ANÁLISE DOS DADOS COM R

Visualização dos Dados e EDA

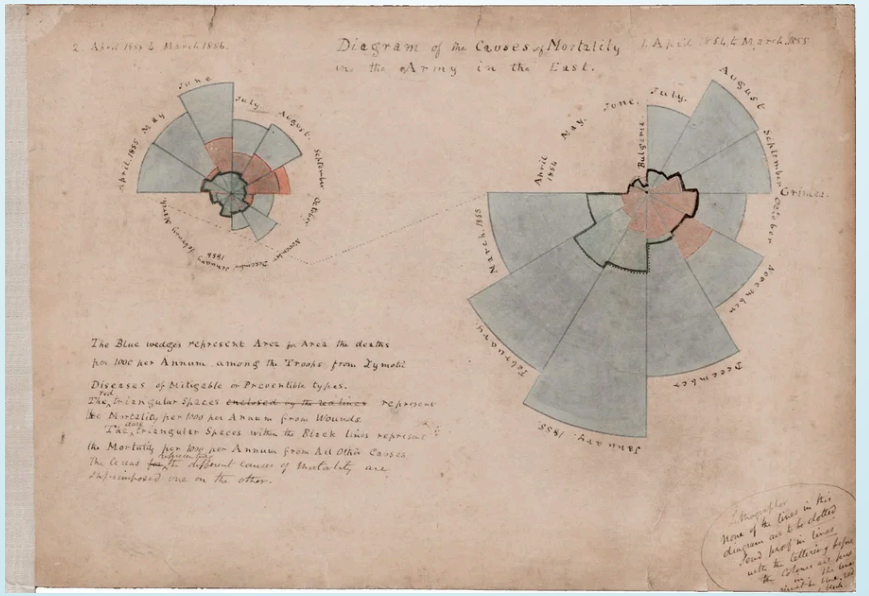
James R. Hunter, PhD
Retrovirologia, EPM, UNIFESP

2024-11-19

VISUALIZAÇÃO DOS DADOS

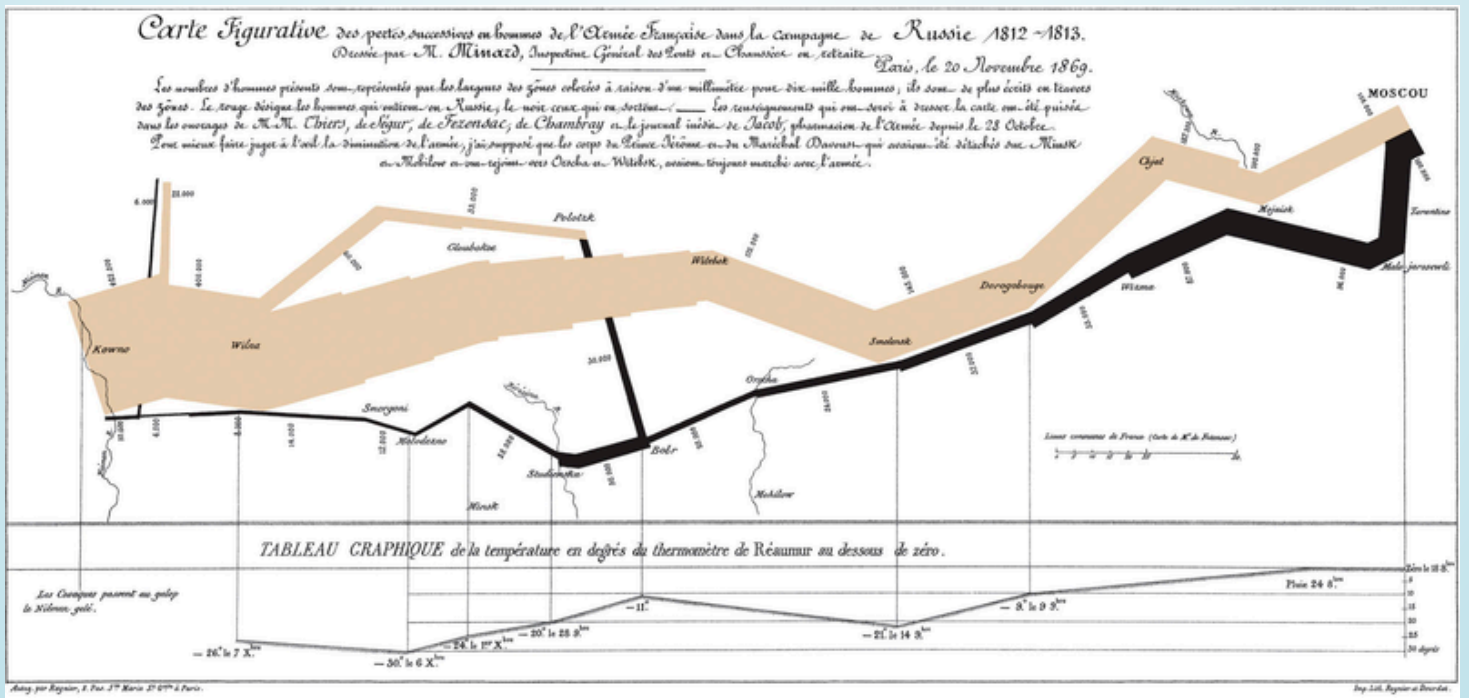
FLORENCE NIGHTINGALE - DIAGRAMMA DE ROSA

- Demonstrou claramente as causas de mortalidade entre soldados ingleses na guerra de Crimeia
- Gráfico hoje conhecido como "*polar axis histogram*"



CHARLES MINARD - DIAGRAMA DE FLUXO DO EXÉRCITO DE NAPOLEÃO

- Publicado em 1869
- Diagrama “Sankey” de fluxos



MAPA DE CÓLERA NO SOHO DE LONDRES

- Bomba de agua de Broad Street
- Mapa de John Snow definitivamente mostrou que este surto de cólera veio da agua infectada desta bomba.

JOHN TUKEY ON VISUALIZATION

> The simple graph has brought more information to the data analyst's mind than any other device.

> O gráfico simples trouxe mais informações à mente do analista dos dados do que qualquer outro dispositivo.

Tópico de suprema importância

ANÁLISE EXPLORATÓRIO DOS DADOS

EXPLORAÇÃO INICIAL DOS DADOS

- Onde queremos tentamos achar o que os dados estão dizendo
- Principal uso de visualizações
- Série de medidas e gráficos que mostram as variáveis
- Exploração das variáveis
  - Uma por vez (univariada)
  - Tabulações cruzadas de conjuntos de variáveis
- Sempre procurando valores de dados estranhos

DADOS: fute_mod.rds

- Conjunto dos dados sobre lesões relacionadas ao futebol nos EUA

```
fm <- readRDS(here("fute_mod_2020.rds")) %>%
  mutate(age_grp = factor(case_when(
    age < 18 ~ "youth",
    age < 60 ~ "adult",
    TRUE ~ "elderly"
  ))) %>%
# use relevel to change order of levels
  mutate(age_grp = fct_relevel(age_grp, c("youth", "adult", "elderly")))
glimpse(head(fm))
```

```
Rows: 6
Columns: 10
$ case_num    <chr> "160102033", "160106032", "160107304", "160109914", "16011…
$ trmt_date   <date> 2016-01-02, 2016-01-02, 2016-01-01, 2016-01-01, 2016-01-0…
$ age         <dbl> 27, 14, 9, 16, 17, 33
$ sex         <fct> Male, Male, Male, Female, Female, Male
$ body_part   <fct> Foot, Knee, Toe, Wrist, Wrist, Knee
$ diag        <fct> "Contusion Or Abrasion", "Fracture", "Fracture", "Strain, …
$ disposition <fct> Released, Released, Released, Released, Released, Released
$ psu         <fct> 63, 61, 8, 20, 73, 61
$ narrative   <chr> "27YOM PLAYING SOCCER COLLIDED WITH ANOTHER PLAYER CONTUSI…
$ age_grp     <fct> adult, youth, youth, youth, youth, adult
```

VARIÁVEL age

```
1  summarytools::descr(fm$age, stats = "common")
```

```
Descriptive Statistics
fm$age
N: 7603

                      age
--------------- ---------
           Mean     16.38
        Std.Dev      8.92
            Min      0.00
         Median     14.00
            Max     85.00
        N.Valid   7603.00
      Pct.Valid    100.00
```

MIN = 0.00 ?

```
summarytools::descr(fm$age)



## Descriptive Statistics
## fm$age
## N: 7603
##
##                       age
## ---------------- ---------
##             Mean     16.38
##          Std.Dev      8.92
##              Min      0.00
##               Q1     11.00
##           Median     14.00
##               Q3     17.00
##              Max     85.00
```

QUEM É ESSA PESSOA COM age = 0?

- UNK AGE MALE WAS HEADBUTTED BY ANOTHER PLAYER WHILE PLAYING SOCCERDX NOSE FX
- Não é um bebezinho; pessoa de idade desconhecida
- Mudar age = 0 para NA
- Existem outros casos com age = 0 ou próximo?

QUANTOS CASOS TÊM IDADE MENOS DE 5 ANOS

- Idade em que crianças começam escola

```
fm %>%
  filter(age < 5) %>%
  summarise(n = n())
```

```
# A tibble: 1 × 1
      n
  <int>
1    82
```

MEDIDAS DE TENDÊNCIA CENTRAL

INTERESSE EM PESSOAS QUE JOGAM FUTEBOL

- Quais tipos de lesões sofrem **amadores** jogando futebol
- Eliminar casos com idades menos de 5 anos

```
fm_mk2 <- fm %>%
  filter(age >= 5)
summarytools::descr(fm_mk2$age, stats = "common")
```

```
Descriptive Statistics
fm_mk2$age
N: 7521

                      age
--------------- ---------
           Mean     16.53
        Std.Dev      8.86
            Min      5.00
         Median     14.00
            Max     85.00
        N.Valid   7521.00
      Pct.Valid    100.00
```

