Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-07-27

Contents

Introduction

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

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

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

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

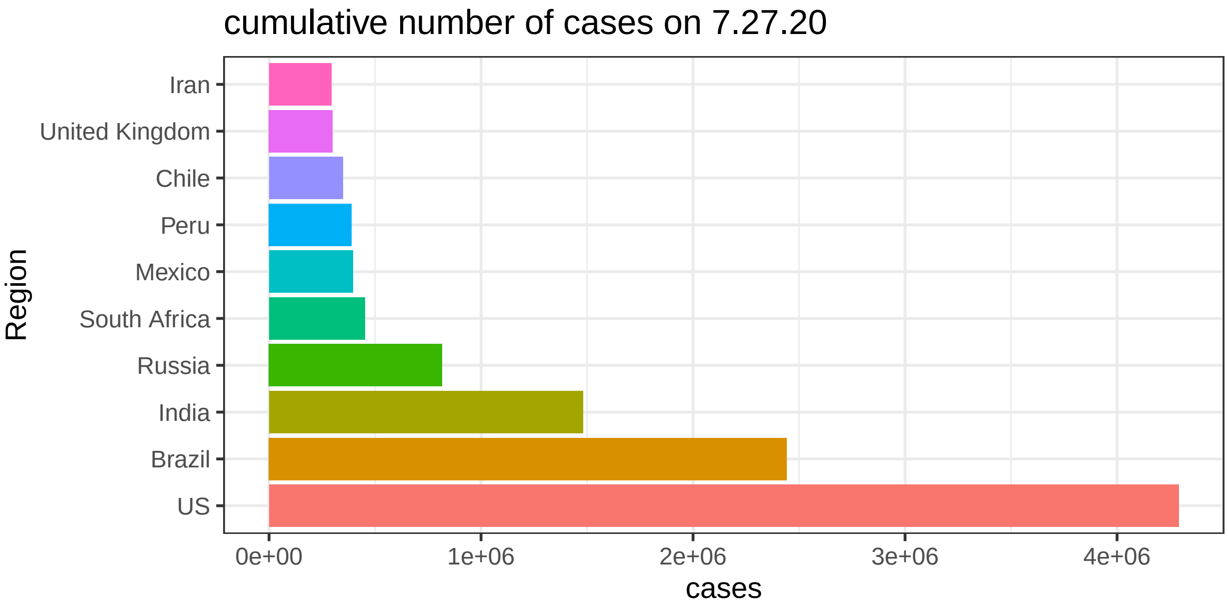
JHU

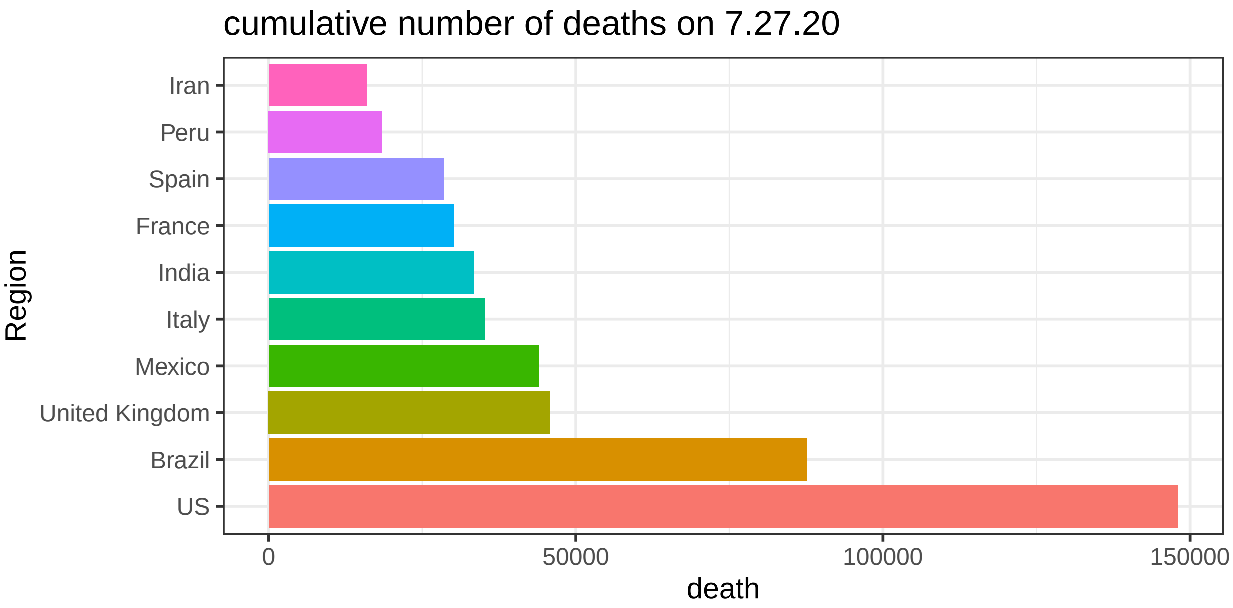
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

time series data

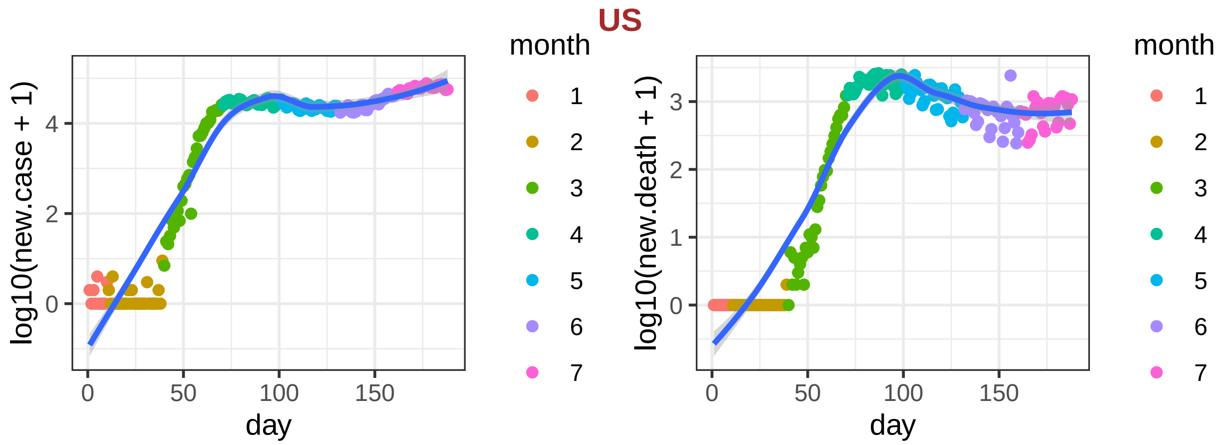
The time series provide counts (e.g., confirmed cases, deaths) starting from Jan 22nd, 2020 for 253 locations. Currently there is no data of individual US state in these time series data files.

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

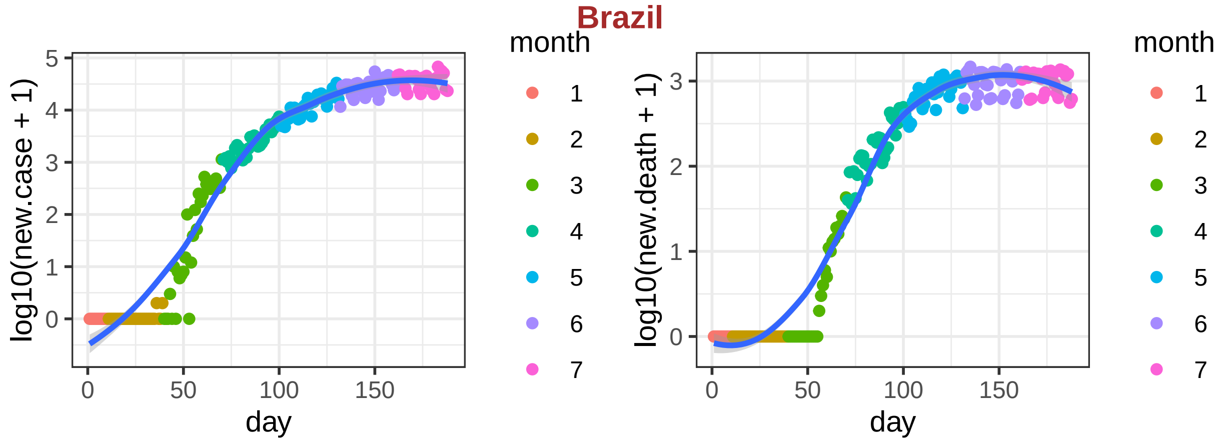




Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

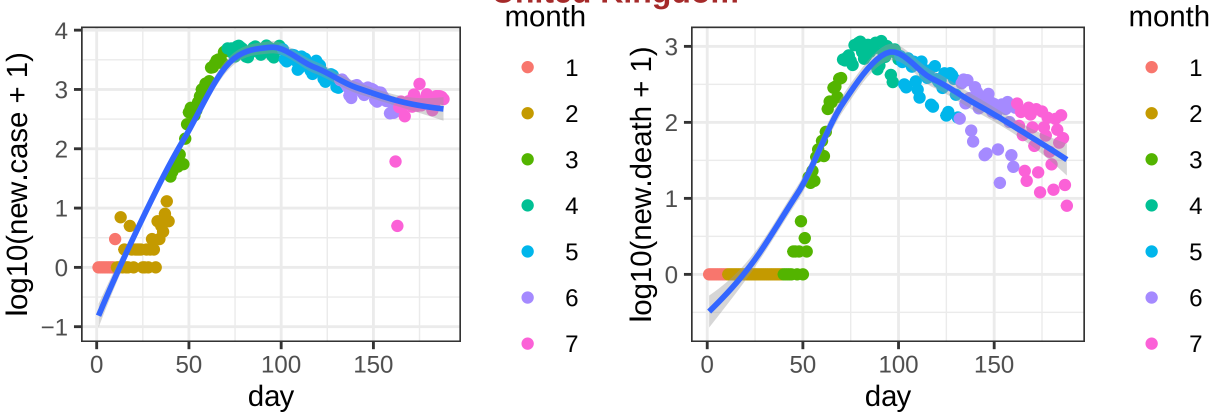


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



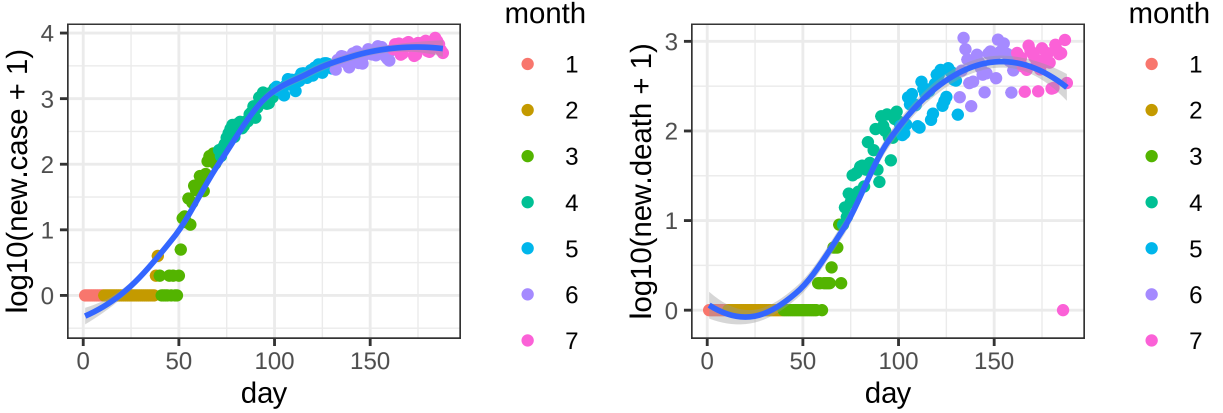
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



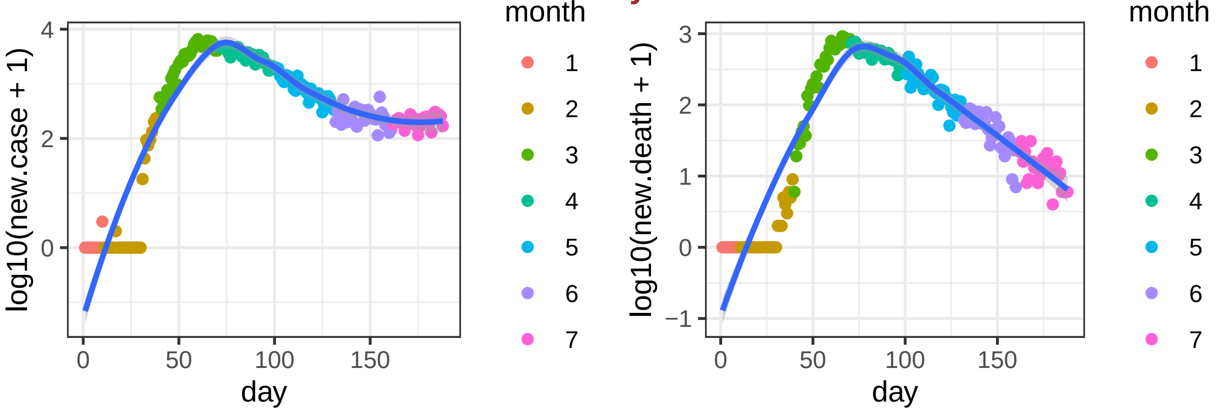
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Mexico



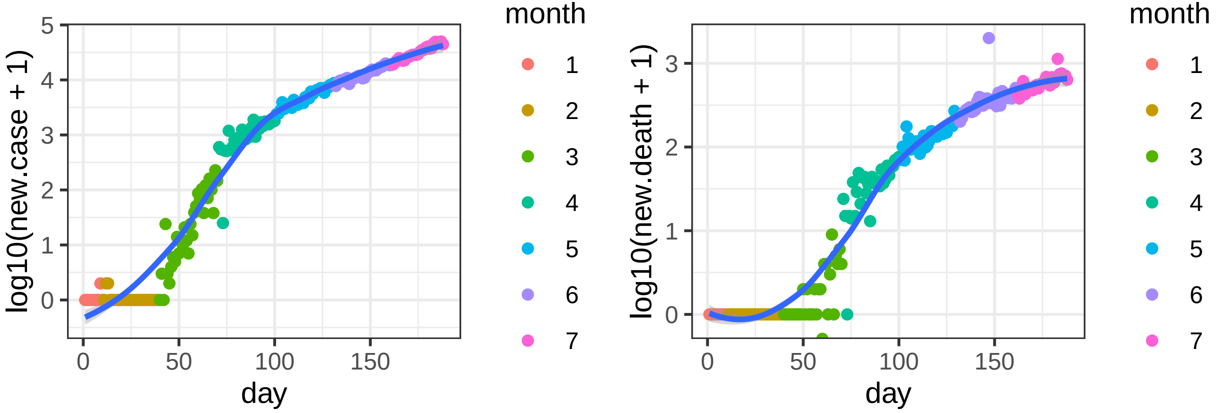
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



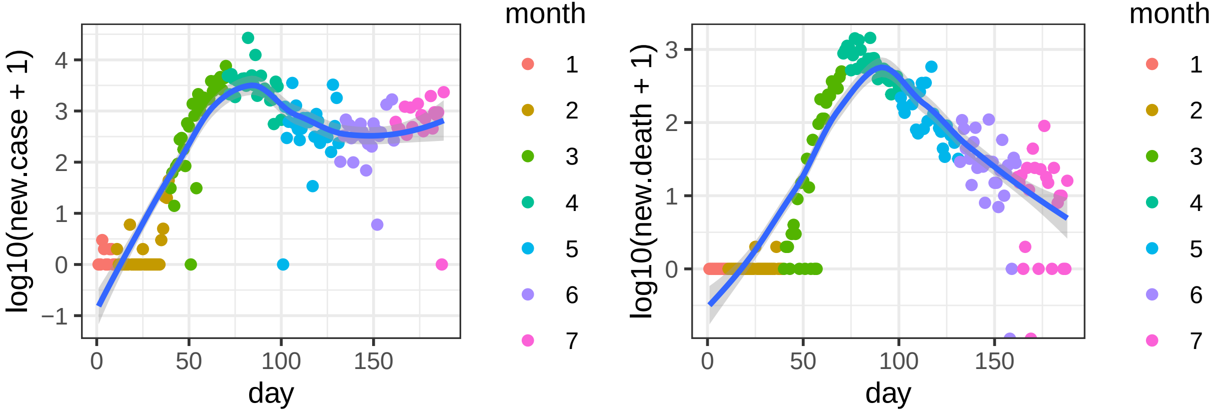
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

India



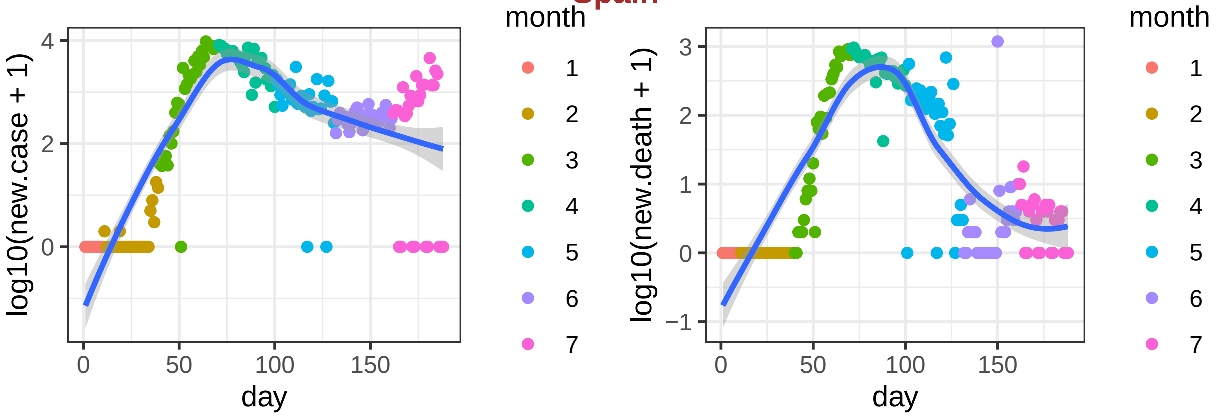
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France

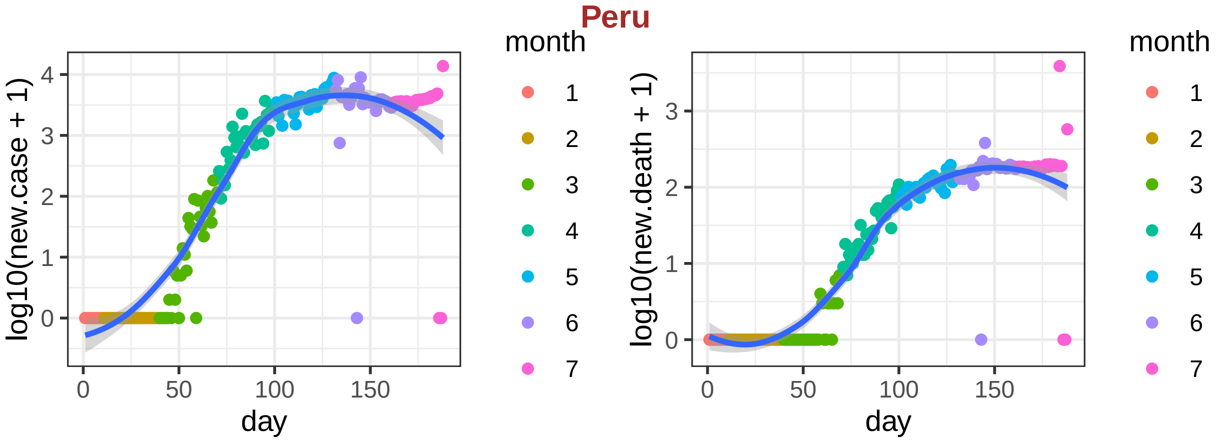


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

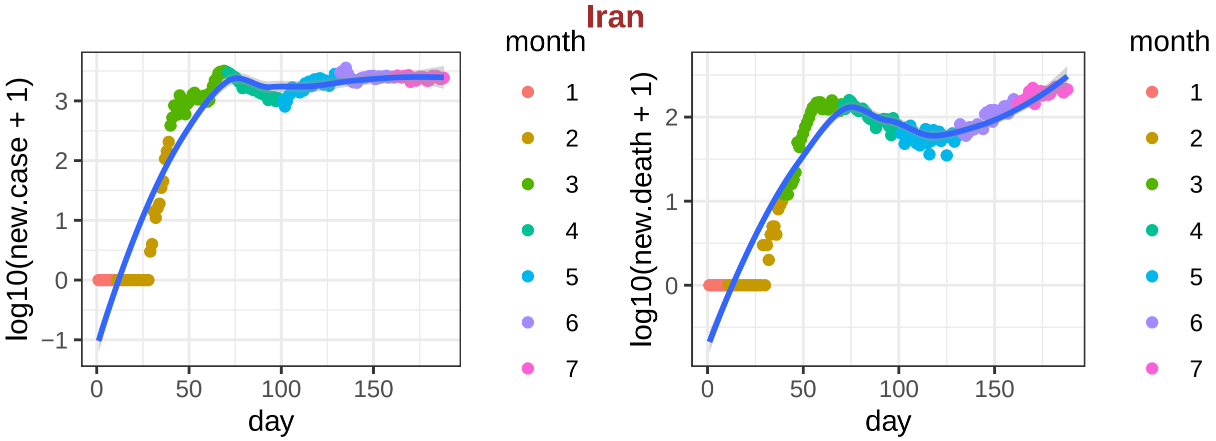
Spain



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



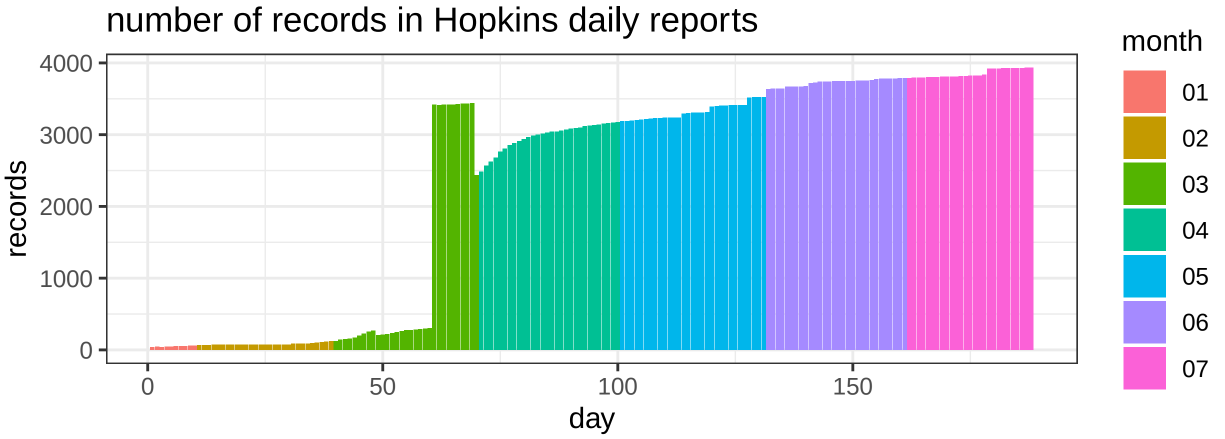
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



