Statistical Analysis of Temporal and Spatial Trends in US Covid-19 Cases and Deaths

Jason Gong and Micah Swann

Introduction

Covid-19 is a novel, highly contagious, acute respiratory virus that was first identified in December 2019 in Wuhan, China. Over the course of the following 14 months, this virus spread rapidly to every corner of the globe, becoming one of the deadliest pandemics in recorded history. In the United States, the first confirmed Covid-19 case was identified in January 2020 and by mid-March there were confirmed cases in every single state and North American territory. In the midst of this rapid pandemic spread, epidemiologists and modelers struggled to accurately forecast the spatial and temporal trends in cases and deaths. However, with regularly updated, publicly-available covid tracking data, a sufficient amount of data now exists to retroactively examine how cases and deaths evolved over the course of this 14 month period. This study utilizes the New York Times Covid Tracking Data to statistically analyze trends in the timeseries of Covid-19 cases and deaths as well as the spatial development of cases at the state level across the united states using cluster analysis.

Data Description

Data Sources

Due to the fragmented nature of the US public health system, there is no centralized governmental data repository that is updated daily with Covid-19 case and death data. Instead, this study obtained data from the New York Times (NYTimes) Covid-19 Tracking Project (https://github.com/nytimes/covid-19-data). The NYTimes relies on dozens of reporters across multiple time zones to regularly update this tracking database with new information from press conferences, report releases, and local databases. Datasets utilized in this analysis reported the daily cumulative case and death counts in the US aggregated at the national, state and county level (US.csv, US-states.csv, US-counties.csv), respectively. Demographic data on state populations were also obtained from the US census bureau to compare per capita rates.

Data Formatting

All data analysis and visualization for this study was conducted in RStudio. The dataset was filtered to only examine cases and deaths reported from the beginning of March 2020 through the middle of February 2021. Raw data was reported as cumulative cases and deaths through time. To examine daily statistics, a filtering function was applied to calculate the finite difference between each consecutive reporting day.

Exploratory Data Analysis

Timeseries Visualization

An initial exploratory data analysis (EDA) was conducted to both elucidate trends and characteristics of the dataset and to guide the model development process. To better understand the temporal evolution of daily new cases and deaths in the US, a timeseries for both of these parameters was first generated (Figure 1). The timeseries of daily US Covid-19 cases depicts four distinct regimes in the change in daily covid 19 cases throughout the course of the pandemic. From March through the end of May 2020, the number of cases grew logistically; growing exponentially in March before asymptoting at a maximum daily new case load of 25,000-30,000 individuals through April and May. Similar but larger magnitude growth trends are evident in June through August, asymptoting around 65,000 daily new cases, and October through December 2020, asymptoting around 250,000 daily new cases. Note that the sharp drop in cases around the end of December is likely a reflection of a decline in reporting around the winter holidays and not a reflection of the actual drop in the real case load. From January 2021 onwards, the case trend differs from the early regimes, with a noticeable linear decline in the reported case numbers through December. Another interesting aspect of this dataset is the seven-day oscillation in the case numbers. New case numbers always tend to be lower on Saturdays and Sundays than during weekdays, reflecting the fact that many labs do not report case numbers on the weekend. The timeseries of daily Covid-19 deaths shows a similar logistic growth rate to the case rate in the early spring 2020. However, the number of deaths, drops significantly from mid-May 2020 and oscillates around 1000 cases a days until November 2020 when the number of daily new deaths rises again, fluctuating around 3000 deaths per days. The seven oscillations, observed in the timeseries of new cases, is even more prominent in the death data, with significant drops in reported deaths during the weekend. This initial visualization and review makes evident that that there are clear similarities and differences in the functional trends between both datasets.

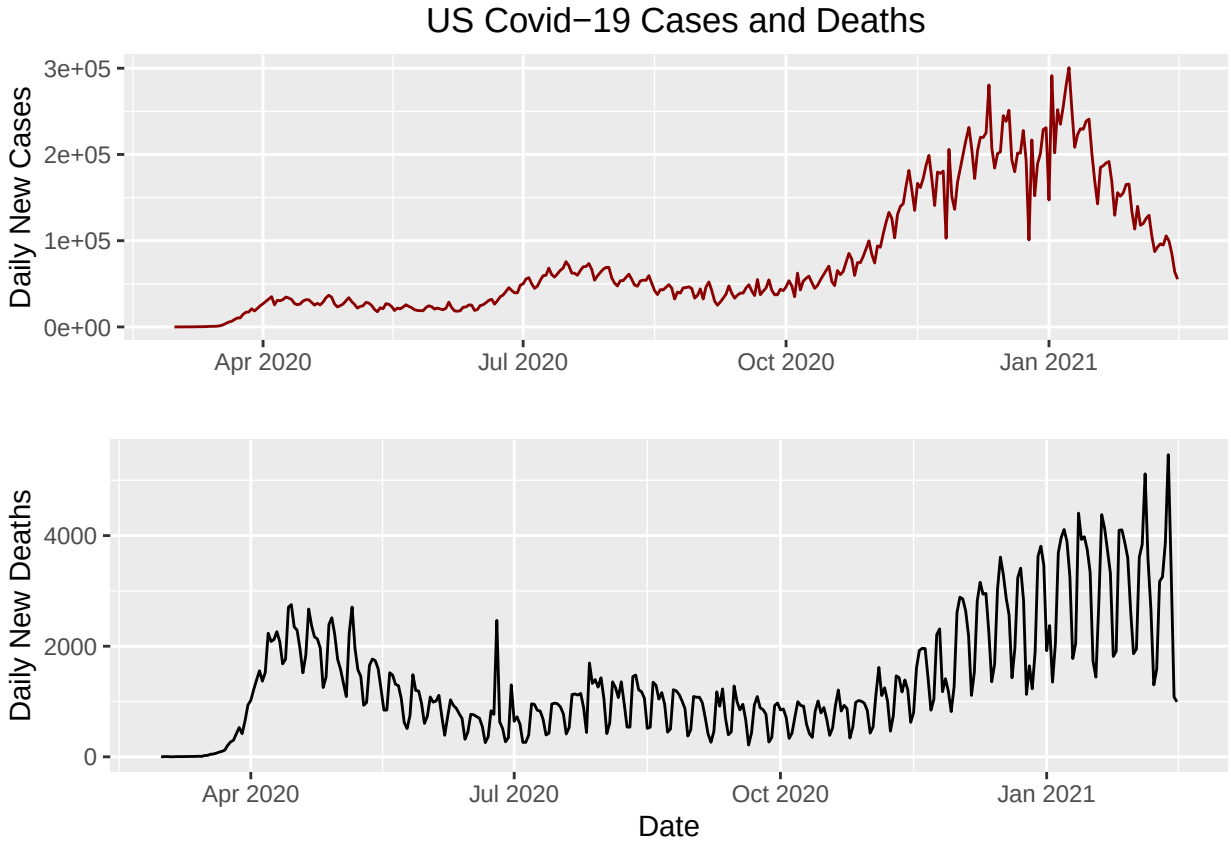


Figure 1 - Timeseries of daily new Covid-19 cases (top) and deaths (bottom) in the United States, March

2020 - February 2021.

Case-Death Scatter Plot

To further examine how the relationship between daily new US covid cases and deaths changed through time, a scatter plot of these two variables was generated (Figure 2). Having previously identified four distinct regimes in the case growth rate, these points were colored by the season for each observation. The scatter plot further highlights the different trends in the daily case to death ratios during each of these time periods. In the spring 2020, there is an significantly evident positive correlation in the case to death ratio. The points in this season are all closely clustered with the deaths growing exponentially with cases. In the summer 2020 data, there is no clear positive or negative trend in the correlation between the two parameters with the points scatter roughly in a circle. In the fall season, again a positive correlation is evident, however the case to death ratio is four to five times that observed in the spring. Finally in the winter (December 2020 – February 2021), the ratio of cases to deaths decreases but is still positive.

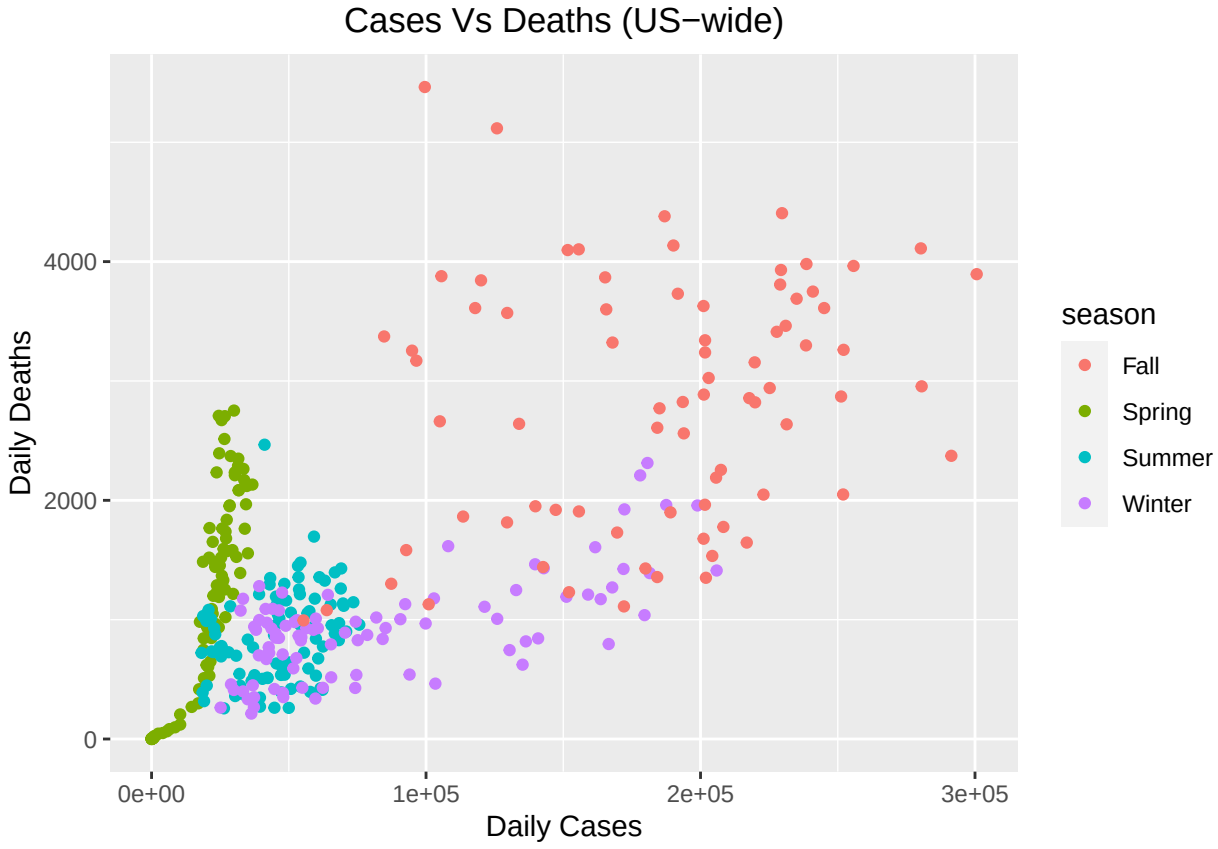


Figure 2 - Scatterplot of daily new Covid-19 cases vs deaths in United States. Point colors indicate season

