Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-07-10

Contents

Introduction

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

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

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

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

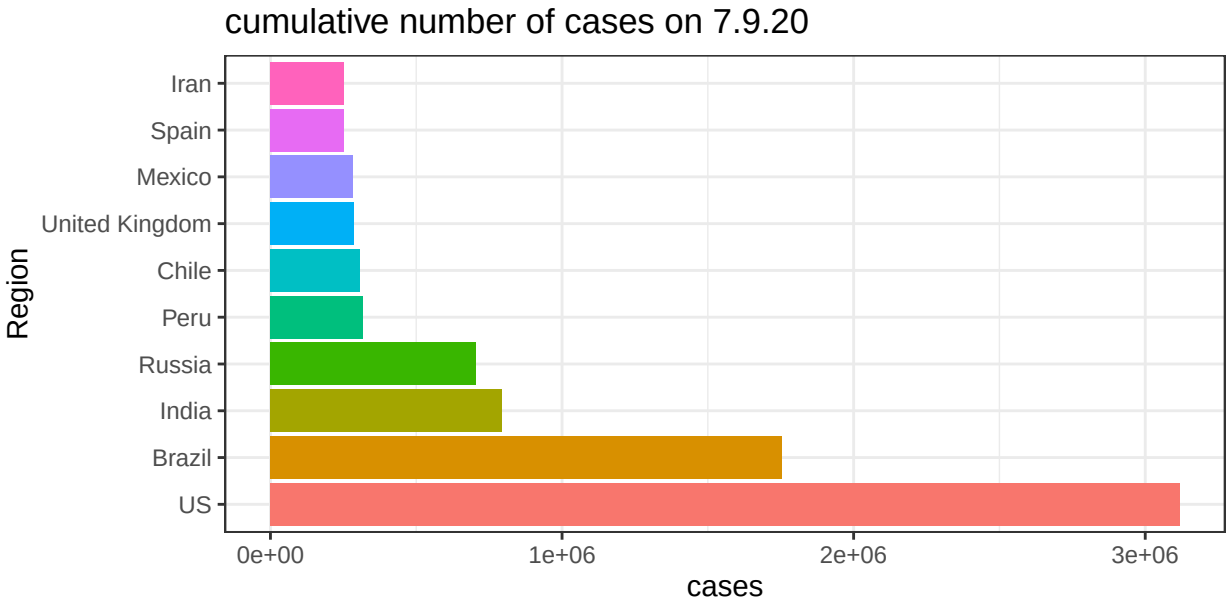
JHU

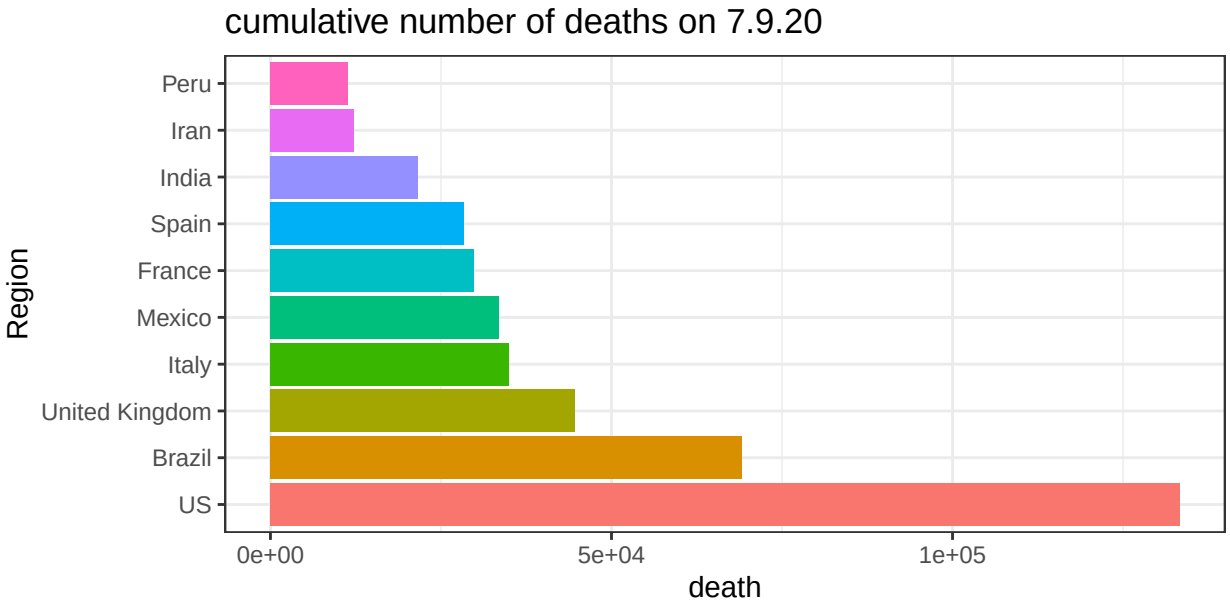
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

time series data

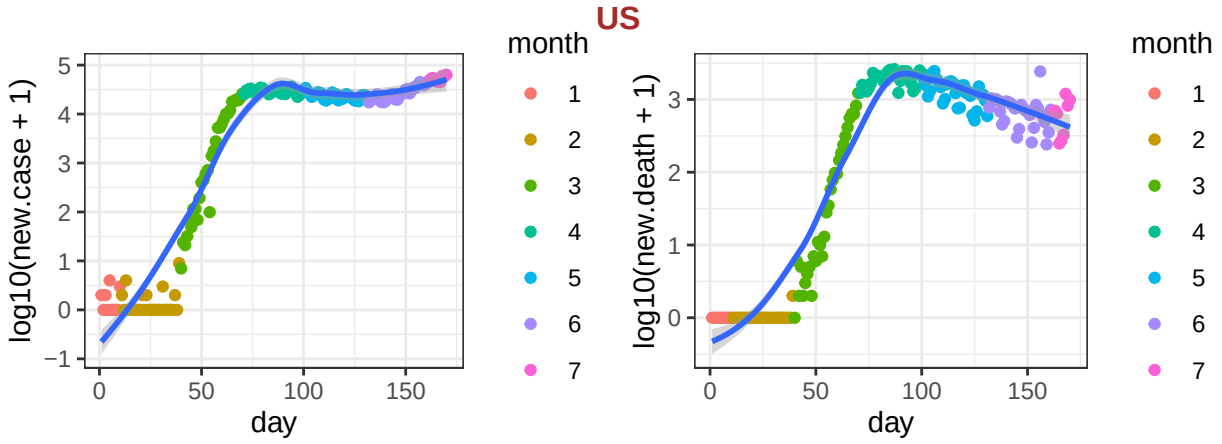
The time series provide counts (e.g., confirmed cases, deaths) starting from Jan 22nd, 2020 for 253 locations. Currently there is no data of individual US state in these time series data files.

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

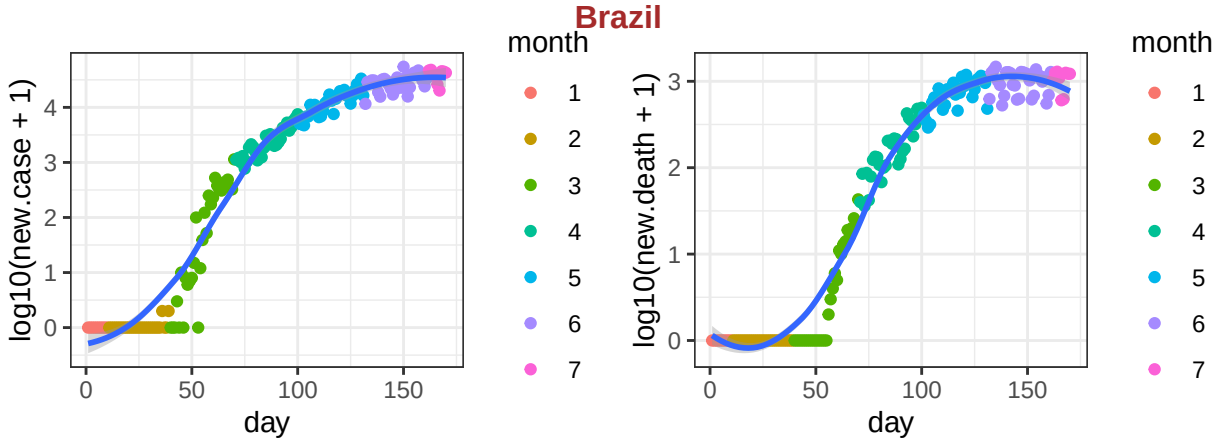




Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

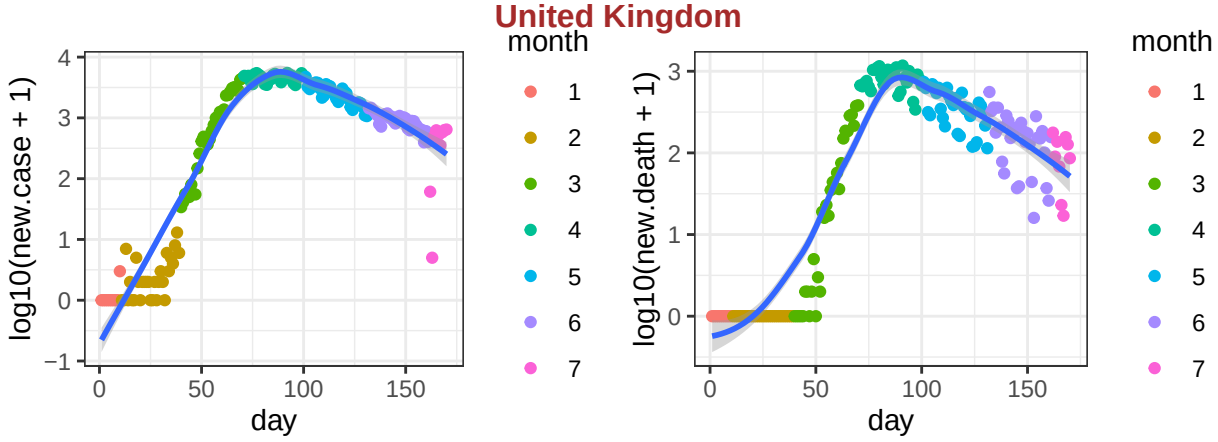


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



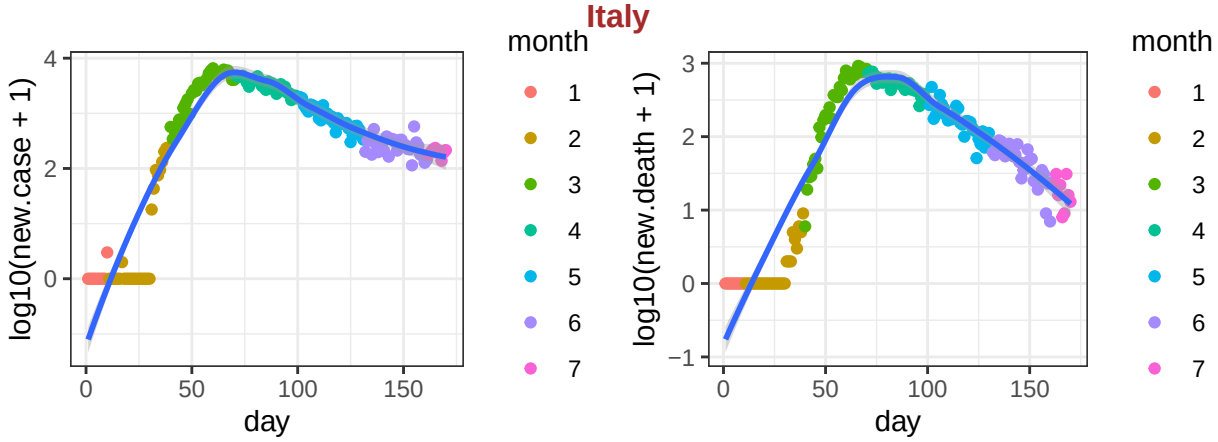
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



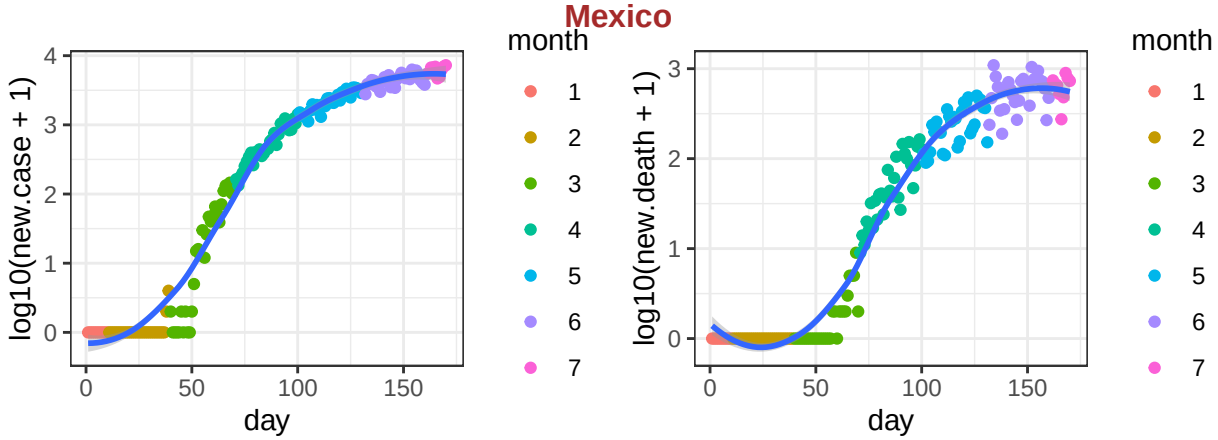
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



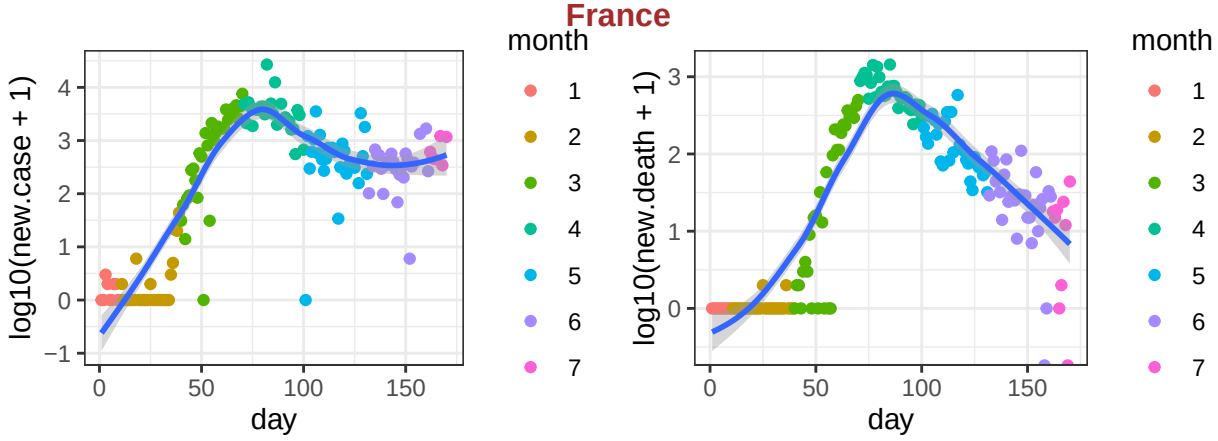
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Mexico



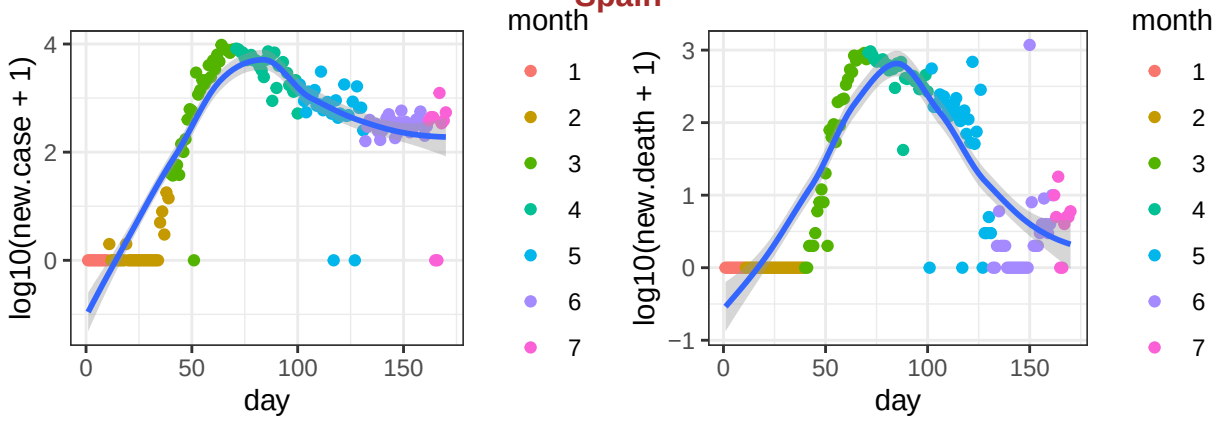
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



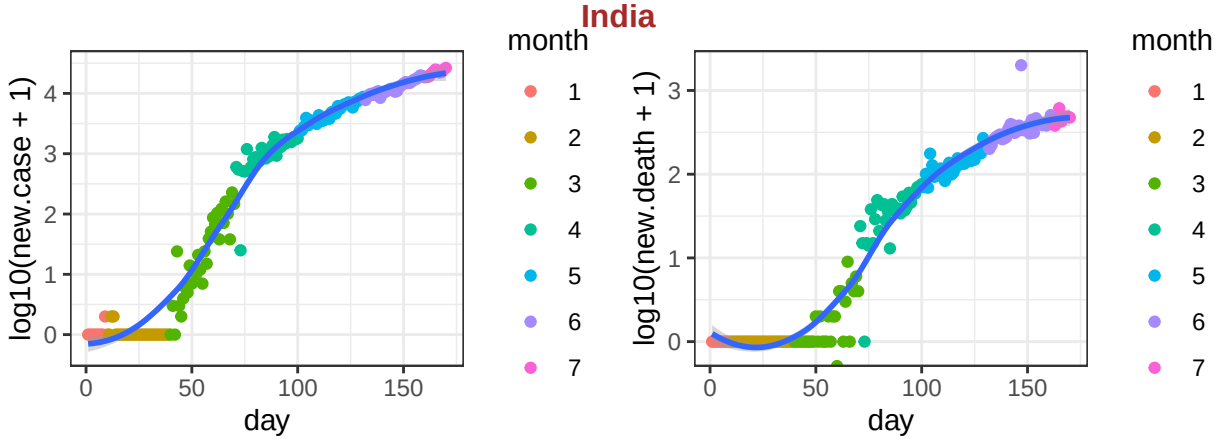
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain

