Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-10-11

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

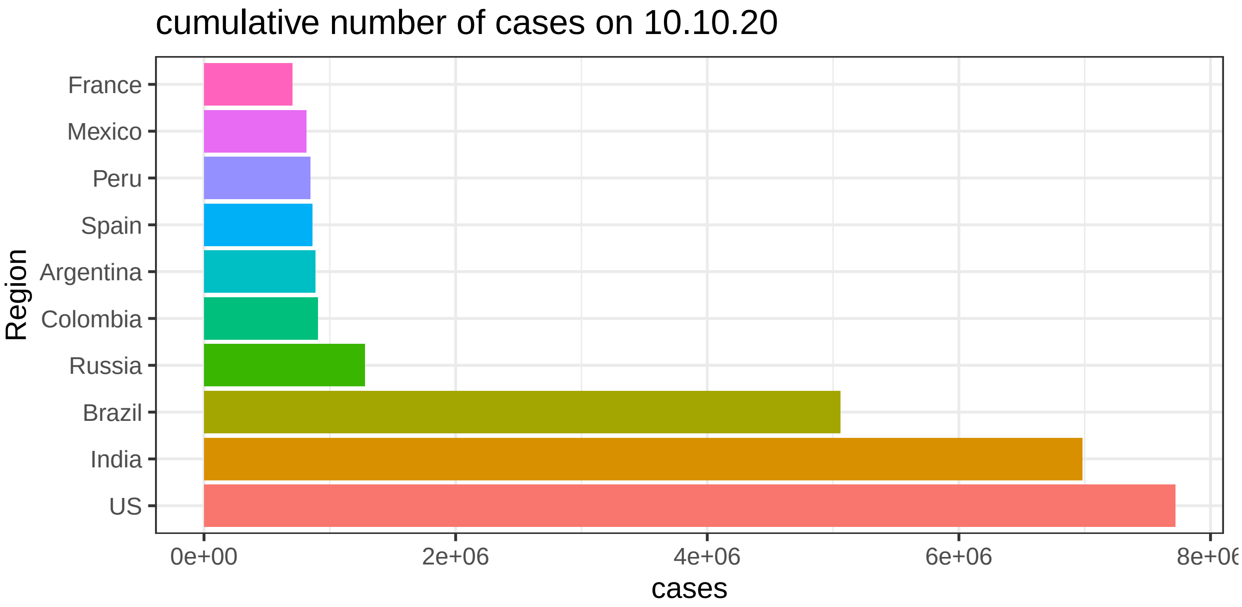
JHU

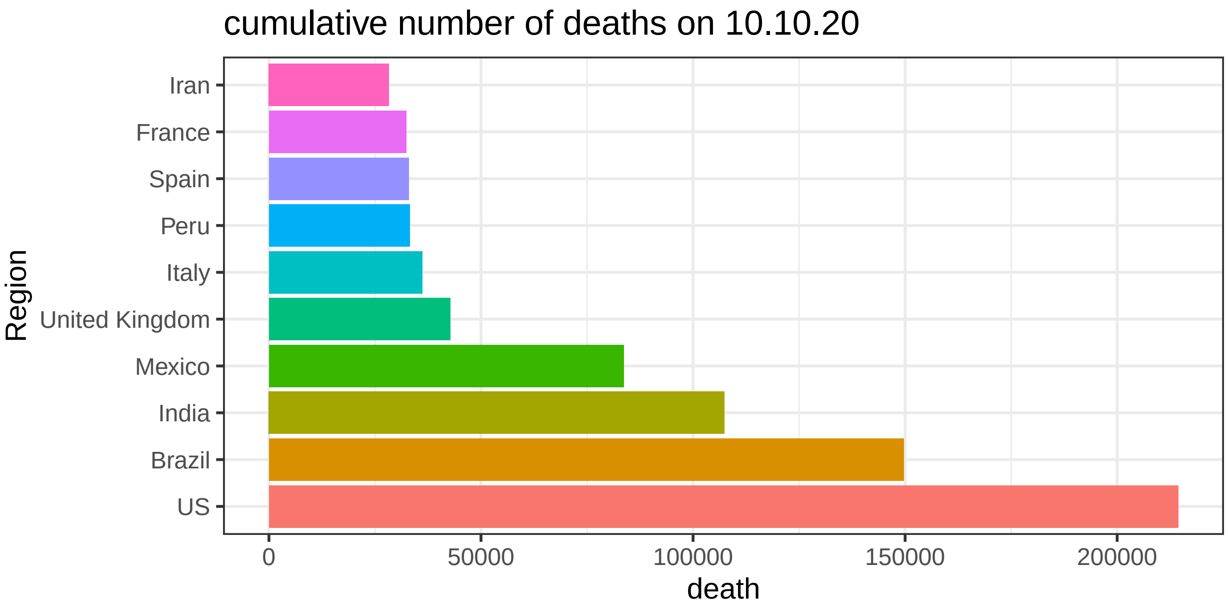
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

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.

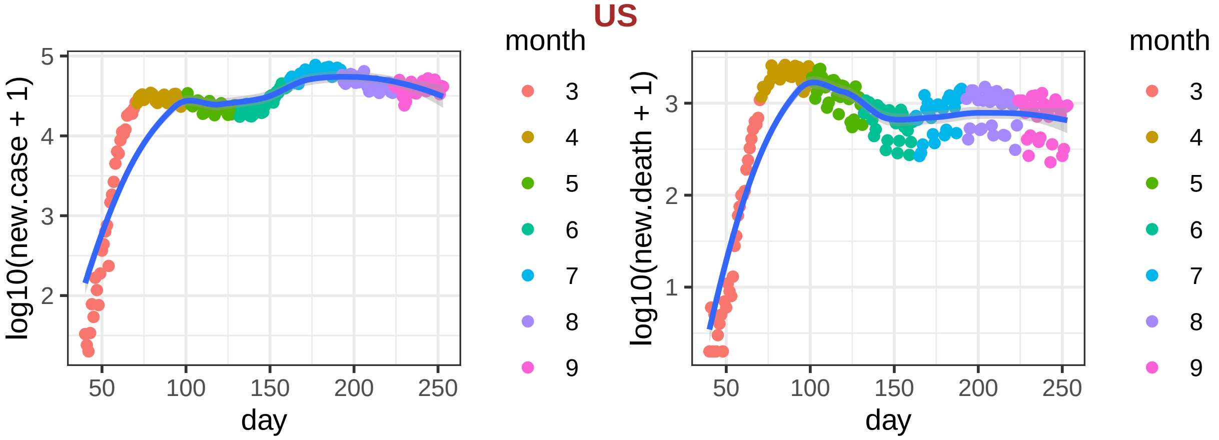
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





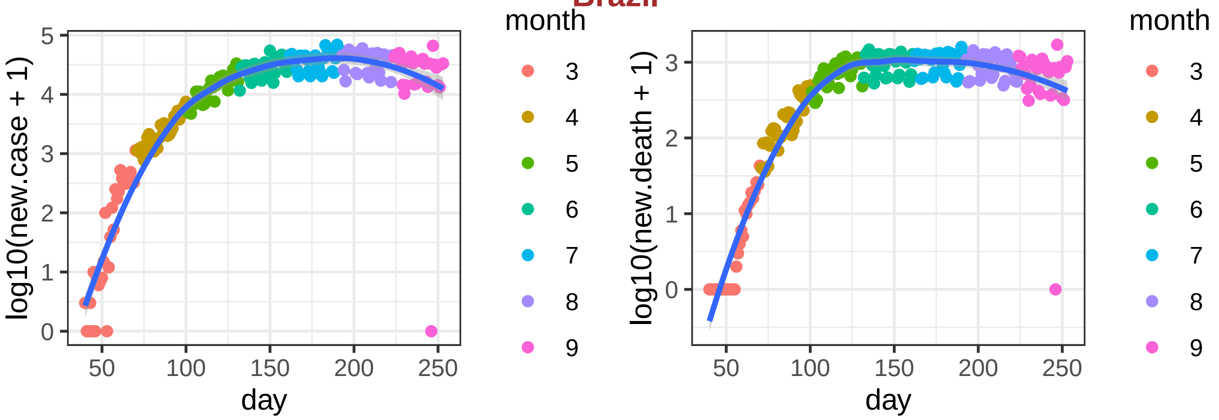
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US

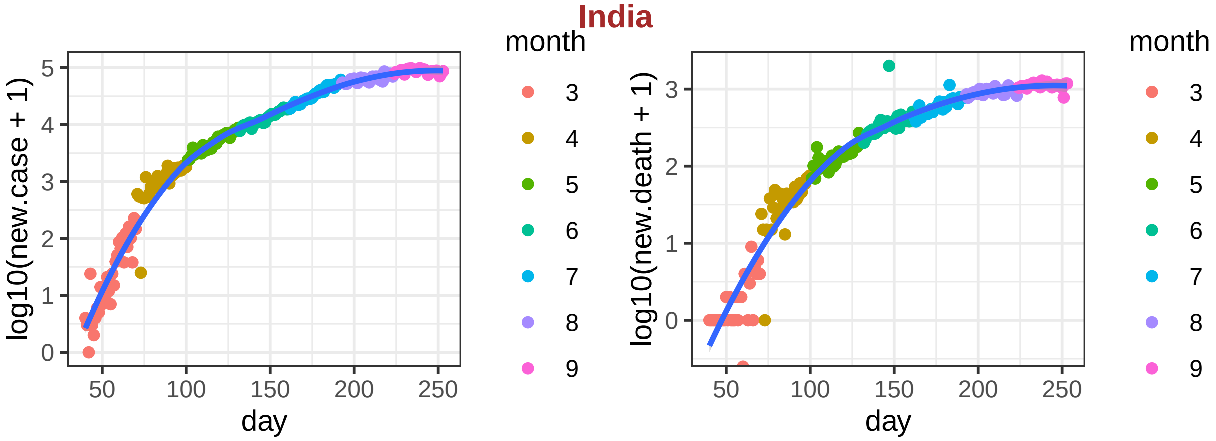


data source: https://github.com/CSSEGISandData/COVID−19

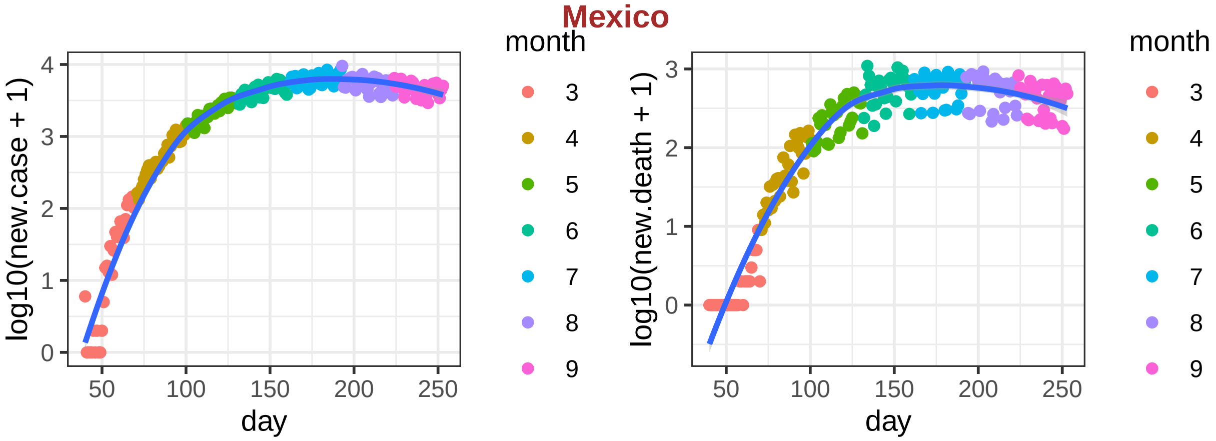
Brazil



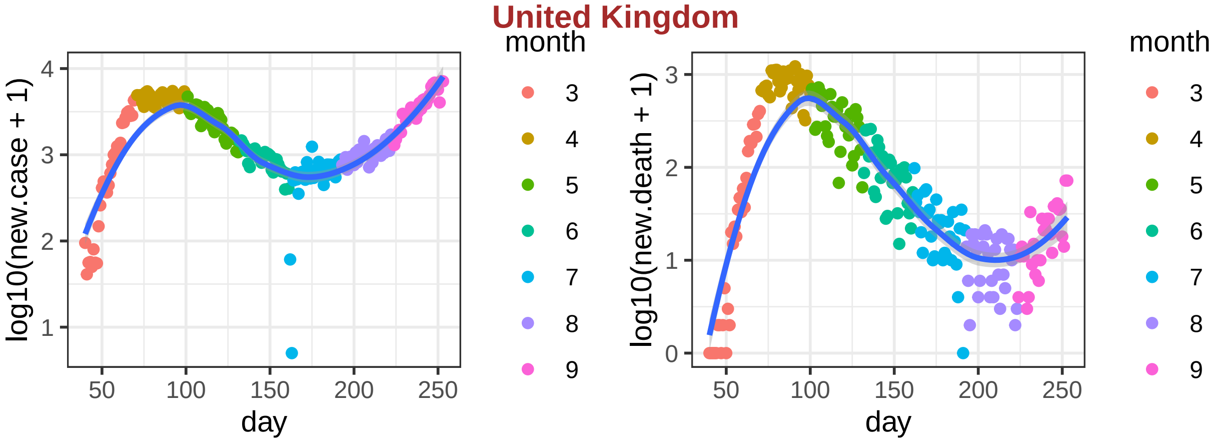
data source: https://github.com/CSSEGISandData/COVID−19



data source: https://github.com/CSSEGISandData/COVID−19

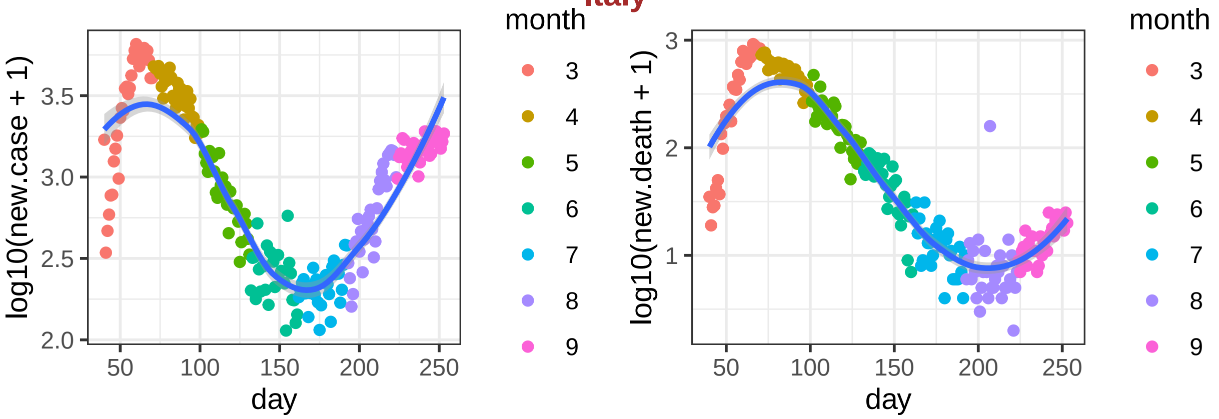


data source: https://github.com/CSSEGISandData/COVID−19



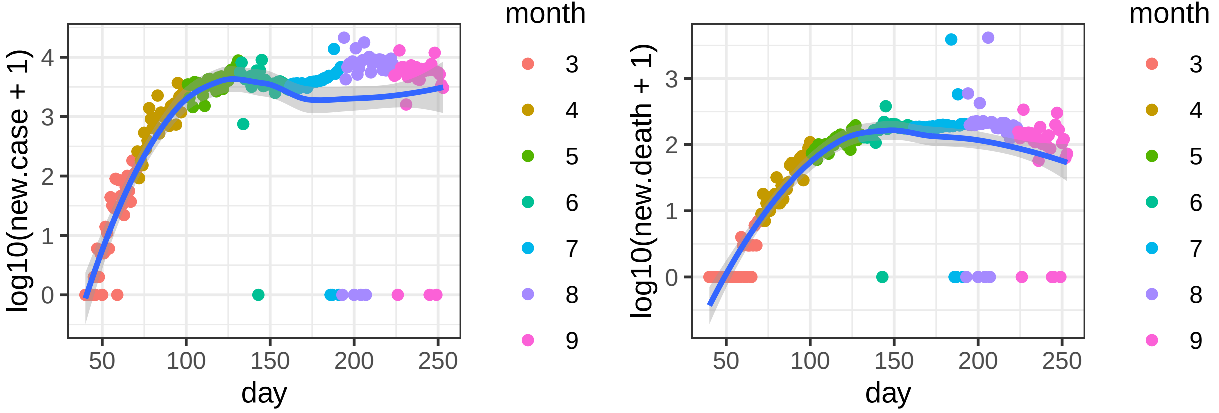
data source: https://github.com/CSSEGISandData/COVID−19

Italy



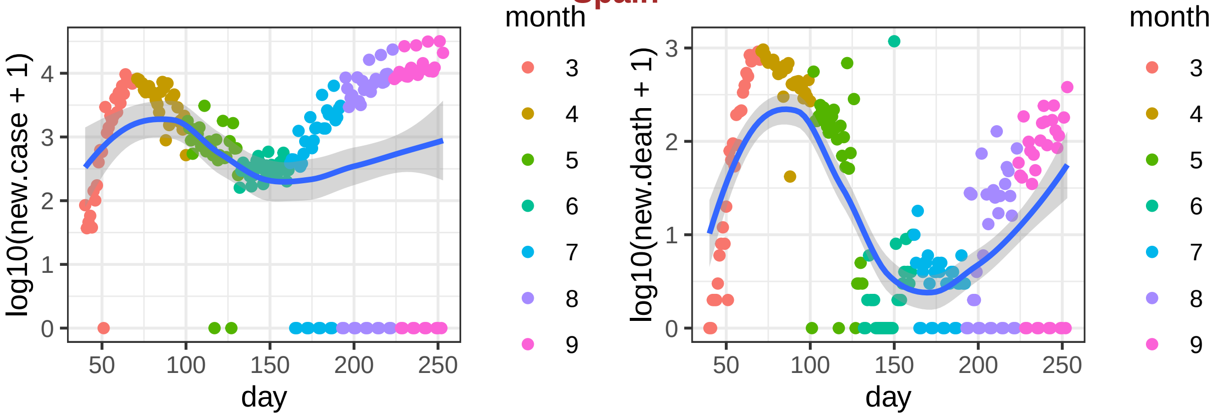
data source: https://github.com/CSSEGISandData/COVID−19

Peru

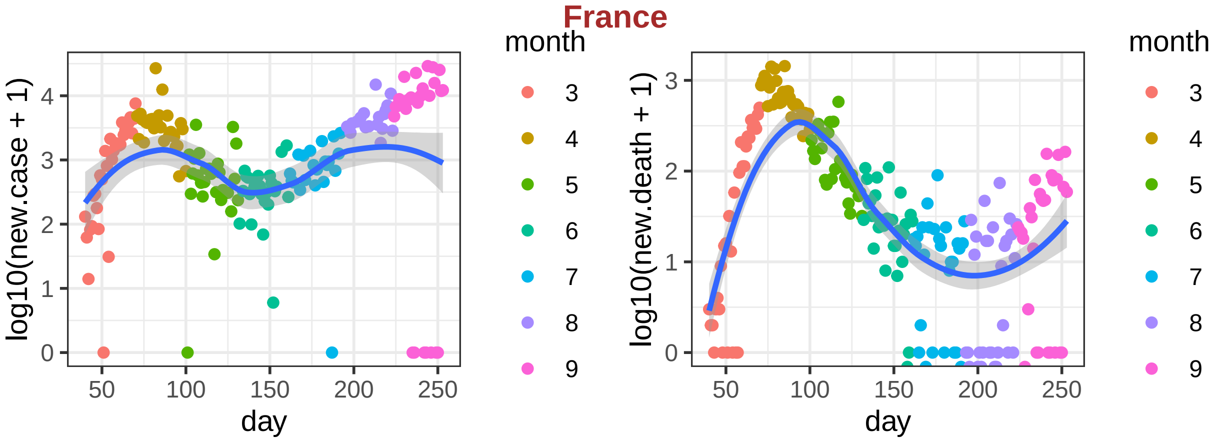


data source: https://github.com/CSSEGISandData/COVID−19

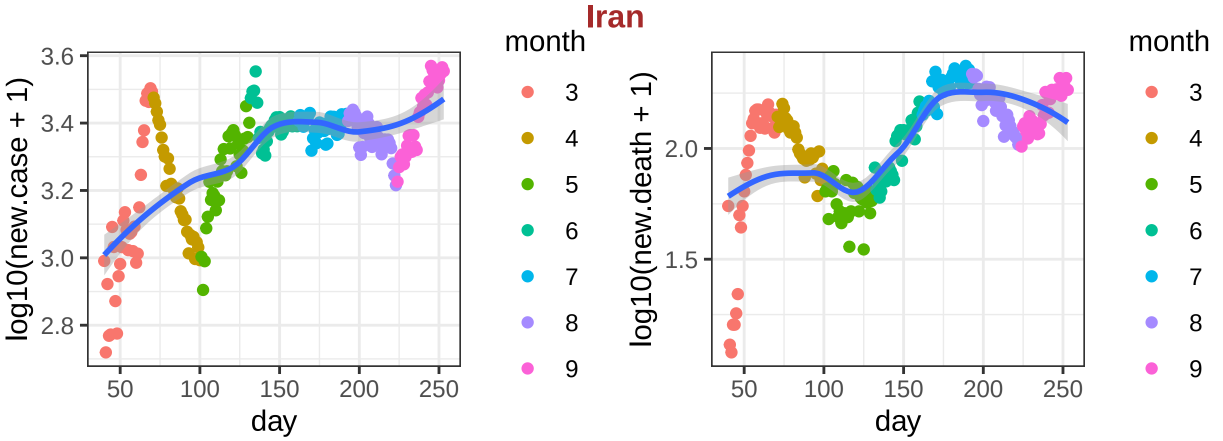
Spain



data source: https://github.com/CSSEGISandData/COVID−19



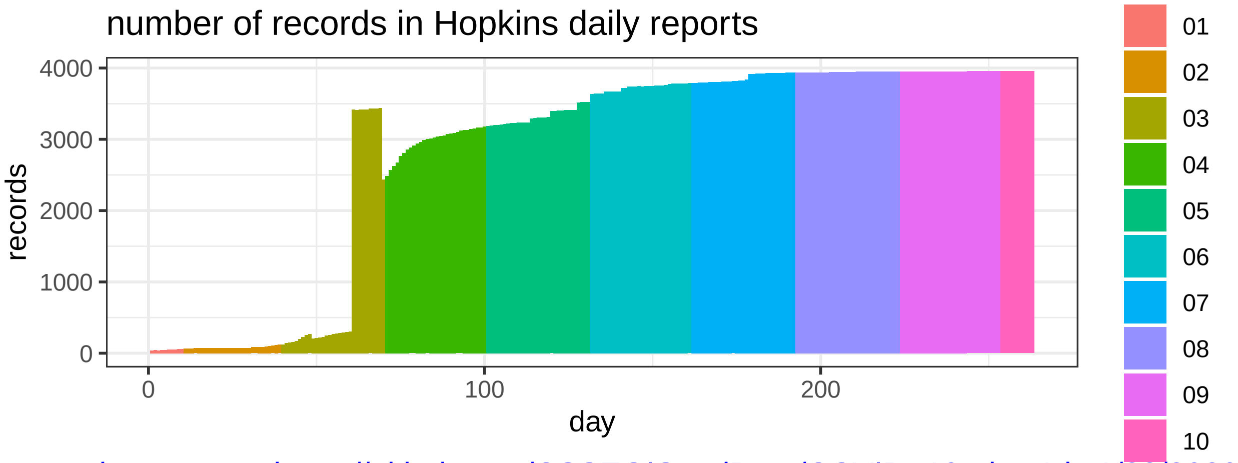
data source: https://github.com/CSSEGISandData/COVID−19



