Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-05-08

Contents

Introduction

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

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

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

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

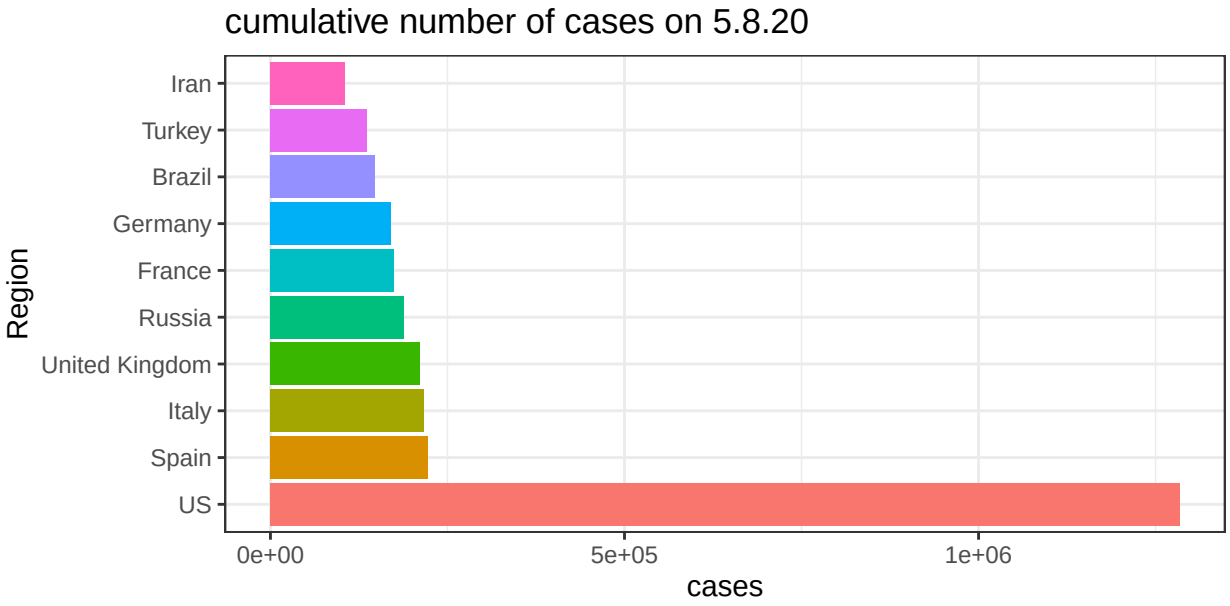
JHU

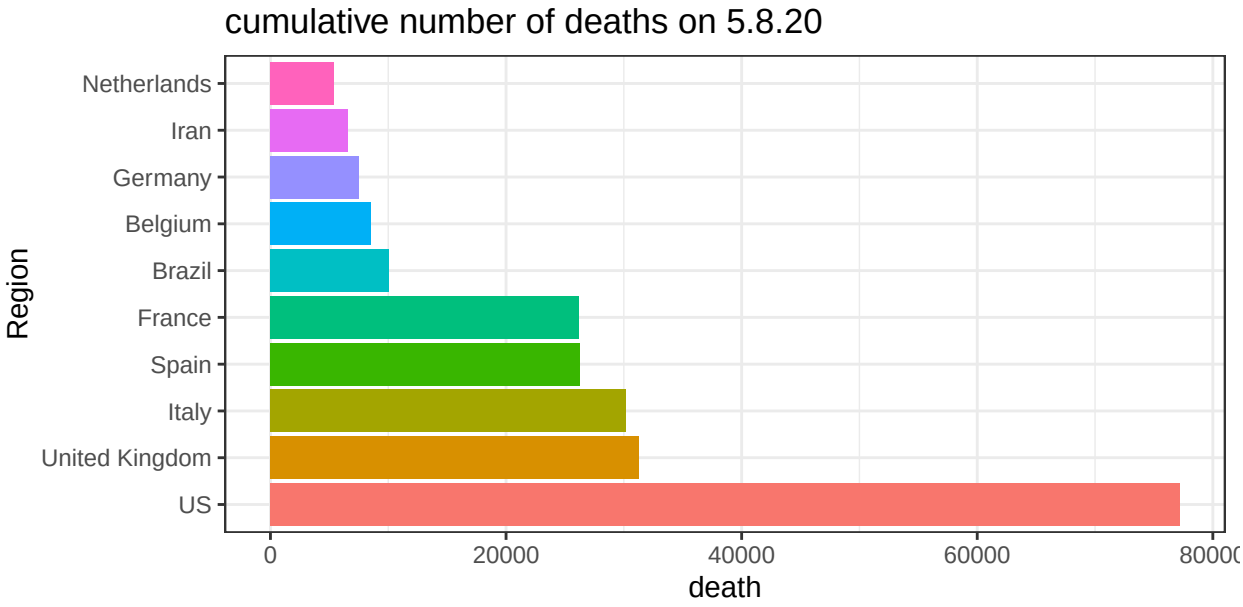
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

time series data

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

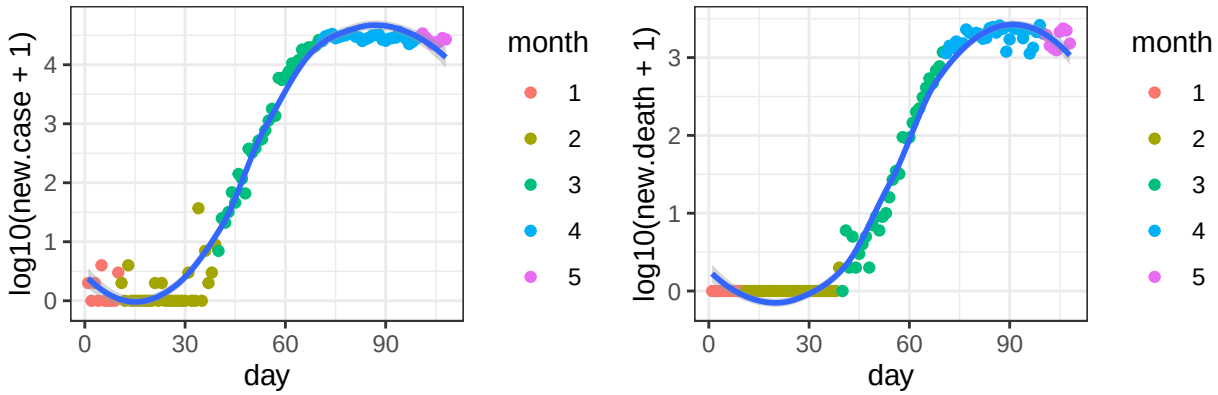
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





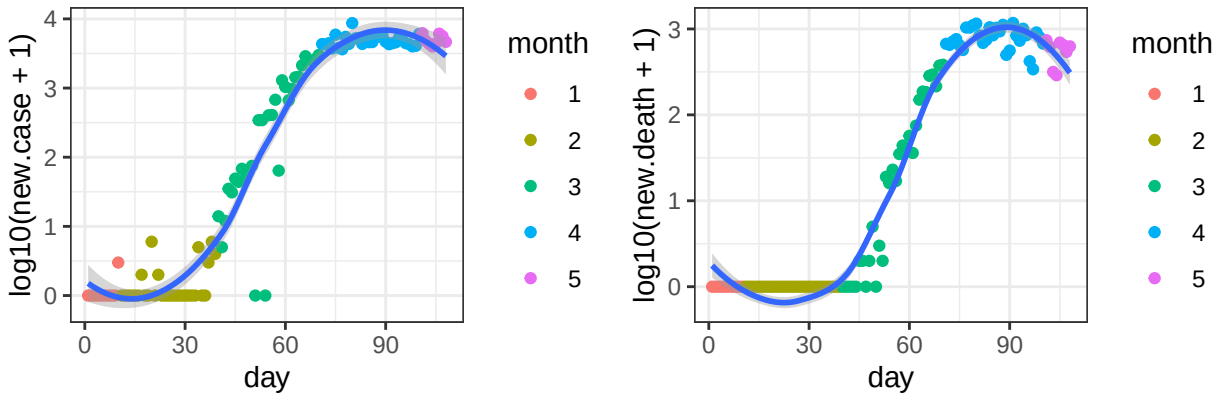
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



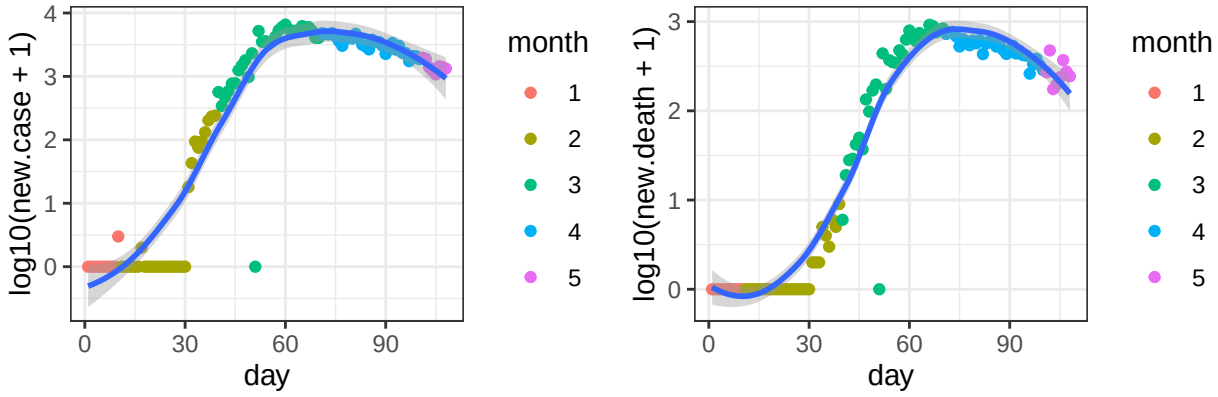
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



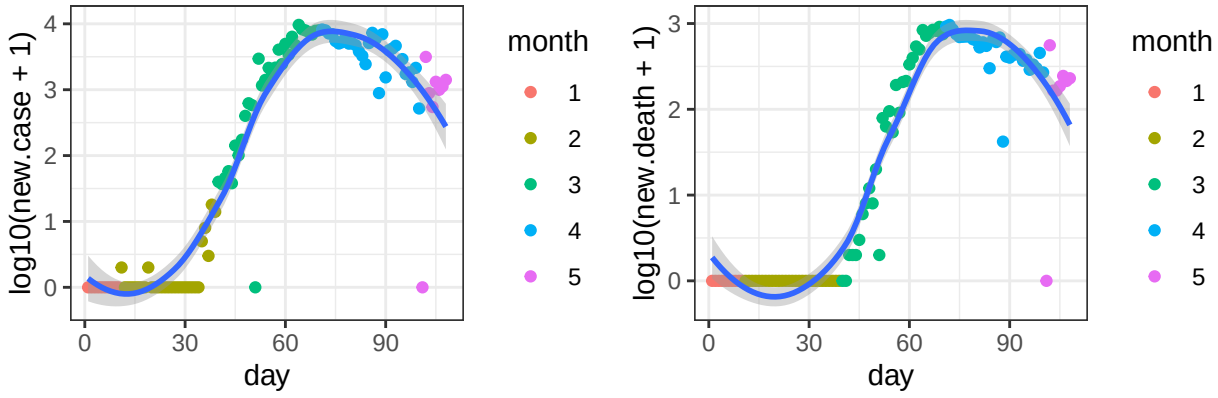
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



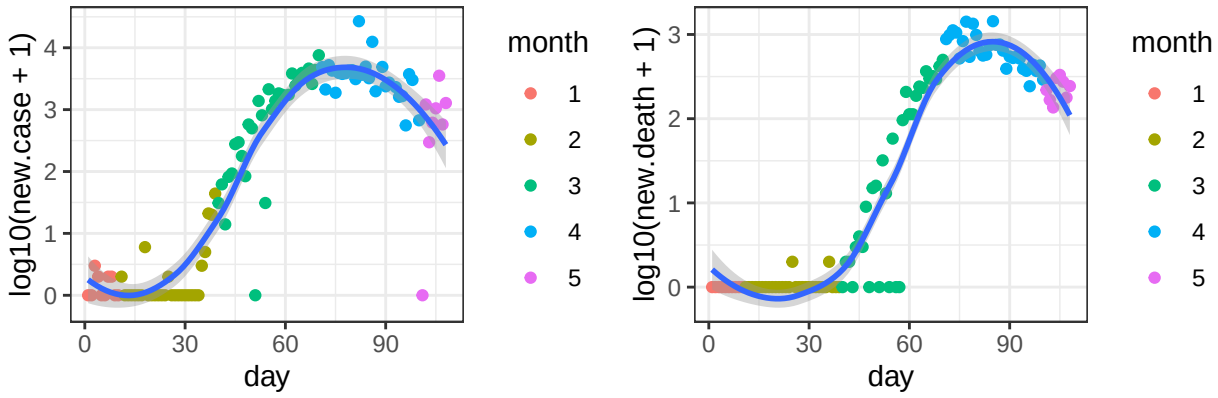
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain



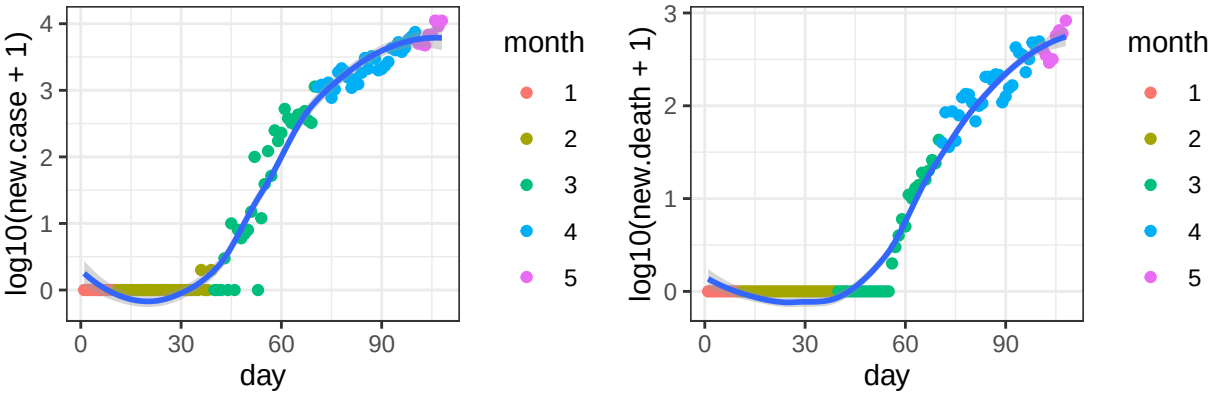
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



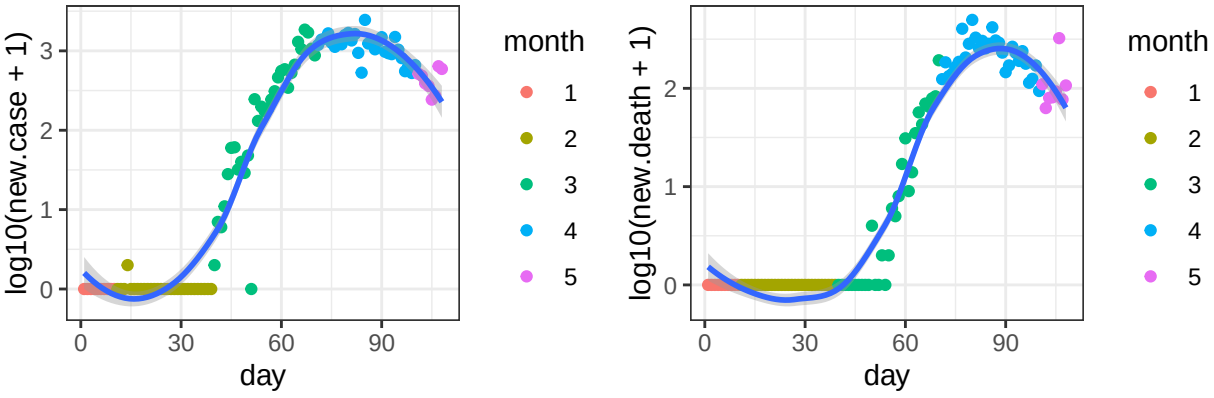
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Brazil



