ZikaVirus

Breno Dutra

09/02/2020

Mini-Projeto - Mapeando a Ocorrência do Vírus Zika

Esse projeto faz parte do curso formação cientista de dados da DataScience Academy (www.datascienceacademy.com.br). Criaremos um mapa interativo mostrando a incidência da doença por estado Brasileiro. Houve um surto do vírus em 2016, usaremos também dados do IBGE sobre grau de instrução e tamanho da população para buscar insights. Os dados do IBGE são de 2010, assumiremos que não houve grandes mudanças de 2010 para 2016.

Etapa 1 - Carregando as bibliotecas e funções

```
#============================== CARREGANDO AS BIBLIOTECAS ==========================
============

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(stringr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```r
library(tidyr)
library(readxl)

#================================== CARREGANDO AS funções ========================
=============

source("./functions.R")
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
# função retirada do http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(gg
plot2)/
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  library(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols))
  }

  if (numPlots==1) {
    print(plots[[1]])

  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}


getLatLong <- function(x, specialSymbol = '_', column = "location"){
  for(indice in 1:dim(x)[1])
  {
    if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "acre"){
      x[indice, 'LAT'] <- -8.77
      x[indice, 'LON'] <- -70.55
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "alagoa
s"){
      x[indice, 'LAT'] <- -9.62
      x[indice, 'LON'] <- -36.82
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "amazona
s"){
      x[indice, 'LAT'] <- -3.47
      x[indice, 'LON'] <- -65.10
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "amapa")
```

```
{
      x[indice, 'LAT'] <- -1.41
      x[indice, 'LON'] <- -51.77
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "bahia")
{
      x[indice, 'LAT'] <- -13.29
      x[indice, 'LON'] <-  -41.71
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "ceara")
{
      x[indice, 'LAT'] <- -5.20
      x[indice, 'LON'] <- -39.53
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "distrit
ofederal"){
      x[indice, 'LAT'] <- -15.83
      x[indice, 'LON'] <-  -47.86
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "espirit
osanto"){
      x[indice, 'LAT'] <- -19.19
      x[indice, 'LON'] <-  -40.34
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "goias")
{
      x[indice, 'LAT'] <- -15.98
      x[indice, 'LON'] <-  -49.86
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "maranha
o"){
      x[indice, 'LAT'] <- -5.42
      x[indice, 'LON'] <- -45.44
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "matogro
sso"){
      x[indice, 'LAT'] <- -12.64
      x[indice, 'LON'] <-  -55.42
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "matogro
ssodosul"){
      x[indice, 'LAT'] <- -20.51
      x[indice, 'LON'] <-  -54.5
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "minasge
rais"){
      x[indice, 'LAT'] <- -18.10
      x[indice, 'LON'] <-  -44.38
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "para"){
      x[indice, 'LAT'] <- -3.79
      x[indice, 'LON'] <- -52.48
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "paraib
a"){
      x[indice, 'LAT'] <- -7.28
      x[indice, 'LON'] <- -36.72
    }
```

```
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "parana"
){
      x[indice, 'LAT'] <- -24.8
      x[indice, 'LON'] <-  -51.5
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "pernamb
uco"){
      x[indice, 'LAT'] <- -8.38
      x[indice, 'LON'] <- -37.86
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "piaui")
{
      x[indice, 'LAT'] <- -6.60
      x[indice, 'LON'] <- -42.28
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "riodeja
neiro"){
      x[indice, 'LAT'] <- -22.2
      x[indice, 'LON'] <-  -42.6
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "riogran
dedonorte"){
      x[indice, 'LAT'] <- -5.81
      x[indice, 'LON'] <- -36.59
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "rondoni
a"){
      x[indice, 'LAT'] <- -10.83
      x[indice, 'LON'] <-  -63.34
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "riogran
dedosul"){
      x[indice, 'LAT'] <- -30.17
      x[indice, 'LON'] <-  -53.50
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "roraim
a"){
      x[indice, 'LAT'] <- -1.99
      x[indice, 'LON'] <- -61.33
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "santaca
tarina"){
      x[indice, 'LAT'] <- -27.45
      x[indice, 'LON'] <-  -50.95
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "sergip
e"){
      x[indice, 'LAT'] <- -10.57
      x[indice, 'LON'] <-  -37.45
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "saopaul
o"){
      x[indice, 'LAT'] <- -22.19
      x[indice, 'LON'] <-  -48.79
    }
    else if( tolower(unique(str_replace_all(x[indice, column], '_', ''))) == "tocanti
ns"){
      x[indice, 'LAT'] <- -9.46
```

```
      x[indice, 'LON'] <- -48.26
    }
    else{
      x[indice, 'LAT'] <- NA
      x[indice, 'LON'] <- NA
    }
  }
  return(x)
}
```

