

Probabilidade e Estatística com R

Fernando Náufel

(versão de 02/03/2022)

Sumário

Apresentação	3
Referências recomendadas	5
Exercício	5
1 O Que É Estatística?	6
1.1 Vídeo 1	6
1.2 Exercícios	6
1.3 Vídeo 2	7
1.4 Exercícios	9
2 Introdução a R	10
2.1 Vídeo 1	10
2.2 Vídeo 2	10
2.3 Exercícios	10
3 Introdução ao tidyverse	12
3.1 Criando uma <i>tibble</i>	12
3.2 Operador de <i>pipe</i> (<i>%>%</i>)	14
3.3 Formato <i>tidy</i>	14
3.4 Manipulando os dados	19
4 Visualização com ggplot2	33
4.1 Vídeo 1	33
4.2 Componentes de um gráfico ggplot2	33
4.3 Conjunto de dados	36
4.4 Gráficos de dispersão (<i>scatter plots</i>)	42
4.5 Vídeo 2	59
4.6 Histogramas e cia.	59
4.7 Ogiva	64
4.8 Ramos e folhas	65
4.9 Personalização do tema	66

4.10 Exercícios	69
5 Visualização com ggplot2 (continuação)	73
5.1 Vídeo 1	73
5.2 <i>Boxplots</i>	73
5.3 Vídeo 2	83
5.4 Gráficos de barras e de colunas	83
5.5 Gráficos de linha e séries temporais	97
5.6 Exercícios	101
5.7 Referências sobre visualização e R	102
6 Medidas	103

Apresentação

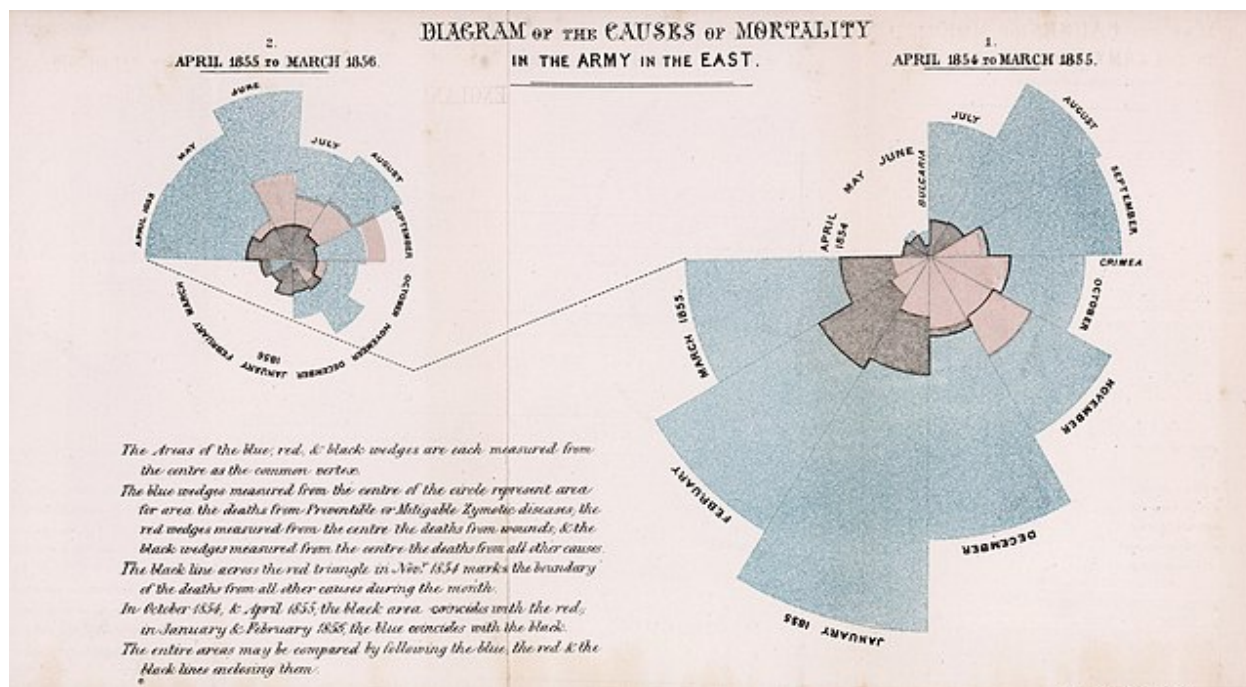
Atenção



Este material ainda está em construção.

Pode haver mudanças a qualquer momento.

Verifique, no rodapé da página *web* ou na capa do arquivo pdf, a data desta versão.



Este livro/site foi iniciado em 2020, durante a pandemia de COVID-19, quando a Universidade Federal Fluminense (UFF) funcionou em regime de ensino remoto durante mais de um ano.

Para atender os alunos do curso de Probabilidade e Estatística do curso de graduação em Ciência da Computação da UFF, decidi gravar aulas em vídeo e disponibilizar os arquivos usados nelas. Foram esses arquivos que deram origem a este livro/site.

Este livro/site foi construído para pessoas que já saibam programar, embora não necessariamente em R.

Para tirar o máximo proveito deste material, você deve fazer o seguinte:

1. Assistir aos vídeos contidos em cada capítulo. A *playlist* completa está em <https://www.youtube.com/playlist?list=PL7SRLwLs7ocaV-Y1vrVU3W7mZnnS0qkWW>.
2. Instalar o R no seu computador ou abrir uma conta no RStudio Cloud, para poder usar o R *online*. Você encontra instruções para fazer isto no [capítulo de introdução a R](#).
3. Baixar, [neste repositório do Github](#), o código-fonte deste livro/site, para poder rodar e alterar os exemplos.
4. Seguir os *links* para outras fontes *online* que abordam assuntos que não são cobertos em detalhes neste curso.
5. Fazer os exercícios. Ao longo do tempo, acrescentarei *links* para vídeos explicando as soluções.



Se você estiver lendo este material na web, você pode clicar nos comandos e funções que aparecem nos blocos de código em R para abrir páginas da documentação sobre eles.

Se você preferir ler este livro em pdf, ou se quiser imprimi-lo, [faça o download do arquivo aqui](#).

Referências recomendadas

Em português

- Sillas Gonzaga, *Introdução a R para Visualização e Apresentação de Dados*, http://sillasgonzaga.com/material/curso_visualizacao/index.html
- Allan Vieira de Castro Quadros, *Introdução à Análise de Dados em R utilizando Tidyverse*, https://allanvc.github.io/book_IADR-T/
- Paulo Felipe de Oliveira, Saulo Guerra, Robert McDonnel, *Ciência de Dados com R – Introdução*, <https://cdr.ibpad.com.br/index.html>
- Curso R, *Ciência de Dados em R*, <https://livro.curso-r.com/>

Em inglês

- Garrett Grolemund, Hadley Wickham, *R for Data Science*, <https://r4ds.had.co.nz/>
- Chester Ismay, Albert Y. Kim, *A ModernDive into R and the Tidyverse*, <https://moderndive.com/>

Exercício

1. Pesquise sobre a imagem do início deste capítulo. Ela foi criada em 1858 por Florence Nightingale.

O Que É Estatística?

1.1

Vídeo 1

https://youtu.be/6Q_XSoLCIpc

1.2

Exercícios

1. Você está interessado em estimar a altura de todos os homens da sua faculdade. Para isso, você decide medir as alturas de todos os homens da sua turma de Estatística.
 - Qual é a amostra?
 - Qual é a população?
2. Um instituto de pesquisa entrevista um grupo de 1000 pessoas, perguntando a cada uma se ela vai votar a favor do candidato A na próxima eleição. Dos entrevistados, 600 responderam que sim. A proporção 0,6 (ou 60%) é uma estatística ou um parâmetro?
3. Você vê alguma diferença entre as cinco situações abaixo? Quais das situações são equivalentes em termos da probabilidade de conseguir 10 cartas do mesmo naipe?
 - a. Usando um baralho normal, você retira 10 cartas e registra as cartas retiradas.

- b. Usando um baralho normal, você repete a seguinte sequência de ações 10 vezes: retirar uma carta do baralho, registrar a carta retirada e repor a carta no baralho.
 - c. Usando uma caixa contendo todas as cartas de 1 milhão de baralhos reunidos, você retira 10 cartas e registra as cartas retiradas.
 - d. Usando uma caixa contendo todas as cartas de 1 milhão de baralhos reunidos, você repete a seguinte sequência de ações 10 vezes: retirar uma carta da caixa, registrar a carta retirada e repor a carta na caixa.
 - e. Usando um baralho *infinito*, você retira 10 cartas e registra as cartas retiradas.
 - f. Usando um baralho *infinito*, você repete a seguinte sequência de ações 10 vezes: retirar uma carta do baralho, registrar a carta retirada e repor a carta no baralho.
4. Qual a graça dos quadrinhos na Figura 1.1, que também [aparecem no vídeo](#)?



Figura 1.1: <http://xkcd.com/552/>

5. Qual a graça dos quadrinhos na Figura 1.2?
6. Veja este vídeo sobre o cavalo Hans:

<https://youtu.be/G3VkCmdUfZE>

Qual a relação entre esta história e a necessidade de duplo cegamento?

1.3

Vídeo 2

<https://youtu.be/492VASxIDRo>



Figura 1.2: <http://xkcd.com/1462/>

1.4

Exercícios

1. Por que não faz sentido calcular a média dos CEPs de um grupo de pessoas?
2. Uma temperatura de -40 graus Celsius é igual a uma temperatura de -40 graus Fahrenheit?
3. Uma temperatura de zero graus Celsius é igual a uma temperatura de zero graus Fahrenheit?
4. Uma variação de temperatura de 1 grau Celsius é igual a uma variação de temperatura de 1 grau Fahrenheit?
5. Um saldo bancário de zero reais é igual a um saldo bancário de zero dólares?
6. Um produto de 1 milhão de reais custa o mesmo que um produto de 1 milhão de dólares?
7. Meses representados por números de 1 a 12 são dados de que nível?

Introdução a R

2.1

Vídeo 1

<https://youtu.be/1kXQDNqm41c>

2.2

Vídeo 2

<https://youtu.be/3GEc1oiKDrU>

2.3

Exercícios

1. Para criar sua conta no RStudio Cloud, acesse <https://rstudio.cloud/>.
2. Se você preferir instalar o R no seu computador, acesse
 - <https://cran.r-project.org/> para baixar e instalar o R, e
 - <https://rstudio.com/products/rstudio/download/> para baixar e instalar o RStudio, um IDE específico para R.

3. Abra o RStudio Cloud ou o seu RStudio instalado localmente.
4. Crie um novo projeto. **Sempre trabalhe em projetos para ter seus arquivos organizados.**
5. Para instalar o `swirl` (pacote do R para exercícios interativos), execute o seguinte comando no console do RStudio:

```
install.packages("swirl")
```

6. Para instalar os exercícios de introdução a R, execute os seguintes comandos no console do RStudio:

```
library(swirl)  
install_course_github('fnaufel', 'introR')
```

7. Mude o idioma para português e execute o `swirl`.

```
select_language('portuguese', append_rprofile = TRUE)  
swirl()
```

8. Na primeira execução, você vai precisar se identificar (qualquer nome serve). Com essa identificação, o `swirl` vai registrar o seu progresso nas lições.
9. No `swirl`, as perguntas são mostradas no console. Você também deve responder no console.
10. Às vezes, um *script* será aberto no editor de textos para que você complete um programa. Quando seu programa estiver pronto, salve o arquivo e digite `submit()` no console para o `swirl` processar o *script*.
11. O `swirl` dá instruções claras no console. Na dúvida, digite `info()` no *prompt* do R (`>`).
12. Se, em vez do *prompt* do R, o console mostrar reticências (`...`), tecla *Enter*.
13. Se nada funcionar, tecla *ESC*.
14. Para sair do `swirl()`, digite `bye()` no *prompt* do R.
15. Para voltar para os exercícios, digite

```
library(swirl)  
swirl()
```

Introdução ao tidyverse



Busque mais informações sobre os pacotes que compõem o tidyverse [nas referências recomendadas](#).

3.1

Criando uma *tibble*

- Cada coluna é um vetor:

```
cores <- tibble(  
  pessoa = c('João', 'Maria', 'Pedro', 'Ana'),  
  'cor favorita' = c('azul', 'rosa', 'preto', 'branco')  
)  
  
cores  
## # A tibble: 4 x 2  
##   pessoa `cor favorita`  
##   <chr>   <chr>  
## 1 João   azul  
## 2 Maria  rosa  
## 3 Pedro  preto  
## 4 Ana    branco
```

- A função `tribble` permite a entrada de forma mais natural. **Lembre-se de usar `~` antes dos nomes das colunas.**

```
cores <- tribble(
  ~pessoa, ~'cor favorita',
  "João",   "azul",
  "Maria",  "rosa",
  "Pedro",  "preto",
  "Ana",    "branco"
)

cores
## # A tibble: 4 x 2
##   pessoa `cor favorita`
##   <chr>  <chr>
## 1 João   azul
## 2 Maria  rosa
## 3 Pedro  preto
## 4 Ana    branco
```

- Se uma coluna não puder ser armazenada em um vetor, a coluna será uma lista:

```
cores <- tibble(
  pessoa = c('João', 'Maria', 'Pedro', 'Ana'),
  'cor favorita' = list(
    c('azul', 'roxo'),
    c('rosa', 'magenta'),
    NA,
    'branco'
  )
)

cores
## # A tibble: 4 x 2
##   pessoa `cor favorita`
##   <chr>  <list>
## 1 João   <chr [2]>
## 2 Maria <chr [2]>
## 3 Pedro <lgl [1]>
## 4 Ana   <chr [1]>
```

- Use `View()` para examinar interativamente o conteúdo de uma coluna-lista:

```
cores %>% View()
```

3.2

Operador de *pipe* (%>%)

- O tidyverse inclui o pacote `magrittr`¹, que contém este operador.
- A ideia é facilitar a leitura de **composições de comandos**. O código

```
y <- h(g(f(x)))
```

pode ser escrito como

```
y <- x %>% f() %>% g() %>% h()
```

- Esta segunda versão é mais fiel à ordem em que as operações acontecem.
- Na verdade, R tem um operador de **atribuição para a direita**, mas poucas pessoas recomendam usá-lo:

```
x %>% f() %>% g() %>% h() -> y
```

- Se `f`, `g` e `h` forem funções de um argumento só, os parênteses podem ser omitidos:

```
y <- x %>% f %>% g %>% h
```

- Se a função `f` tiver outros argumentos, escreva-os normalmente na chamada a `f`:

```
y <- x %>% mean(na.rm = TRUE)
```

- O `pipe` `EXP %>% f(...)` sempre insere o resultado da expressão `EXP` do lado esquerdo como o **primeiro argumento da função `f`**.
- Se você precisar que o resultado da expressão `EXP` seja inserido em outra posição na lista de argumentos de `f`, use um ponto “.” para isso:

```
x %>% consultar(df, .)
```

3.3

Formato *tidy*

- Nossa última versão da `tibble` `cores` é um pouco mais complexa do que deveria ser:

```
cores  
## # A tibble: 4 x 2
```

¹Por que o nome do pacote e o nome do operador (*pipe*) formam um trocadilho?

```
##  pessoa `cor favorita`
##  <chr> <list>
## 1 João  <chr [2]>
## 2 Maria <chr [2]>
## 3 Pedro <lgl [1]>
## 4 Ana   <chr [1]>
```

- O formato *tidy* exige que
 1. Cada linha da *tibble* corresponda a uma observação sobre um indivíduo,
 2. Cada coluna corresponda a uma variável observada, e
 3. Cada célula contenha um valor da variável.
- Na *tibble* *cores*, a primeira e a segunda exigências são satisfeitas, mas a terceira não, pois algumas células contêm valores múltiplos.
- A *tibble* não está no formato *tidy*.
- Podemos “extrair” estes vetores “aninhados” usando o comando `unnest`, do pacote *tidyr*:

```
cores <- cores %>%
  unnest(`cor favorita`)

cores
## # A tibble: 6 x 2
##  pessoa `cor favorita`
##  <chr> <chr>
## 1 João  azul
## 2 João  roxo
## 3 Maria rosa
## 4 Maria magenta
## 5 Pedro <NA>
## 6 Ana   branco
```

- A maioria das funções do *tidyverse* exige que as *tibbles* estejam neste formato *tidy*.
- Um exemplo mais complexo é o *dataset* *billboard*:

```
billboard %>% glimpse()
## Rows: 317
## Columns: 79
## $ artist      <chr> "2 Pac", "2Ge+her", "3 Doors Down", "3 Doors Dow~
## $ track       <chr> "Baby Don't Cry (Keep...", "The Hardest Part Of ~
## $ date.entered <date> 2000-02-26, 2000-09-02, 2000-04-08, 2000-10-21,~
## $ wk1         <dbl> 87, 91, 81, 76, 57, 51, 97, 84, 59, 76, 84, 57, ~
## $ wk2         <dbl> 82, 87, 70, 76, 34, 39, 97, 62, 53, 76, 84, 47, ~
## $ wk3         <dbl> 72, 92, 68, 72, 25, 34, 96, 51, 38, 74, 75, 45, ~
```


## \$ wk4	<dbl> 77, NA, 67, 69, 17, 26, 95, 41, 28, 69, 73, 29, ~
## \$ wk5	<dbl> 87, NA, 66, 67, 17, 26, 100, 38, 21, 68, 73, 23,~
## \$ wk6	<dbl> 94, NA, 57, 65, 31, 19, NA, 35, 18, 67, 69, 18, ~
## \$ wk7	<dbl> 99, NA, 54, 55, 36, 2, NA, 35, 16, 61, 68, 11, 2~
## \$ wk8	<dbl> NA, NA, 53, 59, 49, 2, NA, 38, 14, 58, 65, 9, 17~
## \$ wk9	<dbl> NA, NA, 51, 62, 53, 3, NA, 38, 12, 57, 73, 9, 17~
## \$ wk10	<dbl> NA, NA, 51, 61, 57, 6, NA, 36, 10, 59, 83, 11, 1~
## \$ wk11	<dbl> NA, NA, 51, 61, 64, 7, NA, 37, 9, 66, 92, 1, 17,~
## \$ wk12	<dbl> NA, NA, 51, 59, 70, 22, NA, 37, 8, 68, NA, 1, 3,~
## \$ wk13	<dbl> NA, NA, 47, 61, 75, 29, NA, 38, 6, 61, NA, 1, 3,~
## \$ wk14	<dbl> NA, NA, 44, 66, 76, 36, NA, 49, 1, 67, NA, 1, 7,~
## \$ wk15	<dbl> NA, NA, 38, 72, 78, 47, NA, 61, 2, 59, NA, 4, 10~
## \$ wk16	<dbl> NA, NA, 28, 76, 85, 67, NA, 63, 2, 63, NA, 8, 17~
## \$ wk17	<dbl> NA, NA, 22, 75, 92, 66, NA, 62, 2, 67, NA, 12, 2~
## \$ wk18	<dbl> NA, NA, 18, 67, 96, 84, NA, 67, 2, 71, NA, 22, 2~
## \$ wk19	<dbl> NA, NA, 18, 73, NA, 93, NA, 83, 3, 79, NA, 23, 2~
## \$ wk20	<dbl> NA, NA, 14, 70, NA, 94, NA, 86, 4, 89, NA, 43, 4~
## \$ wk21	<dbl> NA, NA, 12, NA, NA, NA, NA, NA, 5, NA, NA, 44, 4~
## \$ wk22	<dbl> NA, NA, 7, NA, NA, NA, NA, NA, 5, NA, NA, NA, 50~
## \$ wk23	<dbl> NA, NA, 6, NA, NA, NA, NA, NA, 6, NA, NA, NA, NA~
## \$ wk24	<dbl> NA, NA, 6, NA, NA, NA, NA, NA, 9, NA, NA, NA, NA~
## \$ wk25	<dbl> NA, NA, 6, NA, NA, NA, NA, NA, 13, NA, NA, NA, N~
## \$ wk26	<dbl> NA, NA, 5, NA, NA, NA, NA, NA, 14, NA, NA, NA, N~
## \$ wk27	<dbl> NA, NA, 5, NA, NA, NA, NA, NA, 16, NA, NA, NA, N~
## \$ wk28	<dbl> NA, NA, 4, NA, NA, NA, NA, NA, 23, NA, NA, NA, N~
## \$ wk29	<dbl> NA, NA, 4, NA, NA, NA, NA, NA, 22, NA, NA, NA, N~
## \$ wk30	<dbl> NA, NA, 4, NA, NA, NA, NA, NA, 33, NA, NA, NA, N~
## \$ wk31	<dbl> NA, NA, 4, NA, NA, NA, NA, NA, 36, NA, NA, NA, N~
## \$ wk32	<dbl> NA, NA, 3, NA, NA, NA, NA, NA, 43, NA, NA, NA, N~
## \$ wk33	<dbl> NA, NA, 3, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## \$ wk34	<dbl> NA, NA, 3, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## \$ wk35	<dbl> NA, NA, 4, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## \$ wk36	<dbl> NA, NA, 5, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## \$ wk37	<dbl> NA, NA, 5, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## \$ wk38	<dbl> NA, NA, 9, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## \$ wk39	<dbl> NA, NA, 9, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## \$ wk40	<dbl> NA, NA, 15, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## \$ wk41	<dbl> NA, NA, 14, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## \$ wk42	<dbl> NA, NA, 13, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## \$ wk43	<dbl> NA, NA, 14, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## \$ wk44	<dbl> NA, NA, 16, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## \$ wk45	<dbl> NA, NA, 17, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## \$ wk46	<dbl> NA, NA, 21, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## \$ wk47	<dbl> NA, NA, 22, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## \$ wk48	<dbl> NA, NA, 24, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## \$ wk49	<dbl> NA, NA, 28, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## \$ wk50	<dbl> NA, NA, 33, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~

```
## $ wk51      <dbl> NA, NA, 42, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk52      <dbl> NA, NA, 42, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk53      <dbl> NA, NA, 49, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk54      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk55      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk56      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk57      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk58      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk59      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk60      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk61      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk62      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk63      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk64      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk65      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk66      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk67      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk68      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk69      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk70      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk71      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk72      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk73      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk74      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk75      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ wk76      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

- Vamos renomear as colunas:

