Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-11-06

Contents

Introduction 1

JHU 2

time series data 2

daily reports data 6

NY Times 7

state level data 7

county level data 18

COVID Trackng 36

Session information 39

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

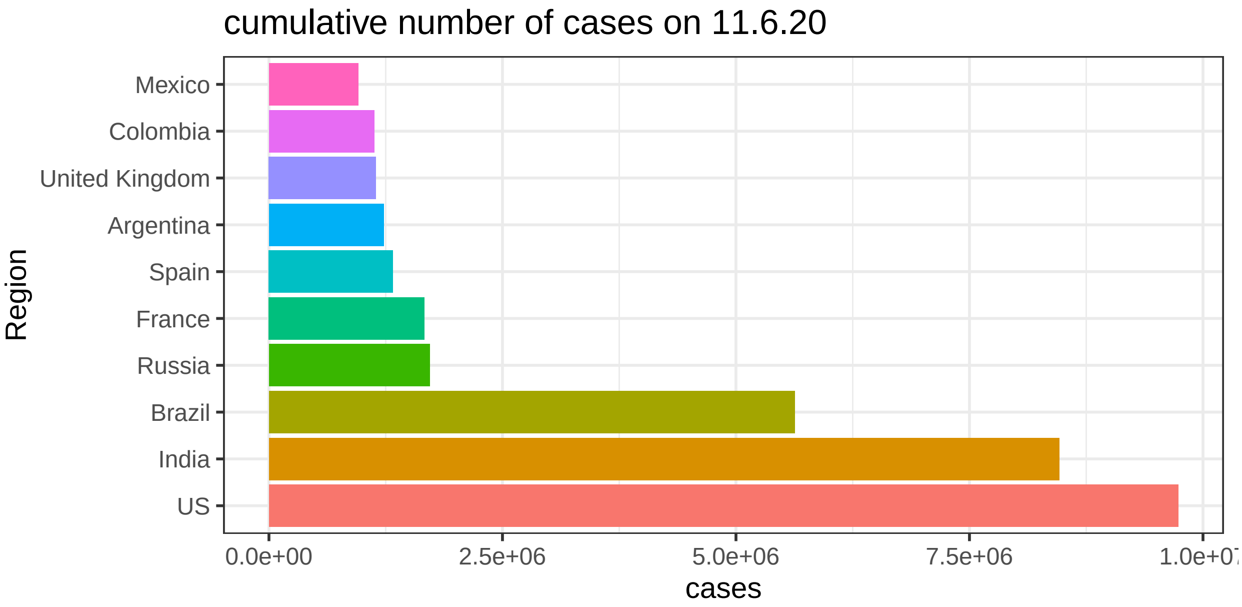
JHU

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

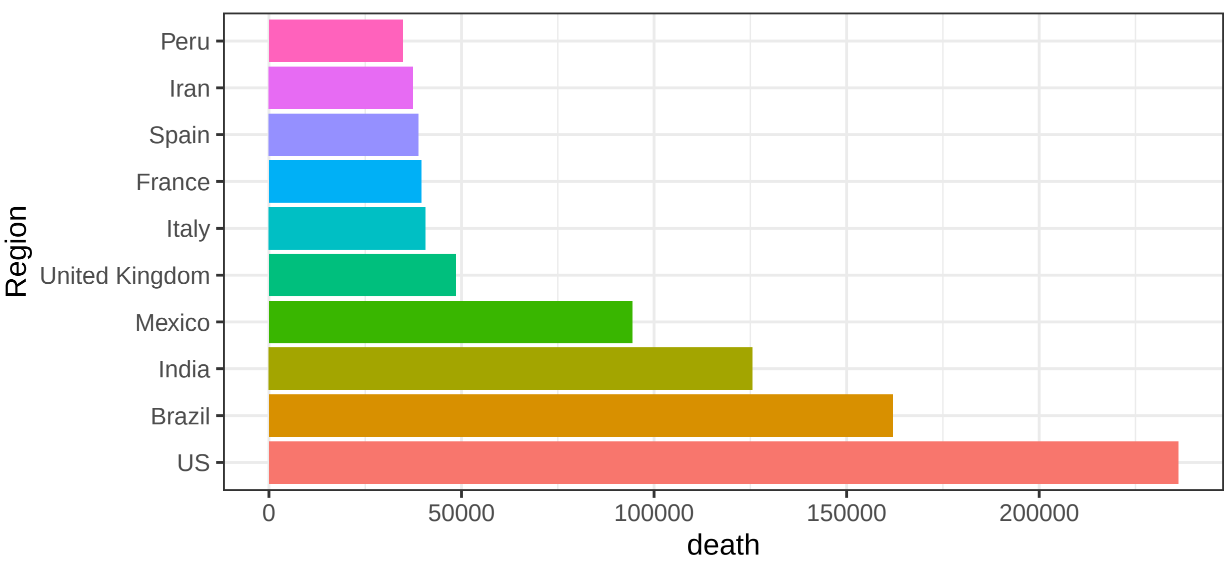
time series data

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.

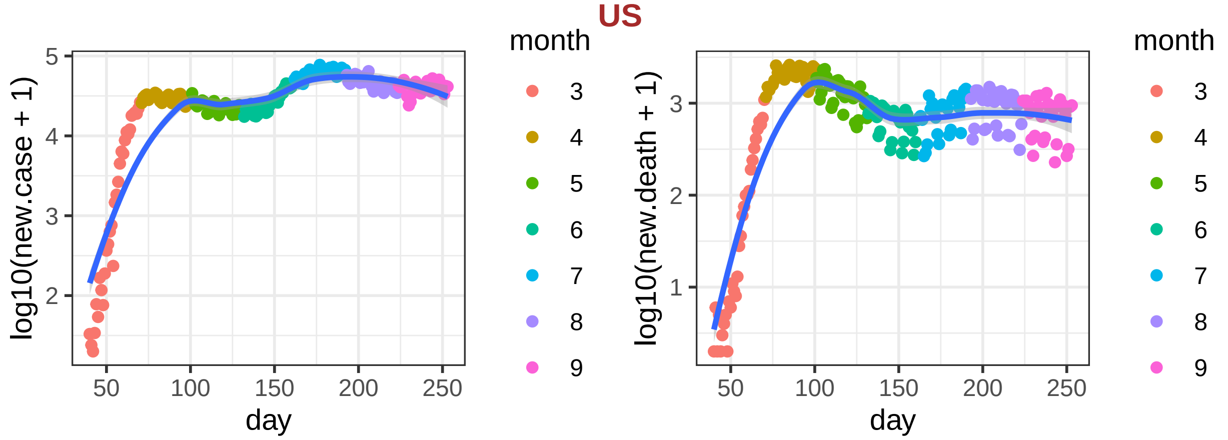


cumulative number of deaths on 11.6.20



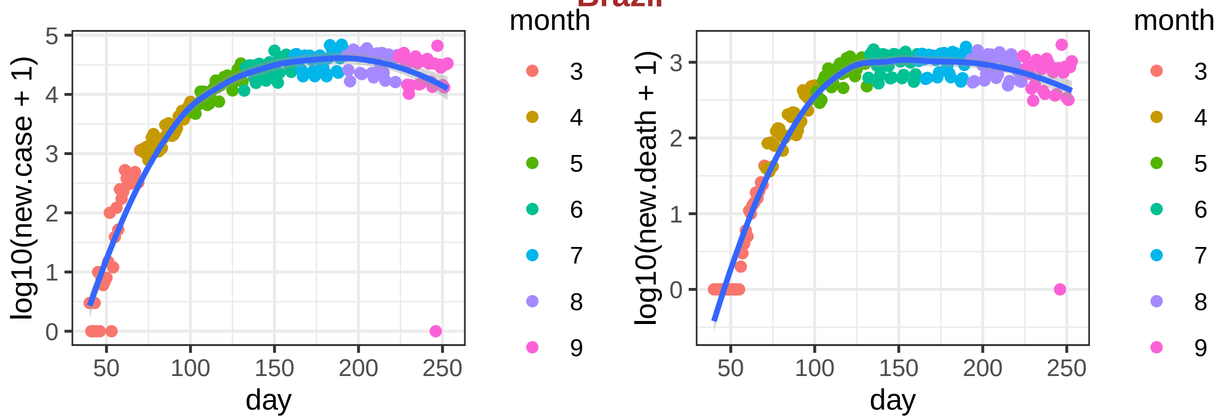
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US

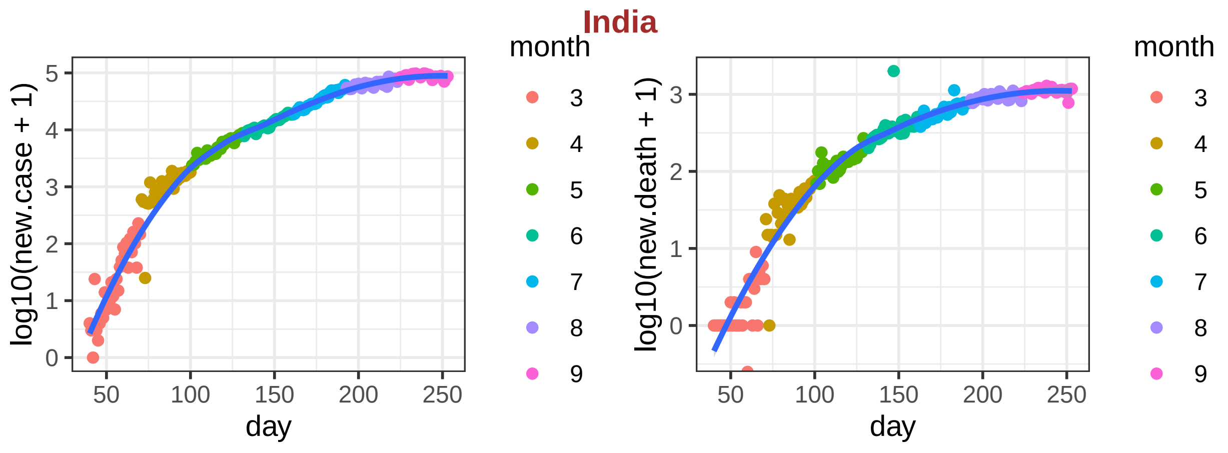


data source: https://github.com/CSSEGISandData/COVID−19

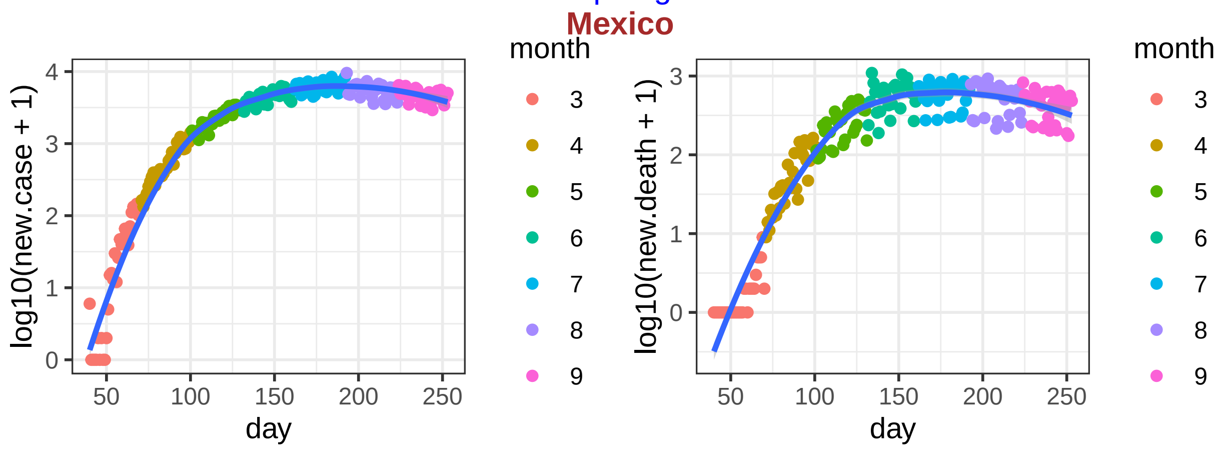
Brazil



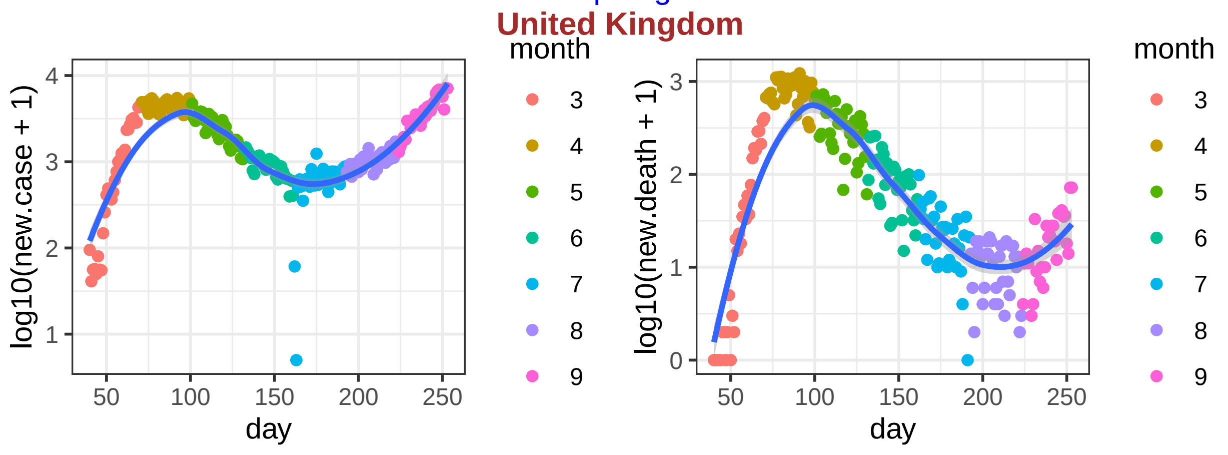
data source: https://github.com/CSSEGISandData/COVID−19



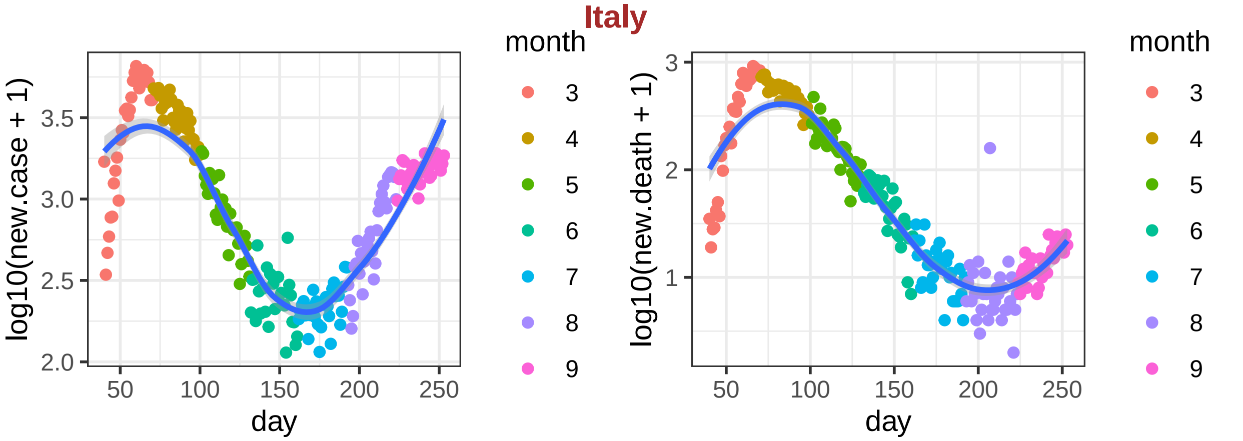
data source: https://github.com/CSSEGISandData/COVID−19



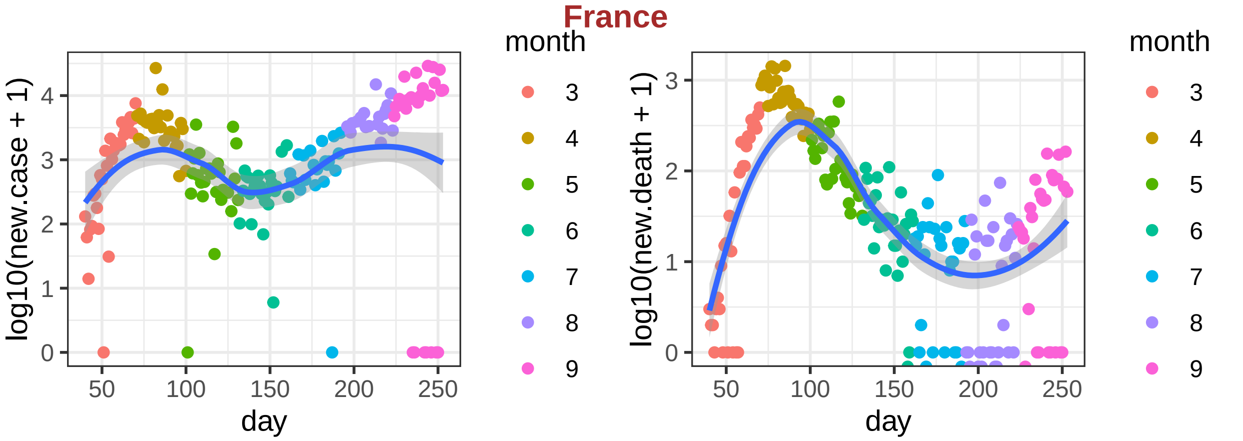
data source: https://github.com/CSSEGISandData/COVID−19



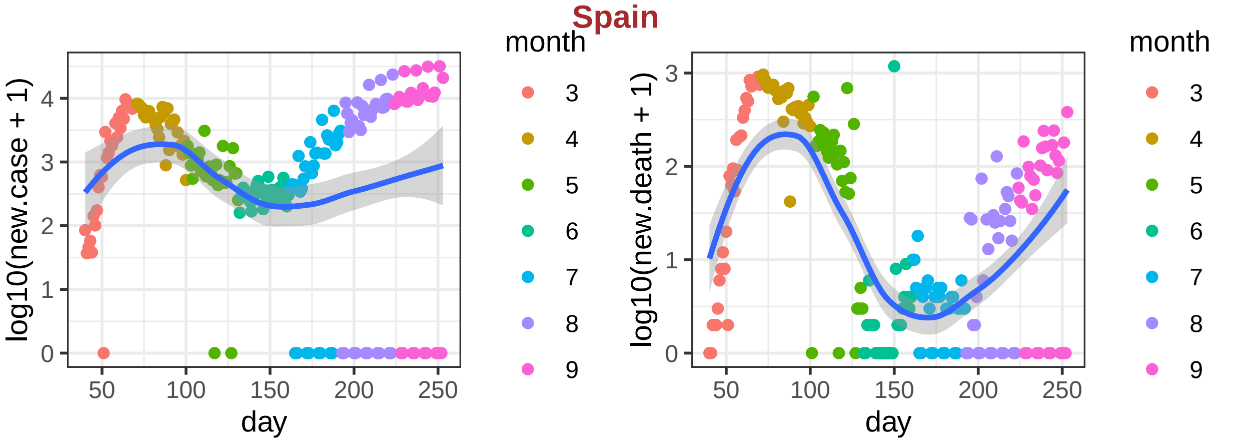
data source: https://github.com/CSSEGISandData/COVID−19



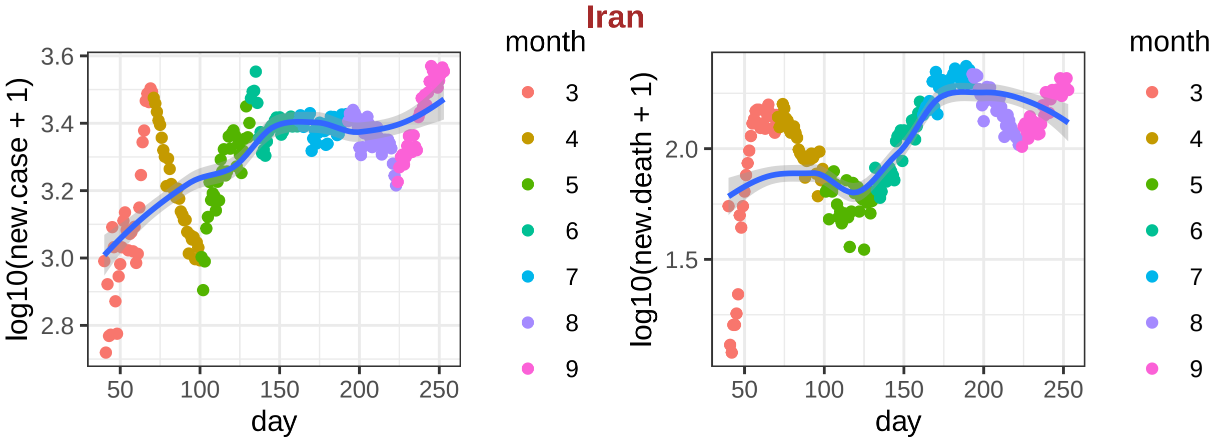
data source: https://github.com/CSSEGISandData/COVID−19



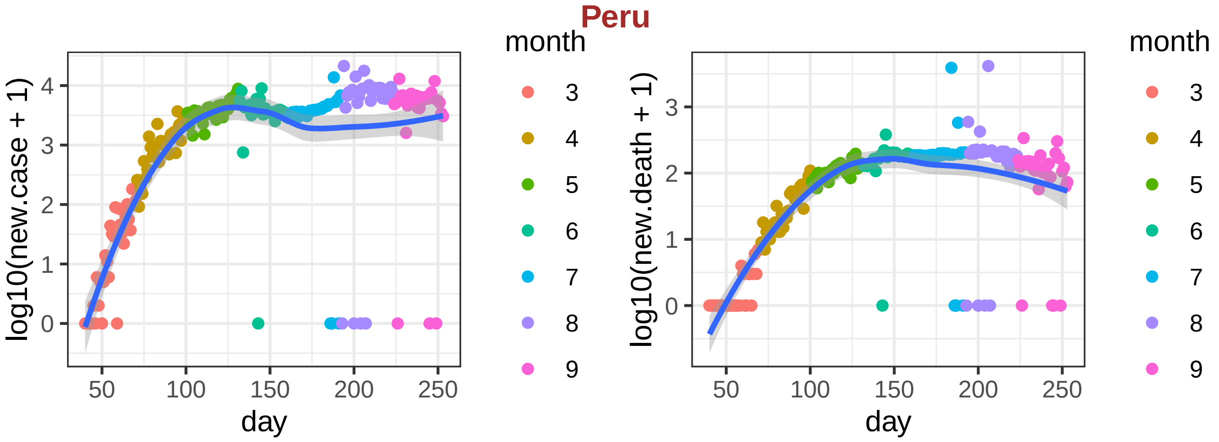
data source: https://github.com/CSSEGISandData/COVID−19



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



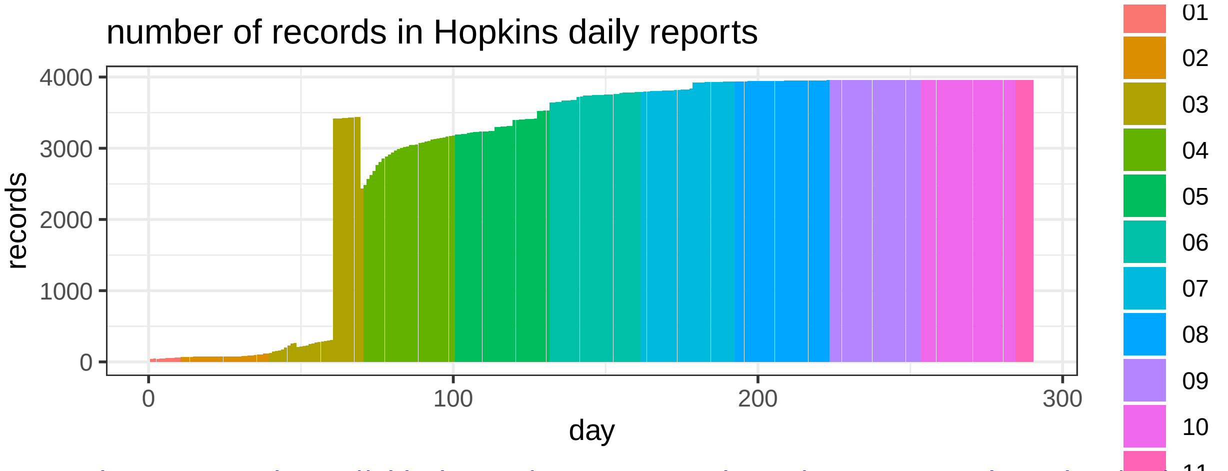
data source: https://github.com/CSSEGISandData/COVID−19



