Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-05-24

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

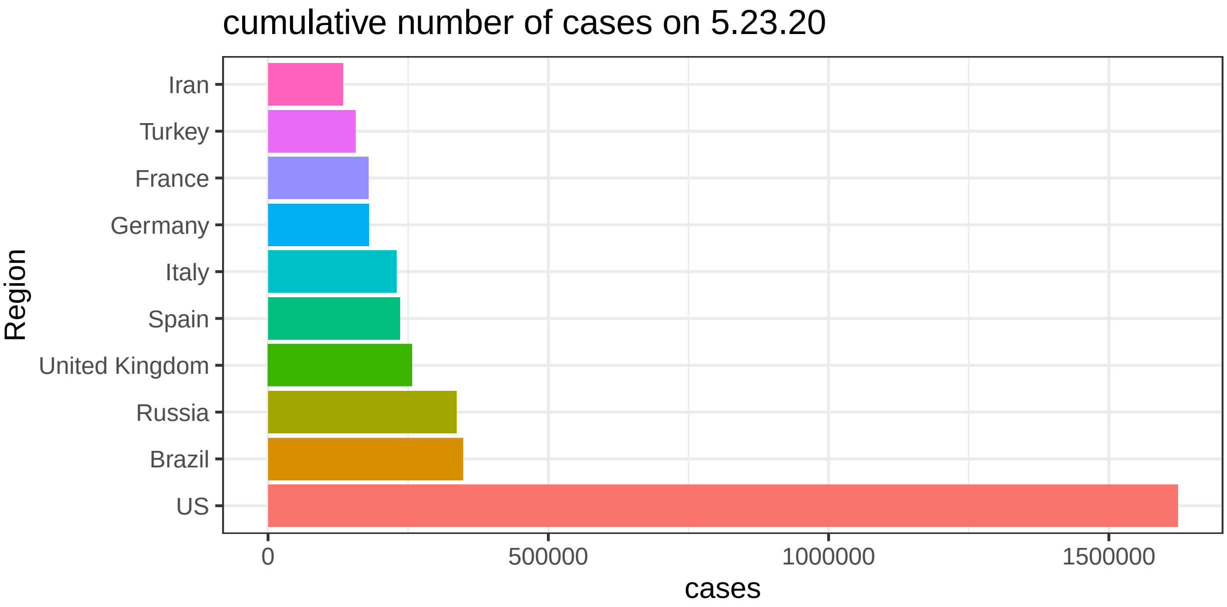
JHU

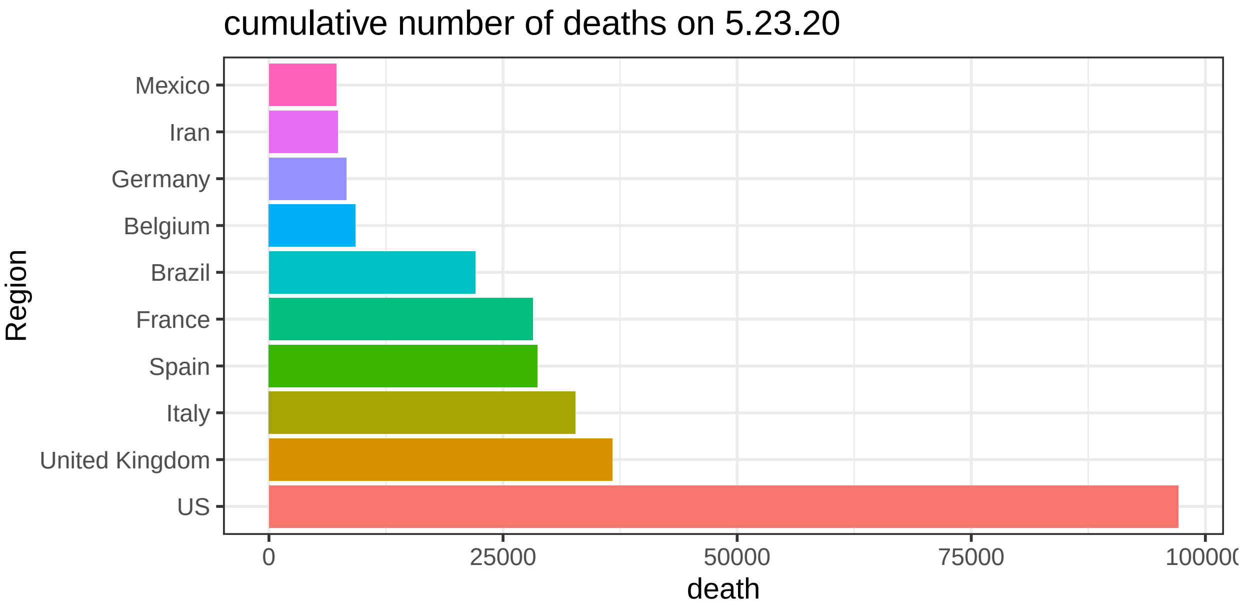
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

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.

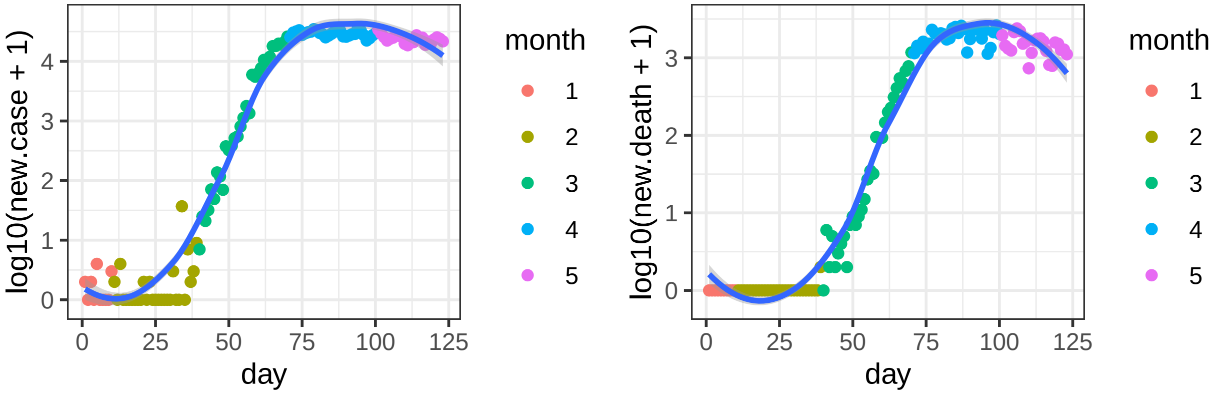
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.




cumulative number of deaths on 5.23.20

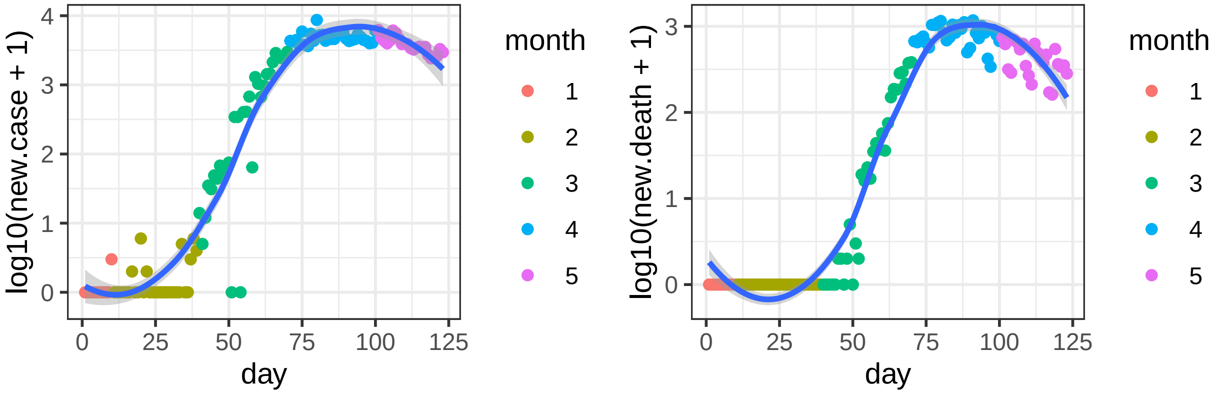
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



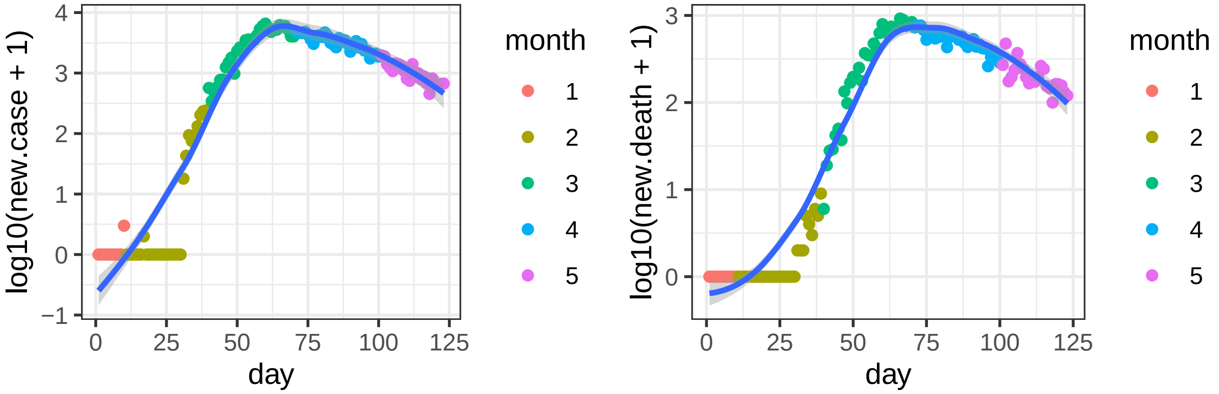
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



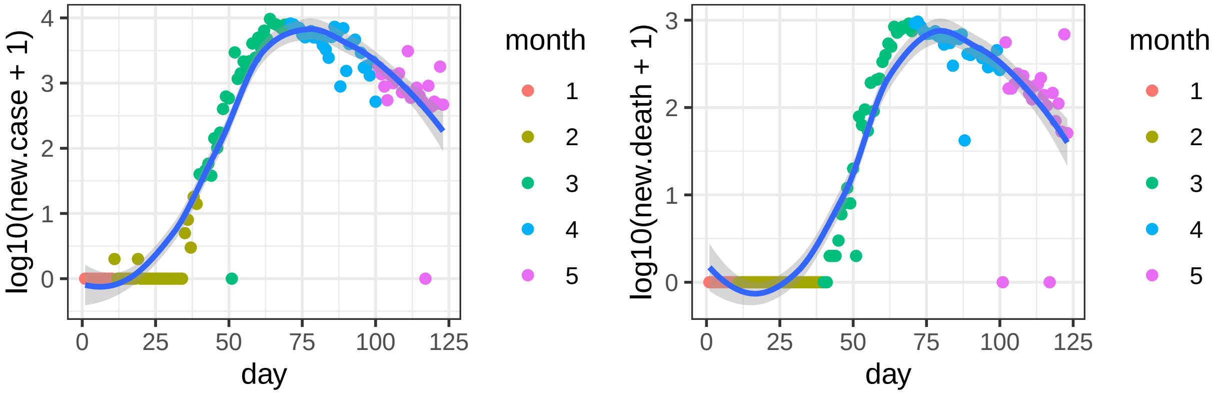
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



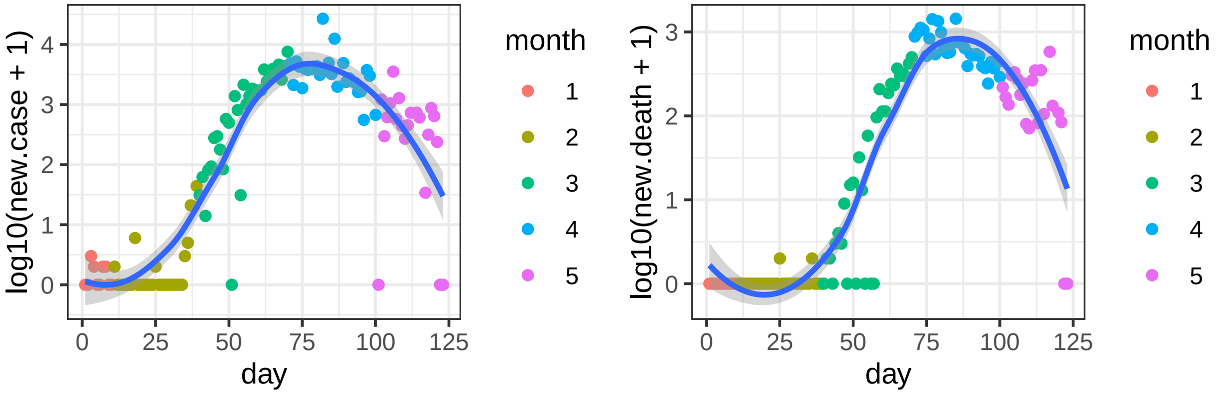
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain



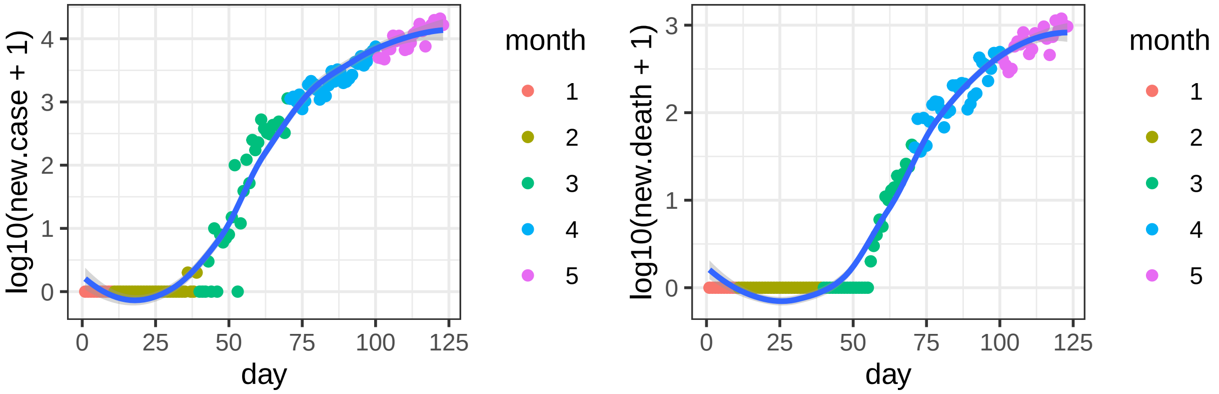
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



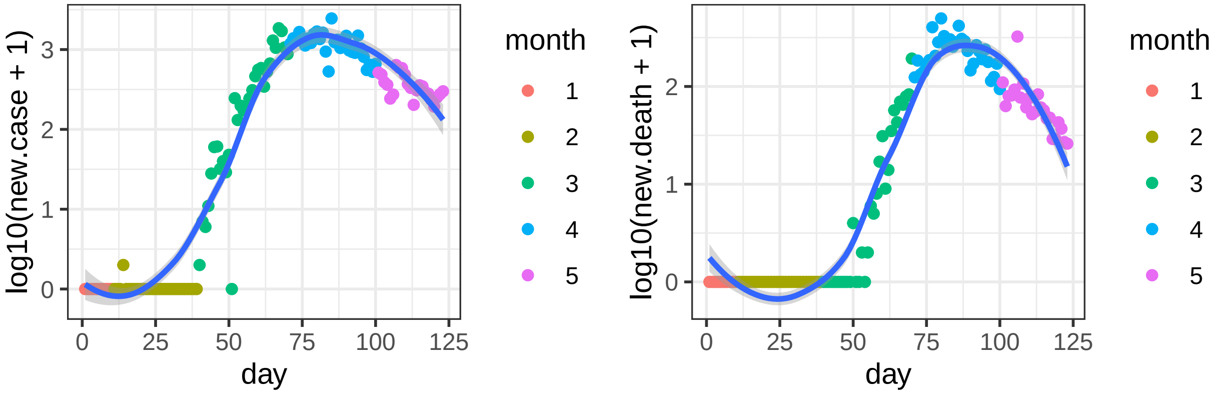
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Brazil