```
bb <- billboard %>%
  rename(
    artista = artist,
    musica = track,
    entrou = date.entered
  )
```

```
bb %>% head()
## # A tibble: 6 x 79
##   artista  musica entrou      wk1    wk2    wk3    wk4    wk5    wk6    wk7
##   <chr>    <chr> <date>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 2 Pac    Baby ~ 2000-02-26    87    82    72    77    87    94    99
## 2 2Ge+her  The H~ 2000-09-02    91    87    92    NA    NA    NA    NA
## 3 3 Doors ~ Krypt~ 2000-04-08    81    70    68    67    66    57    54
## 4 3 Doors ~ Loser 2000-10-21    76    76    72    69    67    65    55
## 5 504 Boyz Wobbl~ 2000-04-15    57    34    25    17    17    31    36
## 6 98~0     Give ~ 2000-08-19    51    39    34    26    26    19    2
## # ... with 69 more variables: wk8 <dbl>, wk9 <dbl>, wk10 <dbl>,
```

```
## #   wk11 <dbl>, wk12 <dbl>, wk13 <dbl>, wk14 <dbl>, wk15 <dbl>,
## #   wk16 <dbl>, wk17 <dbl>, wk18 <dbl>, wk19 <dbl>, wk20 <dbl>,
## #   wk21 <dbl>, wk22 <dbl>, wk23 <dbl>, wk24 <dbl>, wk25 <dbl>,
## #   wk26 <dbl>, wk27 <dbl>, wk28 <dbl>, wk29 <dbl>, wk30 <dbl>,
## #   wk31 <dbl>, wk32 <dbl>, wk33 <dbl>, wk34 <dbl>, wk35 <dbl>,
## #   wk36 <dbl>, wk37 <dbl>, wk38 <dbl>, wk39 <dbl>, wk40 <dbl>, ...
```

- **O que é uma observação** neste conjunto de dados?
Uma música que esteve no *top* 100 da *Billboard* durante o ano 2000.
- **Quais são as variáveis** que qualificam cada observação?
 - O artista,
 - O título da música,
 - A colocação da música no *top* 100 da *Billboard* em cada uma das 76 semanas depois que ela entrou na lista.
- Este último item é complexo, e o criador da *tibble* decidiu criar uma coluna por semana.
- **Uma decisão ruim, pois existe informação embutida nos nomes das colunas.** A coluna `wk68` corresponde à posição da música na semana 68 após ela entrar na lista.
- Isto **nunca** deve acontecer. **A informação deve sempre estar nas células.**
- Vamos simplificar as coisas criando duas colunas:
 - `semana`, com o número da semana; perceba que esta informação vem dos nomes das colunas,
 - `pos`, com a posição da música naquela semana.
- A *tibble*, que antes era larga, vai ser mais estreita e mais longa.
- A função `pivot_longer`, do pacote `tidyr`, vai fazer o trabalho — inclusive extraíndo os números das semanas dos nomes das colunas:

```
bb_tidy <- bb %>%
  pivot_longer(
    wk1:wk76,
    names_to = 'semana',
    names_prefix = 'wk',
    names_transform = list(
      semana = as.integer
    ),
    values_to = 'pos'
  )

bb_tidy
## # A tibble: 24.092 x 5
##   artista musica          entrou  semana  pos
```

```
##   <chr>   <chr>                <date>      <int> <dbl>
## 1 2 Pac   Baby Don't Cry (Keep... 2000-02-26      1    87
## 2 2 Pac   Baby Don't Cry (Keep... 2000-02-26      2    82
## 3 2 Pac   Baby Don't Cry (Keep... 2000-02-26      3    72
## 4 2 Pac   Baby Don't Cry (Keep... 2000-02-26      4    77
## 5 2 Pac   Baby Don't Cry (Keep... 2000-02-26      5    87
## 6 2 Pac   Baby Don't Cry (Keep... 2000-02-26      6    94
## # ... with 24.086 more rows
```

- Existem linhas onde pos tem o valor NA. São resultado da organização original dos dados. No novo formato, não servem mais. Vamos eliminá-las.

```
bb_tidy <- bb_tidy %>%
  filter(!is.na(pos))

bb_tidy
## # A tibble: 5.307 x 5
##   artista musica          entrou      semana  pos
##   <chr>   <chr>          <date>      <int> <dbl>
## 1 2 Pac   Baby Don't Cry (Keep... 2000-02-26      1    87
## 2 2 Pac   Baby Don't Cry (Keep... 2000-02-26      2    82
## 3 2 Pac   Baby Don't Cry (Keep... 2000-02-26      3    72
## 4 2 Pac   Baby Don't Cry (Keep... 2000-02-26      4    77
## 5 2 Pac   Baby Don't Cry (Keep... 2000-02-26      5    87
## 6 2 Pac   Baby Don't Cry (Keep... 2000-02-26      6    94
## # ... with 5.301 more rows
```

3.4

Manipulando os dados

3.4.1

Criando novas colunas: mutate, transmute

- O *data frame* cars tem dados (de 1920!) sobre as distâncias de frenagem (em pés) de um carro viajando a diversas velocidades (em milhas por hora):

```
cars
## # A tibble: 50 x 2
##   speed  dist
##   <dbl> <dbl>
## 1     4     2
## 2     4    10
## 3     7     4
## 4     7    22
## 5     8    16
```

```
## 6      9      10
## # ... with 44 more rows
```

- Vamos criar colunas novas com os valores convertidos para km/h e metros; além disso, uma coluna com a taxa de frenagem:

```
cars %>%
  mutate(
    velocidade = speed * 1.6,
    distancia = dist * .33,
    taxa = velocidade / distancia
  )
## # A tibble: 50 x 5
##   speed dist velocidade distancia taxa
##   <dbl> <dbl>      <dbl>      <dbl> <dbl>
## 1     4     2        6.4        0.66  9.70
## 2     4    10        6.4        3.3   1.94
## 3     7     4       11.2        1.32  8.48
## 4     7    22       11.2        7.26  1.54
## 5     8    16       12.8        5.28  2.42
## 6     9    10       14.4        3.3   4.36
## # ... with 44 more rows
```

- Perceba que as colunas antigas continuam lá. Se quiser manter apenas as colunas novas, use transmute:

```
cars %>%
  transmute(
    velocidade = speed * 1.6,
    distancia = dist * .33,
    taxa = velocidade / distancia
  )
## # A tibble: 50 x 3
##   velocidade distancia taxa
##       <dbl>      <dbl> <dbl>
## 1        6.4        0.66  9.70
## 2        6.4        3.3   1.94
## 3       11.2        1.32  8.48
## 4       11.2        7.26  1.54
## 5       12.8        5.28  2.42
## 6       14.4        3.3   4.36
## # ... with 44 more rows
```

- Ou use o argumento `.keep` de `mutate` para escolher com mais precisão. Veja a ajuda de `mutate`.

3.4.2

Selecionando colunas: select, distinct, pull

- Vamos voltar à nossa *tibble* dos *top 100* da *Billboard*.
- Para ver só a coluna de artistas:

```
bb_tidy %>%  
  select(artista)  
## # A tibble: 5.307 x 1  
##   artista  
##   <chr>  
## 1 2 Pac  
## 2 2 Pac  
## 3 2 Pac  
## 4 2 Pac  
## 5 2 Pac  
## 6 2 Pac  
## # ... with 5.301 more rows
```

- Para eliminar as repetições:

```
bb_tidy %>%  
  select(artista) %>%  
  distinct()  
## # A tibble: 228 x 1  
##   artista  
##   <chr>  
## 1 2 Pac  
## 2 2Ge+her  
## 3 3 Doors Down  
## 4 504 Boyz  
## 5 98~0  
## 6 A*Teens  
## # ... with 222 more rows
```

- Artistas e músicas:

```
bb_tidy %>%  
  select(artista, musica) %>%  
  distinct()  
## # A tibble: 317 x 2  
##   artista      musica  
##   <chr>      <chr>  
## 1 2 Pac      Baby Don't Cry (Keep...  
## 2 2Ge+her    The Hardest Part Of ...  
## 3 3 Doors Down Kryptonite
```

```
## 4 3 Doors Down Loser
## 5 504 Boyz      Wobble Wobble
## 6 98~0         Give Me Just One Nig...
## # ... with 311 more rows
```

- Para especificar colunas a não mostrar:

```
bb_tidy %>%
  select(-entrou)
## # A tibble: 5.307 x 4
##   artista musica          semana  pos
##   <chr>   <chr>          <int> <dbl>
## 1 2 Pac    Baby Don't Cry (Keep...     1    87
## 2 2 Pac    Baby Don't Cry (Keep...     2    82
## 3 2 Pac    Baby Don't Cry (Keep...     3    72
## 4 2 Pac    Baby Don't Cry (Keep...     4    77
## 5 2 Pac    Baby Don't Cry (Keep...     5    87
## 6 2 Pac    Baby Don't Cry (Keep...     6    94
## # ... with 5.301 more rows
```

- Para extrair uma coluna na forma de vetor (unique é uma função do R base, aplicável a vetores):

```
bb_tidy %>%
  pull(artista) %>%
  unique()
##   [1] "2 Pac"                "2Ge+her"
##   [3] "3 Doors Down"         "504 Boyz"
##   [5] "98~0"                 "A*Teens"
##   [7] "Aaliyah"              "Adams, Yolanda"
##   [9] "Adkins, Trace"        "Aguilera, Christina"
##  [11] "Alice DeeJay"         "Allan, Gary"
##  [13] "Amber"                "Anastacia"
##  [15] "Anthony, Marc"        "Avant"
##  [17] "BBMak"                "Backstreet Boys, The"
##  [19] "Badu, Erkyah"         "Baha Men"
##  [21] "Barenaked Ladies"     "Beenie Man"
##  [23] "Before Dark"          "Bega, Lou"
##  [25] "Big Punisher"         "Black Rob"
##  [27] "Black, Clint"         "Blaque"
##  [29] "Blige, Mary J."       "Blink-182"
##  [31] "Bloodhound Gang"      "Bon Jovi"
##  [33] "Braxton, Toni"         "Brock, Chad"
##  [35] "Brooks & Dunn"        "Brooks, Garth"
##  [37] "Byrd, Tracy"          "Cagle, Chris"
##  [39] "Cam'ron"              "Carey, Mariah"
##  [41] "Carter, Aaron"        "Carter, Torrey"
```

## [43]	"Changing Faces"	"Chesney, Kenny"
## [45]	"Clark Family Experience"	"Clark, Terri"
## [47]	"Common"	"Counting Crows"
## [49]	"Creed"	"Cyrus, Billy Ray"
## [51]	"D'Angelo"	"DMX"
## [53]	"Da Brat"	"Davidson, Clay"
## [55]	"De La Soul"	"Destiny's Child"
## [57]	"Diffie, Joe"	"Dion, Celine"
## [59]	"Dixie Chicks, The"	"Dr. Dre"
## [61]	"Drama"	"Dream"
## [63]	"Eastsidaz, The"	"Eiffel 65"
## [65]	"Elliott, Missy \"Misdemeanor\""	"Eminem"
## [67]	"En Vogue"	"Estefan, Gloria"
## [69]	"Evans, Sara"	"Eve"
## [71]	"Everclear"	"Fabian, Lara"
## [73]	"Fatboy Slim"	"Filter"
## [75]	"Foo Fighters"	"Fragma"
## [77]	"Funkmaster Flex"	"Ghostface Killah"
## [79]	"Gill, Vince"	"Gilman, Billy"
## [81]	"Ginuwine"	"Goo Goo Dolls"
## [83]	"Gray, Macy"	"Griggs, Andy"
## [85]	"Guy"	"Hanson"
## [87]	"Hart, Beth"	"Heatherly, Eric"
## [89]	"Henley, Don"	"Herndon, Ty"
## [91]	"Hill, Faith"	"Hoku"
## [93]	"Hollister, Dave"	"Hot Boys"
## [95]	"Houston, Whitney"	"IMx"
## [97]	"Ice Cube"	"Ideal"
## [99]	"Iglesias, Enrique"	"J-Shin"
## [101]	"Ja Rule"	"Jackson, Alan"
## [103]	"Jagged Edge"	"Janet"
## [105]	"Jay-Z"	"Jean, Wyclef"
## [107]	"Joe"	"John, Elton"
## [109]	"Jones, Donell"	"Jordan, Montell"
## [111]	"Juvenile"	"Kandi"
## [113]	"Keith, Toby"	"Kelis"
## [115]	"Kenny G"	"Kid Rock"
## [117]	"Kravitz, Lenny"	"Kumbia Kings"
## [119]	"LFO"	"LL Cool J"
## [121]	"Larrieux, Amel"	"Lawrence, Tracy"
## [123]	"Levert, Gerald"	"Lil Bow Wow"
## [125]	"Lil Wayne"	"Lil' Kim"
## [127]	"Lil' Mo"	"Lil' Zane"
## [129]	"Limp Bizkit"	"Lonestar"
## [131]	"Lopez, Jennifer"	"Loveless, Patty"
## [133]	"Lox"	"Lucy Pearl"
## [135]	"Ludacris"	"M2M"

## [137] "Madison Avenue"	"Madonna"
## [139] "Martin, Ricky"	"Mary Mary"
## [141] "Master P"	"McBride, Martina"
## [143] "McEntire, Reba"	"McGraw, Tim"
## [145] "McKnight, Brian"	"Messina, Jo Dee"
## [147] "Metallica"	"Montgomery Gentry"
## [149] "Montgomery, John Michael"	"Moore, Chante"
## [151] "Moore, Mandy"	"Mumba, Samantha"
## [153] "MusiQ"	"Mya"
## [155] "Mystikal"	"N'Sync"
## [157] "Nas"	"Nelly"
## [159] "Next"	"Nine Days"
## [161] "No Doubt"	"Nu Flavor"
## [163] "Offspring, The"	"Paisley, Brad"
## [165] "Papa Roach"	"Pearl Jam"
## [167] "Pink"	"Price, Kelly"
## [169] "Profyle"	"Puff Daddy"
## [171] "Q-Tip"	"R.E.M."
## [173] "Rascal Flatts"	"Raye, Collin"
## [175] "Red Hot Chili Peppers"	"Rimes, LeAnn"
## [177] "Rogers, Kenny"	"Ruff Endz"
## [179] "Sammie"	"Santana"
## [181] "Savage Garden"	"SheDaisy"
## [183] "Sheist, Shade"	"Shyne"
## [185] "Simpson, Jessica"	"Sisqo"
## [187] "Sister Hazel"	"Smash Mouth"
## [189] "Smith, Will"	"Son By Four"
## [191] "Sonique"	"SoulDecision"
## [193] "Spears, Britney"	"Spencer, Tracie"
## [195] "Splendor"	"Sting"
## [197] "Stone Temple Pilots"	"Stone, Angie"
## [199] "Strait, George"	"Sugar Ray"
## [201] "TLC"	"Tamar"
## [203] "Tamia"	"Third Eye Blind"
## [205] "Thomas, Carl"	"Tippin, Aaron"
## [207] "Train"	"Trick Daddy"
## [209] "Trina"	"Tritt, Travis"
## [211] "Tuesday"	"Urban, Keith"
## [213] "Usher"	"Vassar, Phil"
## [215] "Vertical Horizon"	"Vitamin C"
## [217] "Walker, Clay"	"Wallflowers, The"
## [219] "Westlife"	"Williams, Robbie"
## [221] "Wills, Mark"	"Worley, Darryl"
## [223] "Wright, Chely"	"Yankee Grey"
## [225] "Yearwood, Trisha"	"Ying Yang Twins"
## [227] "Zombie Nation"	"matchbox twenty"

3.4.3

Filtrando linhas: filter, slice

- Apenas as músicas da Britney Spears:

```
bb_tidy %>%
  filter(artista == 'Spears, Britney')
## # A tibble: 51 x 5
##   artista          musica          entrou    semana    pos
##   <chr>            <chr>          <date>    <int> <dbl>
## 1 Spears, Britney From The Bottom Of M... 2000-01-29      1     76
## 2 Spears, Britney From The Bottom Of M... 2000-01-29      2     59
## 3 Spears, Britney From The Bottom Of M... 2000-01-29      3     52
## 4 Spears, Britney From The Bottom Of M... 2000-01-29      4     52
## 5 Spears, Britney From The Bottom Of M... 2000-01-29      5     14
## 6 Spears, Britney From The Bottom Of M... 2000-01-29      6     14
## # ... with 45 more rows
```

- Apenas músicas que chegaram à posição 1:

```
bb_tidy %>%
  filter(pos == 1) %>%
  select(-pos)
## # A tibble: 55 x 4
##   artista          musica          entrou    semana
##   <chr>            <chr>          <date>    <int>
## 1 Aaliyah          Try Again      2000-03-18      14
## 2 Aguilera, Christina Come On Over Baby (A... 2000-08-05      11
## 3 Aguilera, Christina Come On Over Baby (A... 2000-08-05      12
## 4 Aguilera, Christina Come On Over Baby (A... 2000-08-05      13
## 5 Aguilera, Christina Come On Over Baby (A... 2000-08-05      14
## 6 Aguilera, Christina What A Girl Wants      1999-11-27       8
## # ... with 49 more rows
```

- Apenas músicas que chegaram à posição 1 em menos de 10 semanas:

```
bb_tidy %>%
  filter(pos == 1, semana < 10) %>%
  distinct(artista, musica)
## # A tibble: 5 x 2
##   artista          musica
##   <chr>            <chr>
## 1 Aguilera, Christina What A Girl Wants
## 2 Destiny's Child  Independent Women Pa...
## 3 Madonna          Music
## 4 Santana          Maria, Maria
## 5 Sisqo            Incomplete
```

- As funções da família `slice` filtram linhas de diversas maneiras.
- De acordo com seus índices (números de linha):

```
bb_tidy %>%
  slice(c(1, 1000, 5000))
## # A tibble: 3 x 5
##   artista          musica          entrou      semana  pos
##   <chr>          <chr>          <date>      <int> <dbl>
## 1 2 Pac          Baby Don't Cry (Keep~ 2000-02-26      1    87
## 2 Clark Family Experience Meanwhile Back At Th~ 2000-11-18      3    81
## 3 Vassar, Phil    Carlene          2000-03-04      3    64
```

```
bb_tidy %>%
  slice_head(n = 4)
## # A tibble: 4 x 5
##   artista musica          entrou      semana  pos
##   <chr>   <chr>          <date>      <int> <dbl>
## 1 2 Pac    Baby Don't Cry (Keep... 2000-02-26      1    87
## 2 2 Pac    Baby Don't Cry (Keep... 2000-02-26      2    82
## 3 2 Pac    Baby Don't Cry (Keep... 2000-02-26      3    72
## 4 2 Pac    Baby Don't Cry (Keep... 2000-02-26      4    77
```

```
bb_tidy %>%
  slice_tail(n = 4)
## # A tibble: 4 x 5
##   artista          musica entrou      semana  pos
##   <chr>          <chr> <date>      <int> <dbl>
## 1 matchbox twenty Bent    2000-04-29     36    37
## 2 matchbox twenty Bent    2000-04-29     37    38
## 3 matchbox twenty Bent    2000-04-29     38    38
## 4 matchbox twenty Bent    2000-04-29     39    48
```

- De acordo com a ordenação de uma coluna ou de uma função das colunas:

```
bb_tidy %>%
  slice_min(pos)
## # A tibble: 55 x 5
##   artista          musica          entrou      semana  pos
##   <chr>          <chr>          <date>      <int> <dbl>
## 1 Aaliyah          Try Again          2000-03-18     14     1
## 2 Aguilera, Christina Come On Over Baby (A... 2000-08-05     11     1
## 3 Aguilera, Christina Come On Over Baby (A... 2000-08-05     12     1
## 4 Aguilera, Christina Come On Over Baby (A... 2000-08-05     13     1
## 5 Aguilera, Christina Come On Over Baby (A... 2000-08-05     14     1
## 6 Aguilera, Christina What A Girl Wants      1999-11-27      8     1
## # ... with 49 more rows
```

```
bb_tidy %>%
  slice_max(semana)
## # A tibble: 1 x 5
##   artista musica entrou      semana  pos
##   <chr>   <chr> <date>      <int> <dbl>
## 1 Creed   Higher 1999-09-11      65    49
```

- Aleatoriamente, criando uma amostra:

```
bb_tidy %>%
  slice_sample(n = 5)
## # A tibble: 5 x 5
##   artista      musica      entrou      semana  pos
##   <chr>      <chr>      <date>      <int> <dbl>
## 1 Sisqo      Incomplete 2000-06-24      7    2
## 2 3 Doors Down Kryptonite 2000-04-08     23    6
## 3 Lonestar    Amazed     1999-06-05     64   50
## 4 Jones, Donell Where I Wanna Be 2000-04-22      9   38
## 5 Dream      He Loves U Not 2000-09-30      5   47
```

- Veja a ajuda de `slice` para saber mais sobre estas funções. Por exemplo:
 - `slice_min` e `slice_max` podem considerar ou não empates.
 - Você pode especificar uma proporção de linhas (usando `prop`) em vez da quantidade de linhas (`n`).
 - Você pode fazer amostragem com reposição, ou com probabilidades diferentes para cada linha.

3.4.4

Ordenando linhas: `arrange`

- Por título, sem repetições:

```
bb_tidy %>%
  select(musica) %>%
  distinct() %>%
  arrange(musica)
## # A tibble: 316 x 1
##   musica
##   <chr>
## 1 (Hot S**t) Country G...
## 2 3 Little Words
## 3 911
## 4 A Country Boy Can Su...
## 5 A Little Gasoline
```

```
## 6 A Puro Dolor (Purest...
## # ... with 310 more rows
```

- Por título, sem repetições, em ordem inversa:

```
bb_tidy %>%
  select(musica) %>%
  distinct() %>%
  arrange(desc(musica))
## # A tibble: 316 x 1
##   musica
##   <chr>
## 1 Your Everything
## 2 You're A God
## 3 You'll Always Be Lov...
## 4 You Won't Be Lonely ...
## 5 You Should've Told M...
## 6 You Sang To Me
## # ... with 310 more rows
```

3.4.5

Contando linhas: count

- Quantas semanas cada artista ficou nos *top 100*? Duas músicas na mesma semana contam como duas semanas.