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

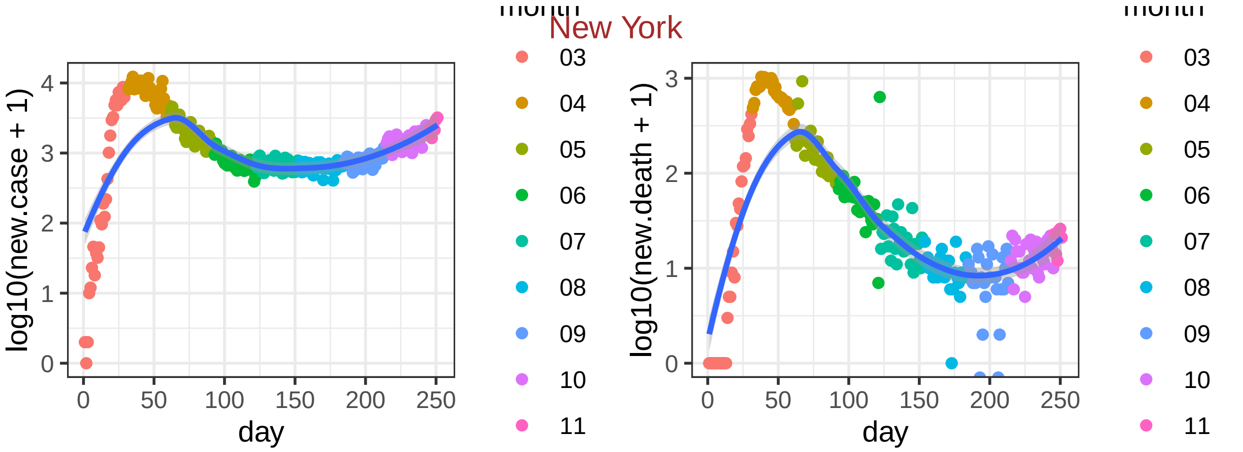
```
## [1] "2020-11-06"
```

state level data

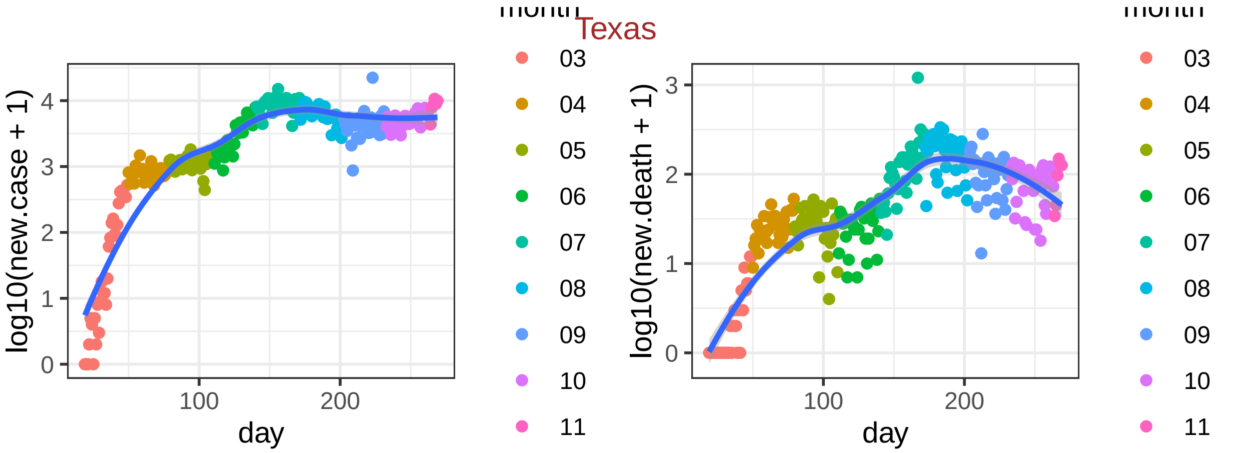
First check the 30 states with the largest number of deaths.

```
##             date          state fips   cases deaths
## 13688 2020-11-06       New York   36  526767  33267
## 13701 2020-11-06          Texas   48 1003056  19125
## 13659 2020-11-06     California    6  967597  17939
## 13664 2020-11-06        Florida   12  832617  17013
## 13686 2020-11-06     New Jersey   34  251180  16416
## 13669 2020-11-06       Illinois   17  466884  10412
## 13677 2020-11-06  Massachusetts   25  167274  10106
## 13695 2020-11-06   Pennsylvania   42  229346   9052
## 13665 2020-11-06        Georgia   13  387202   8389
## 13678 2020-11-06       Michigan   26  222423   7883
## 13657 2020-11-06        Arizona    4  254961   6111
## 13674 2020-11-06      Louisiana   22  191715   6016
## 13692 2020-11-06           Ohio   39  240178   5494
## 13661 2020-11-06    Connecticut    9   78125   4671
## 13689 2020-11-06 North Carolina   37  289124   4608
## 13670 2020-11-06        Indiana   18  203332   4547
## 13676 2020-11-06       Maryland   24  151964   4194
## 13698 2020-11-06 South Carolina   45  182872   4005
## 13705 2020-11-06       Virginia   51  188770   3682
## 13700 2020-11-06      Tennessee   47  269292   3508
## 13680 2020-11-06    Mississippi   28  124854   3419
## 13681 2020-11-06       Missouri   29  209962   3217
## 13655 2020-11-06        Alabama    1  200714   3049
## 13679 2020-11-06      Minnesota   27  170361   2645
## 13706 2020-11-06     Washington   53  120123   2552
## 13660 2020-11-06       Colorado    8  125397   2404
## 13708 2020-11-06      Wisconsin   55  270074   2340
## 13658 2020-11-06       Arkansas    5  119230   2056
## 13684 2020-11-06         Nevada   32  107258   1846
## 13671 2020-11-06           Iowa   19  146267   1828
```

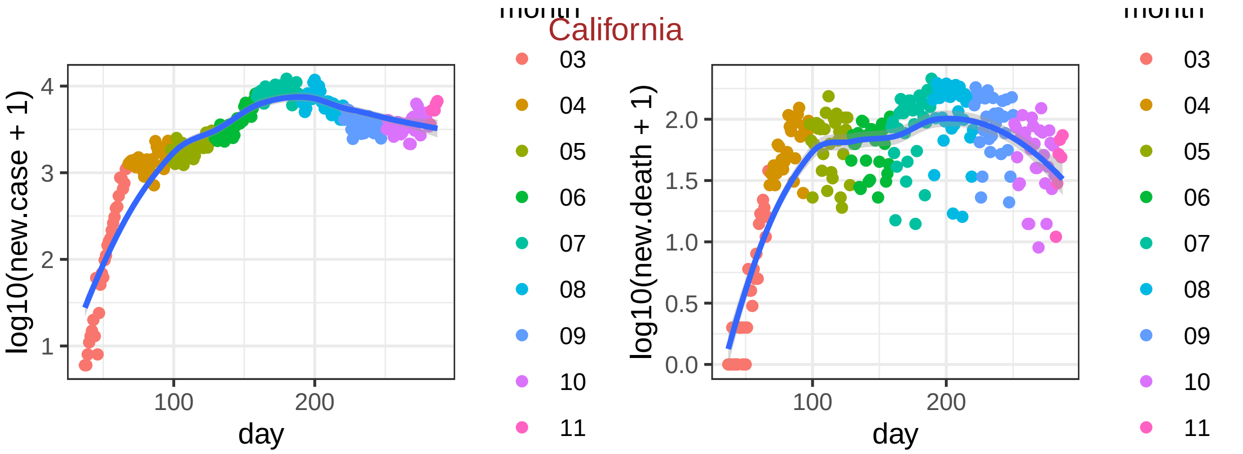
For these 30 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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



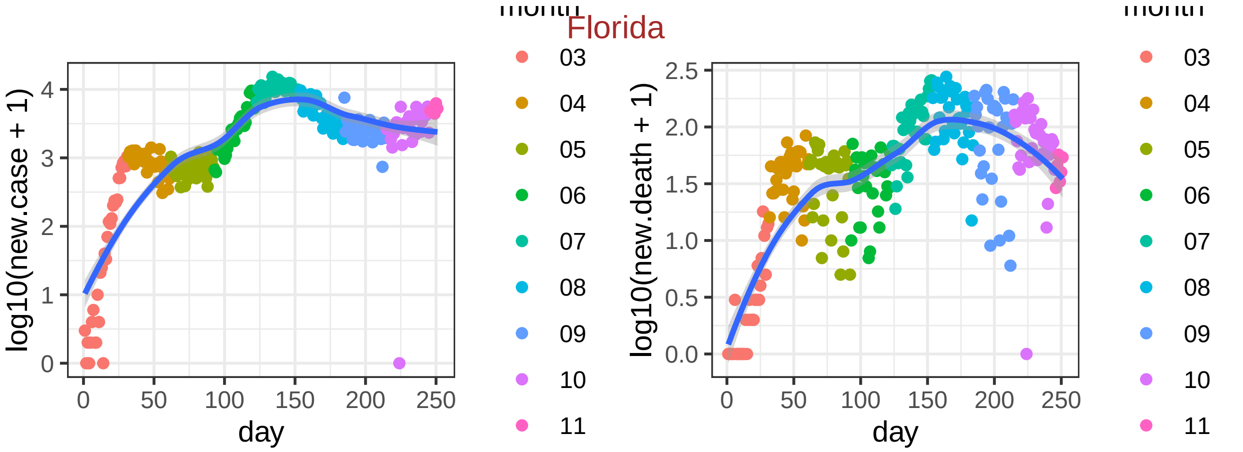
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



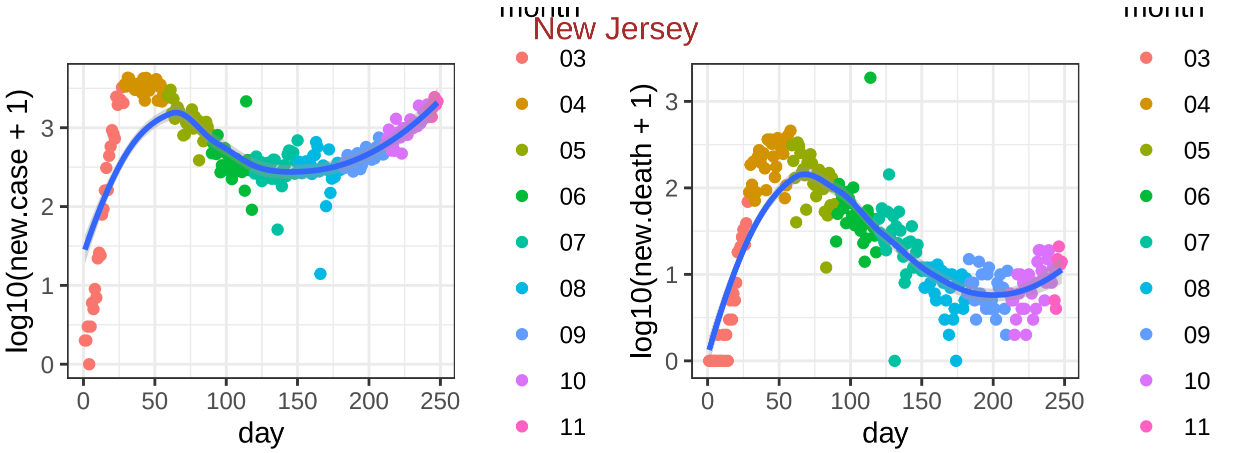
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



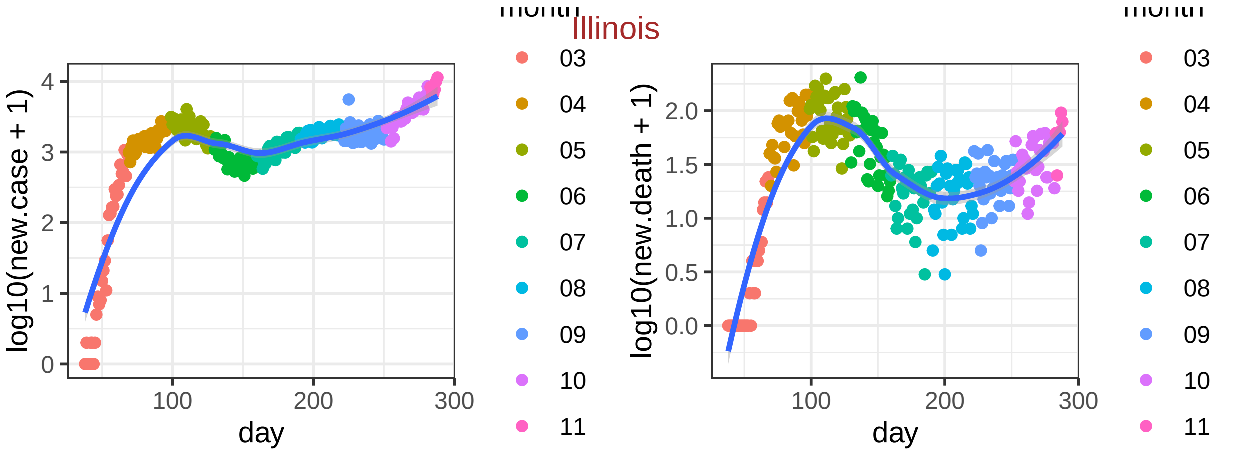
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01


Florida

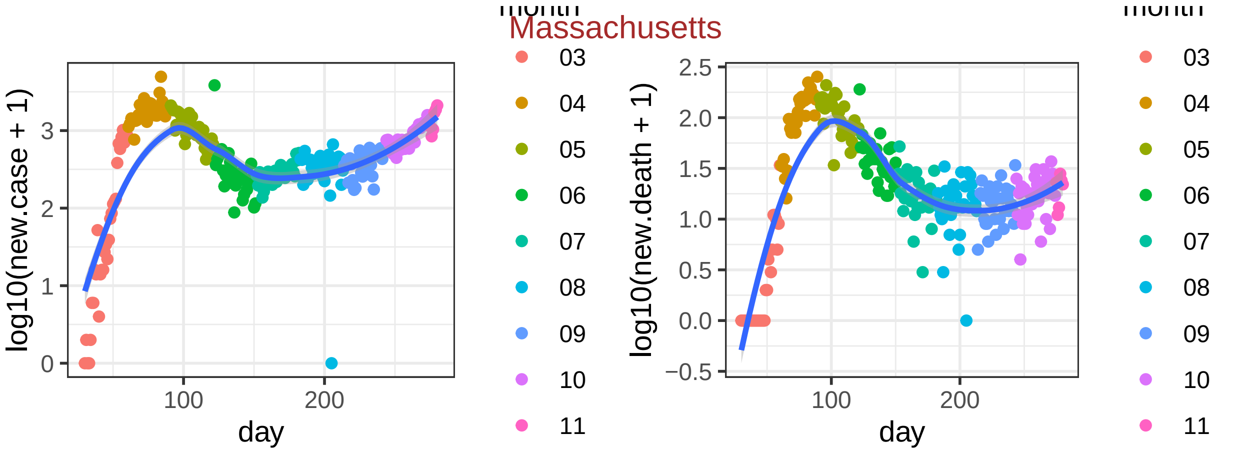
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01


New Jersey

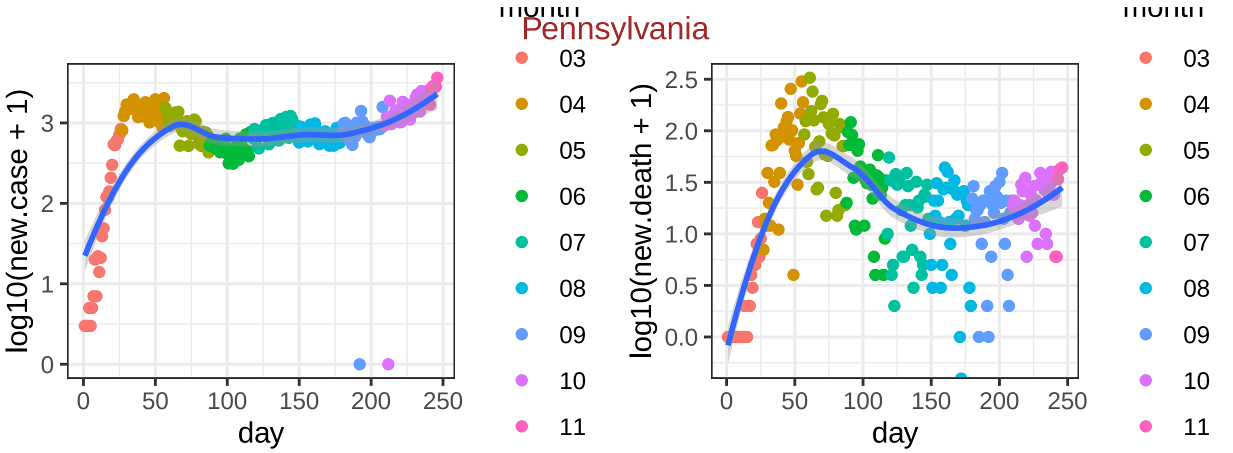
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04


Illinois

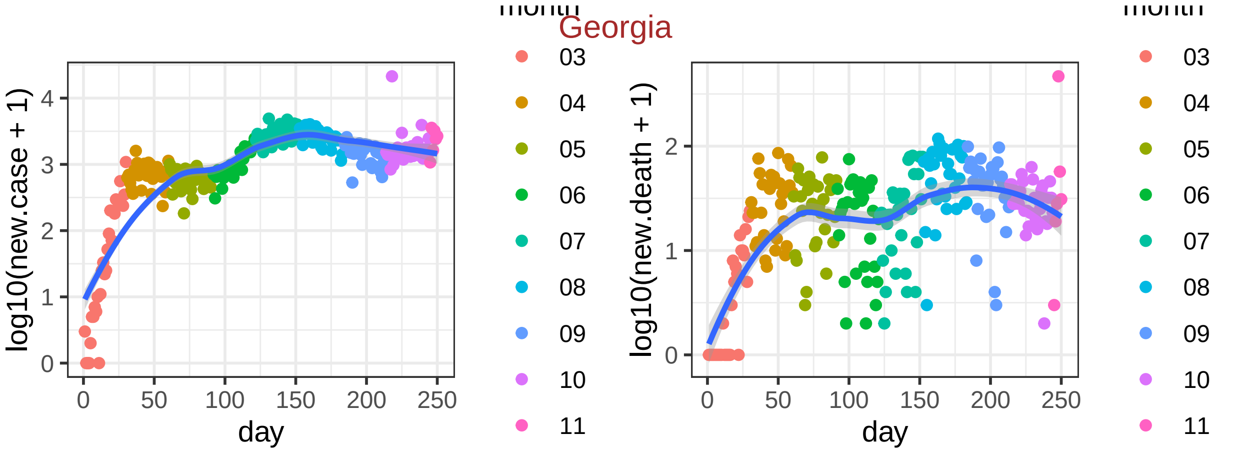
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01


Massachusetts

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

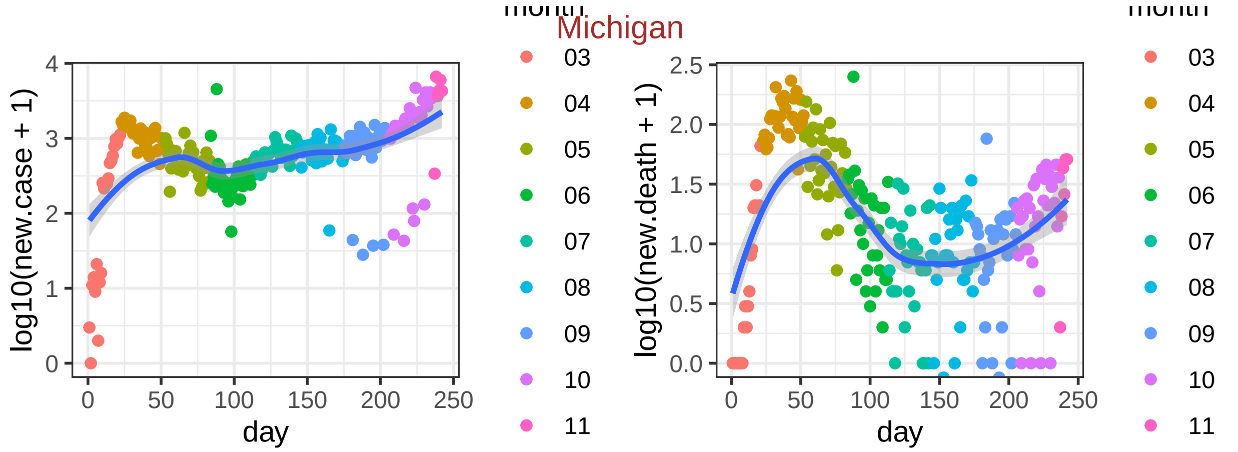

Pennsylvania

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


Georgia

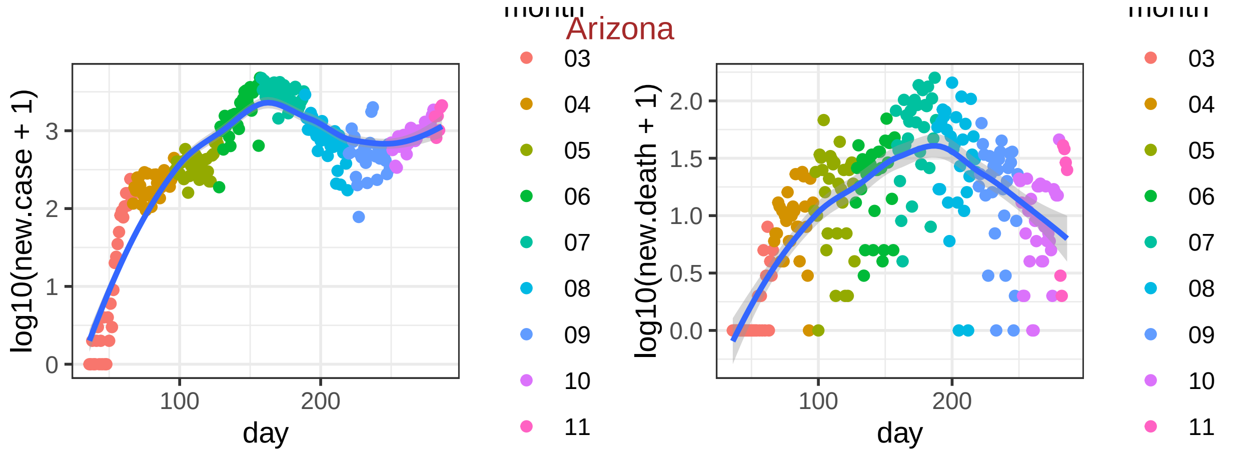
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Michigan



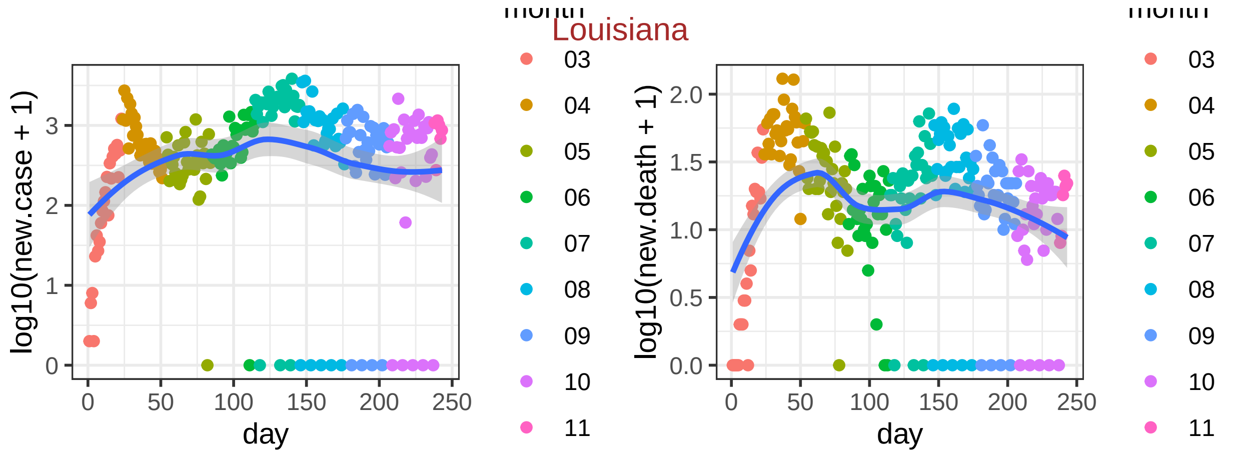
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Arizona



