

COVID19 Data Hub, A curated COVID19 R Package

BARUG Meeting

Eric A. Suess

2/16/2021

COVID19 Data Hub

Today we will introduce and discuss the [COVID19 Hub](#) an R Package that provides access to current numbers related to COVID19.

The [COVID19 Data Hub](#) tries to provide access to a curated collection of data from as many countries around the world as possible. It is an open source package that encourages user suggestions and contributions.

```
> install.packages("COVID19")
```

It is one of the 15 “covid” packages that is currently available on CRAN.

```
library(pkgsearch)  
pkg_search("covid")
```

```
## - "covid" ----- 20 packages in 0.01 seconds  
-  
## # package version by @ title  
## 1 100 covid19jp 0.1.0 Koji Higuchi 1M Japanese Covid-19
```

| | | | | | | | |
|----------|----|----|-----|-------------------|--------|-------------------|-----------------------|
| Da... | ## | 2 | 100 | covid19france | 0.1.0 | Amanda Dobbyn | 10M Cases of COVID-19 |
| in... | ## | 3 | 92 | covid19us | 0.1.7 | Amanda Dobbyn | 5M Cases of COVID-19 |
| in... | ## | 4 | 92 | covid19br | 0.1.1 | Fabio Demarqui | 3M Brazilian COVID-19 |
| P... | ## | 5 | 92 | covidregionaldata | 0.8.2 | Sam Abbott | 2M Subnational Data |
| for... | ## | 6 | 89 | covid19swiss | 0.1.0 | Rami Krispin | 5M COVID-19 Cases in |
| Sw... | ## | 7 | 86 | covidprobability | 0.1.0 | Eric Brown | 6d Estimate the Unit- |
| Wi... | ## | 8 | 86 | oxcovid19 | 0.1.2 | Ernest Guevarra | 4M An R API to the |
| Oxfo... | ## | 9 | 86 | COVID19 | 2.3.2 | Emanuele Guidotti | 1M R Interface to |
| COVID... | ## | 10 | 78 | coronavirus | 0.3.21 | Rami Krispin | 17d The 2019 Novel |
| Coron... | | | | | | | |

About me

I am a Professor at [CSU East Bay](#) in [Statistics and Biostatistics](#), jointly appointed in the [Engineering](#). I have taught classes in Economics, Marketing, and Analytics for the College of Business. I am 5+ years former Chair, after 3 terms, so 9 years (or 14).

I am the Chief Statistician at [machineVantage](#) an AI and ML Neuroscience Marketing start-up company located in Berkeley, CA, Chennai and Bangalore, India, London, England. I am a ≤ 10 hour per week employee. Apply ML and AI algorithms for clients.

Now I am starting to work on the [COVID19 Data Hub](#) with Emanuele Guidotti and David Ardia. Emanuele is located in Switzerland and David is located in Montreal.

Why?

Well at the start of the Covid lock-down I decided *not* to say **No** to any project that came my way. I am now working on many interesting projects. This is the one that is likely to influence my teaching the most in terms of technical skills.

Joe asked and I said **Yes**.

I am hoping this effort is beneficial to:

1. The developers of the package.
2. The R community.
3. The R Consortium Covid19 Working Group.
4. My CSU East Bay colleagues, Ayona Chatterjee and Eric Fox.
5. My current students who are working on Covid19 data projects.
6. Me. Hopefully I can develop more "developer" skills that I can pass on to my students.

COVID19 Data Hub

The COVID19 Data Hub is an R package that pulls data from a curated collection of data [sources](#) that is updated hourly. The data is downloaded and merged together into one file once an hour and can be access through one function in R (or using other frontends).

```
> library(COVID19)
> x_USA <- covid19("USA")
> x_USA
```

The [data](#) is downloaded from many many data sources by code running on a GCP server in the Cloud. The data is processed from the various sources to populate [three levels of data](#). At the end of each day a vintage dataset is made a available.

The levels:

- `administrative_area_level_1` = *Country* level data, totals
- `administrative_area_level_2` = *State* level data

- administrative_area_level_3 = *County* level data

COVID19 Data Hub

There are so many different sources of COVID19 data. Every country, every state and every city has its own data. There are many different government websites, many universities, and many companies.

- [Our World Data](#)
- [The Covid Tracking Project](#)
- [John Hopkins University](#)
- [New York Times](#)

It is going to be an ongoing challenge to maintain all of the connections to the original sources. It is already the case that some of the original sources will be ending their efforts soon.

What can you do with the data?

Below are some examples of the use of some possible uses of the data. I am currently teaching a Time Series course using the [fpp3](#) book and a graduate Statistical Learning class using the [mdsr2e](#) book. So the examples that follow use of of the R packages used in these books.

There is also an excellent tutorial posted on Medium's Toward Data Science [COVID-19 Data Acquisition in R](#) that give further details on how to extend the dataset in real time.

```
library(pacman)
p_load(COVID19, tidyverse, fpp3, naniar)
```

Load the country level data for the United States.

```
x_USA <- covid19("USA", verbose = FALSE)
```

```
## Warning in id(x$country, iso = "ISO", ds = "jhucsse_git", level = 1): missing
## id: Micronesia
```

Time plot of the cumulative deaths.

```
x_USA %>% select(date, deaths) %>%  
  as_tsibble() %>%  
  autoplot() +  
  labs(title = "USA Covid19 Deaths")
```

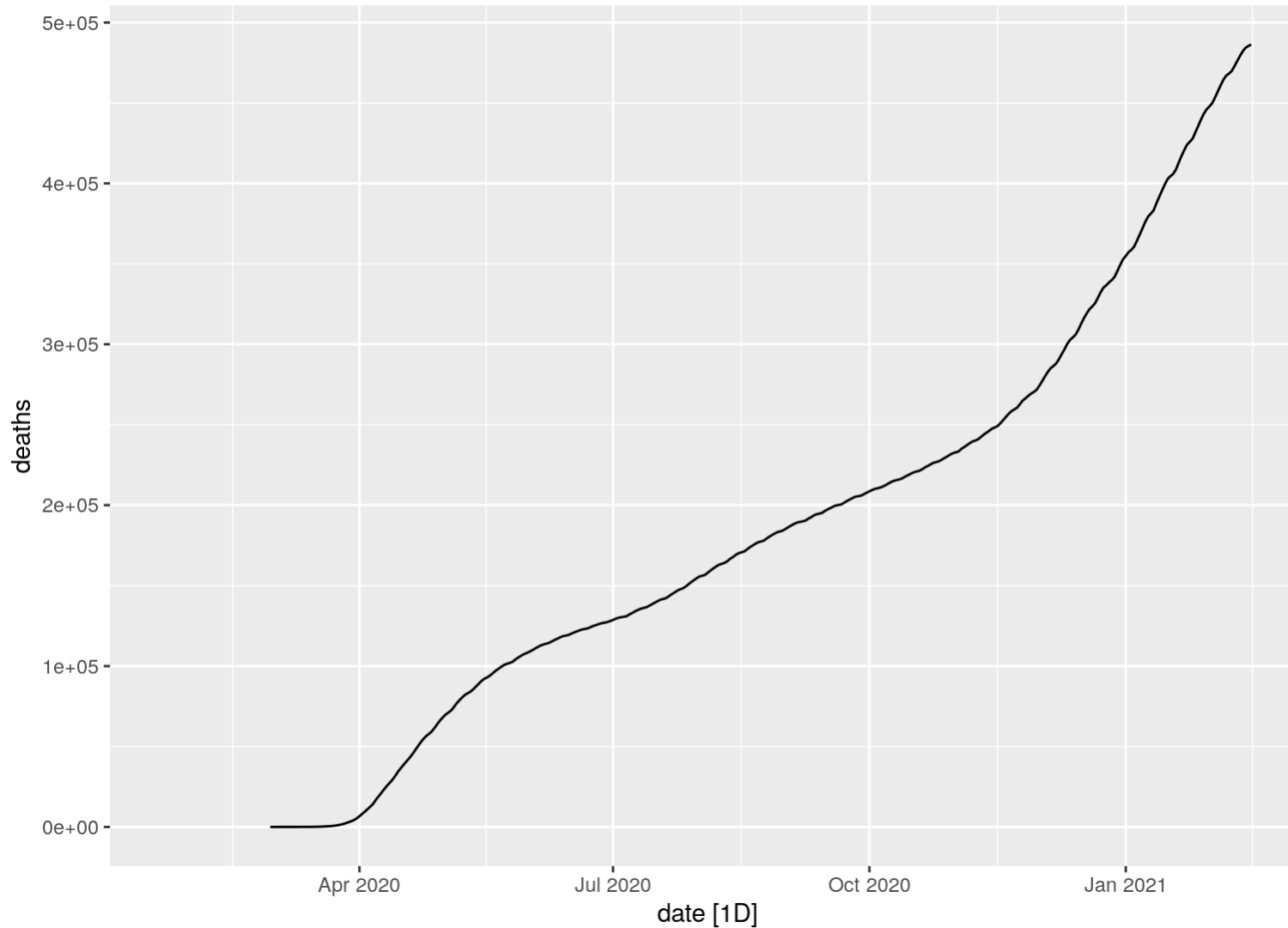
```
## Adding missing grouping variables: `id`
```

```
## Using `date` as index variable.
```

```
## Plot variable not specified, automatically selected `.vars = deaths`
```

```
## Warning: Removed 38 row(s) containing missing values (geom_path).
```

USA Covid19 Deaths



Using the lag() function we can determine daily counts.

```
x_USA %>% select(date, deaths) %>%  
  mutate(daily_deaths = deaths - lag(deaths)) %>%
```

```
as_tsibble() %>%  
tail(10)
```

```
## Adding missing grouping variables: `id`
```

```
## Using `date` as index variable.
```

```
## # A tsibble: 10 x 4 [1D]  
## # Groups:   id [1]  
##   id     date      deaths daily_deaths  
##   <chr> <date>      <dbl>      <dbl>  
## 1 USA   2021-02-06  466890      2546  
## 2 USA   2021-02-07  468204      1314  
## 3 USA   2021-02-08  469786      1582  
## 4 USA   2021-02-09  472818      3032  
## 5 USA   2021-02-10  476100      3282  
## 6 USA   2021-02-11  479257      3157  
## 7 USA   2021-02-12  482142      2885  
## 8 USA   2021-02-13  484301      2159  
## 9 USA   2021-02-14  485384      1083  
## 10 USA  2021-02-15  486325       941
```

Plotting the daily counts reveals a weekly seasonal pattern in the time series.

