Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-05-30

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

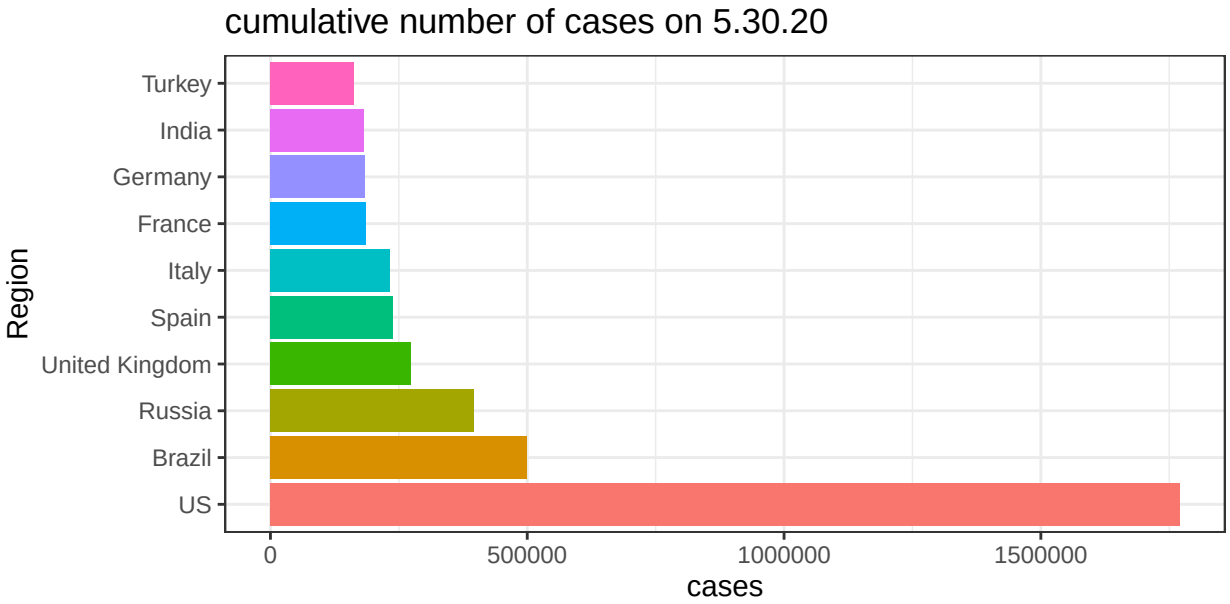
JHU

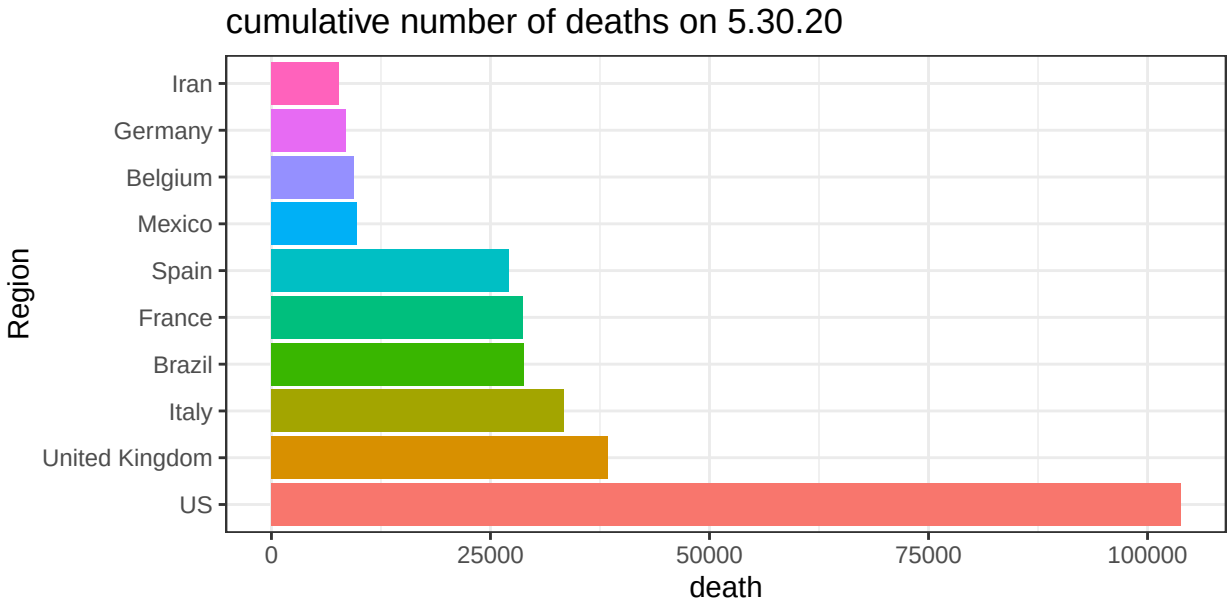
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

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.

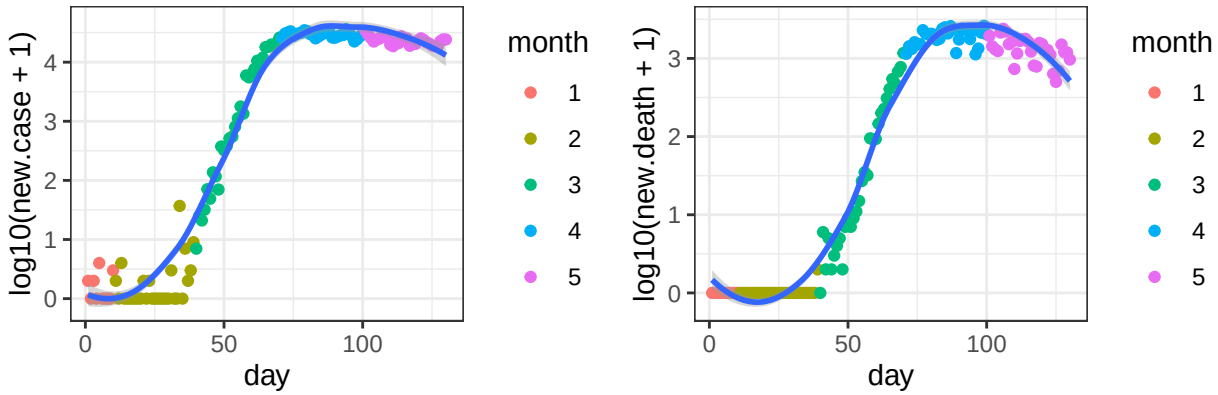
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





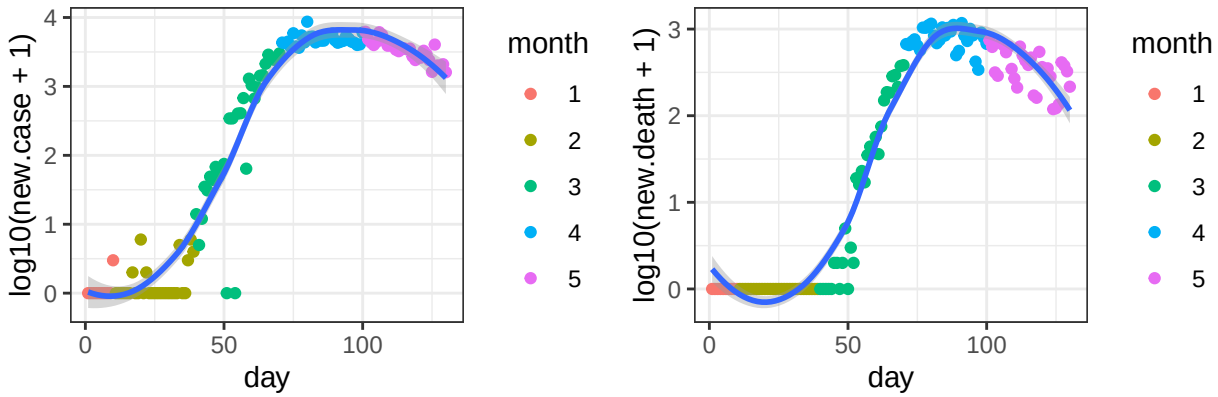
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



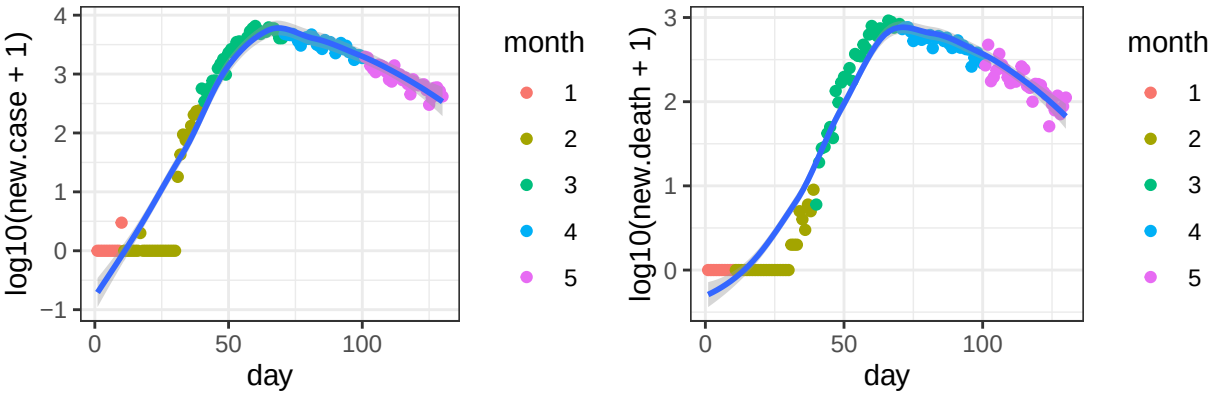
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



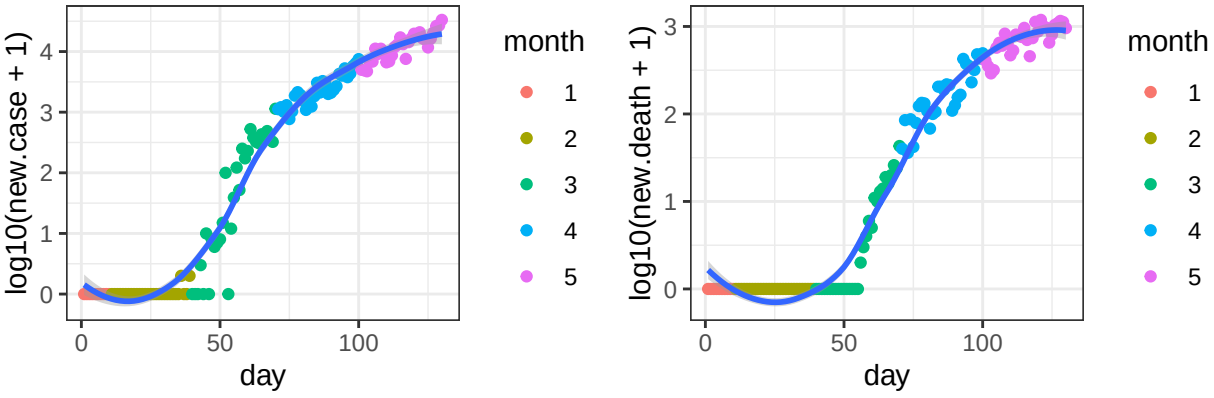
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



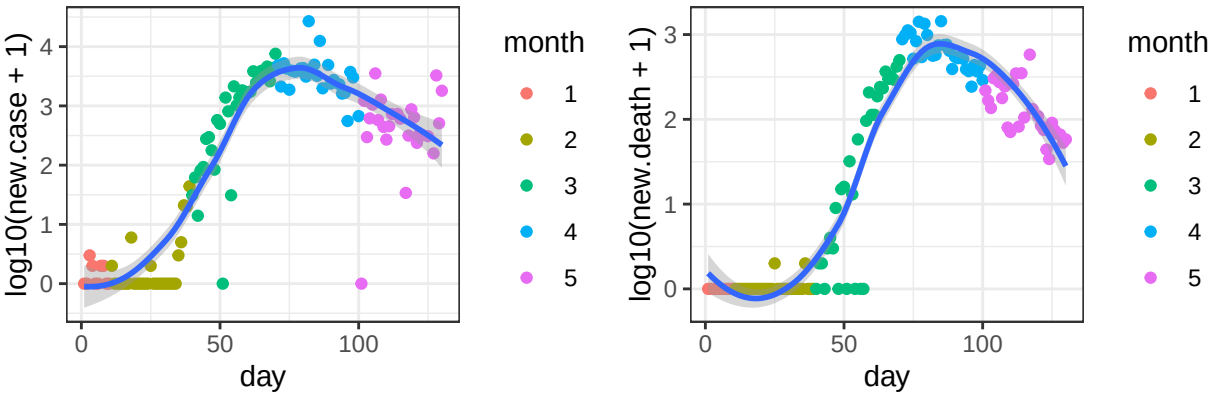
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Brazil



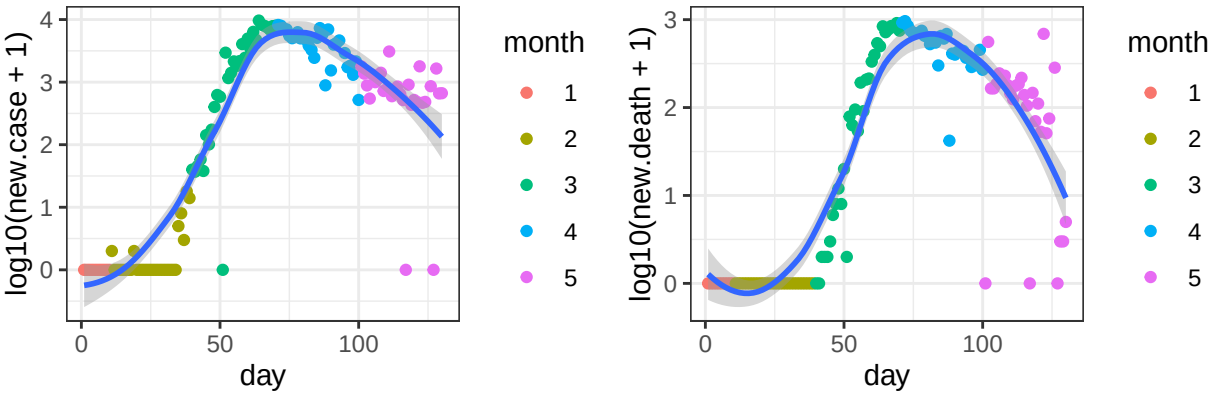
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



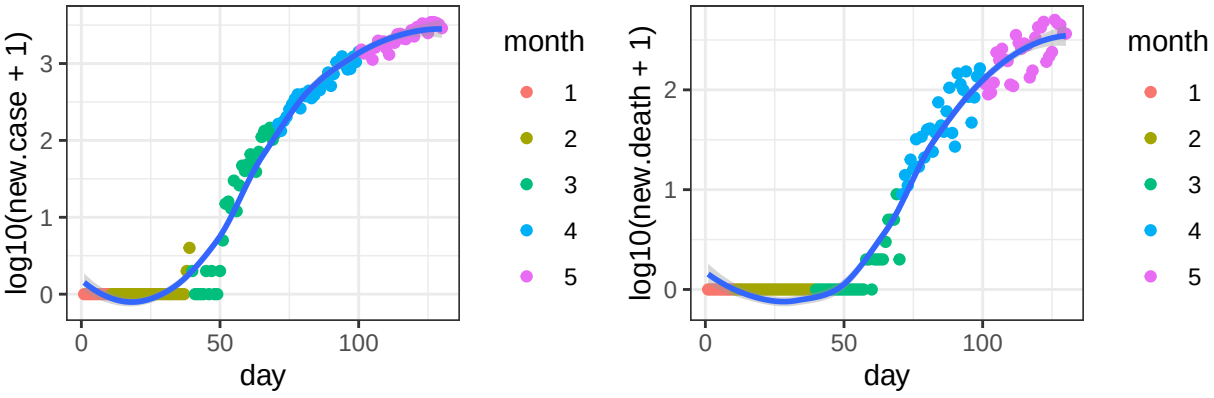
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain



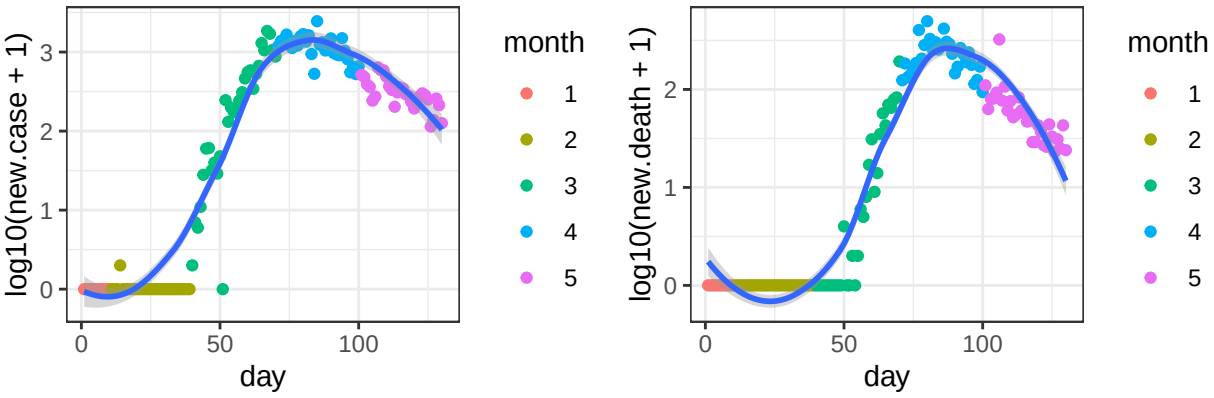
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Mexico



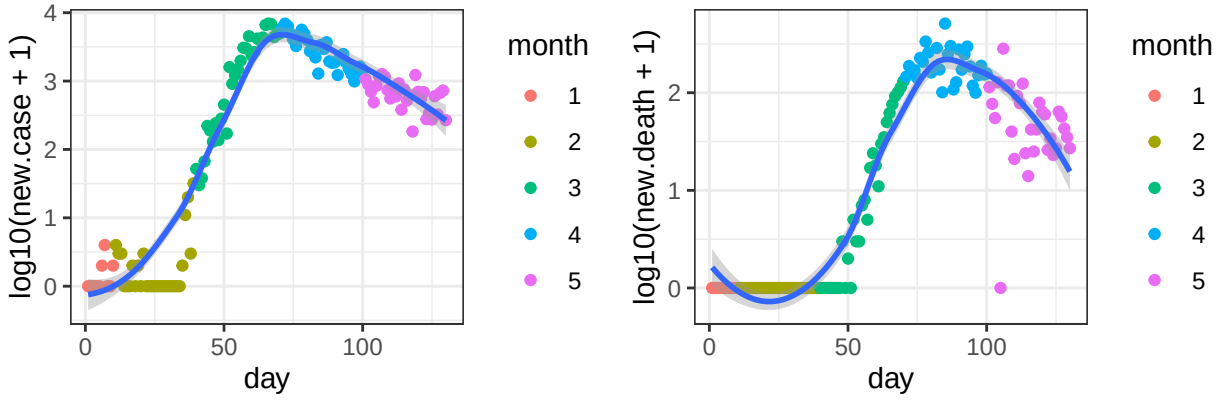
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Belgium



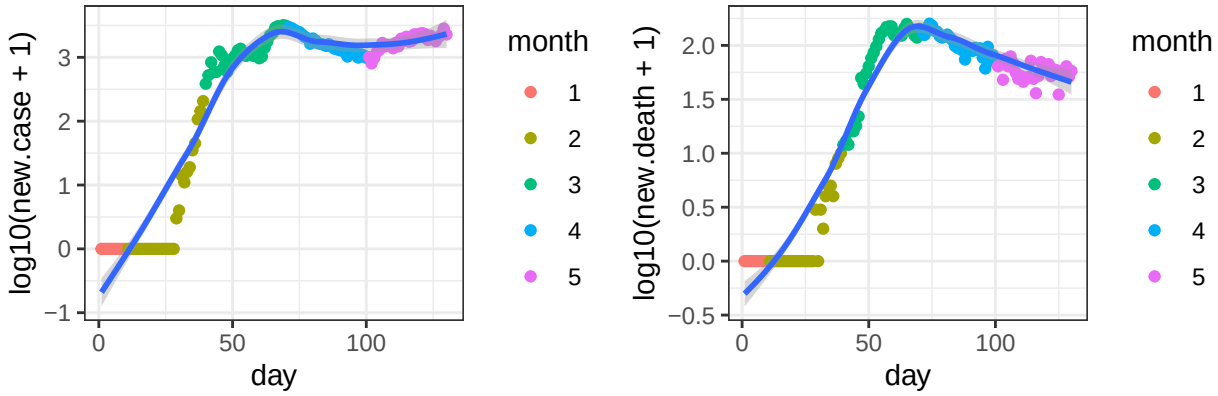
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Germany



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

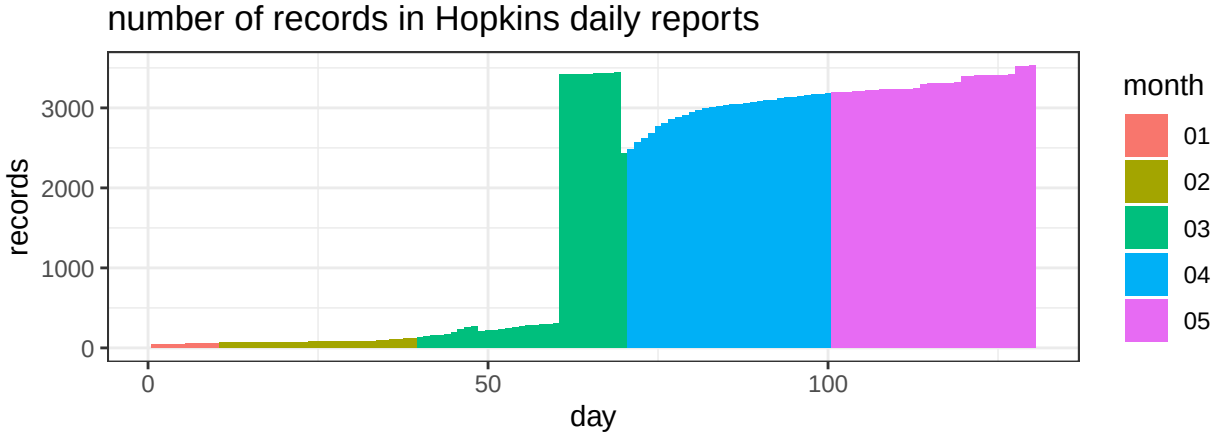
Iran



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

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] "2020-05-29"
```

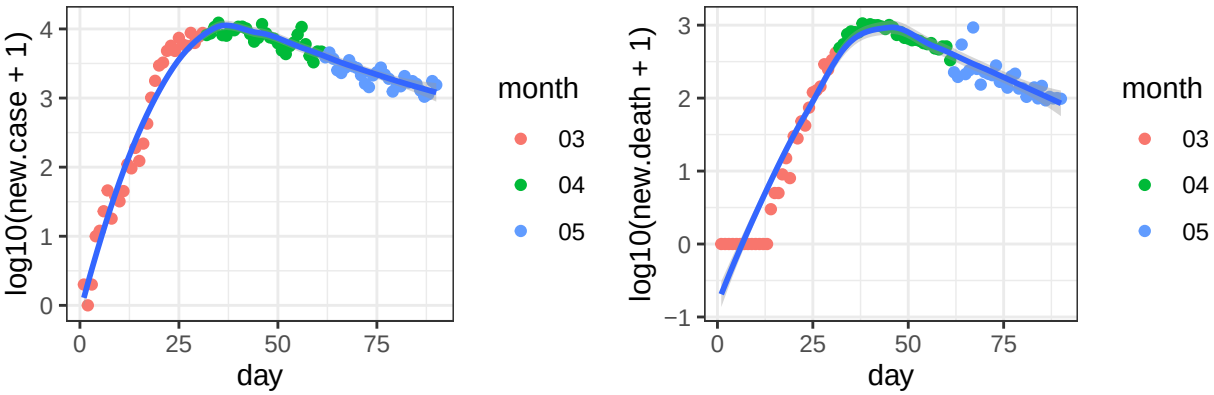
state level data

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