data source: https://github.com/CSSEGISandData/COVID−19

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

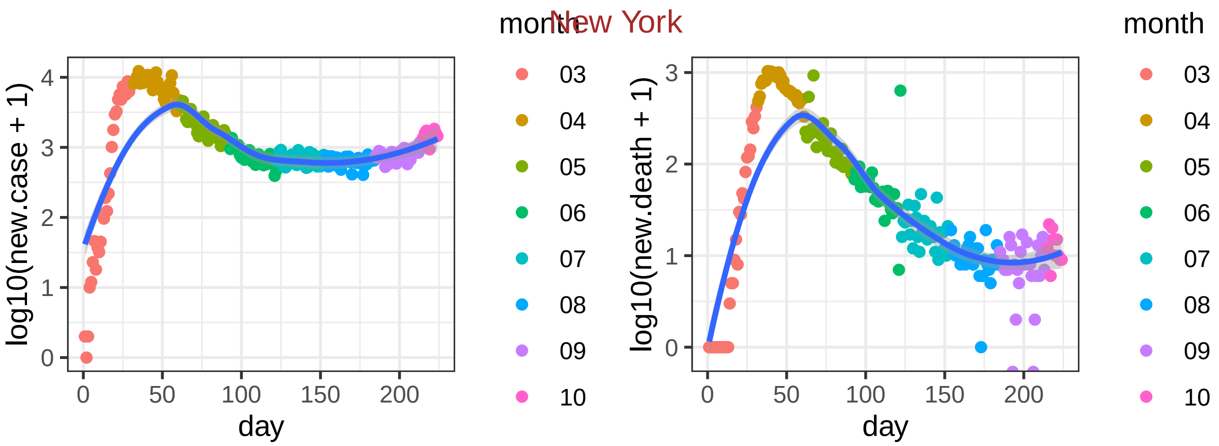
```
## [1] "2020-10-10"
```

state level data

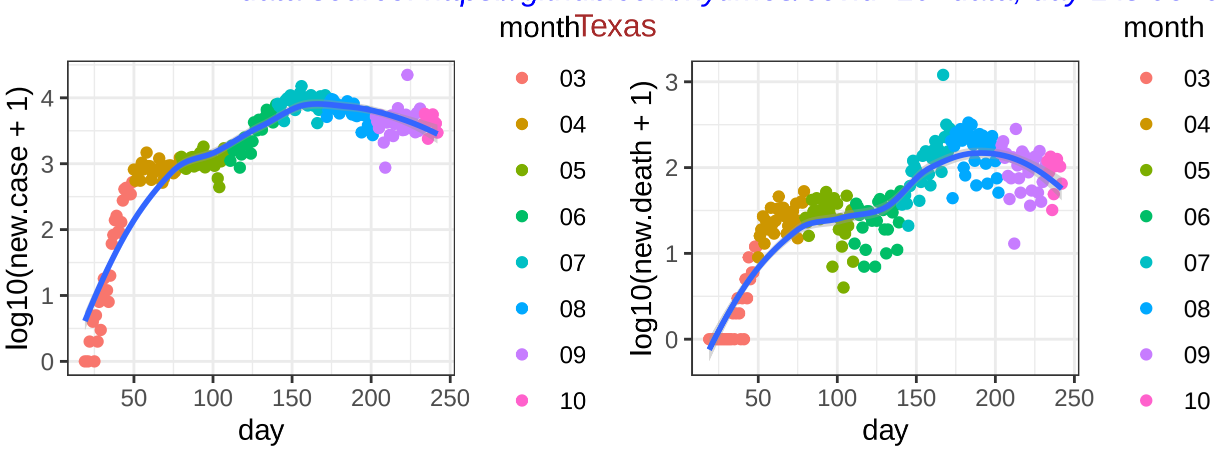
First check the 30 states with the largest number of deaths.

```
##             date          state fips  cases deaths
## 12203 2020-10-10       New York   36 477870  32875
## 12216 2020-10-10          Texas   48 827665  17019
## 12174 2020-10-10     California    6 854302  16568
## 12201 2020-10-10     New Jersey   34 214665  16171
## 12179 2020-10-10        Florida   12 728913  15185
## 12192 2020-10-10  Massachusetts   25 138340   9587
## 12184 2020-10-10       Illinois   17 320283   9235
## 12210 2020-10-10   Pennsylvania   42 175804   8421
## 12180 2020-10-10        Georgia   13 314743   7223
## 12193 2020-10-10       Michigan   26 149406   7221
## 12172 2020-10-10        Arizona    4 224985   5759
## 12189 2020-10-10      Louisiana   22 173406   5635
## 12207 2020-10-10           Ohio   39 167458   4997
## 12176 2020-10-10    Connecticut    9  60038   4530
## 12191 2020-10-10       Maryland   24 131169   3995
## 12204 2020-10-10 North Carolina   37 229959   3786
## 12185 2020-10-10        Indiana   18 135789   3782
## 12213 2020-10-10 South Carolina   45 156621   3551
## 12220 2020-10-10       Virginia   51 157905   3354
## 12195 2020-10-10    Mississippi   28 104638   3096
## 12215 2020-10-10      Tennessee   47 209593   2729
## 12170 2020-10-10        Alabama    1 164526   2664
## 12196 2020-10-10       Missouri   29 148199   2483
## 12221 2020-10-10     Washington   53  97223   2289
## 12194 2020-10-10      Minnesota   27 110881   2184
## 12175 2020-10-10       Colorado    8  78047   2121
## 12199 2020-10-10         Nevada   32  85463   1659
## 12173 2020-10-10       Arkansas    5  92220   1552
## 12223 2020-10-10      Wisconsin   55 155752   1470
## 12186 2020-10-10           Iowa   19  99042   1459
```

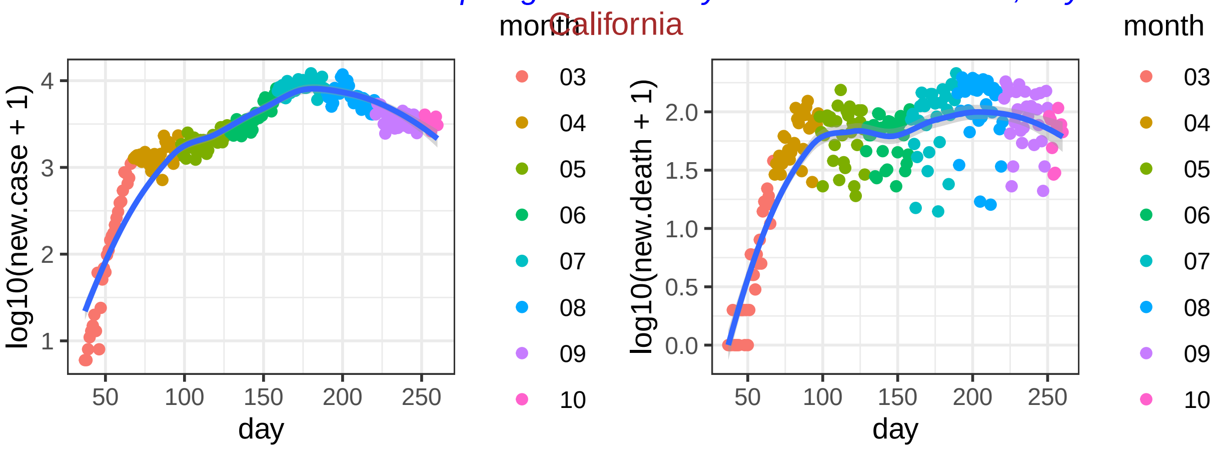
For these 30 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the similarities and differences across states.



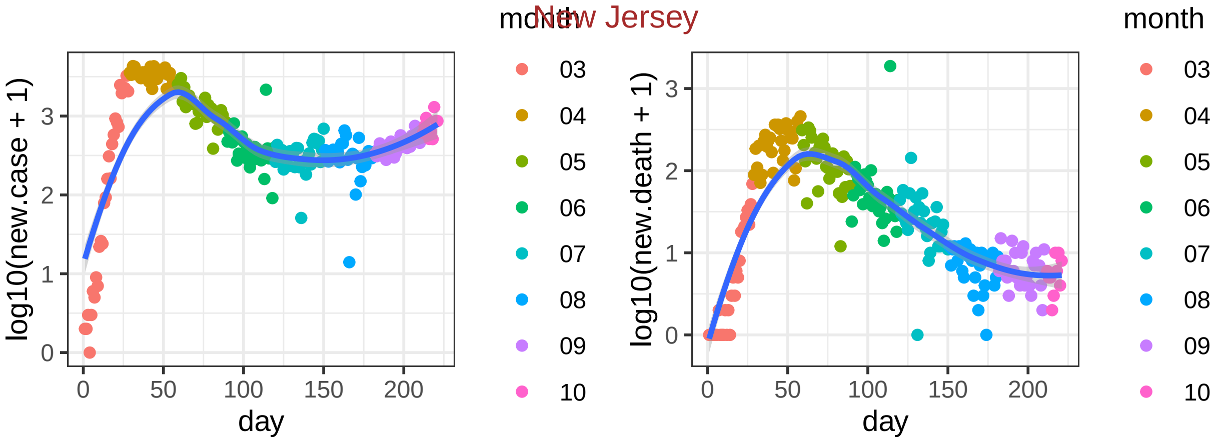
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



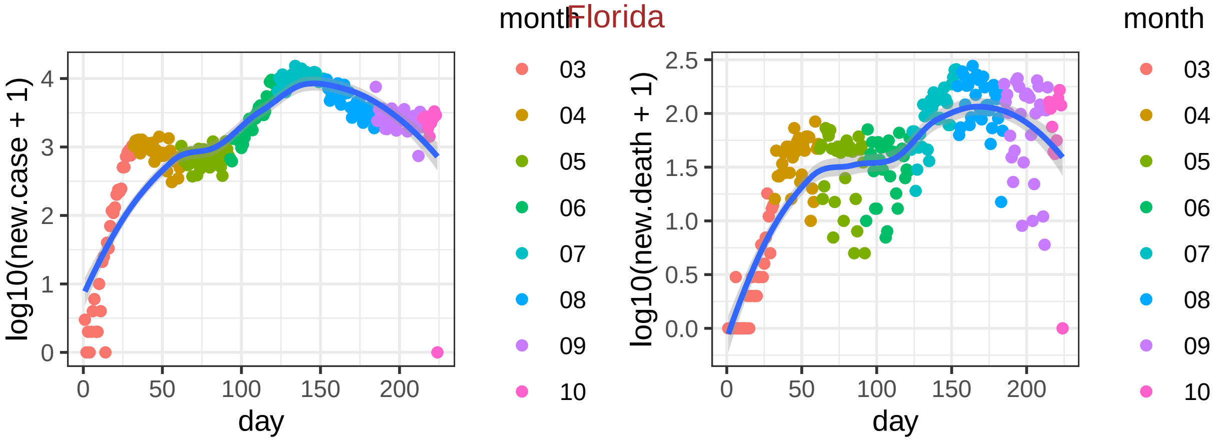
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



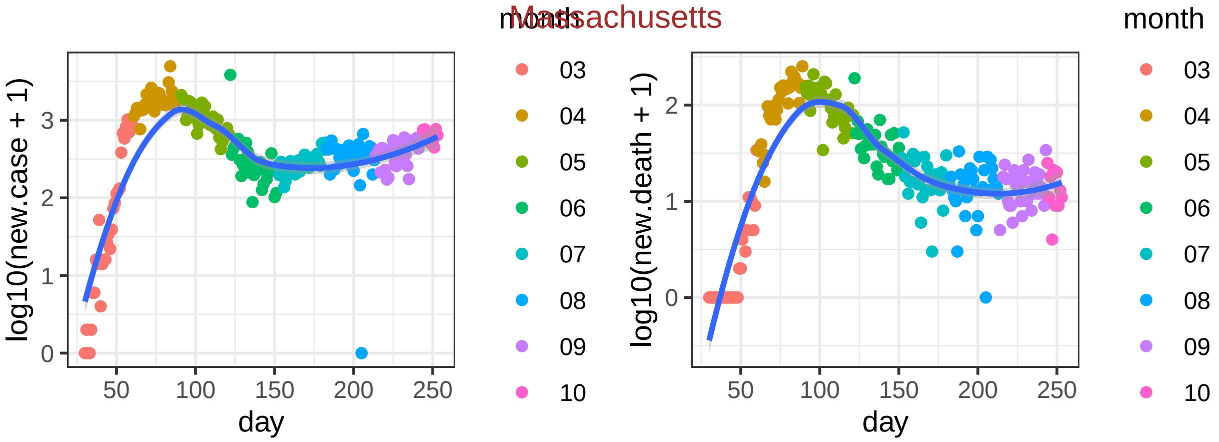
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



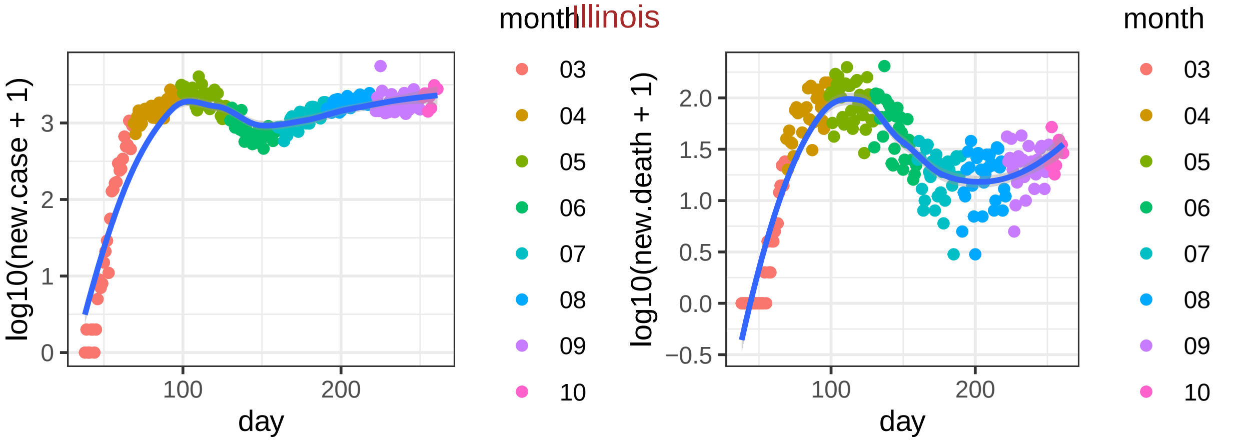
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04



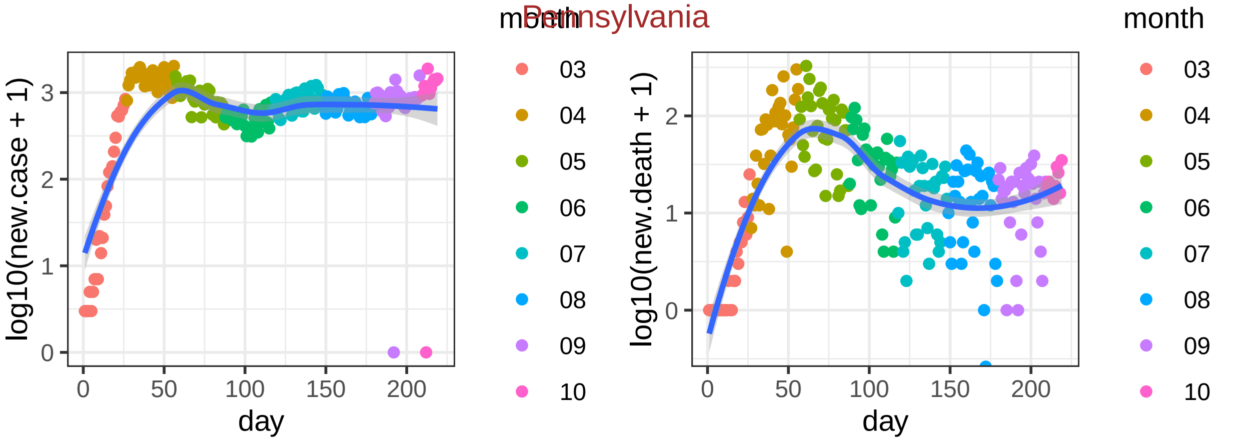
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



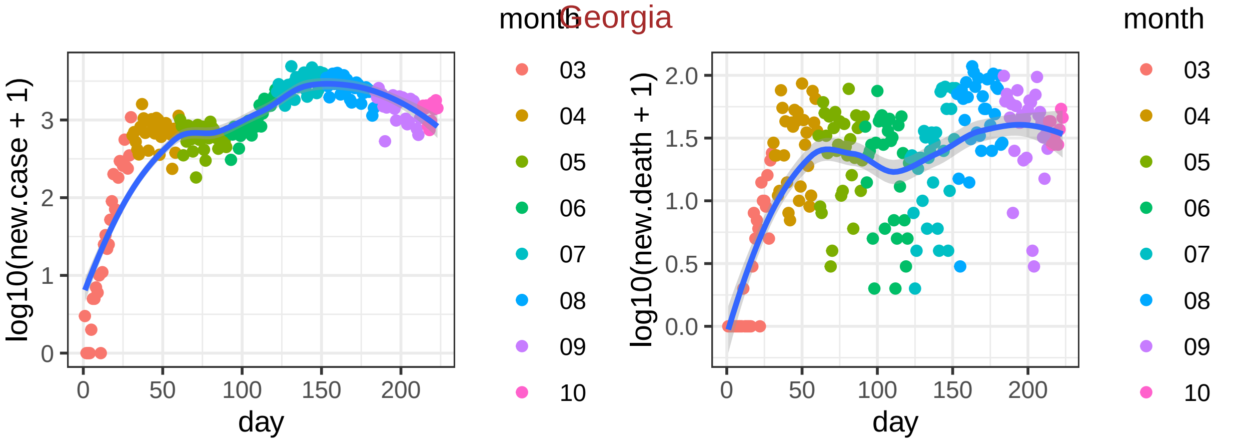
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



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

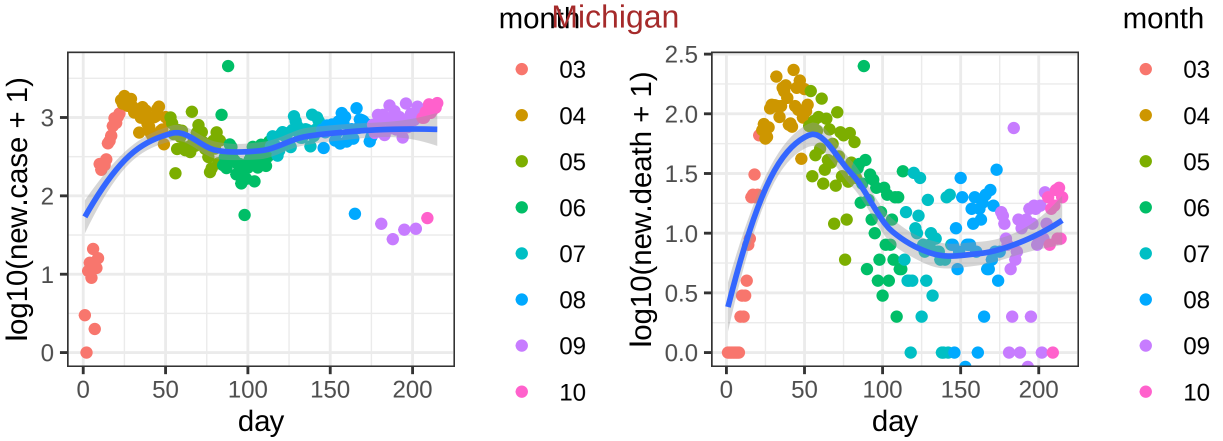


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



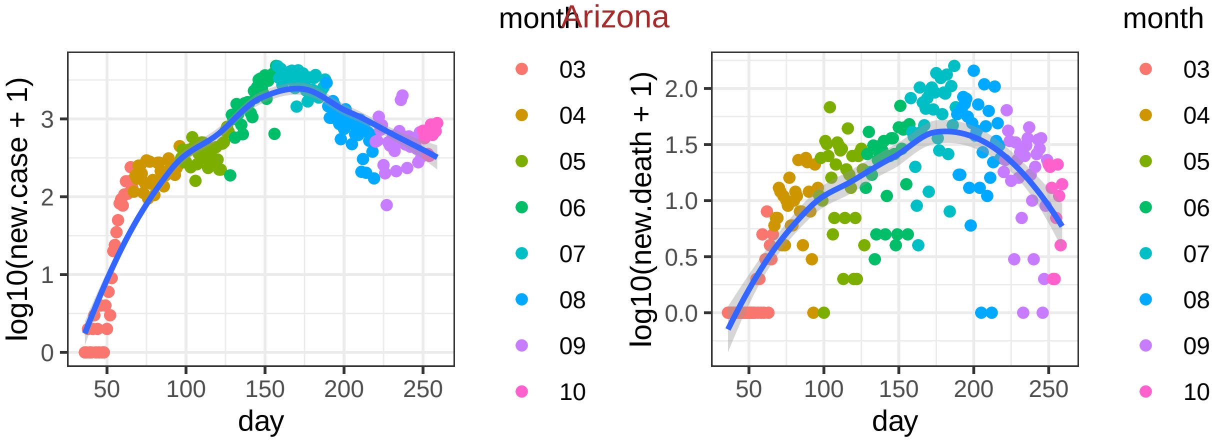
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Michigan



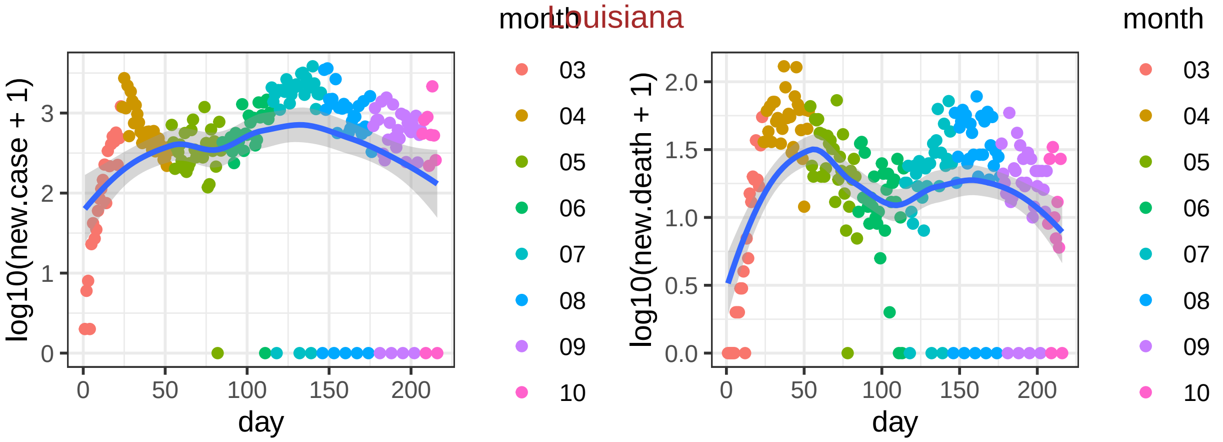
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Arizona

