# Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-07-21

## Contents



## Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by a new type of coronavirus: severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The outbreak first started in Wuhan, China in December 2019. The first kown case of COVID-19 in the U.S. was confirmed on January 20, 2020, in a 35-year-old man who teturned to Washington State on January 15 after traveling to Wuhan. Starting around the end of Feburary, evidence emerge for community spread in the US.

We, as all of us, are indebted to the heros who fight COVID-19 across the whole world in different ways. For this data exploration, I am grateful to many data science groups who have collected detailed COVID-19 outbreak data, including the number of tests, confirmed cases, and deaths, across countries/regions, states/provnices (administrative division level 1, or admin1), and counties (admin2). Specifically, I used the data from these three resources:

- JHU (https://coronavirus.jhu.edu/)
  - The Center for Systems Science and Engineering (CSSE) at John Hopkins University.
  - World-wide counts of coronavirus cases, deaths, and recovered ones.
  - https://github.com/CSSEGISandData/COVID-19
- NY Times (https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html)
  - The New York Times
  - "cumulative counts of coronavirus cases in the United States, at the state and county level, over time"
  - https://github.com/nytimes/covid-19-data

- COVID Trackng (https://covidtracking.com/)
  - COVID Tracking Project
  - "collects information from 50 US states, the District of Columbia, and 5 other US territories to provide the most comprehensive testing data"
  - https://github.com/COVID19Tracking/covid-tracking-data

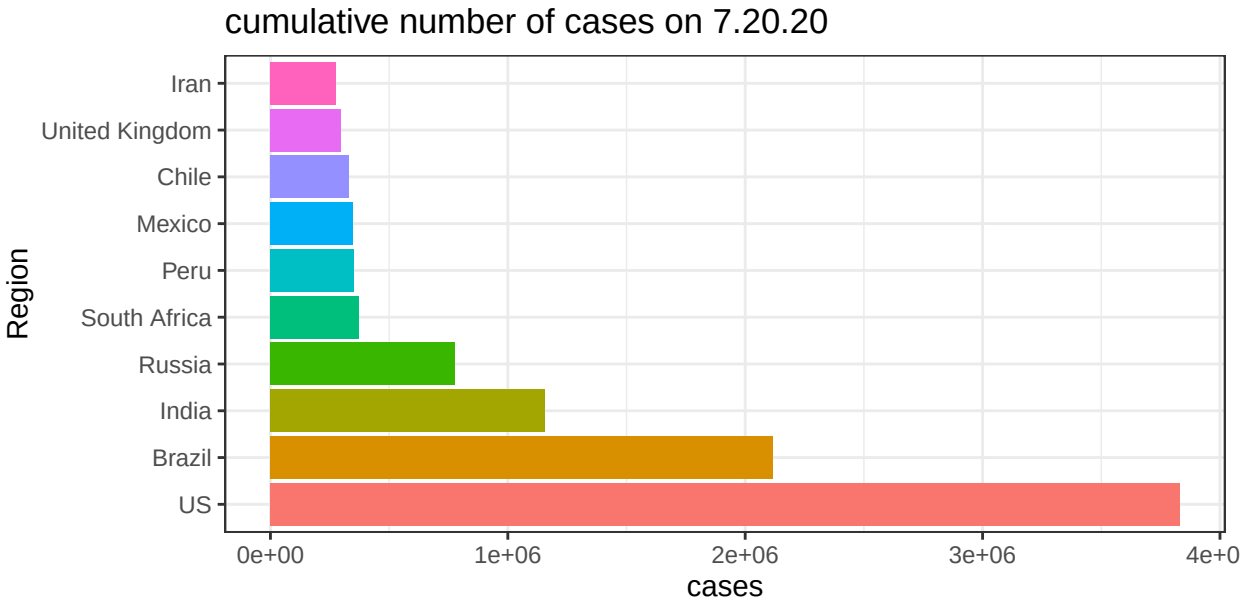
# JHU

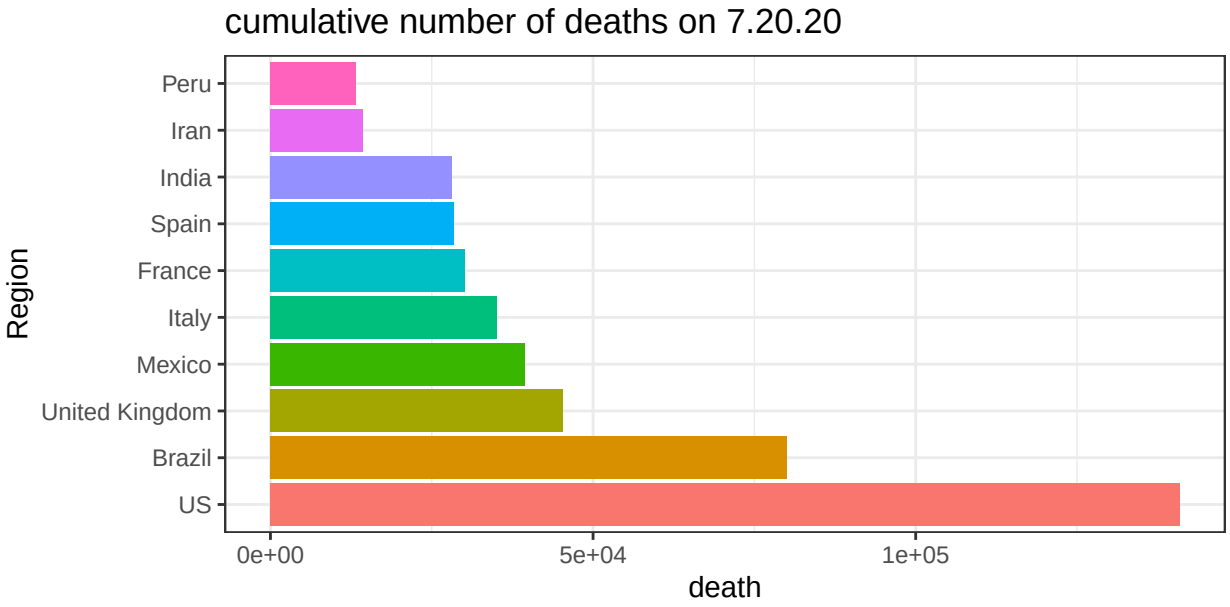
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

### time series data

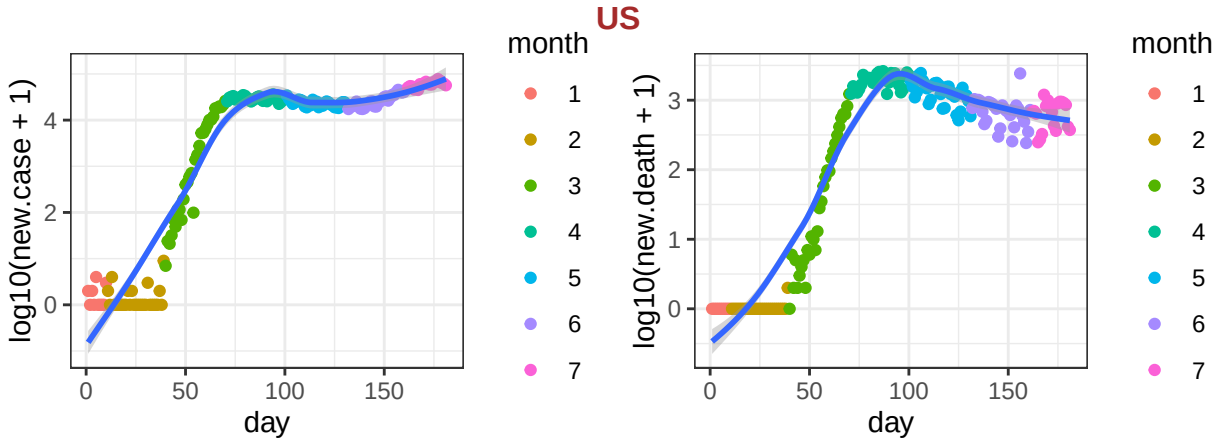
The time series provide counts (e.g., confirmed cases, deaths) starting from Jan 22nd, 2020 for 253 locations. Currently there is no data of individual US state in these time series data files.

Here is the list of 10 records with the largest number of cases or deaths on the most recent date.