State Case Rate Box Plots

Description of box plots blah blah blah blah blah

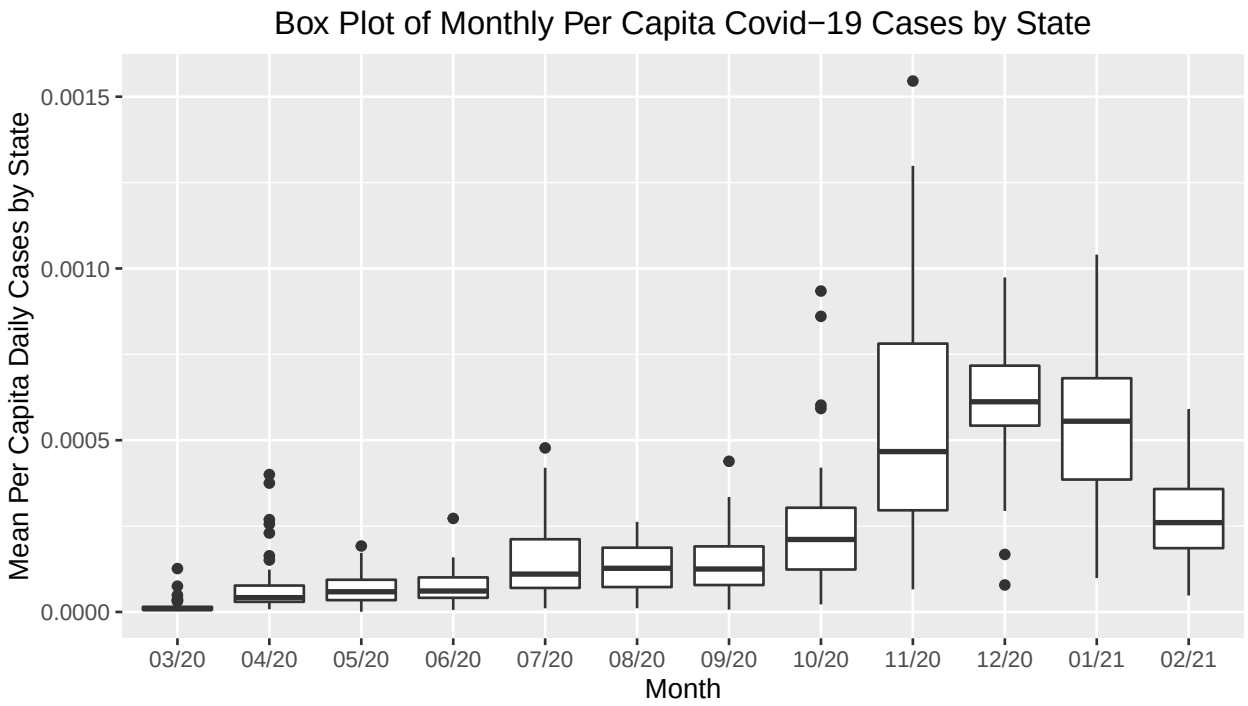


Figure 3 - Boxplot of daily average new Covid-19 cases by state across US. Data is binned by month

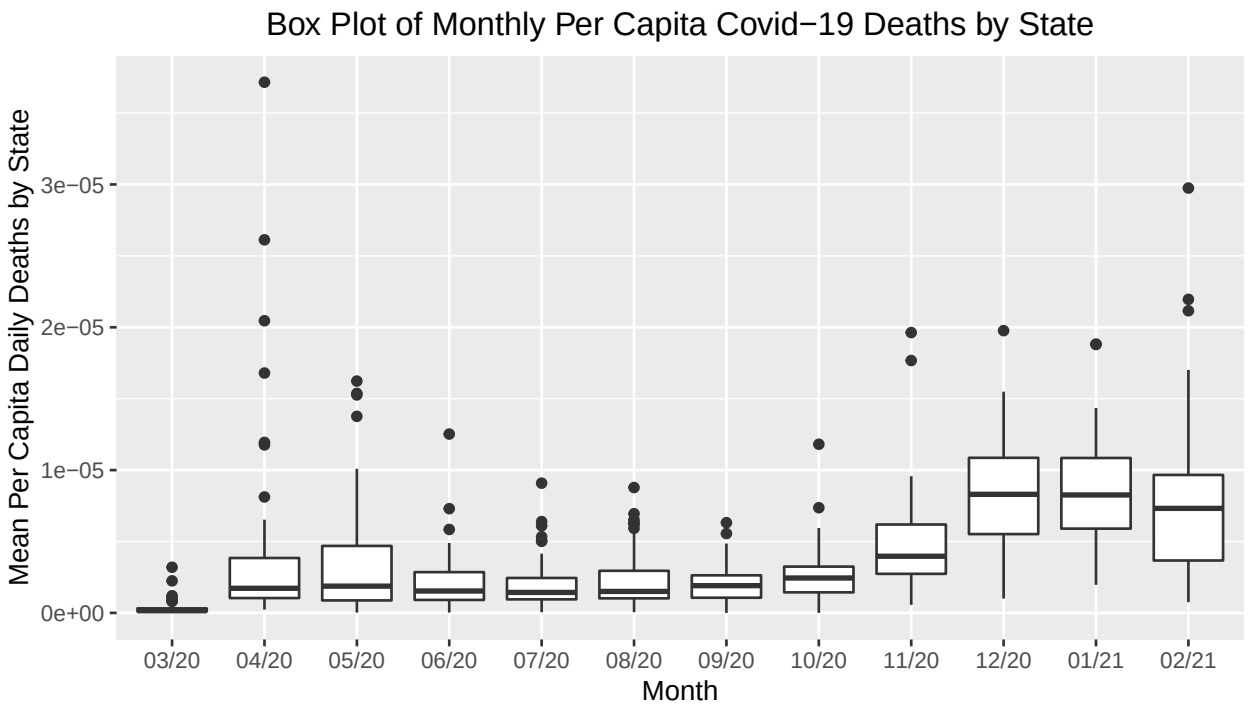


Figure 4 - Boxplot of daily average new Covid-19 deaths by state across US. Data is binned by month

Regional Trends

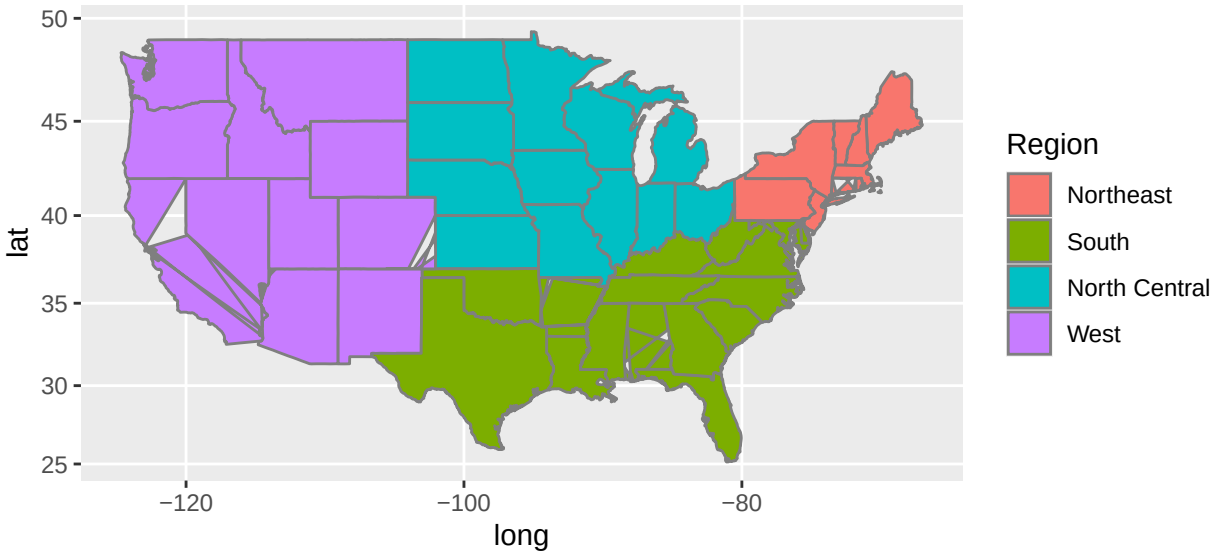


Figure 5 - Map of US States grouped by region

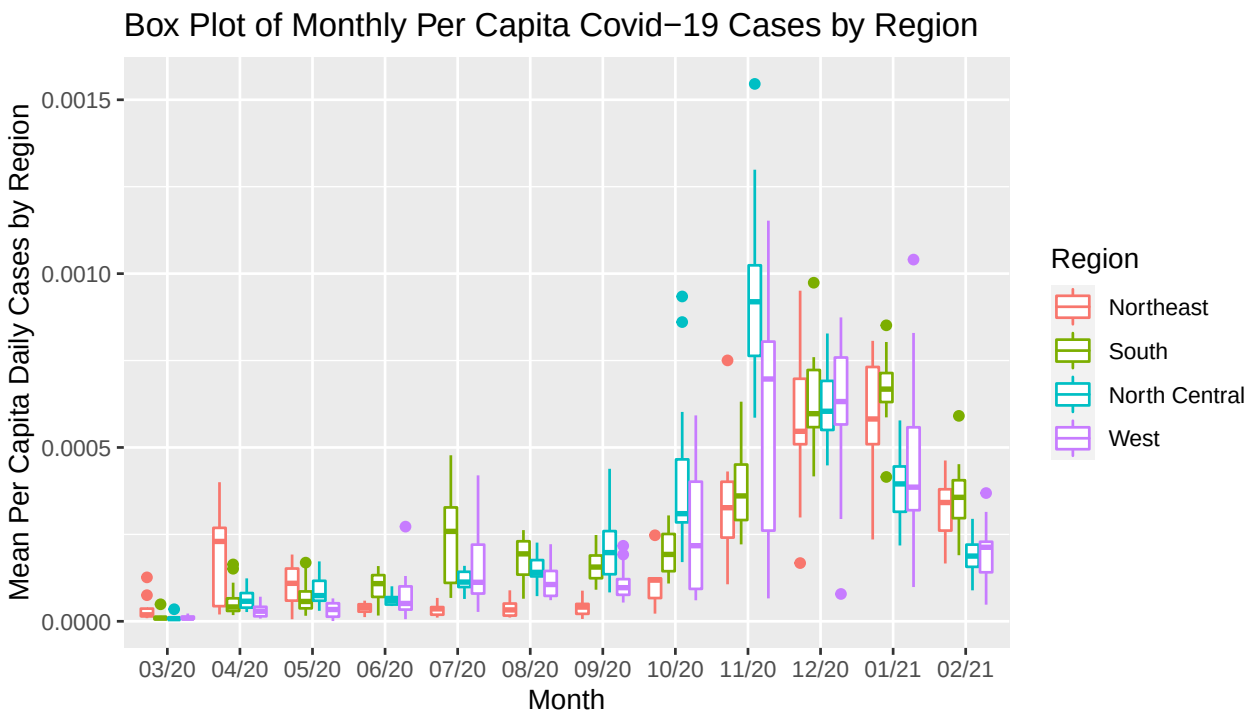


Figure 6 - Boxplot of daily average new Covid-19 cases by state across US. Data is binned by month and categorized by region.

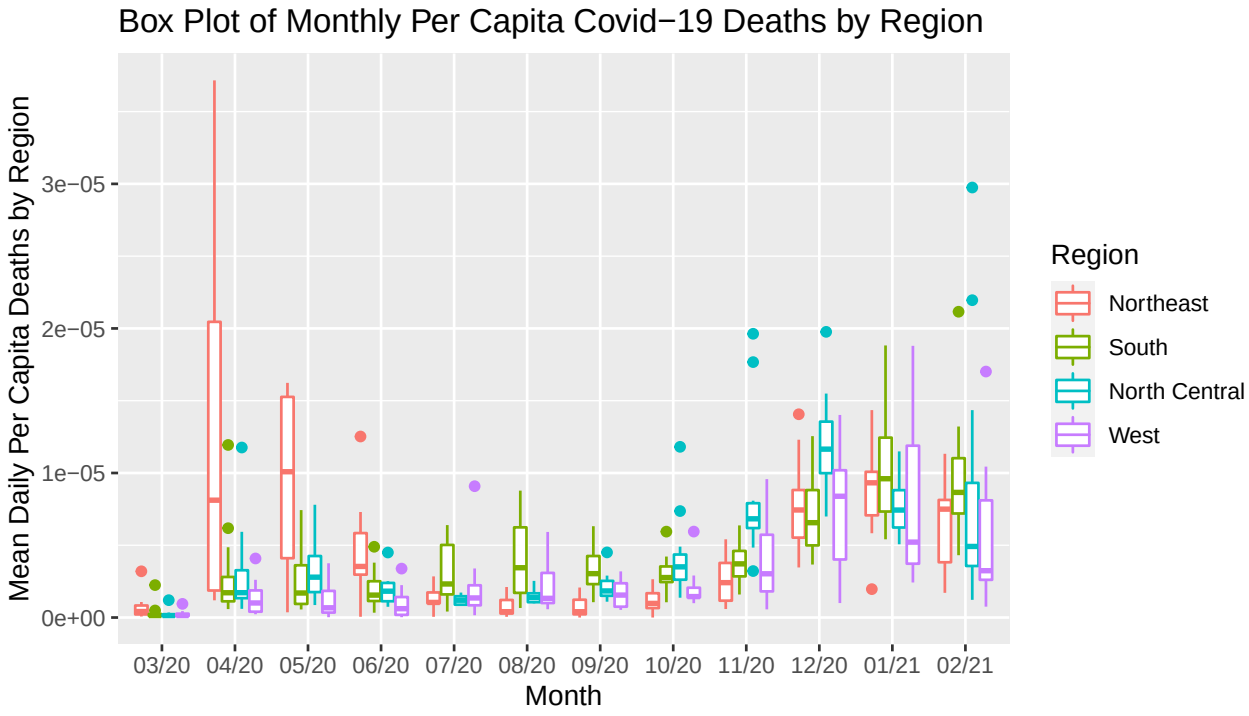


Figure 7 - Boxplot of daily average new Covid-19 deaths by state across US. Data is binned by month and categorized by region.

