Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-06-07

Contents

Introduction

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

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

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

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

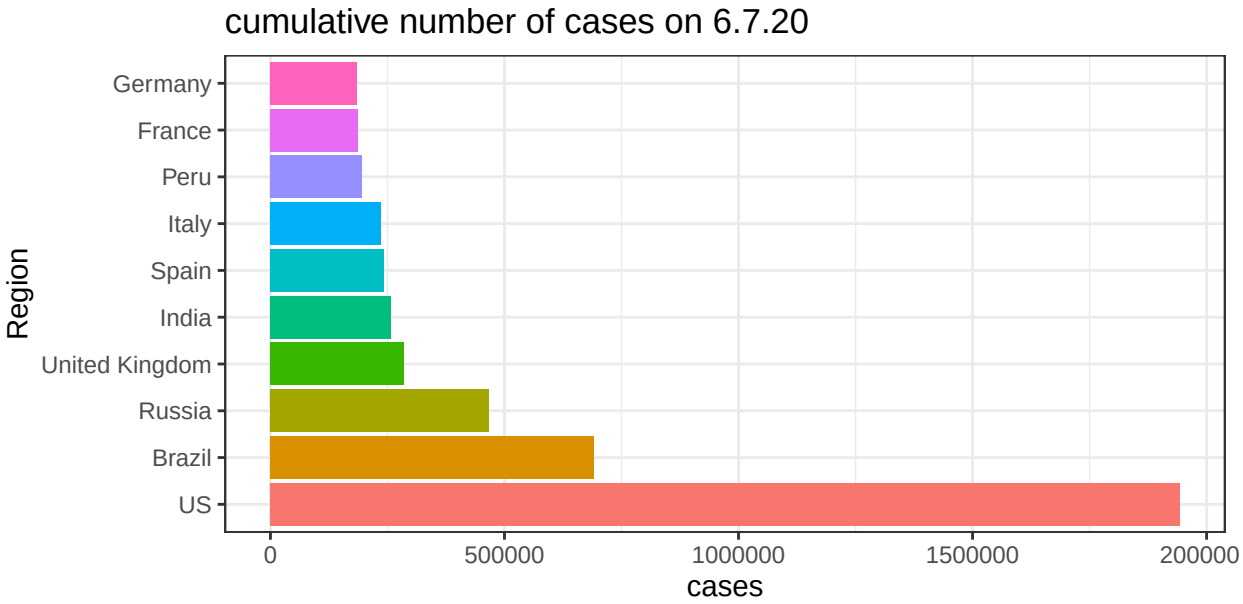
JHU

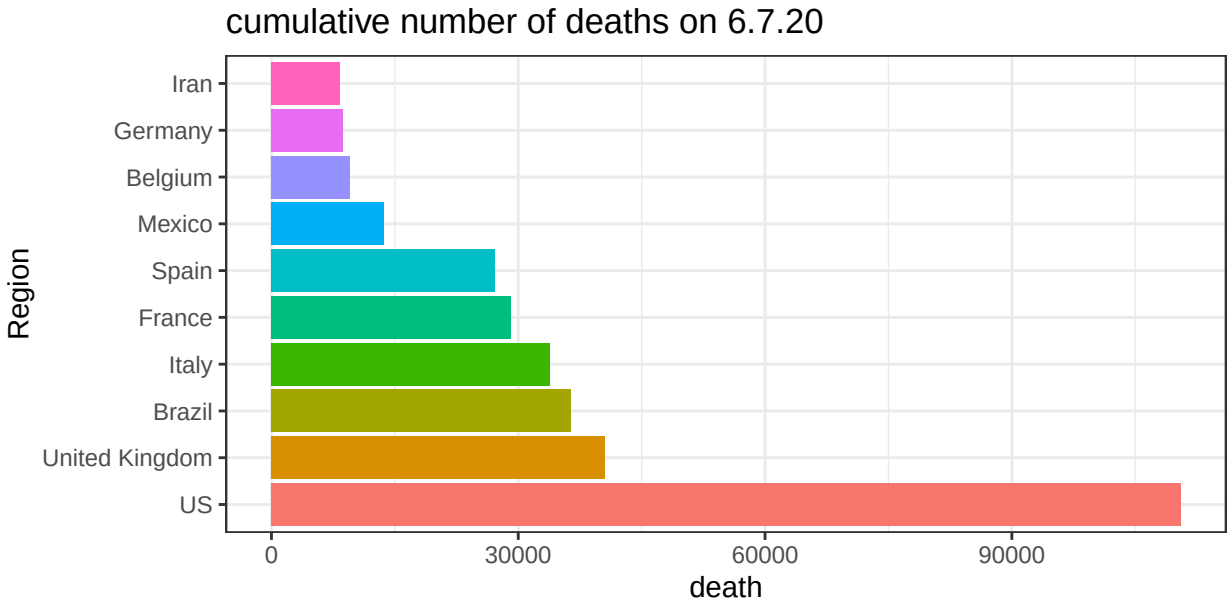
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

time series data

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

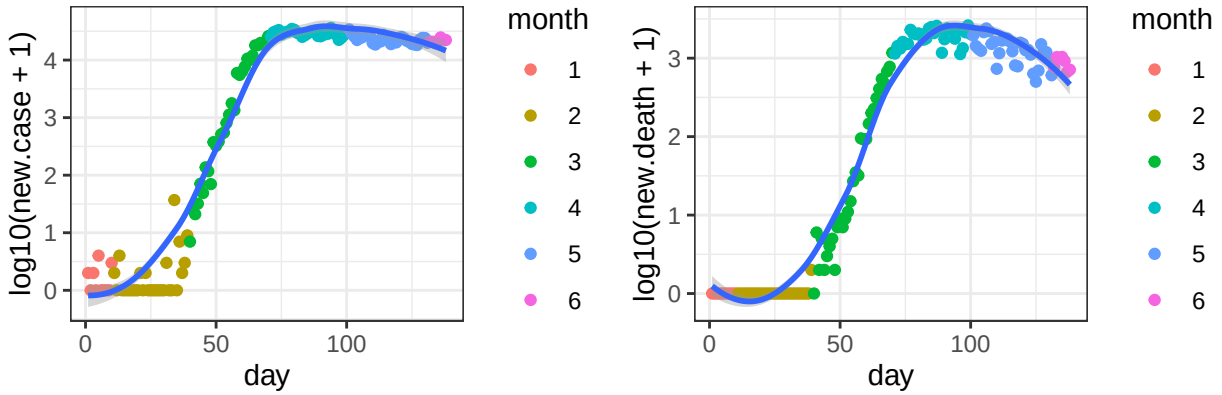
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





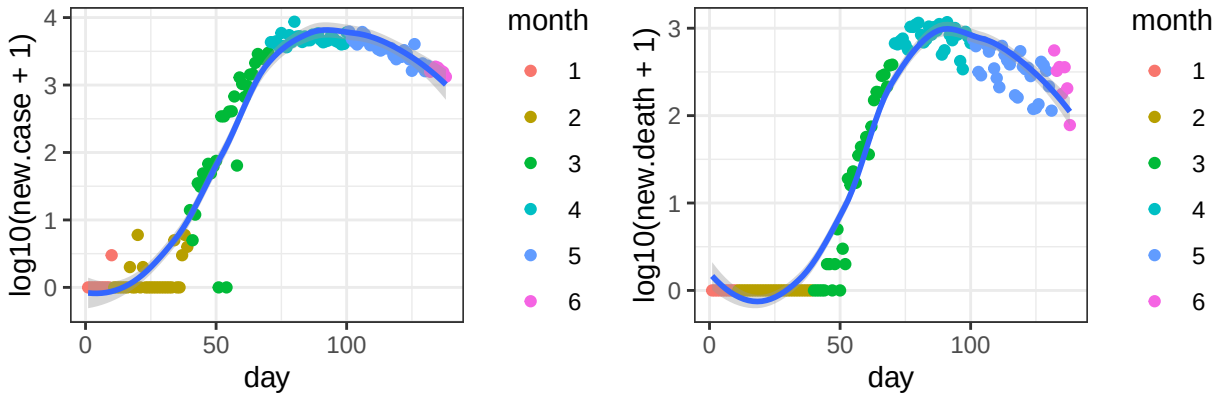
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



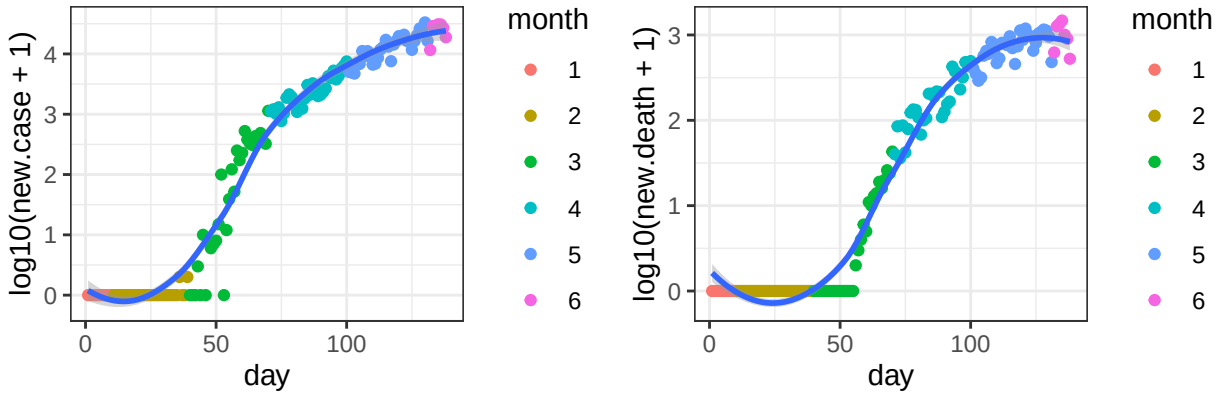
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



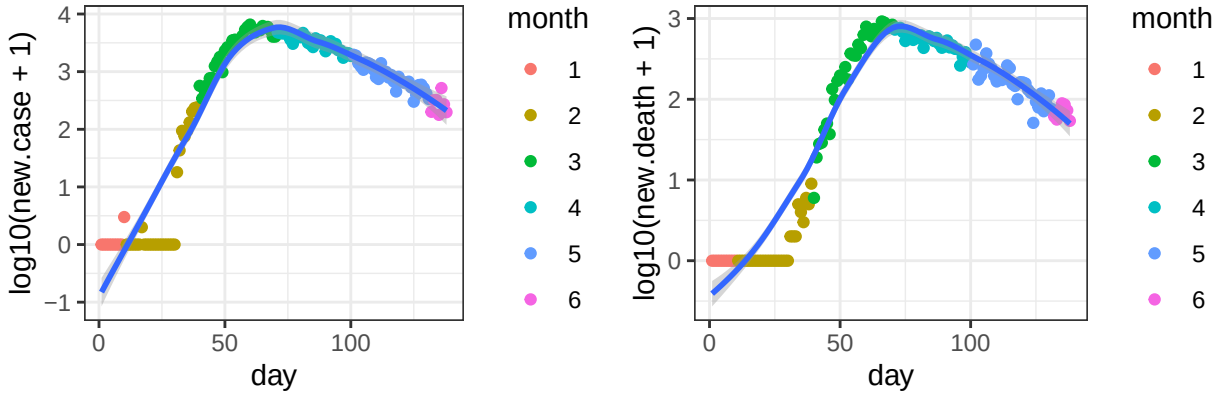
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Brazil



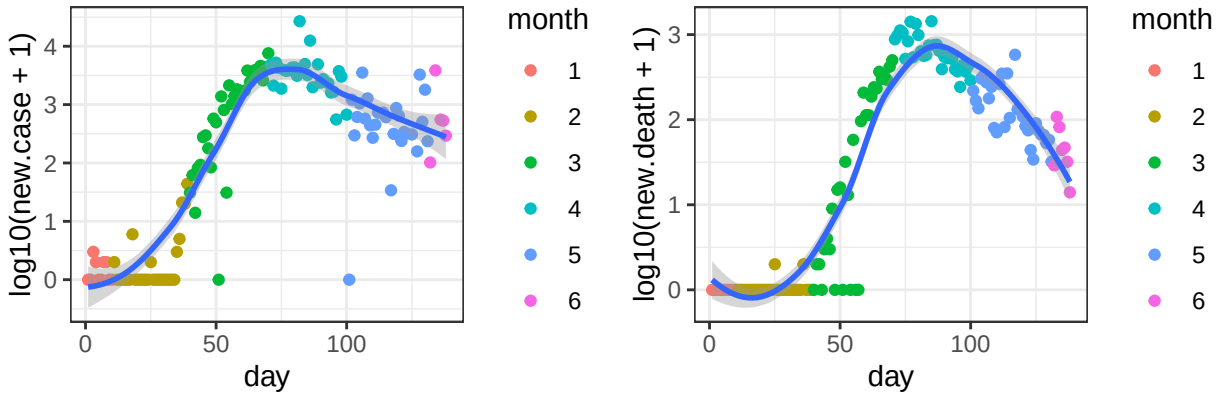
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



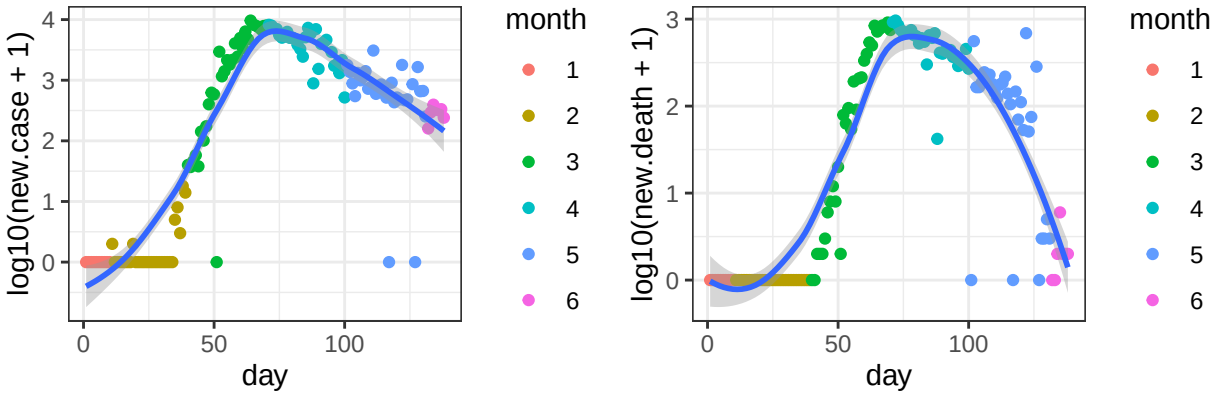
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



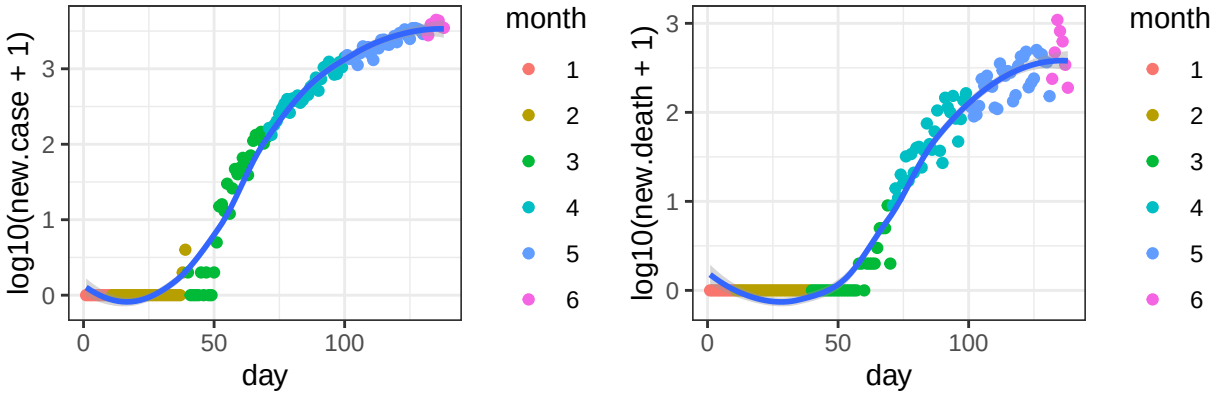
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain



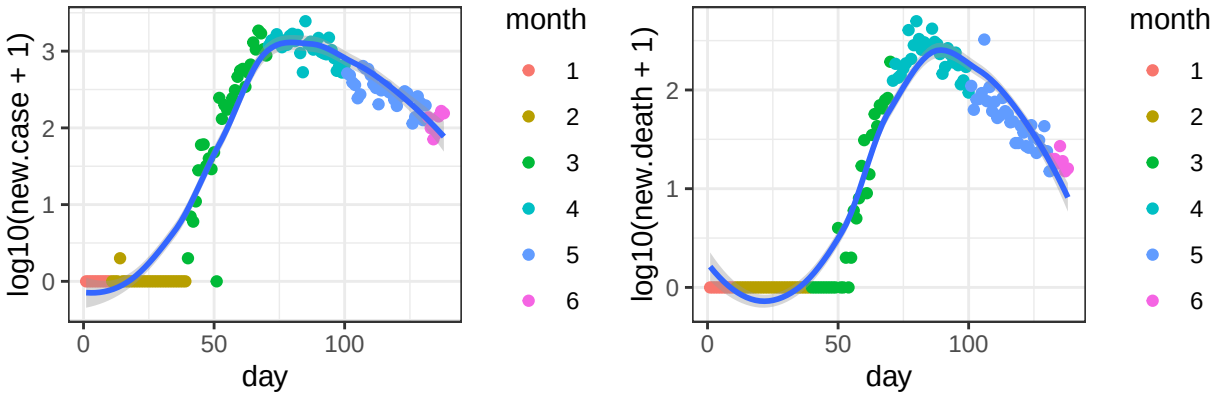
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Mexico

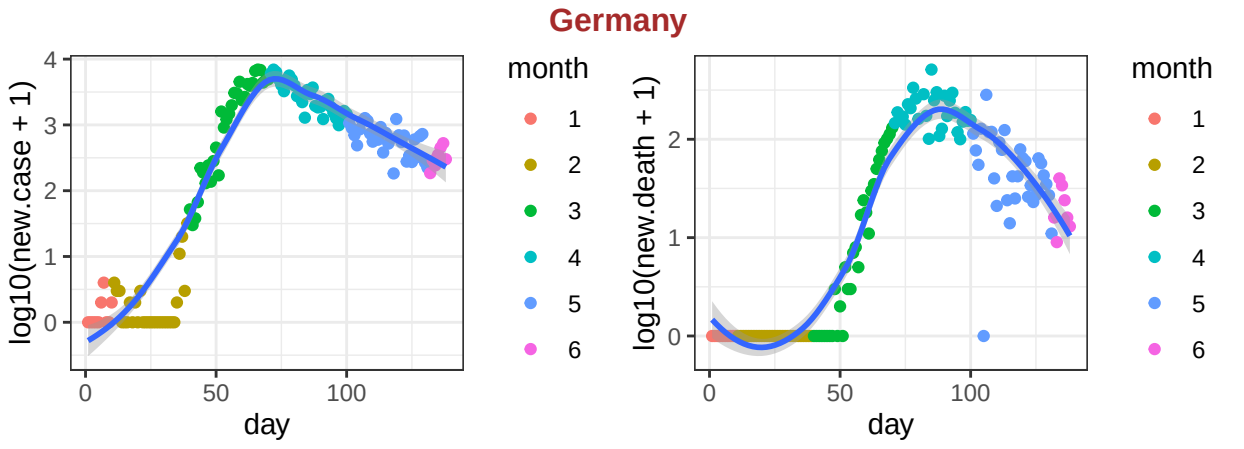


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

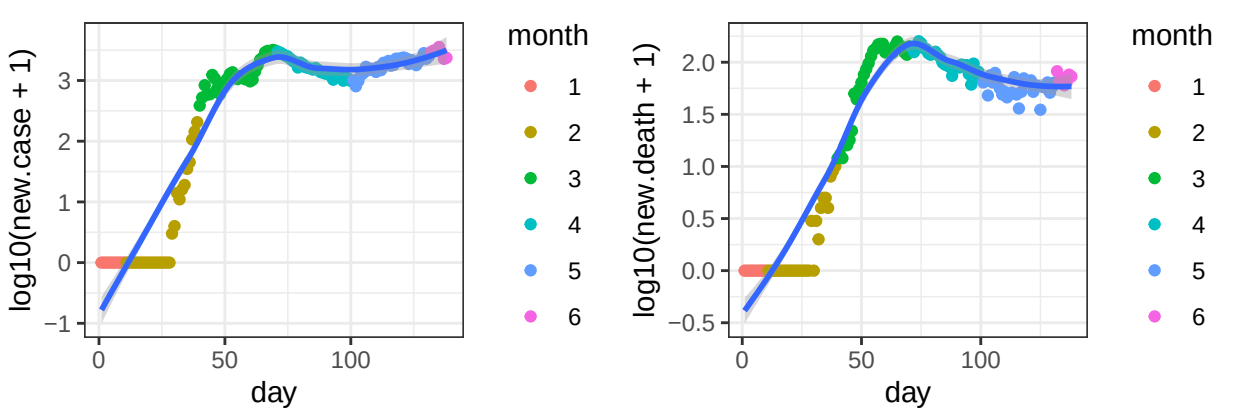
Belgium



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



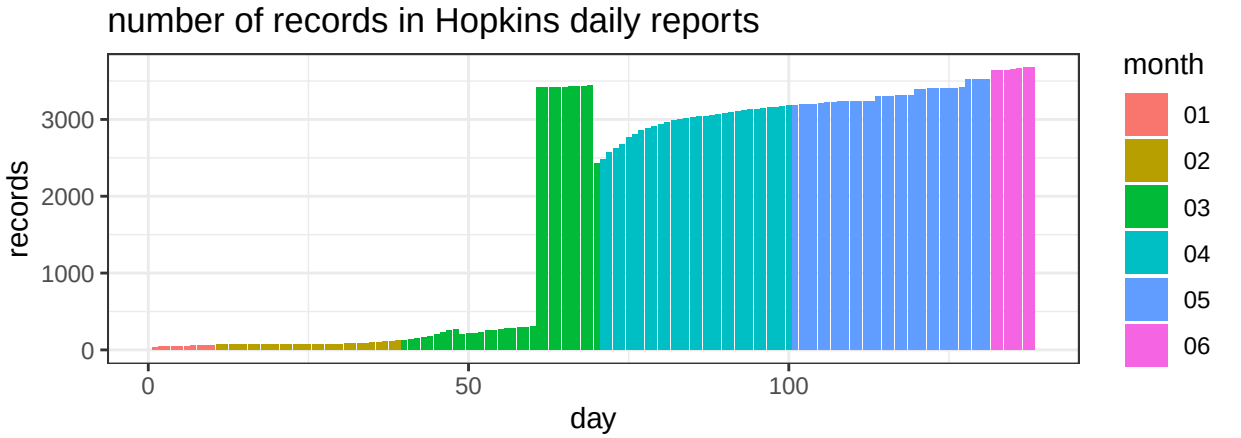
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