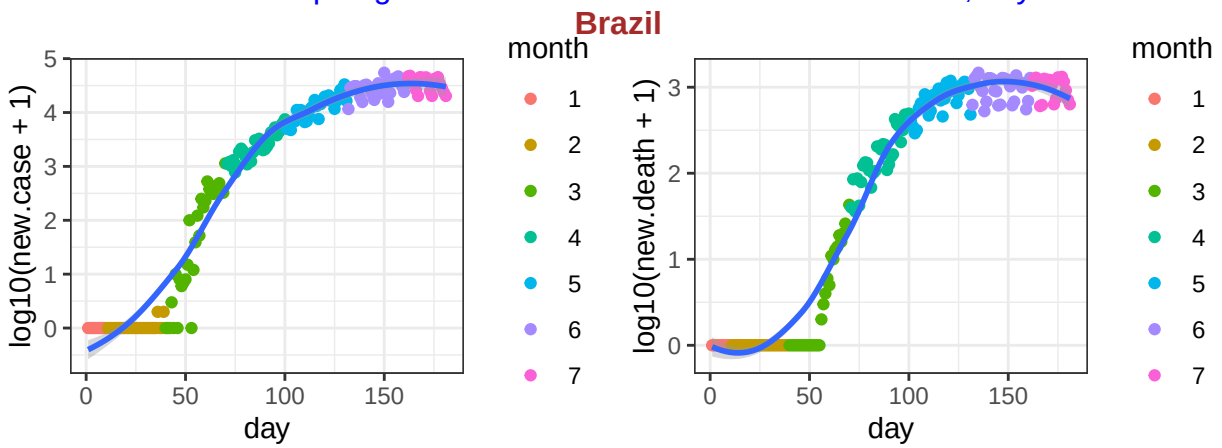





Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

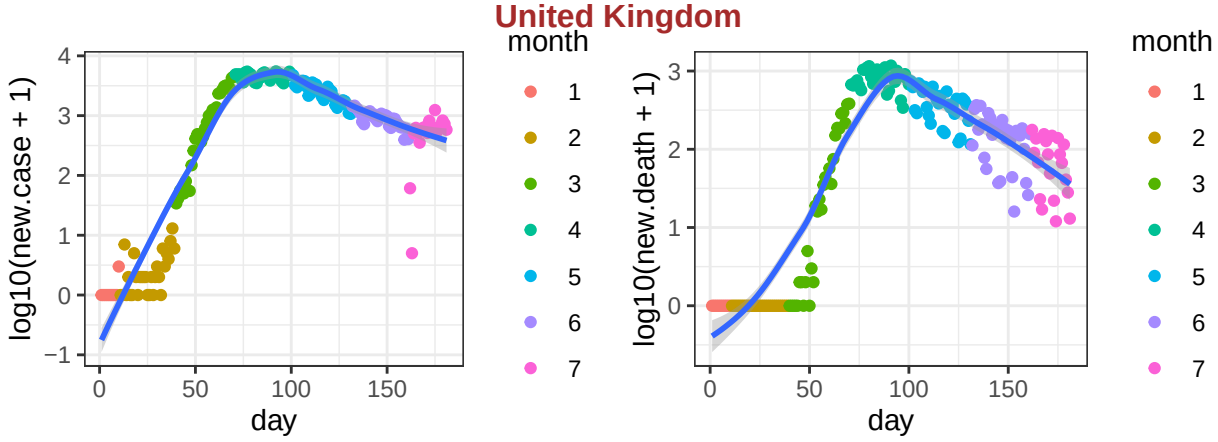



data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



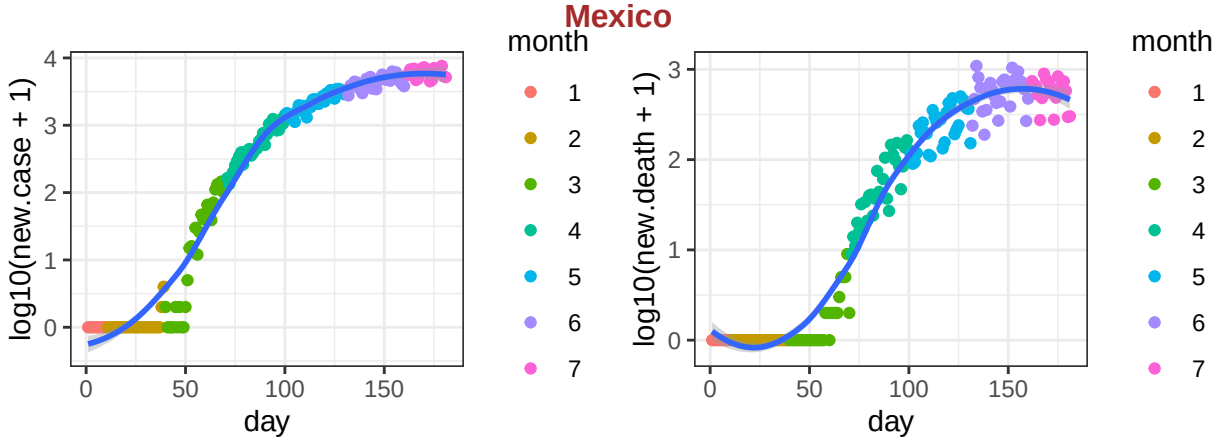

data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

## United Kingdom



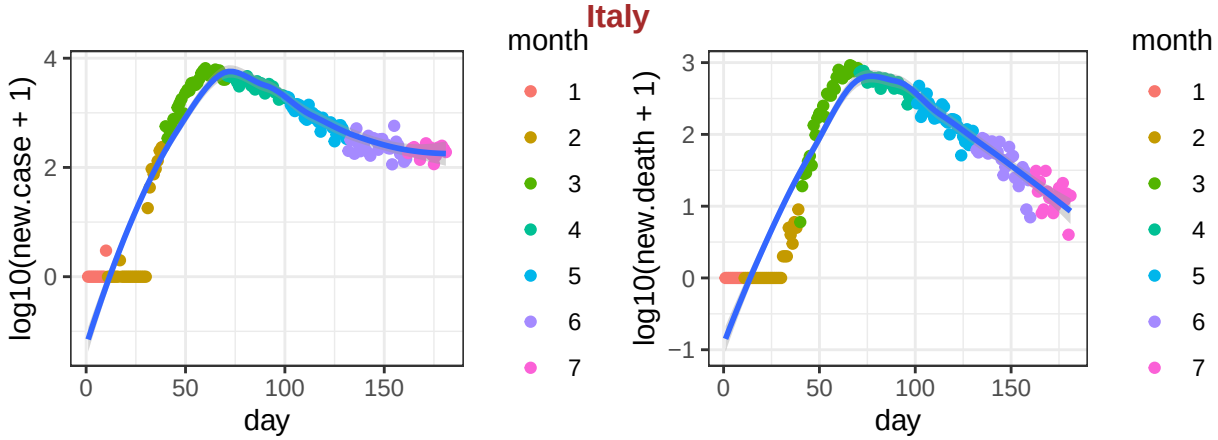

data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

## Mexico



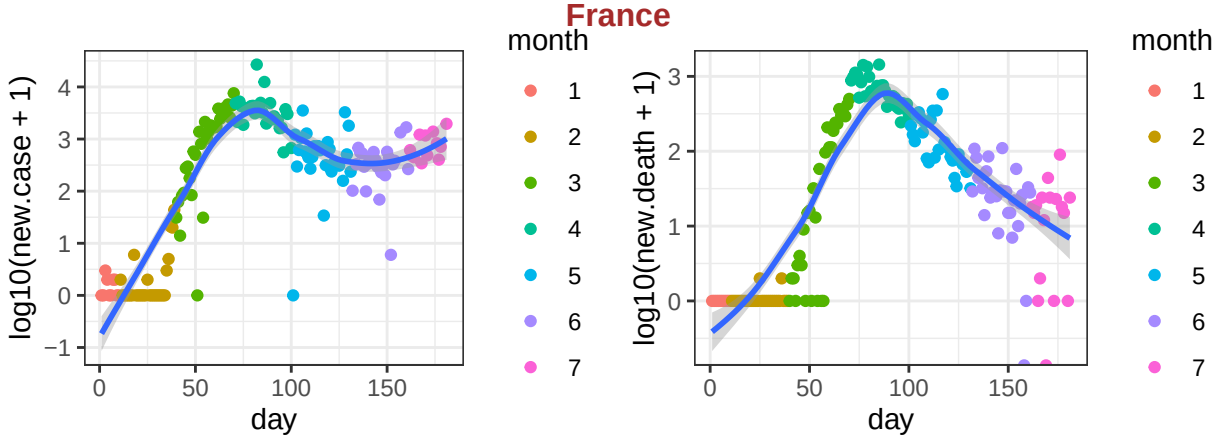

data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

## Italy



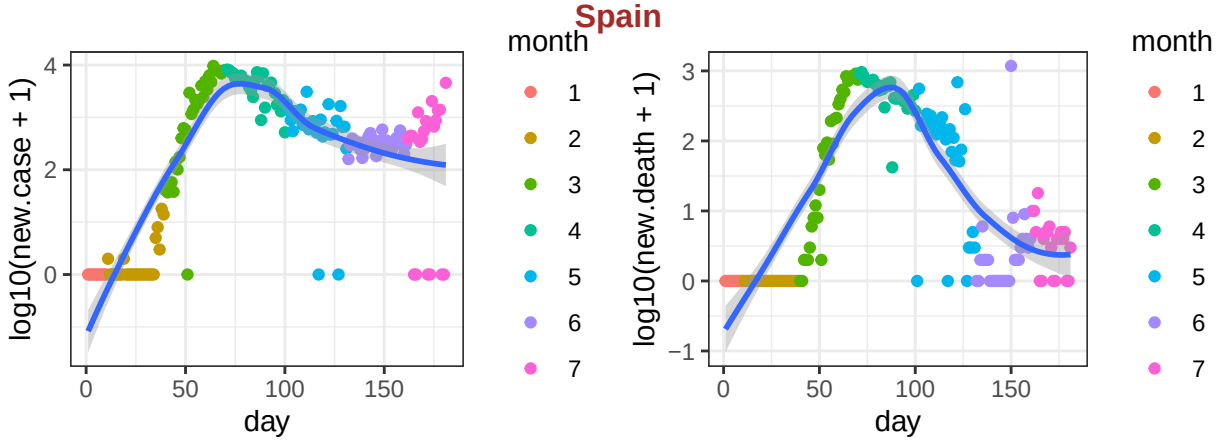

data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

**France**



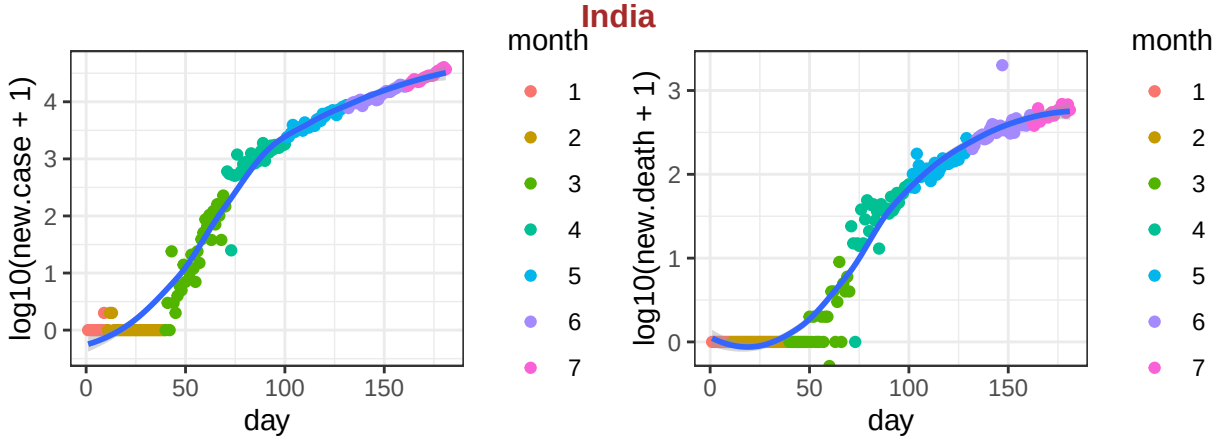

data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

**Spain**

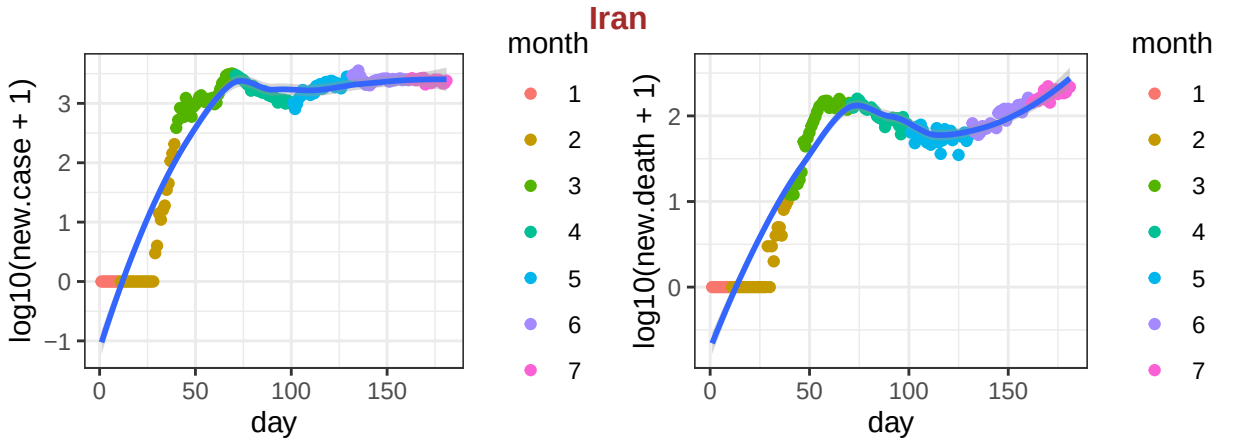



data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

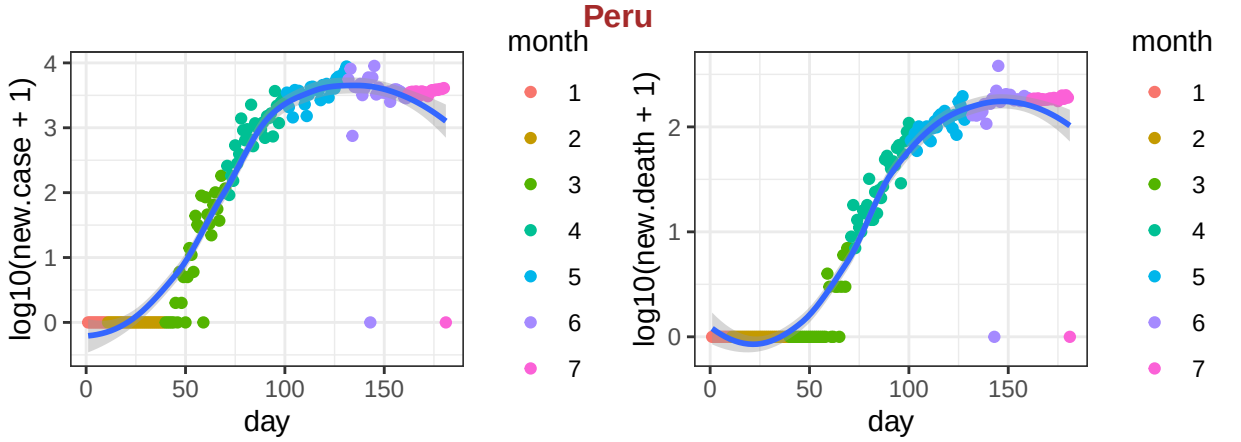
**India**




data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



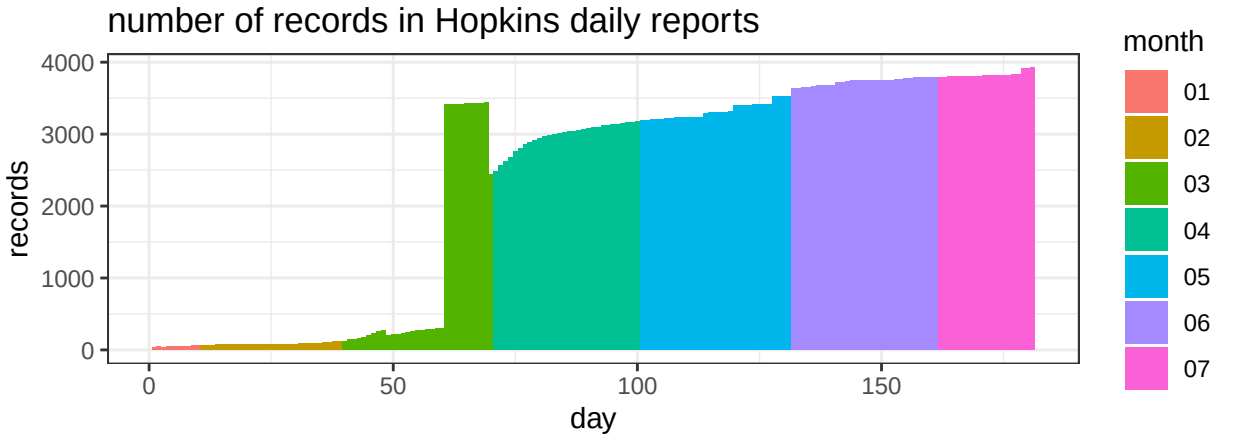

data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020




data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

### daily reports data

The raw data from Hopkins are in the format of daily reports with one file per day. More recent files (since March 22nd) inlcude information from individual states of US or individual counties, as shown in the following figure. So I turn to NY Times data for informatoin of individual states or counties.




data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

# NY Times

The data from NY Times are saved in two text files, one for state level information and the other one for county level information.

The currente date is

```
## [1] "2020-07-20"
```

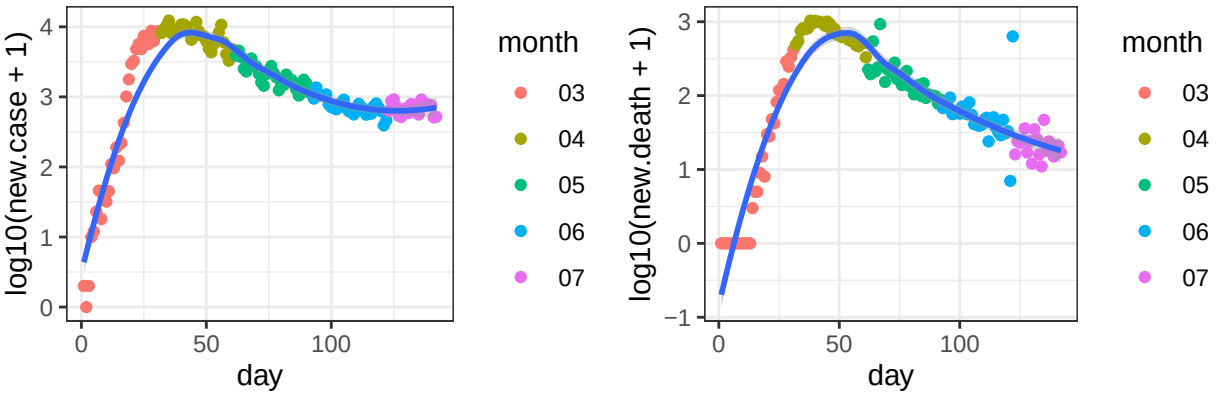
## state level data

First check the 30 states with the largest number of deaths.

```
##            date          state fips  cases deaths
## 7693 2020-07-20       New York   36 412034  32203
## 7691 2020-07-20     New Jersey   34 178937  15715
## 7682 2020-07-20  Massachusetts   25 113789   8433
## 7664 2020-07-20     California    6 400195   7764
## 7674 2020-07-20       Illinois   17 164224   7500
## 7700 2020-07-20   Pennsylvania   42 106498   7076
## 7683 2020-07-20       Michigan   26  82486   6377
## 7669 2020-07-20        Florida   12 360386   5071
## 7666 2020-07-20    Connecticut    9  48055   4406
## 7706 2020-07-20          Texas   48 345672   4160
## 7679 2020-07-20      Louisiana   22  95002   3572
## 7681 2020-07-20       Maryland   24  79251   3382
## 7697 2020-07-20           Ohio   39  76168   3189
## 7670 2020-07-20        Georgia   13 132788   3113
## 7675 2020-07-20        Indiana   18  58607   2825
## 7662 2020-07-20        Arizona    4 145320   2795
## 7710 2020-07-20       Virginia   51  78375   2031
## 7665 2020-07-20       Colorado    8  40649   1759
## 7694 2020-07-20 North Carolina   37 101286   1676
## 7684 2020-07-20      Minnesota   27  47147   1585
## 7711 2020-07-20     Washington   53  49949   1521
## 7685 2020-07-20    Mississippi   28  43889   1358
## 7660 2020-07-20        Alabama    1  68891   1291
## 7686 2020-07-20       Missouri   29  36048   1164
## 7703 2020-07-20 South Carolina   45  71445   1164
## 7702 2020-07-20   Rhode Island   44  17904    995
## 7713 2020-07-20      Wisconsin   55  46754    855
## 7705 2020-07-20      Tennessee   47  77944    838
## 7676 2020-07-20           Iowa   19  39272    797
## 7678 2020-07-20       Kentucky   21  23978    694
```

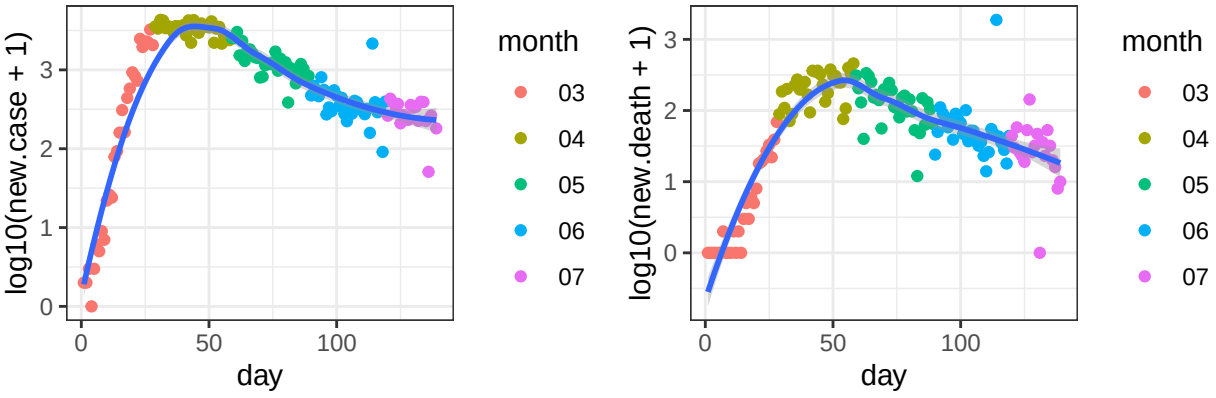
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the similarities and differences across states.

