Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-06-22

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

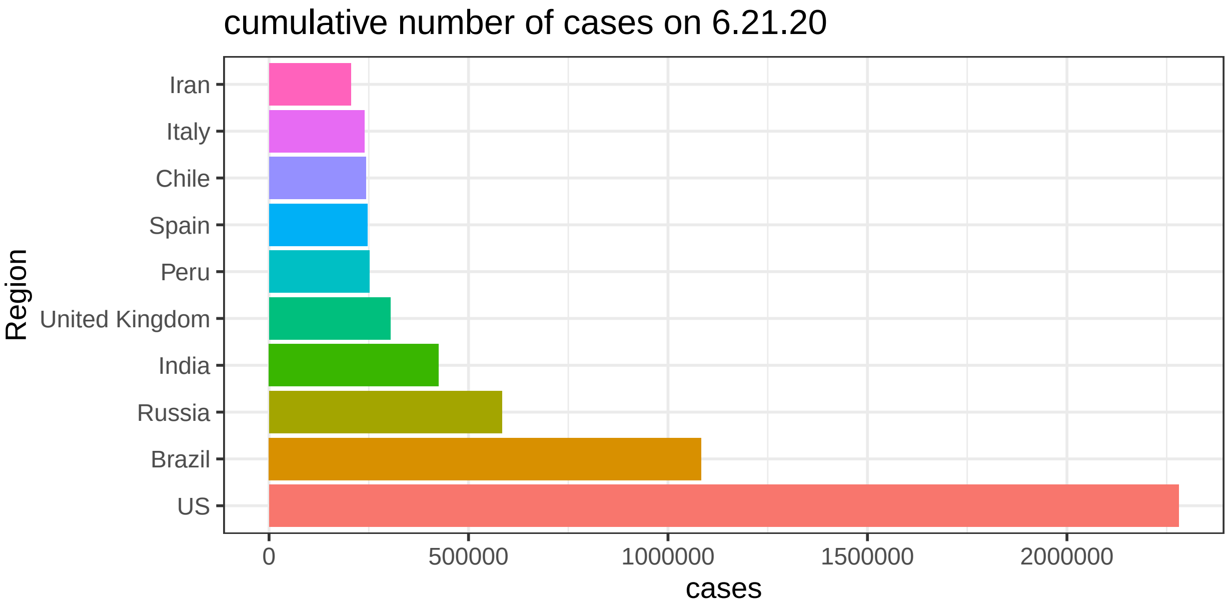
JHU

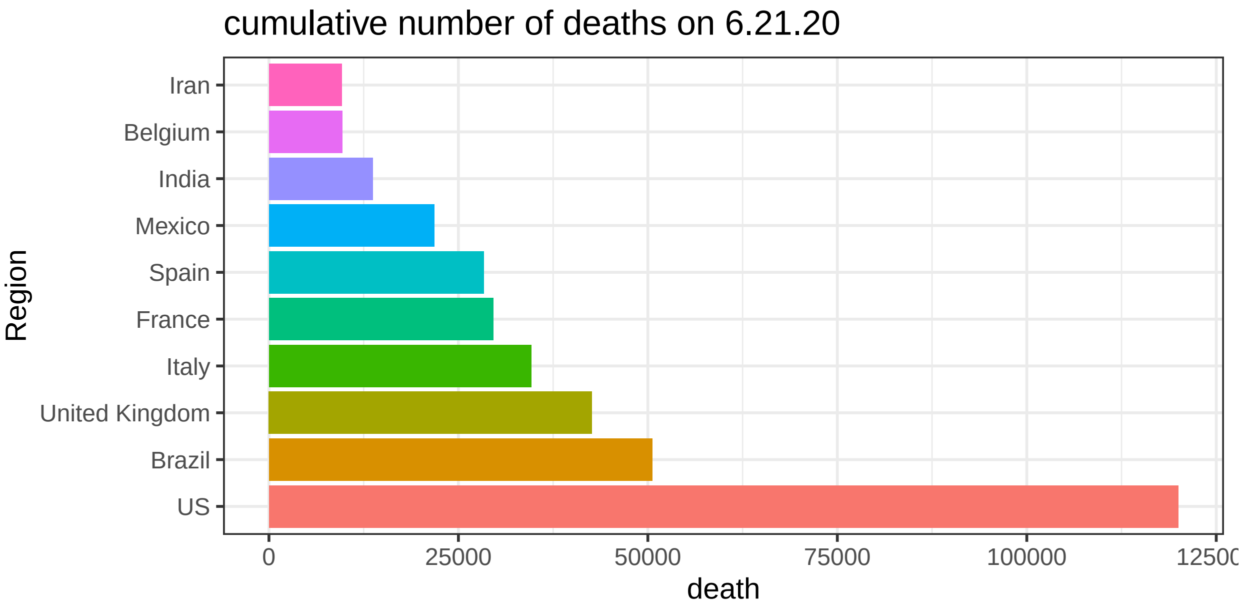
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

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.

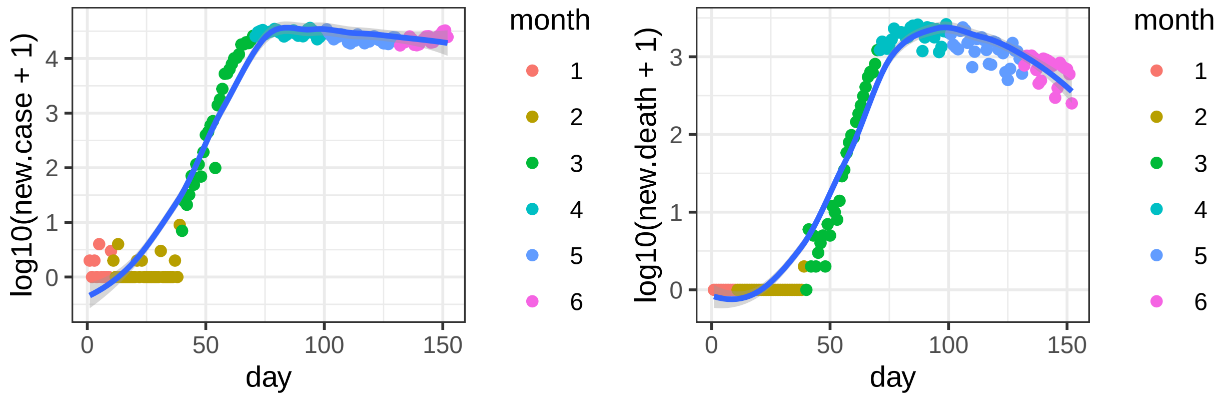
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





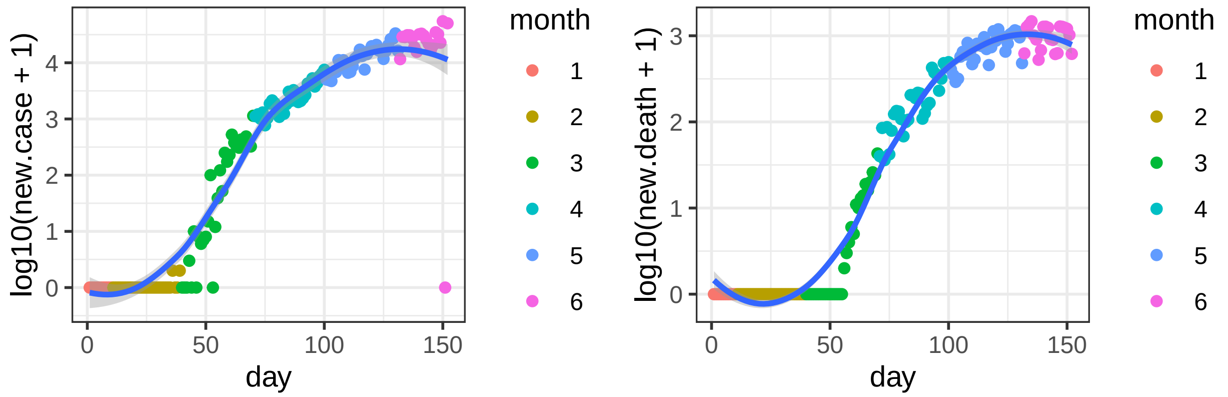
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



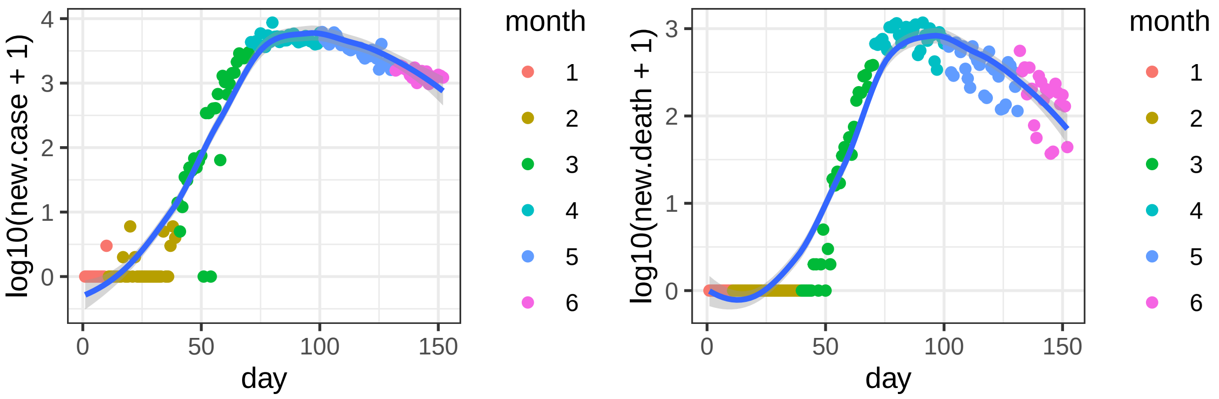
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Brazil



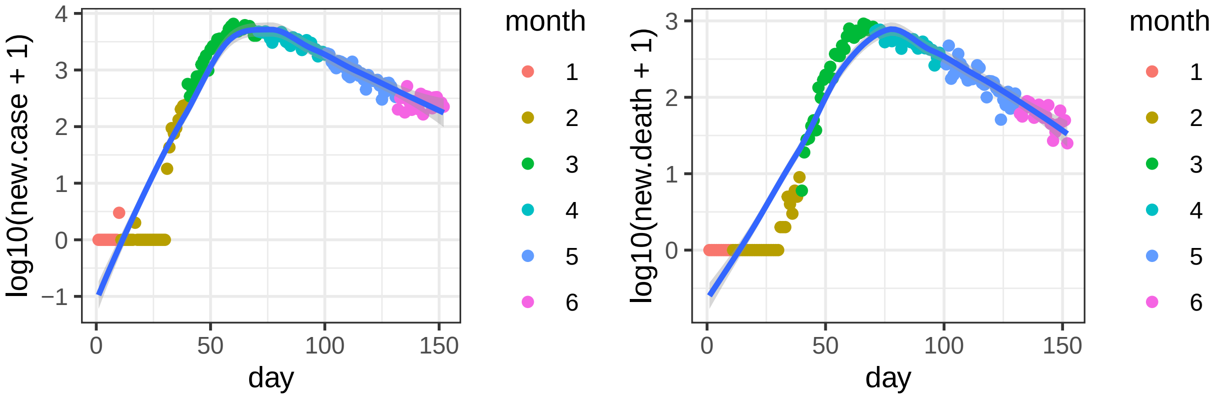
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



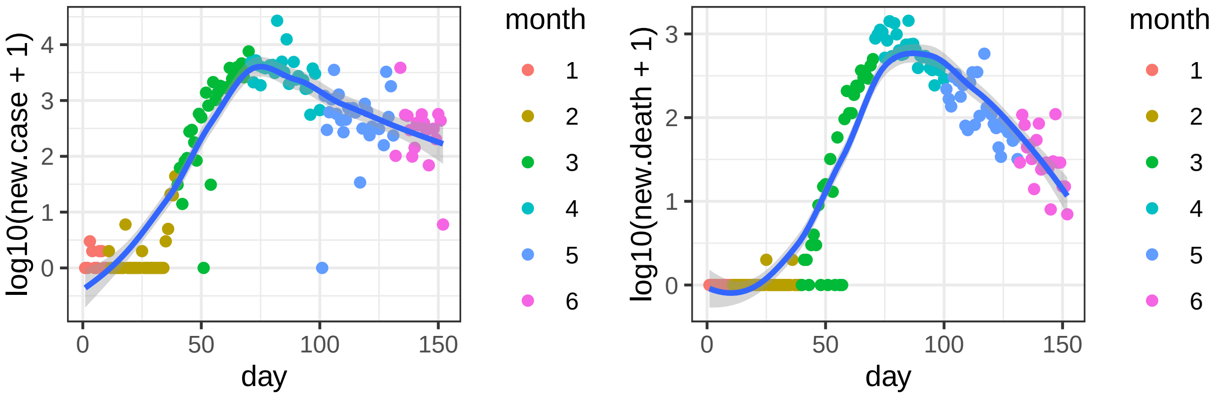
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



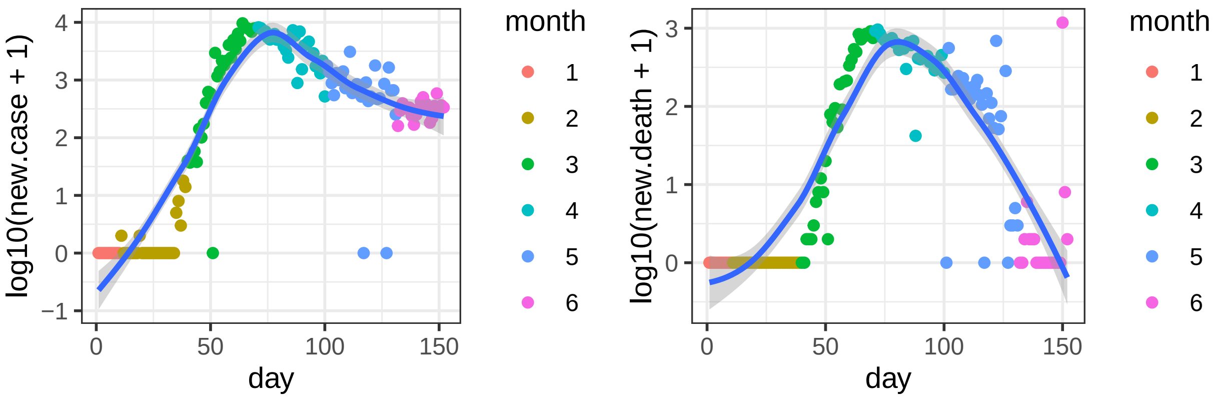
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



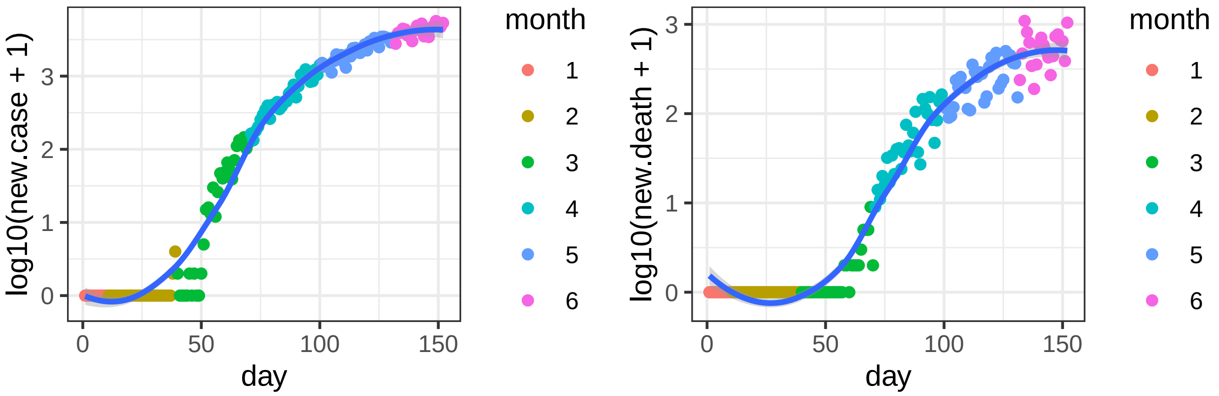
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain



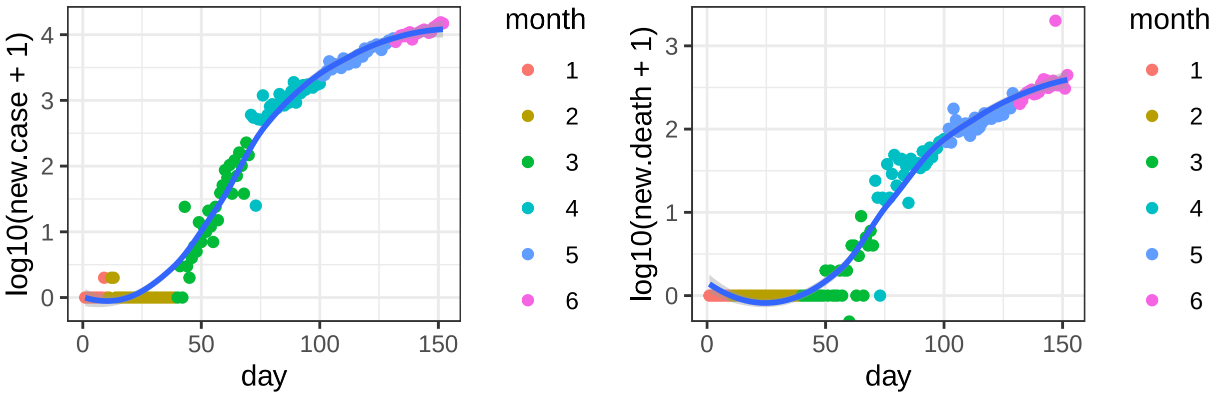
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Mexico

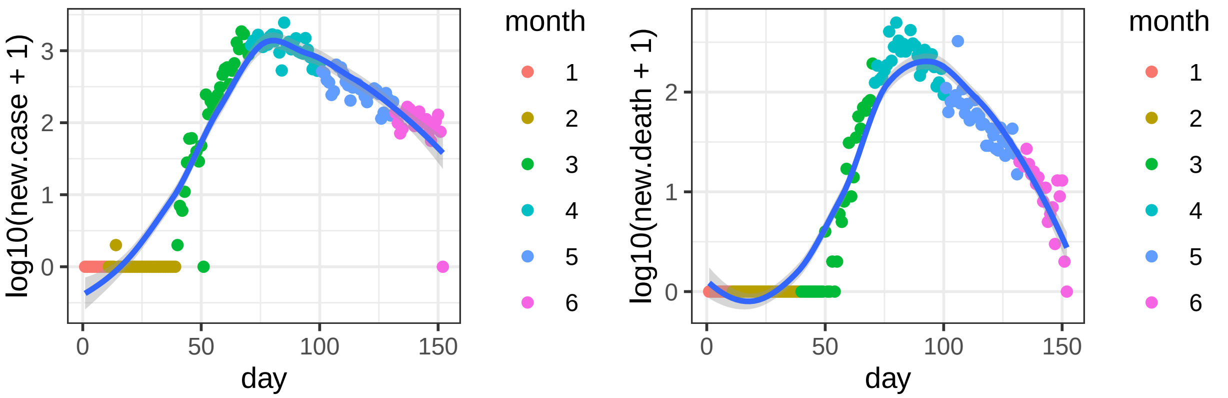


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

India

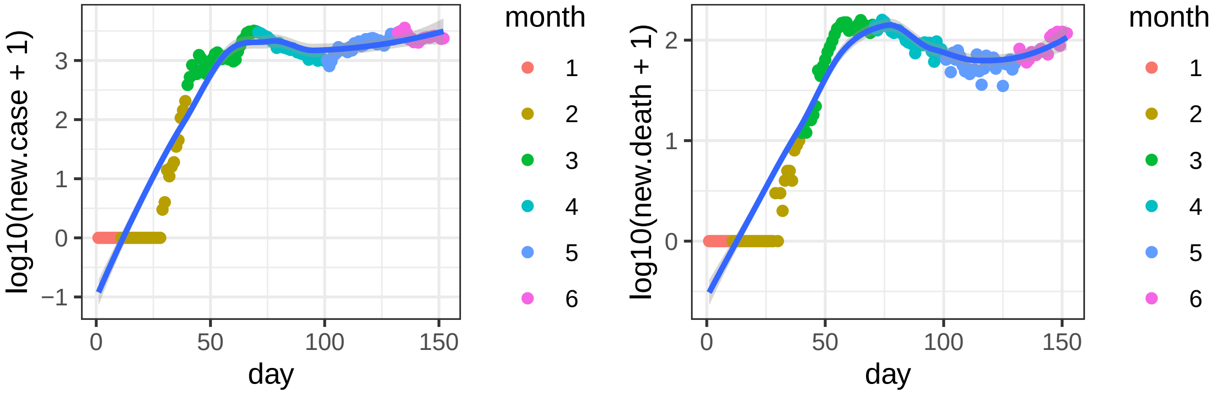


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



Belgium

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

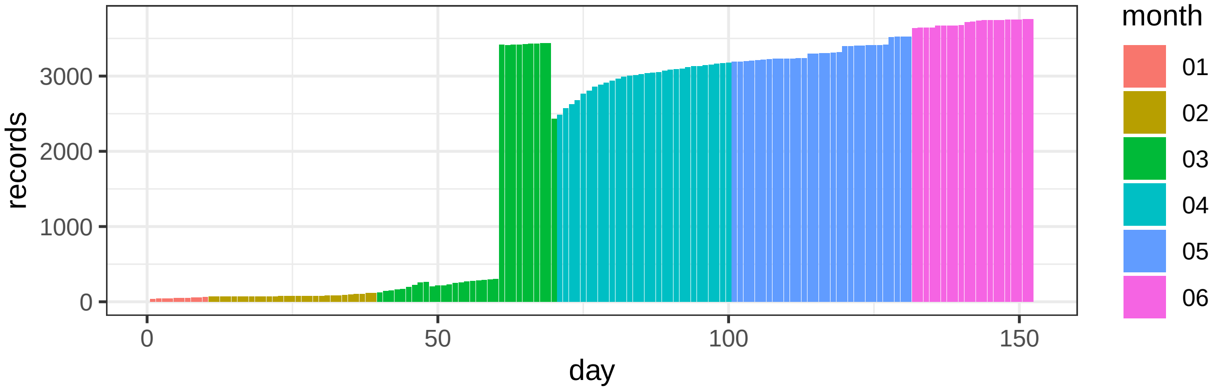


Iran

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.