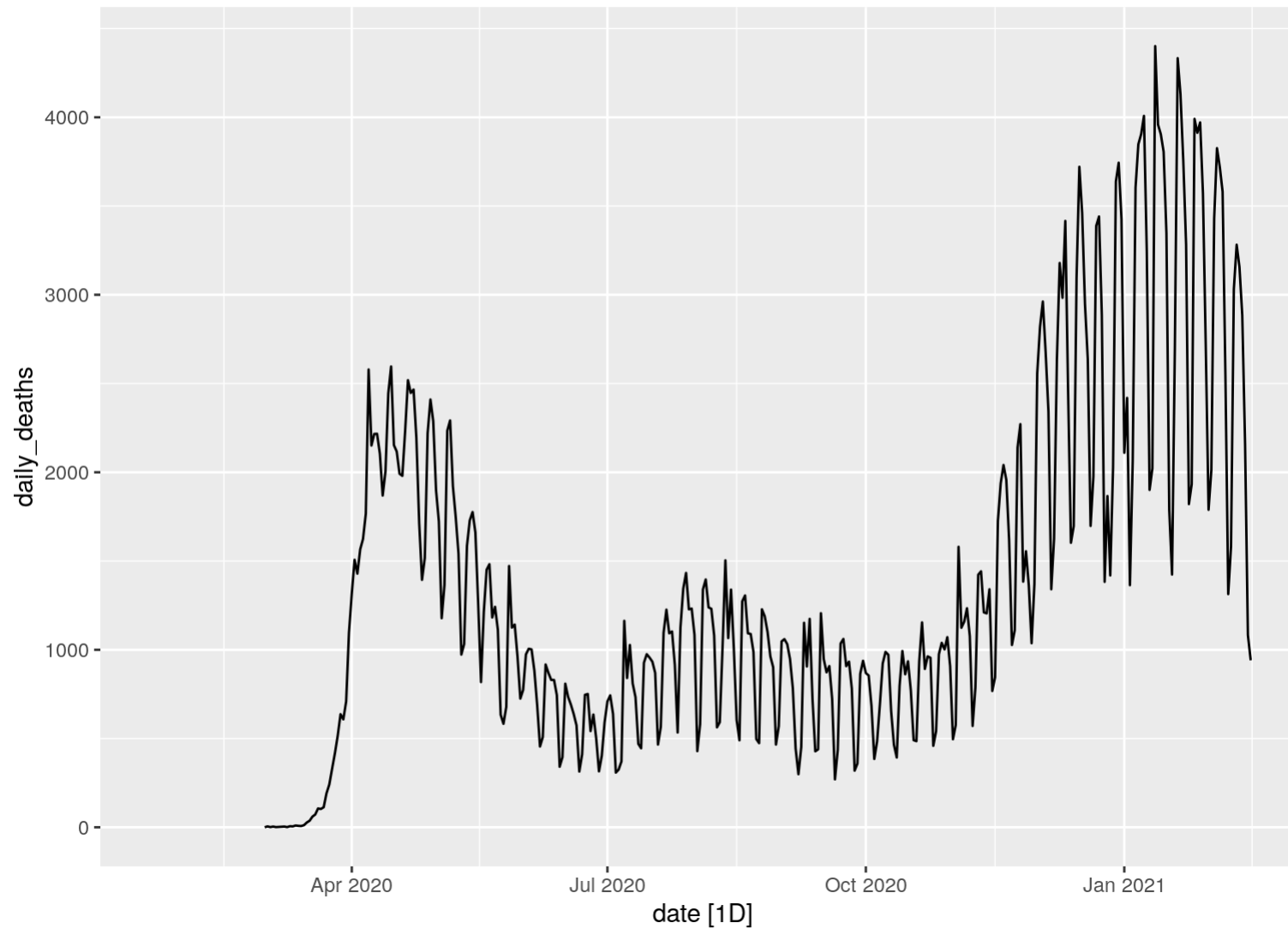
```
x_USA %>% select(date, deaths) %>%  
  mutate(daily_deaths = deaths - lag(deaths)) %>%  
  as_tsibble() %>%  
  autoplot(daily_deaths) +  
  labs(title = "USA Covid19 Daily Deaths")
```

```
## Adding missing grouping variables: `id`
```

```
## Using `date` as index variable.
```

```
## Warning: Removed 39 row(s) containing missing values (geom_path).
```

USA Covid19 Daily Deaths



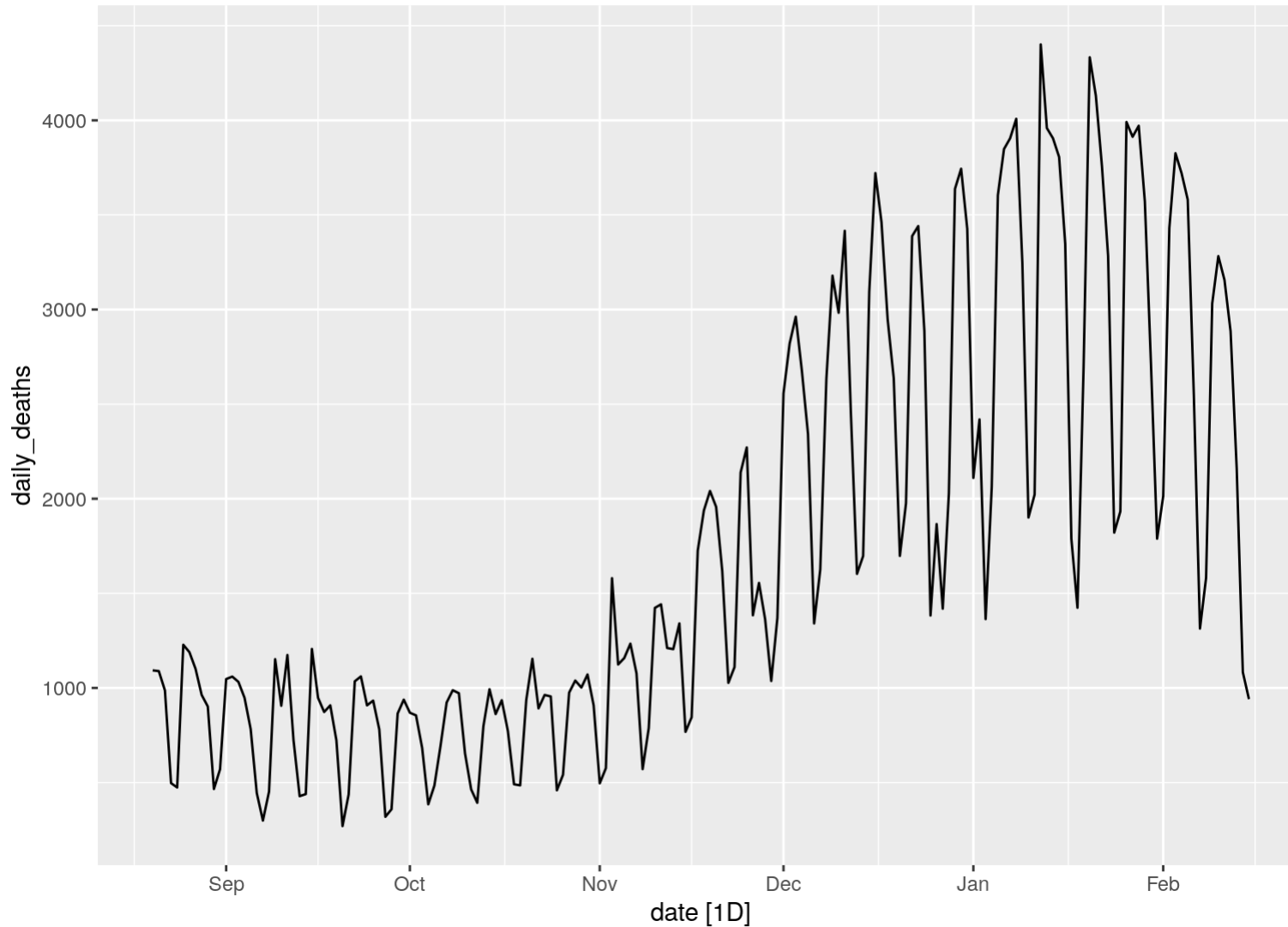
Looking at the last 6 months.

```
x_USA %>% select(date, deaths) %>%  
  mutate(daily_deaths = deaths - lag(deaths)) %>%  
  as_tsibble() %>%  
  tail(180) %>%  
  autoplot(daily_deaths) +  
  labs(title = "USA Covid19 Daily Deaths")
```

```
## Adding missing grouping variables: `id`
```

```
## Using `date` as index variable.
```

