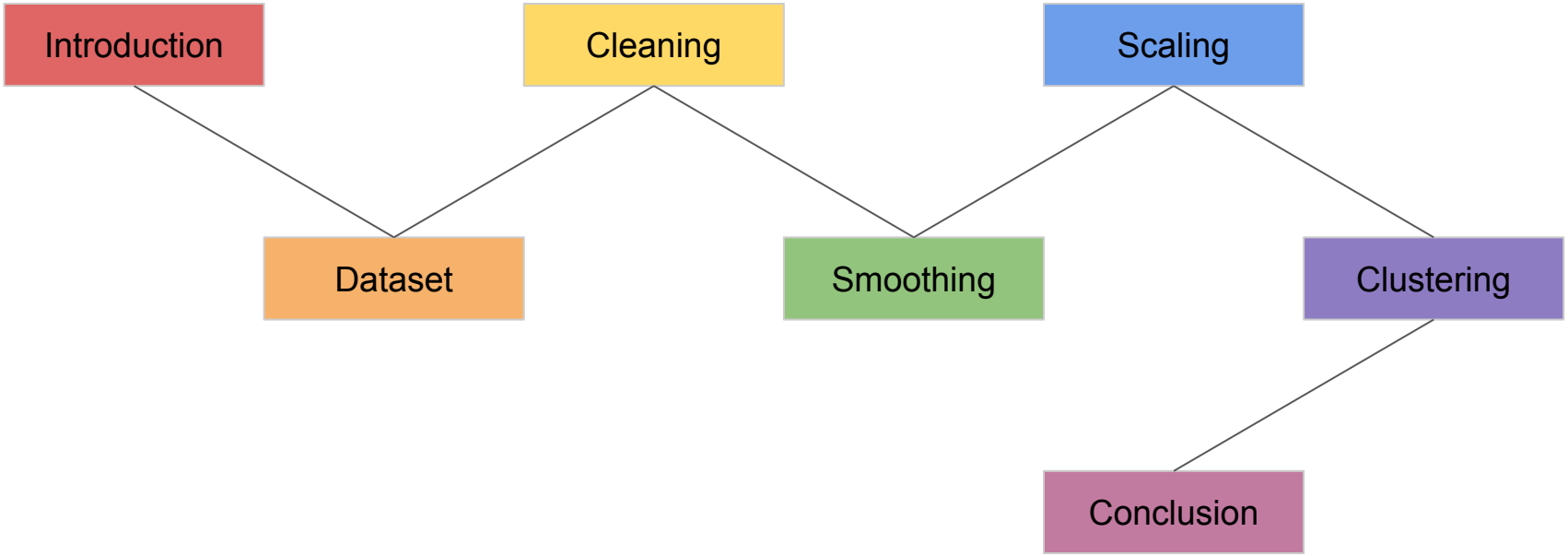
Multivariate Time Series Clustering of US States Using COVID-19 Data

Brandon Vittetoe

Roadmap


Introduction
Cleaning
Scaling
Dataset
Smoothing
Clustering
Conclusion

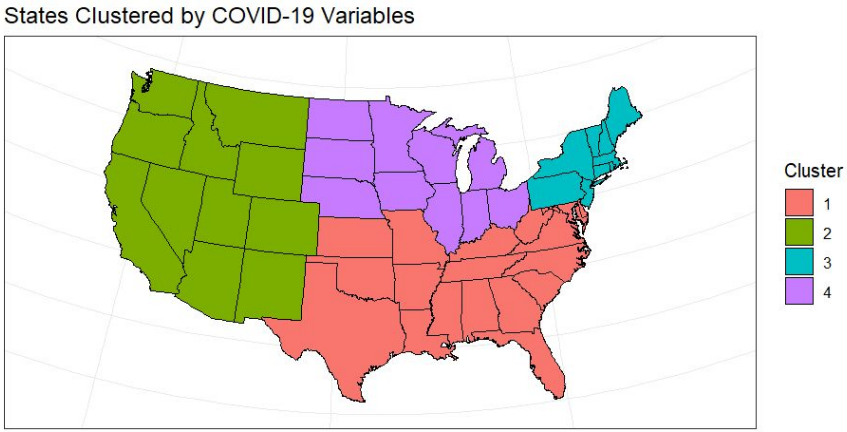
Introduction

Multivariate Time Series Clustering of US States Using COVID-19 Data

- Multivariate Data. Data with multiple variables.
- Time Series Data. Data with observations over time.
- Clustering. The process of grouping similar objects.

In this project, we will use hierarchical clustering to group US states based on trends in multiple COVID-related epidemiological variables per day.

Our goal is turn the messy data into a map like this:



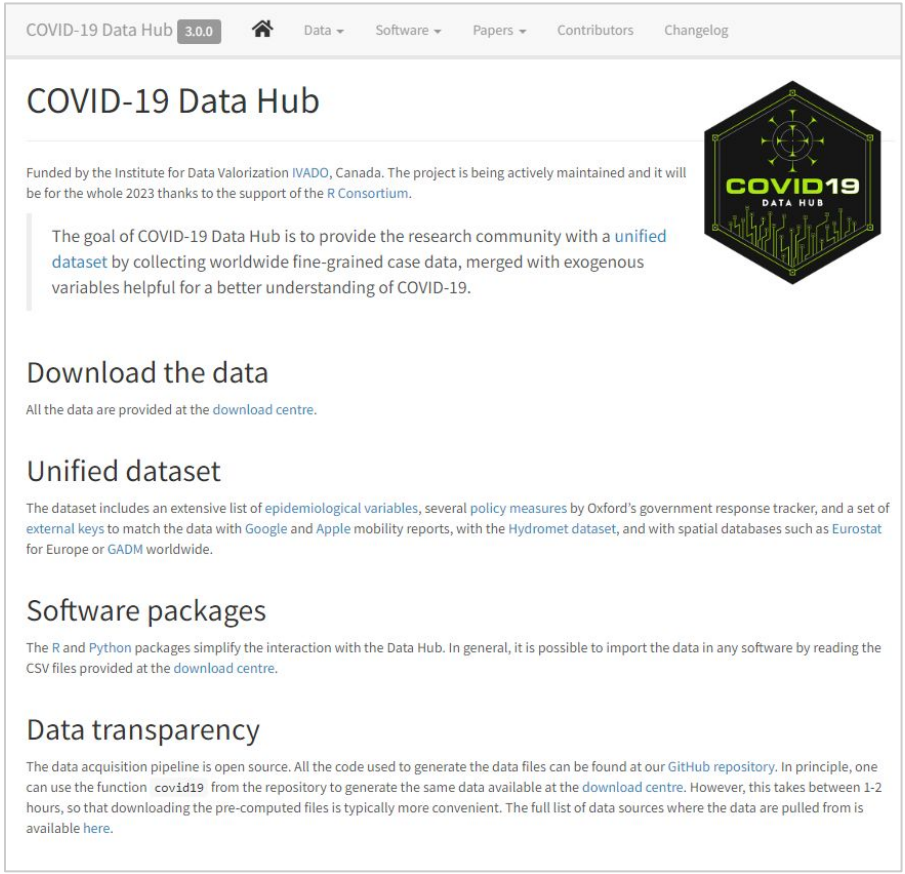
Dataset (1 of 4) Introduction

The `COVID19` package is an interface to the COVID-19 Data Hub.

It provides a daily summary of COVID-19 cases, deaths, recoveries, tests, vaccinations, and hospitalizations for over 230 countries, 760 regions, and 12,000 administrative divisions of lower level.

The package includes policy measures, mobility data, and geospatial identifiers.

The data is collected from various government sources around the world and combined into a unified database.



COVID-19 Data Hub 3.0.0 | Data | Software | Papers | Contributors | Changelog

COVID-19 Data Hub

Funded by the Institute for Data Valorization IVADO, Canada. The project is being actively maintained and it will be for the whole 2023 thanks to the support of the R Consortium.

> The goal of COVID-19 Data Hub is to provide the research community with a unified dataset by collecting worldwide fine-grained case data, merged with exogenous variables helpful for a better understanding of COVID-19.

Download the data

All the data are provided at the download centre.

Unified dataset

The dataset includes an extensive list of epidemiological variables, several policy measures by Oxford's government response tracker, and a set of external keys to match the data with Google and Apple mobility reports, with the Hydromet dataset, and with spatial databases such as Eurostat for Europe or GADM worldwide.

Software packages

The R and Python packages simplify the interaction with the Data Hub. In general, it is possible to import the data in any software by reading the CSV files provided at the download centre.

Data transparency

The data acquisition pipeline is open source. All the code used to generate the data files can be found at our GitHub repository. In principle, one can use the function `covid19` from the repository to generate the same data available at the download centre. However, this takes between 1-2 hours, so that downloading the pre-computed files is typically more convenient. The full list of data sources where the data are pulled from is available here.

Screenshot of https://covid19datahub.io/, 12/9/2023.

Dataset (2 of 4) Preview

```r
covid19(country = c("US"), level = 2)
```

Dataset (2 of 4) Preview

```r
covid19(country = c("US"), level = 2)
```

```
# A tibble: 77,261 × 47
   id     date       confirmed deaths recovered tests vaccines people_vaccinated
   <chr>  <date>         <dbl>  <dbl>     <dbl> <dbl>    <dbl>              <dbl>
 1 10b6… 2020-03-16        NA     NA        NA    NA       NA                 NA
 2 10b6… 2020-03-17        NA     NA        NA    NA       NA                 NA
 3 10b6… 2020-03-18        NA     NA        NA    NA       NA                 NA
 4 10b6… 2020-03-19        NA     NA        NA    NA       NA                 NA
 5 10b6… 2020-03-20        NA     NA        NA    NA       NA                 NA
 6 10b6… 2020-03-21        NA     NA        NA    NA       NA                 NA
 7 10b6… 2020-03-22        NA     NA        NA    NA       NA                 NA
 8 10b6… 2020-03-23        NA     NA        NA    NA       NA                 NA
 9 10b6… 2020-03-24        NA     NA        NA    NA       NA                 NA
10 10b6… 2020-03-25        NA     NA        NA    NA       NA                 NA
# ℹ 77,251 more rows
# ℹ 39 more variables: people_fully_vaccinated <dbl>, hosp <dbl>, icu <dbl>, …
```

