Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-06-15

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

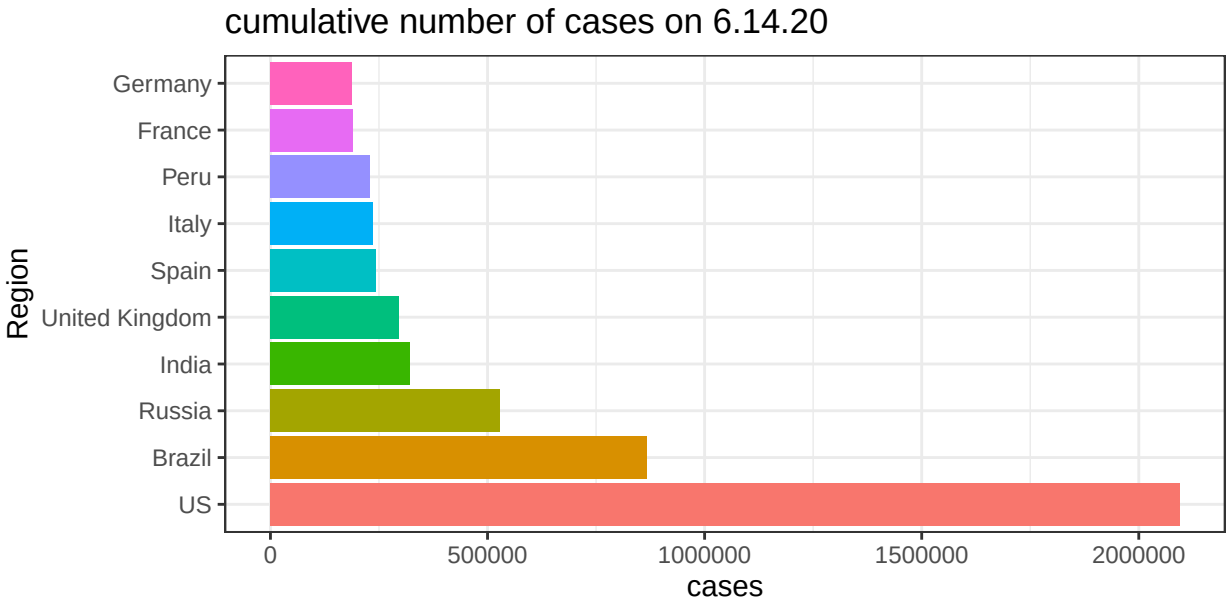
JHU

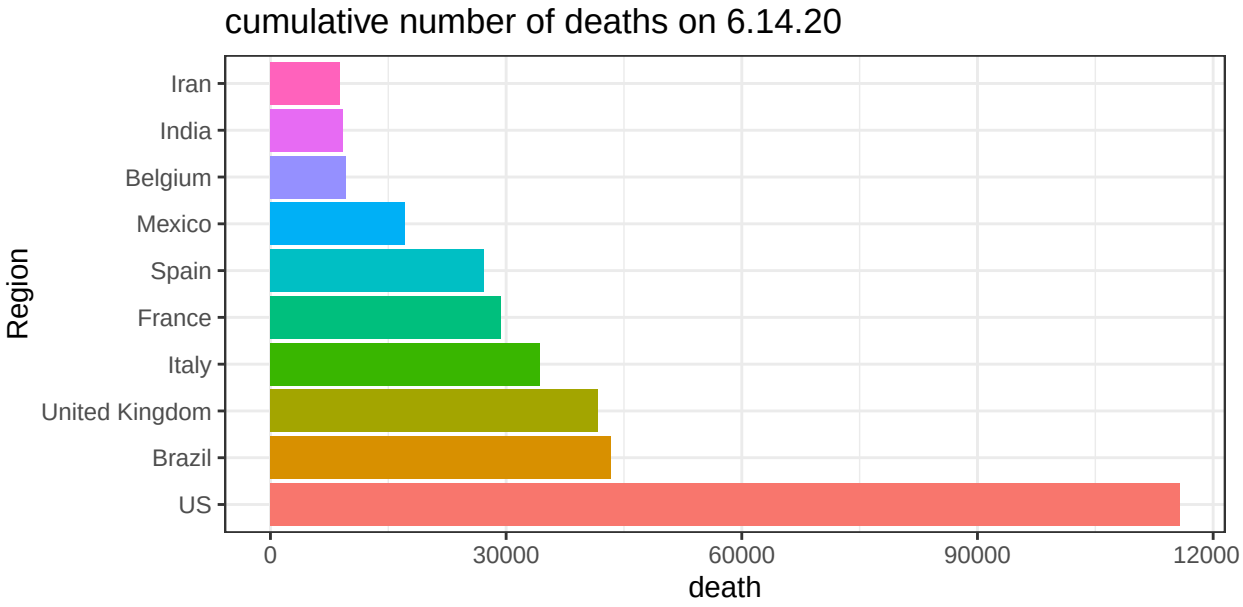
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

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.

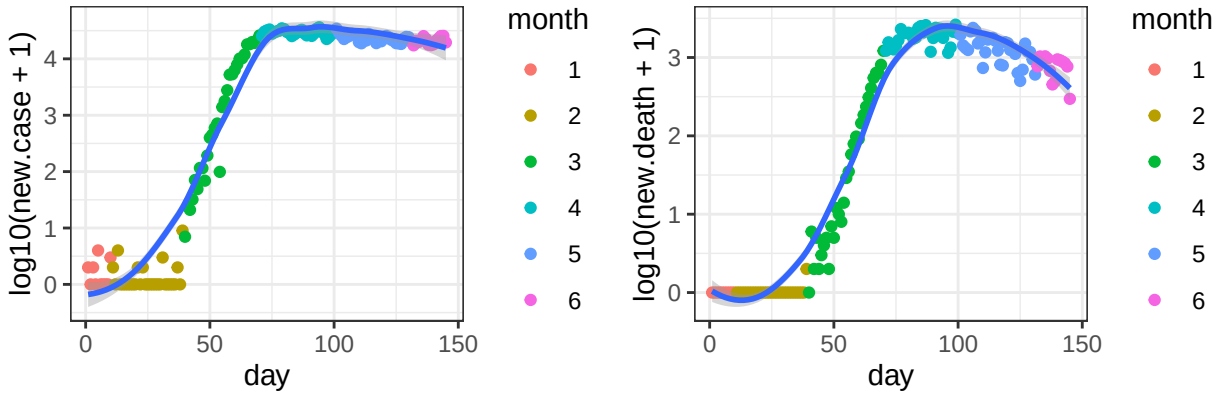
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





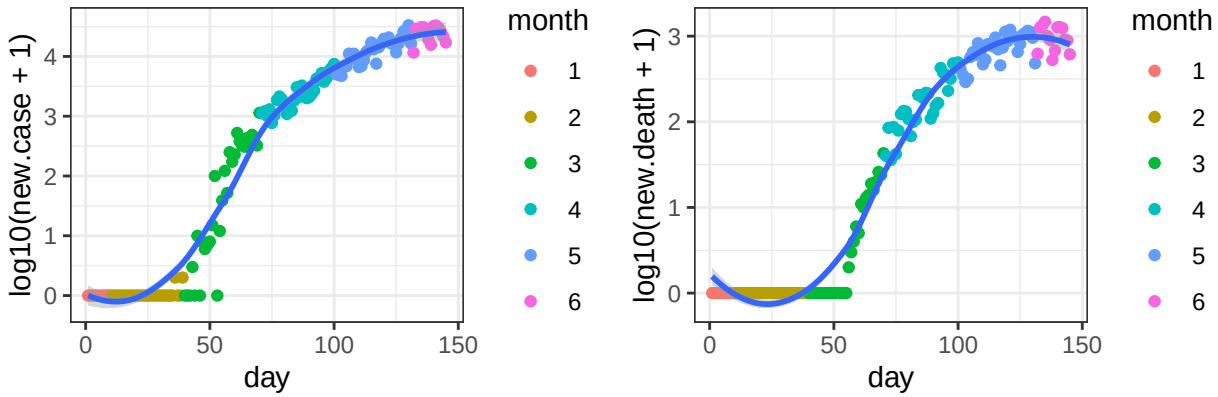
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



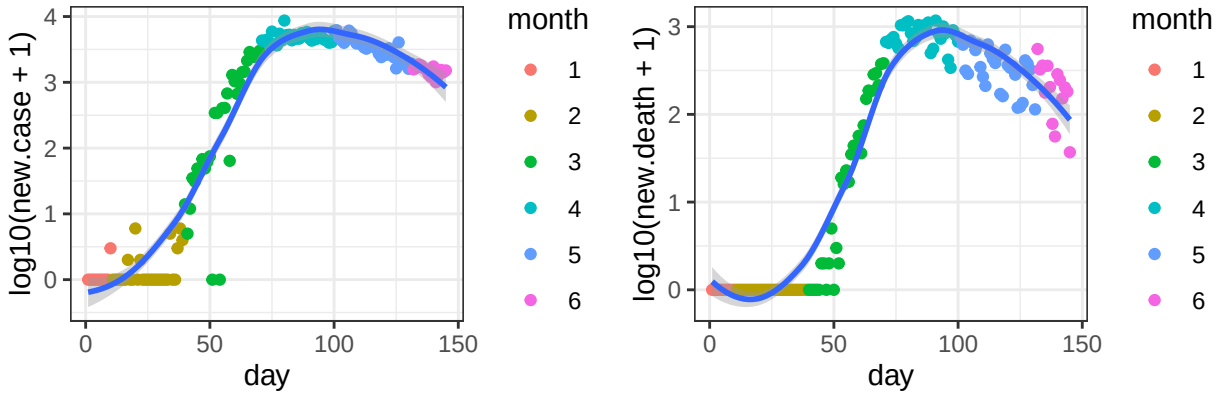
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Brazil



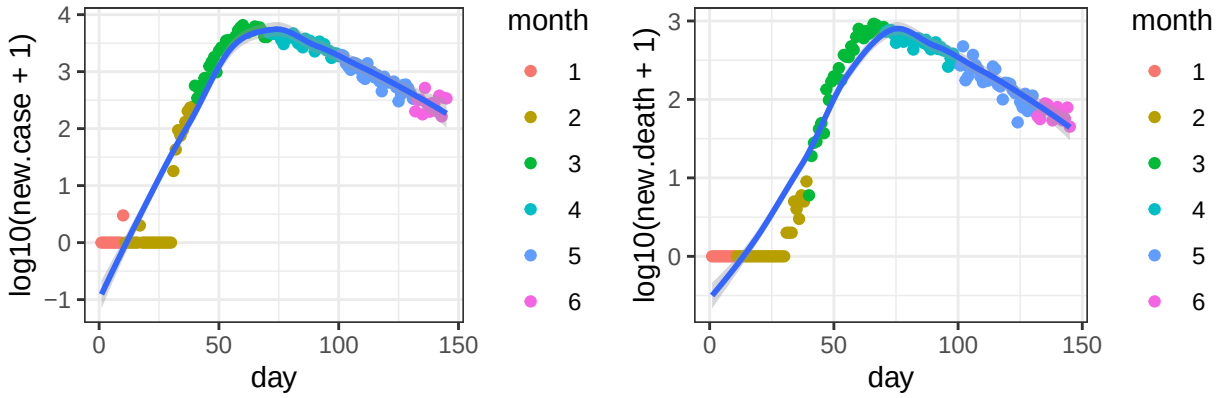
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



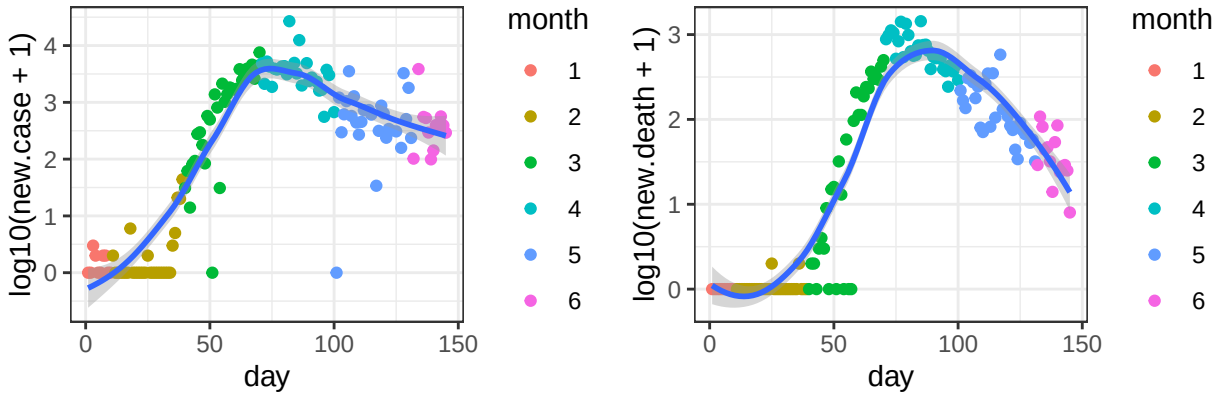
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



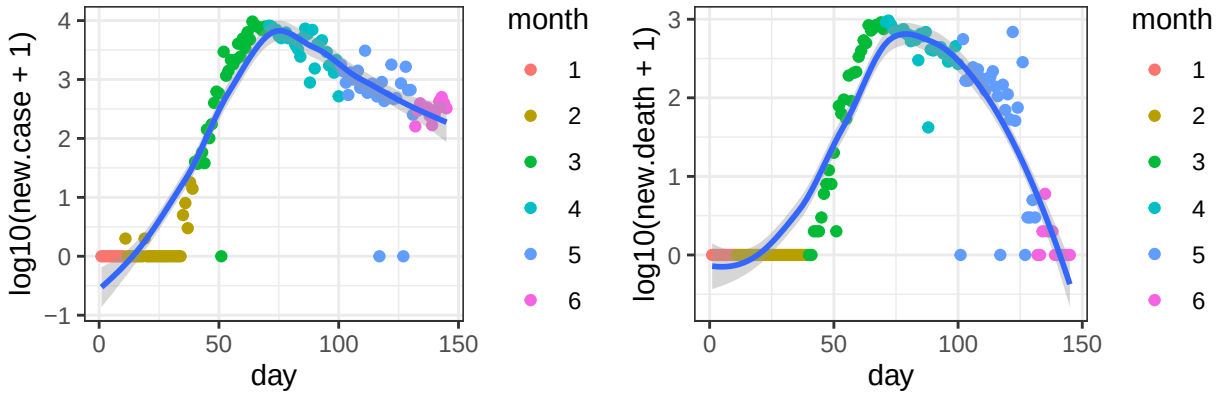
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



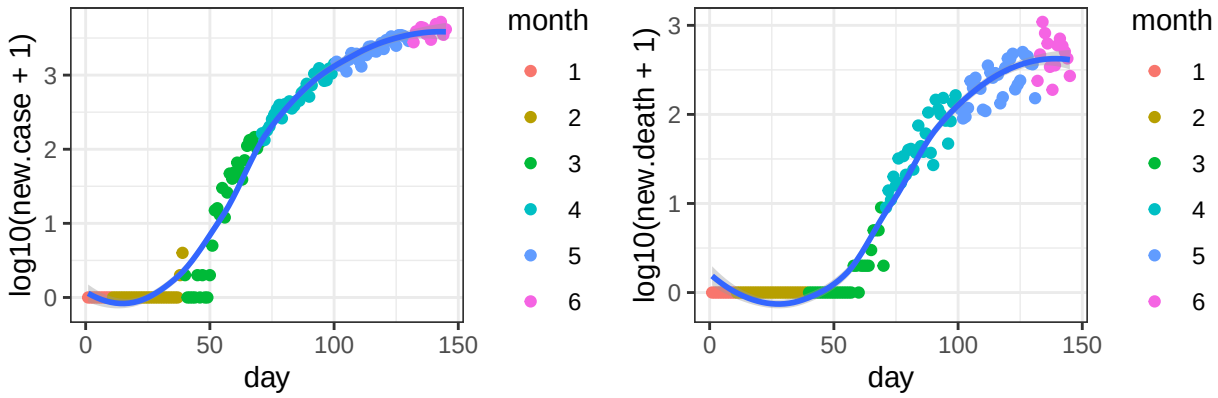
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain



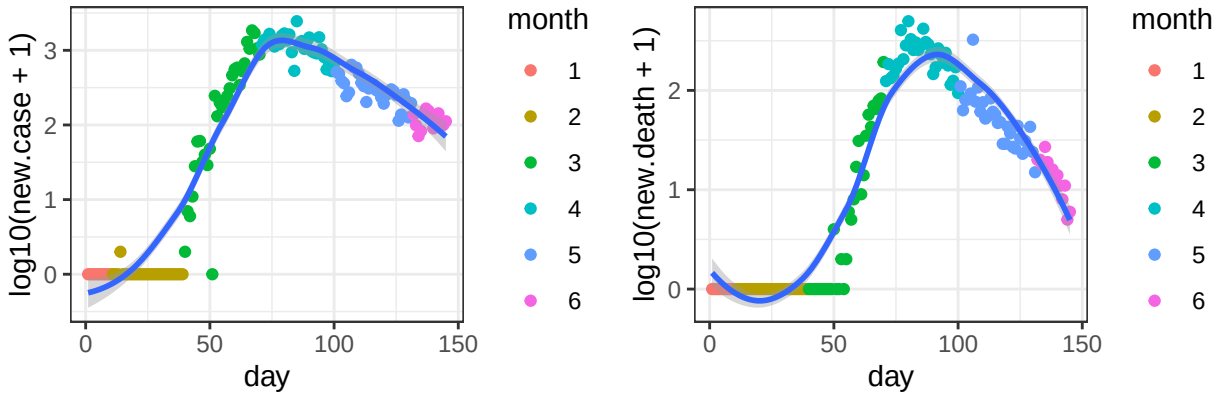
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Mexico



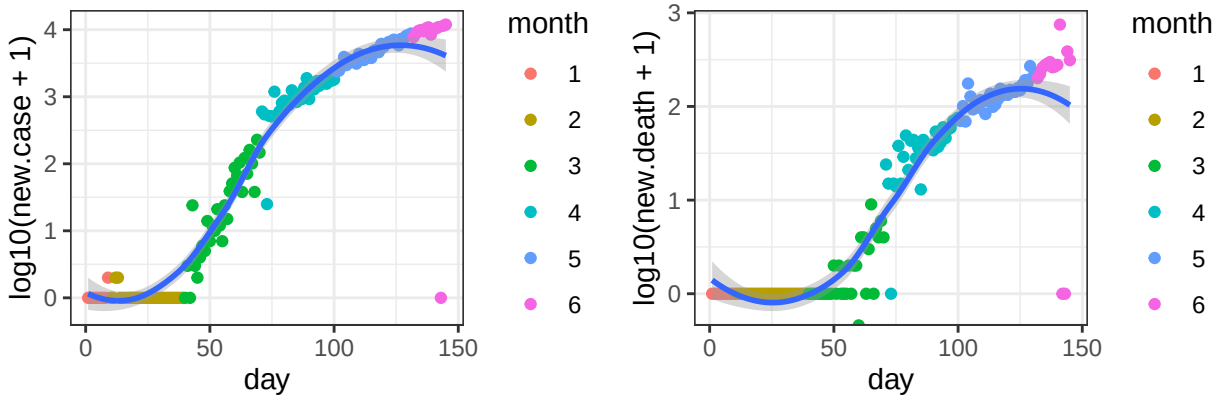
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Belgium



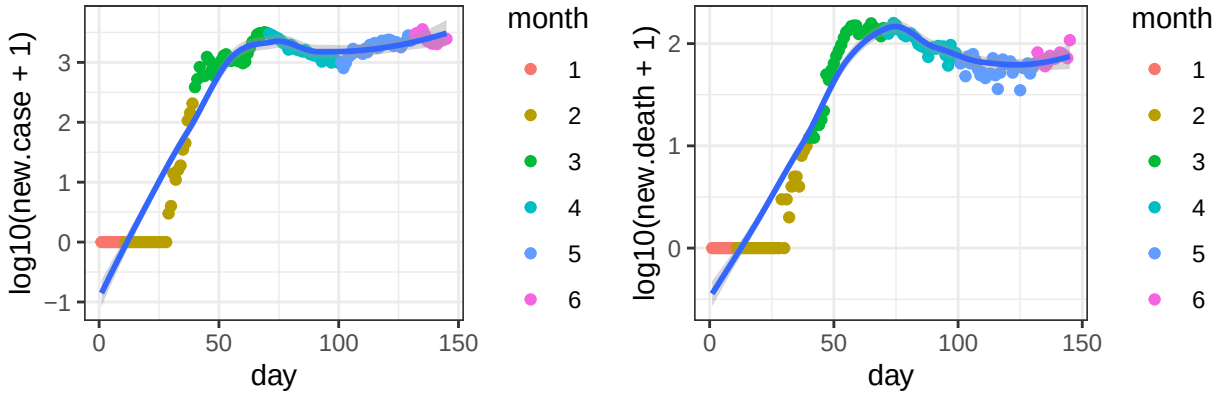
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

