Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-03-30

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

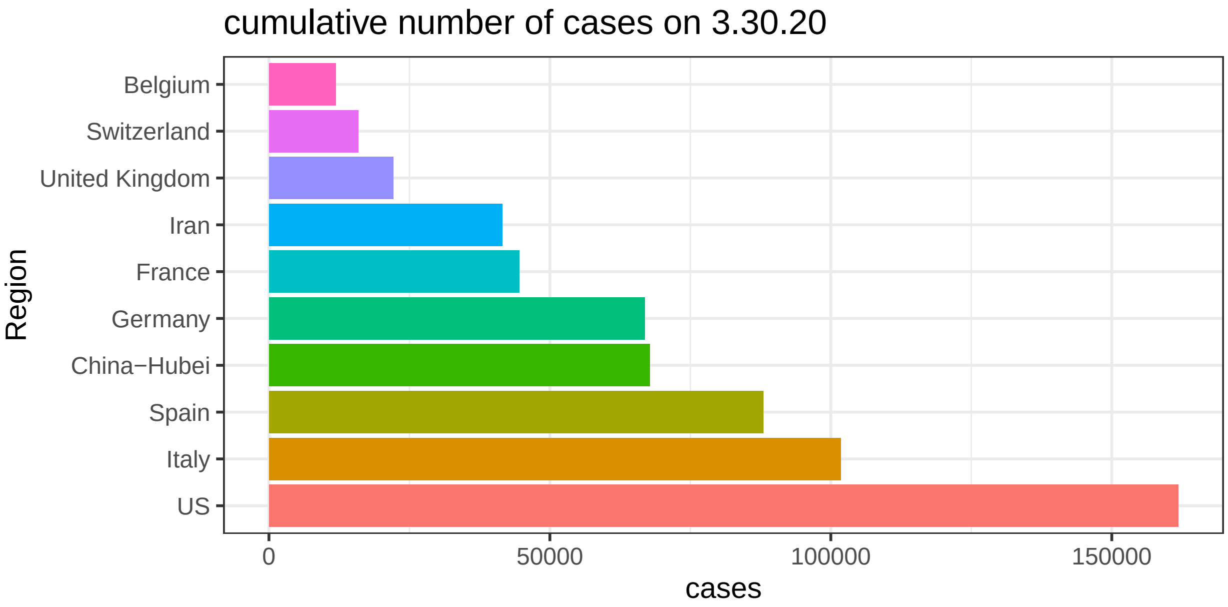
JHU

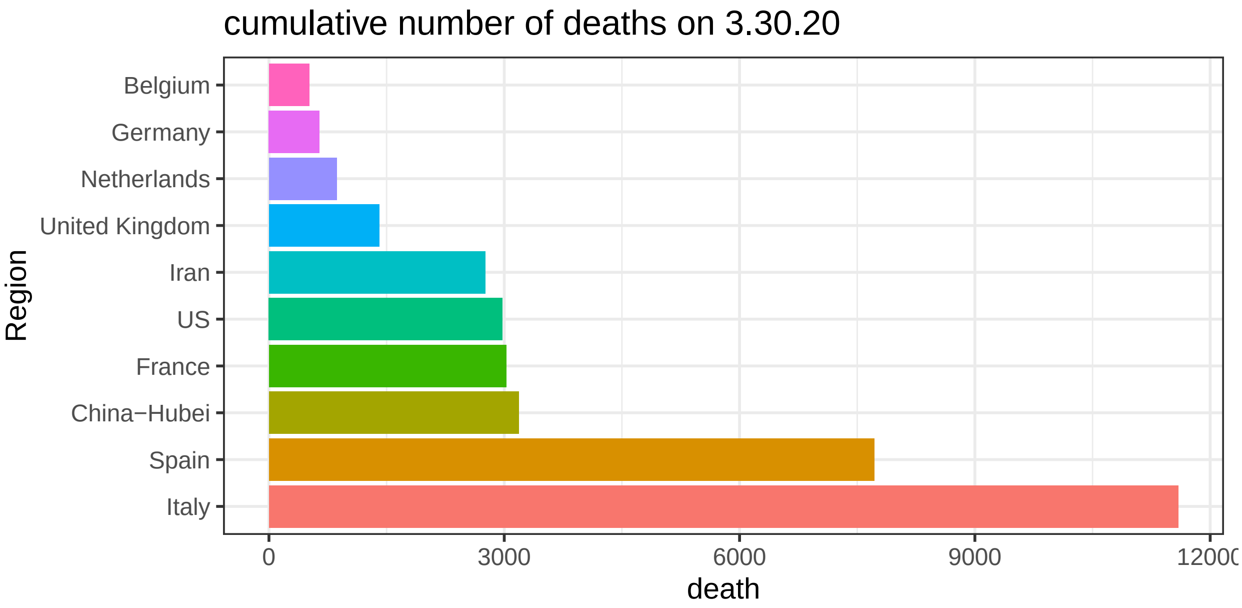
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

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.

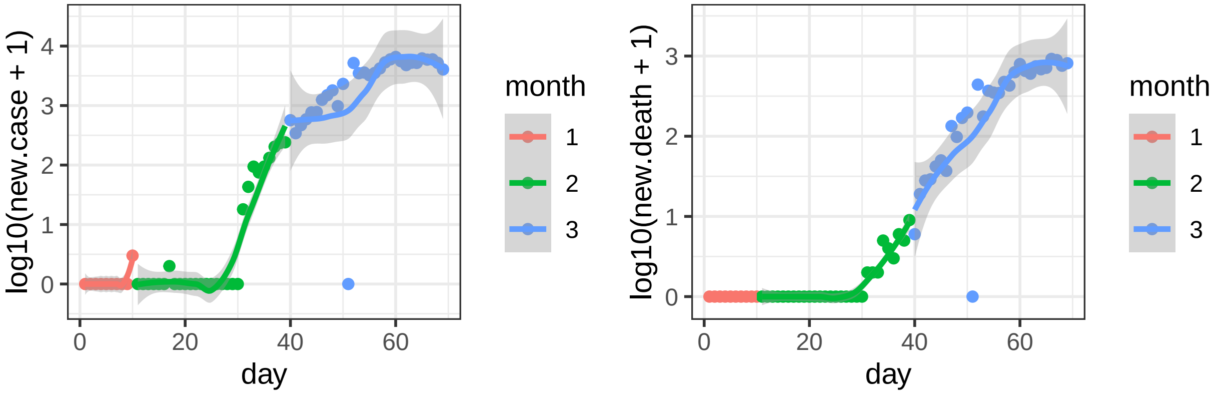
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





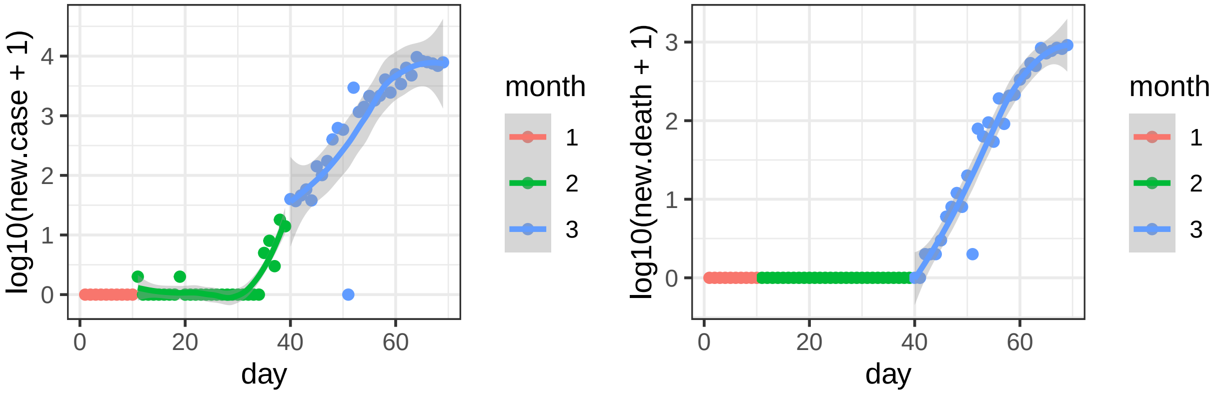
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

