Projeto - Causas de mortes - Brasil

André Campos da Silva

7 de Novembro, 2021

Projeto - Analise de causas de mortes de doenças no Brasil 2019/2020.

Realizar uma analise exploratória das ocorrências de mortes causadas por diferentes doenças no Brasil.

Os cartórios do Brasil decidiram disponibilizar alguns prontuários sobre as mortes registradas no Brasil desde a pandemia do coronavírus. Através dos pedidos de visualização destes gráficos é possível obter os dados brutos e fazer as nossas próprias análises ou visualizações.

Esses dados contêm o número de óbitos registrados por dia, estado, sexo, idade, cor da pele e causa da morte (principalmente com foco em covid-19 e doenças cardiovasculares) ocorridos entre 01-01-2019 e 15-09-2020.

Esses dados foram coletados entre 14/09/2020 e 16/09/2020 e podem ser atualizados, pois pode demorar alguns dias até que o óbito seja registrado pela família, no cartório e posteriormente disponibilizado na plataforma.

Link do dataset https://www.kaggle.com/amandalk/cause-of-death-in-brazil-20192020 (https://www.kaggle.com/amandalk/cause-of-death-in-brazil-20192020)

Carregando os pacotes

```
# Pacotes usados no projeto

library('tidyverse')
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("plyr")
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
library('lubridate')
```

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

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

```
library('gridExtra')
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

Carregando os Dados

```
# Carrego o dataset para análise

data <- read_csv('Dados/death_cause_brazil.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   date = col_date(format = ""),
##   state = col_character(),
##   gender = col_character(),
##   age = col_character(),
##   color = col_character(),
##   cause = col_character(),
##   total = col_double()
## )
```

```
data <- as.data.frame(data)
```

Tratamento dos dados

```
# Vou extrair o dia, mês e ano da variável date para variáveis separadas, para poder fazer uma a
nálise exploratória mais detalhada.

# Crio a função que vai fazer a extração e criação de novas variáveis.
# Crio uma amostra para testar.
data_test <- data[5000:10000,]
data_test$day <- sapply(data_test$date, day)
data_test$month <- sapply(data_test$date, month)
data_test$year <- sapply(data_test$date, year)
head(data_test)
```

```
##            date state gender     age color               cause total day month
## 5000 2019-01-02    PI      M 40 - 49 White          Septicemia     1   2     1
## 5001 2019-01-02    PI      M 50 - 59 White              Others     1   2     1
## 5002 2020-01-02    PI      M 50 - 59 White              Others     1   2     1
## 5003 2019-01-02    PI      M 60 - 69 White       Hearth attack     1   2     1
## 5004 2019-01-02    PI      M 70 - 79 White       Hearth attack     1   2     1
## 5005 2019-01-02    PI      M 70 - 79 White Respiratory failure     1   2     1
##      year
## 5000 2019
## 5001 2019
## 5002 2020
## 5003 2019
## 5004 2019
## 5005 2019
```

```
# Verificando que esta tudo ok, faço para todo o dataset.
data$day <- sapply(data$date, day)
data$month <- sapply(data$date, month)
data$year <- sapply(data$date, year)
head(data)
```

```
##         date state gender     age      color             cause total day month
## 1 2020-01-01    AC      F 60 - 69 East asian        Septicemia     1   1     1
## 2 2019-01-01    AC      F 80 - 89      White     Hearth attack     1   1     1
## 3 2019-01-01    AC      F 30 - 39 Indigenous            Others     1   1     1
## 4 2019-01-01    AC      F 70 - 79      Mixed Cardiogenic shock     1   1     1
## 5 2020-01-01    AC      F 70 - 79      Mixed         Pneumonia     1   1     1
## 6 2020-01-01    AC      F     < 9      Mixed         Pneumonia     1   1     1
##   year
## 1 2020
## 2 2019
## 3 2019
## 4 2019
## 5 2020
## 6 2020
```

```
# Irei adicionar uma nova coluna com as regiões para a fase de análise.

# Verifico cada valor único dos estados
unique(data$state)
```

```
##  [1] "AC" "AL" "AM" "AP" "BA" "CE" "DF" "ES" "GO" "MA" "MG" "MS" "MT" "PA" "PB"
## [16] "PE" "PI" "PR" "RJ" "RN" "RO" "RR" "RS" "SC" "SE" "SP" "TO"
```

```
# Crio as regiões contendo as siglas dos estados para usar na programação de atribuição.
Norte <- c('AC','AP','AM','PA','RO','RR','TO')
Sudeste <- c('ES','MG','RJ','SP')
Nordeste <- c('AL','BA','CE','MA','PB','PE','PI','RN','SE')
Centro_Oeste <- c('DF','GO','MT','MS')
Sul <- c('PR','RS','SC')

# Crio a variável region com o valor NA
data$region <- NA
head(data)
```

```
##         date state gender     age      color             cause total day month
## 1 2020-01-01    AC      F 60 - 69 East asian        Septicemia     1   1     1
## 2 2019-01-01    AC      F 80 - 89      White     Hearth attack     1   1     1
## 3 2019-01-01    AC      F 30 - 39 Indigenous            Others     1   1     1
## 4 2019-01-01    AC      F 70 - 79      Mixed Cardiogenic shock     1   1     1
## 5 2020-01-01    AC      F 70 - 79      Mixed         Pneumonia     1   1     1
## 6 2020-01-01    AC      F     < 9      Mixed         Pneumonia     1   1     1
##   year region
## 1 2020     NA
## 2 2019     NA
## 3 2019     NA
## 4 2019     NA
## 5 2020     NA
## 6 2020     NA
```

```
# Aplico a função adicionando o valor respondente para a região na coluna criada anteriormente.
# Caso o pc tenha muita memória rode esse que está comentando pois ja fara todas as regiões em u
ma so programação, porem caso tenha pouca memória ou queria acompanhar o processo feito por regi
ão, rode o segundo que será o que vou usar, um por vez.

"for (i in 1:length(data$state)){
  if (data$state[i] %in% Norte){
    data$region[i] = 'Norte'

  }else if
  (data$state[i] %in% Nordeste){
    data$region[i] = 'Nordeste'

  }else if
  (data$state[i] %in% Sudeste){
    data$region[i] = 'Sudeste'

  }else if
  (data$state[i] %in% Sul){
    data$region[i] = 'Sul'

  }else if
  (data$state %in% Centro_Oeste){
    data$region[i] = 'Centro_Oeste'
  }
}"
```

```
## [1] "for (i in 1:length(data$state)){\n  if (data$state[i] %in% Norte){\n    data$region[i] =
'Norte'\n    \n  }else if \n  (data$state[i] %in% Nordeste){\n    data$region[i] = 'Nordeste'\n
\n  }else if \n  (data$state[i] %in% Sudeste){\n    data$region[i] = 'Sudeste'\n    \n  }else if
\n  (data$state[i] %in% Sul){\n    data$region[i] = 'Sul'\n    \n  }else if\n  (data$state %in%
Centro_Oeste){\n    data$region[i] = 'Centro_Oeste'  \n  }    \n}"
```

```
# Descomente e execute cada loop for por vez.

"
for (i in 1:length(data$state)){
  if (data$state[i] %in% Nordeste){
    data$region[i] = 'Nordeste'
  }}

for (i in 1:length(data$state)){
  if (data$state[i] %in% Sudeste){
    data$region[i] = 'Sudeste'
  }}

for (i in 1:length(data$state)){
  if (data$state[i] %in% Sul){
    data$region[i] = 'Sul'
  }}

for (i in 1:length(data$state)){
  if (data$state[i] %in% Centro_Oeste){
    data$region[i] = 'Centro_Oeste'
  }}

for (i in 1:length(data$state)){
  if (data$state[i] %in% Norte){
    data$region[i] = 'Norte'
  }}
"
```

```
## [1] "\nfor (i in 1:length(data$state)){\n  if (data$state[i] %in% Nordeste){\n    data$region
[i] = 'Nordeste'\n  }}\n\nfor (i in 1:length(data$state)){\n  if (data$state[i] %in% Sudeste){\n
data$region[i] = 'Sudeste'\n  }}  \n\nfor (i in 1:length(data$state)){\n  if (data$state[i] %in%
Sul){\n    data$region[i] = 'Sul'\n  }}\n\nfor (i in 1:length(data$state)){\n  if (data$state[i]
%in% Centro_Oeste){\n    data$region[i] = 'Centro_Oeste'\n  }}\n\nfor (i in 1:length(data$stat
e)){\n  if (data$state[i] %in% Norte){\n    data$region[i] = 'Norte'\n  }}\n"
```

```
# Salvo o dataset já tratado para análise.
# write_csv(data,'Dados/death_cause_brazil_treated.csv')

# Carrego o dataset tratado sobreponto o data.
data <- read_csv('Dados/death_cause_brazil_treated.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   date = col_date(format = ""),
##   state = col_character(),
##   gender = col_character(),
##   age = col_character(),
##   color = col_character(),
##   cause = col_character(),
##   total = col_double(),
##   day = col_double(),
##   month = col_double(),
##   year = col_double(),
##   region = col_character()
## )
```

```
data <- as.data.frame(data)
```

Análise Exploratória de Dados

```
# Verifico os formatos dos dados
glimpse(data)
```

```
## Rows: 1,098,241
## Columns: 11
## $ date   <date> 2020-01-01, 2019-01-01, 2019-01-01, 2019-01-01, 2020-01-01,...
## $ state  <chr> "AC", "AC", "AC", "AC", "AC", "AC", "AC", "AC", "AC", "AC", ...
## $ gender <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "M", "M", ...
## $ age    <chr> "60 - 69", "80 - 89", "30 - 39", "70 - 79", "70 - 79", "< 9"...
## $ color  <chr> "East asian", "White", "Indigenous", "Mixed", "Mixed", "Mixe...
## $ cause  <chr> "Septicemia", "Hearth attack", "Others", "Cardiogenic shock"...
## $ total  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, ...
## $ day    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ month  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ year   <dbl> 2020, 2019, 2019, 2019, 2020, 2020, 2020, 2020, 2020, 2020, ...
## $ region <chr> "Norte", "Norte", "Norte", "Norte", "Norte", "Norte", "Norte...
```

```
# Verifico as primeiras linhas
head(data)
```

```
##         date state gender     age      color             cause total day month
## 1 2020-01-01    AC      F 60 - 69 East asian        Septicemia     1   1     1
## 2 2019-01-01    AC      F 80 - 89      White     Hearth attack     1   1     1
## 3 2019-01-01    AC      F 30 - 39 Indigenous            Others     1   1     1
## 4 2019-01-01    AC      F 70 - 79      Mixed Cardiogenic shock     1   1     1
## 5 2020-01-01    AC      F 70 - 79      Mixed         Pneumonia     1   1     1
## 6 2020-01-01    AC      F     < 9      Mixed         Pneumonia     1   1     1
##   year region
## 1 2020  Norte
## 2 2019  Norte
## 3 2019  Norte
## 4 2019  Norte
## 5 2020  Norte
## 6 2020  Norte
```

