Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-07-08

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

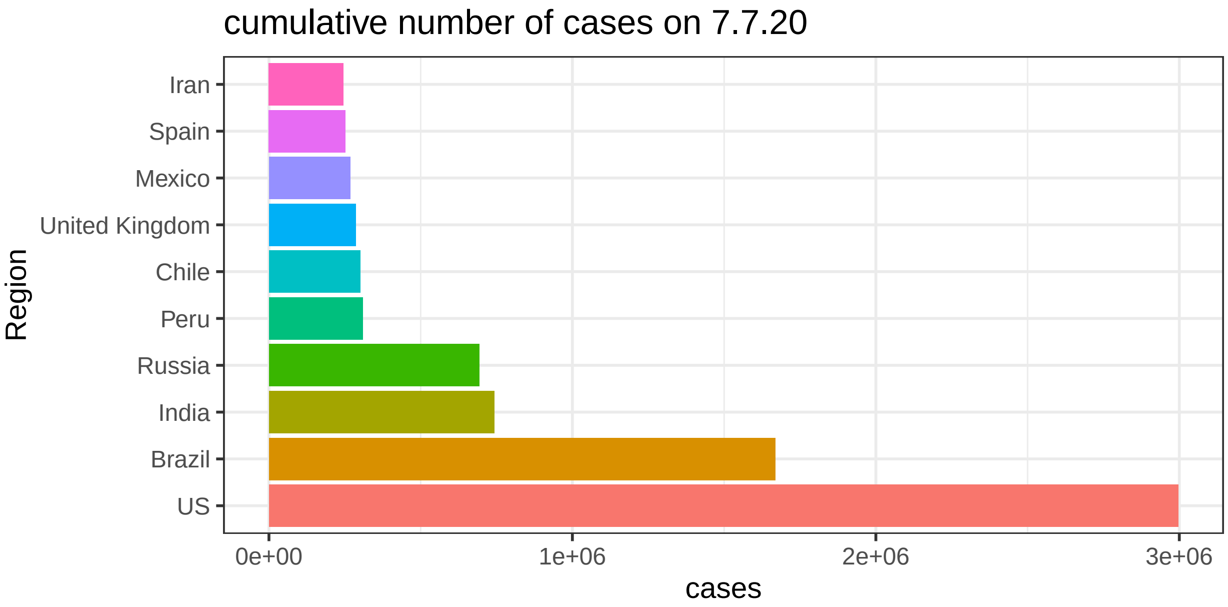
JHU

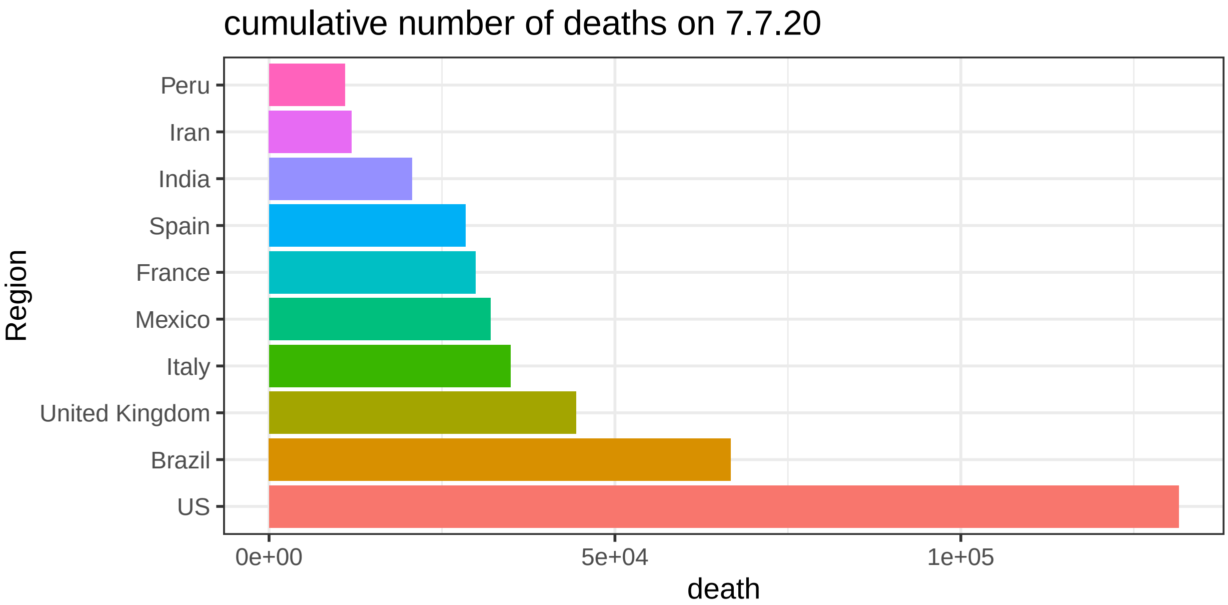
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

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.

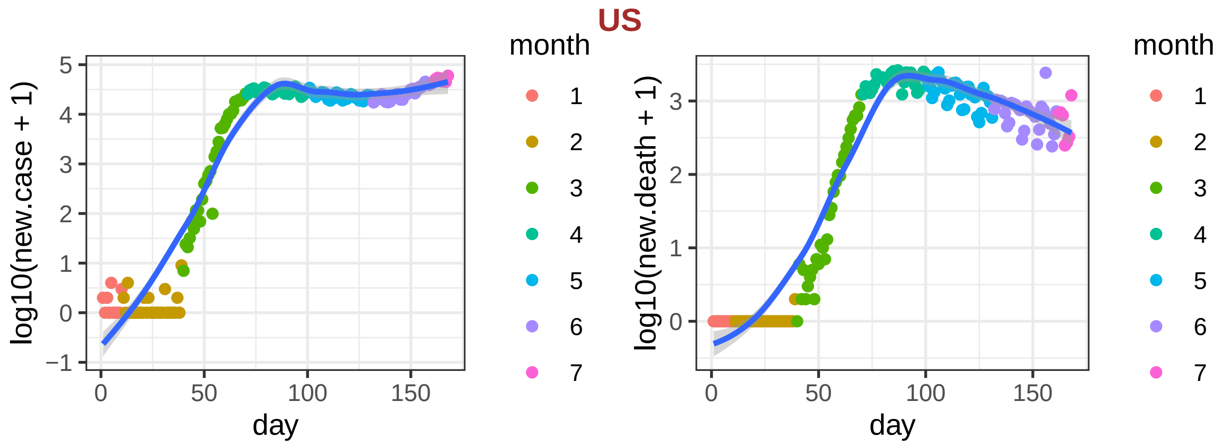
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





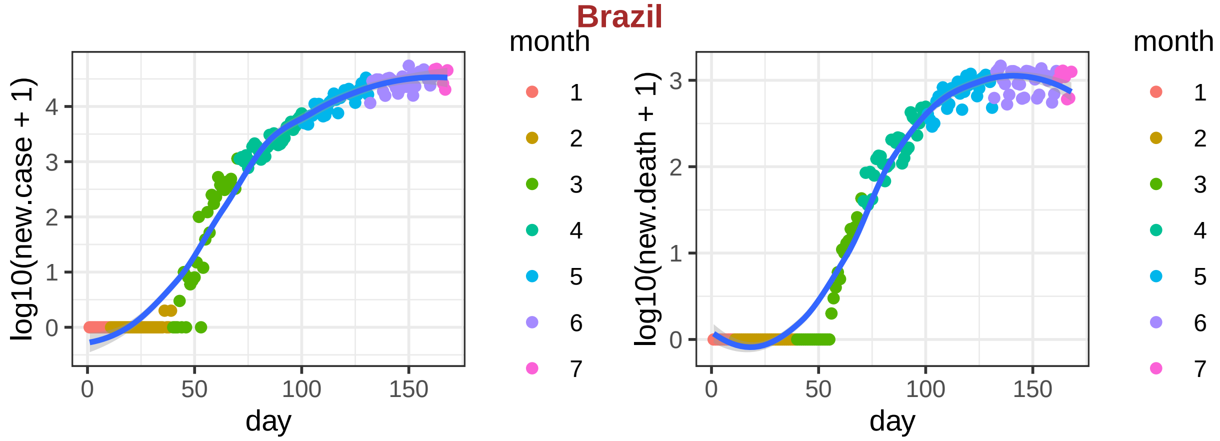
cumulative number of deaths on 7.7.20

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

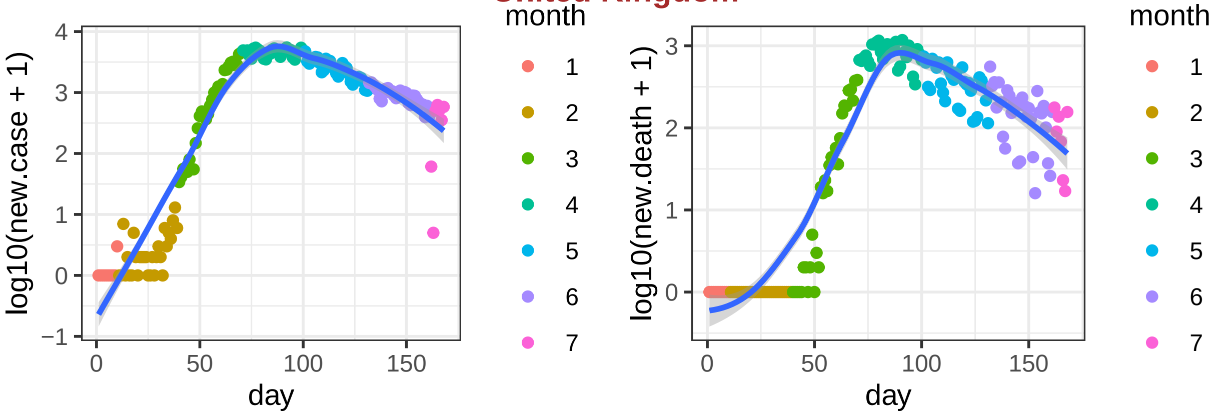
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



Brazil

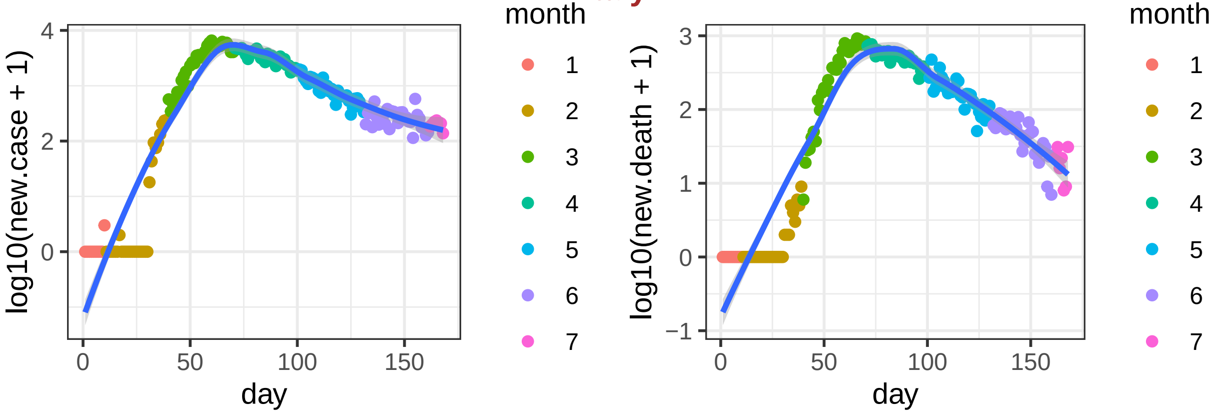
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



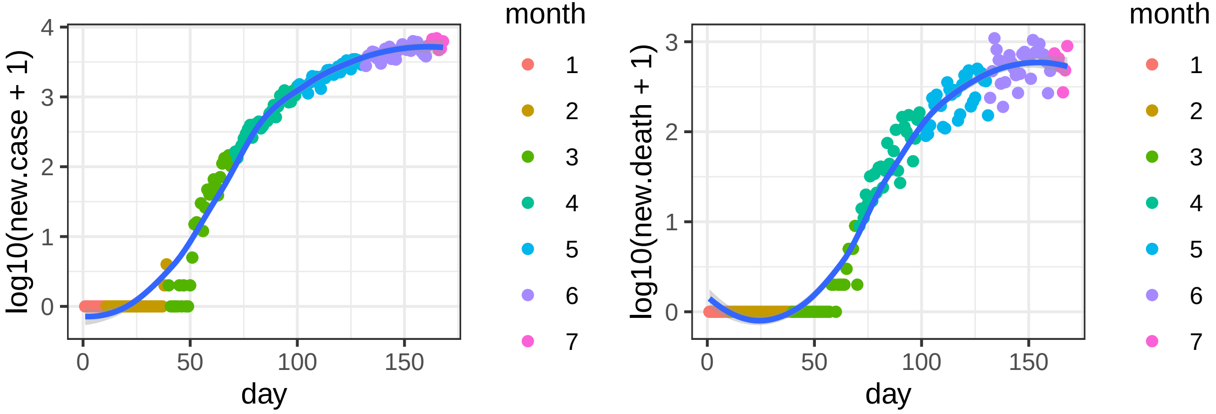
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



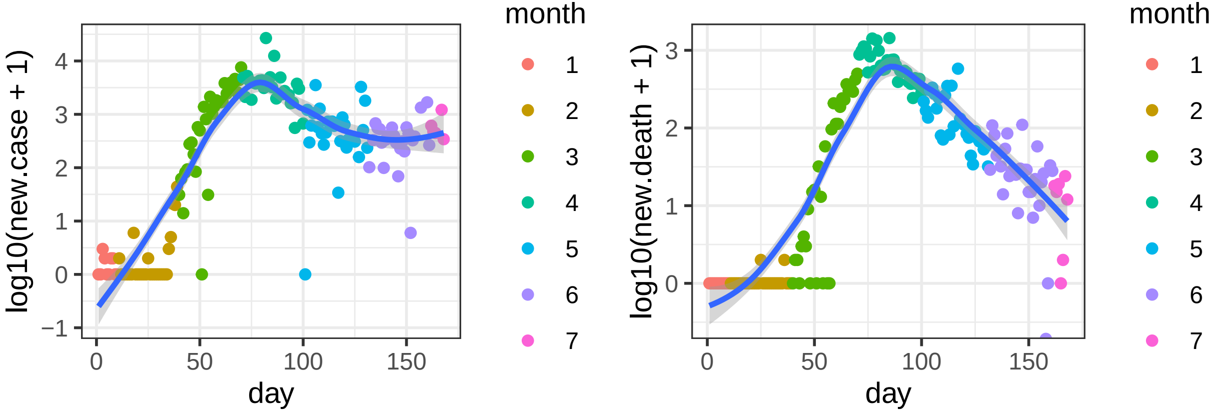
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Mexico



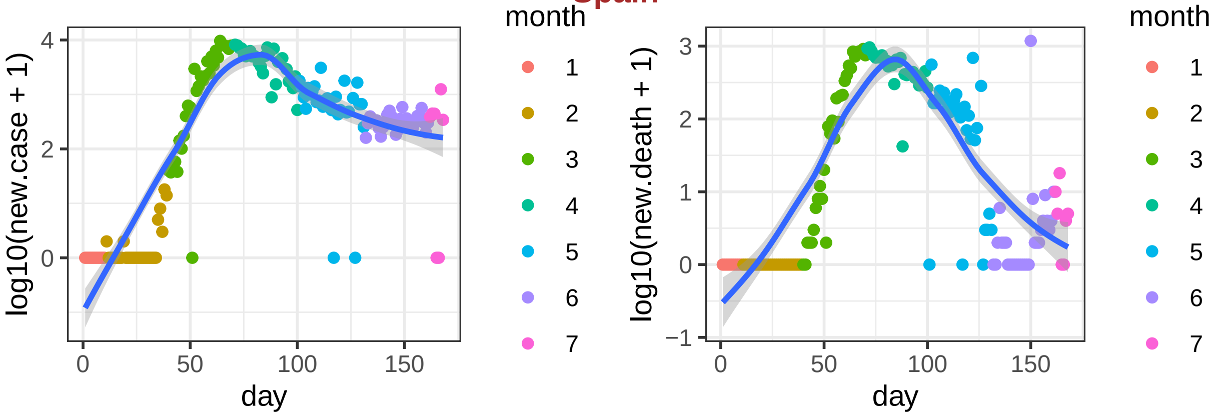
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



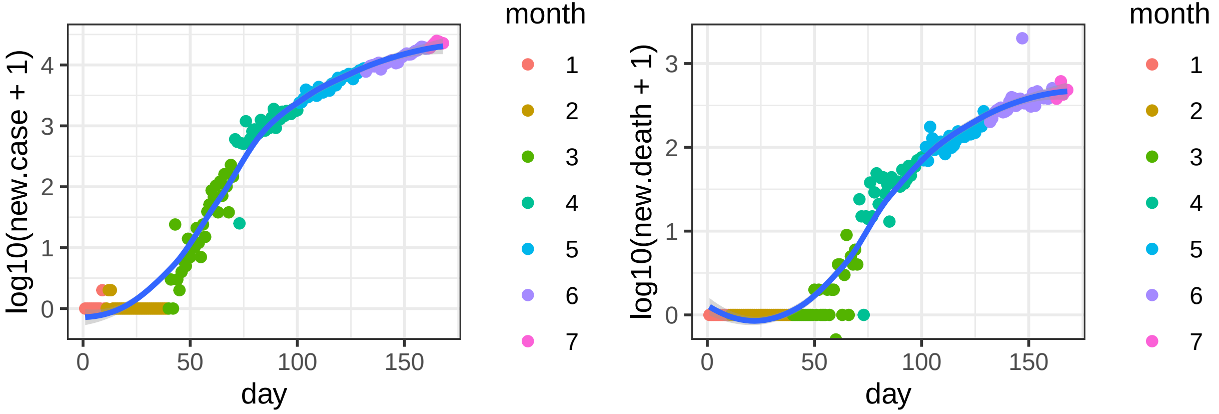
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain

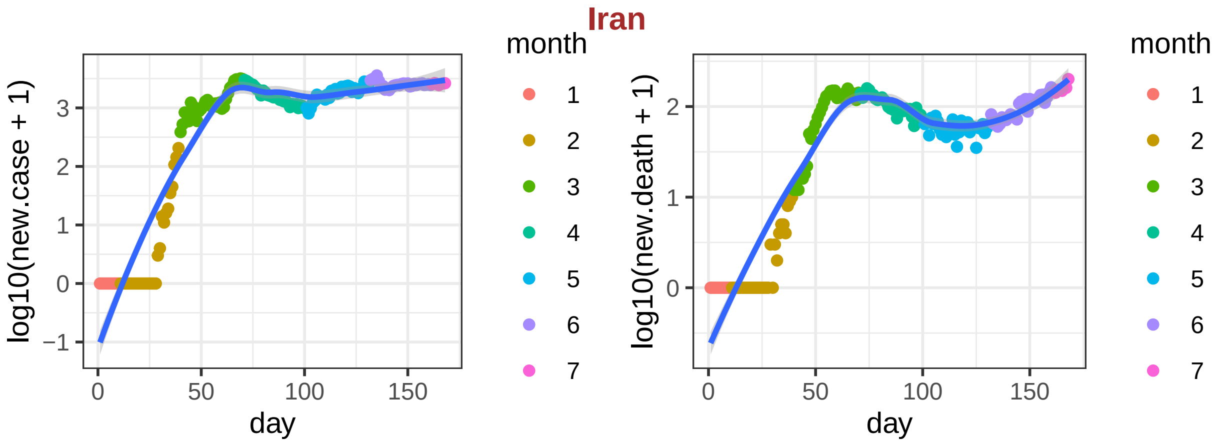


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

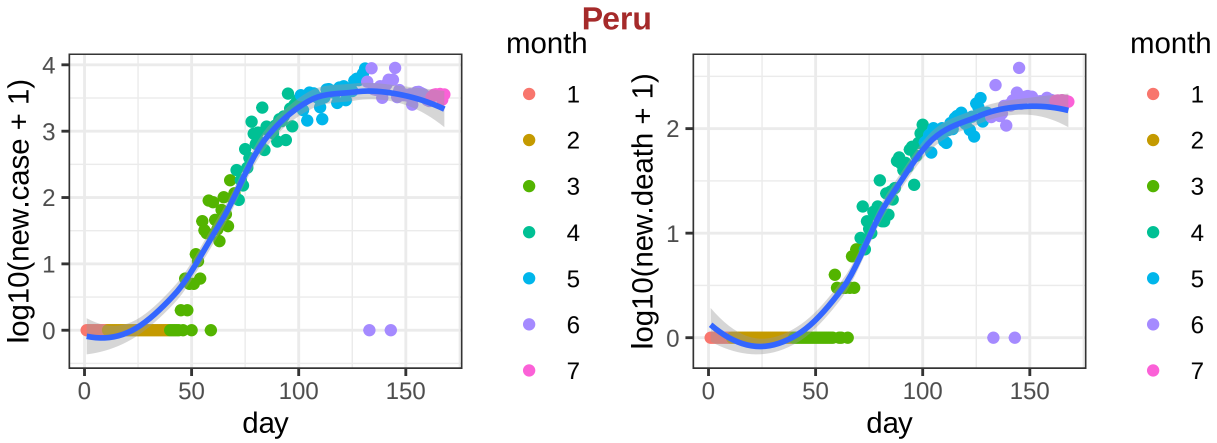
India



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



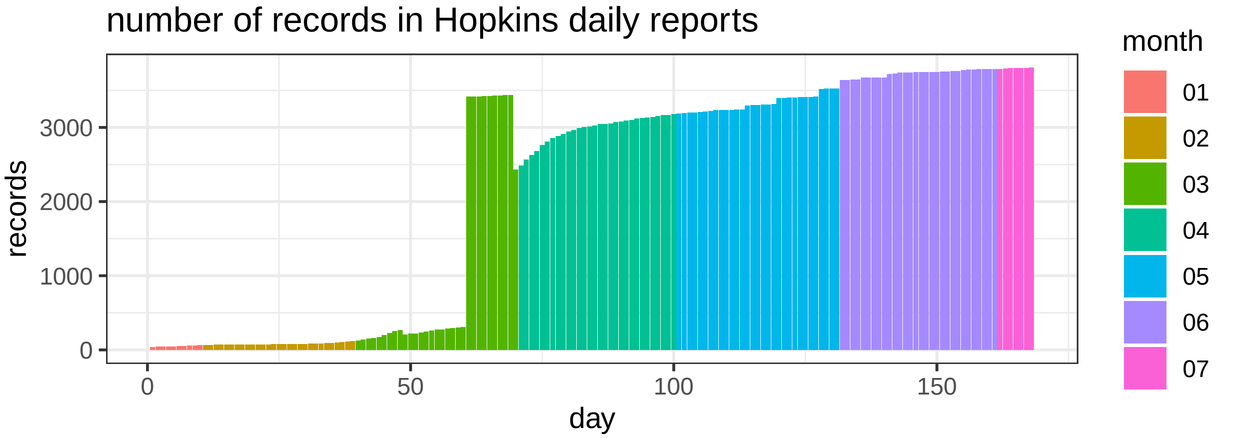
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



