Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-08-03

Contents

Introduction **1**

JHU **2**
time series data 2
daily reports data 6

NY Times **7**
state level data 7
county level data 18

COVID Trackng **36**

Session information **39**

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

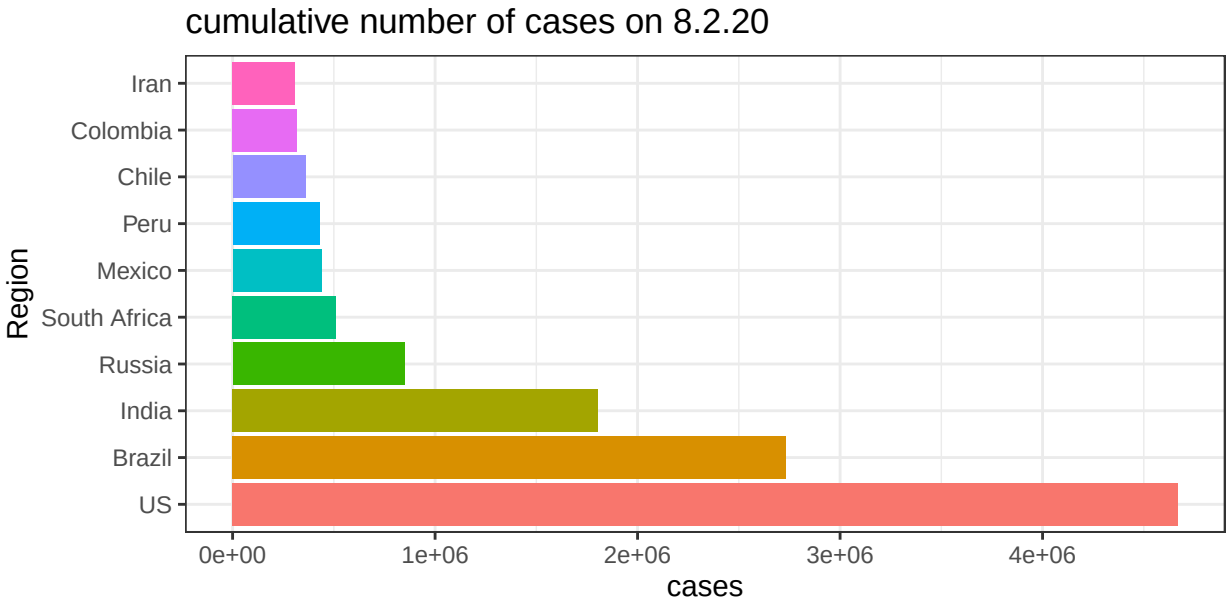
JHU

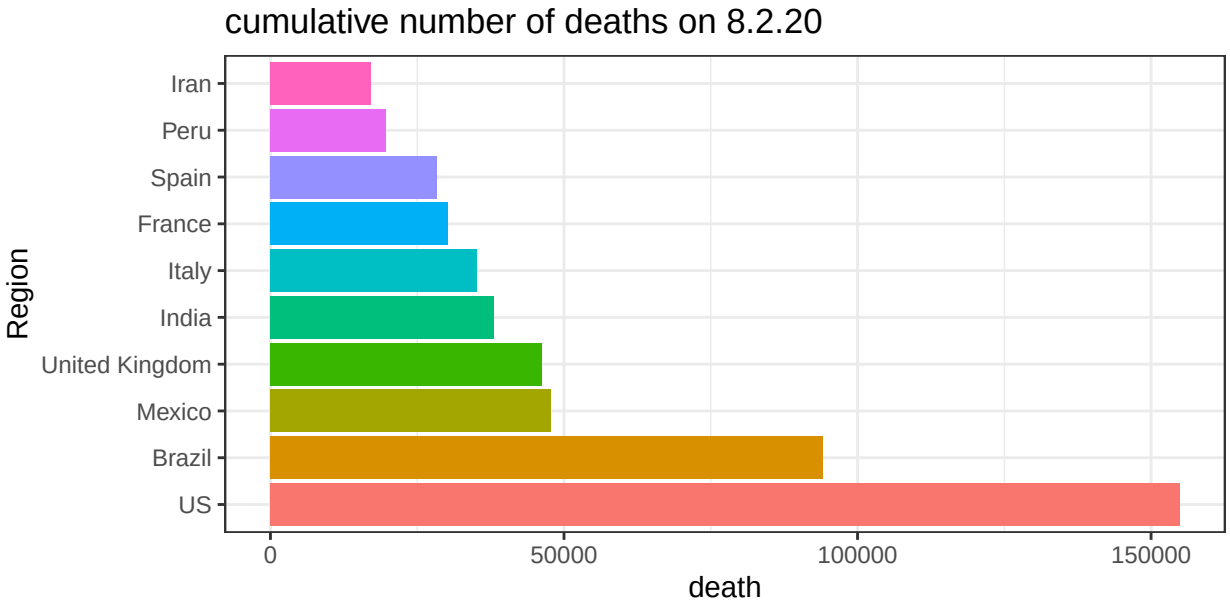
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

time series data

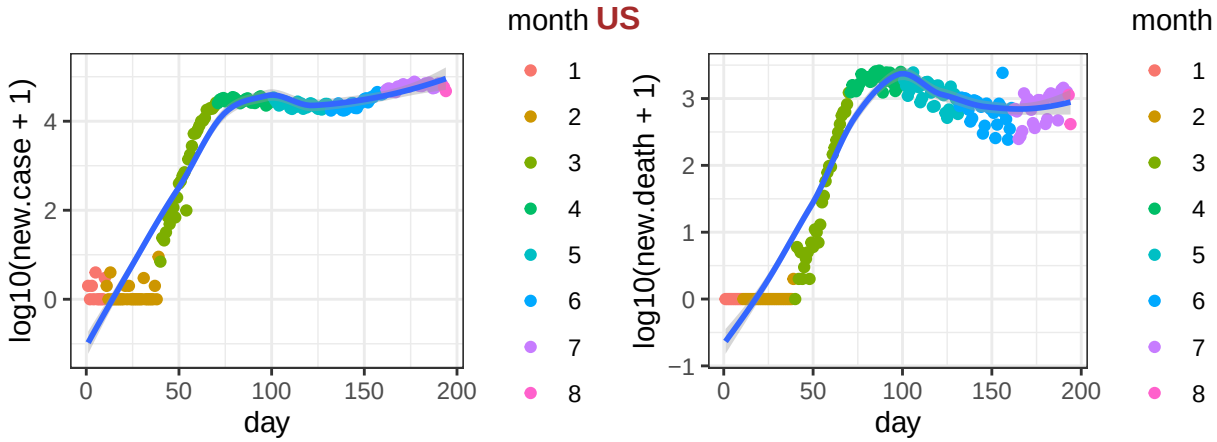
The time series provide counts (e.g., confirmed cases, deaths) starting from Jan 22nd, 2020 for 253 locations. Currently there is no data of individual US state in these time series data files.

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

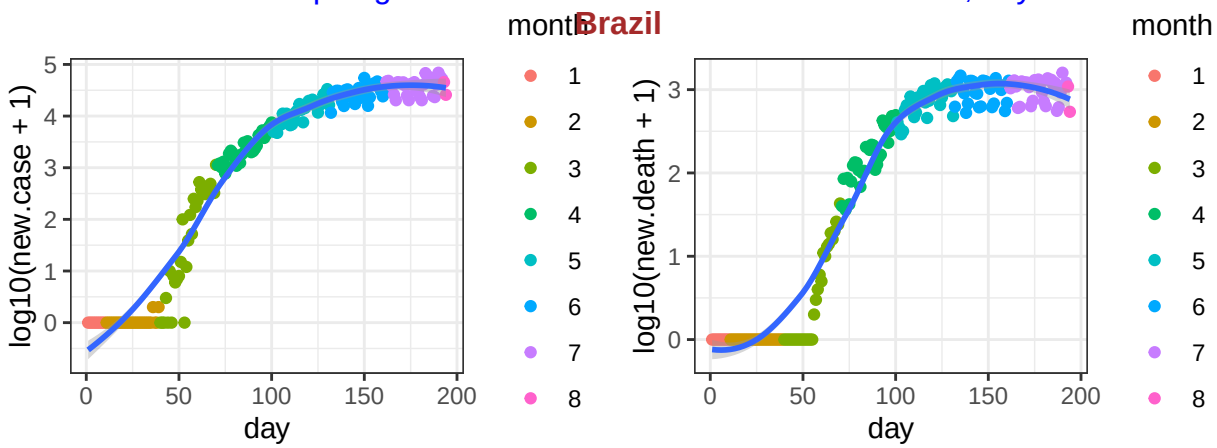




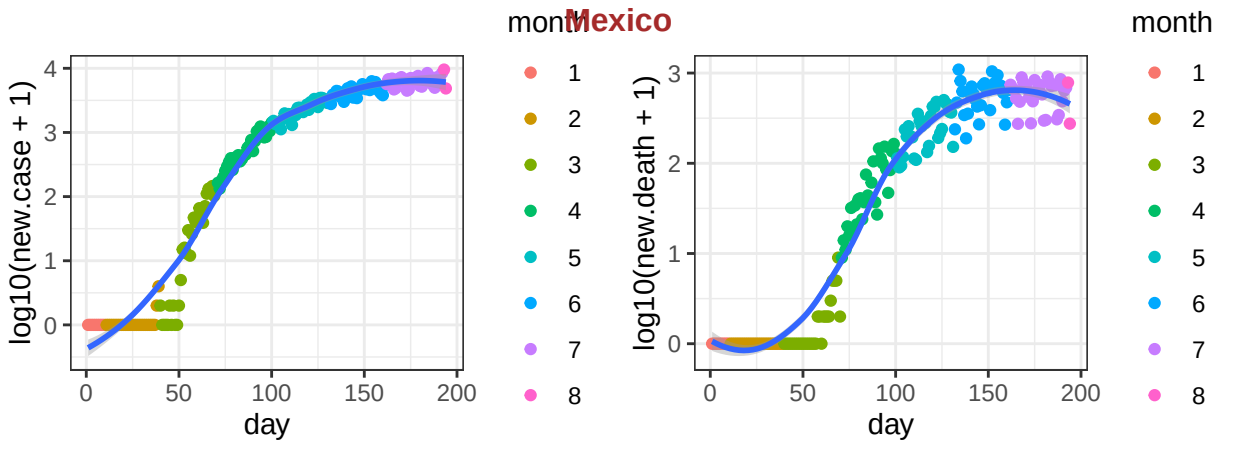
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.



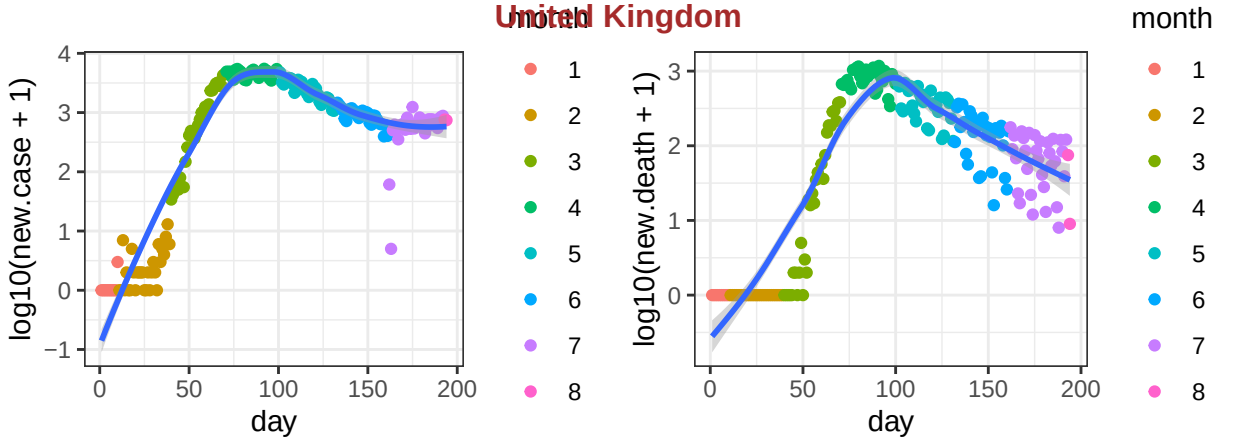
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



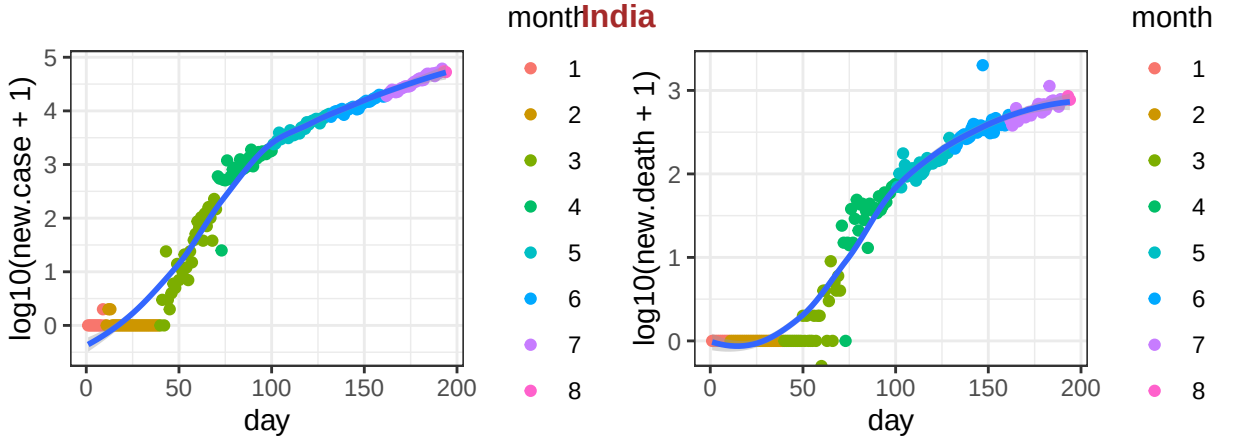
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



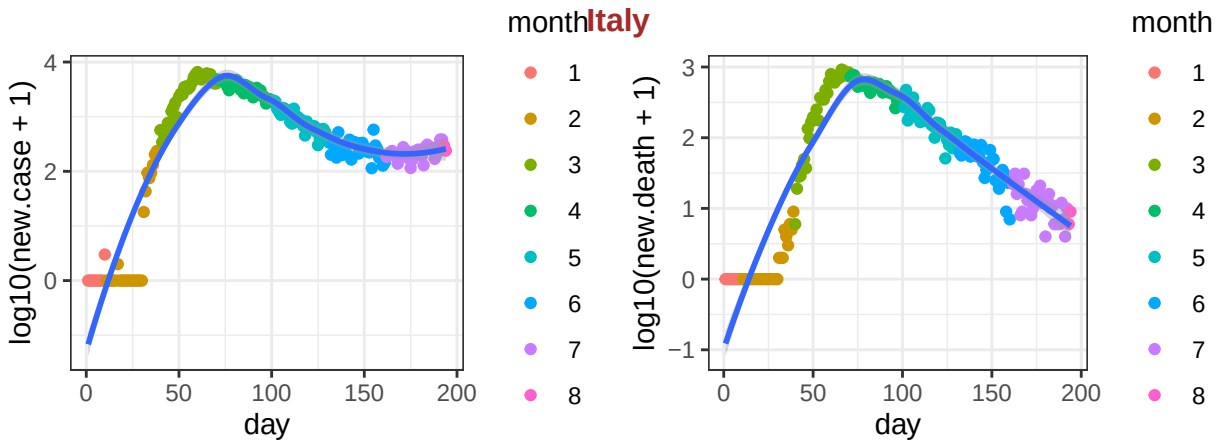
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



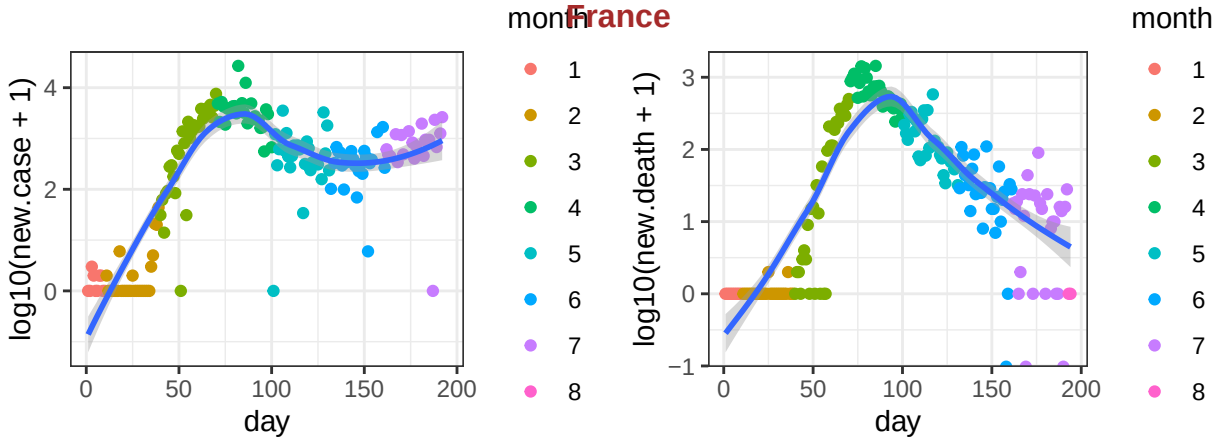
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



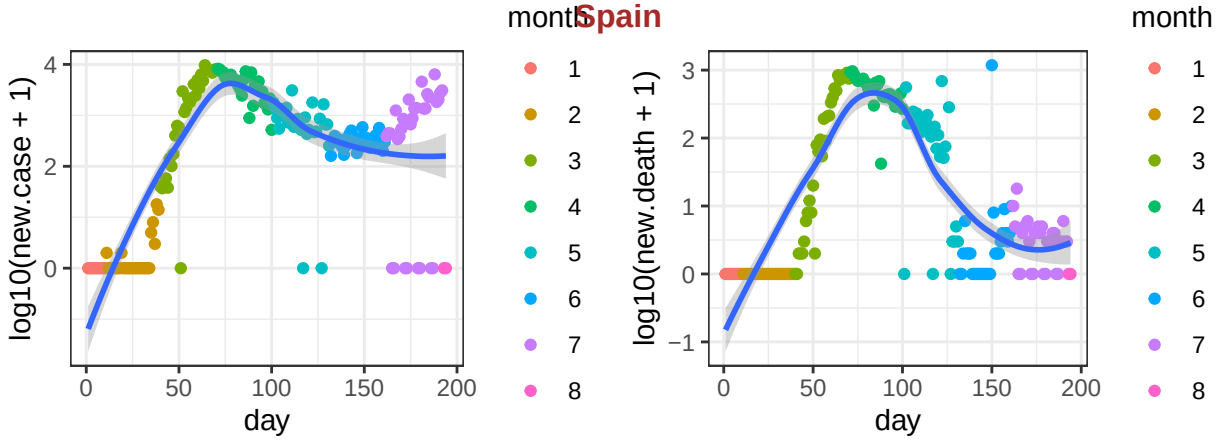
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



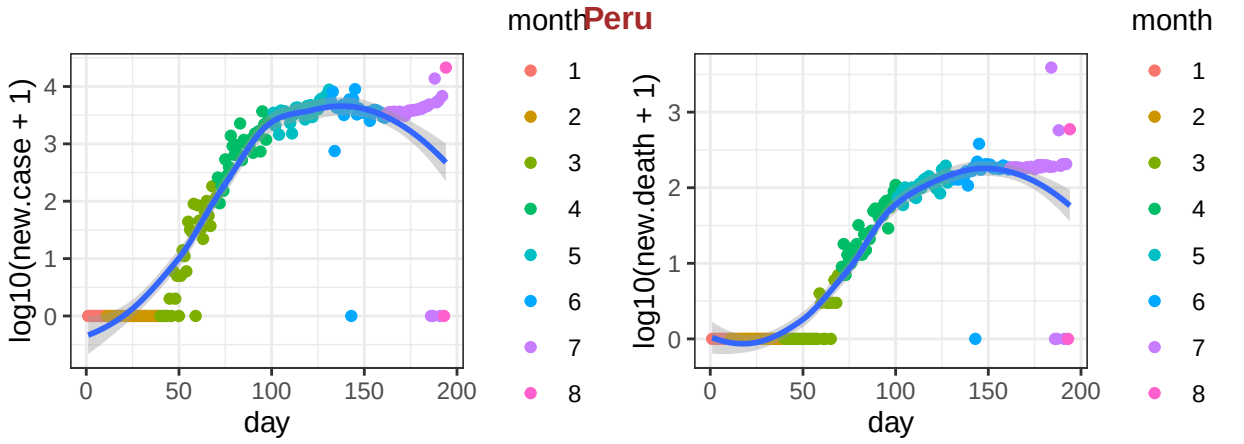
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



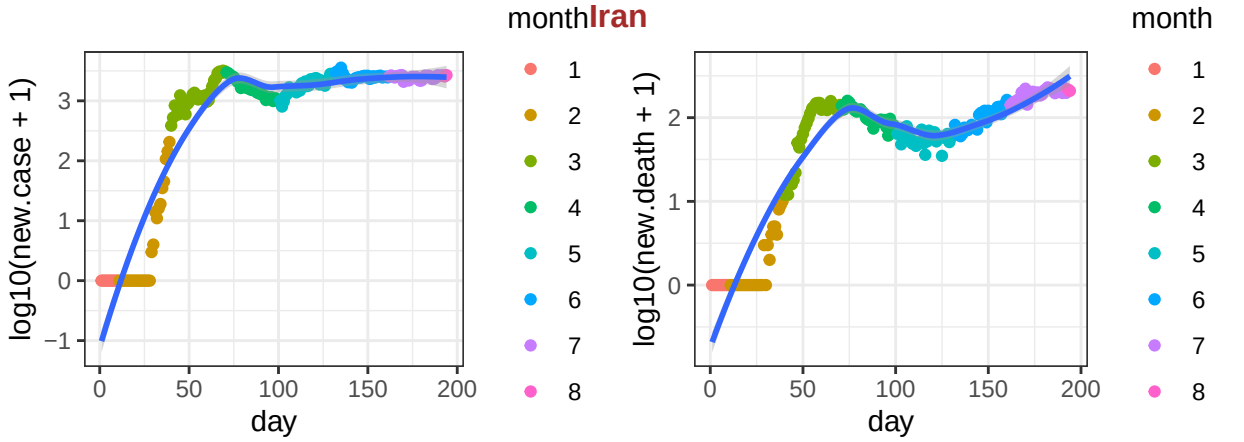
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