```
# Faço um resumo dos dados
summary(data)
```

```
##       date               state              gender              age
##  Min.   :2019-01-01   Length:1098241     Length:1098241     Length:1098241
##  1st Qu.:2019-06-17   Class :character   Class :character   Class :character
##  Median :2019-11-29   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2019-11-22
##  3rd Qu.:2020-05-06
##  Max.   :2020-09-15
##     color              cause               total             day
##  Length:1098241     Length:1098241     Min.   : 1.000   Min.   : 1.00
##  Class :character   Class :character   1st Qu.: 1.000   1st Qu.: 8.00
##  Mode  :character   Mode  :character   Median : 1.000   Median :15.00
##                                        Mean   : 1.872   Mean   :15.58
##                                        3rd Qu.: 2.000   3rd Qu.:23.00
##                                        Max.   :43.000   Max.   :31.00
##      month             year         region
##  Min.   : 1.000   Min.   :2019   Length:1098241
##  1st Qu.: 3.000   1st Qu.:2019   Class :character
##  Median : 6.000   Median :2019   Mode  :character
##  Mean   : 5.807   Mean   :2019
##  3rd Qu.: 8.000   3rd Qu.:2020
##  Max.   :12.000   Max.   :2020
```

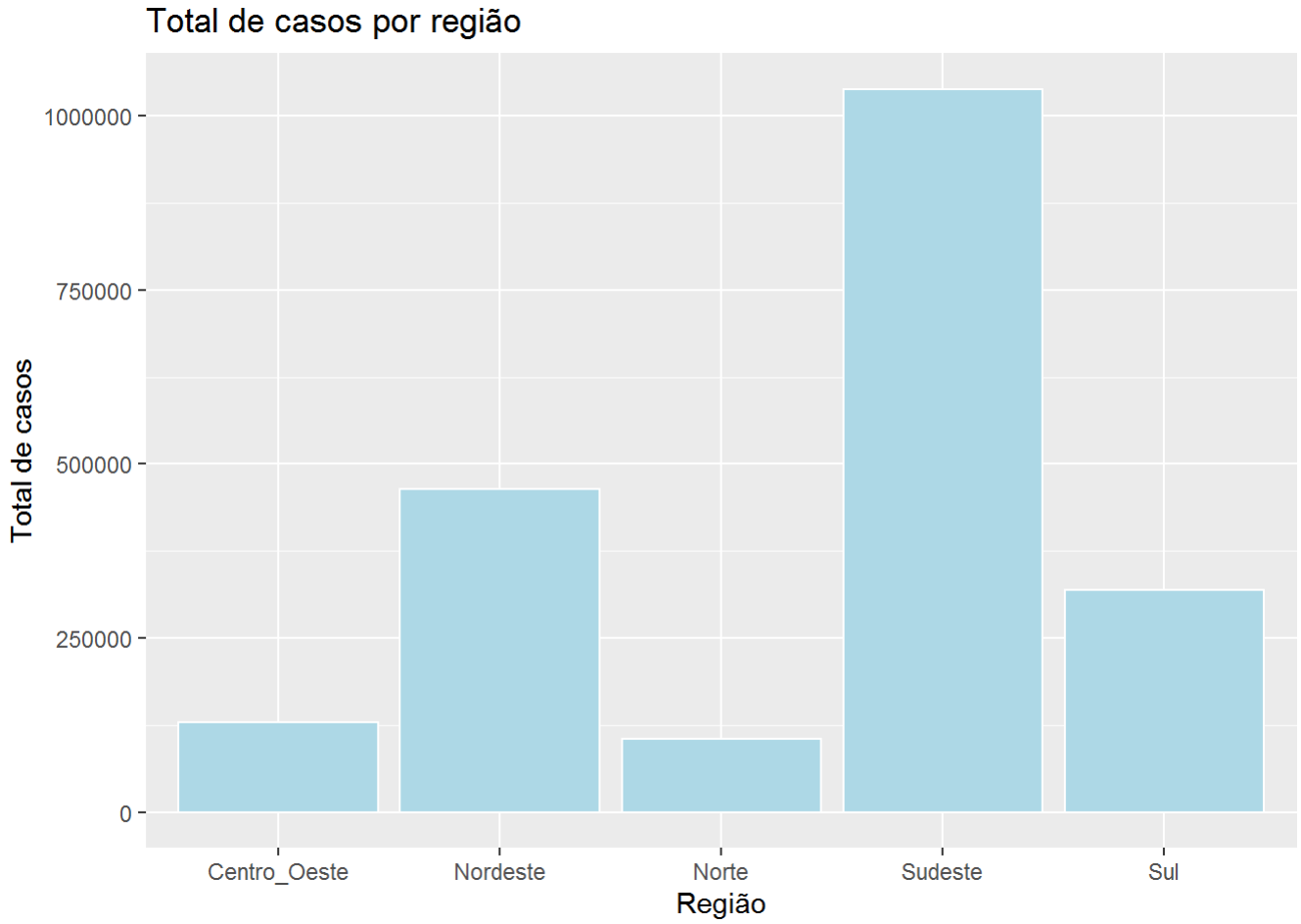
```
# Verifico se existe valores nulos nos dados
sum(is.na(data))
```

```
## [1] 0
```

Analise gráfica

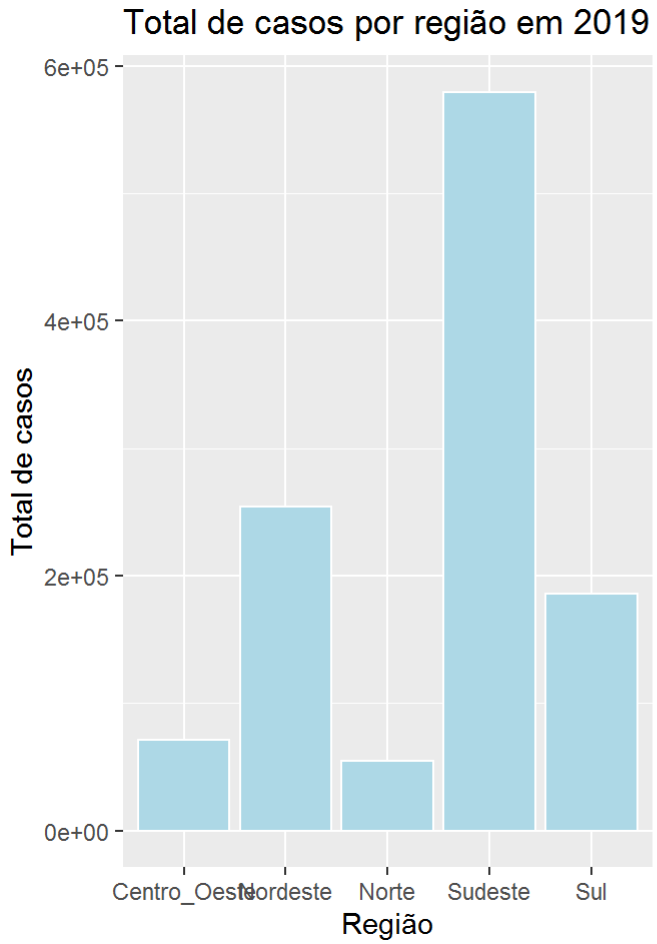
Análise mais ampla pegando as informações por região.

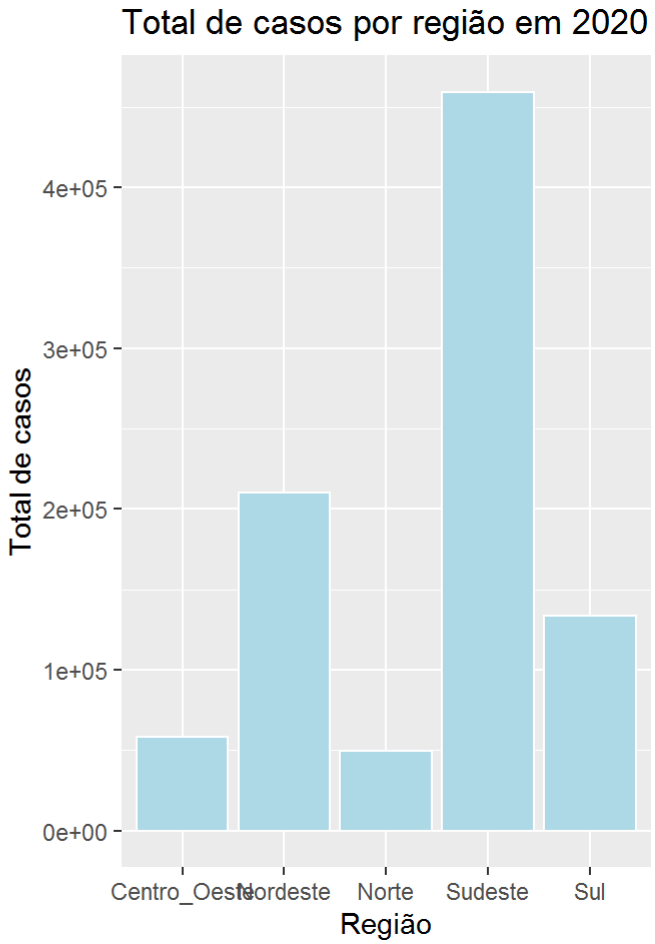
```
data %>%
  ddply(.(region),
  summarize,
  Total = sum(total))%>%
  ggplot(aes(x = region, y = Total))+
  geom_bar(stat = "identity",color = "white", fill = "lightblue")+
  ggtitle('Total de casos por região') + xlab('Região') + ylab('Total de casos')
```



A região sudeste tem o maior número de casos, uma vez que ela possui quase metade do total populacional do brasil.

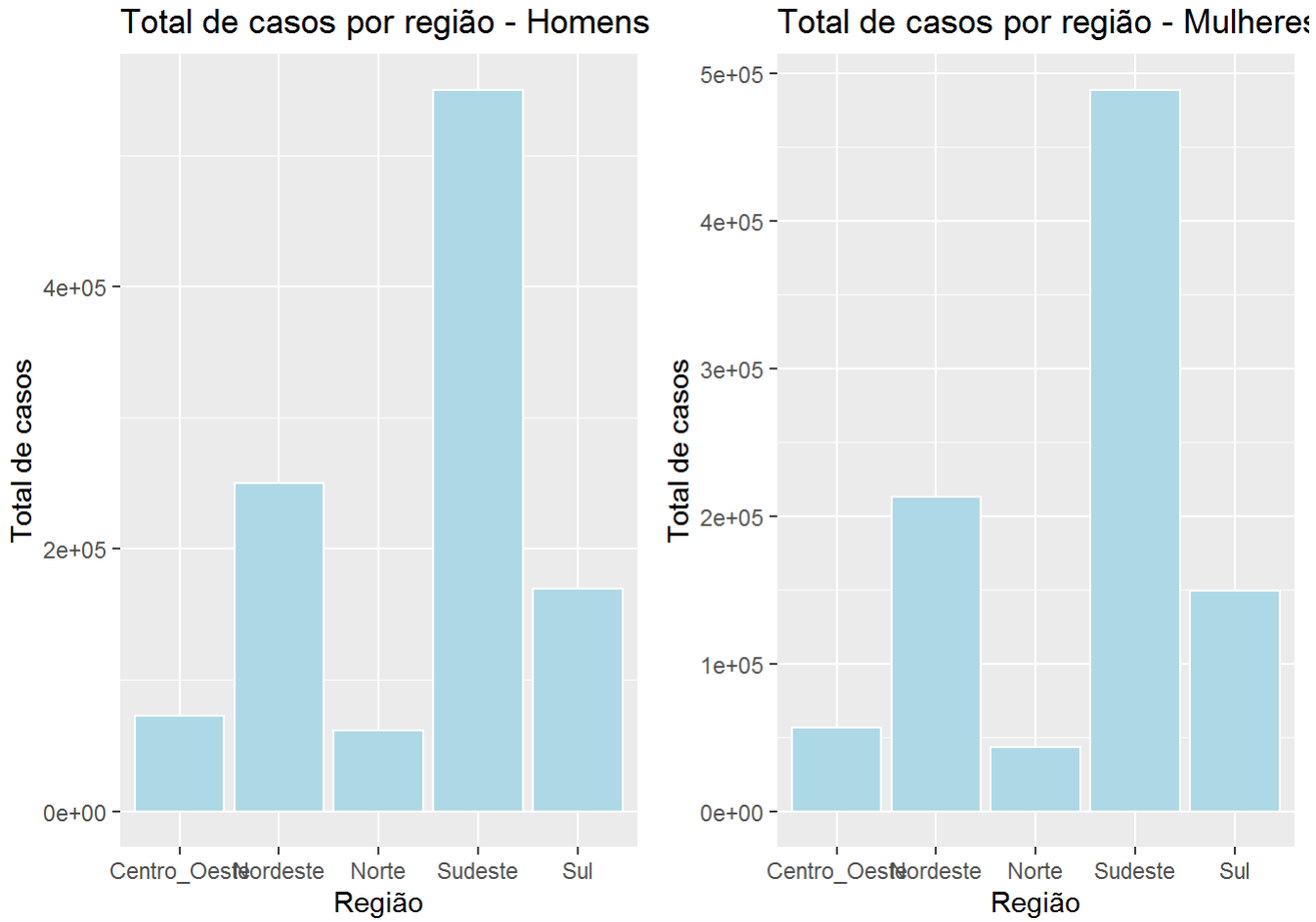
```
pl1 <- data %>%
  filter(year == '2019')%>%
  ddply(.(region),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = region, y = Total))+
  geom_bar(stat = "identity",color = "white", fill = "lightblue")+
  ggtitle('Total de casos por região em 2019') + xlab('Região') + ylab('Total de casos')

