

# ANÁLISE DOS DADOS COM R

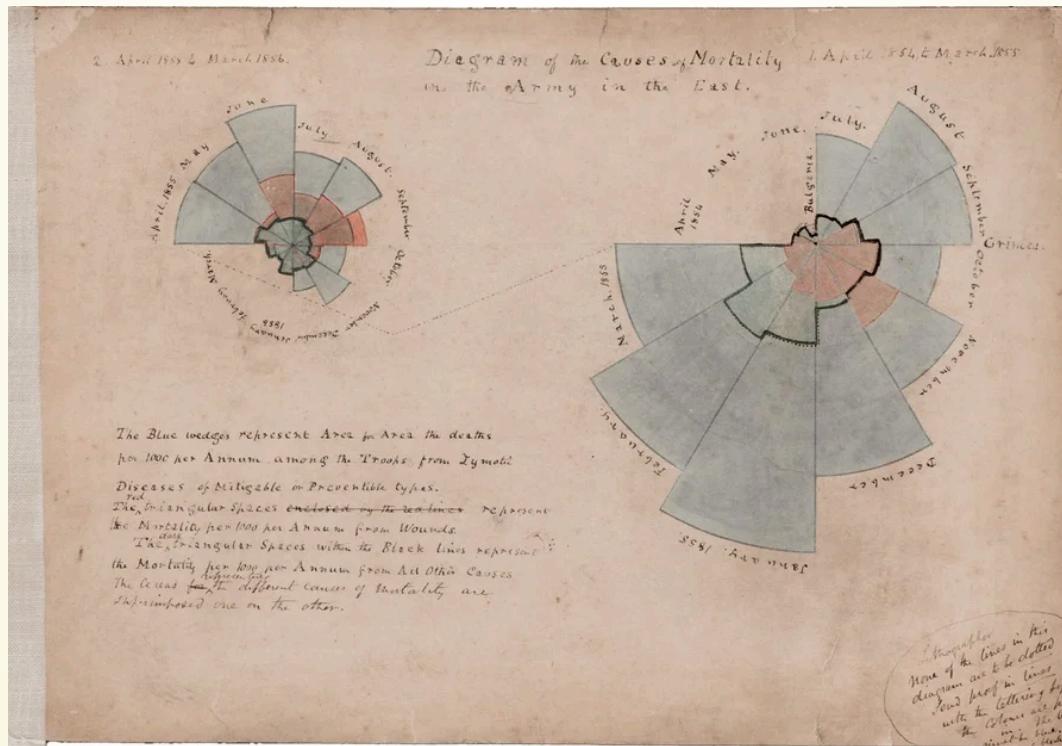
Visualização dos Dados e EDA

James R. Hunter, PhD  
Retrovírologia, EPM, UNIFESP

# VISUALIZAÇÃO DOS DADOS

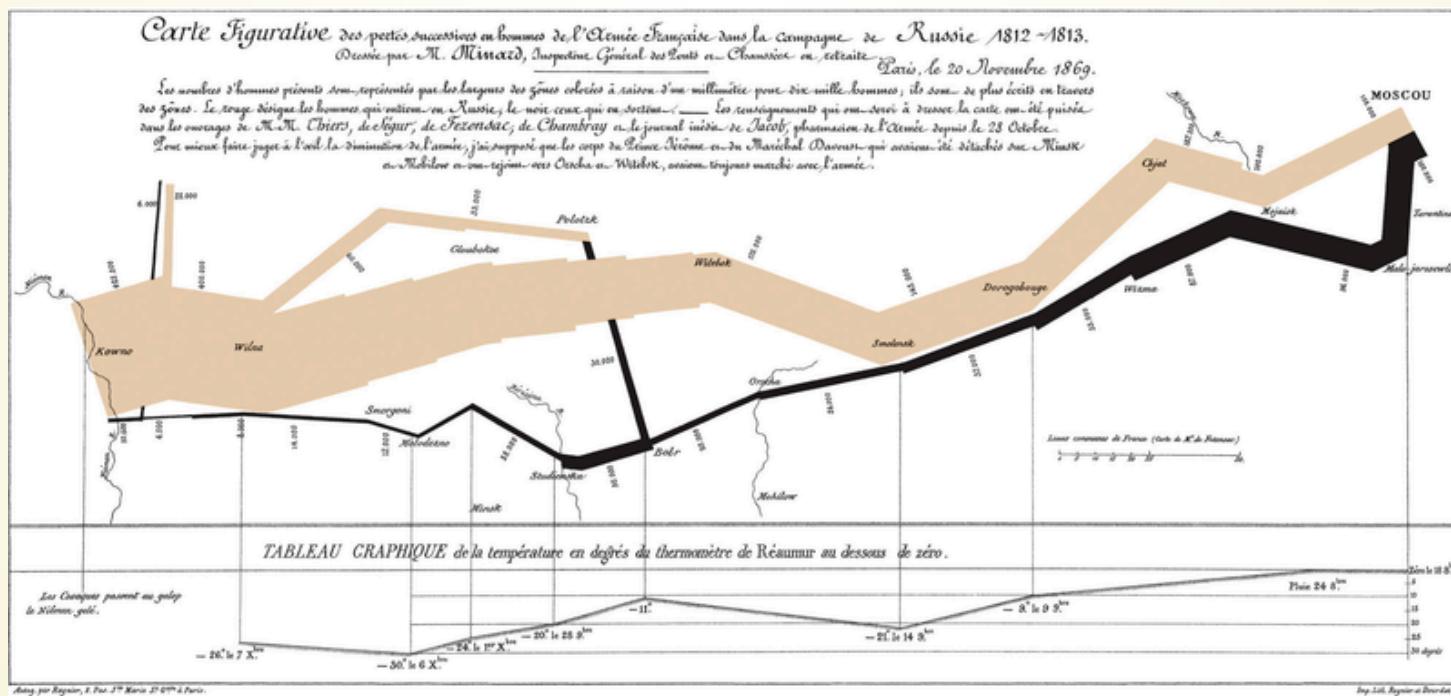
# FLORENCE NIGHTINGALE - DIAGRAMMA DE ROSA

- 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

```
1 fm <- readRDS(here("fute_mod_2020.rds")) %>%
2   mutate(age_grp = factor(case_when(
3     age < 18 ~ "youth",
4     age < 60 ~ "adult",
5     TRUE ~ "elderly"
6   ))) %>%
7   # use relevel to change order of levels
8   mutate(age_grp = fct_relevel(age_grp, c("youth", "adult", "elderly")))
9 glimpse(head(fm))
```