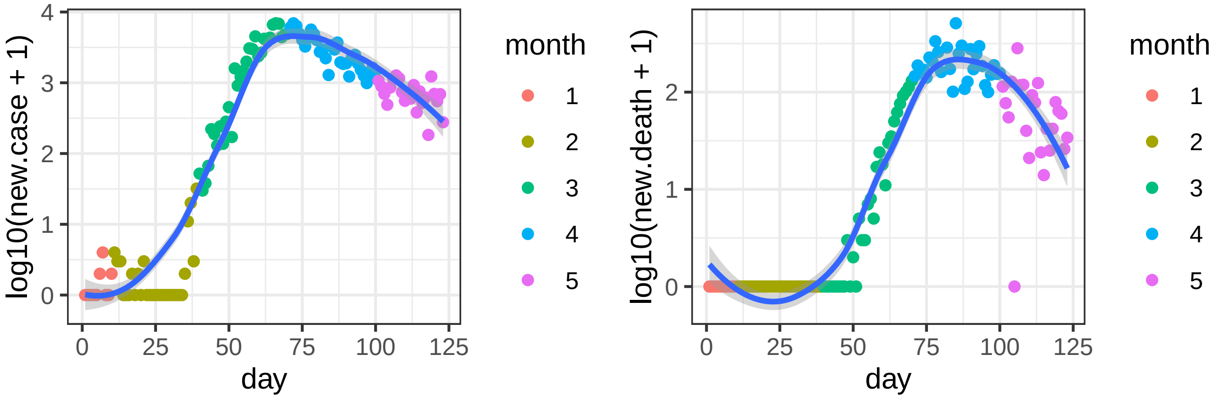
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Belgium

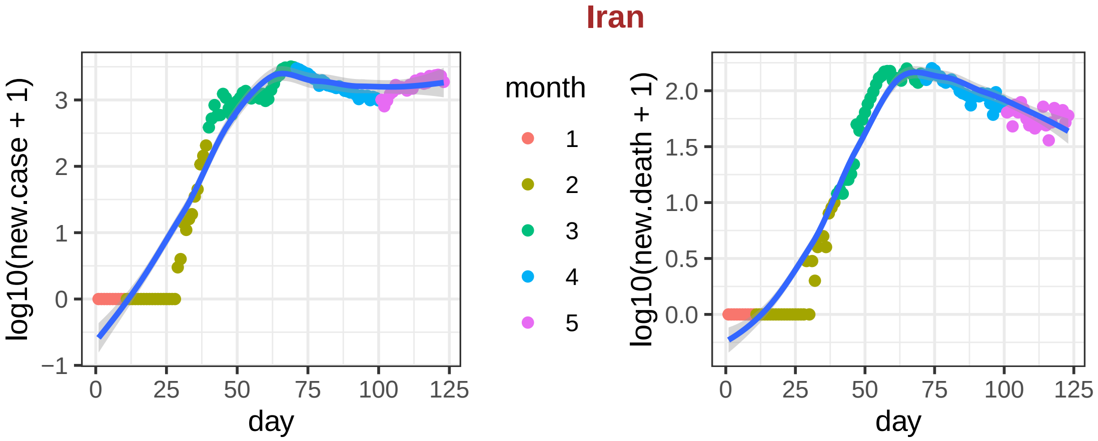


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

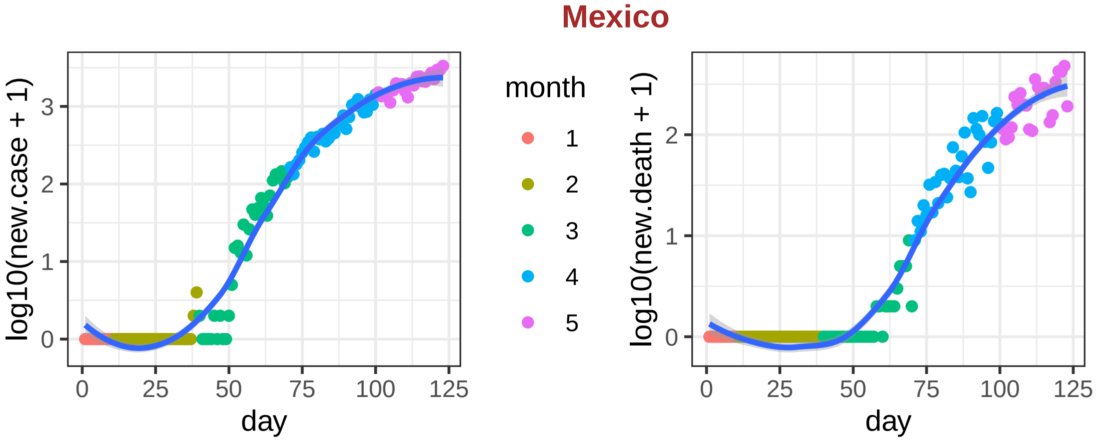
Germany



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



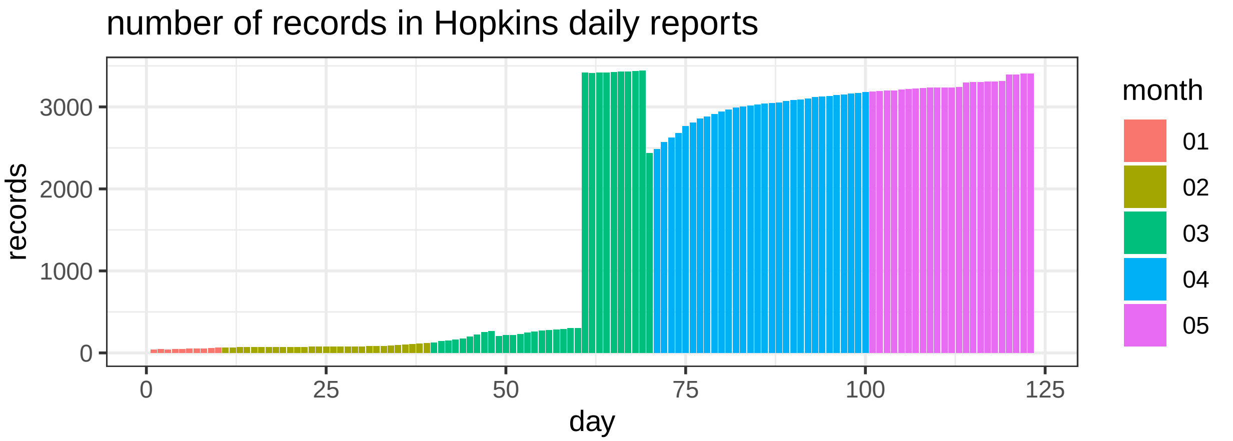
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



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] "2020-05-22"
```

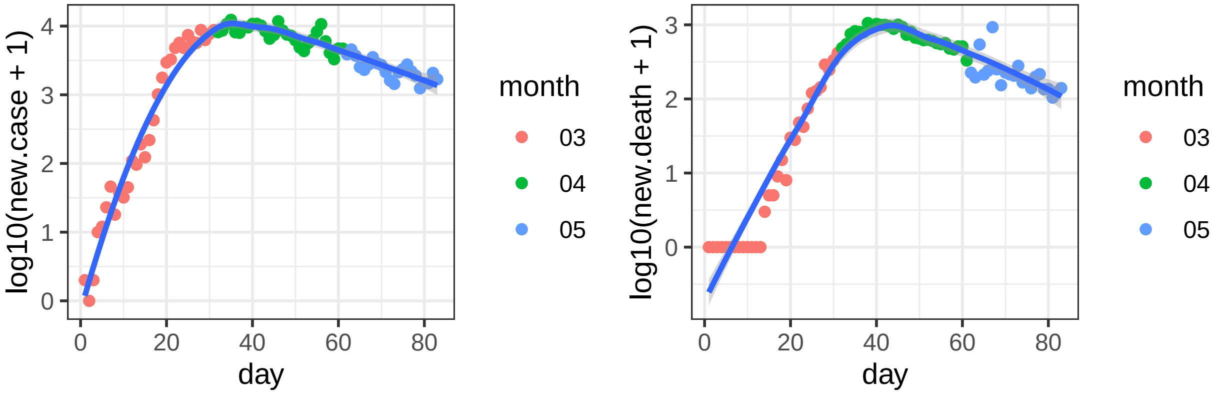
state level data

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