```
bb_tidy %>%
  count(artista, sort = TRUE)
## # A tibble: 228 x 2
##   artista          n
##   <chr>          <int>
## 1 Creed           104
## 2 Lonestar         95
## 3 Destiny's Child  92
## 4 N'Sync           74
## 5 Sisqo            74
## 6 3 Doors Down     73
## # ... with 222 more rows
```

- Quantas semanas cada música ficou nos *top 100*?

```
bb_tidy %>%
  count(musica, sort = TRUE)
## # A tibble: 316 x 2
##   musica          n
##   <chr>          <int>
```

```
## 1 Higher 57
## 2 Amazed 55
## 3 Breathe 53
## 4 Kryptonite 53
## 5 With Arms Wide Open 47
## 6 I Wanna Know 44
## # ... with 310 more rows
```

- Se houve músicas com o mesmo nome, mas de artistas diferentes, o código acima está errado. O certo é

```
bb_tidy %>%
  count(musica, artista, sort = TRUE)
## # A tibble: 317 x 3
##   musica      artista      n
##   <chr>      <chr>    <int>
## 1 Higher    Creed      57
## 2 Amazed    Lonestar   55
## 3 Breathe    Hill, Faith 53
## 4 Kryptonite 3 Doors Down 53
## 5 With Arms Wide Open Creed      47
## 6 I Wanna Know Joe      44
## # ... with 311 more rows
```

3.4.6

Agrupando linhas: group_by e summarize

- Qual foi a melhor posição que cada artista alcançou?

```
bb_tidy %>%
  group_by(artista) %>%
  summarize(melhor = min(pos)) %>%
  arrange(melhor)
## # A tibble: 228 x 2
##   artista      melhor
##   <chr>      <dbl>
## 1 Aaliyah      1
## 2 Aguilera, Christina 1
## 3 Carey, Mariah      1
## 4 Creed          1
## 5 Destiny's Child    1
## 6 Iglesias, Enrique   1
## # ... with 222 more rows
```

- Qual foi a melhor posição que cada música alcançou?

```
bb_tidy %>%
  group_by(artista, musica) %>%
  summarize(melhor = min(pos)) %>%
  arrange(melhor)
## `summarize()` has grouped output by 'artista'. You can override using the
## `.groups` argument.
## # A tibble: 317 x 3
##   artista          musica          melhor
##   <chr>          <chr>          <dbl>
## 1 Aaliyah        Try Again          1
## 2 Aguilera, Christina Come On Over Baby (A... 1
## 3 Aguilera, Christina What A Girl Wants      1
## 4 Carey, Mariah   Thank God I Found Yo... 1
## 5 Creed          With Arms Wide Open    1
## 6 Destiny's Child Independent Women Pa... 1
## # ... with 311 more rows
```

- Quando usamos `summarize`, o agrupamento mais interno é desfeito. Isto significa que o resultado acima ainda está agrupado por artista.
- Quantas semanas cada artista ficou na posição 1?

```
bb_tidy %>%
  filter(pos == 1) %>%
  group_by(artista) %>%
  summarize(semanas = n()) %>%
  arrange(desc(semanas))
## # A tibble: 15 x 2
##   artista          semanas
##   <chr>          <int>
## 1 Destiny's Child      14
## 2 Santana              10
## 3 Aguilera, Christina   6
## 4 Madonna              4
## 5 Savage Garden        4
## 6 Iglesias, Enrique     3
## # ... with 9 more rows
```

- Perceba que `count`, que vimos mais acima, faz agrupamentos do mesmo modo:

```
bb_tidy %>%
  filter(pos == 1) %>%
  count(artista, sort = TRUE)
## # A tibble: 15 x 2
##   artista          n
##   <chr>          <int>
## 1 Destiny's Child      14
```

```
## 2 Santana 10
## 3 Aguilera, Christina 6
## 4 Madonna 4
## 5 Savage Garden 4
## 6 Iglesias, Enrique 3
## # ... with 9 more rows
```

- Uma pergunta diferente: quais são os artistas cujas músicas apareceram nos *top* 100 mais tempo depois do lançamento da música?

```
bb_tidy %>%
  group_by(artista) %>%
  summarize(semanas = max(semana)) %>%
  arrange(desc(semanas))
## # A tibble: 228 x 2
##   artista      semanas
##   <chr>        <int>
## 1 Creed        65
## 2 Lonestar     64
## 3 3 Doors Down 53
## 4 Hill, Faith  53
## 5 Joe          44
## 6 Vertical Horizon 41
## # ... with 222 more rows
```

- Qual a posição média de cada música? Lembre-se de que eliminamos as linhas com NA; logo, a média é sobre a quantidade de semanas em que a música esteve na lista.

```
media1 <- bb_tidy %>%
  group_by(artista, musica) %>%
  summarize(media = mean(pos), .groups = 'drop') %>%
  arrange(media)

media1
## # A tibble: 317 x 3
##   artista      musica      media
##   <chr>        <chr>    <dbl>
## 1 "Santana"     Maria, Maria  10.5
## 2 "Madonna"     Music        13.5
## 3 "N'Sync"      Bye Bye Bye  14.3
## 4 "Elliott, Missy \"Misdemeanor\"" Hot Boyz    14.3
## 5 "Destiny's Child" Independent Women Pa... 14.8
## 6 "Iglesias, Enrique" Be With You  15.8
## # ... with 311 more rows
```

- E se quisermos a média sobre o número de semanas desde a entrada da música até a última semana em que a música apareceu na lista?


```
media2 <- bb_tidy %>%
  group_by(artista, musica) %>%
  summarize(media = sum(pos)/max(semana), .groups = 'drop') %>%
  arrange(media)

media2
## # A tibble: 317 x 3
##   artista                musica                media
##   <chr>                  <chr>                  <dbl>
## 1 "Santana"              Maria, Maria             10.5
## 2 "Madonna"              Music                    13.5
## 3 "N'Sync"               Bye Bye Bye              14.3
## 4 "Elliott, Missy \"Misdemeanor\"" Hot Boyz                 14.3
## 5 "Destiny's Child"      Independent Women Pa...  14.8
## 6 "Iglesias, Enrique"    Be With You              15.8
## # ... with 311 more rows
```

```
identical(media1, media2)
## [1] FALSE
```

Visualização com ggplot2



Busque mais informações sobre os pacotes `tidyverse` e `ggplot2` nas referências recomendadas.

4.1

Vídeo 1

<https://youtu.be/OBpNjqIIyhI>

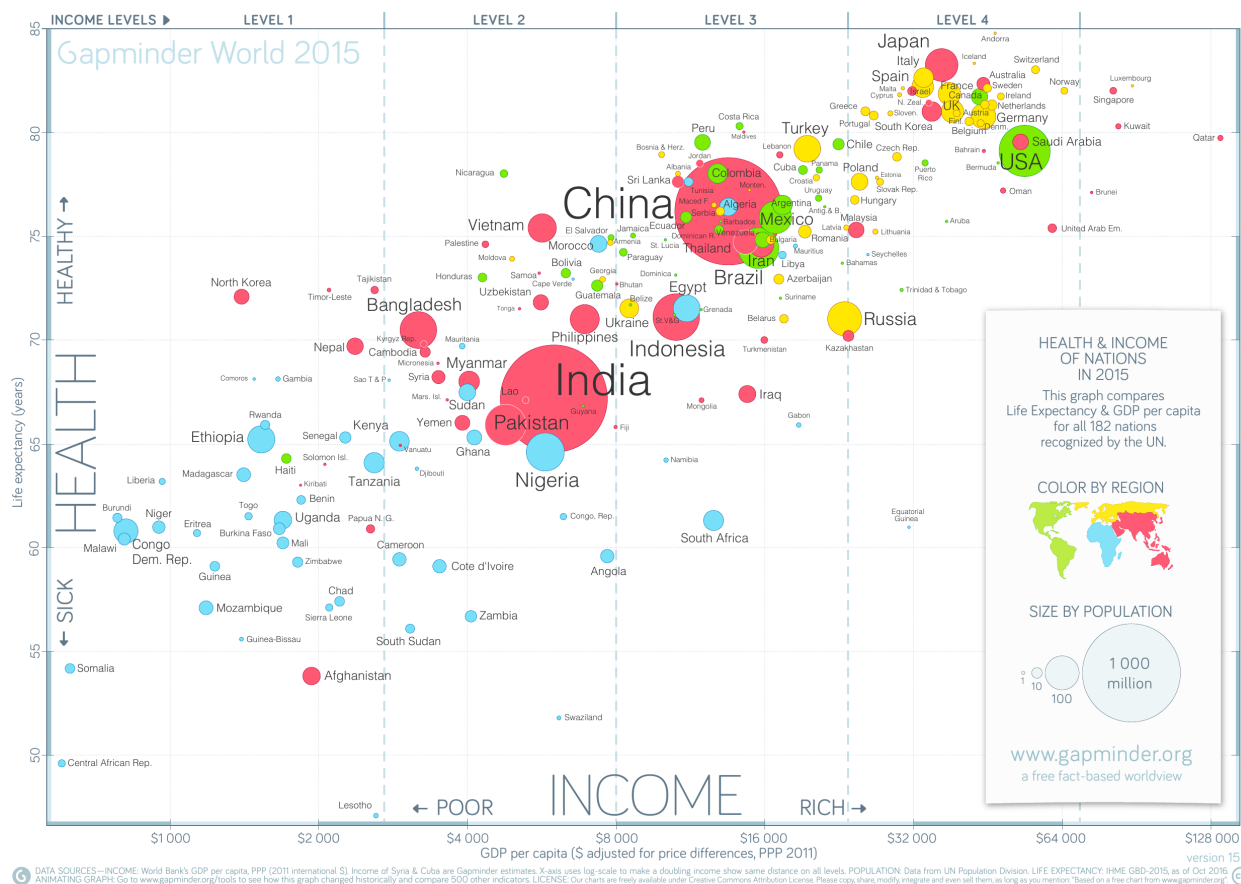
4.2

Componentes de um gráfico ggplot2

4.2.1

Geometrias e mapeamentos estéticos (*mappings*)

- Observe o gráfico abaixo, obtido de <https://www.gapminder.org/downloads/updated-gapminder-world-poster-2015/>.



- O gráfico mostra como, em cada país, a saúde (mais precisamente, a expectativa de vida) se relaciona com a riqueza (mais precisamente, o PIB *per capita*).
- Além da expectativa de vida e o do PIB *per capita*, o gráfico traz mais informações sobre cada país.
- Cada país é representado por um ponto (a **geometria**).
- Informações sobre cada país são representadas por características do ponto correspondente (as **estéticas**):

Variável	Geometria	Estética
PIB <i>per capita</i>	ponto	posição x
Expectativa de vida	ponto	posição y
População	ponto	tamanho
Continente	ponto	cor

- Você pode usar outras estéticas para representar informações:
 - Cor de preenchimento.
 - Cor do traço.
 - Tipo do traço (sólido, pontilhado, tracejado etc.).
 - Forma (círculo, quadrado, triângulo etc.).

- Opacidade.
- etc.
- Você pode usar outras geometrias:
 - Linhas.
 - Barras ou colunas.
 - Caixas.
 - etc.

4.2.2

Escalas (*scales*)

- As escalas controlam os detalhes da aparência da geometria e do mapeamento (eixos, cores etc.).
- Os eixos do gráfico acima são escalas **contínuas**, com valores reais.
- Observe o eixo horizontal. Os valores não aumentam linearmente, mas sim exponencialmente: cada passo à direita equivale a *dobrar* o valor do PIB. O eixo horizontal segue uma **escala logarítmica**.
- Os tamanhos dos pontos formam uma escala **discreta**, com 4 valores possíveis (veja a legenda no canto inferior direito do gráfico).
- As cores também formam uma escala discreta.

4.2.3

Rótulos (*labels*)

- O gráfico também representa informação na forma de texto.
- Além de rótulos (por exemplo, o texto que identifica cada eixo), **o texto também pode, ele mesmo, ser uma geometria, com suas próprias estéticas**: observe como o nome de cada país é escrito em um tamanho proporcional à sua população.

4.2.4

Outros componentes

- Coordenadas:
 - Este gráfico usa **coordenadas cartesianas**, com eixos x e y .
 - Existem gráficos que usam um sistema de **coordenadas polares**.
- Temas:
 - Incluem todos os elementos “decorativos”: cor de fundo, linhas de grade, etc. Ajudam a facilitar a leitura e a interpretação.

- No gráfico acima, um detalhe interessante do tema é a divisão de cada eixo em segmentos claros e segmentos escuros.
- Legendas (*guides*).
- Facetas:
 - Às vezes, um gráfico é composto por múltiplos subgráficos.
 - Cada subgráfico é uma **faceta**.
 - Facetas evitam que informações demais sejam apresentadas no mesmo lugar.

4.3

Conjunto de dados

- Nossos exemplos de gráficos vão usar dados sobre o sono de diversos mamíferos.
- O conjunto de dados se chama `msleep` e está incluído no pacote `ggplot2`.
- Para ver a documentação, digite

```
library(ggplot2)
?msleep
```

- Vamos atribuir o conjunto de dados à variável `df`:

```
df <- msleep
df
## # A tibble: 83 x 11
##   name          genus vore  order conservation sleep_total sleep_rem
##   <chr>         <chr> <chr> <chr> <chr>          <dbl>     <dbl>
## 1 Cheetah      Acin~ carni Carn~ lc             12.1      NA
## 2 Owl monkey  Aotus omni  Prim~ <NA>          17        1.8
## 3 Mountain beaver Aplo~ herbi Rode~ nt          14.4      2.4
## 4 Greater short-t~ Blar~ omni  Sori~ lc          14.9      2.3
## 5 Cow          Bos   herbi Arti~ domesticated    4        0.7
## 6 Three-toed sloth Brad~ herbi Pilo~ <NA>        14.4      2.2
## # ... with 77 more rows, and 4 more variables: sleep_cycle <dbl>,
## #   awake <dbl>, brainwt <dbl>, bodywt <dbl>
```

- Vamos examinar a estrutura — usando R base:

```
str(df)
## tibble [83 x 11] (S3: tbl_df/tbl/data.frame)
##  $ name          : chr [1:83] "Cheetah" "Owl monkey" "Mountain beaver" ...
##  $ genus          : chr [1:83] "Acinonyx" "Aotus" "Aplodontia" ...
##  $ vore           : chr [1:83] "carni" "omni" "herbi" ...
##  $ order          : chr [1:83] "Carnivora" "Primates" "Rodentia" ...
```

```
## $ conservation: chr [1:83] "lc" NA "nt" ...
## $ sleep_total : num [1:83] 12,1 17 14,4 14,9 4 14,4 8,7 7 ...
## $ sleep_rem : num [1:83] NA 1,8 2,4 2,3 0,7 2,2 1,4 NA ...
## $ sleep_cycle : num [1:83] NA NA NA 0,133 ...
## $ awake : num [1:83] 11,9 7 9,6 9,1 20 9,6 15,3 17 ...
## $ brainwt : num [1:83] NA 0,0155 NA 0,00029 0,423 NA NA NA ...
## $ bodywt : num [1:83] 50 0,48 1,35 0,019 ...
```

- Podemos usar `glimpse`, uma função do `tidyverse`:

```
glimpse(df)
## Rows: 83
## Columns: 11
## $ name <chr> "Cheetah", "Owl monkey", "Mountain beaver", "Gre~
## $ genus <chr> "Acinonyx", "Aotus", "Aplodontia", "Blarina", "B~
## $ vore <chr> "carni", "omni", "herbi", "omni", "herbi", "herb~
## $ order <chr> "Carnivora", "Primates", "Rodentia", "Soricomorp~
## $ conservation <chr> "lc", NA, "nt", "lc", "domesticated", NA, "vu", ~
## $ sleep_total <dbl> 12,1, 17,0, 14,4, 14,9, 4,0, 14,4, 8,7, 7,0, 10,~
## $ sleep_rem <dbl> NA, 1,8, 2,4, 2,3, 0,7, 2,2, 1,4, NA, 2,9, NA, 0~
## $ sleep_cycle <dbl> NA, NA, NA, 0,1333333, 0,6666667, 0,7666667, 0,3~
## $ awake <dbl> 11,9, 7,0, 9,6, 9,1, 20,0, 9,6, 15,3, 17,0, 13,9~
## $ brainwt <dbl> NA, 0,01550, NA, 0,00029, 0,42300, NA, NA, NA, 0~
## $ bodywt <dbl> 50,000, 0,480, 1,350, 0,019, 600,000, 3,850, 20,~
```

- Para examinar só as primeiras linhas do *data frame*:

```
head(df)
## # A tibble: 6 x 11
##   name          genus vore order conservation sleep_total sleep_rem
##   <chr>         <chr> <chr> <chr> <chr>          <dbl>      <dbl>
## 1 Cheetah      Acin~ carni Carn~ lc              12.1      NA
## 2 Owl monkey   Aotus omni Prim~ <NA>          17        1.8
## 3 Mountain beaver Aplo~ herbi Rode~ nt              14.4      2.4
## 4 Greater short-t~ Blar~ omni Sori~ lc              14.9      2.3
## 5 Cow          Bos   herbi Arti~ domesticated    4        0.7
## 6 Three-toed sloth Brad~ herbi Pilo~ <NA>          14.4      2.2
## # ... with 4 more variables: sleep_cycle <dbl>, awake <dbl>,
## #   brainwt <dbl>, bodywt <dbl>
```

- Para examinar o *data frame* interativamente:

```
view(df)
```

- Podemos produzir um sumário dos dados usando o pacote *summarytools* (que já foi carregado neste documento):

```
df %>% dfSummary() %>% print()
```

Variável	Estatísticas / Valores	Freqs (% de Válidos)	Faltante
name [character]	1. African elephant 2. African giant pouched rat 3. African striped mouse 4. Arctic fox 5. Arctic ground squirrel 6. Asian elephant 7. Baboon 8. Big brown bat 9. Bottle-nosed dolphin 10. Brazilian tapir [73 outros]	1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 73 (88,0%)	0 (0,0%)
genus [character]	1. Panthera 2. Spermophilus 3. Equus 4. Vulpes 5. Acinonyx 6. Aotus 7. Aplodontia 8. Blarina 9. Bos 10. Bradypus [67 outros]	3 (3,6%) 3 (3,6%) 2 (2,4%) 2 (2,4%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 67 (80,7%)	0 (0,0%)
vore [character]	1. carni 2. herbi 3. insecti 4. omni	19 (25,0%) 32 (42,1%) 5 (6,6%) 20 (26,3%)	7 (8,4%)
order [character]	1. Rodentia 2. Carnivora 3. Primates 4. Artiodactyla 5. Soricomorpha 6. Cetacea 7. Hyracoidea 8. Perissodactyla 9. Chiroptera 10. Cingulata [9 outros]	22 (26,5%) 12 (14,5%) 12 (14,5%) 6 (7,2%) 5 (6,0%) 3 (3,6%) 3 (3,6%) 3 (3,6%) 2 (2,4%) 2 (2,4%) 13 (15,7%)	0 (0,0%)
conservation [character]	1. cd 2. domesticated 3. en 4. lc 5. nt 6. vu	2 (3,7%) 10 (18,5%) 4 (7,4%) 27 (50,0%) 4 (7,4%) 7 (13,0%)	29 (34,9%)

Variável	Estatísticas / Valores	Freqs (% de Válidos)	Faltante
sleep_total [numeric]	Média (dp) : 10,4 (4,5) mín < mediana < máx: 1,9 < 10,1 < 19,9 IQE (CV) : 5,9 (0,4)	65 valores distintos	0 (0,0%)
sleep_rem [numeric]	Média (dp) : 1,9 (1,3) mín < mediana < máx: 0,1 < 1,5 < 6,6 IQE (CV) : 1,5 (0,7)	32 valores distintos	22 (26,5%)
sleep_cycle [numeric]	Média (dp) : 0,4 (0,4) mín < mediana < máx: 0,1 < 0,3 < 1,5 IQE (CV) : 0,4 (0,8)	22 valores distintos	51 (61,4%)
awake [numeric]	Média (dp) : 13,6 (4,5) mín < mediana < máx: 4,1 < 13,9 < 22,1 IQE (CV) : 5,9 (0,3)	65 valores distintos	0 (0,0%)
brainwt [numeric]	Média (dp) : 0,3 (1) mín < mediana < máx: 0 < 0 < 5,7 IQE (CV) : 0,1 (3,5)	53 valores distintos	27 (32,5%)
bodywt [numeric]	Média (dp) : 166,1 (786,8) mín < mediana < máx: 0 < 1,7 < 6654 IQE (CV) : 41,6 (4,7)	82 valores distintos	0 (0,0%)

- Vemos que há muitos NA em diversas variáveis. Para nossos exemplos simples de visualização, vamos usar as colunas

- name
- genus
- order
- sleep_total
- awake
- bodywt
- brainwt

- Mas... a coluna que mostra a dieta (vore) tem só 7 NA. Quais são?

