

Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-04-10

Contents

Introduction	1
JHU	2
time series data	2
daily reports data	6
NY Times	7
state level data	7
county level data	11
COVID Trackng	15
Session information	15

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 Tracking (<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>

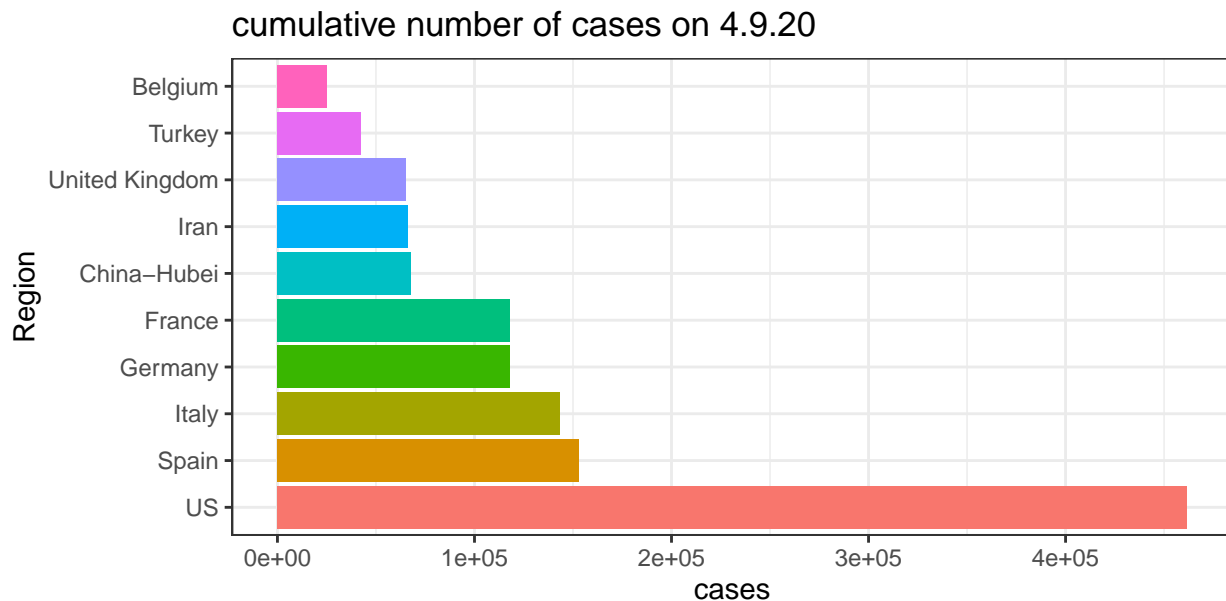
JHU

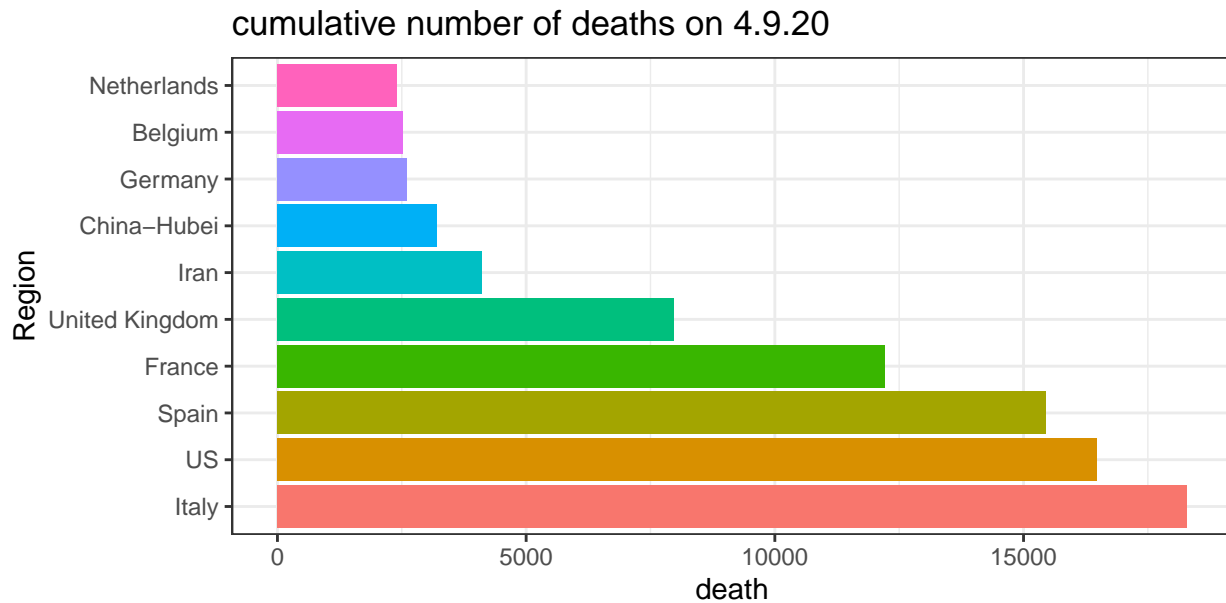
Assume you have cloned the JHU Github repository on your local machine at “../COVID-19”.

time series data

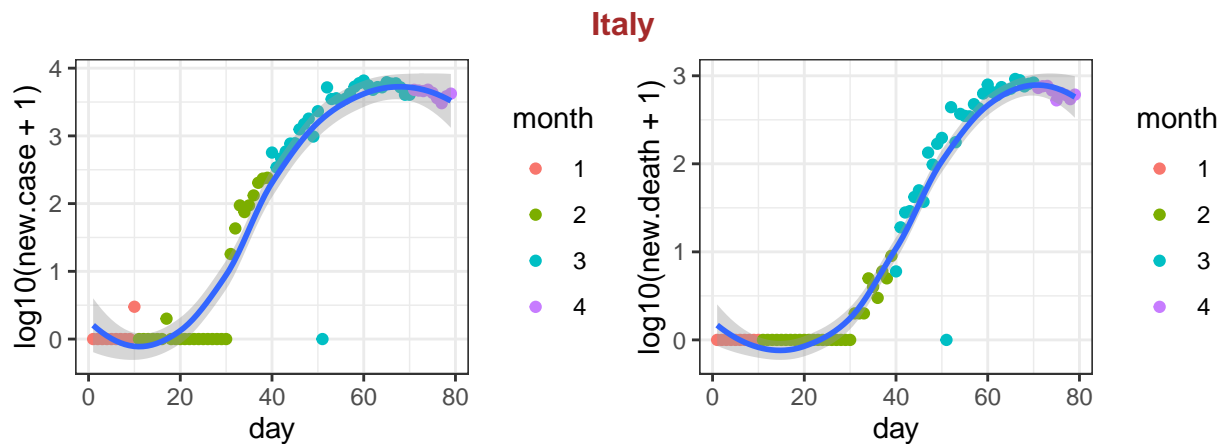
The time series provide counts (e.g., confirmed cases, deaths) starting from Jan 22nd, 2020 for 253 locations. Currently there is no data of individual US state in these time series data files.