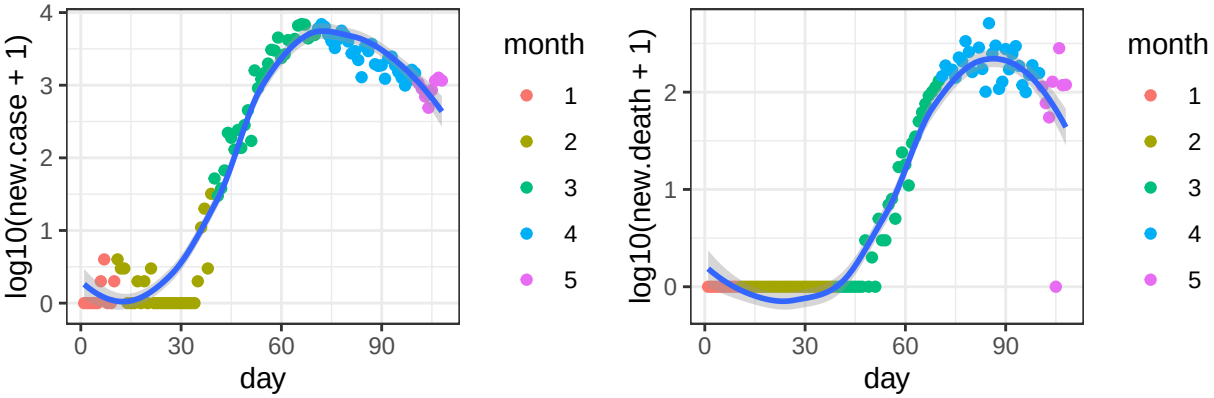
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Belgium



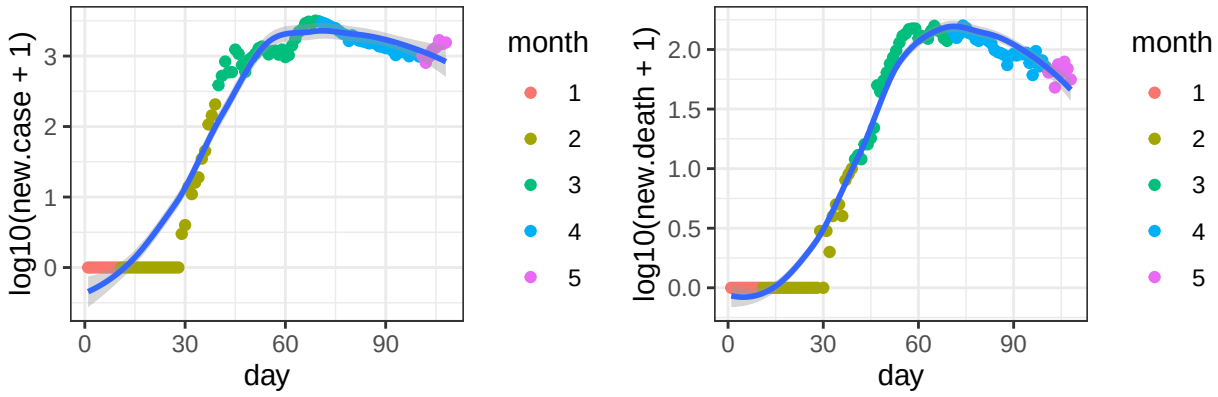
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Germany



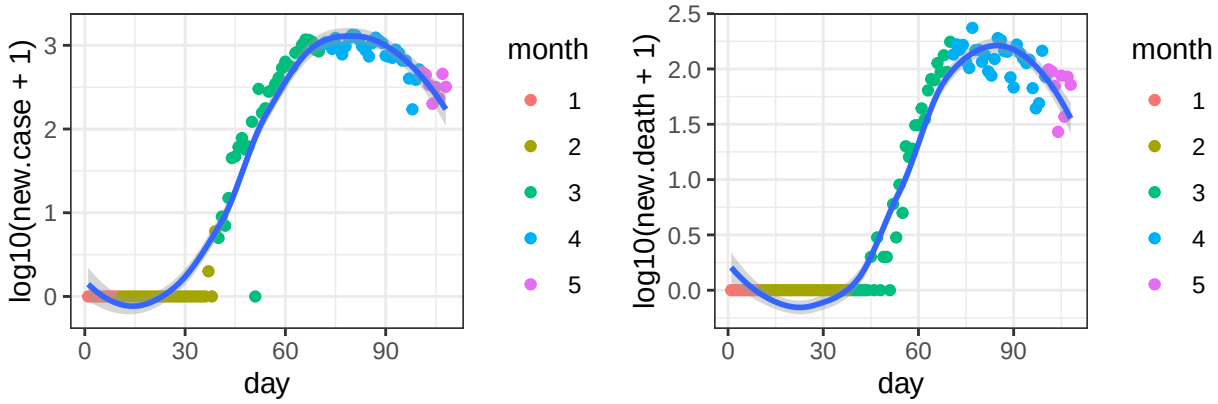
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Iran



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

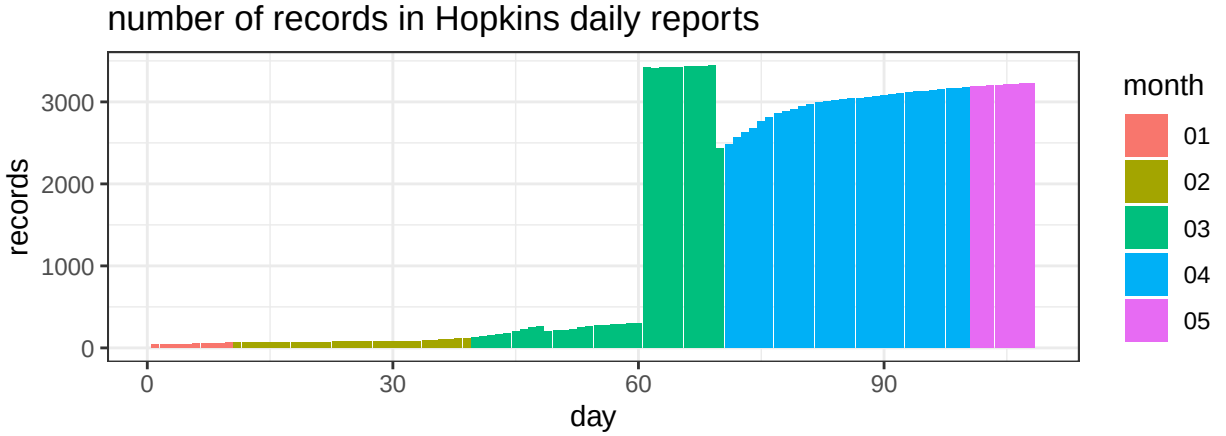
Netherlands



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] "2020-05-07"
```

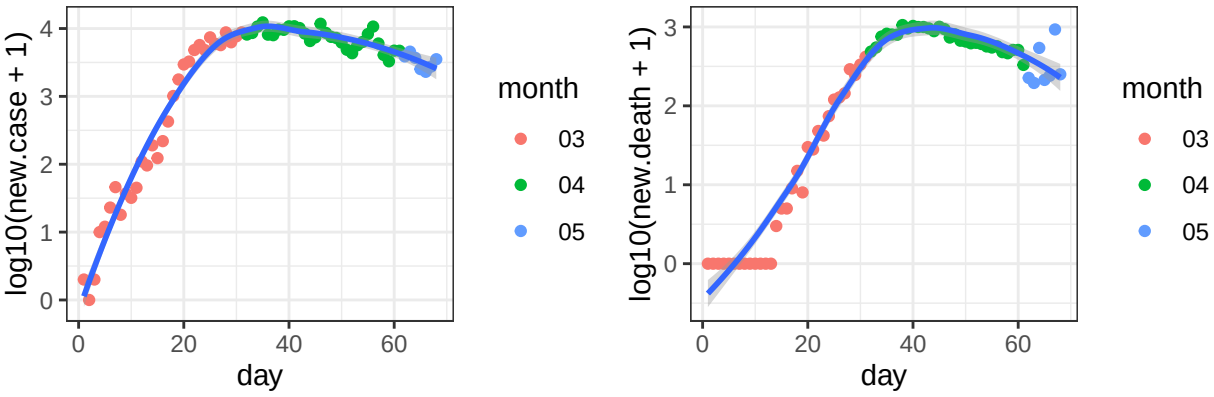
state level data

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