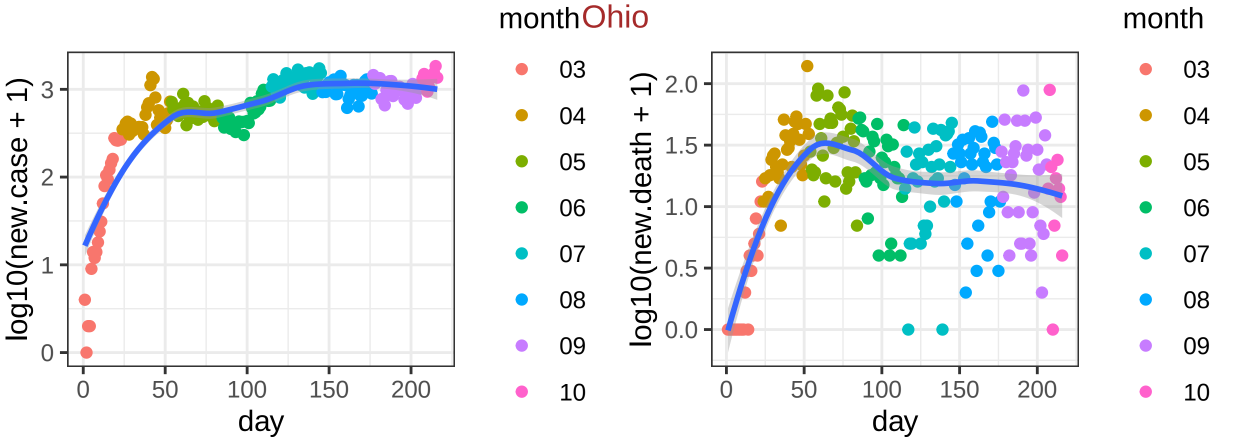


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

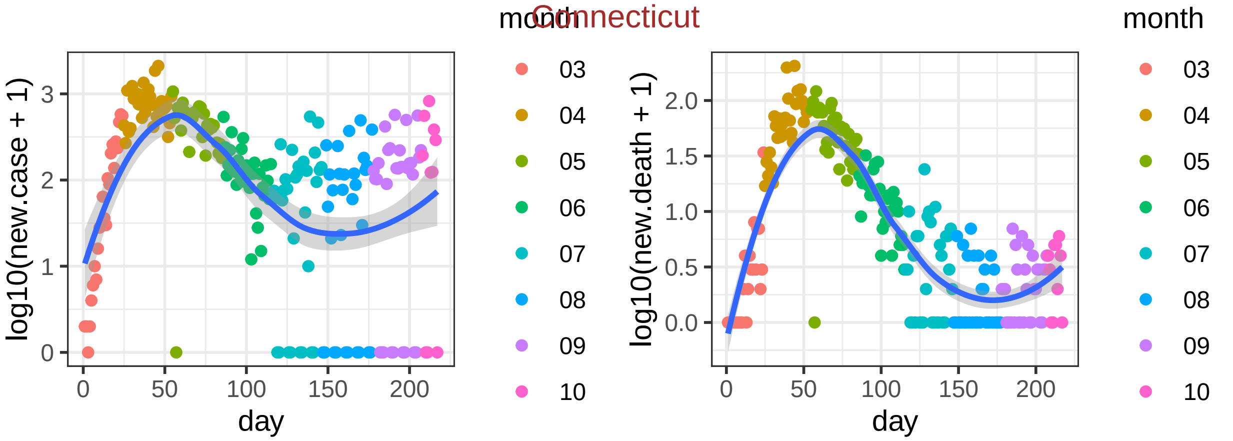
Louisiana



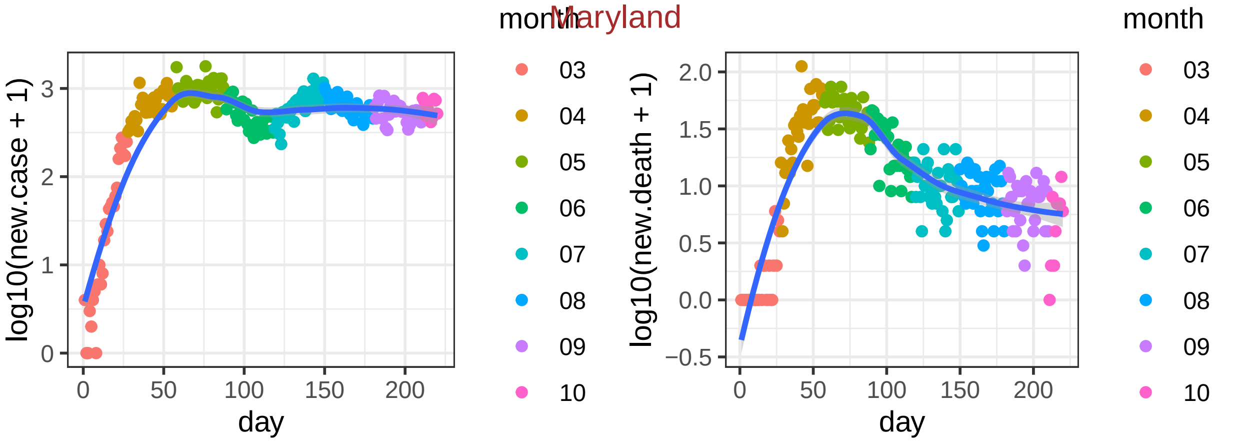
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09


Ohio

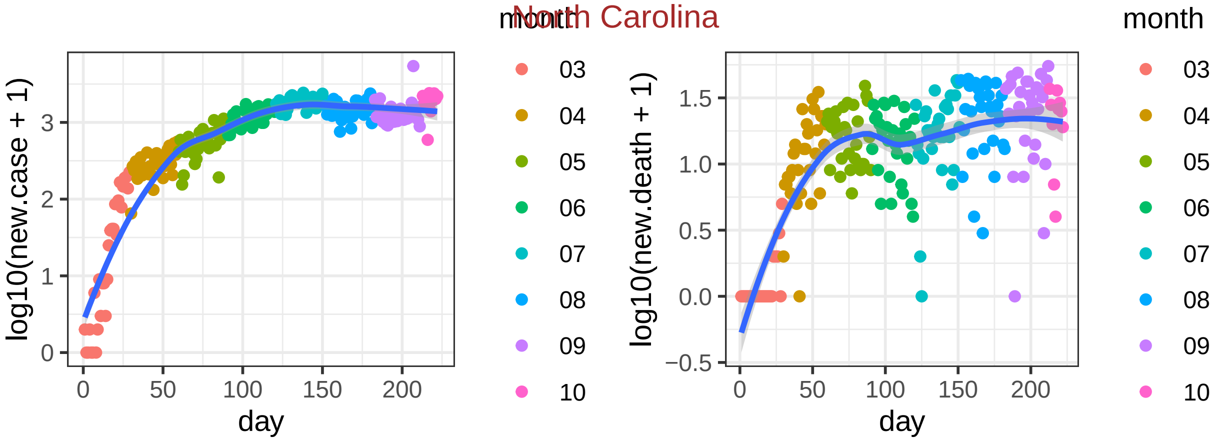
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09


Connecticut

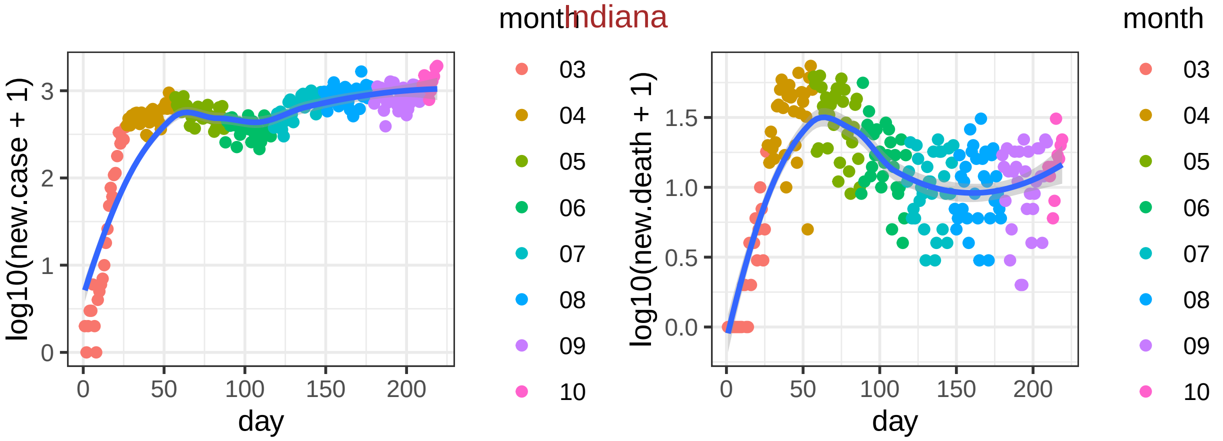
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08


Maryland

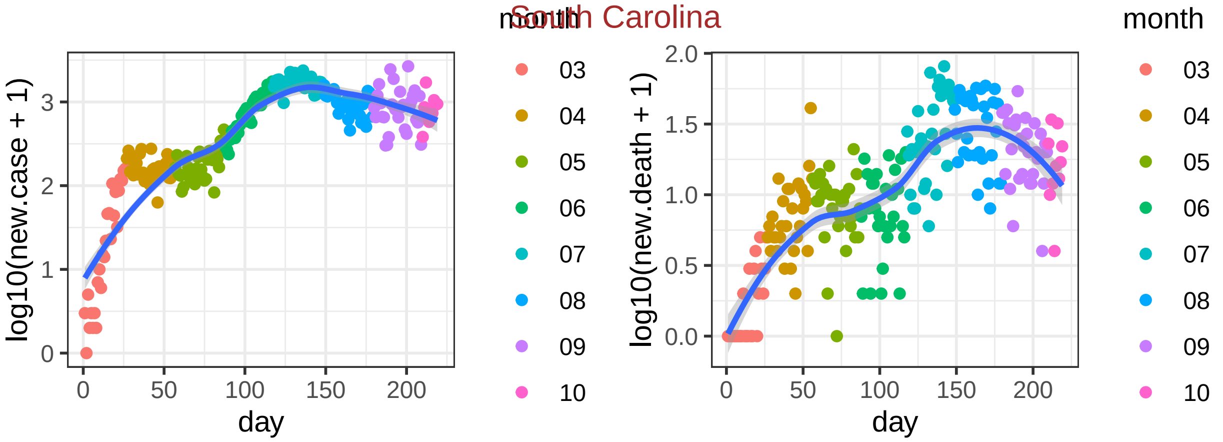
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05


North Carolina

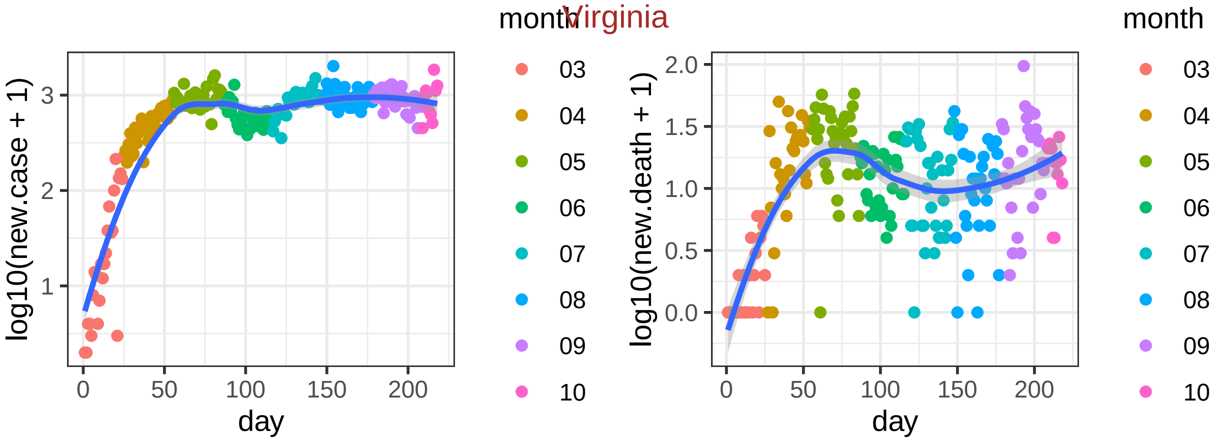
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03


Indiana

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

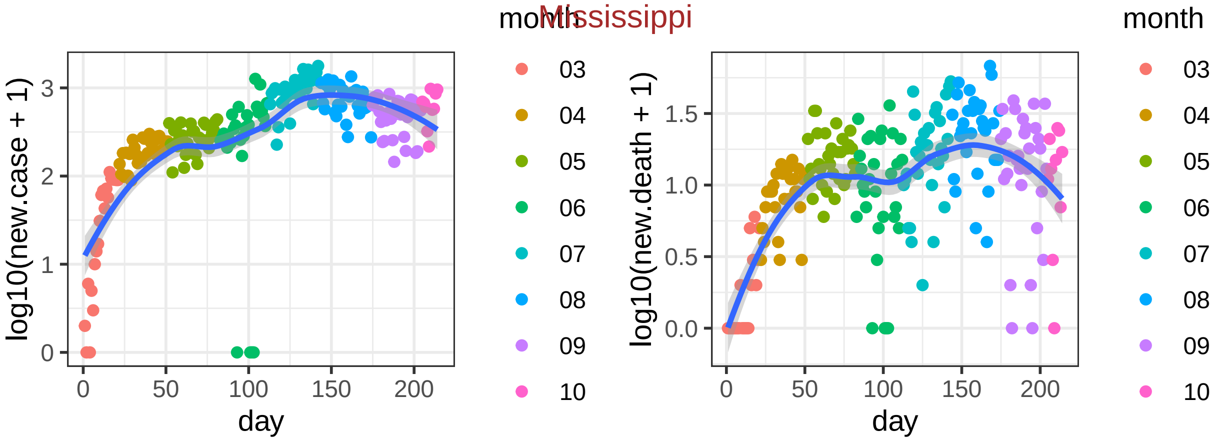

South Carolina

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



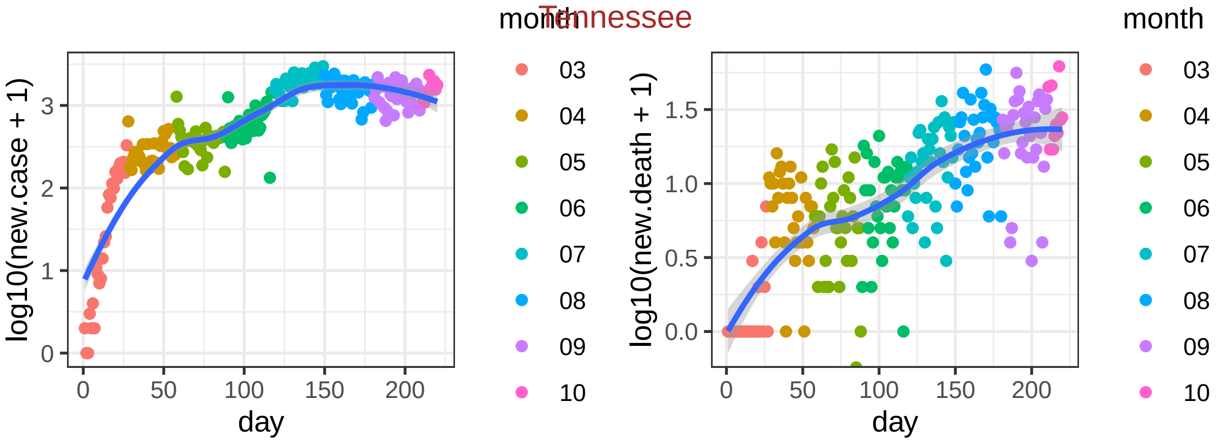
Virginia

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



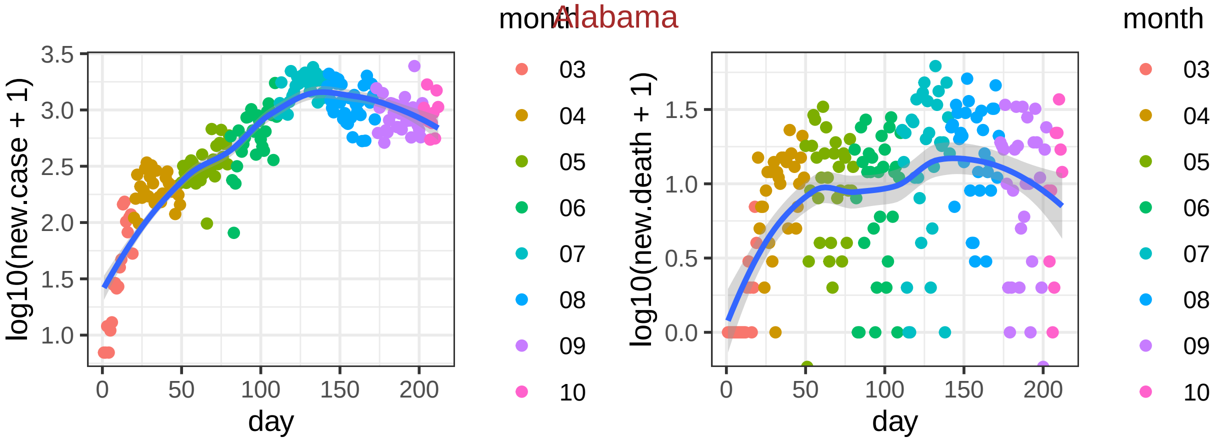
Mississippi

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

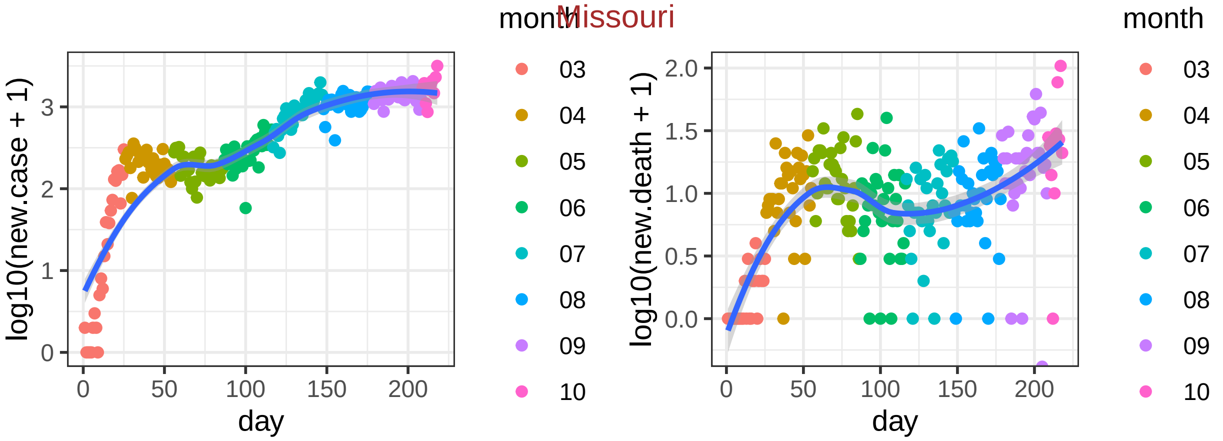


Tennessee

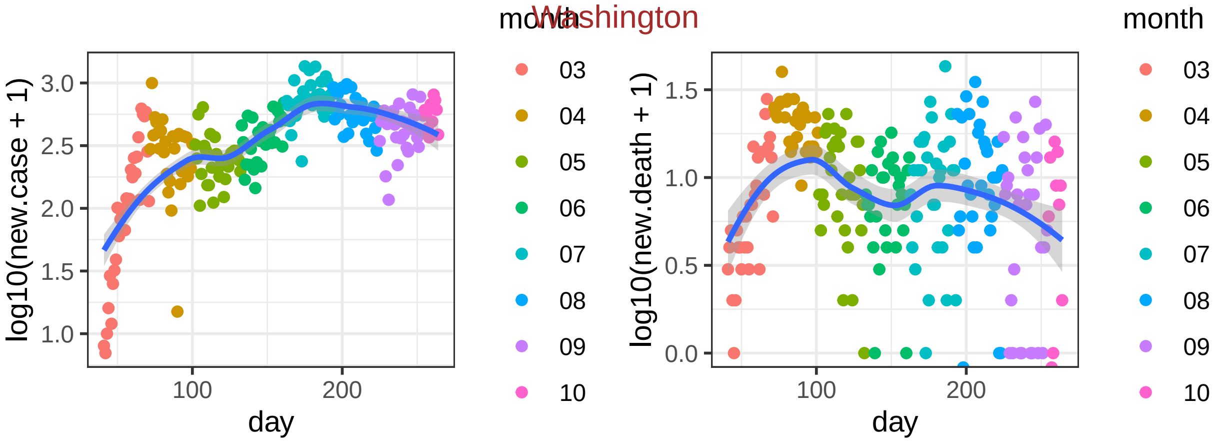
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



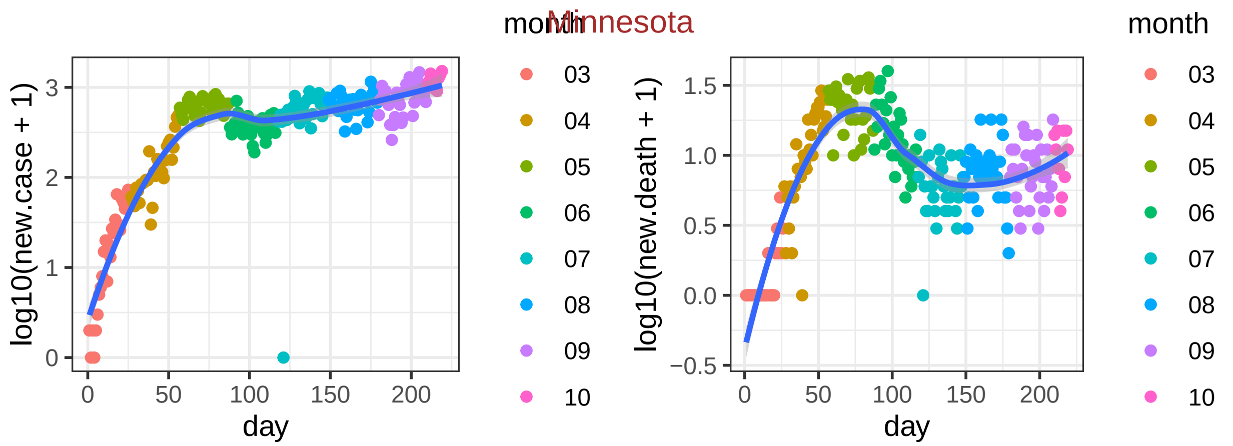
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13



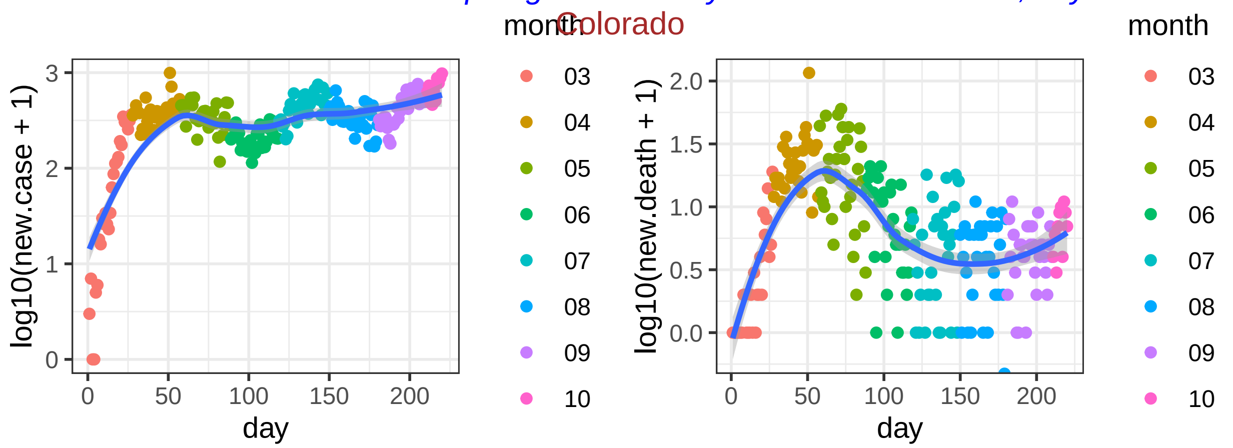
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07