```
##            date                state fips  cases deaths
## 4833 2020-05-29             New York   36 373108  29535
## 4831 2020-05-29           New Jersey   34 158844  11531
## 4822 2020-05-29        Massachusetts   25  95512   6718
## 4840 2020-05-29         Pennsylvania   42  75073   5467
## 4823 2020-05-29             Michigan   26  56589   5406
## 4814 2020-05-29             Illinois   17 117794   5308
## 4804 2020-05-29           California    6 107043   4144
## 4806 2020-05-29          Connecticut    9  41762   3868
## 4819 2020-05-29            Louisiana   22  38907   2766
## 4821 2020-05-29             Maryland   24  51631   2466
## 4809 2020-05-29              Florida   12  54489   2412
## 4837 2020-05-29                 Ohio   39  34566   2131
## 4815 2020-05-29              Indiana   18  34399   2110
## 4810 2020-05-29              Georgia   13  43888   1953
## 4846 2020-05-29                Texas   48  62062   1645
## 4805 2020-05-29             Colorado    8  25602   1437
## 4850 2020-05-29             Virginia   51  42533   1358
## 4851 2020-05-29           Washington   53  22109   1120
## 4824 2020-05-29            Minnesota   27  23541   1006
## 4834 2020-05-29       North Carolina   37  26735    888
## 4802 2020-05-29              Arizona    4  18465    885
## 4826 2020-05-29             Missouri   29  12949    740
## 4825 2020-05-29          Mississippi   28  14790    710
## 4842 2020-05-29         Rhode Island   44  14635    693
## 4800 2020-05-29              Alabama    1  17031    610
## 4853 2020-05-29            Wisconsin   55  17784    569
## 4816 2020-05-29                 Iowa   19  19019    525
## 4843 2020-05-29       South Carolina   45  11131    483
## 4808 2020-05-29 District of Columbia   11   8538    460
## 4818 2020-05-29             Kentucky   21   9688    434
```

