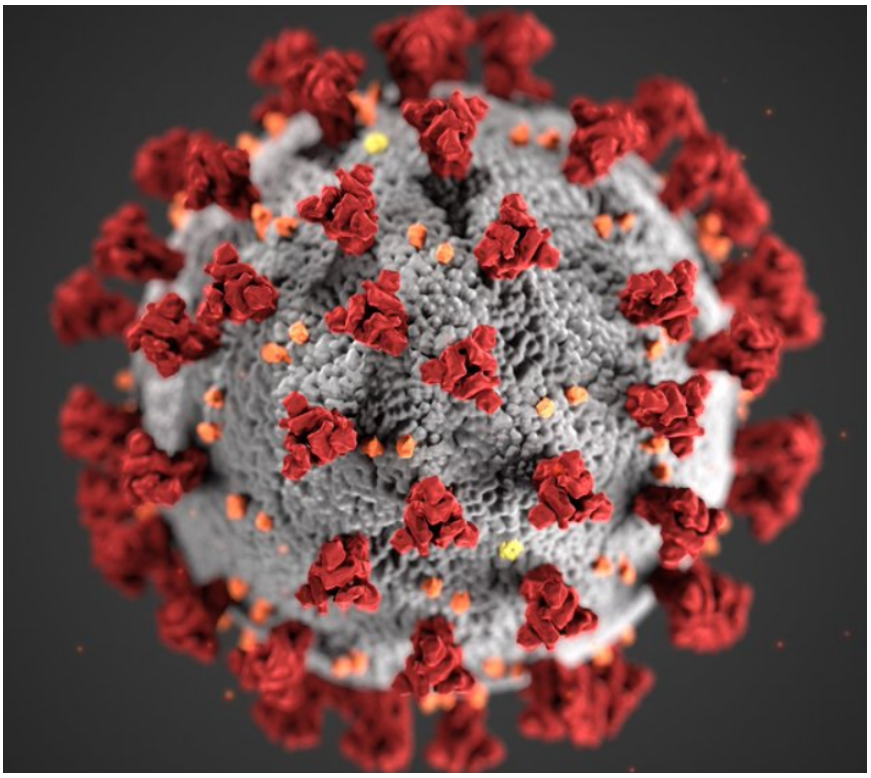


R ile Covid19 Analizi

Muhammed Fatih TÜZEN

15 01 2022

İÇİNDEKİLER

1 Giriş

Veri analizi kapsamında uygulama yapmak için son yılların gündemini oldukça meşgul eden COVID-19 verileri kullanılacaktır. Bu kapsamda hazırlanmış olan **coronavirus** paketi kullanılarak güncel vaka sayıları ve aşılama verilerine erişmek mümkündür.

Bu paket, 2019 Novel Coronavirus COVID-19 (2019-nCoV) salgını ve ülkelere göre aşılama çabalarına ilişkin düzenli bir format veri kümesi sağlar. Ham veriler, Johns Hopkins Üniversitesi Sistem Bilimi ve Mühendisliği Merkezi (JHU CCSE) Coronavirüs deposundan alınmaktadır. Pakete ilişkin detaylara https://github.com/RamiKrispin/coronavirus adresinden ulaşılabilir.

Paketi github sayfasından yükleyebilirsiniz:

```
devtools::install_github("RamiKrispin/coronavirus")
```

2 Veri ile Tanışma

```
# coronavirus verisi
library(coronavirus)
data(coronavirus)

head(coronavirus)
```

```
##         date province country     lat      long      type cases   uid iso2 iso3
## 1 2020-01-22  Alberta  Canada 53.9333 -116.5765 confirmed     0 12401   CA  CAN
## 2 2020-01-23  Alberta  Canada 53.9333 -116.5765 confirmed     0 12401   CA  CAN
## 3 2020-01-24  Alberta  Canada 53.9333 -116.5765 confirmed     0 12401   CA  CAN
## 4 2020-01-25  Alberta  Canada 53.9333 -116.5765 confirmed     0 12401   CA  CAN
## 5 2020-01-26  Alberta  Canada 53.9333 -116.5765 confirmed     0 12401   CA  CAN
## 6 2020-01-27  Alberta  Canada 53.9333 -116.5765 confirmed     0 12401   CA  CAN
##   code3    combined_key population continent_name continent_code
## 1   124 Alberta, Canada    4413146  North America             NA
## 2   124 Alberta, Canada    4413146  North America             NA
## 3   124 Alberta, Canada    4413146  North America             NA
## 4   124 Alberta, Canada    4413146  North America             NA
## 5   124 Alberta, Canada    4413146  North America             NA
## 6   124 Alberta, Canada    4413146  North America             NA
```

```
nrow(coronavirus)
```

```
## [1] 585750
```

```
str(coronavirus)
```

```
## 'data.frame':    585750 obs. of  15 variables:
##  $ date          : Date, format: "2020-01-22" "2020-01-23" ...
##  $ province      : chr  "Alberta" "Alberta" "Alberta" "Alberta" ...
##  $ country       : chr  "Canada" "Canada" "Canada" "Canada" ...
##  $ lat           : num  53.9 53.9 53.9 53.9 53.9 ...
##  $ long          : num  -117 -117 -117 -117 -117 ...
##  $ type          : chr  "confirmed" "confirmed" "confirmed" "confirmed" ...
##  $ cases         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ uid           : num  12401 12401 12401 12401 12401 ...
##  $ iso2          : chr  "CA" "CA" "CA" "CA" ...
##  $ iso3          : chr  "CAN" "CAN" "CAN" "CAN" ...
##  $ code3         : num  124 124 124 124 124 124 124 124 124 124 ...
##  $ combined_key  : chr  "Alberta, Canada" "Alberta, Canada" "Alberta, Canada" "Albert
##  $ population    : num  4413146 4413146 4413146 4413146 4413146 ...
##  $ continent_name: chr  "North America" "North America" "North America" "North Americ
##  $ continent_code: chr  "NA" "NA" "NA" "NA" ...
```

```
summary(coronavirus)
```

```
##       date               province           country               lat        
##  Min.   :2020-01-22   Length:585750      Length:585750      Min.   :-51.796  
##  1st Qu.:2020-07-17   Class :character   Class :character   1st Qu.:  4.562  
##  Median :2021-01-10   Mode  :character   Mode  :character   Median : 21.008  
##  Mean   :2021-01-10                                         Mean   : 19.697  
##  3rd Qu.:2021-07-07                                         3rd Qu.: 40.069  
##  Max.   :2021-12-31                                         Max.   : 71.707  
##                                                             NA's   :3550     
##       long              type                cases                uid          
##  Min.   :-178.12   Length:585750      Min.   :-30974748   Min.   :    4.0  
##  1st Qu.: -15.31   Class :character   1st Qu.:        0   1st Qu.:  267.5  
##  Median :  21.75   Mode  :character   Median :        0   Median :  531.0  
##  Mean   :  23.50                      Mean   :      501   Mean   : 2840.8  
##  3rd Qu.:  88.09                      3rd Qu.:       31   3rd Qu.:  804.0  
##  Max.   : 178.06                      Max.   :  1123456   Max.   :15699.0  
##  NA's   :3550                                             NA's   :3550     
##      iso2               iso3               code3       combined_key      
##  Length:585750      Length:585750      Min.   :  4    Length:585750      
##  Class :character   Class :character   1st Qu.:156    Class :character   
##  Mode  :character   Mode  :character   Median :336    Mode  :character   
##                                        Mean   :375                       
##                                        3rd Qu.:598
```

```
##                                      Max.   :894
##                                      NA's   :9940
##    population         continent_name     continent_code
##  Min.   :       809   Length:585750      Length:585750
##  1st Qu.:    771612   Class :character   Class :character
##  Median :   6880000   Mode  :character   Mode  :character
##  Mean   :  28788798
##  3rd Qu.:  29136808
##  Max.   :1380004385
##  NA's   :14910
```

```
library(dplyr)
glimpse(coronavirus)
```

```
## Rows: 585,750
## Columns: 15
## $ date           <date> 2020-01-22, 2020-01-23, 2020-01-24, 2020-01-25, 2020-0~
## $ province       <chr> "Alberta", "Alberta", "Alberta", "Alberta", "Alberta", ~
## $ country        <chr> "Canada", "Canada", "Canada", "Canada", "Canada", "Cana~
## $ lat            <dbl> 53.9333, 53.9333, 53.9333, 53.9333, 53.9333, 53.9333, 5~
## $ long           <dbl> -116.5765, -116.5765, -116.5765, -116.5765, -116.5765, ~
## $ type           <chr> "confirmed", "confirmed", "confirmed", "confirmed", "co~
## $ cases          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ uid            <dbl> 12401, 12401, 12401, 12401, 12401, 12401, 12401, 12401,~
## $ iso2           <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "~
## $ iso3           <chr> "CAN", "CAN", "CAN", "CAN", "CAN", "CAN", "CAN", "CAN",~
## $ code3          <dbl> 124, 124, 124, 124, 124, 124, 124, 124, 124, 124, 124, ~
## $ combined_key   <chr> "Alberta, Canada", "Alberta, Canada", "Alberta, Canada"~
## $ population     <dbl> 4413146, 4413146, 4413146, 4413146, 4413146, 4413146, 4~
## $ continent_name <chr> "North America", "North America", "North America", "Nor~
## $ continent_code <chr> "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "~
```