pl2 <- data %>%
  filter(year == '2020')%>%
  ddply(.(region),
        summarize,
        Total = sum(total))%>%
    ggplot(aes(x = region, y = Total))+
    geom_bar(stat = "identity",color = "white", fill = "lightblue")+
    ggtitle('Total de casos por região em 2020') + xlab('Região') + ylab('Total de casos')

grid.arrange(pl1,pl2, nrow=1,ncol=2)
```





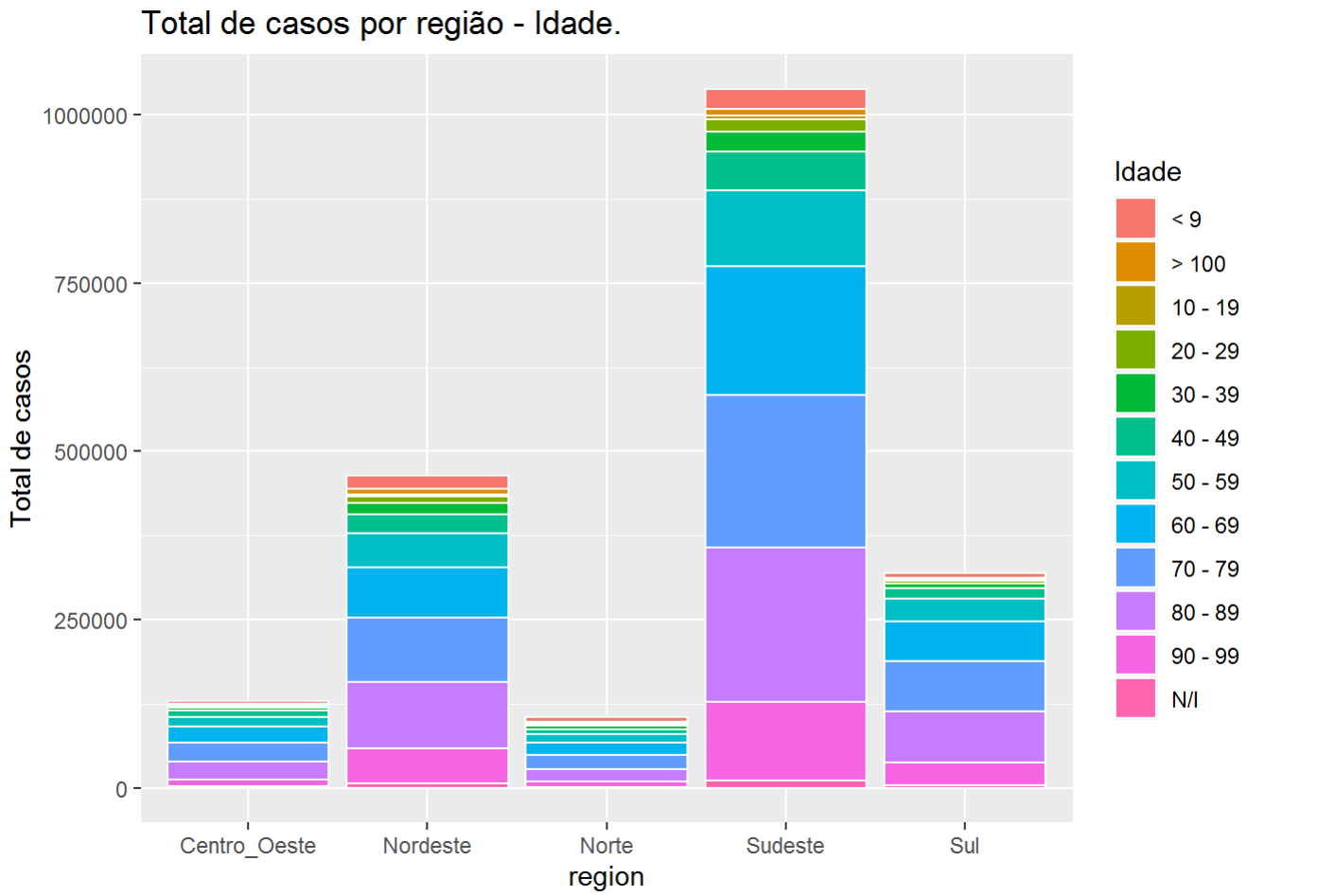
Realizando uma subdivisão dos casos entre os anos 2019 e 2020 que são os anos reportados nesses dados, podemos constatar que a quantidade de casos se mantém praticamente em um padrão constante.

```
pl1 <- data %>%
  filter(gender == 'M')%>%
  ddply(.(region),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = region, y = Total))+
  geom_bar(stat = "identity",color = "white", fill = "lightblue")+
  ggtitle('Total de casos por região - Homens') + xlab('Região') + ylab('Total de casos')

pl2 <- data %>%
  filter(gender == 'F')%>%
  ddply(.(region),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = region, y = Total))+
  geom_bar(stat = "identity",color = "white", fill = "lightblue")+
  ggtitle('Total de casos por região - Mulheres') + xlab('Região') + ylab('Total de casos')

grid.arrange(pl1,pl2, nrow=1,ncol=2)
```



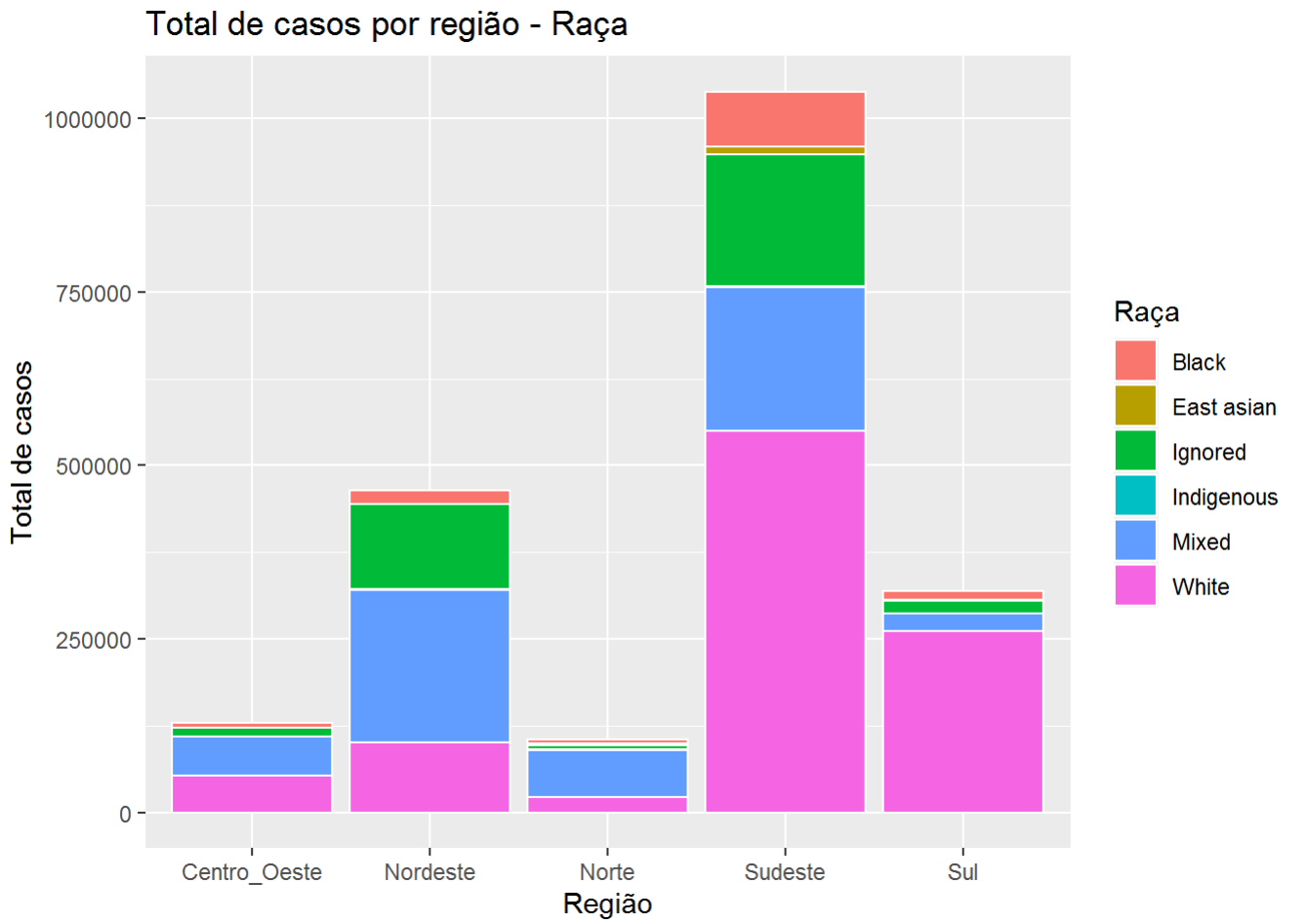
Realizando uma subdivisão dos casos entre o sexo, podemos constatar também que existe um padrão entre os dois sexos, porem um pouco mais de casos do sexo feminino em relação ao masculino em cada região.

```
data %>%
 ddply(.(region,age),
       summarize,
       Total = sum(total))%>%
 ggplot(aes(x = region, y = Total,fill = age))+
 geom_bar(stat = "identity",color = "white")+
 labs(title ='Total de casos por região - Idade.', xlab = 'Região', y = 'Total de casos', fill
= 'Idade')
```



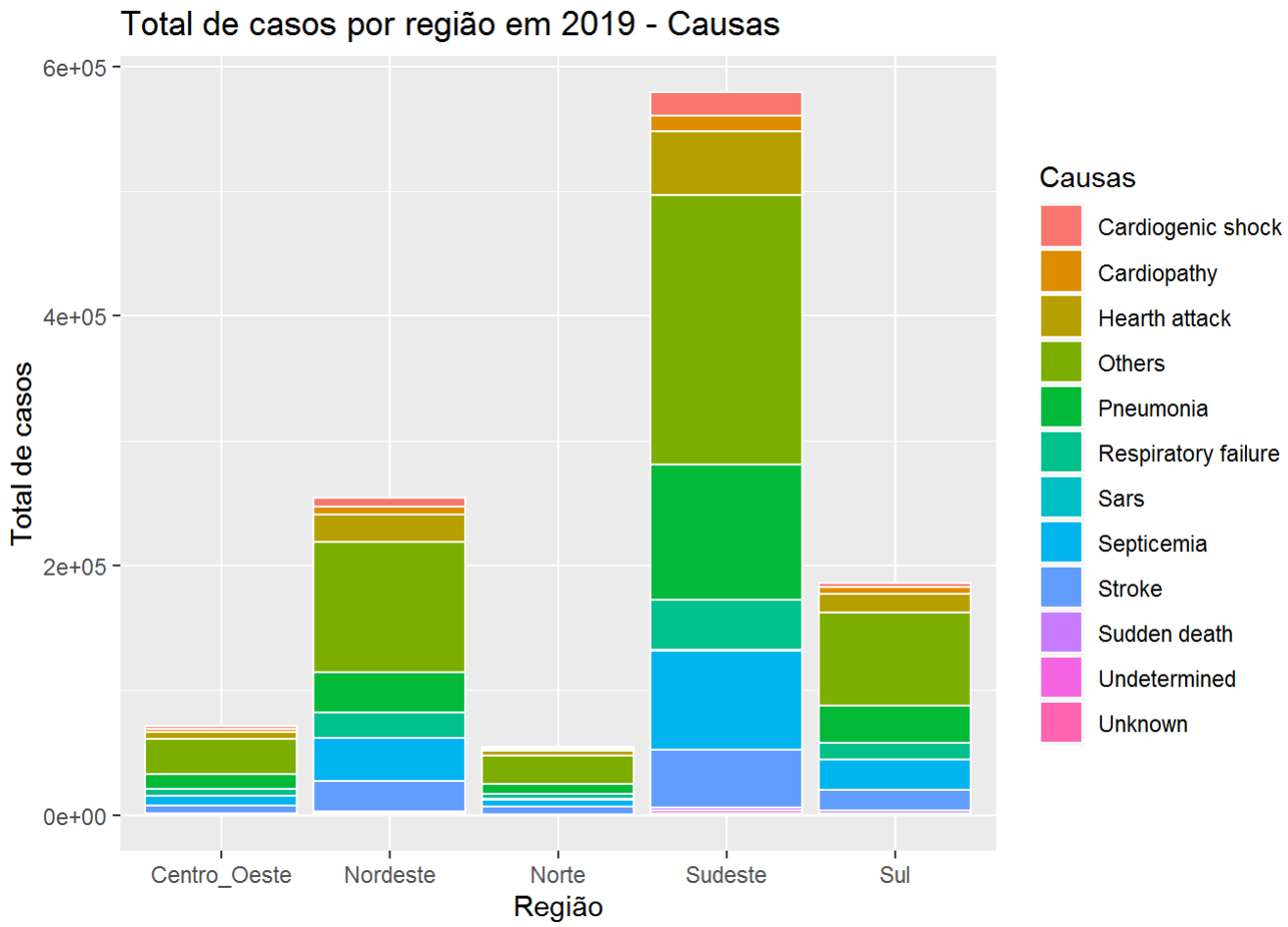
Neste gráfico podemos constatar que há um padrão em todas as regiões onde o número de mortes ocorrem em maior quantidade a partir dos 60 anos.