```
##            date                state fips  cases deaths
## 4448 2020-05-22             New York   36 362991  28802
## 4446 2020-05-22           New Jersey   34 152719  10985
## 4437 2020-05-22        Massachusetts   25  90889   6228
## 4438 2020-05-22             Michigan   26  53865   5158
## 4455 2020-05-22         Pennsylvania   42  70305   5032
## 4429 2020-05-22             Illinois   17 105710   4740
## 4419 2020-05-22           California    6  90801   3690
## 4421 2020-05-22          Connecticut    9  39640   3637
## 4434 2020-05-22            Louisiana   22  37048   2669
## 4436 2020-05-22             Maryland   24  44539   2207
## 4424 2020-05-22              Florida   12  49443   2189
## 4430 2020-05-22              Indiana   18  31165   1941
## 4452 2020-05-22                 Ohio   39  30795   1872
## 4425 2020-05-22              Georgia   13  39734   1779
## 4461 2020-05-22                Texas   48  54369   1498
## 4420 2020-05-22             Colorado    8  23456   1324
## 4465 2020-05-22             Virginia   51  34950   1136
## 4466 2020-05-22           Washington   53  20274   1061
## 4439 2020-05-22            Minnesota   27  19014    851
## 4417 2020-05-22              Arizona    4  15608    775
## 4449 2020-05-22       North Carolina   37  21661    754
## 4441 2020-05-22             Missouri   29  11797    681
## 4440 2020-05-22          Mississippi   28  12624    596
## 4457 2020-05-22         Rhode Island   44  13736    579
## 4415 2020-05-22              Alabama    1  13670    541
## 4468 2020-05-22            Wisconsin   55  14557    496
## 4431 2020-05-22                 Iowa   19  16510    441
## 4458 2020-05-22       South Carolina   45   9638    419
## 4423 2020-05-22 District of Columbia   11   7893    418
## 4433 2020-05-22             Kentucky   21   8688    398
```