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

daily reports data

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



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

NY Times

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

The currente date is

```
## [1] "2020-07-27"
```

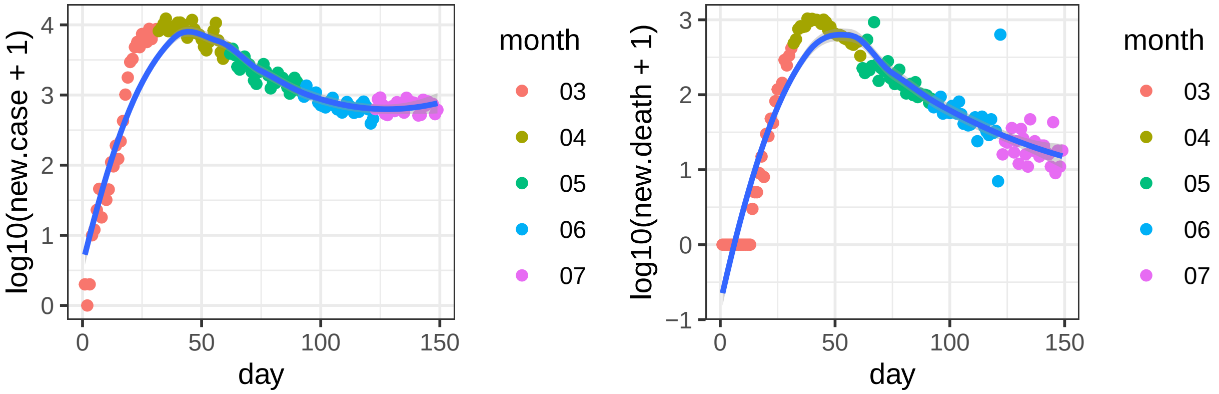
state level data

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

```
##            date          state fips  cases deaths
## 8078 2020-07-27       New York   36 417056  32322
## 8076 2020-07-27     New Jersey   34 181732  15804
## 8049 2020-07-27     California    6 467056   8544
## 8067 2020-07-27  Massachusetts   25 115926   8536
## 8059 2020-07-27       Illinois   17 174442   7619
## 8085 2020-07-27   Pennsylvania   42 112995   7174
## 8068 2020-07-27       Michigan   26  87329   6407
## 8091 2020-07-27          Texas   48 402295   6292
## 8054 2020-07-27        Florida   12 432739   5930
## 8051 2020-07-27    Connecticut    9  48983   4418
## 8064 2020-07-27      Louisiana   22 110029   3786
## 8066 2020-07-27       Maryland   24  85436   3447
## 8055 2020-07-27        Georgia   13 155907   3435
## 8082 2020-07-27           Ohio   39  85177   3344
## 8047 2020-07-27        Arizona    4 163918   3320
## 8060 2020-07-27        Indiana   18  64417   2906
## 8095 2020-07-27       Virginia   51  86072   2082
## 8079 2020-07-27 North Carolina   37 114689   1815
## 8050 2020-07-27       Colorado    8  44723   1800
## 8069 2020-07-27      Minnesota   27  51843   1616
## 8096 2020-07-27     Washington   53  55548   1611
## 8088 2020-07-27 South Carolina   45  82417   1506
## 8070 2020-07-27    Mississippi   28  52957   1501
## 8045 2020-07-27        Alabama    1  81115   1491
## 8071 2020-07-27       Missouri   29  44813   1245
## 8087 2020-07-27   Rhode Island   44  18515   1004
## 8090 2020-07-27      Tennessee   47  93869    965
## 8098 2020-07-27      Wisconsin   55  53323    905
## 8061 2020-07-27           Iowa   19  42674    834
## 8074 2020-07-27         Nevada   32  43880    739
```

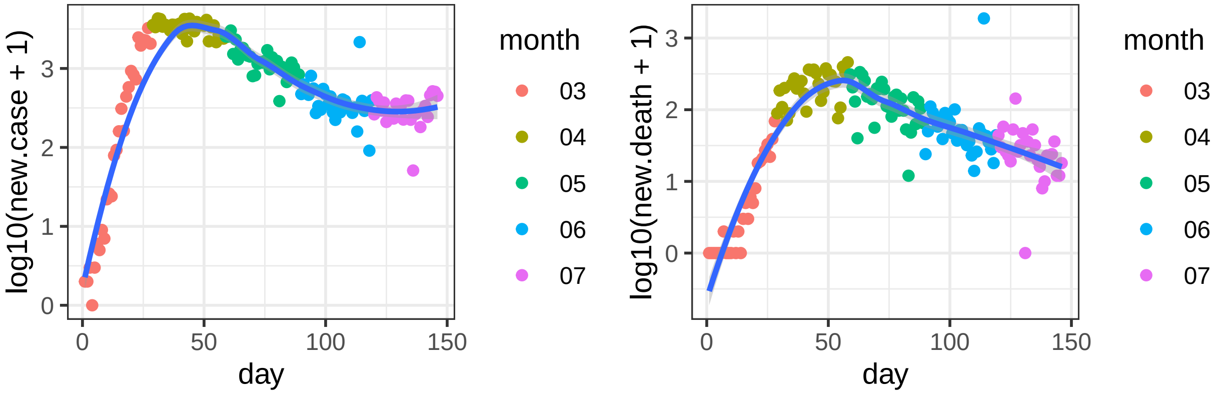
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



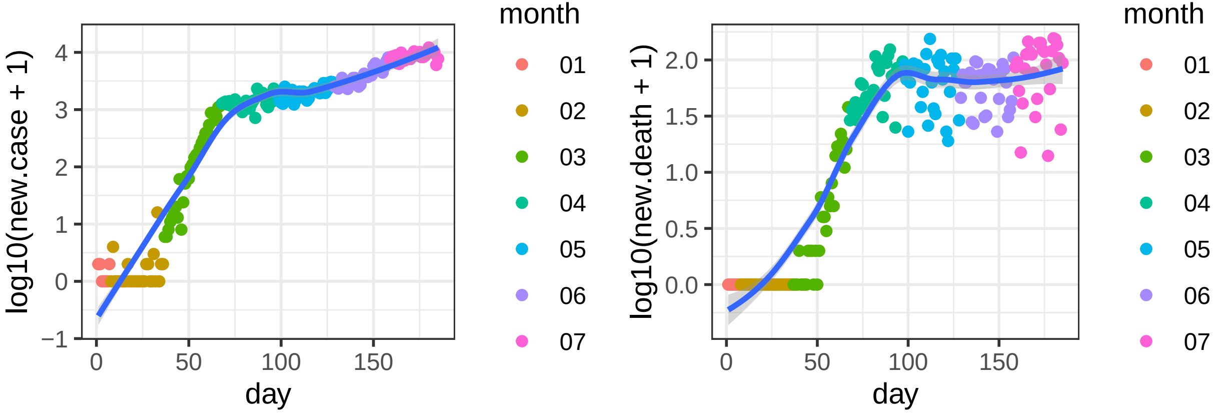
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



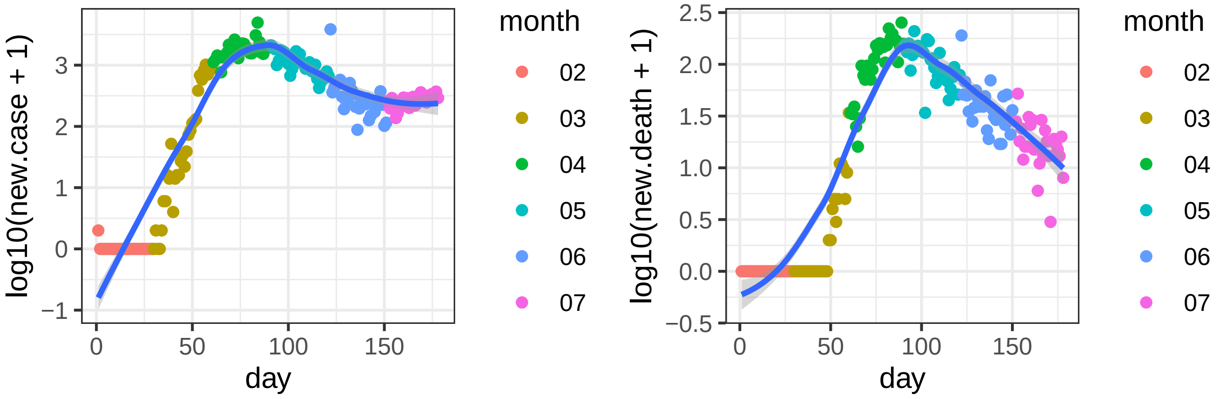
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

California



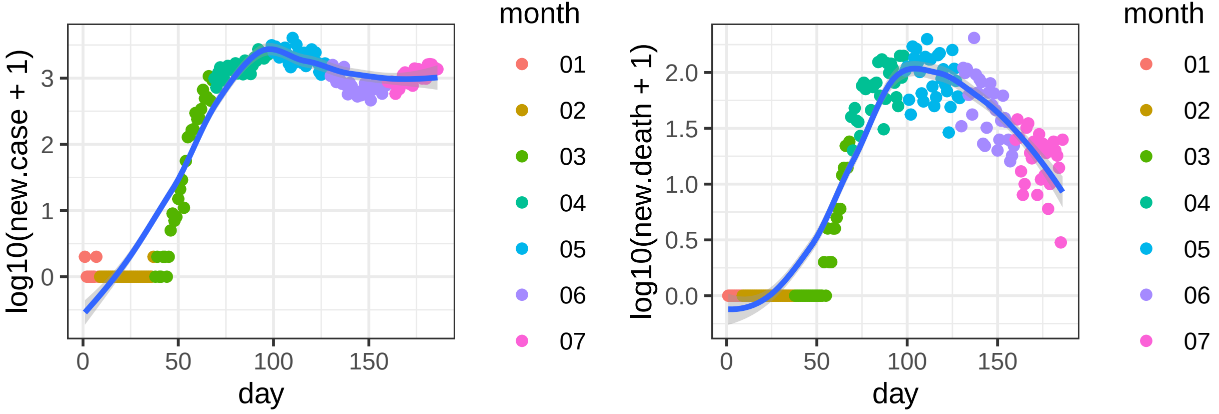
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Massachusetts



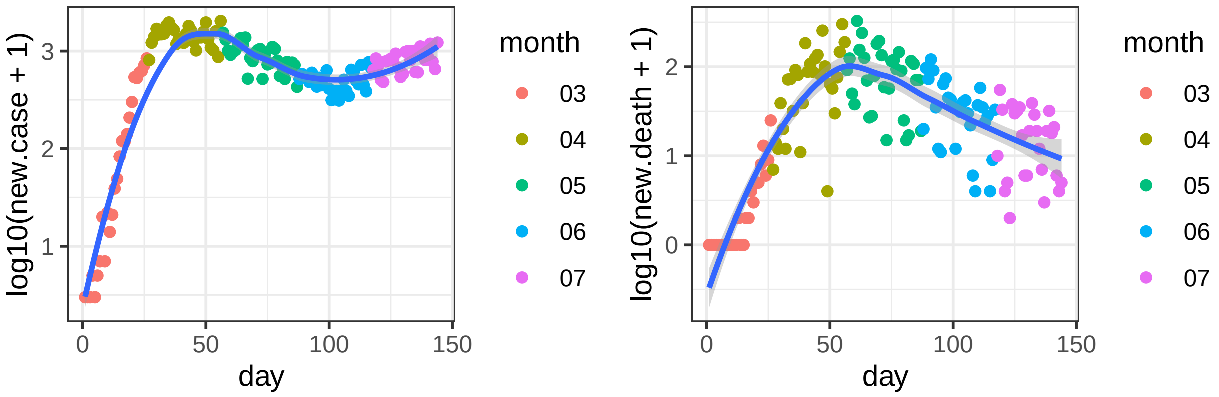
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Illinois



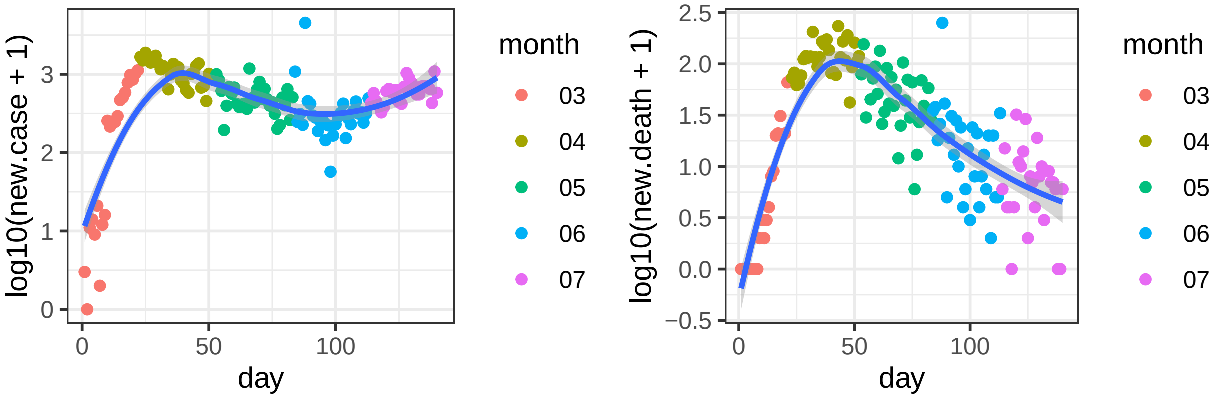
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Pennsylvania



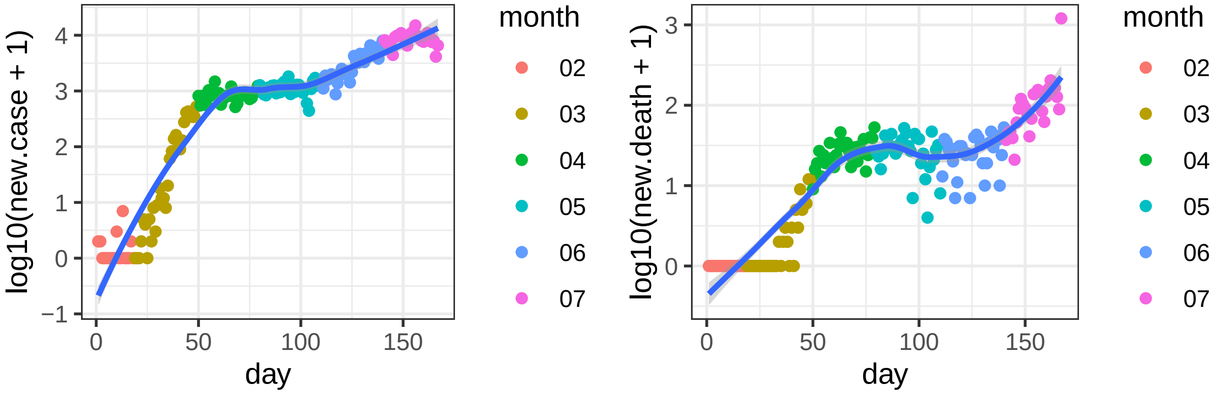
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Michigan



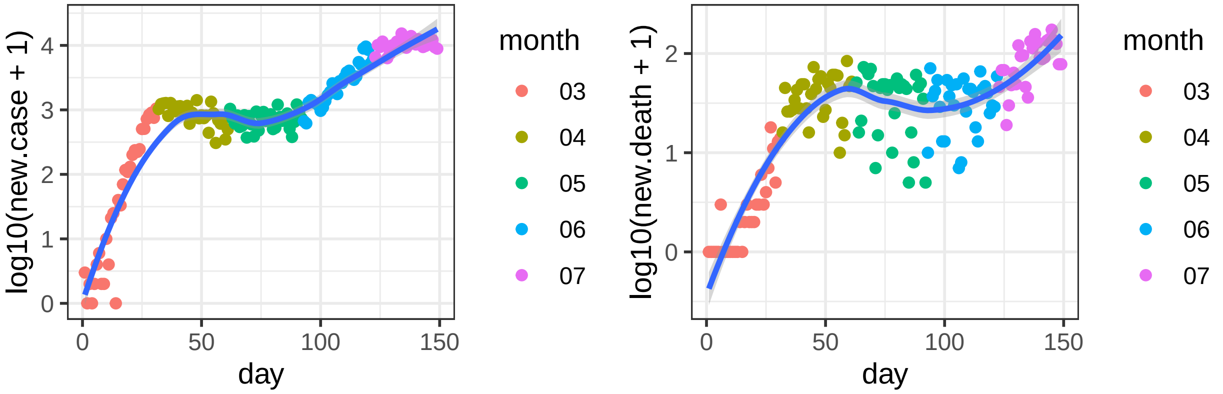
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Texas



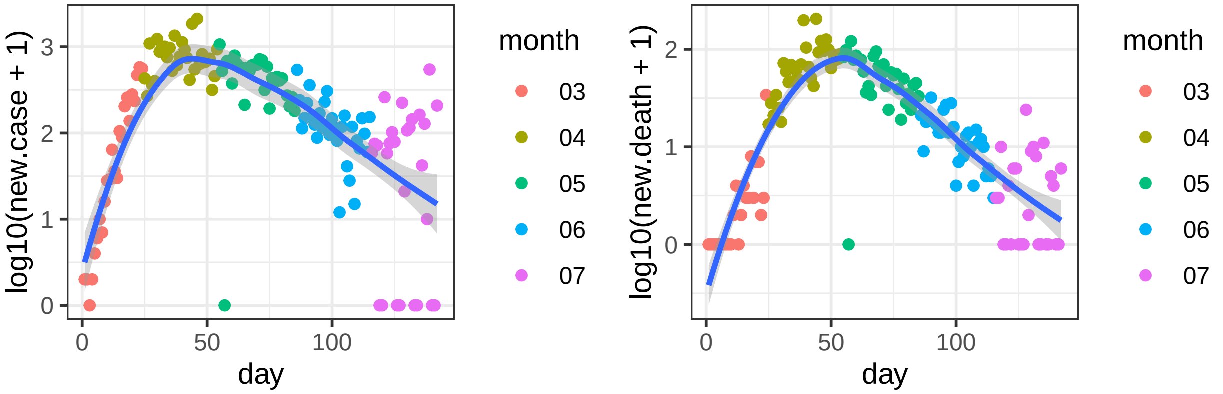
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Florida