Italy



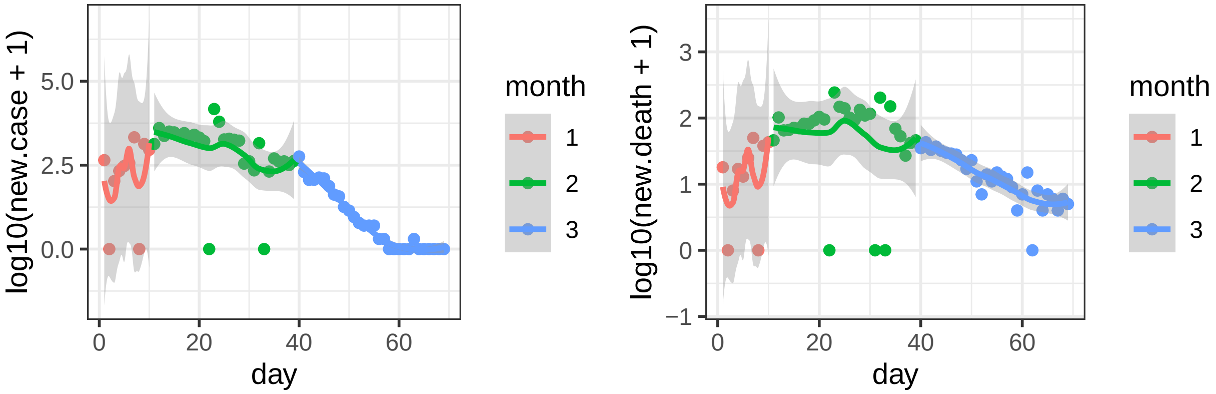
day 1 is Jan 22nd, 2020

Spain



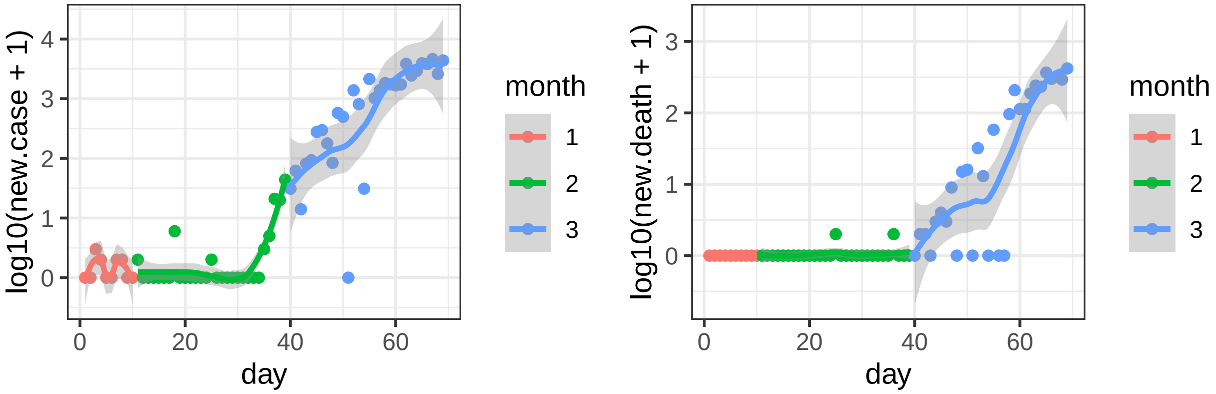
day 1 is Jan 22nd, 2020

China−Hubei



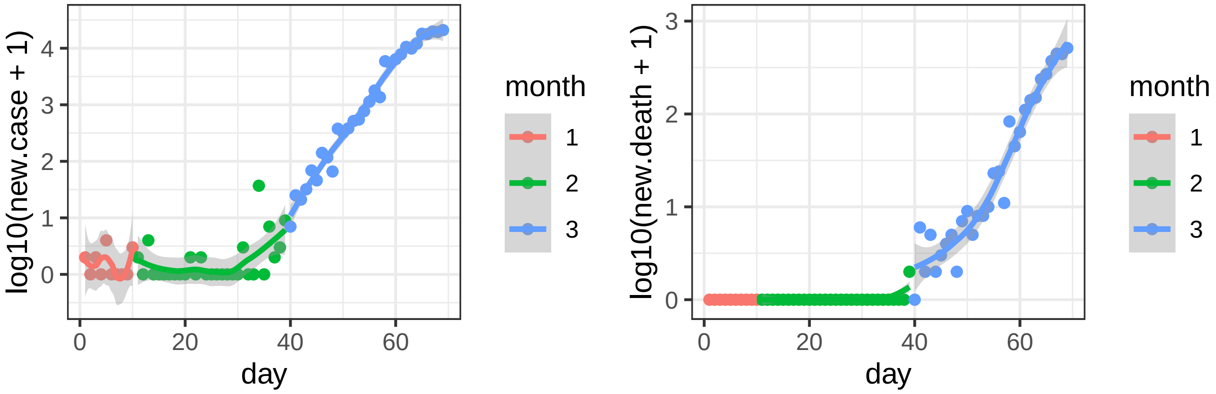
day 1 is Jan 22nd, 2020

France



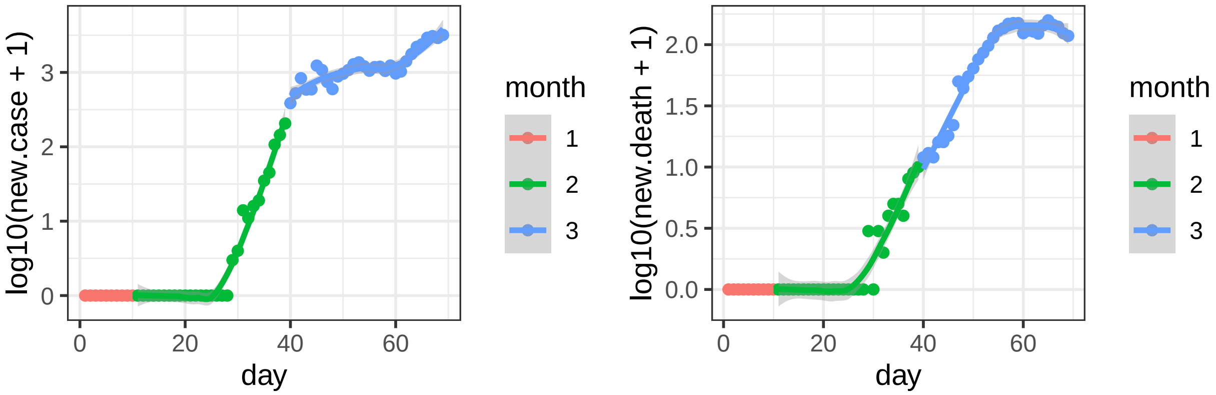
day 1 is Jan 22nd, 2020

US



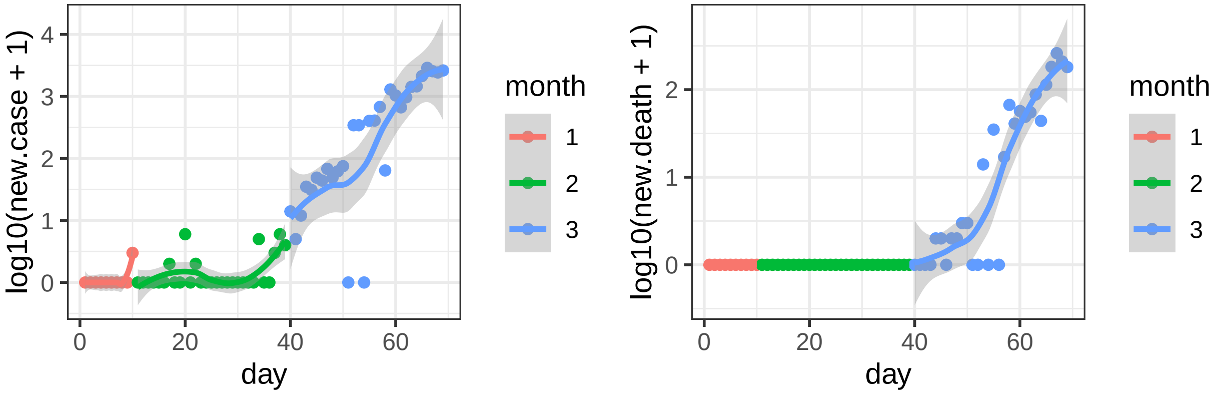
day 1 is Jan 22nd, 2020

Iran



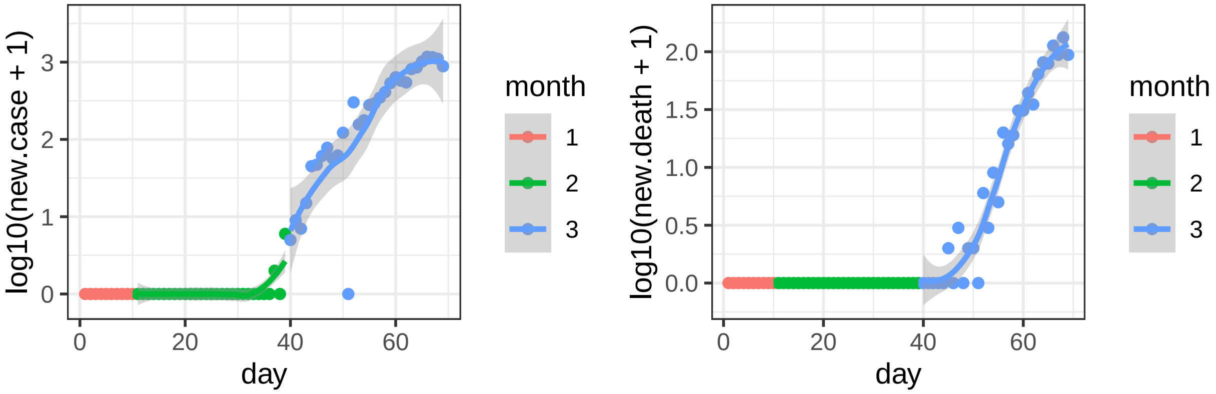
day 1 is Jan 22nd, 2020

United Kingdom



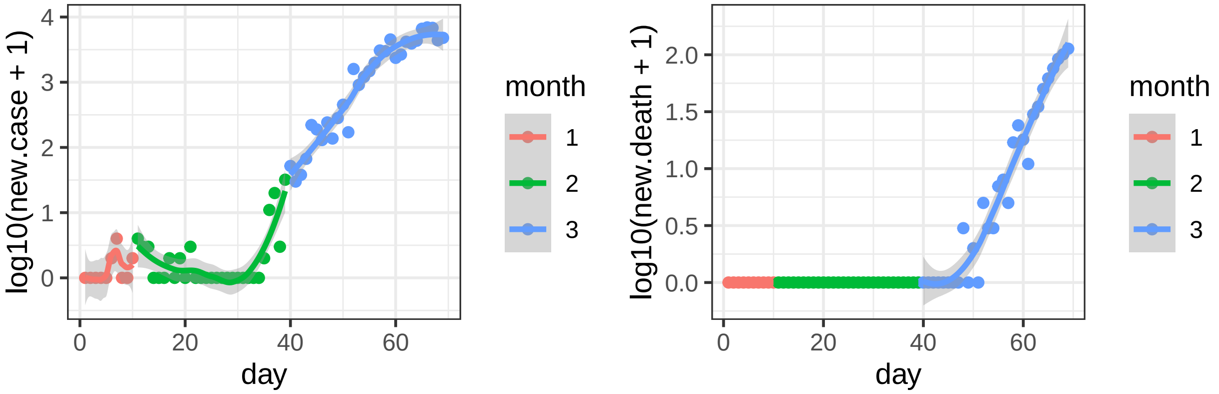
day 1 is Jan 22nd, 2020

Netherlands



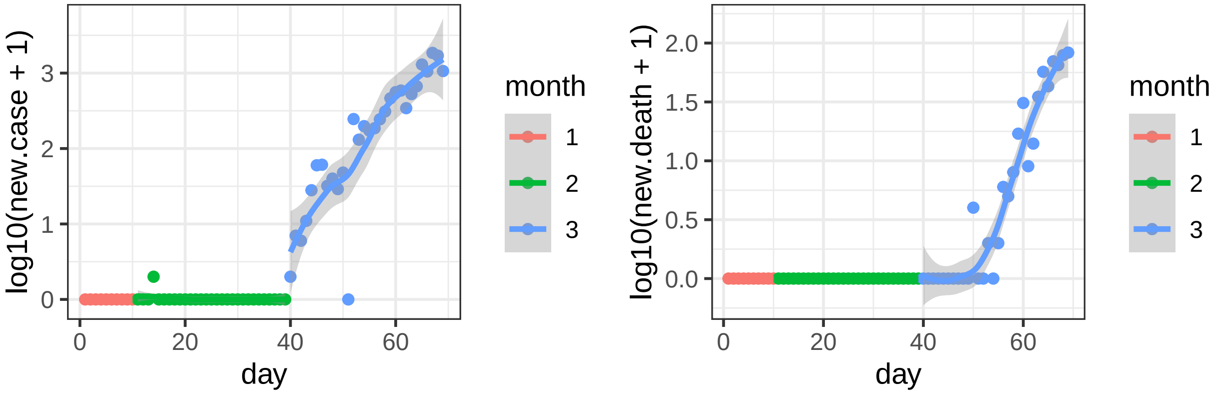
day 1 is Jan 22nd, 2020