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

daily reports data

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



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

NY Times

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

The currente date is

```
## [1] "2020-07-06"
```

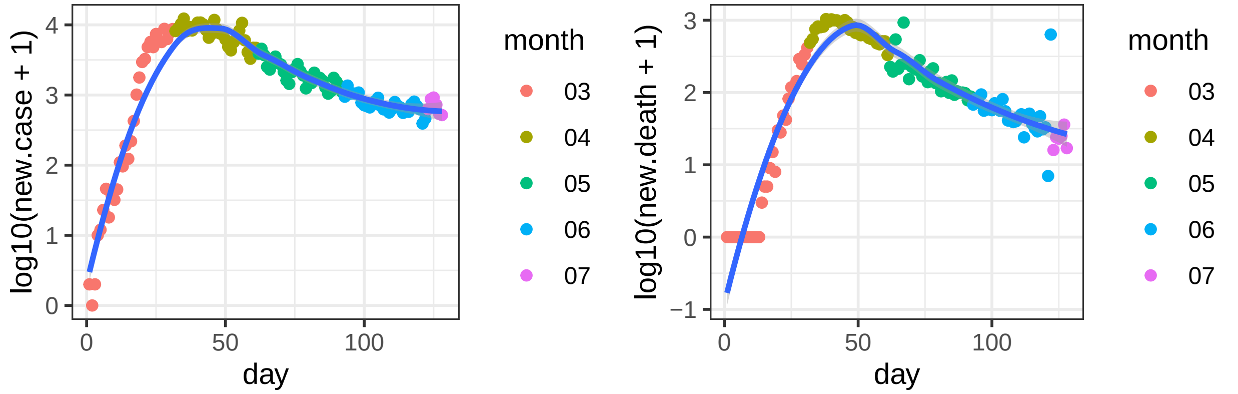
state level data

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

```
##            date          state fips  cases deaths
## 6923 2020-07-06       New York   36 402338  31911
## 6921 2020-07-06     New Jersey   34 175467  15229
## 6912 2020-07-06  Massachusetts   25 110137   8198
## 6904 2020-07-06       Illinois   17 149428   7250
## 6930 2020-07-06   Pennsylvania   42  95187   6798
## 6894 2020-07-06     California    6 277869   6452
## 6913 2020-07-06       Michigan   26  73403   6225
## 6896 2020-07-06    Connecticut    9  46976   4338
## 6899 2020-07-06        Florida   12 206439   3777
## 6909 2020-07-06      Louisiana   22  66435   3296
## 6911 2020-07-06       Maryland   24  70497   3246
## 6927 2020-07-06           Ohio   39  57956   2927
## 6900 2020-07-06        Georgia   13  91015   2829
## 6936 2020-07-06          Texas   48 209319   2726
## 6905 2020-07-06        Indiana   18  49560   2698
## 6940 2020-07-06       Virginia   51  66102   1853
## 6892 2020-07-06        Arizona    4 101542   1832
## 6895 2020-07-06       Colorado    8  34316   1704
## 6914 2020-07-06      Minnesota   27  38606   1511
## 6924 2020-07-06 North Carolina   37  74930   1424
## 6941 2020-07-06     Washington   53  38517   1370
## 6915 2020-07-06    Mississippi   28  31257   1114
## 6916 2020-07-06       Missouri   29  24850   1068
## 6890 2020-07-06        Alabama    1  44878   1007
## 6932 2020-07-06   Rhode Island   44  16991    960
## 6933 2020-07-06 South Carolina   45  46380    827
## 6943 2020-07-06      Wisconsin   55  35318    805
## 6906 2020-07-06           Iowa   19  31764    725
## 6935 2020-07-06      Tennessee   47  51509    646
## 6908 2020-07-06       Kentucky   21  17464    621
```

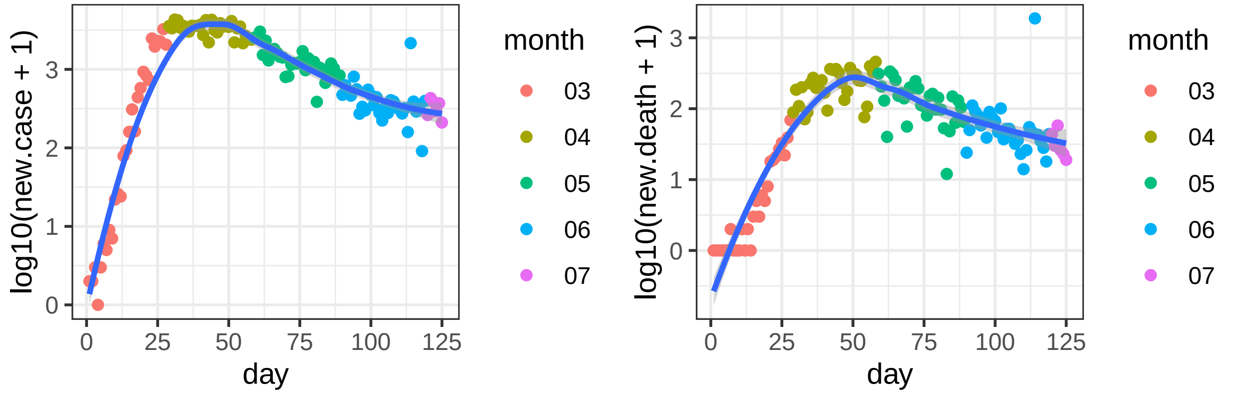
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



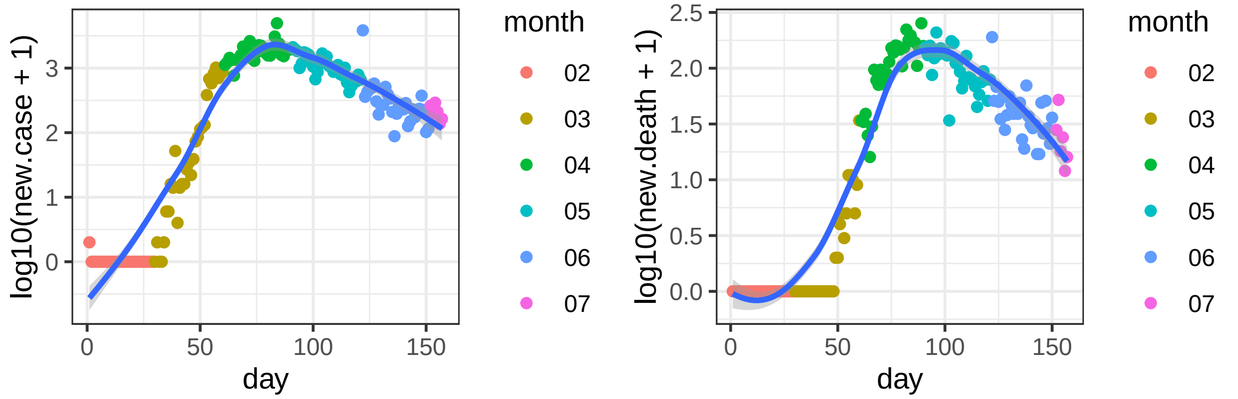
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



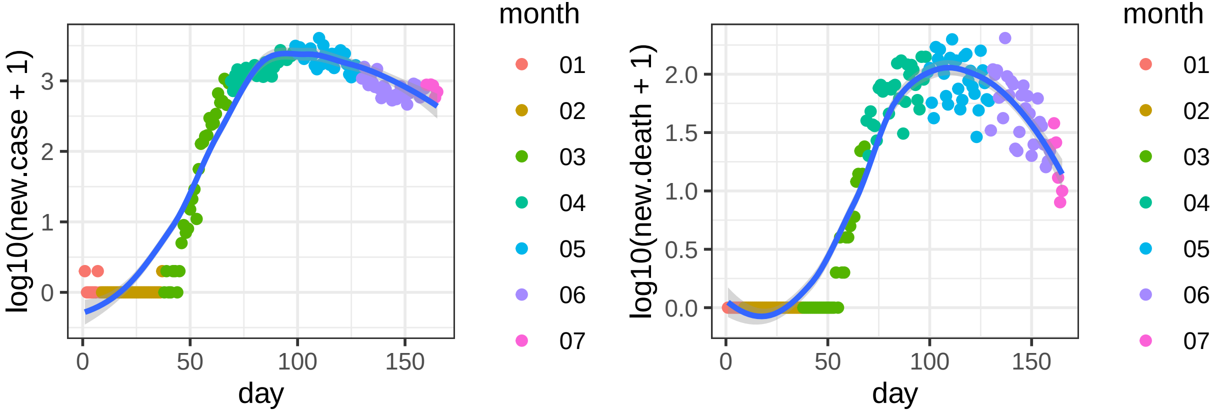
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Massachusetts



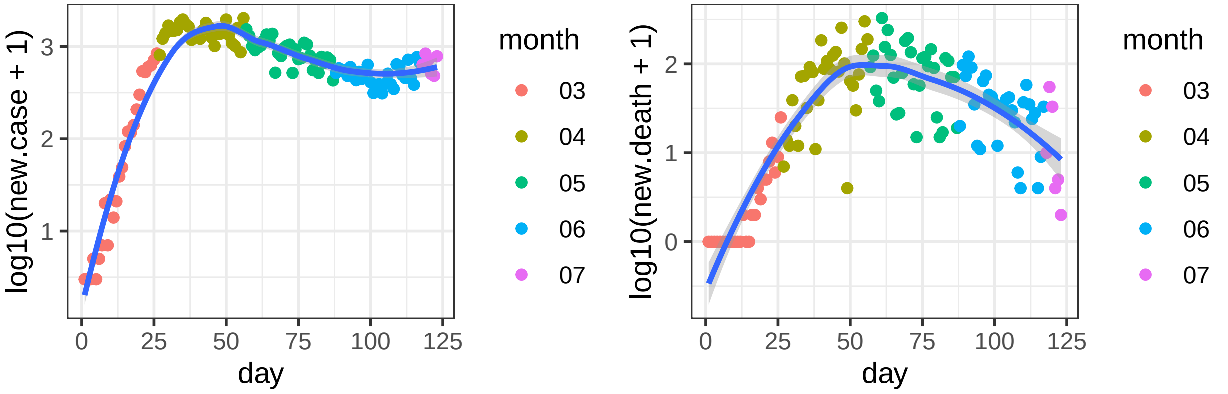
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Illinois



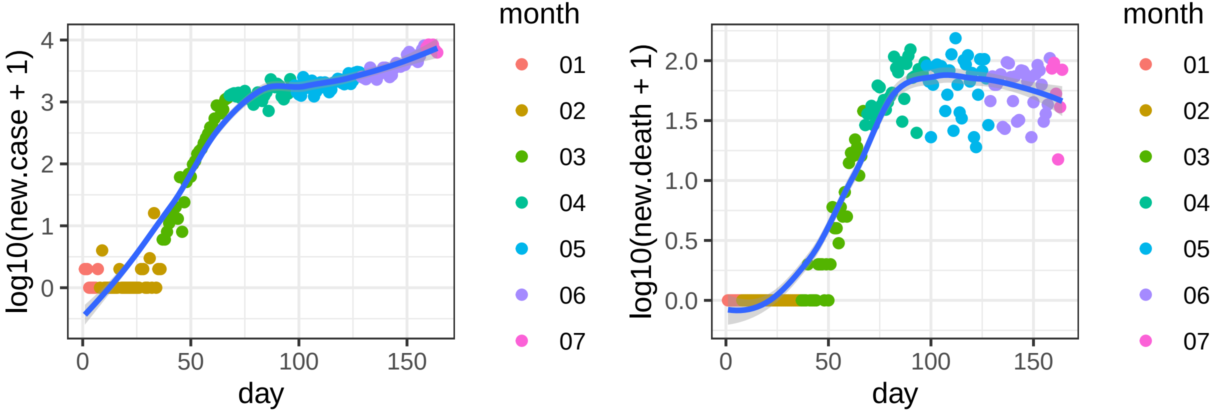
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Pennsylvania



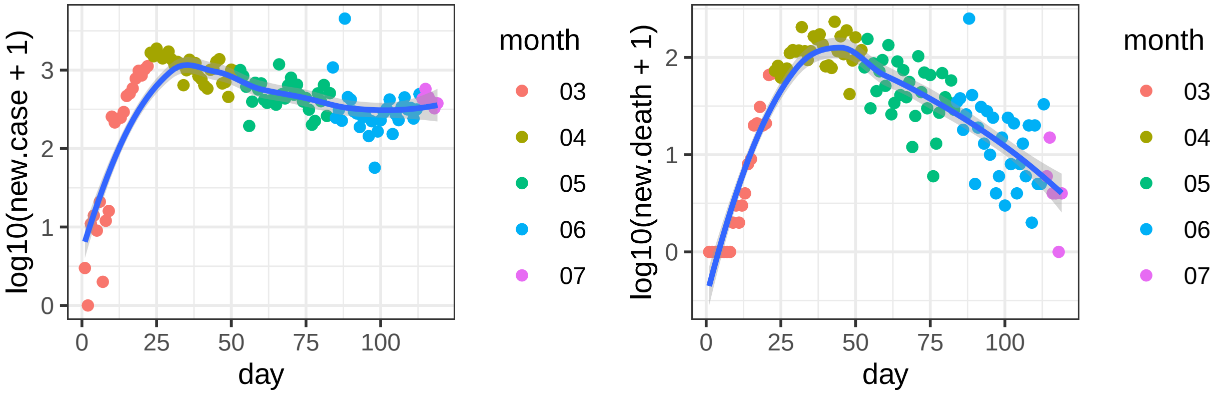
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

California



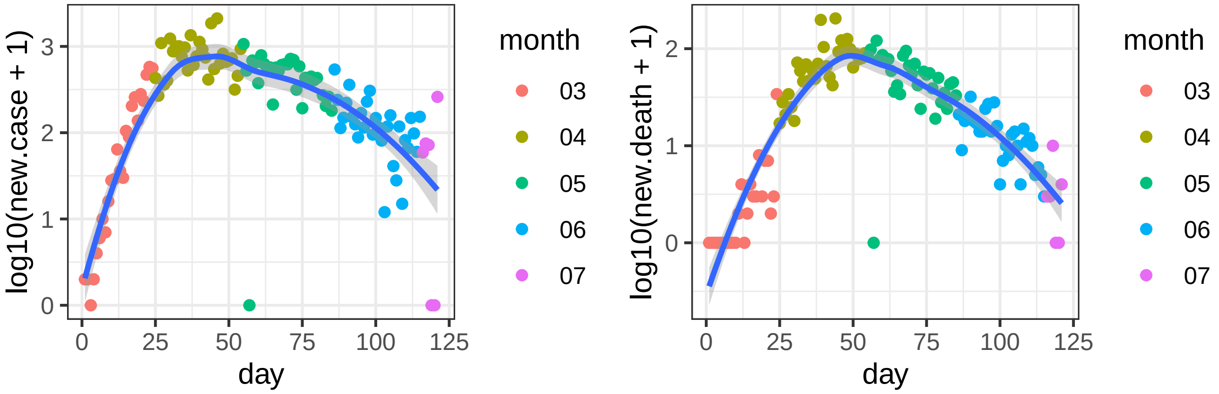
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Michigan



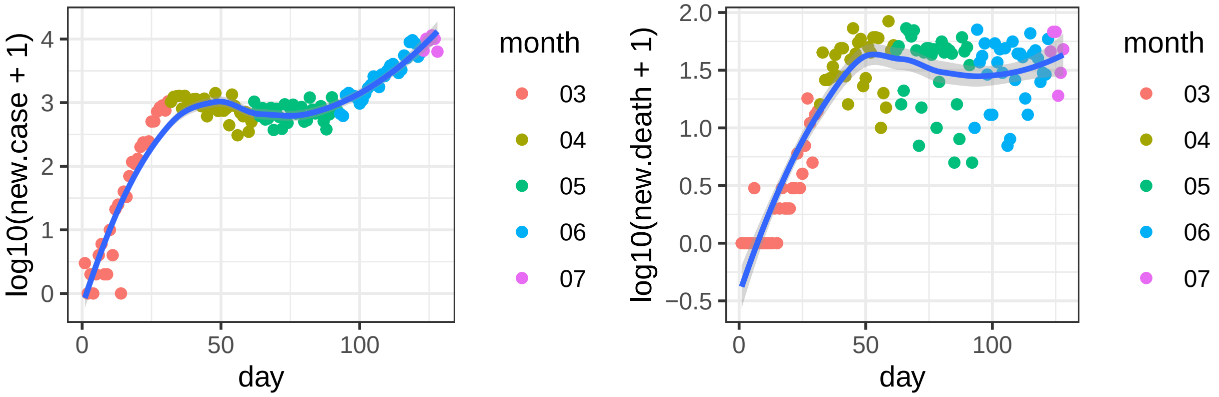
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Connecticut



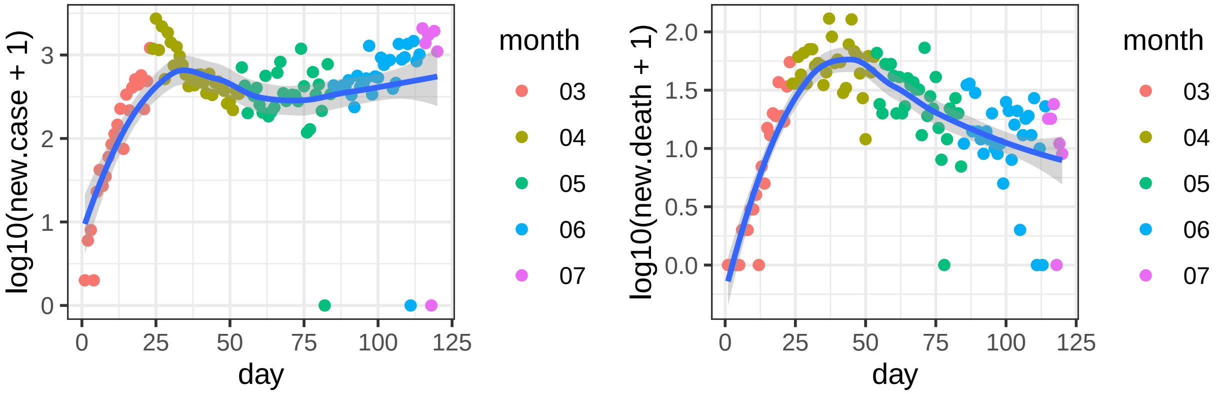
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Florida