MÉDIAS DE DUAS DISTRIBUIÇÕES

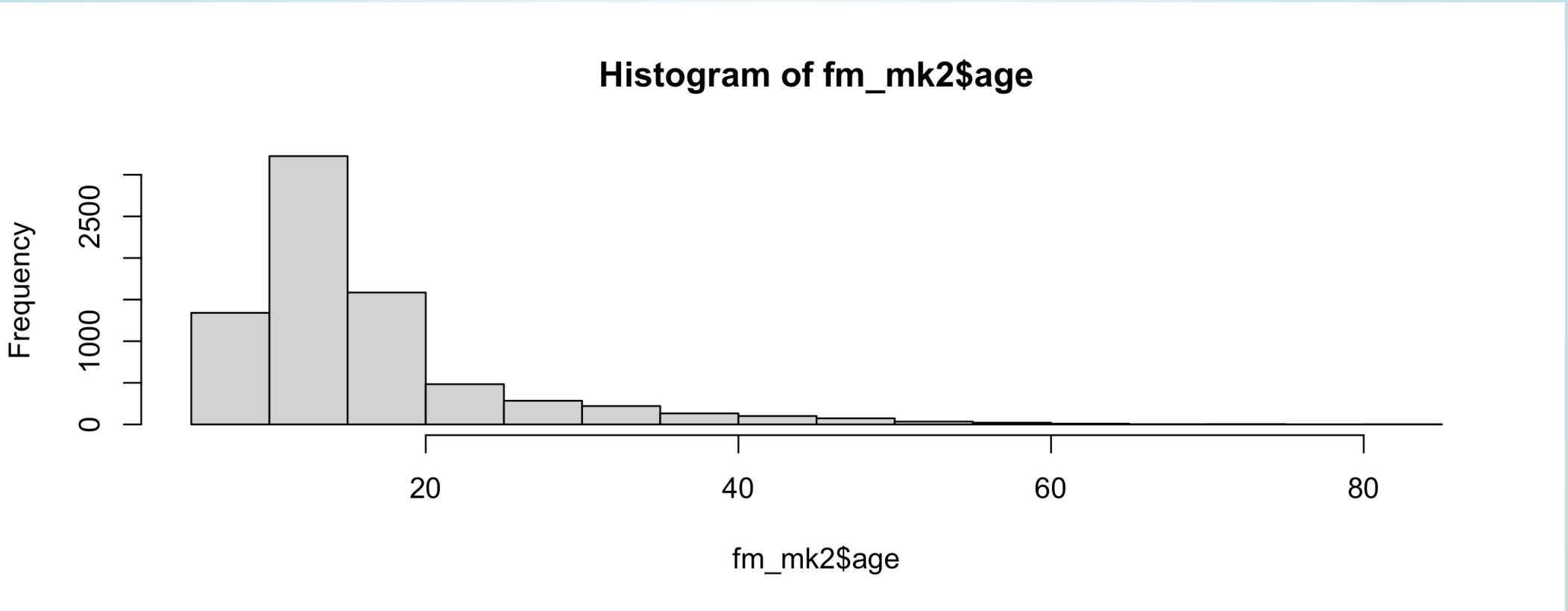
- Média de `fm` (com pequenas crianças): 16.3786225
- `Média de fm_mk2` (sem pequenas crianças): 16.5261268
- Se removêssemos 82 casos, porque a diferença não é maior?

O QUE É A MÉDIA?

- Um das medidas de **tendência central**
  - Valores que ficam no meio da distribuição
  - Valores populares
- O **centro aritmético** de uma distribuição
- Sensível aos valores extremos

VISUALIZAÇÃO CLÁSSICA DE UMA DISTRIBUIÇÃO - HISTOGRAMA

```
1 hist(fm_mk2$age)
```



HISTOGRAMA FOI ÚTIL?

- Não deu muita informação
- Problema de `bins`
- Apresentação muito feia

SISTEMA GRÁFICO ALTERNATIVO

GRAMMAR OF GRAPHICS - ggplot2

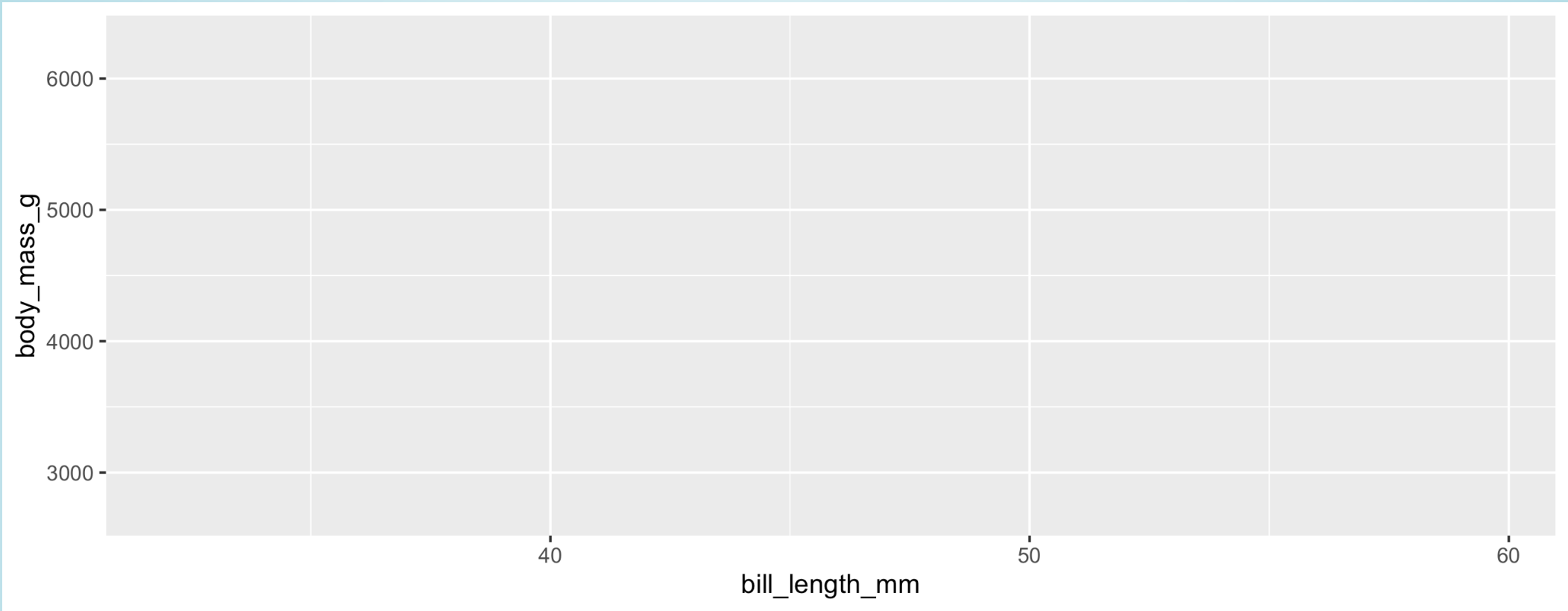
- Um sistema para **construir** gráficos (que se comunicam muito melhor)
- Um dos primeiros produtos de Hadley Wickham
- Construir seu gráfico camada por camada
- Começar por especificar um conjunto de dados: `penguin`
  - Variáveis `bill_length_mm` e `body_mass_g`

```
Rows: 6
Columns: 5
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, 36.7, 39.3, 38.9
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, 19.3, 20.6, 17.8
$ flipper_length_mm <dbl> 181, 186, 195, 193, 190, 181
$ body_mass_g       <dbl> 3750, 3800, 3250, 3450, 3650, 3625
$ species           <chr> "Adelie", "Adelie", "Adelie", "Adelie", "Adelie", "A…
```

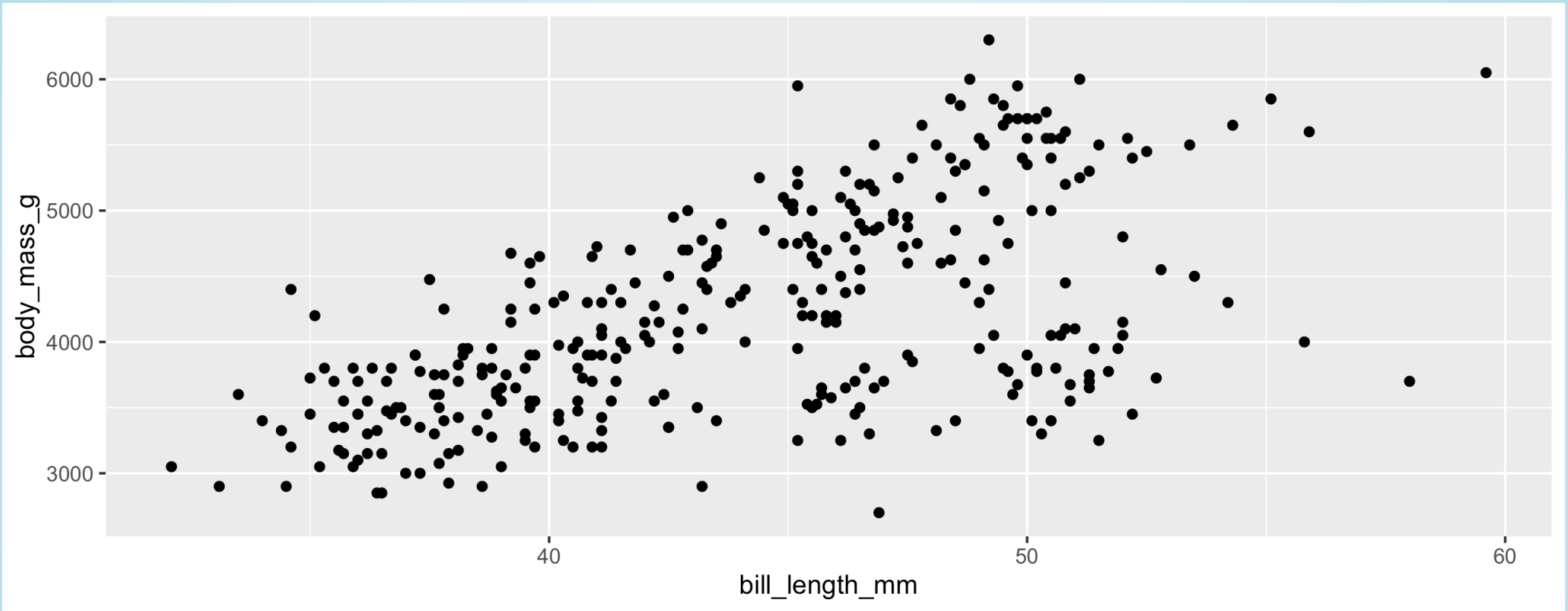
```
1 ggplot()
```

```
1  ggplot(data = pd)
```