USA Covid19 Daily Deaths



Trying a multiplicative Classical Decomposition Model to see the Trend and Seasonal components in the time series.

```
x_USA %>% select(date, deaths) %>%  
  mutate(daily_deaths = deaths - lag(deaths)) %>%  
  as_tsibble() %>%  
  tail(180) %>%  
  model(classical_decomposition(daily_deaths, type = "multiplicative")) %>%  
  components() %>%  
  autoplot() +  
  labs(title = "Classical multiplicative decomposition of USA Covid19 Daily Deaths")
```

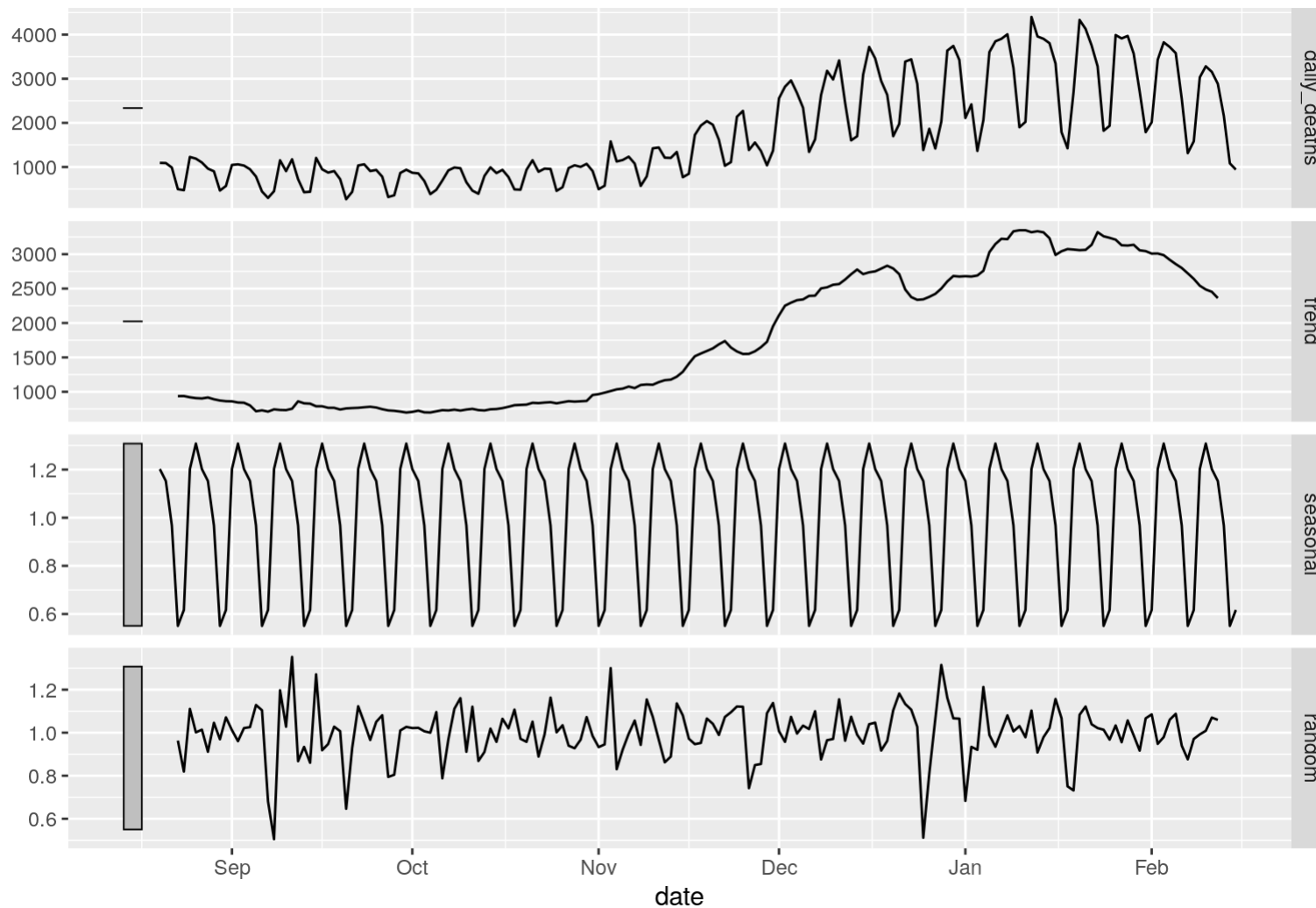
```
## Adding missing grouping variables: `id`
```

```
## Using `date` as index variable.
```

```
## Warning: Removed 3 row(s) containing missing values (geom_path).
```


Classical multiplicative decomposition of USA Covid19 Daily Deaths

$\text{daily_deaths} = \text{trend} * \text{seasonal} * \text{random}$



Computing some features of the time series.

```
x_USA %>% select(date, deaths) %>%  
  mutate(daily_deaths = deaths - lag(deaths)) %>%
```

```
as_tsibble() %>%
tail(180) %>%
select(date, daily_deaths) %>%
features(daily_deaths, feat_stl)
```

```
## Adding missing grouping variables: `id`
```

```
## Using `date` as index variable.
```

```
## Adding missing grouping variables: `id`
```

```
## # A tibble: 1 x 9
##   trend_strength seasonal_streng~ seasonal_peak_w~ seasonal_trough~ spikiness
##   <dbl>           <dbl>           <dbl>           <dbl>         <dbl>
## 1         0.954         0.873             0             4        576880.
## # ... with 4 more variables: linearity <dbl>, curvature <dbl>,
## #   stl_e_acf1 <dbl>, stl_e_acf10 <dbl>
```

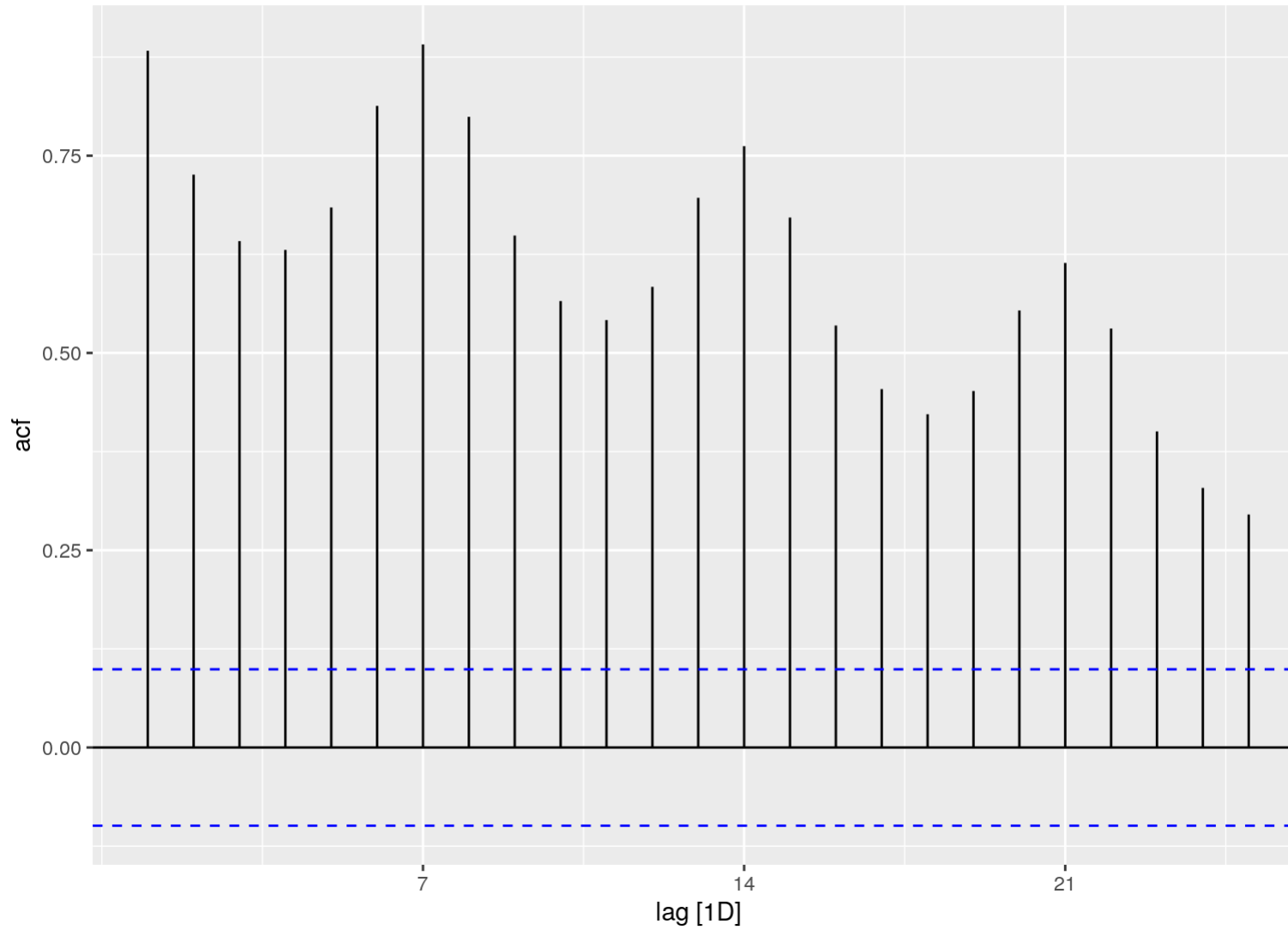
Autocorrelation plot. (See Allison Horst's new series on the ACF post on Twitter @allison_horst yesterday.)

```
x_USA %>% select(date, deaths) %>%
mutate(daily_deaths = deaths - lag(deaths)) %>%
as_tsibble() %>%
ACF(daily_deaths) %>%
autoplot() +
labs(title = "USA Covid19 Daily Deaths")
```

```
## Adding missing grouping variables: `id`
```

```
## Using `date` as index variable.
```