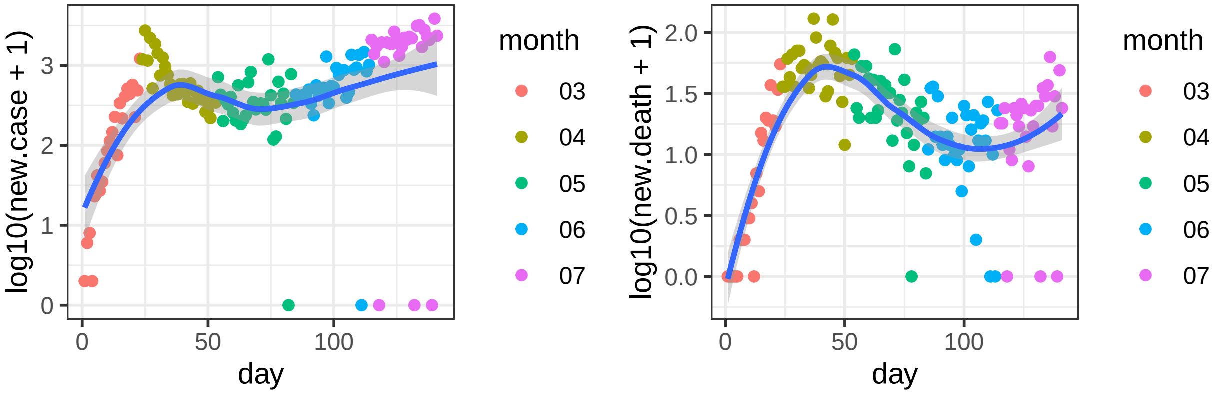
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Connecticut



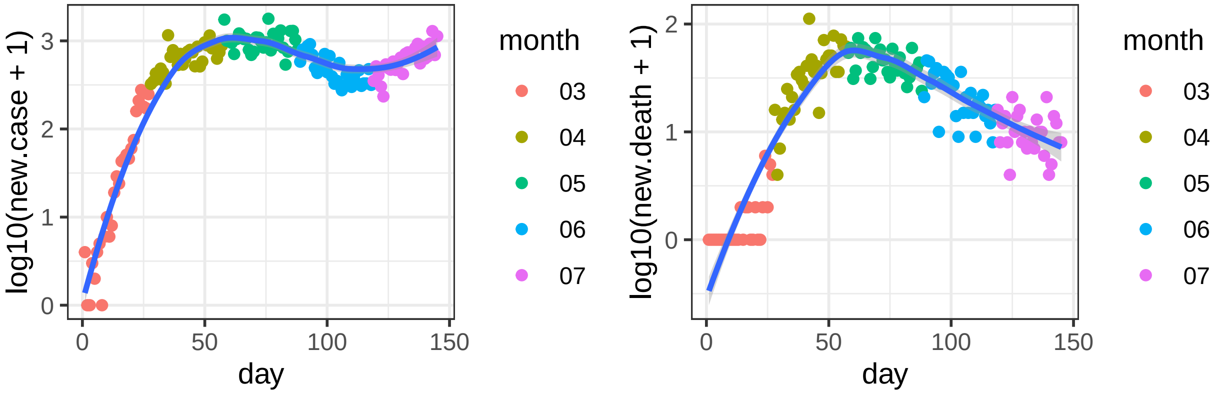
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Louisiana



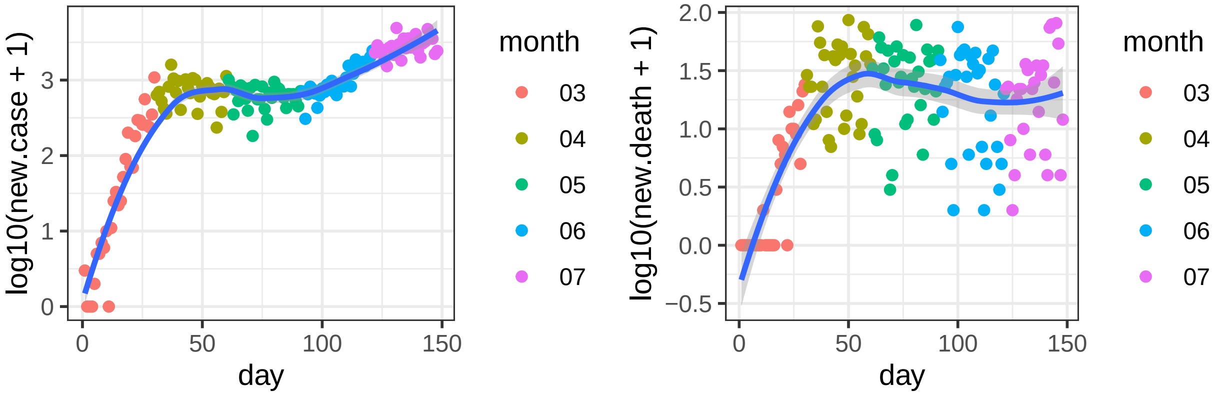
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



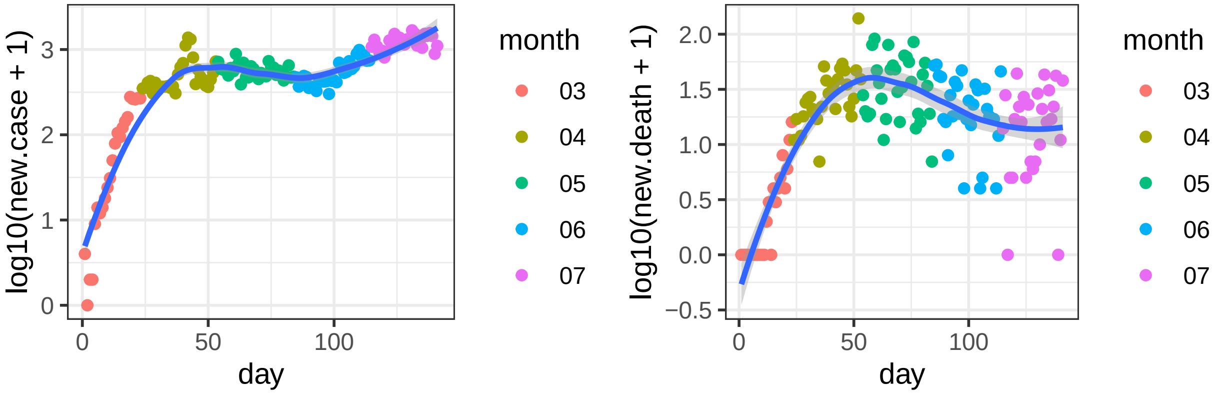
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Georgia



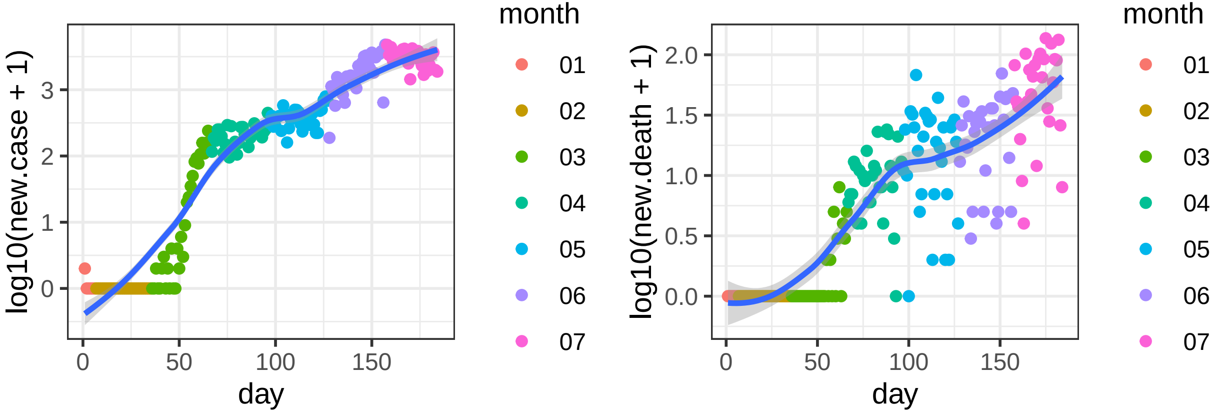
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Ohio



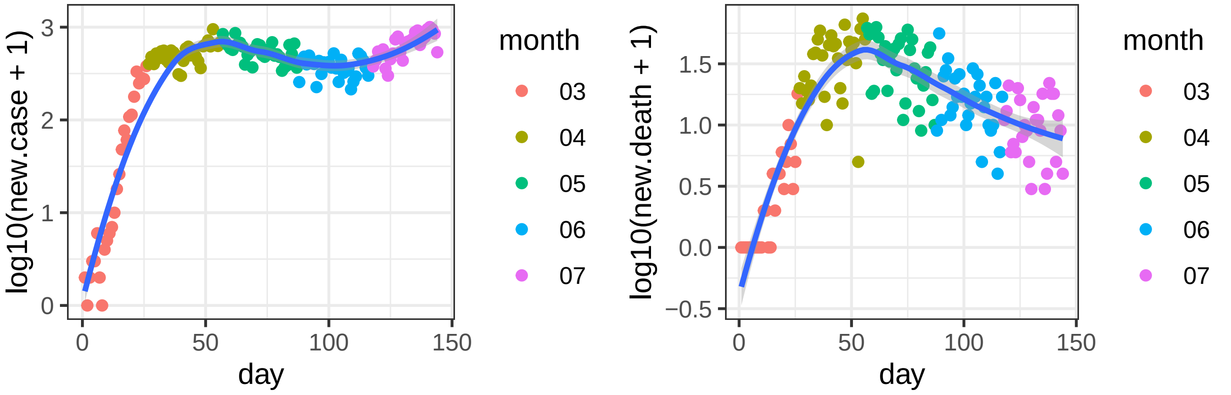
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Arizona



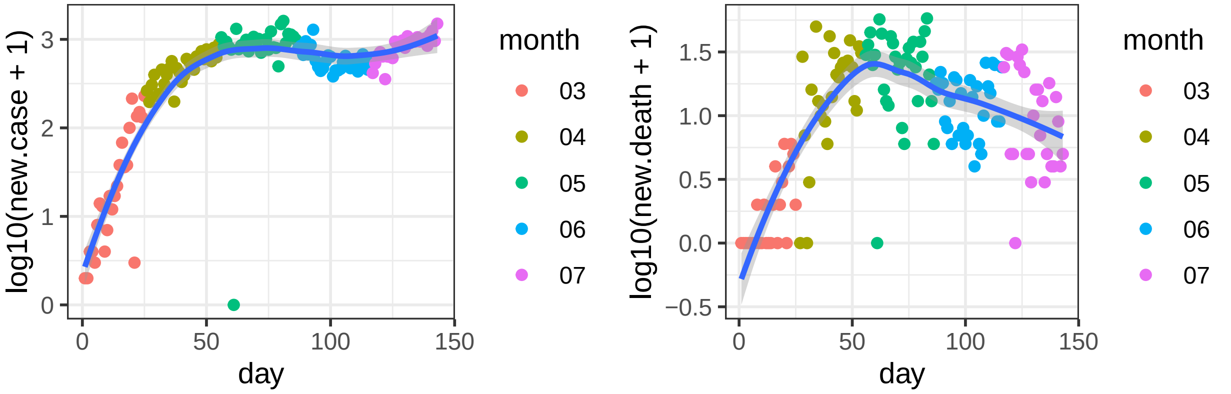
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Indiana



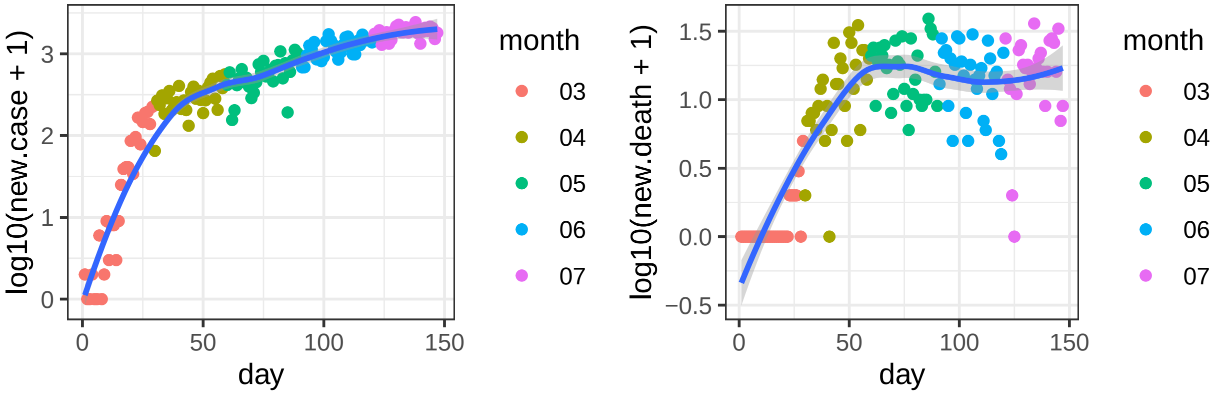
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Virginia



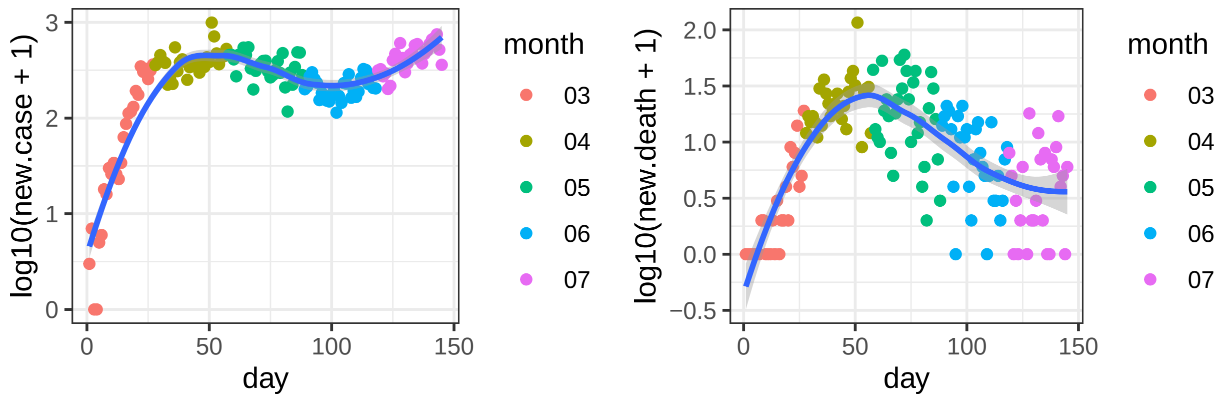
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

North Carolina



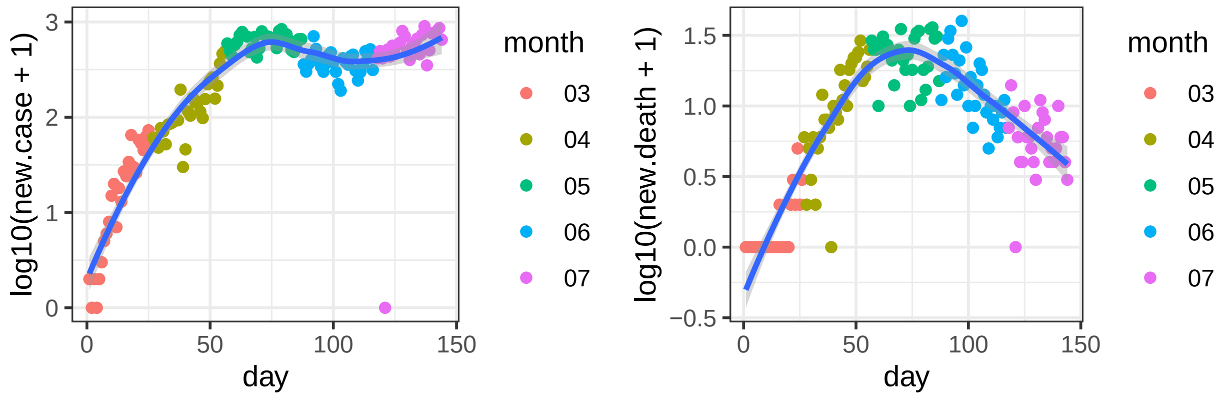
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Colorado



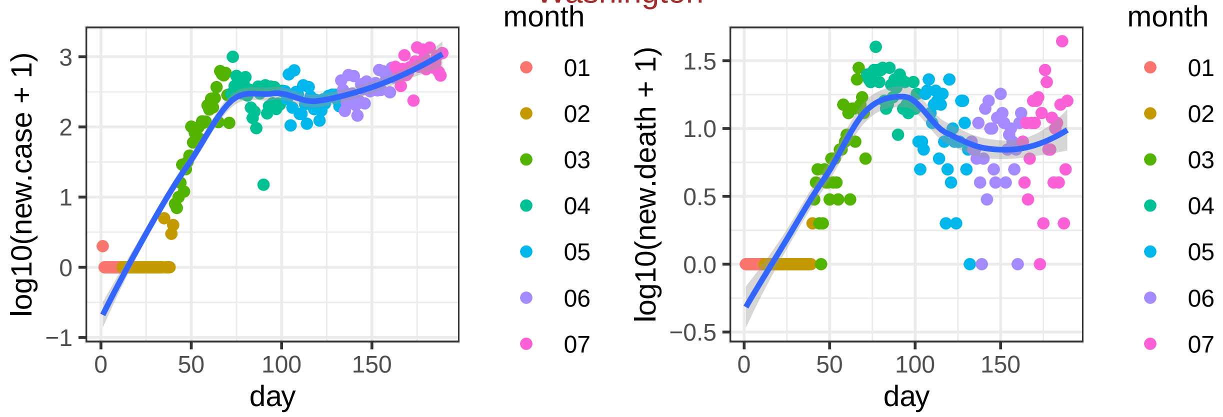
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Minnesota



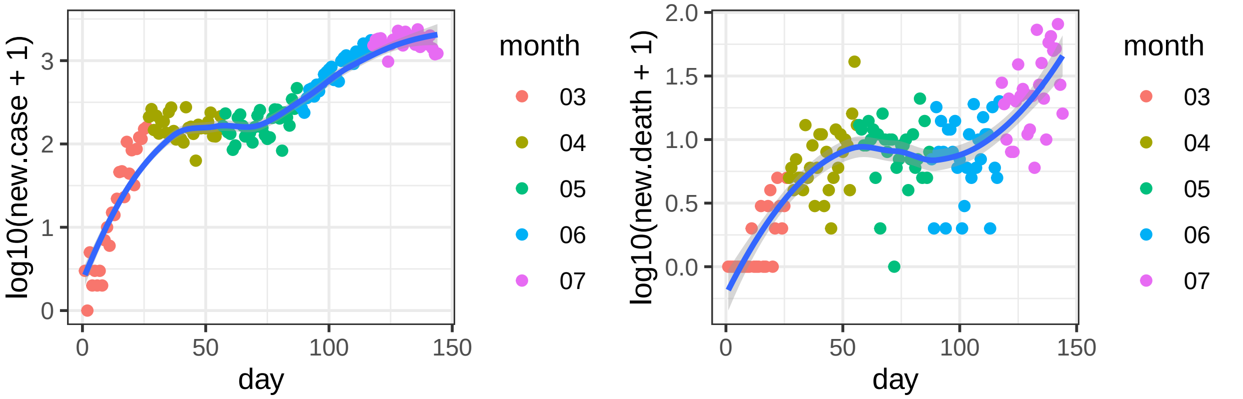
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Washington



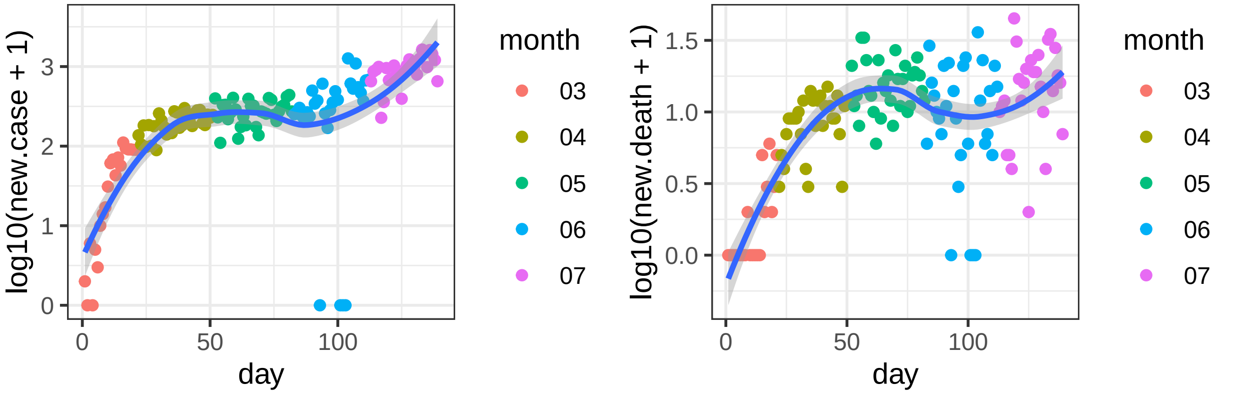
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