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



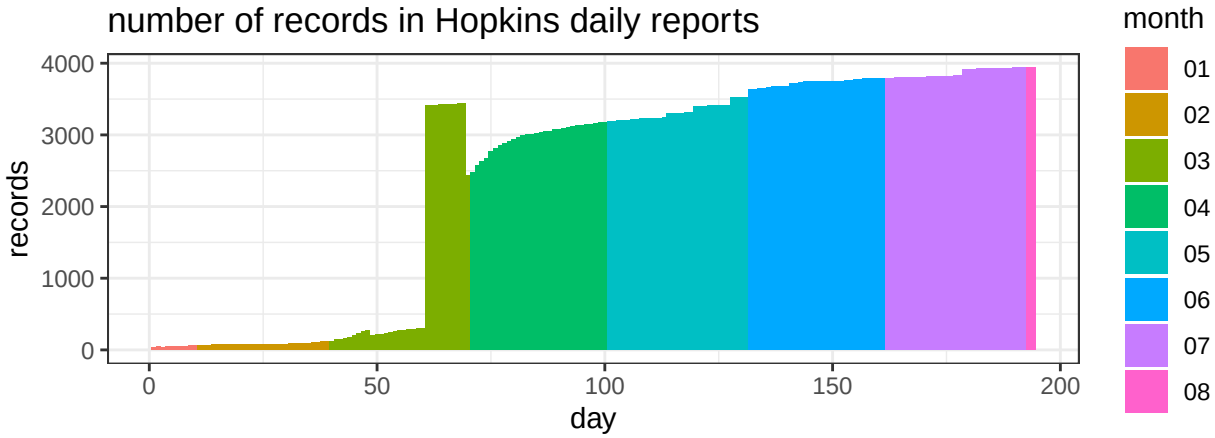
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



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] "2020-08-01"
```

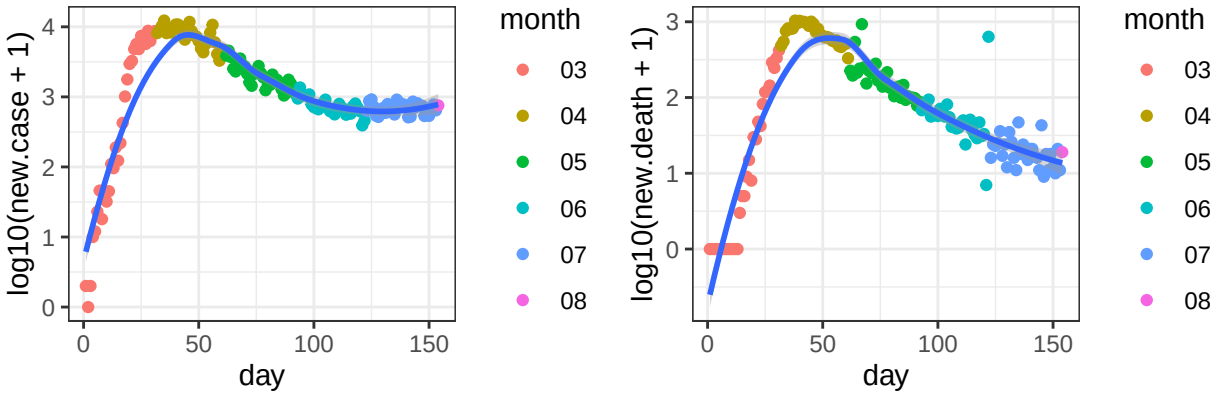
state level data

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