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

daily reports data

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



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

NY Times

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

The currente date is

```
## [1] "2020-06-06"
```

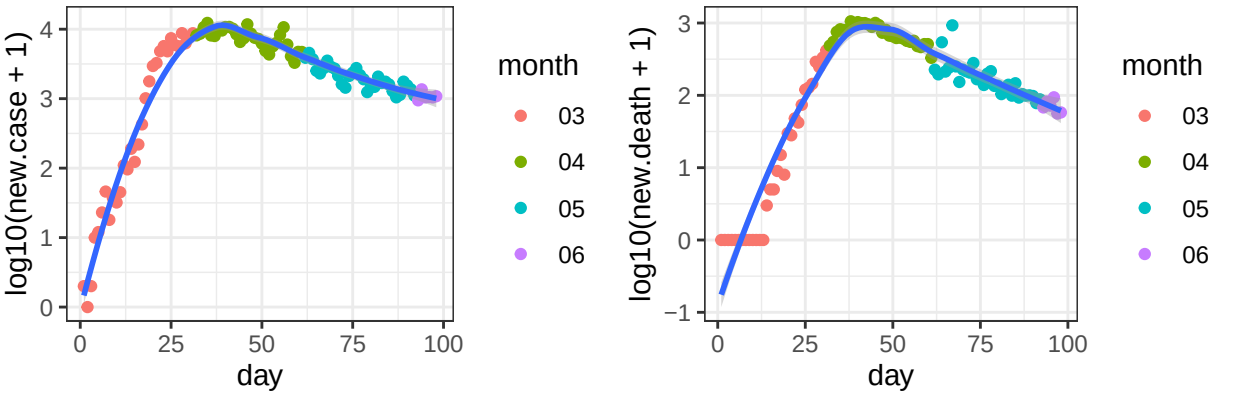
state level data

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

```
##            date                state fips  cases deaths
## 5273 2020-06-06             New York   36 382102  30123
## 5271 2020-06-06           New Jersey   34 163893  12106
## 5262 2020-06-06        Massachusetts   25 103132   7289
## 5280 2020-06-06         Pennsylvania   42  79507   5986
## 5254 2020-06-06             Illinois   17 127251   5898
## 5263 2020-06-06             Michigan   26  64196   5894
## 5244 2020-06-06           California    6 129147   4626
## 5246 2020-06-06          Connecticut    9  43818   4055
## 5259 2020-06-06            Louisiana   22  42597   2925
## 5261 2020-06-06             Maryland   24  58099   2740
## 5249 2020-06-06              Florida   12  62750   2687
## 5277 2020-06-06                 Ohio   39  38111   2370
## 5255 2020-06-06              Indiana   18  37928   2292
## 5250 2020-06-06              Georgia   13  48943   2147
## 5286 2020-06-06                Texas   48  75077   1840
## 5245 2020-06-06             Colorado    8  27834   1527
## 5290 2020-06-06             Virginia   51  49397   1460
## 5264 2020-06-06            Minnesota   27  27512   1181
## 5291 2020-06-06           Washington   53  24486   1163
## 5242 2020-06-06              Arizona    4  25517   1046
## 5274 2020-06-06       North Carolina   37  34809   1020
## 5266 2020-06-06             Missouri   29  14659    823
## 5265 2020-06-06          Mississippi   28  17034    811
## 5282 2020-06-06         Rhode Island   44  15441    772
## 5240 2020-06-06              Alabama    1  20043    689
## 5293 2020-06-06            Wisconsin   55  20701    646
## 5256 2020-06-06                 Iowa   19  21527    602
## 5283 2020-06-06       South Carolina   45  13916    545
## 5248 2020-06-06 District of Columbia   11   9269    483
## 5258 2020-06-06             Kentucky   21  11359    480
```

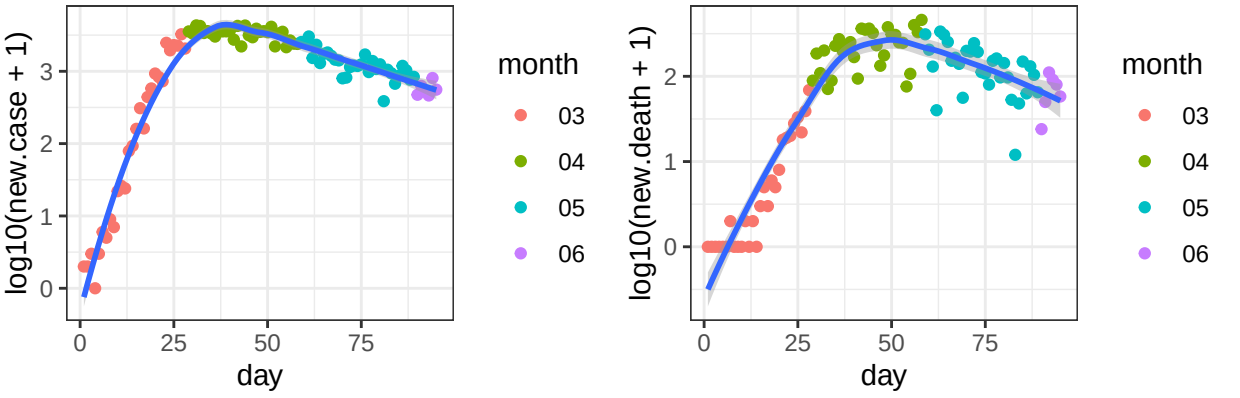
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



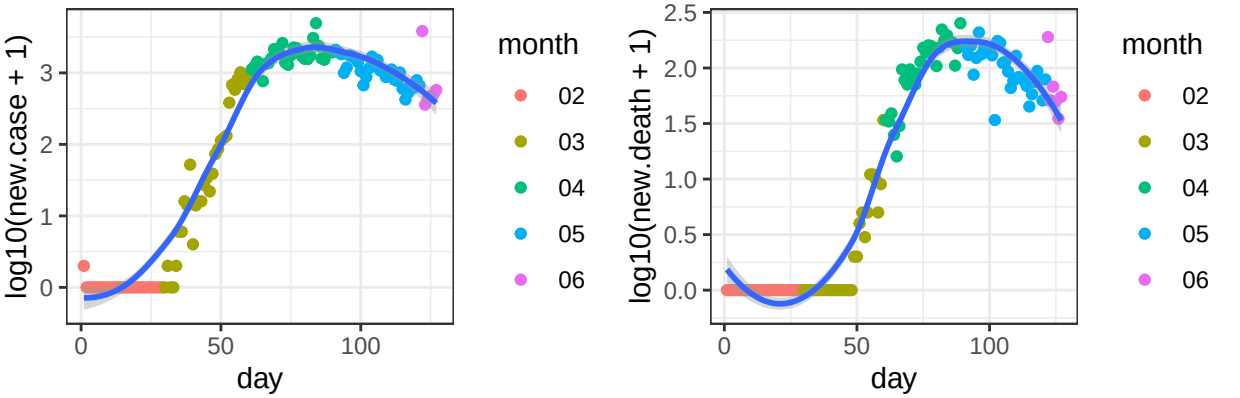
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



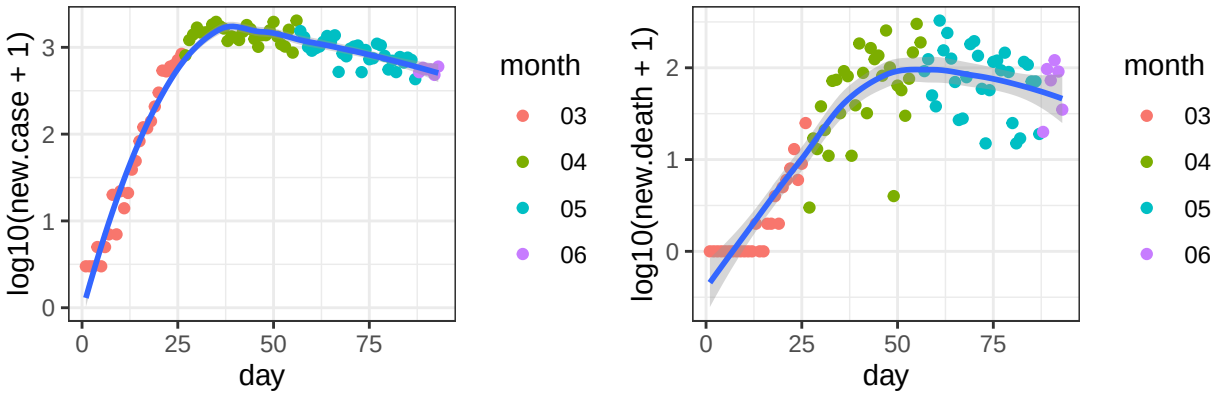
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Massachusetts



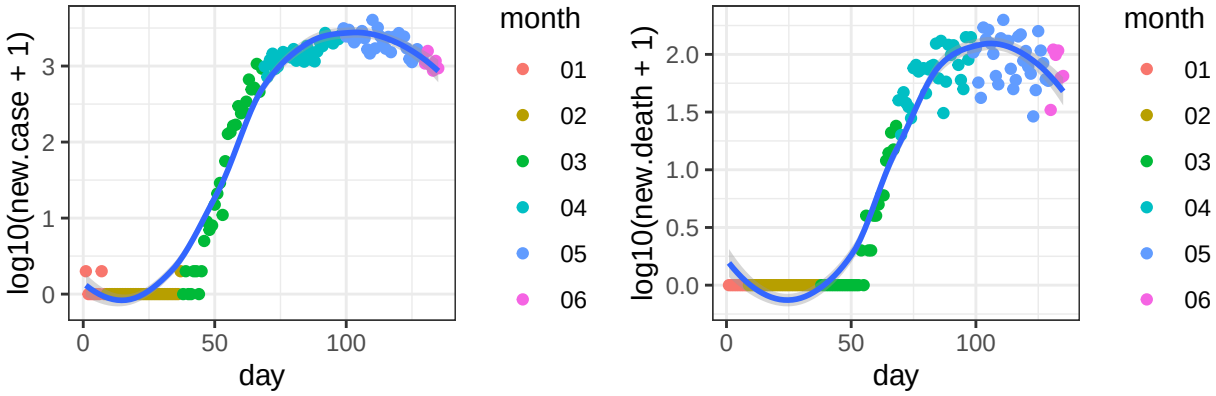
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Pennsylvania



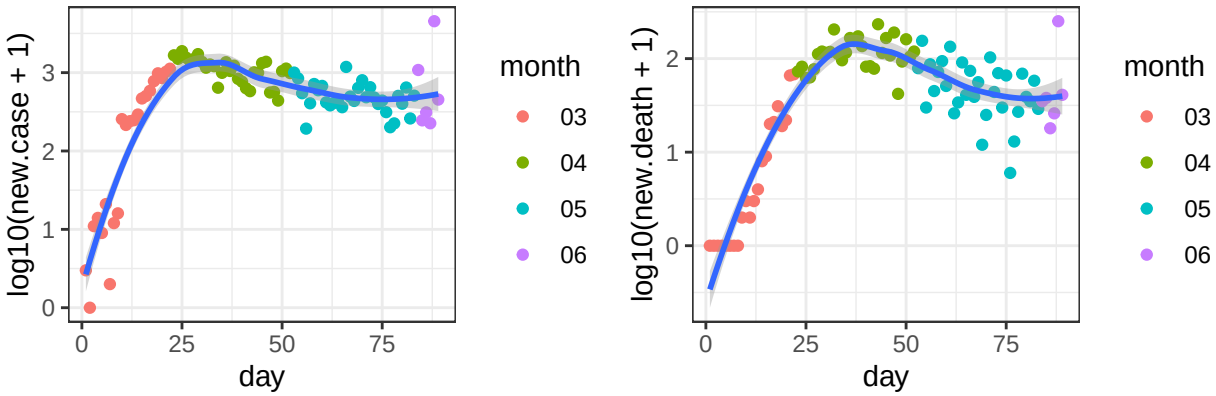
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Illinois



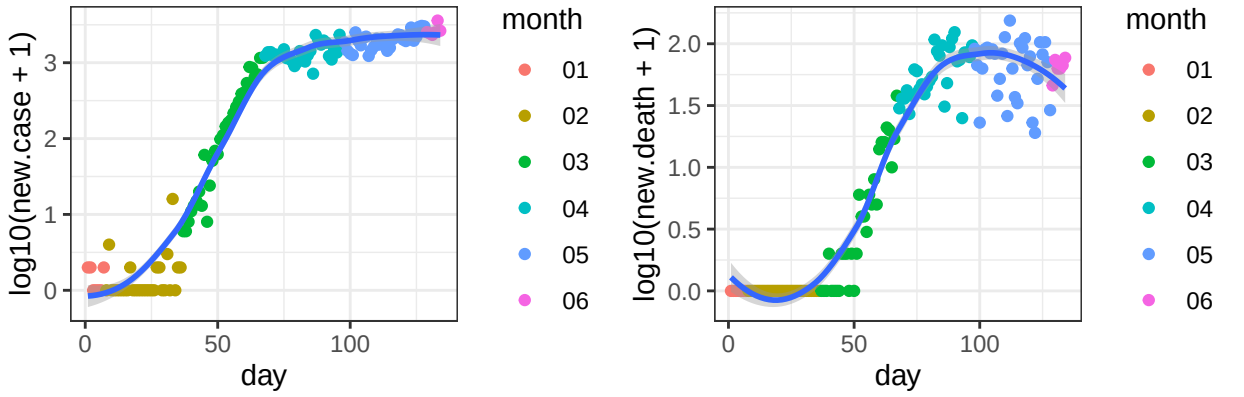
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Michigan



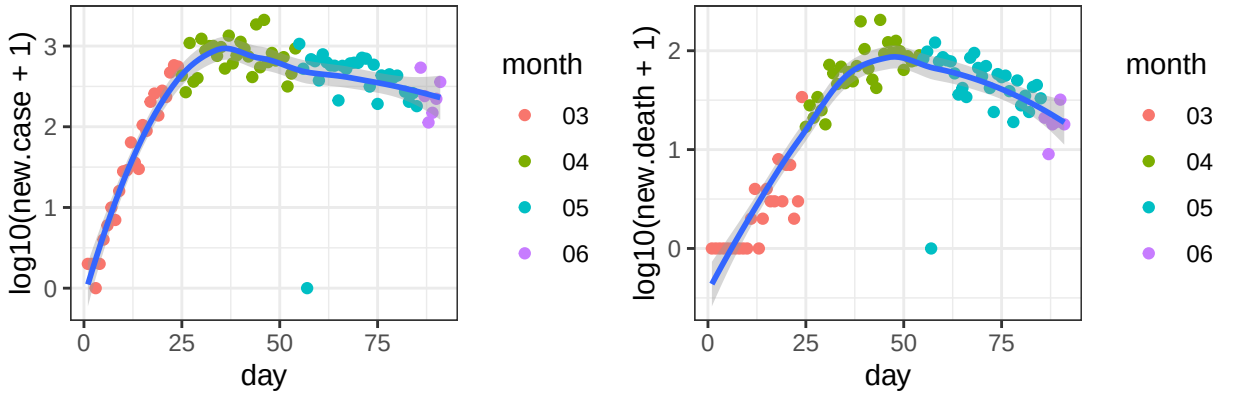
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

California



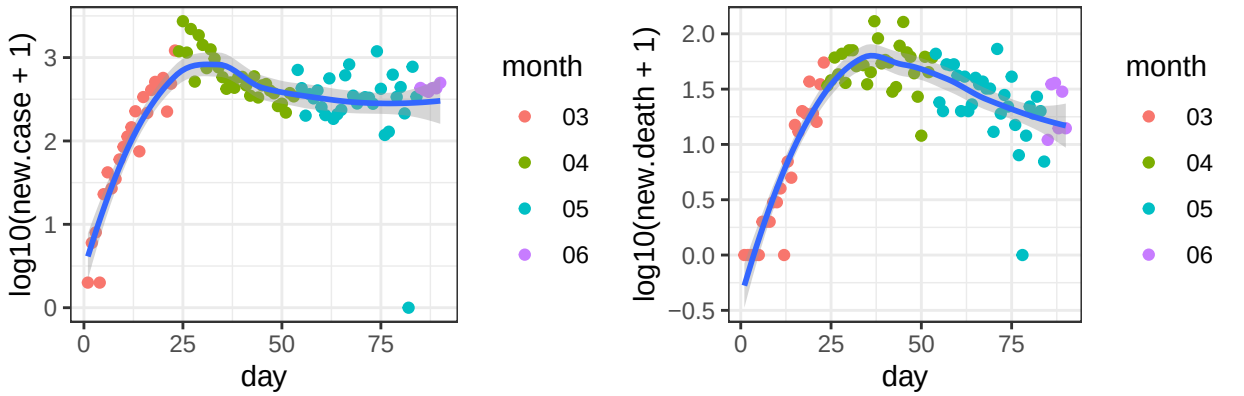
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Connecticut



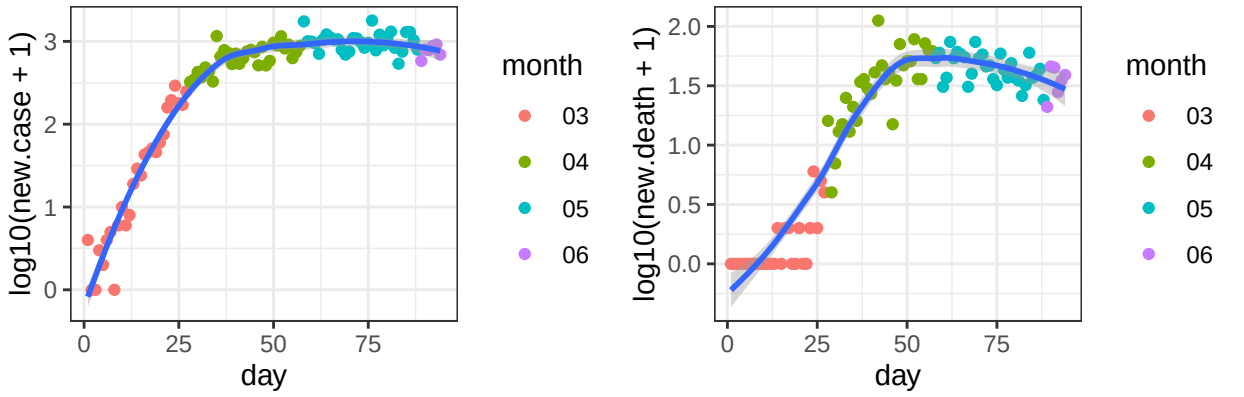
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Louisiana