South Carolina



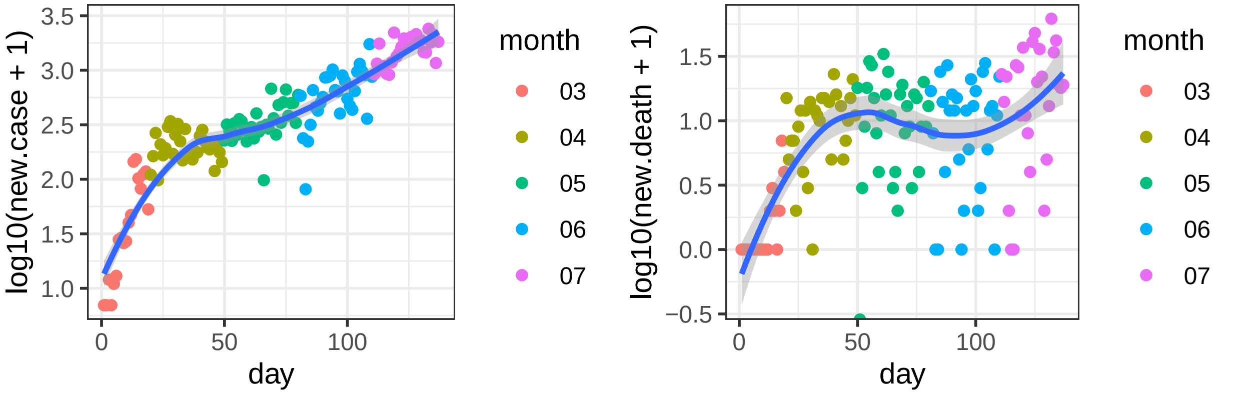
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Mississippi



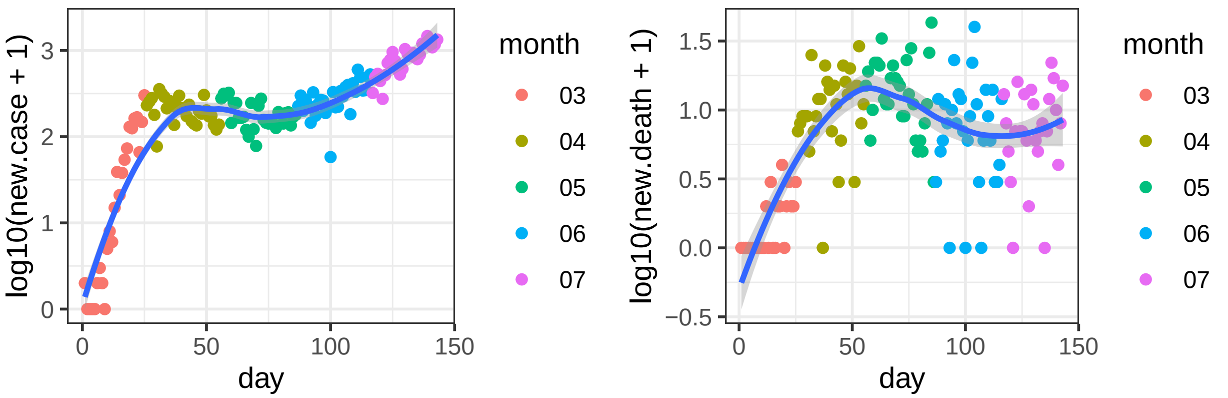
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Alabama



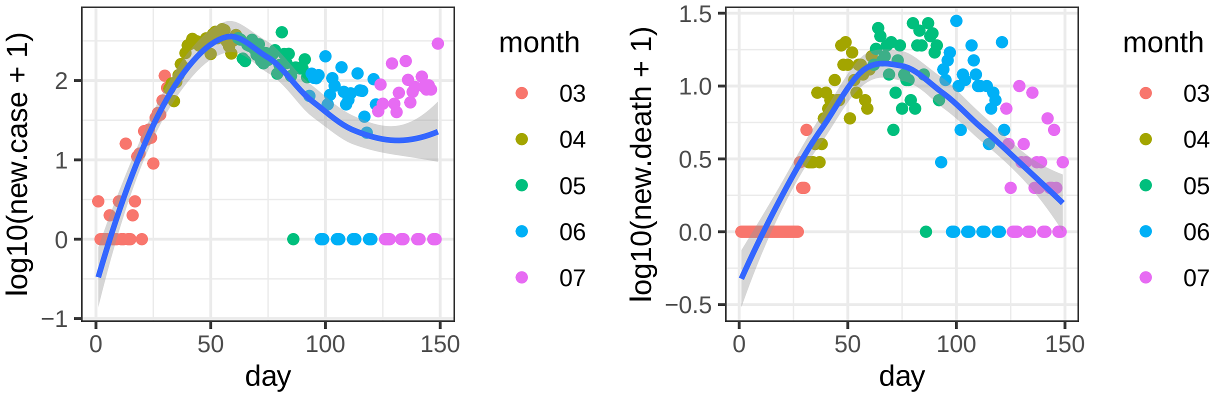
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Missouri



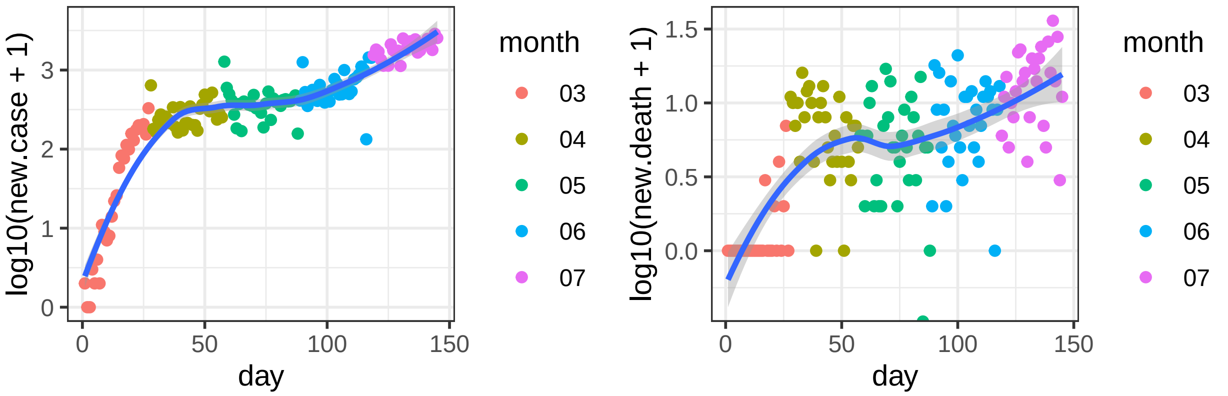
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Rhode Island



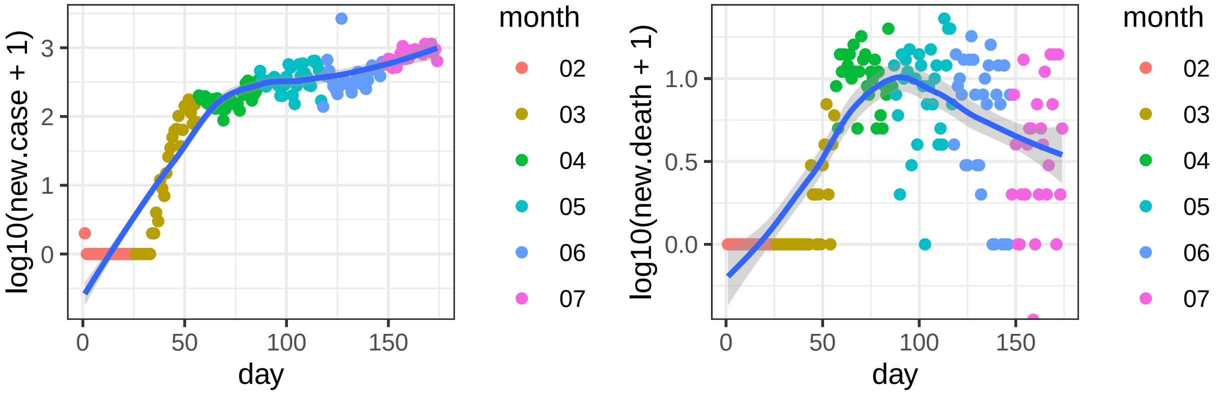
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Tennessee



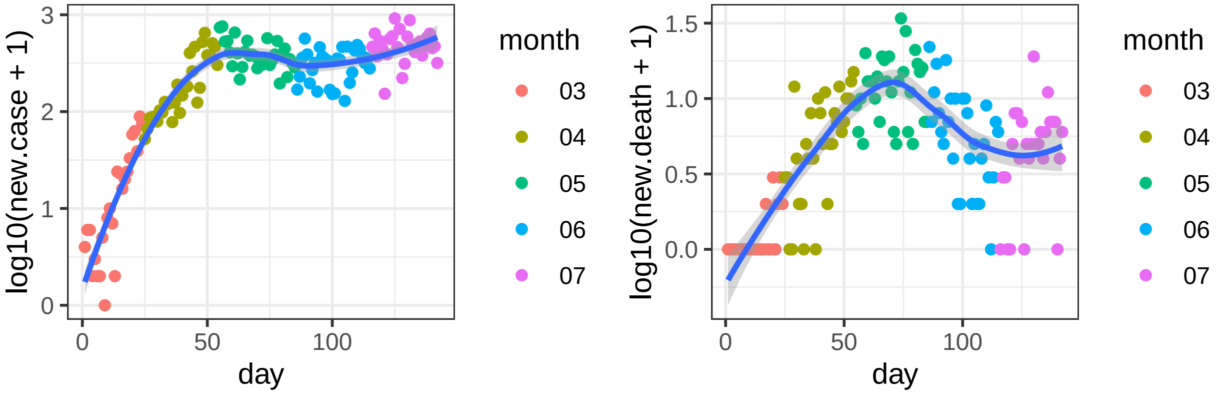
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Wisconsin



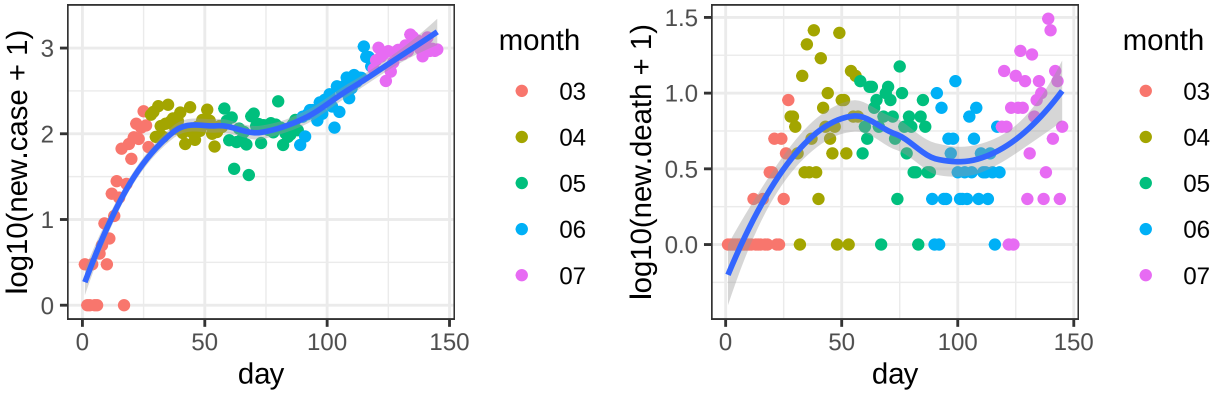
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

Iowa



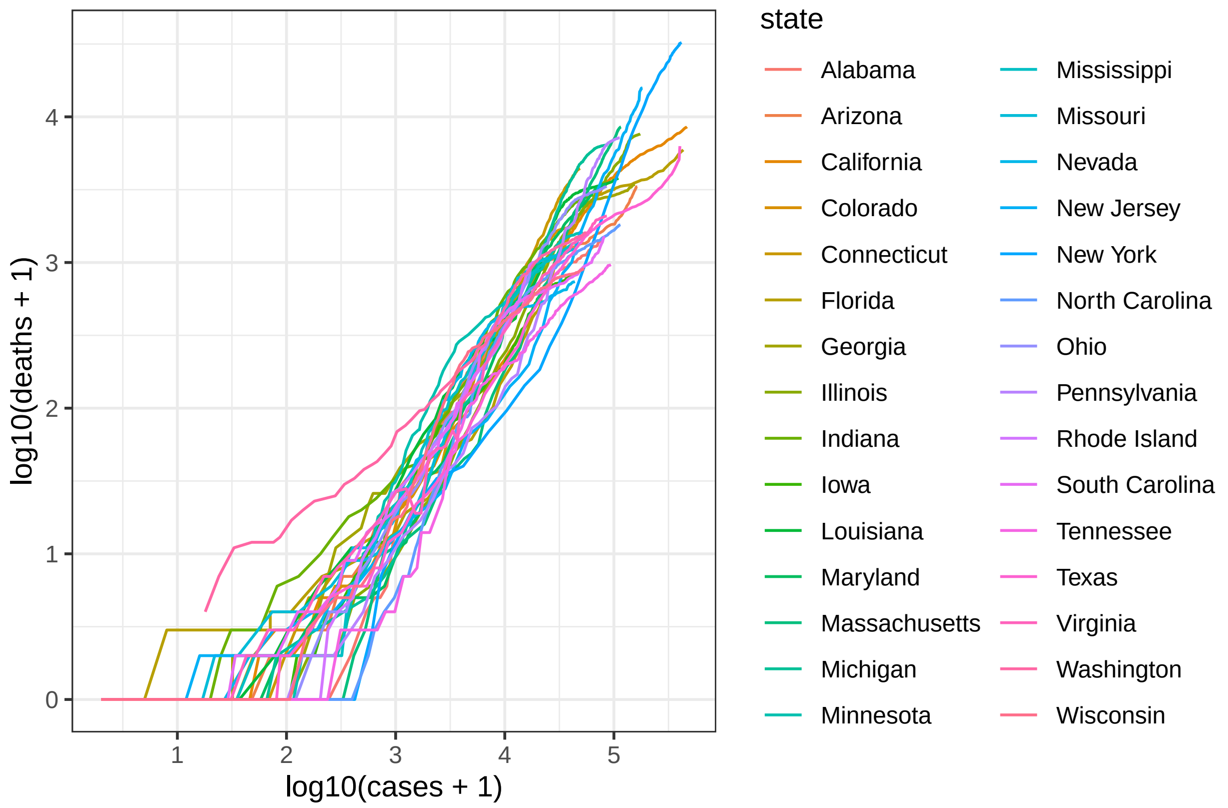
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

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



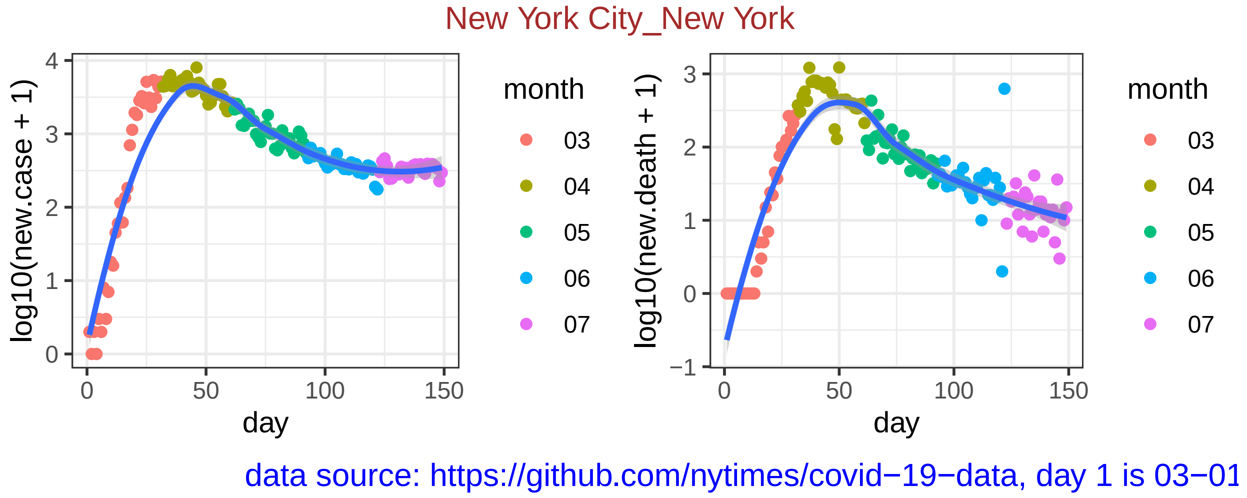
data source: https://github.com/nytimes/covid−19−data

county level data

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

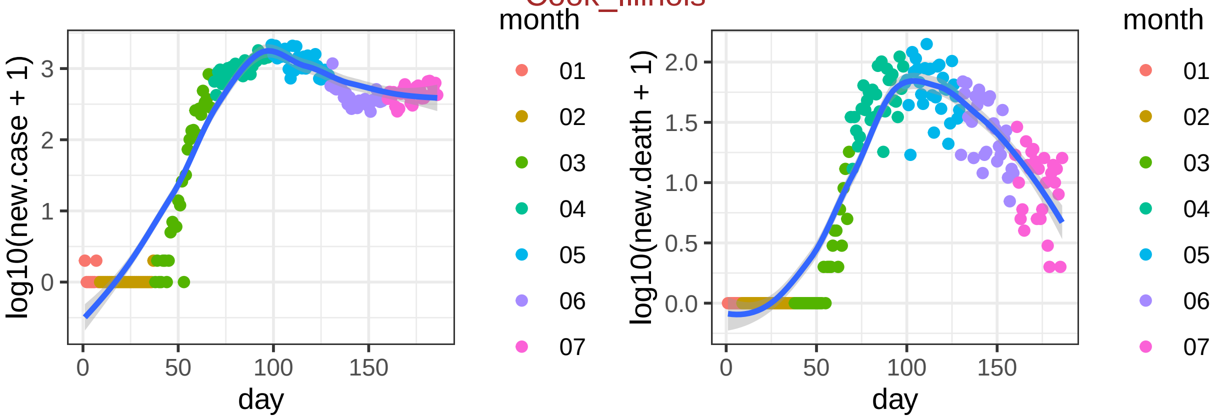
```
##              date         county        state  fips  cases deaths
## 374970 2020-07-27  New York City     New York    NA 228740  22970
## 373730 2020-07-27           Cook     Illinois 17031 103008   4845
## 373325 2020-07-27    Los Angeles   California  6037 176028   4375
## 374437 2020-07-27          Wayne     Michigan 26163  26186   2785
## 374969 2020-07-27         Nassau     New York 36059  43017   2706
## 374894 2020-07-27          Essex   New Jersey 34013  19582   2103
## 374889 2020-07-27         Bergen   New Jersey 34003  20543   2045
## 374989 2020-07-27        Suffolk     New York 36103  42967   2043
## 374348 2020-07-27      Middlesex Massachusetts 25017  25396   1959
## 373223 2020-07-27       Maricopa      Arizona  4013 109988   1807
## 375406 2020-07-27   Philadelphia Pennsylvania 42101  29803   1678
## 374997 2020-07-27    Westchester     New York 36119  35798   1577
## 374896 2020-07-27         Hudson   New Jersey 34017  19562   1502
## 373428 2020-07-27       Hartford  Connecticut  9003  12390   1410
## 373483 2020-07-27     Miami-Dade      Florida 12086 107314   1404
## 374899 2020-07-27      Middlesex   New Jersey 34023  17769   1404
## 373427 2020-07-27      Fairfield  Connecticut  9001  17459   1402
## 374907 2020-07-27          Union   New Jersey 34039  16551   1347
## 374903 2020-07-27        Passaic   New Jersey 34031  17483   1247
## 374344 2020-07-27          Essex Massachusetts 25009  16990   1169
## 374417 2020-07-27        Oakland     Michigan 26125  14111   1127
## 375814 2020-07-27         Harris        Texas 48201  66195   1100
## 373431 2020-07-27            New Haven          Connecticut  9009  12941   1097
## 374352 2020-07-27              Suffolk        Massachusetts 25025  20914   1048
## 374902 2020-07-27                Ocean           New Jersey 34029  10308   1012
## 374354 2020-07-27            Worcester        Massachusetts 25027  13140    980
## 374350 2020-07-27              Norfolk        Massachusetts 25021  10013    977
## 374404 2020-07-27               Macomb             Michigan 26099   9224    942
## 374900 2020-07-27             Monmouth           New Jersey 34025  10008    855
## 375401 2020-07-27           Montgomery         Pennsylvania 42091   9555    844
## 374901 2020-07-27               Morris           New Jersey 34027   7229    828
## 374465 2020-07-27             Hennepin            Minnesota 27053  16506    807
## 375505 2020-07-27           Providence         Rhode Island 44007  14033    801
## 374330 2020-07-27           Montgomery             Maryland 24031  17203    787
## 373866 2020-07-27               Marion              Indiana 18097  13953    760
## 373490 2020-07-27           Palm Beach              Florida 12099  30956    758
## 374331 2020-07-27      Prince George's             Maryland 24033  22224    731
## 375378 2020-07-27             Delaware         Pennsylvania 42045   8371    720
## 374351 2020-07-27             Plymouth        Massachusetts 25023   9000    705
## 374346 2020-07-27              Hampden        Massachusetts 25013   7309    688
## 373338 2020-07-27            Riverside           California  6065  35636    671
## 376161 2020-07-27                 King           Washington 53033  14638    664
## 374712 2020-07-27            St. Louis             Missouri 29189  11507    636
## 374955 2020-07-27                 Erie             New York 36029   8327    619
## 374342 2020-07-27              Bristol        Massachusetts 25005   8907    617
## 374898 2020-07-27               Mercer           New Jersey 34021   7977    612
## 373446 2020-07-27              Broward              Florida 12011  50784    607
## 375770 2020-07-27               Dallas                Texas 48113  47239    607
## 374864 2020-07-27                Clark               Nevada 32003  37492    606
## 373440 2020-07-27 District of Columbia District of Columbia 11001  11858    582
```

