Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-10-23

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

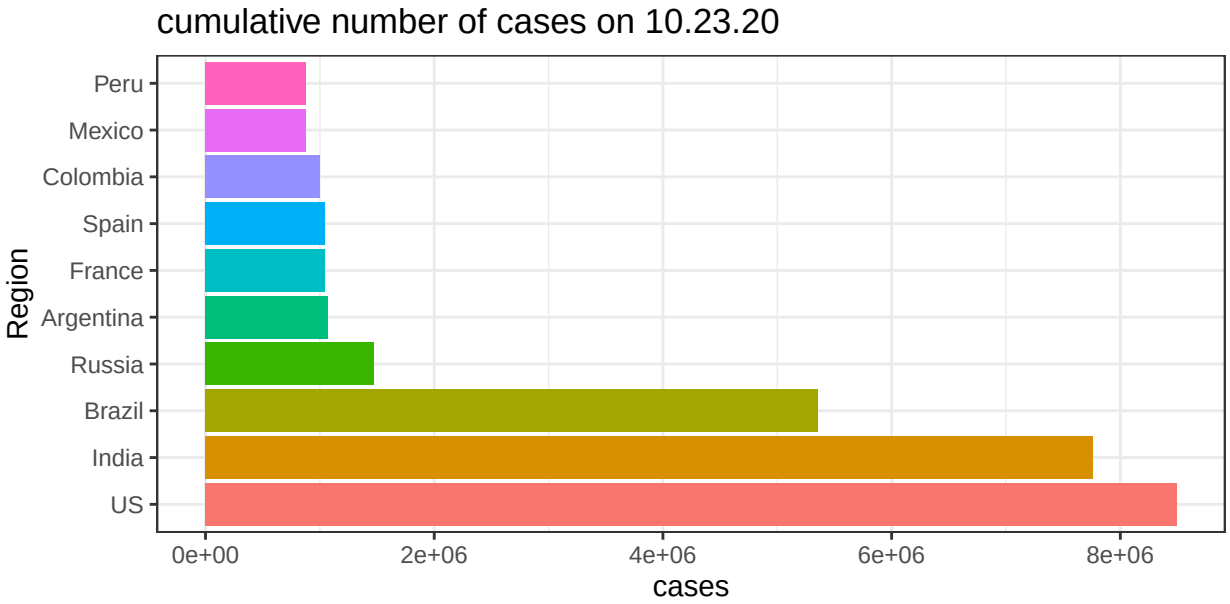
JHU

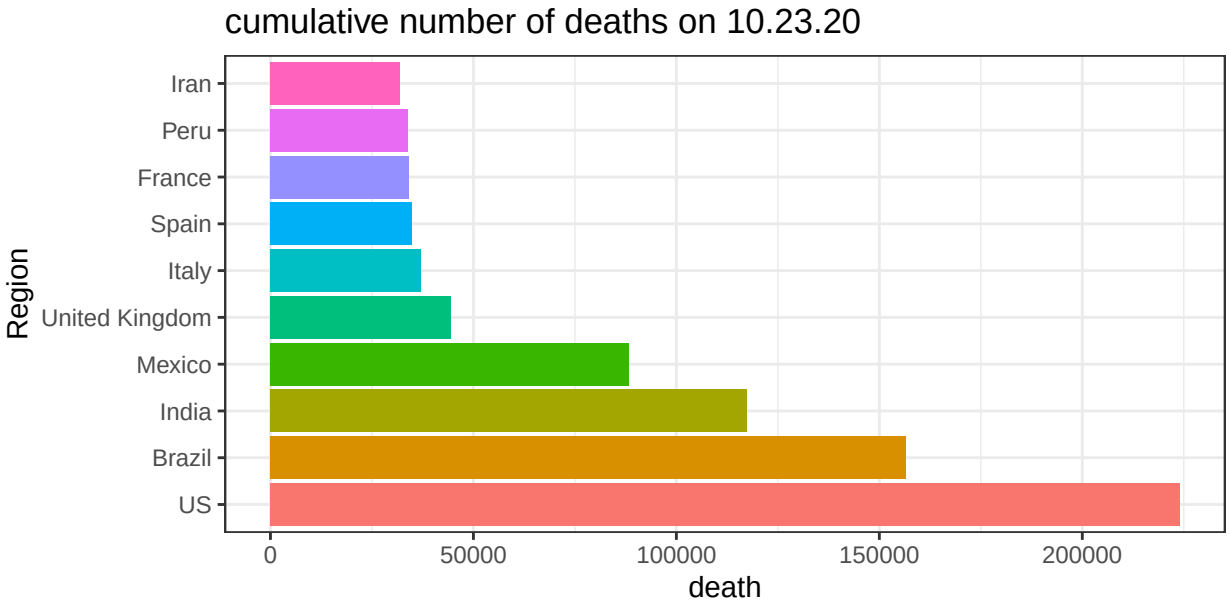
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

time series data

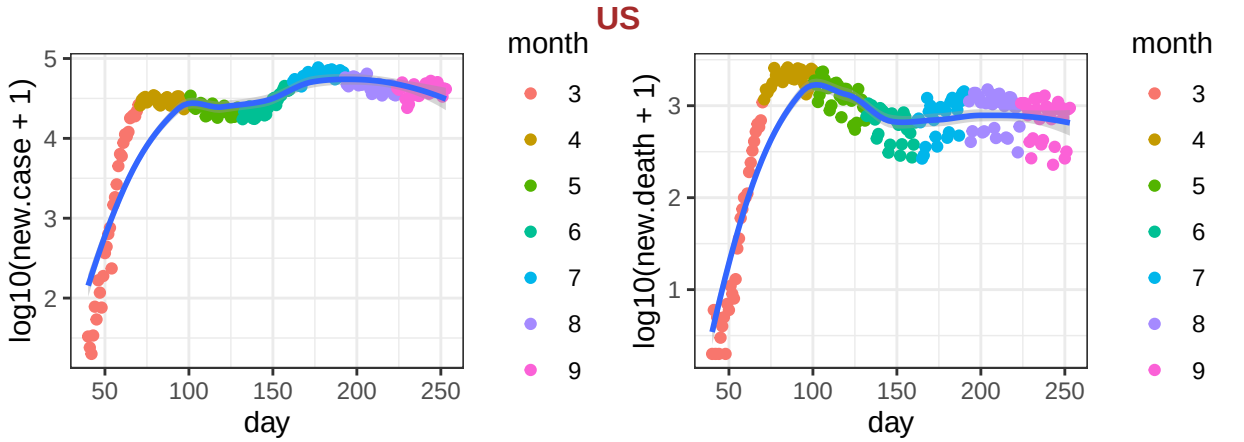
The time series provide counts (e.g., confirmed cases, deaths) starting from Jan 22nd, 2020 for 253 locations. Currently there is no data of individual US state in these time series data files.

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

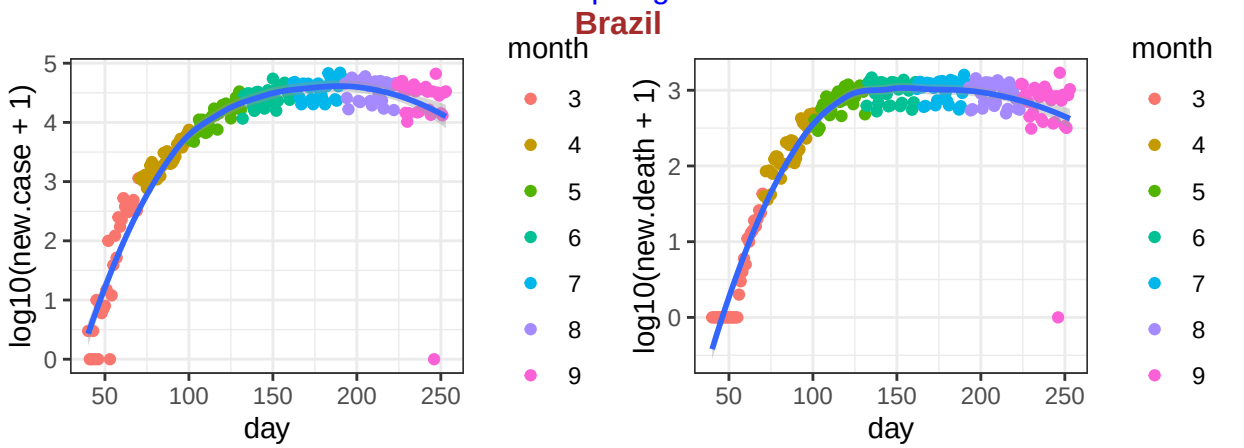




Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

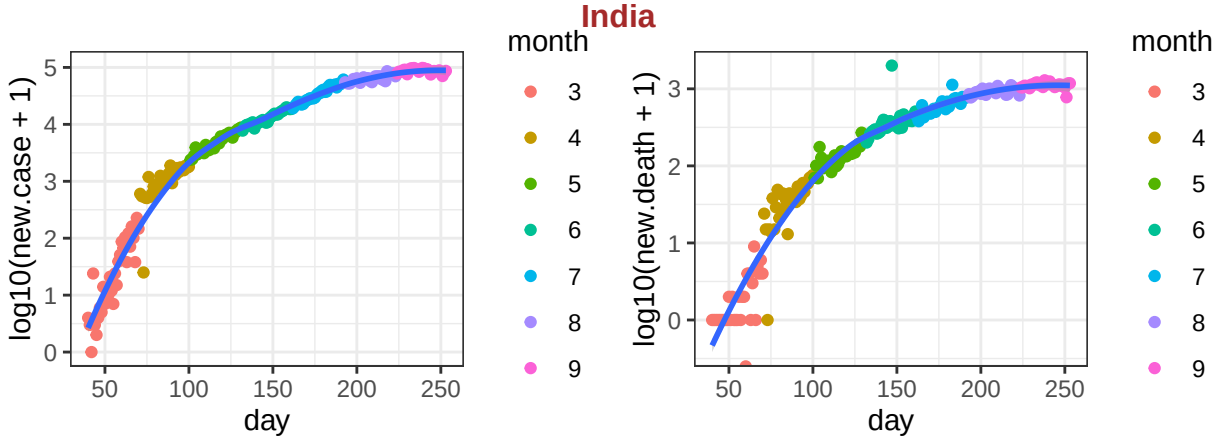


data source: https://github.com/CSSEGISandData/COVID−19



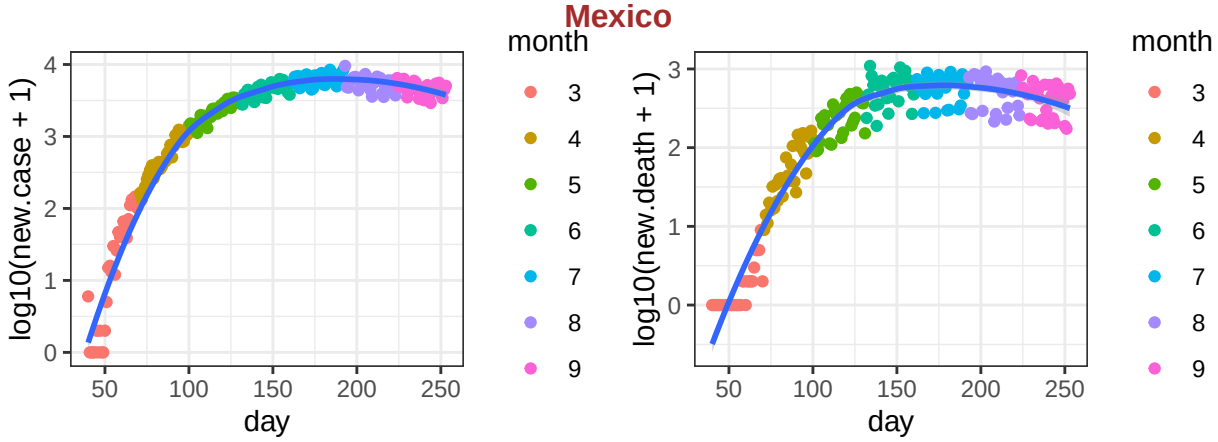
data source: https://github.com/CSSEGISandData/COVID−19

India



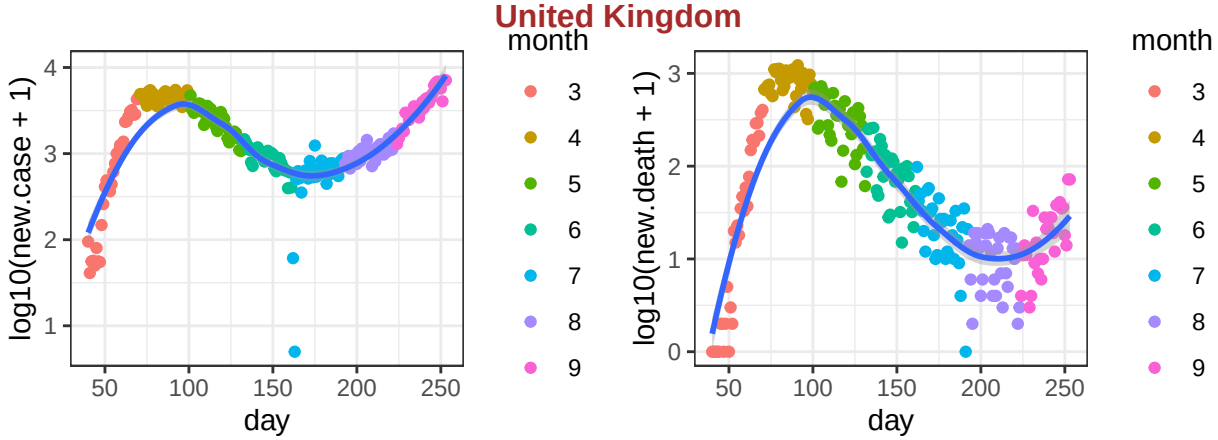
data source: https://github.com/CSSEGISandData/COVID−19

Mexico



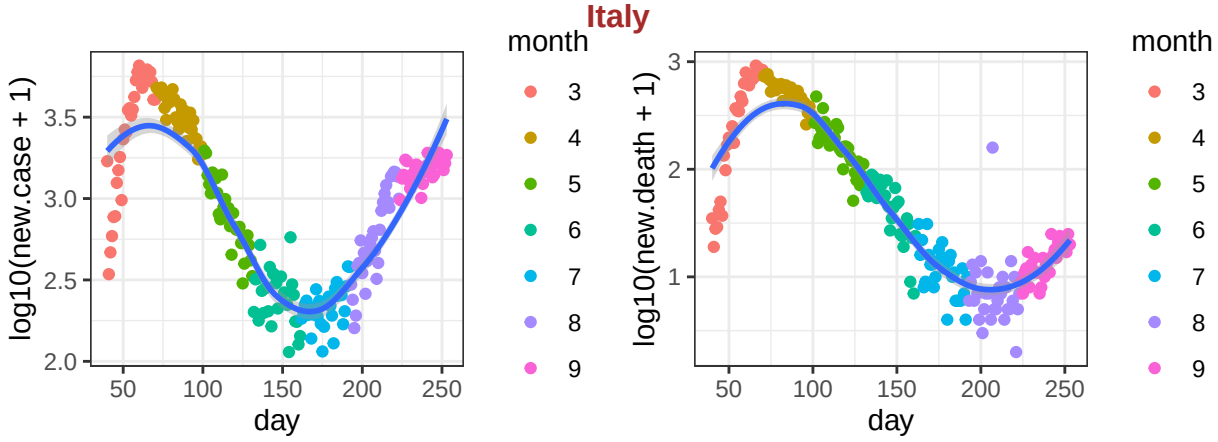
data source: https://github.com/CSSEGISandData/COVID−19

United Kingdom



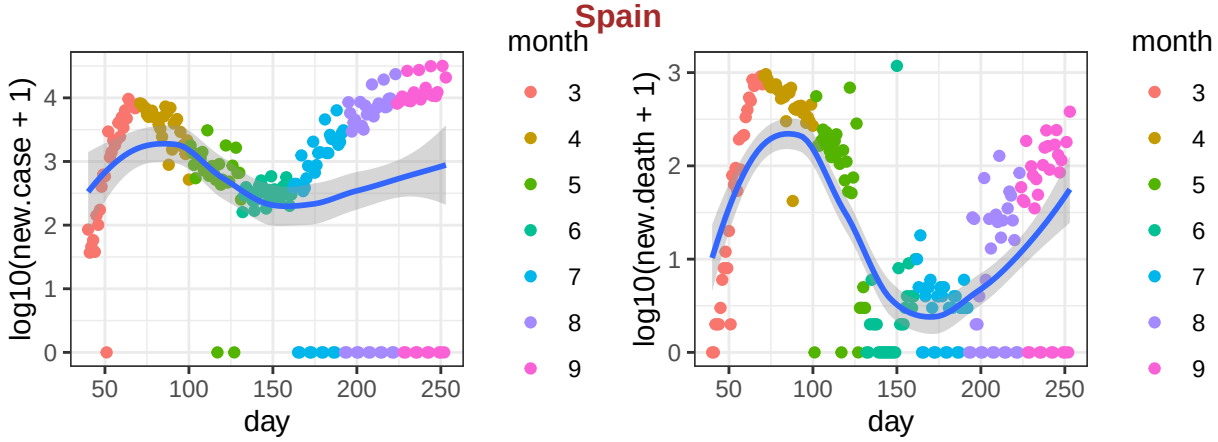
data source: https://github.com/CSSEGISandData/COVID−19

Italy



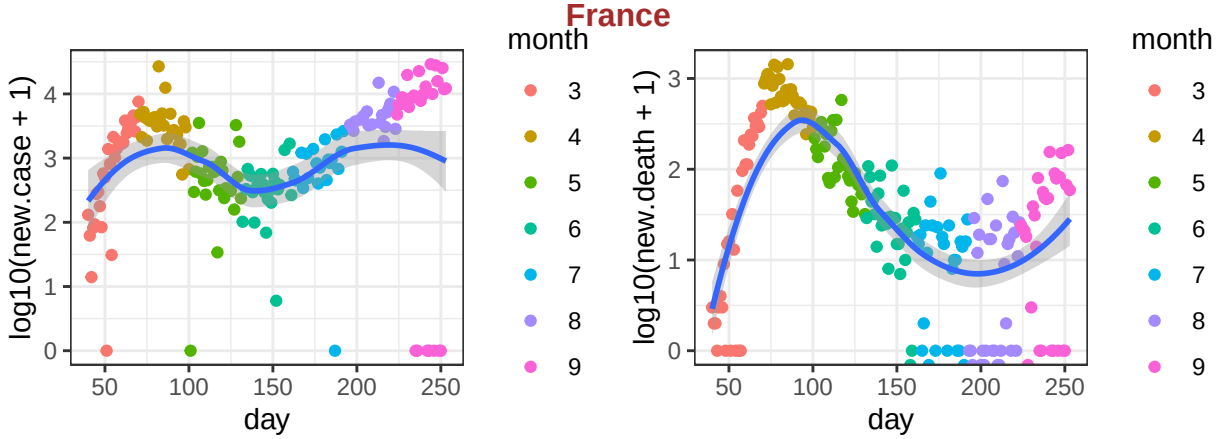
data source: https://github.com/CSSEGISandData/COVID−19

Spain

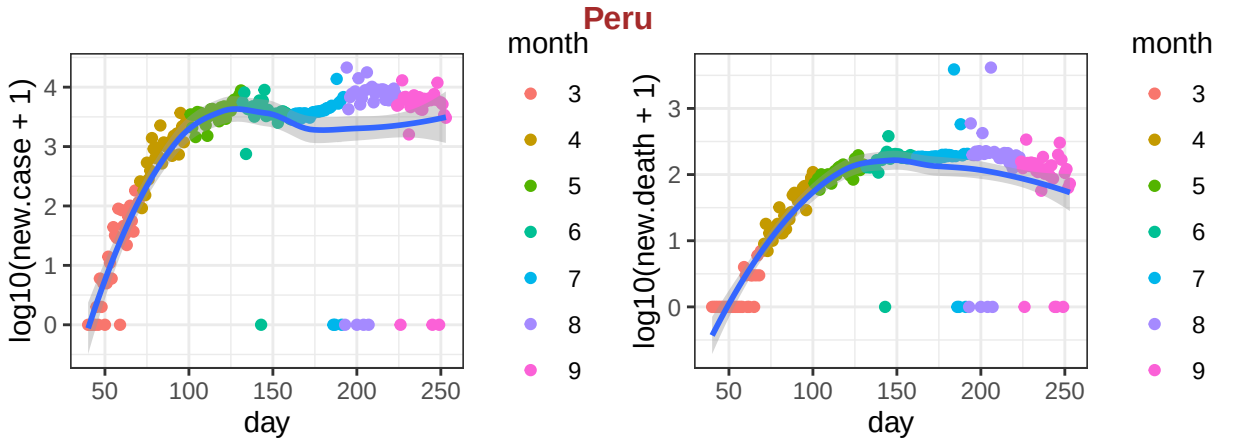


data source: https://github.com/CSSEGISandData/COVID−19

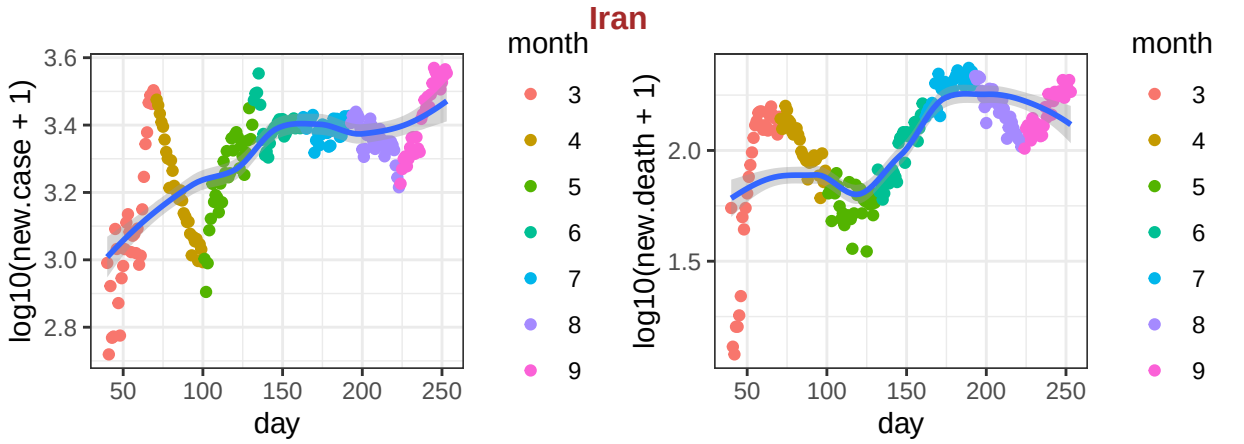
France



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



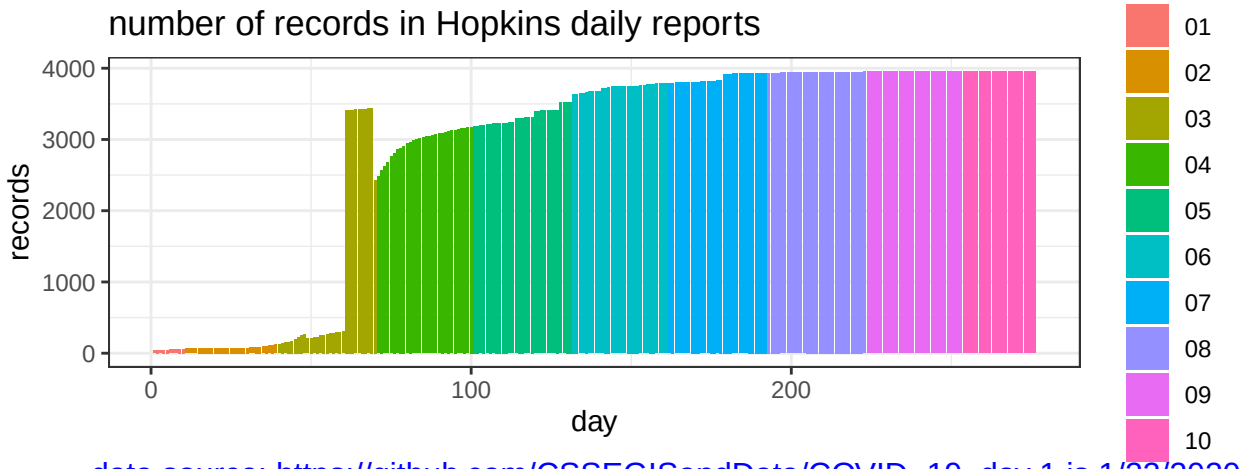
data source: https://github.com/CSSEGISandData/COVID−19