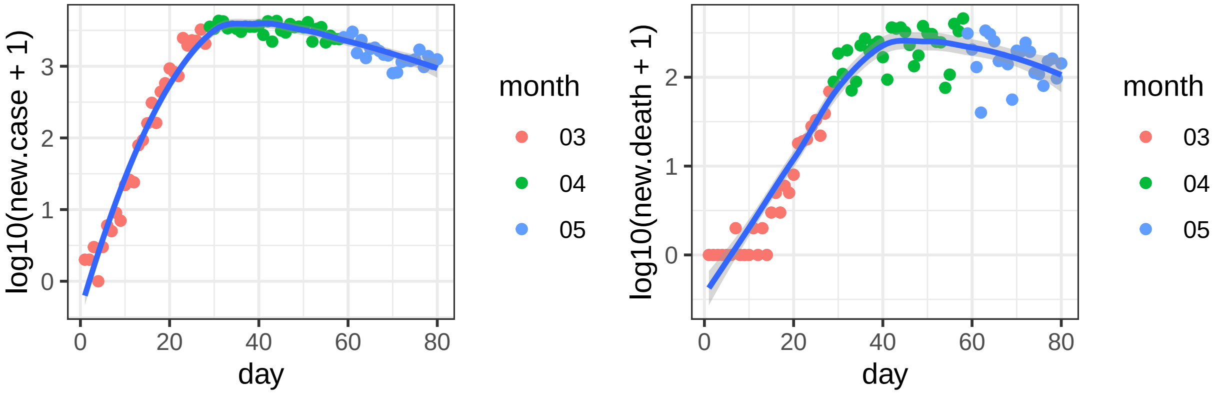
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



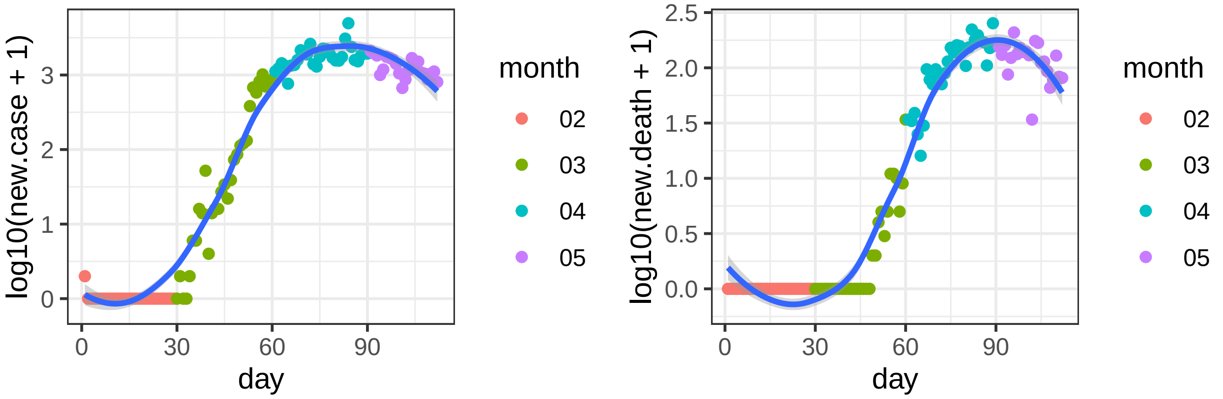
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey

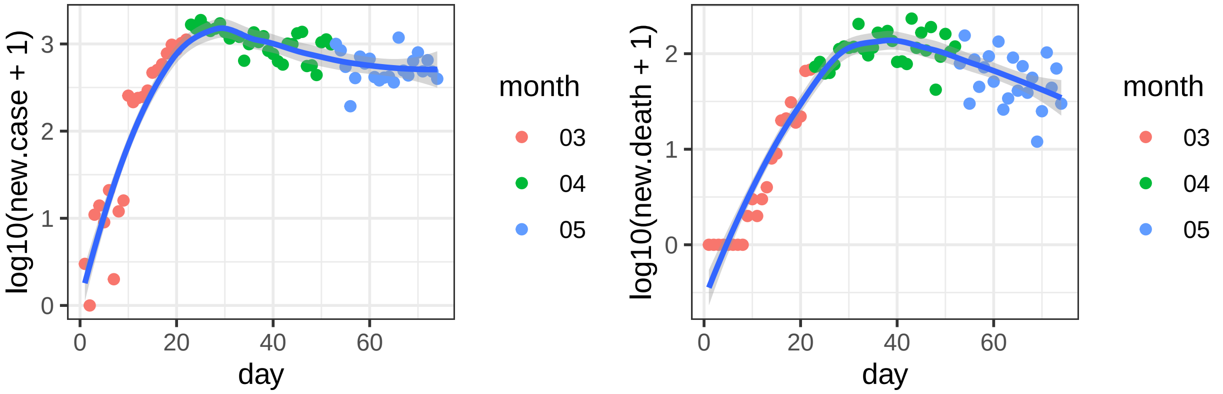


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

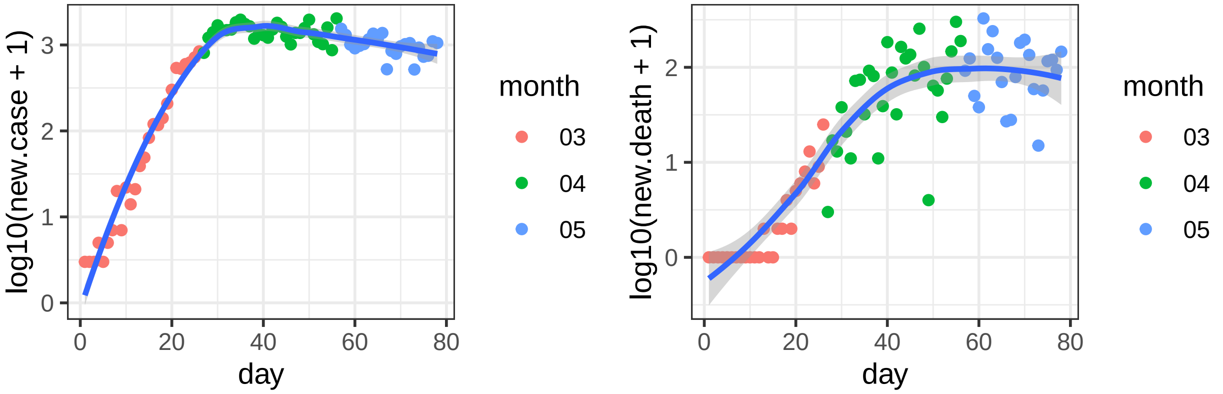
Massachusetts



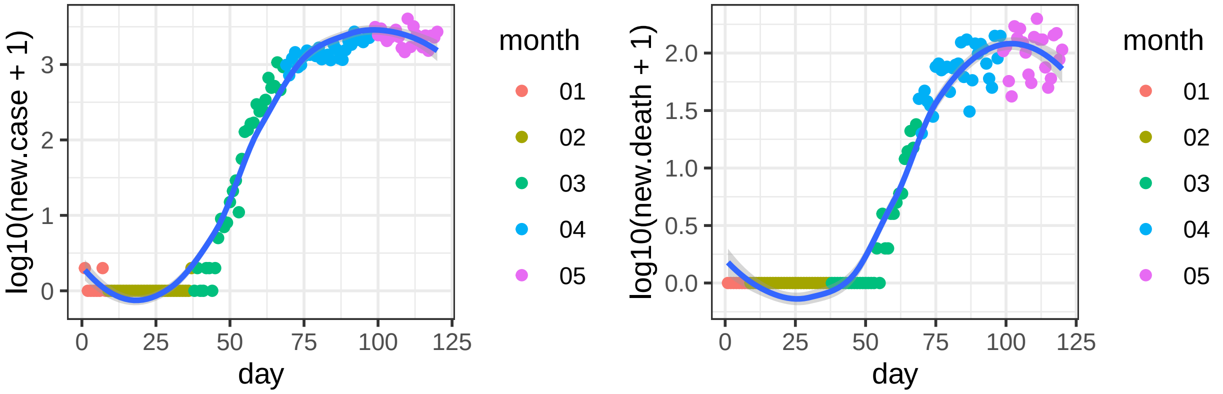
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01


Michigan

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

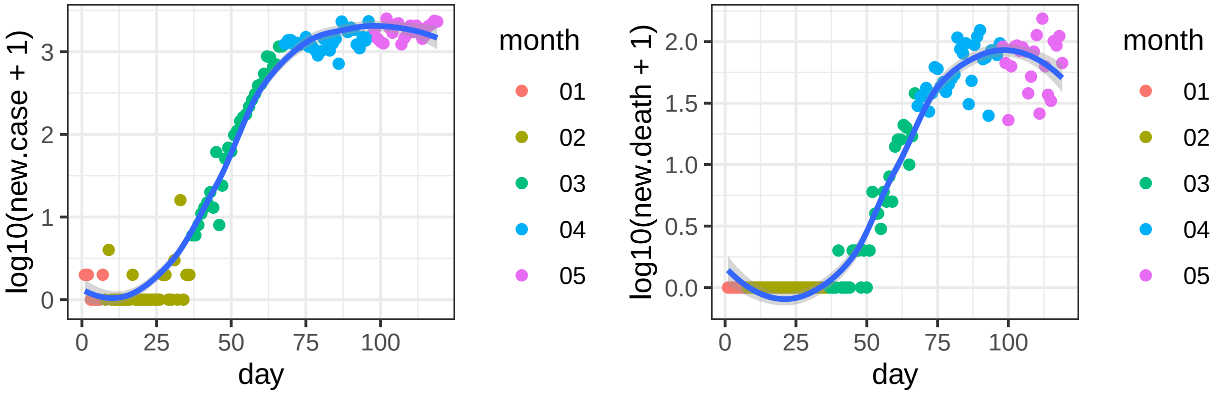

Pennsylvania

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


Illinois

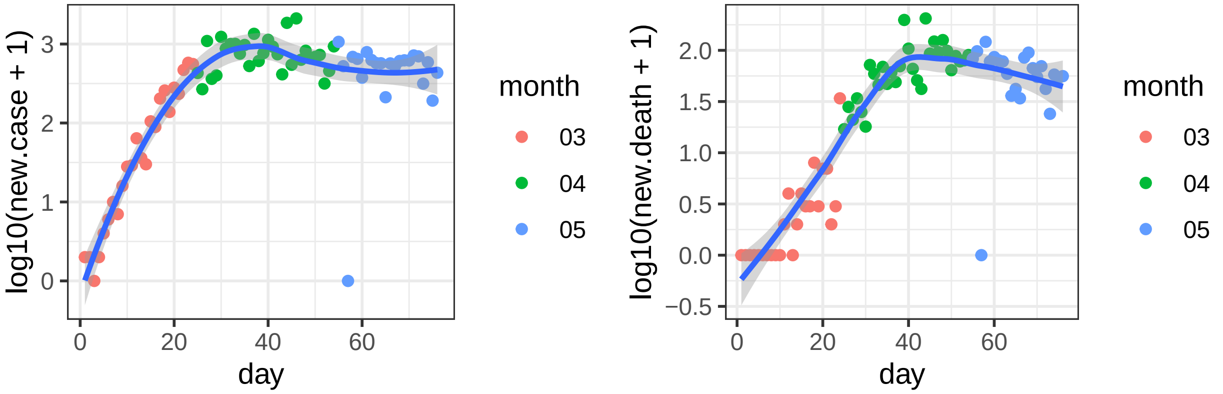
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

California



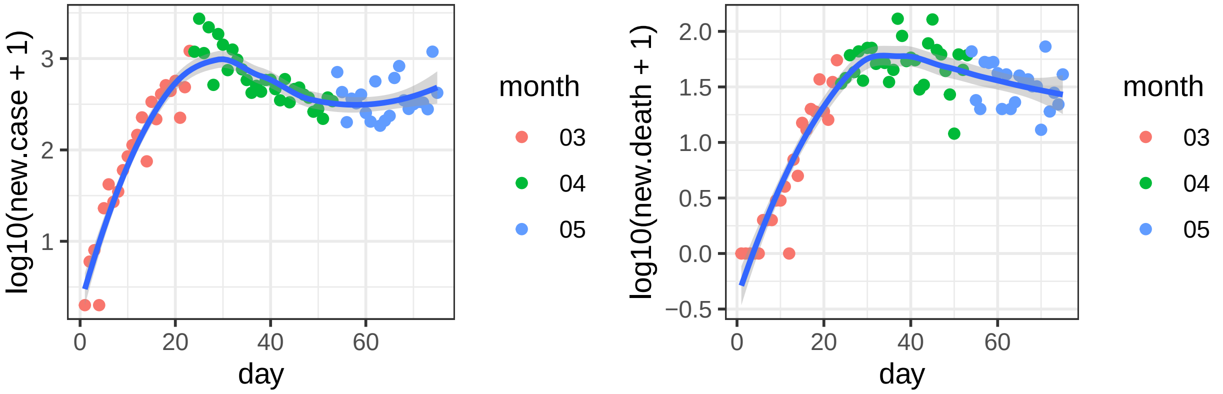
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Connecticut



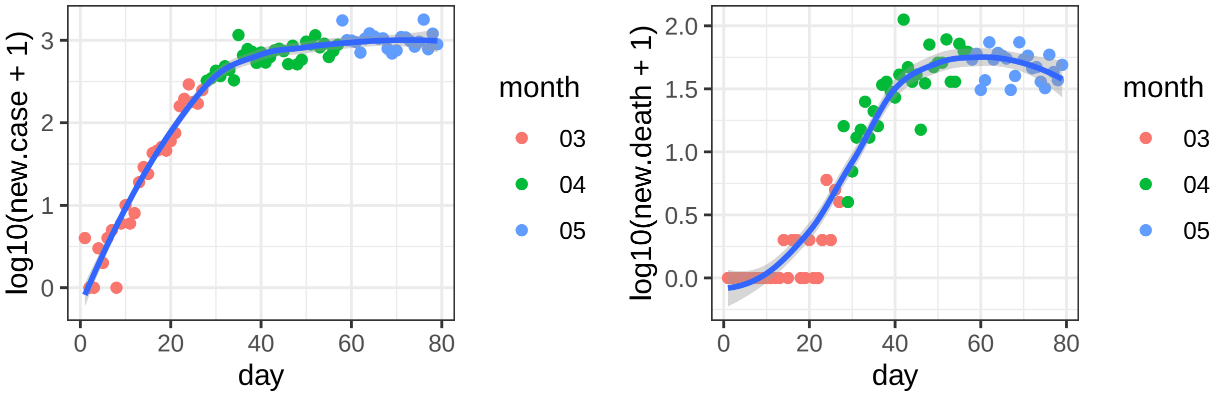
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Louisiana



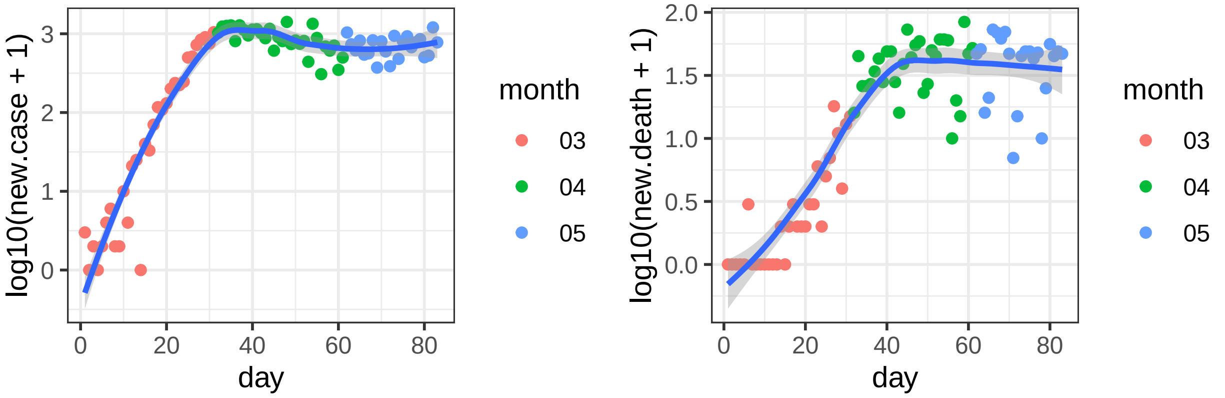
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



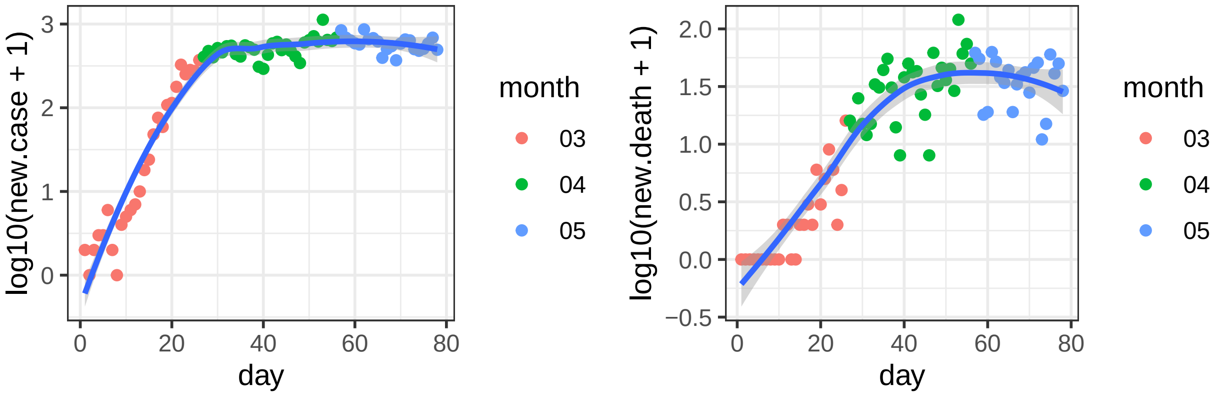
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Florida



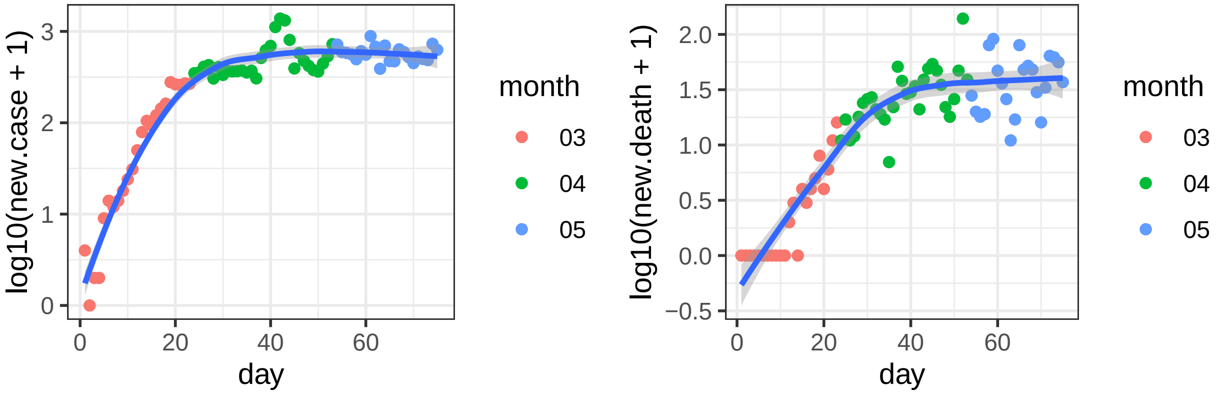
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Indiana



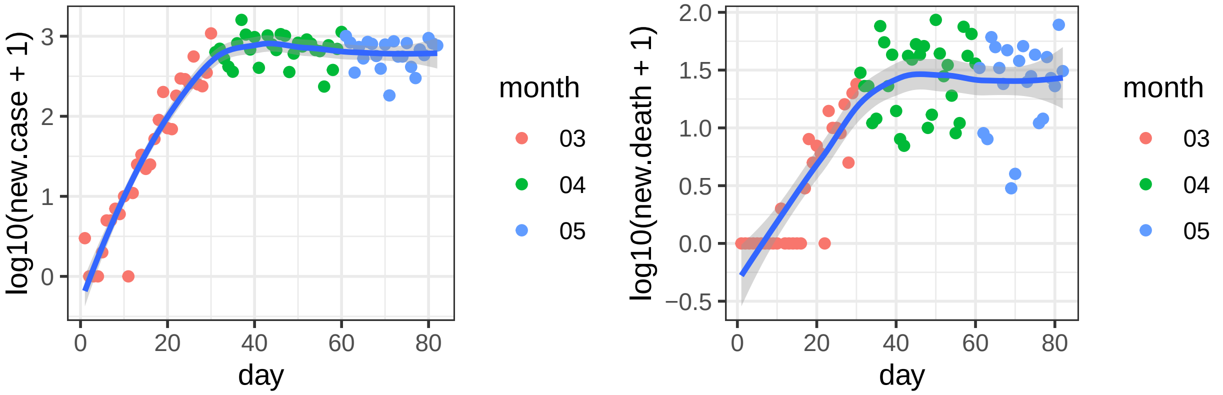
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Ohio



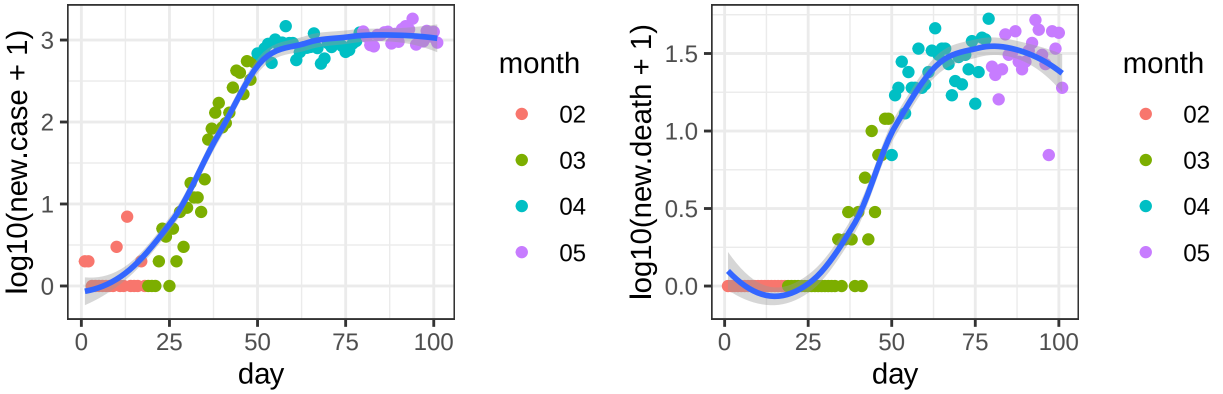
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Georgia



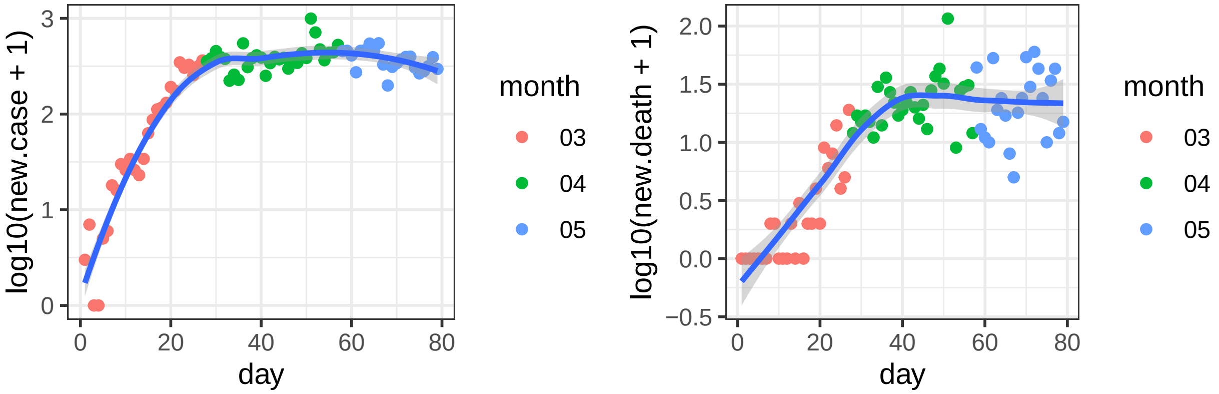
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Texas



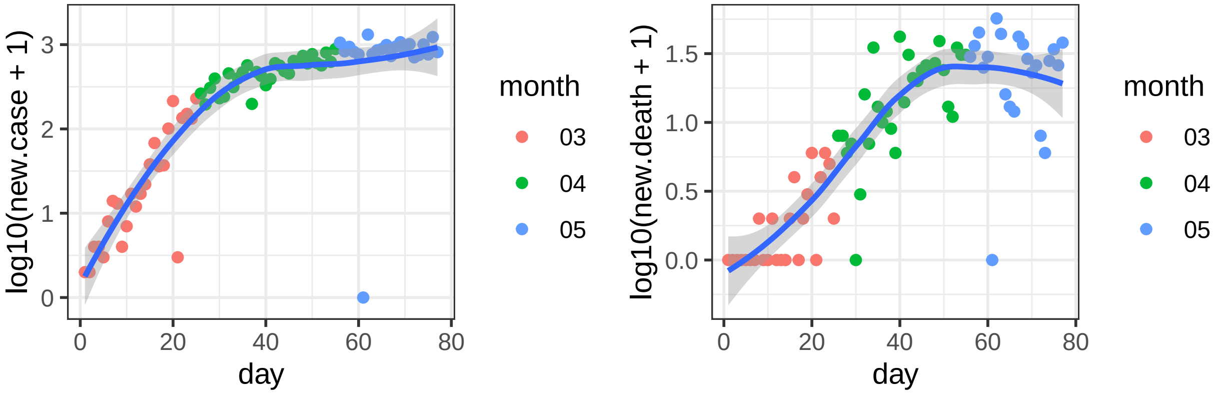