```
data %>%
   ddply(.(region,color),
         summarize,
         Total = sum(total))%>%
   ggplot(aes(x = region, y = Total,fill = color))+
   geom_bar(stat = "identity",color = "white")+
   labs(title = 'Total de casos por região - Raça',x = 'Região',y = 'Total de casos', fill = 'Ra
ça')
```



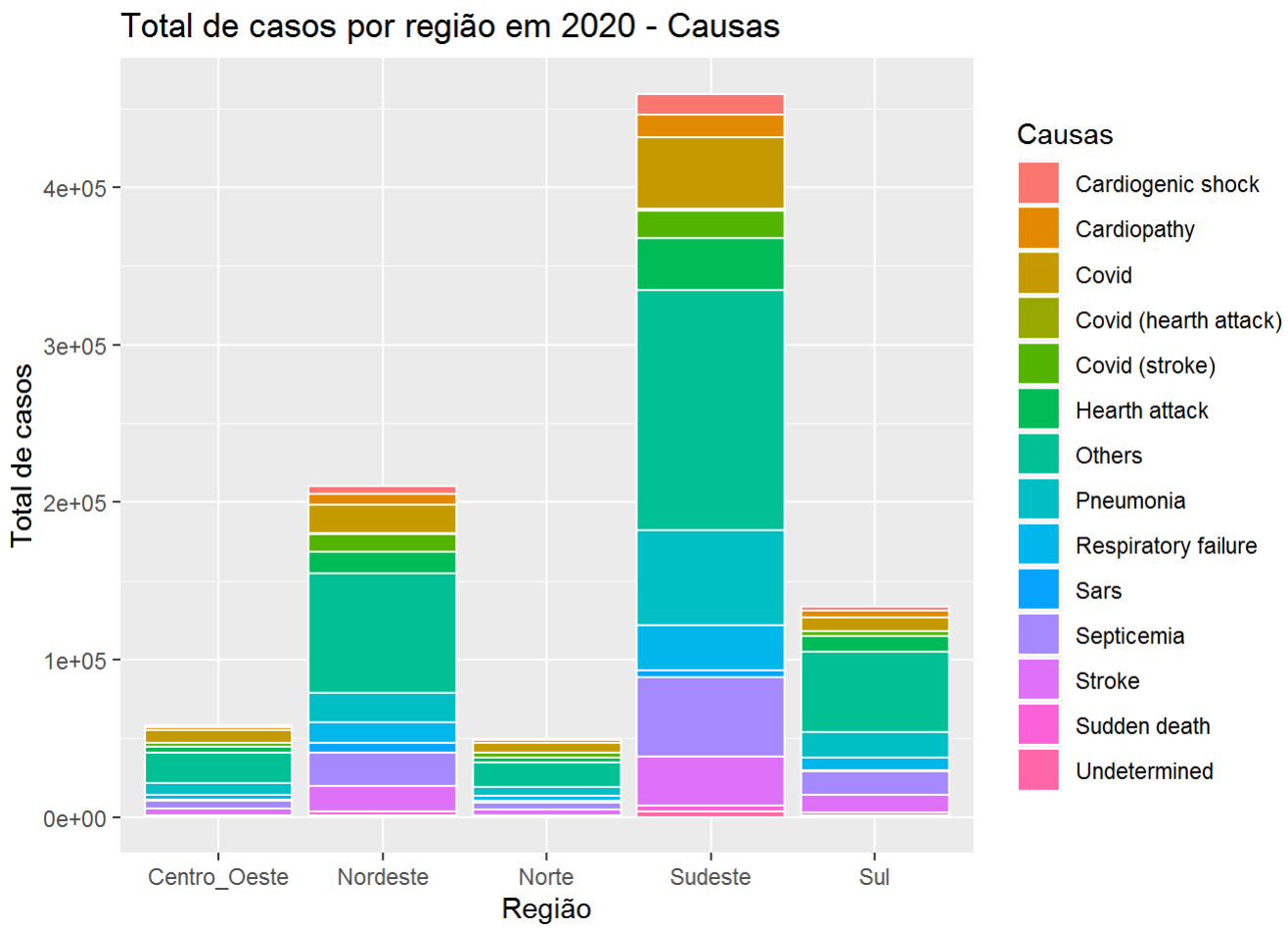
Nas regiões Sudeste e Sul, os maiores números de casos aparem em pessoal brancas, já nas demais aparecem em pessoas consideradas misturadas segundo a descrição do dataset.