Rows: 6  
Columns: 10

	case_num	trmt_date	age	sex	body_part	diag	disposition	psu	narrative	age_grp
1	"160102033", "160106032", "160107304", "160109914", "16011...	2016-01-02, 2016-01-02, 2016-01-01, 2016-01-01, 2016-01-0...	27, 14, 9, 16, 17, 33	Male, Male, Male, Female, Female, Male	Foot, Knee, Toe, Wrist, Wrist, Knee	"Contusion Or Abrasion", "Fracture", "Fracture", "Strain, ...	Released, Released, Released, Released, Released, Released	63, 61, 8, 20, 73, 61	"27YOM PLAYING SOCCER COLLIDED WITH ANOTHER PLAYER CONTUSI..."	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

```
1 fm %>%
2   filter(age < 5) %>%
3   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

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

Descriptive Statistics

fm\_mk2\$age

N: 7521

age	
<hr/>	
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

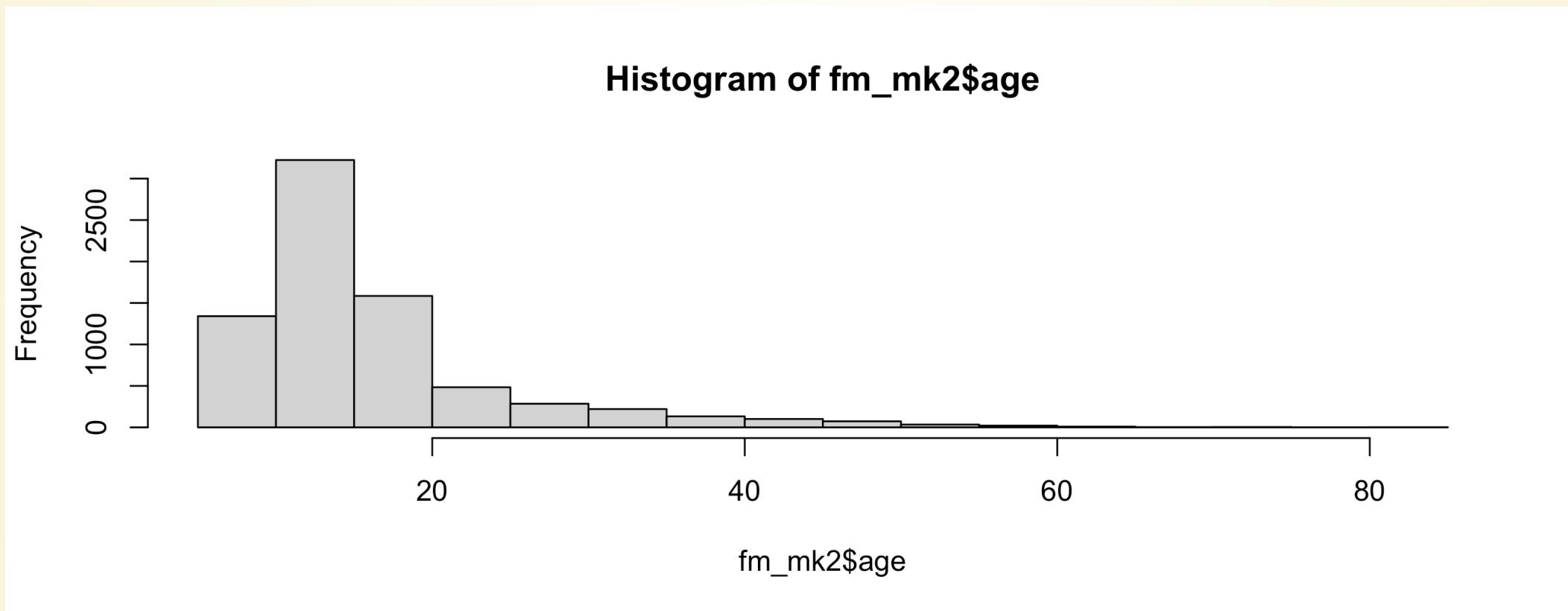
- Média de `fm` (com pequenas crianças): 16.3786225
- Média de `fm_mk2` (sem pequenas crianças): 16.5261268
- Se removêssemos 82 casos, porque a diferença não é maior?

# 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**

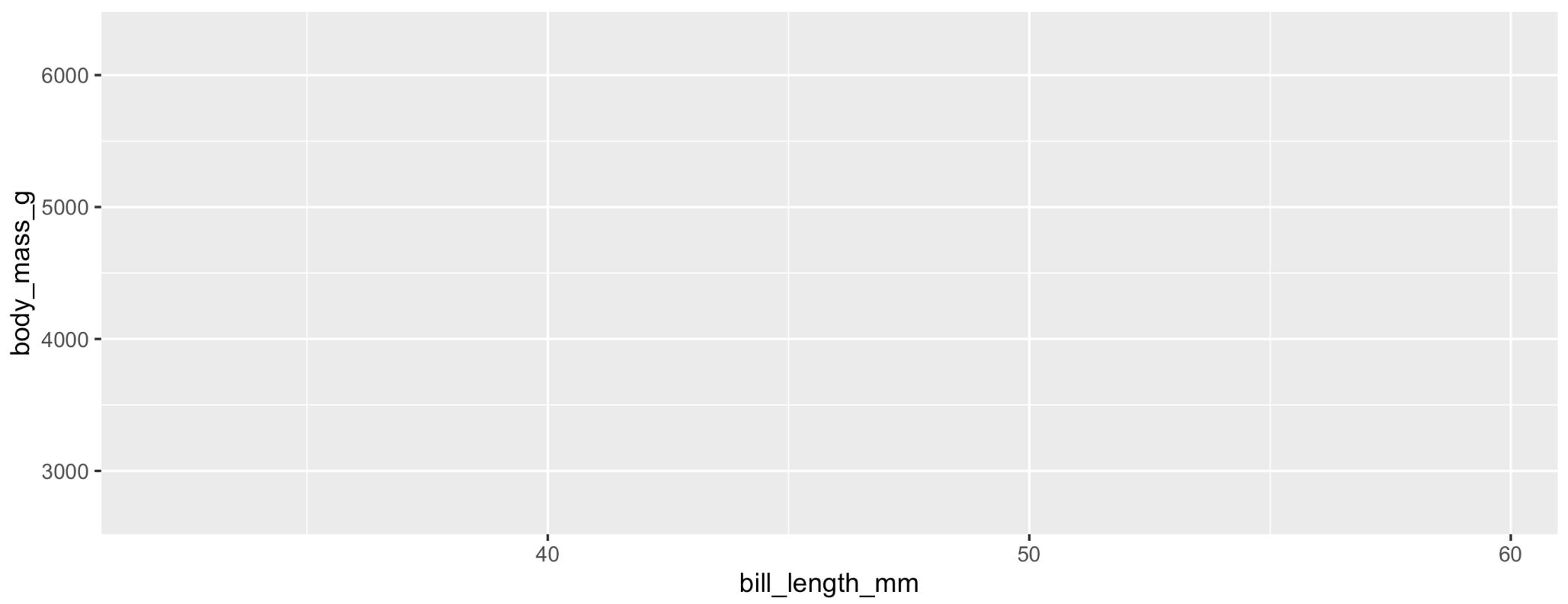
- Um sistema para **construir** gráficos (que se comunicam muito melhor)
- Um dos primeiros produtos de Hadley Wickham
- Construir seu gráfico camada por camada
- Começar por especificar um conjunto de dados: **penguin**
  - Variáveis **bill\_length\_mm** e **body\_mass\_g**