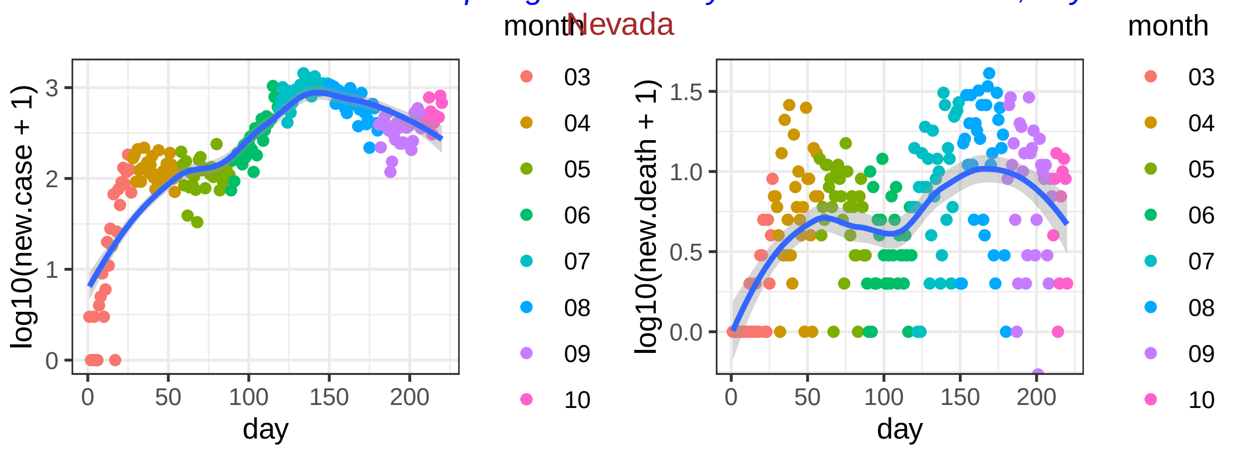
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



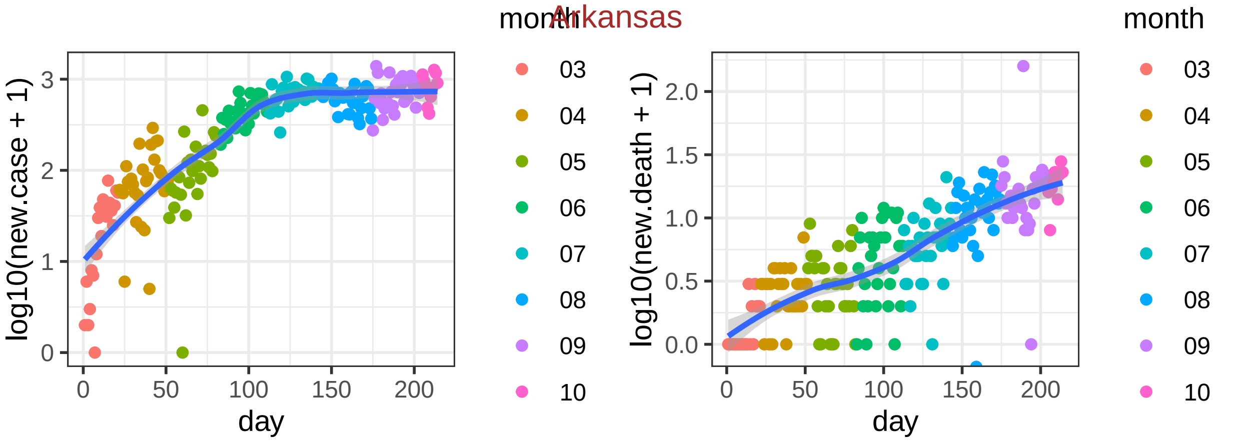
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



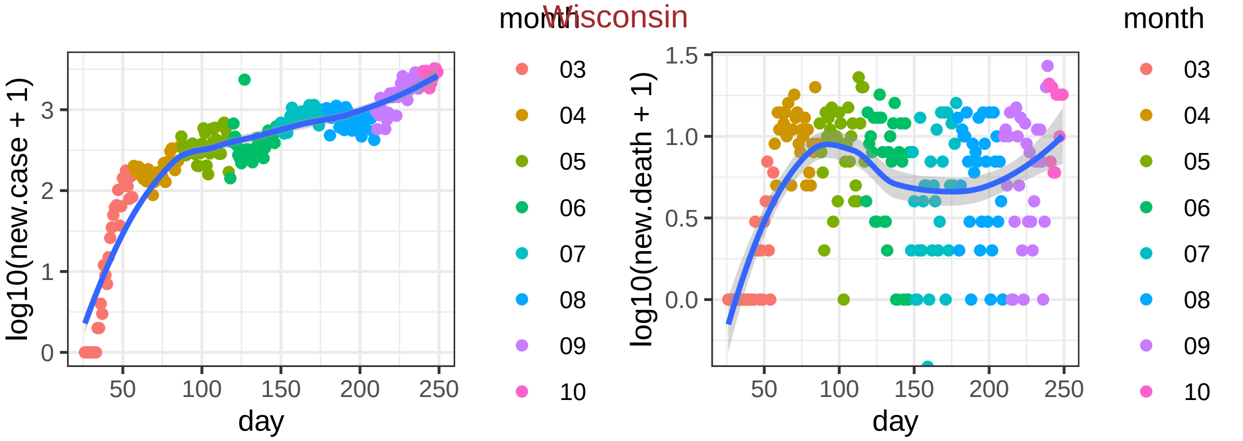
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



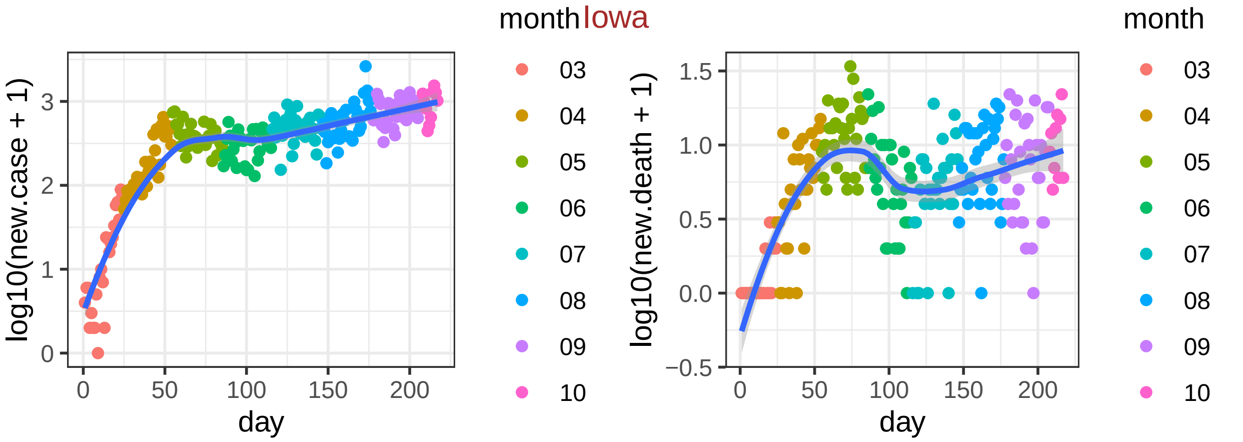
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



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

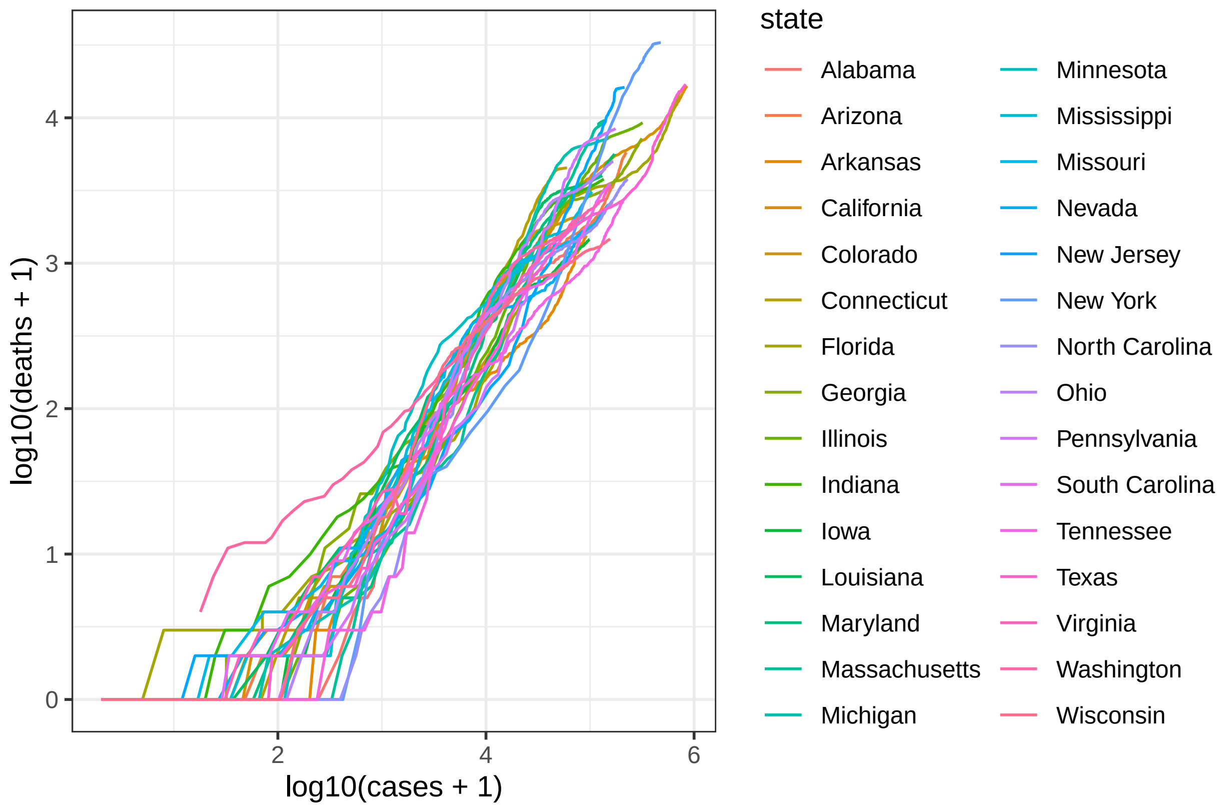


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



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

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



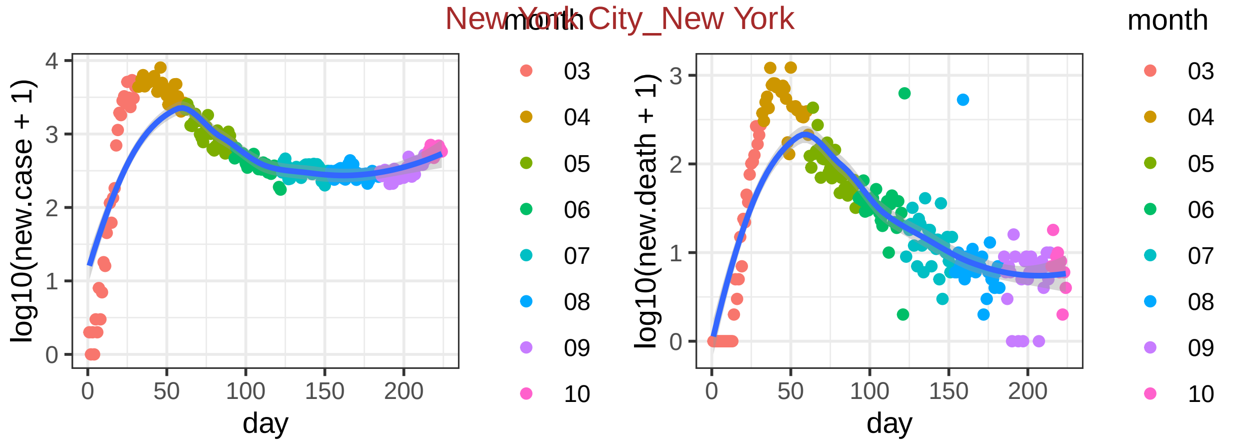
data source: https://github.com/nytimes/covid−19−data

county level data

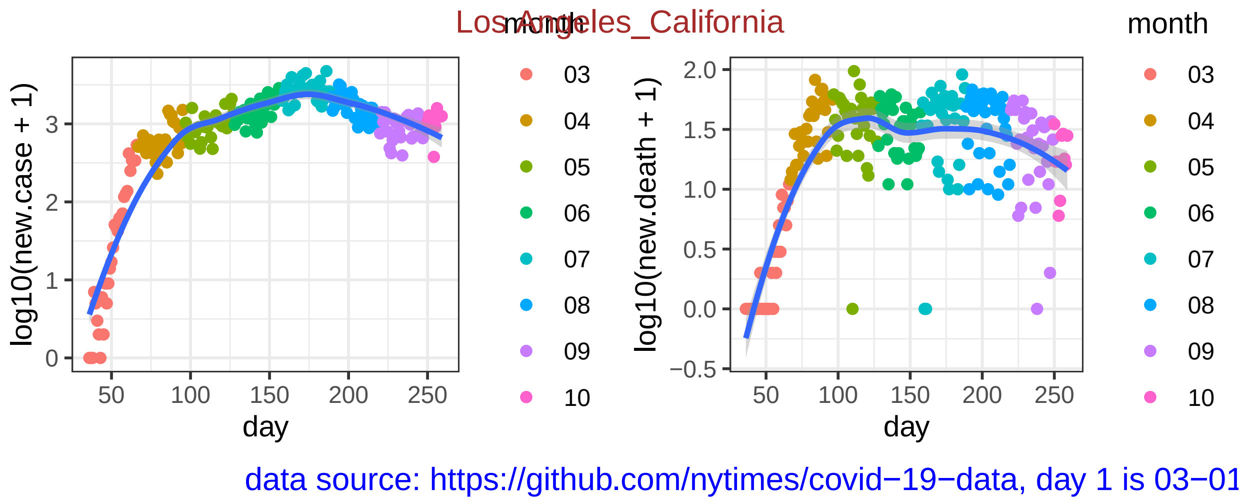
First check the 50 counties with the largest number of deaths.

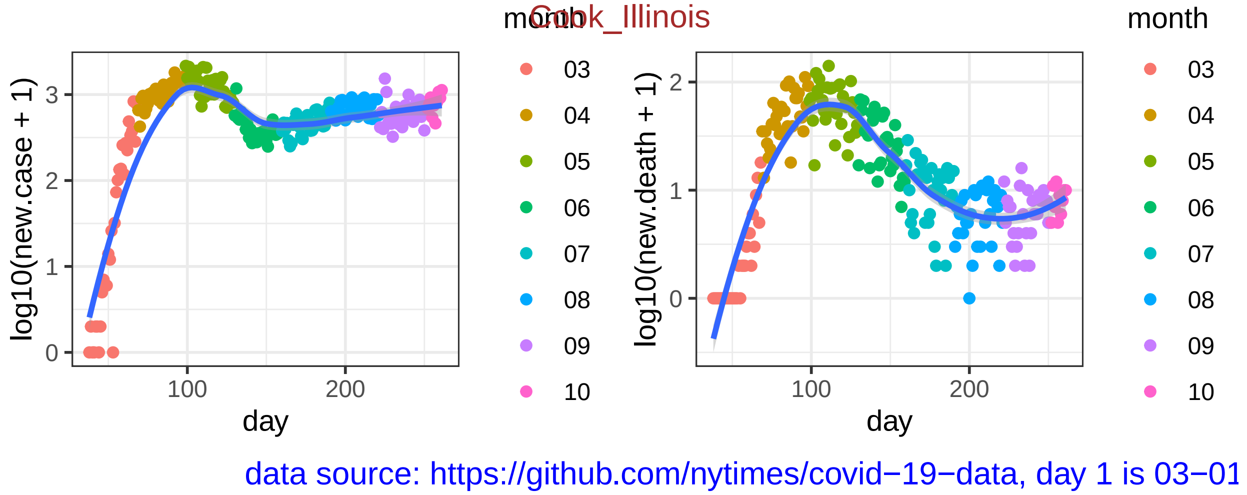
```
##              date         county         state  fips  cases deaths
## 617351 2020-10-10  New York City      New York    NA 254599  23882
## 615683 2020-10-10    Los Angeles    California  6037 281165   6768
## 616093 2020-10-10           Cook      Illinois 17031 153514   5298
## 615581 2020-10-10       Maricopa       Arizona  4013 145847   3466
## 615843 2020-10-10     Miami-Dade       Florida 12086 174111   3409
## 616802 2020-10-10          Wayne      Michigan 26163  36900   3008
## 618198 2020-10-10         Harris         Texas 48201 150626   2675
## 617350 2020-10-10         Nassau      New York 36059  47841   2204
## 616713 2020-10-10      Middlesex Massachusetts 25017  28917   2186
## 617274 2020-10-10          Essex    New Jersey 34013  21918   2128
## 617269 2020-10-10         Bergen    New Jersey 34003  23349   2051
## 617370 2020-10-10        Suffolk      New York 36103  47357   2016
## 617788 2020-10-10   Philadelphia  Pennsylvania 42101  38767   1854
## 618205 2020-10-10        Hidalgo         Texas 48215  33352   1841
## 617276 2020-10-10         Hudson    New Jersey 34017  21300   1519
## 617378 2020-10-10    Westchester      New York 36119  38846   1461
## 615806 2020-10-10        Broward       Florida 12011  78795   1455
## 615788 2020-10-10       Hartford   Connecticut  9003  15520   1442
## 617243 2020-10-10          Clark        Nevada 32003  71325   1438
## 617279 2020-10-10      Middlesex    New Jersey 34023  20506   1433
## 615787 2020-10-10      Fairfield   Connecticut  9001  20752   1426
## 615850 2020-10-10     Palm Beach       Florida 12099  47646   1425
## 617287 2020-10-10           Union    New Jersey 34039  18330    1361
## 618112 2020-10-10           Bexar         Texas 48029  59526    1360
## 615694 2020-10-10          Orange    California  6059  57225    1340
## 616709 2020-10-10           Essex Massachusetts 25009  21078    1307
## 615697 2020-10-10       Riverside    California  6065  61824    1256
## 617283 2020-10-10         Passaic    New Jersey 34031  19702    1255
## 616782 2020-10-10         Oakland      Michigan 26125  21878    1220
## 618154 2020-10-10          Dallas         Texas 48113  88646    1167
## 616717 2020-10-10         Suffolk Massachusetts 25025  25497    1145
## 616719 2020-10-10       Worcester Massachusetts 25027  15022    1127
## 615791 2020-10-10       New Haven   Connecticut  9009  14877    1118
## 616715 2020-10-10         Norfolk Massachusetts 25021  10792    1080
## 617282 2020-10-10           Ocean    New Jersey 34029  15119    1052
## 616769 2020-10-10          Macomb      Michigan 26099  16156    1046
## 618128 2020-10-10         Cameron         Texas 48061  23312    1044
## 615700 2020-10-10  San Bernardino    California  6071  57834     986
## 616830 2020-10-10        Hennepin     Minnesota 27053  29929     951
## 617887 2020-10-10      Providence  Rhode Island 44007  19636     902
## 617783 2020-10-10      Montgomery  Pennsylvania 42091  12774     887
## 617280 2020-10-10        Monmouth    New Jersey 34025  12851     867
## 616695 2020-10-10      Montgomery      Maryland 24031  23545     855
## 617076 2020-10-10       St. Louis      Missouri 29189  26134     839
## 616696 2020-10-10 Prince George's      Maryland 24033  30874     833
## 617281 2020-10-10          Morris    New Jersey 34027   8269     832
## 615701 2020-10-10       San Diego    California  6073  50206     825
## 616229 2020-10-10          Marion       Indiana 18097  23211     824
## 618547 2020-10-10            King    Washington 53033  23898     803
## 617760 2020-10-10        Delaware  Pennsylvania 42045  12062     799
```

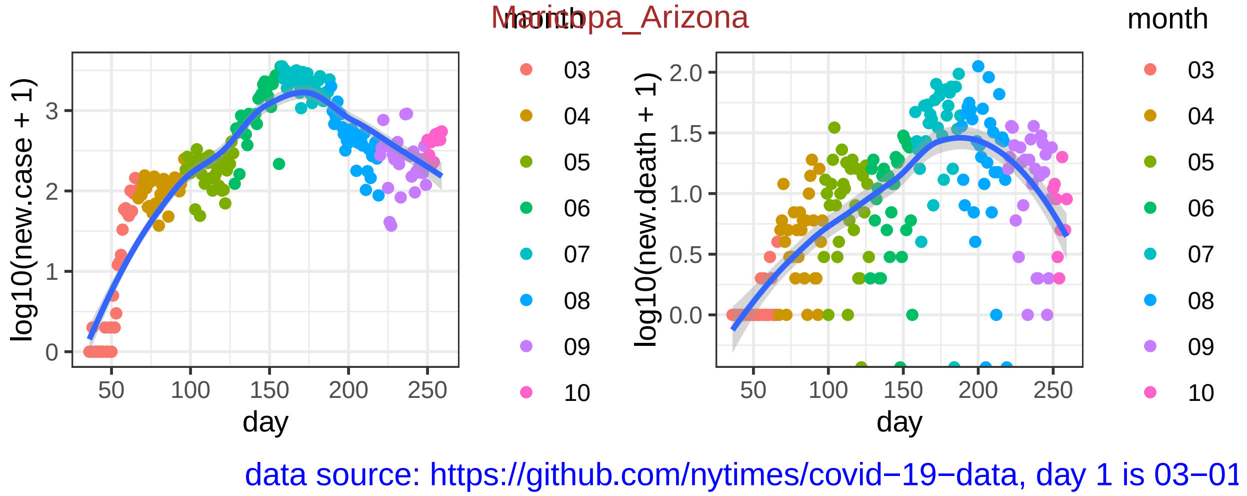
For these 50 counties, I check the number of new cases and the number of new deaths.