```
df %>%
  filter(is.na(vore)) %>%
  select(name)
## # A tibble: 7 x 1
##   name
##   <chr>
## 1 Vesper mouse
```



```
## 2 Desert hedgehog
## 3 Deer mouse
## 4 Phalanger
## 5 Rock hyrax
## 6 Mole rat
## # ... with 1 more row
```

- OK. Vamos manter a coluna `vore` também, apesar dos NA. Quando formos usar esta variável, tomaremos cuidado.
- Também... a coluna `bodywt` tem 0 como valor mínimo. Como assim?

```
df %>%
  filter(bodywt < 1) %>%
  select(name, bodywt) %>%
  arrange(bodywt)
## # A tibble: 35 x 2
##   name                bodywt
##   <chr>              <dbl>
## 1 Lesser short-tailed shrew 0.005
## 2 Little brown bat        0.01
## 3 Greater short-tailed shrew 0.019
## 4 Deer mouse              0.021
## 5 House mouse             0.022
## 6 Big brown bat           0.023
## # ... with 29 more rows
```

- Ah, sem problema. A função `dfSummary` arredondou estes pesos para 0. Os valores de verdade ainda estão na *tibble*.
- Vamos criar uma *tibble* nova, só com as colunas que nos interessam:

```
sono <- df %>%
  select(
    name, order, genus, vore, bodywt,
    brainwt, awake, sleep_total
  )
```

- Vamos ver o sumário:

```
sono %>% dfSummary() %>% print()
```

Variável	Estatísticas / Valores	Freqs (% de Válidos)	Faltante
name [character]	1. African elephant 2. African giant pouched rat 3. African striped mouse 4. Arctic fox 5. Arctic ground squirrel 6. Asian elephant 7. Baboon 8. Big brown bat 9. Bottle-nosed dolphin 10. Brazilian tapir [73 outros]	1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 73 (88,0%)	0 (0,0%)
order [character]	1. Rodentia 2. Carnivora 3. Primates 4. Artiodactyla 5. Soricomorpha 6. Cetacea 7. Hyracoidea 8. Perissodactyla 9. Chiroptera 10. Cingulata [9 outros]	22 (26,5%) 12 (14,5%) 12 (14,5%) 6 (7,2%) 5 (6,0%) 3 (3,6%) 3 (3,6%) 3 (3,6%) 2 (2,4%) 2 (2,4%) 13 (15,7%)	0 (0,0%)
genus [character]	1. Panthera 2. Spermophilus 3. Equus 4. Vulpes 5. Acinonyx 6. Aotus 7. Aplodontia 8. Blarina 9. Bos 10. Bradypus [67 outros]	3 (3,6%) 3 (3,6%) 2 (2,4%) 2 (2,4%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 1 (1,2%) 67 (80,7%)	0 (0,0%)
vore [character]	1. carni 2. herbi 3. insecti 4. omni	19 (25,0%) 32 (42,1%) 5 (6,6%) 20 (26,3%)	7 (8,4%)
bodywt [numeric]	Média (dp) : 166,1 (786,8) mín < mediana < máx: 0 < 1,7 < 6654 IQE (CV) : 41,6 (4,7)	82 valores distintos	0 (0,0%)
brainwt [numeric]	Média (dp) : 0,3 (1) mín < mediana < máx: 0 < 0 < 5,7 IQE (CV) : 0,1 (3,5)	53 valores distintos	27 (32,5%)

Variável	Estatísticas / Valores	Freqs (% de Válidos)	Faltante
awake [numeric]	Média (dp) : 13,6 (4,5) mín < mediana < máx: 4,1 < 13,9 < 22,1 IQE (CV) : 5,9 (0,3)	65 valores distintos	0 (0,0%)
sleep_total [numeric]	Média (dp) : 10,4 (4,5) mín < mediana < máx: 1,9 < 10,1 < 19,9 IQE (CV) : 5,9 (0,4)	65 valores distintos	0 (0,0%)

4.4

Gráficos de dispersão (*scatter plots*)

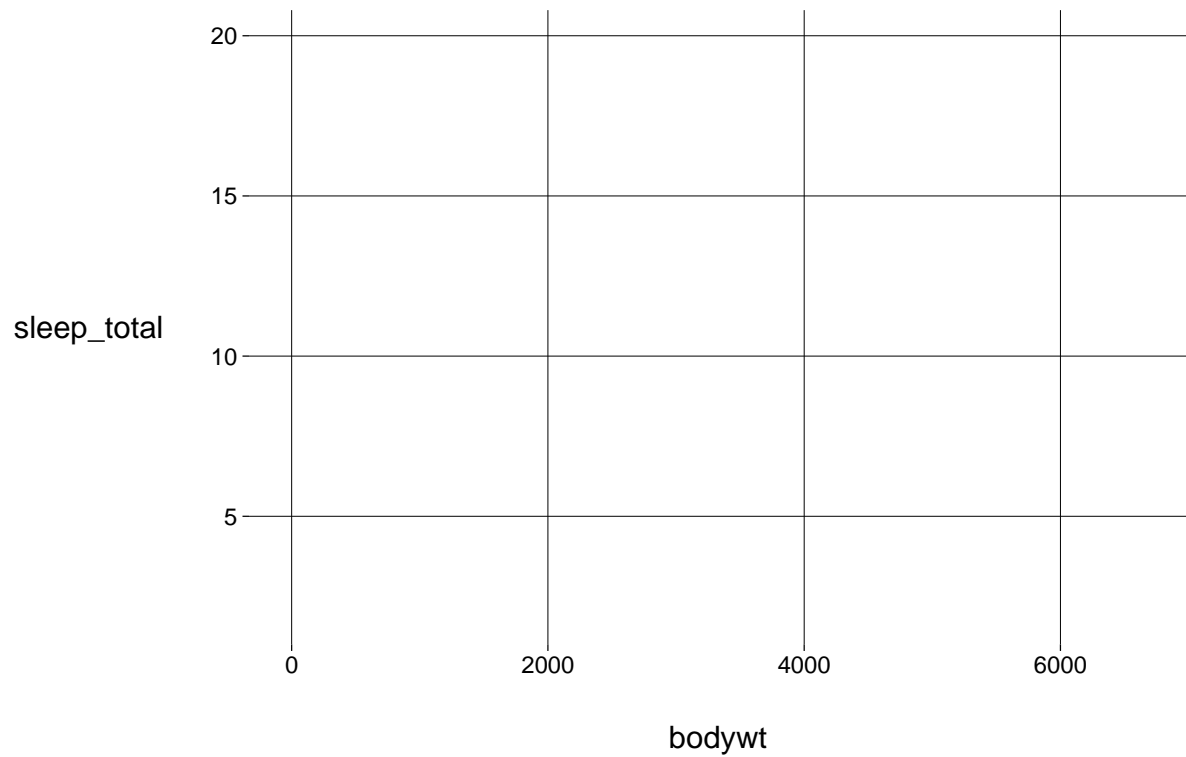
- Servem para visualizar a *relação* entre **duas variáveis quantitativas**.
- **Essa relação não é necessariamente de causa e efeito.**
- Isto é, a variável do eixo horizontal não determina, necessariamente, os valores da variável do eixo vertical.
- Pense em **associação**, **correlação**, não em causalidade.
- Troque as variáveis de eixo, se ajudar a deixar isto claro.

4.4.1

Horas de sono e peso corporal

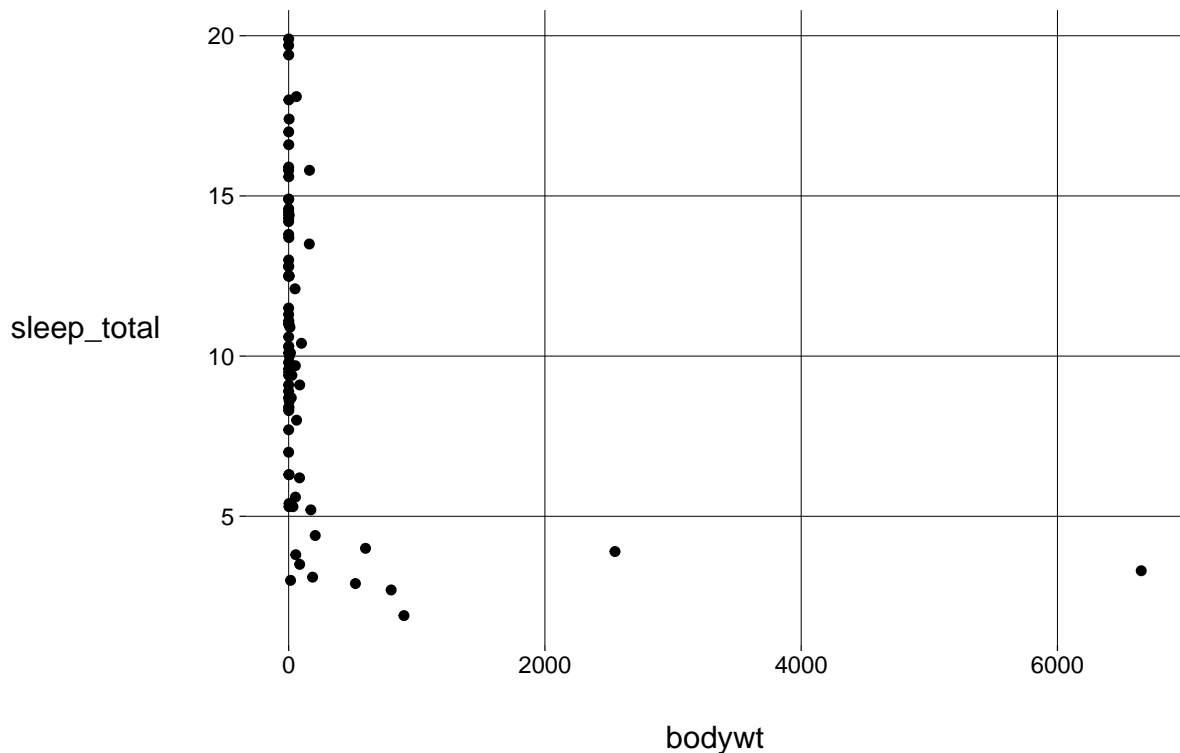
- Como as variáveis `sleep_total` e `bodywt` estão relacionadas?

```
sono %>%
  ggplot(aes(x = bodywt, y = sleep_total))
```



- O que houve? Cadê os pontos?
- O problema foi que só especificamos o mapeamento estético (com `aes`, que são as iniciais de *aesthetics*). **Faltou a geometria.**

```
sono %>%  
  ggplot(aes(x = bodywt, y = sleep_total)) +  
  geom_point()
```



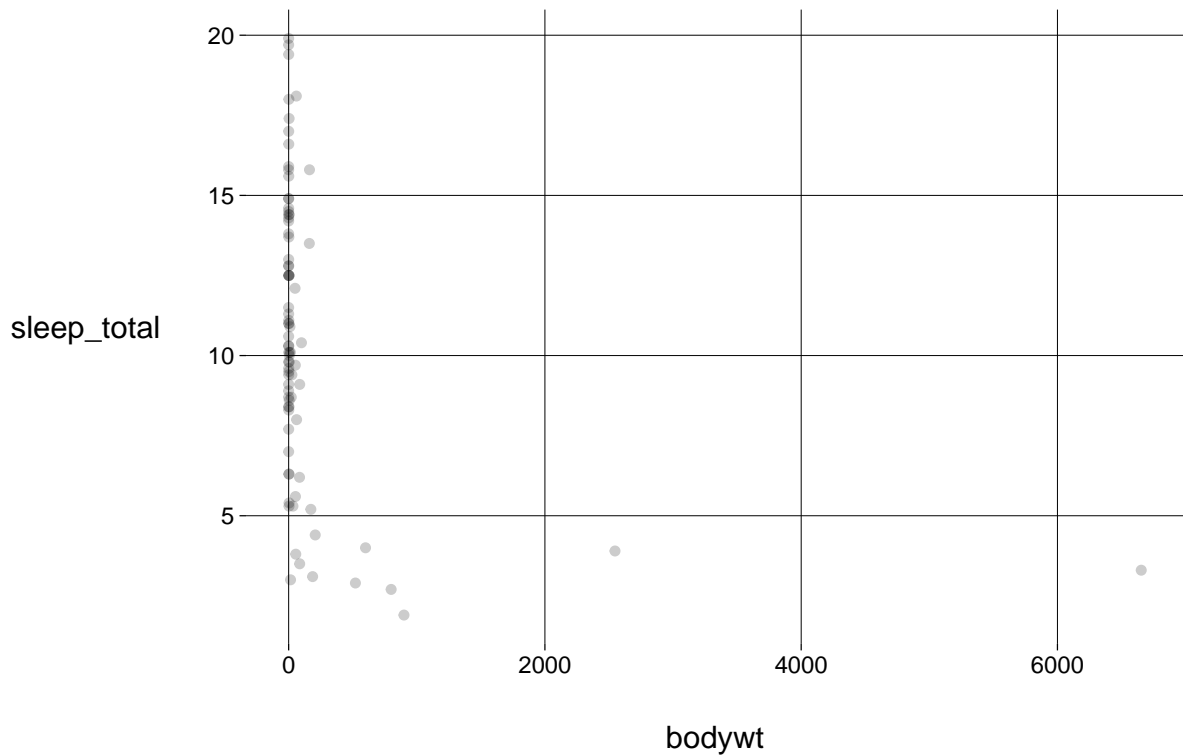
- Que horror.
- A única coisa que percebemos aqui é que os mamíferos muito pesados dormem menos de 5 horas por noite.
- Estes animais muito pesados estão estragando a escala do eixo x .
- Que animais são estes?

```
sono %>%
  filter(bodywt > 250) %>%
  select(name, bodywt) %>%
  arrange(bodywt)
## # A tibble: 6 x 2
##   name          bodywt
##   <chr>         <dbl>
## 1 Horse          521
## 2 Cow            600
## 3 Pilot whale   800
## 4 Giraffe       900.
## 5 Asian elephant 2547
## 6 African elephant 6654
```

- Além disso, há muitos pontos sobrepostos. Em bom português, temos um problema de *overplotting*.
- Existem diversas maneiras de lidar com isso.
- A primeira delas é **alterando a opacidade dos pontos**. Isto é um ajuste na geometria

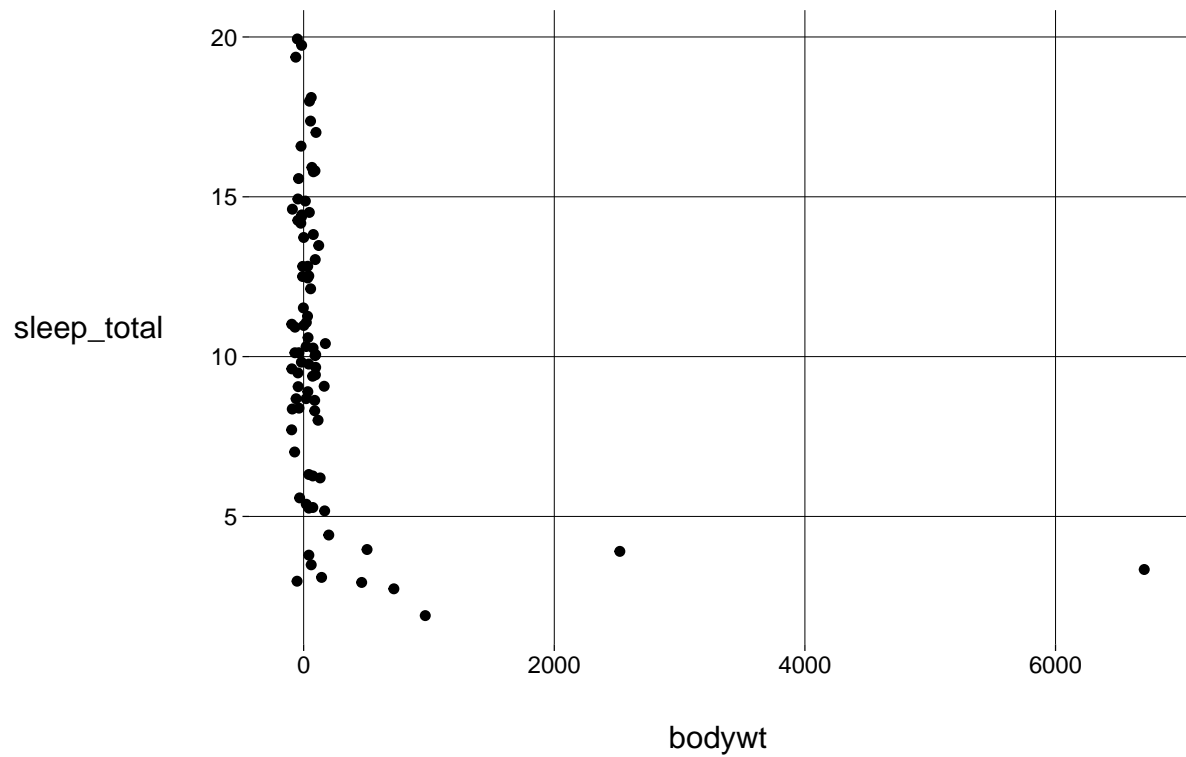
apenas, pois a opacidade, aqui, não representa informação nenhuma.

```
sono %>%  
  ggplot(aes(x = bodywt, y = sleep_total)) +  
    geom_point(alpha = 0.2)
```



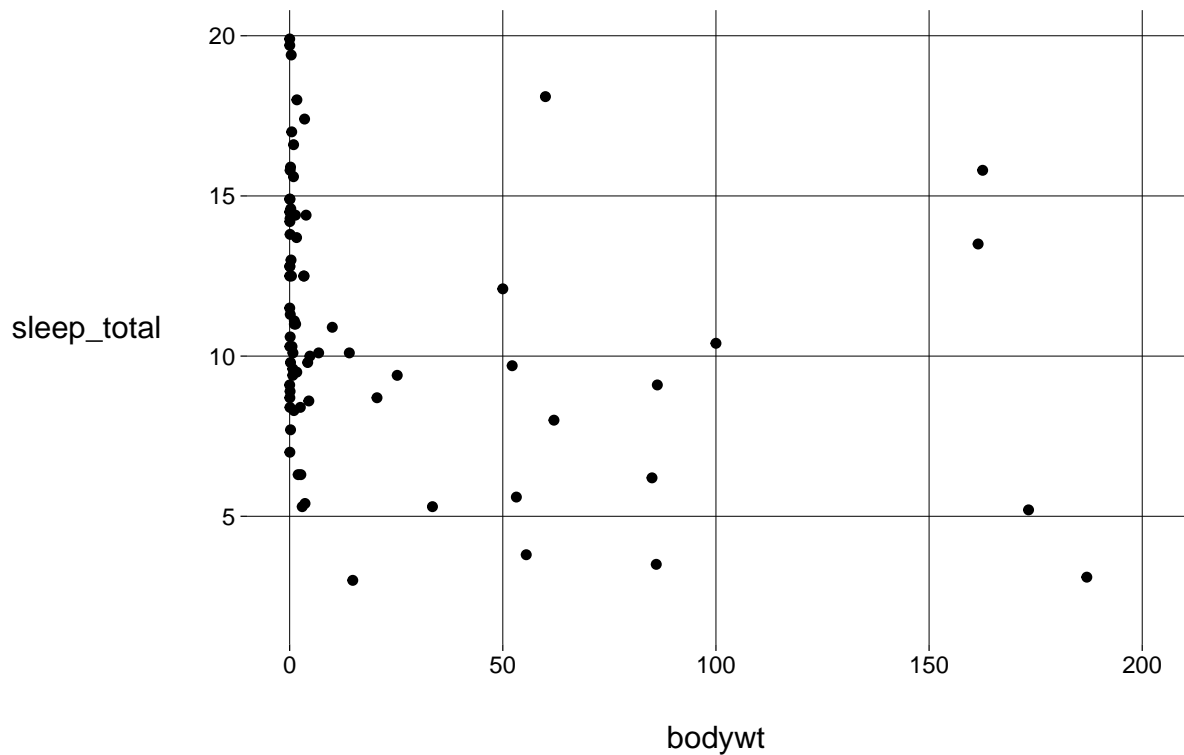
- Outra maneira é usar `geom_jitter` em vez de `geom_point`. “*Jitter*” significa “tremer”. As posições dos pontos são ligeiramente perturbadas, para evitar colisões. Perdemos precisão, mas a visualização fica melhor.