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


New York City_New York

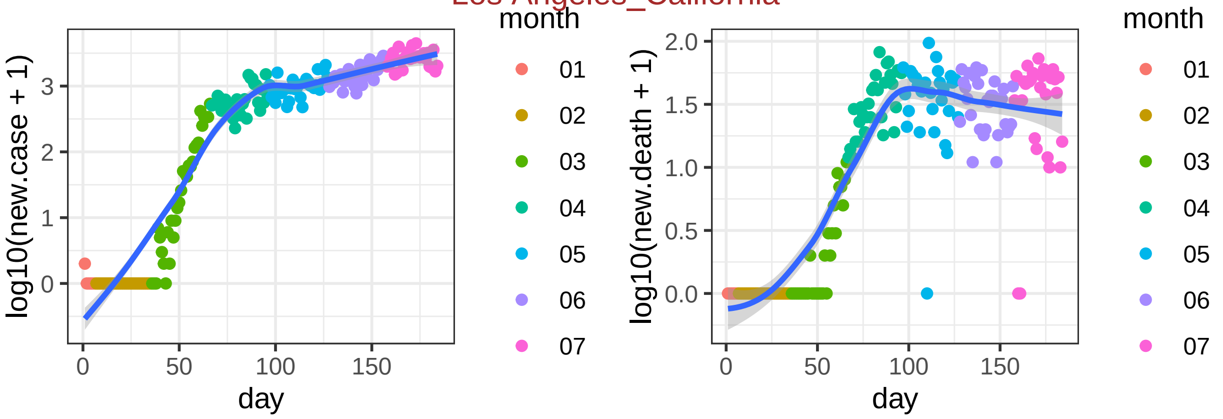
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



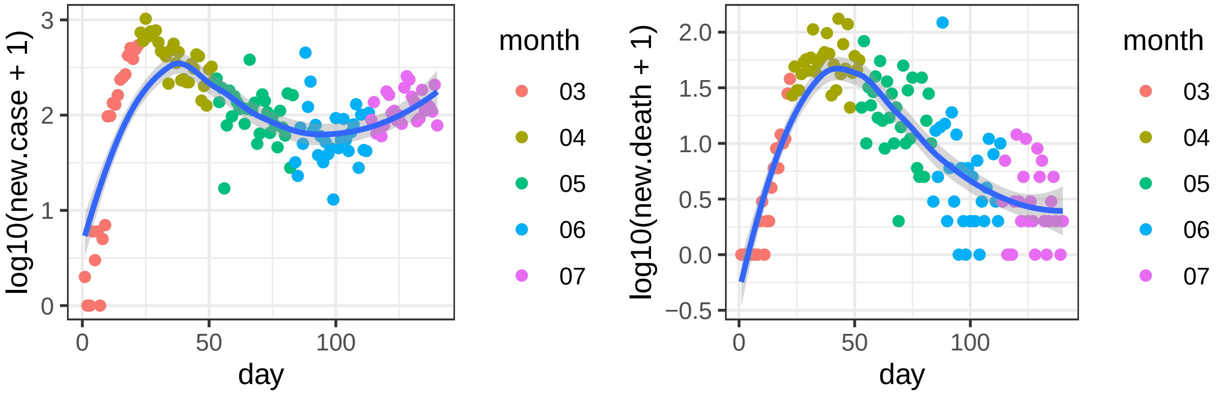
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Los Angeles_California



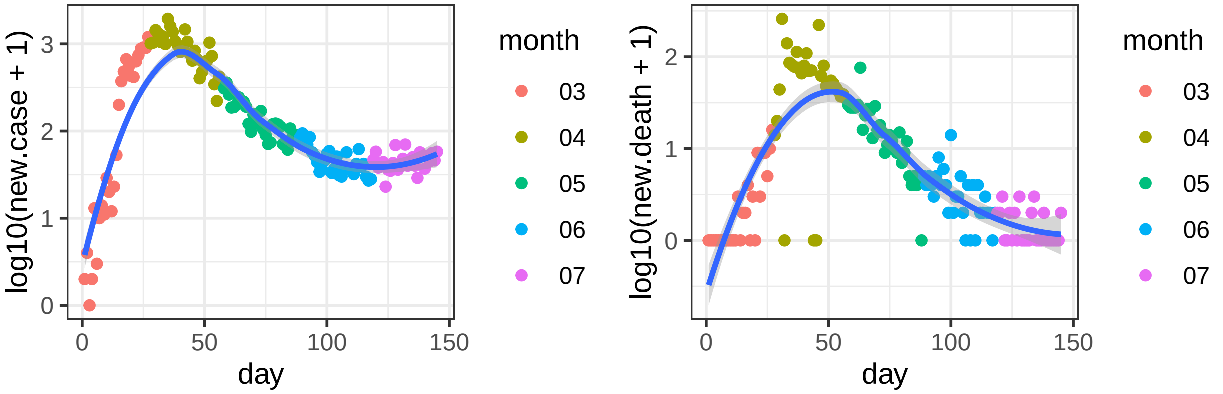
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Wayne_Michigan



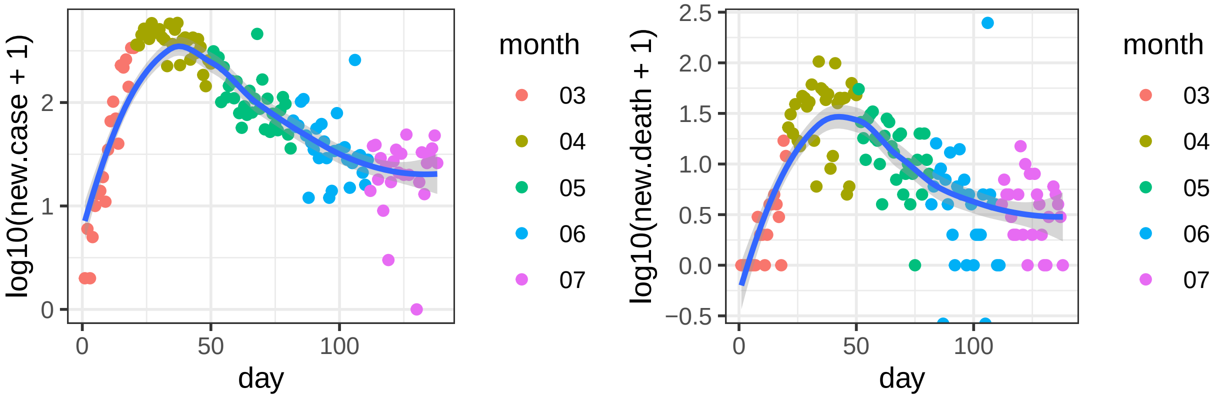
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Nassau_New York



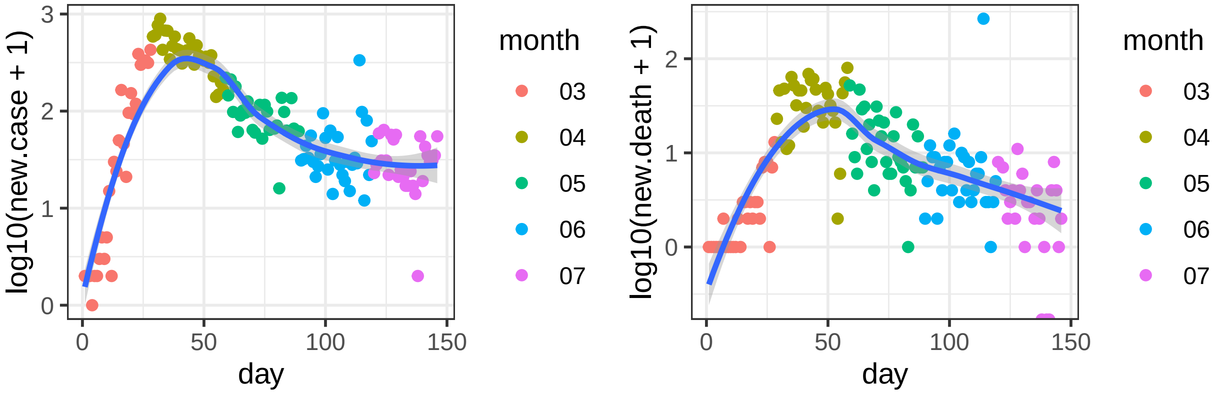
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



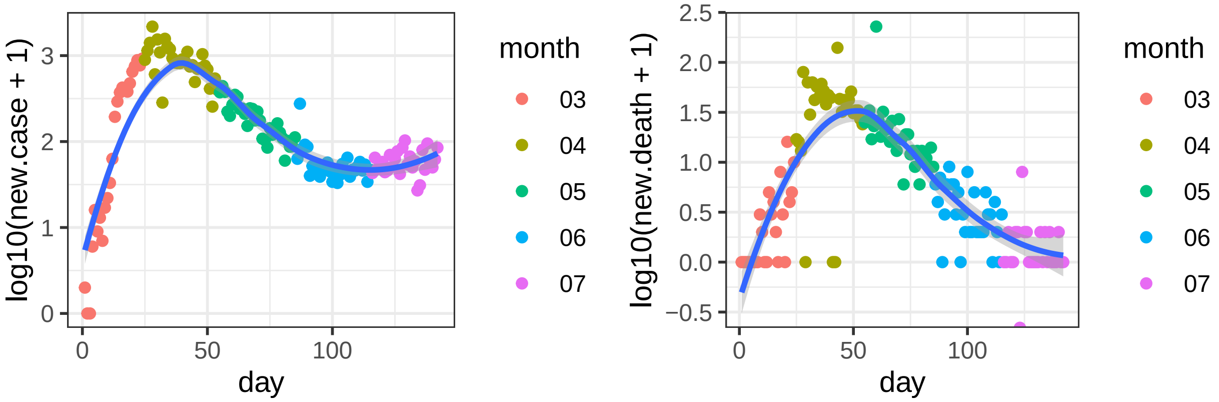
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Bergen_New Jersey



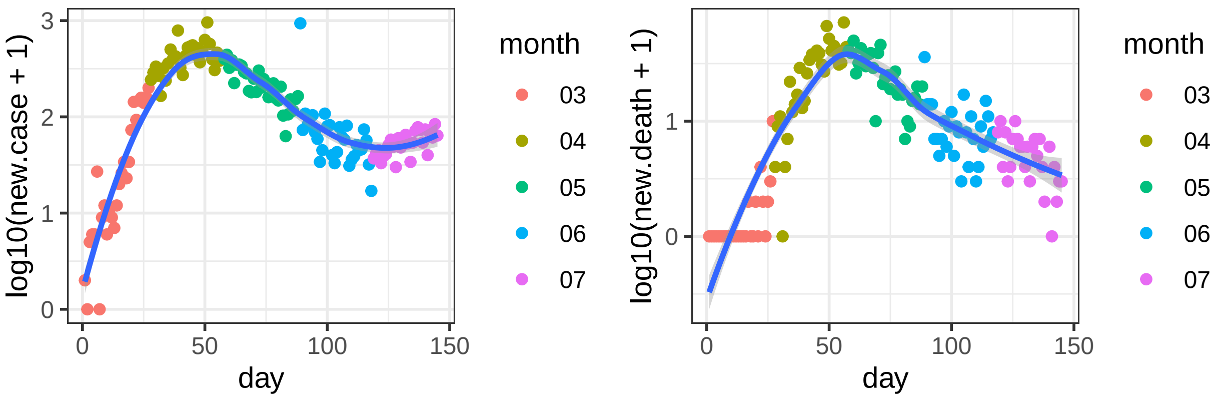
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Suffolk_New York



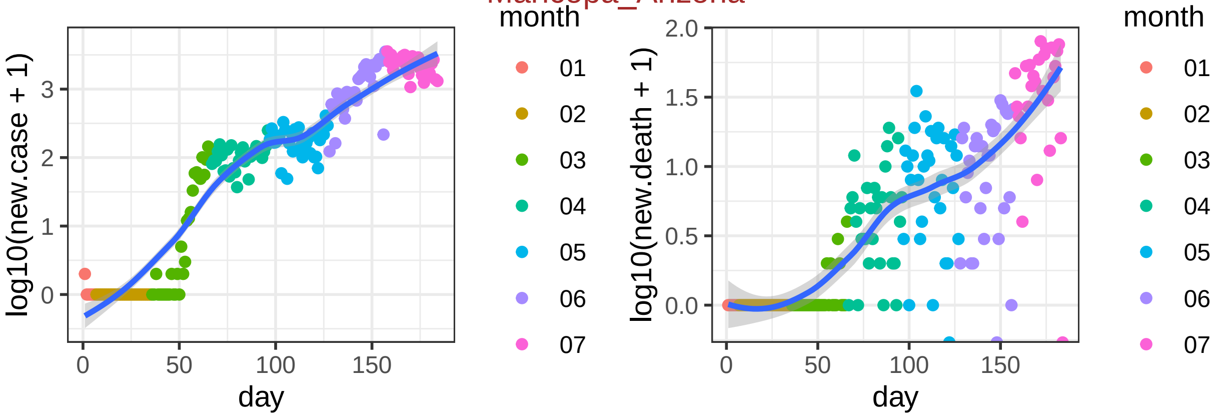
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Middlesex_Massachusetts



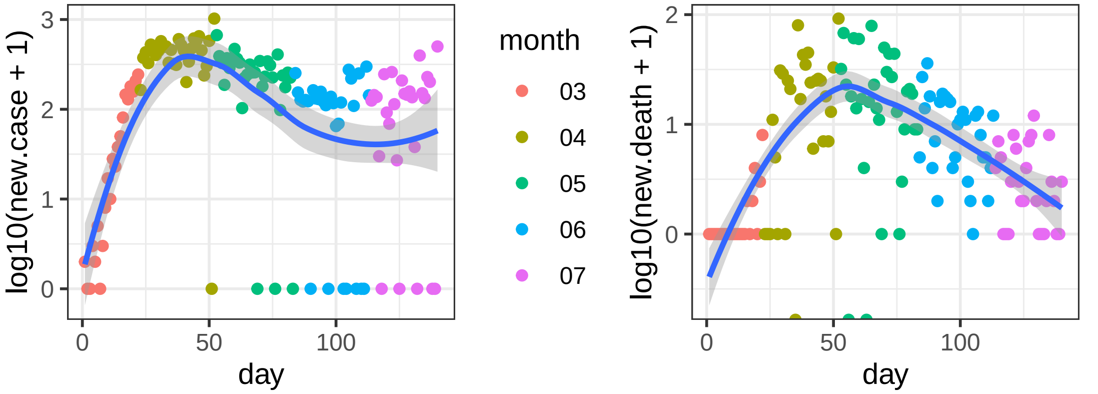
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Maricopa_Arizona



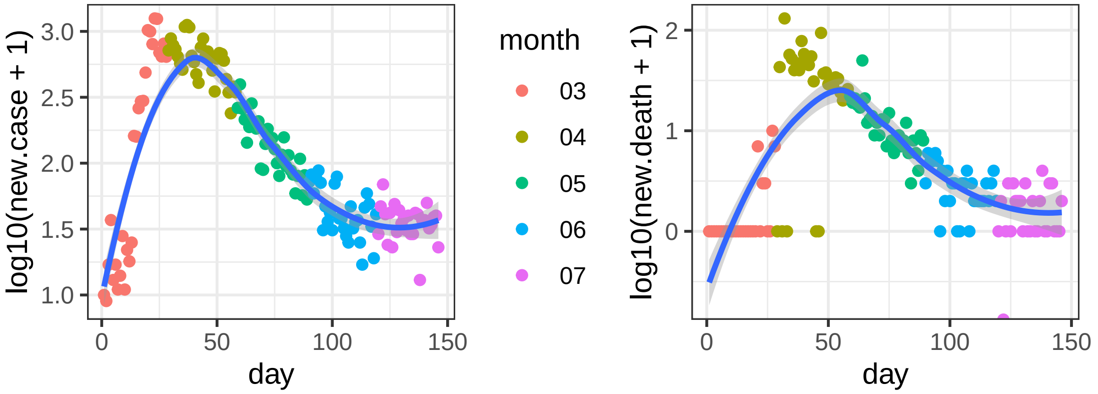
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Philadelphia_Pennsylvania



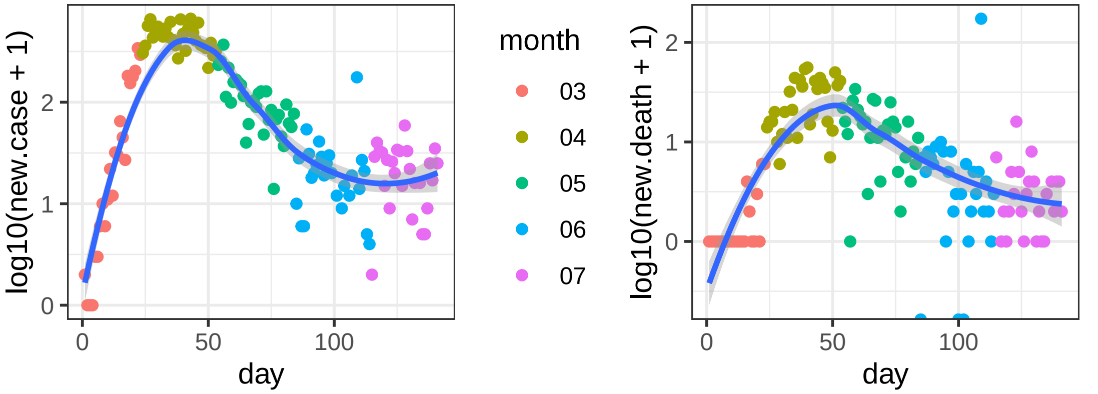
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Westchester_New York



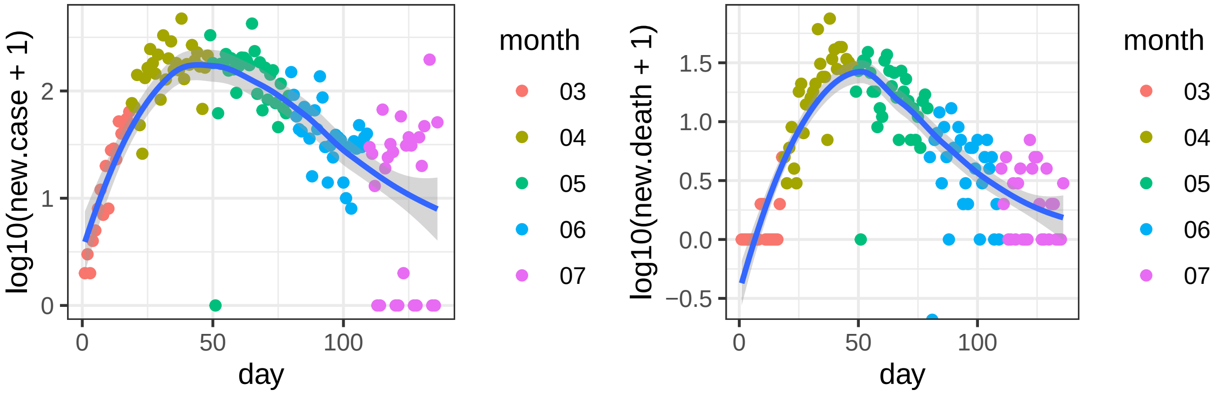
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Hudson_New Jersey



