Tutorial, Analise Exploratorios de Dados

Code

Lourdes M.M Villavicencio

8/26/2019

Este tutorial, apresenta uma analise exploratoria dos dados do zika virus, Você encontrara o dataset no site da Kaggel https://www.kaggle.com/cdc/zika-virus-epidemic (https://www.kaggle.com/cdc/zika-virus-epidemic)

Carregamos as librerias necessarias

```
library(dplyr)
library(data.table)
library(ggplot2)
library(RColorBrewer)
library(rworldmap)
library(tidyr)
```

Carregamos os dados

```
setwd("~/Milagros/cursosR/cdc_zika.csv")
zika <- read.csv('cdc_zika.csv',stringsAsFactors = F, header = T)
```

```
class(zika)
```

```
[1] "data.frame"
```

```
dim(zika)
```

```
[1] 107619      9
```

```
colnames(zika)
```

```
[1] "report_date"      "location"         "location_type"
[4] "data_field"       "data_field_code"  "time_period"
[7] "time_period_type" "value"            "unit"
```

```
str(zika)
```

```
'data.frame':   107619 obs. of  9 variables:
 $ report_date     : chr  "2016-03-19" "2016-03-19" "2016-03-19" "2016-03-1
9" ...
 $ location        : chr  "Argentina-Buenos_Aires" "Argentina-Buenos_Aires"
"Argentina-Buenos_Aires" "Argentina-Buenos_Aires" ...
 $ location_type   : chr  "province" "province" "province" "province" ...
 $ data_field      : chr  "cumulative_confirmed_local_cases" "cumulative_pro
bable_local_cases" "cumulative_confirmed_imported_cases" "cumulative_probabl
e_imported_cases" ...
 $ data_field_code : chr  "AR0001" "AR0002" "AR0003" "AR0004" ...
 $ time_period     : logi  NA NA NA NA NA NA ...
 $ time_period_type: logi  NA NA NA NA NA NA ...
 $ value           : chr  "0" "0" "2" "1" ...
 $ unit            : chr  "cases" "cases" "cases" "cases" ...
```

```
zika <- zika %>%
separate(col = "report_date", into = c("year", "month"), sep = "-") %>%
separate(col = "location", into = c("country", "state"), sep = "-")
```

```
Expected 2 pieces. Additional pieces discarded in 107372 rows [1, 2, 3, 4,
5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].Expected 2 p
ieces. Missing pieces filled with `NA` in 90 rows [104408, 104409, 104410, 1
04411, 104412, 104413, 104414, 104415, 104416, 104417, 104418, 104419, 10442
0, 104421, 104422, 104423, 104424, 104425, 104426, 104427, ...].Expected 2 p
ieces. Additional pieces discarded in 88610 rows [6305, 6306, 6307, 6308, 63
09, 6310, 6311, 6312, 6313, 6314, 6315, 6316, 6317, 6318, 6319, 6320, 6321,
6322, 6323, 6324, ...].Expected 2 pieces. Missing pieces filled with `NA` i
n 1584 rows [6074, 6082, 6092, 6097, 6106, 6107, 6115, 6125, 6130, 6139, 614
0, 6148, 6158, 6163, 6172, 6173, 6181, 6191, 6196, 6205, ...].
```

```
zika$value <- as.character(zika$value)
zika$value <- as.numeric(zika$value,na.rm=TRUE)
```