```
##            date          state fips  cases deaths
## 8353 2020-08-01       New York   36 420477  32390
## 8351 2020-08-01     New Jersey   34 183904  15830
## 8324 2020-08-01     California    6 509507   9365
## 8342 2020-08-01  Massachusetts   25 118040   8626
## 8334 2020-08-01       Illinois   17 182232   7707
## 8366 2020-08-01          Texas   48 448182   7471
## 8360 2020-08-01   Pennsylvania   42 117468   7270
## 8329 2020-08-01        Florida   12 480020   7021
## 8343 2020-08-01       Michigan   26  91450   6460
## 8326 2020-08-01    Connecticut    9  49810   4432
## 8339 2020-08-01      Louisiana   22 116394   3949
## 8322 2020-08-01        Arizona    4 177019   3753
## 8330 2020-08-01        Georgia   13 174834   3744
## 8357 2020-08-01           Ohio   39  92087   3515
## 8341 2020-08-01       Maryland   24  89925   3506
## 8335 2020-08-01        Indiana   18  68773   2971
## 8370 2020-08-01       Virginia   51  90801   2215
## 8354 2020-08-01 North Carolina   37 124078   1989
## 8325 2020-08-01       Colorado    8  47357   1846
## 8363 2020-08-01 South Carolina   45  90599   1751
## 8345 2020-08-01    Mississippi   28  59881   1693
## 8371 2020-08-01     Washington   53  59649   1676
## 8344 2020-08-01      Minnesota   27  55228   1646
## 8320 2020-08-01        Alabama    1  89349   1603
## 8346 2020-08-01       Missouri   29  51985   1311
## 8365 2020-08-01      Tennessee   47 105455   1056
## 8362 2020-08-01   Rhode Island   44  19022   1007
## 8373 2020-08-01      Wisconsin   55  58064    955
## 8336 2020-08-01           Iowa   19  45293    874
## 8349 2020-08-01         Nevada   32  49207    832
```

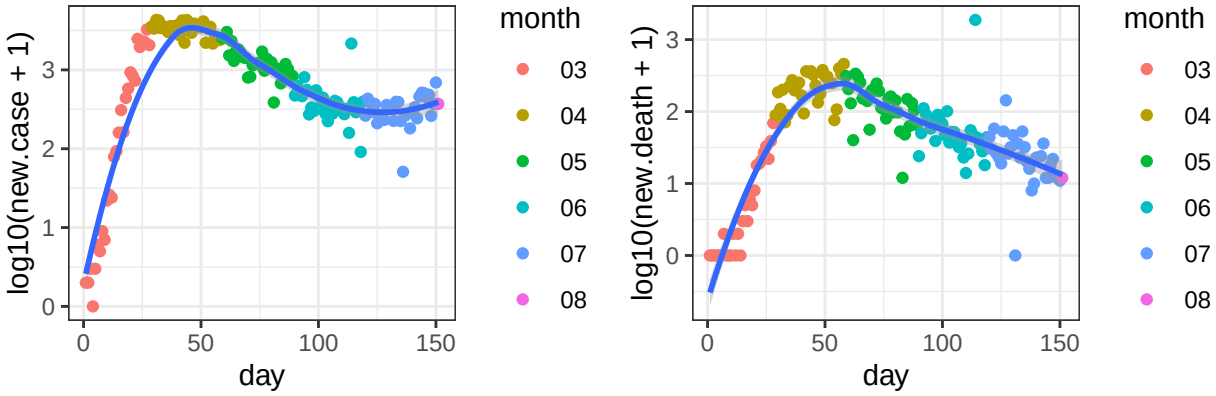
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



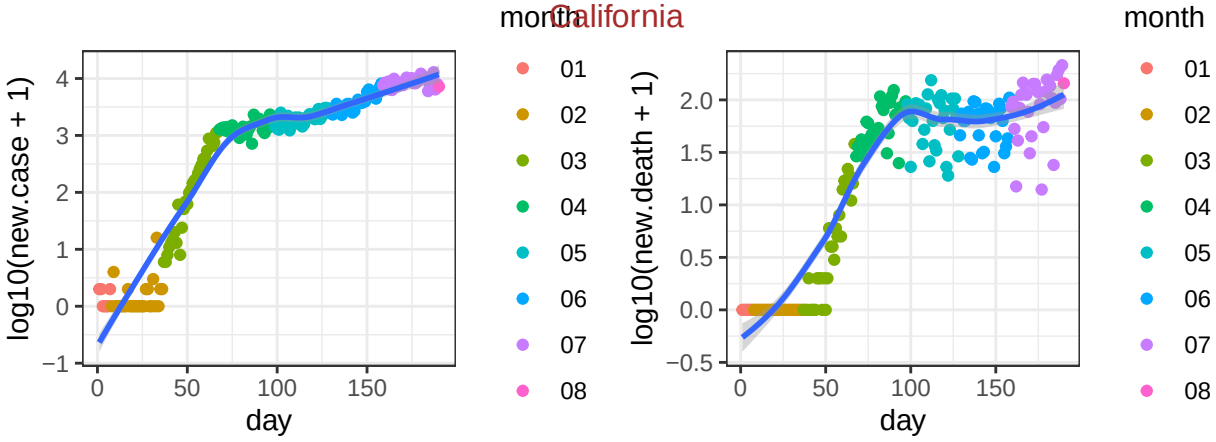
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



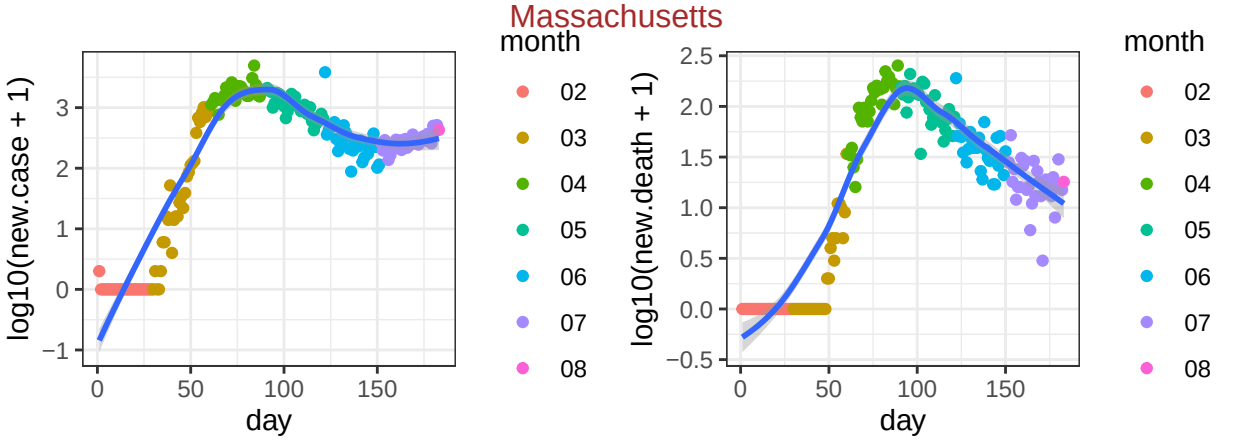
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

California



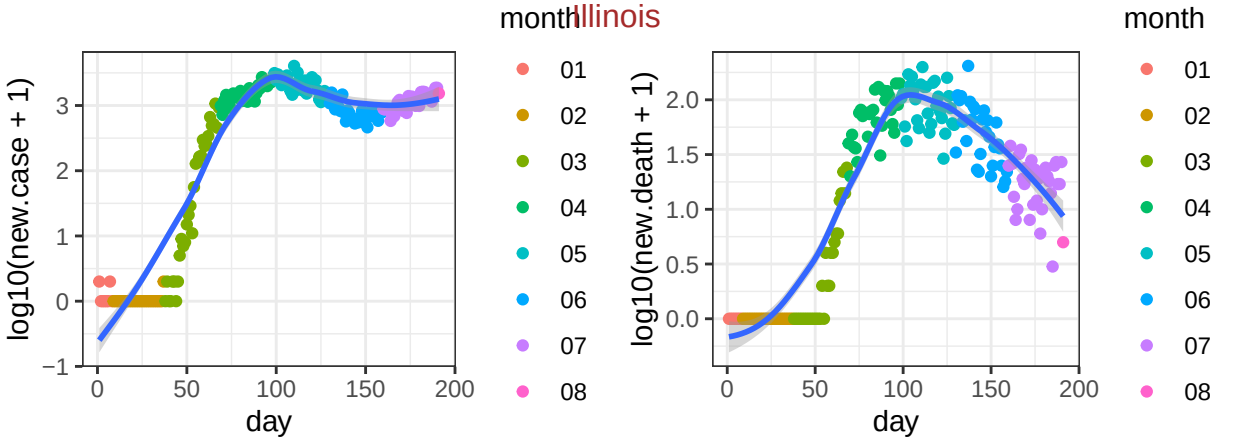
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Massachusetts



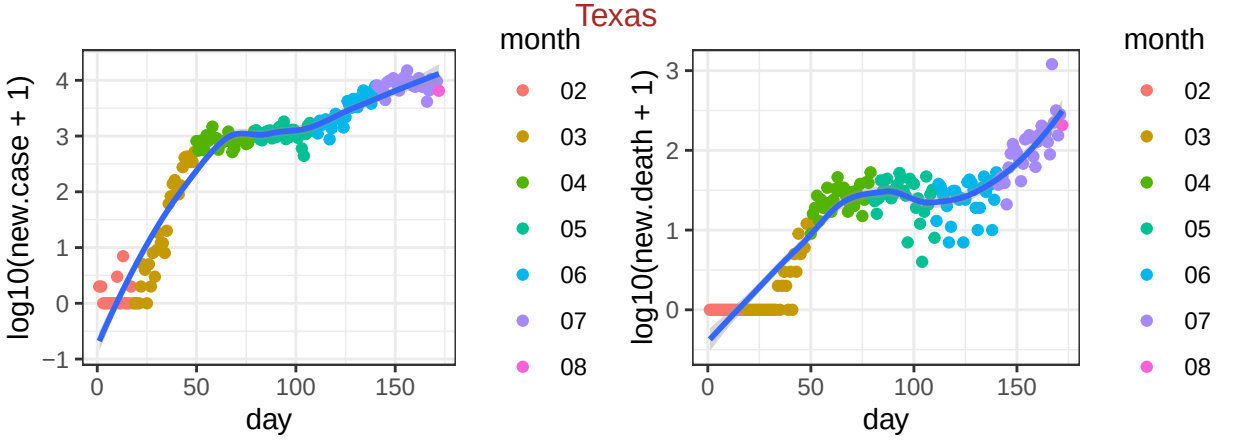
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Illinois



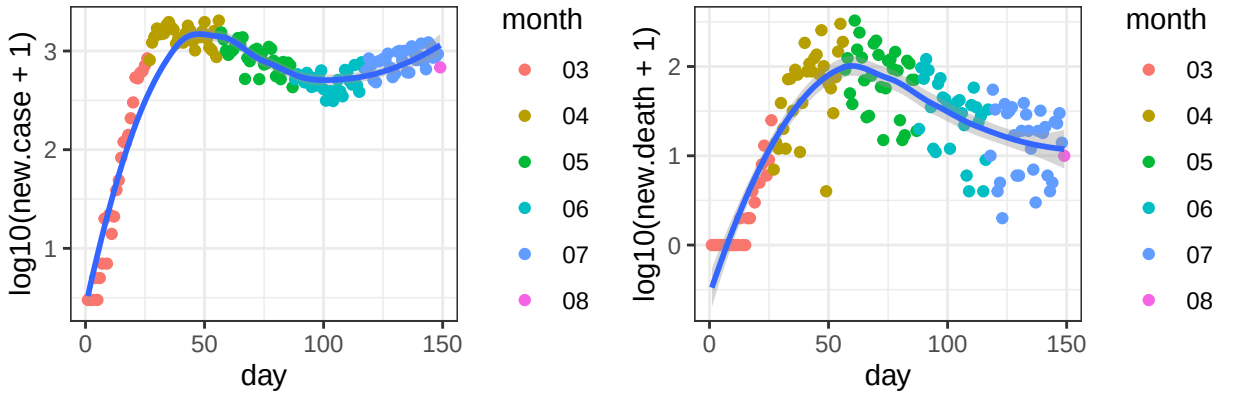
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Texas



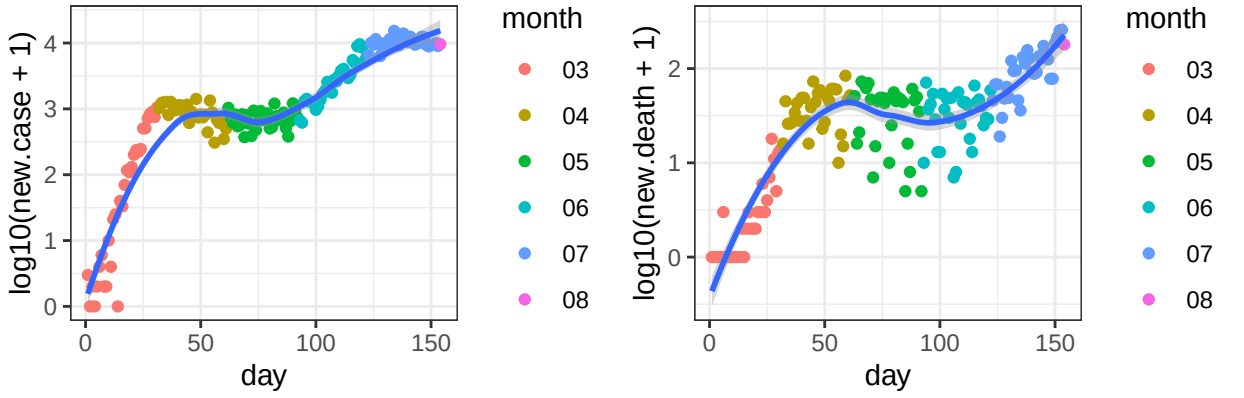
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Pennsylvania



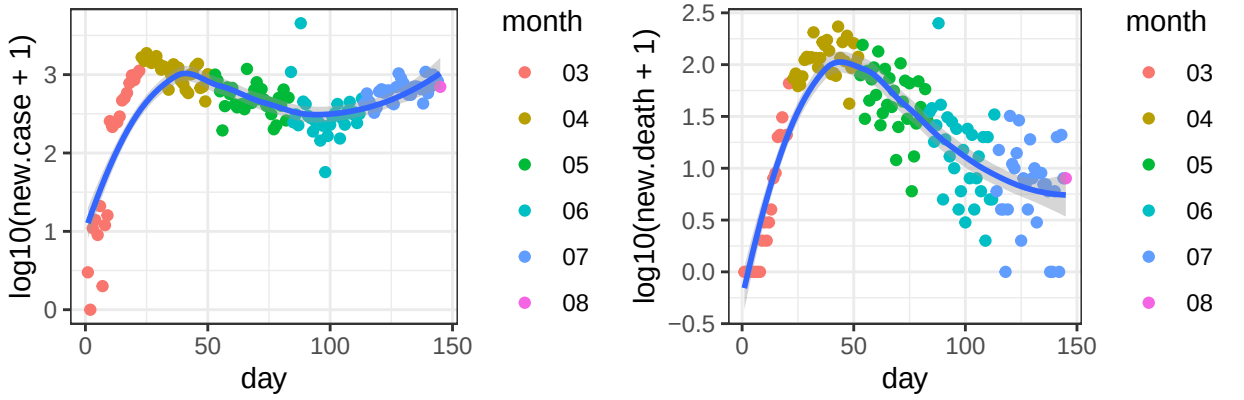
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Florida



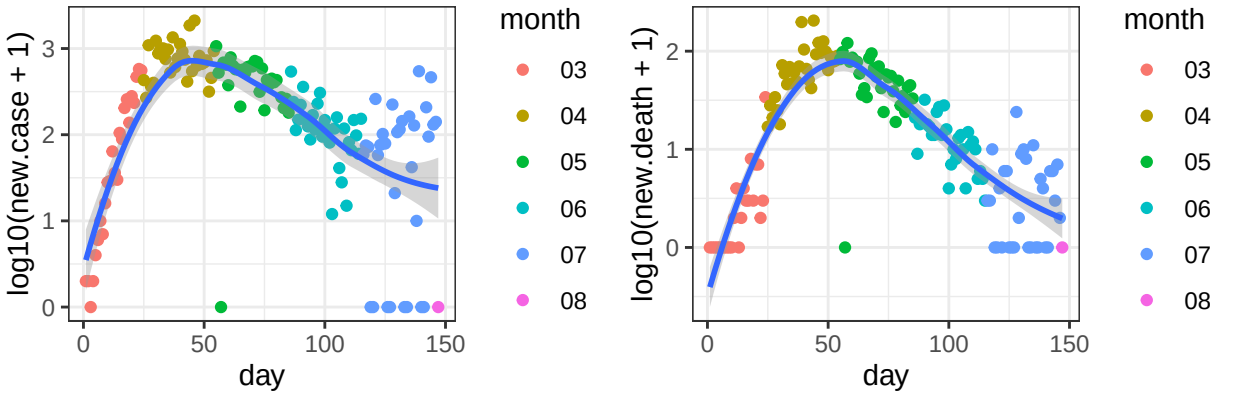