India



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

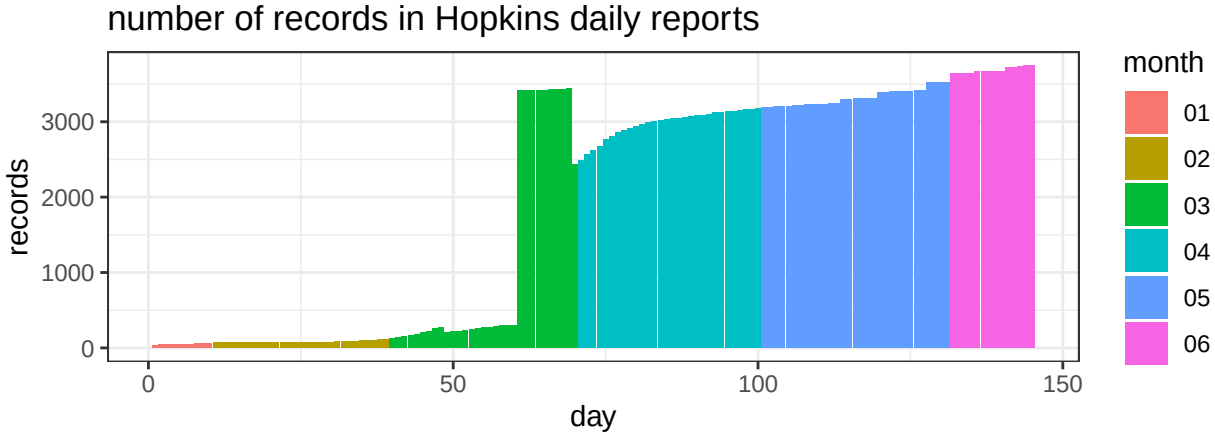
Iran



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] "2020-06-13"
```

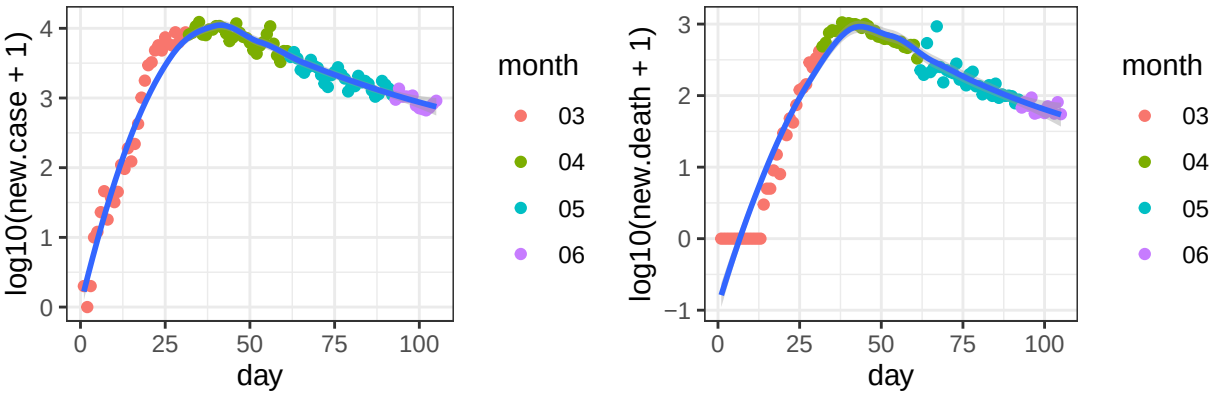
state level data

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