Etapa 2 - Carregando os dados

```
# unindo vários arquivos csv em um dataframe
files <- list.files("./Dados", pattern = "*.csv")

zika <- (lapply(paste("./Dados/", files, sep=''), read.csv, stringsAsFactors = FALSE
))

zika <- joinByRowDataF(zika)
```

Etapa 3 - Explorando e organizando os dados

```
# explorando e organizando os dados
str(zika)
```

```
## 'data.frame':    264 obs. of  9 variables:
##  $ report_date     : chr  "2016-04-02" "2016-04-02" "2016-04-02" "2016-04-02" ...
##  $ location        : chr  "Norte" "Brazil-Rondonia" "Brazil-Acre" "Brazil-Amazona
s" ...
##  $ location_type   : chr  "region" "state" "state" "state" ...
##  $ data_field      : chr  "zika_reported" "zika_reported" "zika_reported" "zika_re
ported" ...
##  $ data_field_code : chr  "BR0011" "BR0011" "BR0011" "BR0011" ...
##  $ time_period     : logi  NA NA NA NA NA NA ...
##  $ time_period_type: logi  NA NA NA NA NA NA ...
##  $ value           : int  6295 618 375 1520 44 771 74 2893 30286 1202 ...
##  $ unit            : chr  "cases" "cases" "cases" "cases" ...
```

```
summary(zika)
```

```
##  report_date          location         location_type       data_field
##  Length:264         Length:264         Length:264         Length:264
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  data_field_code    time_period    time_period_type     value
##  Length:264         Mode:logical   Mode:logical     Min.   :     7.0
##  Class :character   NA's:264       NA's:264         1st Qu.:   498.5
##  Mode  :character                                   Median :  2028.5
##                                                     Mean   : 12405.3
##                                                     3rd Qu.:  8460.2
##                                                     Max.   :165932.0
##      unit
##  Length:264
##  Class :character
##  Mode  :character
##
##
##
```

```
sapply(zika,
       function(x){
         sum(is.na(x))
       })
```

```
##      report_date         location    location_type       data_field
##                0                0                0                0
##  data_field_code      time_period time_period_type            value
##                0              264              264                0
##             unit
##                0
```

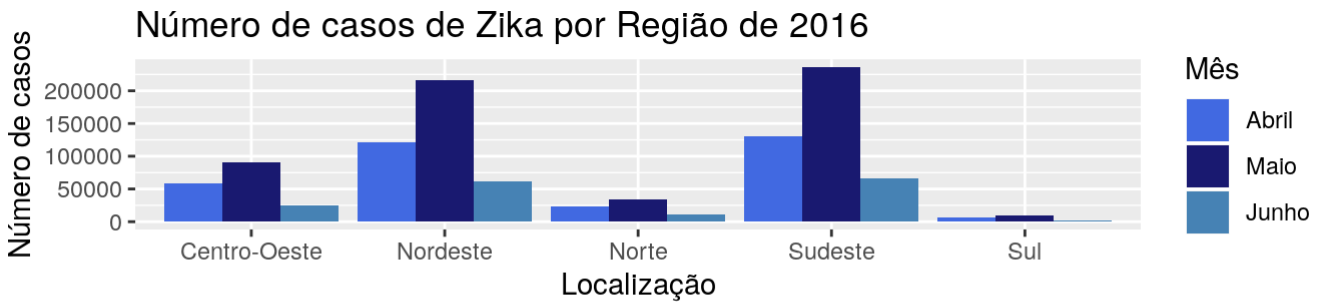
```
sapply(zika,
       function(x){
         unique(x)
       })
```