3 Veri Analizi

3.1 Zamana göre vaka sayıları

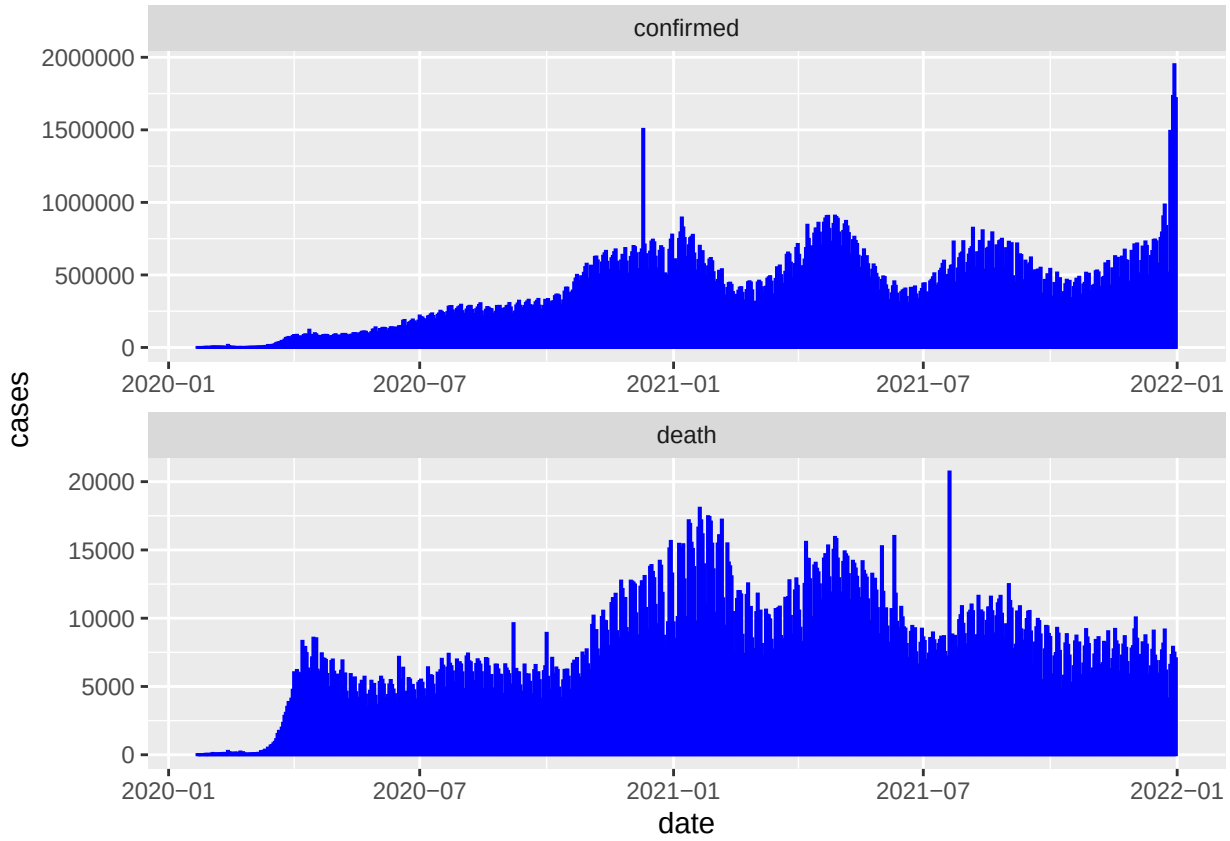
```
# Zamana göre vaka sayıları

coronavirus %>%
  filter(type == "confirmed") %>%
  group_by(date) %>%
```

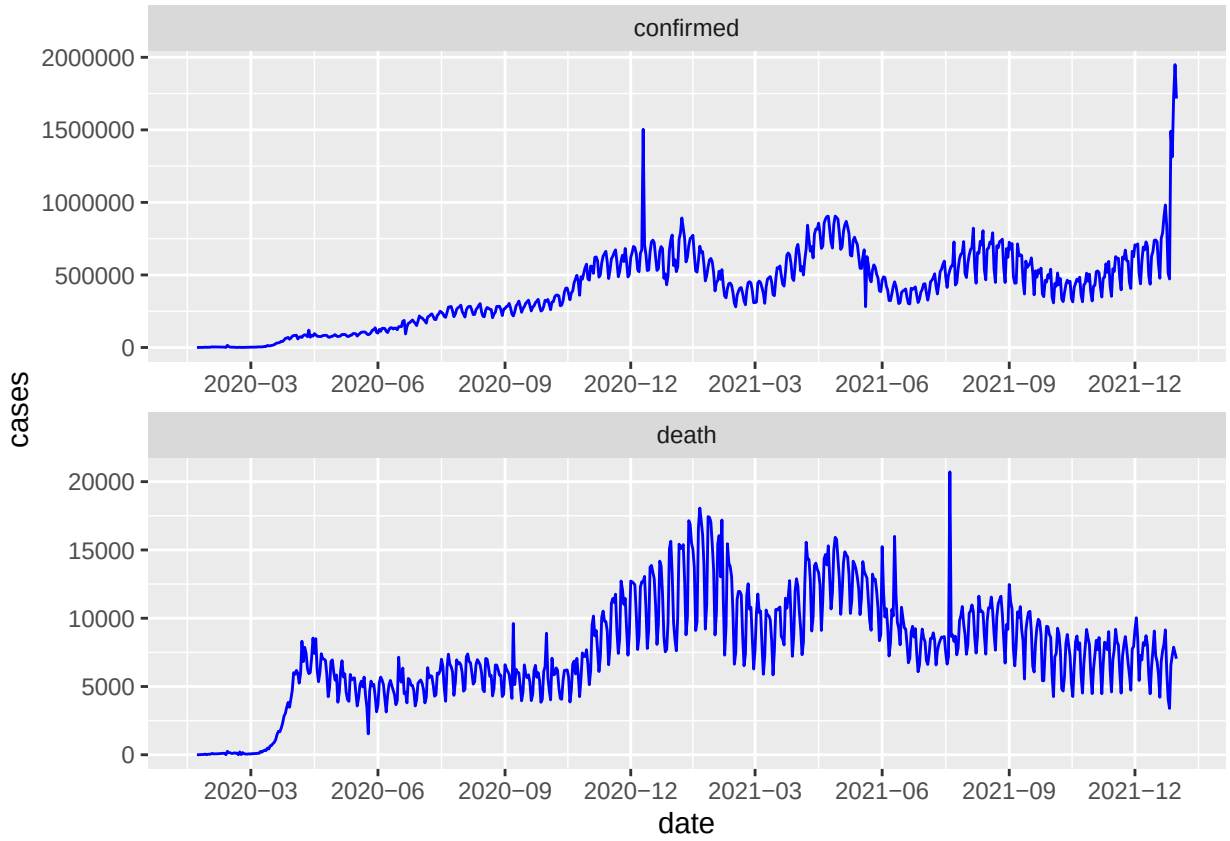
```
  summarise(cases = sum(cases)) %>%
  arrange(desc(date))
```

```
## # A tibble: 710 x 2
##    date         cases
##    <date>       <int>
##  1 2021-12-31 1715785
##  2 2021-12-30 1949468
##  3 2021-12-29 1730636
##  4 2021-12-28 1313713
##  5 2021-12-27 1490304
##  6 2021-12-26  472215
##  7 2021-12-25  512541
##  8 2021-12-24  834103
##  9 2021-12-23  982453
## 10 2021-12-22  900818
## # ... with 700 more rows
```

