Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-06-16

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 **37**

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

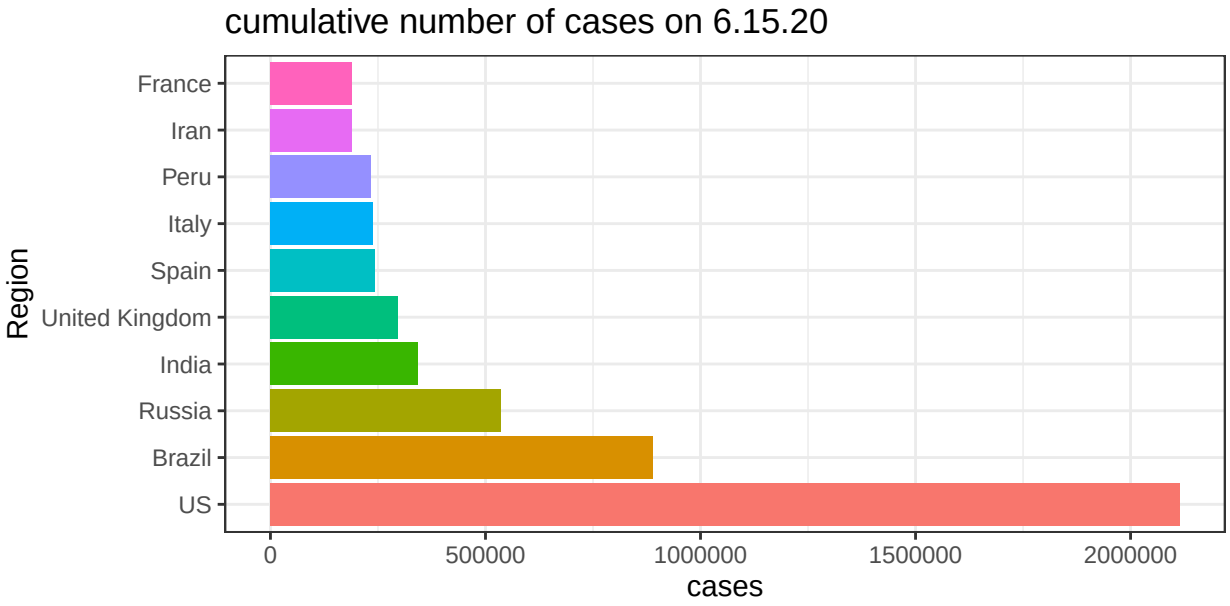
JHU

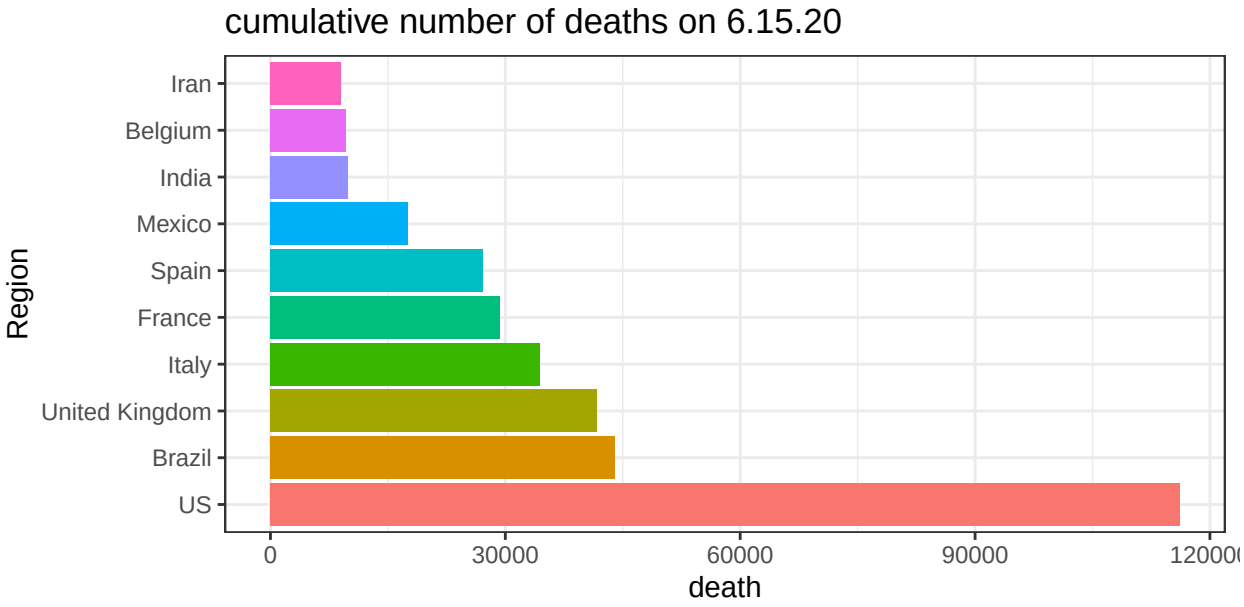
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

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.

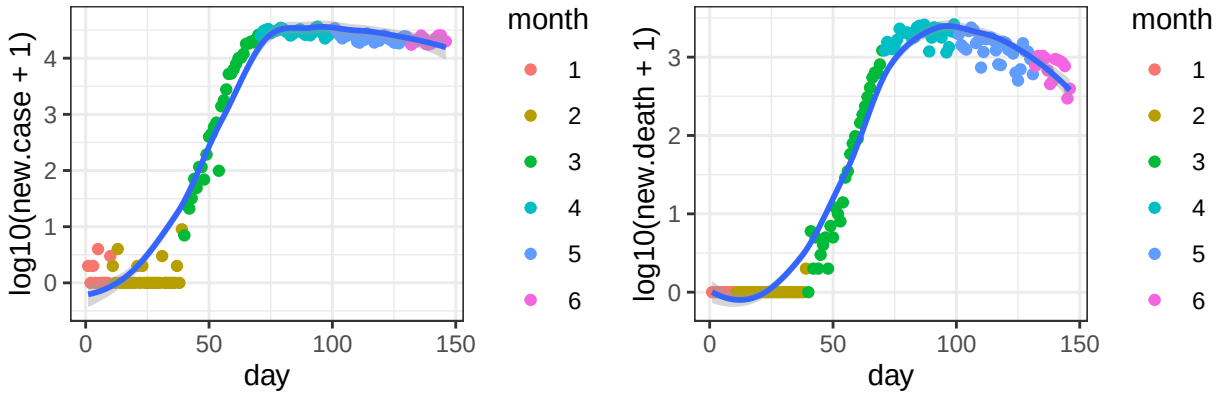
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





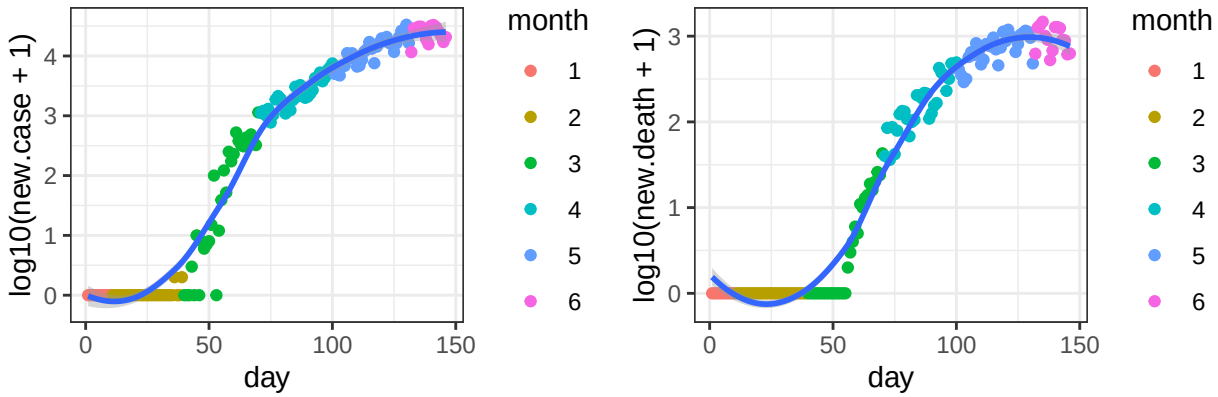
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



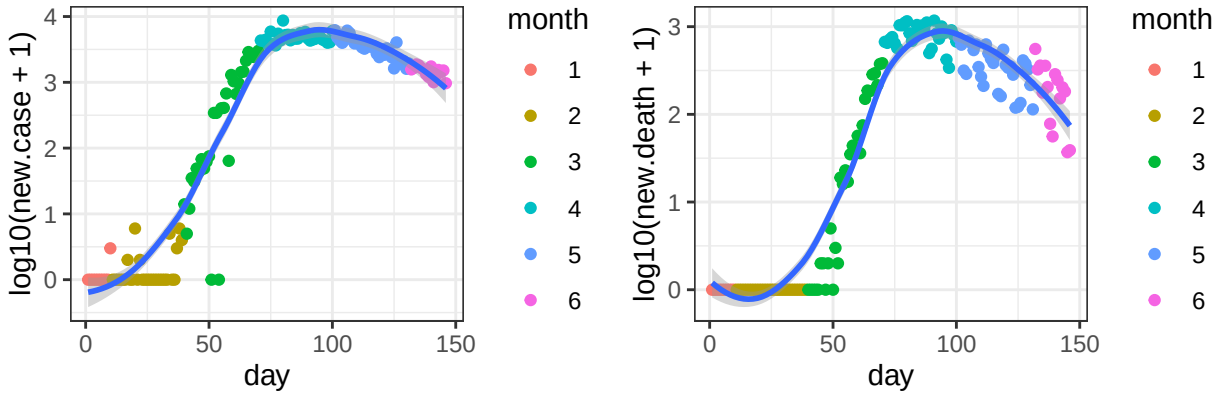
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Brazil



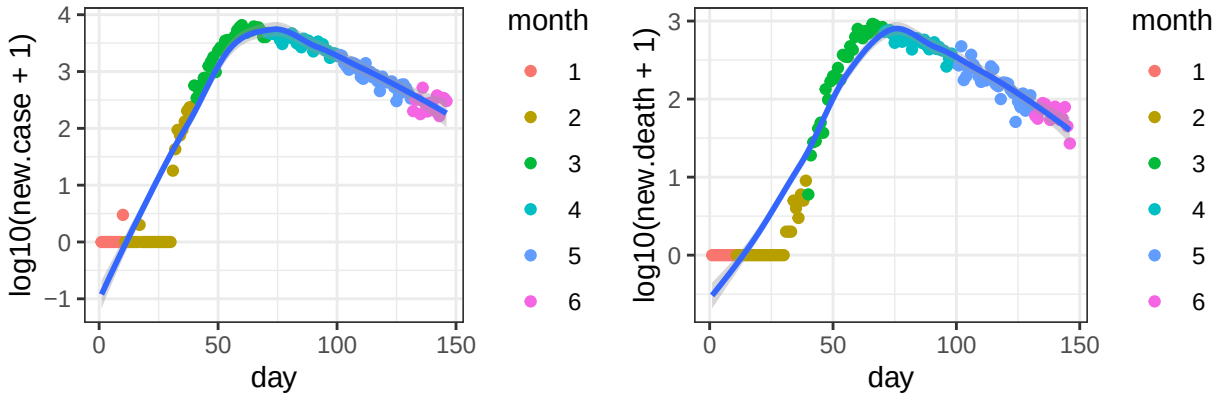
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



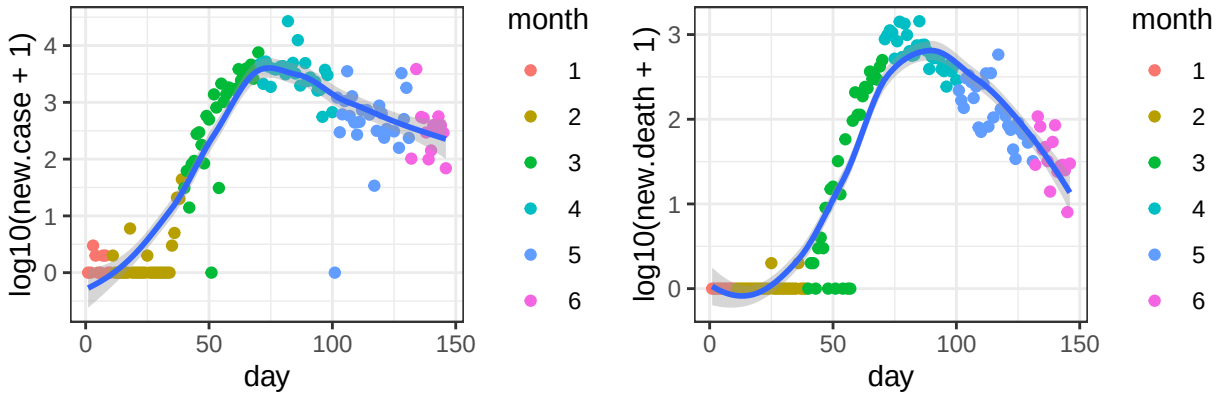
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



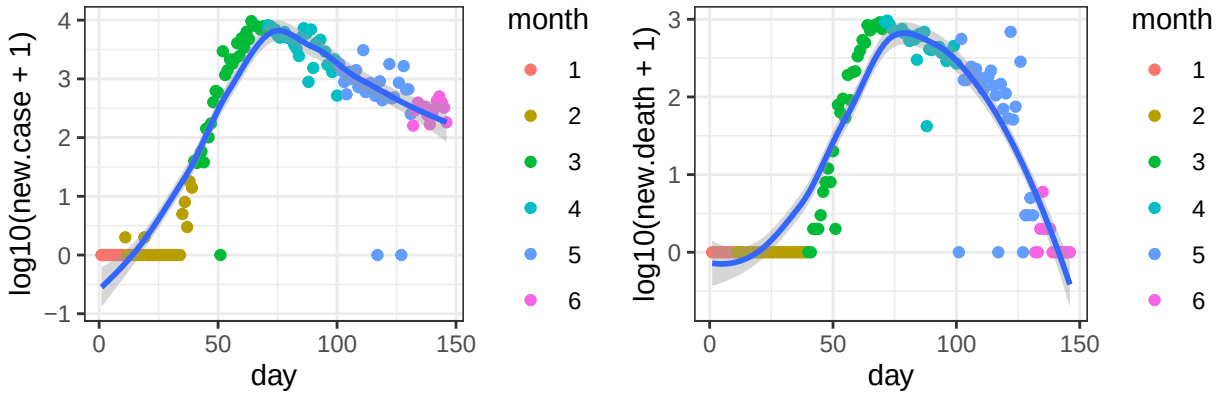
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



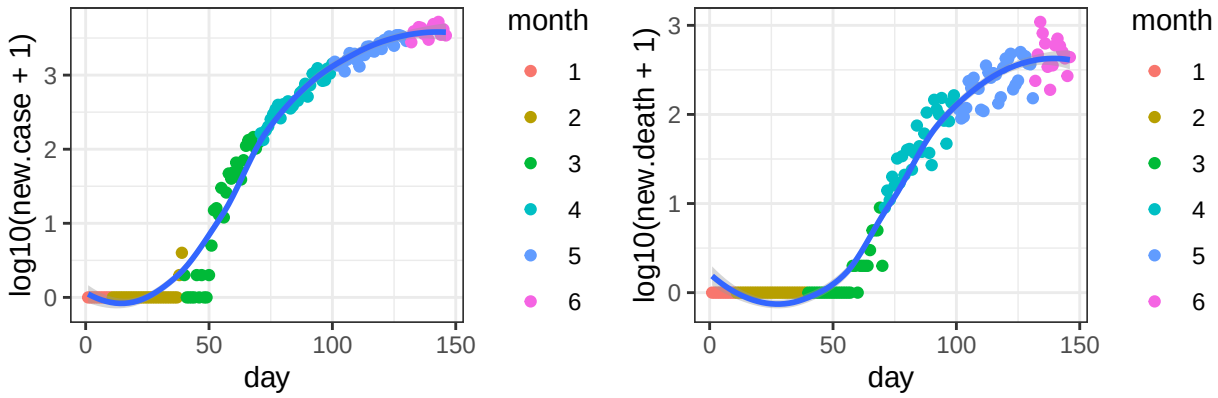
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain



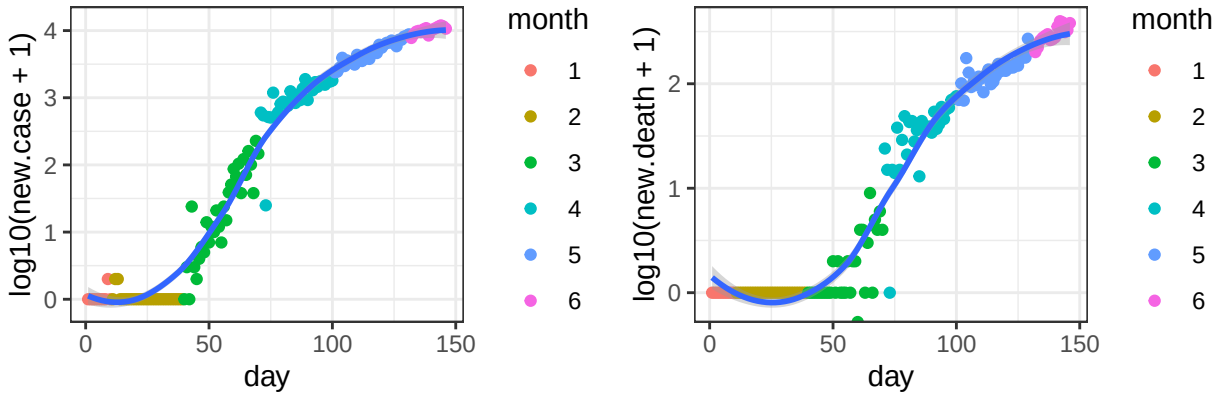
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Mexico

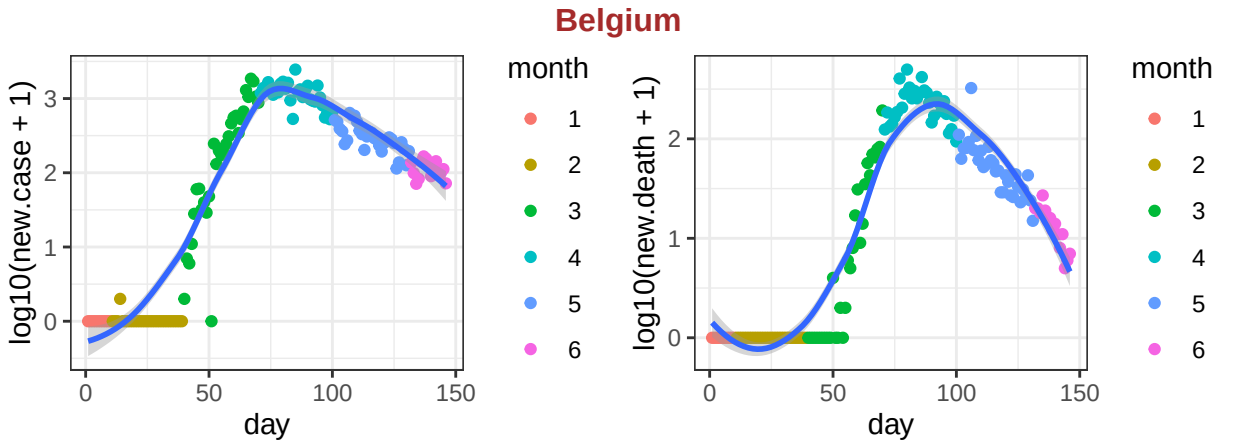


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

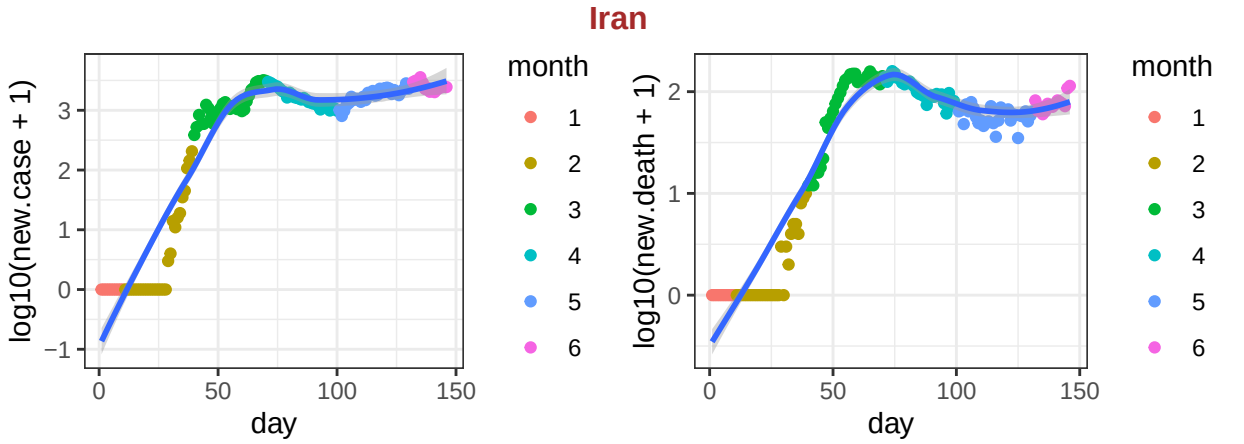
India



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



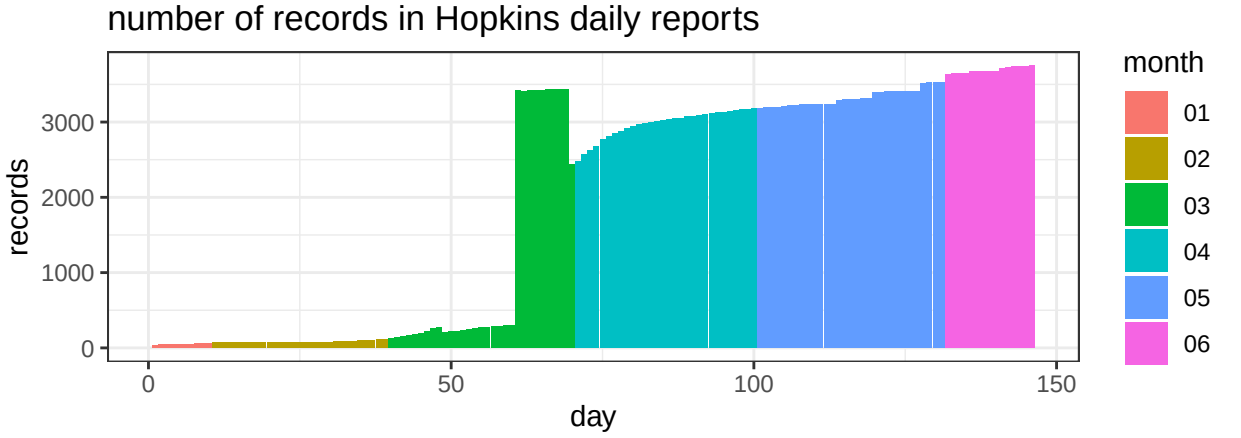
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