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

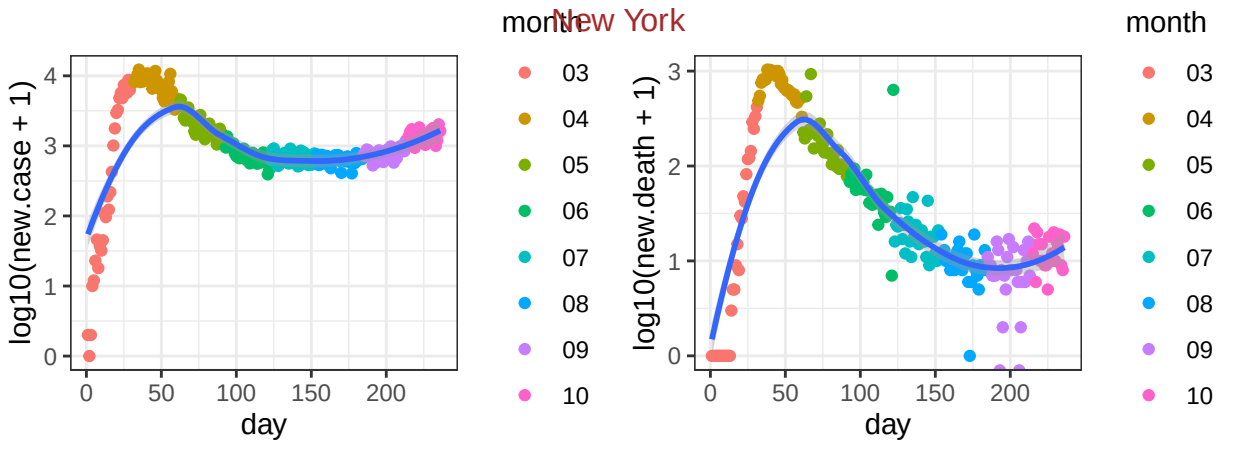
```
## [1] "2020-10-22"
```

state level data

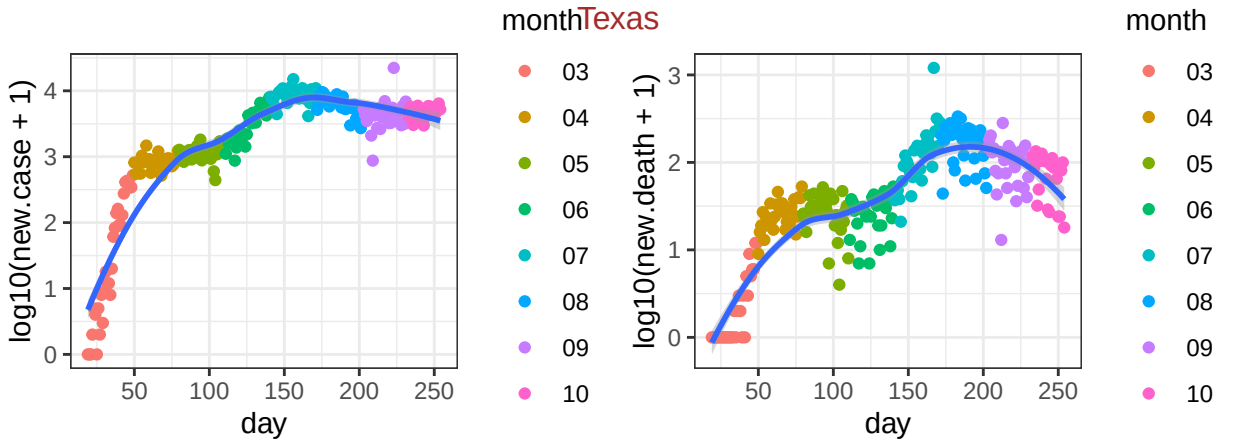
First check the 30 states with the largest number of deaths.

```
##             date          state fips  cases deaths
## 12863 2020-10-22       New York   36 494874  33022
## 12876 2020-10-22          Texas   48 891074  17760
## 12834 2020-10-22     California    6 896424  17266
## 12839 2020-10-22        Florida   12 768083  16266
## 12861 2020-10-22     New Jersey   34 226174  16263
## 12852 2020-10-22  Massachusetts   25 147215   9810
## 12844 2020-10-22       Illinois   17 365120   9663
## 12870 2020-10-22   Pennsylvania   42 193401   8660
## 12840 2020-10-22        Georgia   13 331316   7547
## 12853 2020-10-22       Michigan   26 170102   7465
## 12832 2020-10-22        Arizona    4 234914   5859
## 12849 2020-10-22      Louisiana   22 181904   5799
## 12867 2020-10-22           Ohio   39 190430   5161
## 12836 2020-10-22    Connecticut    9  65373   4569
## 12864 2020-10-22 North Carolina   37 253418   4110
## 12851 2020-10-22       Maryland   24 138473   4070
## 12845 2020-10-22        Indiana   18 157678   4065
## 12873 2020-10-22 South Carolina   45 167485   3755
## 12880 2020-10-22       Virginia   51 170104   3524
## 12855 2020-10-22    Mississippi   28 113081   3231
## 12875 2020-10-22      Tennessee   47 234079   2983
## 12830 2020-10-22        Alabama    1 177064   2843
## 12856 2020-10-22       Missouri   29 169311   2734
## 12881 2020-10-22     Washington   53 105364   2389
## 12854 2020-10-22      Minnesota   27 128205   2354
## 12835 2020-10-22       Colorado    8  90639   2224
## 12833 2020-10-22       Arkansas    5 102798   1772
## 12859 2020-10-22         Nevada   32  93028   1736
## 12883 2020-10-22      Wisconsin   55 195853   1730
## 12846 2020-10-22           Iowa   19 112242   1617
```

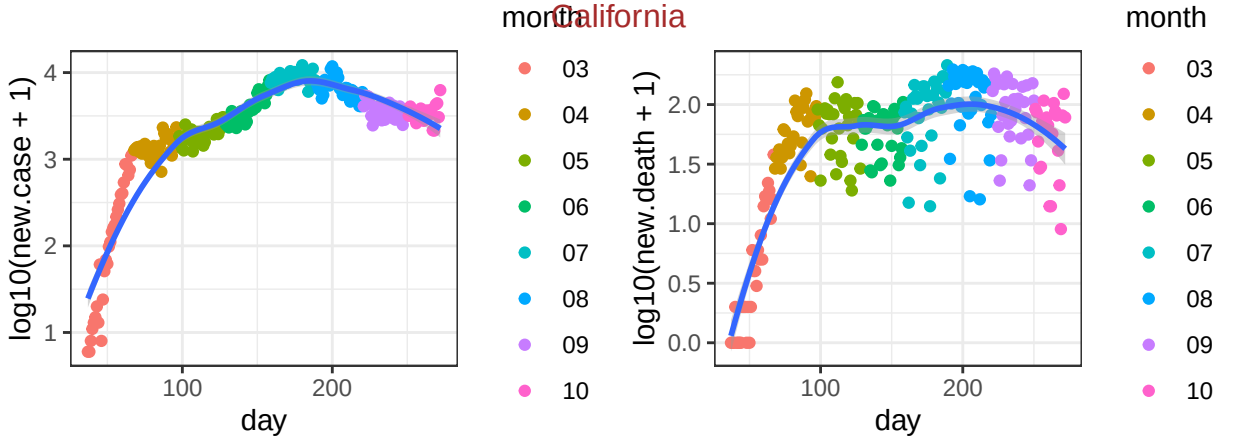
For these 30 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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



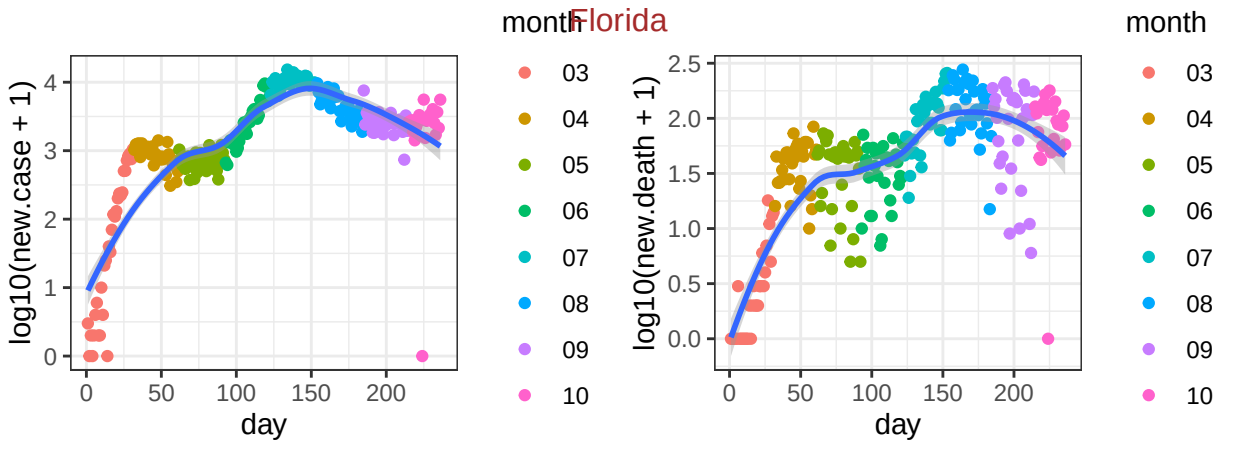
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



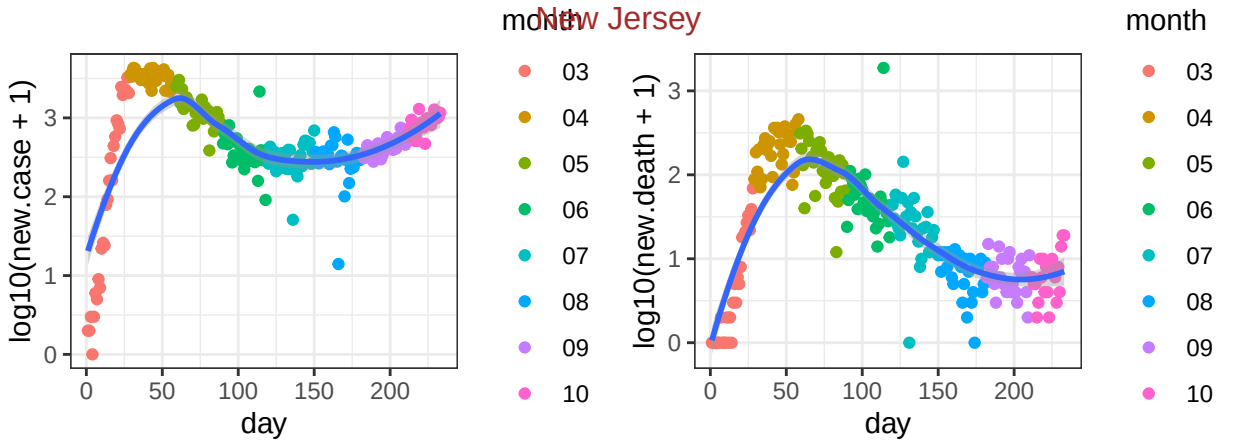
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



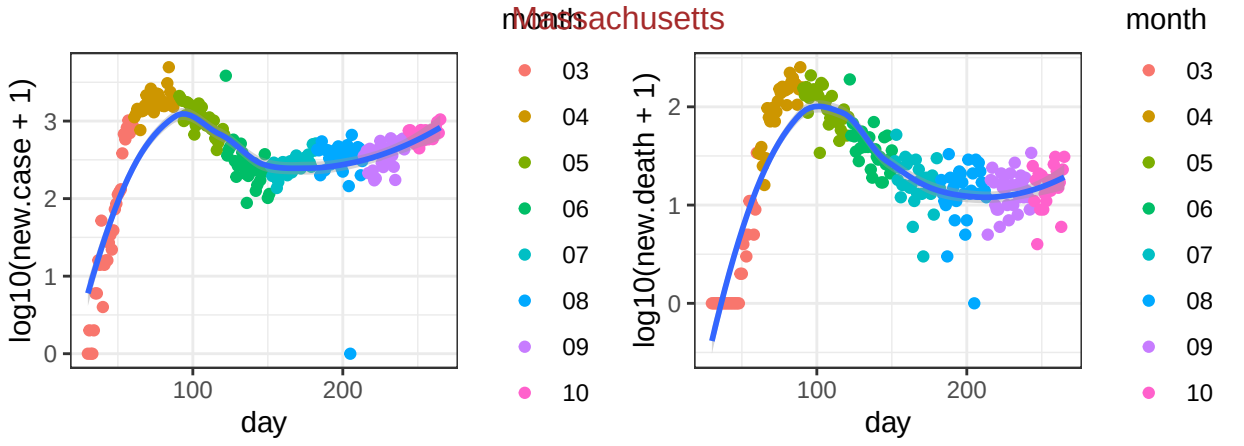
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



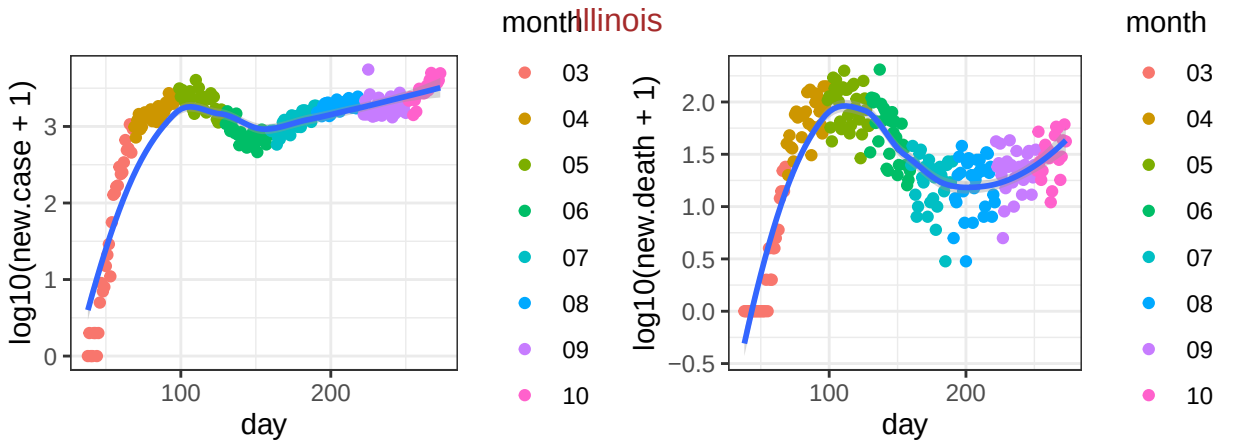
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



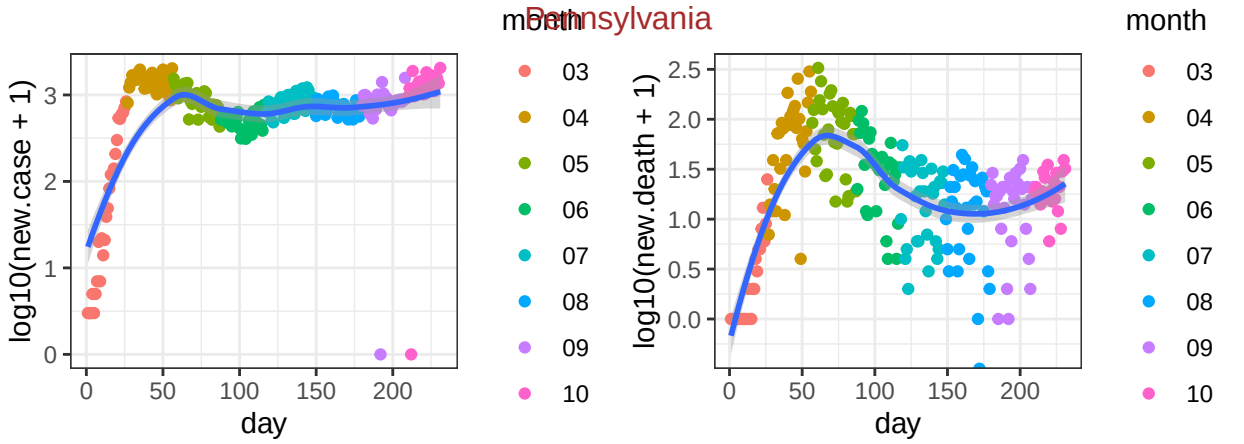
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04



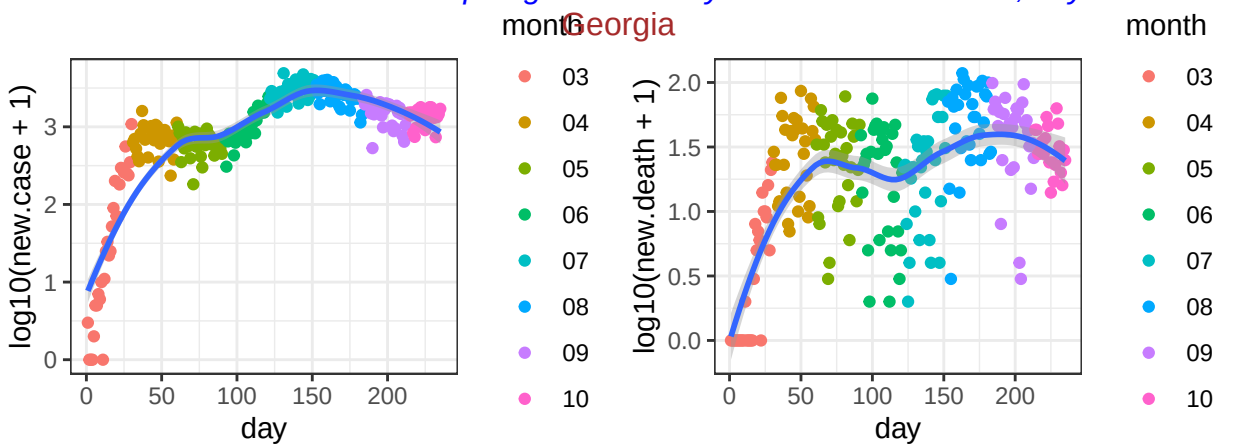
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



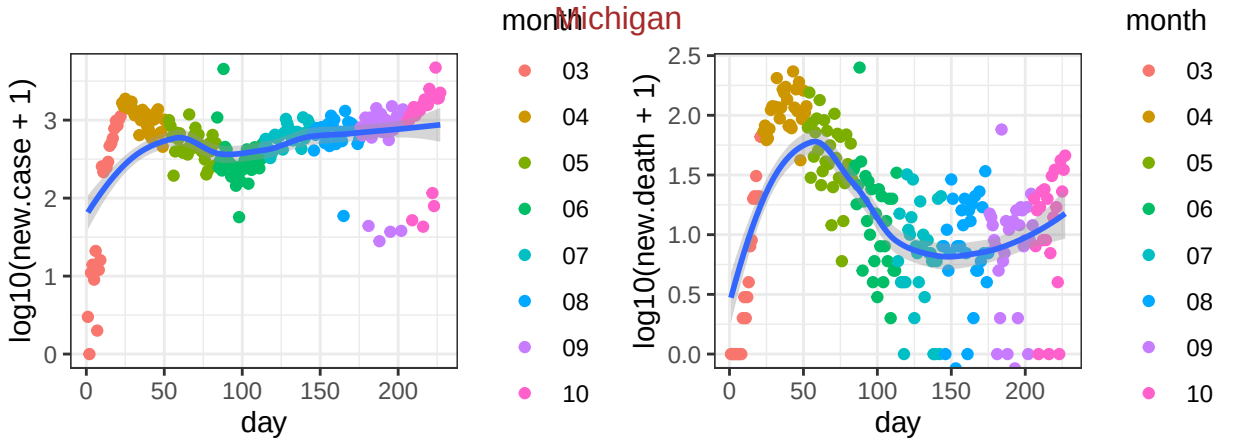
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



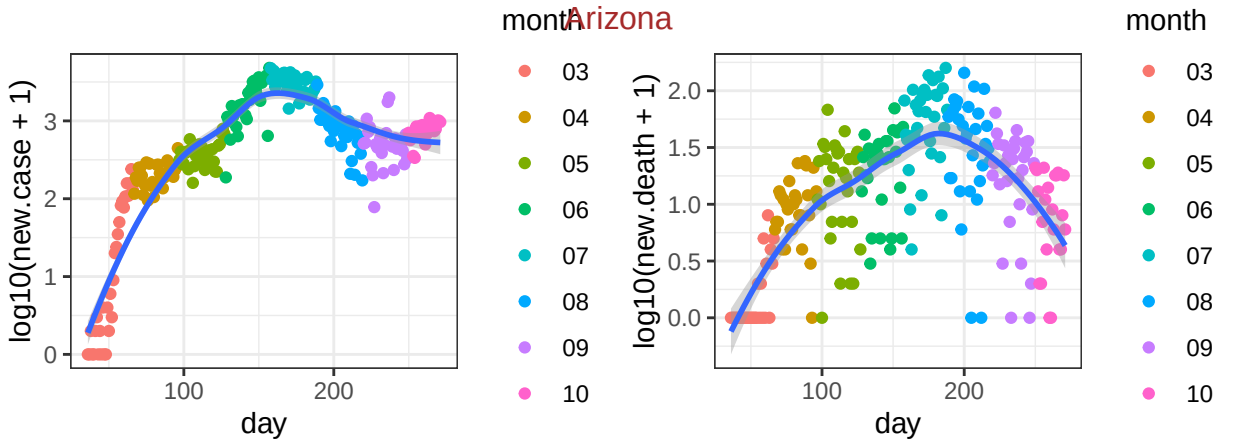
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