New York



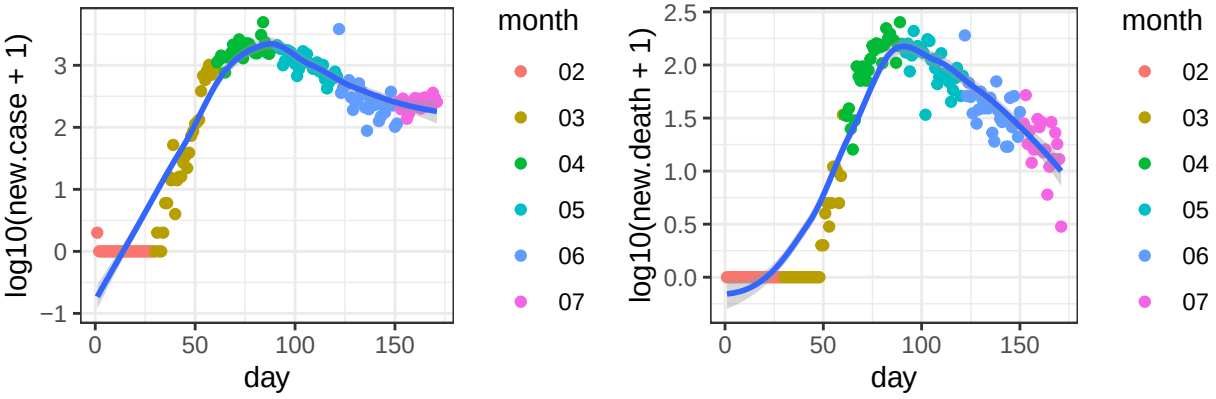

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01*

New Jersey



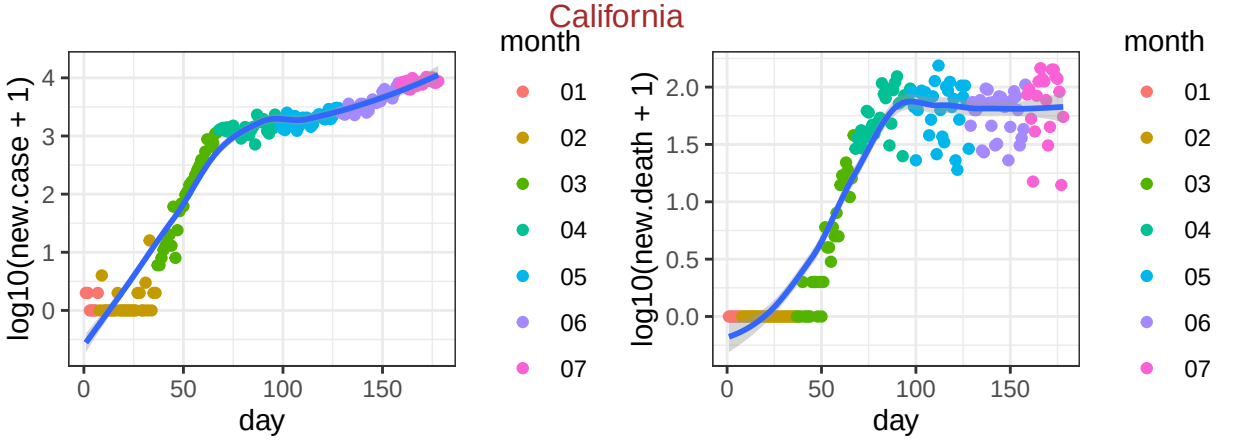

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04*

Massachusetts



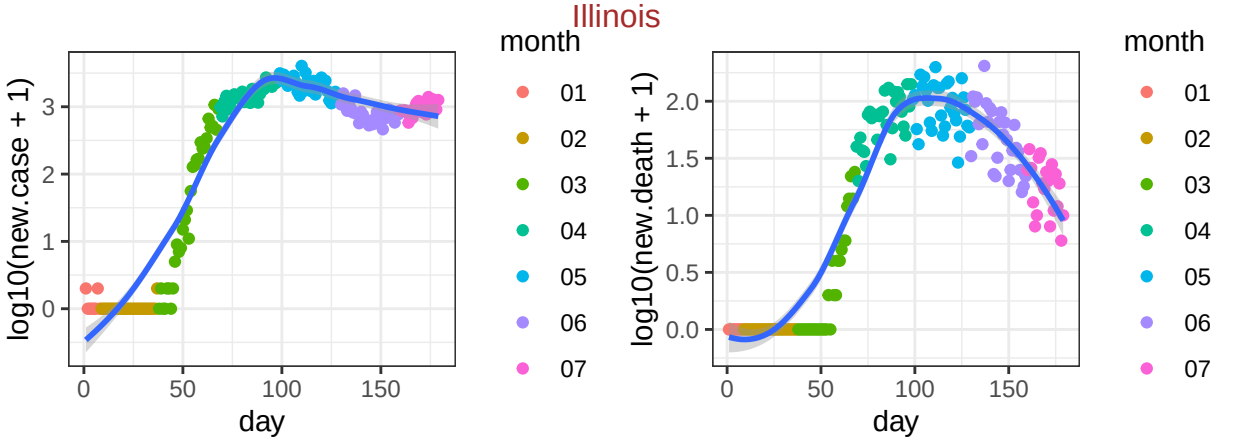

*data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01*

California



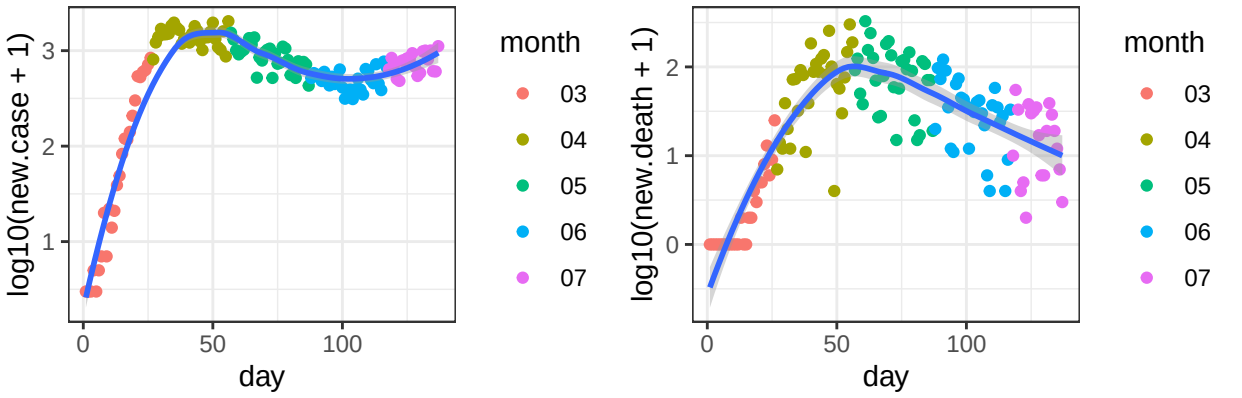

*data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25*

Illinois



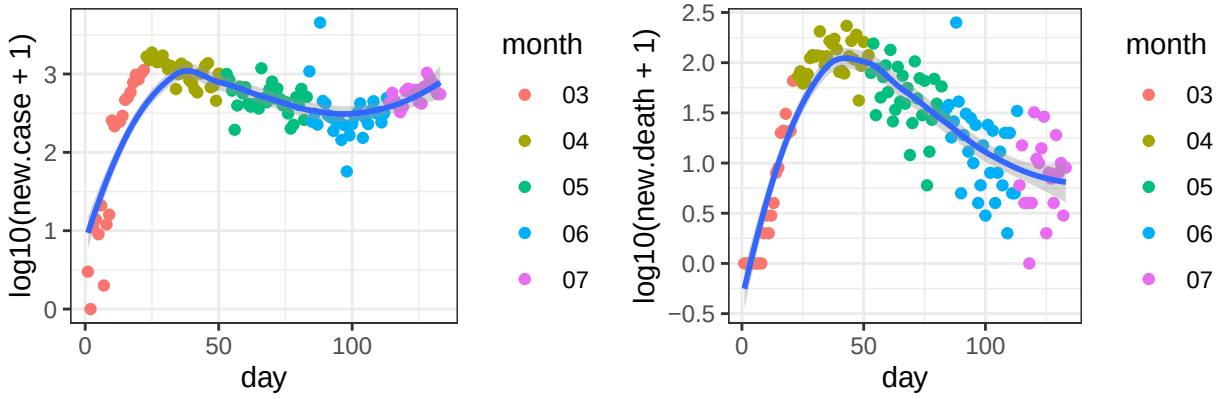

*data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24*

Pennsylvania



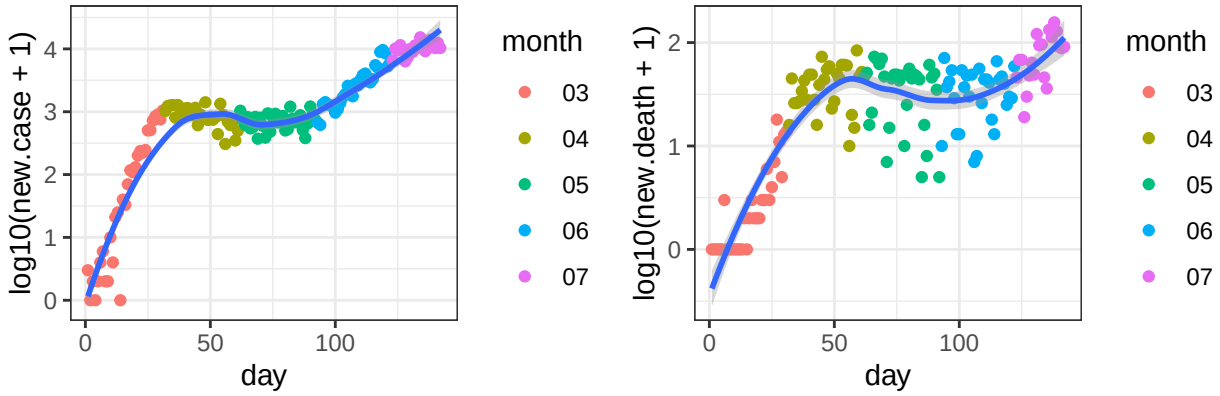

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06*

Michigan



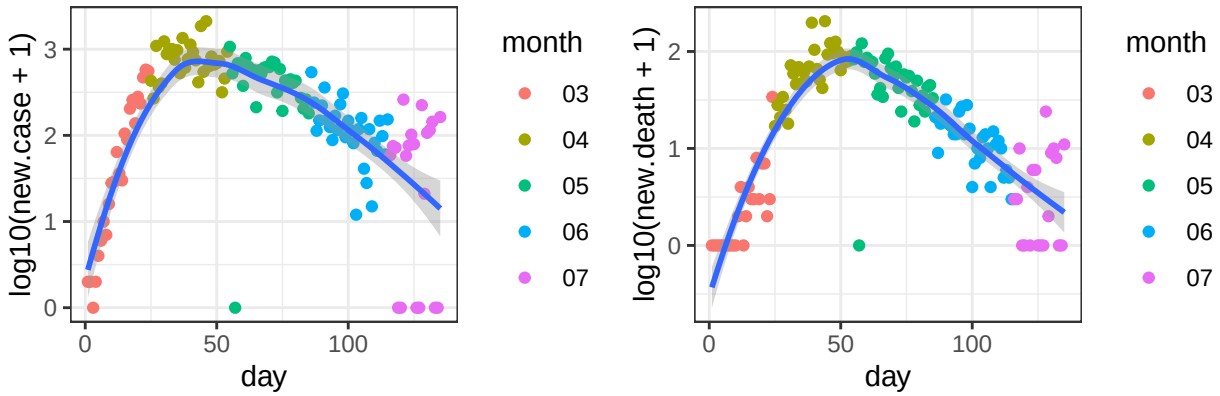

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10*

Florida



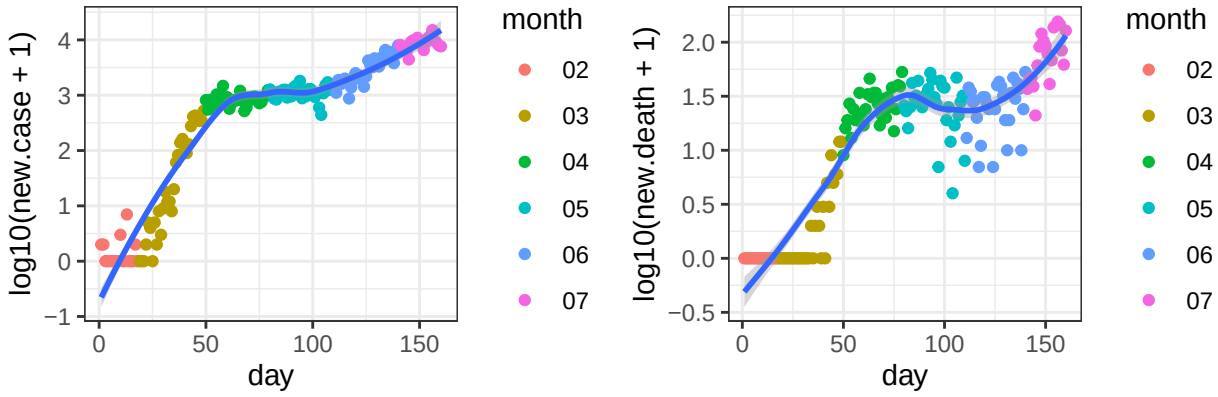

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01*

Connecticut



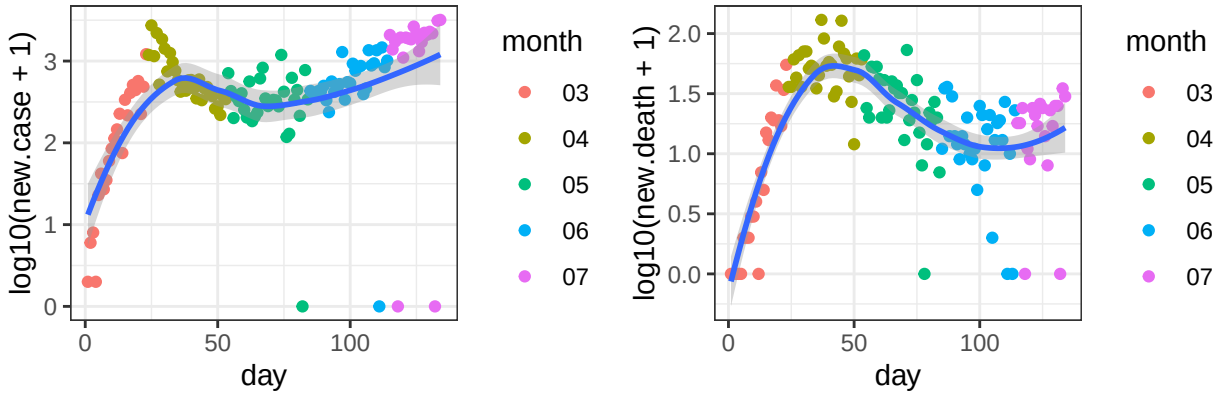

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08*

Texas



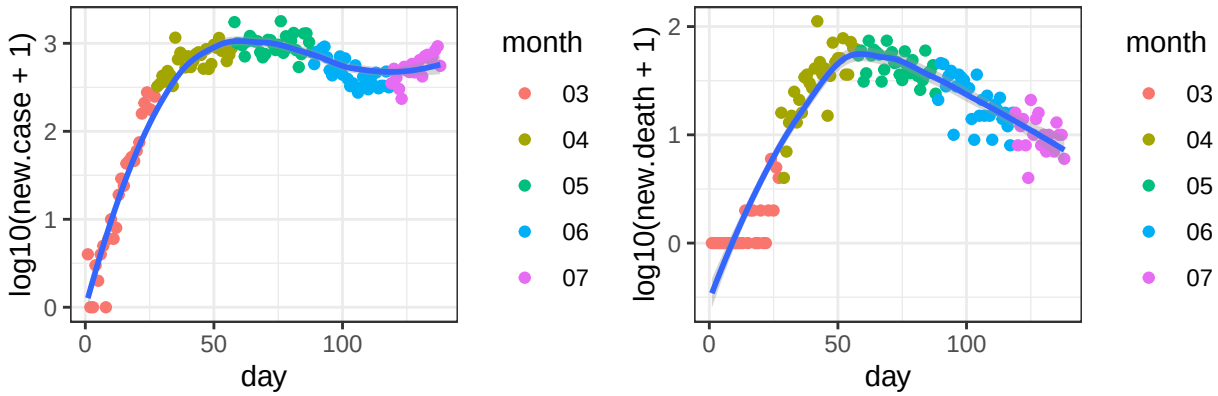

*data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12*

Louisiana



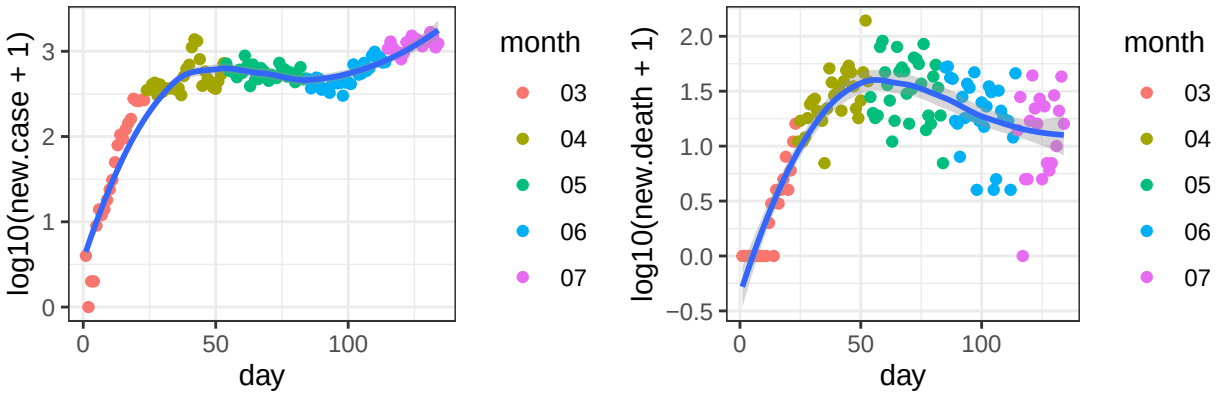

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09*

Maryland



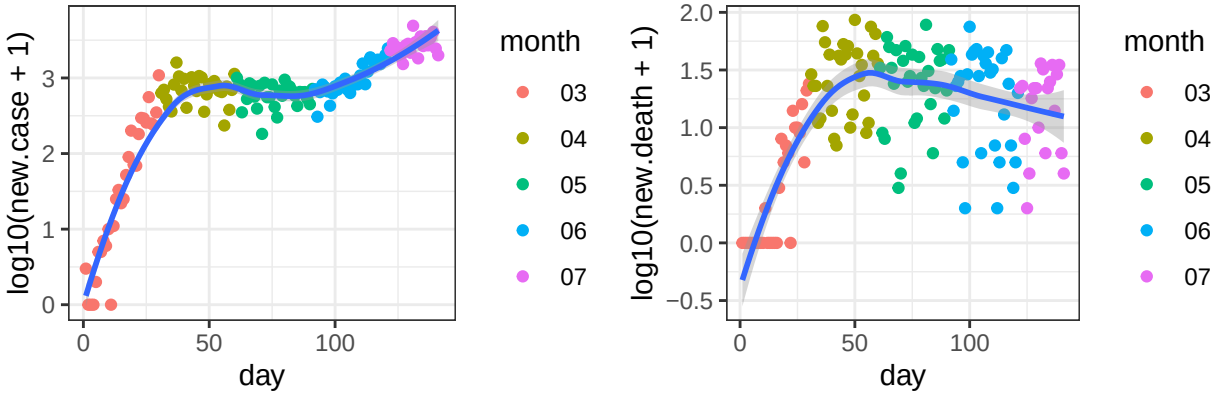

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05*

Ohio



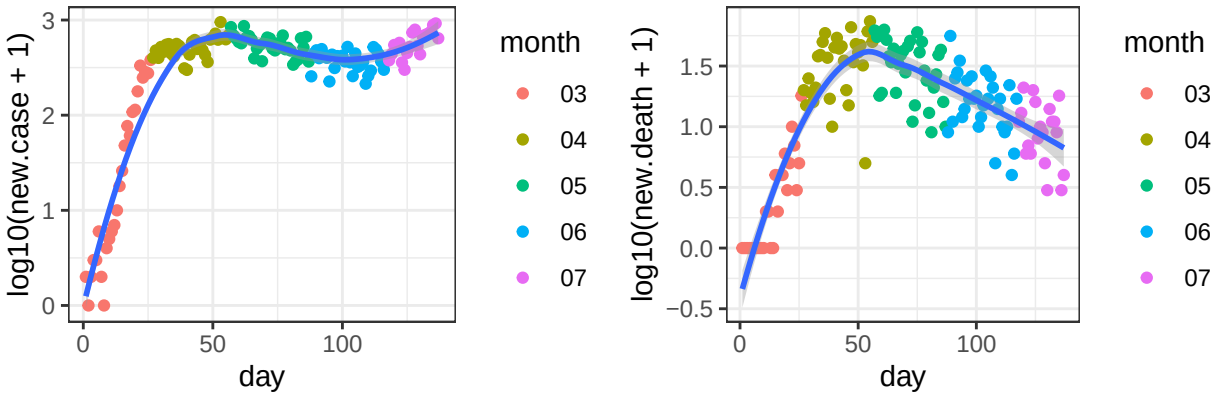

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09*

Georgia



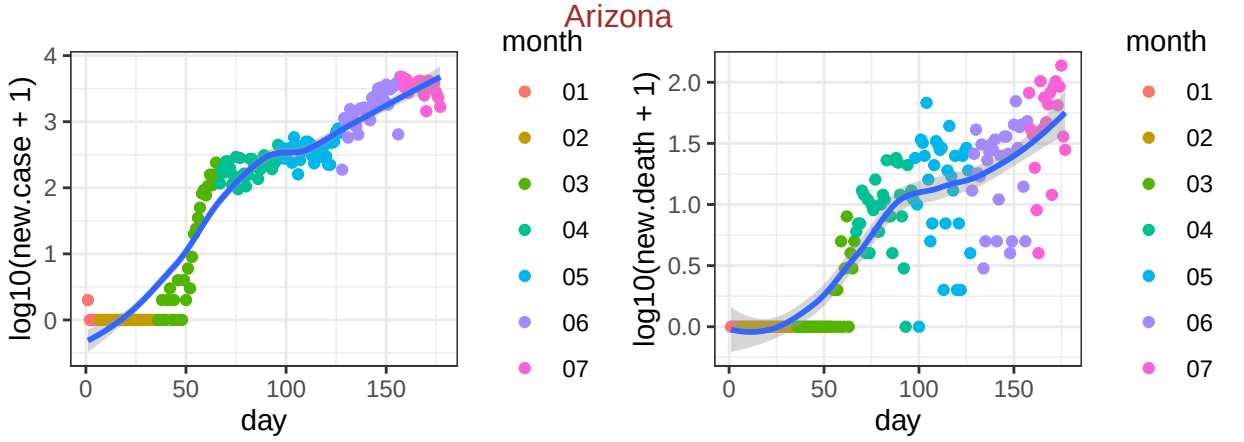

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02*

Indiana



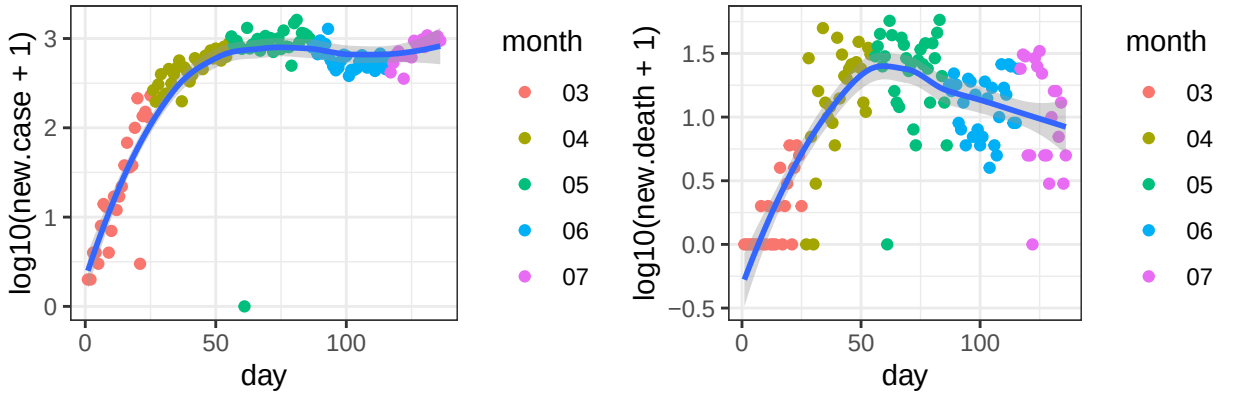

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06*

Arizona



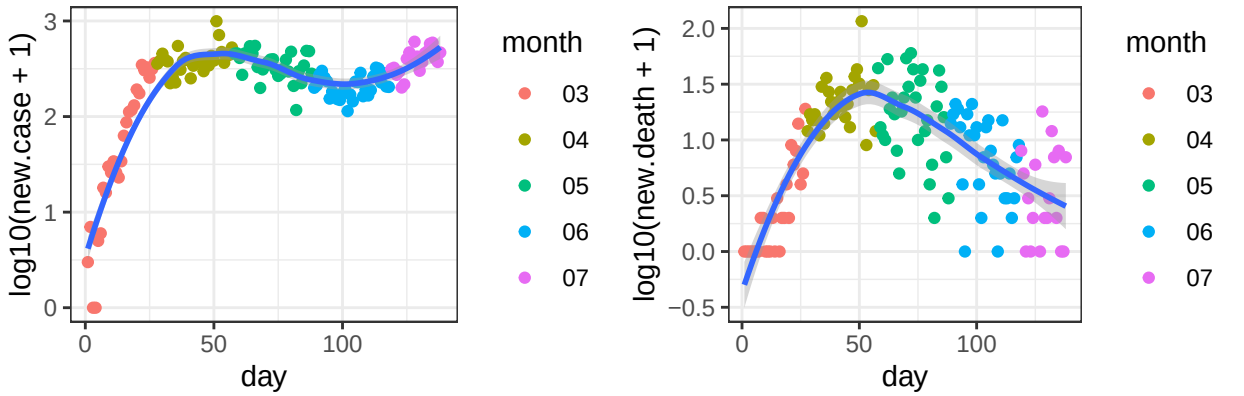

*data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26*

Virginia



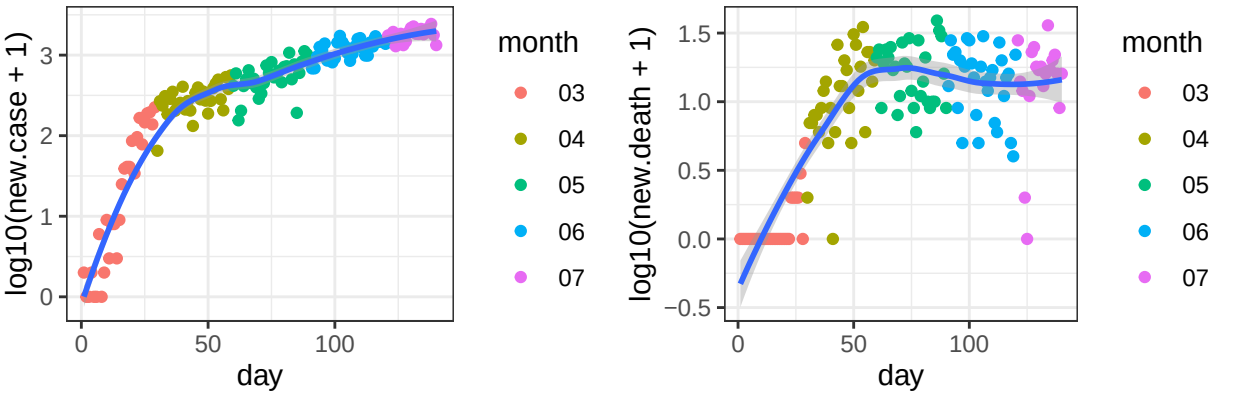

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07*

Colorado



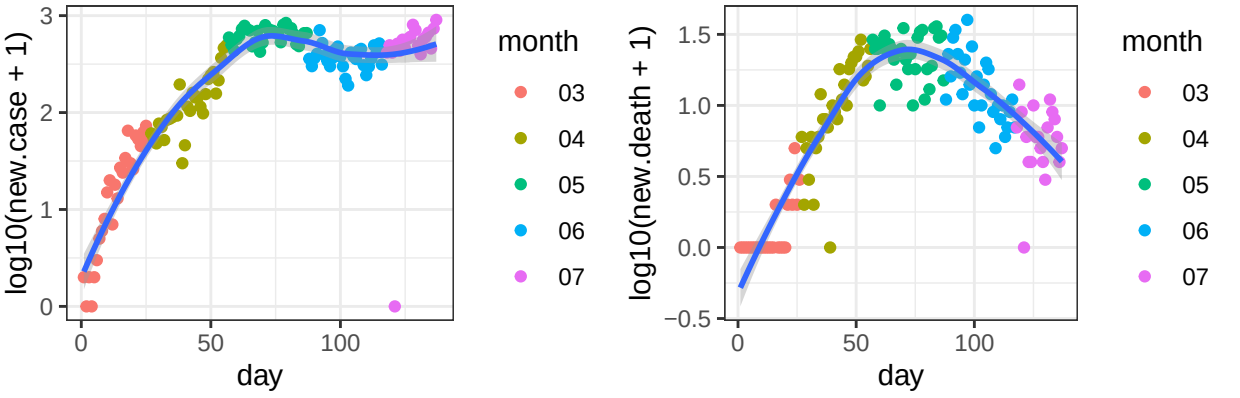

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05*

North Carolina



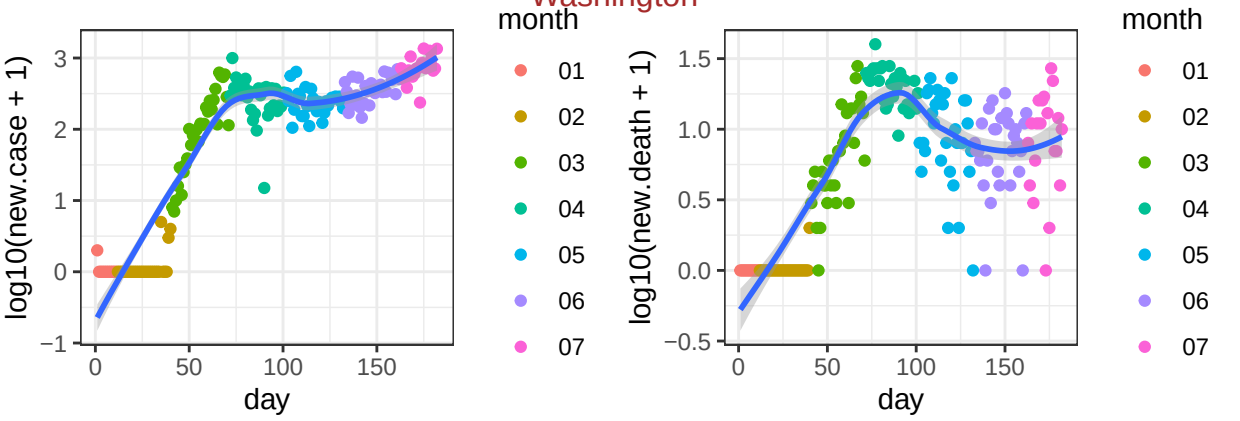

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03*

Minnesota



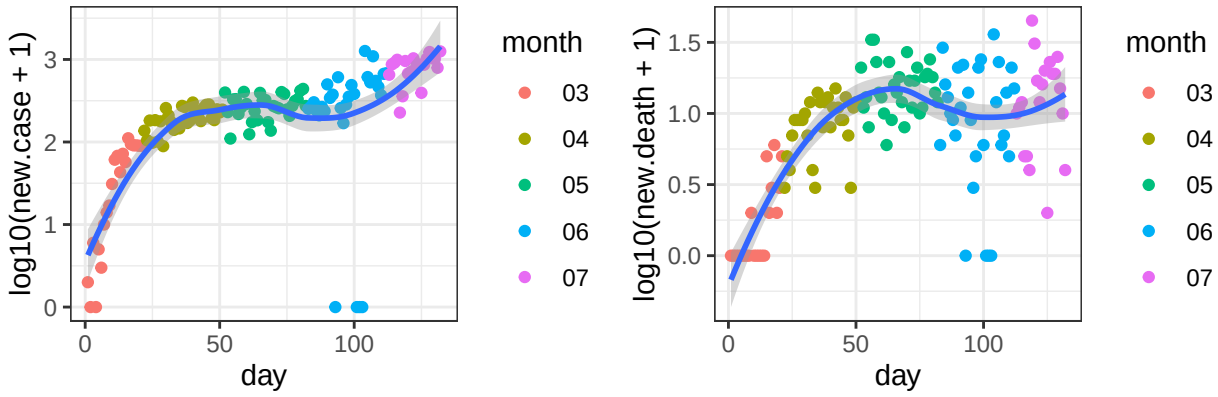

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06*

Washington



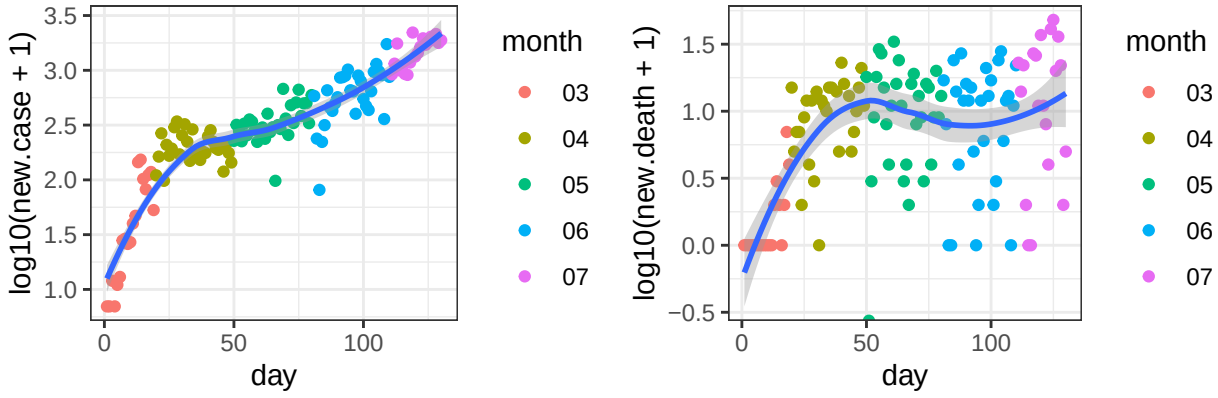

*data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21*

Mississippi



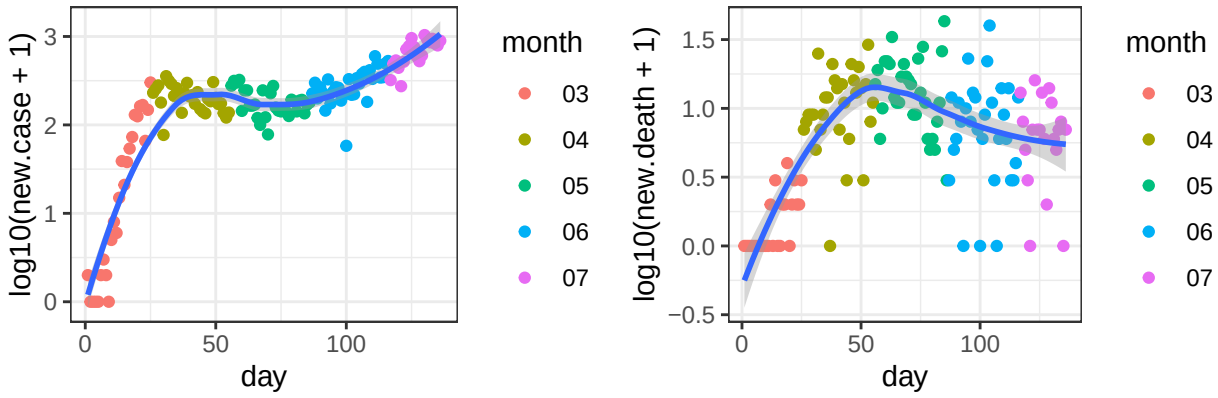

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11*

Alabama



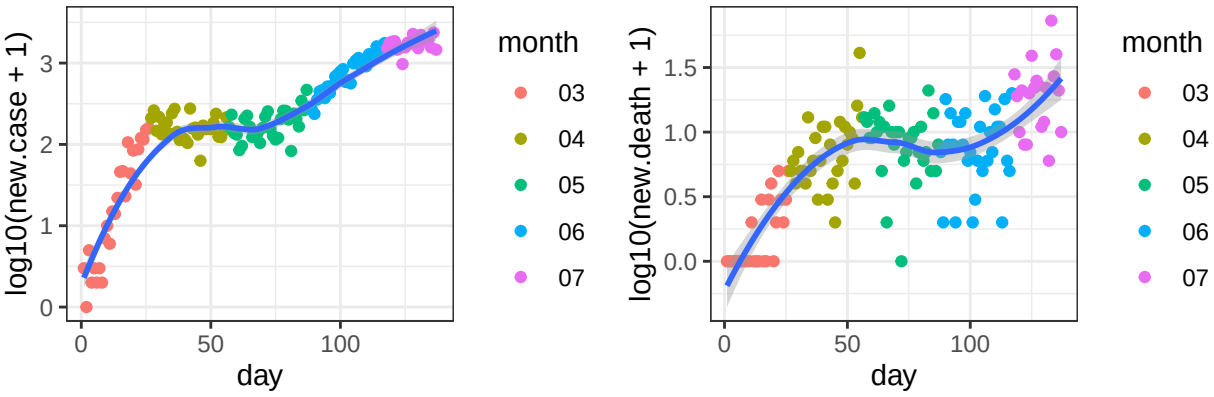

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13*

Missouri



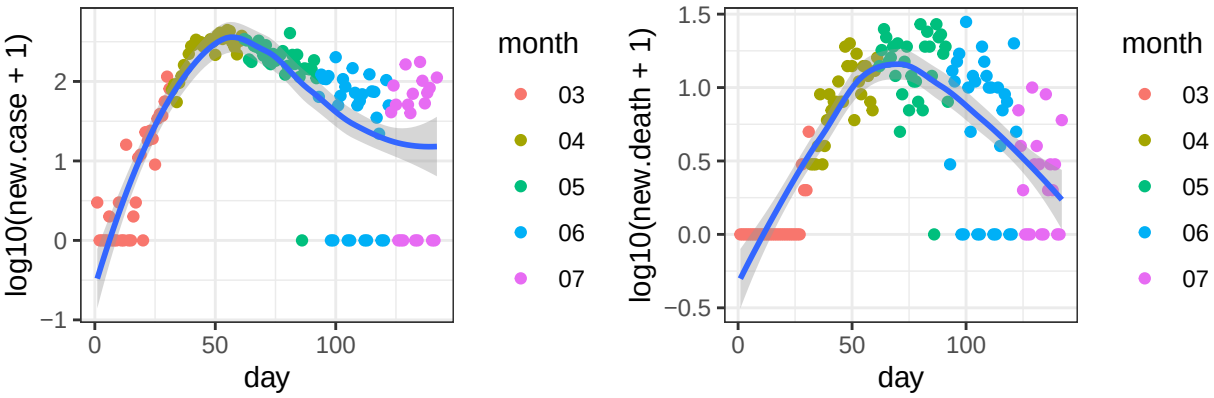

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07*

South Carolina



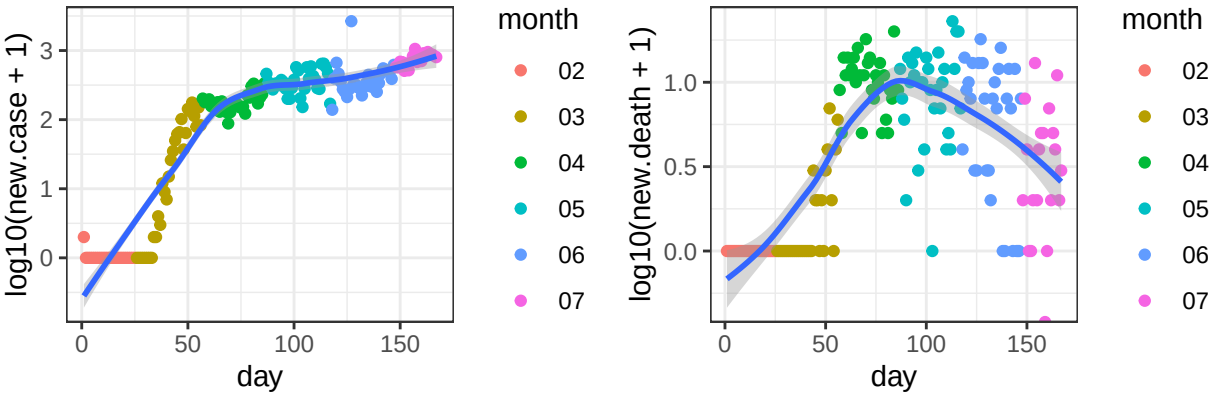

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06*

Rhode Island



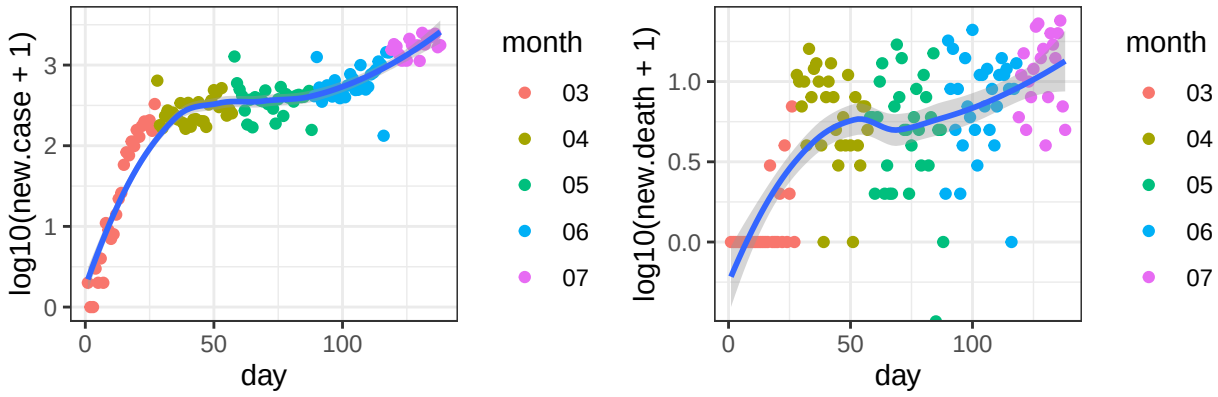

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01*

Wisconsin



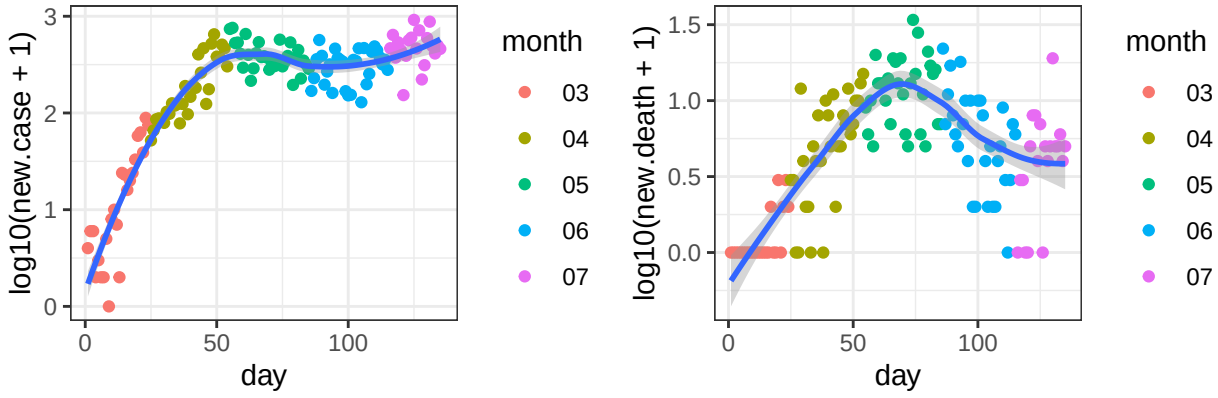

*data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05*

Tennessee



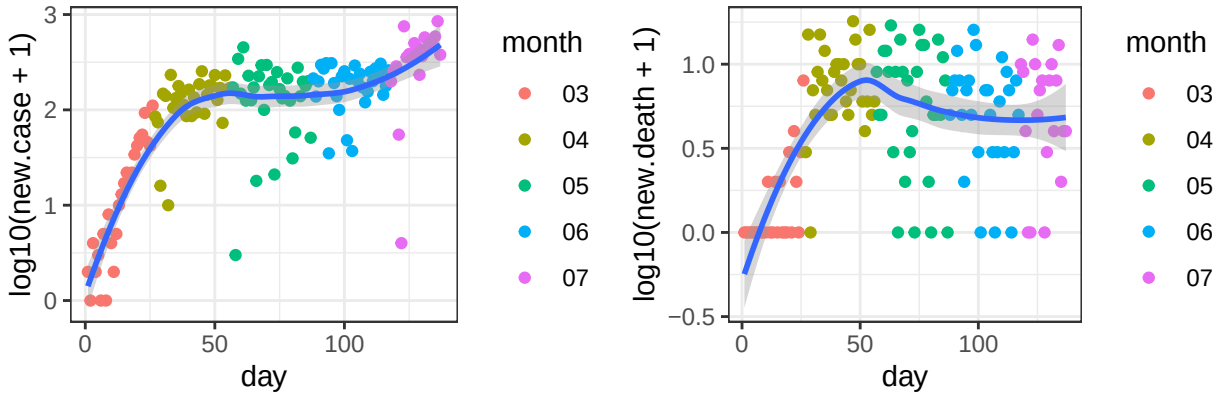

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05*

Iowa



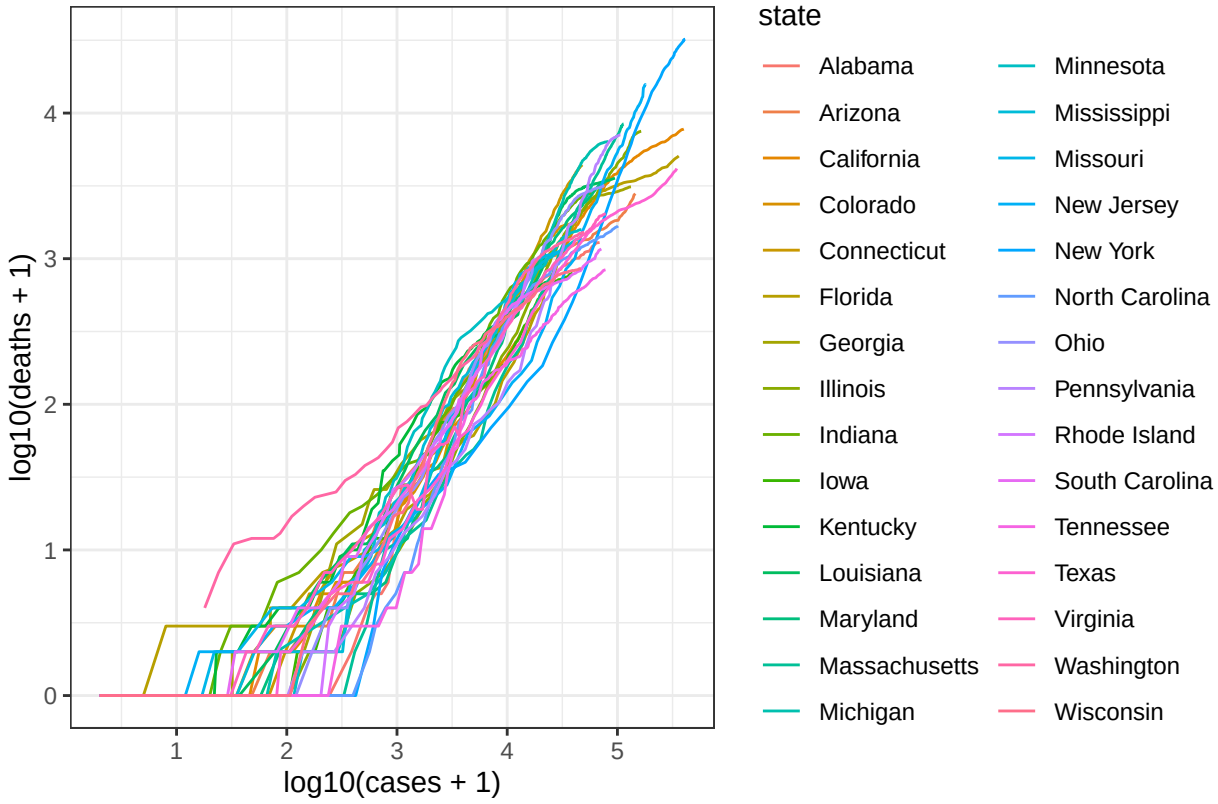

*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08*

Kentucky




*data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06*

Next I check the relation between the **cumulative** number of cases and deaths for these 10 states, starting on March




data source: https://github.com/nytimes/covid−19−data

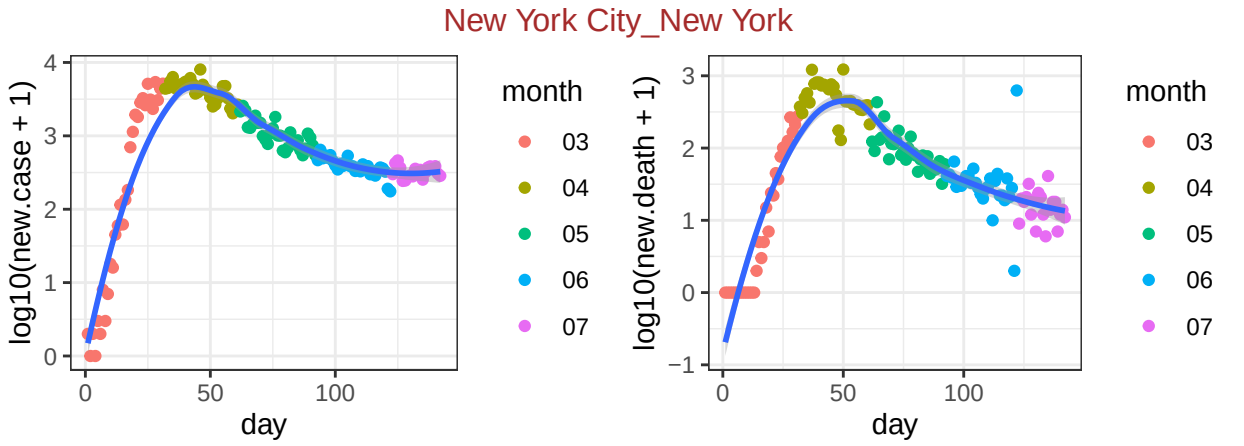
### county level data

First check the 50 counties with the largest number of deaths.

```