Dataset (3 of 4) Epidemiological Variables

```
# A tibble: 10 × 2
   variable                 description
   <chr>                    <chr>
 1 confirmed                Cumulative number of confirmed cases.
 2 deaths                   Cumulative number of deaths.
 3 recovered                Cumulative number of patients released from hospitals or reported recovered.
 4 tests                    Cumulative number of tests.
 5 vaccines                 Cumulative number of total doses administered.
 6 people_vaccinated        Cumulative number of people who received at least one vaccine dose.
 7 people_fully_vaccinated  Cumulative number of people who received all doses prescribed by the vaccination
                            protocol.
 8 hosp                     Number of hospitalized patients on date.
 9 icu                      Number of hospitalized patients in intensive therapy on date.
10 vent                     Number of patients requiring invasive ventilation on date.
```

Dataset (4 of 4) Missing Values

```
data %>%
  filter(administrative_area_level_2 %in% states,
    date >= "2021-01-01", date <= "2022-12-31") %>%
  select(epidemiological_variables) %>%
  summarize(sum(missing_values)) %>%
  gather(variable, missing_values)
```

Dataset (4 of 4) Missing Values

```r
data %>%
  filter(administrative_area_level_2 %in% states,
    date >= "2021-01-01", date <= "2022-12-31") %>%
  select(epidemiological_variables) %>%
  summarize(sum(missing_values)) %>%
  gather(variable, missing_values)
```

```
# A tibble: 10 × 2
   variable                missing_values
   <chr>                            <int>
 1 confirmed                            0
 2 deaths                               0
 3 recovered                        34586
 4 tests                                2
 5 vaccines                             0
 6 people_vaccinated                    0
 7 people_fully_vaccinated              0
 8 hosp                                 0
 9 icu                                  0
10 vent                             34586
```

Cleaning (1 of 8) Load Data

```
data <- covid19(country = c("US"), level = 2)
```

Cleaning (1 of 8) Load Data

```r
data <- covid19(country = c("US"), level = 2)
```

```
# A tibble: 77,261 × 47
   id     date       confirmed deaths recovered    tests vaccines people_vaccinated
   <chr>  <date>         <dbl>  <dbl>     <dbl>    <dbl>    <dbl>             <dbl>
 1 10b6… 2020-03-16        NA     NA        NA       NA       NA                NA
 2 10b6… 2020-03-17        NA     NA        NA       NA       NA                NA
 3 10b6… 2020-03-18        NA     NA        NA       NA       NA                NA
 4 10b6… 2020-03-19        NA     NA        NA       NA       NA                NA
 5 10b6… 2020-03-20        NA     NA        NA       NA       NA                NA
 6 10b6… 2020-03-21        NA     NA        NA       NA       NA                NA
 7 10b6… 2020-03-22        NA     NA        NA       NA       NA                NA
 8 10b6… 2020-03-23        NA     NA        NA       NA       NA                NA
 9 10b6… 2020-03-24        NA     NA        NA       NA       NA                NA
10 10b6… 2020-03-25        NA     NA        NA       NA       NA                NA
# ℹ 77,251 more rows
# ℹ 39 more variables: people_fully_vaccinated <dbl>, hosp <dbl>, icu <dbl>, …
```

Cleaning (2 of 8) Select Columns

```
data <- select(data, columns)
```

```
# A tibble: 77,261 × 47
   id     date       confirmed deaths recovered    tests vaccines people_vaccinated
   <chr>  <date>         <dbl>  <dbl>     <dbl>    <dbl>    <dbl>             <dbl>
 1 10b6… 2020-03-16        NA     NA        NA       NA       NA                NA
 2 10b6… 2020-03-17        NA     NA        NA       NA       NA                NA
 3 10b6… 2020-03-18        NA     NA        NA       NA       NA                NA
 4 10b6… 2020-03-19        NA     NA        NA       NA       NA                NA
 5 10b6… 2020-03-20        NA     NA        NA       NA       NA                NA
 6 10b6… 2020-03-21        NA     NA        NA       NA       NA                NA
 7 10b6… 2020-03-22        NA     NA        NA       NA       NA                NA
 8 10b6… 2020-03-23        NA     NA        NA       NA       NA                NA
 9 10b6… 2020-03-24        NA     NA        NA       NA       NA                NA
10 10b6… 2020-03-25        NA     NA        NA       NA       NA                NA
# ℹ 77,251 more rows
# ℹ 39 more variables: people_fully_vaccinated <dbl>, hosp <dbl>, icu <dbl>, …
```

Cleaning (2 of 8) Select Columns

```
data <- select(data, columns)
```

```
# A tibble: 77,261 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2020-03-16        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 2 2020-03-17        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 3 2020-03-18        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 4 2020-03-19        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 5 2020-03-20        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 6 2020-03-21        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 7 2020-03-22        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 8 2020-03-23        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 9 2020-03-24        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
10 2020-03-25        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
# ℹ 77,251 more rows
```

Cleaning (3 of 8) Filter Rows

```
data <- filter(data, administrative_area_level_2 %in% states, date >= "2021-01-01", date <= "2022-12-31")
```

```
# A tibble: 77,261 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2020-03-16        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 2 2020-03-17        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 3 2020-03-18        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 4 2020-03-19        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 5 2020-03-20        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 6 2020-03-21        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 7 2020-03-22        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 8 2020-03-23        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
 9 2020-03-24        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
10 2020-03-25        NA      NA       NA       NA      NA     NA  Northern Mariana Islands
# ℹ 77,251 more rows
```

Cleaning (3 of 8) Filter Rows

```r
data <- filter(data, administrative_area_level_2 %in% states, date >= "2021-01-01", date <= "2022-12-31")
```

```
# A tibble: 36,500 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-01-01    415361    5382  6837436   108270     984    188  Minnesota
 2 2021-01-02    417891    5436  6851047   110896     941    178  Minnesota
 3 2021-01-03    420603    5489  6868629   113137     937    168  Minnesota
 4 2021-01-04    423747    5502  6893050   124687     964    168  Minnesota
 5 2021-01-05    425320    5520  6954659   142867     940    153  Minnesota
 6 2021-01-06    427655    5596  7046211   161282     905    147  Minnesota
 7 2021-01-07    429638    5640  7108052   178303     870    148  Minnesota
 8 2021-01-08    432012    5688  7175081   195879     822    146  Minnesota
 9 2021-01-09    434481    5731  7223935   199806     766    146  Minnesota
10 2021-01-10    436640    5775  7241208   201839     779    153  Minnesota
# ℹ 36,490 more rows
```

Cleaning (4 of 7) Sort Data

```
data <- arrange(data, date, administrative_area_level_2)
```

```
# A tibble: 36,500 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-01-01    415361    5382  6837436   108270     984    188  Minnesota
 2 2021-01-02    417891    5436  6851047   110896     941    178  Minnesota
 3 2021-01-03    420603    5489  6868629   113137     937    168  Minnesota
 4 2021-01-04    423747    5502  6893050   124687     964    168  Minnesota
 5 2021-01-05    425320    5520  6954659   142867     940    153  Minnesota
 6 2021-01-06    427655    5596  7046211   161282     905    147  Minnesota
 7 2021-01-07    429638    5640  7108052   178303     870    148  Minnesota
 8 2021-01-08    432012    5688  7175081   195879     822    146  Minnesota
 9 2021-01-09    434481    5731  7223935   199806     766    146  Minnesota
10 2021-01-10    436640    5775  7241208   201839     779    153  Minnesota
# ℹ 36,490 more rows
```

Cleaning (4 of 7) Sort Data

```
data <- arrange(data, date, administrative_area_level_2)
```

```
# A tibble: 36,500 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-01-01    365747    4872  3292989    52016    3013    790  Alabama
 2 2021-01-01     46740     198  1412020    23954      69     11  Alaska
 3 2021-01-01    530267    9015  4747564   132257    4661   1017  Arizona
 4 2021-01-01    229442    3711  1966898    59361    1211    353  Arkansas
 5 2021-01-01   2345811   26236 30270167   585717   21121   4556  California
 6 2021-01-01    338357    4936  4262531   136689    1101    313  Colorado
 7 2021-01-01    185708    5995  4739531    83631    1232    243  Connecticut
 8 2021-01-01     58064     930  1180941    17296     472     57  Delaware
 9 2021-01-01   1323307   21672 17399998   324954    7099   1389  Florida
10 2021-01-01    654950   10610  5372138   122211    5067   1200  Georgia
# ℹ 36,490 more rows
```

Cleaning (5 of 7) Handle Missing Values

```
data[!complete.cases(data), ]
```

```
# A tibble: 36,500 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-01-01    365747    4872  3292989    52016    3013    790  Alabama
 2 2021-01-01     46740     198  1412020    23954      69     11  Alaska
 3 2021-01-01    530267    9015  4747564   132257    4661   1017  Arizona
 4 2021-01-01    229442    3711  1966898    59361    1211    353  Arkansas
 5 2021-01-01   2345811   26236 30270167   585717   21121   4556  California
 6 2021-01-01    338357    4936  4262531   136689    1101    313  Colorado
 7 2021-01-01    185708    5995  4739531    83631    1232    243  Connecticut
 8 2021-01-01     58064     930  1180941    17296     472     57  Delaware
 9 2021-01-01   1323307   21672 17399998   324954    7099   1389  Florida
10 2021-01-01    654950   10610  5372138   122211    5067   1200  Georgia
# ℹ 36,490 more rows
```

```
data <- group_by(data, administrative_area_level_2)
data <- mutate(data, tests = na.approx(tests))
data <- ungroup(data)
```

Cleaning (5 of 7) Handle Missing Values

```
data[!complete.cases(data), ]
```

```
# A tibble: 7 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-04-04   1024011   18643 11168023  6387591    1344    290  Ohio
 2 2021-04-05   1026929   18643       NA  6474344    1407    296  Ohio
 3 2021-04-06   1028800   18741       NA  6612341    1415    297  Ohio
 4 2021-04-07   1030864   18741 11215135  6775375    1527    331  Ohio
 5 2021-04-08   1033606   18741 11262156  6943963    1583    319  Ohio
 6 2021-04-09   1035552   18827 11302718  7085171    1538    329  Ohio
 7 2021-04-10   1037600   18827 11337665  7181082    1462    329  Ohio
```

```
data <- group_by(data, administrative_area_level_2)
data <- mutate(data, tests = na.approx(tests))
data <- ungroup(data)
```