```
##            date                state fips  cases deaths
## 5658 2020-06-13             New York   36 387402  30565
## 5656 2020-06-13           New Jersey   34 166605  12589
## 5647 2020-06-13        Massachusetts   25 105395   7576
## 5639 2020-06-13             Illinois   17 133117   6491
## 5665 2020-06-13         Pennsylvania   42  82988   6264
## 5648 2020-06-13             Michigan   26  66024   6017
## 5629 2020-06-13           California    6 150418   5059
## 5631 2020-06-13          Connecticut    9  44994   4186
## 5644 2020-06-13            Louisiana   22  46396   3004
## 5646 2020-06-13             Maryland   24  61935   2926
## 5634 2020-06-13              Florida   12  73544   2924
## 5662 2020-06-13                 Ohio   39  40848   2554
## 5640 2020-06-13              Indiana   18  40535   2413
## 5635 2020-06-13              Georgia   13  54178   2411
## 5671 2020-06-13                Texas   48  88120   1989
## 5630 2020-06-13             Colorado    8  29002   1598
## 5675 2020-06-13             Virginia   51  53869   1541
## 5649 2020-06-13            Minnesota   27  30203   1314
## 5676 2020-06-13           Washington   53  26920   1216
## 5627 2020-06-13              Arizona    4  34773   1190
## 5659 2020-06-13       North Carolina   37  42911   1135
## 5651 2020-06-13             Missouri   29  16327    890
## 5650 2020-06-13          Mississippi   28  19348    889
## 5667 2020-06-13         Rhode Island   44  15947    833
## 5625 2020-06-13              Alabama    1  24601    773
## 5678 2020-06-13            Wisconsin   55  22638    694
## 5641 2020-06-13                 Iowa   19  23792    651
## 5668 2020-06-13       South Carolina   45  17955    599
## 5643 2020-06-13             Kentucky   21  12605    527
## 5633 2020-06-13 District of Columbia   11   9709    511
```

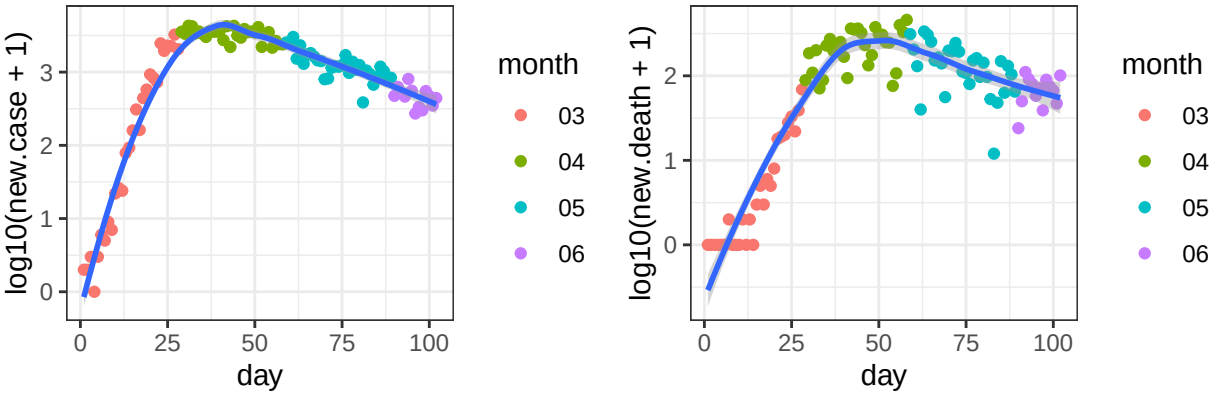
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



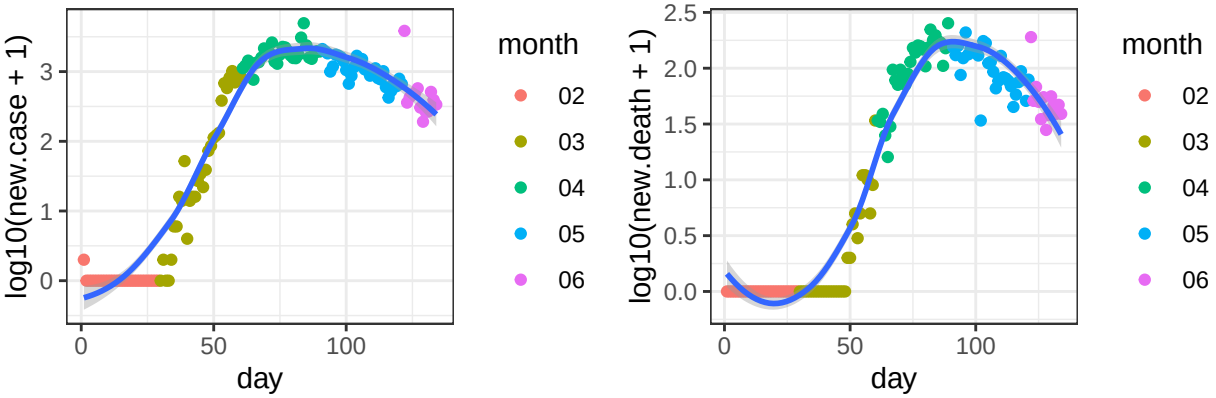
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



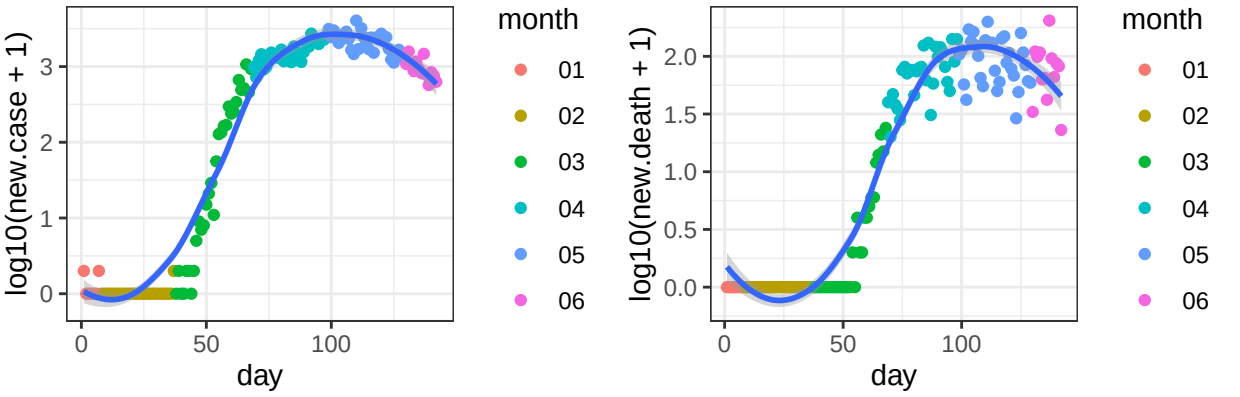
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Massachusetts



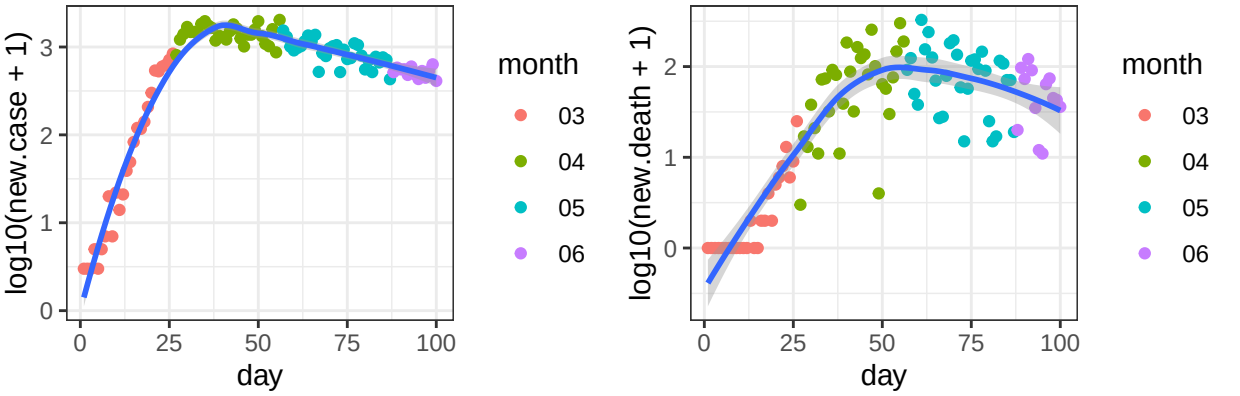
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Illinois



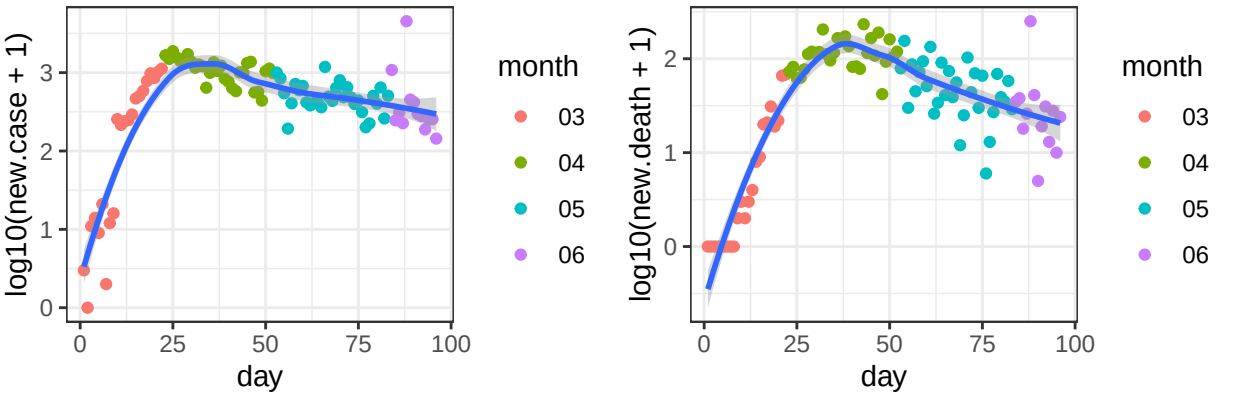
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Pennsylvania



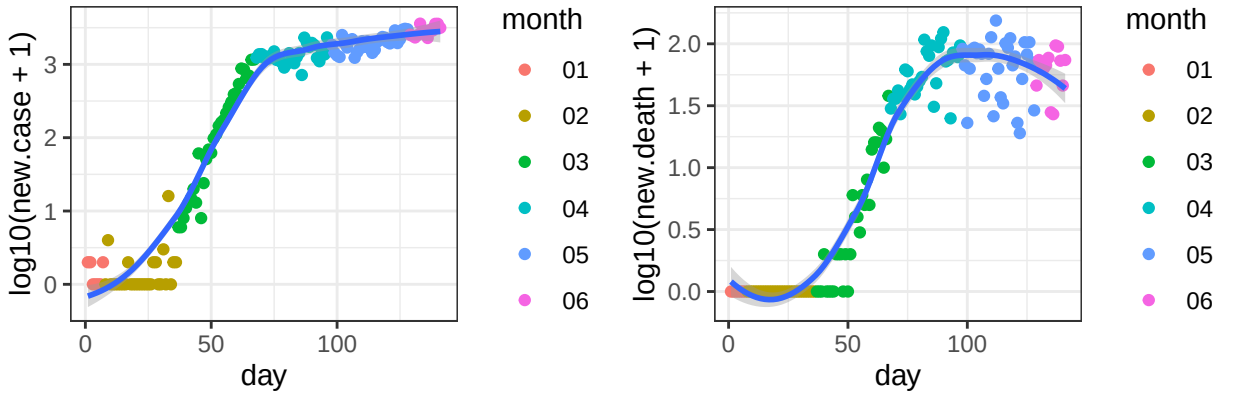
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Michigan



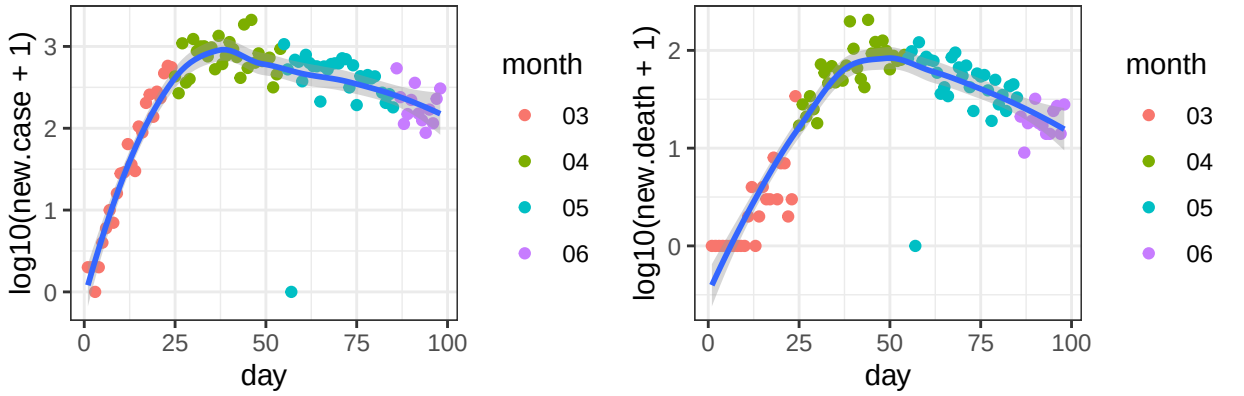
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

California



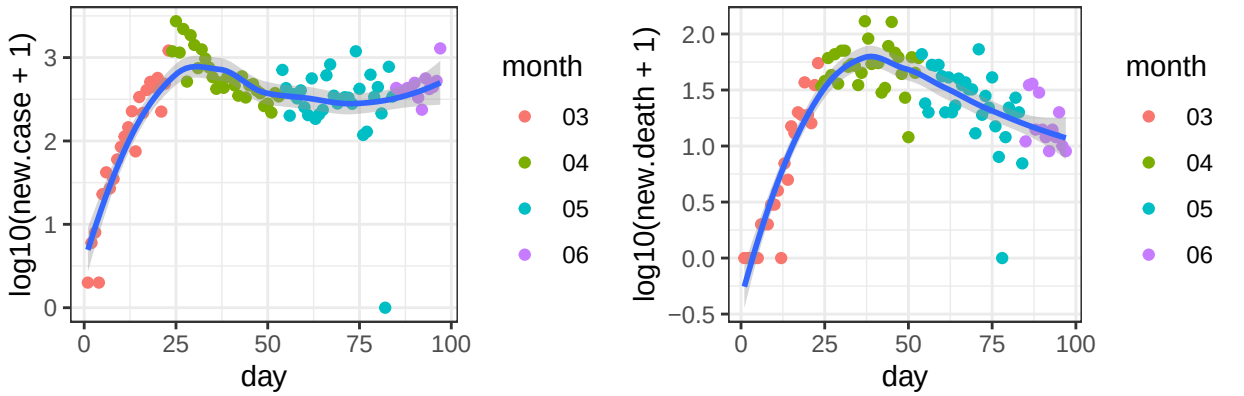