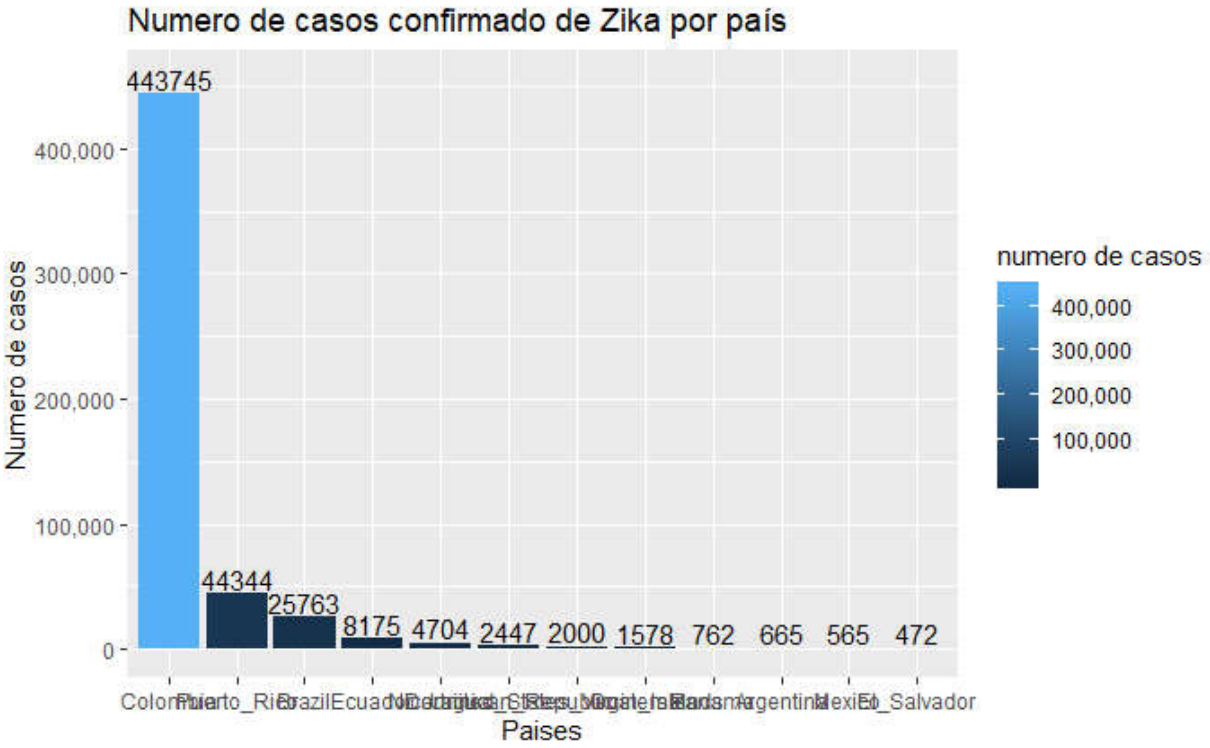
#organizando os meses

```
zika$month_name <- ""
zika$month_name[which(zika$month == "01")] <- "jan"
zika$month_name[which(zika$month == "02")] <- "fev"
zika$month_name[which(zika$month == "03")] <- "mar"
zika$month_name[which(zika$month == "04")] <- "abr"
zika$month_name[which(zika$month == "05")] <- "mai"
zika$month_name[which(zika$month == "06")] <- "jun"
zika$month_name[which(zika$month == "07")] <- "jul"
zika$month_name[which(zika$month == "11")] <- "nov"
zika$month_name[which(zika$month == "12")] <- "dez"
```

```
zika$confirmed <- ifelse(grepl('confirmed', zika$data_field), TRUE, FALSE)
zika$mc <- ifelse(grepl('microcephaly', zika$data_field), TRUE, FALSE)
```