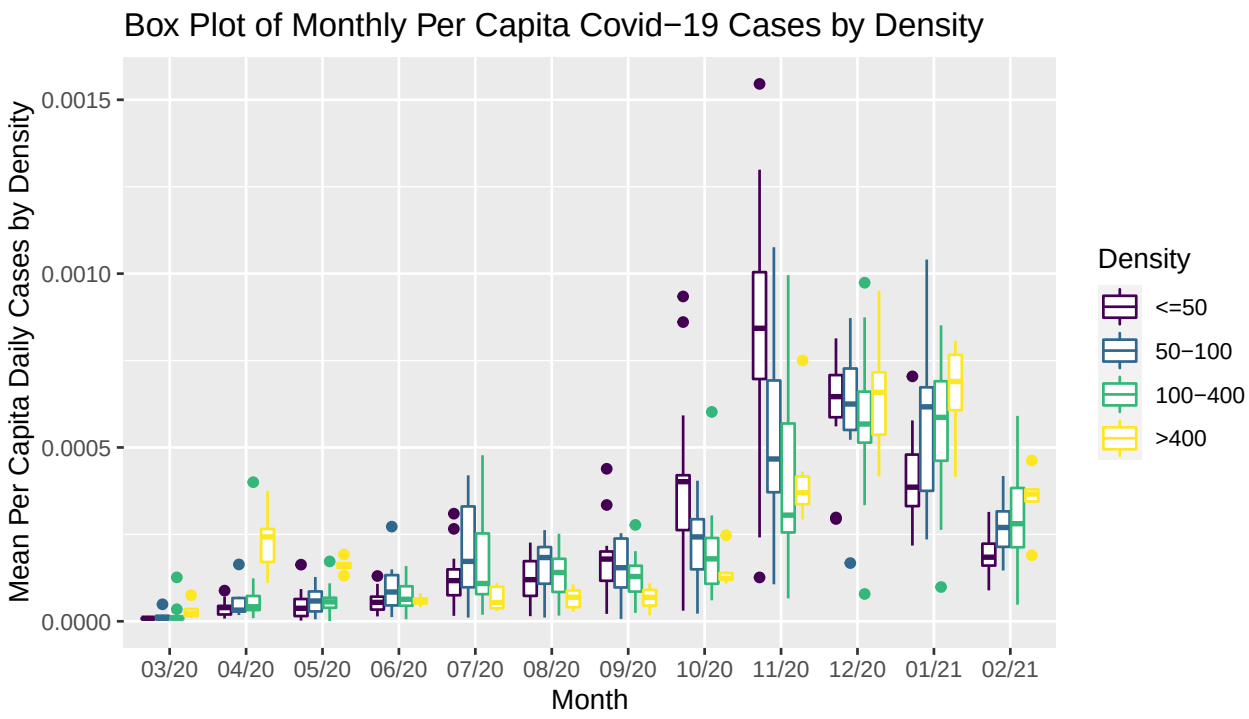


Figure 8 - Density Case Boxplot

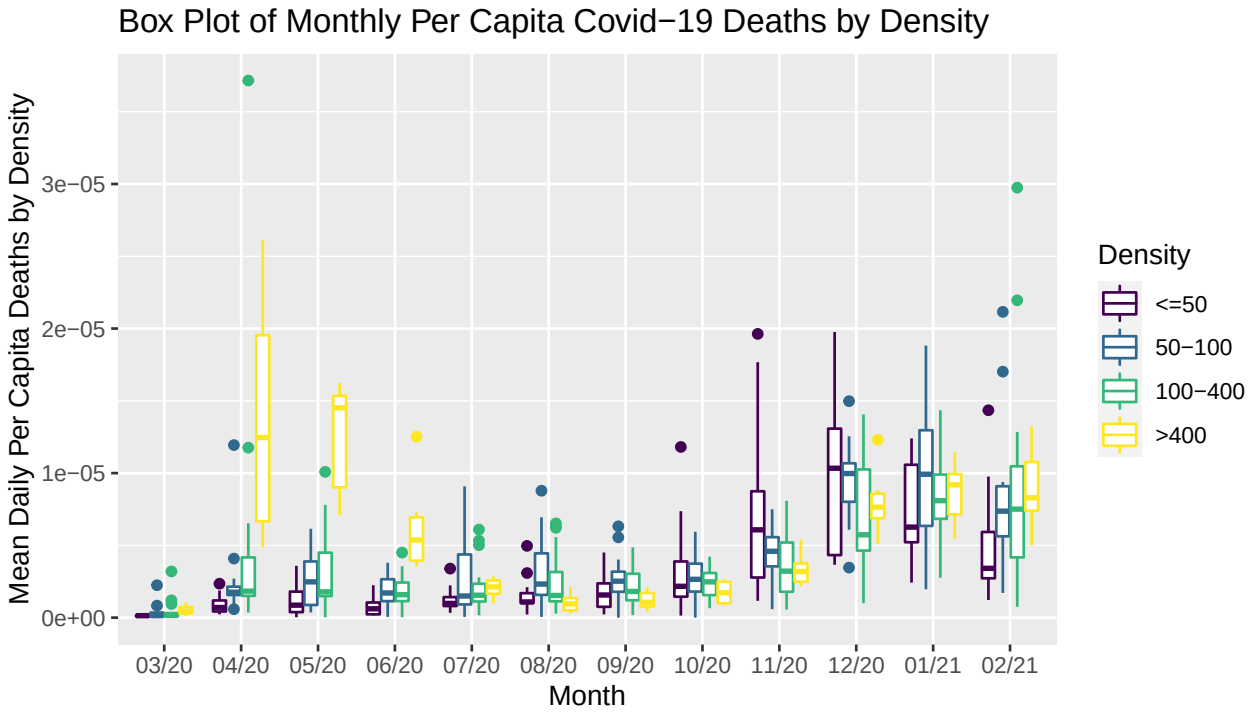


Figure 9 - Density Death Boxplot

Algorithms

ARIMA Forecast Model

Utilizing the timeseries data of US Covid-19 cases and deaths an ARIMA (Autoregressive Integrated Moving Average) forecast model was developed to predict the change in these two parameters over the subsequent 30 days following the final day that data was reported in this dataset, February 15th, 2021. For this analysis, univariate time series forecasting was conducted; only previous values of a single parameter were used to forecast future parameter values. ARIMA is a family of linear regression models that seek to statistically 'explain' trends in time series data and forecast the future. This model combines the mathematical formulations of two less sophisticated models: a simple autoregressive model (AR) and a moving average (MA) model. In a purely autoregressive model, a modeled value at time t (Y_t) is soley a function of previous lags and modeled coefficients.

$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2}$

Where β_1 is the coefficient of lag 1 that the model estimates, α is in the intercept term and ϵ is the model error

Model Development Process

Cluster Analysis

To better understand the spatial distribution of COVID new cases and new deaths increase across the 50 states in the US, our first question is to ask whether or not the state geographical distance influence the patterns how the cases and deaths increases (measured from times series data for cases and deaths). Thus, we applied a K-shape algorithm to calculate the pair-wise distance between any two state-level time series for both cases and deaths. The K-shape distance algorithm is an computationally efficient accurate method to estimate the distances between time-series data.[1] Then we applied a pearson correlation inference testing for the relationship between the geographical distances between states and the K-shape distances of time series between states.

Then, we applied a partition clustering algorithm and a hierarchical clustering algorithm that is offered by *dtwclust*[2] R package to classify the time series data into K groups in an unsupervised way. Specifically, we used a *Dynamic Time Warping* (DTW) algorithm on the z-normalized time series data. The reason for applying the z-normalization preprocessing is because the time series differences between states has large variance simply due to its population (i.e. large population have large cases increase). However, in this study, we are interested in the relative growth pattern for cases and deaths instead of the absolute increase number of cases/deaths.

With the partition cluster classification for each states, we are interested in if we can predict its cluster membership by the state demographical and political statistics. We scrapped the recent state-level consensus data[3], and applied a multinomial regression model to predict the cluster membership (of cases and deaths respectively) of states by the total polutation, percetage of white citizens, per capital income, median age, political party affiliation based on 2020 election state-level results, and the population density.

[1] Paparrizos, John, and Luis Gravano. "k-shape: Efficient and accurate clustering of time series." Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. 2015.

[2] https://github.com/asardaes/dtwclust

[3] https://www.census.gov/data/developers/data-sets.html

$$
\begin{aligned}
f(k,i) = \beta_{0,k} &+ \beta_{1,k} X_{population,i} \\
&+ \beta_{2,k} X_{whitepercentage,i} \\
&+ \beta_{3,k} X_{income,i} + \beta_{4,k} X_{age,i} \\
&+ \beta_{5,k} X_{partyaffiliation,i} + \beta_{6,k} X_{populationdensity,i}
\end{aligned}
$$

Where $f(k,i)$ is the probability that observation i has outcome k, and $\beta_{m,k}$ is the regression coefficient for the mth predictor for the outcome k.

Results

AR Covid Case Forecast Model

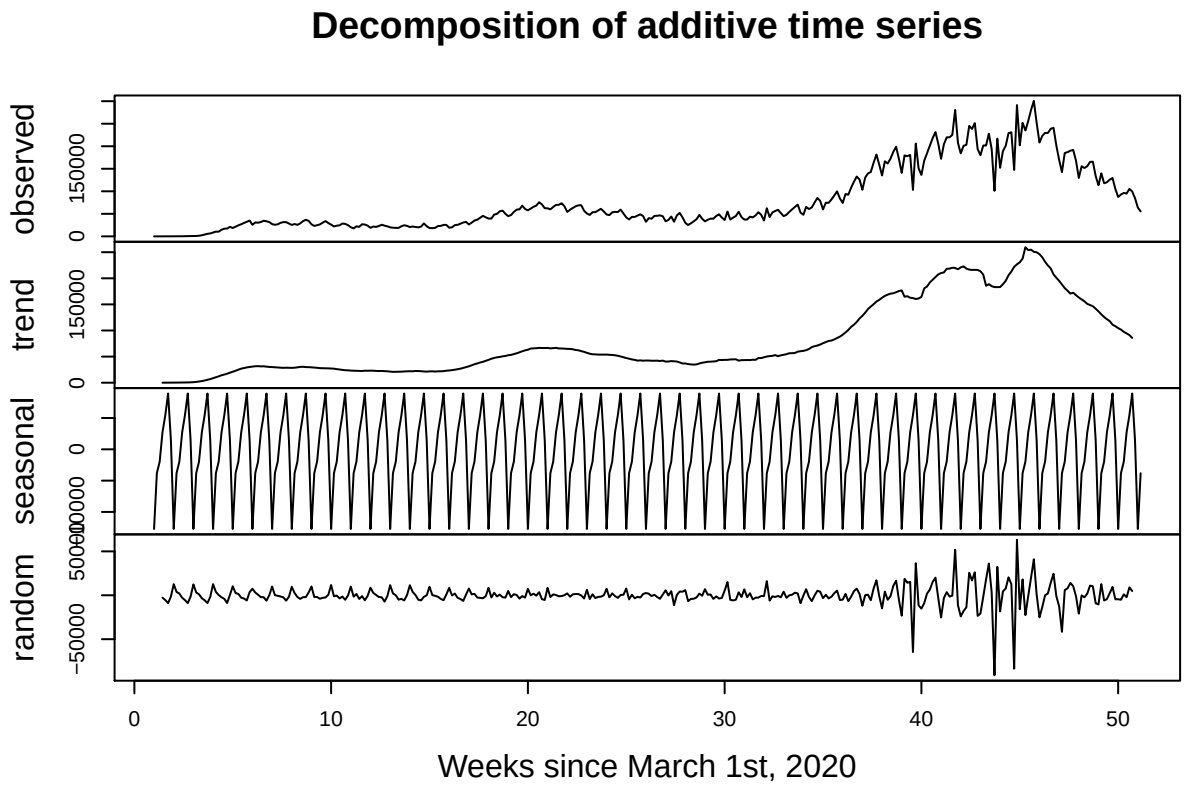


Figure 10 - Decomposition of Covid Case Data

Test for Stationarity

```
# Test for Stationarity
tseries::adf.test(new_cases.ts)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  new_cases.ts
## Dickey-Fuller = -0.68928, Lag order = 7, p-value = 0.971
## alternative hypothesis: stationary
```

It is not stationary!