number of records in Hopkins daily reports

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] "2020-06-21"
```

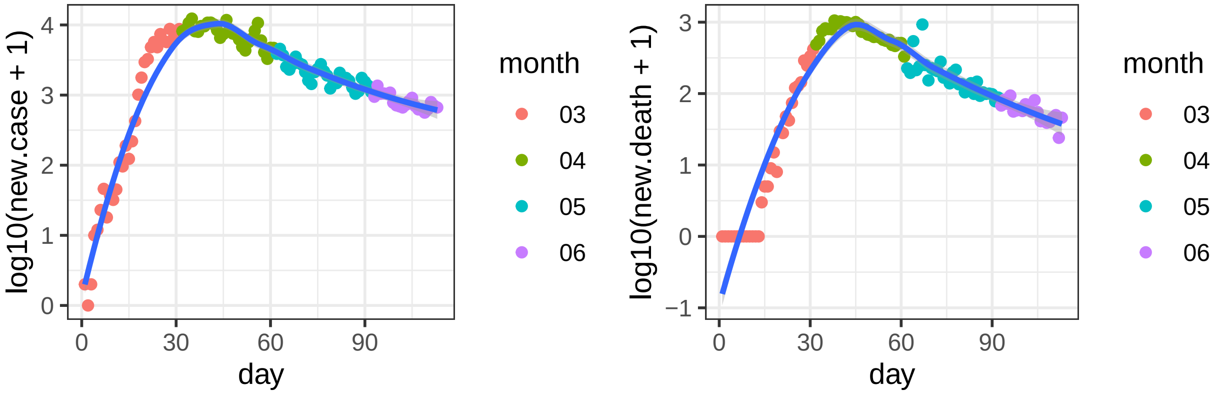
state level data

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