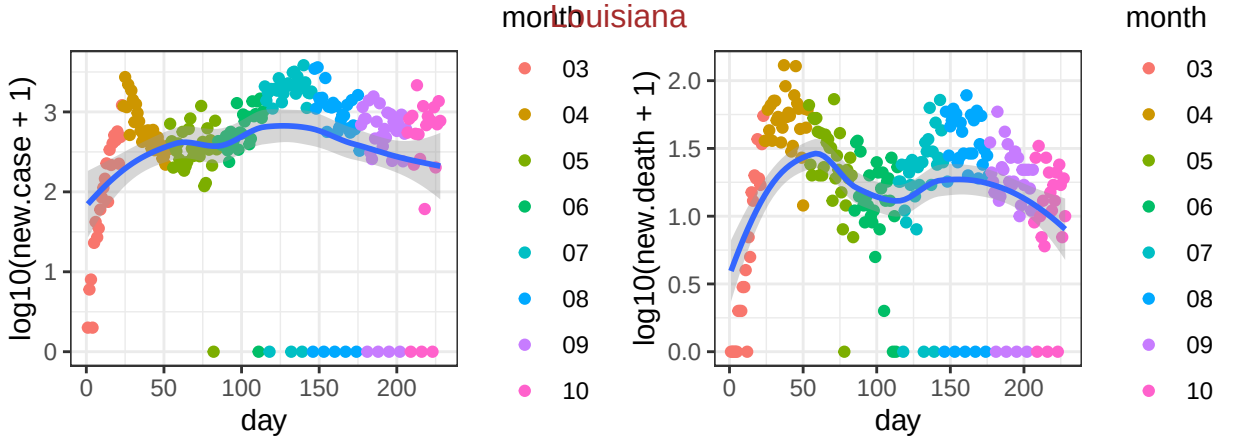
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02



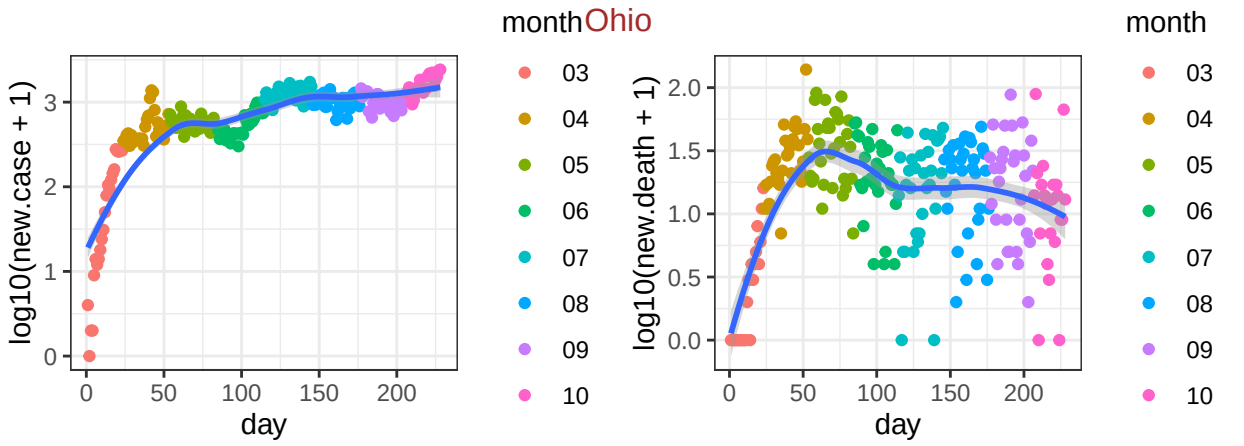
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10



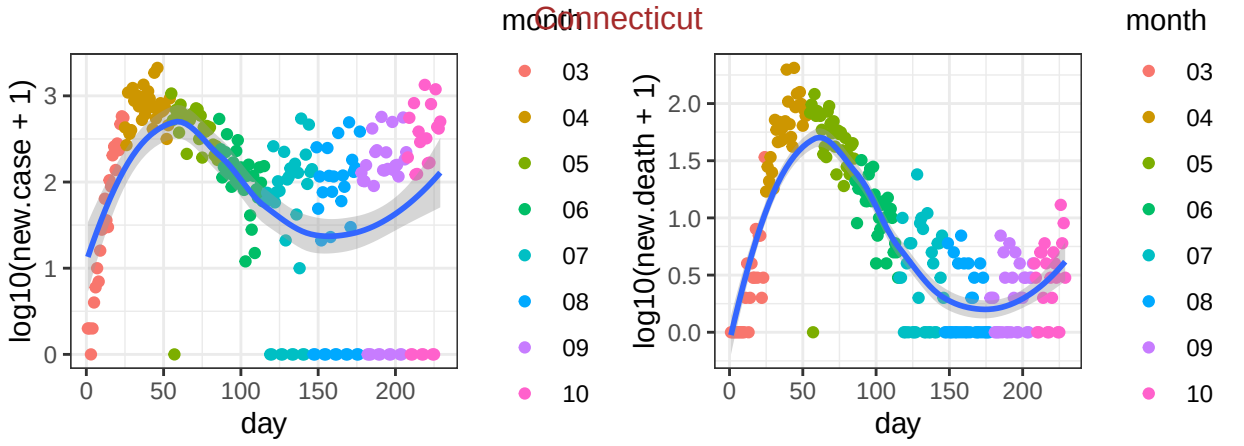
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



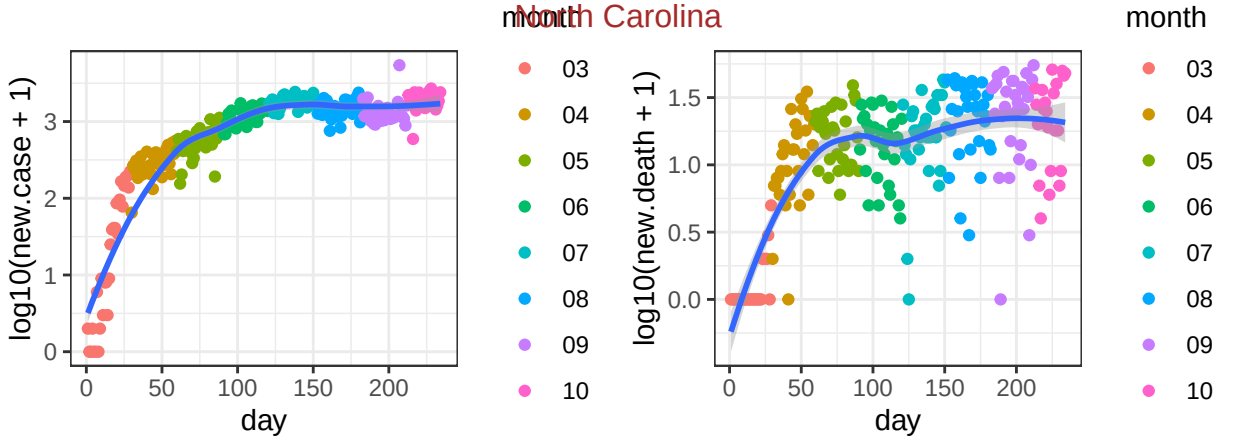
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09



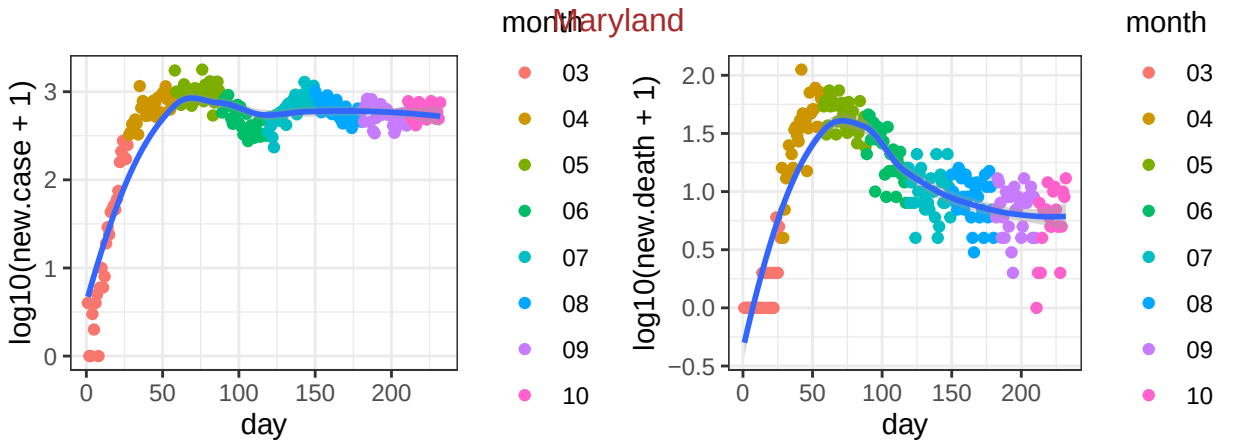
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09



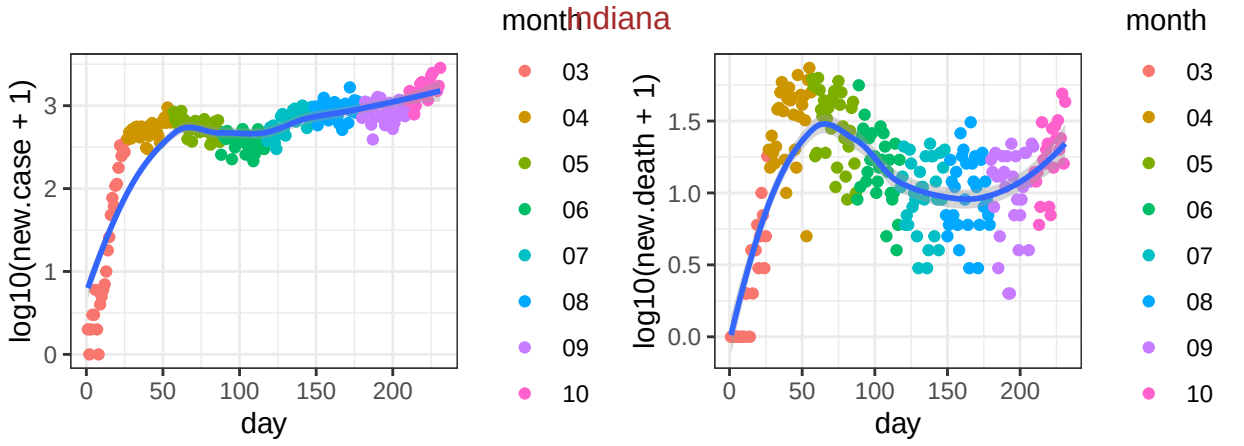
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08



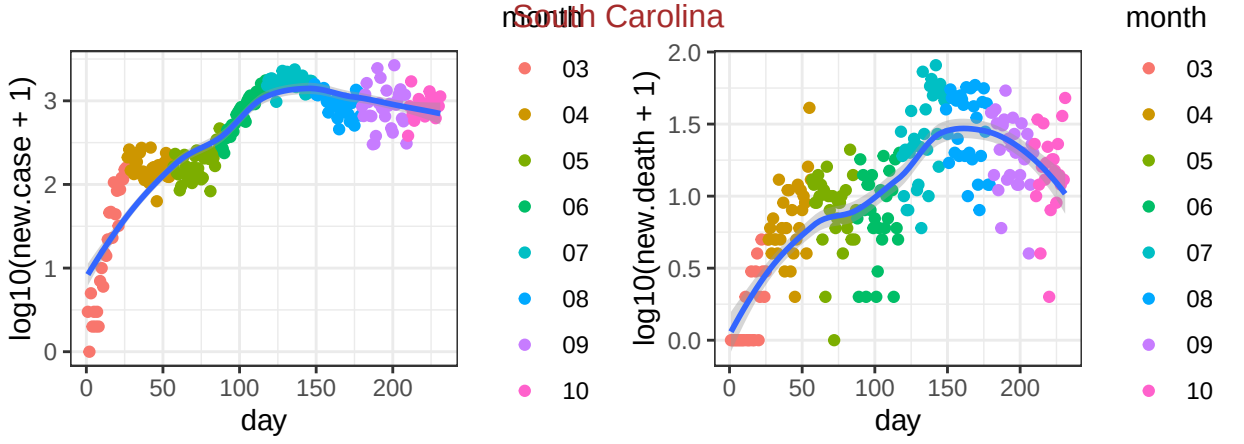
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03



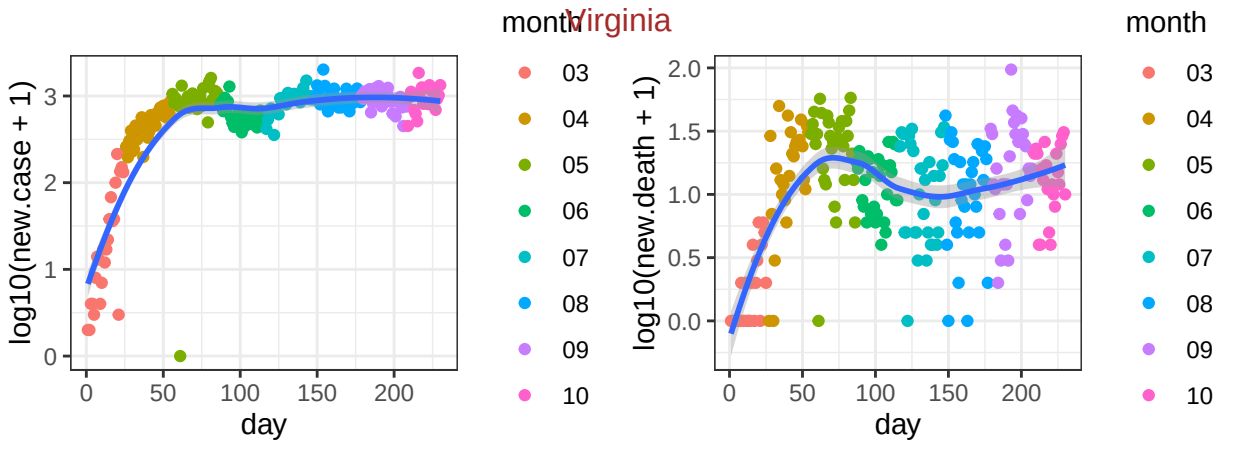
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



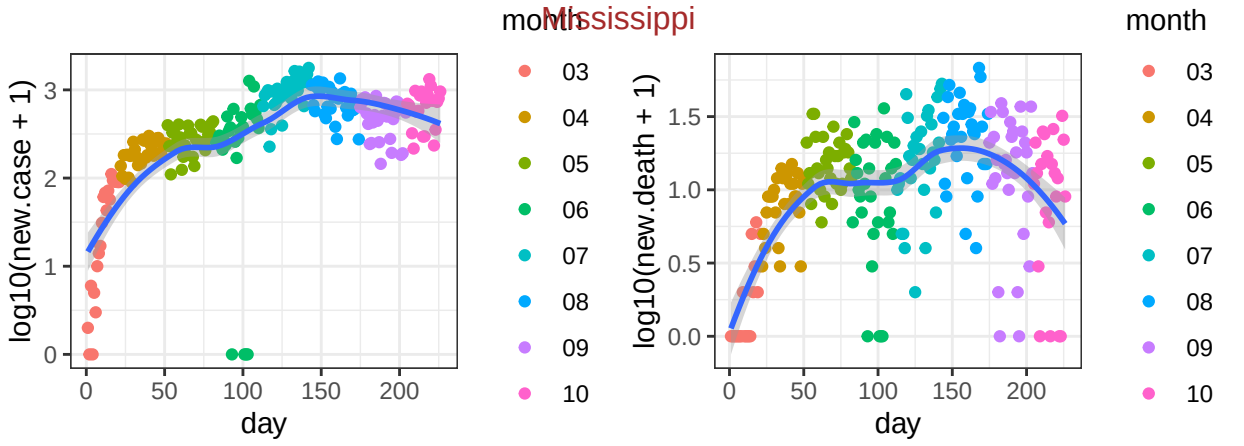
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



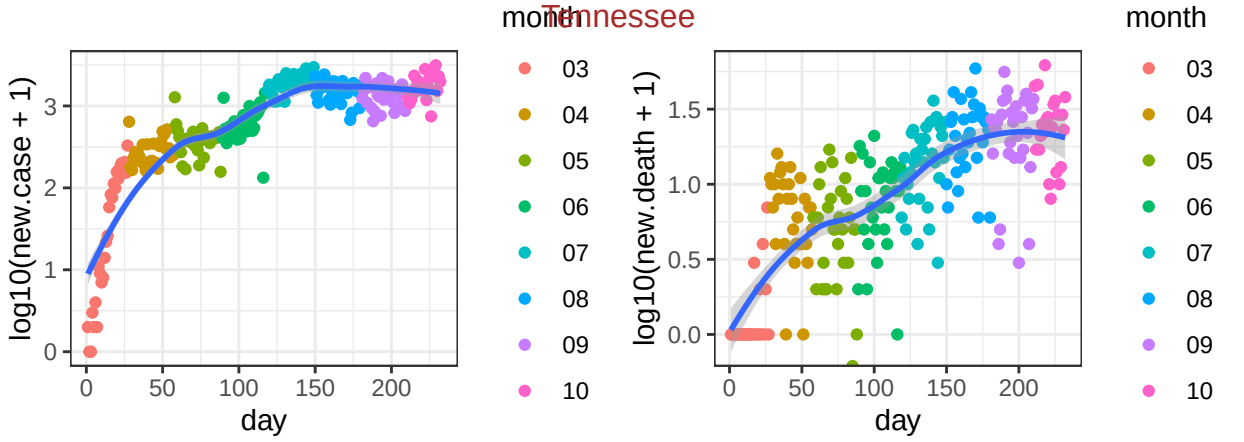
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



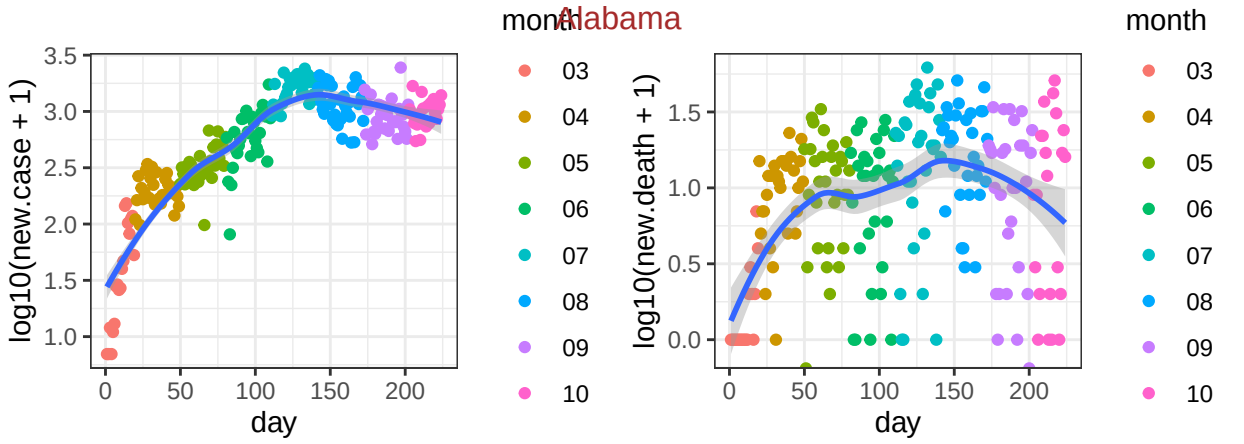
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07



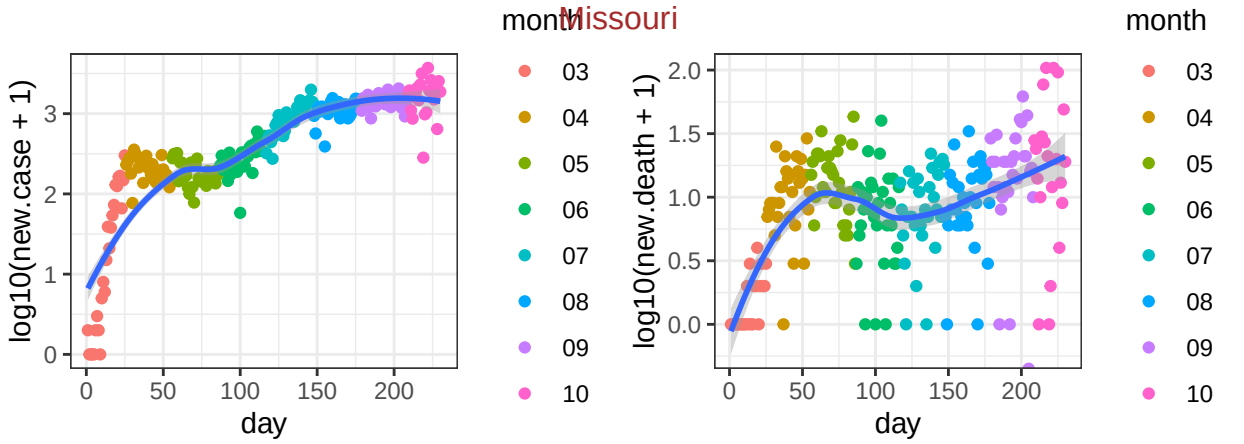
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11



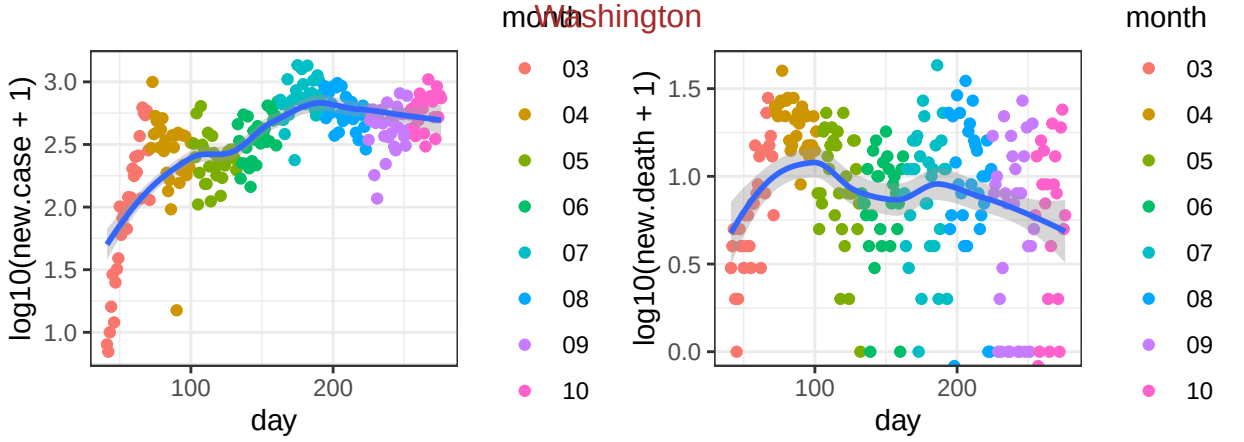
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