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Michigan



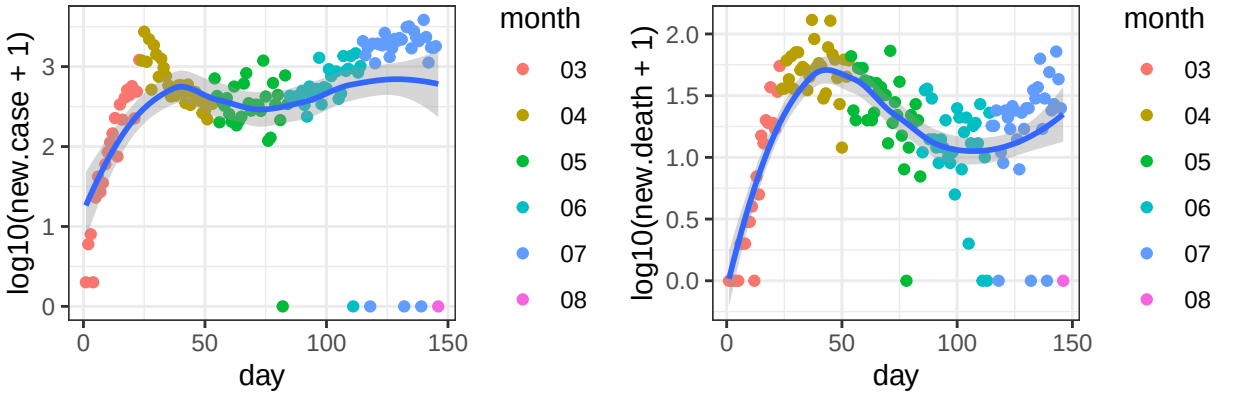
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Connecticut



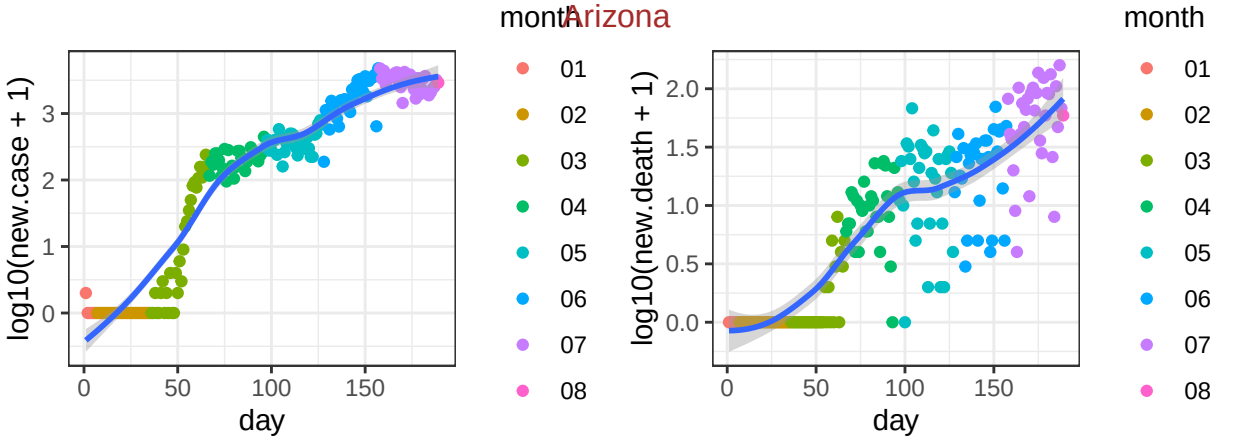
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Louisiana



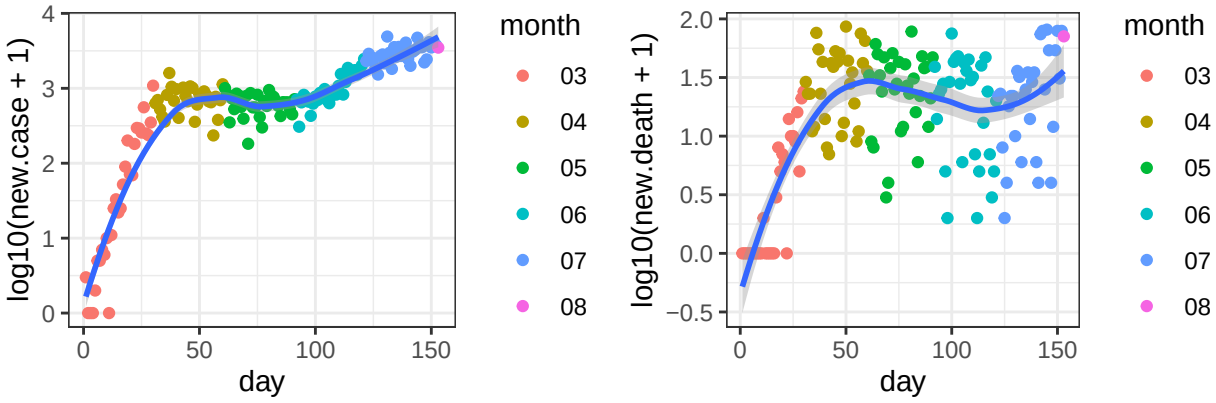
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Arizona



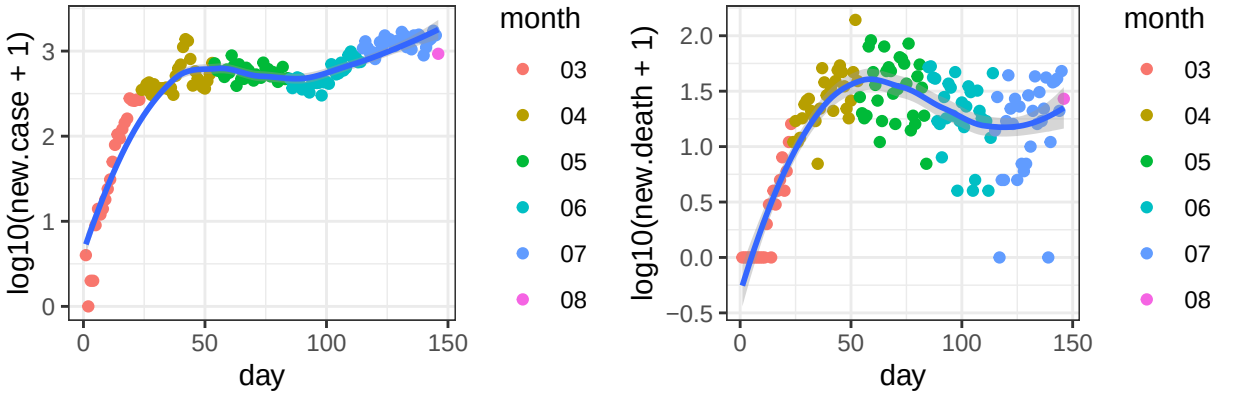
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Georgia



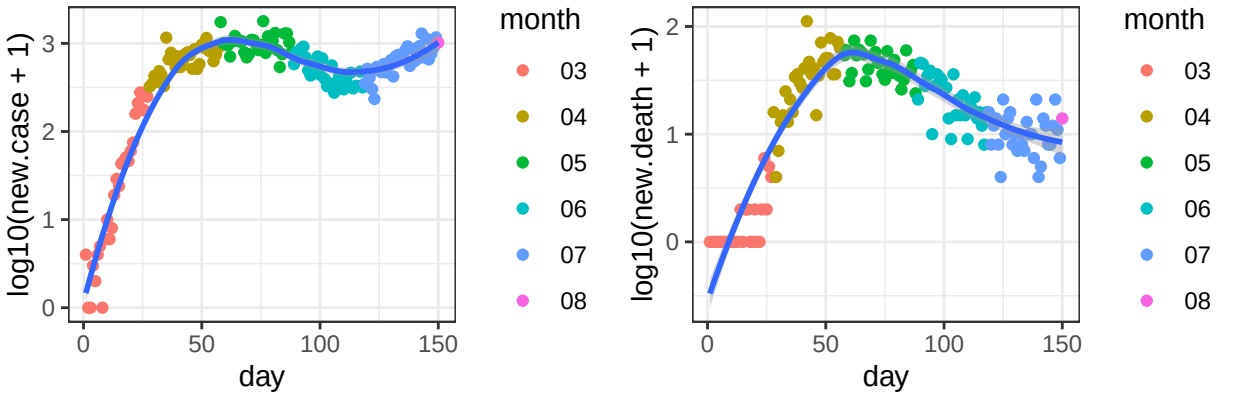
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Ohio



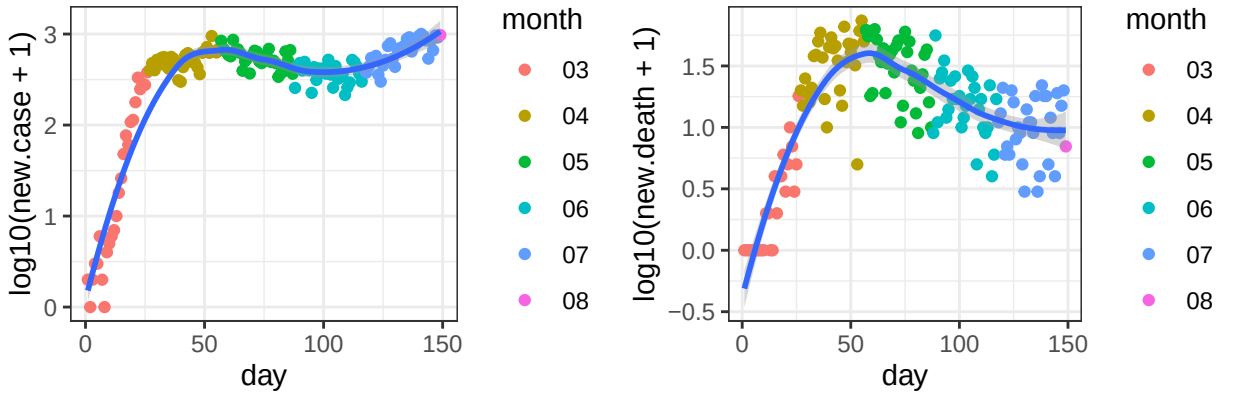
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



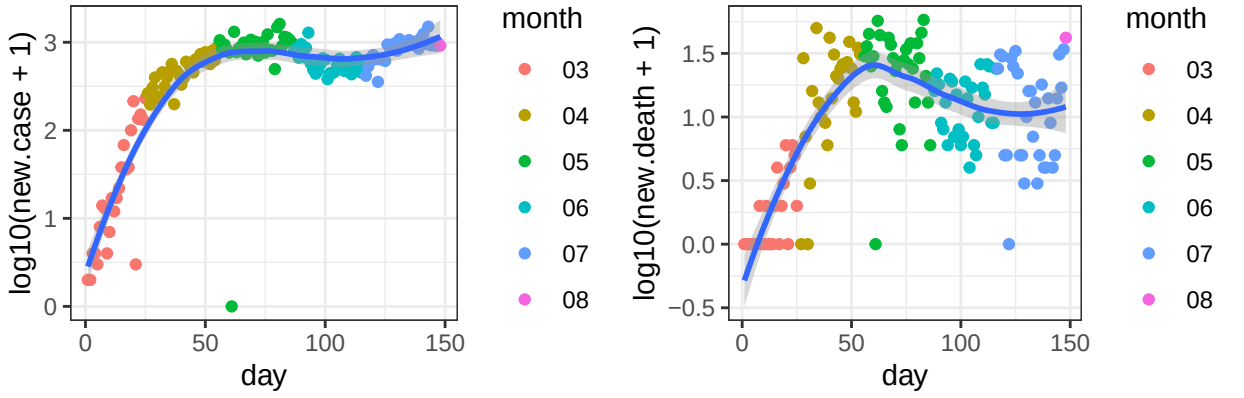
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Indiana



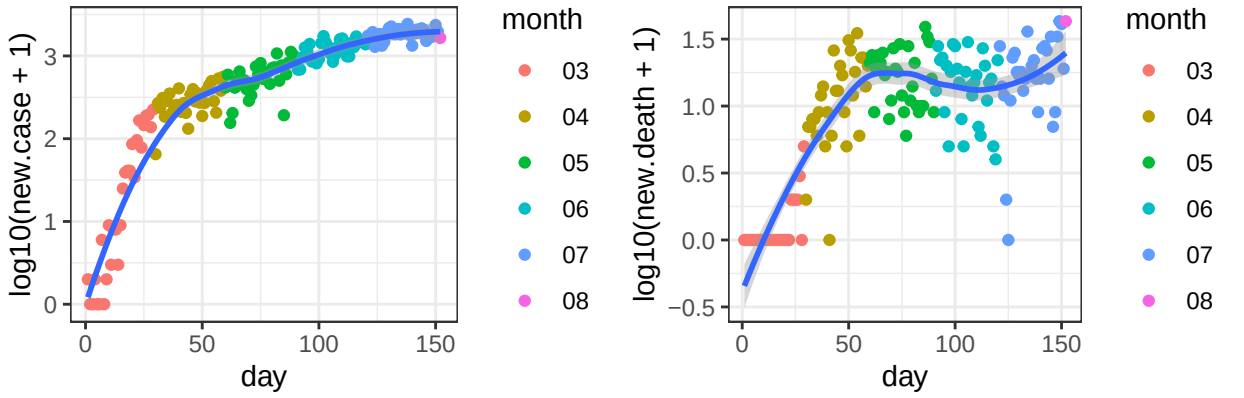
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Virginia



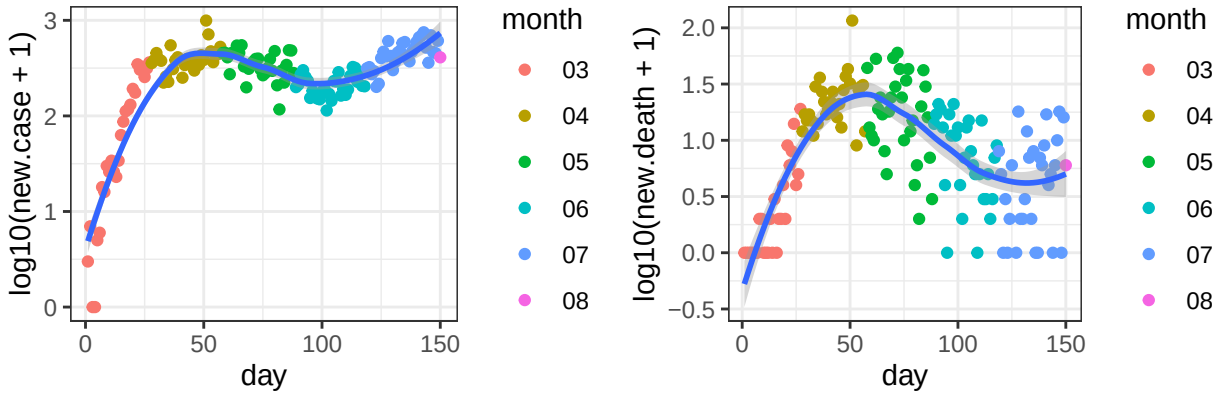
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

North Carolina



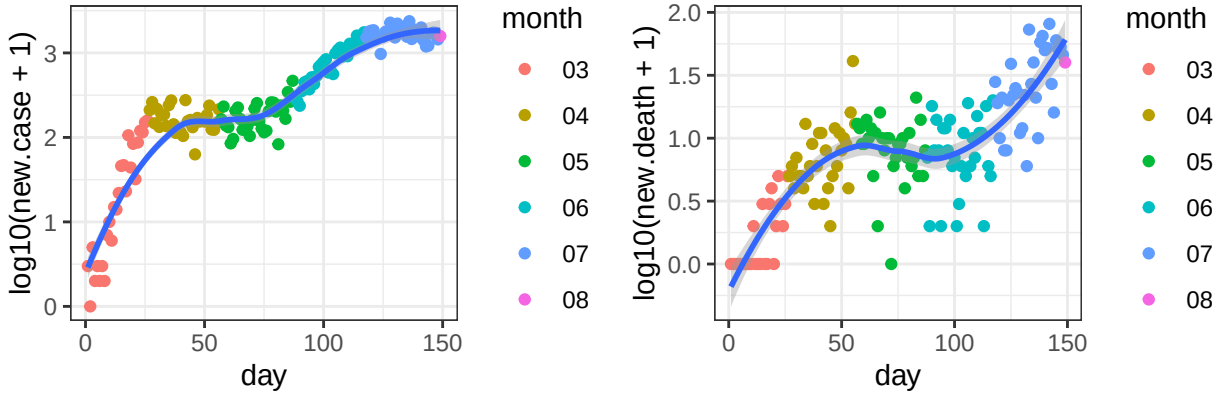
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Colorado



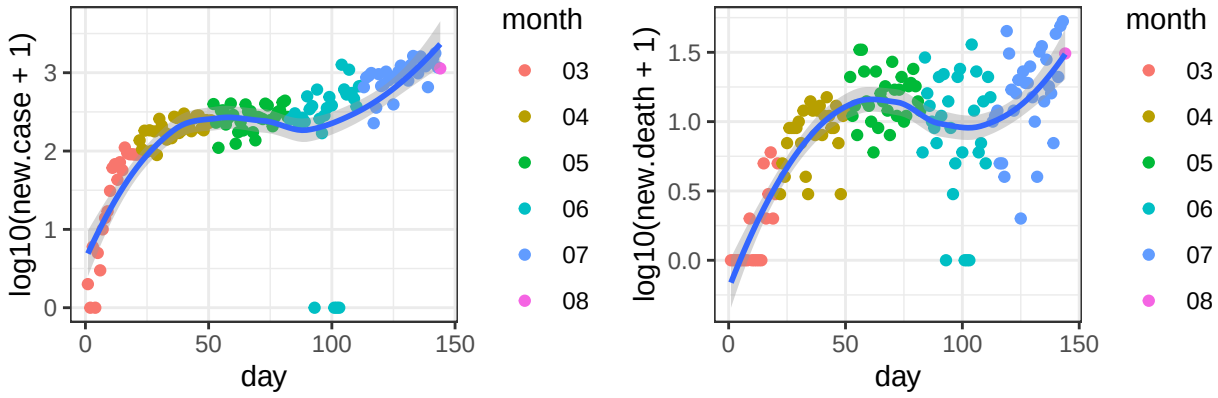
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

South Carolina



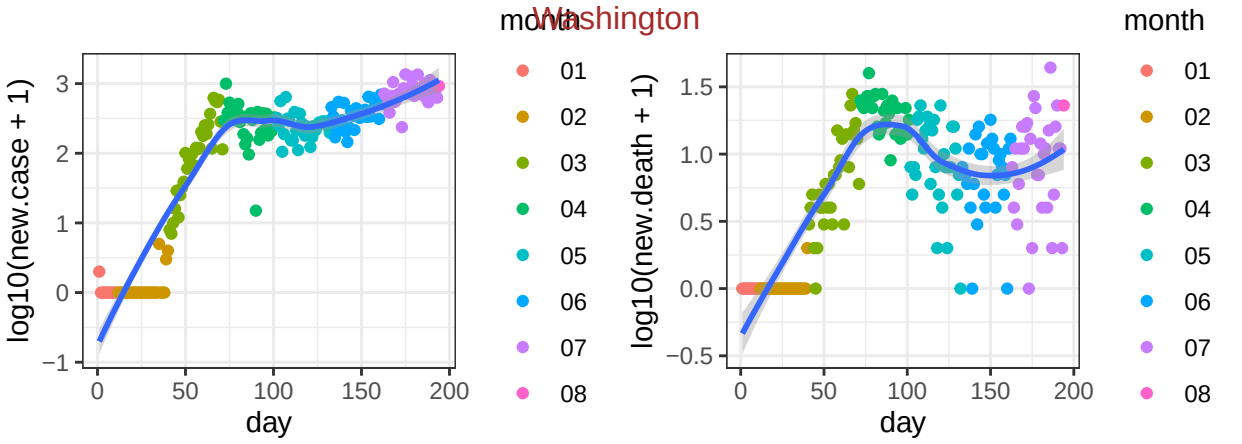
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Mississippi



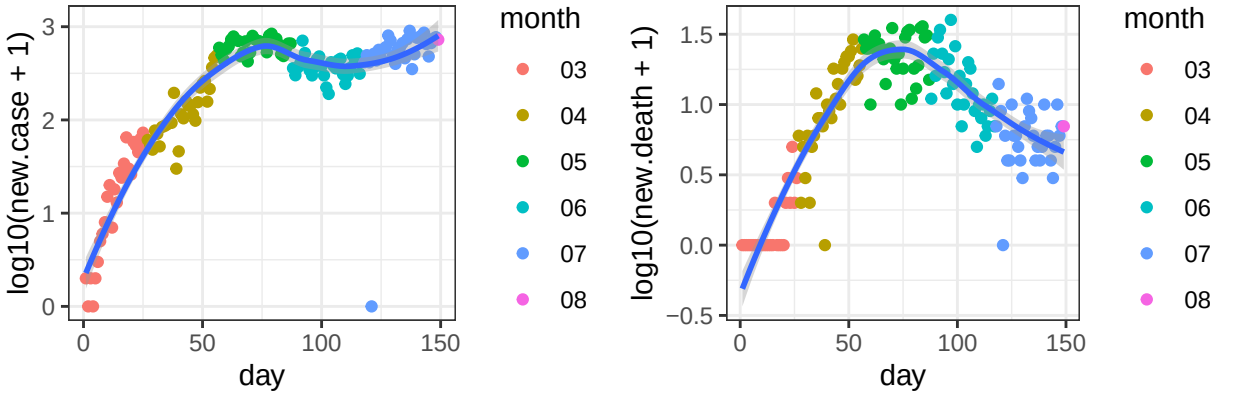
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Washington



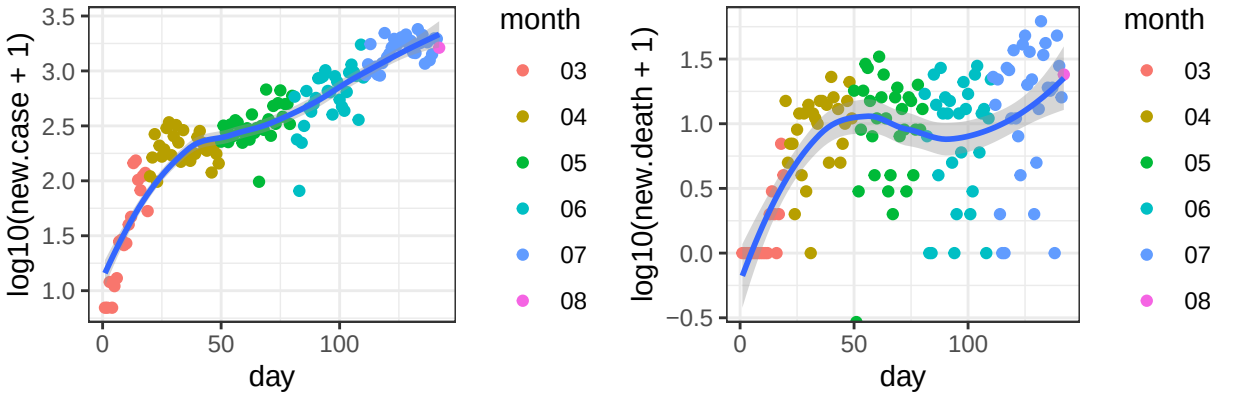
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Minnesota



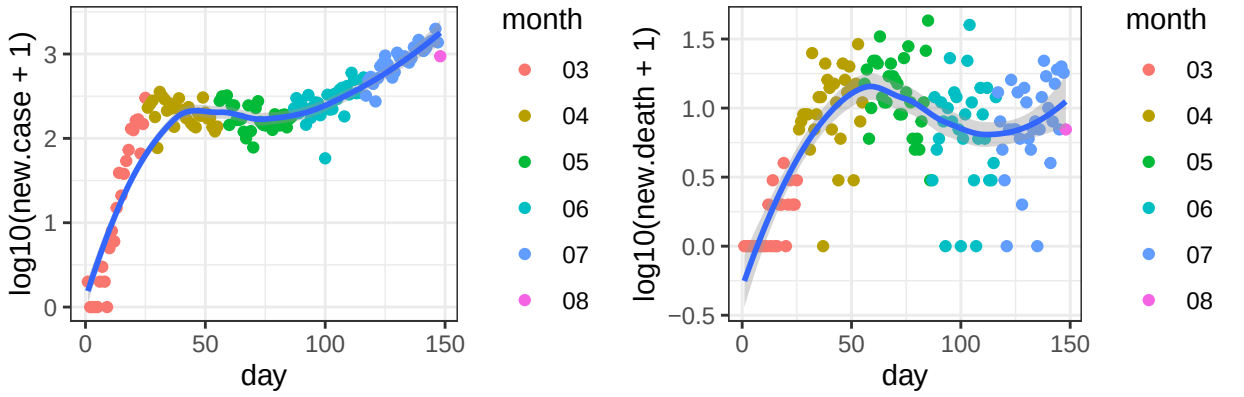
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Alabama



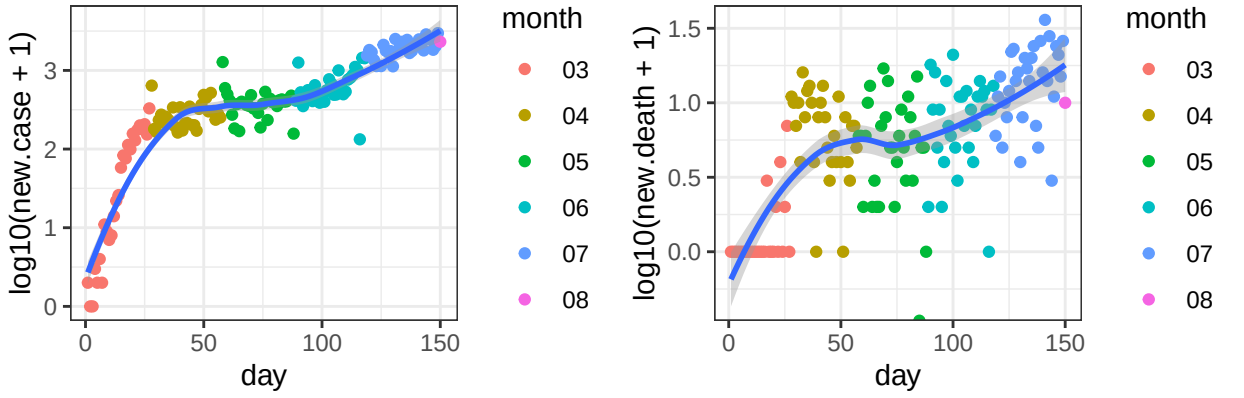
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Missouri



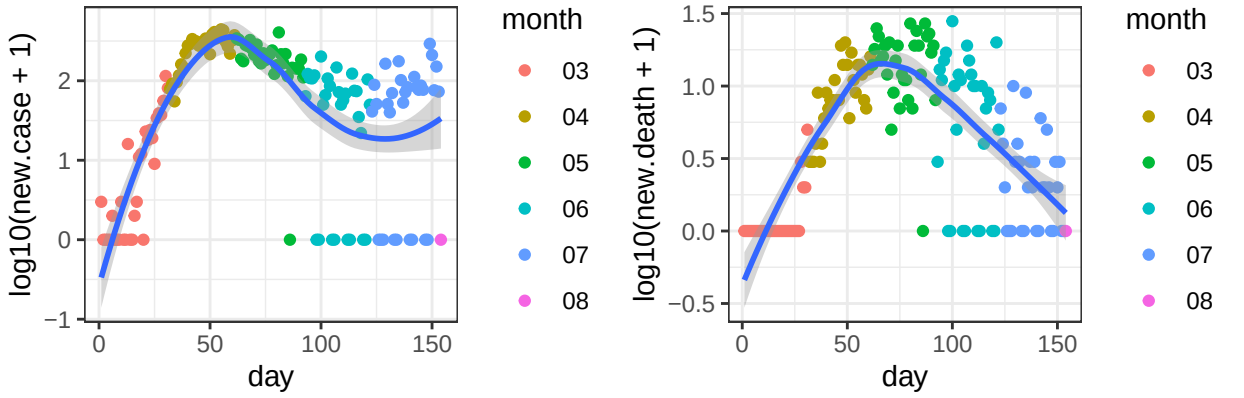
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Tennessee



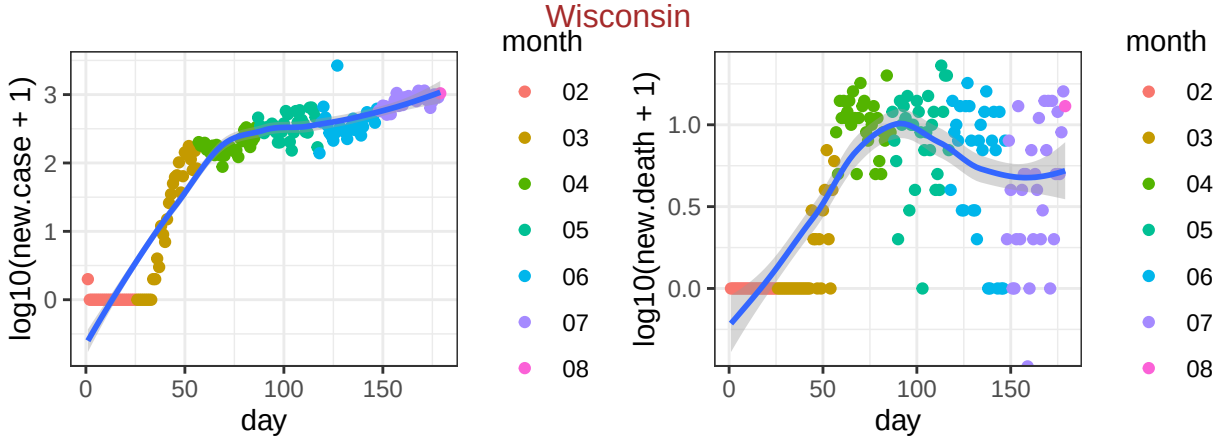
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Rhode Island



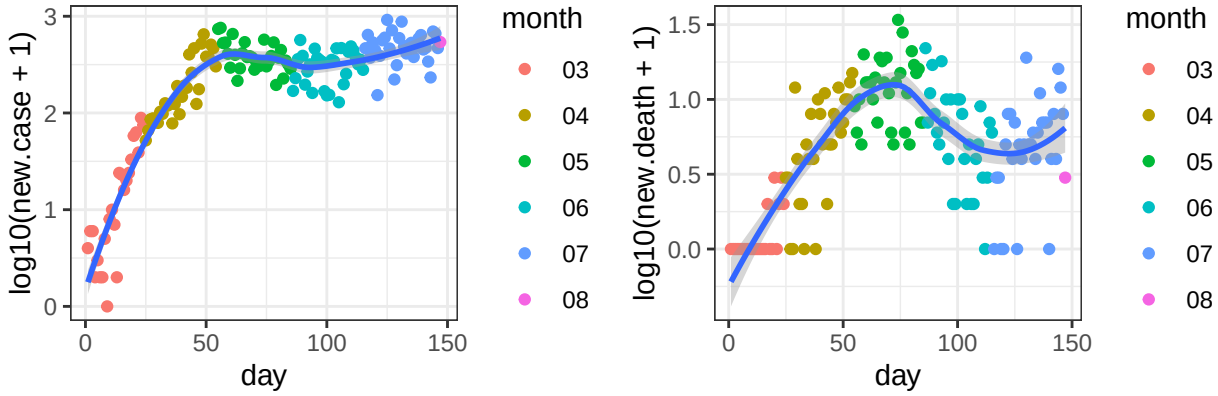
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Wisconsin



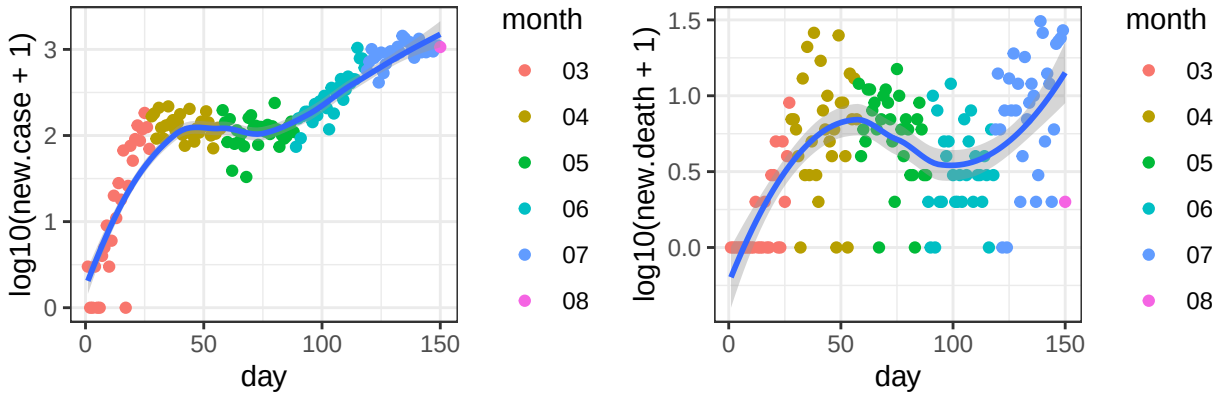
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

Iowa



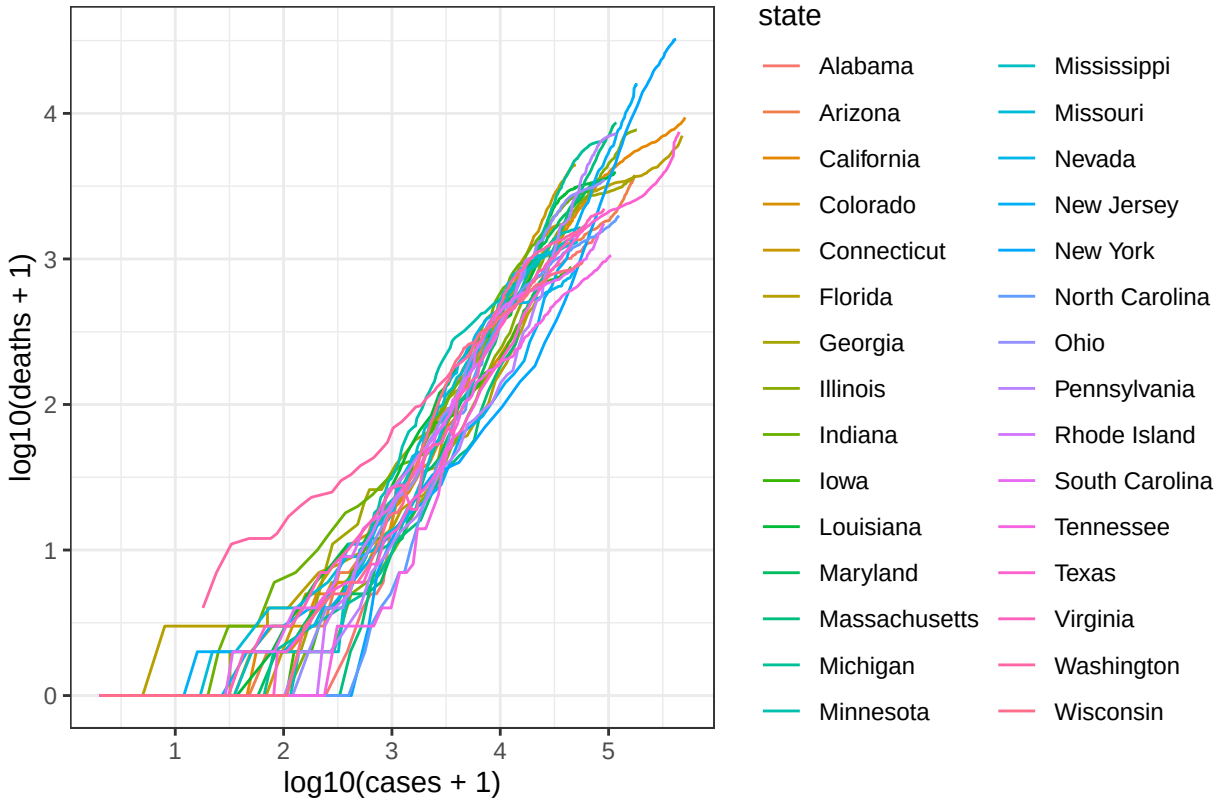
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Nevada



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

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

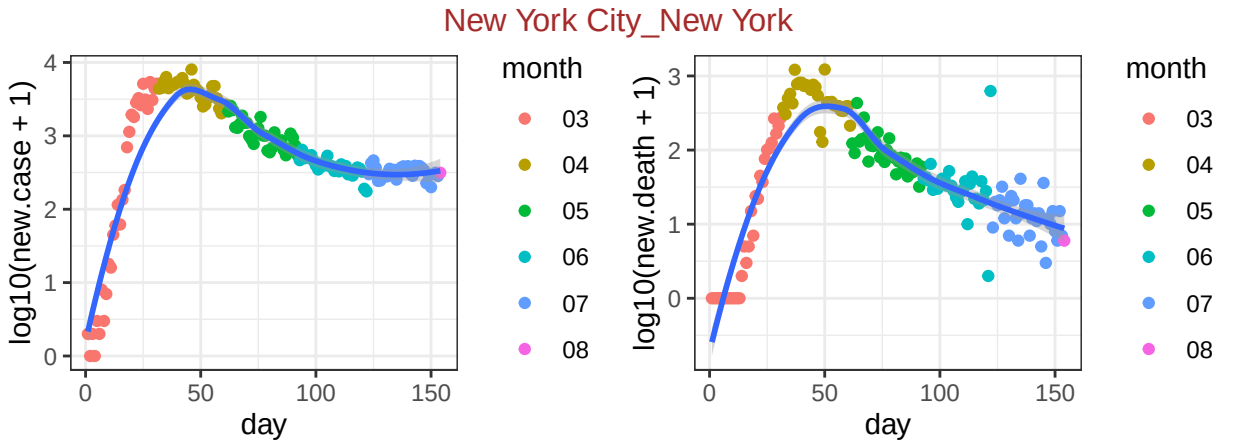
county level data

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