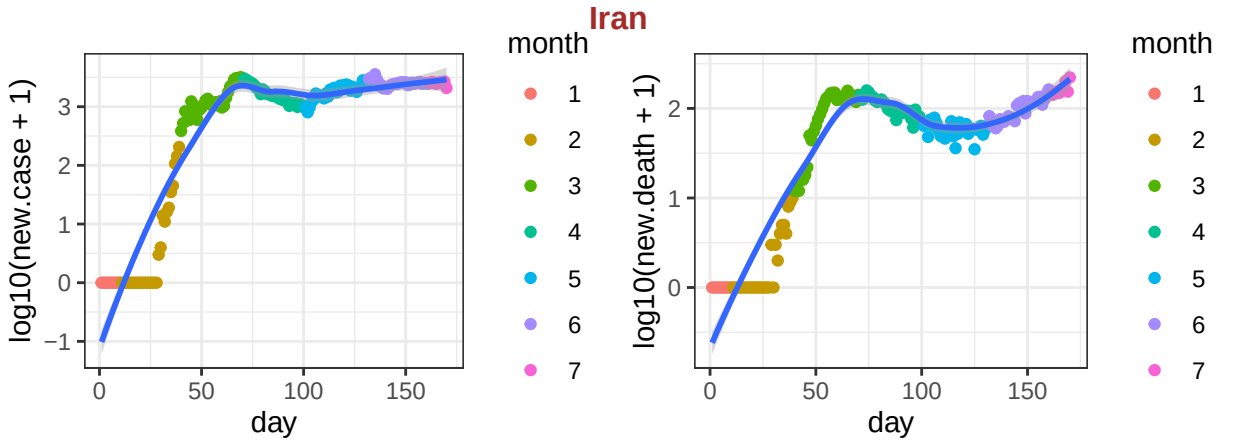


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

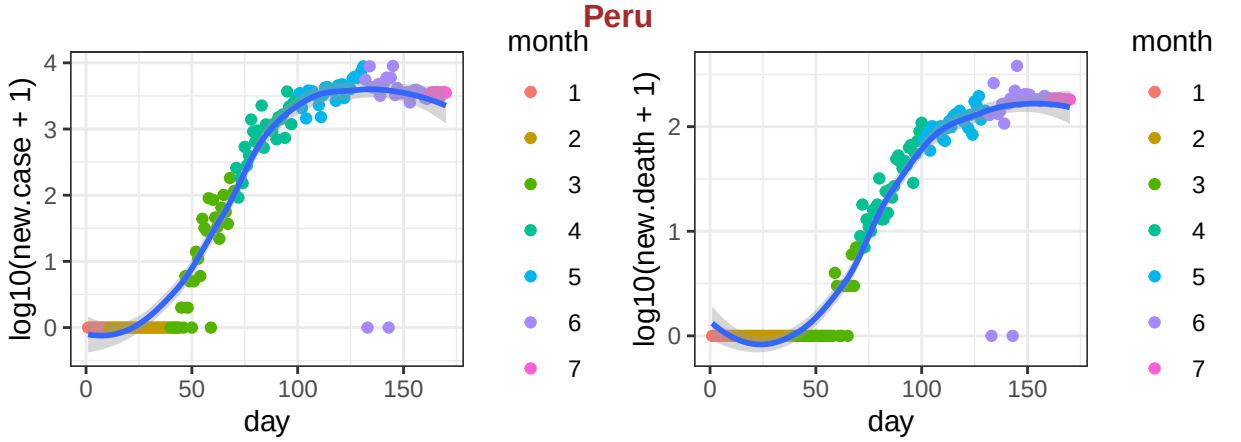
India



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



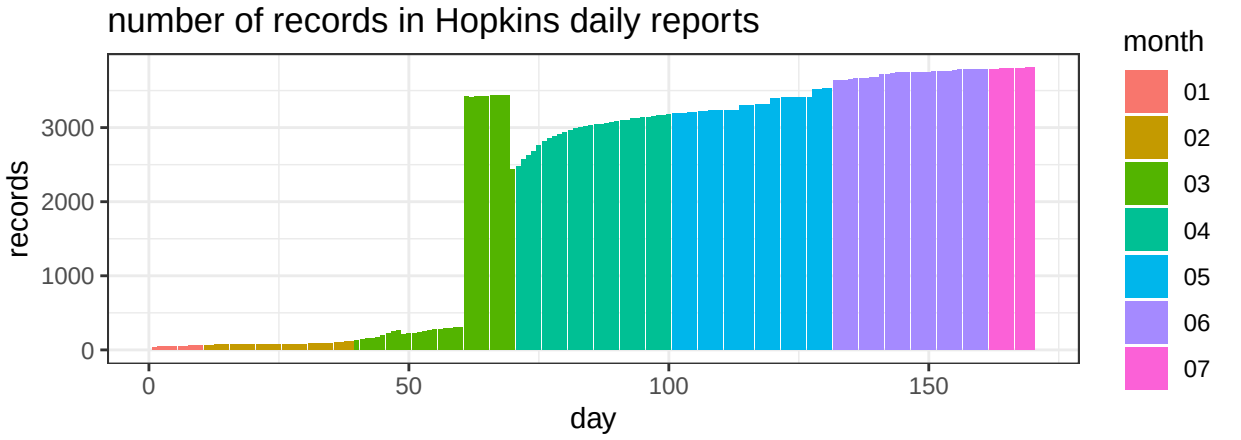
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



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] "2020-07-09"
```

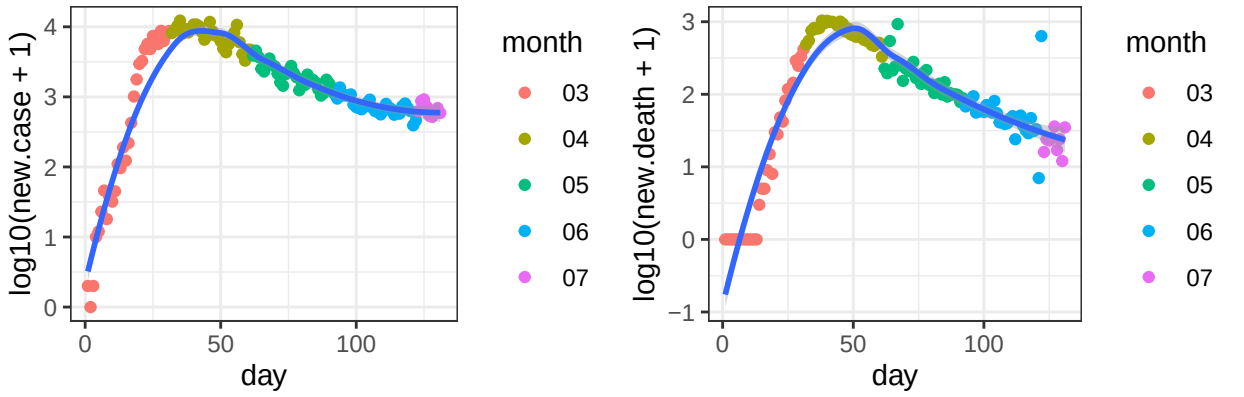
state level data

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