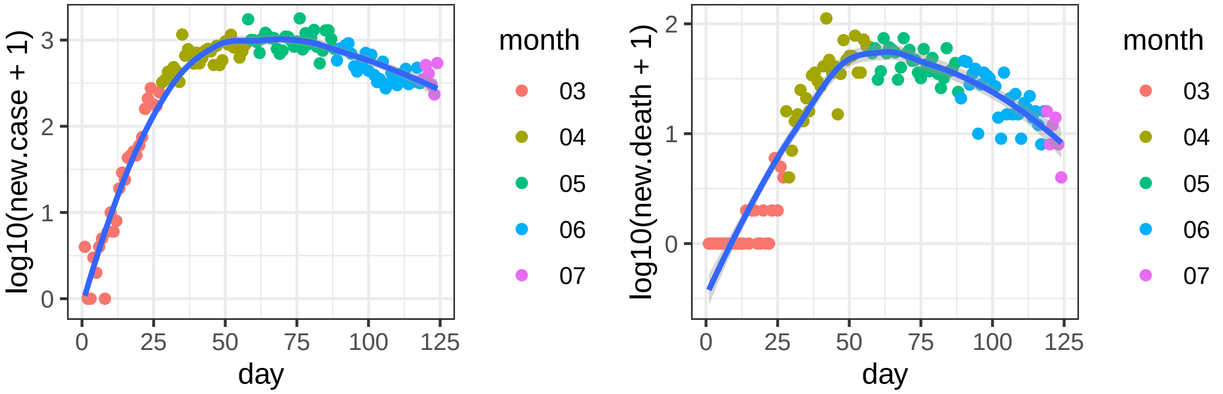
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Louisiana



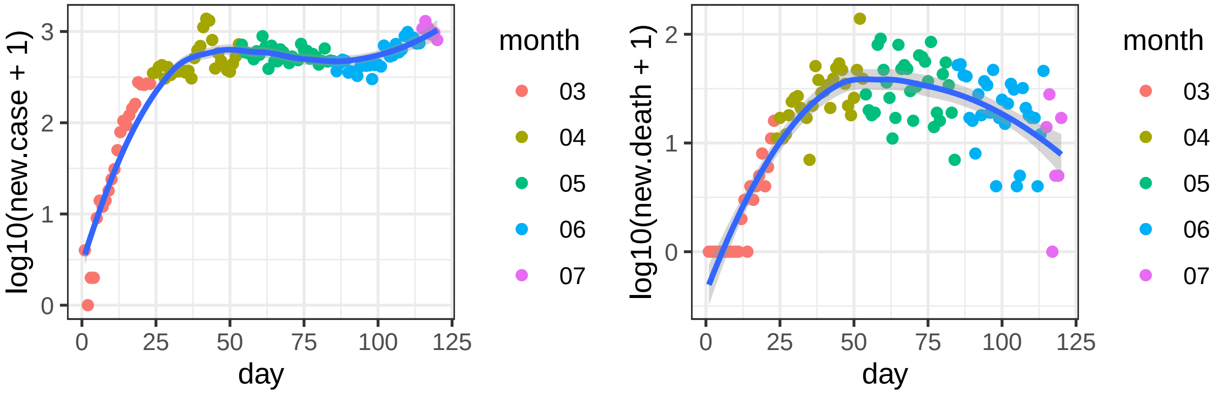
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



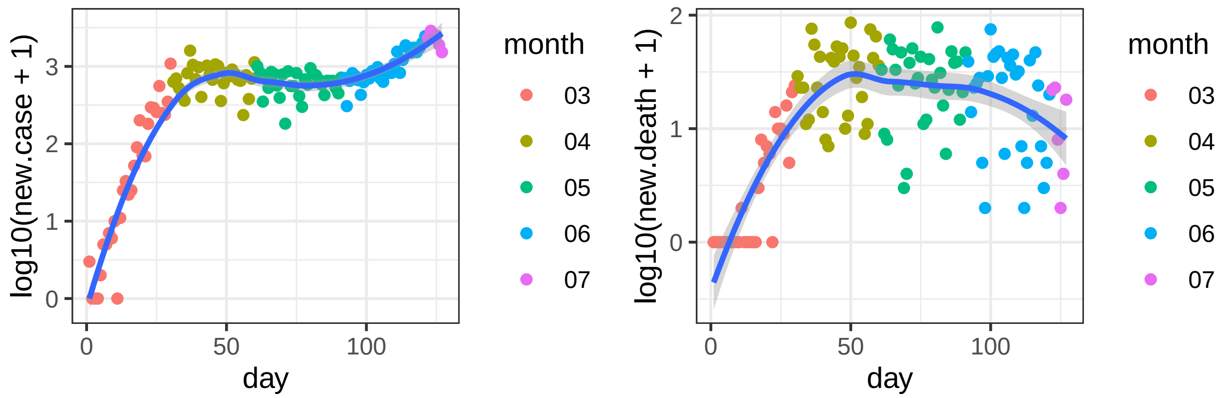
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Ohio



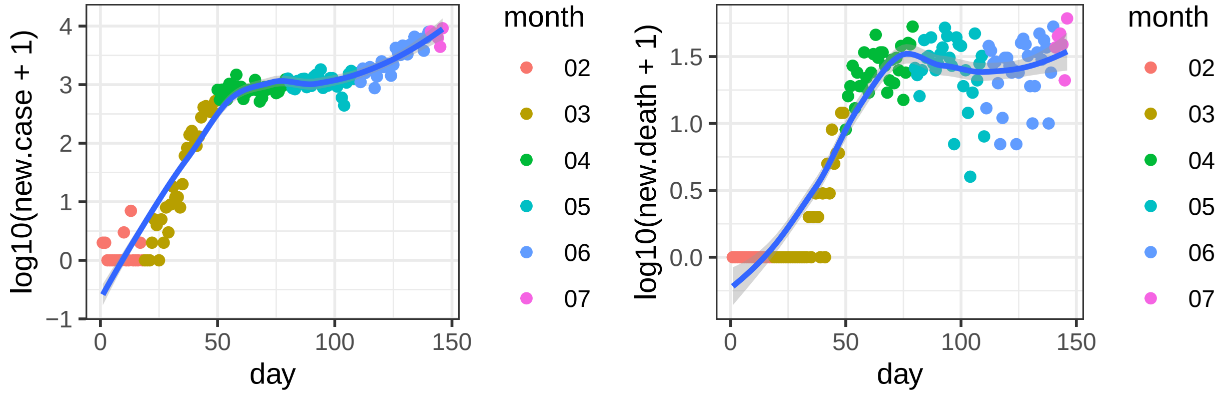
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Georgia



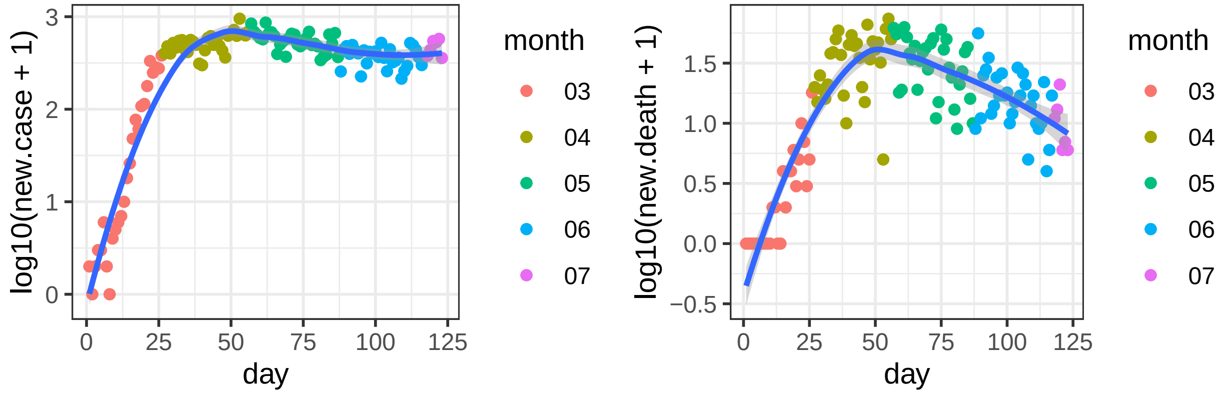
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Texas



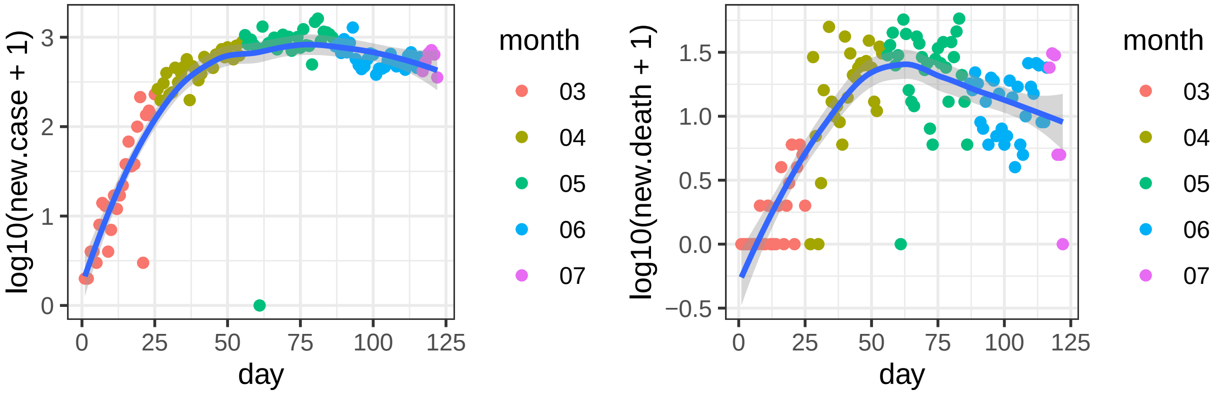
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Indiana



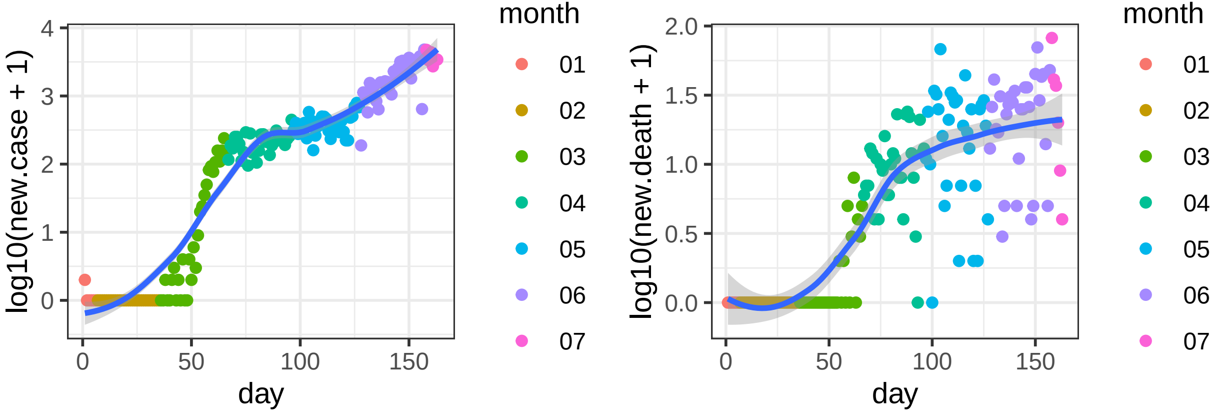
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Virginia



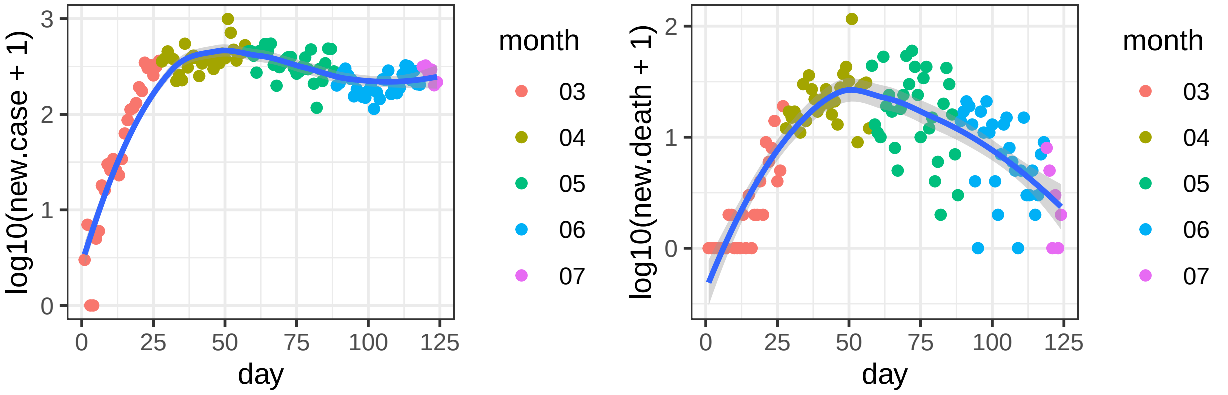
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Arizona



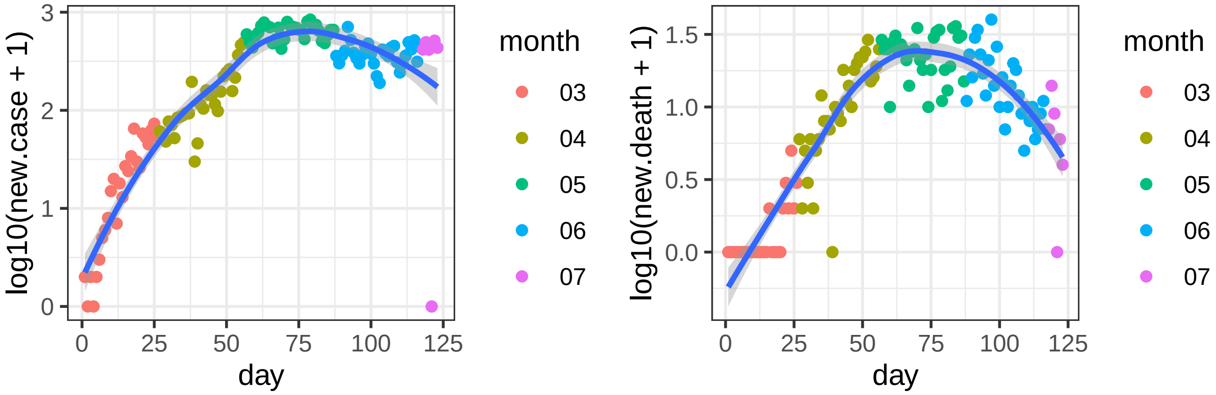
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Colorado



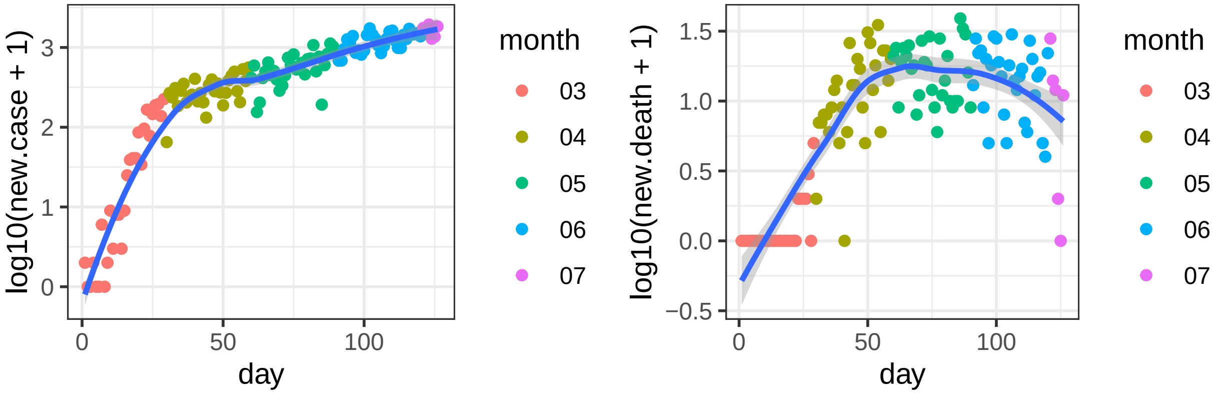
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Minnesota



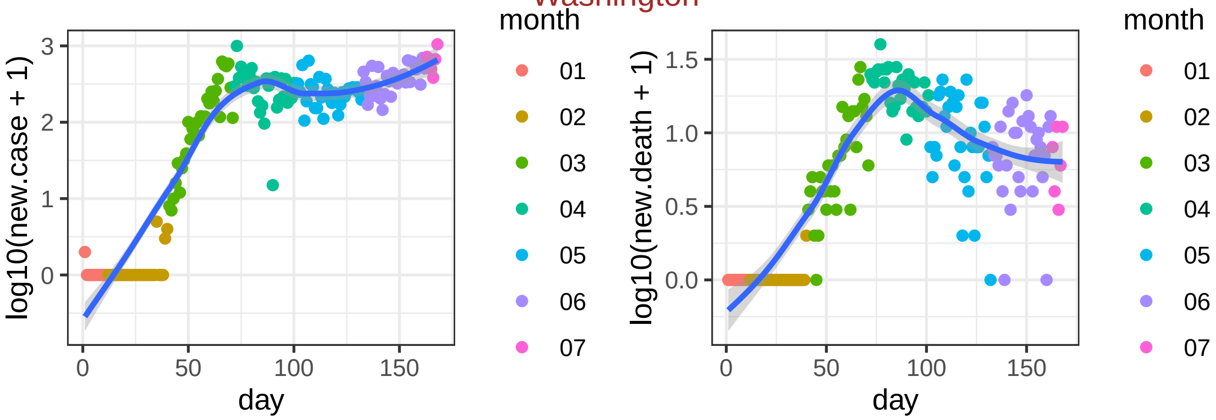
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

North Carolina



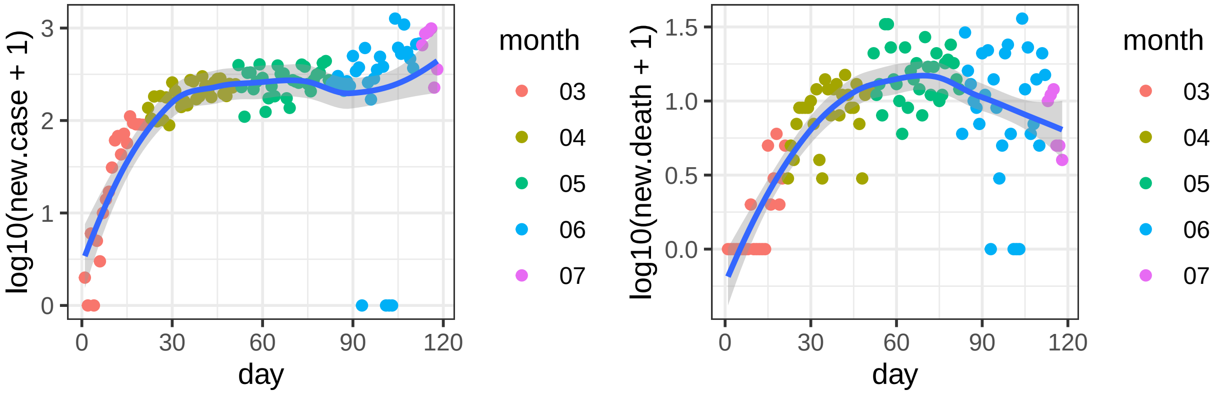
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Washington



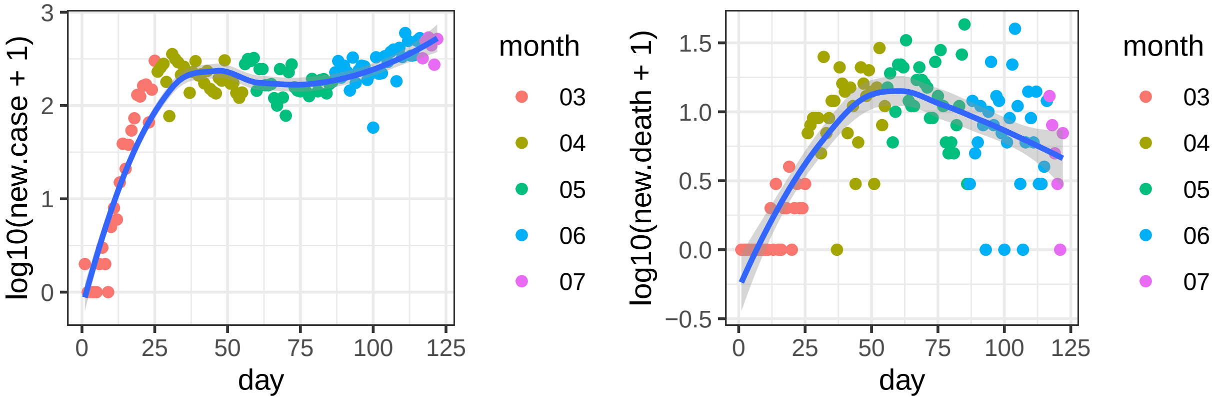
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Mississippi