Germany



day 1 is Jan 22nd, 2020

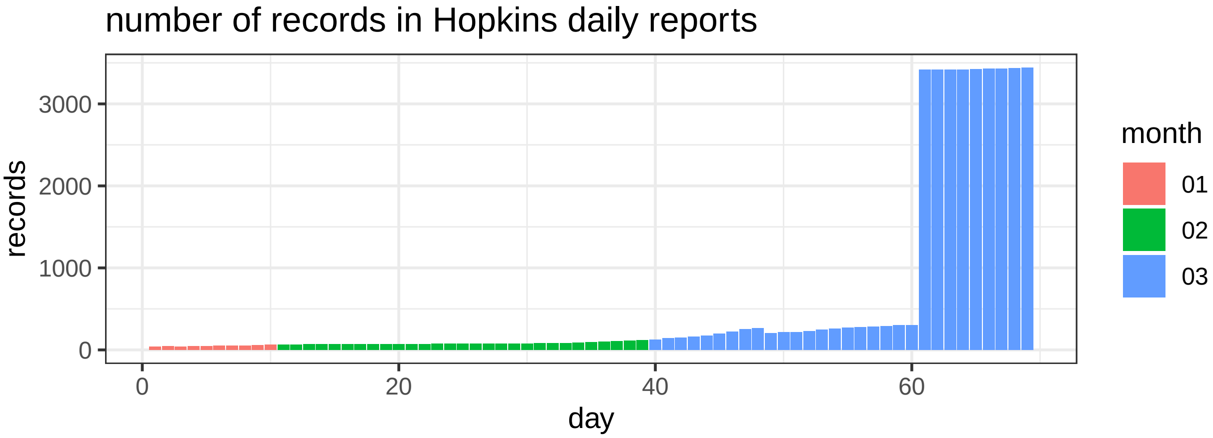
Belgium



day 1 is Jan 22nd, 2020

daily reports data

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


number of records in Hopkins daily reports

day 1 is Jan 22nd, 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

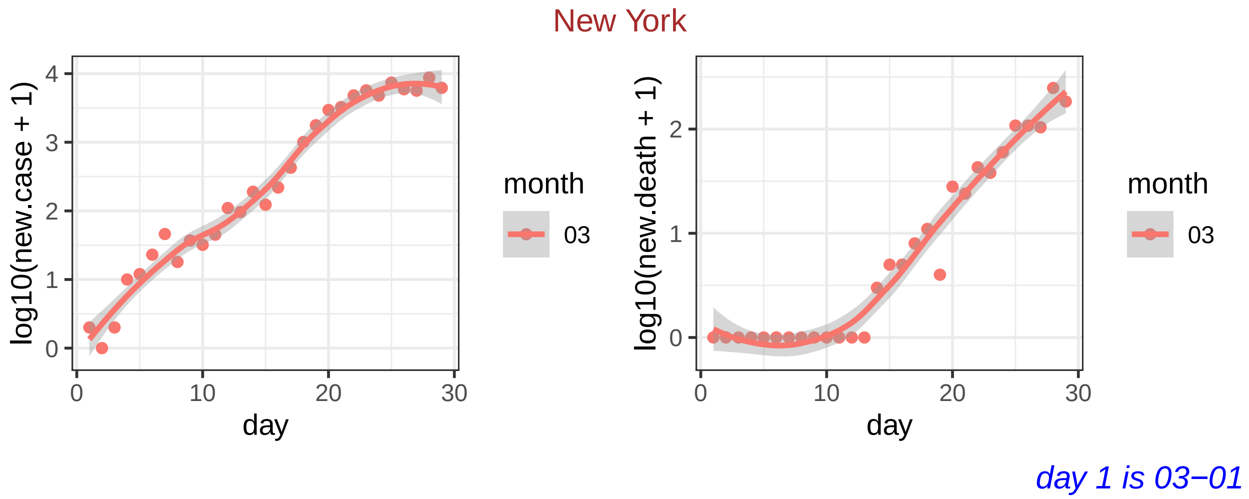
```
## [1] "2020-03-29"
```

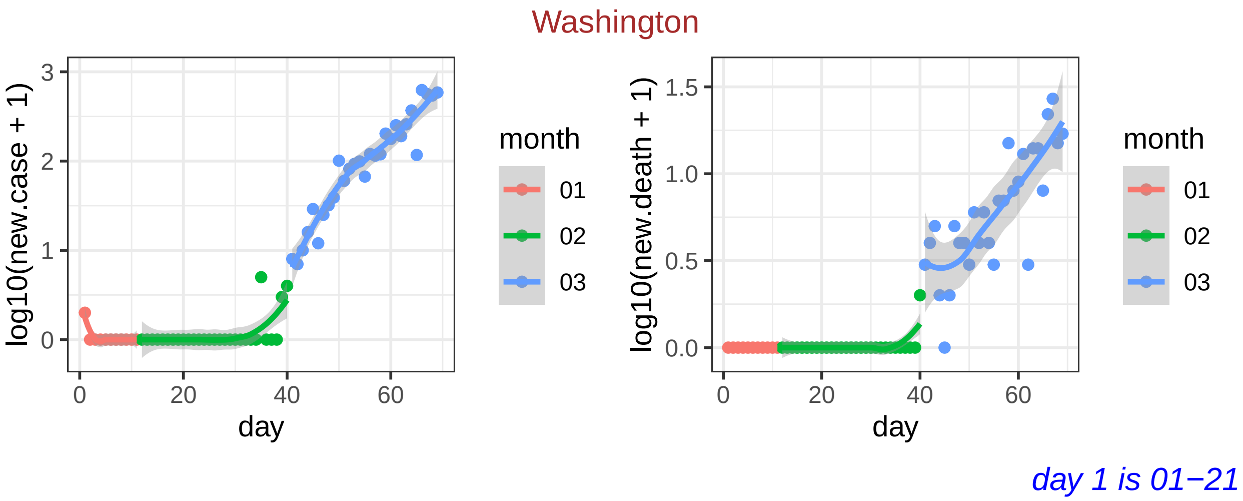
state level data

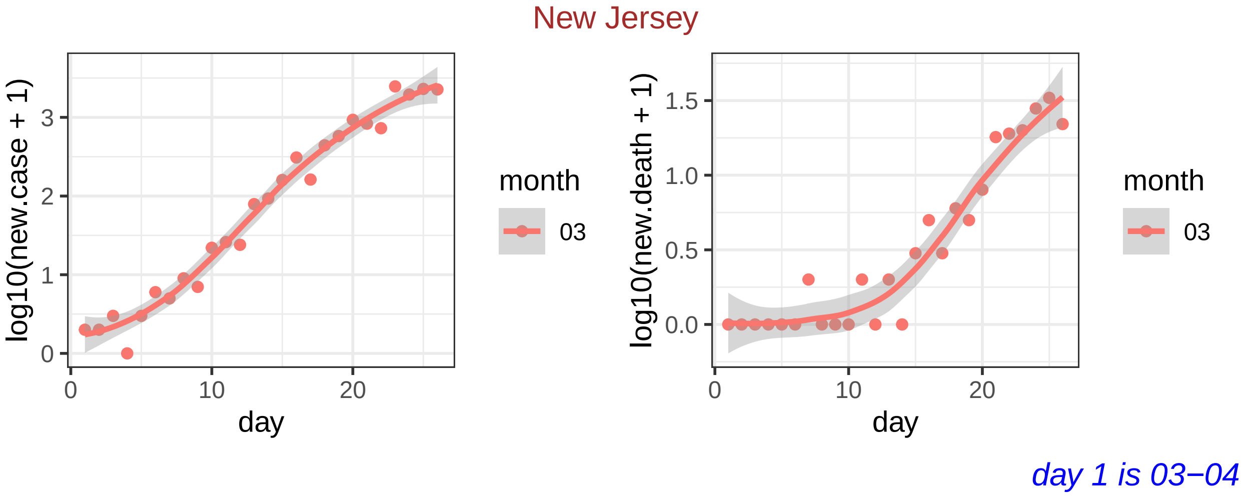
First check the 10 states with the largest number of deaths.

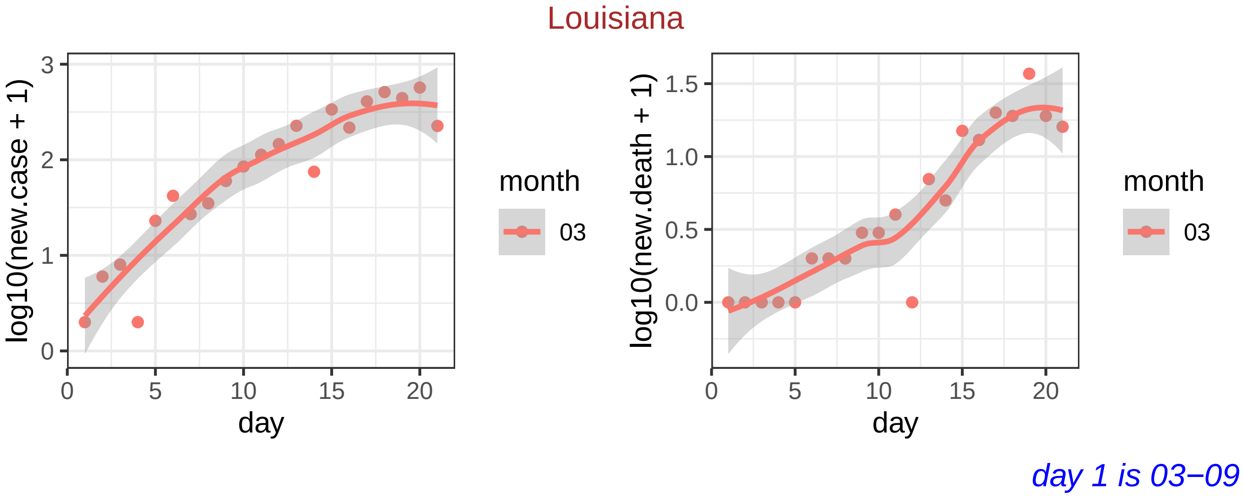
```
##            date         state fips cases deaths
## 1478 2020-03-29      New York   36 59568    965
## 1496 2020-03-29    Washington   53  4896    207
## 1476 2020-03-29    New Jersey   34 13386    161
## 1464 2020-03-29     Louisiana   22  3540    152
## 1468 2020-03-29      Michigan   26  5486    132
## 1449 2020-03-29    California    6  6266    130
## 1455 2020-03-29       Georgia   13  2683     83
## 1459 2020-03-29      Illinois   17  4613     70
## 1454 2020-03-29       Florida   12  4942     59
## 1467 2020-03-29 Massachusetts   25  4955     48
```

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.