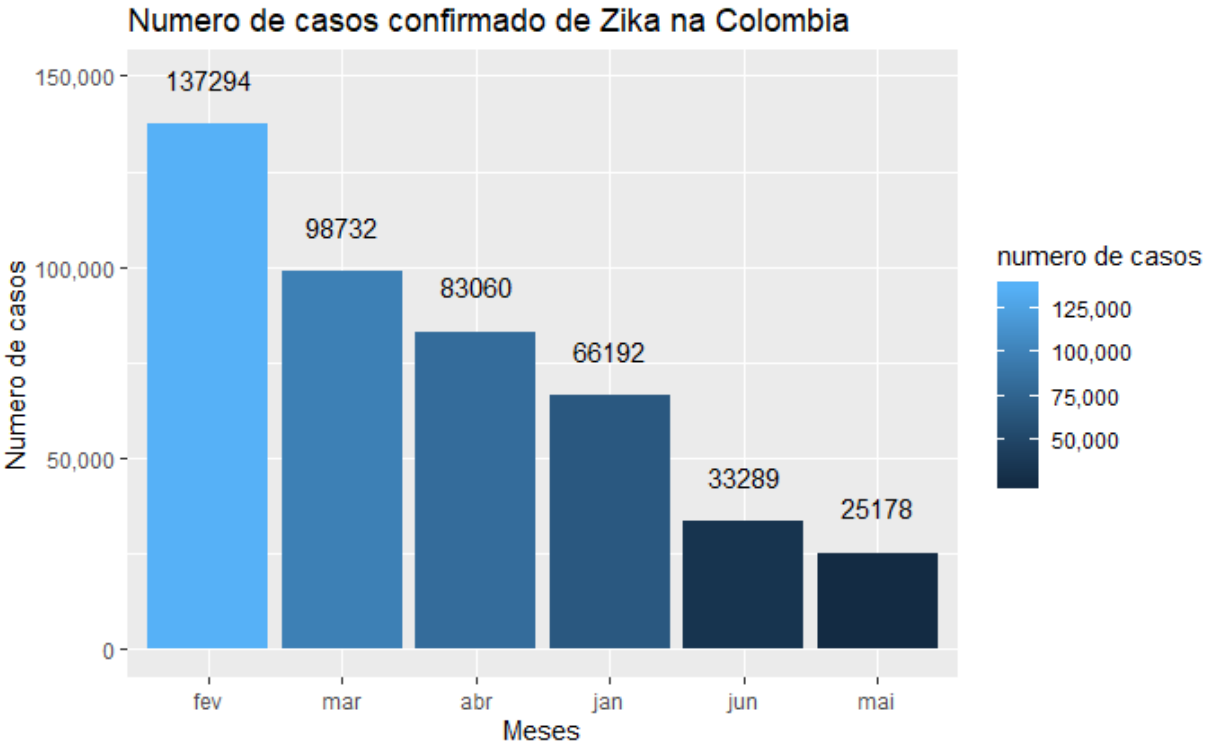
Grafico de barras, dos paises com casos confirmados com zika

```
country_data <- zika %>%
filter(confirmed == T & unit == 'cases' & !is.na(value)) %>%
group_by(country) %>%
summarise(num_cases = sum(value,na.rm = T)) %>%
arrange(desc(num_cases))

library(scales)
ggplot(data=country_data,
     aes(x=reorder(country, desc(num_cases)),
           y=num_cases,
           fill=num_cases))+
  geom_bar(stat='identity') +
  geom_text(aes(x = reorder(country, desc(num_cases)),
                y = num_cases, label = num_cases), nudge_y = 12000) +
  labs(x = "Paises",
       y = "Numero de casos",
       title = "Numero de casos confirmado de Zika por país") +
  scale_fill_continuous(name= "numero de casos", labels=comma) +
  scale_y_continuous(labels =  comma)
```