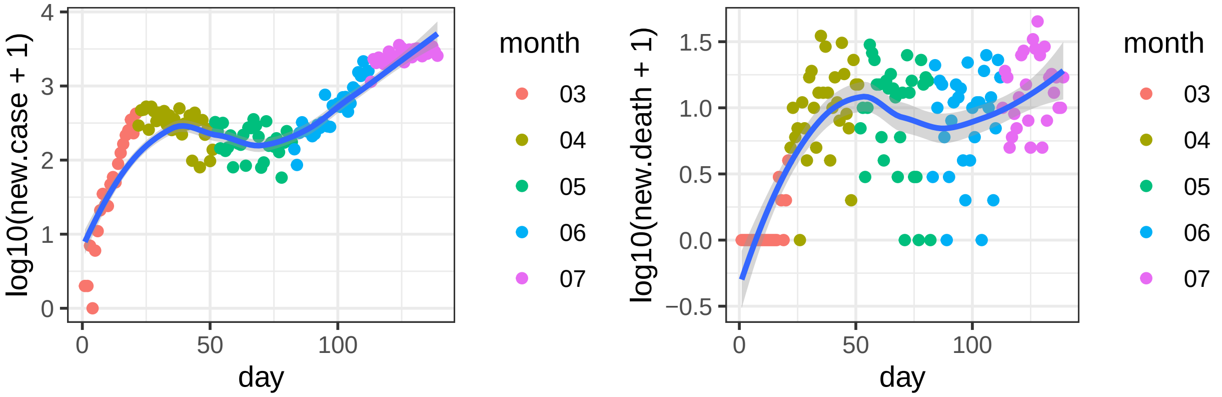
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Hartford_Connecticut



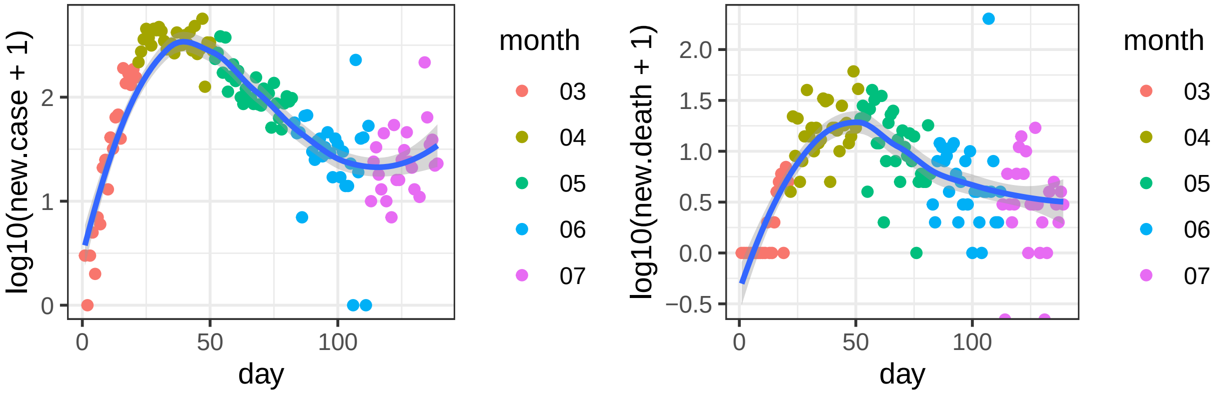
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Miami−Dade_Florida



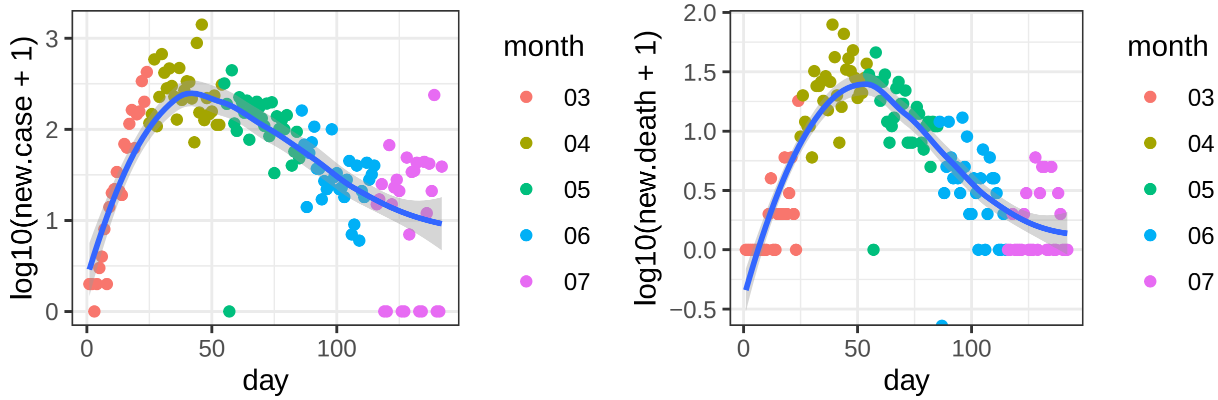
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Middlesex_New Jersey



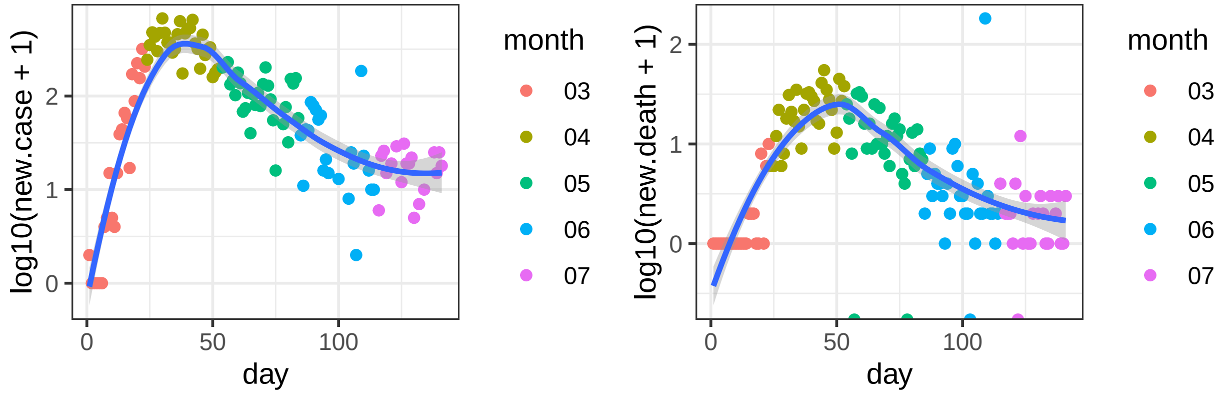
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Fairfield_Connecticut



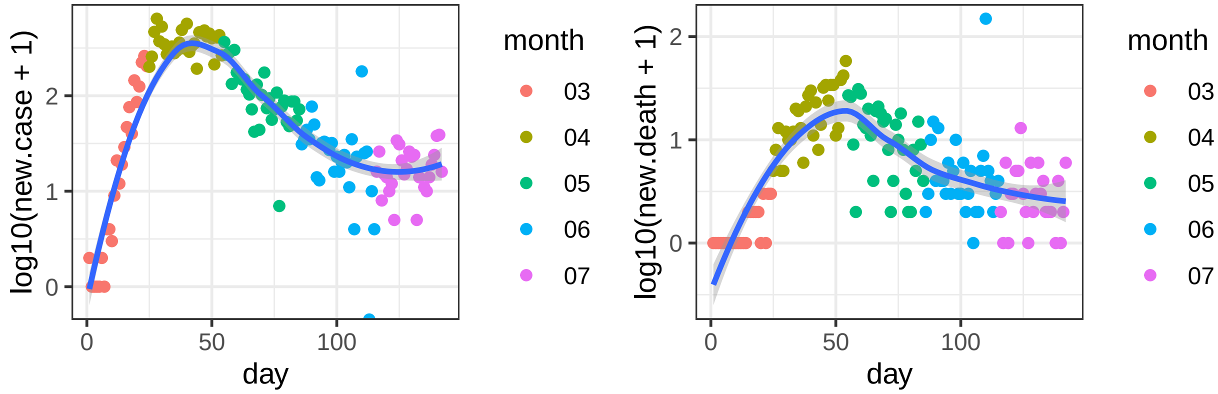
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Union_New Jersey

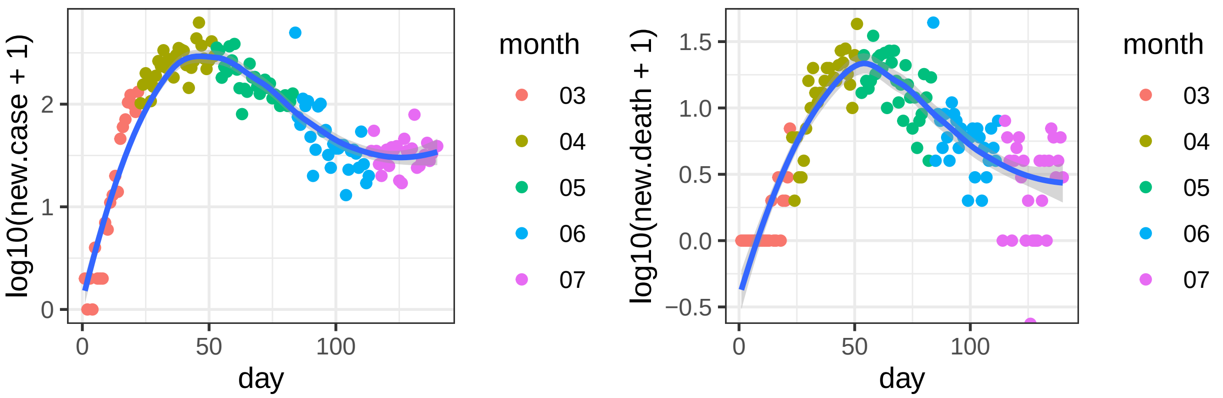


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

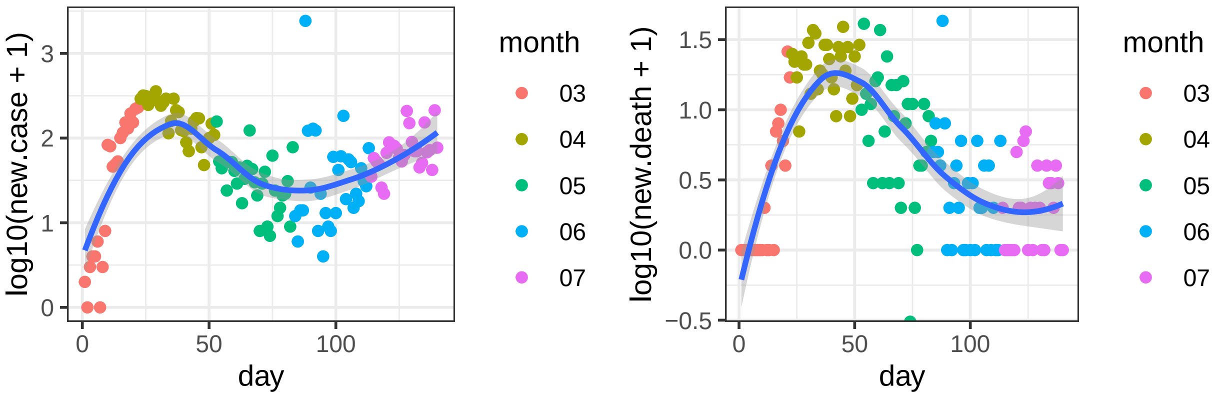
Passaic_New Jersey



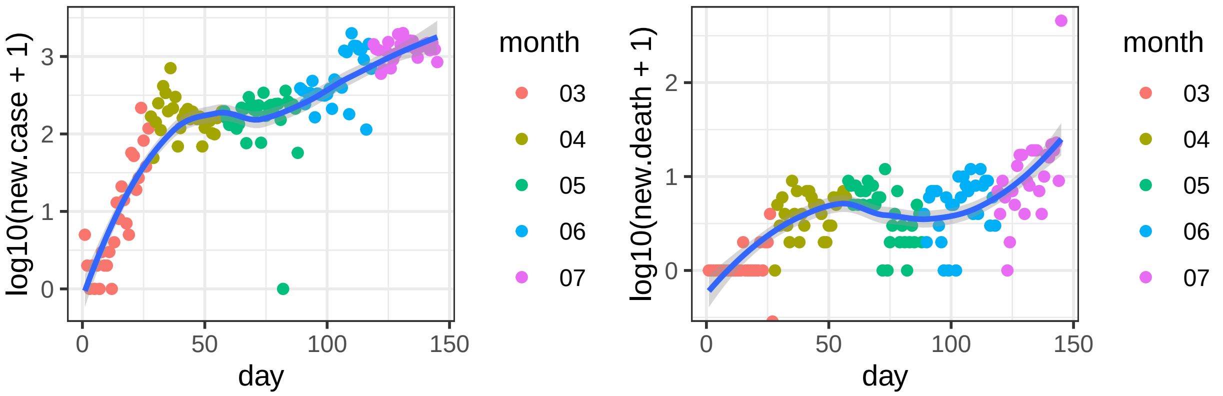
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08


Essex_Massachusetts

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

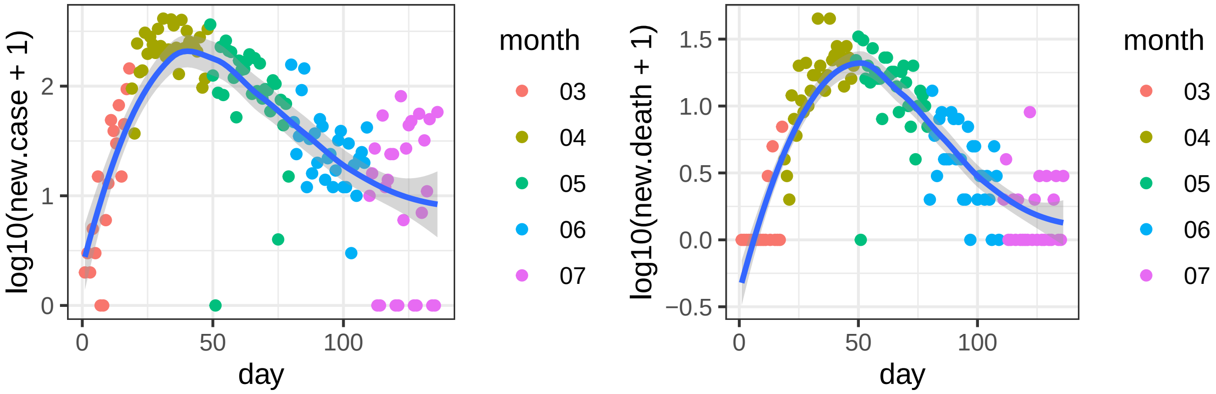

Oakland_Michigan

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


Harris_Texas

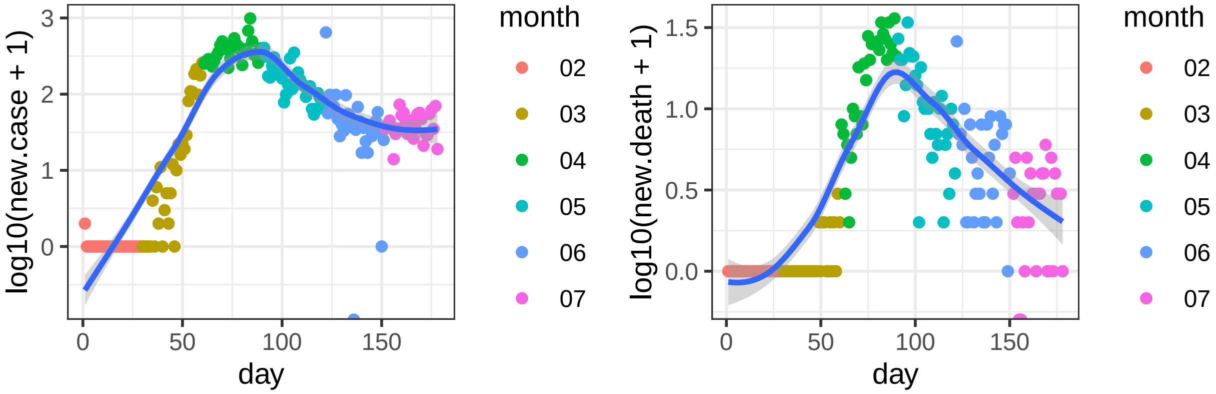
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

New Haven_Connecticut



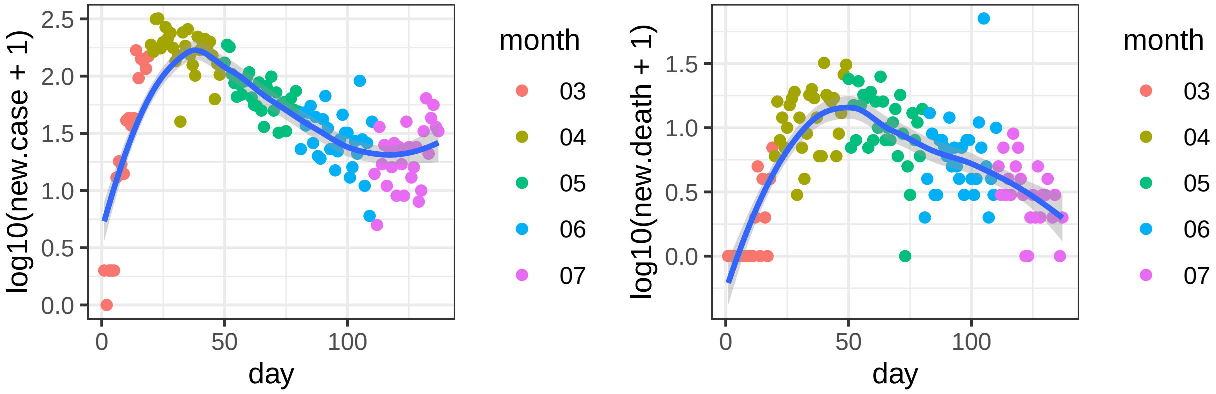
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Suffolk_Massachusetts



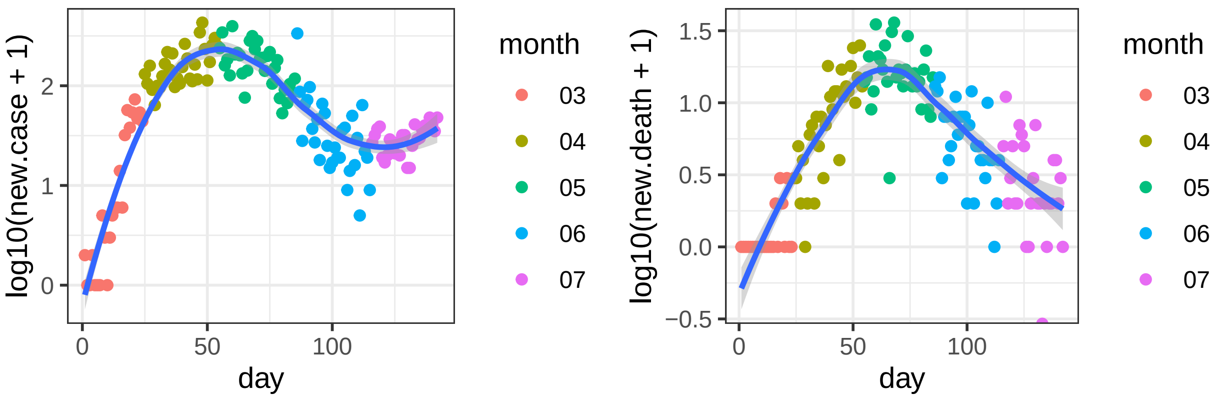
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Ocean_New Jersey



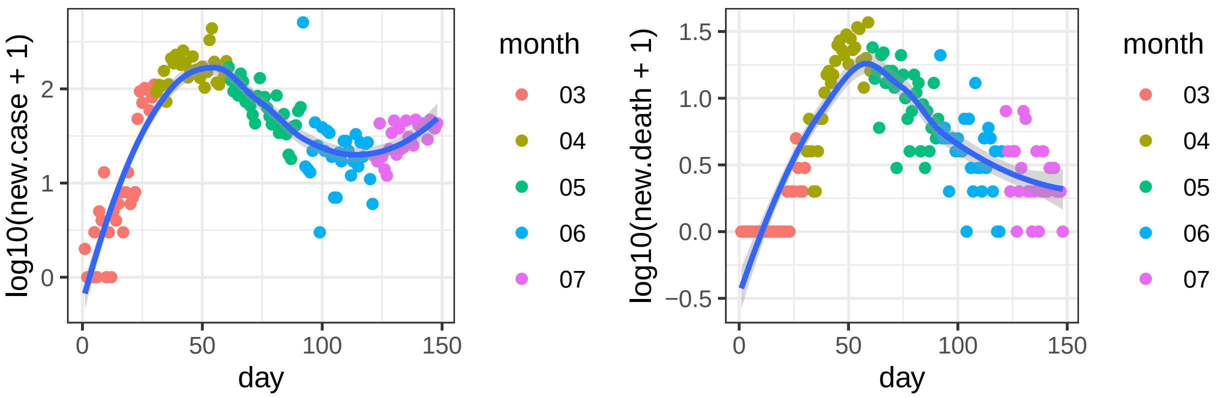
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Worcester_Massachusetts