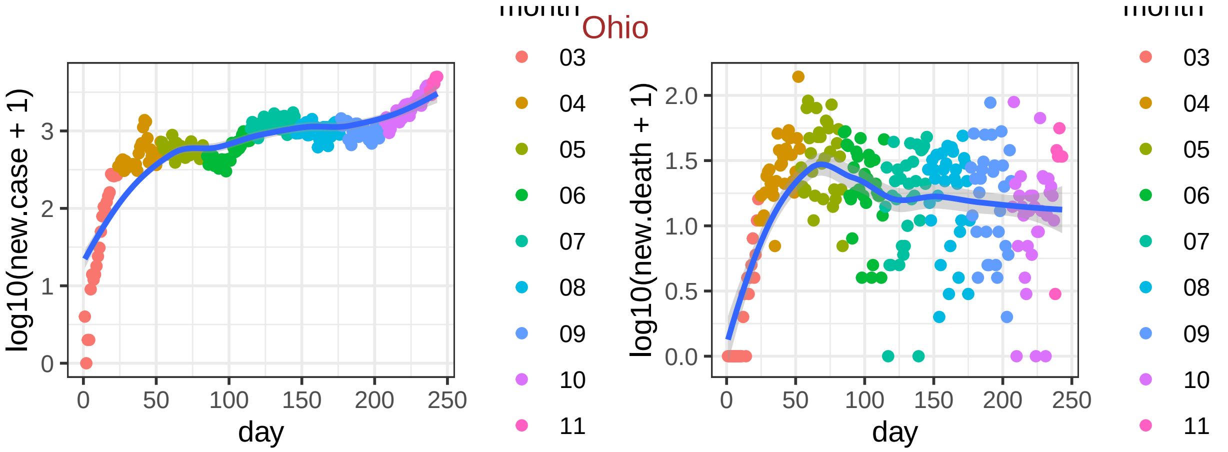
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Louisiana



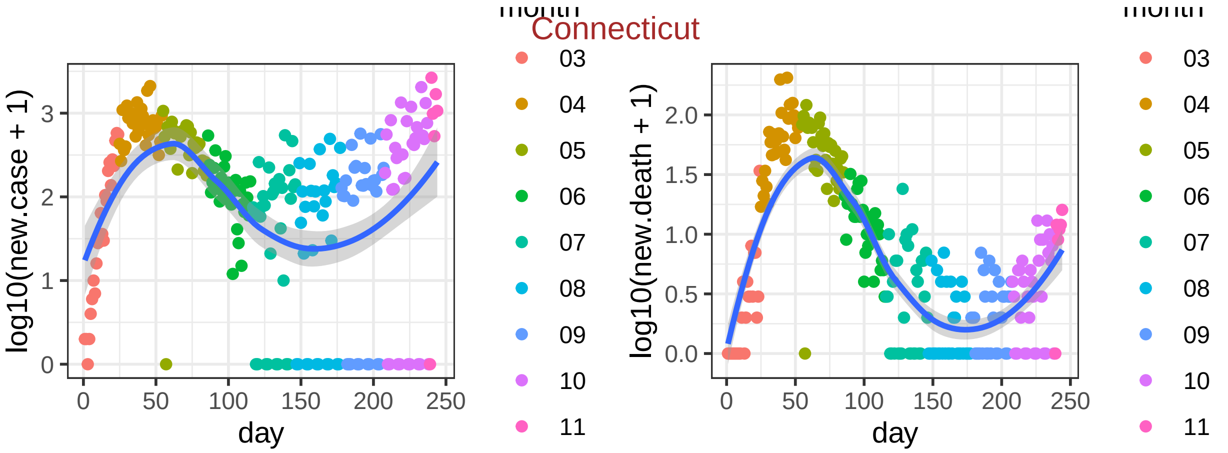
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Ohio



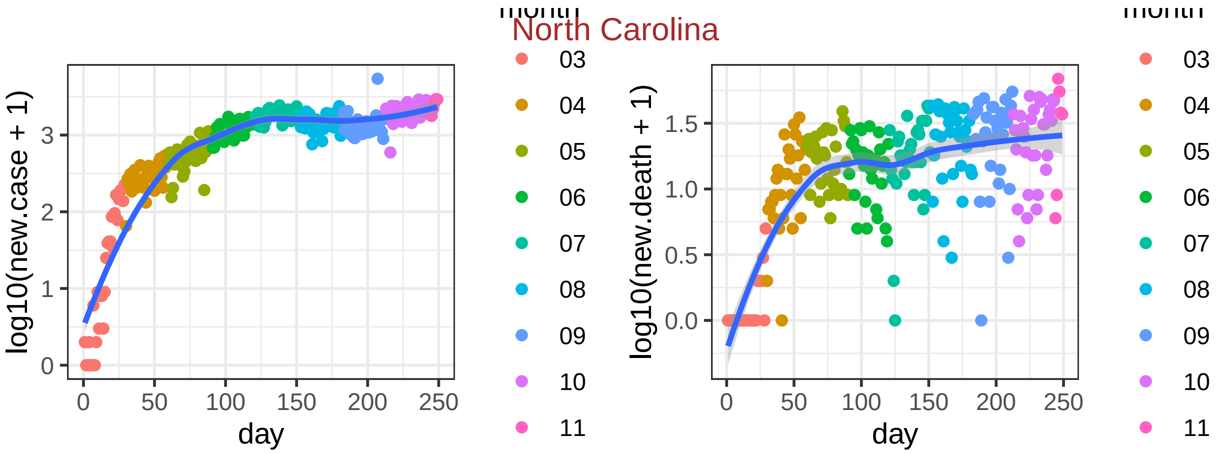
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Connecticut



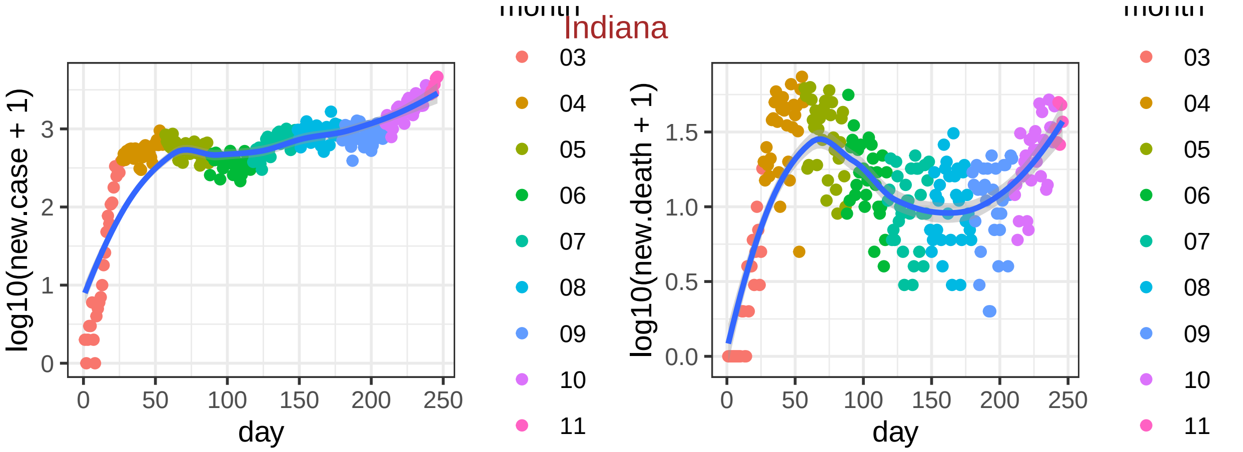
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

North Carolina



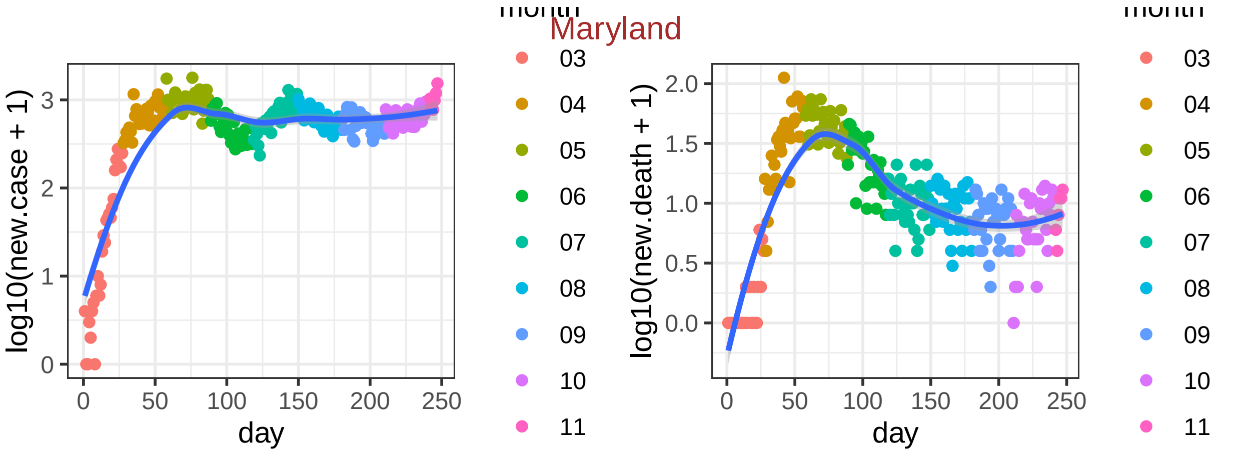
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Indiana



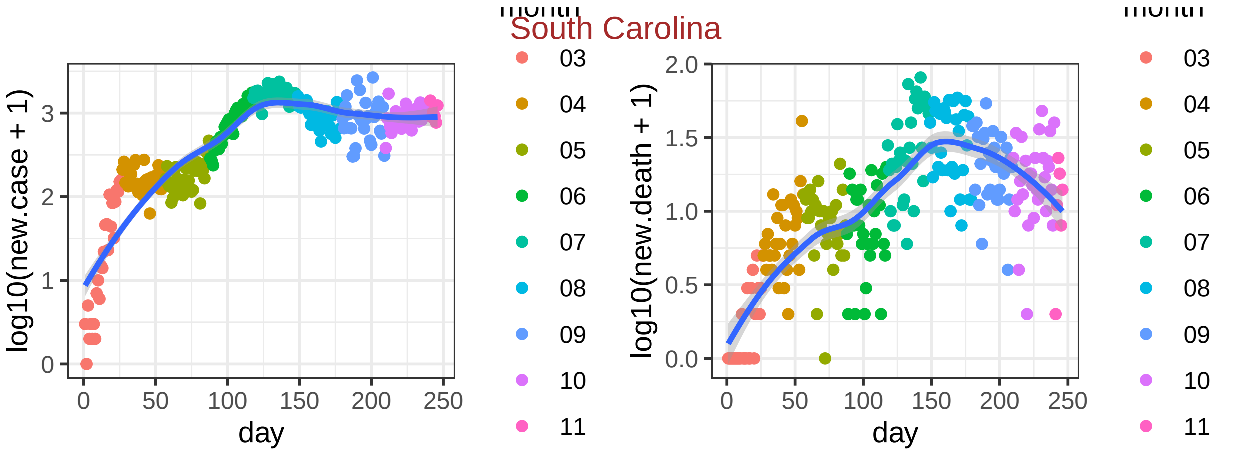
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Maryland

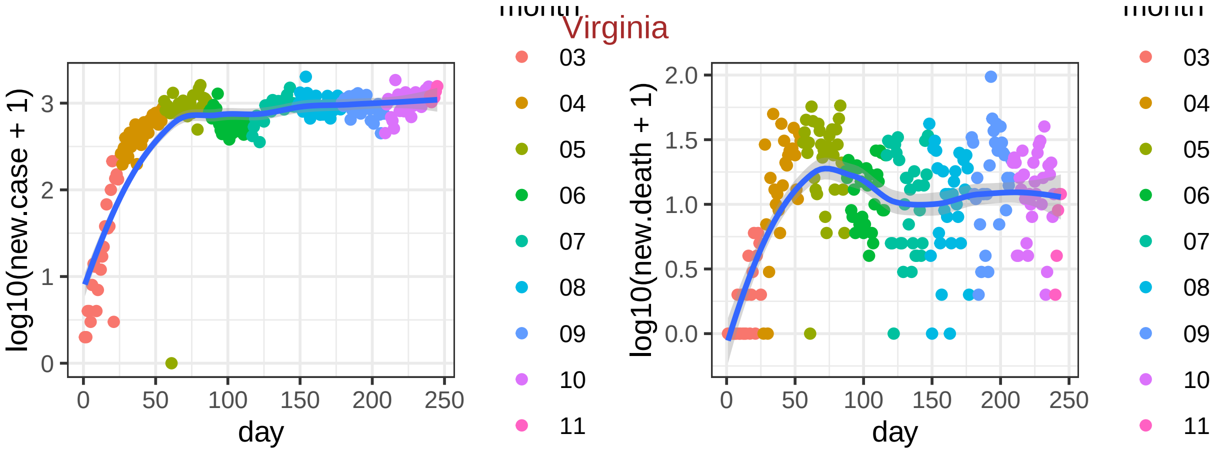


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

South Carolina

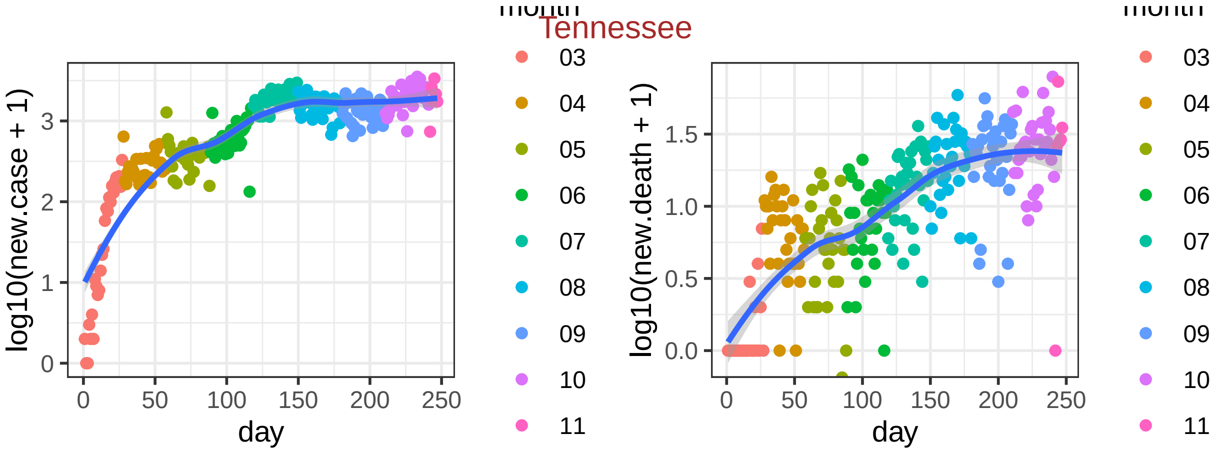


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



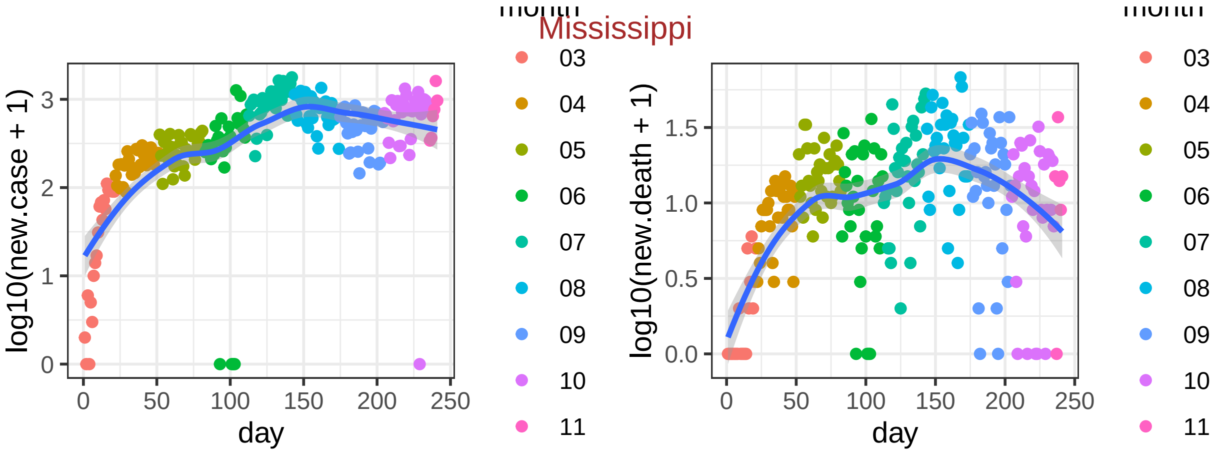
Virginia

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



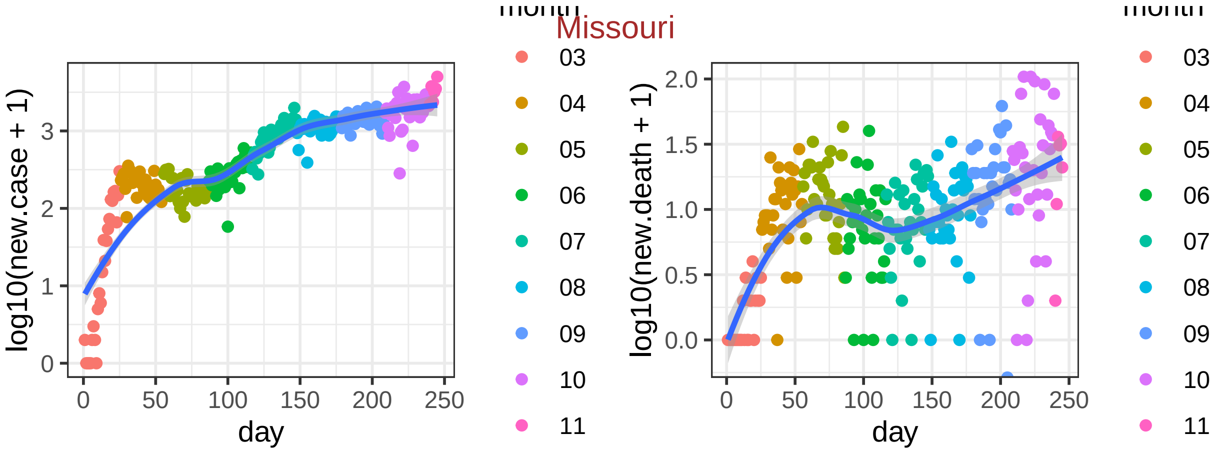
Tennessee

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

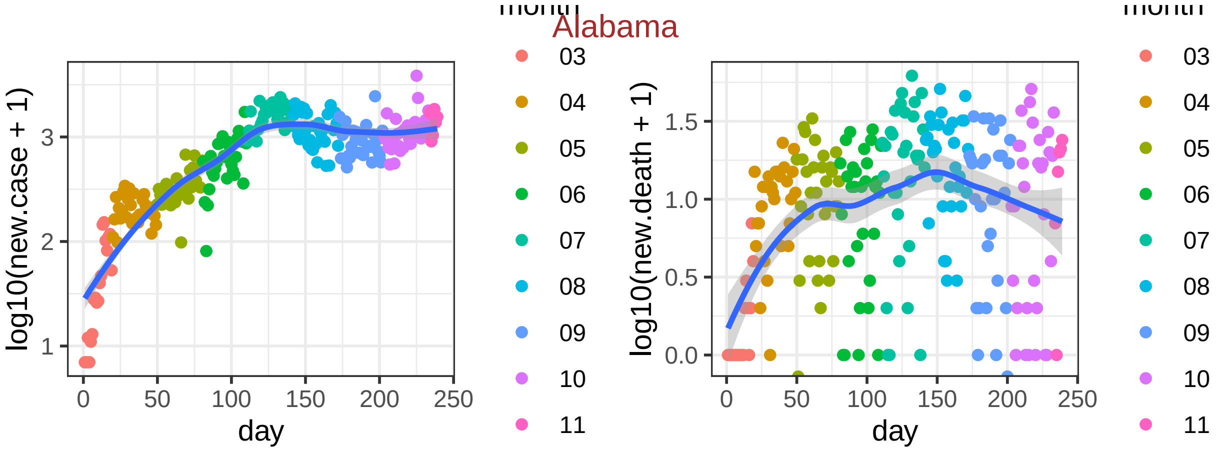


Mississippi

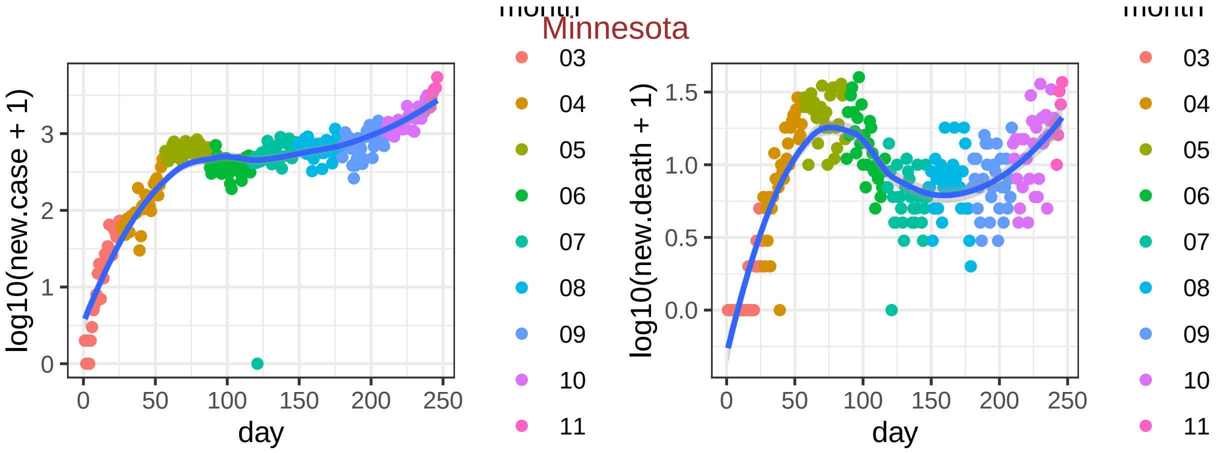
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11



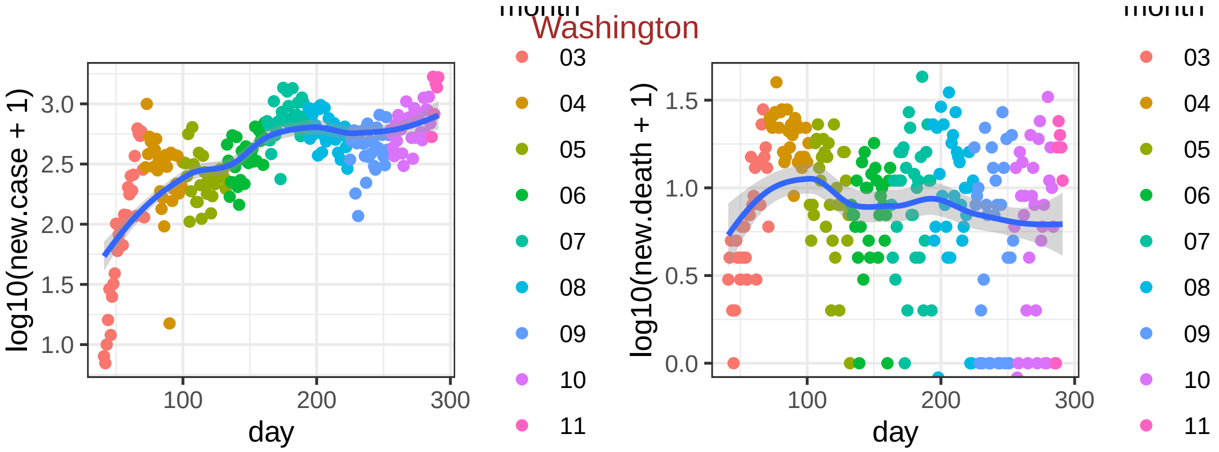
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07