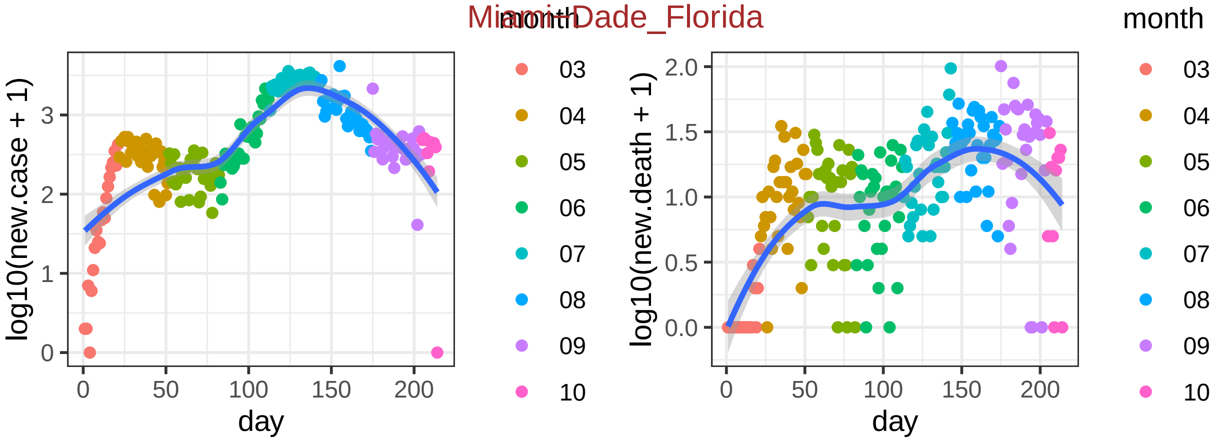


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

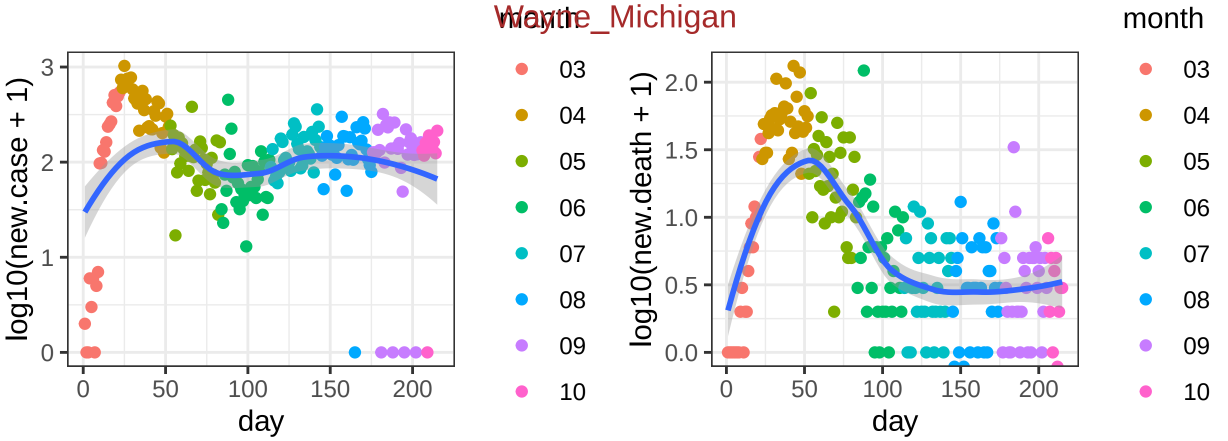

Los_Angeles_California
month
log10(new.case + 1)
0
1
2
3
50
100
150
200
250
day
03
04
05
06
07
08
09
10
month
log10(new.death + 1)
−0.5
0.0
0.5
1.0
1.5
2.0
50
100
150
200
250
day
03
04
05
06
07
08
09
10
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01


Cook_Illinois
month
log10(new.case + 1)
0
1
2
3
100
200
day
03
04
05
06
07
08
09
10
month
log10(new.death + 1)
0
1
2
100
200
day
03
04
05
06
07
08
09
10
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

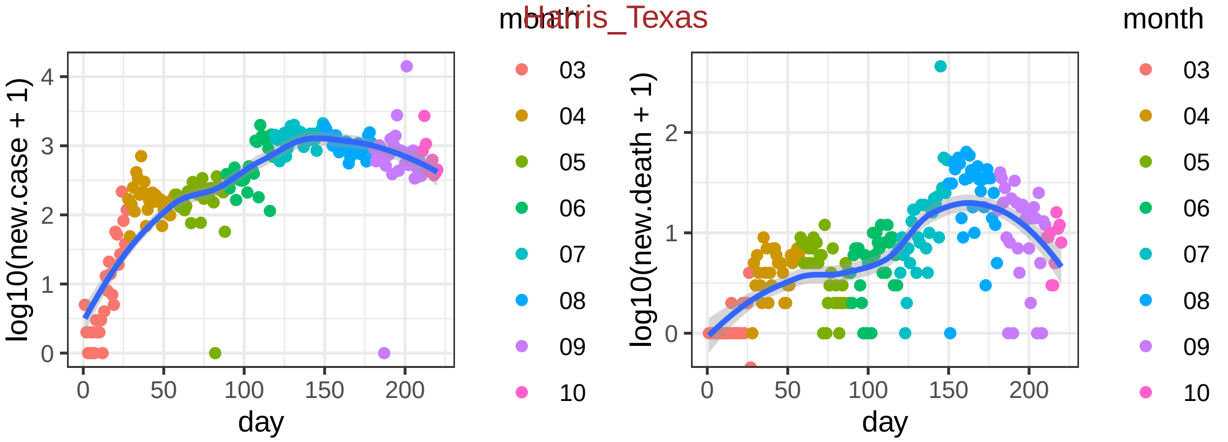

Maricopa_Arizona
month
log10(new.case + 1)
0
1
2
3
50
100
150
200
250
day
03
04
05
06
07
08
09
10
month
log10(new.death + 1)
0.0
0.5
1.0
1.5
2.0
50
100
150
200
250
day
03
04
05
06
07
08
09
10
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



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

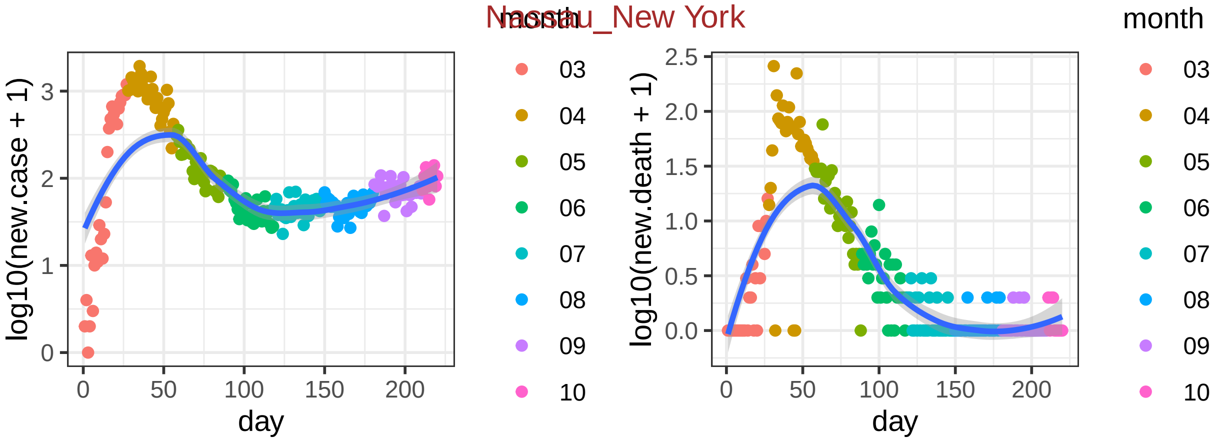


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10



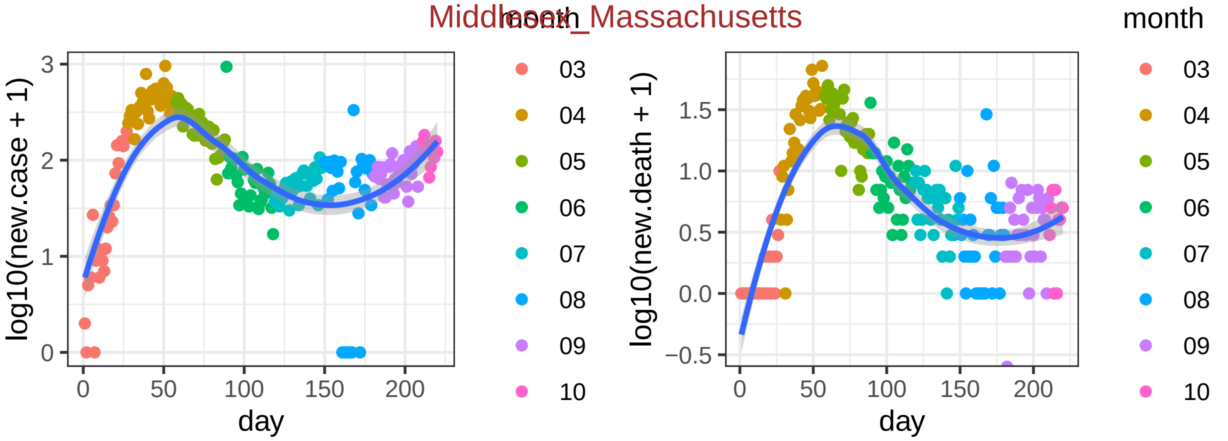
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Nassau_New York



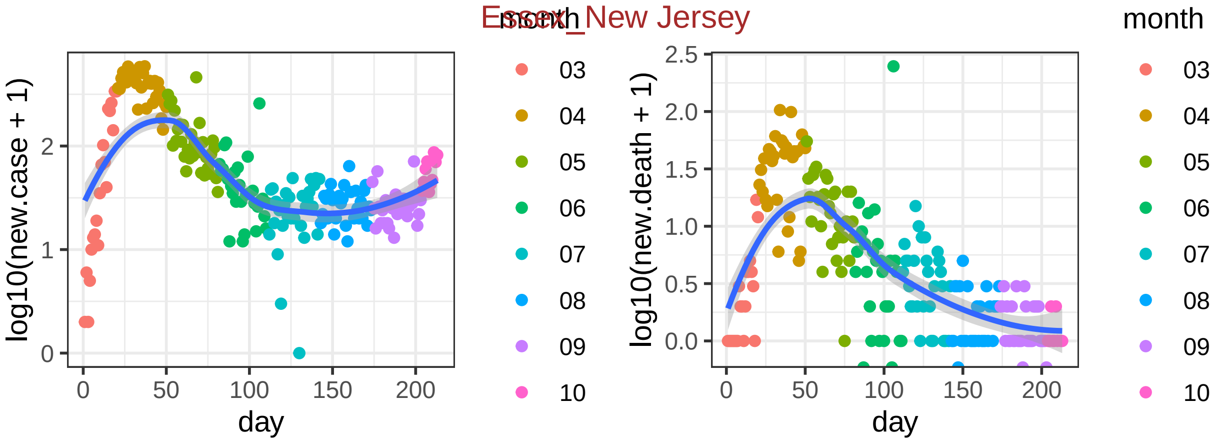
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Middlesex_Massachusetts

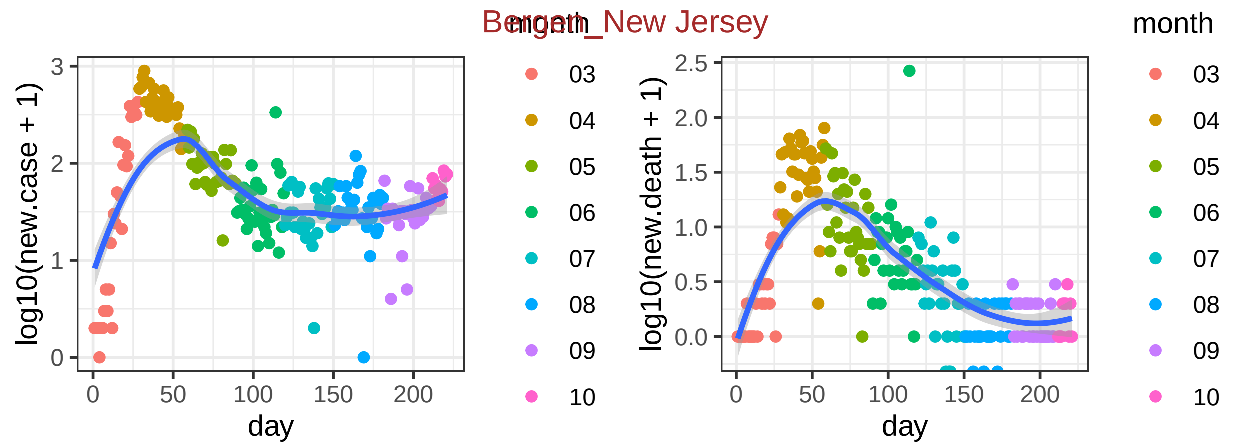


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

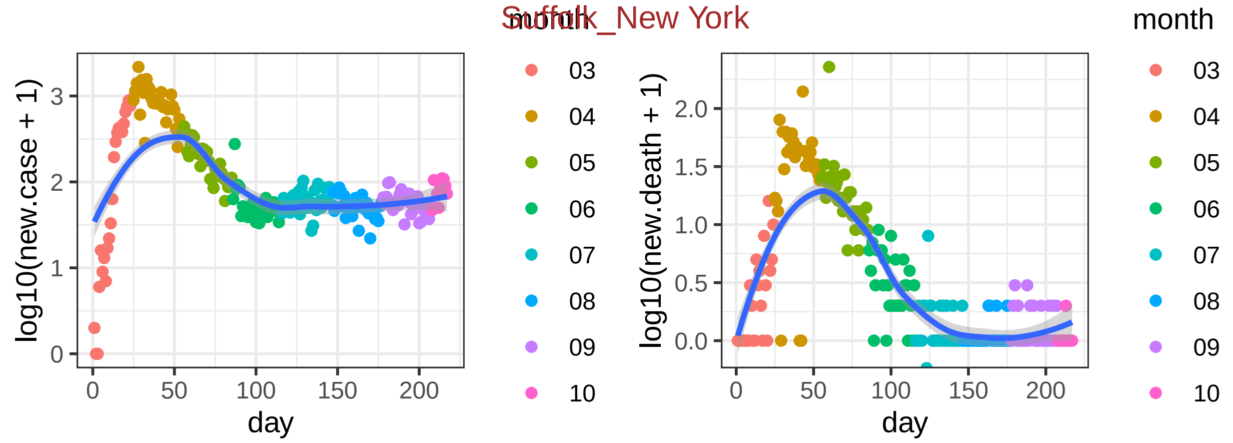
Essex_New Jersey



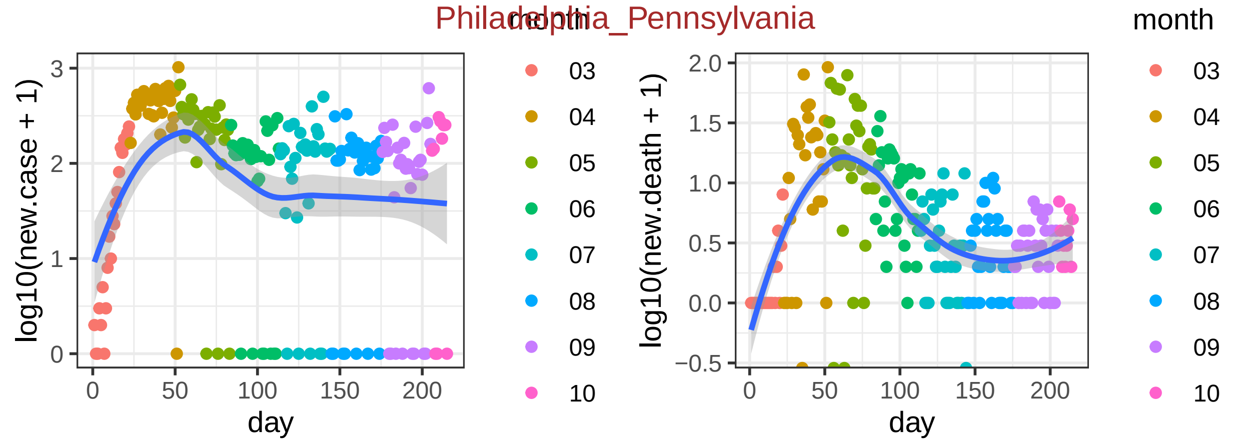
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12



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

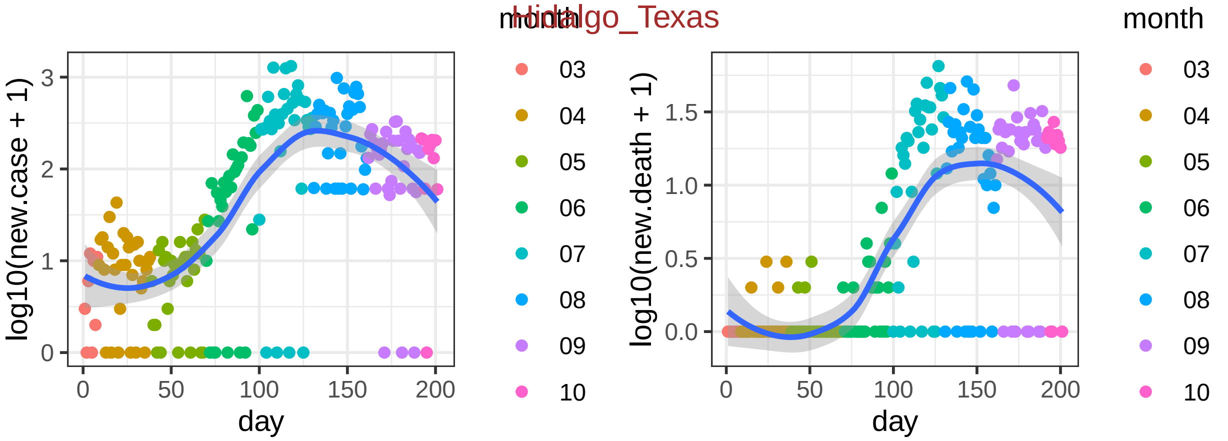


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08



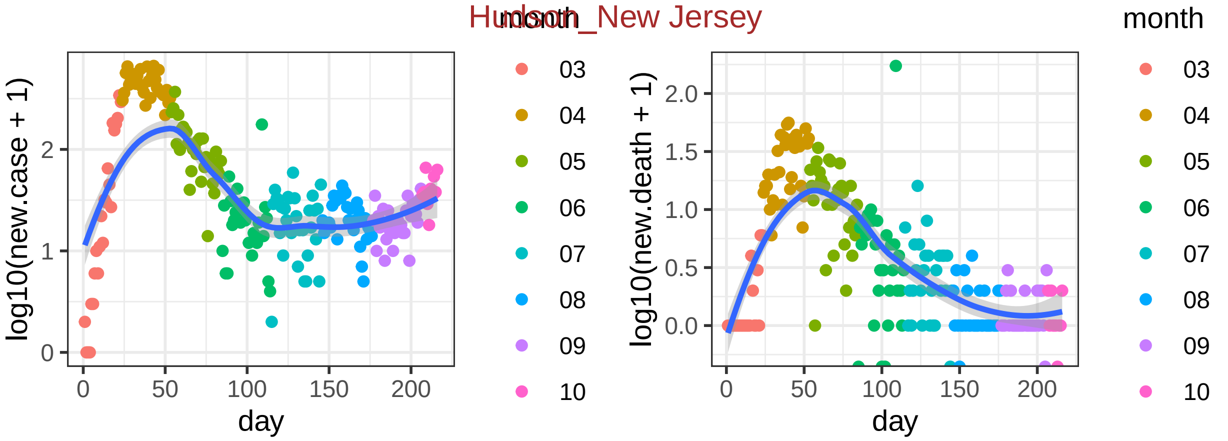
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Hidalgo_Texas



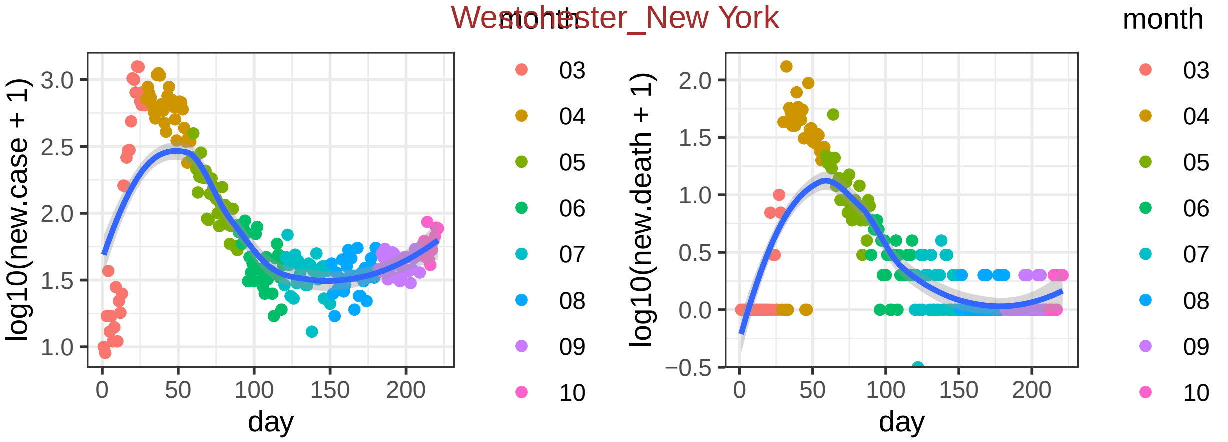
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−24

Hudson_New Jersey

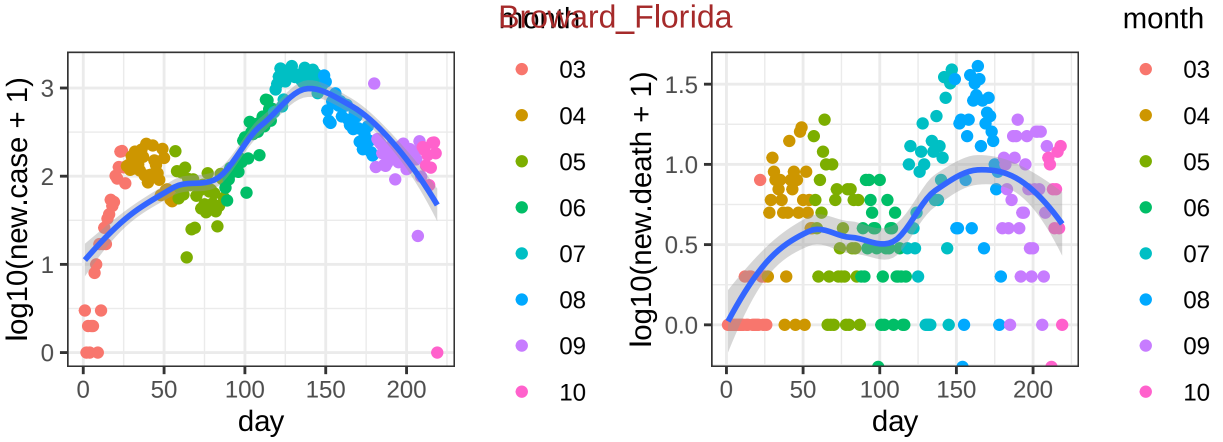


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

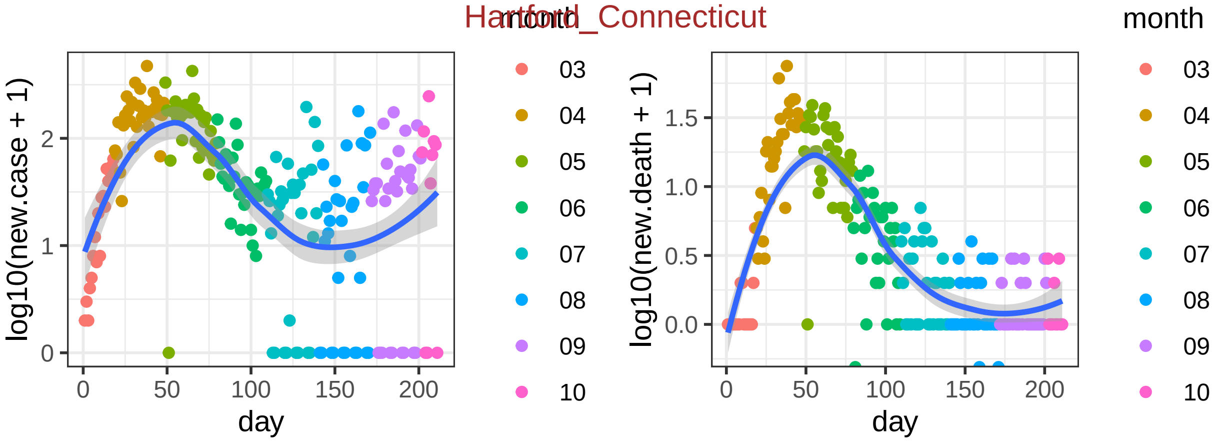
Westchester_New York



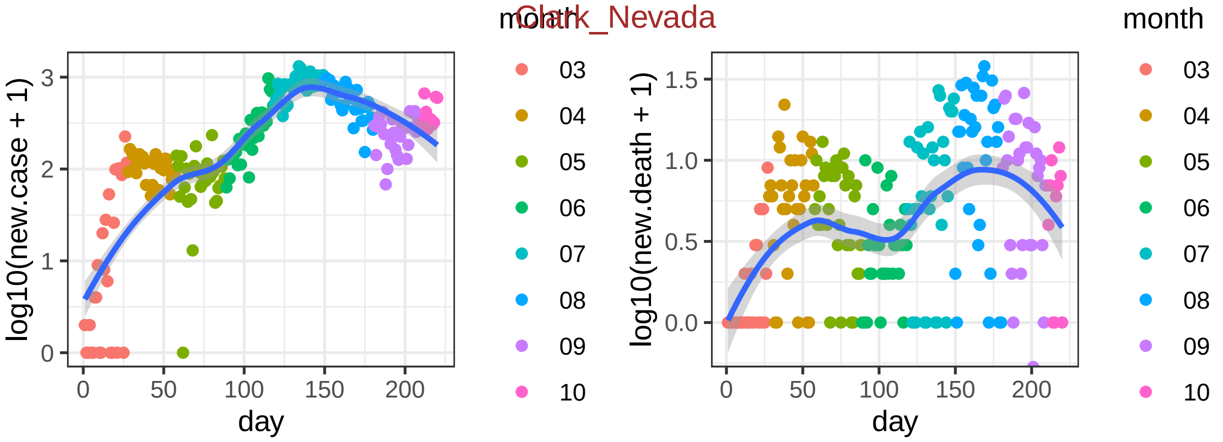
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04