```
##              date          county         state  fips  cases deaths
## 391054 2020-08-01   New York City      New York    NA 230147  23007
## 389813 2020-08-01            Cook      Illinois 17031 106131   4888
## 389406 2020-08-01     Los Angeles    California  6037 190693   4669
## 390520 2020-08-01           Wayne      Michigan 26163  27208   2805
## 391053 2020-08-01          Nassau      New York 36059  43271   2706
## 390977 2020-08-01           Essex    New Jersey 34013  19748   2102
## 389304 2020-08-01        Maricopa       Arizona  4013 119295   2089
## 391073 2020-08-01         Suffolk      New York 36103  43300   2044
## 390972 2020-08-01          Bergen    New Jersey 34003  20749   2040
## 390431 2020-08-01       Middlesex Massachusetts 25017  25801   1983
## 391490 2020-08-01    Philadelphia  Pennsylvania 42101  30354   1690
## 389565 2020-08-01      Miami-Dade       Florida 12086 121206   1647
## 391081 2020-08-01     Westchester      New York 36119  35973   1578
## 390979 2020-08-01          Hudson    New Jersey 34017  19666   1501
## 389510 2020-08-01        Hartford   Connecticut  9003  12645   1412
## 389509 2020-08-01       Fairfield   Connecticut  9001  17793   1406
## 390982 2020-08-01       Middlesex    New Jersey 34023  17906   1404
## 390990 2020-08-01           Union    New Jersey 34039  16716   1347
## 391898 2020-08-01          Harris         Texas 48201  74884   1288
## 390986 2020-08-01         Passaic    New Jersey 34031  17616   1242
## 390427 2020-08-01           Essex Massachusetts 25009  17305   1182
## 390500 2020-08-01         Oakland      Michigan 26125  14721   1126
```

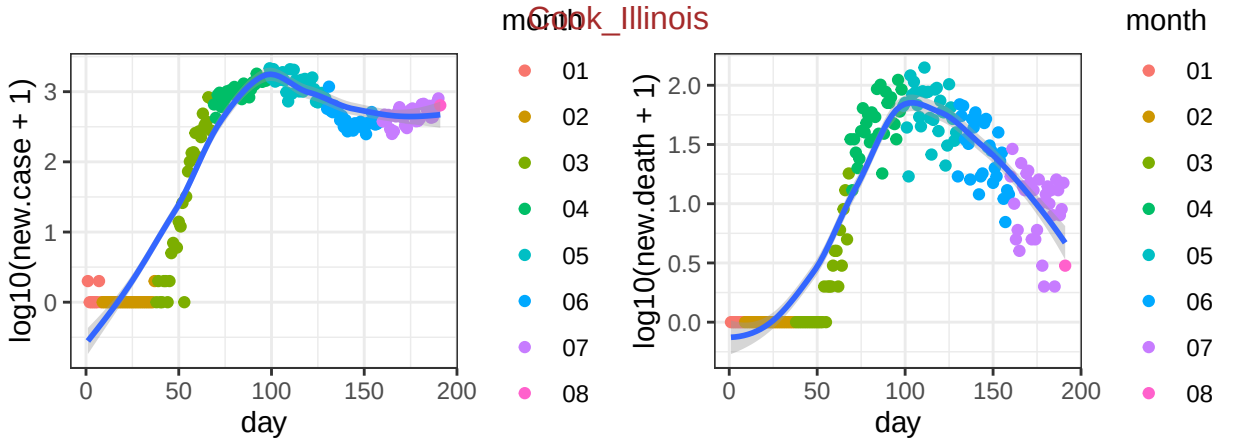
```
## 389513 2020-08-01       New Haven   Connecticut  9009  13041    1101
## 390435 2020-08-01         Suffolk Massachusetts 25025  21279    1057
## 390985 2020-08-01           Ocean    New Jersey 34029  10507    1015
## 390437 2020-08-01       Worcester Massachusetts 25027  13376     991
## 390433 2020-08-01         Norfolk Massachusetts 25021  10305     986
## 390487 2020-08-01          Macomb      Michigan 26099   9806     941
## 390983 2020-08-01        Monmouth    New Jersey 34025  10193     858
## 391485 2020-08-01      Montgomery  Pennsylvania 42091   9813     850
## 389572 2020-08-01      Palm Beach       Florida 12099  33852     833
## 390984 2020-08-01          Morris    New Jersey 34027   7295     829
## 390548 2020-08-01        Hennepin     Minnesota 27053  17547     815
## 391589 2020-08-01      Providence  