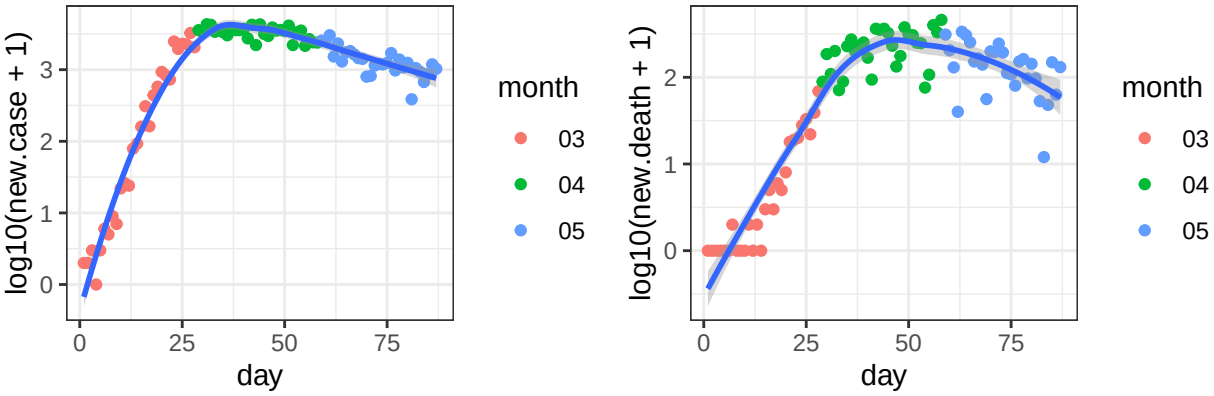
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



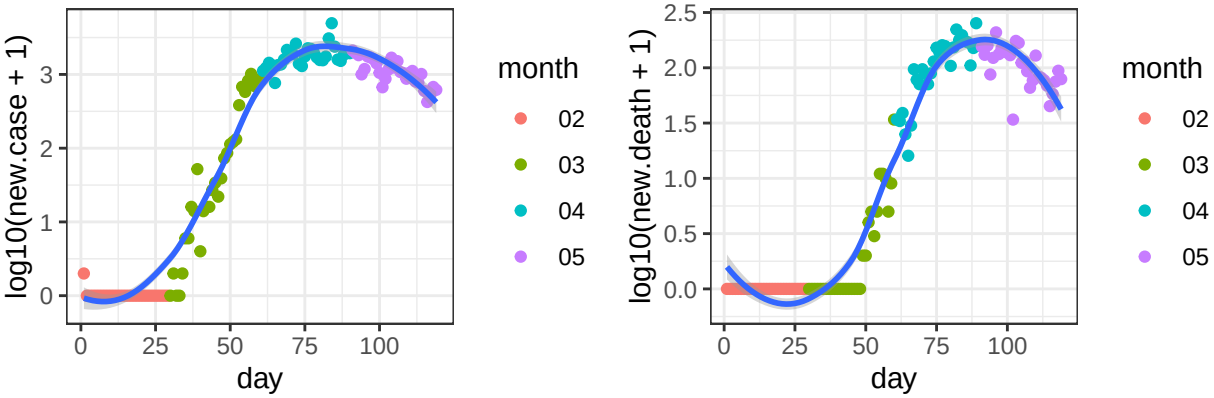
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



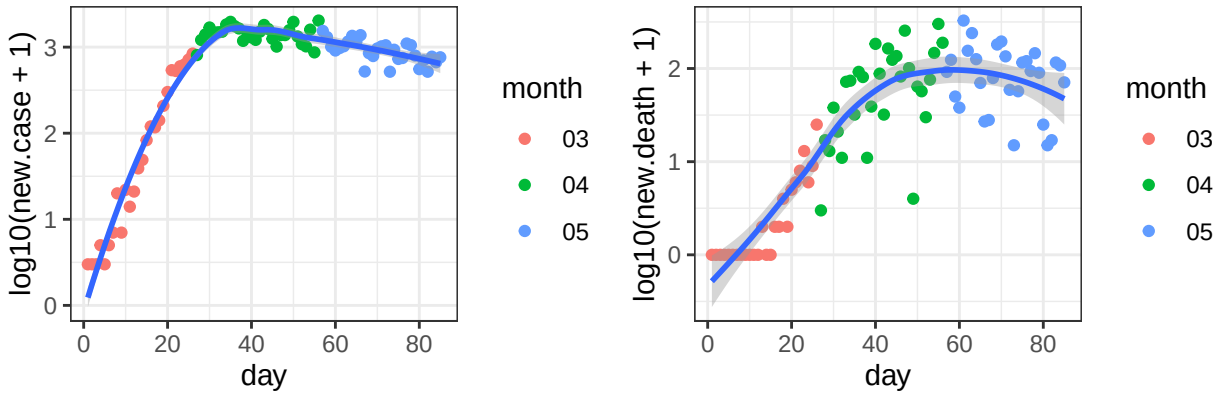
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Massachusetts



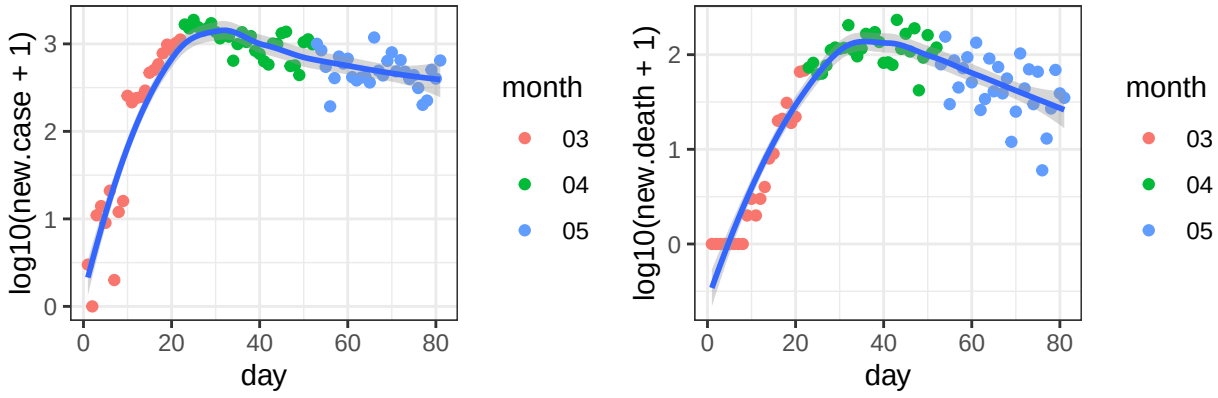
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Pennsylvania



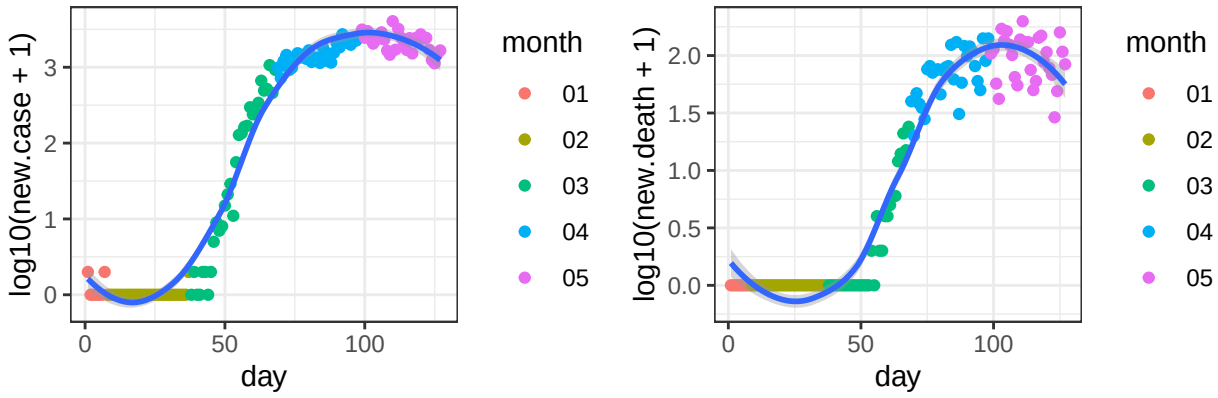
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Michigan



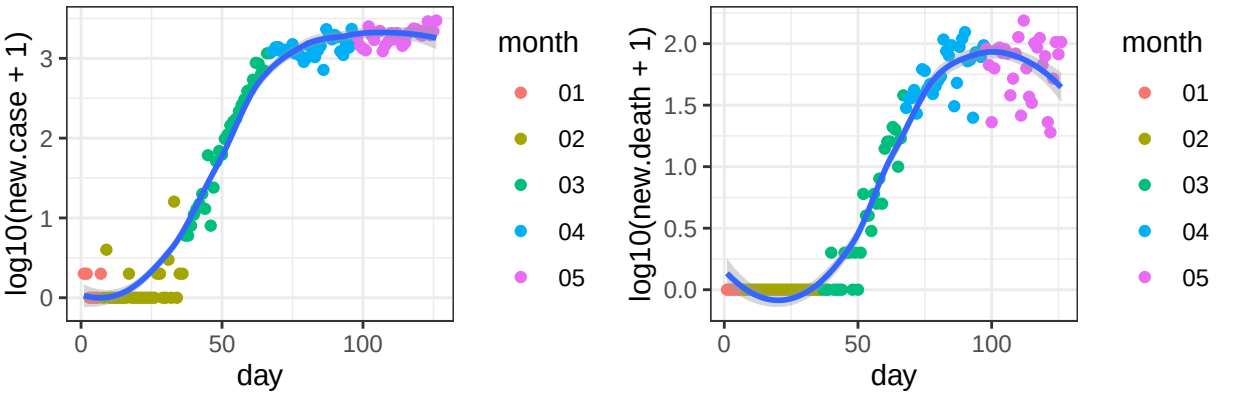
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Illinois



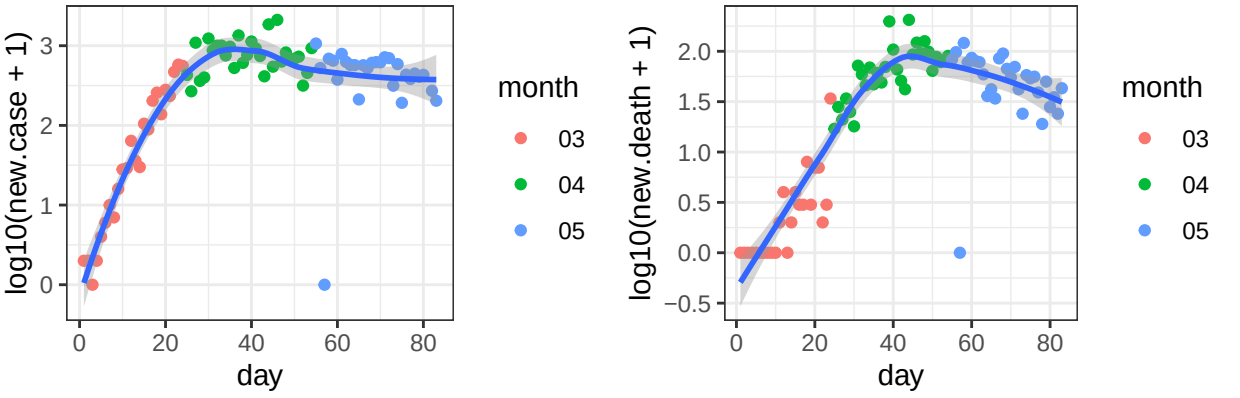
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

California



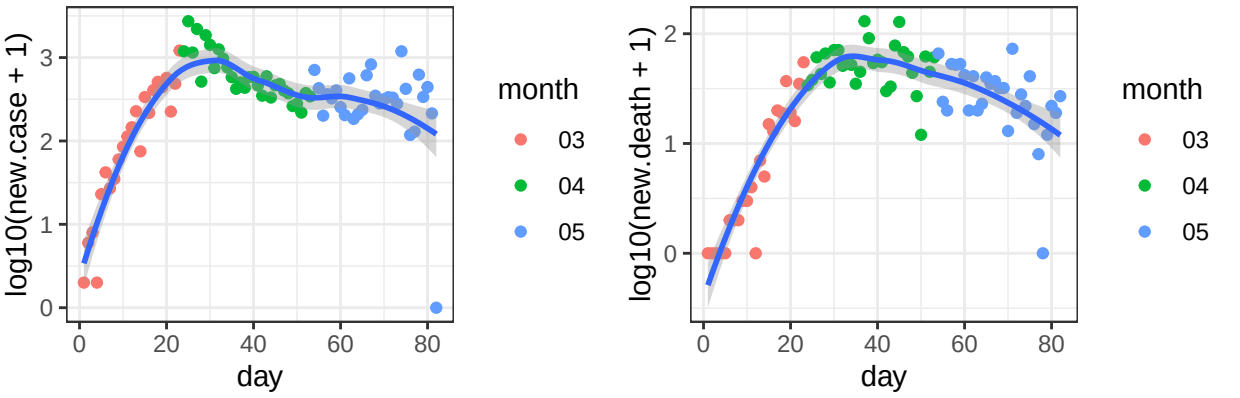
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Connecticut



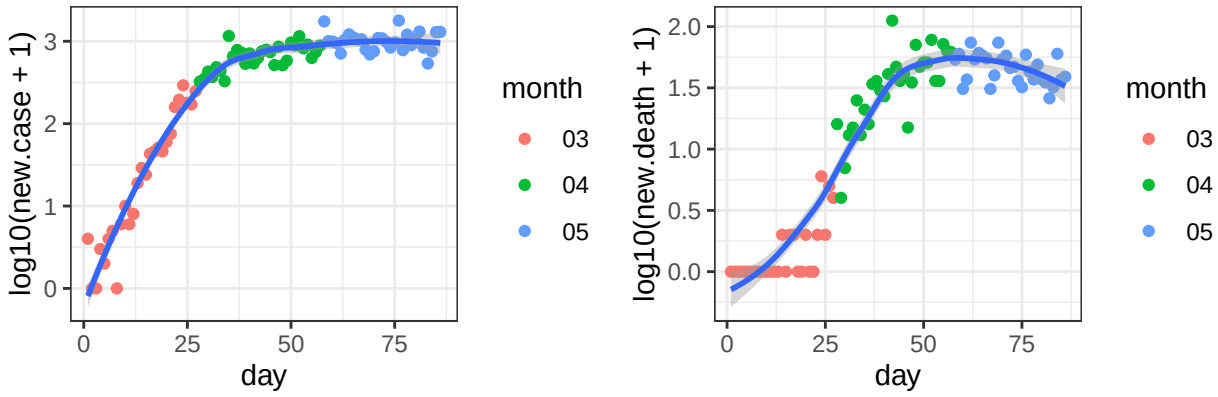
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Louisiana



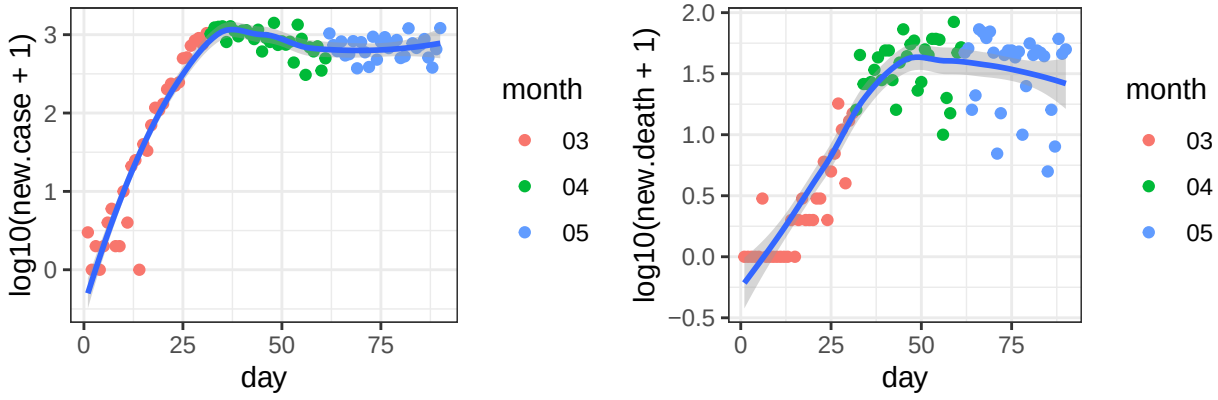
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



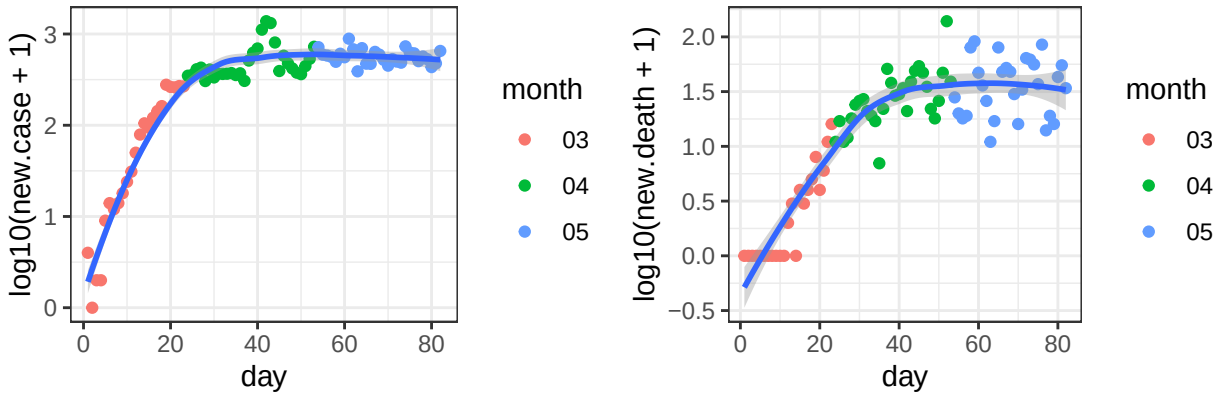
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Florida



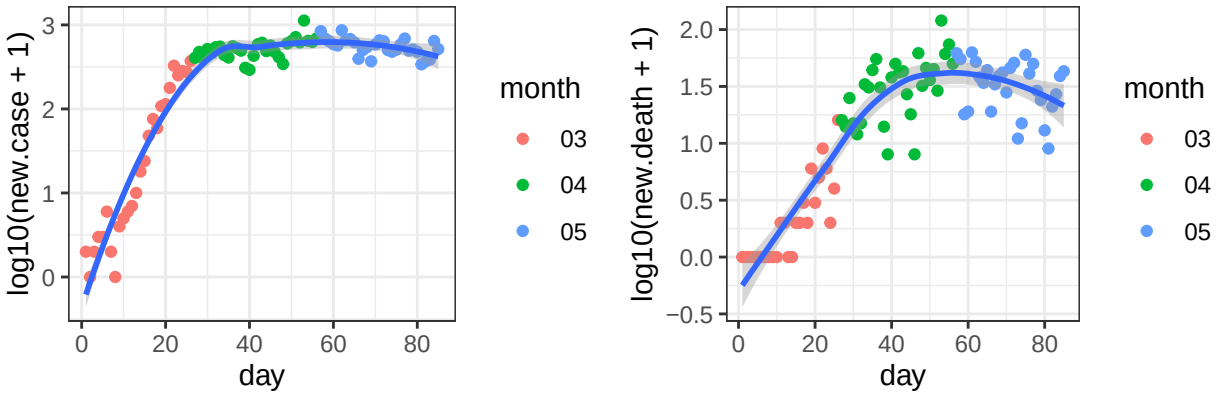
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Ohio



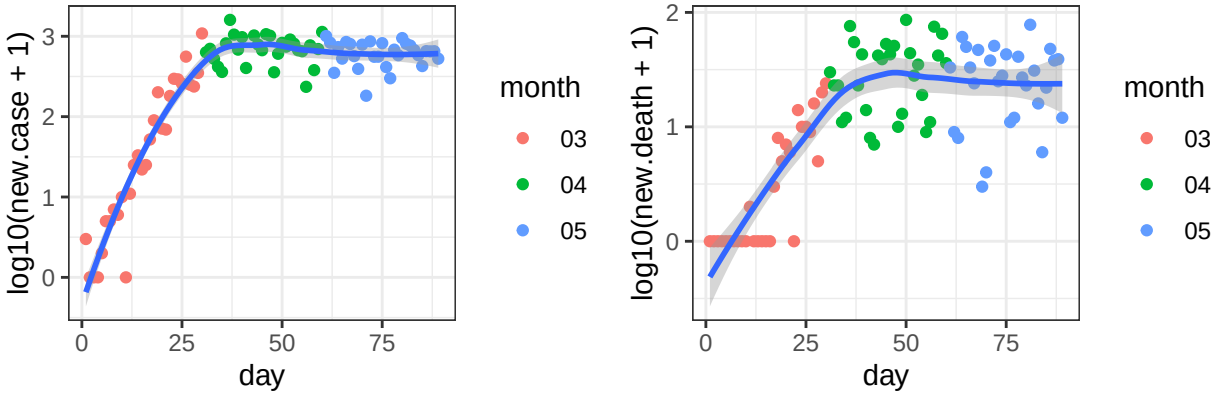
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Indiana



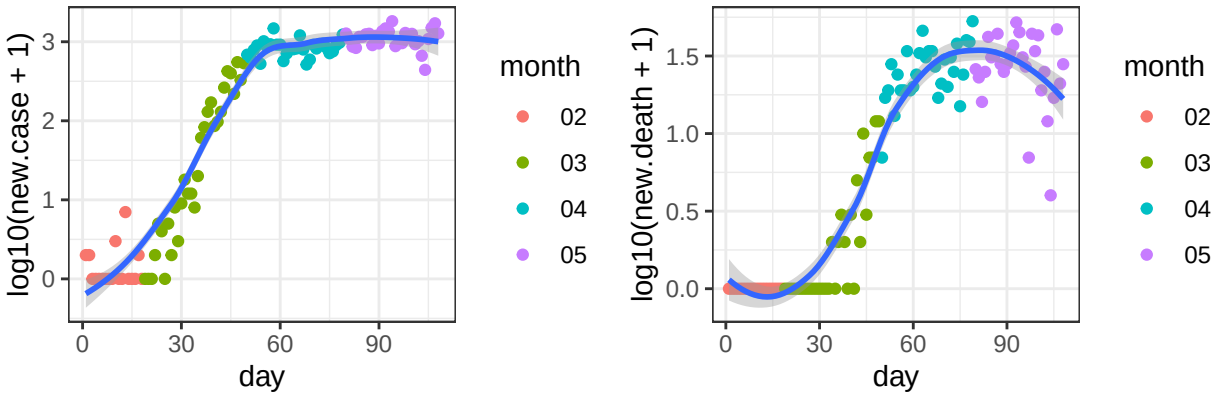
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Georgia



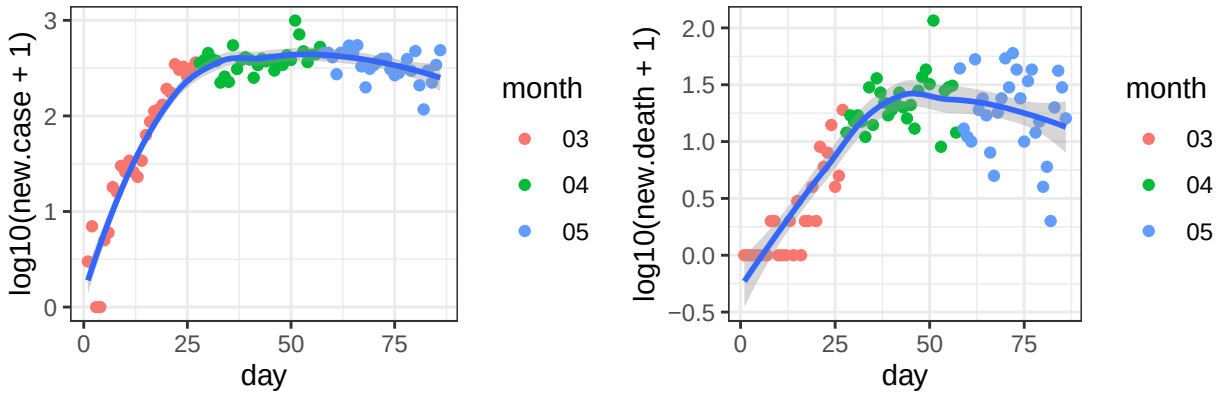
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Texas



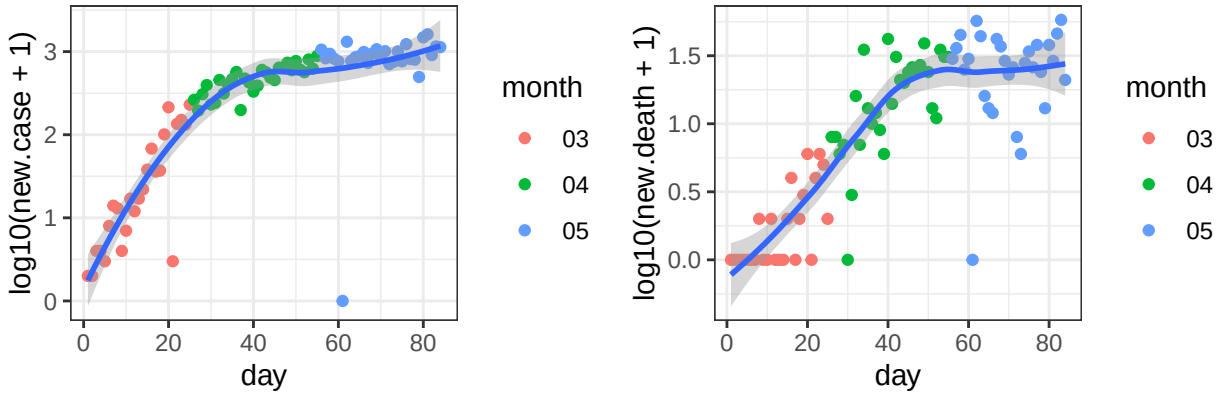