```
##            date          state fips  cases deaths
## 7088 2020-07-09       New York   36 404207  31979
## 7086 2020-07-09     New Jersey   34 176217  15448
## 7077 2020-07-09  Massachusetts   25 110897   8268
## 7069 2020-07-09       Illinois   17 152188   7333
## 7095 2020-07-09   Pennsylvania   42  97634   6895
## 7059 2020-07-09     California    6 303474   6825
## 7078 2020-07-09       Michigan   26  75247   6275
## 7061 2020-07-09    Connecticut    9  47209   4348
## 7064 2020-07-09        Florida   12 232710   4008
## 7074 2020-07-09      Louisiana   22  72102   3355
## 7076 2020-07-09       Maryland   24  72037   3288
## 7101 2020-07-09          Texas   48 240025   3037
## 7092 2020-07-09           Ohio   39  61331   3006
## 7065 2020-07-09        Georgia   13  98693   2880
## 7070 2020-07-09        Indiana   18  50829   2739
## 7057 2020-07-09        Arizona    4 112783   2047
## 7105 2020-07-09       Virginia   51  67988   1937
## 7060 2020-07-09       Colorado    8  35604   1707
## 7079 2020-07-09      Minnesota   27  40201   1528
## 7089 2020-07-09 North Carolina   37  79648   1487
## 7106 2020-07-09     Washington   53  40369   1410
## 7080 2020-07-09    Mississippi   28  33591   1204
## 7081 2020-07-09       Missouri   29  27301   1095
## 7055 2020-07-09        Alabama    1  49174   1068
## 7097 2020-07-09   Rhode Island   44  17243    974
## 7098 2020-07-09 South Carolina   45  50691    905
## 7108 2020-07-09      Wisconsin   55  37358    821
## 7071 2020-07-09           Iowa   19  33254    741
## 7100 2020-07-09      Tennessee   47  56531    700
## 7073 2020-07-09       Kentucky   21  18576    640
```

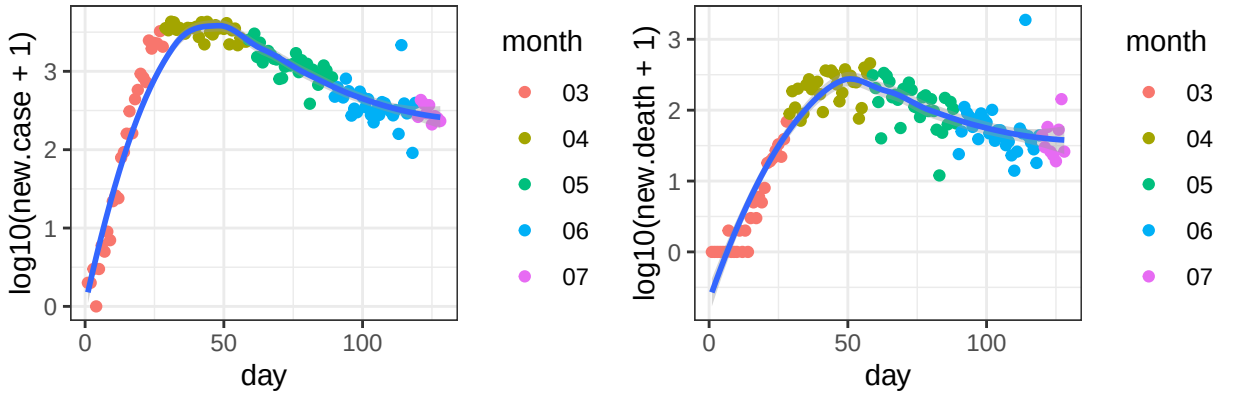
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



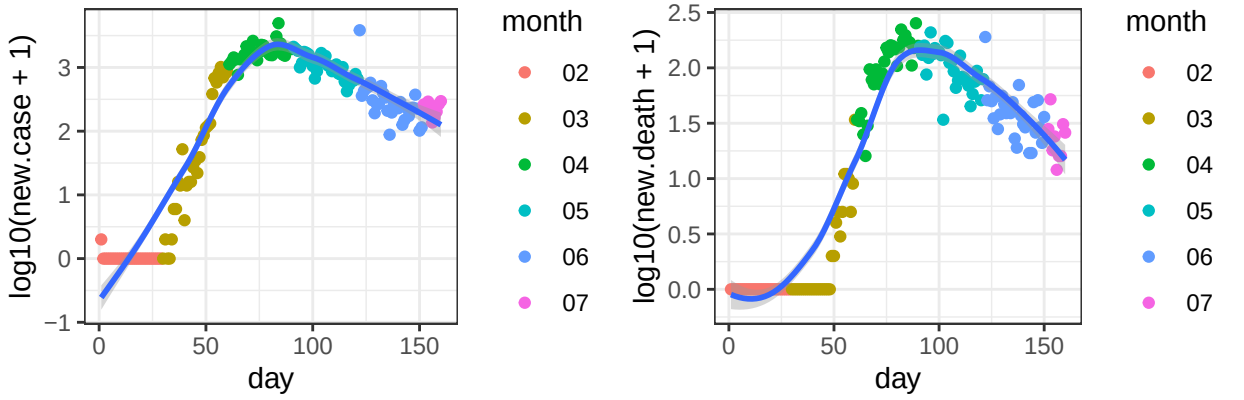
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



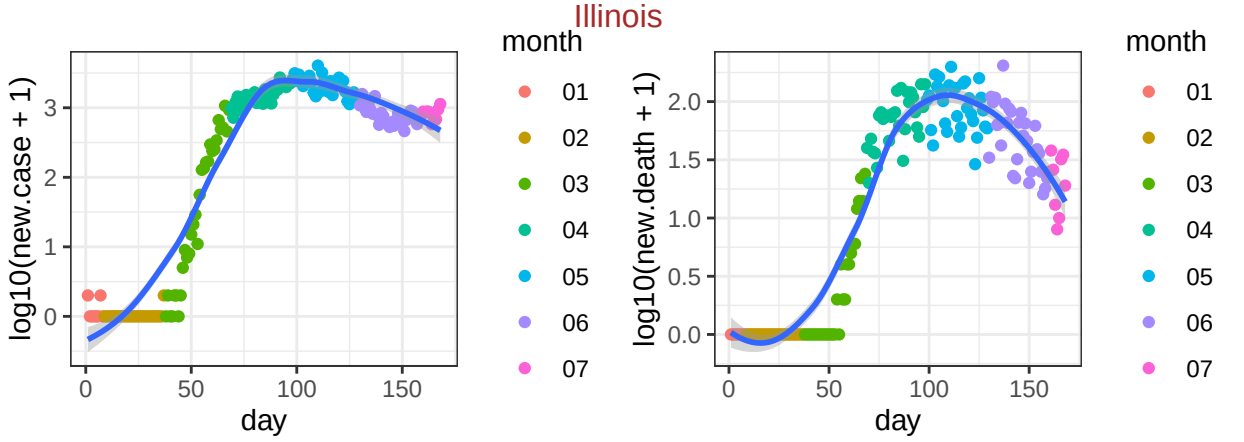
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Massachusetts



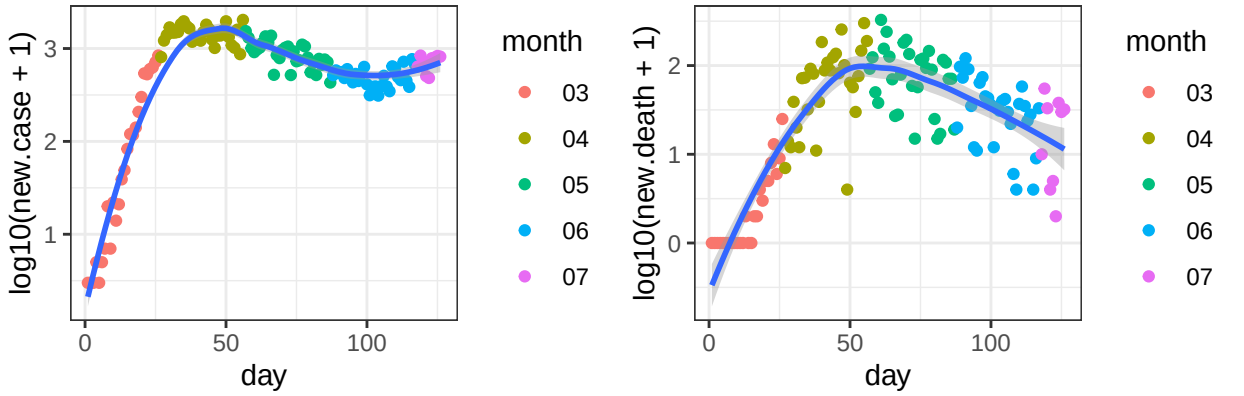
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Illinois



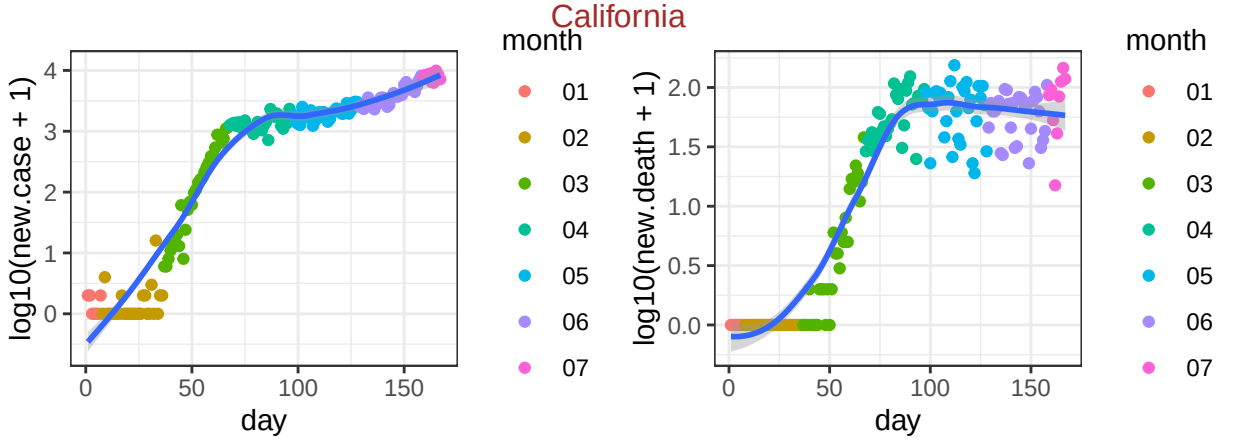
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Pennsylvania



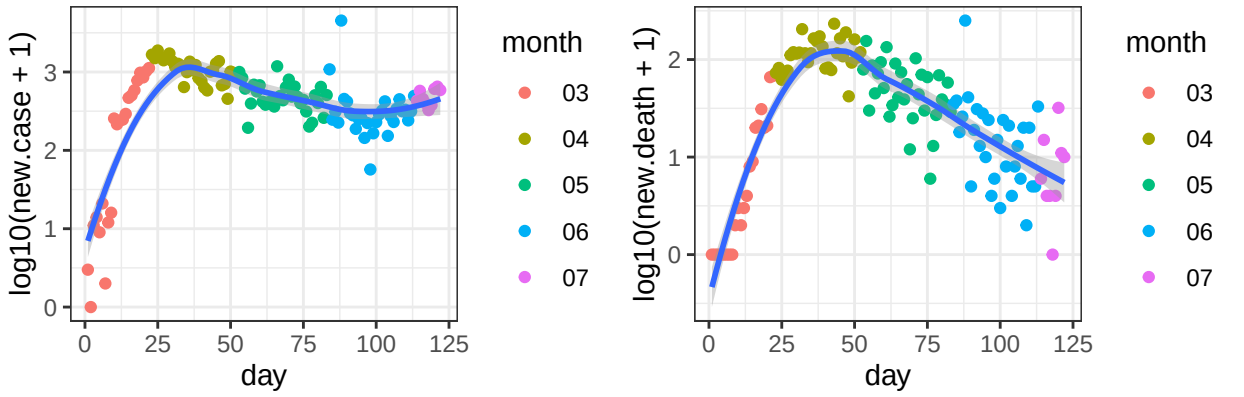
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

California



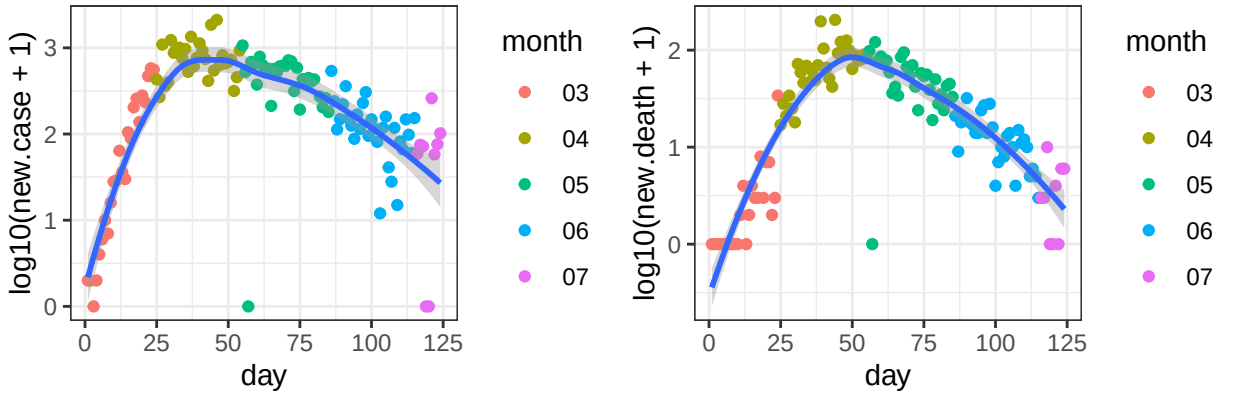
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Michigan



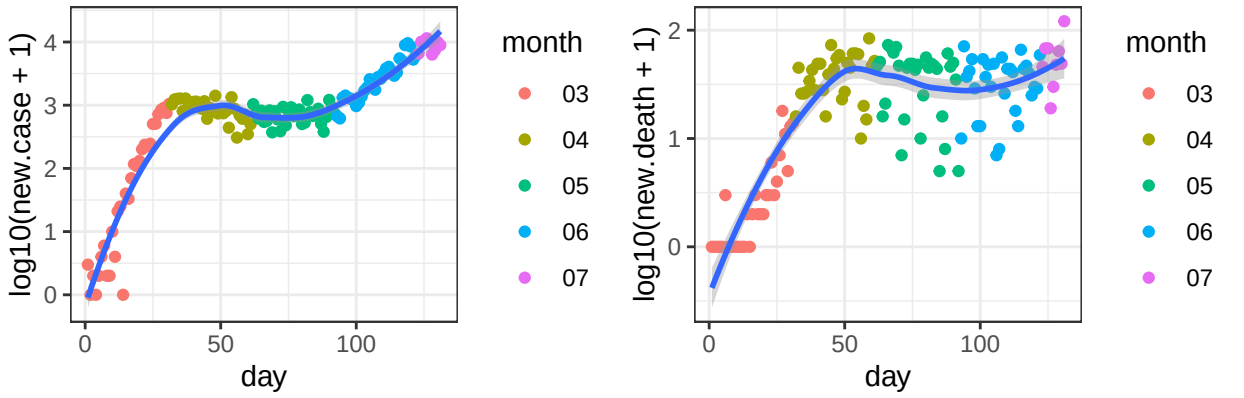
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Connecticut



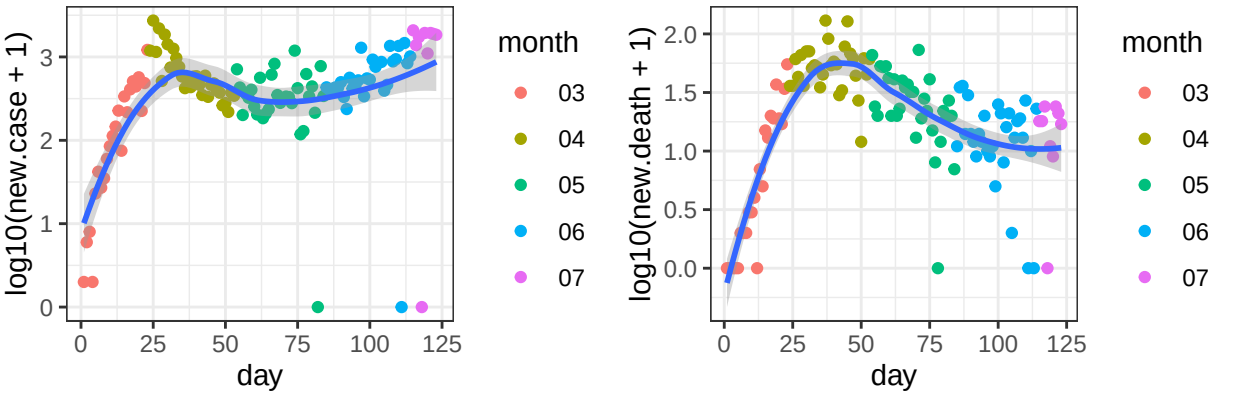