```
##            date                state fips  cases deaths
## 3623 2020-05-07             New York   36 332931  26206
## 3621 2020-05-07           New Jersey   34 133635   8801
## 3612 2020-05-07        Massachusetts   25  73721   4552
## 3613 2020-05-07             Michigan   26  45643   4343
## 3630 2020-05-07         Pennsylvania   42  56149   3599
## 3604 2020-05-07             Illinois   17  70802   3139
## 3596 2020-05-07          Connecticut    9  31784   2797
## 3594 2020-05-07           California    6  62481   2561
## 3609 2020-05-07            Louisiana   22  30652   2135
## 3599 2020-05-07              Florida   12  38820   1599
## 3611 2020-05-07             Maryland   24  29476   1503
## 3605 2020-05-07              Indiana   18  22942   1414
## 3600 2020-05-07              Georgia   13  30524   1333
## 3627 2020-05-07                 Ohio   39  22131   1271
## 3636 2020-05-07                Texas   48  36682   1016
## 3595 2020-05-07             Colorado    8  18264    942
## 3641 2020-05-07           Washington   53  17334    903
## 3640 2020-05-07             Virginia   51  21570    769
## 3624 2020-05-07       North Carolina   37  13431    521
## 3614 2020-05-07            Minnesota   27   9365    508
## 3592 2020-05-07              Arizona    4   9945    450
## 3616 2020-05-07             Missouri   29   9410    449
## 3615 2020-05-07          Mississippi   28   8686    396
## 3632 2020-05-07         Rhode Island   44  10530    388
## 3643 2020-05-07            Wisconsin   55   9215    374
## 3590 2020-05-07              Alabama    1   9046    369
## 3633 2020-05-07       South Carolina   45   7142    316
## 3608 2020-05-07             Kentucky   21   6173    302
## 3619 2020-05-07               Nevada   32   5888    293
## 3598 2020-05-07 District of Columbia   11   5654    285
```