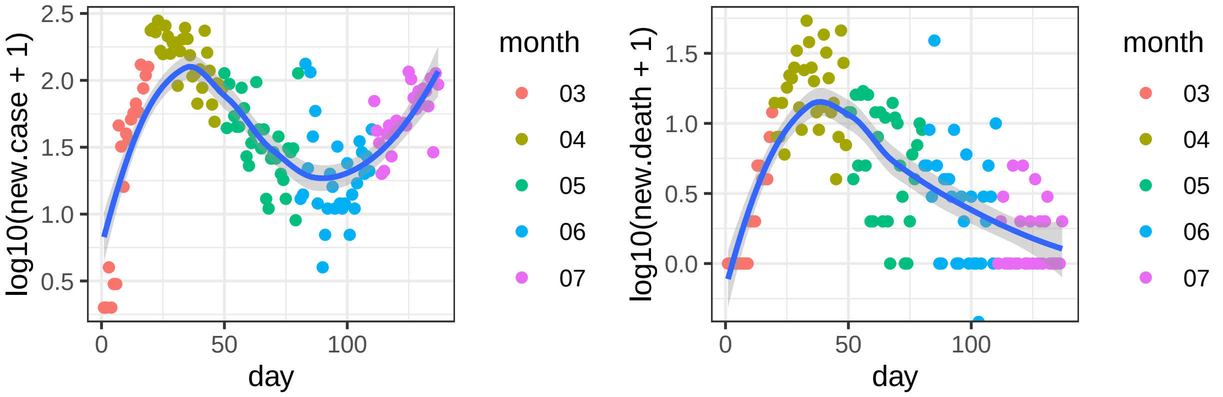
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Norfolk_Massachusetts



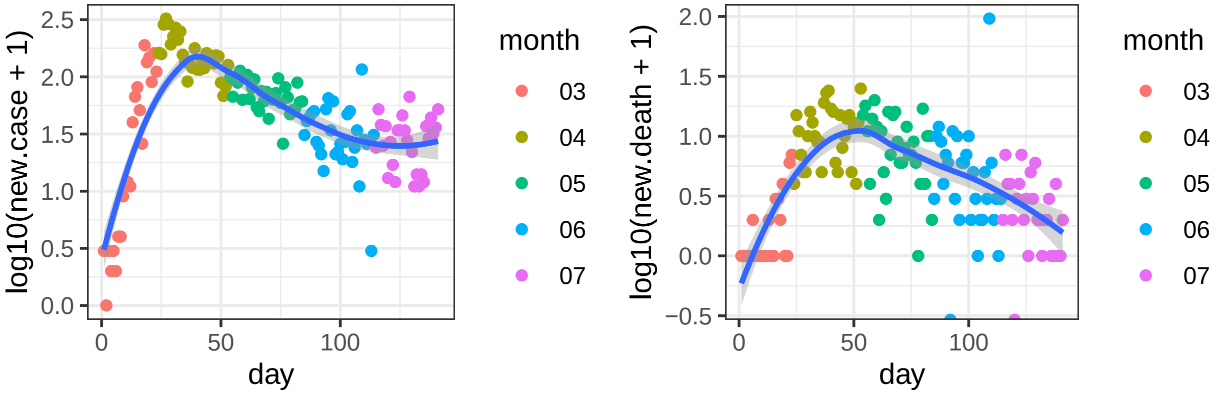
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Macomb_Michigan



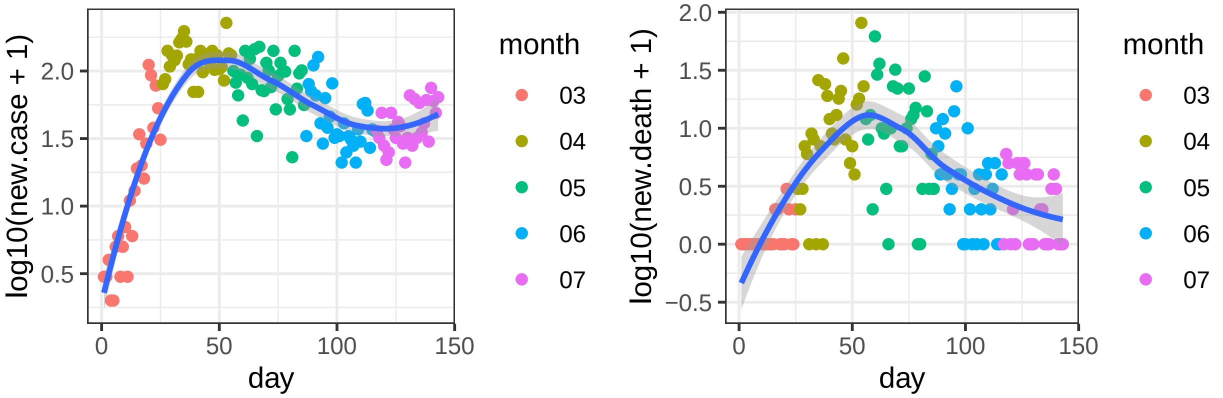
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Monmouth_New Jersey



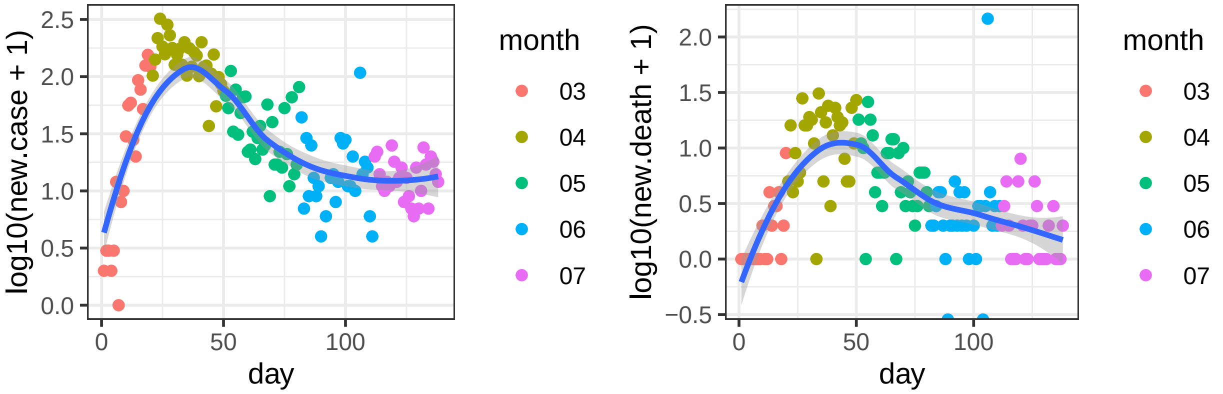
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Montgomery_Pennsylvania



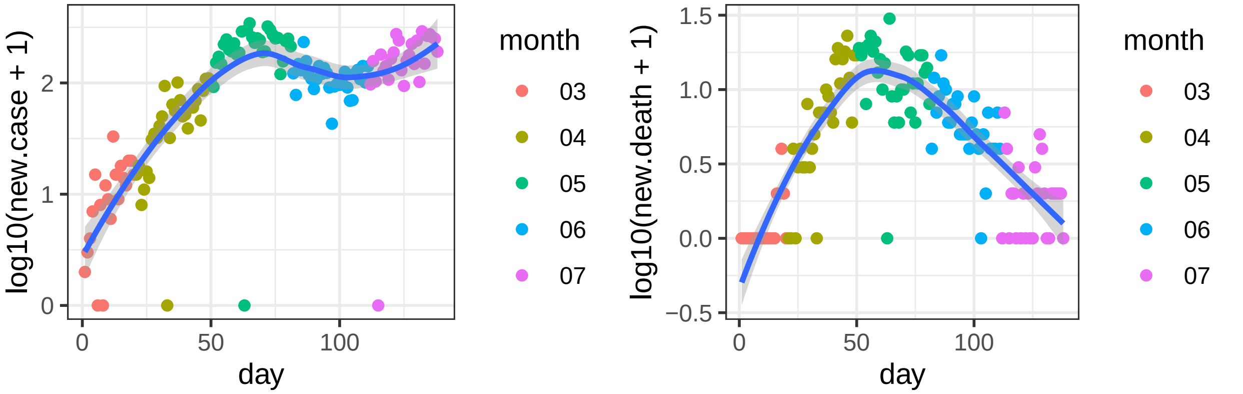
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Morris_New Jersey



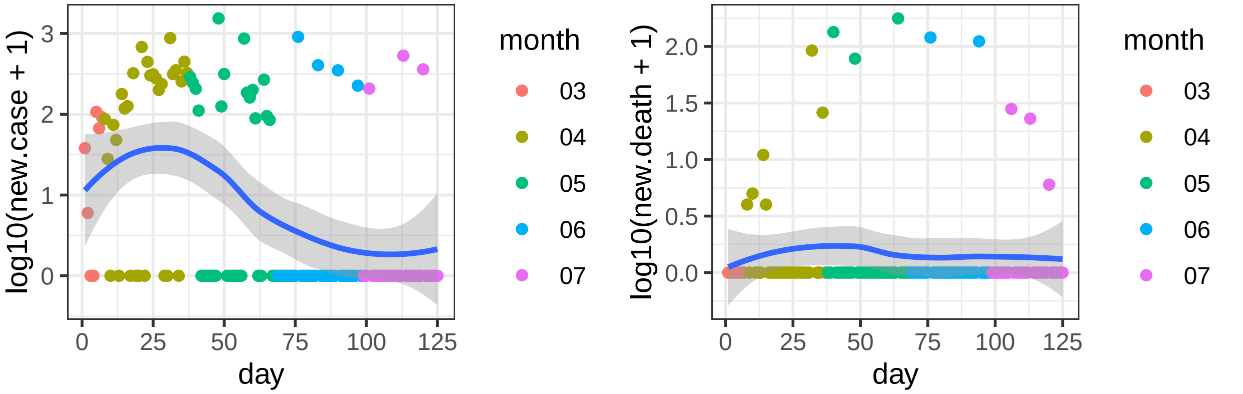
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Hennepin_Minnesota



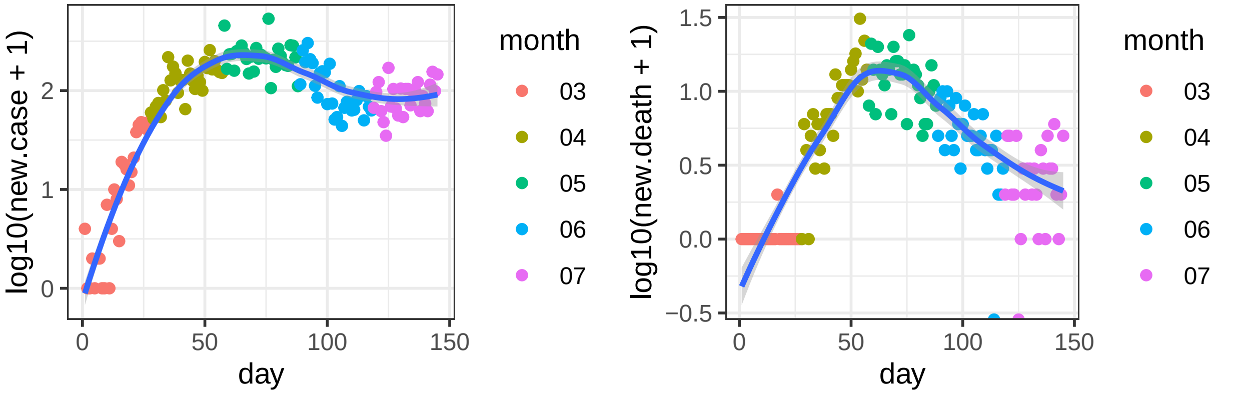
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Providence_Rhode Island



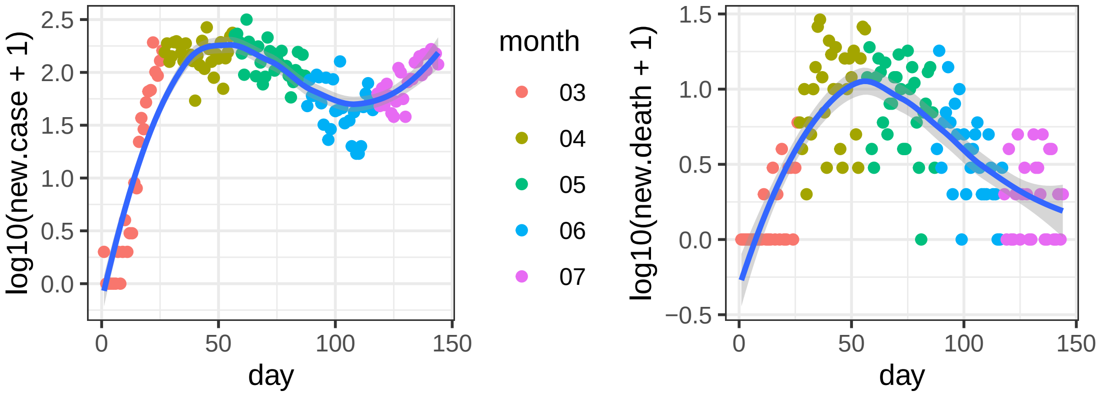
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Montgomery_Maryland



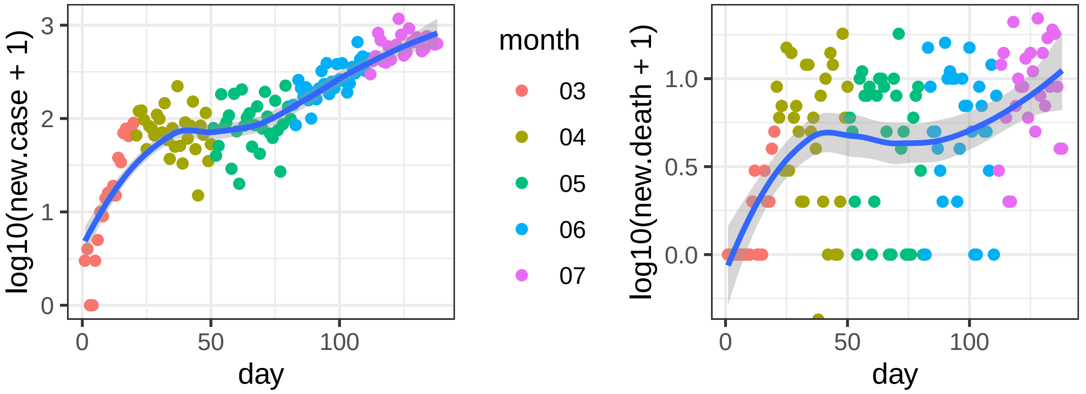
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Marion_Indiana



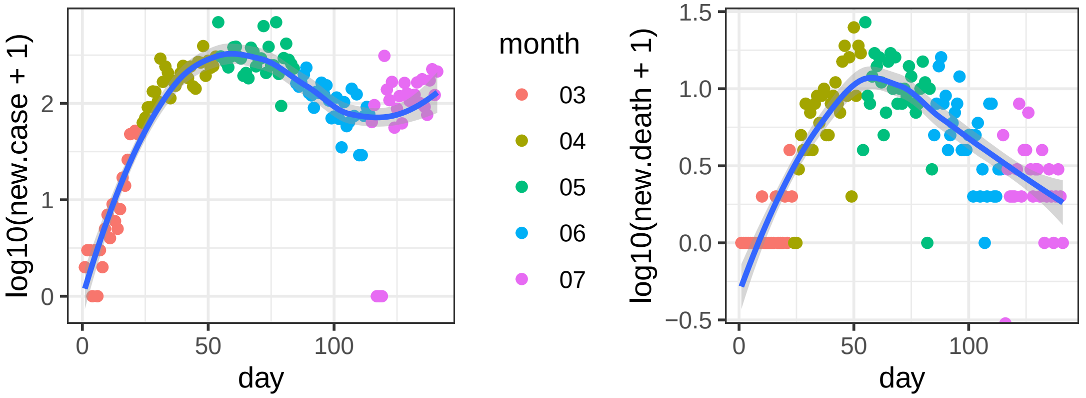
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Palm Beach_Florida



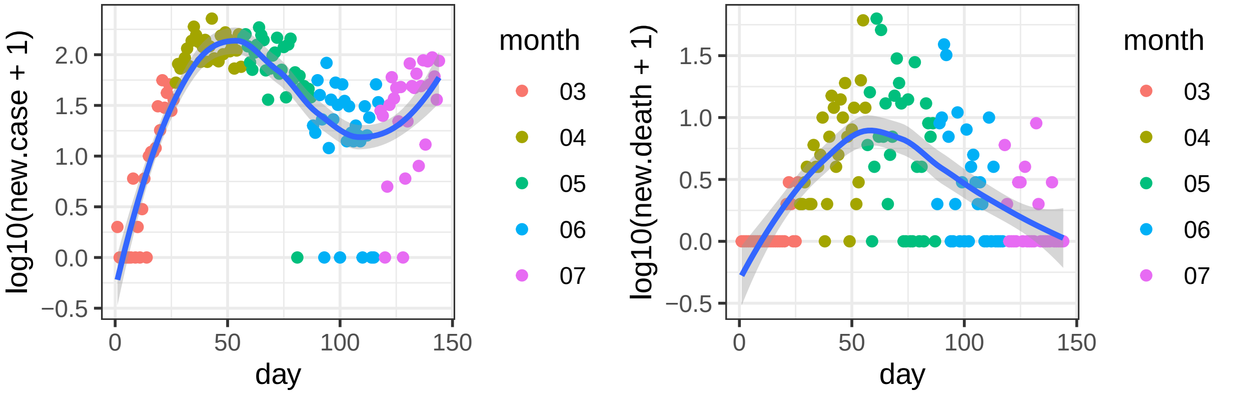
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Prince George's_Maryland



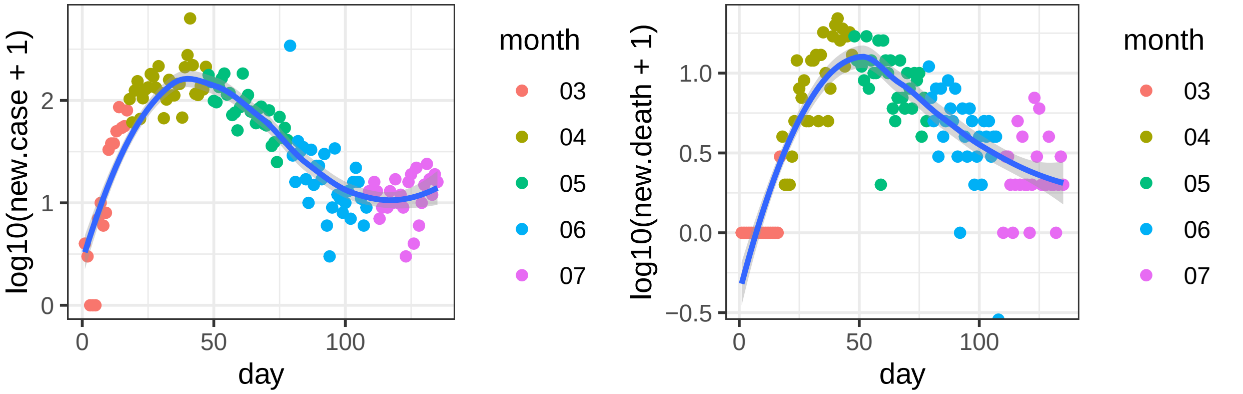
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Delaware_Pennsylvania