Cleaning (5 of 7) Handle Missing Values

```
data[!complete.cases(data), ]
```

```
# A tibble: 7 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-04-04   1024011   18643 11168023  6387591    1344    290  Ohio
 2 2021-04-05   1026929   18643 11183727  6474344    1407    296  Ohio
 3 2021-04-06   1028800   18741 11199431  6612341    1415    297  Ohio
 4 2021-04-07   1030864   18741 11215135  6775375    1527    331  Ohio
 5 2021-04-08   1033606   18741 11262156  6943963    1583    319  Ohio
 6 2021-04-09   1035552   18827 11302718  7085171    1538    329  Ohio
 7 2021-04-10   1037600   18827 11337665  7181082    1462    329  Ohio
```

```
data <- group_by(data, administrative_area_level_2)
data <- mutate(data, tests = na.approx(tests))
data <- ungroup(data)
```

Cleaning (6 of 7) Transform Cumulative Variables

```
data
```

```
# A tibble: 7 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-04-04   1024011   18643 11168023  6387591    1344    290  Ohio
 2 2021-04-05   1026929   18643 11183727  6474344    1407    296  Ohio
 3 2021-04-06   1028800   18741 11199431  6612341    1415    297  Ohio
 4 2021-04-07   1030864   18741 11215135  6775375    1527    331  Ohio
 5 2021-04-08   1033606   18741 11262156  6943963    1583    319  Ohio
 6 2021-04-09   1035552   18827 11302718  7085171    1538    329  Ohio
 7 2021-04-10   1037600   18827 11337665  7181082    1462    329  Ohio
```

```
data <- group_by(data, administrative_area_level_2)
data <- mutate(data, cumulative_features = cumulative_features - lag(cumulative_features))
data <- ungroup(data)
```

Cleaning (6 of 7) Transform Cumulative Variables

```r
data
```

```
# A tibble: 36,500 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-01-01    365747    4872  3292989    52016    3013    790  Alabama
 2 2021-01-01     46740     198  1412020    23954      69     11  Alaska
 3 2021-01-01    530267    9015  4747564   132257    4661   1017  Arizona
 4 2021-01-01    229442    3711  1966898    59361    1211    353  Arkansas
 5 2021-01-01   2345811   26236 30270167   585717   21121   4556  California
 6 2021-01-01    338357    4936  4262531   136689    1101    313  Colorado
 7 2021-01-01    185708    5995  4739531    83631    1232    243  Connecticut
 8 2021-01-01     58064     930  1180941    17296     472     57  Delaware
 9 2021-01-01   1323307   21672 17399998   324954    7099   1389  Florida
10 2021-01-01    654950   10610  5372138   122211    5067   1200  Georgia
# ℹ 36,490 more rows
```

```r
data <- group_by(data, administrative_area_level_2)
data <- mutate(data, cumulative_features = cumulative_features - lag(cumulative_features))
data <- ungroup(data)
```

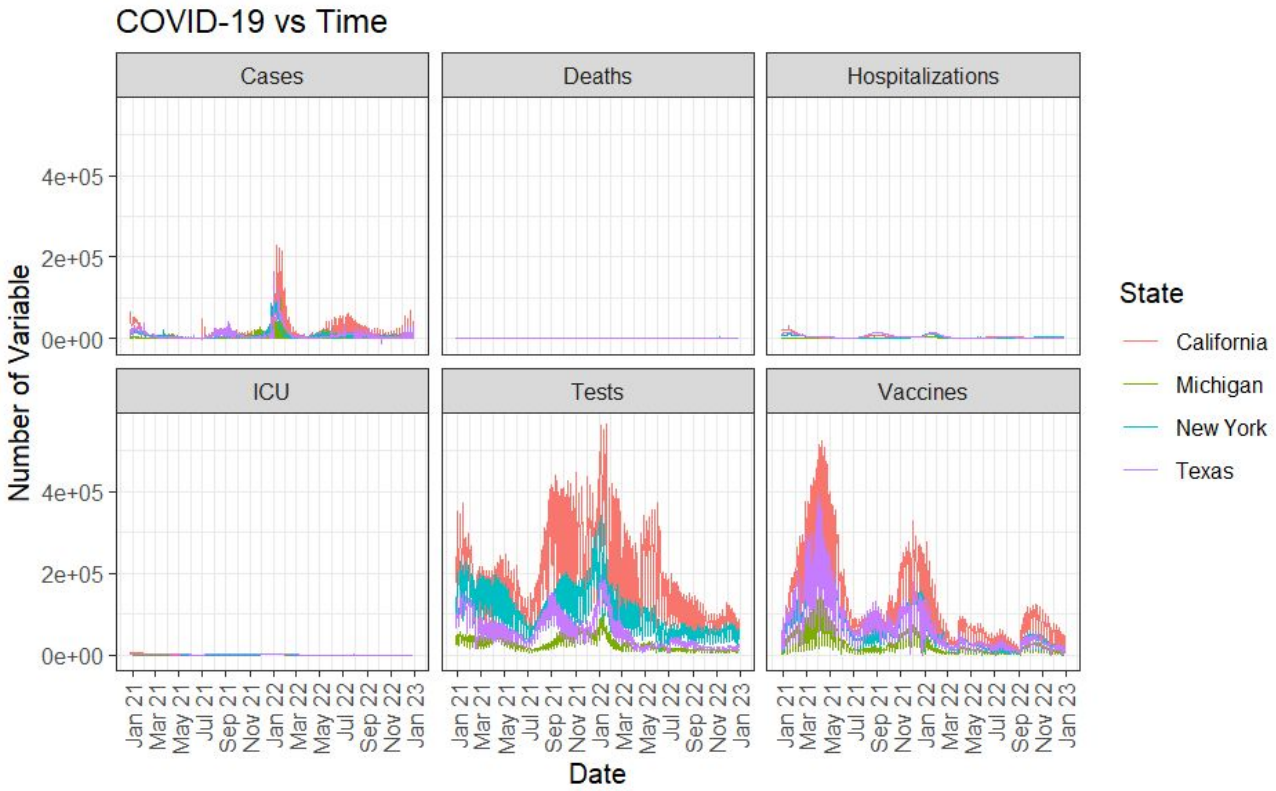
Cleaning (6 of 7) Transform Cumulative Variables

```
data
```

```
# A tibble: 36,500 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-01-01      4521      45    13414     1487    3013    790  Alabama
 2 2021-01-01         0       0     5002      206      69     11  Alaska
 3 2021-01-01      6438     136    33266    10265    4661   1017  Arizona
 4 2021-01-01      4304      35     8359      961    1211    353  Arkansas
 5 2021-01-01     37951     271   240083     5807   21121   4556  California
 6 2021-01-01      2778      57    26709     1063    1101    313  Colorado
 7 2021-01-01         0       0    15516      750    1232    243  Connecticut
 8 2021-01-01       608       4     8159      858     472     57  Delaware
 9 2021-01-01         0       0    60467    18069    7099   1389  Florida
10 2021-01-01     10885      22     5889     6109    5067   1200  Georgia
# ℹ 36,490 more rows
```

```
data <- group_by(data, administrative_area_level_2)
data <- mutate(data, cumulative_features = cumulative_features - lag(cumulative_features))
data <- ungroup(data)
```

Cleaning (7 of 7) Summary



Smoothing (1 of 3) 7 Day Average

```
clean_data
```

```
# A tibble: 36,500 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-01-01      4521      45    13414     1487    3013    790  Alabama
 2 2021-01-01         0       0     5002      206      69     11  Alaska
 3 2021-01-01      6438     136    33266    10265    4661   1017  Arizona
 4 2021-01-01      4304      35     8359      961    1211    353  Arkansas
 5 2021-01-01     37951     271   240083     5807   21121   4556  California
 6 2021-01-01      2778      57    26709     1063    1101    313  Colorado
 7 2021-01-01         0       0    15516      750    1232    243  Connecticut
 8 2021-01-01       608       4     8159      858     472     57  Delaware
 9 2021-01-01         0       0    60467    18069    7099   1389  Florida
10 2021-01-01     10885      22     5889     6109    5067   1200  Georgia
# ℹ 36,490 more rows
```

```
data <- group_by(data, administrative_area_level_2)
data <- mutate(data, features = rollmean(features, 7, align = "right"))
data <- ungroup(data)
```

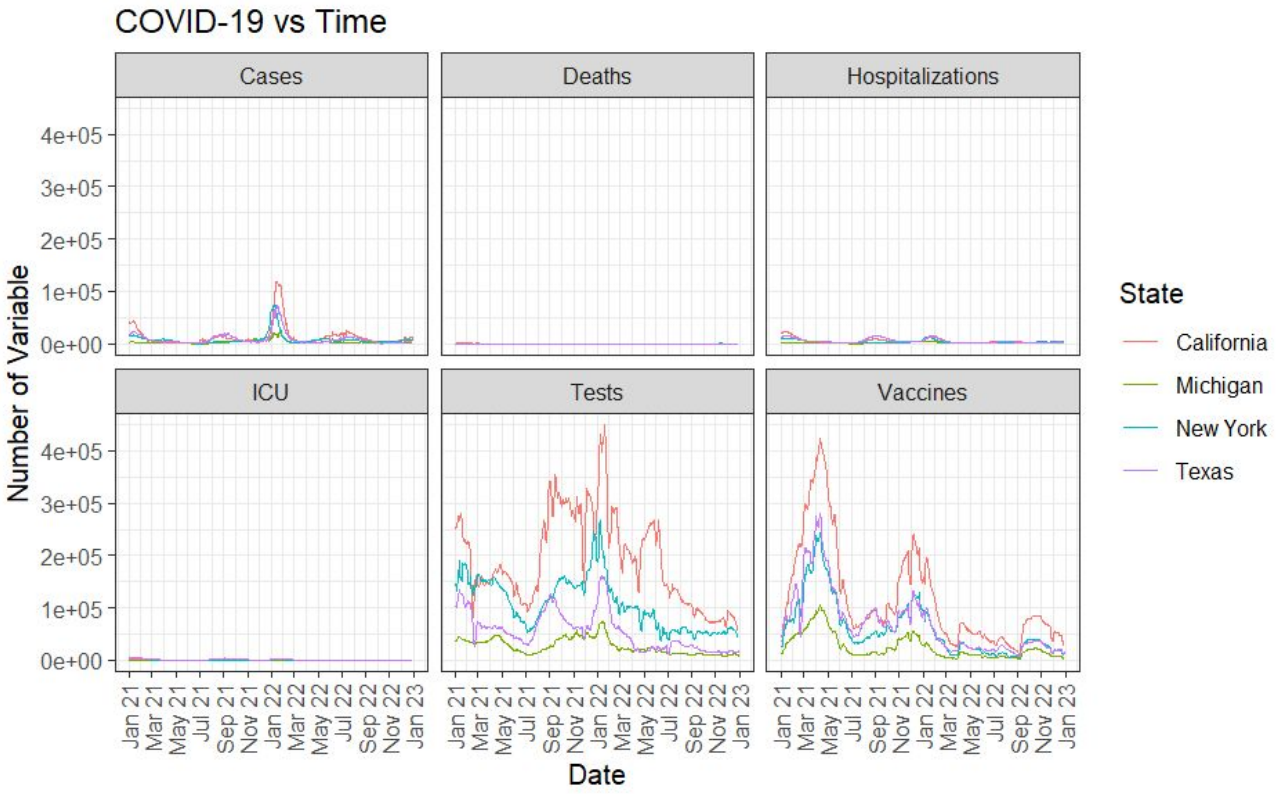
Smoothing (1 of 3) 7 Day Average

```r
clean_data
```

```
# A tibble: 36,500 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-01-01     3332.   27.4    15349.    4081.  2956.   760.   Alabama
 2 2021-01-01      266.    1       6636.    1256     73.3   13.7  Alaska
 3 2021-01-01     6182    86.6    38972.   10767.  4568.  1014.   Arizona
 4 2021-01-01     2311.   39       8458.    3305.  1172.   342.   Arkansas
 5 2021-01-01    40150.  325.    254579.   35954. 20989.  4418.   California
 6 2021-01-01     2147.   43.6    26550.    7880   1188.   336.   Colorado
 7 2021-01-01     1852.   29.1     6378.    6738.  1325.   243    Connecticut
 8 2021-01-01      630.    5.57    7546.    1249.   490.    61.1  Delaware
 9 2021-01-01    10824.   96.9    94638.   25602.  6743.  1352.   Florida
10 2021-01-01     7386.   43.9    31132.    9082.  4871.  1139    Georgia
# ℹ 36,490 more rows
```

```r
data <- group_by(data, administrative_area_level_2)
data <- mutate(data, features = rollmean(features, 7, align = "right"))
data <- ungroup(data)
```