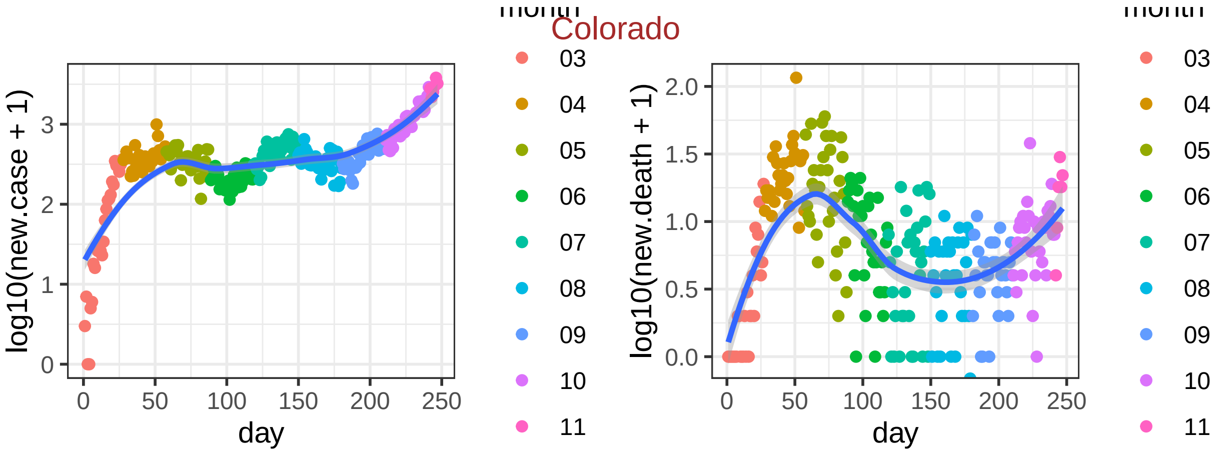
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13



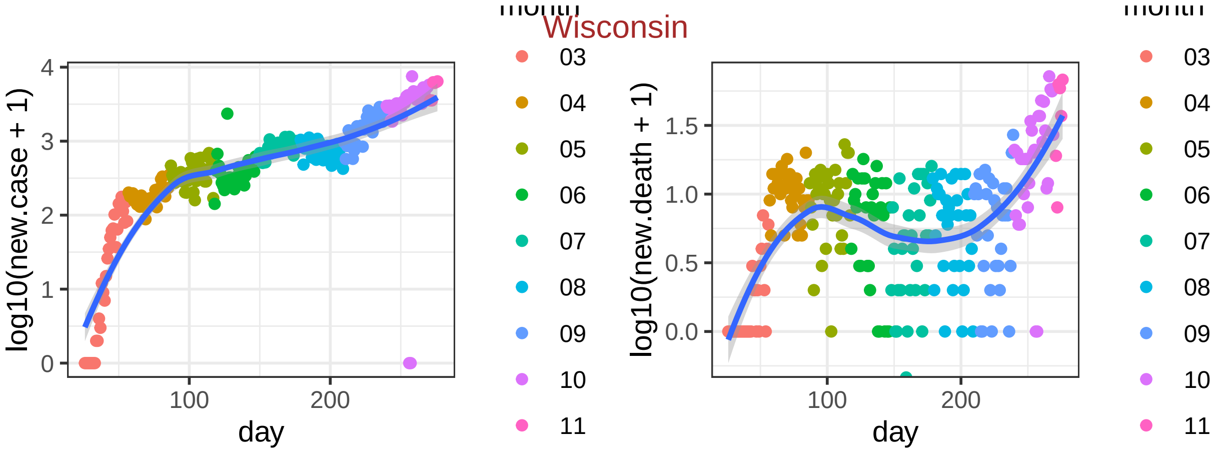
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



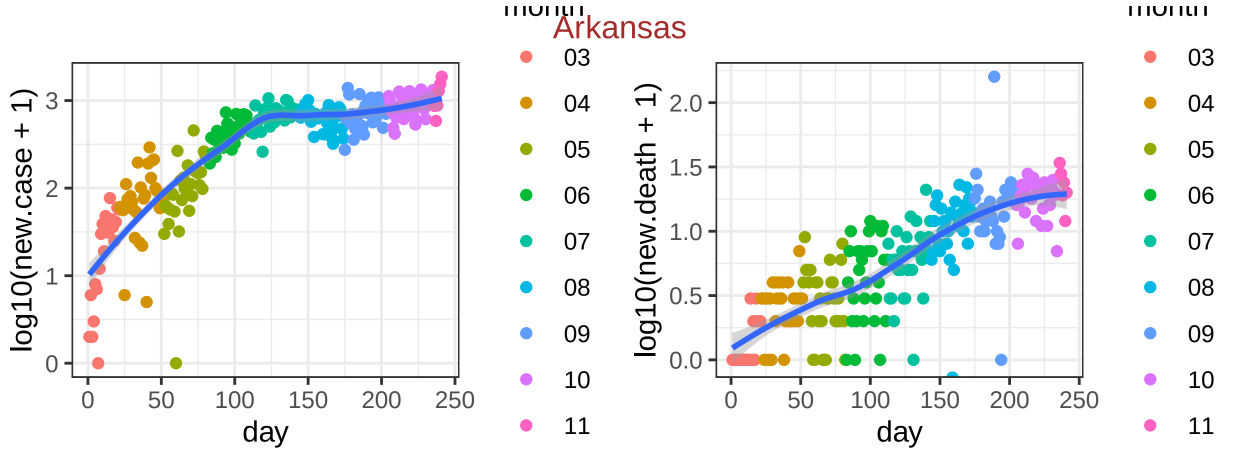
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



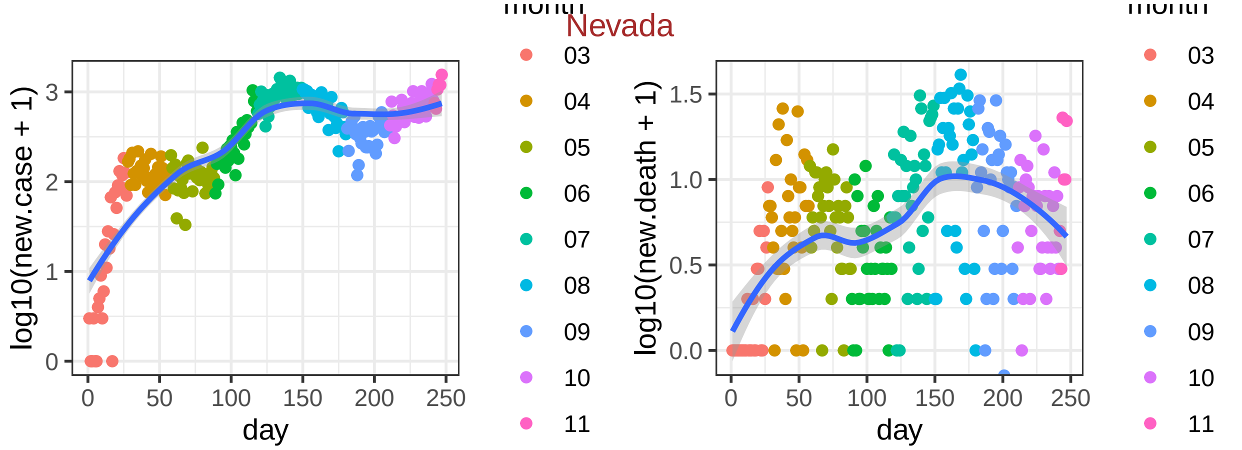
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



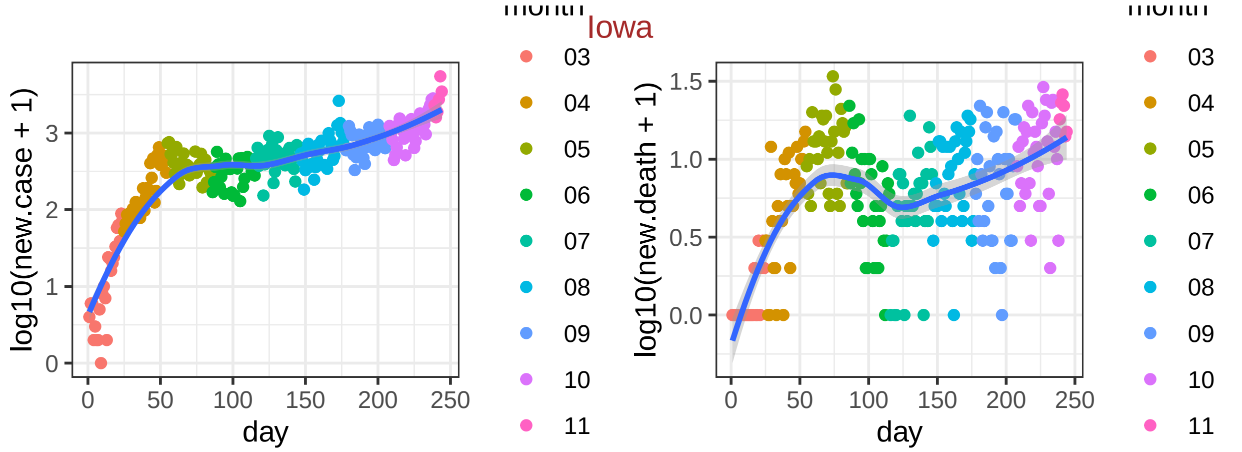
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



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

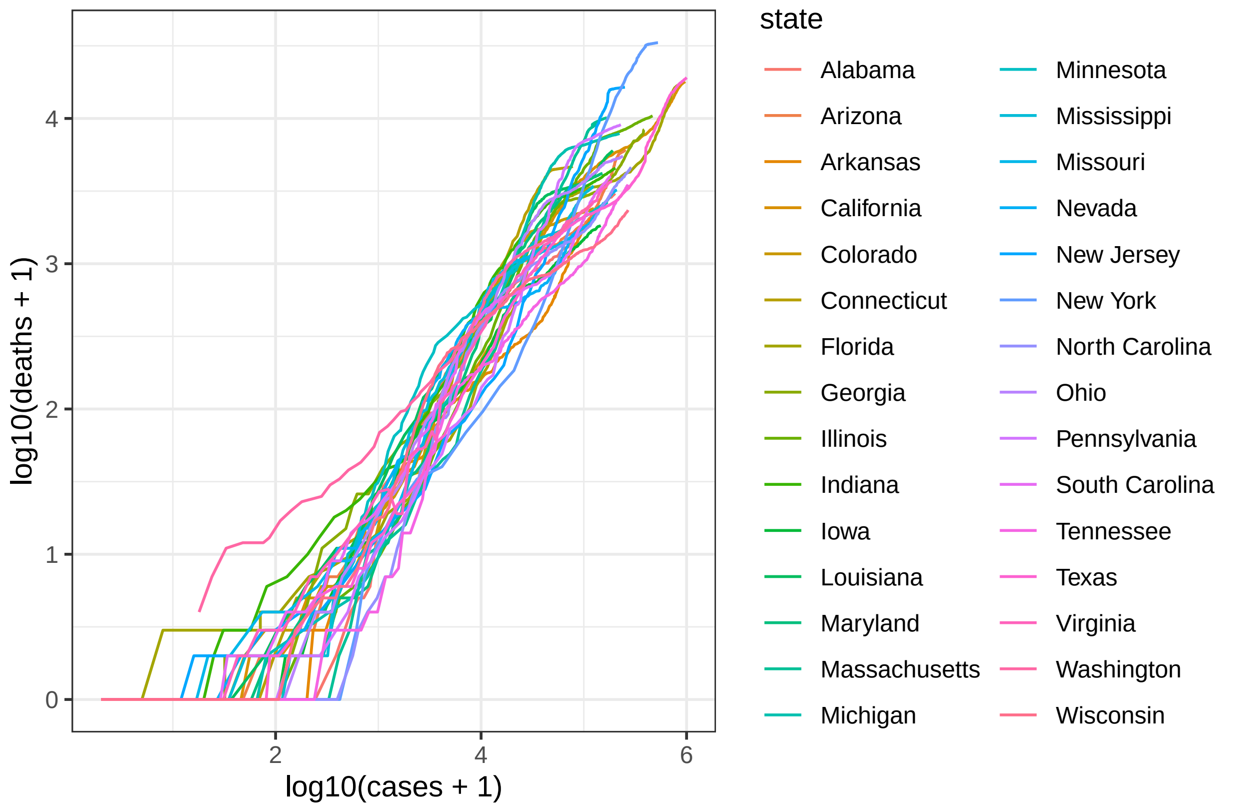


data source: https://github.com/nytimes/covid−19-data, day 1 is 03−05



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



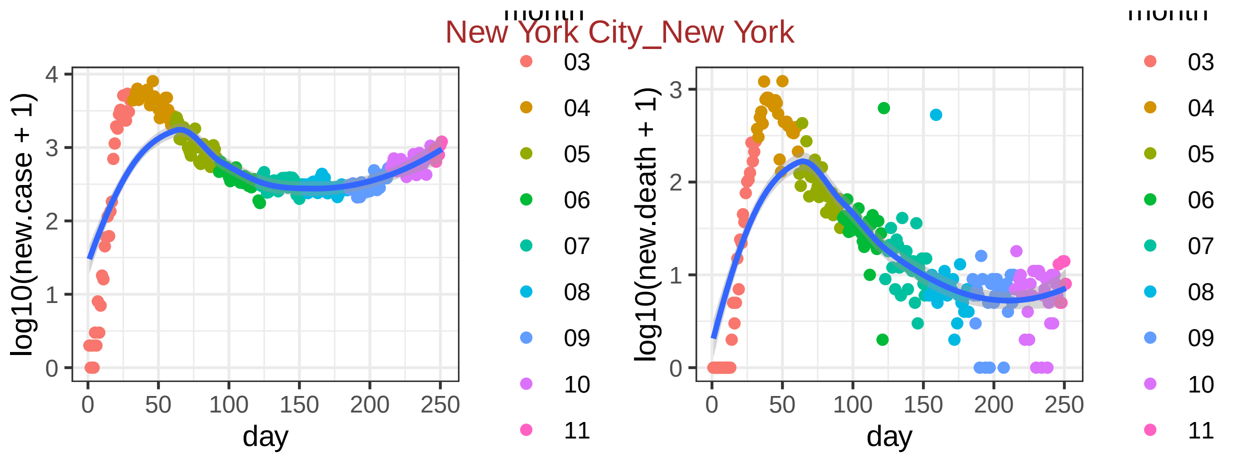
data source: https://github.com/nytimes/covid−19−data

county level data

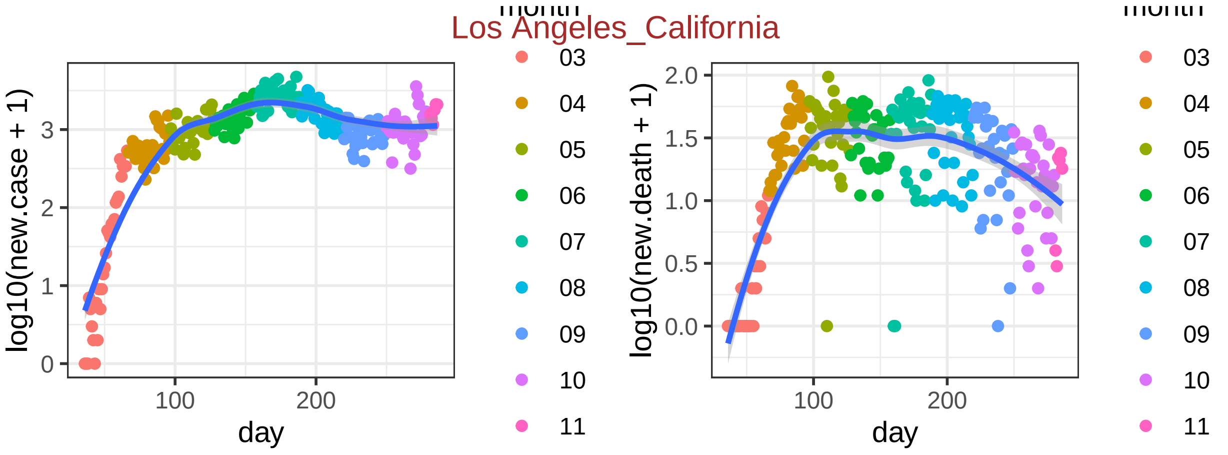
First check the 50 counties with the largest number of deaths.

```
##                  date          county         state  fips  cases deaths
## 704947 2020-11-06   New York City      New York    NA 273386  24054
## 703279 2020-11-06     Los Angeles    California  6037 317727   7157
## 703689 2020-11-06            Cook      Illinois 17031 210266   5742
## 703177 2020-11-06        Maricopa       Arizona  4013 164355   3684
## 703439 2020-11-06      Miami-Dade       Florida 12086 191837   3671
## 704398 2020-11-06           Wayne      Michigan 26163  44914   3080
## 705794 2020-11-06          Harris         Texas 48201 165967   2854
## 704309 2020-11-06       Middlesex Massachusetts 25017  34142   2310
## 704946 2020-11-06          Nassau      New York 36059  51164   2220
## 704870 2020-11-06           Essex    New Jersey 34013  26838   2152
## 704865 2020-11-06          Bergen    New Jersey 34003  26627   2070
## 704966 2020-11-06         Suffolk      New York 36103  50210   2022
## 705801 2020-11-06         Hidalgo         Texas 48215  36627   1978
## 705385 2020-11-06    Philadelphia  Pennsylvania 42101  47675   1905
## 703446 2020-11-06      Palm Beach       Florida 12099  54622   1606
## 704839 2020-11-06           Clark        Nevada 32003  86673   1568
## 703402 2020-11-06         Broward       Florida 12011  89751   1545
## 704872 2020-11-06          Hudson    New Jersey 34017  24652   1537
## 703384 2020-11-06        Hartford   Connecticut  9003  19769   1505
## 703290 2020-11-06          Orange    California  6059  63469   1503
## 704974 2020-11-06     Westchester      New York 36119  41719   1475
## 704875 2020-11-06       Middlesex    New Jersey 34023  23742   1453
## 703383 2020-11-06       Fairfield   Connecticut  9001  26435    1447
## 705708 2020-11-06           Bexar         Texas 48029  67139    1425
## 704883 2020-11-06           Union    New Jersey 34039  21772    1377
## 704305 2020-11-06           Essex Massachusetts 25009  25741    1364
## 703293 2020-11-06       Riverside    California  6065  70696    1333
## 705750 2020-11-06          Dallas         Texas 48113 108631    1307
## 704879 2020-11-06         Passaic    New Jersey 34031  22267    1265
## 704378 2020-11-06         Oakland      Michigan 26125  29154    1249
## 704313 2020-11-06         Suffolk Massachusetts 25025  30062    1191
## 704315 2020-11-06       Worcester Massachusetts 25027  17873    1189
## 703387 2020-11-06       New Haven   Connecticut  9009  19154    1137
## 704311 2020-11-06         Norfolk Massachusetts 25021  12538    1121
## 704365 2020-11-06          Macomb      Michigan 26099  22331    1115
## 703296 2020-11-06  San Bernardino    California  6071  67777    1092
## 705724 2020-11-06         Cameron         Texas 48061  24598    1092
## 704878 2020-11-06           Ocean    New Jersey 34029  17414    1076
## 704426 2020-11-06        Hennepin     Minnesota 27053  40559    1016
## 705484 2020-11-06      Providence  Rhode Island 44007  25309     970
## 703297 2020-11-06       San Diego    California  6073  59179     907
## 705380 2020-11-06      Montgomery  Pennsylvania 42091  15118     898
## 704672 2020-11-06       St. Louis      Missouri 29189  34593     895
## 704291 2020-11-06      Montgomery      Maryland 24031  27037     886
## 704876 2020-11-06        Monmouth    New Jersey 34025  15036     874
## 705912 2020-11-06         Tarrant         Texas 48439  72118     871
## 704292 2020-11-06 Prince George's      Maryland 24033  34406     869
## 703824 2020-11-06          Marion       Indiana 18097  29621     854
## 704312 2020-11-06        Plymouth Massachusetts 25023  12022     848
## 704877 2020-11-06          Morris    New Jersey 34027  10014     841
```

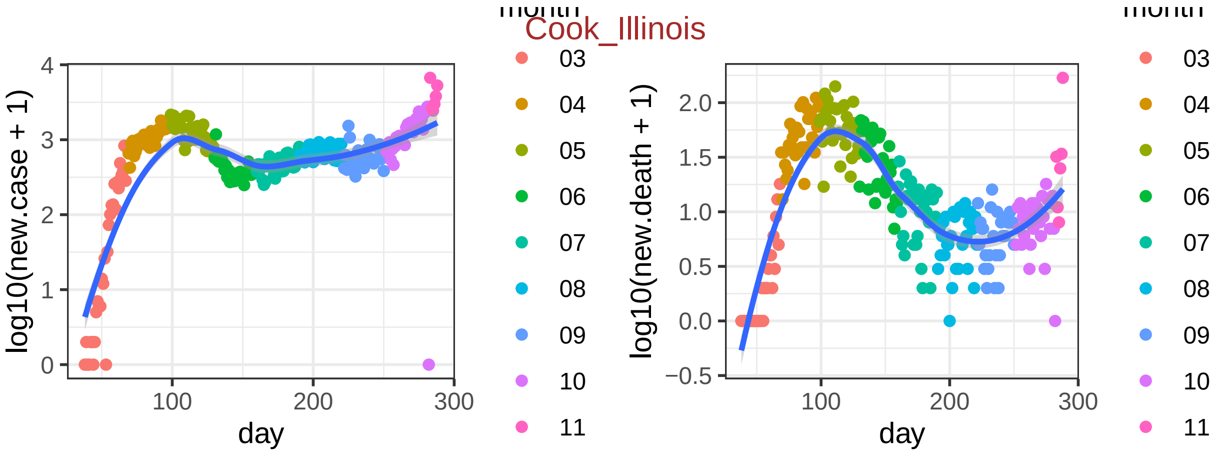
For these 50 counties, I check the number of new cases and the number of new deaths.