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Colorado



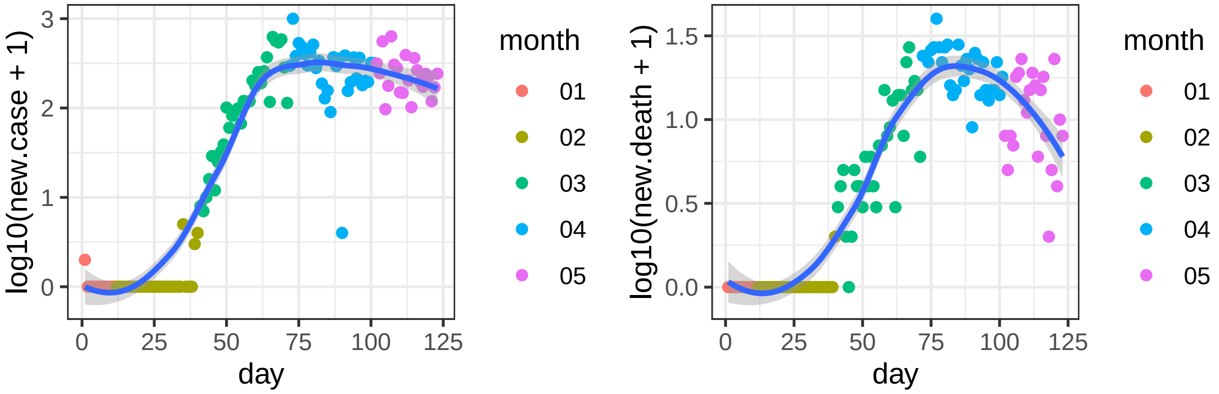
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Virginia



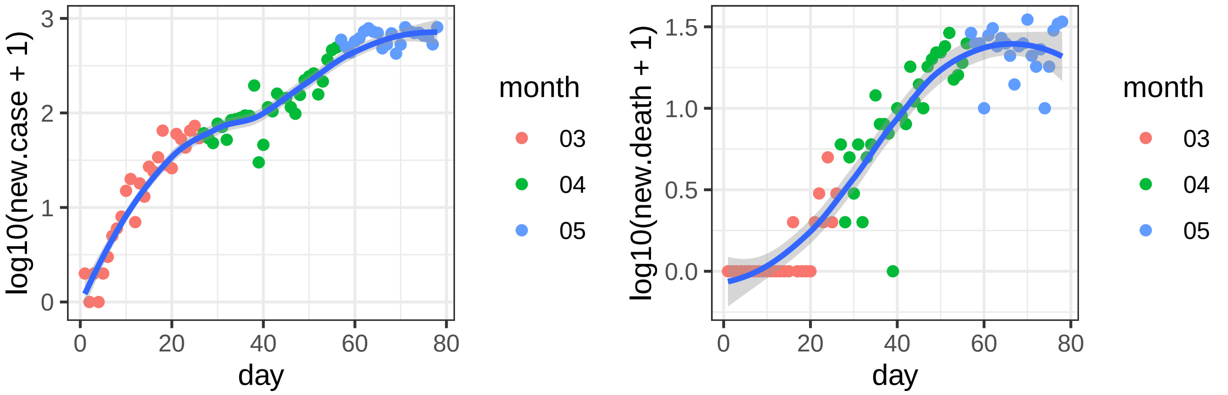
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Washington



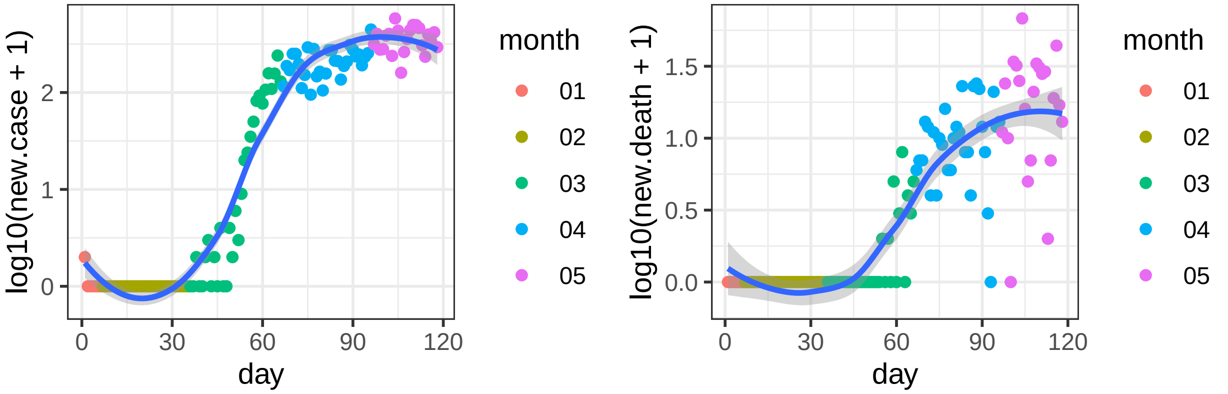
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Minnesota



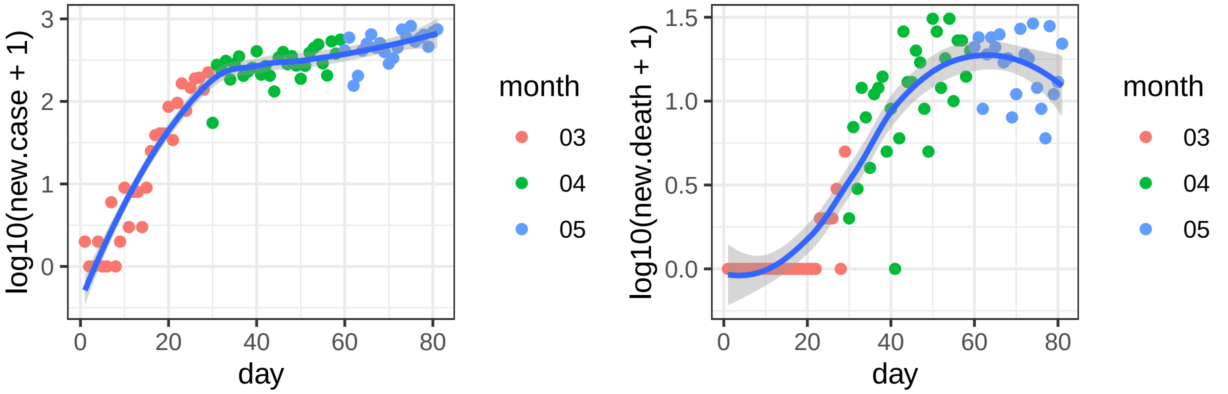
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Arizona



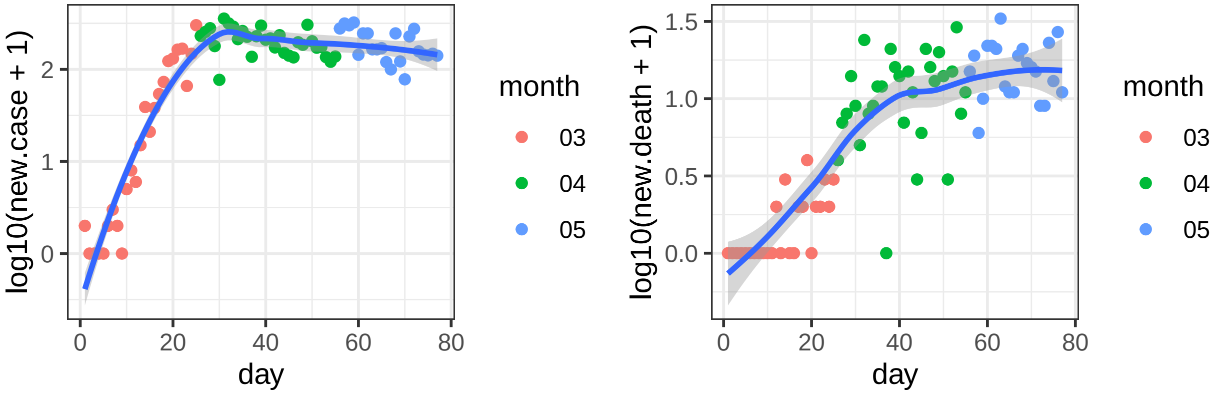
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

North Carolina



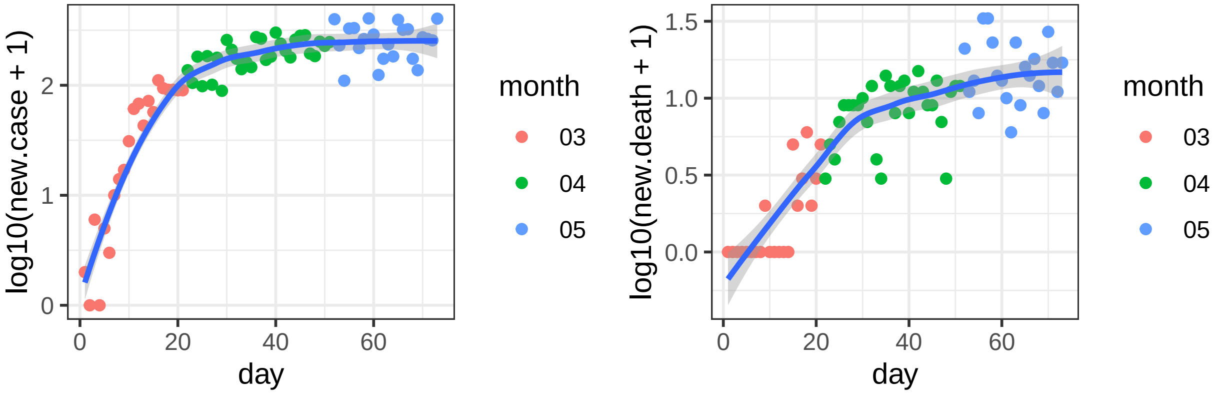
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Missouri



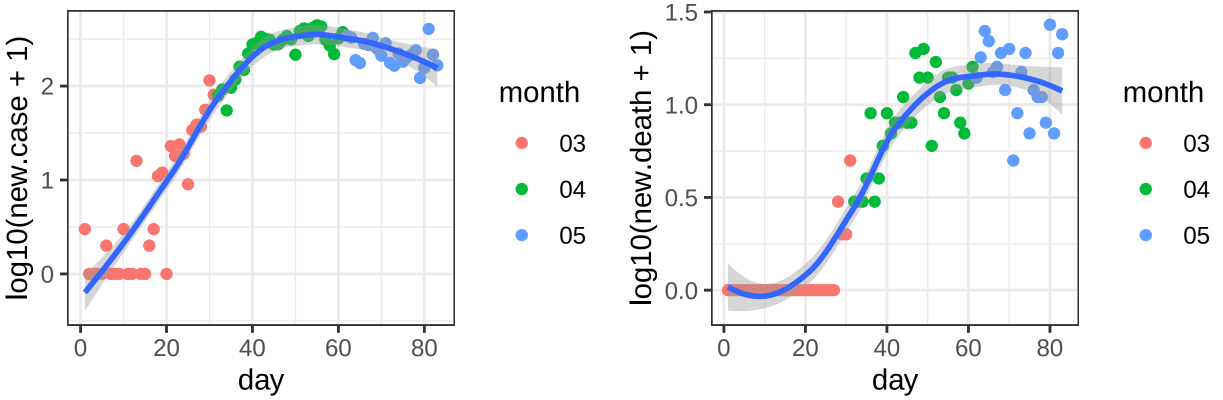
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Mississippi



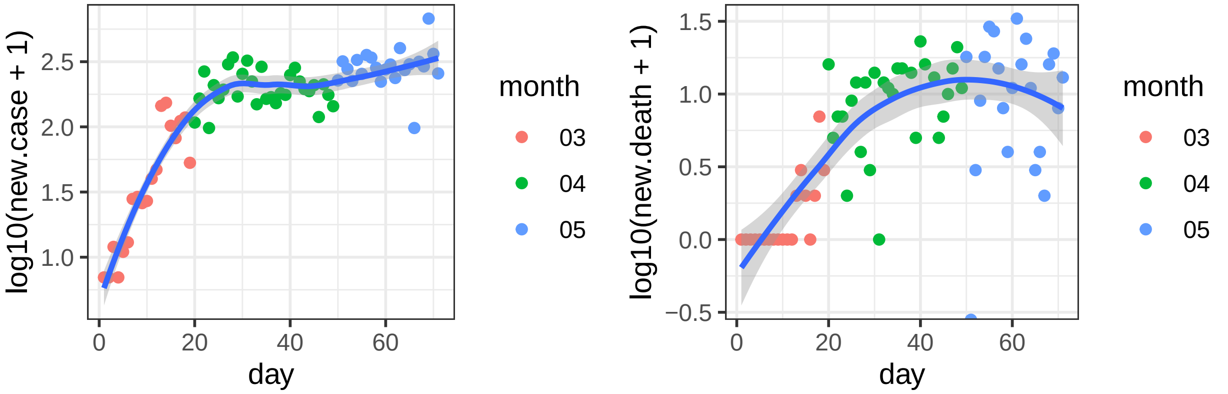
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Rhode Island



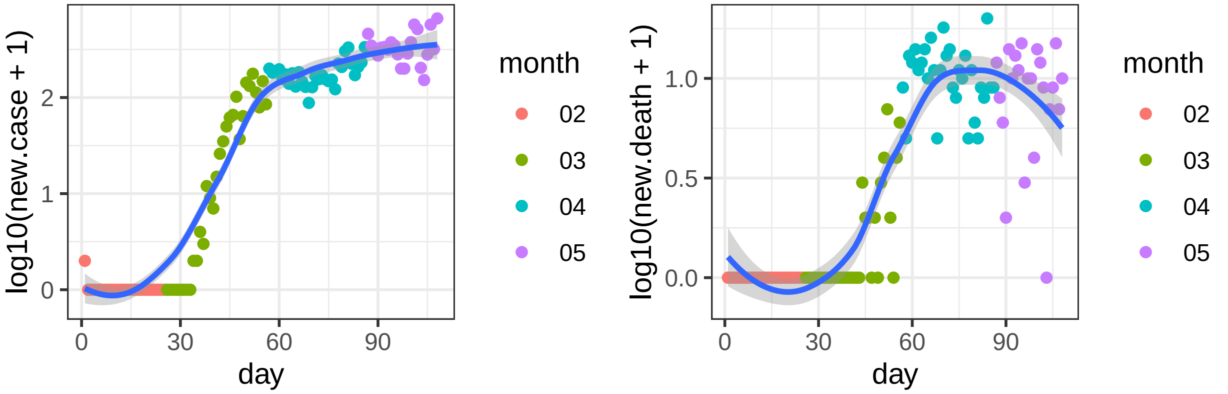
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Alabama



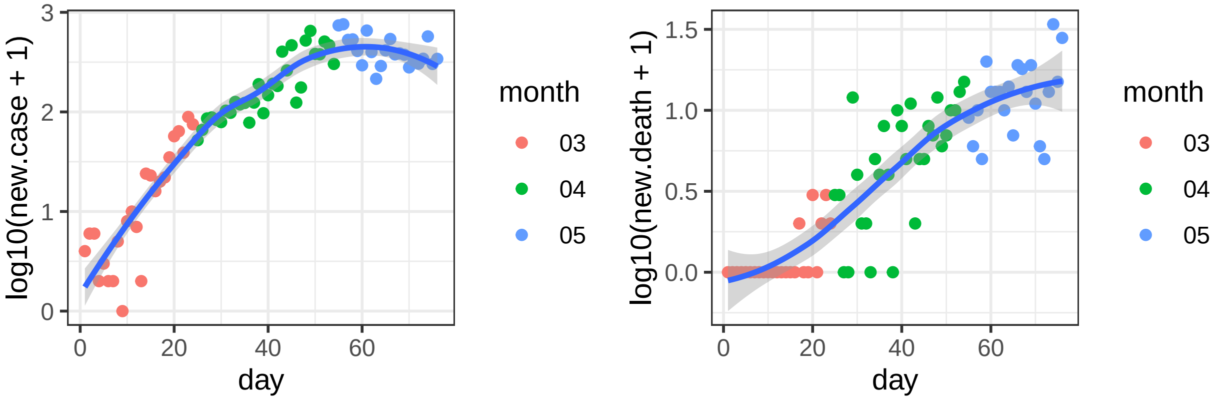
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Wisconsin



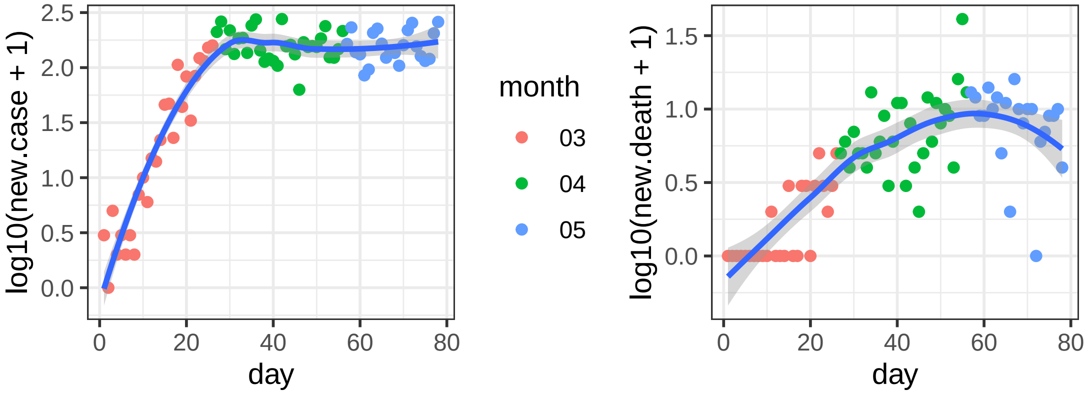
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

Iowa



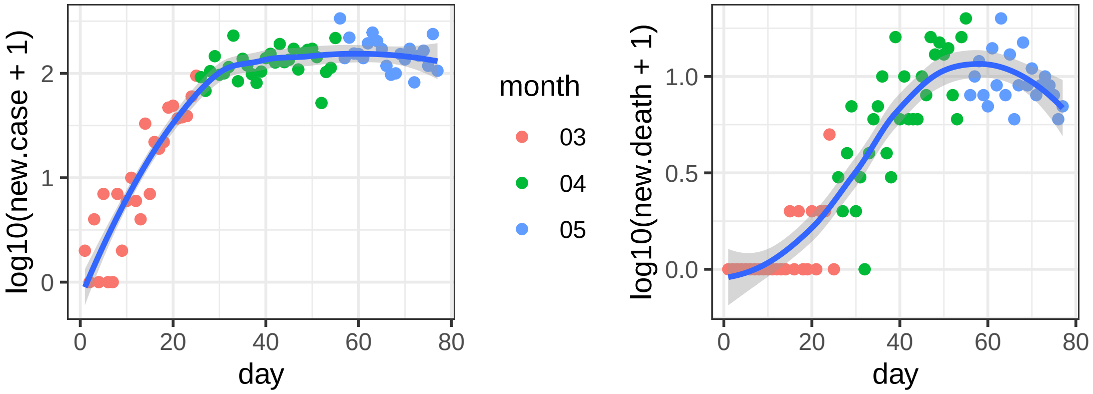
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

South Carolina



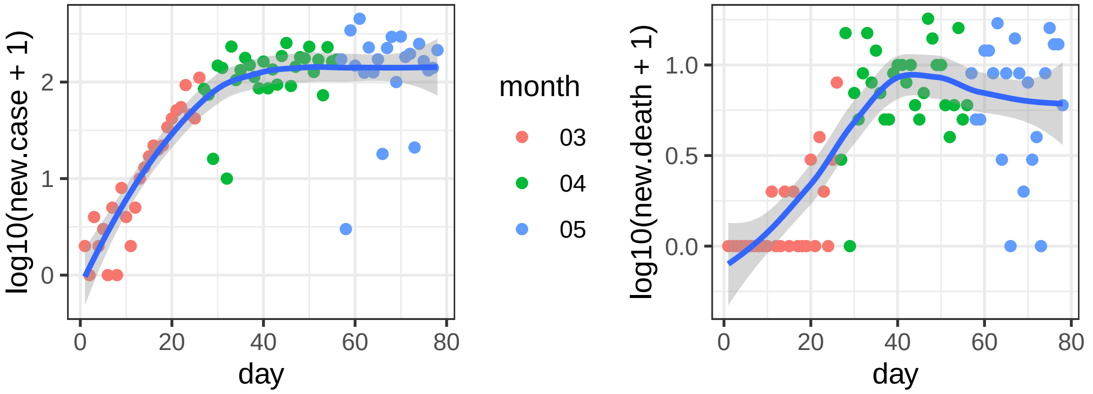
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

District of Columbia



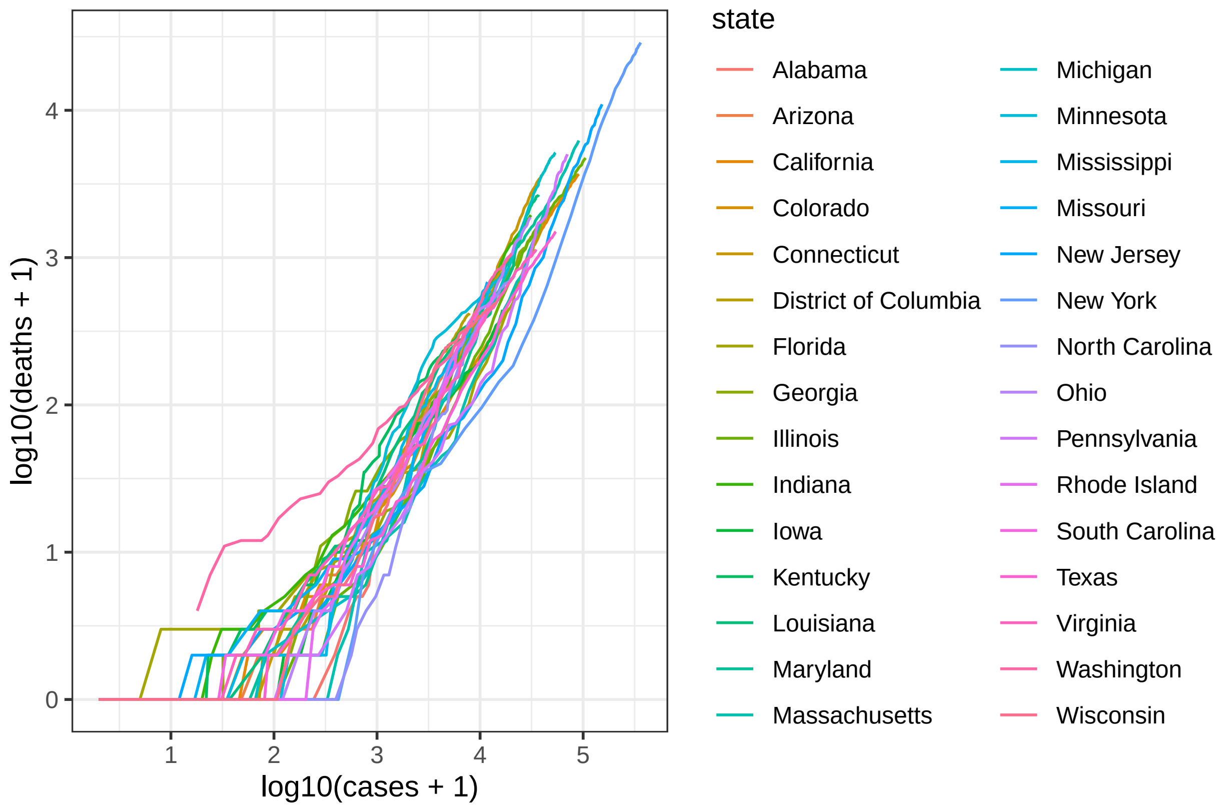
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

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

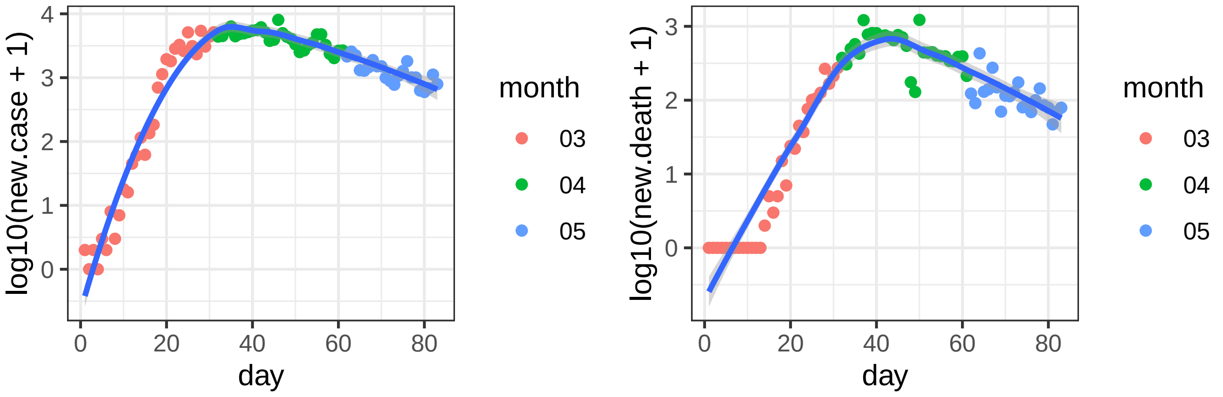
county level data

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