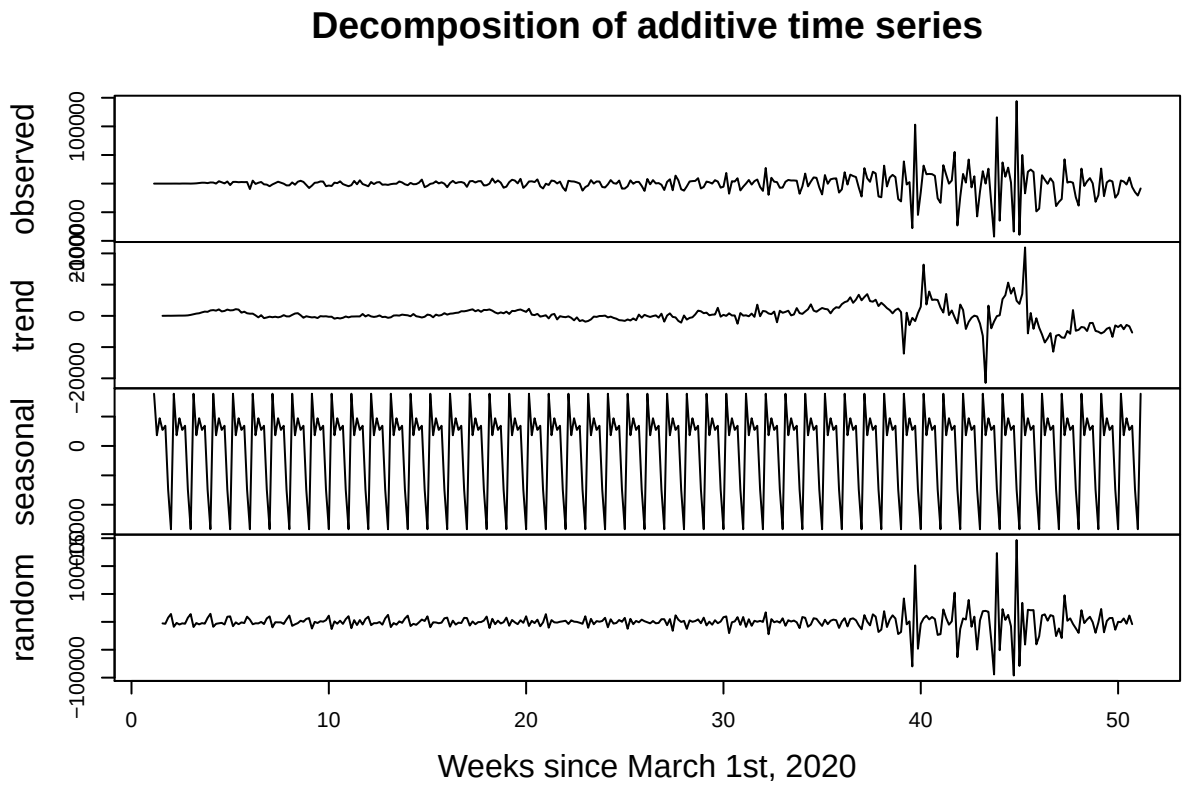


Figure 10 - Decomposition of Difference Covid Case Data

```
# Test for Stationarity with differencing
tseries::adf.test(diff(new_cases.ts))
```

```
## Warning in tseries::adf.test(diff(new_cases.ts)): p-value smaller than printed
## p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  diff(new_cases.ts)
## Dickey-Fuller = -5.2524, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

This data is stationary!

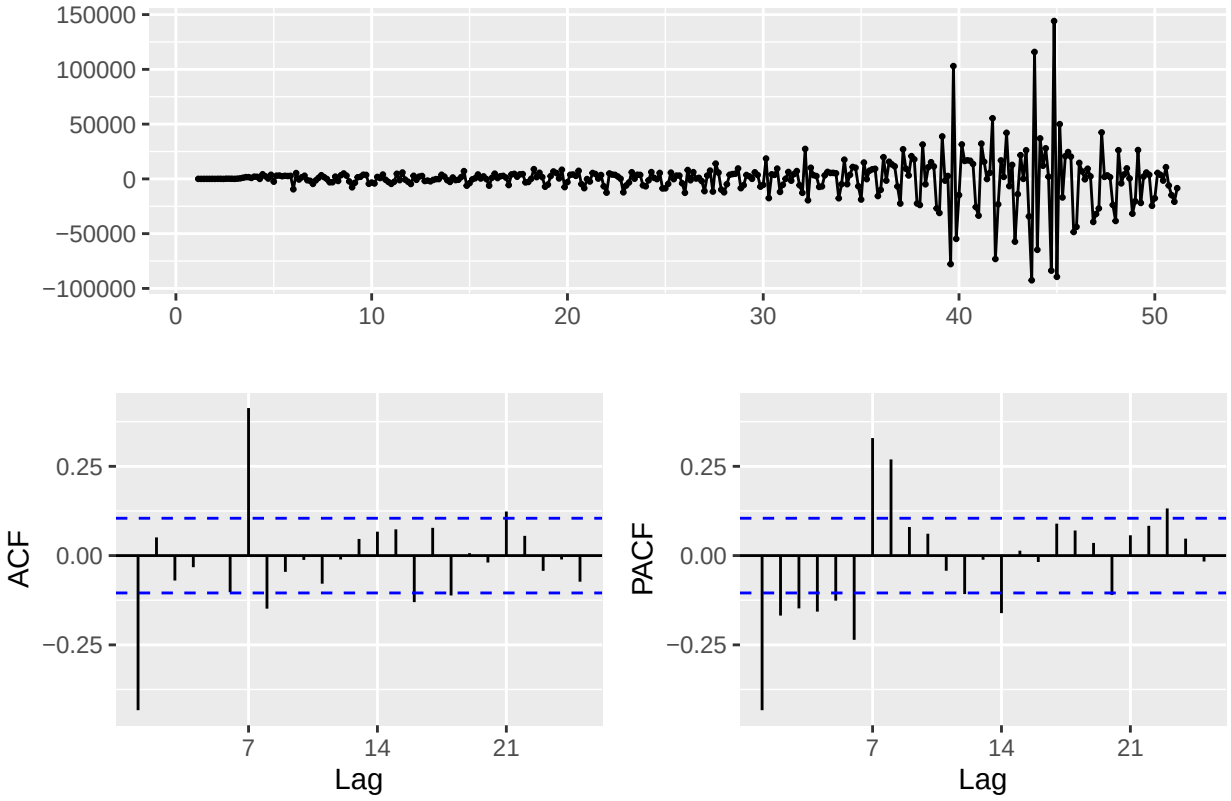


Figure 11 - Autocorelation function and Partial Autocorrelation Function for difference covid case data

Checking Residuals

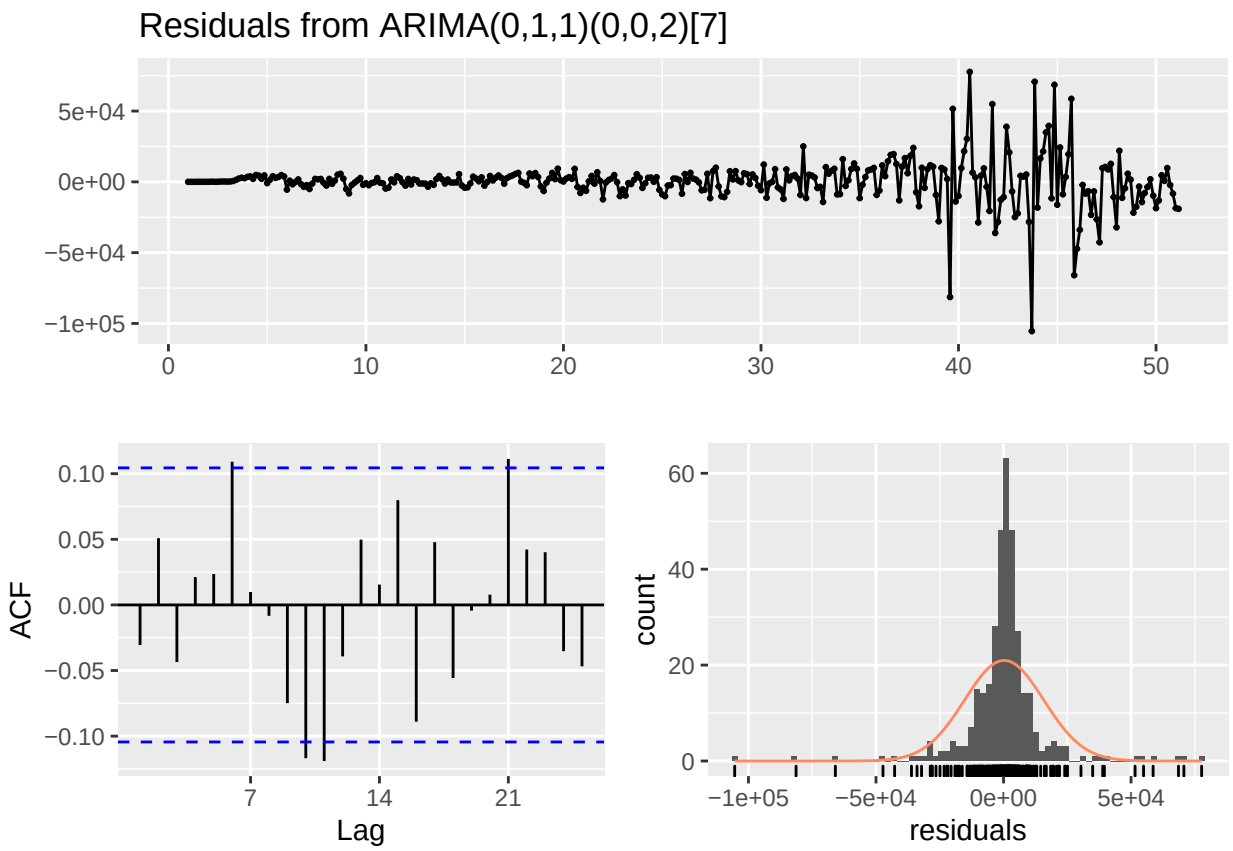


Figure 11 - Autocorelation function and Partial Autocorrelation Function for difference covid case data

Visualizing Forecast

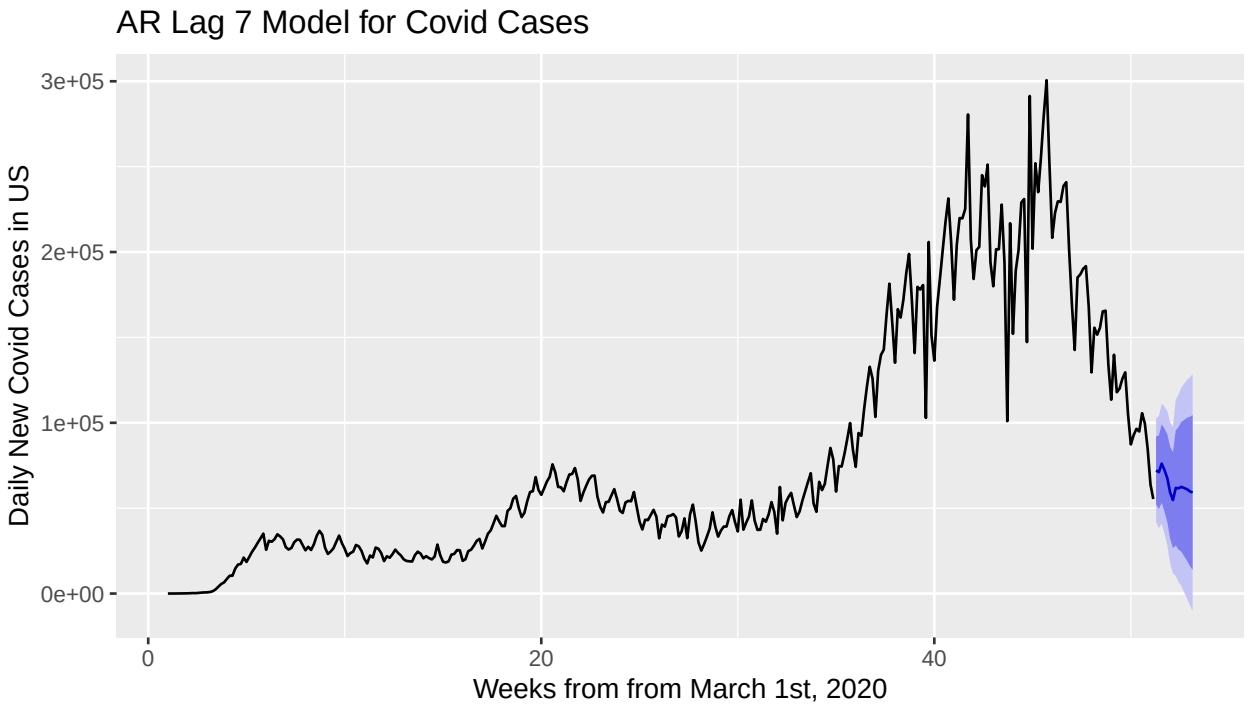
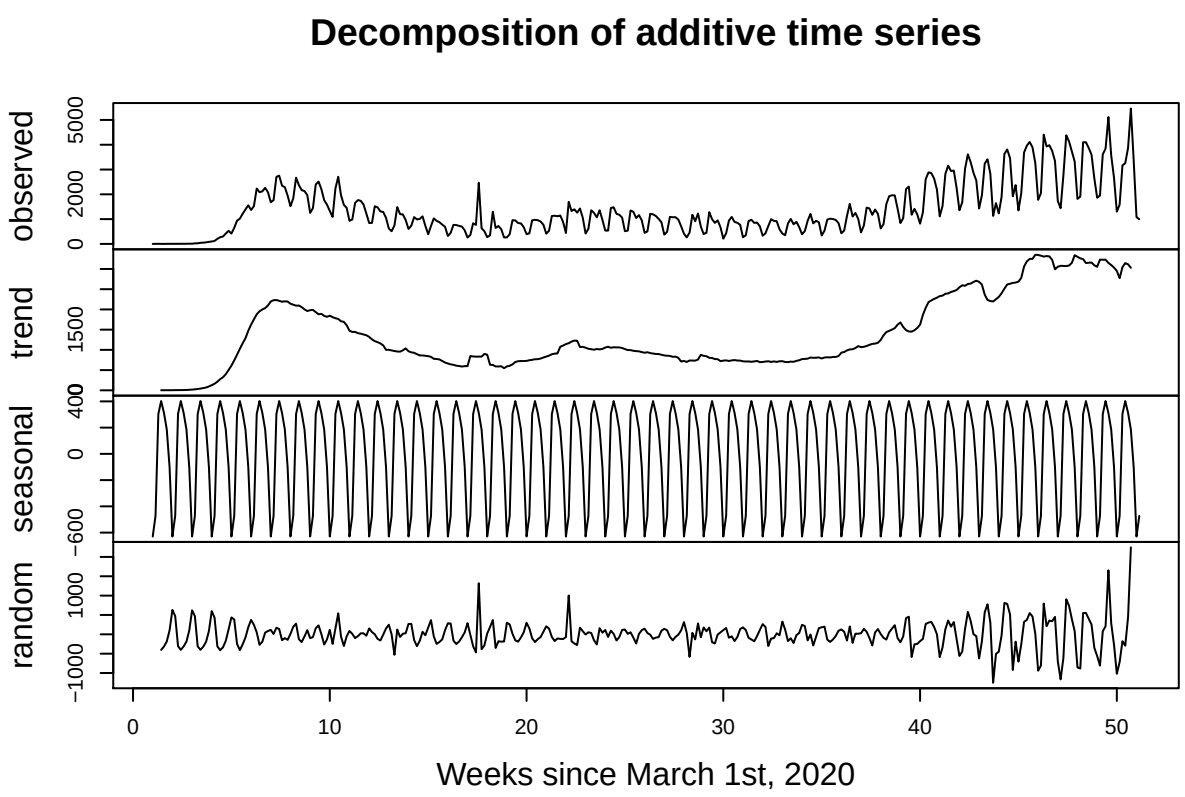


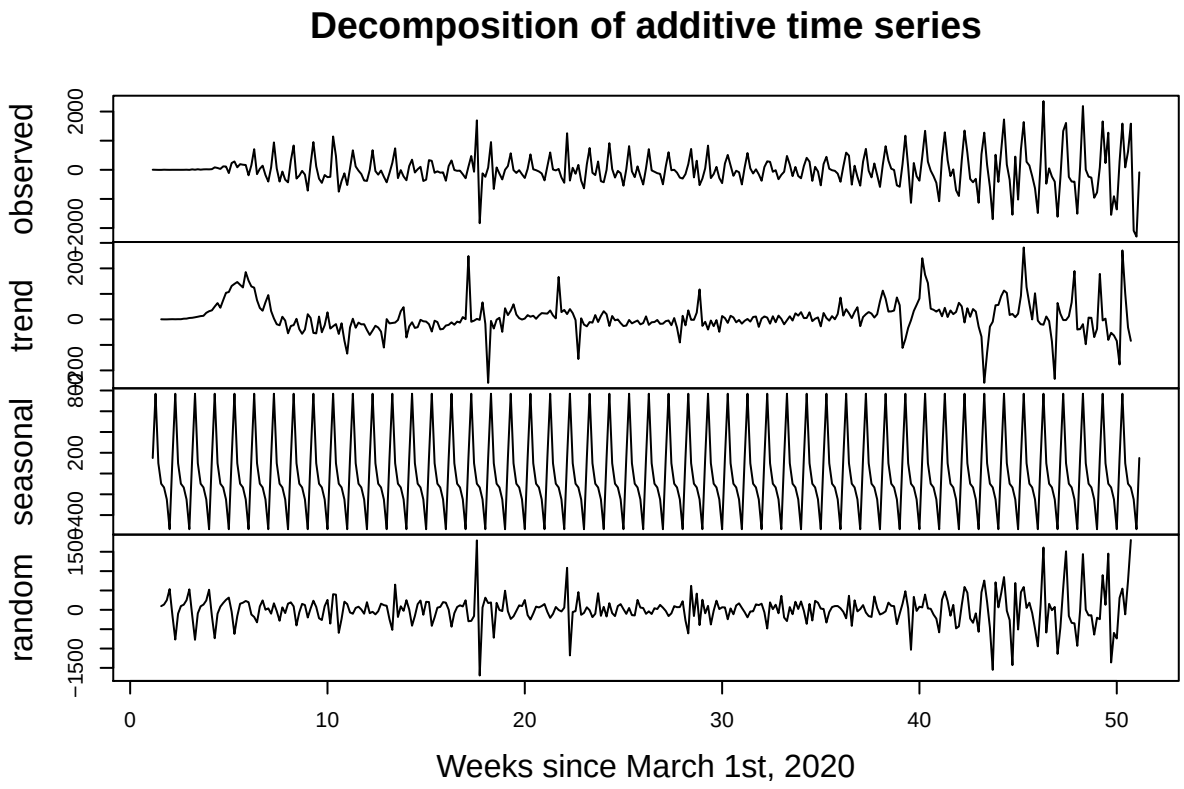
Figure 12 - 30 Day forecast for Covid-19 cases in US using AR Lag-7 Model. Dark blue indicates model results, light blue bars indicate 95

AR Covid Death Forecast Model