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Connecticut



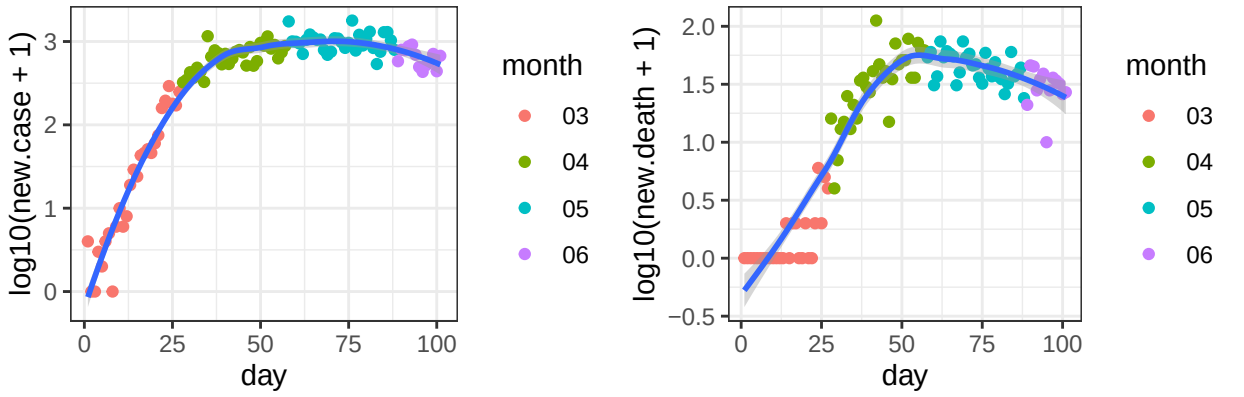
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Louisiana



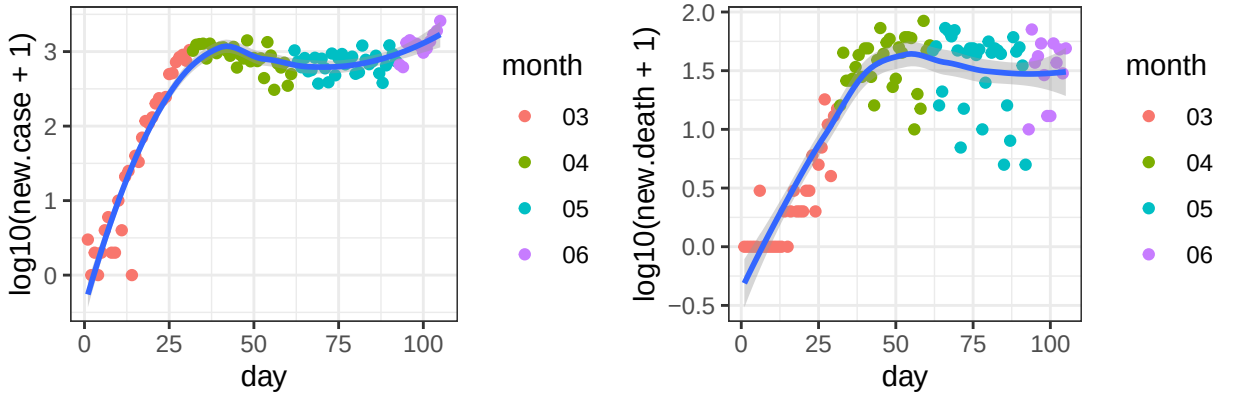
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



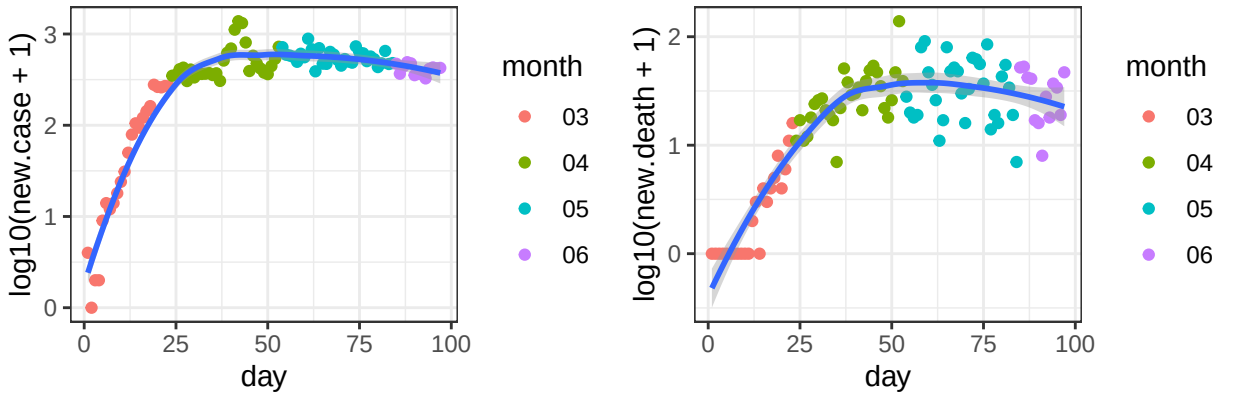
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Florida



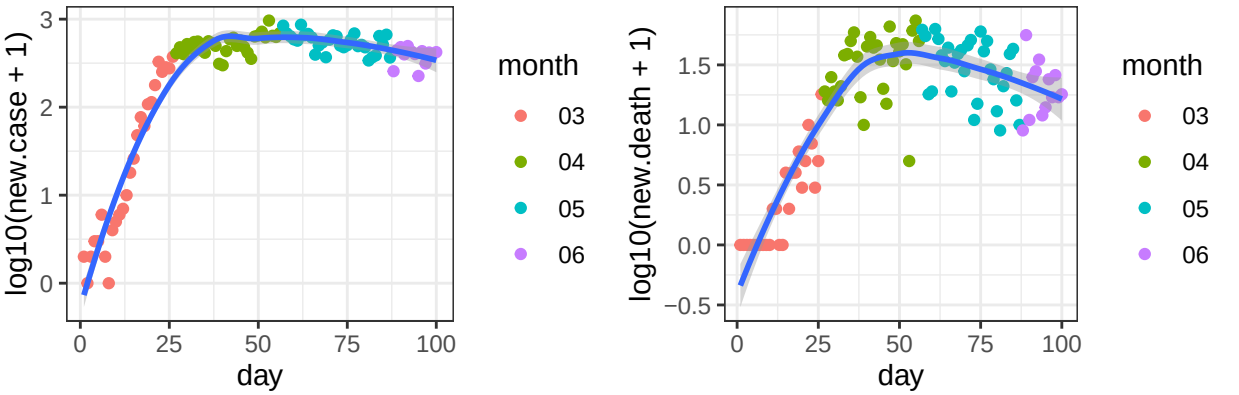
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Ohio



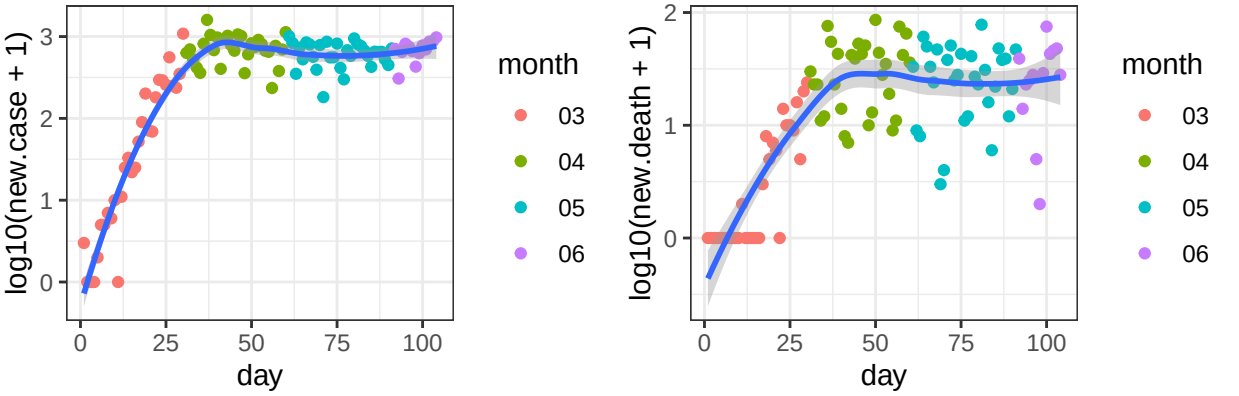
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Indiana



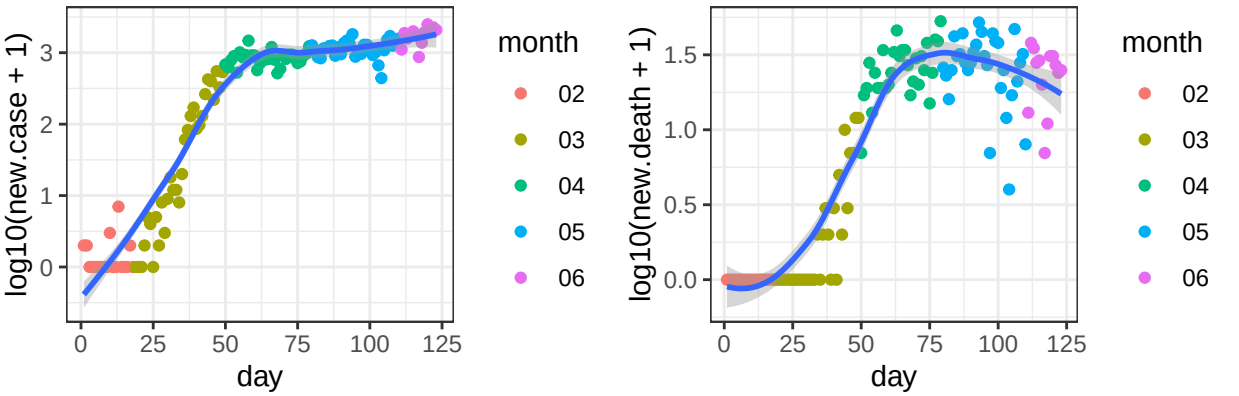
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Georgia



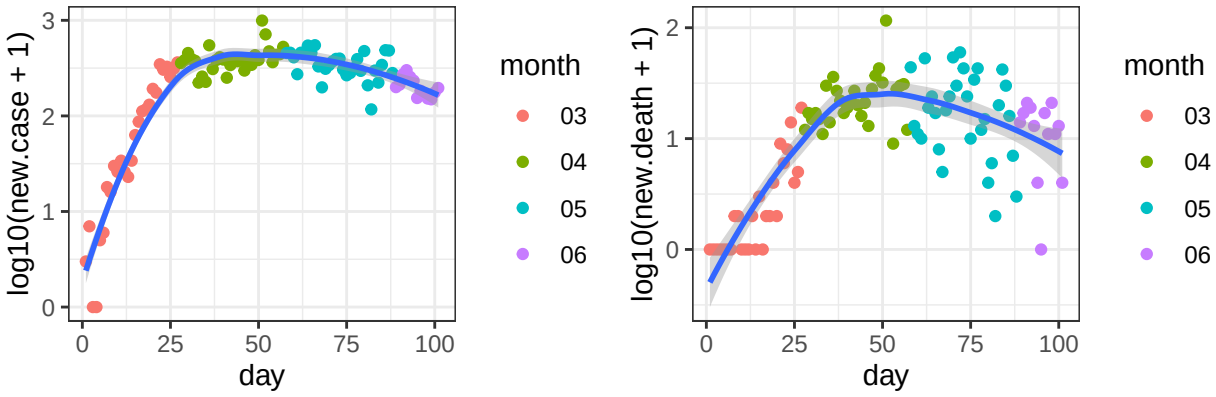
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Texas



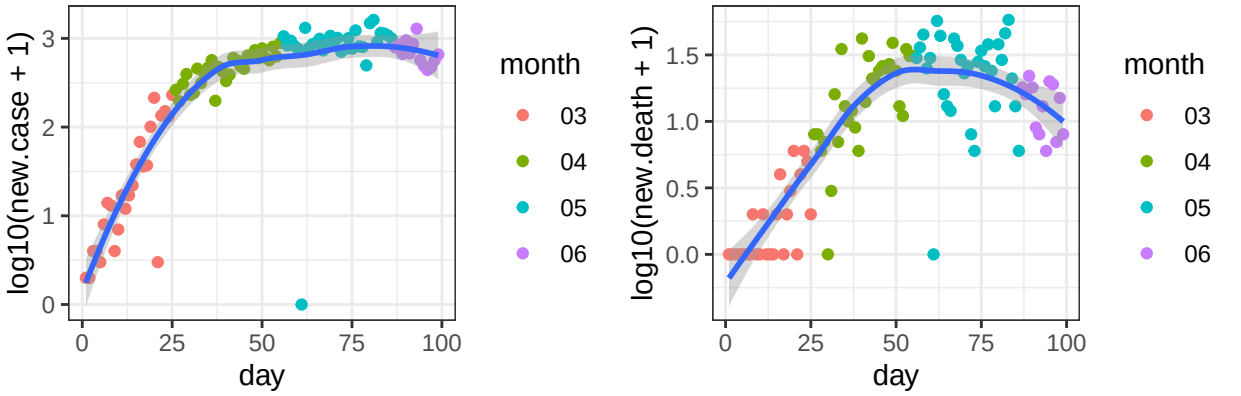
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Colorado



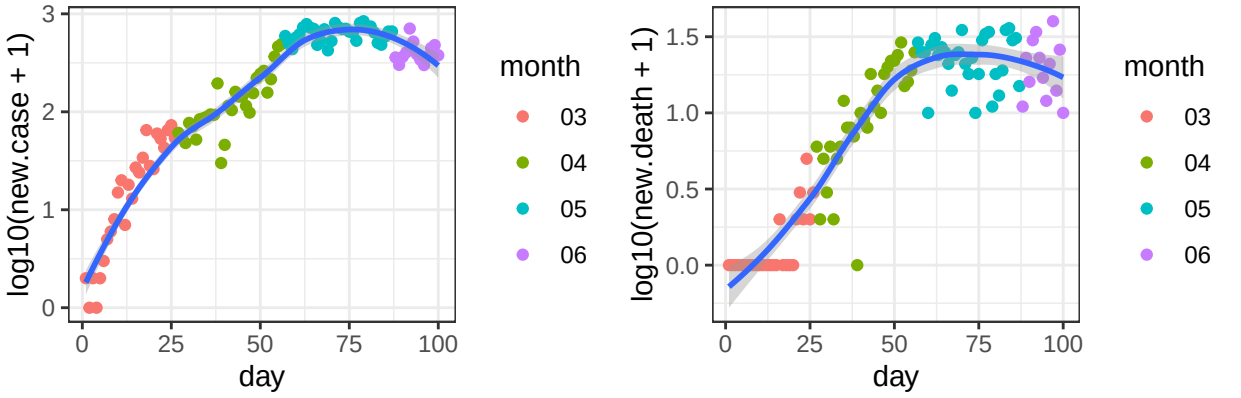
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Virginia



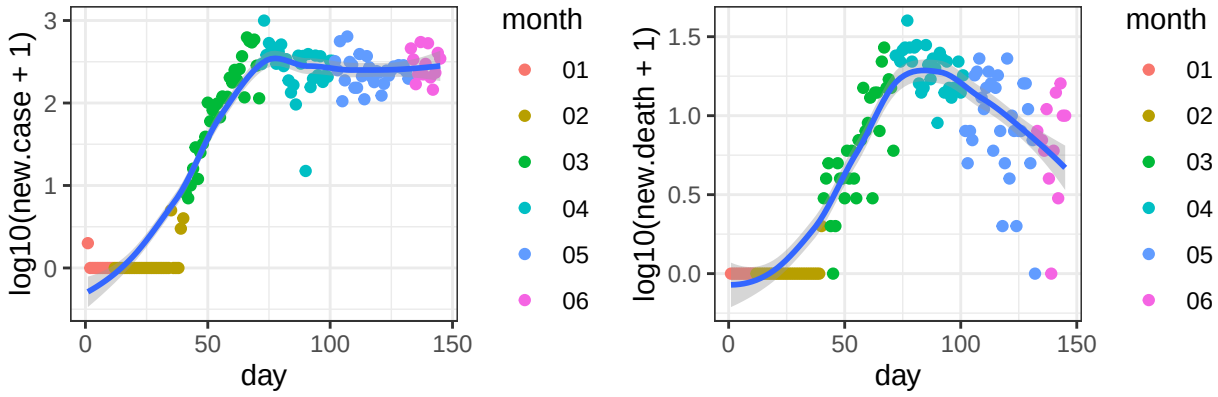
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Minnesota



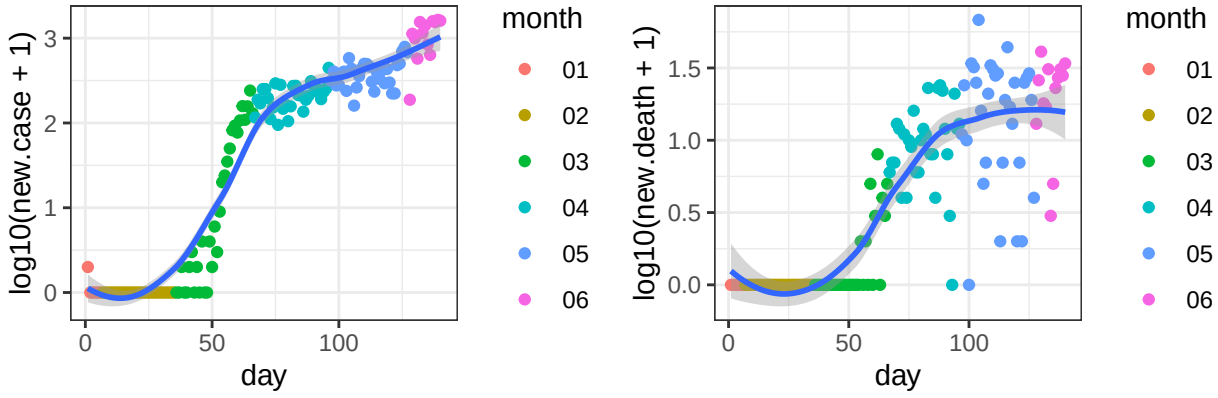
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Washington



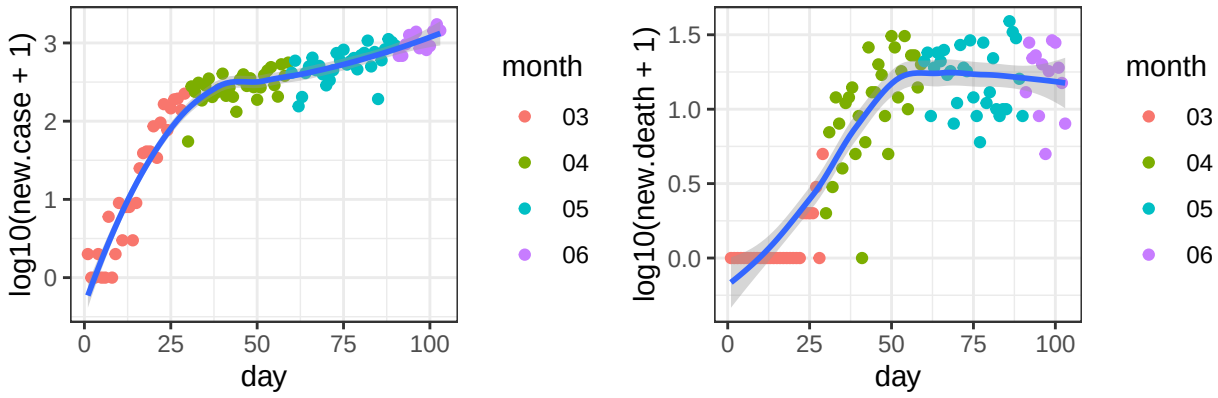
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Arizona



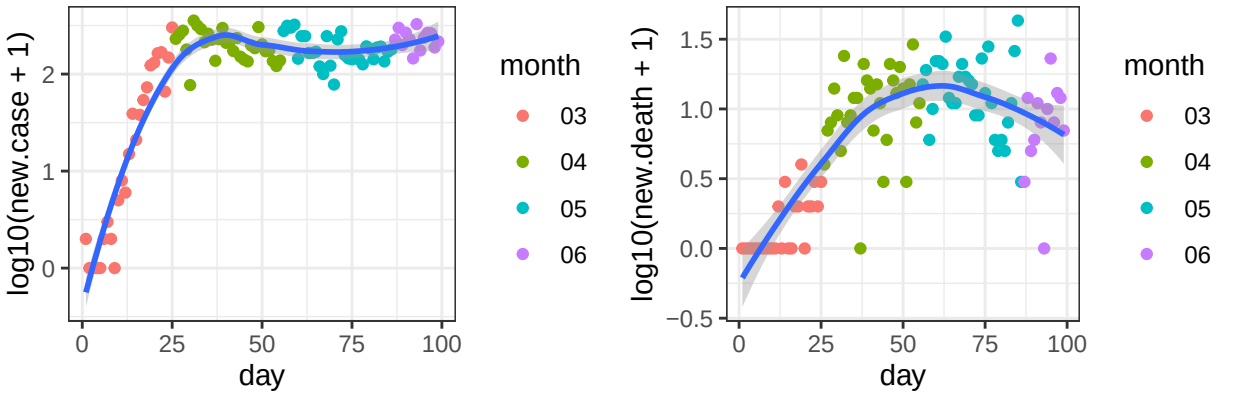
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

North Carolina



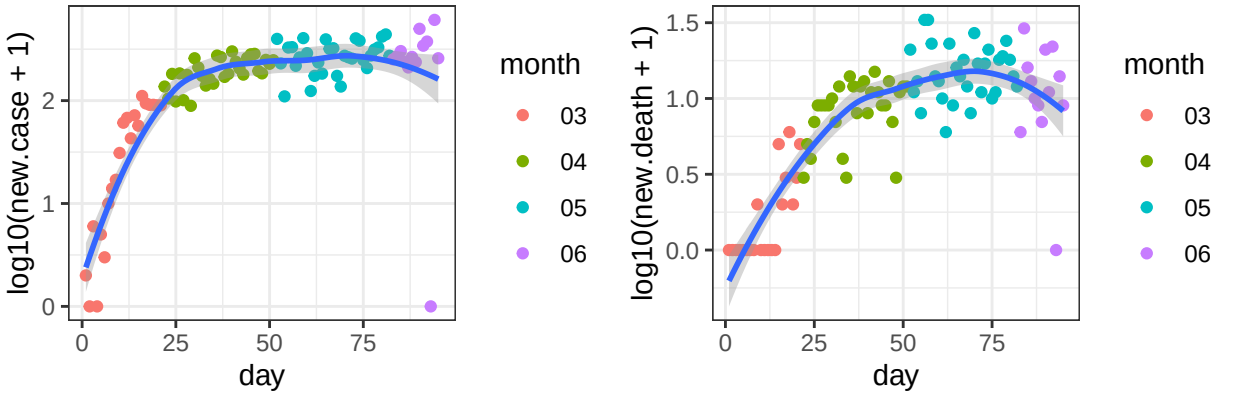
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Missouri



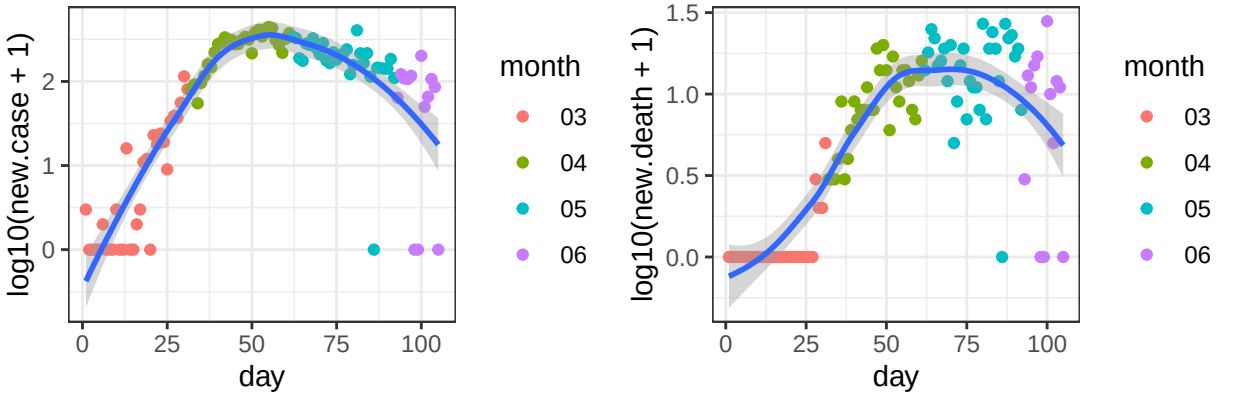
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Mississippi



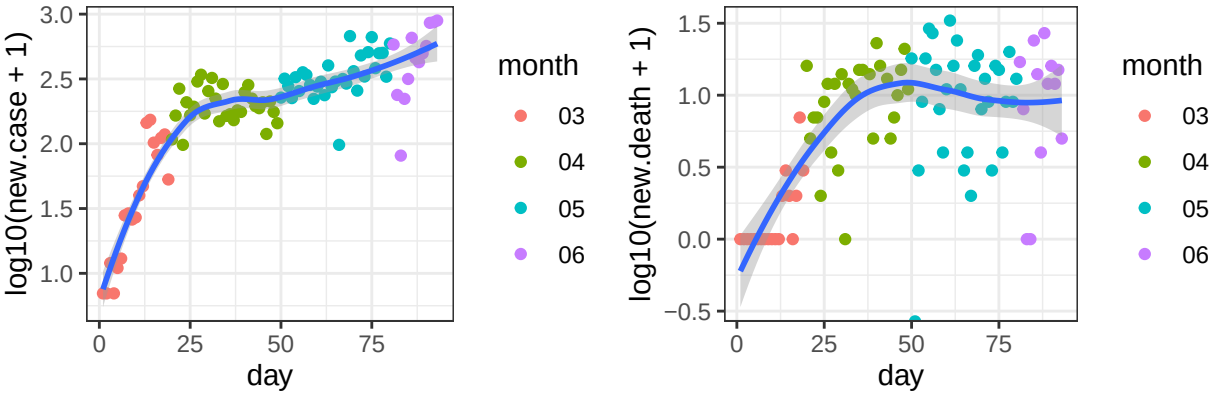
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Rhode Island



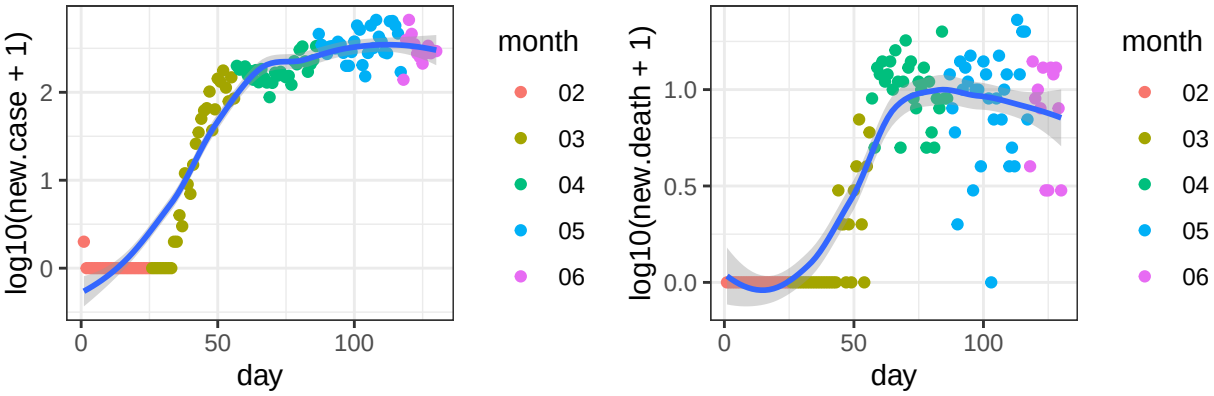
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Alabama



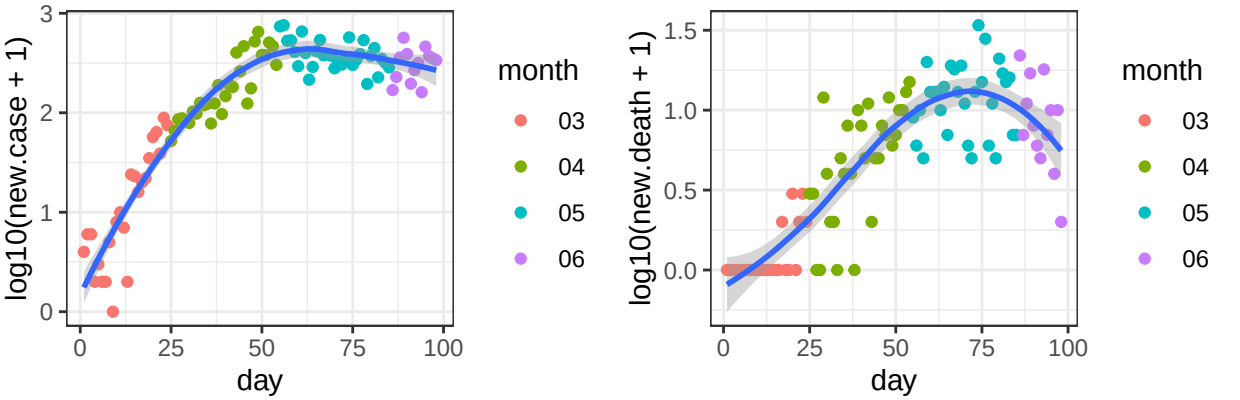
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Wisconsin



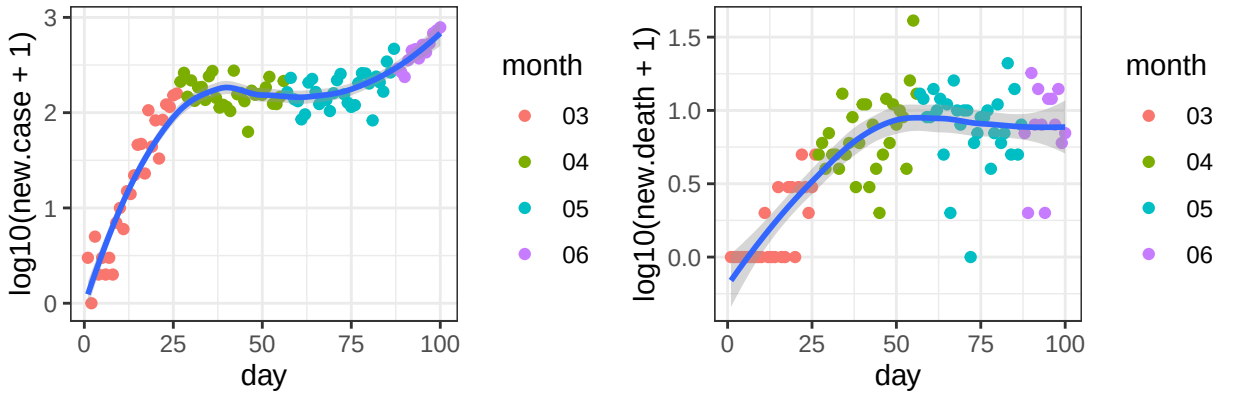
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

Iowa



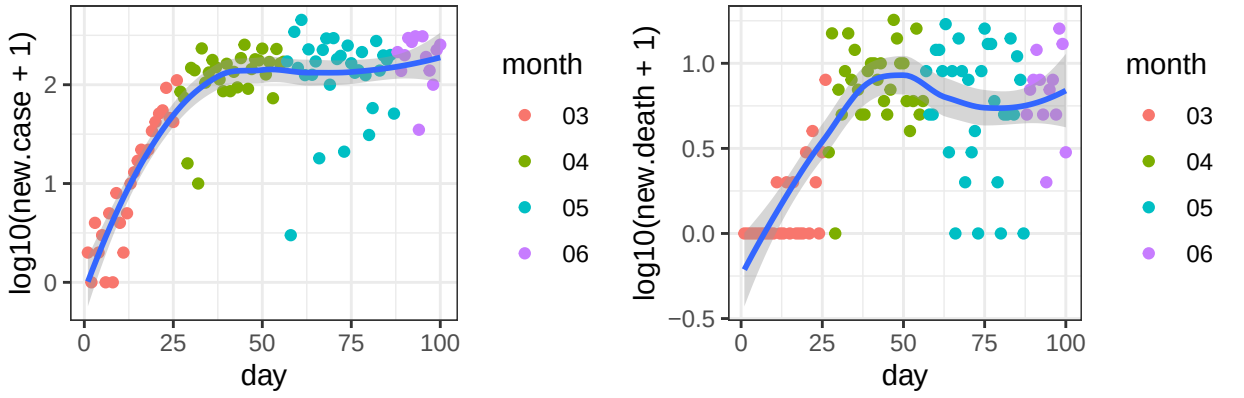
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

South Carolina



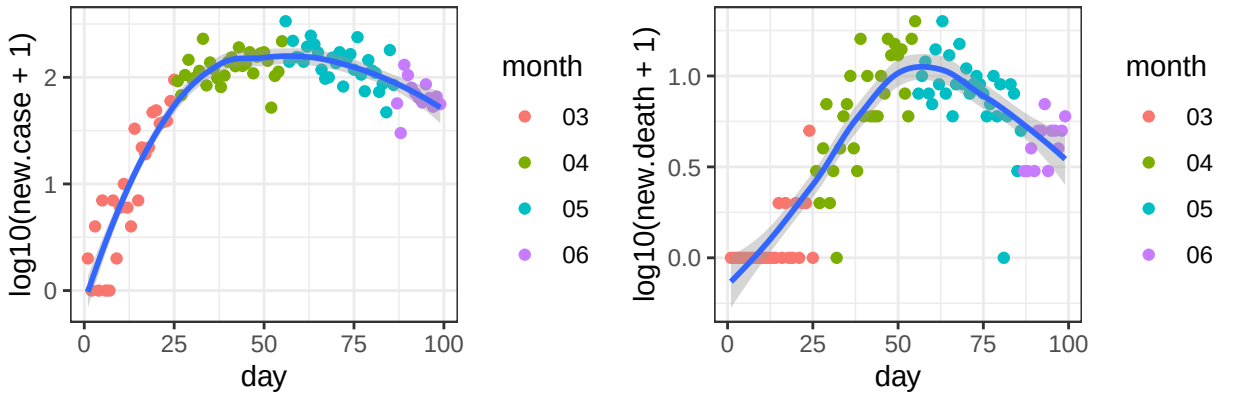
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Kentucky



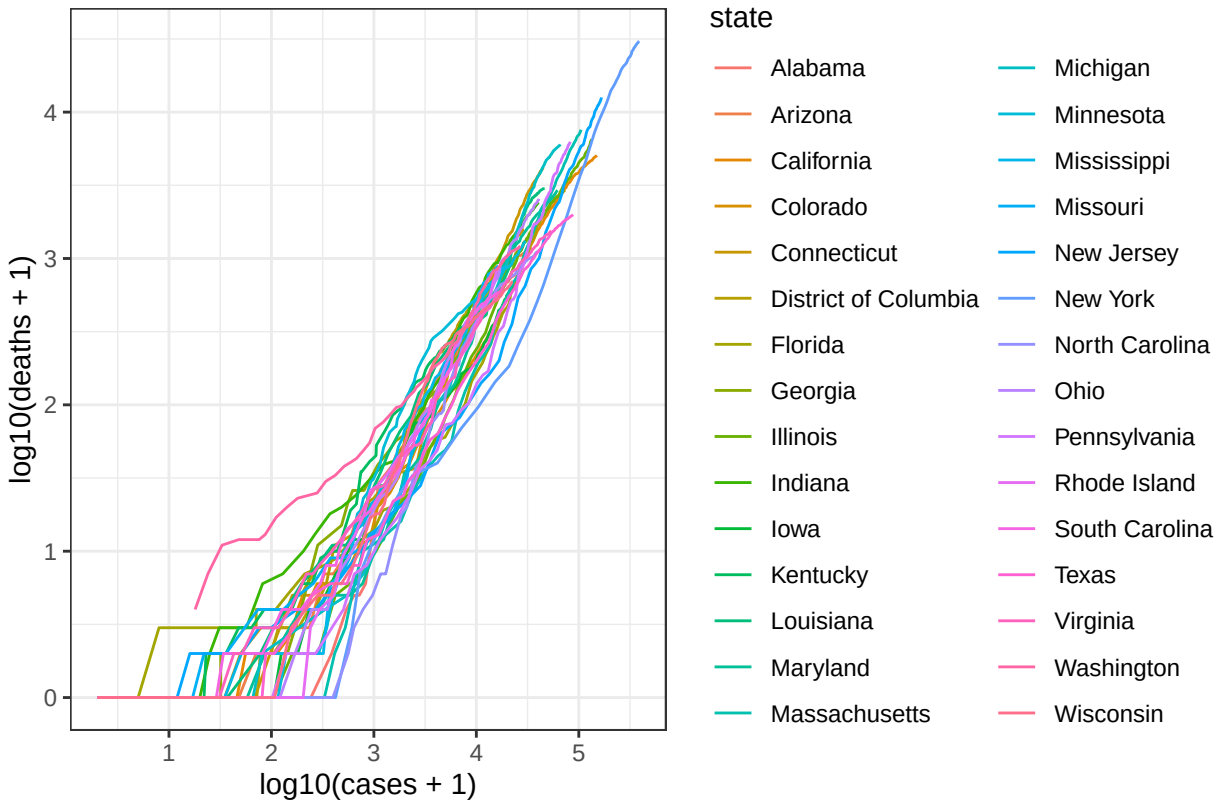
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

District of Columbia



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

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

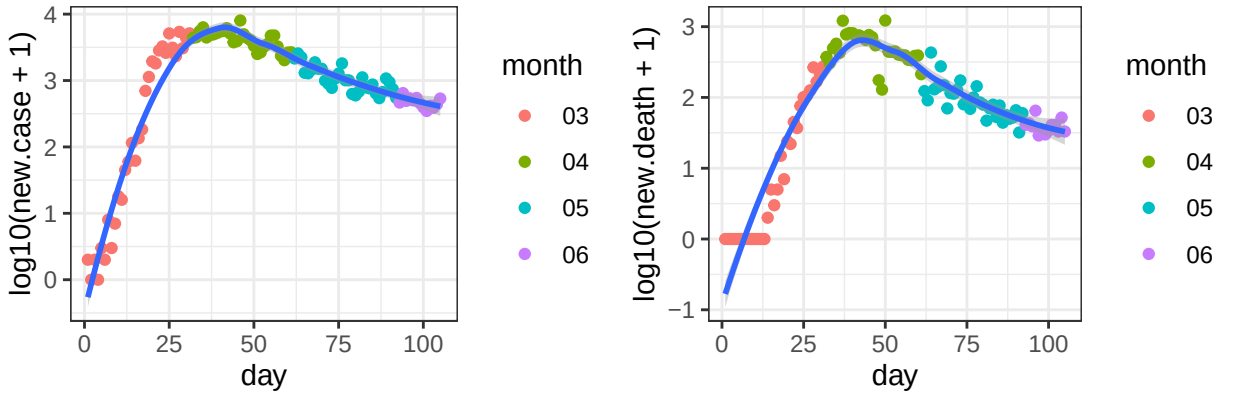
First check the 50 counties with the largest number of deaths.