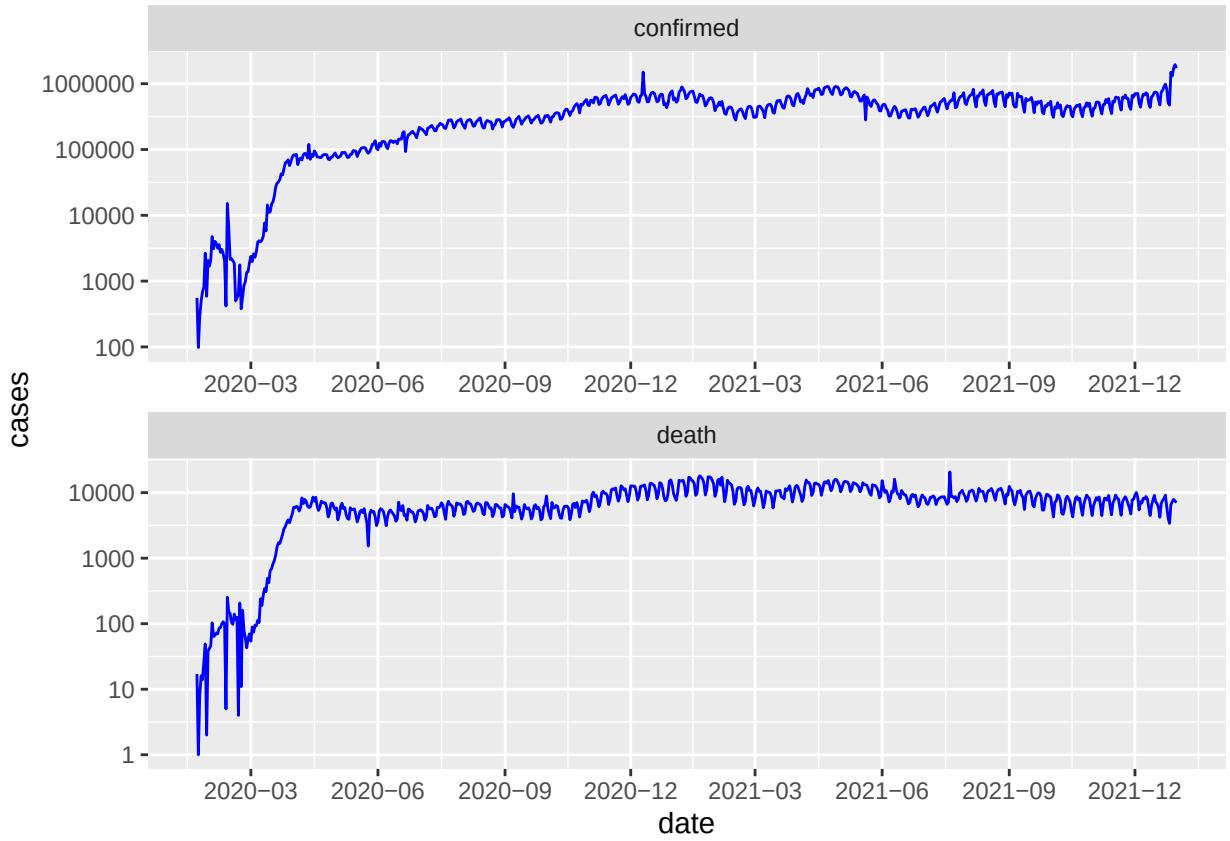
```
library(ggplot2)
coronavirus %>%
  filter(type != "recovery") %>%
  group_by(date,type) %>%
  summarise(cases = sum(cases)) %>%
  ggplot(aes(x=date,y=cases)) +
  geom_col(col="blue") +
  facet_wrap(~type,scales = "free",nrow = 2)
```




```
coronavirus %>%
  filter(type != "recovery") %>%
  group_by(date,type) %>%
  summarise(cases = sum(cases)) %>%
  ggplot(aes(x=date,y=cases)) +
  geom_line(col="blue") +
  facet_wrap(~type,scales = "free",nrow = 2)+
  scale_x_date(date_breaks = "3 month", date_labels = "%Y-%m")
```




```
coronavirus %>%
  filter(type != "recovery") %>%
  group_by(date,type) %>%
  summarise(cases = sum(cases)) %>%
  ggplot(aes(x=date,y=cases)) +
  geom_line(col="blue") +
  scale_y_log10() +
  facet_wrap(~type,scales = "free",nrow = 2) +
  scale_x_date(date_breaks = "3 month", date_labels = "%Y-%m")
```