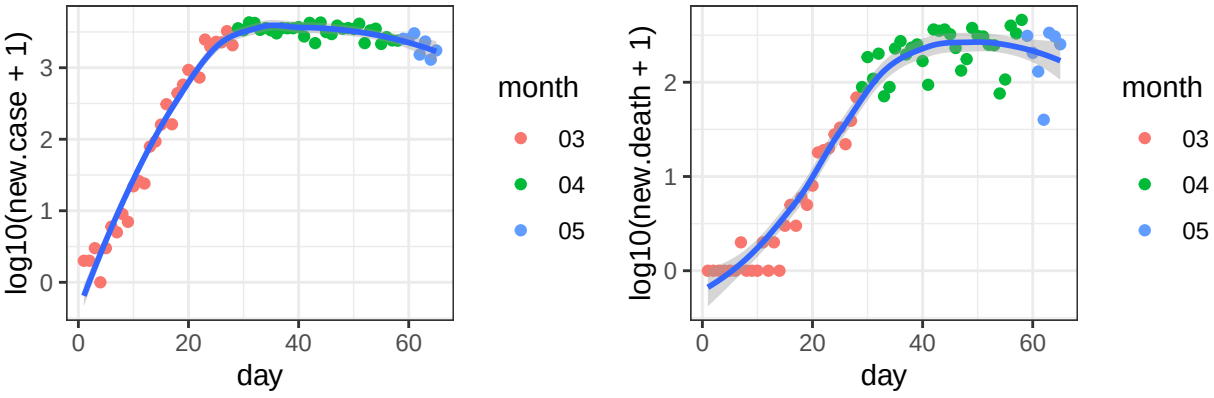
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



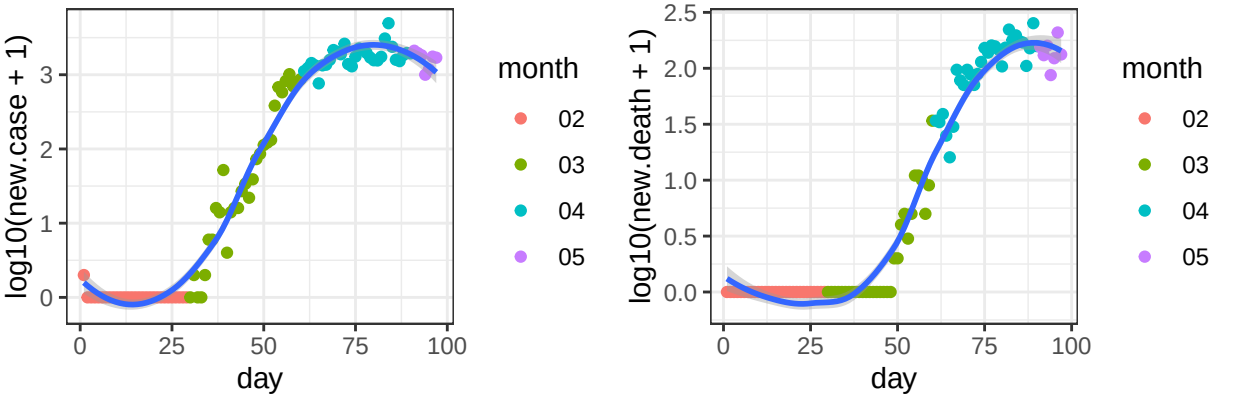
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



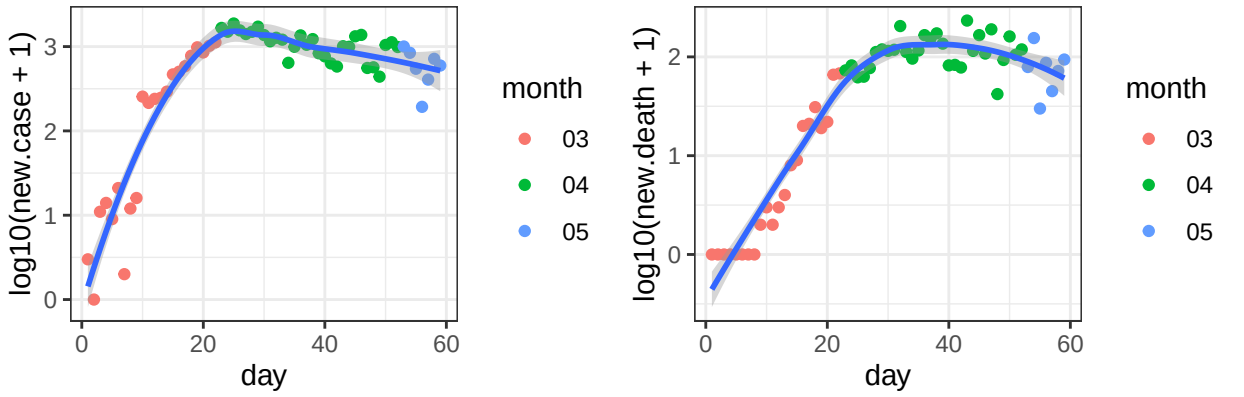
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Massachusetts



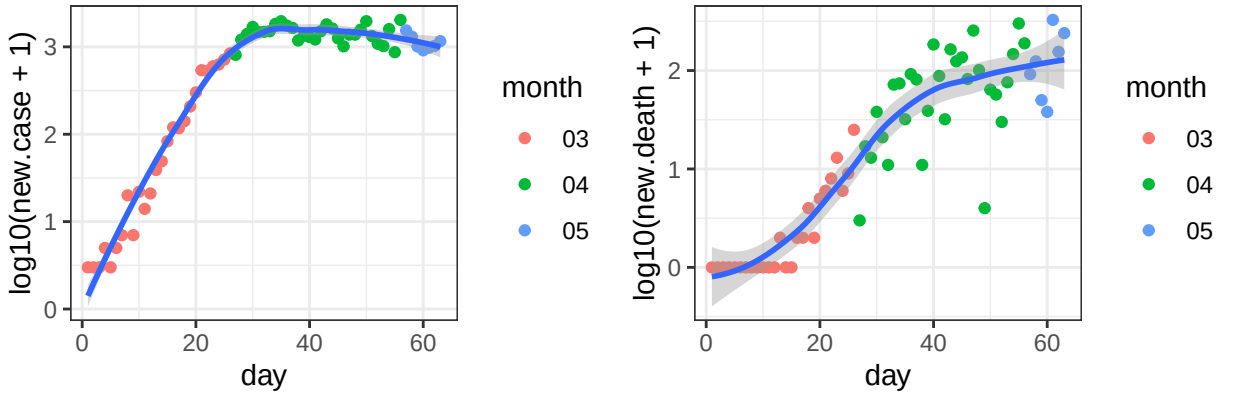
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Michigan



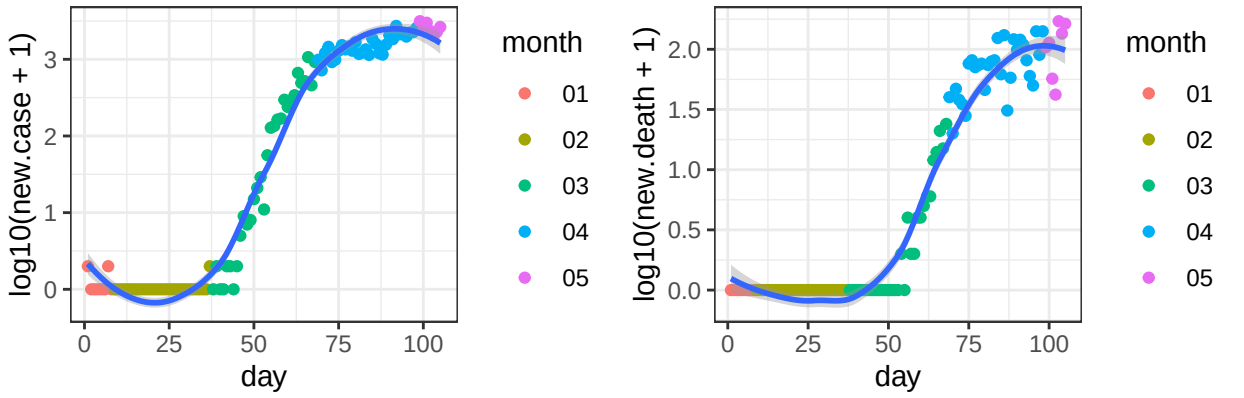
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Pennsylvania



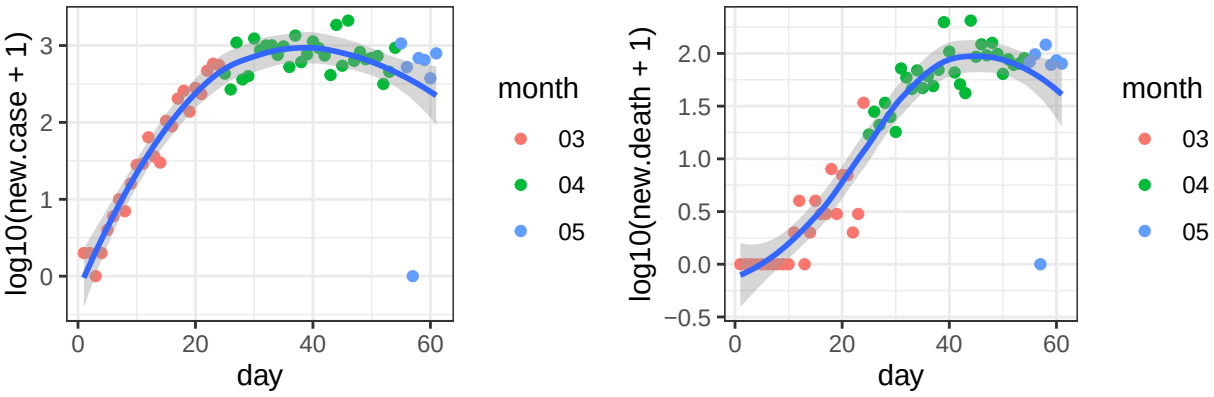
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Illinois



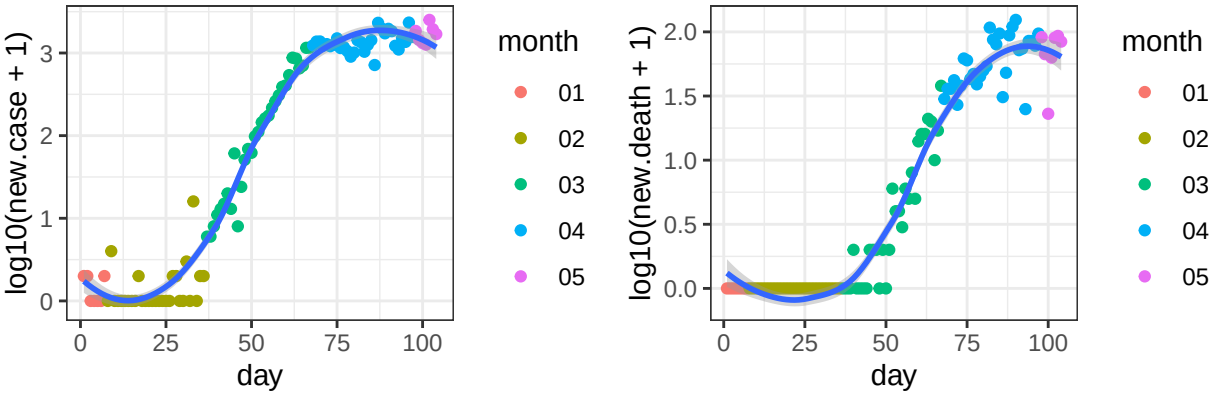
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Connecticut



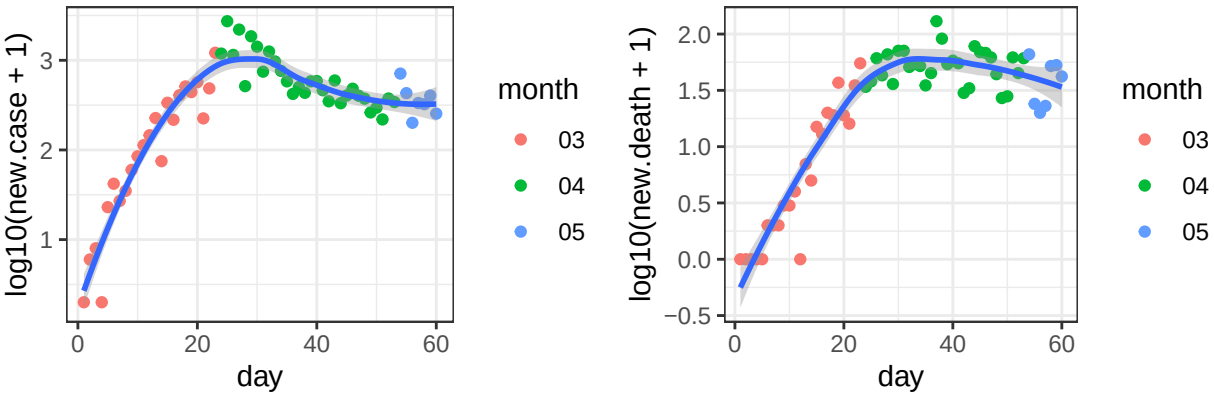
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

California



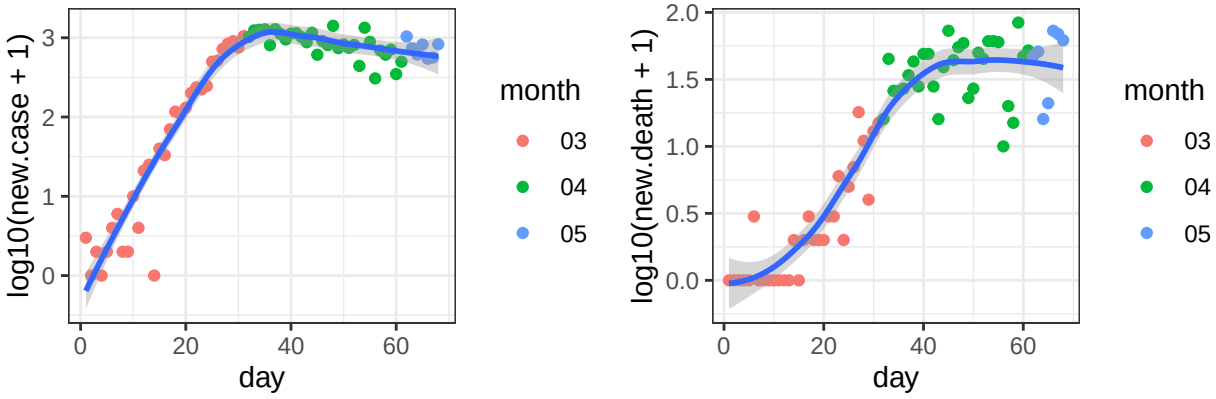
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Louisiana



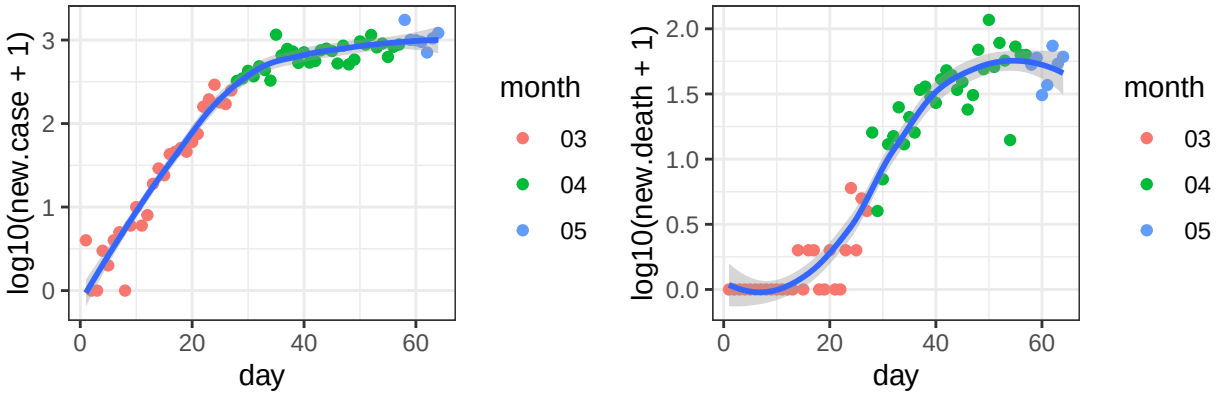
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Florida



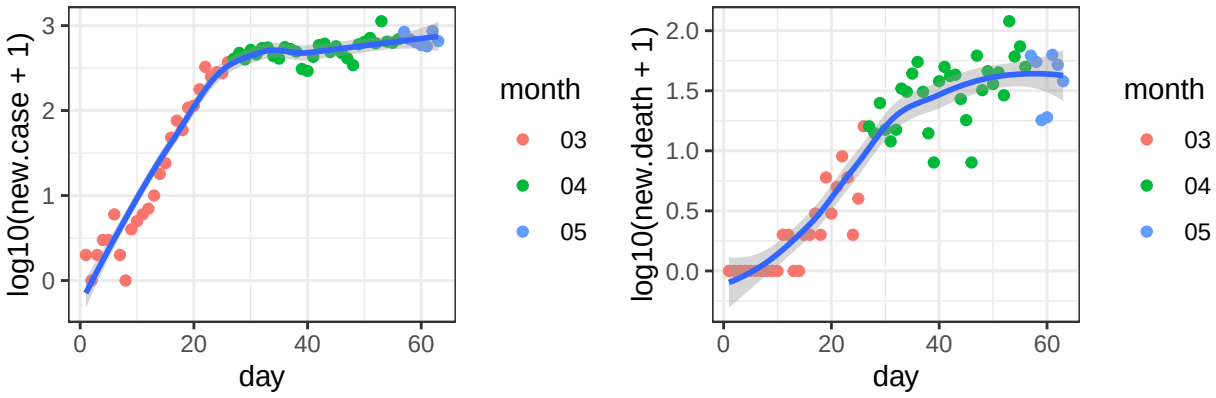
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Maryland



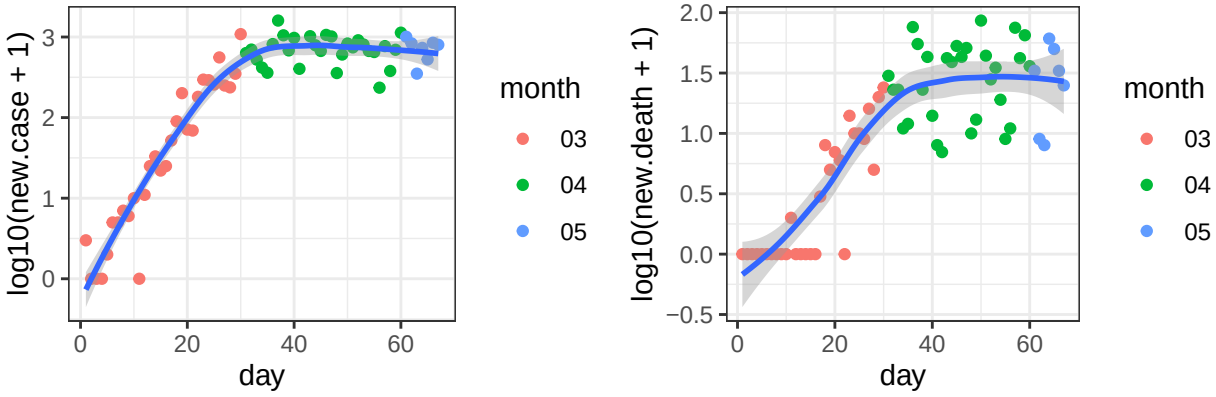
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Indiana



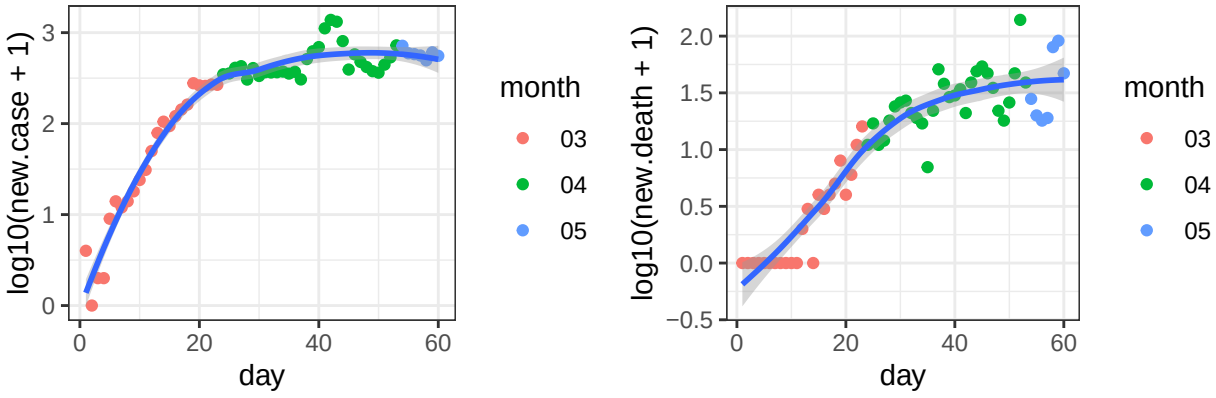
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Georgia



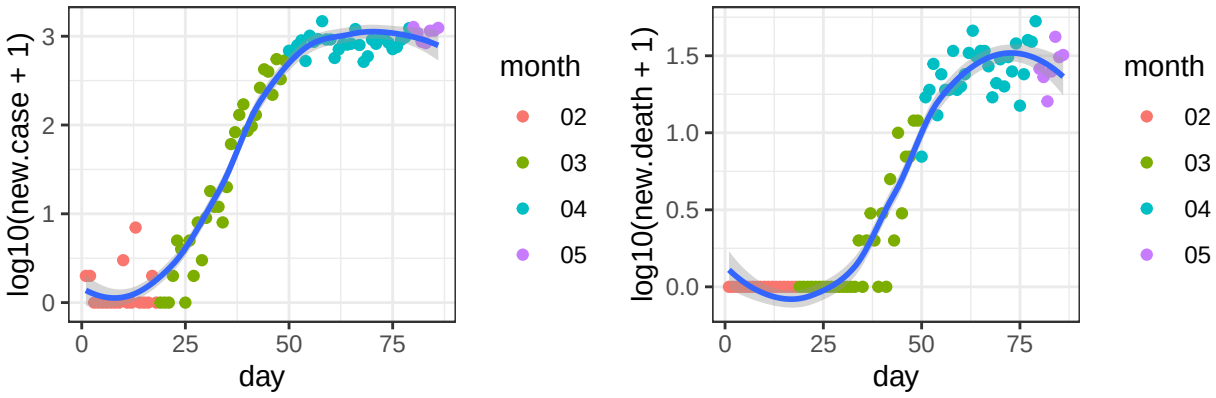
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Ohio



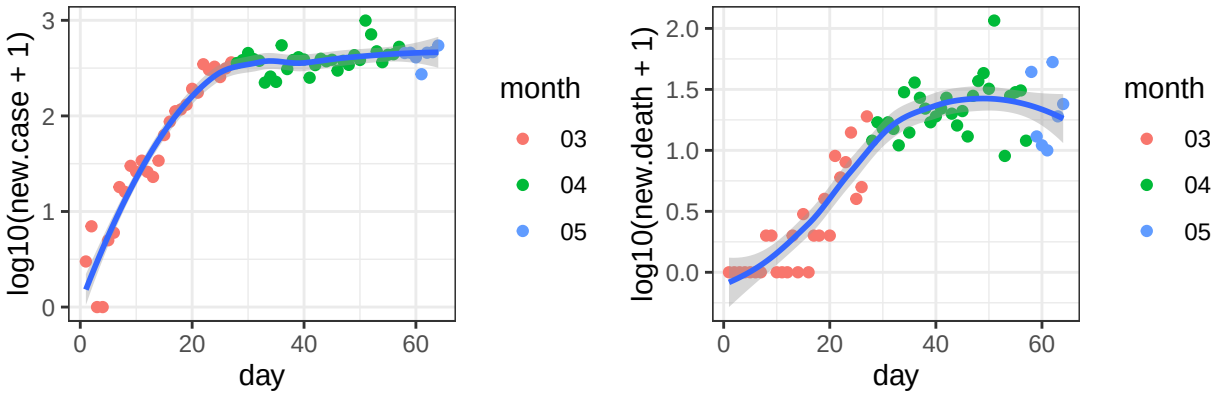
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Texas



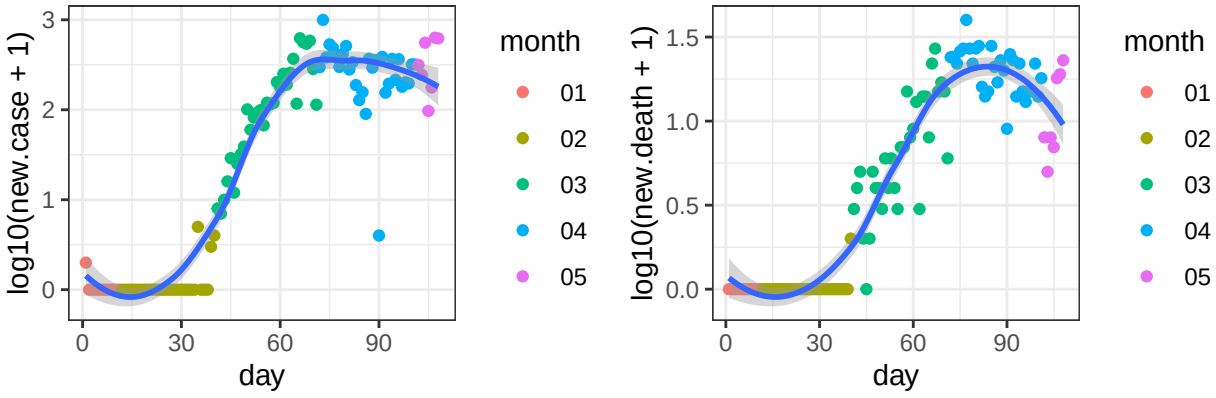
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Colorado



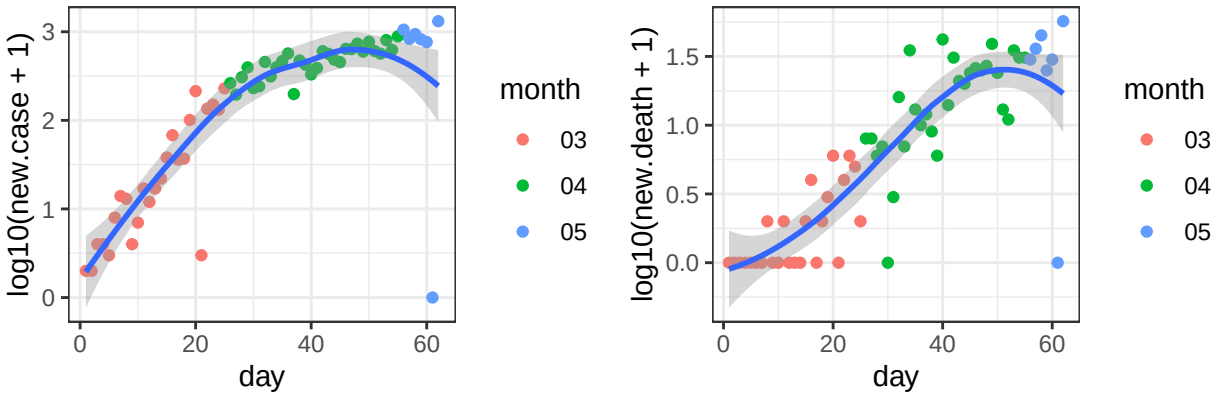
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Washington