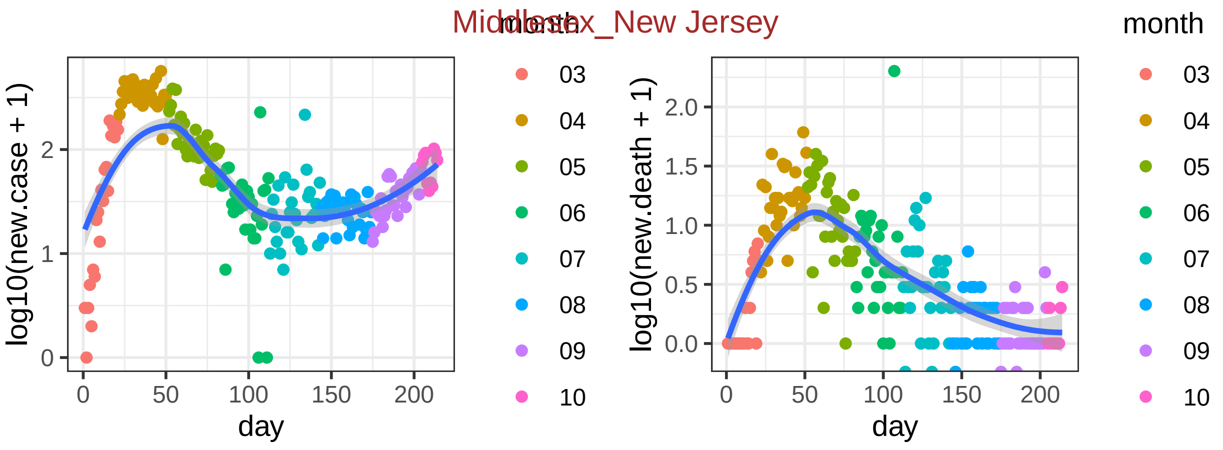
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



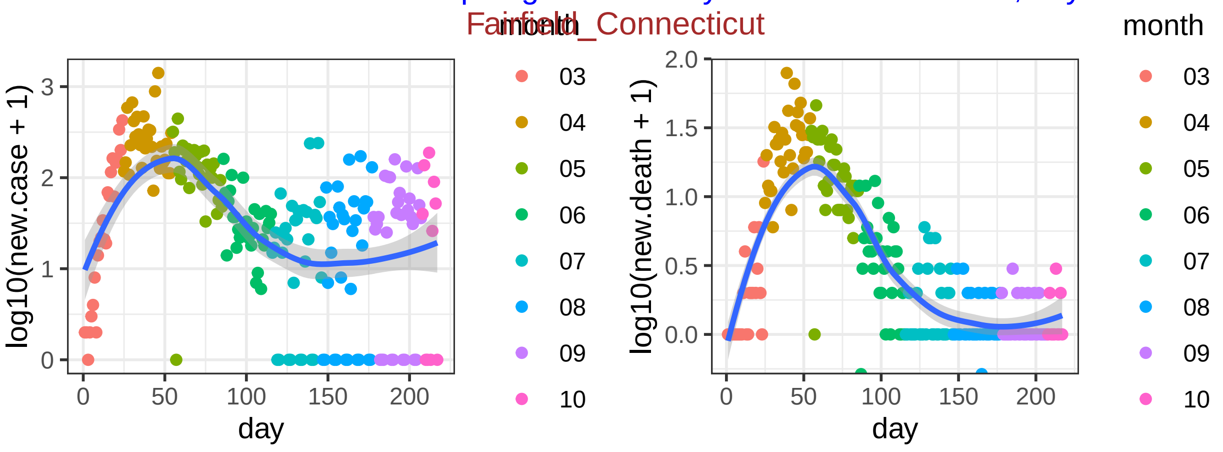
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14



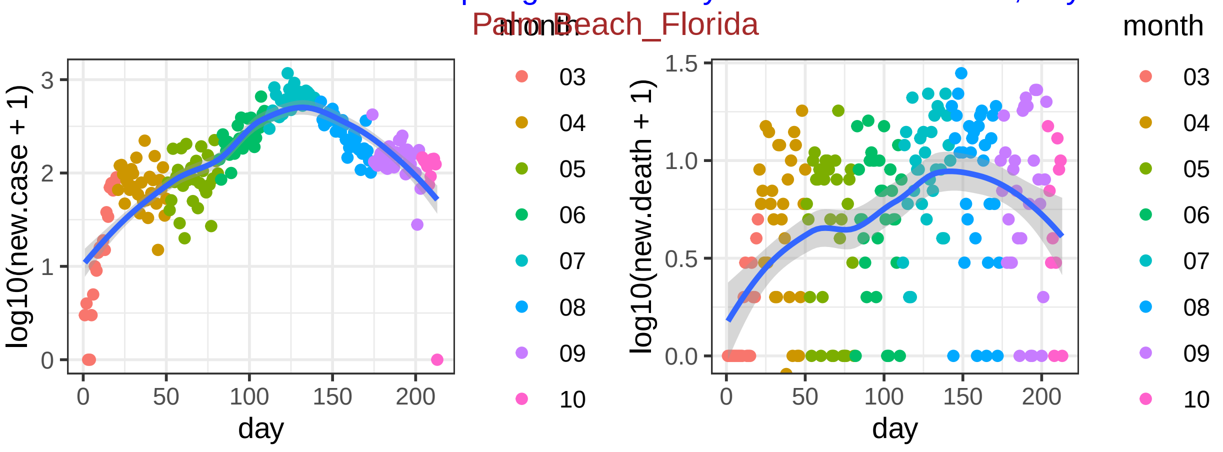
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



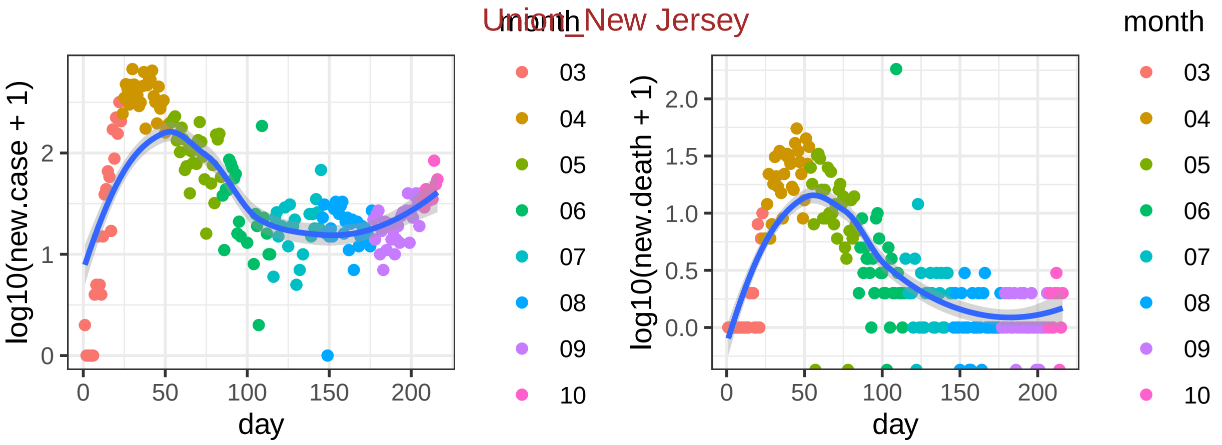
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11



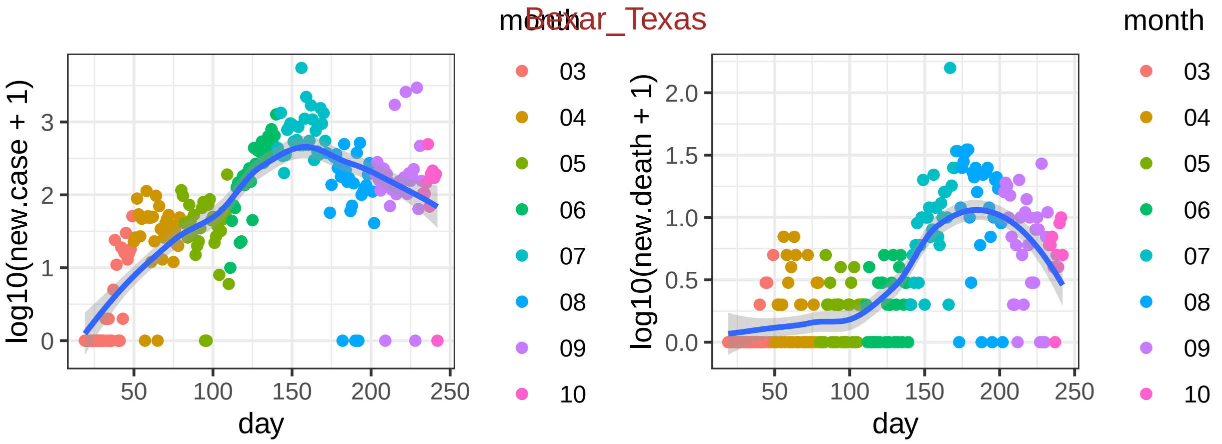
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08



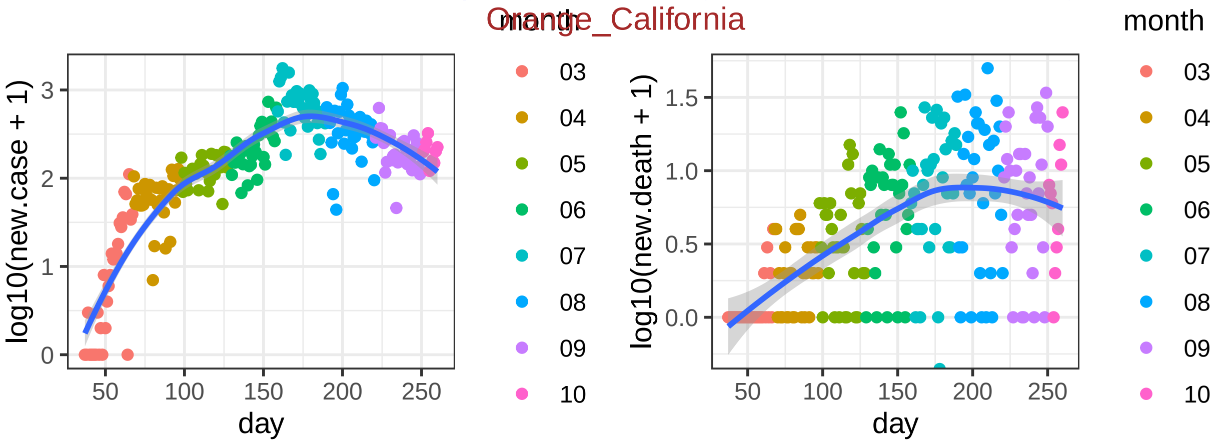
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12



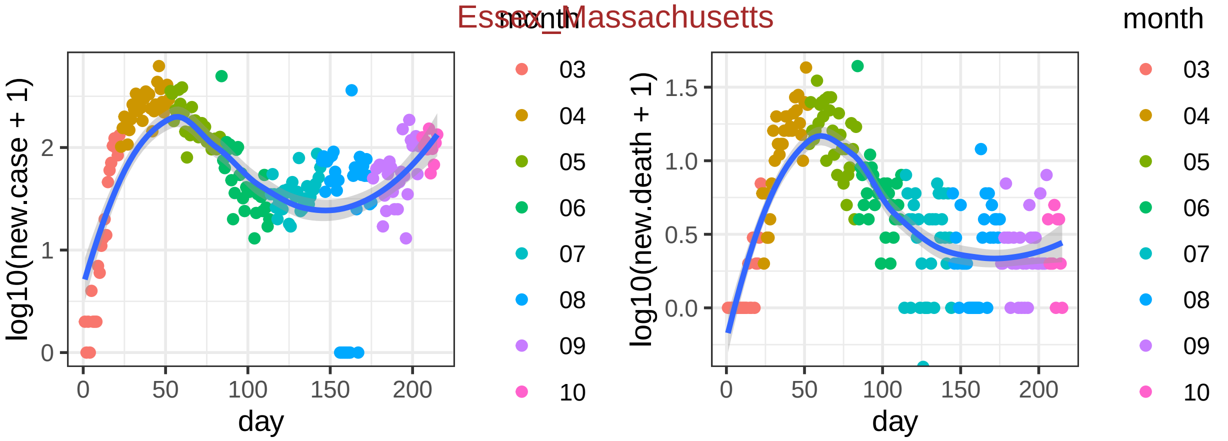
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09



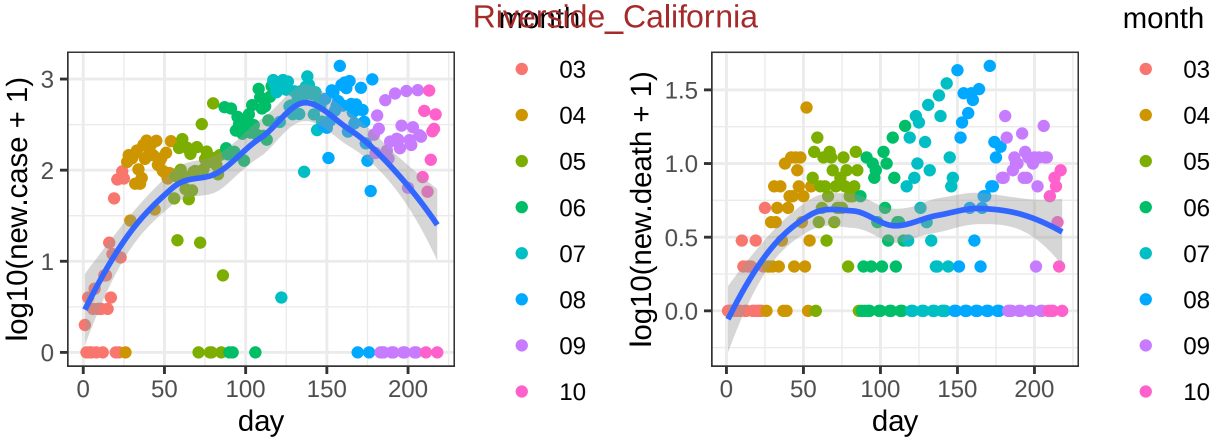
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



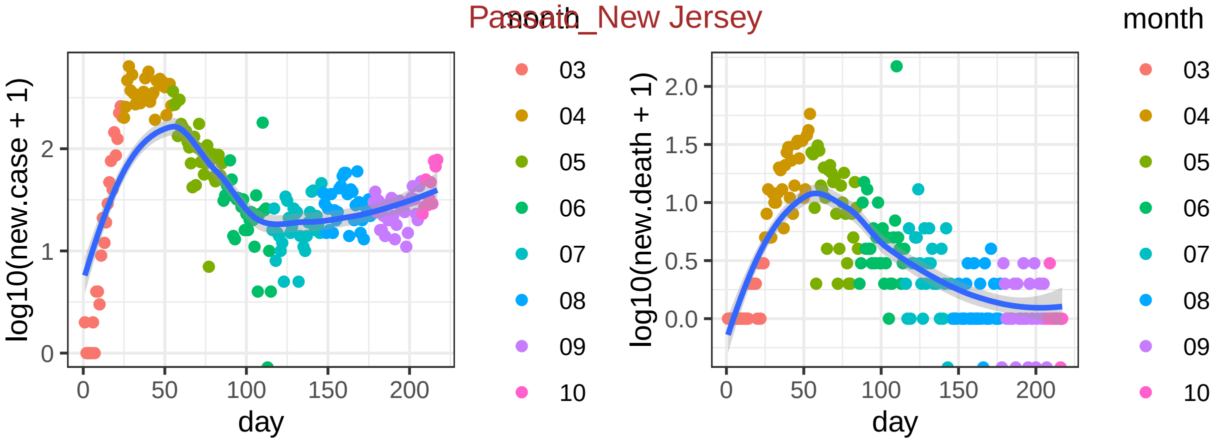
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01


Essex_Massachusetts

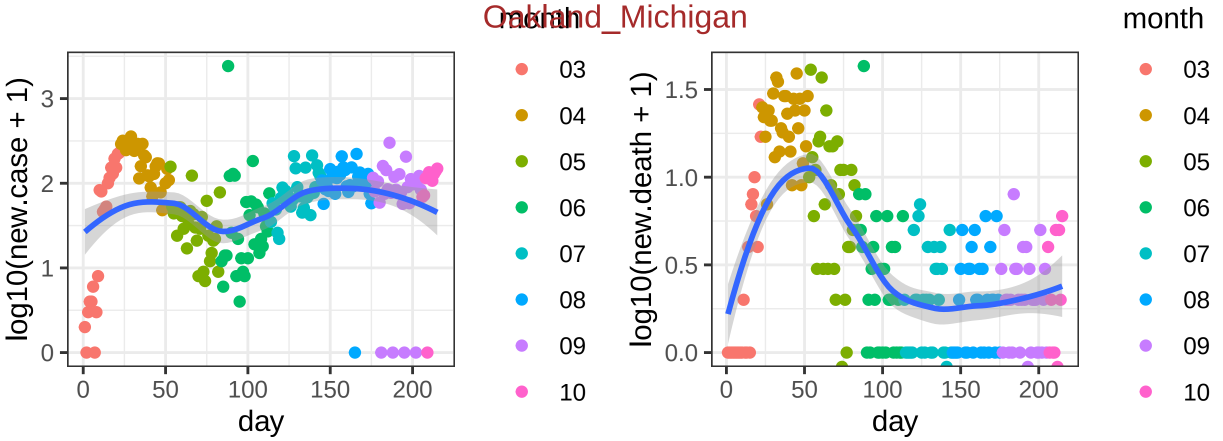
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10


