Case Study

Rifa

12/6/2020

```
library(data.table)
library(magrittr)
library(ggplot2)
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

Load all the tables

```
tables = list.files("data/", full.names = T)
tables
```

```
##  [1] "data//Case.csv"          "data//PatientInfo.csv"
##  [3] "data//Policy.csv"        "data//Region.csv"
##  [5] "data//SearchTrend.csv"   "data//SeoulFloating.csv"
##  [7] "data//Time.csv"          "data//TimeAge.csv"
##  [9] "data//TimeGender.csv"    "data//TimeProvince.csv"
## [11] "data//Weather.csv"
```

Read individual csv files

```
case_dt = fread(tables[1])
pinfo_dt = fread(tables[2])
policy_dt = fread(tables[3])
region_dt = fread(tables[4])
search_dt = fread(tables[5])
seoul_dt = fread(tables[6])
time_dt = fread(tables[7])
tage_dt = fread(tables[8])
tgender_dt = fread(tables[9])
tprovince_dt = fread(tables[10])
weather_dt = fread(tables[11])
```

VISUALIZING TREND IN NUMBER OF CASES FOR VARIOUS PROVINCES

```
head(tprovince_dt,n=5)
```

```
##          date time province confirmed released deceased
## 1: 2020-01-20   16    Seoul         0        0        0
## 2: 2020-01-20   16    Busan         0        0        0
## 3: 2020-01-20   16    Daegu         0        0        0
## 4: 2020-01-20   16  Incheon         1        0        0
## 5: 2020-01-20   16  Gwangju         0         0         0
```

```r
summary(tprovince_dt)
```

```
##       date                 time          province           confirmed     
##  Min.   :2020-01-20   Min.   : 0.000   Length:2771        Min.   :   0.0  
##  1st Qu.:2020-02-29   1st Qu.: 0.000   Class :character   1st Qu.:   9.0  
##  Median :2020-04-10   Median : 0.000   Mode  :character   Median :  42.0  
##  Mean   :2020-04-10   Mean   : 4.123                      Mean   : 444.3  
##  3rd Qu.:2020-05-21   3rd Qu.:16.000                      3rd Qu.: 133.0  
##  Max.   :2020-06-30   Max.   :16.000                      Max.   :6906.0  
##     released         deceased     
##  Min.   :   0.0   Min.   :  0.00  
##  1st Qu.:   1.0   1st Qu.:  0.00  
##  Median :  21.0   Median :  0.00  
##  Mean   : 320.7   Mean   :  9.24  
##  3rd Qu.:  92.0   3rd Qu.:  1.00  
##  Max.   :6700.0   Max.   :189.00
```

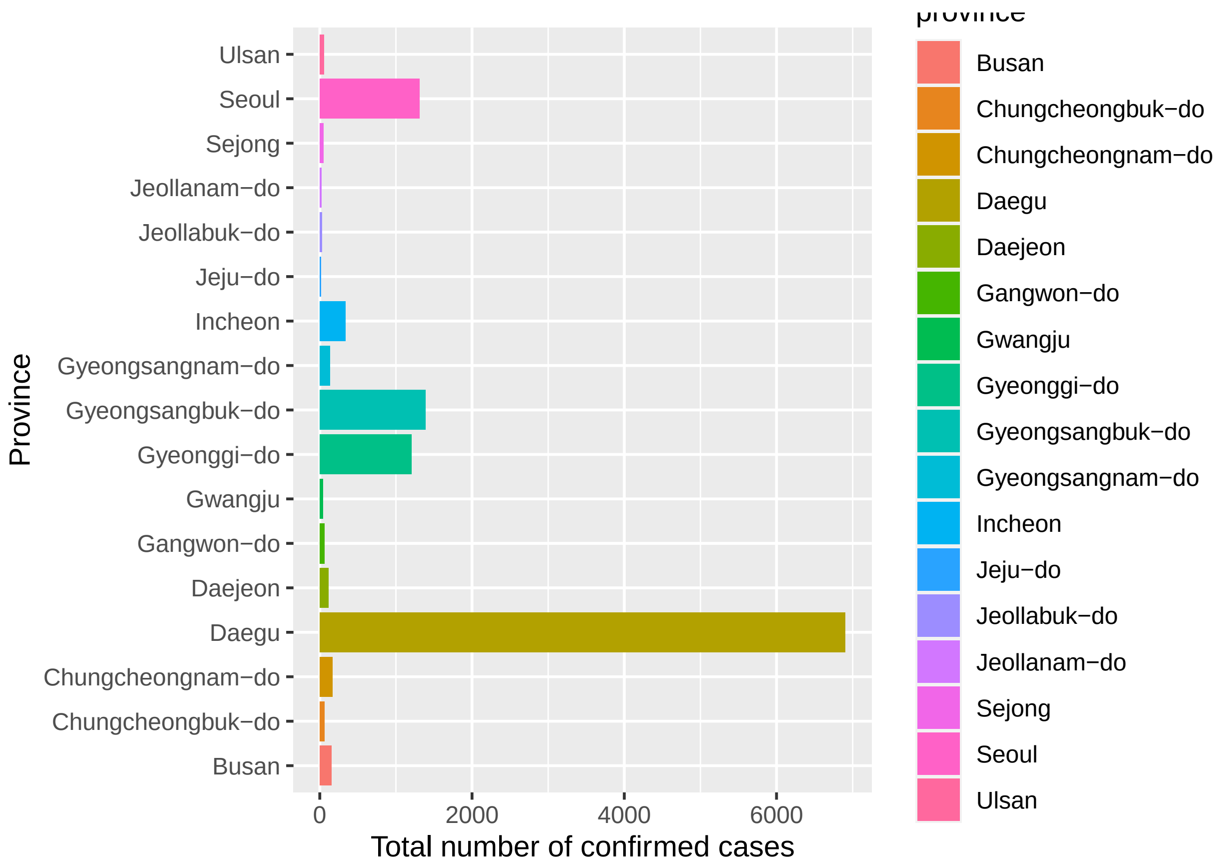
```r
#Printing the name of the various provinces
tprovince_dt[, unique(province)]
```