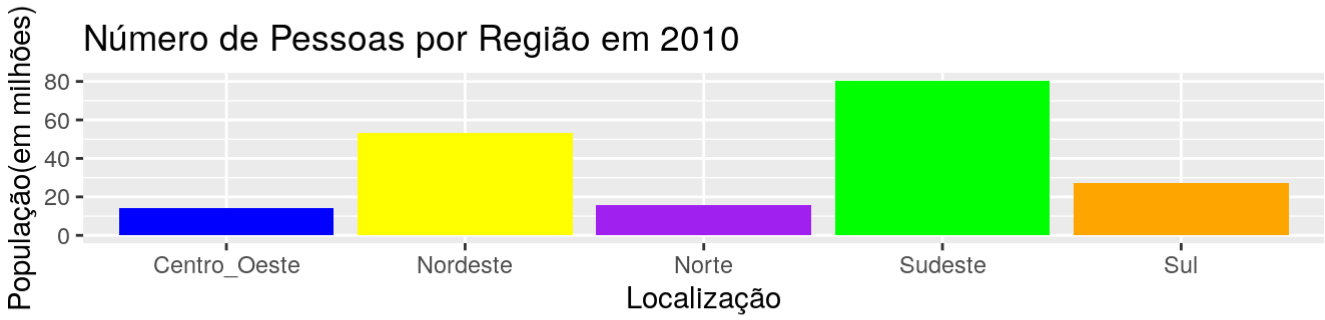
```
## $report_date
## [1] "2016-04-02" "2016-04-23" "2016-04-30" "2016-05-07" "2016-05-14"
## [6] "2016-05-21" "2016-05-28" "2016-06-11"
##
## $location
##  [1] "Norte"                      "Brazil-Rondonia"
##  [3] "Brazil-Acre"                "Brazil-Amazonas"
##  [5] "Brazil-Roraima"             "Brazil-Para"
##  [7] "Brazil-Amapa"               "Brazil-Tocantins"
##  [9] "Nordeste"                   "Brazil-Maranhao"
## [11] "Brazil-Piaui"               "Brazil-Ceara"
## [13] "Brazil-Rio_Grande_do_Norte" "Brazil-Paraiba"
## [15] "Brazil-Pernambuco"          "Brazil-Alagoas"
## [17] "Brazil-Sergipe"             "Brazil-Bahia"
## [19] "Sudeste"                    "Brazil-Minas_Gerais"
## [21] "Brazil-Espirito_Santo"      "Brazil-Rio_de_Janeiro"
## [23] "Brazil-Sao_Paulo"           "Sul"
## [25] "Brazil-Parana"              "Brazil-Santa_Catarina"
## [27] "Brazil-Rio_Grande_do_Sul"   "Centro-Oeste"
## [29] "Brazil-Mato_Grosso_do_Sul"  "Brazil-Mato_Grosso"
## [31] "Brazil-Goias"               "Brazil-Distrito_Federal"
## [33] "Brazil"
##
## $location_type
## [1] "region"  "state"   "country"
##
## $data_field
## [1] "zika_reported"
##
## $data_field_code
## [1] "BR0011"
##
## $time_period
## [1] NA
##
## $time_period_type
## [1] NA
##
## $value
##   [1]   6295    618    375   1520     44    771     74   2893  30286   1202
##  [11]      7    156    640   1060    333   1479    348  25061  35505   6693
##  [21]   1382  25930   1500   1797   1540     62    195  17504    296  16055
##  [31]    920    233  91387   8545    716    127   2172   1079    783     56
##  [41]   3612  43000   1906  34507    676   1877   1745    367     75   1443
##  [51]    404  46318   1790   9669  32312   2547   2197   1847    264     86
##  [61]  20101    276   1907  17391    527 120161   8379    732    122   1249
##  [71]     57   3264  47709   2206  37836    832   1954   2275    450     82
##  [81]   1647    427  48027   1900  10553   3262   2343   1965    287     91
##  [91]  21364    295   2278  18227    564 127822   8053    960    823     79
## [101]   1362   2535  51065   2003     95    887   1757   2452    496   2458
## [111]    497  40420  54803  11237   1918  38196   3452   2431   2025    311
## [121]  21756    621  18226   2604    305 138108   8432    974    923     83
## [131]   1583    162  54165   2276    148   1238   1788   2648   2812    499
## [141]  42260  61309  11670   2080  43516   4043   2491   2045    102    344
## [151]  22508    696   3217    369 148905   9022   1032    843   2176   2032
## [161]    189   2667  59745   2328    217   2144   2312   2865    393   3577
## [171]    490  45419  65328  12449   2150  46027   4702   2463   2001     99
```