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Florida



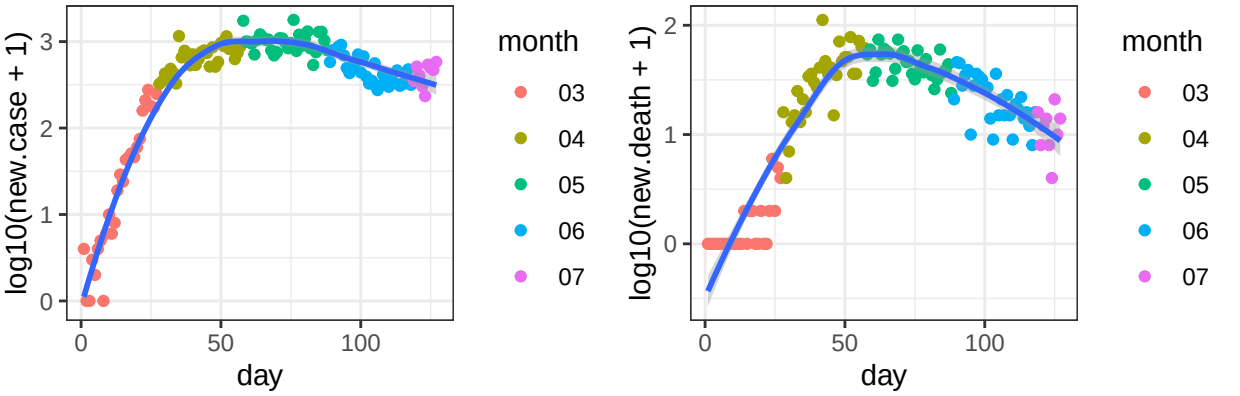
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Louisiana



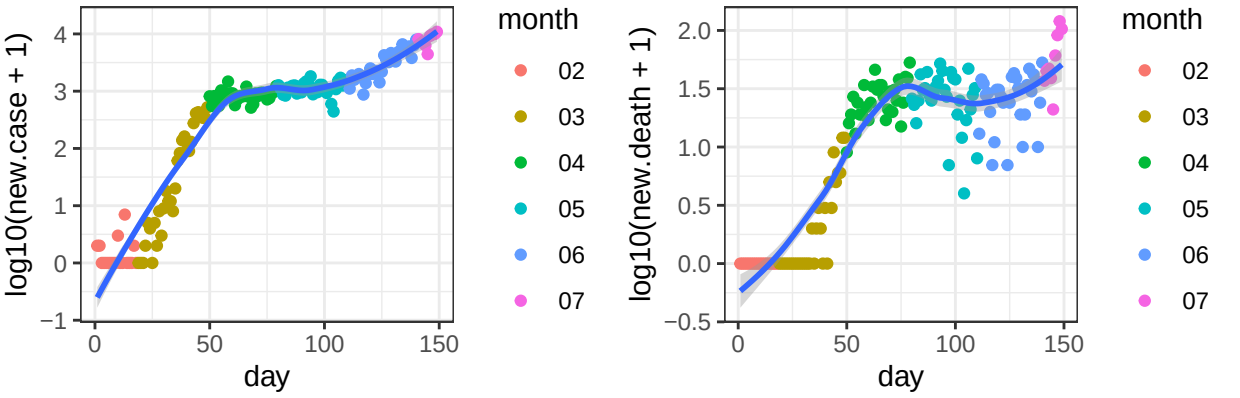
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



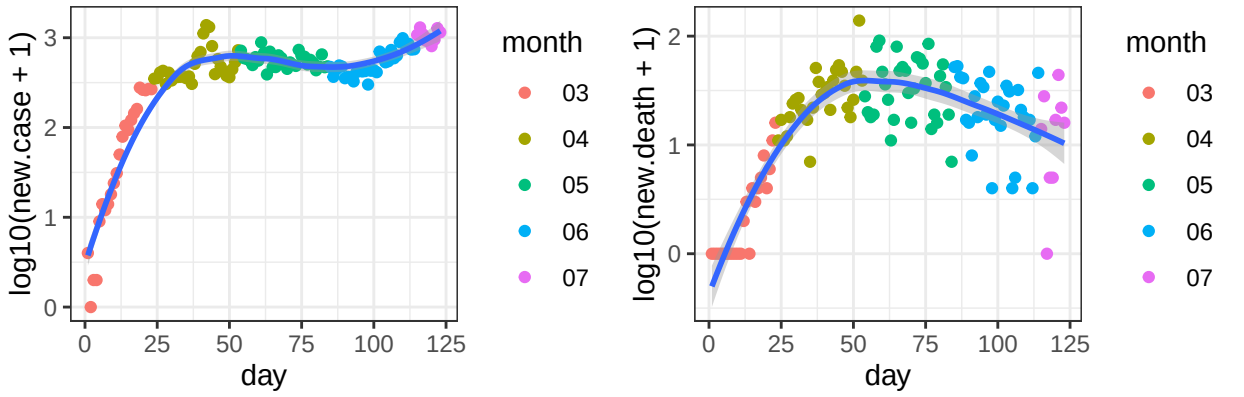
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Texas



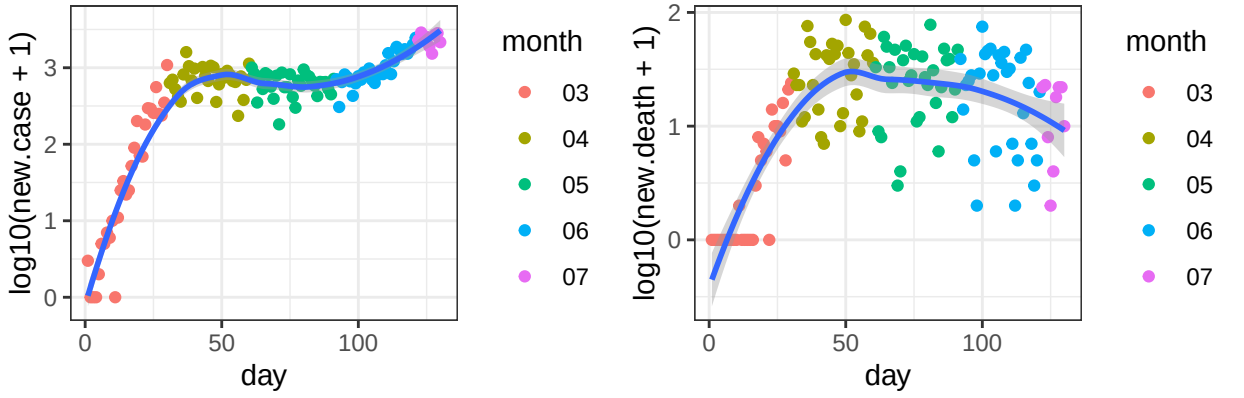
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Ohio



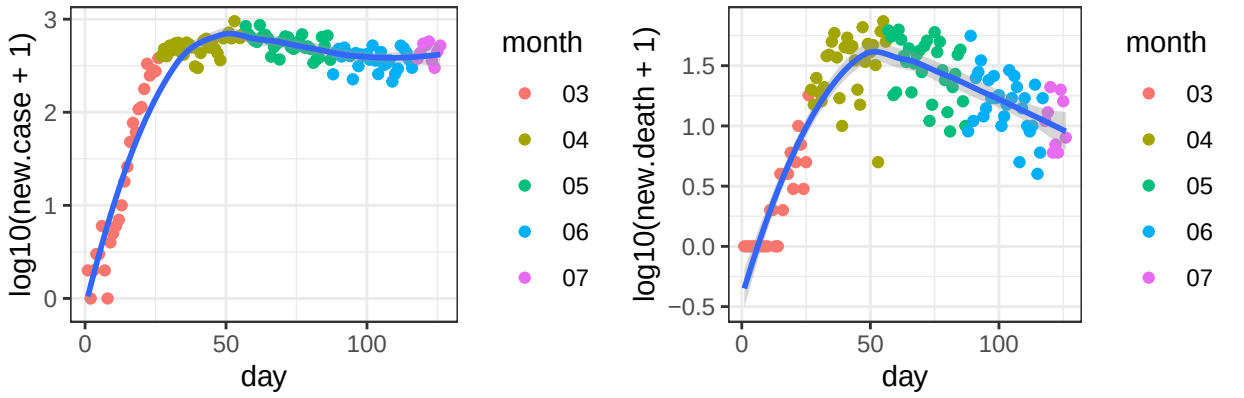
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Georgia



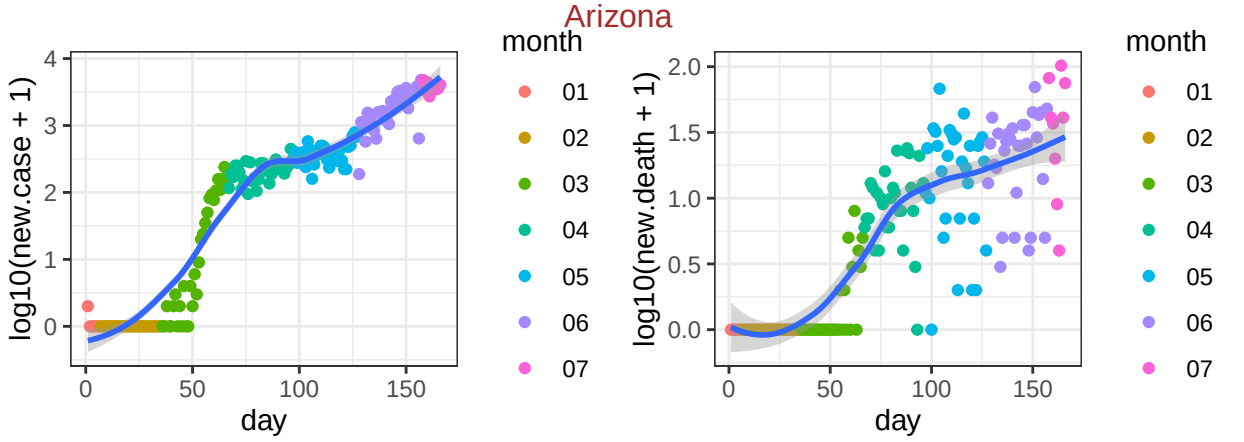
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Indiana



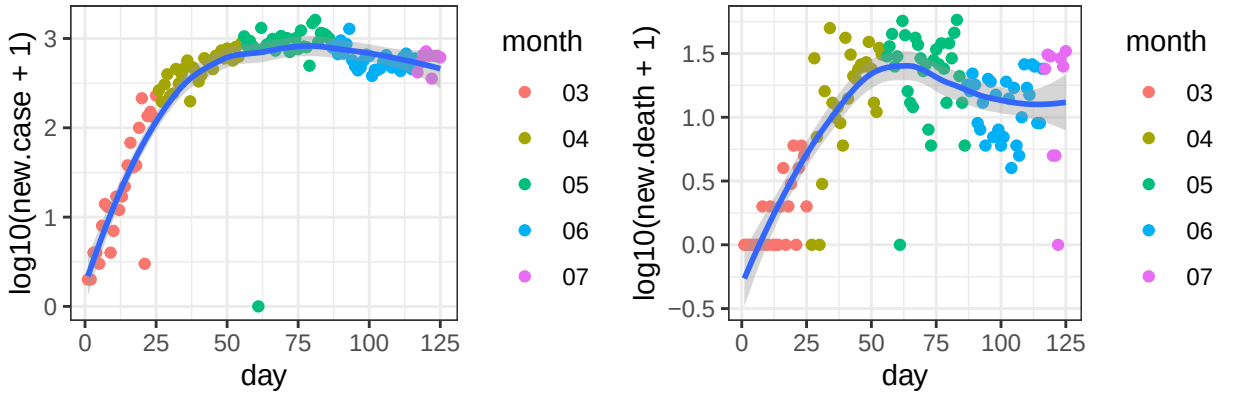
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Arizona



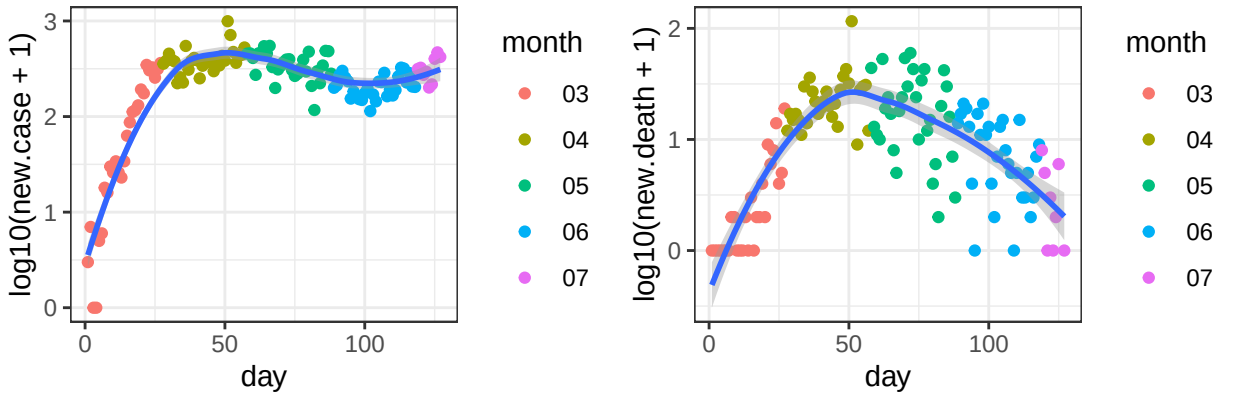
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Virginia



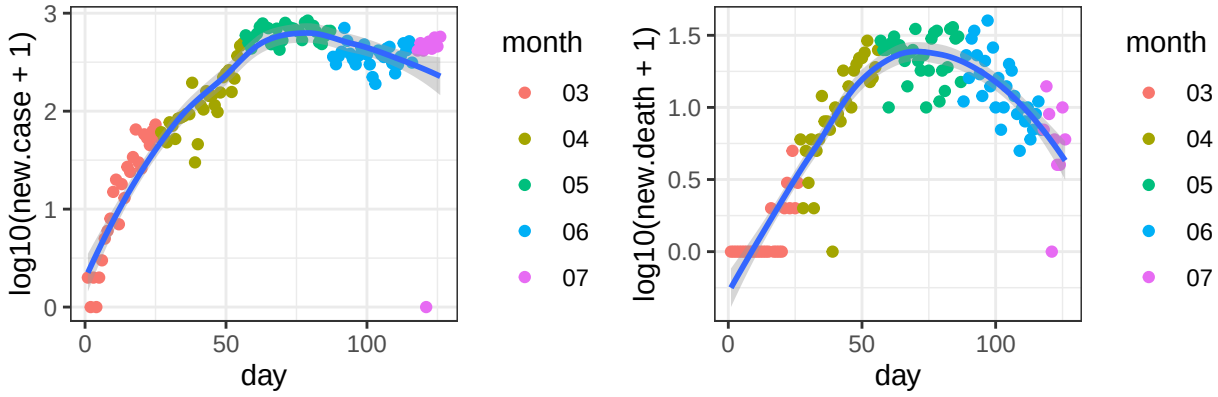
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Colorado



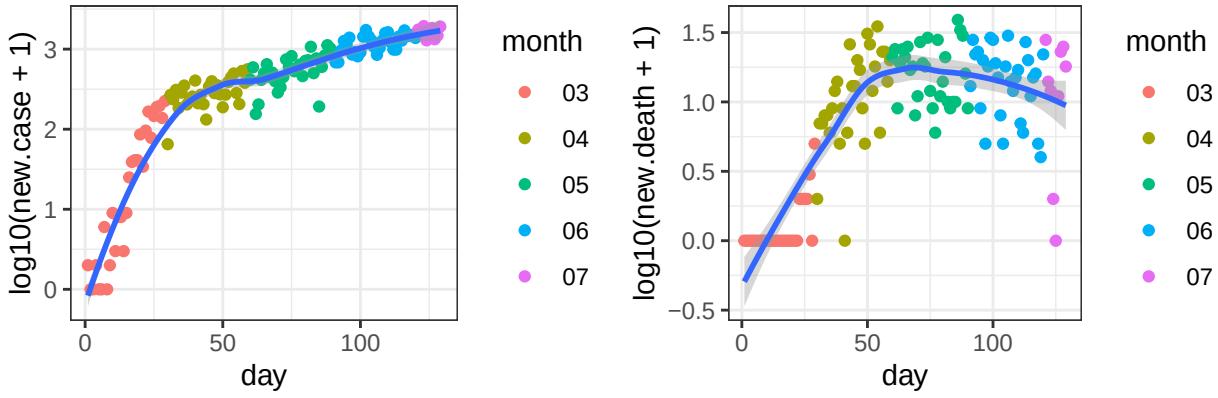
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Minnesota



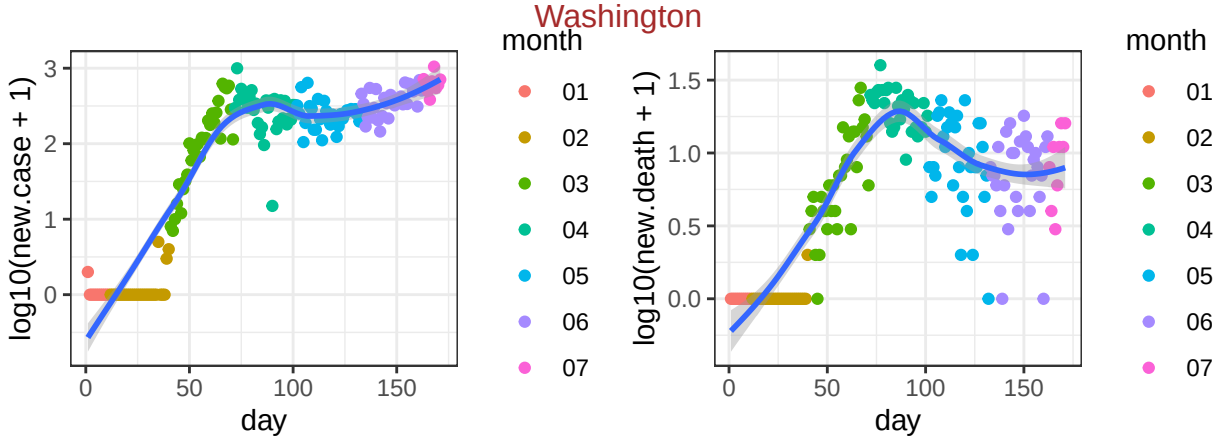
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

North Carolina



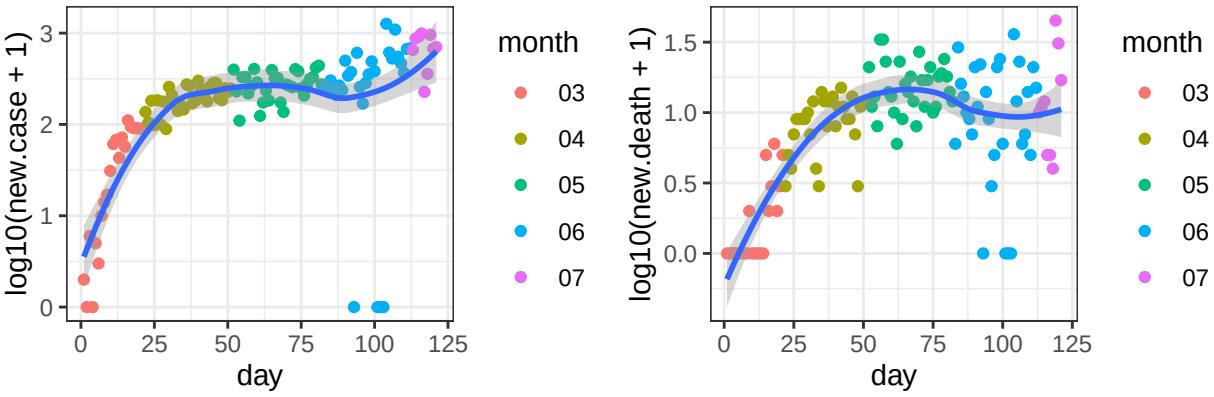
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Washington



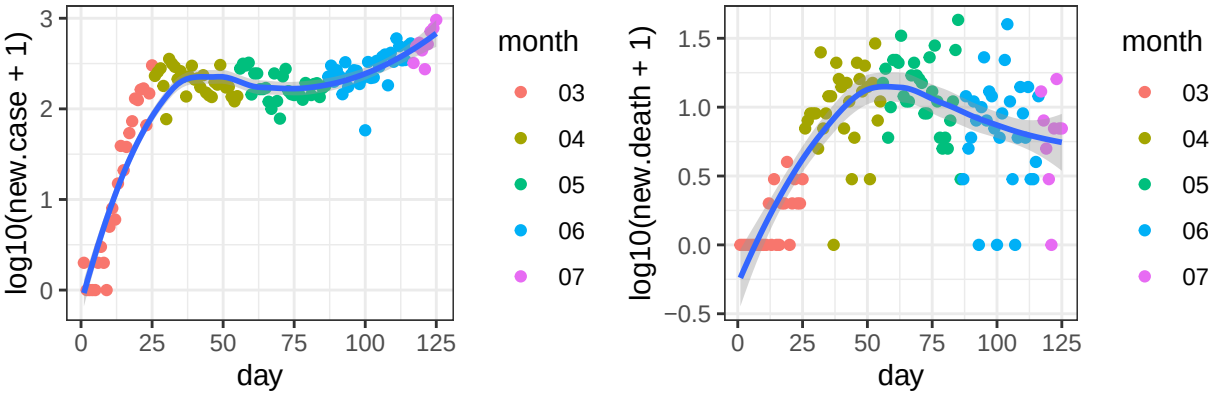
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Mississippi



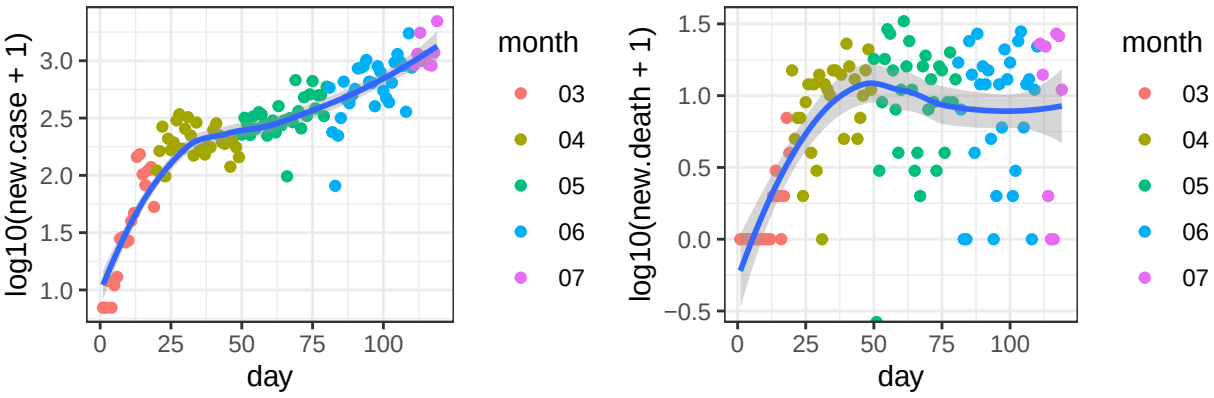
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Missouri



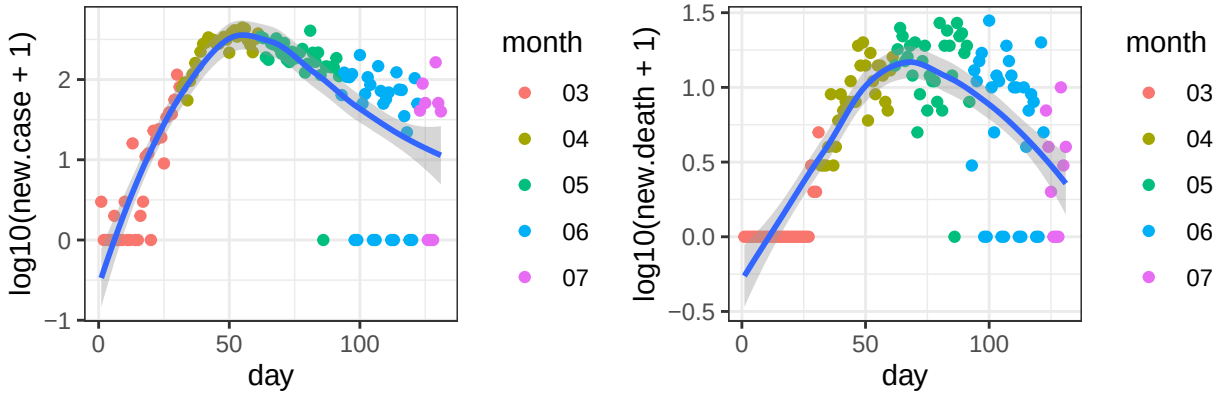
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Alabama



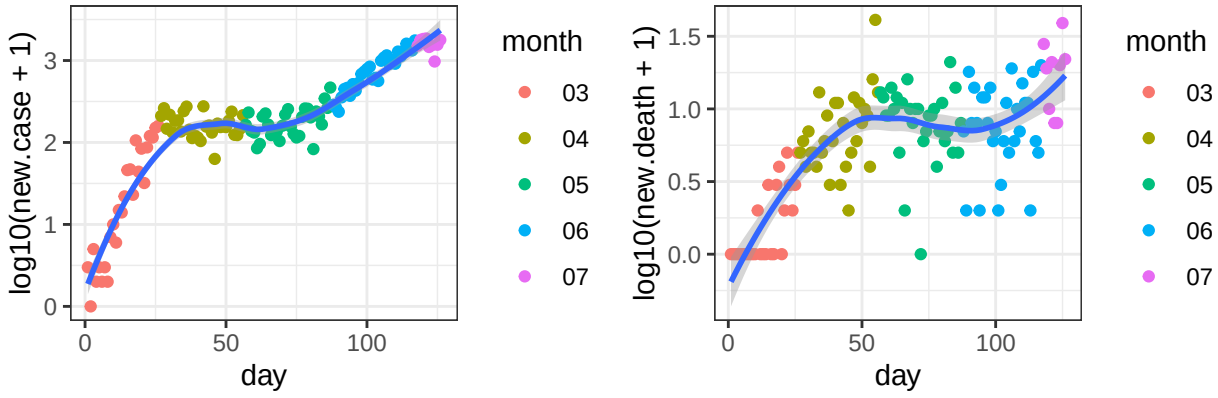
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Rhode Island



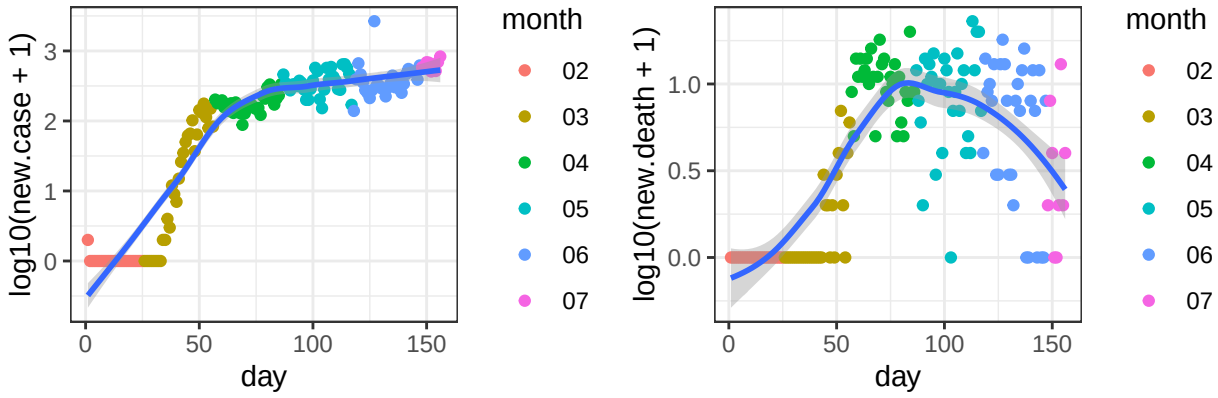
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

South Carolina



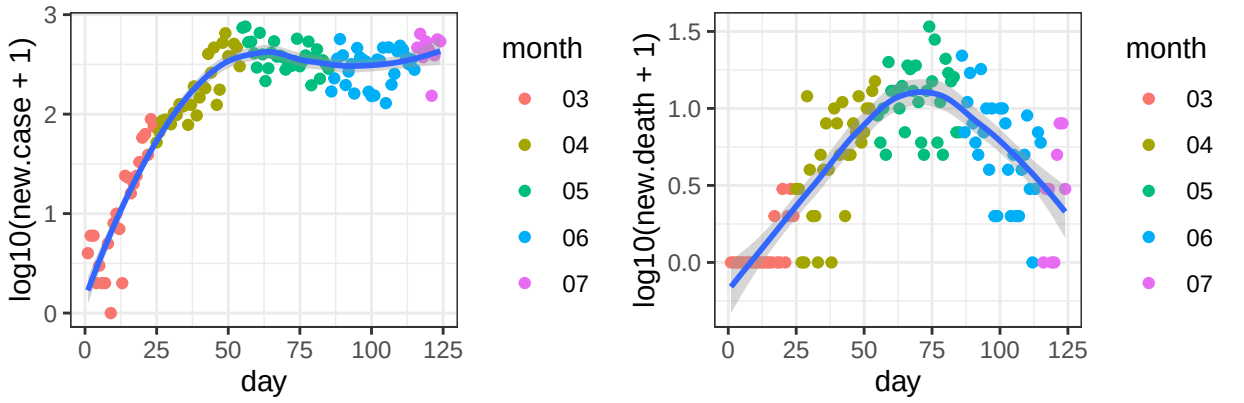
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Wisconsin



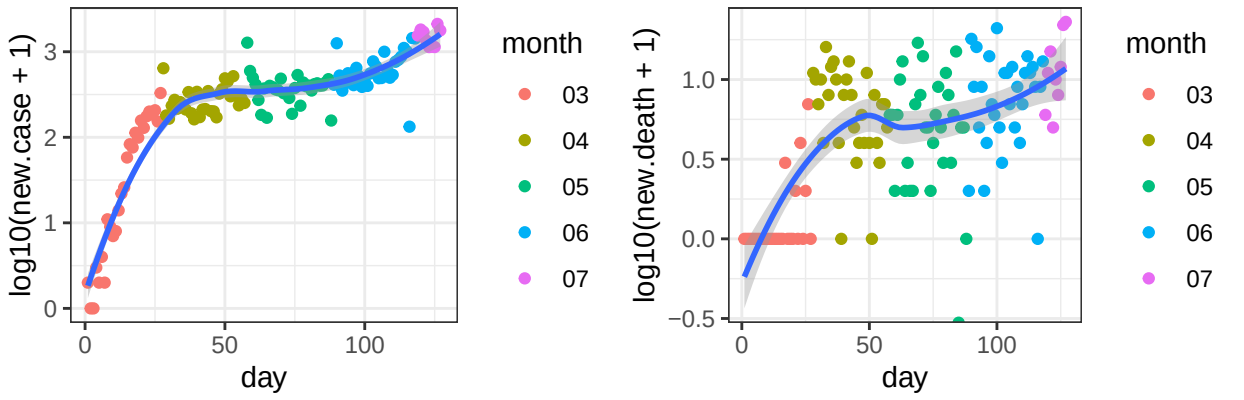
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

Iowa



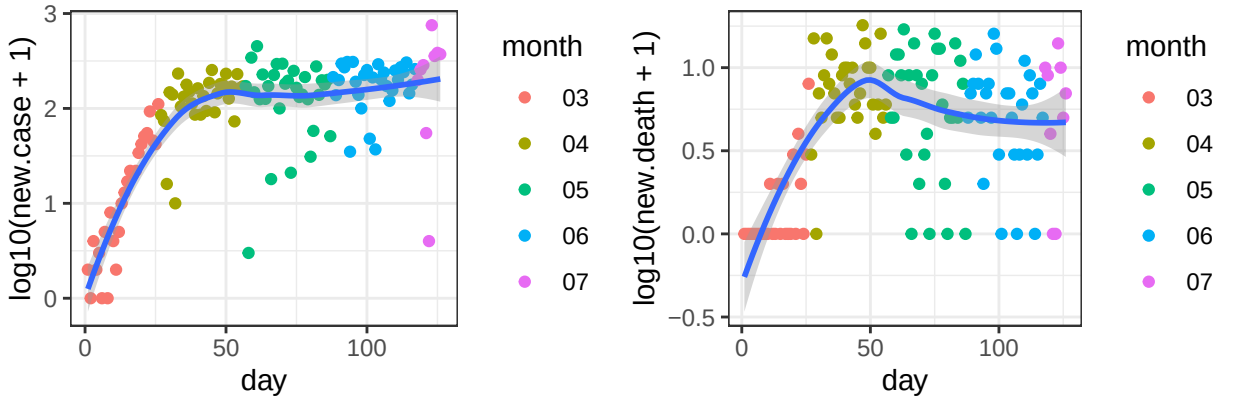
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Tennessee



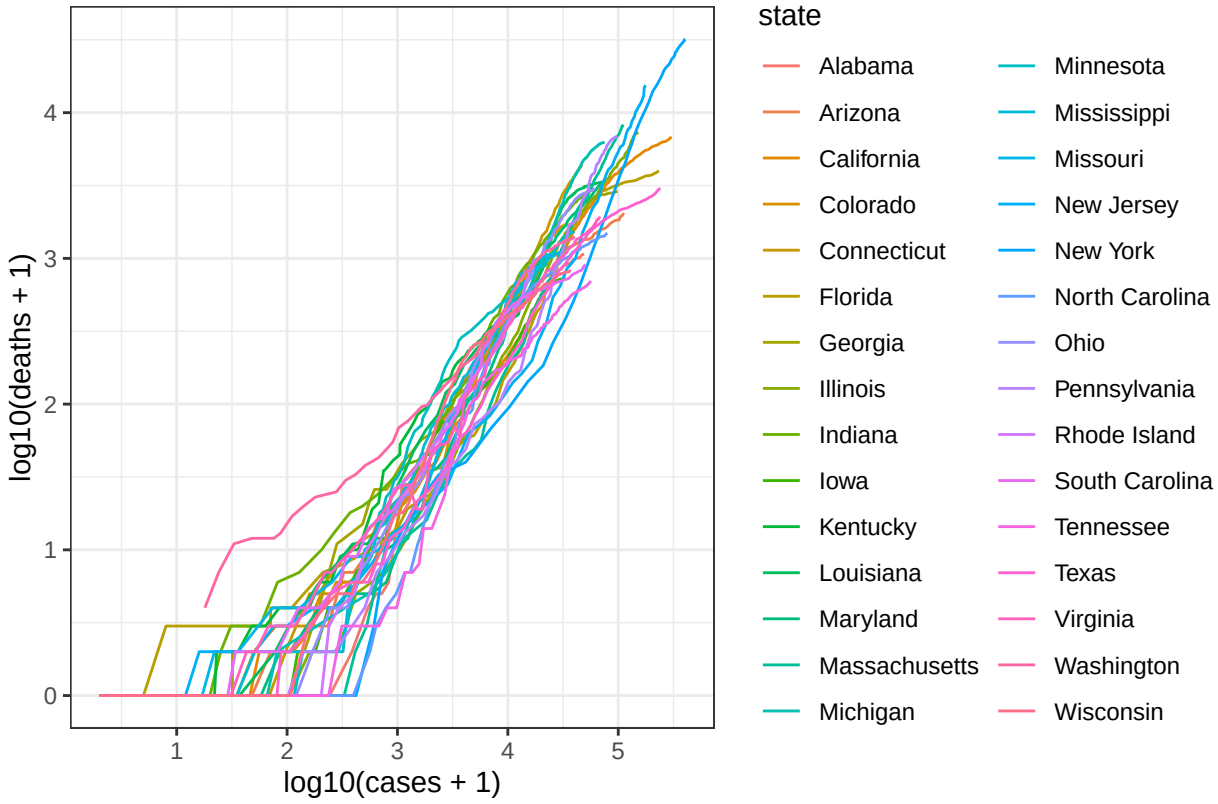
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Kentucky



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

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

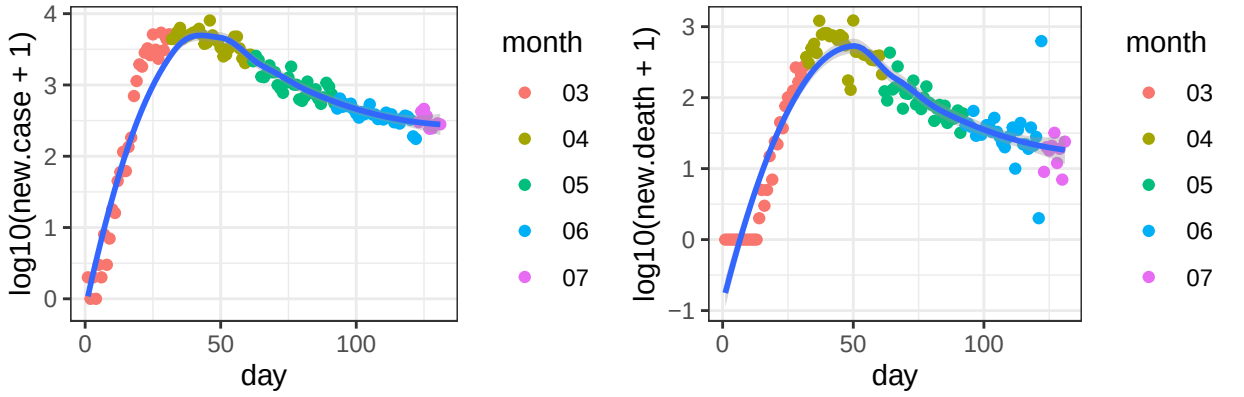
First check the 50 counties with the largest number of deaths.