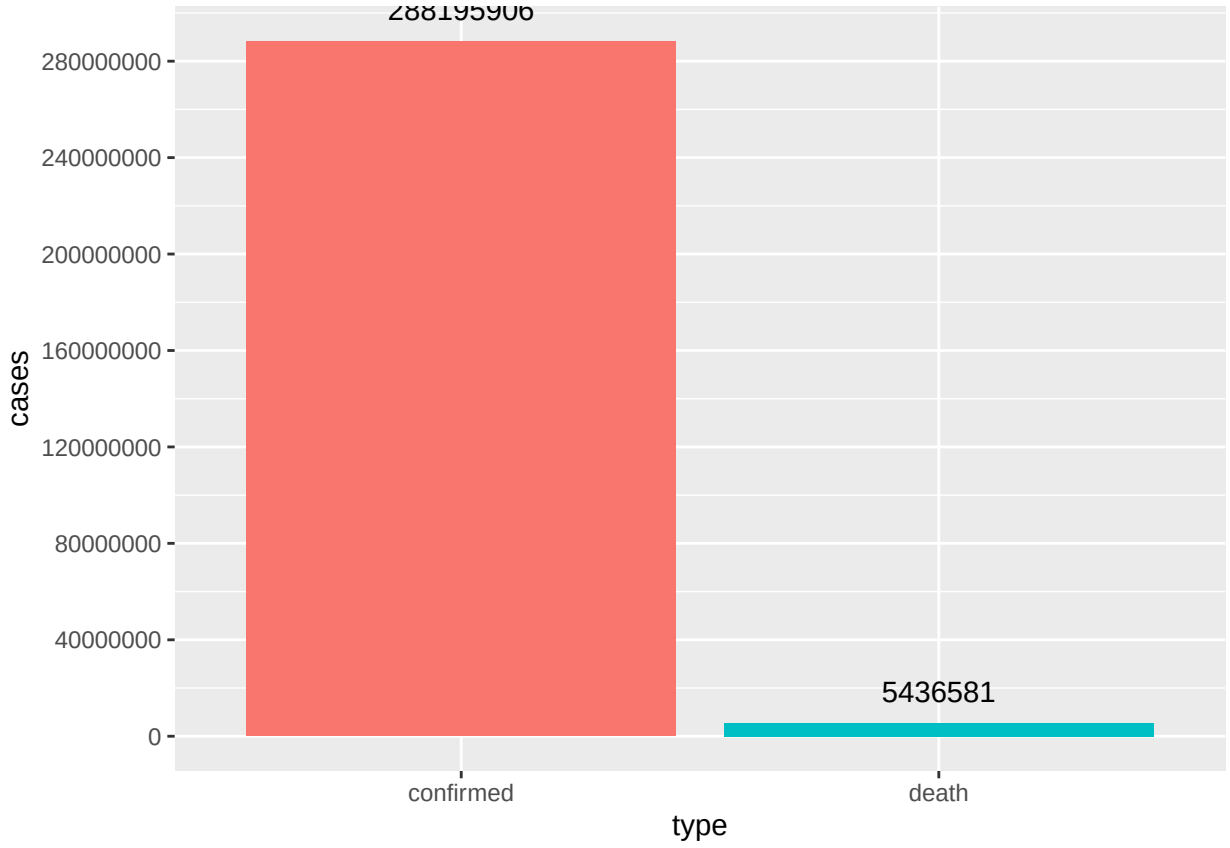
3.2 Vaka tipine göre vaka sayıları

```
# Vaka tipine göre vaka sayıları
coronavirus %>%
  filter(type != "recovery") %>%
  group_by(type) %>%
  summarise(cases = sum(cases))
```

```
## # A tibble: 2 x 2
##   type          cases
##   <chr>         <int>
## 1 confirmed 288195906
## 2 death       5436581
```

```
coronavirus %>%
  filter(type != "recovery") %>%
  group_by(type) %>%
  summarise(cases = sum(cases)) %>%
  ggplot(aes(x=type,y=cases,fill=type)) +
```

```
  geom_col()+
  theme(legend.position = "none") +
  scale_y_continuous(breaks = seq(0,3e+8,by=4e+7),labels = function(x) format(x, scienti
  geom_text(aes(label = cases), hjust=0.5,vjust=-1)
```



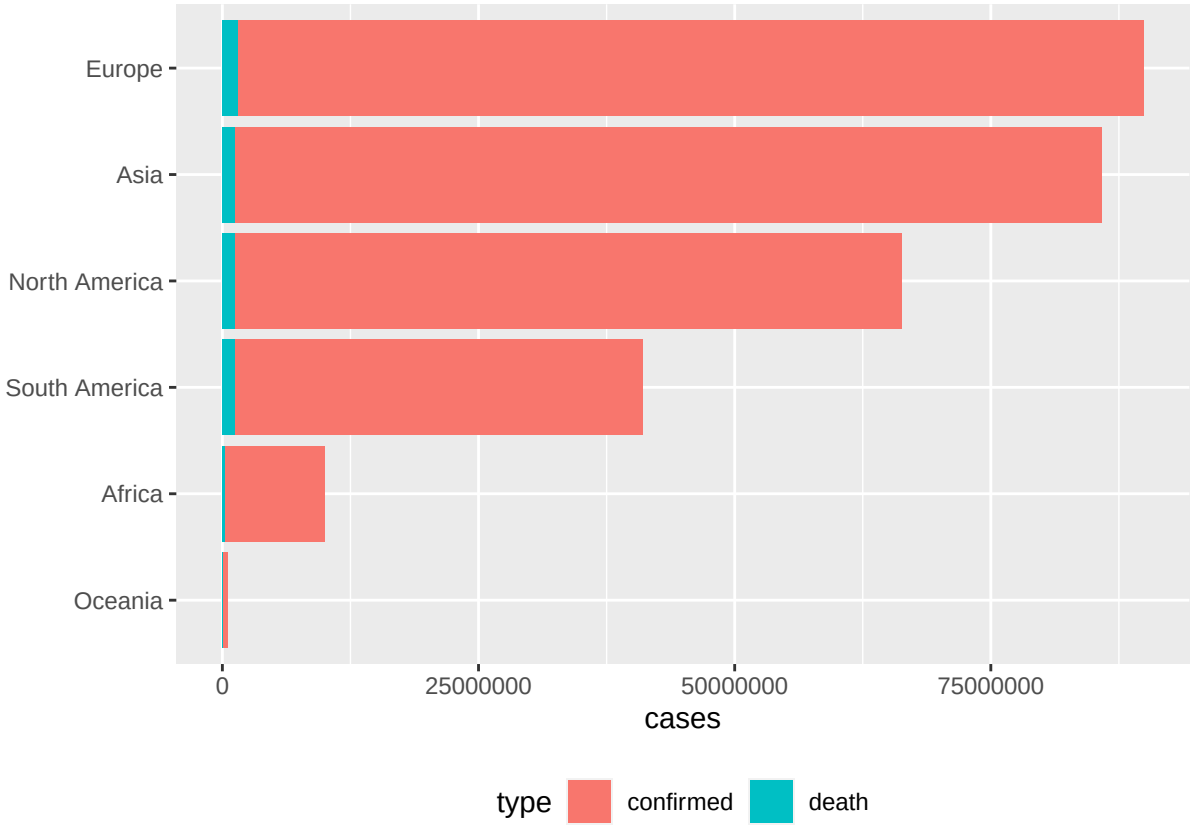
3.3 Kıtalara göre ölüm ve vaka sayıları

```
# kıtalara göre ölüm ve vaka sayıları
library(tidyr)
coronavirus %>%
  filter(type != "recovery") %>%
  group_by(type, continent_name) %>%
  summarise(cases = sum(cases)) %>%
  spread(key = type, value = cases) %>%
  mutate(death_rate = death / confirmed) %>%
  filter(!is.na(continent_name)) %>%
  arrange(-death_rate)
```

```
## # A tibble: 6 x 4
```

```
##   continent_name confirmed   death death_rate
##   <chr>              <int>   <int>      <dbl>
## 1 South America   39815458 1192277    0.0299
## 2 Africa           9809694  228501    0.0233
## 3 North America   65038920 1224062    0.0188
## 4 Europe          88357000 1527831    0.0173
## 5 Asia            84606964 1259383    0.0149
## 6 Oceania           566270    4512    0.00797
```

```
coronavirus %>%
  filter(type != "recovery") %>%
  group_by(type, continent_name) %>%
  summarise(cases = sum(cases)) %>%
  filter(!is.na(continent_name)) %>%
  ggplot(aes(x=reorder(continent_name,cases),y=cases,fill=type)) +
  geom_col()+
  theme(legend.position = "bottom") +
  labs(x="") +
  coord_flip()
```