##              date              county               state  fips  cases deaths
## 352208 2020-07-20       New York City            New York    NA 226388  22882
## 350975 2020-07-20                Cook            Illinois 17031  99052   4777
## 350569 2020-07-20         Los Angeles          California  6037 159045   4104
## 351681 2020-07-20               Wayne            Michigan 26163  25251   2783
## 352207 2020-07-20              Nassau            New York 36059  42678   2705
## 352132 2020-07-20               Essex          New Jersey 34013  19376   2093
## 352227 2020-07-20             Suffolk            New York 36103  42496   2041
## 352127 2020-07-20              Bergen          New Jersey 34003  20302   2033
## 351592 2020-07-20           Middlesex       Massachusetts 25017  24958   1948
## 352643 2020-07-20        Philadelphia        Pennsylvania 42101  28592   1665
## 352235 2020-07-20         Westchester            New York 36119  35550   1572
## 352134 2020-07-20              Hudson          New Jersey 34017  19448   1498
## 350467 2020-07-20             Maricopa             Arizona  4013  96711   1485
## 350672 2020-07-20             Hartford         Connecticut  9003  12082   1406
## 350671 2020-07-20            Fairfield         Connecticut  9001  17112   1399
## 352137 2020-07-20           Middlesex          New Jersey 34023  17389   1397
## 352145 2020-07-20                Union          New Jersey 34039  16696   1344
## 350727 2020-07-20          Miami-Dade             Florida 12086  87034   1309
## 352141 2020-07-20              Passaic          New Jersey 34031  17330   1239
## 351588 2020-07-20               Essex       Massachusetts 25009  16757   1143
## 351661 2020-07-20             Oakland            Michigan 26125  13442   1117
## 350675 2020-07-20           New Haven         Connecticut  9009  12788   1092
```

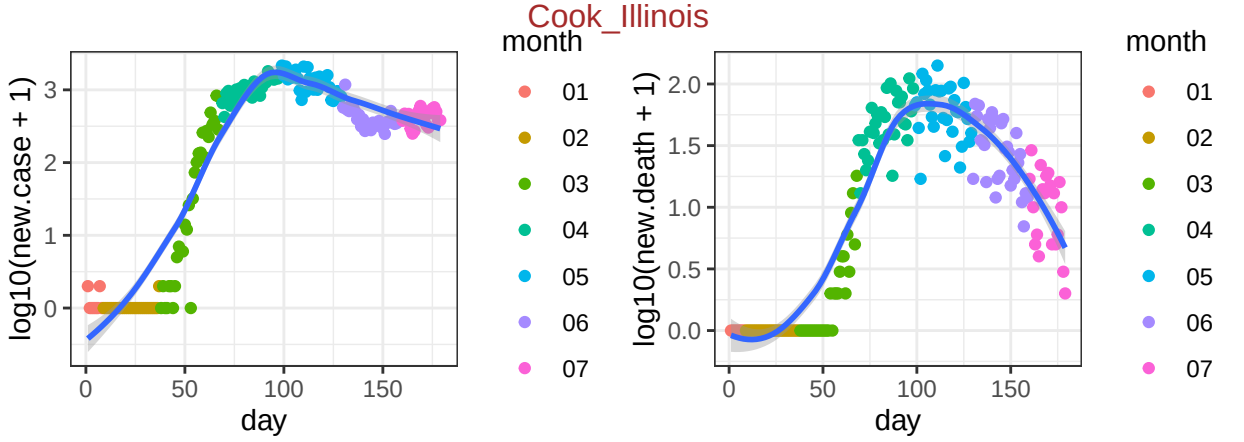
```
## 351596 2020-07-20              Suffolk       Massachusetts 25025  20621    1035
## 352140 2020-07-20               Ocean          New Jersey 34029  10029    1010
## 351598 2020-07-20           Worcester       Massachusetts 25027  12859     969
## 351594 2020-07-20             Norfolk       Massachusetts 25021   9737     968
## 351648 2020-07-20              Macomb            Michigan 26099   8658     939
## 352138 2020-07-20            Monmouth          New Jersey 34025   9773     849
## 352638 2020-07-20          Montgomery        Pennsylvania 42091   9182     837
## 352139 2020-07-20              Morris          New Jersey 34027   7124     826
## 351709 2020-07-20            Hennepin           Minnesota 27053  14835     802
## 352742 2020-07-20          Providence        Rhode Island 44007  13674     796
## 351574 2020-07-20          Montgomery            Maryland 24031  16471     772
## 351111 2020-07-20              Marion             Indiana 18097  12976     752
## 351575 2020-07-20      Prince George's            Maryland 24033  21150     724
## 352615 2020-07-20            Delaware        Pennsylvania 42045   7950     718
## 351595 2020-07-20            Plymouth       Massachusetts 25023   8894     696
## 350734 2020-07-20          Palm Beach             Florida 12099  26424     685
## 351590 2020-07-20             Hampden       Massachusetts 25013   7141     678
## 353396 2020-07-20                King          Washington 53033  13341     657
## 352193 2020-07-20                Erie            New York 36029   8046     613
## 351586 2020-07-20             Bristol       Massachusetts 25005   8719     610
## 351954 2020-07-20           St. Louis            Missouri 29189   9114     610
## 352136 2020-07-20              Mercer          New Jersey 34021   7871     608
## 350582 2020-07-20           Riverside          California  6065  30340     588
## 350684 2020-07-20 District of Columbia District of Columbia 11001  11339     579
## 352601 2020-07-20               Bucks        Pennsylvania 42017   6387     573
## 352129 2020-07-20              Camden          New Jersey 34007   7901     560
## 352143 2020-07-20            Somerset          New Jersey 34035   5150     553
## 351512 2020-07-20             Orleans           Louisiana 22071   9379     546
```

For these 50 counties, I check the number of new cases and the number of new deaths.