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] "2020-06-15"
```

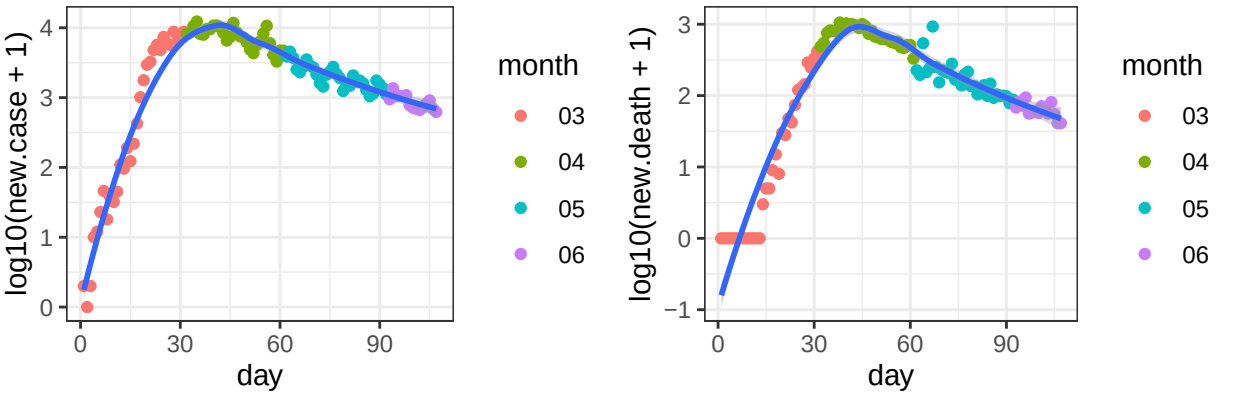
state level data

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