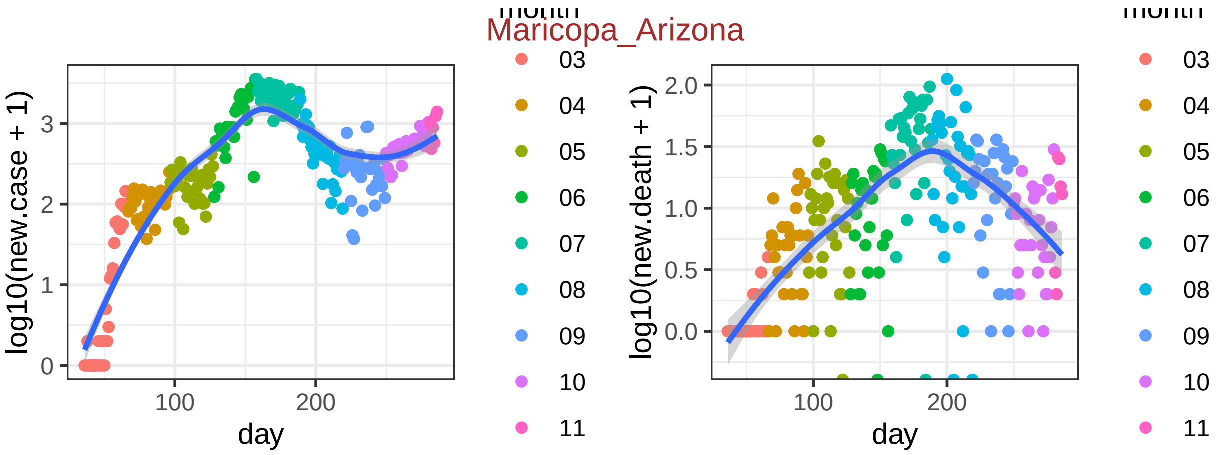
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



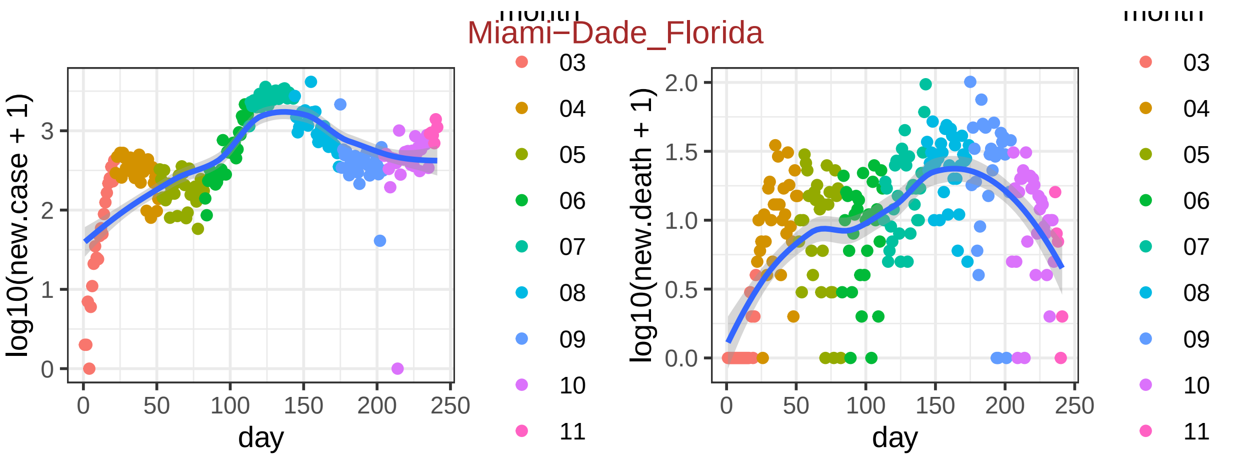
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



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

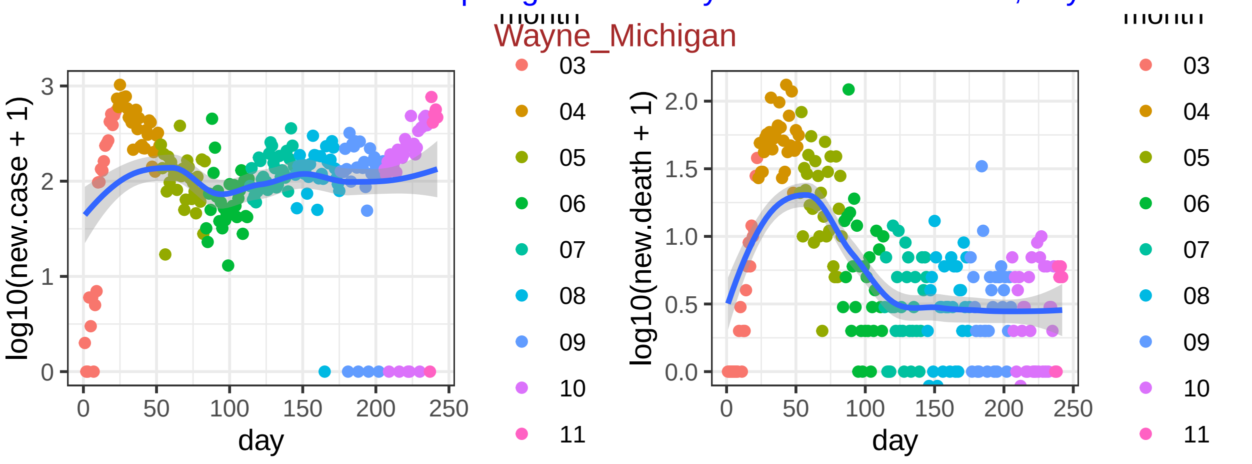


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



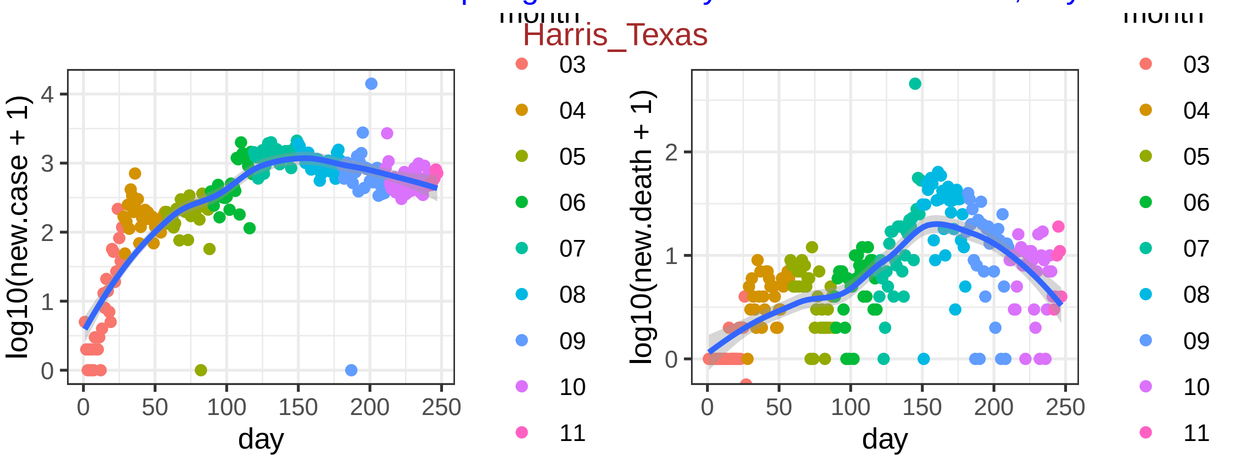
Miami−Dade_Florida

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



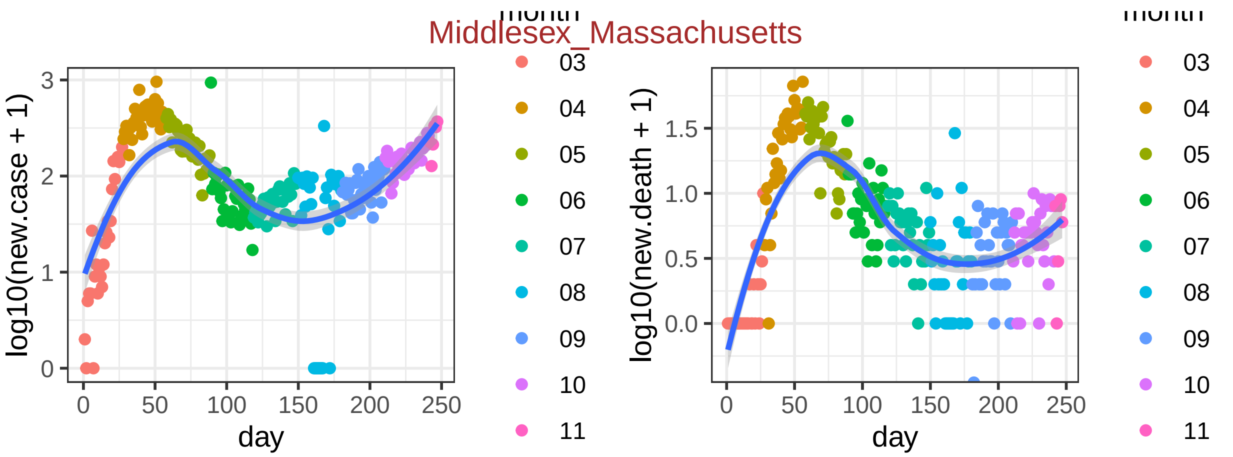
Wayne_Michigan

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

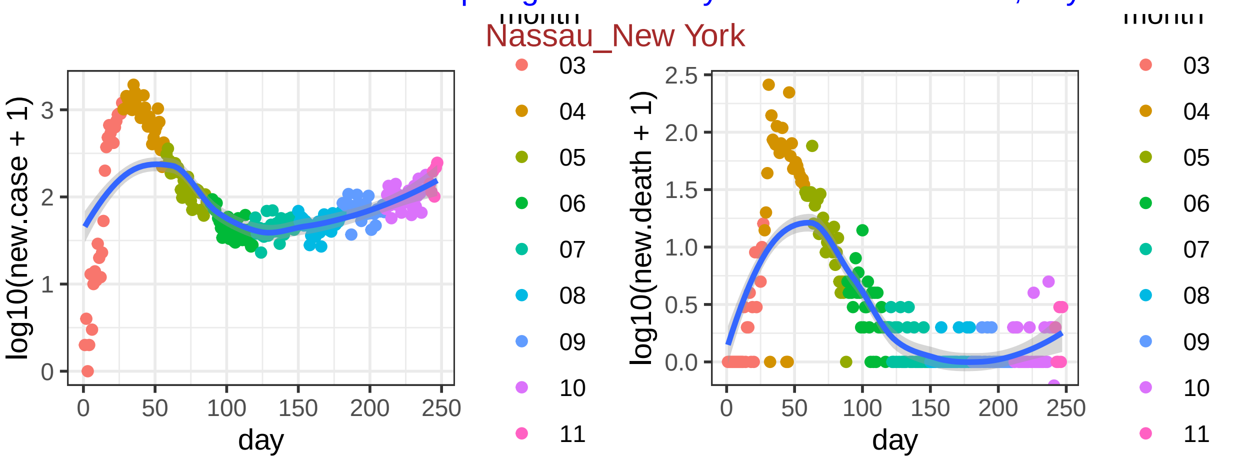


Harris_Texas

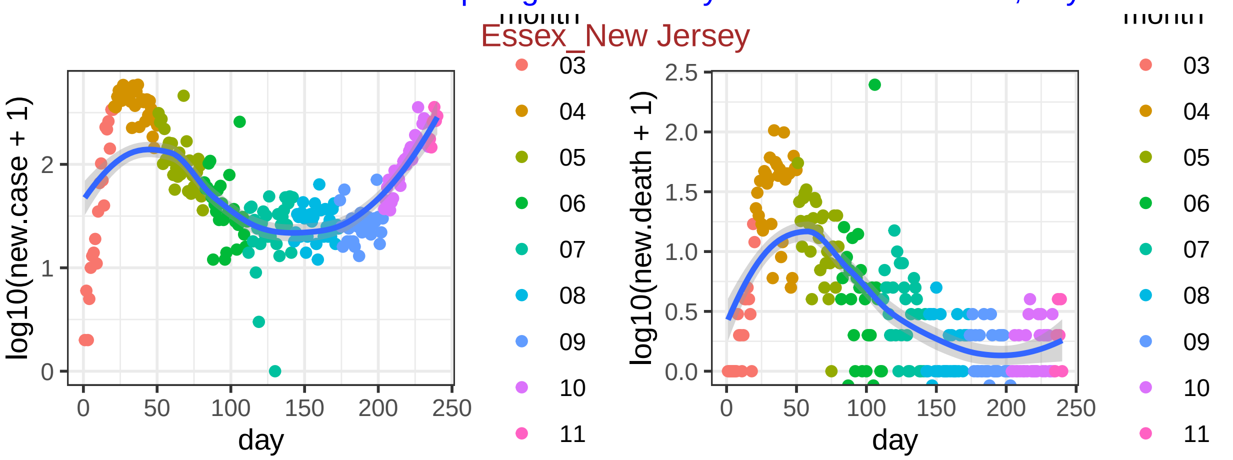
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05


Middlesex_Massachusetts

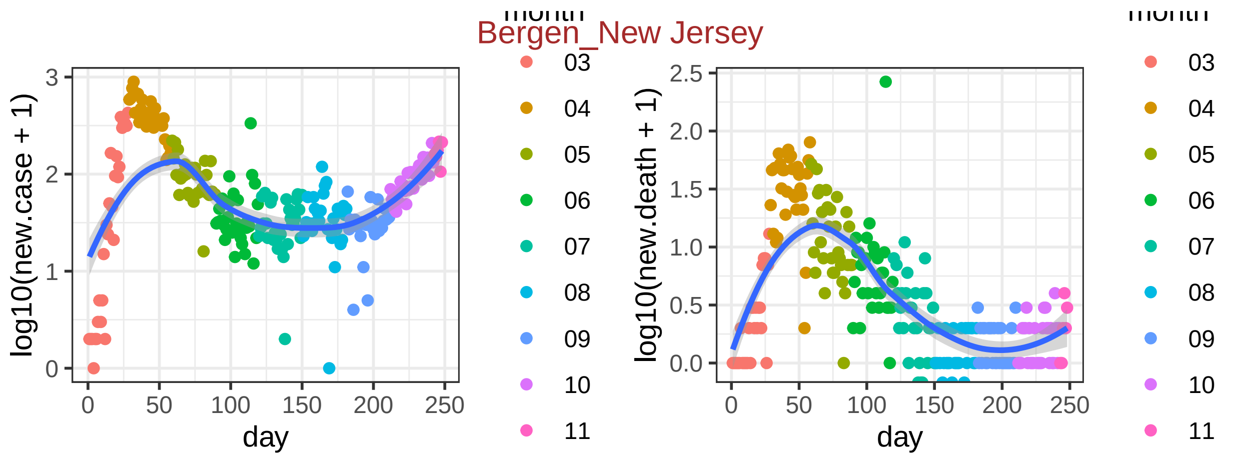
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05


Nassau_New York

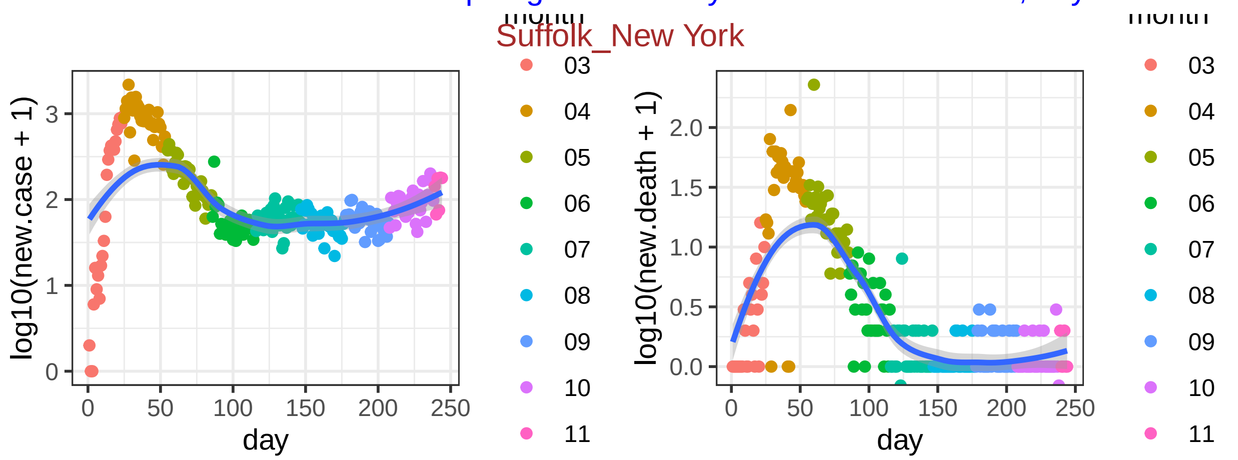
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05


Essex_New Jersey

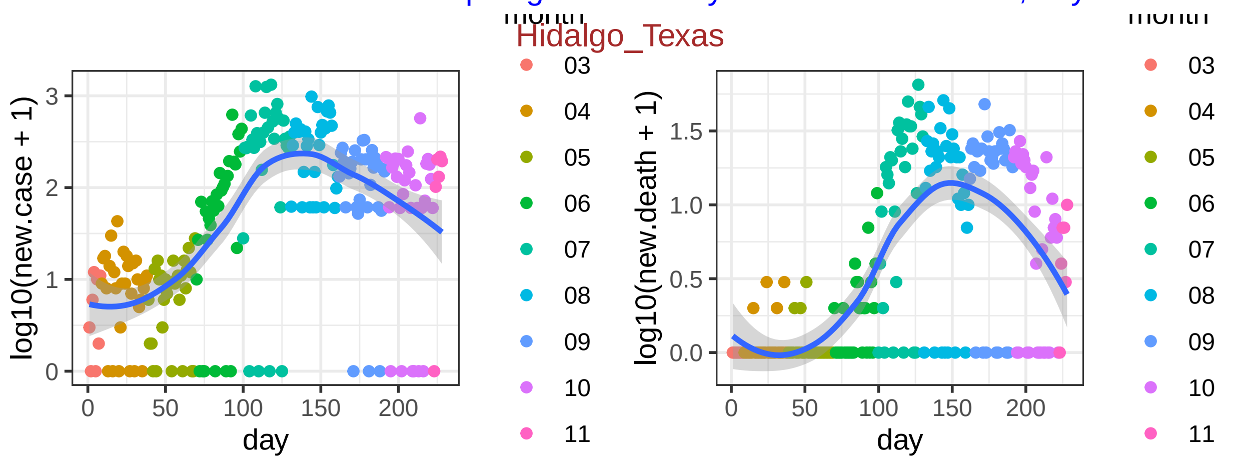
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12



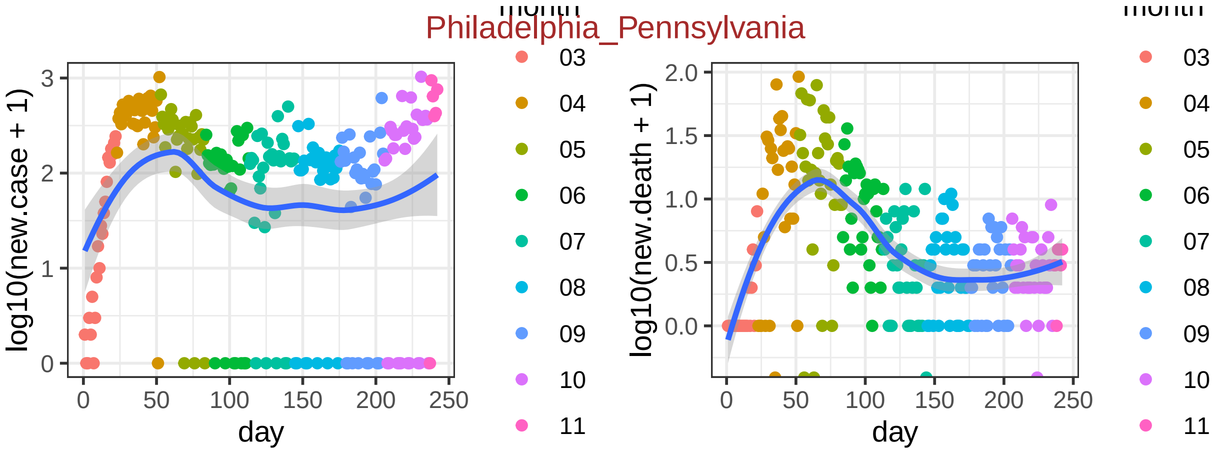
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04



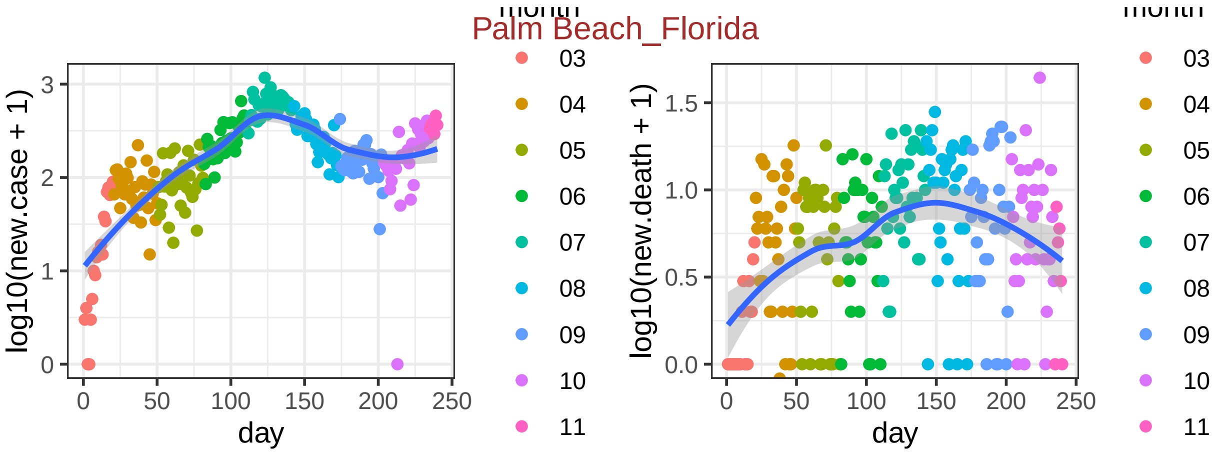
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08



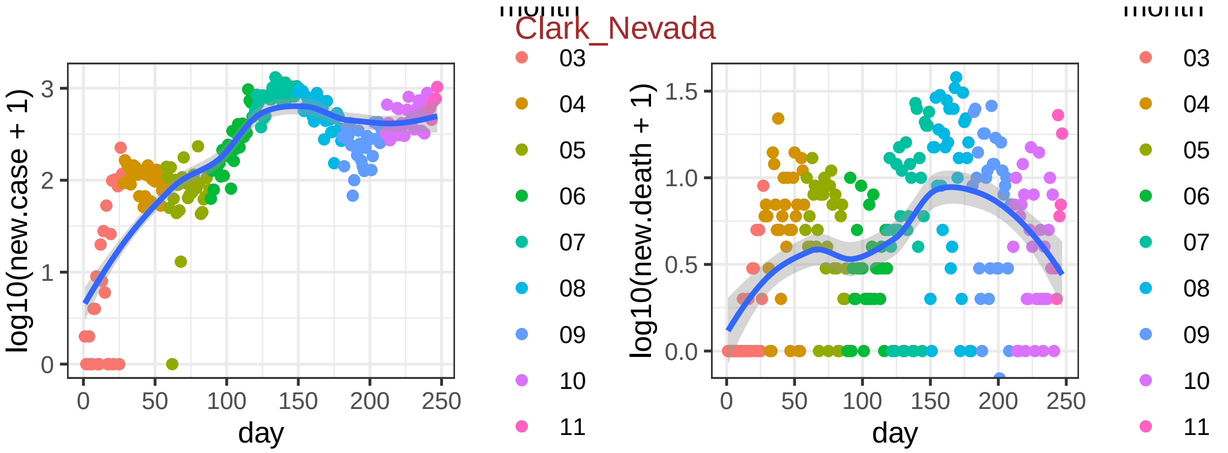
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−24


Philadelphia_Pennsylvania

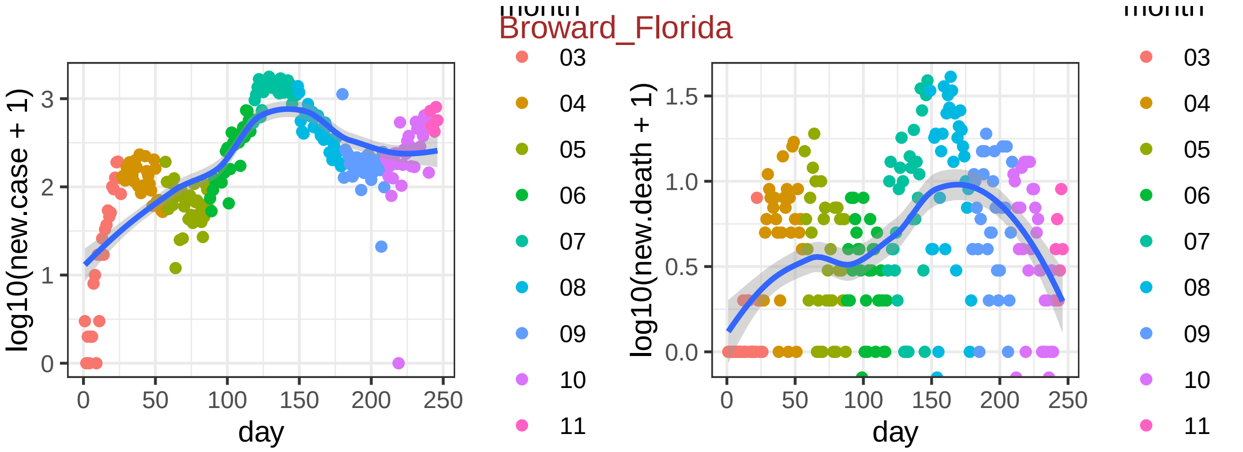
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10


Palm Beach_Florida

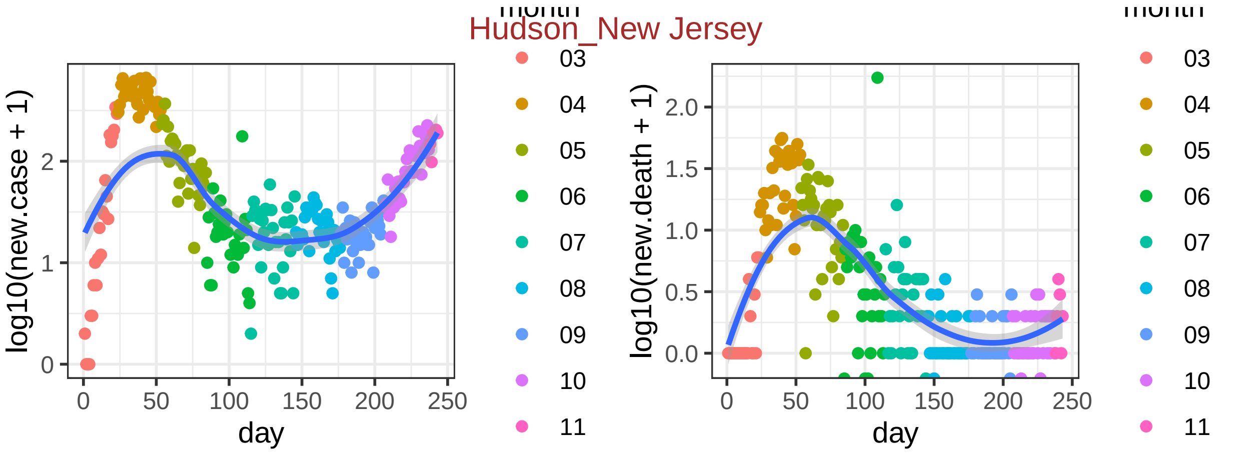
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12