Rhode Island 44007  14549     808
## 390413 2020-08-01      Montgomery      Maryland 24031  17704     789
## 389949 2020-08-01          Marion       Indiana 18097  14721     766
## 389528 2020-08-01         Broward       Florida 12011  56797     742
## 390414 2020-08-01 Prince George's      Maryland 24033  23057     741
## 391462 2020-08-01        Delaware  Pennsylvania 42045   8770     730
## 390434 2020-08-01        Plymouth Massachusetts 25023   9107     711
## 390429 2020-08-01         Hampden Massachusetts 25013   7433     697
## 389420 2020-08-01       Riverside    California  6065  37612     695
## 390947 2020-08-01           Clark        Nevada 32003  42167     688
## 391854 2020-08-01          Dallas         Texas 48113  50590     681
## 392245 2020-08-01            King    Washington 53033  15418     674
## 390794 2020-08-01       St. Louis      Missouri 29189  13162     650
## 389417 2020-08-01          Orange    California  6059  36833     649
## 391905 2020-08-01         Hidalgo         Texas 48215  17006     644
## 390425 2020-08-01         Bristol Massachusetts 25005   9088     623
## 391039 2020-08-01            Erie      New York 36029   8548     621
```

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



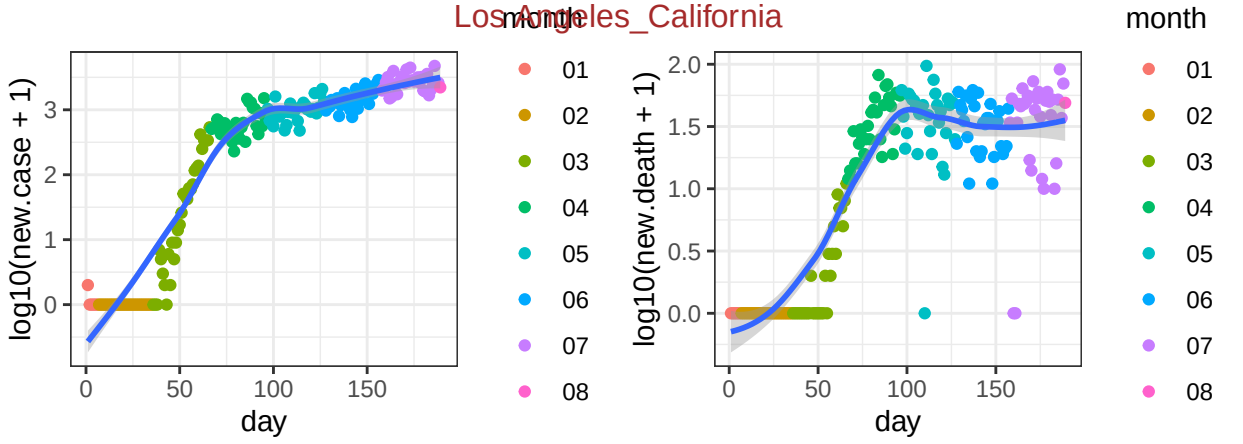
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



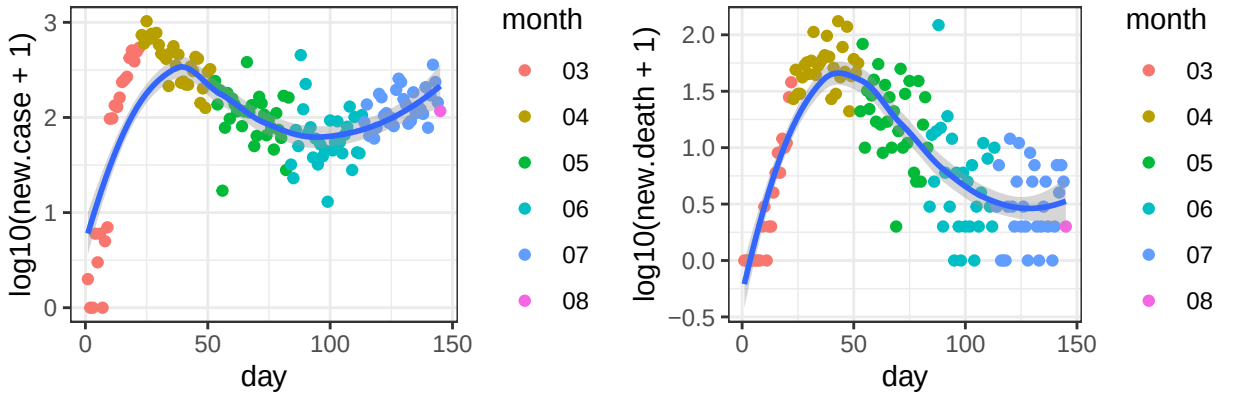
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Los Angeles_California



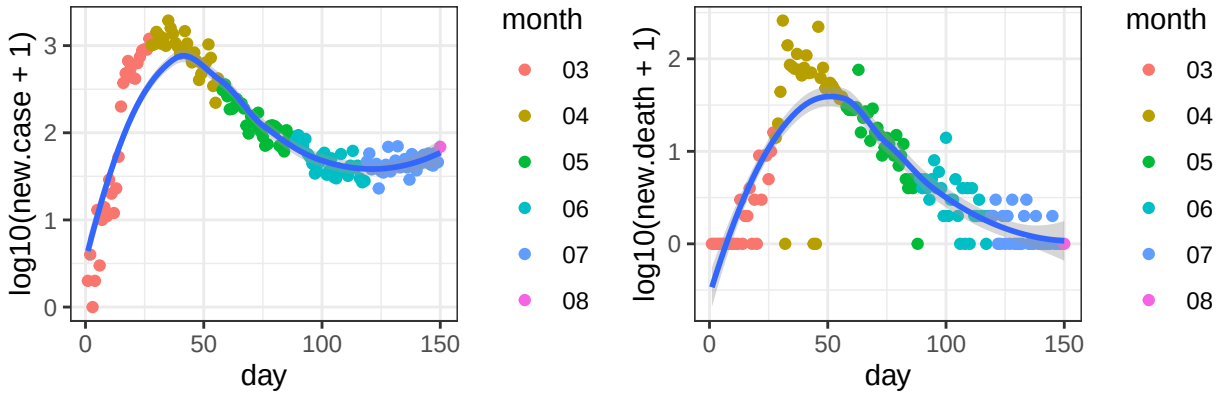
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Wayne_Michigan



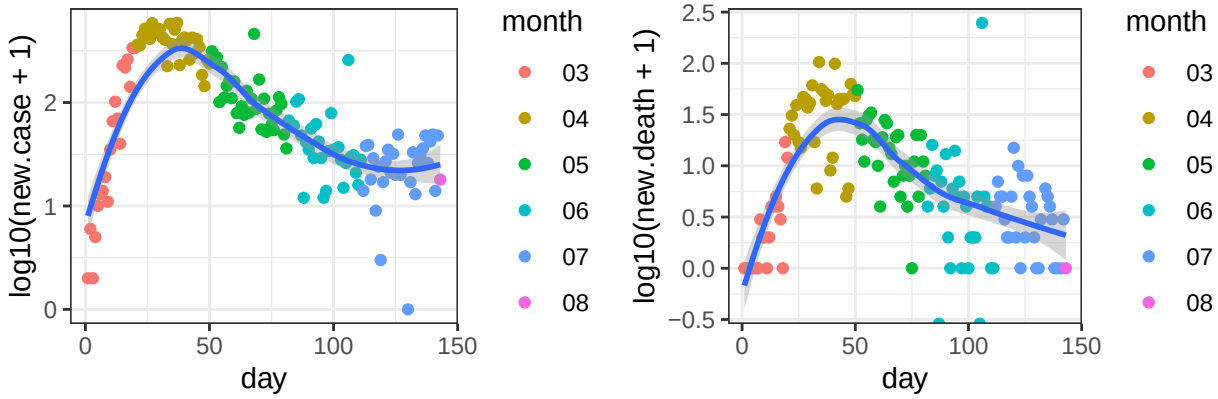
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Nassau_New York



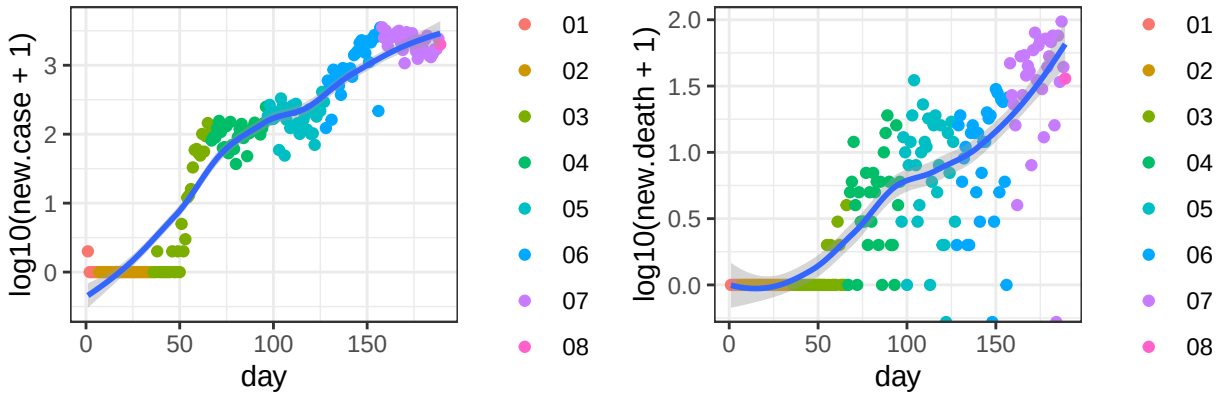
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



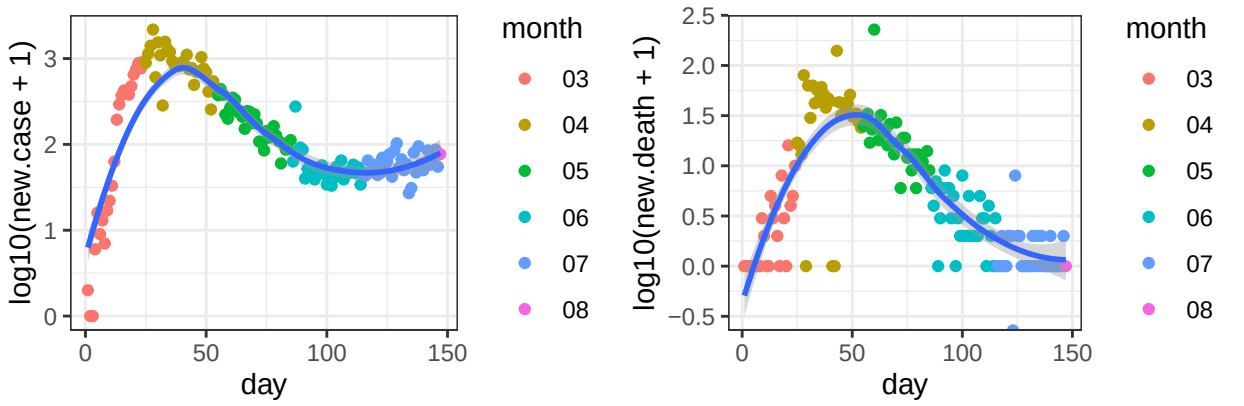
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Maricopa_Arizona



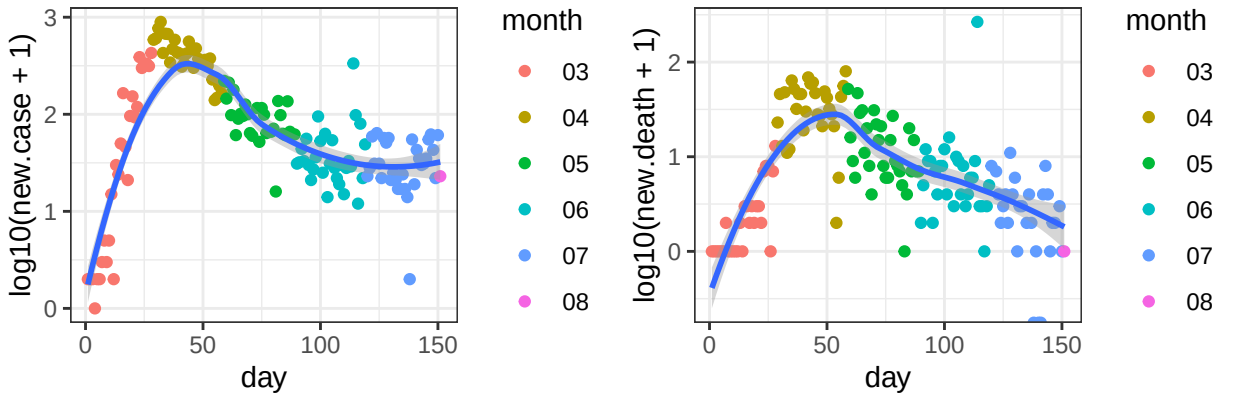
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Suffolk_New York



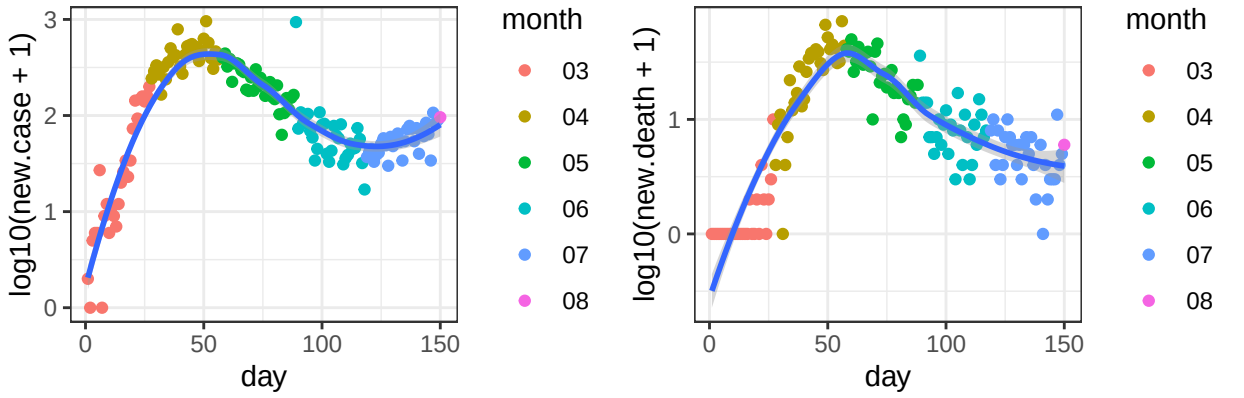
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Bergen_New Jersey



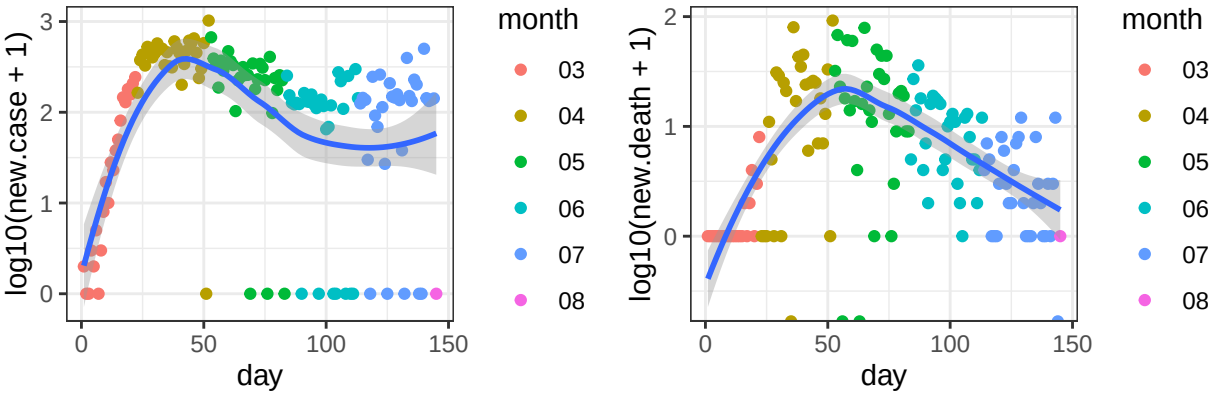
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Middlesex_Massachusetts



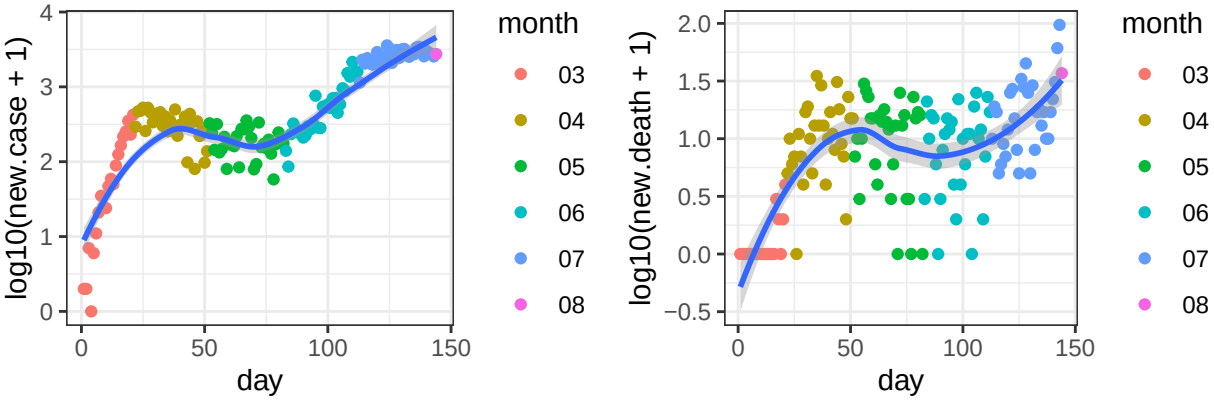
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Philadelphia_Pennsylvania



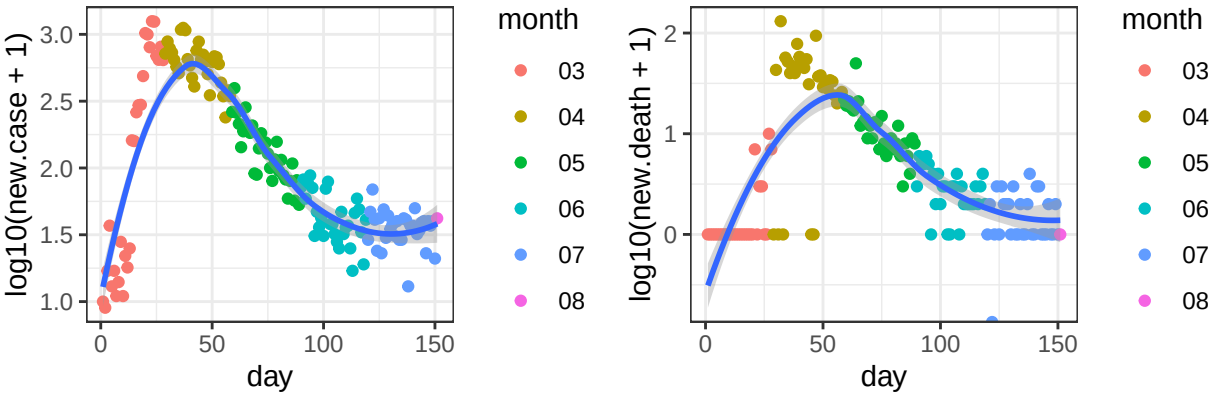
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Miami−Dade_Florida