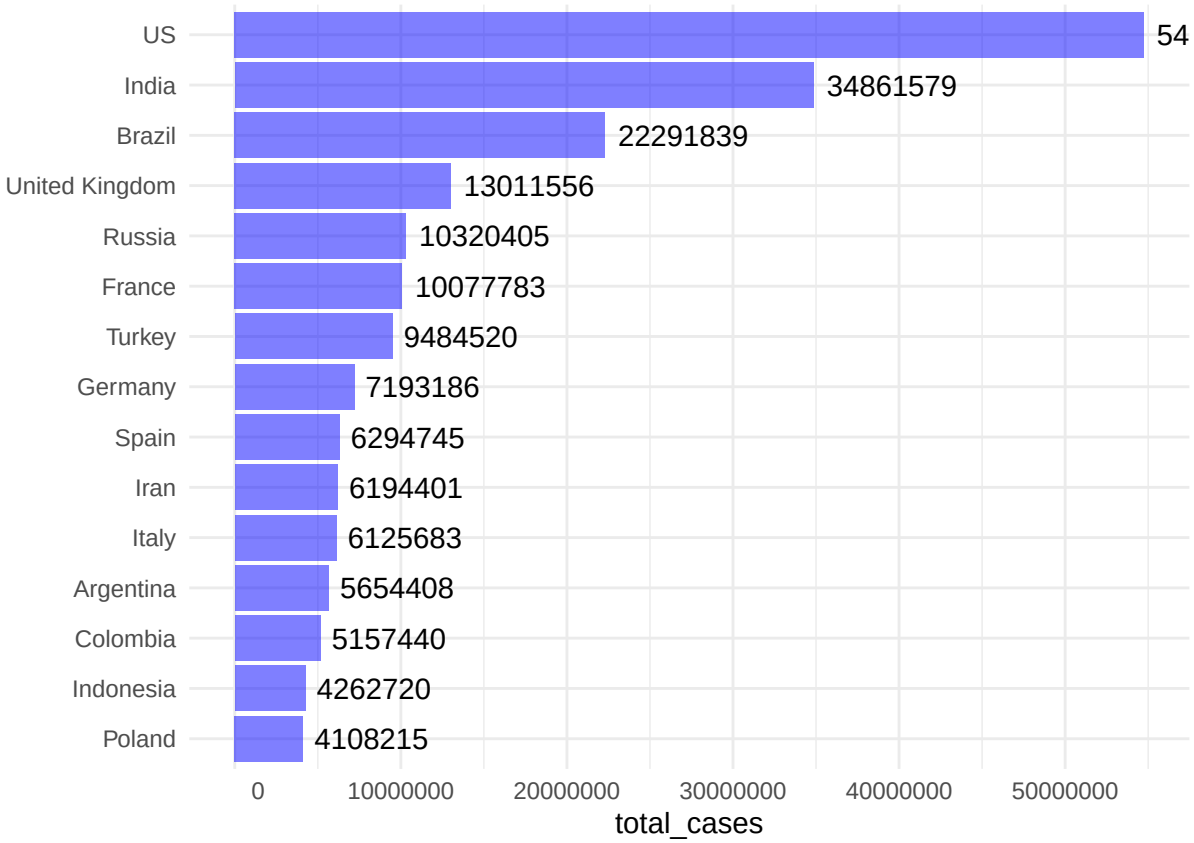


3.4 Ülkelere göre onaylanmış vaka sayıları

```
# Ülkelere göre onaylanmış vaka sayıları
coronavirus %>%
  filter(type == "confirmed") %>%
  group_by(country) %>%
  summarise(total_cases = sum(cases)) %>%
  arrange(-total_cases)
```

```
## # A tibble: 196 x 2
##    country        total_cases
##    <chr>                <int>
##  1 US                54743982
##  2 India             34861579
##  3 Brazil            22291839
##  4 United Kingdom    13011556
##  5 Russia            10320405
##  6 France            10077783
##  7 Turkey             9484520
##  8 Germany            7193186
##  9 Spain              6294745
## 10 Iran               6194401
## # ... with 186 more rows
```

```
coronavirus %>%
  filter(type == "confirmed") %>%
  group_by(country) %>%
  summarise(total_cases = sum(cases)) %>%
  arrange(-total_cases) %>%
  head(15) %>%
  ggplot(aes(x=reorder(country,total_cases),y=total_cases)) +
  geom_col(fill="blue",alpha=0.5)+
  theme(legend.position = "bottom") +
  labs(x="") +
  scale_y_continuous(breaks = seq(0,6e+7,by=1e+7),labels = function(x) format(x, scienti
  coord_flip()+
  geom_text(aes(label = total_cases),hjust=-0.1,vjust=0.5)+
  theme_minimal()
```



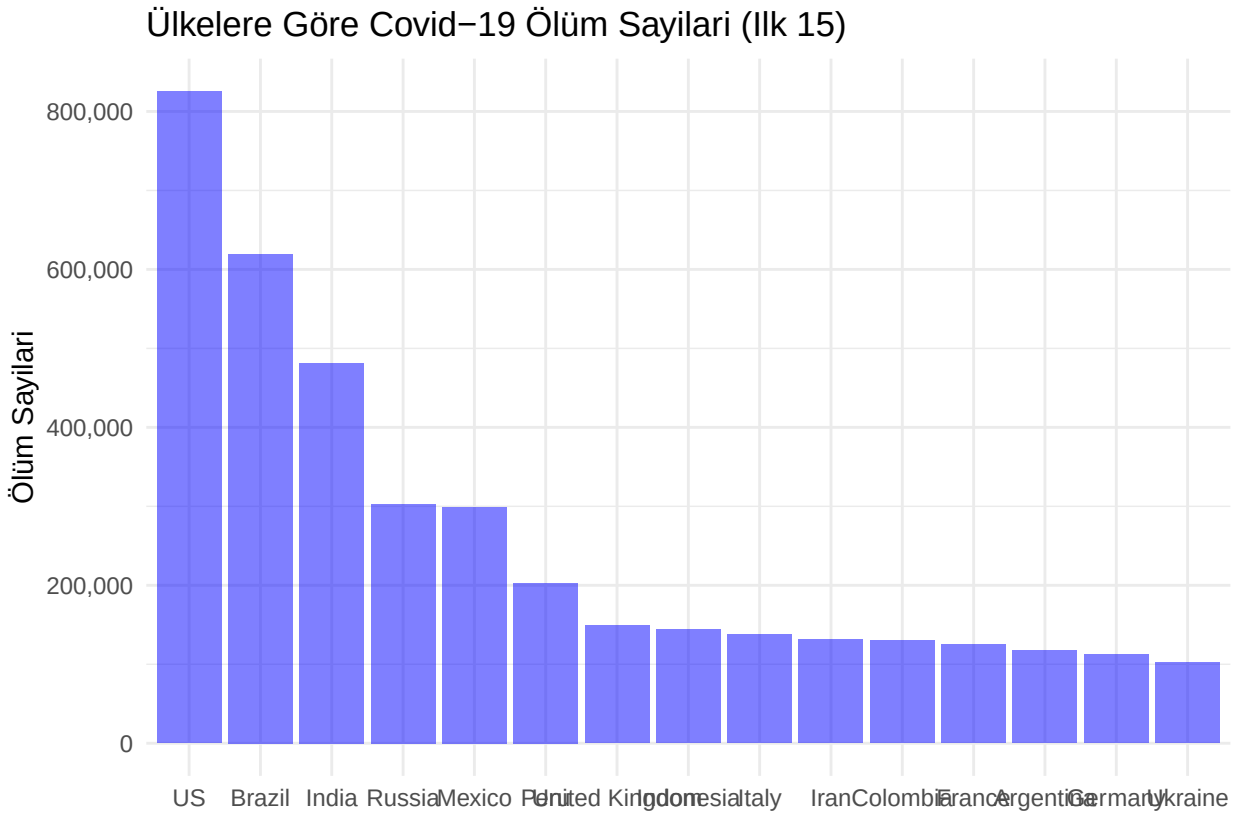
3.5 Ülkelere göre ölüm sayıları

```
# Ülkelere göre ölüm sayıları
coronavirus %>%
  filter(type == "death") %>%
  group_by(country) %>%
  summarise(total_cases = sum(cases)) %>%
  arrange(-total_cases)
```

```
## # A tibble: 196 x 2
##    country        total_cases
##    <chr>                <int>
##  1 US                  825536
##  2 Brazil              619334
##  3 India               481486
##  4 Russia              302671
##  5 Mexico              299428
##  6 Peru                202653
##  7 United Kingdom      149096
```

```
##  8 Indonesia           144094
##  9 Italy               137402
## 10 Iran                131606
## # ... with 186 more rows
```

```
coronavirus %>%
  filter(type == "death") %>%
  group_by(country) %>%
  summarise(total_deaths = sum(cases)) %>%
  arrange(-total_deaths) %>%
  head(15) %>%
  ggplot(aes(x = reorder(country, -total_deaths), y = total_deaths)) +
  geom_col(fill = "blue", alpha = 0.5) +
  scale_y_continuous(labels = scales::comma) +
  theme(legend.position = "bottom") +
  labs( title = "Ülkelere Göre Covid-19 Ölüm Sayıları (İlk 15)",
        x="",
        y="Ölüm Sayıları") +
  theme_minimal()
```