```
##              date              county               state  fips  cases deaths
## 312079 2020-07-09       New York City            New York    NA 222723  22719
## 310863 2020-07-09                Cook            Illinois 17031  94005   4676
## 310460 2020-07-09         Los Angeles          California  6037 124738   3689
## 311565 2020-07-09               Wayne            Michigan 26163  23725   2746
## 312078 2020-07-09              Nassau            New York 36059  42164   2699
## 312004 2020-07-09               Essex          New Jersey 34013  19177   2060
## 312098 2020-07-09             Suffolk            New York 36103  41849   2037
## 311999 2020-07-09              Bergen          New Jersey 34003  20043   2020
## 311476 2020-07-09           Middlesex       Massachusetts 25017  24348   1903
## 312509 2020-07-09        Philadelphia        Pennsylvania 42101  27228   1633
## 312106 2020-07-09         Westchester            New York 36119  35182   1564
## 312006 2020-07-09              Hudson          New Jersey 34017  19216   1477
## 310561 2020-07-09            Hartford         Connecticut  9003  11866   1385
## 310560 2020-07-09           Fairfield         Connecticut  9001  16886   1380
## 312009 2020-07-09           Middlesex          New Jersey 34023  17158   1361
## 312017 2020-07-09               Union          New Jersey 34039  16646   1341
## 312013 2020-07-09             Passaic          New Jersey 34031  17158   1216
## 311472 2020-07-09               Essex       Massachusetts 25009  16379   1133
## 311545 2020-07-09             Oakland            Michigan 26125  12488   1097
## 310616 2020-07-09          Miami-Dade             Florida 12086  55960   1092
## 310564 2020-07-09           New Haven         Connecticut  9009  12509   1079
## 311480 2020-07-09             Suffolk       Massachusetts 25025  20172   1013
```