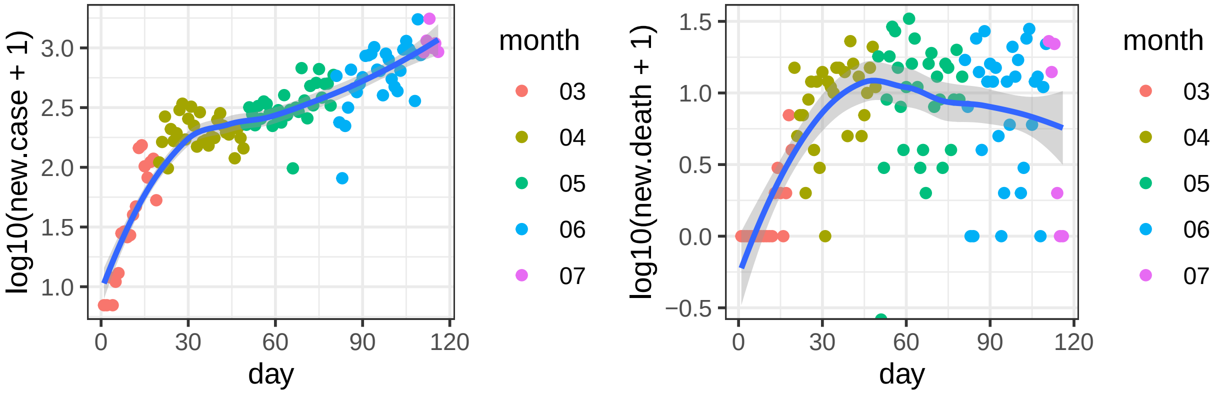
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Missouri



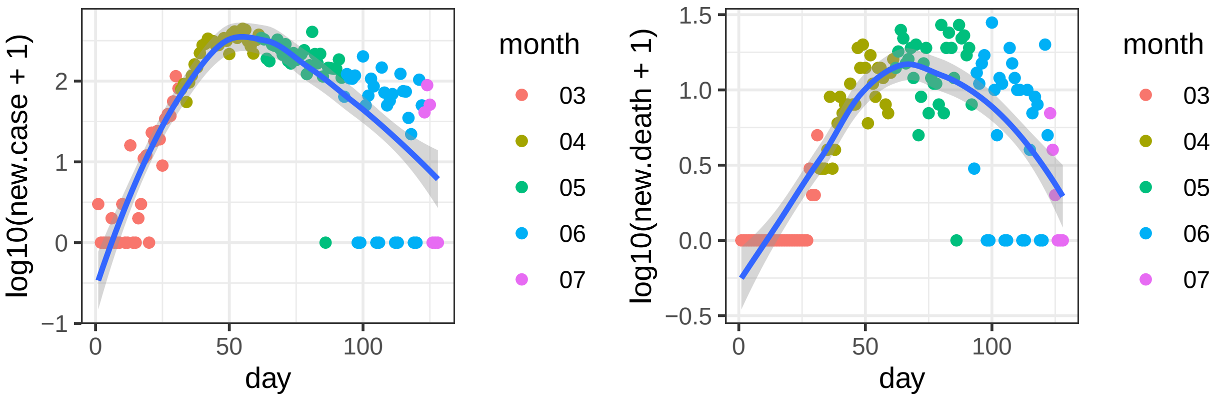
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Alabama



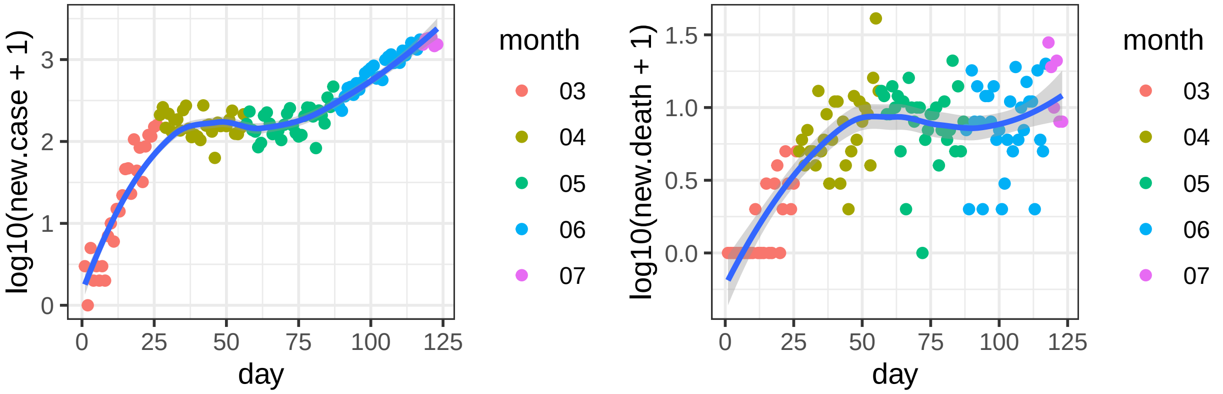
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Rhode Island



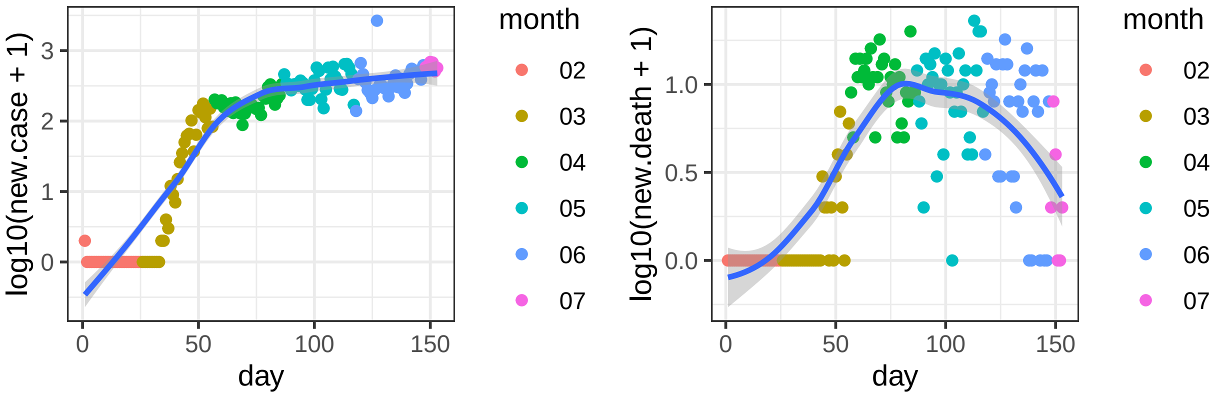
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

South Carolina



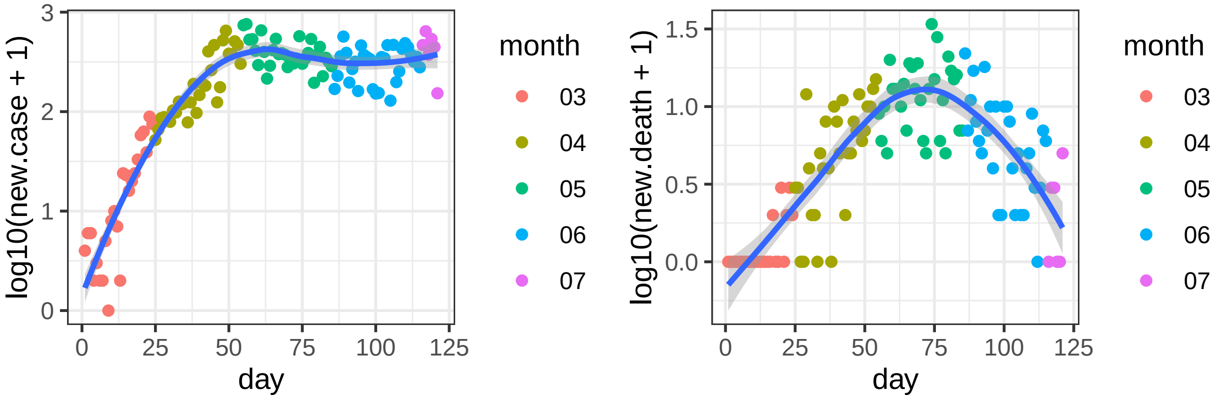
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Wisconsin



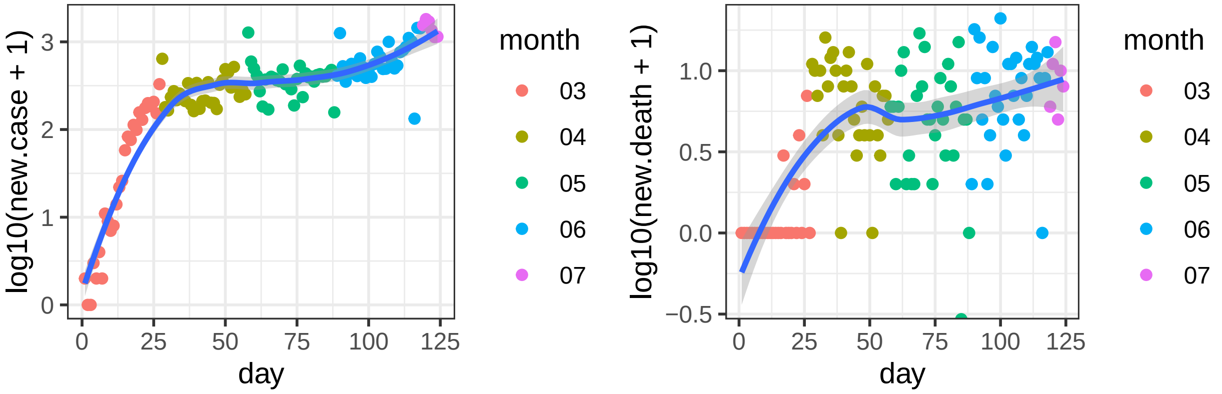
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

Iowa



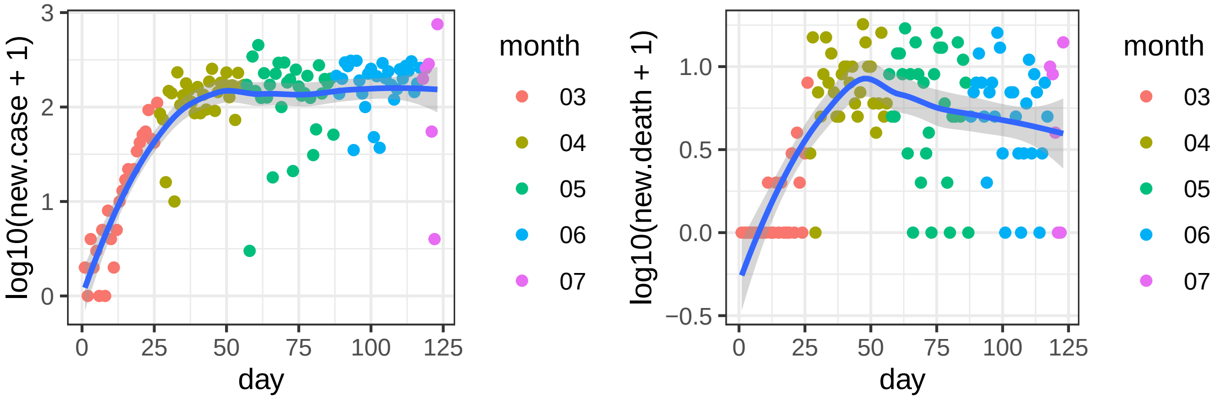
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Tennessee



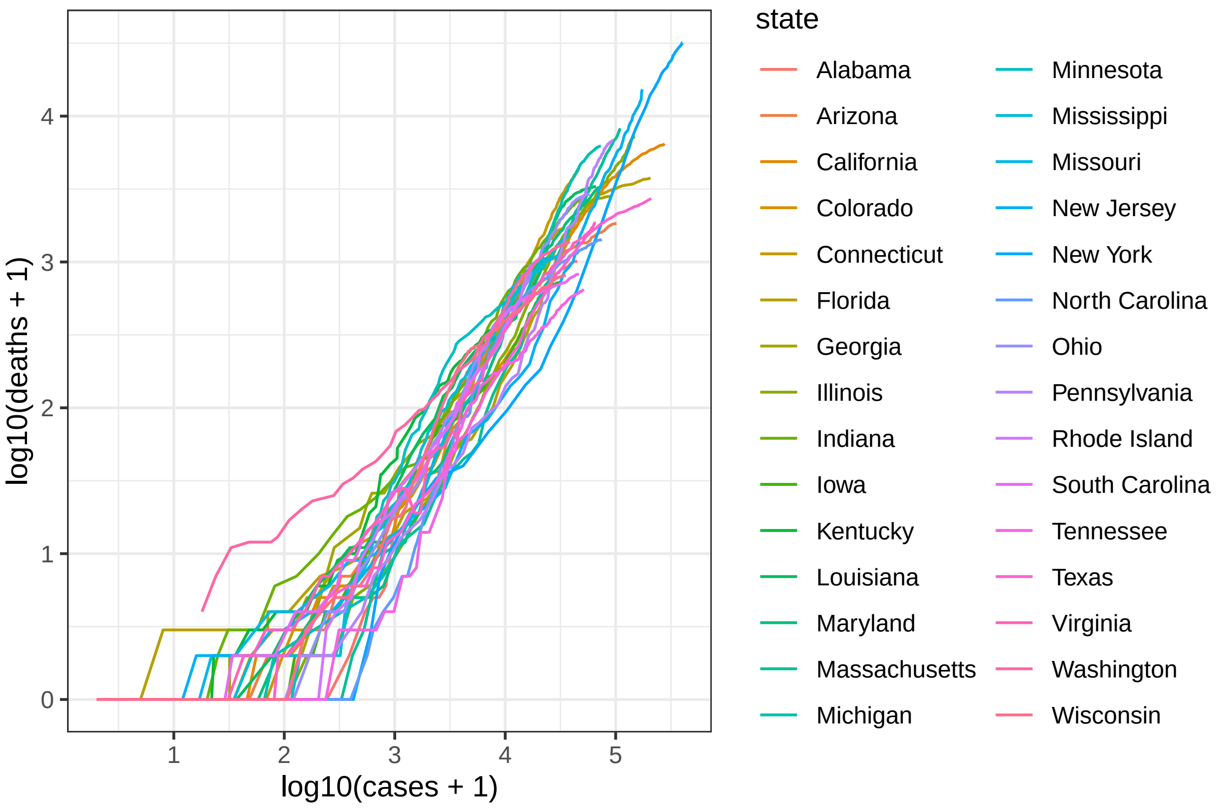
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Kentucky



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

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



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

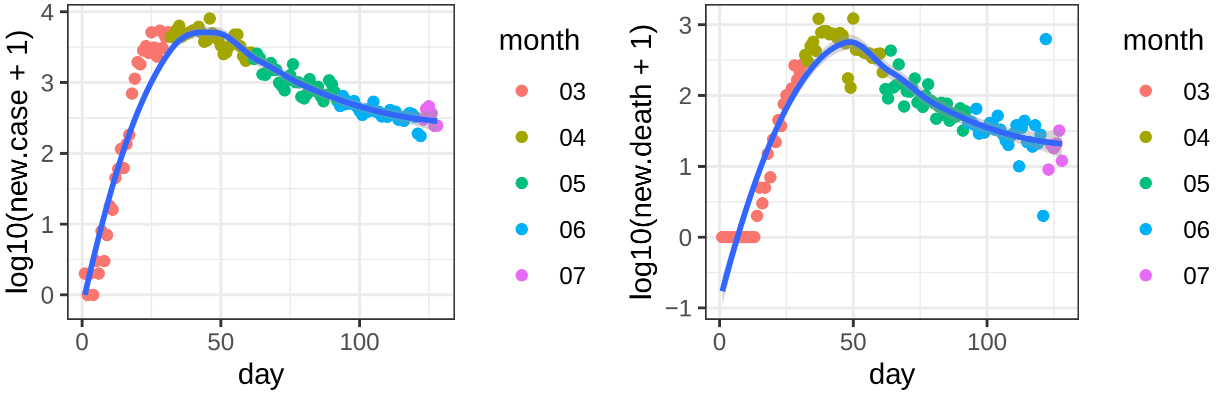
county level data

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

```
##                   date               county                state  fips  cases deaths
## 302828 2020-07-06        New York City             New York    NA 221882  22672
## 301612 2020-07-06                 Cook             Illinois 17031  92781   4630
## 301208 2020-07-06          Los Angeles           California  6037 116570   3534
## 302314 2020-07-06                Wayne             Michigan 26163  23283   2732
## 302827 2020-07-06               Nassau             New York 36059  42053   2698
## 302753 2020-07-06                Essex           New Jersey 34013  19136   2041
## 302847 2020-07-06              Suffolk             New York 36103  41685   2030
## 302748 2020-07-06               Bergen           New Jersey 34003  19916   2006
## 302225 2020-07-06            Middlesex        Massachusetts 25017  24193   1882
## 303258 2020-07-06         Philadelphia         Pennsylvania 42101  26810   1619
## 302855 2020-07-06          Westchester             New York 36119  35083   1560
## 302755 2020-07-06               Hudson           New Jersey 34017  19157   1456
## 301309 2020-07-06             Hartford          Connecticut  9003  11794   1380
## 301308 2020-07-06            Fairfield          Connecticut  9001  16823   1377
## 302758 2020-07-06            Middlesex           New Jersey 34023  17153   1333
## 302766 2020-07-06                Union           New Jersey 34039  16607   1328
## 302762 2020-07-06              Passaic           New Jersey 34031  17110   1196
## 302221 2020-07-06                Essex        Massachusetts 25009  16283   1122
## 302294 2020-07-06              Oakland             Michigan 26125  12254   1091
## 301312 2020-07-06            New Haven          Connecticut  9009  12462   1078
## 301364 2020-07-06           Miami-Dade              Florida 12086  48991   1051
## 302229 2020-07-06              Suffolk        Massachusetts 25025  20014   1008
## 302761 2020-07-06                Ocean           New Jersey 34029   9784    974
## 302231 2020-07-06            Worcester        Massachusetts 25027  12515    942
## 302227 2020-07-06              Norfolk        Massachusetts 25021   9284    940
## 302281 2020-07-06               Macomb             Michigan 26099   7820    924
## 301106 2020-07-06             Maricopa              Arizona  4013  64915    881
## 302759 2020-07-06             Monmouth           New Jersey 34025   9416    820
## 303253 2020-07-06           Montgomery         Pennsylvania 42091   8634    811
## 302760 2020-07-06               Morris           New Jersey 34027   6967    806
## 302342 2020-07-06             Hennepin            Minnesota 27053  12456    787
## 302207 2020-07-06           Montgomery             Maryland 24031  15201    753
## 303279 2020-07-06           Providence         Rhode Island 44007  13144    747
## 301748 2020-07-06               Marion              Indiana 18097  11814    731
## 303230 2020-07-06             Delaware         Pennsylvania 42045   7393    702
## 302208 2020-07-06      Prince George's             Maryland 24033  19531    690
## 302223 2020-07-06              Hampden        Massachusetts 25013   6883    670
## 302228 2020-07-06             Plymouth        Massachusetts 25023   8748    668
## 303923 2020-07-06                 King           Washington 53033  11142    624
## 302813 2020-07-06                 Erie             New York 36029   7500    599
## 302219 2020-07-06              Bristol        Massachusetts 25005   8331    592
## 302757 2020-07-06               Mercer           New Jersey 34021   7788    591
## 302586 2020-07-06            St. Louis             Missouri 29189   6931    584
## 303216 2020-07-06                Bucks         Pennsylvania 42017   5912    568
## 301321 2020-07-06 District of Columbia District of Columbia 11001  10515    561
## 301371 2020-07-06           Palm Beach              Florida 12099  17240    543
## 302145 2020-07-06              Orleans            Louisiana 22071   8143    534
## 302764 2020-07-06             Somerset           New Jersey 34035   5023    531
## 302750 2020-07-06               Camden           New Jersey 34007   7547    525
## 303812 2020-07-06              Fairfax             Virginia 51059  14205    495
```

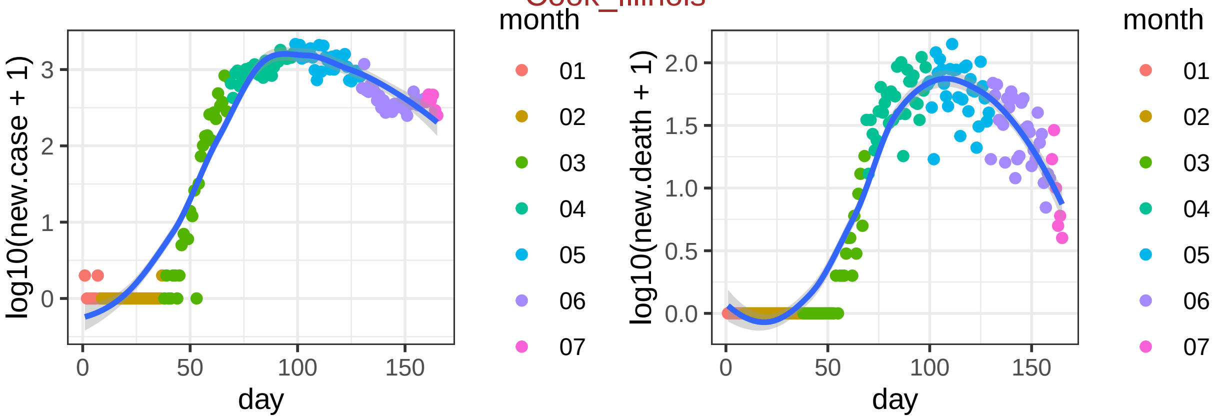
For these 50 counties, I check the number of new cases and the number of new deaths.