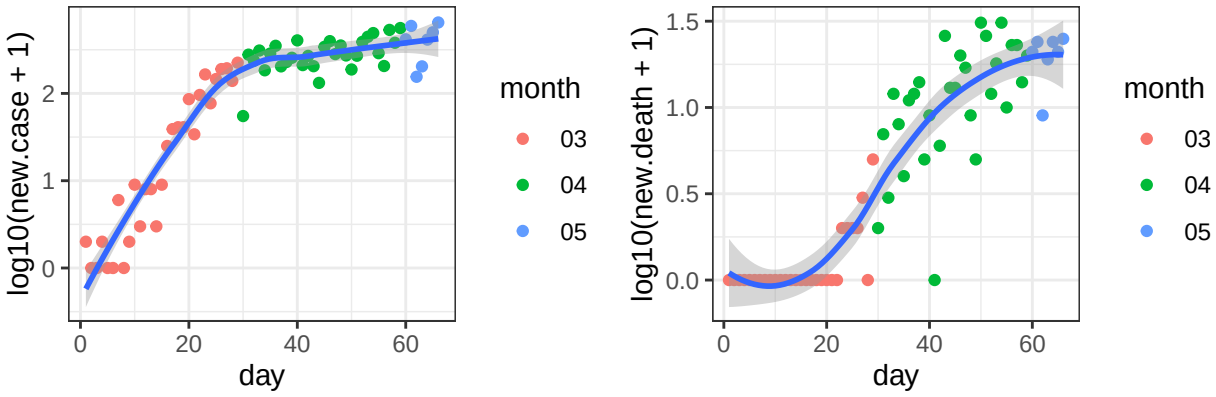
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Virginia



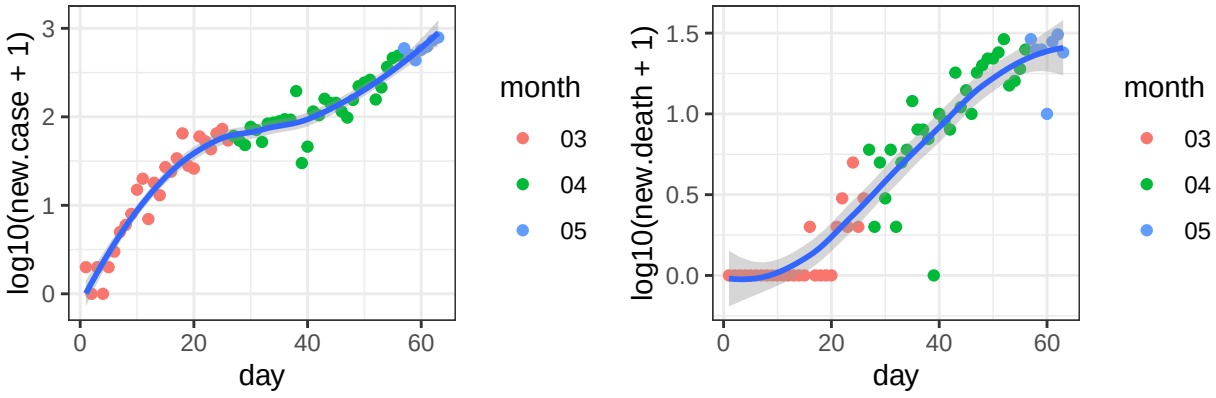
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

North Carolina



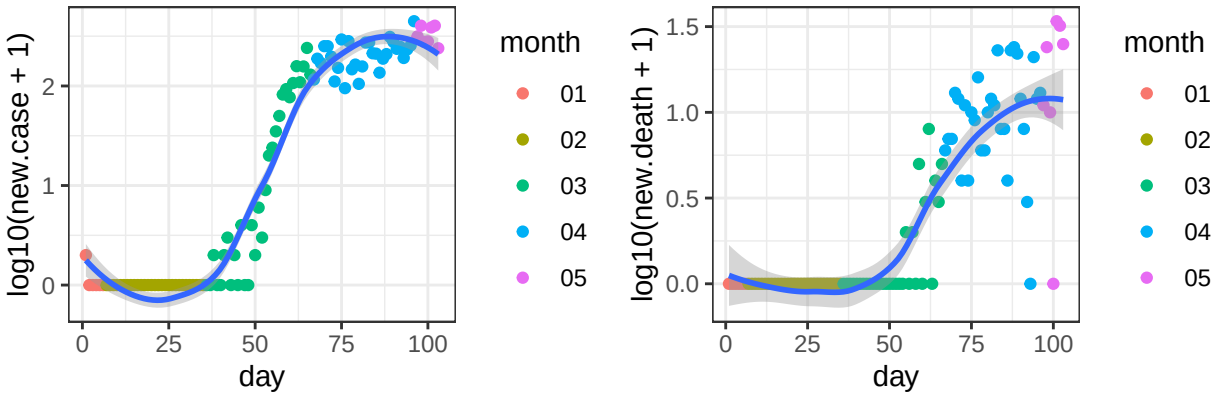
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Minnesota



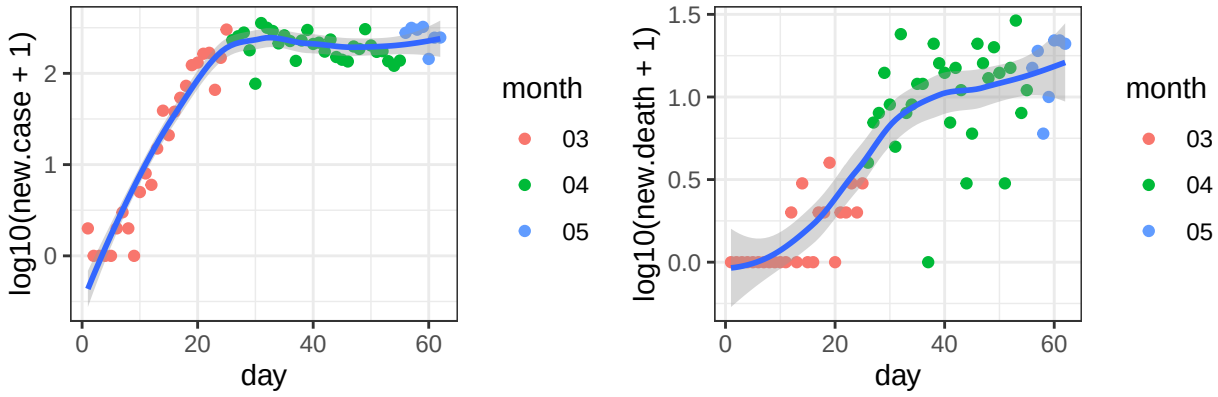
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Arizona



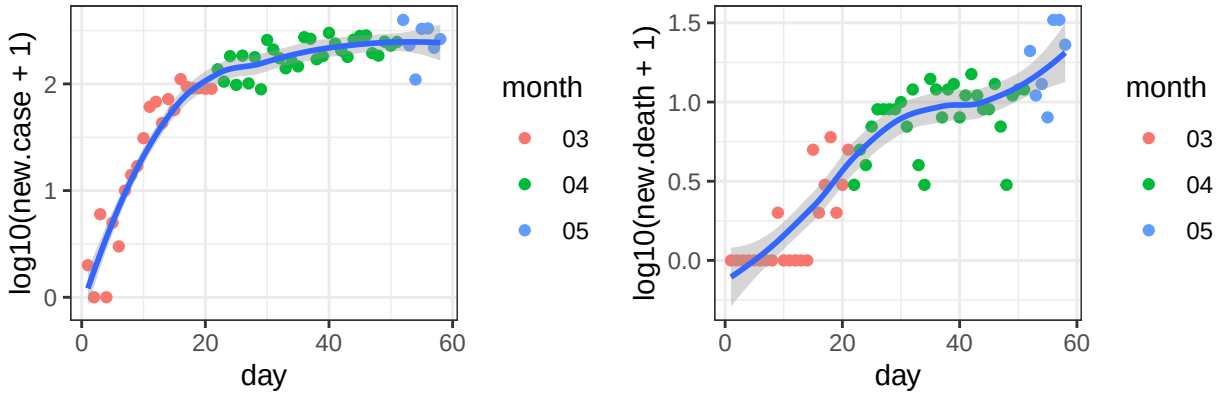
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Missouri



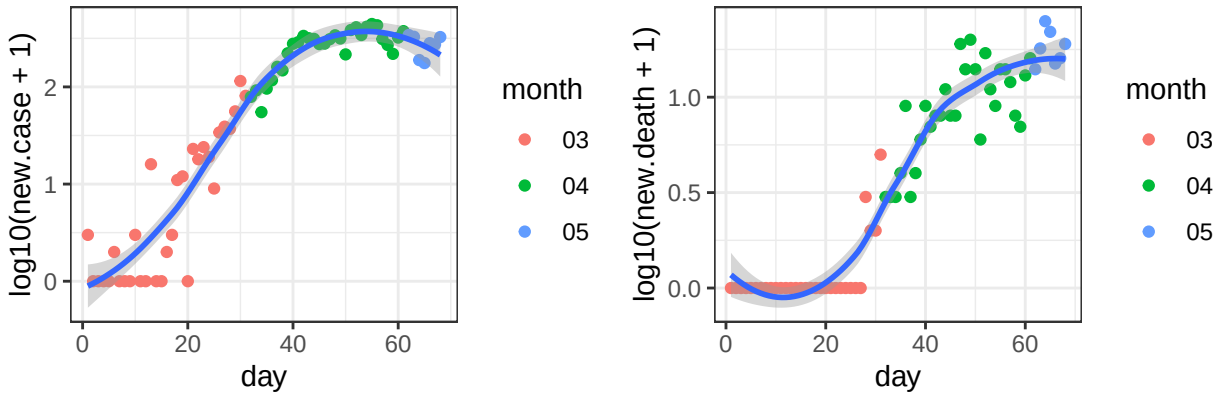
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Mississippi



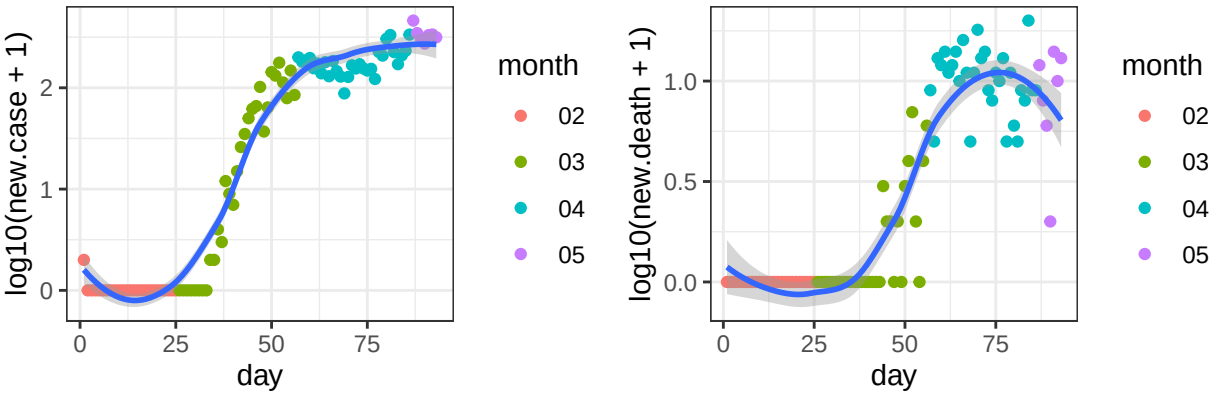
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Rhode Island



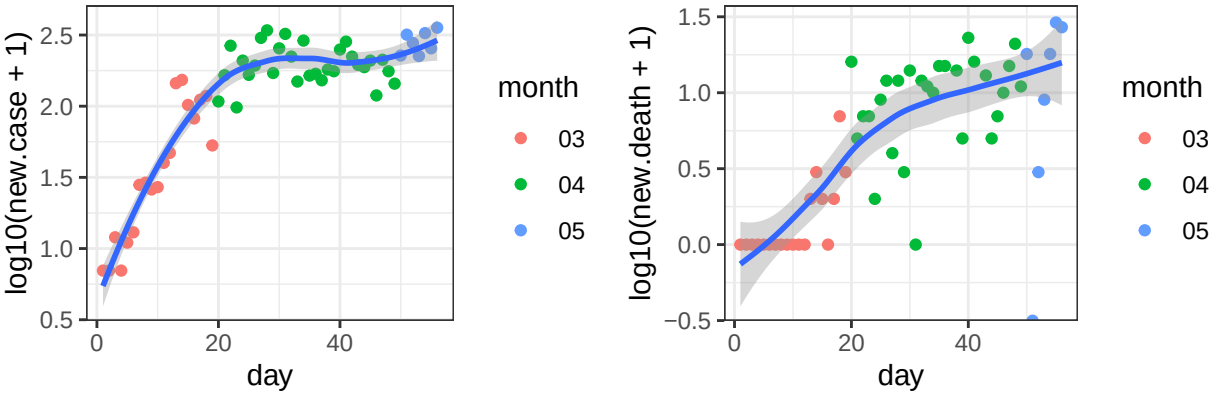
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Wisconsin



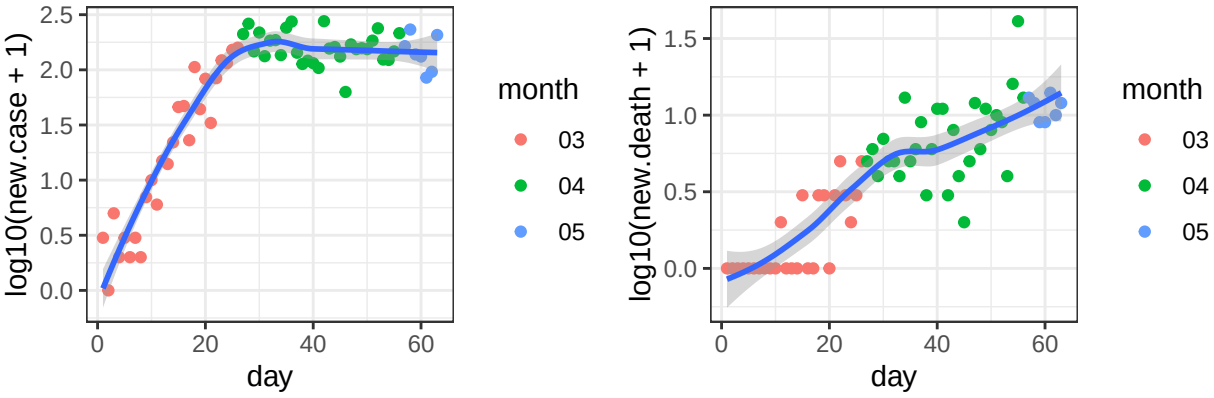
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

Alabama



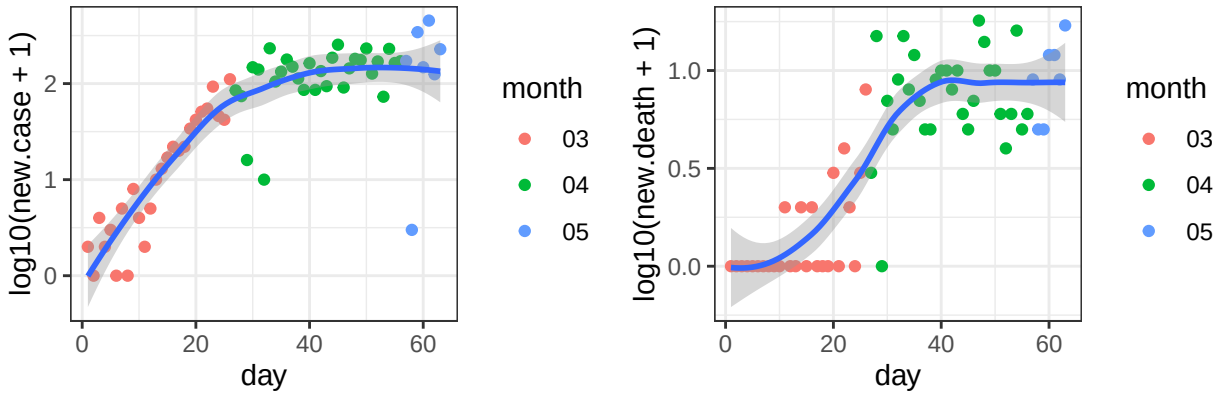
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

South Carolina



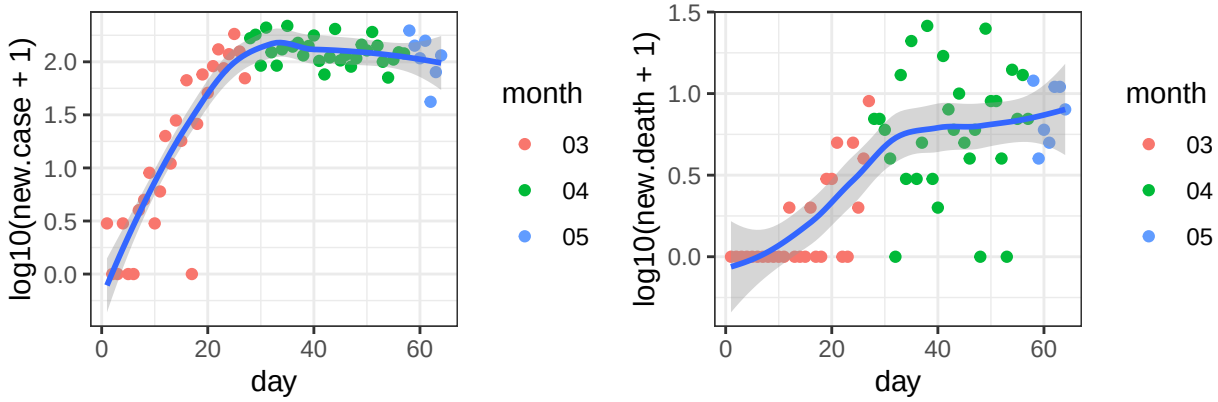
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Kentucky



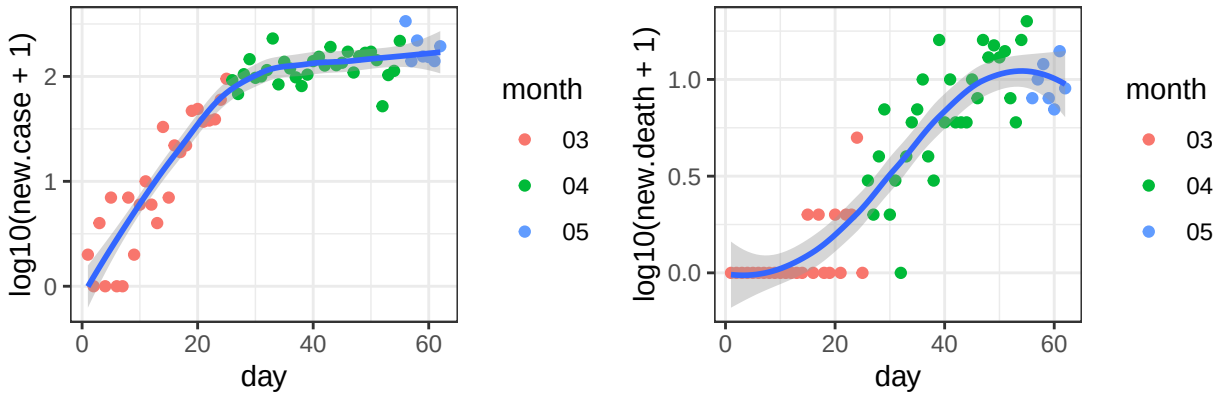
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Nevada



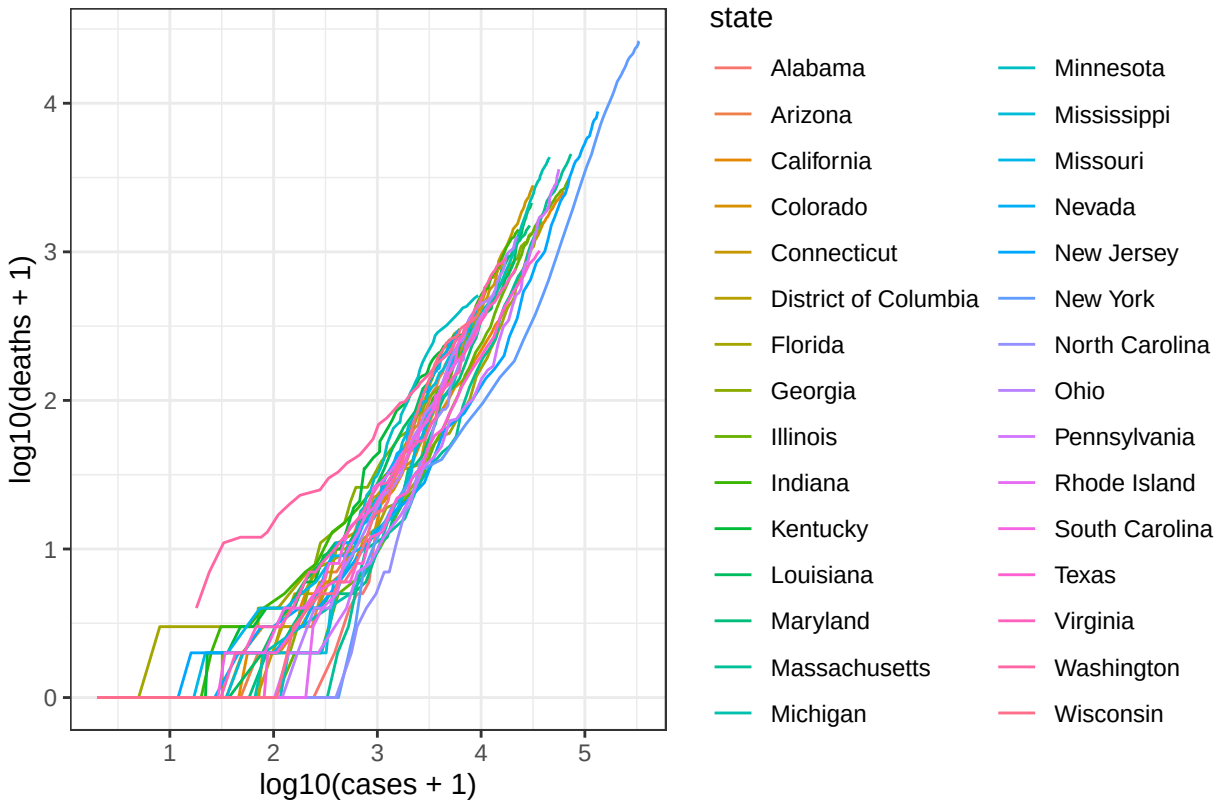
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

District of Columbia



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

Next I check the relation between the **cumulative** number of cases and deaths for these 10 states, starting on March



data source: https://github.com/nytimes/covid−19−data

county level data

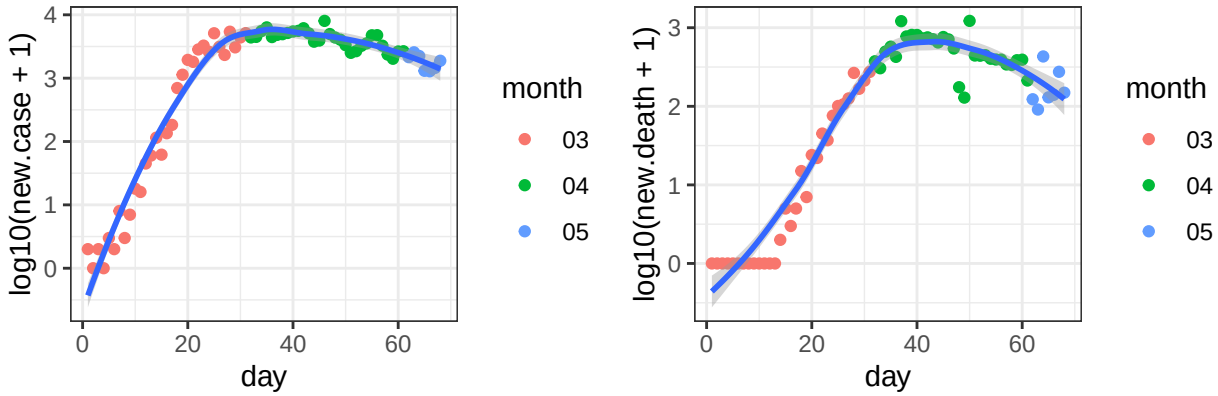
First check the 30 counties with the largest number of deaths.