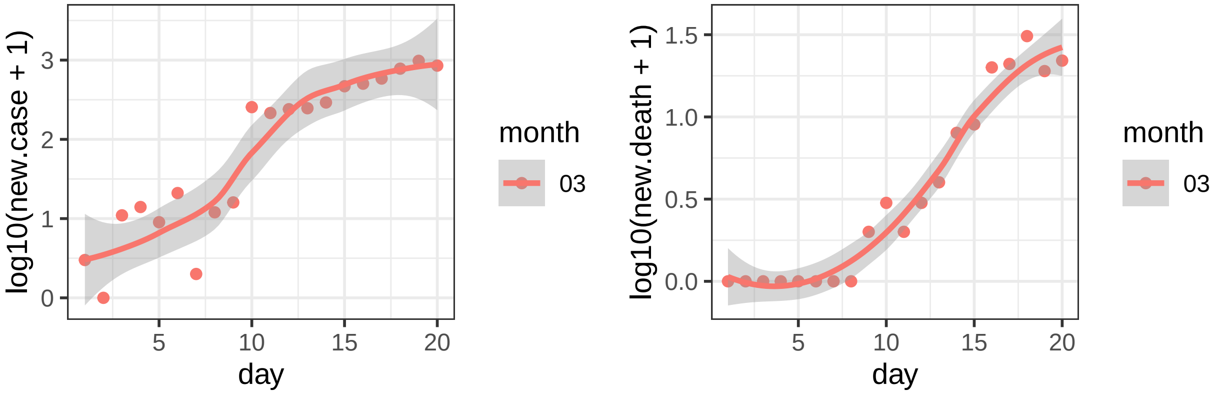



Washington
log10(new.case + 1)
0
1
2
3
0
20
40
60
day
month
01
02
03
log10(new.death + 1)
0.0
0.5
1.0
1.5
0
20
40
60
day
month
01
02
03
day 1 is 01−21


New Jersey
log10(new.case + 1)
0
1
2
3
0
10
20
day
month
03
log10(new.death + 1)
0.0
0.5
1.0
1.5
0
10
20
day
month
03
day 1 is 03−04

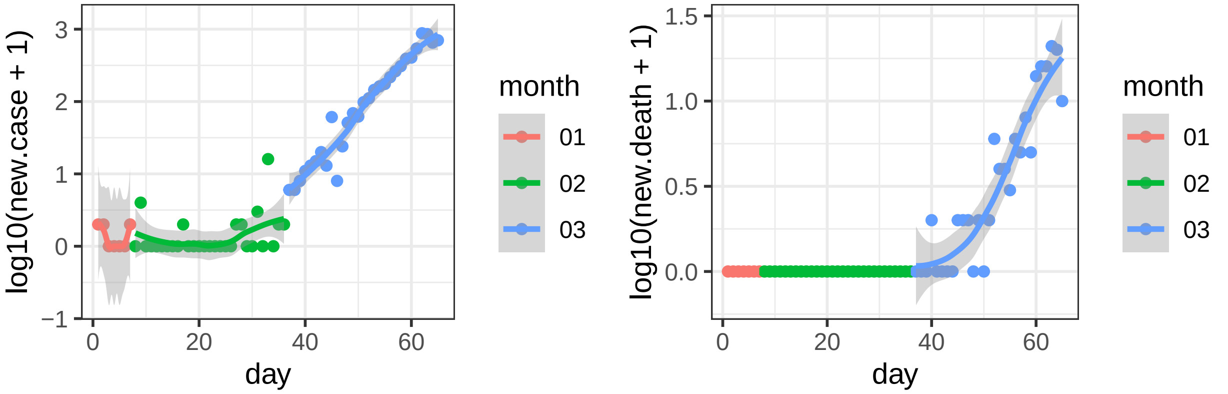

Louisiana
log10(new.case + 1)
0
1
2
3
0
5
10
15
20
day
month
03
log10(new.death + 1)
0.0
0.5
1.0
1.5
0
5
10
15
20
day
month
03
day 1 is 03−09