```
## 310358 2020-07-09             Maricopa                  Arizona  4013  73165    1012
## 312012 2020-07-09                Ocean               New Jersey 34029   9846     992
## 311482 2020-07-09            Worcester            Massachusetts 25027  12583     954
## 311478 2020-07-09              Norfolk            Massachusetts 25021   9384     950
## 311532 2020-07-09              Macomb                 Michigan 26099   7930     928
## 312010 2020-07-09            Monmouth               New Jersey 34025   9469     831
## 312504 2020-07-09          Montgomery             Pennsylvania 42091   8749     822
## 312011 2020-07-09              Morris               New Jersey 34027   7018     817
## 311593 2020-07-09            Hennepin                Minnesota 27053  12867     789
## 312530 2020-07-09          Providence             Rhode Island 44007  13144     774
## 311458 2020-07-09          Montgomery                 Maryland 24031  15541     754
## 310999 2020-07-09              Marion                  Indiana 18097  11943     736
## 312481 2020-07-09            Delaware             Pennsylvania 42045   7496     706
## 311459 2020-07-09      Prince George's                Maryland 24033  19942     700
## 311479 2020-07-09            Plymouth            Massachusetts 25023   8777     674
## 311474 2020-07-09             Hampden            Massachusetts 25013   6932     671
## 313176 2020-07-09                King               Washington 53033  11488     634
## 312064 2020-07-09                Erie                 New York 36029   7624     601
## 312008 2020-07-09              Mercer               New Jersey 34021   7809     596
## 311470 2020-07-09             Bristol            Massachusetts 25005   8399     594
## 311838 2020-07-09           St. Louis                 Missouri 29189   7469     593
## 310623 2020-07-09          Palm Beach                  Florida 12099  18654     578
## 312467 2020-07-09               Bucks             Pennsylvania 42017   6003     571
## 310573 2020-07-09 District of Columbia District of Columbia 11001  10679     568
## 312015 2020-07-09            Somerset               New Jersey 34035   5038     544
## 312001 2020-07-09              Camden               New Jersey 34007   7652     540
## 311396 2020-07-09             Orleans                Louisiana 22071   8344     539
## 310473 2020-07-09           Riverside               California  6065  24042     533
```

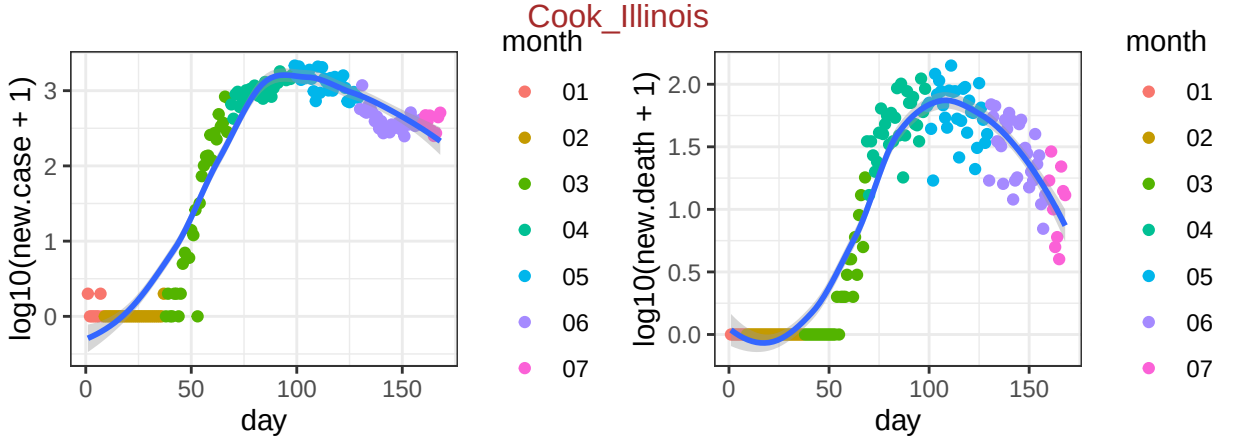
For these 50 counties, I check the number of new cases and the number of new deaths.

New York City_New York



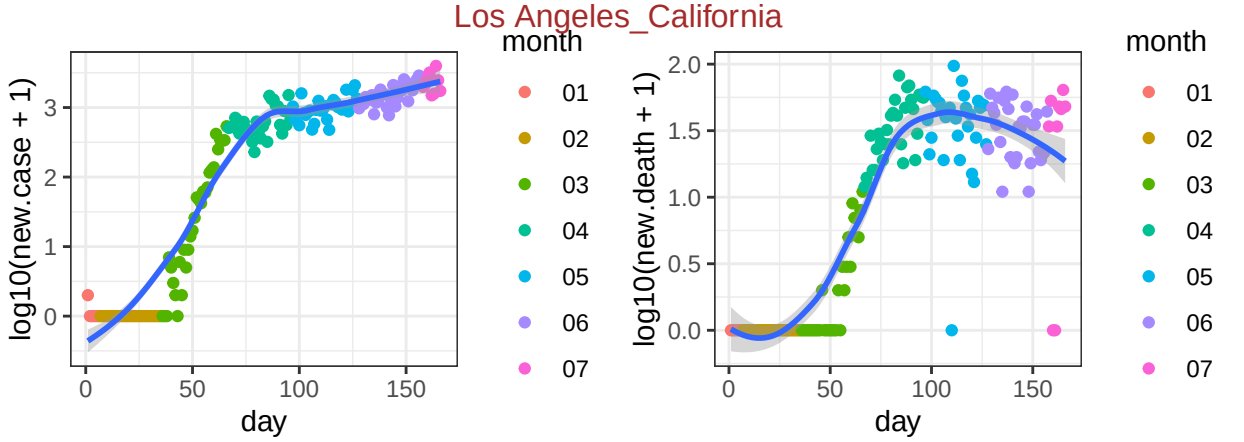
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



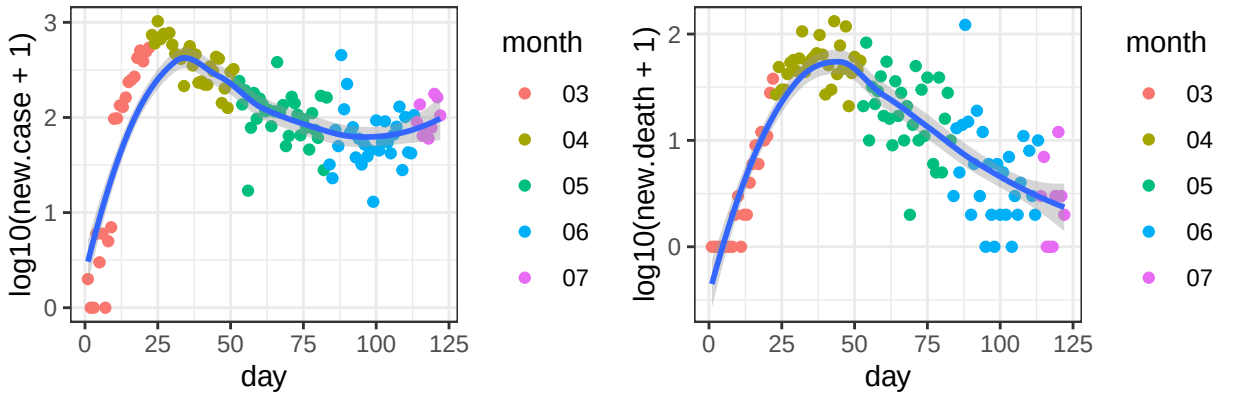
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Los Angeles_California



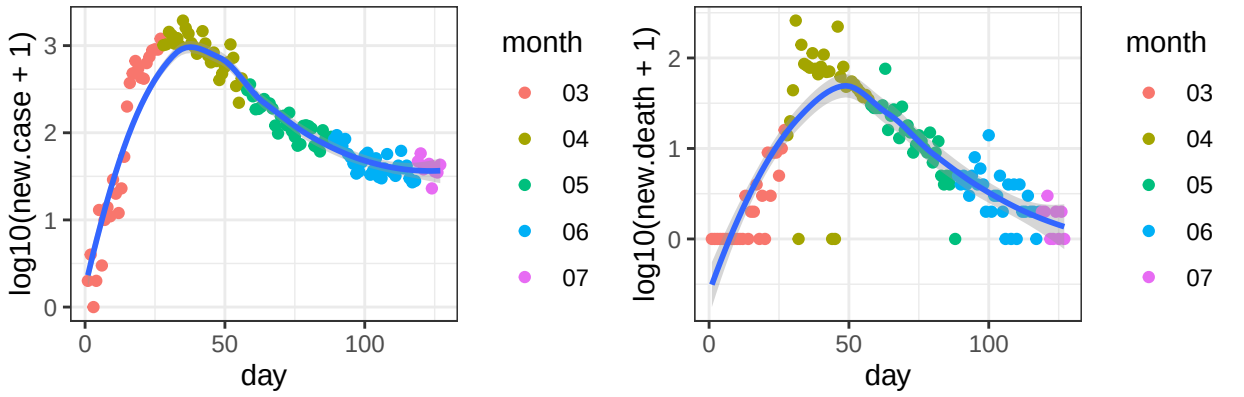
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Wayne_Michigan