USA Covid19 Daily Deaths



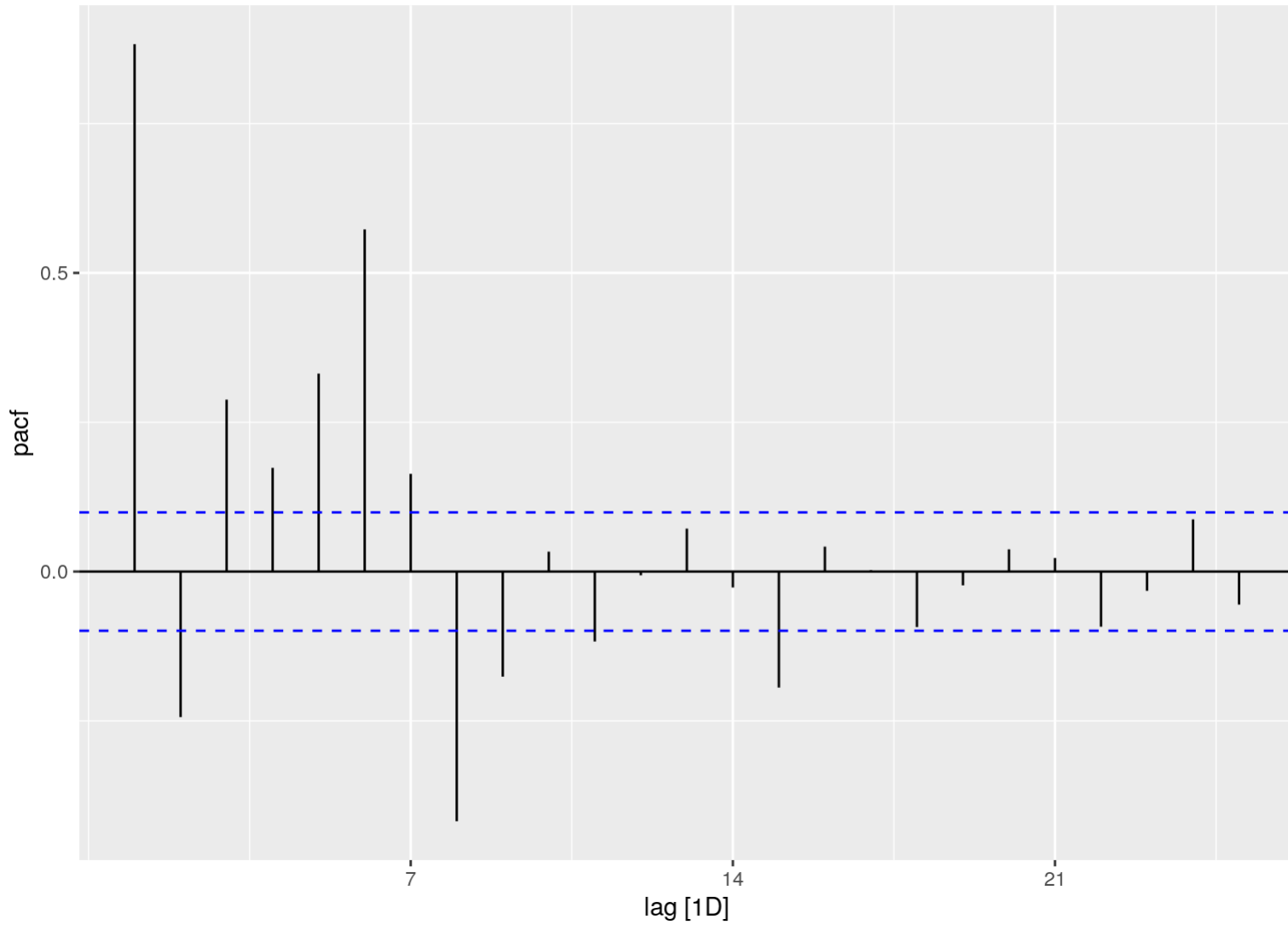
PACF

```
x_USA %>% select(date, deaths) %>%  
  mutate(daily_deaths = deaths - lag(deaths)) %>%  
  as_tsibble() %>%  
  PACF(daily_deaths) %>%  
  autoplot() +  
  labs(title = "USA Covid19 Daily Deaths")
```

```
## Adding missing grouping variables: `id`
```

```
## Using `date` as index variable.
```

USA Covid19 Daily Deaths



Note: The time series is not stationary, so need to take another difference.

Comparisons

Brazil

```
x_BRA <- covid19("BRA", verbose = FALSE)
tail(x_BRA, 10)
```

```
## # A tibble: 10 x 36
## # Groups:   id [1]
##   id      date      vaccines tests confirmed recovered deaths hosp vent
##   <chr> <date>      <dbl> <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl>
##   <dbl>
## 1 BRA    2021-02-06  3401383    NA    9447165    8428992 230034    NA    NA
##   NA
## 2 BRA    2021-02-07  3553681    NA    9524640    8467982 231534    NA    NA
##   NA
## 3 BRA    2021-02-08  3605538    NA    9524640    8478818 231534    NA    NA
##   NA
## 4 BRA    2021-02-09  3820207    NA    9599565    8577207 233520    NA    NA
##   NA
## 5 BRA    2021-02-10  4120332    NA    9659167    8616282 234850    NA    NA
##   NA
## 6 BRA    2021-02-11  4406835    NA    9713909    8637050 236201    NA    NA
##   NA
## 7 BRA    2021-02-12  4696136    NA    9765455    8691664 237489    NA    NA
##   NA
## 8 BRA    2021-02-13  5125206    NA    9809754    8740445 238532    NA    NA
##   NA
## 9 BRA    2021-02-14  5236943    NA    9834513    8765048 239245    NA    NA
```

```
NA
## 10 BRA 2021-02-15 5293979 NA 9866710 8821887 239773 NA NA
NA
## # ... with 26 more variables: population <dbl>, school_closing <int>,
## # workplace_closing <int>, cancel_events <int>,
## # gatherings_restrictions <int>, transport_closing <int>,
## # stay_home_restrictions <int>, internal_movement_restrictions <int>,
## # international_movement_restrictions <int>, information_campaigns <int>,
## # testing_policy <int>, contact_tracing <int>, stringency_index <dbl>,
## # iso_alpha_3 <chr>, iso_alpha_2 <chr>, iso_numeric <int>, currency <chr>,
## # administrative_area_level <chr>, administrative_area_level_1 <chr>,
## # administrative_area_level_2 <chr>, administrative_area_level_3 <chr>,
## # latitude <dbl>, longitude <dbl>, key <lgl>, key_apple_mobility <chr>,
## # key_google_mobility <chr>
```