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

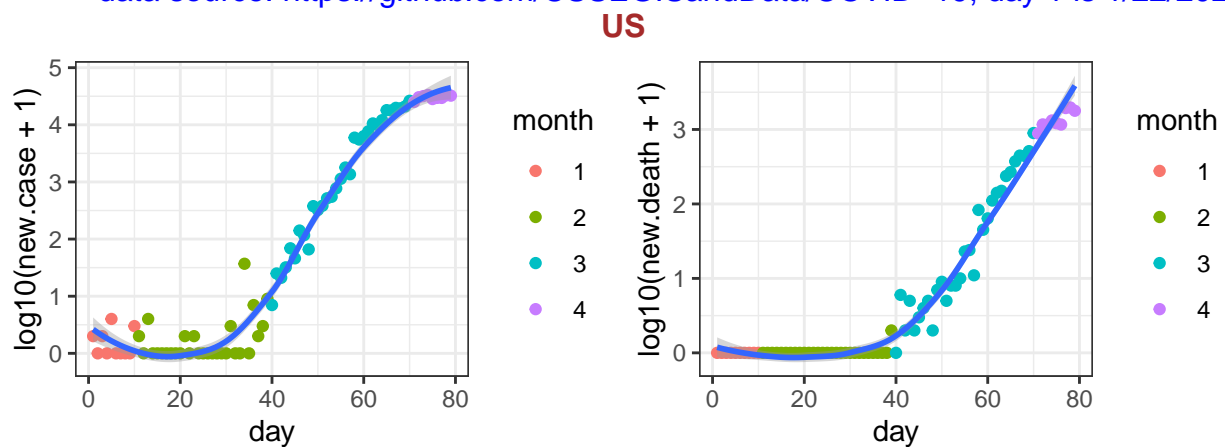




Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

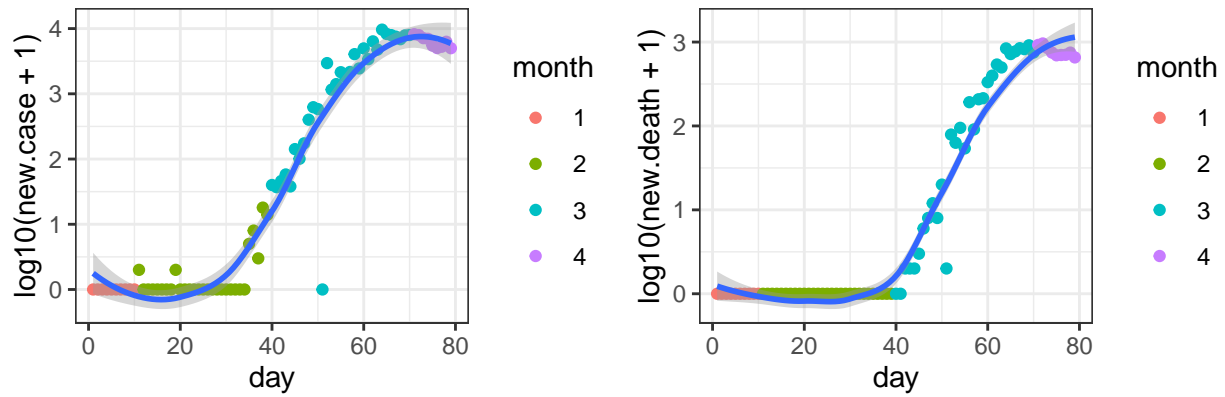


data source: <https://github.com/CSSEGISandData/COVID-19>, day 1 is 1/22/2020



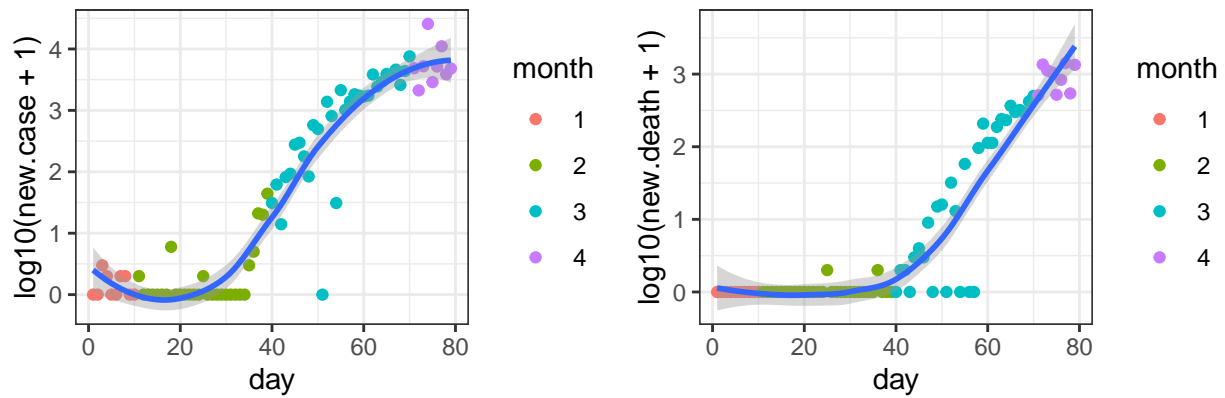
data source: <https://github.com/CSSEGISandData/COVID-19>, day 1 is 1/22/2020

Spain



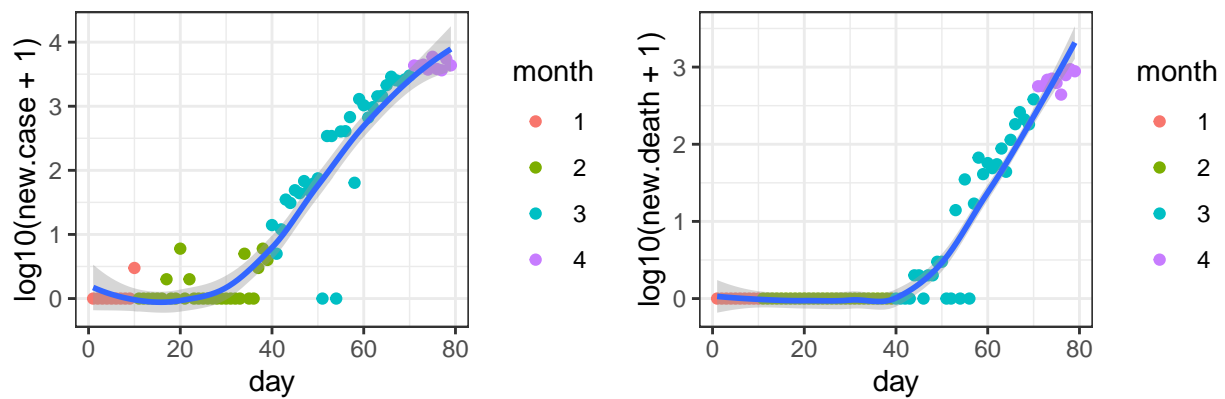
data source: <https://github.com/CSSEGISandData/COVID-19>, day 1 is 1/22/2020

France



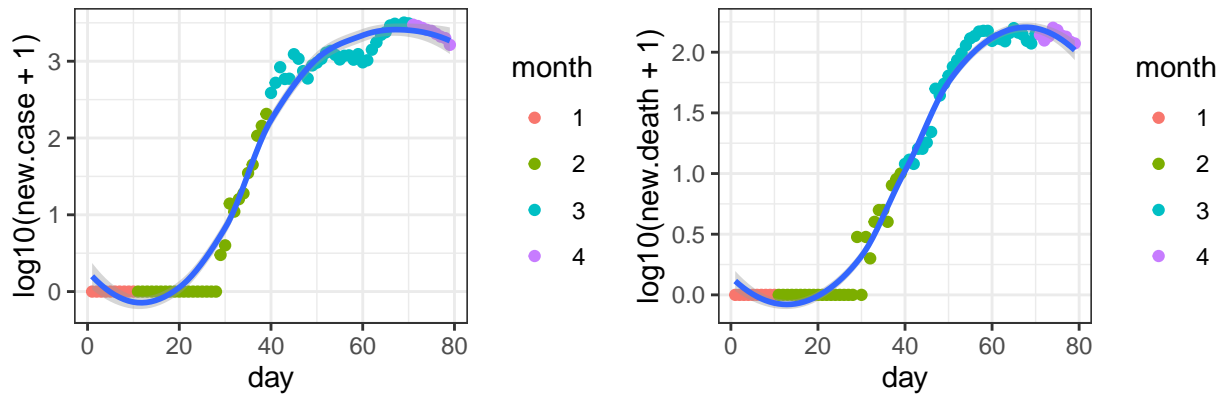
data source: <https://github.com/CSSEGISandData/COVID-19>, day 1 is 1/22/2020

United Kingdom



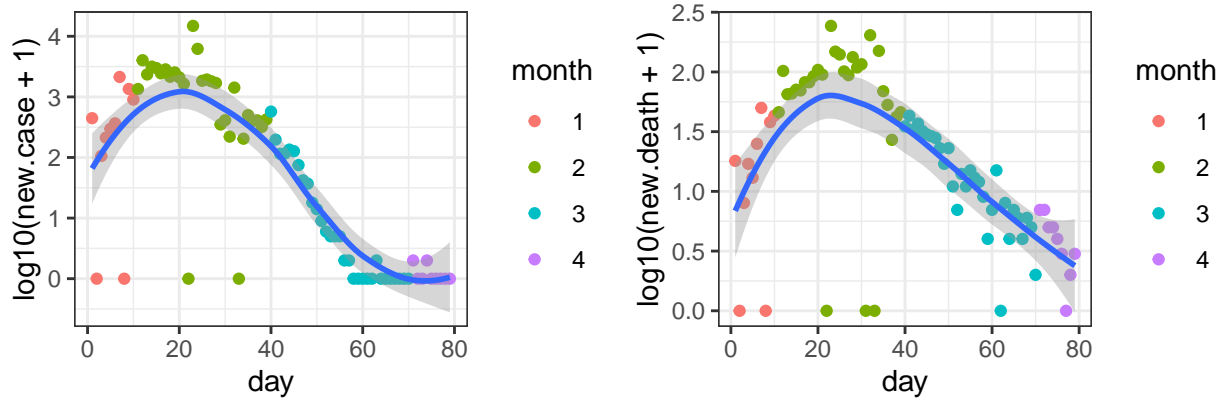
data source: <https://github.com/CSSEGISandData/COVID-19>, day 1 is 1/22/2020