```
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...
```

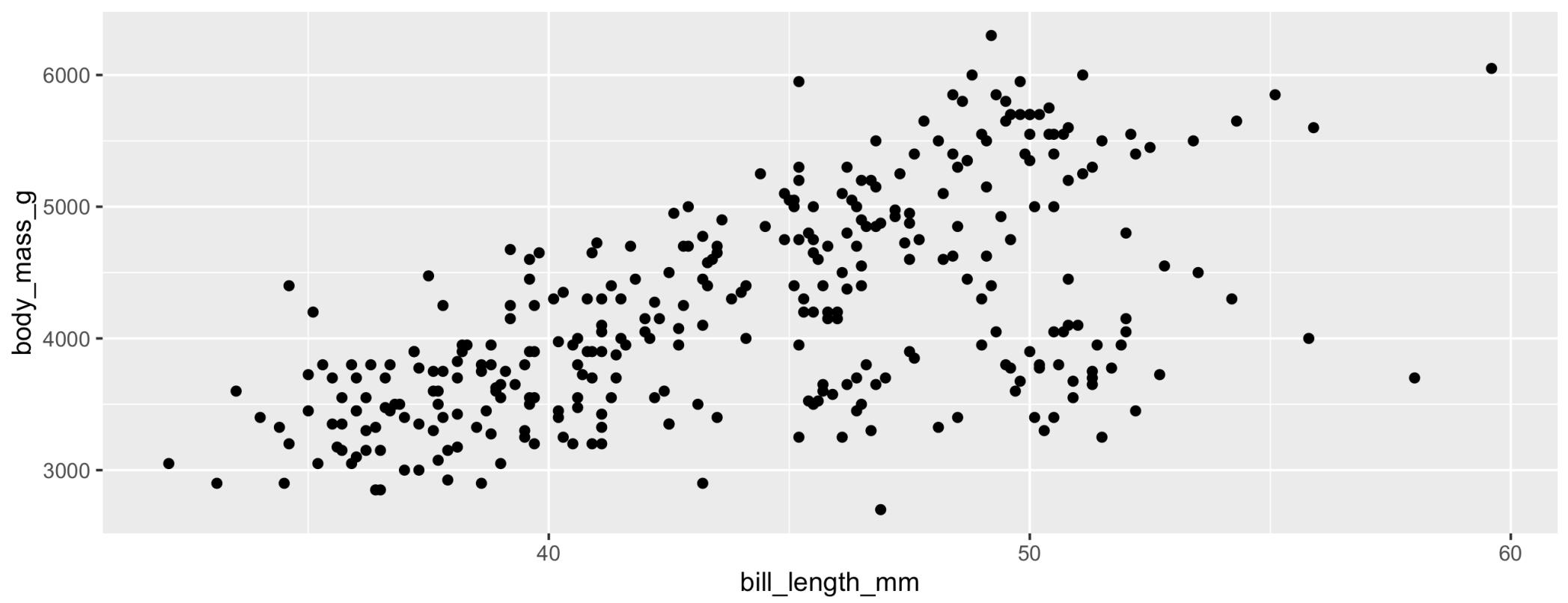
```
1 ggplot()
```

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

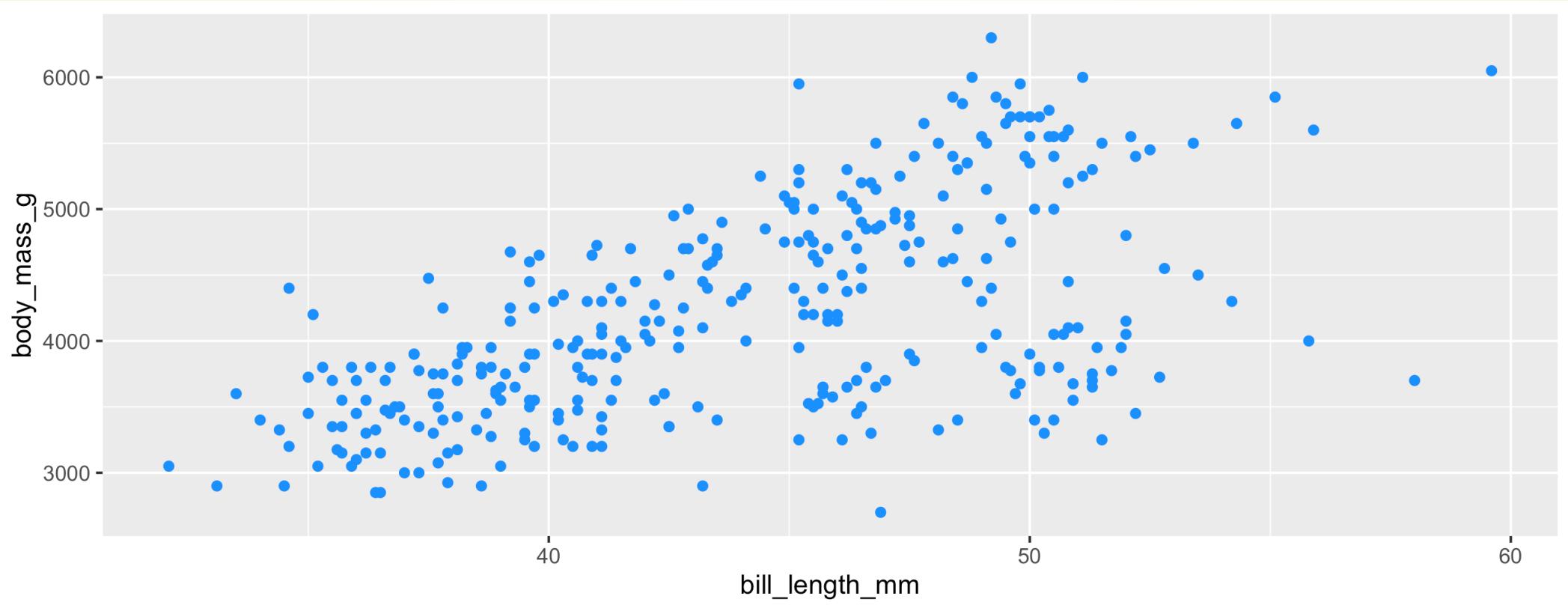
```
1 ggplot(data = pd, aes(x = bill_length_mm, body_mass_g ))
```



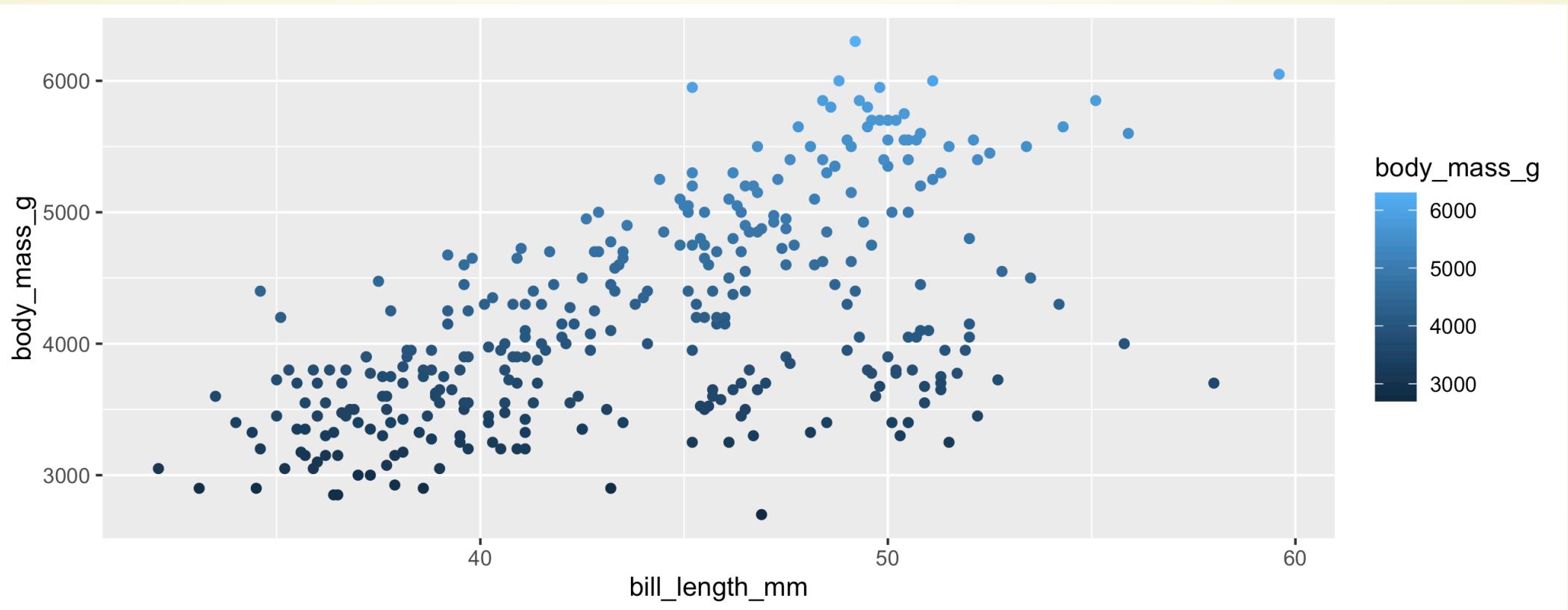
```