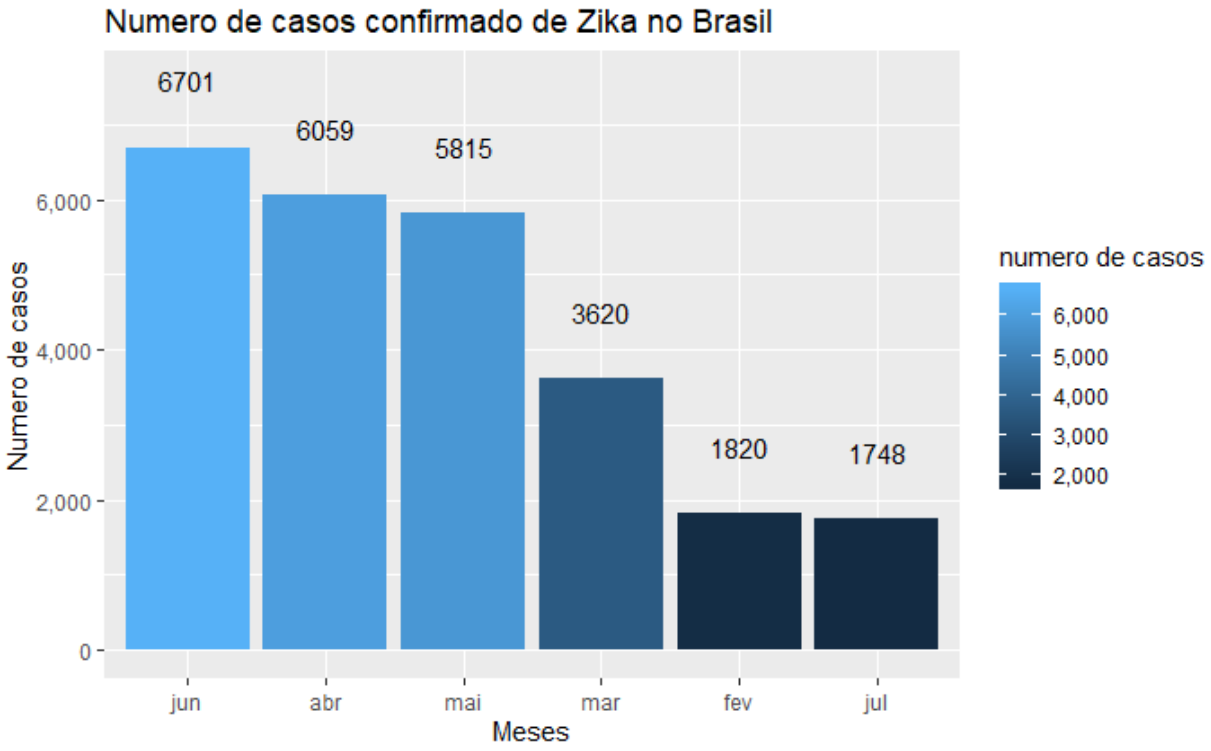


Gráficos de barras, mostrando por mes, as incidencias de Zika nos paises onde houve maior incidencia de Zika no ano de 2016.

```
colombia_2016 <- zika %>%
filter(confirmed == T & unit == 'cases' & !is.na(value)) %>%
select(year, month, country, value, month_name) %>%
filter(country == "Colombia") %>%
group_by(month, month_name) %>%
summarise(num_cases = sum(value,na.rm = T)) %>%
arrange(desc(num_cases))

ggplot(data=colombia_2016,
        aes(x=reorder(month_name, desc(num_cases)),
            y=num_cases,
            fill=num_cases))+
   geom_bar(stat='identity') +
   geom_text(aes(x = reorder(month_name, desc(num_cases)),
                 y = num_cases, label = num_cases), nudge_y = 12000) +
   labs(x = "Meses",
        y = "Numero de casos",
        title = "Numero de casos confirmado de Zika na Colombia") +
   scale_fill_continuous(name= "numero de casos", labels=comma) +
   scale_y_continuous(labels =  comma)
```



```
Brasil2016 <- zika %>%
  filter(confirmed == T & unit == 'cases' & !is.na(value)) %>%
  select(year, month, country, value, month_name) %>%
  filter(country == "Brazil") %>%
  group_by(month, month_name) %>%
  summarise(num_cases = sum(value,na.rm = T)) %>%
  arrange(desc(num_cases))

ggplot(data=Brasil2016,
       aes(x=reorder(month_name, desc(num_cases)),
           y=num_cases,
           fill=num_cases))+
  geom_bar(stat='identity') +
  geom_text(aes(x = reorder(month_name, desc(num_cases)),
                y = num_cases, label = num_cases), nudge_y = 900) +
  labs(x = "Meses",
       y = "Numero de casos",
       title = "Numero de casos confirmado de Zika no Brasil") +
  scale_fill_continuous(name= "numero de casos", labels=comma) +
  scale_y_continuous(labels =  comma)
```



