Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-08-08

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

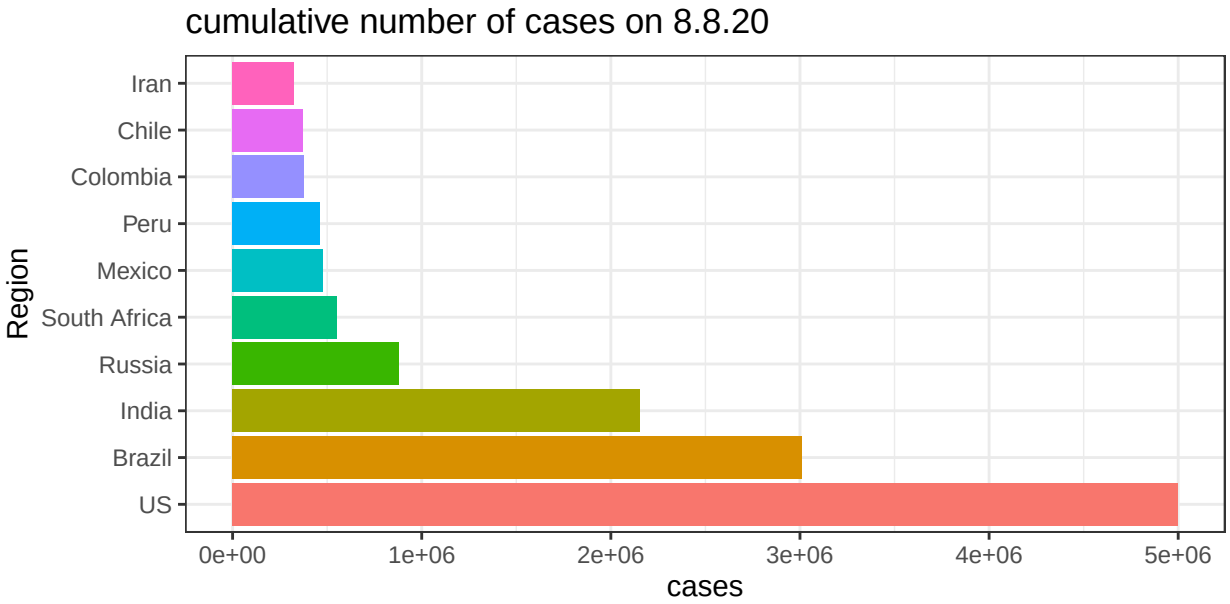
JHU

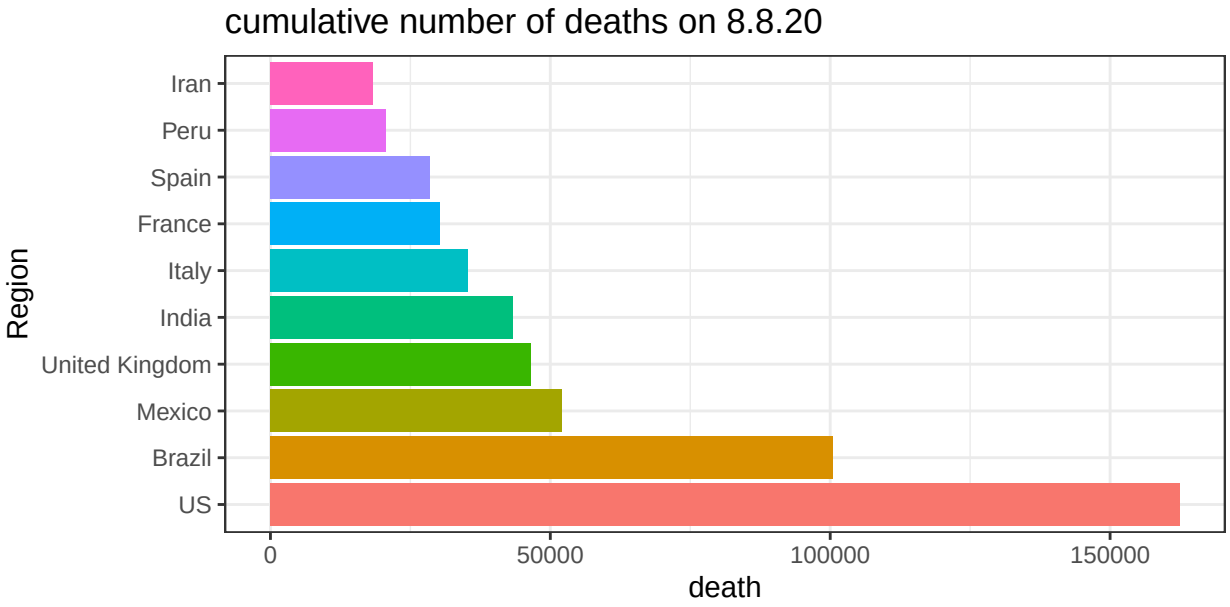
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

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.

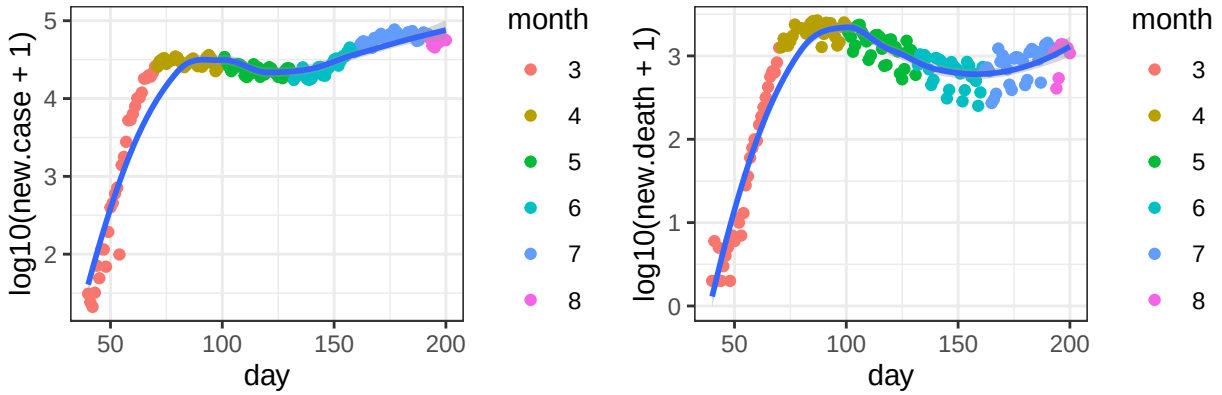
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





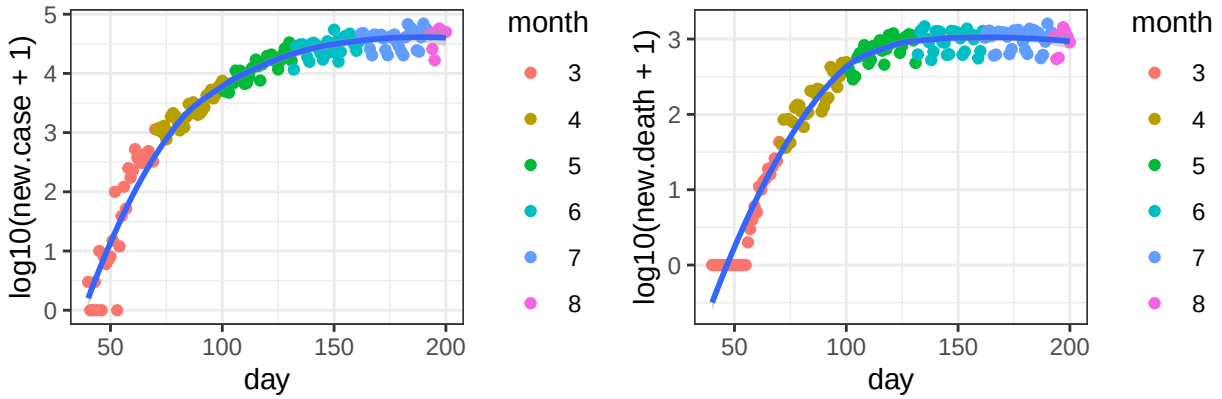
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



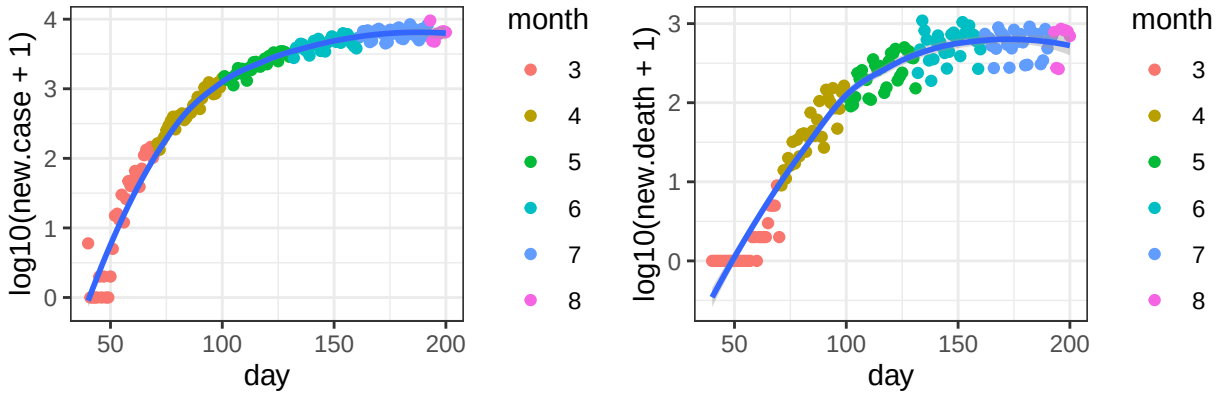
data source: https://github.com/CSSEGISandData/COVID−19

Brazil



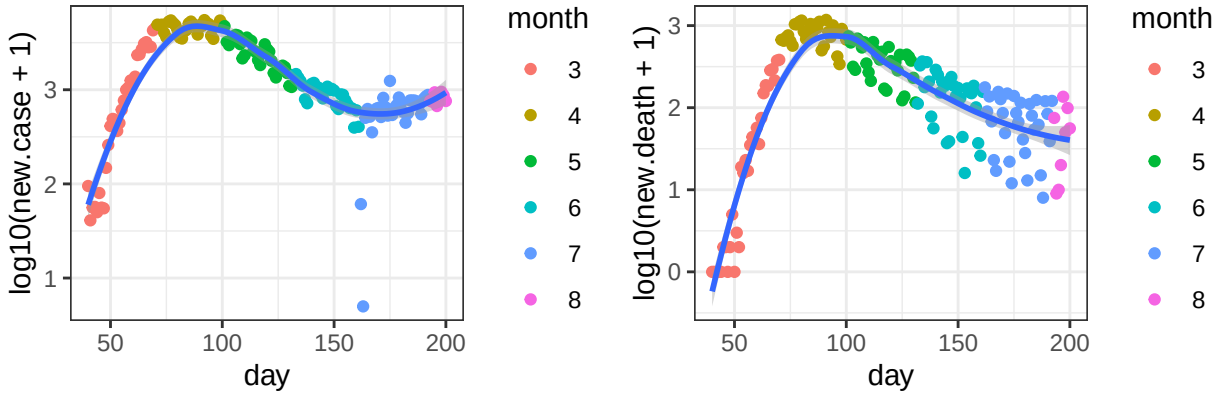
data source: https://github.com/CSSEGISandData/COVID−19

Mexico



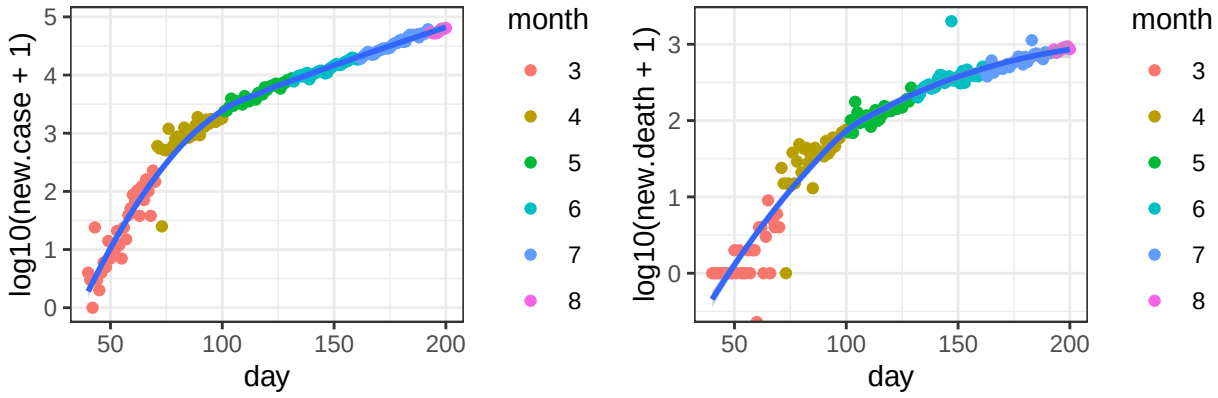
data source: https://github.com/CSSEGISandData/COVID−19

United Kingdom



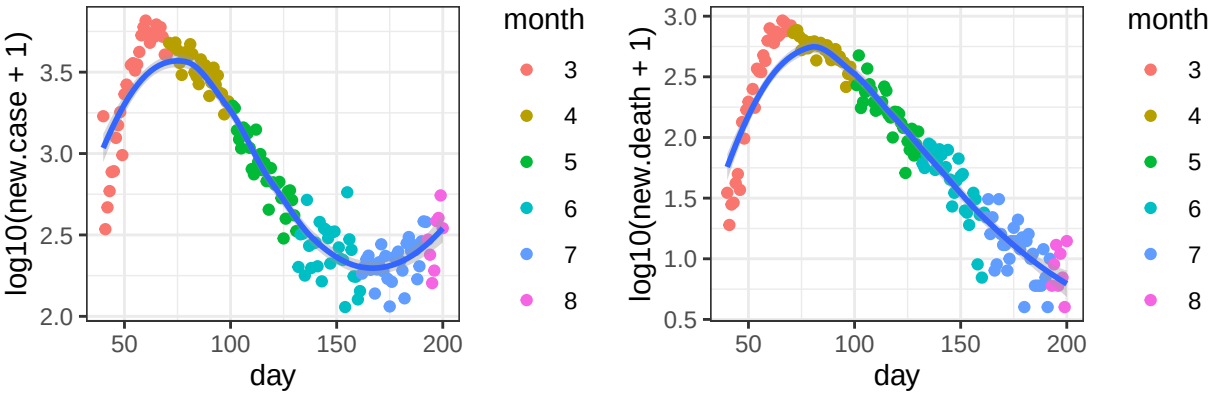
data source: https://github.com/CSSEGISandData/COVID−19

India



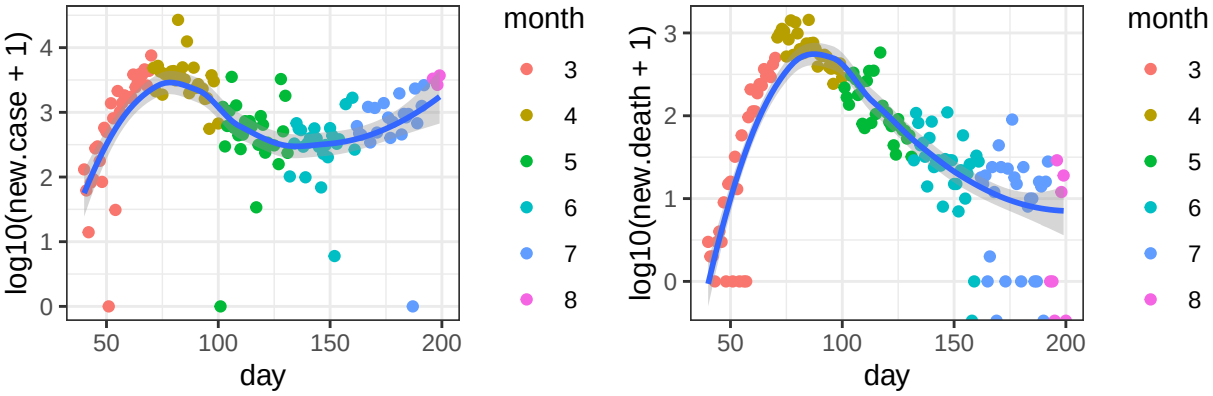
data source: https://github.com/CSSEGISandData/COVID−19

Italy



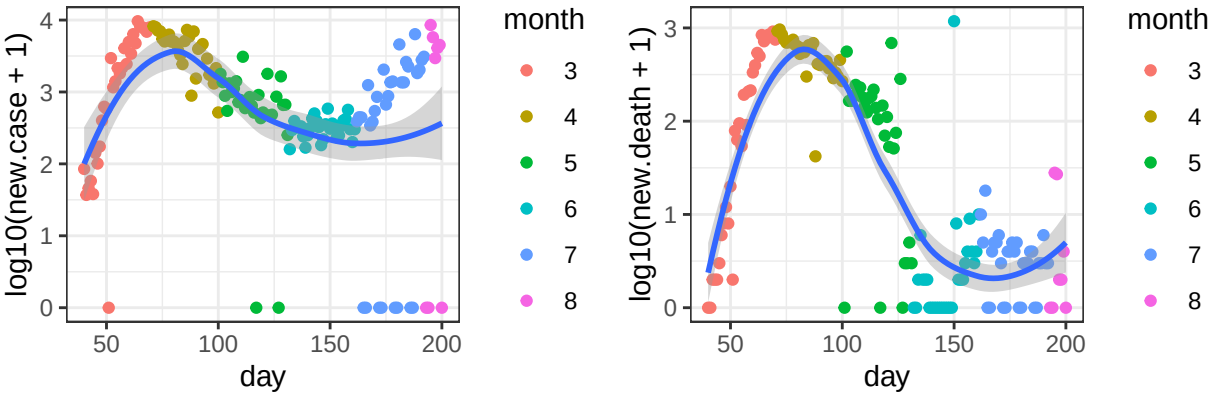
data source: https://github.com/CSSEGISandData/COVID−19

France



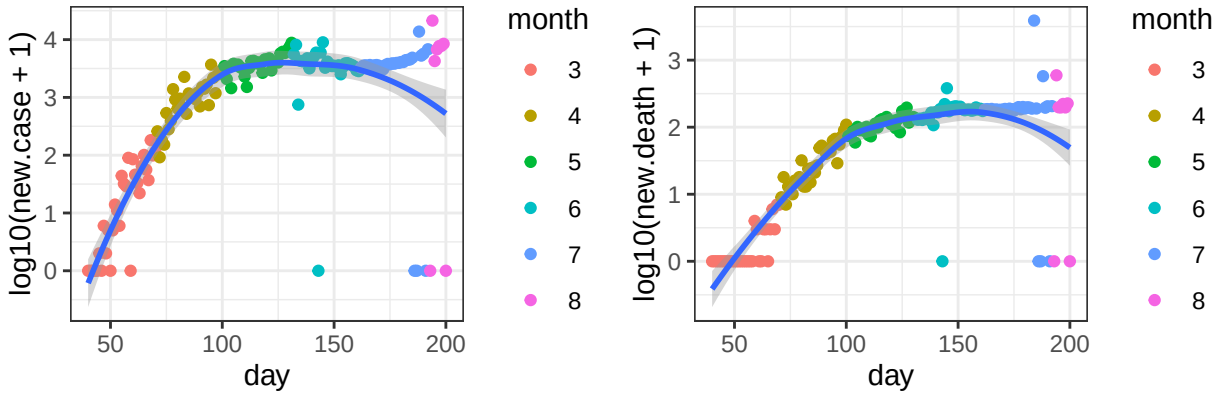
data source: https://github.com/CSSEGISandData/COVID−19

Spain



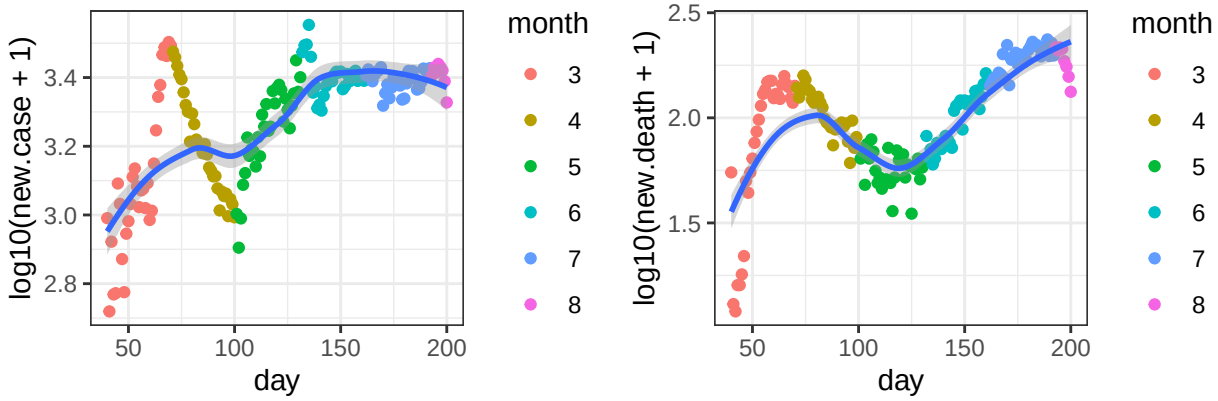
data source: https://github.com/CSSEGISandData/COVID−19