1 ggplot(data = pd, aes(x = bill_length_mm, body_mass_g )) +  
2   geom_point()
```



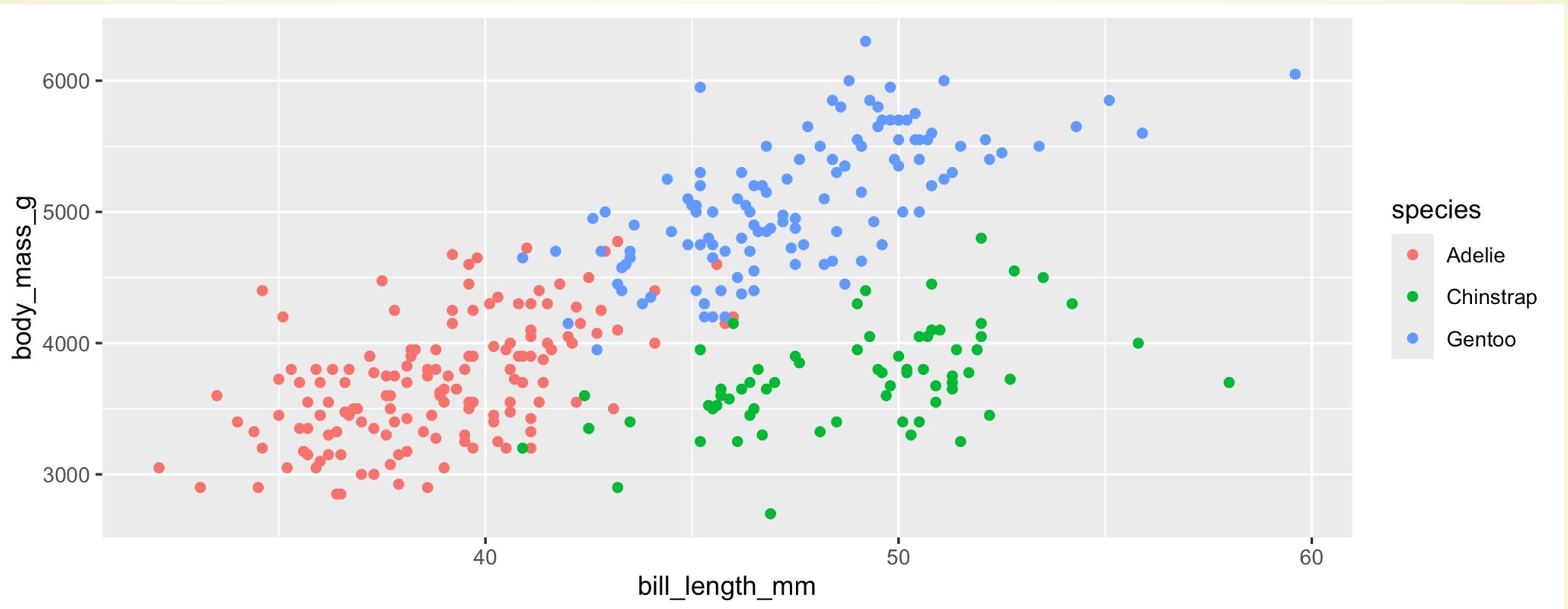
```
1 ggplot(data = pd, aes(x = bill_length_mm, body_mass_g )) +  
2   geom_point(color = "dodgerblue")
```



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



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

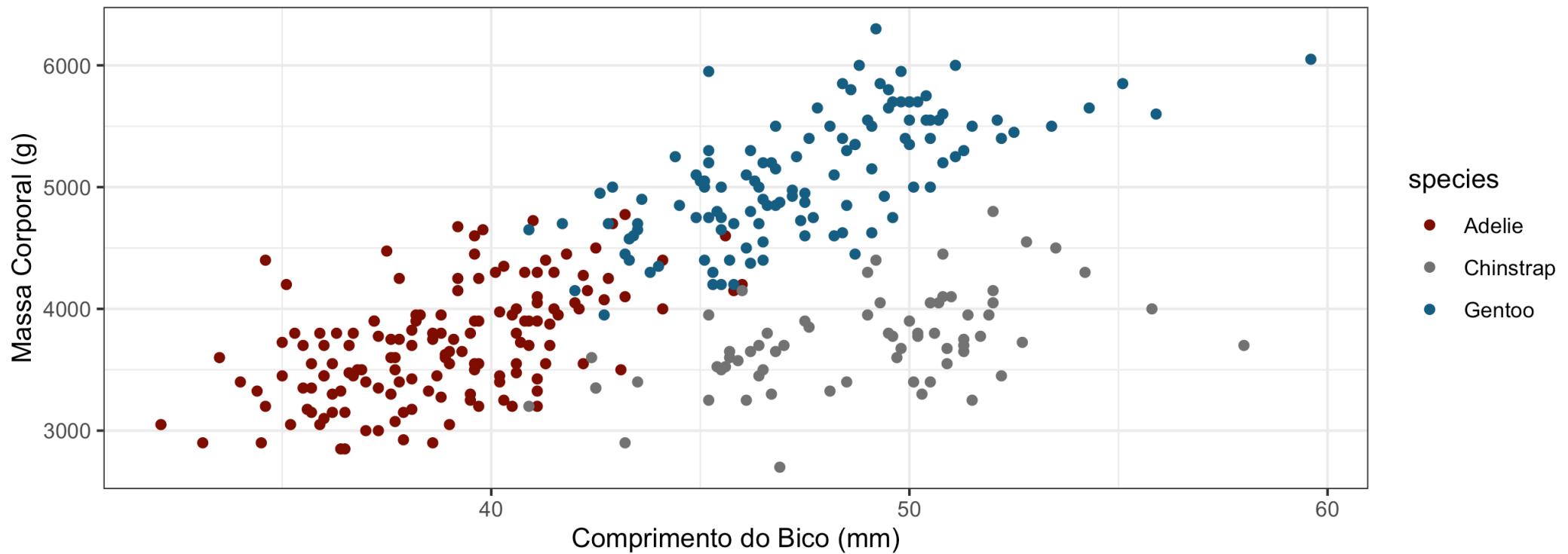