Smoothing (1 of 3) 7 Day Average



Smoothing (2 of 3) LOESS

```r
clean_data
```

```
# A tibble: 36,500 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-01-01      4521      45    13414     1487    3013    790  Alabama
 2 2021-01-01         0       0     5002      206      69     11  Alaska
 3 2021-01-01      6438     136    33266    10265    4661   1017  Arizona
 4 2021-01-01      4304      35     8359      961    1211    353  Arkansas
 5 2021-01-01     37951     271   240083     5807   21121   4556  California
 6 2021-01-01      2778      57    26709     1063    1101    313  Colorado
 7 2021-01-01         0       0    15516      750    1232    243  Connecticut
 8 2021-01-01       608       4     8159      858     472     57  Delaware
 9 2021-01-01         0       0    60467    18069    7099   1389  Florida
10 2021-01-01     10885      22     5889     6109    5067   1200  Georgia
# ℹ 36,490 more rows
```

```r
data <- group_by(data, administrative_area_level_2)
data <- mutate(data, features = loess(features ~ date))
data <- ungroup(data)
```

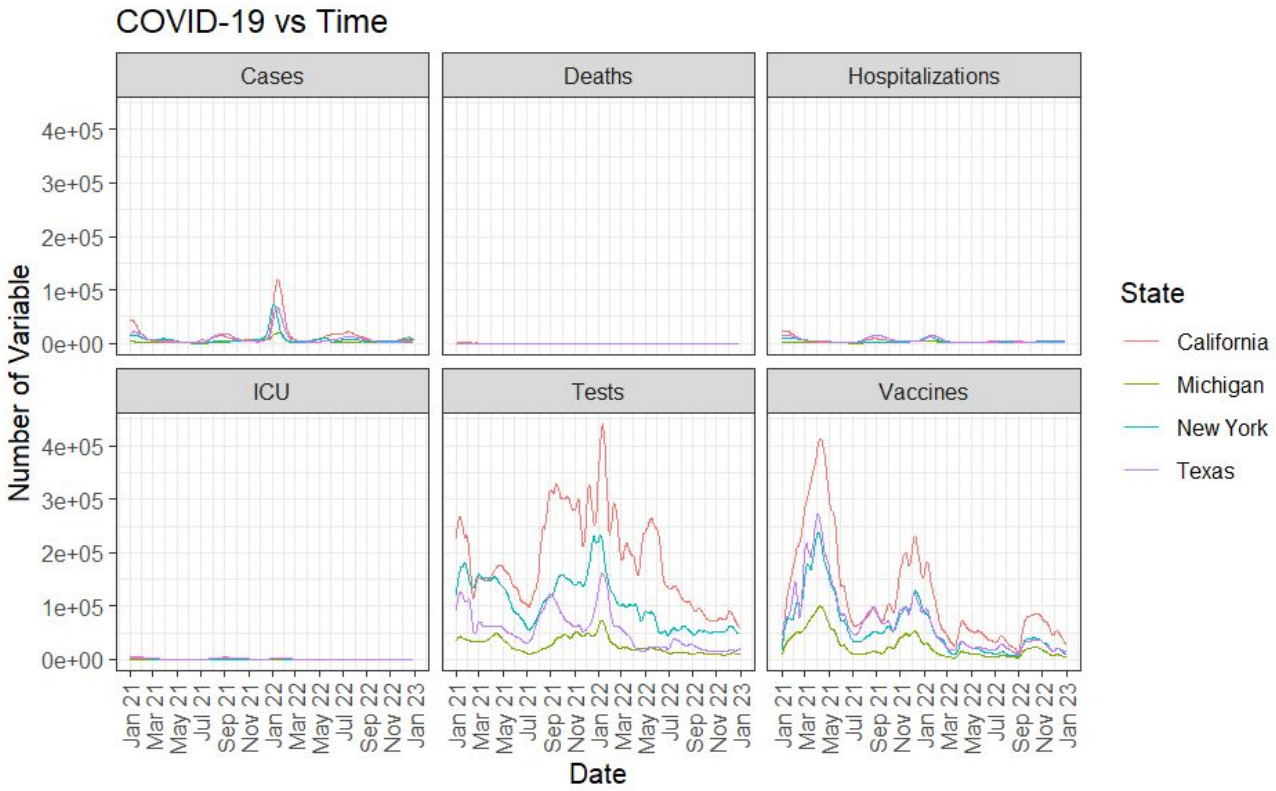
Smoothing (2 of 3) LOESS

```r
clean_data
```

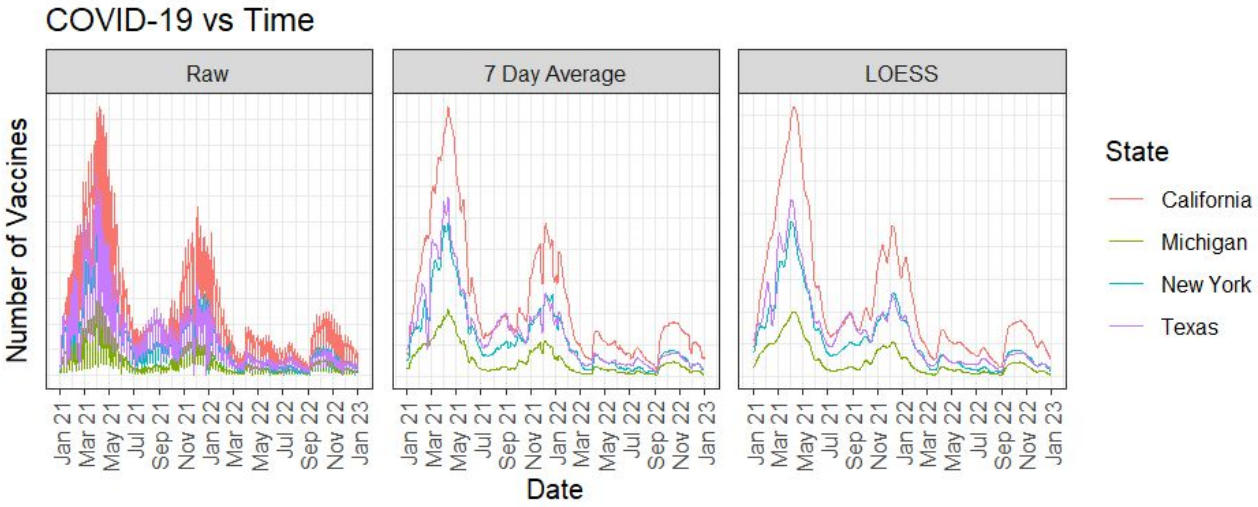
```
# A tibble: 36,500 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-01-01     4157.   17.1    11275.    1932.  3137.   793.   Alabama
 2 2021-01-01      341.    2.25    6610.     610.    85.5   14.1  Alaska
 3 2021-01-01     9699.   89.6    33379.    6628.  4688.  1027.   Arizona
 4 2021-01-01     3061.   31.6     8079.    3716.  1230.   350.   Arkansas
 5 2021-01-01    41347.  289.    225499.   15517. 21173.  4539.   California
 6 2021-01-01     2769.   43.9    28912.    7253.  1056.   286.   Colorado
 7 2021-01-01     2051.   39.4    28466.    4280.  1275.   243.   Connecticut
 8 2021-01-01      729.    3.32    8941.     750.   467.    54.8  Delaware
 9 2021-01-01    14458.   90.4    83793.   25169.  7442.  1392.   Florida
10 2021-01-01     8009.    5.00    7514.    3679.  5138.  1199.   Georgia
# ℹ 36,490 more rows
```

```r
data <- group_by(data, administrative_area_level_2)
data <- mutate(data, features = loess(features ~ date))
data <- ungroup(data)
```