```
data %>%
  filter(year == '2019')%>%
   ddply(.(region,cause),
         summarize,
         Total = sum(total))%>%
   ggplot(aes(x = region, y = Total,fill = cause))+
   geom_bar(stat = "identity",color = "white")+
   labs(title ='Total de casos por região em 2019 - Causas', x = 'Região',y = 'Total de casos',
 fill = 'Causas')
```



Existe um padrão dos casos por região em 2019, tirando a opção de outros que tem o maior número de casos, porém sem uma causa exata divulgada no dadaset, dos nomes informados das causas de mortes o ataque cardíaco e pneumonia são as que mais tem casos.

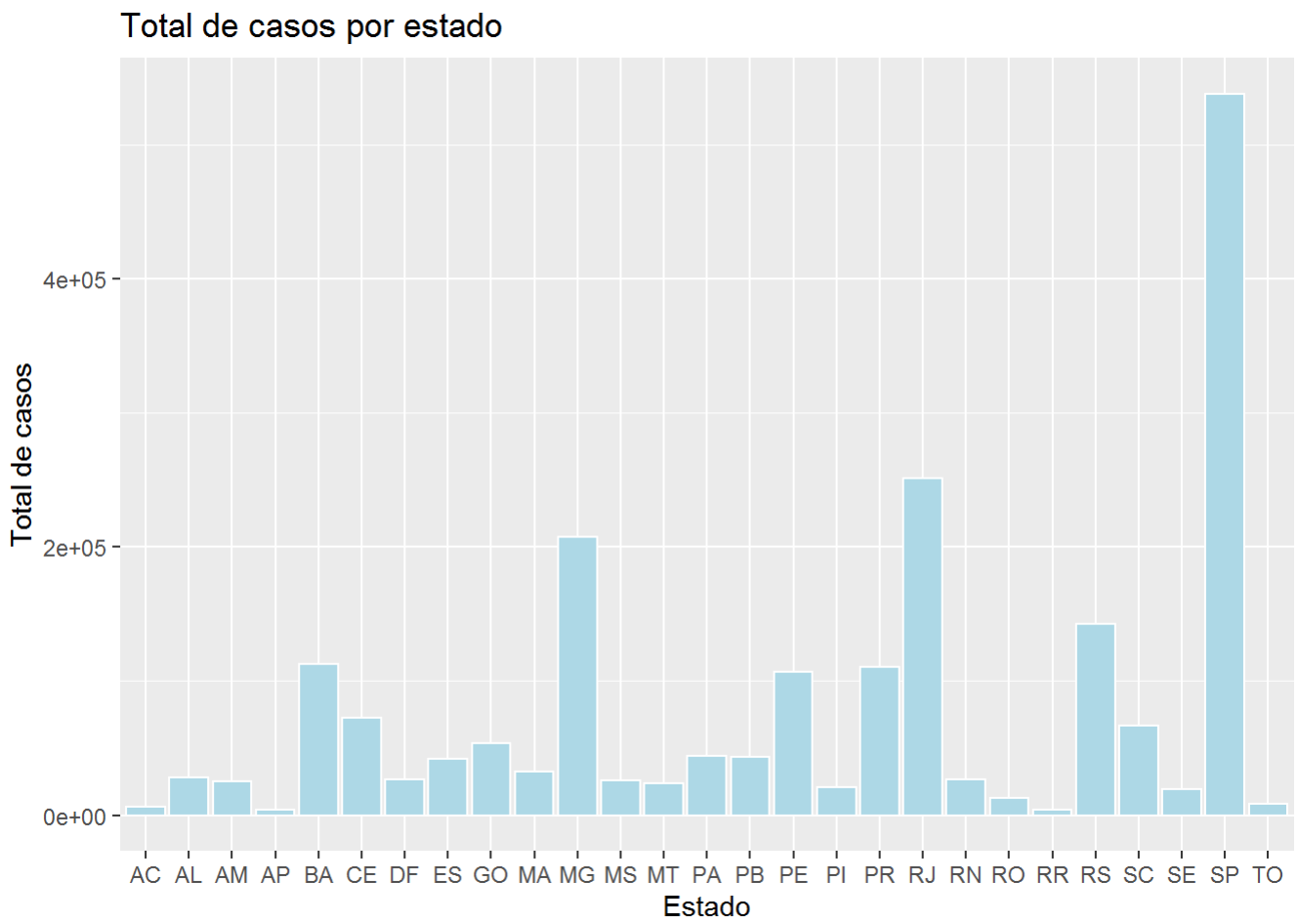
```
data %>%
  filter(year == '2020')%>%
  ddply(.(region,cause),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = region, y = Total,fill = cause))+
  geom_bar(stat = "identity",color = "white")+
  labs(title ='Total de casos por região em 2020 - Causas', x = 'Região',y = 'Total de casos', f
ill = 'Causas')
```



Segue o praticamente o mesmo padrão de 2019, porem com o surgimento da pandemia do Covid-19 no início de 2020, já podemos notar casos de mortes pela pandemia, porém ainda em uma escala pequena, isso se deve provavelmente pois a coleta dos dados para esse dataset foi feito no início da pandemia.

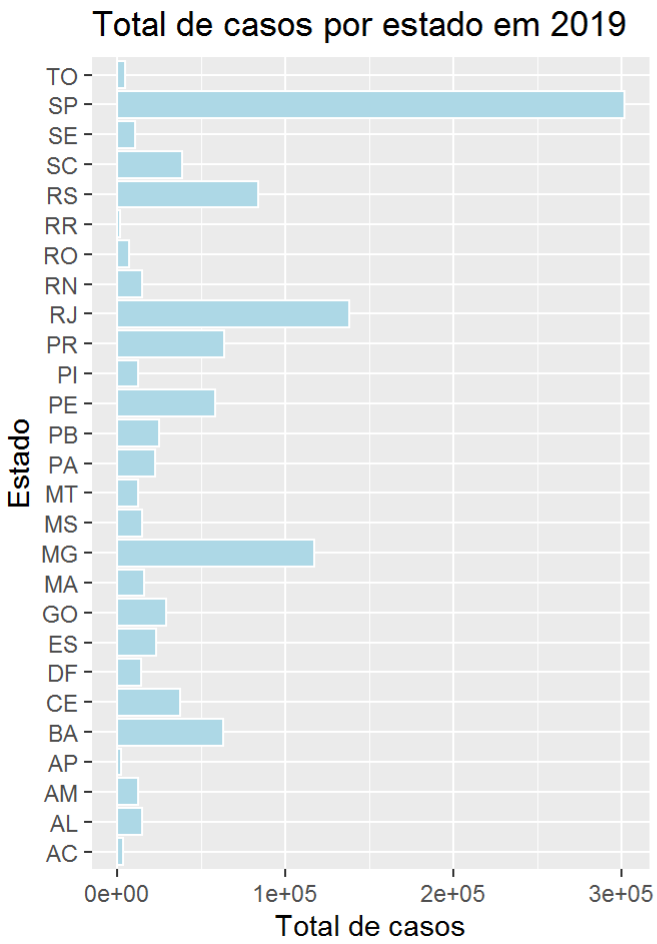
Agora uma análise mais especifica baseado por estado. Farei basicamente a mesma análise feita por região, mas agora segmentando por estado.

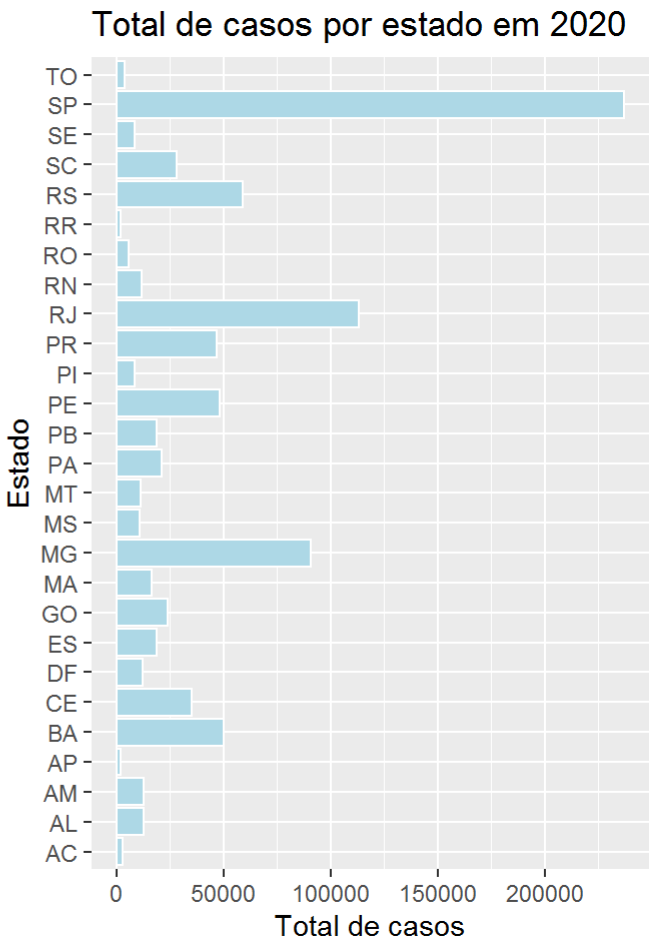
```
data %>%
  ddply(.(state),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = state, y = Total))+
  geom_bar(stat = "identity",color = "white", fill = "lightblue")+
  ggtitle('Total de casos por estado') + xlab('Estado') + ylab('Total de casos')
```



Como vimos que a região sudeste é a que tem maior número de casos, temos os 3 estados dessa região em destaque em número de casos, São Paulo, Rio de Janeiro e Minas Gerais respectivamente.

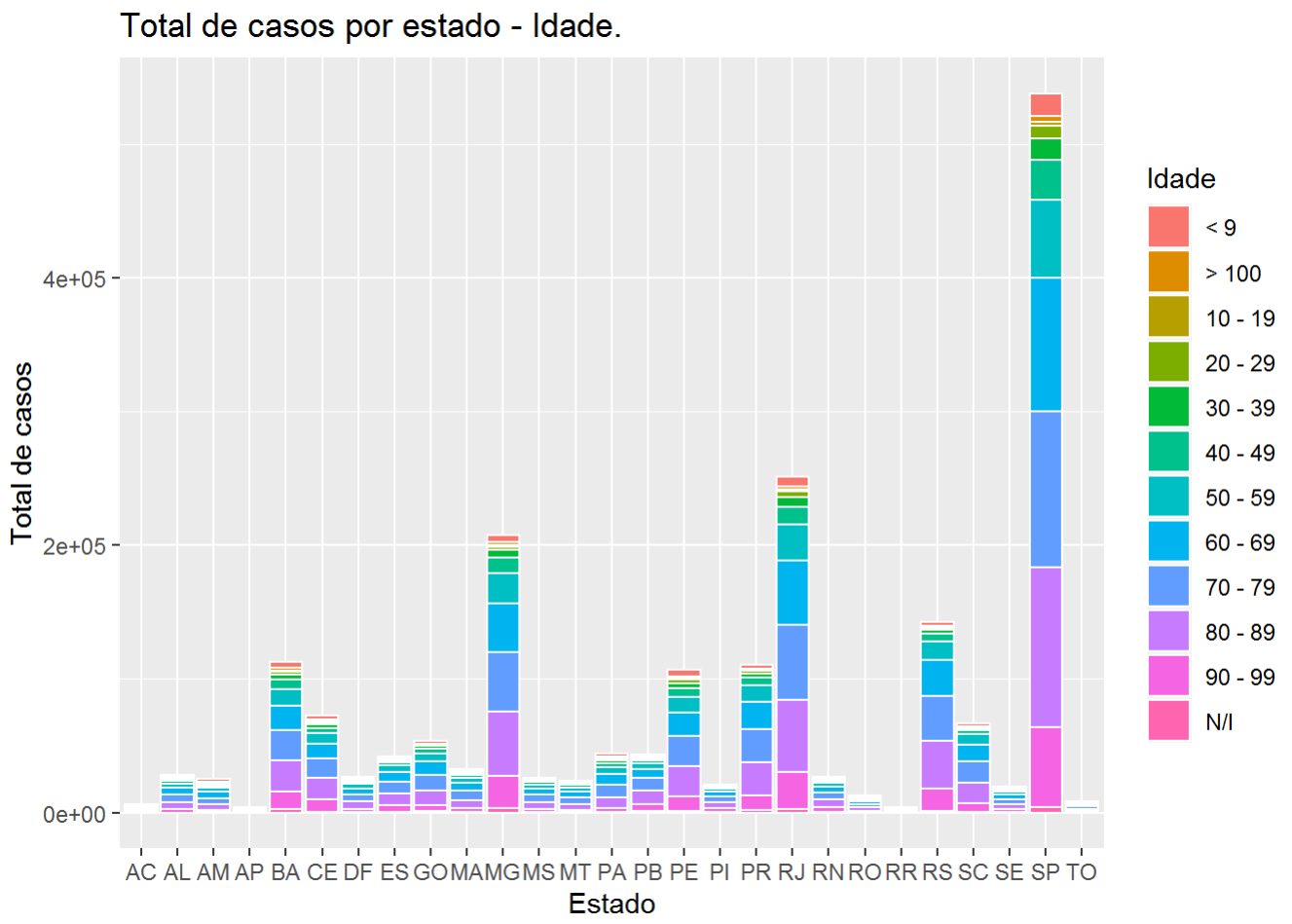
```
pl1 <- data %>%
  filter(year == '2019')%>%
  ddply(.(state),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = Total, y = state))+
  geom_bar(stat = "identity",color = "white", fill = "lightblue")+
  ggtitle('Total de casos por estado em 2019') + xlab('Total de casos') + ylab('Estado')

pl2 <- data %>%
  filter(year == '2020')%>%
  ddply(.(state),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = Total, y = state))+
  geom_bar(stat = "identity",color = "white", fill = "lightblue")+
  ggtitle('Total de casos por estado em 2020') + xlab('Total de casos') + ylab('Estado')

grid.arrange(pl1,pl2, nrow=1,ncol=2)
```





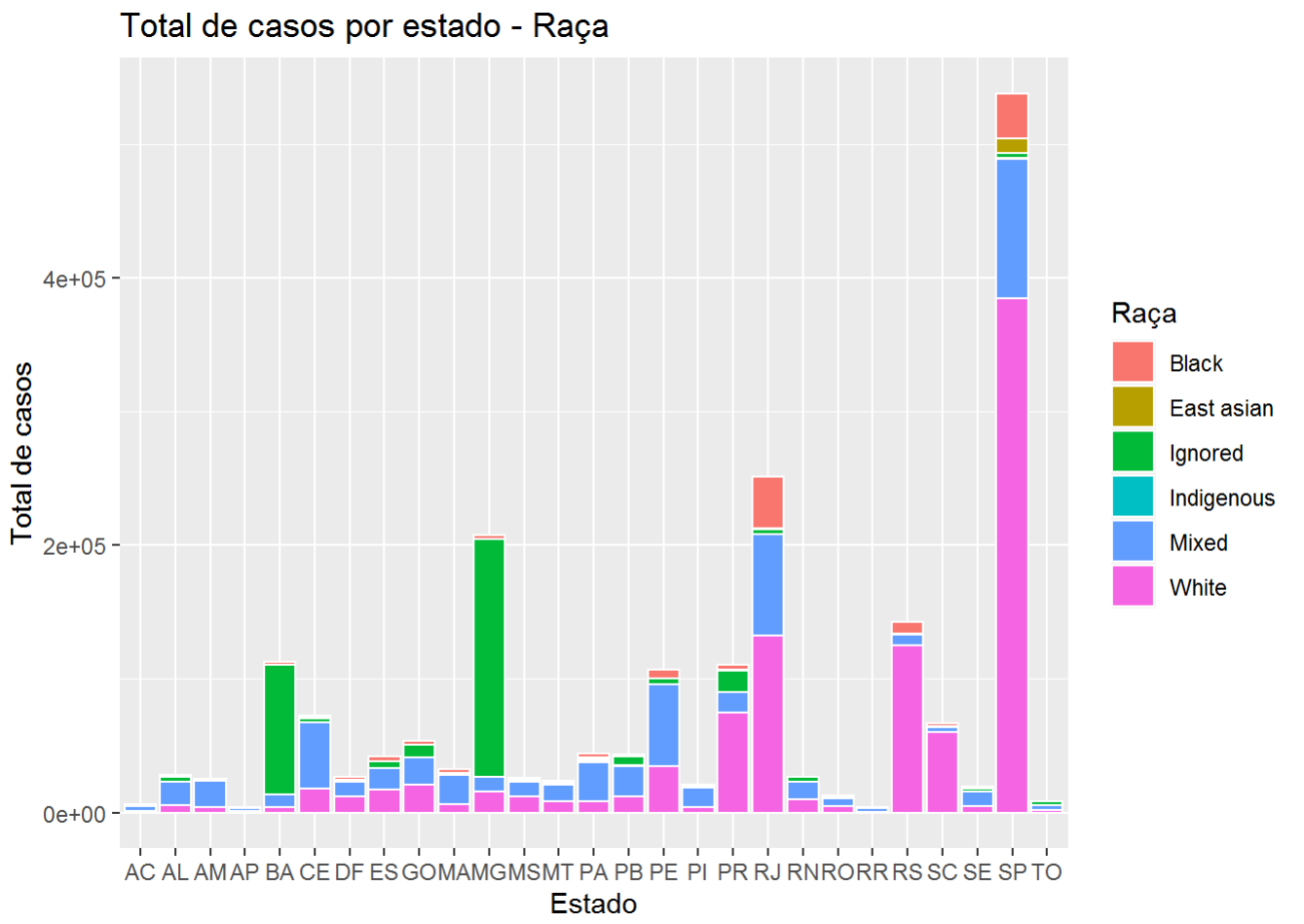
Existe um padrão entre o número de casso separando por ano, o ano de 2020 tem um pouco menos de casos para cada estado, isso se deve pelo fato da data de extração dos dados para a análise, temos um ano inteiro de 2019 e uma parte de 2020, como a diferença é pequena podemos considerar proporcionalmente que ao final do ano de 2020 os casos em cada estado podem ter superado os de 2019.