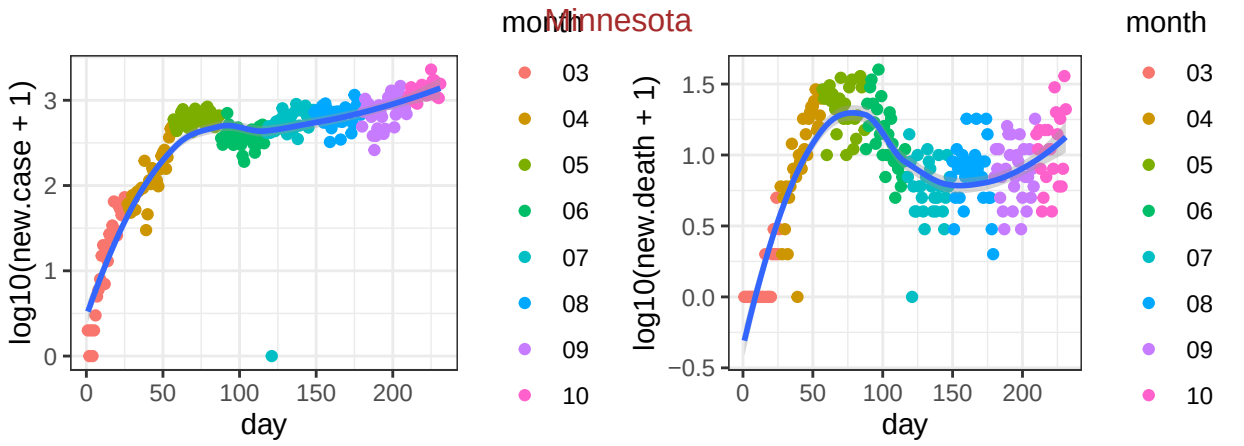
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13



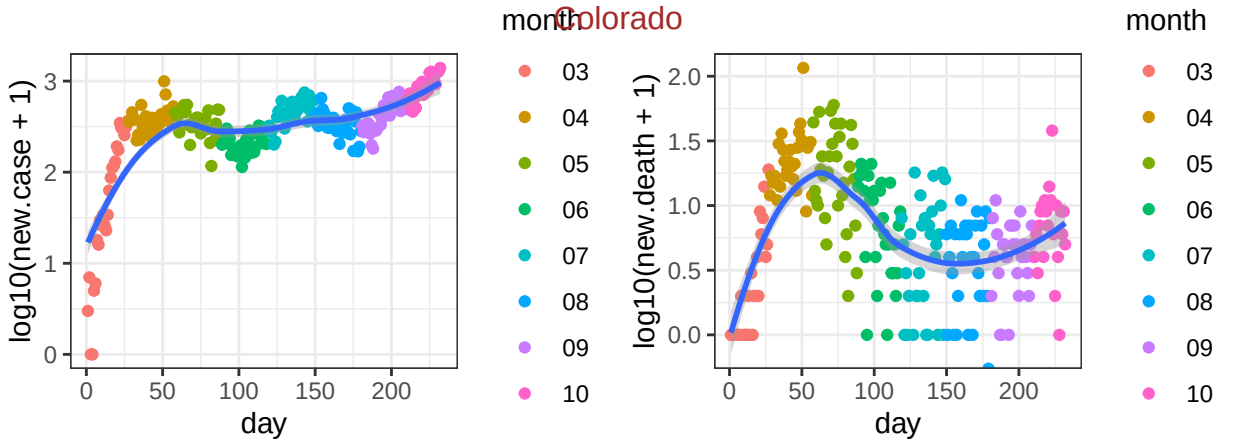
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07



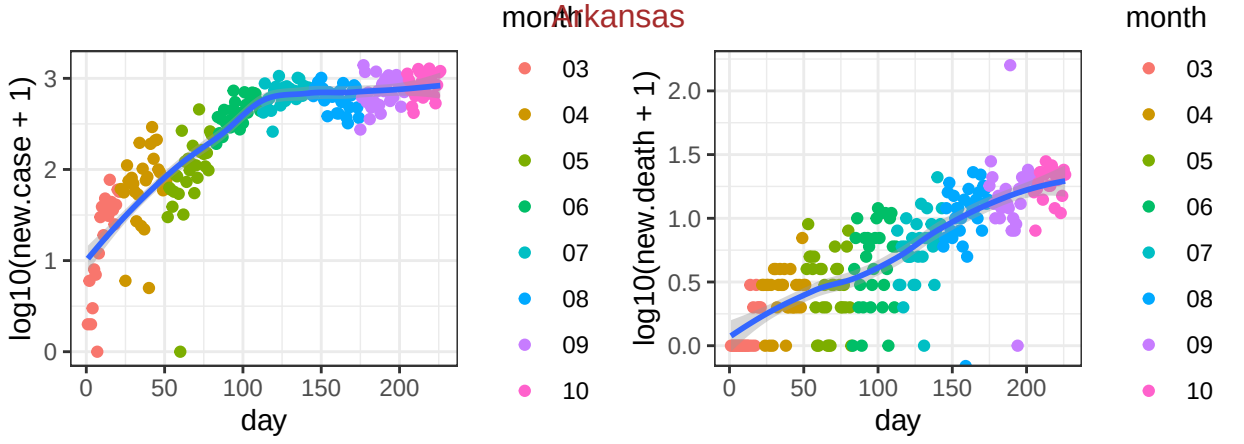
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



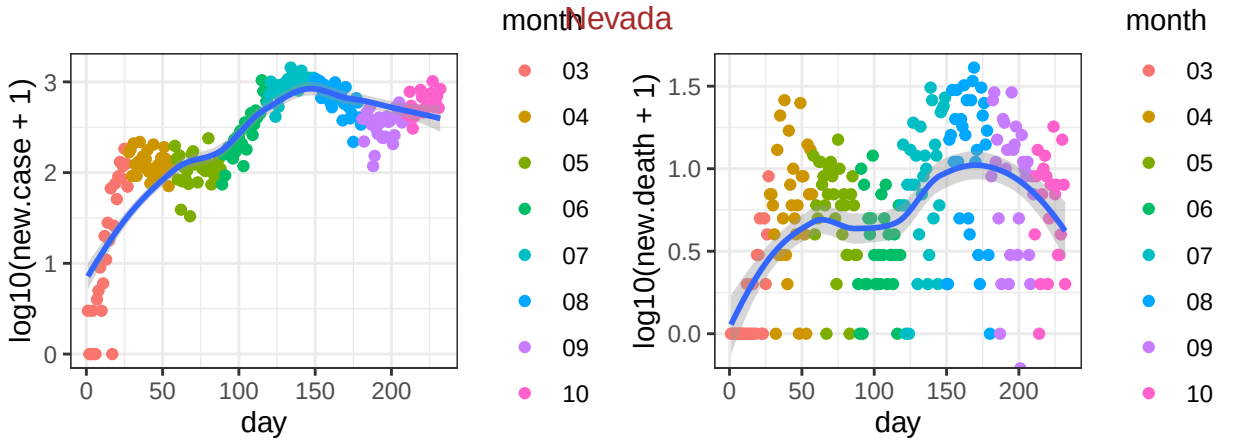
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



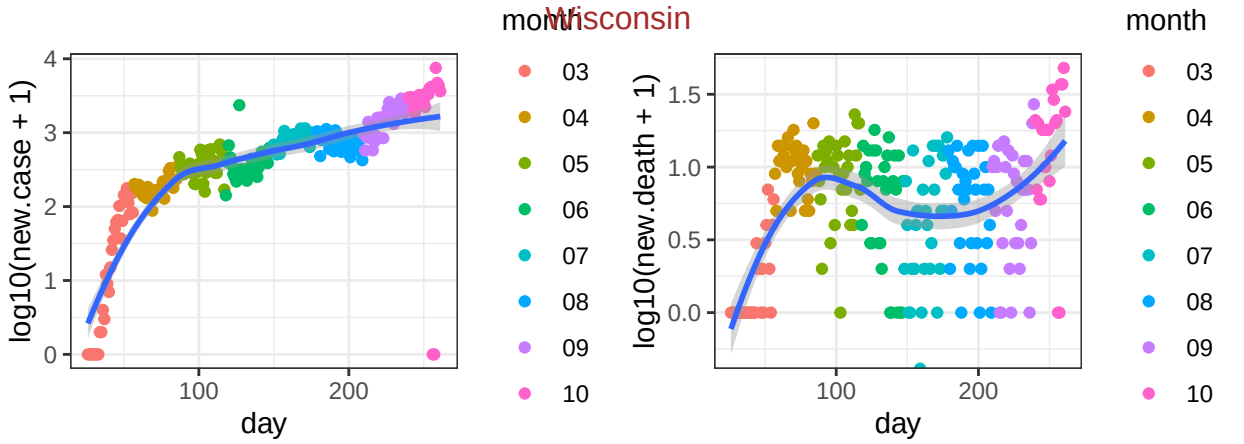
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



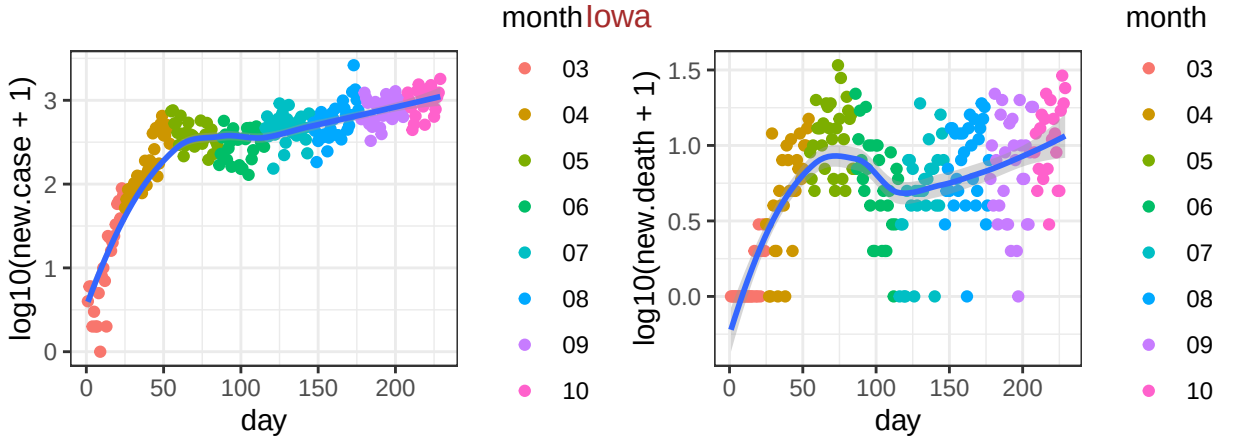
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11



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

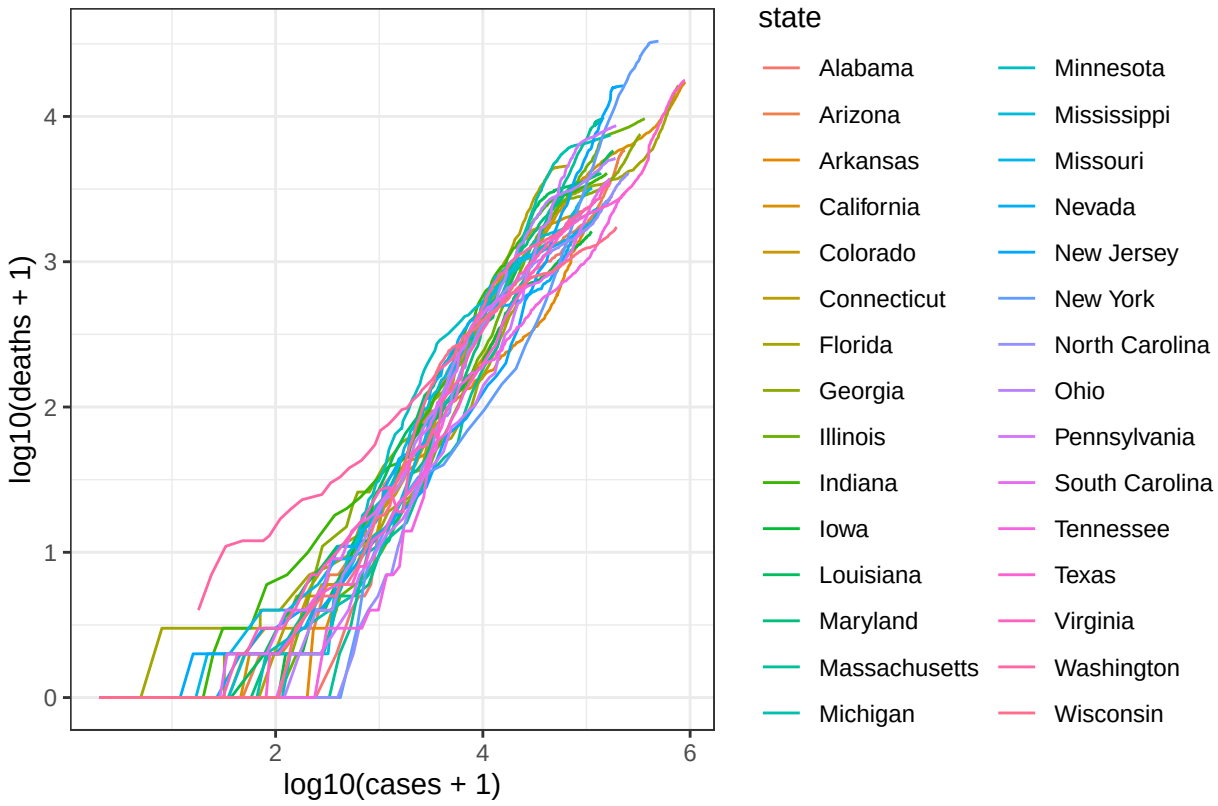


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



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

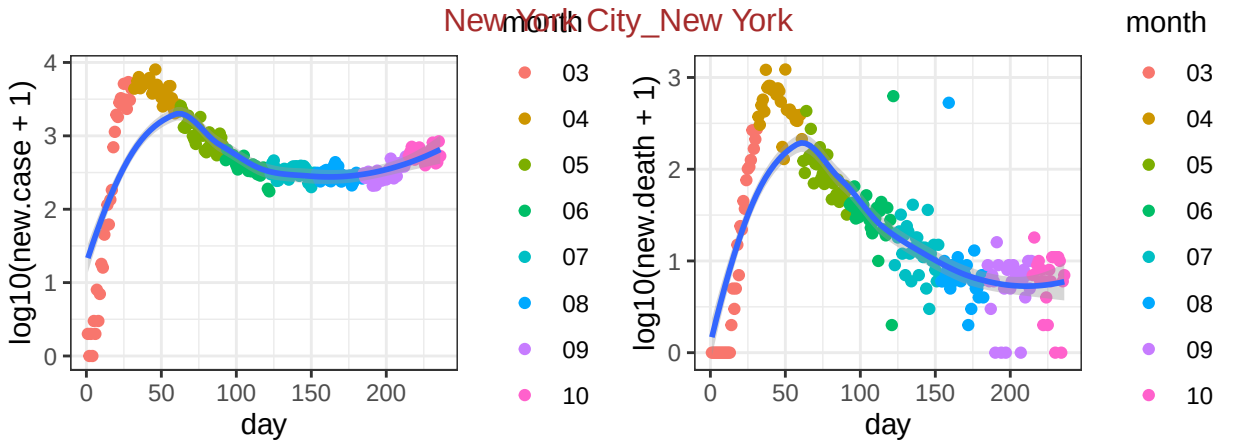
county level data

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

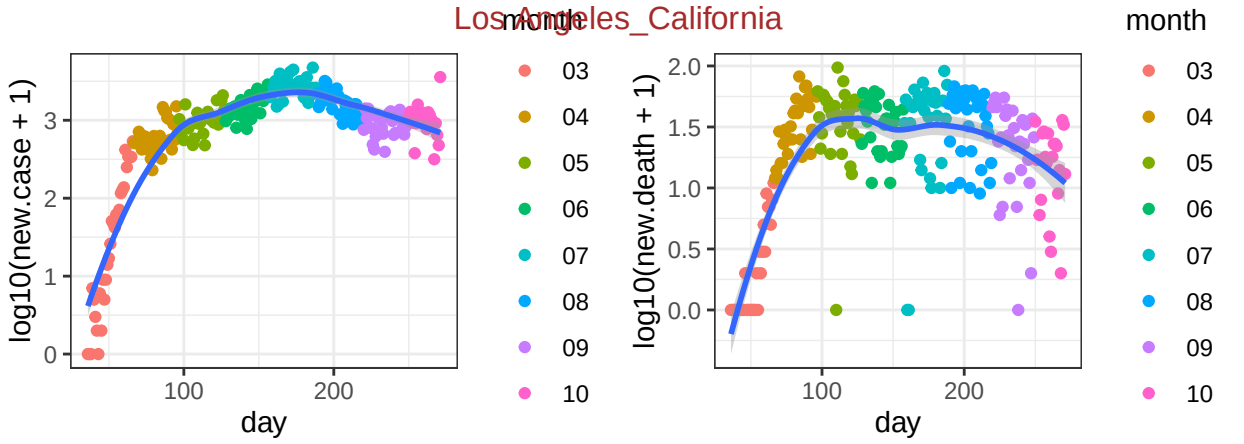
```
##              date          county         state  fips  cases deaths
## 656287 2020-10-22   New York City      New York    NA 261293  23955
## 654618 2020-10-22     Los Angeles    California  6037 294065   6956
## 655028 2020-10-22           Cook      Illinois 17031 170039   5390
## 654778 2020-10-22      Miami-Dade       Florida 12086 180496   3585
## 654516 2020-10-22        Maricopa       Arizona  4013 152101   3536
## 655738 2020-10-22          Wayne      Michigan 26163  39126   3042
## 657135 2020-10-22         Harris         Texas 48201 156742   2750
## 655649 2020-10-22      Middlesex Massachusetts 25017  30656   2240
## 656286 2020-10-22         Nassau      New York 36059  48941   2208
## 656210 2020-10-22          Essex    New Jersey 34013  23298   2139
## 656205 2020-10-22         Bergen    New Jersey 34003  24415   2056
## 656306 2020-10-22        Suffolk      New York 36103  48308   2019
## 657142 2020-10-22         Hidalgo         Texas 48215  34343   1899
## 656725 2020-10-22    Philadelphia  Pennsylvania 42101  41561   1871
## 654785 2020-10-22     Palm Beach       Florida 12099  49757   1549
## 656212 2020-10-22         Hudson    New Jersey 34017  22273   1525
## 654741 2020-10-22        Broward       Florida 12011  82250   1519
## 656179 2020-10-22          Clark        Nevada 32003  76817   1488
## 656314 2020-10-22     Westchester      New York 36119  39788   1470
## 654723 2020-10-22        Hartford   Connecticut  9003  16911   1459
## 656215 2020-10-22       Middlesex    New Jersey 34023  21525   1443
## 654629 2020-10-22          Orange    California  6059  59697   1434
```

```
## 654722 2020-10-22       Fairfield   Connecticut  9001  22088   1429
## 657049 2020-10-22           Bexar         Texas 48029  64026   1385
## 656223 2020-10-22           Union    New Jersey 34039  19328   1367
## 655645 2020-10-22           Essex Massachusetts 25009  22596   1333
## 654632 2020-10-22       Riverside    California  6065  65386   1279
## 656219 2020-10-22         Passaic    New Jersey 34031  20352   1260
## 655718 2020-10-22         Oakland      Michigan 26125  23827   1233
## 657091 2020-10-22          Dallas         Texas 48113  96483   1205
## 655653 2020-10-22         Suffolk Massachusetts 25025  26954   1166
## 655655 2020-10-22       Worcester Massachusetts 25027  15778   1156
## 654726 2020-10-22       New Haven   Connecticut  9009  15952   1122
## 655651 2020-10-22         Norfolk Massachusetts 25021  11336   1097
## 655705 2020-10-22          Macomb      Michigan 26099  17802   1079
## 657065 2020-10-22         Cameron         Texas 48061  23937   1075
## 654635 2020-10-22  San Bernardino    California  6071  61550   1070
## 656218 2020-10-22           Ocean    New Jersey 34029  16368   1061
## 655766 2020-10-22        Hennepin     Minnesota 27053  33024    980
## 656824 2020-10-22      Providence  Rhode Island 44007  21543    930
## 656720 2020-10-22      Montgomery  Pennsylvania 42091  13346    893
## 656216 2020-10-22        Monmouth    New Jersey 34025  13647    870
## 654636 2020-10-22       San Diego    California  6073  53561    866
## 656012 2020-10-22       St. Louis      Missouri 29189  28842    865
## 655631 2020-10-22      Montgomery      Maryland 24031  24812    863
## 655632 2020-10-22 Prince George's      Maryland 24033  32216    852
## 655164 2020-10-22          Marion       Indiana 18097  25324    839
## 656217 2020-10-22          Morris    New Jersey 34027   8803    834
## 657485 2020-10-22            King    Washington 53033  25759    817
## 655652 2020-10-22        Plymouth Massachusetts 25023  10895    814
```

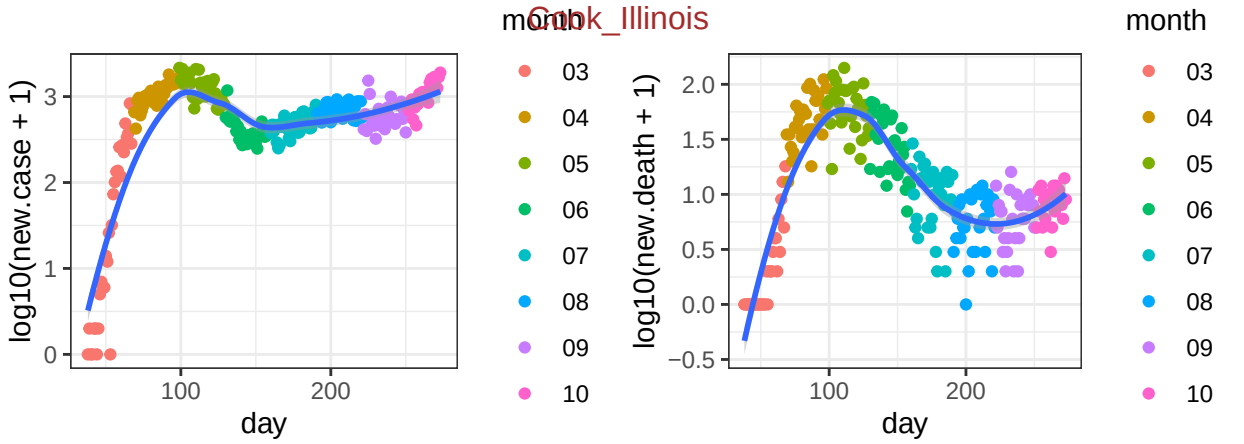
For these 50 counties, I check the number of new cases and the number of new deaths.