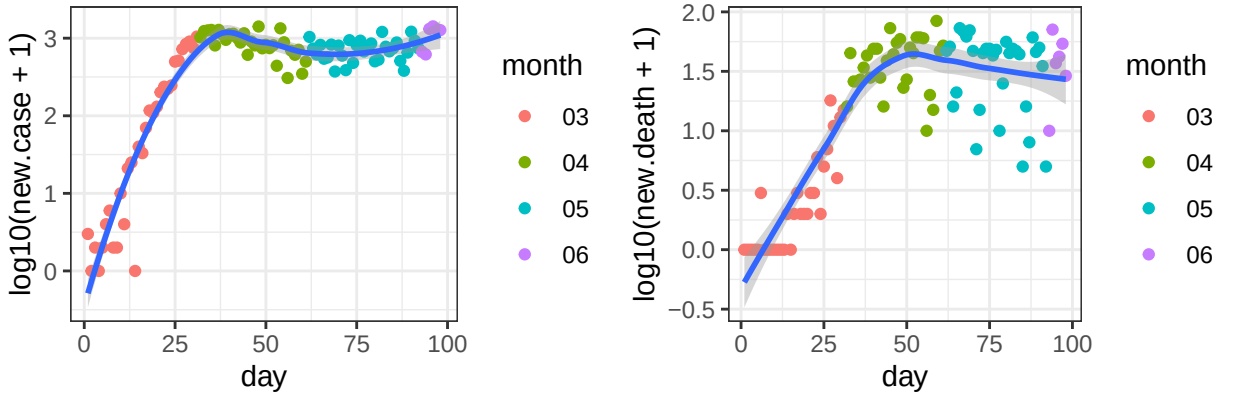
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



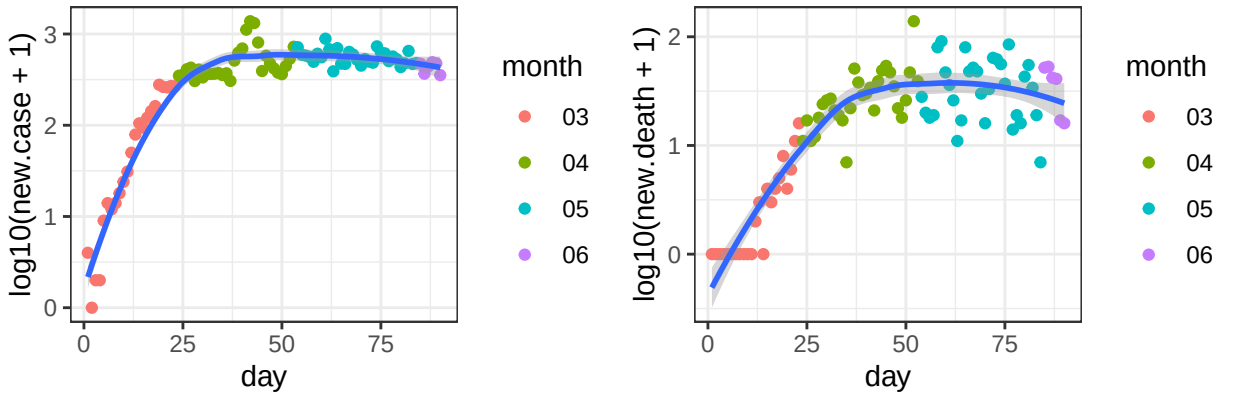
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Florida



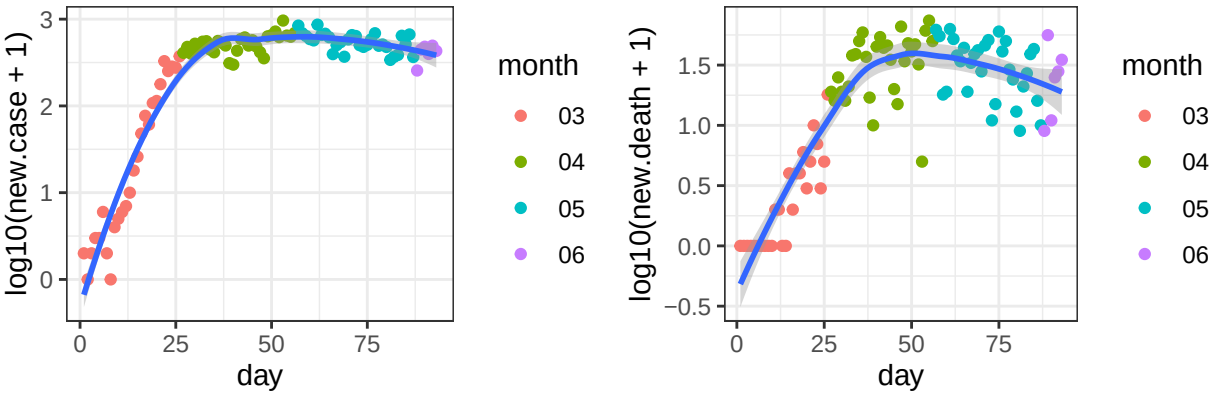
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Ohio



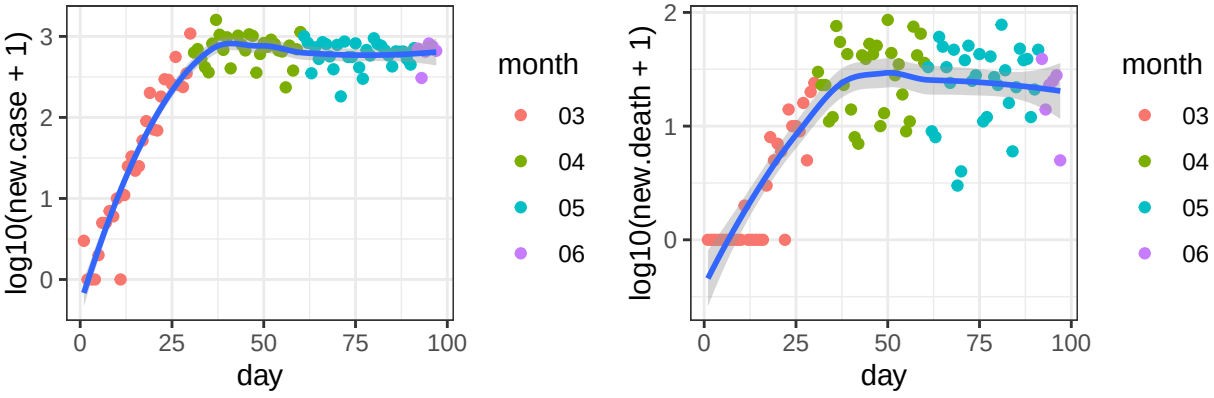
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Indiana



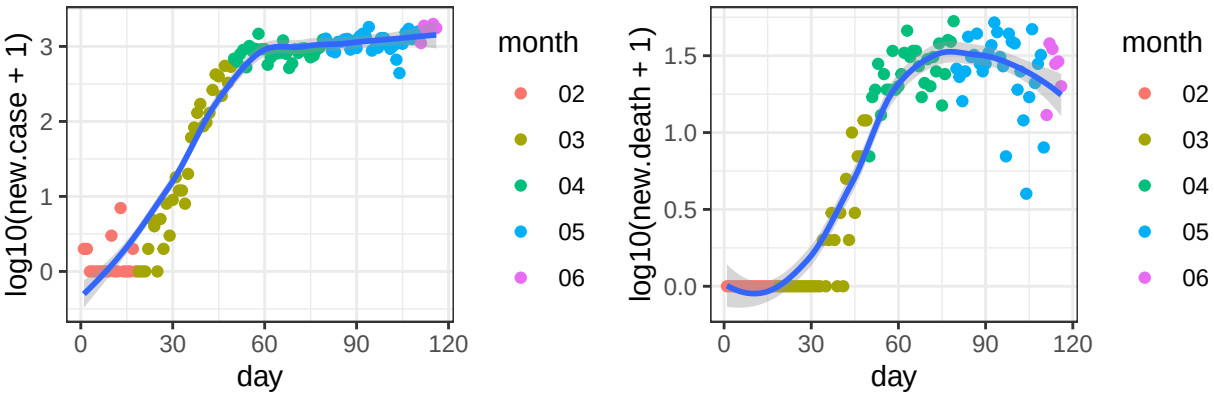
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Georgia



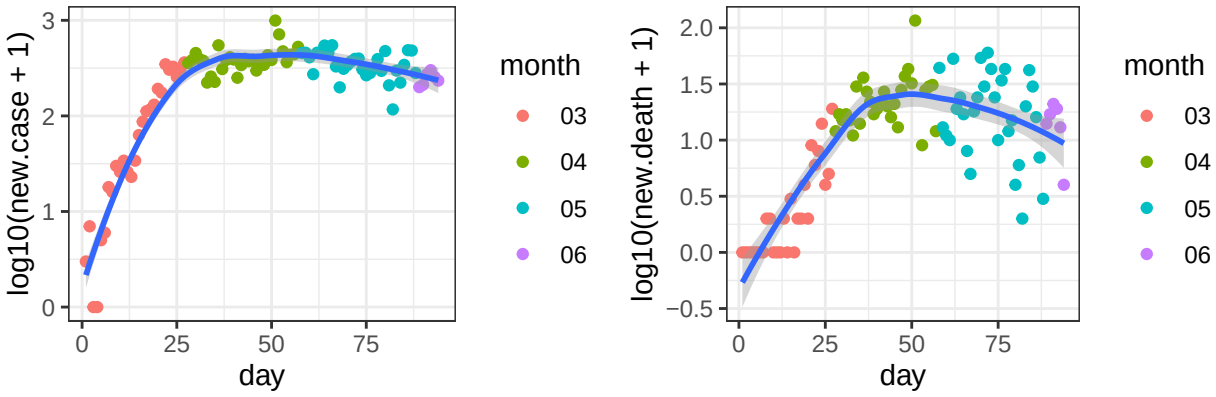
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Texas



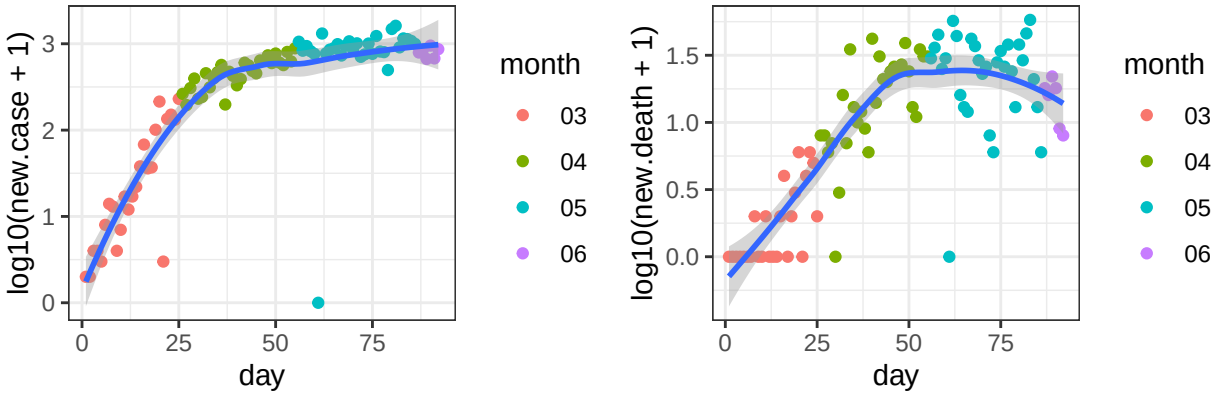
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Colorado



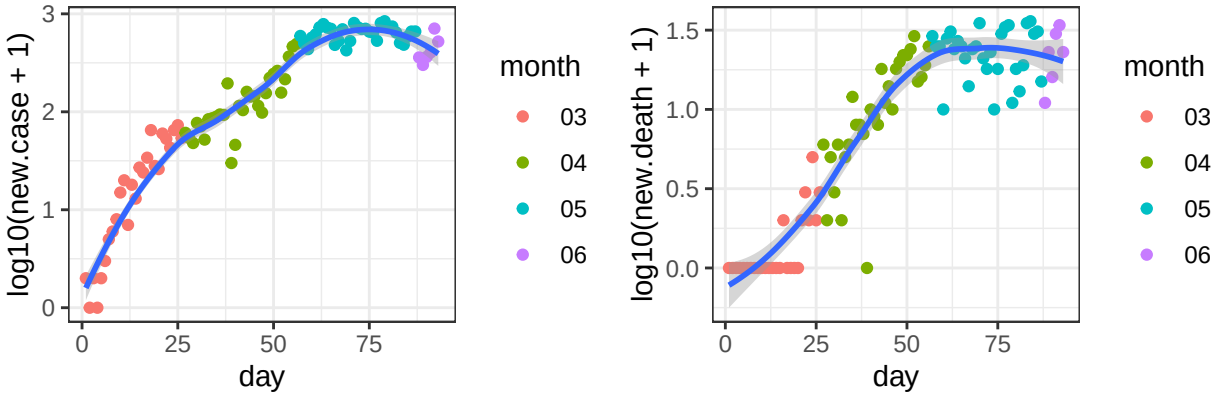
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Virginia



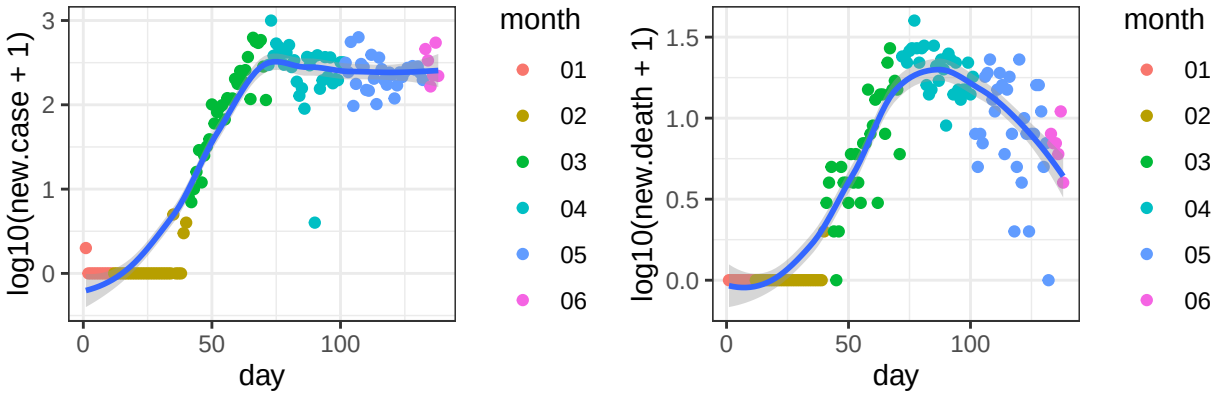
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Minnesota



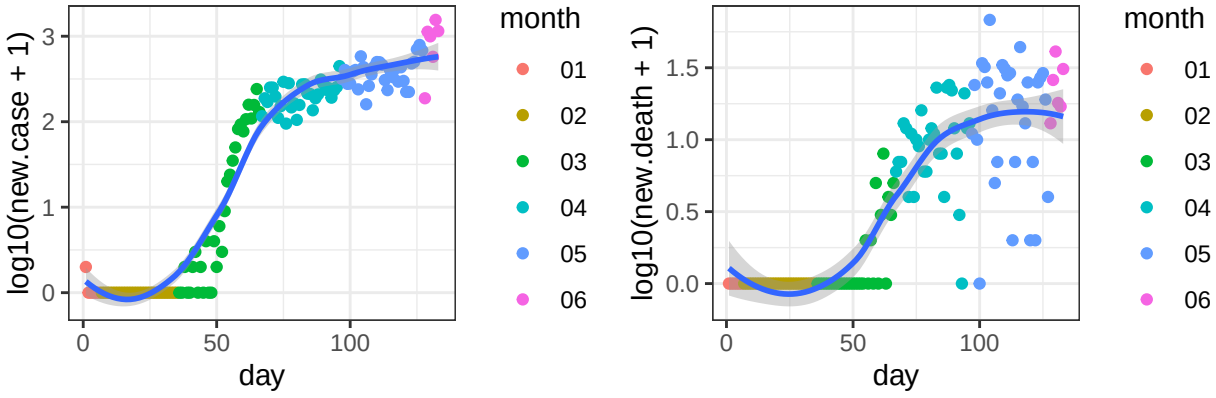
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Washington



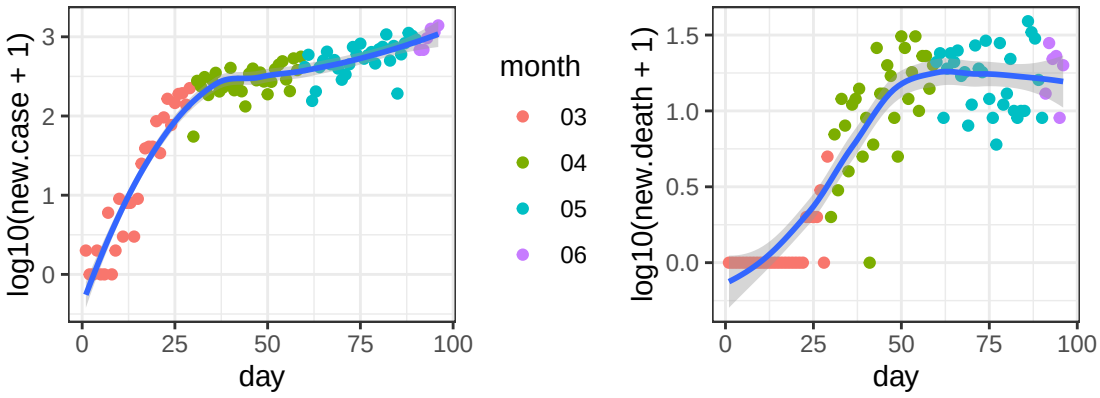
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Arizona



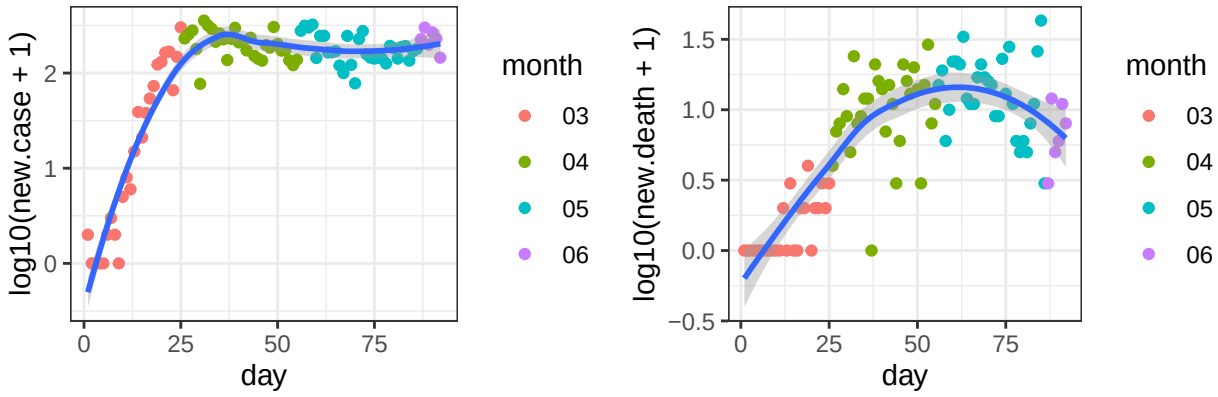
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

North Carolina



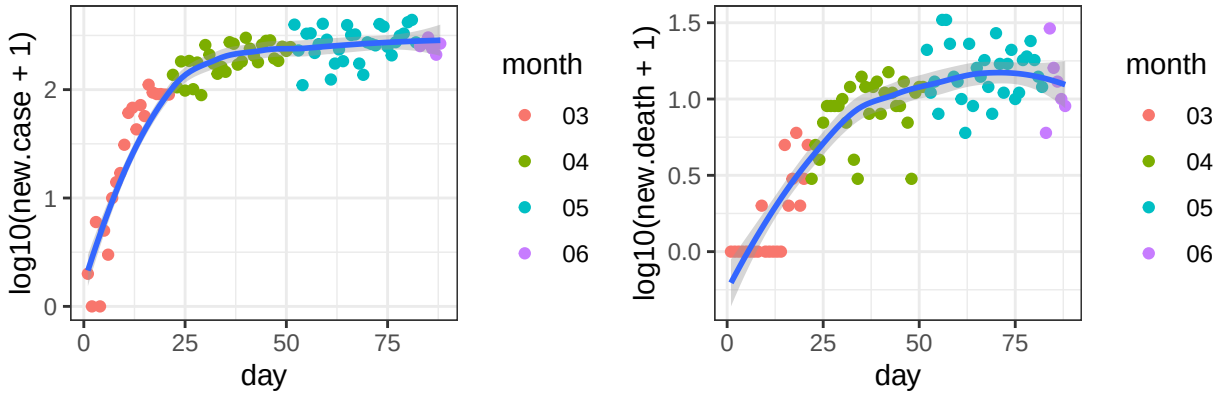
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Missouri