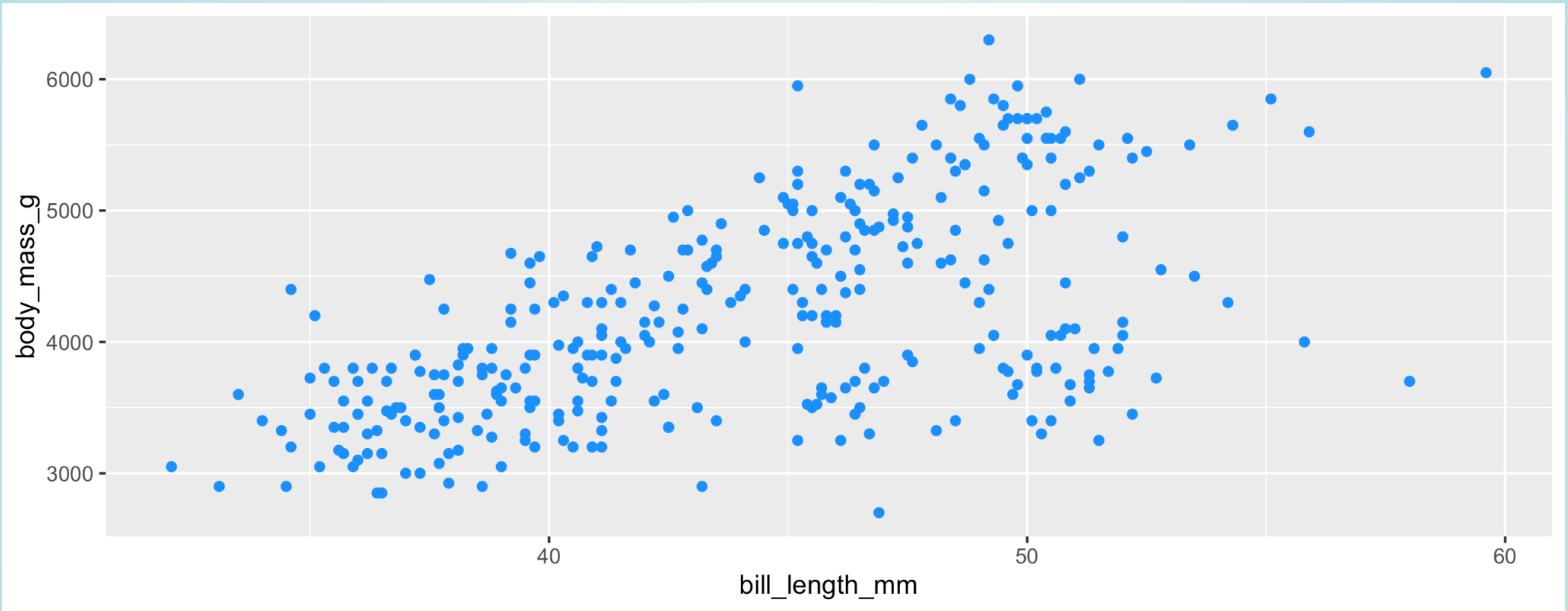
```
1  ggplot(data = pd, aes(x = bill_length_mm, body_mass_g ))
```



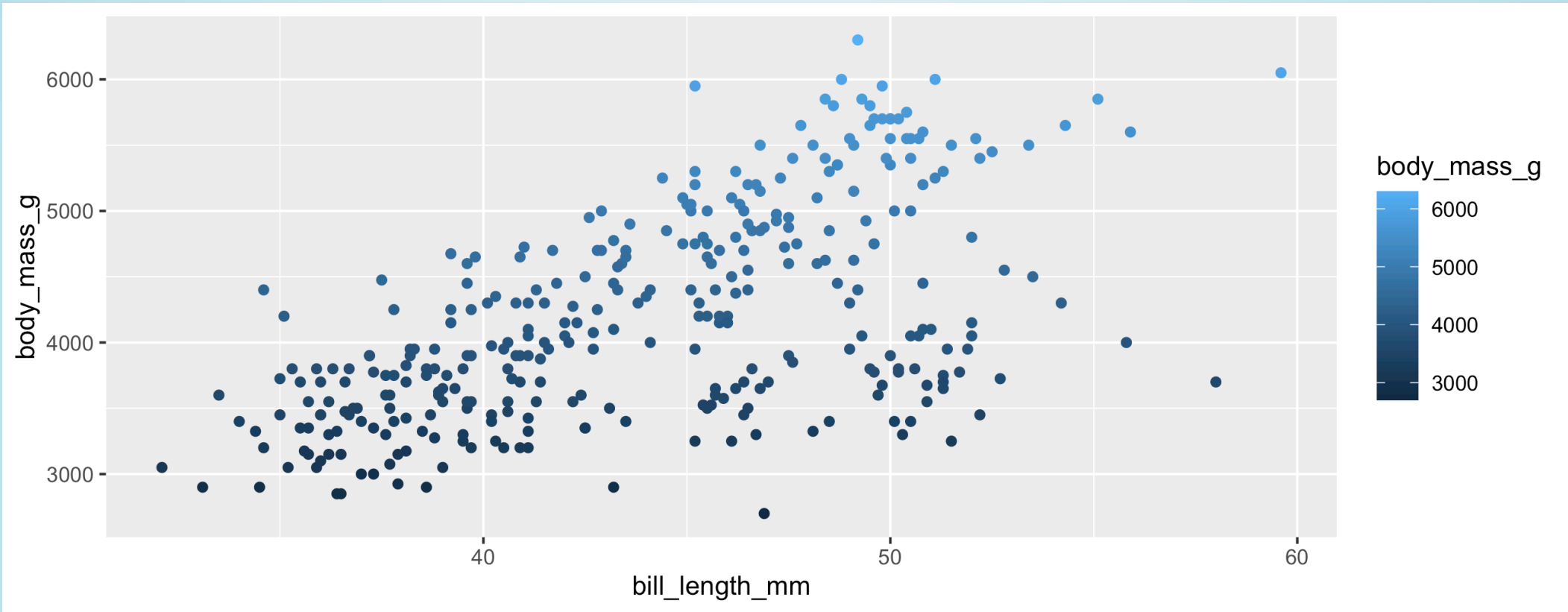
```
1  ggplot(data = pd, aes(x = bill_length_mm, body_mass_g )) +
2    geom_point()
```



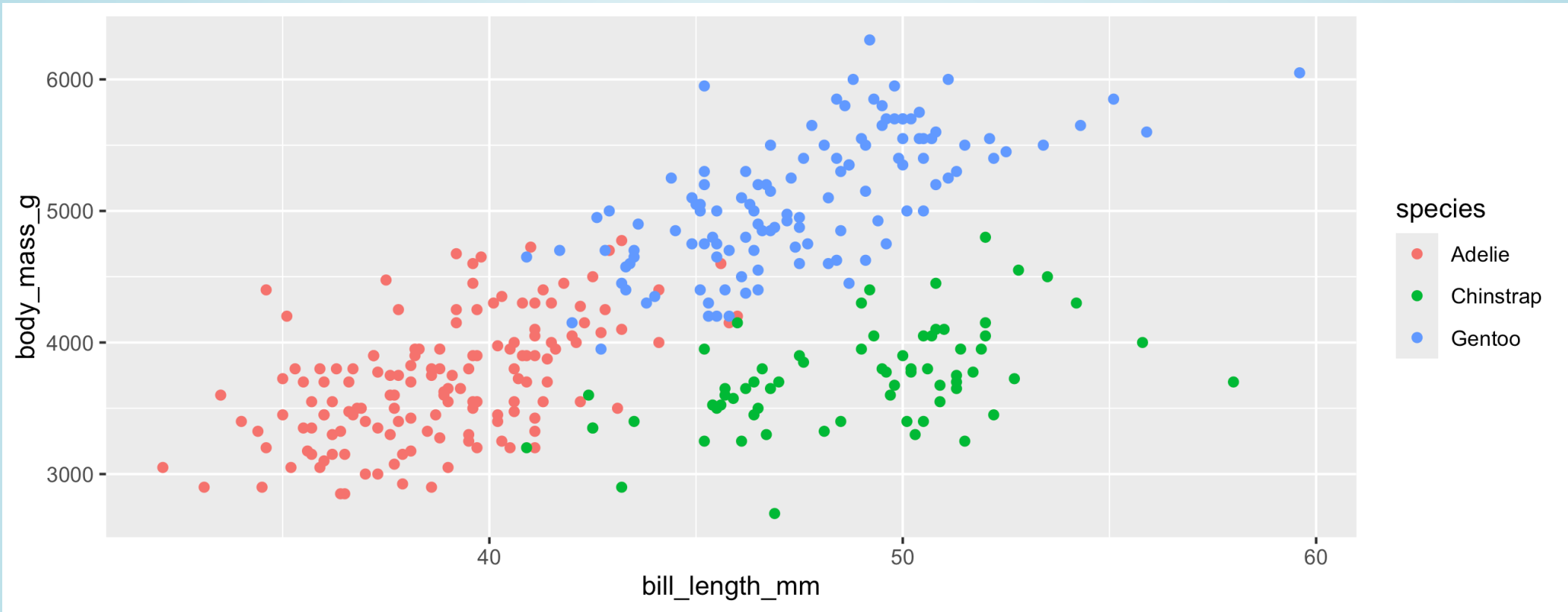
```
1  ggplot(data = pd, aes(x = bill_length_mm, body_mass_g )) +
2    geom_point(color = "dodgerblue")
```



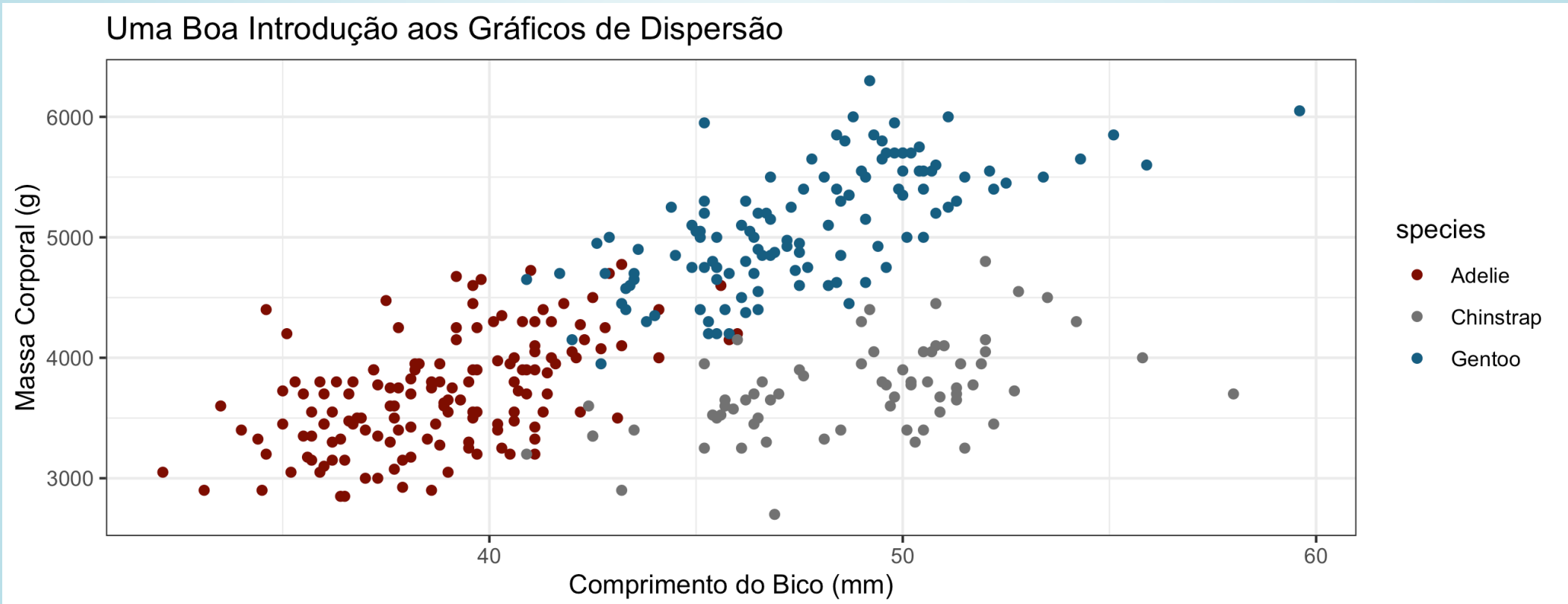
```
ggplot(data = pd, aes(x = bill_length_mm, body_mass_g, color = body_mass_g)
  geom_point()
```



```
ggplot(data = pd, aes(x = bill_length_mm, body_mass_g, color = species )) +
  geom_point()
```



```r
ggplot(data = pd, aes(x = bill_length_mm, body_mass_g, color = species)) +
  geom_point() +
  labs(title = "Uma Boa Introdução aos Gráficos de Dispersão", x = "Comprim
       y = "Massa Corporal (g)") +
  scale_colour_manual(values = c("#800000FF", "#767676FF", "#155F83FF")) +
  theme_bw()
```