Peru



data source: https://github.com/CSSEGISandData/COVID−19

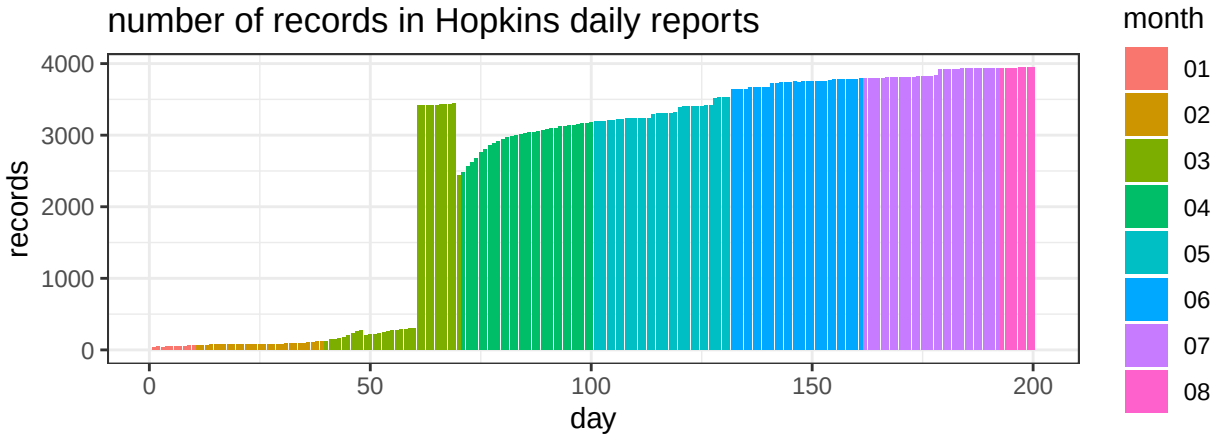
Iran



data source: https://github.com/CSSEGISandData/COVID−19

daily reports data

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



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

NY Times

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

The currente date is

```
## [1] "2020-08-08"
```

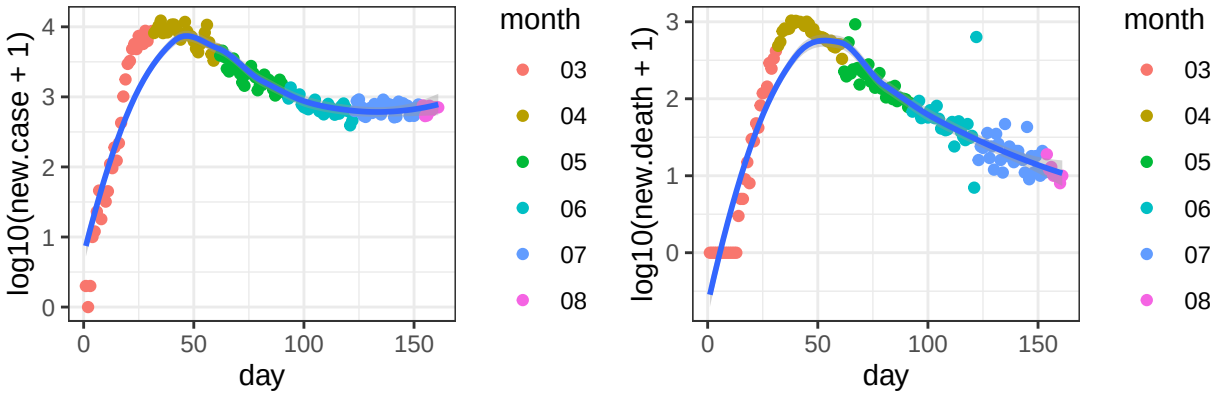
state level data

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

```
##            date          state fips  cases deaths
## 8738 2020-08-08       New York   36 425055  32345
## 8736 2020-08-08     New Jersey   34 186282  15869
## 8709 2020-08-08     California    6 556158  10299
## 8751 2020-08-08          Texas   48 502533   8879
## 8727 2020-08-08  Massachusetts   25 120711   8721
## 8714 2020-08-08        Florida   12 526569   8108
## 8719 2020-08-08       Illinois   17 194532   7846
## 8745 2020-08-08   Pennsylvania   42 122666   7372
## 8728 2020-08-08       Michigan   26  96328   6521
## 8711 2020-08-08    Connecticut    9  50320   4441
## 8724 2020-08-08      Louisiana   22 128864   4207
## 8707 2020-08-08        Arizona    4 186226   4140
## 8715 2020-08-08        Georgia   13 197308   4095
## 8742 2020-08-08           Ohio   39  99969   3668
## 8726 2020-08-08       Maryland   24  95157   3577
## 8720 2020-08-08        Indiana   18  75025   3036
## 8755 2020-08-08       Virginia   51  99189   2322
## 8739 2020-08-08 North Carolina   37 134968   2184
## 8748 2020-08-08 South Carolina   45  99460   2007
## 8730 2020-08-08    Mississippi   28  66646   1874
## 8710 2020-08-08       Colorado    8  50491   1862
## 8705 2020-08-08        Alabama    1 100173   1755
## 8756 2020-08-08     Washington   53  64347   1752
## 8729 2020-08-08      Minnesota   27  60142   1689
## 8731 2020-08-08       Missouri   29  58673   1381
## 8750 2020-08-08      Tennessee   47 117725   1201
## 8747 2020-08-08   Rhode Island   44  19738   1014
## 8758 2020-08-08      Wisconsin   55  64231   1007
## 8734 2020-08-08         Nevada   32  55481    949
## 8721 2020-08-08           Iowa   19  48311    926
```

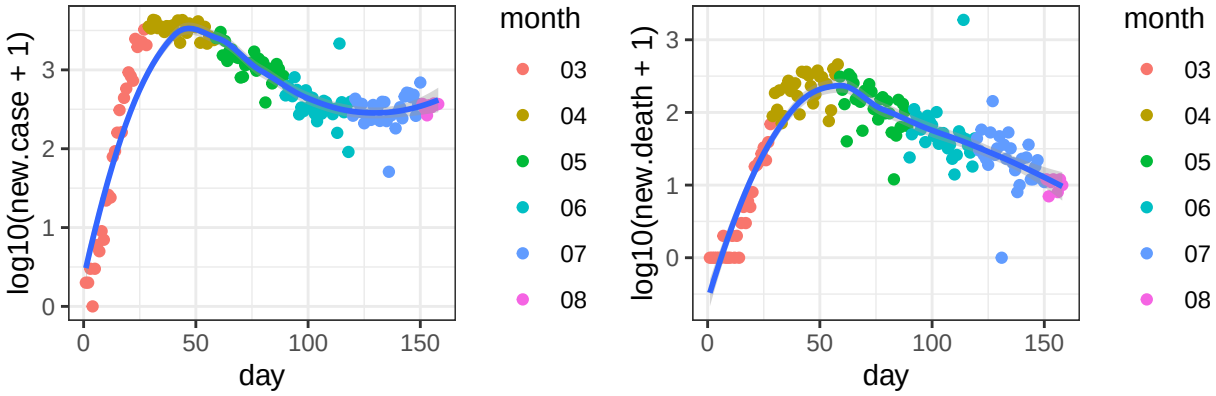
For these 30 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the similarities and differences across states.

New York



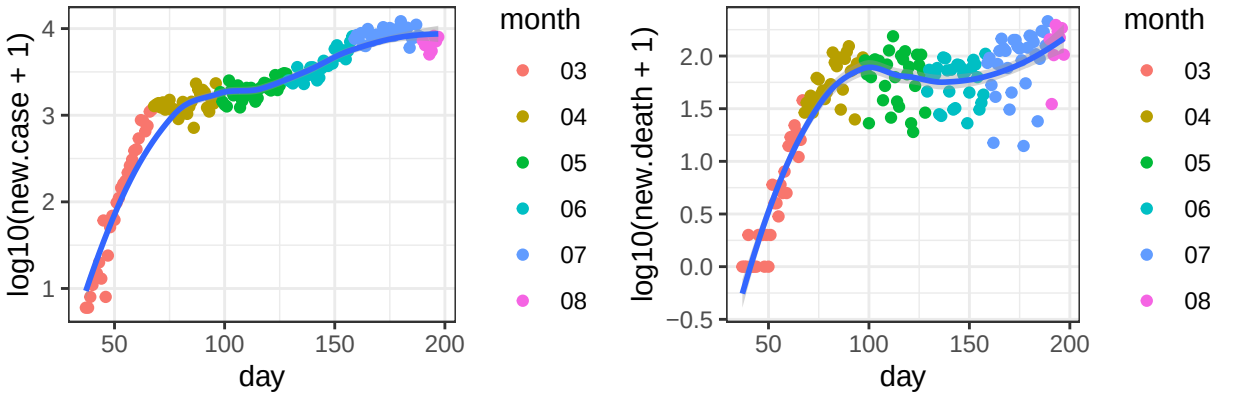
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



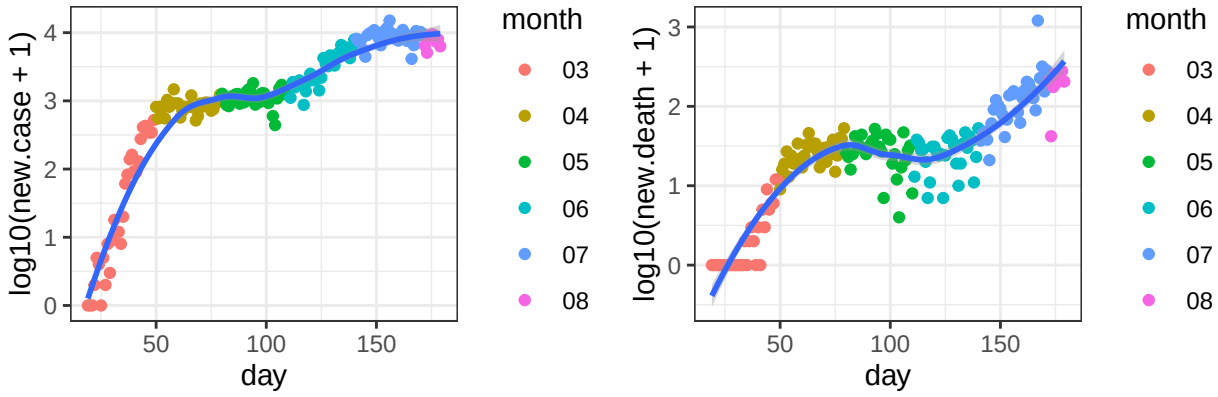
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

California



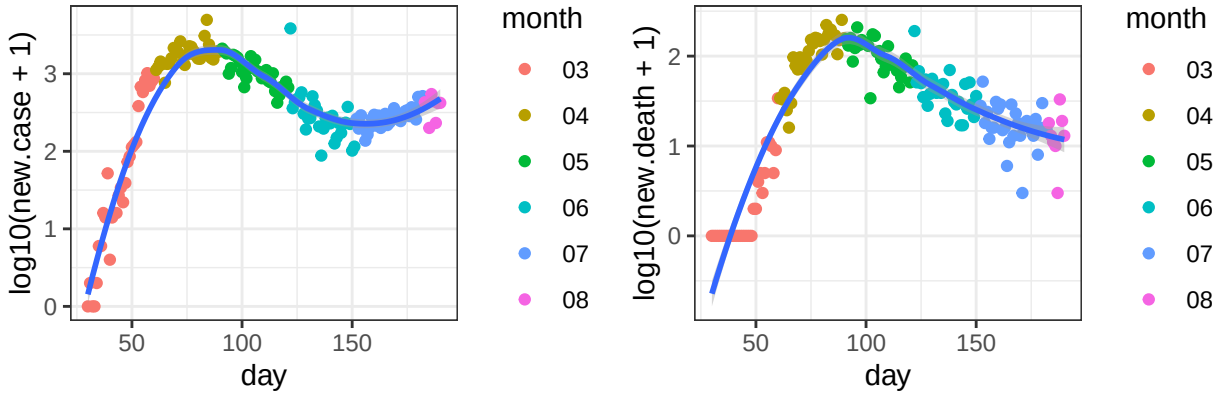
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Texas



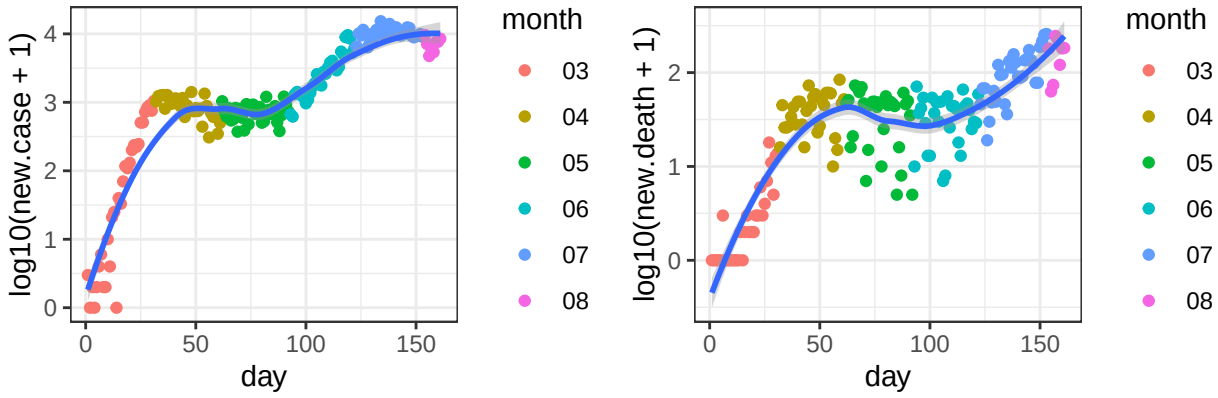
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Massachusetts



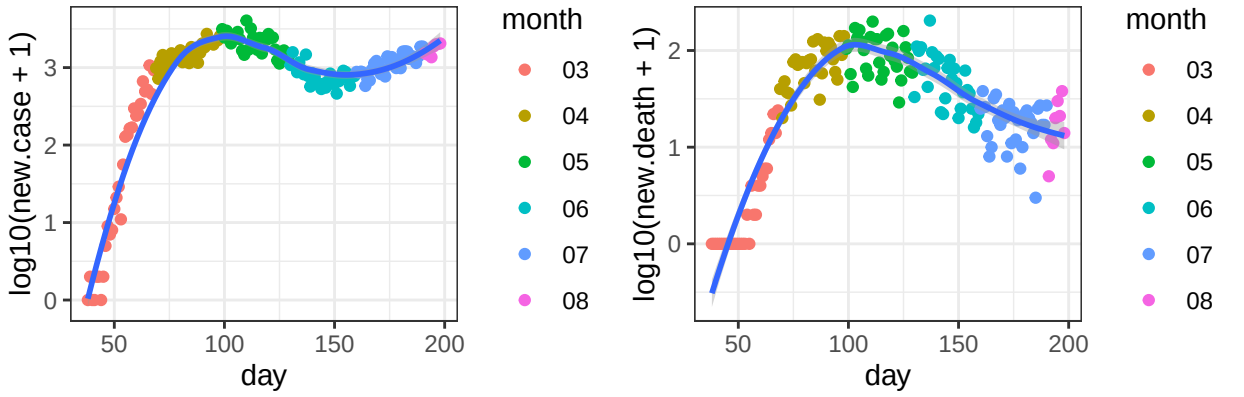
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Florida



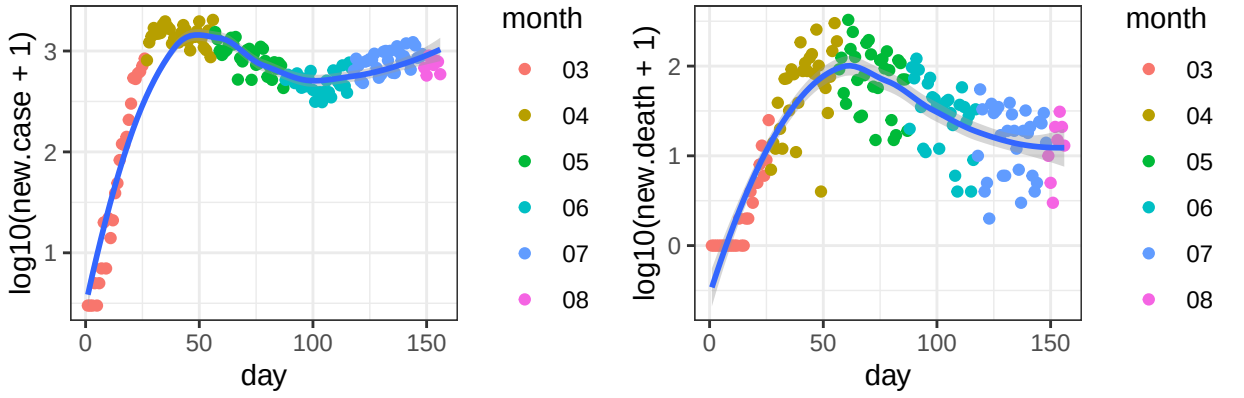
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Illinois



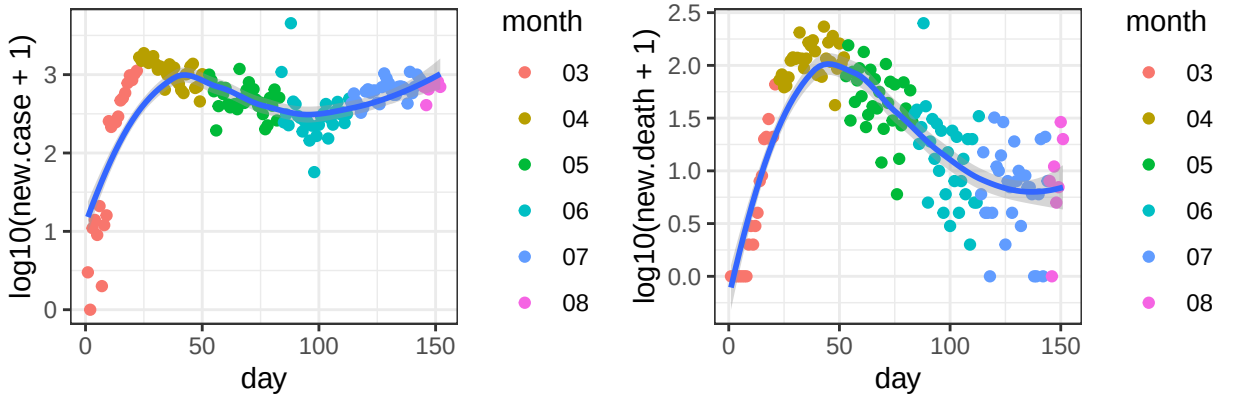
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Pennsylvania



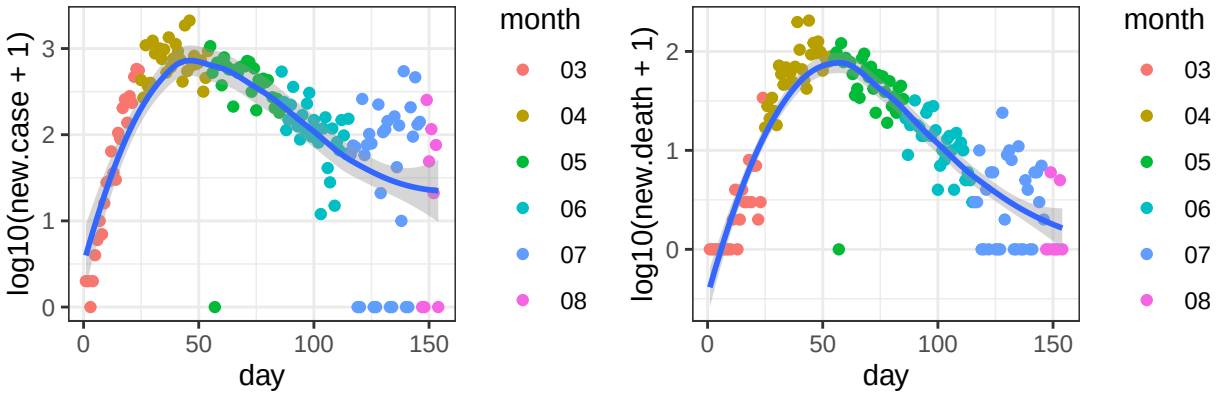
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Michigan



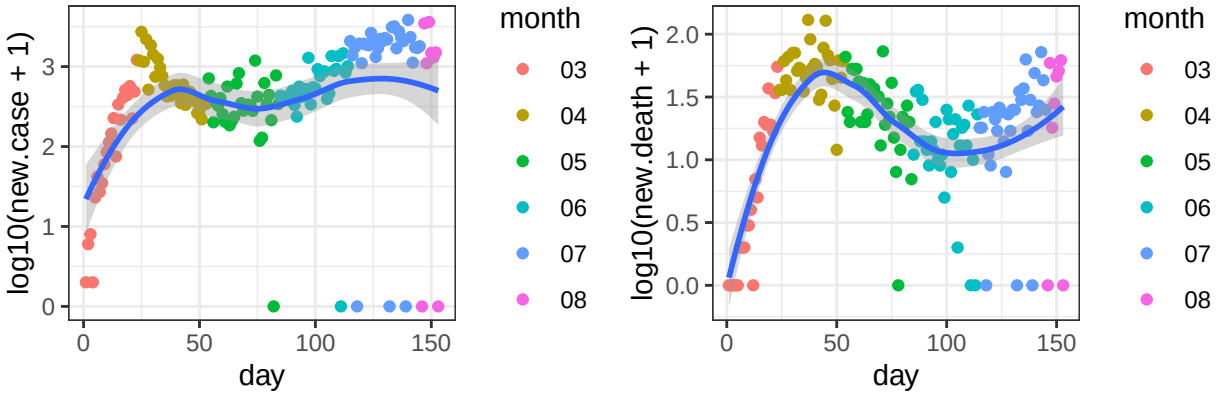
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Connecticut



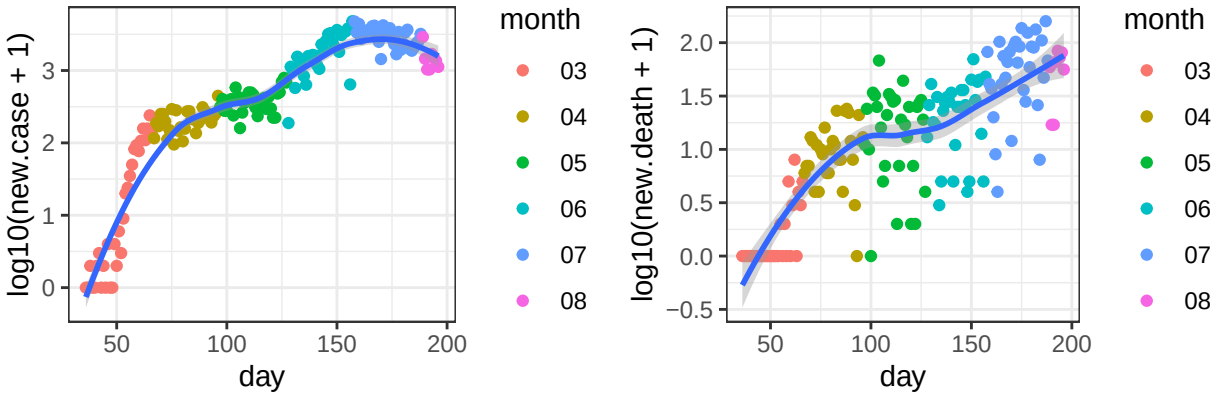
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Louisiana



