population as well as reducing deaths for those infected. The forecasting model created varied in performance over time, but could be a useful tool to predict infection rates for not only COVID-19, but a pandemic or endemic by another infectious agent in the future.

Next Steps

The data for this analysis stopped being available before the pandemic concluded. The missing data could be gathered to further study case counts and infection rates. Other attributes such as hospitalizations and vaccination rates could be included as well. More plots for infection rates could be made for the states with most and least cases and compared for significant differences, although it is assumed that the four surges would be apparent in any state. High density urban counties could also be compared with sparsely populated rural counties. As far as forecasting, more data would improve the model. A non-linear model can be tested as well and other potential holidays could be identified.