Iran



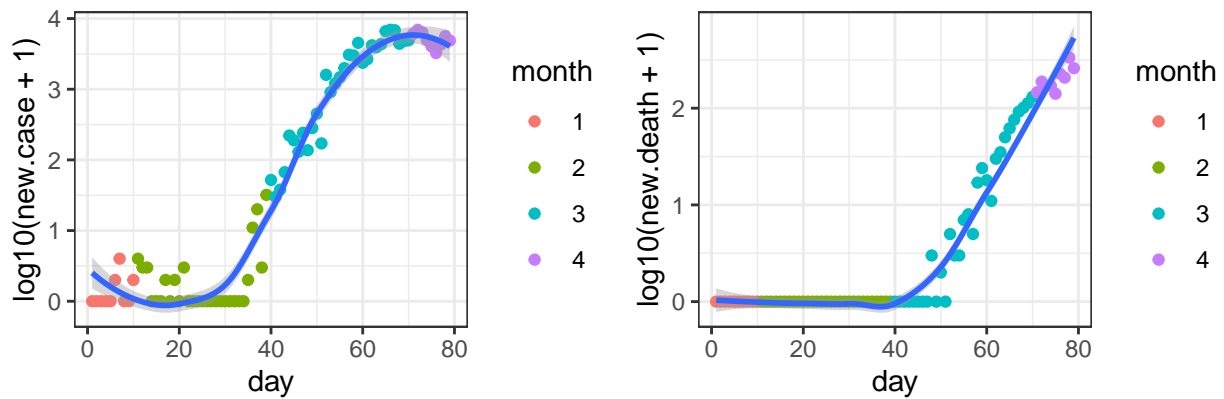
data source: <https://github.com/CSSEGISandData/COVID-19>, day 1 is 1/22/2020

China-Hubei

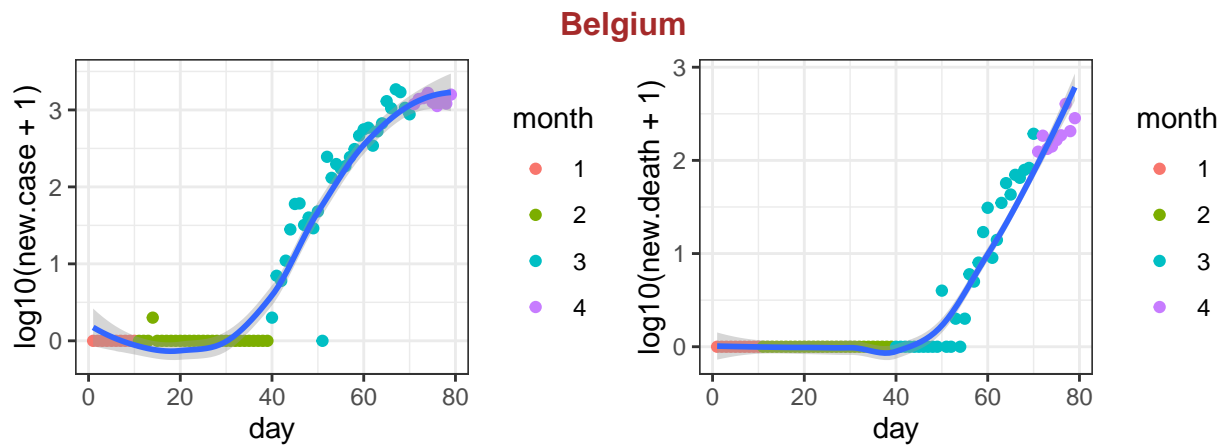


data source: <https://github.com/CSSEGISandData/COVID-19>, day 1 is 1/22/2020

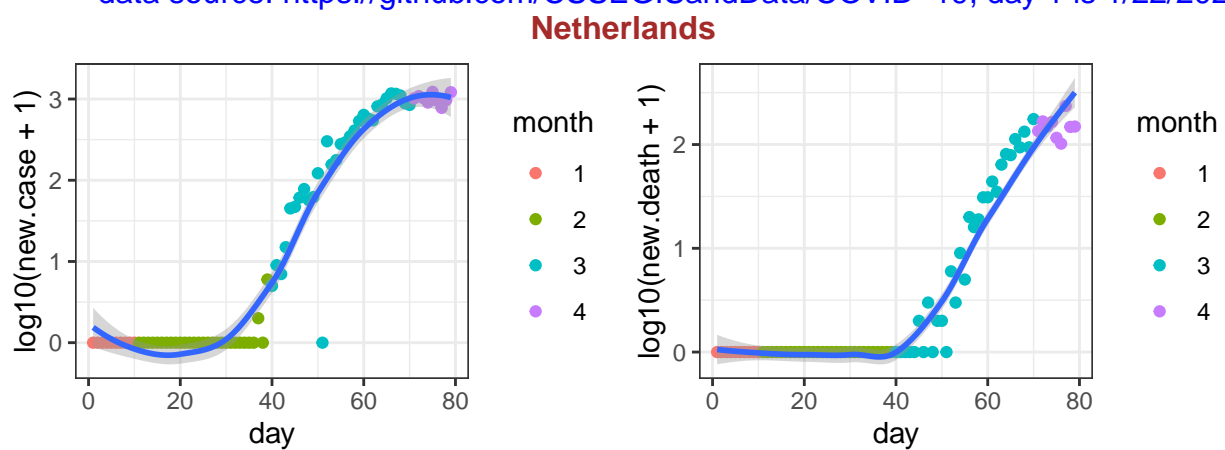
Germany



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



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

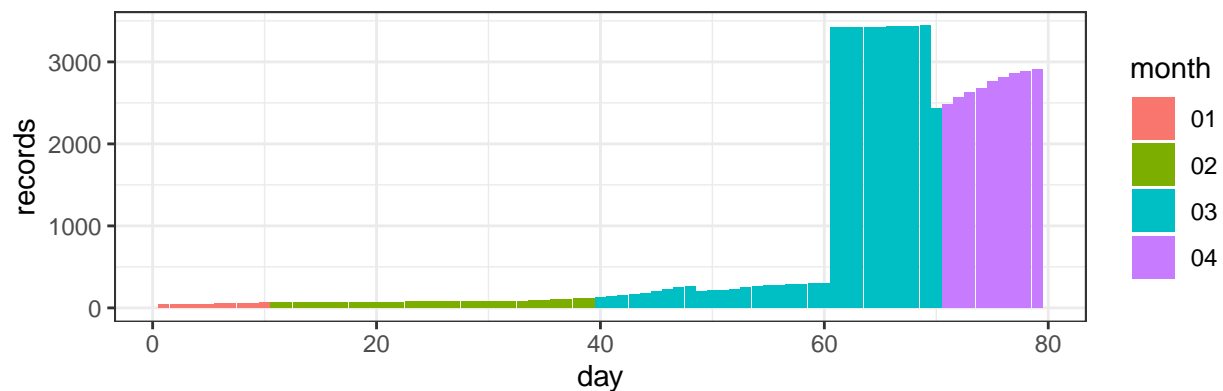


data source: <https://github.com/CSSEGISandData/COVID-19>, day 1 is 1/22/2020

daily reports data

The raw data from Hopkins are in the format of daily reports with one file per day. More recent files (since March 22nd) include 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.