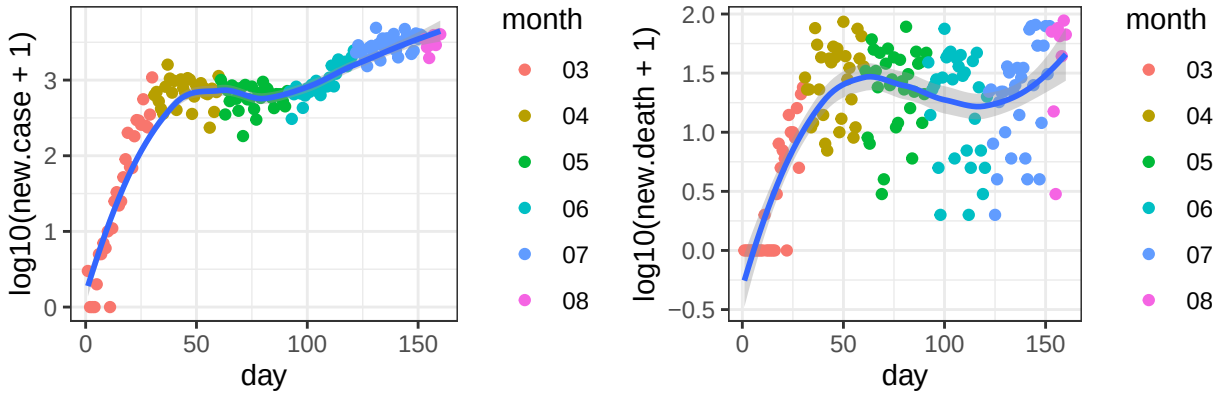
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Arizona



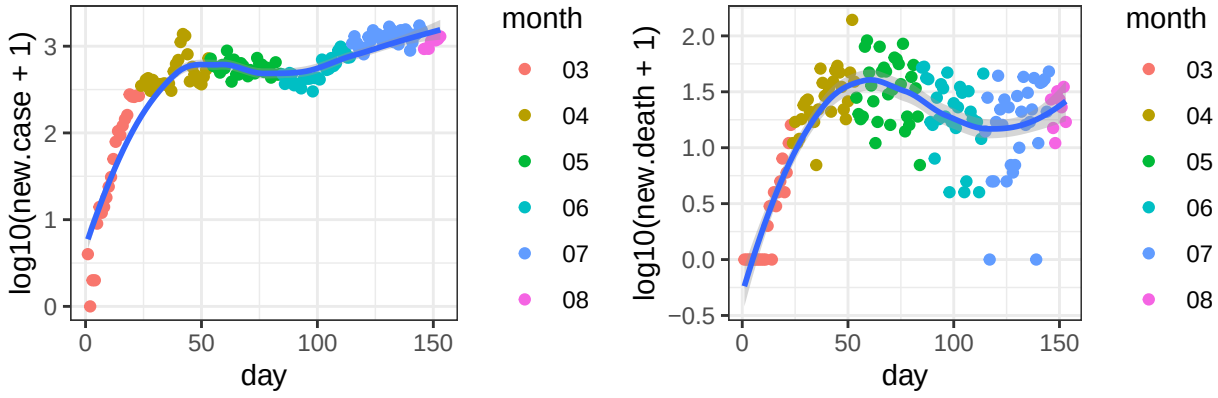
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Georgia



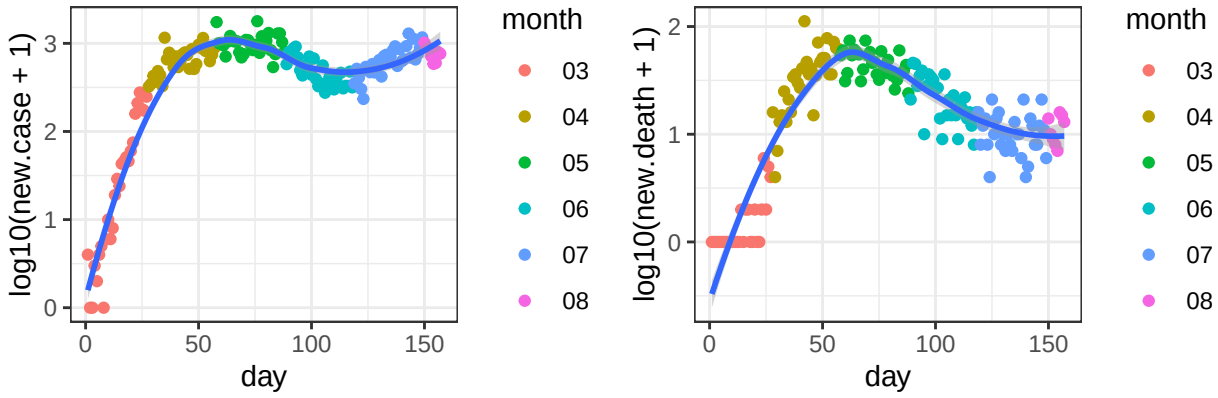
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Ohio



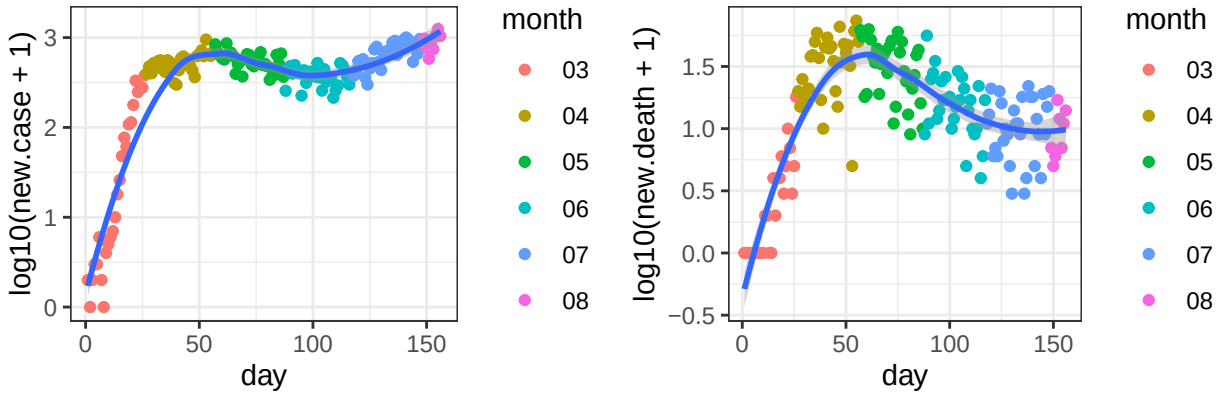
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



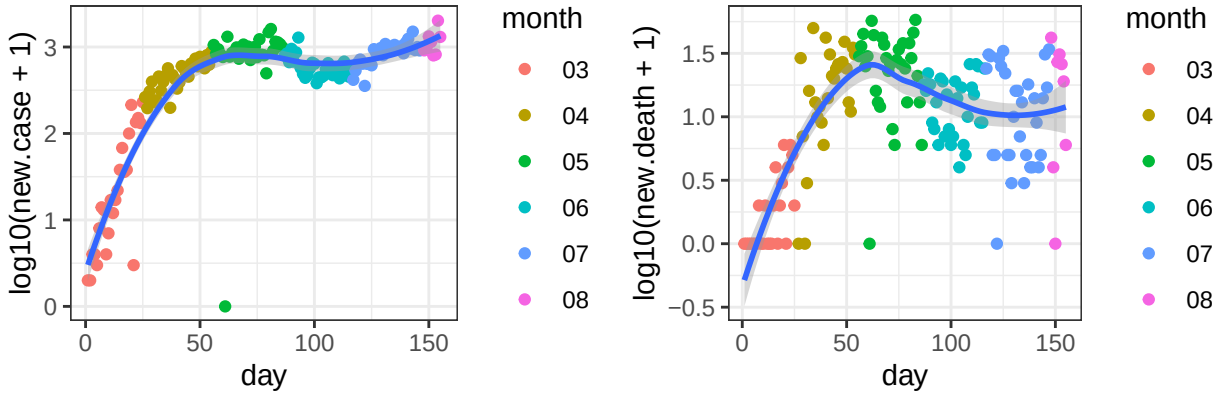
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Indiana



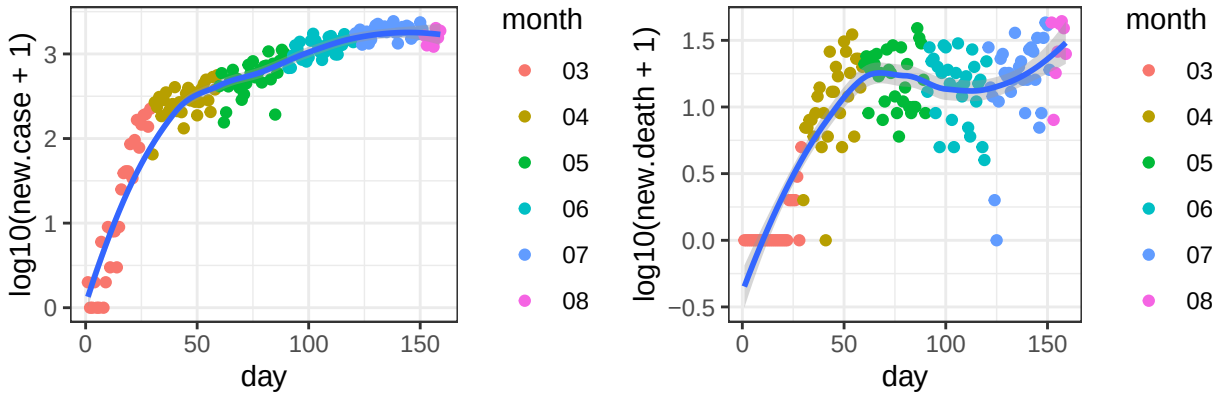
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Virginia



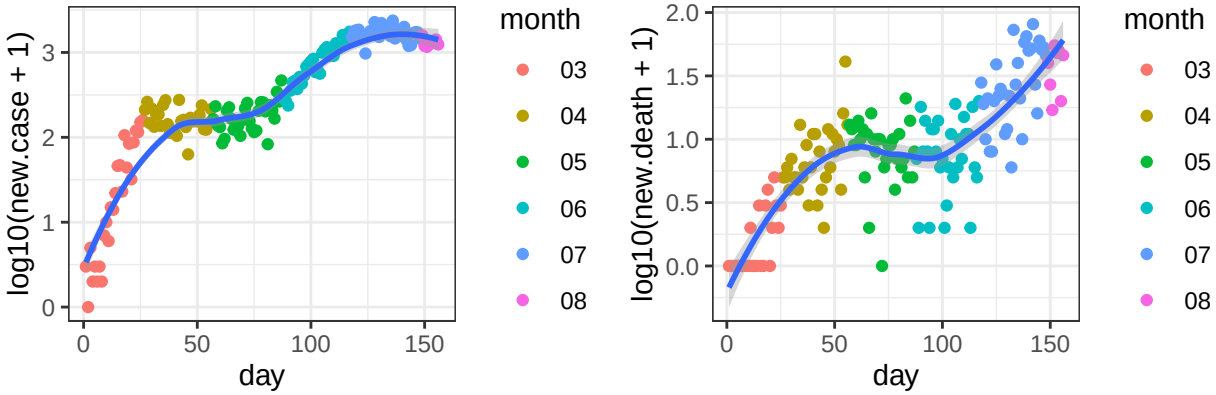
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

North Carolina



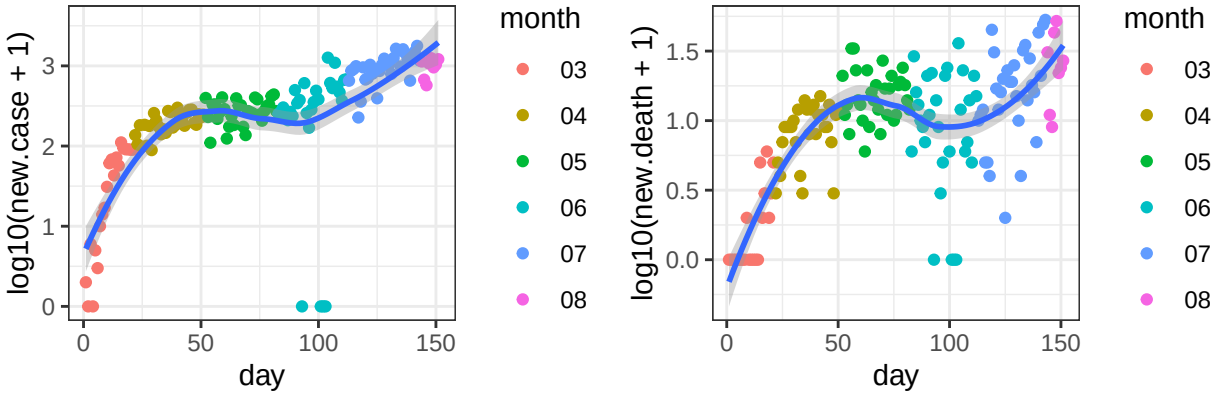
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

South Carolina



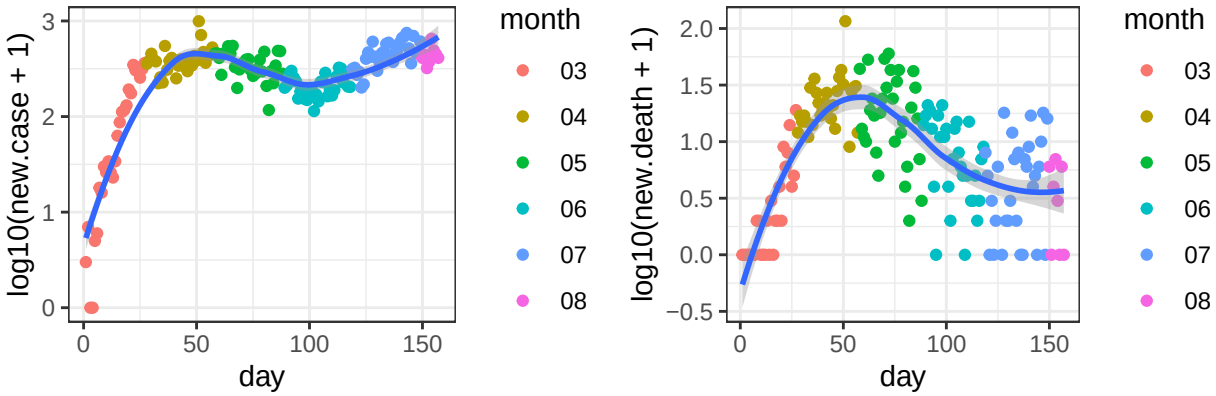
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Mississippi



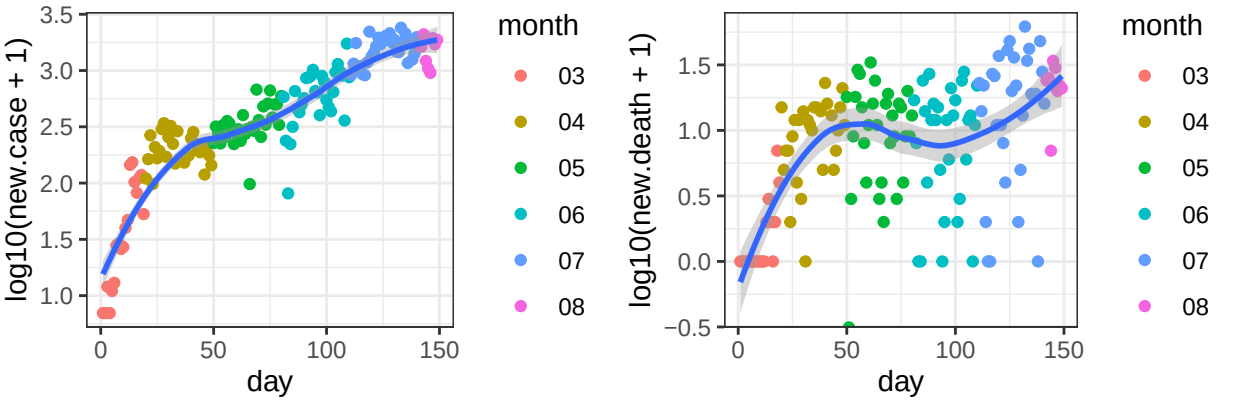
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Colorado



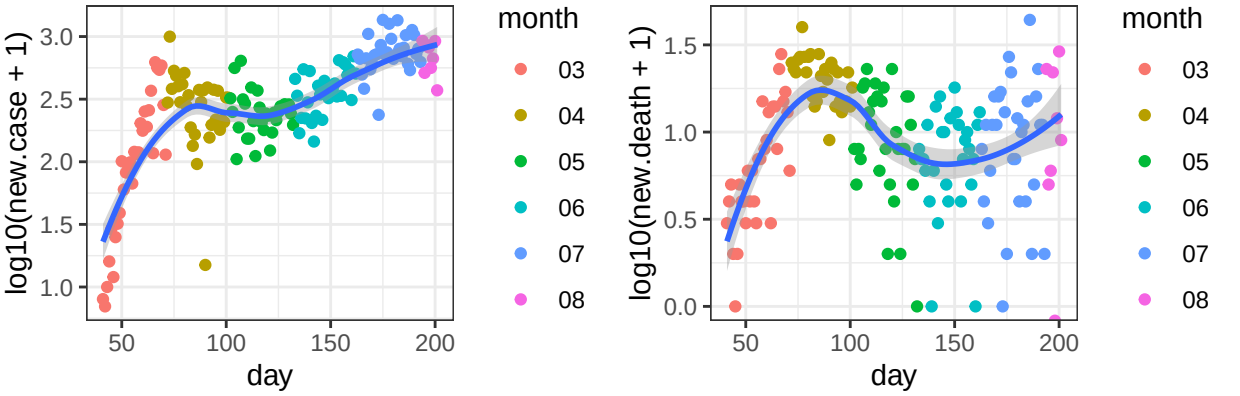
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Alabama



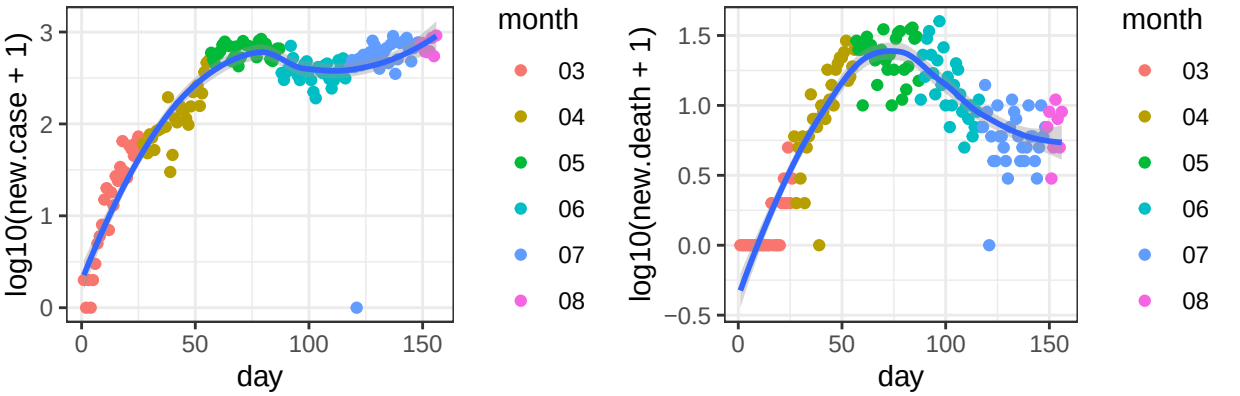
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Washington



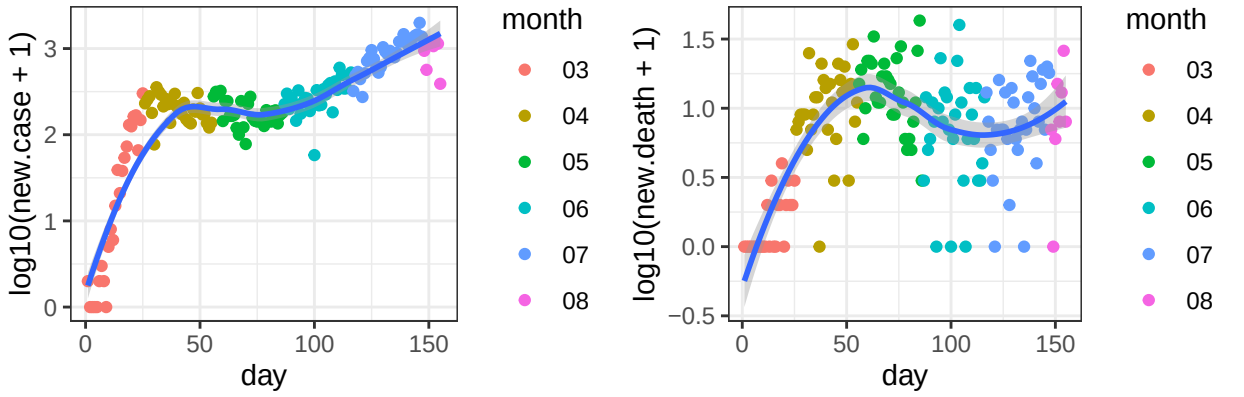
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Minnesota