```
##              date              county               state  fips  cases deaths
## 232498 2020-06-13       New York City            New York    NA 214242  21551
## 231311 2020-06-13                Cook            Illinois 17031  84581   4173
## 230915 2020-06-13         Los Angeles          California  6037  72023   2890
## 232002 2020-06-13               Wayne            Michigan 26163  21711   2669
## 232497 2020-06-13              Nassau            New York 36059  41172   2668
## 232517 2020-06-13             Suffolk            New York 36103  40615   1996
## 231914 2020-06-13           Middlesex       Massachusetts 25017  23156   1748
## 232423 2020-06-13               Essex          New Jersey 34013  18336   1741
## 232418 2020-06-13              Bergen          New Jersey 34003  18805   1664
## 232525 2020-06-13          Westchester            New York 36119  34252   1535
## 232922 2020-06-13        Philadelphia        Pennsylvania 42101  24338   1502
## 231014 2020-06-13           Fairfield         Connecticut  9001  16277   1345
## 231015 2020-06-13            Hartford         Connecticut  9003  11189   1321
## 232425 2020-06-13              Hudson          New Jersey 34017  18717   1253
## 232436 2020-06-13               Union          New Jersey 34039  16337   1121
## 232428 2020-06-13           Middlesex          New Jersey 34023  16385   1074
## 231983 2020-06-13             Oakland            Michigan 26125  11298   1067
## 231018 2020-06-13           New Haven         Connecticut  9009  12021   1041
## 231910 2020-06-13               Essex       Massachusetts 25009  15573   1041
## 232432 2020-06-13             Passaic          New Jersey 34031  16612    997
## 231918 2020-06-13             Suffolk       Massachusetts 25025  19299    943
## 231970 2020-06-13              Macomb            Michigan 26099   7035    889
```