Smoothing (2 of 3) LOESS



Smoothing (3 of 3) Summary



Scaling (1 of 2)

```r
data
```

```
# A tibble: 36,500 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-01-01     4157.   17.1    11275.    1932.  3137.   793.   Alabama
 2 2021-01-01      341.    2.25    6610.     610.    85.5   14.1  Alaska
 3 2021-01-01     9699.   89.6    33379.    6628.  4688.  1027.   Arizona
 4 2021-01-01     3061.   31.6     8079.    3716.  1230.   350.   Arkansas
 5 2021-01-01    41347.  289.    225499.   15517. 21173.  4539.   California
 6 2021-01-01     2769.   43.9    28912.    7253.  1056.   286.   Colorado
 7 2021-01-01     2051.   39.4    28466.    4280.  1275.   243.   Connecticut
 8 2021-01-01      729.    3.32    8941.     750.   467.    54.8  Delaware
 9 2021-01-01    14458.   90.4    83793.   25169.  7442.  1392.   Florida
10 2021-01-01     8009.    5.00    7514.    3679.  5138.  1199.   Georgia
# ℹ 36,490 more rows
```

```r
data <- group_by(data, administrative_area_level_2)
data <- mutate(data, features = scale(features))
data <- ungroup(data)
```

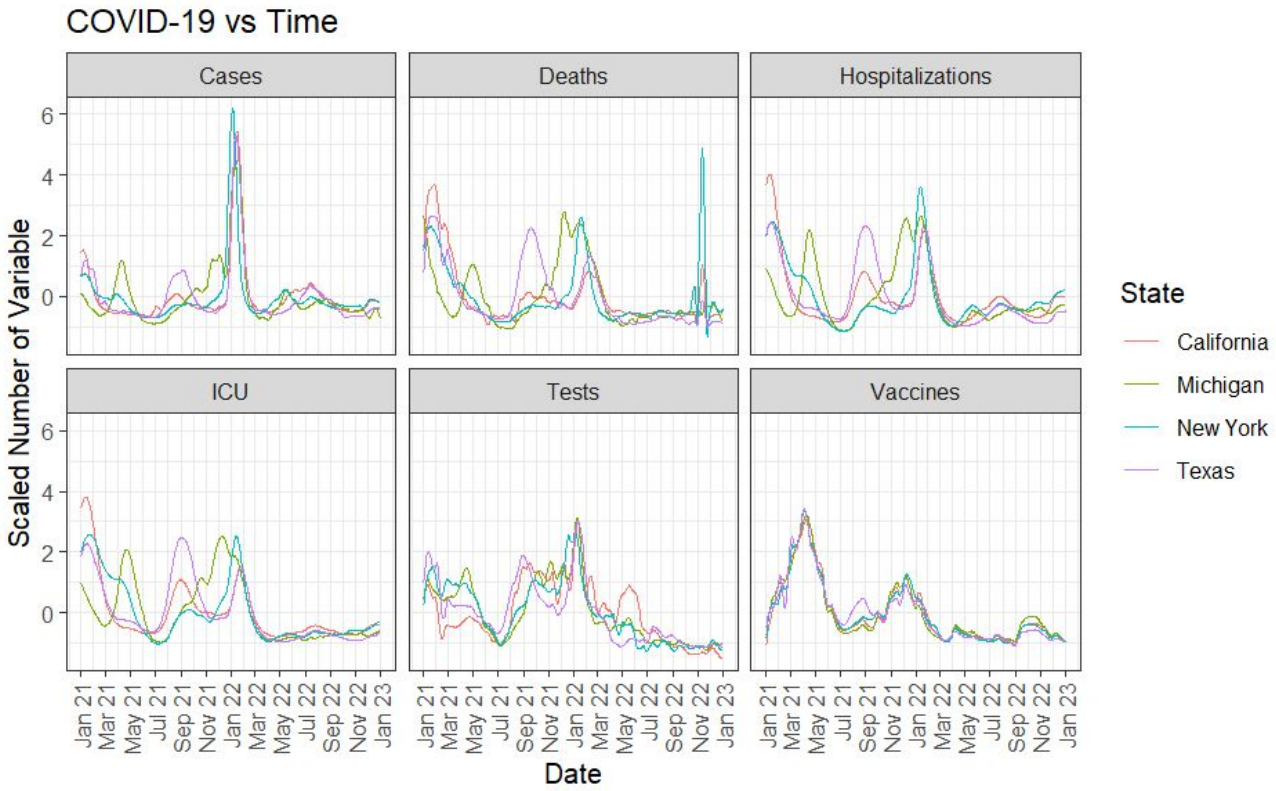
Scaling (1 of 2)

```r
data
```

```
# A tibble: 36,500 × 8
   date       confirmed  deaths    tests vaccines    hosp    icu  administrative_area_level_2
   <date>         <dbl>   <dbl>    <dbl>    <dbl>   <dbl>  <dbl>  <chr>
 1 2021-01-01    1.15   -0.193    0.971   -0.900   2.55   2.36   Alabama
 2 2021-01-01   -0.0195  0.296    0.733   -0.698   0.296  0.0662 Alaska
 3 2021-01-01    2.13    1.72     1.64    -0.788   3.08   2.96   Arizona
 4 2021-01-01    1.41    1.75     1.15    -0.506   1.76   1.57   Arkansas
 5 2021-01-01    1.46    1.59     0.438   -1.06    3.65   3.47   California
 6 2021-01-01    0.408   2.90     0.858   -0.683   1.11   0.837  Colorado
 7 2021-01-01    0.734   3.62     1.26    -0.742   2.24   2.61   Connecticut
 8 2021-01-01    0.789   0.0697   2.01    -0.864   2.04   2.21   Delaware
 9 2021-01-01    0.600   0.0663   0.492   -0.639   0.915  0.798  Florida
10 2021-01-01    1.31   -1.02    -1.20    -0.950   1.85   1.81   Georgia
# ℹ 36,490 more rows
```

```r
data <- group_by(data, administrative_area_level_2)
data <- mutate(data, features = scale(features))
data <- ungroup(data)
```

Scaling (2 of 2) Summary



Clustering (1 of 5) Calculating Distances

```r
# Initialize distances
distances <- 0

# Calculate distances
for(i in features) {
  temp <- data
  temp <- select(temp, date,
                 administrative_area_level_2, i)
  temp <- spread(temp, administrative_area_level_2, i)
  temp <- select(temp, -date)
  temp <- t(temp)
  temp <- dist(temp)
  distances <- distances + temp
}

# Hierarchical clustering
hc <- hclust(distances)
```

Clustering (1 of 5) Calculating Distances

```r
# Initialize distances
distances <- 0

# Calculate distances
for(i in features) {
  temp <- data
  temp <- select(temp, date,
                 administrative_area_level_2, i)
  temp <- spread(temp, administrative_area_level_2, i)
  temp <- select(temp, -date)
  temp <- t(temp)
  temp <- dist(temp)
  distances <- distances + temp
}

# Hierarchical clustering
hc <- hclust(distances)
```

```
# A tibble: 36,500 × 3
# Feature: vaccines
   date       administrative_area_level_2 vaccines
   <date>     <chr>                         <dbl>
 1 2021-01-01 Alabama                      -0.900
 2 2021-01-01 Alaska                       -0.698
 3 2021-01-01 Arizona                      -0.788
 4 2021-01-01 Arkansas                     -0.506
 5 2021-01-01 California                   -1.06
 6 2021-01-01 Colorado                     -0.683
 7 2021-01-01 Connecticut                  -0.742
 8 2021-01-01 Delaware                     -0.864
 9 2021-01-01 Florida                      -0.639
10 2021-01-01 Georgia                      -0.950
# ℹ 36,490 more rows
```

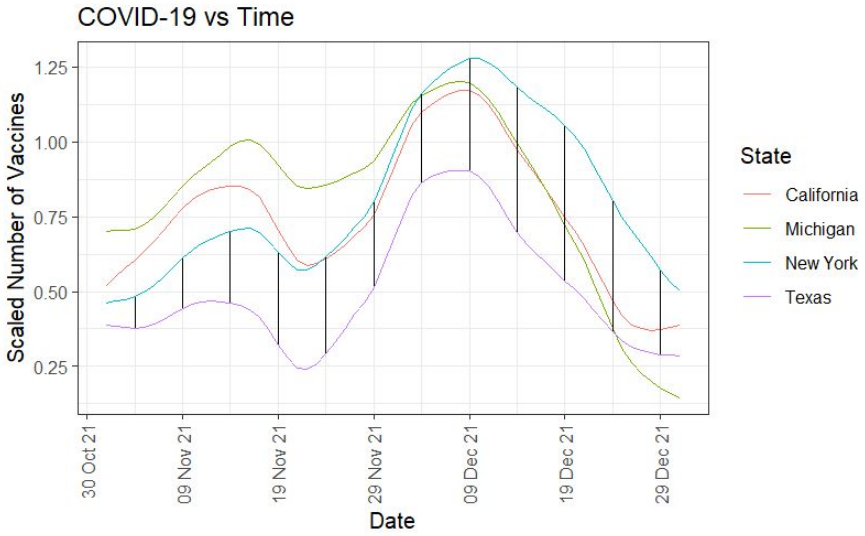
Clustering (1 of 5) Calculating Distances

```r
# Initialize distances
distances <- 0

# Calculate distances
for(i in features) {
  temp <- data
  temp <- select(temp, date,
                 administrative_area_level_2, i)
  temp <- spread(temp, administrative_area_level_2, i)
  temp <- select(temp, -date)
  temp <- t(temp)
  temp <- dist(temp)
  distances <- distances + temp
}

# Hierarchical clustering
hc <- hclust(distances)
```

```
# A tibble: 50 × 730
# Feature: vaccines
                T1     T2      T3      T4      T5
             <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
Alabama     -0.900 -0.789 -0.682  -0.578  -0.477
Alaska      -0.698 -0.313  0.0442  0.371   0.666
Arizona     -0.788 -0.711 -0.635  -0.557  -0.480
Arkansas    -0.506 -0.357 -0.217  -0.0843  0.0388
California  -1.06  -0.964 -0.867  -0.774  -0.685
Colorado    -0.683 -0.578 -0.480  -0.389  -0.305
Connecticut -0.742 -0.649 -0.560  -0.476  -0.396
Delaware    -0.864 -0.771 -0.678  -0.585  -0.492
Florida     -0.639 -0.482 -0.335  -0.198  -0.0731
Georgia     -0.950 -0.798 -0.654  -0.520  -0.394
# ℹ 40 more rows
# ℹ 725 more variables: T6 <dbl>, T7 <dbl>, …
```

Clustering (1 of 5) Calculating Distances

```r
# Initialize distances
distances <- 0

# Calculate distances
for(i in features) {
  temp <- data
  temp <- select(temp, date,
                 administrative_area_level_2, i)
  temp <- spread(temp, administrative_area_level_2, i)
  temp <- select(temp, -date)
  temp <- t(temp)
  temp <- dist(temp)
  distances <- distances + temp
}

# Hierarchical clustering
hc <- hclust(distances)
```