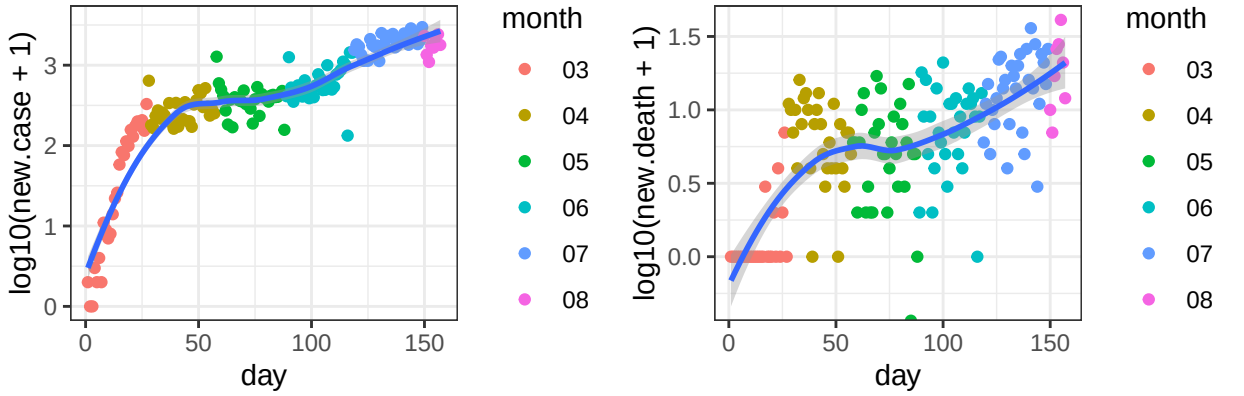
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Missouri



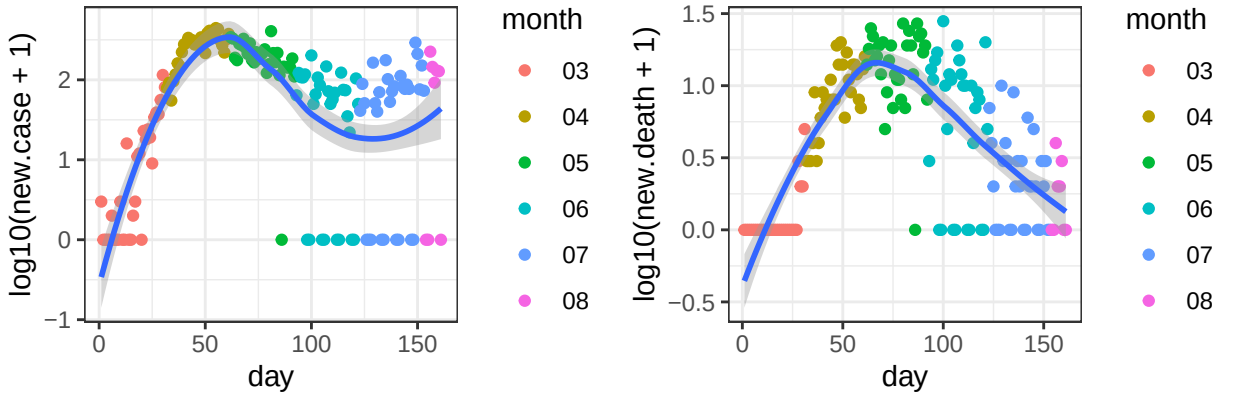
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Tennessee



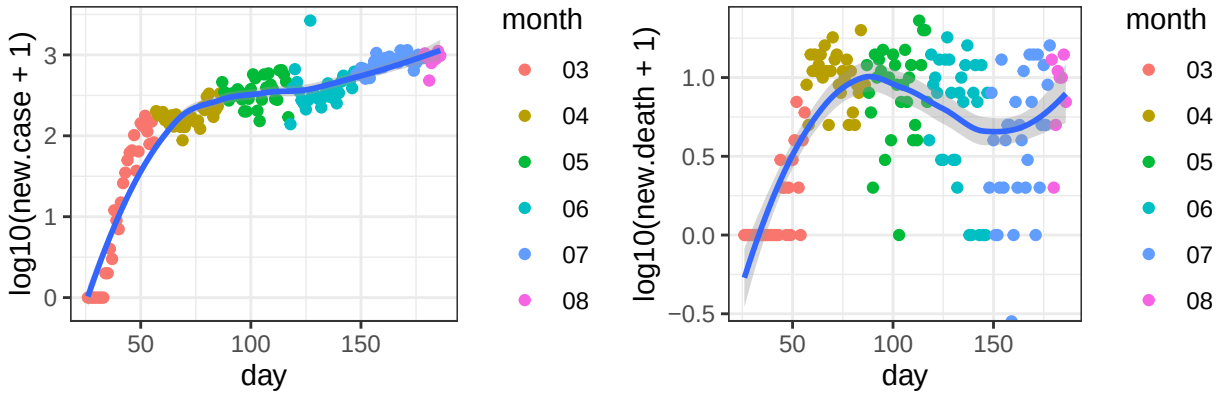
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Rhode Island



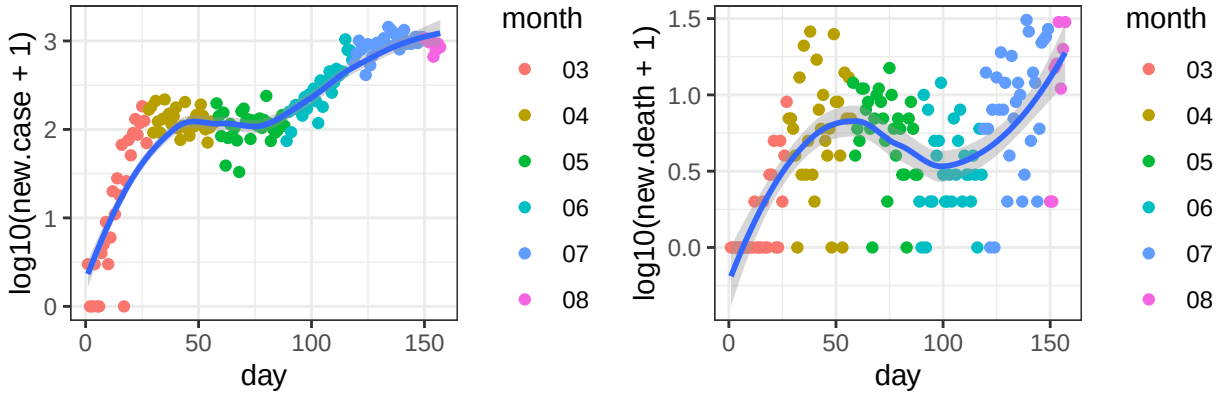
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Wisconsin



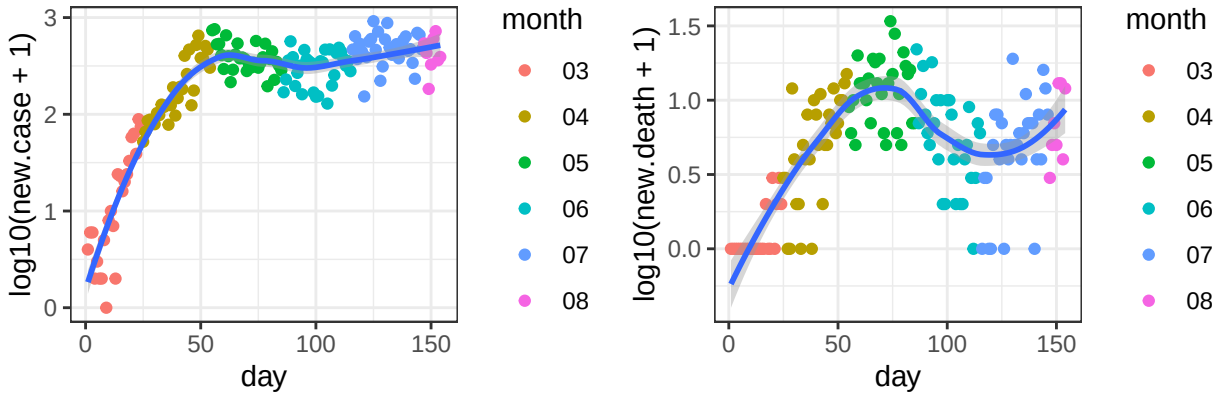
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Nevada



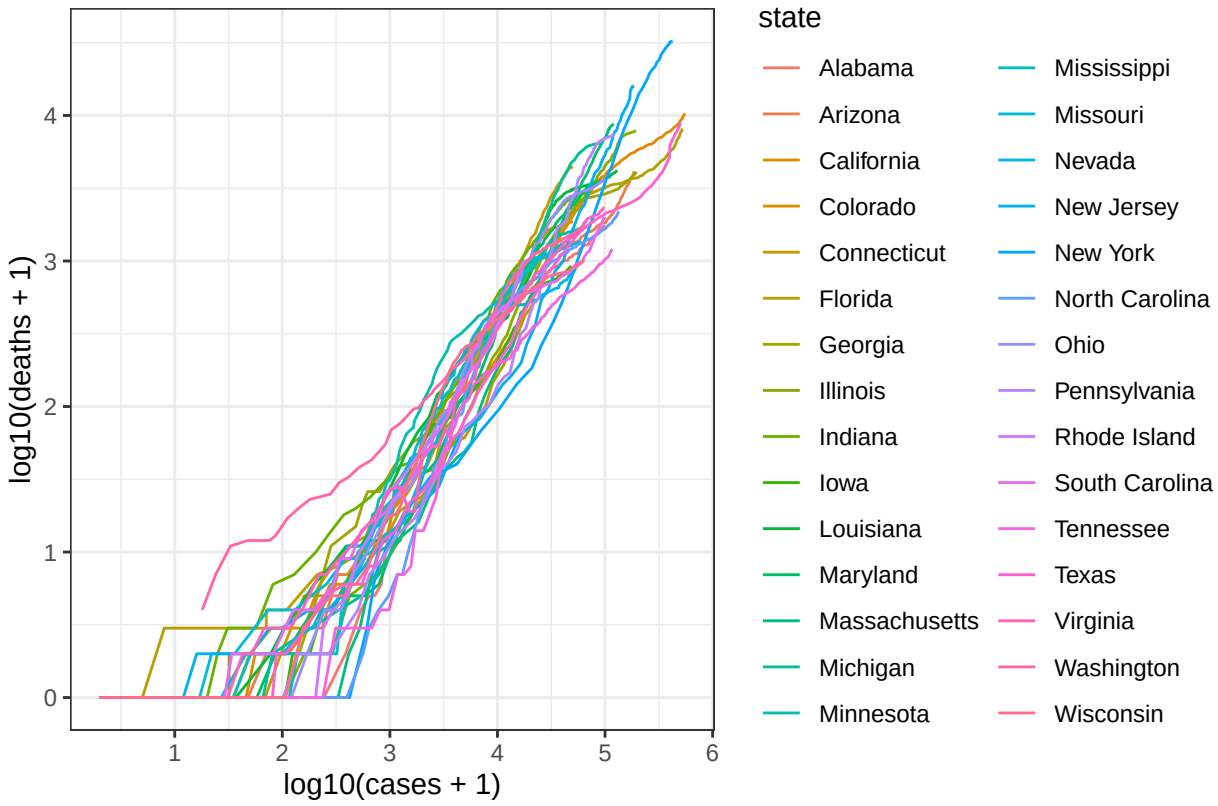
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Iowa



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

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

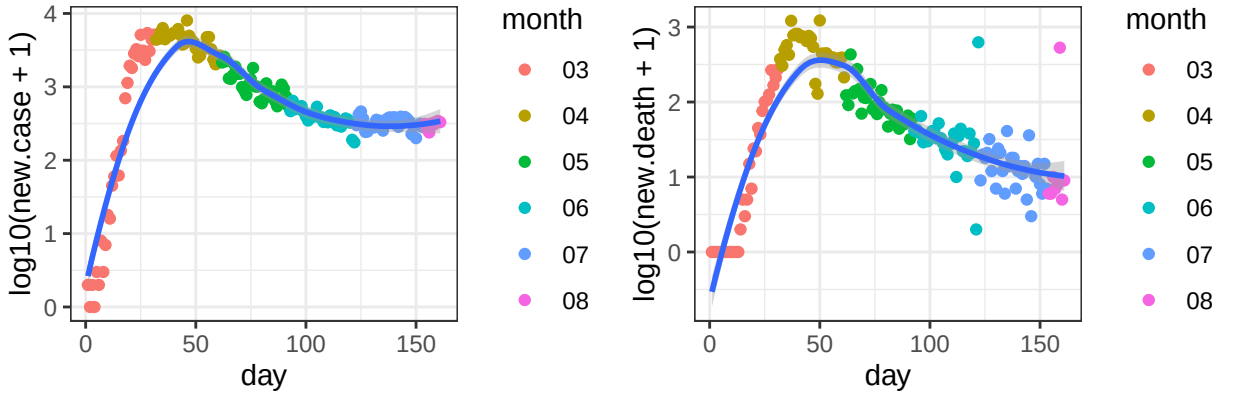
First check the 50 counties with the largest number of deaths.

```
##              date          county         state  fips  cases deaths
## 413583 2020-08-08   New York City      New York    NA 232271  23575
## 411931 2020-08-08     Los Angeles    California  6037 206761   4967
## 412338 2020-08-08            Cook      Illinois 17031 110865   4924
## 413045 2020-08-08           Wayne      Michigan 26163  28115   2828
## 411829 2020-08-08        Maricopa       Arizona  4013 125545   2347
## 413582 2020-08-08          Nassau      New York 36059  43628   2194
## 413506 2020-08-08           Essex    New Jersey 34013  19954   2111
## 413501 2020-08-08          Bergen    New Jersey 34003  21002   2038
## 412956 2020-08-08       Middlesex Massachusetts 25017  26345   2002
## 413602 2020-08-08         Suffolk      New York 36103  43749   1998
## 412090 2020-08-08      Miami-Dade       Florida 12086 131216   1838
## 414019 2020-08-08    Philadelphia  Pennsylvania 42101  31120   1703
## 414427 2020-08-08          Harris         Texas 48201  84600   1562
## 413508 2020-08-08          Hudson    New Jersey 34017  19808   1504
## 413610 2020-08-08     Westchester      New York 36119  36179   1447
## 412035 2020-08-08        Hartford   Connecticut  9003  12761   1415
## 412034 2020-08-08       Fairfield   Connecticut  9001  17956   1408
## 413511 2020-08-08       Middlesex    New Jersey 34023  18088   1403
## 413519 2020-08-08           Union    New Jersey 34039  16833   1350
## 413515 2020-08-08         Passaic    New Jersey 34031  17771   1239
## 412952 2020-08-08           Essex Massachusetts 25009  17789   1192
## 413025 2020-08-08         Oakland      Michigan 26125  15427   1129
```

```
## 412038 2020-08-08       New Haven   Connecticut  9009  13145    1104
## 412960 2020-08-08         Suffolk Massachusetts 25025  21778    1073
## 413514 2020-08-08           Ocean    New Jersey 34029  10652    1018
## 412962 2020-08-08       Worcester Massachusetts 25027  13602    1000
## 412958 2020-08-08         Norfolk Massachusetts 25021  10601     997
## 413012 2020-08-08          Macomb      Michigan 26099  10580     951
## 412097 2020-08-08      Palm Beach       Florida 12099  36598     929
## 414014 2020-08-08      Montgomery  Pennsylvania 42091  10077     857
## 413512 2020-08-08        Monmouth    New Jersey 34025  10381     856
## 413073 2020-08-08        Hennepin     Minnesota 27053  19057     831
## 413513 2020-08-08          Morris    New Jersey 34027   7383     828
## 414118 2020-08-08      Providence  Rhode Island 44007  15047     813
## 412938 2020-08-08      Montgomery      Maryland 24031  18299     802
## 411945 2020-08-08       Riverside    California  6065  40902     799
## 413475 2020-08-08           Clark        Nevada 32003  47739     799
## 414434 2020-08-08         Hidalgo         Texas 48215  19534     790
## 412053 2020-08-08         Broward       Florida 12011  61614     789
## 412474 2020-08-08          Marion       Indiana 18097  15849     773
## 414342 2020-08-08           Bexar         Texas 48029  42543     773
## 412939 2020-08-08 Prince George's      Maryland 24033  24019     751
## 414383 2020-08-08          Dallas         Texas 48113  53831     751
## 413991 2020-08-08        Delaware  Pennsylvania 42045   9171     744
## 411942 2020-08-08          Orange    California  6059  39076     720
## 412959 2020-08-08        Plymouth Massachusetts 25023   9226     720
## 412954 2020-08-08         Hampden Massachusetts 25013   7582     704
## 414774 2020-08-08            King    Washington 53033  16491     694
## 413319 2020-08-08       St. Louis      Missouri 29189  14666     663
## 412950 2020-08-08         Bristol Massachusetts 25005   9324     632
```

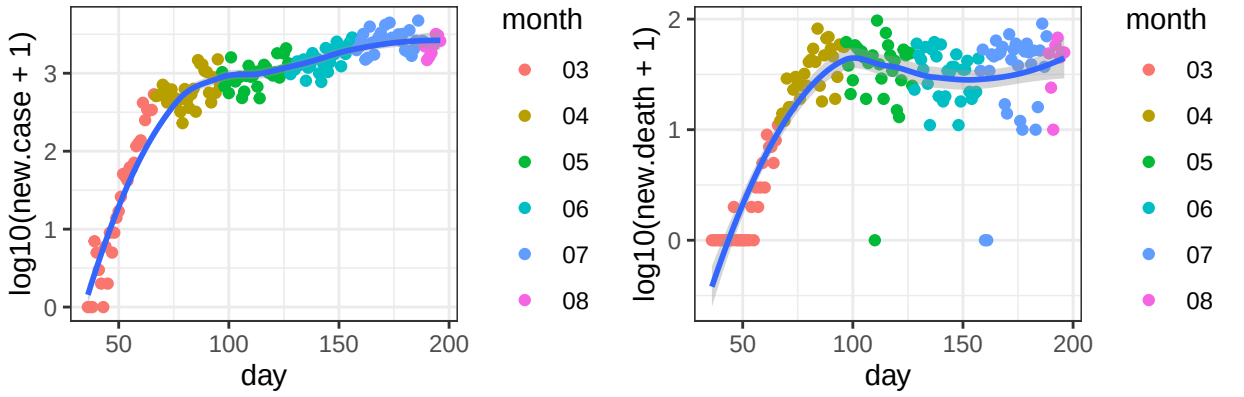
For these 50 counties, I check the number of new cases and the number of new deaths.

New York City_New York



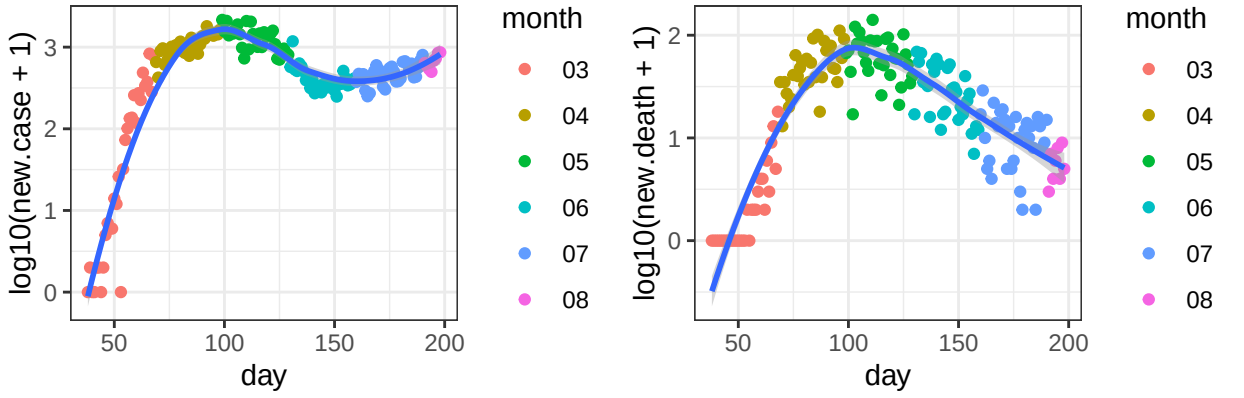
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Los Angeles_California



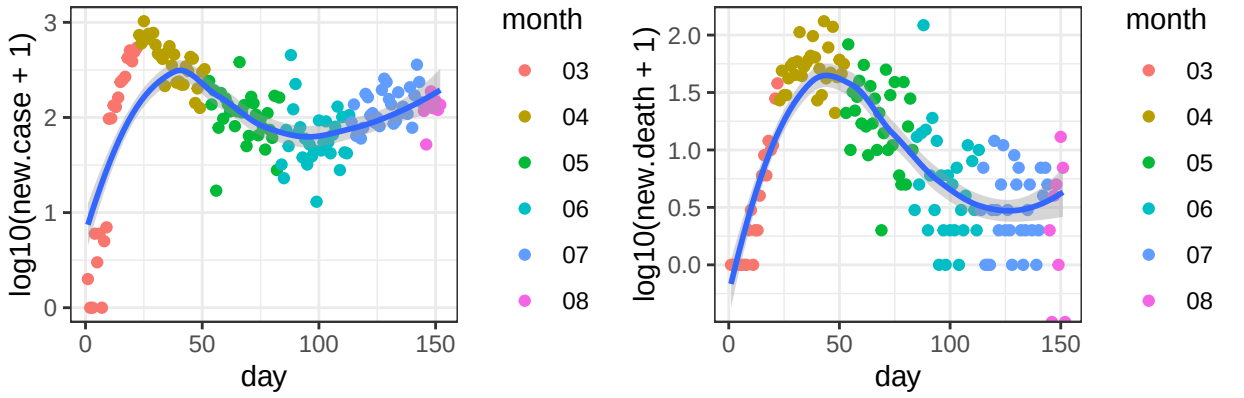
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