New York City_New York



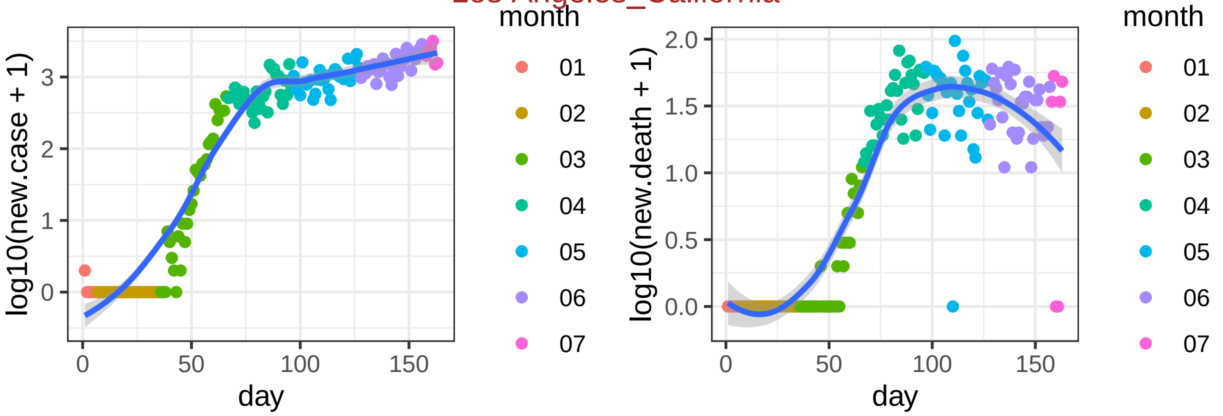
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



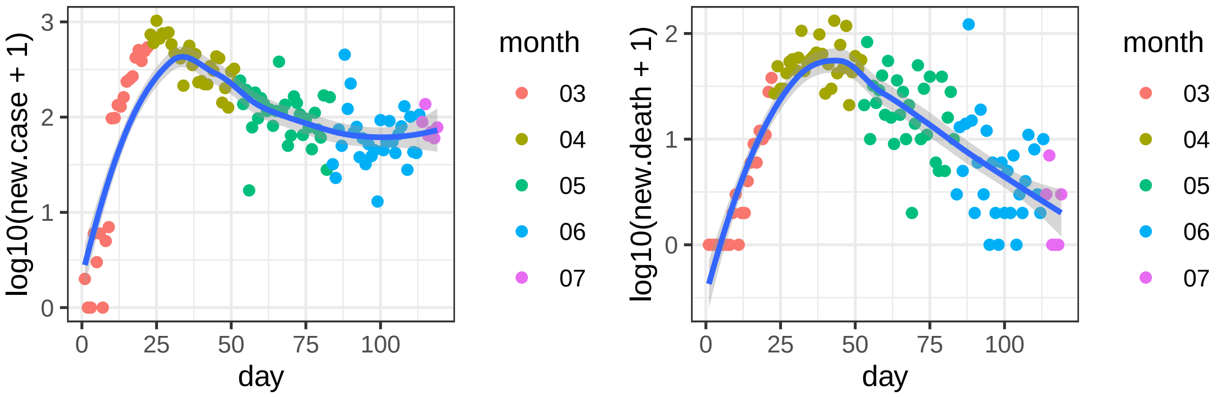
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Los Angeles_California



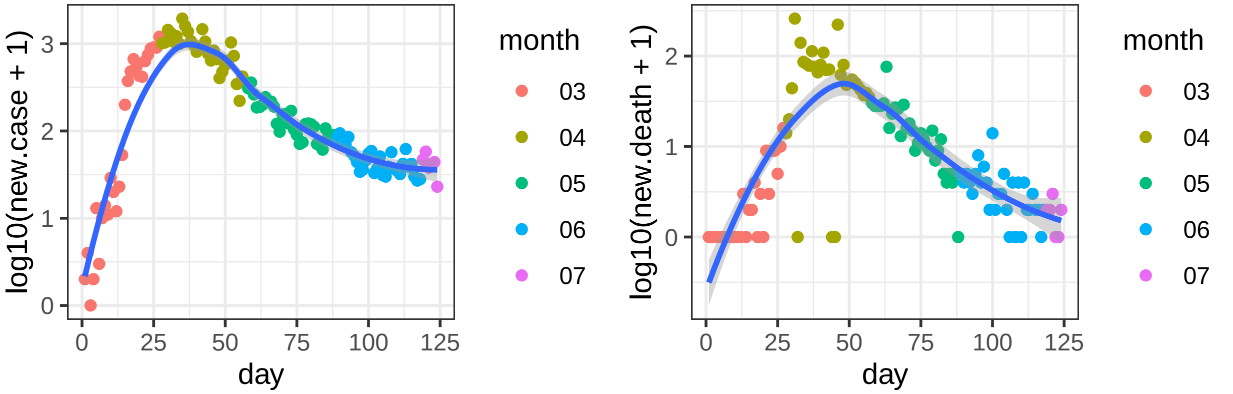
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Wayne_Michigan



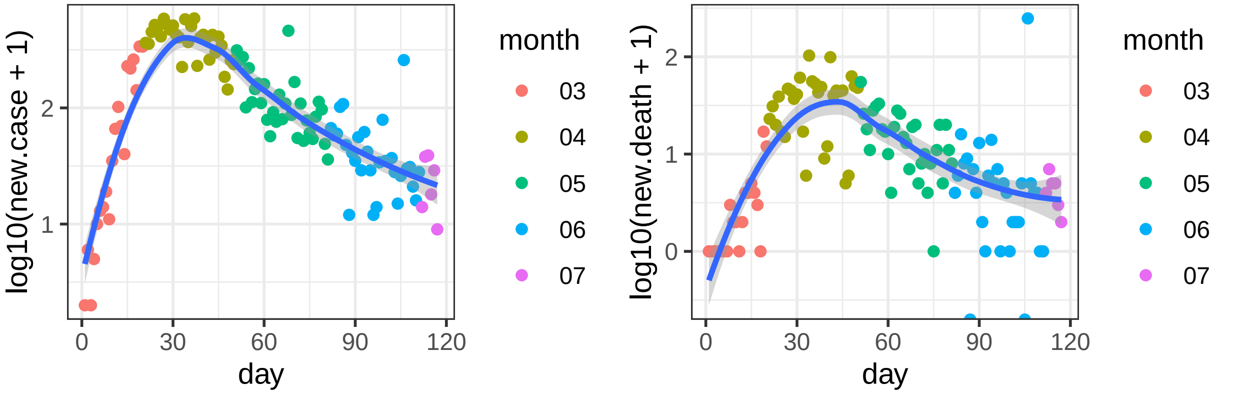
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Nassau_New York



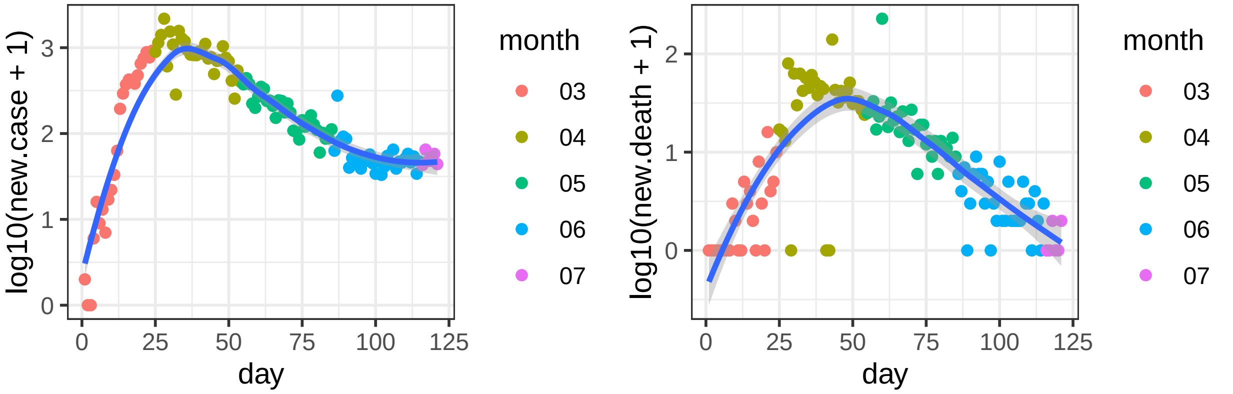
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



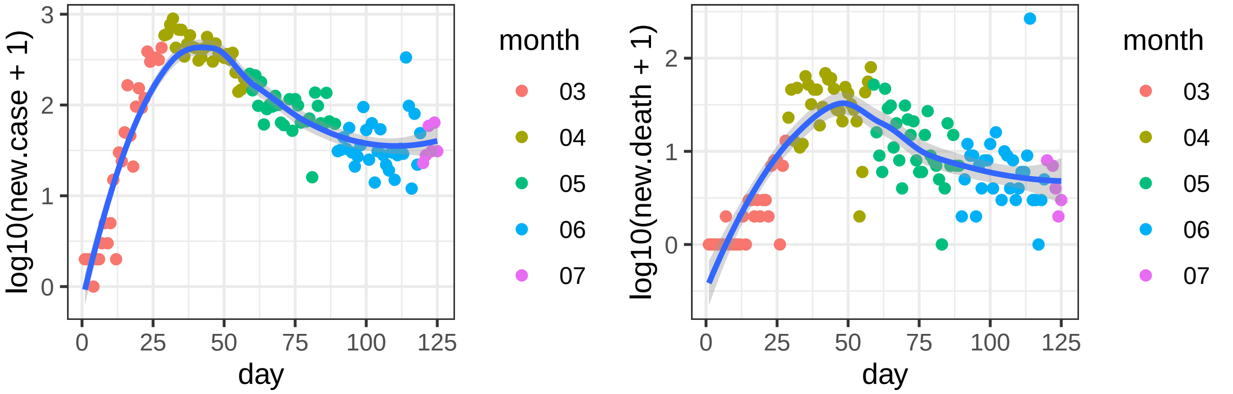
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Suffolk_New York



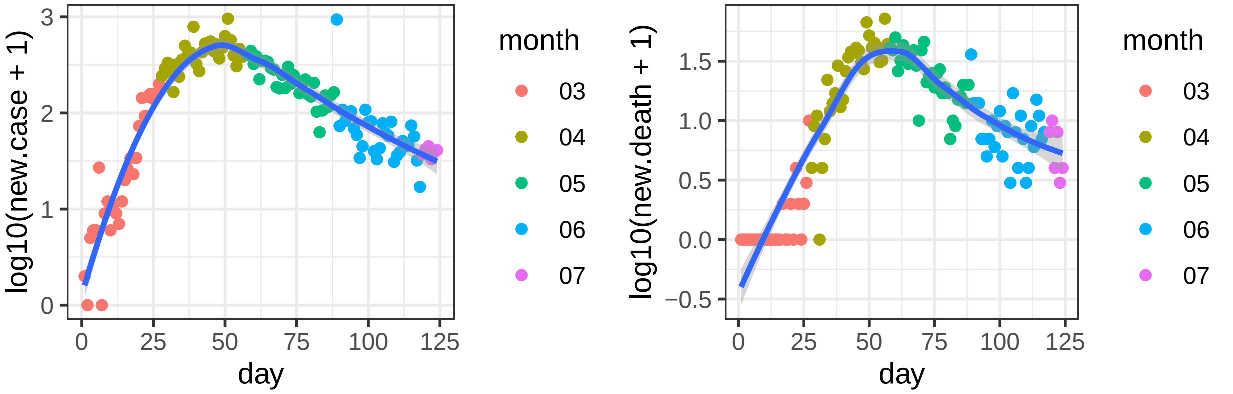
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Bergen_New Jersey



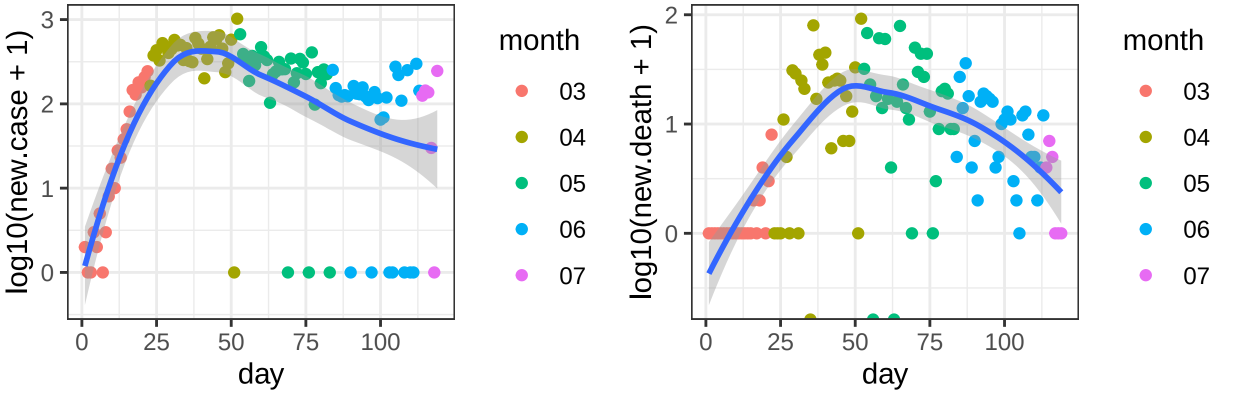
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Middlesex_Massachusetts



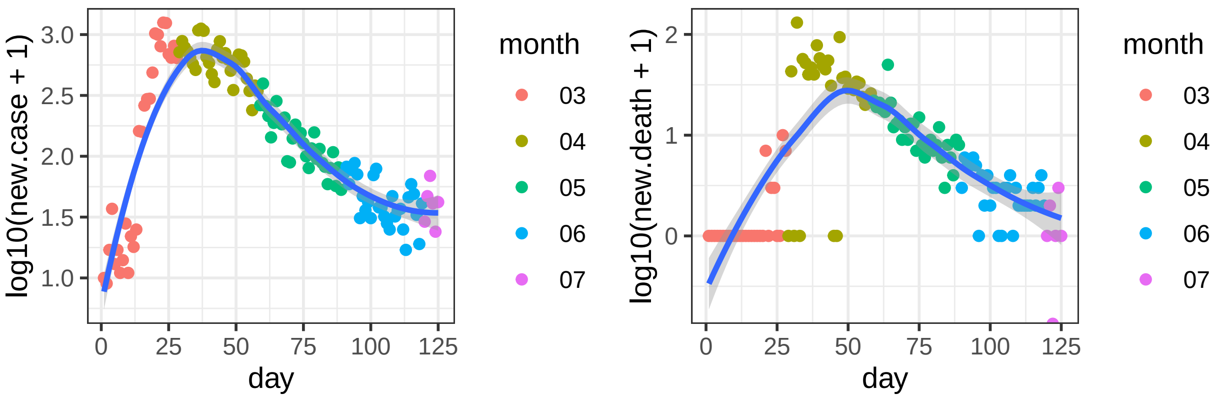
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Philadelphia_Pennsylvania



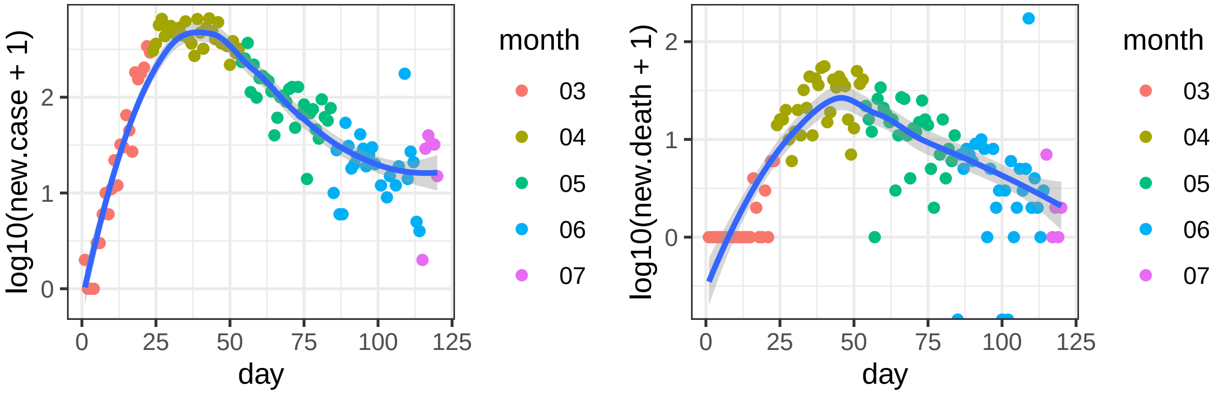
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Westchester_New York



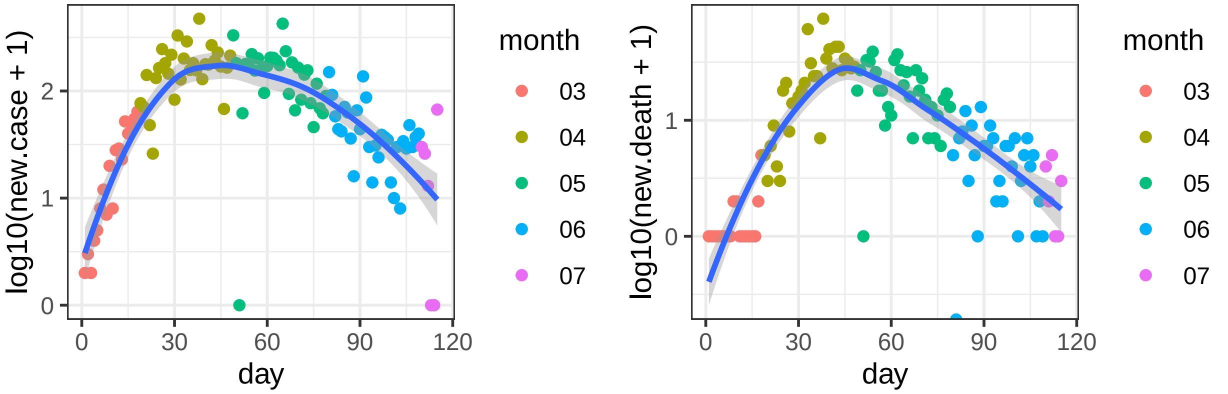
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Hudson_New Jersey



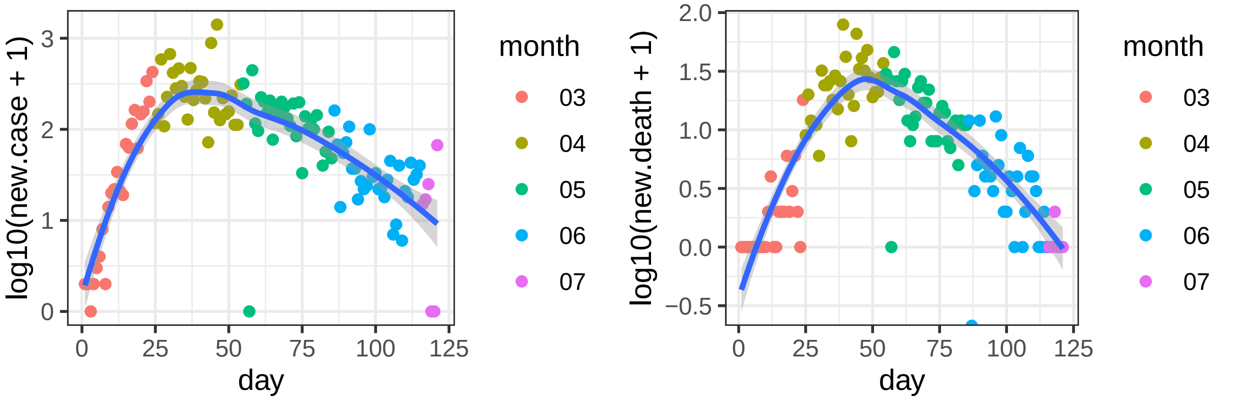
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Hartford_Connecticut