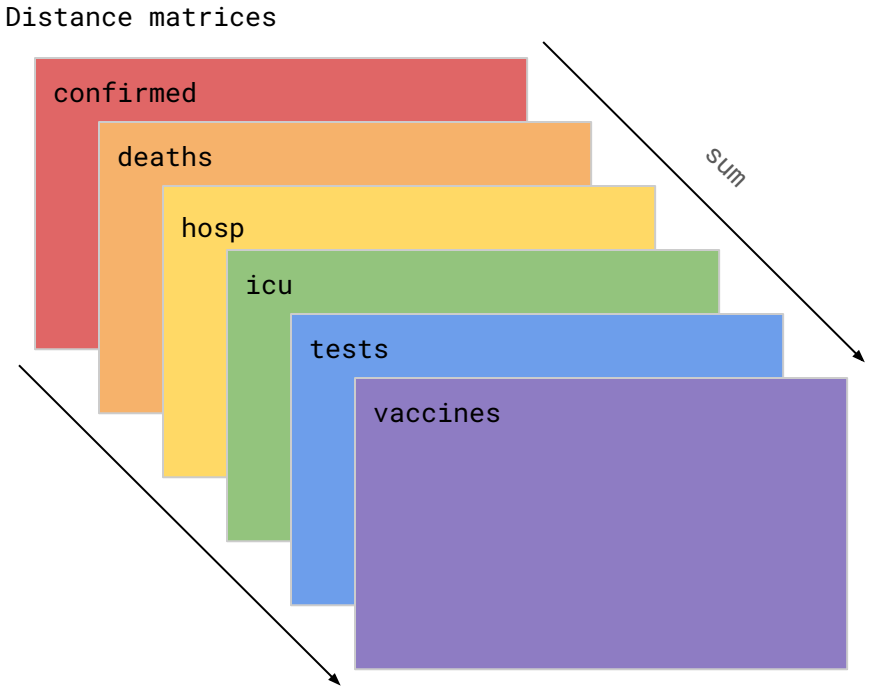
Clustering (1 of 5) Calculating Distances

```r
# Initialize distances
distances <- 0

# Calculate distances
for(i in features) {
  temp <- data
  temp <- select(temp, date,
                 administrative_area_level_2, i)
  temp <- spread(temp, administrative_area_level_2, i)
  temp <- select(temp, -date)
  temp <- t(temp)
  temp <- dist(temp)
  distances <- distances + temp
}

# Hierarchical clustering
hc <- hclust(distances)
```

```
# A tibble: 50 × 50
# Feature: vaccines
          Alabama Alaska Arizona Arkansas California
            <dbl>  <dbl>   <dbl>    <dbl>      <dbl>
Alabama       0
Alaska       14.6      0
Arizona       9.19   11.2     0
Arkansas      7.42   13.7     9.66      0
California   11.4    16.9     7.37     13.3        0
Colorado     12.4    16.9     9.22     14.2        5.35
Connecticut  12.7    17.2     8.82     14.3        3.82
Delaware     11.9    16.3     7.85     12.0        5.33
Florida       6.56   15.2     8.63      9.28       8.20
Georgia       6.67   16.2     9.00      9.88       8.17
# ℹ 40 more rows
# ℹ 45 more variables: Colorado <dbl>, …
```

Clustering (1 of 5) Calculating Distances

```r
# Initialize distances
distances <- 0

# Calculate distances
for(i in features) {
  temp <- data
  temp <- select(temp, date,
                 administrative_area_level_2, i)
  temp <- spread(temp, administrative_area_level_2, i)
  temp <- select(temp, -date)
  temp <- t(temp)
  temp <- dist(temp)
  distances <- distances + temp
}

# Hierarchical clustering
hc <- hclust(distances)
```



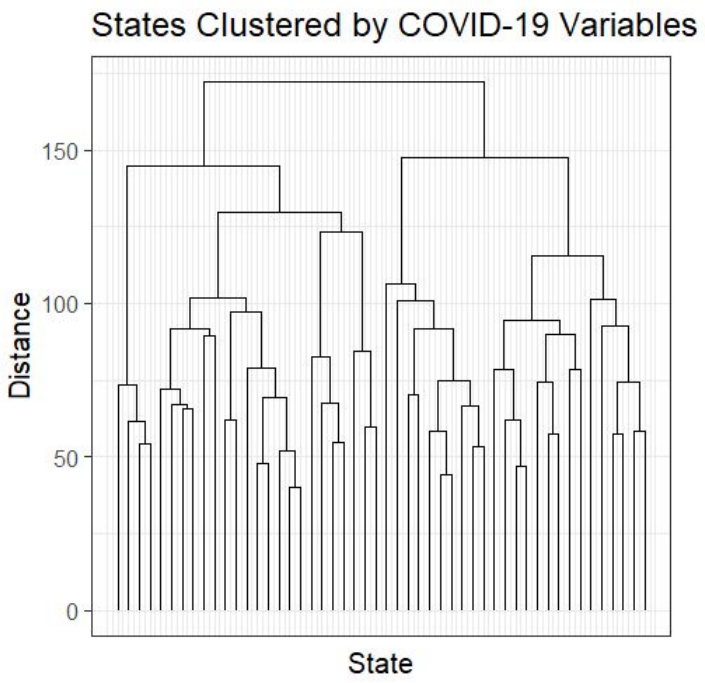
Clustering (1 of 5) Calculating Distances

```r
# Initialize distances
distances <- 0

# Calculate distances
for(i in features) {
  temp <- data
  temp <- select(temp, date,
                 administrative_area_level_2, i)
  temp <- spread(temp, administrative_area_level_2, i)
  temp <- select(temp, -date)
  temp <- t(temp)
  temp <- dist(temp)
  distances <- distances + temp
}

# Hierarchical clustering
hc <- hclust(distances)
```

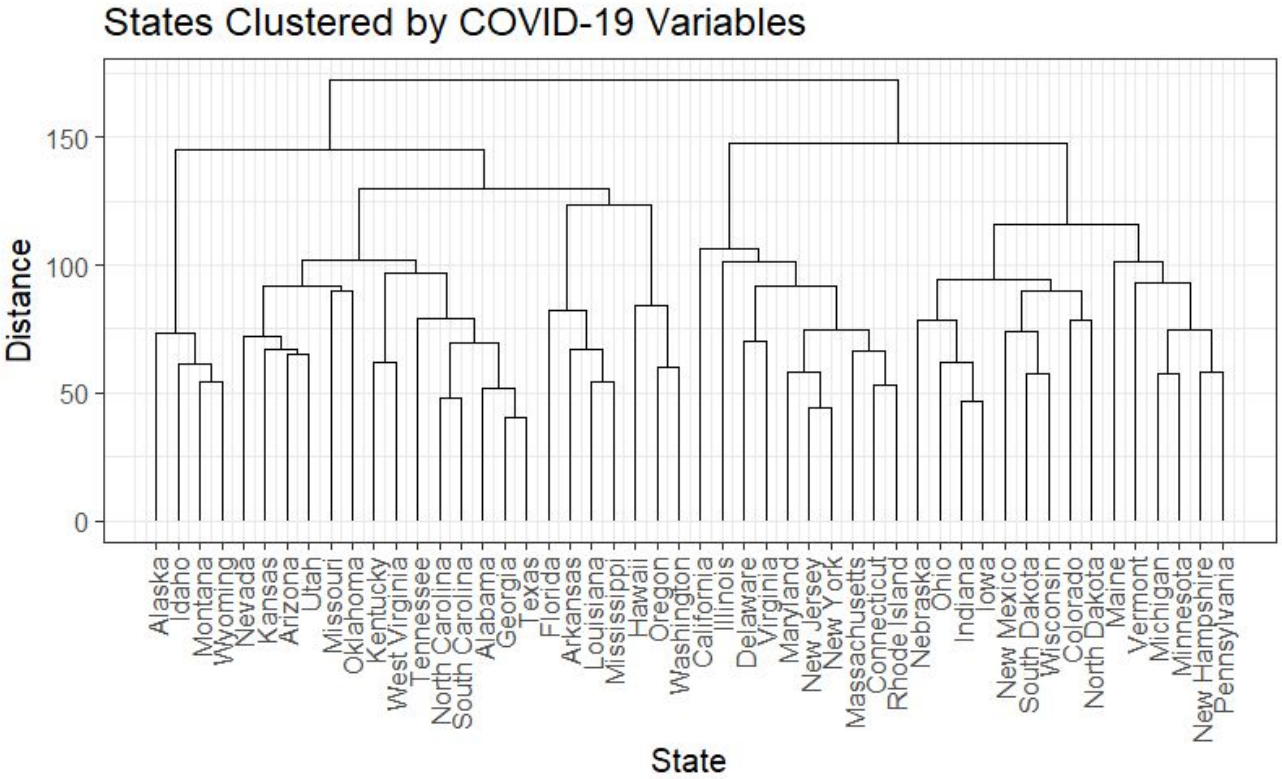
```
# A tibble: 50 × 50
# Features: sum(all)
          Alabama  Alaska Arizona Arkansas California
            <dbl>   <dbl>   <dbl>    <dbl>      <dbl>
Alabama       0
Alaska      127.        0
Arizona      88.0     127.      0
Arkansas     64.1     126.    88.5       0
California   89.2     153.    72.9     103.          0
Colorado    135.      123.    97.8     123.       121.
Connect.    129.      168.   101.      132.       103.
Delaware    117.      148.    92.0     119.       101.
Florida      81.2     135.   122.       82.4      119.
Georgia      52.0     132.    91.4      78.1       87.1
# ℹ 40 more rows
# ℹ 45 more variables: Colorado <dbl>, …
```