```
sono %>%  
  ggplot(aes(x = bodywt, y = sleep_total)) +  
    geom_jitter(width = 100)
```



- Vamos mudar os limites do gráfico para nos concentrarmos nos animais menos pesados. Observe que **isto é um ajuste na escala.**

```
sono %>%  
  ggplot(aes(x = bodywt, y = sleep_total)) +  
    geom_point() +  
    scale_x_continuous(limits = c(0, 200))  
## Warning: Removed 7 rows containing missing values (geom_point).
```



- Nestes limites, a relação entre horas de sono e peso não é mais tão pronunciada.

4.4.2

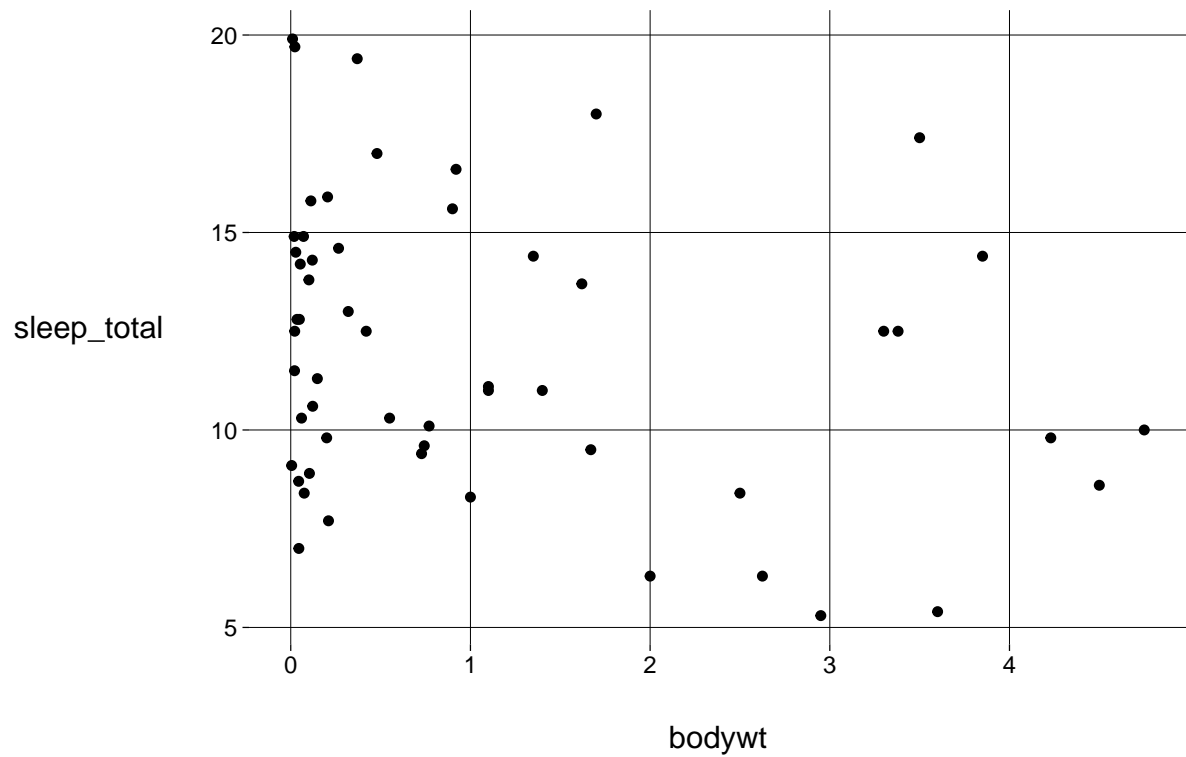
Horas de sono e peso corporal para animais pequenos

- Vamos restringir o gráfico a animais com no máximo 5kg.

```
limite <- 5
```

- Em vez de mudar a escala do gráfico, vamos filtrar as linhas do *data frame*:

```
sono %>%
  filter(bodywt < limite) %>%
  ggplot(aes(x = bodywt, y = sleep_total)) +
  geom_point()
```

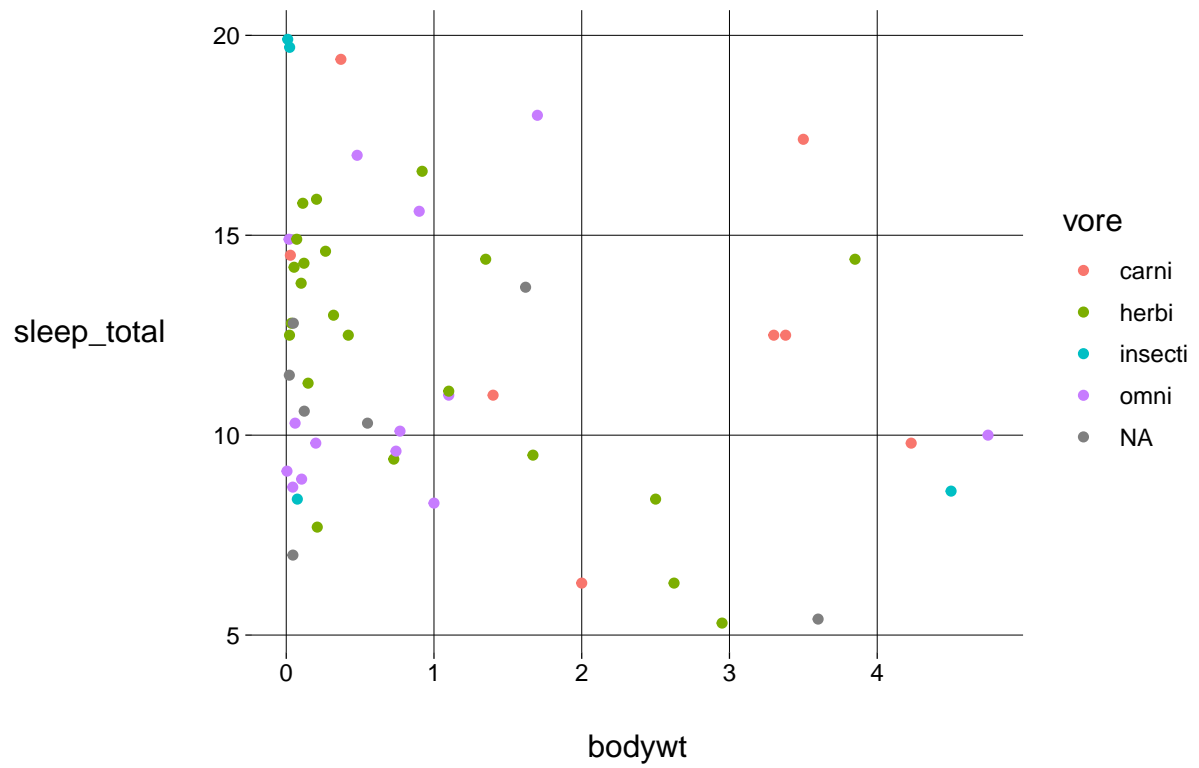



4.4.3

Incluindo a dieta

- Com a estética `color`. Observe como a legenda aparece automaticamente.

```
sono %>%  
  filter(bodywt < limite) %>%  
  ggplot(aes(x = bodywt, y = sleep_total, color = vore)) +  
    geom_point()
```

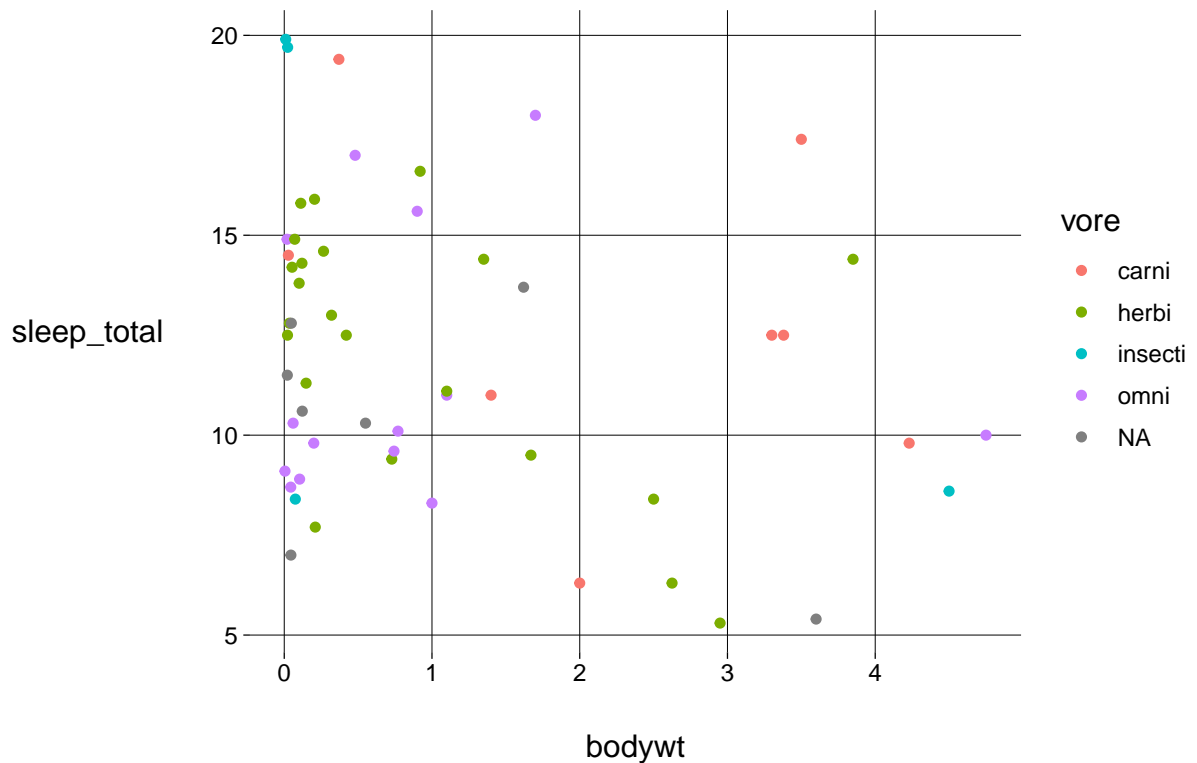


4.4.4

A estética pode ser especificada na `geom`

- Compare com o código anterior.

```
sono %>%
  filter(bodywt < limite) %>%
  ggplot() +
    geom_point(aes(x = bodywt, y = sleep_total, color = vore))
```



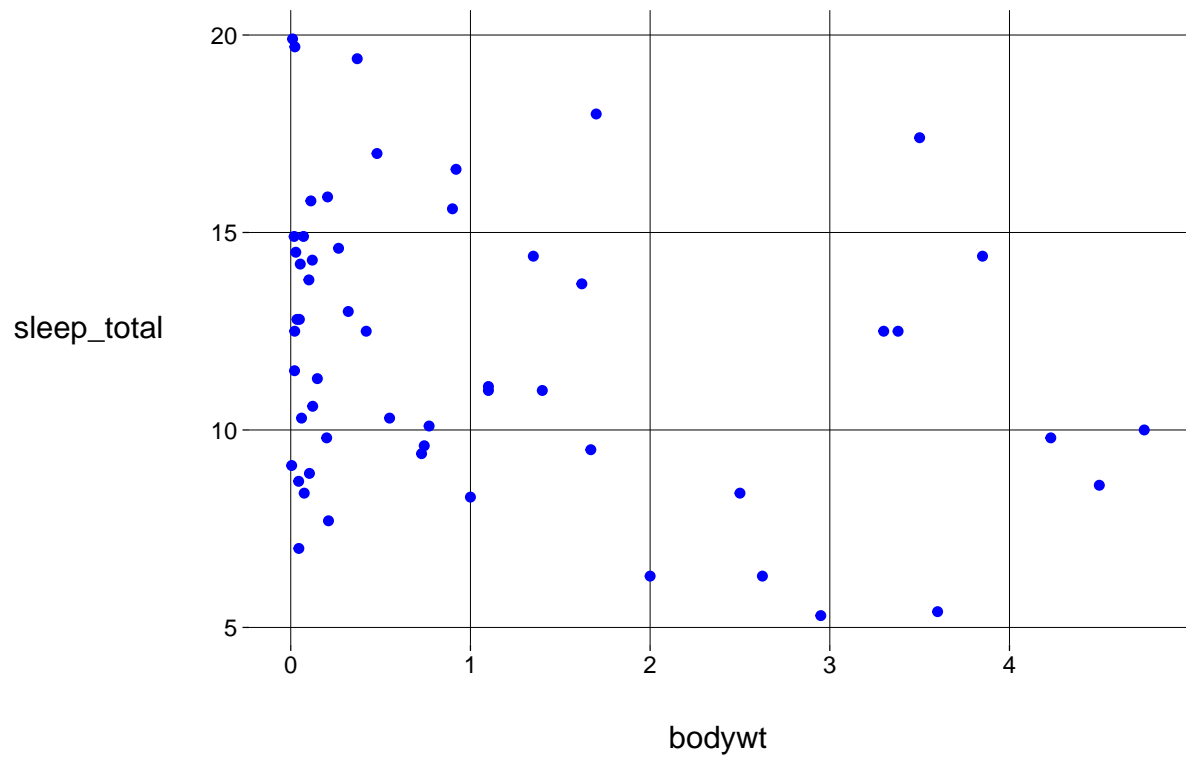
- Fazendo deste modo, a estética só vale para uma geometria. Se você acrescentar outras geometrias (linhas, por exemplo), a estética não valerá para elas.

4.4.5

Aparência fixa ou dependendo de variável?

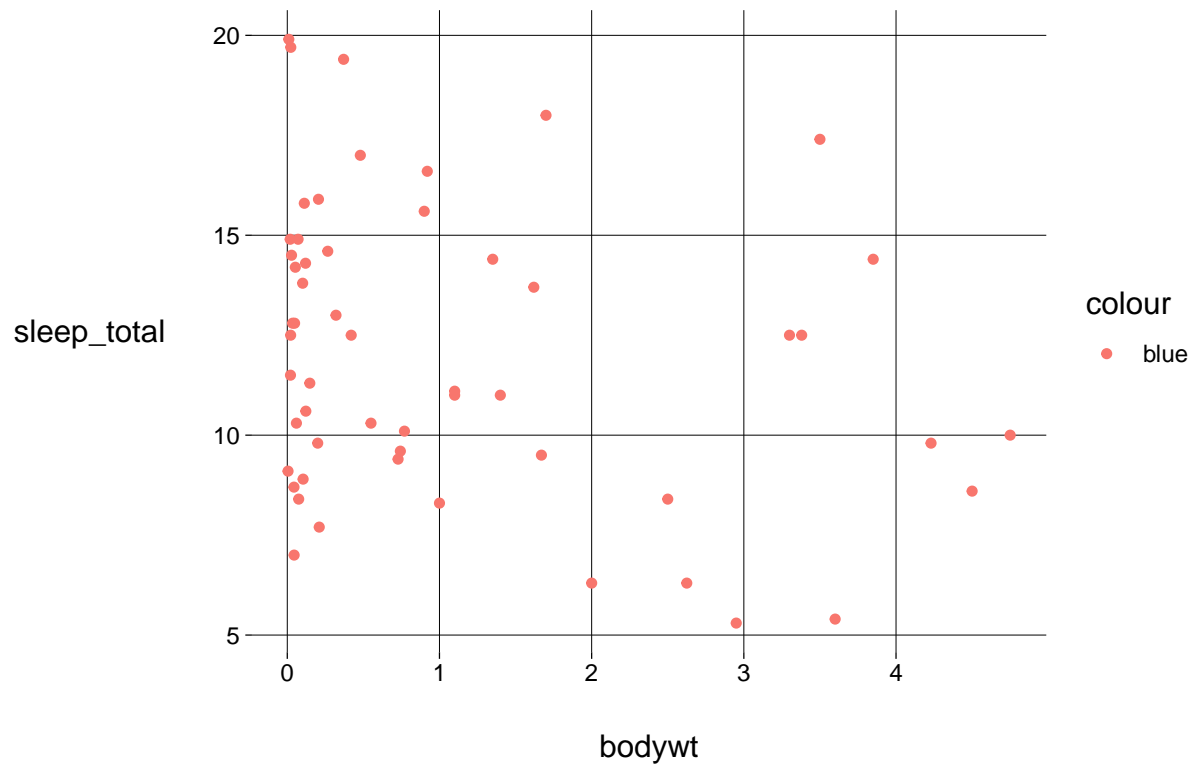
- Se for fixa, não é estética. Não representa informação.
- Se depender de variável, é estética. Representa informação.
- Compare o último *chunk* acima com:

```
sono %>%
  filter(bodywt < limite) %>%
  ggplot() +
    geom_point(aes(x = bodywt, y = sleep_total), color = 'blue')
```



- Se for uma estética, precisa estar associada a uma variável, não a um valor fixo. Um erro comum seria fazer:

```
sono %>%  
  filter(bodywt < limite) %>%  
  ggplot() +  
    geom_point(aes(x = bodywt, y = sleep_total, color = 'blue'))
```

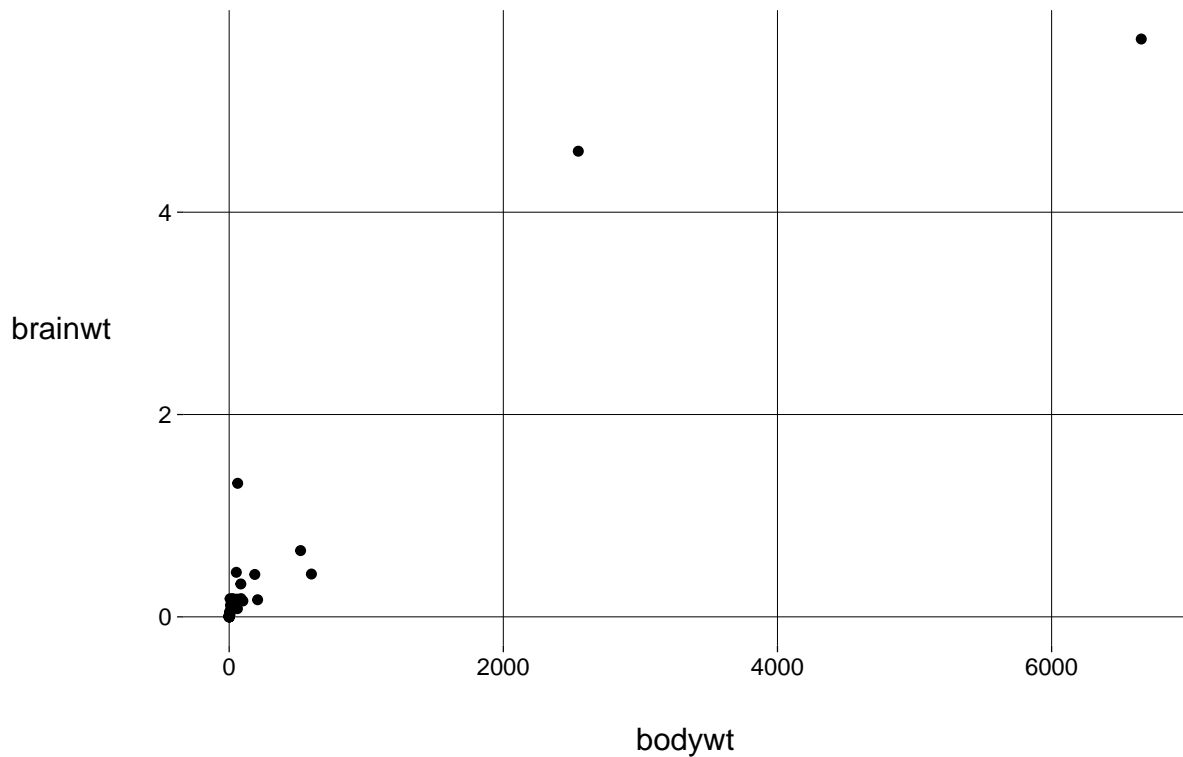


4.4.6

Uma correlação mais clara

- Peso cerebral versus peso corporal:

```
sono %>%  
  ggplot(aes(x = bodywt, y = brainwt)) +  
    geom_point()  
## Warning: Removed 27 rows containing missing values (geom_point).
```



- A mensagem de aviso (*warning*) diz que há 27 valores faltantes (NA) em bodywt ou brainwt. De fato:

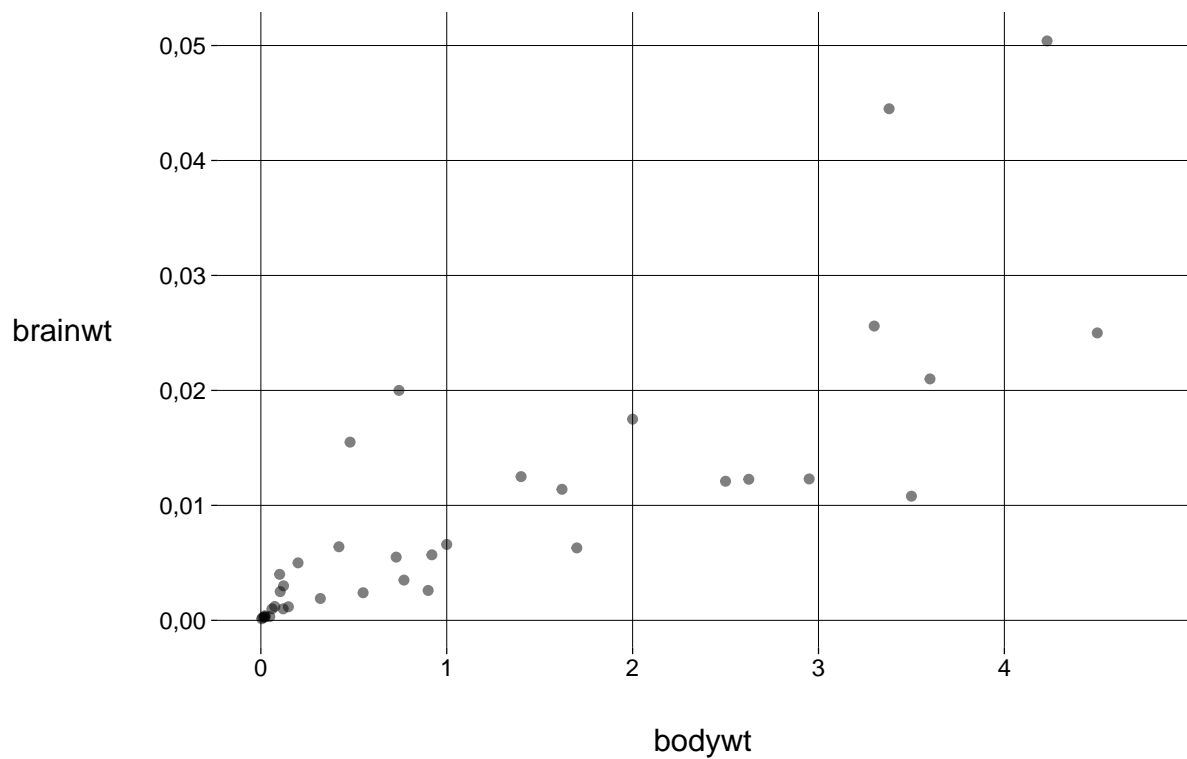
```
sono %>%
  filter(is.na(bodywt)) %>%
  count()
## # A tibble: 1 x 1
##       n
##   <int>
## 1     0
```

```
sono %>%
  filter(is.na(brainwt)) %>%
  count()
## # A tibble: 1 x 1
##       n
##   <int>
## 1    27
```

- Vamos restringir aos animais mais leves e mudar a opacidade:

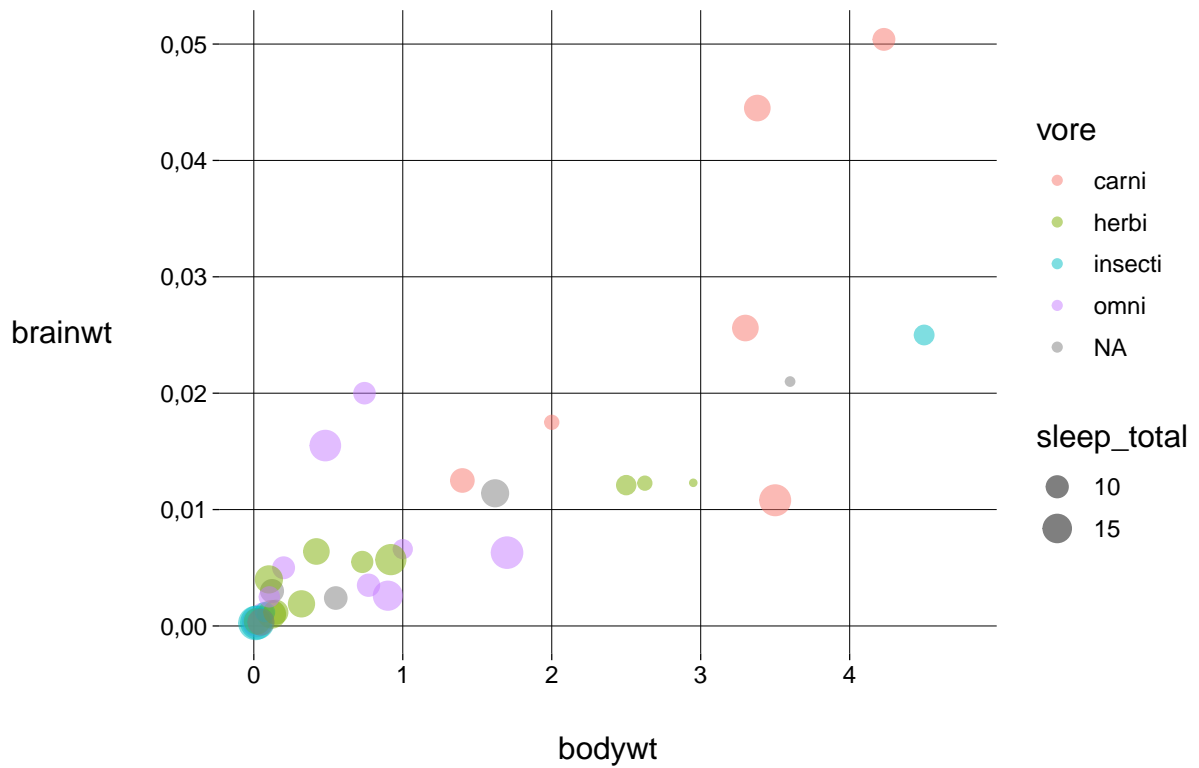
```
sono %>%
  filter(bodywt < limite) %>%
  ggplot(aes(x = bodywt, y = brainwt)) +
  geom_point(alpha = .5)
```

```
## Warning: Removed 18 rows containing missing values (geom_point).
```



- Vamos incluir horas de sono e dieta. Observe as estéticas usadas.

```
sono %>%  
  filter(bodywt < limite) %>%  
  ggplot(  
    aes(  
      x = bodywt,  
      y = brainwt,  
      size = sleep_total,  
      color = vore  
    )  
  ) +  
  geom_point(alpha = .5)  
## Warning: Removed 18 rows containing missing values (geom_point).
```



- Vamos mudar a escala dos tamanhos e incluir rótulos:

```
grafico <- sono %>%
  filter(bodywt < limite) %>%
  ggplot(
    aes(
      x = bodywt,
      y = brainwt,
      size = sleep_total,
      color = vore
    )
  ) +
  geom_point(alpha = .5) +
  scale_size(
    breaks = seq(0, 24, 4)
  ) +
  labs(
    title = 'Peso do cérebro versus peso corporal',
    subtitle = paste0(
      'para mamíferos com menos de ',
      limite,
      ' kg'
    ),
    caption = 'Fonte: dataset `msleep`',
    x = 'Peso corporal (kg)',
  )
```



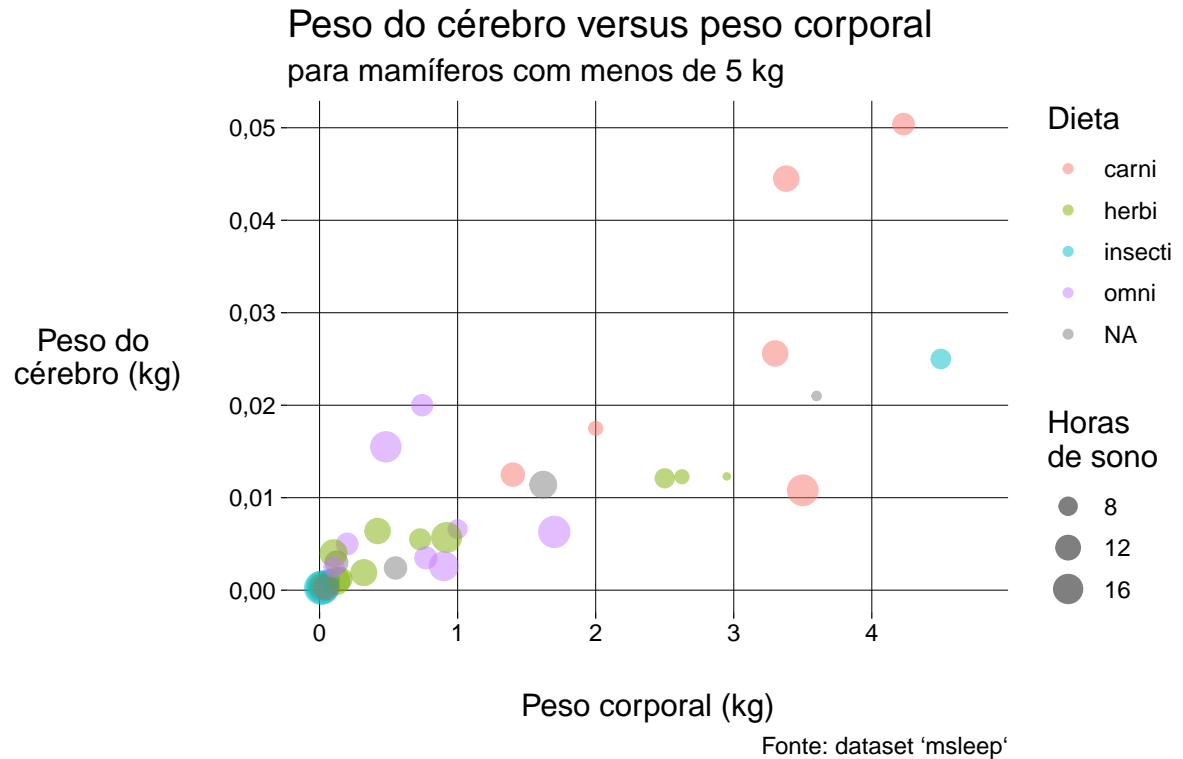
```

y = 'Peso do\n cérebro (kg)',
color = 'Dieta',
size = 'Horas\nde sono'
)

```

grafico

Warning: Removed 18 rows containing missing values (geom_point).



- Vamos mudar as cores usadas para a dieta, usando uma escala diferente.

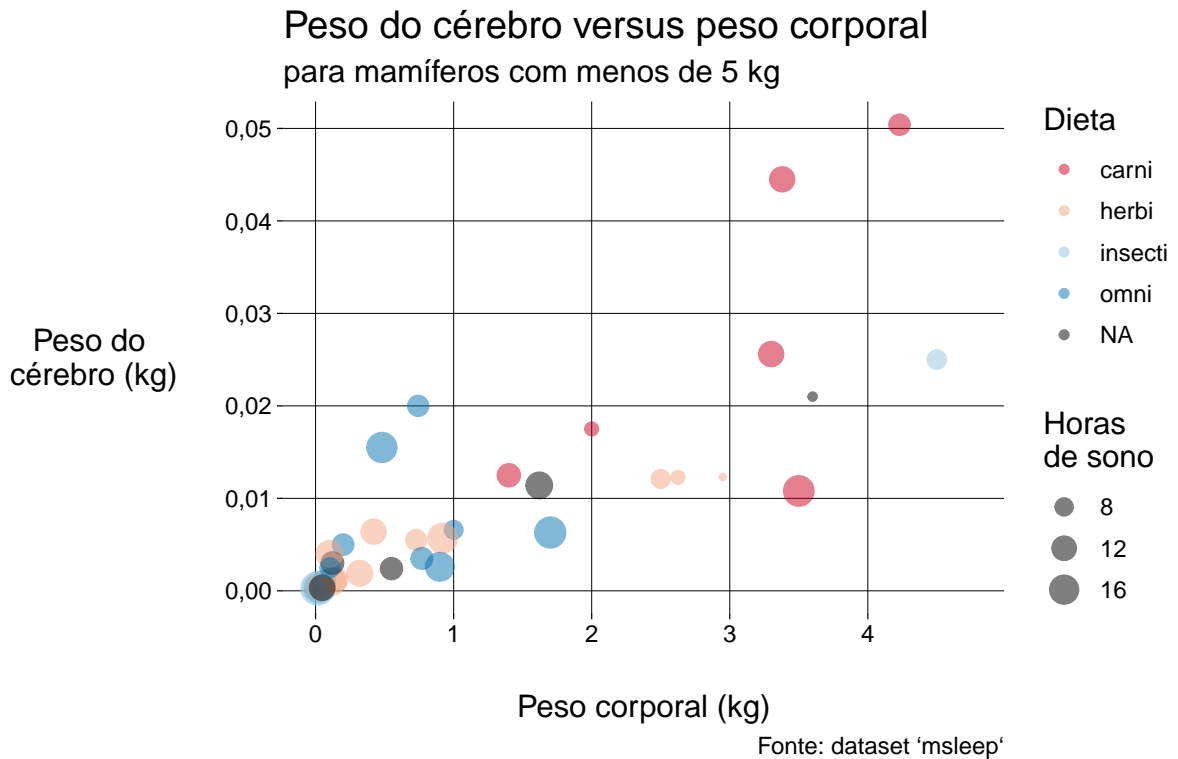
```

grafico2 <- grafico +
  scale_color_discrete(
    palette = 'RdBu',
    na.value = 'black',
    type = scale_color_brewer
  )

```

grafico2

Warning: Removed 18 rows containing missing values (geom_point).

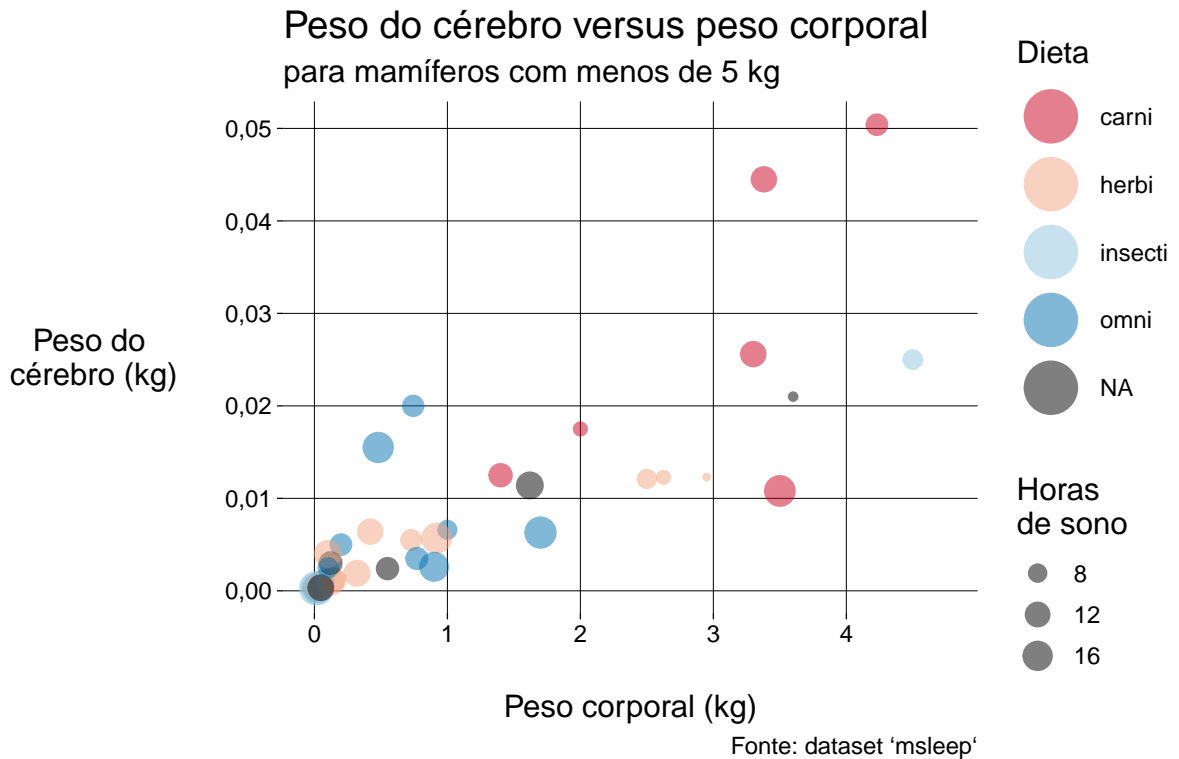


- Observe como usamos o gráfico já salvo na variável `grafico` e simplesmente acrescentamos a nova escala. Este tipo de “montagem” de gráficos `ggplot2` é bem conveniente, para evitar repetição de código.
- Um último ajuste na aparência: os pontos na legenda “Dieta” estão pequenos demais. Quase não identificamos as cores deles.

Vamos usar a função `guides` para modificar (*override*) a estética `color` — apenas na legenda, não nos pontos mostrados no gráfico, cujos tamanhos representam o número de horas de sono — tornando o tamanho maior. [Leia mais sobre `override.aes` neste link \(em inglês\)](#).

```
grafico3 <- grafico2 +
  guides(color = guide_legend(override.aes = list(size = 10)))

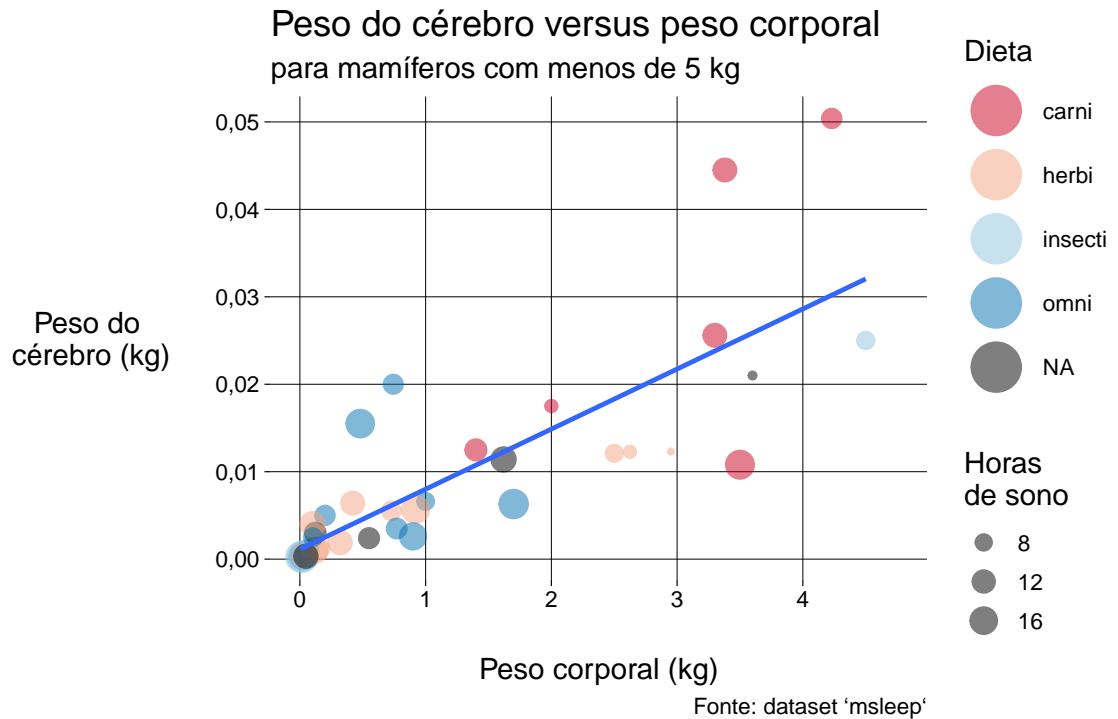
grafico3
## Warning: Removed 18 rows containing missing values (geom_point).
```



- Agora podemos finalmente comentar sobre a informação que o gráfico mostra sobre os dados:
 - De fato, existe uma correlação entre peso cerebral e peso corporal: quanto maior o peso corporal, maior o peso cerebral. Nada surpreendente.
 - Podemos fazer o `ggplot2` traçar uma reta de regressão com a geometria `geom_smooth`. Vamos falar mais sobre correlação **em um capítulo futuro**.