```
# Test for Stationarity
tseries::adf.test(new_deaths.ts)
```

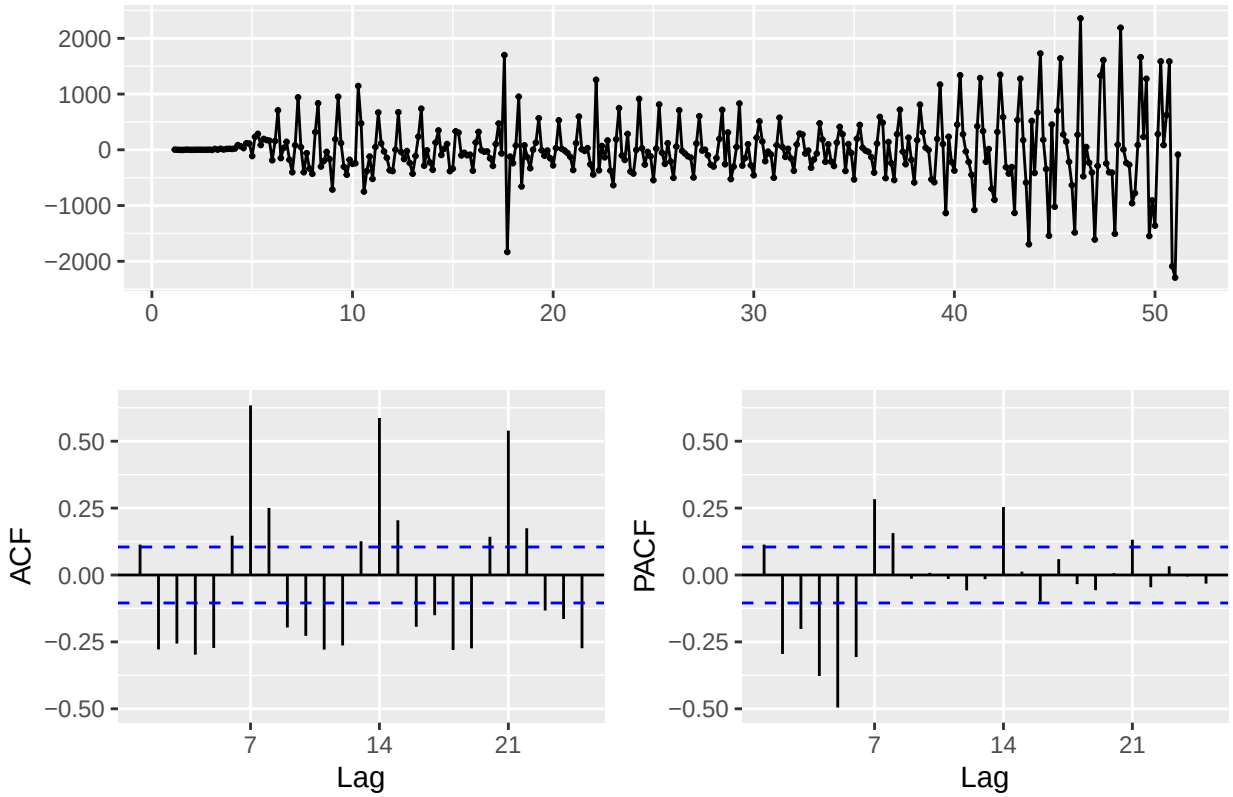
```
##
##  Augmented Dickey-Fuller Test
##
## data:  new_deaths.ts
## Dickey-Fuller = -1.1801, Lag order = 7, p-value = 0.9087
## alternative hypothesis: stationary
```



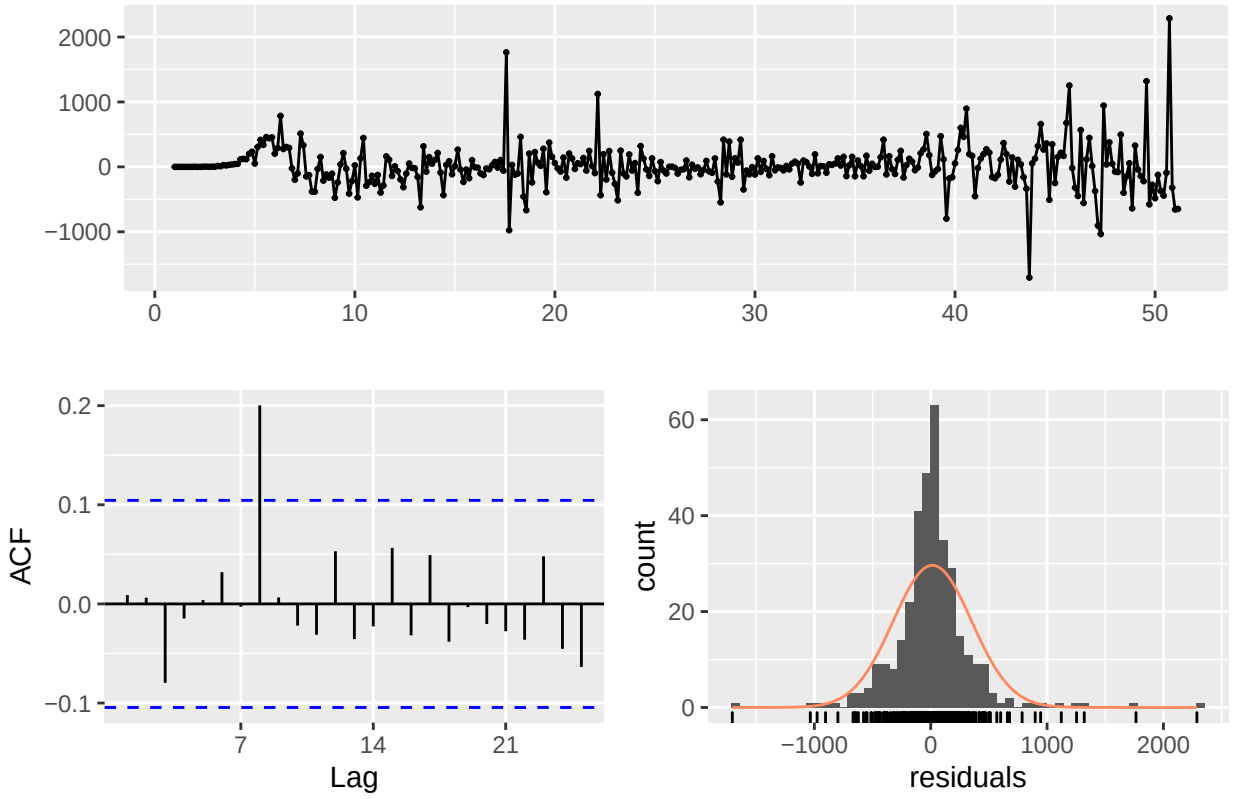