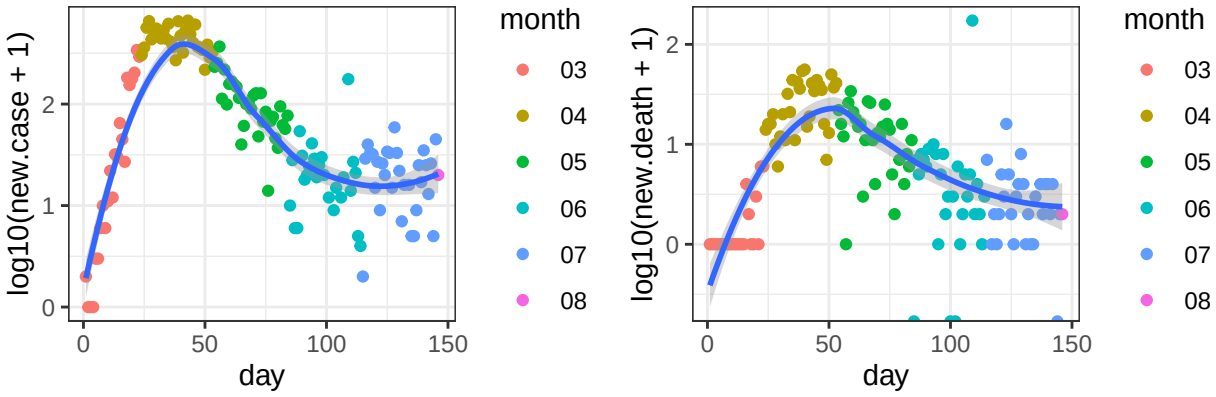
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Westchester_New York



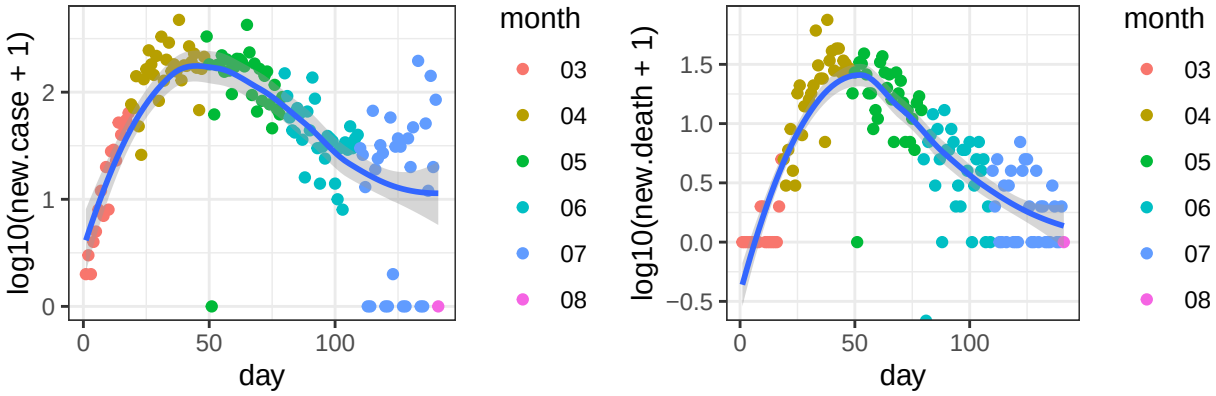
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Hudson_New Jersey



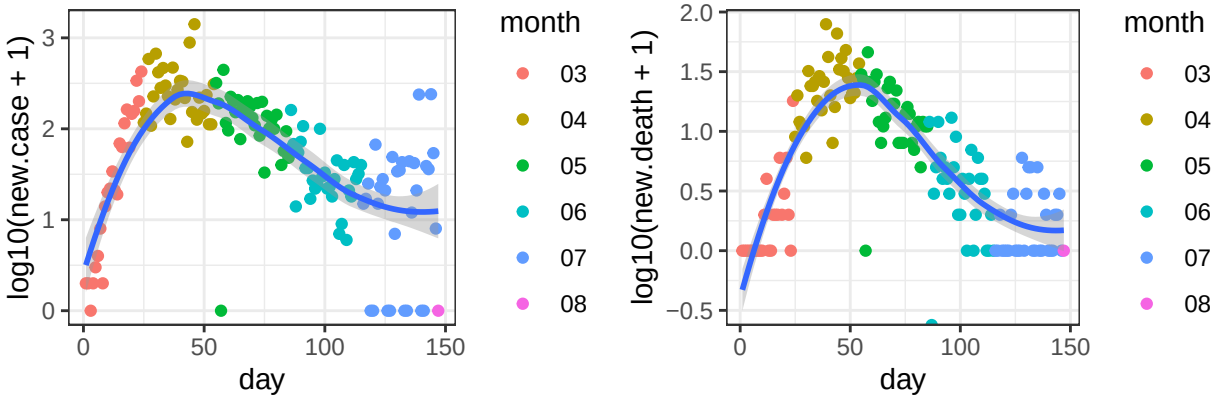
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Hartford_Connecticut



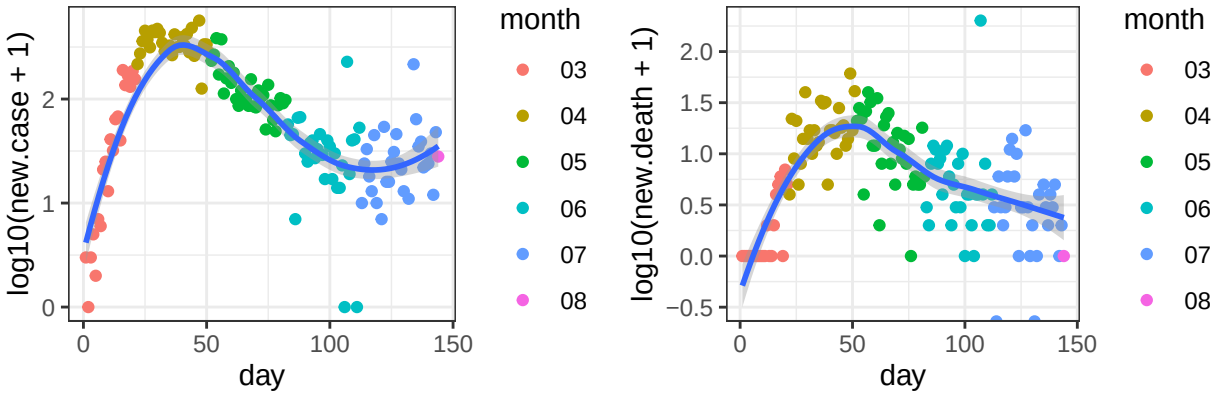
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Fairfield_Connecticut



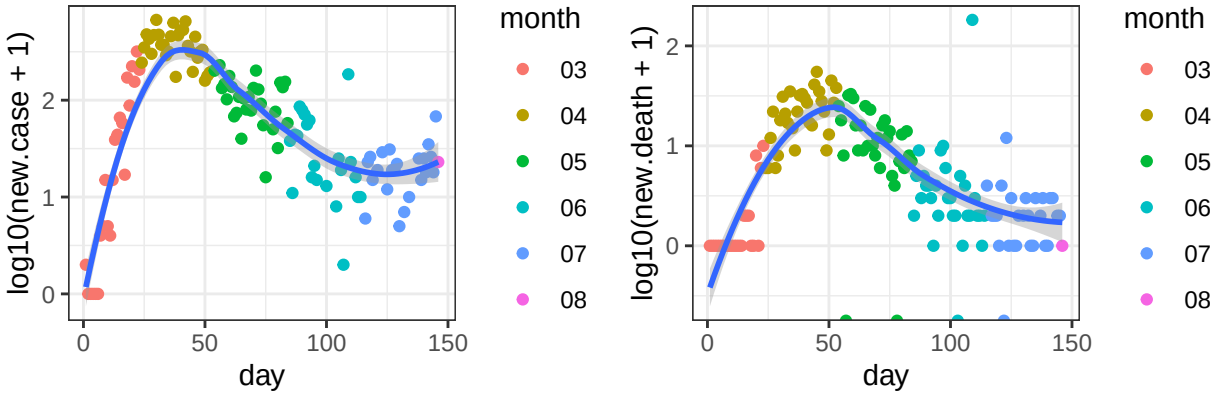
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Middlesex_New Jersey



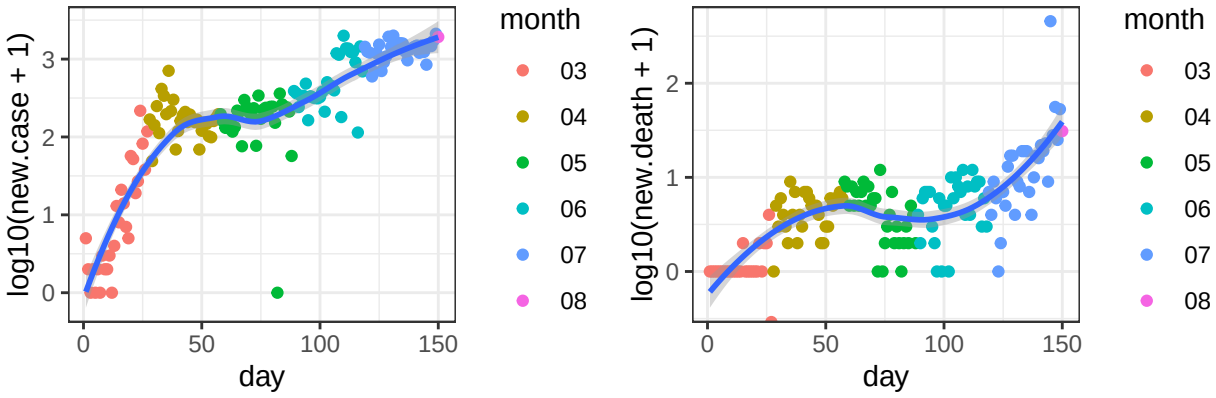
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Union_New Jersey



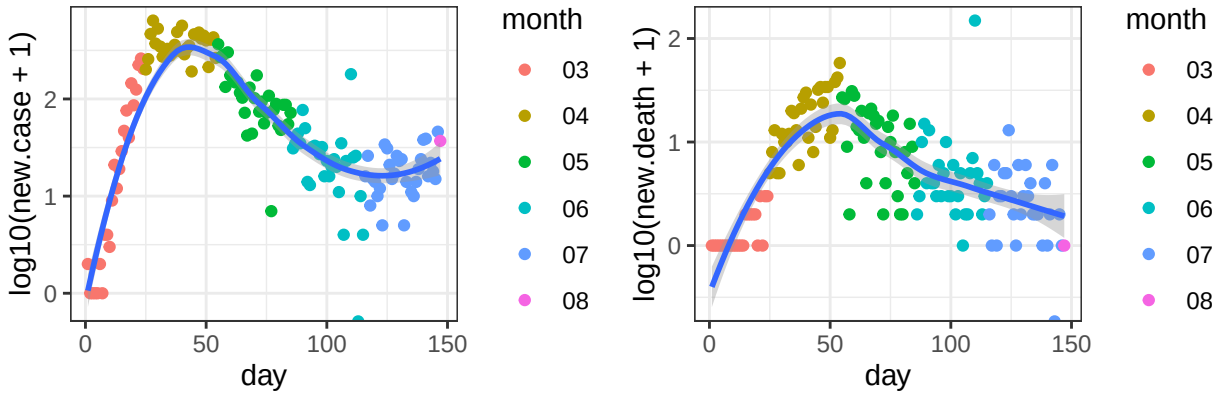
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Harris_Texas



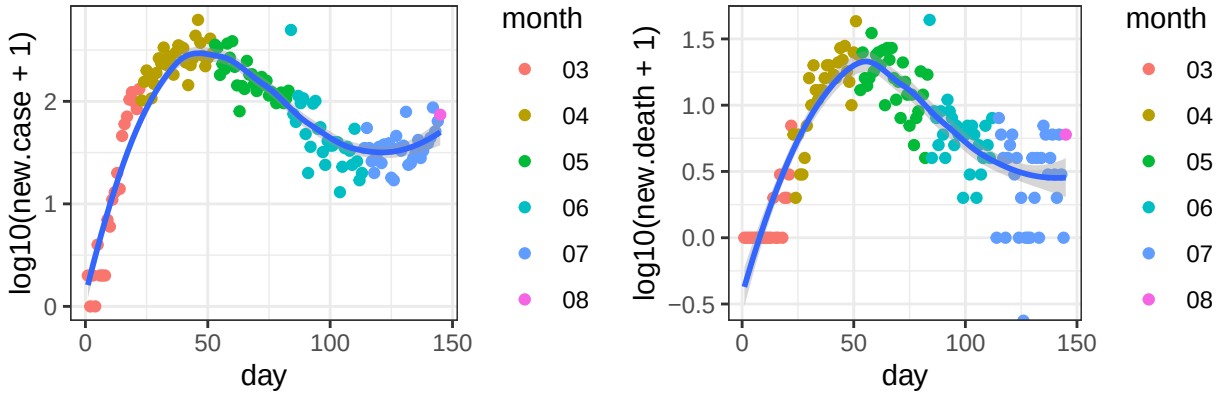
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Passaic_New Jersey



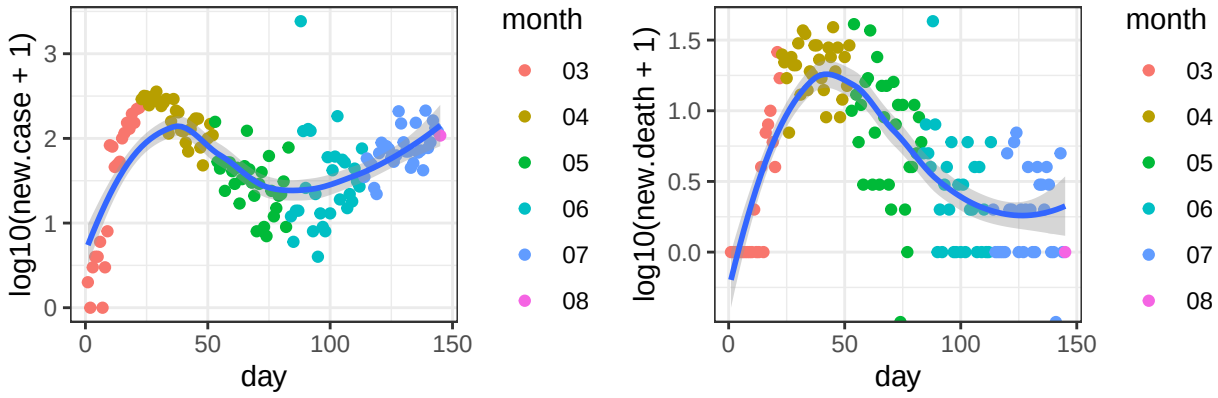
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Essex_Massachusetts



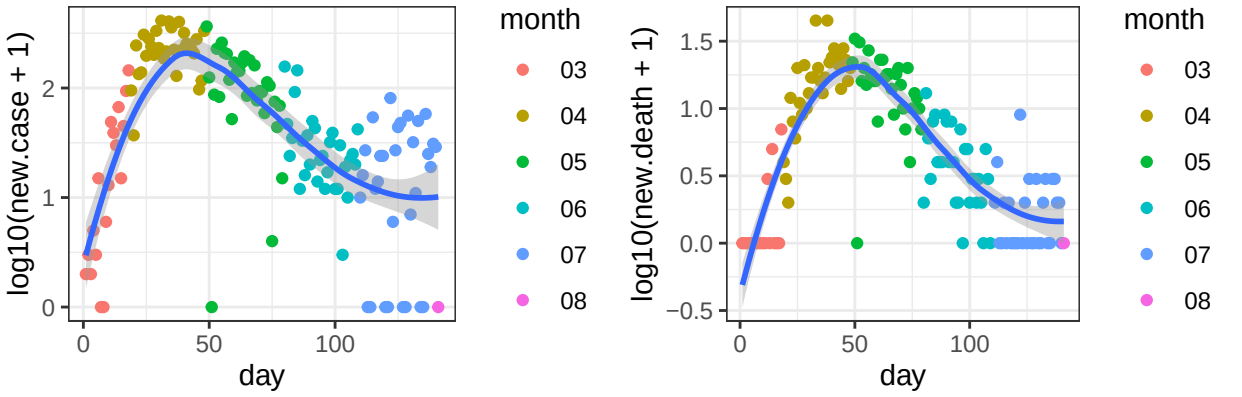
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Oakland_Michigan



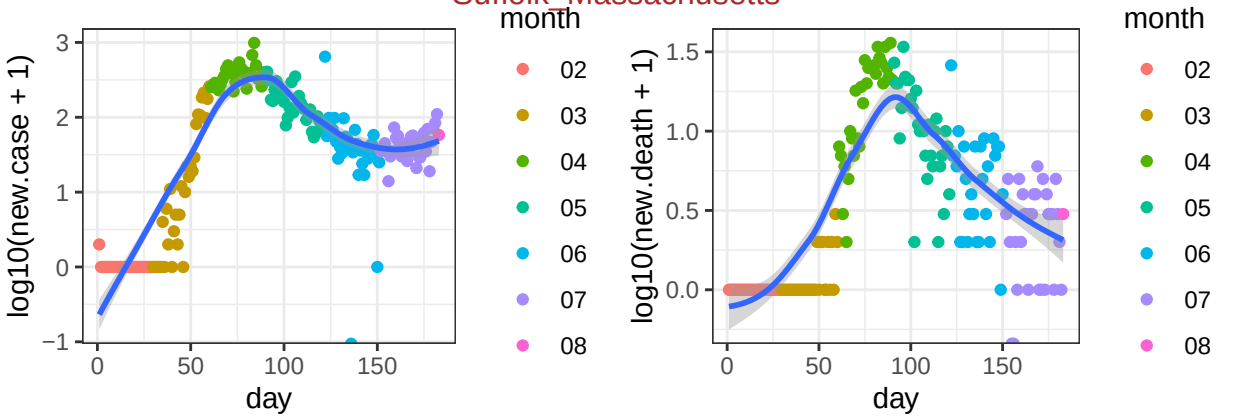
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

New Haven_Connecticut



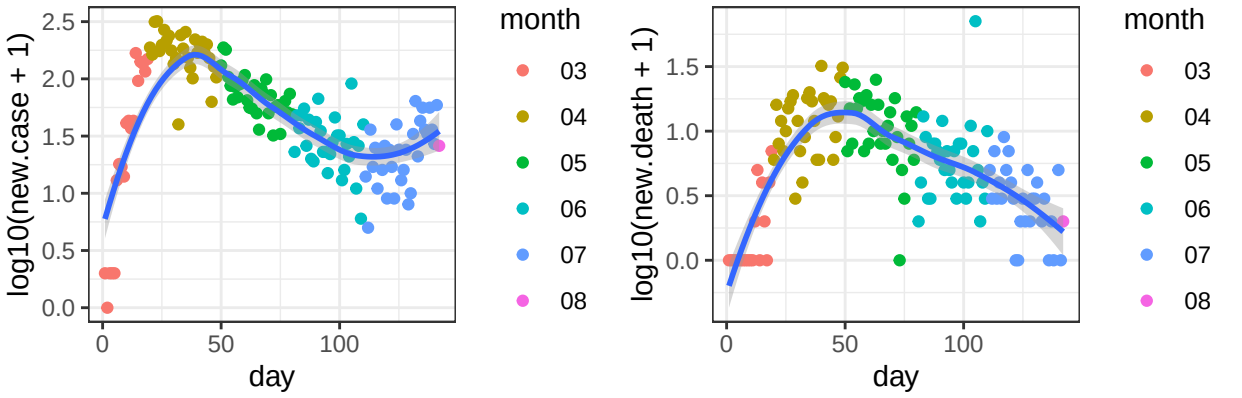
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Suffolk_Massachusetts



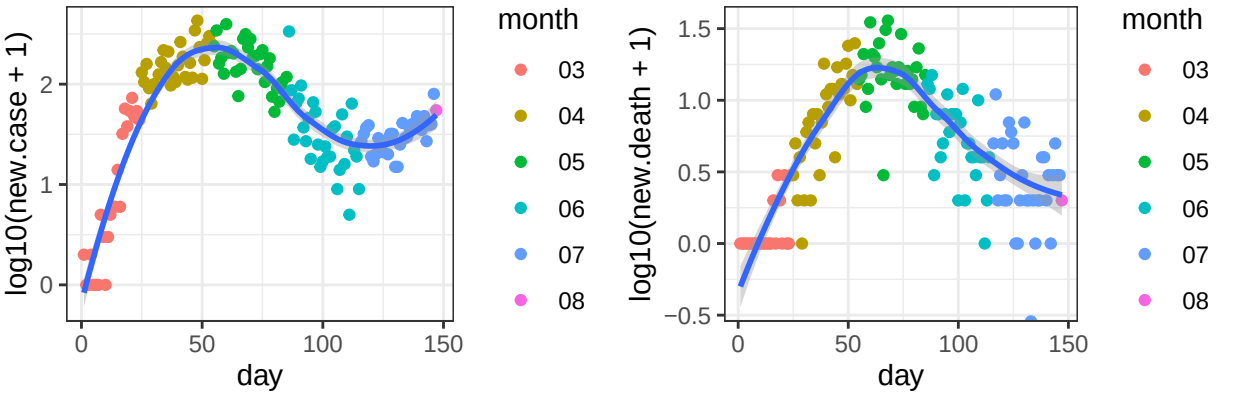
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Ocean_New Jersey



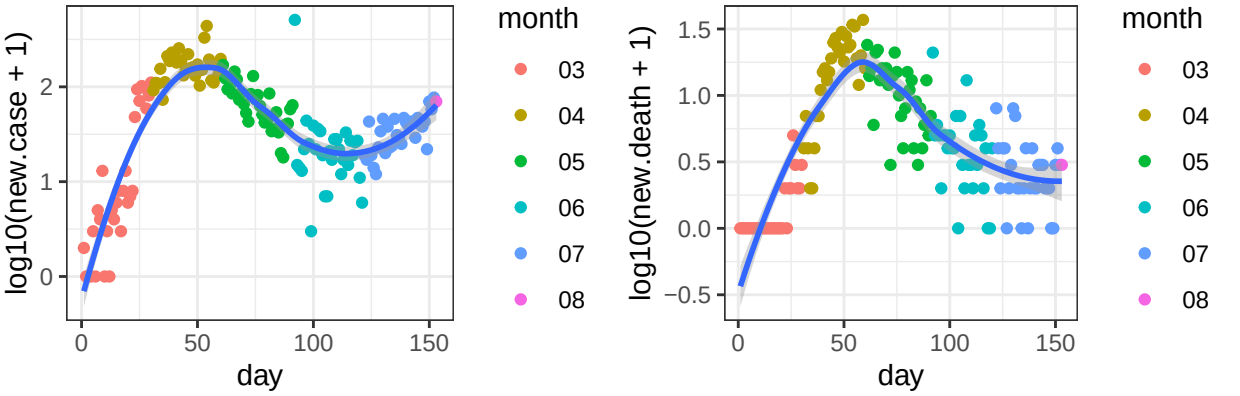
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Worcester_Massachusetts



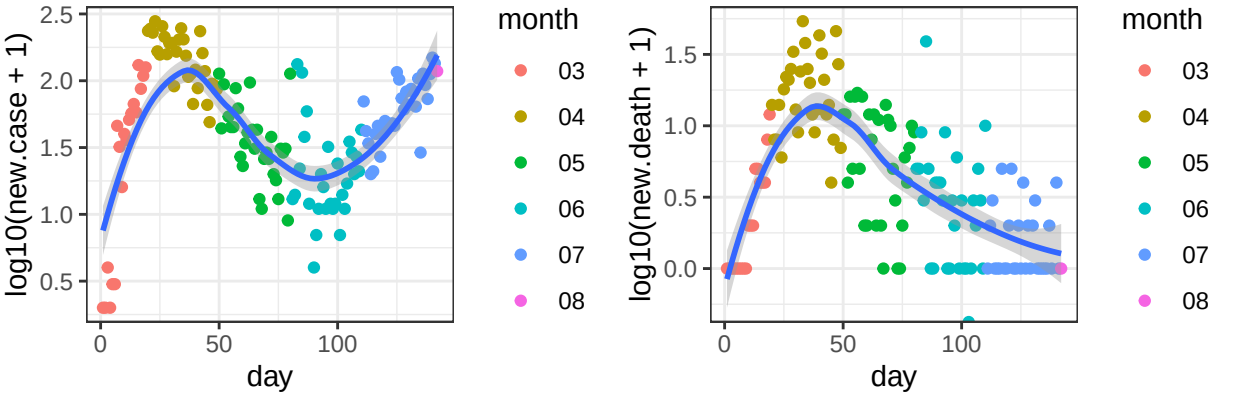
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Norfolk_Massachusetts



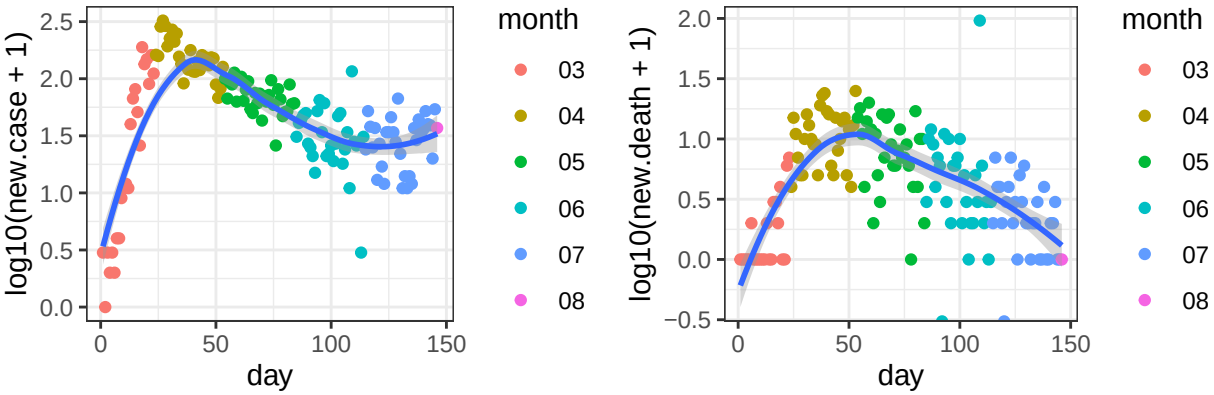
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Macomb_Michigan



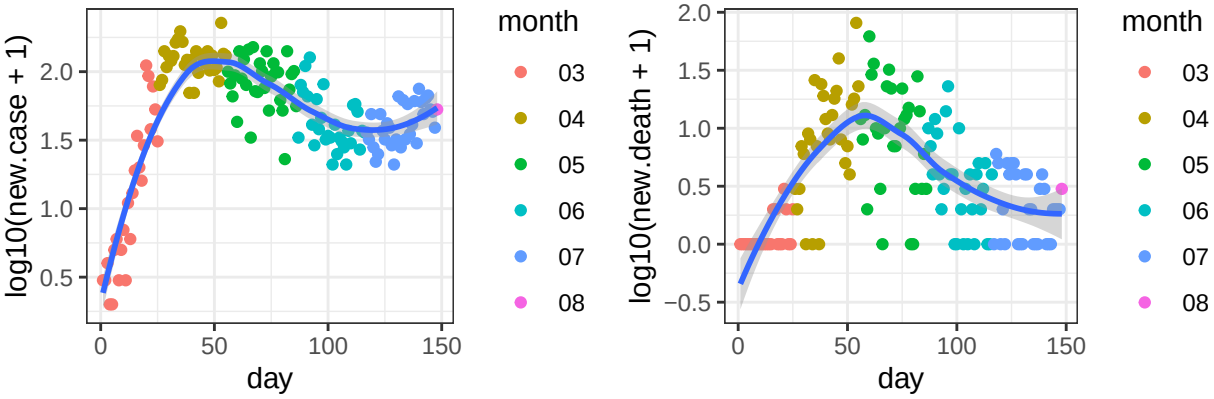
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Monmouth_New Jersey



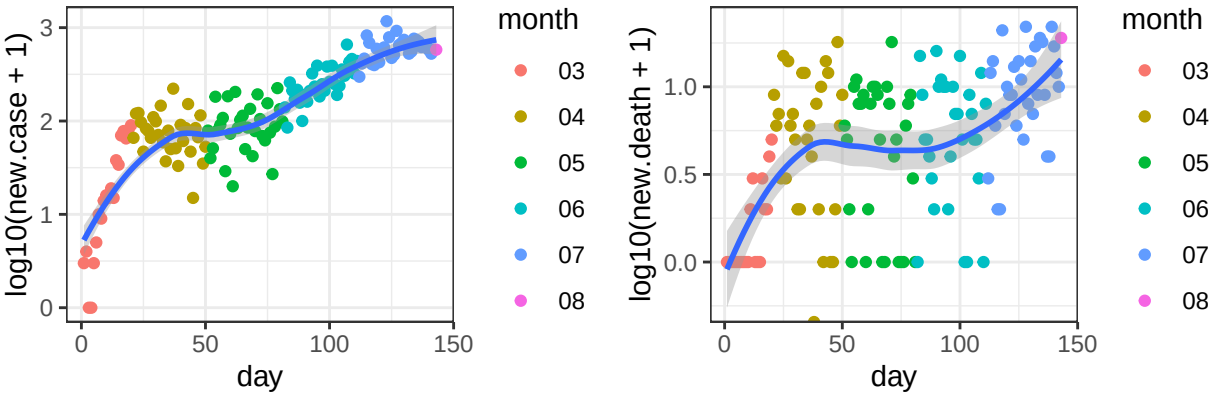
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Montgomery_Pennsylvania



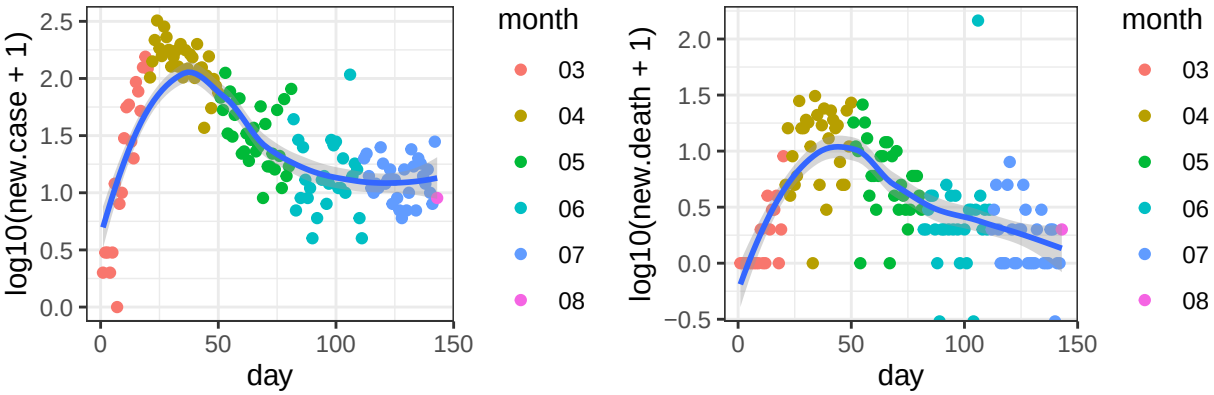
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Palm Beach_Florida



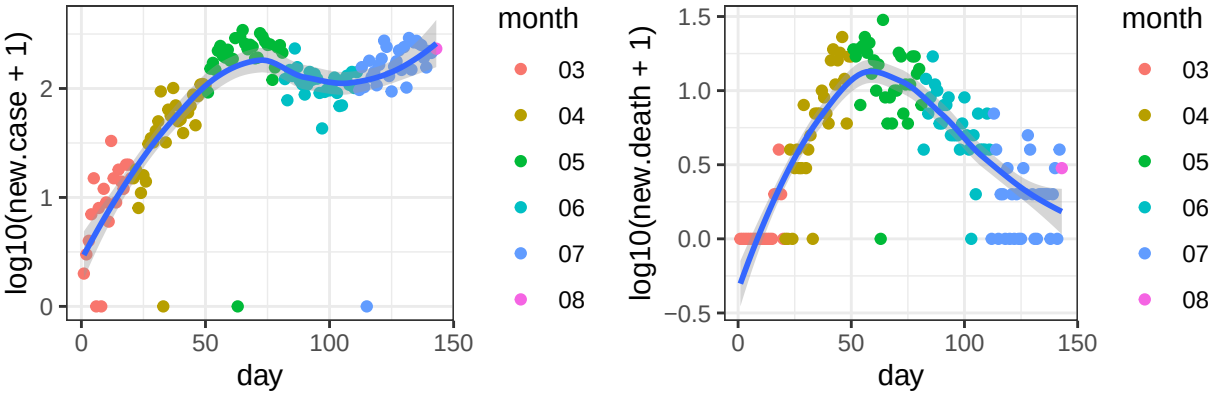
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Morris_New Jersey



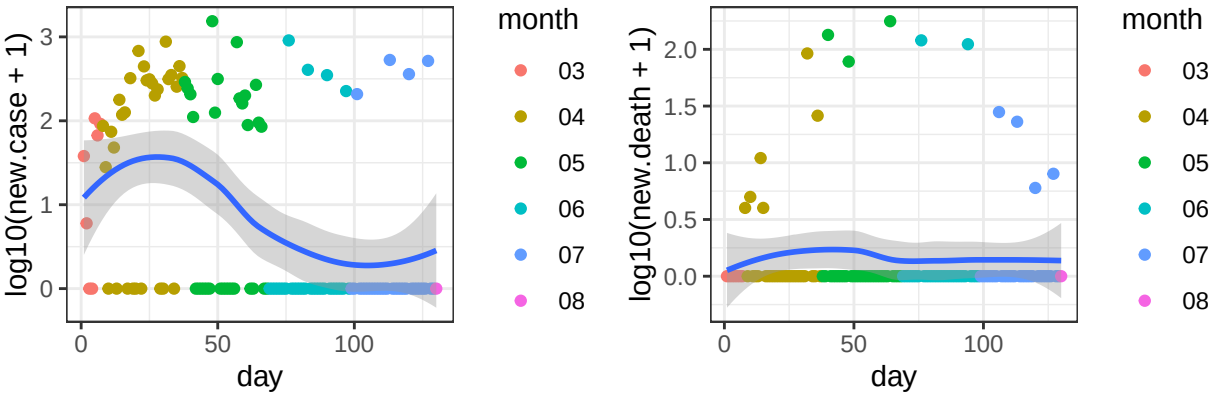
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Hennepin_Minnesota



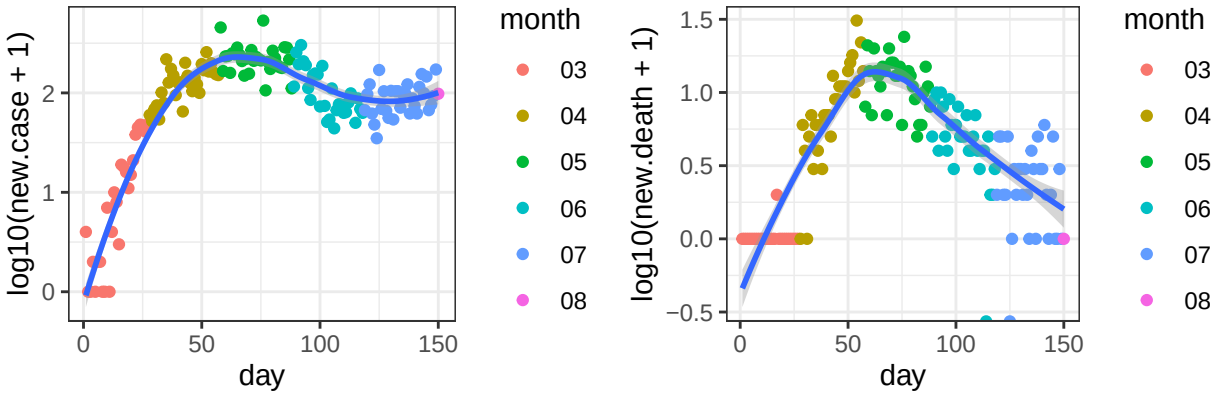