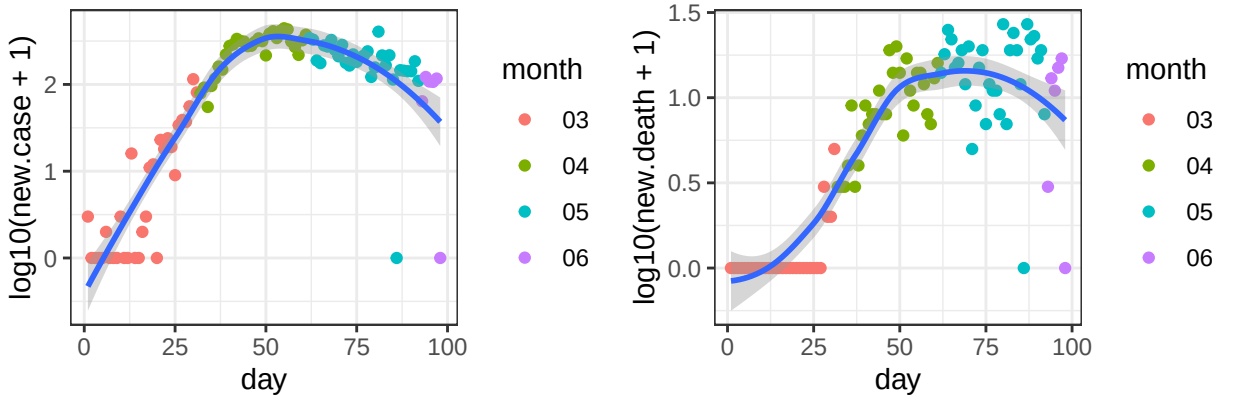
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Mississippi



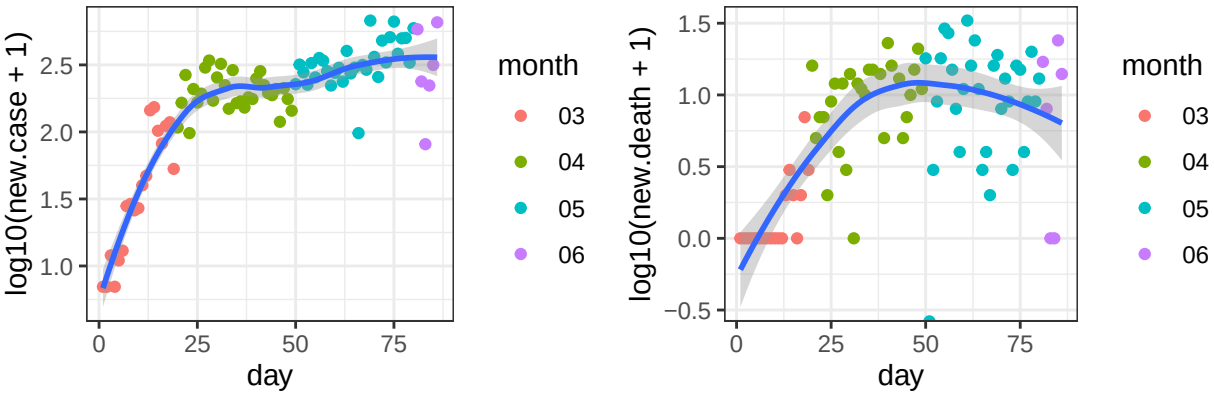
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Rhode Island



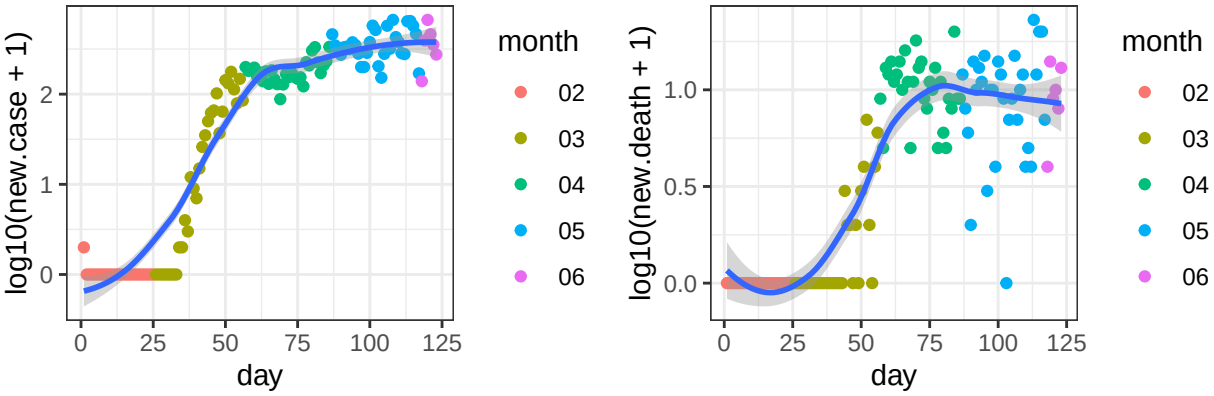
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Alabama



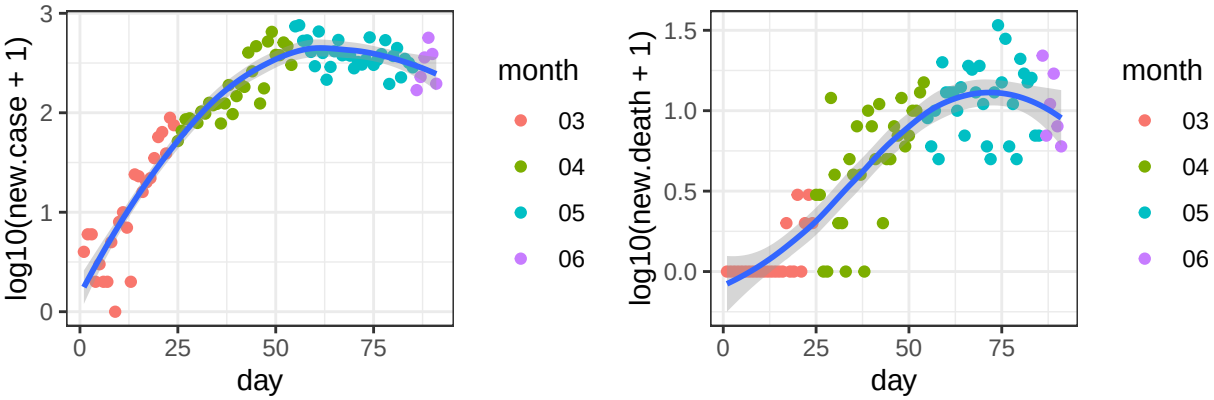
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Wisconsin



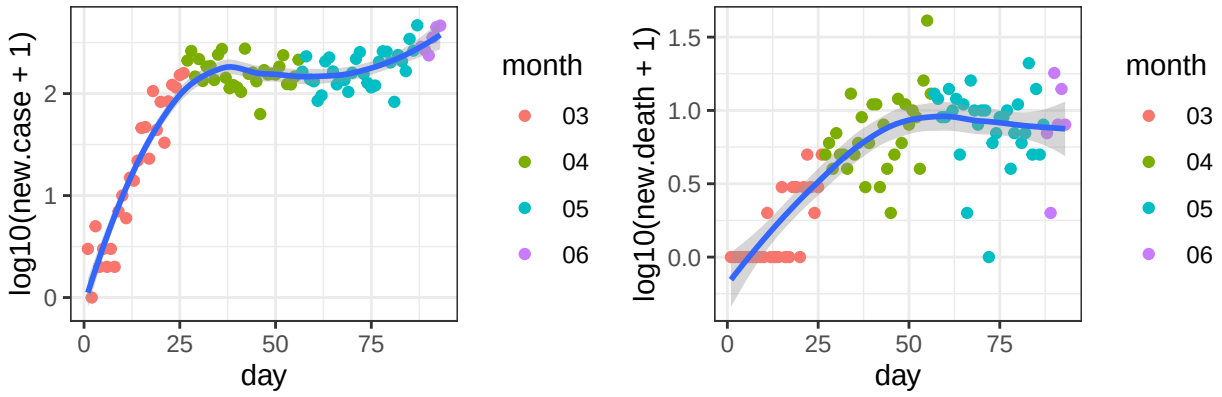
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

Iowa



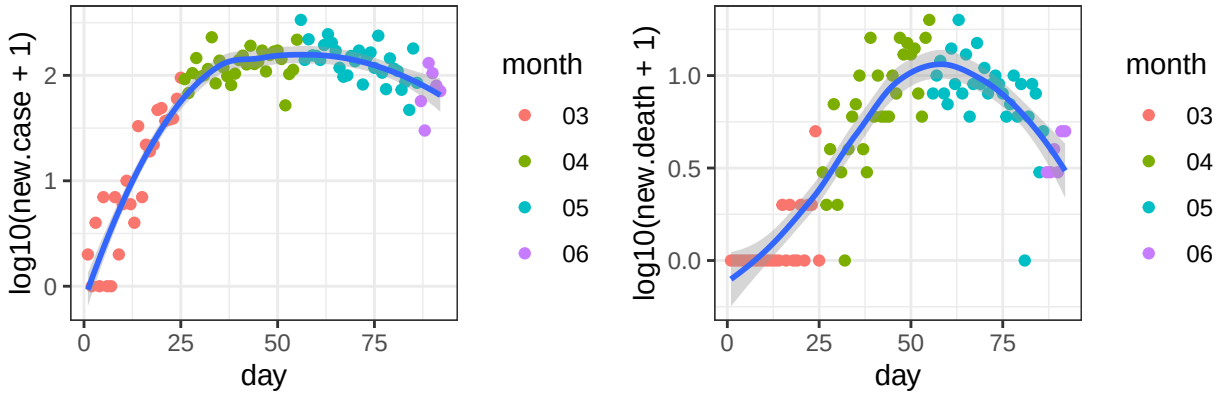
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

South Carolina



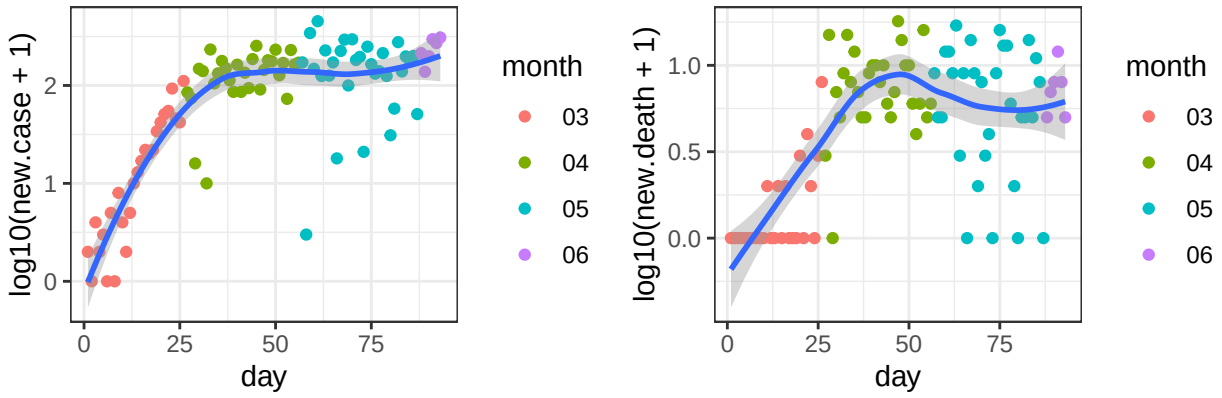
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

District of Columbia



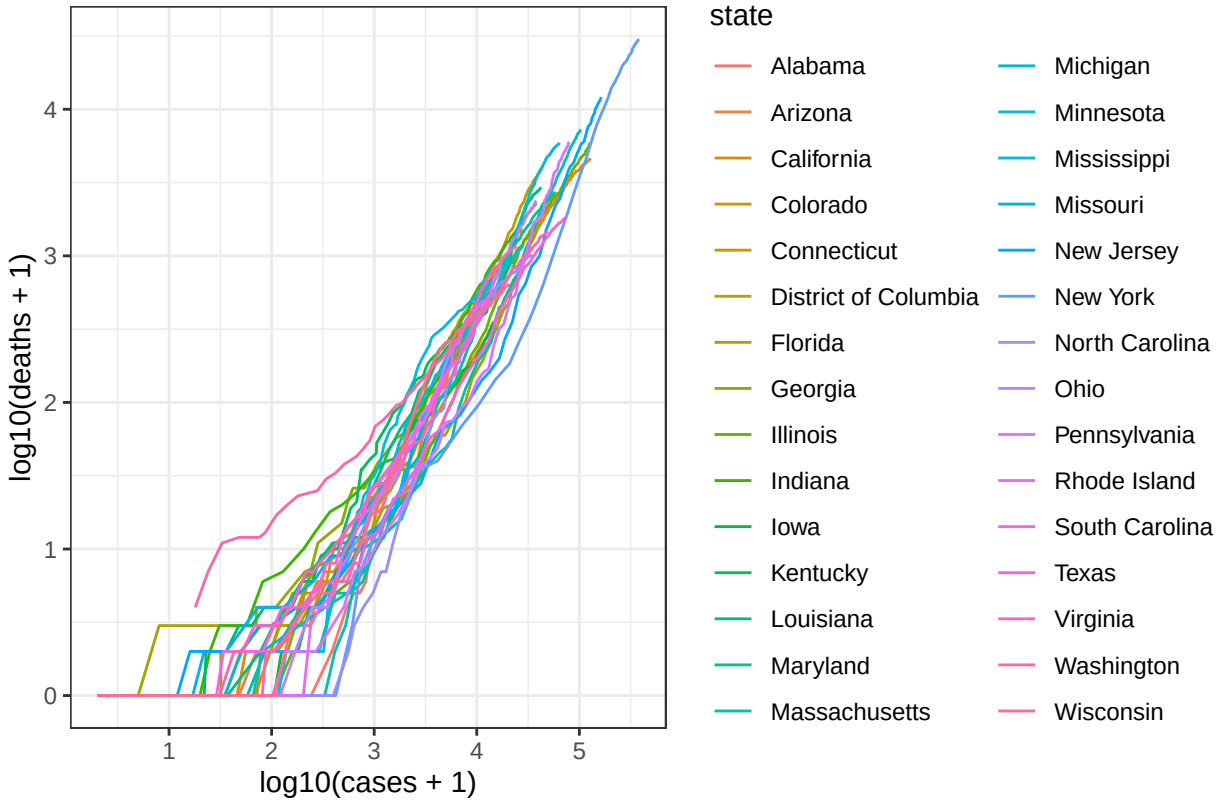
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

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

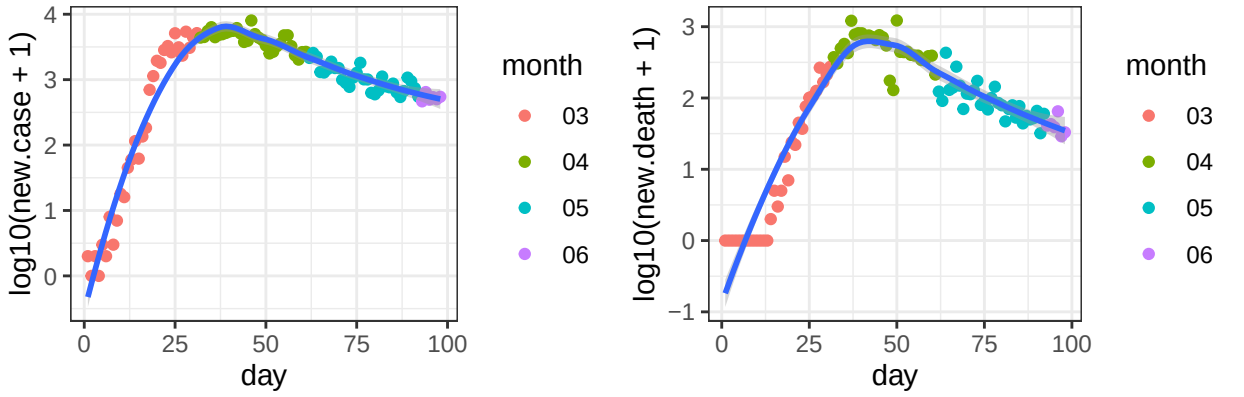
First check the 50 counties with the largest number of deaths.

```
##             date              county               state  fips  cases deaths
## 211368 2020-06-06       New York City            New York    NA 211274  21294
## 210192 2020-06-06                Cook            Illinois 17031  81924   3913
## 211367 2020-06-06              Nassau            New York 36059  40853   2635
## 210878 2020-06-06               Wayne            Michigan 26163  21163   2627
## 209796 2020-06-06         Los Angeles          California  6037  62338   2620
## 211387 2020-06-06             Suffolk            New York 36103  40278   1970
## 210792 2020-06-06           Middlesex       Massachusetts 25017  22686   1701
## 211293 2020-06-06               Essex          New Jersey 34013  18066   1701
## 211288 2020-06-06              Bergen          New Jersey 34003  18492   1612
## 211395 2020-06-06          Westchester            New York 36119  33923   1523
## 211791 2020-06-06        Philadelphia        Pennsylvania 42101  23529   1414
## 209895 2020-06-06           Fairfield         Connecticut  9001  16020   1309
## 209896 2020-06-06            Hartford         Connecticut  9003  10747   1279
## 211295 2020-06-06              Hudson          New Jersey 34017  18548   1210
## 211306 2020-06-06               Union          New Jersey 34039  16116   1095
## 210859 2020-06-06             Oakland            Michigan 26125  10980   1055
## 211298 2020-06-06           Middlesex          New Jersey 34023  16203   1032
## 209899 2020-06-06           New Haven         Connecticut  9009  11817   1007
## 210788 2020-06-06               Essex       Massachusetts 25009  15170    998
## 211302 2020-06-06             Passaic          New Jersey 34031  16436    969
## 210796 2020-06-06             Suffolk       Massachusetts 25025  18955    923
## 210846 2020-06-06              Macomb            Michigan 26099   6940    870
```

```
## 210794 2020-06-06              Norfolk        Massachusetts 25021   8689    859
## 210798 2020-06-06            Worcester        Massachusetts 25027  11696    820
## 211301 2020-06-06                Ocean           New Jersey 34029   8979    767
## 209951 2020-06-06           Miami-Dade              Florida 12086  19298    765
## 211786 2020-06-06           Montgomery         Pennsylvania 42091   7542    724
## 210905 2020-06-06             Hennepin            Minnesota 27053   9255    667
## 210326 2020-06-06               Marion              Indiana 18097  10390    663
## 210774 2020-06-06           Montgomery             Maryland 24031  12662    652
## 211763 2020-06-06             Delaware         Pennsylvania 42045   6661    651
## 211299 2020-06-06             Monmouth           New Jersey 34025   8454    636
## 211300 2020-06-06               Morris           New Jersey 34027   6584    626
## 210790 2020-06-06              Hampden        Massachusetts 25013   6337    618
## 210775 2020-06-06      Prince George's             Maryland 24033  16838    595
## 210795 2020-06-06             Plymouth        Massachusetts 25023   8347    588
## 212436 2020-06-06                 King           Washington 53033   8419    578
## 211353 2020-06-06                 Erie             New York 36029   6429    547
## 211749 2020-06-06                Bucks         Pennsylvania 42017   5243    529
## 211812 2020-06-06            Providence         Rhode Island 44007  11052    518
## 210713 2020-06-06              Orleans            Louisiana 22071   7222    512
## 211297 2020-06-06               Mercer           New Jersey 34021   7148    500
## 209695 2020-06-06              Maricopa              Arizona  4013  12761    489
## 209908 2020-06-06 District of Columbia District of Columbia 11001   9269    483
## 210786 2020-06-06              Bristol        Massachusetts 25005   7635    467
## 211379 2020-06-06             Rockland             New York 36087  13315    465
## 211142 2020-06-06            St. Louis             Missouri 29189   5029    460
## 210703 2020-06-06            Jefferson            Louisiana 22051   7831    458
## 211304 2020-06-06              Somerset           New Jersey 34035   4664    425
## 212325 2020-06-06              Fairfax             Virginia 51059  12056    413
```

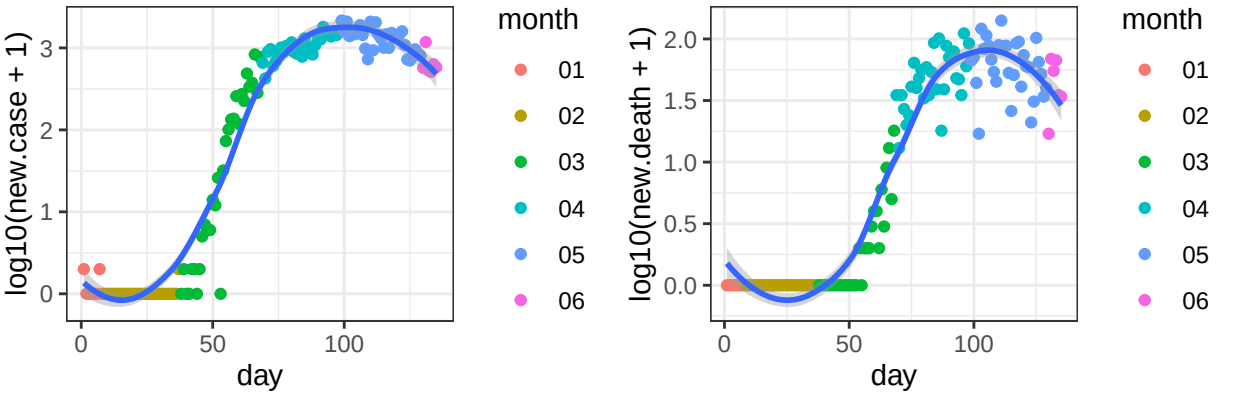
For these 50 counties, I check the number of new cases and the number of new deaths.