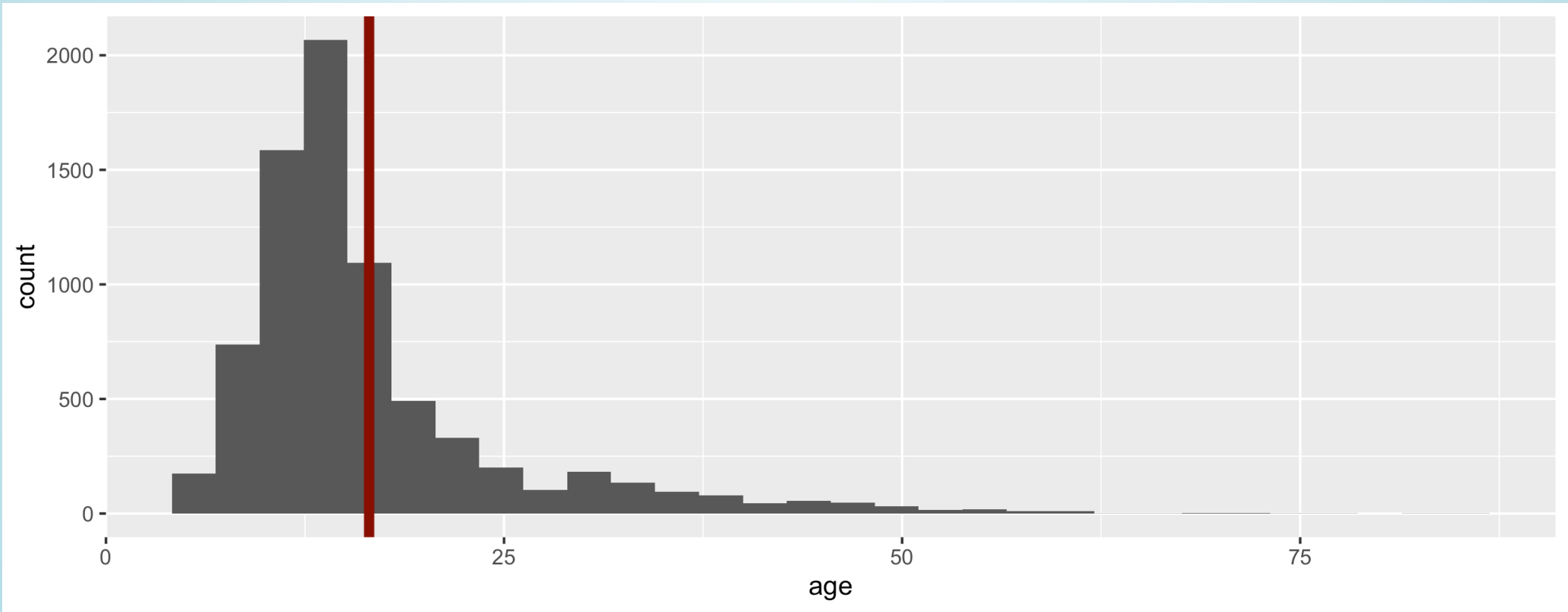


RECURSOS - ggplot

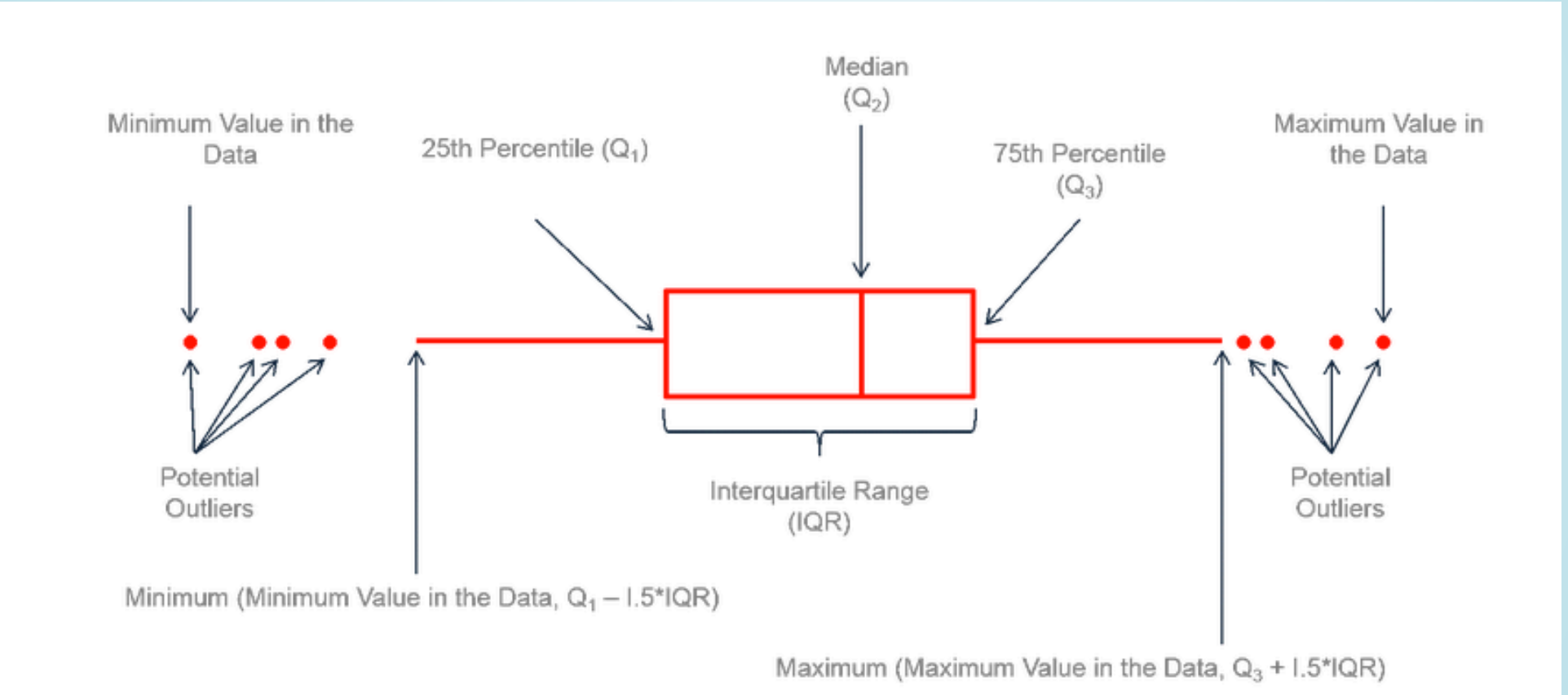
- Winston Chang, **R Graphics Cookbook**, 2Ed., https://r-graphics.org
- Kieran Healy, **Data Visualization: A Practical Introduction**, https://socviz.co
- https://r-graph_gallery.com - examples of many types of graphs with explanations and code
- ggplot cheat sheet

HISTOGRAMA DE age

```
avg_age <- mean(fm_mk2$age)
ggplot(data = fm_mk2, aes(x = age)) +
  geom_histogram(bins = 30) +
  geom_vline(xintercept = avg_age, colour = "darkred", size = 2)
```