```
##              date        county         state  fips  cases deaths
## 166626 2020-05-22 New York City      New York    NA 201298  20569
## 165468 2020-05-22          Cook      Illinois 17031  68949   3187
## 166625 2020-05-22        Nassau      New York 36059  39608   2572
## 166149 2020-05-22         Wayne      Michigan 26163  19602   2323
## 165073 2020-05-22   Los Angeles    California  6037  43052   2049
## 166645 2020-05-22       Suffolk      New York 36103  38672   1863
## 166551 2020-05-22         Essex    New Jersey 34013  17014   1585
## 166546 2020-05-22        Bergen    New Jersey 34003  17653   1515
## 166063 2020-05-22     Middlesex Massachusetts 25017  20085   1496
## 166653 2020-05-22   Westchester      New York 36119  32766   1444
## 167043 2020-05-22  Philadelphia  Pennsylvania 42101  21009   1221
## 165173 2020-05-22     Fairfield   Connecticut  9001  14889   1195
## 165174 2020-05-22      Hartford   Connecticut  9003   9463   1155
## 166553 2020-05-22        Hudson    New Jersey 34017  17897   1134
## 166564 2020-05-22         Union    New Jersey 34039  15191   1018
## 166129 2020-05-22       Oakland      Michigan 26125   8131    944
## 166556 2020-05-22     Middlesex    New Jersey 34023  15165    935
## 165177 2020-05-22     New Haven   Connecticut  9009  10756    888
## 166560 2020-05-22       Passaic    New Jersey 34031  15604    881
## 166059 2020-05-22         Essex Massachusetts 25009  13221    842
## 166067 2020-05-22       Suffolk Massachusetts 25025  17180    818
## 166116 2020-05-22        Macomb      Michigan 26099   6445    776
## 166065 2020-05-22       Norfolk Massachusetts 25021    7724     771
## 166559 2020-05-22         Ocean    New Jersey 34029    8285     678
## 166069 2020-05-22     Worcester Massachusetts 25027   10101     652
## 167038 2020-05-22    Montgomery  Pennsylvania 42091    6366     619
## 165229 2020-05-22    Miami-Dade       Florida 12086   16521     614
## 166558 2020-05-22        Morris    New Jersey 34027    6171     587
## 165601 2020-05-22        Marion       Indiana 18097    9024     564
## 167670 2020-05-22          King    Washington 53033    7699     544
```