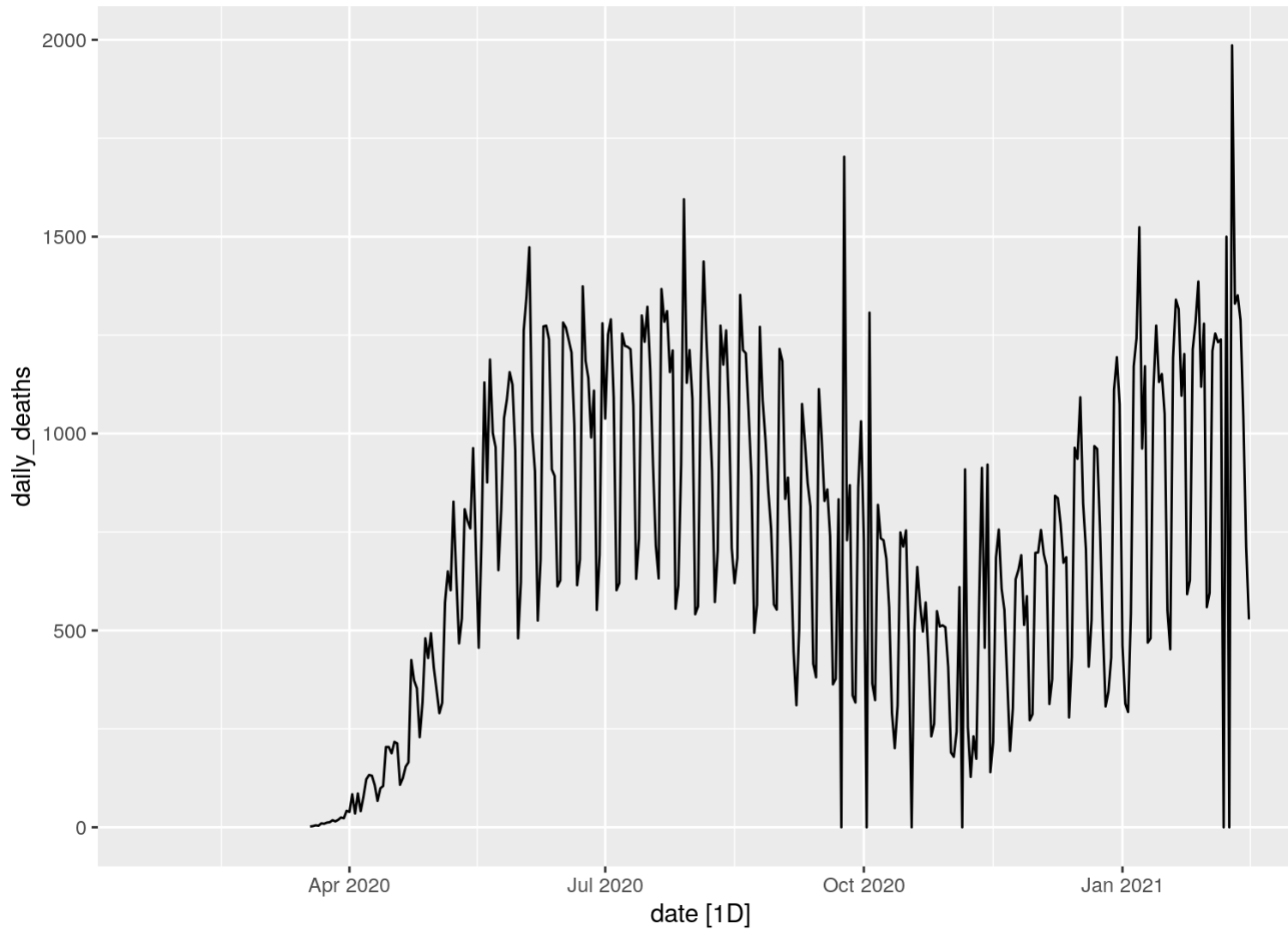
```
x_BRA %>% select(date, deaths) %>%
  mutate(daily_deaths = deaths - lag(deaths)) %>%
  as_tsibble() %>%
  autoplot(daily_deaths) +
  labs(title = "Brazil Covid19 Daily Deaths")
```

```
## Adding missing grouping variables: `id`
```

```
## Using `date` as index variable.
```

```
## Warning: Removed 56 row(s) containing missing values (geom_path).
```


Brazil Covid19 Daily Deaths



```
x_USA_BRA <- covid19(c("USA","BRA"), verbose = FALSE)
tail(x_USA_BRA, 10)
```

```
## # A tibble: 10 x 36
## # Groups:   id [1]
##   id      date      vaccines    tests confirmed recovered deaths  hosp  vent
##   <chr> <date>      <dbl>    <dbl>    <dbl>    <dbl>  <dbl> <dbl> <dbl>
## 1 USA    2021-02-06 39037964  3.11e8  26917787      NA 466890 84233   NA
## 2 USA    2021-02-07 41210937  3.11e8  27007368      NA 468204 81439   NA
## 3 USA    2021-02-08 42417617  3.12e8  27097095      NA 469786 80055   NA
## 4 USA    2021-02-09 43206190  3.12e8  27192455      NA 472818 79179   NA
## 5 USA    2021-02-10 44769970  NA      27287159      NA 476100 76979   NA
## 6 USA    2021-02-11 46390270  NA      27392512      NA 479257 74225   NA
## 7 USA    2021-02-12 48410558  NA      27492023      NA 482142    NA     NA
## 8 USA    2021-02-13 50641884  NA      27575344      NA 484301    NA     NA
## 9 USA    2021-02-14 52884356  NA      27640282      NA 485384    NA     NA
## 10 USA    2021-02-15      NA NA      27694165      NA 486325    NA     NA
## # ... with 27 more variables: icu <dbl>, population <dbl>,
## # school_closing <int>, workplace_closing <int>, cancel_events <int>,
## # gatherings_restrictions <int>, transport_closing <int>,
## # stay_home_restrictions <int>, internal_movement_restrictions <int>,
## # international_movement_restrictions <int>, information_campaigns <int>,
## # testing_policy <int>, contact_tracing <int>, stringency_index <dbl>,
## # iso_alpha_3 <chr>, iso_alpha_2 <chr>, iso_numeric <int>, currency <chr>,
## # administrative_area_level <chr>, administrative_area_level_1 <chr>,
## # administrative_area_level_2 <chr>, administrative_area_level_3 <chr>,
## # latitude <dbl>, longitude <dbl>, key <lgl>, key_apple_mobility <chr>,
## # key_google_mobility <chr>
```

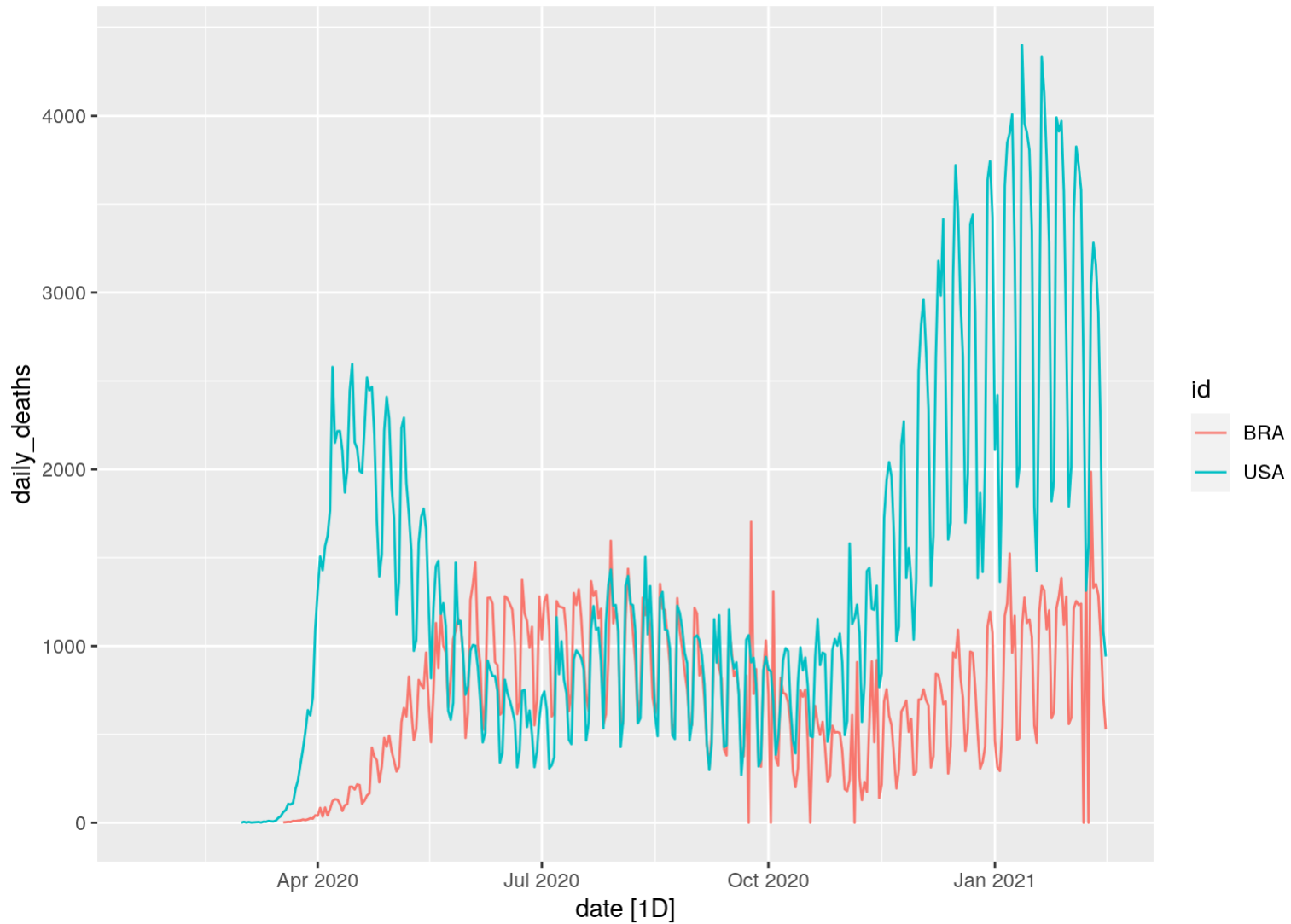
```
x_USA_BRA %>% select(date, deaths) %>%
  mutate(daily_deaths = deaths - lag(deaths)) %>%
  as_tsibble(key = id, index = date) %>%
  autoplot(daily_deaths) +
  labs(title = "USA and Brazil Covid19 Daily Deaths")
```

```
## Adding missing grouping variables: `id`
```

```
## `mutate_if()` ignored the following grouping variables:  
## Column `id`
```

```
## Warning: Removed 95 row(s) containing missing values (geom_path).
```