Clustering (2 of 5) Hierarchical Clustering

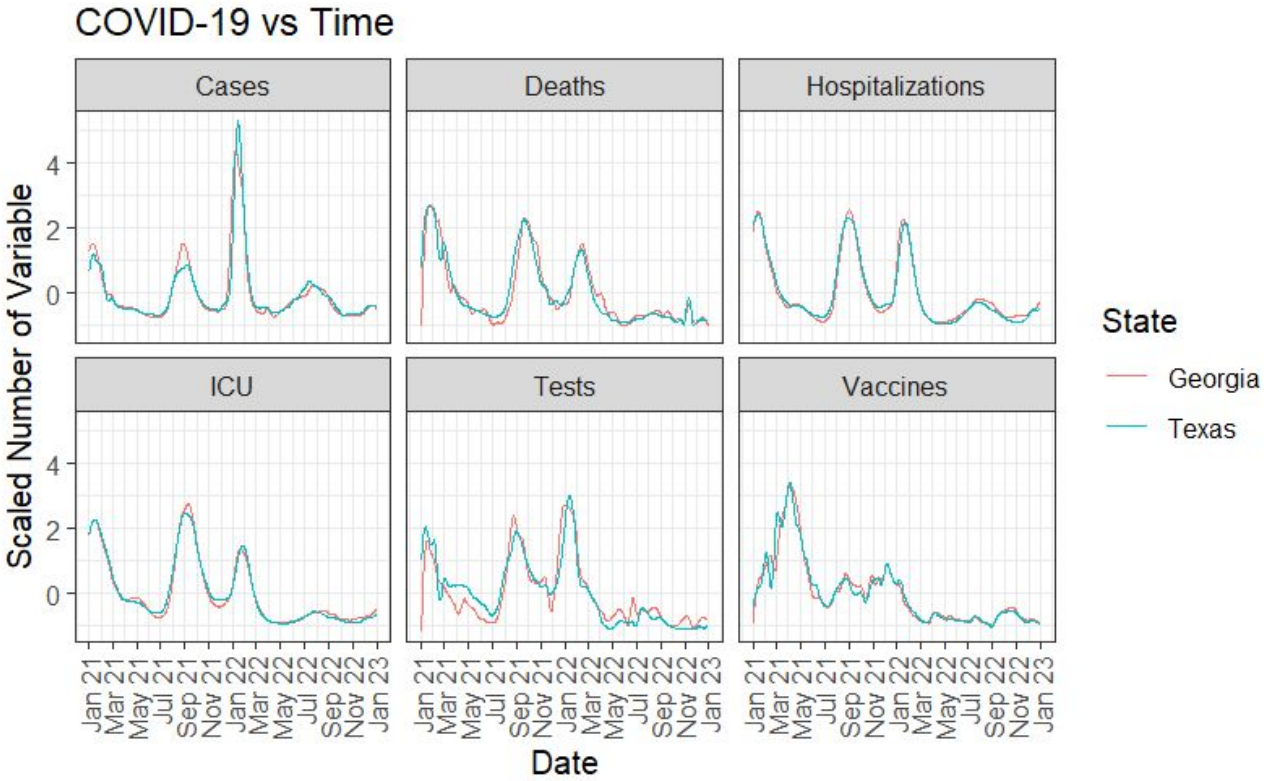
```r
# Initialize distances
distances <- 0

# Calculate distances
for(i in features) {
  temp <- data
  temp <- select(temp, date,
                 administrative_area_level_2, i)
  temp <- spread(temp, administrative_area_level_2, i)
  temp <- select(temp, -date)
  temp <- t(temp)
  temp <- dist(temp)
  distances <- distances + temp
}

# Hierarchical clustering
hc <- hclust(distances)
```



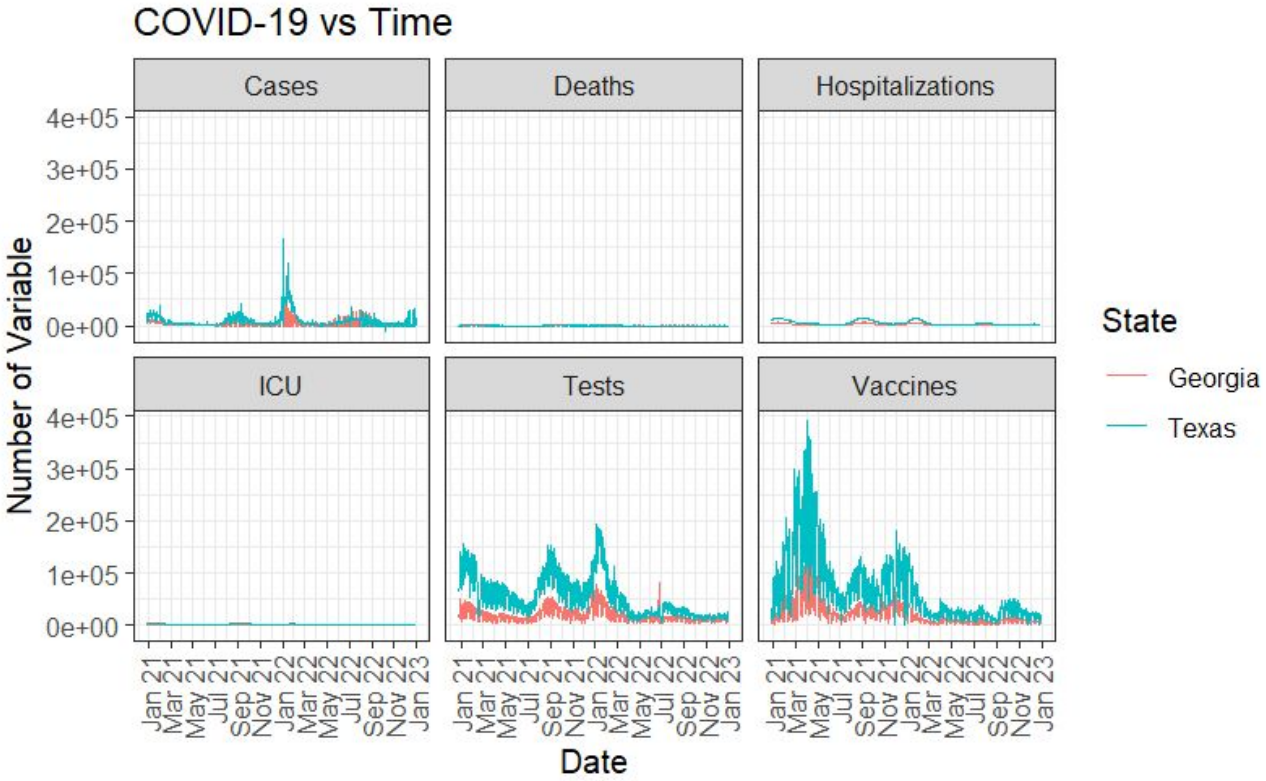
Clustering (2 of 5) Hierarchical Clustering



Clustering (2 of 5) Hierarchical Clustering



Clustering (2 of 5) Hierarchical Clustering

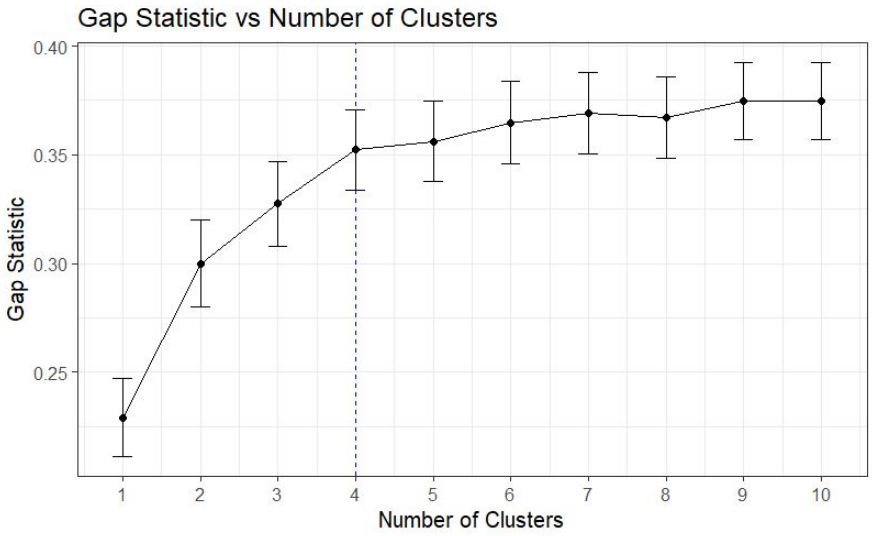


Clustering (3 of 5) Gap Statistic

```r
# Calculate gap statistics
clusGap(distances, FUN = hclust)

# Cut tree
clusters <- cutree(hc, k = 4)

# Plot clusters
states <- map_data("state")
ggplot(states, aes(x = longitude, y = latitude)) +
  geom_polygon(aes(fill = clusters))
```

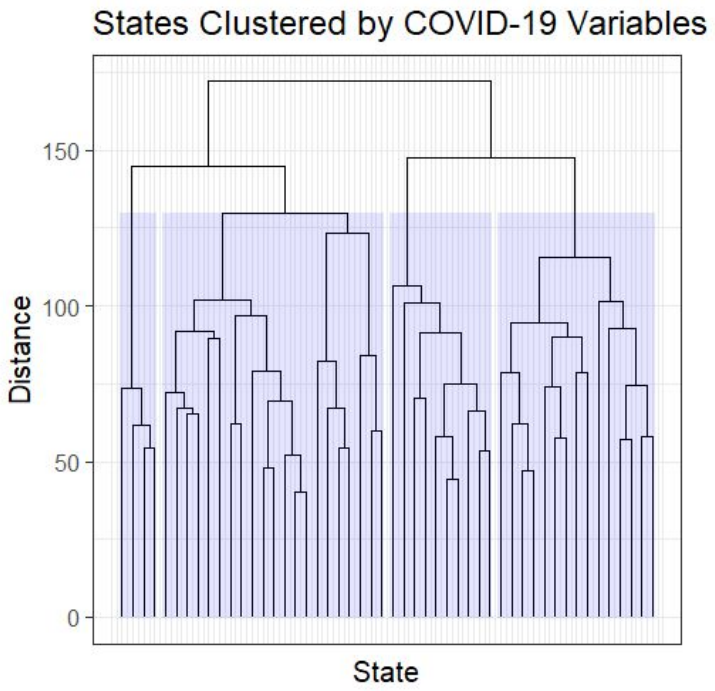
Clustering (3 of 5) Gap Statistic

```r
# Calculate gap statistics
clusGap(distances, FUN = hclust)

# Cut tree
clusters <- cutree(hc, k = 4)

# Plot clusters
states <- map_data("state")
ggplot(states, aes(x = longitude, y = latitude)) +
  geom_polygon(aes(fill = clusters))
```