New York City_New York



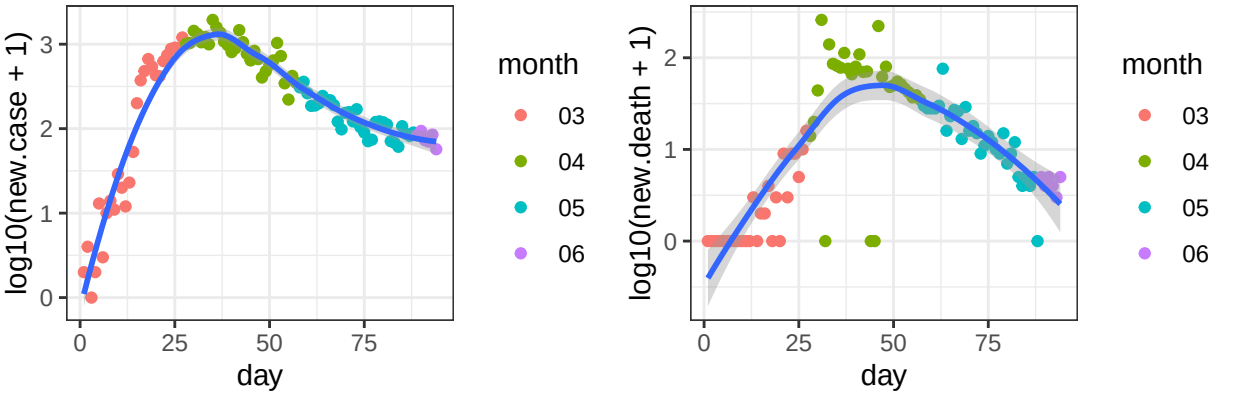
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



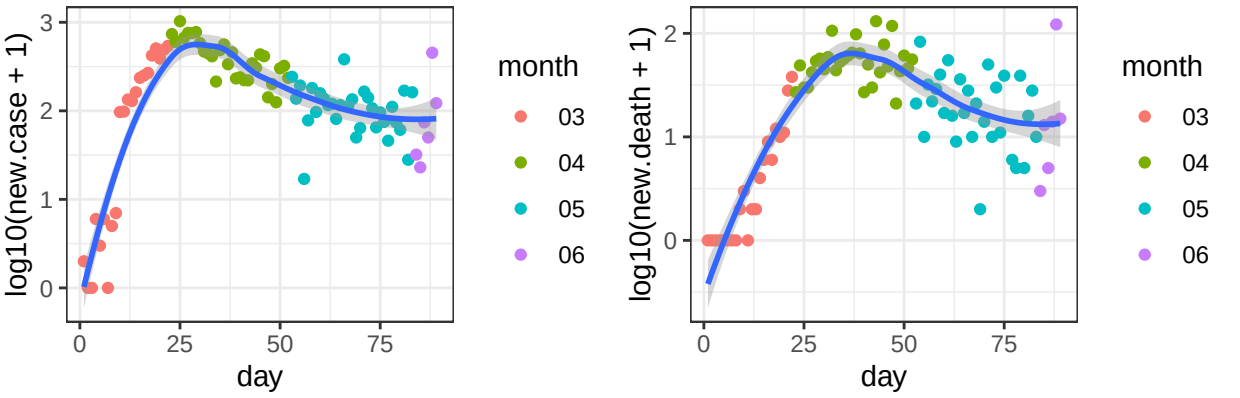
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Nassau_New York



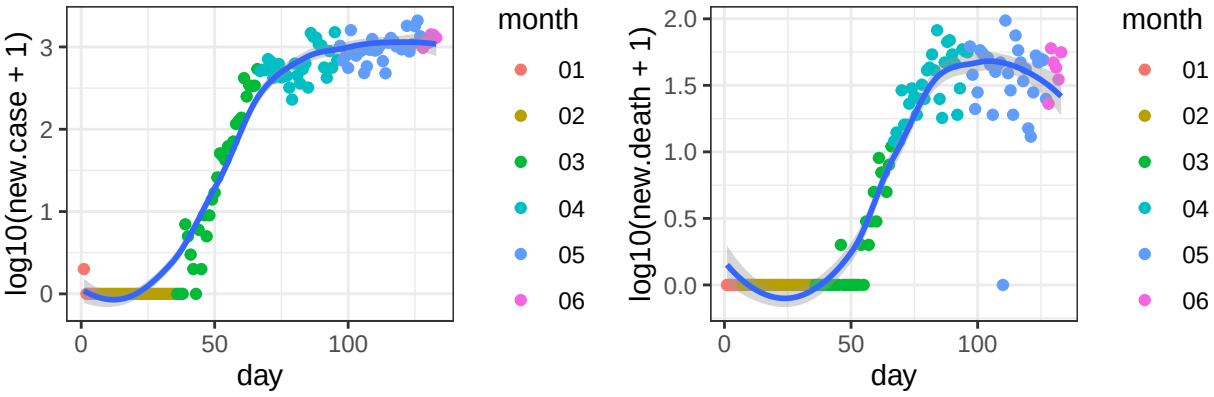
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Wayne_Michigan



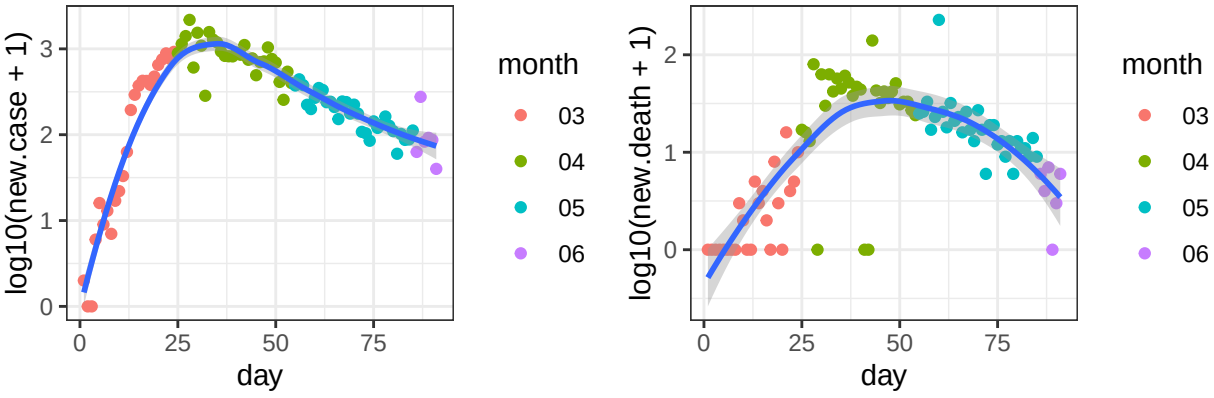
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Los Angeles_California



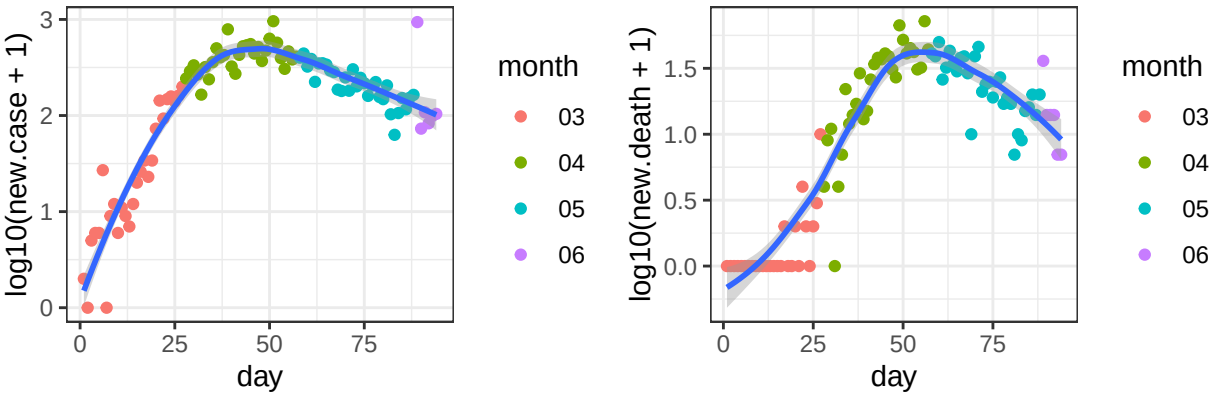
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Suffolk_New York



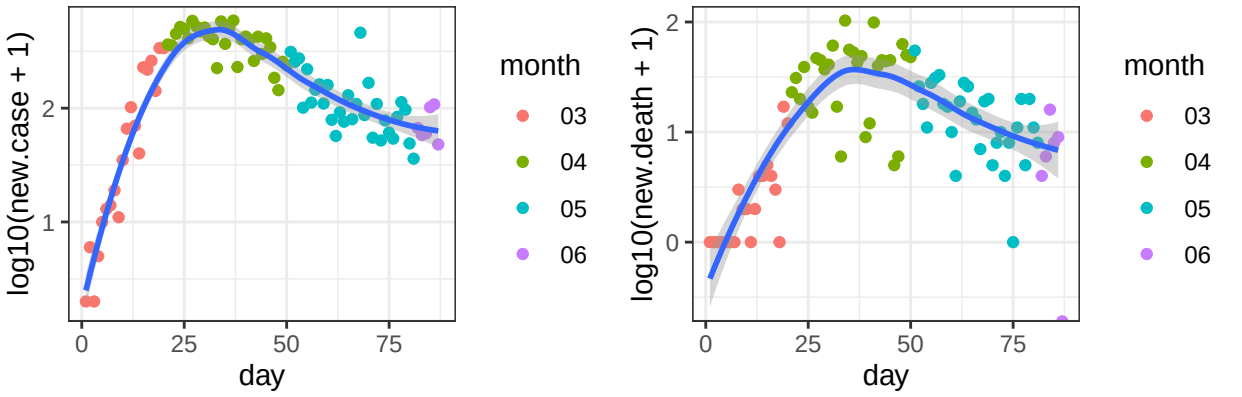
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Middlesex_Massachusetts



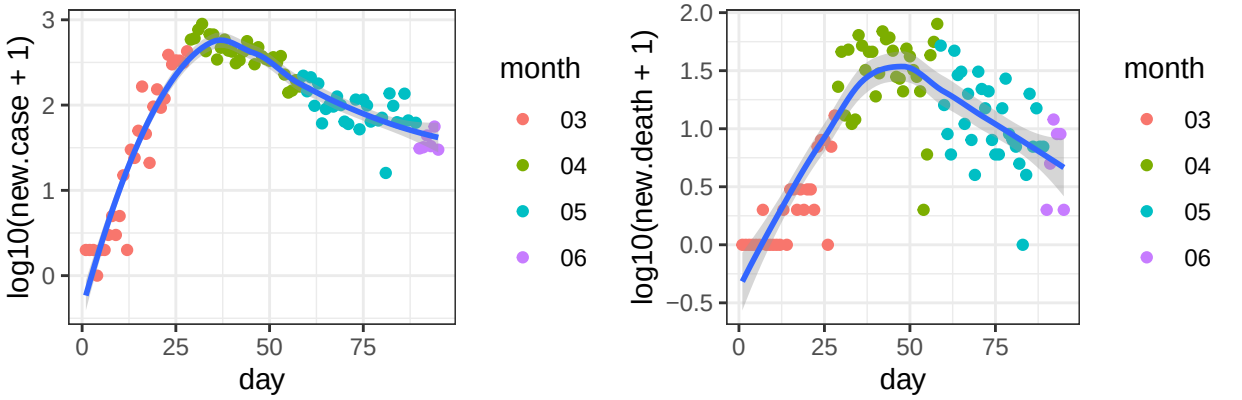
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



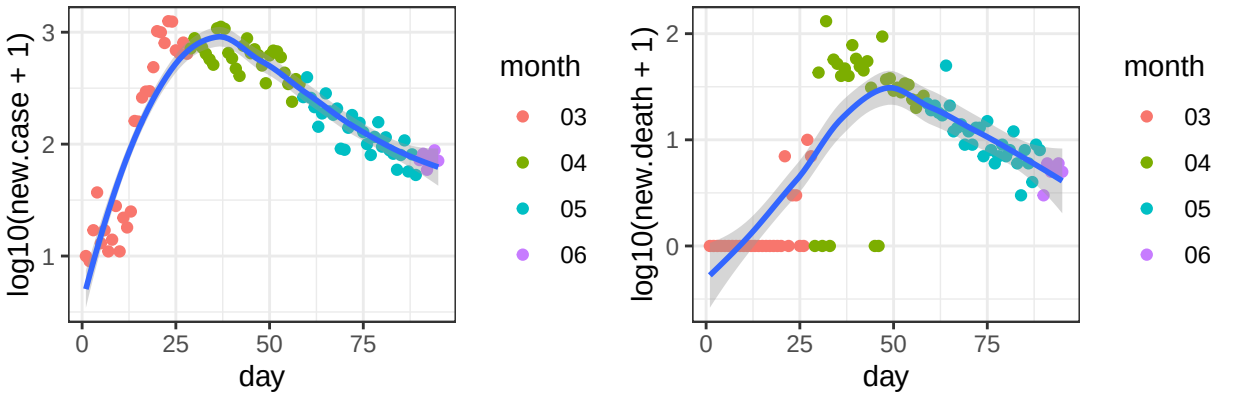
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Bergen_New Jersey



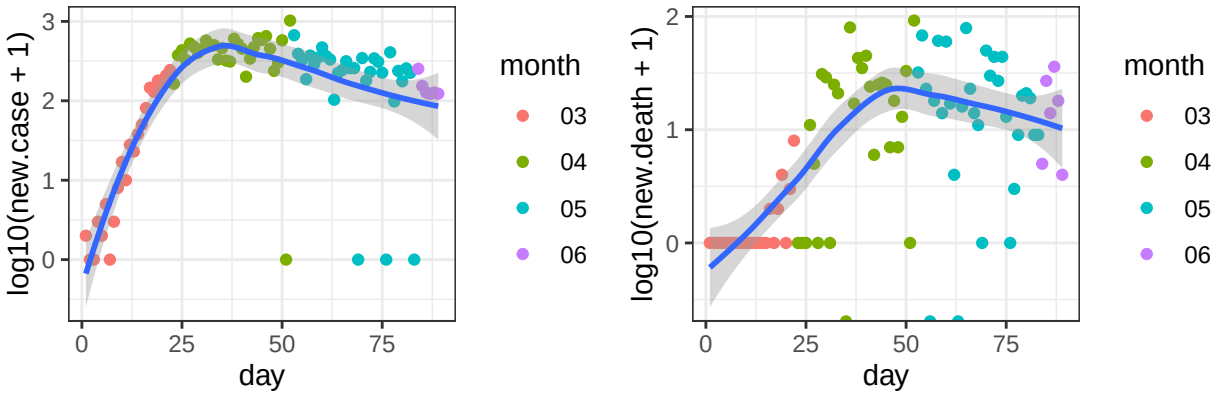
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Westchester_New York



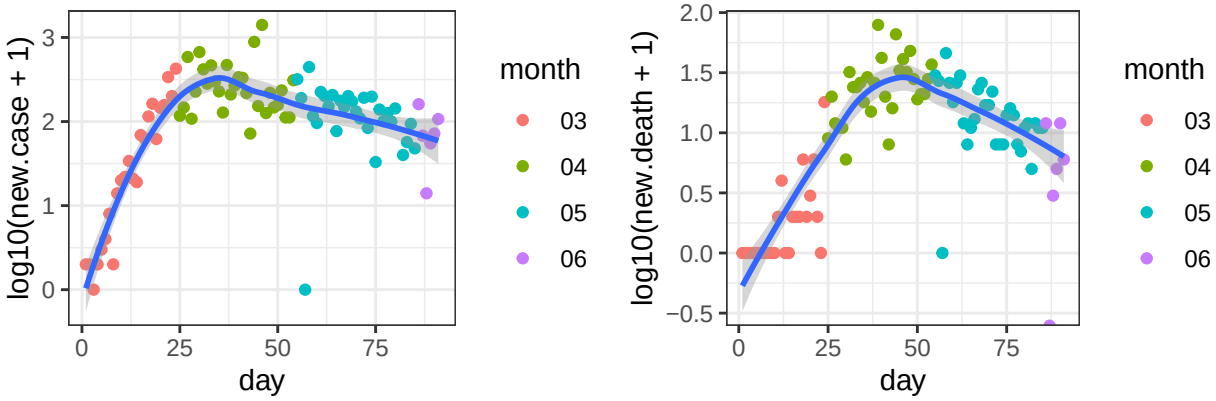
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Philadelphia_Pennsylvania



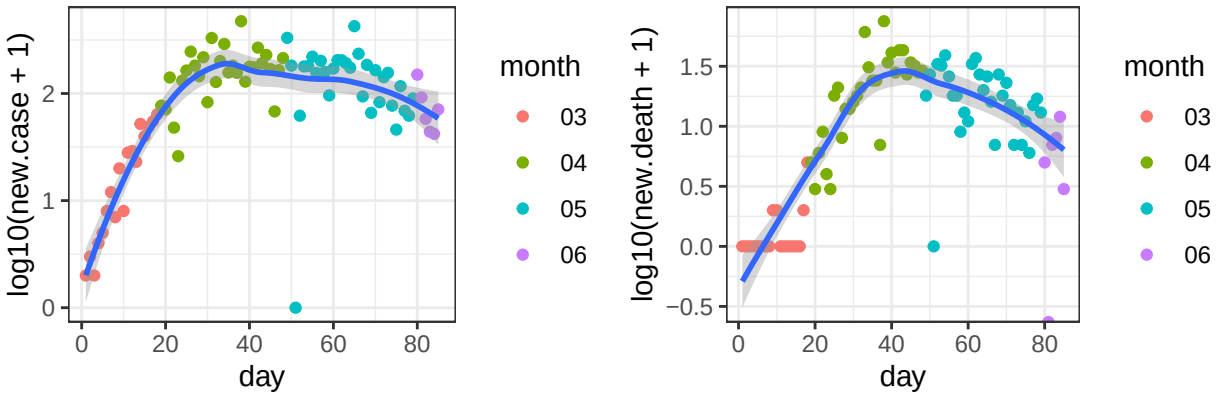
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Fairfield_Connecticut



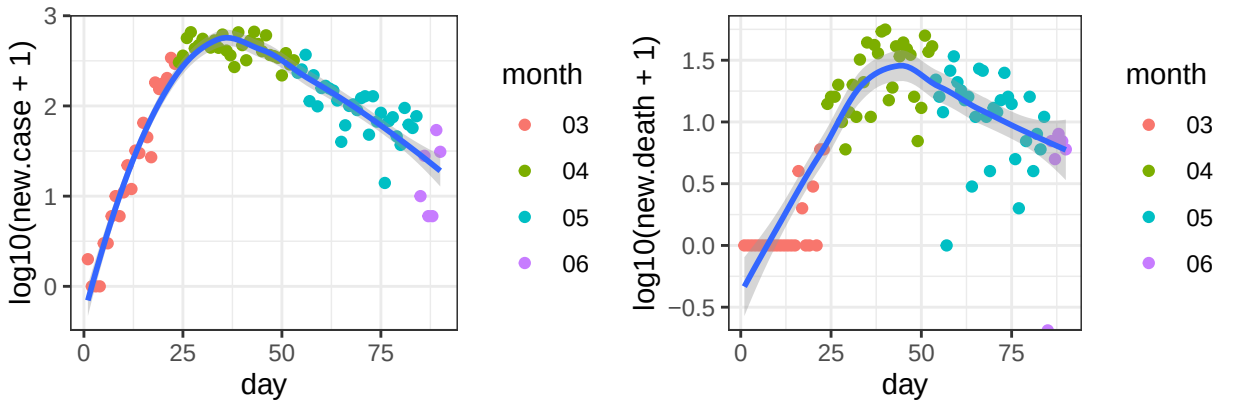
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Hartford_Connecticut