number of records in Hopkins daily reports



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 current date is

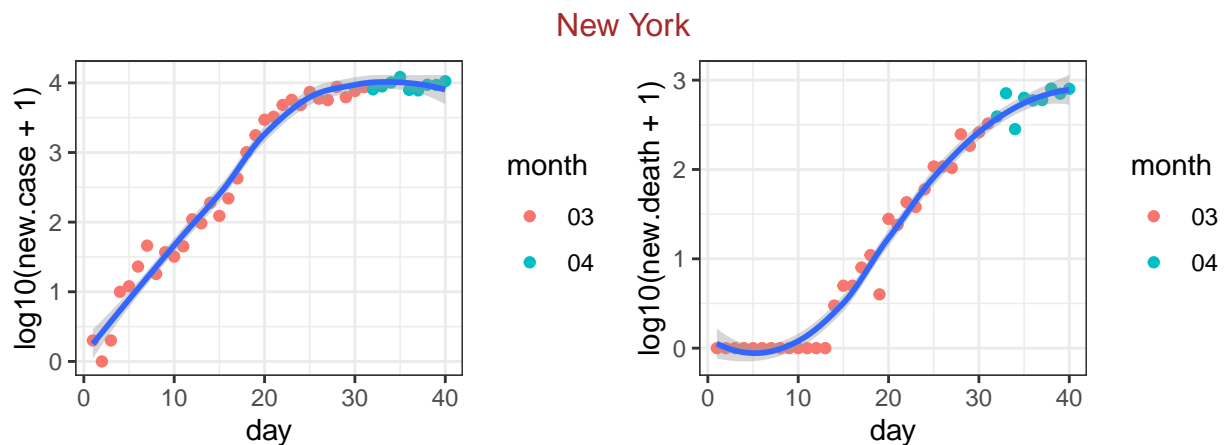
```
## [1] "2020-04-09"
```

state level data

First check the 10 states with the largest number of deaths.

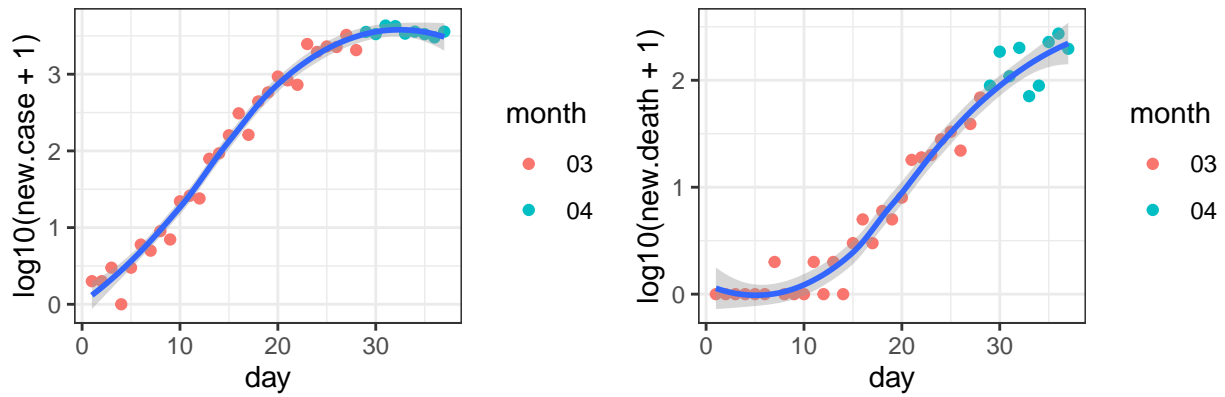
##	date	state	fips	cases	deaths
## 2084	2020-04-09	New York	36	159937	7067
## 2082	2020-04-09	New Jersey	34	51027	1700
## 2074	2020-04-09	Michigan	26	21375	1076
## 2070	2020-04-09	Louisiana	22	18283	702
## 2055	2020-04-09	California	6	20191	548
## 2065	2020-04-09	Illinois	17	16422	534
## 2073	2020-04-09	Massachusetts	25	18941	503
## 2102	2020-04-09	Washington	53	9608	456
## 2061	2020-04-09	Georgia	13	10885	412
## 2057	2020-04-09	Connecticut	9	9784	380

For these 10 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.



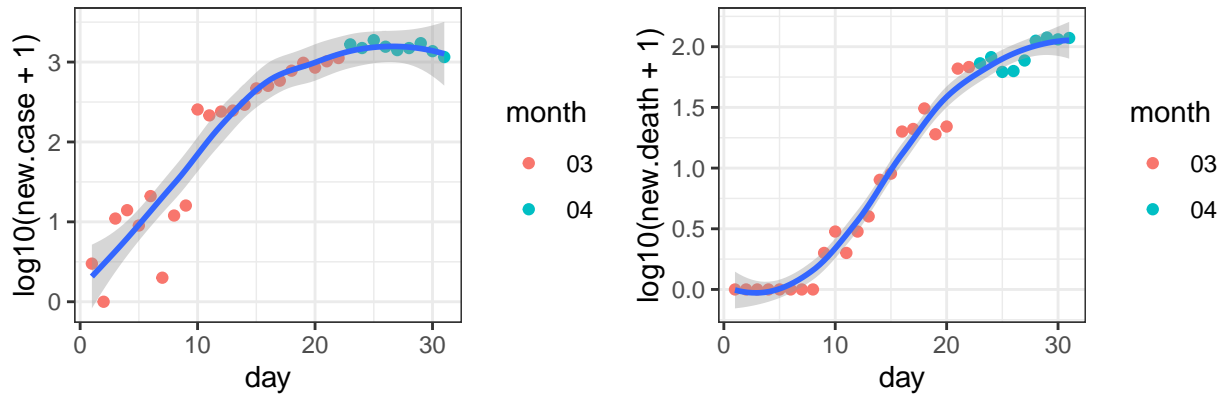
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-01

New Jersey



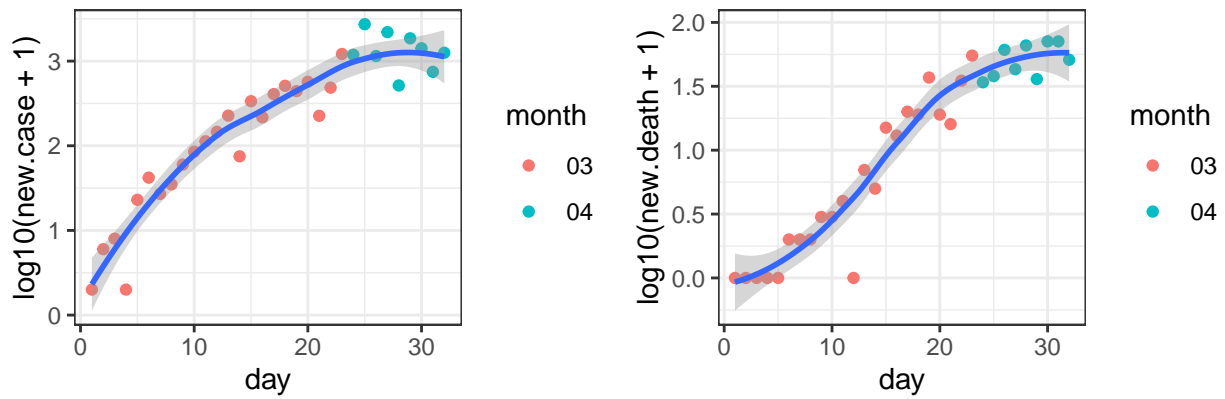
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-04

Michigan



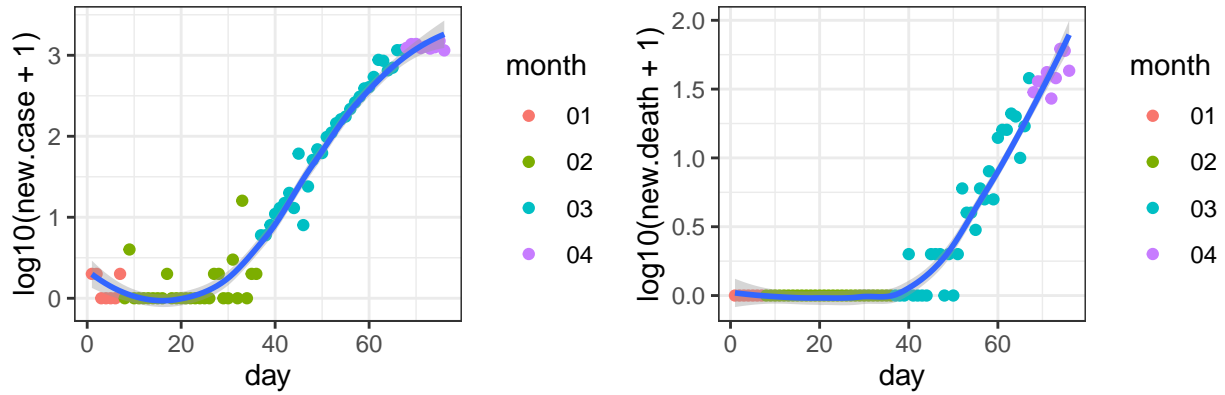
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-10