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Colorado



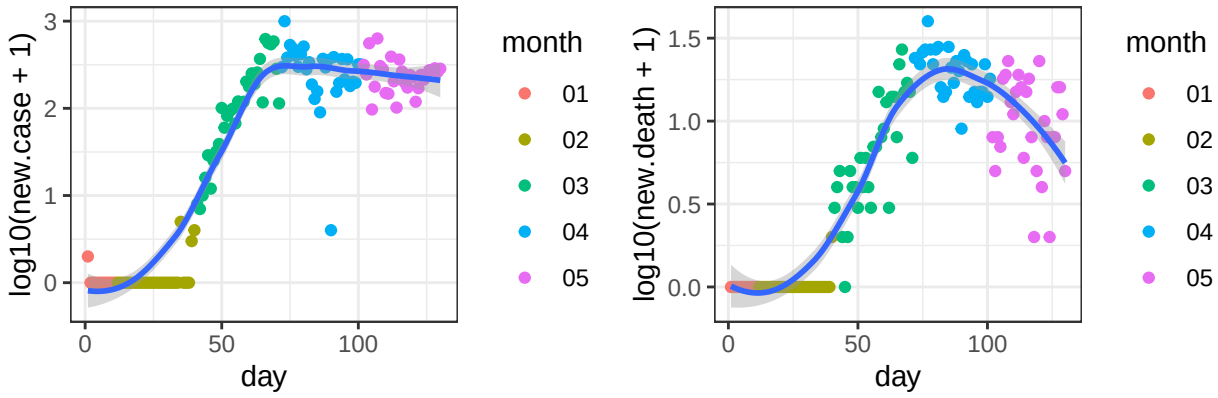
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Virginia



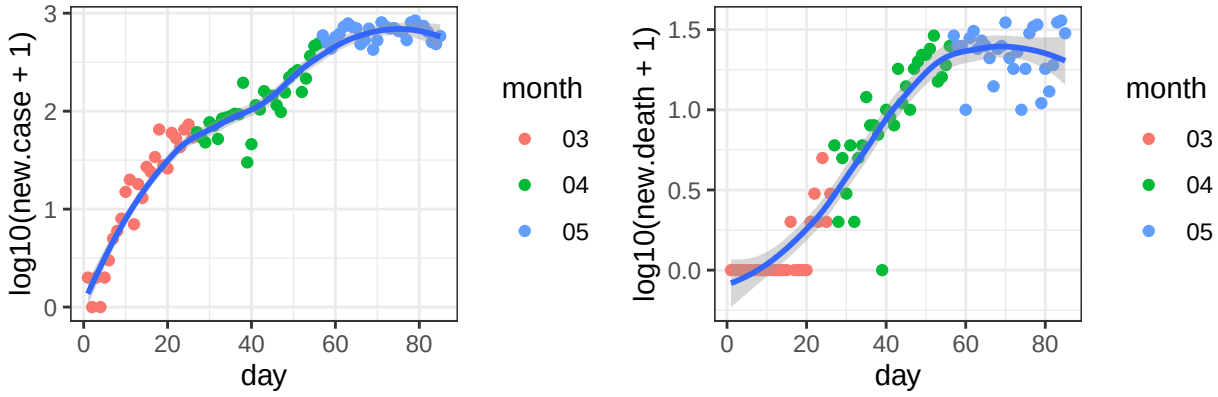
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Washington



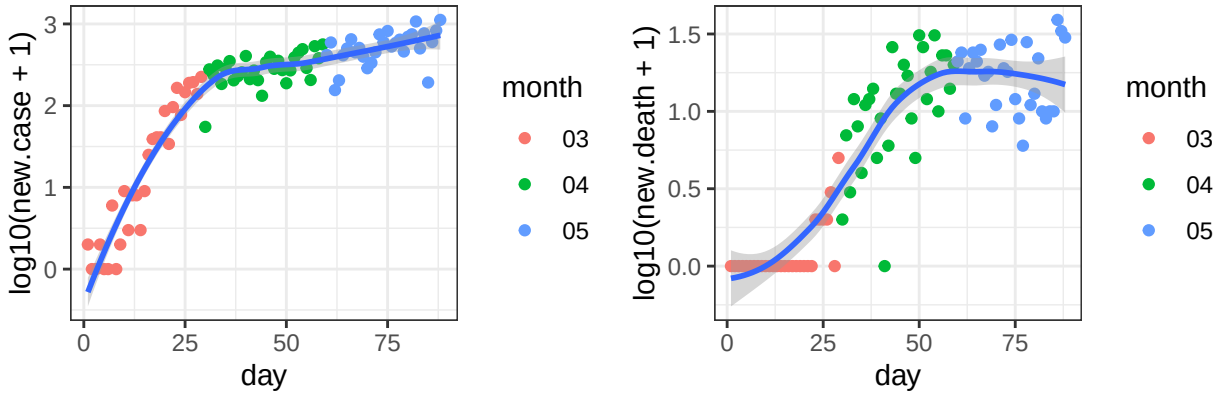
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Minnesota



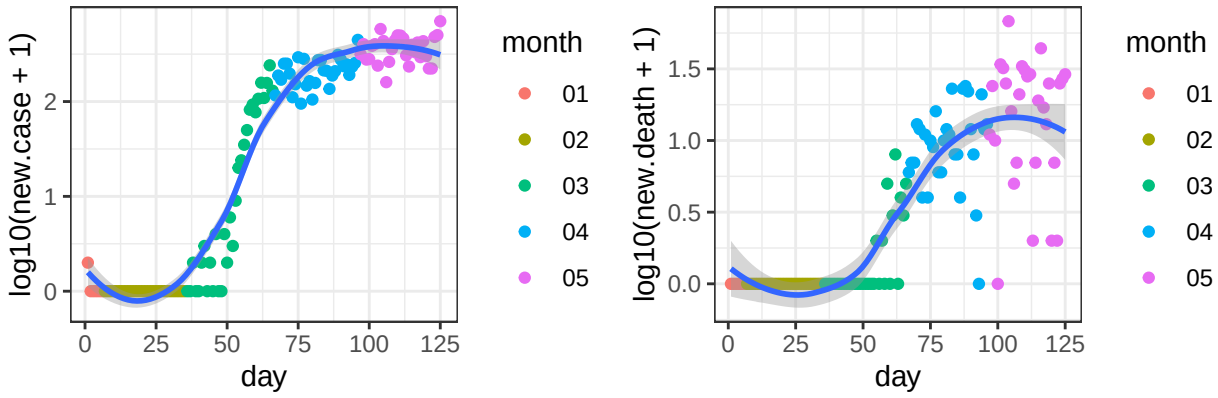
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

North Carolina



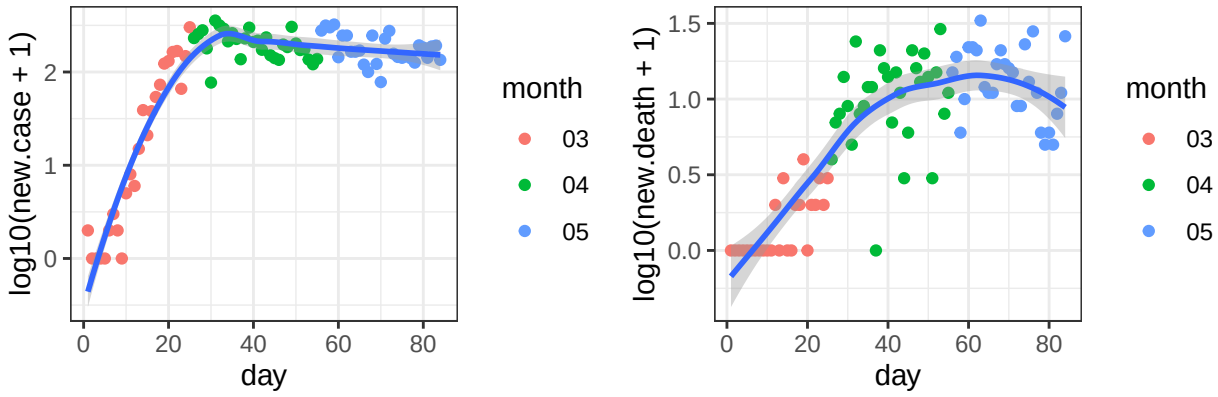
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Arizona



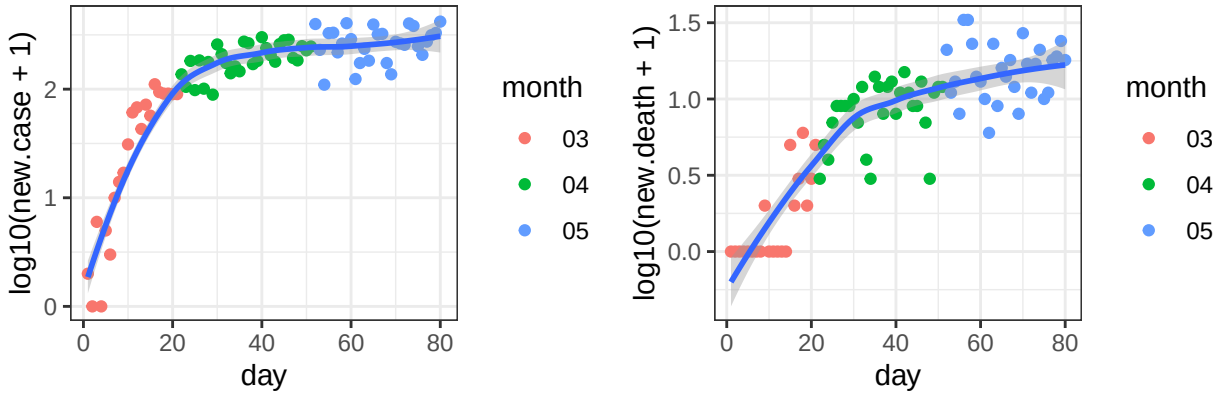
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Missouri



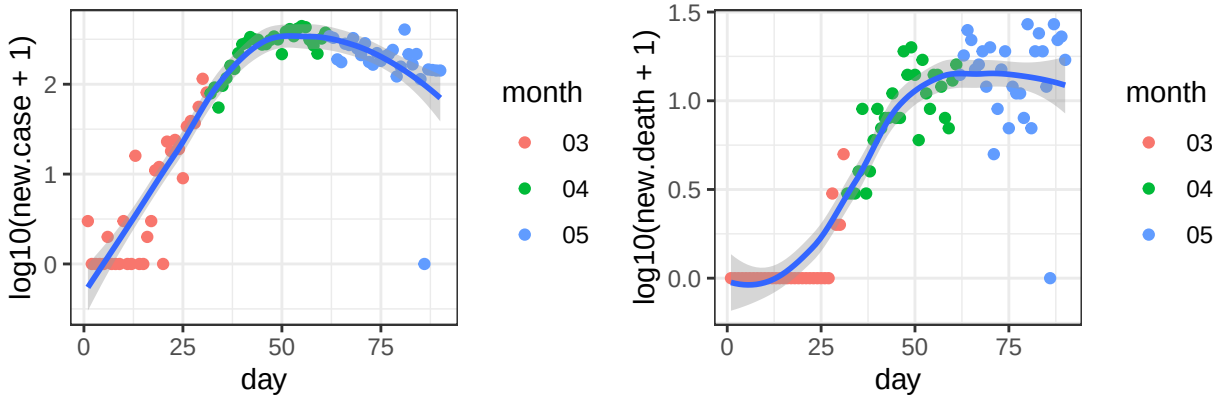
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Mississippi



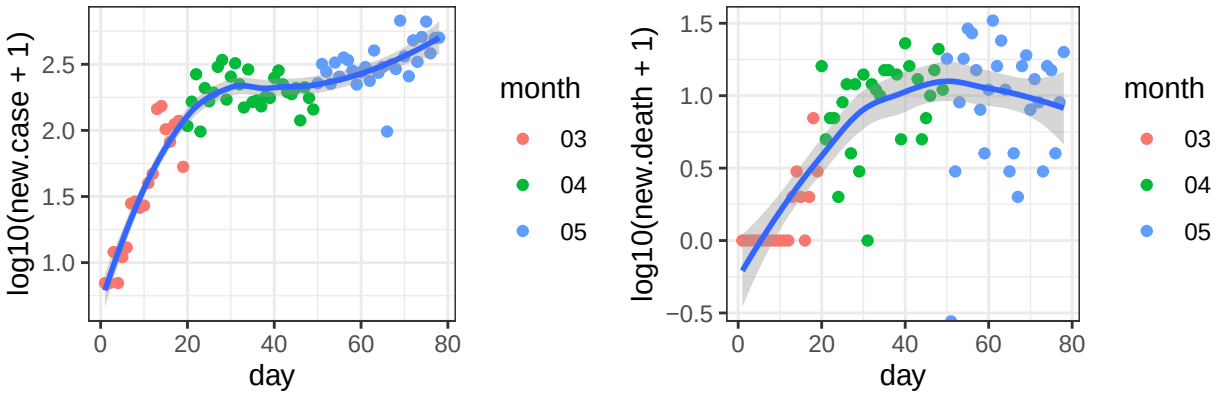
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Rhode Island



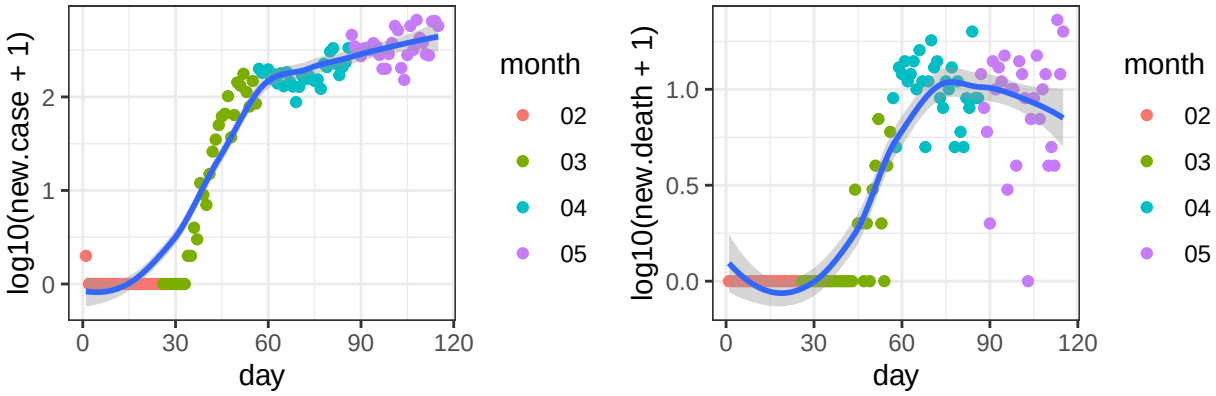
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Alabama



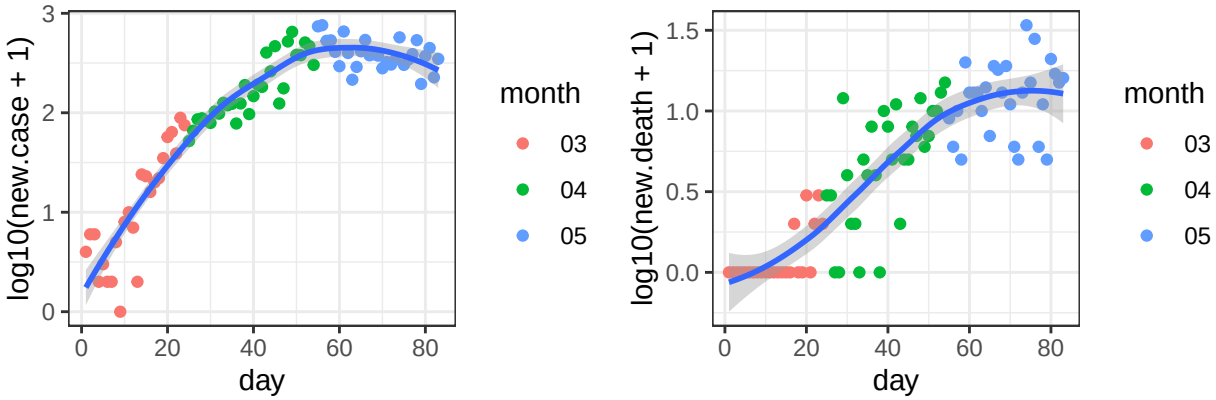
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Wisconsin



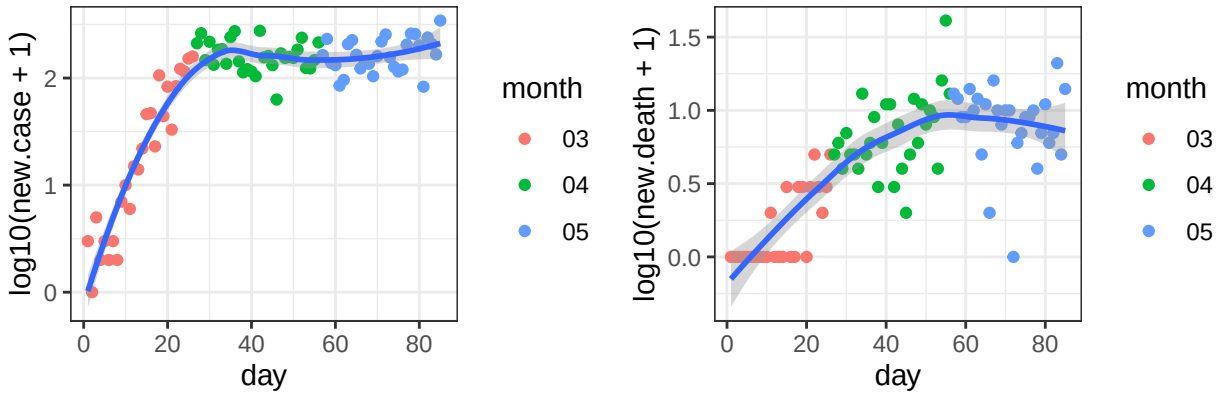
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

Iowa



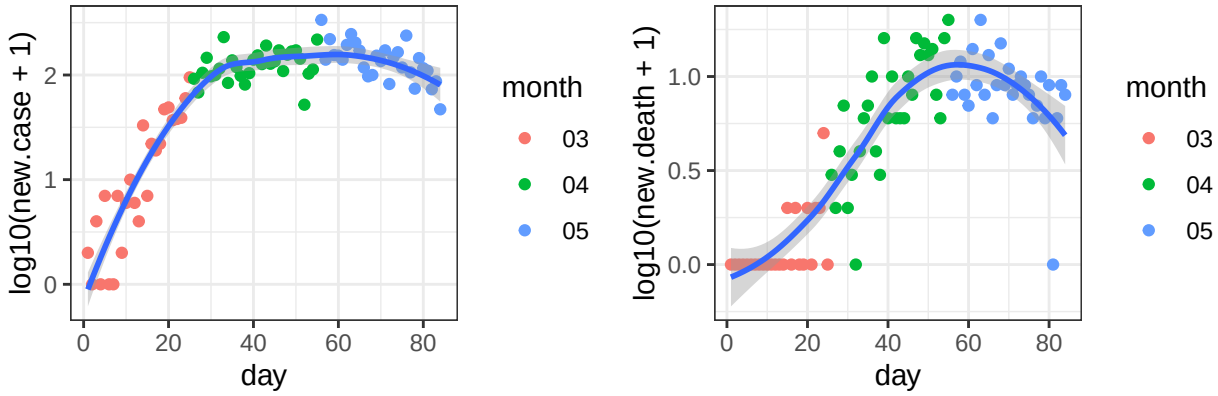
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

South Carolina



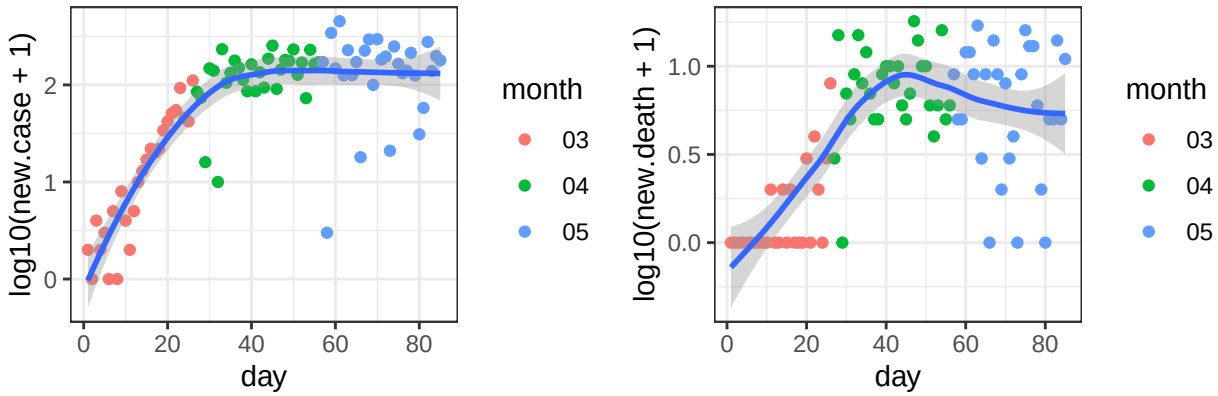
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

District of Columbia



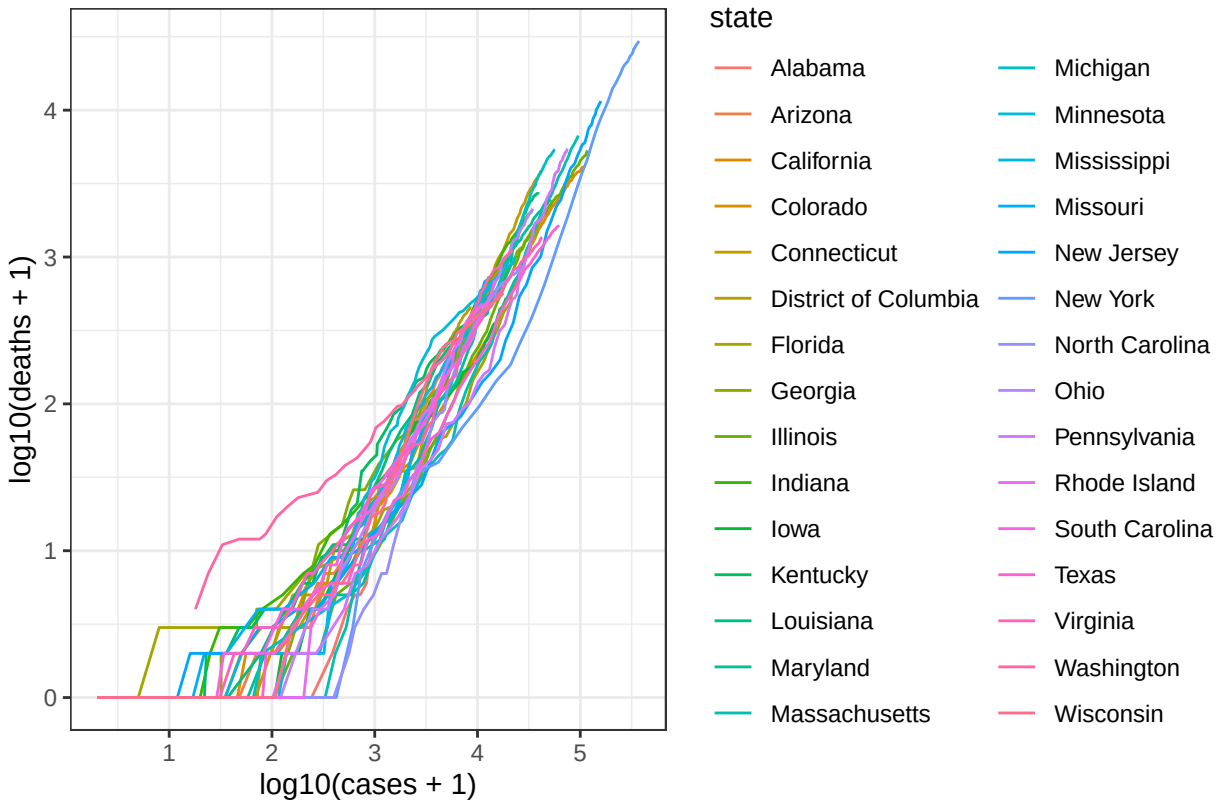
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

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

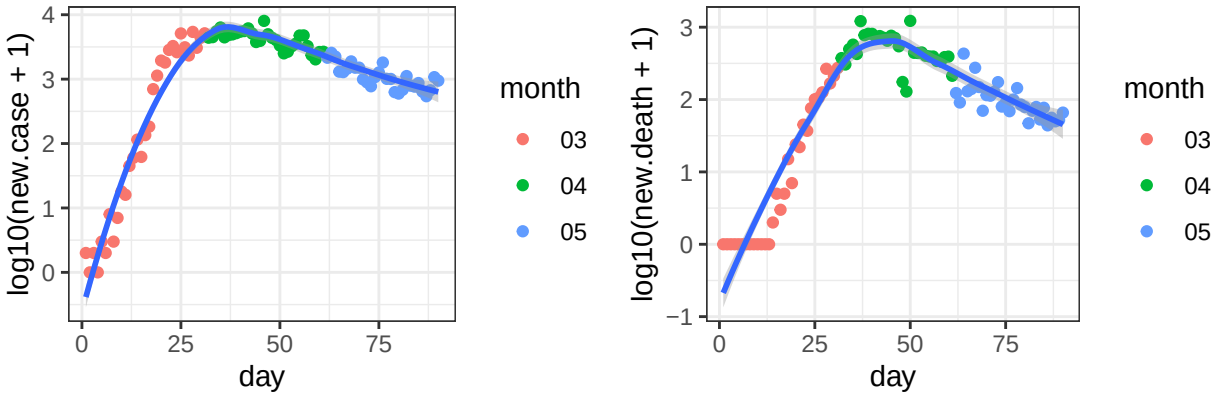
First check the 30 counties with the largest number of deaths.