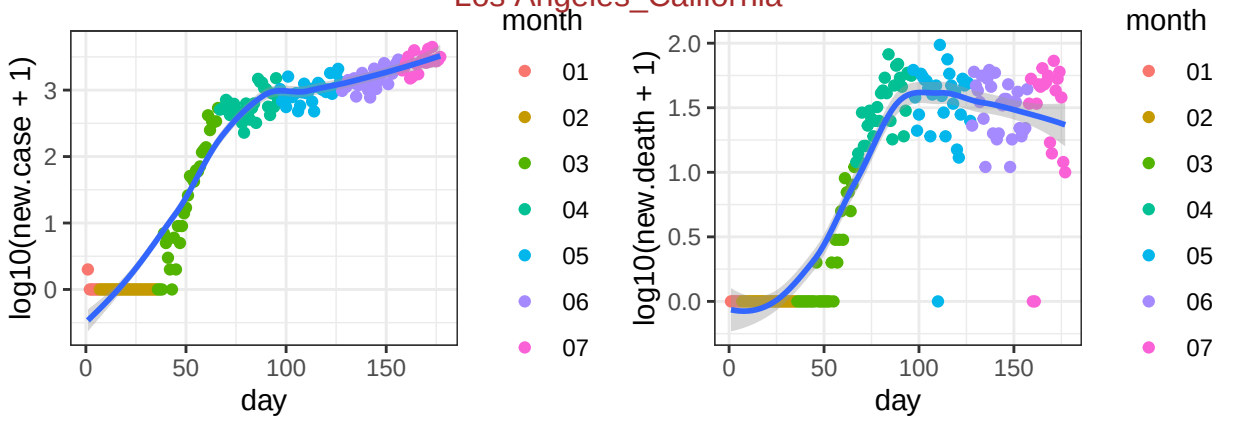

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



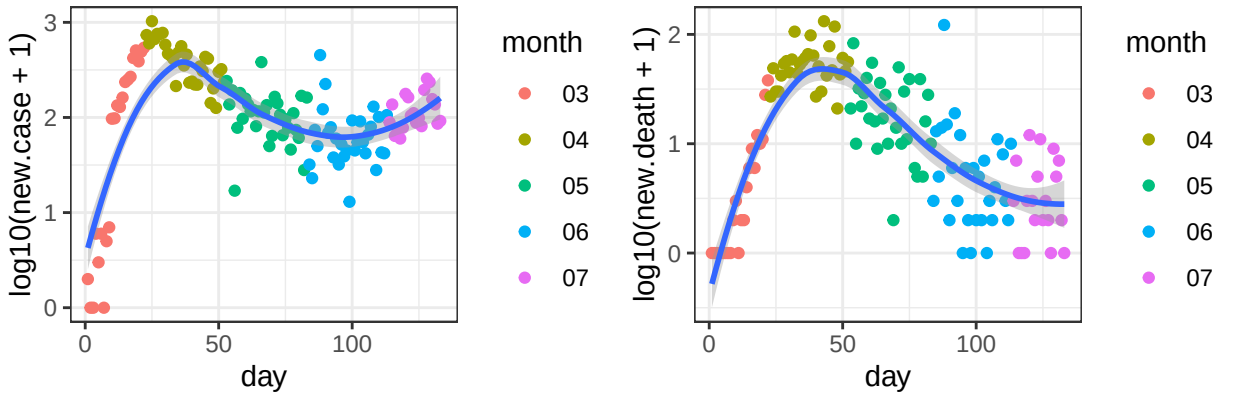

data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Los Angeles_California



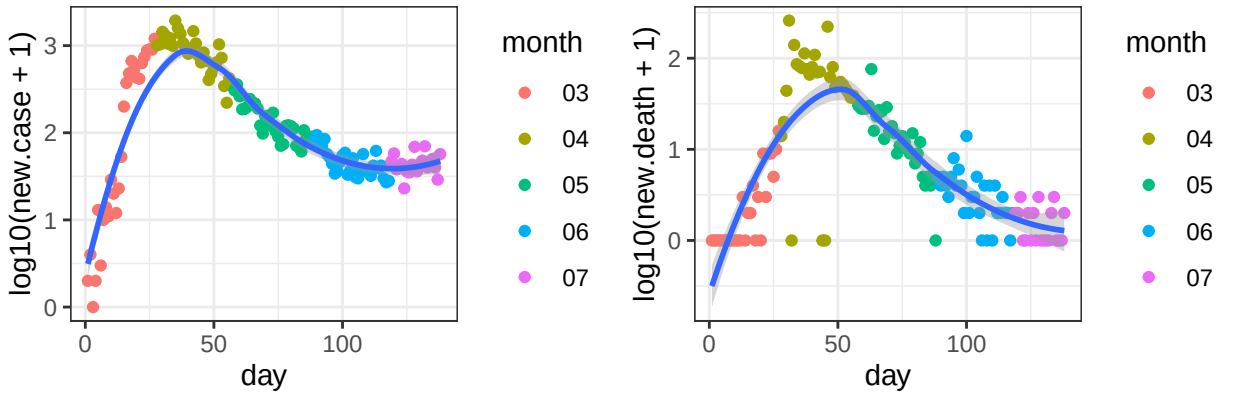

data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Wayne_Michigan



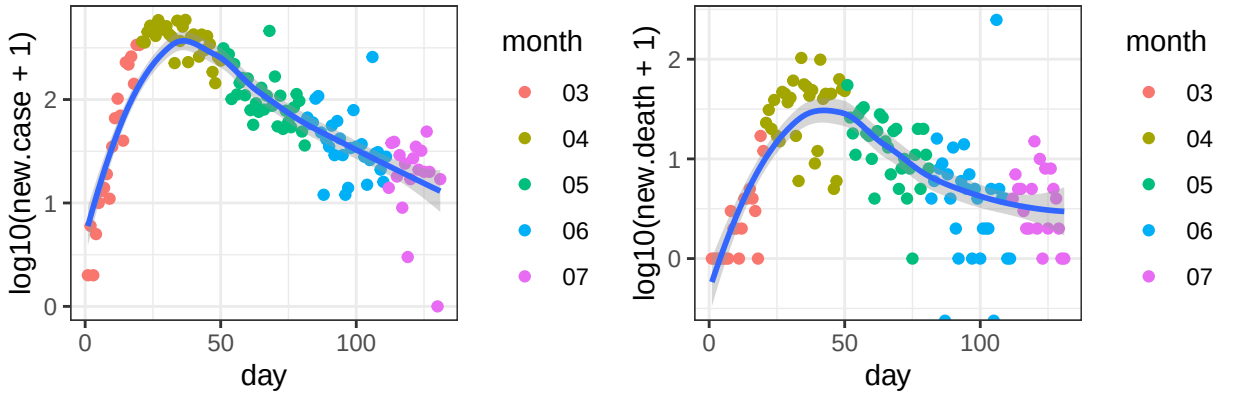

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Nassau_New York



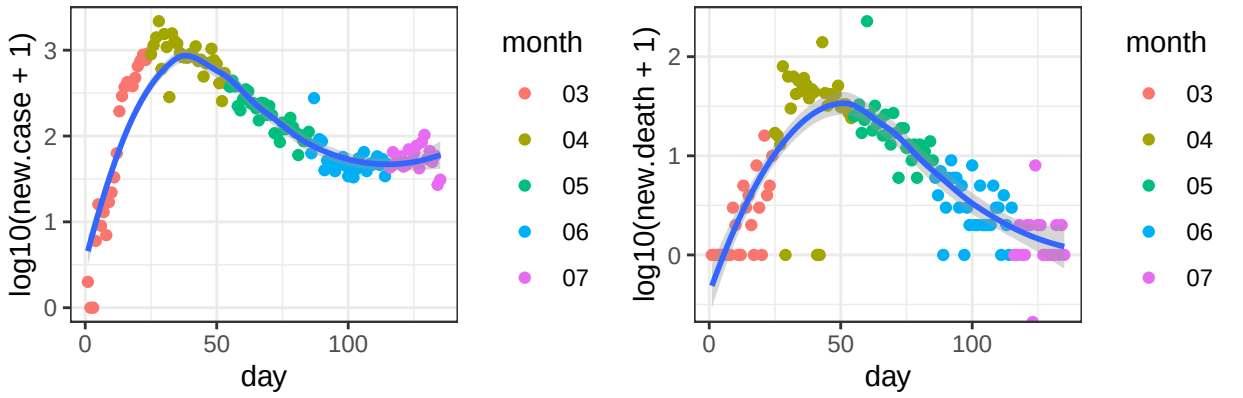

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



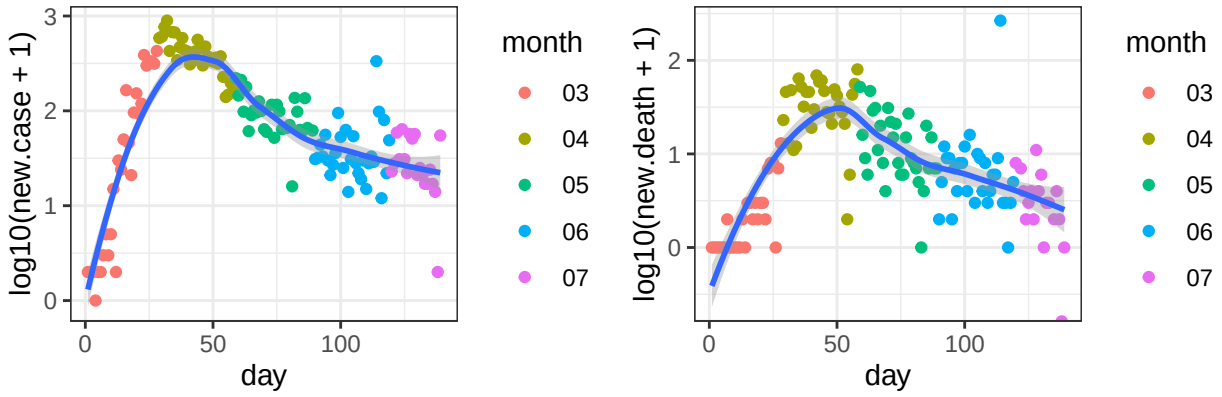

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Suffolk_New York



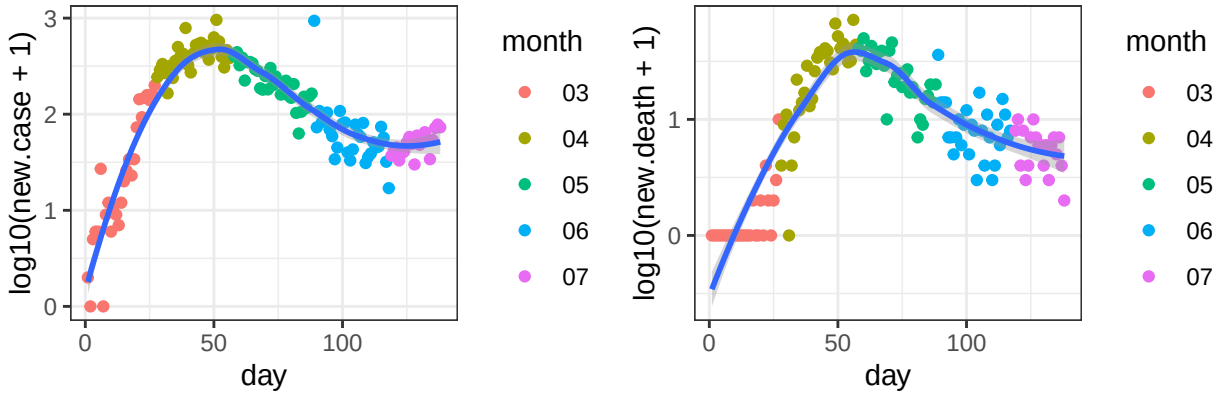

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Bergen_New Jersey



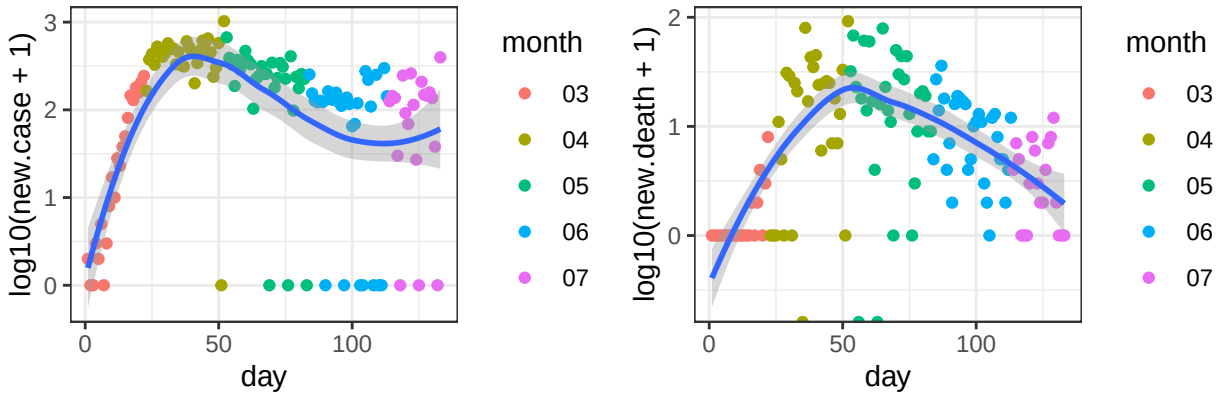

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Middlesex_Massachusetts



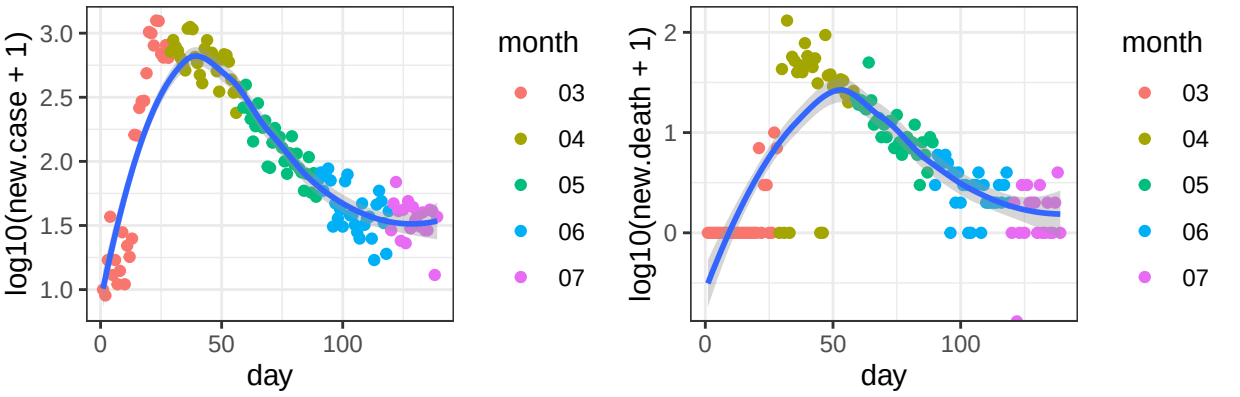

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Philadelphia_Pennsylvania



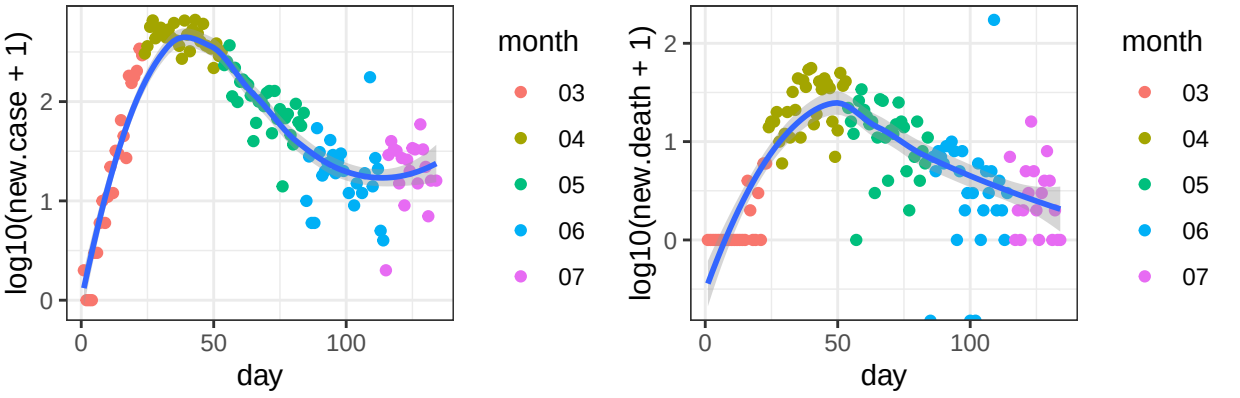

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Westchester_New York



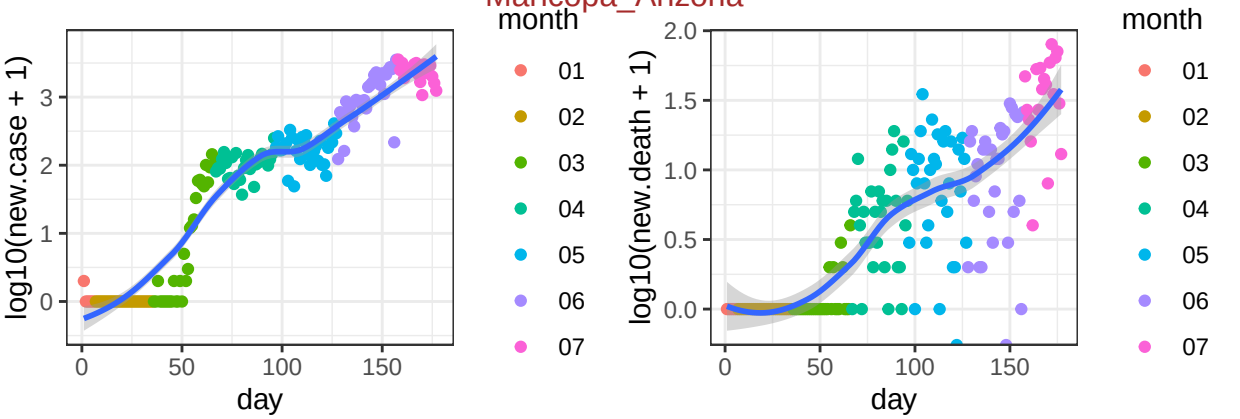

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Hudson_New Jersey



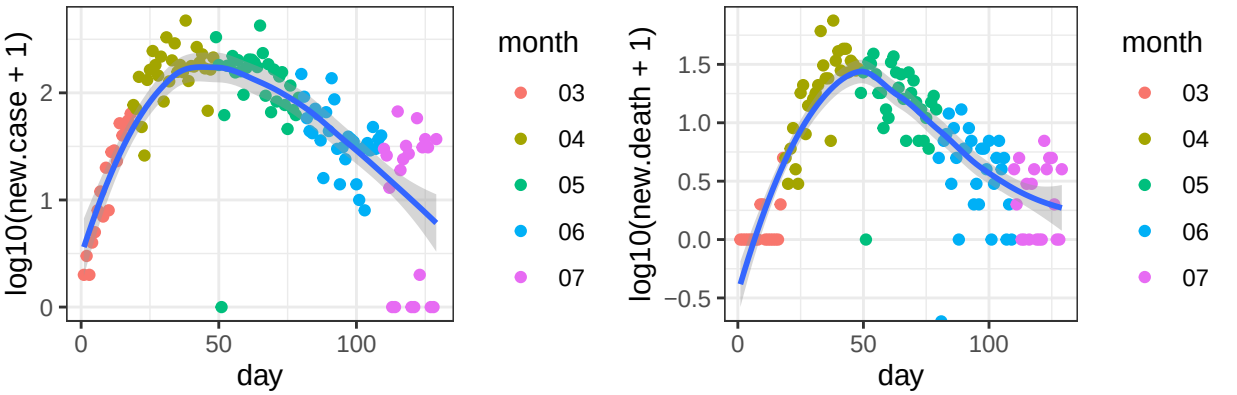

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maricopa_Arizona



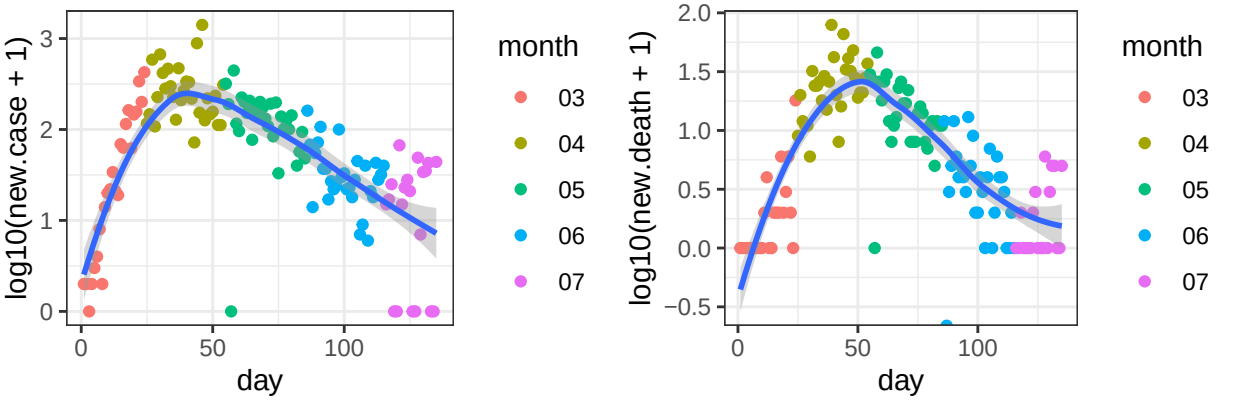

data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Hartford_Connecticut



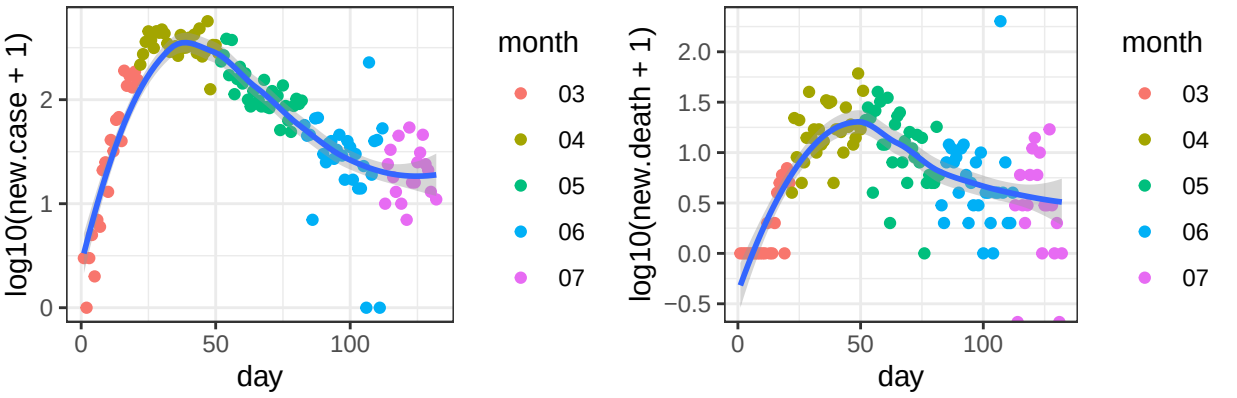

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Fairfield_Connecticut



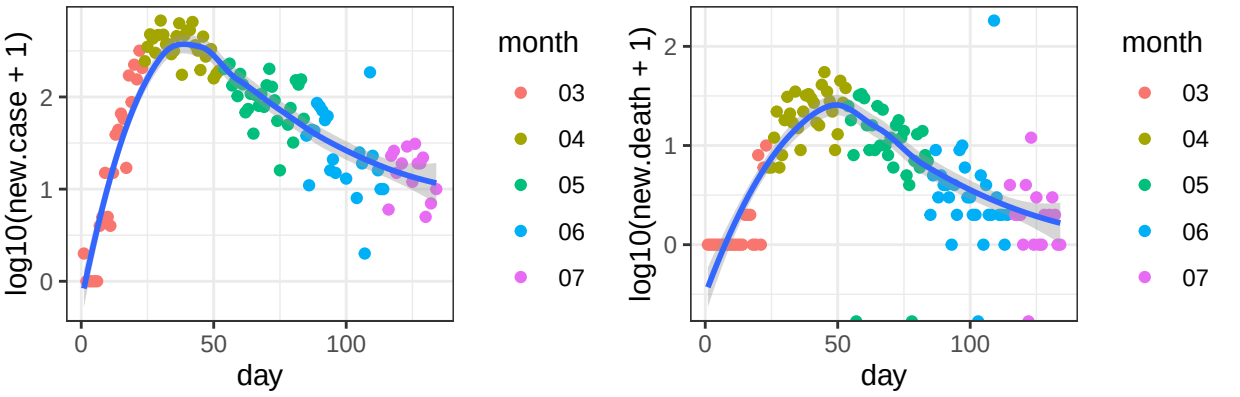

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Middlesex_New Jersey



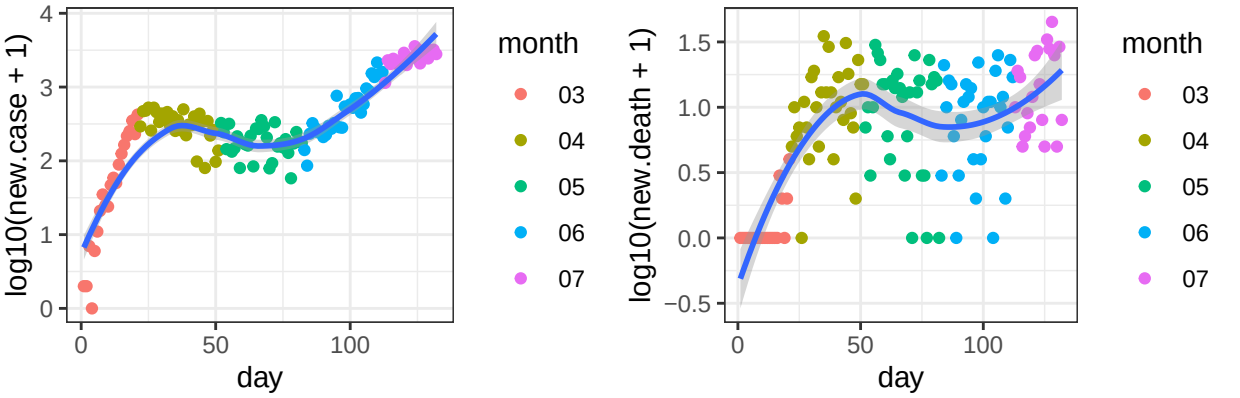

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Union_New Jersey



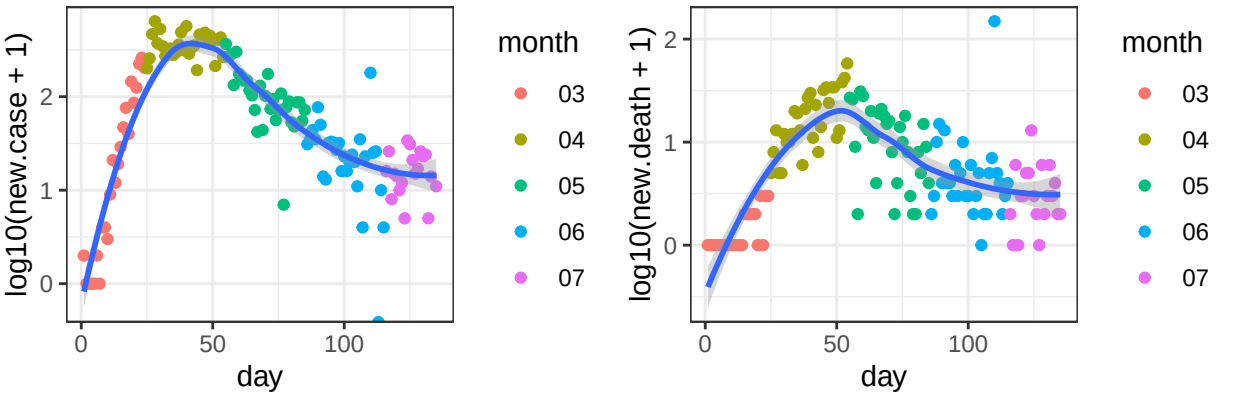

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Miami−Dade_Florida