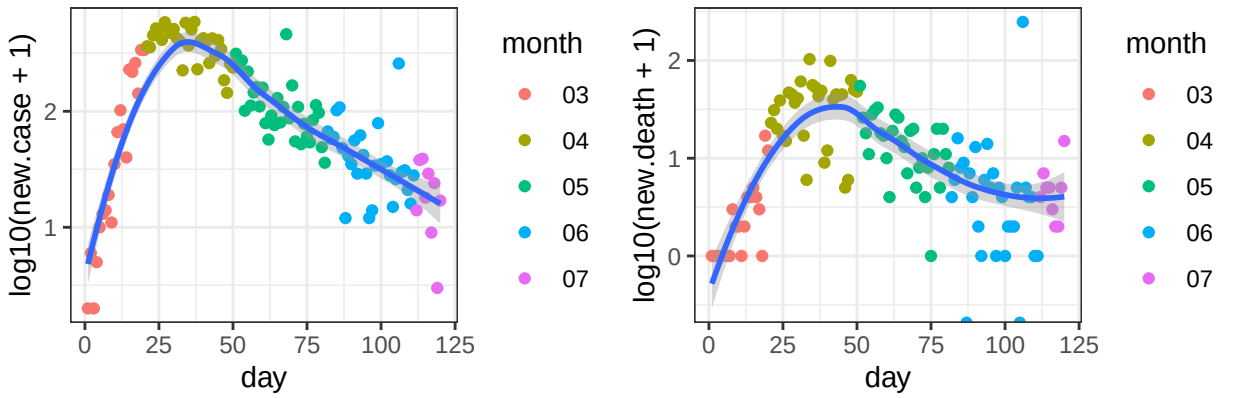
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Nassau_New York



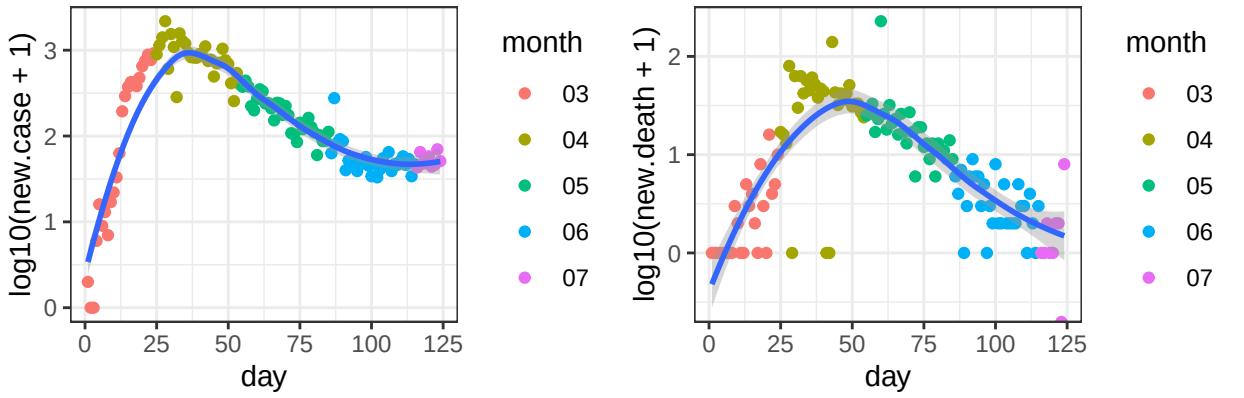
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



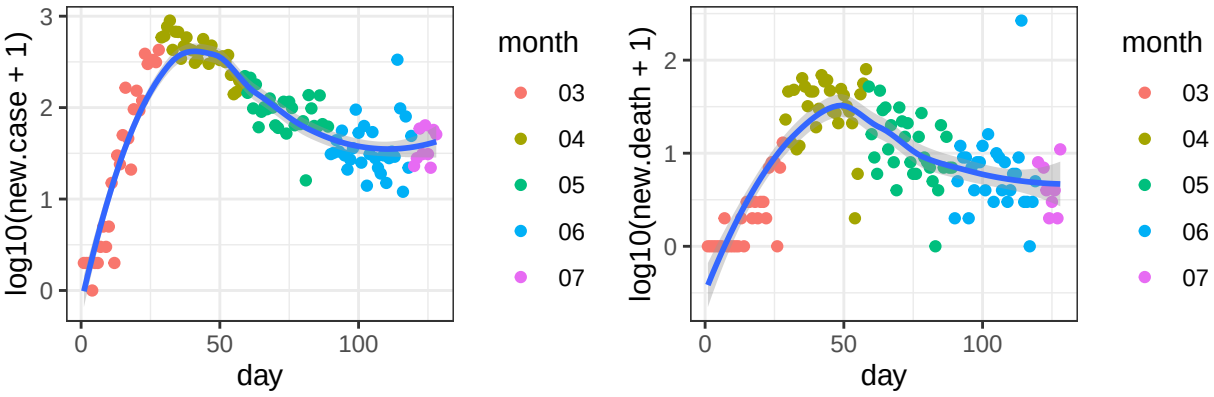
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Suffolk_New York



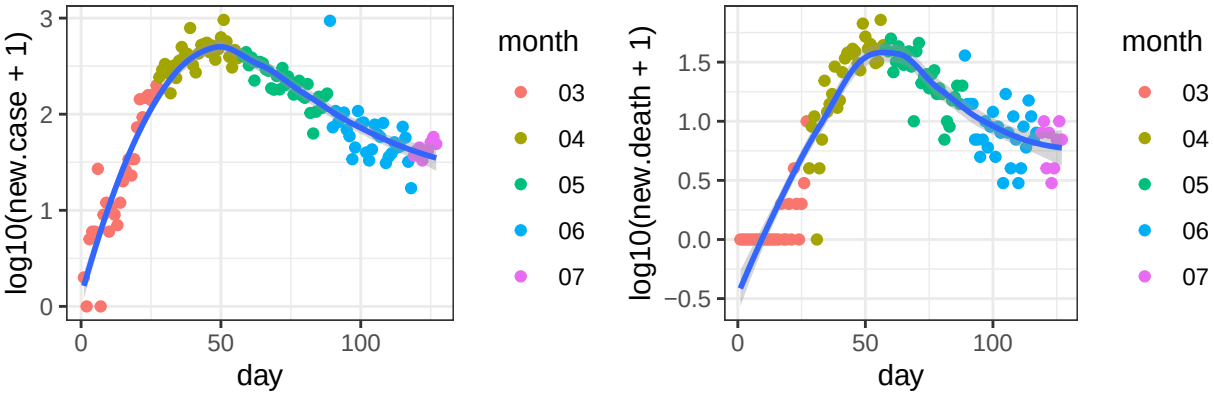
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Bergen_New Jersey



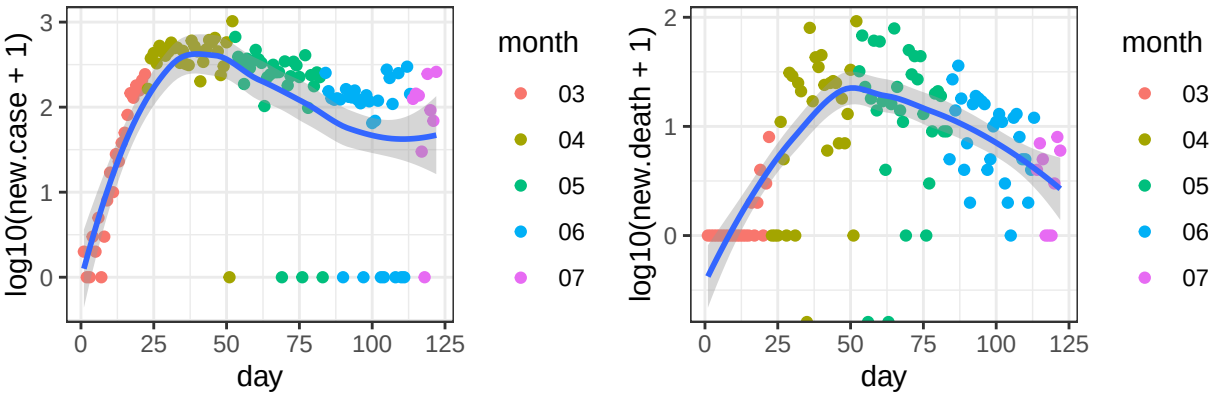
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Middlesex_Massachusetts



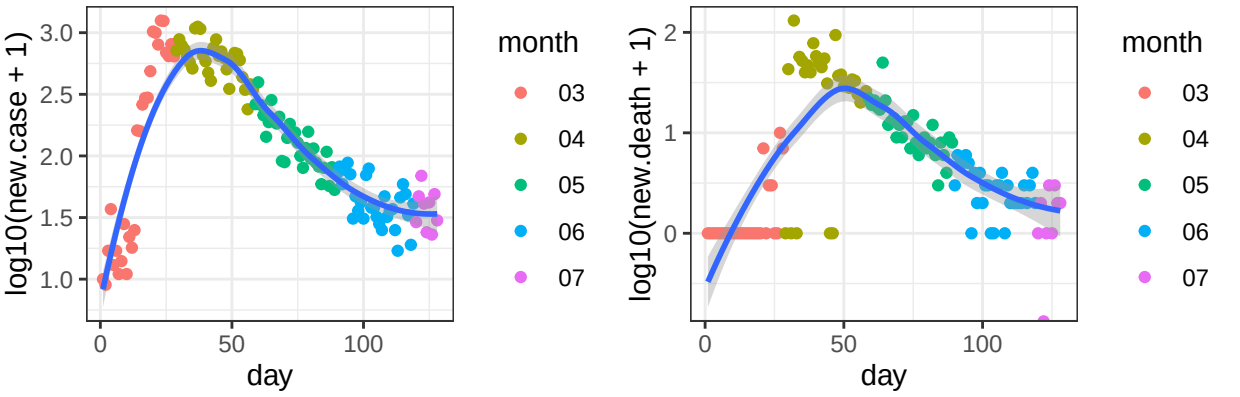
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Philadelphia_Pennsylvania



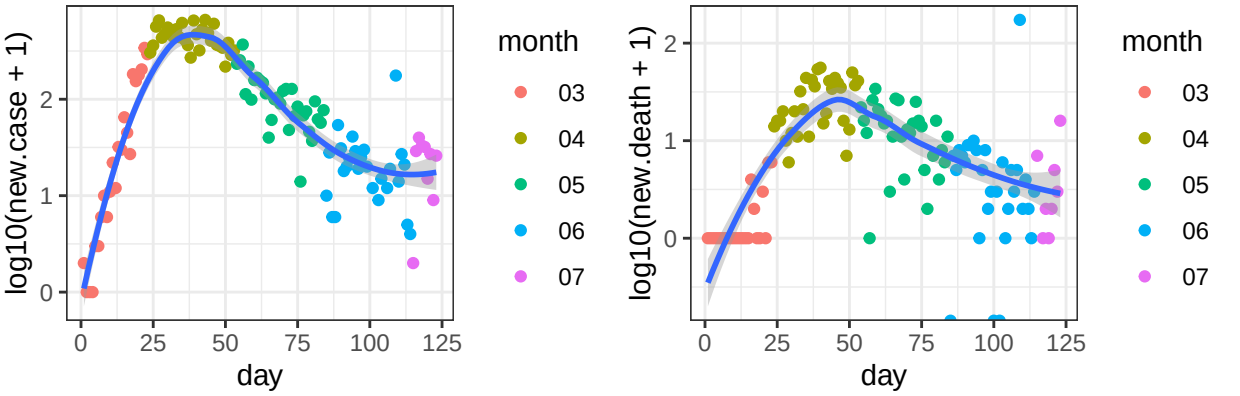
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Westchester_New York



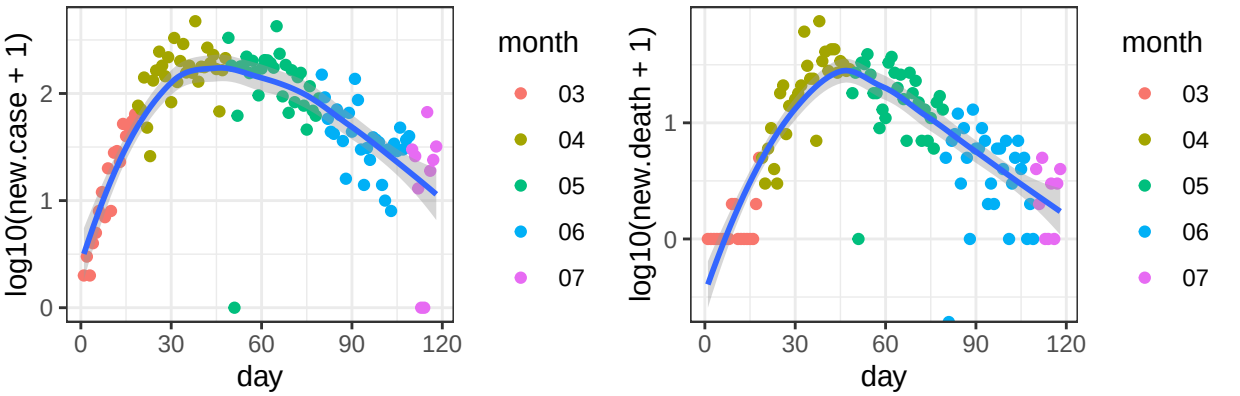
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Hudson_New Jersey



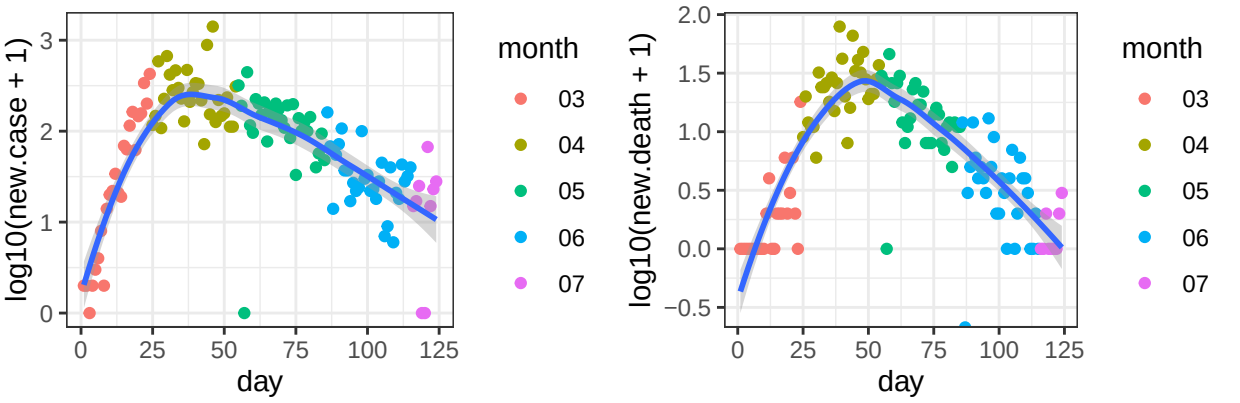
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Hartford_Connecticut



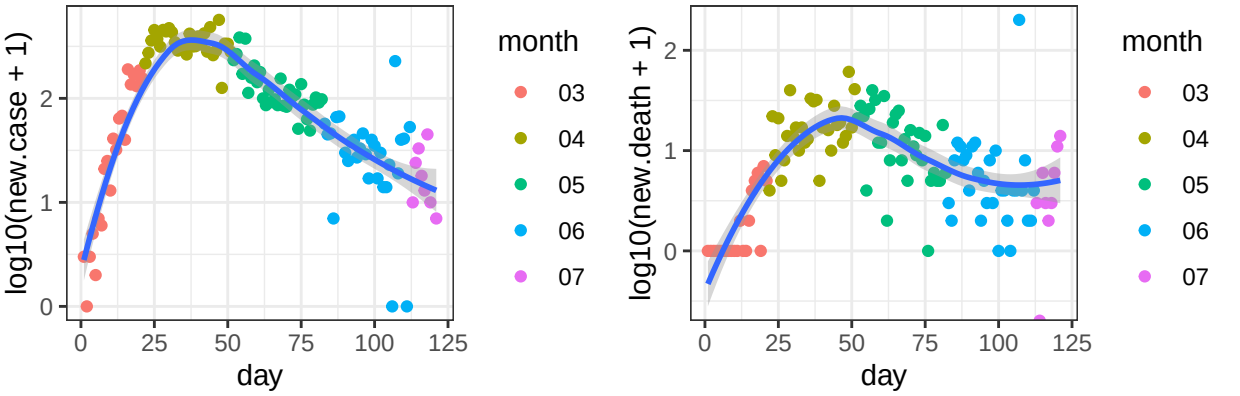
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Fairfield_Connecticut



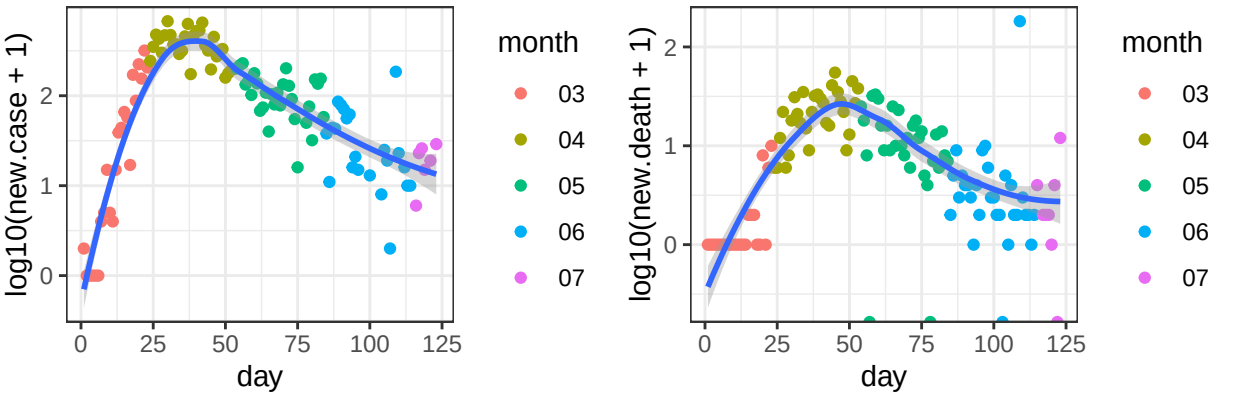
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Middlesex_New Jersey



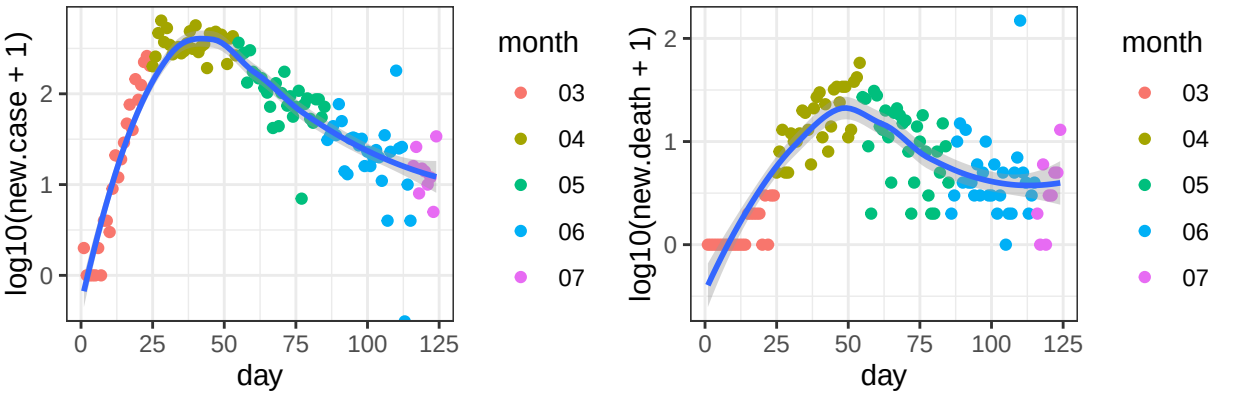
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Union_New Jersey



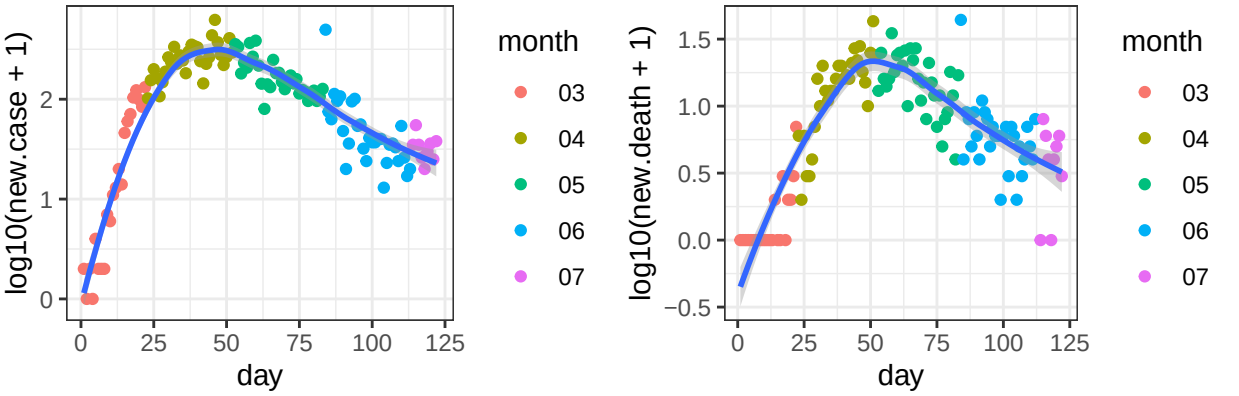
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Passaic_New Jersey



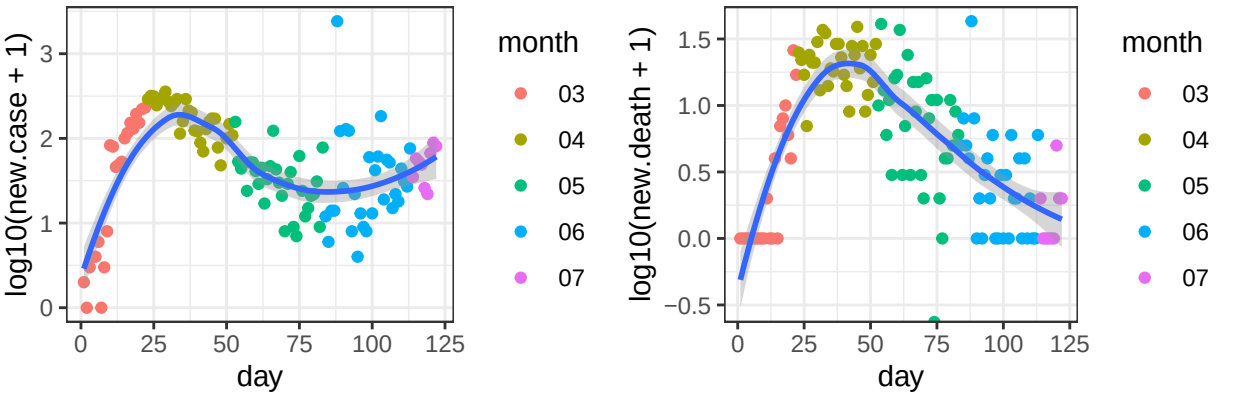
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Essex_Massachusetts



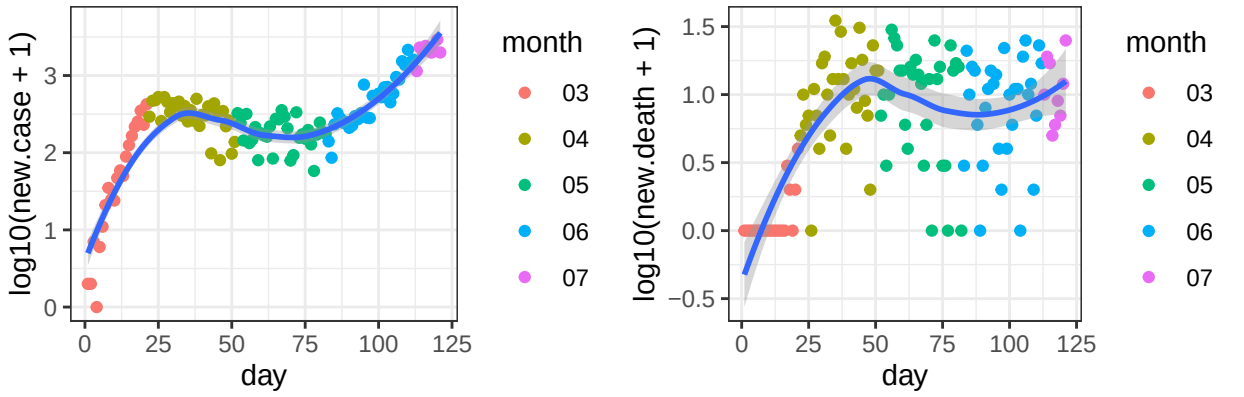
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Oakland_Michigan



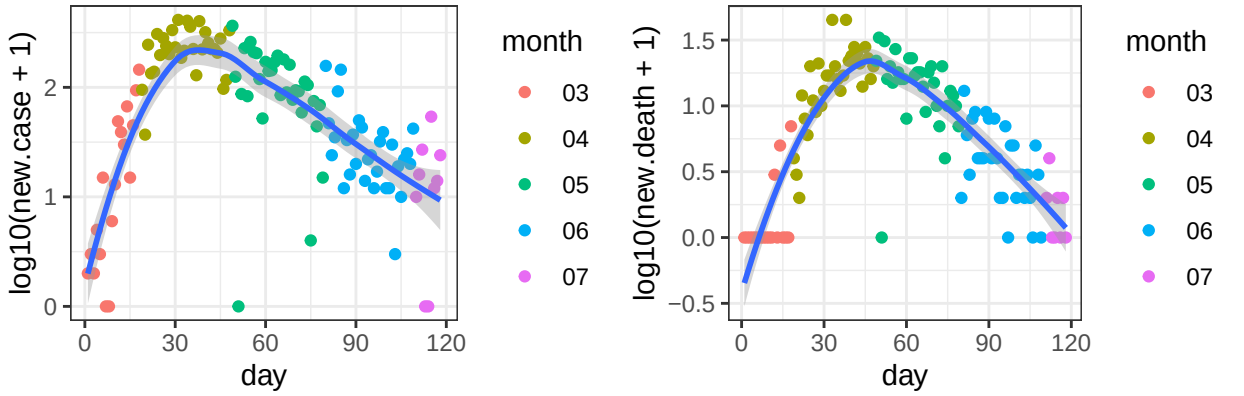
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Miami−Dade_Florida



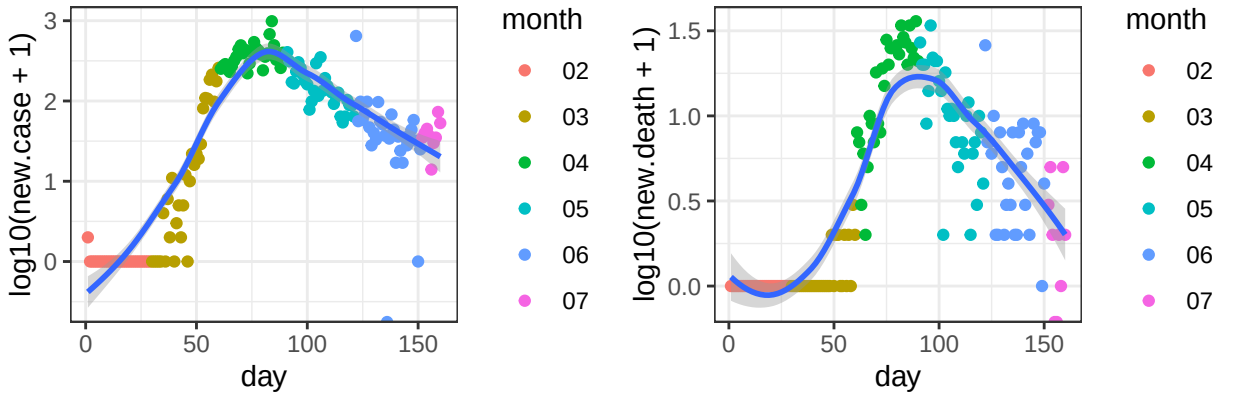
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

New Haven_Connecticut



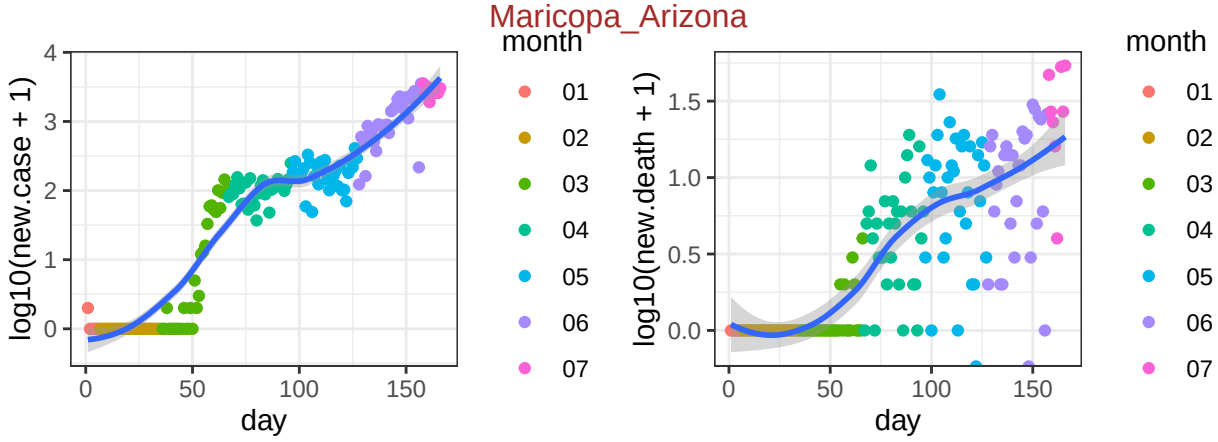
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Suffolk_Massachusetts



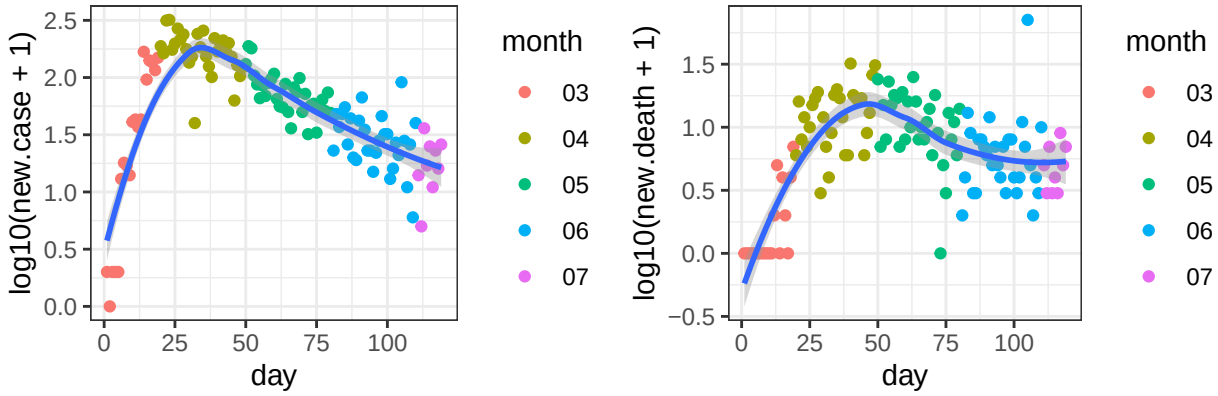
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Maricopa_Arizona



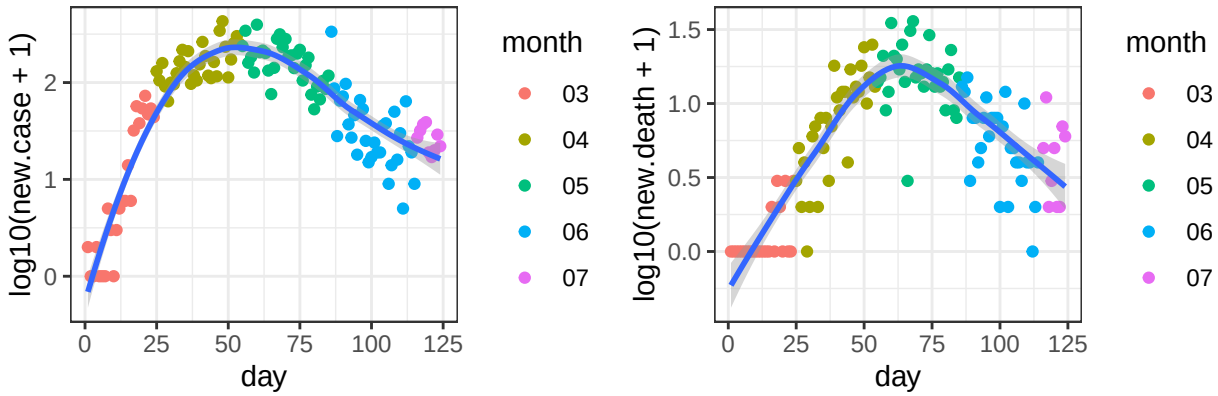
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Ocean_New Jersey



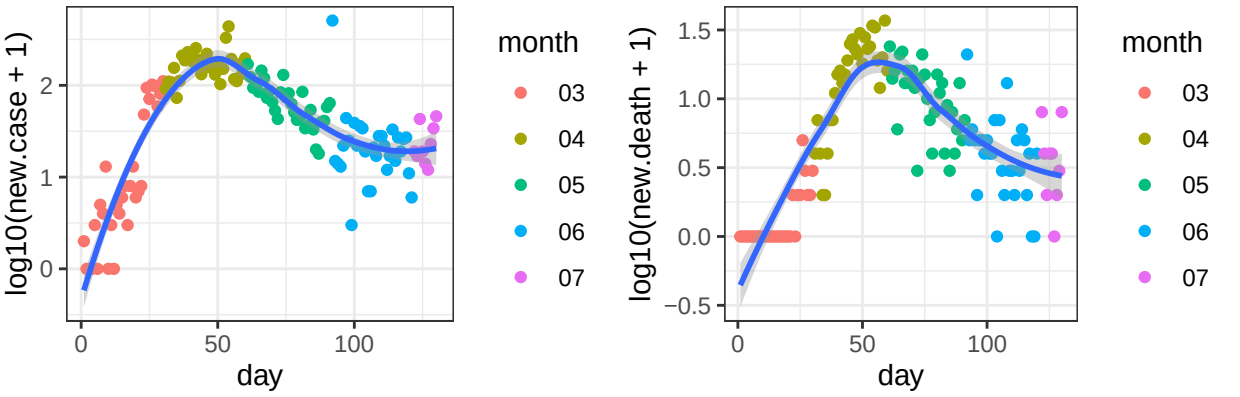
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Worcester_Massachusetts



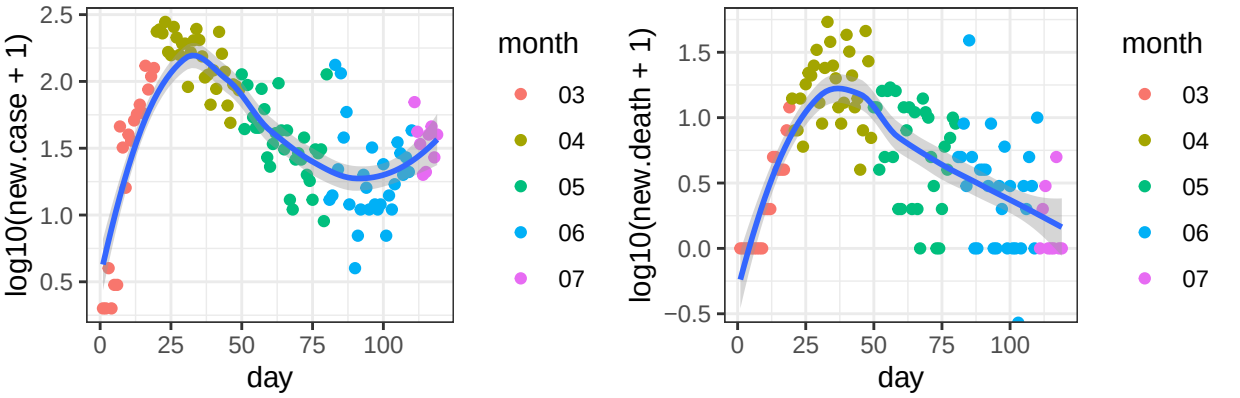
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Norfolk_Massachusetts



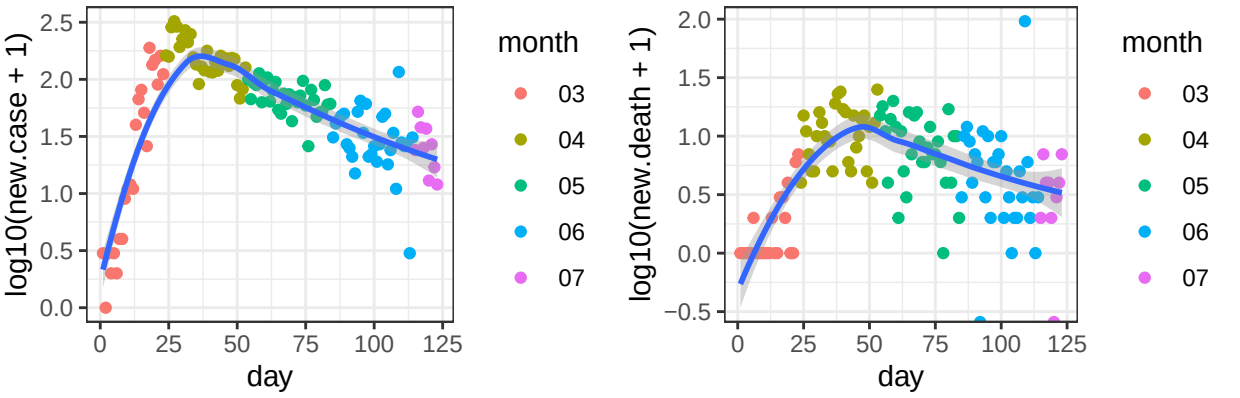
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Macomb_Michigan



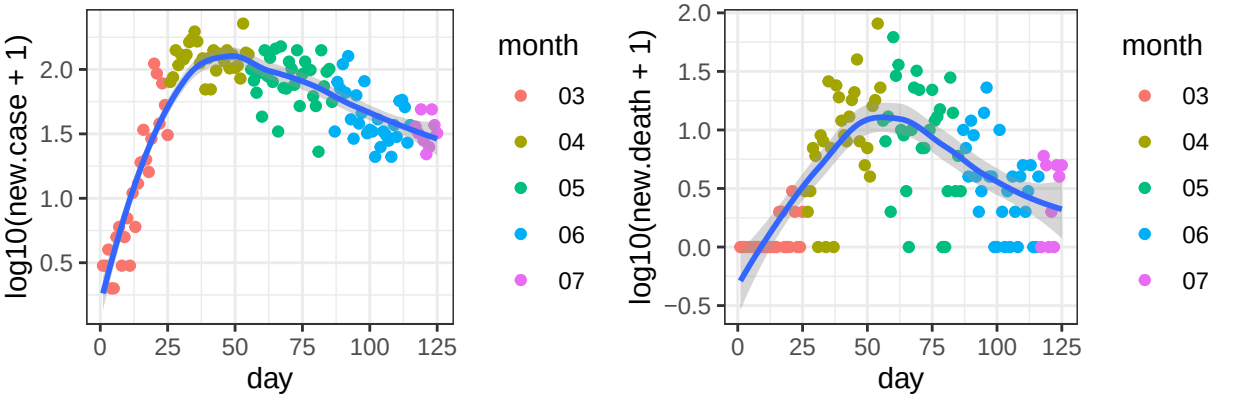
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Monmouth_New Jersey



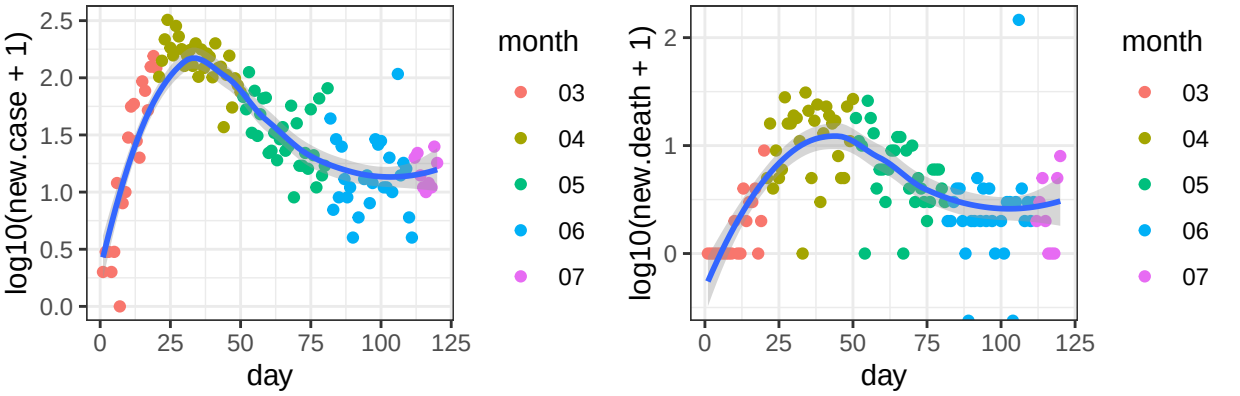
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Montgomery_Pennsylvania