```
##              date        county         state  fips  cases deaths
## 187406 2020-05-29 New York City      New York    NA 206800  20960
## 186238 2020-05-29          Cook      Illinois 17031  76266   3570
## 187405 2020-05-29        Nassau      New York 36059  40226   2611
## 186922 2020-05-29         Wayne      Michigan 26163  20227   2425
## 185842 2020-05-29   Los Angeles    California  6037  51562   2290
## 187425 2020-05-29       Suffolk      New York 36103  39445   1928
## 187331 2020-05-29         Essex    New Jersey 34013  17546   1647
## 186837 2020-05-29     Middlesex Massachusetts 25017  20972   1583
## 187326 2020-05-29        Bergen    New Jersey 34003  18223   1567
## 187433 2020-05-29   Westchester      New York 36119  33348   1484
## 187826 2020-05-29  Philadelphia  Pennsylvania 42101  22405   1300
## 185941 2020-05-29     Fairfield   Connecticut  9001  15409   1257
## 185942 2020-05-29      Hartford   Connecticut  9003  10146   1222
## 187333 2020-05-29        Hudson    New Jersey 34017  18287   1168
## 187344 2020-05-29         Union    New Jersey 34039  15610   1060
## 186903 2020-05-29       Oakland      Michigan 26125   8311    975
## 187336 2020-05-29     Middlesex    New Jersey 34023  15734    972
## 185945 2020-05-29     New Haven   Connecticut  9009  11241    957
## 187340 2020-05-29       Passaic    New Jersey 34031  16045    917
## 186833 2020-05-29         Essex Massachusetts 25009  13994    906
## 186841 2020-05-29       Suffolk Massachusetts 25025  17786    861
## 186839 2020-05-29       Norfolk Massachusetts 25021   7959    811
```