USA and Brazil Covid19 Daily Deaths



Estonia, Lithuania, and Latvia

```
x_three <- covid19(c("EST","LTU","LVA"), verbose = FALSE)
tail(x_three, 10)
```

```
## # A tibble: 10 x 36
## # Groups:   id [1]
##   id     date      vaccines  tests confirmed recovered deaths  hosp  vent
##   <chr> <date>      <dbl>  <dbl>   <dbl>    <dbl> <dbl> <dbl> <dbl>
##   <dbl>
## 1 LVA    2021-02-07    32250 1.24e6    71800    59897  1339   NA    NA
##   NA
## 2 LVA    2021-02-08    32661 1.25e6    72088    60117  1347   NA    NA
##   NA
## 3 LVA    2021-02-09    32845 1.26e6    72869    60798  1363   NA    NA
##   NA
## 4 LVA    2021-02-10    33452 1.27e6    73859    61889  1395   NA    NA
##   NA
## 5 LVA    2021-02-11    35098 1.28e6    74701    62844  1416   NA    NA
##   NA
## 6 LVA    2021-02-12    36644 1.30e6    75509    62844  1431   NA    NA
##   NA
## 7 LVA    2021-02-13    37043 1.31e6    76282    64528  1443   NA    NA
##   NA
## 8 LVA    2021-02-14    37063 1.31e6    76706    65046  1451   NA    NA
##   NA
## 9 LVA    2021-02-15         NA 1.31e6    76984    65450  1468   NA    NA
##   NA
## 10 LVA    2021-02-16         NA 1.33e6    77697         NA  1486   NA    NA
##   NA
## # ... with 26 more variables: population <dbl>, school_closing <int>,
## #   workplace_closing <int>, cancel_events <int>,
```

```
## # gatherings_restrictions <int>, transport_closing <int>,  
## # stay_home_restrictions <int>, internal_movement_restrictions <int>,  
## # international_movement_restrictions <int>, information_campaigns <int>,  
## # testing_policy <int>, contact_tracing <int>, stringency_index <dbl>,  
## # iso_alpha_3 <chr>, iso_alpha_2 <chr>, iso_numeric <int>, currency <chr>,  
## # administrative_area_level <chr>, administrative_area_level_1 <chr>,  
## # administrative_area_level_2 <chr>, administrative_area_level_3 <chr>,  
## # latitude <dbl>, longitude <dbl>, key <lgl>, key_apple_mobility <chr>,  
## # key_google_mobility <chr>
```

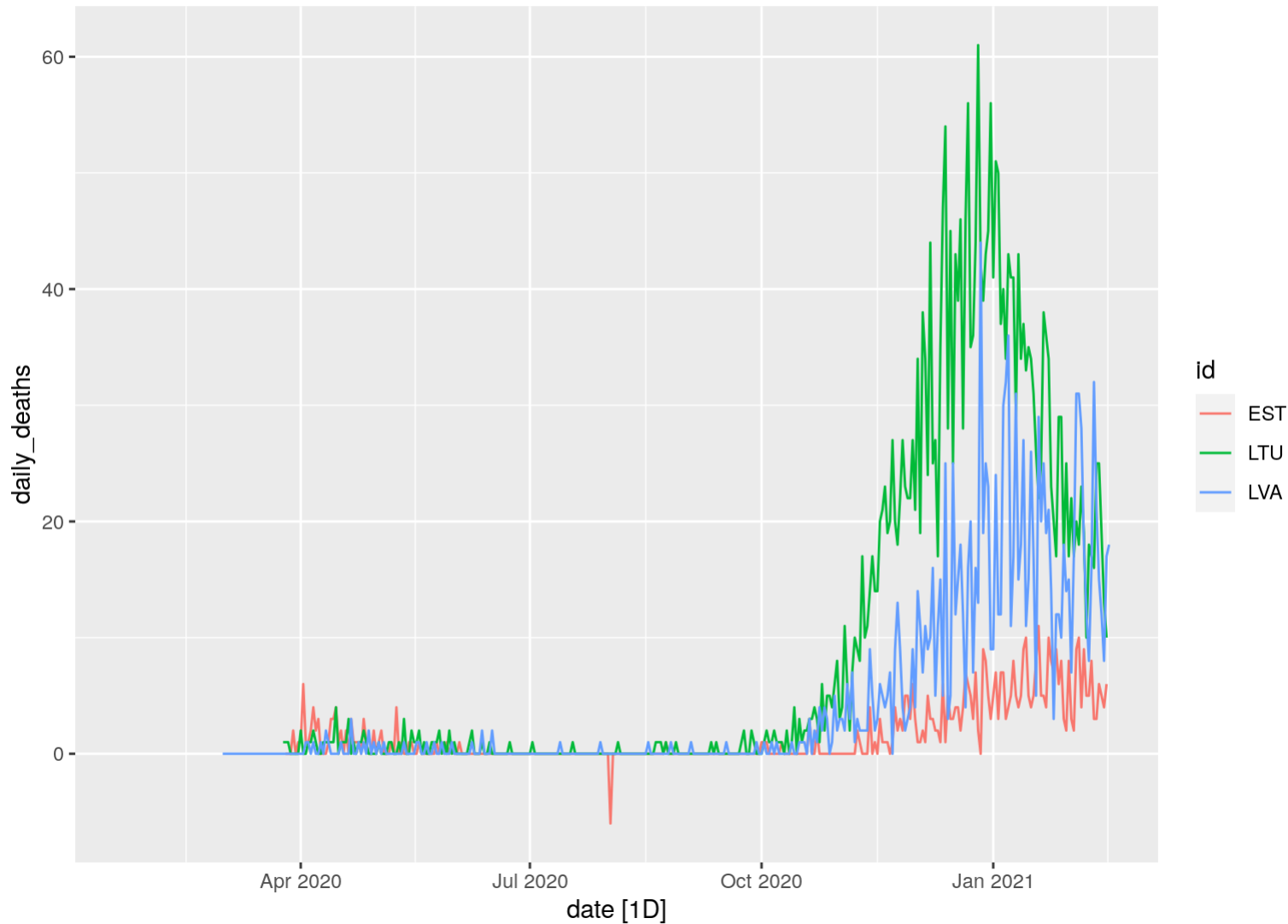
```
x_three %>% select(date, deaths) %>%  
  mutate(daily_deaths = deaths - lag(deaths)) %>%  
  as_tsibble(key = id, index = date) %>%  
  autoplot(daily_deaths) +  
  labs(title = "Covid19 Daily Deaths")
```

```
## Adding missing grouping variables: `id`
```

```
## `mutate_if()` ignored the following grouping variables:  
## Column `id`
```

```
## Warning: Removed 166 row(s) containing missing values (geom_path).
```

Covid19 Daily Deaths



Summarize the data weekly.