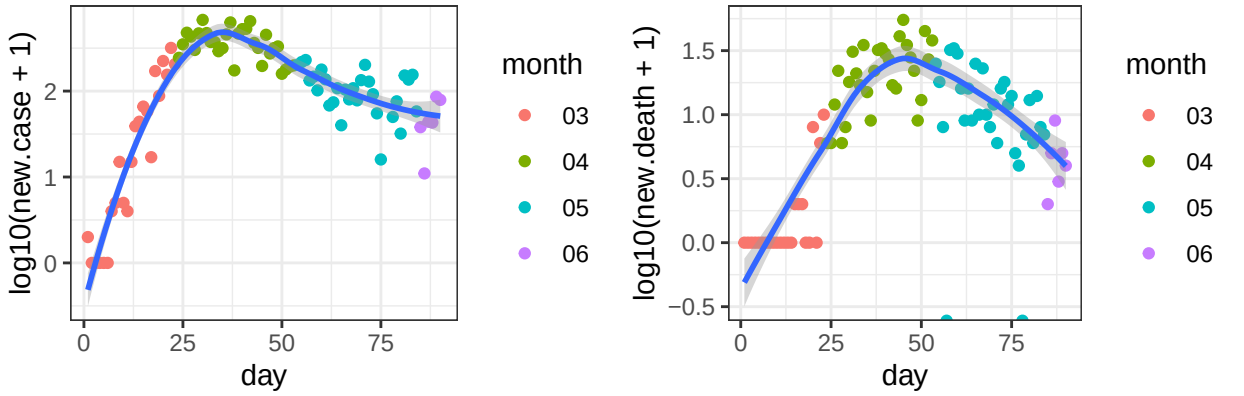
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Hudson_New Jersey



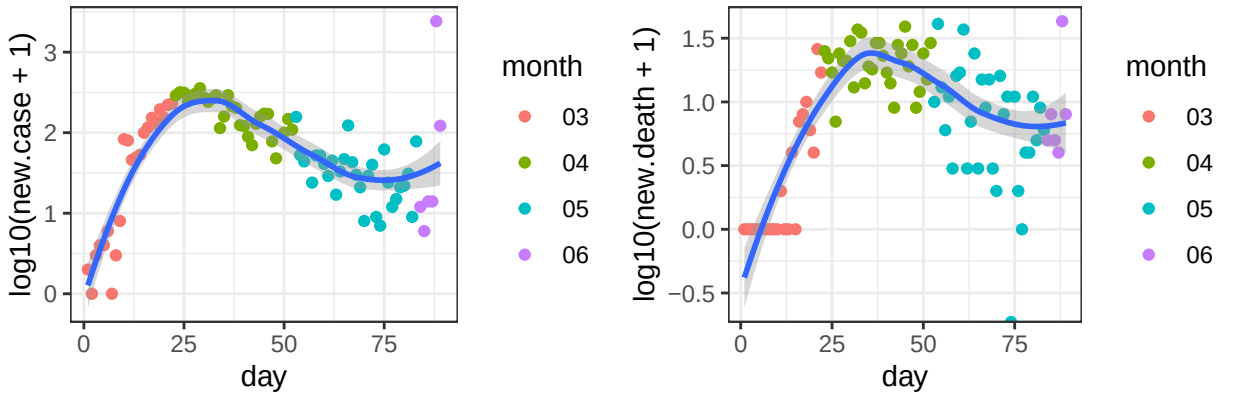
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Union_New Jersey



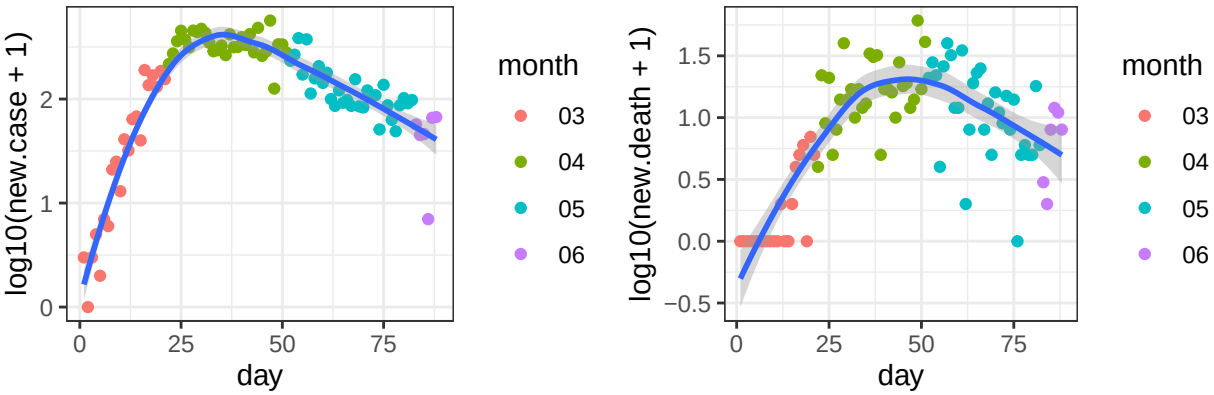
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Oakland_Michigan



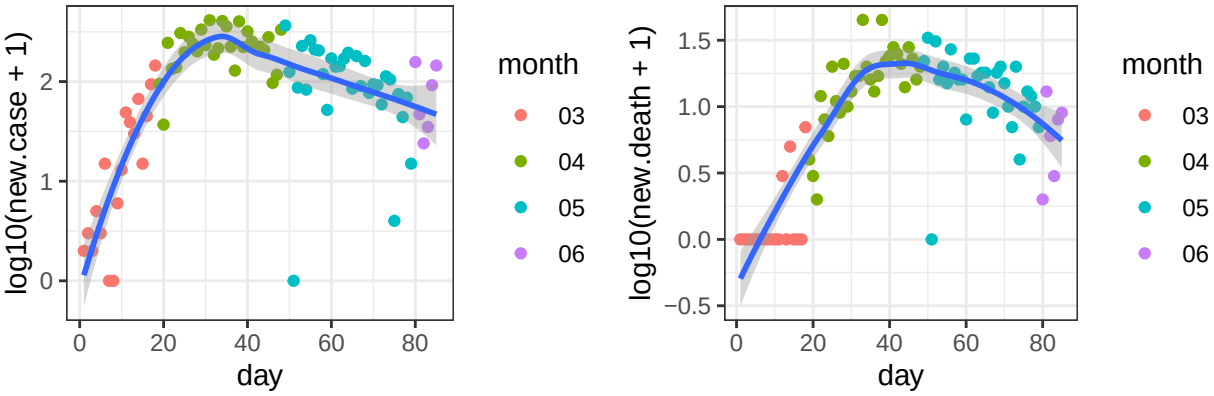
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Middlesex_New Jersey



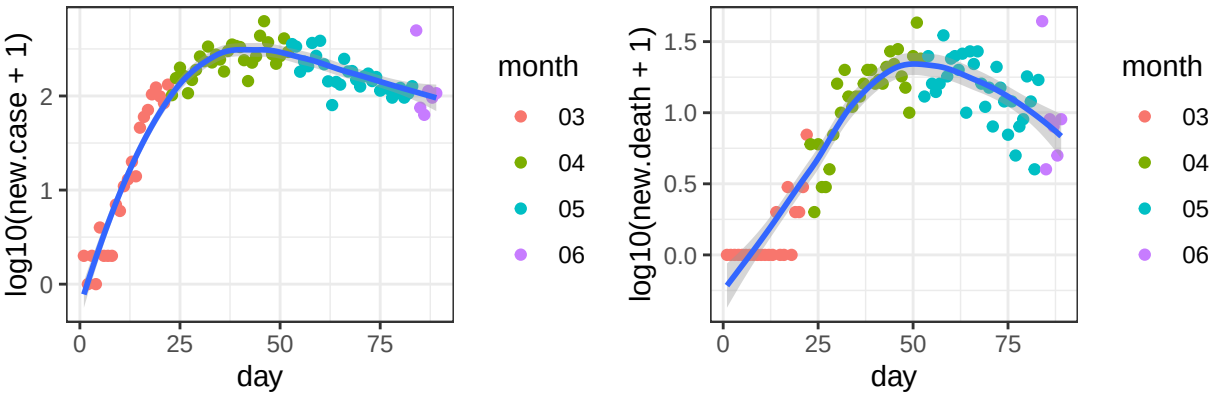
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

New Haven_Connecticut



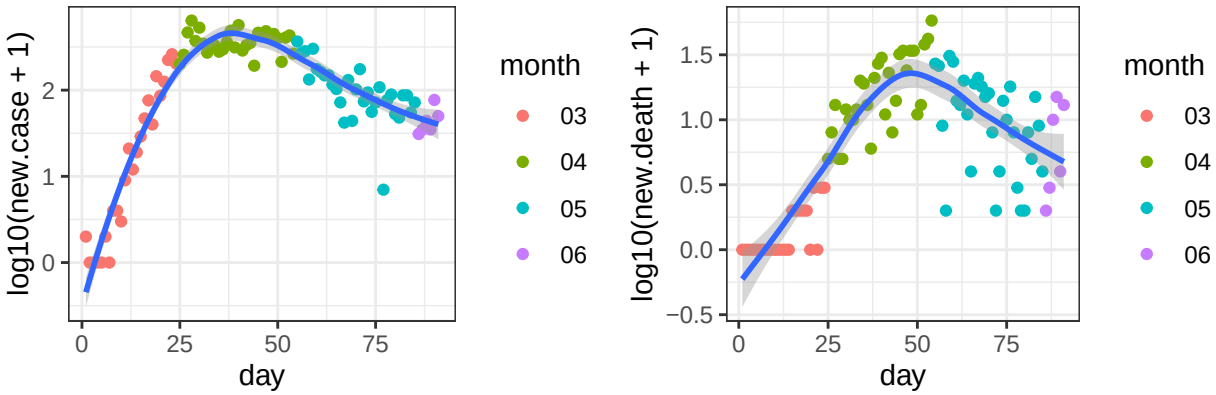
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Essex_Massachusetts



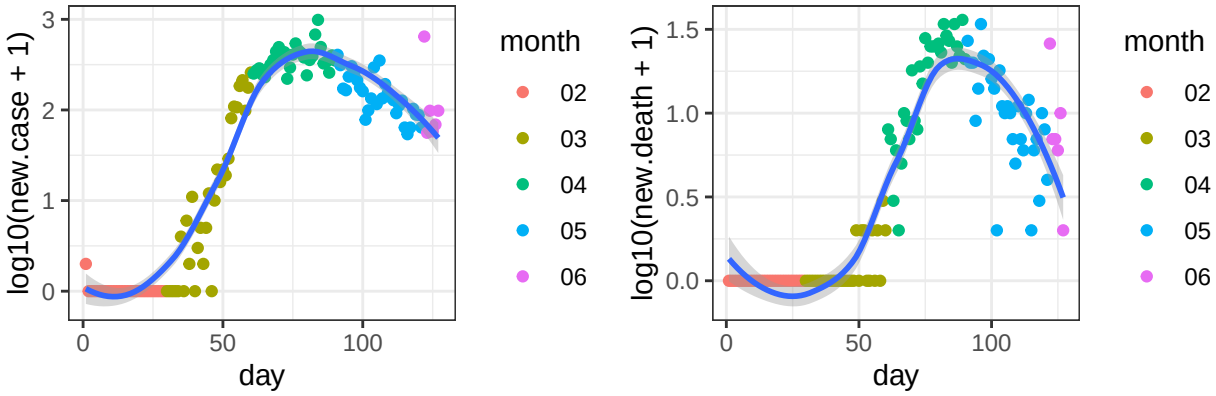
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Passaic_New Jersey



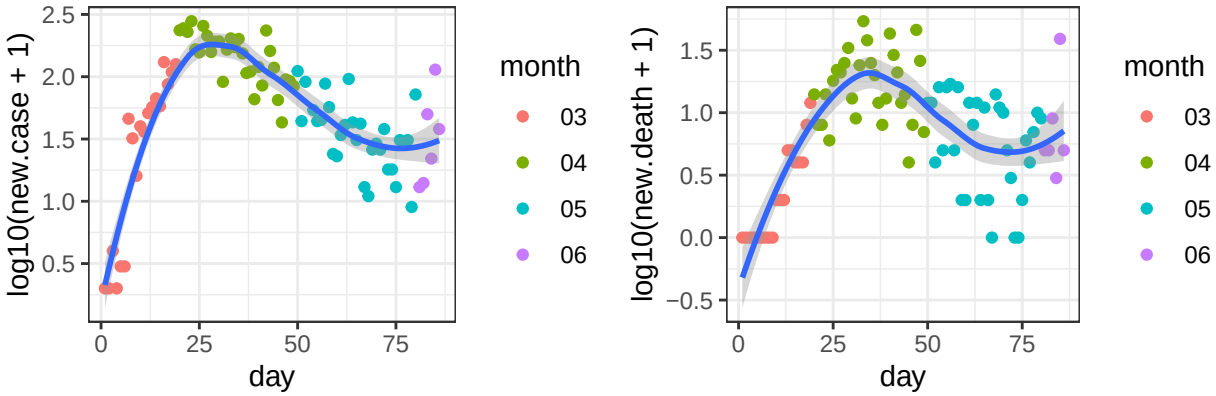
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Suffolk_Massachusetts



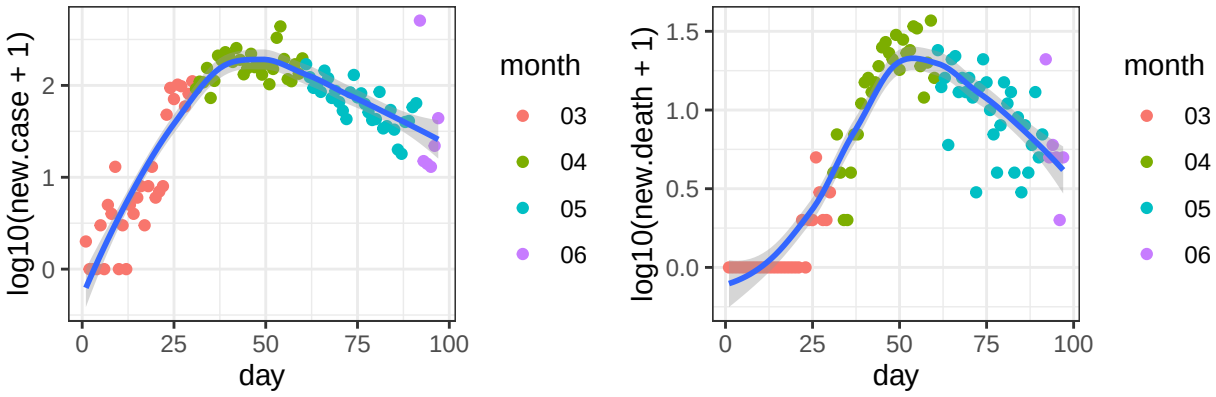
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Macomb_Michigan



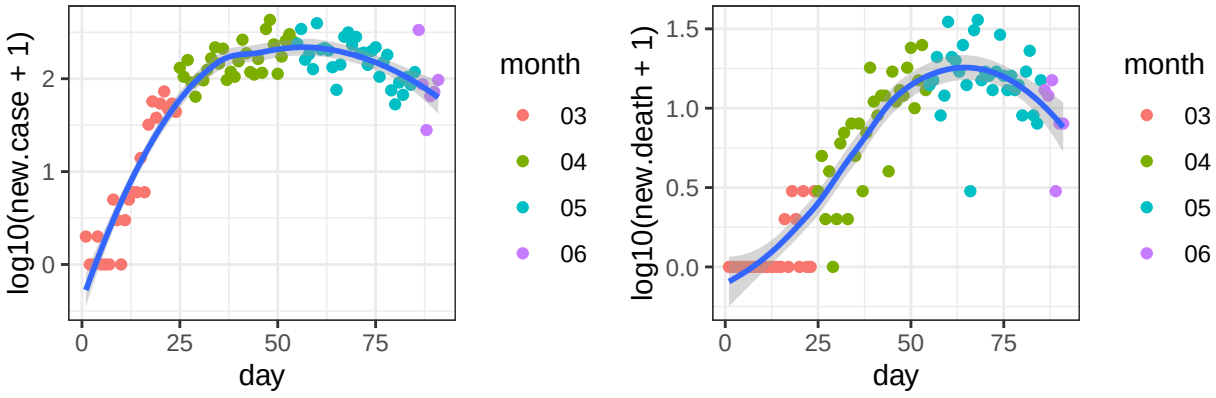
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Norfolk_Massachusetts



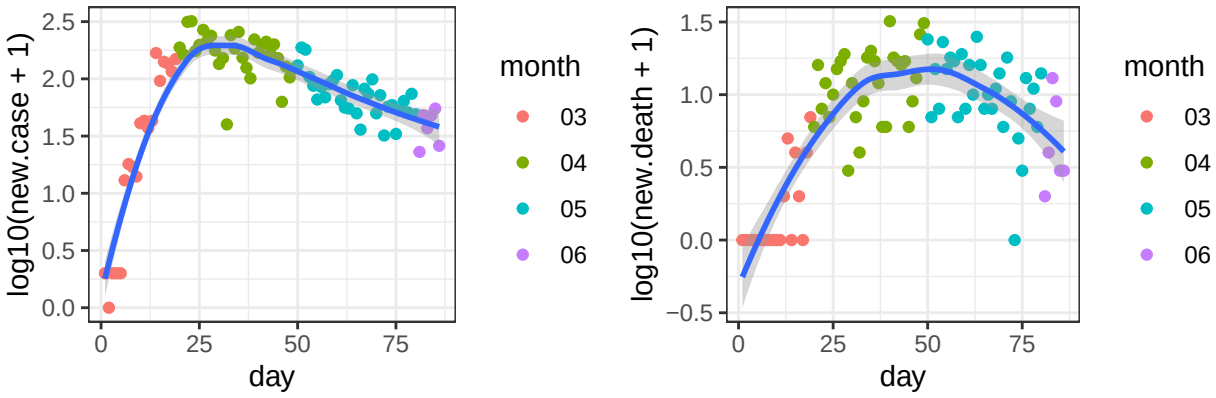
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Worcester_Massachusetts



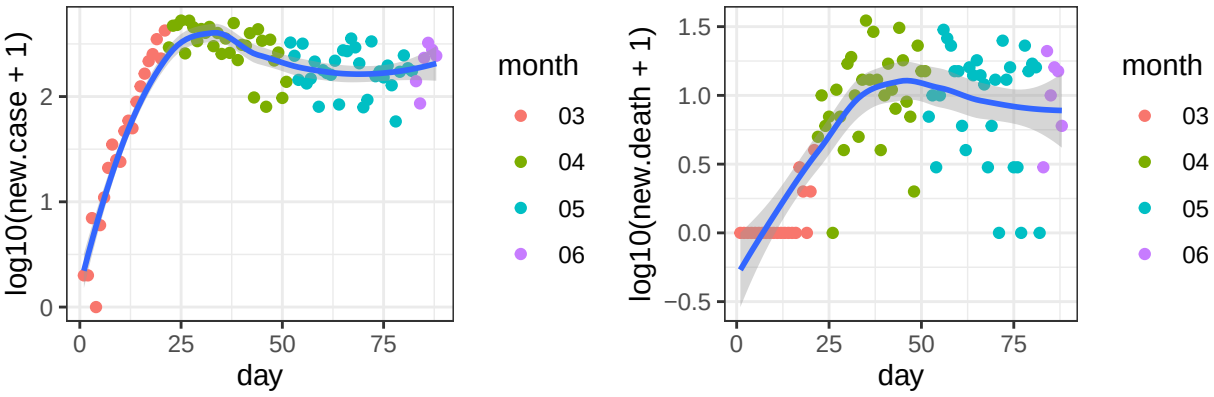
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Ocean_New Jersey