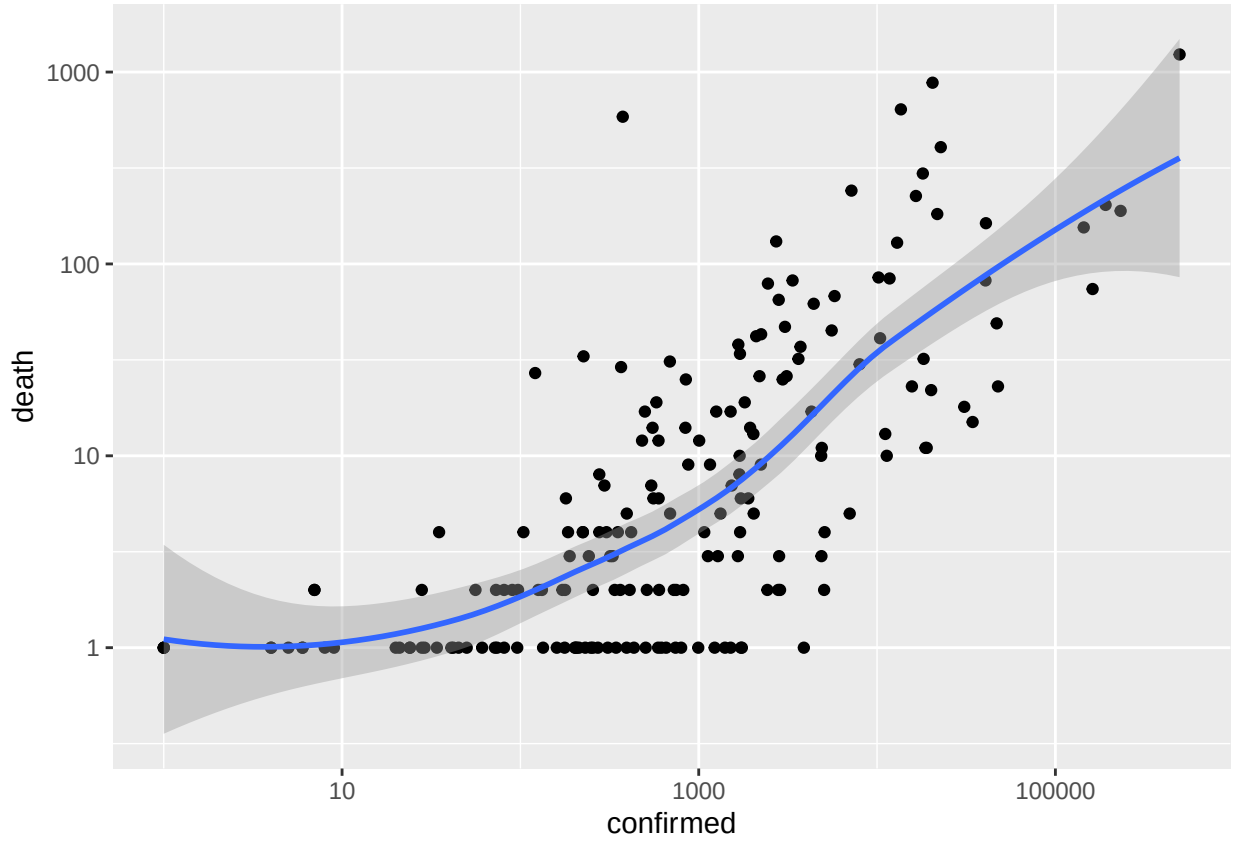
3.6 Son 24 saatteki vaka ve ölüm sayıları

```
# son 24 saatteki vaka ve ölüm sayıları
coronavirus %>%
  filter(cases!=0,type != "recovery") %>%
  select(date,country, type, cases) %>%
  group_by(country, type) %>%
  filter(date==max(date)) %>%
  summarise(total_cases=sum(cases)) %>%
  spread(key = type,value = total_cases) %>%
  arrange(-confirmed) %>%
  head(20)
```

```
## # A tibble: 20 x 3
## # Groups:   country [20]
##    country        confirmed death
##    <chr>              <int> <int>
##  1 US                497151  1235
##  2 France            232200   189
##  3 United Kingdom    191197   203
##  4 Spain             161688    74
##  5 Italy             144255   155
##  6 Argentina          47663    23
##  7 Canada             46868    49
##  8 Turkey             40786   163
##  9 Greece             40560    82
## 10 Australia          34353    15
## 11 Portugal           30829    18
## 12 India              22775   406
## 13 Germany            21764   182
## 14 Russia             20482   880
## 15 Ireland            20110    22
## 16 Switzerland        18987    11
## 17 Denmark            18696    11
## 18 Netherlands        18244    32
## 19 Mexico             18061   296
## 20 Vietnam            16515   226
```

```
coronavirus %>%
  filter(cases!=0,type != "recovery") %>%
  select(date,country, type, cases) %>%
  group_by(country, type) %>%
  filter(date==max(date)) %>%
```

```
  summarise(total_cases=sum(cases)) %>%
  spread(key = type,value = total_cases) %>%
  ggplot(aes(x=confirmed,y=death))+
  geom_point() +
  scale_x_log10()+
  scale_y_log10() +
  geom_smooth()
```



3.7 Türkiye'de Covid-19

```
# Türkiye'de Covid-19

coronavirus %>%
  filter(type != "recovery",country=="Turkey",cases!=0) %>%
  select(date,type,cases) %>%
  summary()
```

```
##       date                type                cases
```