Clark_Nevada

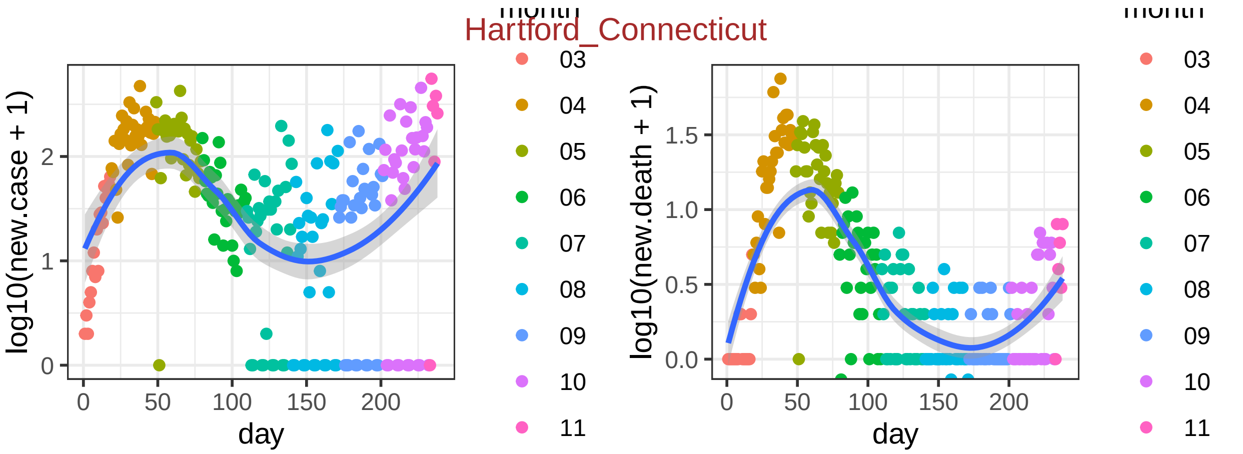
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05


Broward_Florida

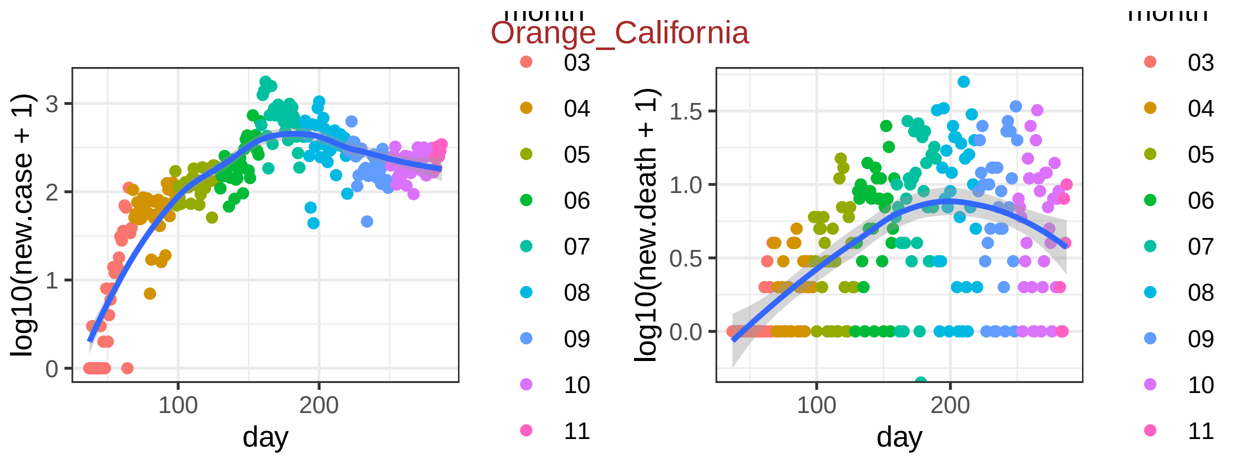
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06


Hudson_New Jersey

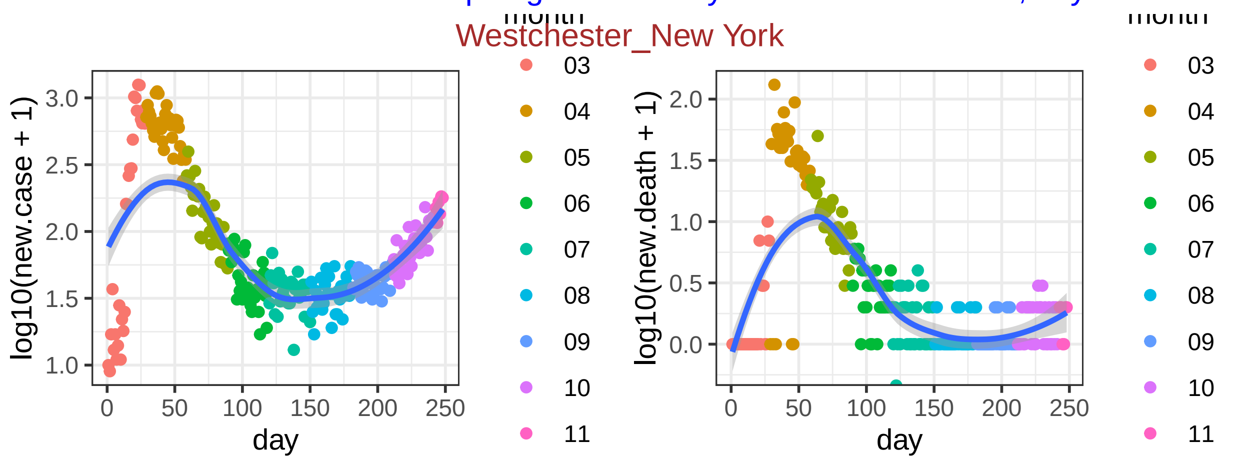
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09


Hartford_Connecticut

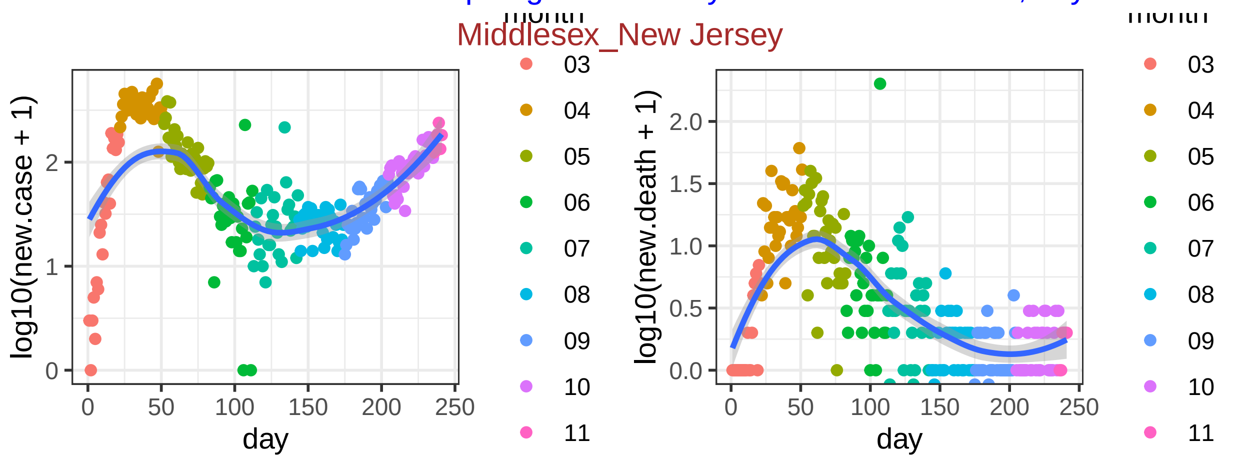
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14


Orange_California

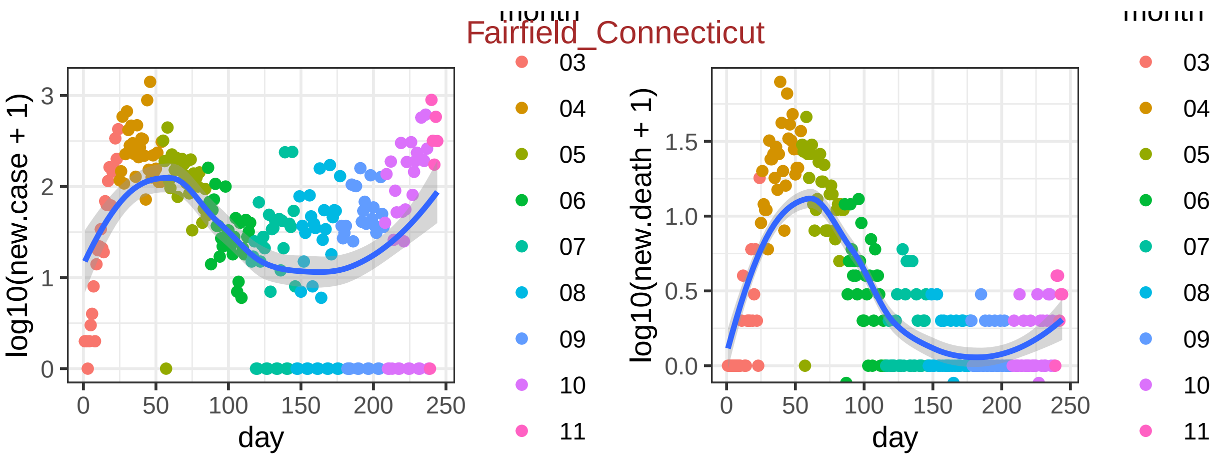
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01


Westchester_New York

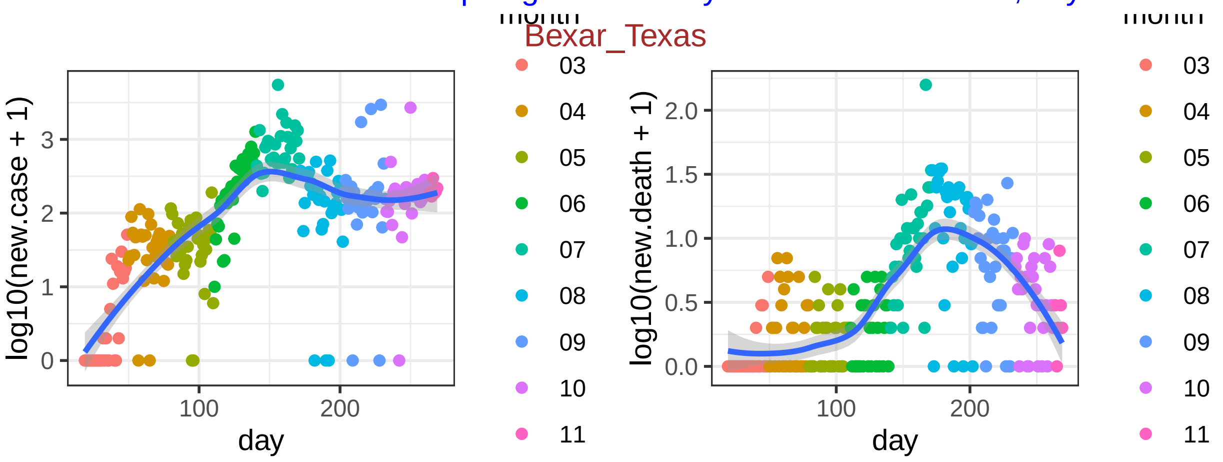
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04


Middlesex_New Jersey

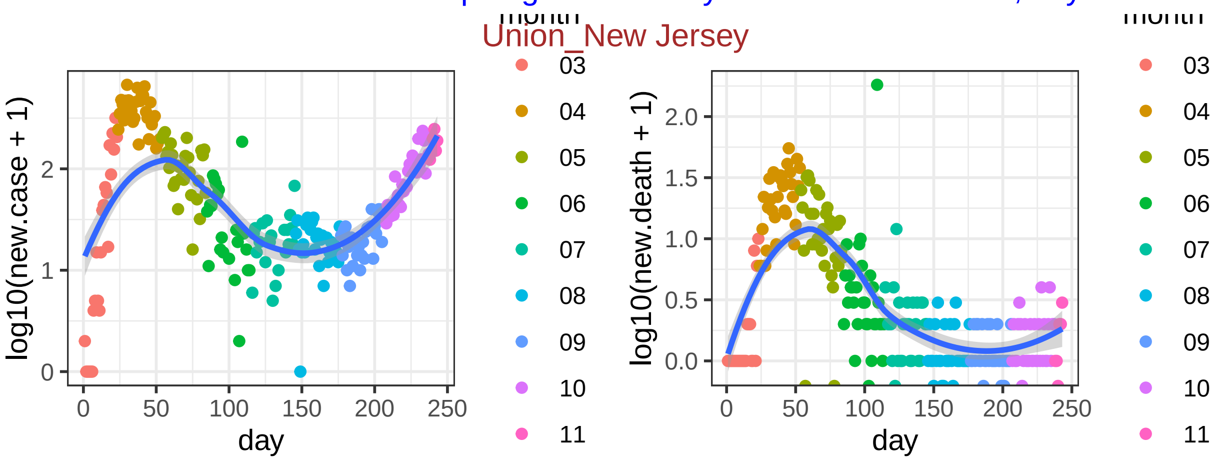
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11



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

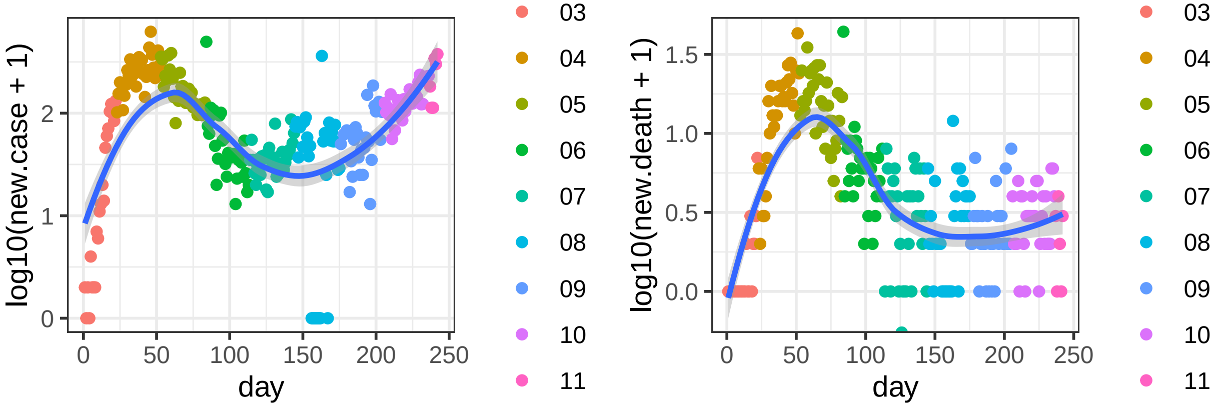


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



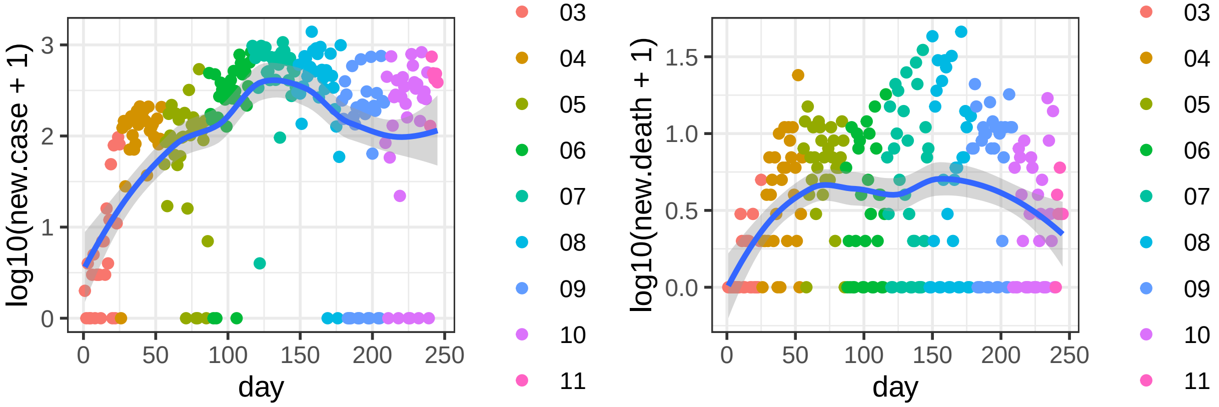
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Essex_Massachusetts



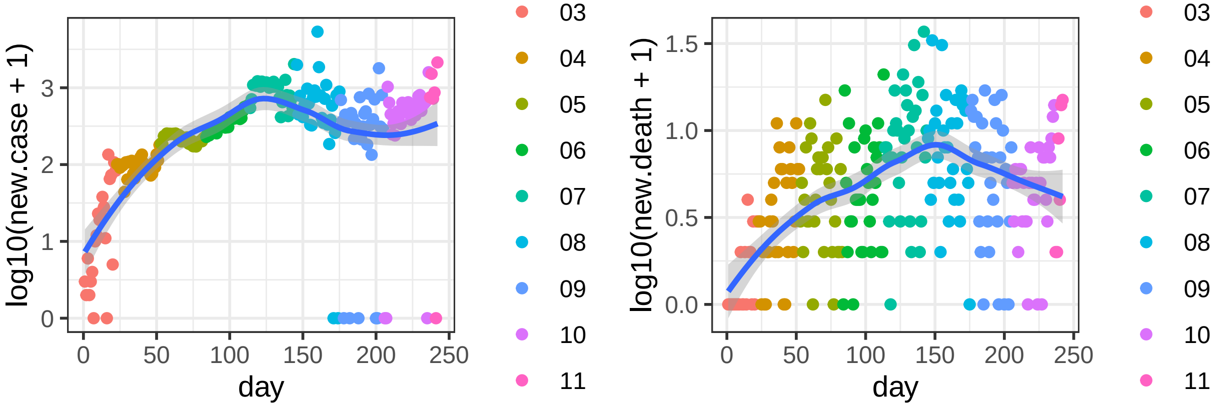
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Riverside_California

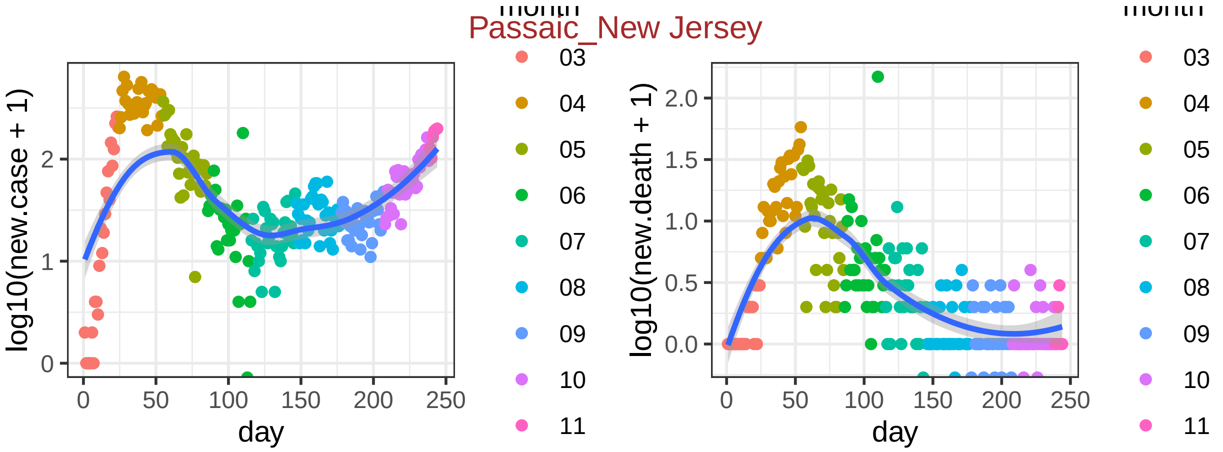


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

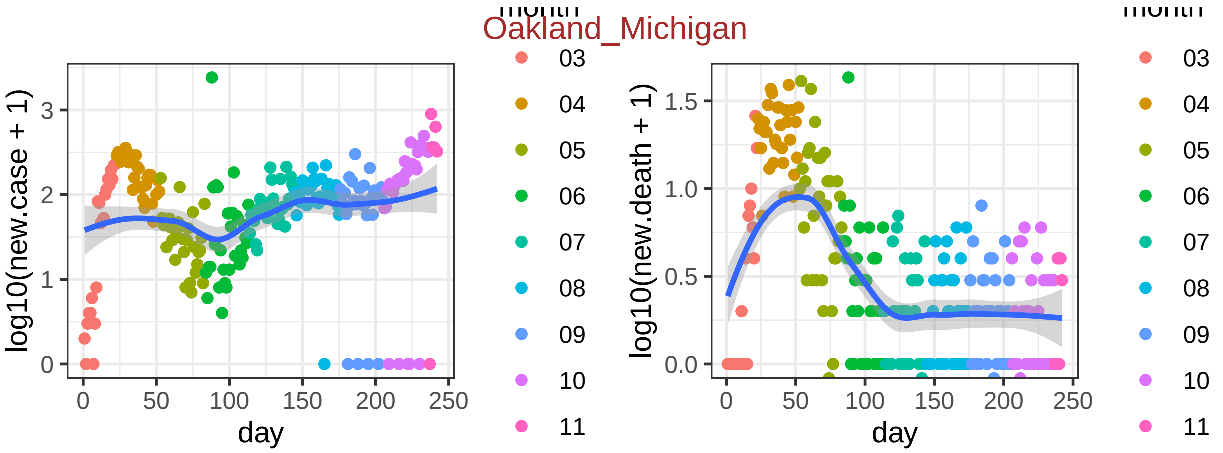
Dallas_Texas



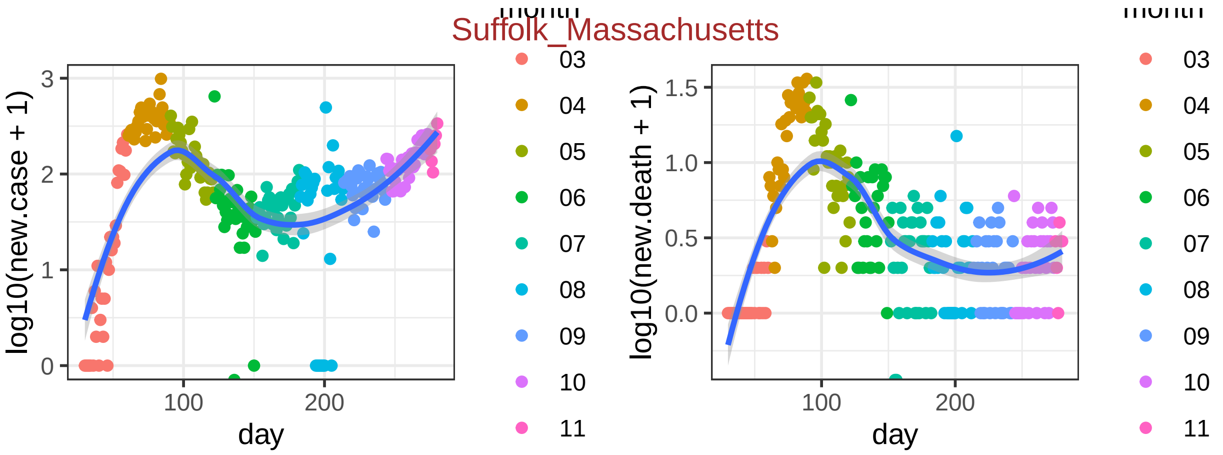
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10