```
##            date                state fips  cases deaths
## 5768 2020-06-15             New York   36 388719  30645
## 5766 2020-06-15           New Jersey   34 167103  12676
## 5757 2020-06-15        Massachusetts   25 105690   7646
## 5749 2020-06-15             Illinois   17 134247   6543
## 5775 2020-06-15         Pennsylvania   42  83687   6307
## 5758 2020-06-15             Michigan   26  66302   6025
## 5739 2020-06-15           California    6 155636   5120
## 5741 2020-06-15          Connecticut    9  45235   4204
## 5754 2020-06-15            Louisiana   22  47284   3018
## 5756 2020-06-15             Maryland   24  62653   2947
## 5744 2020-06-15              Florida   12  77318   2937
## 5772 2020-06-15                 Ohio   39  41576   2573
## 5745 2020-06-15              Georgia   13  55505   2457
## 5750 2020-06-15              Indiana   18  41422   2433
## 5781 2020-06-15                Texas   48  91727   2018
## 5740 2020-06-15             Colorado    8  29284   1605
## 5785 2020-06-15             Virginia   51  54886   1552
## 5759 2020-06-15            Minnesota   27  30724   1335
## 5786 2020-06-15           Washington   53  27577   1223
## 5737 2020-06-15              Arizona    4  37005   1204
## 5769 2020-06-15       North Carolina   37  45257   1156
## 5760 2020-06-15          Mississippi   28  19799    895
## 5761 2020-06-15             Missouri   29  16712    895
## 5777 2020-06-15         Rhode Island   44  16093    851
## 5735 2020-06-15              Alabama    1  26272    774
## 5788 2020-06-15            Wisconsin   55  23072    697
## 5751 2020-06-15                 Iowa   19  24110    661
## 5778 2020-06-15       South Carolina   45  19378    602
## 5753 2020-06-15             Kentucky   21  12863    524
## 5743 2020-06-15 District of Columbia   11   9799    515
```

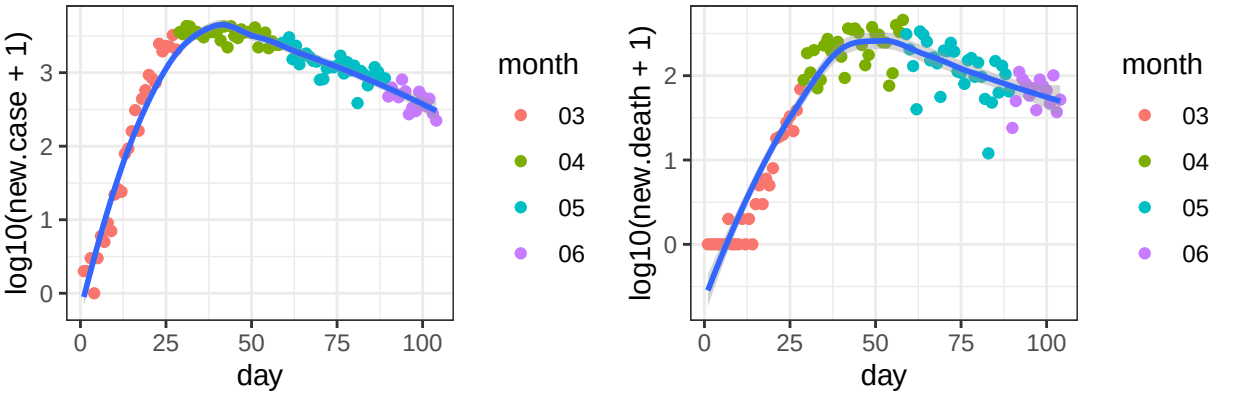
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



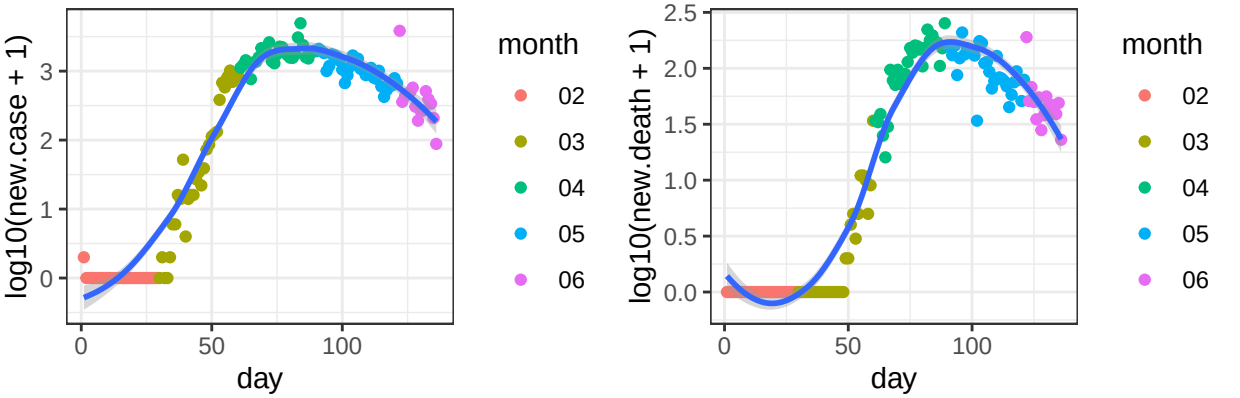
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



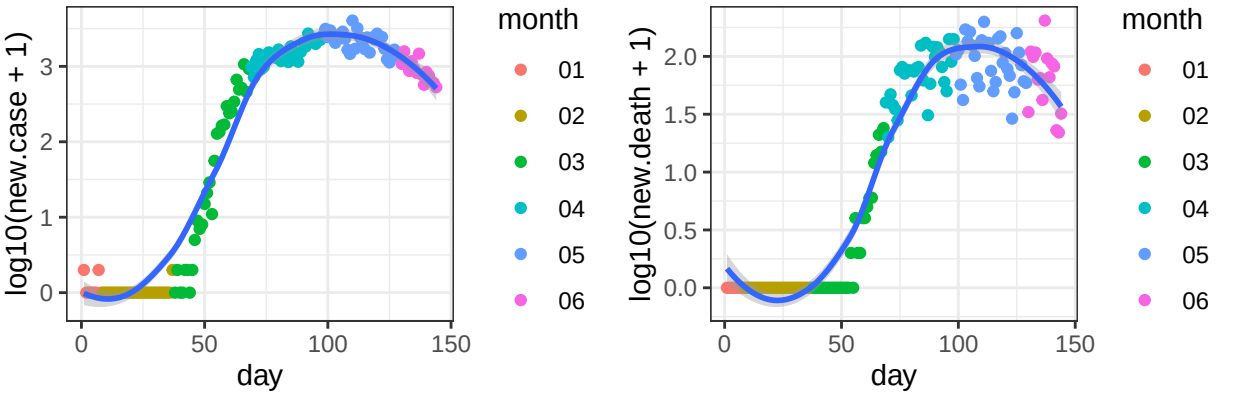
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Massachusetts



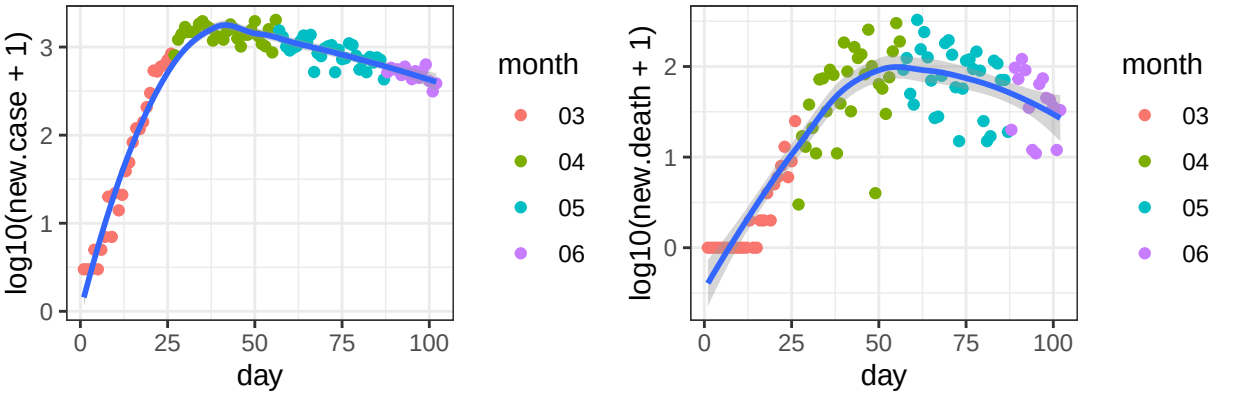
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Illinois



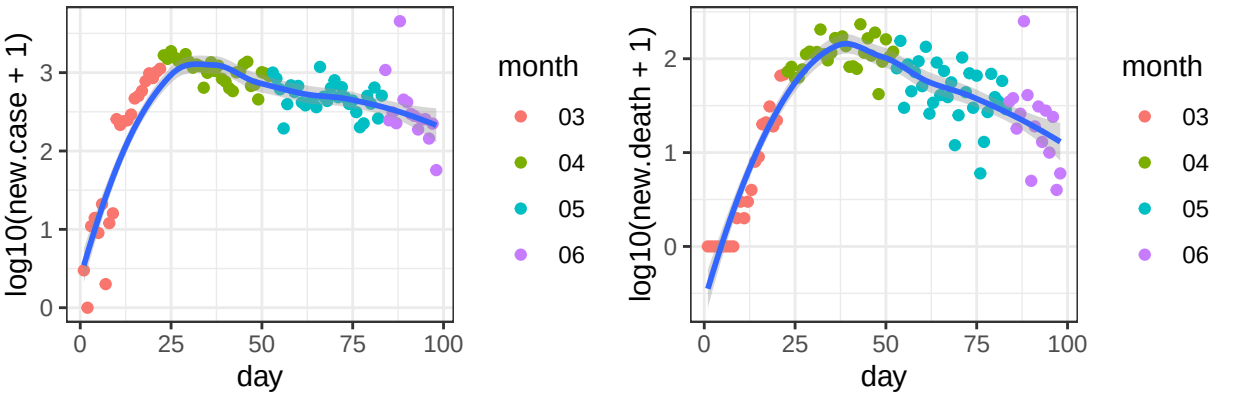
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Pennsylvania



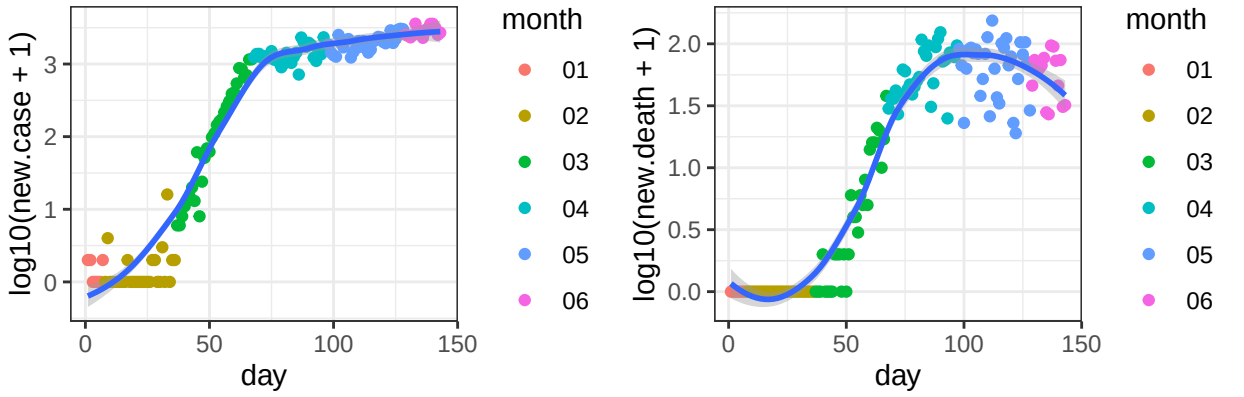
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Michigan



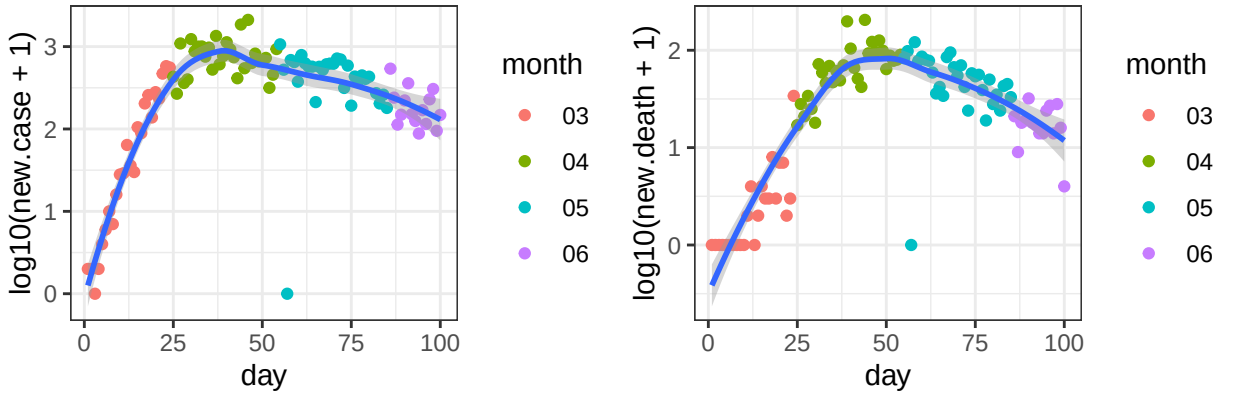
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

California