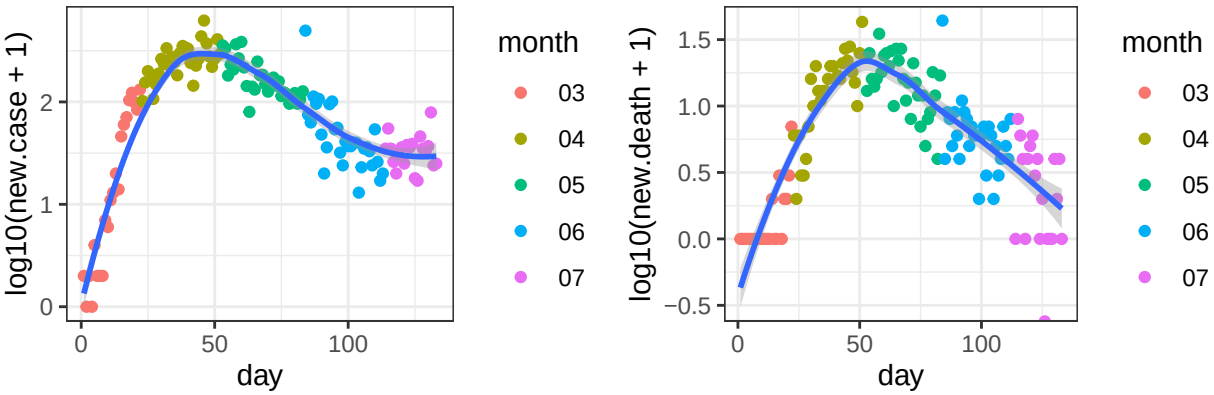

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Passaic_New Jersey



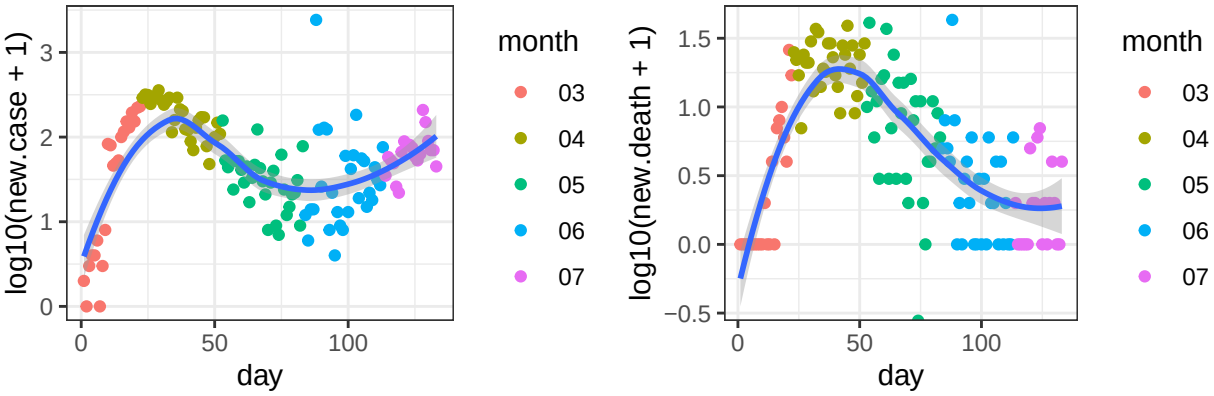

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Essex_Massachusetts



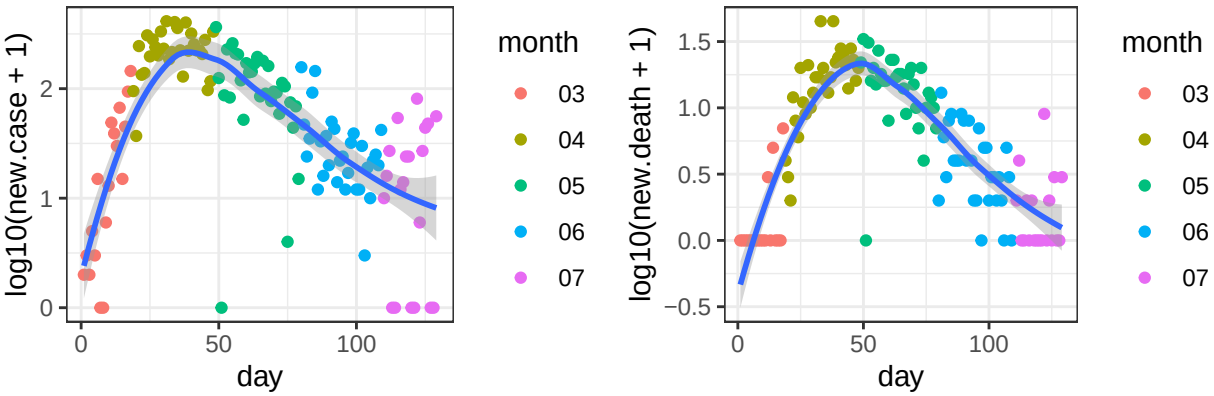

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Oakland_Michigan



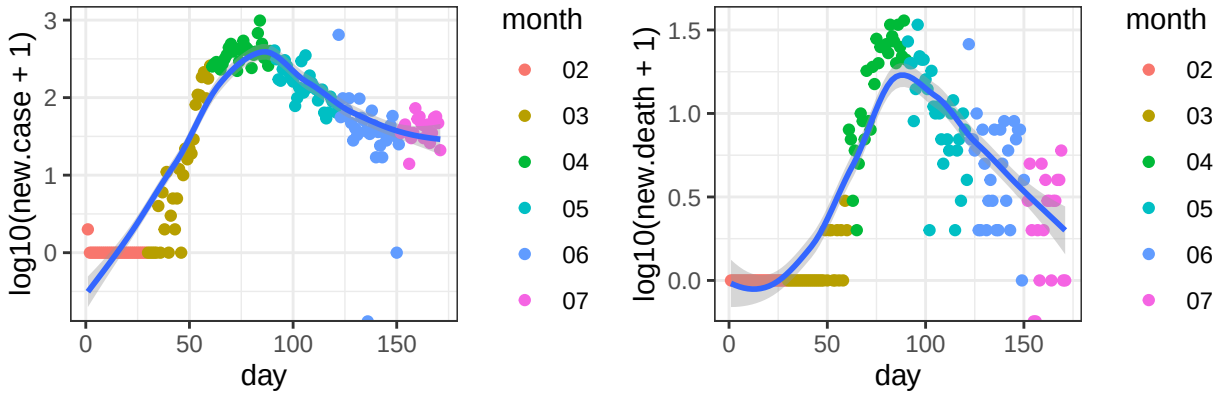

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

New Haven_Connecticut



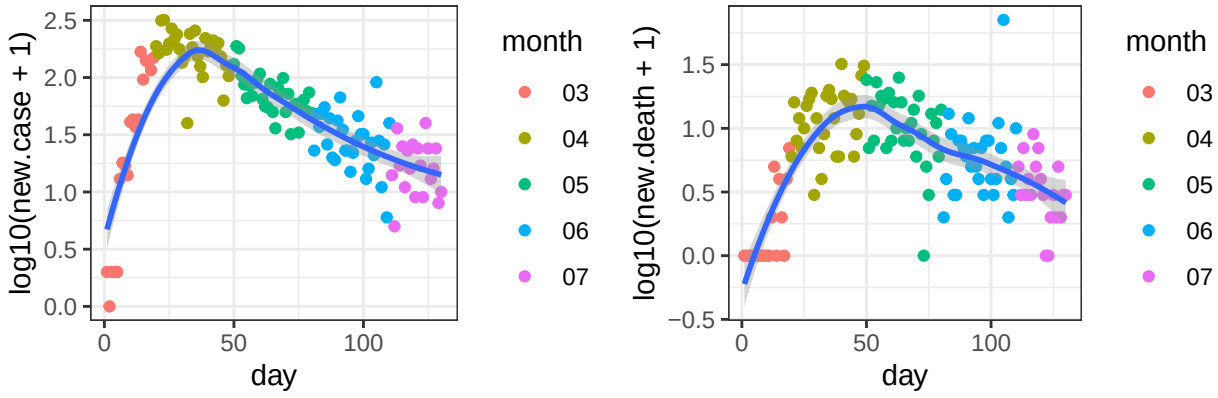

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Suffolk_Massachusetts



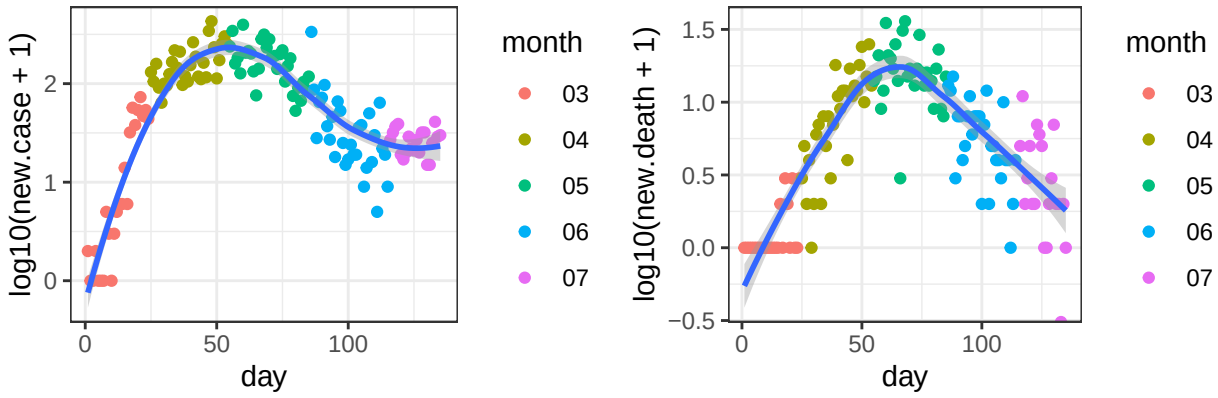

data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Ocean_New Jersey



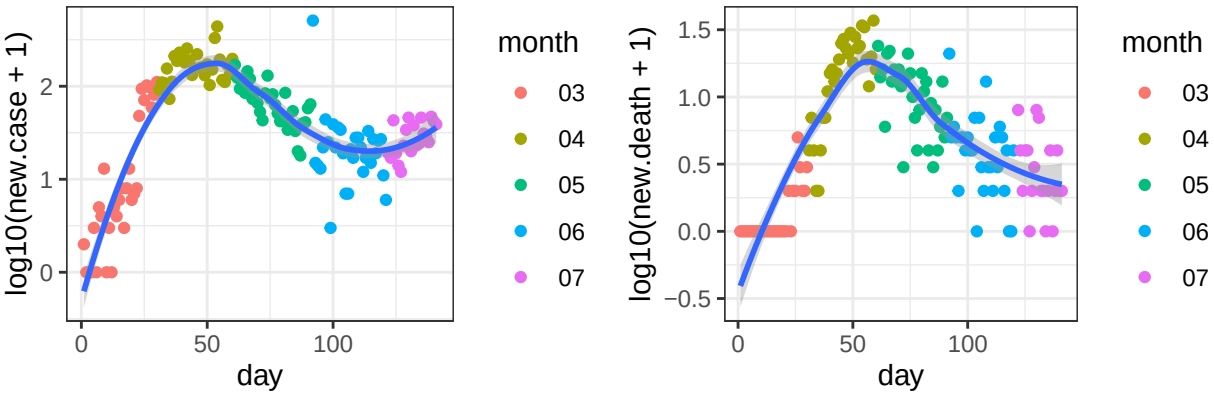

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Worcester_Massachusetts



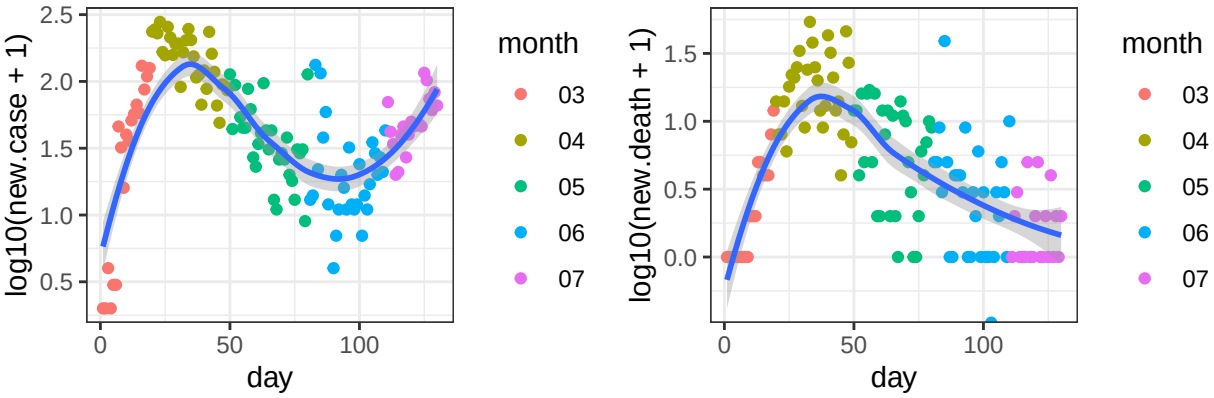

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Norfolk_Massachusetts



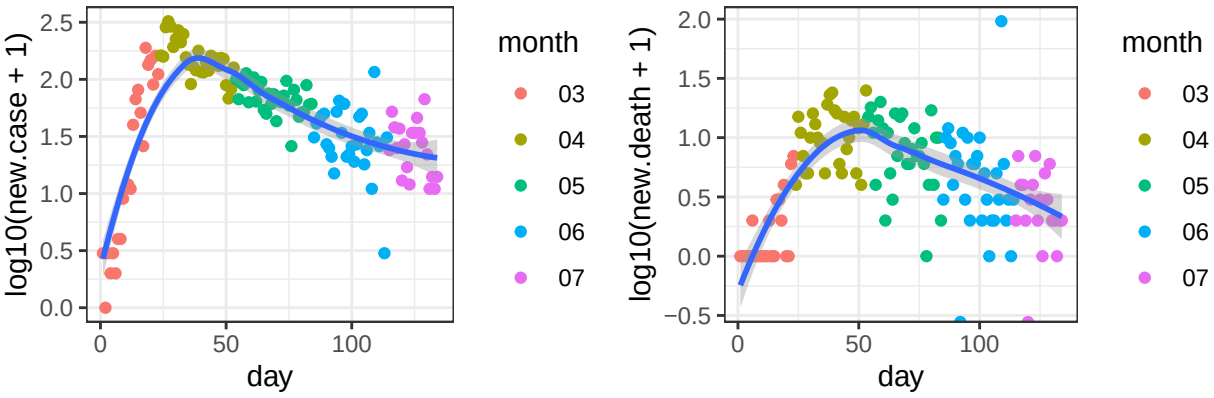

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Macomb_Michigan



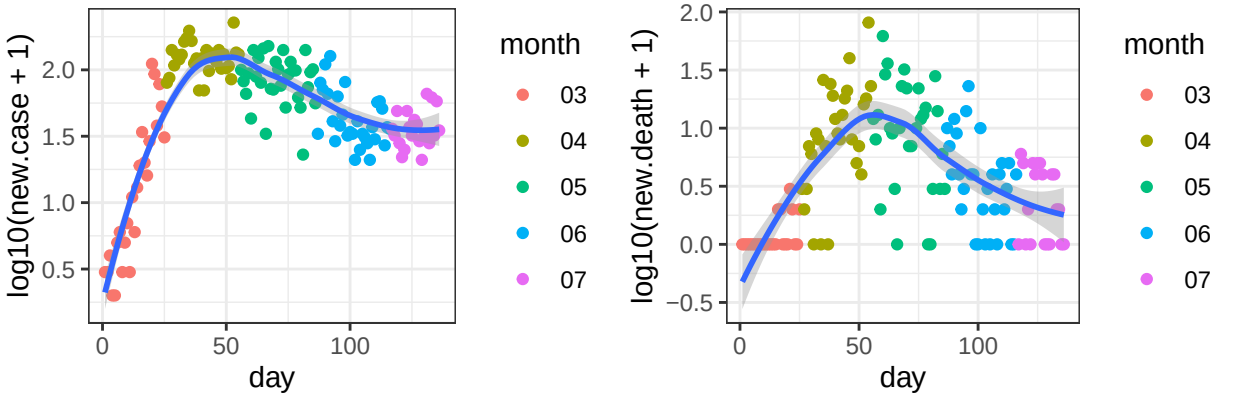

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Monmouth_New Jersey



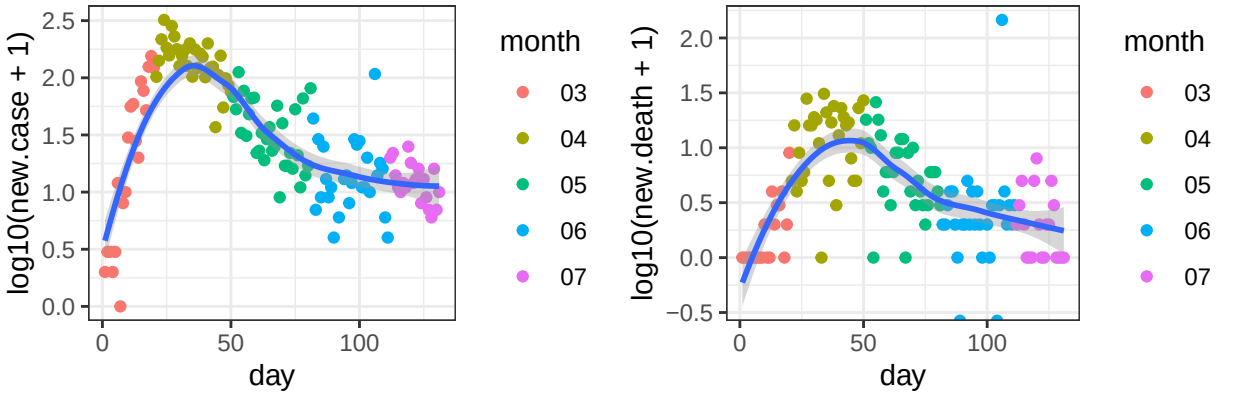

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Montgomery_Pennsylvania



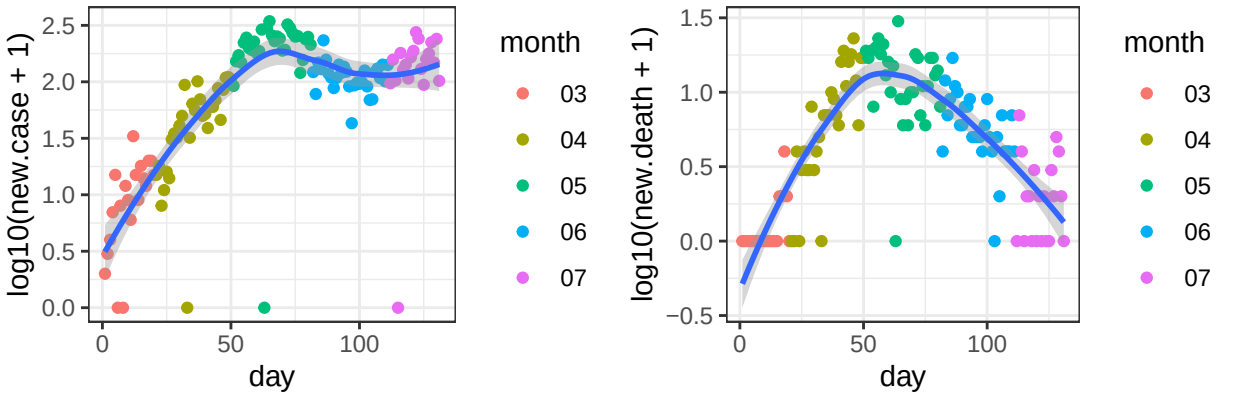

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Morris_New Jersey



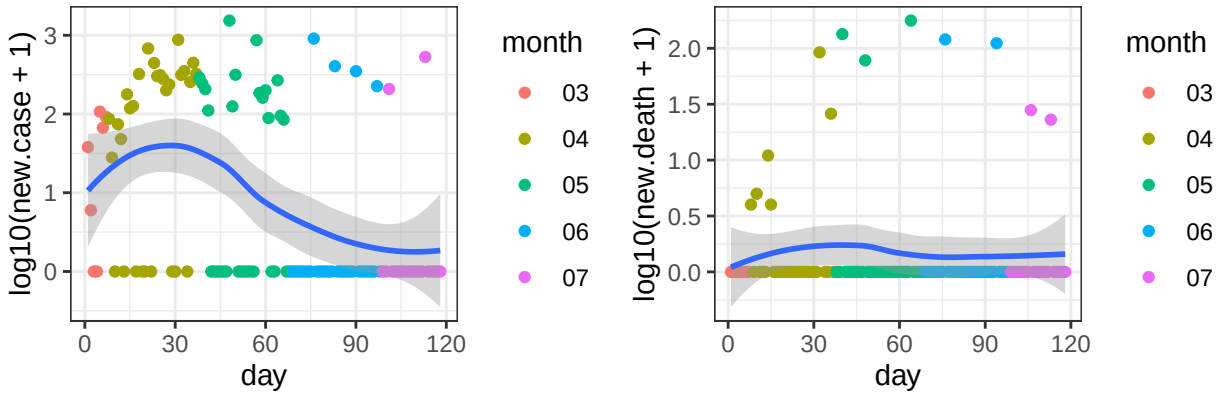

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Hennepin_Minnesota



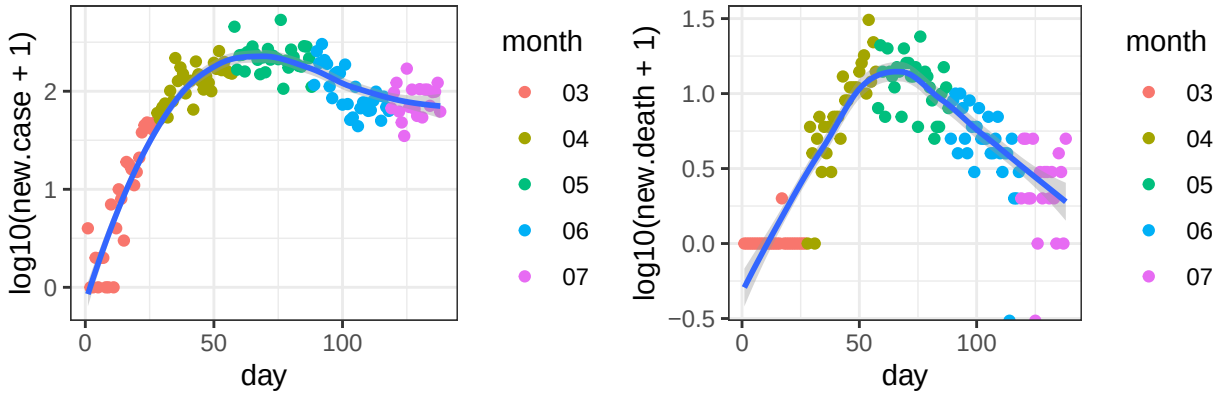

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Providence_Rhode Island



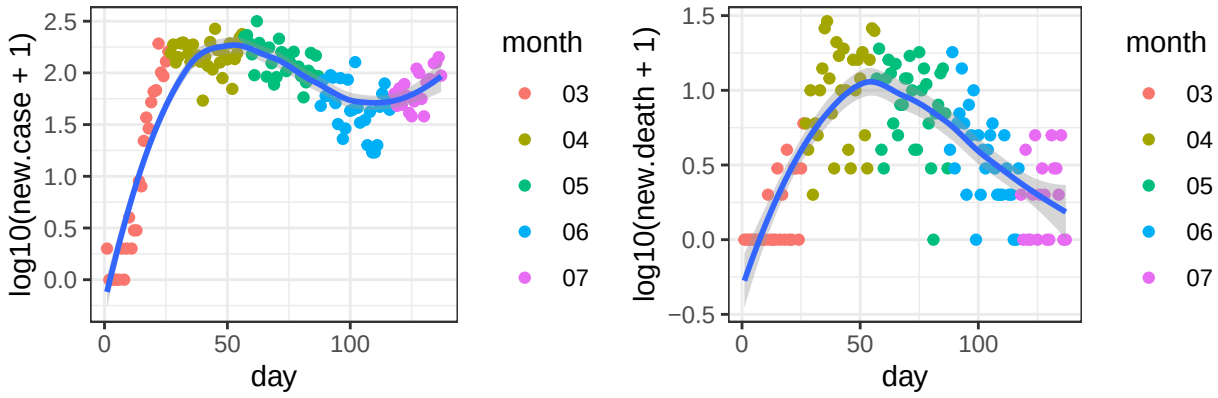

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Montgomery_Maryland



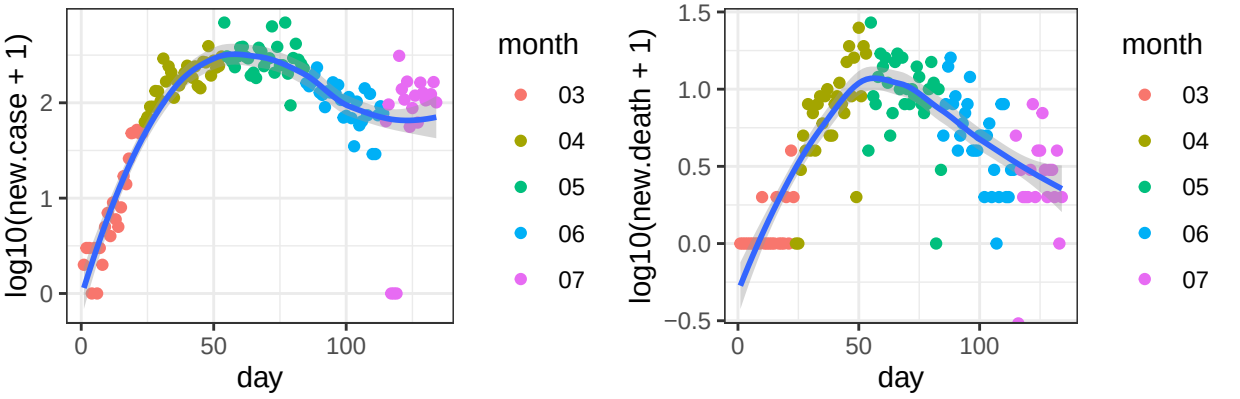

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Marion_Indiana



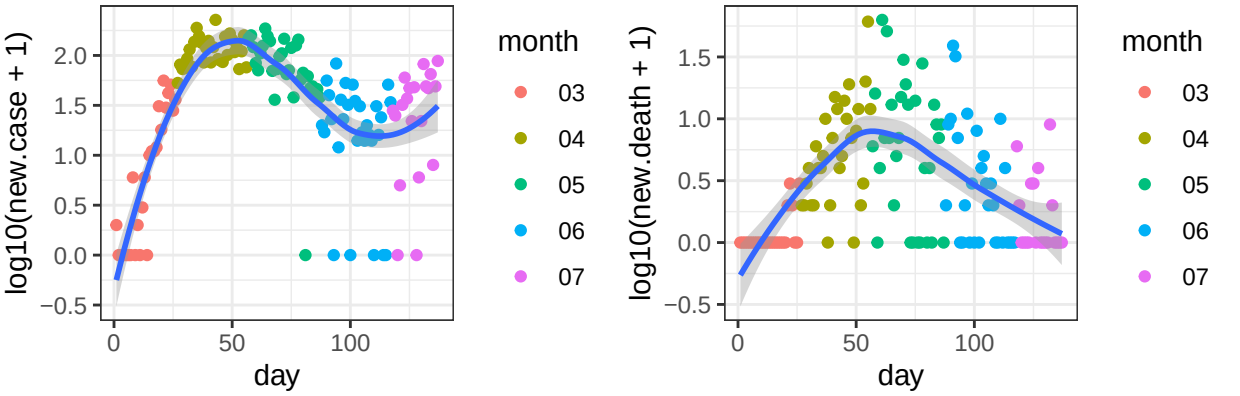