Louisiana



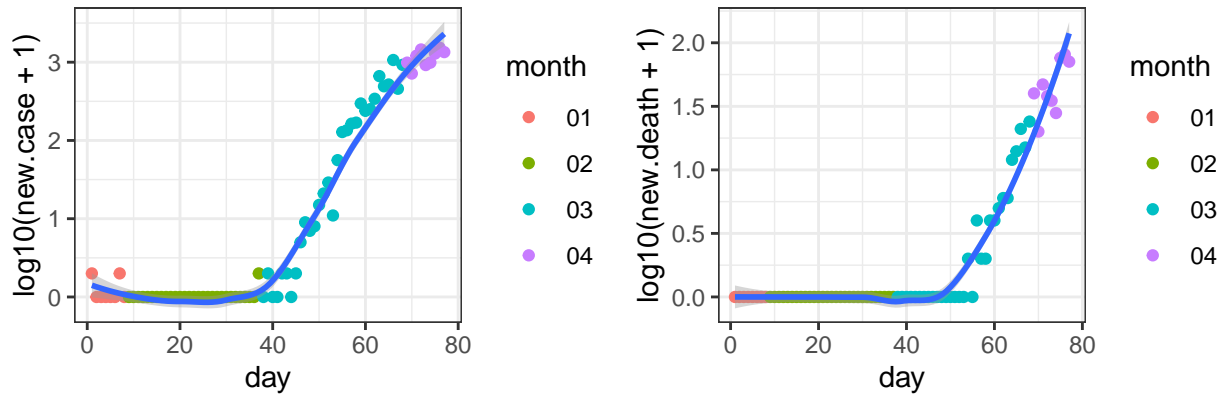
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-09

California



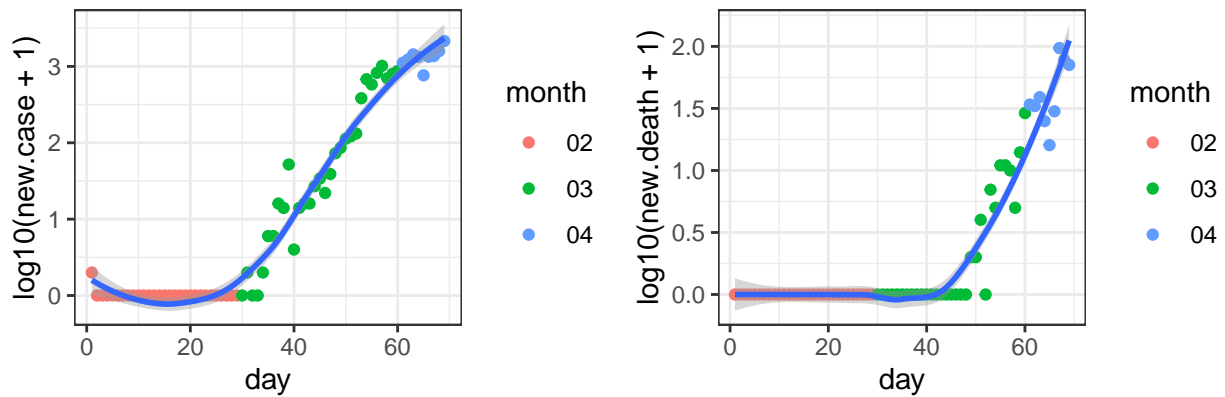
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 01-25

Illinois



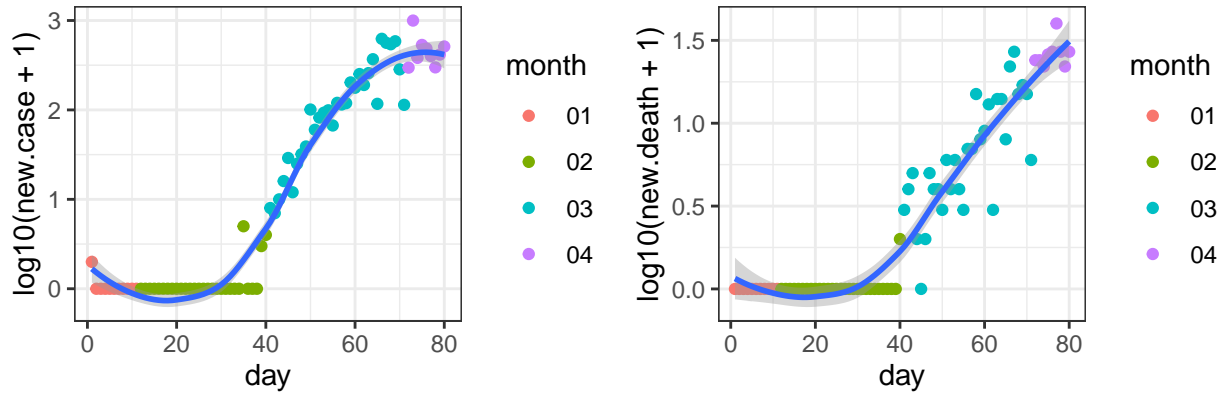
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 01-24

Massachusetts



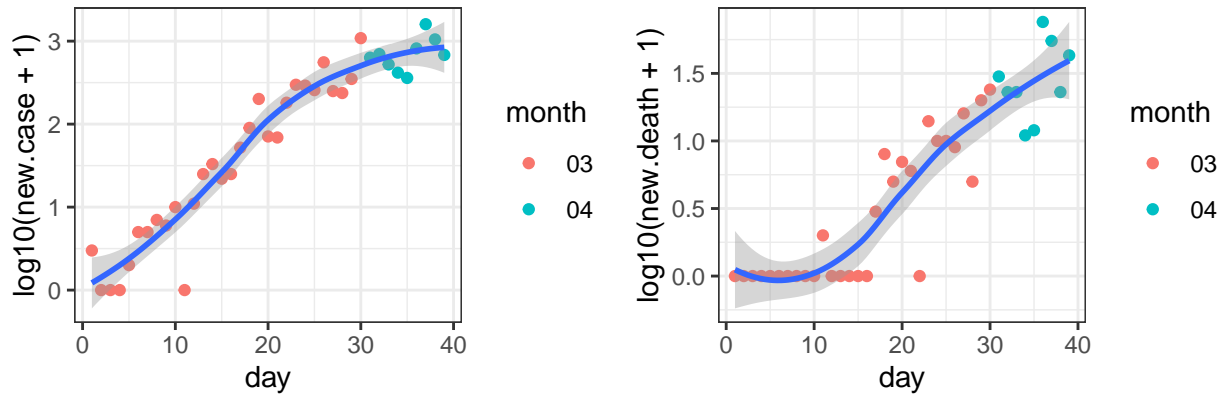
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 02-01

Washington



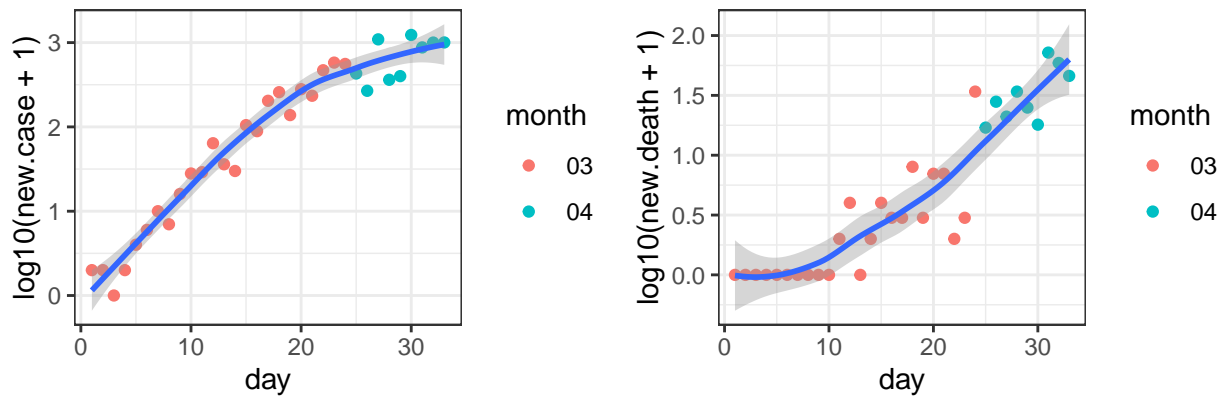
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 01-21