Michigan



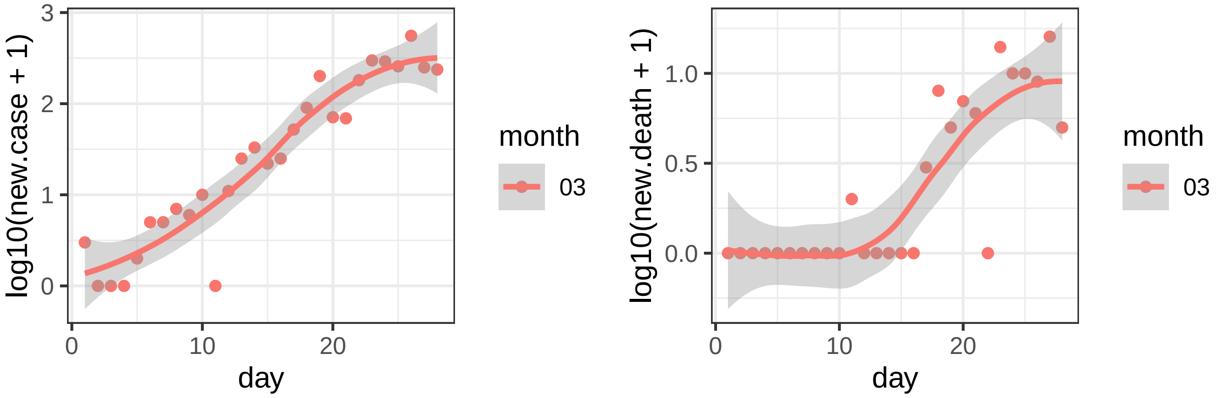
day 1 is 03−10

California

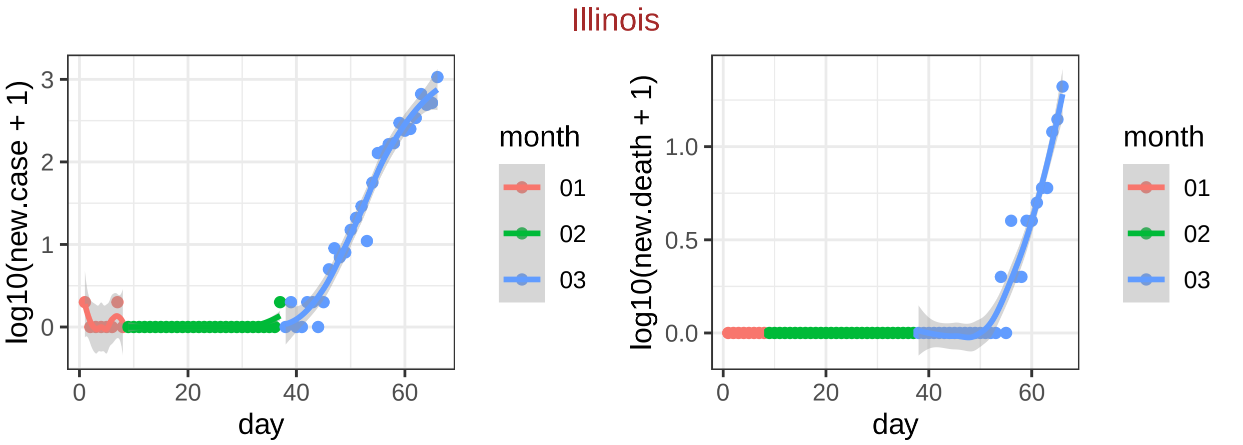


day 1 is 01−25

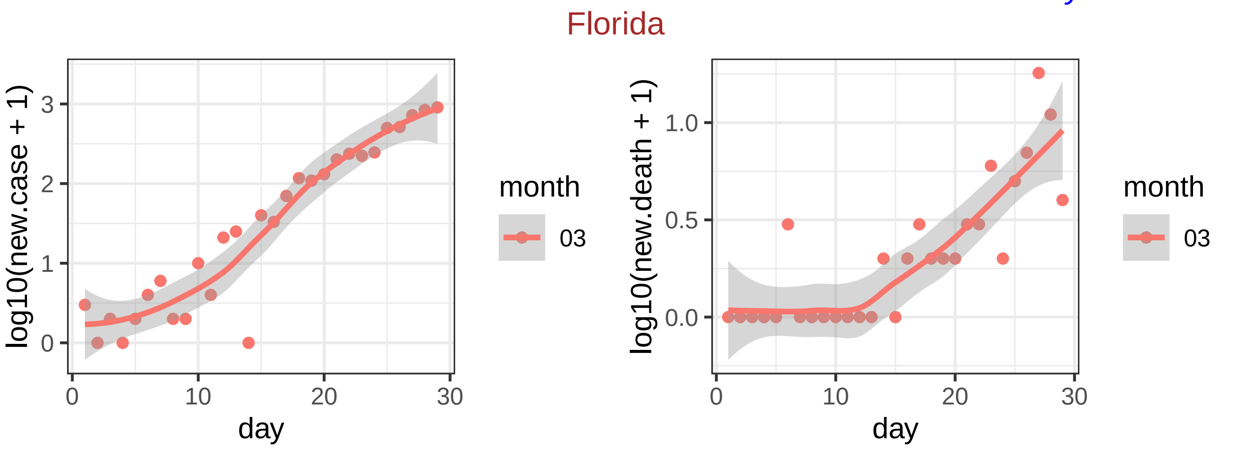
Georgia



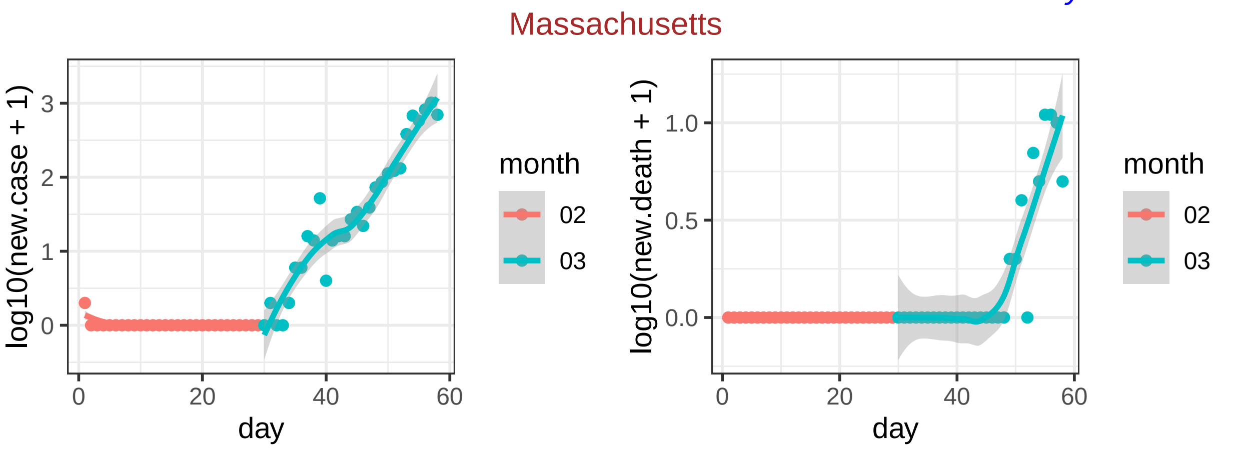
day 1 is 03−02



day 1 is 01−24

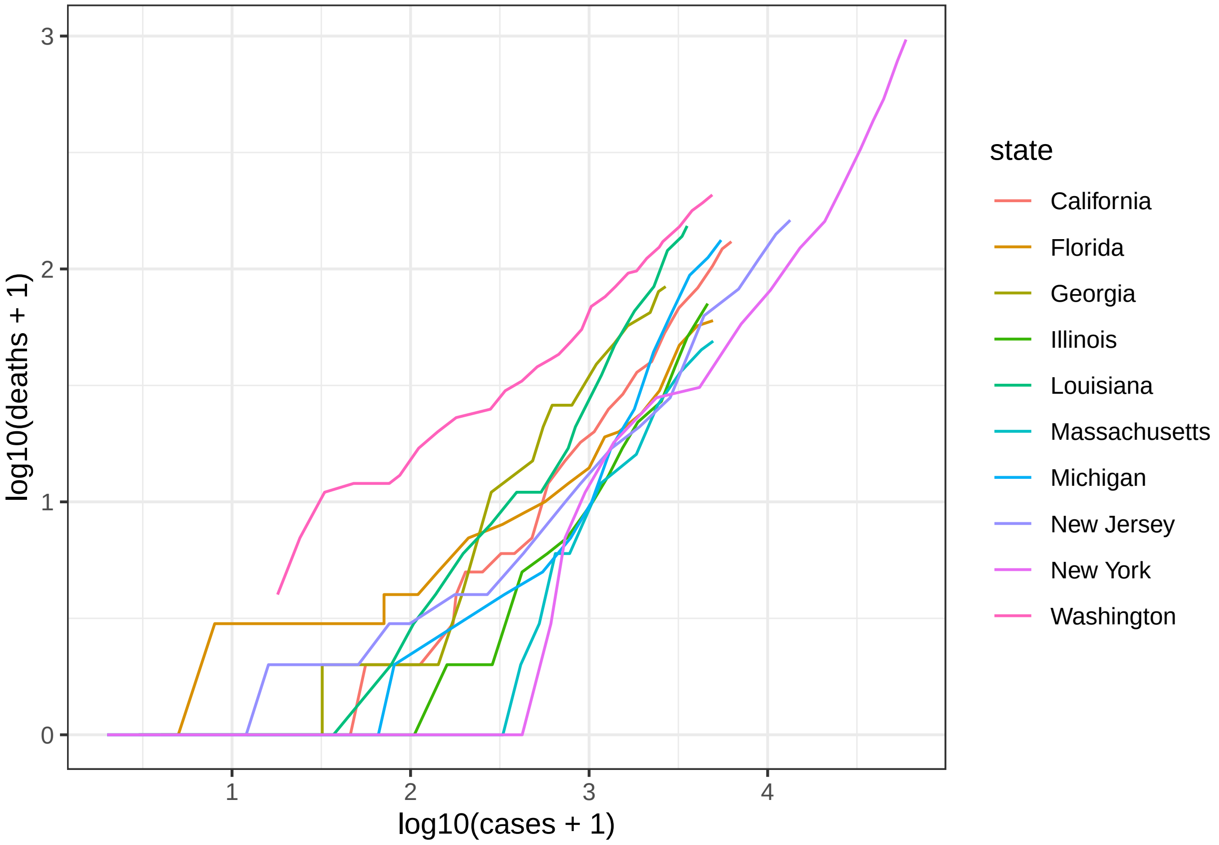


day 1 is 03−01



day 1 is 02−01

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



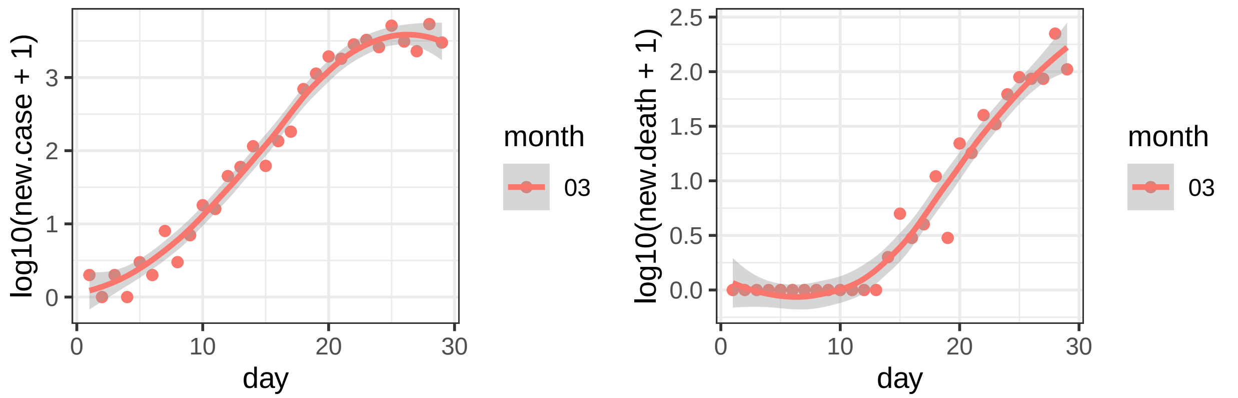
county level data

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

```
##              date        county      state  fips cases deaths
## 18902 2020-03-29 New York City   New York    NA 33768    776
## 19611 2020-03-29          King Washington 53033  2163    146
## 18465 2020-03-29       Orleans  Louisiana 22071  1350     73
## 18600 2020-03-29         Wayne   Michigan 26163  2704     56
## 18166 2020-03-29          Cook   Illinois 17031  3445     40
## 18920 2020-03-29       Suffolk   New York 36103  5023     40
## 18901 2020-03-29        Nassau   New York 36059  6445     39
## 17870 2020-03-29   Los Angeles California  6037  2136     37
## 18836 2020-03-29        Bergen New Jersey 34003  2169     35
## 18585 2020-03-29       Oakland   Michigan 26125  1170     34
```

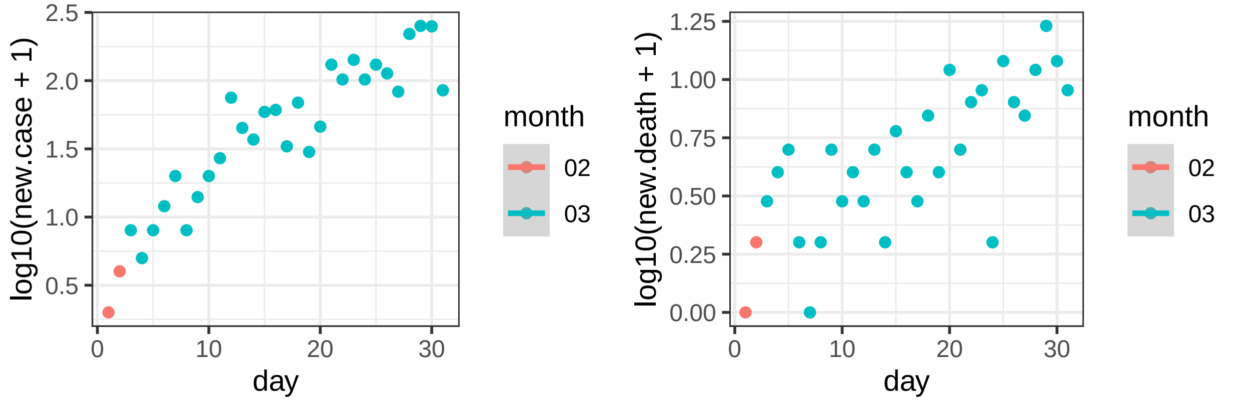
For these 10 counties, I check the number of new cases and the number of new deaths.