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Prince George's_Maryland



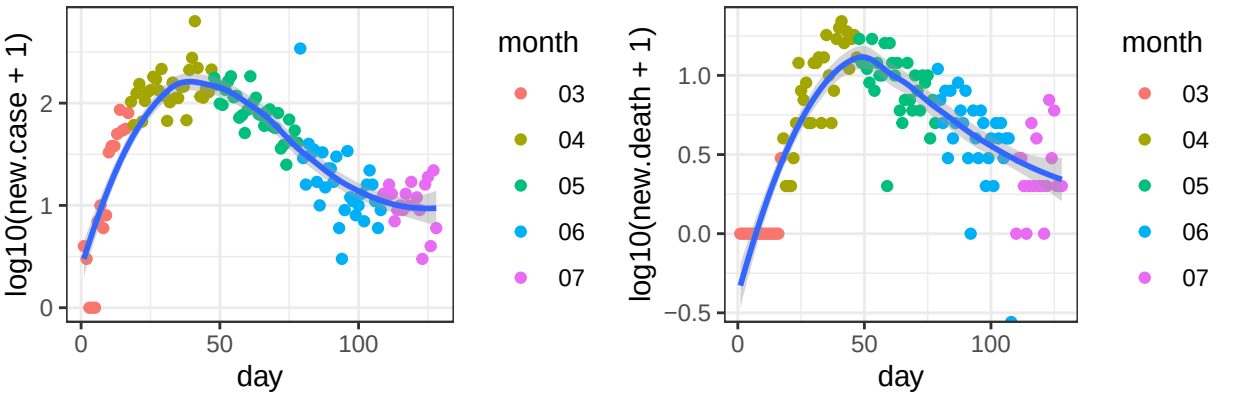

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Delaware_Pennsylvania



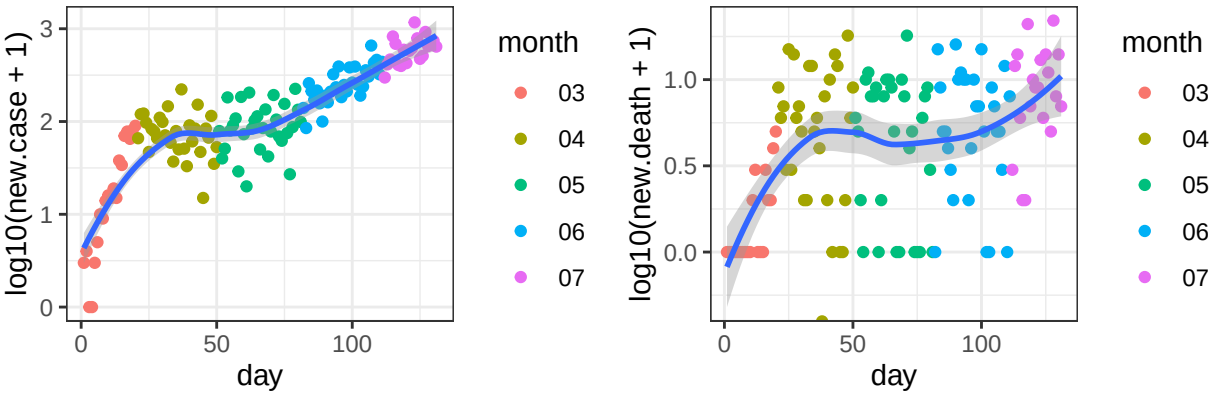

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Plymouth_Massachusetts



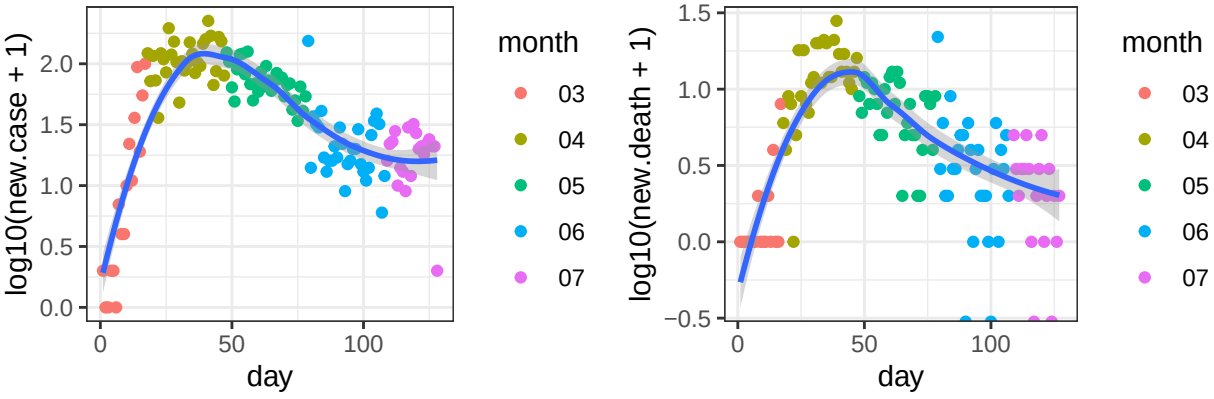

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Palm Beach_Florida



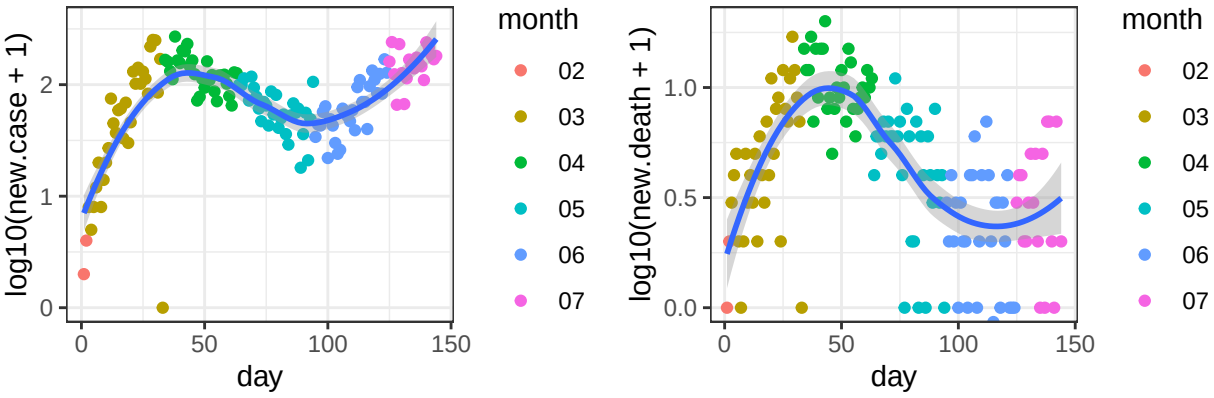

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Hampden_Massachusetts



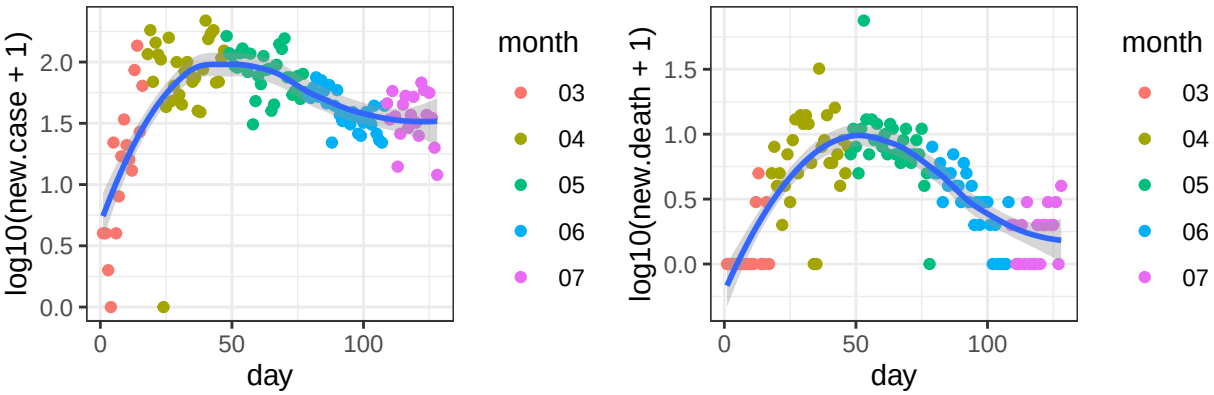

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

King_Washington



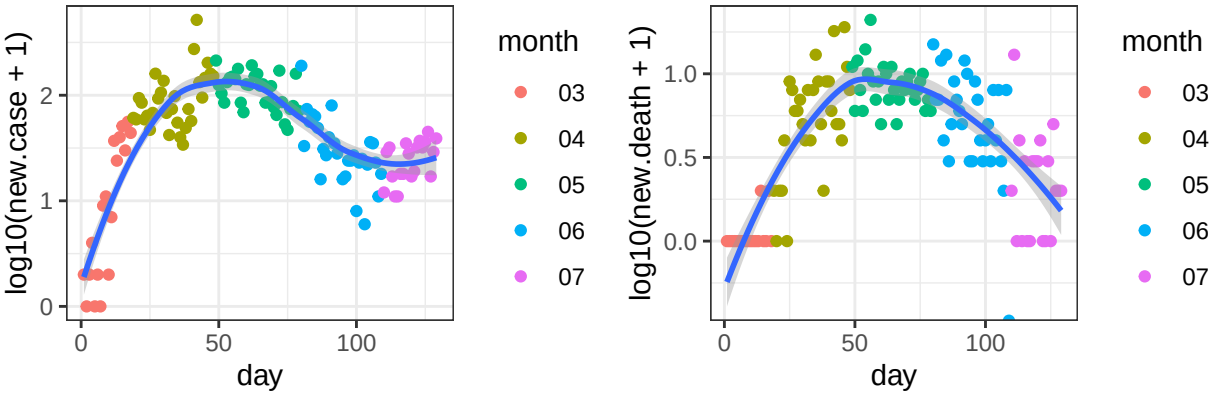

data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

Erie_New York



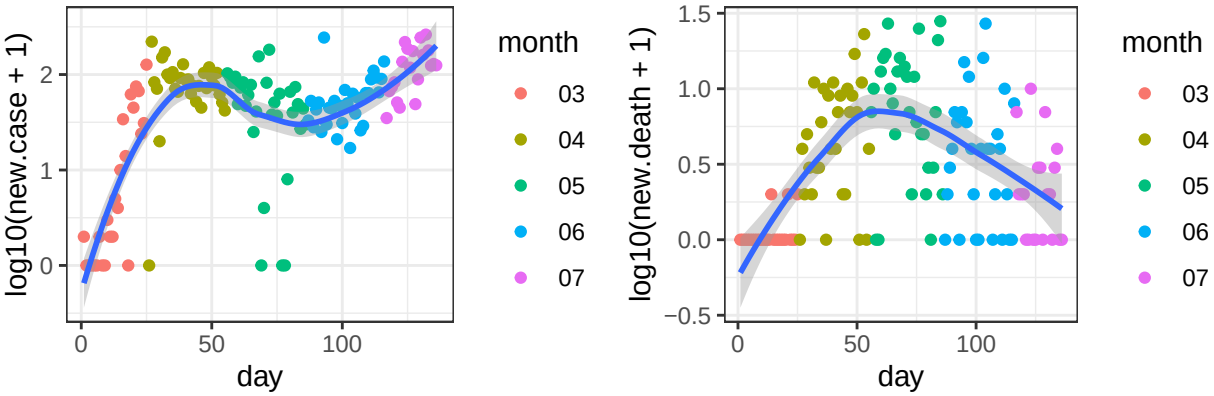

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Bristol_Massachusetts



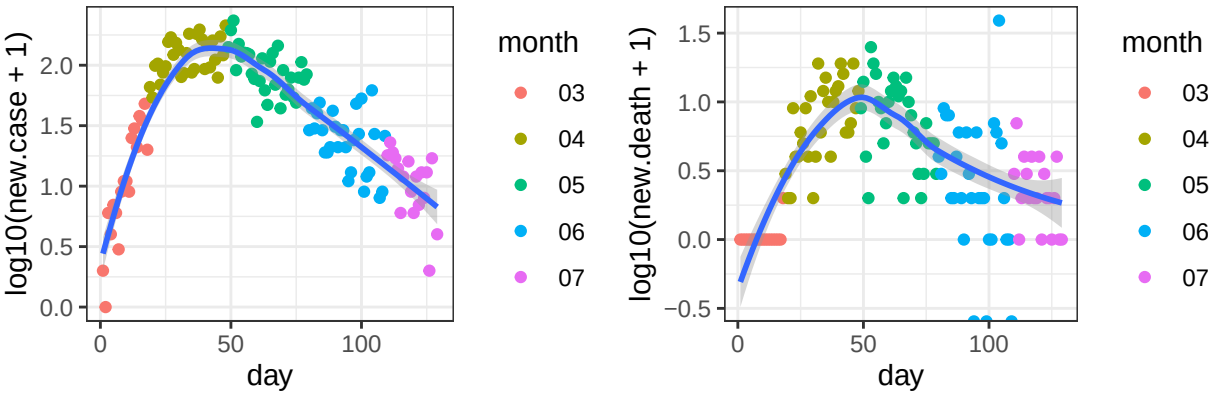

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

St. Louis_Missouri



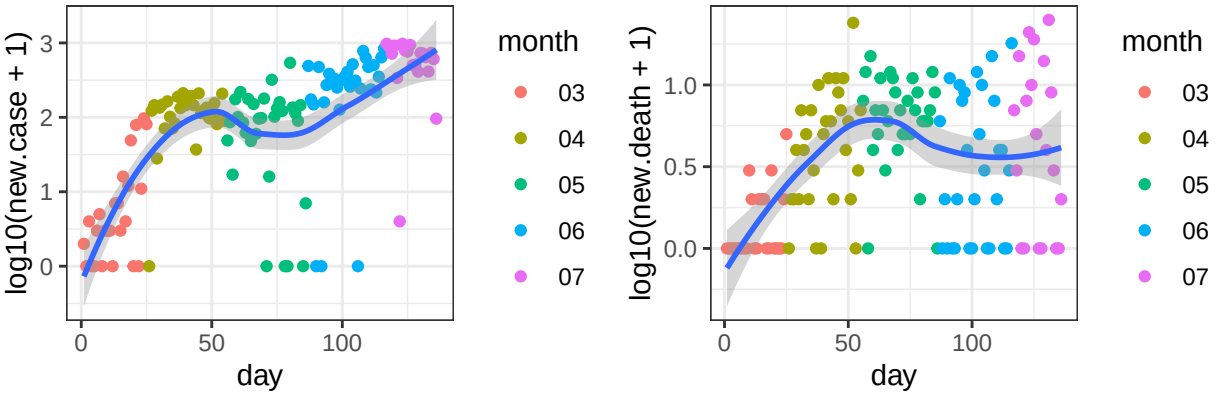

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Mercer_New Jersey



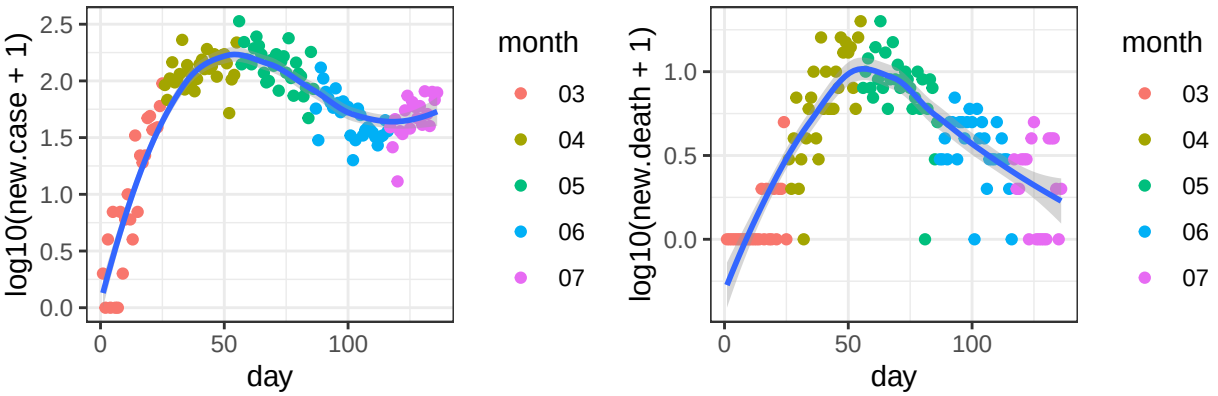

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Riverside_California



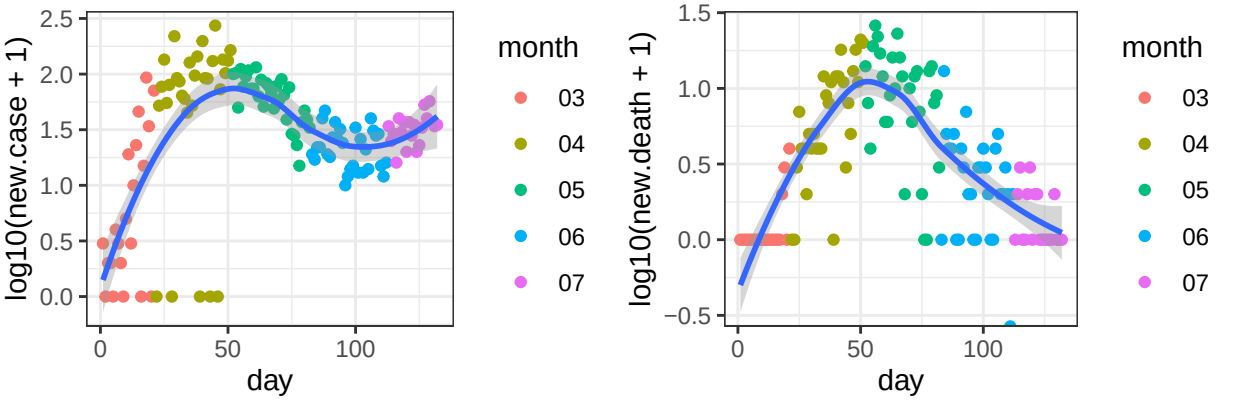

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

District of Columbia_District of Columbia



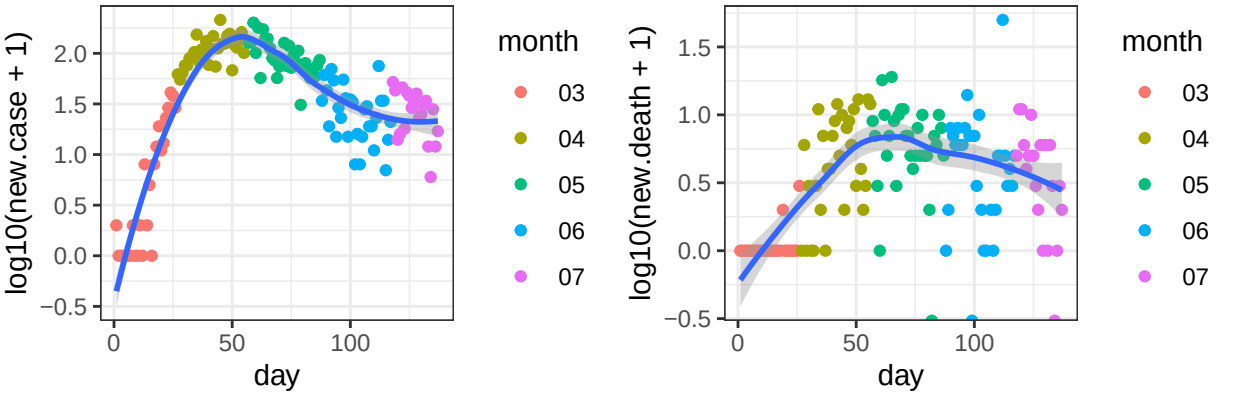

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Bucks_Pennsylvania



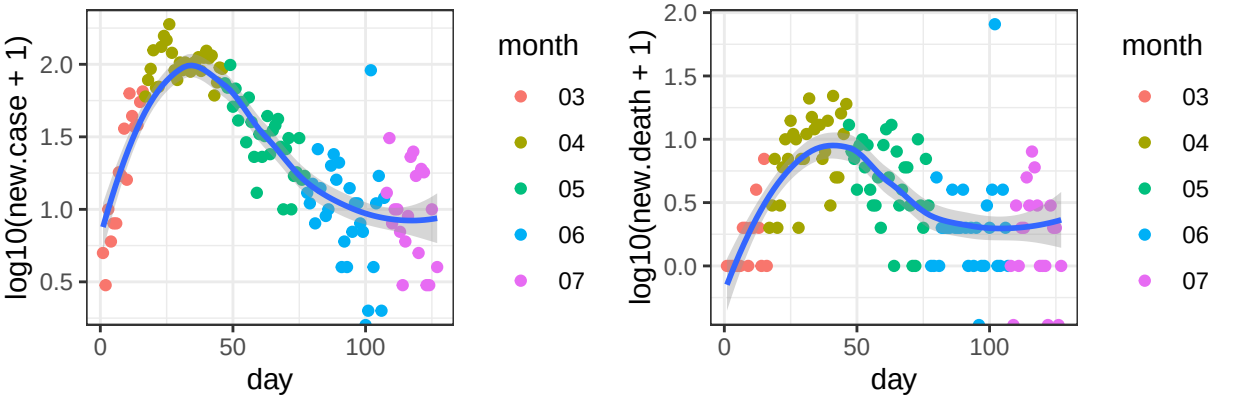

data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Camden_New Jersey

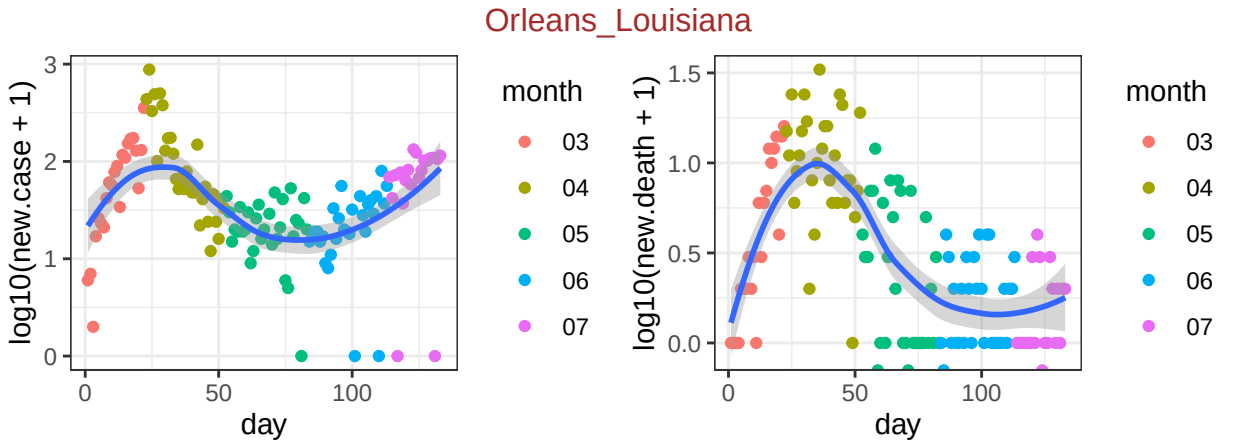



data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Somerset_New Jersey




data source: https://github.com/nytimes/covid−19−data, day 1 is 03−16

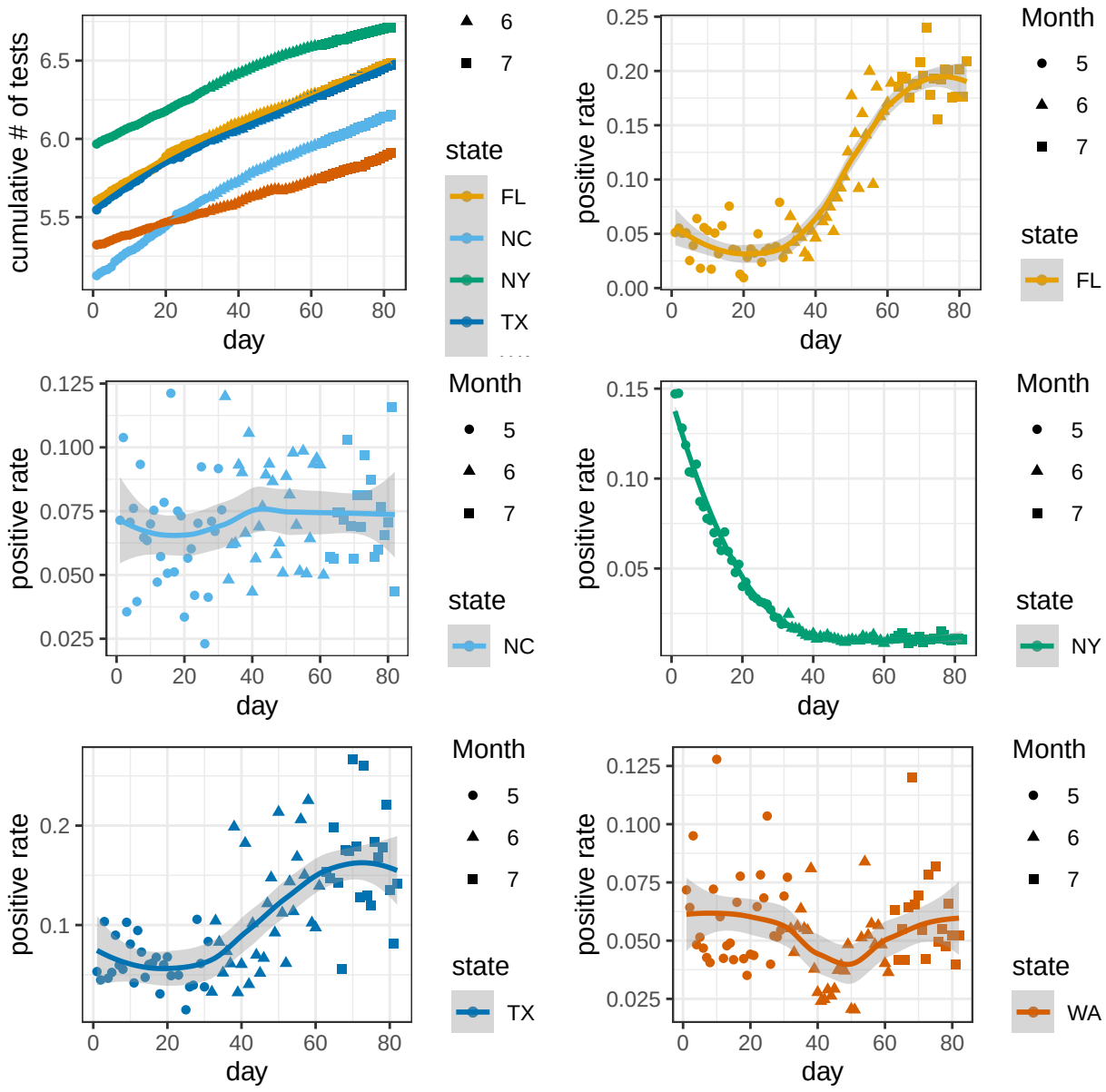



data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