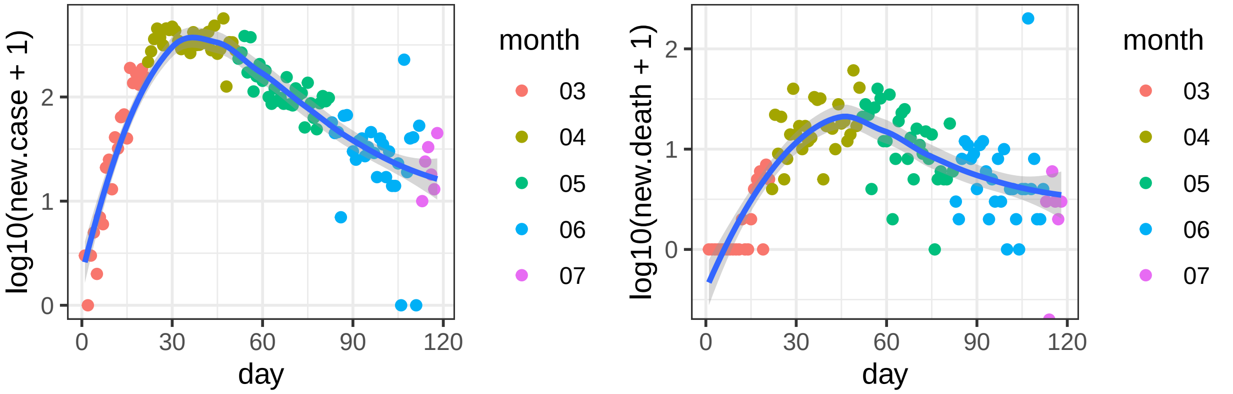
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Fairfield_Connecticut



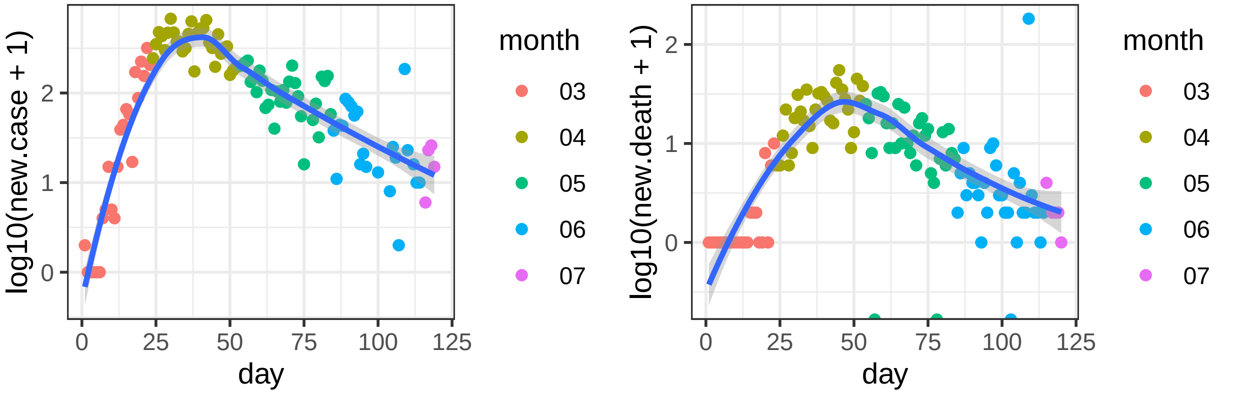
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Middlesex_New Jersey



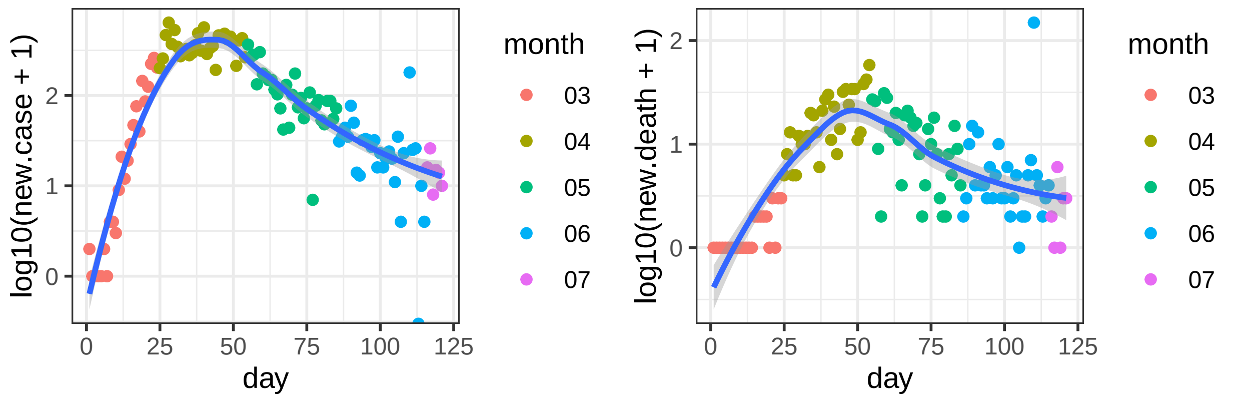
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Union_New Jersey



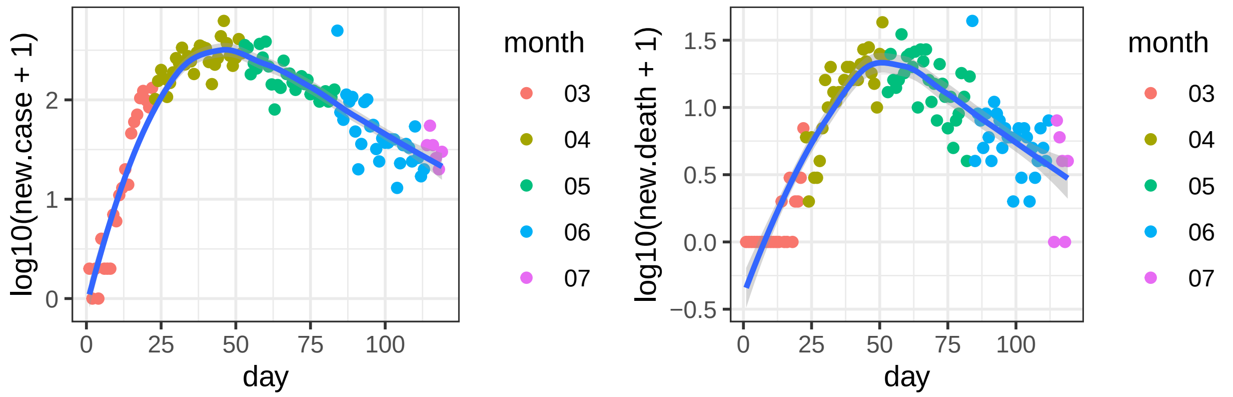
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Passaic_New Jersey



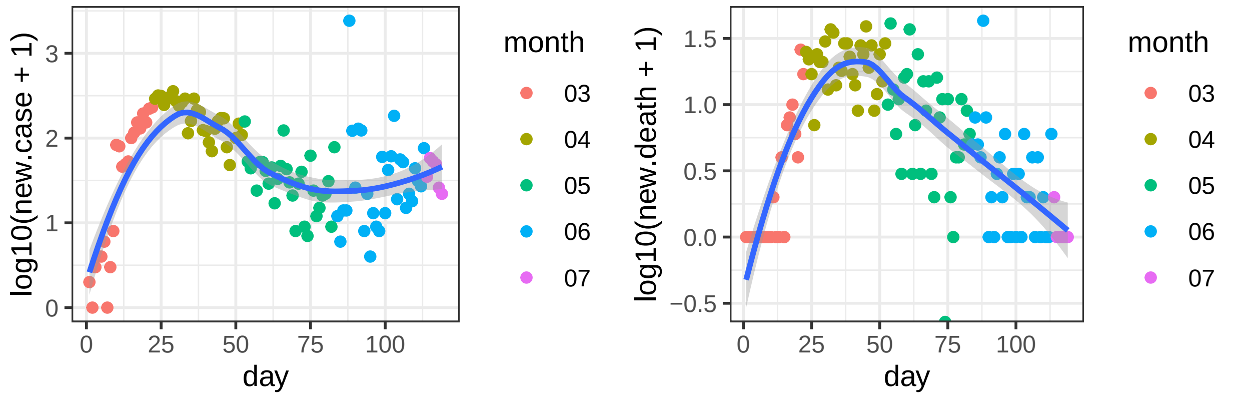
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Essex_Massachusetts



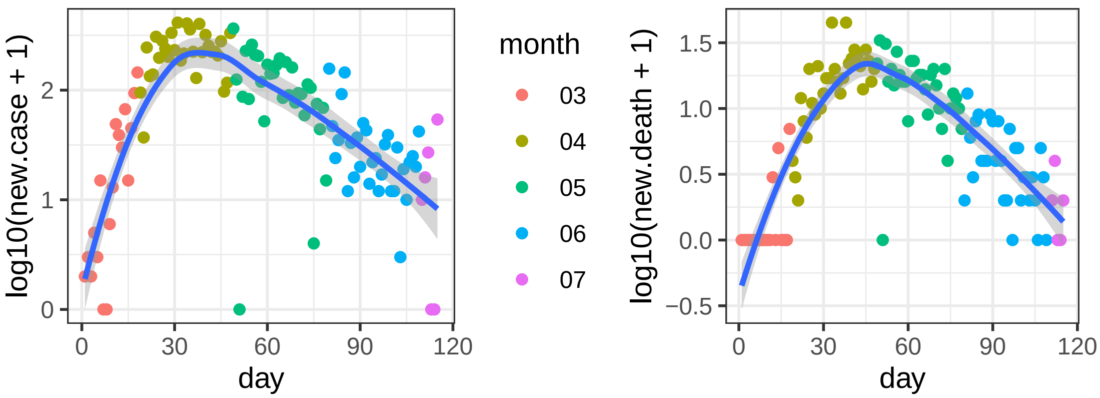
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Oakland_Michigan



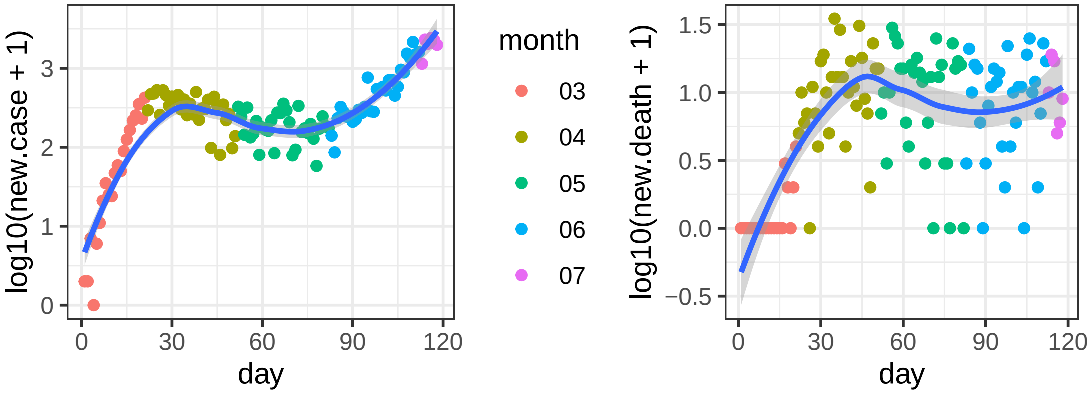
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

New Haven_Connecticut



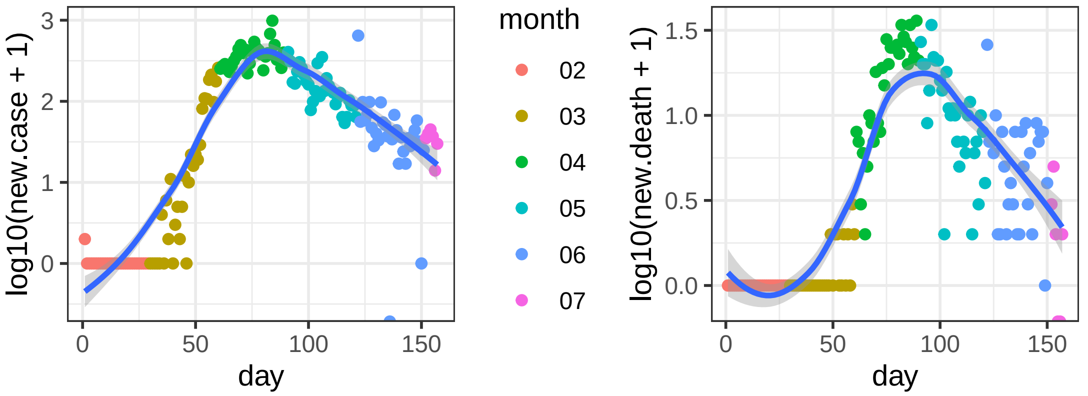
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Miami−Dade_Florida



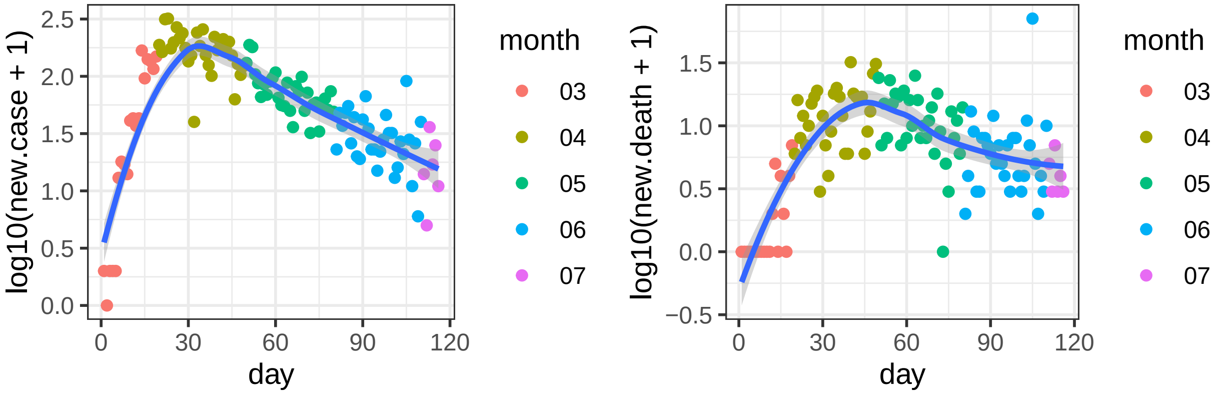
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Suffolk_Massachusetts



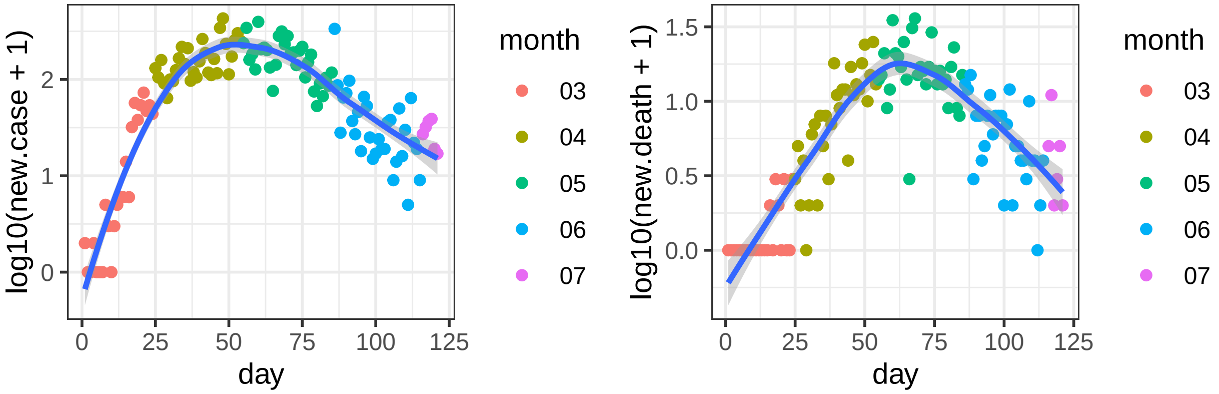
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Ocean_New Jersey



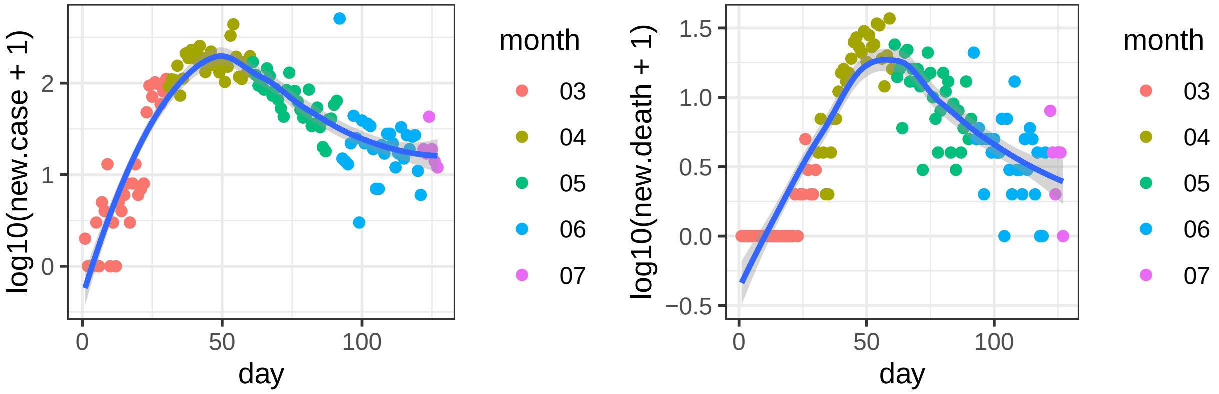
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Worcester_Massachusetts



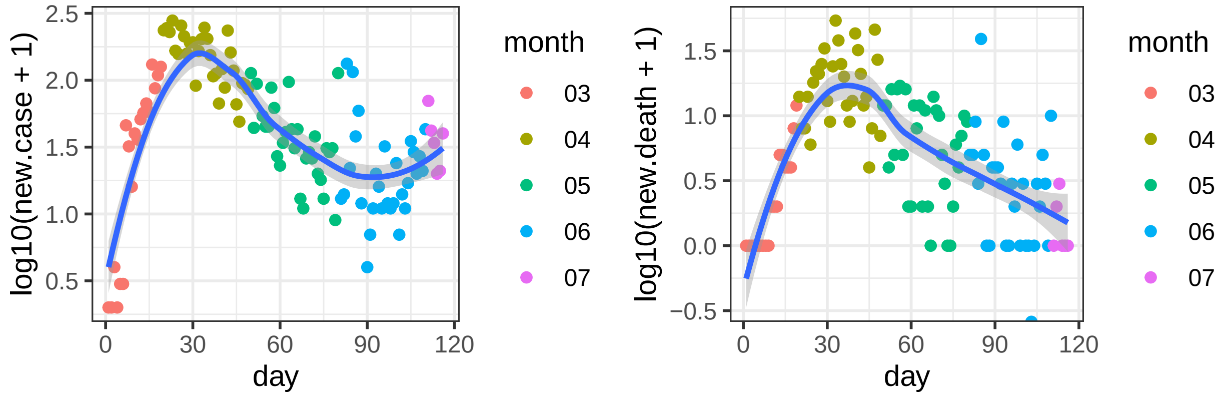
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Norfolk_Massachusetts



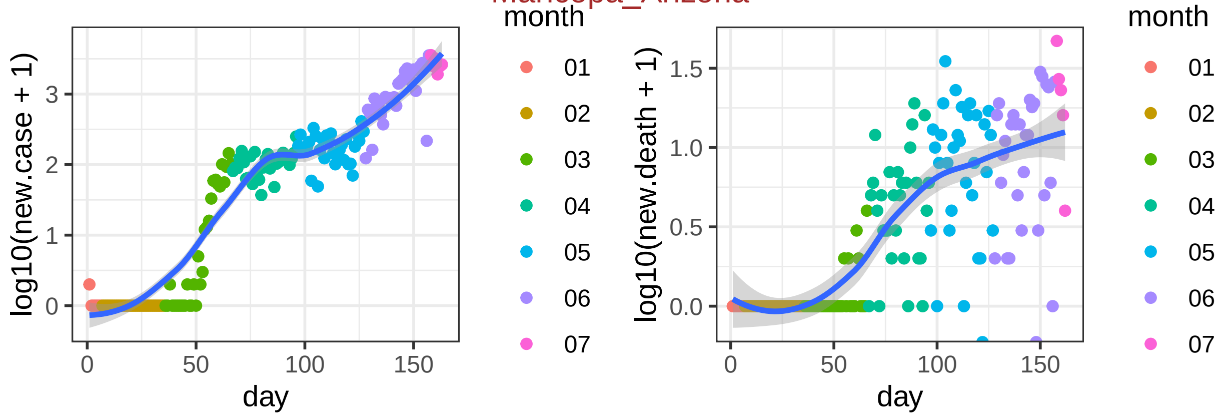
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Macomb_Michigan