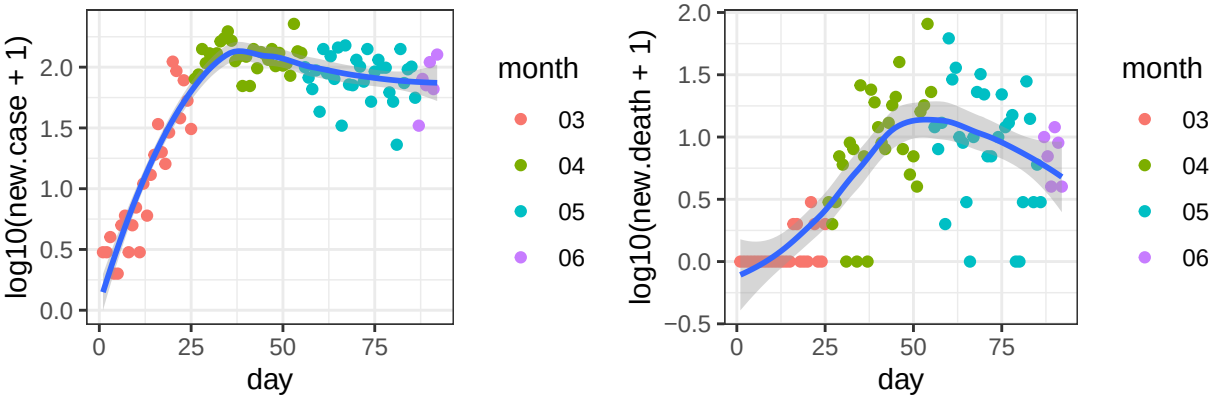
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Miami−Dade_Florida



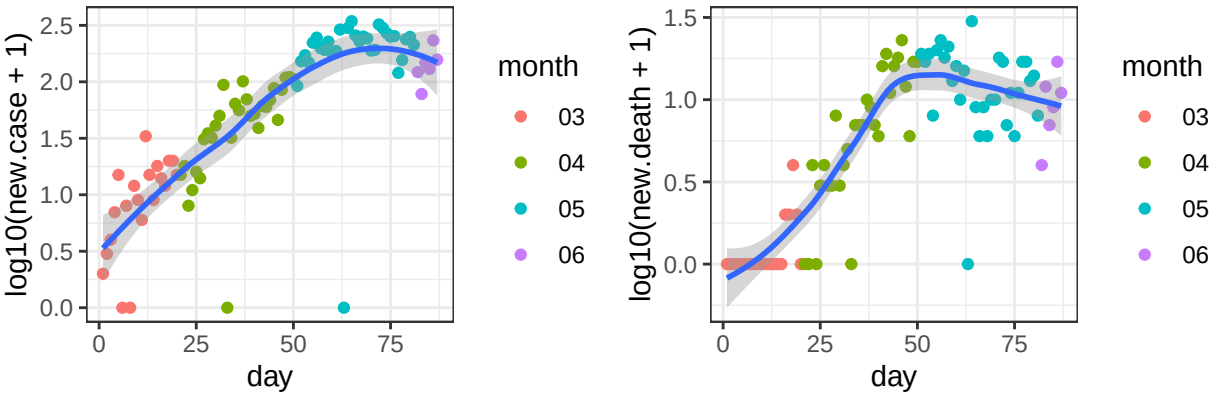
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Montgomery_Pennsylvania



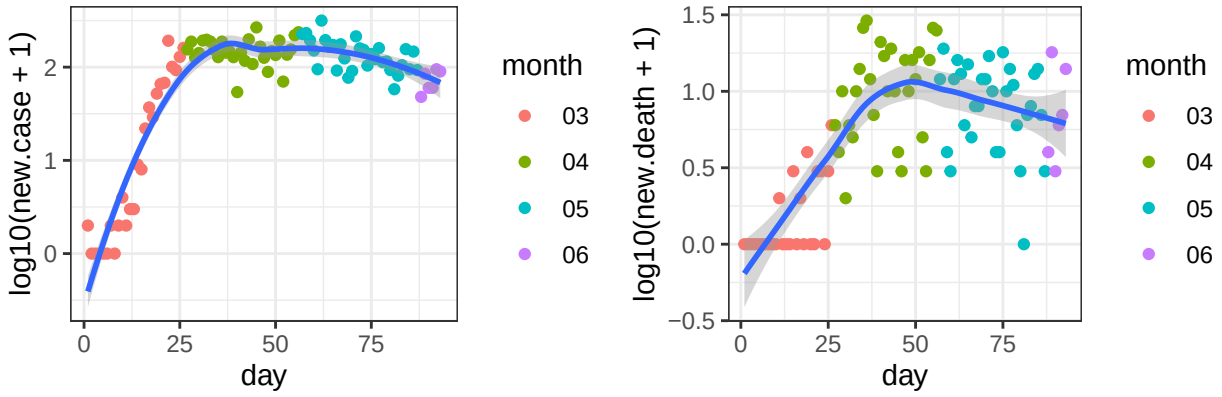
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Hennepin_Minnesota



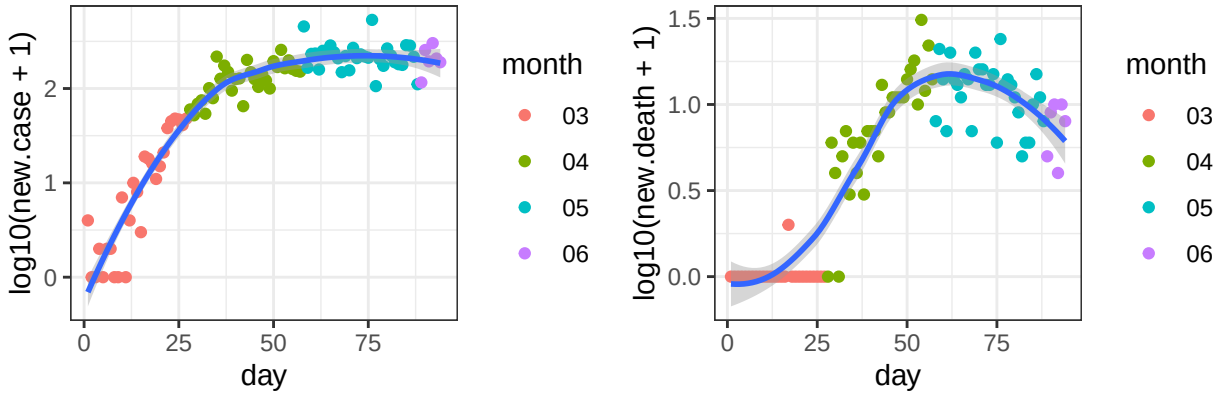
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Marion_Indiana



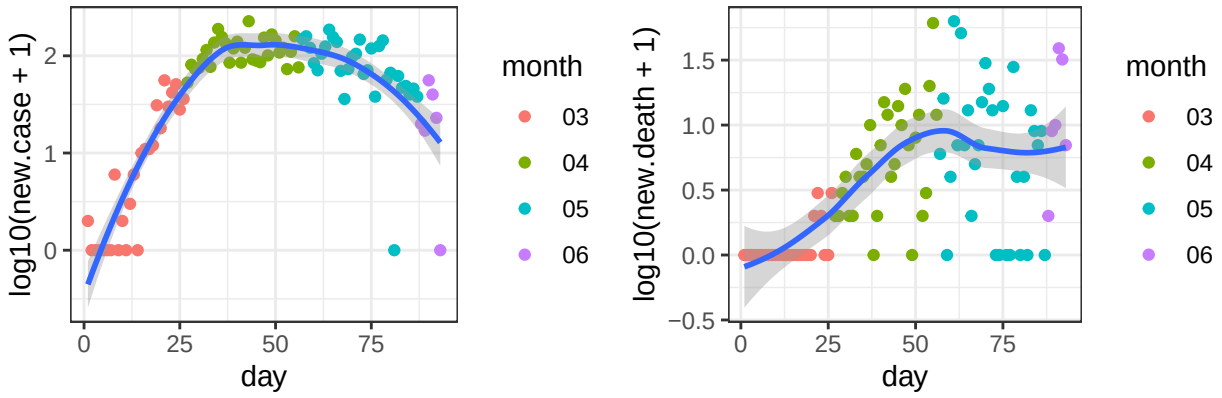
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Montgomery_Maryland



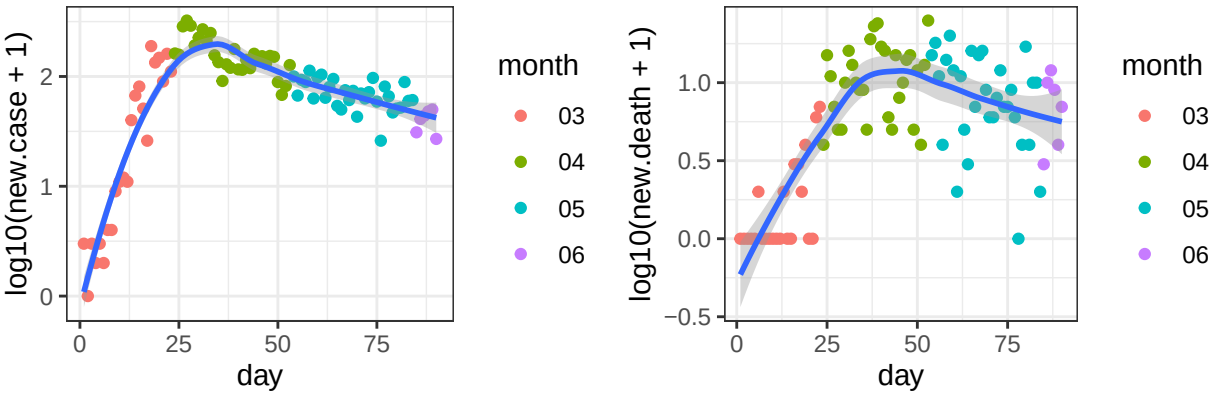
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Delaware_Pennsylvania



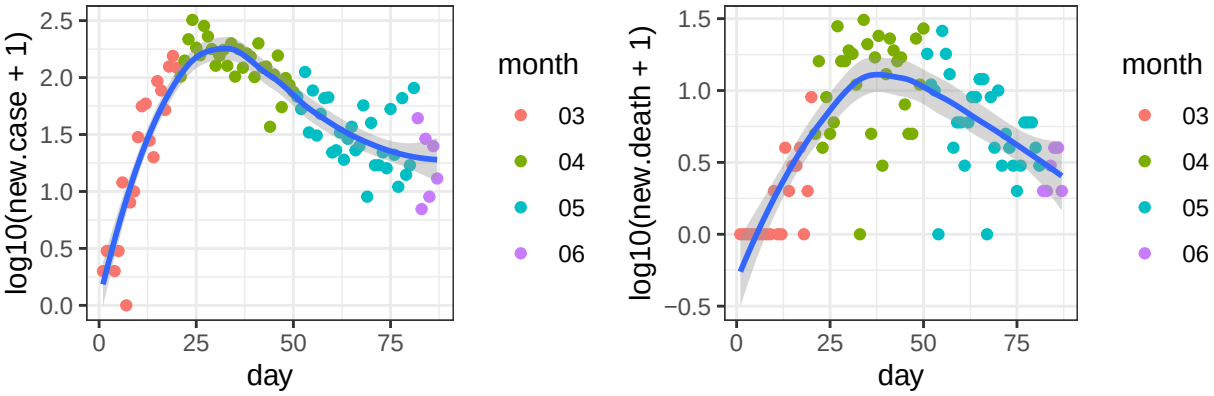
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Monmouth_New Jersey



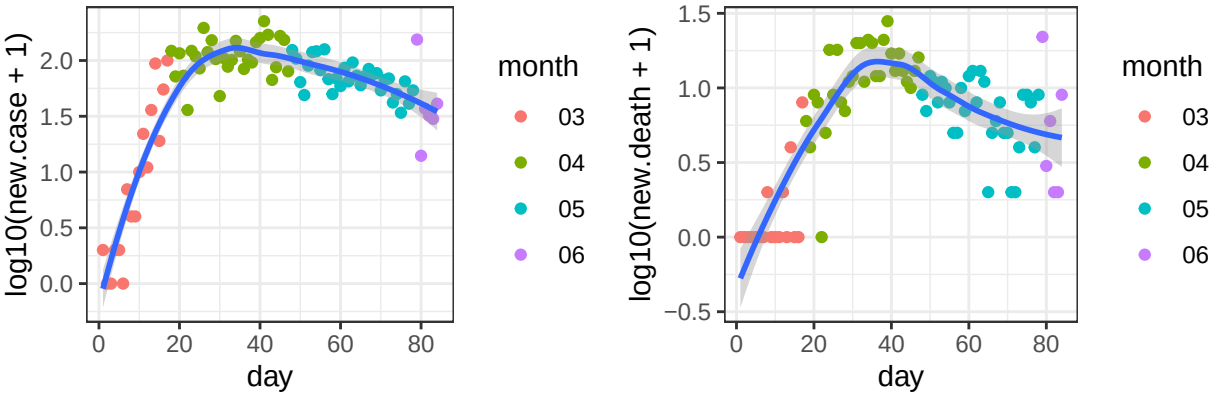
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Morris_New Jersey



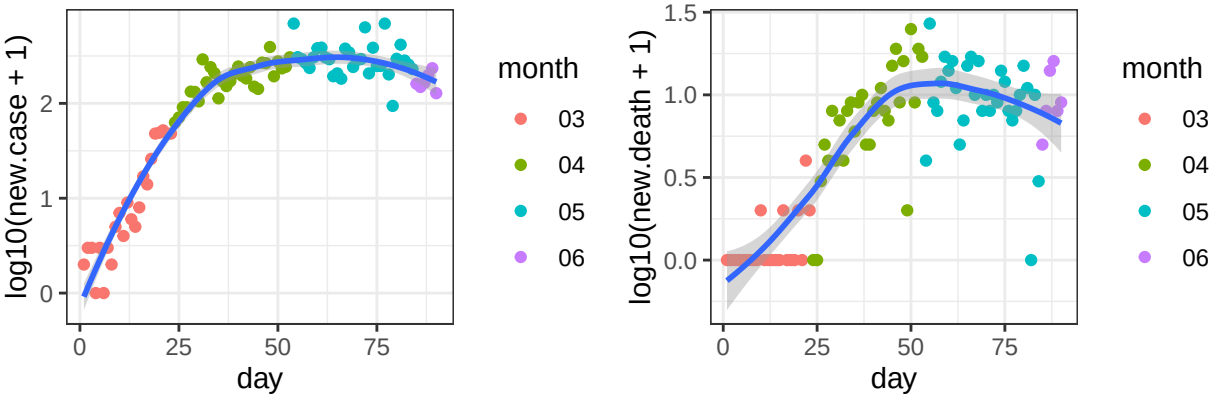
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Hampden_Massachusetts



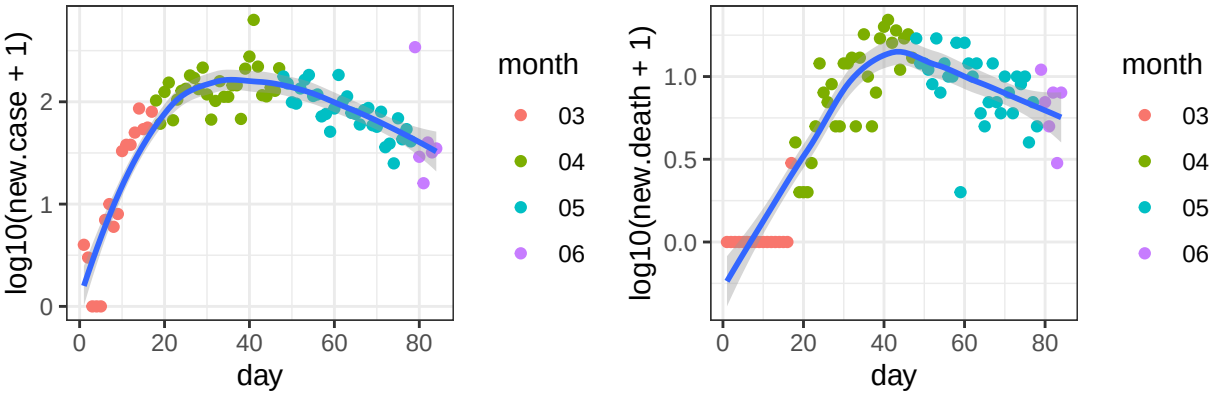
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Prince George's_Maryland



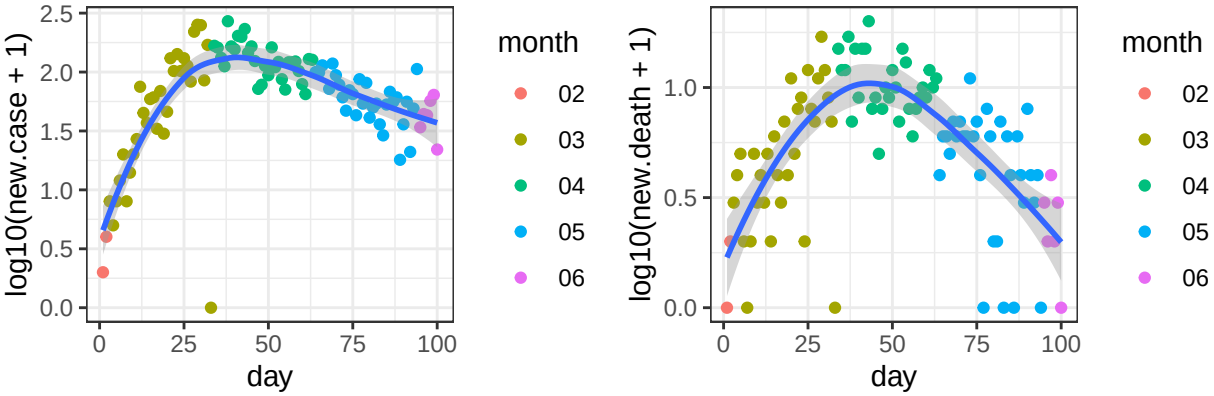
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Plymouth_Massachusetts



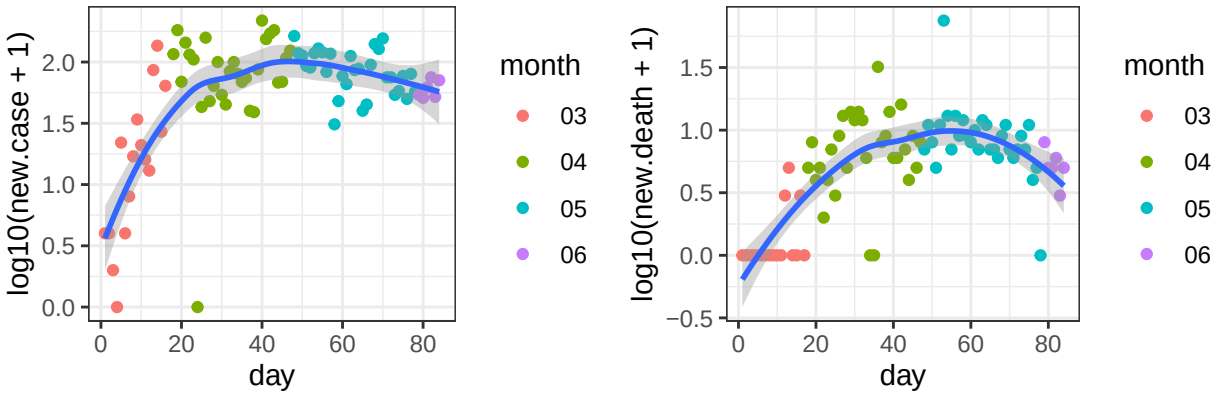
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

King_Washington



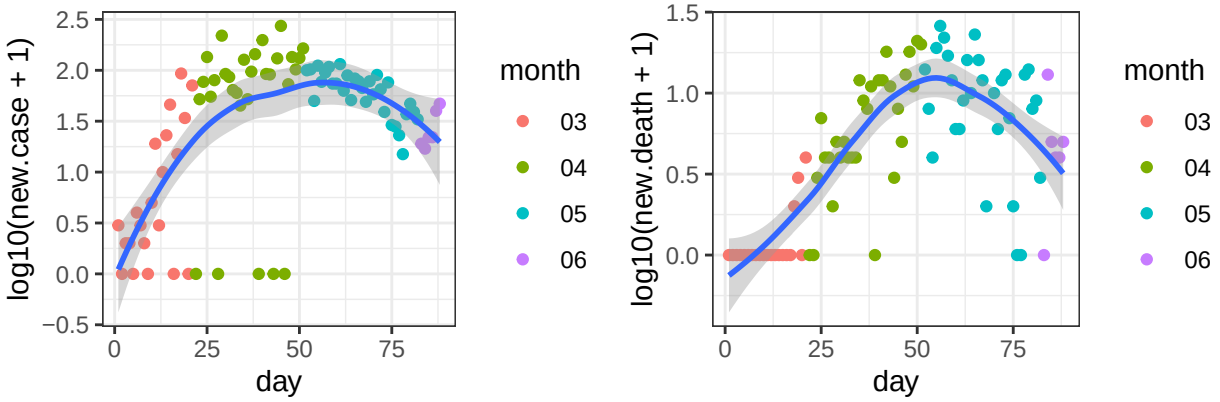
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