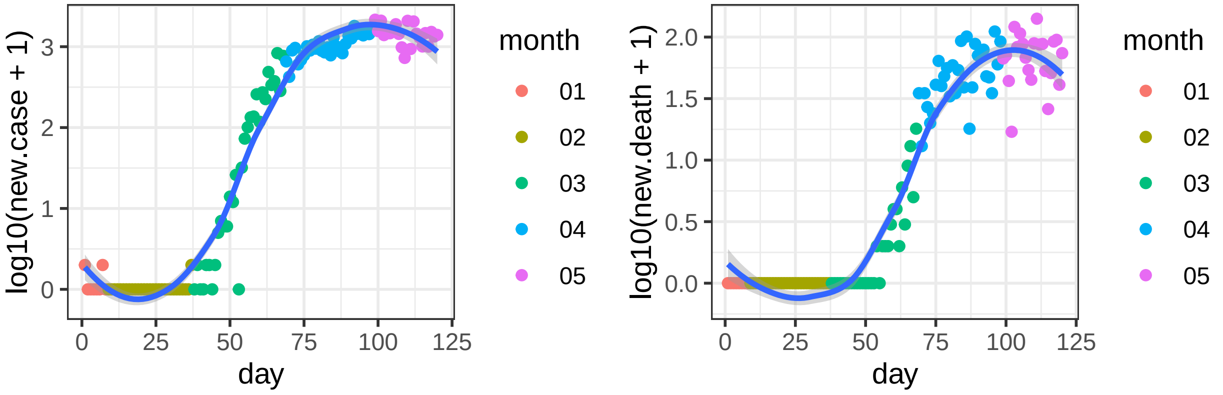
For these 30 counties, I check the number of new cases and the number of new deaths.

New York City_New York



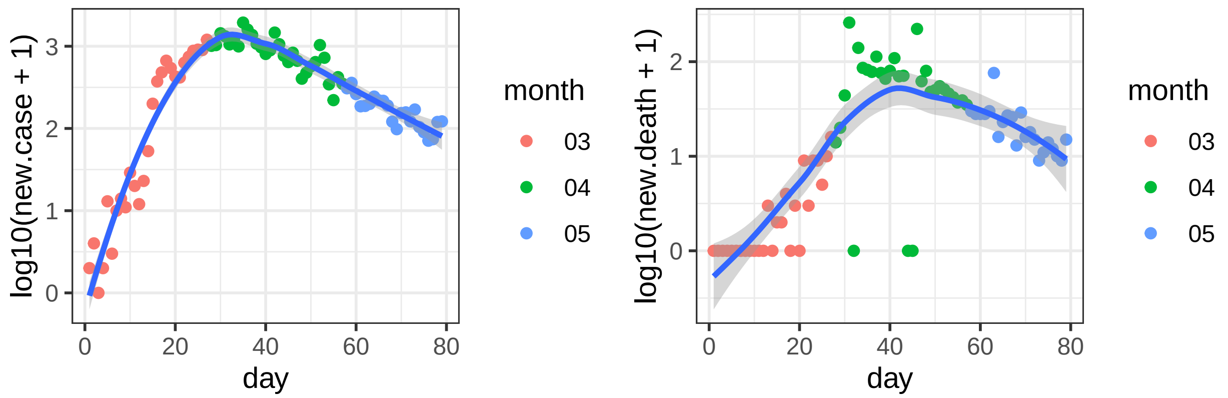
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



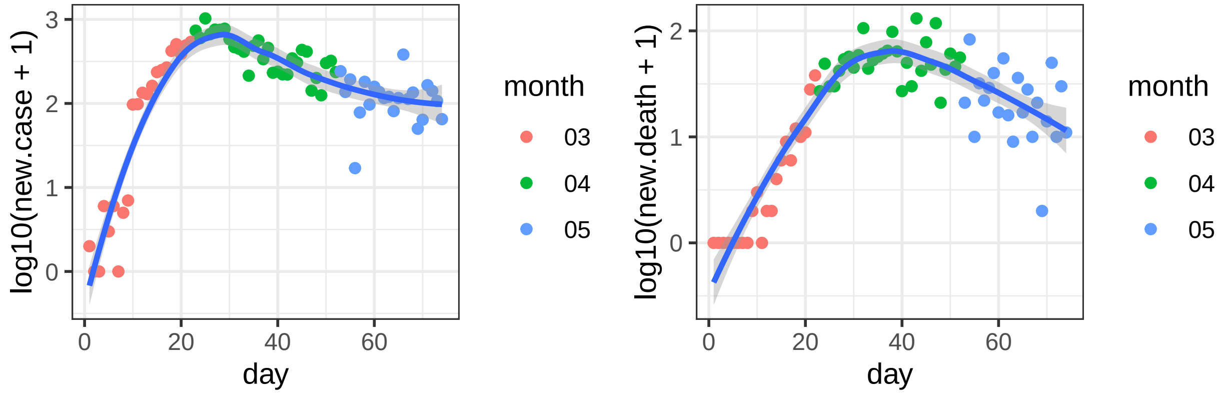
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Nassau_New York



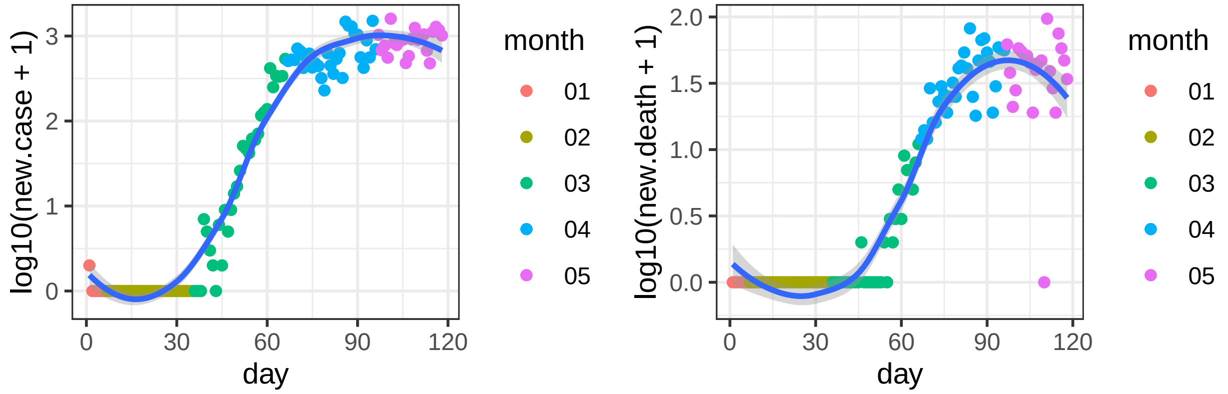
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Wayne_Michigan



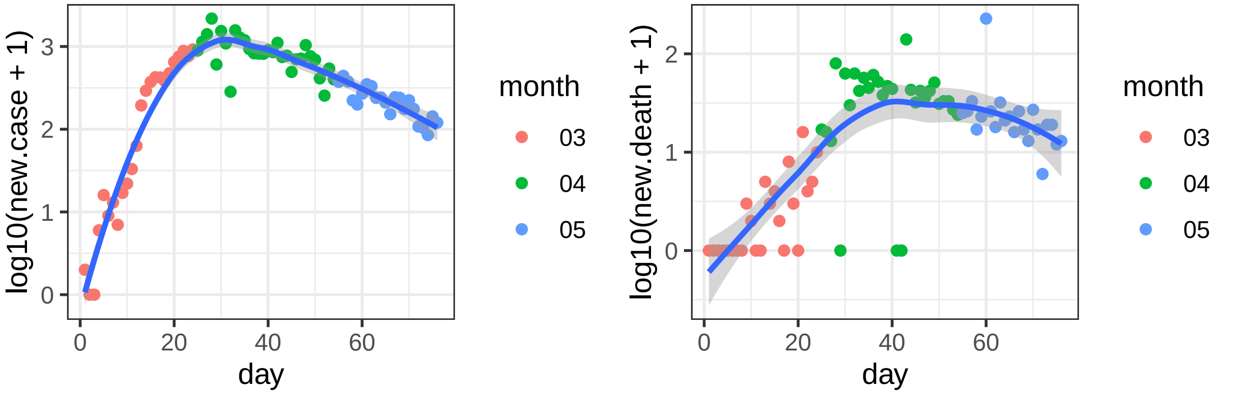
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Los Angeles_California



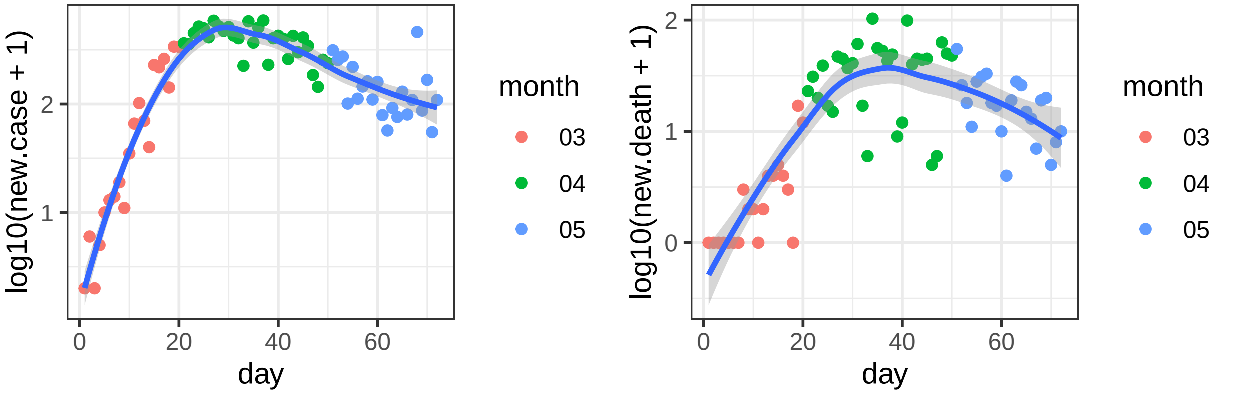
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Suffolk_New York



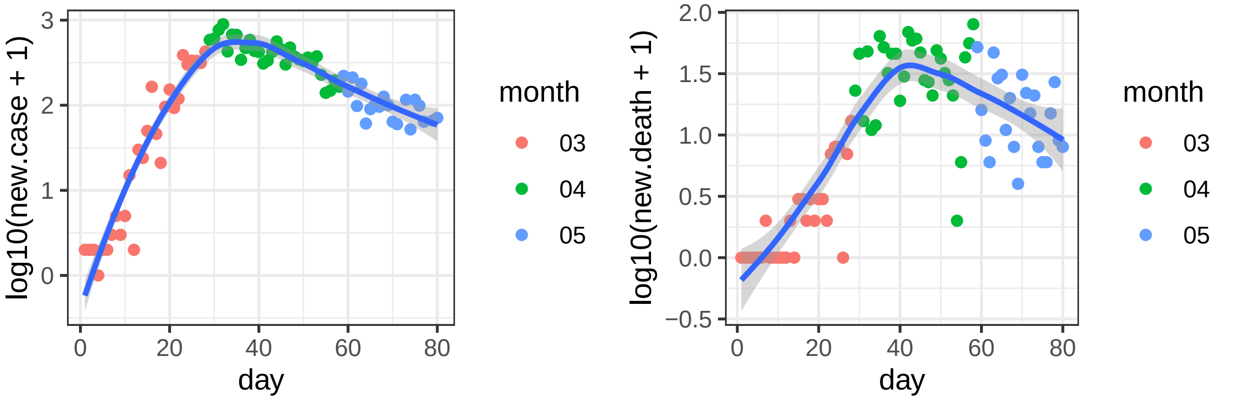
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Essex_New Jersey



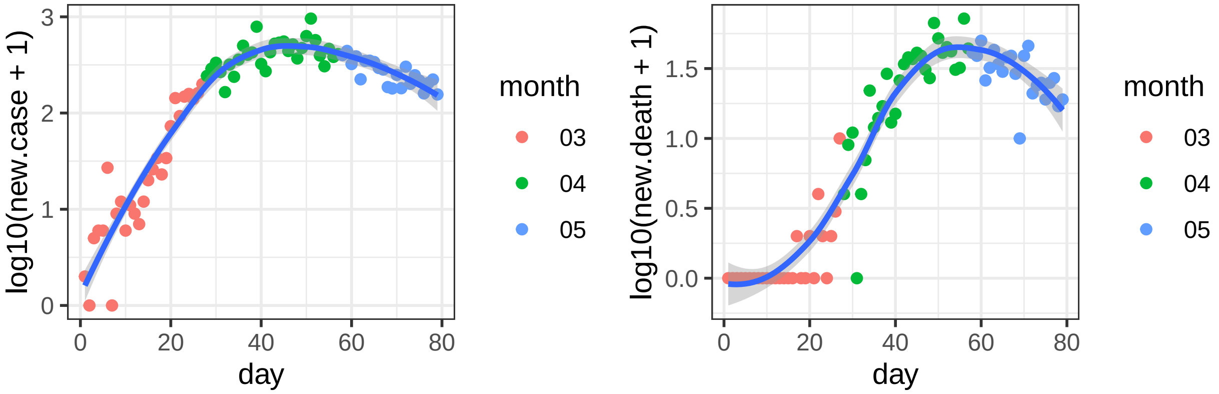
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Bergen_New Jersey



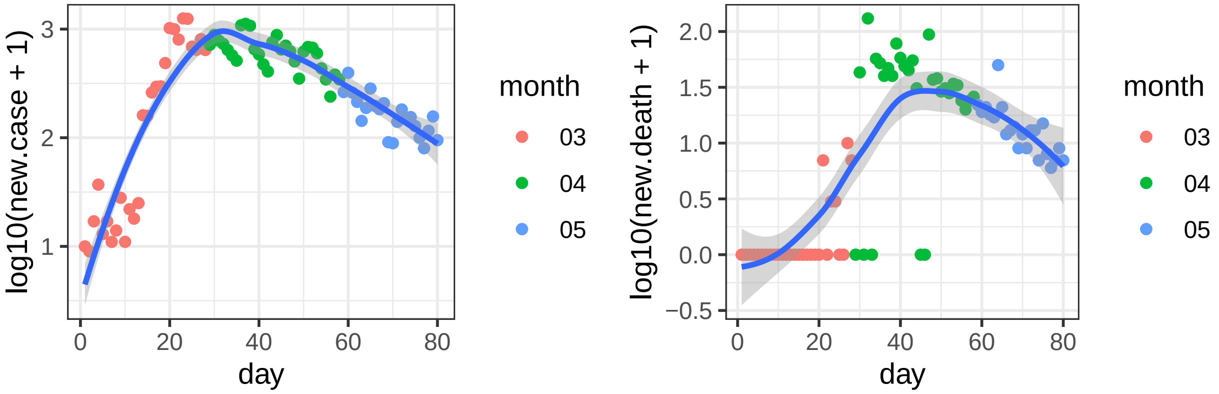
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Middlesex_Massachusetts



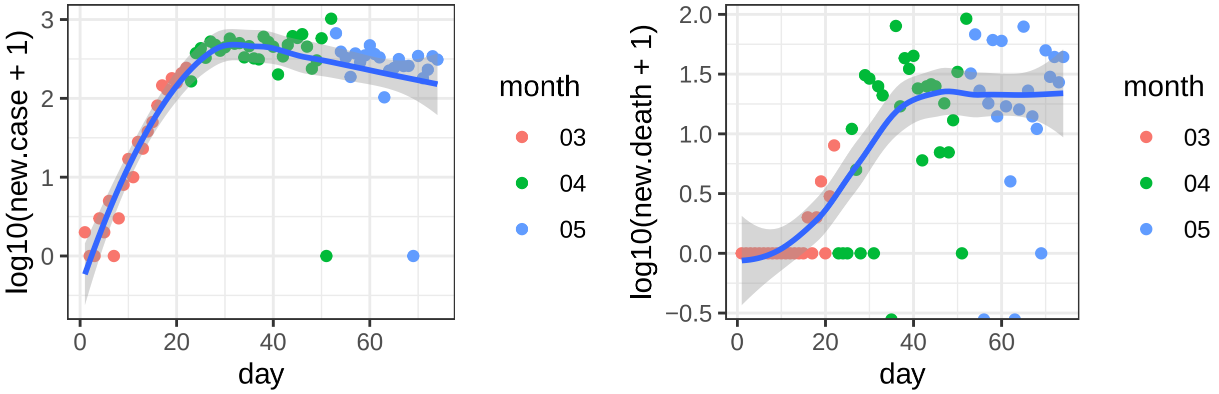
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Westchester_New York



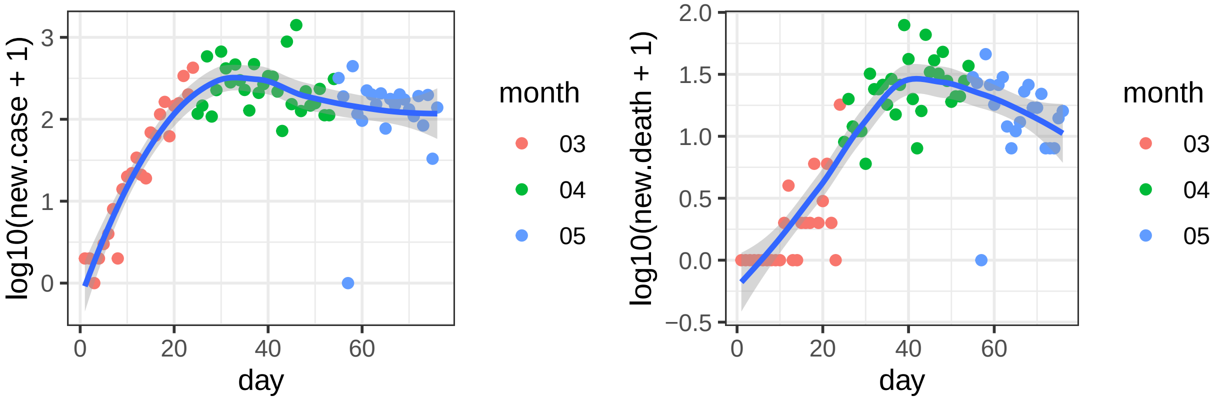
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Philadelphia_Pennsylvania



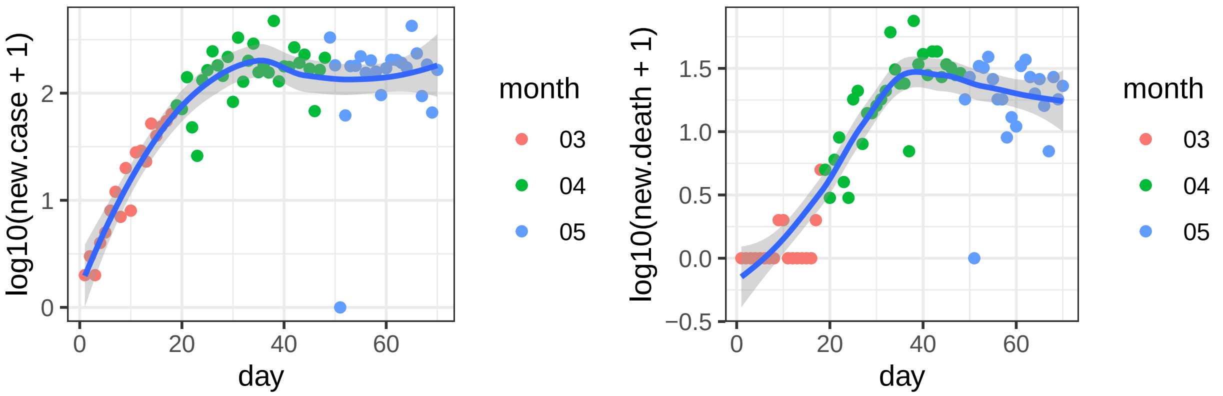
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Fairfield_Connecticut



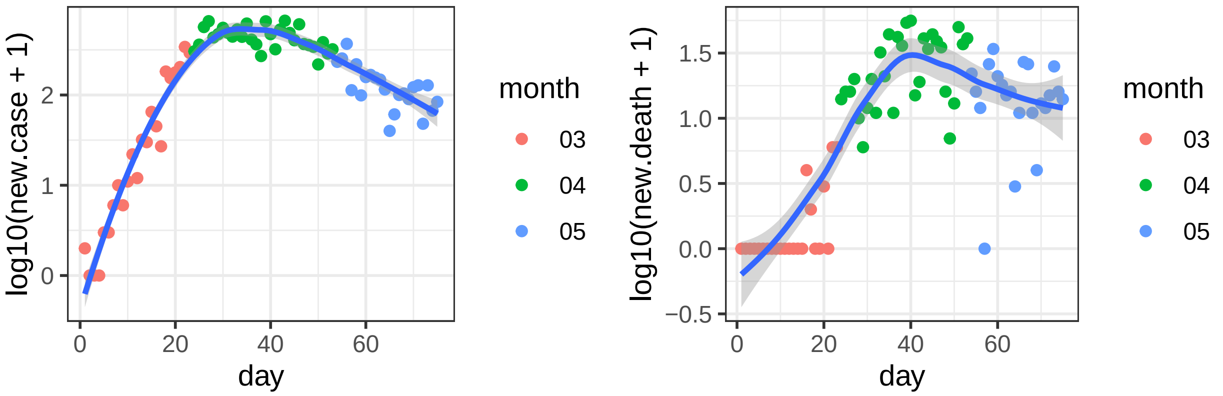
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Hartford_Connecticut



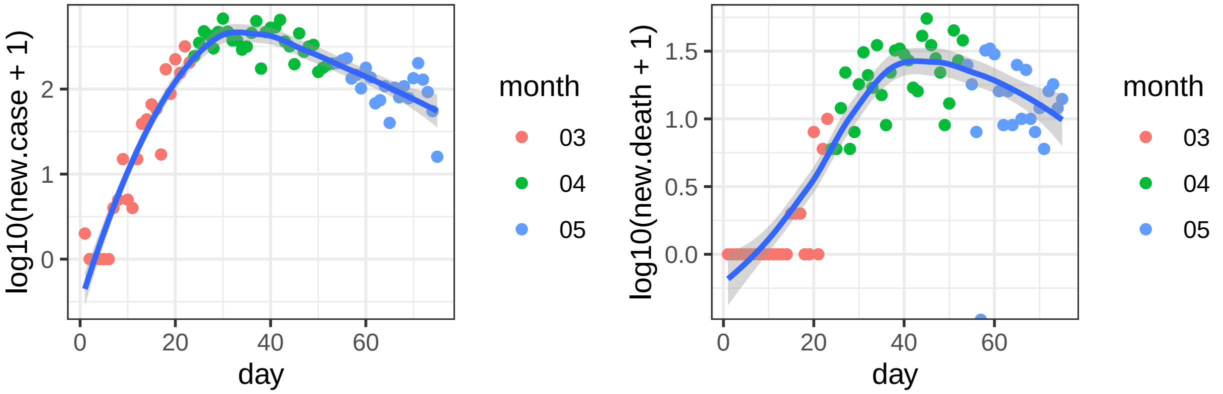
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Hudson_New Jersey



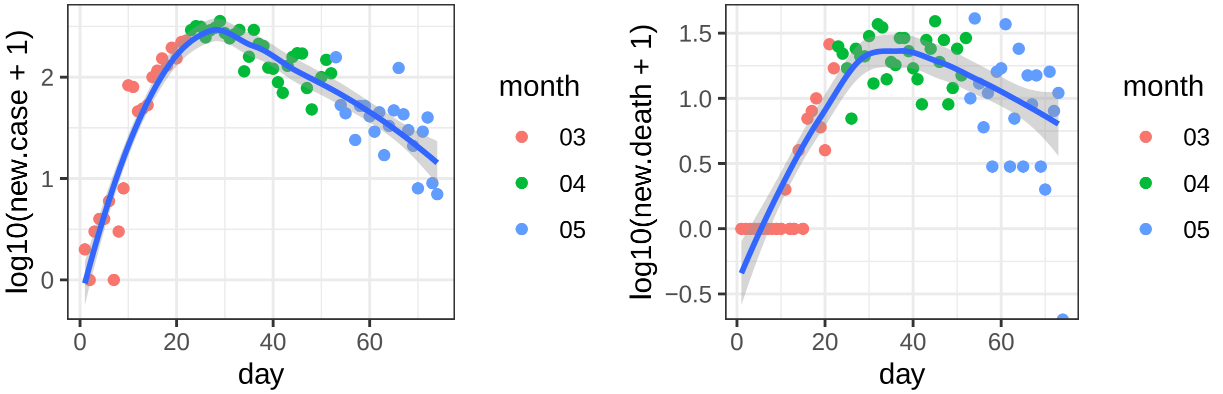
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Union_New Jersey



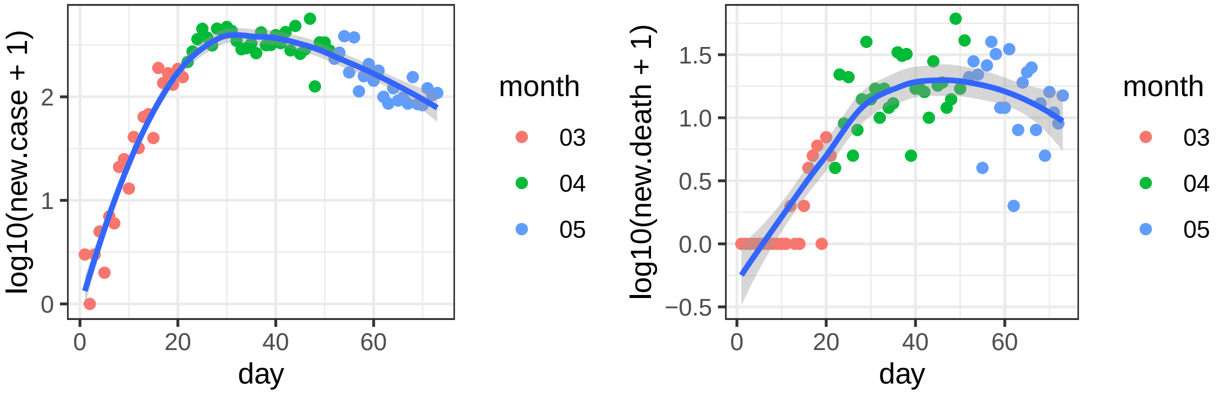
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Oakland_Michigan



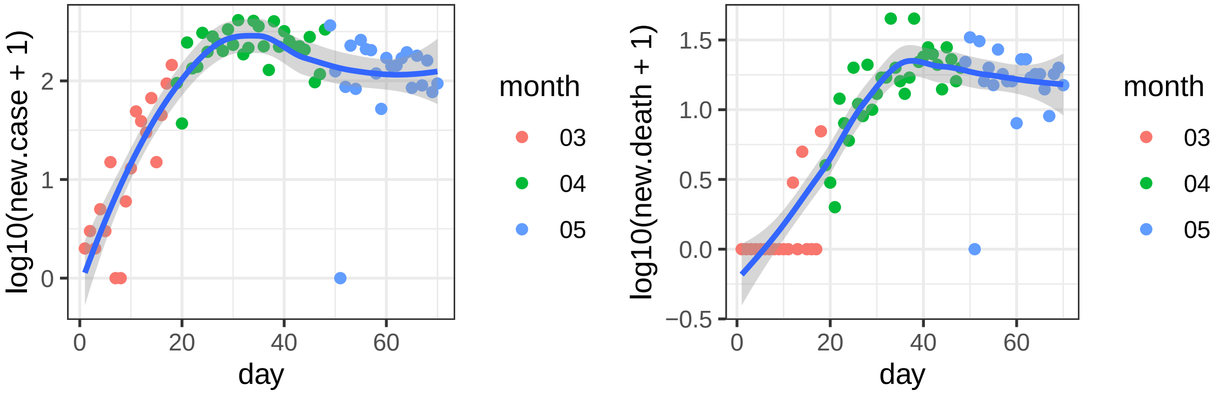
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Middlesex_New Jersey



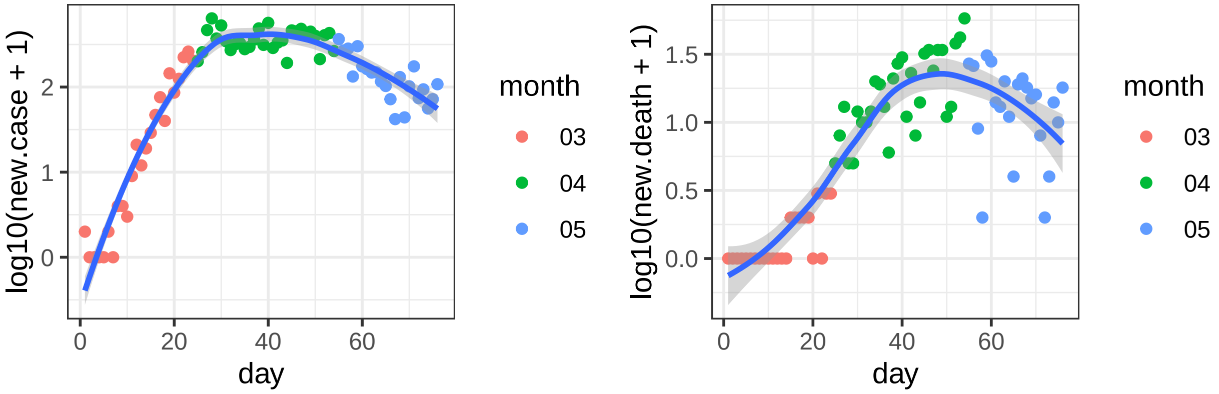
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

New Haven_Connecticut



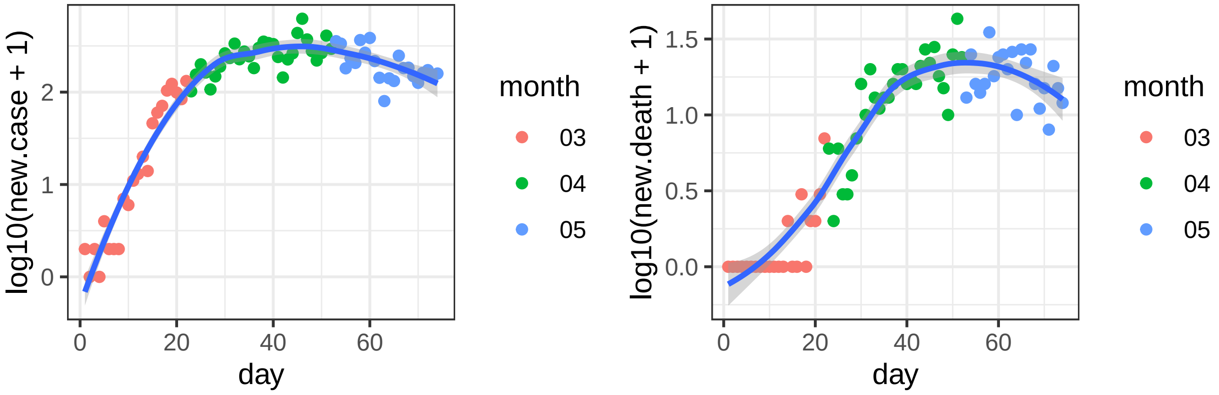