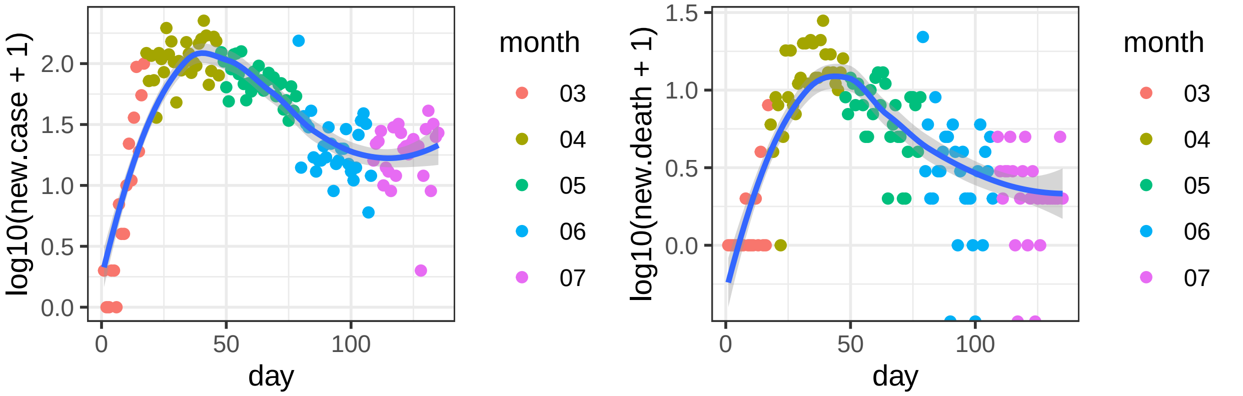
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Plymouth_Massachusetts



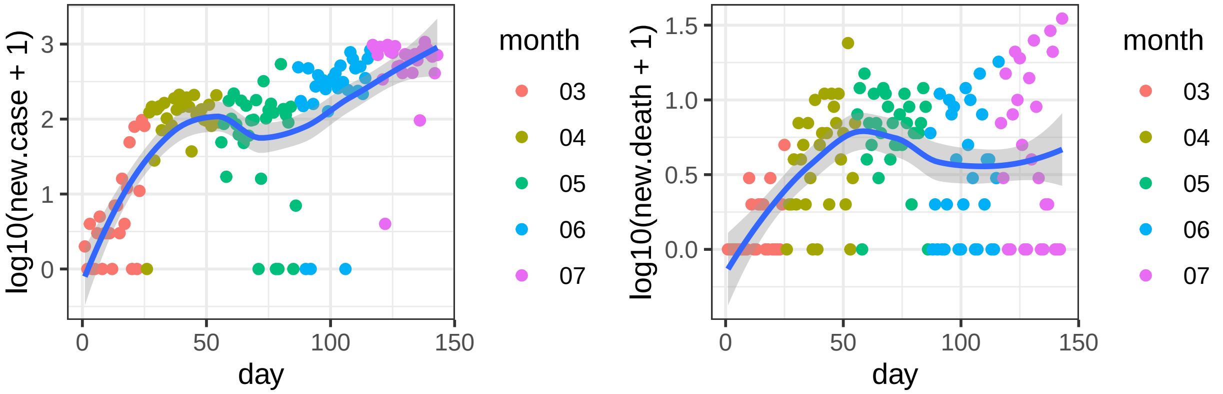
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Hampden_Massachusetts



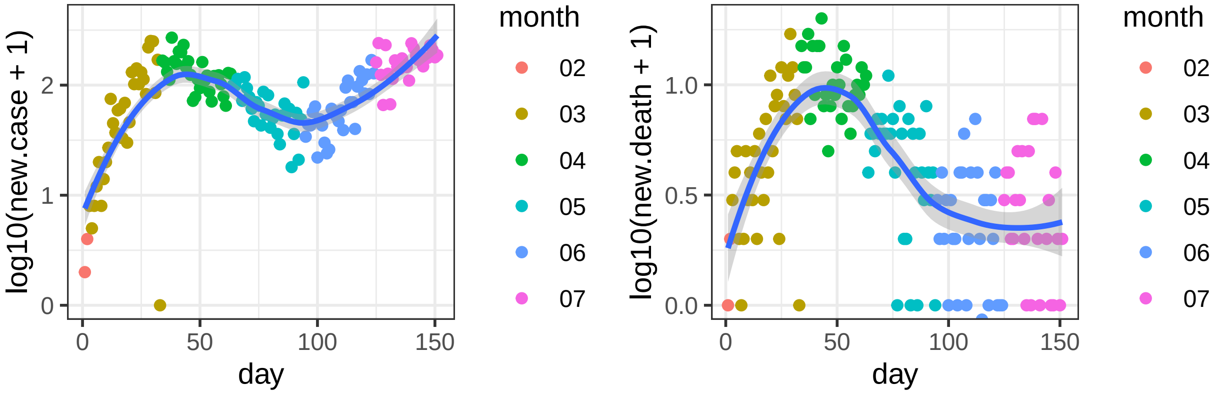
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Riverside_California



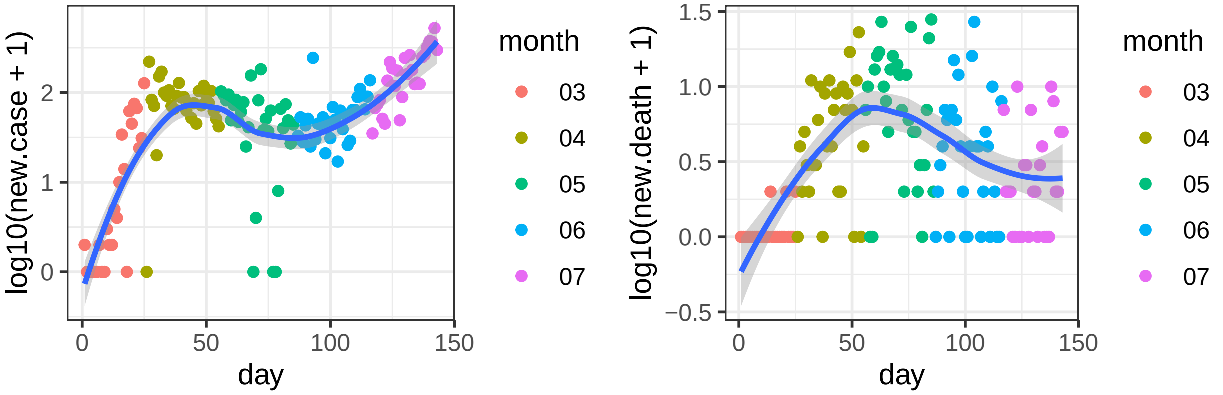
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

King_Washington



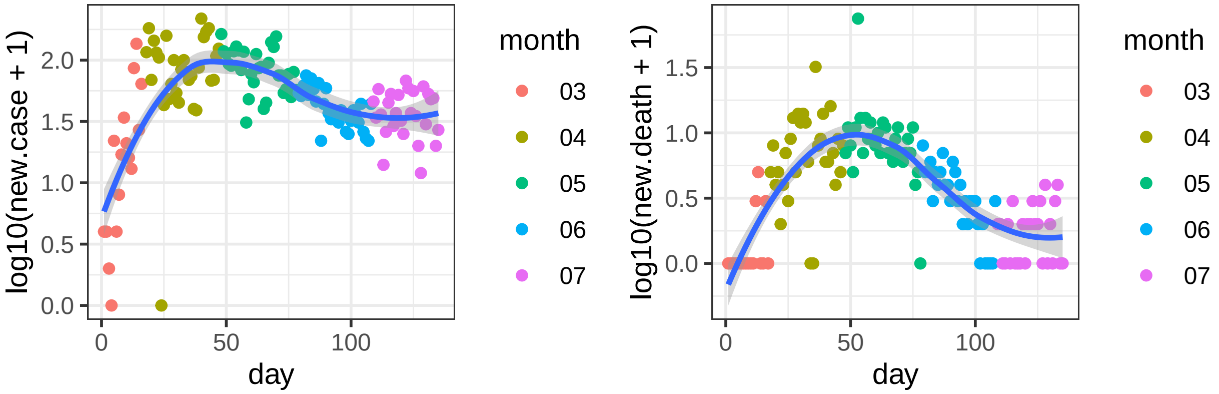
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

St. Louis_Missouri



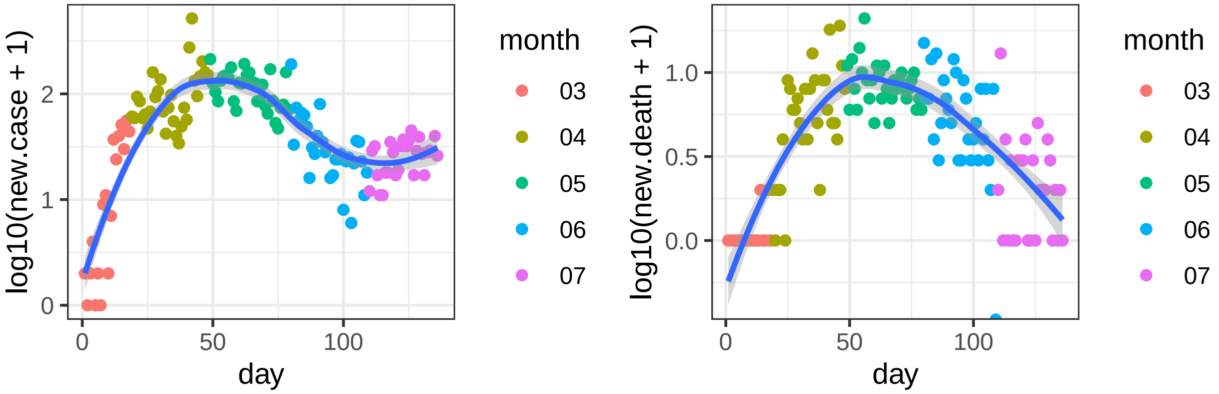
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Erie_New York



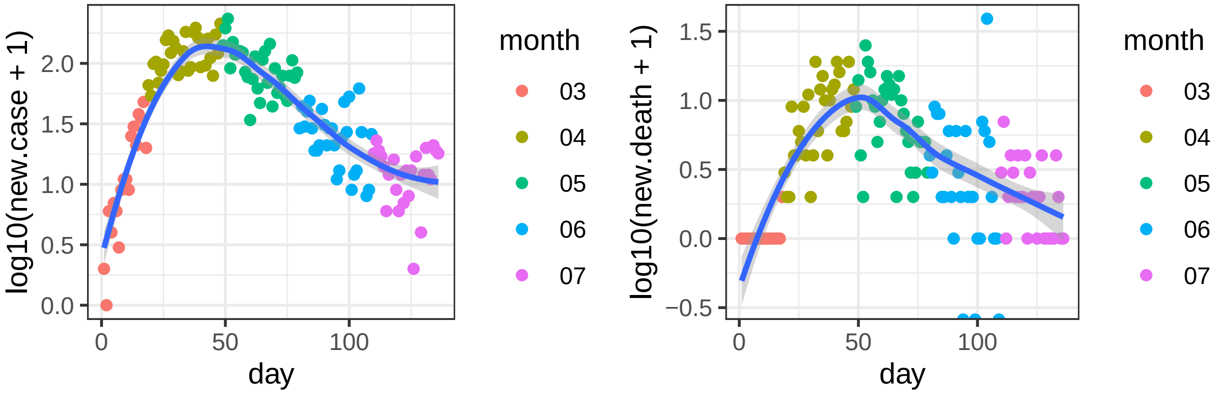
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Bristol_Massachusetts



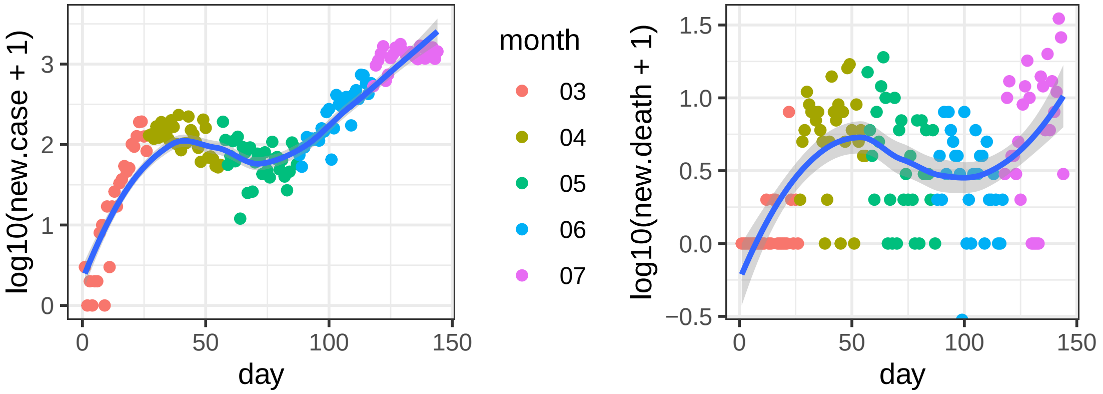
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Mercer_New Jersey



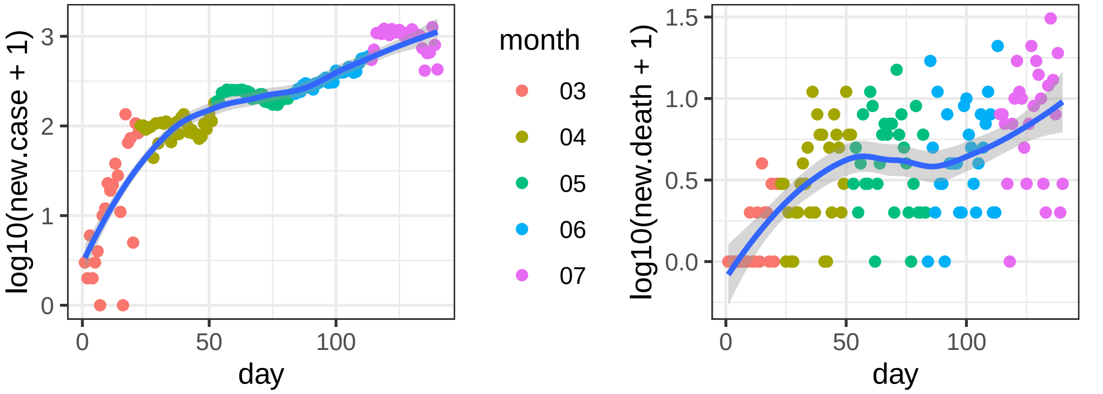
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Broward_Florida



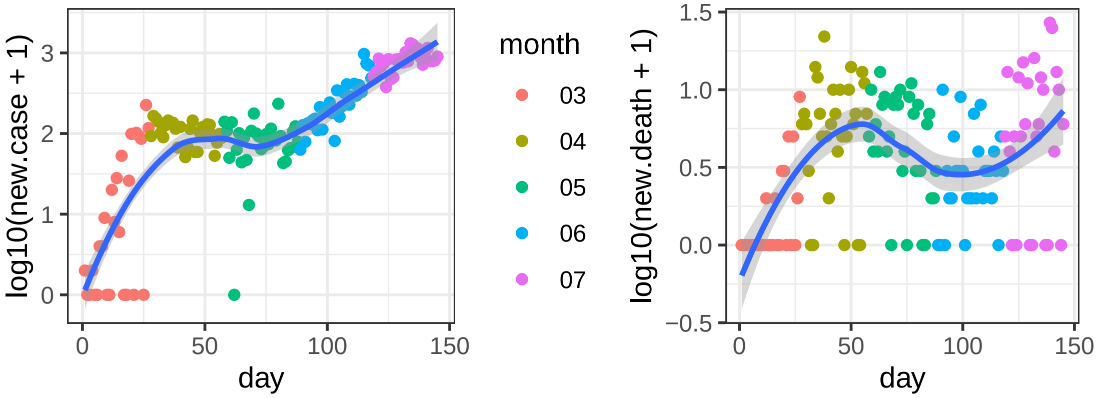
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Dallas_Texas



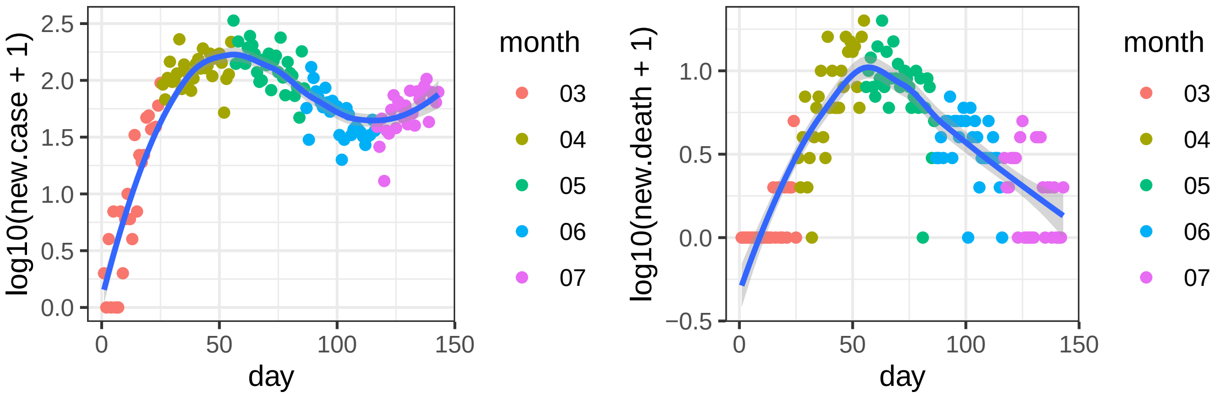
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Clark_Nevada



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

District of Columbia_District of Columbia

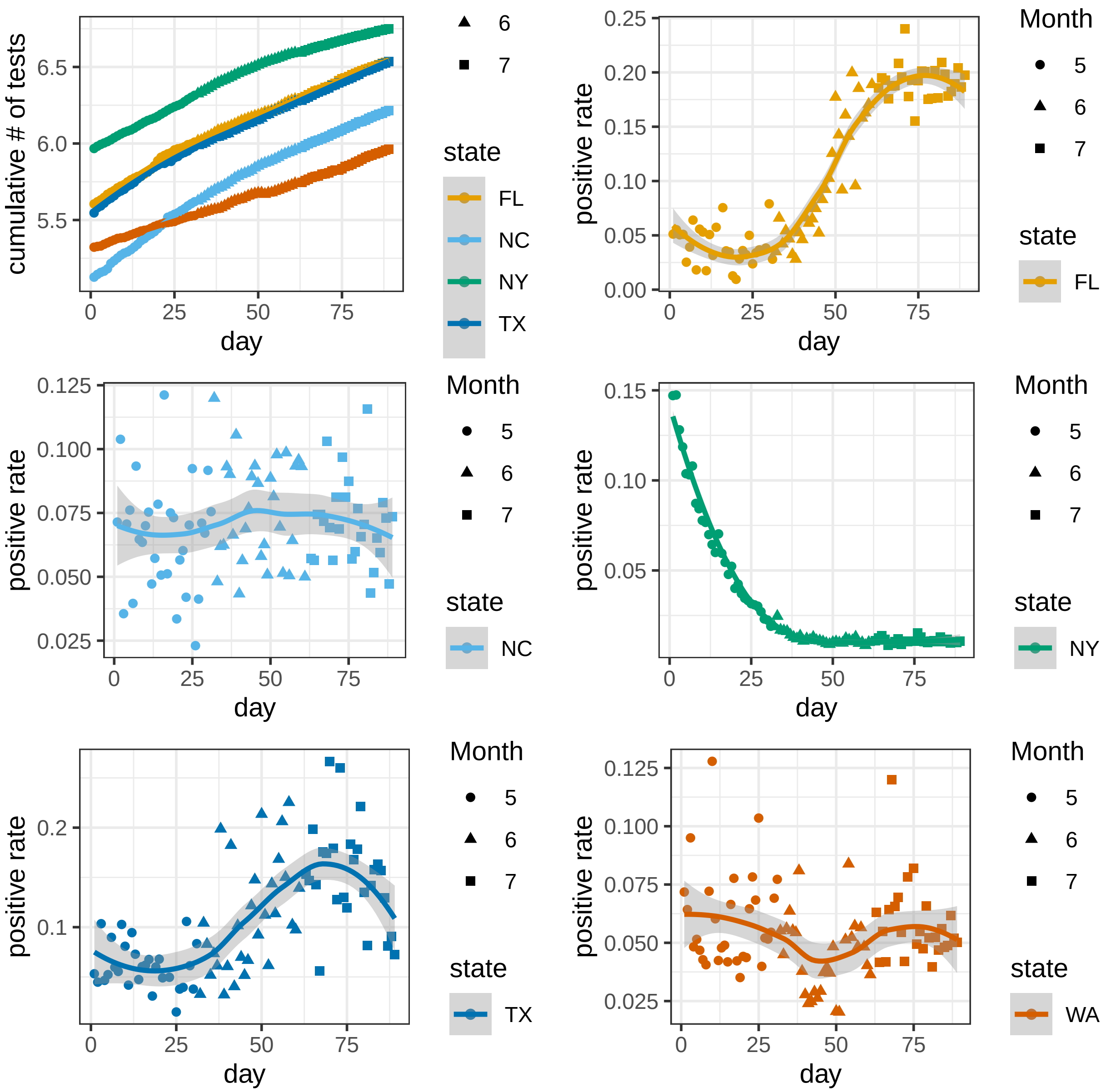


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

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 0727: 0.20(FL) 0.07(NC) 0.01(NY) 0.07(TX) 0.05(WA)

Session information

```
sessionInfo()
```

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

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