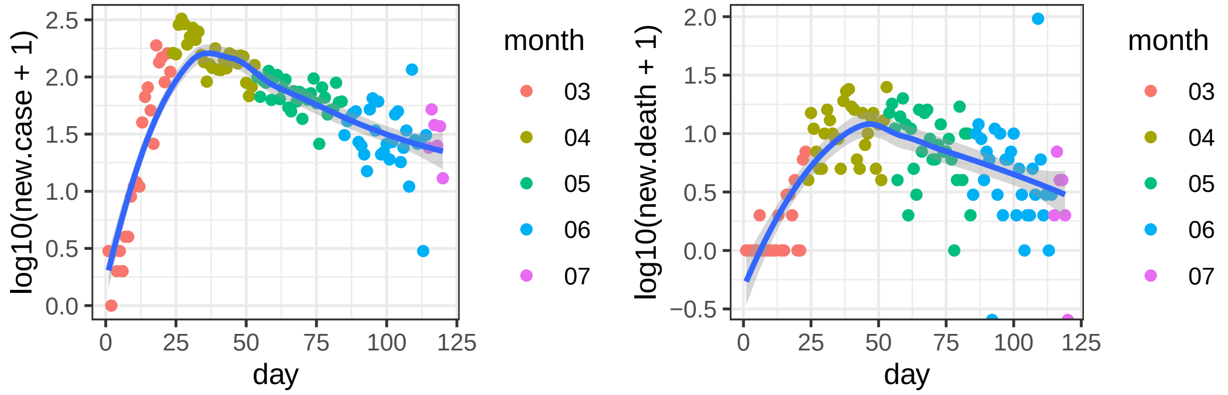
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Maricopa_Arizona



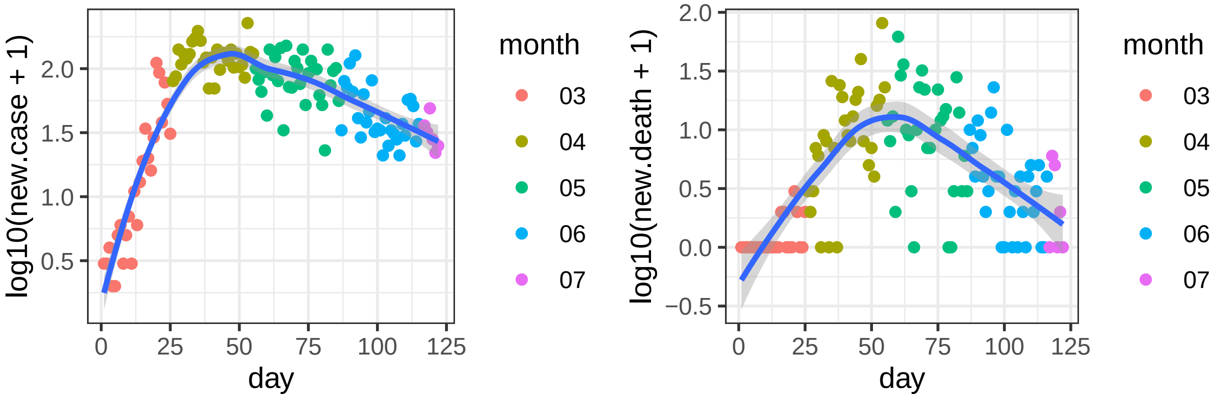
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Monmouth_New Jersey



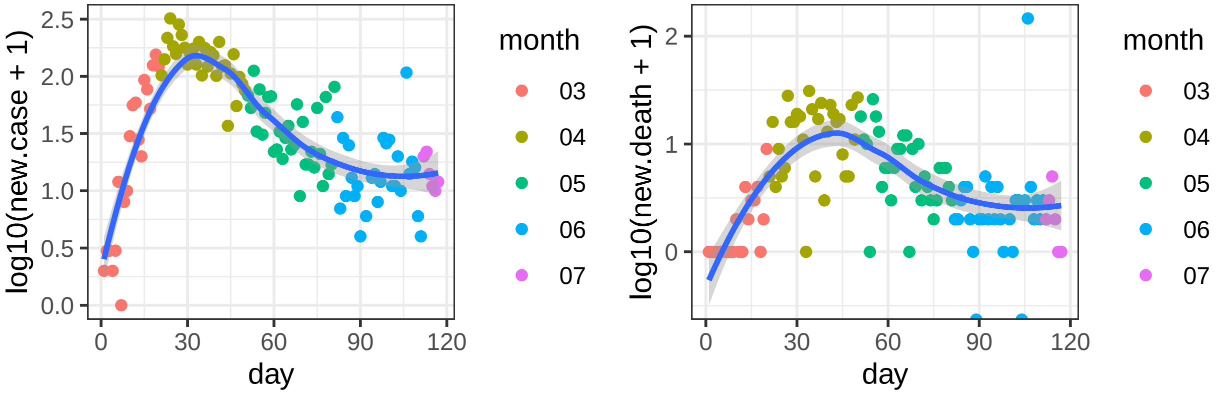
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Montgomery_Pennsylvania



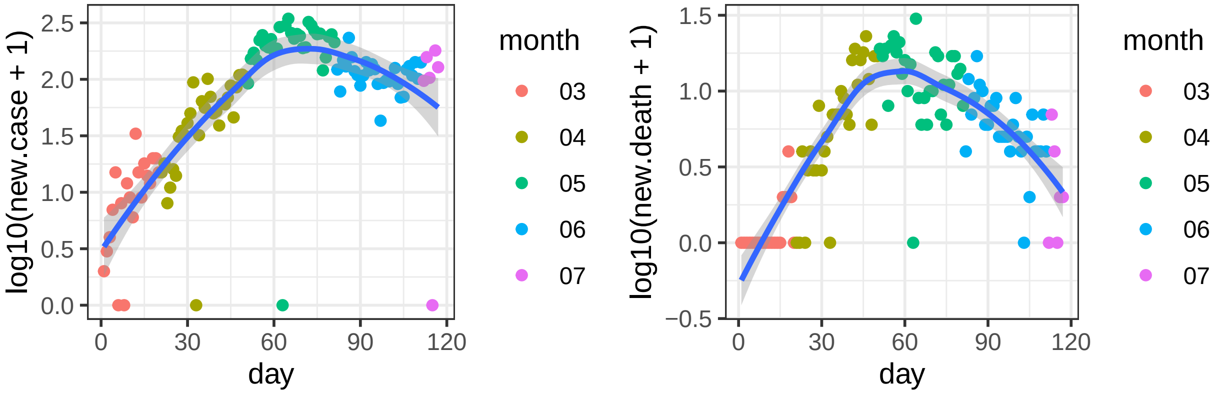
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Morris_New Jersey



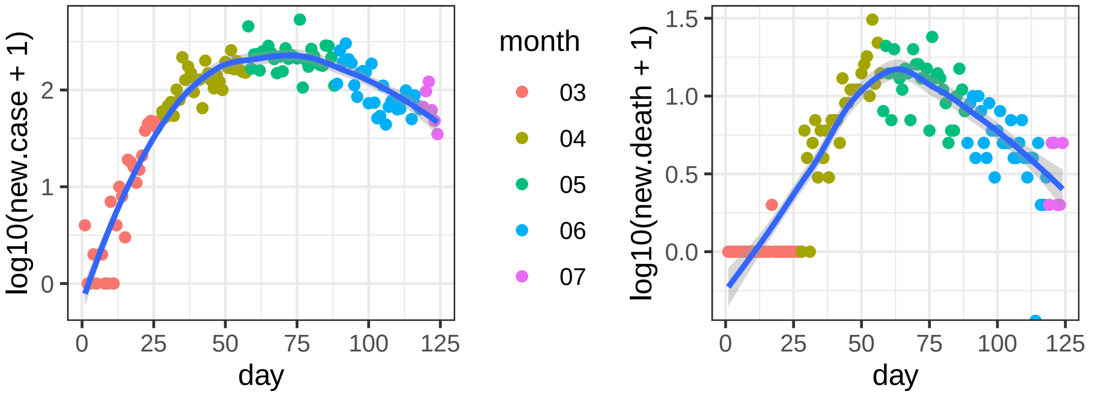
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Hennepin_Minnesota



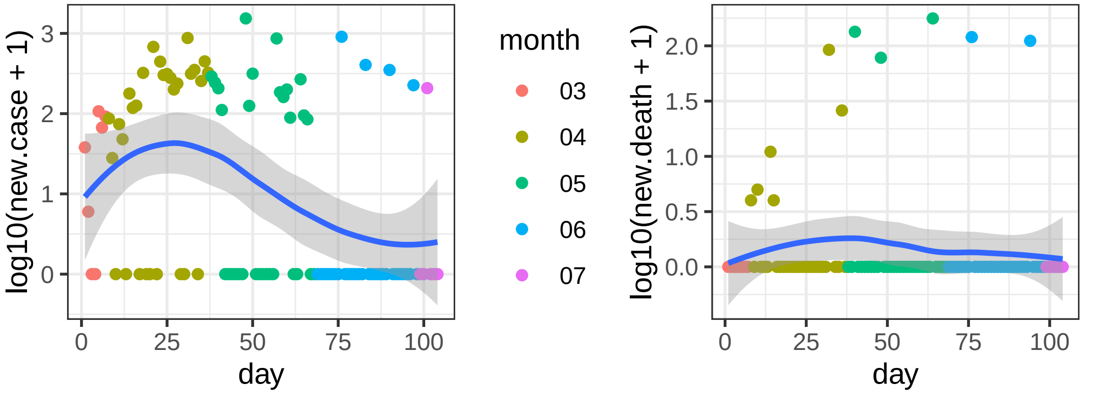
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Montgomery_Maryland



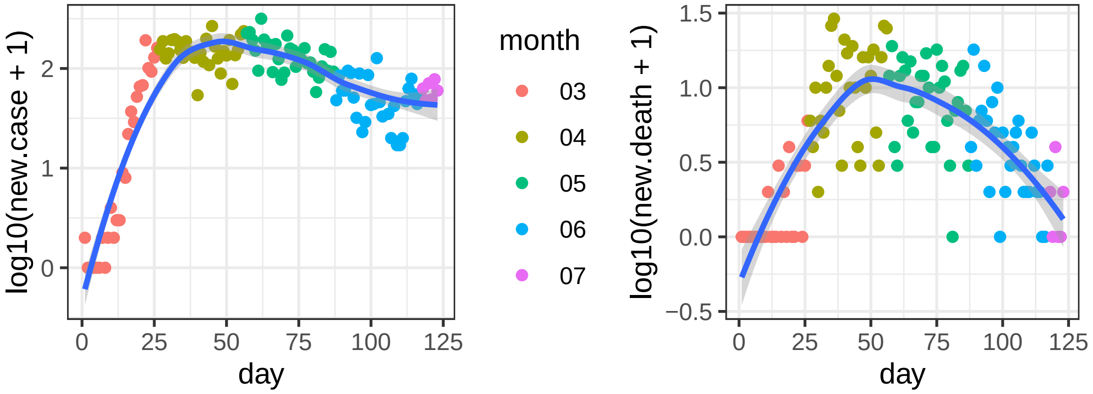
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Providence_Rhode Island



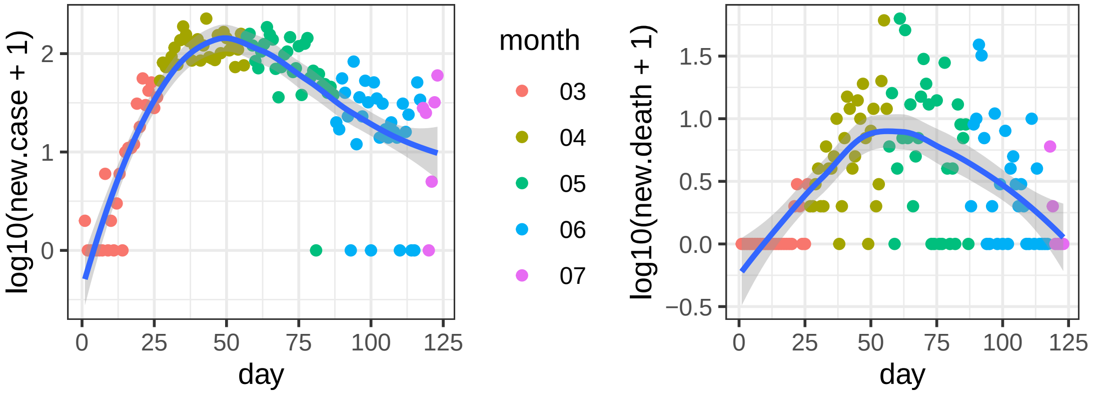
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Marion_Indiana



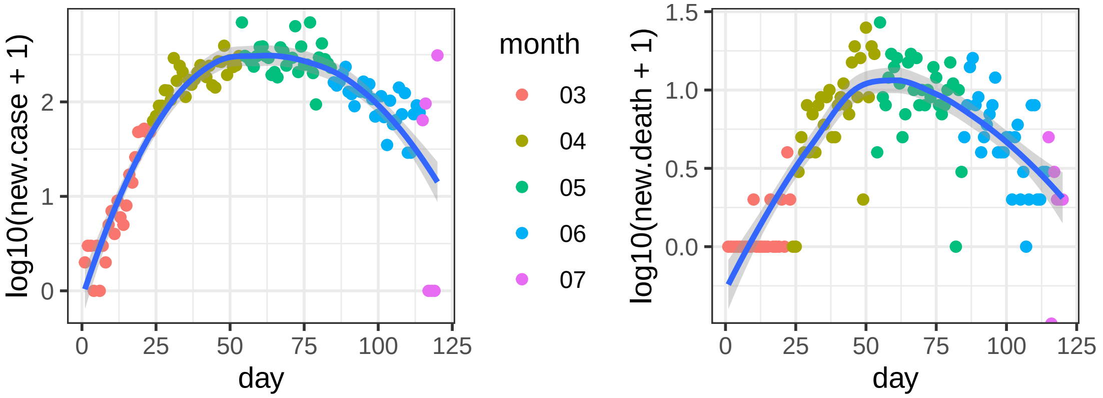
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Delaware_Pennsylvania



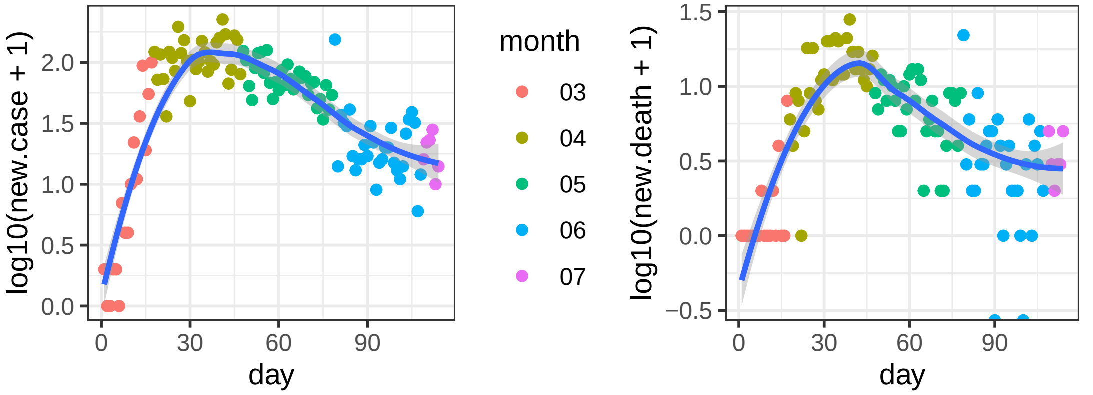
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Prince George's_Maryland



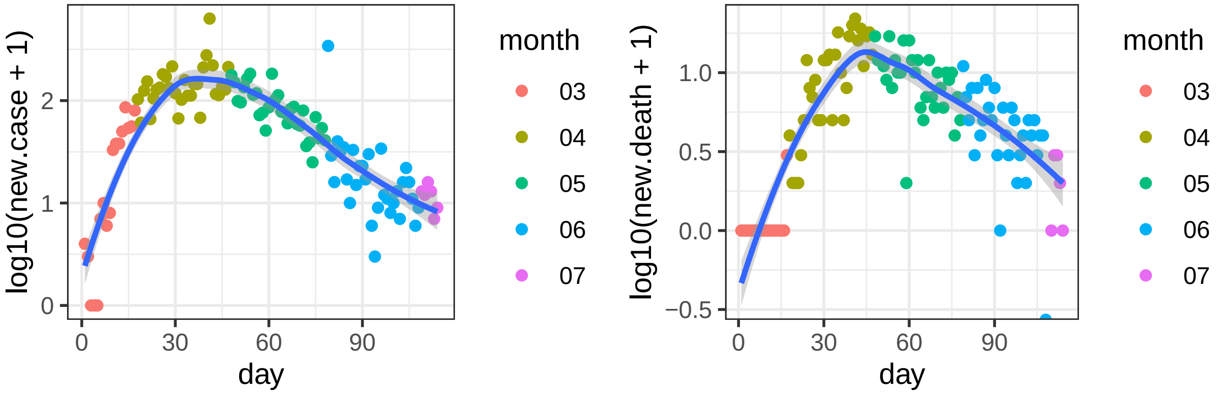
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Hampden_Massachusetts



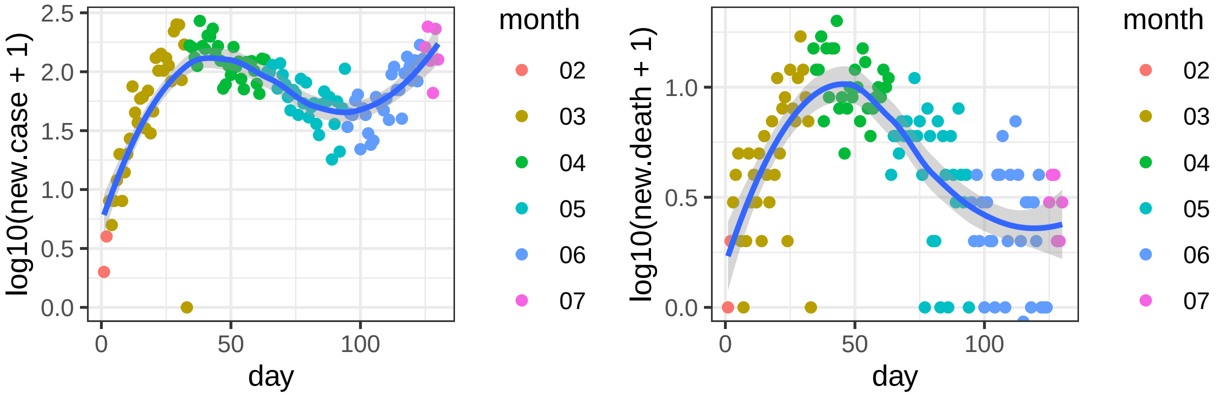
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Plymouth_Massachusetts



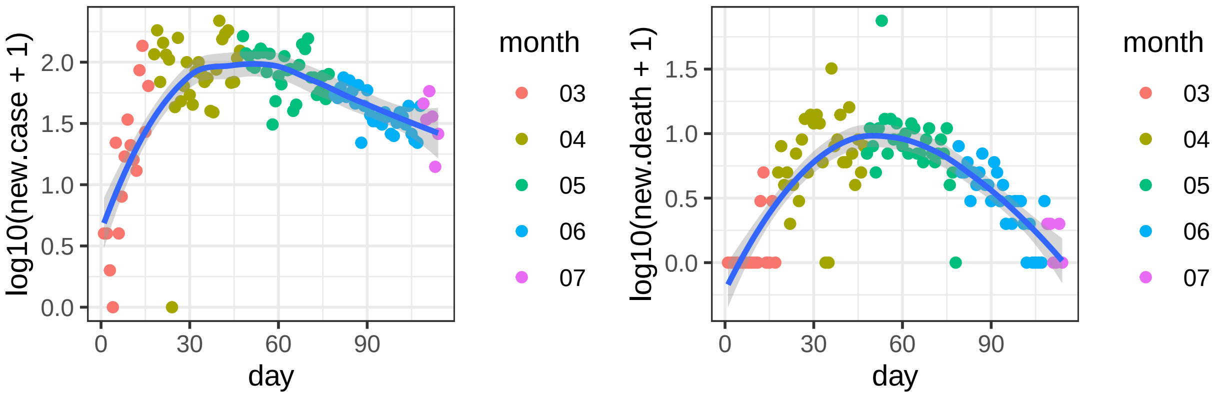
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