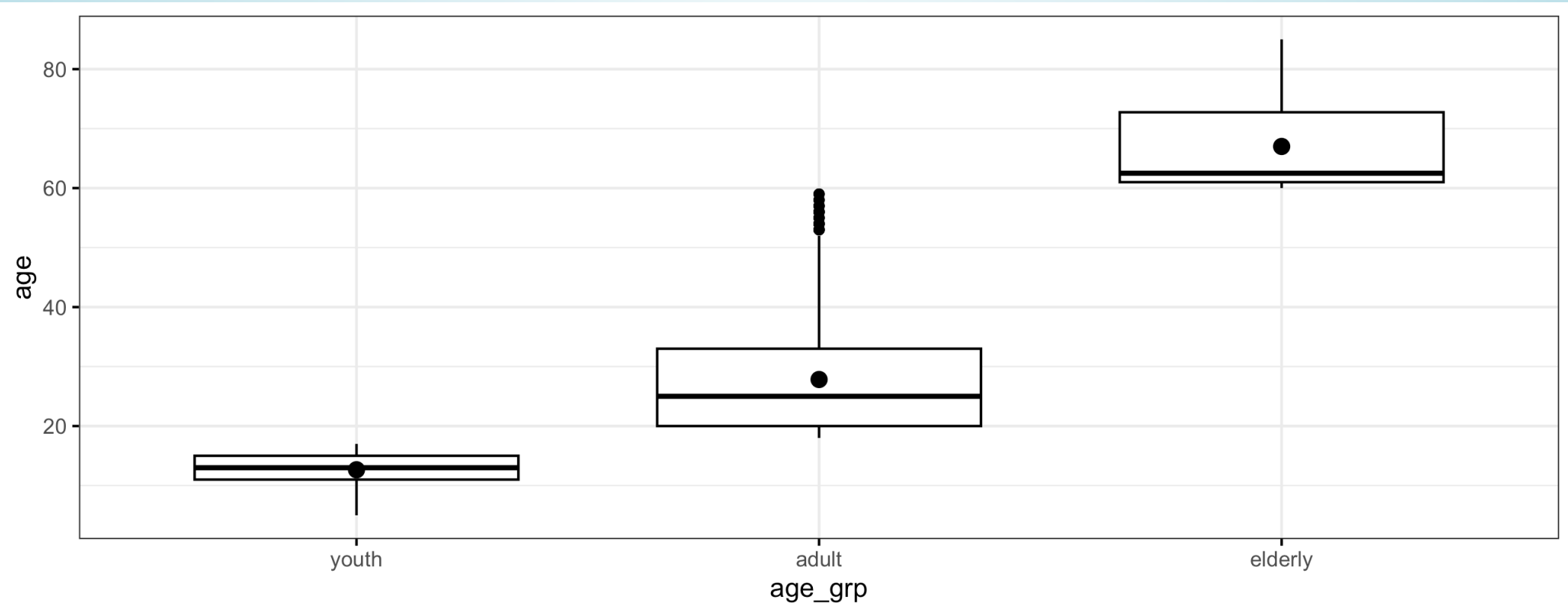
SEGUNDO GRÁFICO QUE MOSTRA DISTRIBUIÇÕES BEM - *BOXPLOT*



- source: https://r-graph-gallery.com

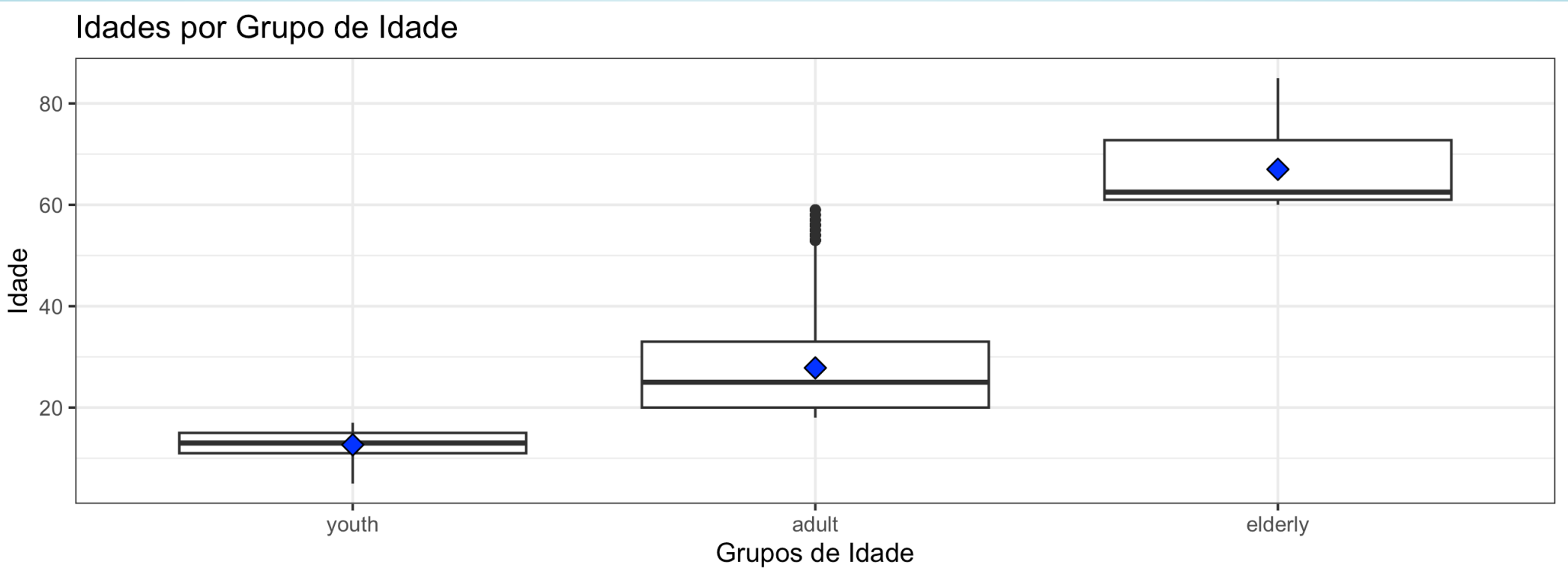
BOXPLOT COM OS DADOS DE FUTEBOL

```
fm_mk2 |> ggpubr::ggboxplot(x = "age_grp",
                            y = "age",
                            add = "mean",
                            ggtheme = theme_bw())
```



BOXPLOT DE FUTEBOL COM ggplot

```
fm_mk2 |>
  ggplot(mapping = aes(x = age_grp,  y = age,)) +
  geom_boxplot() +
  stat_summary(fun = "mean", geom = "point", shape = 23,
               size = 3, fill = "blue") +
  labs(title = "Idades por Grupo de Idade",
       x = "Grupos de Idade",
       y = "Idade",
       caption = "Texto que explica o gráfico.") +
  theme_bw()
```



Texto que explica o gráfico.

DE ONDE VEIO ESTA INFORMAÇÃO SOBRE A MÉDIA?

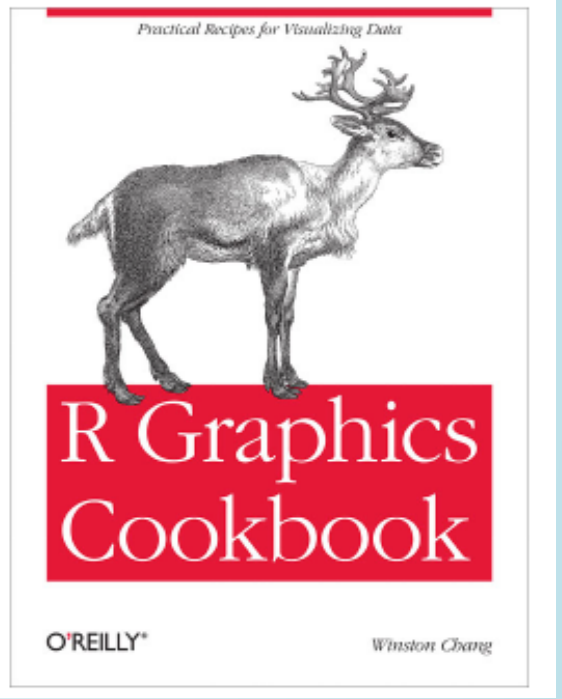
6.8.1 Problem

You want to add markers for the mean to a box plot.

6.8.2 Solution

Use `stat_summary()`. The mean is often shown with a diamond, so we'll use shape 23 with a white fill. We'll also make the diamond slightly larger by setting `size = 3` (Figure 6.21):

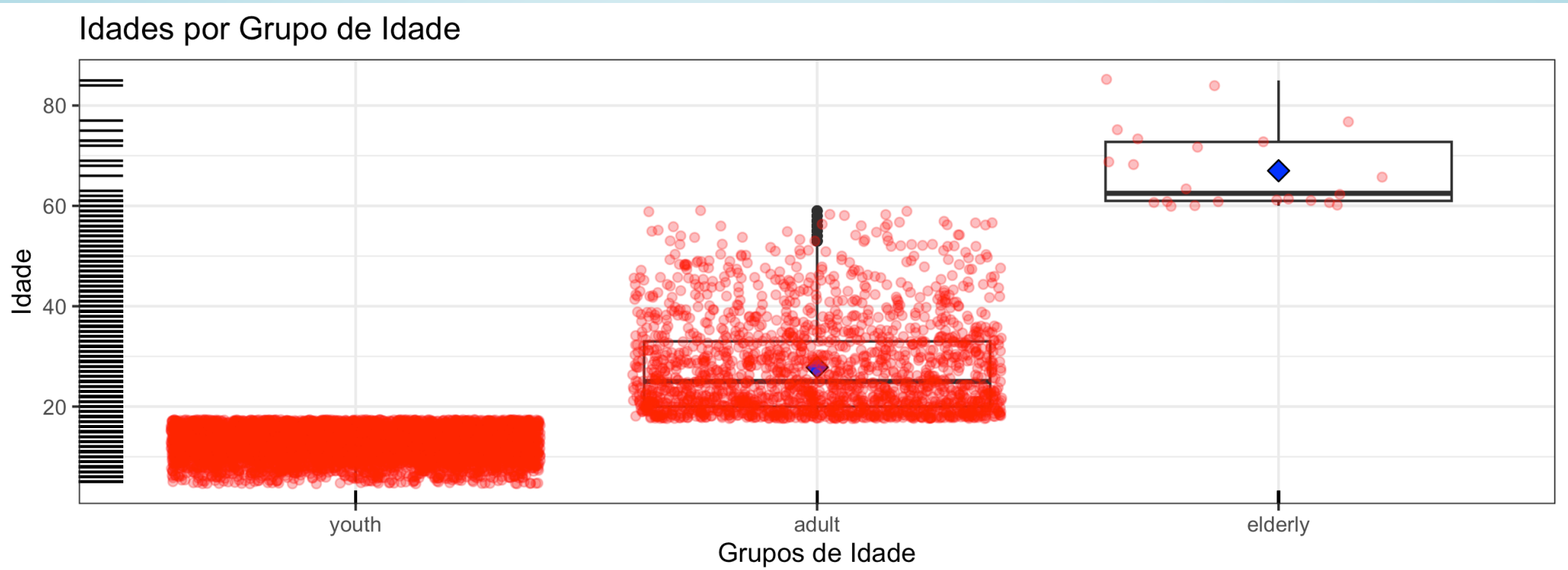
```
library(MASS) # Load MASS for the birthwt data set

ggplot(birthwt, aes(x = factor(race), y = bwt)) +
  geom_boxplot() +
  stat_summary(fun.y = "mean", geom = "point", shape = 23, size = 3, fill = "white")
#> Warning: The `fun.y` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
#> ℹ Please use the `fun` argument instead.
#> This warning is displayed once every 8 hours.
#> Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
#> generated.
```



ggplot BOXPLOT - 2

- Gostaria de saber onde cai os pontos
  - `geom_jitter` - mostra todos os pontos com um pouco de variação
  - `geom_rug` - mostra os casos individuais numa certa dimensão

```
fm_mk2 |>
  ggplot(mapping = aes(x = age_grp,  y = age,)) +
  geom_boxplot() +
  stat_summary(fun = "mean", geom = "point", shape = 23, size = 3, fill = "blue") +
  geom_jitter(alpha = .3, color = "red") +
  geom_rug() +
  labs(title = "Idades por Grupo de Idade",
       x = "Grupos de Idade",
       y = "Idade",
       caption = "Texto que explica o gráfico.") +
  theme_bw()
```

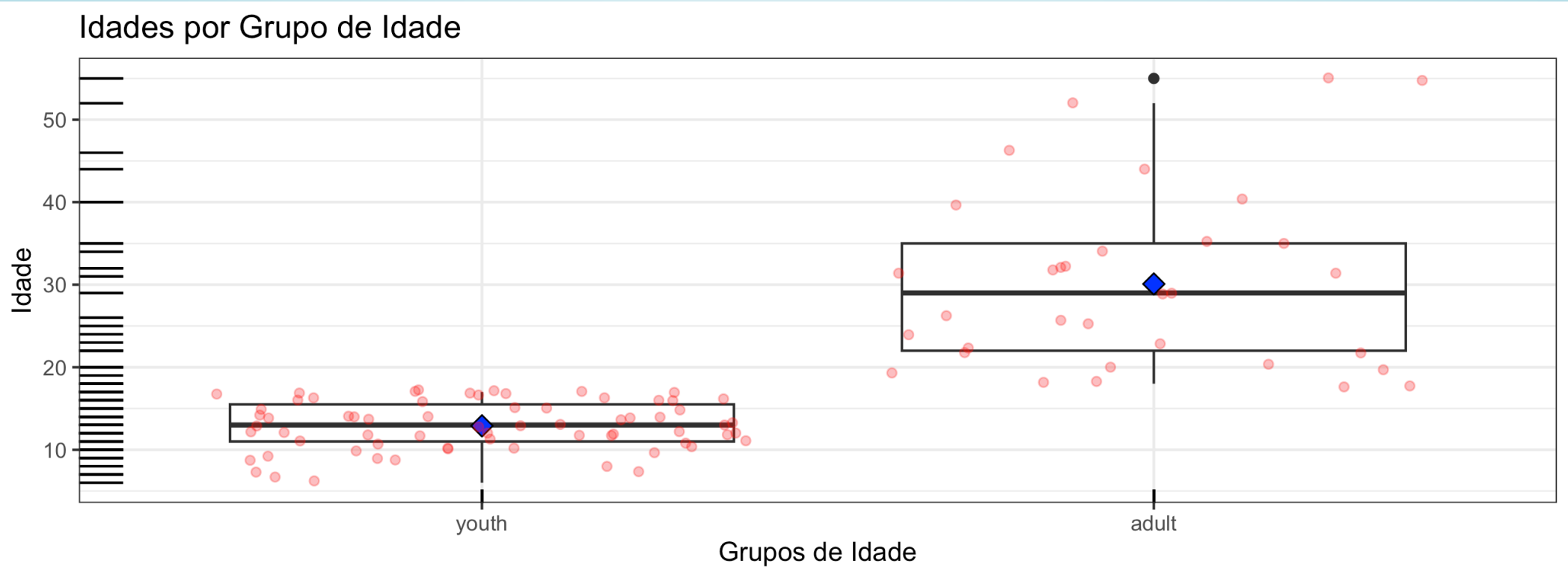


Texto que explica o gráfico.