Georgia



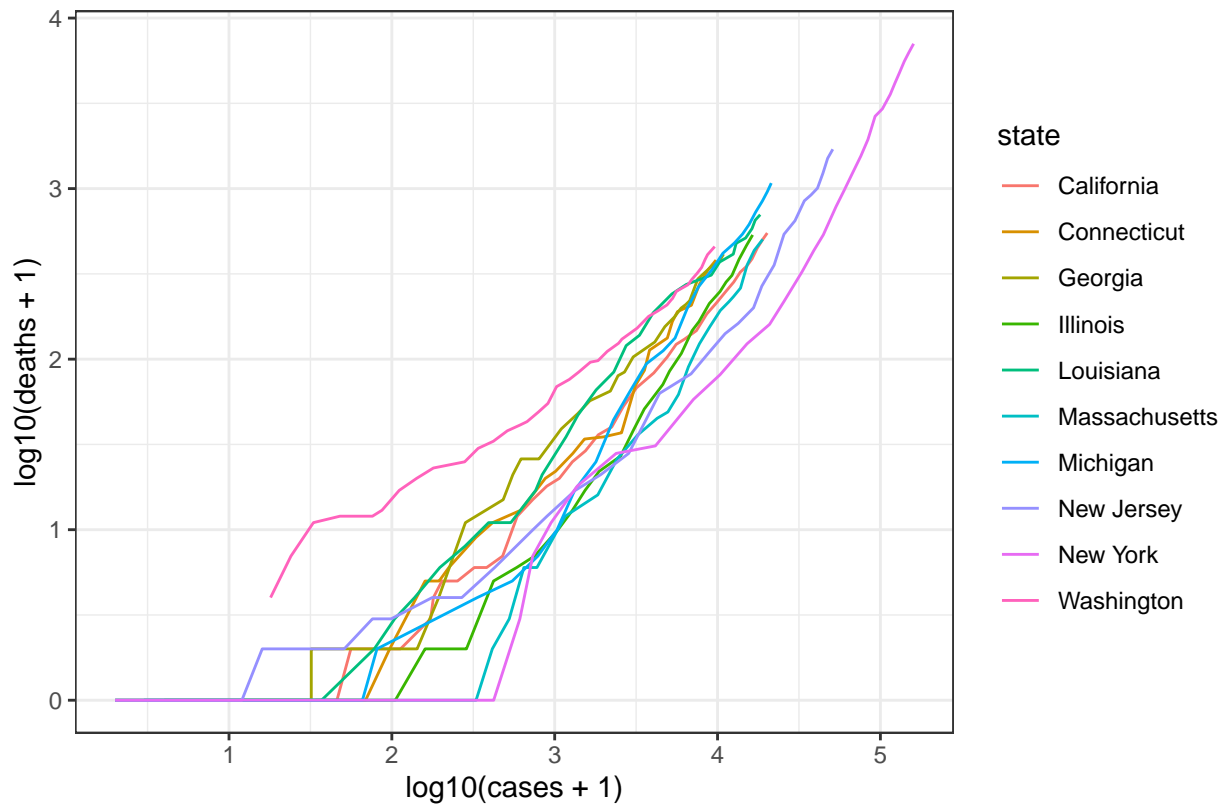
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-02

Connecticut



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>

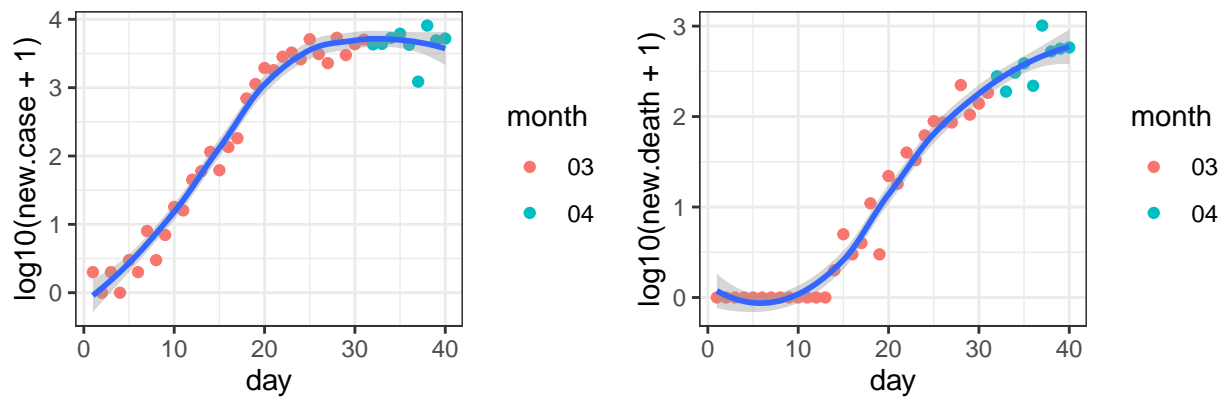
county level data

First check the 10 counties with the largest number of deaths.

##	date	county	state	fips	cases	deaths
## 44818	2020-04-09	New York City	New York	NA 87028	5150	
## 44817	2020-04-09	Nassau	New York	36059	20140	778
## 44424	2020-04-09	Wayne	Michigan	26163	10093	504
## 44845	2020-04-09	Westchester	New York	36119	17004	389
## 44837	2020-04-09	Suffolk	New York	36103	17413	369
## 43842	2020-04-09	Cook	Illinois	17031	11415	351
## 44745	2020-04-09	Bergen	New Jersey	34003	8343	345
## 44750	2020-04-09	Essex	New Jersey	34013	6069	312
## 45734	2020-04-09	King	Washington	53033	3888	260
## 44405	2020-04-09	Oakland	Michigan	26125	4247	246

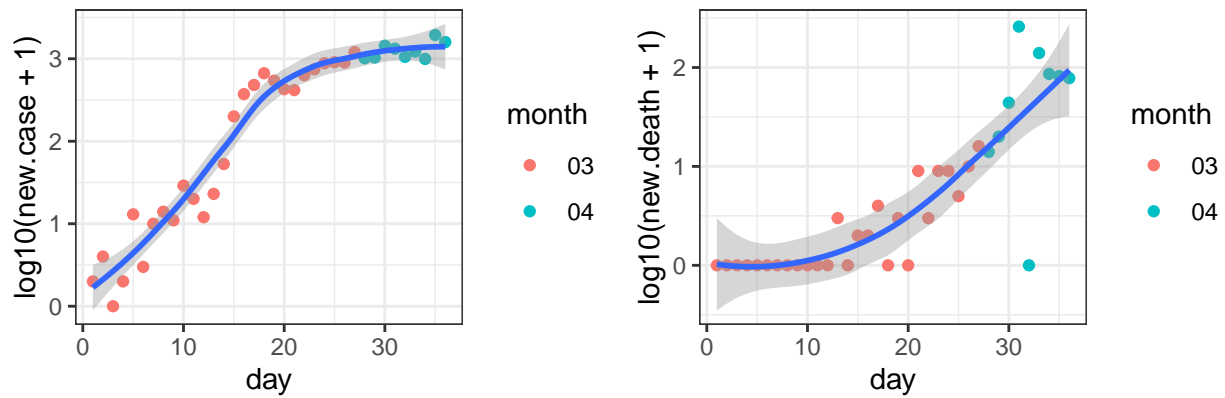
For these 10 counties, I check the number of new cases and the number of new deaths.