New York City_New York



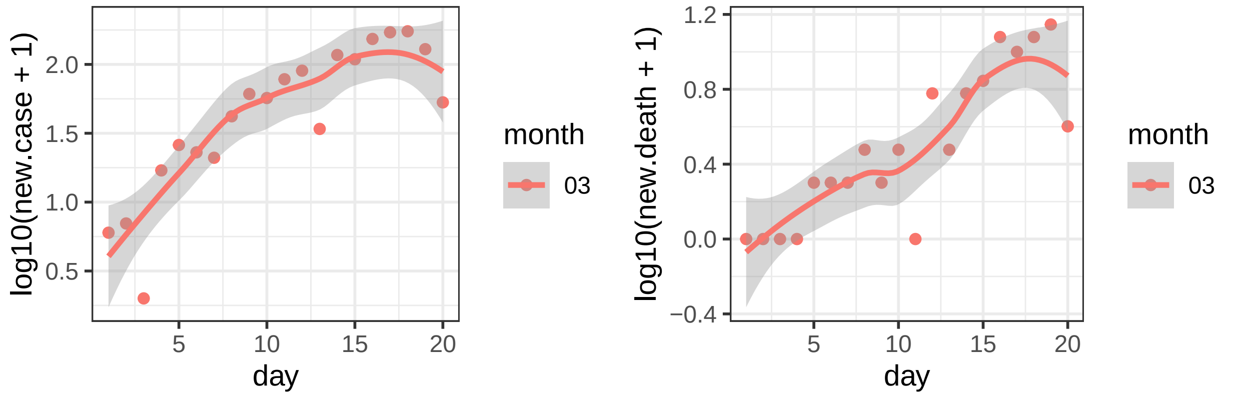
day 1 is 03−01

King_Washington



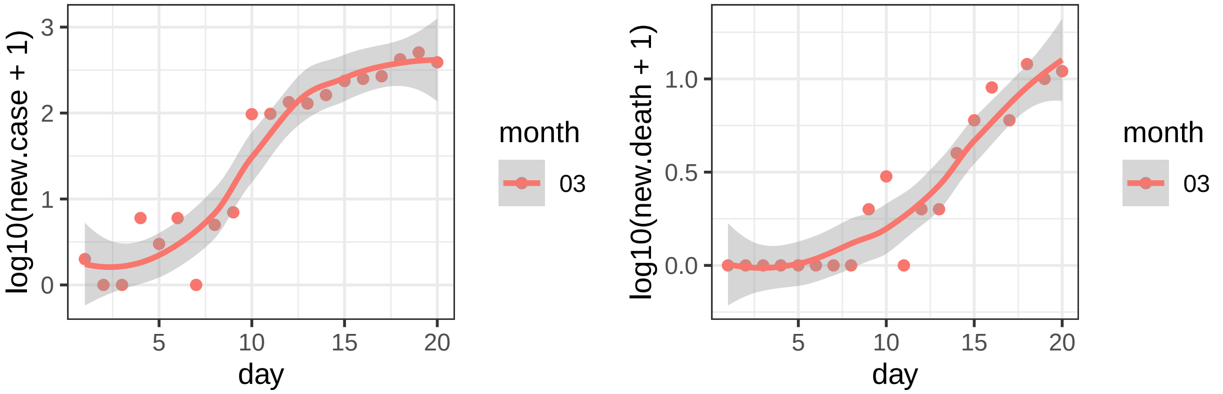
day 1 is 02−28

Orleans_Louisiana



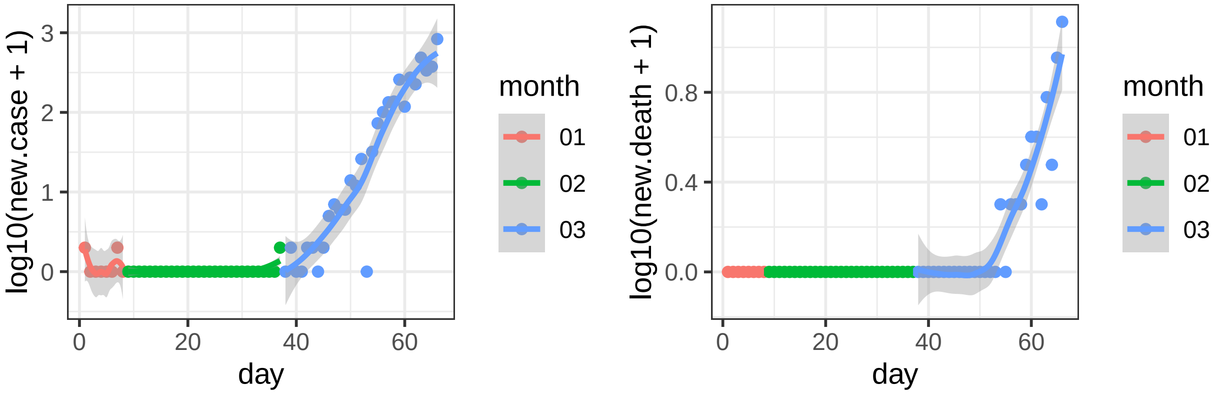
day 1 is 03−10

Wayne_Michigan



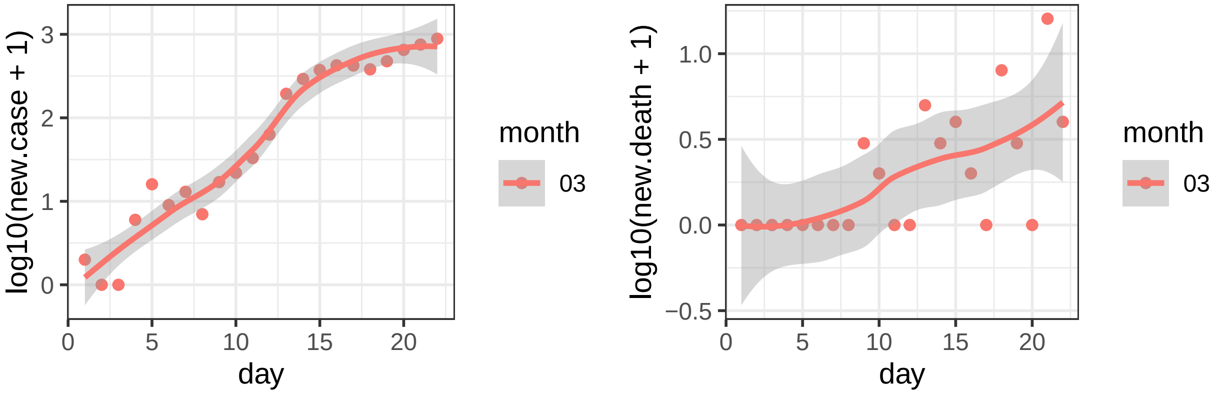
day 1 is 03−10

Cook_Illinois



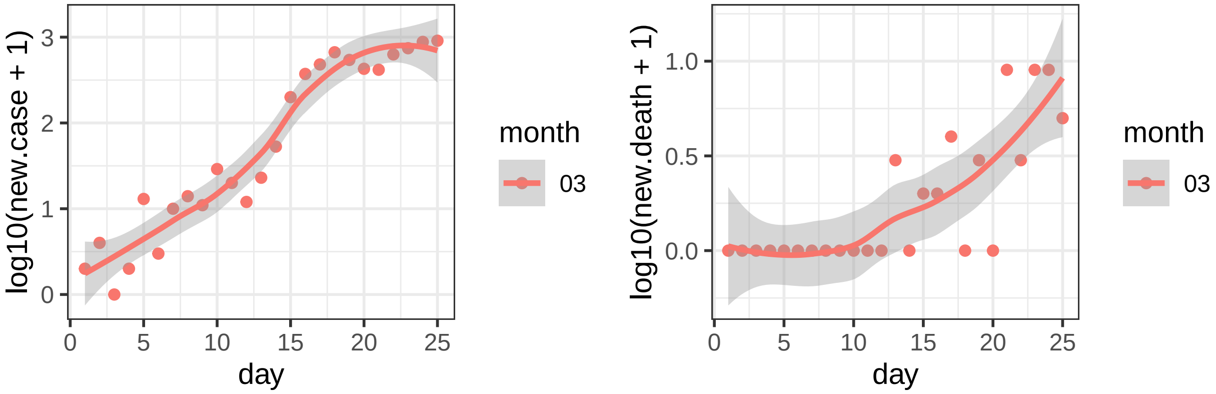
day 1 is 01−24

Suffolk_New York



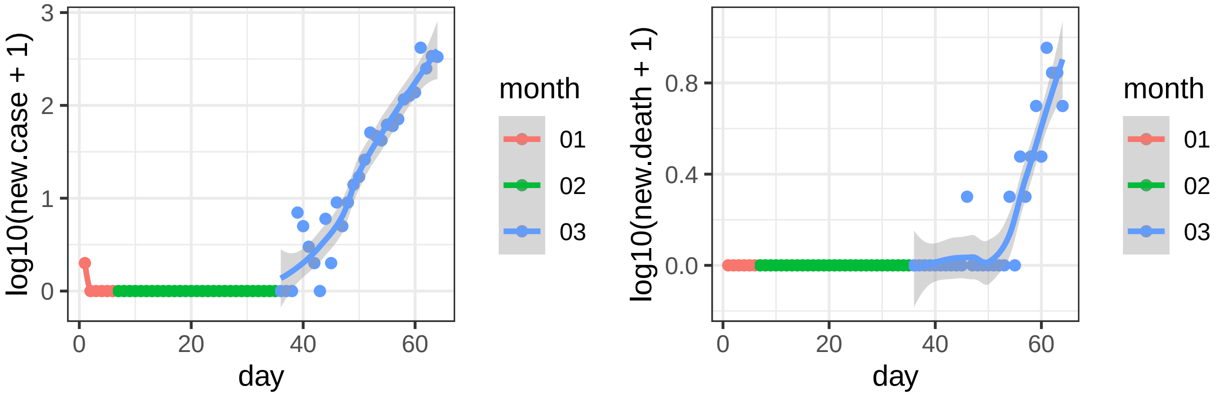
day 1 is 03−08

Nassau_New York



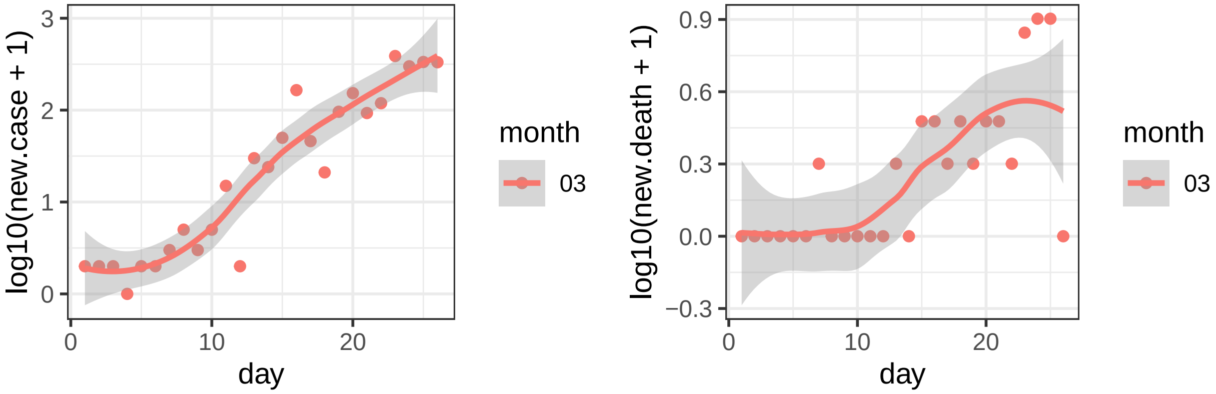
day 1 is 03−05

Los Angeles_California