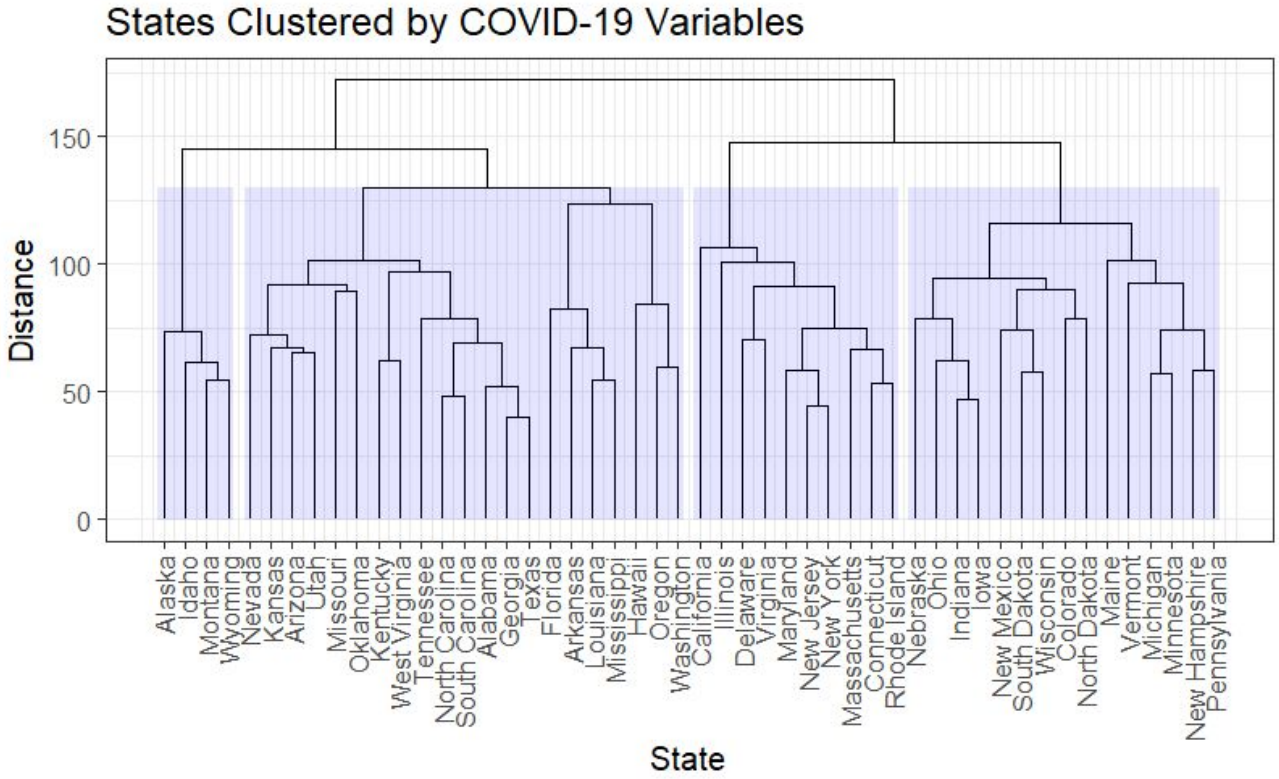


Clustering (4 of 5) Cut Tree

```r
# Calculate gap statistics
clusGap(distances, FUN = hclust)

# Cut tree
clusters <- cutree(hc, k = 4)

# Plot clusters
states <- map_data("state")
ggplot(states, aes(x = longitude, y = latitude)) +
  geom_polygon(aes(fill = clusters))
```

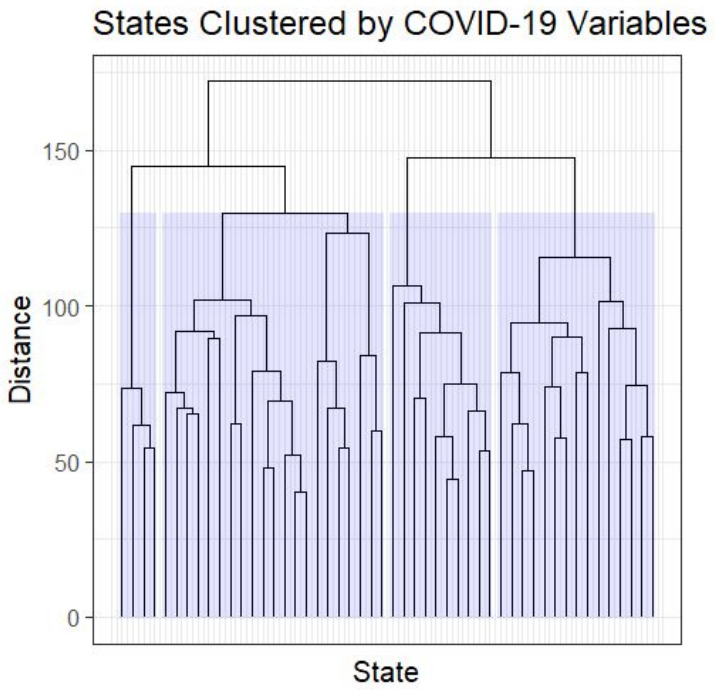


Clustering (4 of 5) Cut Tree



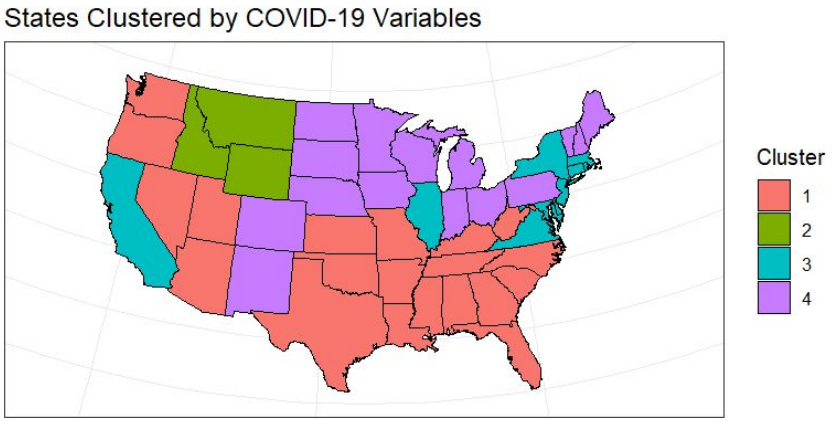
Clustering (5 of 5) Plot Clusters!

```r
# Calculate gap statistics
clusGap(distances, FUN = hclust)

# Cut tree
clusters <- cutree(hc, k = 4)

# Plot clusters
states <- map_data("state")
ggplot(states, aes(x = longitude, y = latitude)) +
  geom_polygon(aes(fill = clusters))
```

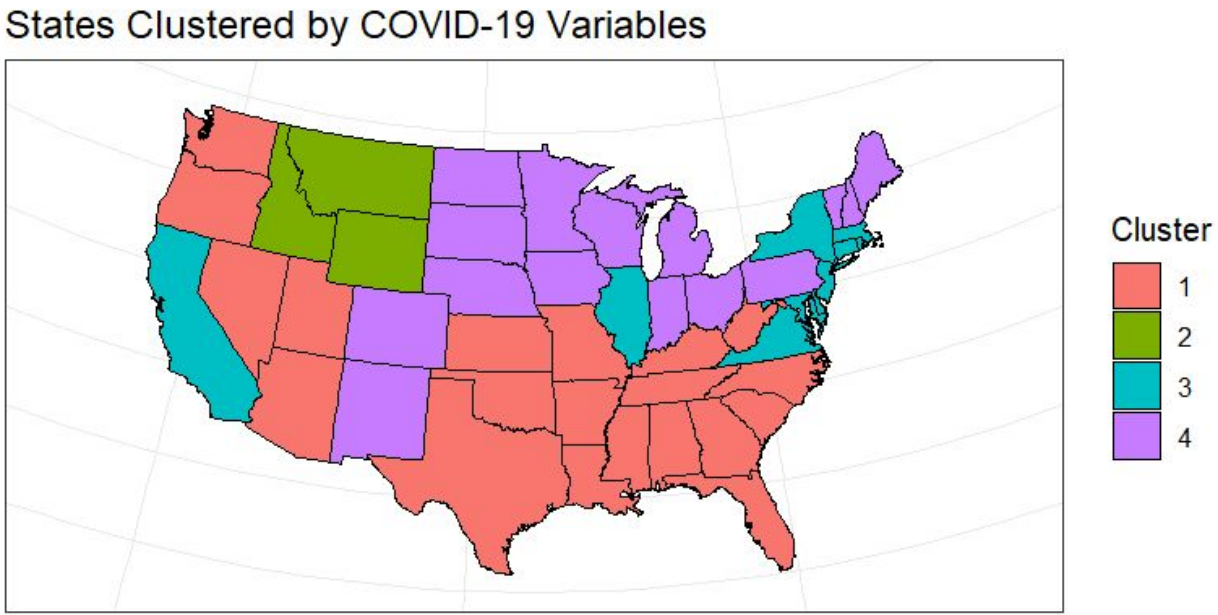


Clustering (5 of 5) Plot Clusters!

```r
# Calculate gap statistics
clusGap(distances, FUN = hclust)

# Cut tree
clusters <- cutree(hc, k = 4)

# Plot Clusters
states <- map_data("state")
ggplot(states, aes(x = longitude, y = latitude)) +
  geom_polygon(aes(fill = clusters))
```



Clustering (5 of 5) Plot Clusters!



Note: Alaska = 2, Hawaii = 1.

Conclusion

We utilized hierarchical clustering on multivariate time series data of COVID-19 metrics to uncover distinct regional patterns of the pandemic’s impact throughout the US.

Our approach included data cleaning, preprocessing, and employing the gap statistic to determine the optimal number of clusters.

Pronounced clusters in the South and Midwest.

A smaller cluster in the Northwest

A distinctive grouping that included several Northeastern states, Illinois, and California.

These results offer critical insights for shaping targeted public health policies and underscore the pivotal role of data-driven analysis in comprehending and addressing the challenges of the pandemic.

References

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). The new S language. Wadsworth & Brooks.

Cleveland, W. S., Grosse, E., & Shyu, W. M. (2017). Local regression models. In W. S. Cleveland & R. McGill (Eds.), Statistical models in S (pp. 309–376). Routledge.

Guidotti, E. (2022). A worldwide epidemiological database for COVID-19 at fine-grained spatial resolution. Scientific Data, 9(1), 112. https://doi.org/10.1038/s41597-022-01245-1

Guidotti, E., & Ardia, D. (2020). COVID-19 data hub. Journal of Open Source Software, 5(51), 2376. https://doi.org/10.21105/joss.02376

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411–423.