New York City_New York



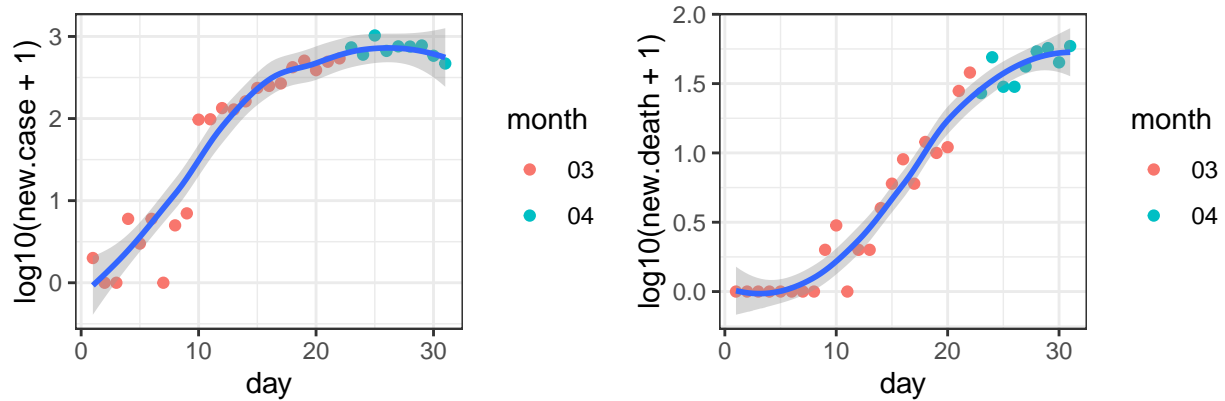
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-01

Nassau_New York



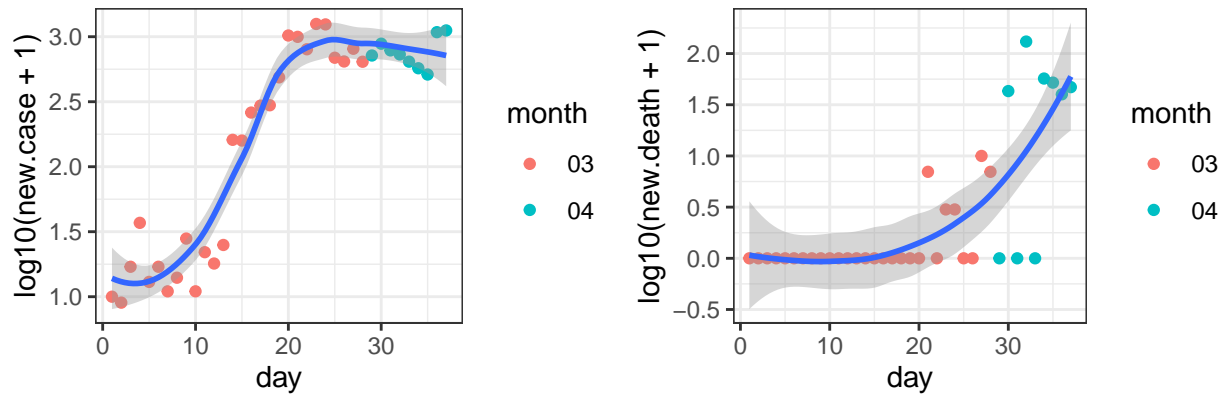
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-05

Wayne_Michigan



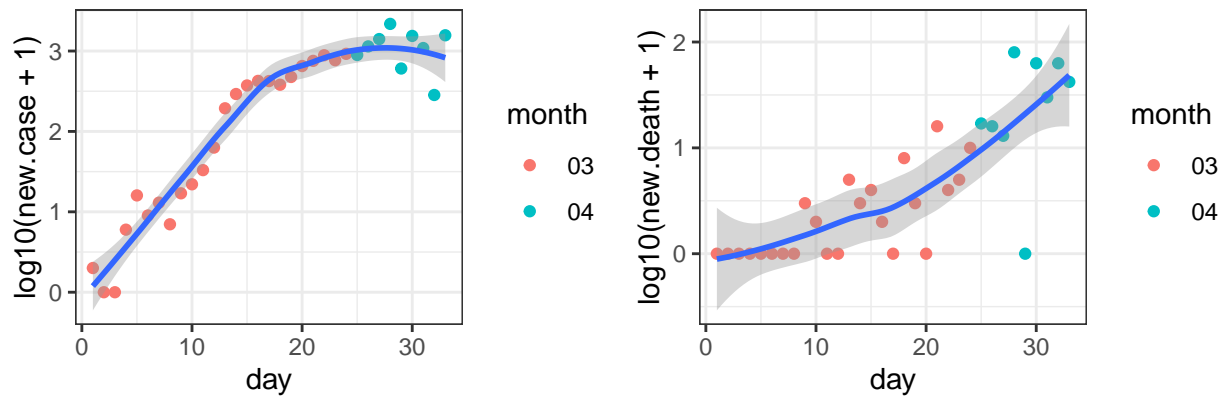
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-10

Westchester_New York



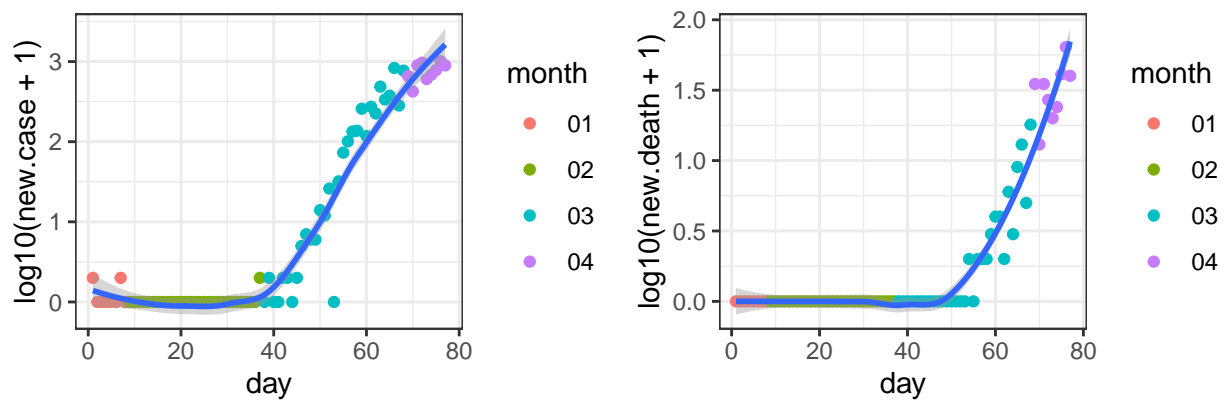
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-04

Suffolk_New York



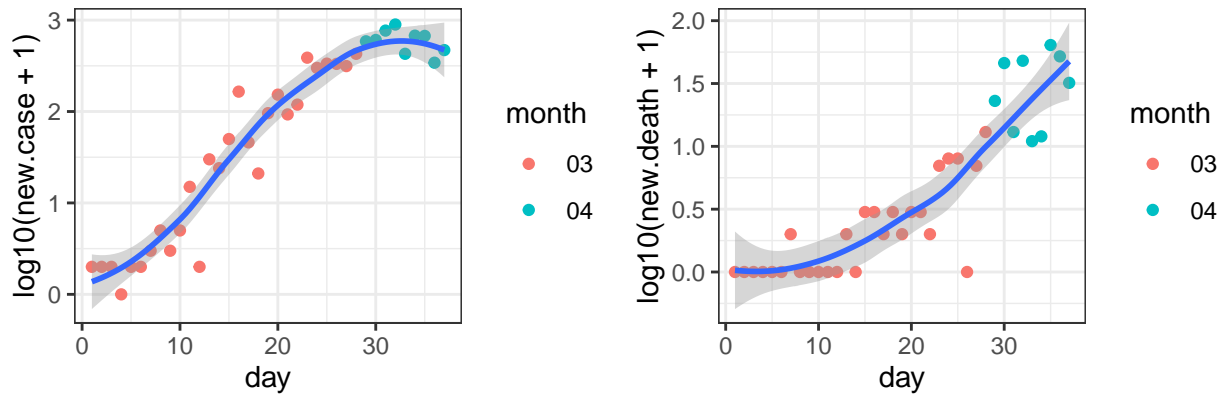
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-08