## COVID Trackng

The positive rates of testing can be an indicator on how much the COVID-19 has spread. However, they can be much more noisy data since the negative testing resutls are often not reported and the tests are almost surely taken on a non-representative random sample of the population. The COVID traking project proides a grade per state: "If you are calculating positive rates, it should only be with states that have an A grade. And be careful going back in time because almost all the states have changed their level of reporting at different times." (https://covidtracking.com/about-tracker/). The data are also availalbe for both counties and states, here I only look at state level data.

The grades of the states may change over timea and I strongly recommend checking their webiste before puting serious interpretation on the following plot.




*github.com/COVID19Tracking/, positive rate on 0720: 0.21(FL) 0.04(NC) 0.01(NY) 0.14(TX) 0.05(WA)*

## Session information

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
```

```
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] httr_1.4.1    ggpubr_0.2.5  magrittr_1.5  ggplot2_3.3.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3       pillar_1.4.3     compiler_3.6.2   tools_3.6.2
##  [5] digest_0.6.23    lattice_0.20-38  nlme_3.1-144     evaluate_0.14
##  [9] lifecycle_0.2.0  tibble_3.0.1     gtable_0.3.0     mgcv_1.8-31
## [13] pkgconfig_2.0.3  rlang_0.4.6      Matrix_1.2-18    yaml_2.2.1
## [17] xfun_0.12        gridExtra_2.3    withr_2.1.2      stringr_1.4.0
## [21] dplyr_0.8.4      knitr_1.28       vctrs_0.3.0      cowplot_1.0.0
## [25] grid_3.6.2       tidyselect_1.0.0 glue_1.3.1       R6_2.4.1
## [29] rmarkdown_2.1    purrr_0.3.3      farver_2.0.3     splines_3.6.2
## [33] scales_1.1.0     ellipsis_0.3.0   htmltools_0.4.0  assertthat_0.2.1
## [37] colorspace_1.4-1 ggsignif_0.6.0   labeling_0.3     stringi_1.4.5
## [41] munsell_0.5.0    crayon_1.3.4
```