```
## 186890 2020-05-29        Macomb      Michigan 26099   6616    793
## 186843 2020-05-29     Worcester Massachusetts 25027  10816    746
## 187339 2020-05-29         Ocean    New Jersey 34029   8627    721
## 185997 2020-05-29    Miami-Dade       Florida 12086  17640    685
## 187821 2020-05-29    Montgomery  Pennsylvania 42091   6906    677
## 187338 2020-05-29        Morris    New Jersey 34027   6367    610
## 186372 2020-05-29        Marion       Indiana 18097   9720    609
## 186819 2020-05-29    Montgomery      Maryland 24031  11075    595
```

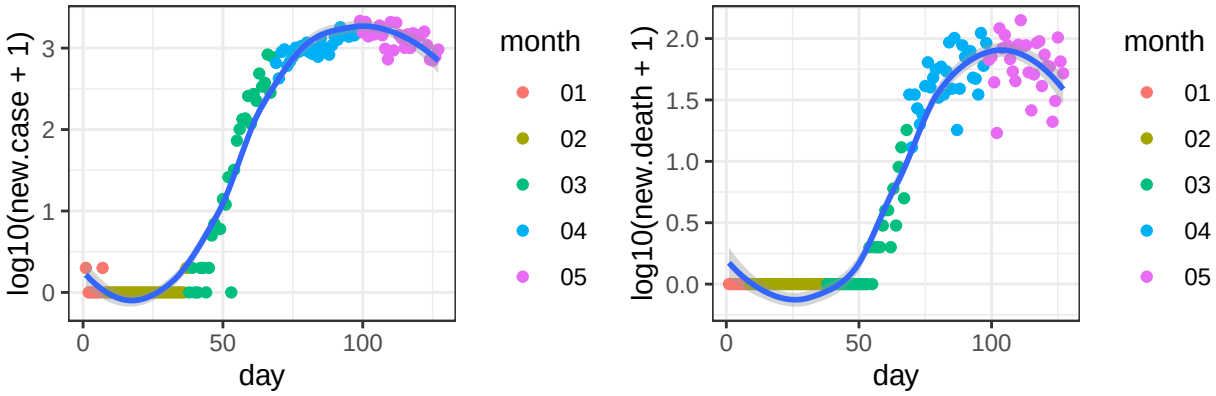
For these 30 counties, I check the number of new cases and the number of new deaths.

New York City_New York



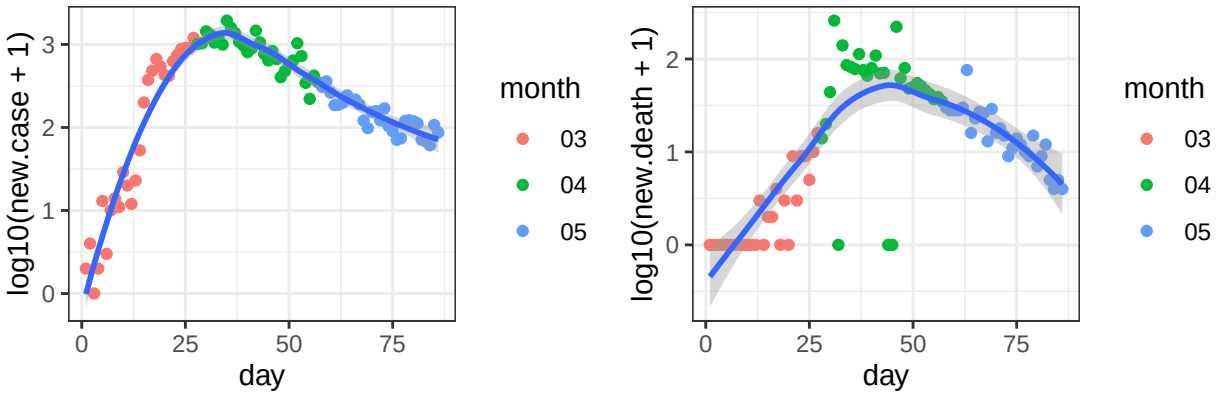
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



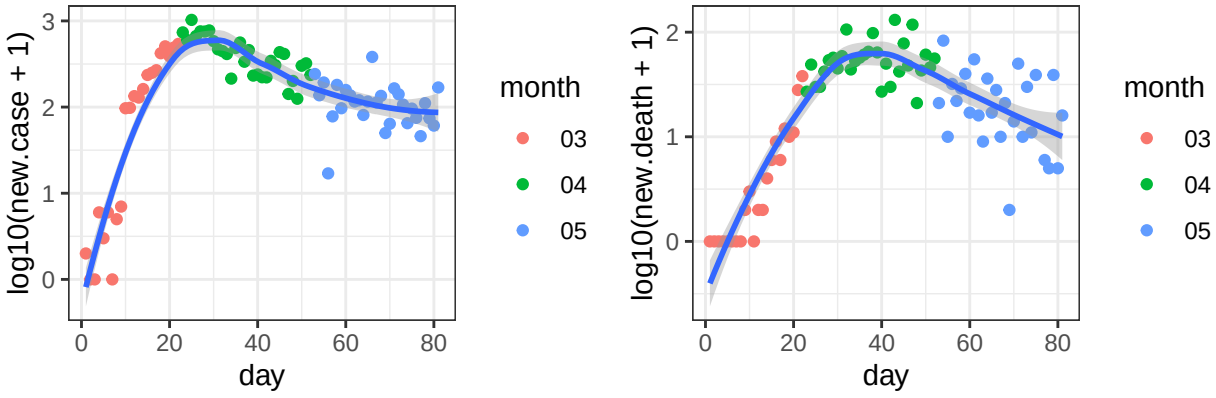
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Nassau_New York



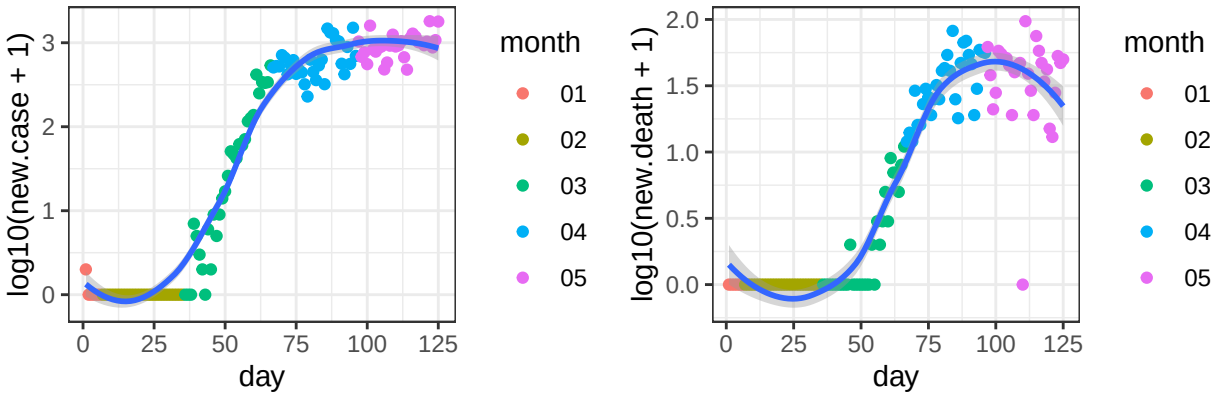
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Wayne_Michigan



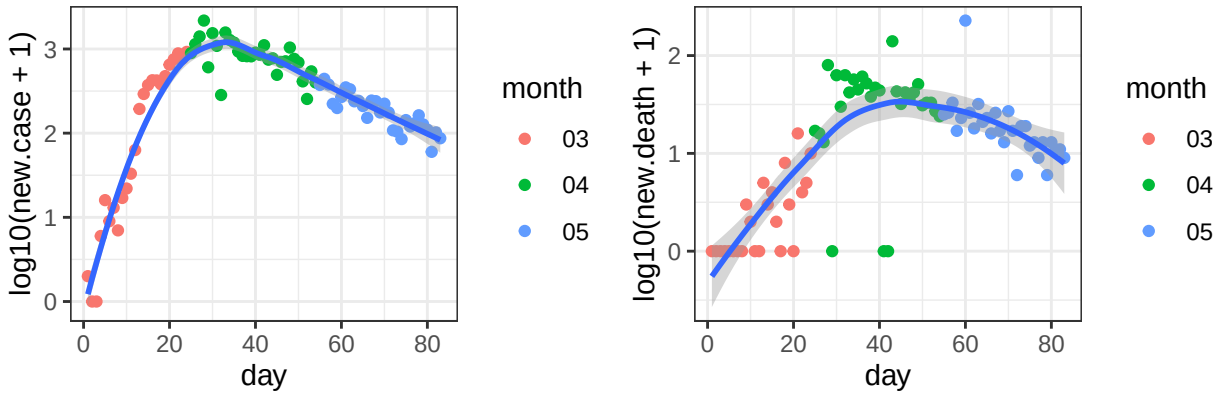
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Los Angeles_California



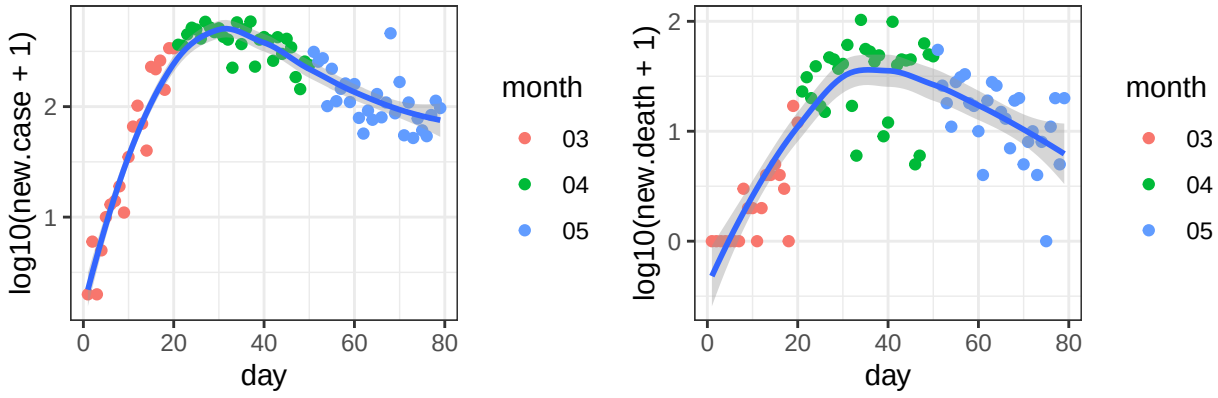
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Suffolk_New York



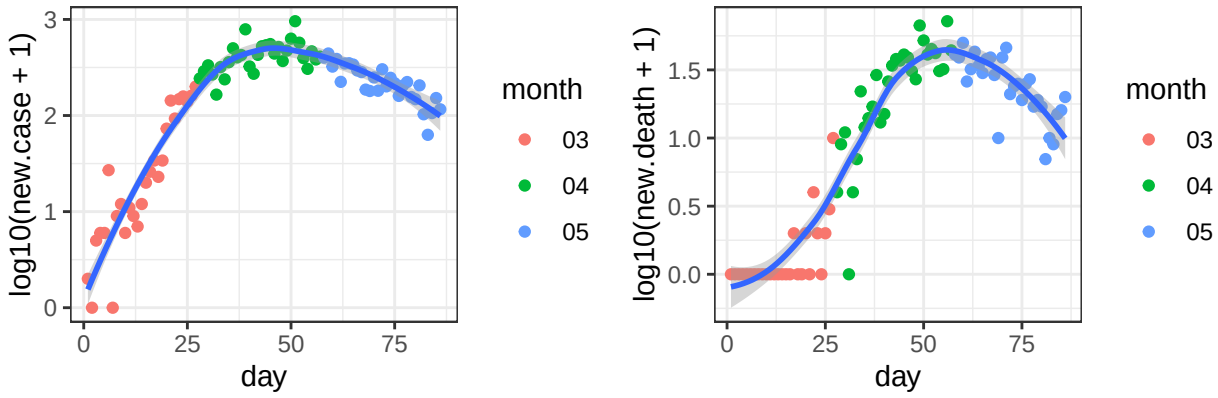
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Essex_New Jersey



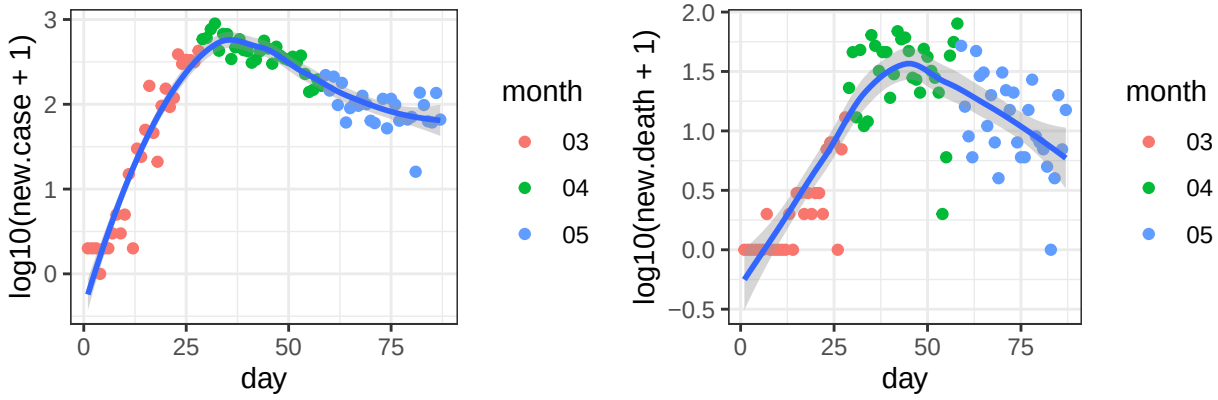
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Middlesex_Massachusetts



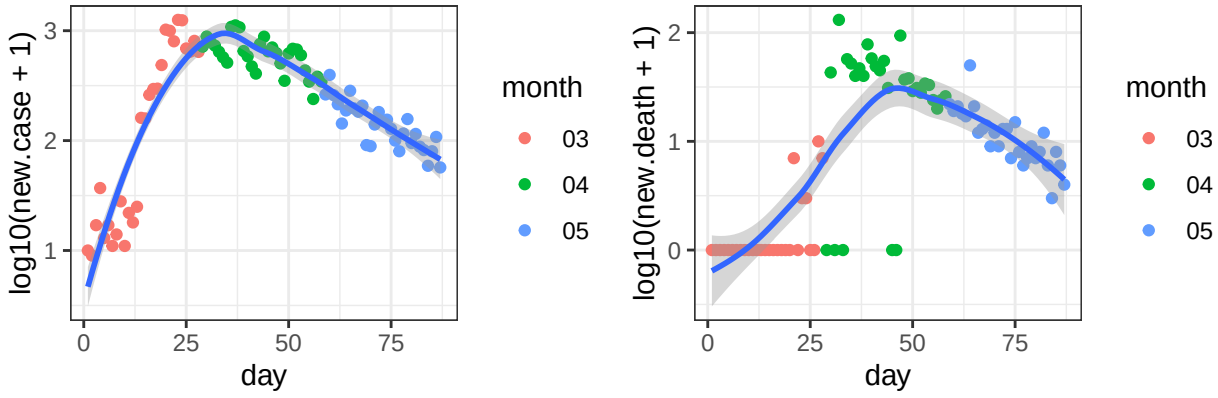
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Bergen_New Jersey



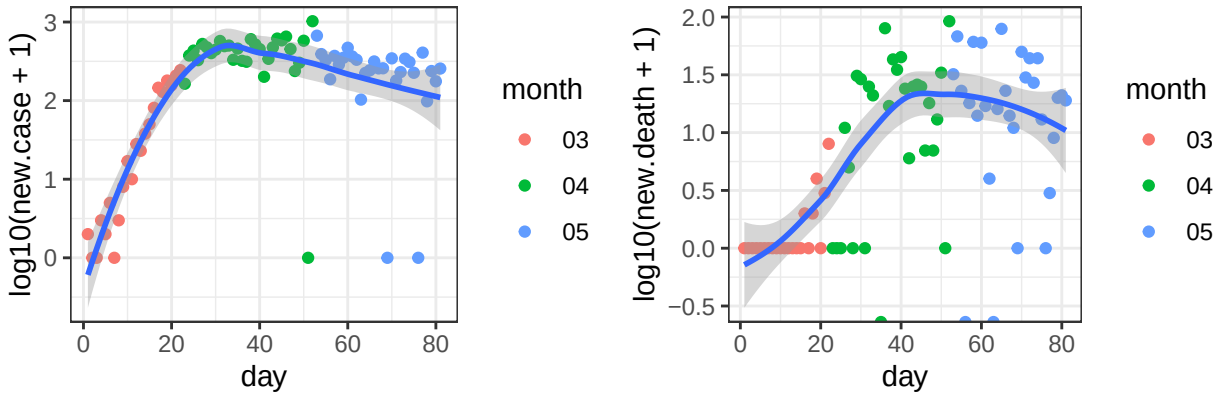
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Westchester_New York



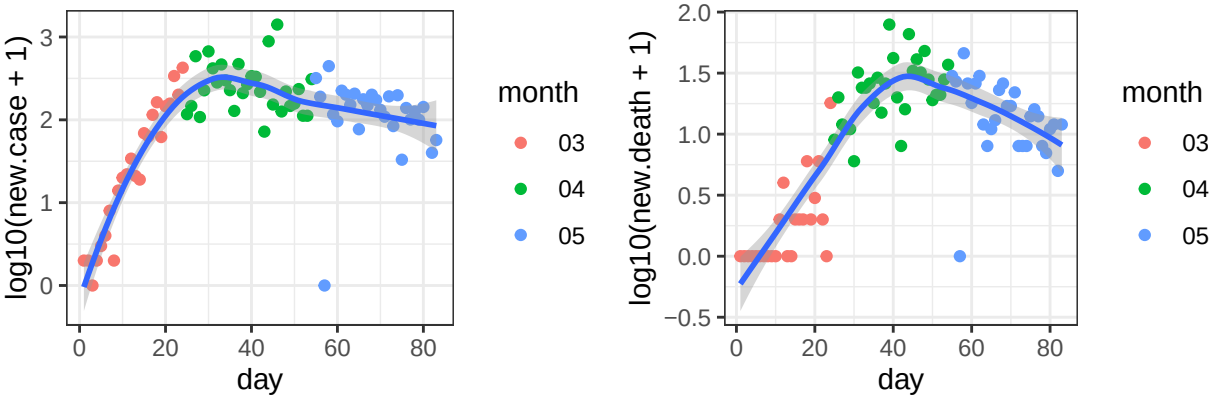
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Philadelphia_Pennsylvania



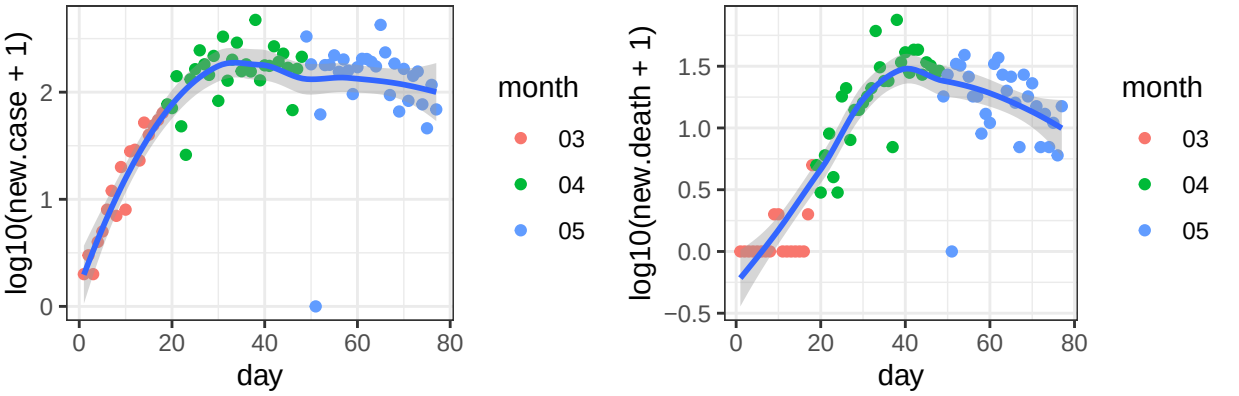
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Fairfield_Connecticut



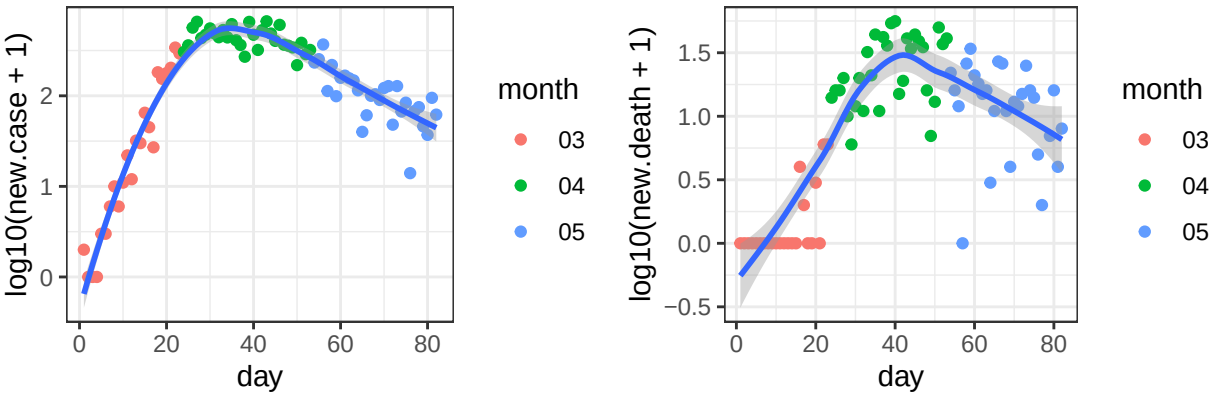
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Hartford_Connecticut



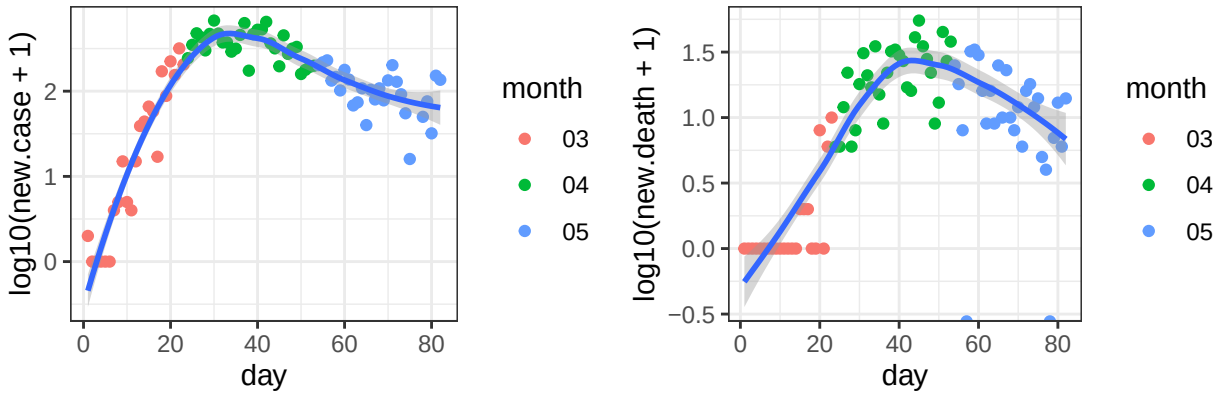
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Hudson_New Jersey



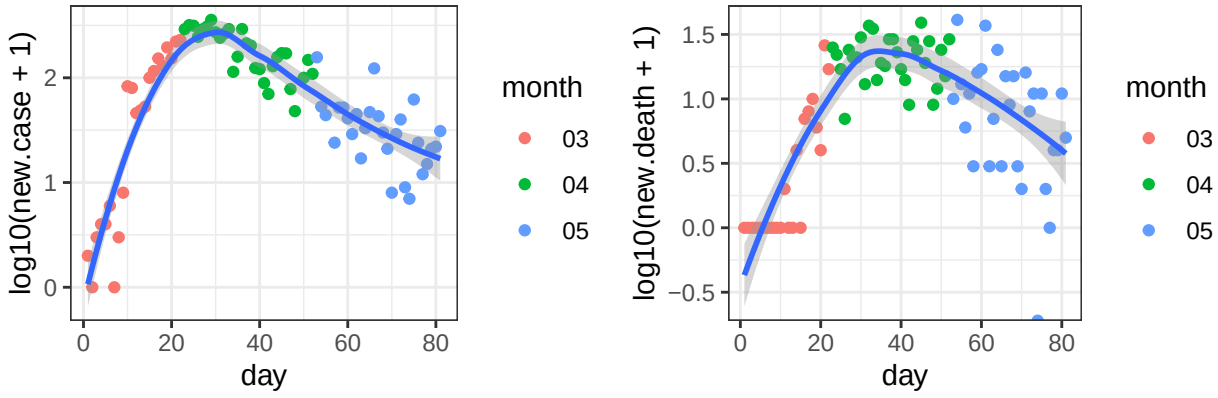
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Union_New Jersey



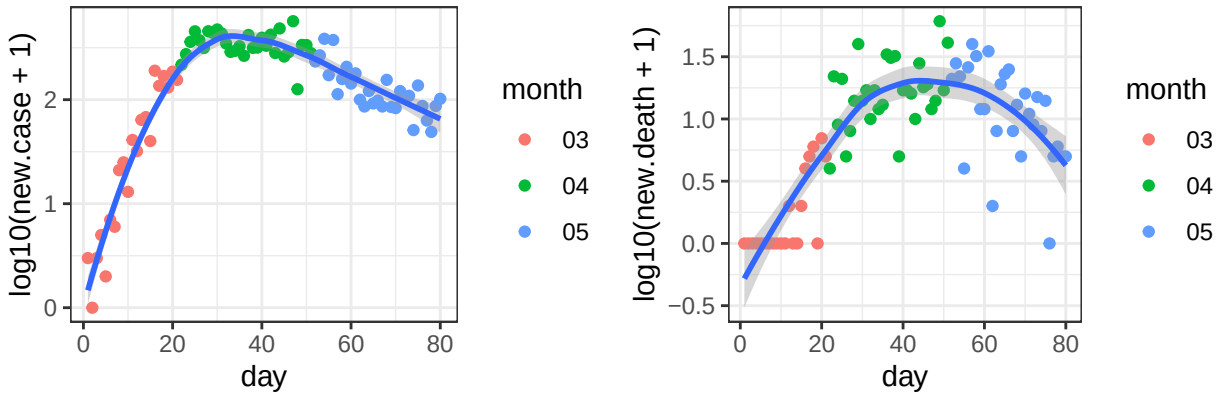
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Oakland_Michigan