```

1 ggplot(data = pd, aes(x = bill_length_mm, body_mass_g, color = species)) +
2   geom_point() +
3   labs(title = "Uma Boa Introdução aos Gráficos de Dispersão", x = "Comprimento do Bico (mm)",
4        y = "Massa Corporal (g)") +
5   scale_colour_manual(values = c("#800000FF", "#767676FF", "#155F83FF")) +
6   theme_bw()

```

Uma Boa Introdução aos Gráficos de Dispersão

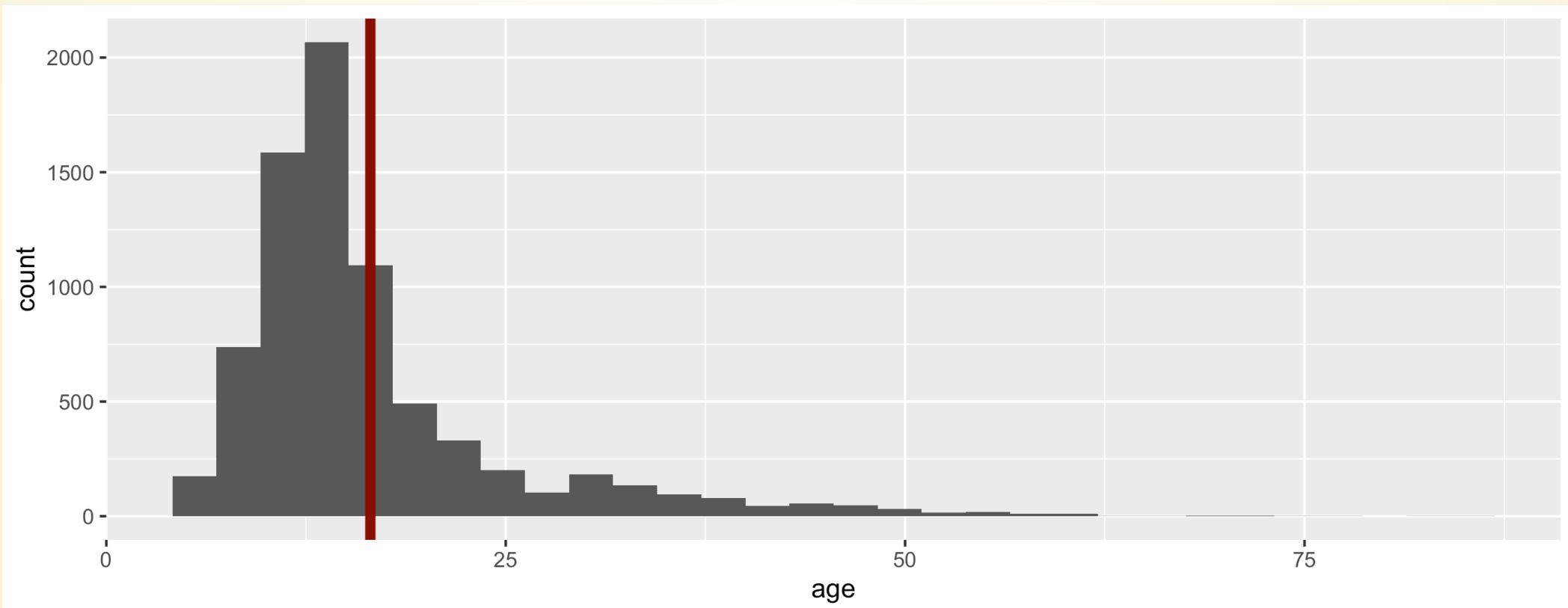


# RECURSOS - `ggplot`

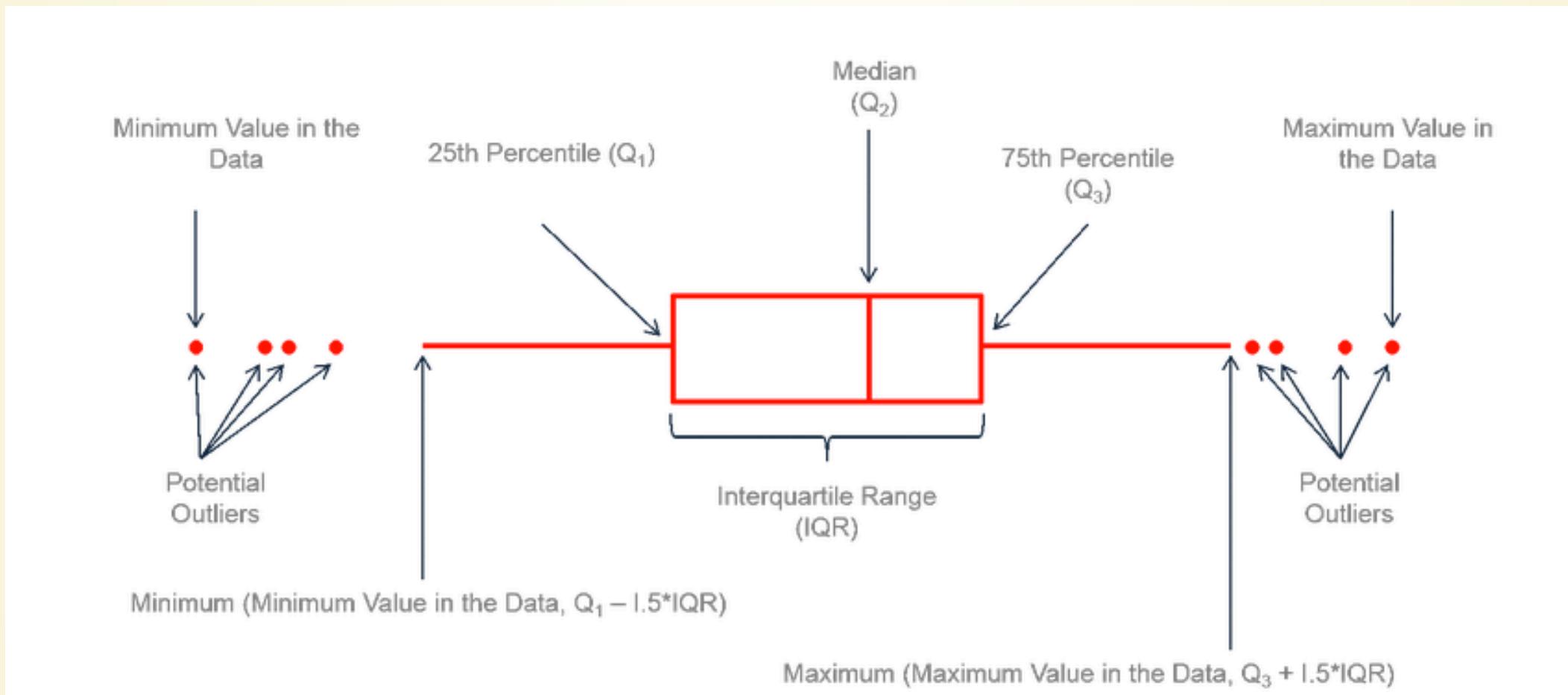
- Winston Chang, **R Graphics Cookbook**, 2Ed., <https://r-graphics.org>
- Kieran Healy, **Data Visualization: A Practical Introduction**, <https://socviz.co>
- [https://r-graph\\_gallery.com](https://r-graph_gallery.com) - examples of many types of graphs with explanations and code
- `ggplot` cheat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

# HISTOGRAMA DE age

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

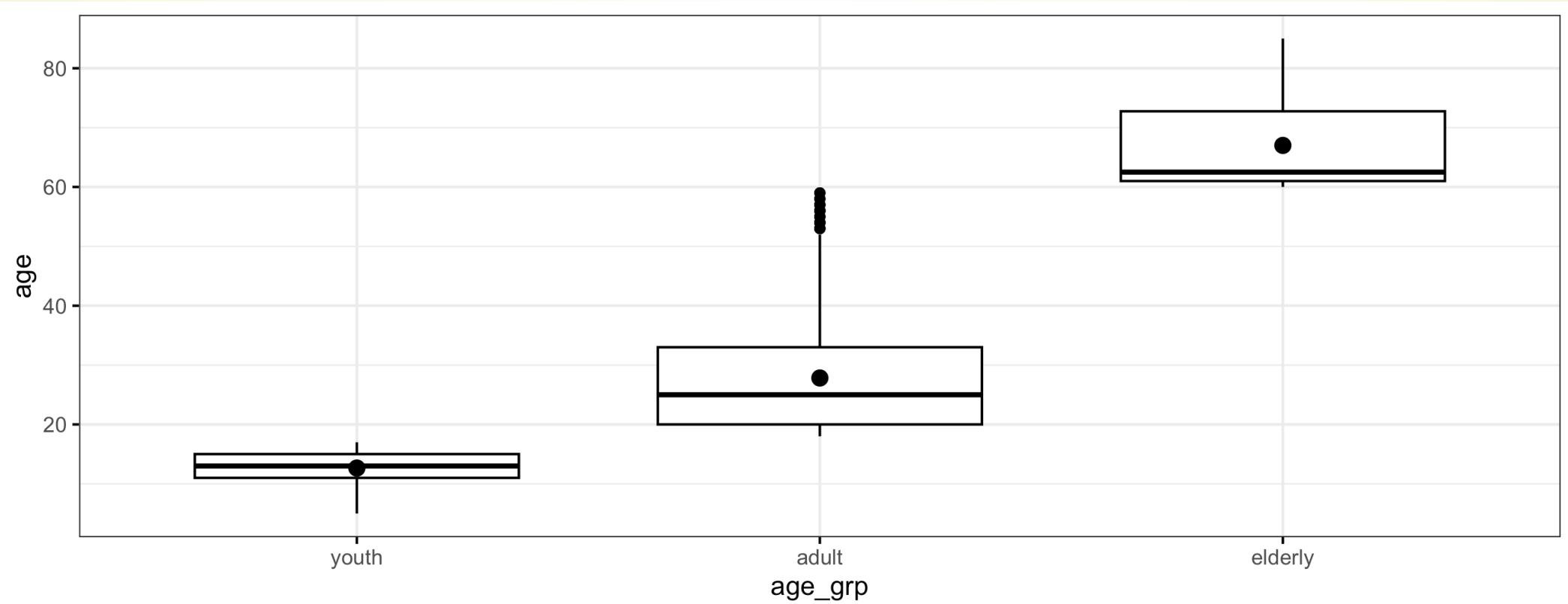


# SEGUNDO GRÁFICO QUE MOSTRA DISTRIBUIÇÕES BEM - BOXPLOT



# BOXPLOT COM OS DADOS DE FUTEBOL

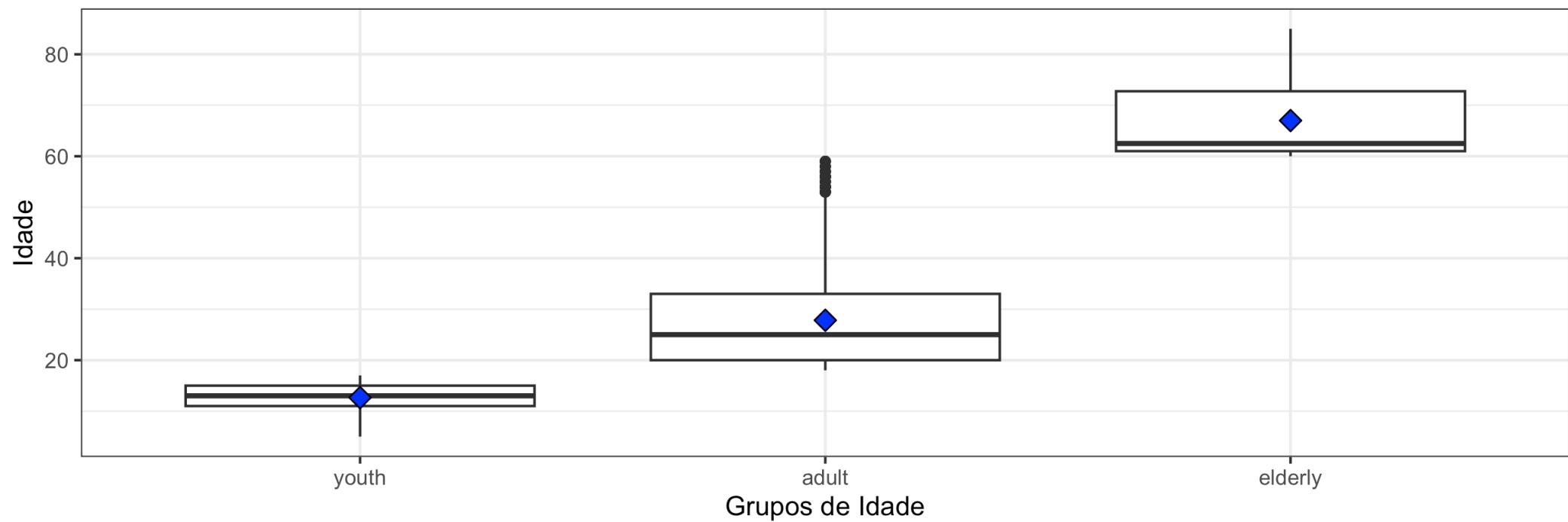
```
1 fm_mk2 |> ggpubr::ggboxplot(x = "age_grp",
2                               y = "age",
3                               add = "mean",
4                               ggtheme = theme_bw())
```



# BOXPLOT DE FUTEBOL COM ggplot

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

## Idades por Grupo de Idade



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.

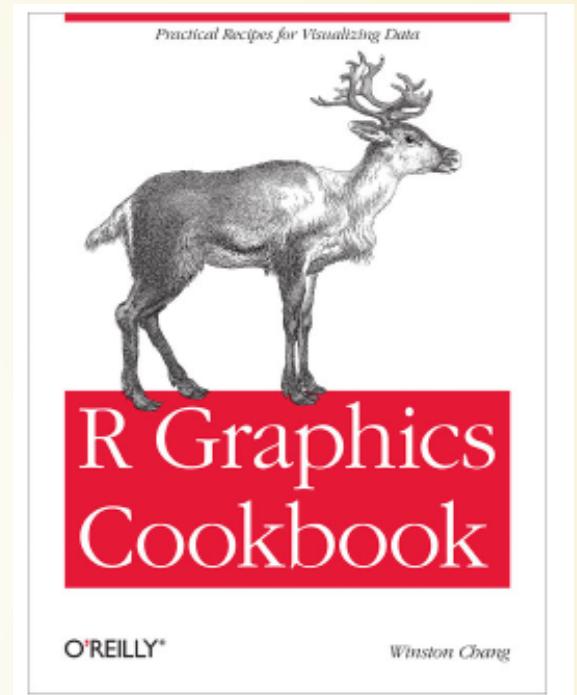
## 6.8.2 Solution

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.
```