```
##              date        county         state  fips  cases deaths
## 122759 2020-05-07 New York City      New York    NA 185653  19141
## 122758 2020-05-07        Nassau      New York 36059  37593   2340
## 121624 2020-05-07          Cook      Illinois 17031  48341   2110
## 122289 2020-05-07         Wayne      Michigan 26163  17667   2012
## 122778 2020-05-07       Suffolk      New York 36103  35892   1599
## 121231 2020-05-07   Los Angeles    California  6037  29427   1418
## 122684 2020-05-07         Essex    New Jersey 34013  15095   1381
## 122679 2020-05-07        Bergen    New Jersey 34003  16609   1319
## 122786 2020-05-07   Westchester      New York 36119  30709   1305
## 122204 2020-05-07     Middlesex Massachusetts 25017  16676   1103
## 121329 2020-05-07     Fairfield   Connecticut  9001  12679    977
## 122686 2020-05-07        Hudson    New Jersey 34017  16354    923
## 121330 2020-05-07      Hartford   Connecticut  9003   6750    867
## 122697 2020-05-07         Union    New Jersey 34039  13781    829
## 123170 2020-05-07  Philadelphia  Pennsylvania 42101  17047    816
## 122270 2020-05-07       Oakland      Michigan 26125   7624    789
## 122689 2020-05-07     Middlesex    New Jersey 34023  13411    737
## 122693 2020-05-07       Passaic    New Jersey 34031  14133    703
## 122257 2020-05-07         Macomb      Michigan 26099   5876    678
## 122208 2020-05-07       Suffolk Massachusetts 25025  14732    663
## 121333 2020-05-07     New Haven   Connecticut  9009   8678    643
## 122206 2020-05-07       Norfolk Massachusetts 25021   6729    608
```