```
## 231916 2020-06-13              Norfolk        Massachusetts 25021   8860    882
## 231920 2020-06-13            Worcester        Massachusetts 25027  11961    863
## 231070 2020-06-13           Miami-Dade              Florida 12086  21632    822
## 232431 2020-06-13                Ocean           New Jersey 34029   9222    813
## 232917 2020-06-13           Montgomery         Pennsylvania 42091   7865    768
## 232030 2020-06-13             Hennepin            Minnesota 27053  10069    712
## 231446 2020-06-13               Marion              Indiana 18097  10736    693
## 231896 2020-06-13           Montgomery             Maryland 24031  13573    686
## 232429 2020-06-13             Monmouth           New Jersey 34025   8720    667
## 232894 2020-06-13             Delaware         Pennsylvania 42045   6894    664
## 231912 2020-06-13              Hampden        Massachusetts 25013   6460    637
## 232943 2020-06-13           Providence         Rhode Island 44007  11959    637
## 232430 2020-06-13               Morris           New Jersey 34027   6568    635
## 231897 2020-06-13      Prince George's             Maryland 24033  17745    634
## 231917 2020-06-13             Plymouth        Massachusetts 25023   8478    621
## 233570 2020-06-13                 King           Washington 53033   8702    593
## 232483 2020-06-13                 Erie             New York 36029   6753    573
## 230814 2020-06-13              Maricopa              Arizona  4013  17791    549
## 232880 2020-06-13                Bucks         Pennsylvania 42017   5419    542
## 232427 2020-06-13               Mercer           New Jersey 34021   7323    517
## 231834 2020-06-13              Orleans            Louisiana 22071   7374    516
## 231027 2020-06-13 District of Columbia District of Columbia 11001   9709    511
## 231908 2020-06-13              Bristol        Massachusetts 25005   7906    507
## 232270 2020-06-13            St. Louis             Missouri 29189   5506    500
## 232509 2020-06-13             Rockland             New York 36087  13411    466
## 231824 2020-06-13            Jefferson            Louisiana 22051   8339    465
## 232434 2020-06-13             Somerset           New Jersey 34035   4756    436
## 231317 2020-06-13                DuPage             Illinois 17043   8402    430
```

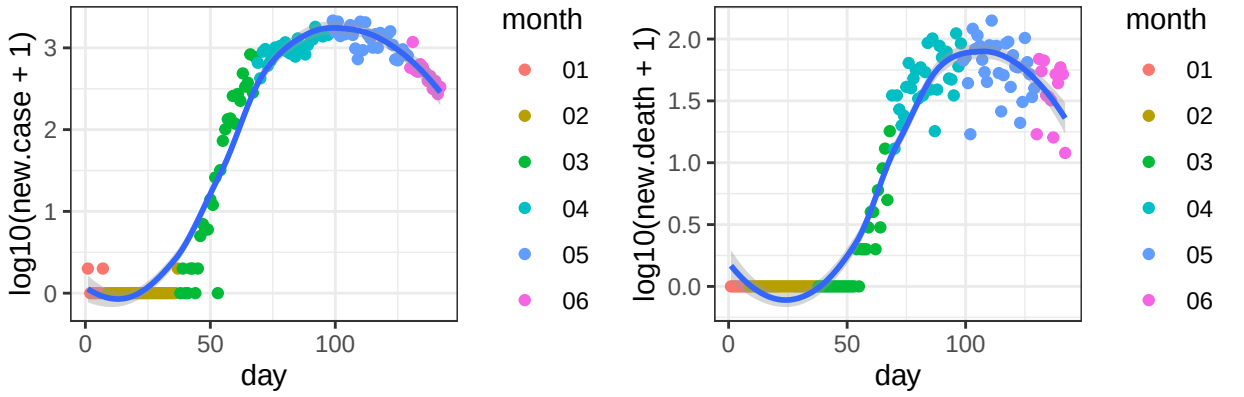
For these 50 counties, I check the number of new cases and the number of new deaths.

New York City_New York



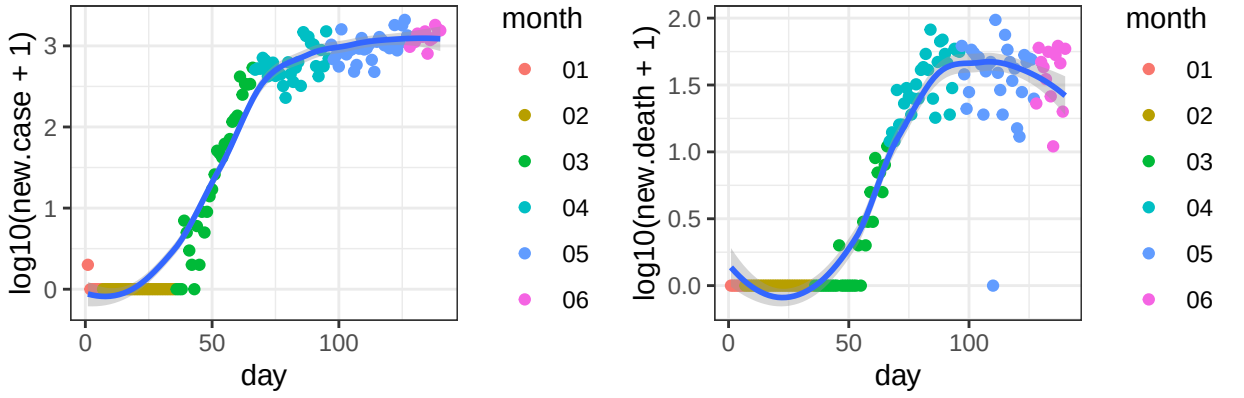
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



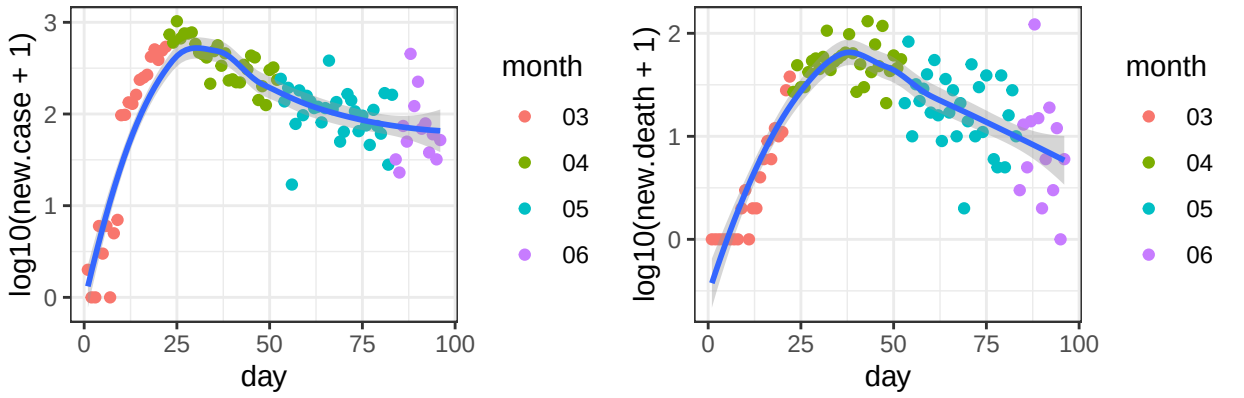
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Los Angeles_California



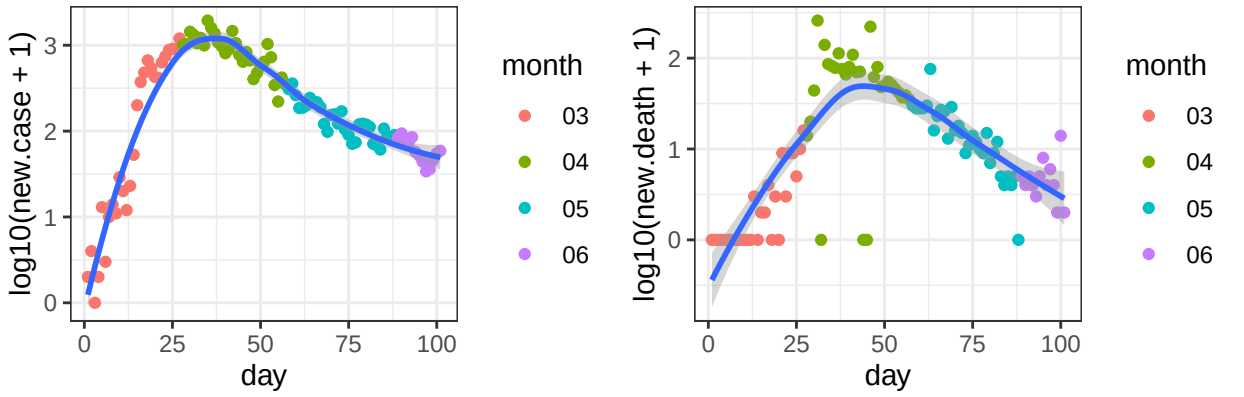
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Wayne_Michigan