```
# using differencing method
tseries::adf.test(diff(new_deaths.ts))
```

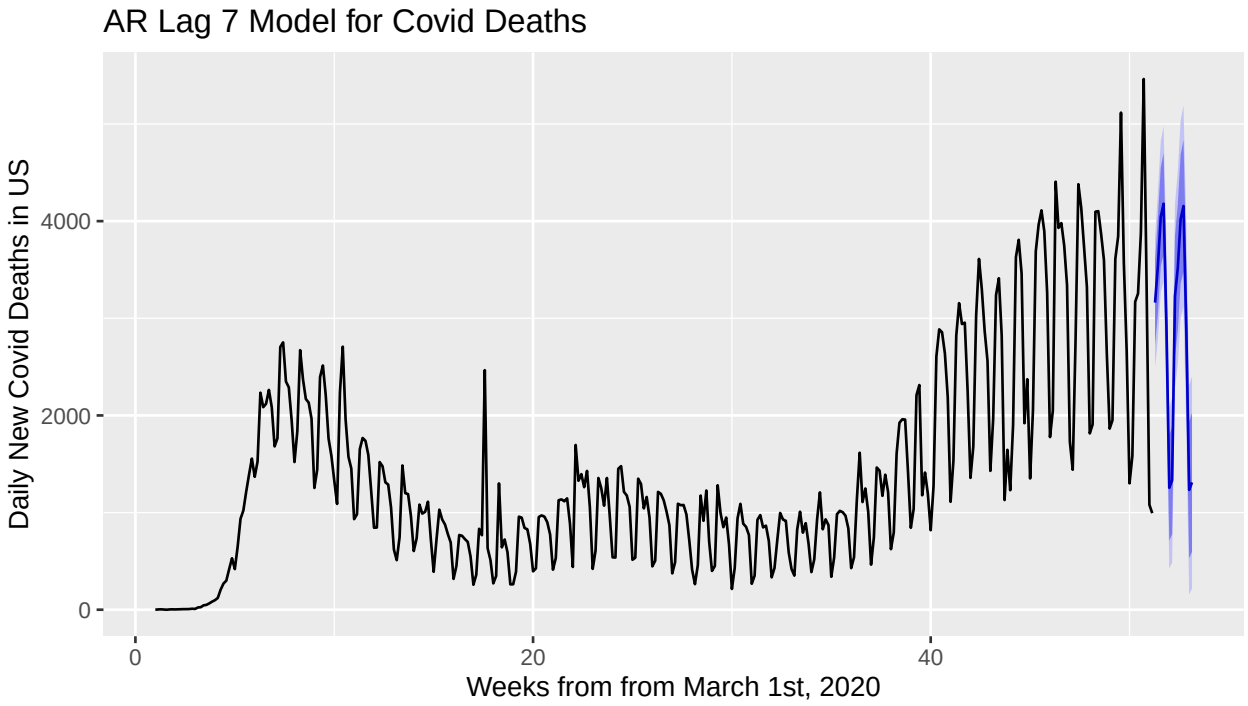
```
## Warning in tseries::adf.test(diff(new_deaths.ts)): p-value smaller than printed
## p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  diff(new_deaths.ts)
## Dickey-Fuller = -8.1672, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```



Residuals from ARIMA(1,0,2)(0,1,1)[7]




AR Lag 7 Model for Covid Deaths
Daily New Covid Deaths in US
4000
2000
0
0
20
40
Weeks from from March 1st, 2020

Spatial analysis

We first smoothed the time-series data for better esitimate the K-shape distance and clustering analysis. Below we plotted the new cases/deaths time series (log transformed) data for each states.

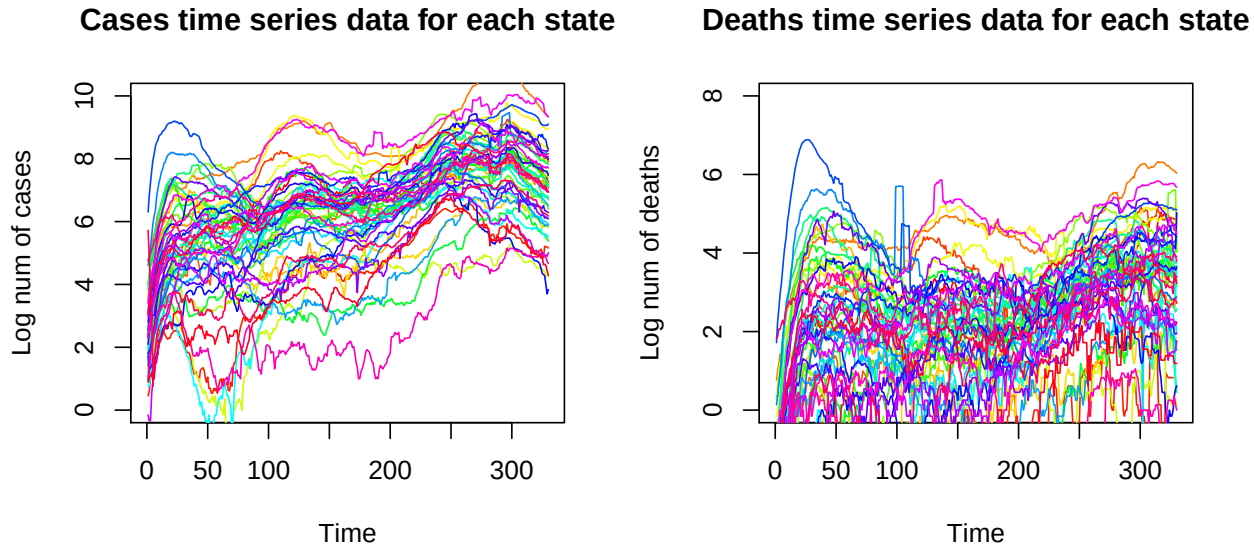
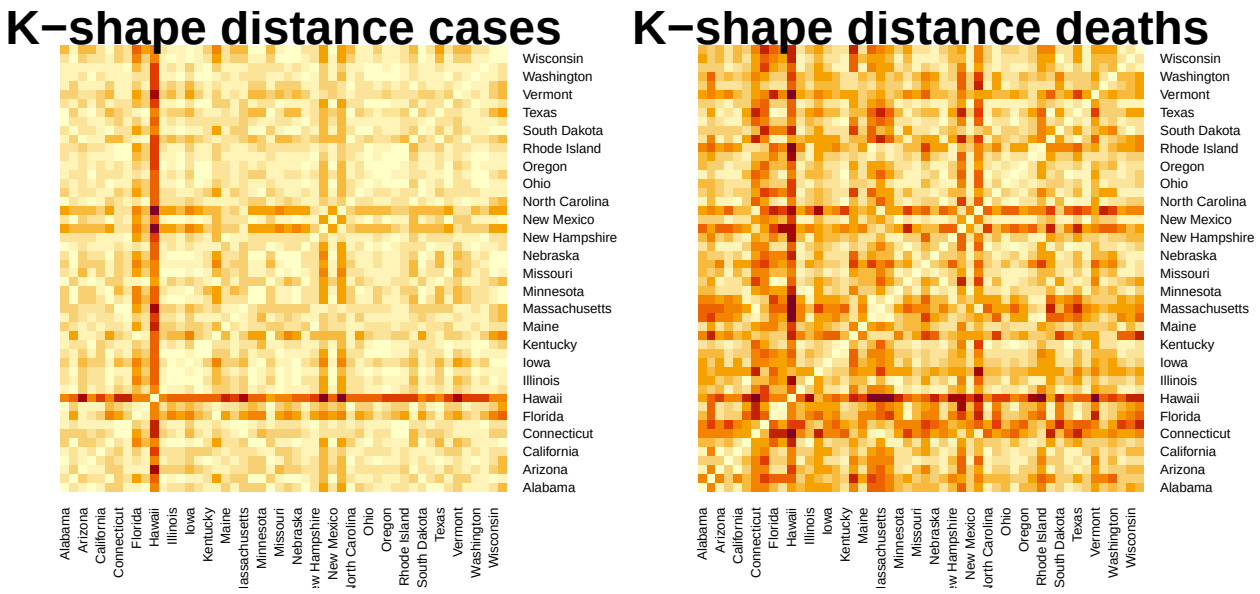


Figure <PLACE HOLDER> - Timeseries of smoothed and log transformed daily Covid-19 cases (left) and deaths (right) in the United States.

We hypothesized that the spreading of the COVID-19 virus over states in the US is geographical-dependent. Thus we want to know the relationship between the shape of the cases/deaths growth and the geographical distance among states. First, we calculated the K-shape distance between the time series of any two states. Then, the esimated K-shape distances for cases and deaths as well as the geographical distances between states is then plotted in a adjacency matrices.



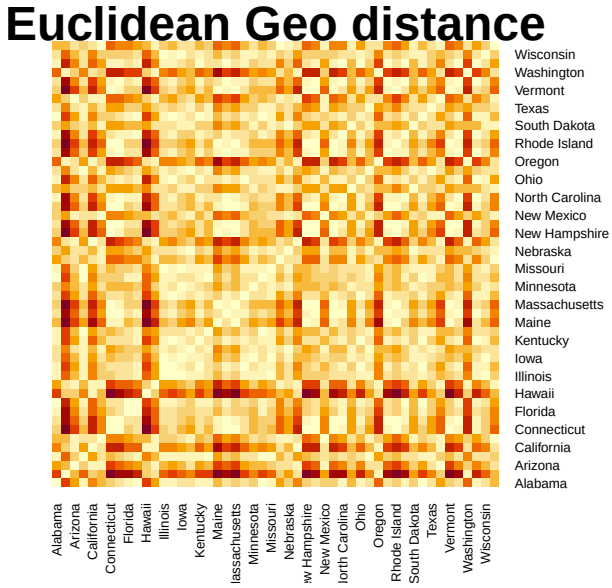


Figure <PLACE HOLDER> - Heapmaps for the K-shape distances for cases and deaths and geographical distances between states

We tested the correlation between the geographical distances among states and the K-shape distance among the states. The results show that:

```
##
##  Pearson's product-moment correlation
##
## data:  df$case and df$geo_dis
## t = 11.184, df = 1223, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2529146 0.3545684
## sample estimates:
##       cor
## 0.3046086
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$death and df$geo_dis
## t = 10.6, df = 1223, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2379327 0.3405521
## sample estimates:
##      cor
## 0.290076
```

The results supported our hypotheses that the geographical distances amongs states is correlated with the growth pattern of cases ($r = 0.305$, $p < 0.001$) and deaths ($r = 0.290$, $p < 0.001$), as shown in Figure below.