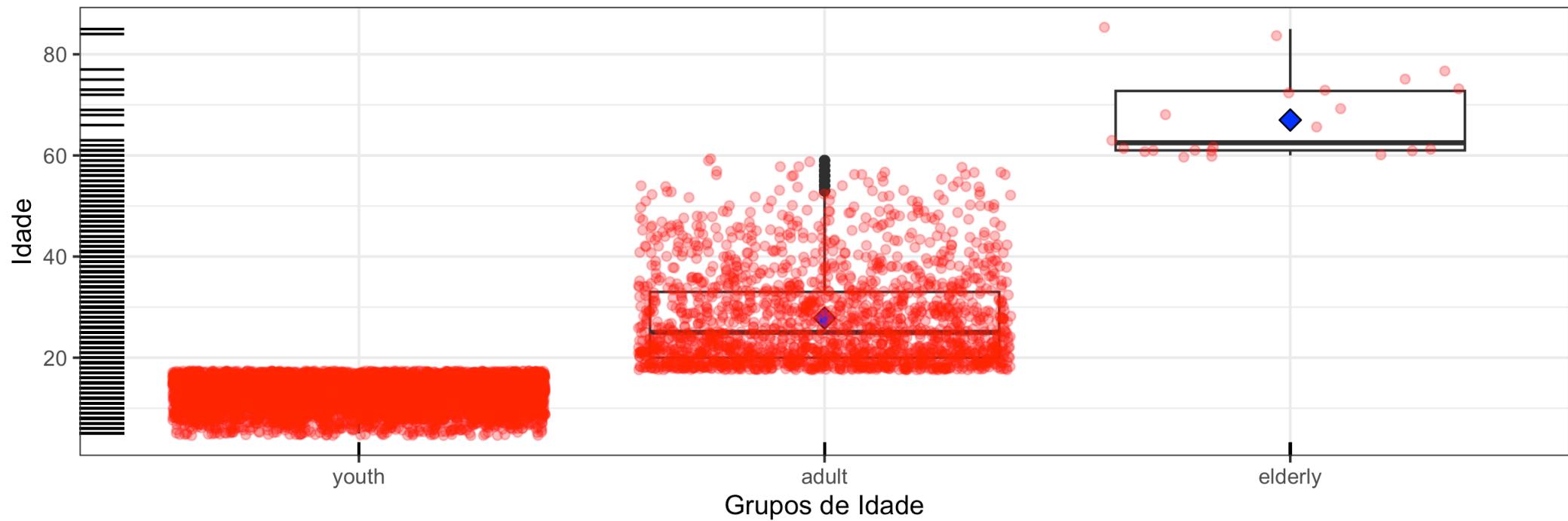


# ggplot *BOXPLOT*- 2

- 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

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

## Idades por Grupo de Idade



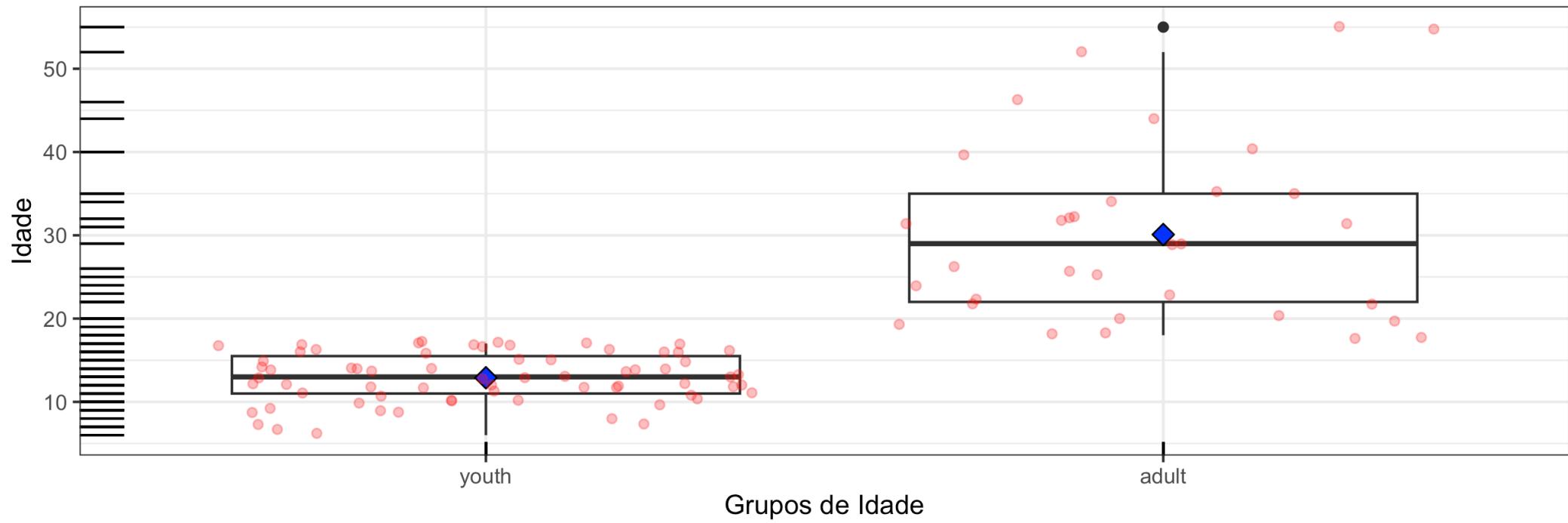
Texto que explica o gráfico.

# LIÇÃO DISSO

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

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

## Idades por Grupo de Idade



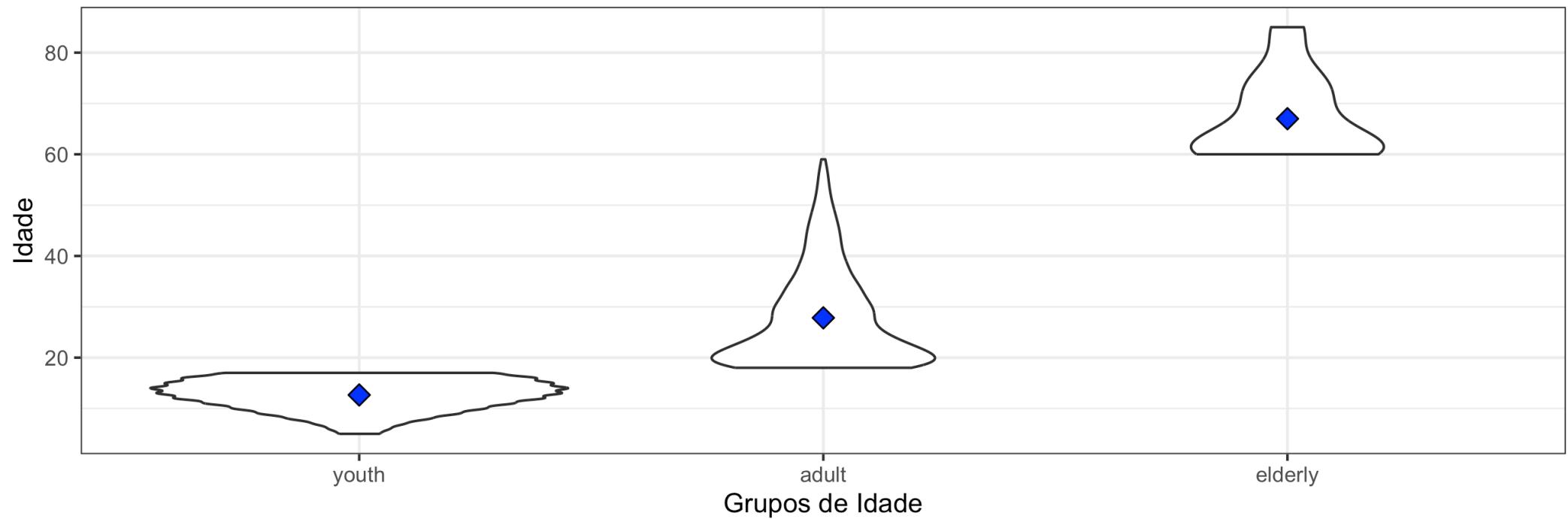
Texto que explica o gráfico.

# ALTERNATIVA DE BOXPLOT – GRÁFICO DE VIOLETO

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

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

## Idades por Grupo de Idade



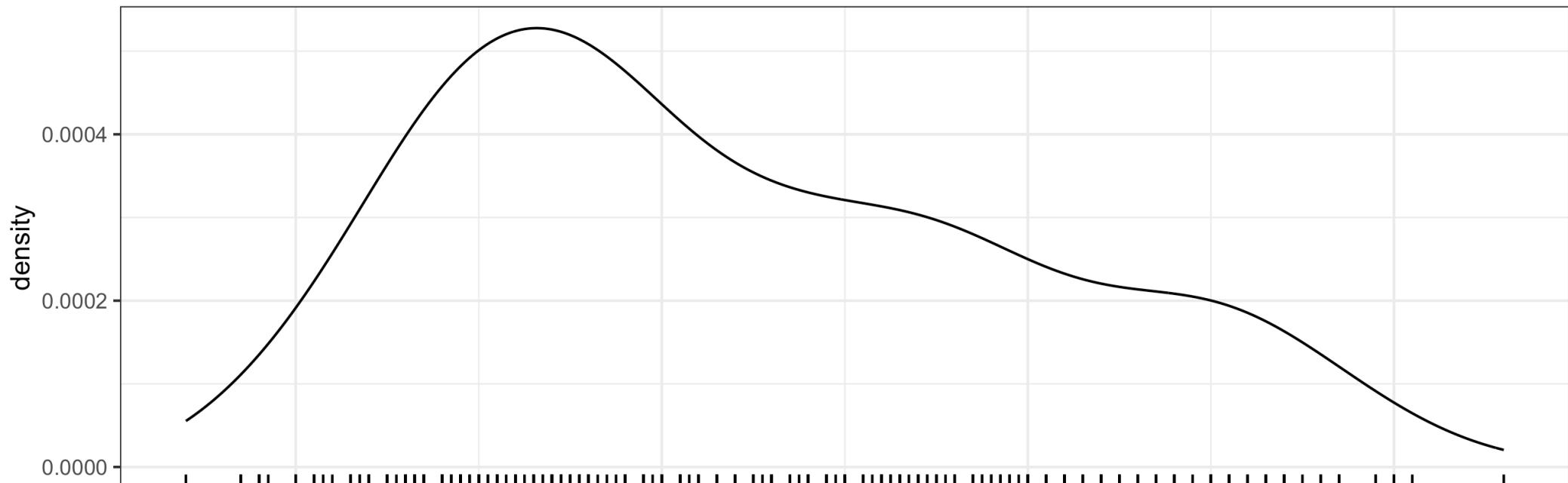
Texto que explica o gráfico.