Box Plot, para o numero de casos por região no Brasil.

```
region<- zika%>%
select(country, year, value,month, month_name)%>%
filter(country %in% c("Norte","Nordeste","Sudeste","Sul","Centro"))%>%
group_by(year,month, month_name, country) %>%
summarise(num_cases = sum(value,na.rm  = T)) %>%
arrange(desc(num_cases))

ggplot(region, aes(x=country, y=num_cases)) +
geom_boxplot( colour = "black", fill = "#56B4E9") +
labs(title="BoxPlot de casos de Zika por Região do Brasil",x="Região", y =
"Numero de casos")
```

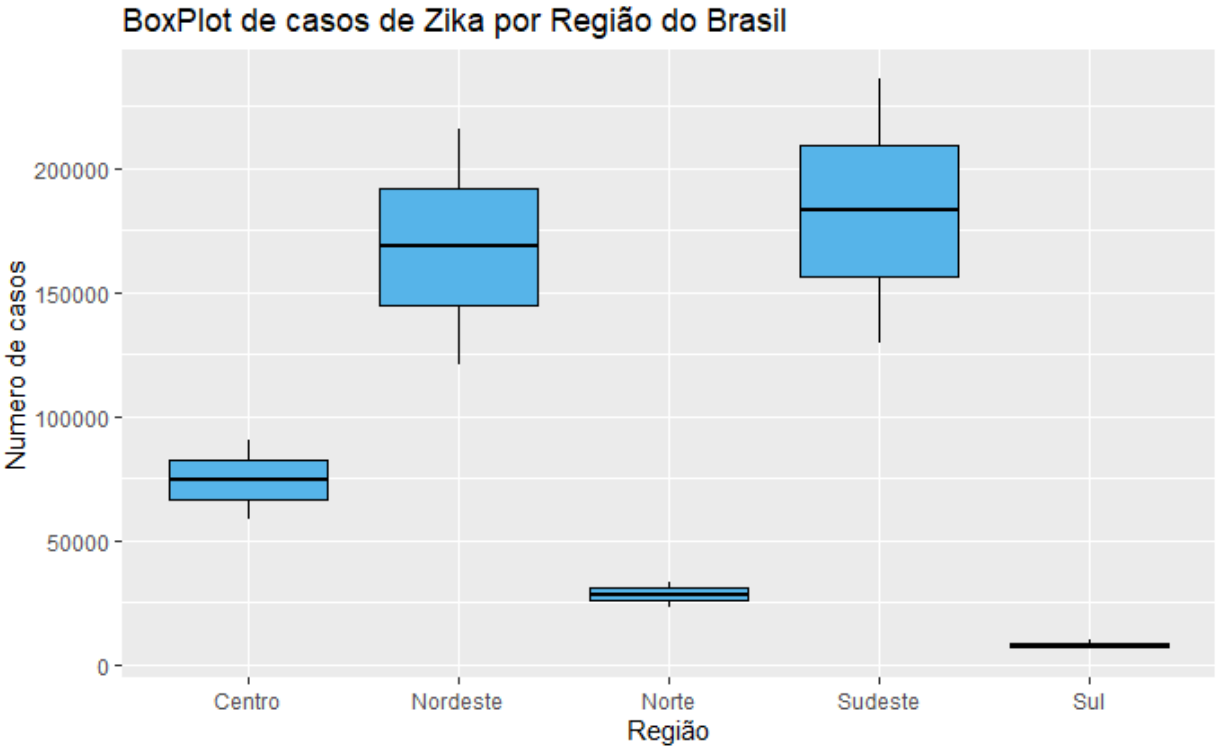


Grafico de barras, mostrando a incidencia total de casos confirmados no brasil por estado e por mes.

```
BR_uf <- zika %>%
filter(confirmed == T & unit == 'cases' & !is.na(value)) %>%
select(year, month, month_name,state, country, state, value) %>%
filter(country == "Brazil") %>%
group_by(month_name, month,state) %>%
drop_na(state) %>%
summarise(num_cases = sum(value,na.rm  = T)) %>%
arrange(desc(num_cases))
ggplot(BR_uf ,
       aes(x = month_name, y = num_cases, fill = state)) +
      geom_bar(stat='identity') +
      labs(x = "Mes",
       y = "Numero de casos",
       title = "Numero de casos confirmados com Zika por estados do Brasil")
```