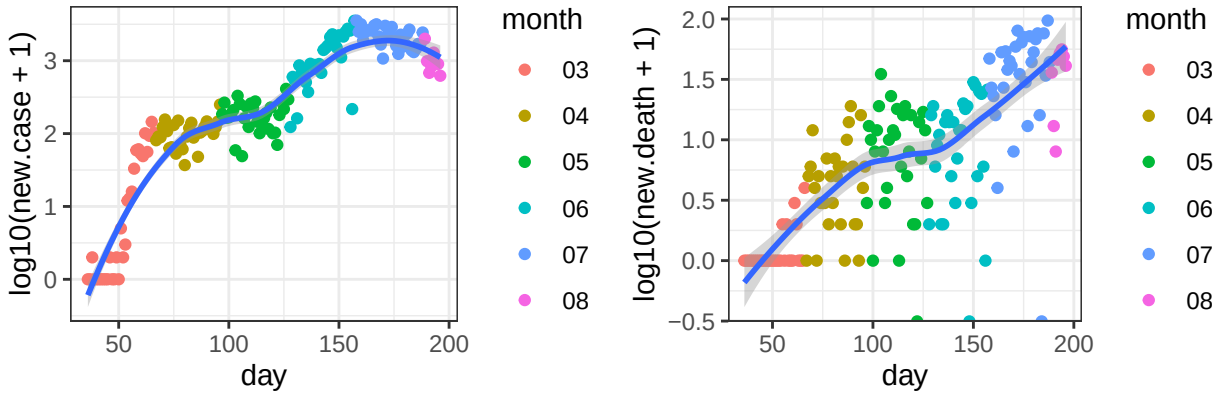
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Wayne_Michigan



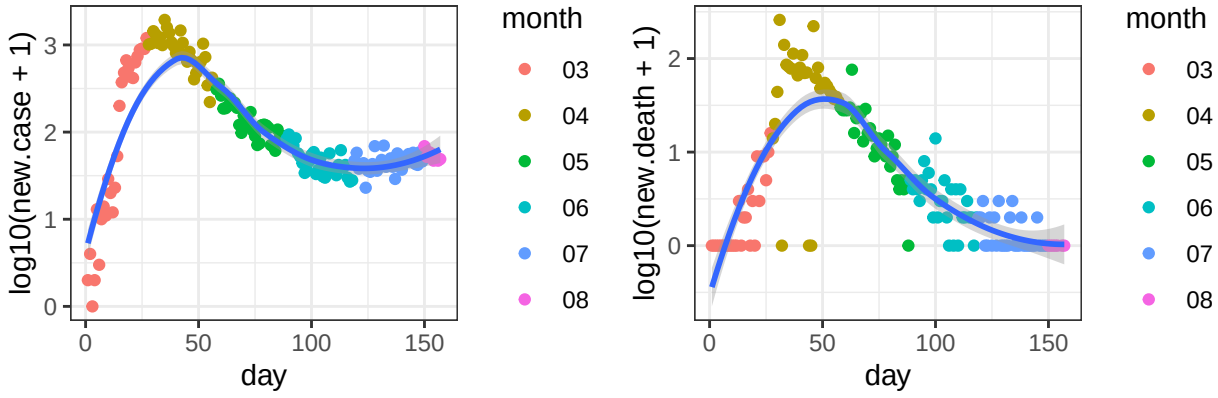
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Maricopa_Arizona



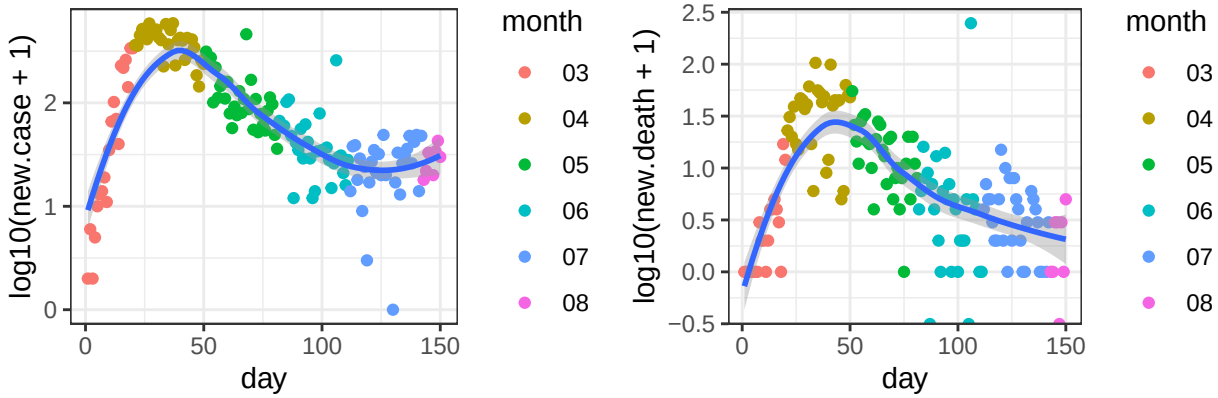
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Nassau_New York



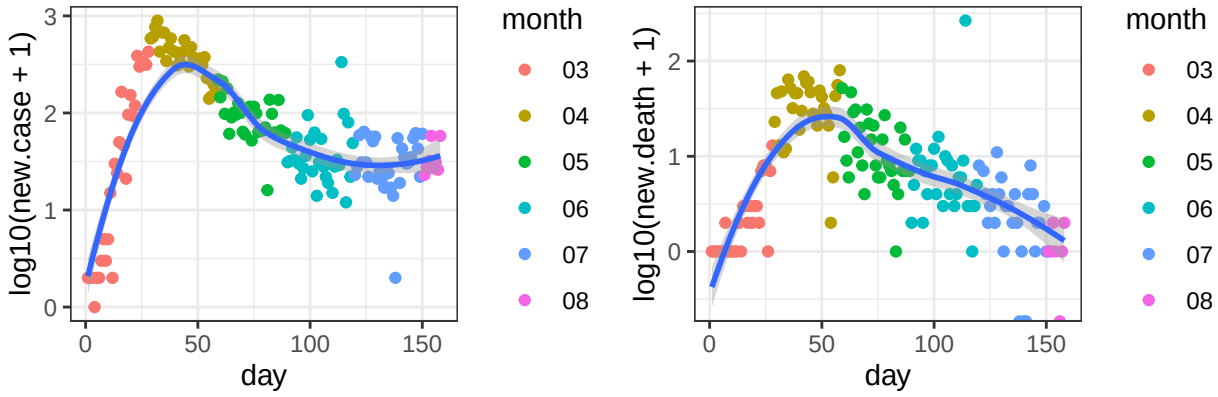
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



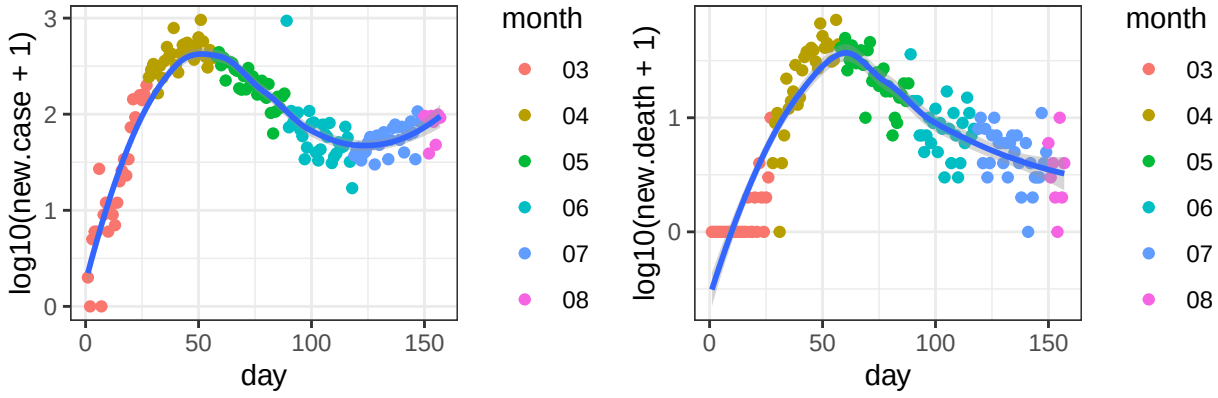
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Bergen_New Jersey



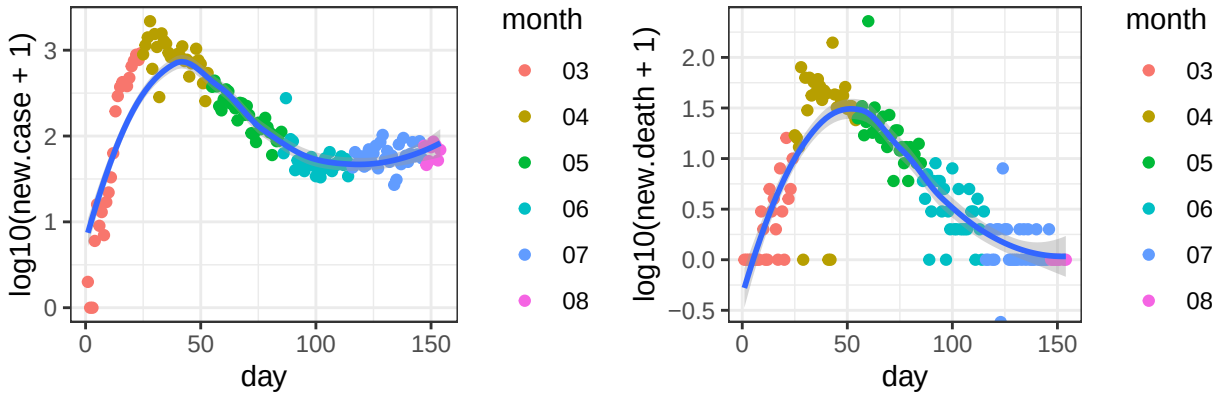
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Middlesex_Massachusetts



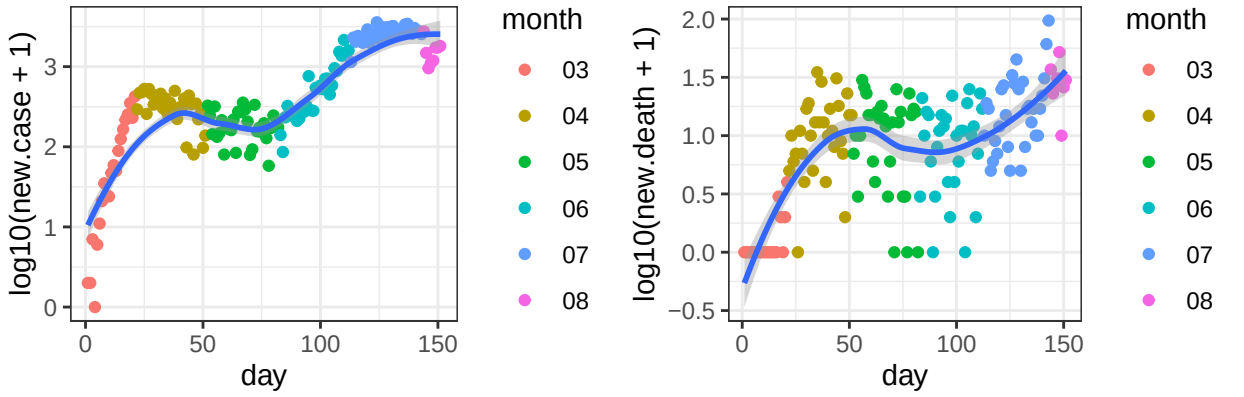
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Suffolk_New York



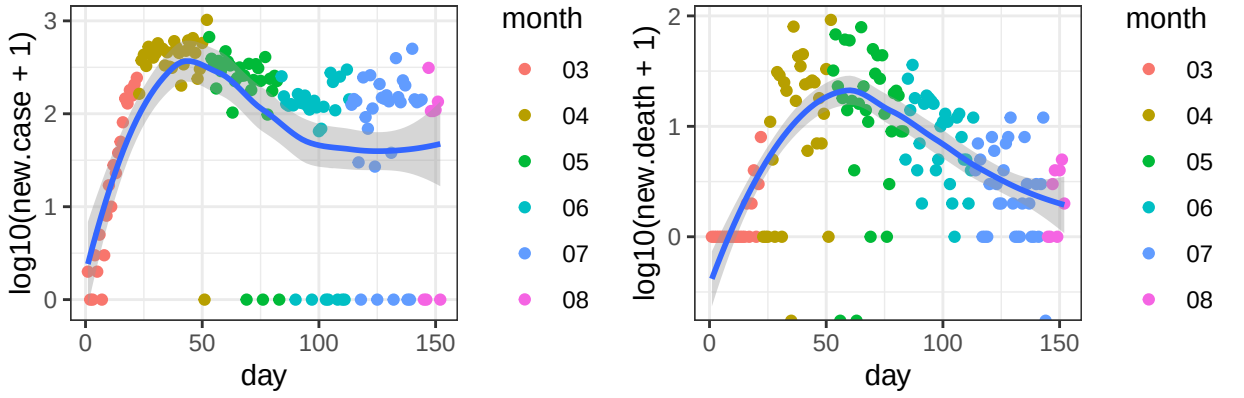
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Miami−Dade_Florida



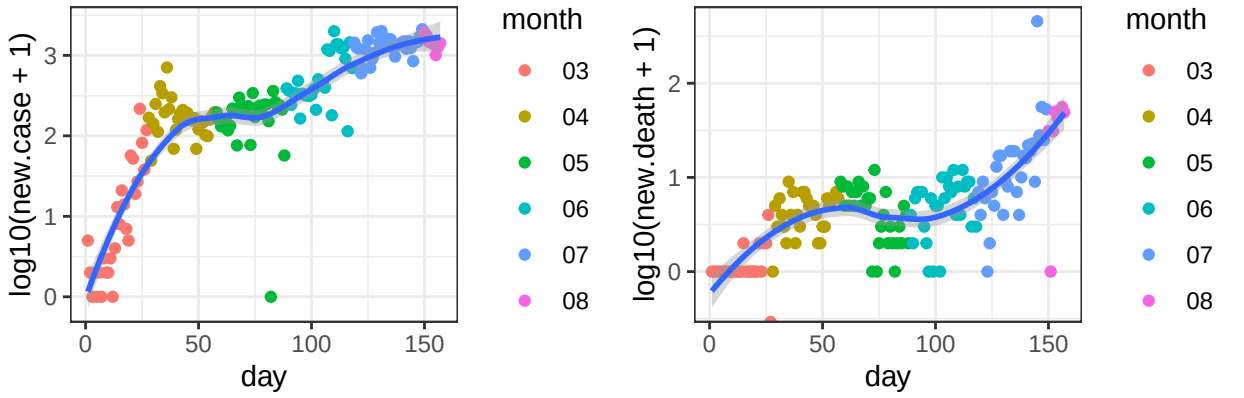
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Philadelphia_Pennsylvania



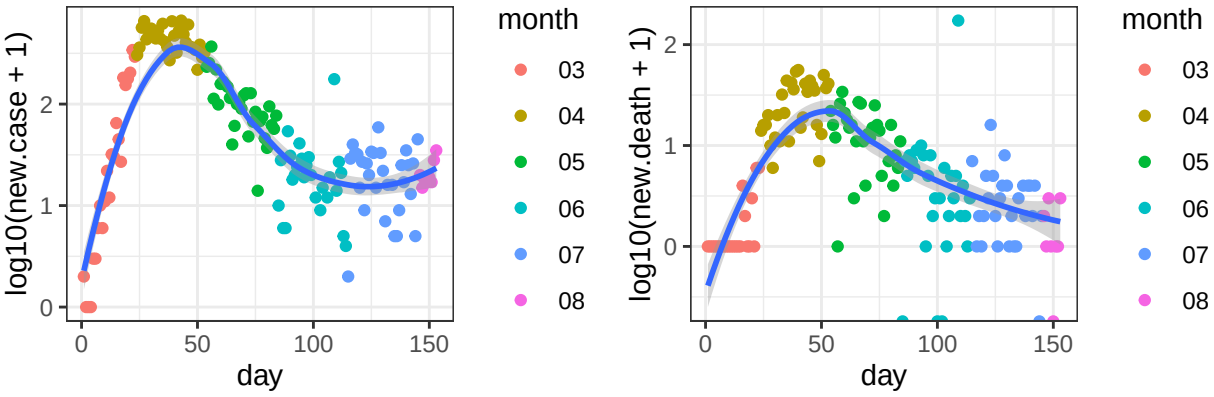
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Harris_Texas



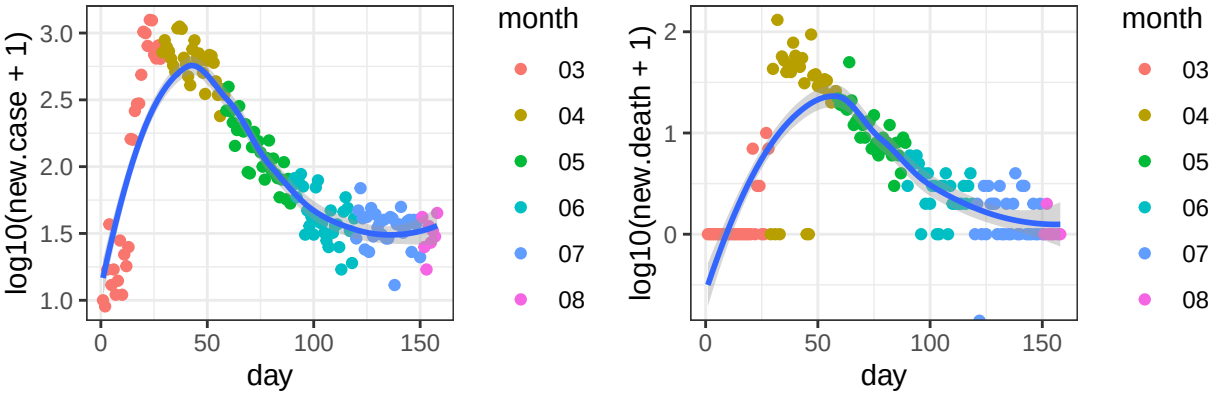
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Hudson_New Jersey



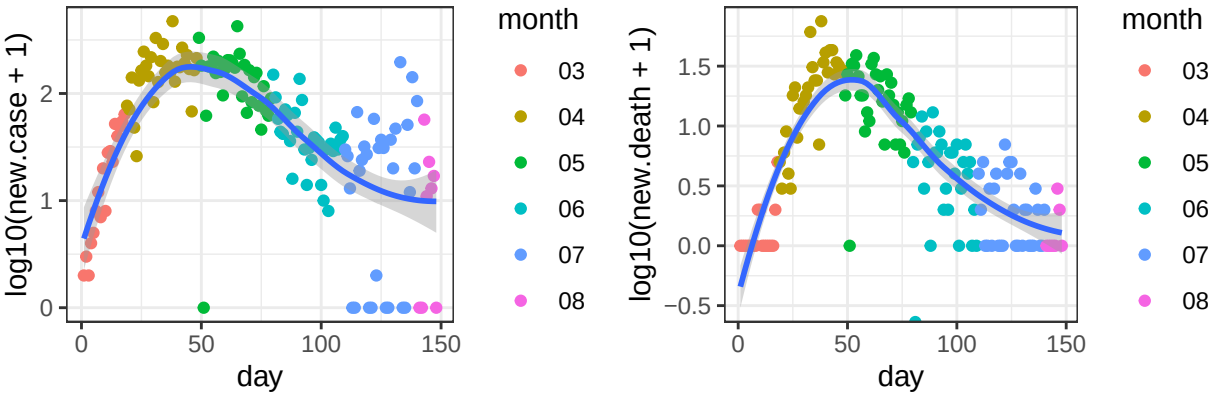
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Westchester_New York



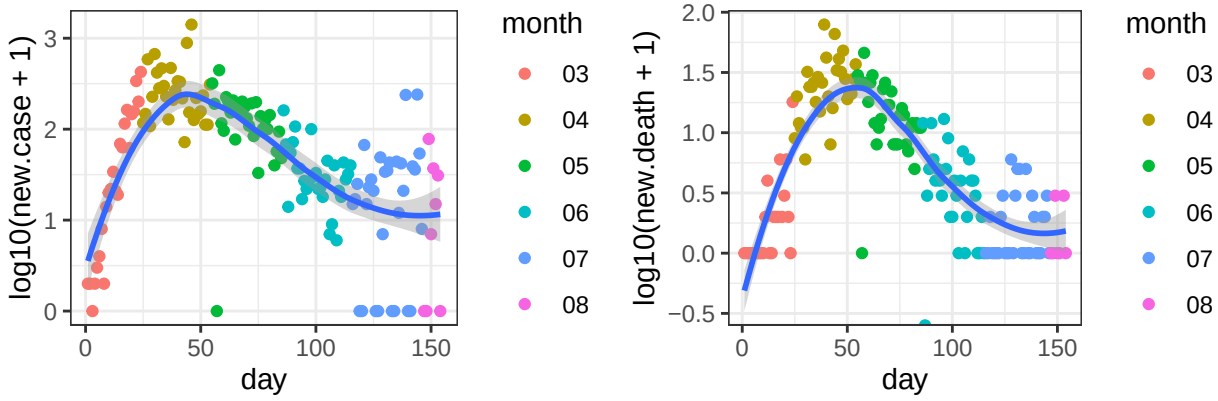
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Hartford_Connecticut



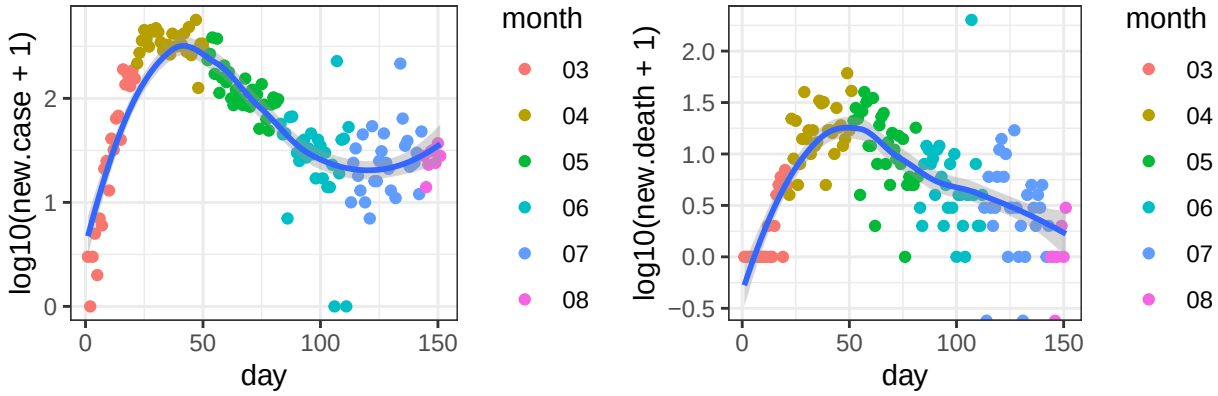
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Fairfield_Connecticut