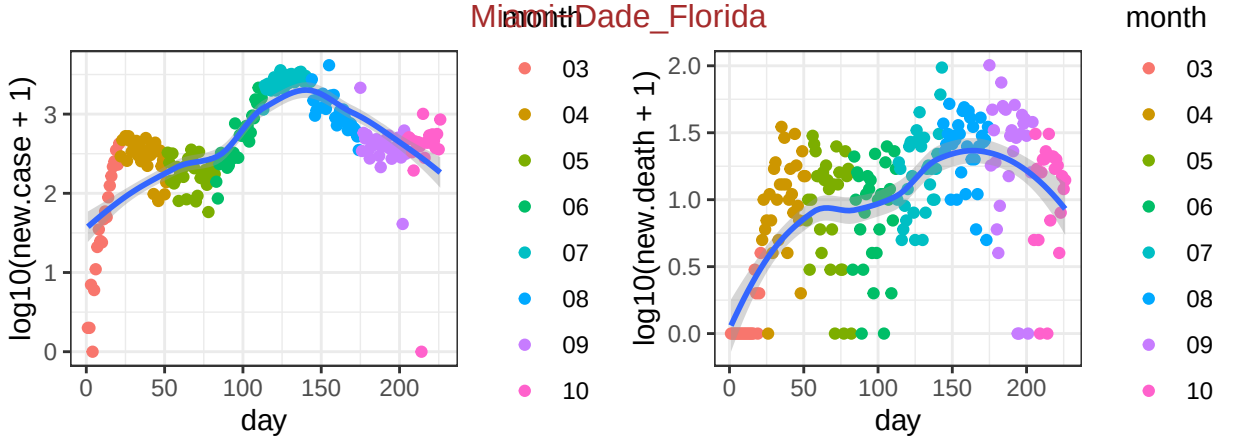
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



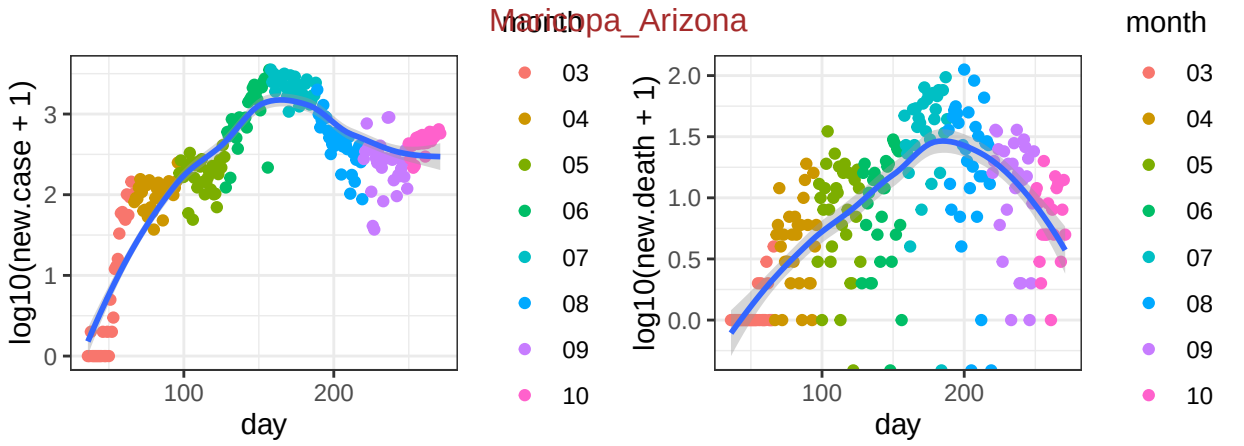
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



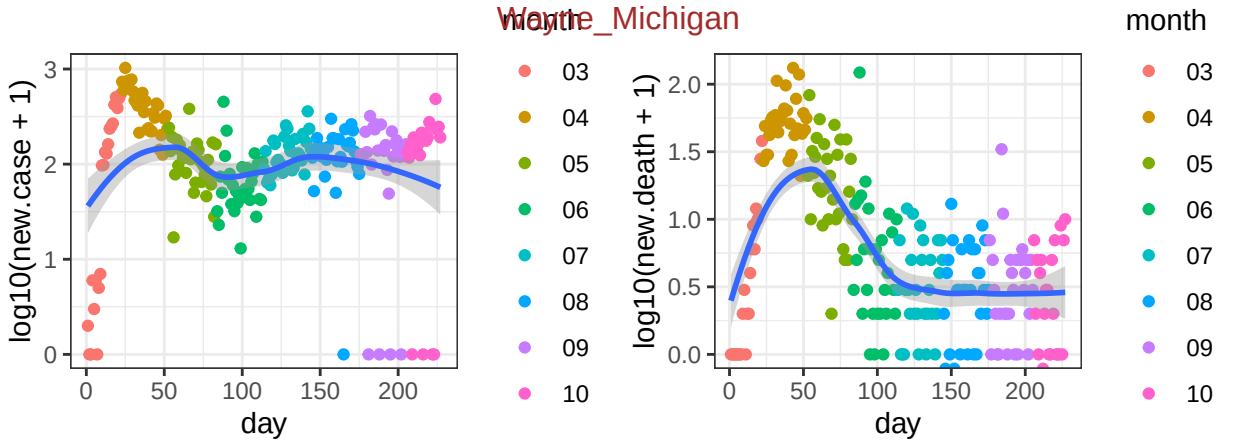
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



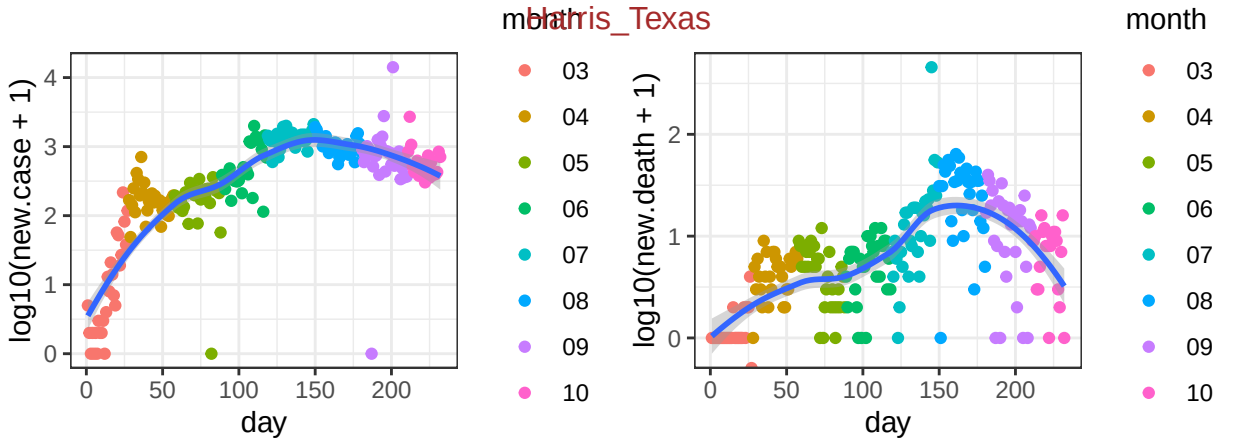
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11



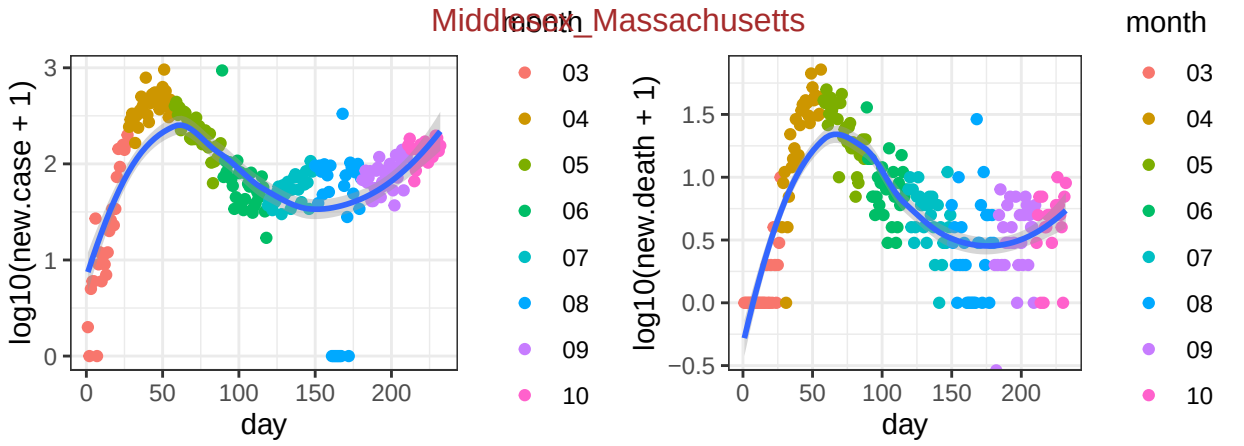
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



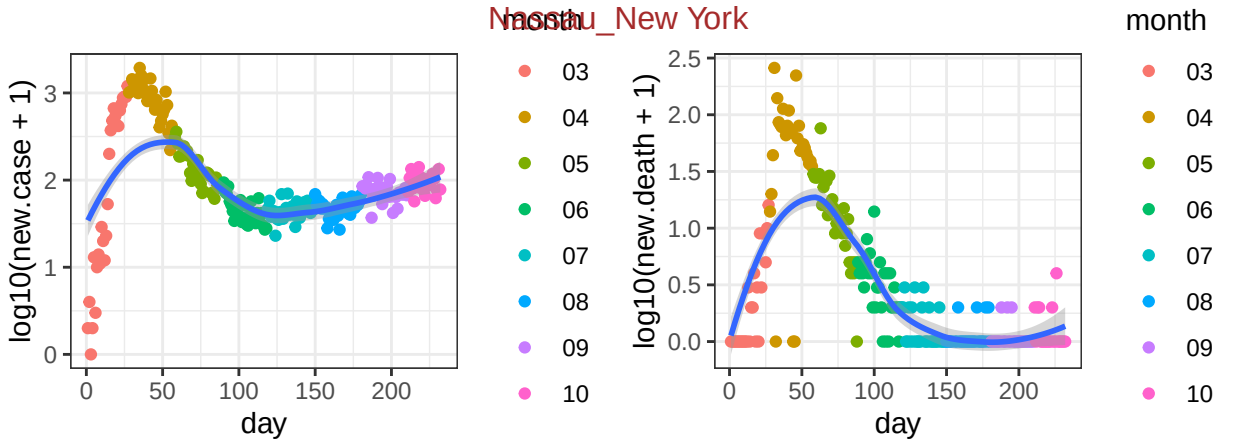
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10



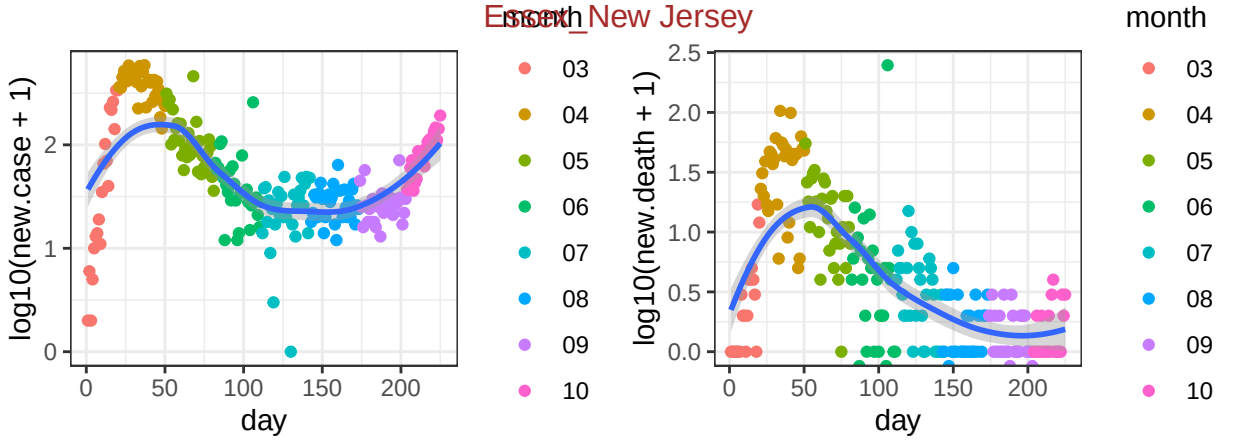
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



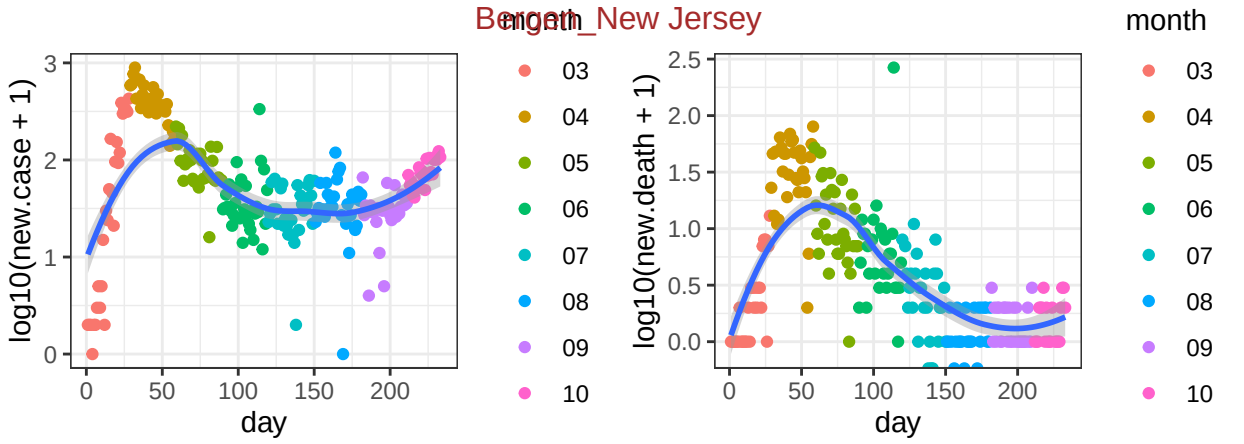
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



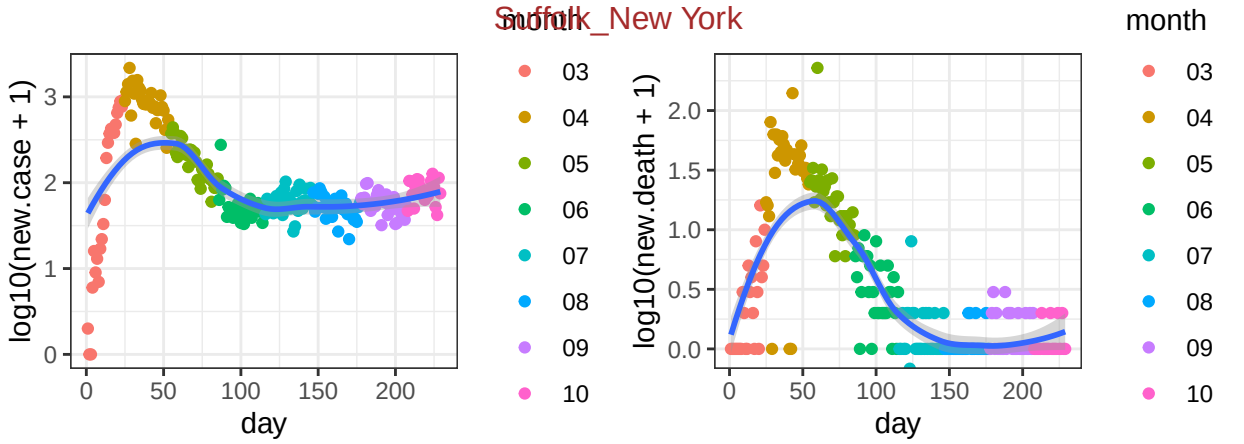
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



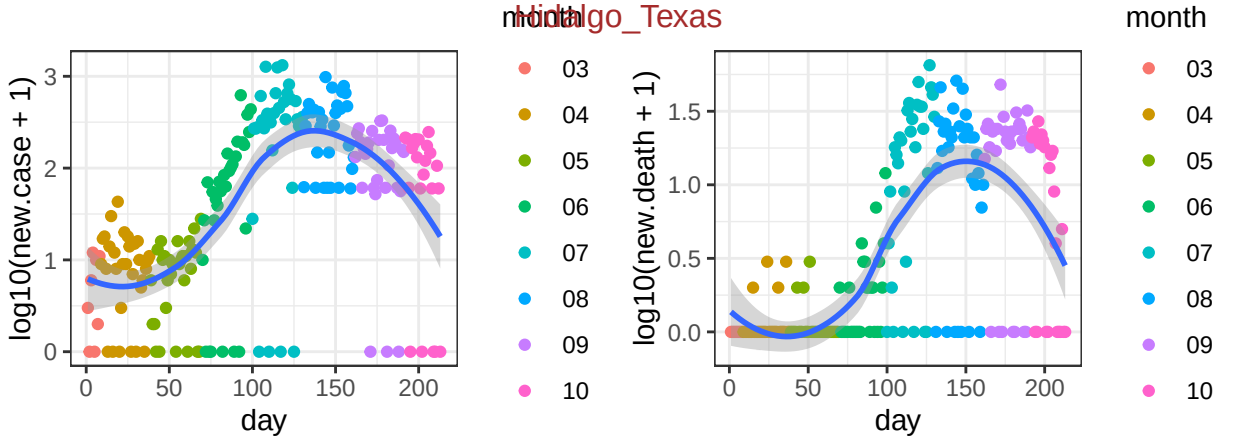
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12



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

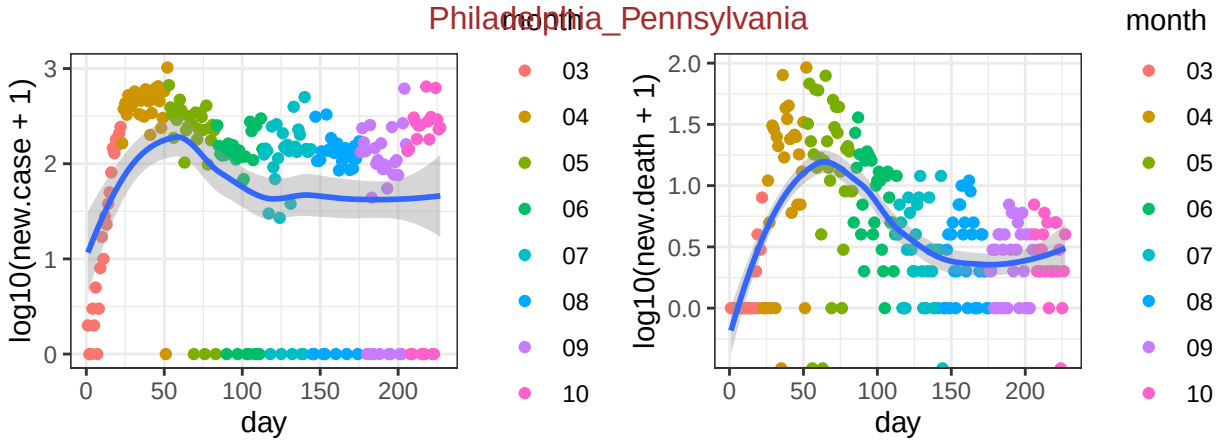


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08



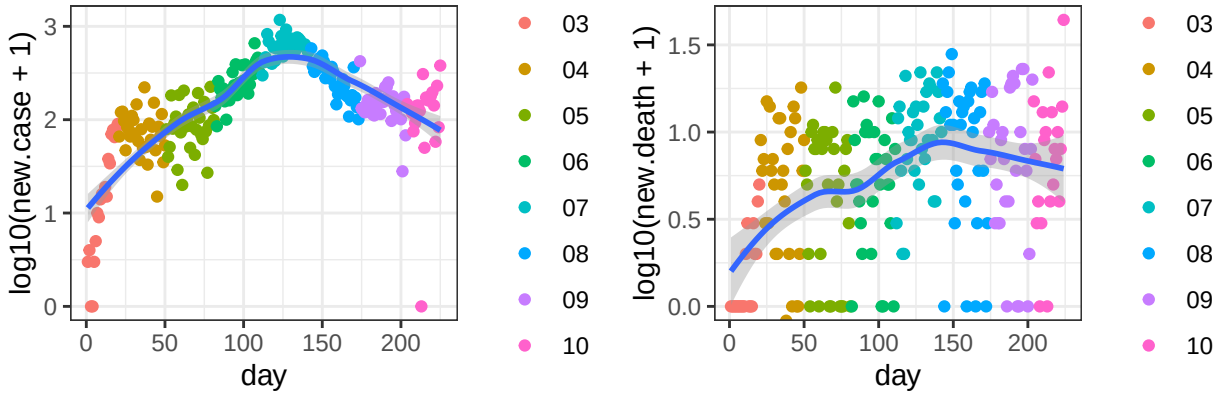
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−24

Philadelphia_Pennsylvania



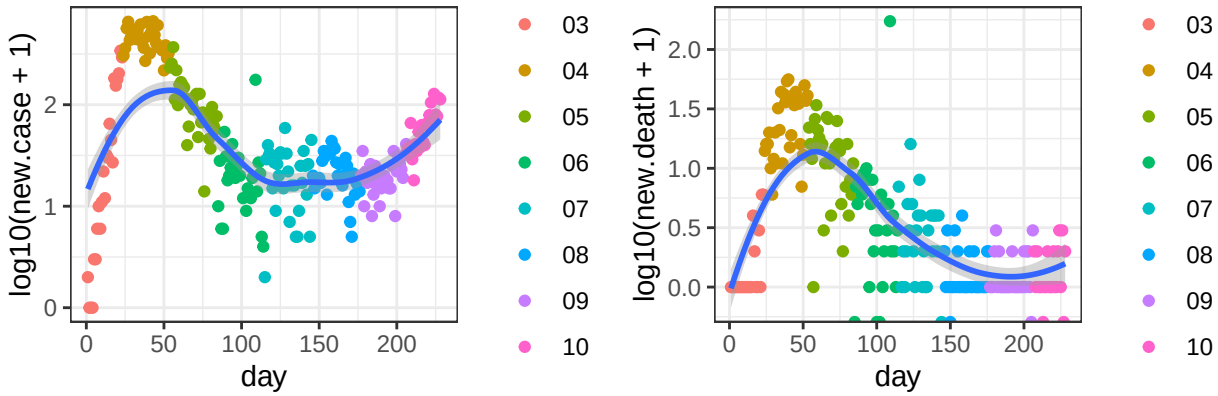
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Palm Beach_Florida



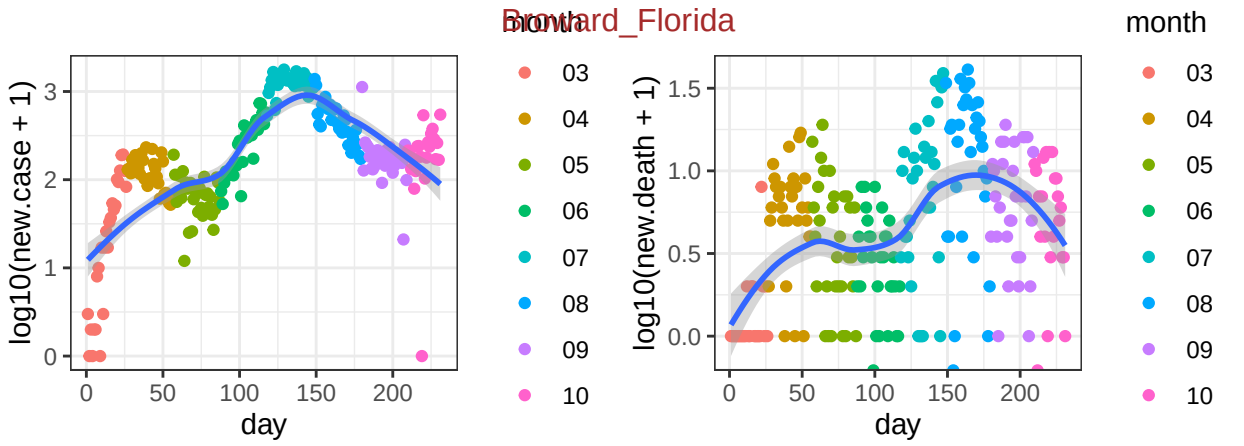
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Hudson_New Jersey