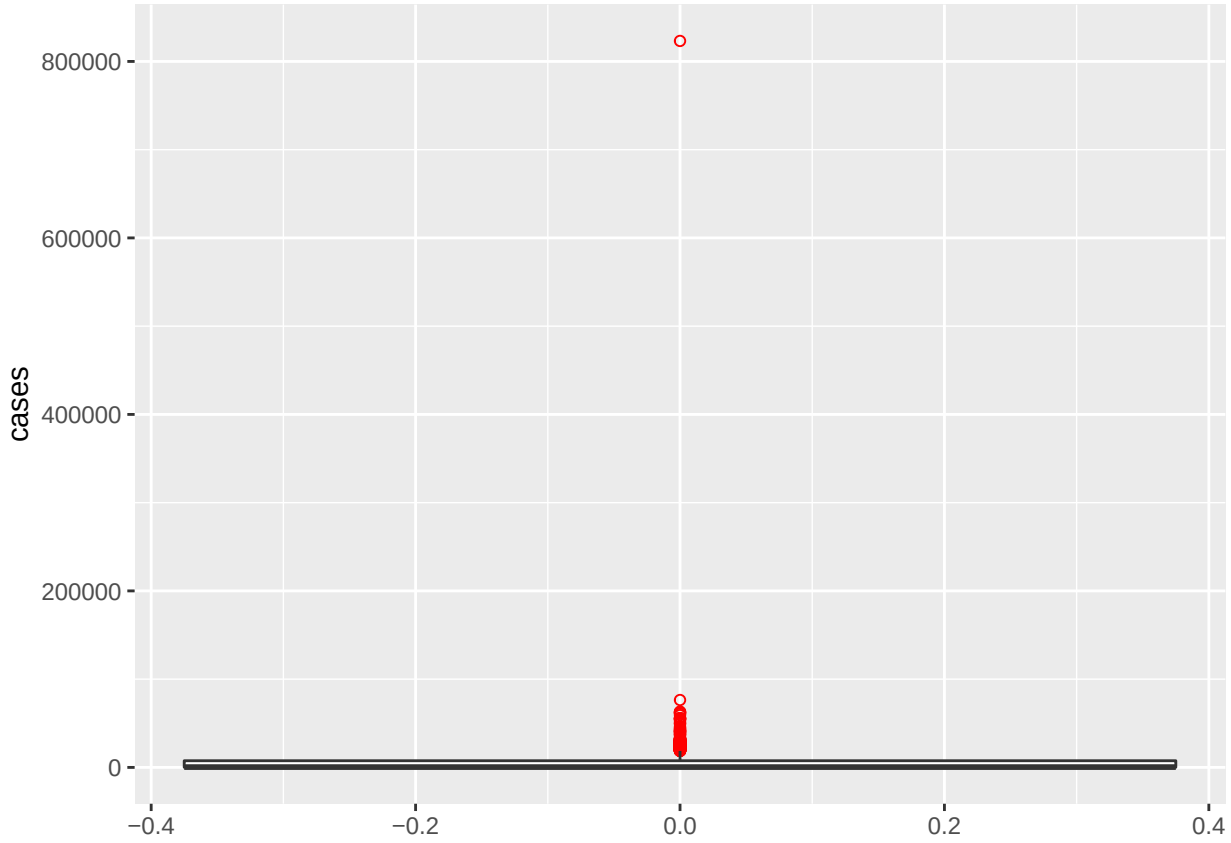
```
##  Min.   :2020-03-11   Length:1305        Min.   :     1
##  1st Qu.:2020-08-25   Class :character   1st Qu.:    95
##  Median :2021-02-04   Mode  :character   Median :   346
##  Mean   :2021-02-04                      Mean   :  7331
##  3rd Qu.:2021-07-17                      3rd Qu.:  7550
##  Max.   :2021-12-31                      Max.   :823225
```

```
# Türkiye'de 823225 vaka olmadı. Veri girişi hatası olabilir.
# bunu kaldırabiliriz.

# vaka sayısına göre sıralayalım
coronavirus %>%
  filter(type != "recovery",country=="Turkey",cases!=0) %>%
  select(date,type,cases) %>%
  arrange(-cases) %>%
  head(20)
```

```
##          date      type  cases
## 1  2020-12-10 confirmed 823225
## 2  2021-12-30 confirmed  76365
## 3  2021-04-16 confirmed  63082
## 4  2021-04-14 confirmed  62797
## 5  2021-04-17 confirmed  62606
## 6  2021-04-21 confirmed  61967
## 7  2021-04-15 confirmed  61400
## 8  2021-04-20 confirmed  61028
## 9  2021-04-13 confirmed  59187
## 10 2021-04-08 confirmed  55941
## 11 2021-04-18 confirmed  55802
## 12 2021-04-09 confirmed  55791
## 13 2021-04-19 confirmed  55149
## 14 2021-04-22 confirmed  54791
## 15 2021-04-07 confirmed  54740
## 16 2021-04-12 confirmed  54562
## 17 2021-04-10 confirmed  52676
## 18 2021-04-11 confirmed  50678
## 19 2021-04-06 confirmed  49584
## 20 2021-04-23 confirmed  49438
```

```
# boxplot çizdirelim
coronavirus %>%
  filter(type != "recovery",country=="Turkey",cases!=0) %>%
  select(date,type,cases) %>%
  ggplot(aes(y=cases)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 1)
```



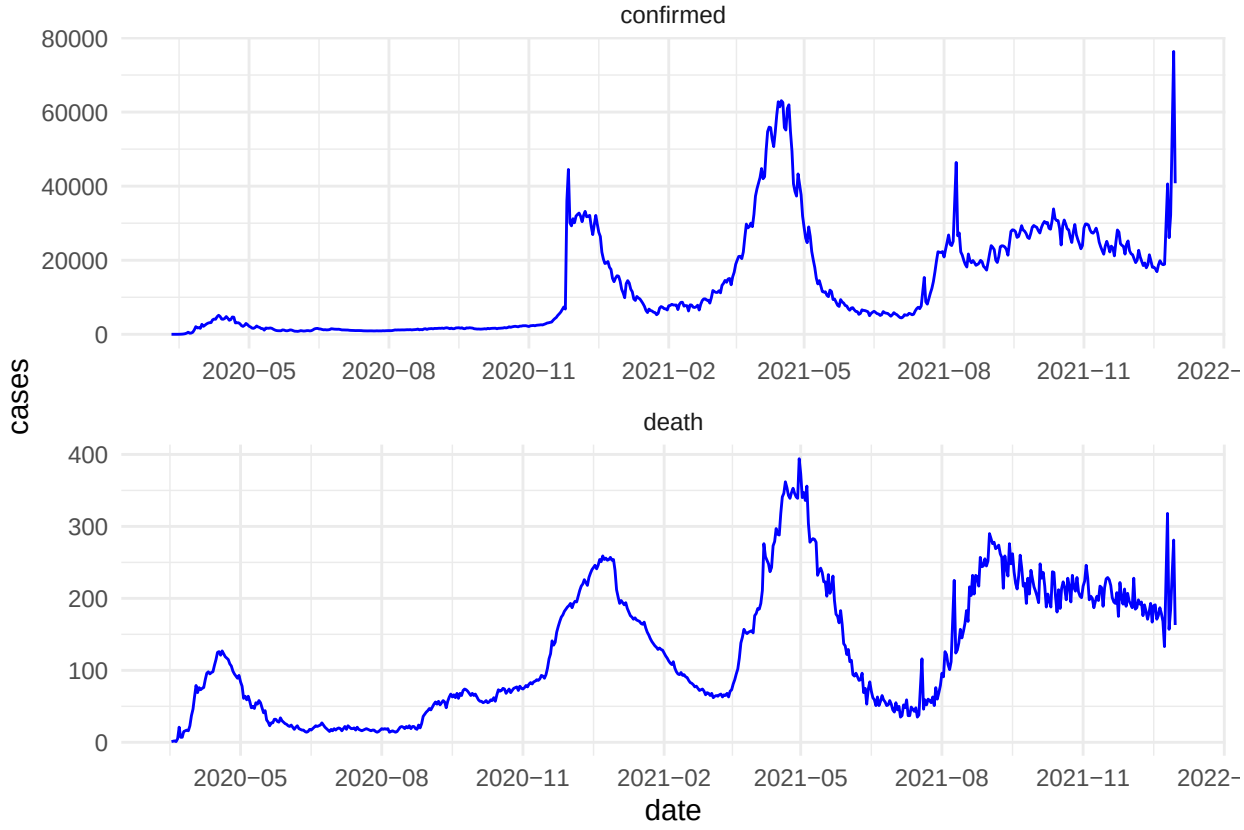

```
is_outlier <- function(x) {
  return(x < quantile(x, 0.0025) | x > quantile(x, 0.9975))
}

coronavirus %>%
  filter(type != "recovery",country=="Turkey",cases!=0) %>%
  select(date,type,cases) %>%
  group_by(type) %>%
  mutate(outlier=ifelse(is_outlier(cases), TRUE, FALSE)) %>%
  ungroup() %>%
  arrange(-cases) %>%
  head(10)
```

```
## # A tibble: 10 x 4
##    date       type       cases outlier
##    <date>     <chr>      <int> <lgl>
##  1 2020-12-10 confirmed 823225 TRUE
##  2 2021-12-30 confirmed  76365 TRUE
##  3 2021-04-16 confirmed  63082 FALSE
##  4 2021-04-14 confirmed  62797 FALSE
##  5 2021-04-17 confirmed  62606 FALSE
```