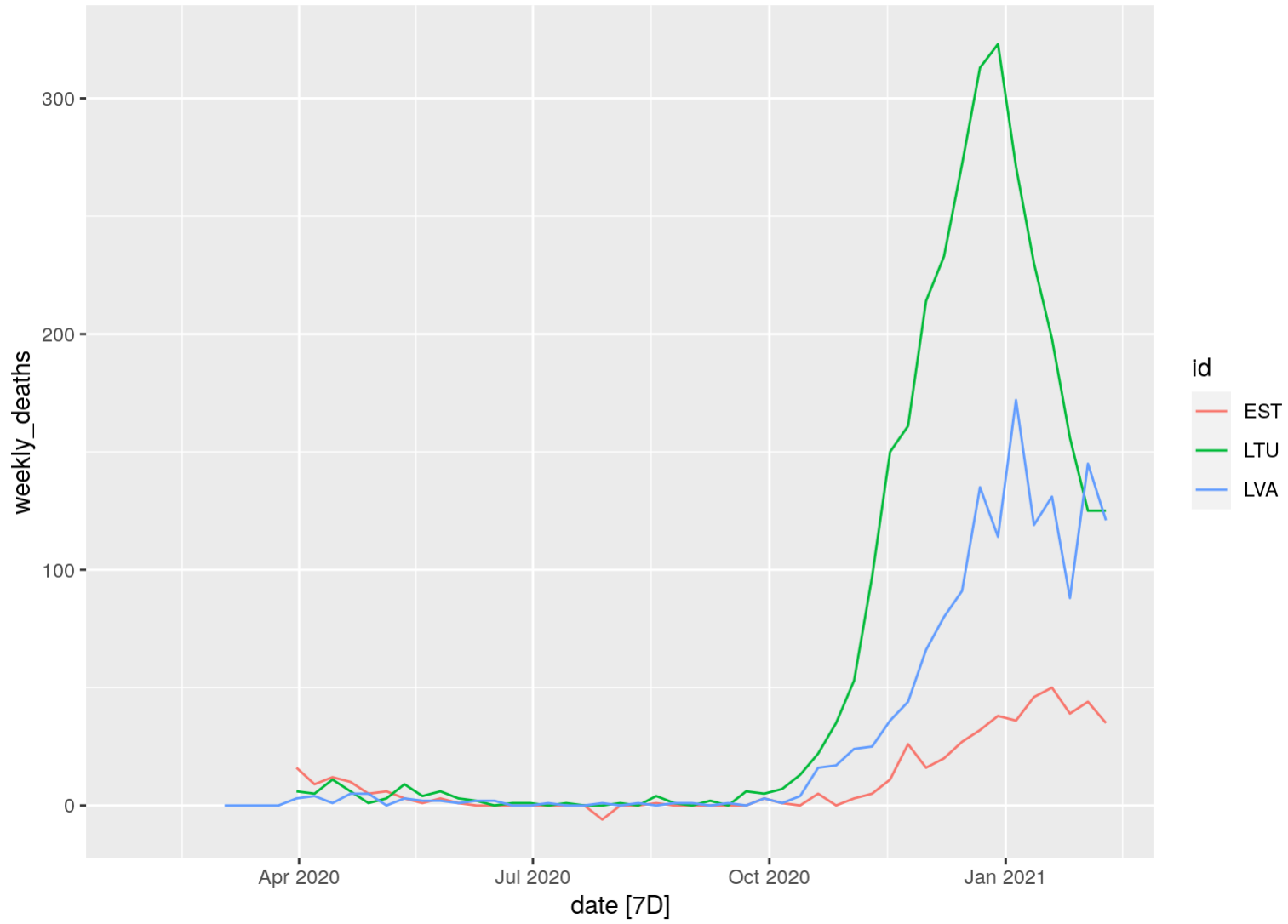
```
x_three %>% select(date, deaths) %>%  
  mutate(daily_deaths = deaths - lag(deaths)) %>%
```

```
as_tsibble(key = id, index = date) %>%  
# Currently only supports daily data  
index_by(date) %>%  
summarise(weekly_deaths = sum(daily_deaths)) %>%  
# Compute weekly aggregates  
fabletools::aggregate_index("1 week", weekly_deaths = sum(weekly_deaths)) %>%  
autoplot(weekly_deaths) +  
labs(title = "Covid19 Weekly Deaths")
```

```
## Adding missing grouping variables: `id`
```

```
## Warning: Removed 23 row(s) containing missing values (geom_path).
```

Covid19 Weekly Deaths



Completeness of the data

We can do a data availability study.

Estonia, Lithuania, and Latvia

```
x_three %>% anyNA()
```

```
## [1] TRUE
```

```
x_three %>% n_miss()
```

```
## [1] 7924
```

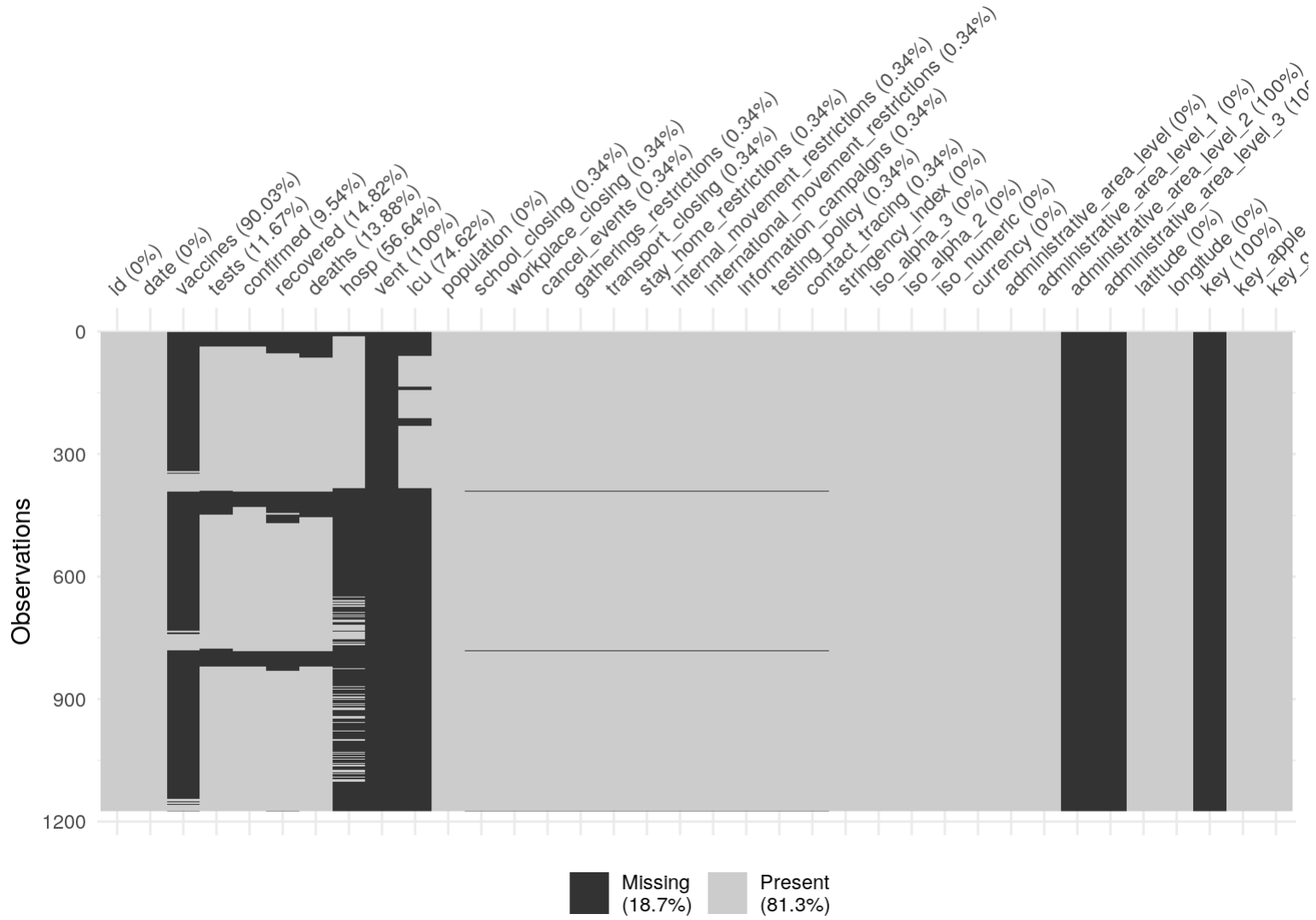
```
x_three %>% prop_miss()
```

```
## [1] 0.1874882
```

Visualize the missing values.

```
library(visdat)
```

```
x_three %>% group_by(id) %>%  
  vis_miss()
```