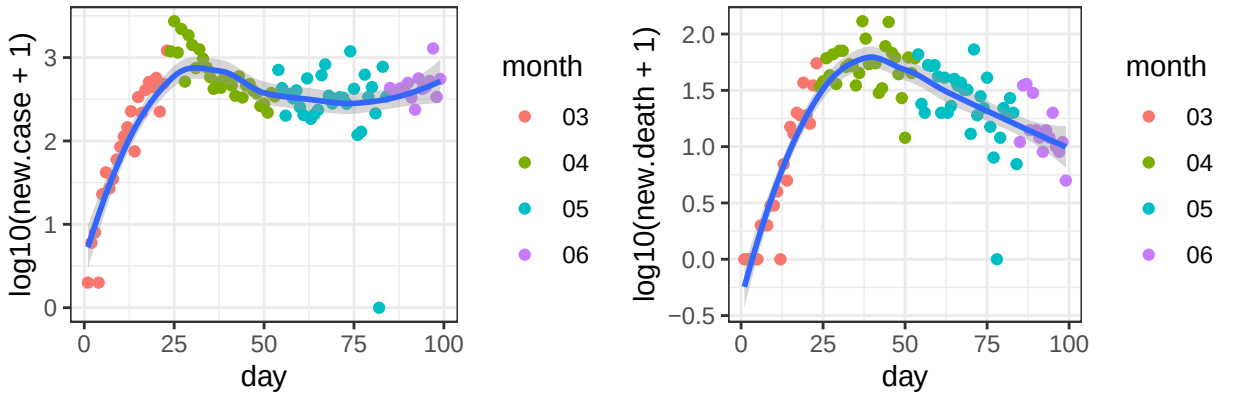
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Connecticut



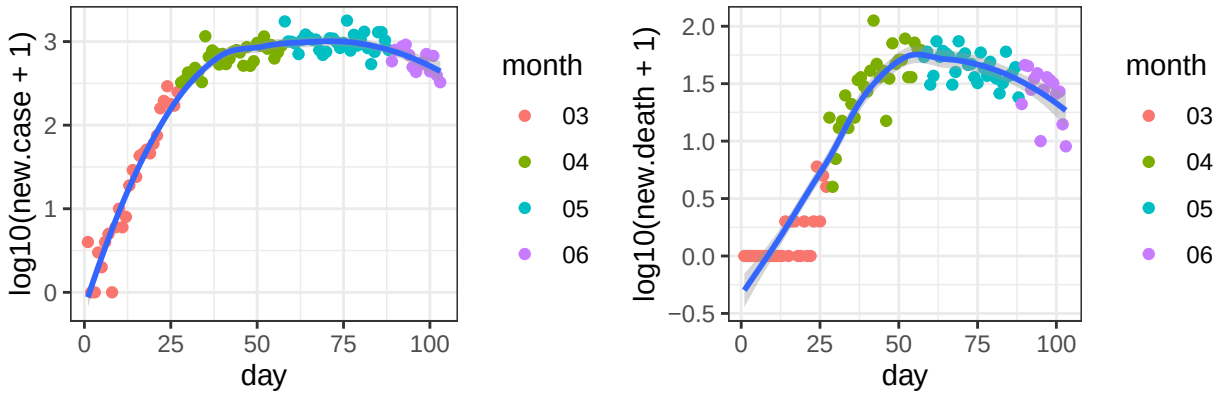
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Louisiana



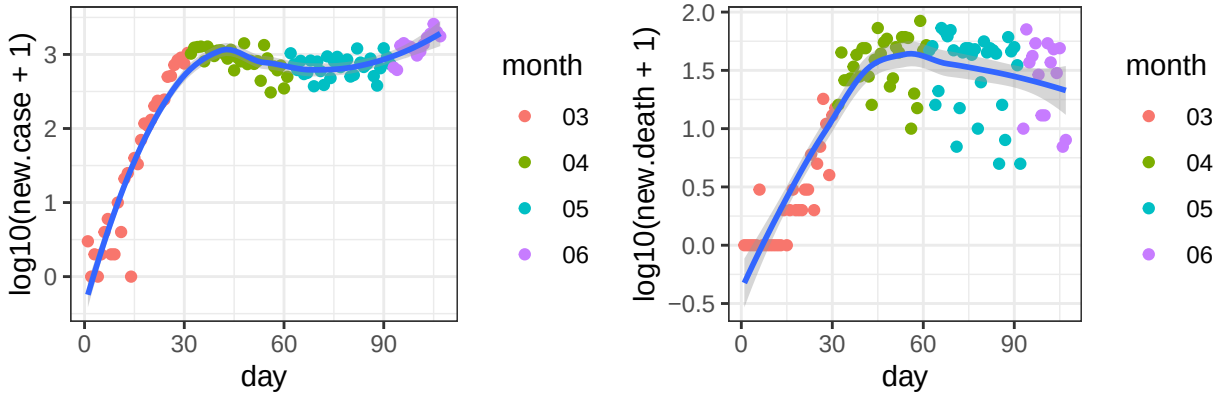
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



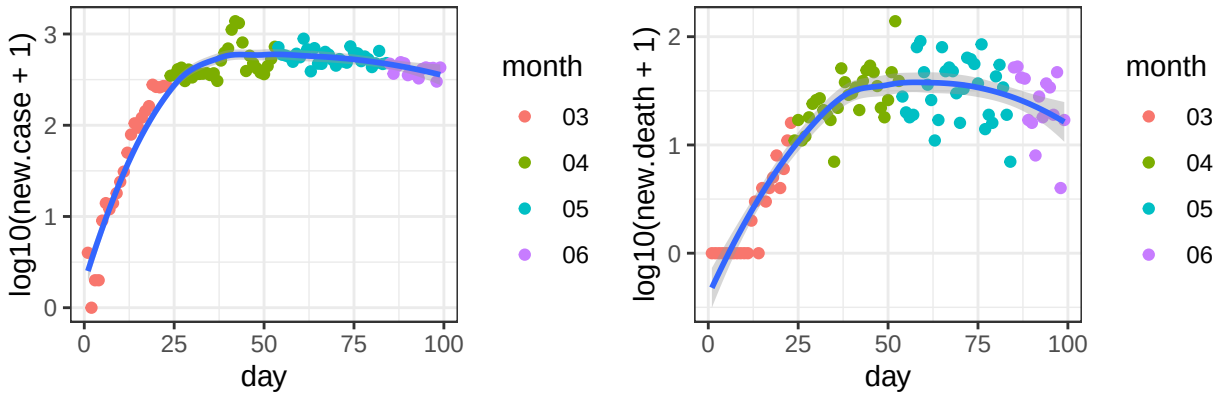
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Florida



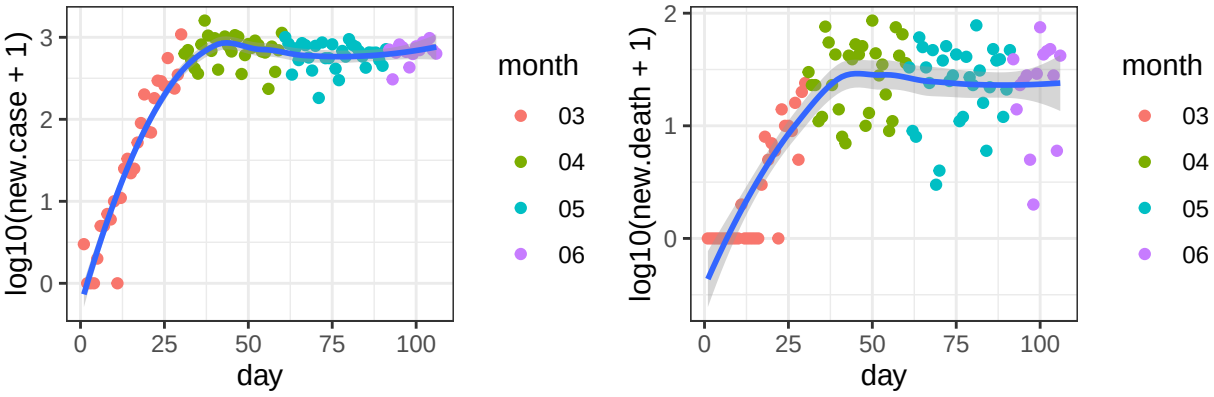
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Ohio



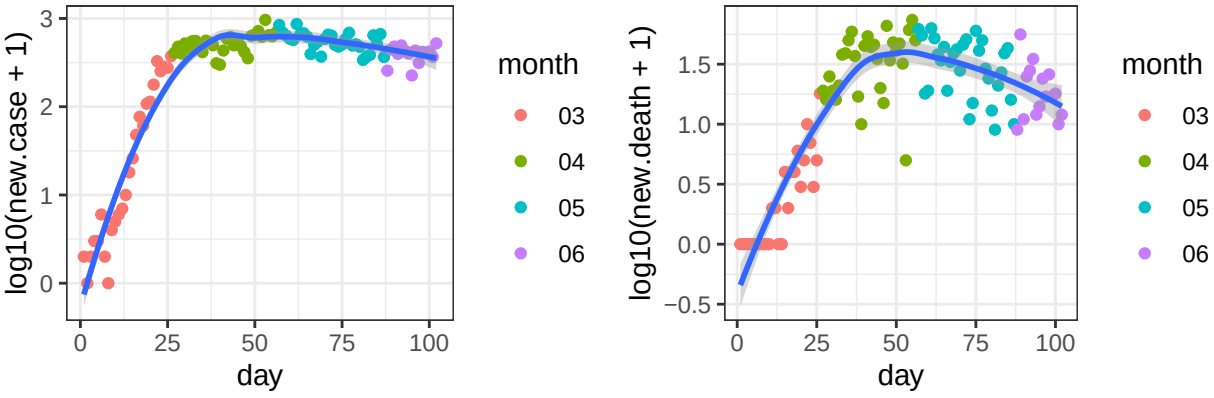
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Georgia



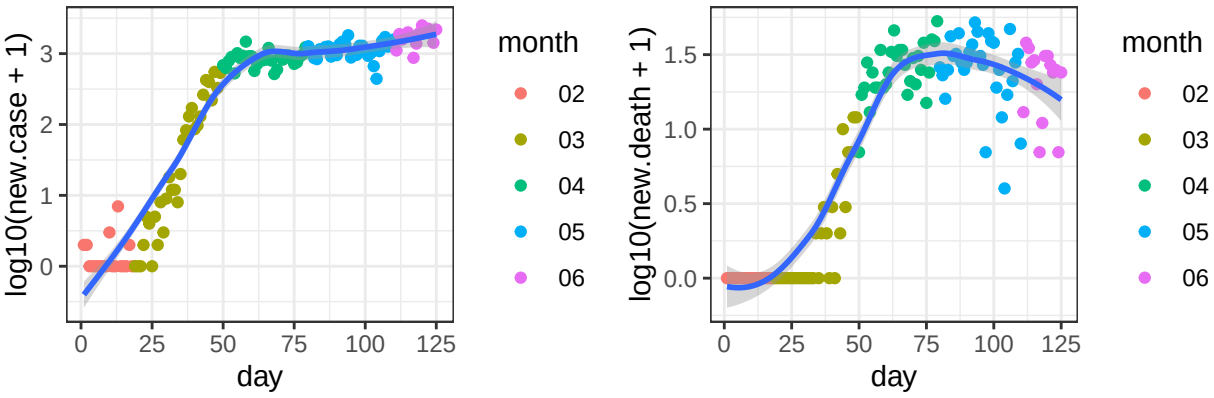
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Indiana



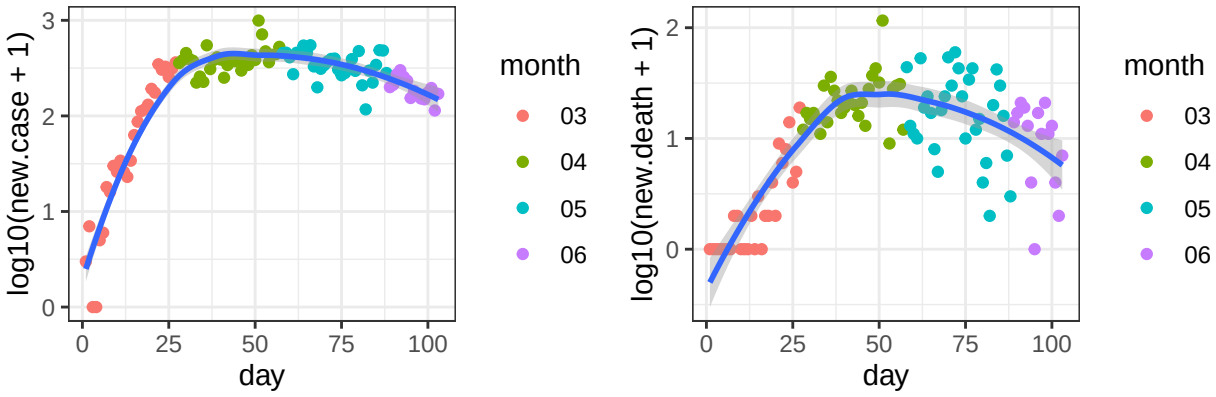
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Texas



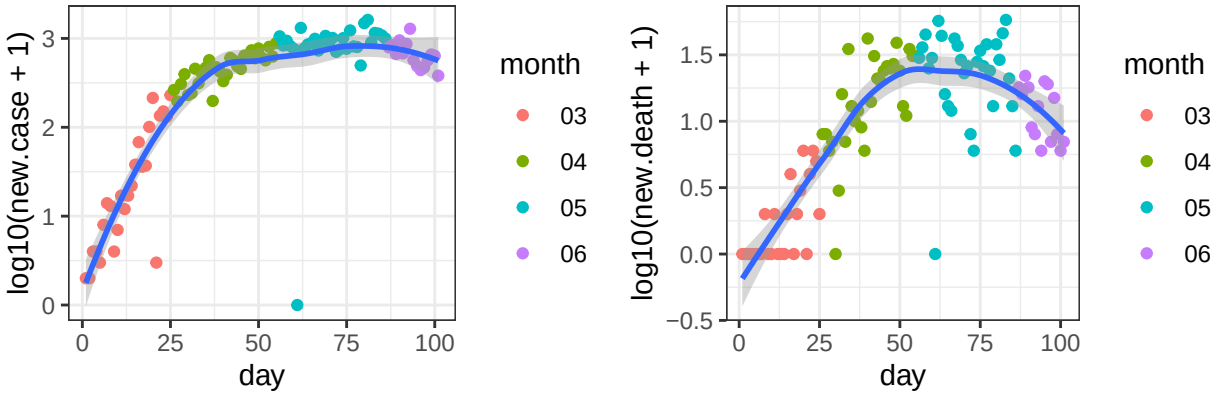
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Colorado



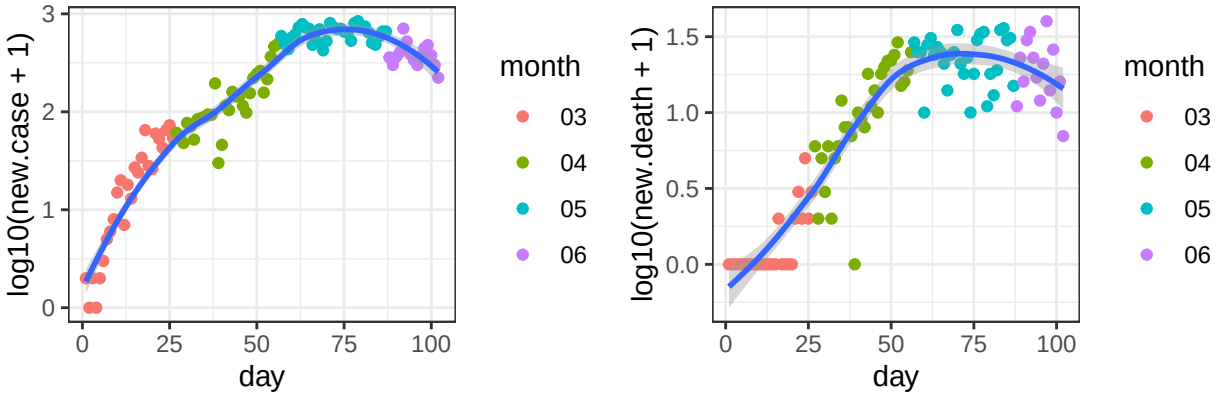
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Virginia



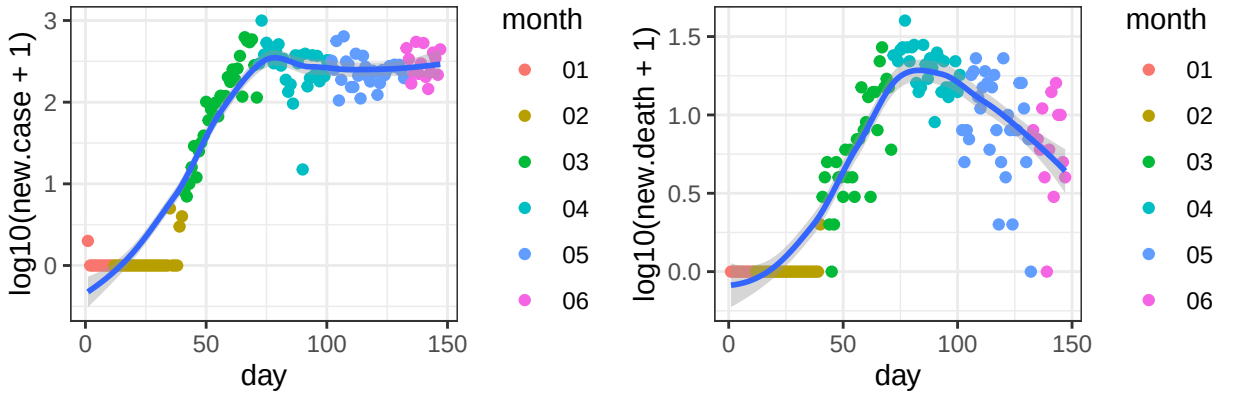
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Minnesota



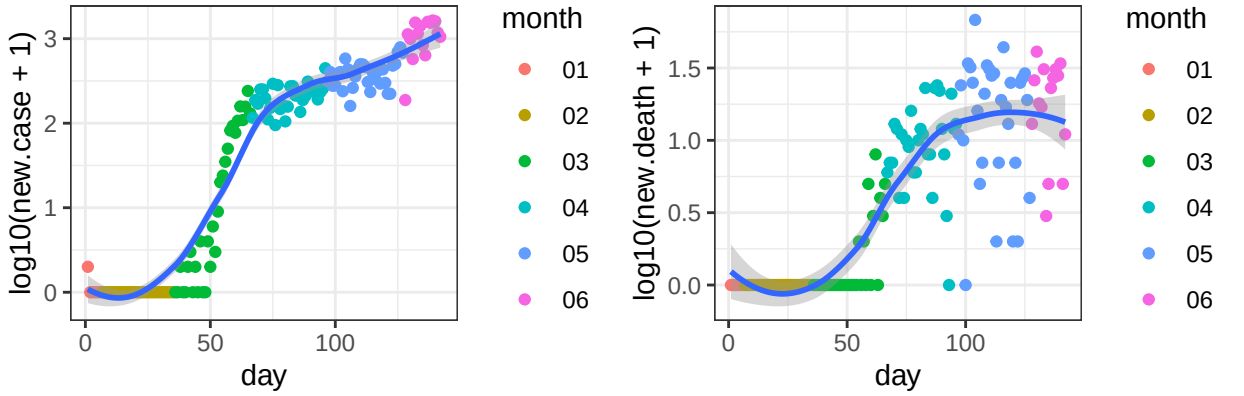
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Washington



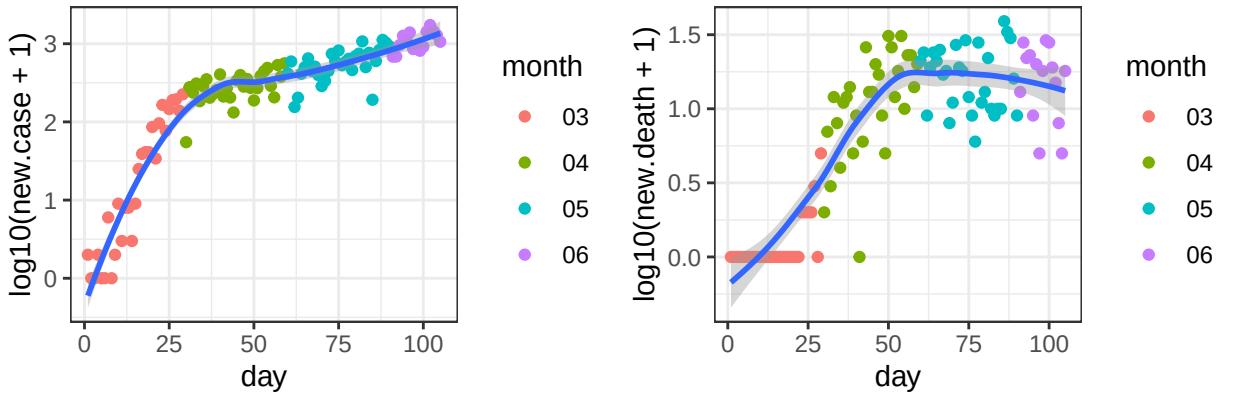
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Arizona



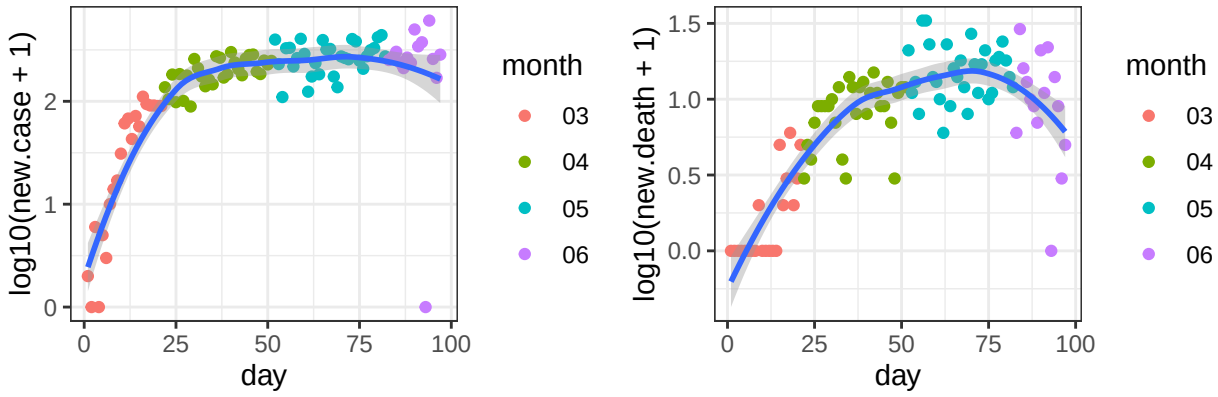
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

North Carolina



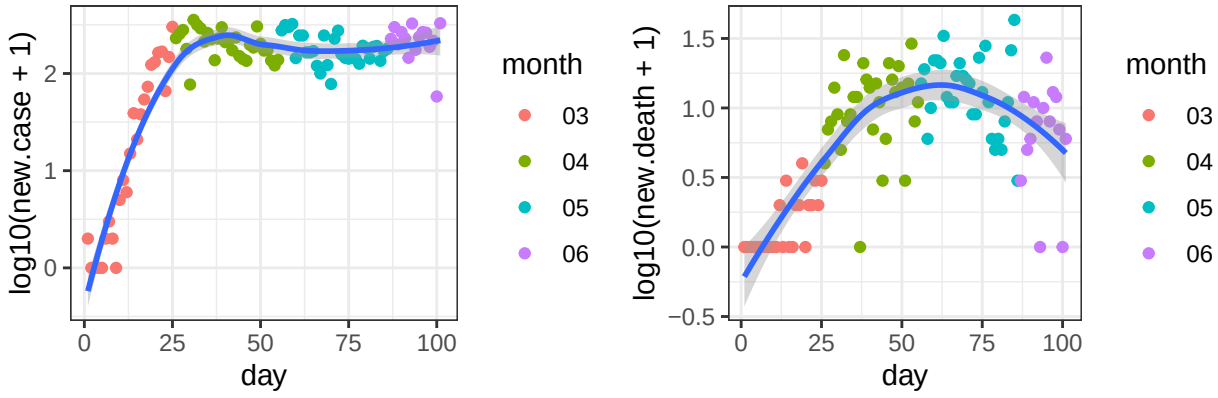
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Mississippi



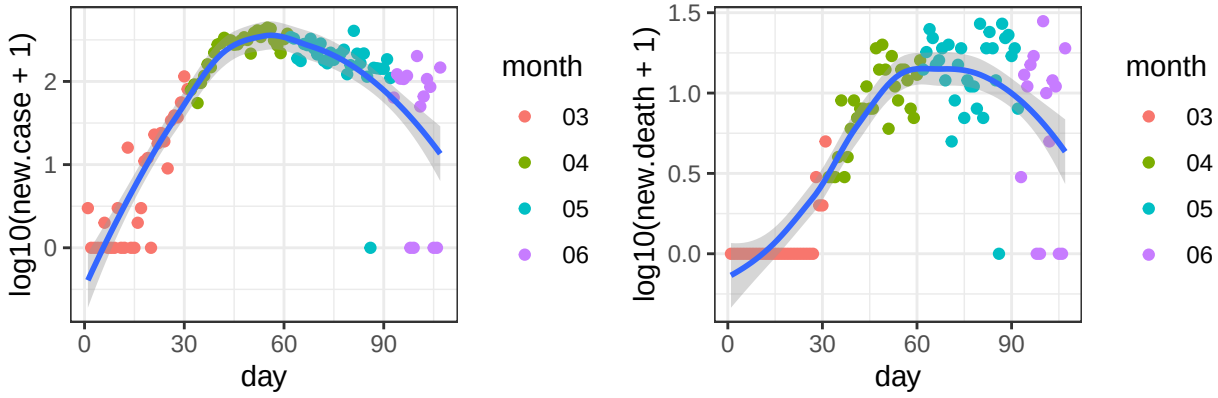
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Missouri



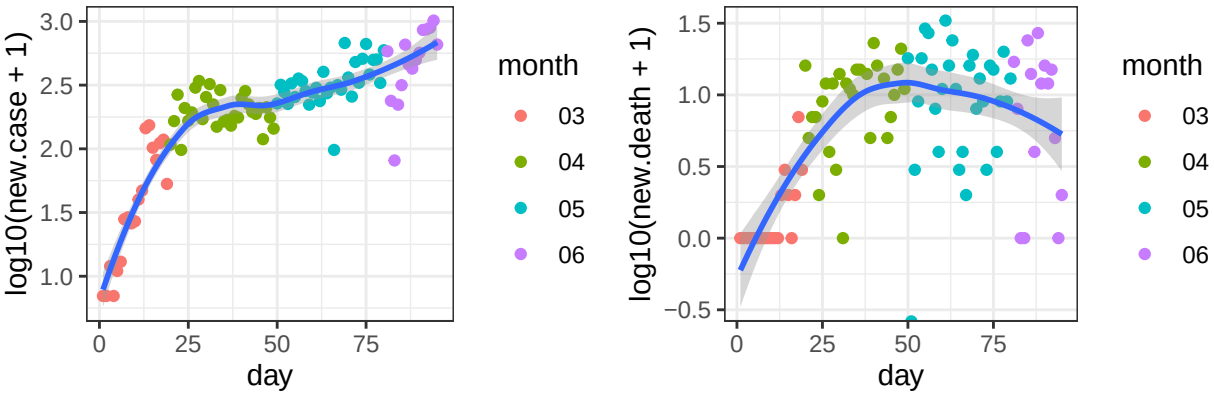
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Rhode Island



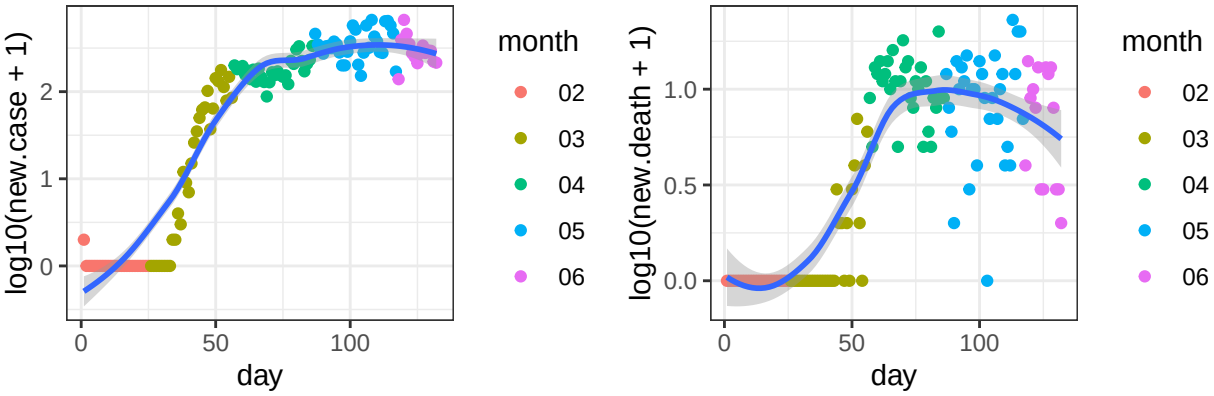
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Alabama



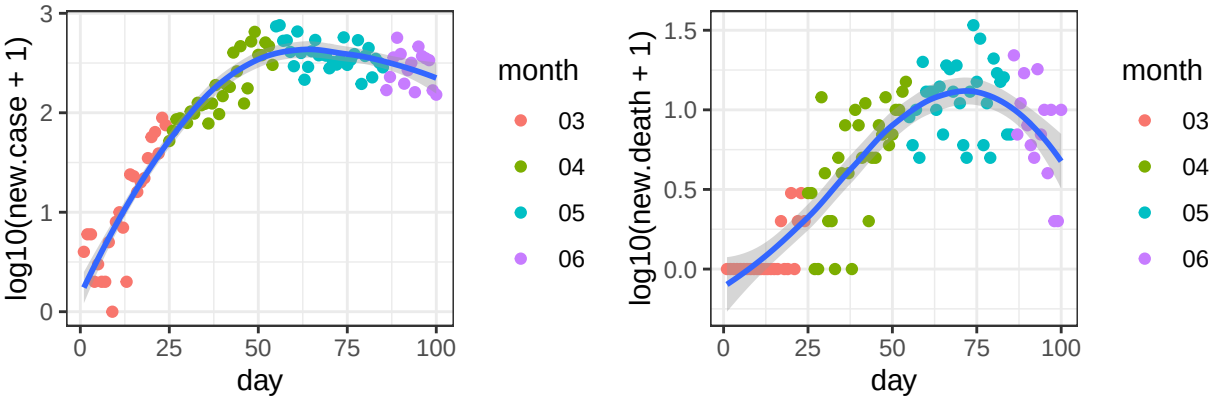
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Wisconsin



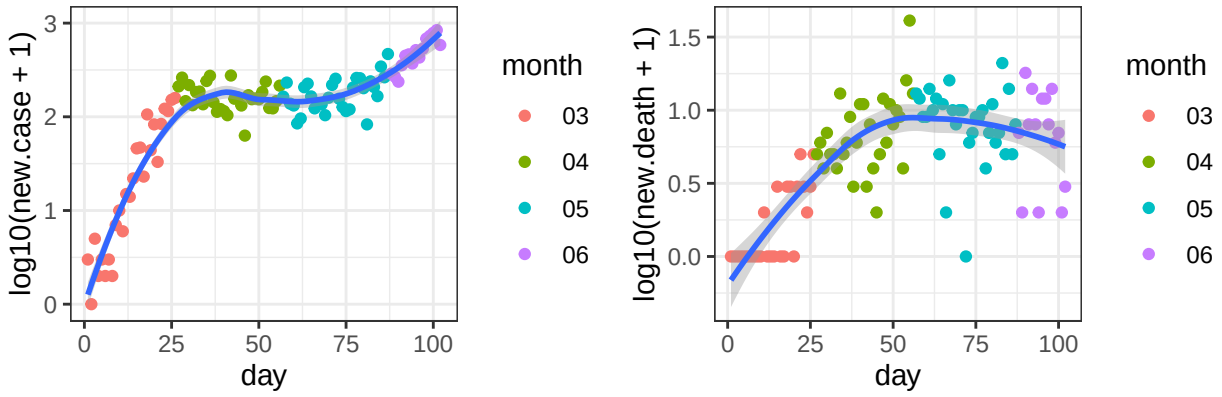
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

Iowa



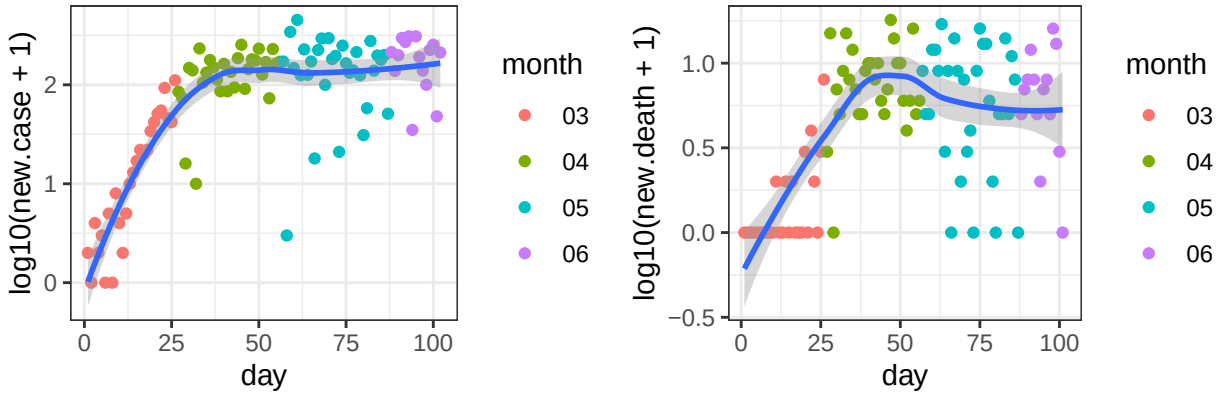
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

South Carolina



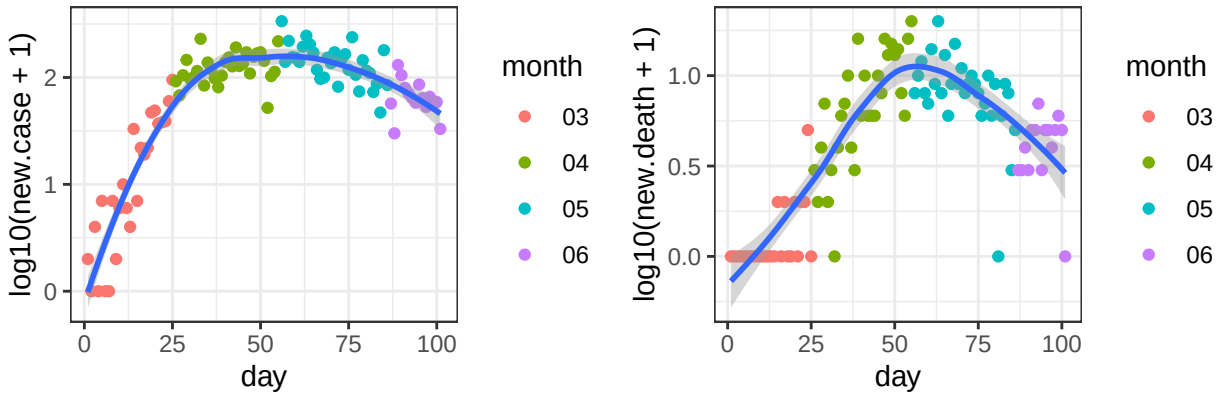
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Kentucky



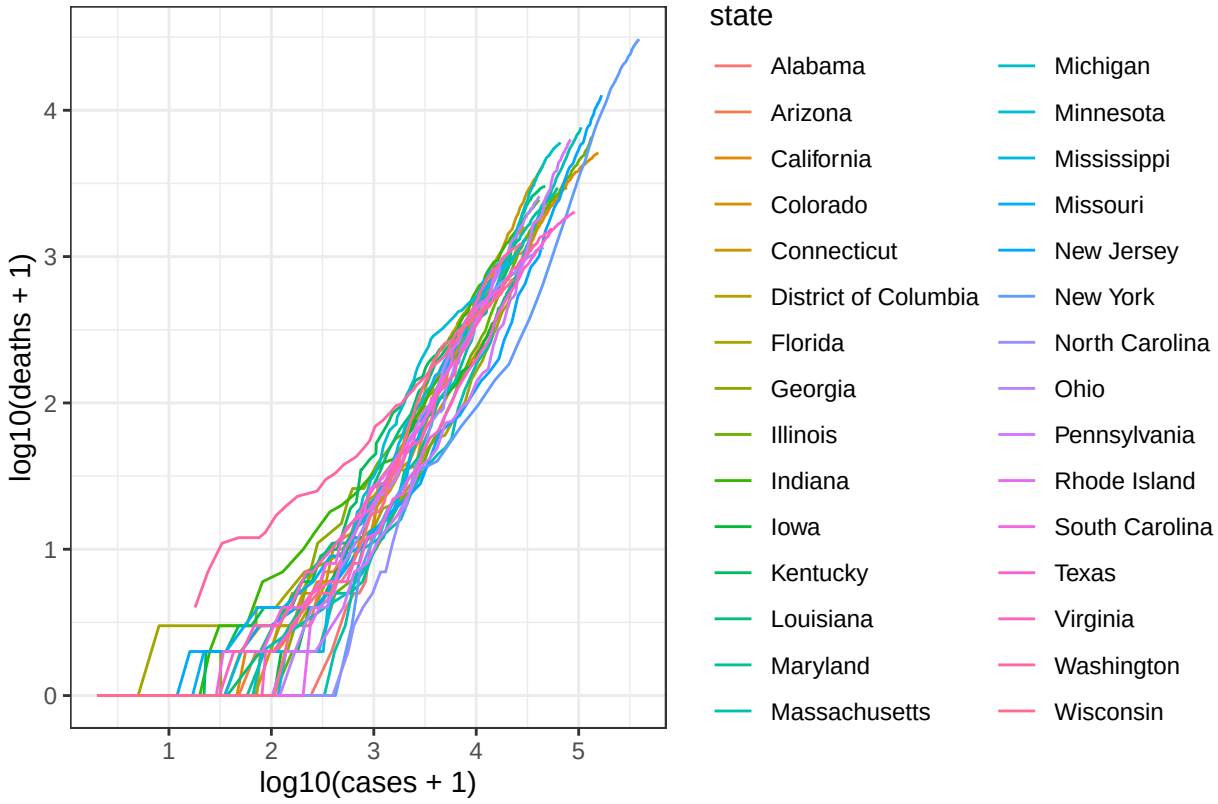
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

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

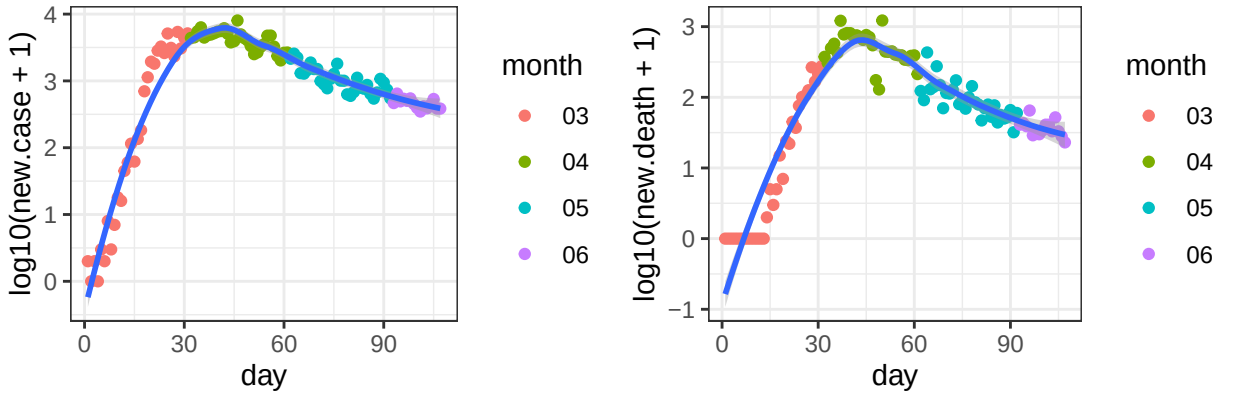
First check the 50 counties with the largest number of deaths.