```
##            date                state fips  cases deaths
## 6098 2020-06-21             New York   36 392702  30884
## 6096 2020-06-21           New Jersey   34 169142  12870
## 6087 2020-06-21        Massachusetts   25 107061   7857
## 6079 2020-06-21             Illinois   17 138154   6865
## 6105 2020-06-21         Pennsylvania   42  86024   6472
## 6088 2020-06-21             Michigan   26  67873   6094
## 6069 2020-06-21           California    6 178807   5517
## 6071 2020-06-21          Connecticut    9  45755   4260
## 6074 2020-06-21              Florida   12  97283   3160
## 6084 2020-06-21            Louisiana   22  49890   3105
## 6086 2020-06-21             Maryland   24  64903   3066
## 6102 2020-06-21                 Ohio   39  44808   2700
## 6075 2020-06-21              Georgia   13  61493   2603
## 6080 2020-06-21              Indiana   18  43496   2540
## 6111 2020-06-21                Texas   48 114886   2195
## 6070 2020-06-21             Colorado    8  30524   1647
## 6115 2020-06-21             Virginia   51  57994   1611
## 6089 2020-06-21            Minnesota   27  32952   1412
## 6067 2020-06-21              Arizona    4  52666   1350
## 6116 2020-06-21           Washington   53  29797   1271
## 6099 2020-06-21       North Carolina   37  52854   1245
## 6091 2020-06-21             Missouri   29  18552    975
## 6090 2020-06-21          Mississippi   28  21022    943
## 6107 2020-06-21         Rhode Island   44  16337    894
## 6065 2020-06-21              Alabama    1  30021    839
## 6118 2020-06-21            Wisconsin   55  24920    744
## 6081 2020-06-21                 Iowa   19  26020    686
## 6108 2020-06-21       South Carolina   45  24693    653
## 6083 2020-06-21             Kentucky   21  13919    544
## 6073 2020-06-21 District of Columbia   11  10020    533
```

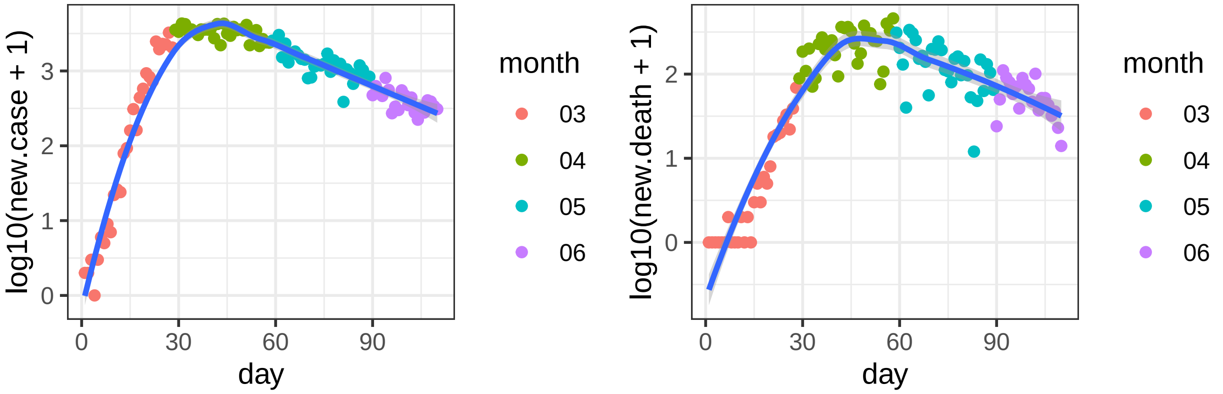
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



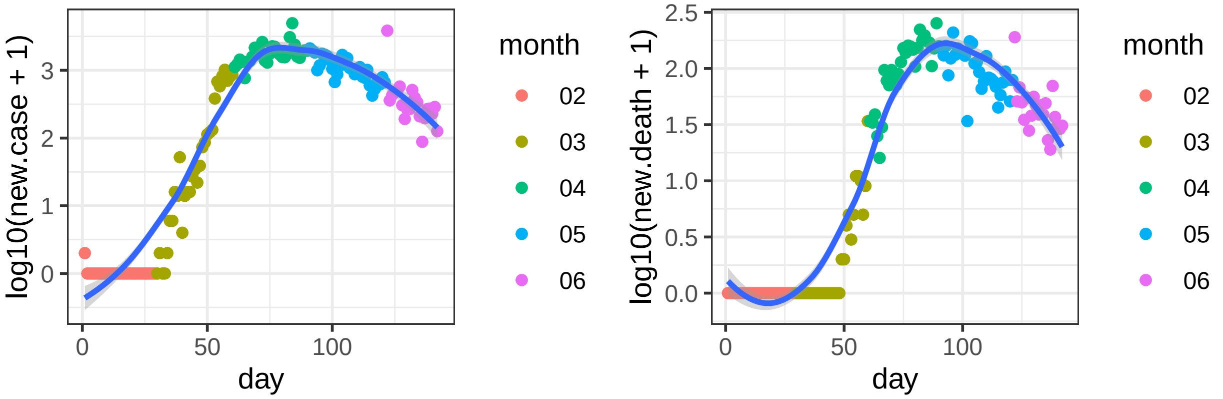
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



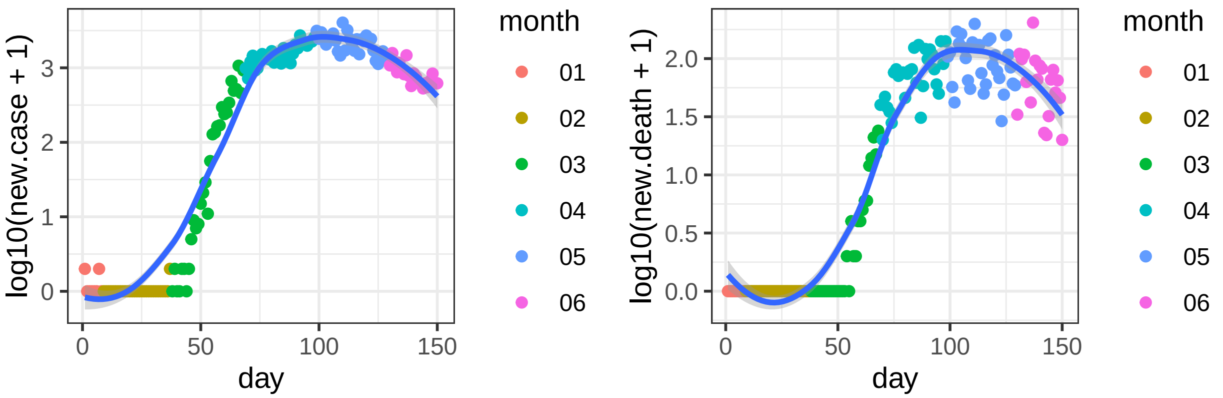
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Massachusetts



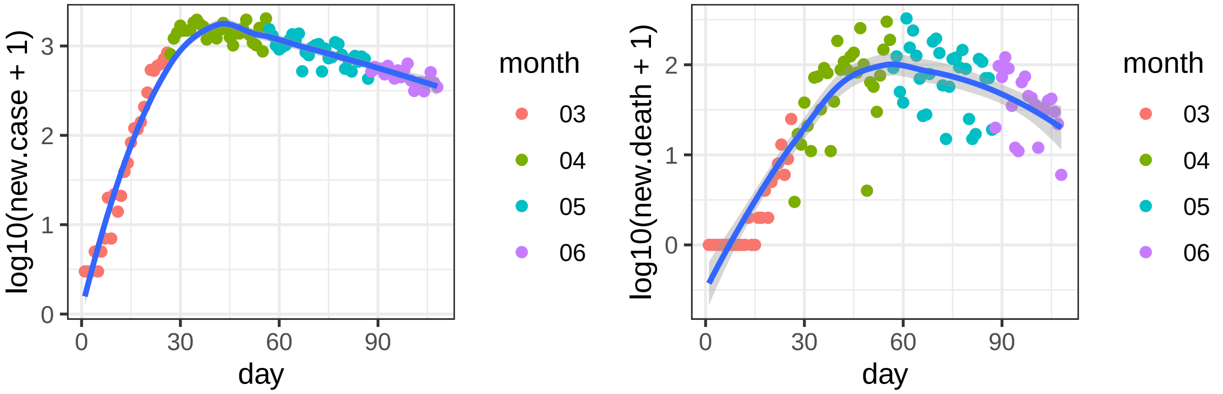
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Illinois



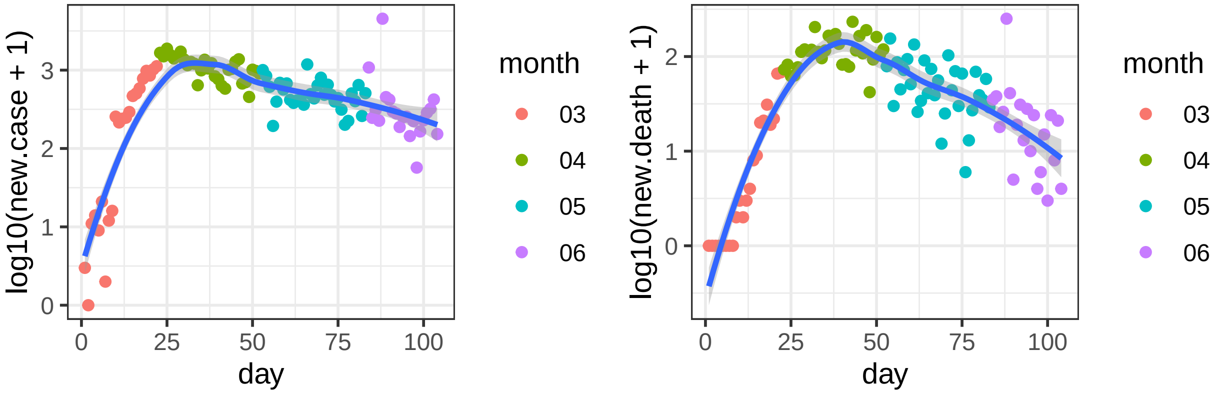
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Pennsylvania



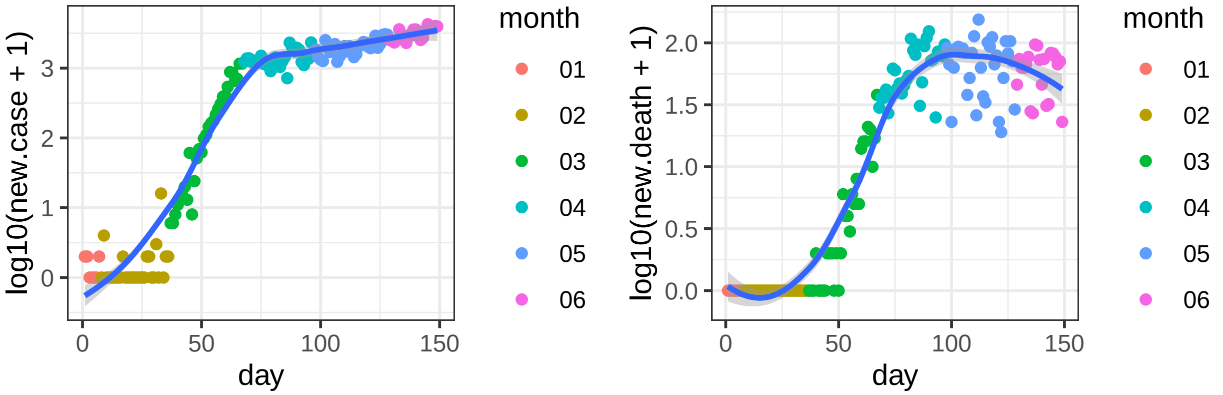
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Michigan



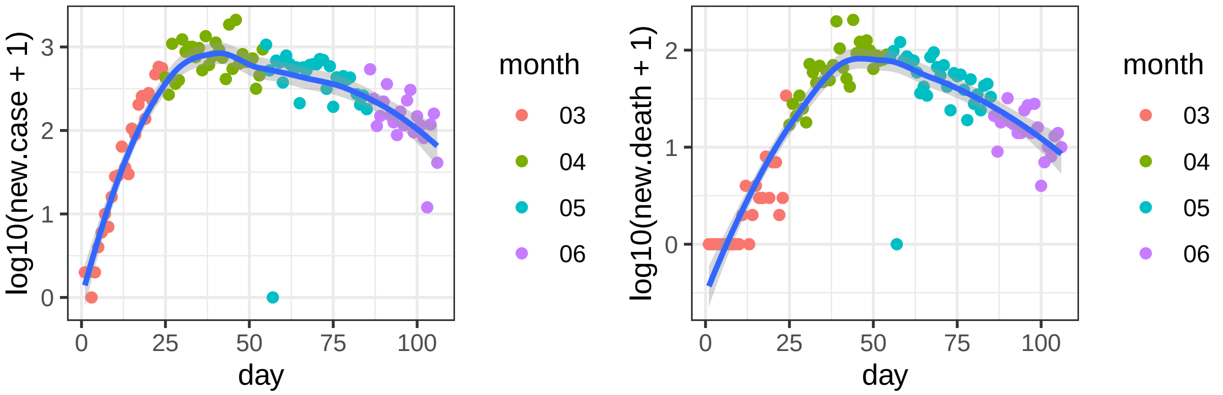
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

California