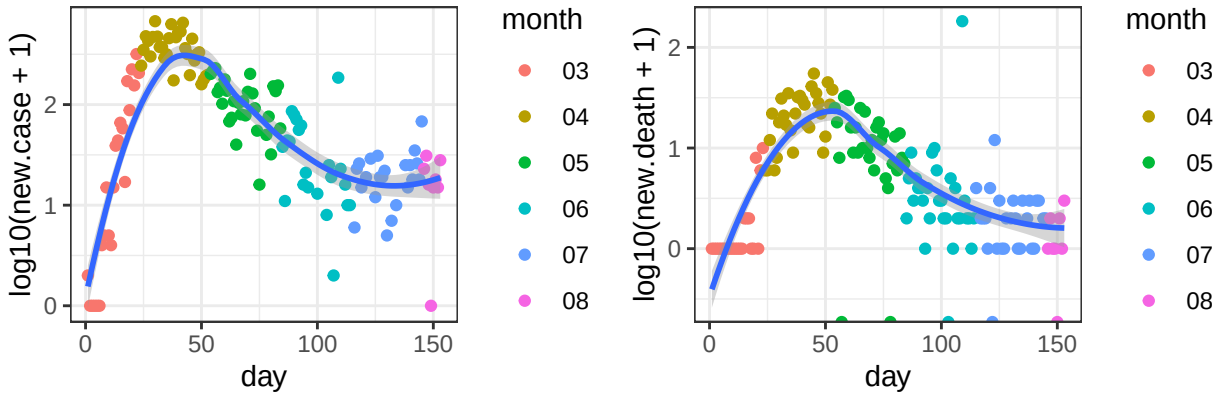
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Middlesex_New Jersey



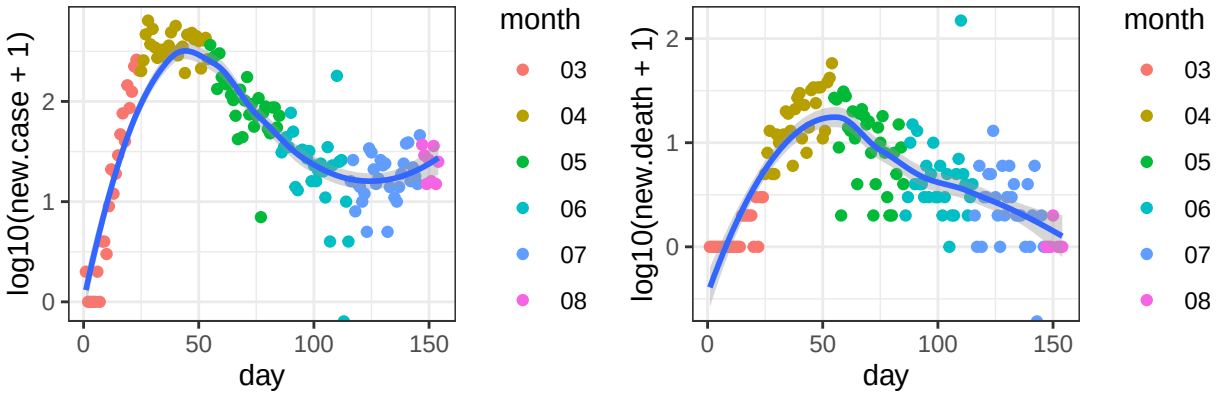
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Union_New Jersey



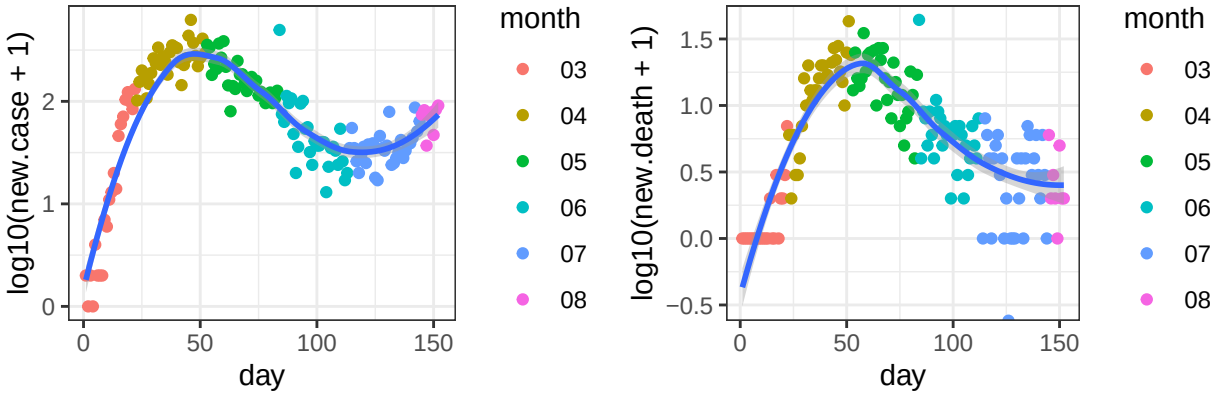
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Passaic_New Jersey



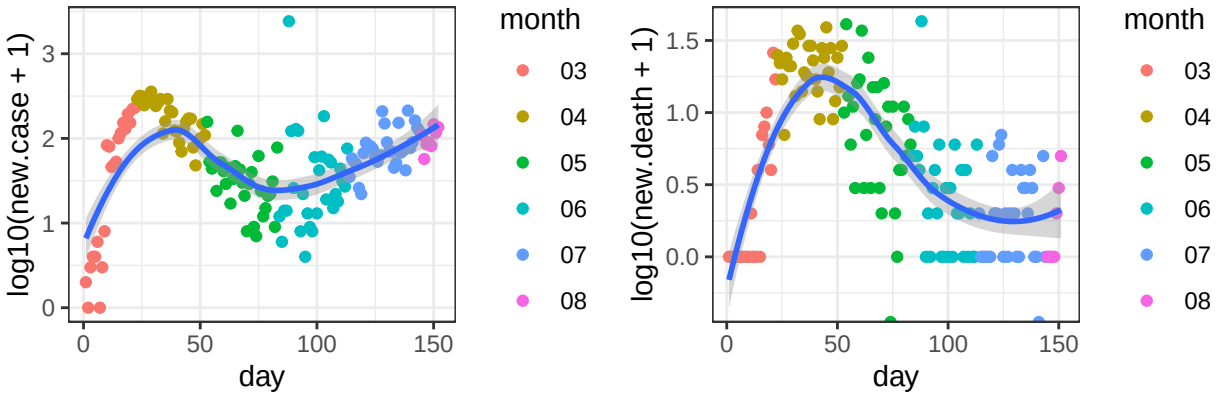
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Essex_Massachusetts



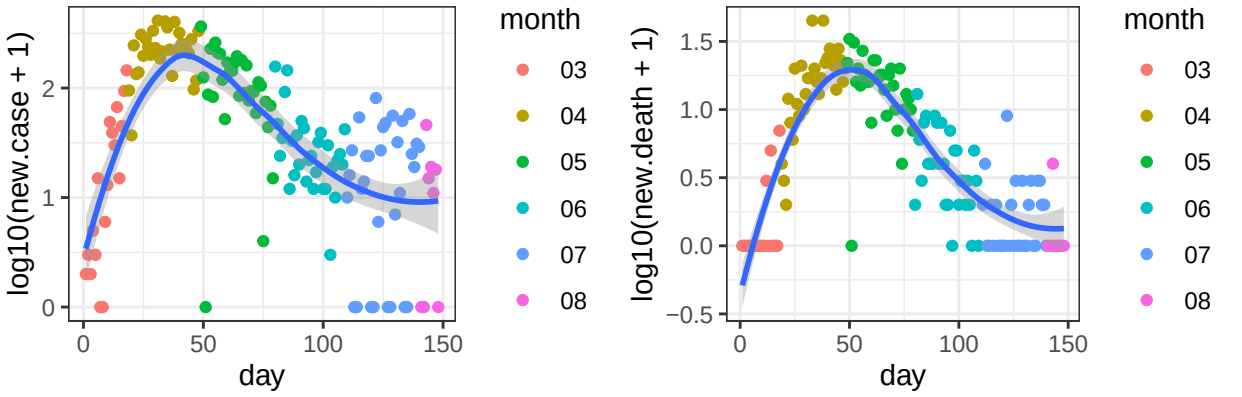
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Oakland_Michigan



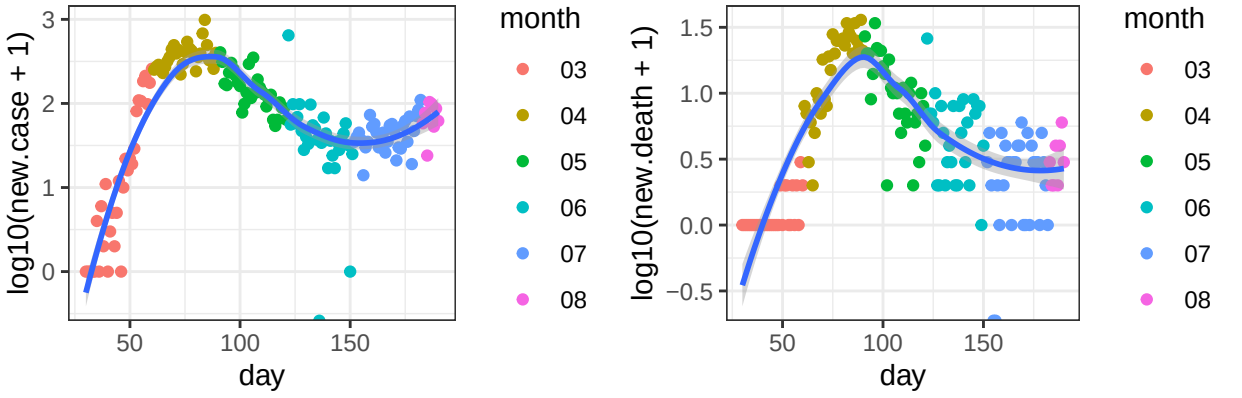
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

New Haven_Connecticut



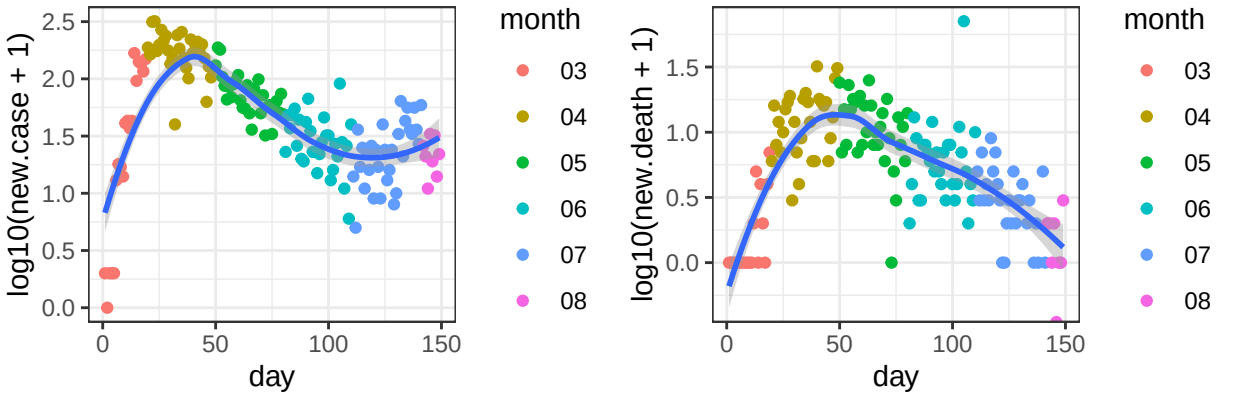
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Suffolk_Massachusetts



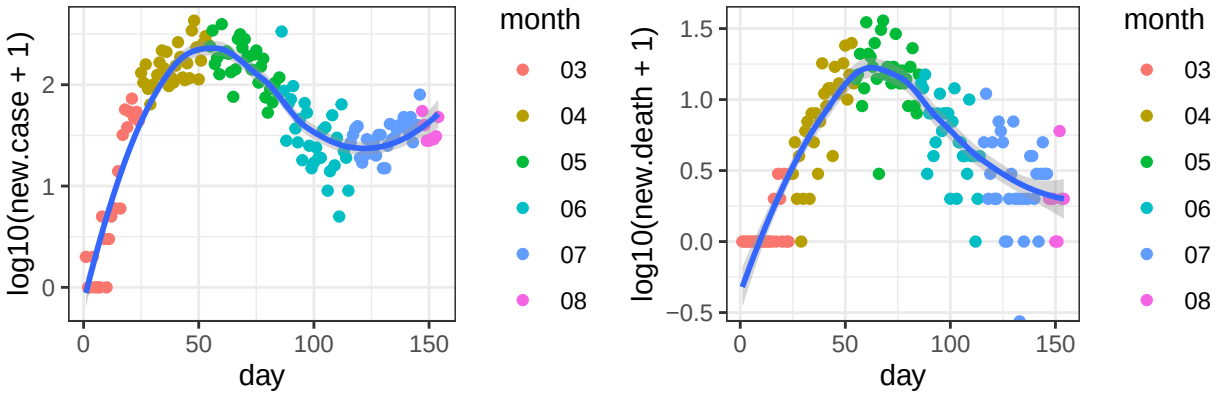
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Ocean_New Jersey



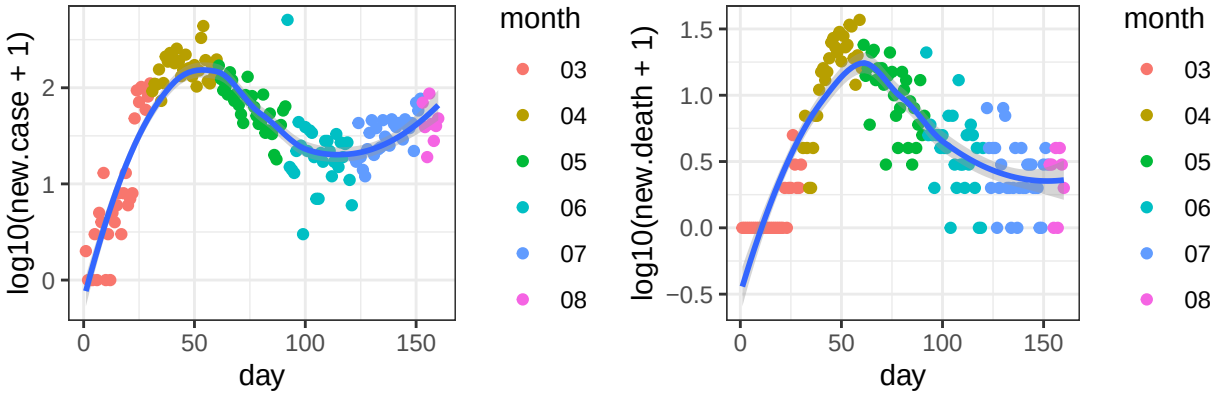
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Worcester_Massachusetts



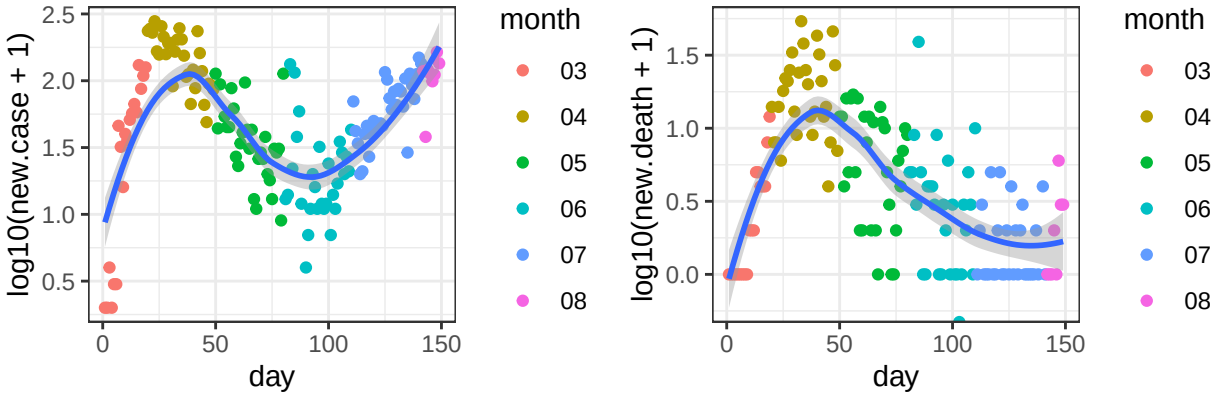
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Norfolk_Massachusetts



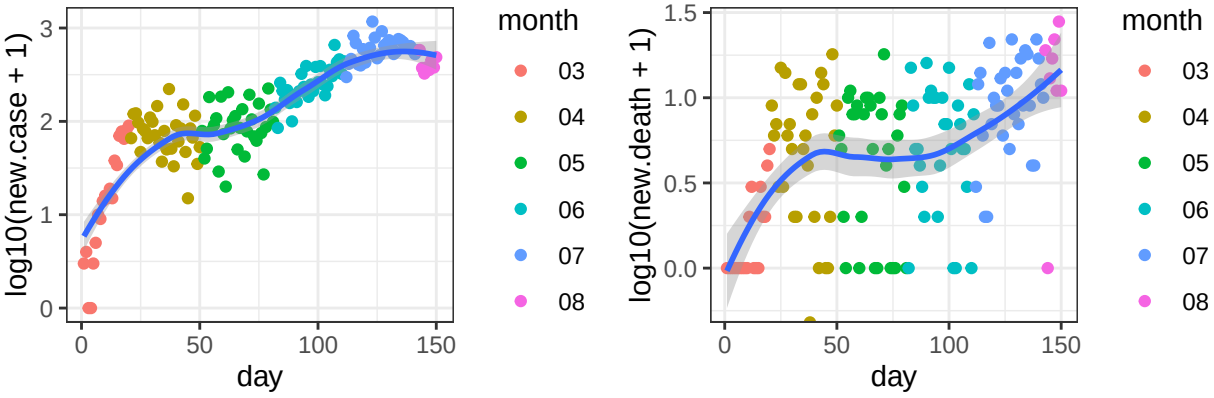
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Macomb_Michigan



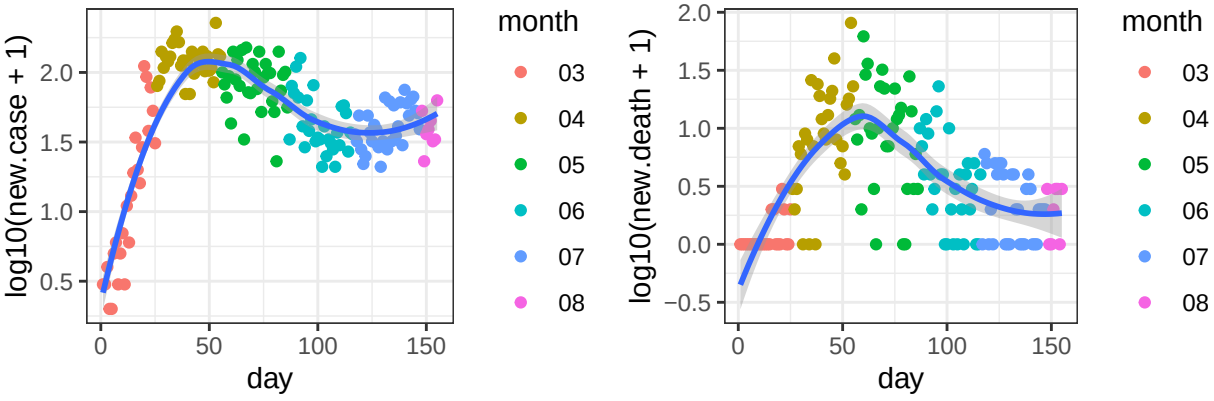
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Palm Beach_Florida



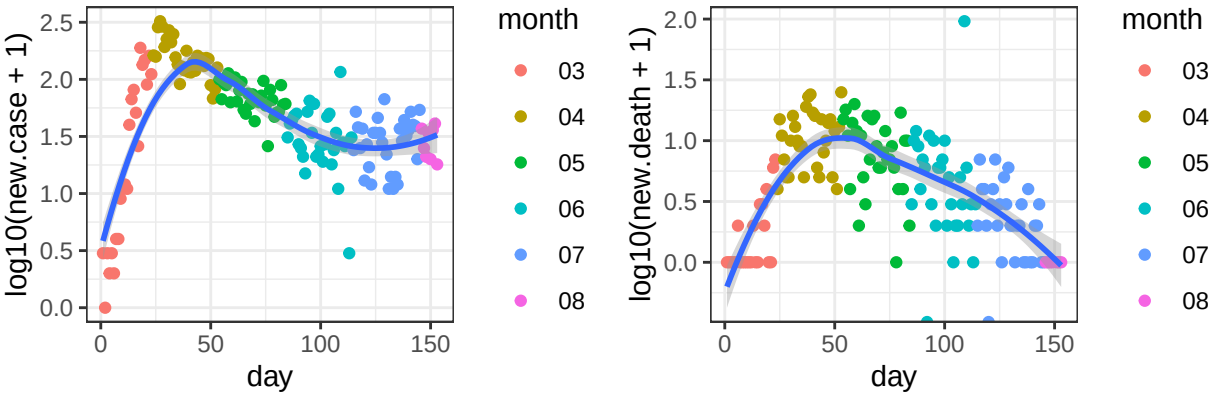
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Montgomery_Pennsylvania