LIÇÃO DISSO

- Com muitos pontos, impossível ver a distribuição
- Criar uma versão com só 100 pontos

```
fm_mk2 |>
  slice_sample(n = 100) |>
  ggplot(mapping = aes(x = age_grp,  y = age,)) +
  geom_boxplot() +
  stat_summary(fun = "mean", geom = "point", shape = 23, size = 3, fill = "
  geom_jitter(alpha = .3, color = "red") +
  geom_rug() +
  labs(title = "Idades por Grupo de Idade",
       x = "Grupos de Idade",
       y = "Idade",
       caption = "Texto que explica o gráfico.") +
  theme_bw()
```

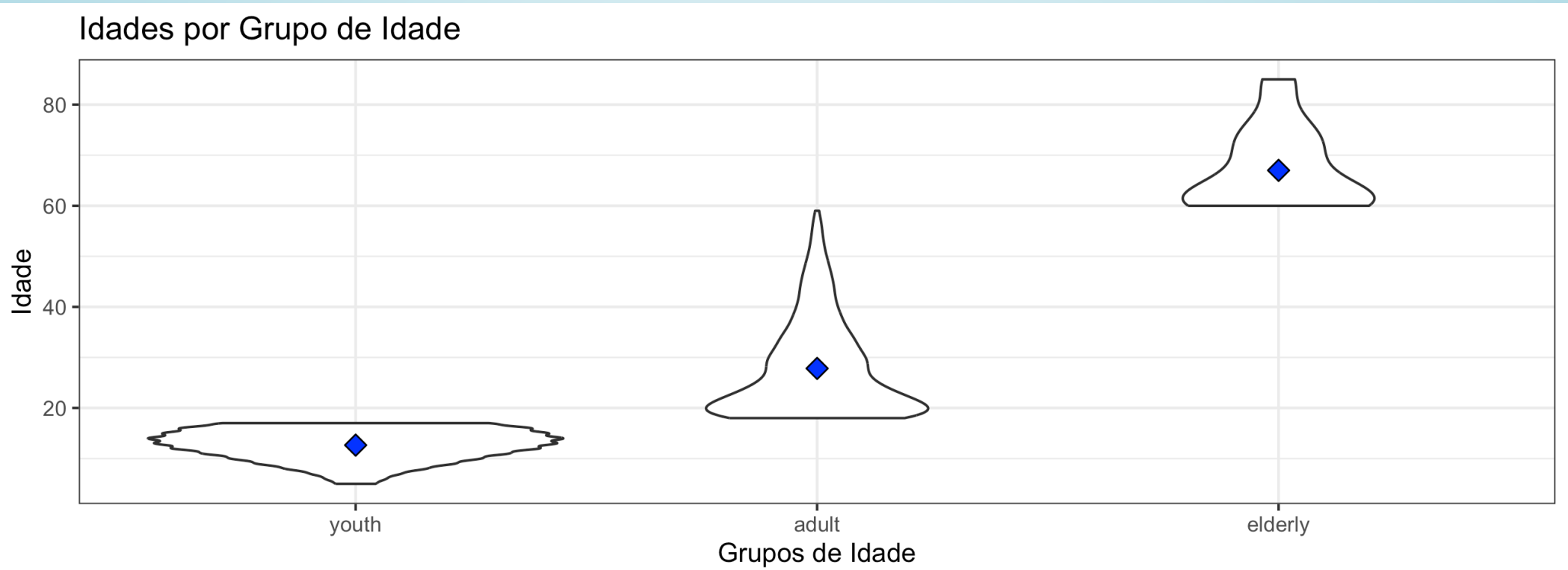


Texto que explica o gráfico.

ALTERNATIVA DE BOXPLOT – GRÁFICO DE VIOLINO

- Mostra concentrações de dados pela largura da caixa invés de *jitter* e *rug*

```
fm_mk2 |>
  ggplot(mapping = aes(x = age_grp,  y = age,)) +
  geom_violin() +
  stat_summary(fun = "mean", geom = "point", shape = 23, size = 3, fill = "
  labs(title = "Idades por Grupo de Idade",
       x = "Grupos de Idade",
       y = "Idade",
       caption = "Texto que explica o gráfico.") +
  theme_bw()
```

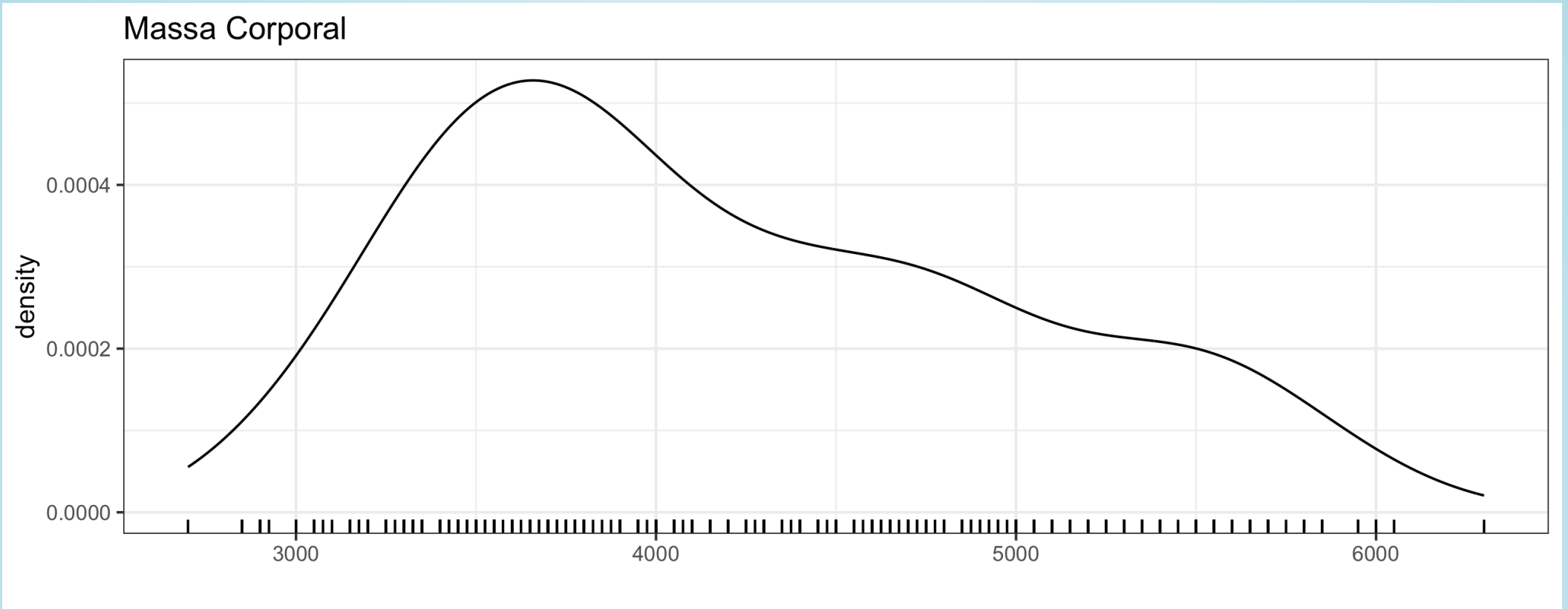


Texto que explica o gráfico.

PLOTAGEM DE DENSIDADE

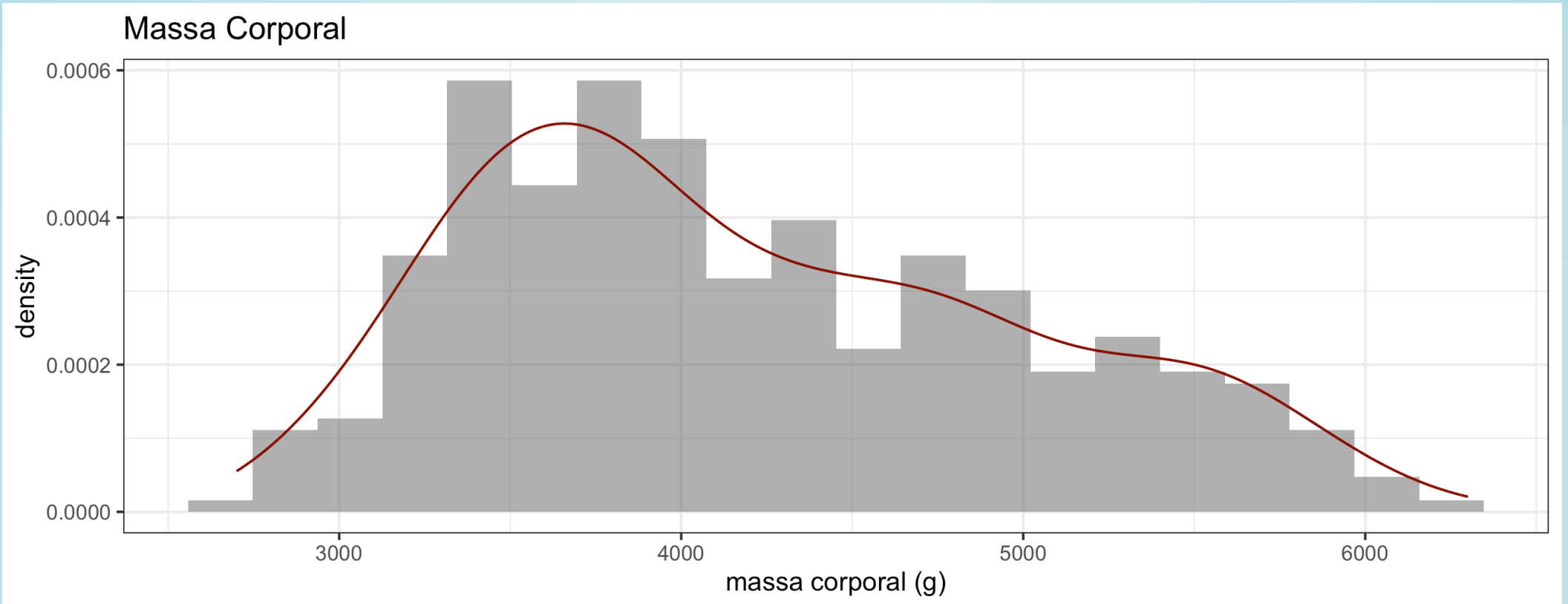
- Tecnicamente, uma plotagem de densidades de *kernels*
  - *Kernel Density* divide a distribuição em partes e calcular a densidade em cada região não-linearmente e recombinar elas para compôr uma curva suave
- Usar `body_mass_g` para ilustrar

```
set.seed(42)
pd |> # dataframe penguins |>
  ggplot(mapping = aes(x = body_mass_g)) +
  geom_density() +
  labs(title = "Massa Corporal",
       x = "") +
  geom_rug()+
  theme_bw()
```