```
data %>%
  ddply(.(state,age),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = state, y = Total,fill = age))+
  geom_bar(stat = "identity",color = "white")+
  labs(title ='Total de casos por estado - Idade.', x = 'Estado', y = 'Total de casos', fill =
'Idade')
```



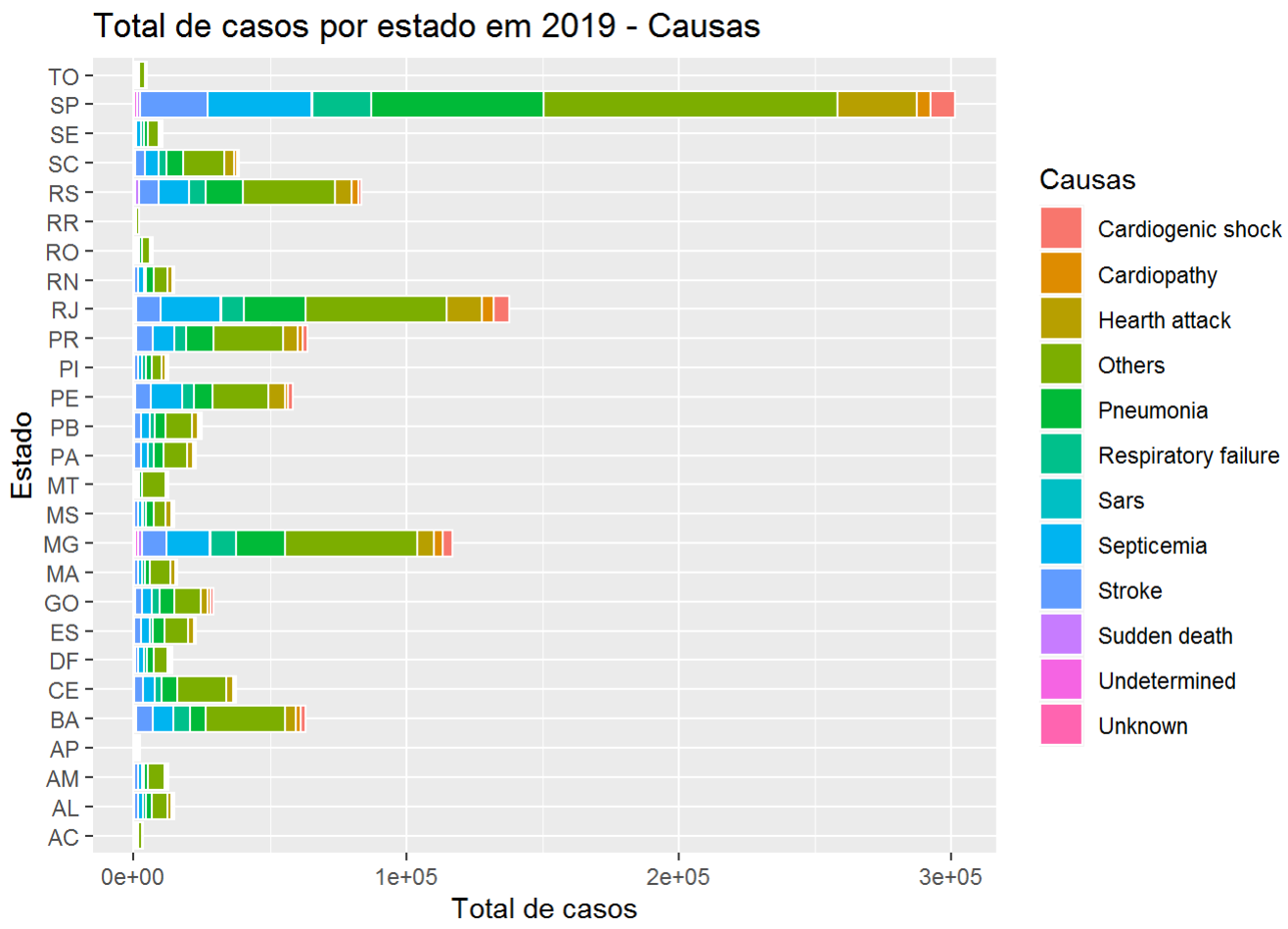
Assim como por região a taxa de mortalidade se mantem igual em todos os estados, onde os maiores casos de mortes por doença são a partir dos 50 anos.

```
data %>%
  ddply(.(state,color),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = state, y = Total,fill = color))+
  geom_bar(stat = "identity",color = "white")+
  labs(title = 'Total de casos por estado - Raça',x = 'Estado',y = 'Total de casos', fill = 'Raç
a')
```



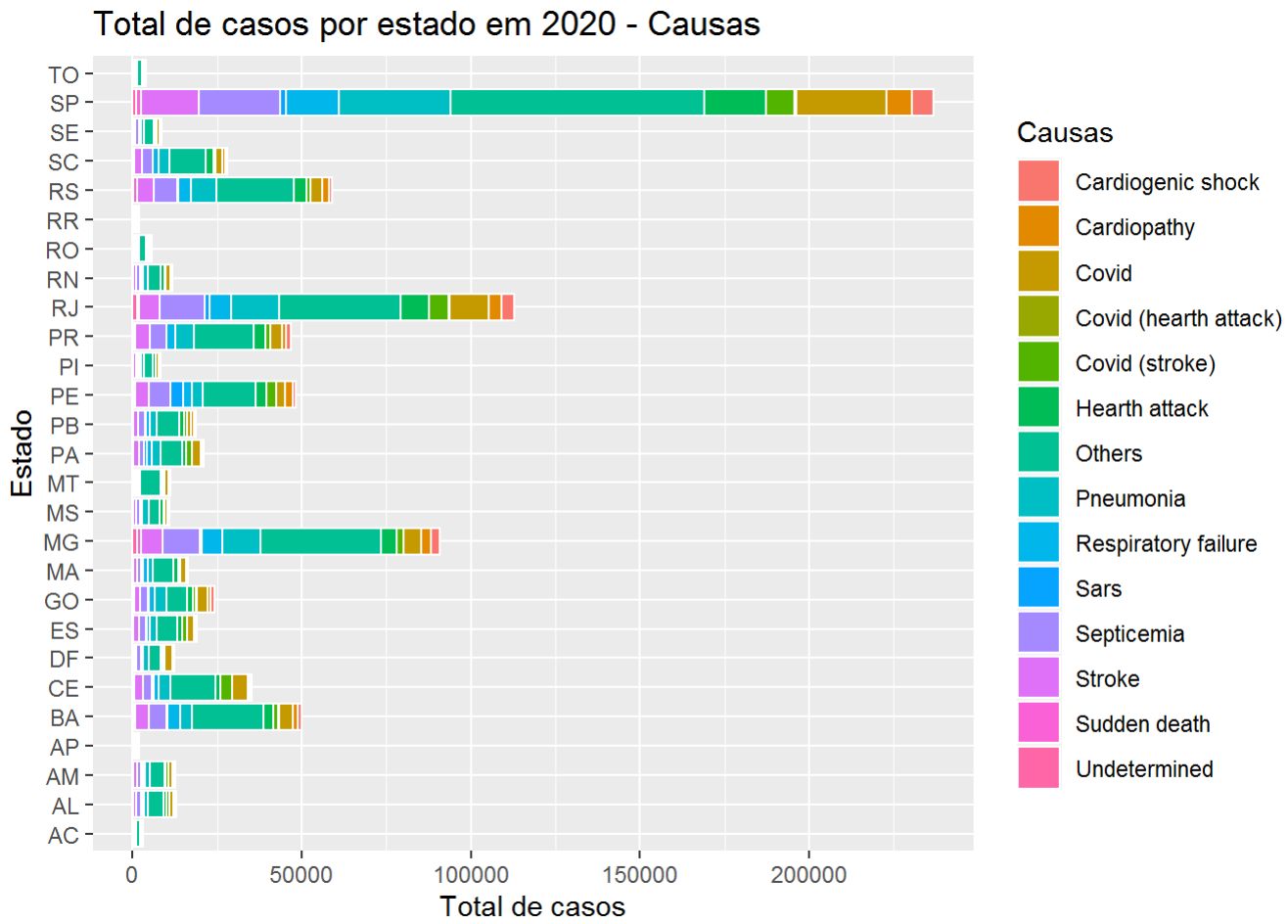
Nesse gráfico podemos ver com mais detalhes o que foi visto no gráfico por região, onde estados como SP, SC, RS, RJ e PR os casos em grandes partes são de pessoas brancas, MG e BA constam um grande número onde esse dado foi ignorado,

```
data %>%
  filter(year == '2019')%>%
  ddply(.(state,cause),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = Total, y = state,fill = cause))+
  geom_bar(stat = "identity",color = "white")+
  labs(title ='Total de casos por estado em 2019 - Causas', x = 'Total de casos',y = 'Estado', f
ill = 'Causas')
```



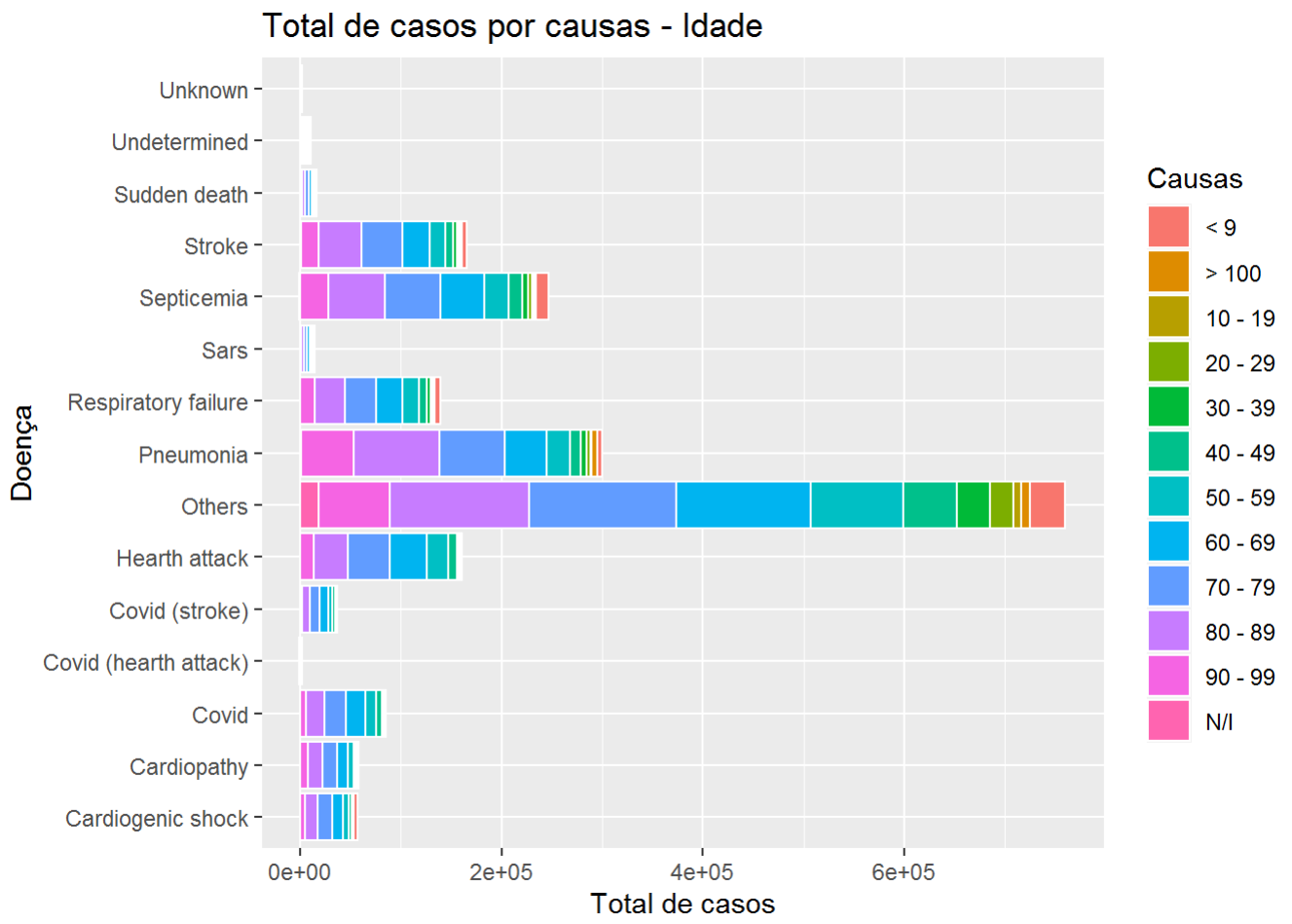
A análise em 2019 por estado segue o mesmo padrão das regiões, com um destaque agora nos estados de SP, RJ e MG, onde casos de septicemia aparecem também em destaque além do ataque cardíaco e pneumonia.