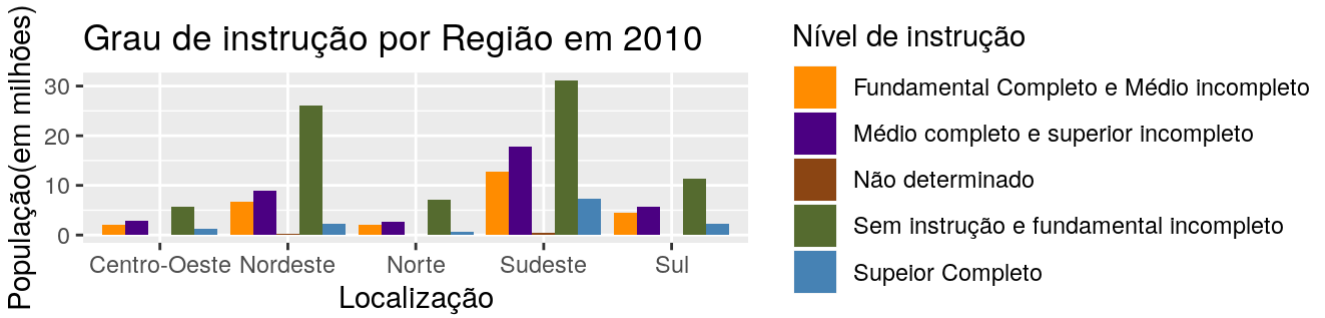
```
## [181]    363  24683    698  19662   3954 161241  10645    898    846   3713
## [191]   2121   2795  61829   2840    241   2358   2342   2889    394   3847
## [201]    491  46427  65820  12891   2166   4736   2392   1935     97    360
## [211]  25246    762  19985   4132 165932
##
## $unit
## [1] "cases"
```

```
zika <- deleteColumns(zika, c("time_period", "time_period_type"))

zika$location <- sapply(zika$location, function(x, pattern = "Brazil-", novoTexto =
""){
  return(str_replace(x, pattern = pattern, novoTexto))
})

zika$report_date <- as.Date(zika$report_date)

zika$Dia <- day(zika$report_date)
zika$Mes <- month(zika$report_date)
zika$Ano <- year(zika$report_date)


zika <- zika %>%
  select(c(-4, -5, -7))

# plots para analisar os dados
zika2 <- zika %>%
  select(c(report_date, location, location_type, value, Mes)) %>%
  filter(location_type == "region") %>%
  group_by(location, Mes) %>%
  summarise(casos = sum(value))

g1 <- ggplot(data = zika2, aes(x = as.factor(location), y = casos, fill = as.factor(M
es))) +
  geom_bar(stat = "identity", position = 'dodge') +
  ggtitle("Número de casos de Zika por Região de 2016") +
  xlab("Localização") +
  ylab("Número de casos") +
  labs(fill = "Mês") +
  scale_fill_manual(values=c("#4169E1", "#191970",  "#4682B4"),  labels = c("Abril",
"Maio", "Junho"))

# censo: https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografic
o-2010.html?edicao=9673&t=downloads
populacao2010 <- data.frame(Regiao = c("Sul", "Centro_Oeste", "Sudeste", "Norte", "No
rdeste"),
                            Populacao = c(27386891, 14058094, 80364410, 15864454, 530
81950))

color <- c("blue", "yellow", "purple", "green", "orange")

g2 <- ggplot(data = populacao2010, aes(x = Regiao, y = Populacao/1000000)) +
  geom_bar(stat = "identity", position = 'dodge', fill = color) +
  ggtitle("Número de Pessoas por Região em 2010") +
  xlab("Localização") +
  ylab("População(em milhões)")

library(readxl)

# pessoas acima de 10 anos de idade por grau de instrução
escolaridade <- read_xls("./Dados/EscolaridadeCenso2010.xls")

escolaridade$Total <- NULL

escolaridade2 <- escolaridade %>%
```