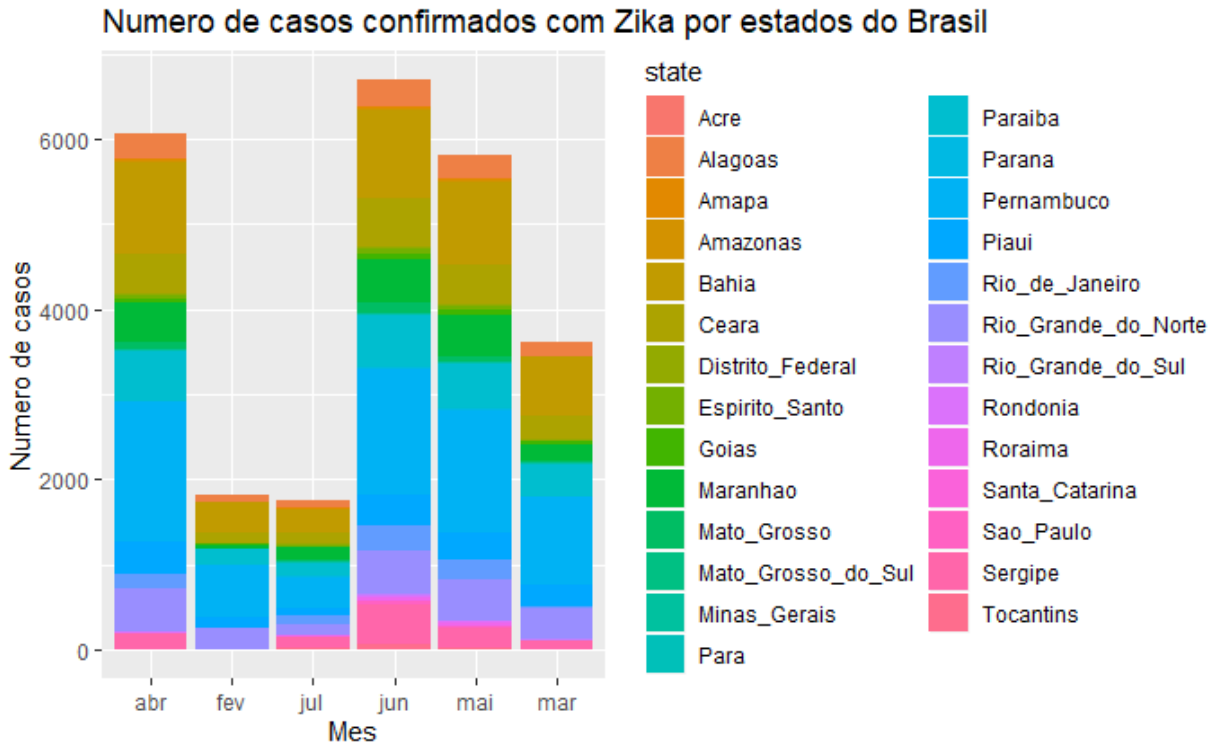
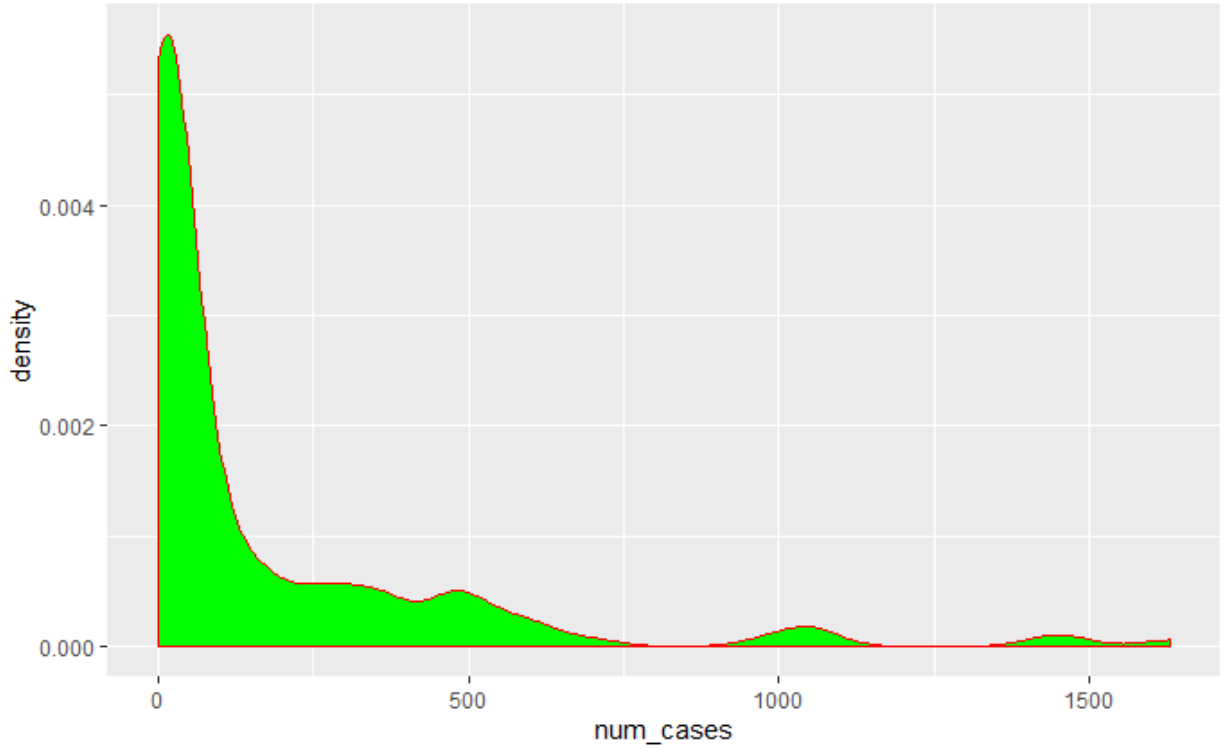