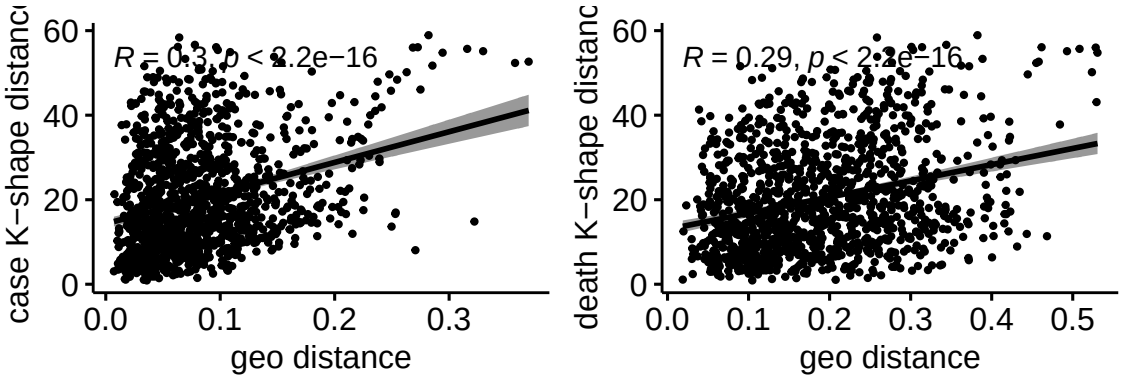


Figure <PLACE HOLDER> - Scatter plot and fitted regression for the geo distances against the K-shape distances of time series for cases and deaths

In order to find out the best clusters for the clustering analysis on the time series. We repeated fitted the partition clustering algorithm and estimate the Clustering Validity Indices (CVI) for difference size of clusters K. As shown below:

```
## [1] "Partition clustering for cases"
```

```
##                 k_5          k_6          k_7          k_8          k_9
## Sil    2.427002e-01 2.281518e-01 9.721031e-02 1.070423e-01 1.952345e-01
## SF     1.719163e-05 6.259975e-05 7.007379e-07 4.652564e-07 4.077302e-07
## CH     1.952300e+01 1.514802e+01 1.178695e+01 1.050432e+01 9.819983e+00
## DB     1.243350e+00 1.097612e+00 1.636101e+00 1.436401e+00 1.133072e+00
## DBstar 1.608476e+00 1.251101e+00 2.147506e+00 2.005032e+00 1.438085e+00
## D      1.086919e-01 1.782714e-01 7.636712e-02 7.733620e-02 2.242977e-01
## COP    1.145331e-01 1.075595e-01 1.181281e-01 1.160932e-01 9.090545e-02
```

```
## [1] "Partition clustering for deaths"
```

```
##                 k_5          k_6          k_7          k_8          k_9
## Sil    2.280162e-01 1.695672e-01 1.095635e-01 1.228939e-01 1.273682e-01
## SF     1.510344e-08 3.165868e-11 4.665157e-13 2.797762e-14 2.442491e-15
## CH     1.318438e+01 1.137754e+01 9.563044e+00 8.840770e+00 8.449628e+00
## DB     8.198211e-01 1.348596e+00 1.374002e+00 1.373717e+00 1.320057e+00
## DBstar 9.447234e-01 1.401937e+00 1.763292e+00 1.637294e+00 1.495256e+00
## D      1.629750e-01 1.759680e-01 1.942755e-01 1.955405e-01 2.256621e-01
## COP    2.099462e-01 1.864132e-01 1.893187e-01 1.722067e-01 1.565607e-01
```

This analysis indicates that for the partition clustering analysis, the best cluter size is around 6 for both cases and deaths. Thus we choose $K = 6$. We applied a partition clustering algorithm based on estimated DTW distances between the time-series. The classification for time series at different states into distinct clusters are shown in Figure . We also applied a hierarchical clustering analysis on the time series, the results are shown in Figure . In order to intrepret the growth pattern for each cluster, we plotted the centroid time series (measured by DTW barycenter averaging[4]). As shown in Figure , different clusters for cases and deaths are different for 1) the time point for different waves of COVID-19 spreading. 2) the relative

[4]Petitjean, François, Alain Ketterlin, and Pierre Gançarski. "A global averaging method for dynamic time warping, with applications to clustering." Pattern recognition 44.3 (2011): 678-693.

cases/deaths increase for each waves. These growth patterns systematically differ among states based on their geographical position.

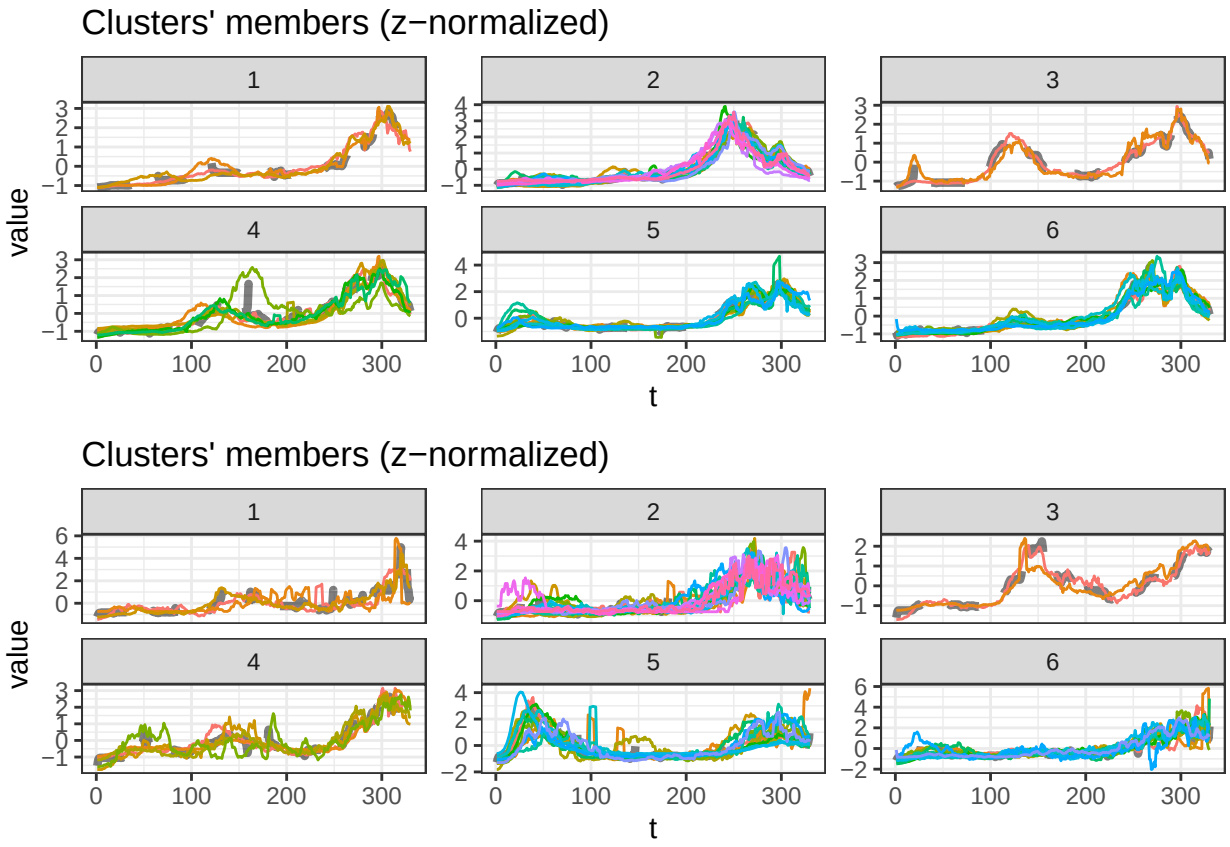


Figure <PLACE HOLDER> - Clustering analysis for cases (Upper) and death (Bottom) with k=6

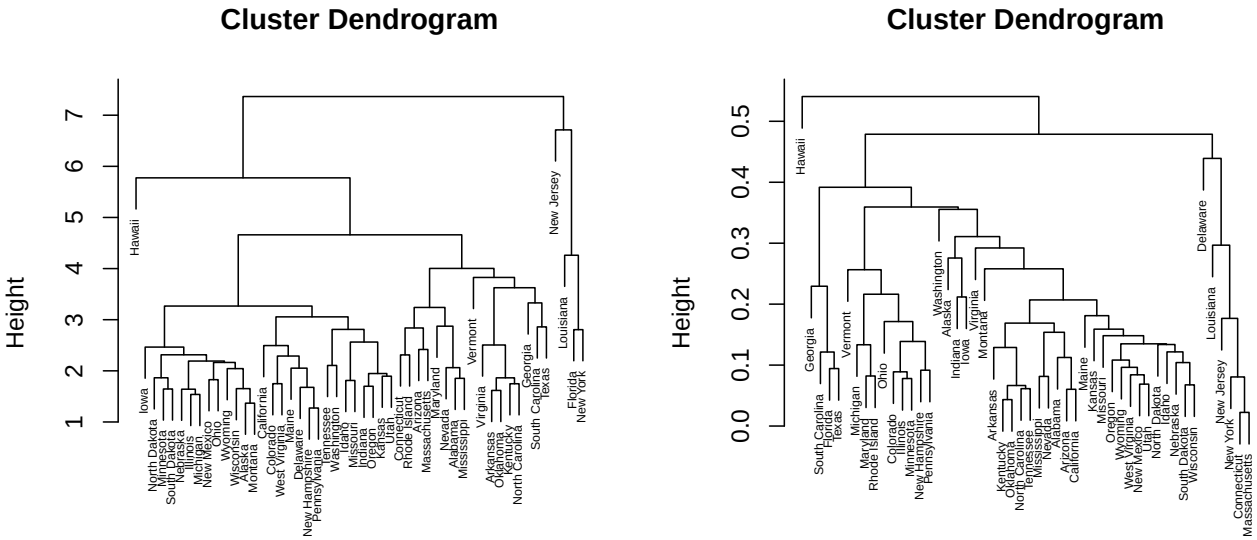


Figure <PLACE HOLDER> - Hierarchical clustering analysis results - binary tree plot for cases (Left) and death (Right)

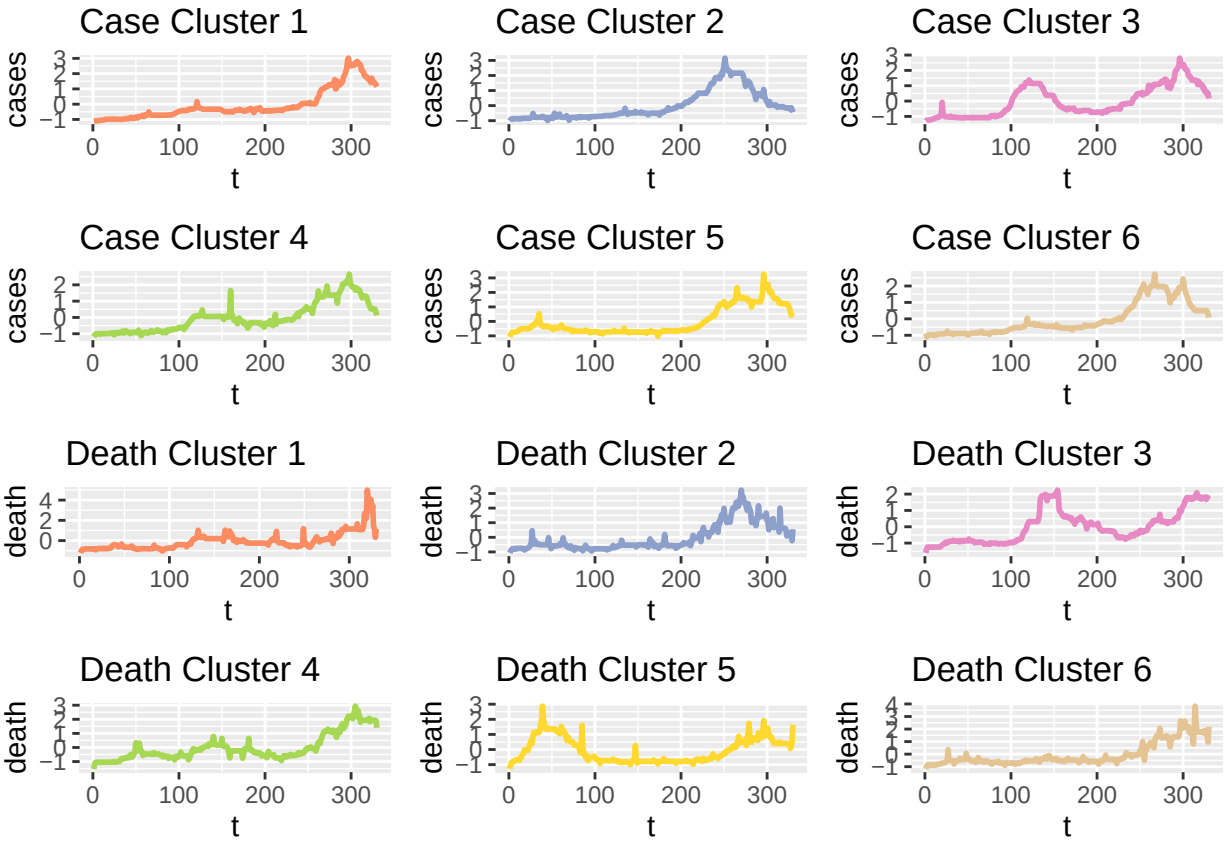


Figure <PLACE HOLDER> - The centroid time series for different clusters of cases (upper two rows) and deaths (bottom two rows)

With the clutering results, we plotted the cluster membership for each state in a US map. As shown in Figure , the 6 clusters is spatially distributed in a systematic and predictable manner.