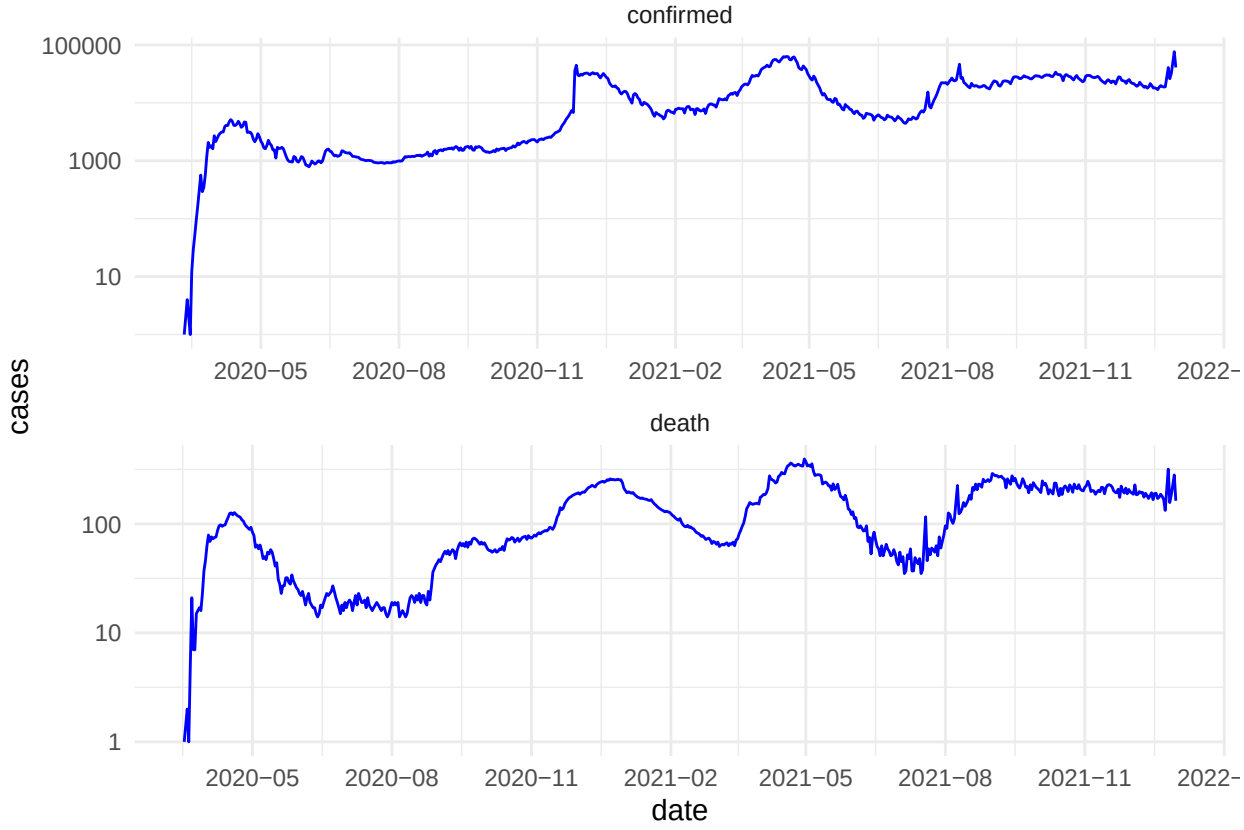
```
##  6 2021-04-21 confirmed  61967 FALSE
##  7 2021-04-15 confirmed  61400 FALSE
##  8 2021-04-20 confirmed  61028 FALSE
##  9 2021-04-13 confirmed  59187 FALSE
## 10 2021-04-08 confirmed  55941 FALSE
```

```
coronavirus %>%
  filter(type != "recovery",country=="Turkey",between(cases,1,100000)) %>%
  ggplot(aes(x=date,y=cases)) +
  geom_line(col="blue") +
  facet_wrap(~type,scales = "free",nrow = 2) +
  scale_x_date(date_breaks = "3 month", date_labels = "%Y-%m") +
  theme_minimal()
```




```
coronavirus %>%
  filter(type != "recovery",country=="Turkey",between(cases,1,100000)) %>%
  ggplot(aes(x=date,y=cases)) +
  geom_line(col="blue") +
  scale_y_log10() +
  facet_wrap(~type,scales = "free",nrow = 2) +
```

```
  scale_x_date(date_breaks = "3 month", date_labels = "%Y-%m") +
  theme_minimal()
```




```
# büyüme hesaplayalım
coronavirus %>%
  filter(type == "confirmed",country=="Turkey",between(cases,1,100000)) %>%
  select(date,cases) %>%
  mutate(growth=cases/lag(cases,1)*100-100) %>%
  arrange(desc(date)) %>%
  head(10)
```

```
##          date cases     growth
## 1  2021-12-31 40786 -46.590716
## 2  2021-12-30 76365 137.335281
## 3  2021-12-28 32176  23.284417
## 4  2021-12-27 26099 -35.729413
## 5  2021-12-26 40608 114.743522
## 6  2021-12-24 18910   0.740504
## 7  2021-12-23 18771  -1.696779
## 8  2021-12-22 19095  -3.847122
```

```
## 9  2021-12-21 19859   5.846925
## 10 2021-12-20 18762  10.952099
```

```
coronavirus %>%
  filter(type == "confirmed",country=="Turkey",between(cases,1,100000)) %>%
  select(date,cases) %>%
  mutate(growth=cases/lag(cases,1)*100-100) %>%
  filter(lubridate::year(date)>2020) %>%
  ggplot(aes(x=date,y=growth)) +
  geom_line(size=1,color="blue")
```