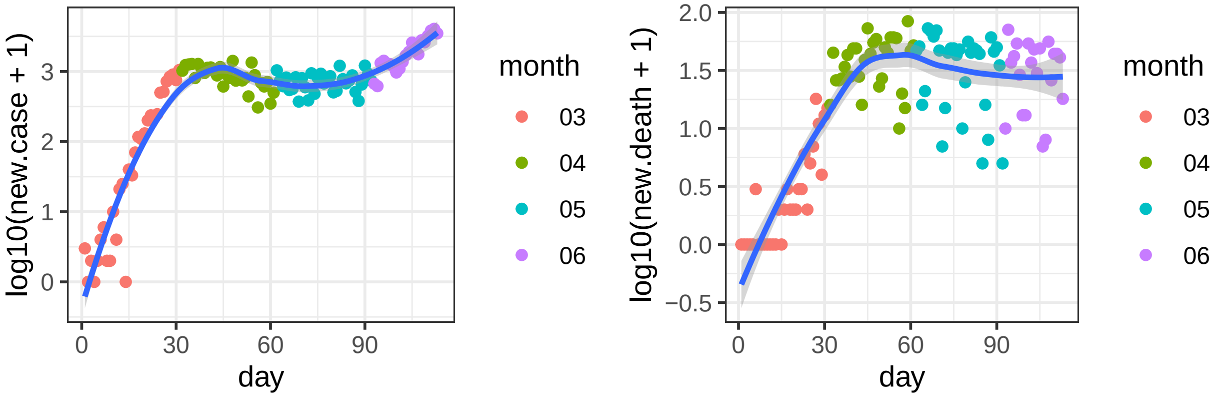
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Connecticut



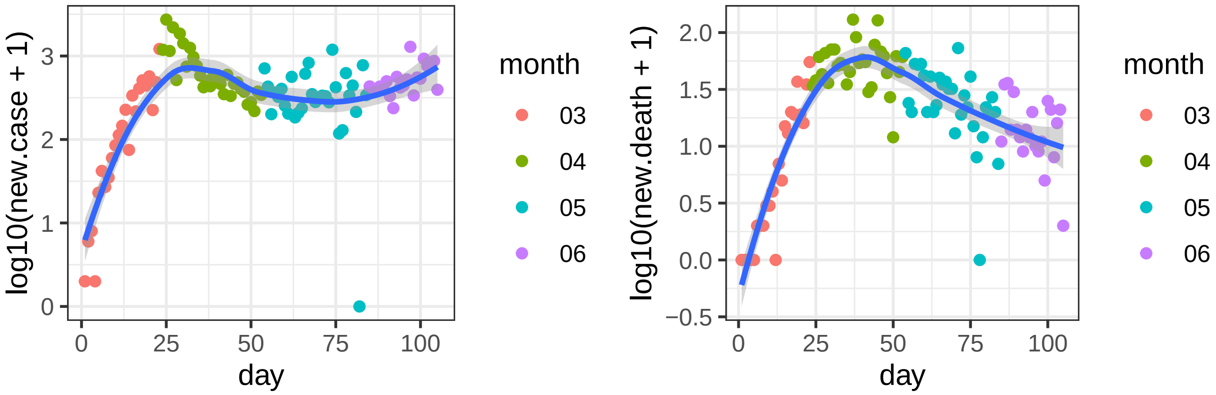
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Florida



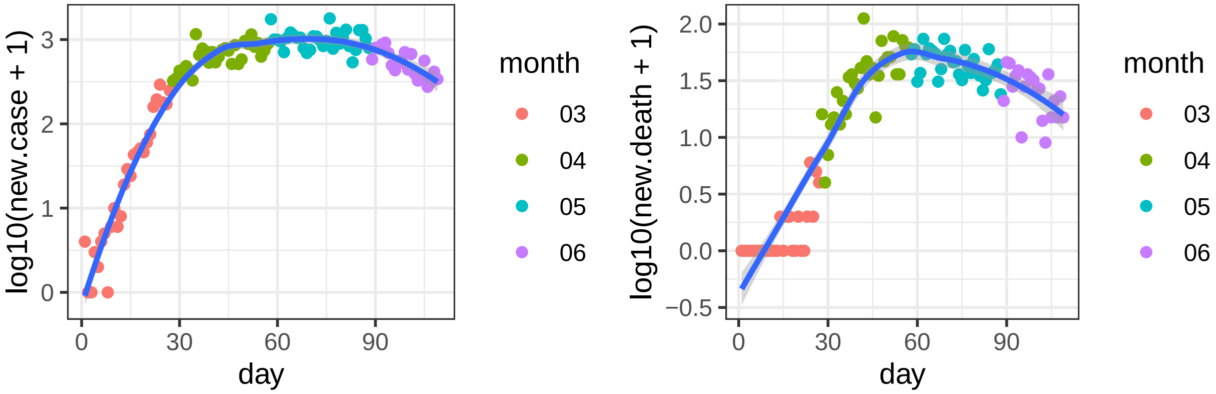
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Louisiana



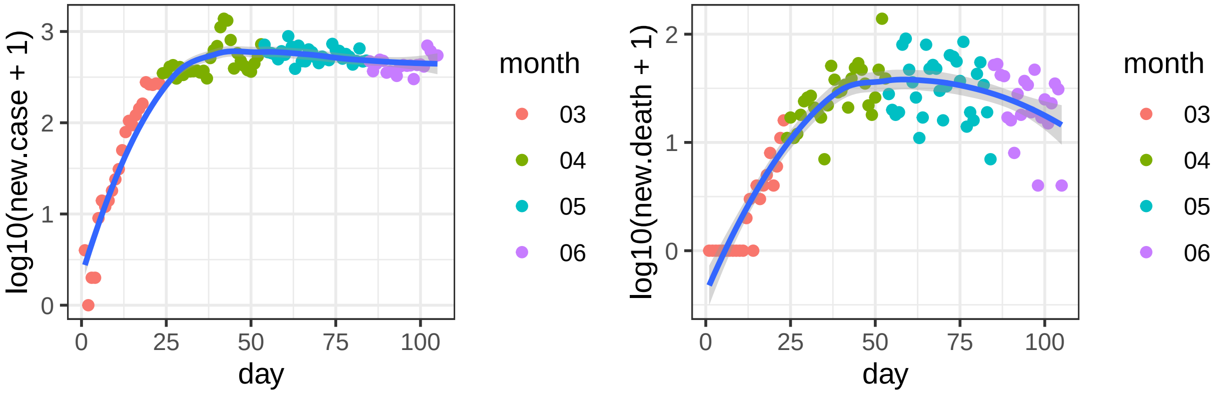
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland

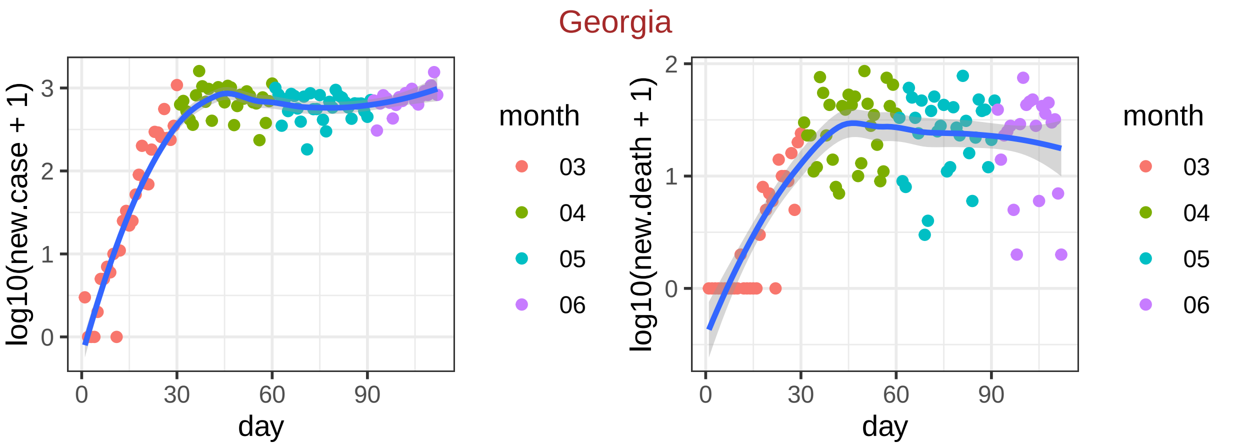


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

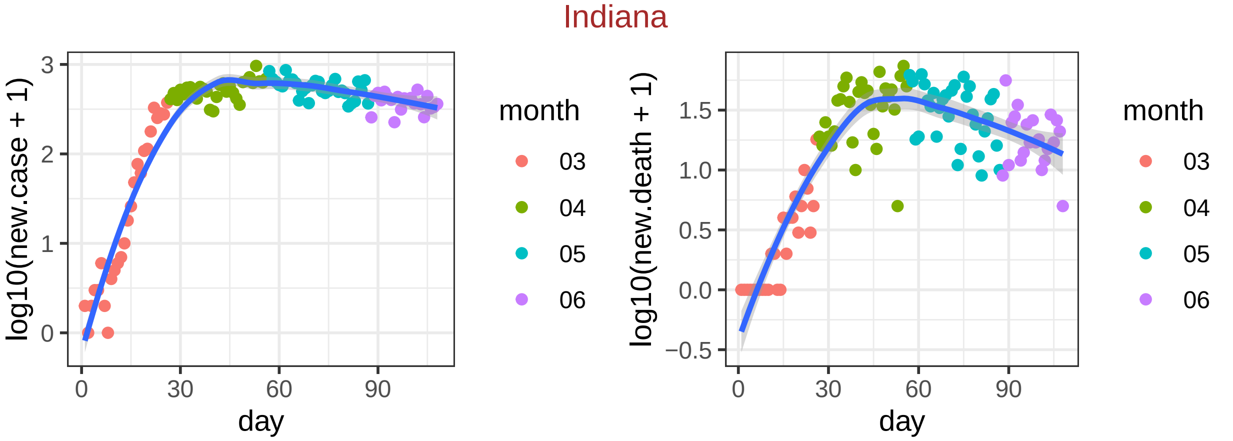
Ohio



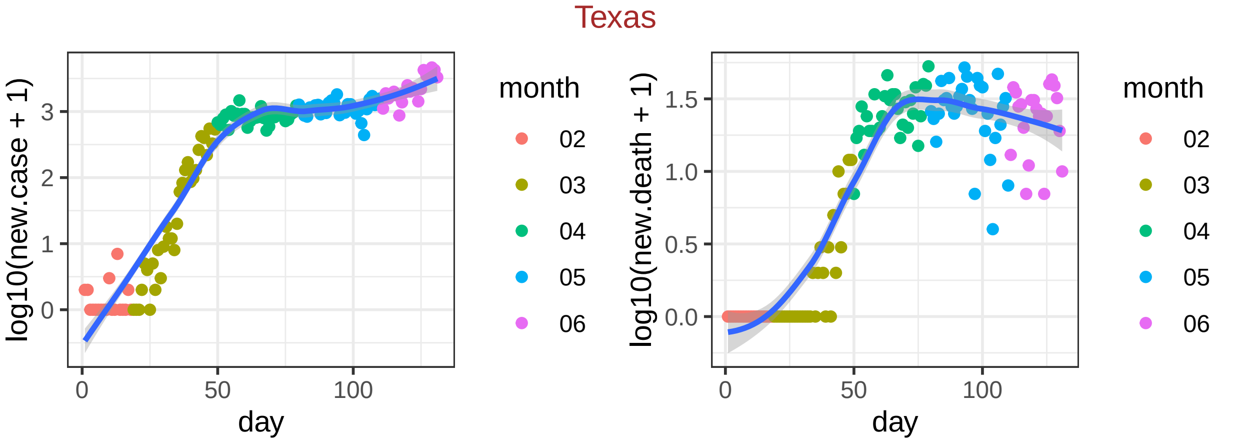
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09



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

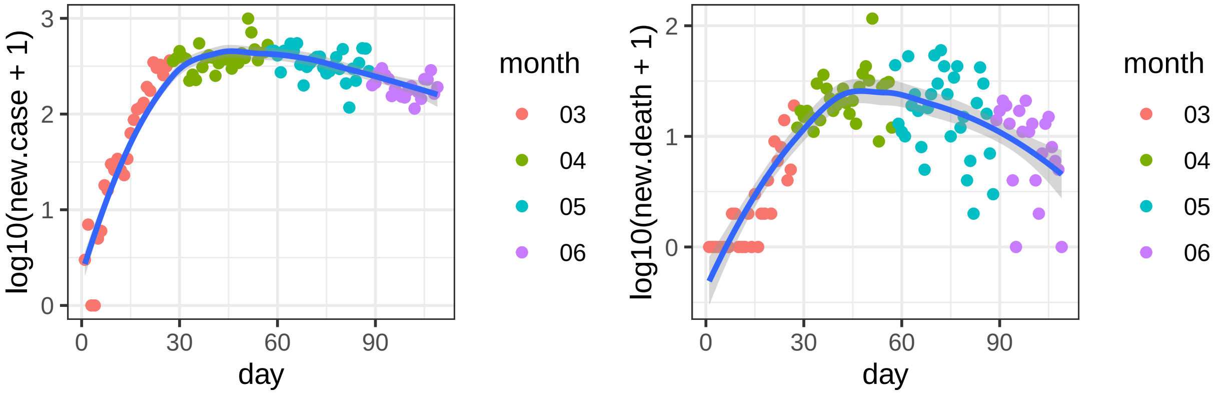


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



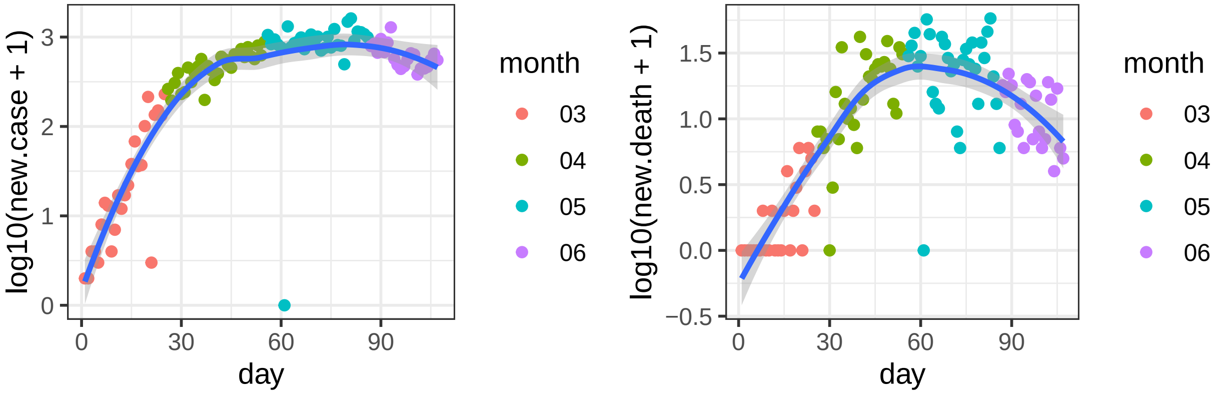
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Colorado



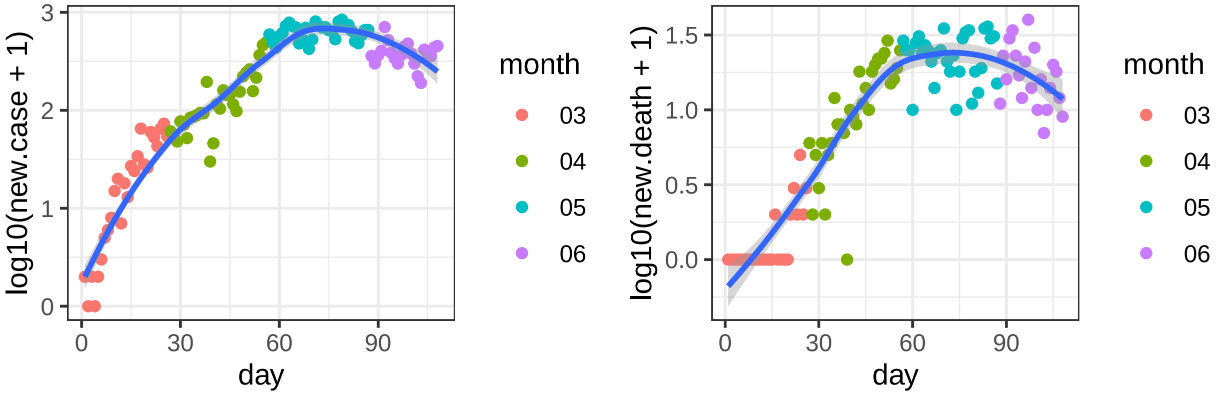
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Virginia



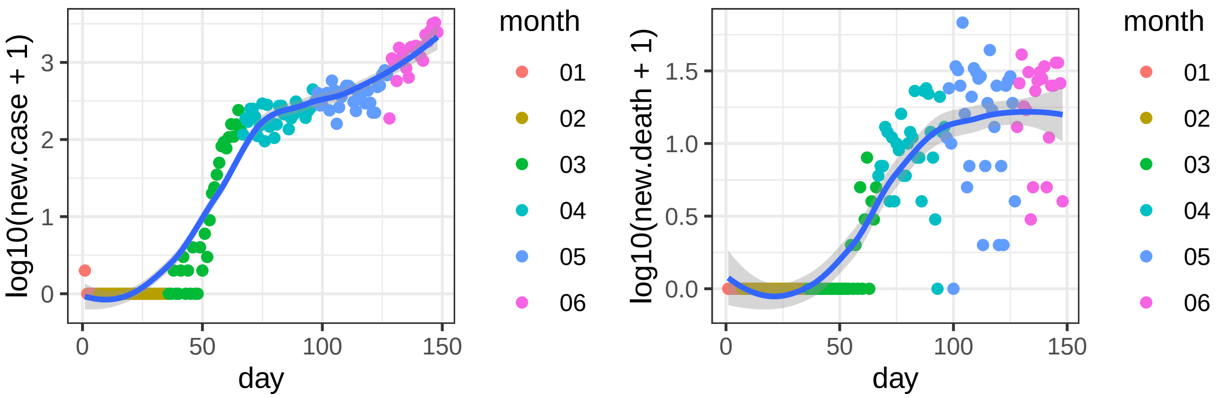
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Minnesota



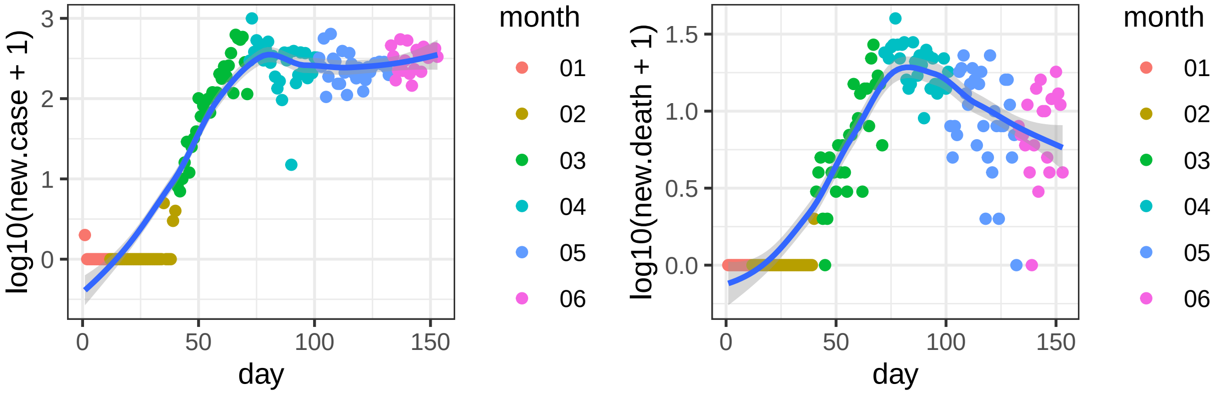
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Arizona



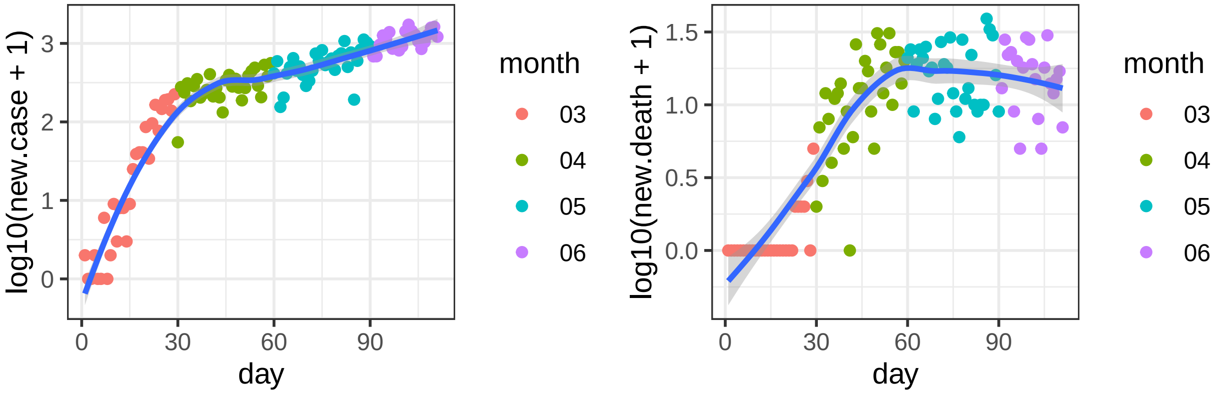
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Washington



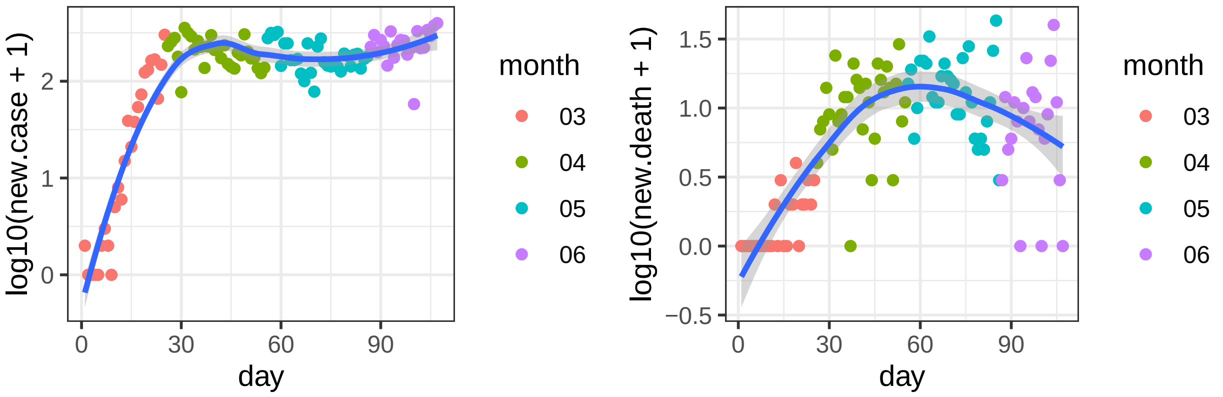
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

North Carolina



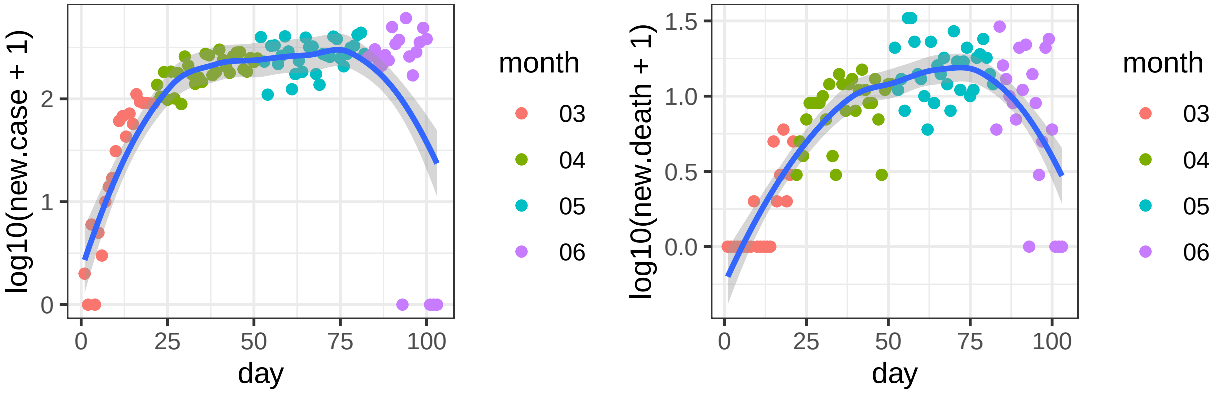
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Missouri



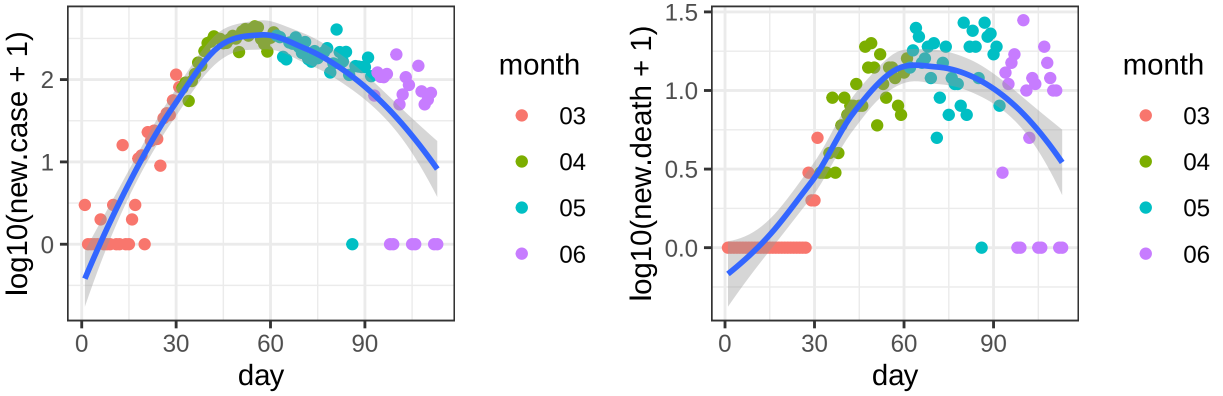
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Mississippi



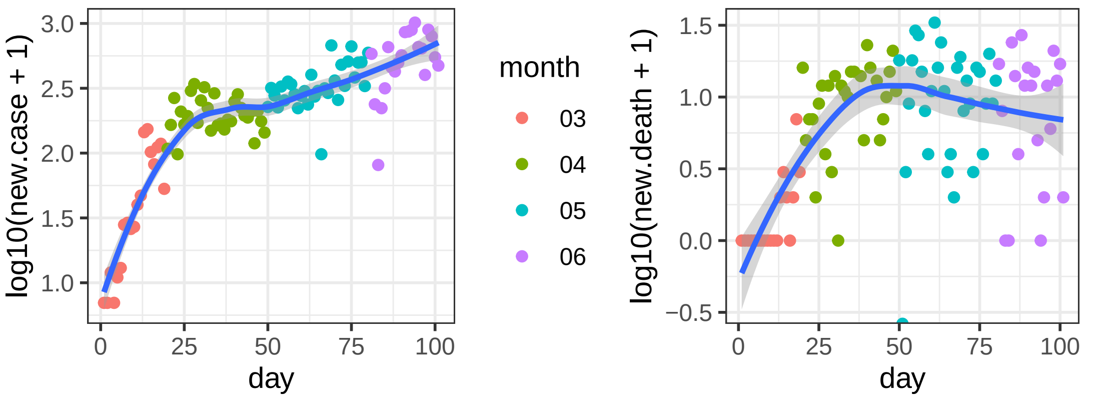
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Rhode Island



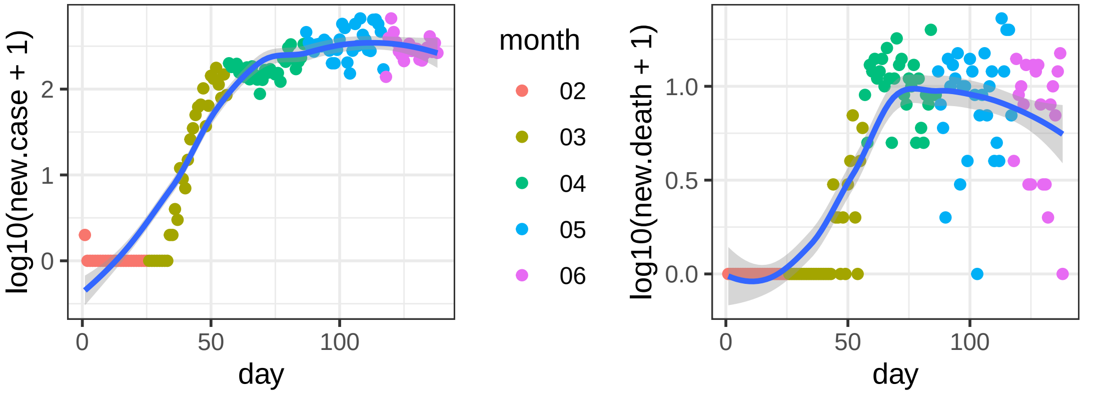
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Alabama



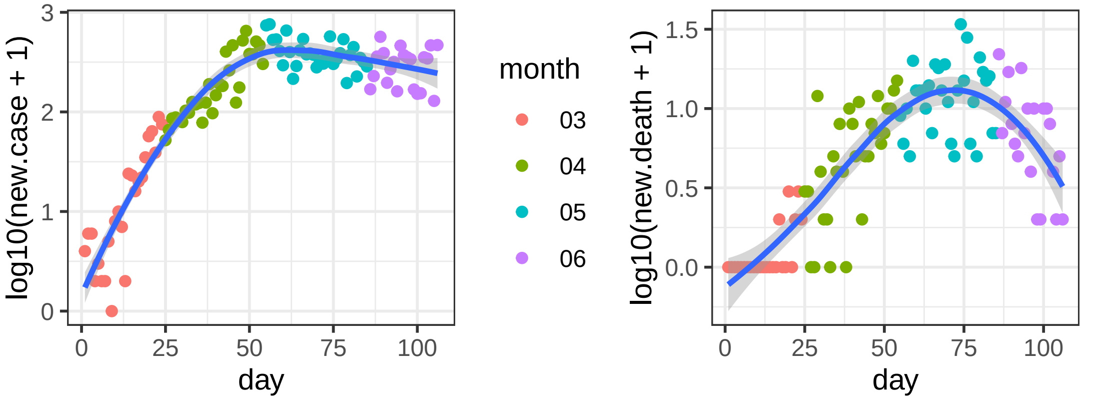
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Wisconsin

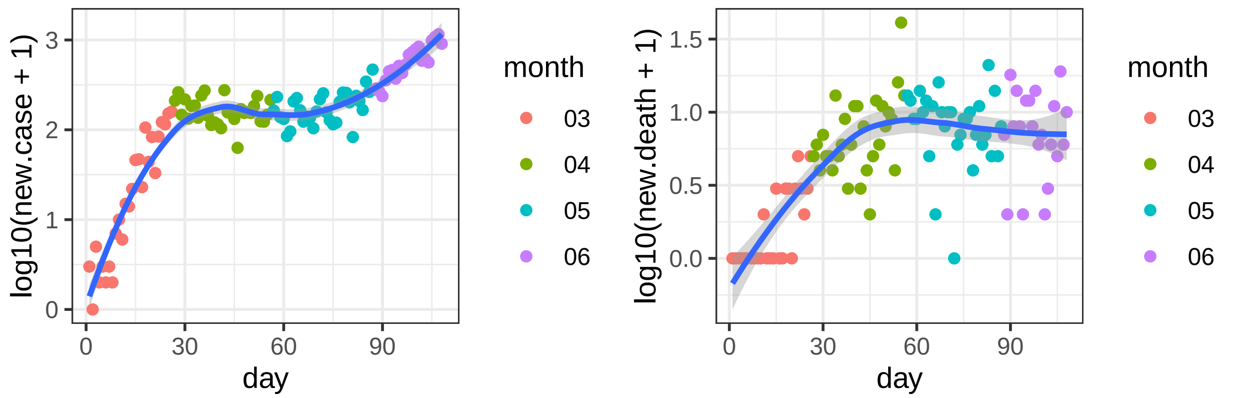


data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

Iowa

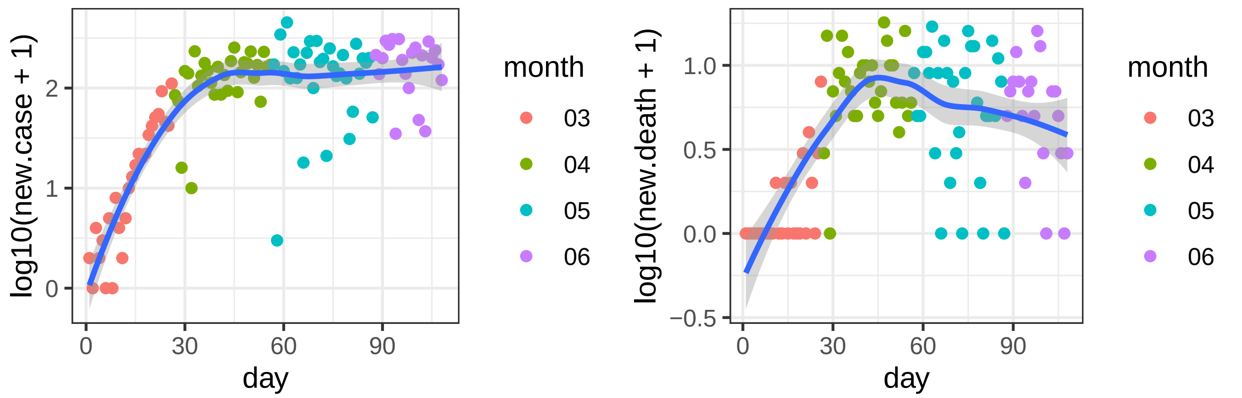


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08



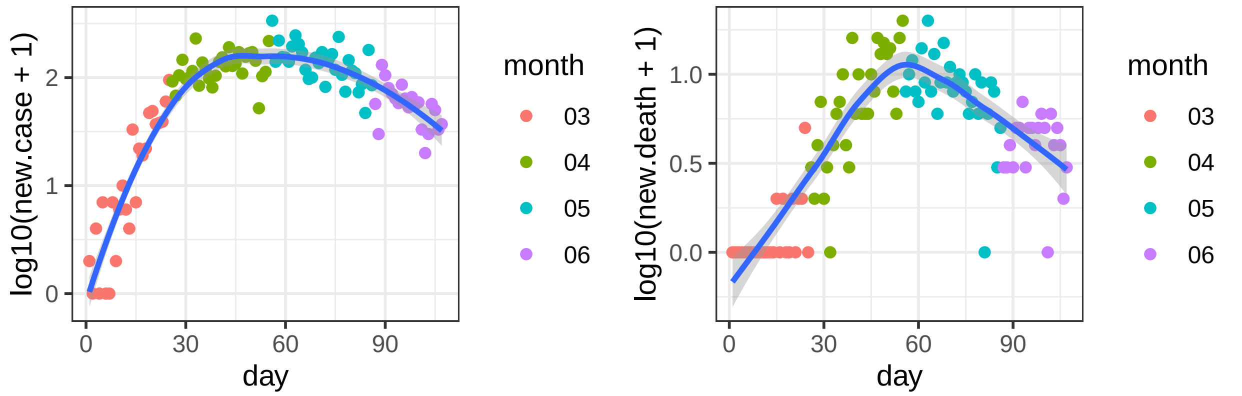
South Carolina

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



Kentucky

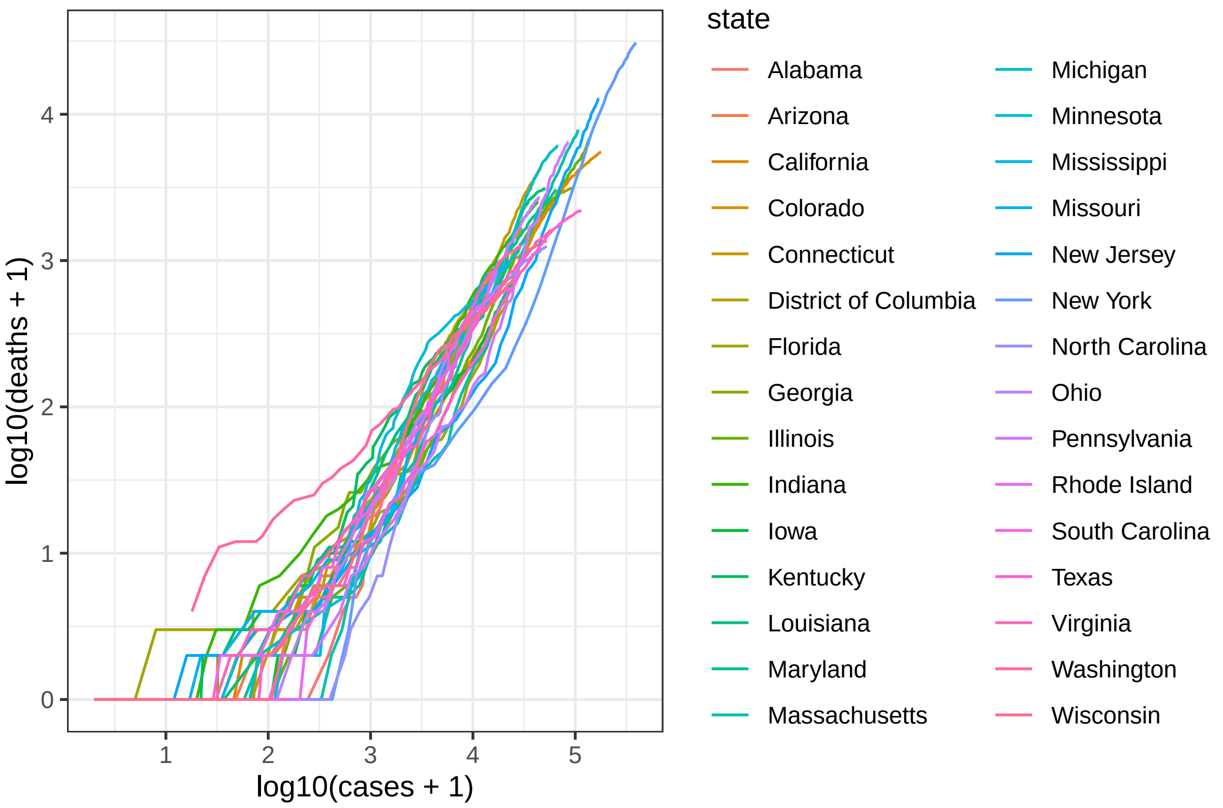
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



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

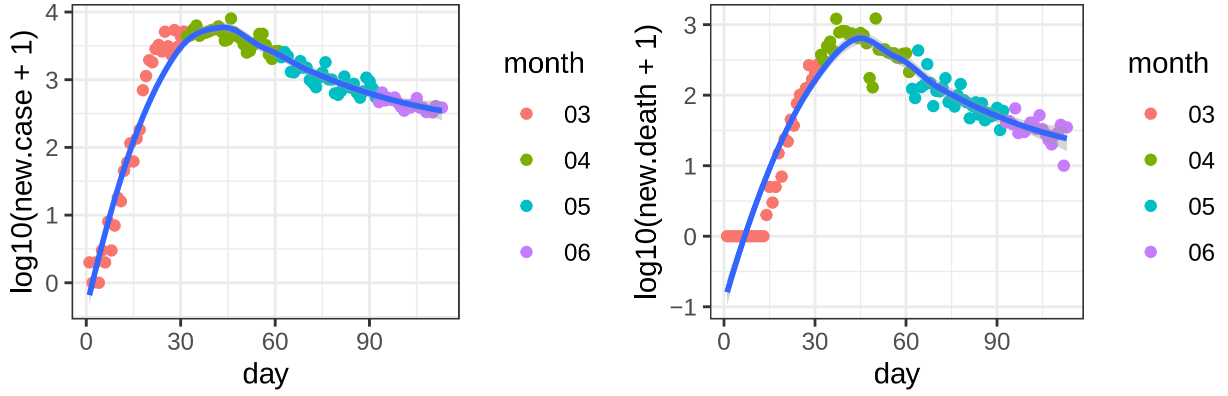
county level data

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