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

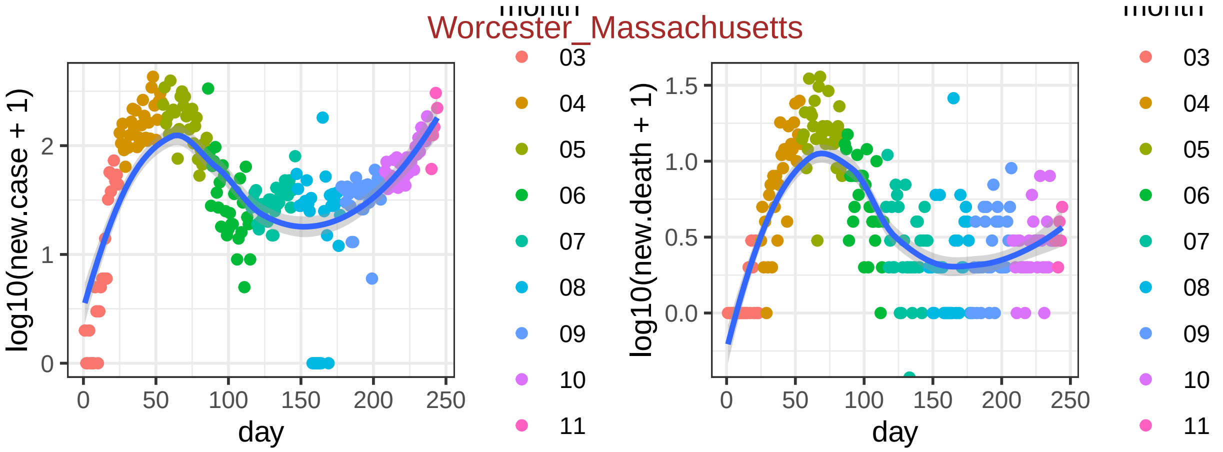


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10



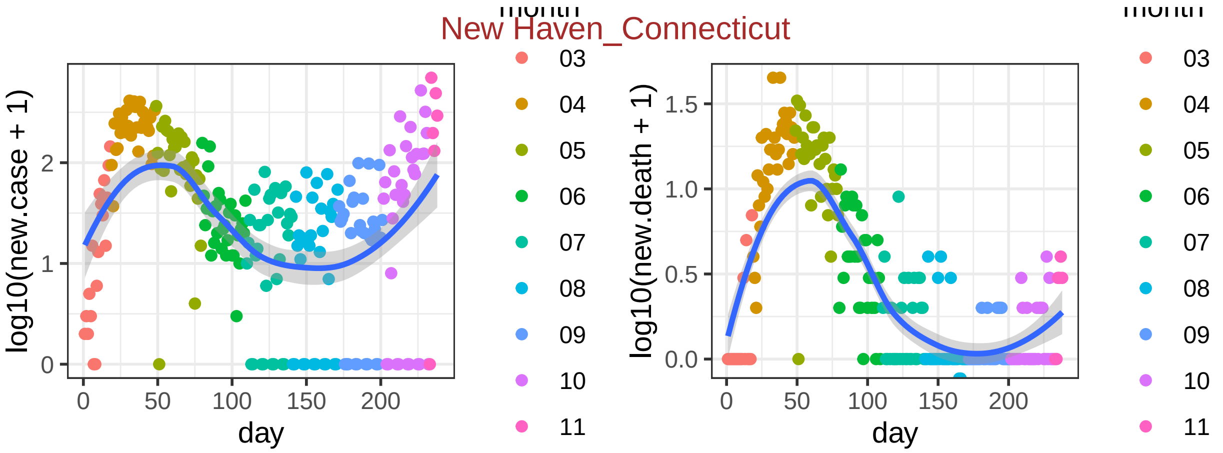
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Worcester_Massachusetts



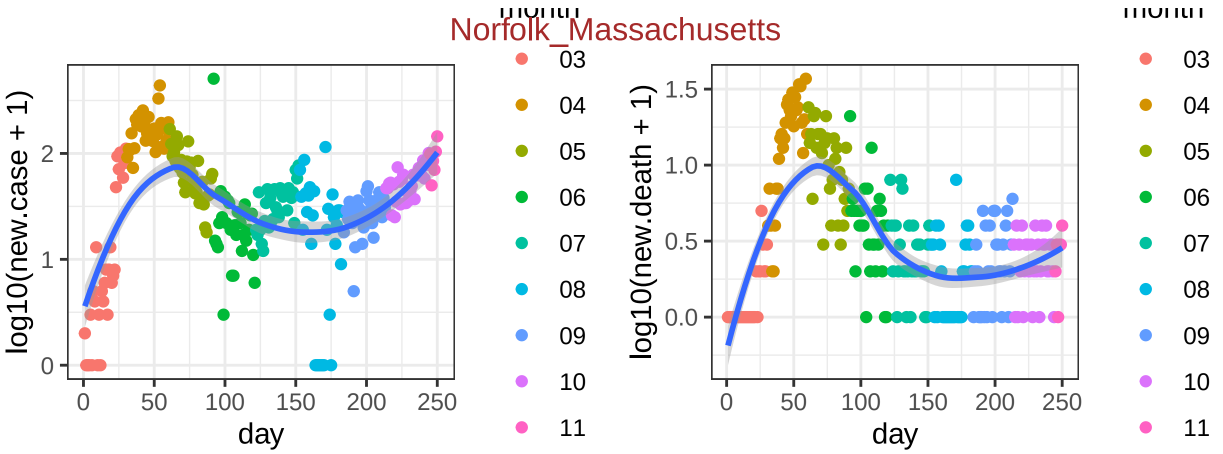
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

New Haven_Connecticut

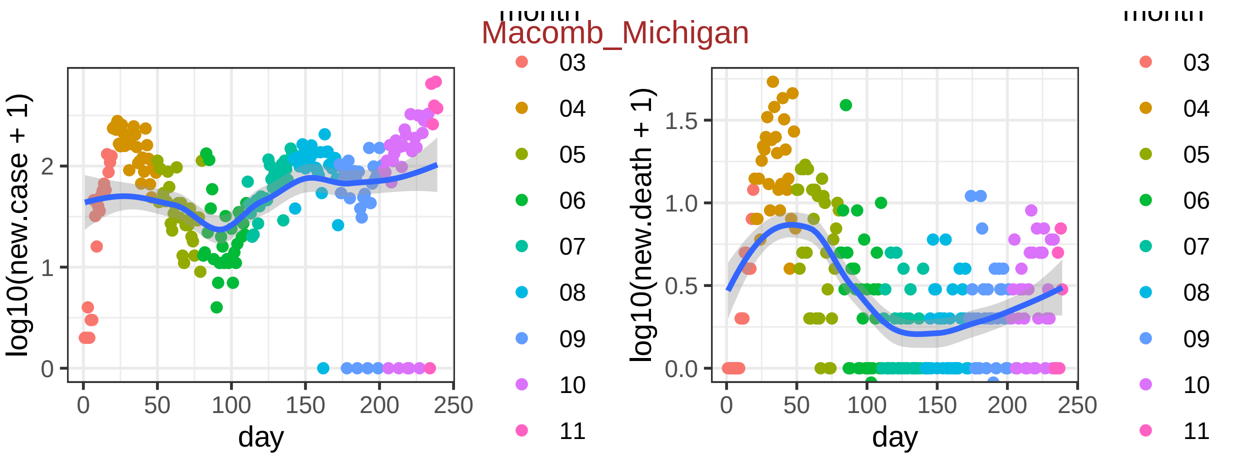


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

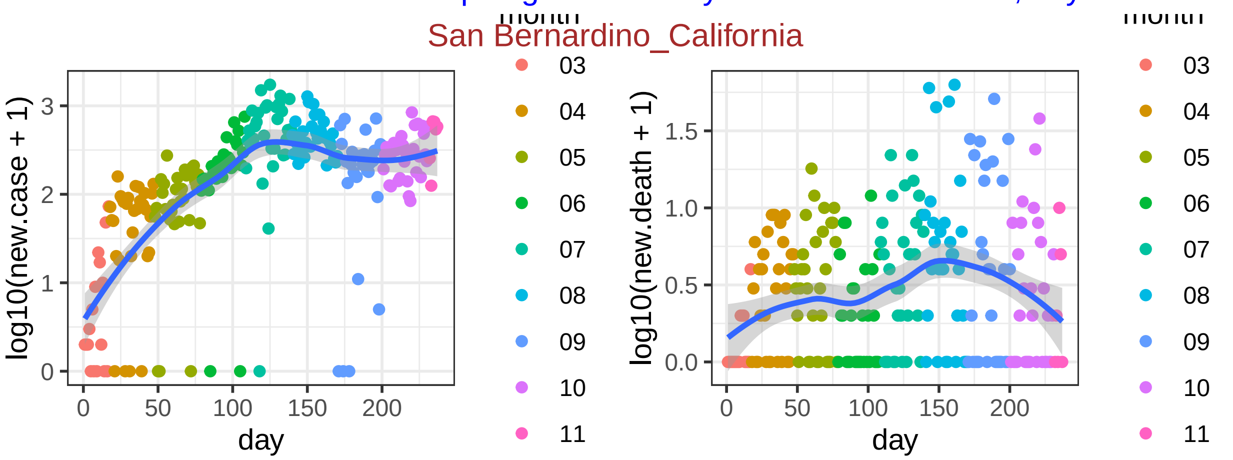
Norfolk_Massachusetts



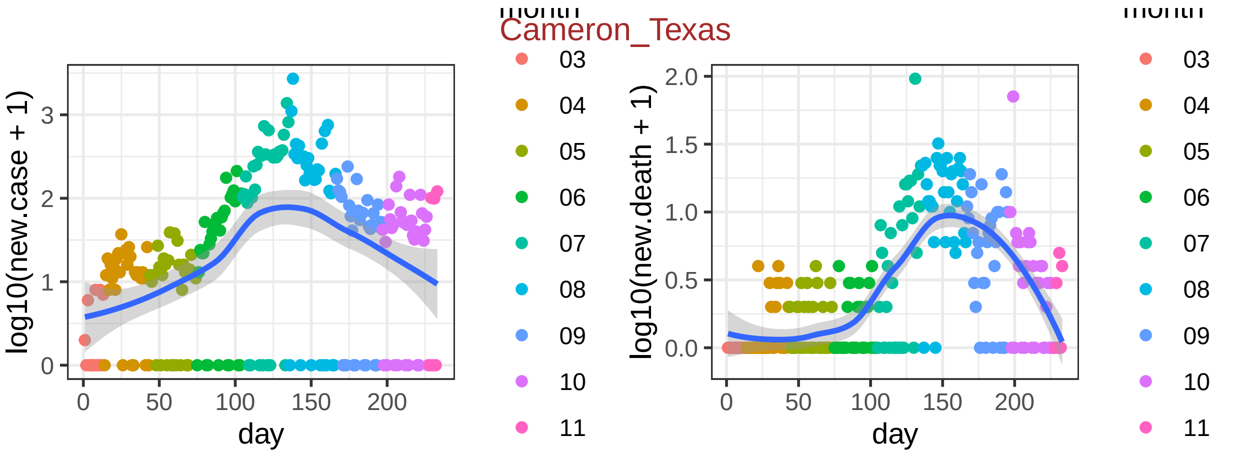
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02


Macomb_Michigan

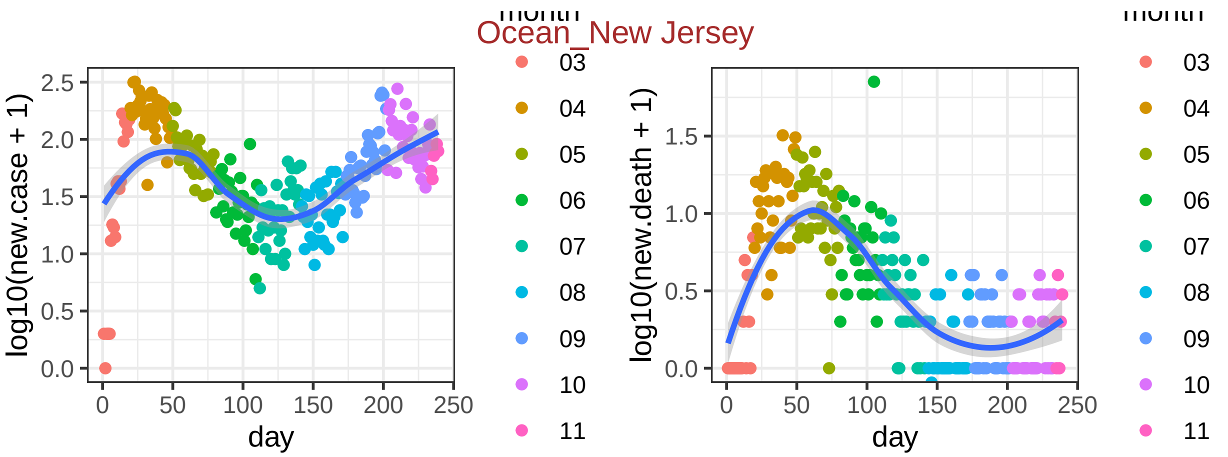
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13


San Bernardino_California

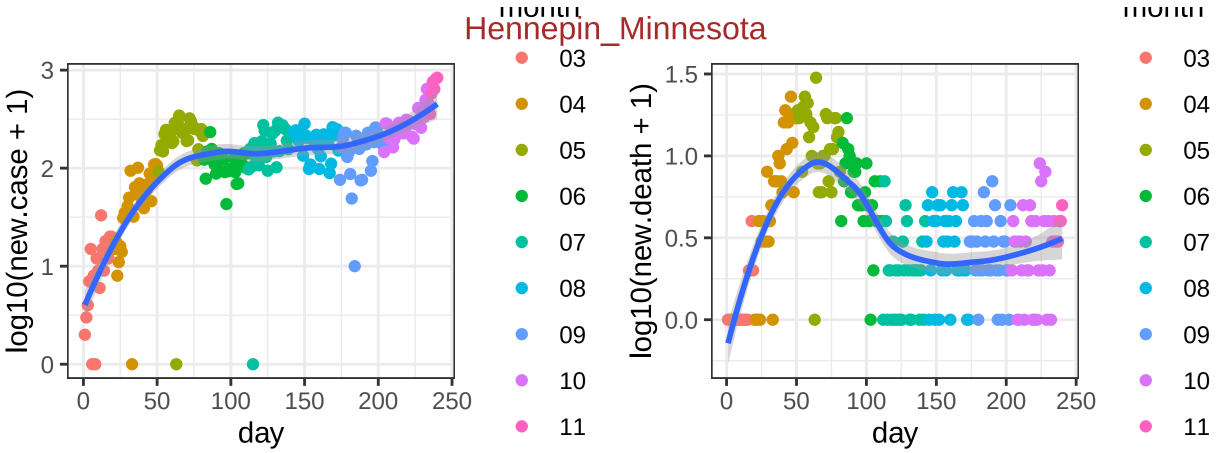
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15


Cameron_Texas

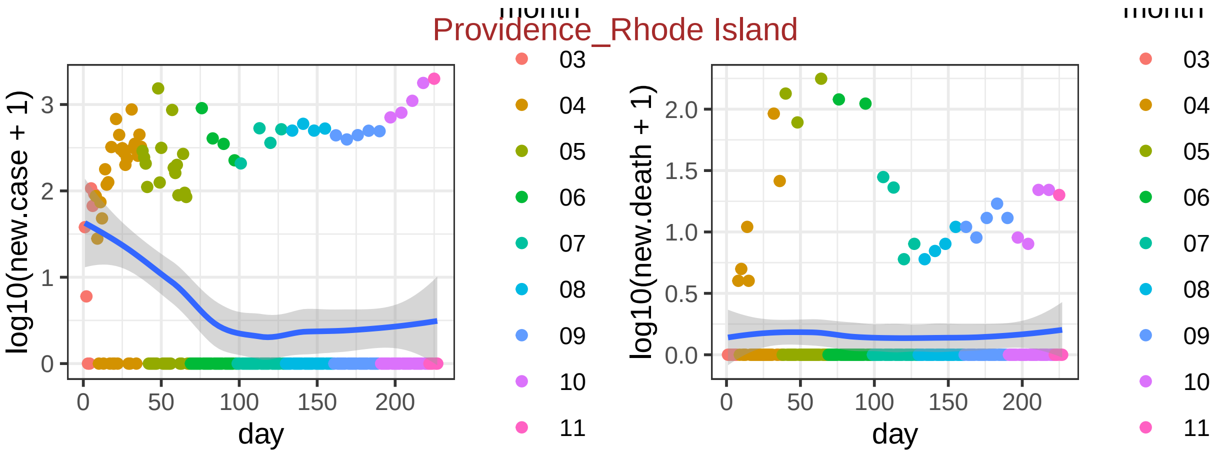
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−19



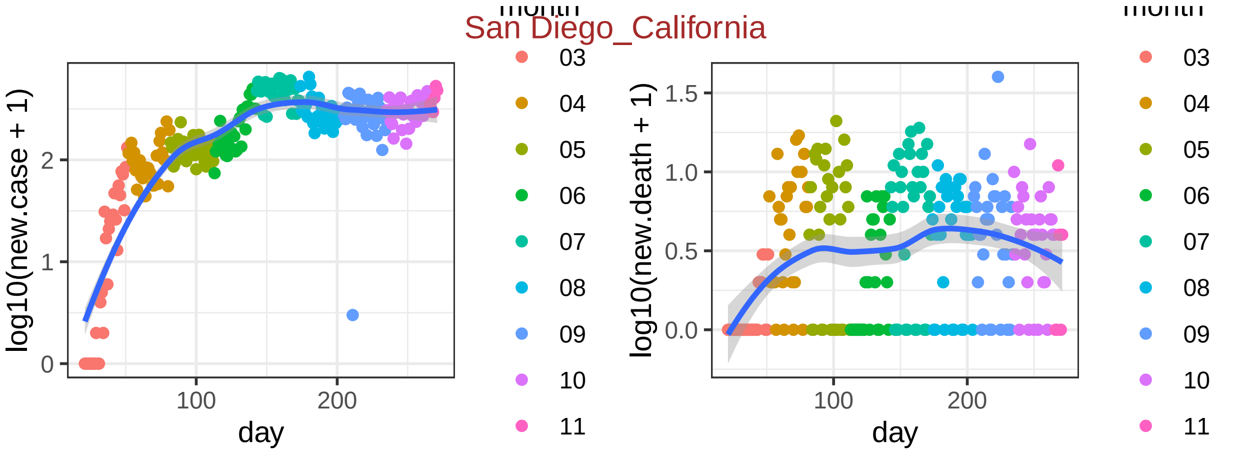
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13



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

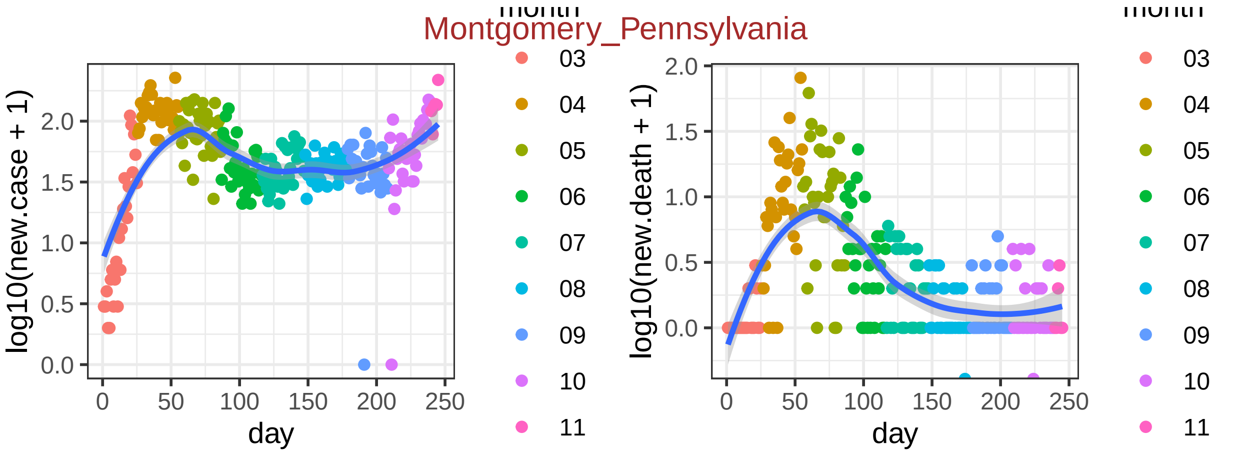


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25



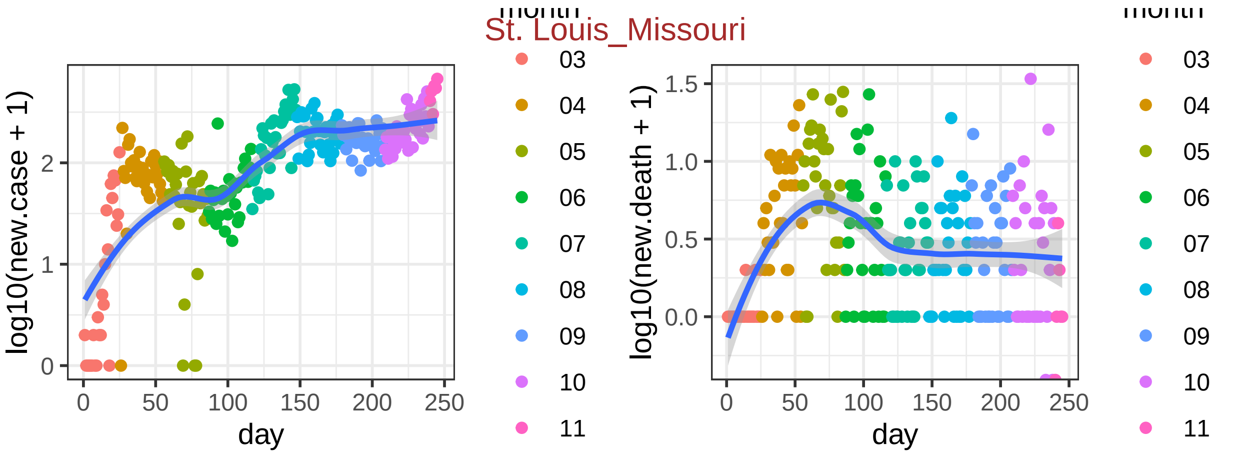
San Diego_California

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



Montgomery_Pennsylvania

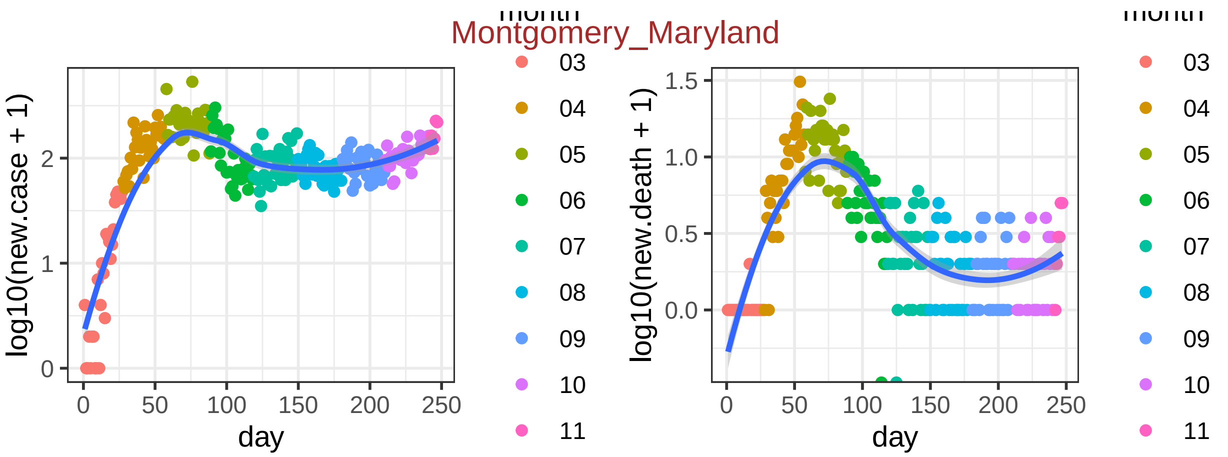
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07



St. Louis_Missouri

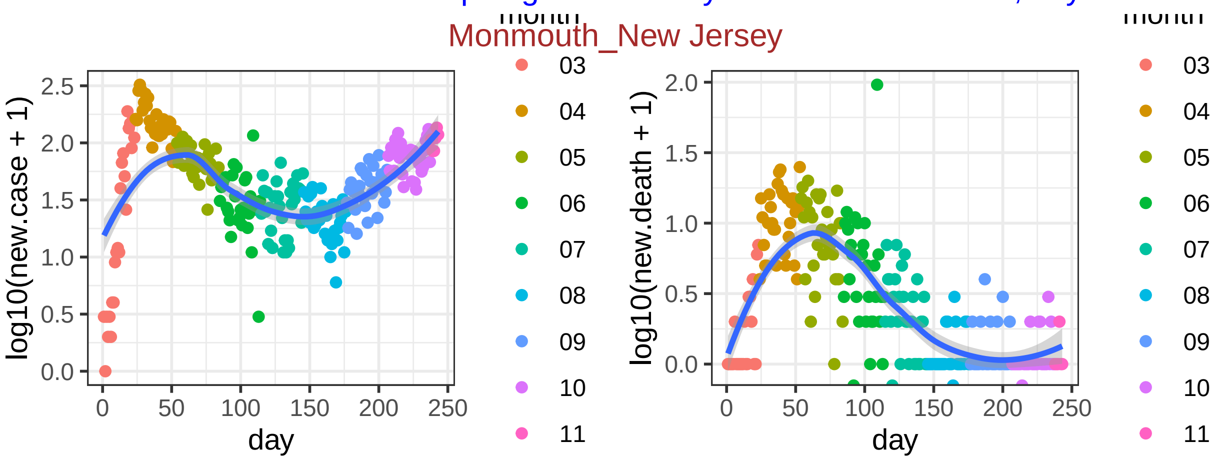
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Montgomery_Maryland