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Passaic_New Jersey



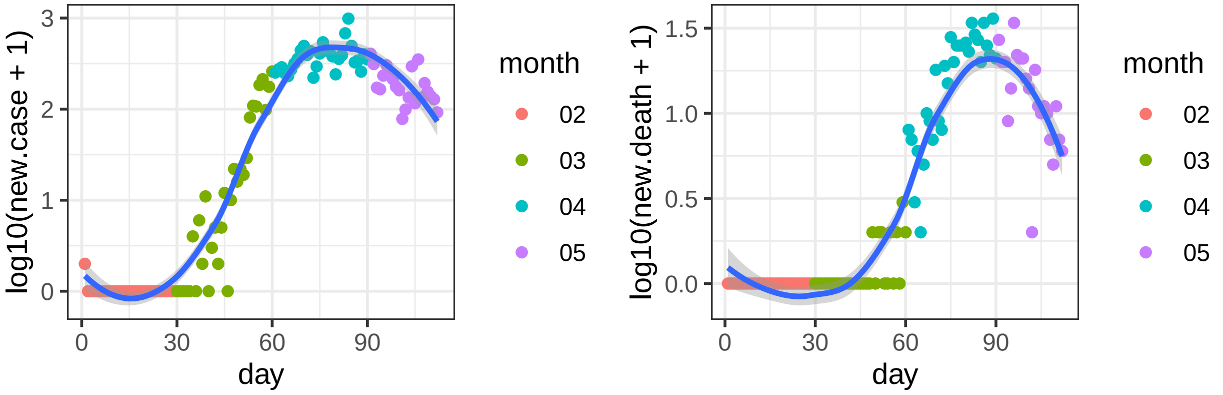
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Essex_Massachusetts



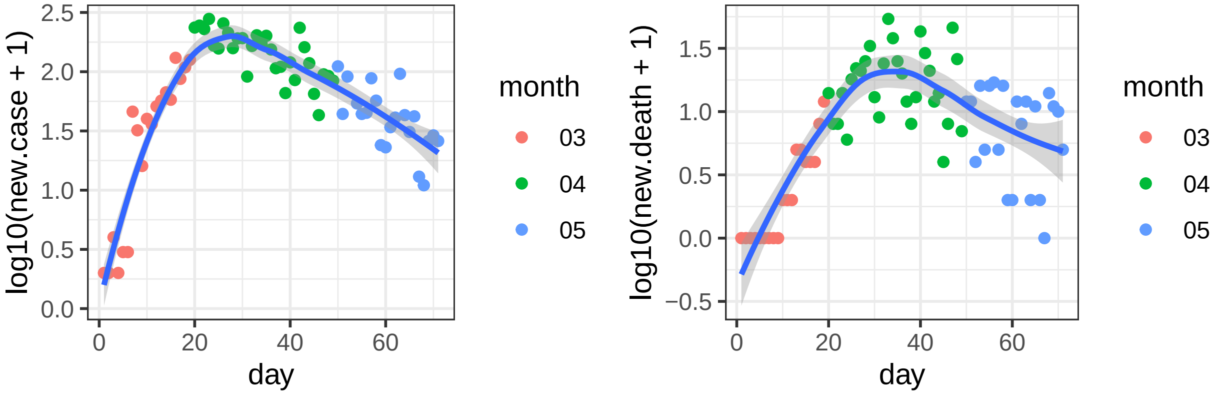
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Suffolk_Massachusetts



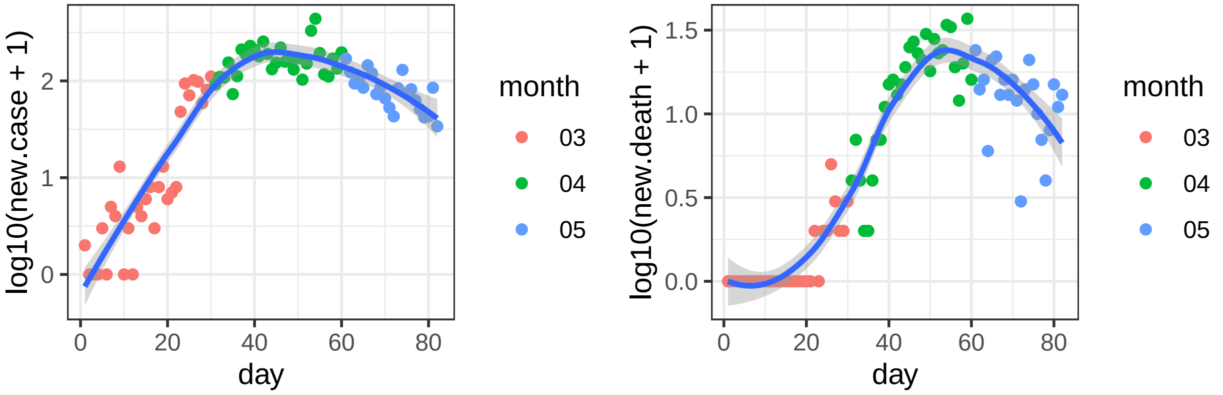
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Macomb_Michigan



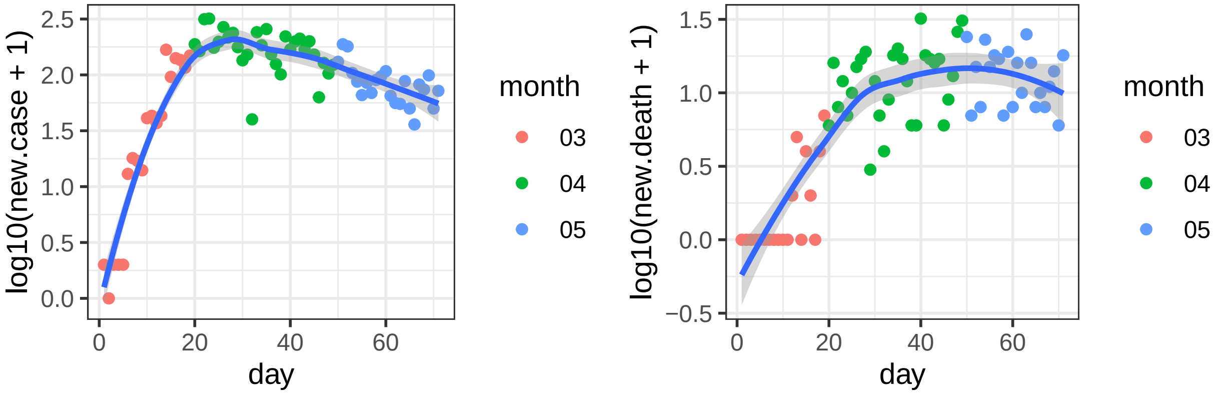
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Norfolk_Massachusetts



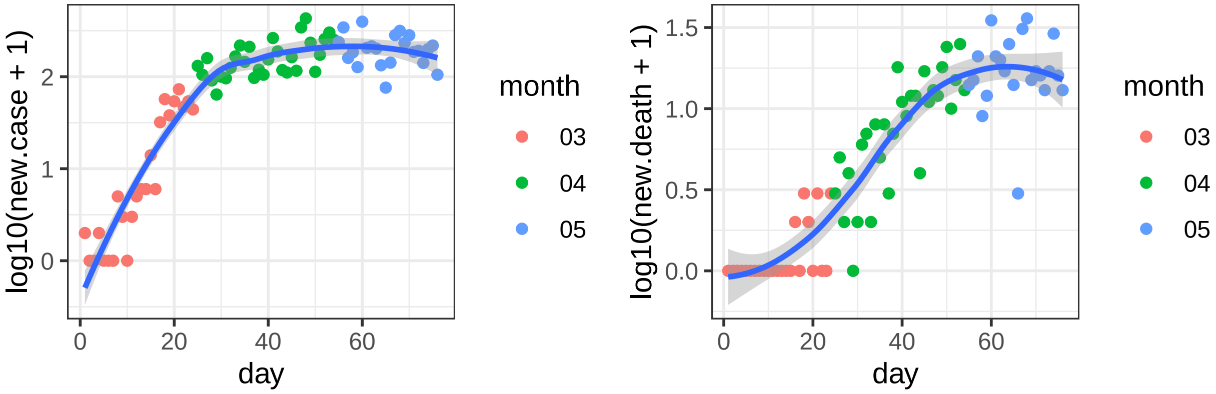
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Ocean_New Jersey



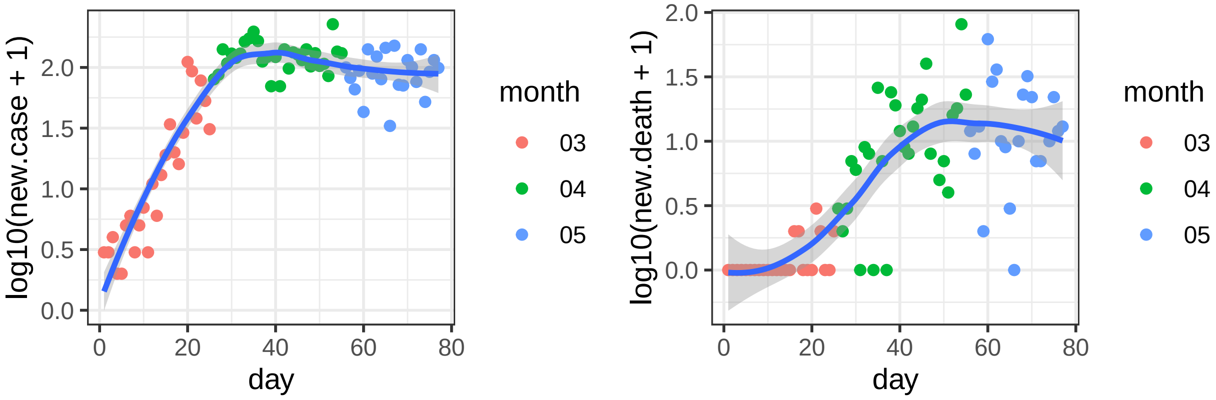
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Worcester_Massachusetts

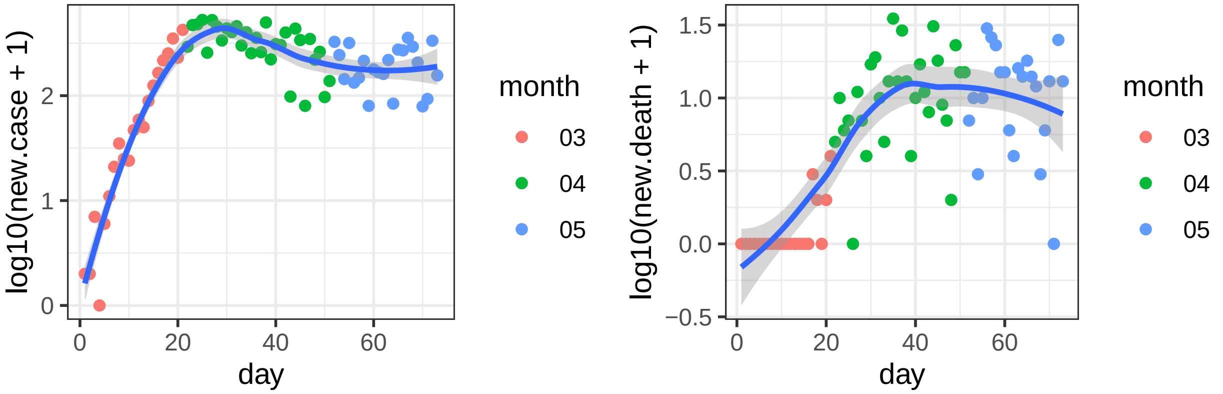


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

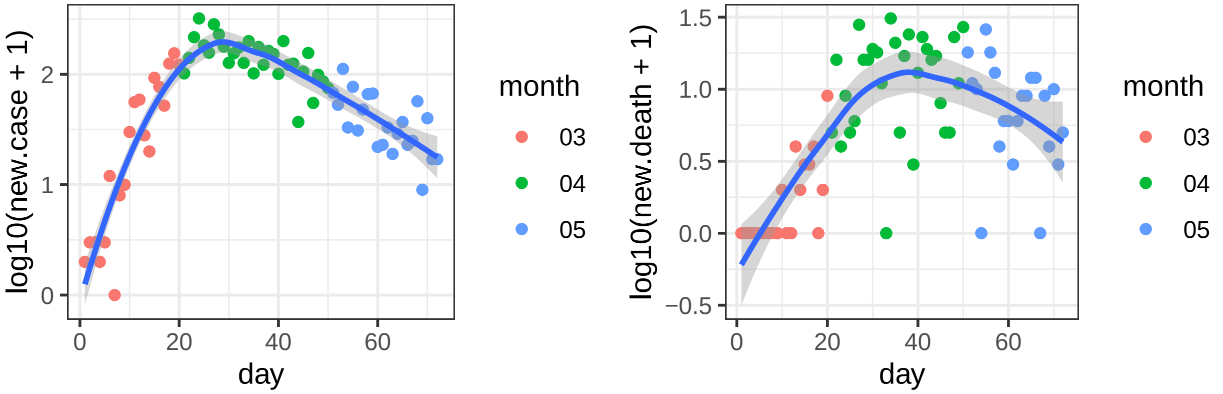
Montgomery_Pennsylvania



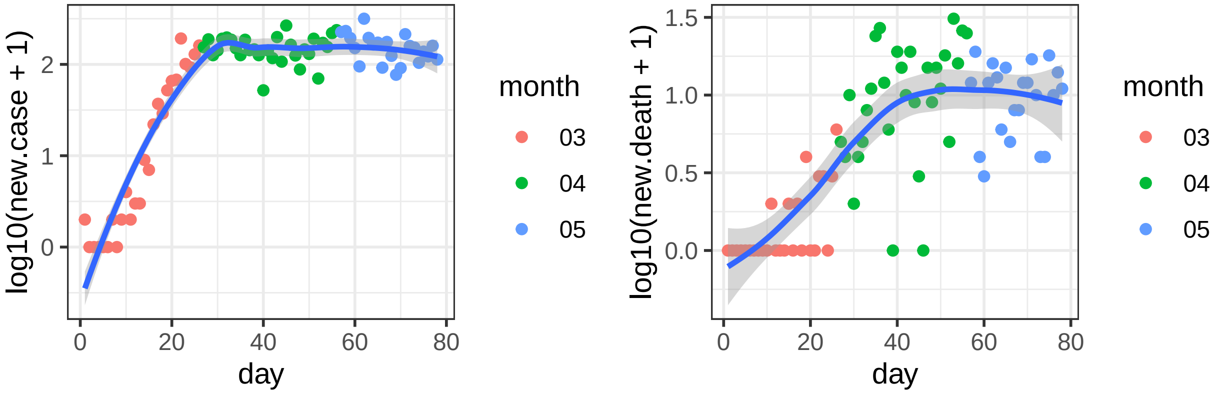
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07


Miami−Dade_Florida

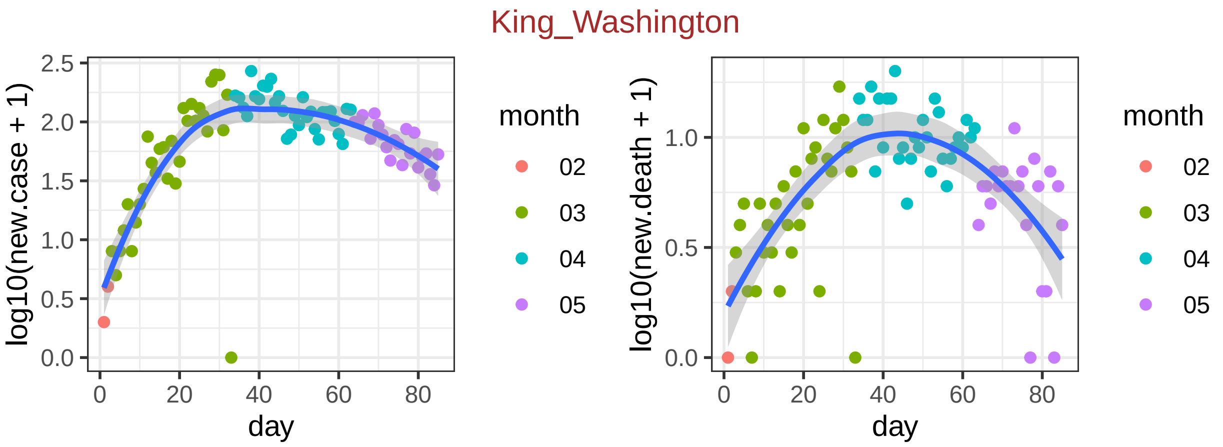
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11


Morris_New Jersey

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

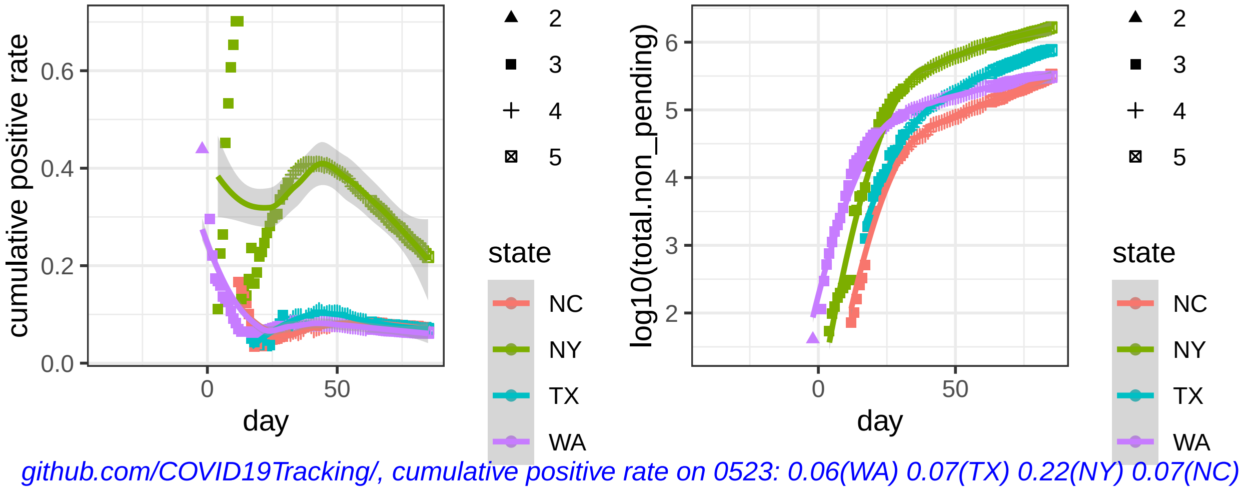


data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

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 0523: 0.06(WA) 0.07(TX) 0.22(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 10.15.4
```

```
##
## 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_3.2.1
##
## 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    evaluate_0.14    lifecycle_0.1.0  tibble_2.1.3
##  [9] gtable_0.3.0     pkgconfig_2.0.3  rlang_0.4.4      yaml_2.2.1
## [13] xfun_0.12        gridExtra_2.3    withr_2.1.2      dplyr_0.8.4
## [17] stringr_1.4.0    knitr_1.28       grid_3.6.2       tidyselect_1.0.0
## [21] cowplot_1.0.0    glue_1.3.1       R6_2.4.1         rmarkdown_2.1
## [25] purrr_0.3.3      farver_2.0.3     scales_1.1.0     htmltools_0.4.0
## [29] assertthat_0.2.1 colorspace_1.4-1 ggsignif_0.6.0   labeling_0.3
## [33] stringi_1.4.5    lazyeval_0.2.2   munsell_0.5.0    crayon_1.3.4
```