Riverside_California

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

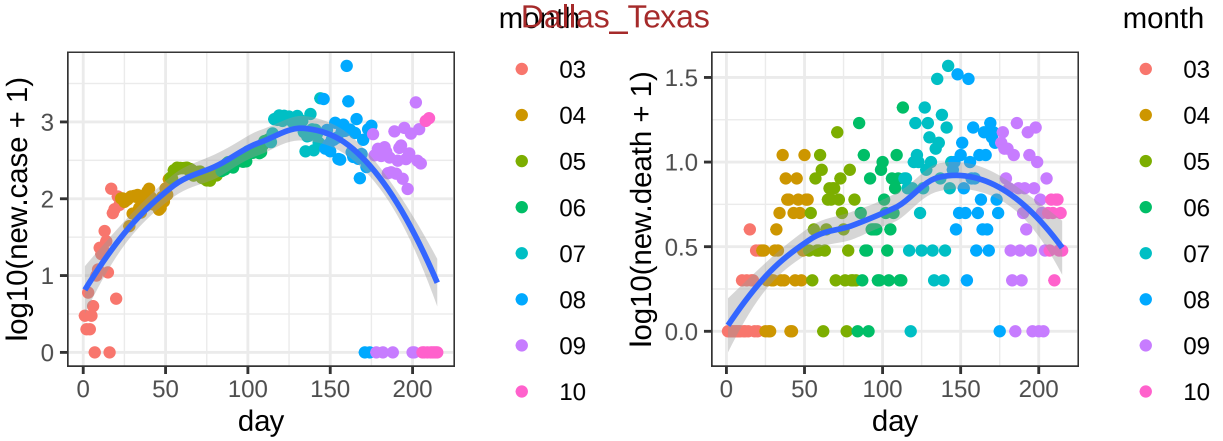

Passaic_New Jersey

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



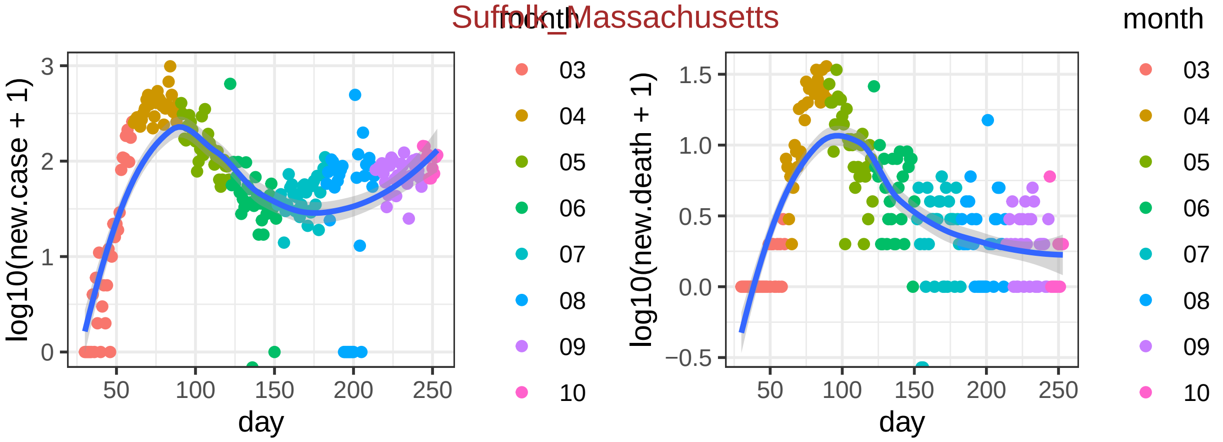
Oakland_Michigan

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



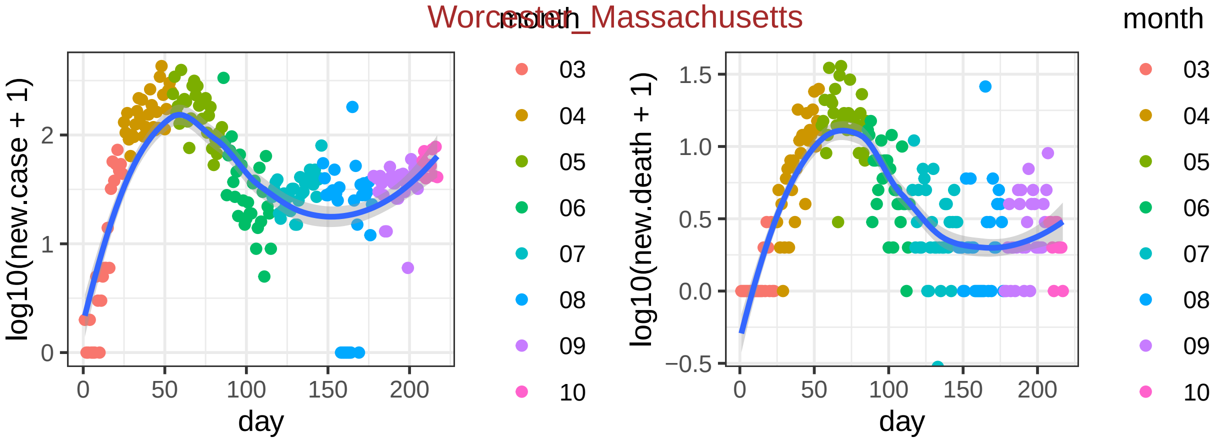
Dallas_Texas

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

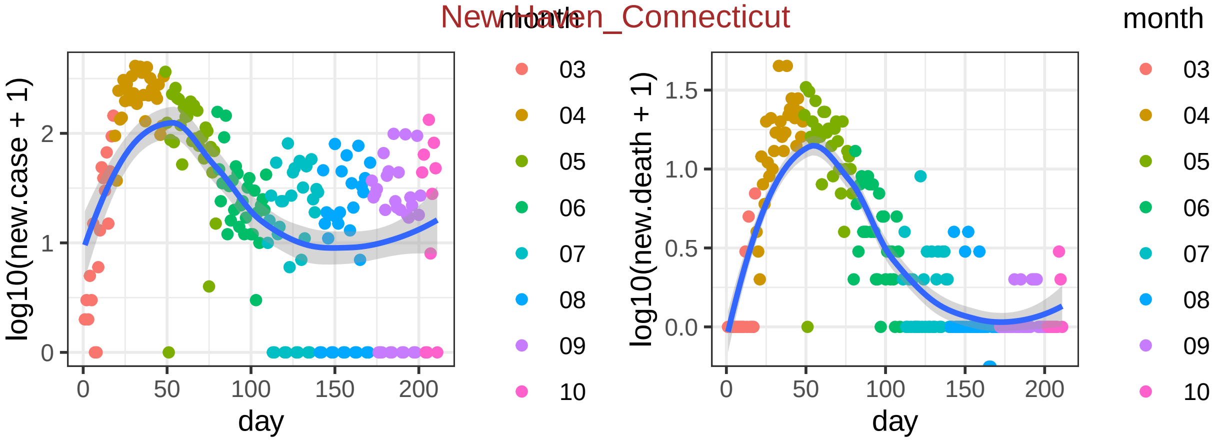


Suffolk_Massachusetts

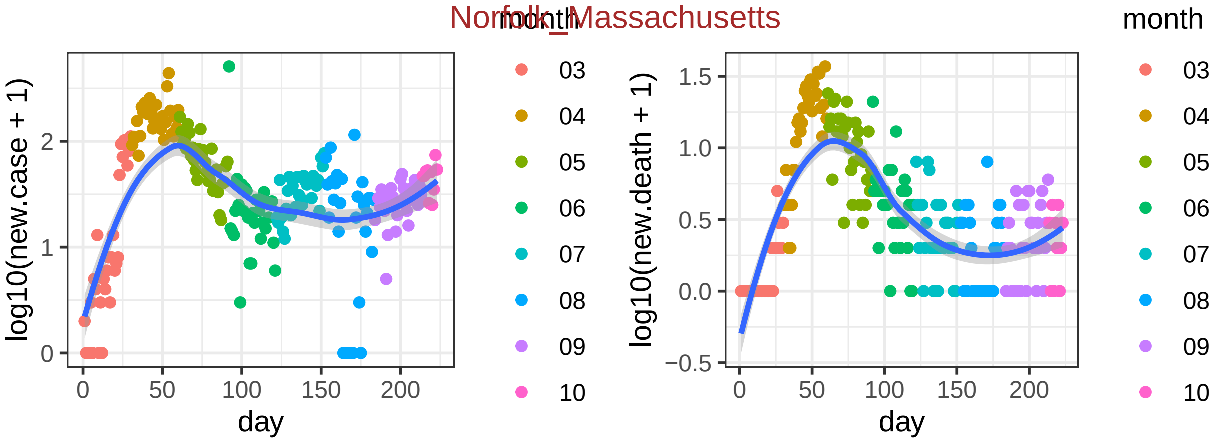
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01


Worcester_Massachusetts

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

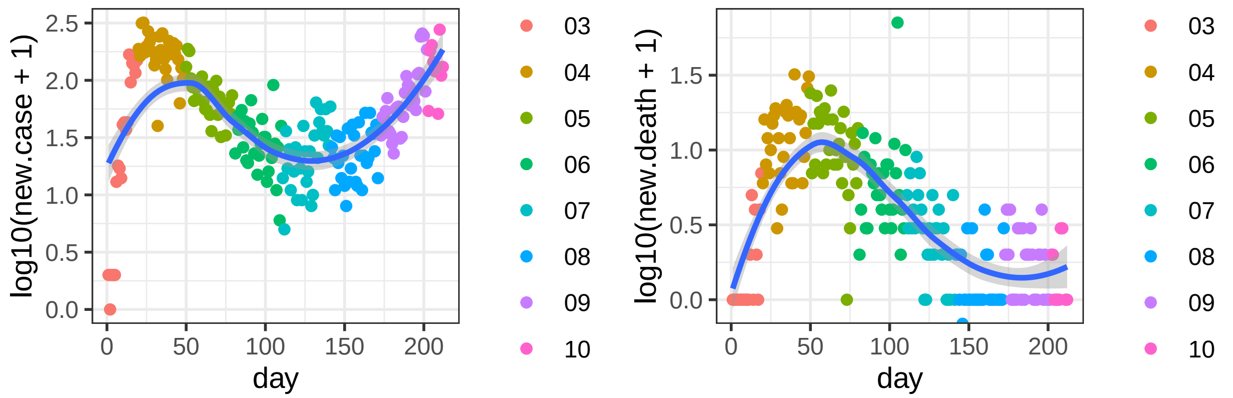

NewHaven_Connecticut

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


Norfolk_Massachusetts

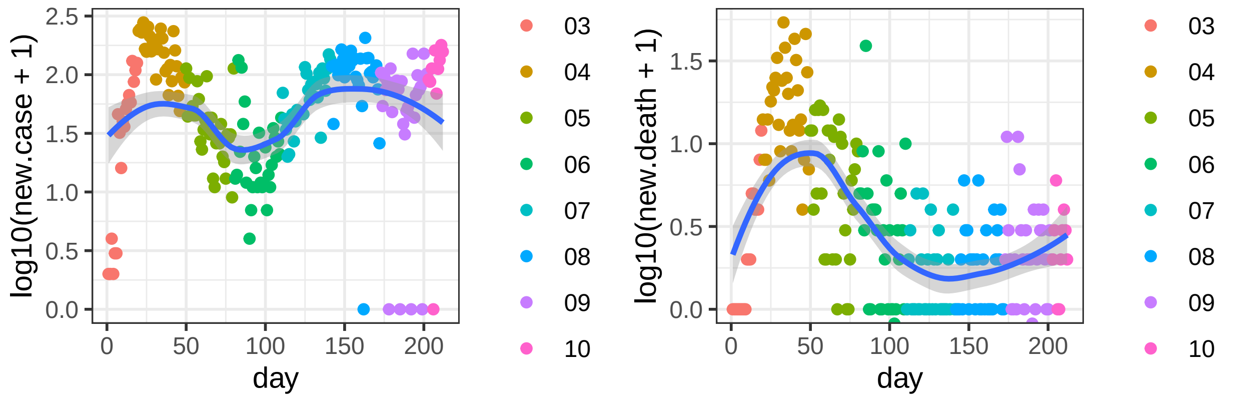
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Ocean_New Jersey



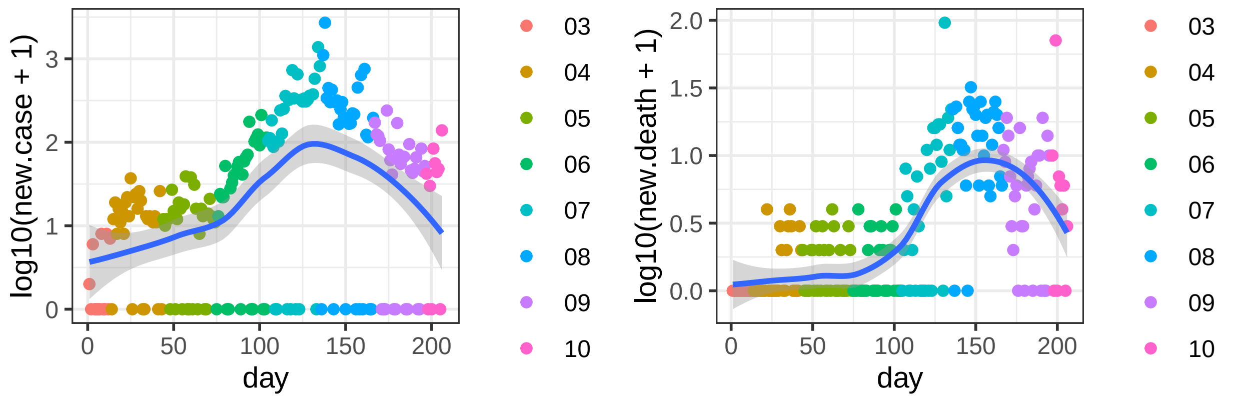
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Macomb_Michigan

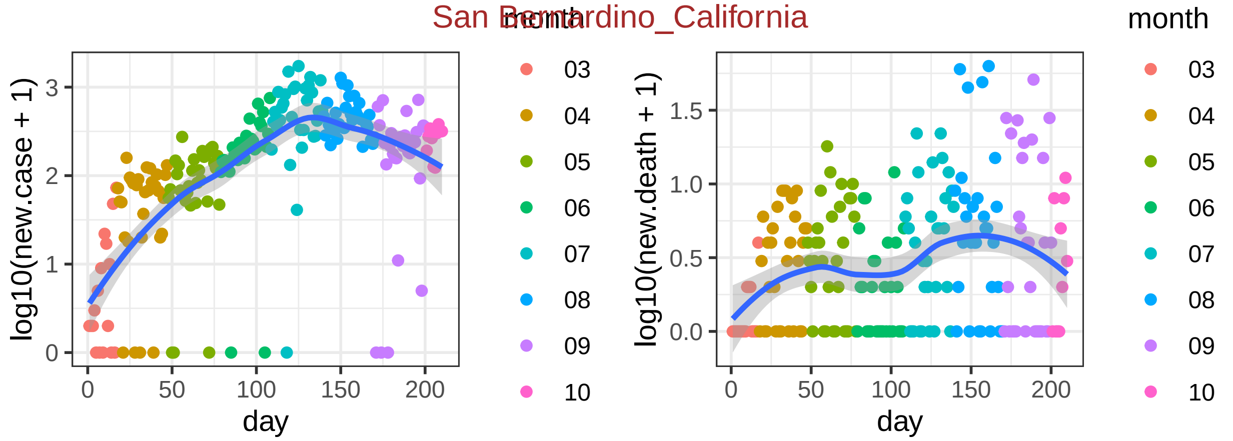


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

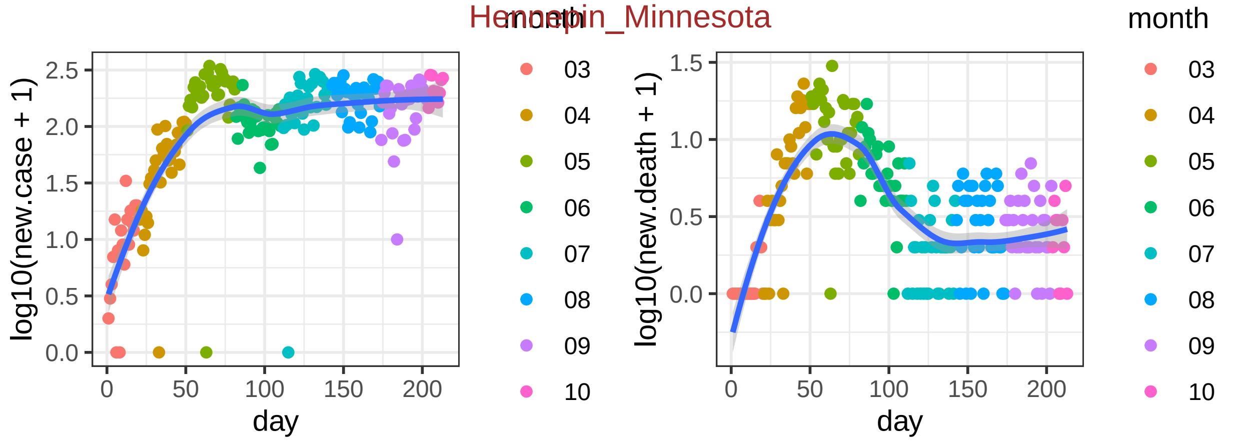
Cameron_Texas



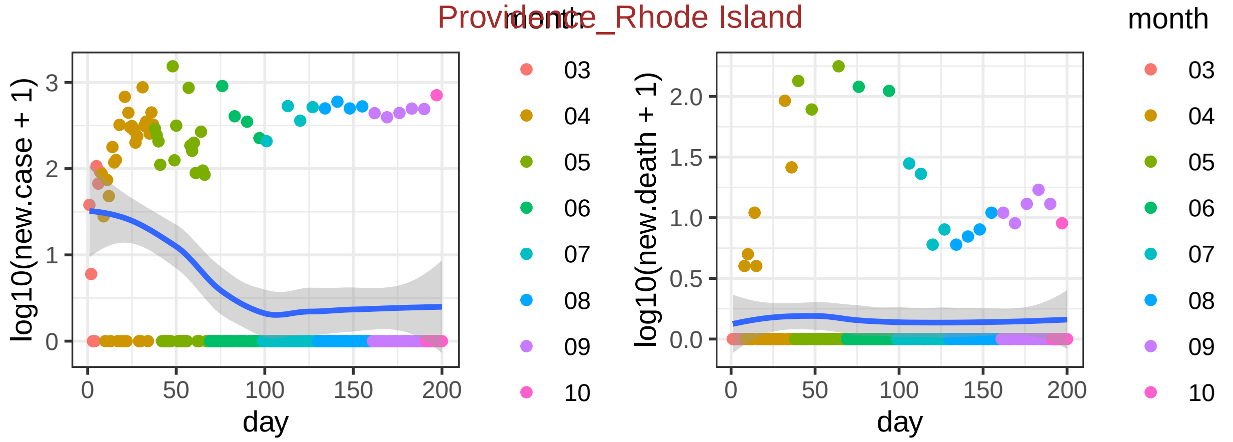
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−19


San Bernardino_California

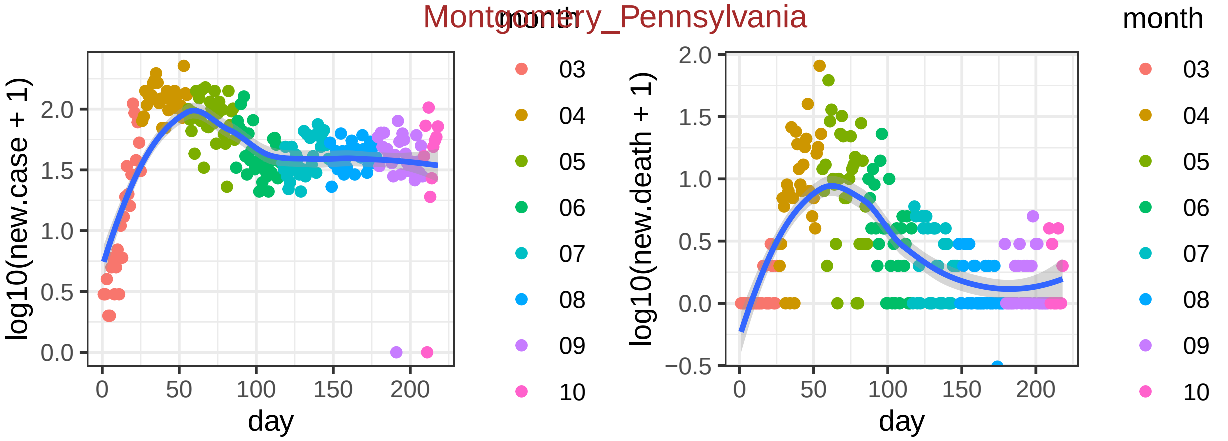
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15


Hennepin_Minnesota

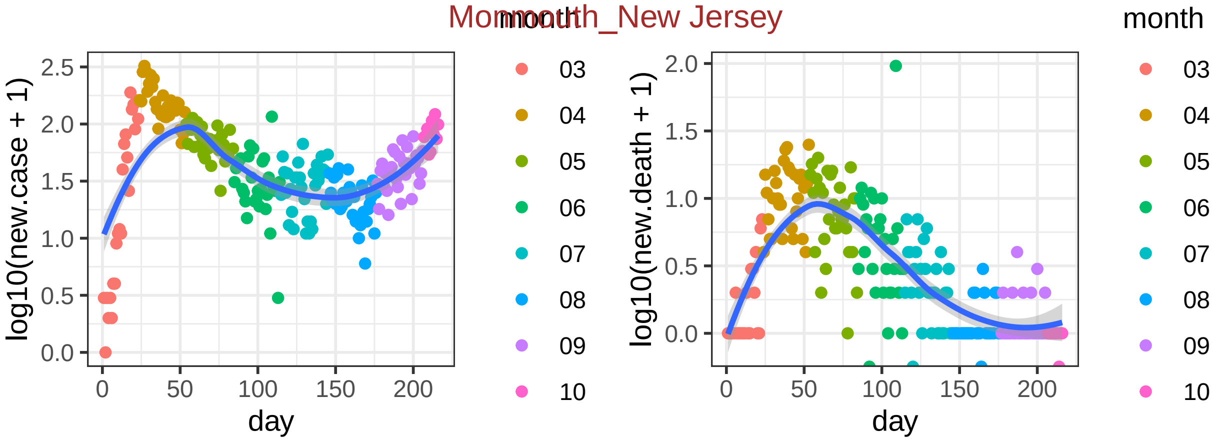
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12


Providence_Rhode Island

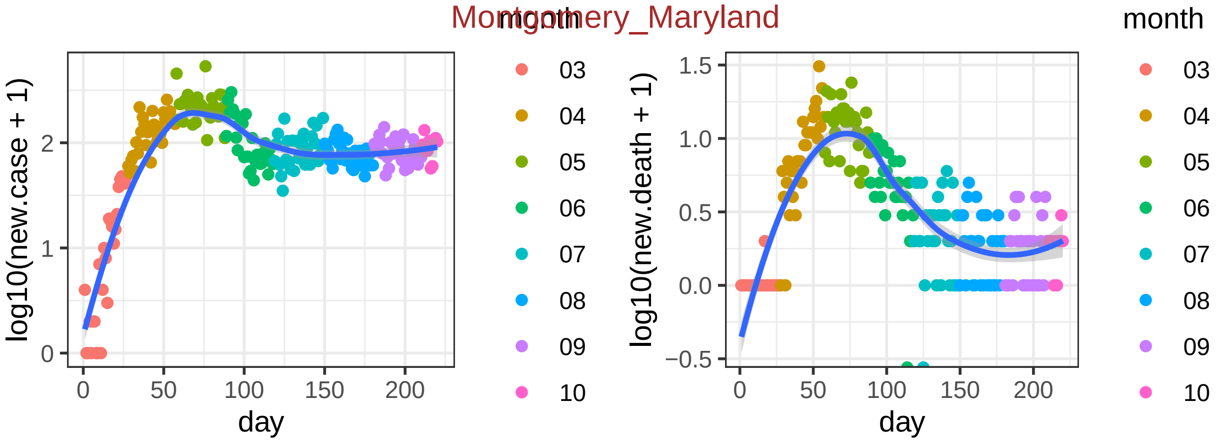
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25


Montgomery_Pennsylvania

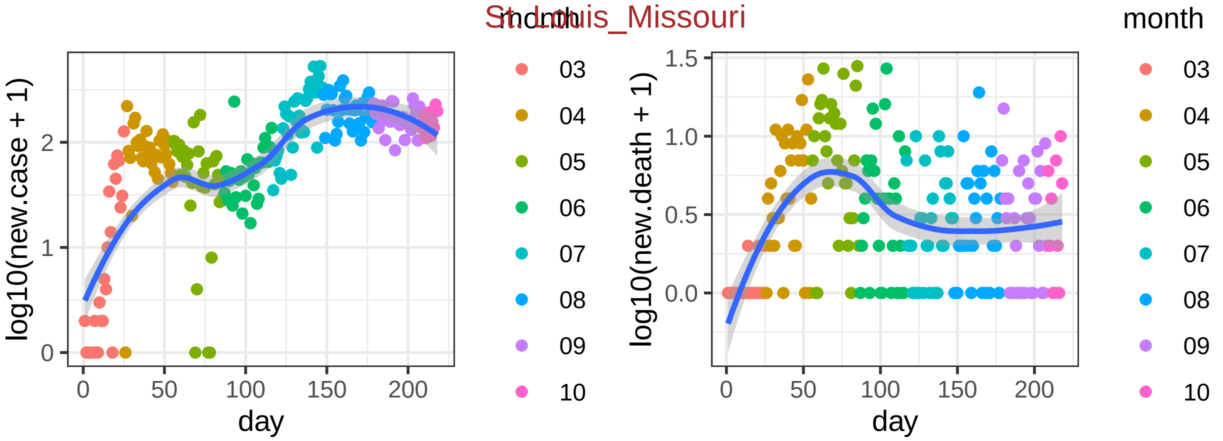
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07


Monmouth_New Jersey

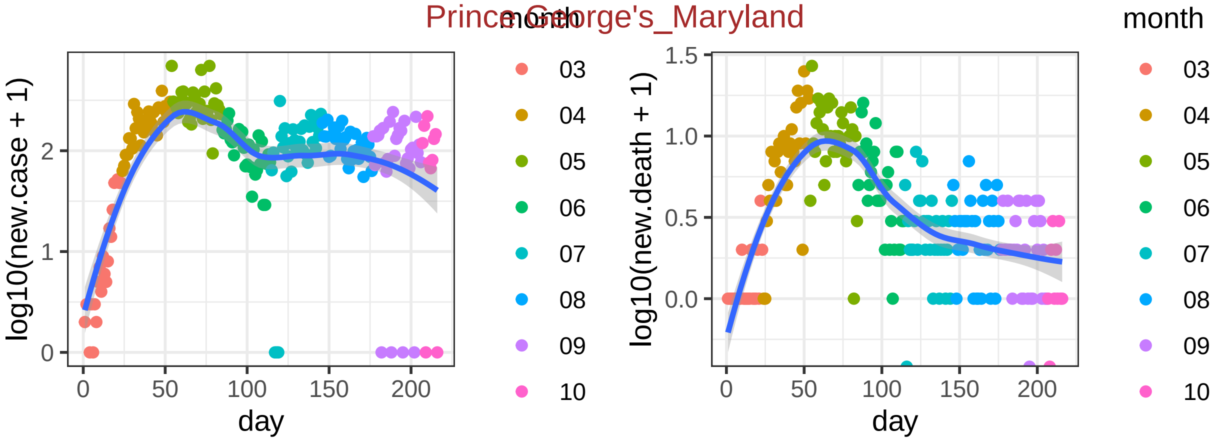
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09