```
## 122200 2020-05-07          Essex Massachusetts 25009  10610    578
## 123165 2020-05-07    Montgomery  Pennsylvania 42091   4915    506
## 122691 2020-05-07        Morris    New Jersey 34027   5702    503
## 122692 2020-05-07         Ocean    New Jersey 34029   7209    500
## 123785 2020-05-07          King    Washington 53033   7182    482
## 122124 2020-05-07       Orleans     Louisiana 22071   6626    463
## 121385 2020-05-07    Miami-Dade       Florida 12086  13584    454
## 122202 2020-05-07       Hampden Massachusetts 25013   4441    434
```

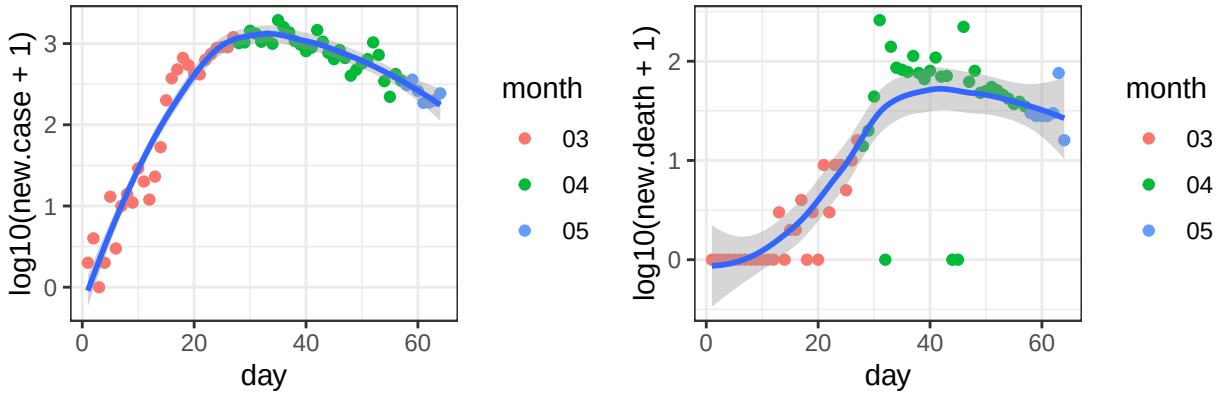
For these 30 counties, I check the number of new cases and the number of new deaths.

New York City_New York



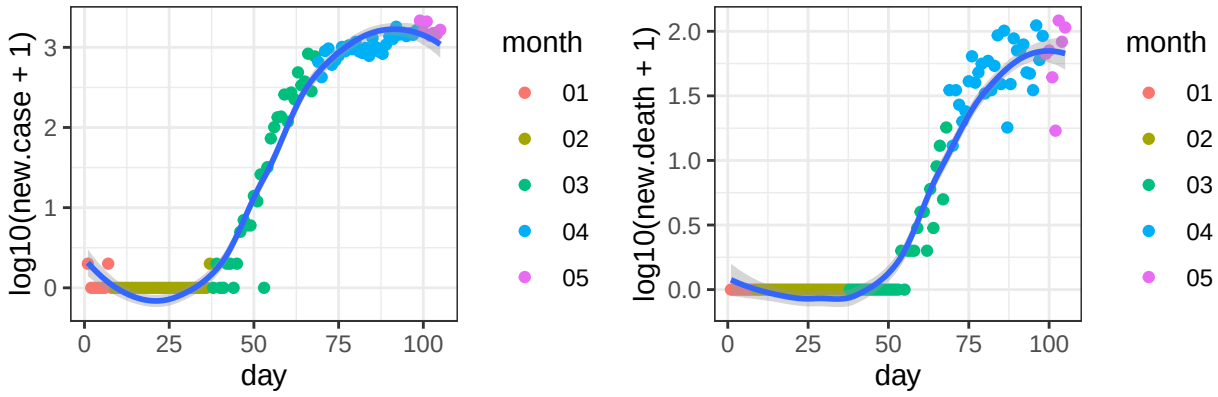
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Nassau_New York



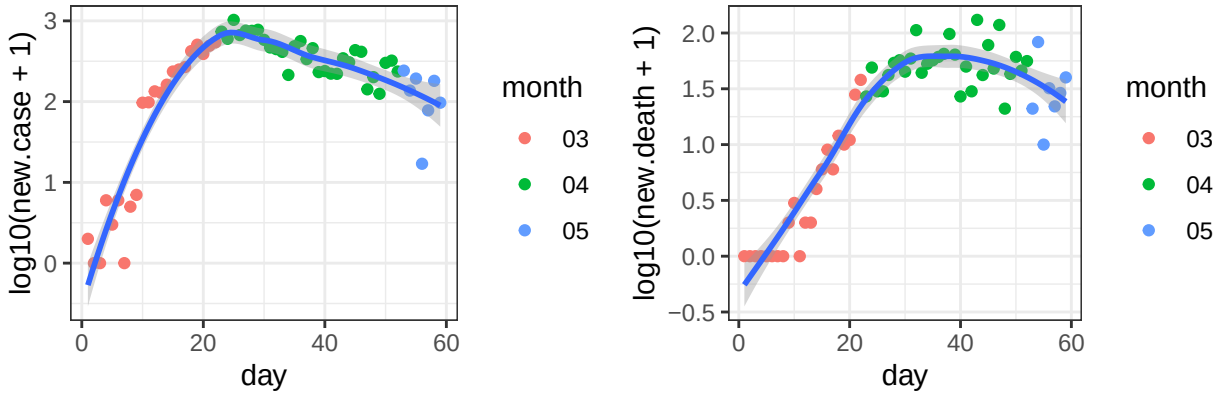
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Cook_Illinois



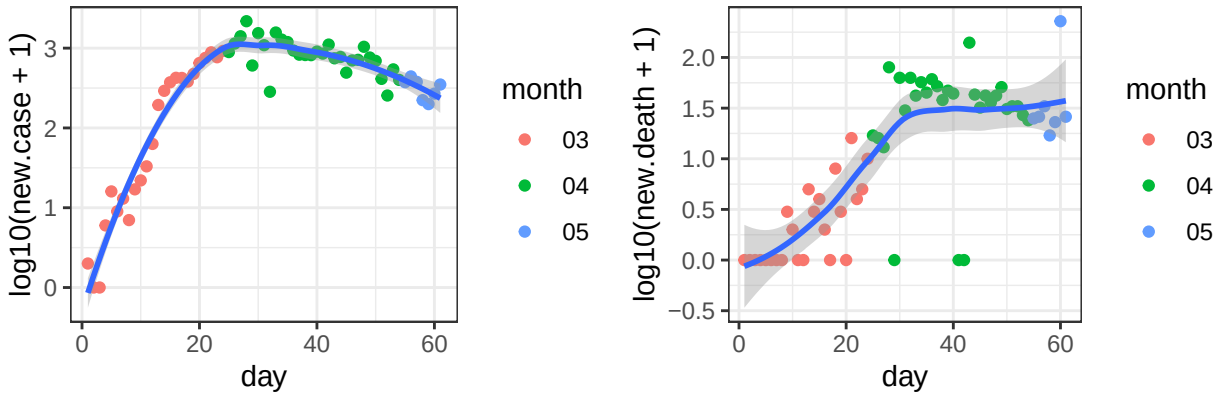
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Wayne_Michigan



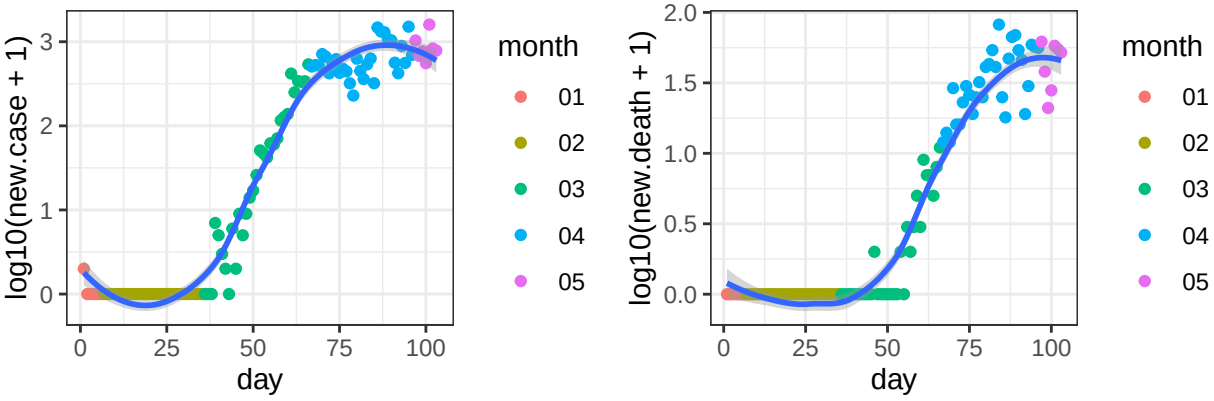
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Suffolk_New York



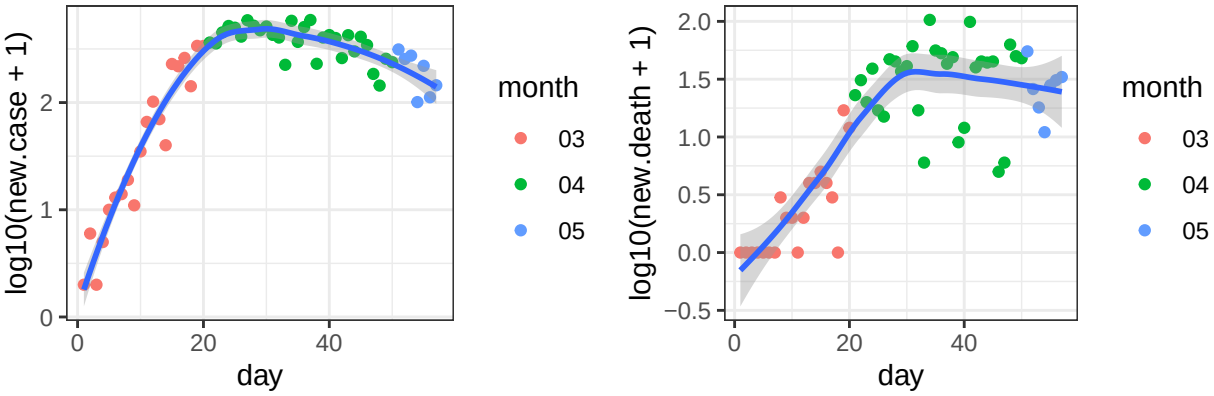
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Los Angeles_California



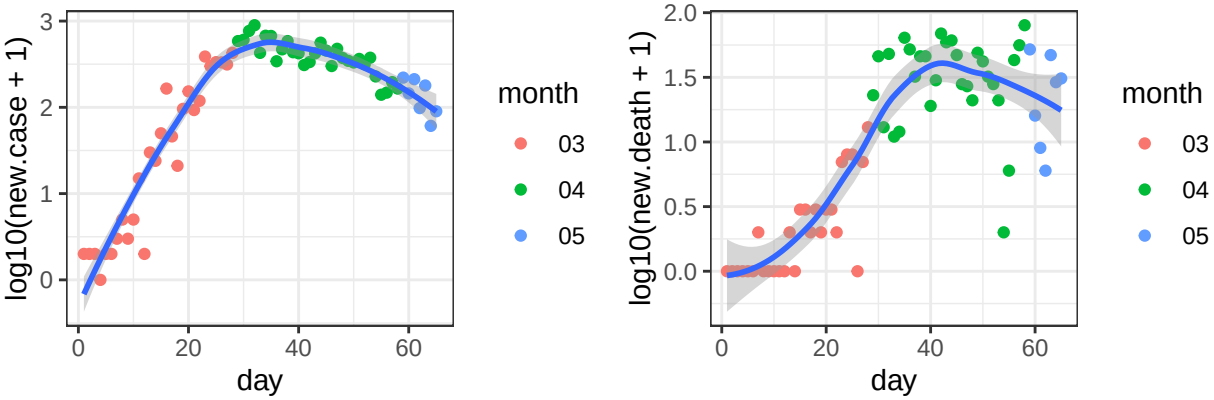
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Essex_New Jersey



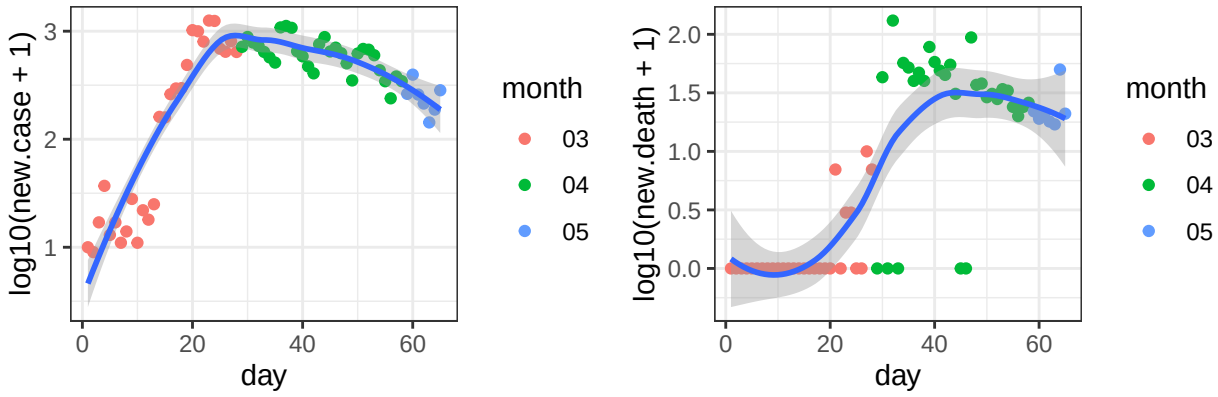
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Bergen_New Jersey



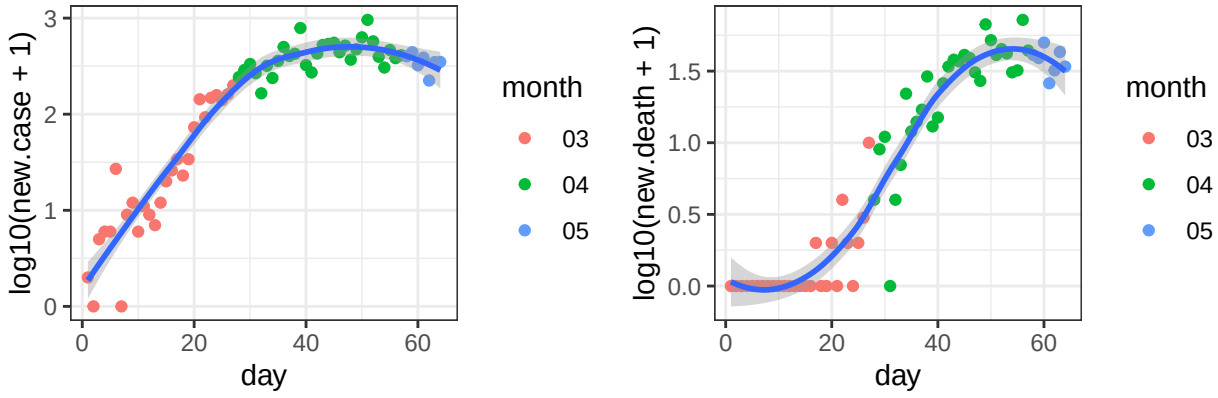
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Westchester_New York



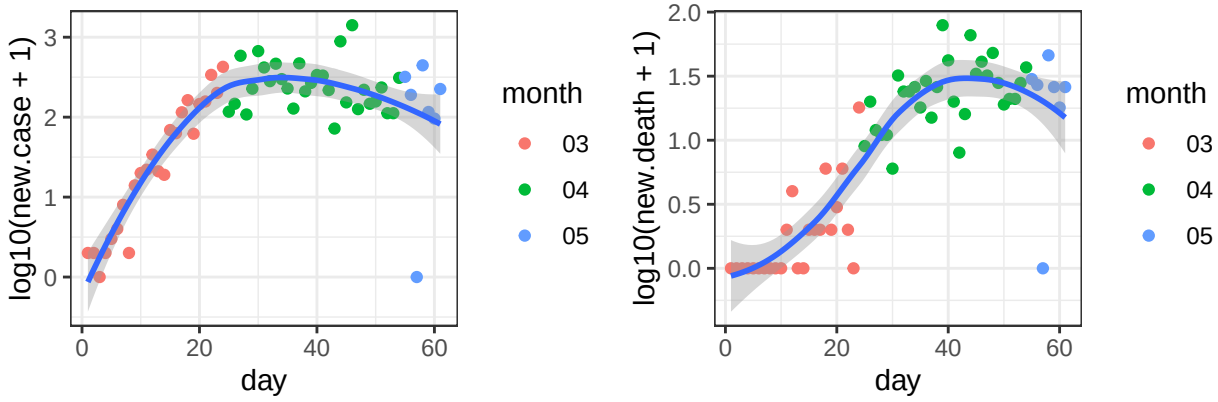
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Middlesex_Massachusetts



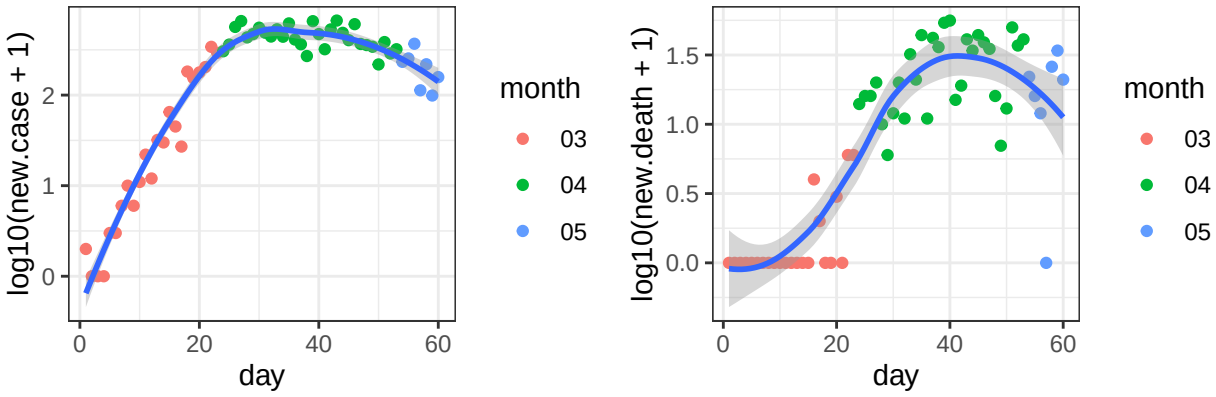
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Fairfield_Connecticut



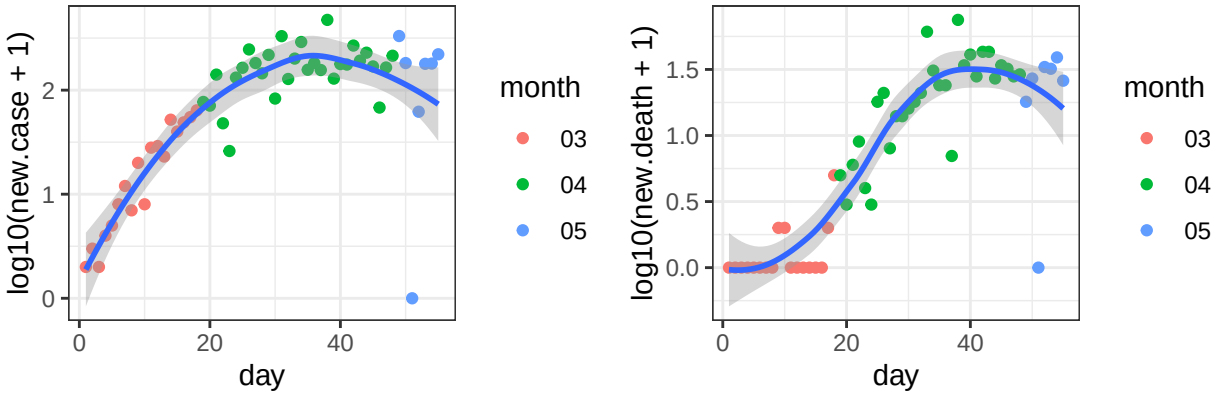
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Hudson_New Jersey



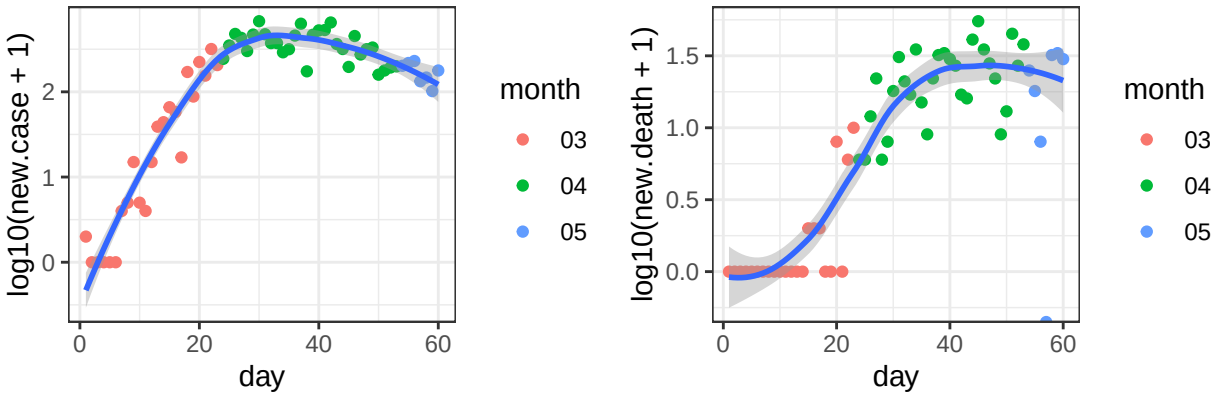
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Hartford_Connecticut



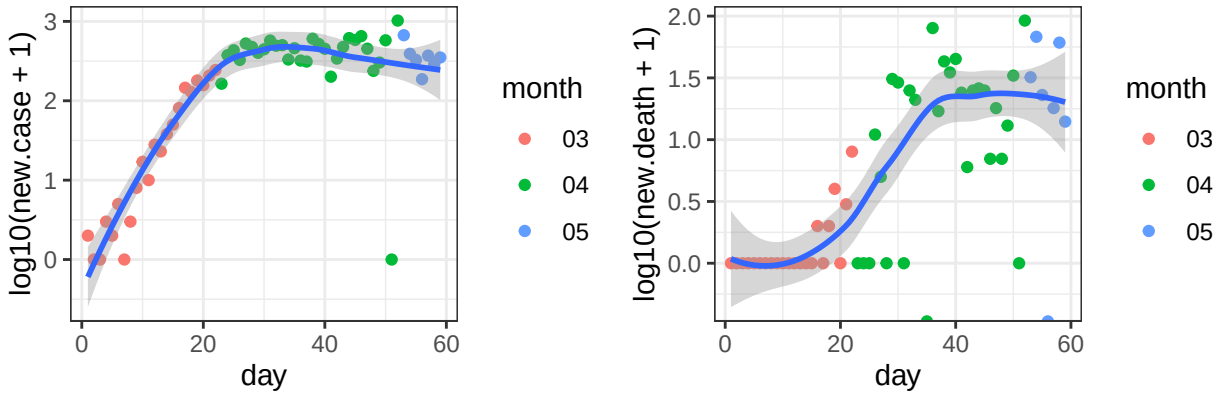