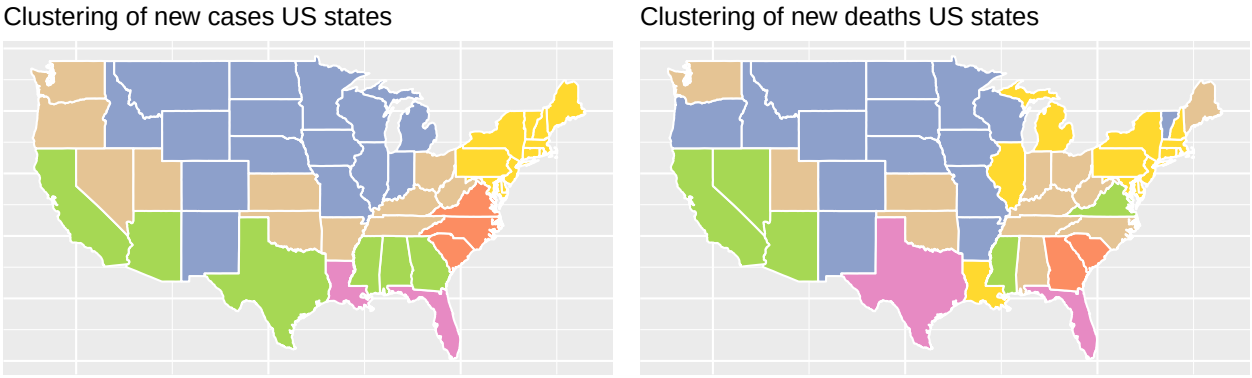


Figure <PLACE HOLDER> - The US state maps color filled by the cluster memebership for case and death growth

Thus, we predicted that the pattern of cases and deaths growth of each state should be predicted by the demographical and political information for each state. We fitted a multinomical statistical testing to predict

the cluster membership. The predictors include total population, percent white, median age, per capita income, political party affiliation, and the population density. For the summary of the two full models.

```
## [1] "The summary for multinomial regression model for case"
```

```
## Call:
## multinom(formula = cluster_case ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale,
##     data = us_data_2020_subset, model = TRUE)
##
## Coefficients:
##   (Intercept) total_population_scale percent_white_scale
## 2   -3.523898             -10.680197           7.9839377
## 3   -9.598203             -20.183907         -29.3839346
## 4   -4.764977              -1.101475          -0.1842134
## 5  -73.707712              -3.122757          18.4151516
## 6   -3.266475             -11.689754           7.7792892
##   per_capita_income_scale     vote1  density median_age_scale
## 2               -9.847059  19.93123 13.07925       -10.479705
## 3              -12.885659 -93.78982 89.96034       -29.357739
## 4              -10.201007  10.48114  8.53031        -9.285471
## 5               23.645953  46.25958 52.80814        31.840571
## 6              -10.903175  20.34255 16.09125       -10.683255
##
## Std. Errors:
##   (Intercept) total_population_scale percent_white_scale
## 2    2.517193              10.528552            5.723430
## 3   13.220036              13.958305           27.453271
## 4    3.040129               1.793308            1.503059
## 5   13.645758               3.006947            7.252484
## 6    2.477410              10.588449            5.691011
##   per_capita_income_scale     vote1   density median_age_scale
## 2                7.667594  14.40289 15.370556         7.764069
## 3               31.442918 106.92157 96.270527        34.985872
## 4                6.651477   7.91356  7.734387         6.885724
## 5               13.596642  13.64580 80.976307        16.681916
## 6                7.700071  14.44687 15.397997         7.771521
##
## Residual Deviance: 53.95988
## AIC: 123.9599
```

```
## [1] "The summary for multinomial regression model for death"
```

```
## Call:
## multinom(formula = cluster_death ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale,
##     data = us_data_2020_subset, model = TRUE)
##
## Coefficients:
##   (Intercept) total_population_scale percent_white_scale
## 2  -2.8809563              4.0251518          3.02372906
## 3 -10.7677462              8.3127728         -1.70953107
## 4   0.6042385              3.3340421          0.02640449
```

```
## 5   1.3217932             -0.5889042          2.19332741
## 6   3.1916647              4.2271236          4.18973048
##   per_capita_income_scale      vote1     density median_age_scale
## 2               3.4365575  -2.840577 -24.8976874      -0.38632168
## 3              -3.4590219 -23.559925   2.0428774      -1.83600671
## 4               3.5075138  -2.541278 -13.5788881      -0.02911417
## 5               0.8952054   1.575004   0.6199005      -0.08120682
## 6               2.9362365  -4.530928 -11.9762720      -0.75874714
##
## Std. Errors:
##   (Intercept) total_population_scale percent_white_scale
## 2    3.909727               5.229011            2.157974
## 3   66.659743             199.450147          194.118629
## 4    2.760000               4.074744            2.260767
## 5    2.748174               1.932406            1.760382
## 6    2.684460               4.175240            2.137636
##   per_capita_income_scale     vote1     density median_age_scale
## 2                2.859498  4.110357   10.663620         2.900057
## 3              123.779581 89.389101 1028.559697       361.214404
## 4                2.919025  4.191116    8.349647         3.072747
## 5                2.424181  3.102355    2.804091         2.574971
## 6                2.792203  3.968053    8.045517         2.850869
##
## Residual Deviance: 56.73566
## AIC: 126.7357
```

With the coefficients and the standard error extracted from the fitted model, we calcualted the p values for each of the predictors. See the results below.

```
## [1] "The p values for each predictors for case"
```

```
##    (Intercept) total_population_scale percent_white_scale
## 2 1.615338e-01              0.3103904          0.16302886
## 3 4.678176e-01              0.1481732          0.28447284
## 4 1.170305e-01              0.5390740          0.90245635
## 5 6.608195e-08              0.2990308          0.01111218
## 6 1.873349e-01              0.2695888          0.17164308
##   per_capita_income_scale        vote1   density median_age_scale
## 2              0.19905670 0.1664090716 0.3948089       0.17708985
## 3              0.68194447 0.3803870759 0.3500700       0.40139565
## 4              0.12511690 0.1853524109 0.2700675       0.17749426
## 5              0.08201675 0.0006988677 0.5143089       0.05630236
## 6              0.15678031 0.1591033656 0.2960125       0.16923465
```

```
## [1] "The p values for each predictors for death"
```

```
##   (Intercept) total_population_scale percent_white_scale
## 2   0.4612021              0.4414345          0.16115760
## 3   0.8716736              0.9667550          0.99297342
## 4   0.8267069              0.4132309          0.99068137
## 5   0.6305368              0.7605552          0.21278708
## 6   0.2344629              0.3113342          0.04999773
##   per_capita_income_scale     vote1    density median_age_scale
```

```
## 2               0.2294392 0.4895165 0.01955256        0.8940259
## 3               0.9777060 0.7921144 0.99841528        0.9959445
## 4               0.2295167 0.5442832 0.10388950        0.9924402
## 5               0.7119179 0.6116778 0.82503789        0.9748413
## 6               0.2929903 0.2535156 0.13660206        0.7901268
```

Moreover, we applied a chi square testing for the full model for both cases and deaths. The results suggest that our model can significantly predict the cluster membership for cases growth ($p < 0.001$) and deaths growth ($p < 0.001$) based on the demographical and political information of each state. For the results of chi square testing, see below:

```
## [1] "Chi square testing for the full model, case"
```

```
##
##  Pearson's Chi-squared test
##
## data:  us_data_2020_subset$cluster_case and predict(multi_mo_case)
## X-squared = 172.27, df = 25, p-value < 2.2e-16
```

```
## [1] "Chi square testing for the full model, death"
```

```
##
##  Pearson's Chi-squared test
##
## data:  us_data_2020_subset$cluster_death and predict(multi_mo_death)
## X-squared = 163.33, df = 25, p-value < 2.2e-16
```

Finally, we are interested in whether or not each individual predictor is a significant predictor for the cases or deaths growth for each state. Then we compared the Likelihood Ratio testing between the reduced model and the full model with one predictor dropped each time. Based on the results, we found that the percentage of white polulation ($p < 0.001$) and the political party affliation ($p < 0.05$) are the two significant preditors for predicting case. The total population ($p = 0.139$), median age ($p = 0.663$), and per capita income ($p = 0.252$) are not significant predictor for predicting the case growth pattern. We also found that, for death growth pattern, only the polulation density is the significant predictor. The total population ($p = 0.707$), median age ($p = 0.993$), per capita income ($p = 0.721$), and political party affliation ($p = 0.469$) are not significant predictor for predicting the death growth pattern. However, the percentage of white polulation ($p = 0.051$) is the weak predictor for predicting the death growth pattern. See the full results below:

```
## [1] "Case total_population_scale"
```

```
## Likelihood ratio test
##
## Model 1: cluster_case ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale
## Model 2: cluster_case ~ percent_white_scale + per_capita_income_scale +
##     vote + density + median_age_scale
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  35 -26.980
## 2  30 -31.144 -5 8.3273     0.1391
```

```
## [1] "Case percent_white_scale"
```

```
## Likelihood ratio test
##
## Model 1: cluster_case ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale
## Model 2: cluster_case ~ total_population_scale + per_capita_income_scale +
##     vote + density + median_age_scale
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  35 -26.980
## 2  30 -37.922 -5 21.884  0.0005509 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Case median_age_scale"
```

```
## Likelihood ratio test
##
## Model 1: cluster_case ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale
## Model 2: cluster_case ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  35 -26.980
## 2  30 -28.599 -5 3.2389     0.6632
```

```
## [1] "Case per_capita_income_scale"
```

```
## Likelihood ratio test
##
## Model 1: cluster_case ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale
## Model 2: cluster_case ~ total_population_scale + percent_white_scale +
##     vote + density + median_age_scale
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  35 -26.980
## 2  30 -30.278 -5 6.5953     0.2525
```

```
## [1] "Case vote"
```

```
## Likelihood ratio test
##
## Model 1: cluster_case ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale
## Model 2: cluster_case ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + density + median_age_scale
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  35 -26.980
## 2  30 -33.299 -5 12.638    0.02702 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Case density"
```

```
## Likelihood ratio test
##
## Model 1: cluster_case ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale
## Model 2: cluster_case ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + median_age_scale
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  35 -26.980
## 2  30 -30.211 -5 6.4627     0.2638
```

```
## [1] "Death total_population_scale"
```

```
## Likelihood ratio test
##
## Model 1: cluster_death ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale
## Model 2: cluster_death ~ percent_white_scale + per_capita_income_scale +
##     vote + density + median_age_scale
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  35 -28.368
## 2  30 -29.844 -5 2.9519     0.7074
```

```
## [1] "Death percent_white_scale"
```

```
## Likelihood ratio test
##
## Model 1: cluster_death ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale
## Model 2: cluster_death ~ total_population_scale + per_capita_income_scale +
##     vote + density + median_age_scale
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  35 -28.368
## 2  30 -33.888 -5 11.041    0.05057 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Death median_age_scale"
```

```
## Likelihood ratio test
##
## Model 1: cluster_death ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale
## Model 2: cluster_death ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1  35 -28.368
## 2  30 -28.598 -5  0.46     0.9935
```

```
## [1] "Death per_capita_income_scale"
```

```
## Likelihood ratio test
##
```

```
## Model 1: cluster_death ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale
## Model 2: cluster_death ~ total_population_scale + percent_white_scale +
##     vote + density + median_age_scale
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  35 -28.368
## 2  30 -29.800 -5 2.8651     0.7208
```

```
## [1] "Death vote"
```

```
## Likelihood ratio test
##
## Model 1: cluster_death ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale
## Model 2: cluster_death ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + density + median_age_scale
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  35 -28.368
## 2  30 -30.658 -5 4.5796     0.4693
```

```
## [1] "Death density"
```

```
## Likelihood ratio test
##
## Model 1: cluster_death ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + density + median_age_scale
## Model 2: cluster_death ~ total_population_scale + percent_white_scale +
##     per_capita_income_scale + vote + median_age_scale
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  35 -28.368
## 2  30 -38.726 -5 20.717   0.000916 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusions

Temporal Trends

Spatial Trends

Based on our analysis, we found out that: 1. The new cases growth pattern and the new deaths growth patterns among states are correlated with their geographical location. This indicates that geographically adjacent states will have similar cases and deaths growth pattern. This might suggest that the spreading of the virus is constraint by the geographical distance and might be diffused adjacently.

2. We found that the cases and deaths time series could be clustered into six clusters. Each cluster indicates an unique growth pattern for case or death. The distinction between different clusters could be 1) the timing of waves of spreading, and 2) the relative number of cases/deaths among the three waves (some states have low first wave cases, but have high second and third wave cases, but other states have high first wave cases, but have relatively low second and third wave cases).

3. We also found that the clustering of states are also geographically clustered. This could be due to 1) the adjancent spreading of the virus; 2) the natural environment of different states (e.g. temperature, weather, etc.); 3) the way people live (e.g., urban vs. rural).

4. We built a multinomial logistic regression model to predict the cluster membership for both case and death, with predictors as total population, percent white, median age, per capita income, political party affiliation, and the population density. We found that: 1) the demographical information of races can predict both death and cases, which might indicate the existence of racial inequality at times of a pandemic. 2) The political party affiation is the significant predictor for cases, which might indicate that different party affiliation influences people's attitude and coping strategy toward COVID-19. 3) The population density is the major predictor for death, which might indicate that the growth of death is mainly determined by the avaliability of hospitalization resources. When in a population dense area, the rapid spreading of the virus will run out of the hospital resources and eventually have causal effect on patients' death.