```
grafico4 <- grafico3 +
  geom_smooth(
    aes(group = 1),
    show.legend = FALSE,
    method = 'lm',
    se = FALSE
  )

grafico4
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 18 rows containing non-finite values (stat_smooth).
## Warning: Removed 18 rows containing missing values (geom_point).
```



- Todos os carnívoros têm peso corporal maior que 1kg e peso cerebral maior ou igual a 10g.
- Só um carnívoro dorme 8 horas ou menos. Qual?
- Todos os insetívoros — com exceção de um (qual?) — são muito leves e dormem muito.
- Todos os onívoros têm menos de 2kg de peso corporal e 20g ou menos de peso cerebral.

4.5

Vídeo 2

<https://youtu.be/c-LoZ9e8xWc>

4.6

Histogramas e cia.

- A idéia agora é agrupar indivíduos em classes, dependendo do valor de uma variável quantitativa.

4.6.1

Distribuições de frequência

- Vamos nos concentrar nas horas de sono.

```
sono$sleep_total
## [1] 12,1 17,0 14,4 14,9 4,0 14,4 8,7 7,0 10,1 3,0 5,3 9,4 10,0
## [14] 12,5 10,3 8,3 9,1 17,4 5,3 18,0 3,9 19,7 2,9 3,1 10,1 10,9
## [27] 14,9 12,5 9,8 1,9 2,7 6,2 6,3 8,0 9,5 3,3 19,4 10,1 14,2
## [40] 14,3 12,8 12,5 19,9 14,6 11,0 7,7 14,5 8,4 3,8 9,7 15,8 10,4
## [53] 13,5 9,4 10,3 11,0 11,5 13,7 3,5 5,6 11,1 18,1 5,4 13,0 8,7
## [66] 9,6 8,4 11,3 10,6 16,6 13,8 15,9 12,8 9,1 8,6 15,8 4,4 15,6
## [79] 8,9 5,2 6,3 12,5 9,8
```

- Antes de montar o histograma, vamos construir uma **distribuição de frequência**.
- A **amplitude** é a diferença entre o valor máximo e o valor mínimo. A função `range` não retorna a amplitude, mas sim os valores mínimo e máximo:

```
sono$sleep_total %>% range()
## [1] 1,9 19,9
```

- Vamos decidir que cada classe vai ter 2 horas. A função `cut` substitui os valores do vetor pelos nomes das classes:

```
sono$sleep_total %>%
  cut(breaks = seq(0, 20, 2), right = FALSE)
## [1] [12,14) [16,18) [14,16) [14,16) [4,6) [14,16) [8,10) [6,8)
## [9] [10,12) [2,4) [4,6) [8,10) [10,12) [12,14) [10,12) [8,10)
## [17] [8,10) [16,18) [4,6) [18,20) [2,4) [18,20) [2,4) [2,4)
## [25] [10,12) [10,12) [14,16) [12,14) [8,10) [0,2) [2,4) [6,8)
## [33] [6,8) [8,10) [8,10) [2,4) [18,20) [10,12) [14,16) [14,16)
## [41] [12,14) [12,14) [18,20) [14,16) [10,12) [6,8) [14,16) [8,10)
## [49] [2,4) [8,10) [14,16) [10,12) [12,14) [8,10) [10,12) [10,12)
## [57] [10,12) [12,14) [2,4) [4,6) [10,12) [18,20) [4,6) [12,14)
## [65] [8,10) [8,10) [8,10) [10,12) [10,12) [16,18) [12,14) [14,16)
## [73] [12,14) [8,10) [8,10) [14,16) [4,6) [14,16) [8,10) [4,6)
## [81] [6,8) [12,14) [8,10)
## 10 Levels: [0,2) [2,4) [4,6) [6,8) [8,10) [10,12) [12,14) ... [18,20)
```

- A função `table` faz a contagem dos elementos de cada classe:

```
sono$sleep_total %>%
  cut(breaks = seq(0, 20, 2), right = FALSE) %>%
  table(dnn = 'Horas de sono') %>%
  as.data.frame()
## # A tibble: 10 x 2
```

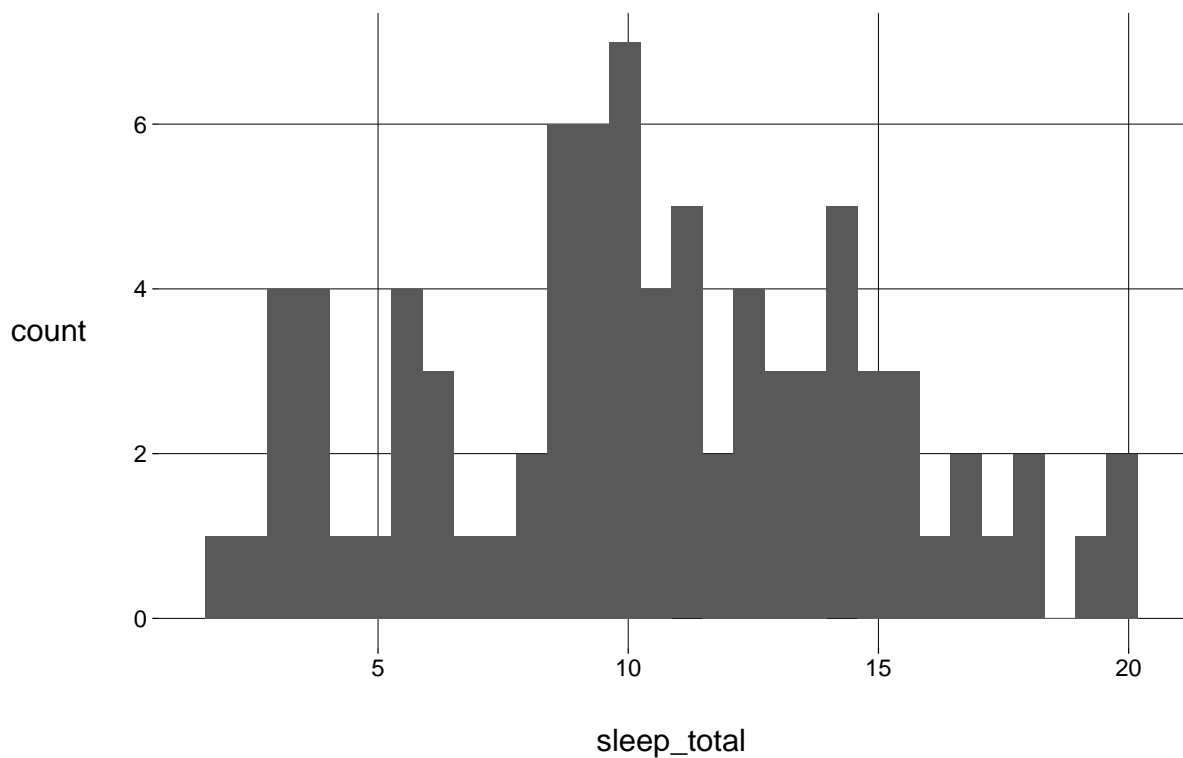
```
##   Horas.de.sono   Freq
##   <fct>         <int>
## 1 [0,2)          1
## 2 [2,4)          8
## 3 [4,6)          7
## 4 [6,8)          5
## 5 [8,10)         17
## 6 [10,12)        14
## # ... with 4 more rows
```

4.6.2

Histograma

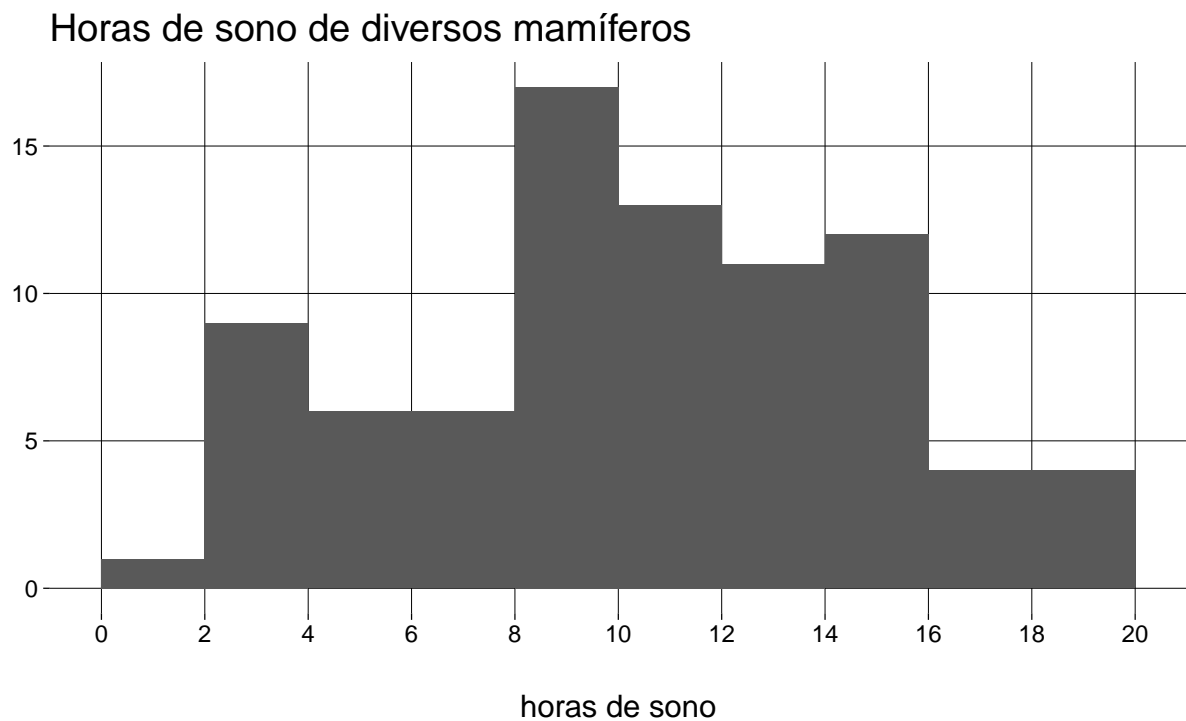
- Na verdade, o ggplot2 já faz esses cálculos para nós.
- O *default* é criar 30 classes (*bins*):

```
sono %>%
  ggplot(aes(x = sleep_total)) +
  geom_histogram()
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- Vamos mudar isto passando um vetor de limites das classes (*breaks*). Vamos acrescentar rótulos também:

```
sono %>%
  ggplot(aes(x = sleep_total)) +
    geom_histogram(breaks = seq(0, 20, 2)) +
    scale_x_continuous(breaks = seq(0, 20, 2)) +
    labs(
      title = 'Horas de sono de diversos mamíferos',
      x = 'horas de sono',
      y = NULL,
      caption = 'Fonte: dataset `msleep`'
    )
)
```



Fonte: dataset 'msleep'

- Nossas impressões:

- A classe que mais tem elementos é a de 8 a 10 horas.
- A distribuição é mais ou menos simétrica.
- A distribuição tem forma aproximada de sino: há poucos mamíferos com valores extremos de horas de sono; a maioria está próxima do valor médio:

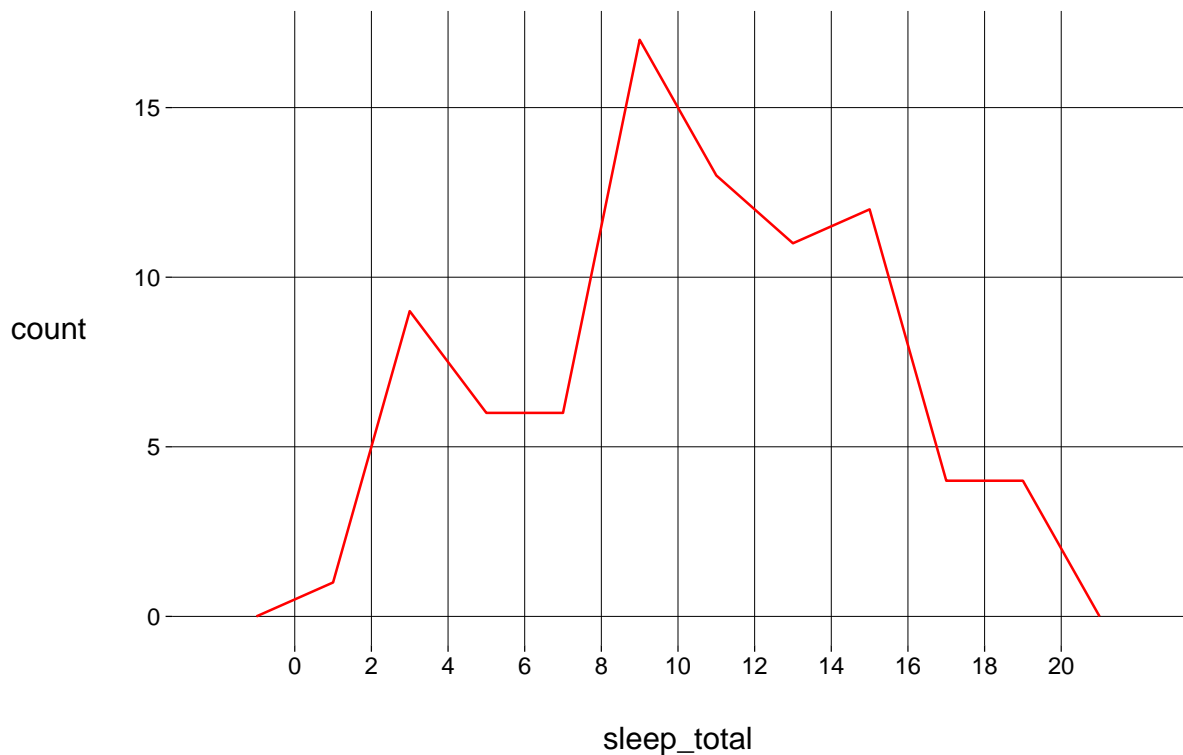
```
mean(sono$sleep_total)
## [1] 10,43373
```

4.6.3

Polígono de frequência

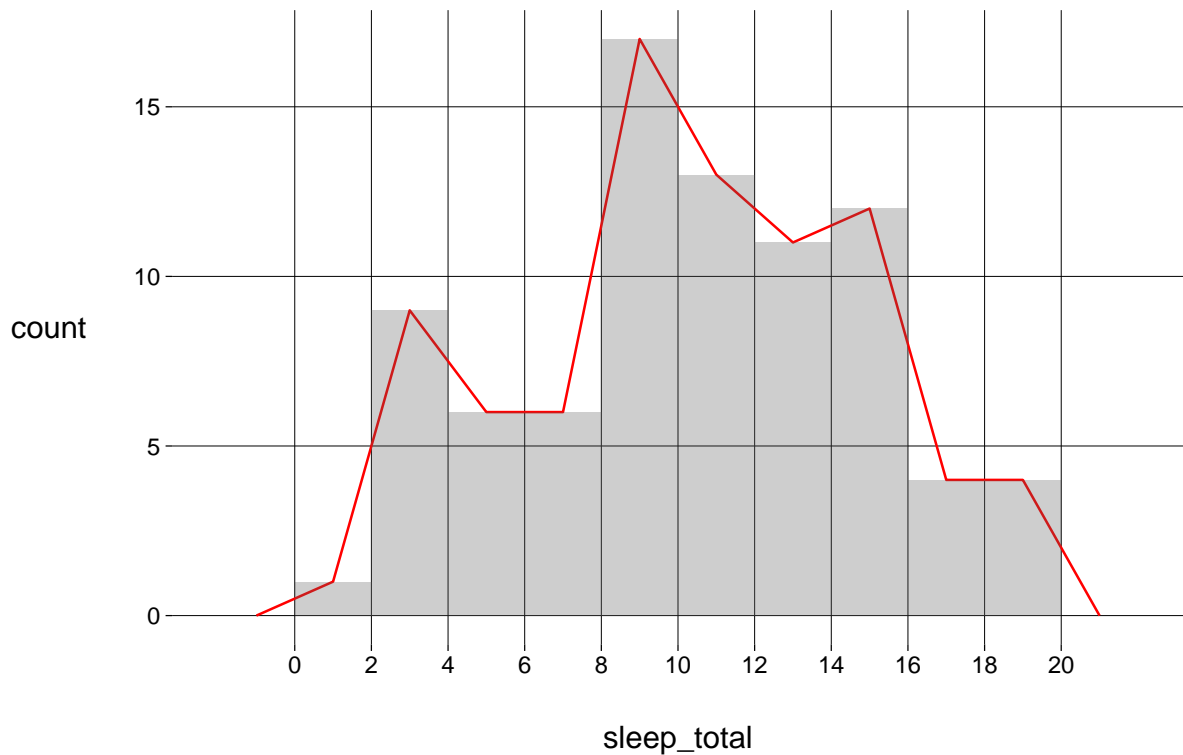
- Em vez das barras do histograma, podemos desenhar uma linha ligando seus topos.
- O resultado é um **polígono de frequência**.

```
pf <- sono %>%  
  ggplot(aes(x = sleep_total)) +  
    geom_freqpoly(breaks = seq(0, 20, 2), color = 'red') +  
    scale_x_continuous(breaks = seq(0, 20, 2))  
  
pf
```



- Vamos sobrepor o polígono de frequência ao histograma, para deixar claro o que está acontecendo:

```
pf + geom_histogram(breaks = seq(0, 20, 2), alpha = .3)
```

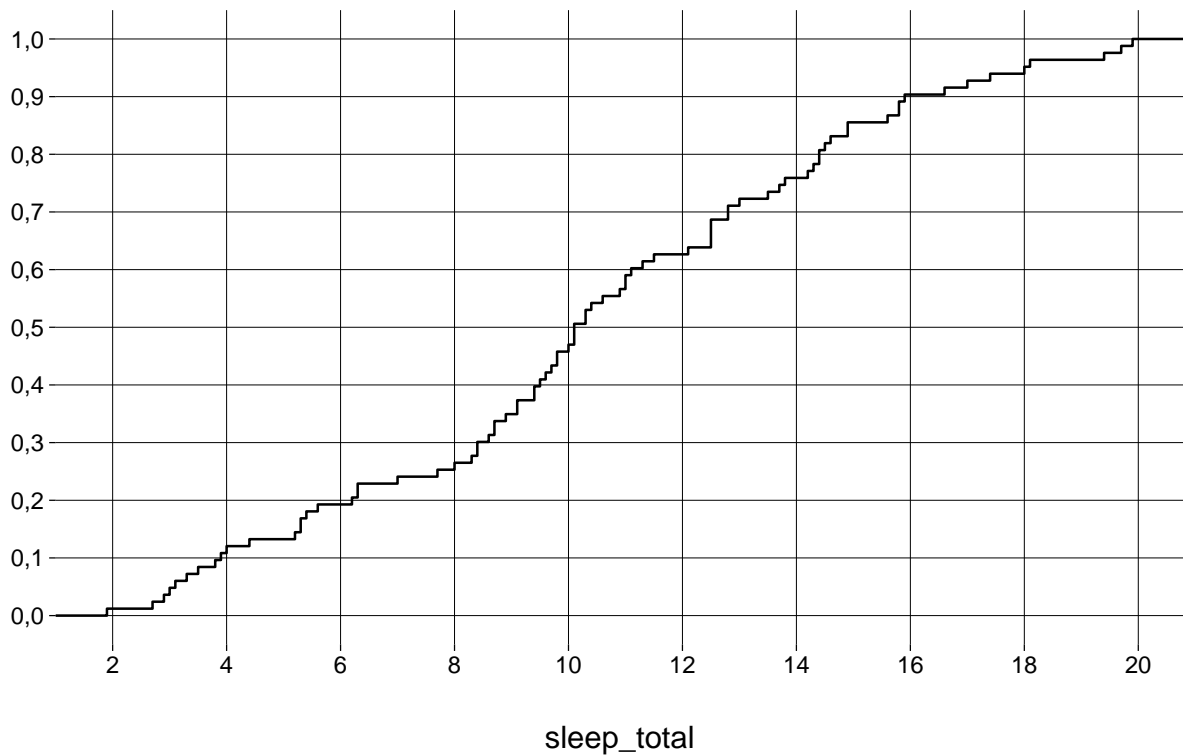



4.7

Ogiva

- A ogiva é um gráfico que mostra a **frequência acumulada**: para cada valor v da variável no eixo x , a proporção de indivíduos com valor menor ou igual a v .
- A geometria `geom_step` gera o gráfico de uma **função degrau**.
- Cada geometria está ligada a uma **stat**, um algoritmo para computar o que vai ser desenhado. Aqui, passamos para a geometria **a função `ecdf` (*empirical cumulative distribution function*)**, do pacote `stats`, que calcula as frequências acumuladas.

```
sono %>%
  ggplot(aes(x = sleep_total)) +
    geom_step(stat = 'ecdf') +
    scale_x_continuous(breaks = seq(0, 20, 2)) +
    scale_y_continuous(breaks = seq(0, 1, .1)) +
    labs(y = NULL)
```



- Com a ogiva, podemos obter informações difíceis de visualizar no histograma. Por exemplo:
 - Cerca de 20% dos mamíferos têm menos de 6 horas de sono.
 - Cerca de metade dos mamíferos têm menos de 10 horas de sono.
 - Cerca de 10% dos mamíferos têm mais de 16 horas de sono.