```
data %>%
  filter(year == '2020')%>%
  ddply(.(state,cause),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = Total, y = state,fill = cause))+
  geom_bar(stat = "identity",color = "white")+
  labs(title ='Total de casos por estado em 2020 - Causas', x = 'Total de casos',y = 'Estado', f
ill = 'Causas')
```



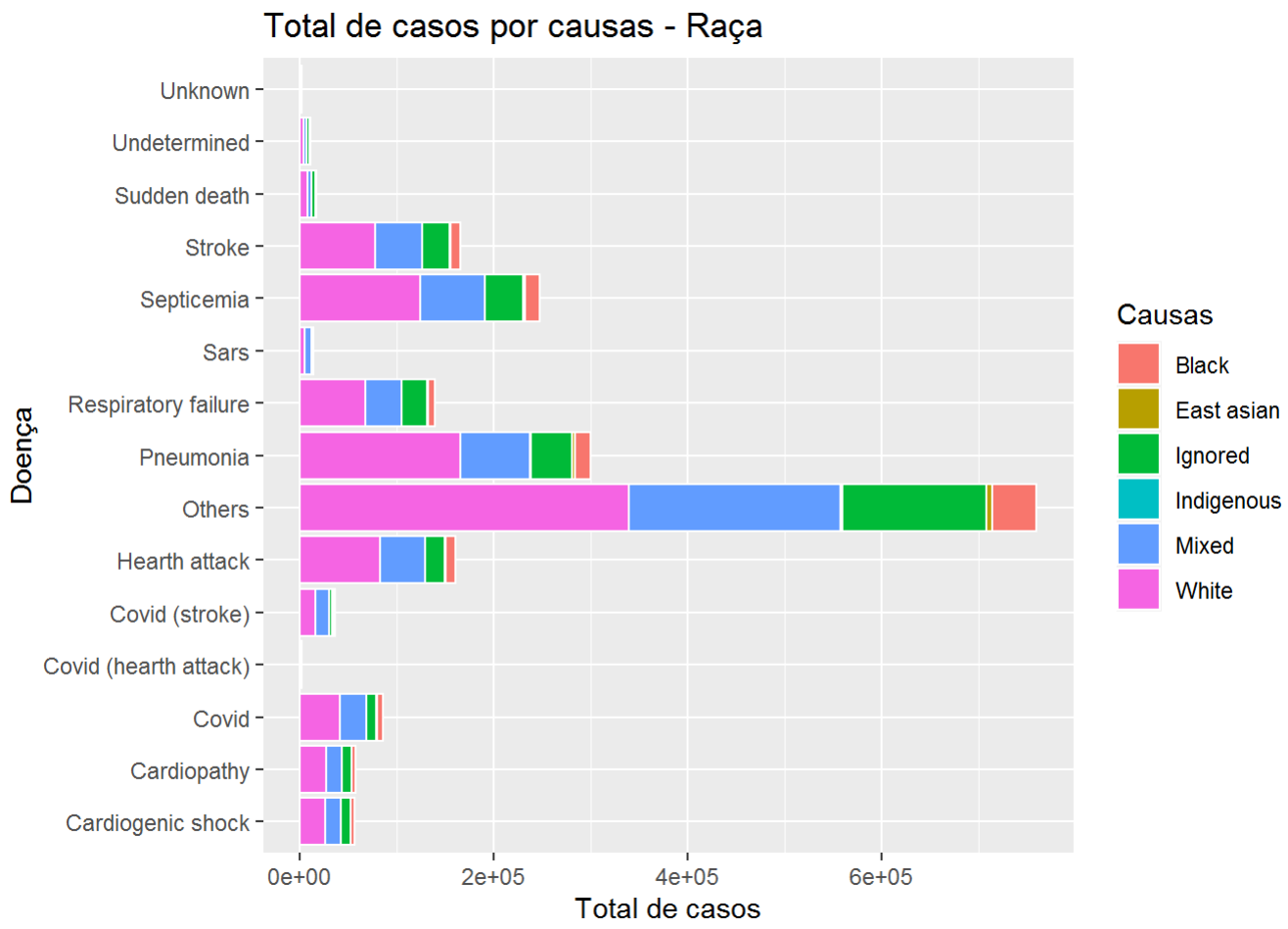
A análise em 2020 já podemos ver alguns casos de covid em alguns estados em destaque os mais populosos como SP e RJ assim como um aumento nos casos de septicemia.

```
data %>%
  ddply(.(cause,age),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = Total, y = cause,fill = age))+
  geom_bar(stat = "identity",color = "white")+
  labs(title ='Total de casos por causas - Idade', x = 'Total de casos',y = 'Doença', fill = 'Ca
usas')
```



Doenças cardíacas, vasculares, respiratórias e covid por exemplo, são doenças com mais mortalidade a partida da faixa de 60 anos, bem relevante uma vez que com o passar dos anos e com a idade o corpo fica mais frágil e tais doenças e sua recuperação são mais complicadas devido à idade

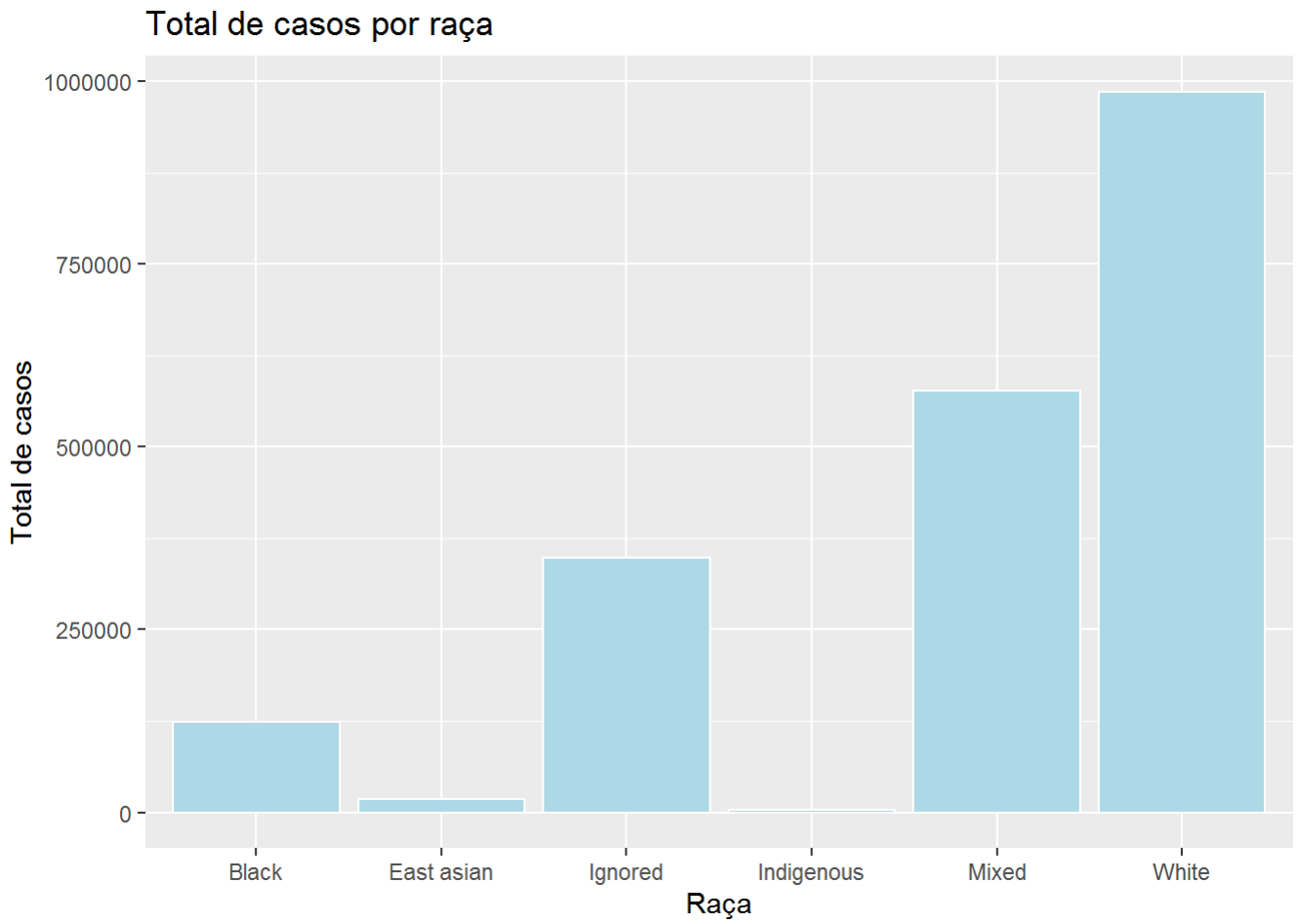
```
data %>%
  ddply(.(cause,color),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = Total, y = cause,fill = color))+
  geom_bar(stat = "identity",color = "white")+
  labs(title ='Total de casos por causas - Raça', x = 'Total de casos',y = 'Doença', fill = 'Cau
sas')
```



Seguem o mesmo padrão da análise por região e estado, maiores casos de doenças em pessoas brancas e consideras misturadas.

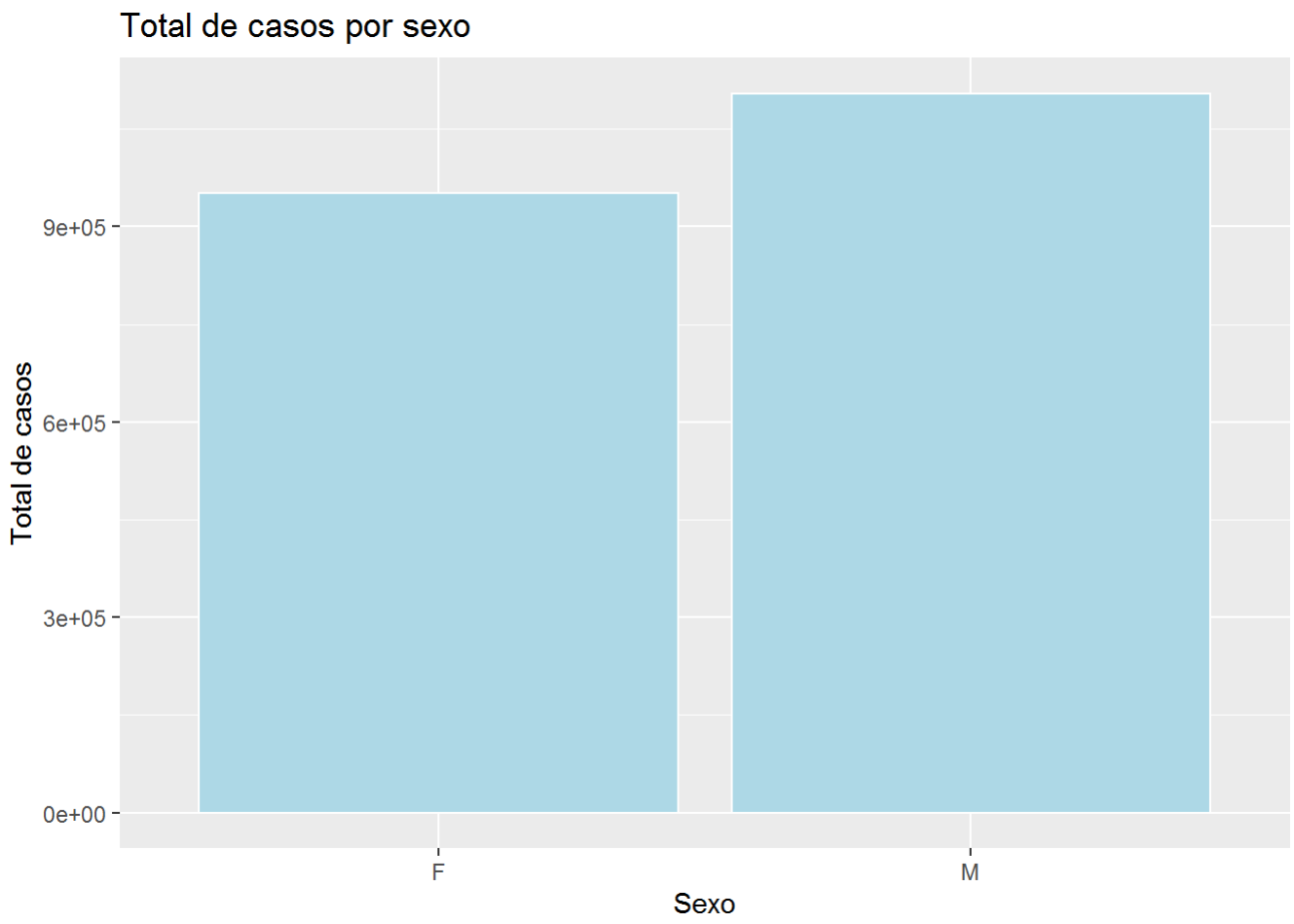
Analise individual das demais variáveis