# 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

```
1 set.seed(42)
2 pd |> # dataframe penguins |>
3   ggplot(mapping = aes(x = body_mass_g)) +
4   geom_density() +
5   labs(title = "Massa Corporal",
6        x = "") +
7   geom_rug()+
8   theme_bw()
```

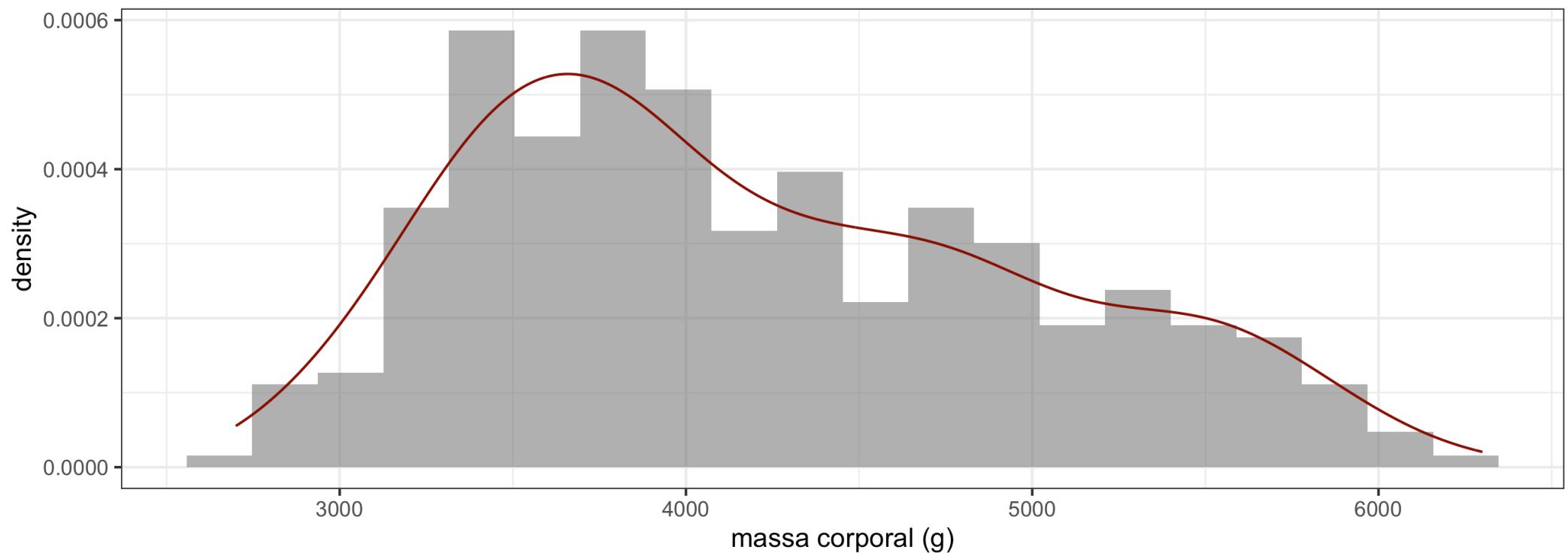
## Massa Corporal



# DENSIDADE COM HISTOGRAMA

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

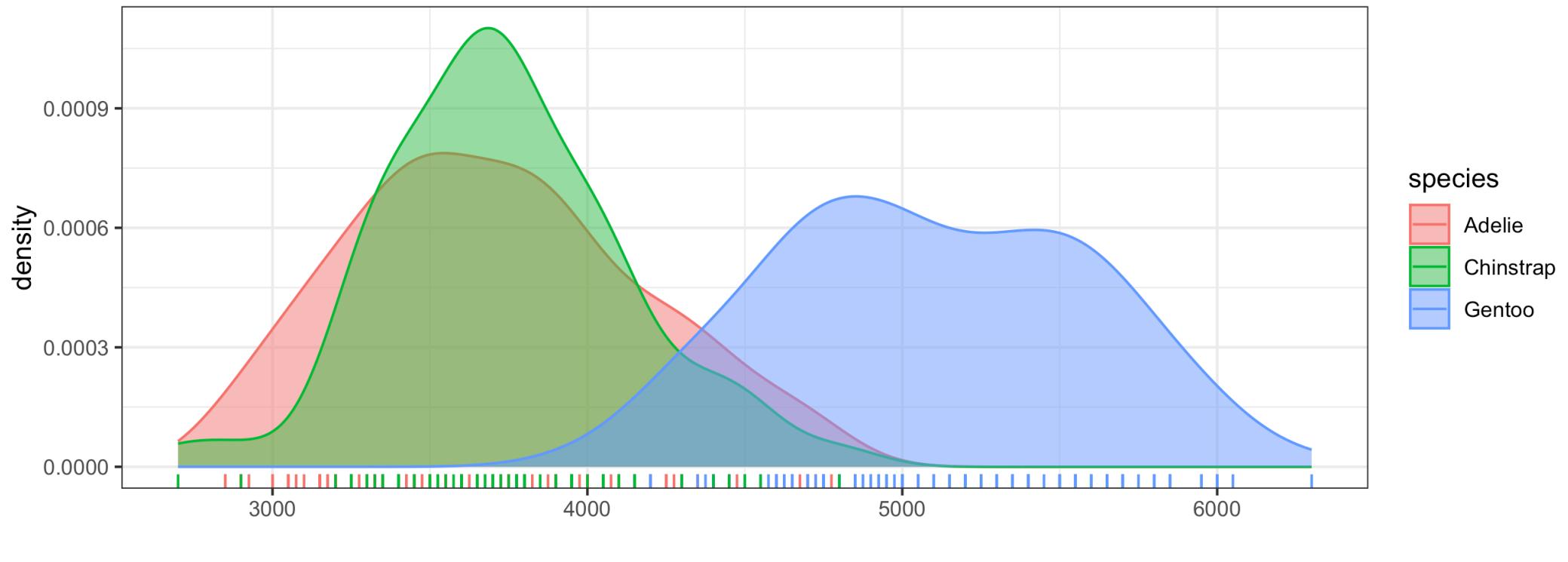
## Massa Corporal



# COMPARAR AS ESPÉCIES

```
1 set.seed(42)
2 pd |> # dataframe penguins |>
3   group_by(species) |>
4   ggplot(mapping = aes(x = body_mass_g, colour = species, fill = species))
5   geom_density(alpha = 0.5) +
6   labs(title = "Massa Corporal",
7        x = "") +
8   geom_rug()+
9   theme_bw()
```

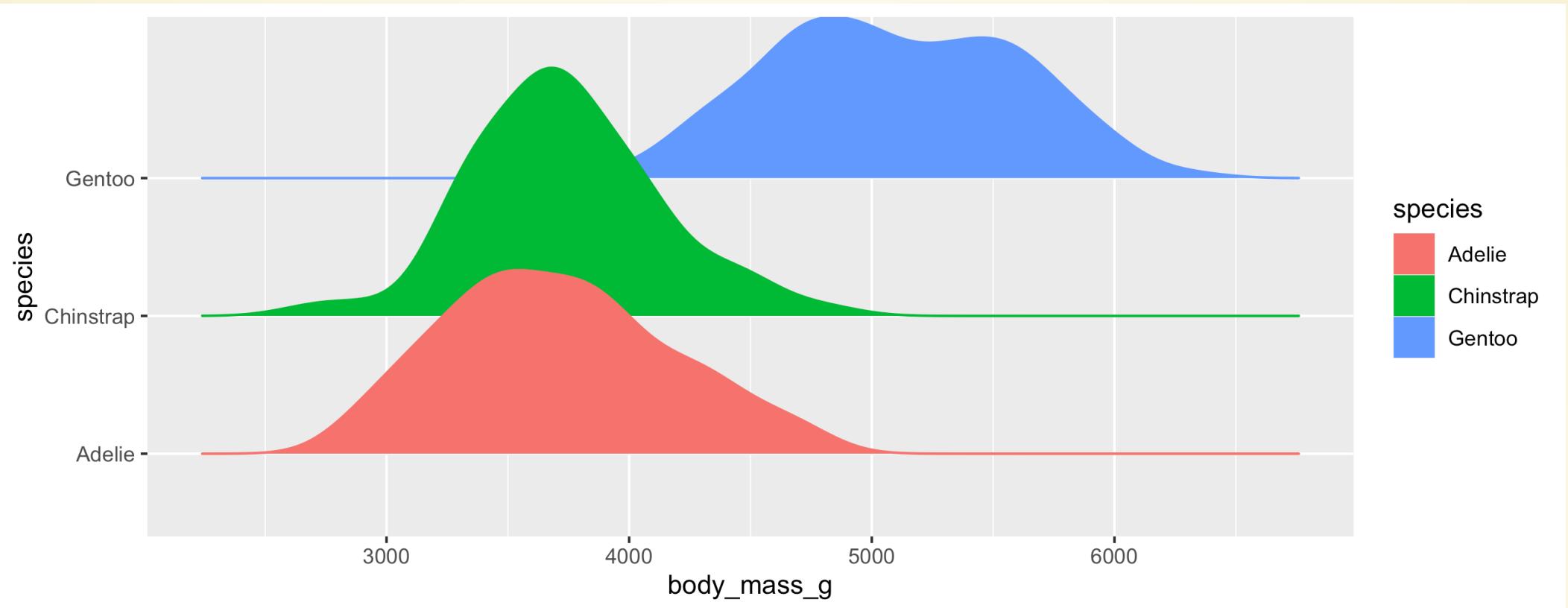
## Massa Corporal



# GRÁFICO *RIDGELINE*- EXTENSÃO DE DENSIDADE

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

```
1 library(ggridges)
2
3 pd |>
4   ggplot(aes(x = body_mass_g,
5             y = species,
6             colour = species,
7             fill = species)) +
8   geom_density_ridges() +
9   theme_gray() # default theme for ggplot
```