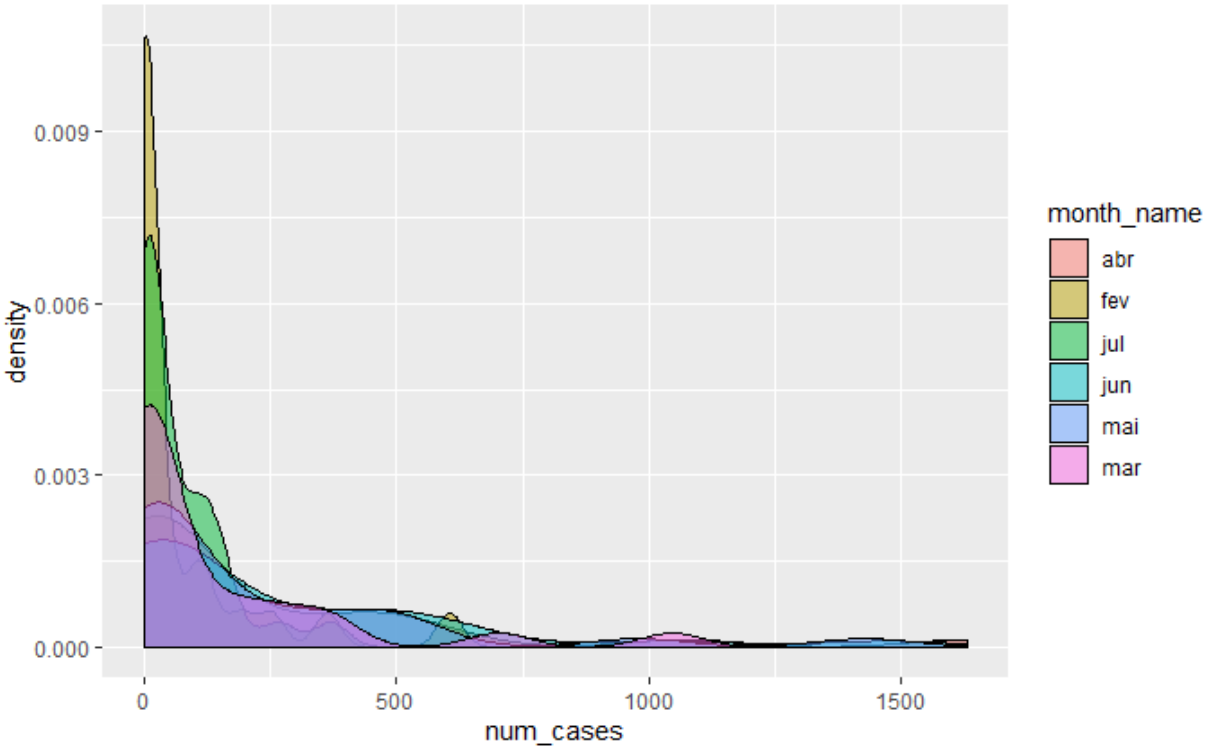