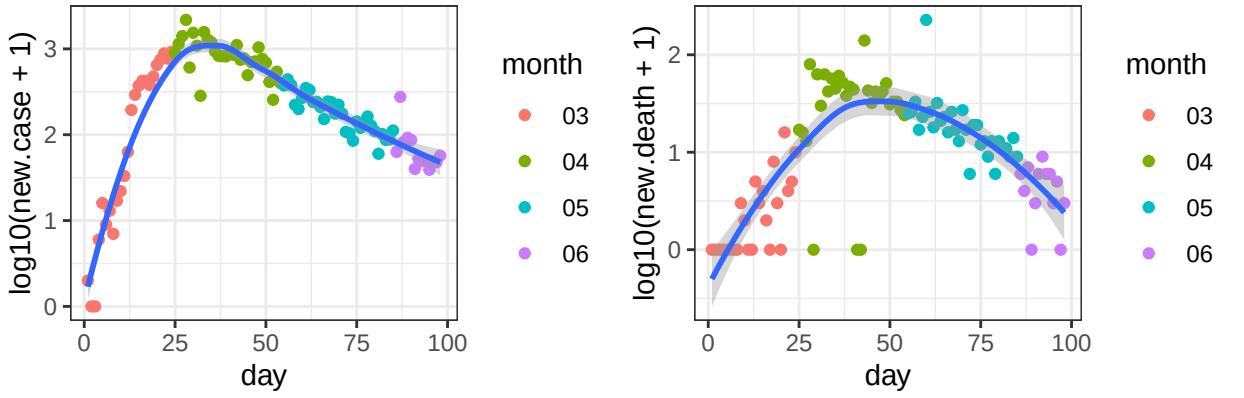
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Nassau_New York



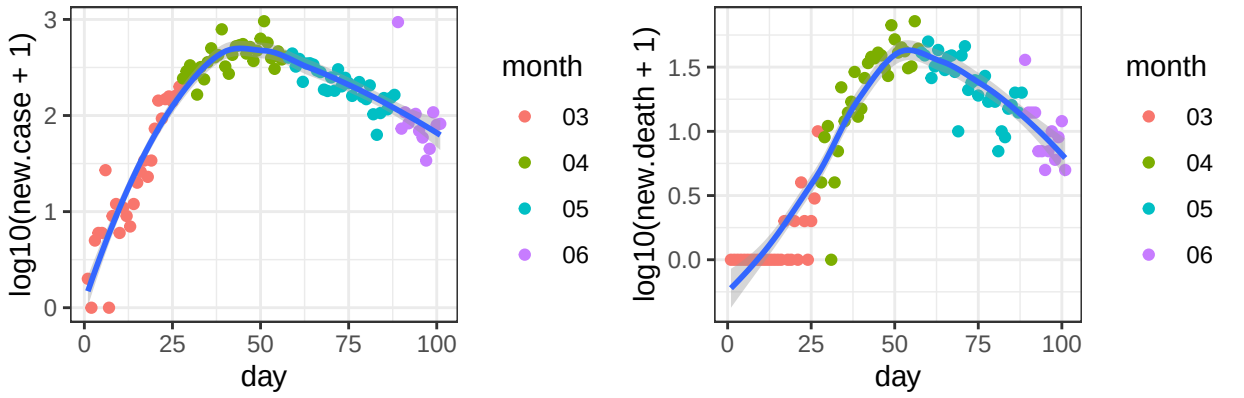
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Suffolk_New York



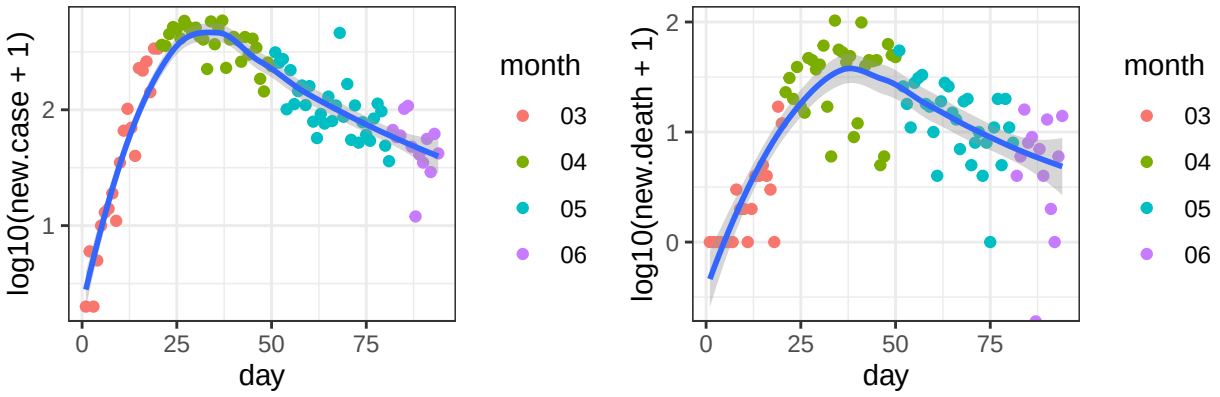
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Middlesex_Massachusetts



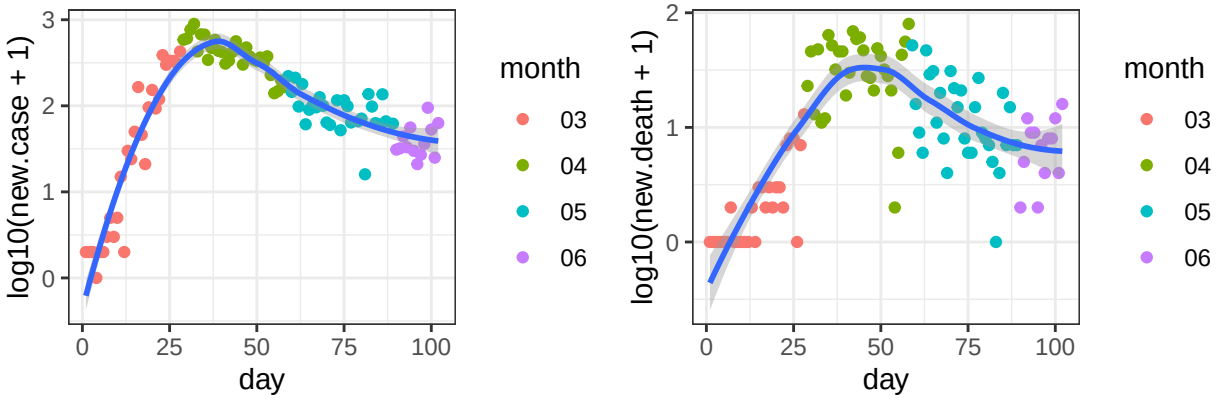
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



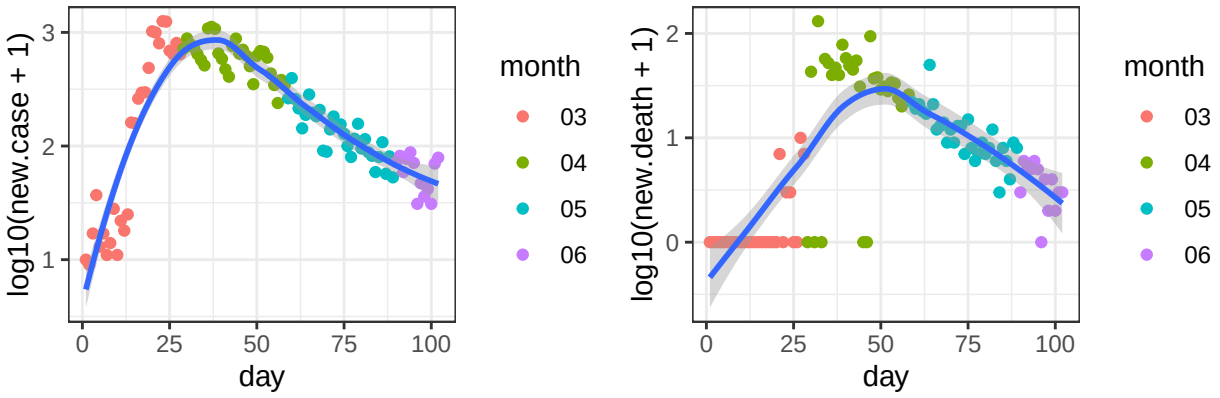
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Bergen_New Jersey



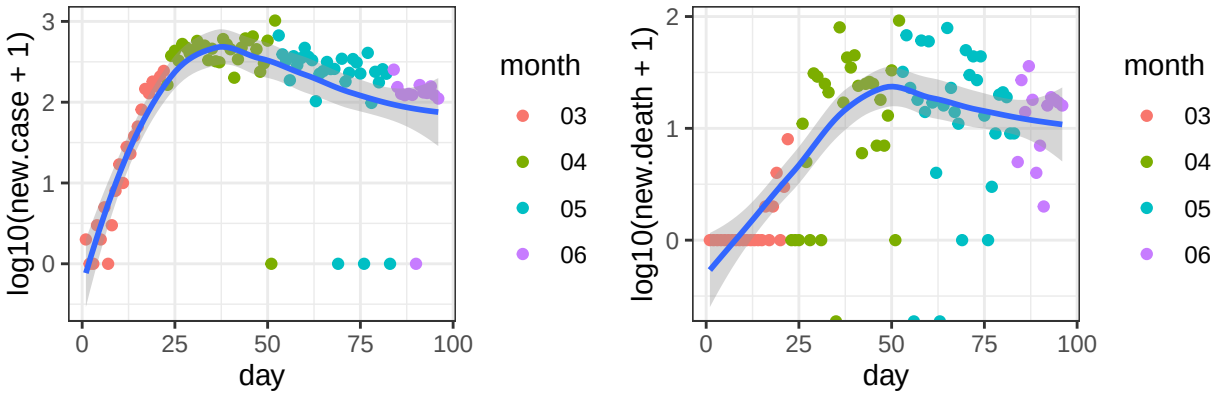
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Westchester_New York



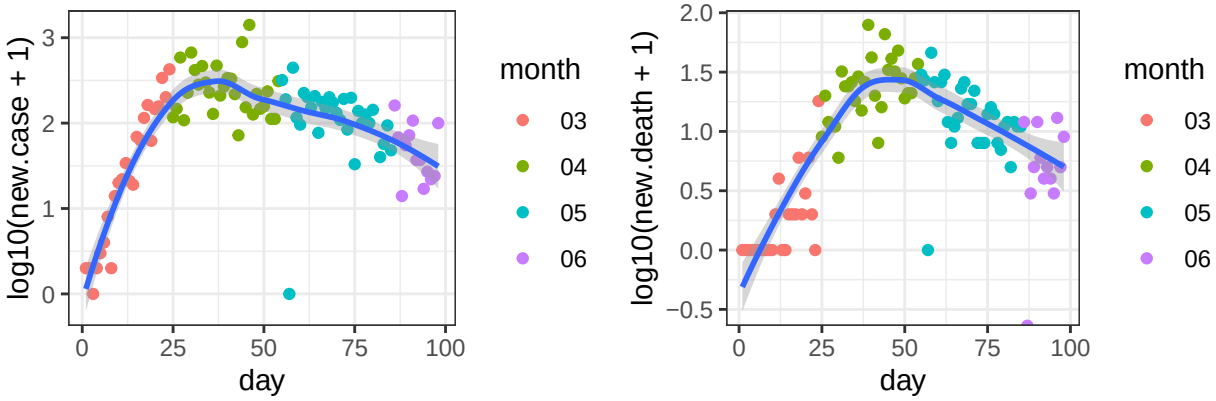
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Philadelphia_Pennsylvania



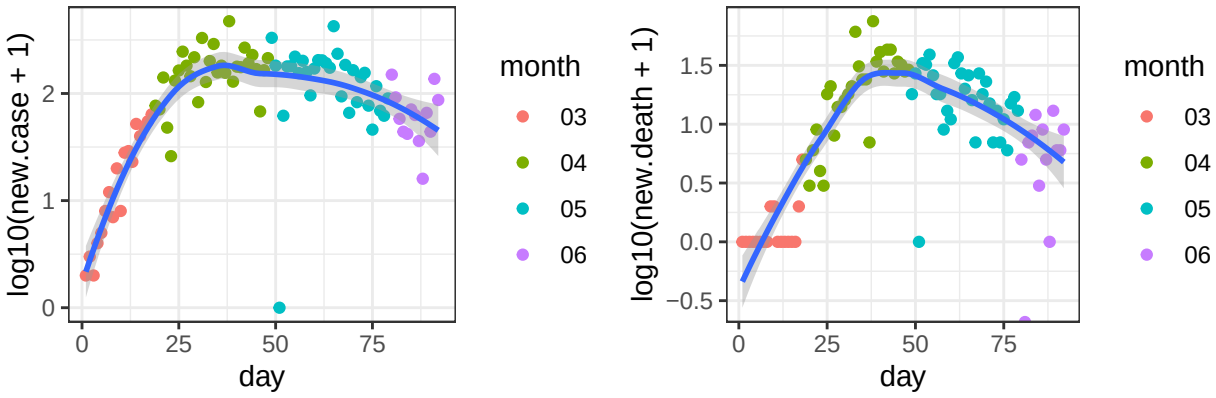
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Fairfield_Connecticut



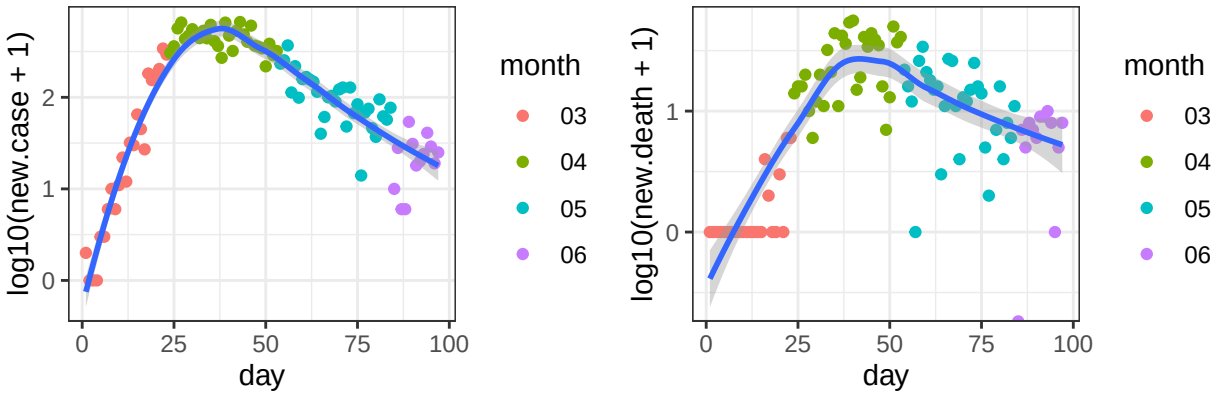
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Hartford_Connecticut



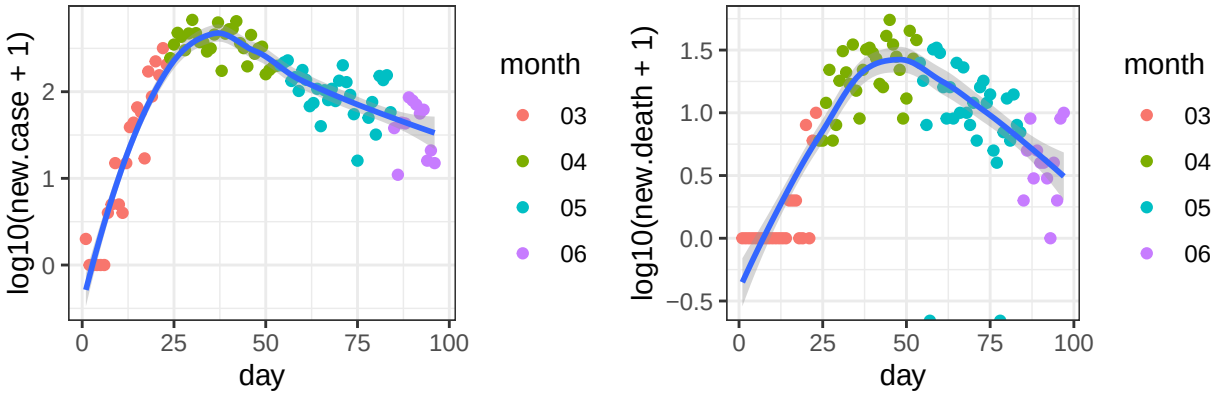
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Hudson_New Jersey



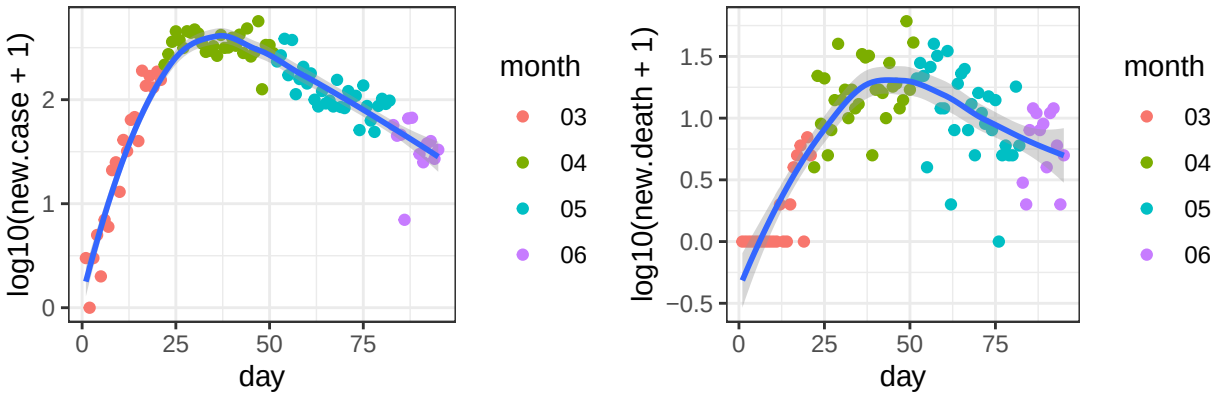
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Union_New Jersey



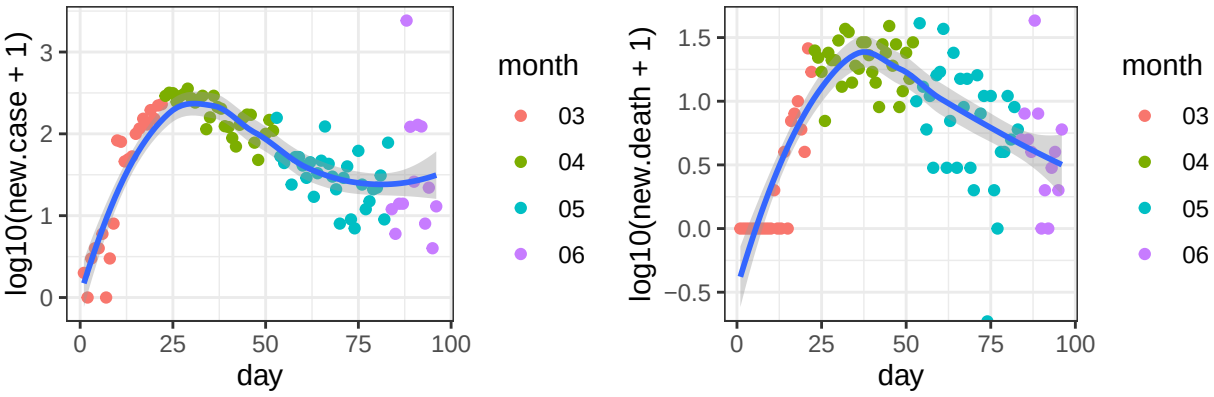
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Middlesex_New Jersey



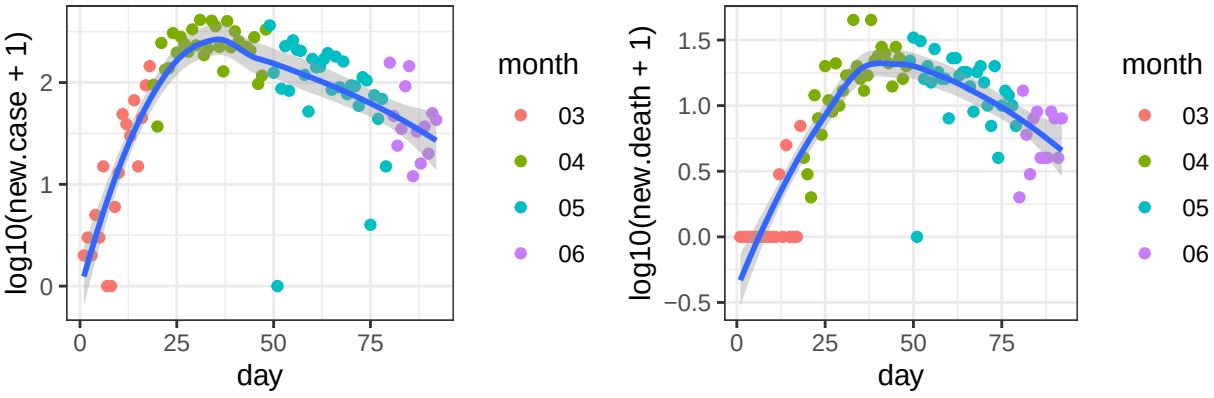
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Oakland_Michigan



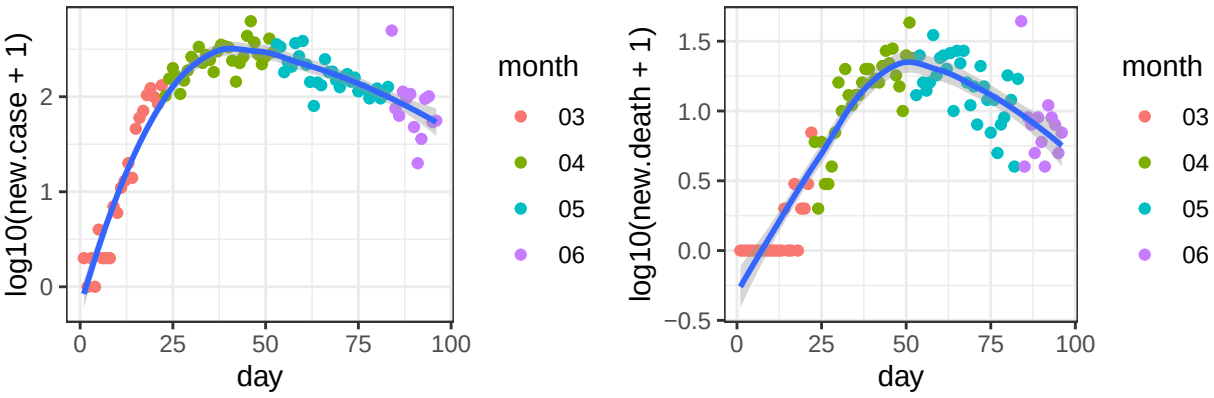
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

New Haven_Connecticut



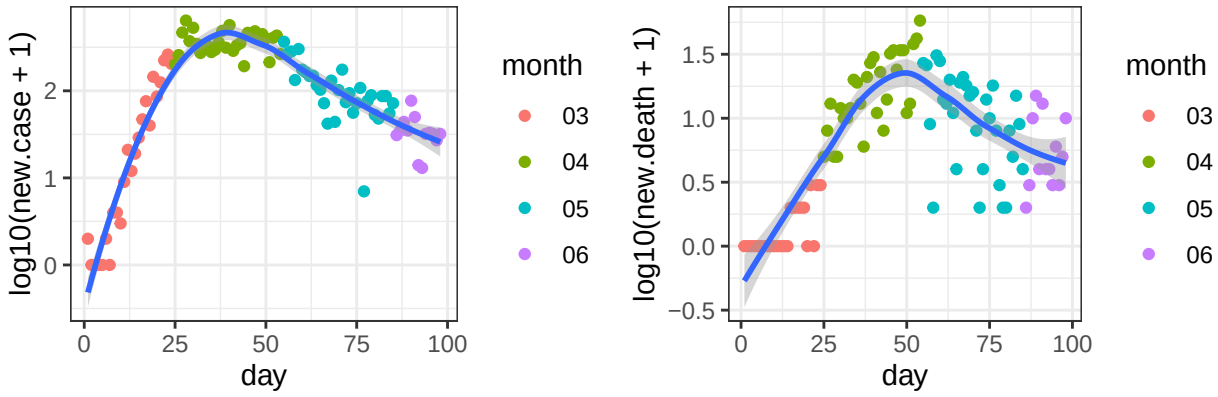
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Essex_Massachusetts



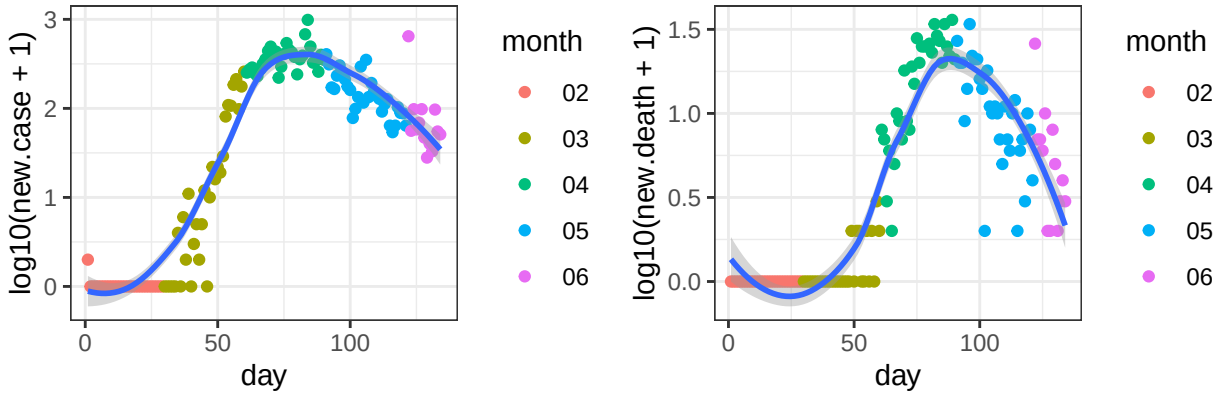
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Passaic_New Jersey



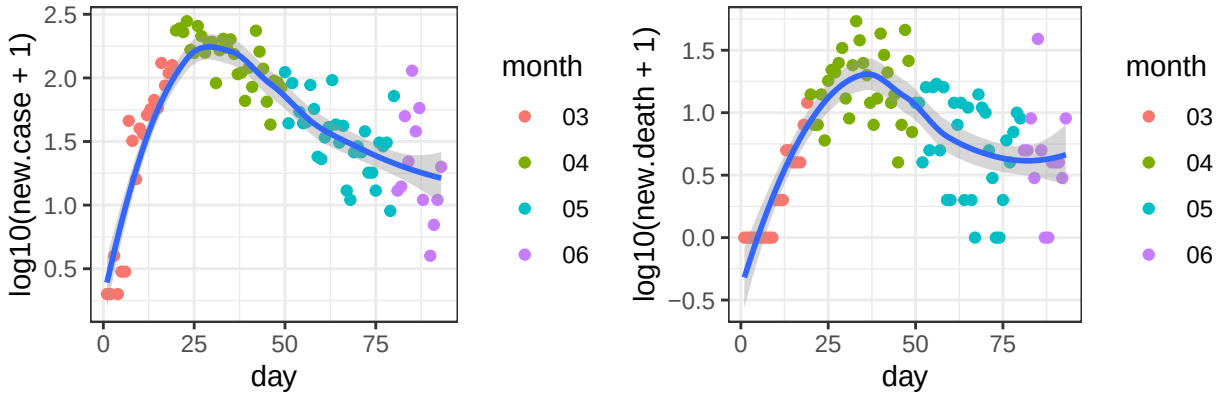
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Suffolk_Massachusetts



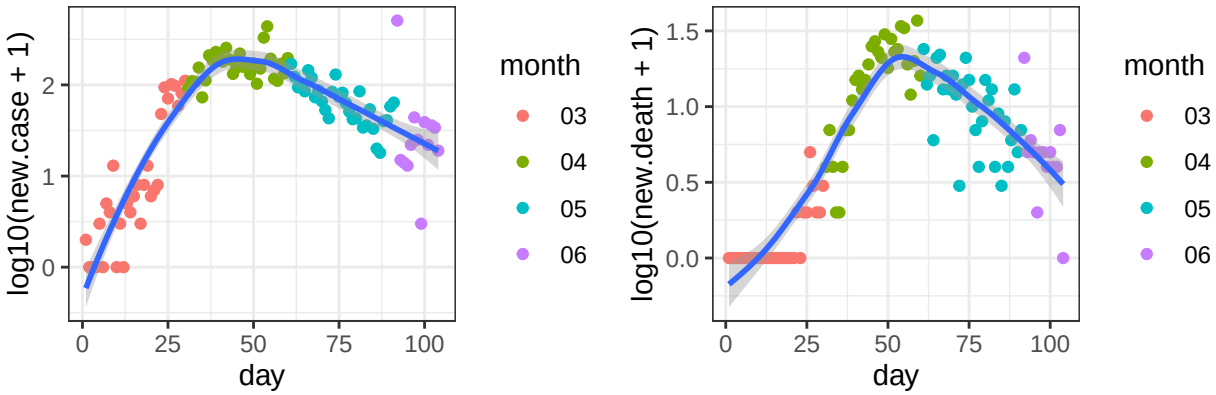
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Macomb_Michigan



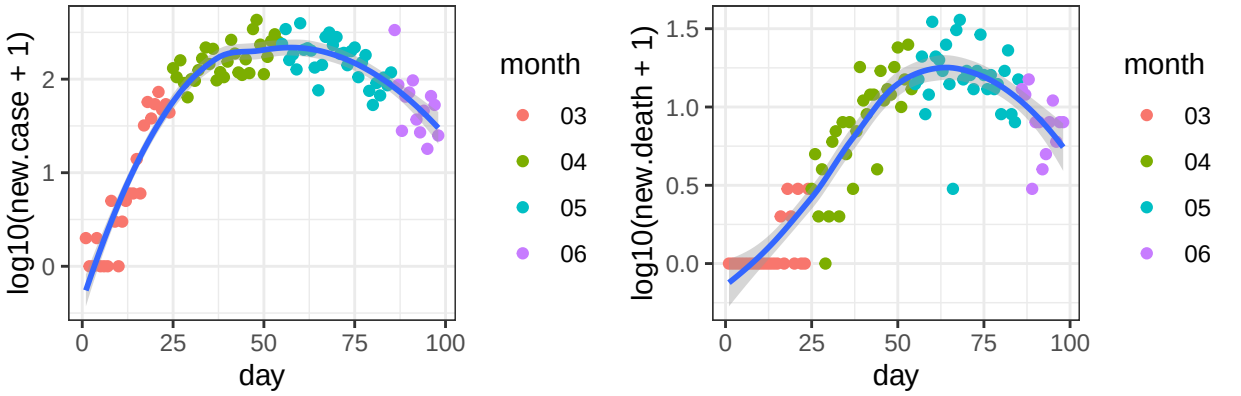
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Norfolk_Massachusetts



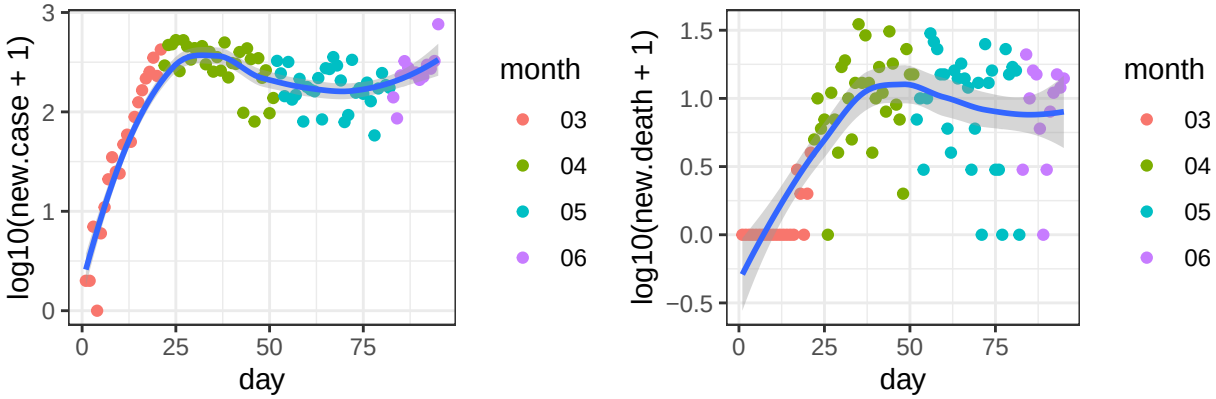
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Worcester_Massachusetts



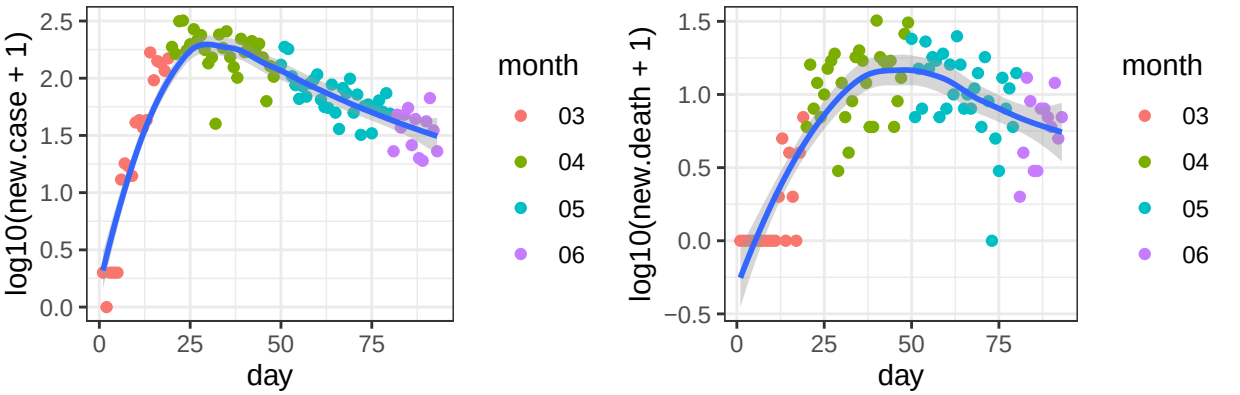
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Miami−Dade_Florida



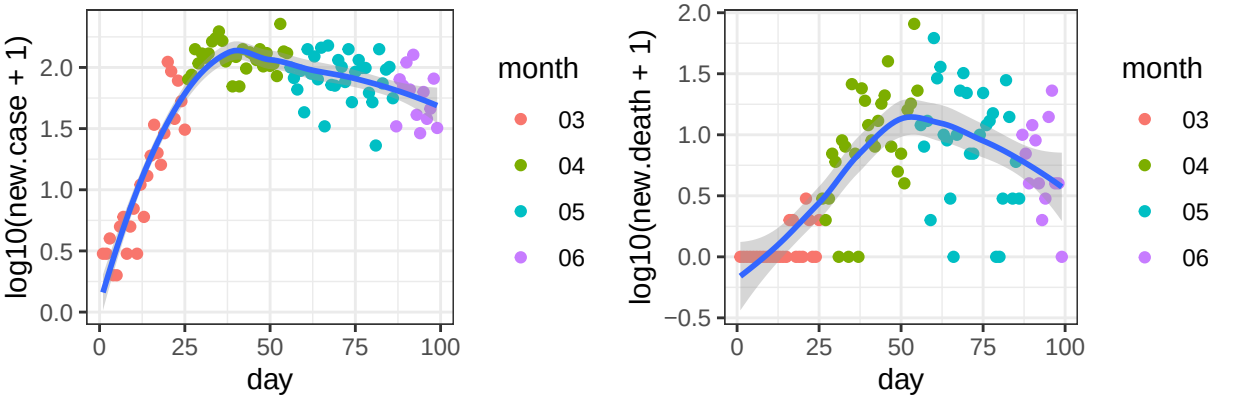
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Ocean_New Jersey



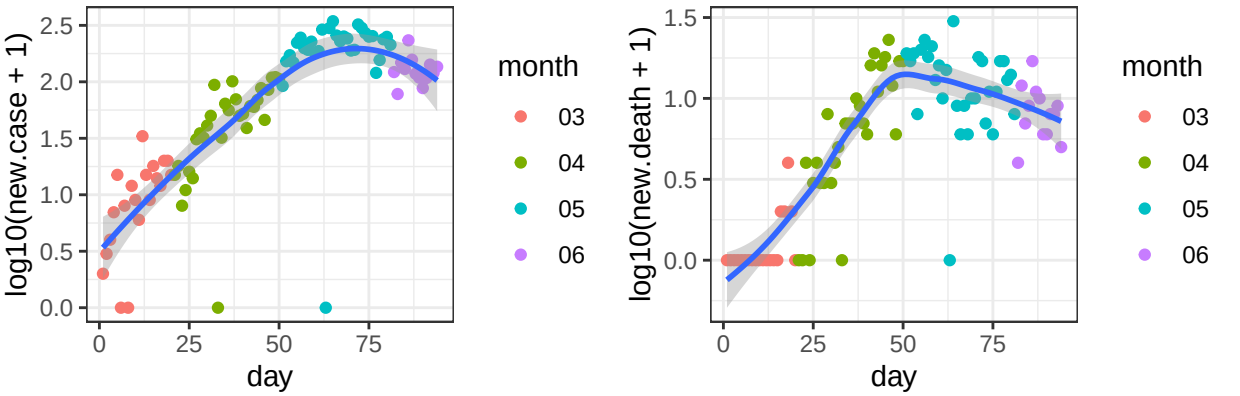
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Montgomery_Pennsylvania



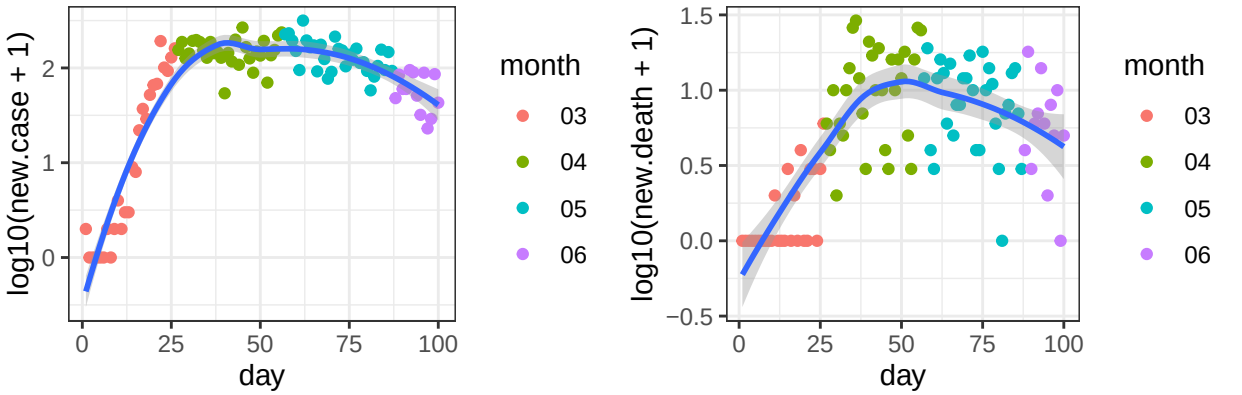
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Hennepin_Minnesota