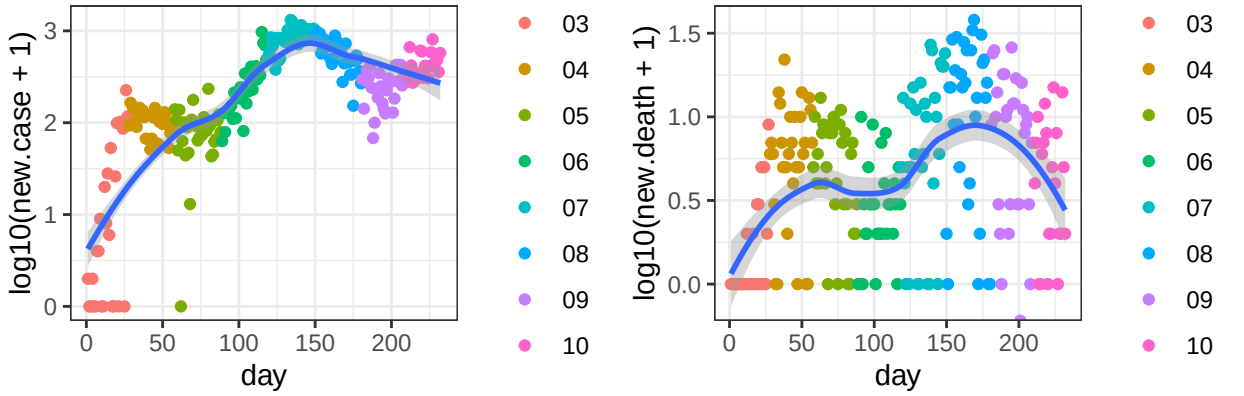
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Broward_Florida



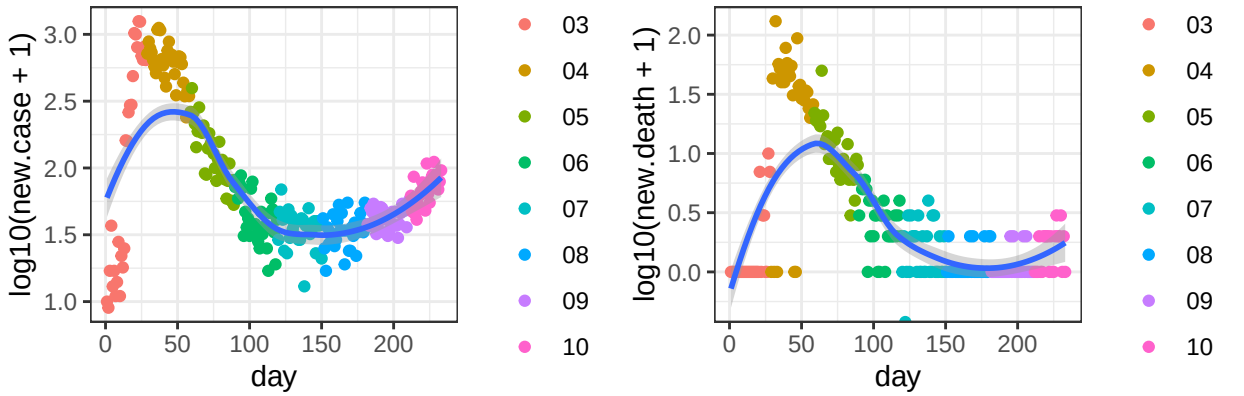
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Clark_Nevada

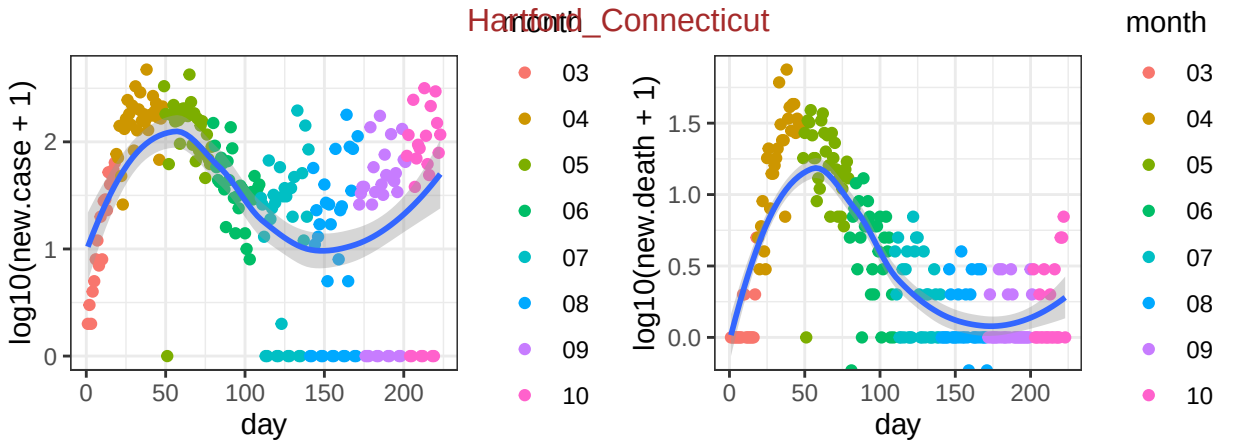


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

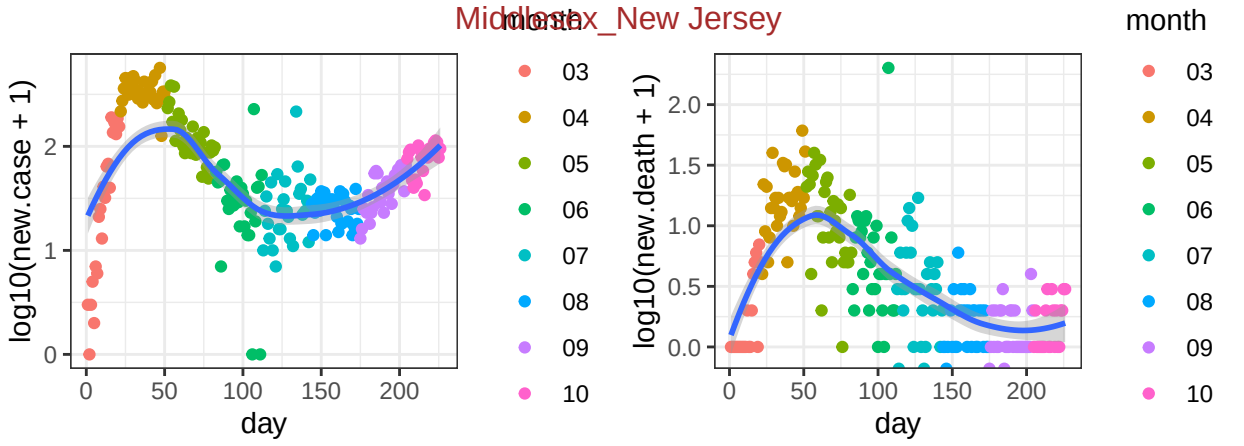
Westchester_New York



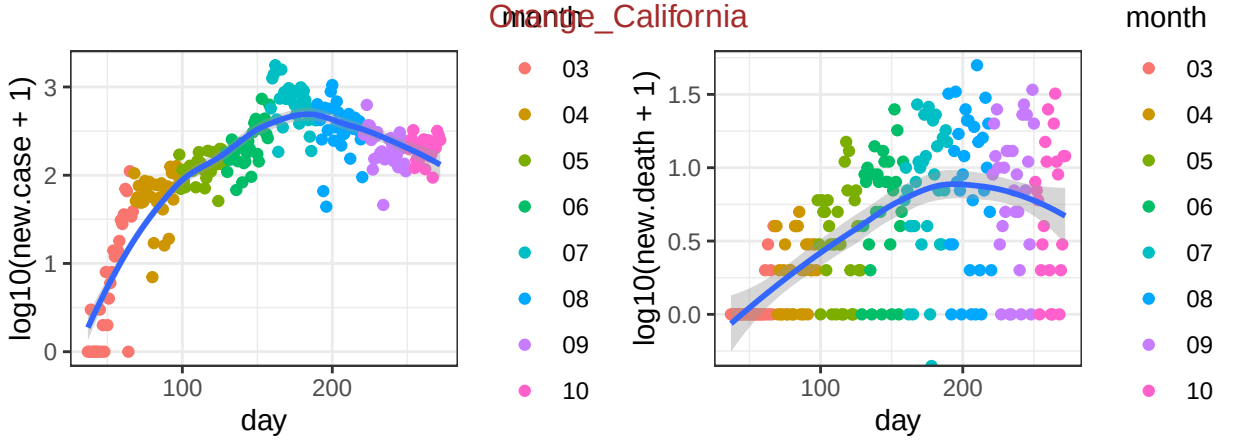
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04



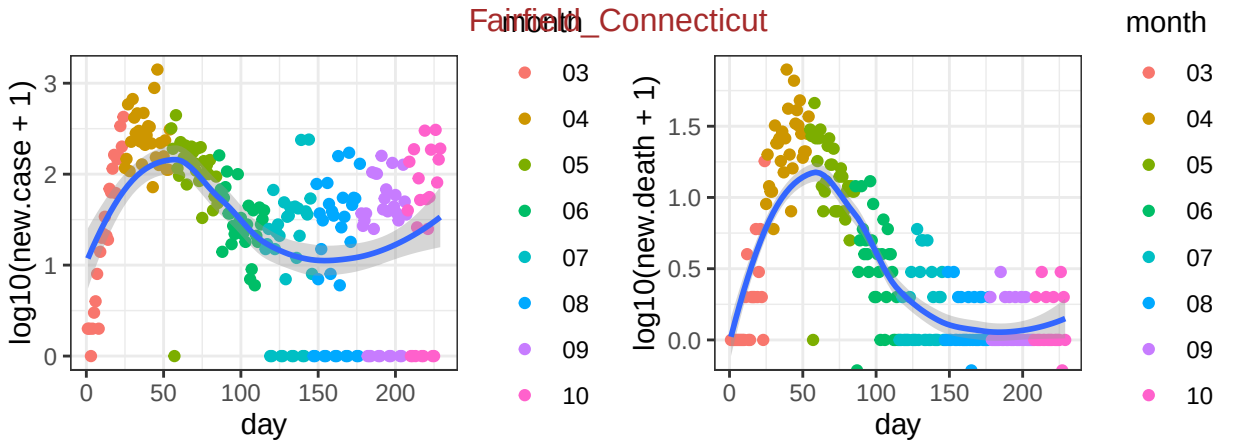
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14



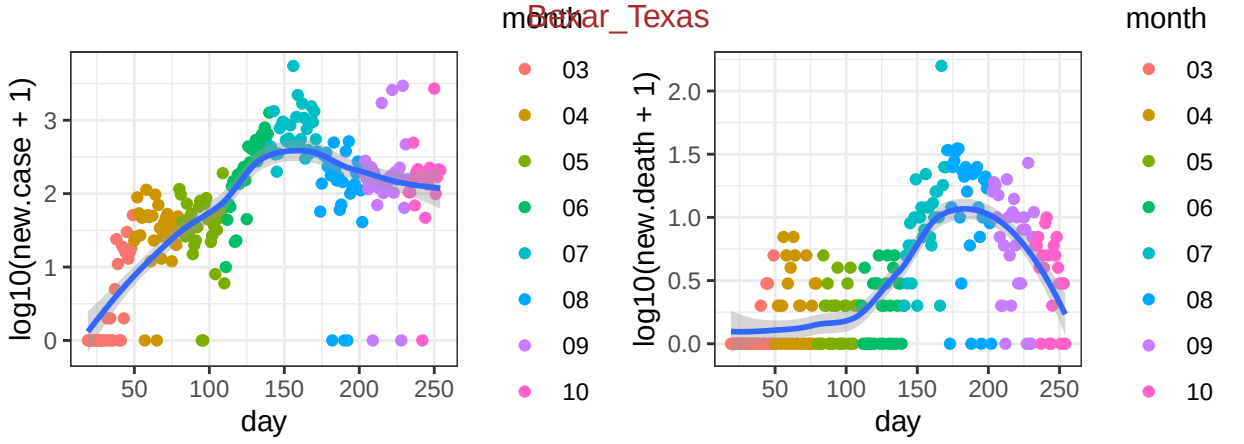
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11



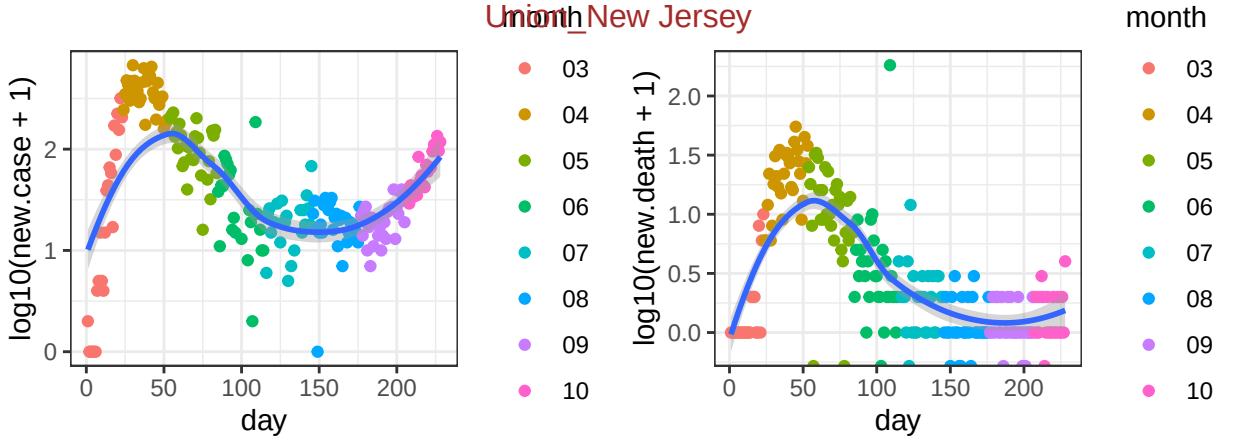
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



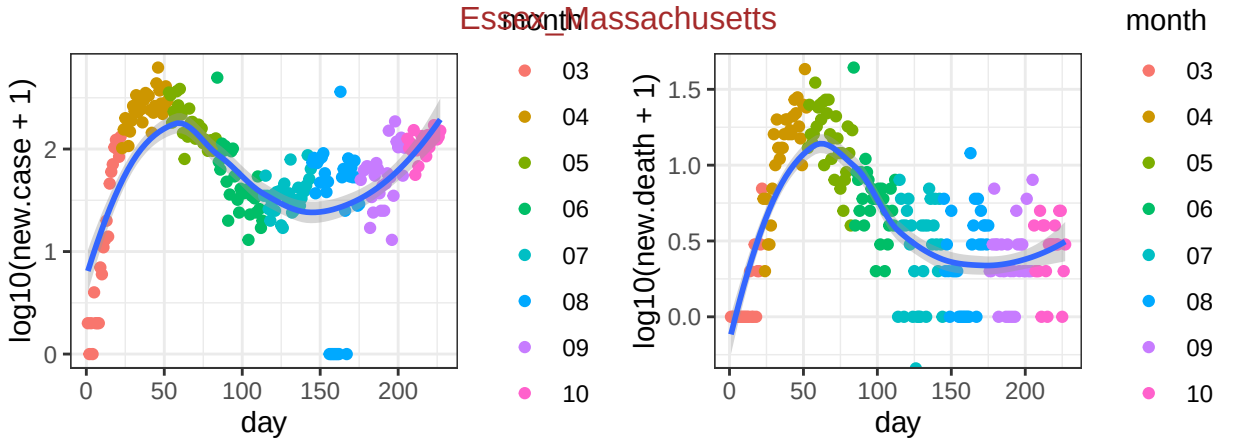
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08



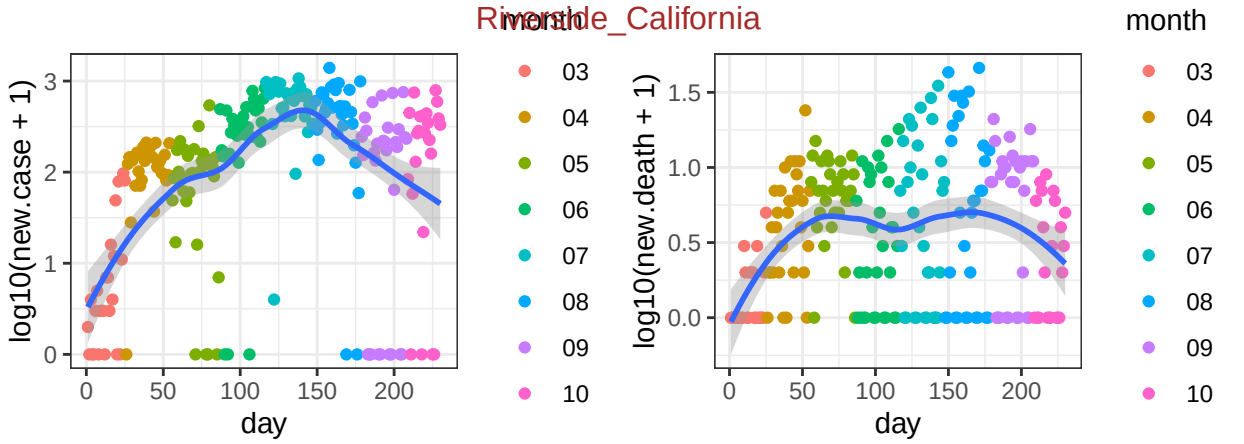
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



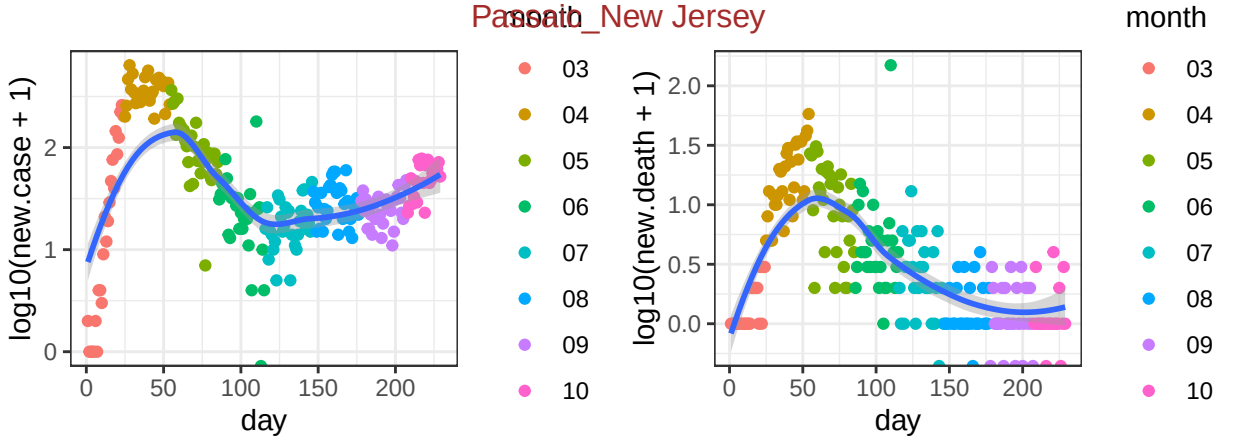
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09



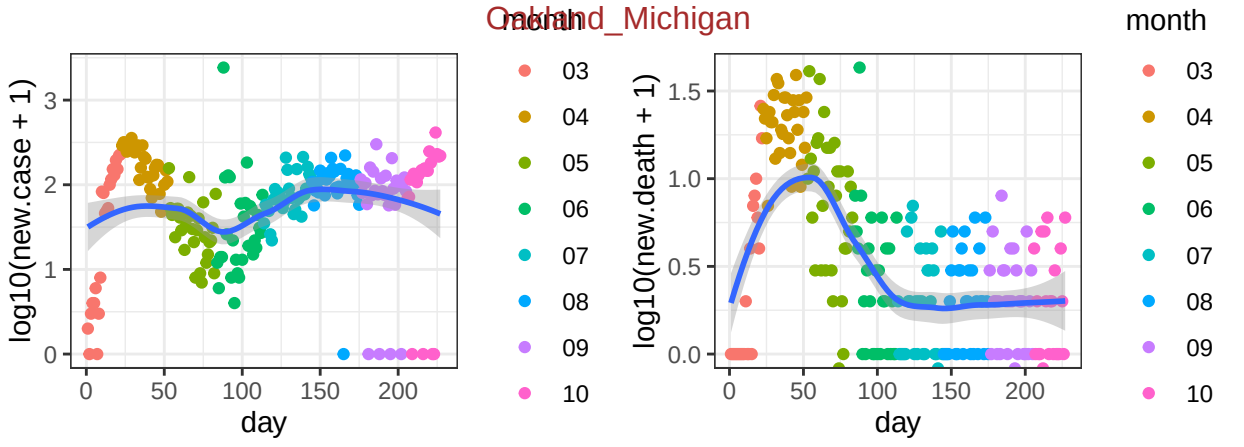
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10



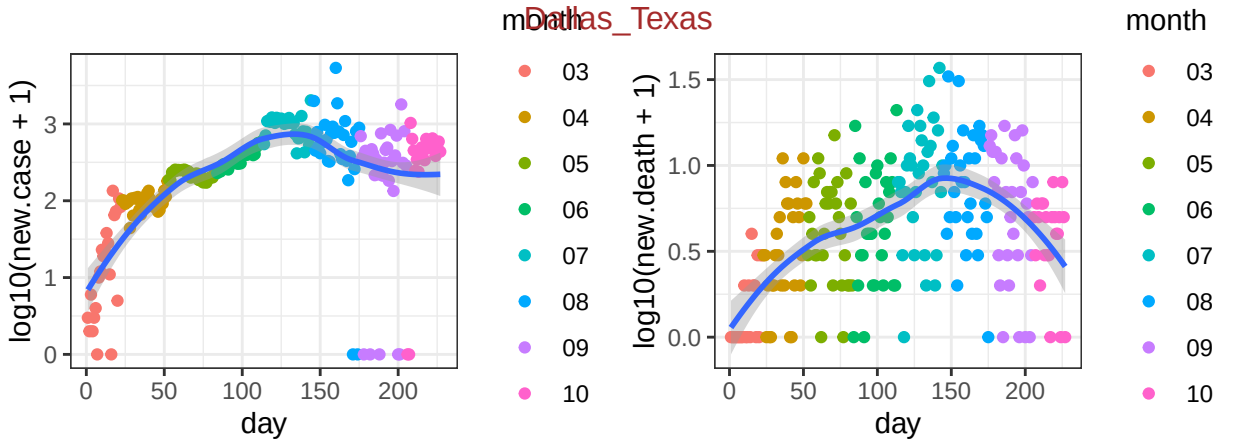
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07