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Providence_Rhode Island



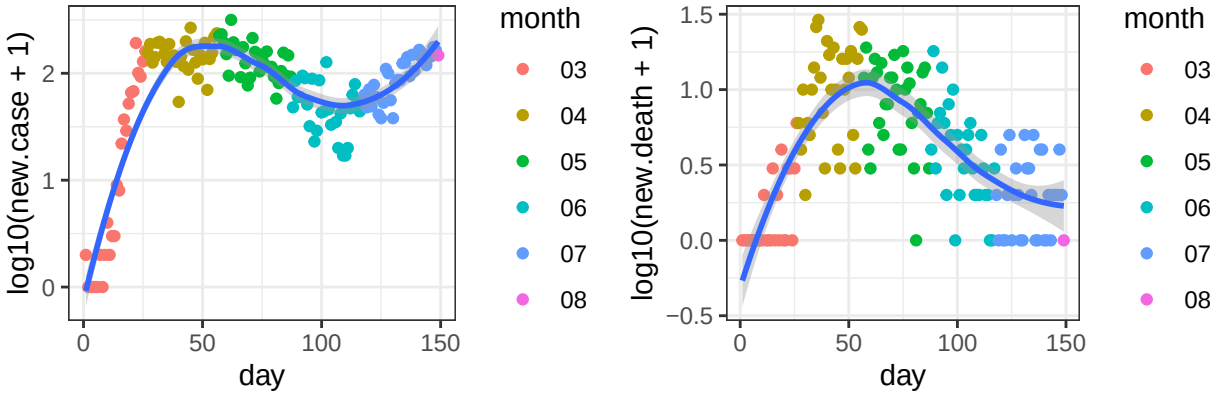
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Montgomery_Maryland



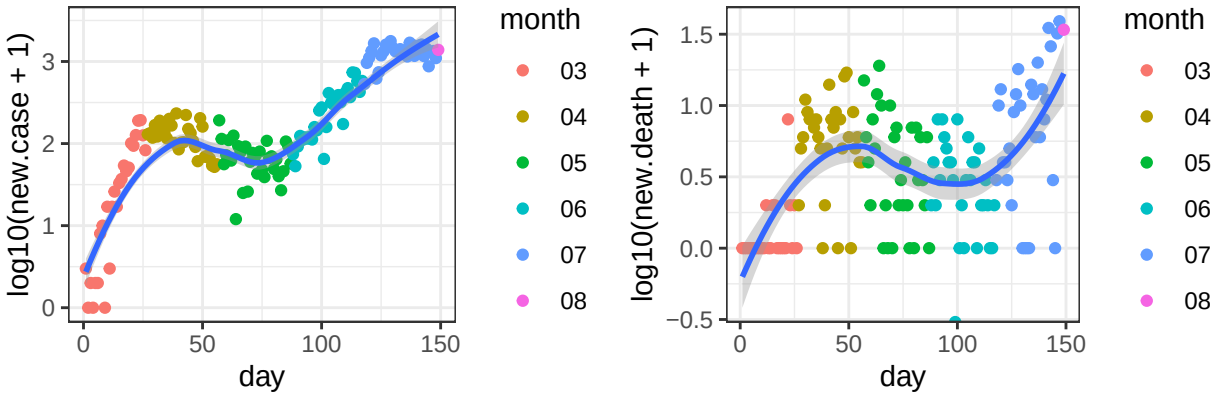
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Marion_Indiana



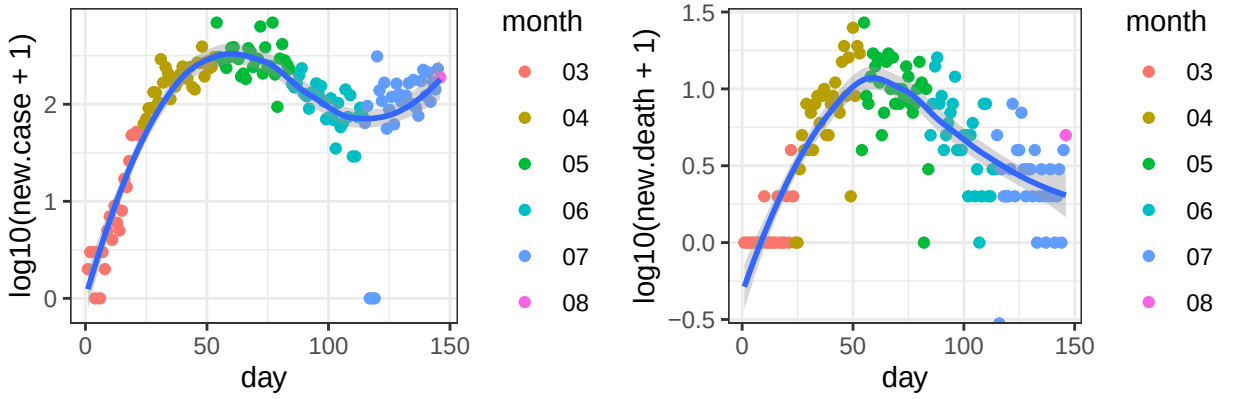
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Broward_Florida



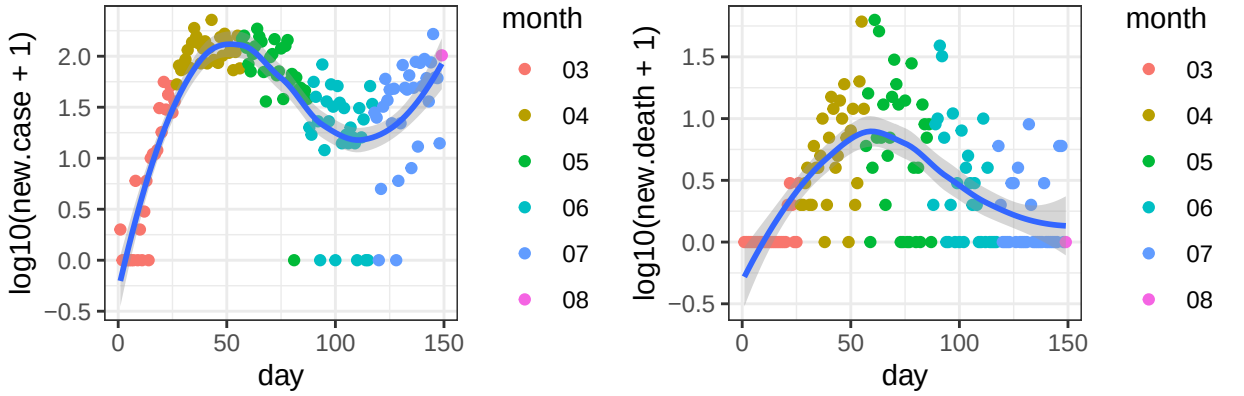
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Prince George's_Maryland



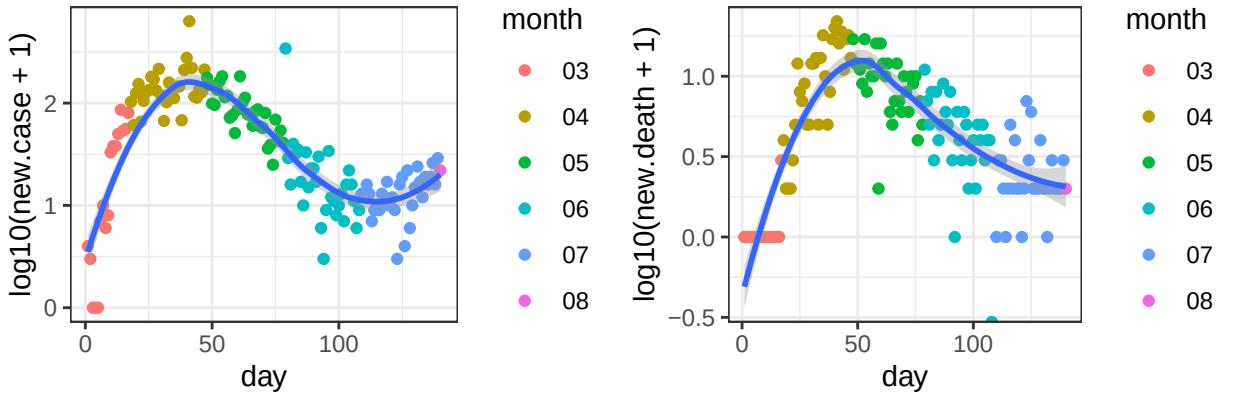
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Delaware_Pennsylvania



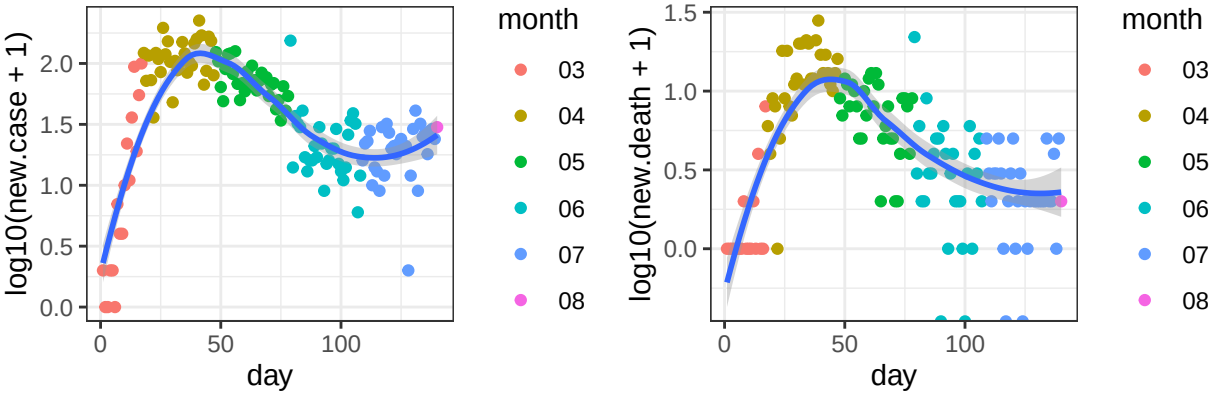
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Plymouth_Massachusetts



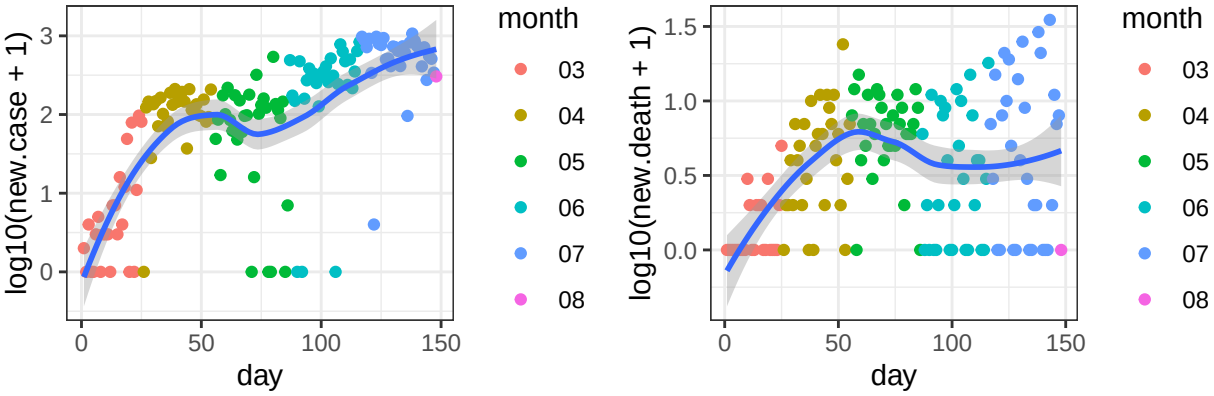
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Hampden_Massachusetts



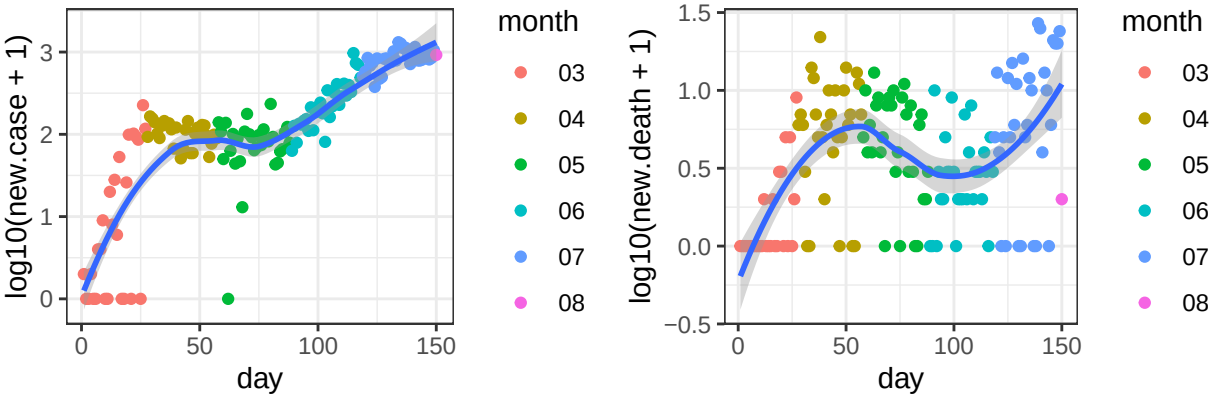
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Riverside_California



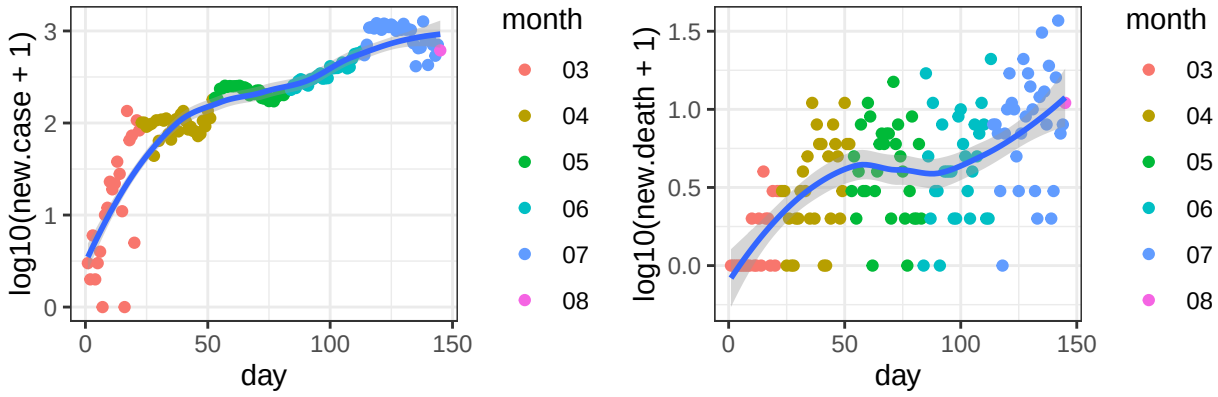
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Clark_Nevada



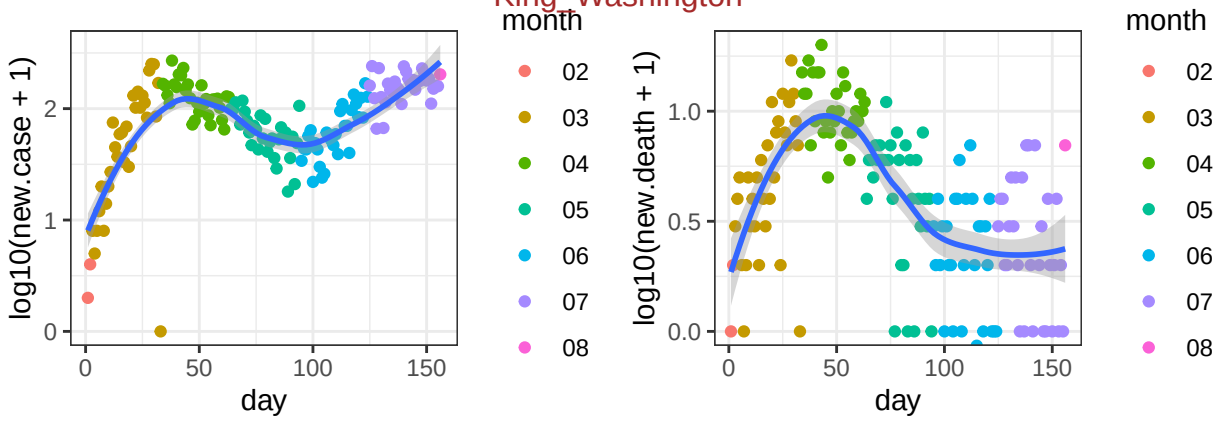
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Dallas_Texas



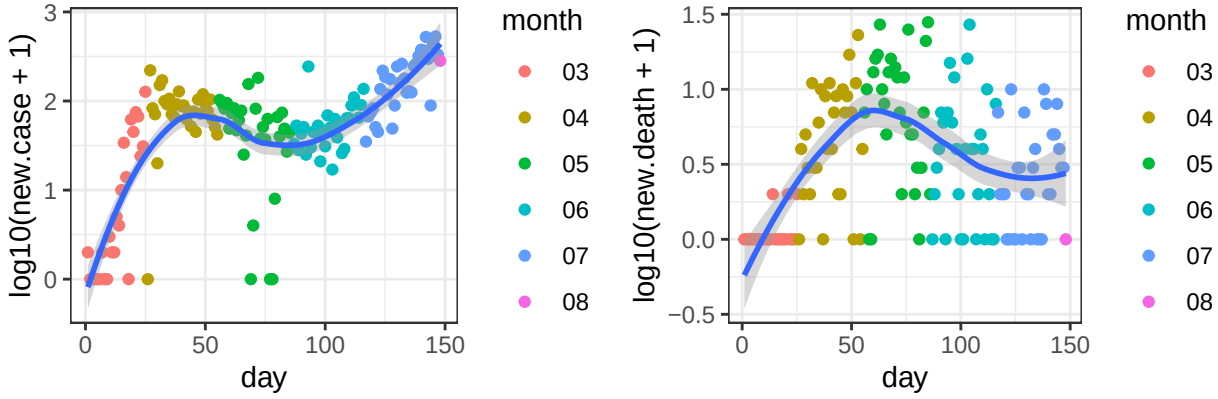
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

King_Washington



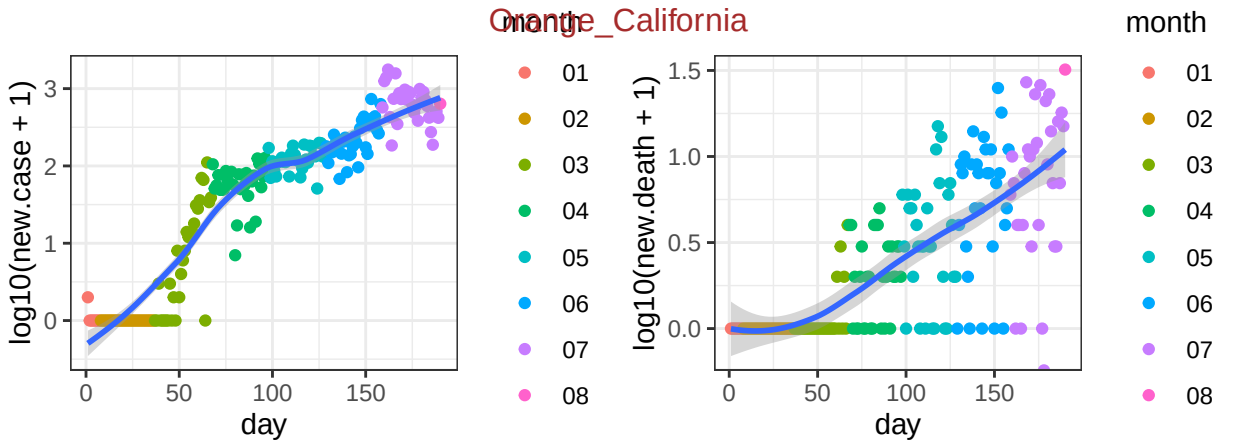
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

St. Louis_Missouri



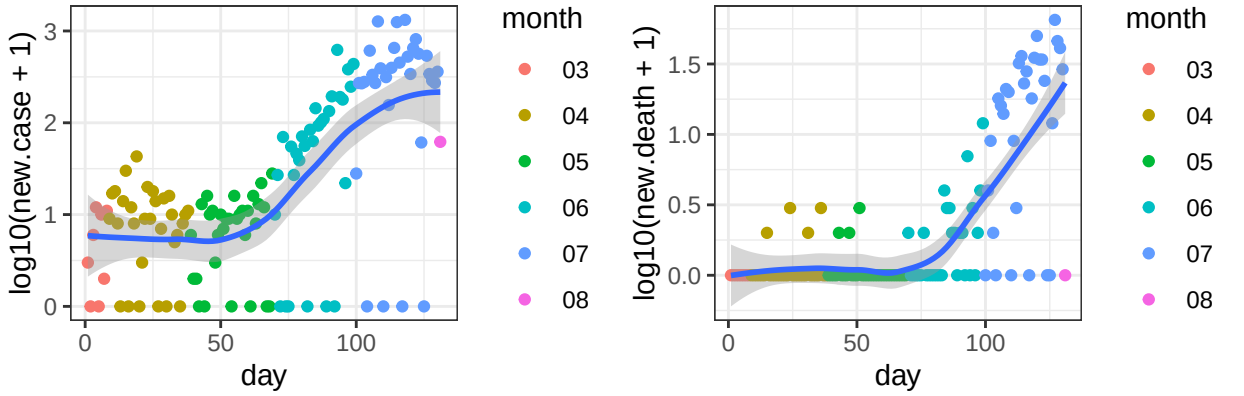
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Orange_California



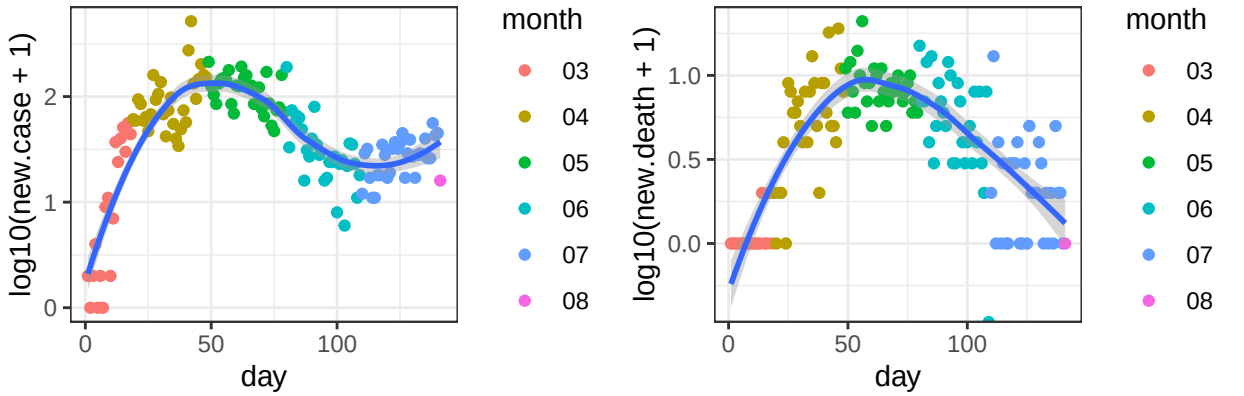
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Hidalgo_Texas

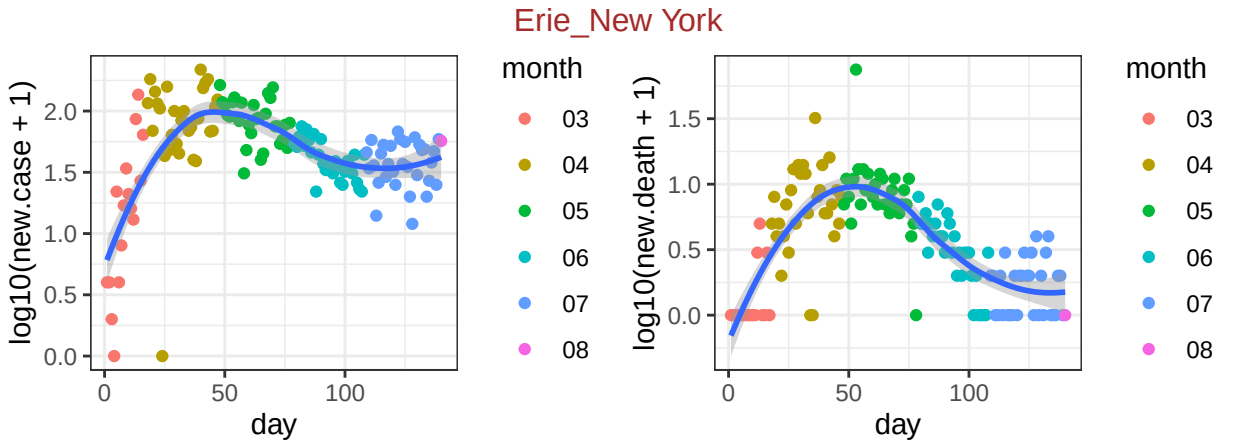


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−24

Bristol_Massachusetts



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

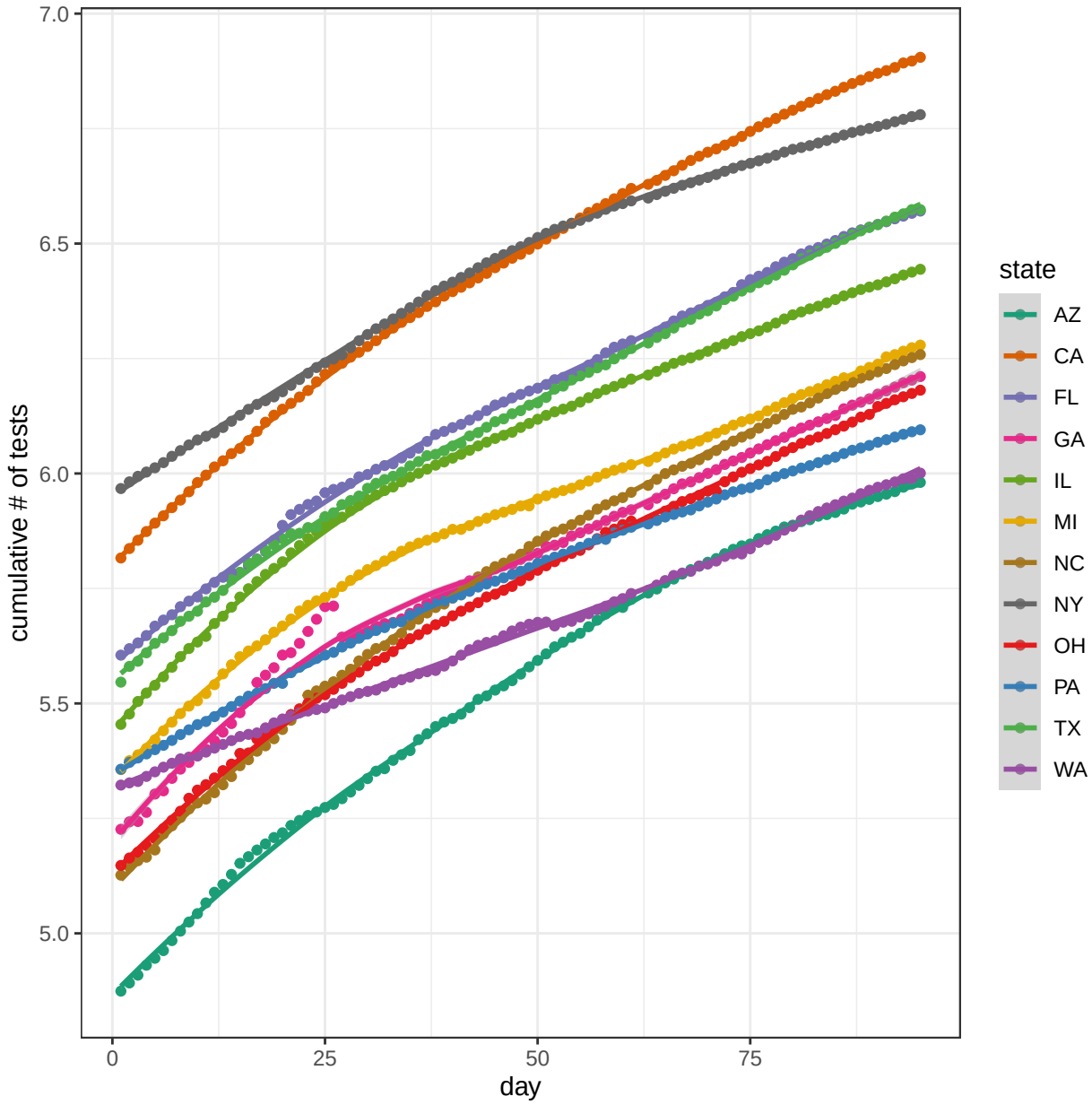


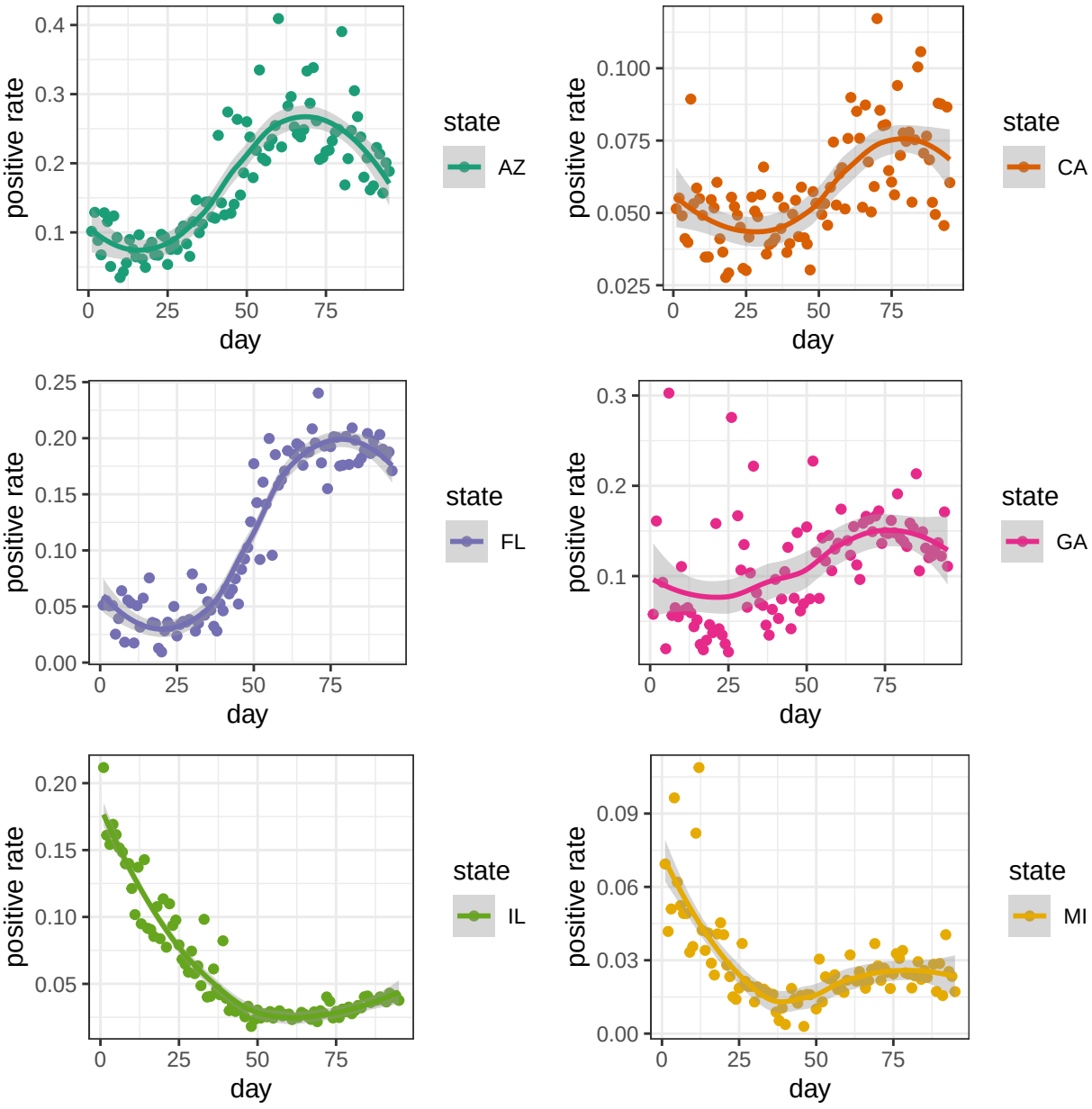
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

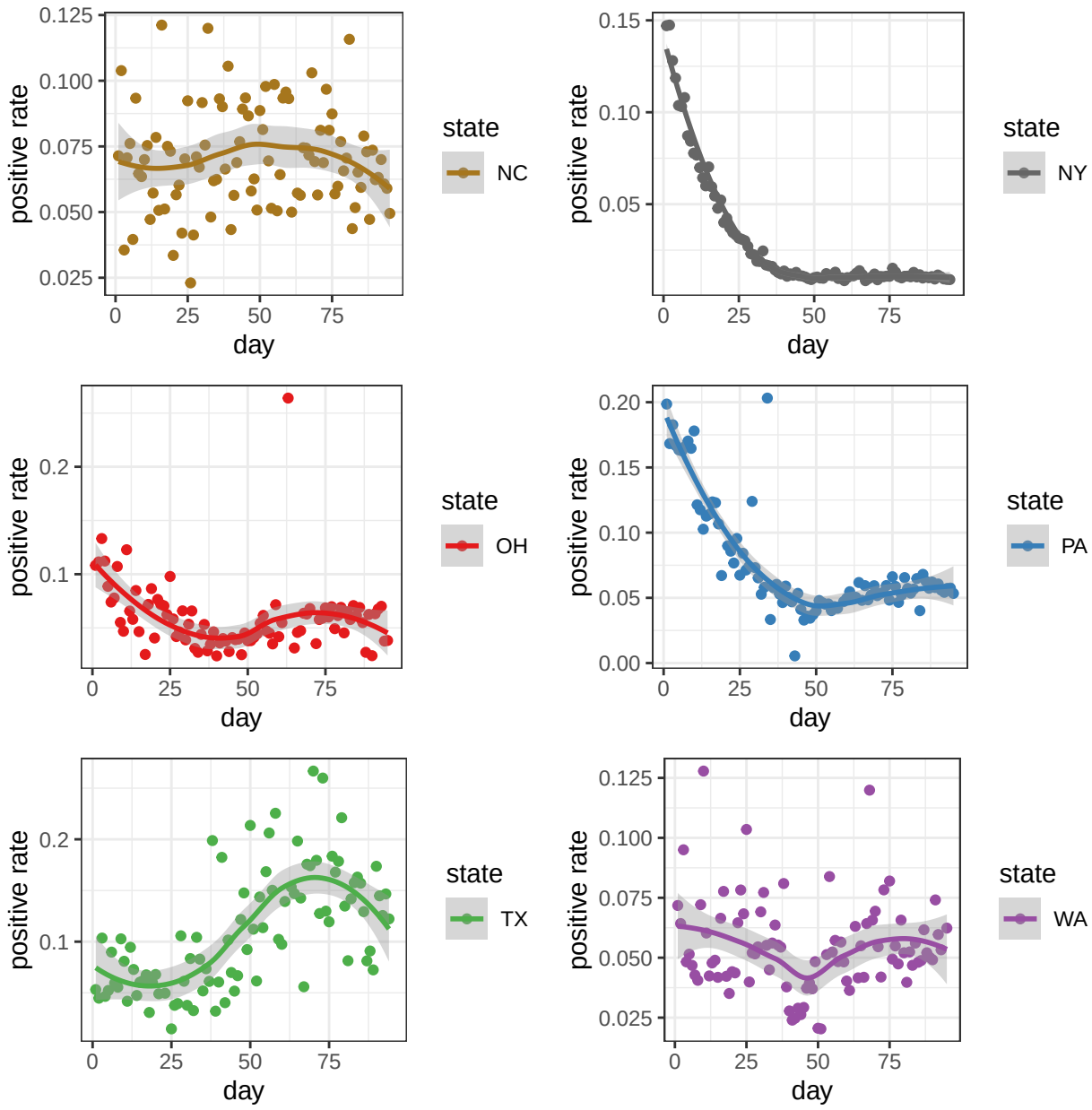
COVID Trackng

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.


7.0
6.5
6.0
5.5
5.0
cumulative # of tests
0
25
50
75
day
state
AZ
CA
FL
GA
IL
MI
NC
NY
OH
PA
TX
WA


positive rate
day
state
AZ
CA
FL
GA
IL
MI



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] RColorBrewer_1.1-2 httr_1.4.1         ggpubr_0.2.5       magrittr_1.5
## [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    farver_2.0.3     purrr_0.3.3      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
```