Massa Corporal
density
0.0004
0.0002
0.0000
3000
4000
5000
6000

DENSIDADE COM HISTOGRAMA

```
pd |> # dataframe penguins |>
  ggplot(mapping = aes(x = body_mass_g)) +
  geom_histogram(aes(y = ..density..), bins = 20, alpha = 0.5) +
  geom_density(colour = "darkred") +
  labs(title = "Massa Corporal",
       x = "massa corporal (g)") +
  theme_bw()
```


Massa Corporal
0.0006
0.0004
0.0002
0.0000
density
3000
4000
5000
6000
massa corporal (g)

COMPARAR AS ESPÉCIES

```
set.seed(42)
pd |> # dataframe penguins |>
  group_by(species) |>
  ggplot(mapping = aes(x = body_mass_g, colour = species, fill = species))
  geom_density(alpha = 0.5) +
  labs(title = "Massa Corporal",
       x = "") +
  geom_rug()+
  theme_bw()
```

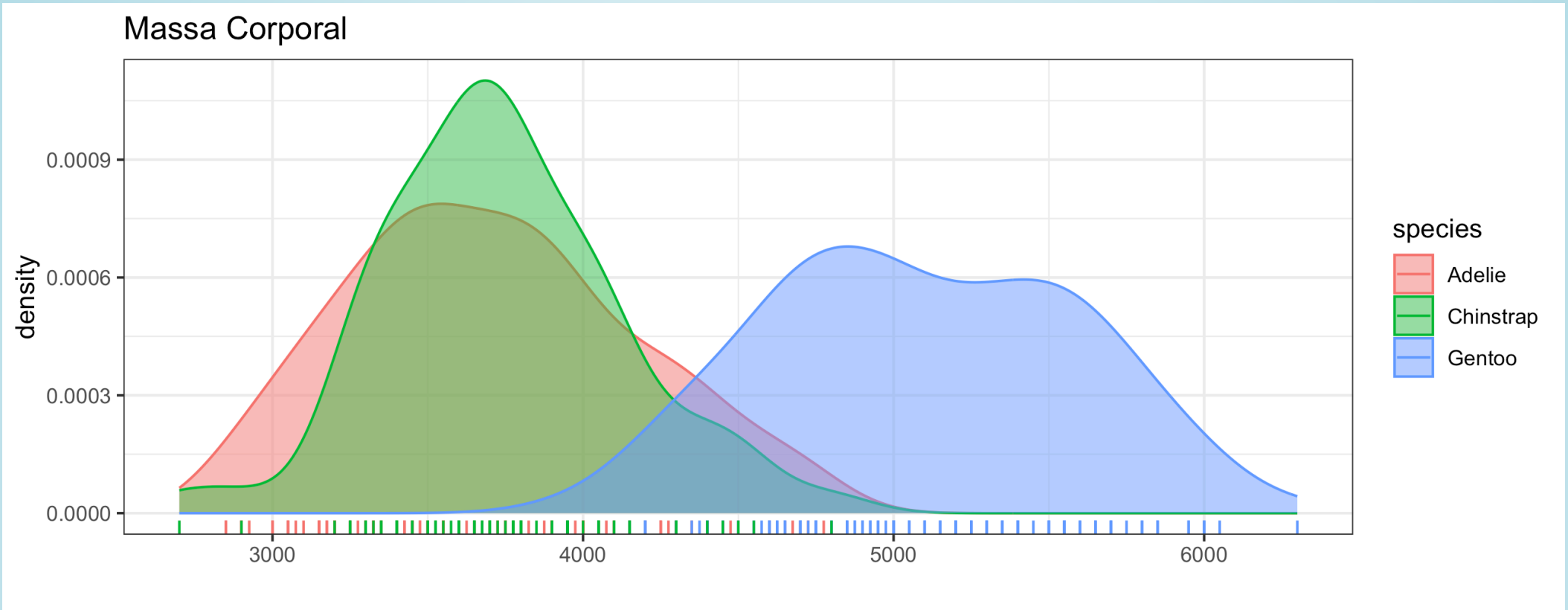
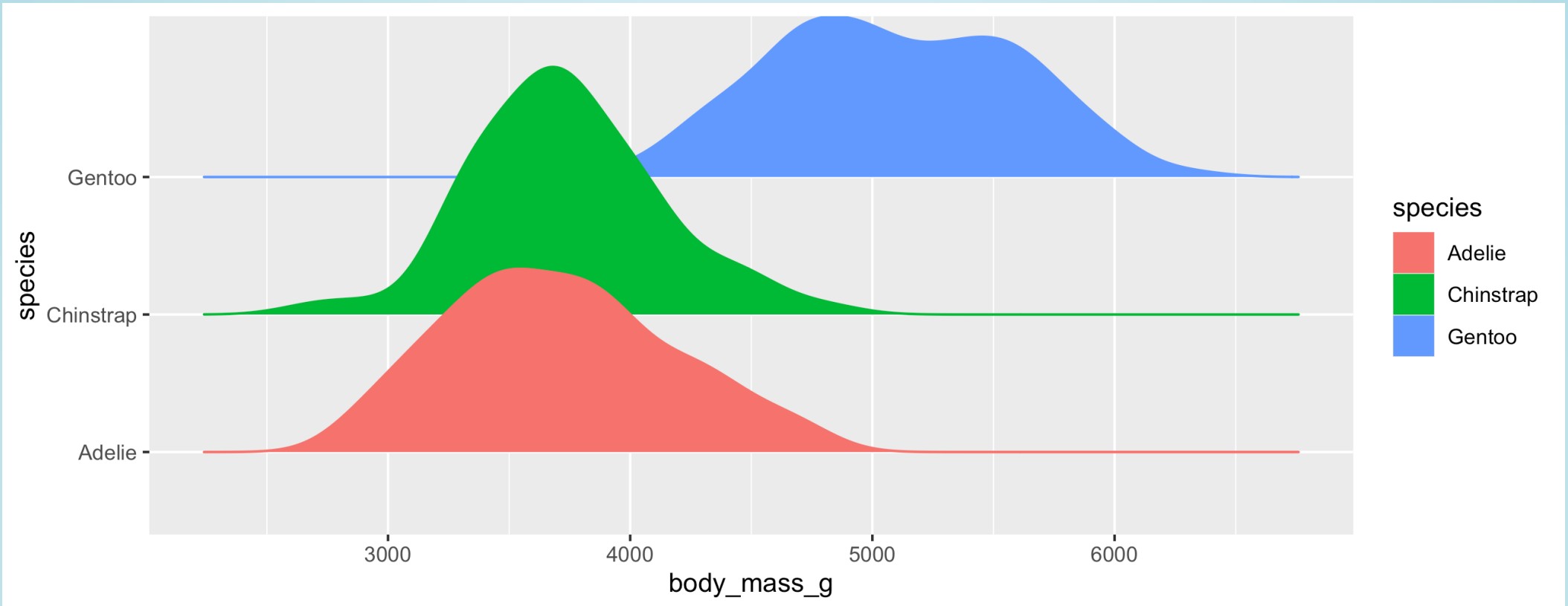

Massa Corporal
density
0.0009
0.0006
0.0003
0.0000
3000
4000
5000
6000
species
Adelie
Chinstrap
Gentoo

GRÁFICO *RIDGELINE* - EXTENSÃO DE DENSIDADE

- Maneira fácil para comparar as densidades de várias categorias das variáveis
- Precisa instalar `ggridges` de CRAN

```
library(ggridges)

pd |>
  ggplot(aes(x = body_mass_g,
             y = species,
             colour = species,
             fill = species)) +
  geom_density_ridges() +
  theme_gray() # default theme for ggplot
```


species
Gentoo
Chinstrap
Adelie
3000
4000
5000
6000
body_mass_g
species
Adelie
Chinstrap
Gentoo

RIDGELINE COM MAIS OOOMPH

- Mudar cores para uma paleta mais agradável
  - Usar `ggsci` paleta `uchicago`
- Tirar a legenda - desnecessária
- Reduz o tamanho das caudas
- Mostrar os quartis nas curvas de densidade

```
library(ggridges)

pd |>
  ggplot(aes(x = body_mass_g,
             y = species,
             fill = species)) +
  stat_density_ridges(quantile_lines = TRUE, rel_min_height = 0.01) +
  scale_fill_uchicago(palette = "default", alpha = 0.8) +
  guides(fill = FALSE) +
  labs(title = "Massa Corporal por Espécie",
       x = "Massa Corporal (g)",
       y = "") +
  theme_gray() # default theme for ggplot
```

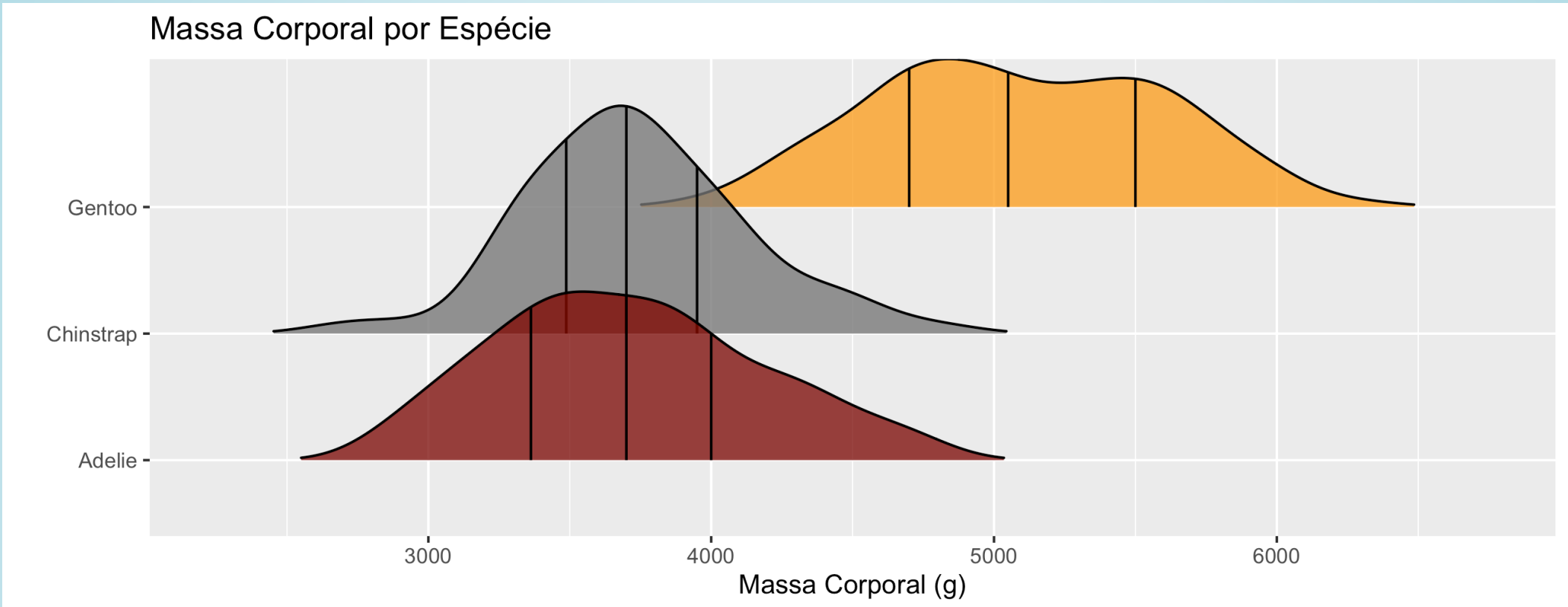

Massa Corporal por Espécie
Gentoo
Chinstrap
Adelie
3000
4000
5000
6000
Massa Corporal (g)