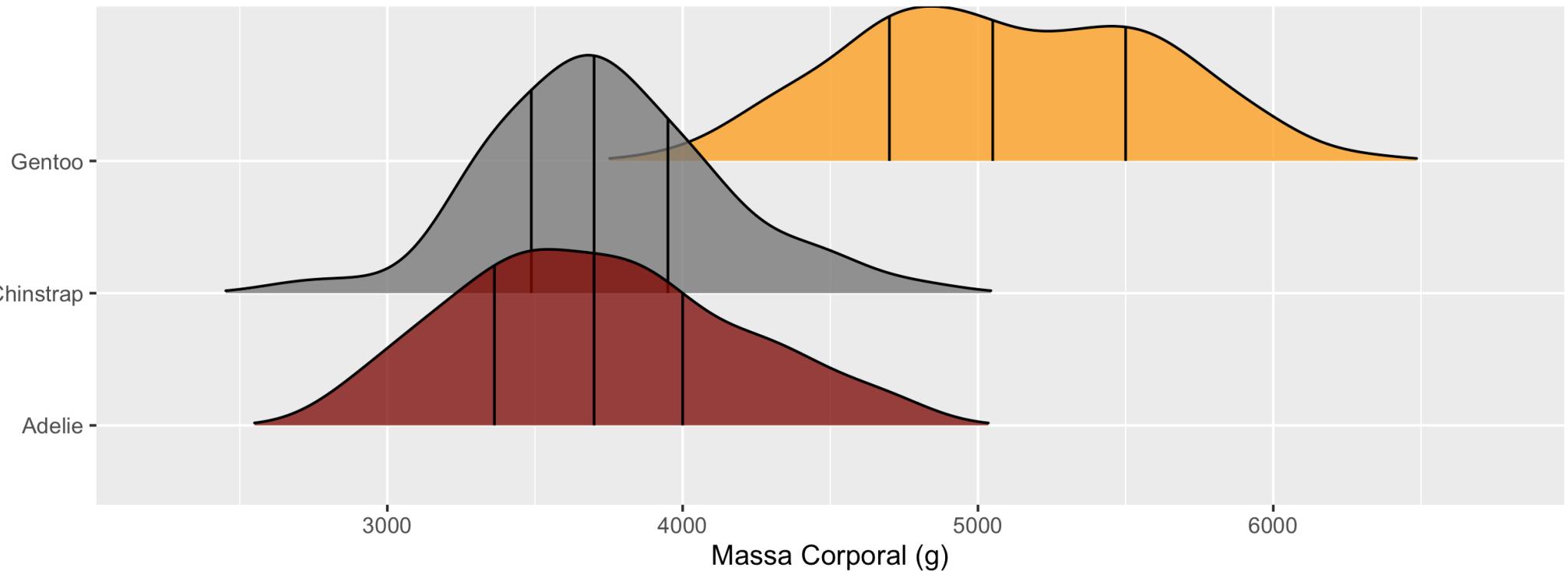


# RIDGE LINE COM MAIS OOMPH

- 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

```
1 library(ggridges)
2
3 pd |>
4   ggplot(aes(x = body_mass_g,
5             y = species,
6             fill = species)) +
7   stat_density_ridges(quantile_lines = TRUE, rel_min_height = 0.01) +
8   scale_fill_uchicago(palette = "default", alpha = 0.8) +
9   guides(fill = FALSE) +
10  labs(title = "Massa Corporal por Espécie",
11        x = "Massa Corporal (g)",
12        y = "") +
13  theme_gray() # default theme for ggplot
```

## Massa Corporal por Espécie



# **GRÁFICO RADAR - OUTRA MANEIRA DE RETRATAR DIMENSÕES**

# 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

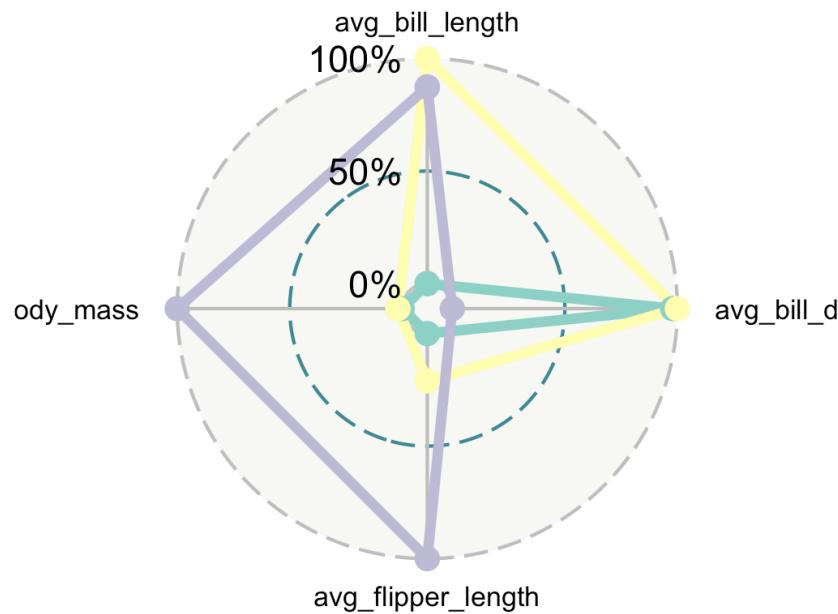
```
1 ggpengrad <- pd_radar %>%
2   ggradar(
3     font.radar = "arial",
4     grid.label.size = 5, # Affects the grid annotations (0%, 50%, etc.)
5     axis.label.size = 3, # Afffects the names of the variables
6     group.point.size = 3, # Simply the size of the point
7     legend.title = "Espécie",
8     plot.title = "Características - Pinguins Palmer",
9   )
```

# RESULTADO

## Características - Pinguins Palmer

### Espécie

- Adelie
- Chinstrap
- Gentoo



# ÚLTIMO TIPO DE HOJE

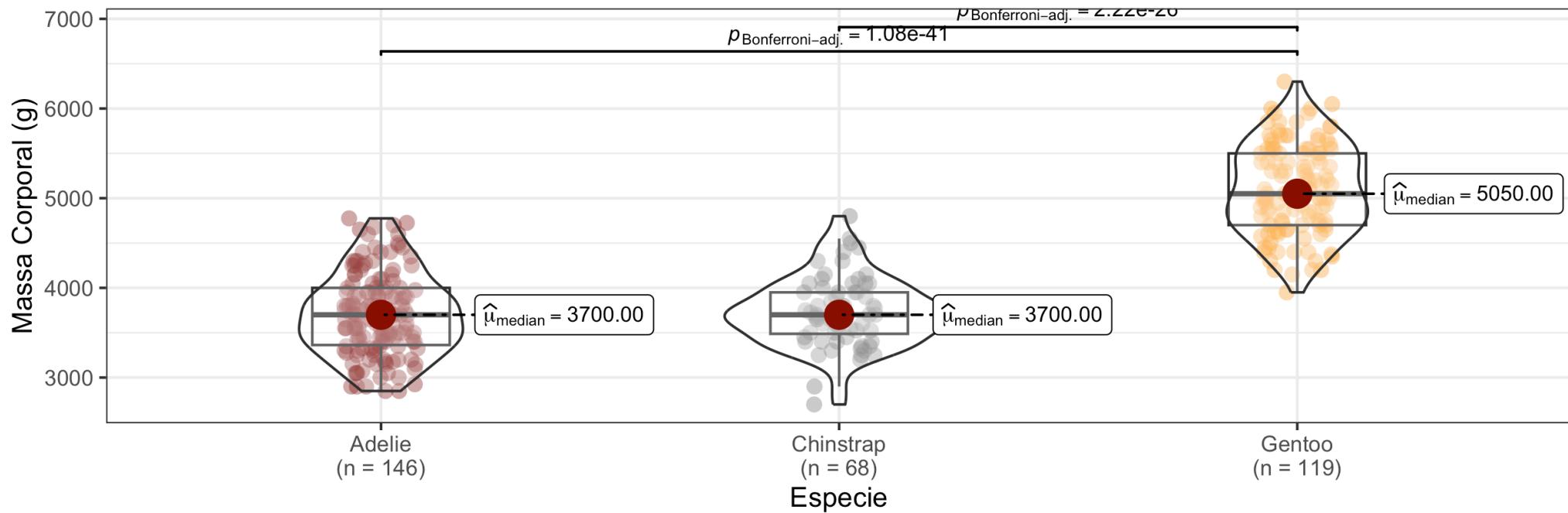
# INCORPORAR GRÁFICOS COM TESTES ESTATÍSTICOS

- Pacote `ggstatsplot` - precisa instalar da CRAN
- Pacote tem muitas combinações por vários testes
- Diferença entre a média de massa corporal das espécies é significativa?

```
1 library(ggstatsplot)
2 pd |>
3   ggbetweenstats(
4     x = species,
5     y = body_mass_g,
6     type = "nonparametric",
7     p.adjust.method = "bonferroni",
8     xlab = "Especie",
9     ylab = "Massa Corporal (g)",
10    title = "Massa Corporal entre Especies",
11    ggtheme = theme_bw(),
12    package = "ggsci",
13    palette = "default_uchicago"
14  )
```

## Massa Corporal entre Especies

$\chi^2_{\text{Kruskal-Wallis}}(2) = 212.09, p = 8.84e-47, \hat{\varepsilon}_{\text{ordinal}}^2 = 0.64, \text{CI}_{95\%} [0.61, 1.00], n_{\text{obs}} = 333$



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