```
##              date          county         state  fips  cases deaths
## 256774 2020-06-21   New York City      New York    NA 217189  21753
## 255577 2020-06-21            Cook      Illinois 17031  87177   4404
## 255181 2020-06-21     Los Angeles    California  6037  83397   3120
## 256269 2020-06-21           Wayne      Michigan 26163  22162   2689
## 256773 2020-06-21          Nassau      New York 36059  41479   2683
## 256793 2020-06-21         Suffolk      New York 36103  40972   2013
## 256181 2020-06-21       Middlesex Massachusetts 25017  23574   1807
## 256698 2020-06-21           Essex    New Jersey 34013  18551   1760
## 256693 2020-06-21          Bergen    New Jersey 34003  19010   1696
## 257202 2020-06-21    Philadelphia  Pennsylvania 42101  24841   1553
## 256801 2020-06-21     Westchester      New York 36119  34520   1545
## 255280 2020-06-21       Fairfield   Connecticut  9001  16475   1361
## 255281 2020-06-21        Hartford   Connecticut  9003  11405   1350
## 256700 2020-06-21          Hudson    New Jersey 34017  18744   1262
## 256711 2020-06-21           Union    New Jersey 34039  16322   1135
## 256703 2020-06-21       Middlesex    New Jersey 34023  16605   1101
## 256250 2020-06-21         Oakland      Michigan 26125  11685   1077
## 256177 2020-06-21           Essex Massachusetts 25009  15829   1076
## 255284 2020-06-21       New Haven   Connecticut  9009  12185   1061
## 256707 2020-06-21         Passaic    New Jersey 34031  16769   1014
## 256185 2020-06-21         Suffolk Massachusetts 25025  19551    978
## 256183 2020-06-21         Norfolk Massachusetts 25021   8994    912
## 256237 2020-06-21              Macomb            Michigan 26099   7391     904
## 256187 2020-06-21           Worcester       Massachusetts 25027  12130     900
## 255336 2020-06-21          Miami-Dade             Florida 12086  25789     884
## 256706 2020-06-21               Ocean          New Jersey 34029   9425     847
## 257197 2020-06-21          Montgomery        Pennsylvania 42091   8103     784
## 256297 2020-06-21            Hennepin           Minnesota 27053  10830     747
## 256163 2020-06-21          Montgomery            Maryland 24031  14119     720
## 255712 2020-06-21              Marion             Indiana 18097  11067     714
## 256704 2020-06-21            Monmouth          New Jersey 34025   8942     695
## 257174 2020-06-21            Delaware        Pennsylvania 42045   7084     684
## 256164 2020-06-21     Prince