Administrative level 2

```
x_USA_state <- covid19("USA", level = 2)
```

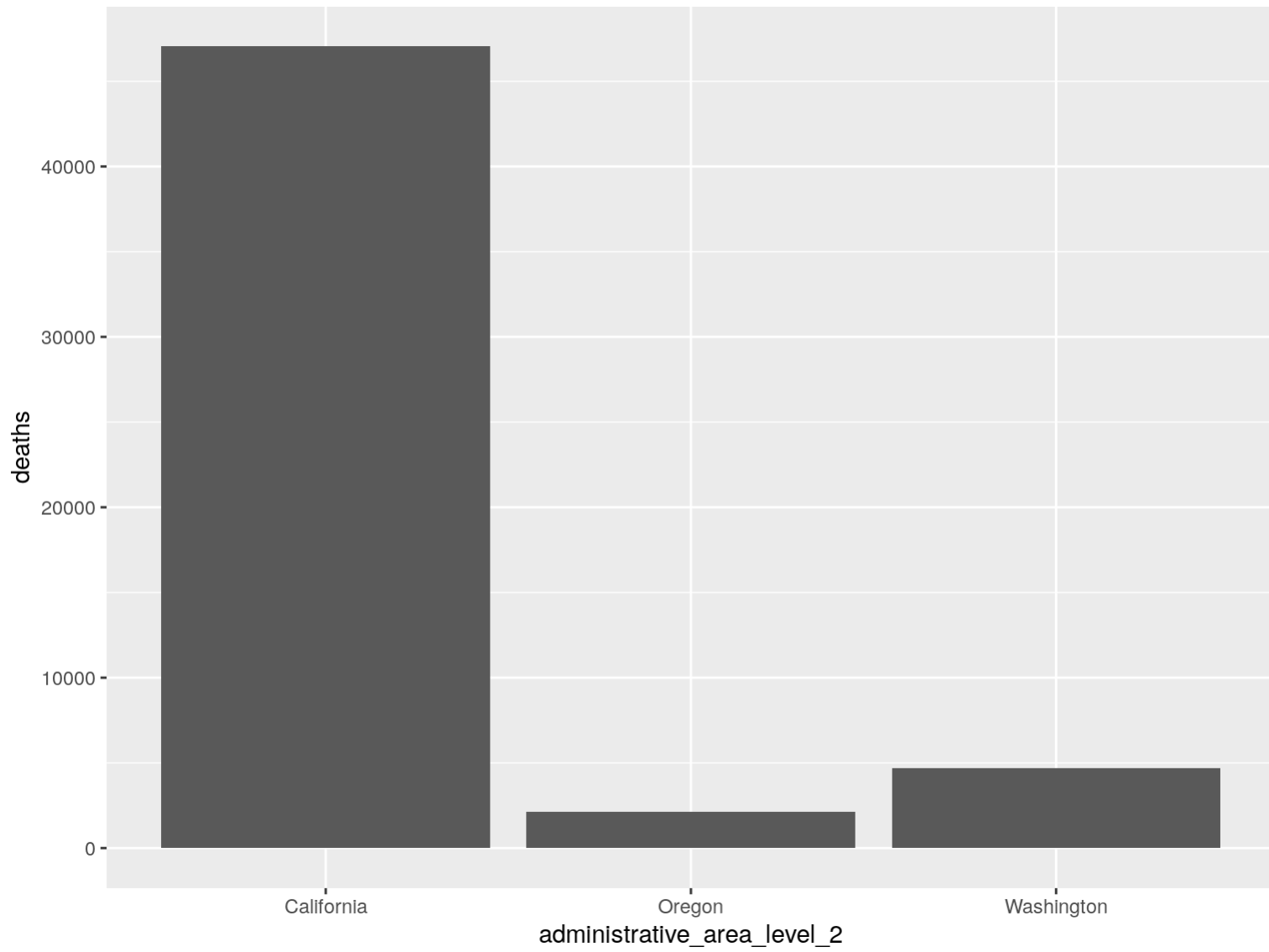
```
## Warning in id(x$state, iso = iso[[1]], ds = "jhucsse_git", level = level):  
## missing id: Nunavut, Repatriated Travellers
```

```
## Warning in id(x$state, iso = iso[[1]], ds = "jhucsse_git", level = level):  
## missing id: Wallis and Futuna
```

```
##  
## Hale Thomas, Sam Webster, Anna Petherick, Toby Phillips, and Beatriz  
## Kira (2020). Oxford COVID-19 Government Response Tracker, Blavatnik  
## School of Government.  
##  
## The COVID Tracking Project (2020), https://covidtracking.com  
##  
## Johns Hopkins Center for Systems Science and Engineering (2020),  
## https://github.com  
##  
## Guidotti, E., Ardia, D., (2020), "COVID-19 Data Hub", Journal of Open  
## Source Software 5(51):2376, doi: 10.21105/joss.02376.  
##  
## To see these entries in BibTeX format, use 'print(<citation>,  
## bibtex=TRUE)', 'toBibtex(.)', or set  
## 'options(citation.bibtex.max=999)'.  
##  
## To hide the data sources use 'verbose = FALSE'.
```

```
x_USA_state %>% select(date, administrative_area_level_2, deaths) %>%  
  filter(date == "2021-02-15") %>%  
  filter(administrative_area_level_2 %in% c("California", "Oregon", "Washington")) %>%  
  ggplot(aes(x = administrative_area_level_2, y = deaths)) +  
  geom_bar(stat="identity")
```

```
## Adding missing grouping variables: `id`
```



Administrative level 3

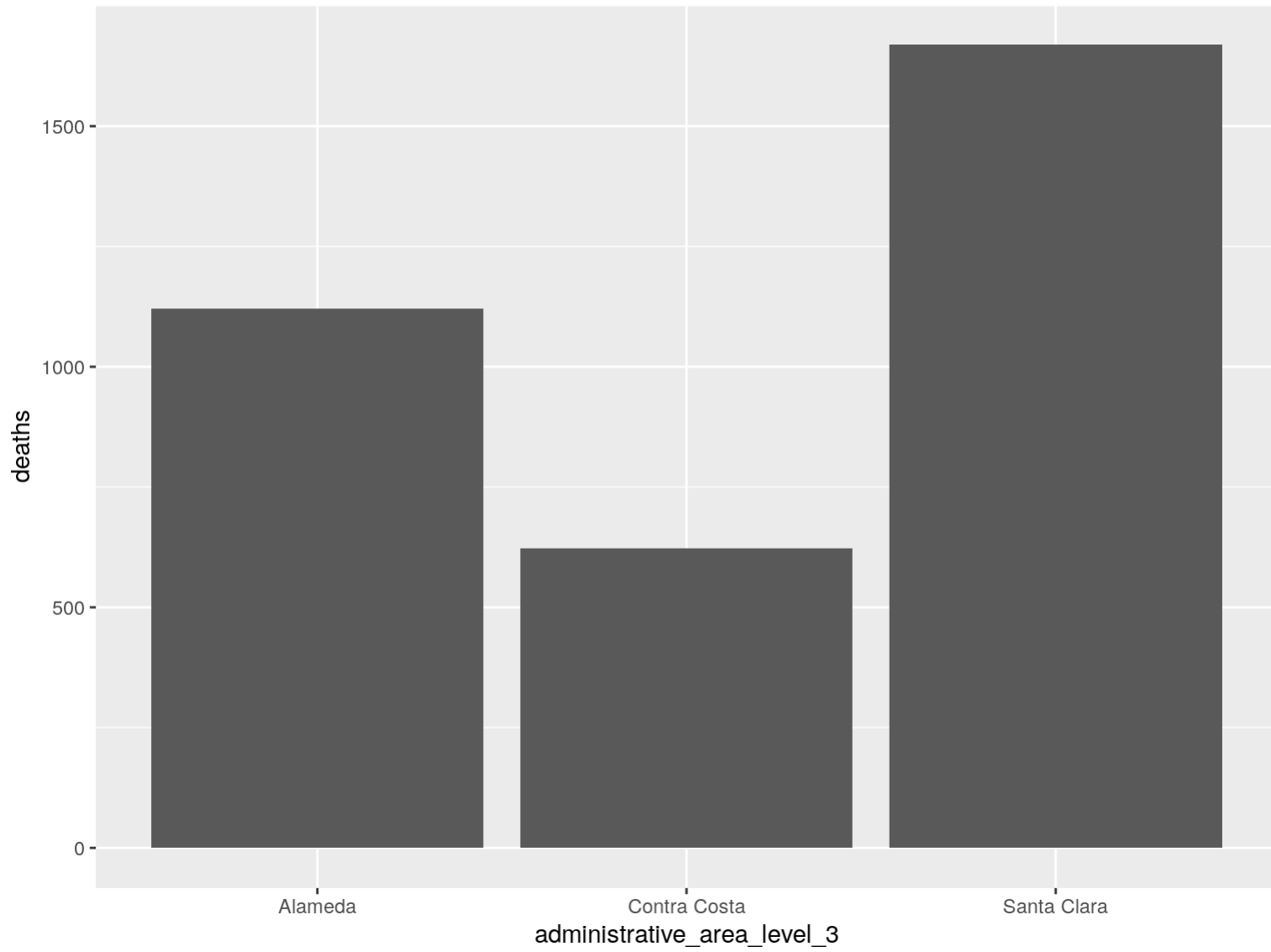
```
x_USA_county <- covid19("USA", level = 3)
```

```
## Warning in id(y$fips, iso = "USA", ds = "nytimes_git", level = level): missing  
## id: 2997, 2998
```

```
##  
## World Bank Open Data (2018), https://data.worldbank.org  
##  
## Hale Thomas, Sam Webster, Anna Petherick, Toby Phillips, and Beatriz  
## Kira (2020). Oxford COVID-19 Government Response Tracker, Blavatnik  
## School of Government.  
##  
## Johns Hopkins Center for Systems Science and Engineering (2020),  
## https://github.com  
##  
## The New York Times (2020), https://github.com  
##  
## Guidotti, E., Ardia, D., (2020), "COVID-19 Data Hub", Journal of Open  
## Source Software 5(51):2376, doi: 10.21105/joss.02376.  
##  
## To see these entries in BibTeX format, use 'print(<citation>,  
## bibtex=TRUE)', 'toBibtex(.)', or set  
## 'options(citation.bibtex.max=999)'.  
##  
## To hide the data sources use 'verbose = FALSE'.
```

```
x_USA_county %>% select(date, administrative_area_level_2, administrative_area_level_3, deaths, vaccines)
%>%
  filter(date == "2021-02-15") %>%
  filter(administrative_area_level_2 %in% c("California")) %>%
  filter(administrative_area_level_3 %in% c("Alameda", "Contra Costa", "Santa Clara")) %>%
  ggplot(aes(x = administrative_area_level_3, y = deaths)) +
  geom_bar(stat="identity")
```

```
## Adding missing grouping variables: `id`
```

Getting into the role

- Checking the [Issues](#) everyday.
- Trying to continue the development new documentation and examples of the use of the data.
- Fully understanding the philosophy of the creators of the project.
- Recruiting others to help out. Maybe just for motivation. Please **star** the COVID19 Data Hub Project on Github.
- Putting in some hours to complete some of the Open Issues.

Please reach out if you have any suggestions.

- On the Project Github Issues page.
- Or by email. eric.suess@csueastbay.edu