```
##  [1] "Seoul"             "Busan"             "Daegu"            
##  [4] "Incheon"           "Gwangju"           "Daejeon"          
##  [7] "Ulsan"             "Sejong"            "Gyeonggi-do"      
## [10] "Gangwon-do"        "Chungcheongbuk-do" "Chungcheongnam-do"
## [13] "Jeollabuk-do"      "Jeollanam-do"      "Gyeongsangbuk-do" 
## [16] "Gyeongsangnam-do"  "Jeju-do"
```

```r
# Filtering out provinces with no cases
province_cases<-tprovince_dt[, .(number_of_cases=max(confirmed)), by='province']
province_cases
```

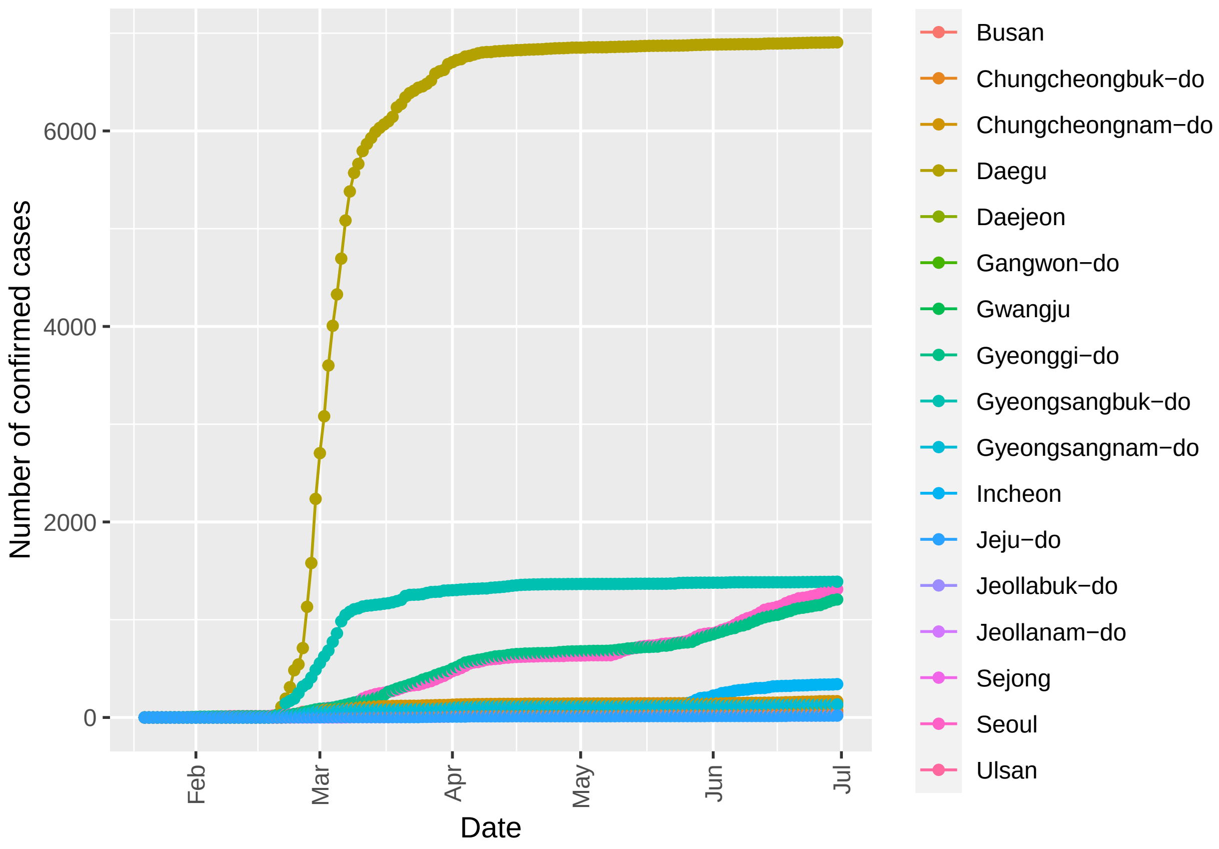
```
##              province number_of_cases
##  1:             Seoul            1312
##  2:             Busan             154
##  3:             Daegu            6906
##  4:           Incheon             341
##  5:           Gwangju              44
##  6:           Daejeon             117
##  7:             Ulsan              55
##  8:            Sejong              50
##  9:       Gyeonggi-do            1207
## 10:        Gangwon-do              65
## 11: Chungcheongbuk-do              65
## 12: Chungcheongnam-do             167
## 13:      Jeollabuk-do              27
## 14:      Jeollanam-do              24
## 15:  Gyeongsangbuk-do            1389
## 16:  Gyeongsangnam-do             134
## 17:           Jeju-do              19
```

```r
# Plotting a bar graph for number of total cases for various provinces
ggplot(province_cases, aes(x=province, y=number_of_cases , fill=province)) + geom_bar(stat='identity')
labs(x='Province', y='Total number of confirmed cases')
```



The number of cases is quite high (in thousands) for some provinces and comparatively smaller (in hundreds or less) for other provinces. So we need to choose the scale of the plot correctly.

```
# Looking at the total number of cases for each province
ggplot(tprovince_dt, aes(x=date, y=confirmed, color=province)) + geom_line() + geom_point() +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
labs(x='Date', y='Number of confirmed cases')
```


Number of confirmed cases
6000
4000
2000
0
Feb
Mar
Apr
May
Jun
Jul
Date
Busan
Chungcheongbuk−do
Chungcheongnam−do
Daegu
Daejeon
Gangwon−do
Gwangju
Gyeonggi−do
Gyeongsangbuk−do
Gyeongsangnam−do
Incheon
Jeju−do
Jeollabuk−do
Jeollanam−do
Sejong
Seoul
Ulsan