George's            Maryland 24033  18367     659
## 256179 2020-06-21             Hampden       Massachusetts 25013   6598     648
## 256184 2020-06-21            Plymouth       Massachusetts 25023   8583     643
## 256705 2020-06-21              Morris          New Jersey 34027   6699     641
## 257223 2020-06-21          Providence        Rhode Island 44007  12363     637
## 255080 2020-06-21            Maricopa             Arizona  4013  30136     632
## 257857 2020-06-21                King          Washington 53033   9236     602
## 256759 2020-06-21                Erie            New York 36029   7004     590
## 257160 2020-06-21               Bucks        Pennsylvania 42017   5547     552
## 256540 2020-06-21           St. Louis            Missouri 29189   5850     550
## 256175 2020-06-21             Bristol       Massachusetts 25005   8035     542
## 255293 2020-06-21 District of Columbia District of Columbia 11001  10020     533
## 256101 2020-06-21             Orleans           Louisiana 22071   7518     529
## 256702 2020-06-21              Mercer          New Jersey 34021   7541     524
## 256091 2020-06-21           Jefferson           Louisiana 22051   8681     477
## 256785 2020-06-21            Rockland            New York 36087  13504     469
## 255343 2020-06-21          Palm Beach             Florida 12099  10752     468
## 256150 2020-06-21           Baltimore            Maryland 24005   7560     451
```

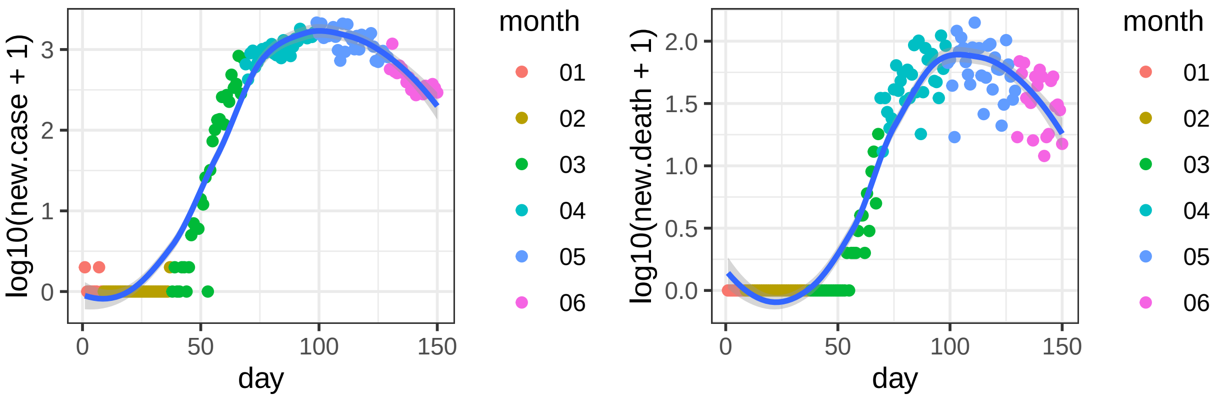
For these 50 counties, I check the number of new cases and the number of new deaths.

New York City_New York



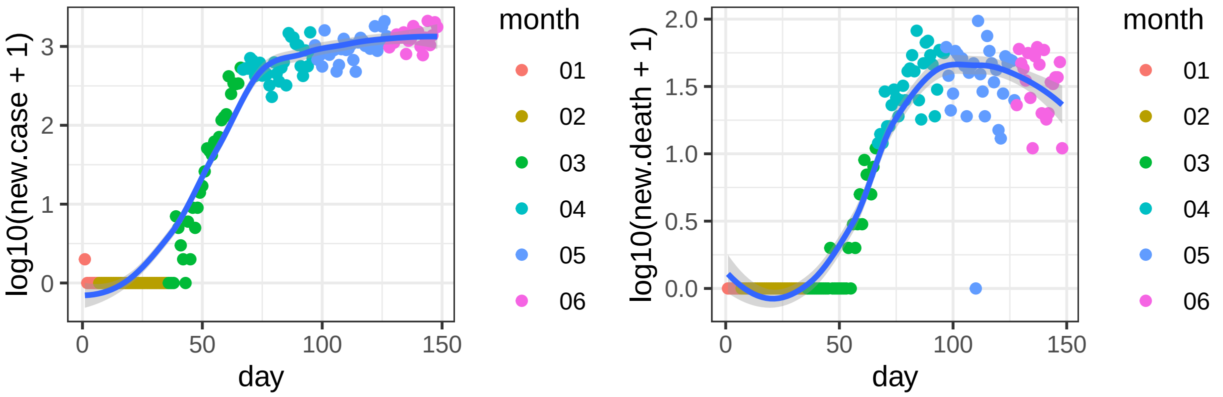
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



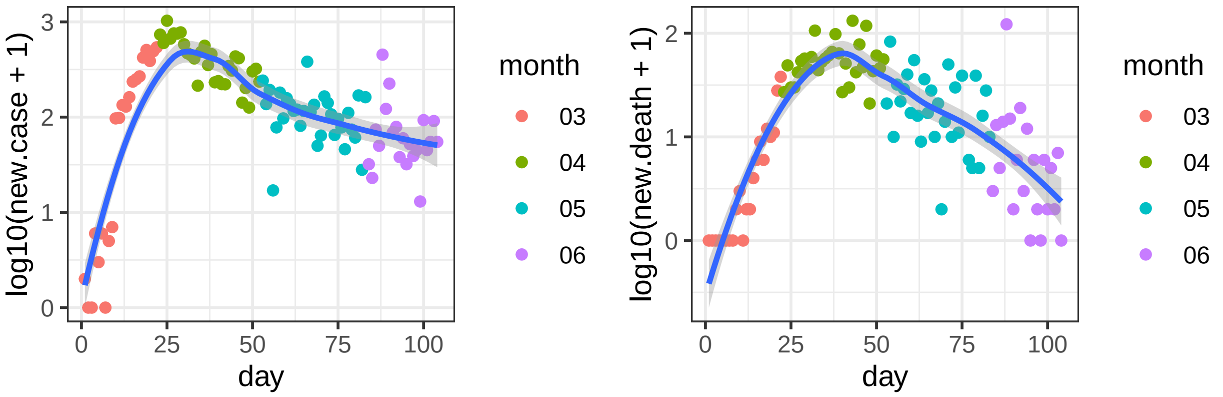
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Los Angeles_California



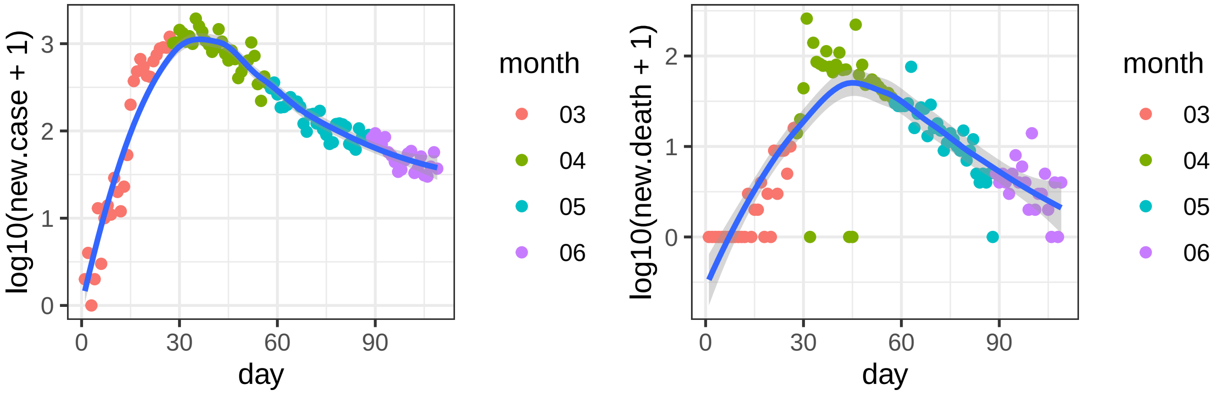
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Wayne_Michigan



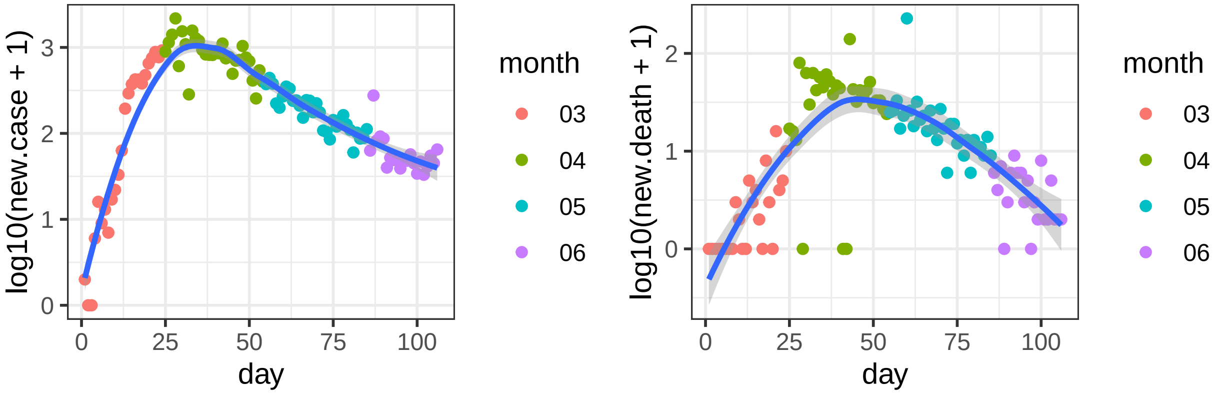
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Nassau_New York



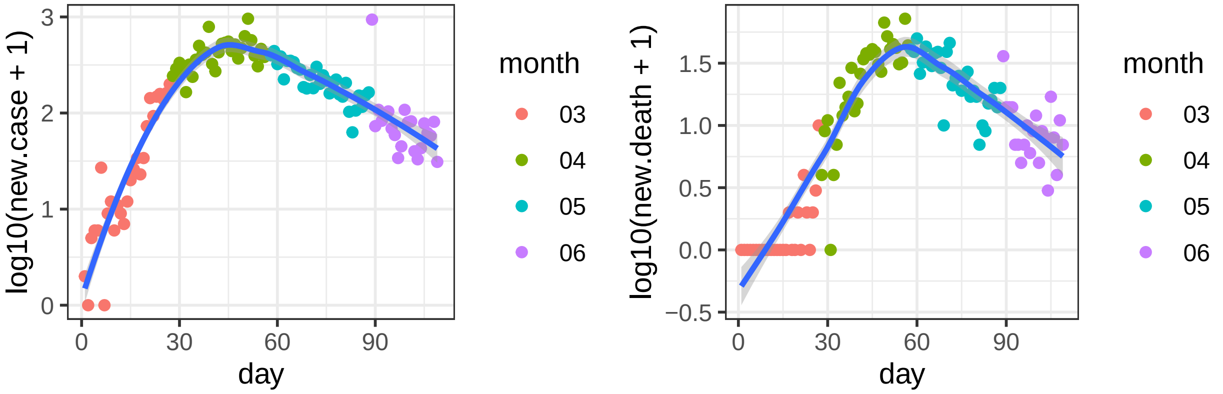
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Suffolk_New York



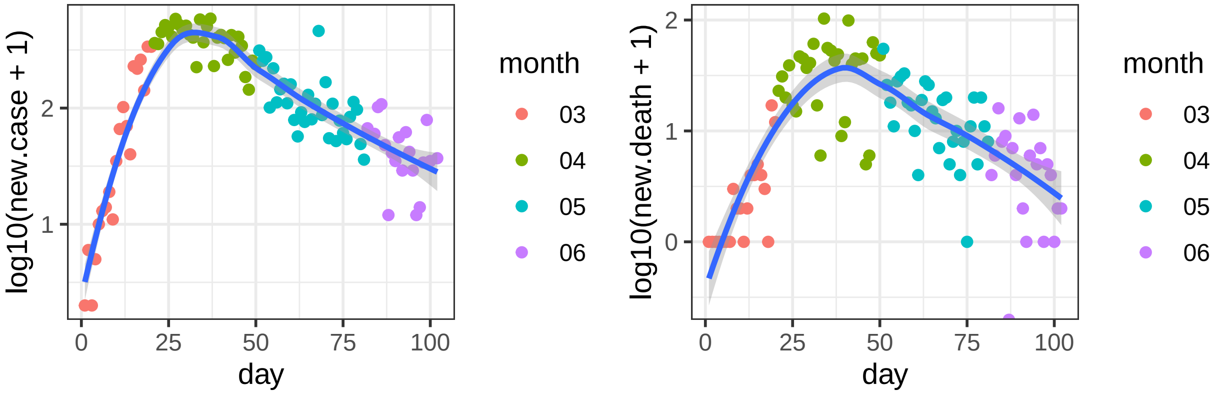
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Middlesex_Massachusetts



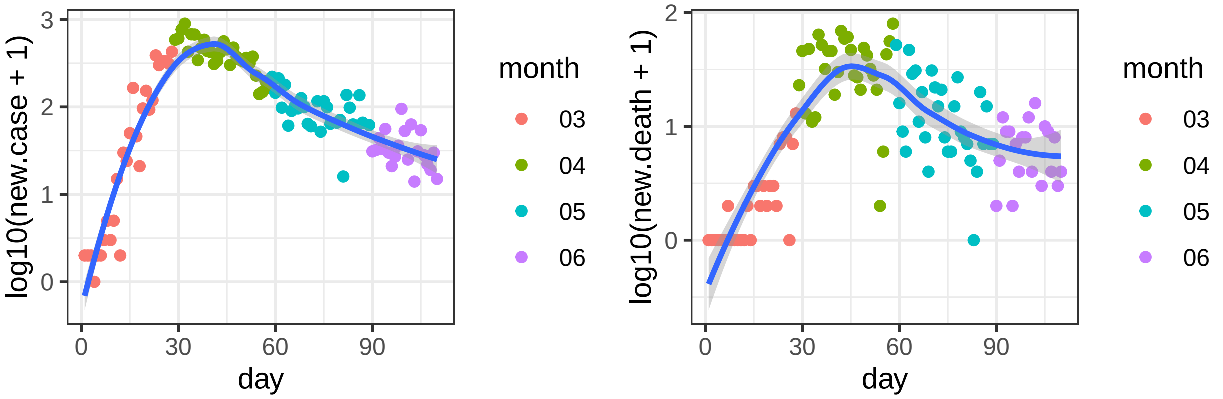
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



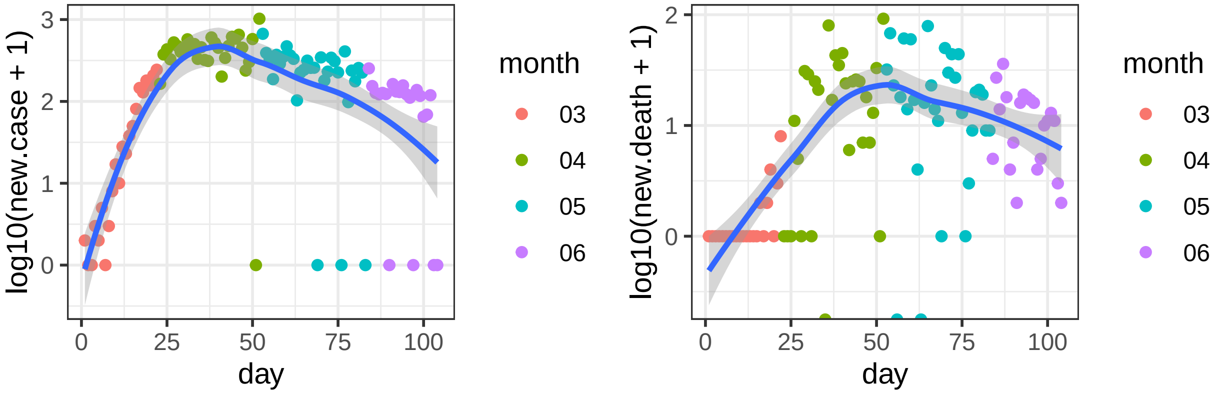
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Bergen_New Jersey