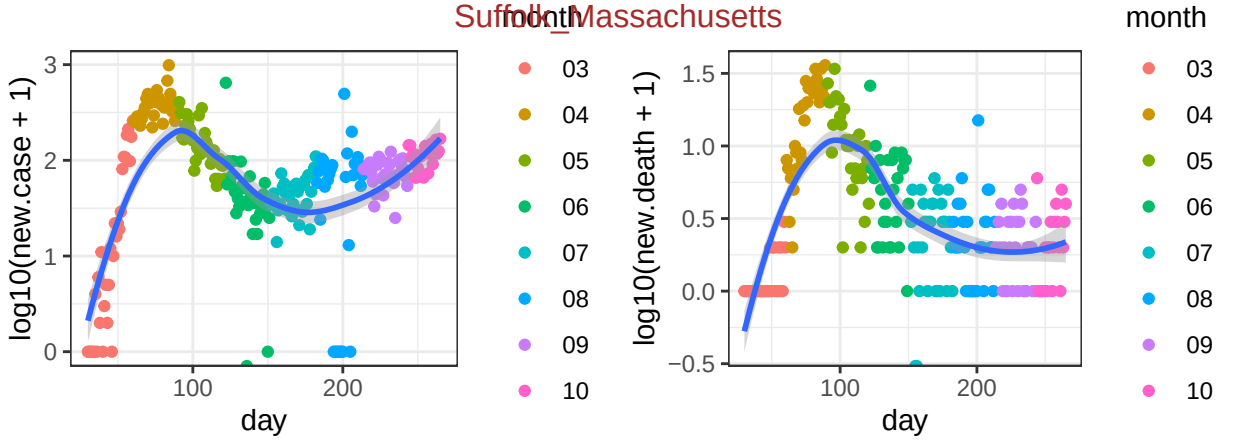
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08



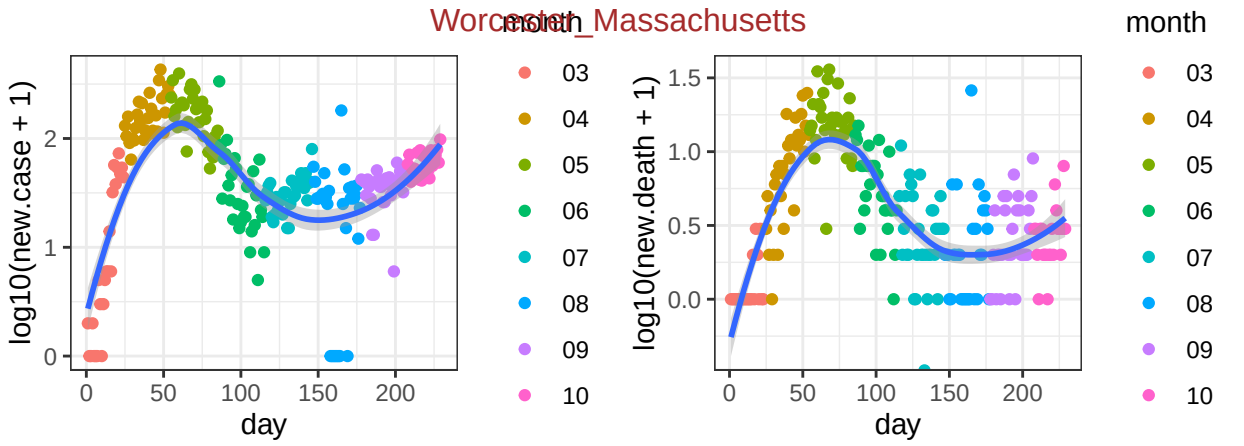
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10



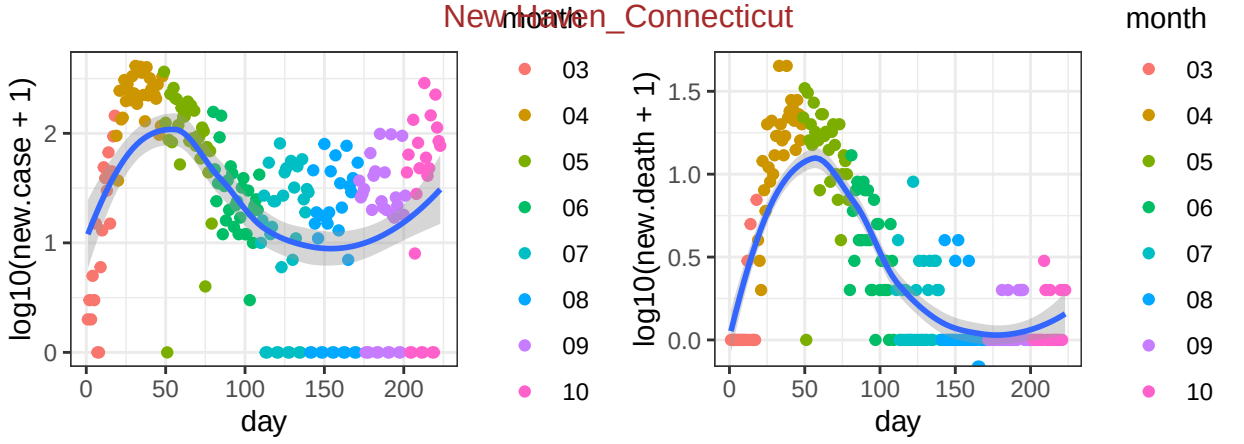
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10



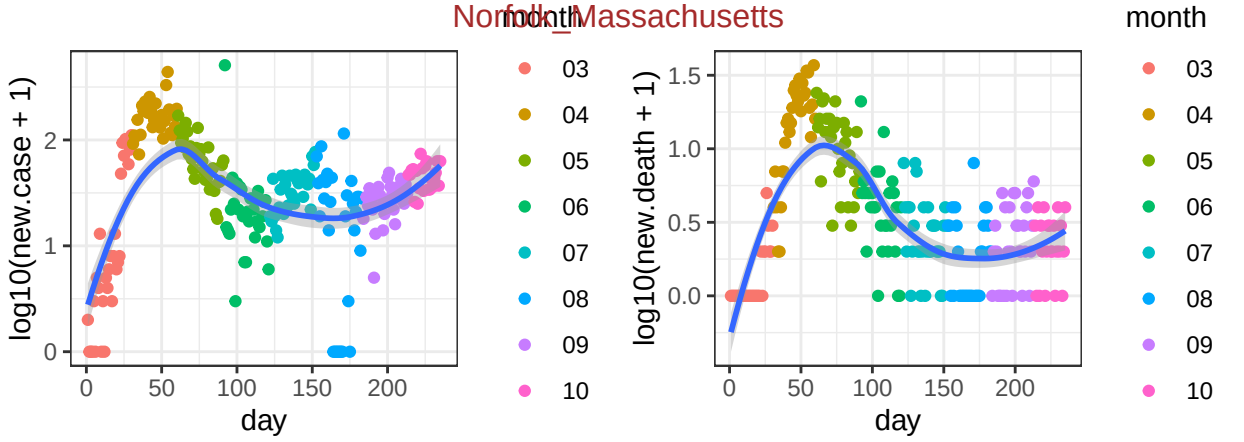
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



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

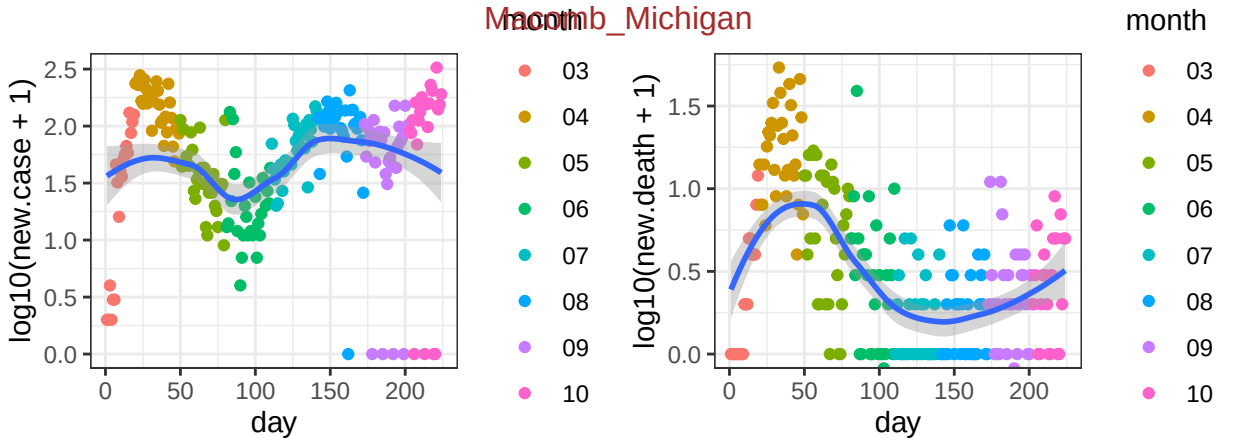


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14



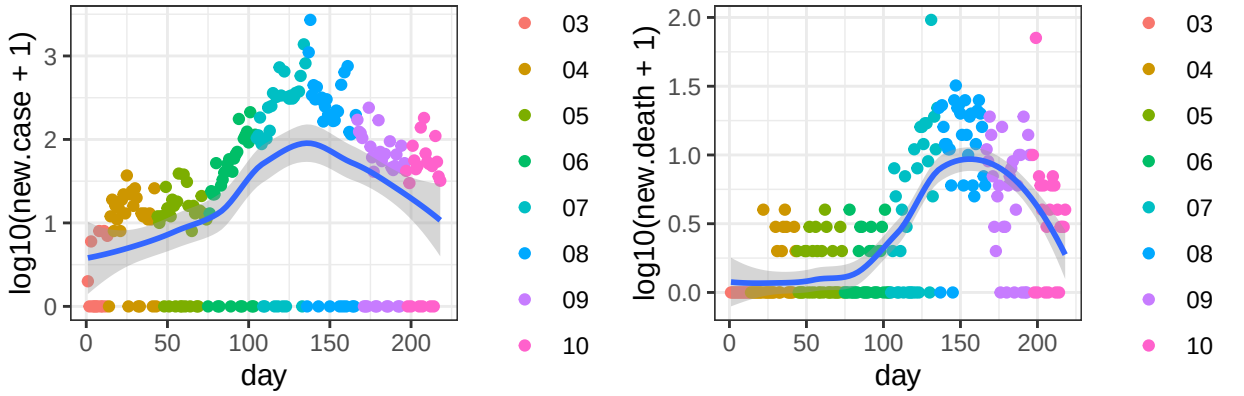
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Macomb_Michigan



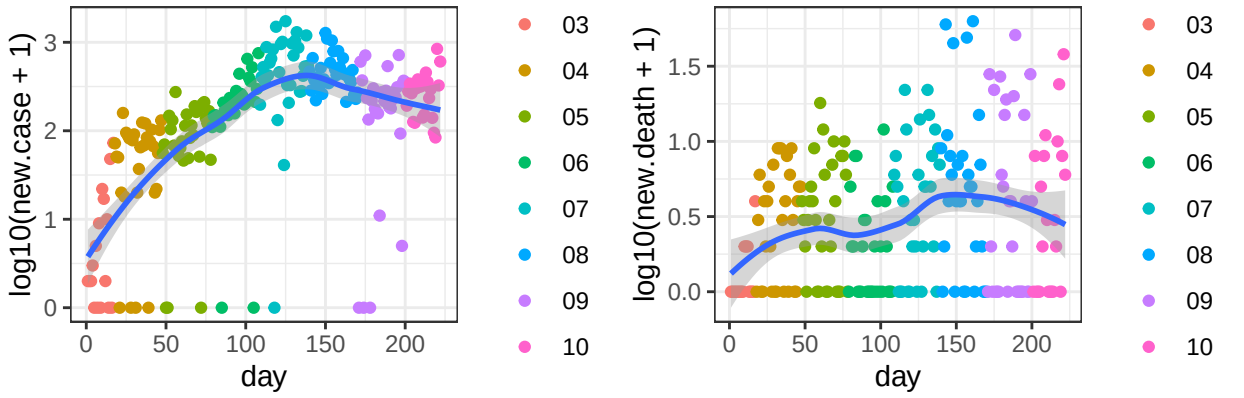
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Cameron_Texas

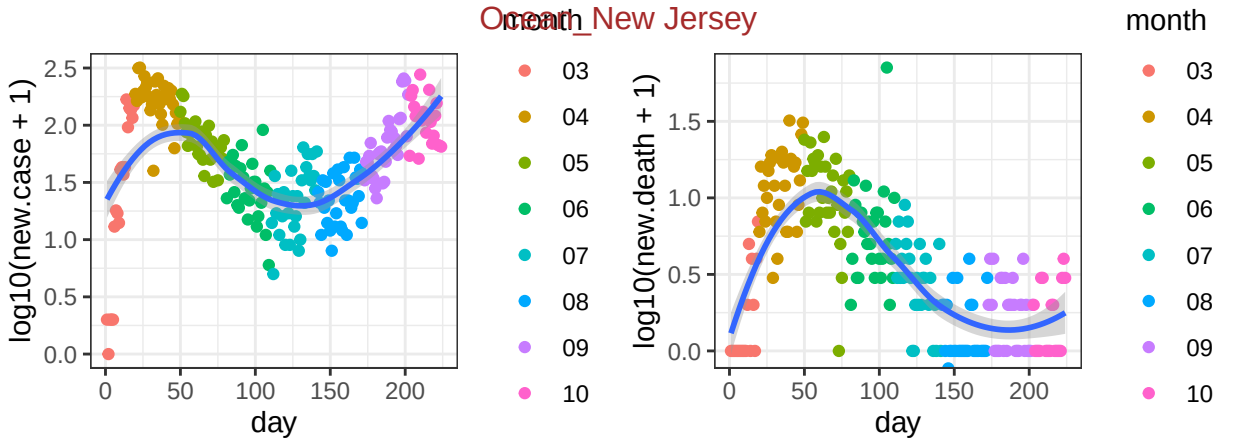


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−19

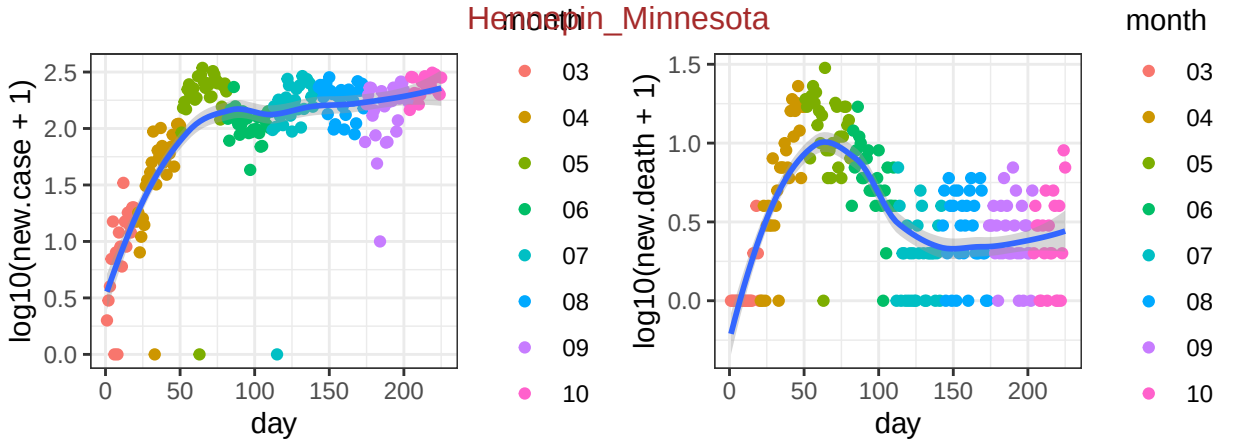
San Bernardino_California



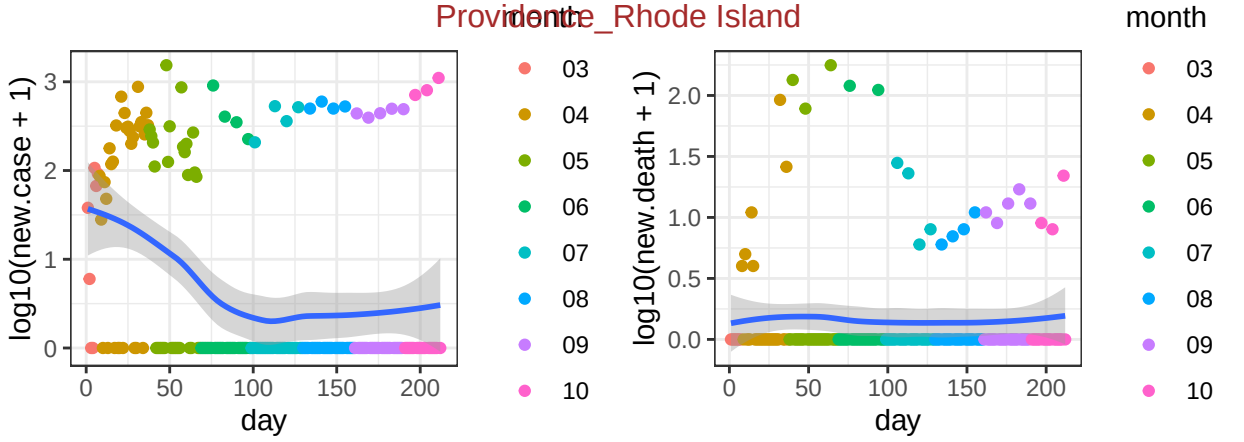
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15



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

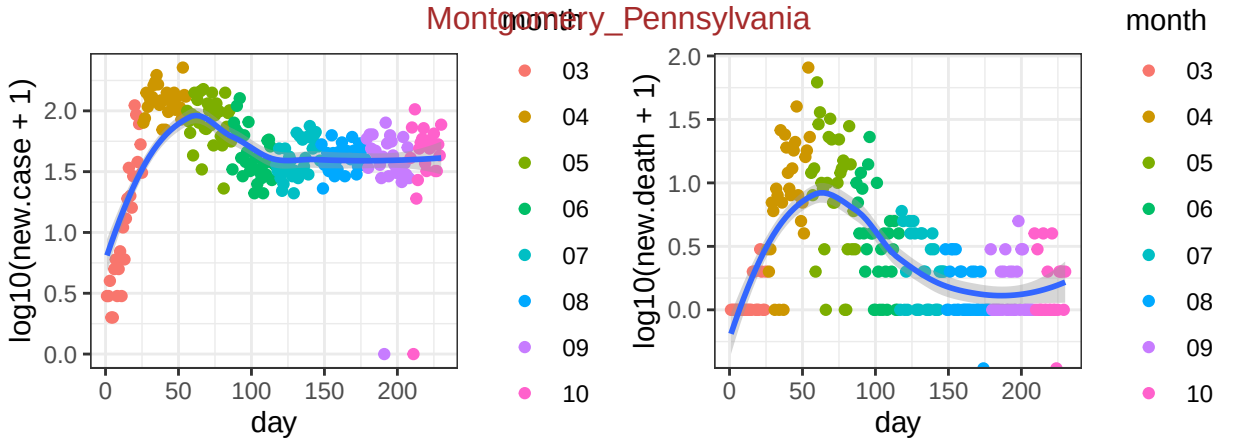


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12



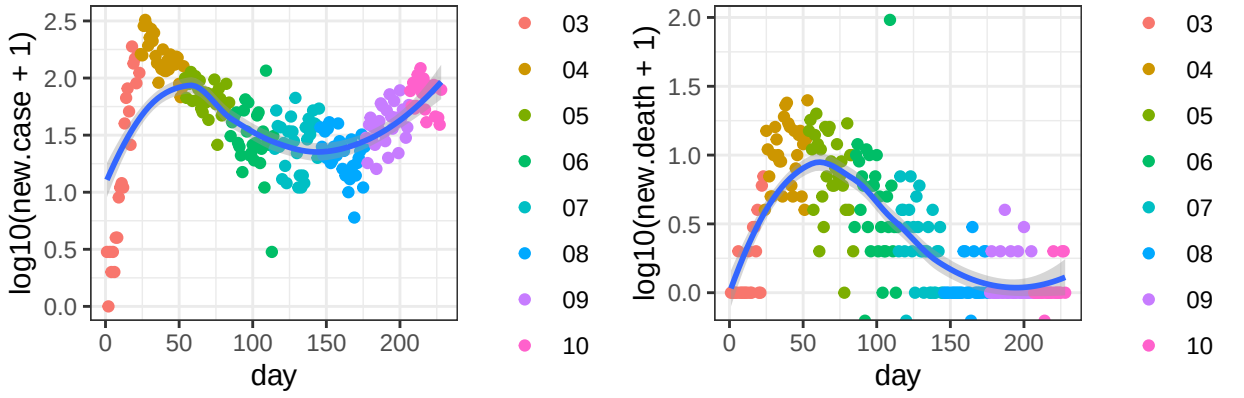
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Montgomery_Pennsylvania



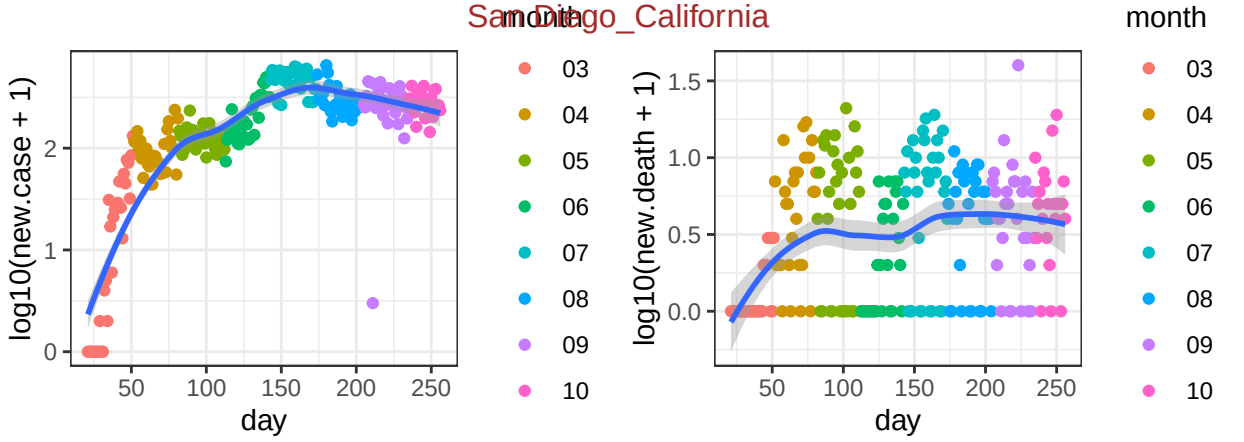
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Monmouth_New Jersey