Erie_New York



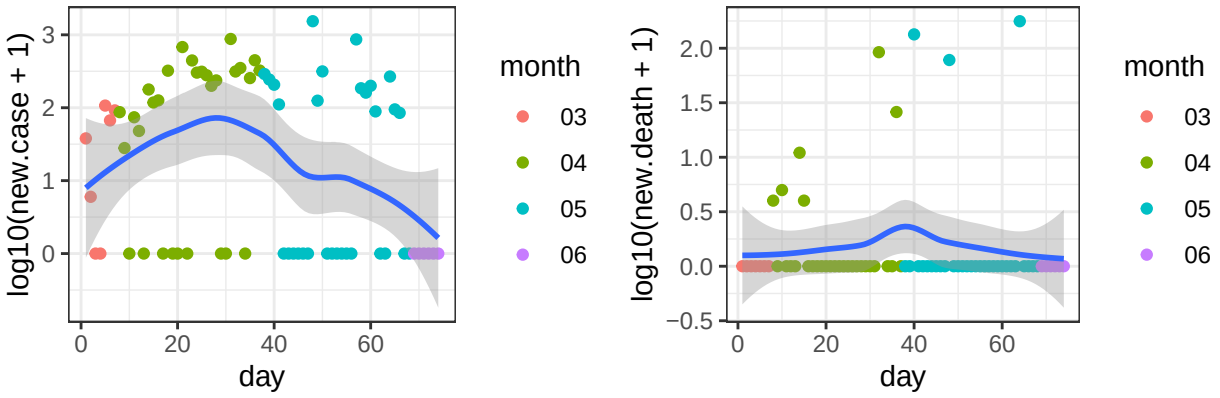
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Bucks_Pennsylvania



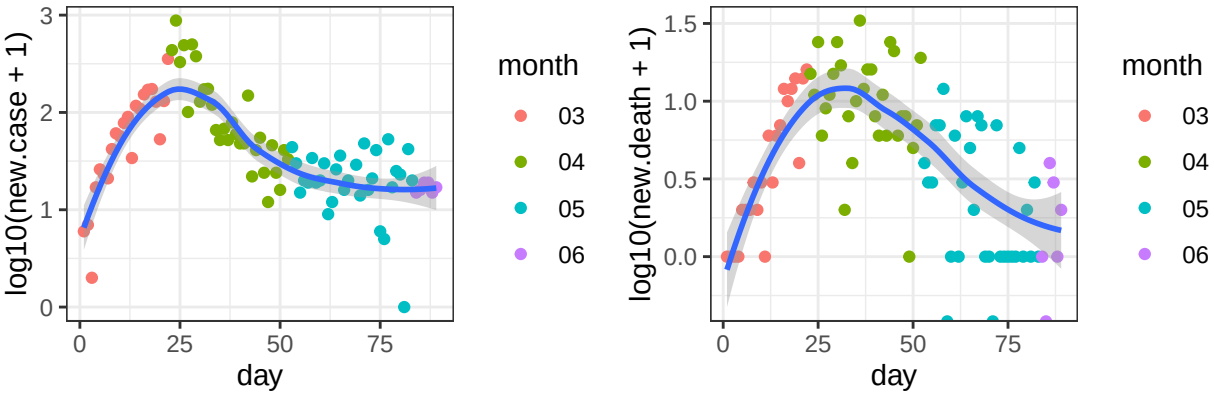
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Providence_Rhode Island



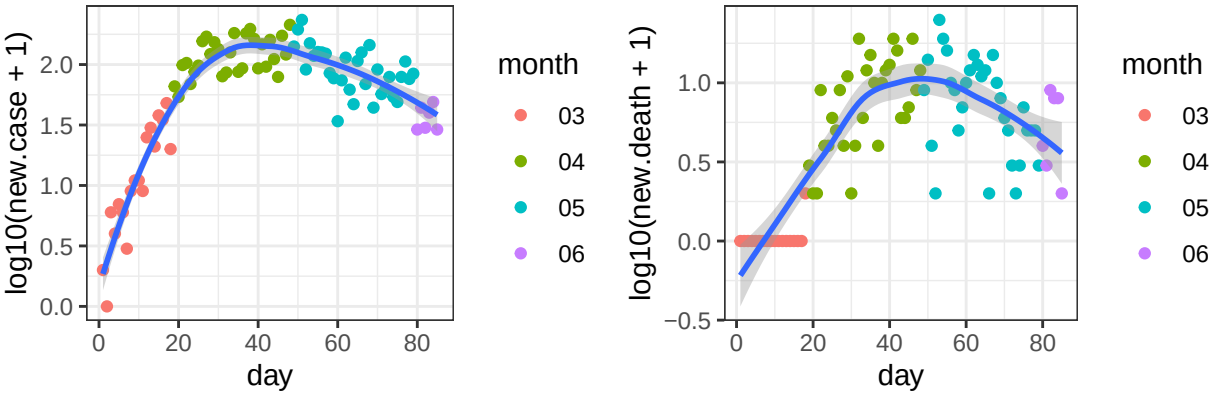
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Orleans_Louisiana



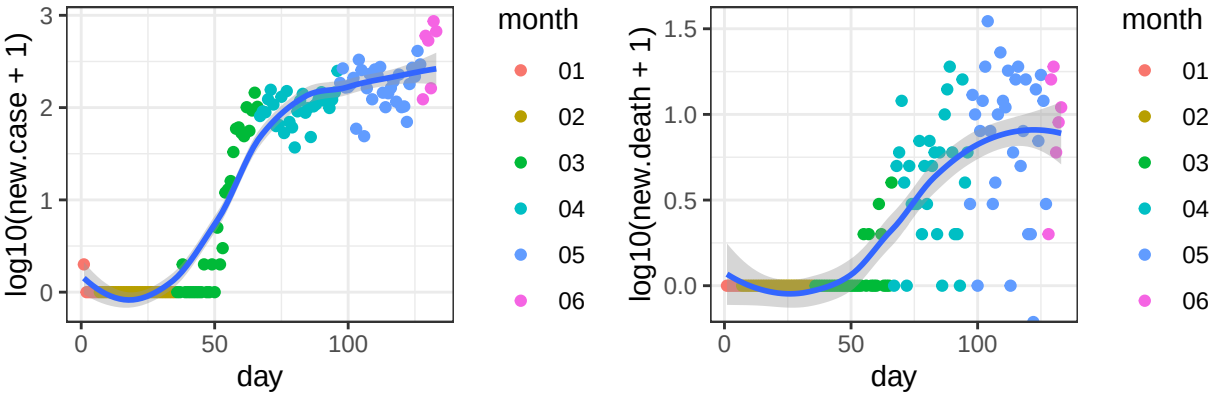
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Mercer_New Jersey



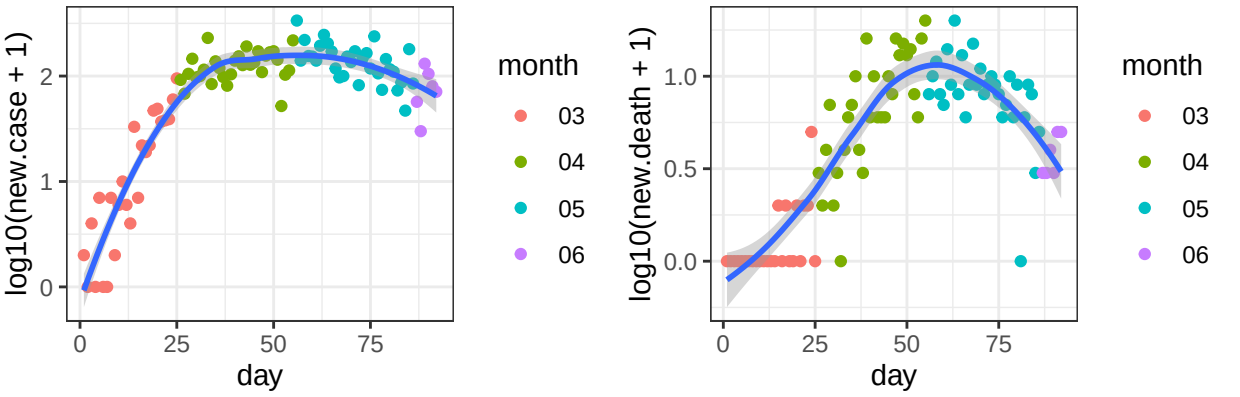
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Maricopa_Arizona



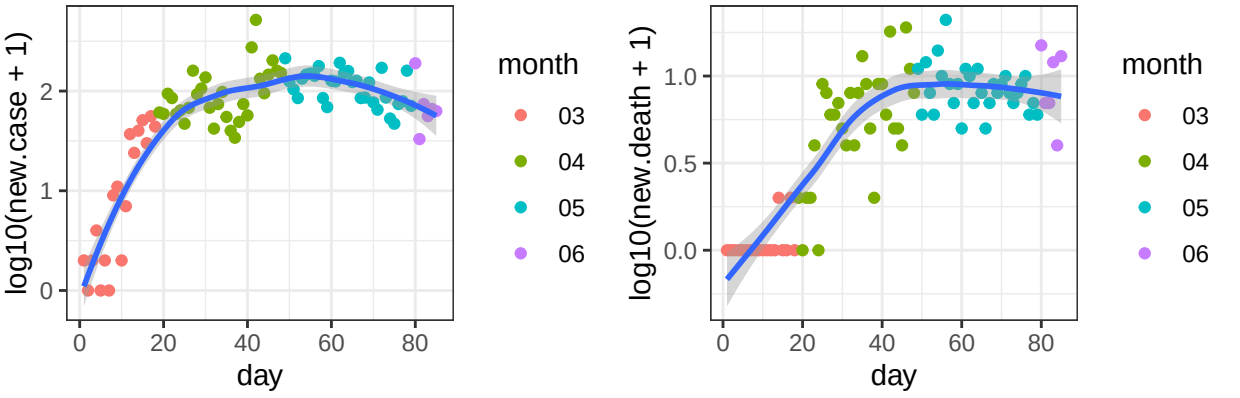
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

District of Columbia_District of Columbia



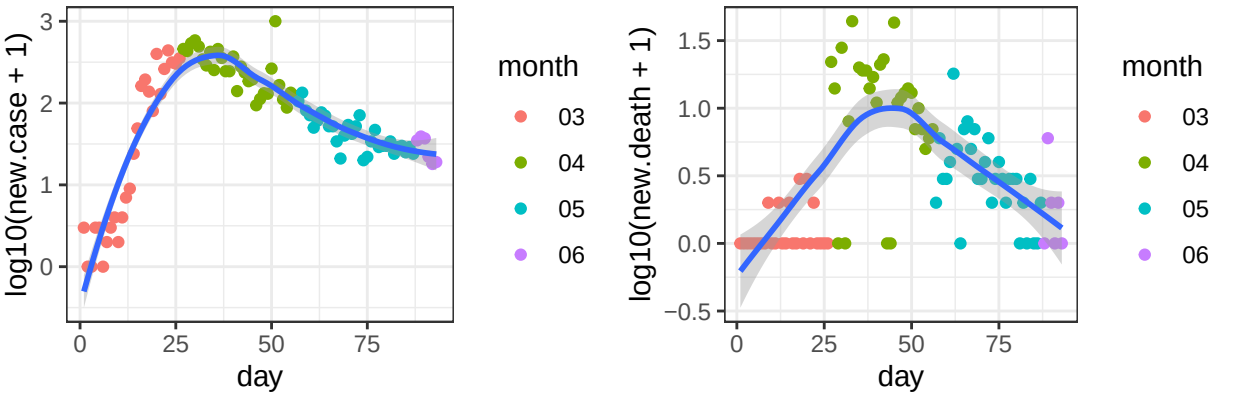
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Bristol_Massachusetts



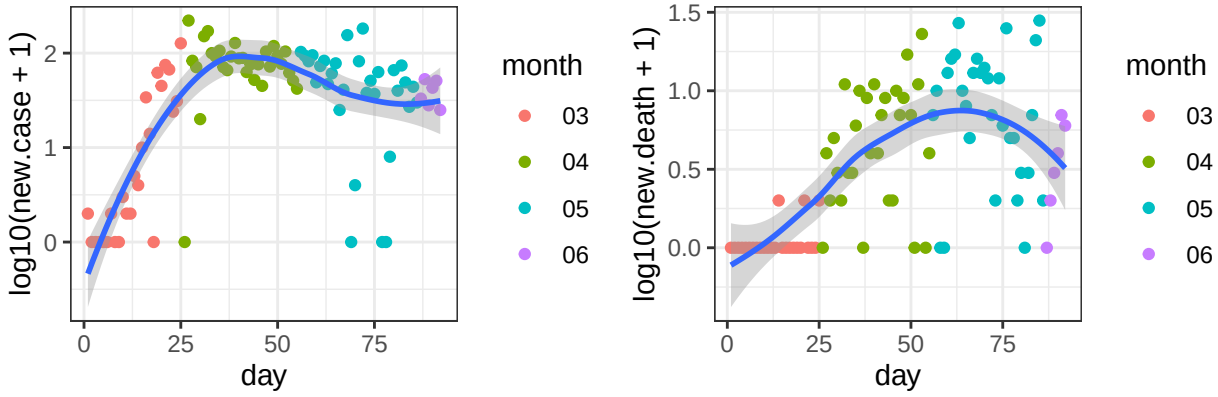
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Rockland_New York



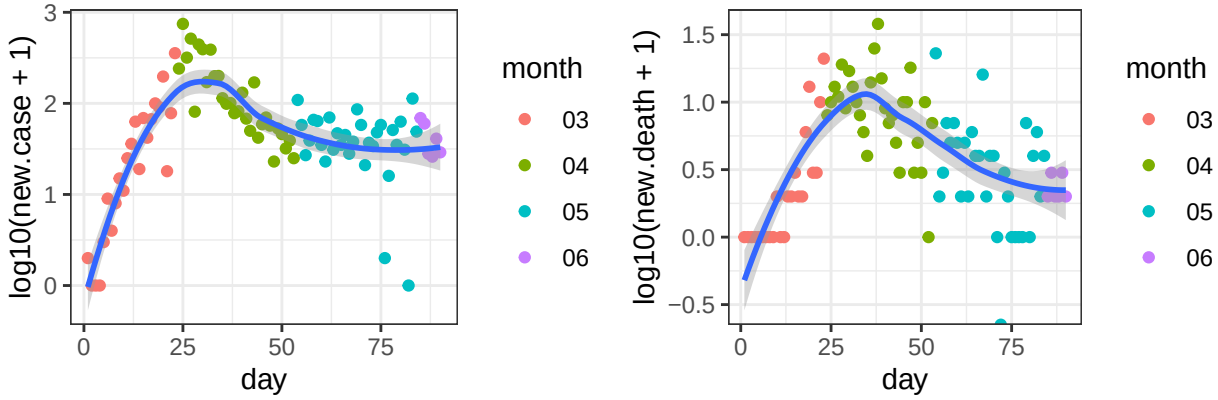
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

St. Louis_Missouri



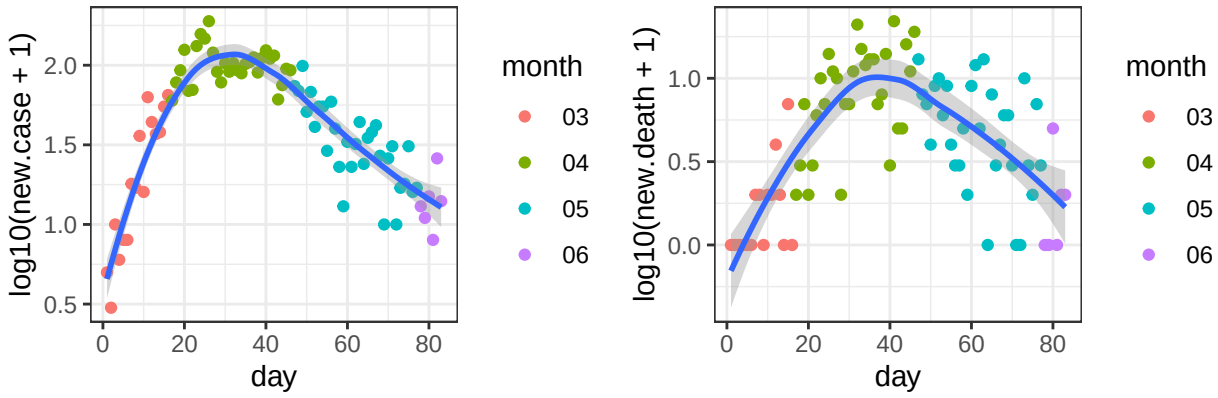
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Jefferson_Louisiana



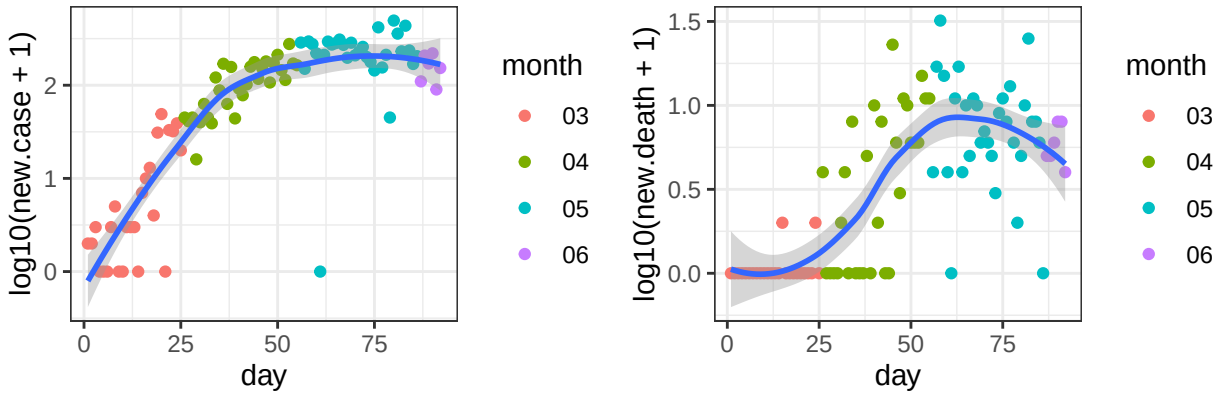
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Somerset_New Jersey



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

Fairfax_Virginia

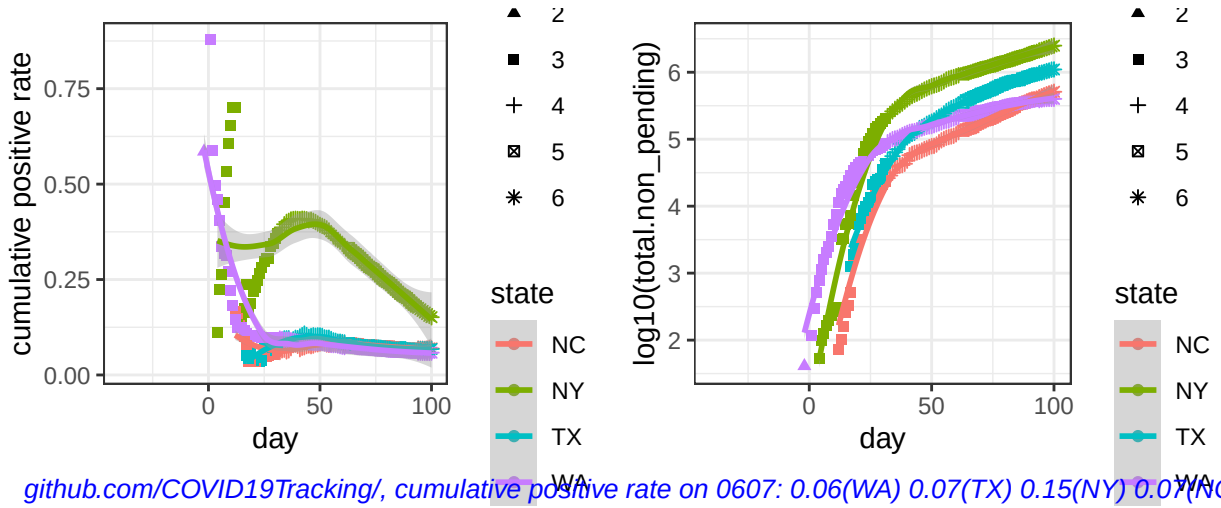


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

COVID Trackng

The positive rates of testing can be an indicator on how much the COVID-19 has spread. However, they are 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.

Since the daily postive rate can fluctuate a lot, here I only illustrae the cumulative positave rate across time, for four states with grade A data. Of course since this is an R markdown file, you can modify the source code and check for other states.



github.com/COVID19Tracking/, cumulative positive rate on 0607: 0.06(WA) 0.07(TX) 0.15(NY) 0.07(NC)

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
```