```
##              date              county               state  fips  cases deaths
## 238550 2020-06-15       New York City            New York    NA 215011  21600
## 237361 2020-06-15                Cook            Illinois 17031  85184   4206
## 236965 2020-06-15         Los Angeles          California  6037  73791   2926
## 238053 2020-06-15               Wayne            Michigan 26163  21816   2672
## 238549 2020-06-15              Nassau            New York 36059  41240   2672
## 238569 2020-06-15             Suffolk            New York 36103  40692   2004
## 237965 2020-06-15           Middlesex       Massachusetts 25017  23227   1763
## 238475 2020-06-15               Essex          New Jersey 34013  18375   1751
## 238470 2020-06-15              Bergen          New Jersey 34003  18848   1664
## 238577 2020-06-15          Westchester            New York 36119  34326   1535
## 238974 2020-06-15        Philadelphia        Pennsylvania 42101  24475   1509
## 237064 2020-06-15           Fairfield         Connecticut  9001  16338   1347
## 237065 2020-06-15            Hartford         Connecticut  9003  11231   1328
## 238477 2020-06-15              Hudson          New Jersey 34017  18765   1256
## 238488 2020-06-15               Union          New Jersey 34039  16308   1128
## 238480 2020-06-15           Middlesex          New Jersey 34023  16458   1083
## 238034 2020-06-15             Oakland            Michigan 26125  11313   1067
## 237961 2020-06-15               Essex       Massachusetts 25009  15627   1051
## 237068 2020-06-15           New Haven         Connecticut  9009  12055   1045
## 238484 2020-06-15             Passaic          New Jersey 34031  16649   1001
## 237969 2020-06-15             Suffolk       Massachusetts 25025  19334    951
## 238021 2020-06-15              Macomb            Michigan 26099   7299    894
```