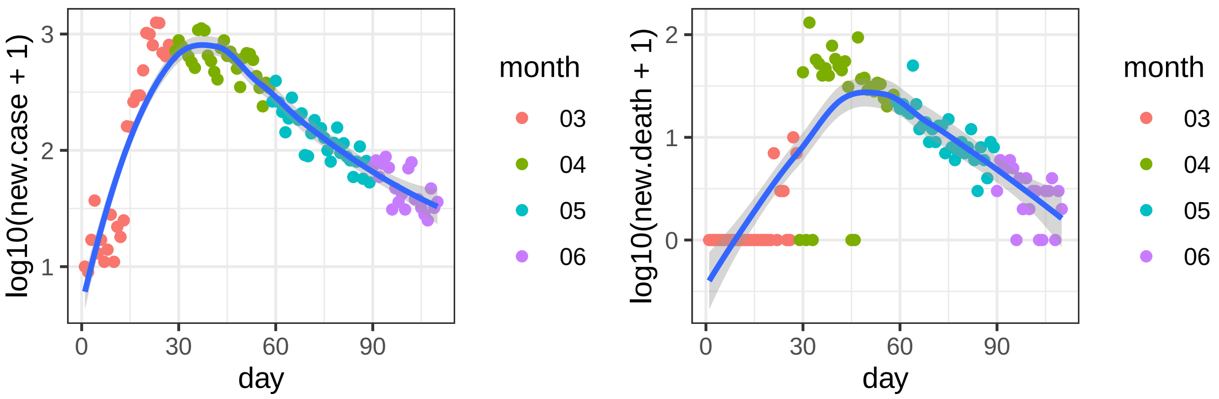
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Philadelphia_Pennsylvania



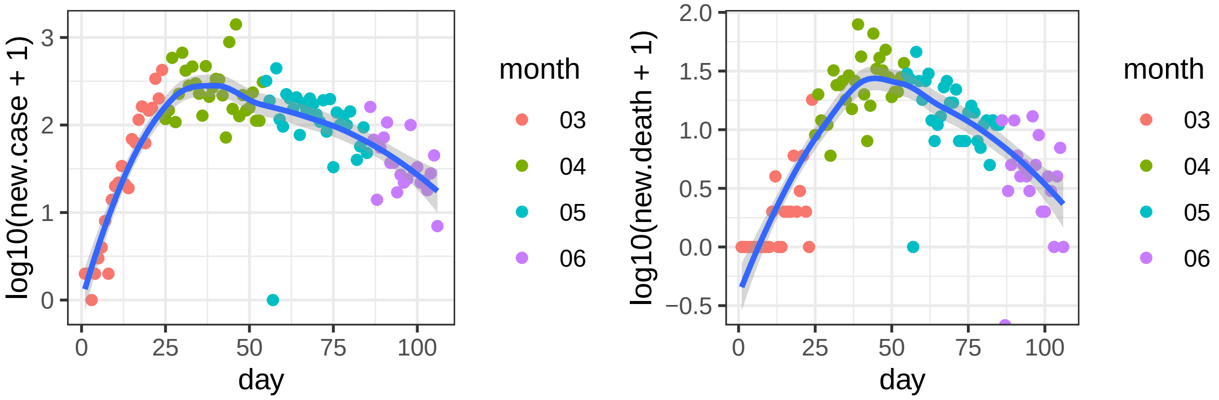
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Westchester_New York



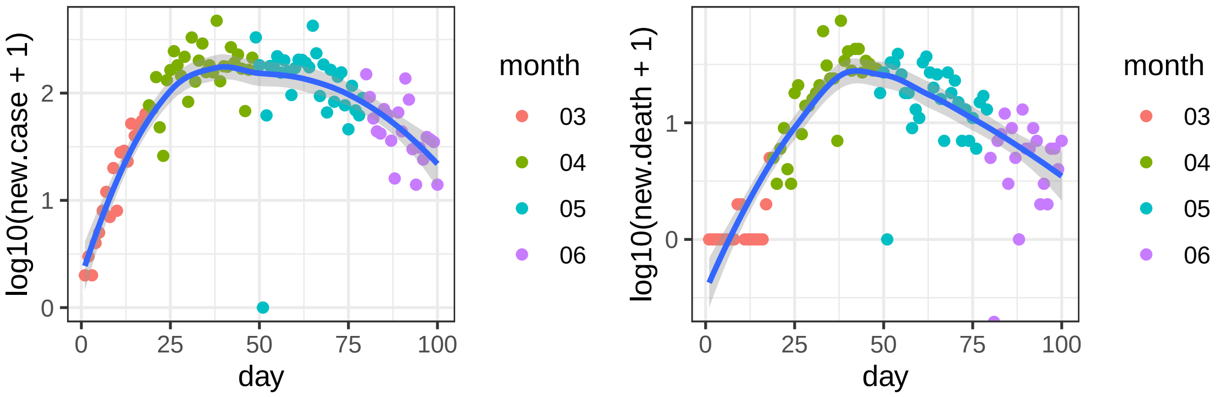
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Fairfield_Connecticut

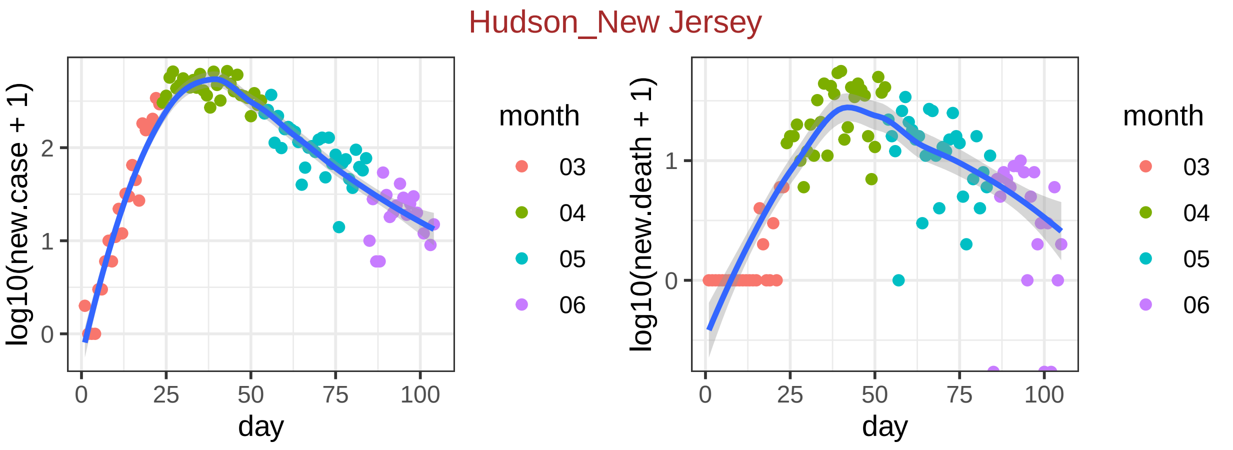


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

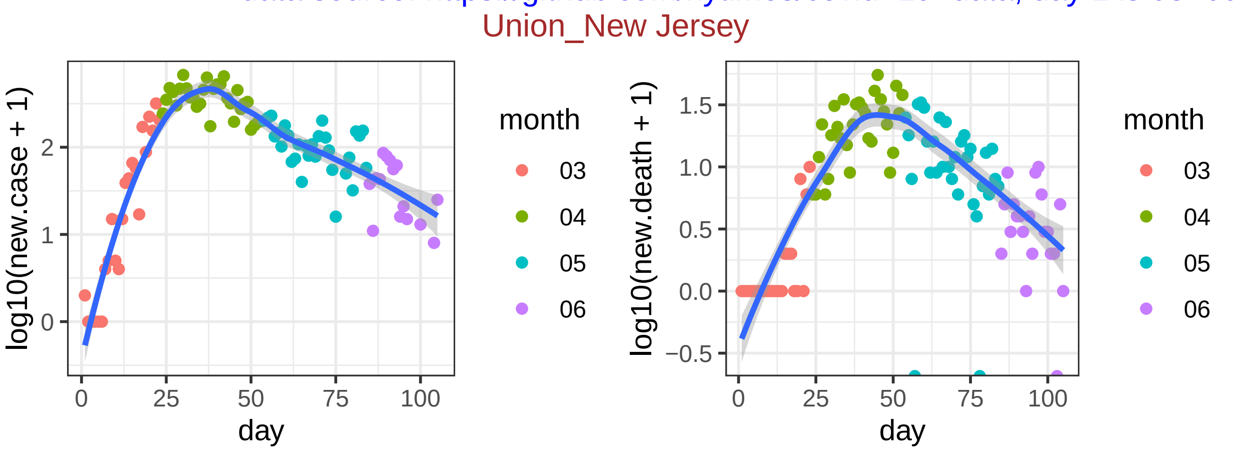
Hartford_Connecticut



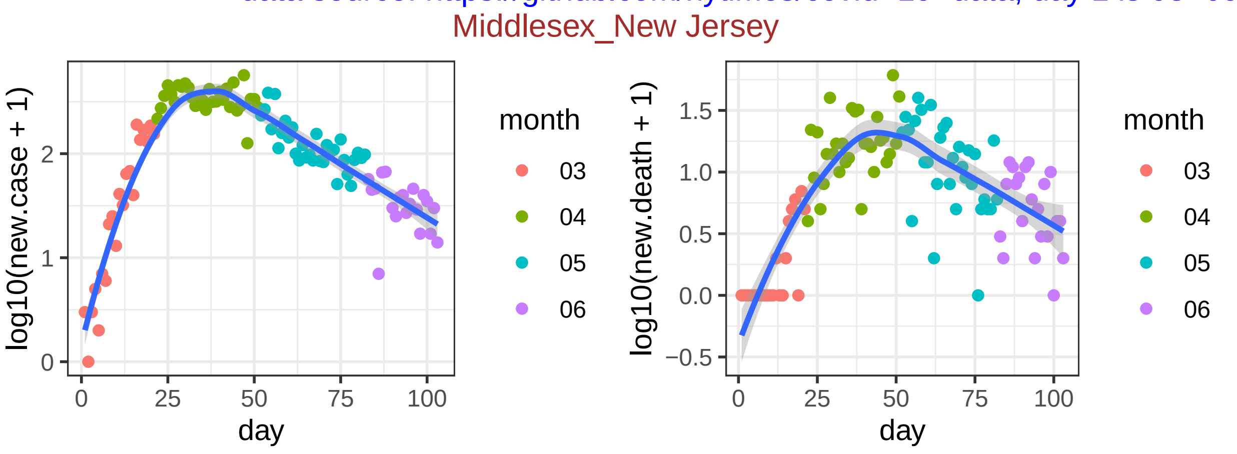
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14



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

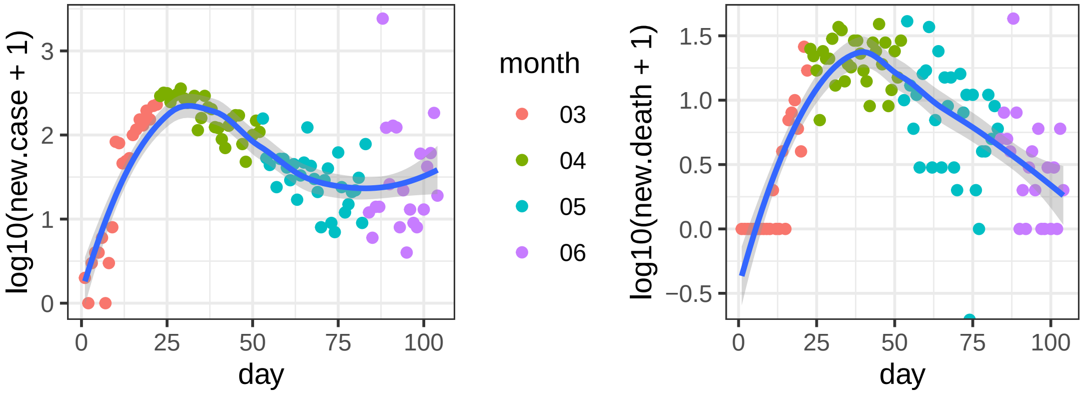


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09



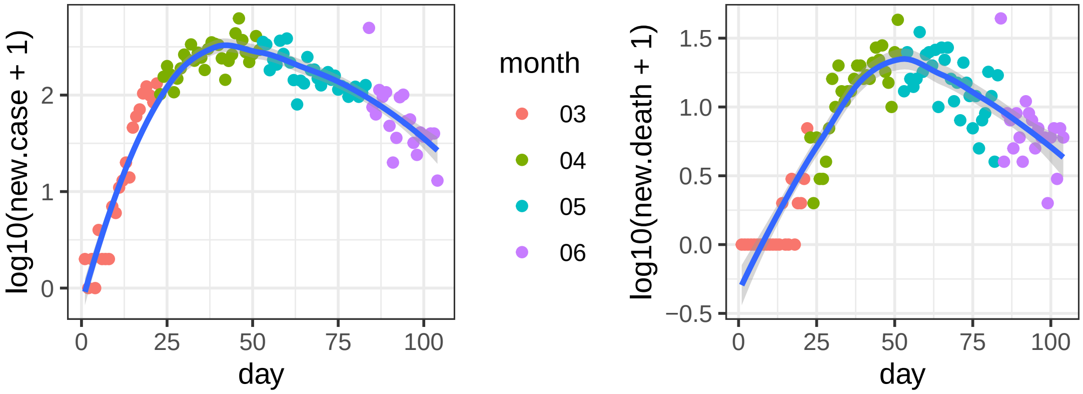
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Oakland_Michigan



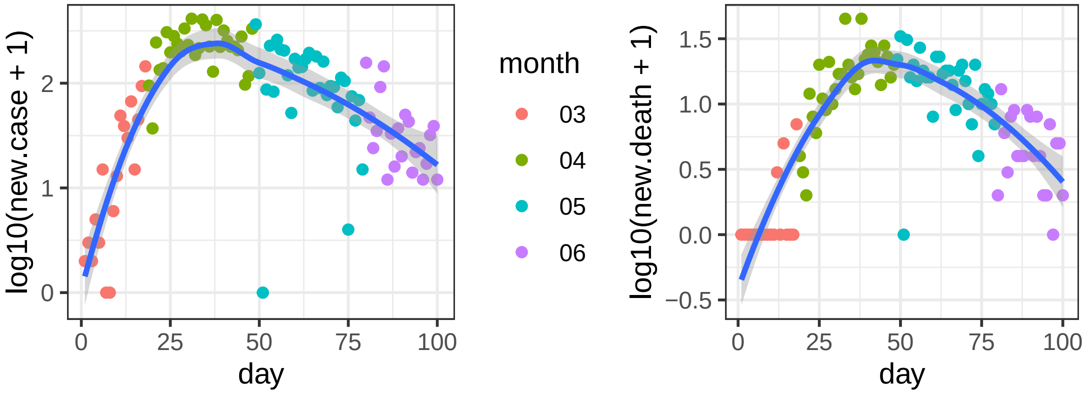
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Essex_Massachusetts



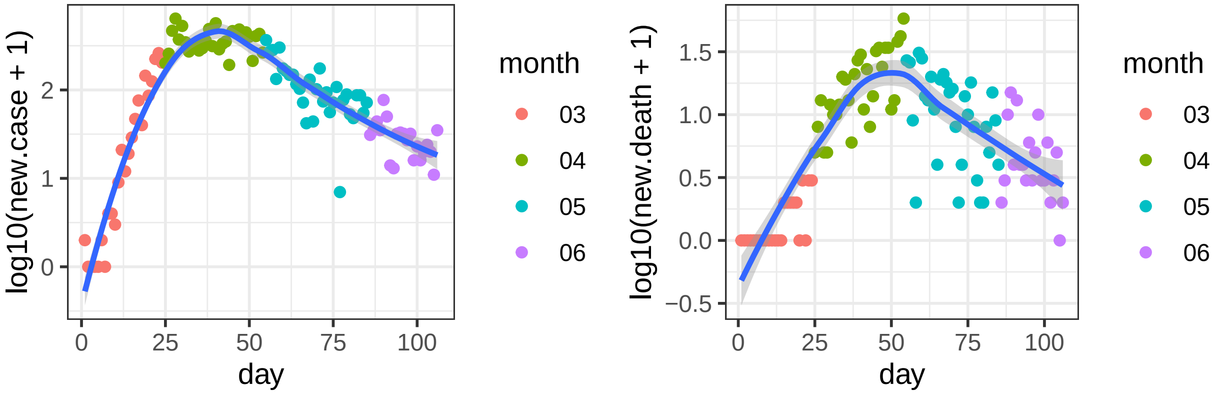
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

New Haven_Connecticut



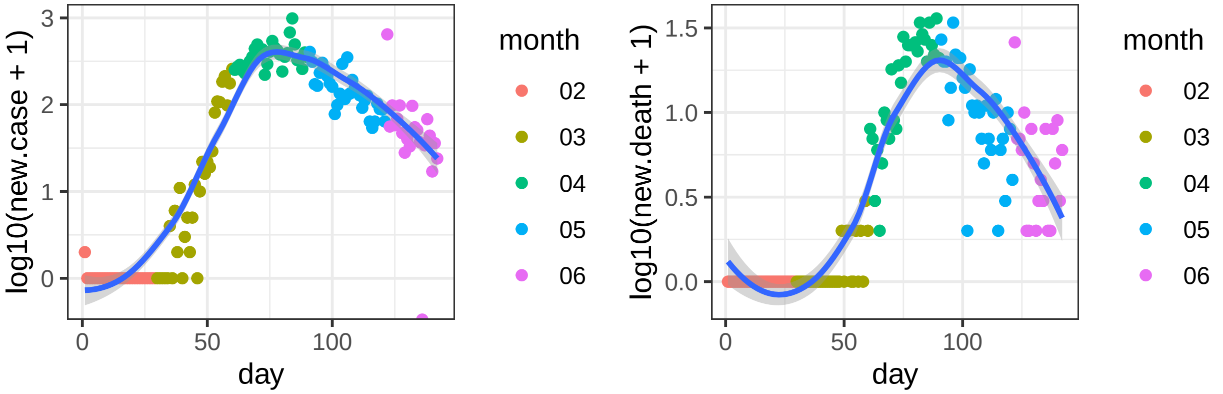
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Passaic_New Jersey



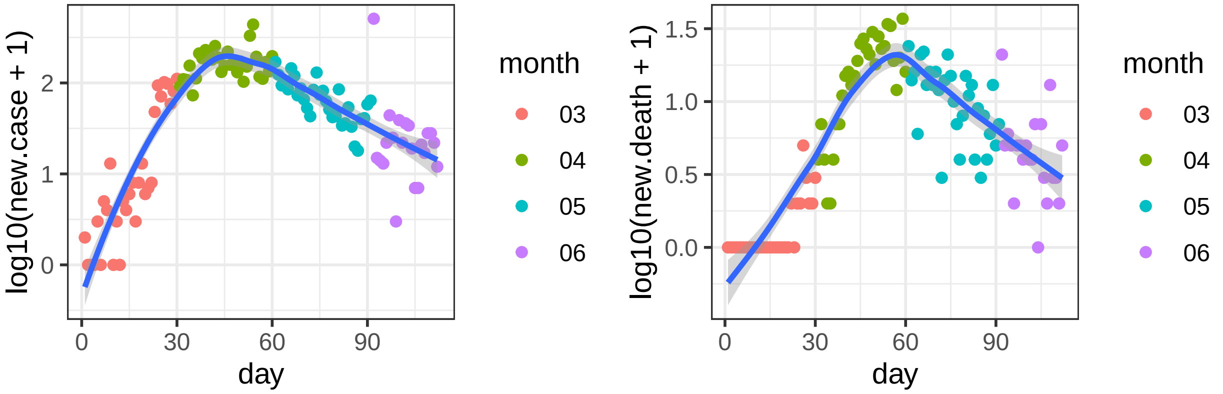
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Suffolk_Massachusetts



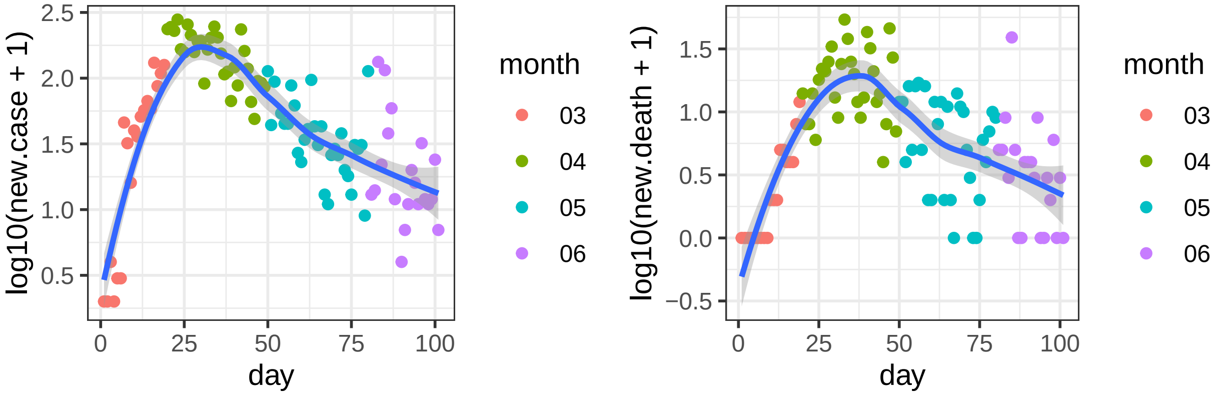
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Norfolk_Massachusetts



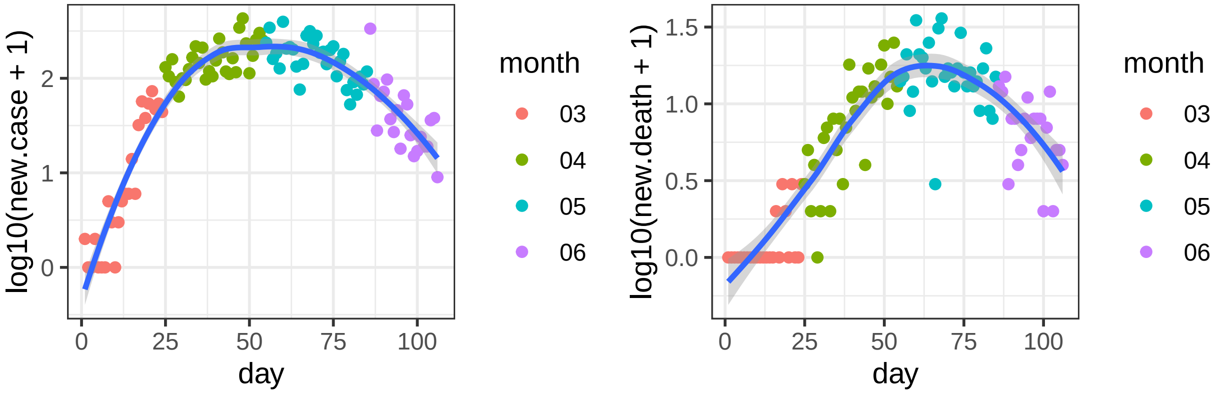
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Macomb_Michigan



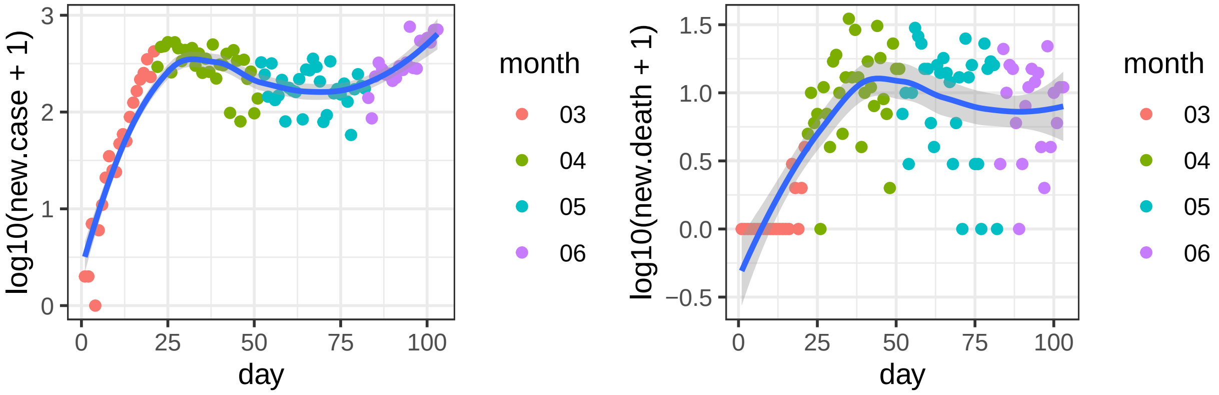
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Worcester_Massachusetts

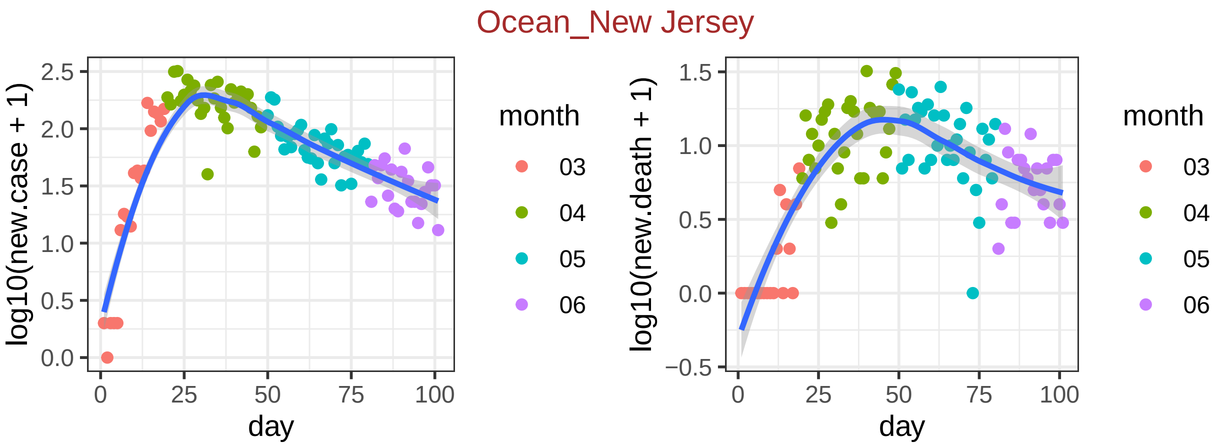


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

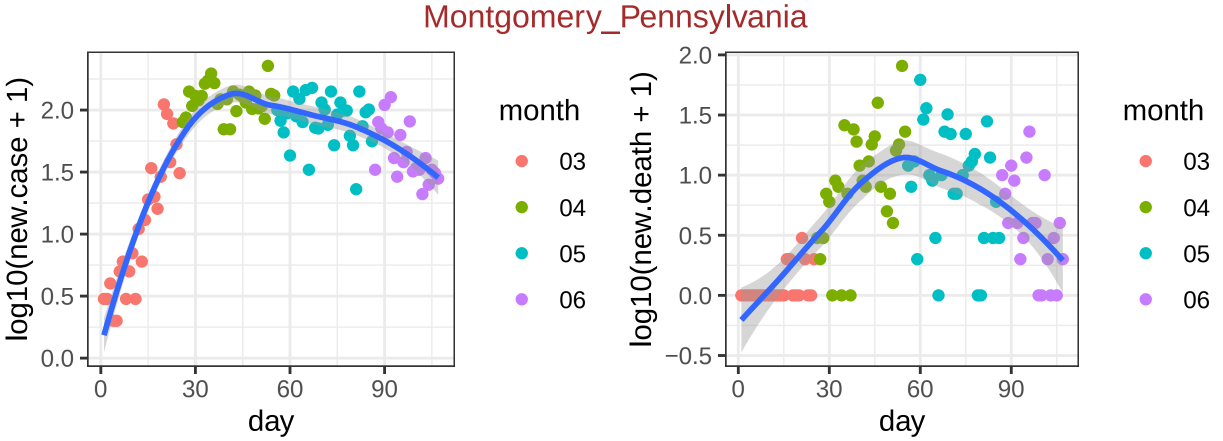
Miami−Dade_Florida



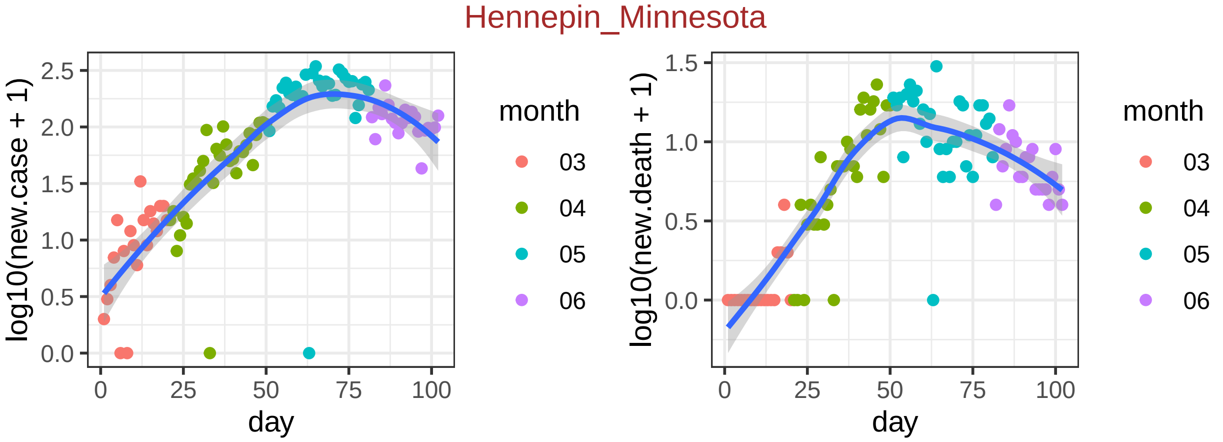
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11


Ocean_New Jersey

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

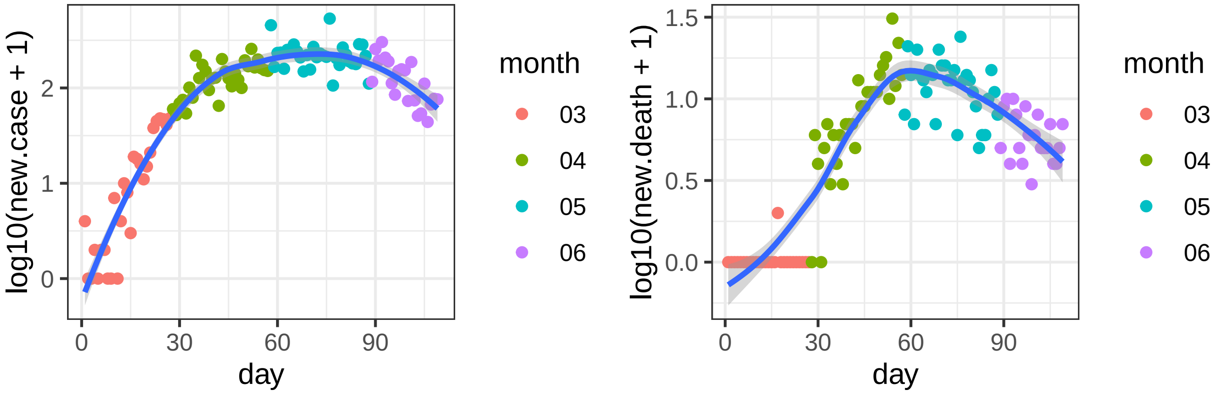

Montgomery_Pennsylvania

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


Hennepin_Minnesota

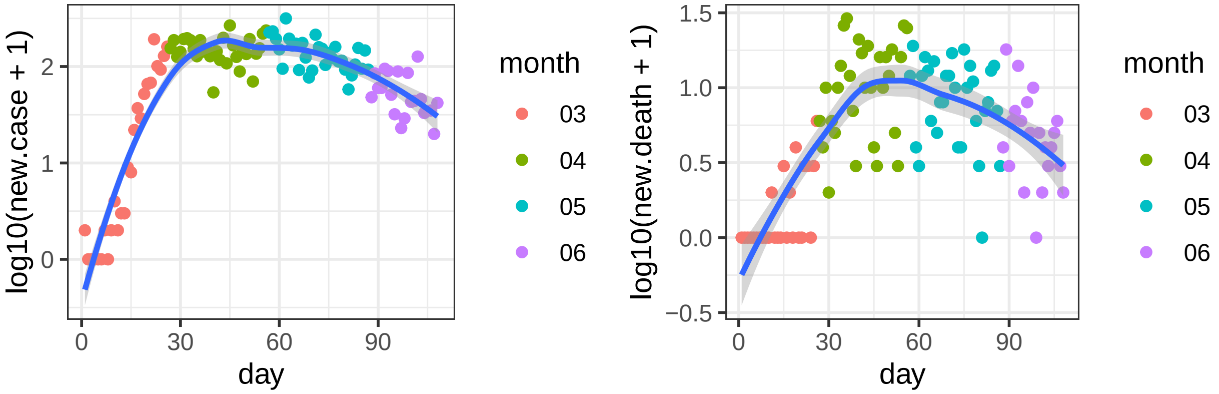
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Montgomery_Maryland



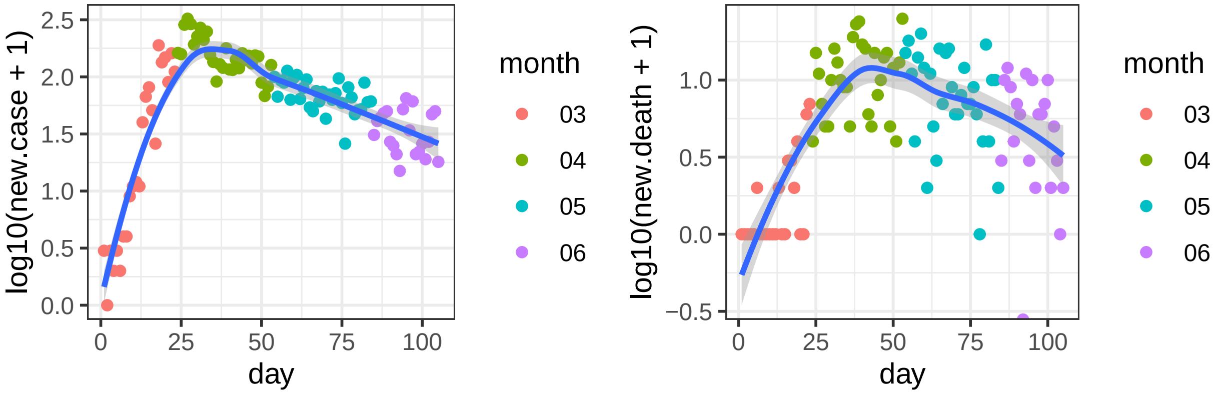
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Marion_Indiana



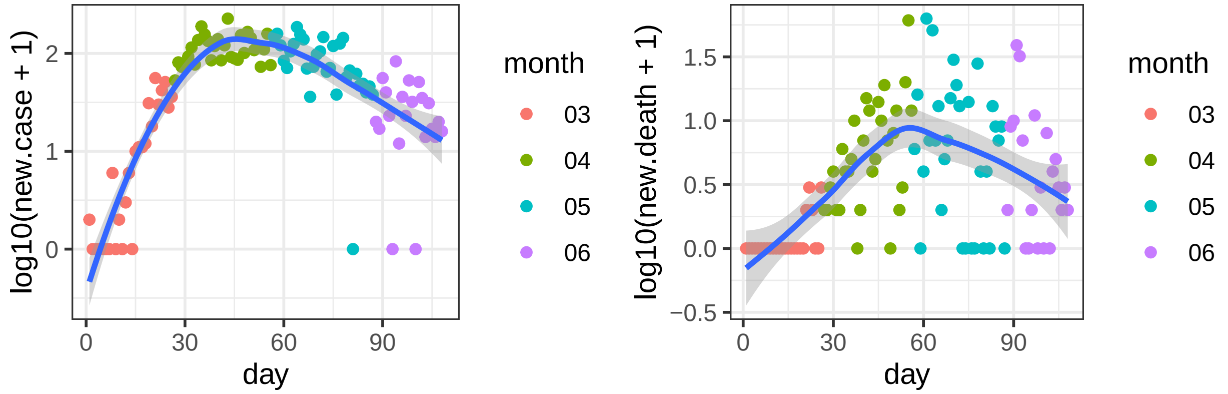