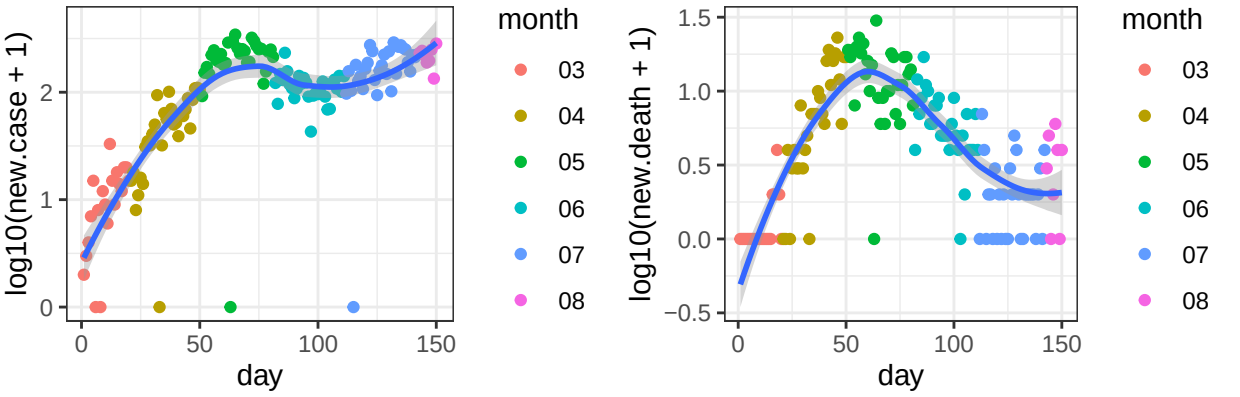
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Monmouth_New Jersey



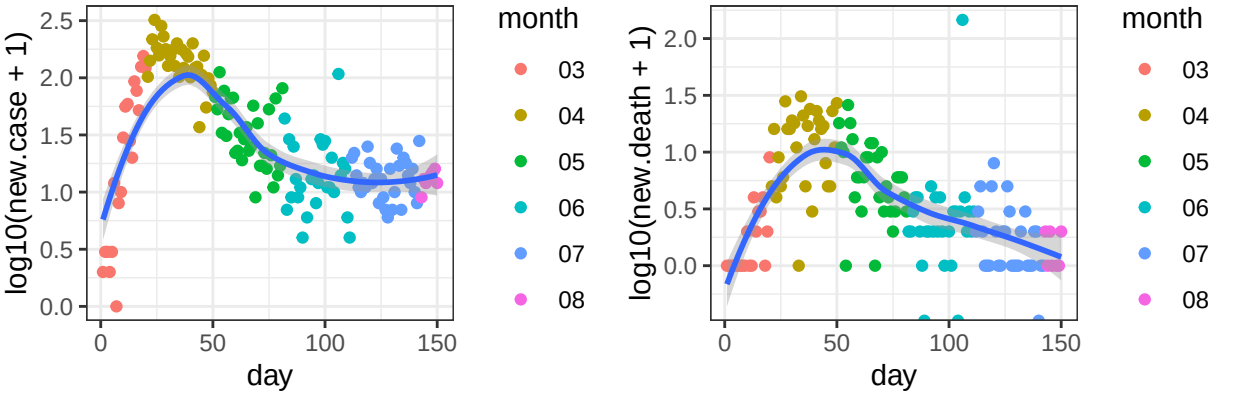
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Hennepin_Minnesota



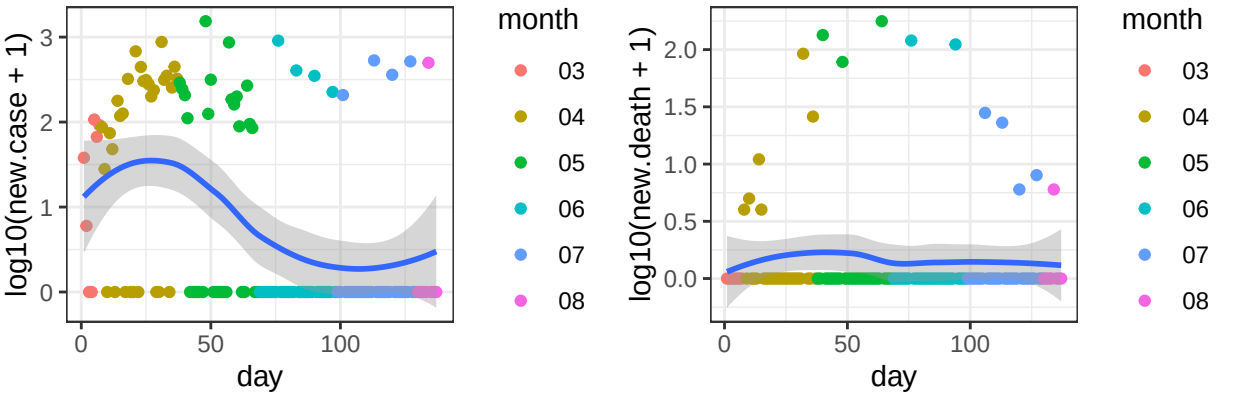
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Morris_New Jersey



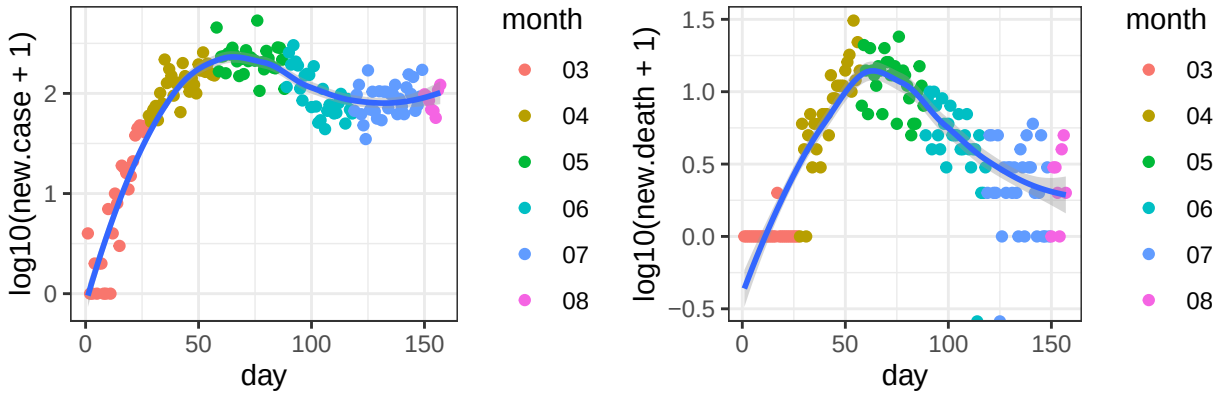
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Providence_Rhode Island



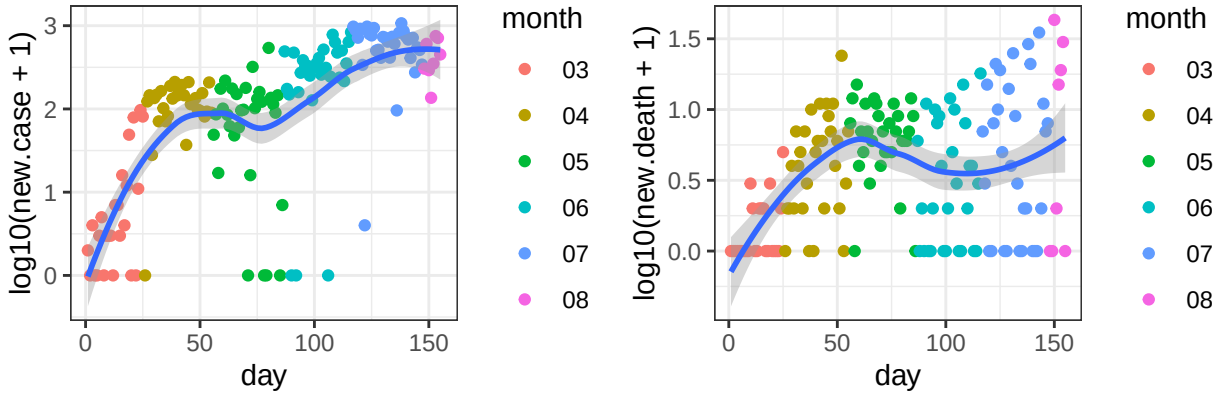
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Montgomery_Maryland



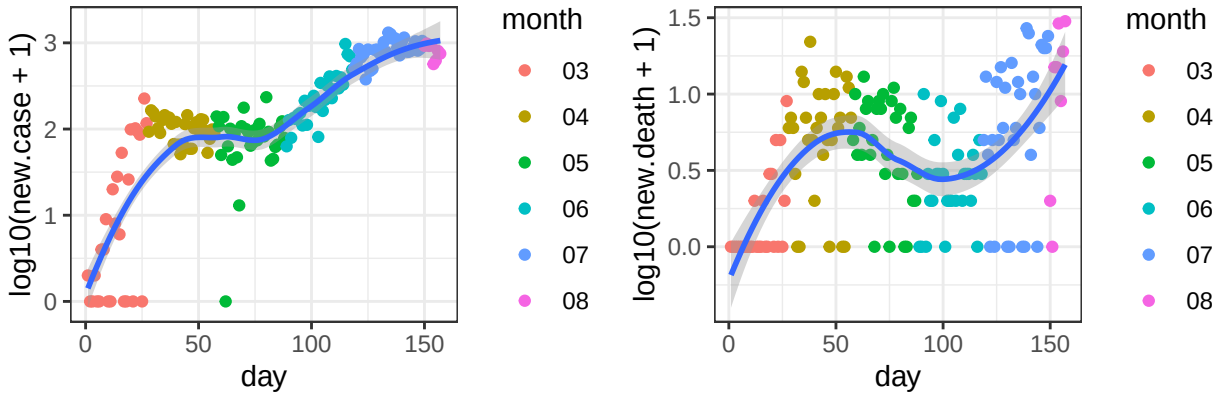
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Riverside_California



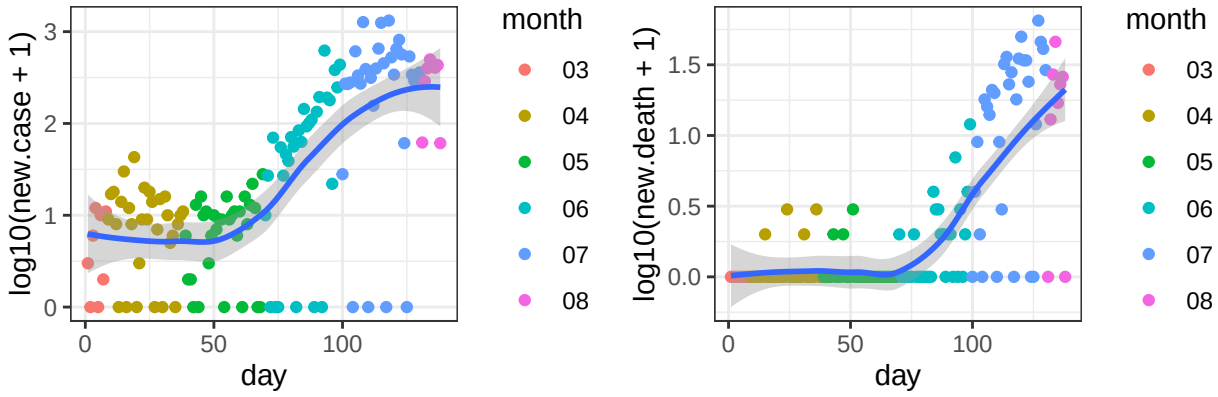
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Clark_Nevada



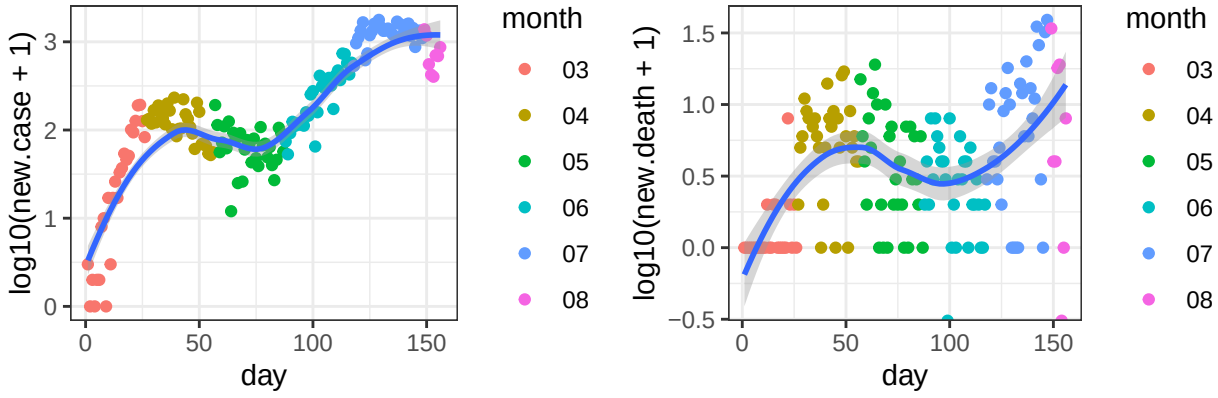
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Hidalgo_Texas



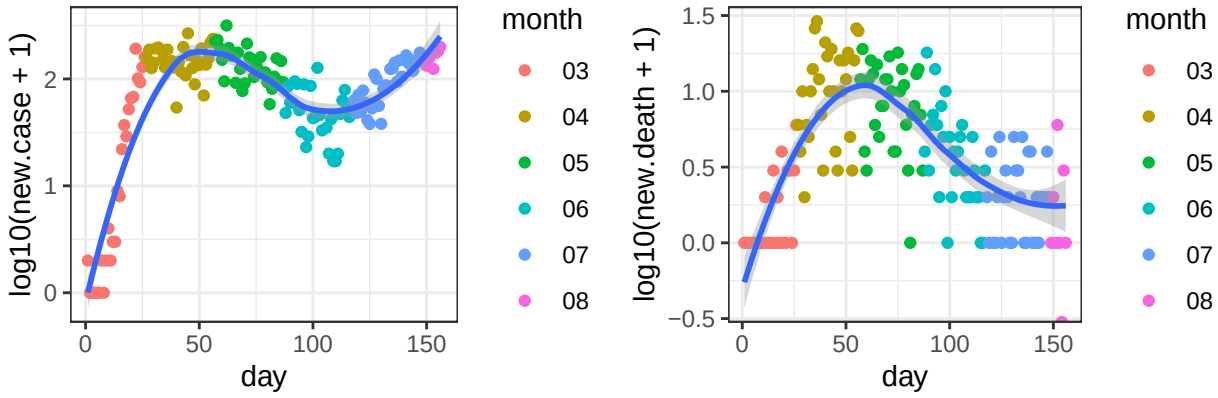
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−24

Broward_Florida



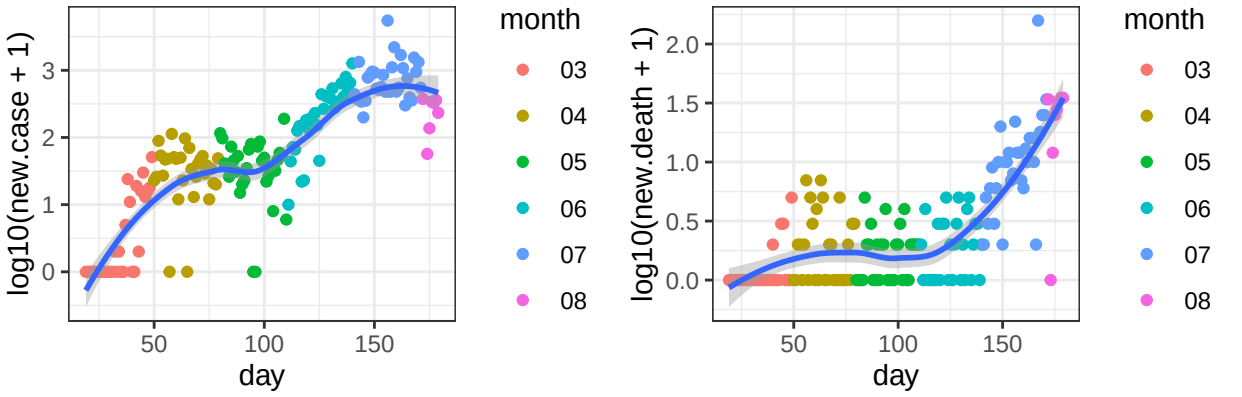
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Marion_Indiana



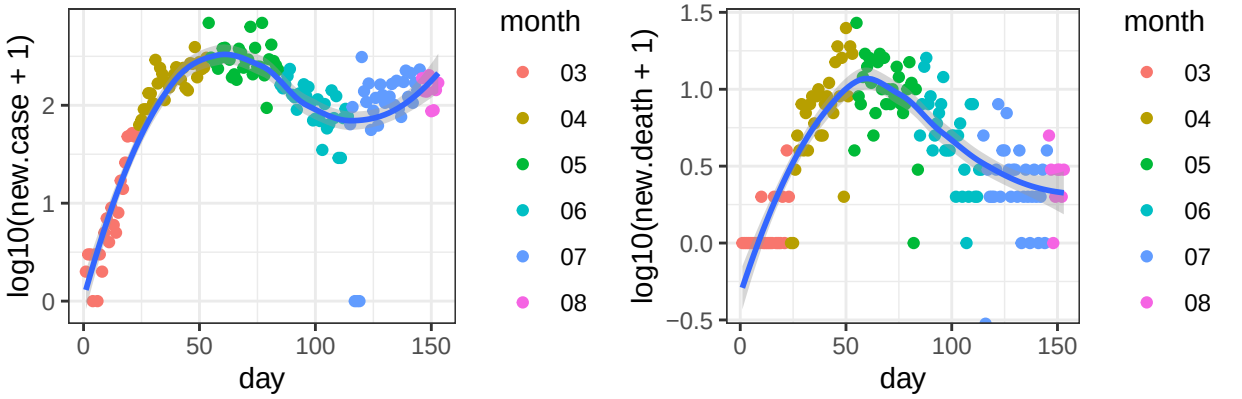
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Bexar_Texas



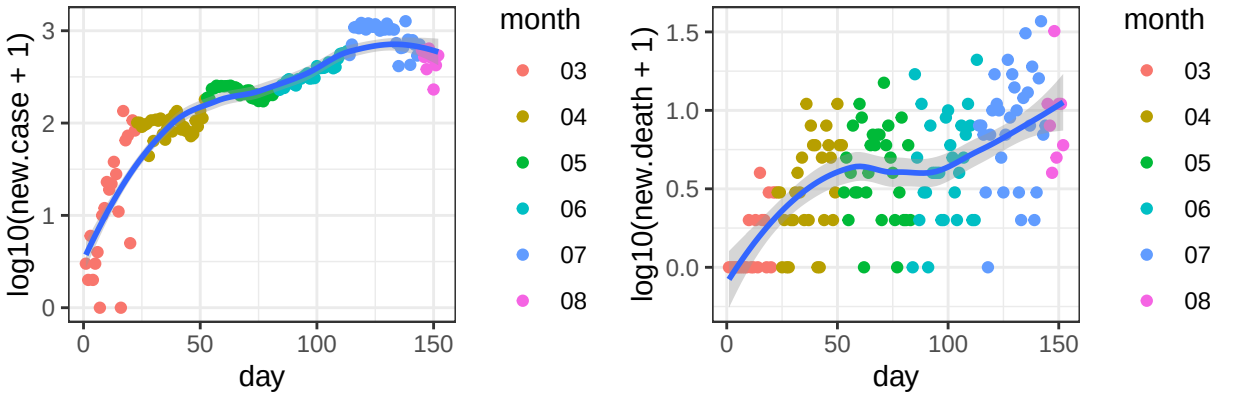
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Prince George's_Maryland



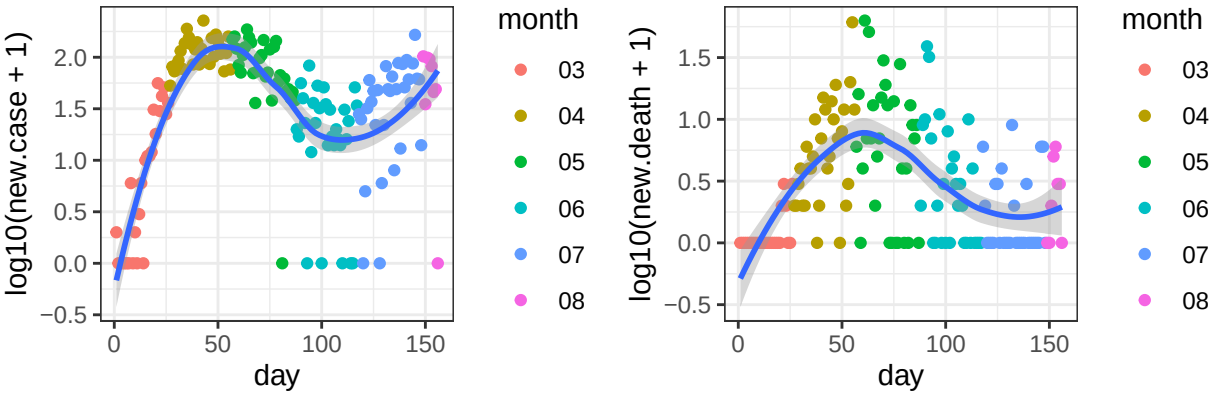
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Dallas_Texas



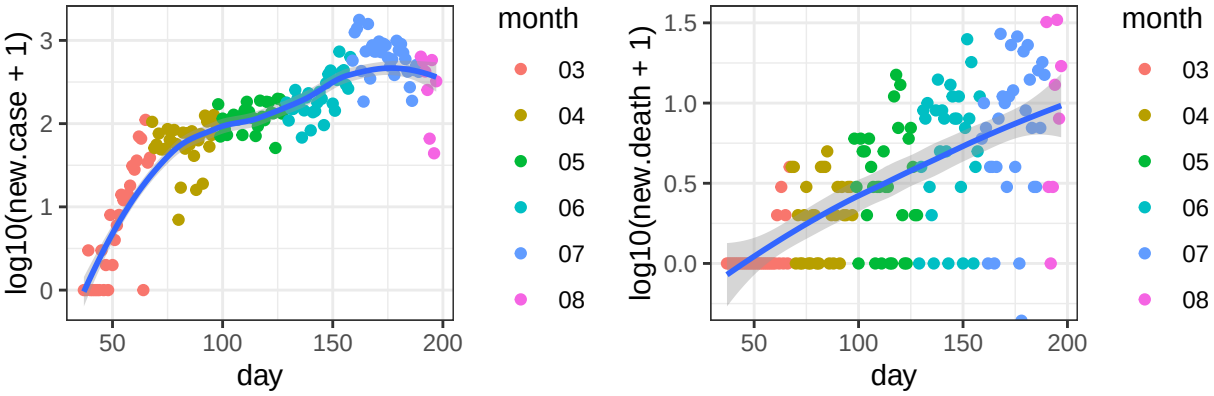
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Delaware_Pennsylvania