```
## 237967 2020-06-15              Norfolk        Massachusetts 25021   8872    890
## 237971 2020-06-15            Worcester        Massachusetts 25027  11991    871
## 237120 2020-06-15           Miami-Dade              Florida 12086  22196    826
## 238483 2020-06-15                Ocean           New Jersey 34029   9258    820
## 238969 2020-06-15           Montgomery         Pennsylvania 42091   7930    777
## 238081 2020-06-15             Hennepin            Minnesota 27053  10281    720
## 237496 2020-06-15               Marion              Indiana 18097  10905    697
## 237947 2020-06-15           Montgomery             Maryland 24031  13696    694
## 238481 2020-06-15             Monmouth           New Jersey 34025   8761    678
## 238946 2020-06-15             Delaware         Pennsylvania 42045   6978    671
## 237948 2020-06-15      Prince George's             Maryland 24033  17920    640
## 237963 2020-06-15              Hampden        Massachusetts 25013   6489    640
## 238482 2020-06-15               Morris           New Jersey 34027   6588    639
## 238995 2020-06-15            Providence         Rhode Island 44007  12363    637
## 237968 2020-06-15             Plymouth        Massachusetts 25023   8512    626
## 239625 2020-06-15                 King           Washington 53033   8799    594
## 238535 2020-06-15                 Erie             New York 36029   6817    579
## 236864 2020-06-15              Maricopa              Arizona  4013  19372    557
## 238932 2020-06-15                Bucks         Pennsylvania 42017   5439    542
## 237885 2020-06-15              Orleans            Louisiana 22071   7411    519
## 237959 2020-06-15              Bristol        Massachusetts 25005   7925    518
## 238479 2020-06-15               Mercer           New Jersey 34021   7371    517
## 237077 2020-06-15 District of Columbia District of Columbia 11001   9799    515
## 238322 2020-06-15            St. Louis             Missouri 29189   5604    500
## 237875 2020-06-15             Jefferson            Louisiana 22051   8416    467
## 238561 2020-06-15              Rockland             New York 36087  13441    466
## 238486 2020-06-15              Somerset           New Jersey 34035   4764    437
## 237367 2020-06-15                 DuPage             Illinois 17043   8465    431
```

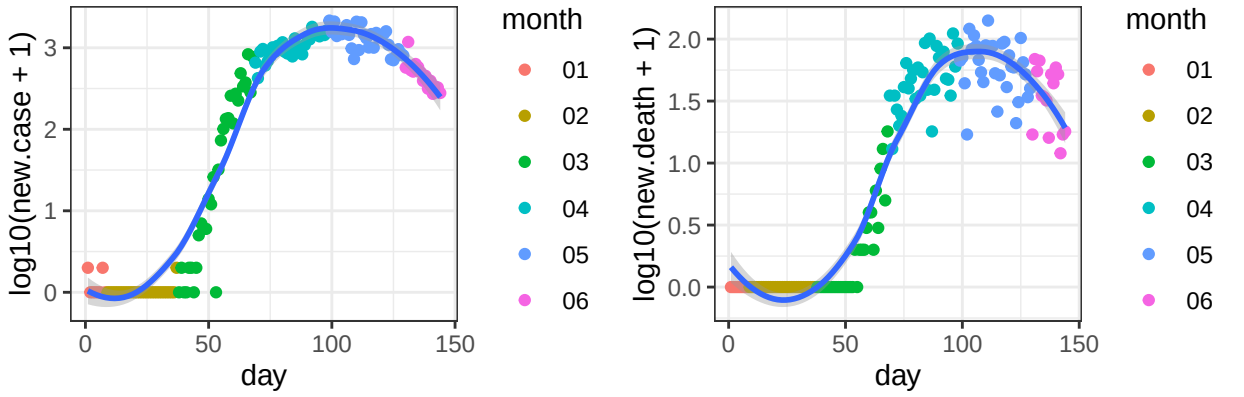
For these 50 counties, I check the number of new cases and the number of new deaths.

New York City_New York



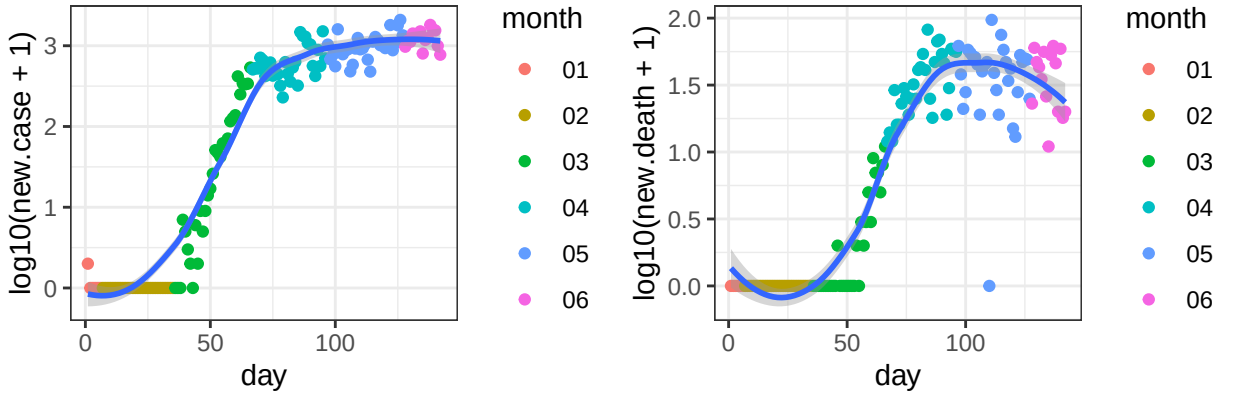
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



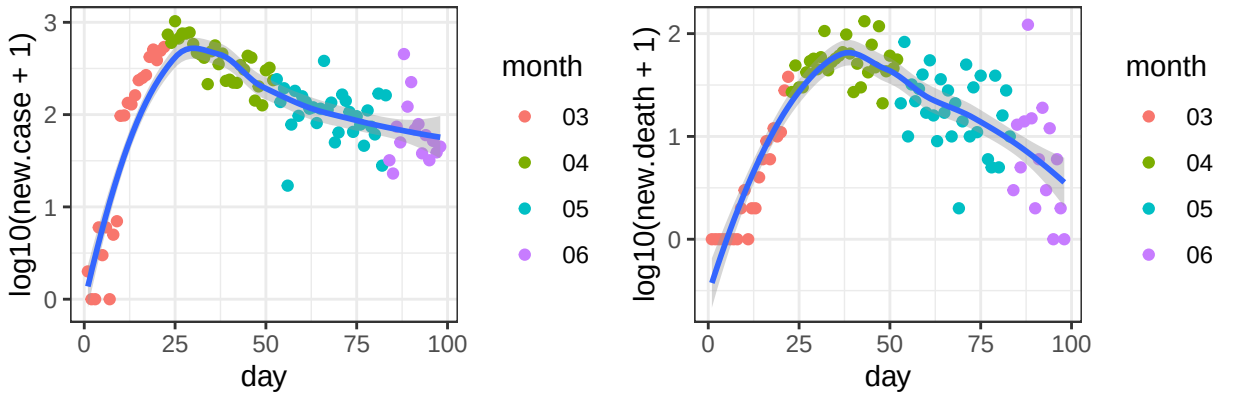
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Los Angeles_California



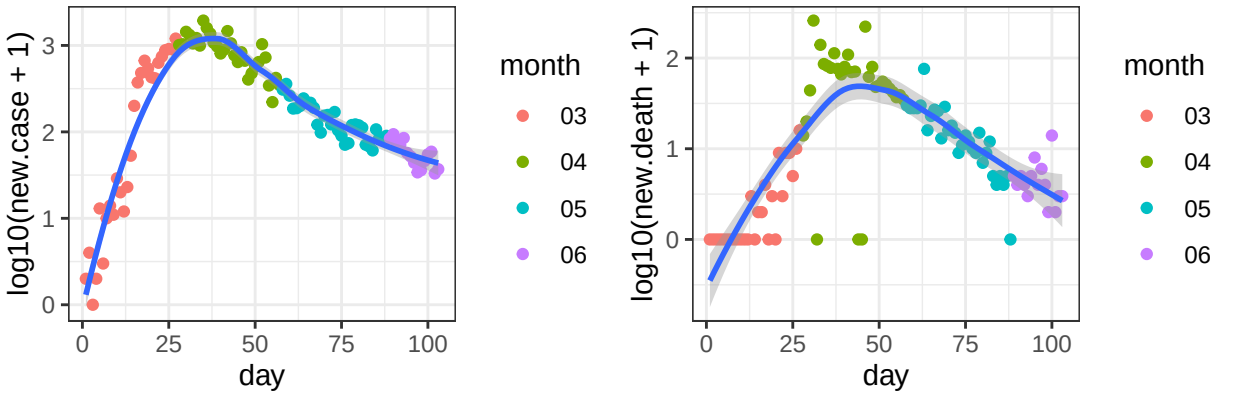
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Wayne_Michigan



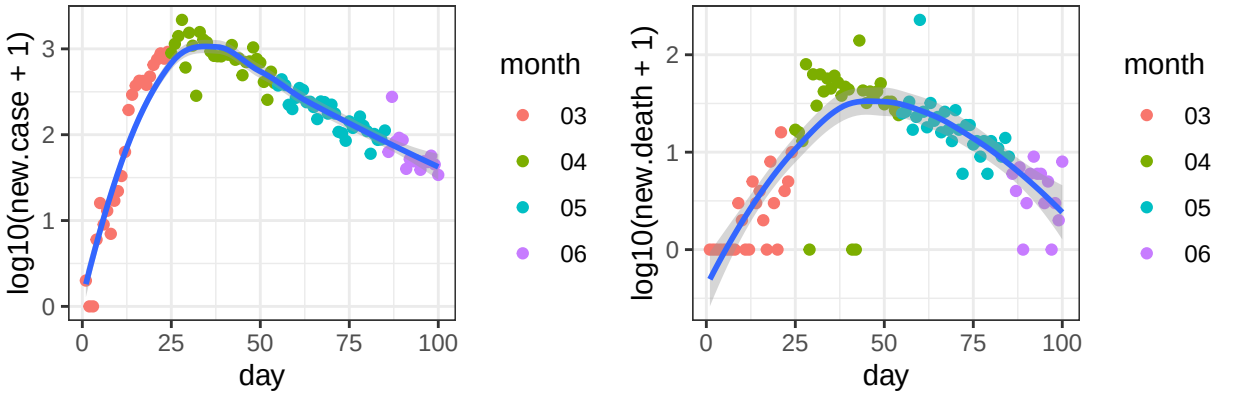