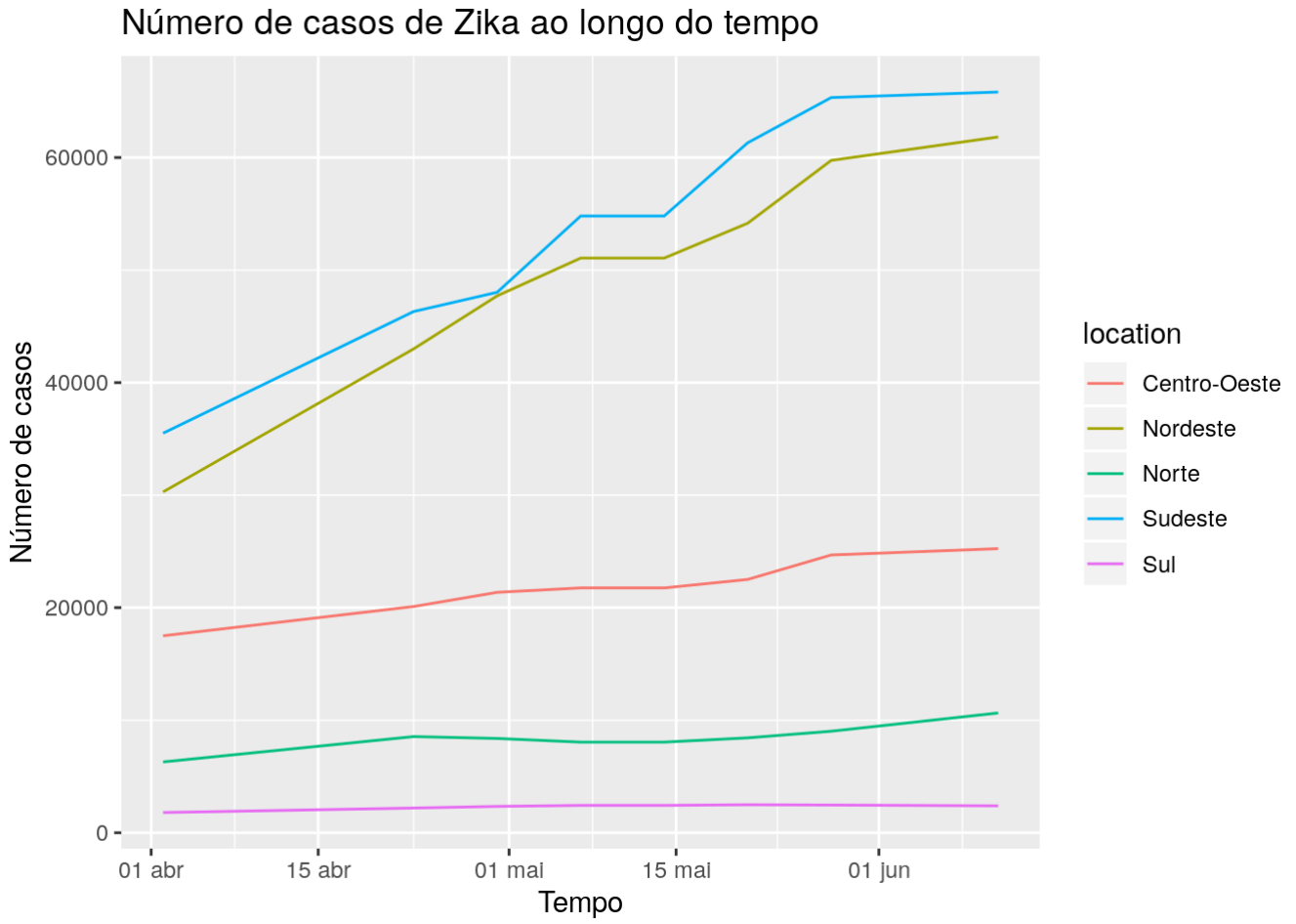
```
  filter(Localizacao %in% c("Norte", "Sudeste", "Centro-Oeste", "Nordeste", "Sul"))

names(escolaridade2) <- c("Localizacao", "Sem_instrução_e_fundamental_incompleto", "F
undamental_completo_e_médio_incompleto", "Médio_completo_e_superior_incompleto", "Sup
erior_completo", "Não_determinado")

# modificando os dados para facilitar na hora do plot
escolaridade2 <- escolaridade2 %>%
  gather(Instrucao, Populacao, -Localizacao)

g3 <- ggplot(data = escolaridade2, aes(x = as.factor(Localizacao), y = Populacao/1000
000, fill = as.factor(Instrucao))) +
  geom_bar(stat = "identity", position = 'dodge') +
  ggtitle("Grau de instrução por Região em 2010") +
  xlab("Localização") +
  ylab("População(em milhões)") +
  labs(fill = "Nível de instrução") +
  scale_fill_manual(values = c("#FF8C00", "#4B0082", "#8B4513", "#556B2F", "#4682B4"
),
                    labels = c("Fundamental Completo e Médio incompleto",
                               "Médio completo e superior incompleto",
                               "Não determinado",
                               "Sem instrução e fundamental incompleto",
                               "Supeior Completo"))

multiplot(g1, g2, g3)
```