King_Washington



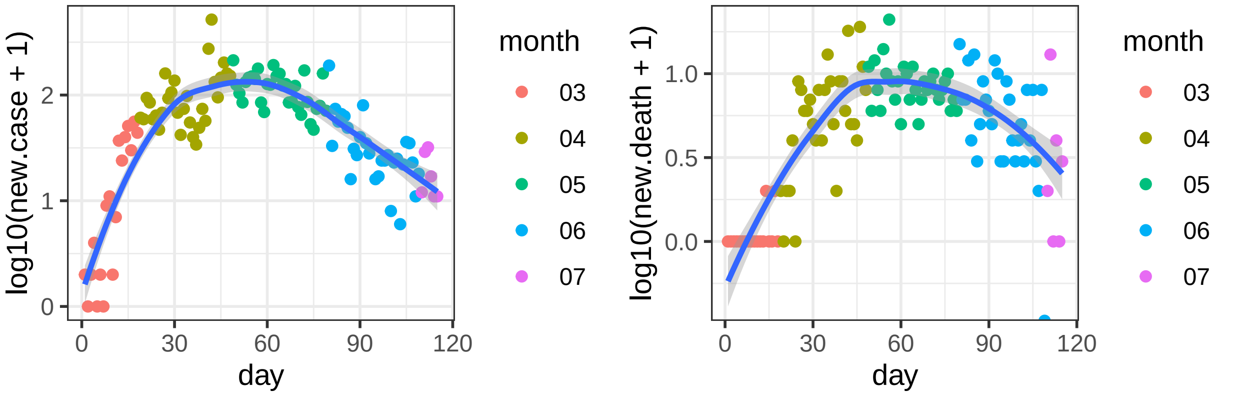
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

Erie_New York



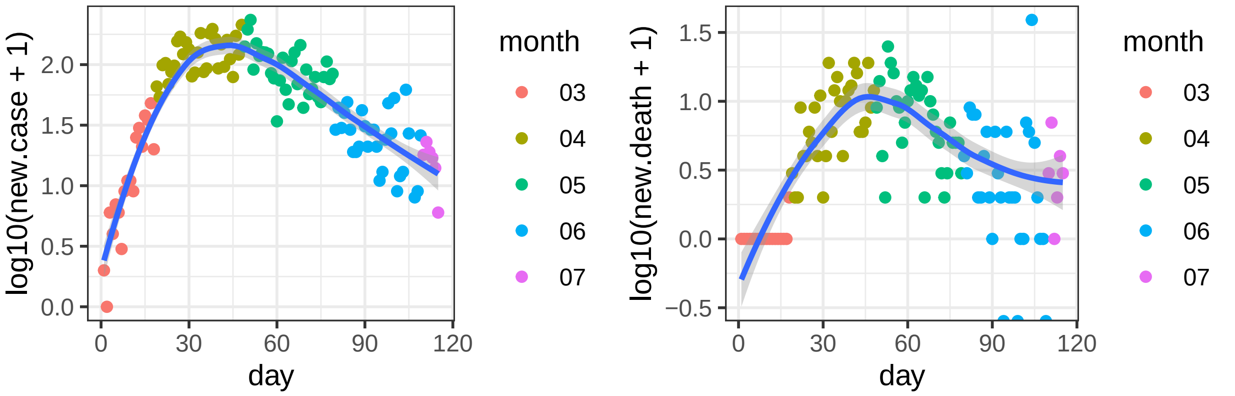
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Bristol_Massachusetts



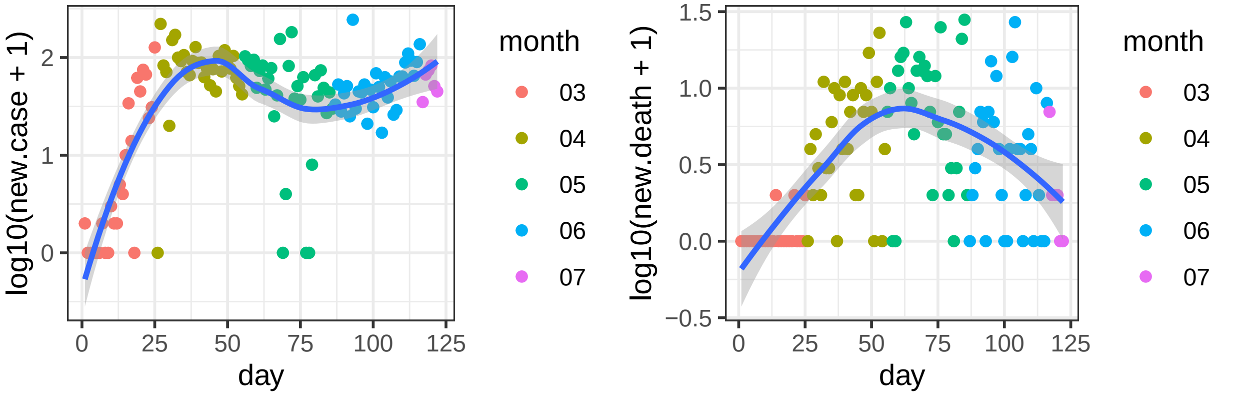
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Mercer_New Jersey



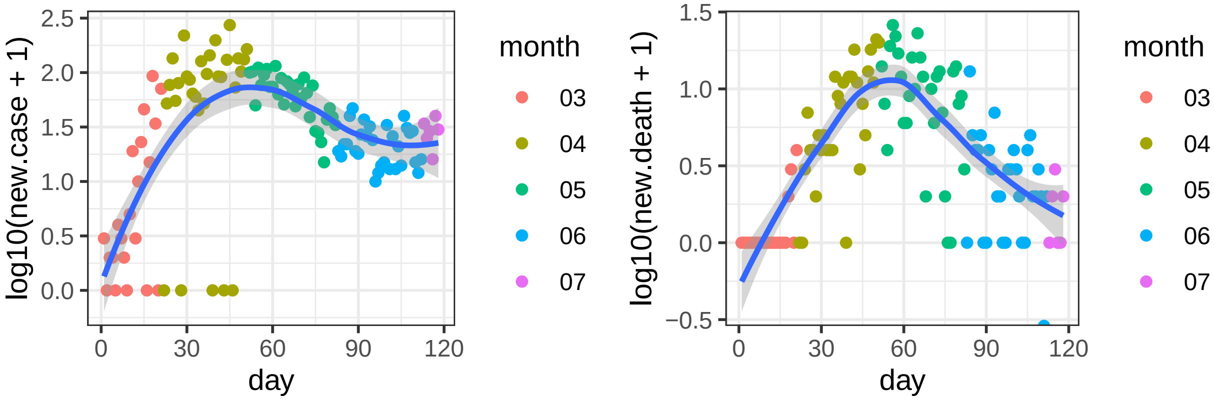
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

St. Louis_Missouri



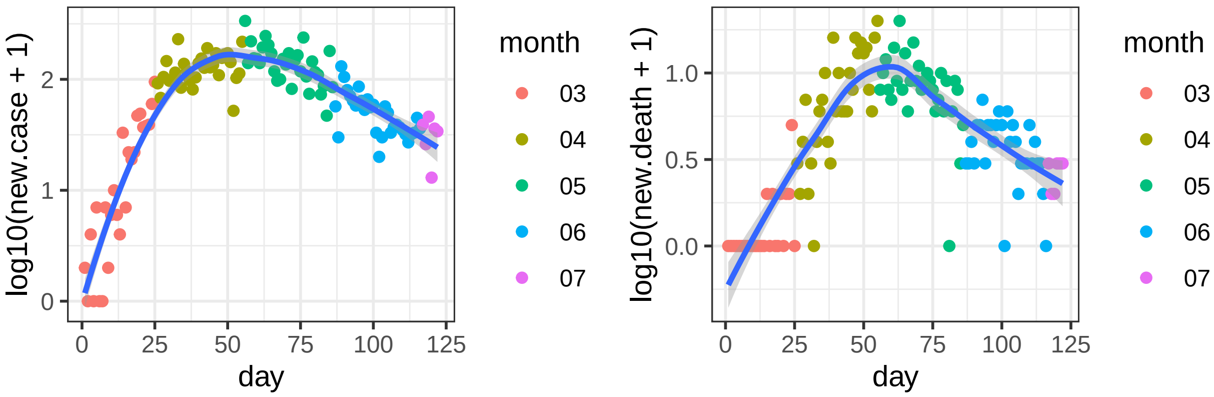
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Bucks_Pennsylvania



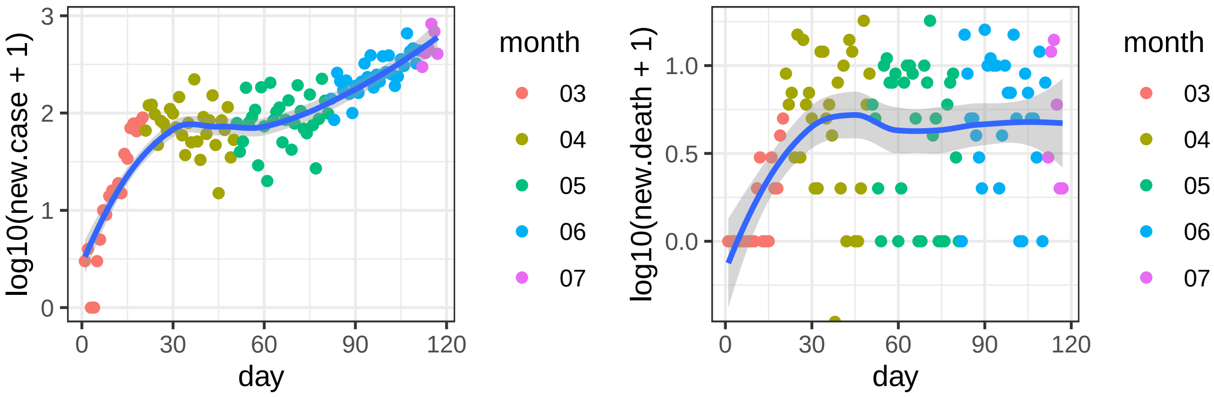
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

District of Columbia_District of Columbia



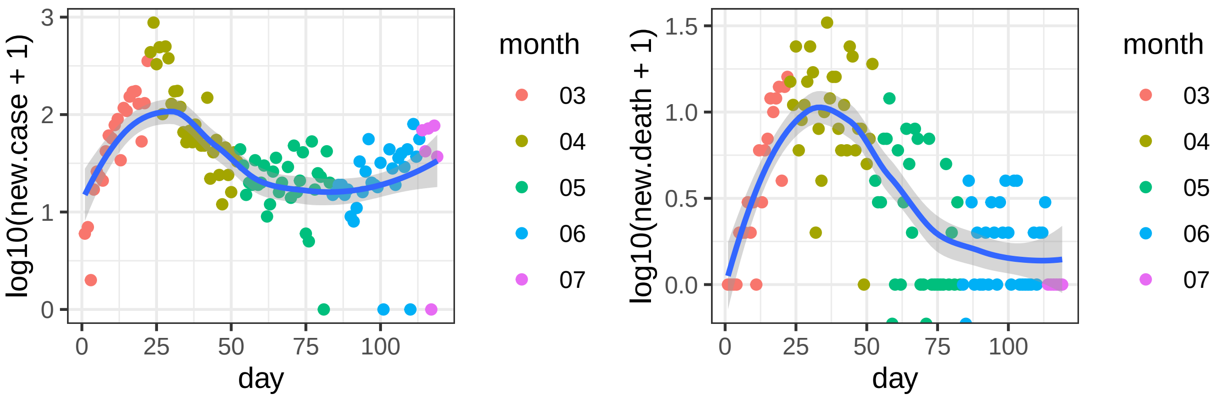
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Palm Beach_Florida



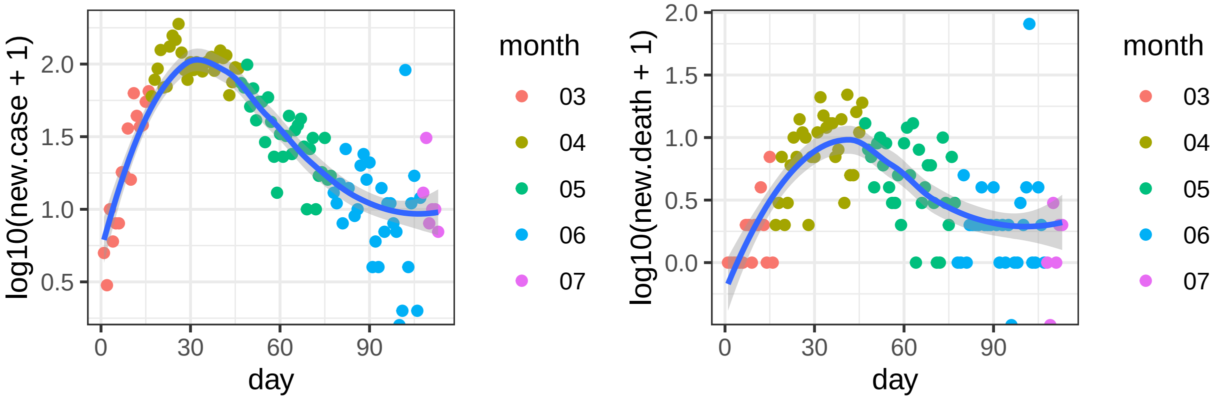
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Orleans_Louisiana



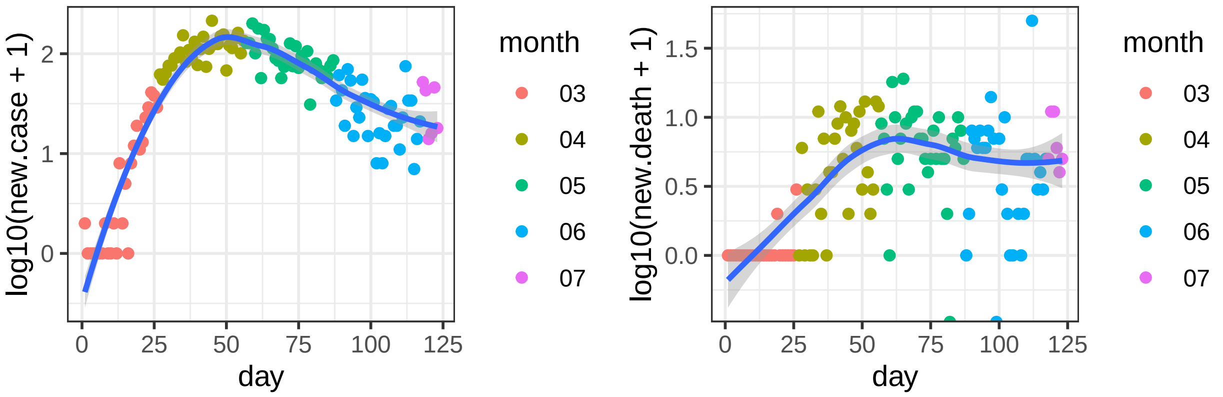
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Somerset_New Jersey



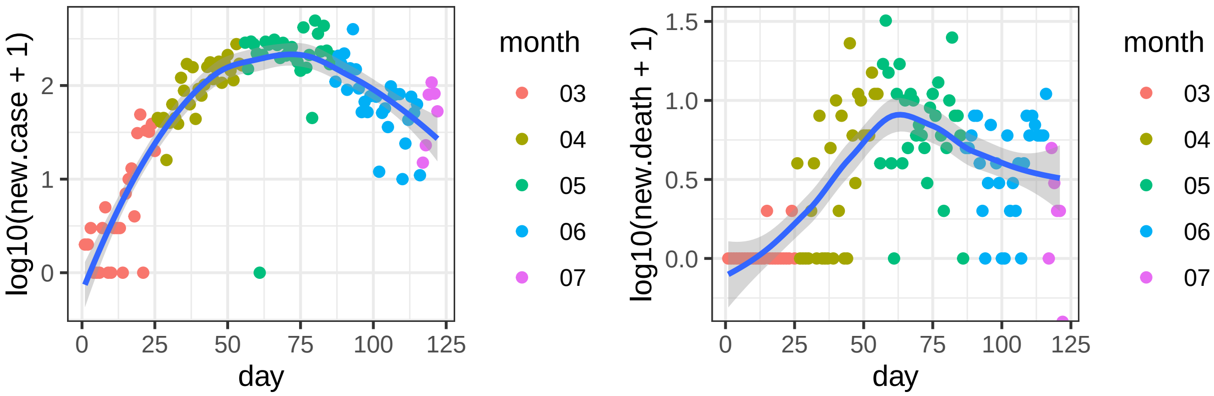
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−16

Camden_New Jersey



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

Fairfax_Virginia

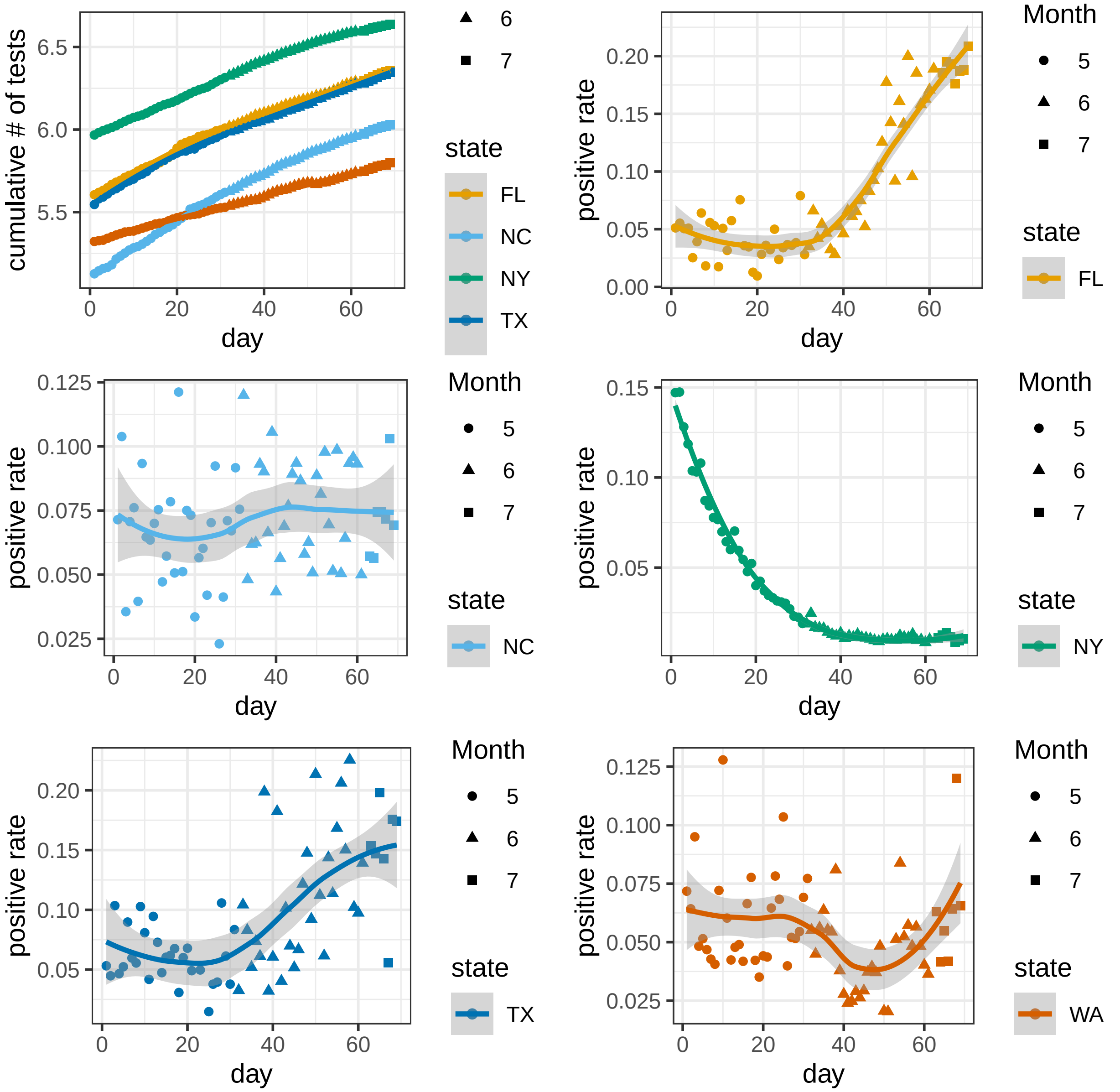


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

COVID Trackng

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

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



github.com/COVID19Tracking/, positive rate on 0707: 0.21(FL) 0.07(NC) 0.01(NY) 0.17(TX) 0.07(WA)

Session information

```
sessionInfo()
```

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

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