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Nassau_New York



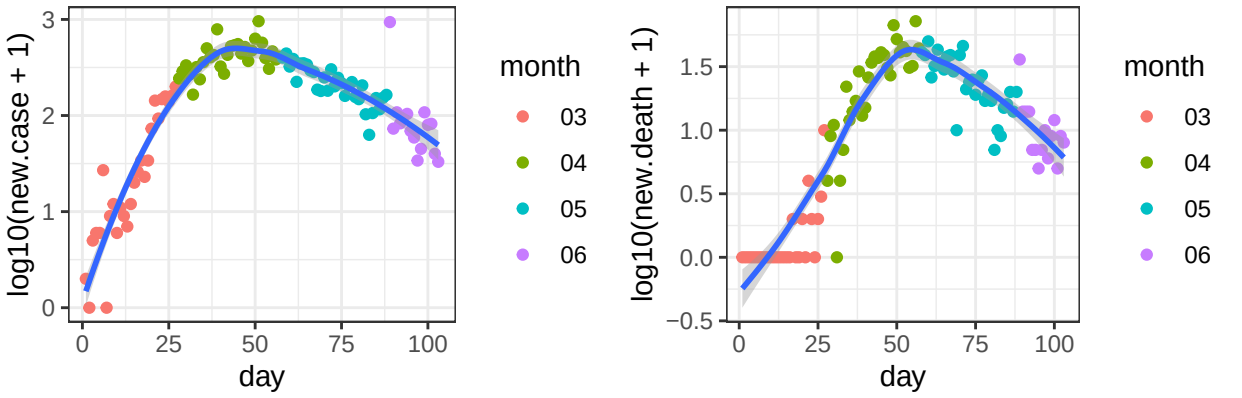
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Suffolk_New York



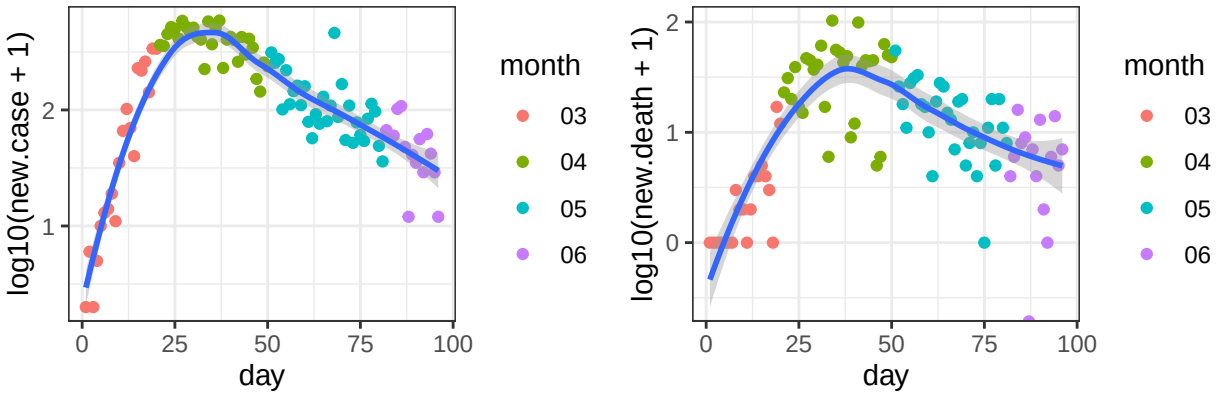
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Middlesex_Massachusetts



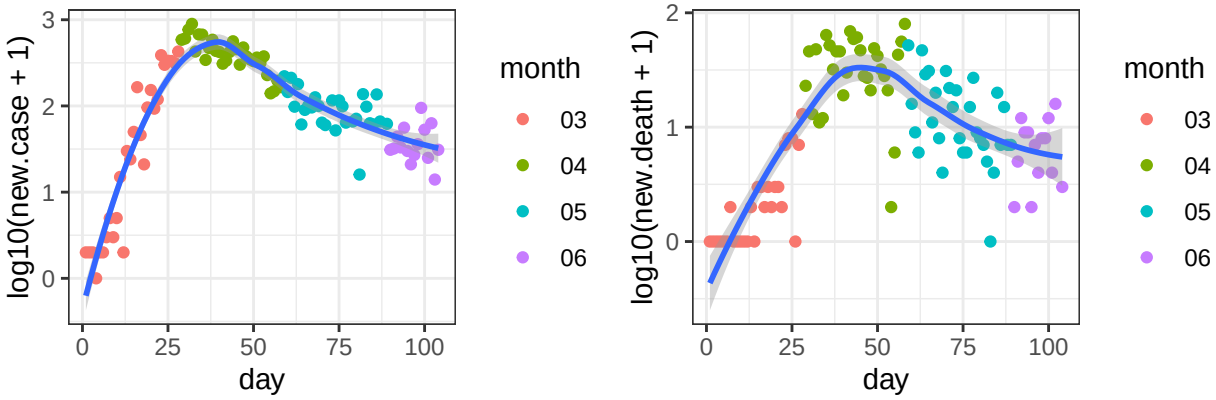
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



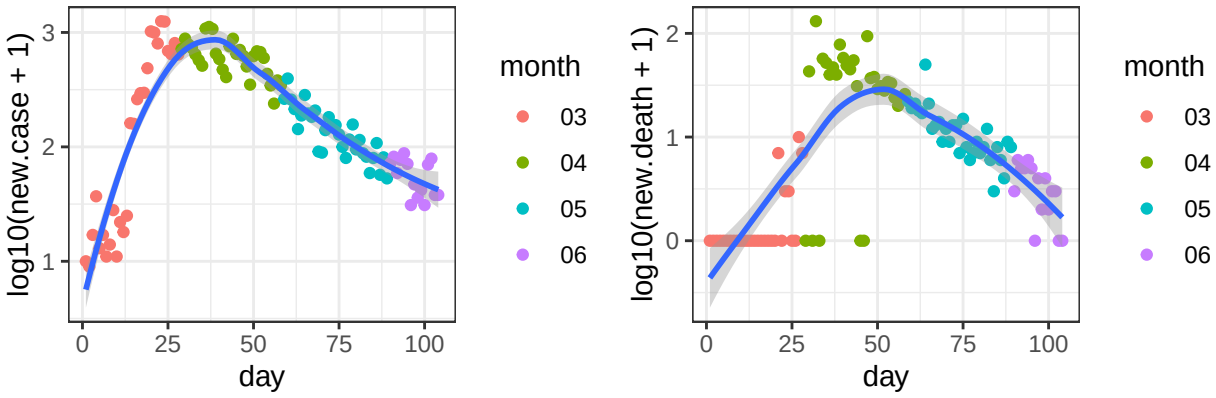
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Bergen_New Jersey



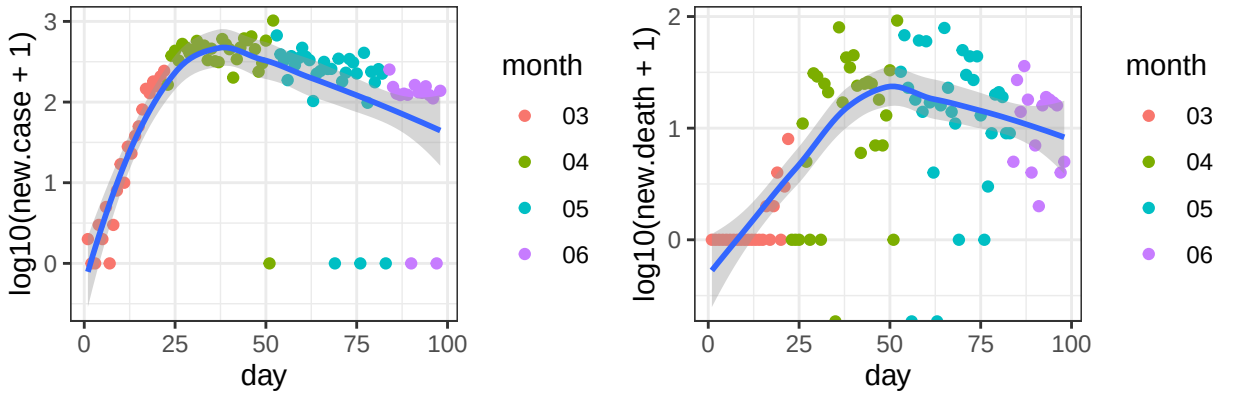
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Westchester_New York



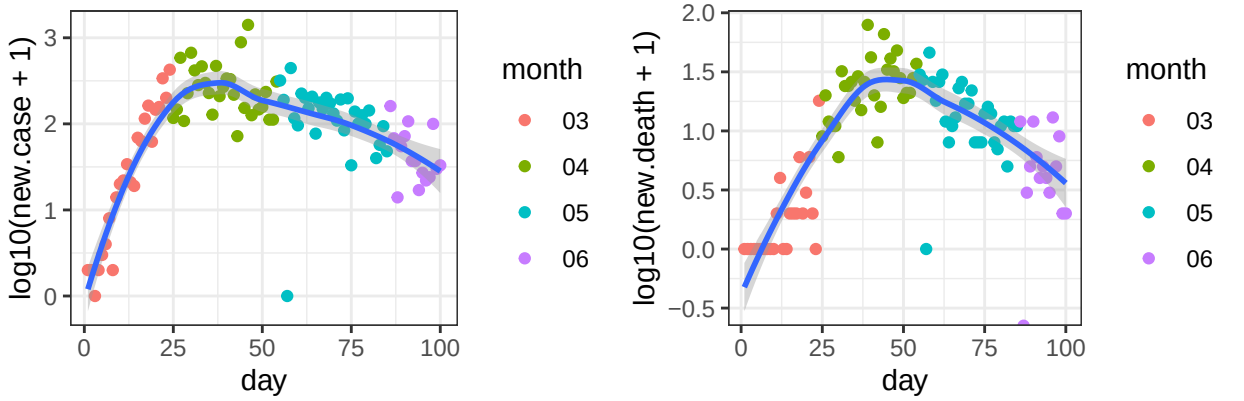
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Philadelphia_Pennsylvania



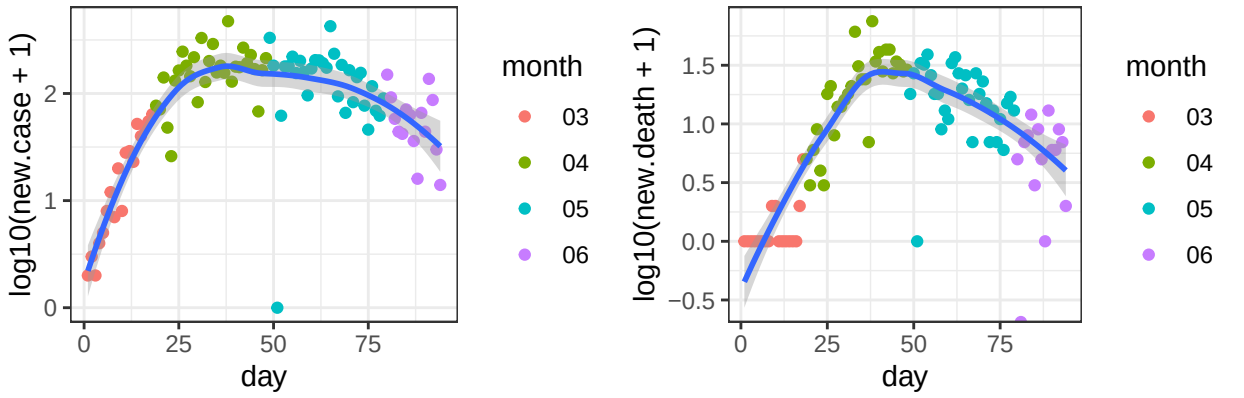
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Fairfield_Connecticut



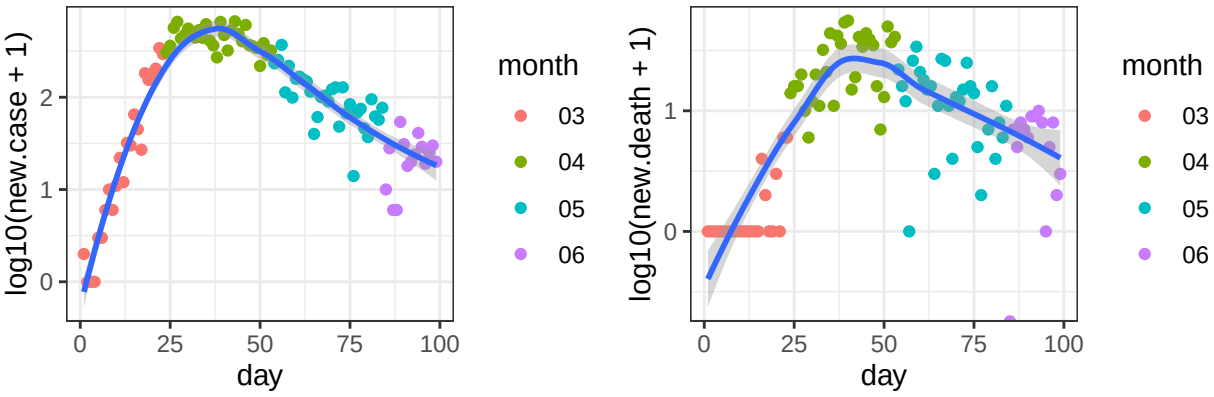
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Hartford_Connecticut



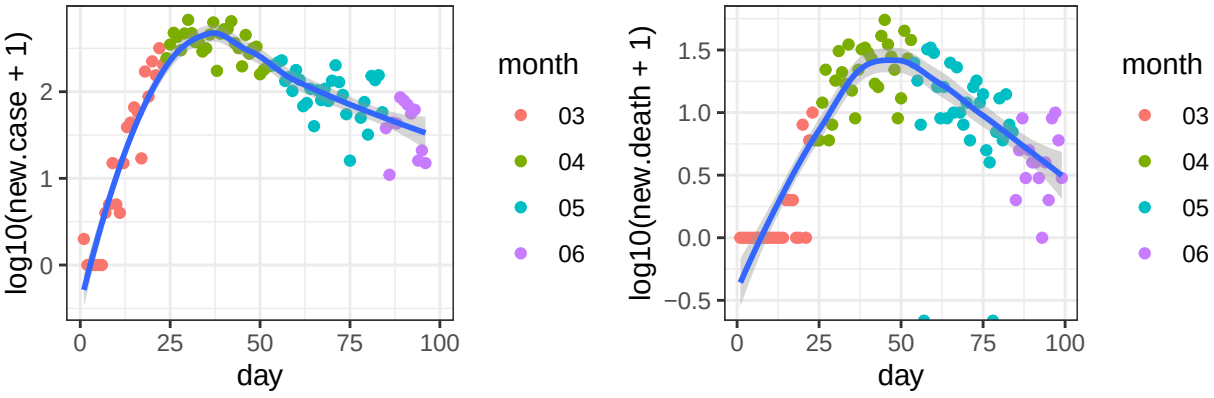
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Hudson_New Jersey



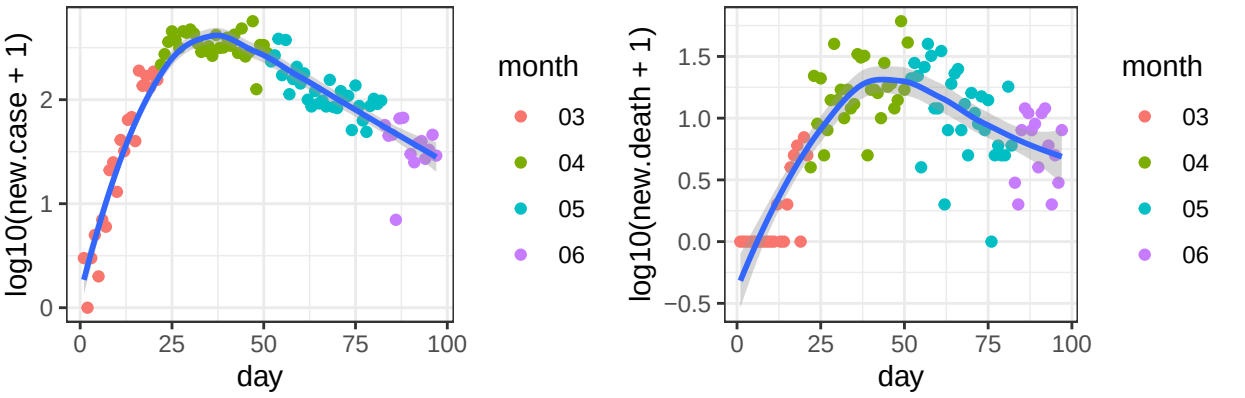
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Union_New Jersey



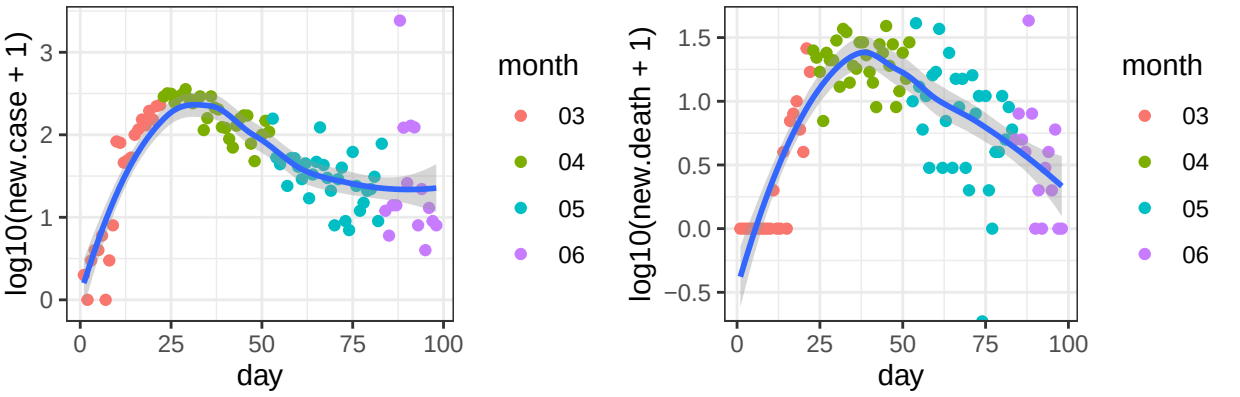
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Middlesex_New Jersey



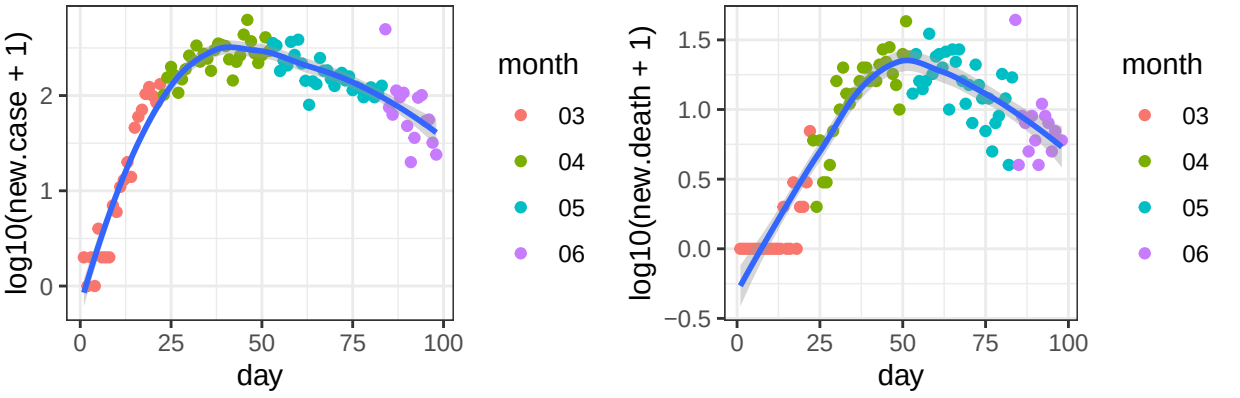
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Oakland_Michigan



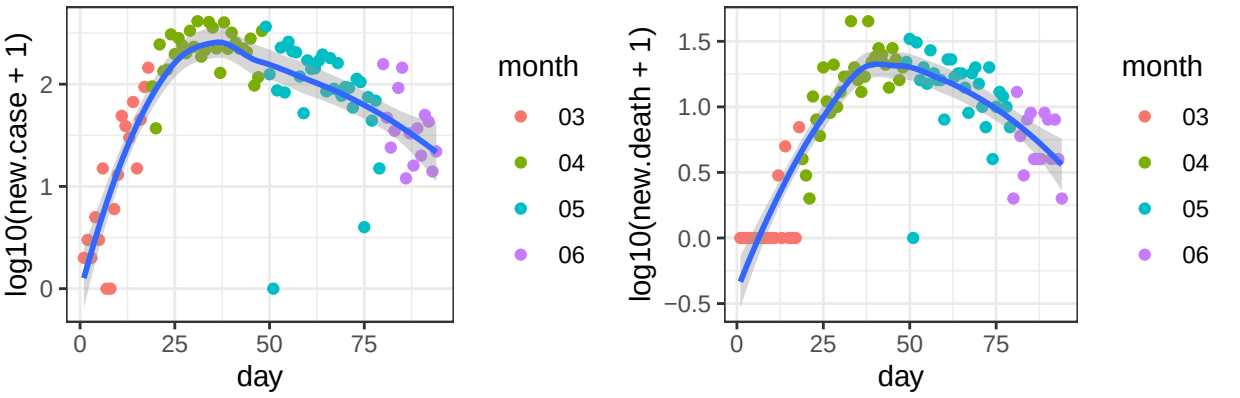
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Essex_Massachusetts



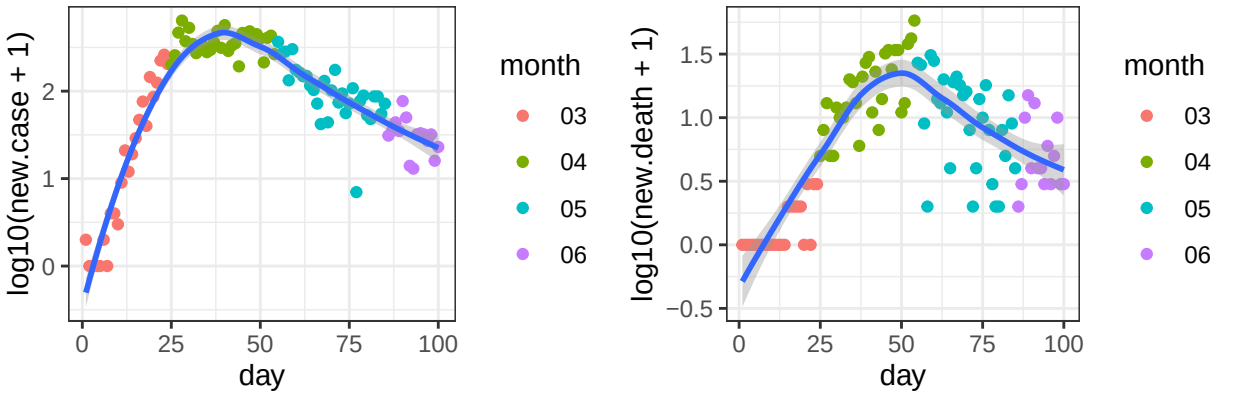
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

New Haven_Connecticut



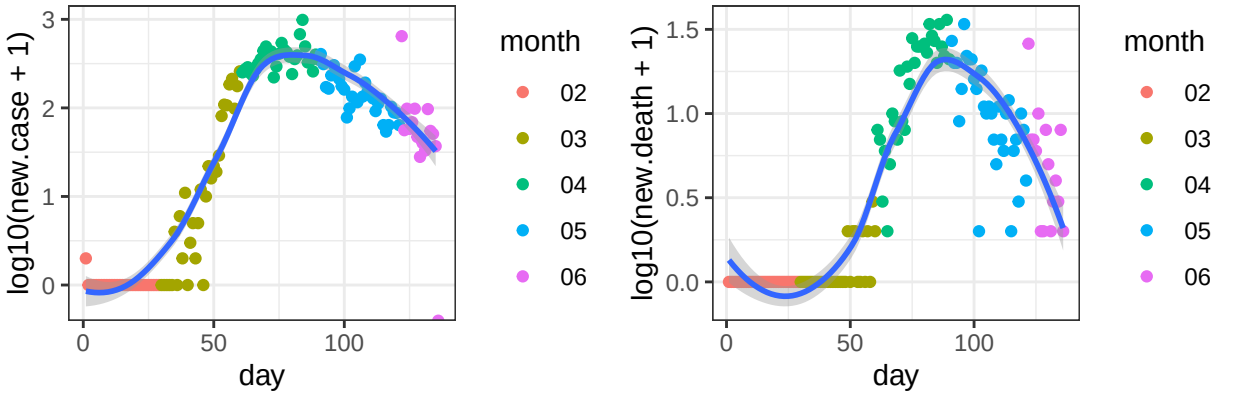
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Passaic_New Jersey



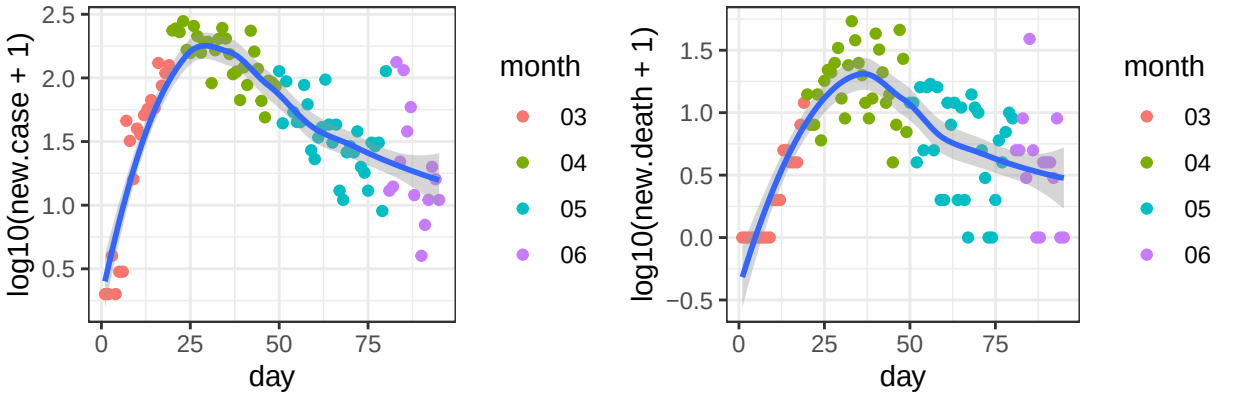
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Suffolk_Massachusetts



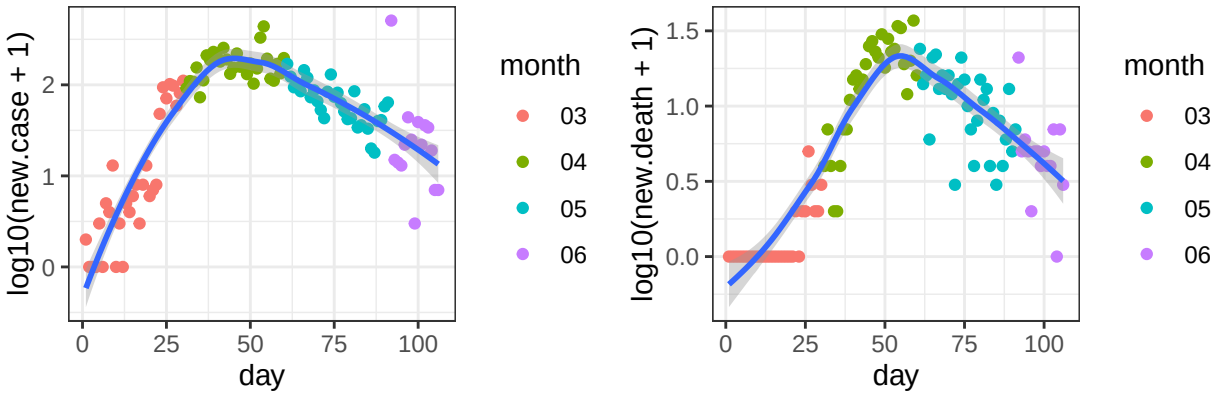
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Macomb_Michigan



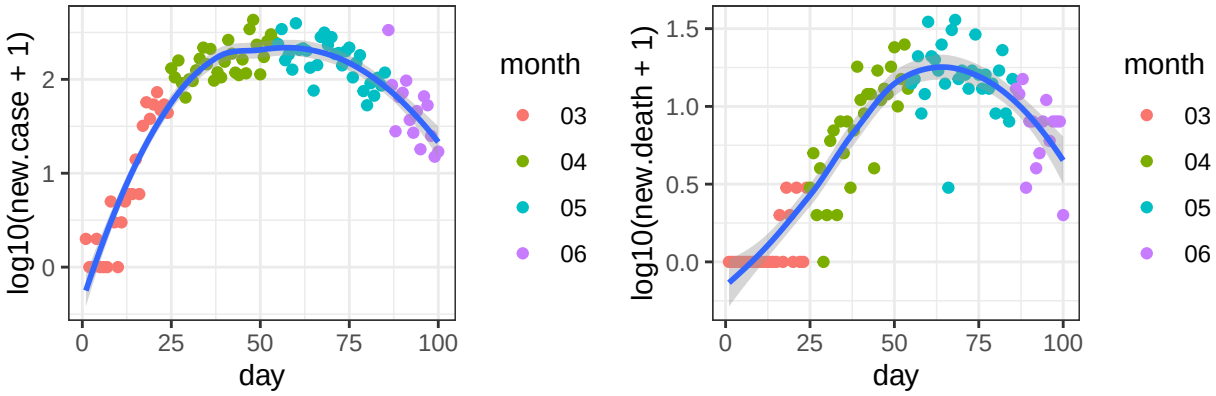
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Norfolk_Massachusetts



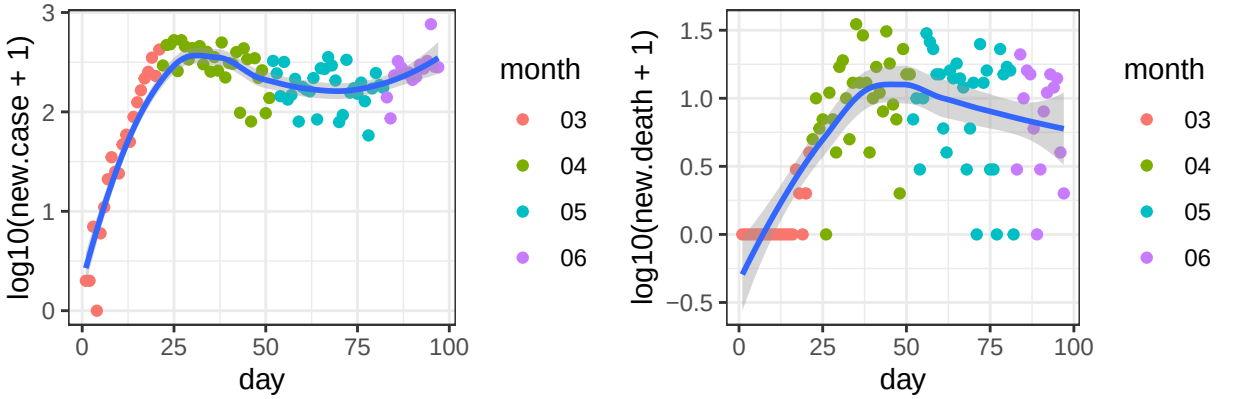
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Worcester_Massachusetts



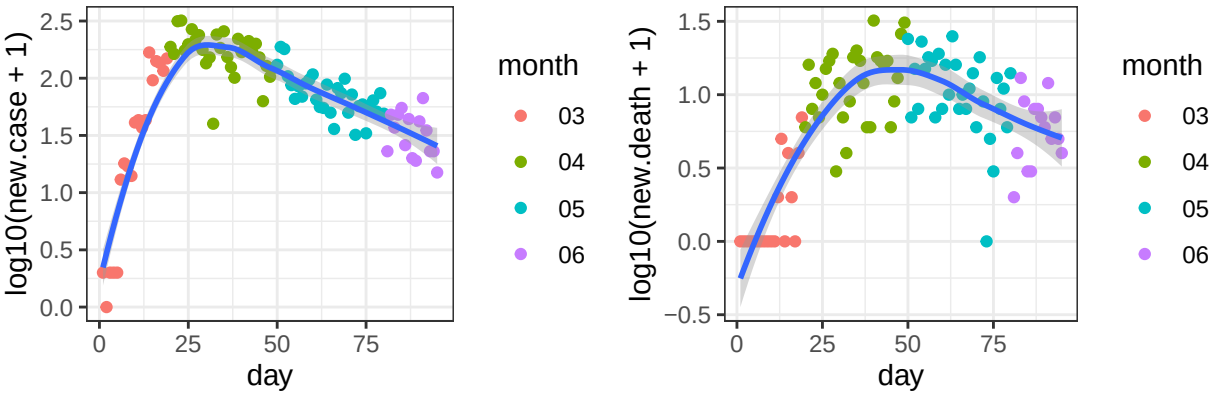
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Miami−Dade_Florida



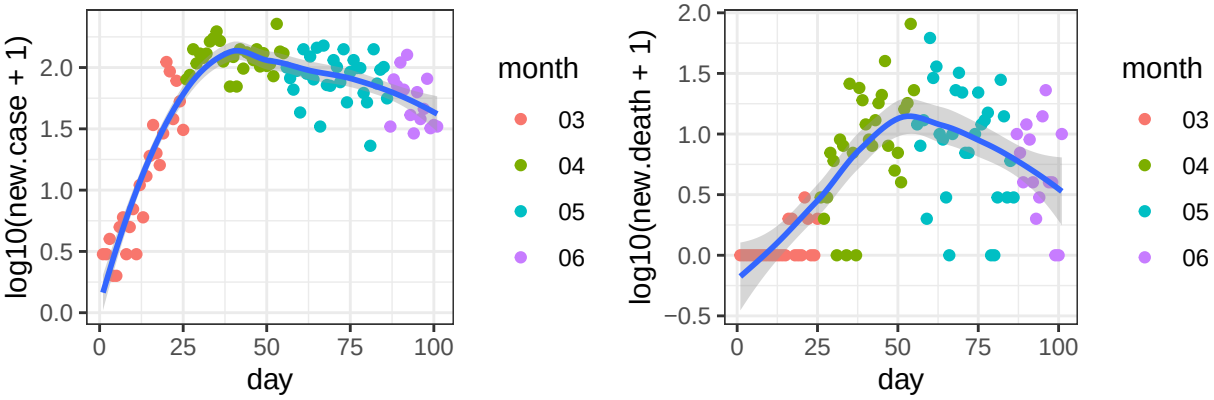
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Ocean_New Jersey



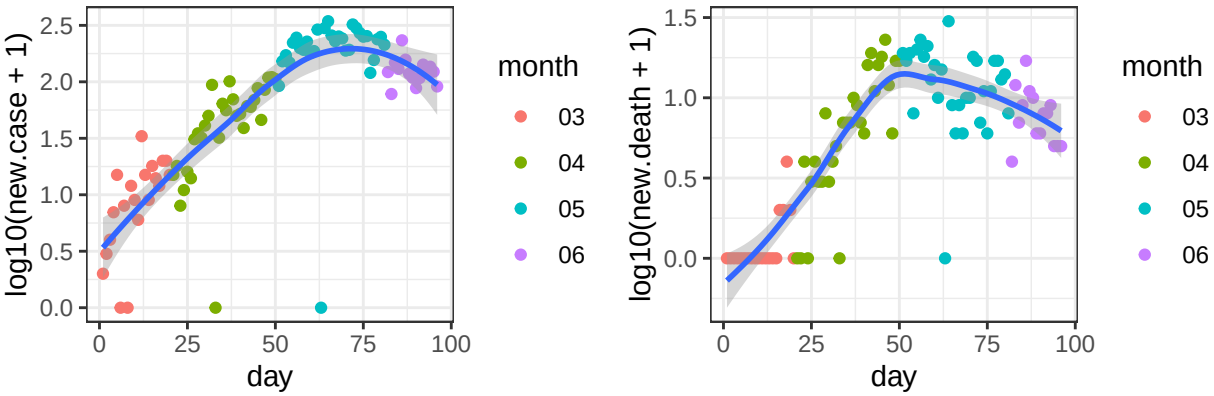
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Montgomery_Pennsylvania



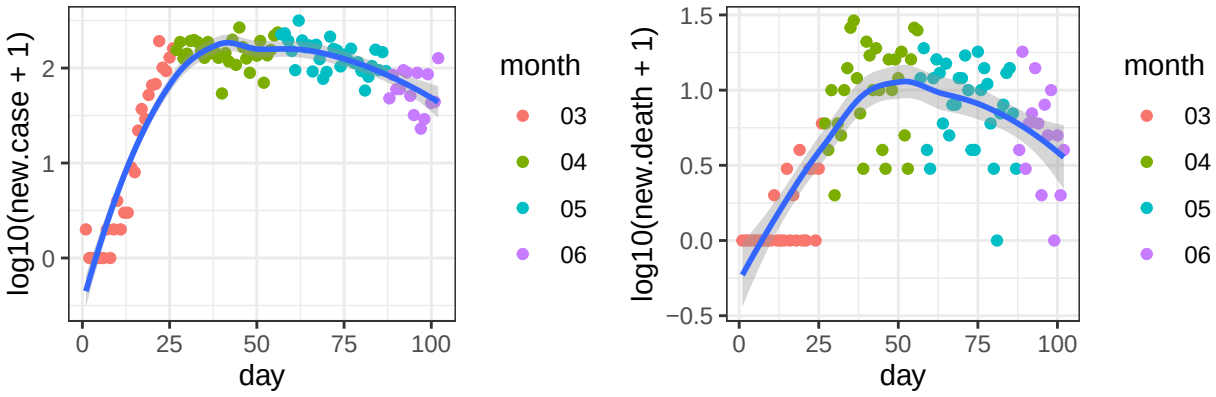
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Hennepin_Minnesota



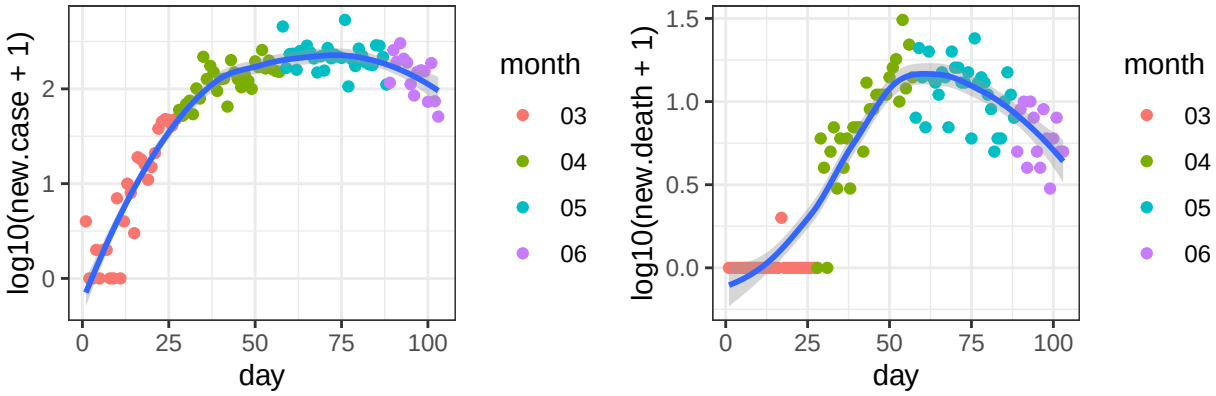
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Marion_Indiana



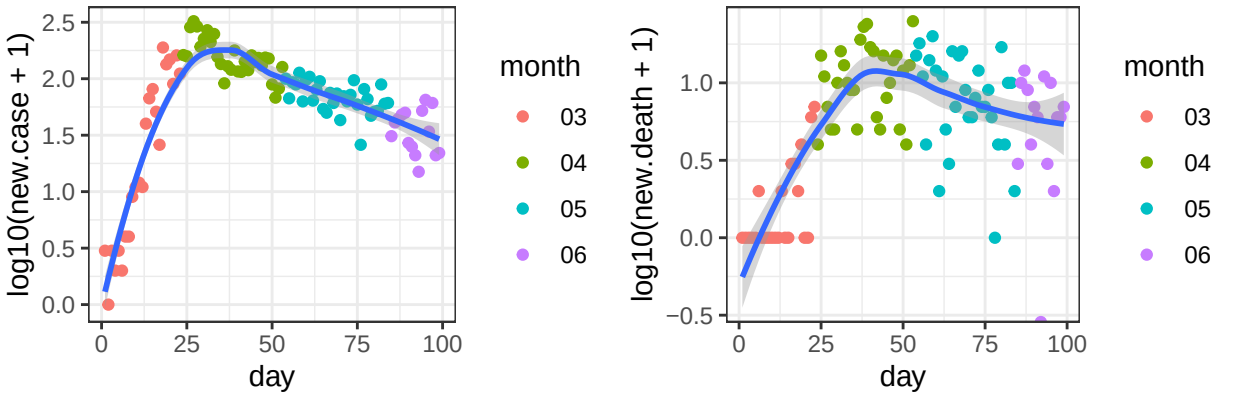