Gráfico de função de densidade de probabilidade, mostra a distribuição de casos de zika.

```
BR_uf %>%
  ggplot(aes(x = num_cases)) +
  geom_density(color = "red", fill = "green")
```



```
BR_uf %>%
  ggplot(aes(x = num_cases))+
  geom_density(aes(fill = month_name),
               # adicionar transparencia
               alpha = 0.5 )
```



Cálculo da media do aparecimento de Zika no brasil no ano de 2016, em cada estado brasileiro reportado na base de dados do Zika Virus. Assim como a mÃ©dia de casos por 100 mil habitantes em cada estado.

```
mean_uf <- zika %>%
    select(year, month, country, state, value) %>%
    filter(country == "Brazil") %>%
    group_by(state) %>%
    drop_na(state) %>%
    summarise(avg_state = round(mean(value, na.rm = T), digits = 0)) %>%
    mutate(avg_pop = avg_state/100000)
mean_uf
```

Apresenta-se uma exemplo simples de um test- t para duas amostras independentes, aqui não levamos em conta a normalidade dos dados. H0: as medias são iguais H1: as medias são deiferentes Se rejeita H0 para um p-valor<0.05

```
Acre<- zika%>%
select(state, year, value,month, month_name)%>%
filter(state == "Acre" ) %>%
group_by(year,month, month_name, state) %>%
summarise(num_cases = sum(value,na.rm  = T))%>%
arrange(desc(num_cases))

Alagoas<- zika%>%
  select(state, year, value,month, month_name)%>%
  filter(state == "Alagoas" ) %>%
  group_by(year,month, month_name, state) %>%
  summarise(num_cases = sum(value,na.rm  = T))%>%
  arrange(desc(num_cases))
```

```
dados <- data.frame( resposta = c(Acre$num_cases,Alagoas$num_cases),
  grupos=c(rep("Acre", 6), rep("Alagoas", 6)))
dados
```

```
teste_t <- t.test(resposta ~ grupos, data=dados, var.equal=TRUE)
teste_t
```

```
    Two Sample t-test

data:  resposta by grupos
t = -1.4415, df = 10, p-value = 0.18
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7845.819  1681.819
sample estimates:
   mean in group Acre mean in group Alagoas
             1023.667              4105.667
```