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Monmouth_New Jersey



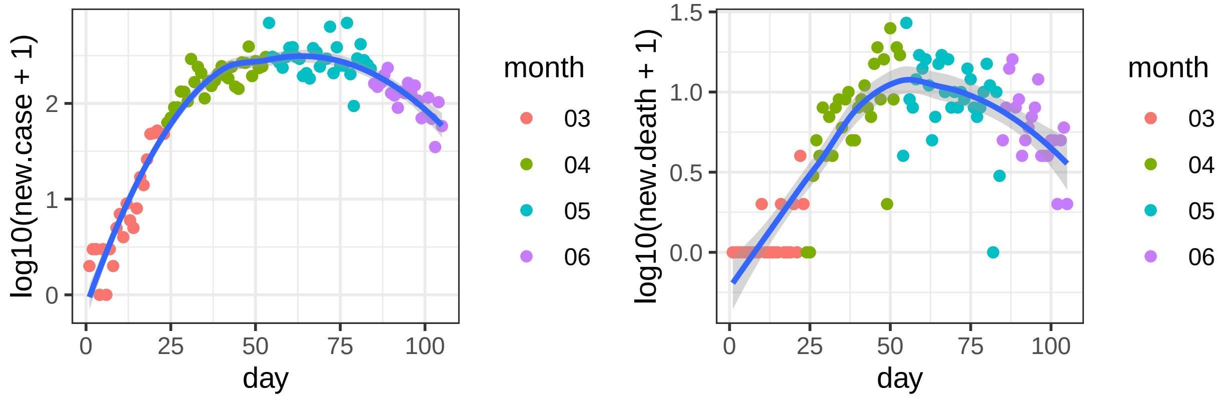
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Delaware_Pennsylvania



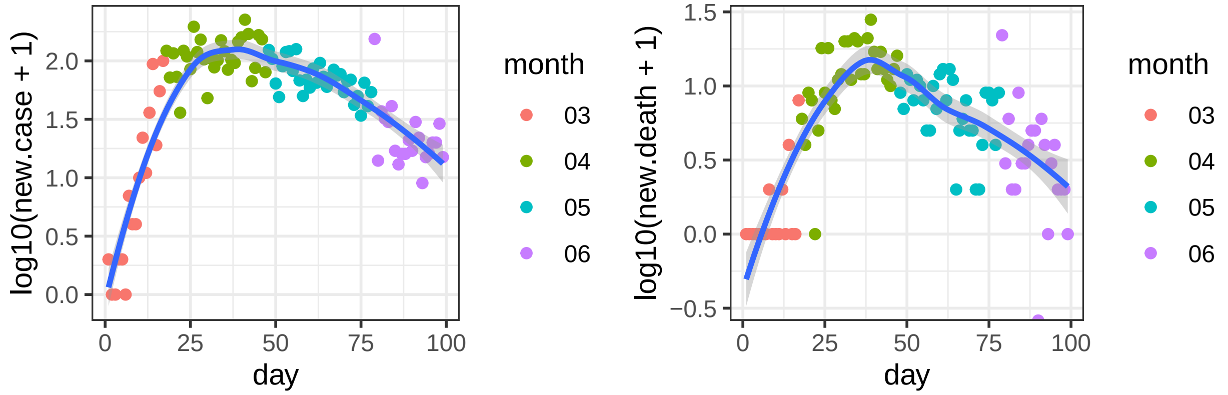
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Prince George's_Maryland



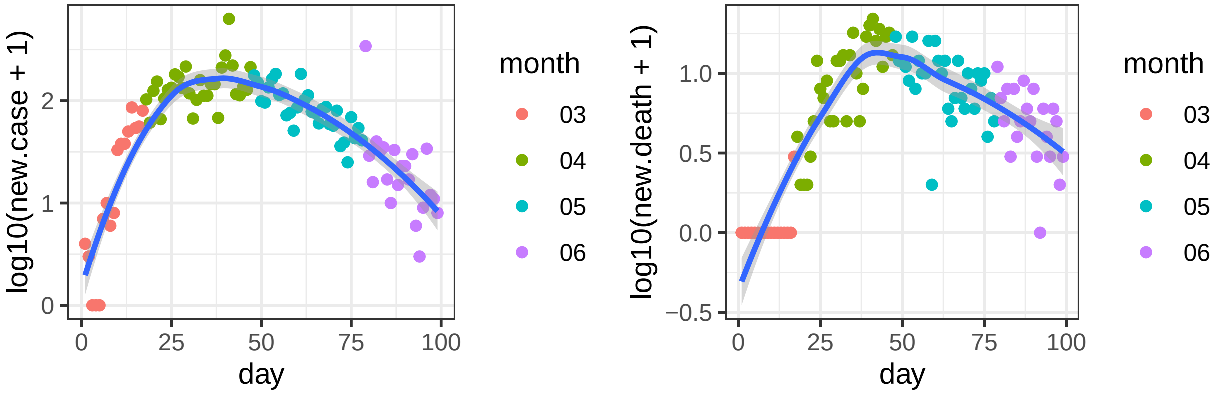
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Hampden_Massachusetts



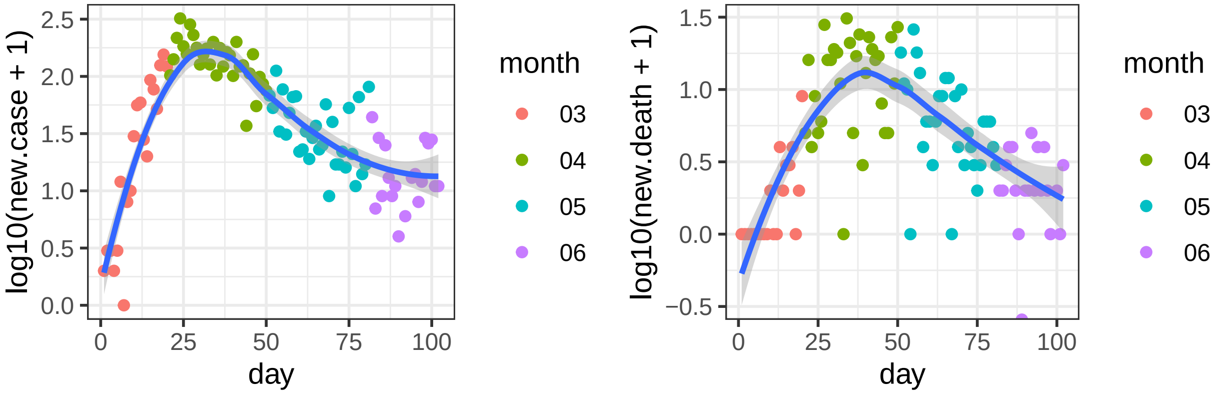
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Plymouth_Massachusetts



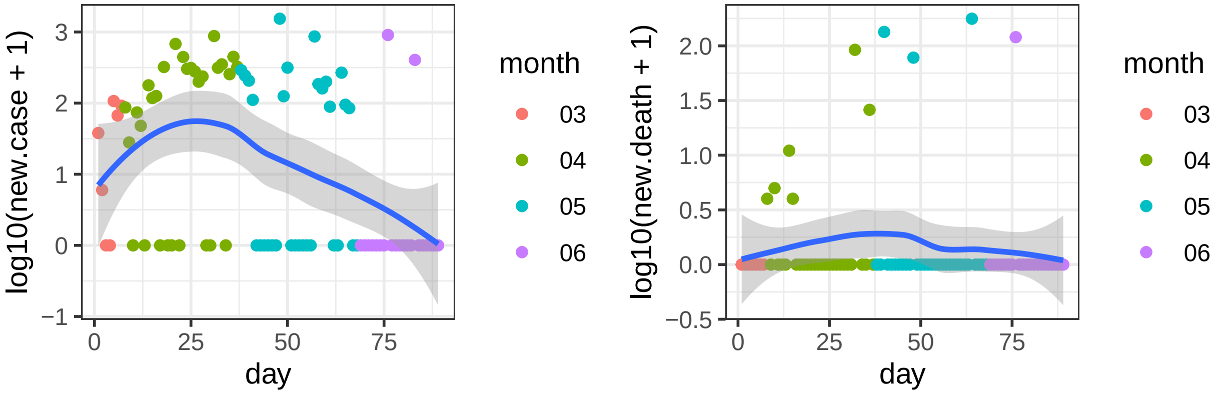
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Morris_New Jersey



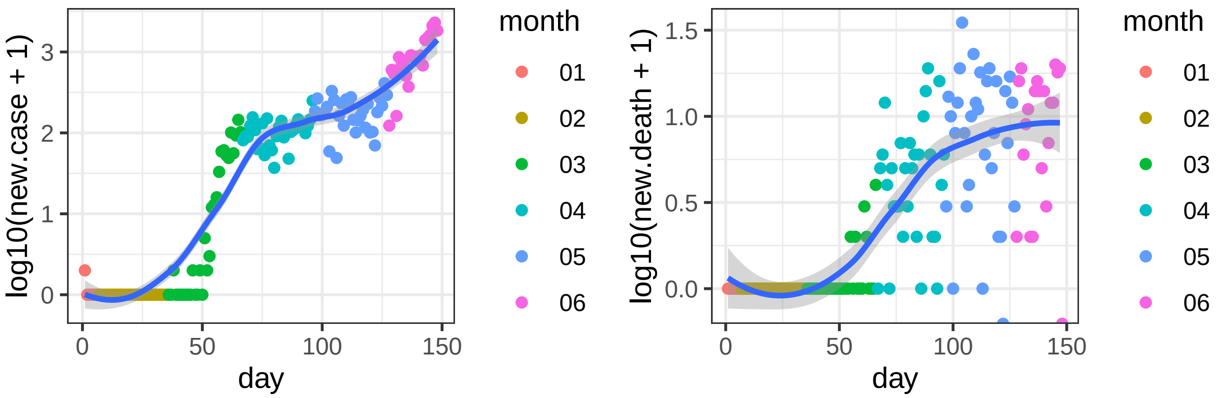
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Providence_Rhode Island



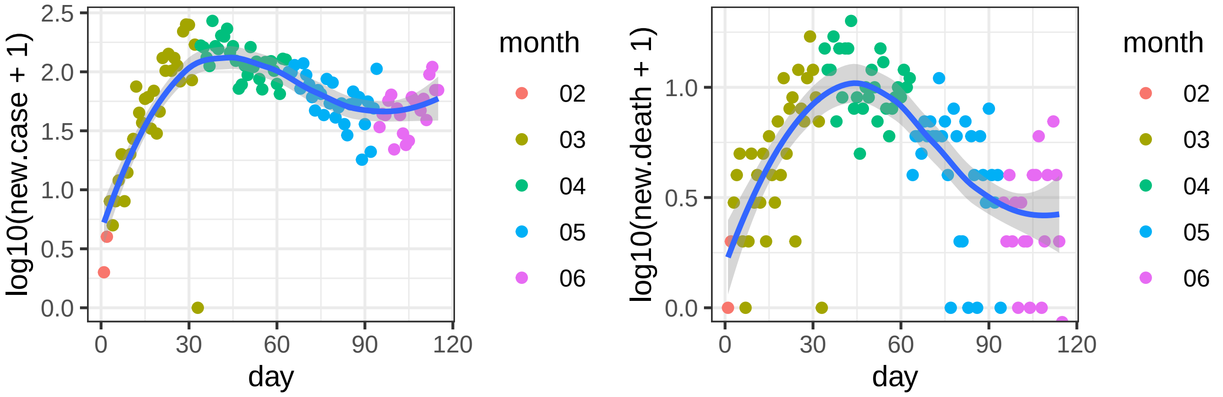
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Maricopa_Arizona



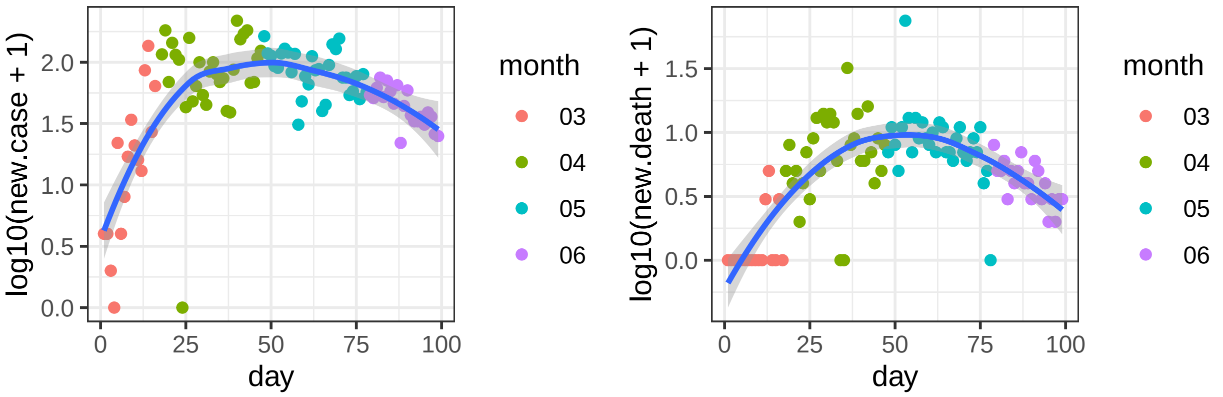
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

King_Washington



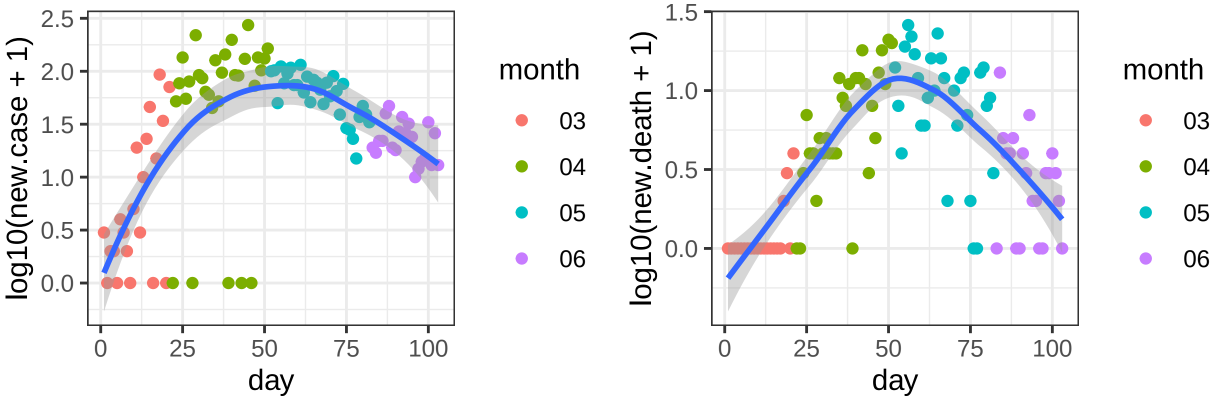
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

Erie_New York



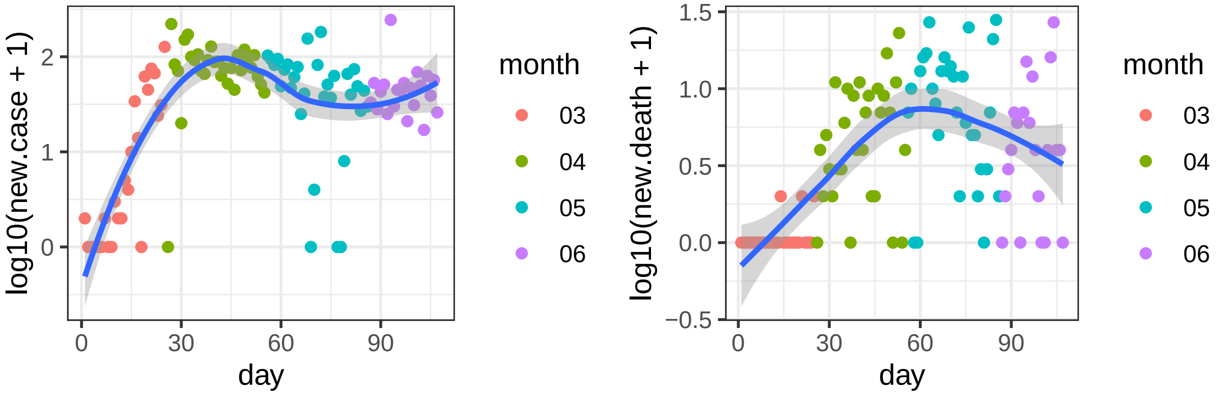
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Bucks_Pennsylvania



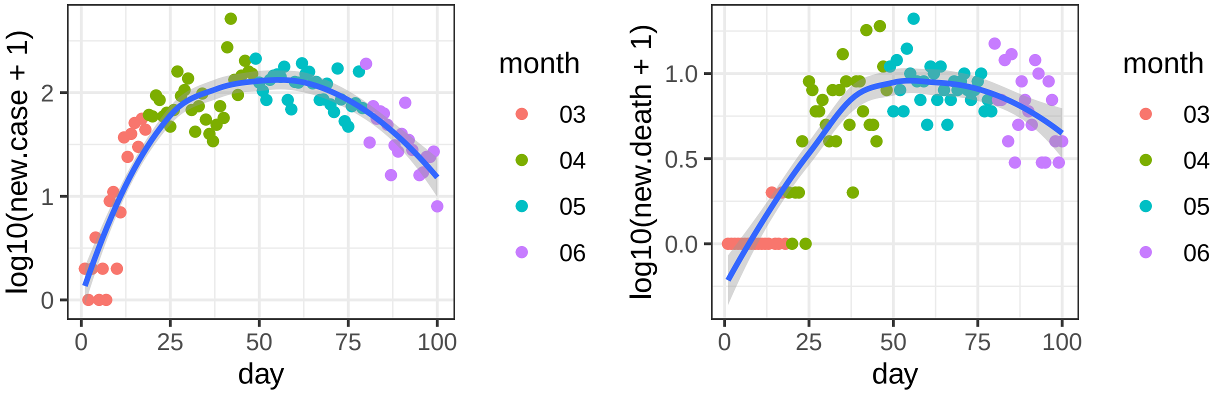
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

St. Louis_Missouri

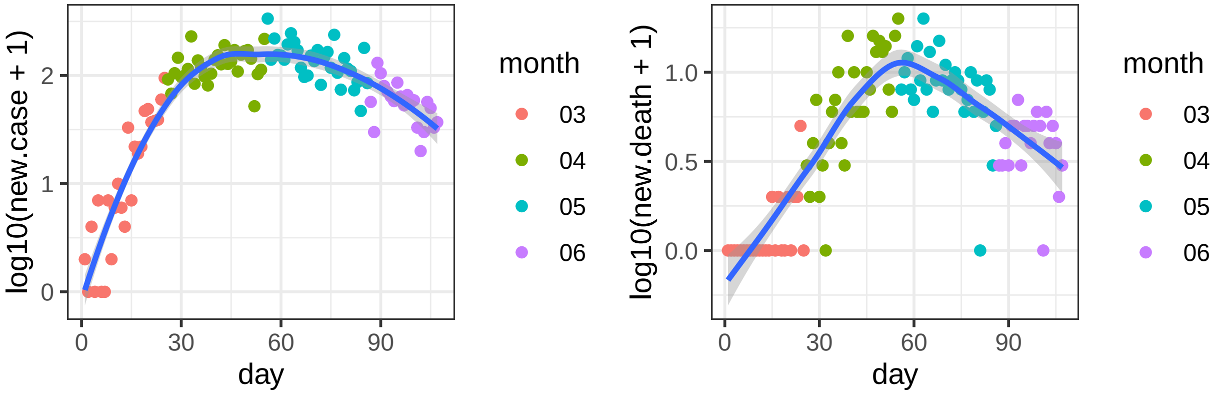


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

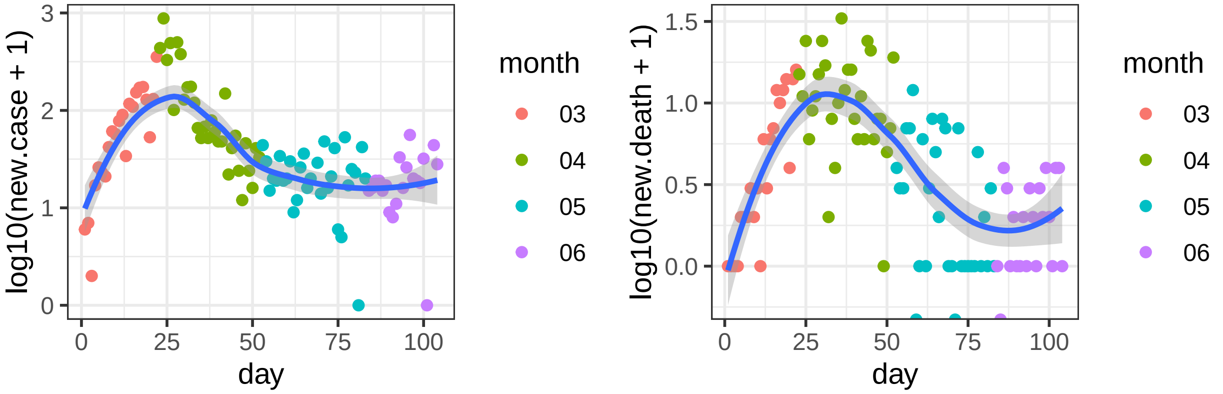
Bristol_Massachusetts



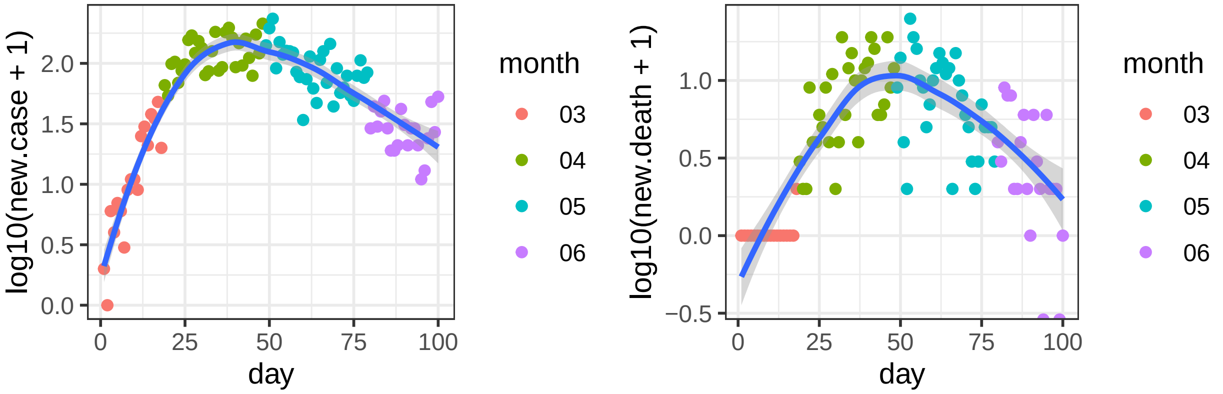
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14


District of Columbia_District of Columbia

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

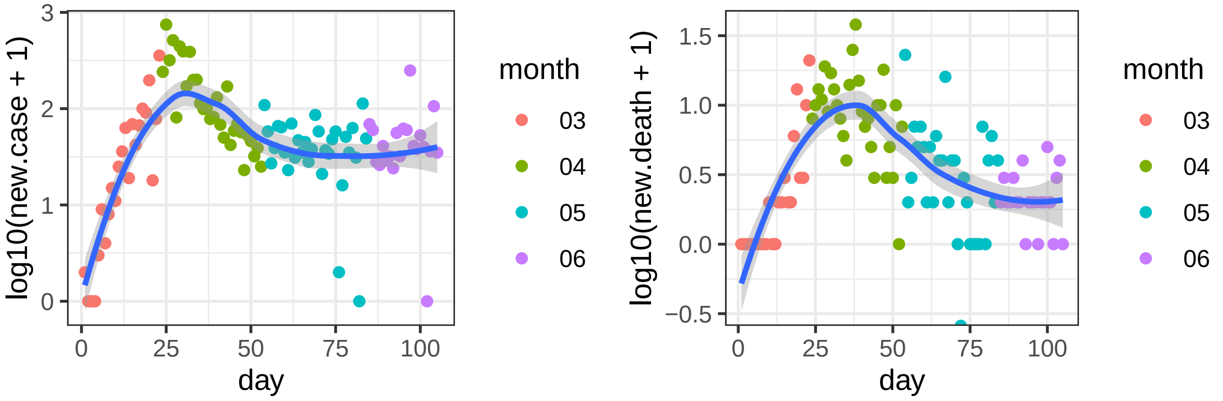

Orleans_Louisiana

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


Mercer_New Jersey

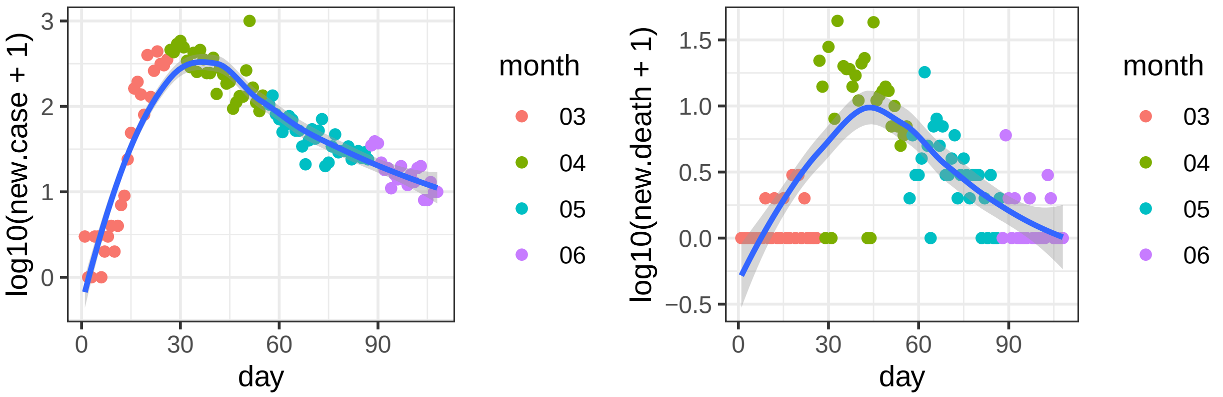
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Jefferson_Louisiana



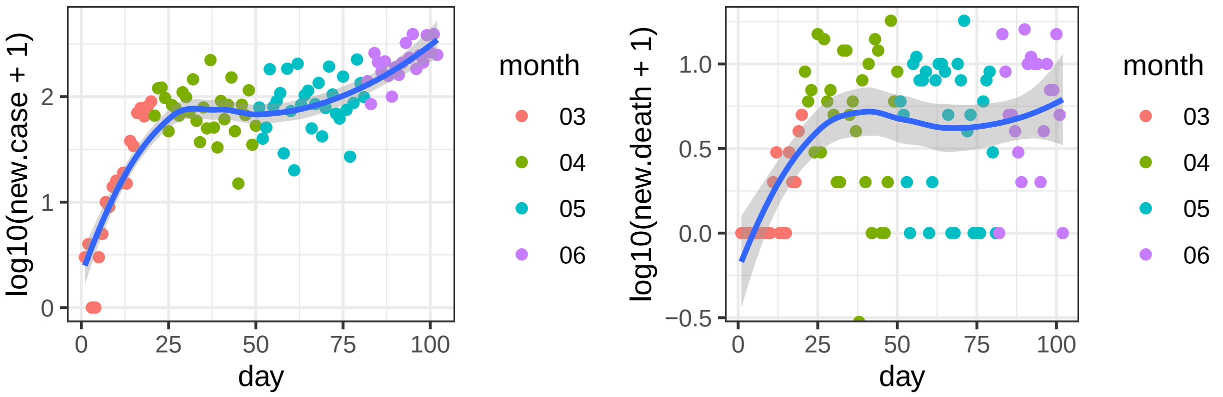
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Rockland_New York

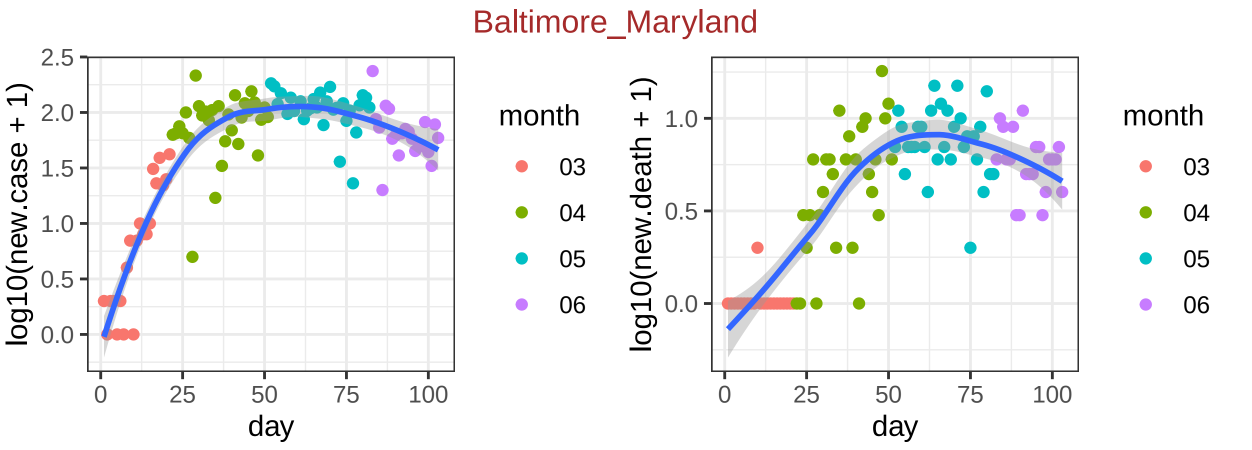


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Palm Beach_Florida



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

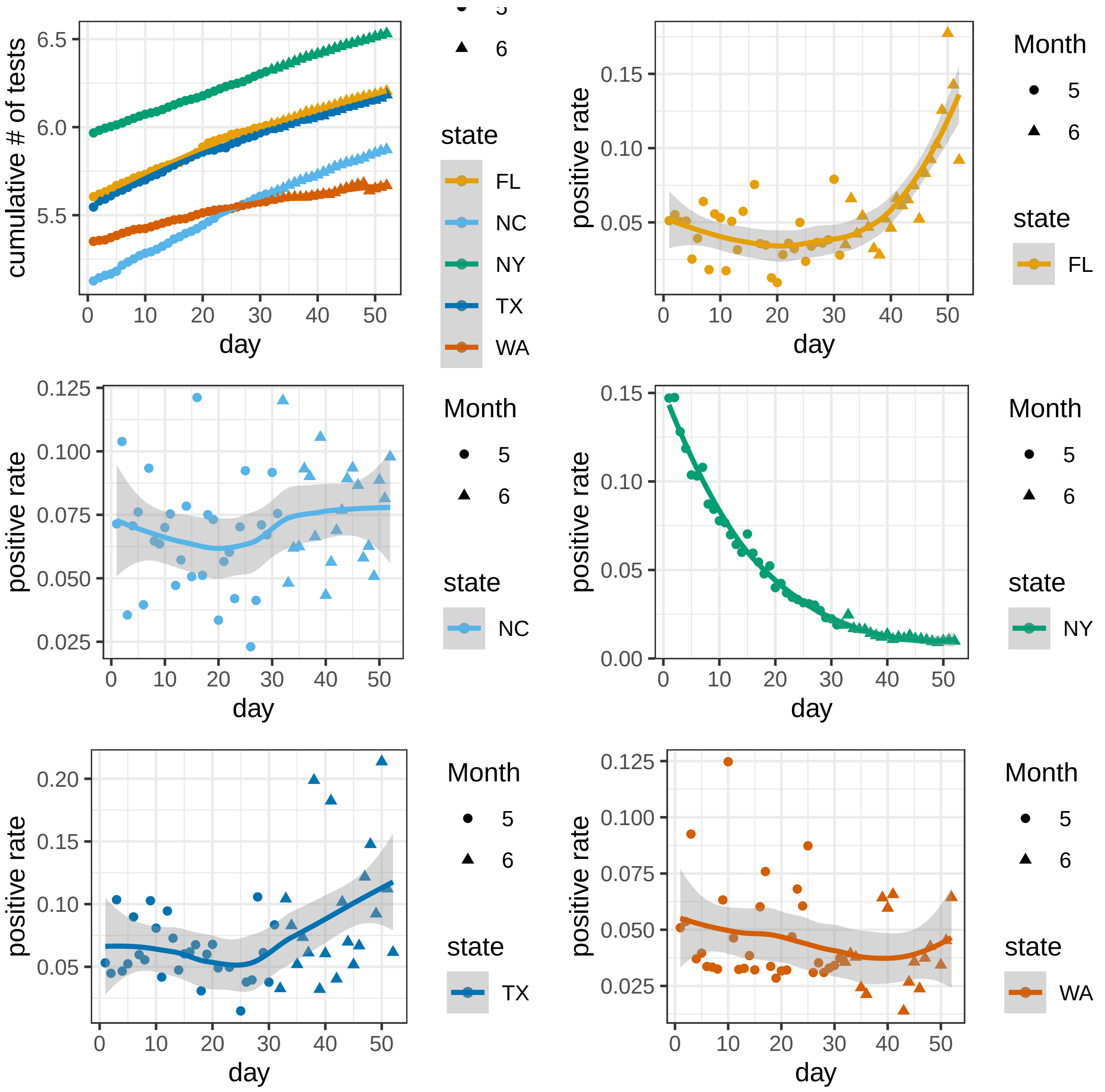


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

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 0621: 0.09(FL) 0.10(NC) 0.01(NY) 0.06(TX) 0.06(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
```