Cook_Illinois



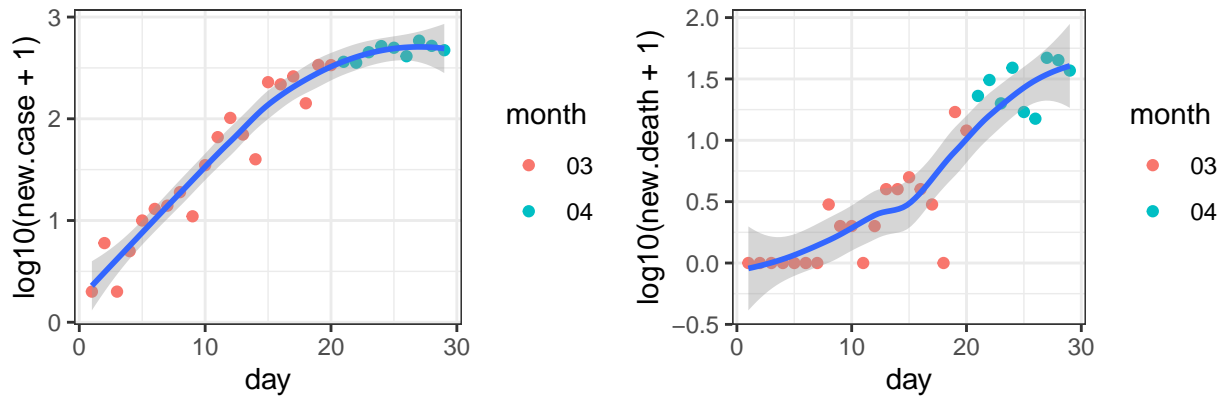
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 01-24

Bergen_New Jersey



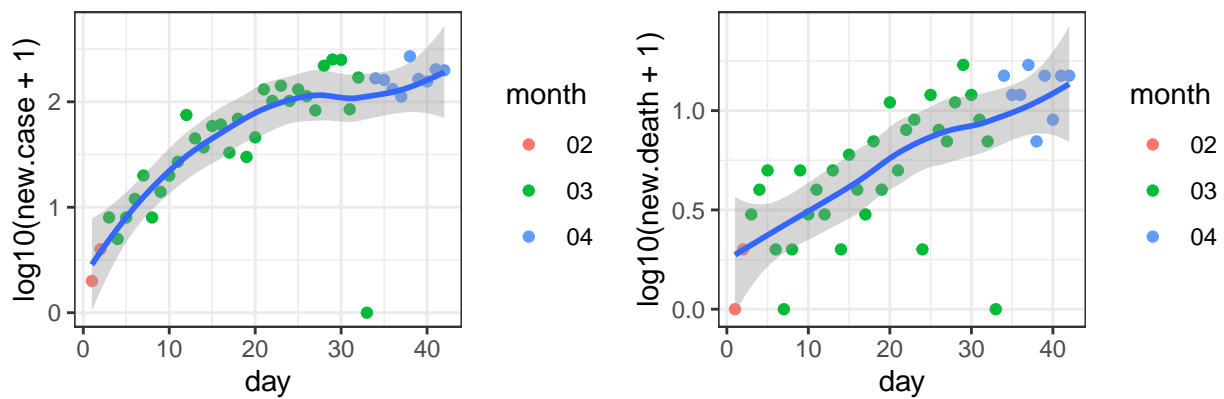
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-04

Essex_New Jersey



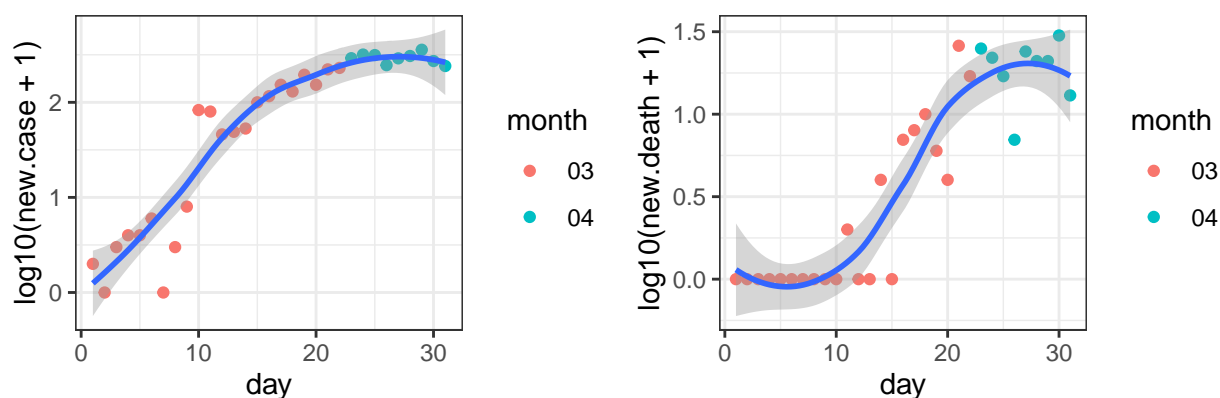
data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-12

King_Washington



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

Oakland_Michigan

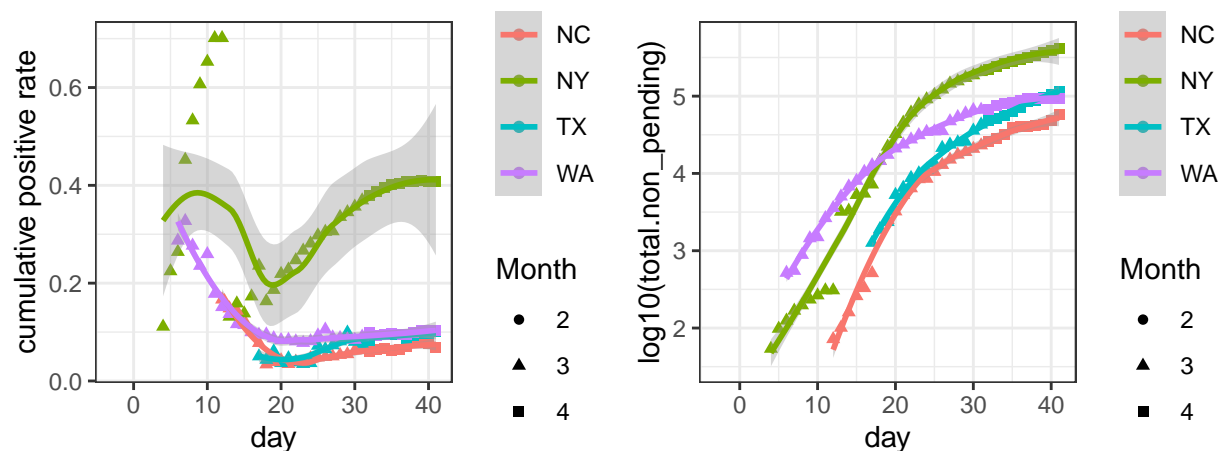


data source: <https://github.com/nytimes/covid-19-data>, day 1 is 03-10

COVID Tracking

The positive rates of testing can be an indicator on how much the COVID-19 has spread. However, they are more noisy data since the negative testing results are often not reported and the tests are almost surely taken on a non-representative random sample of the population. The COVID tracking project provides 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 available for both counties and states, here I only look at state level data.

Since the daily positive rate can fluctuate a lot, here I only illustrate the cumulative positive rate across time, for four states with grade A data. Of course since this is an R markdown file, you can modify the source code and check for other states.



github.com/COVID19Tracking/, cumulative positive rate on 0410: 0.10(WA) 0.10(TX) 0.41(NY) 0.07(NC)

Session information

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
```

```

## Running under: macOS Catalina 10.15.4
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods    base
##
## other attached packages:
## [1] httr_1.4.1    ggpubr_0.2.5  magrittr_1.5  ggplot2_3.2.1
##
## loaded via a namespace (and not attached):
## [1] Rcpp_1.0.3      pillar_1.4.3    compiler_3.6.2  tools_3.6.2
## [5] digest_0.6.23   evaluate_0.14    lifecycle_0.1.0 tibble_2.1.3
## [9] gtable_0.3.0     pkgconfig_2.0.3  rlang_0.4.4     yaml_2.2.1
## [13] xfun_0.12        gridExtra_2.3    withr_2.1.2     dplyr_0.8.4
## [17] stringr_1.4.0    knitr_1.28       grid_3.6.2      tidyselect_1.0.0
## [21] cowplot_1.0.0    glue_1.3.1       R6_2.4.1         rmarkdown_2.1
## [25] purrr_0.3.3      farver_2.0.3     scales_1.1.0     htmltools_0.4.0
## [29] assertthat_0.2.1 colorspace_1.4-1 ggsignif_0.6.0   labeling_0.3
## [33] stringi_1.4.5    lazyeval_0.2.2   munsell_0.5.0    crayon_1.3.4

```