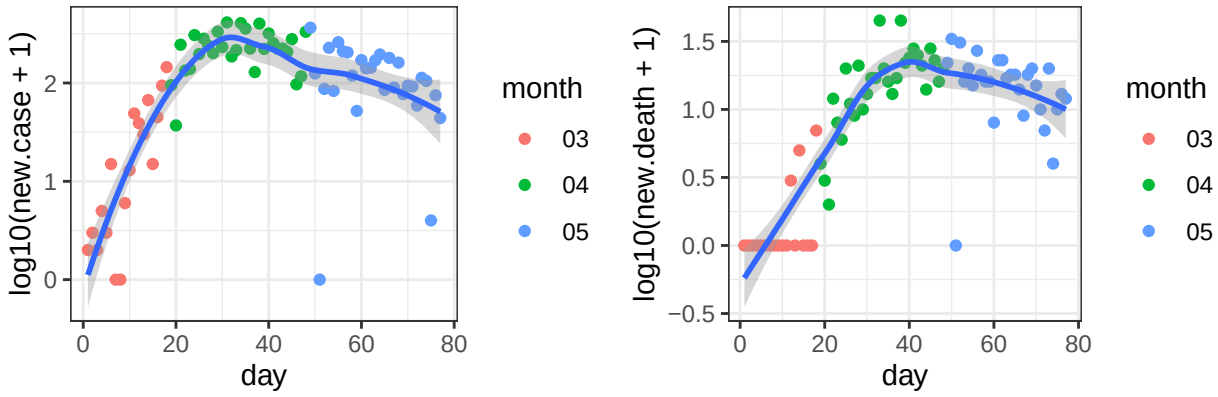
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Middlesex_New Jersey



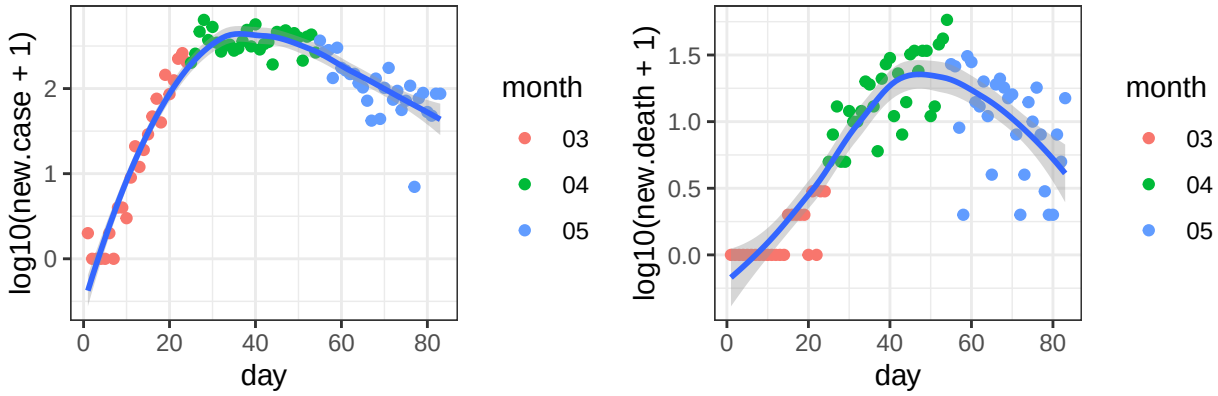
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

New Haven_Connecticut



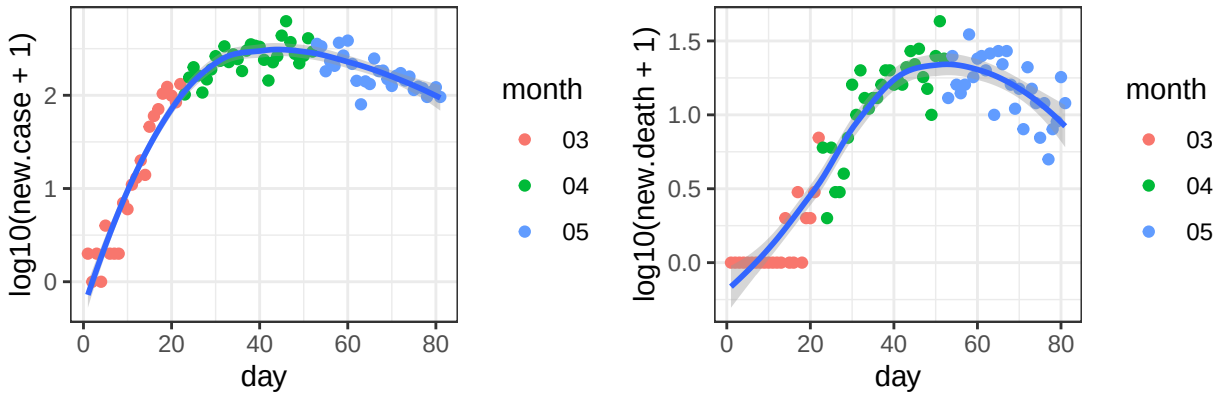
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Passaic_New Jersey



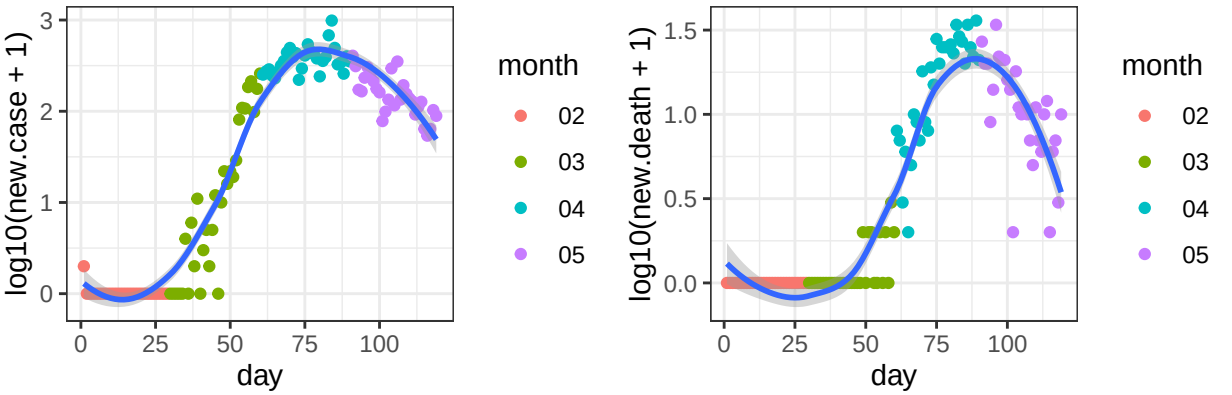
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Essex_Massachusetts



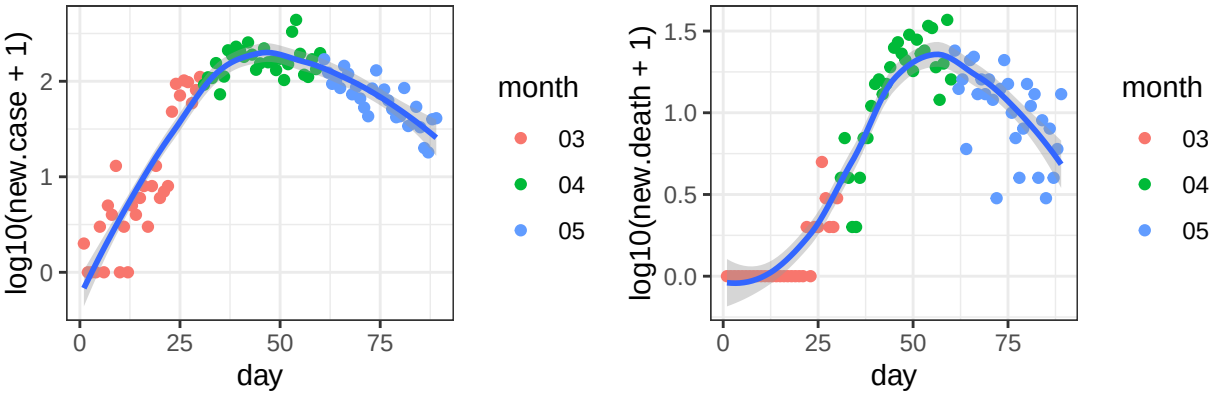
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Suffolk_Massachusetts



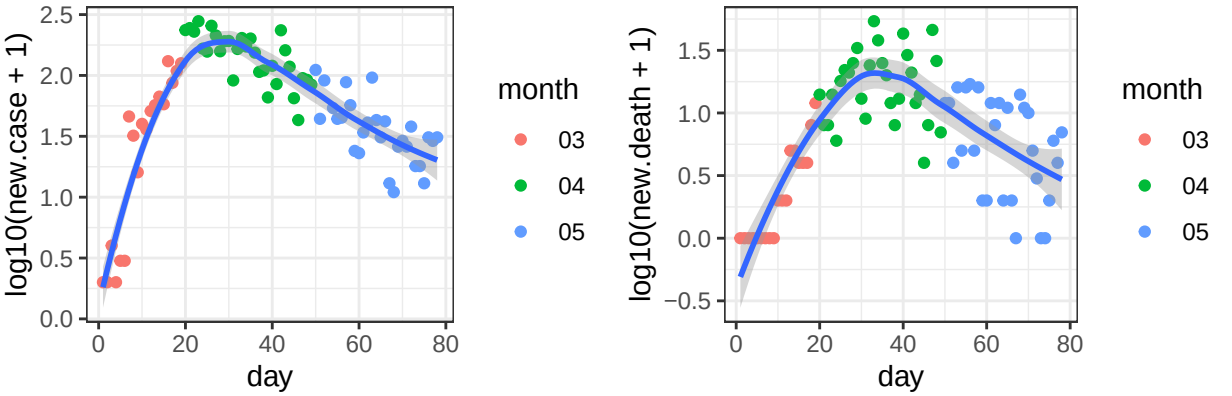
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Norfolk_Massachusetts



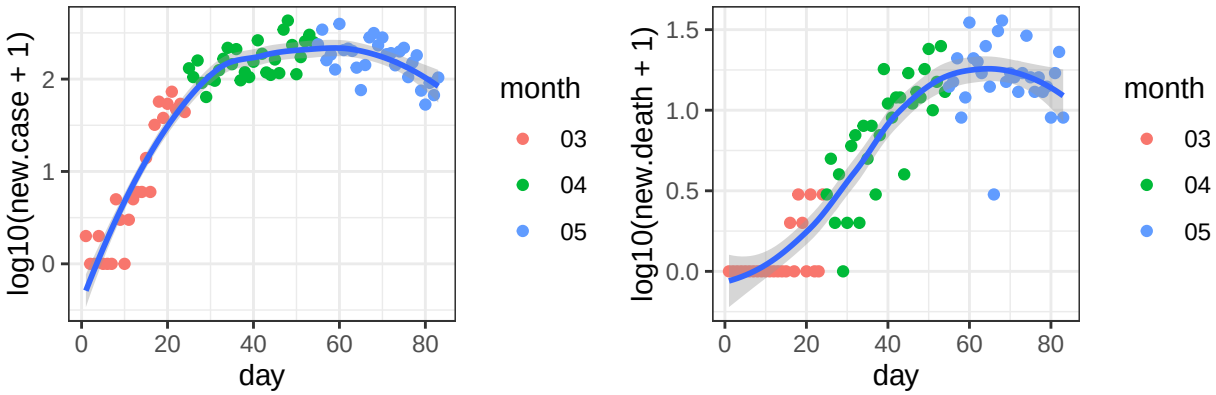
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Macomb_Michigan



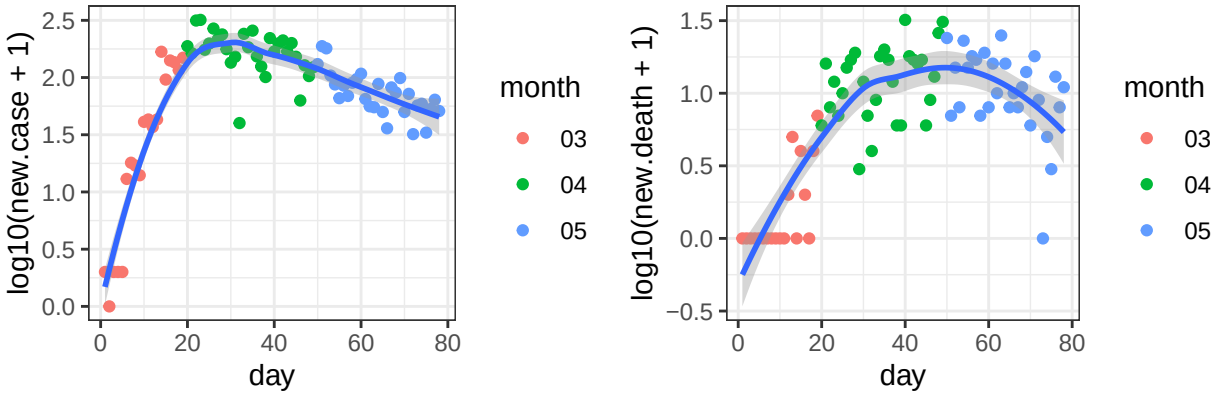
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Worcester_Massachusetts



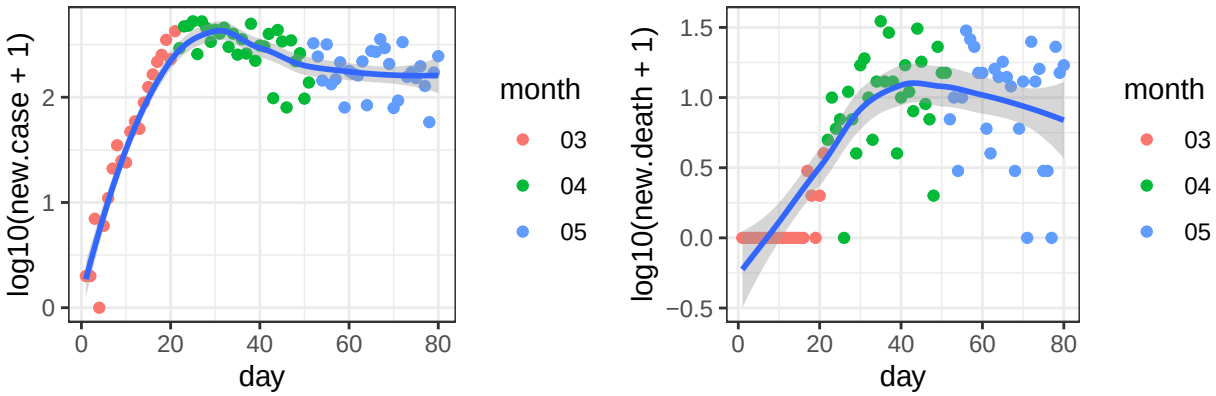
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Ocean_New Jersey



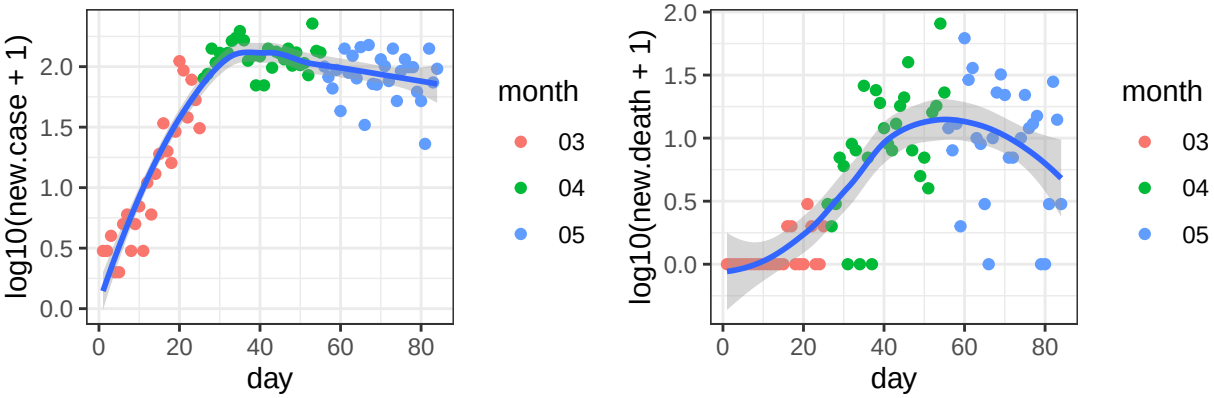
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Miami−Dade_Florida



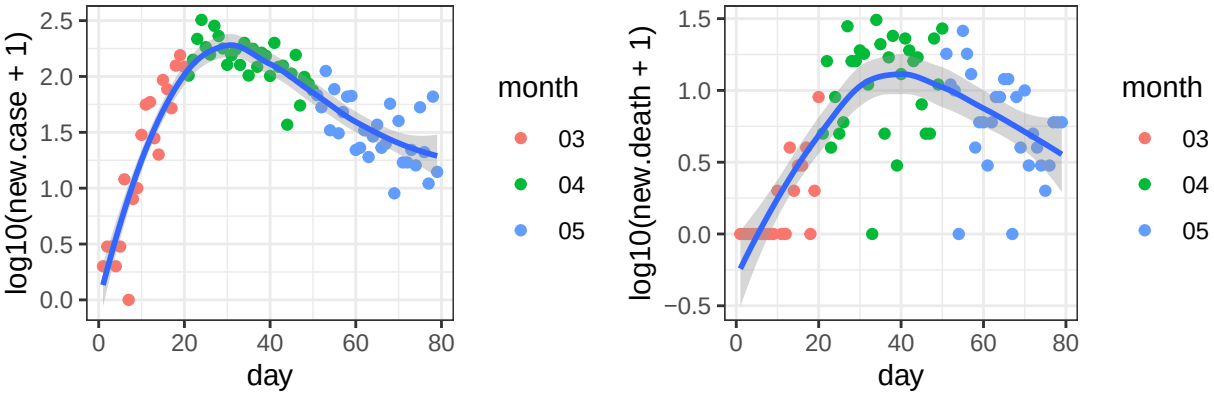
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Montgomery_Pennsylvania



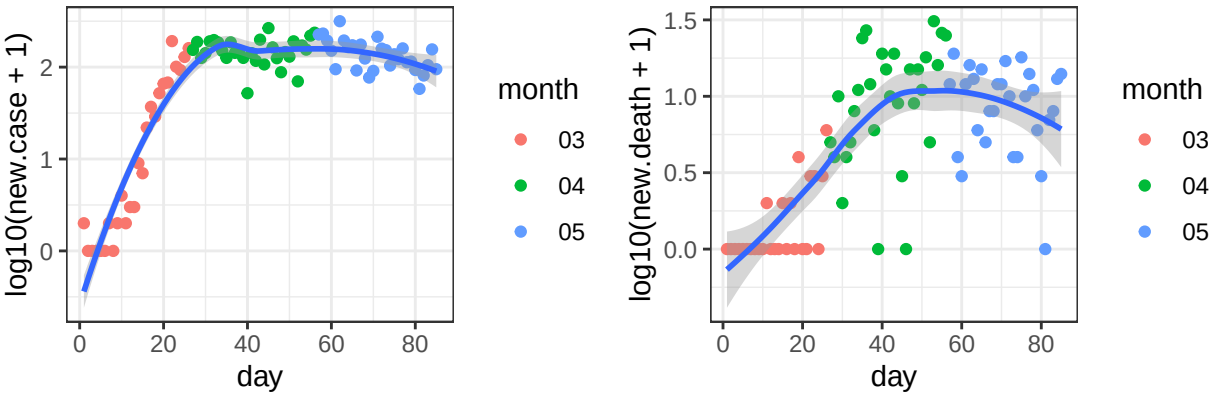
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Morris_New Jersey



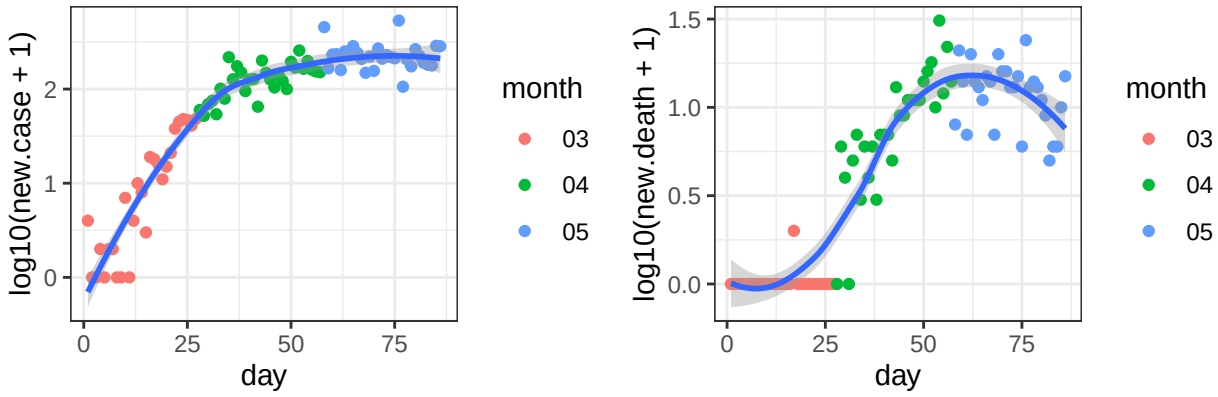
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Marion_Indiana



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

Montgomery_Maryland

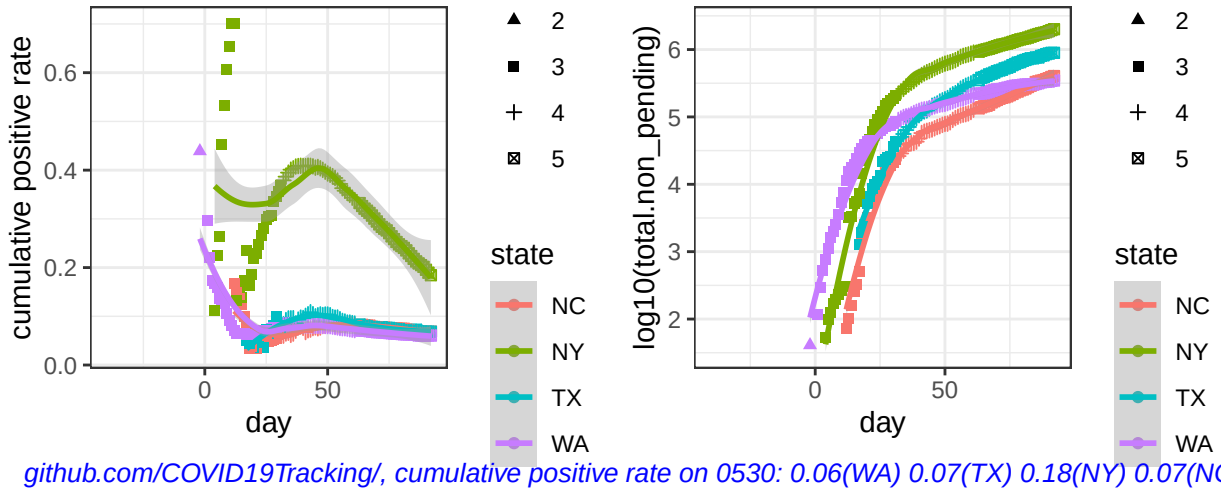


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

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 0530: 0.06(WA) 0.07(TX) 0.18(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_0.2.0  tibble_3.0.1
##  [9] gtable_0.3.0     pkgconfig_2.0.3  rlang_0.4.6      yaml_2.2.1
## [13] xfun_0.12        gridExtra_2.3    withr_2.1.2      stringr_1.4.0
## [17] dplyr_0.8.4      knitr_1.28       vctrs_0.3.0      cowplot_1.0.0
## [21] grid_3.6.2       tidyselect_1.0.0 glue_1.3.1       R6_2.4.1
## [25] rmarkdown_2.1    farver_2.0.3     purrr_0.3.3      scales_1.1.0
## [29] ellipsis_0.3.0   htmltools_0.4.0  assertthat_0.2.1 colorspace_1.4-1
## [33] ggsignif_0.6.0   labeling_0.3     stringi_1.4.5    lazyeval_0.2.2
## [37] munsell_0.5.0    crayon_1.3.4
```