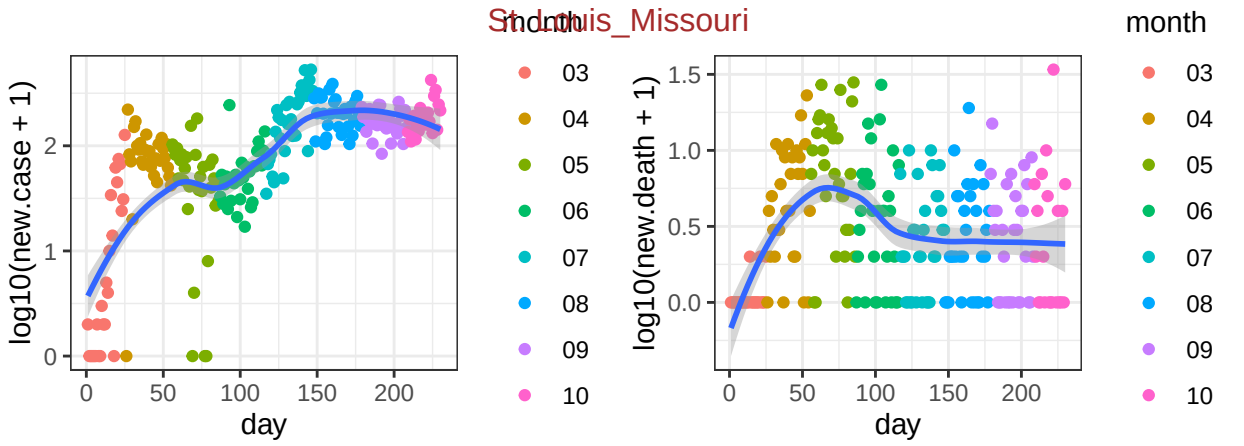
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

San Diego_California



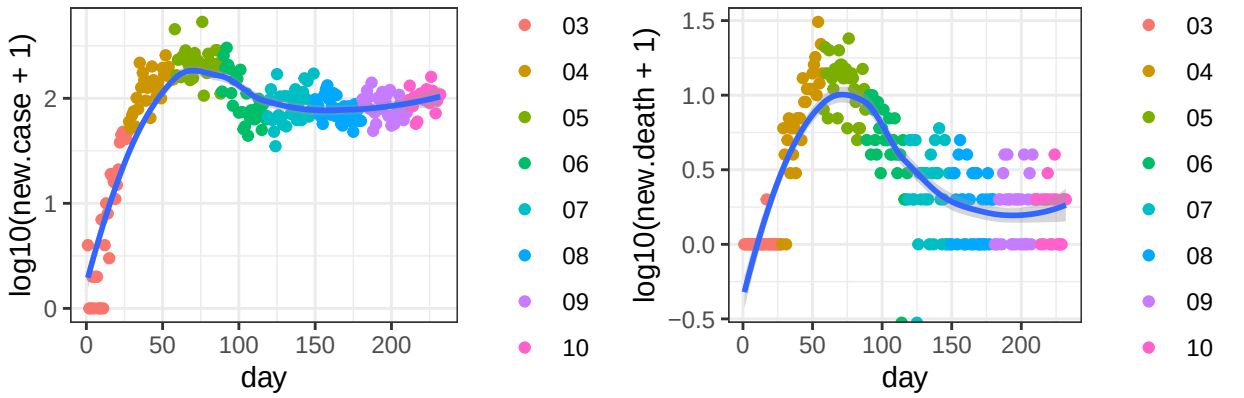
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

St Louis_Missouri



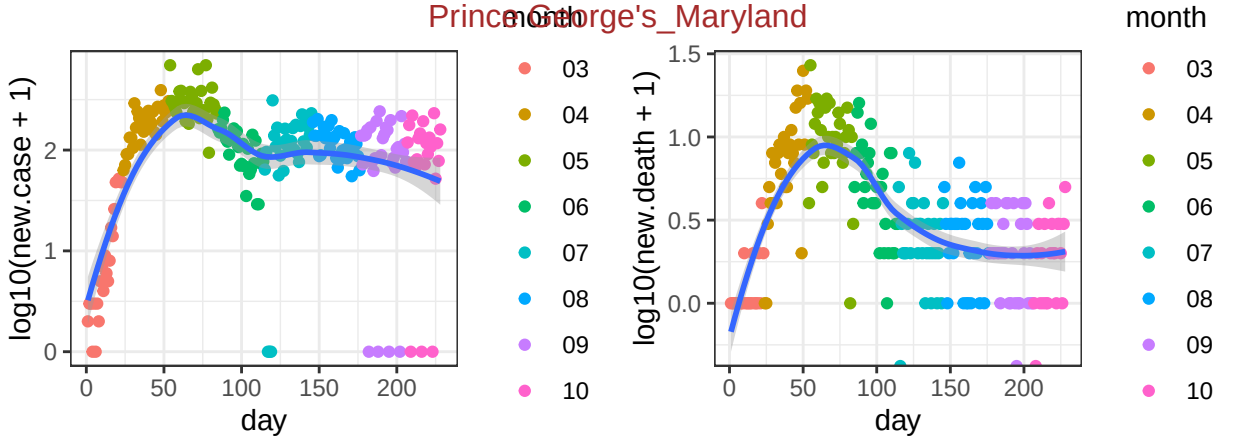
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Montgomery_Maryland

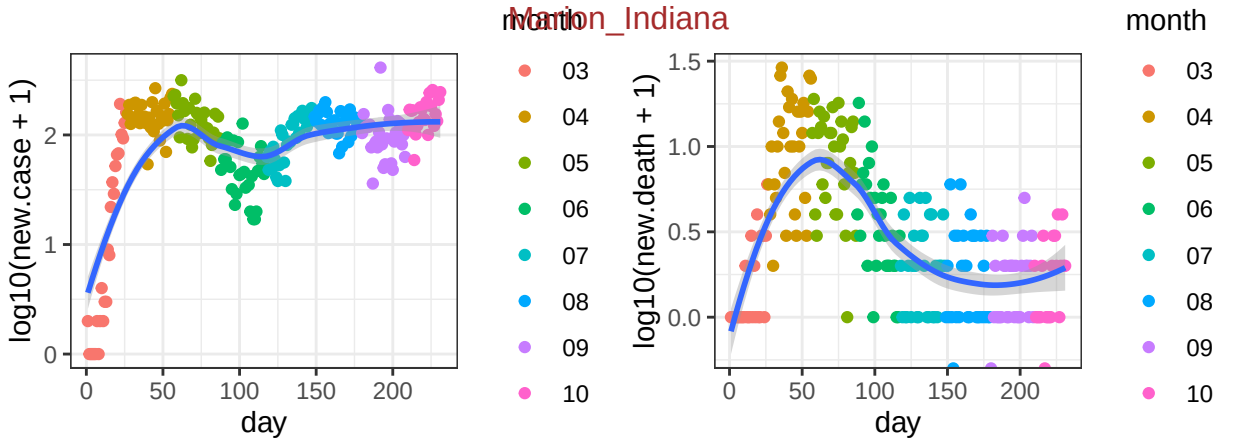


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

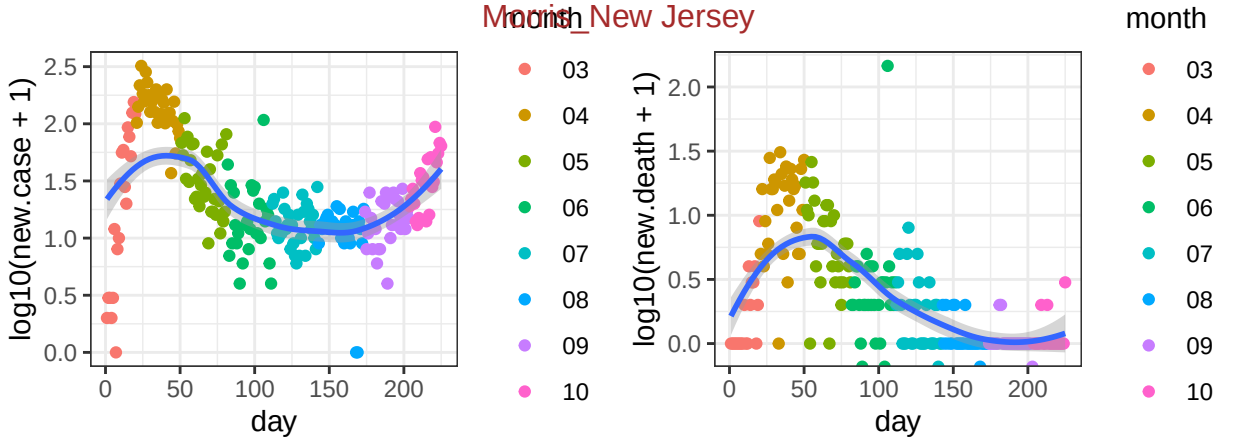
Prince George's_Maryland



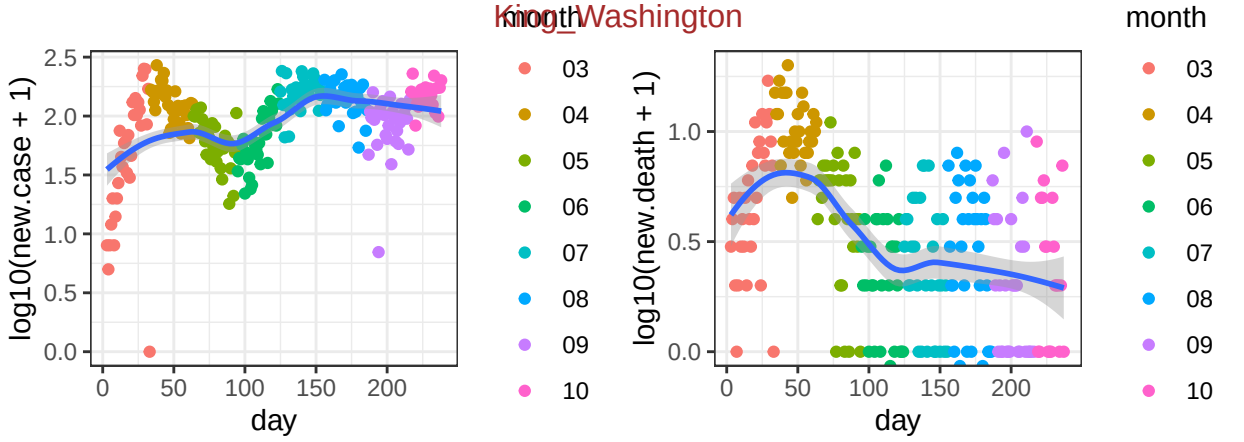
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09



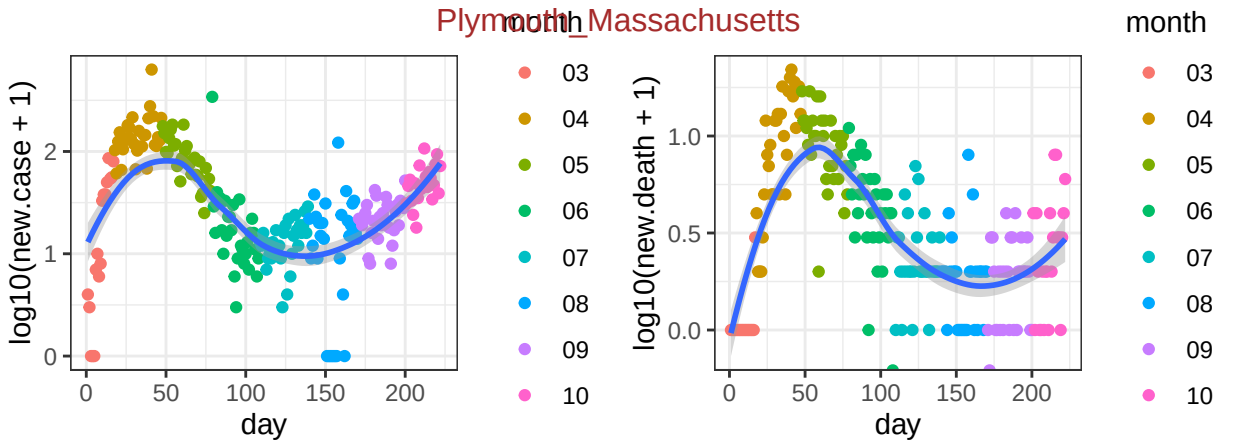
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06



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

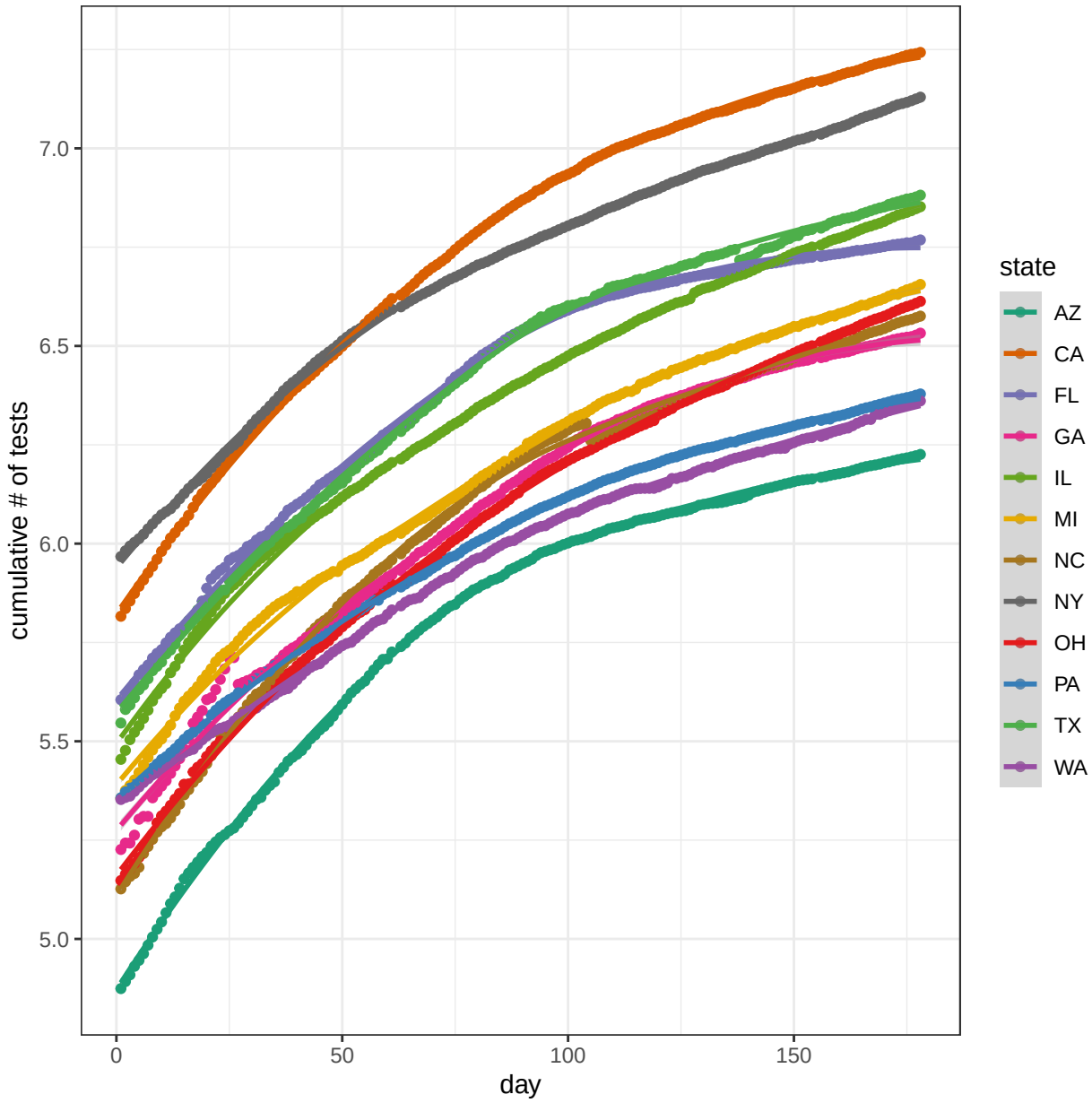


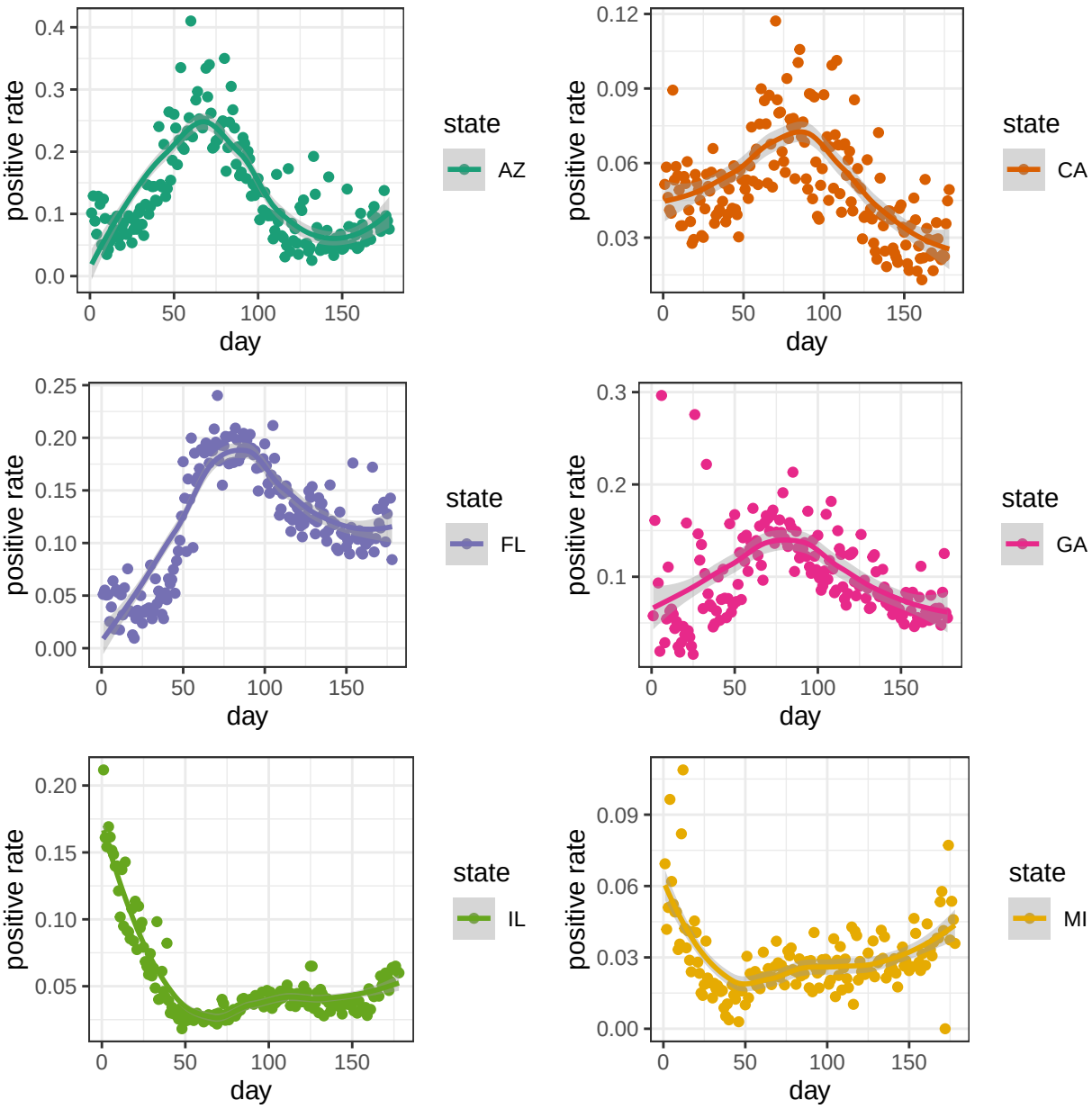
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

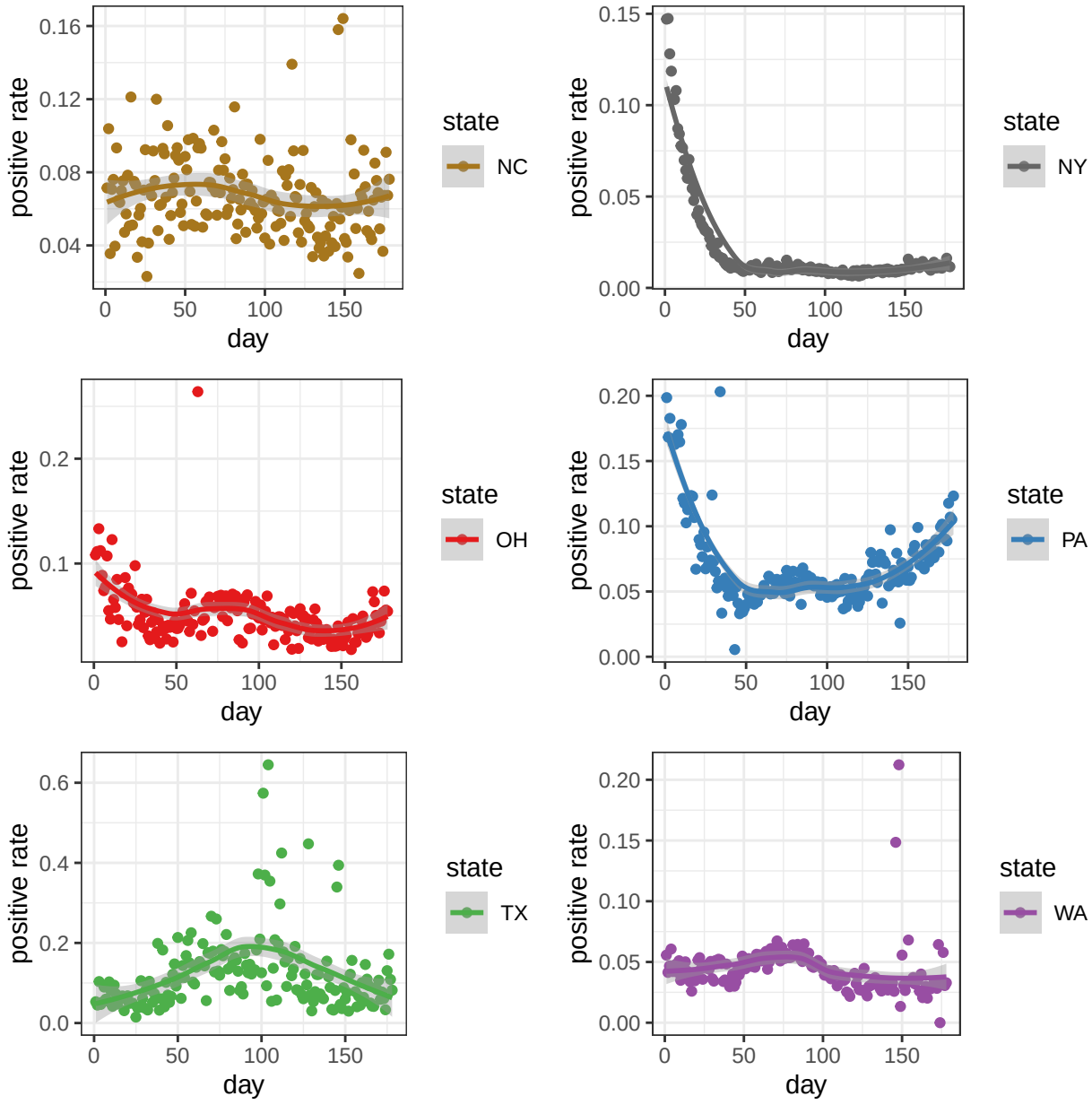
COVID Trackng

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
day
state
AZ
CA
FL
GA
IL
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
```