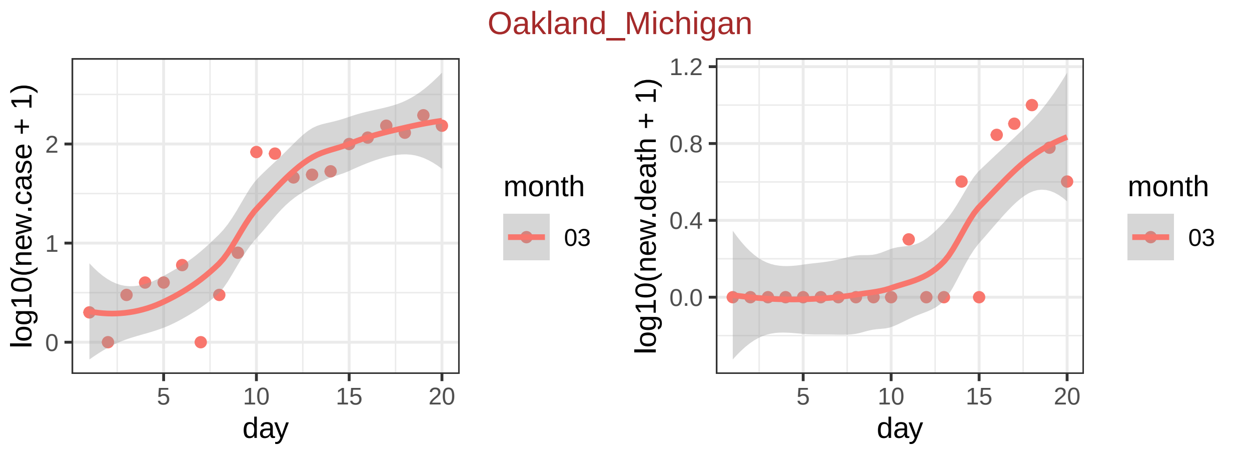


day 1 is 01−26

Bergen_New Jersey



day 1 is 03−04

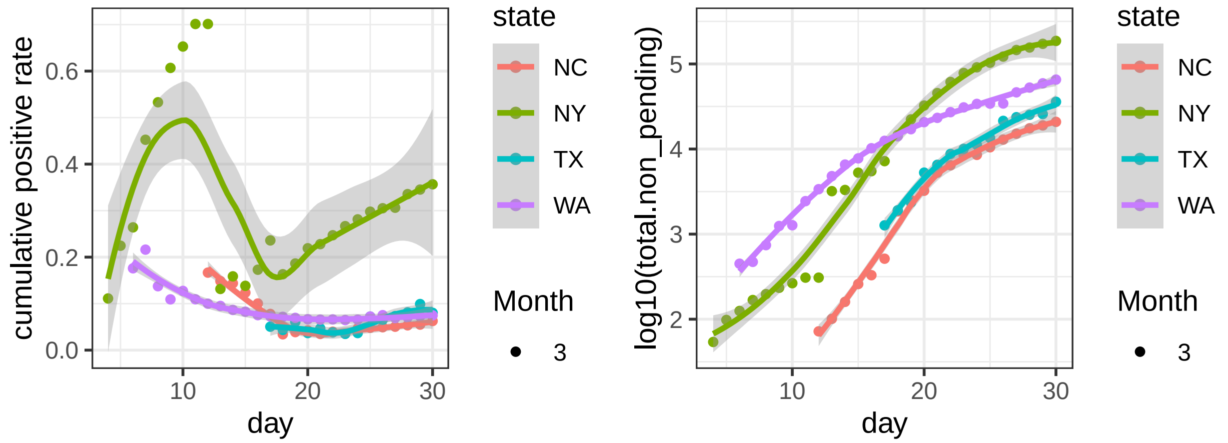


day 1 is 03−10

COVID Trackng

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 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.

Since the daily postive rate can fluctuate a lot, here I only illustrae the cumulative positave 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.



cumulative positive rate on 0330: 0.07(WA) 0.08(TX) 0.36(NY) 0.06(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
```