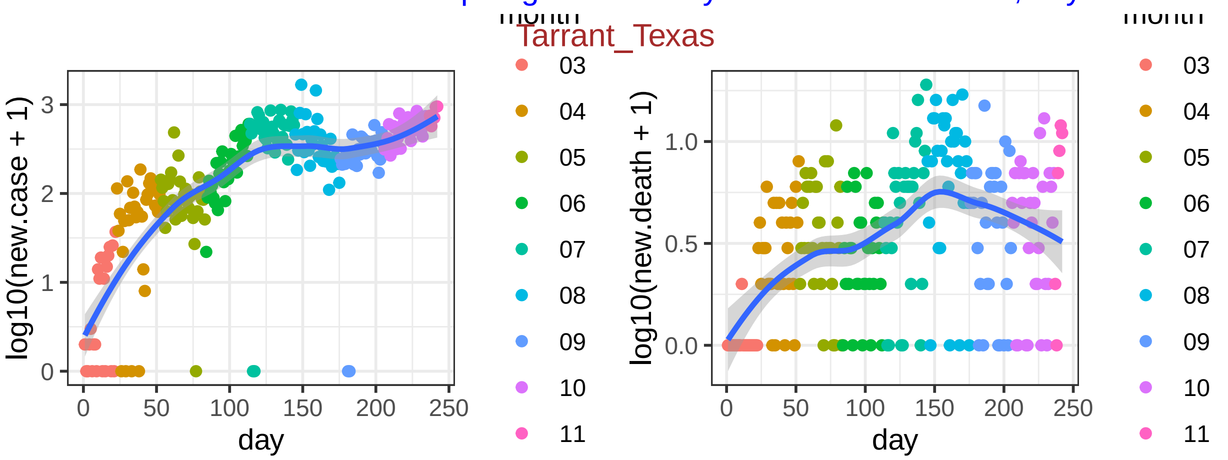
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Monmouth_New Jersey

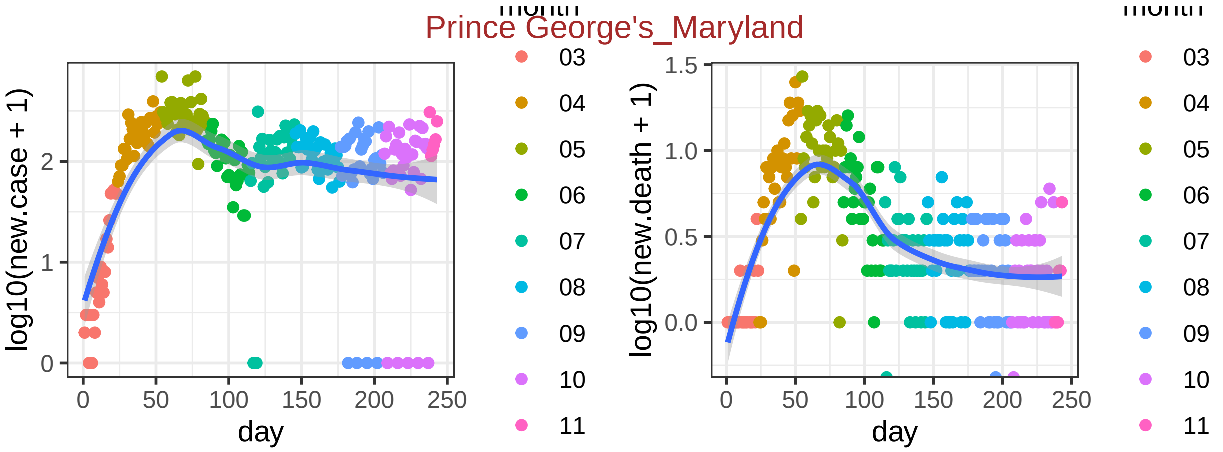


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

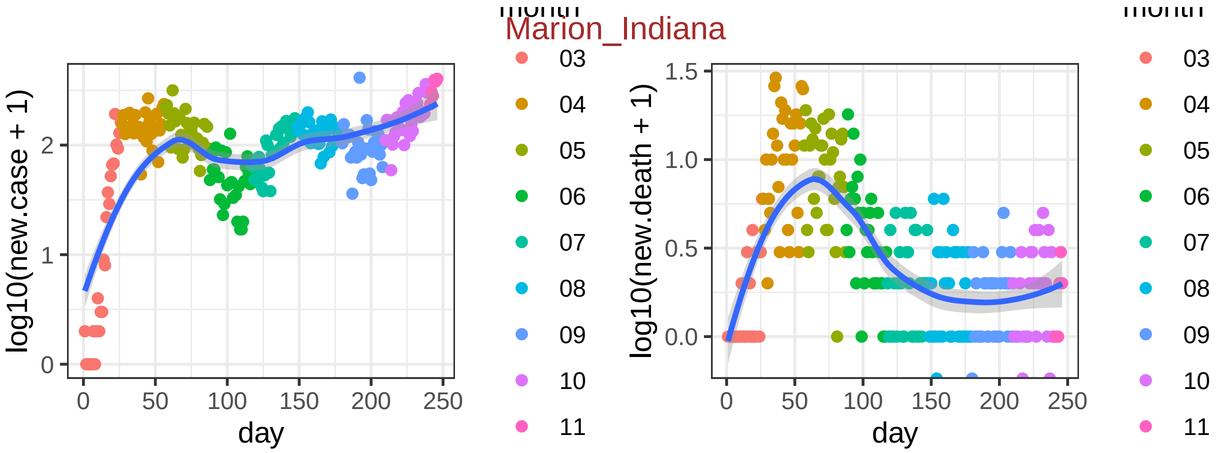
Tarrant_Texas



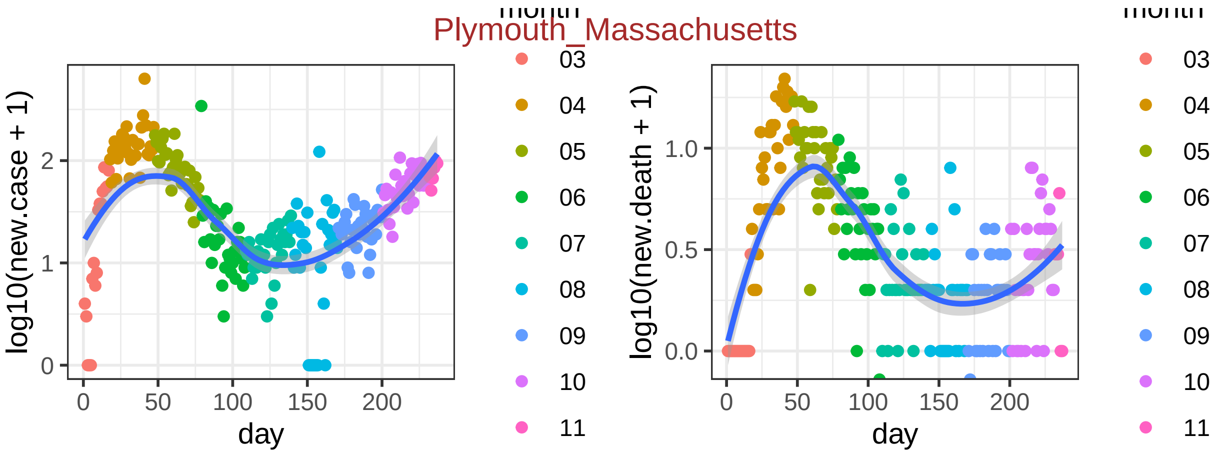
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10


Prince George's_Maryland

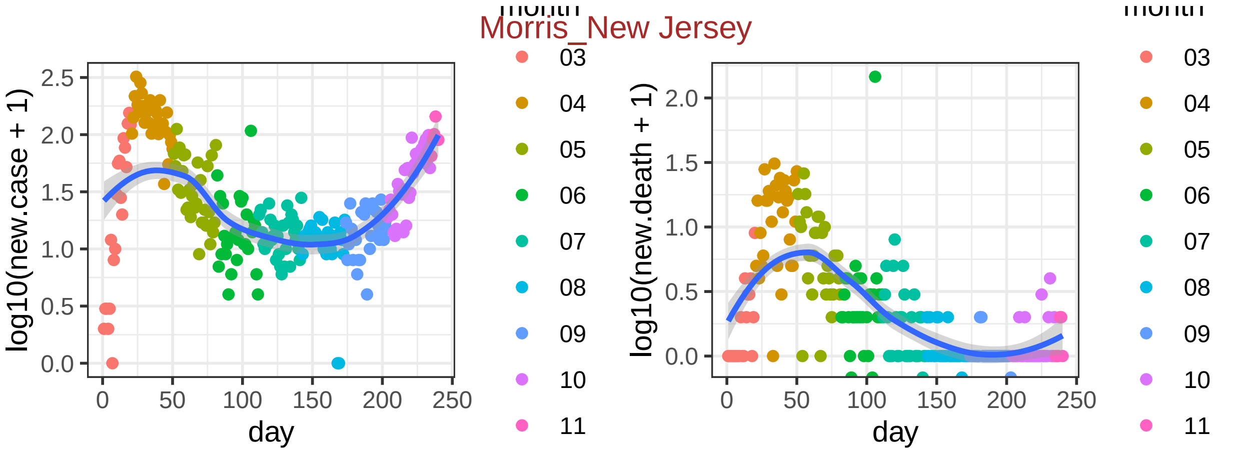
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09


Marion_Indiana

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


Plymouth_Massachusetts

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

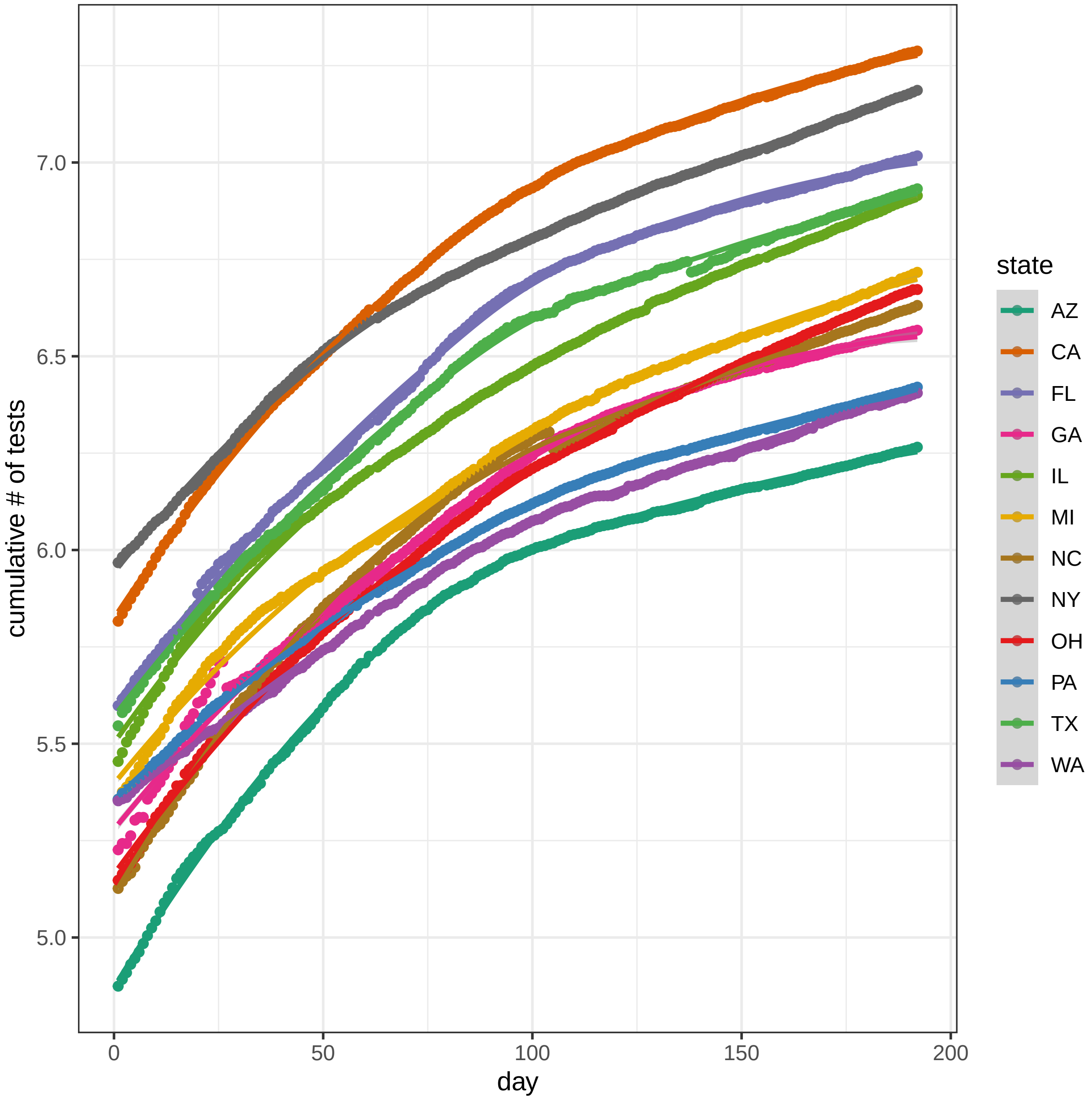


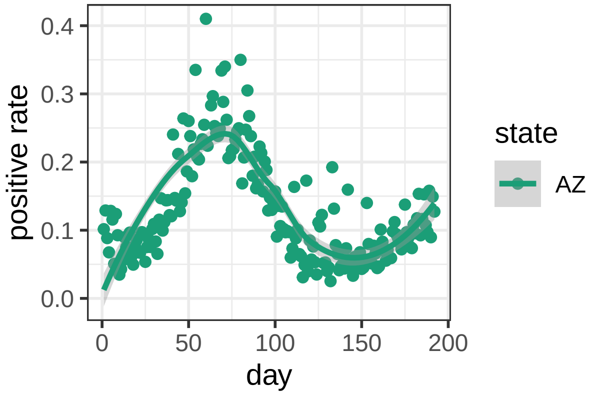
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

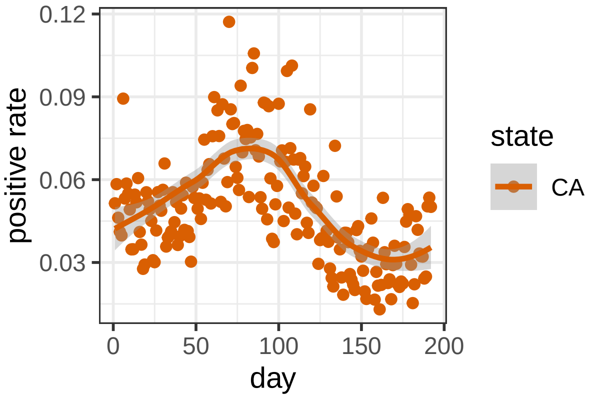
COVID Trackng

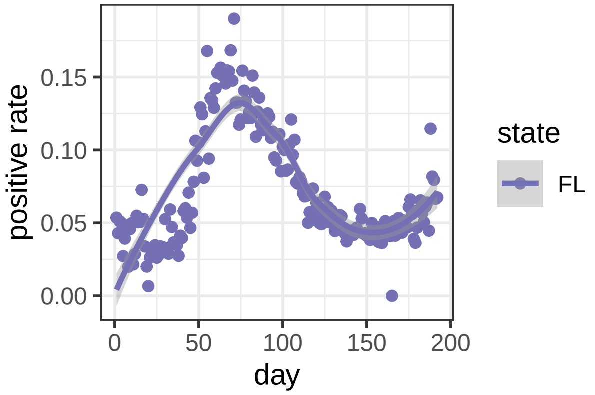
The positive rates of testing can be an indicator on how much the COVID-19 has spread. However, they can be much more noisy data since the negative testing resutls are often not reported and the tests are almost surely taken on a non-representative random sample of the population. The COVID traking project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

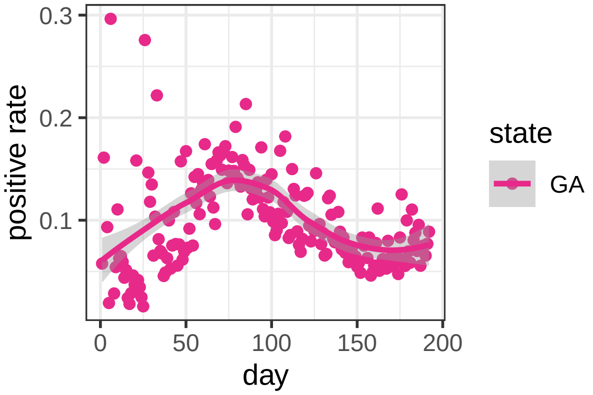
The grades of the states may change over timea and I strongly recommend checking their webiste before puting serious interpretation on the following plot.

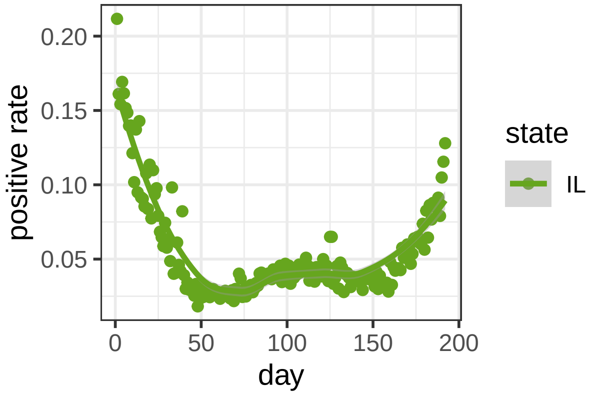

cumulative # of tests
7.0
6.5
6.0
5.5
5.0
0
50
100
150
200
day
state
AZ
CA
FL
GA
IL
MI
NC
NY
OH
PA
TX
WA

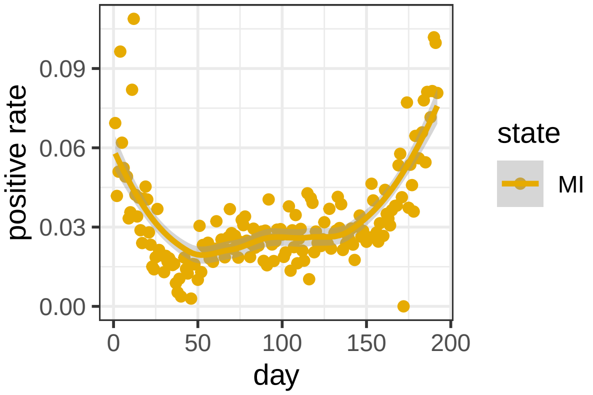

0.4
0.3
0.2
0.1
0.0
positive rate
0
50
100
150
200
day
state
AZ

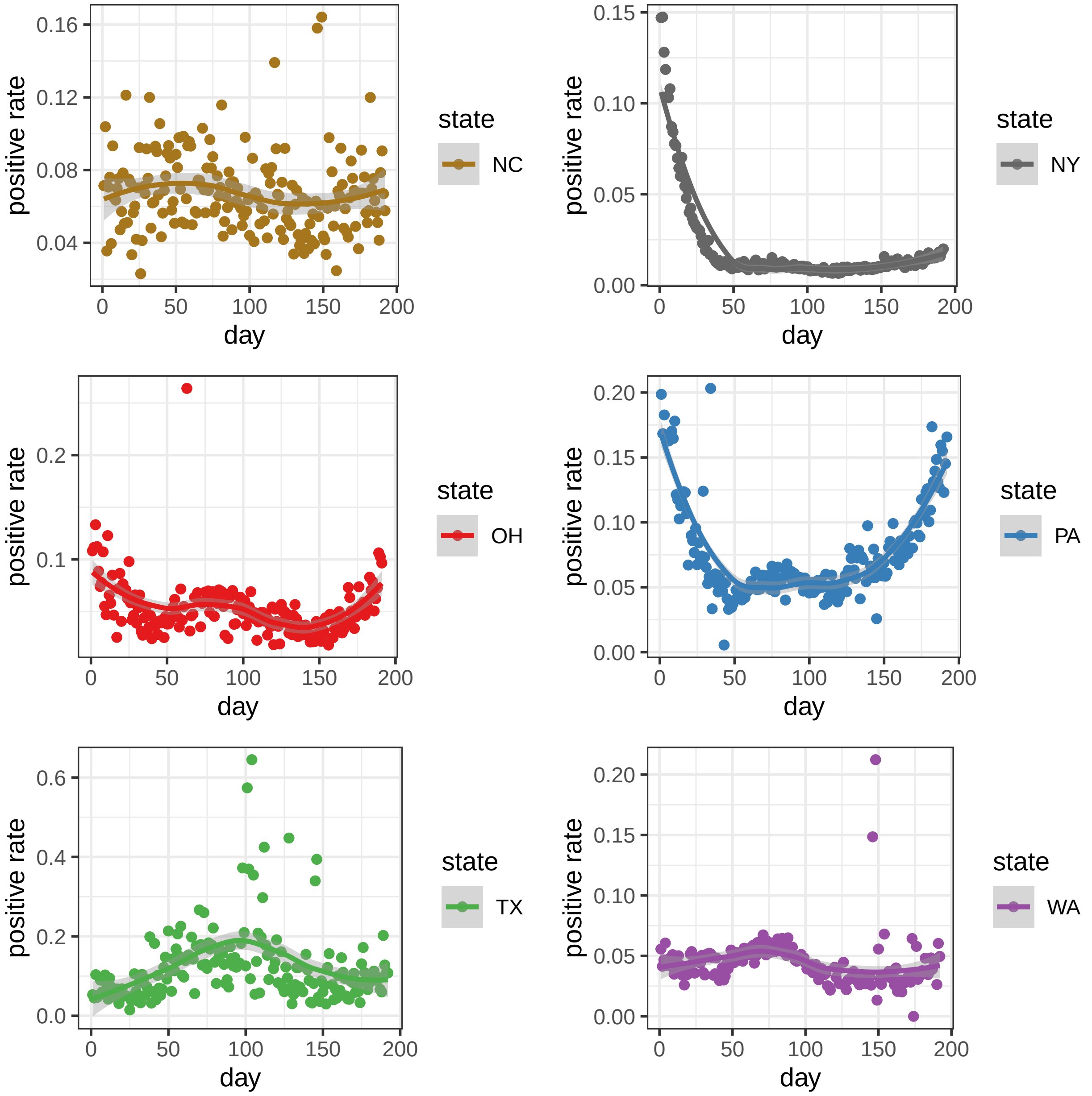

0.12
0.09
0.06
0.03
positive rate
0
50
100
150
200
day
state
CA


0.15
0.10
0.05
0.00
positive rate
0
50
100
150
200
day
state
FL


0.3
0.2
0.1
positive rate
0
50
100
150
200
day
state
GA


0.20
0.15
0.10
0.05
positive rate
0
50
100
150
200
day
state
IL


0.09
0.06
0.03
0.00
positive rate
0
50
100
150
200
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_0.4.0.999   ggplot2_3.3.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.0.0  xfun_0.12         purrr_0.3.3       splines_3.6.2
##  [5] haven_2.3.0       lattice_0.20-38   carData_3.0-4     colorspace_1.4-1
##  [9] vctrs_0.3.0       generics_0.0.2    htmltools_0.4.0   mgcv_1.8-31
## [13] yaml_2.2.1        rlang_0.4.6       pillar_1.4.3      foreign_0.8-75
## [17] glue_1.3.1        withr_2.1.2       readxl_1.3.1      lifecycle_0.2.0
## [21] stringr_1.4.0     munsell_0.5.0     ggsignif_0.6.0    gtable_0.3.0
## [25] cellranger_1.1.0  zip_2.0.4         evaluate_0.14     labeling_0.3
## [29] knitr_1.28        rio_0.5.16        forcats_0.5.0     curl_4.3
## [33] broom_0.5.6       Rcpp_1.0.3        scales_1.1.0      backports_1.1.5
## [37] abind_1.4-5       farver_2.0.3      gridExtra_2.3     hms_0.5.3
## [41] digest_0.6.23     stringi_1.4.5     openxlsx_4.1.5    rstatix_0.6.0
## [45] dplyr_0.8.4       cowplot_1.0.0     grid_3.6.2        tools_3.6.2
## [49] magrittr_1.5      tibble_3.0.1      crayon_1.3.4      tidyr_1.0.2
## [53] car_3.0-8         pkgconfig_2.0.3   Matrix_1.2-18     ellipsis_0.3.0
## [57] data.table_1.12.8 assertthat_0.2.1  rmarkdown_2.1     R6_2.4.1
## [61] nlme_3.1-144      compiler_3.6.2
```