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Montgomery_Maryland



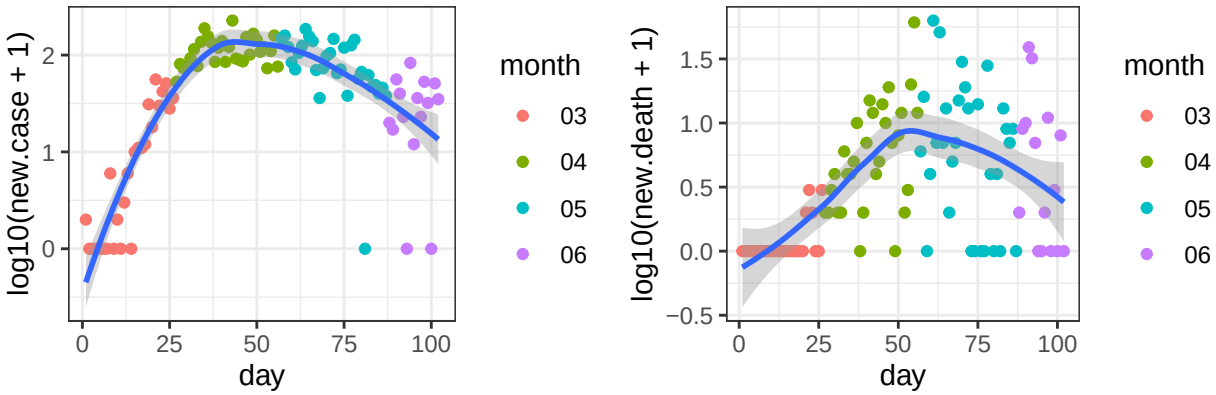
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Monmouth_New Jersey



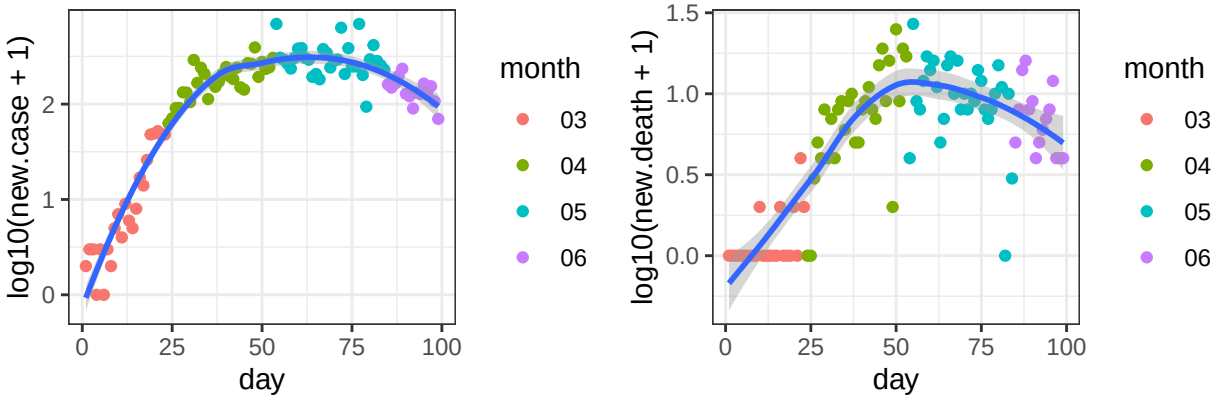
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Delaware_Pennsylvania



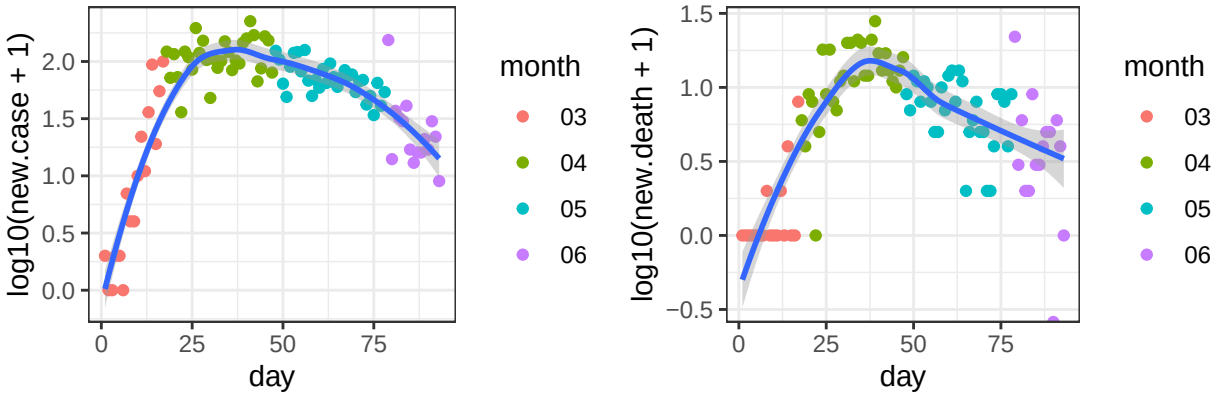
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Prince George's_Maryland



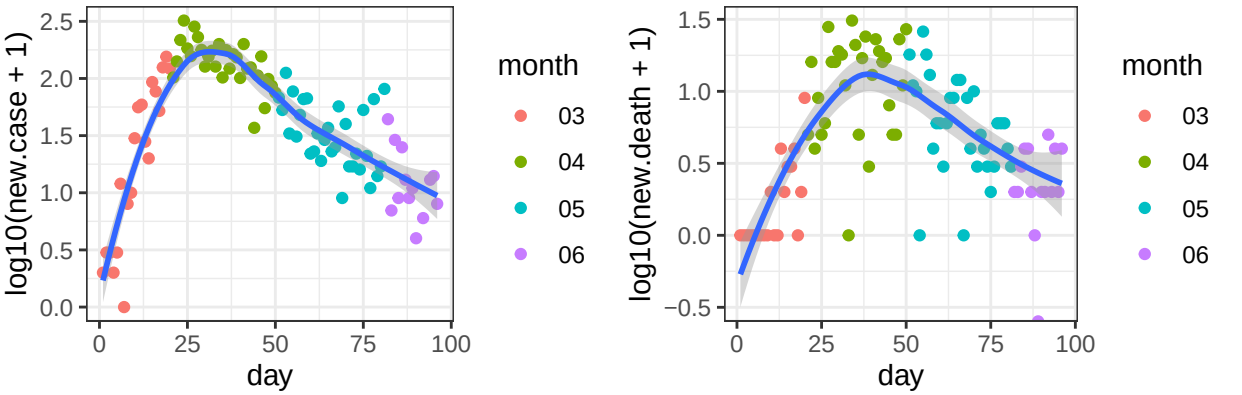
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Hampden_Massachusetts



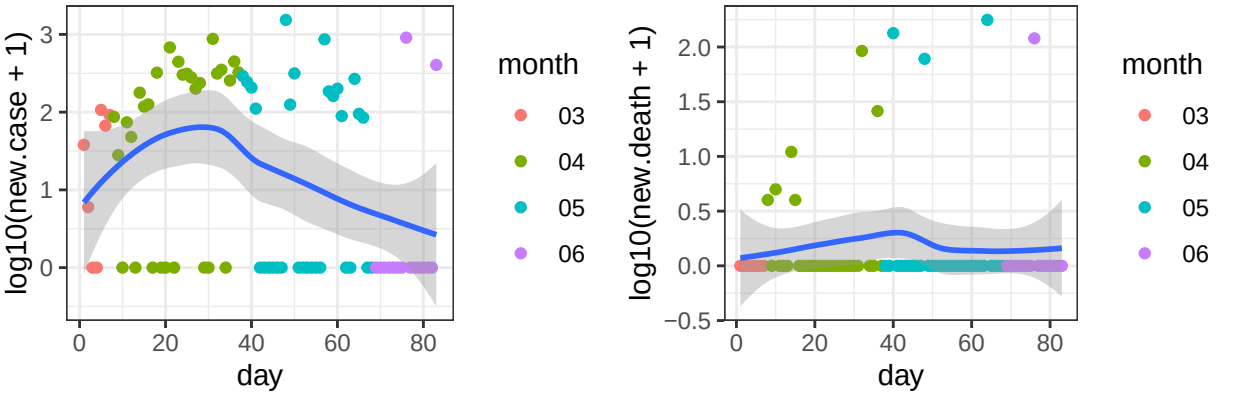
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Morris_New Jersey



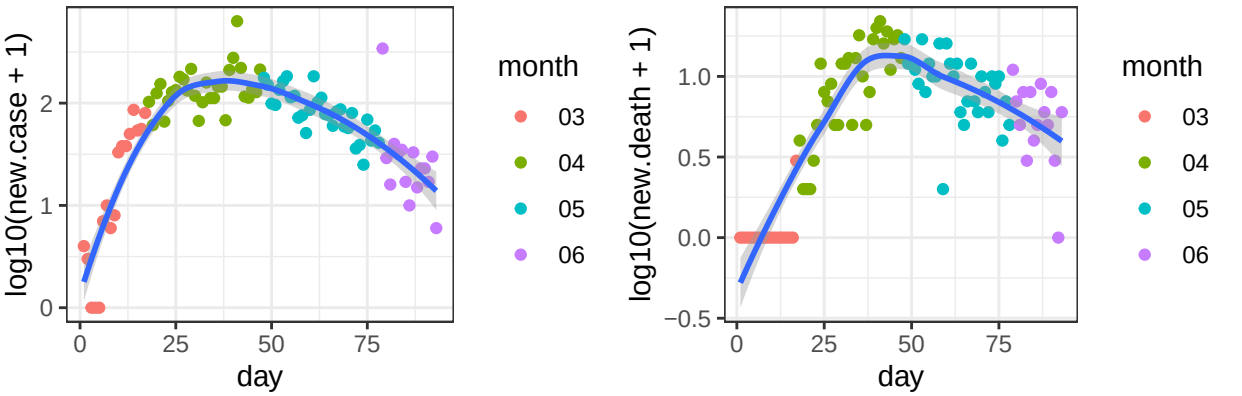
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Providence_Rhode Island



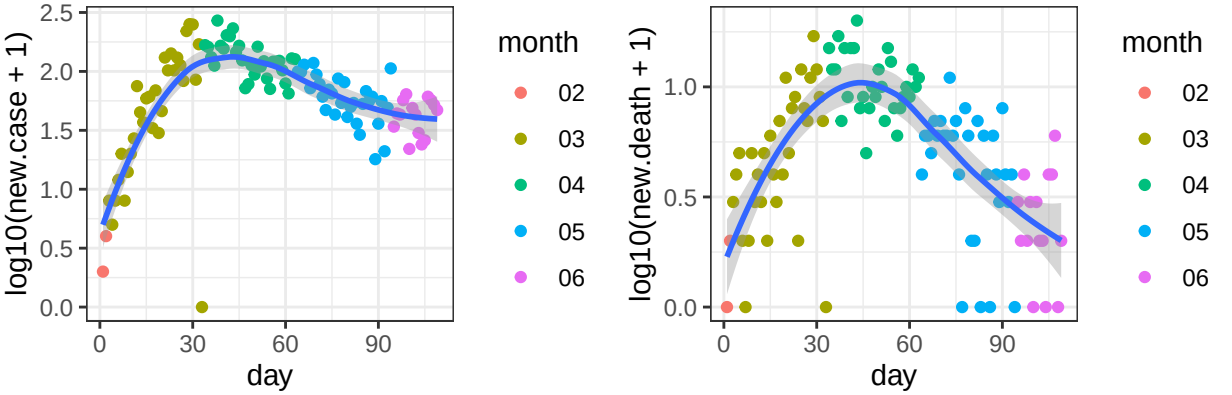
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Plymouth_Massachusetts



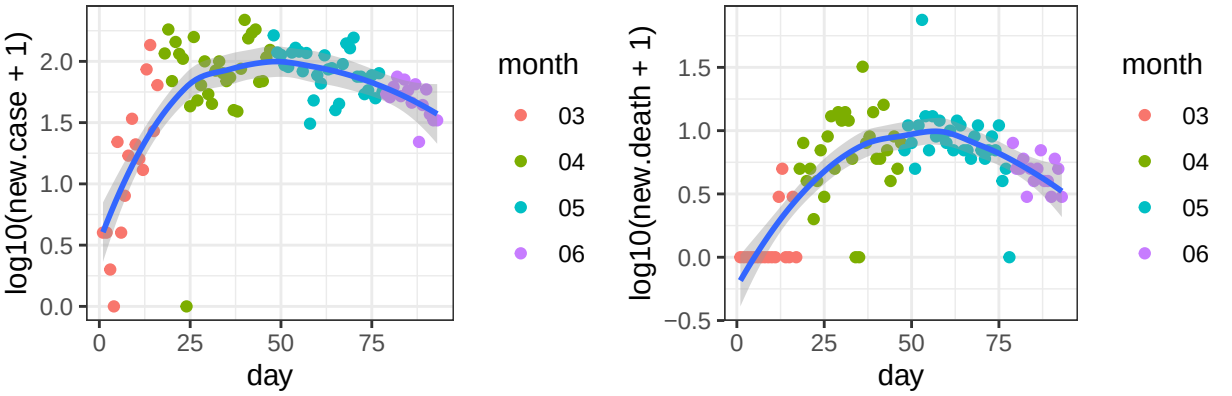
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

King_Washington



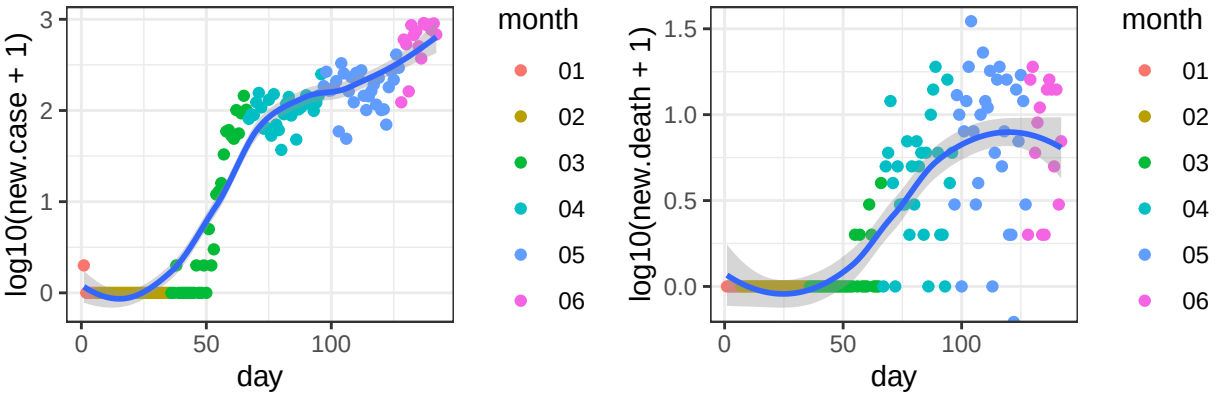
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

Erie_New York



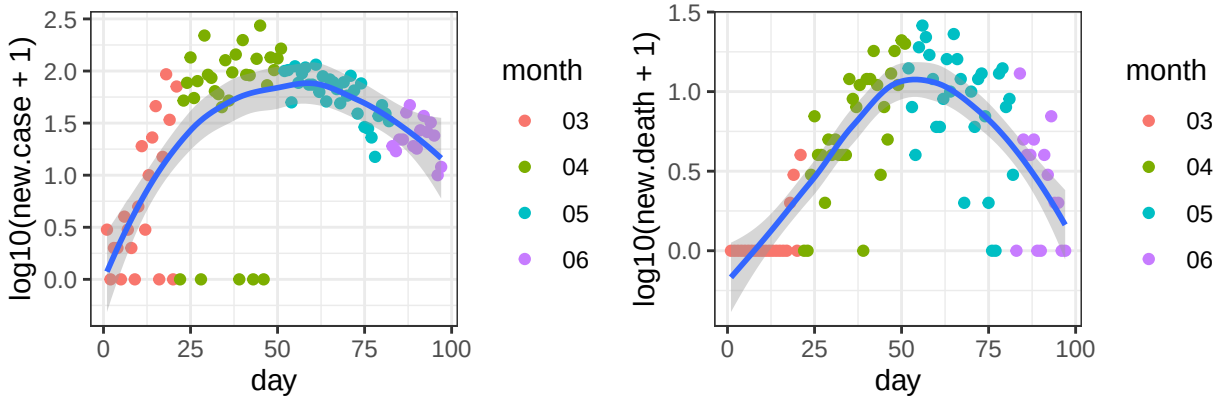
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Maricopa_Arizona



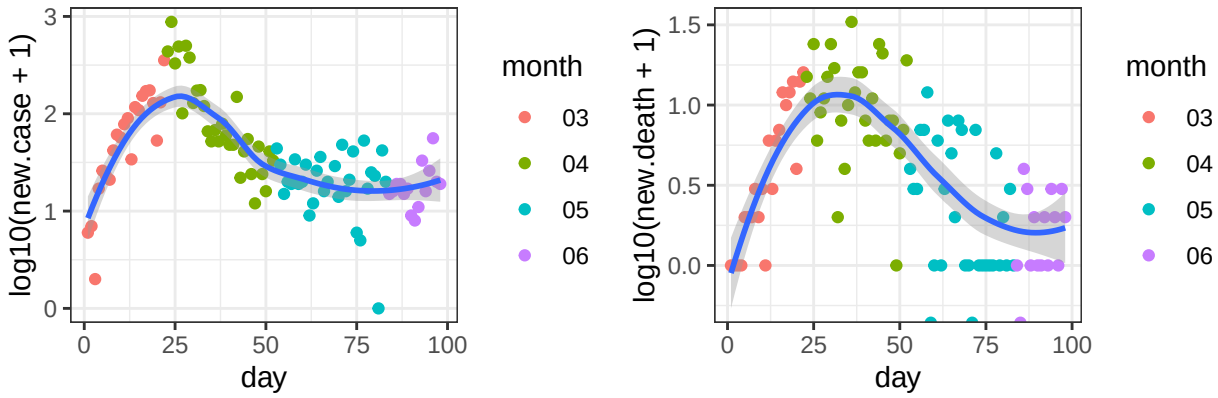
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Bucks_Pennsylvania



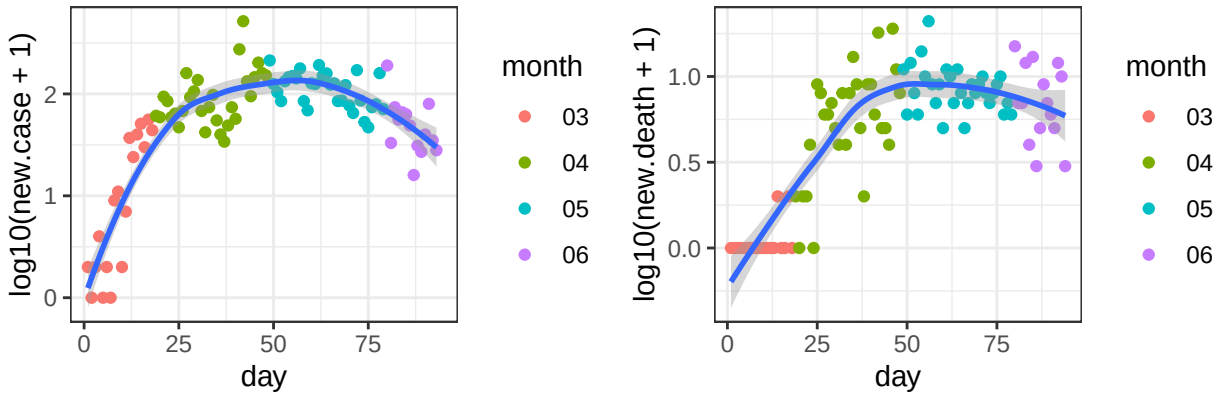
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Orleans_Louisiana



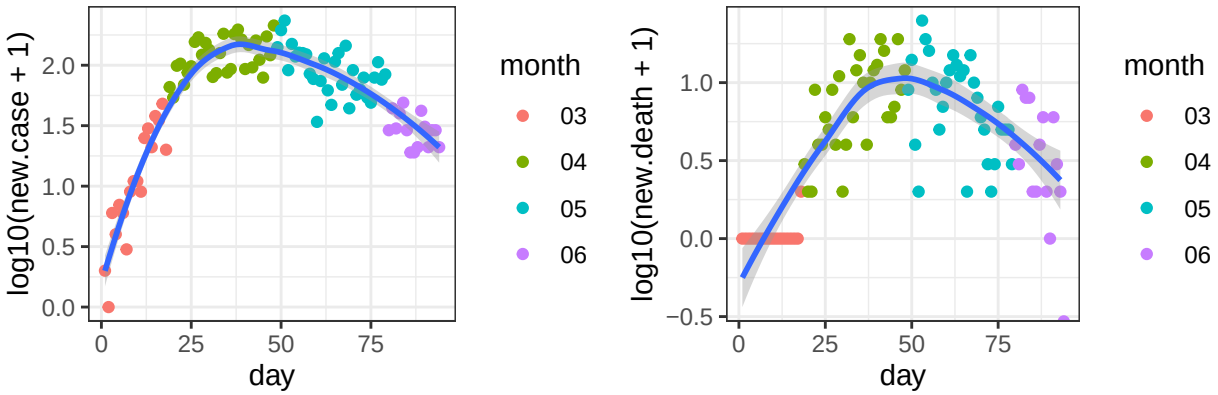
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Bristol_Massachusetts



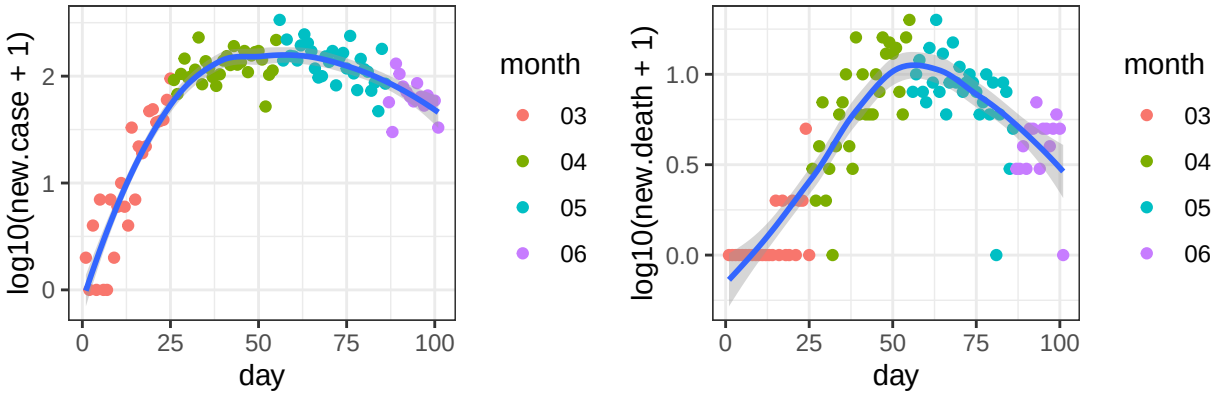
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Mercer_New Jersey



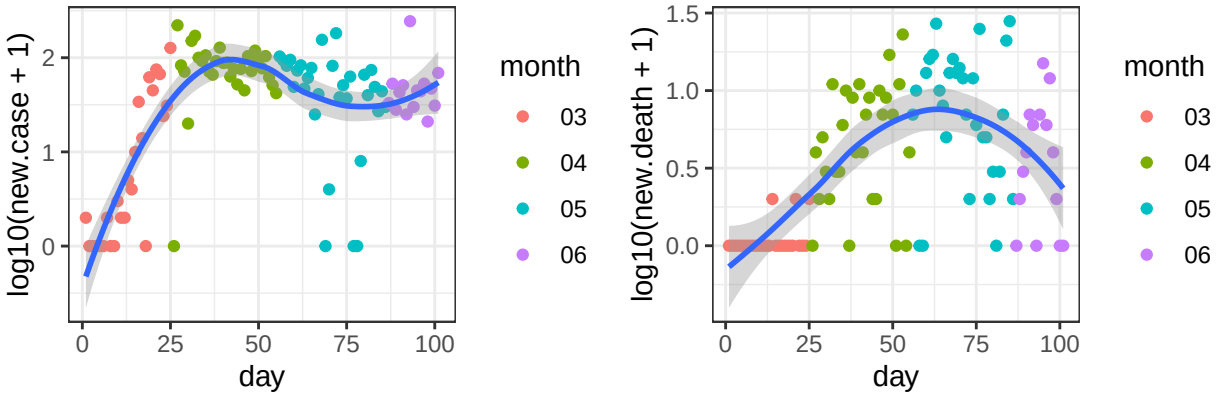
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

District of Columbia_District of Columbia



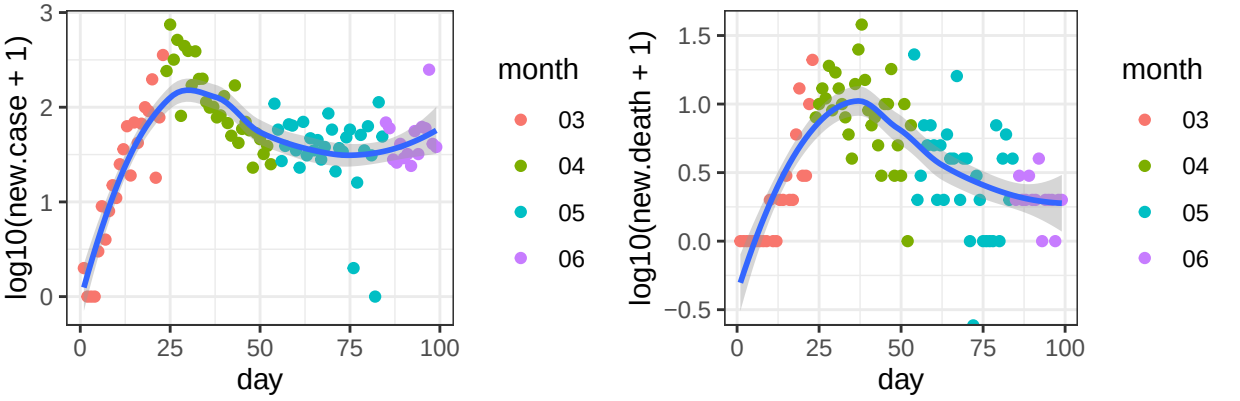
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

St. Louis_Missouri



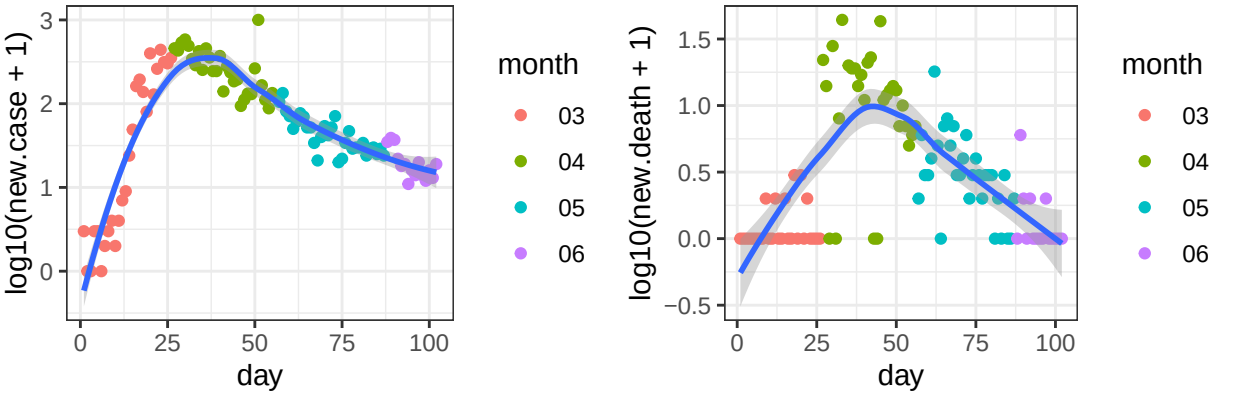
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Jefferson_Louisiana



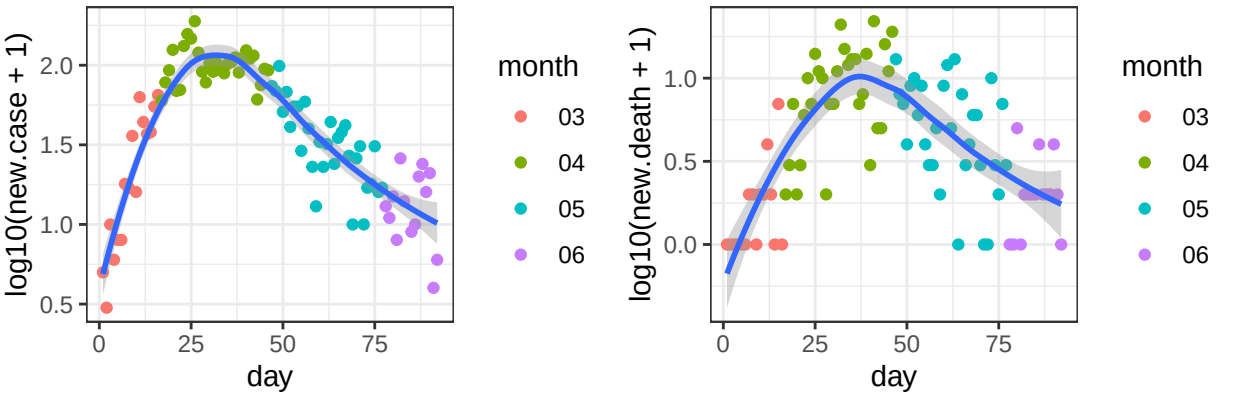
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Rockland_New York

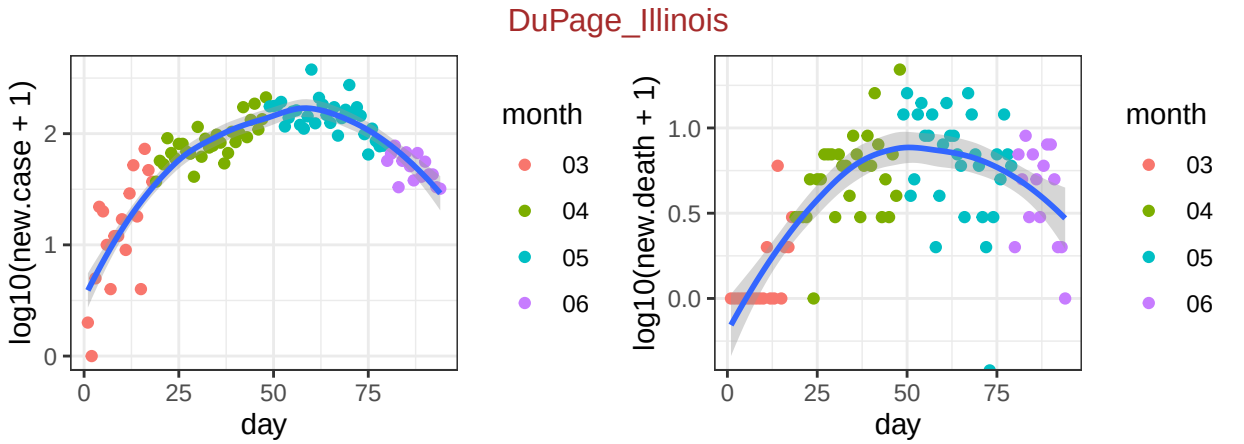


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Somerset_New Jersey



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

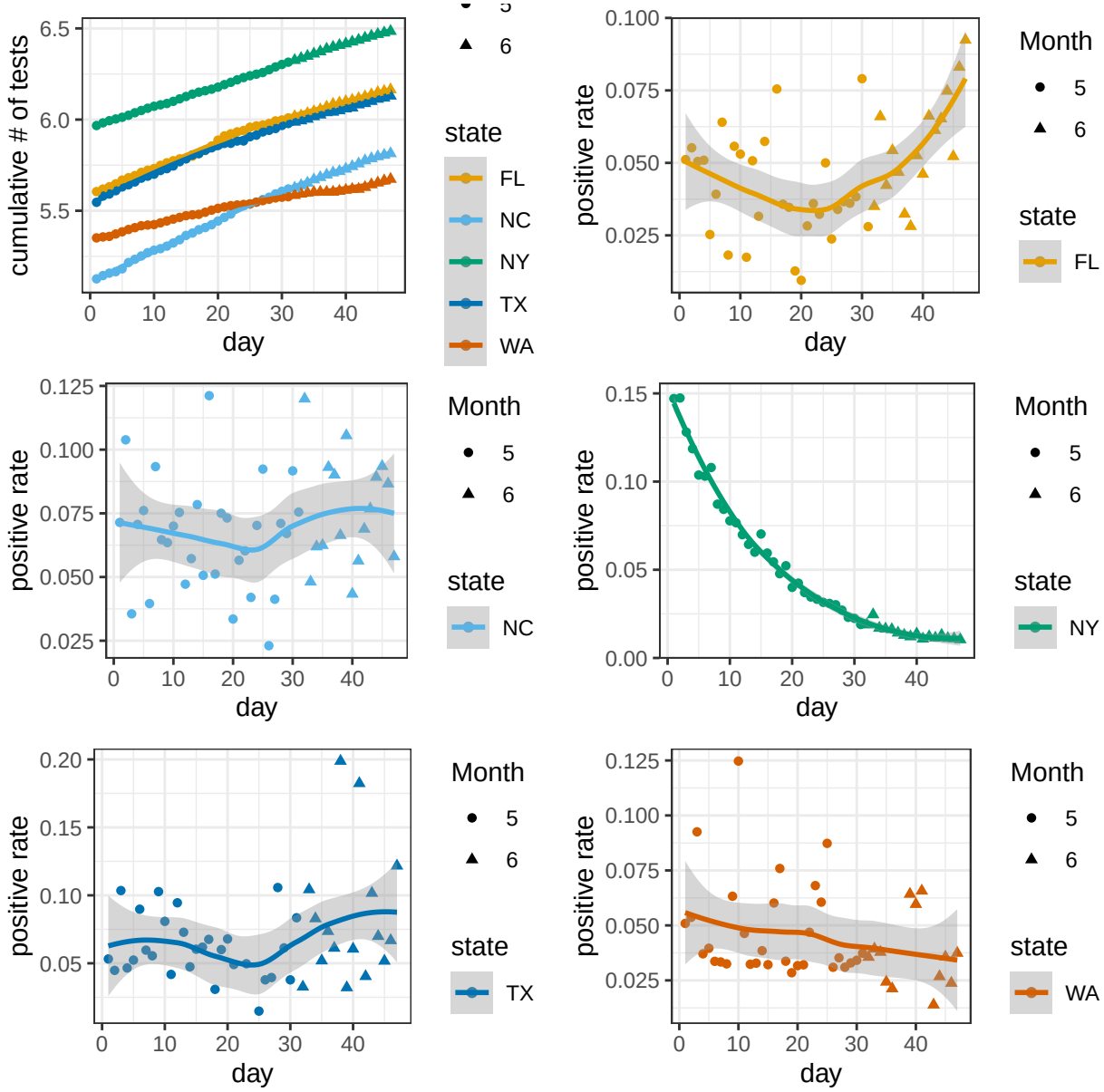


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

COVID Trackng

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

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



github.com/COVID19Tracking/, positive rate on 0616: 0.09(FL) 0.06(NC) 0.01(NY) 0.12(TX) 0.04(WA)

Session information

```
sessionInfo()
```

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

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