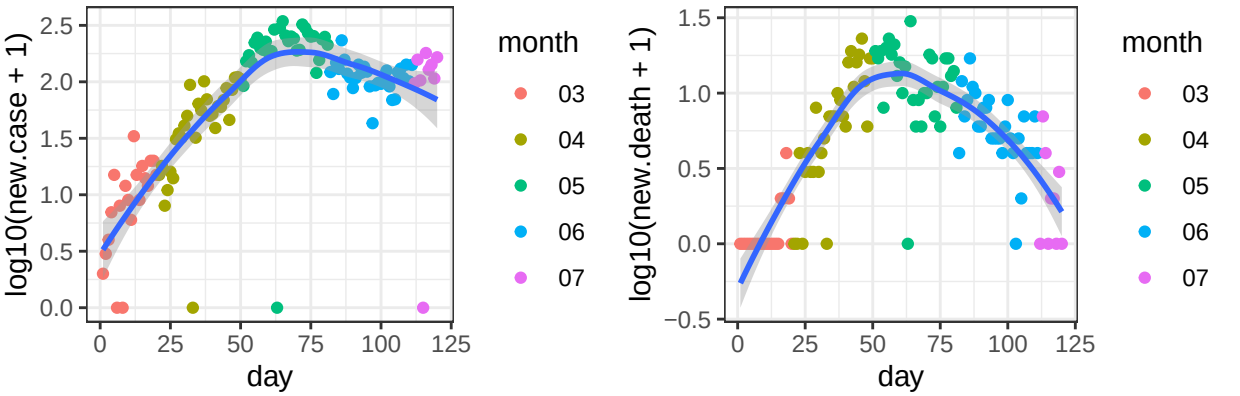
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Morris_New Jersey



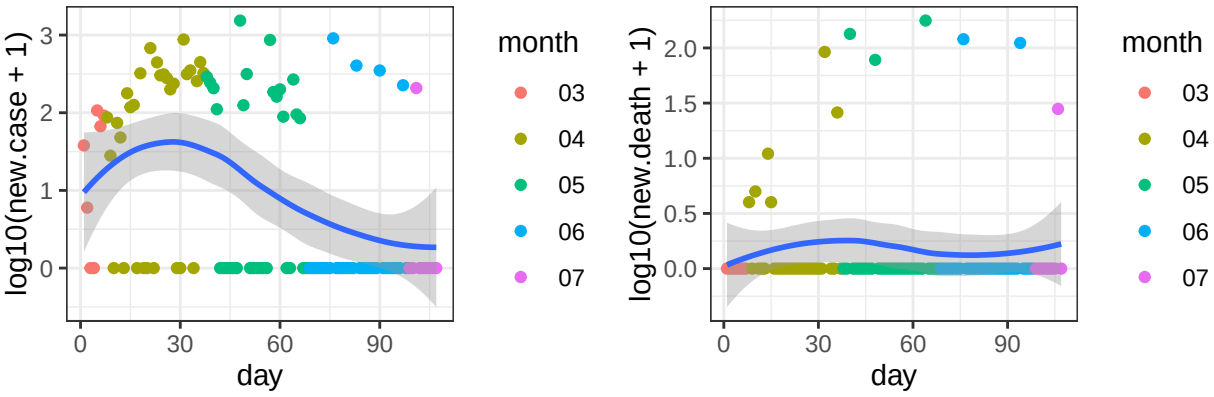
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Hennepin_Minnesota



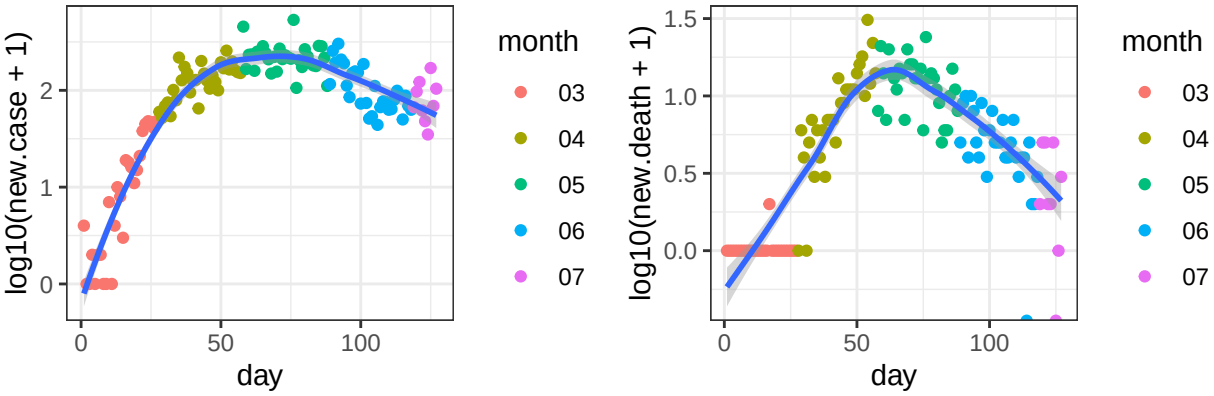
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Providence_Rhode Island



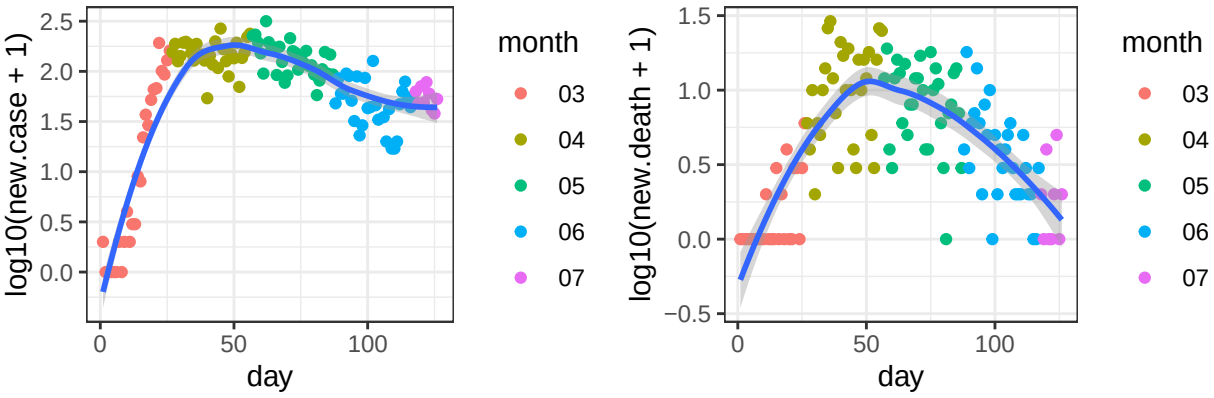
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Montgomery_Maryland



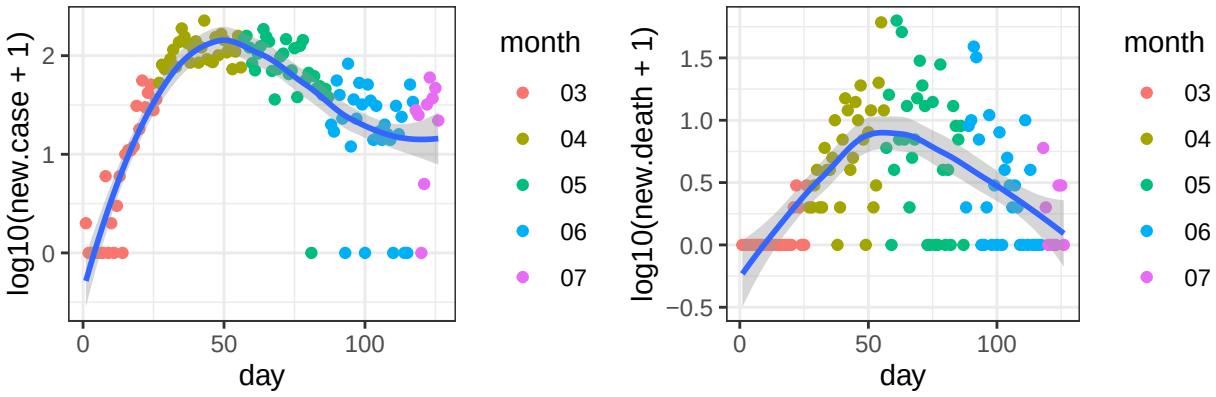
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Marion_Indiana



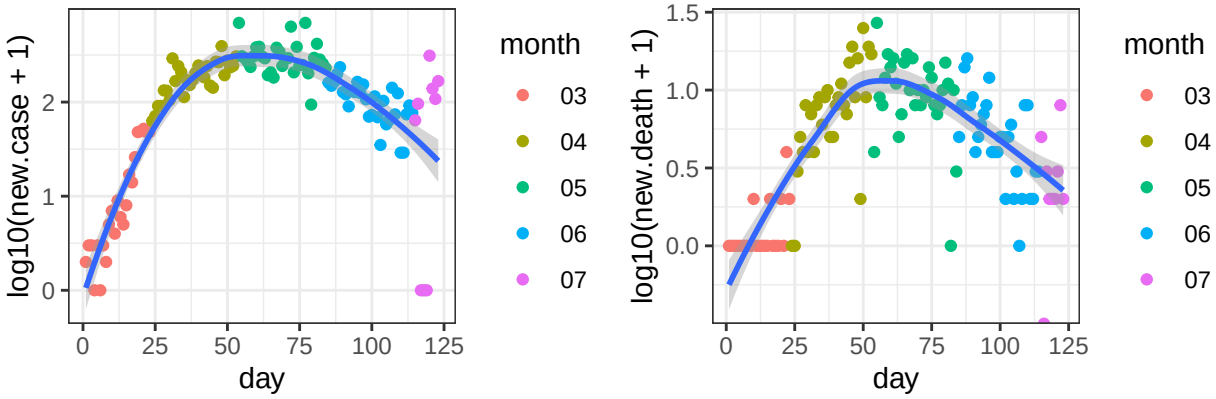
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Delaware_Pennsylvania



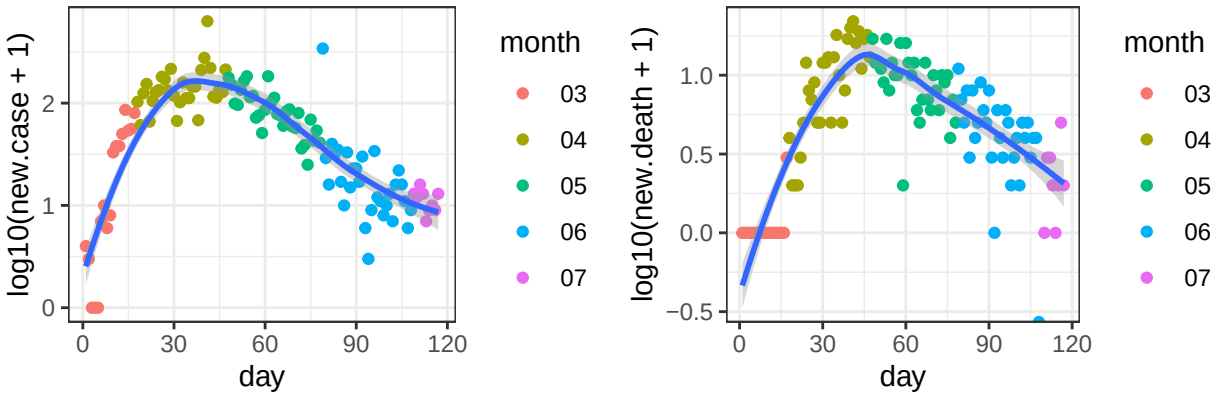
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Prince George's_Maryland



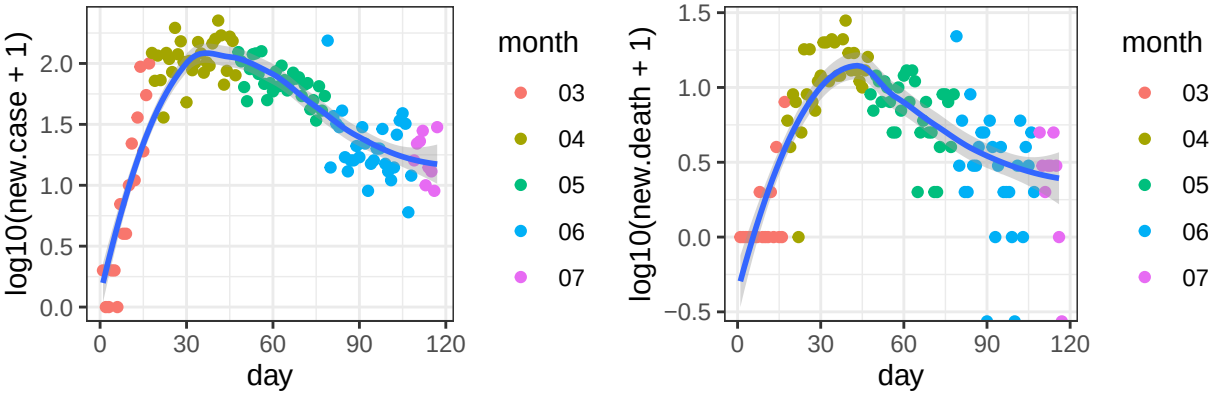
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Plymouth_Massachusetts



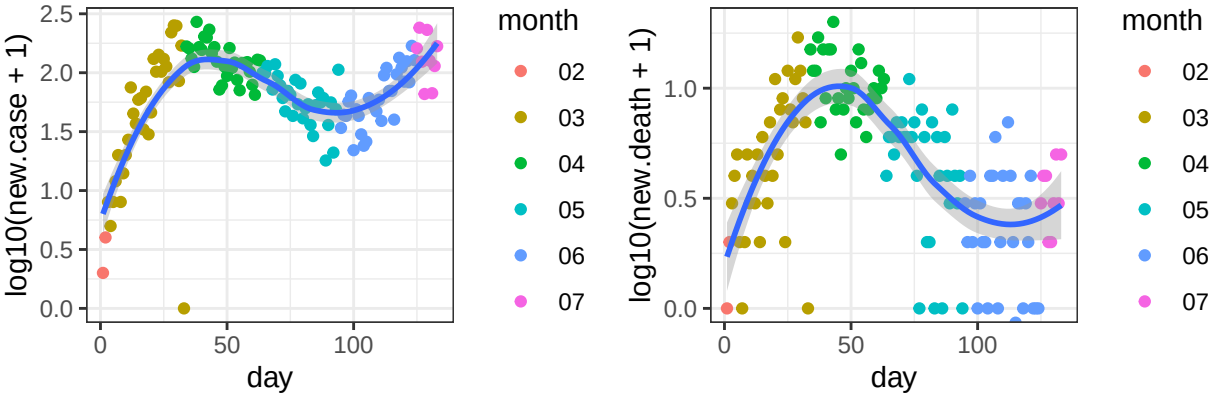
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Hampden_Massachusetts



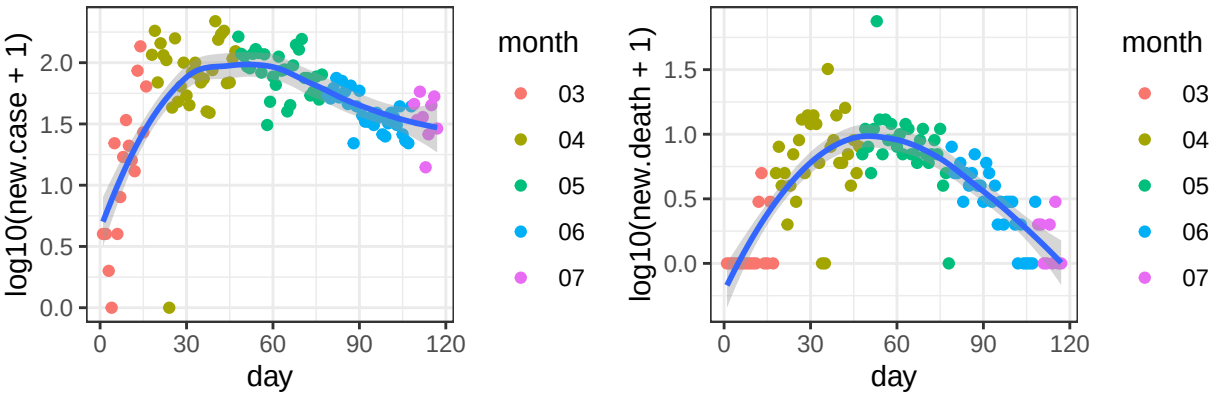
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

King_Washington



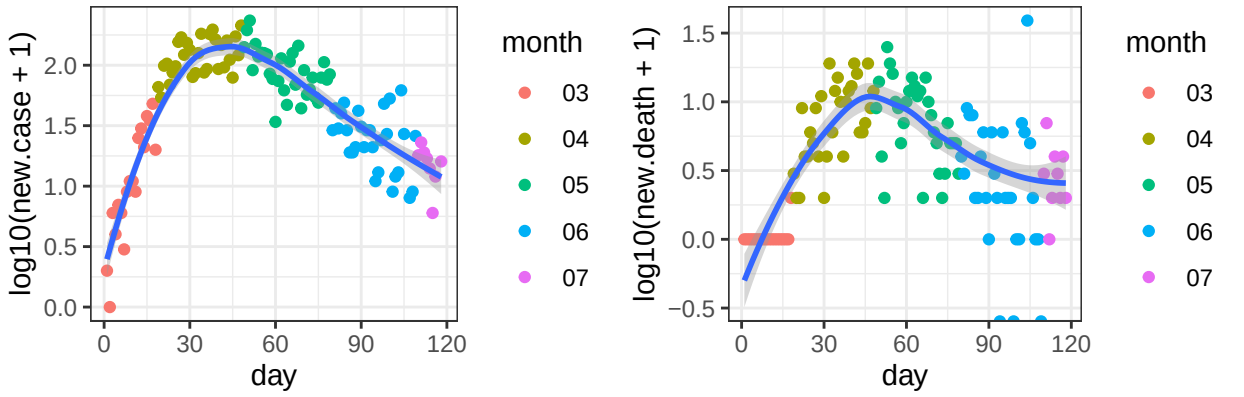
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

Erie_New York



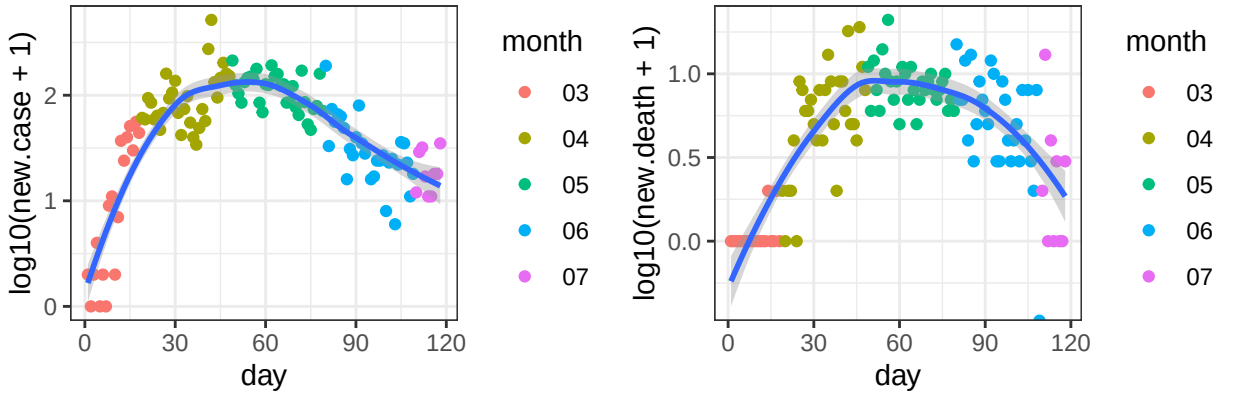
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Mercer_New Jersey



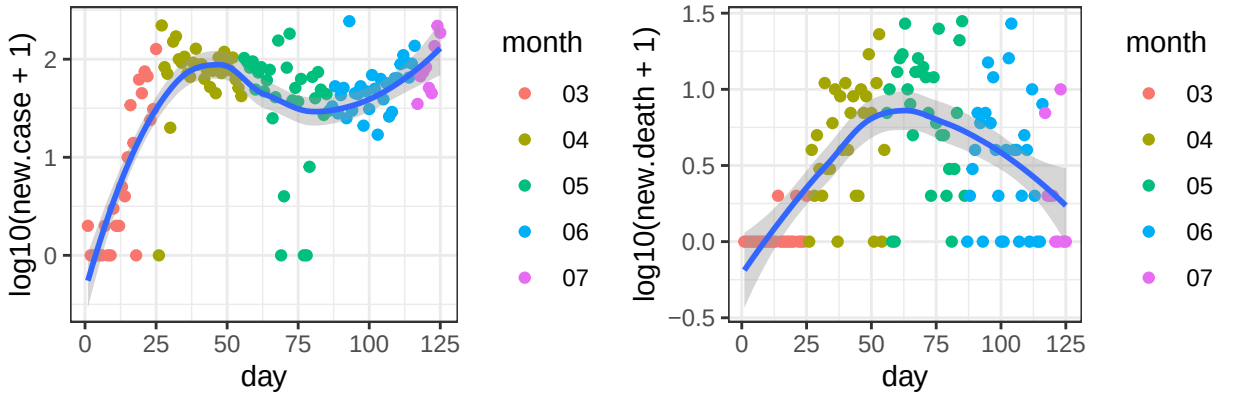
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Bristol_Massachusetts



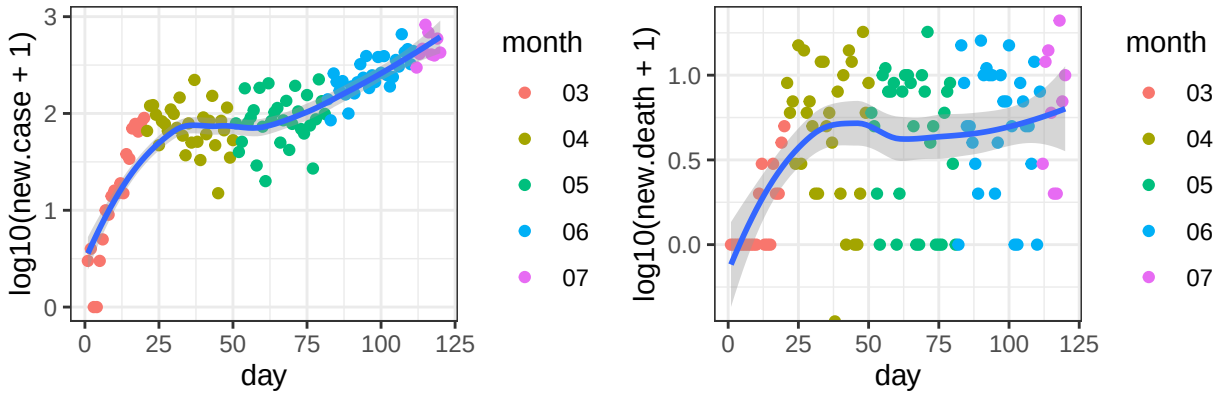
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

St. Louis_Missouri



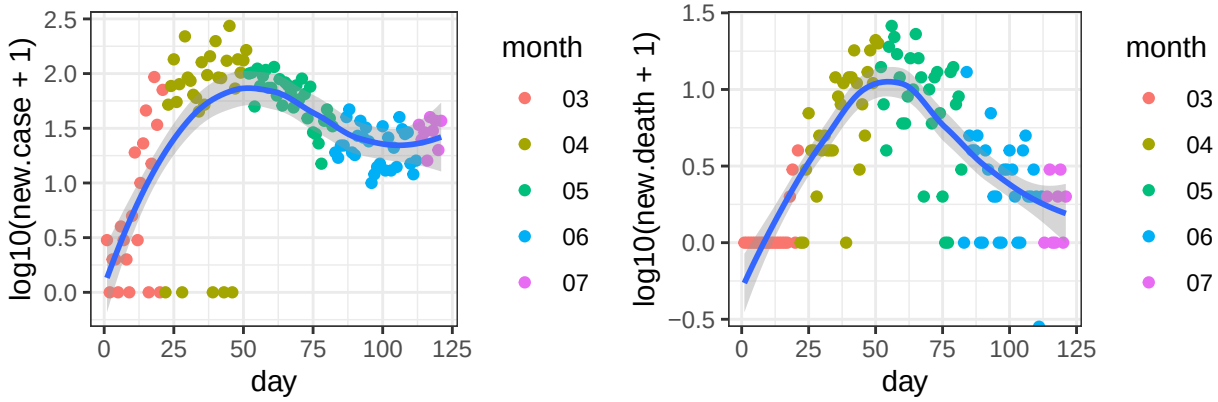
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Palm Beach_Florida



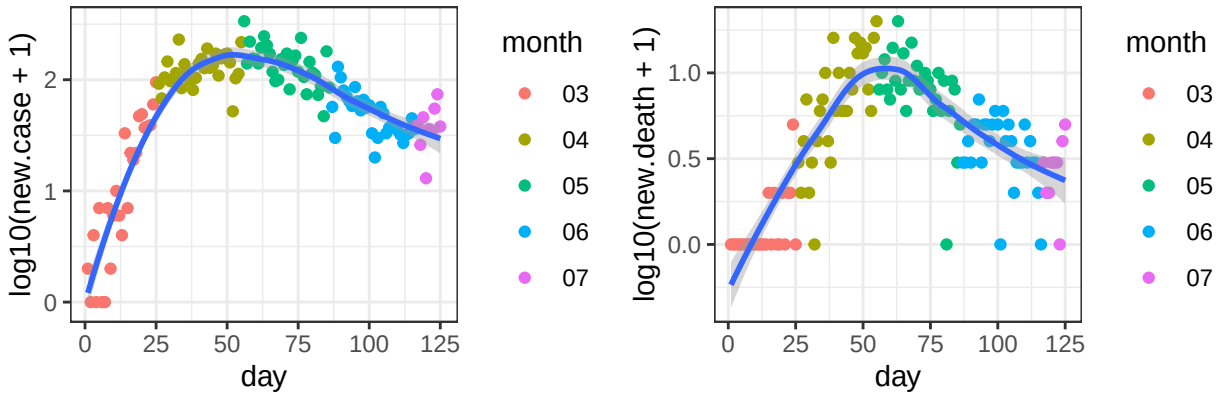
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Bucks_Pennsylvania



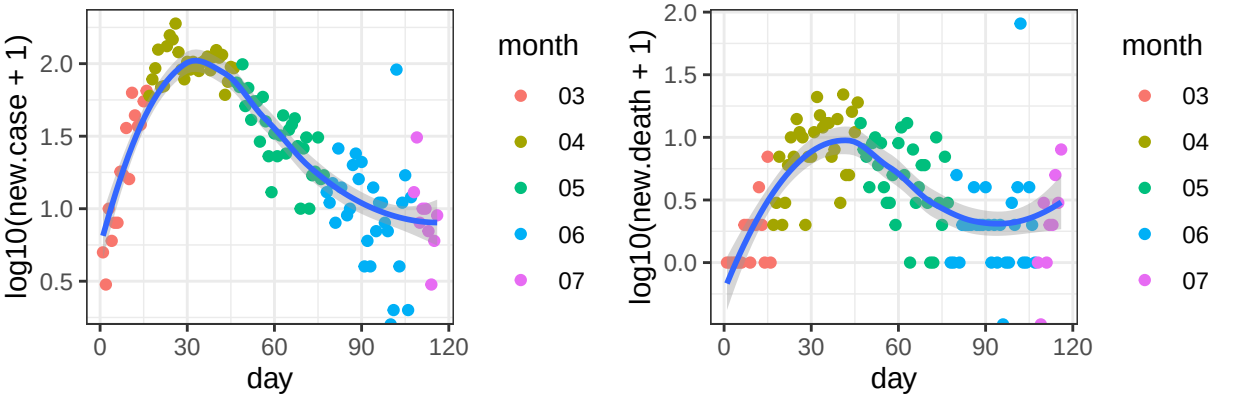
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

District of Columbia_District of Columbia



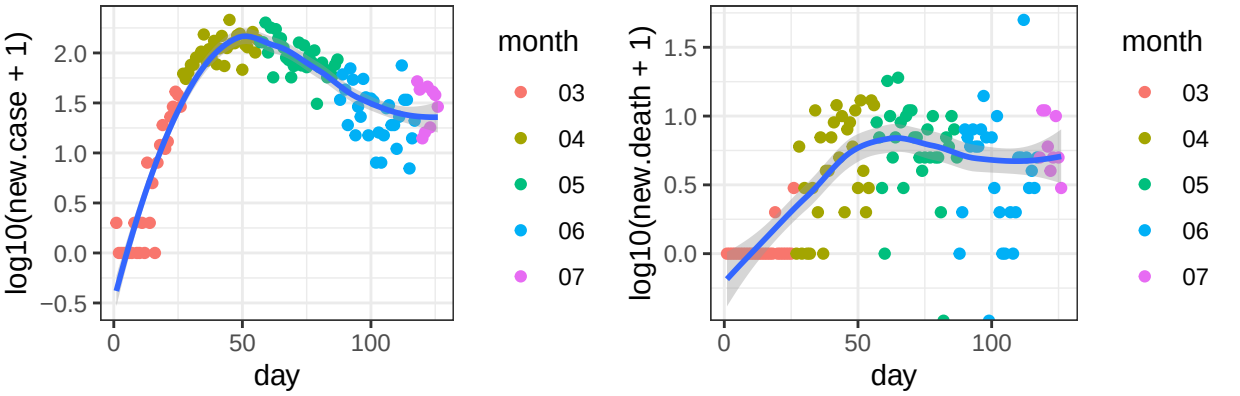
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Somerset_New Jersey



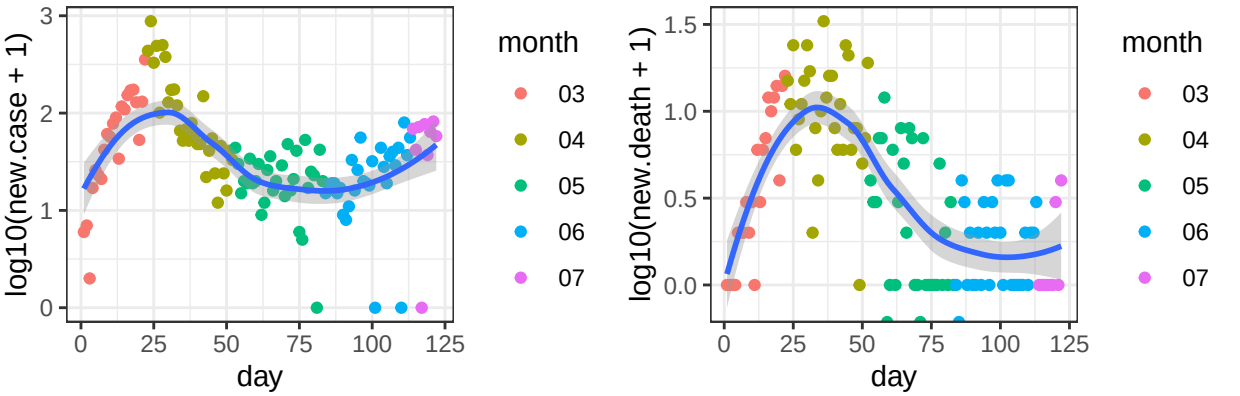
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−16

Camden_New Jersey

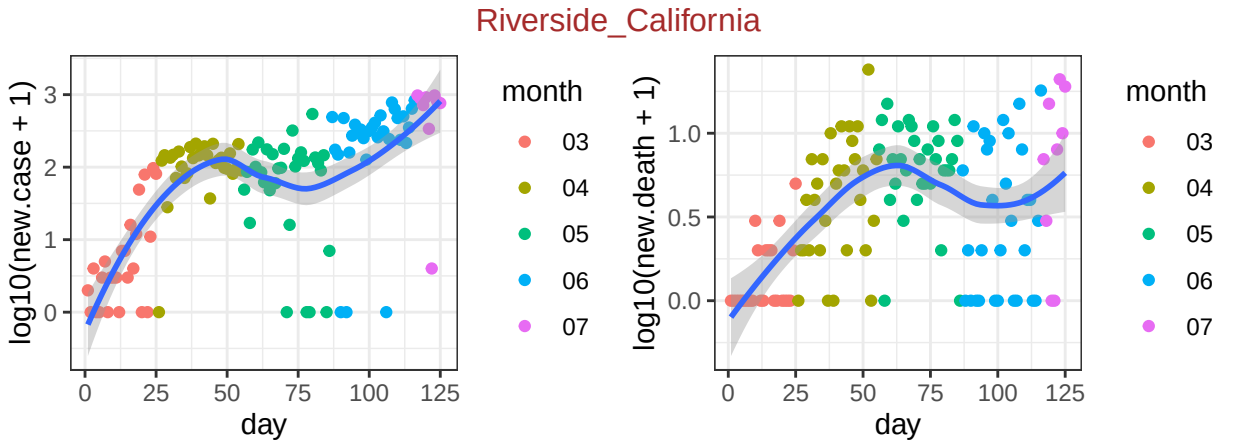


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Orleans_Louisiana



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

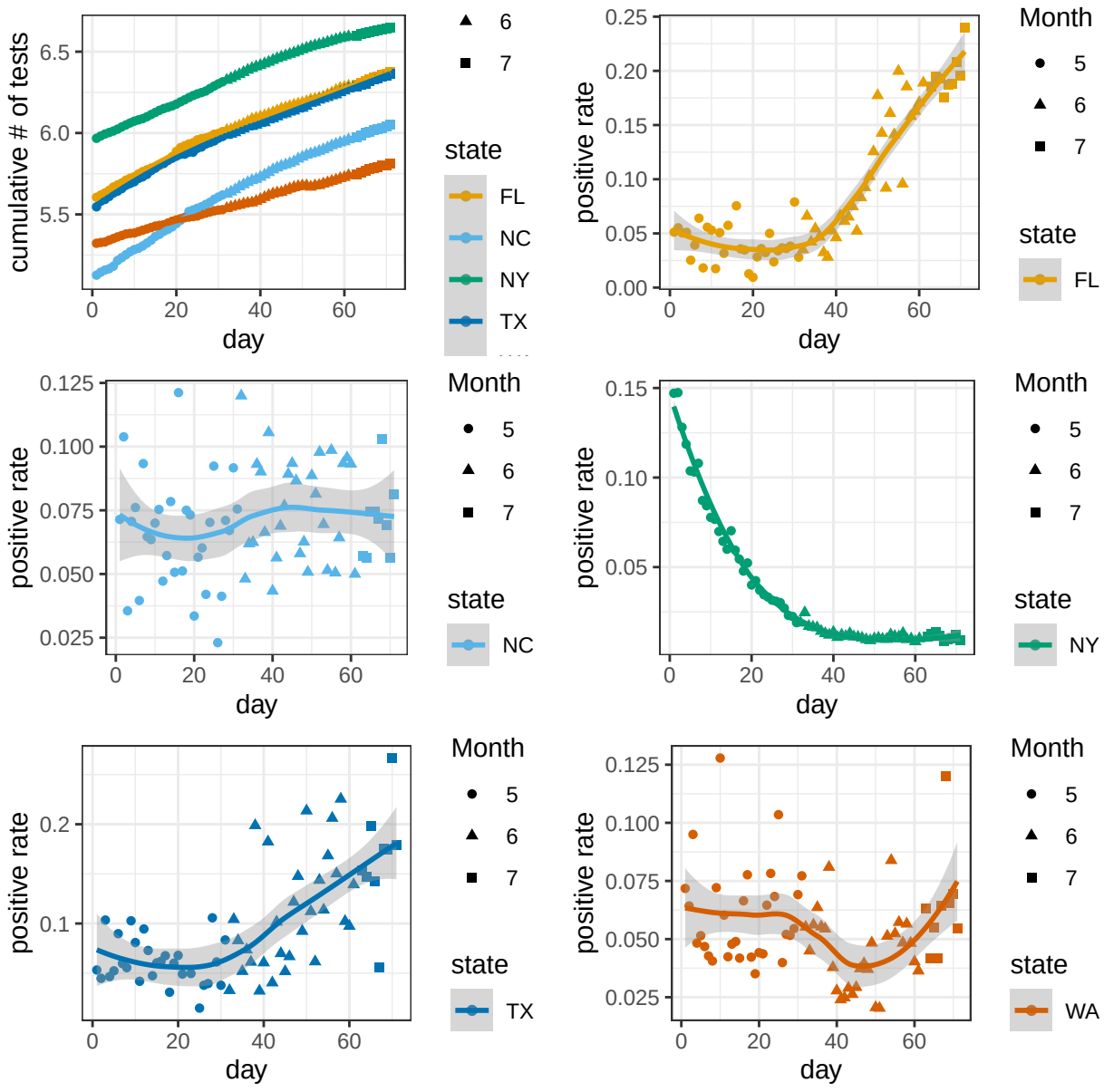


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

COVID Trackng

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

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



github.com/COVID19Tracking/, positive rate on 0709: 0.24(FL) 0.08(NC) 0.01(NY) 0.18(TX) 0.05(WA)

Session information

```
sessionInfo()
```

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

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