Montgomery_Maryland

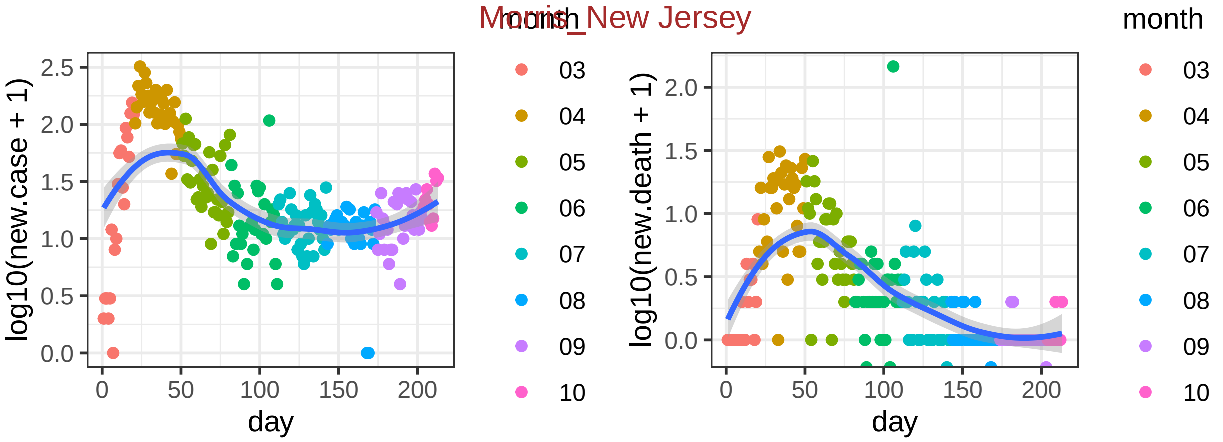
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



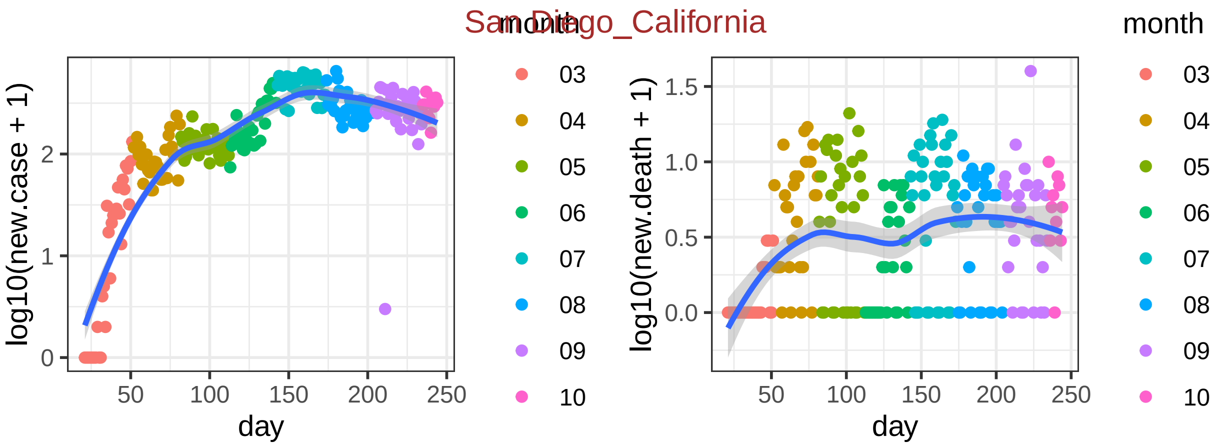
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07



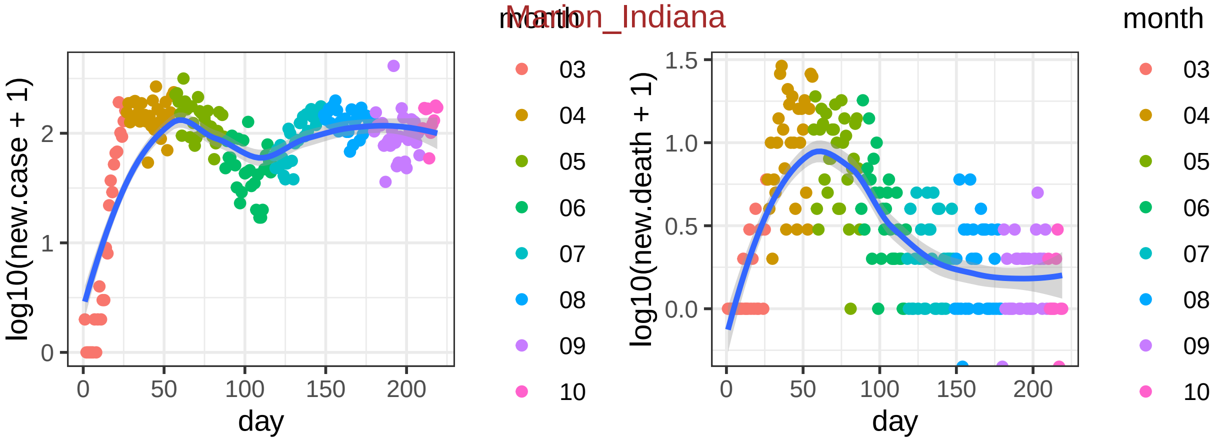
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09



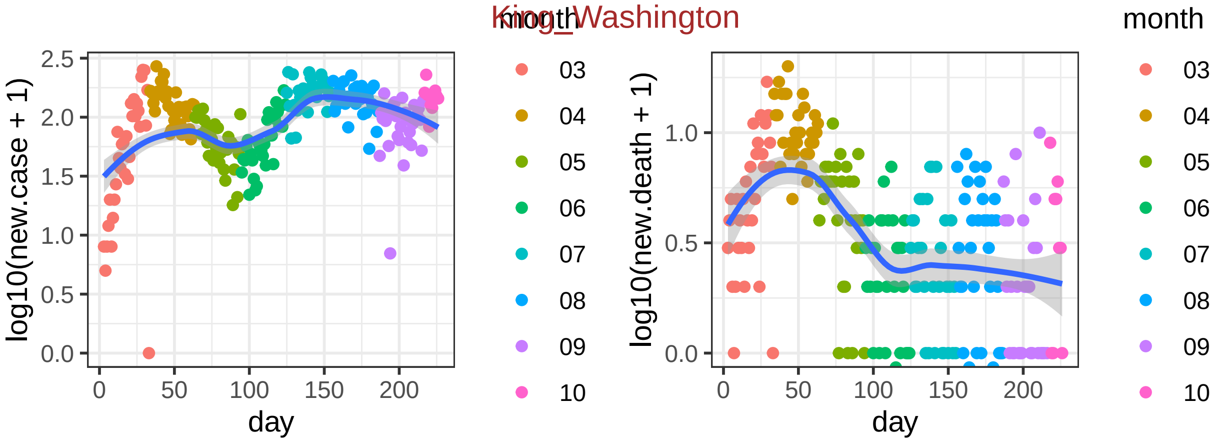
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12



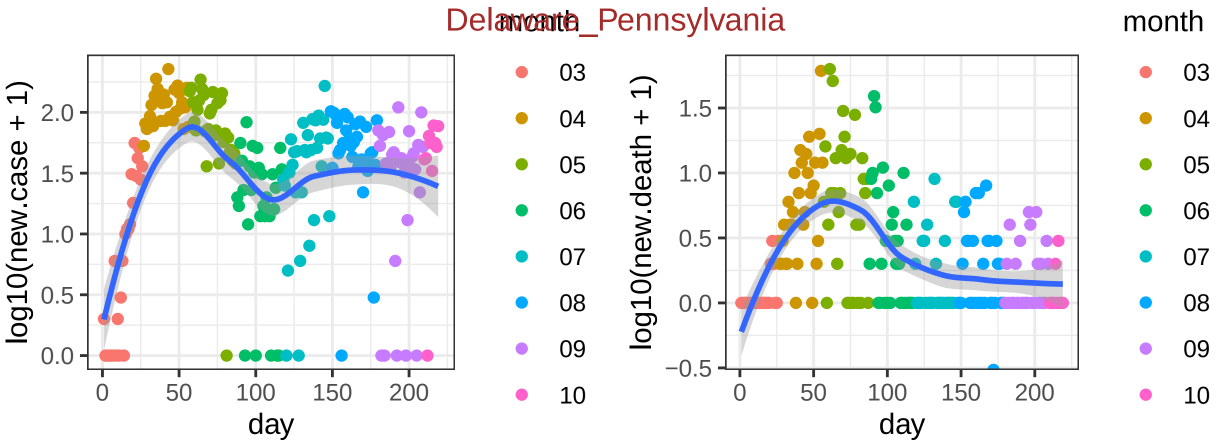
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



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



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

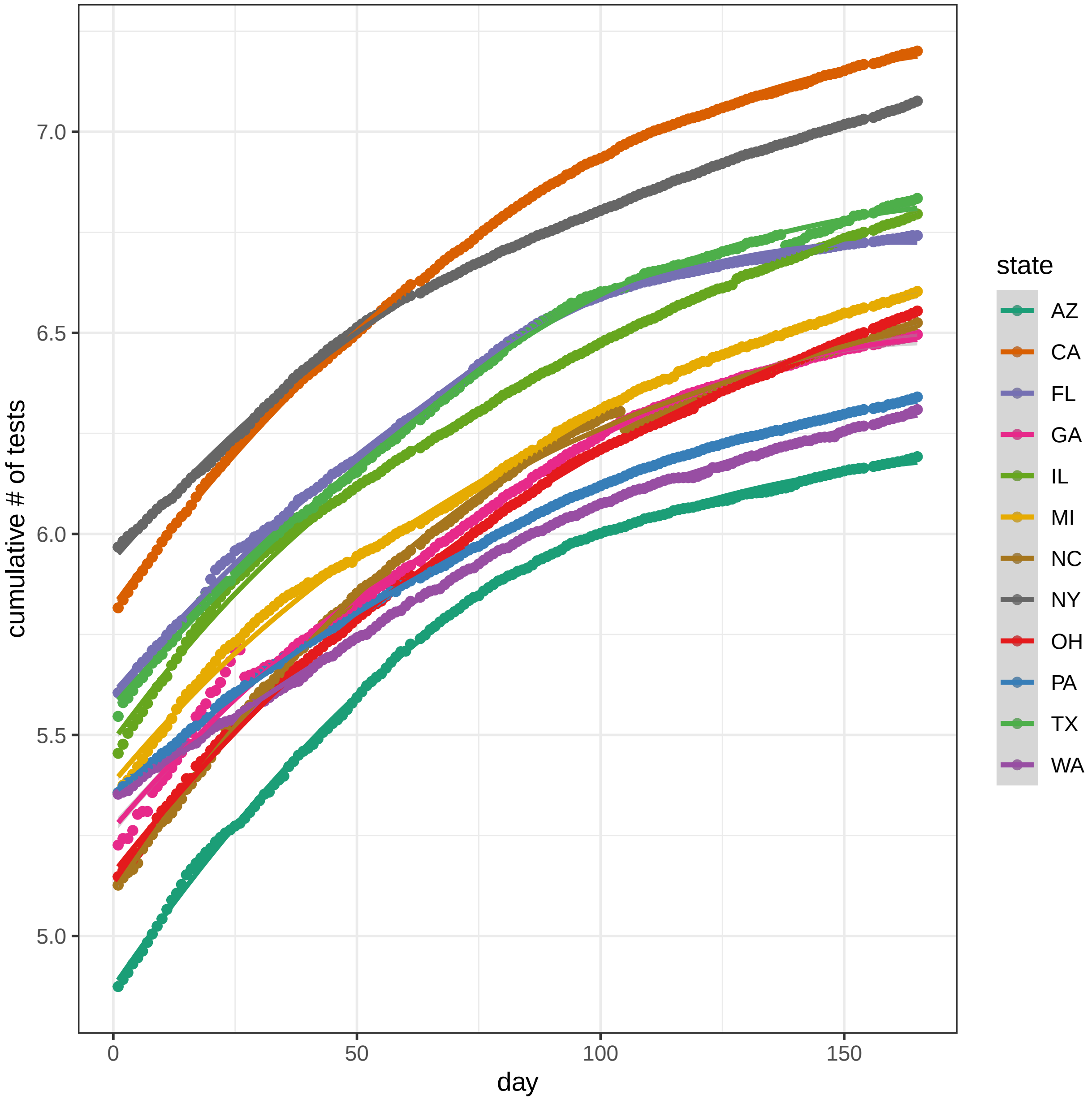


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

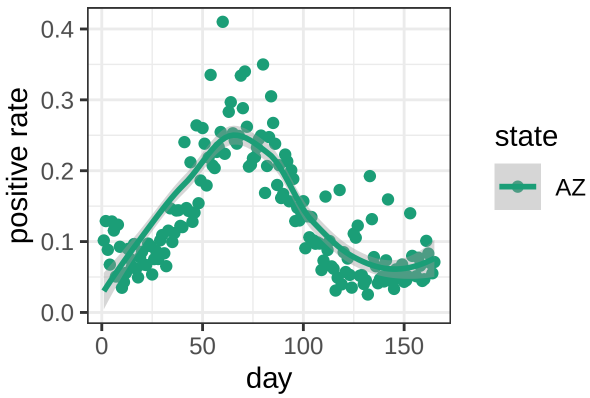
COVID Trackng

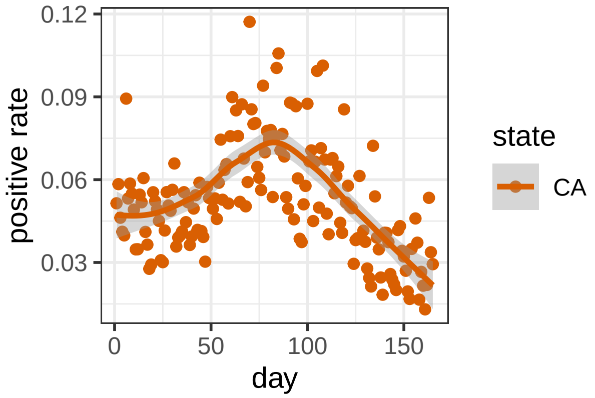
The positive rates of testing can be an indicator on how much the COVID-19 has spread. However, they can be much more noisy data since the negative testing resutls are often not reported and the tests are almost surely taken on a non-representative random sample of the population. The COVID traking project 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.

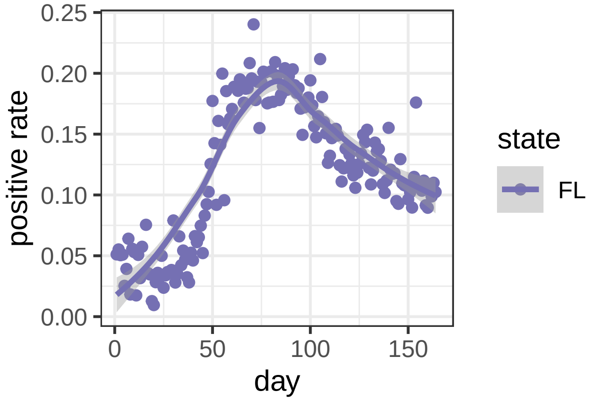
The grades of the states may change over timea and I strongly recommend checking their webiste before puting serious interpretation on the following plot.

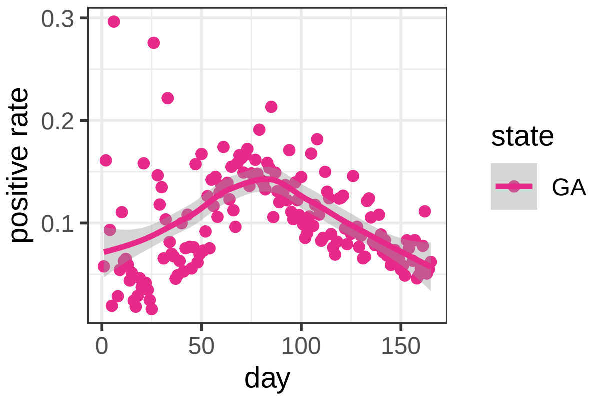


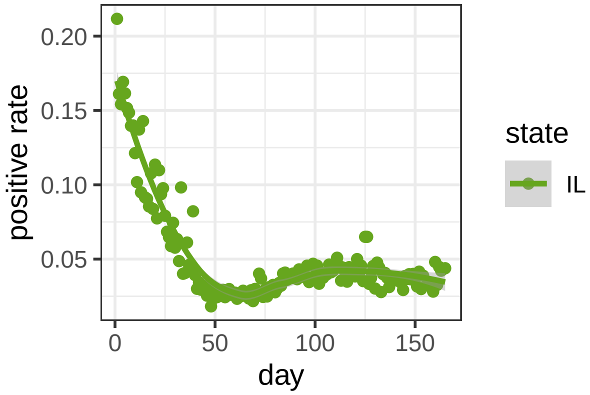
cumulative # of tests
7.0
6.5
6.0
5.5
5.0
0
50
100
150
day
state
AZ
CA
FL
GA
IL
MI
NC
NY
OH
PA
TX
WA

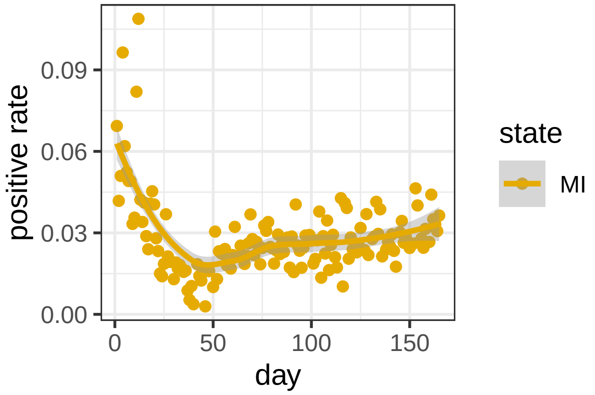

positive rate
0.4
0.3
0.2
0.1
0.0
0
50
100
150
day
state
AZ

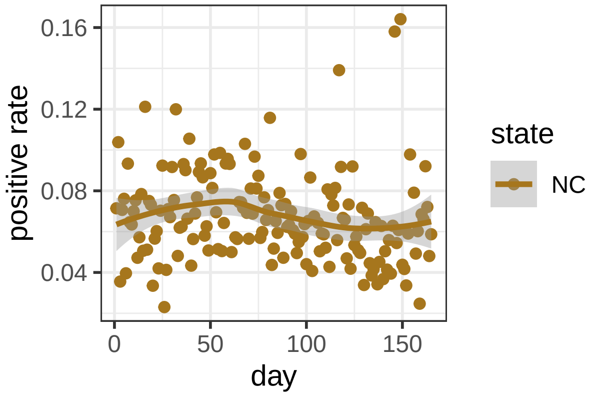

positive rate
0.12
0.09
0.06
0.03
0
50
100
150
day
state
CA

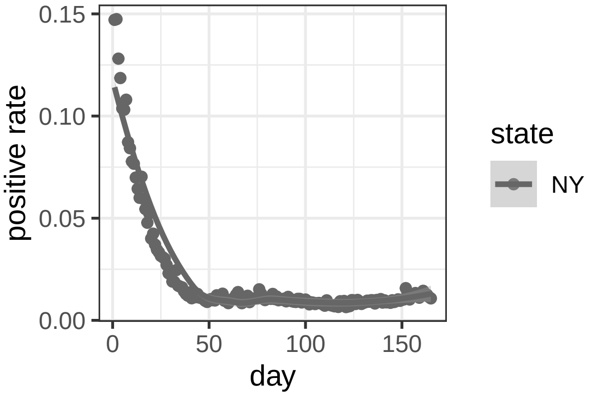

positive rate
0.25
0.20
0.15
0.10
0.05
0.00
0
50
100
150
day
state
FL

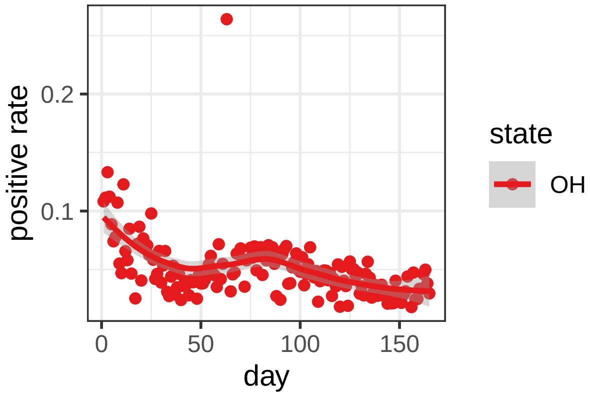

positive rate
0.3
0.2
0.1
0
50
100
150
day
state
GA

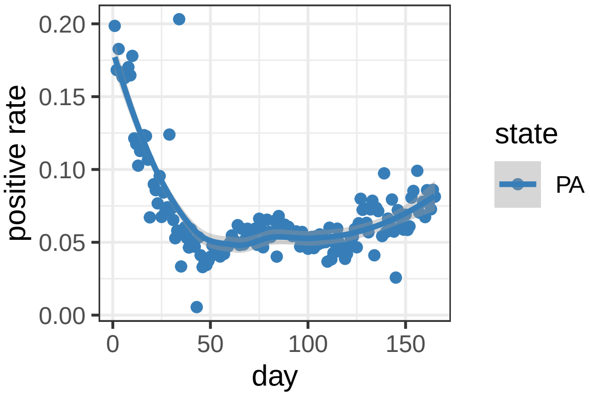

positive rate
0.20
0.15
0.10
0.05
0
50
100
150
day
state
IL

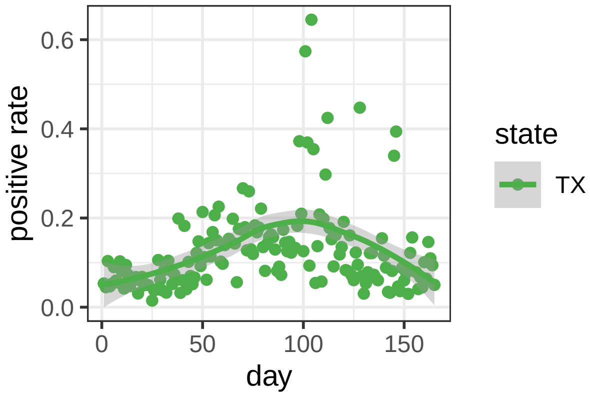

positive rate
0.09
0.06
0.03
0.00
0
50
100
150
day
state
MI

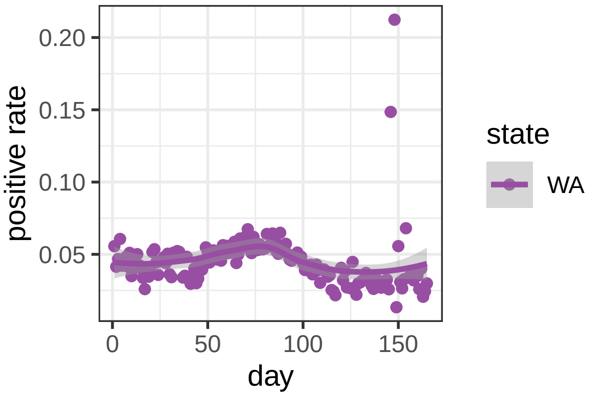












Session information

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.6
##
## 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] RColorBrewer_1.1-2 httr_1.4.1         ggpubr_0.2.5       magrittr_1.5
## [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    farver_2.0.3     purrr_0.3.3      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
```