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Union_New Jersey



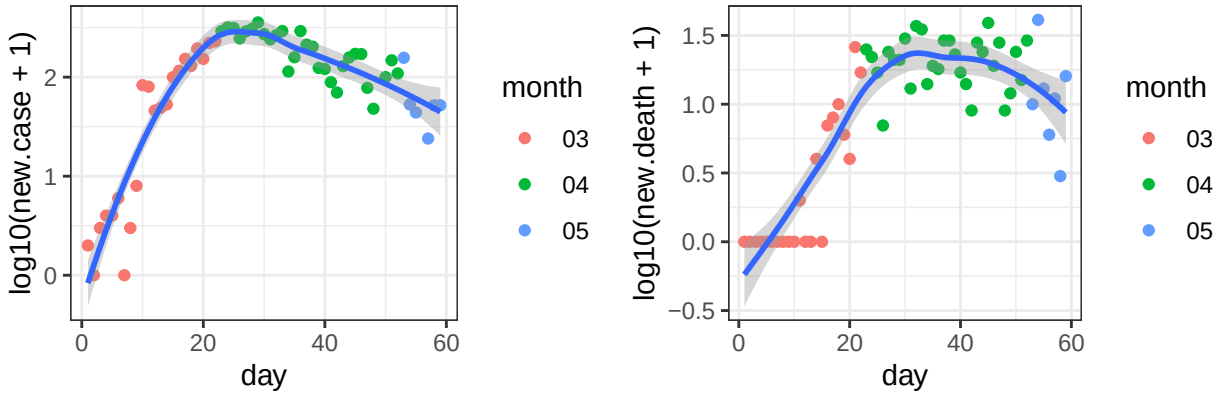
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Philadelphia_Pennsylvania



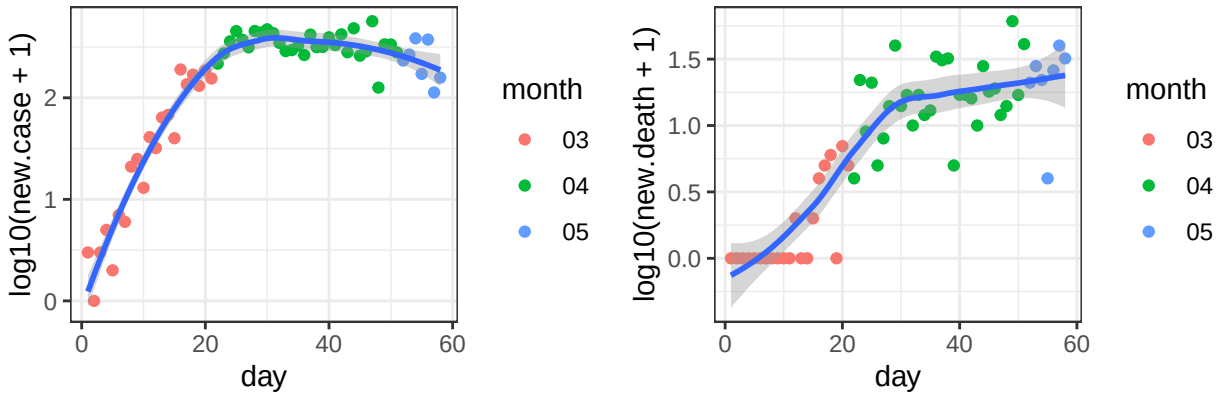
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Oakland_Michigan



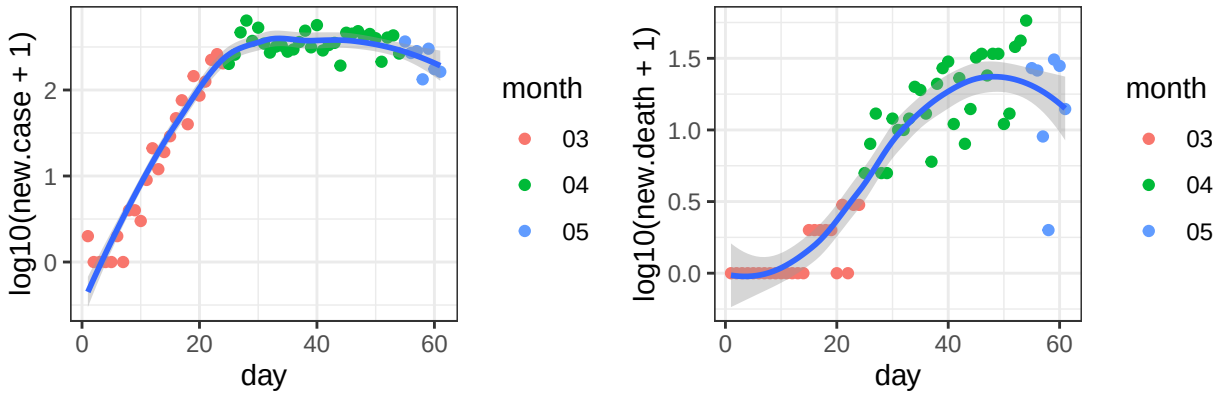
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Middlesex_New Jersey



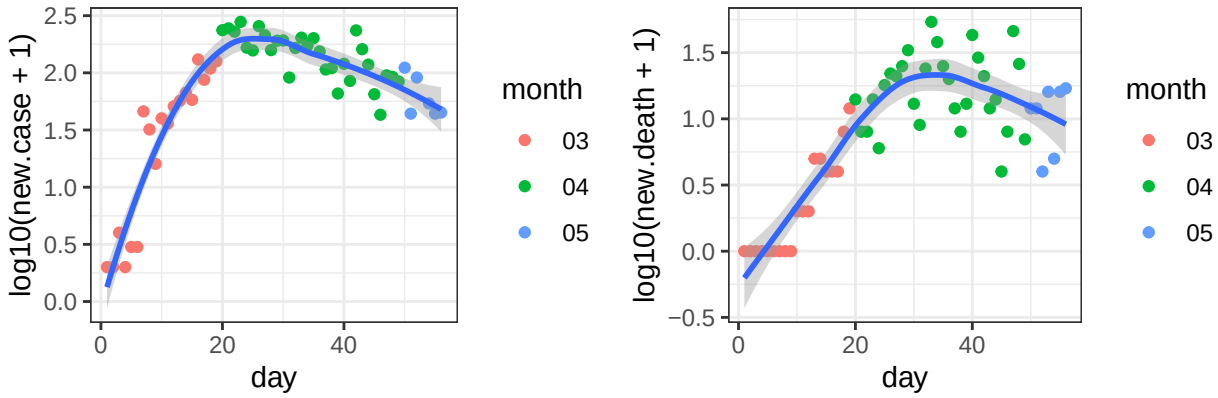
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Passaic_New Jersey



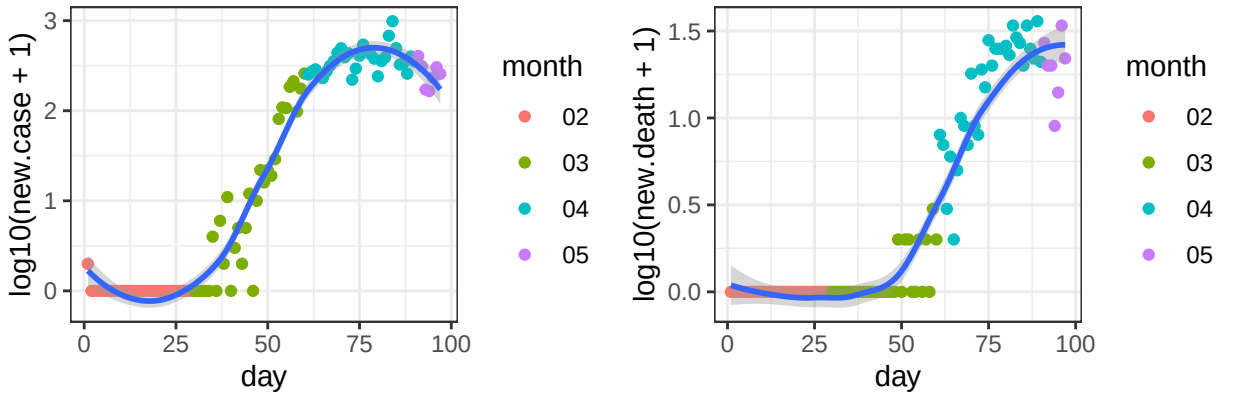
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Macomb_Michigan



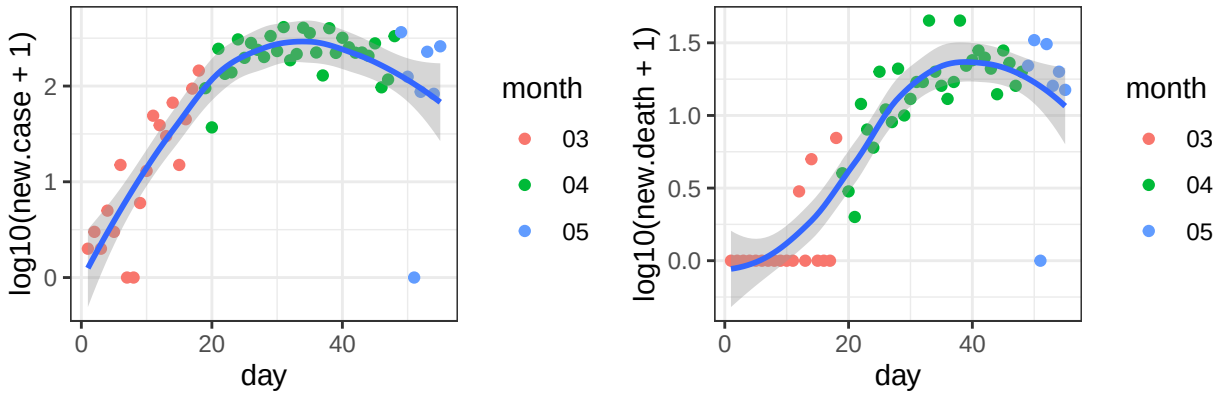
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Suffolk_Massachusetts



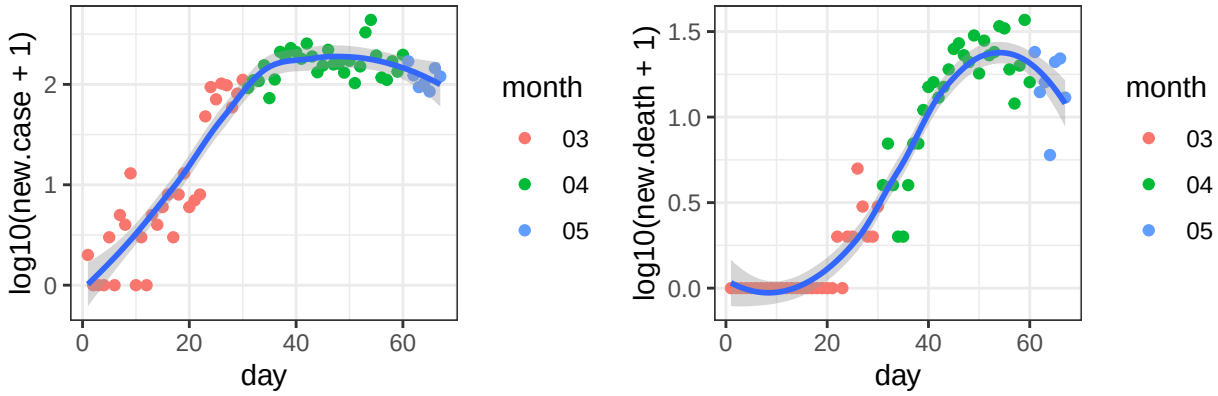
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

New Haven_Connecticut



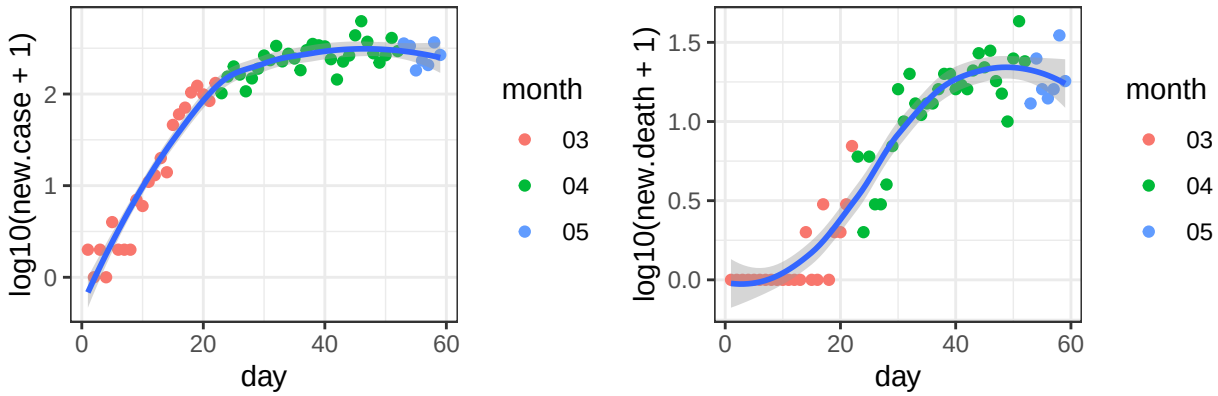
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Norfolk_Massachusetts



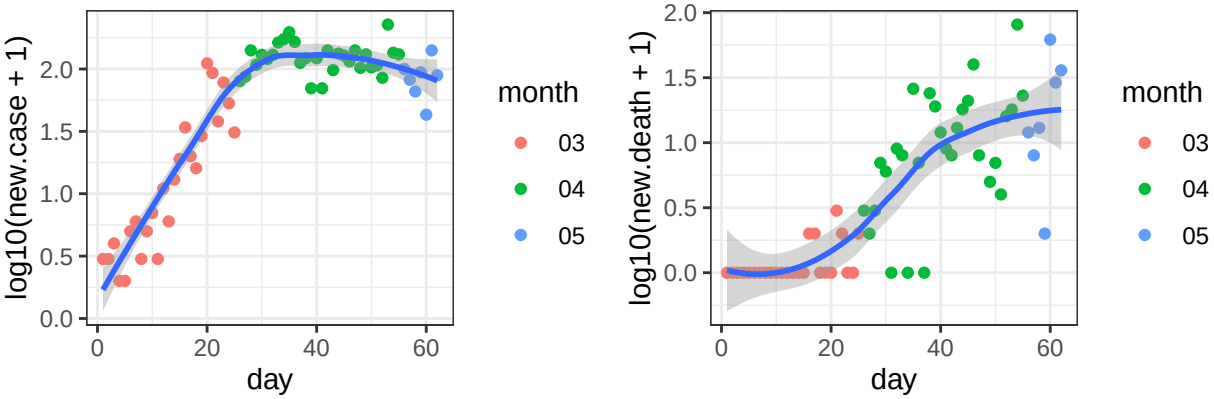
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Essex_Massachusetts



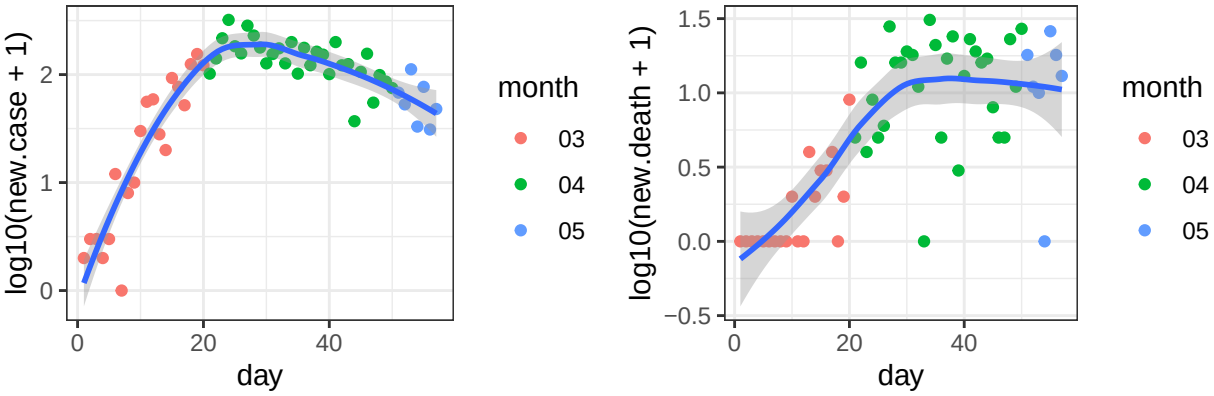
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Montgomery_Pennsylvania



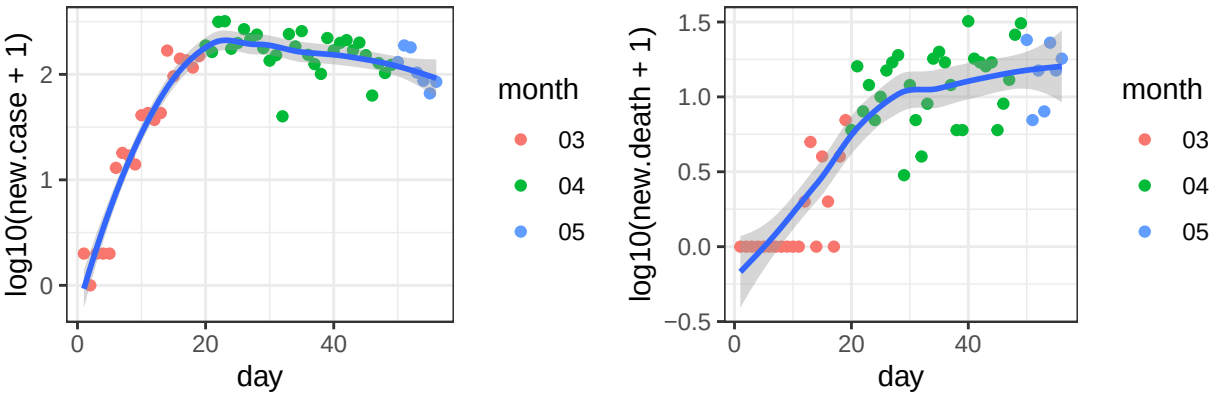
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Morris_New Jersey



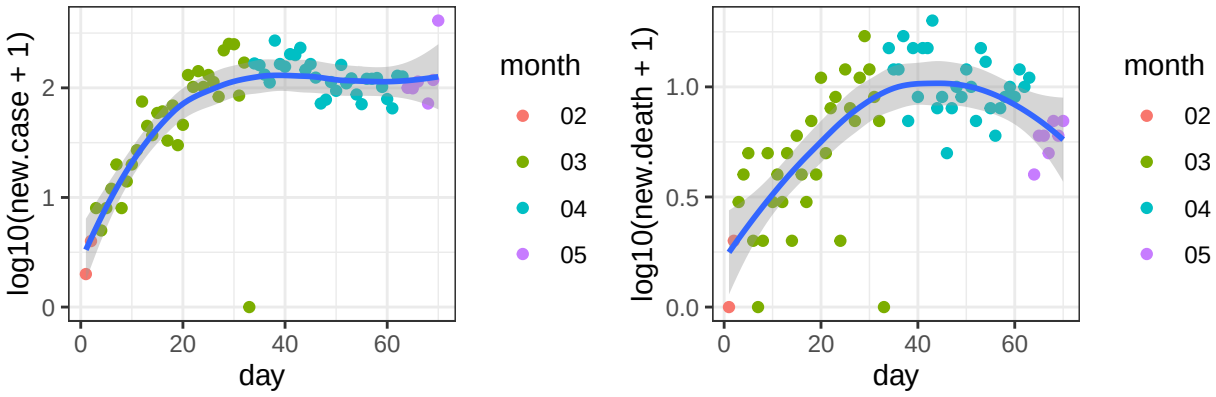
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Ocean_New Jersey



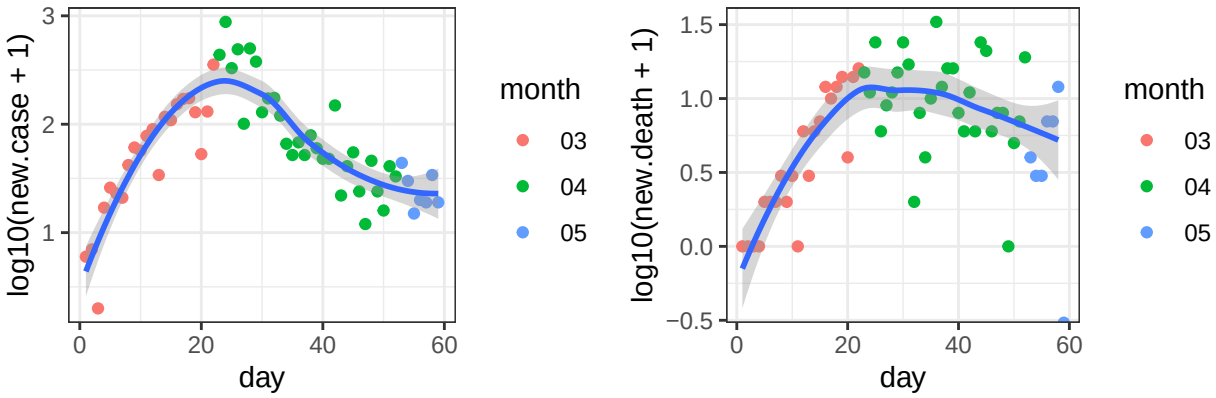
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

King_Washington



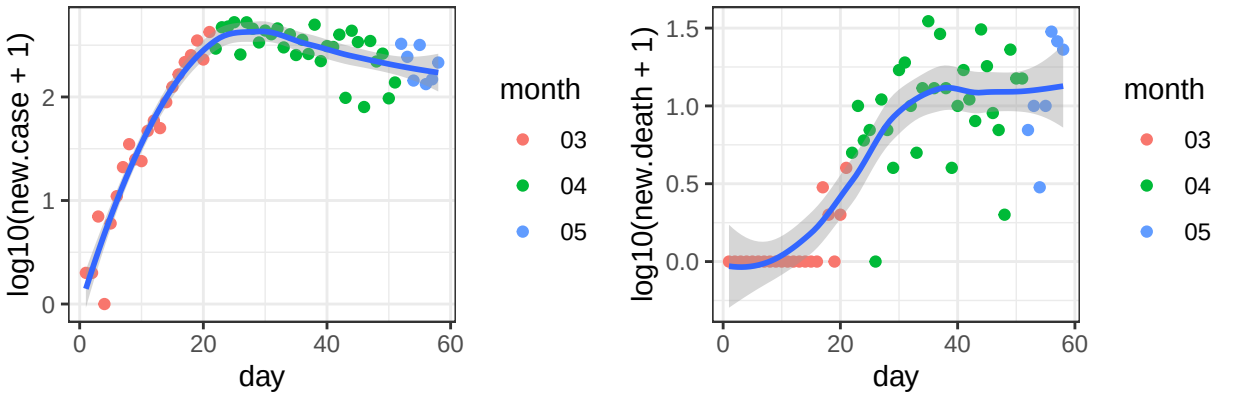
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

Orleans_Louisiana



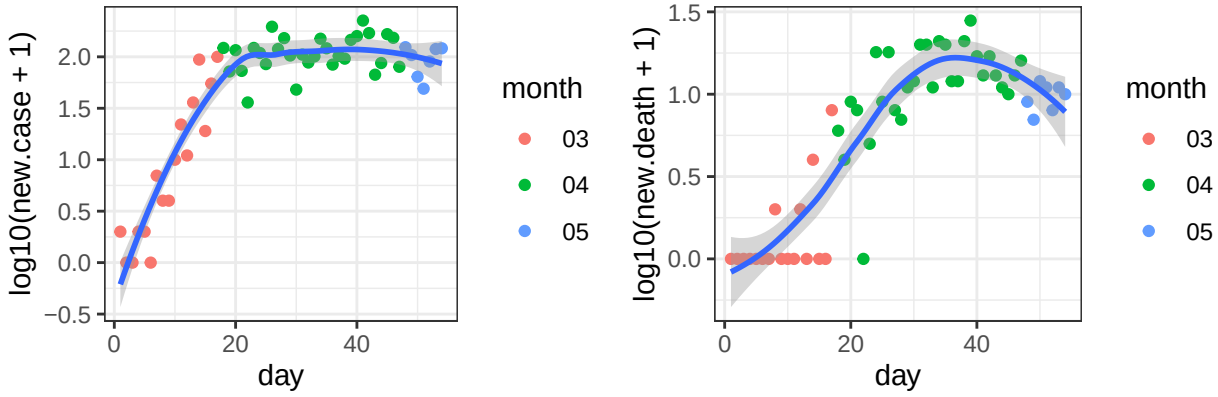
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Miami−Dade_Florida



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

Hampden_Massachusetts

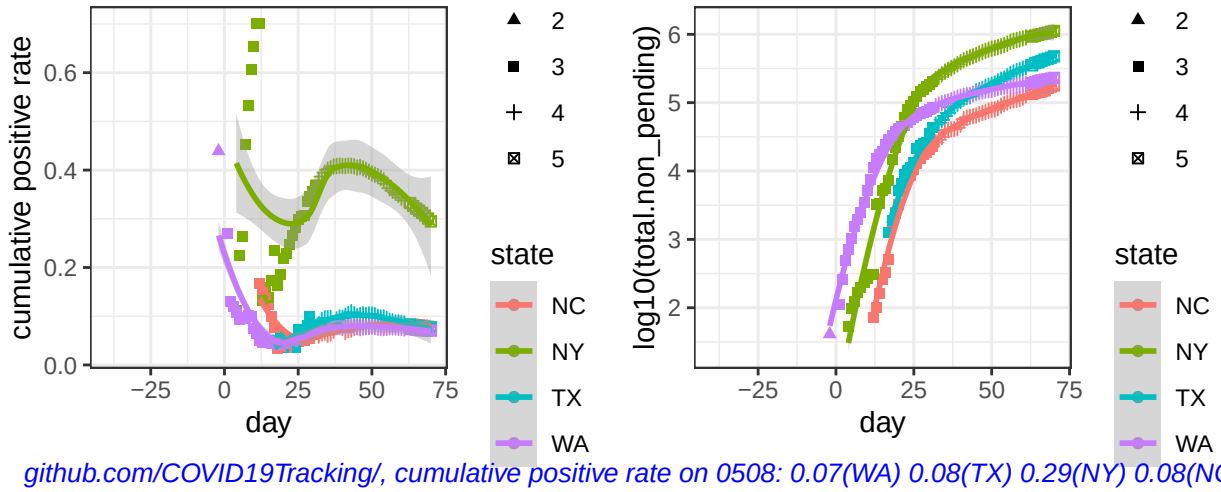


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

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 0508: 0.07(WA) 0.08(TX) 0.29(NY) 0.08(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
```