GRÁFICO *RADAR* - OUTRA MANEIRA DE RETRATAR DIMENSÕES

- Cria um campo circular para mostrar um número de dimensões
- Funciona melhor comparando poucas classes
- Precisa preparar os dados para utilizar este tipo de gráfico
  - Comparando ou a média ou mediana das classes em cada dimensão
  - Vai pôr as dimensões na escala de 0 até 1 utilizando `scales::rescale()`
- **Usa pacote `ggradar`**
  - **remotes::install_github("ricardo-bion/ggradar")**

RADAR - PREPARAÇÃO DOS DADOS

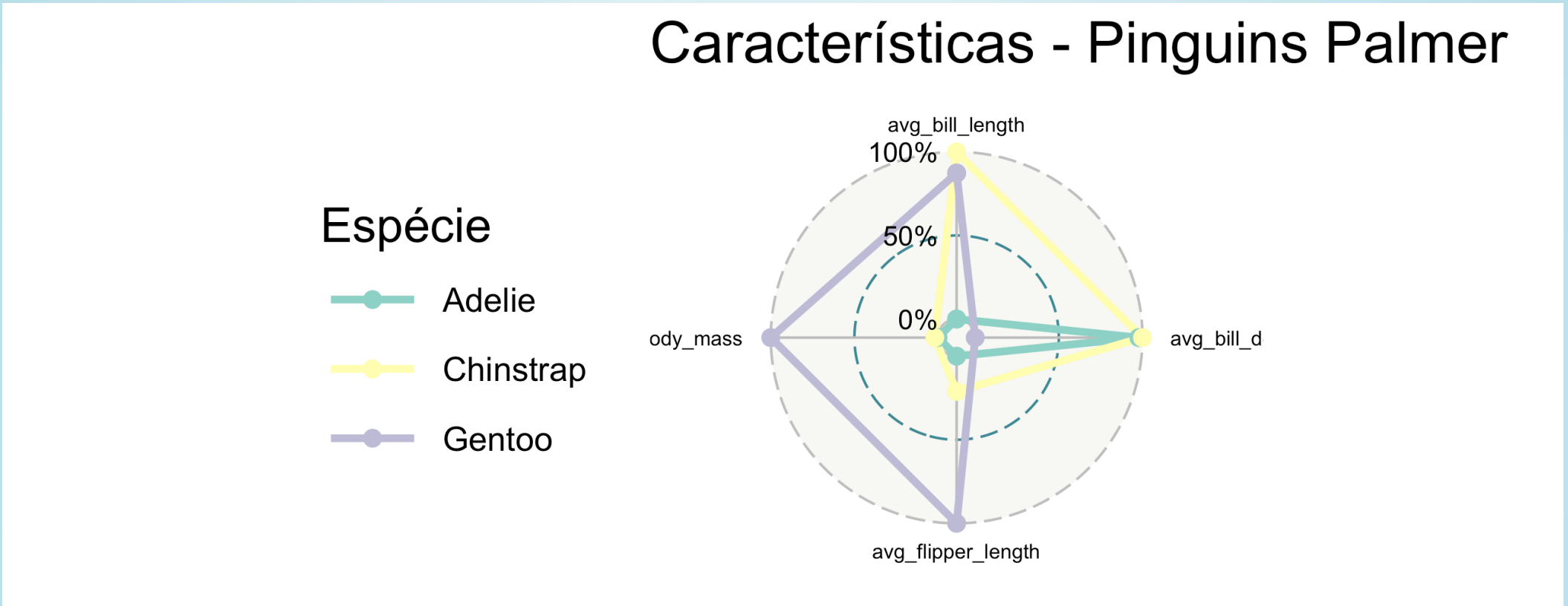
```
 1 pacman::p_load(ggradar, scales)
 2
 3 pd_radar <- pd |>
 4   tidyr::drop_na() |>    # NAs can't be processed in ops below
 5   group_by(species)  |>
 6   summarise(
 7     avg_bill_length = mean(bill_length_mm),
 8     avg_bill_depth = mean(bill_depth_mm),
 9     avg_flipper_length = mean(flipper_length_mm),
10     avg_body_mass = mean(body_mass_g)
11   )  |>
12   ungroup()  |>
13   mutate_at(vars(-species), rescale)
14 pd_radar
```

```
# A tibble: 3 × 5
  species   avg_bill_length avg_bill_depth avg_flipper_length avg_body_mass
  <chr>               <dbl>          <dbl>               <dbl>         <dbl>
1 Adelie              0              0.979               0             0
2 Chinstrap           1              1                   0.211         0.0194
3 Gentoo              0.874          0                   1             1
```

CÓDIGO DO GRÁFICO

```
ggpengrad <- pd_radar %>%
  ggradar(
    font.radar = "arial",
    grid.label.size = 5,  # Affects the grid annotations (0%, 50%, etc.)
    axis.label.size = 3, # Afftects the names of the variables
    group.point.size = 3,  # Simply the size of the point
    legend.title = "Espécie",
    plot.title = "Características - Pinguins Palmer",
  )
```

RESULTADO

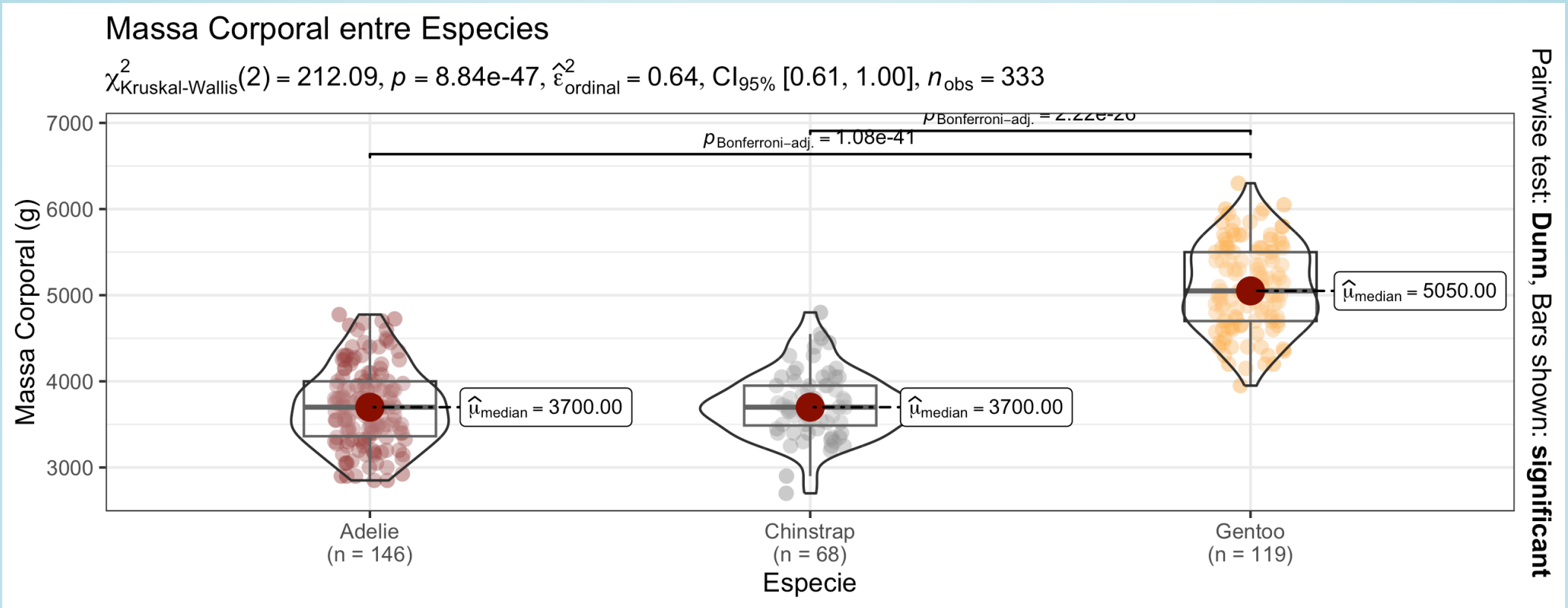

Características - Pinguins Palmer
Espécie
Adelie
Chinstrap
Gentoo
avg_bill_length
100%
50%
0%
ody_mass
avg_bill_d
avg_flipper_length

ÚLTIMO TIPO DE HOJE

INCORPORAR GRÁFICOS COM TESTES ESTATÍSTICOS

- Pacote `ggstatsplot` - precisa instalar da CRAN
- Pacote tem muitos combinações por vários testes
- Diferença entre a média de massa corporal dos especies é significativa?

```
library(ggstatsplot)
pd |>
  ggbetweenstats(
    x = species,
    y = body_mass_g,
    type = "nonparametric",
    p.adjust.method = "bonferroni",
    xlab = "Especie",
    ylab = "Massa Corporal (g)",
    title = "Massa Corporal entre Especies",
    ggtheme = theme_bw(),
    package = "ggsci",
    palette = "default_uchicago"
  )
```


Massa Corporal entre Especies
$\chi^2_{Kruskal-Wallis}(2) = 212.09$, $p = 8.84e-47$, $\hat{\varepsilon}^2_{ordinal} = 0.64$, $CI_{95\%}$ [0.61, 1.00], $n_{obs} = 333$
$p_{Bonferroni-adj.}$ = 2.22e-20
$p_{Bonferroni-adj.}$ = 1.08e-41
Massa Corporal (g)
7000
6000
5000
4000
3000
$\hat{\mu}_{median}$ = 3700.00
$\hat{\mu}_{median}$ = 3700.00
$\hat{\mu}_{median}$ = 5050.00
Adelie
(n = 146)
Chinstrap
(n = 68)
Gentoo
(n = 119)
Especie
Pairwise test: **Dunn**, Bars shown: **significant**

USO DE PLOTAGENS COM STATS

- Eu uso essas plotagens para minha análise das variáveis, não para a apresentação
- Têm muita informação para apresentações; não simplifica a visualização