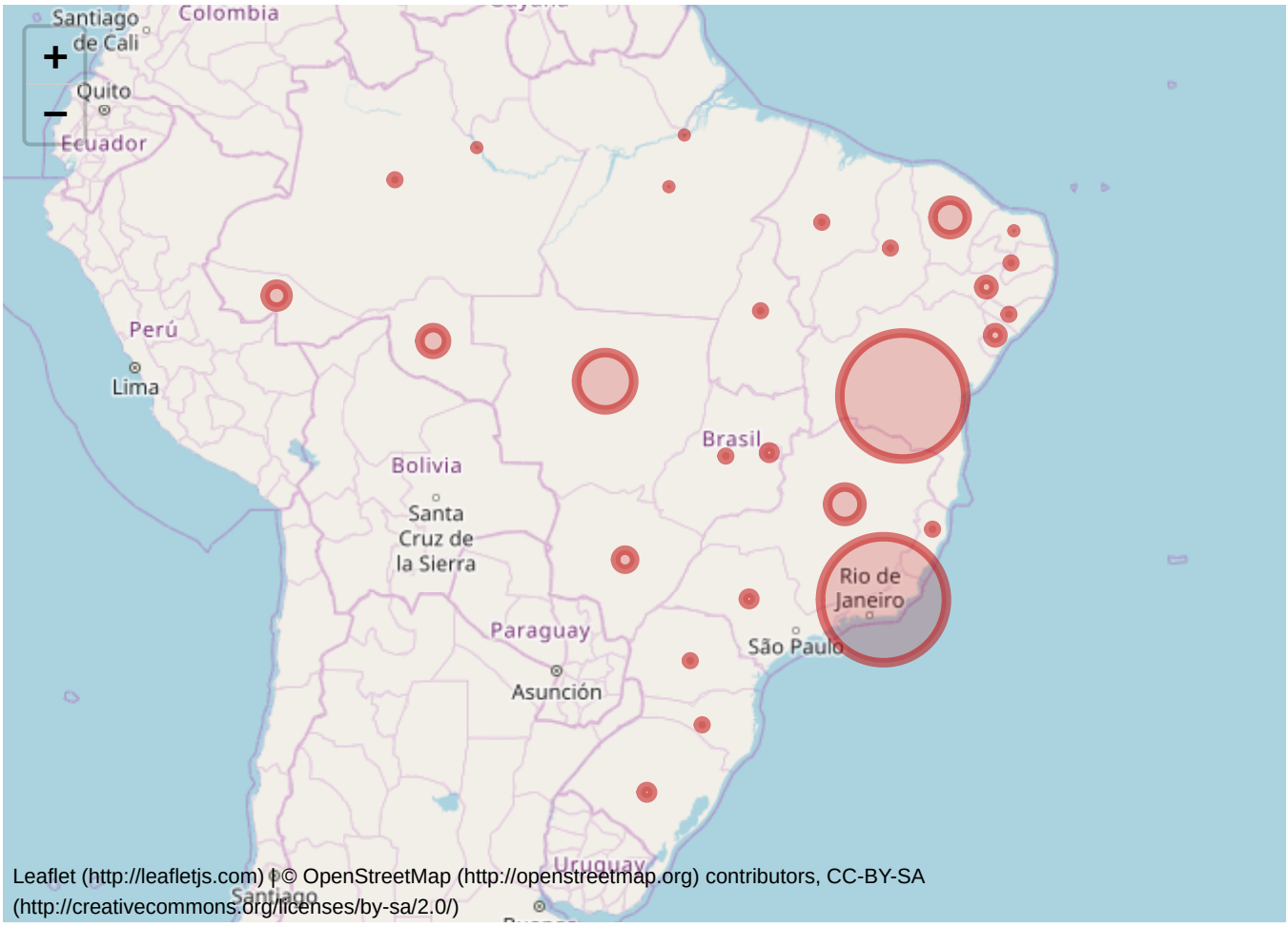


```
zika3 <- zika %>%
  select(c(report_date, location, location_type, value)) %>%
  filter(location_type == "region") %>%
  group_by(report_date, location) %>%
  summarise(casos = sum(value))


ggplot(zika3, aes(x=report_date, y=casos, color=location)) +
  geom_line() +
  ggtitle("Número de casos de Zika ao longo do tempo") +
  xlab("Tempo") +
  ylab("Número de casos") +
  scale_x_date(date_labels = "%d %b")
```



Etapa 4 - Criando os mapas interativos

```
# Mapa interativo das regiões
library(leaflet)

zika <- getLatLong(zika)
zika_map <- (zika %>%
               filter(location_type == 'state') %>%
               group_by(location, LAT, LON) %>%
               summarise(casos = sum(value))
)

leaflet(data = zika_map) %>%
  addTiles() %>%
  addCircles(lat = zika_map$LAT,
             lng = zika_map$LON,
             popup = ~as.character(zika_map$casos),
             fillColor = c('#c00000'),
             color = c('#c00000'),
             label = zika_map$location,
             radius = sapply(zika_map$casos,
                             function(x){
                               if(x <= 10000 && x >= 3000)
                                 return(x*10)
                               else if(x>1000 && x < 3000)
                                 return(x*12)
                               else if(x <= 1000)
                                 return(x*18)
                               else
                                 return(x)
                             }))
```



Fim