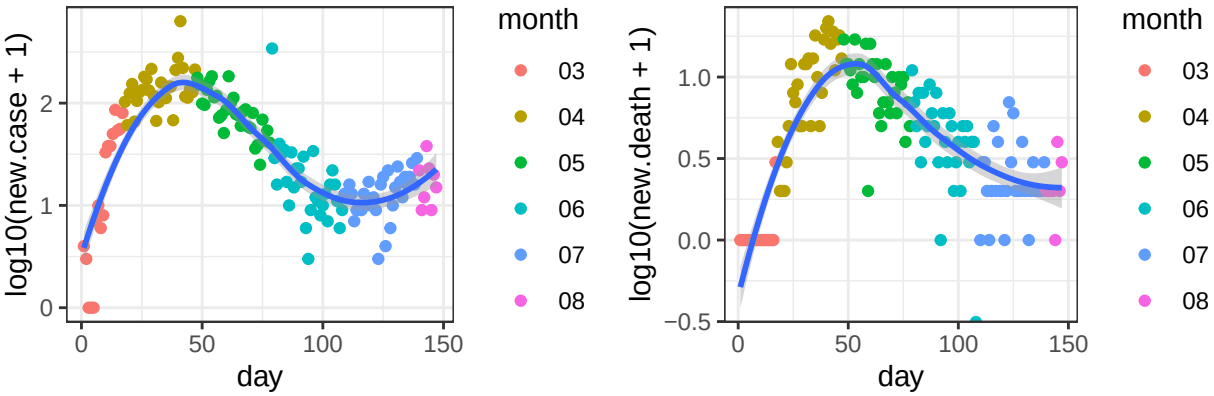
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Orange_California



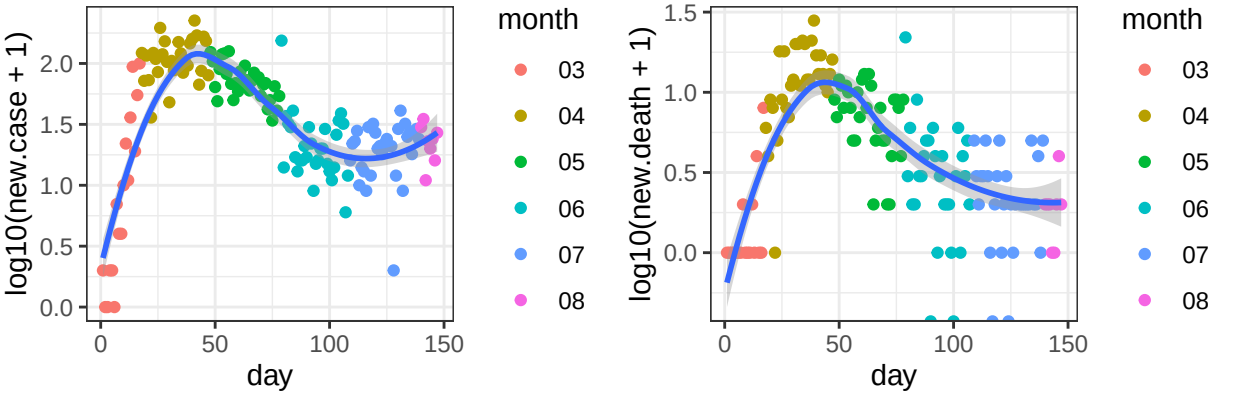
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Plymouth_Massachusetts



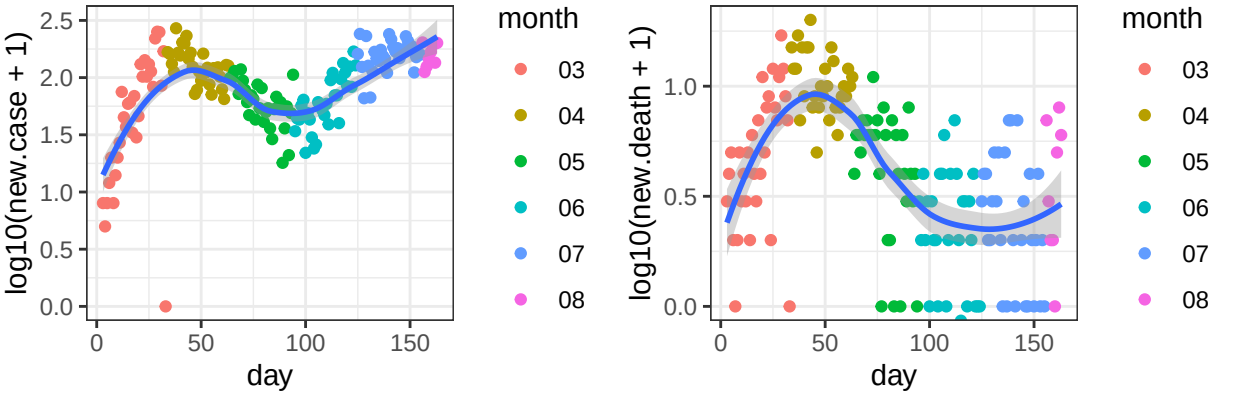
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Hampden_Massachusetts



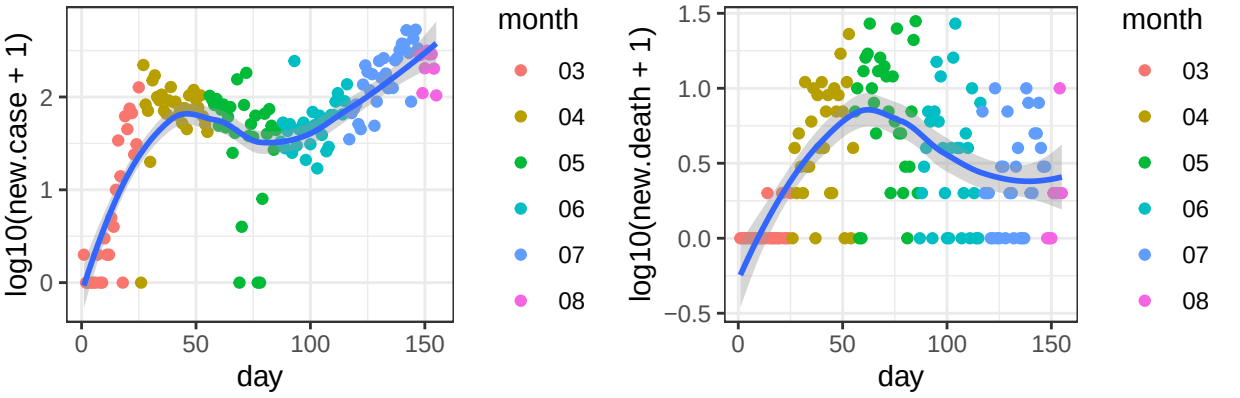
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

King_Washington

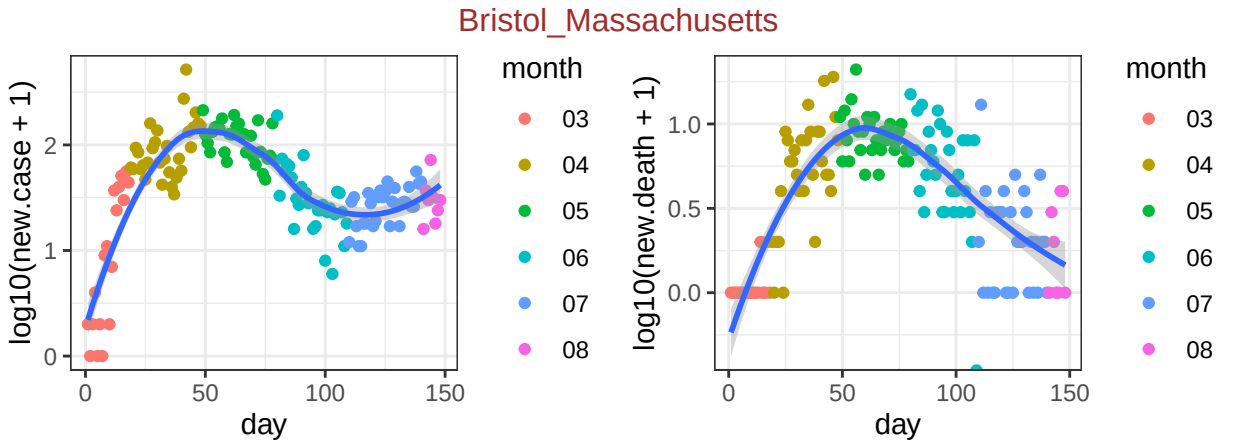


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

St. Louis_Missouri



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

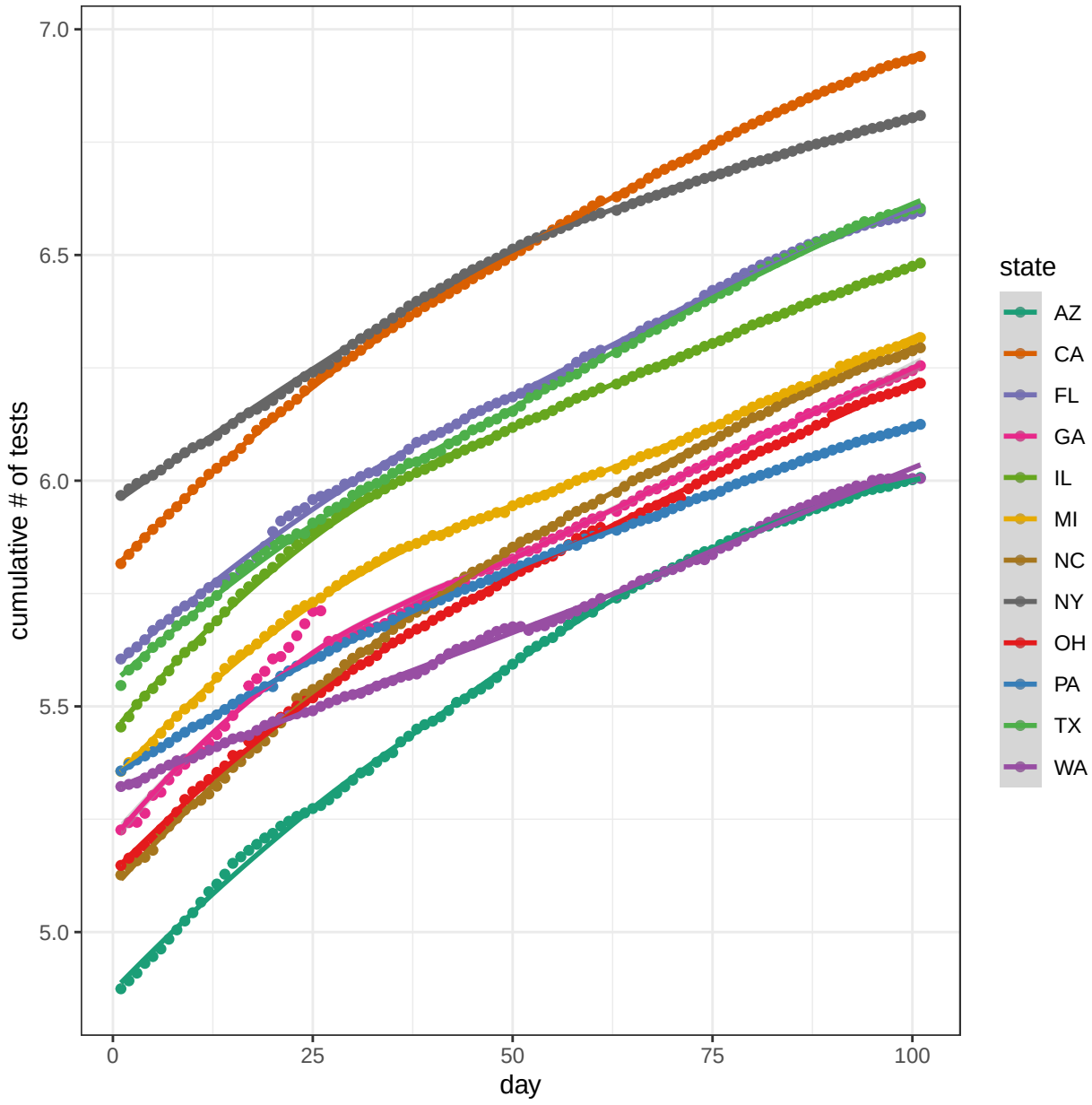


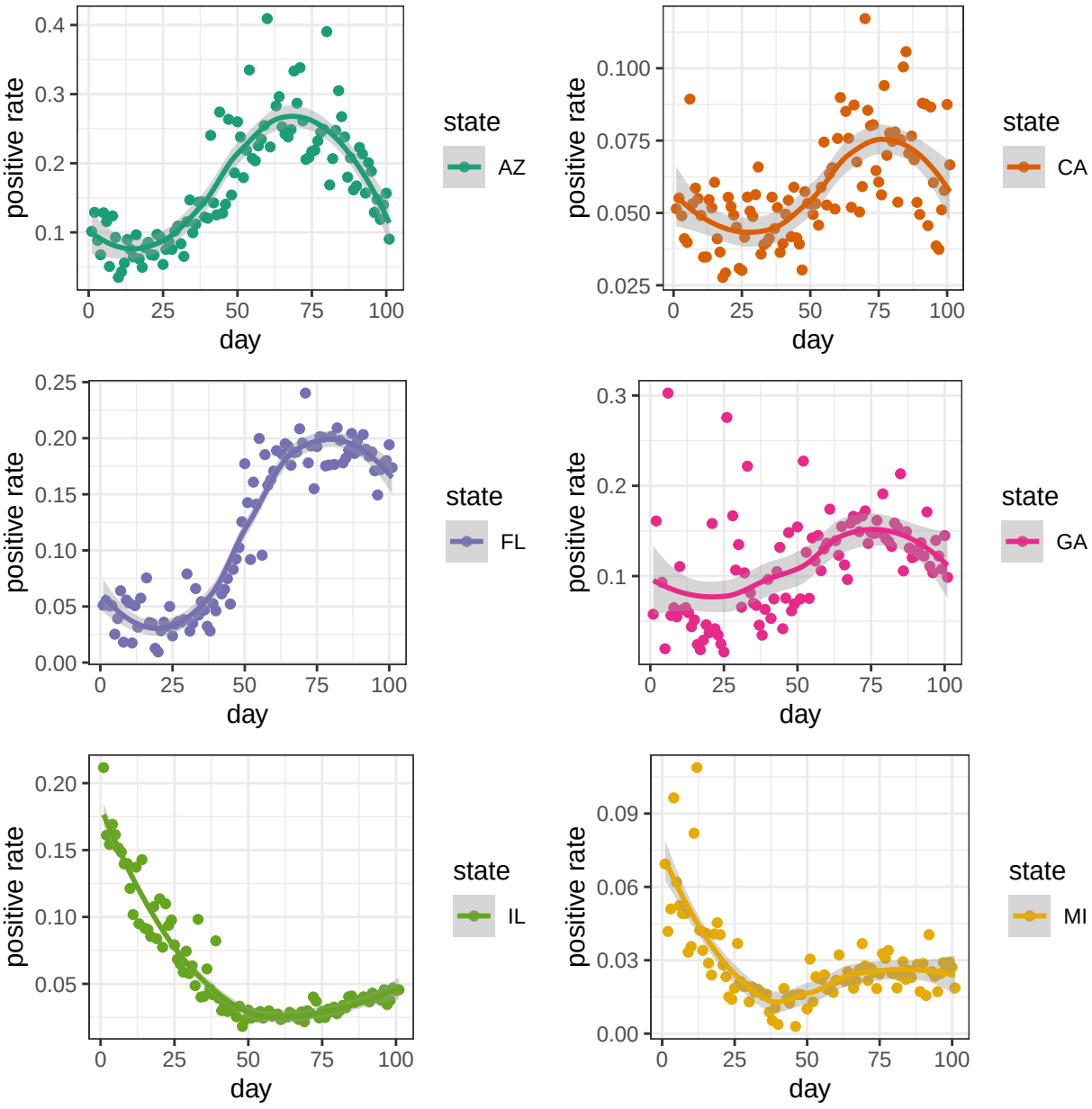
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

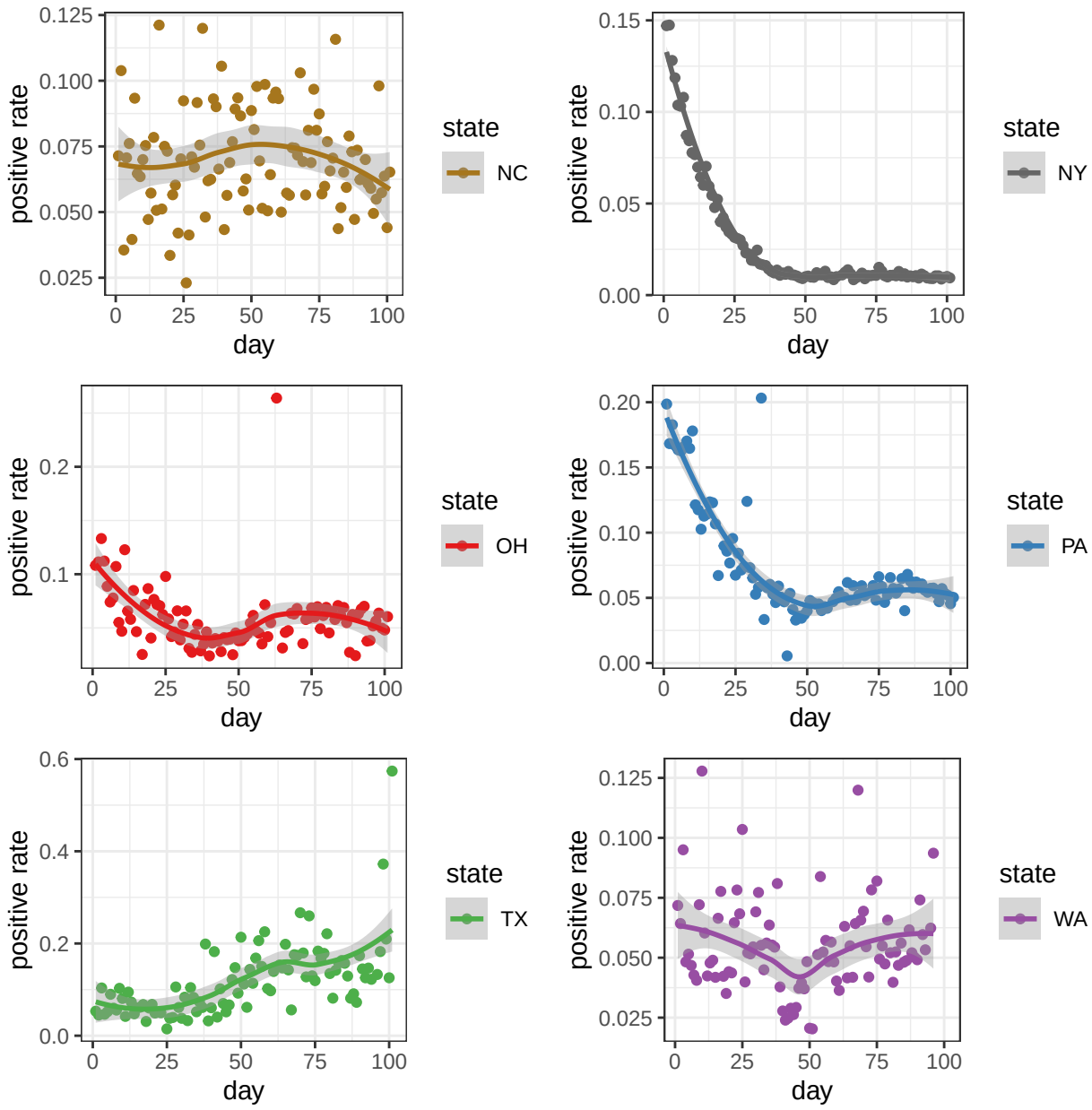
COVID Trackng

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.


7.0
6.5
6.0
5.5
5.0
cumulative # of tests
0
25
50
75
100
day
state
AZ
CA
FL
GA
IL
MI
NC
NY
OH
PA
TX
WA


positive rate
day
state
AZ
CA
FL
GA
IL
MI



Session information

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
```

```
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] RColorBrewer_1.1-2 httr_1.4.1         ggpubr_0.2.5       magrittr_1.5
## [5] ggplot2_3.3.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3       pillar_1.4.3     compiler_3.6.2   tools_3.6.2
##  [5] digest_0.6.23    lattice_0.20-38  nlme_3.1-144     evaluate_0.14
##  [9] lifecycle_0.2.0  tibble_3.0.1     gtable_0.3.0     mgcv_1.8-31
## [13] pkgconfig_2.0.3  rlang_0.4.6      Matrix_1.2-18    yaml_2.2.1
## [17] xfun_0.12        gridExtra_2.3    withr_2.1.2      stringr_1.4.0
## [21] dplyr_0.8.4      knitr_1.28       vctrs_0.3.0      cowplot_1.0.0
## [25] grid_3.6.2       tidyselect_1.0.0 glue_1.3.1       R6_2.4.1
## [29] rmarkdown_2.1    farver_2.0.3     purrr_0.3.3      splines_3.6.2
## [33] scales_1.1.0     ellipsis_0.3.0   htmltools_0.4.0  assertthat_0.2.1
## [37] colorspace_1.4-1 ggsignif_0.6.0   labeling_0.3     stringi_1.4.5
## [41] munsell_0.5.0    crayon_1.3.4
```