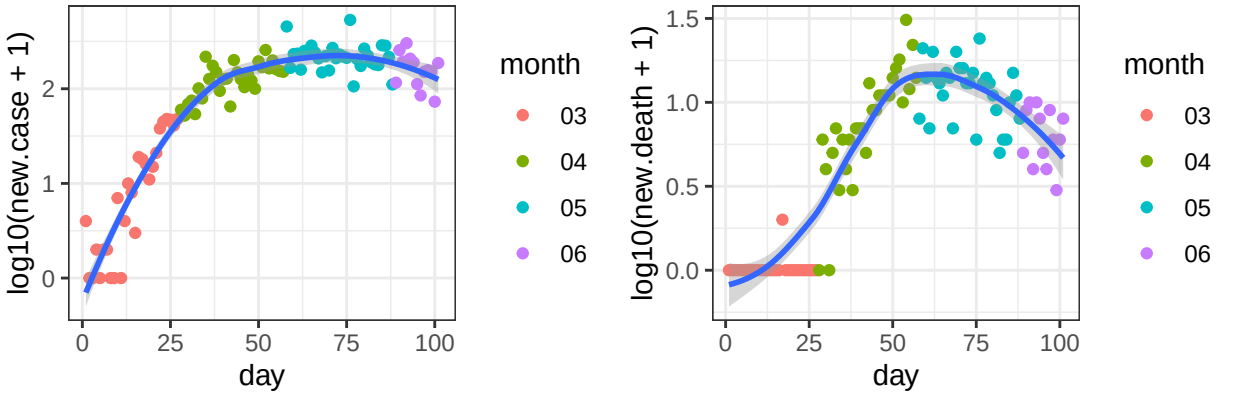
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Marion_Indiana



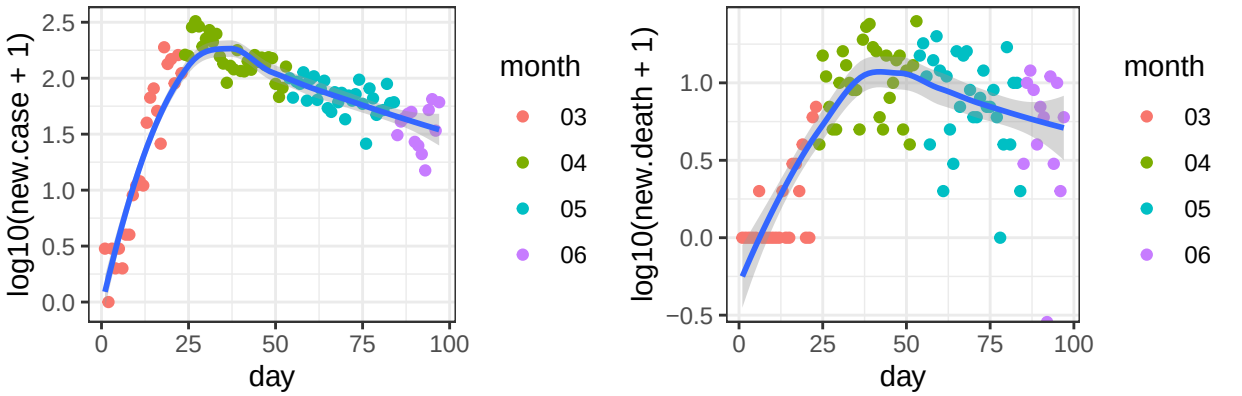
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Montgomery_Maryland



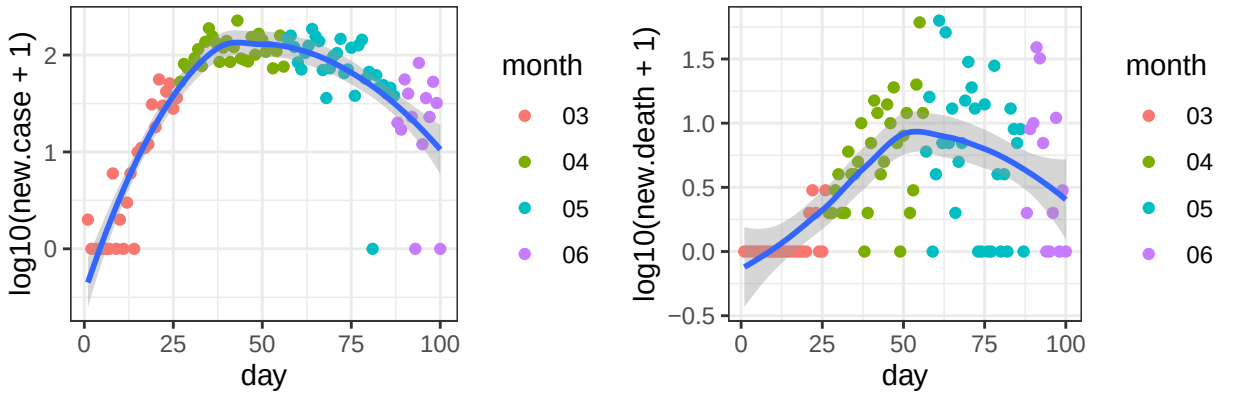
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Monmouth_New Jersey



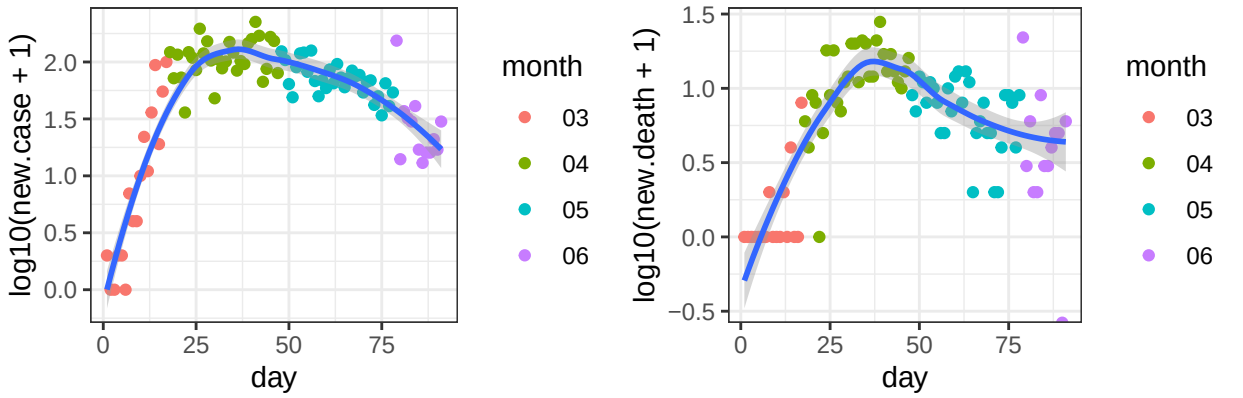
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Delaware_Pennsylvania



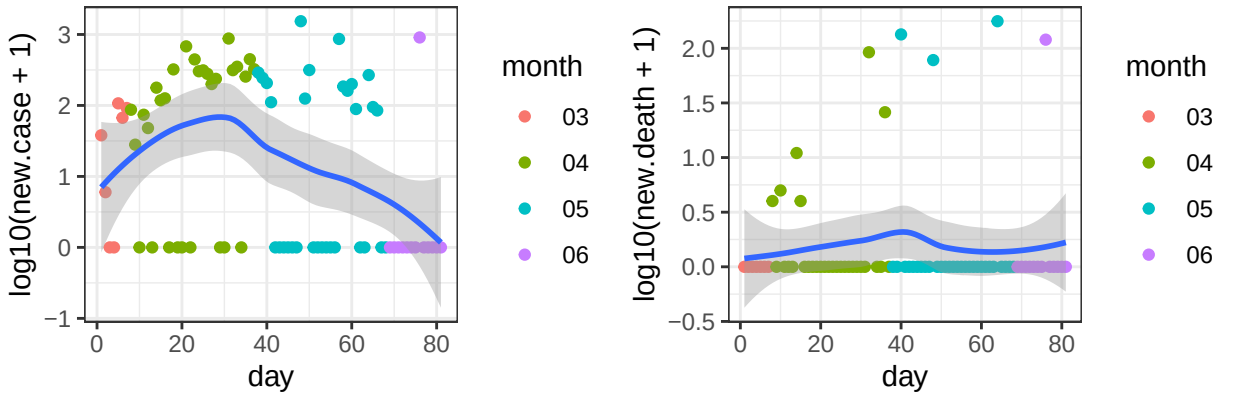
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Hampden_Massachusetts



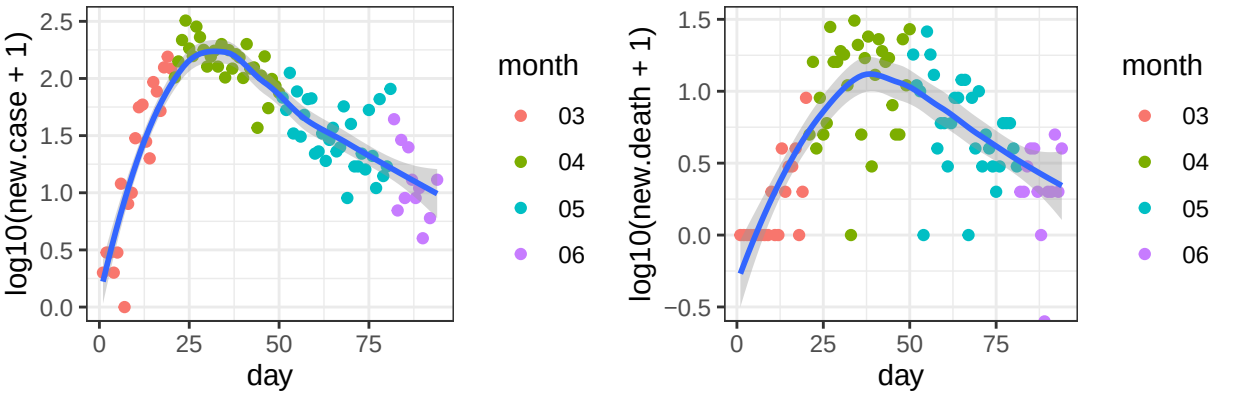
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Providence_Rhode Island



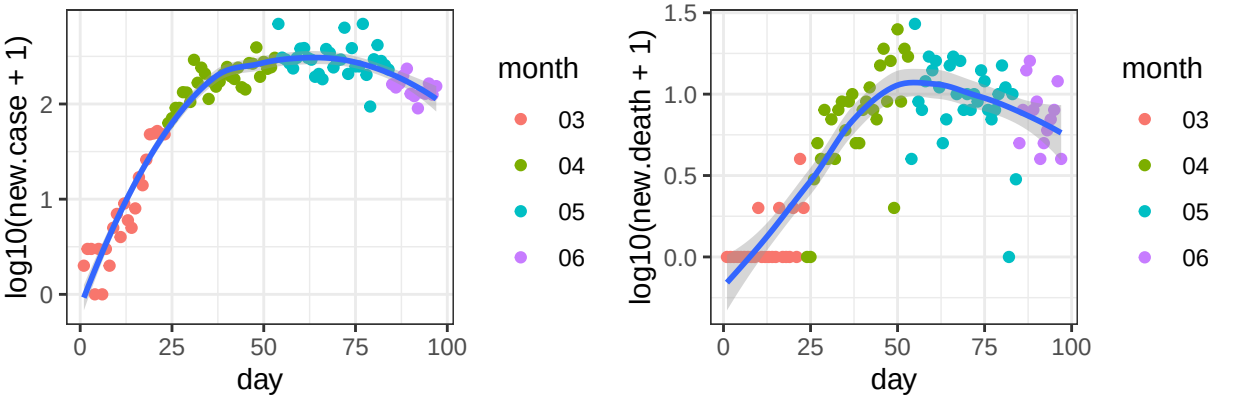
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Morris_New Jersey



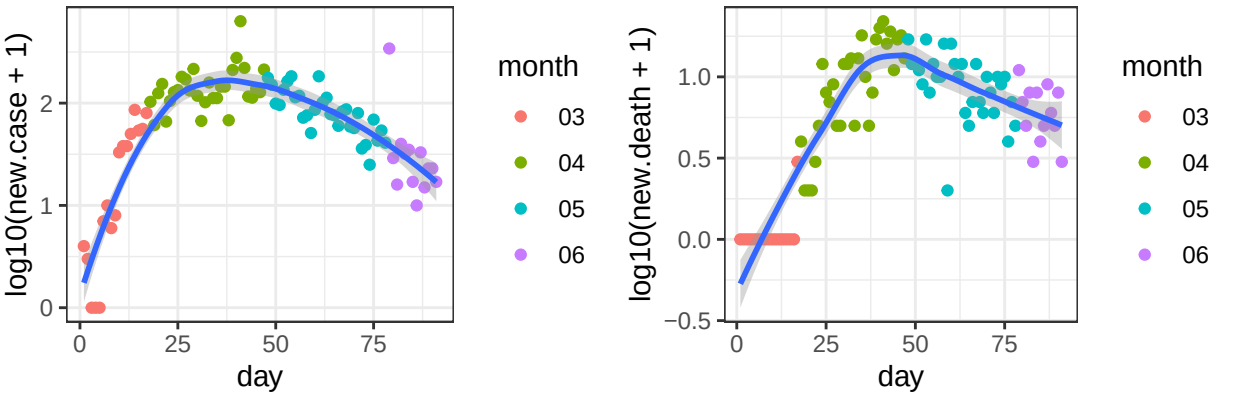
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Prince George's_Maryland



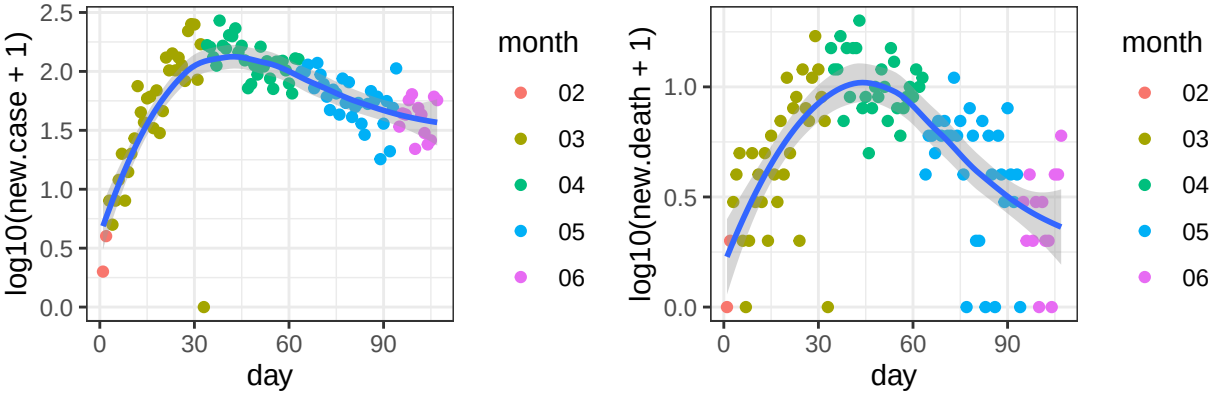
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Plymouth_Massachusetts



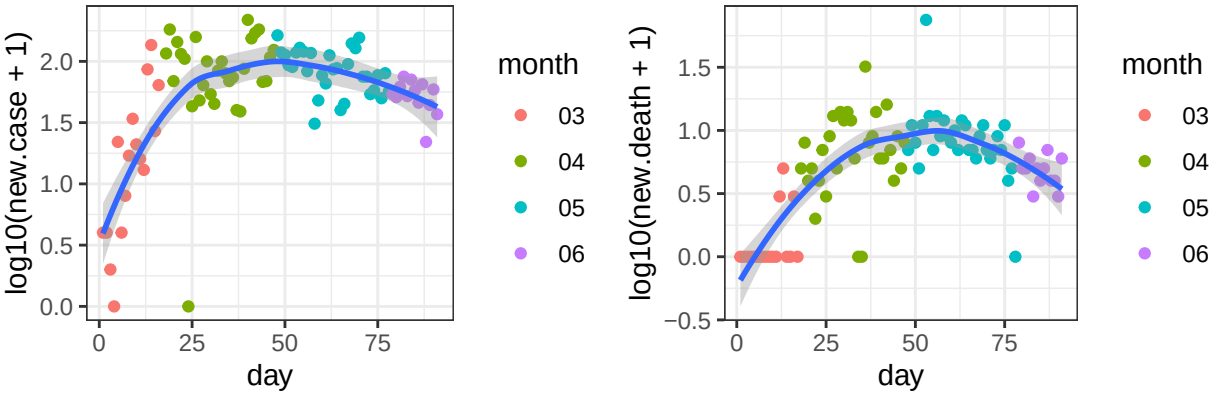
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

King_Washington



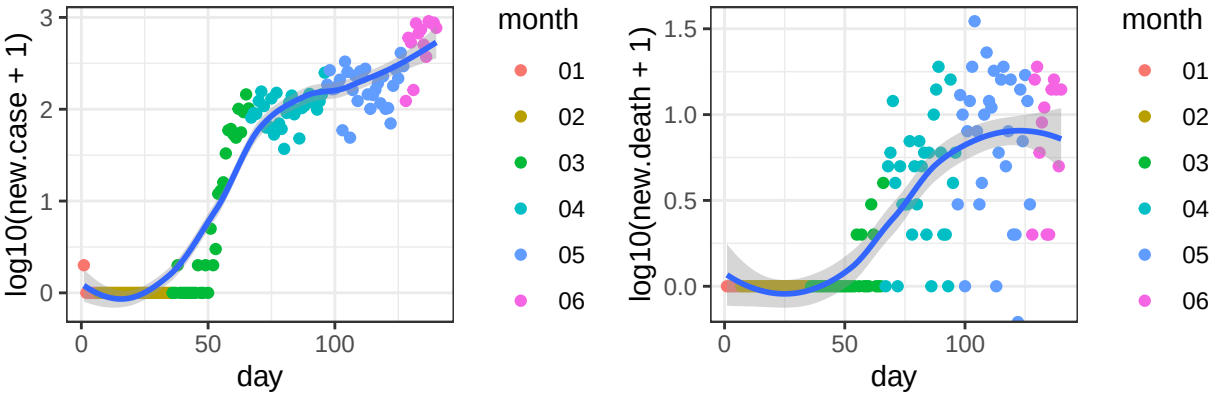
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

Erie_New York



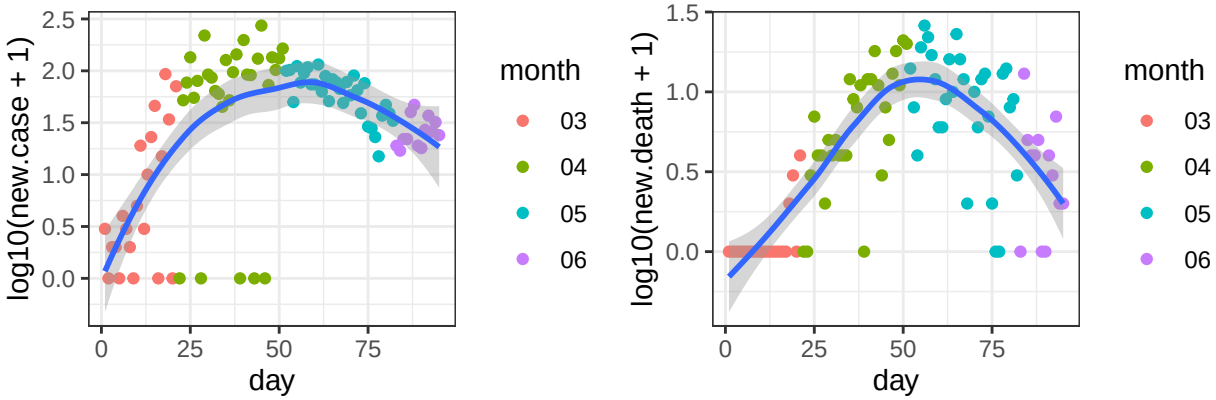
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Maricopa_Arizona



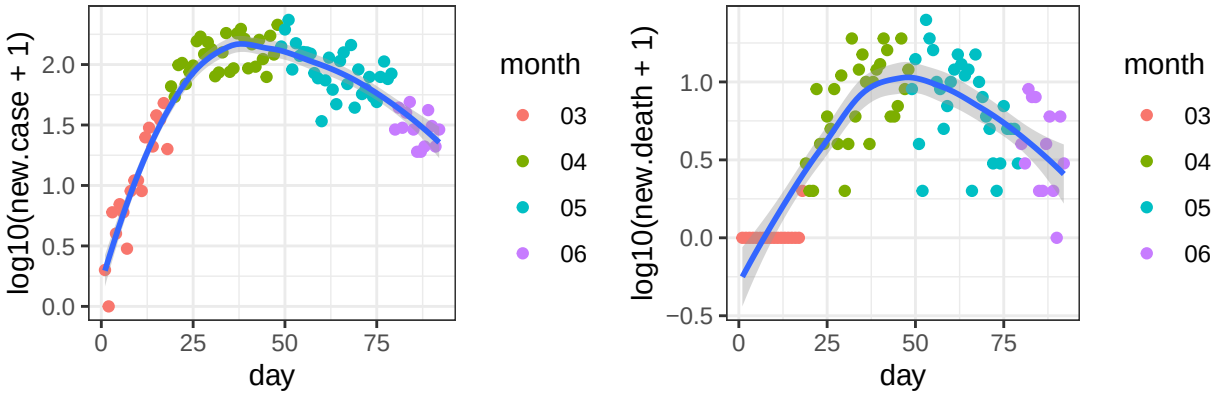
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Bucks_Pennsylvania



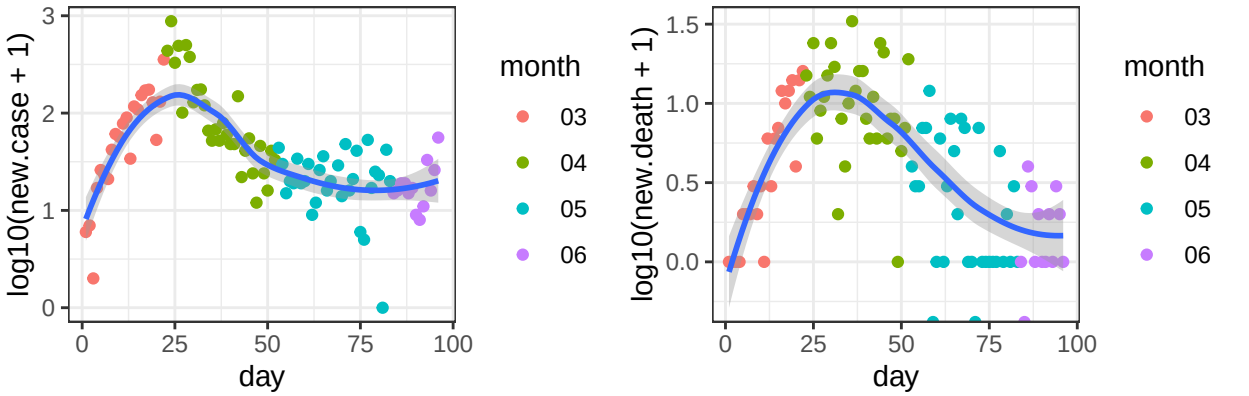
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Mercer_New Jersey



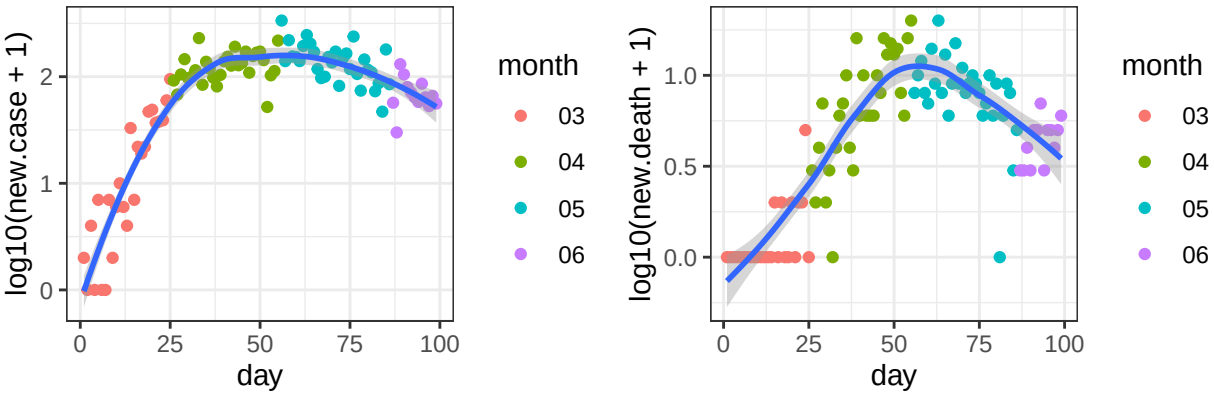
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Orleans_Louisiana



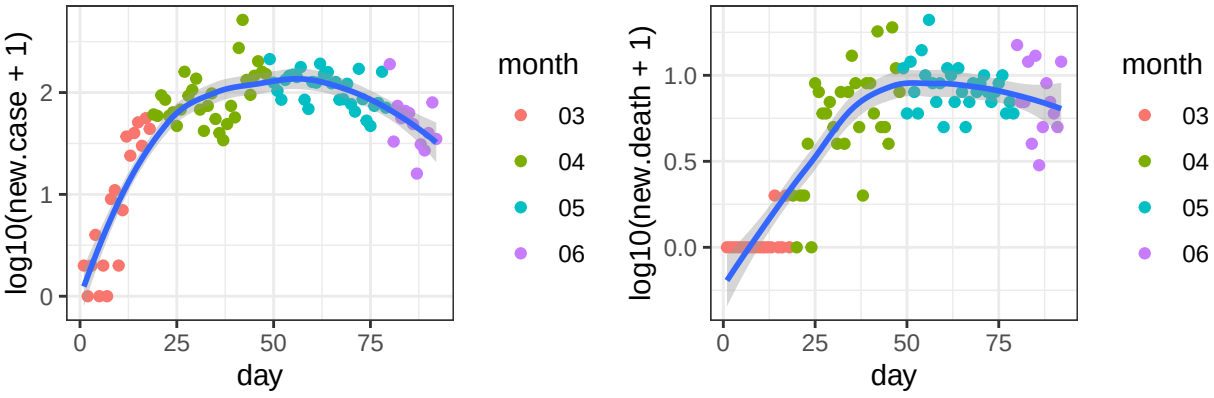
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

District of Columbia_District of Columbia



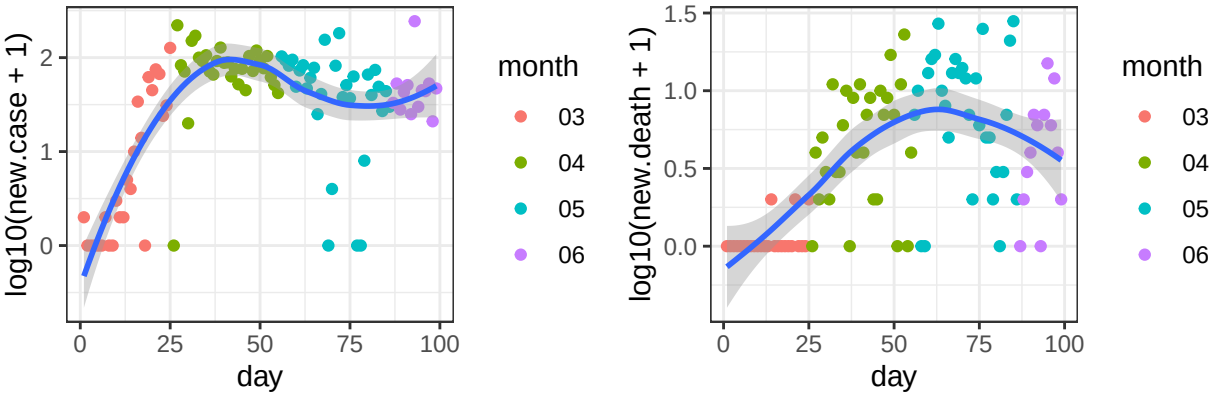
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Bristol_Massachusetts



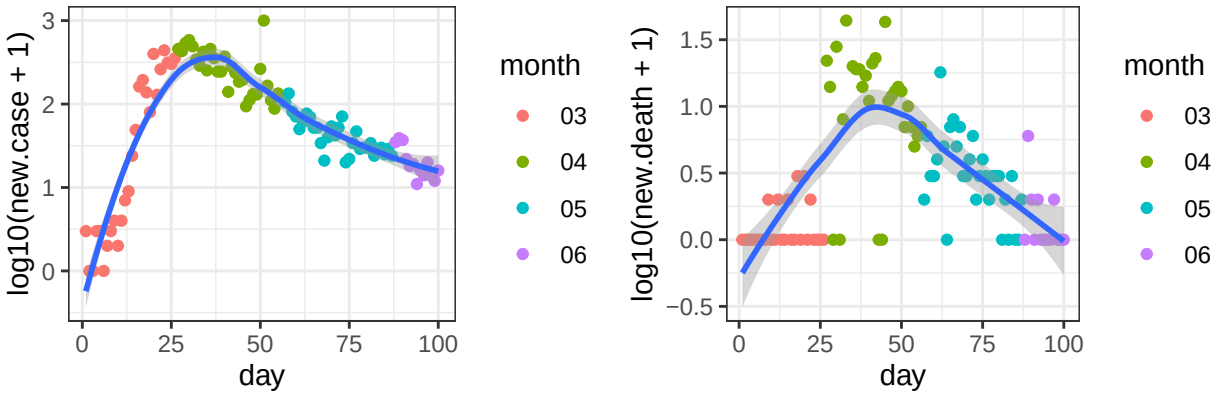
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

St. Louis_Missouri



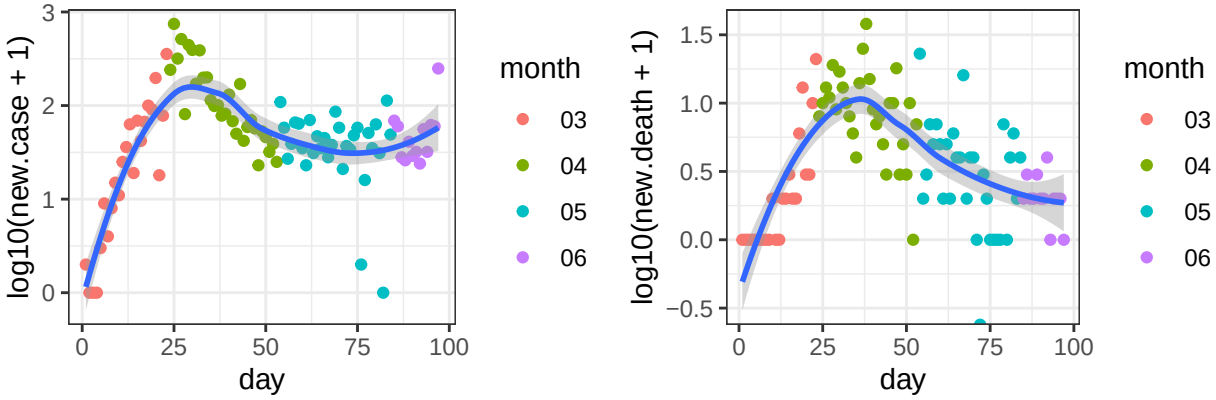
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Rockland_New York



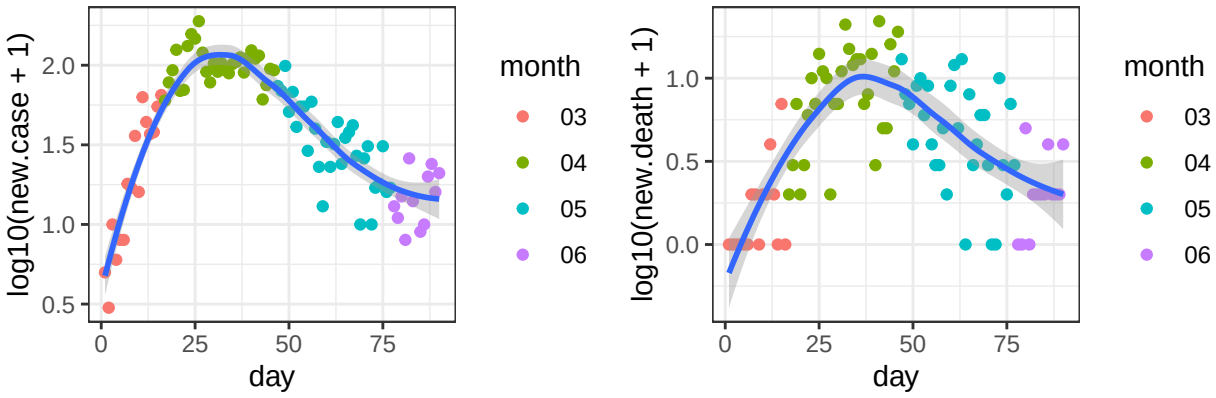
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Jefferson_Louisiana

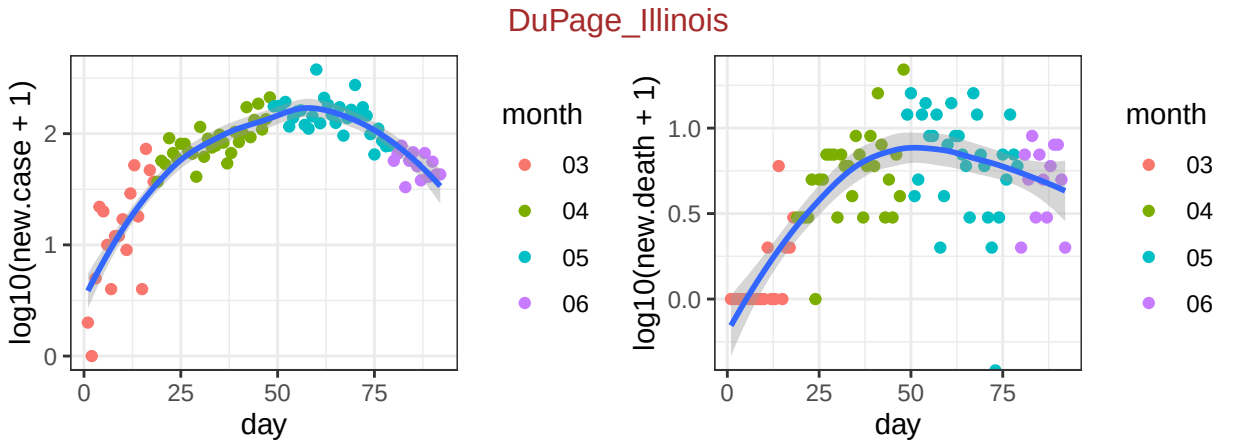


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Somerset_New Jersey



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

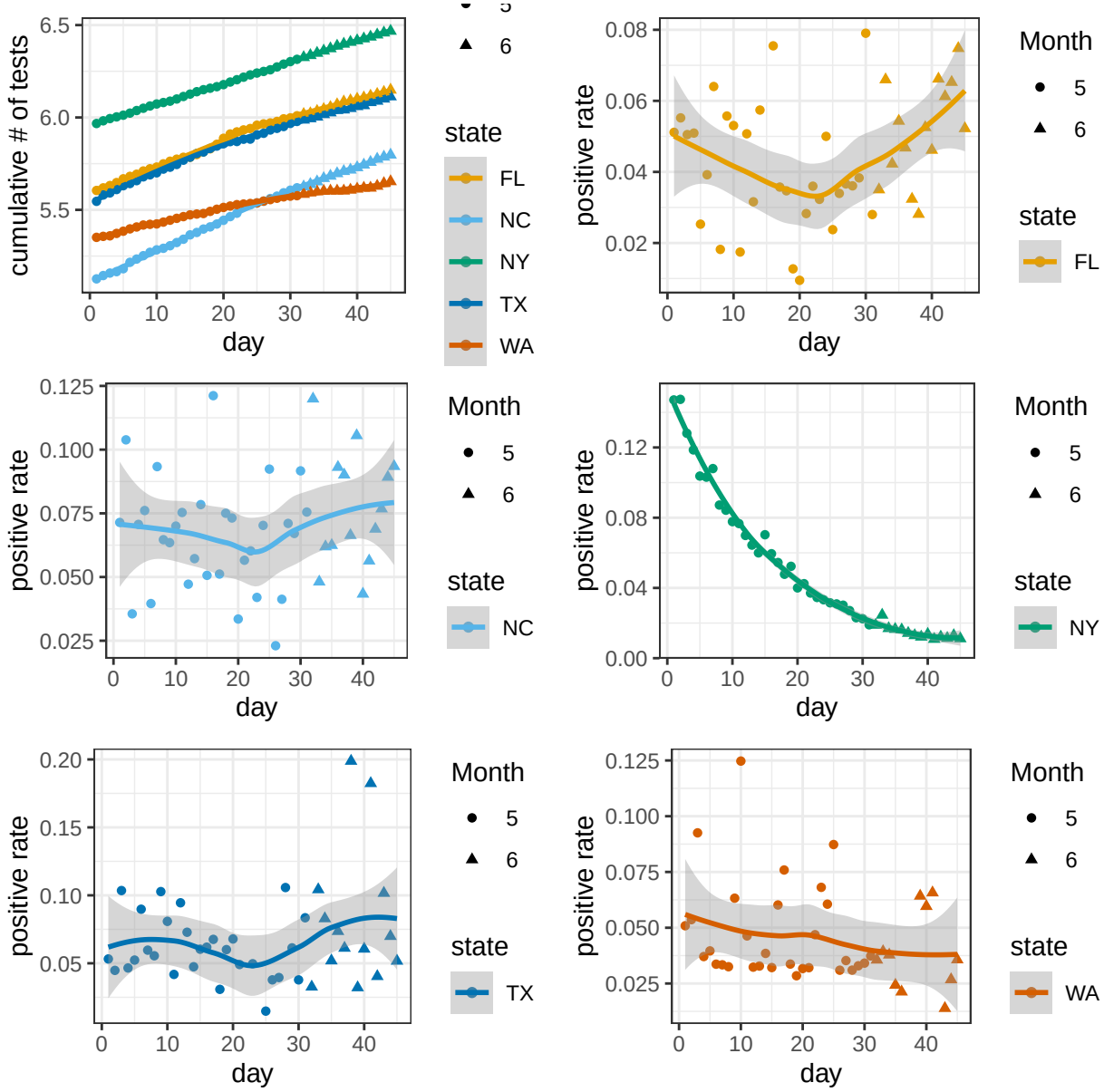


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

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 0614: 0.05(FL) 0.09(NC) 0.01(NY) 0.05(TX) 0.04(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
```