```
data %>%
  ddply(.(color),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = color, y = Total))+
  geom_bar(stat = "identity",color = "white", fill = "lightblue")+
  ggtitle('Total de casos por raça') + xlab('Raça') + ylab('Total de casos')
```



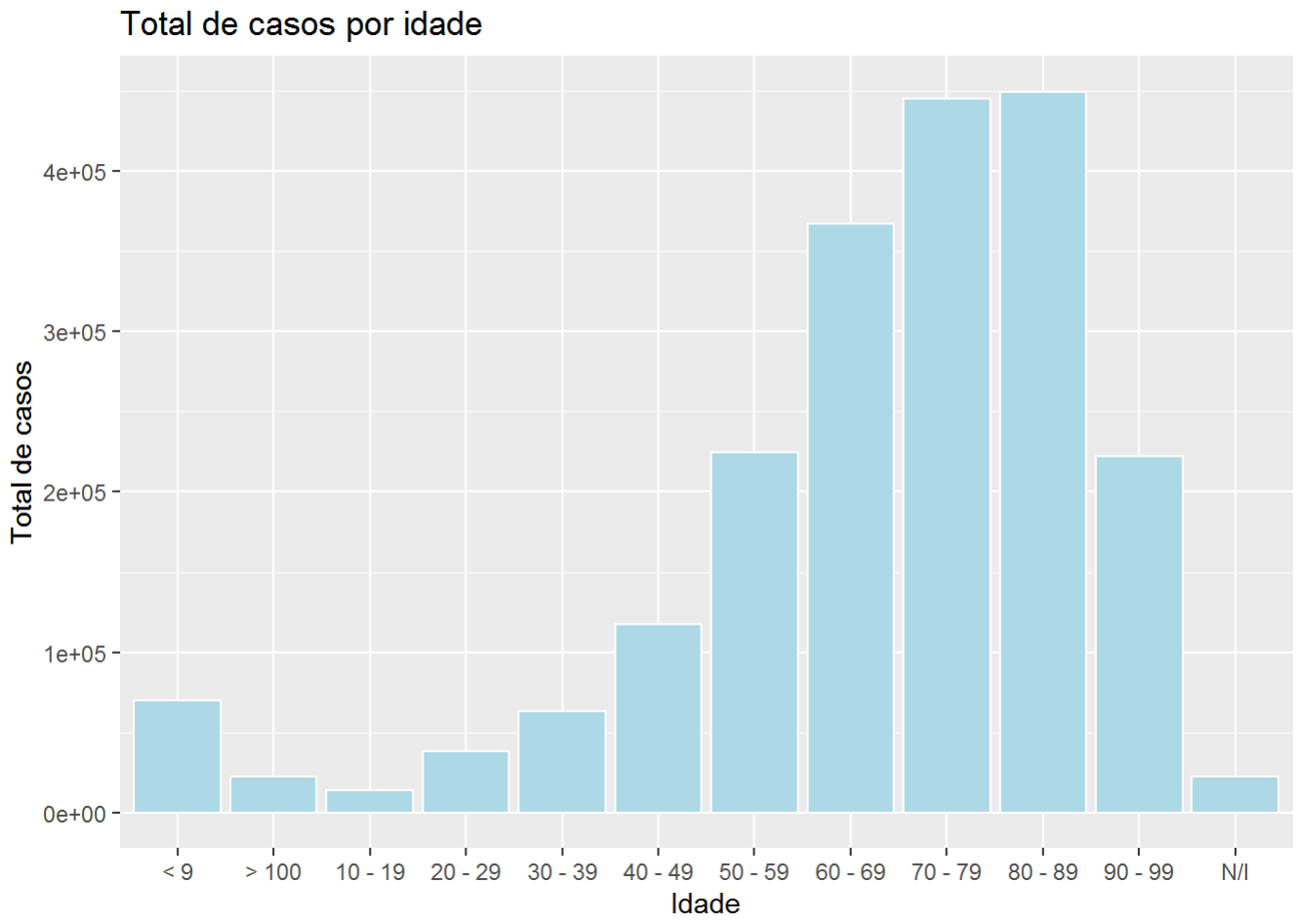
Total geral de casos, temos o que já vimos anteriormente, brancos e misturados com maiores ocorrências.

```
data %>%
  ddply(.(gender),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = gender, y = Total))+
  geom_bar(stat = "identity",color = "white", fill = "lightblue")+
  ggtitle('Total de casos por sexo') + xlab('Sexo') + ylab('Total de casos')
```



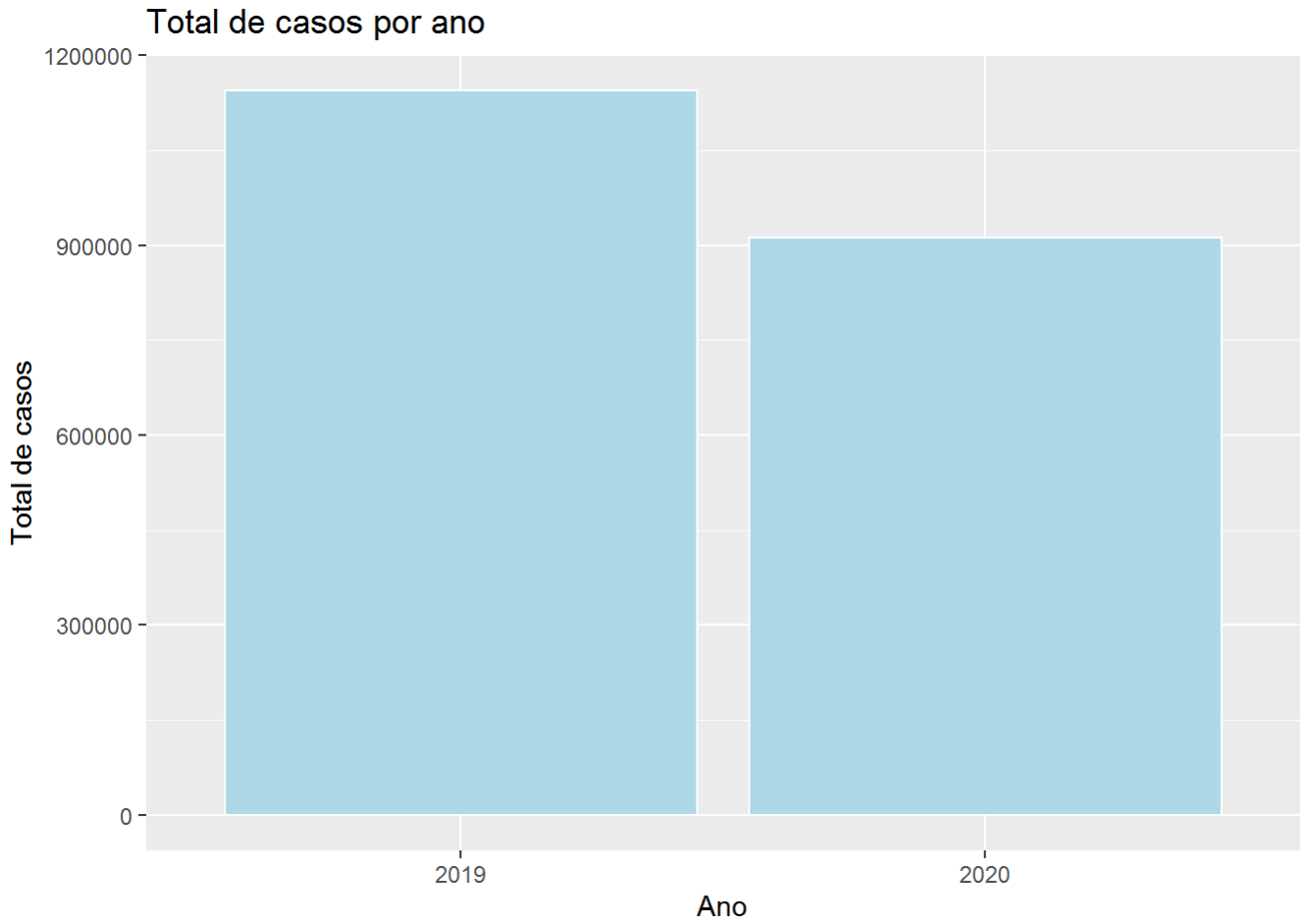
Bem balanceado o total de casos dividido por sexo.

```
data %>%
  ddply(.(age),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = age, y = Total))+
  geom_bar(stat = "identity",color = "white", fill = "lightblue")+
  ggtitle('Total de casos por idade') + xlab('Idade') + ylab('Total de casos')
```



Como já vimos entre 50 a 100 estão entre os maiores casos, com um pico maior entre 70 a 89.

```
data %>%
  ddply(.(year),
        summarize,
        Total = sum(total))%>%
  ggplot(aes(x = as.factor(year), y = Total))+
  geom_bar(stat = "identity",color = "white", fill = "lightblue")+
  ggtitle('Total de casos por ano') + xlab('Ano') + ylab('Total de casos')
```



Como temos dados do ano todo de 2019 e apenas metade do de 2020, normal ter mais casos em 2019, mas podemos constatar mais uma vez que a probabilidade de 2020 passar 2019 e bem grande.

```

# Considerações Finais

Com esses gráficos podemos ter insights como explicado em cada imagem, a respeito de tendência de probabilidades de ocorrência dessas doenças, em vários aspectos, região, estado, raça, idade, e etc, muito possivelmente pegando um dado atualizado com o ano de 2020 vamos ter um grande aumento nos casos de covid.

## Obrigado! Entre em contato comigo acessando meu portifolio (https://campos1989.github.io/ (https://campos1989.github.io/)) no menu contato!
```