4.8

Ramos e folhas

- No início dos anos 1900, quando estatísticas eram feitas à mão, Arthur Bowley criou os **diagramas de ramos e folhas**.
- Um diagrama de ramos e folhas é, basicamente, uma listagem de todos os valores de uma variável, agrupados de maneira que todos os valores de uma classe (i.e., de uma linha) têm os algarismos iniciais dentro de um intervalo.
- Para as horas de sono dos mamíferos:

```
sono$sleep_total %>%
  stem()
##
##   The decimal point is at the |
##
```

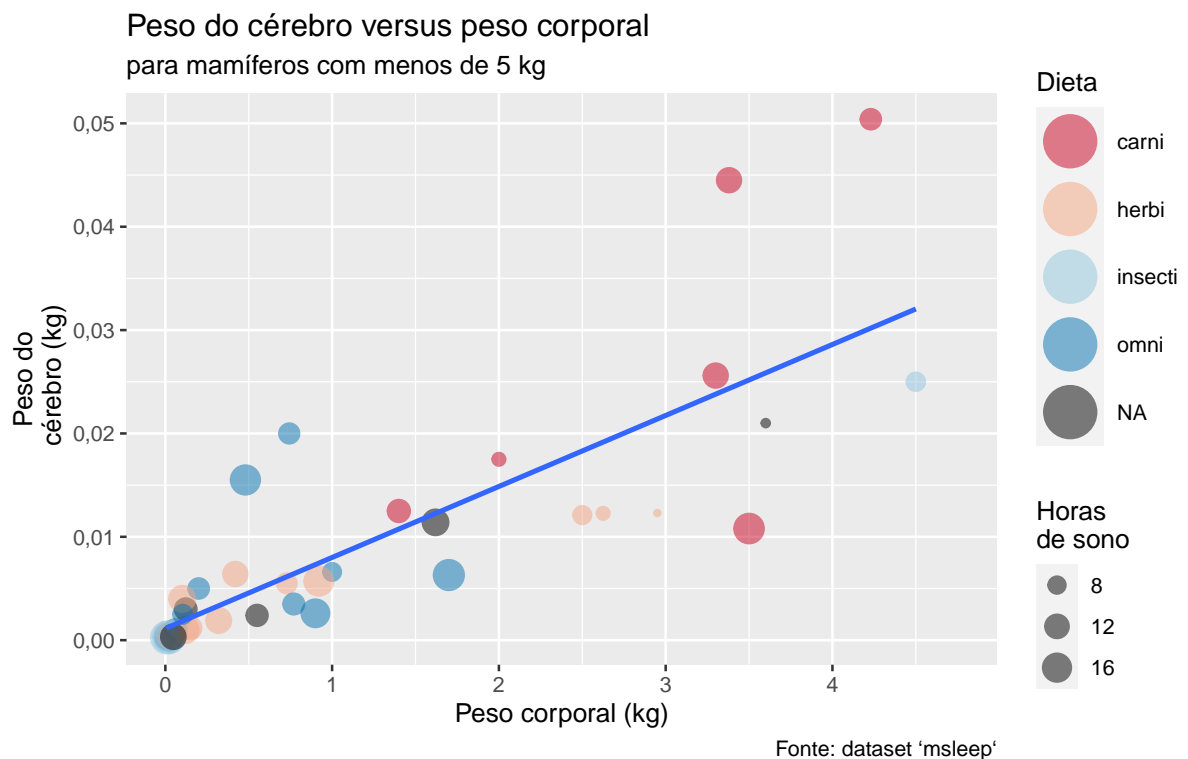
```
## 0 | 9
## 2 | 79013589
## 4 | 0423346
## 6 | 23307
## 8 | 03446779114456788
## 10 | 01113346900135
## 12 | 15555880578
## 14 | 234456996889
## 16 | 604
## 18 | 01479
```

- A primeira linha representa um indivíduo com 0,9 horas de sono.
- A penúltima linha representa 3 valores:
 - 16,6
 - 17,0
 - 17,4

4.9

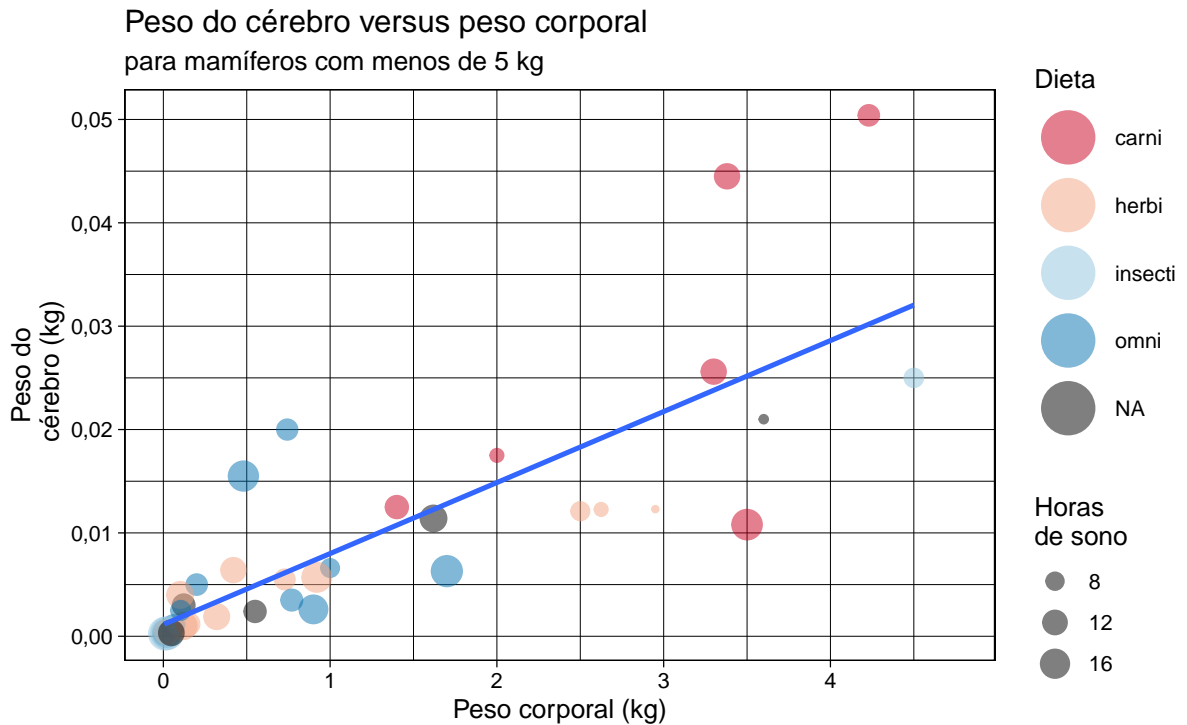
Personalização do tema

- O `ggplot2` tem um tema *default*, chamado `theme_gray`, que gera o *scatterplot* de um exemplo anterior deste capítulo do seguinte modo:

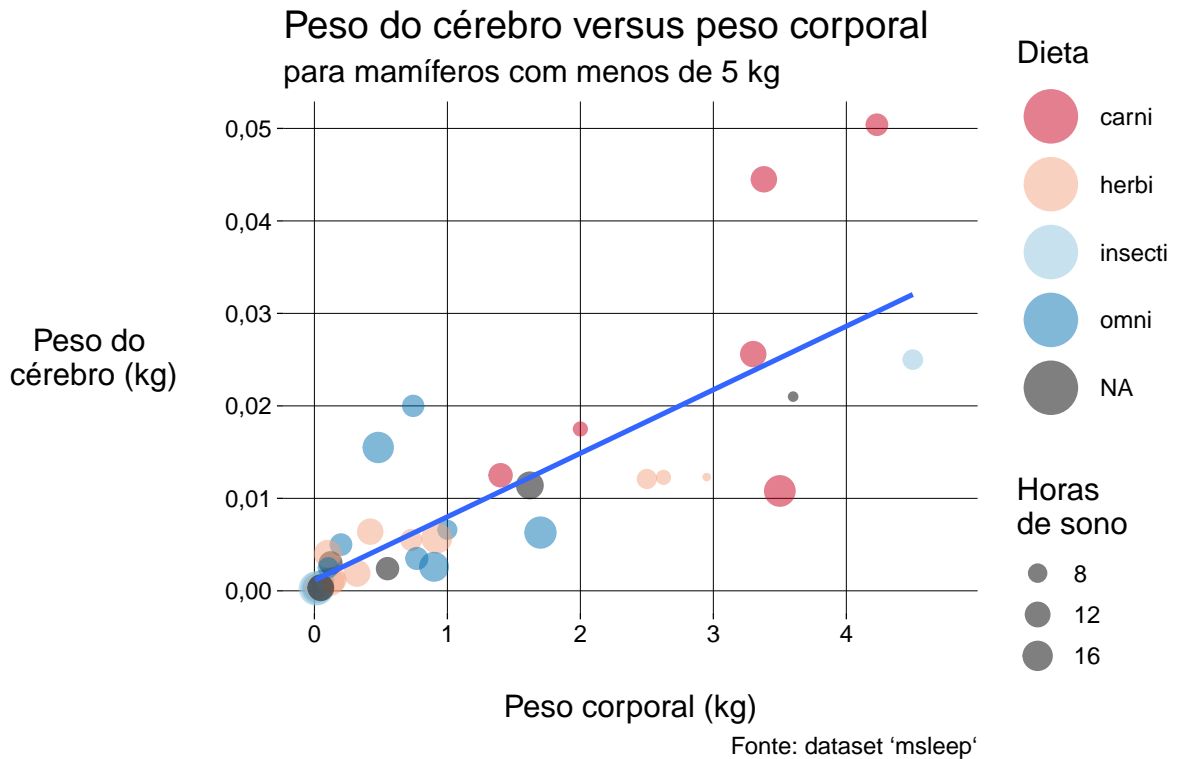


- Para este material, escolhi o tema `theme_linedraw`, que usa linhas pretas sobre fundo

branco:



- Para deixar os gráficos mais leves e facilitar a leitura, fiz as seguintes alterações no tema:
 - Mudei o tamanho do texto dos rótulos.
 - Fiz o rótulo do eixo y aparecer na horizontal; embora isto ocupe um pouco mais de espaço, evita que o leitor tenha que girar a cabeça para ler o rótulo.
 - Eliminei as linhas dos eixos, para o gráfico ficar mais leve.
 - Eliminei a moldura da área de dados, para o gráfico ficar mais leve.
 - Eliminei a grade secundária, para o gráfico ficar mais leve.
- O resultado é



- Os meus comandos para alterar o tema são

```
# Tamanho do texto depende do formato de saída (html ou pdf):
plot_text_size = ifelse(is_html_output(), 12, 13)

# Tema mais leve:
theme_set(
  theme_linedraw() +
  theme(
    # Tamanho do texto
    text = element_text(size = plot_text_size),
    # Eixo y
    axis.title.y.left = element_text(
      # Nunca girar o rótulo do eixo y
      angle = 0,
      # Separar o rótulo do eixo um pouco
      margin = margin(r = 20),
      # Posicionar verticalmente no meio
      vjust = .5
    ),
    # Eixo y secundário (à direita), quando presente
    axis.title.y.right = element_text(
      # Nunca girar o rótulo do eixo y
      angle = 0,
      # Separar o rótulo do eixo um pouco
```

```

margin = margin(l = 20),
# Posicionar verticalmente no meio
vjust = .5
),
# Não colocar marcas no eixo y secundário
axis.ticks.y.right = element_blank(),
# Separar o eixo x do rótulo um pouco mais
axis.title.x.bottom = element_text(
  margin = margin(t = 20)
),
# Eliminar linhas dos eixos
axis.line = element_blank(),
# Eliminar a moldura da área de dados
panel.border = element_blank(),
# Eliminar a grade secundária
panel.grid.minor = element_blank()
)
)

```

4.10

Exercícios



Não se esqueça de incluir títulos nos gráficos e rótulos nos eixos.

4.10.1

Peso cerebral e peso corporal

1. Observe os **comandos que geraram o gráfico grafico4**.
2. O que acontece se você retirar `aes(group = 1)` da chamada a `geom_smooth`? Explique.
3. O que acontece se você mudar `show.legend = FALSE` para `show.legend = TRUE` na chamada a `geom_smooth`? Explique.
4. O que acontece se você mudar `se = FALSE` para `se = TRUE` na chamada a `geom_smooth`? Explique.
5. Acrescente ao gráfico a camada `facet_wrap(~vore)`. O que acontece?
6. Examine o *data frame* `sono` e identifique o único insetívoro com mais de 4kg.
7. Instale o pacote `gg_repel` e acrescente ao gráfico `grafico4` (não facetado) a geometria `geom_label_repel` (consulte a ajuda) para rotular o mamífero insetívoro identificado no item anterior com o seu nome, **sem cobrir outros pontos do gráfico**. Cuidado para não alterar a legenda que já existe.

4.10.2

Peso cerebral e horas de sono

Use o *data frame* `sono` definido como

```
library(ggplot2)

sono <- msleep %>%
  select(
    name, order, genus, vore, bodywt,
    brainwt, awake, sleep_total
  )
```

1. Construa um histograma da variável `brainwt`. Escolha o número de classes que você achar melhor. O que acontece com os valores `NA`?
2. Descubra que função da forma `scale_x_...` usar para fazer com que o eixo x tenha uma escala logarítmica. Gere um novo histograma.
3. Qual dos dois histogramas é melhor para responder a pergunta “Qual a faixa de peso cerebral que tem mais animais?” de forma satisfatória?
4. Construa um *scatter plot* de horas de sono versus peso do cérebro. Você percebe alguma correlação entre estas variáveis? Se precisar, concentre-se em um subconjunto dos dados.
5. Usando `geom_smooth` ([leia a respeito](#)), sobreponha uma reta de regressão ao gráfico de dispersão, usando o método `lm` e sem o erro padrão (i.e., com `se = FALSE`). O que você observa? Discuta.

4.10.3

Igualdade de gênero entre furacões?

[Este artigo](#) tenta achar uma relação entre o gênero do nome de um furacão e a quantidade de vítimas fatais provocadas por ele.

Os dados estão no pacote DAAG, que deve ser instalado:

```
if (!require(DAAG))
  install.packages("DAAG")
```

Vamos usar apenas algumas das variáveis, com nomes em português.

```
df <- hurricNamed %>%
  as_tibble() %>%
  transmute(
    id = paste(Year, Name, sep = '-'),
    nome = Name,
    ano = Year,
    velocidade = LF.WindsMPH * 1.8,      # convertido para km/h
    pressao = LF.PressureMB,             # mbar
    prejuizo = BaseDam2014 %>% round(),   # milhões de dólares de 2014
    mortes = deaths,
    genero = mf
  )
```

1. Crie histogramas para as seguintes variáveis, escolhendo a quantidade de barras que você achar melhor.

- velocidade
- prejuizo
- mortes

Não se esqueça de incluir títulos nos gráficos e rótulos nos eixos.

Comente os histogramas.

2. Os histogramas de prejuízos e mortes não ficaram bons. Vamos gerar histogramas transformados.

No *data frame*, crie duas novas colunas:

- logprejuizo: *logaritmo* do prejuízo (na base 10)
- logmortes: *logaritmo* do número de mortes (na base 10)

Agora, gere histogramas destas duas novas variáveis.

3. O que significa o valor do logaritmo do prejuízo na base 10?
4. O que significa o valor do logaritmo do número de mortes na base 10?
5. Por que o histograma do logaritmo do número de mortes vem com uma mensagem de aviso?
6. Por que isto não acontece com o logaritmo do prejuízo?
7. Faça um gráfico de dispersão com *pressao* no eixo *y* e *velocidade* no eixo *x*.

8. Usando `geom_smooth` ([leia a respeito](#)), sobreponha uma reta de regressão ao gráfico, usando o método `lm` e sem o erro padrão (i.e., com `se = FALSE`). O que você observa? Discuta.
9. Faça um gráfico de dispersão com `logmortes` no eixo y e `pressao` no eixo x .
10. Usando `geom_smooth` ([leia a respeito](#)), sobreponha uma reta de regressão ao gráfico, usando o método `lm` e sem o erro padrão (i.e., com `se = FALSE`). O que você observa? Discuta.
11. Faça um gráfico de dispersão com `logmortes` no eixo y e `pressao` no eixo x , com pontos coloridos de acordo com o gênero do nome do furacão.
12. Usando `geom_smooth` ([leia a respeito](#)), sobreponha retas de regressão ao gráfico, uma para cada gênero, usando o método `lm` e sem o erro padrão (i.e., com `se = FALSE`). O que você observa? Discuta.



Visualizações como esta ajudam a explorar os dados, mas não servem para testar rigorosamente a hipótese de que furacões mulheres matam mais do que furacões homens.

Mais adiante no curso, vamos aprender a fazer testes mais rigorosos sobre hipóteses como esta.

Visualização com ggplot2 (continuação)



Busque mais informações sobre os pacotes `tidyverse` e `ggplot2` nas referências recomendadas.

5.1

Vídeo 1

<https://youtu.be/TjgLDeIQHIc>

5.2

Boxplots

5.2.1

Conjunto de dados

- Vamos continuar a trabalhar com os dados sobre as horas de sono de alguns mamíferos:

```
sono <- msleep %>%  
  select(name, vore, order, sleep_total)
```

```
sono
## # A tibble: 83 x 4
##   name                vore order      sleep_total
##   <chr>              <chr> <chr>      <dbl>
## 1 Cheetah            carni Carnivora      12.1
## 2 Owl monkey         omni  Primates        17
## 3 Mountain beaver    herbi Rodentia      14.4
## 4 Greater short-tailed shrew omni  Soricomorpha    14.9
## 5 Cow                herbi Artiodactyla    4
## 6 Three-toed sloth    herbi Pilosa      14.4
## # ... with 77 more rows
```

5.2.2

Mediana e quartis

- Para entender *boxplots*, precisamos, antes, entender algumas medidas.
- Se tomarmos as quantidades de horas de sono de todos os animais do conjunto de dados e **classificarmos estas quantidades em ordem crescente**, vamos ter:

```
horas <- sono %>%
  pull(sleep_total) %>%
  sort()

horas
## [1] 1,9 2,7 2,9 3,0 3,1 3,3 3,5 3,8 3,9 4,0 4,4 5,2 5,3
## [14] 5,3 5,4 5,6 6,2 6,3 6,3 7,0 7,7 8,0 8,3 8,4 8,4 8,6
## [27] 8,7 8,7 8,9 9,1 9,1 9,4 9,4 9,5 9,6 9,7 9,8 9,8 10,0
## [40] 10,1 10,1 10,1 10,3 10,3 10,4 10,6 10,9 11,0 11,0 11,1 11,3 11,5
## [53] 12,1 12,5 12,5 12,5 12,5 12,8 12,8 13,0 13,5 13,7 13,8 14,2 14,3
## [66] 14,4 14,4 14,5 14,6 14,9 14,9 15,6 15,8 15,8 15,9 16,6 17,0 17,4
## [79] 18,0 18,1 19,4 19,7 19,9
```

- Quantos valores são?

```
length(horas)
## [1] 83
```

- O valor que está **bem no meio desta fila** — i.e., na posição 42 — é a **mediana**:

```
horas[ceiling(length(horas) / 2)]
## [1] 10,1
```

- Em R:

```
median(horas)
## [1] 10,1
```

Mediana e média são coisas muito diferentes.

Por acaso, neste exemplo, a média das horas é próxima da mediana:



```
mean(horas)
## [1] 10,43373
```

Isto costuma acontecer quando a distribuição dos dados é aproximadamente simétrica.

- Os **quartis** são os valores que estão nas posições $\frac{1}{4}$, $\frac{1}{2}$ e $\frac{3}{4}$ da fila. São o **primeiro, segundo e terceiro quartis**, respectivamente.

```
horas[
  c(
    ceiling(length(horas) / 4),
    ceiling(length(horas) / 2),
    ceiling(3 * length(horas) / 4)
  )
]
## [1] 7,7 10,1 13,8
```

- Sim, a mediana é o segundo quartil.
- Em R, a **função quantile** generaliza esta idéia: dado um número q entre 0 e 1, o **quantil (com “N”) q é o elemento que está na posição que corresponde à fração q da fila ordenada.**

```
horas %>% quantile(c(.25, .5, .75))
## 25% 50% 75%
## 7,85 10,10 13,75
```

- Na verdade, R tem 9 algoritmos diferentes para calcular os quantis de uma amostra! Leia a ajuda da função `quantile` para conhecê-los.
- As diferenças entre nossos cálculos “à mão” e os resultados retornados por `quantile` são porque, em algumas situações, `quantile` calcula uma média ponderada entre elementos vizinhos. Por isso, `quantile` pode retornar valores que nem estão no vetor.
- Em R, a **função summary** mostra o **mínimo**, os **quartis (com “R”)**, a **média**, e o **máximo** de um vetor:

```
summary(horas)
## Min. 1st Qu. Median Mean 3rd Qu. Max.
```

```
##      1,90      7,85     10,10     10,43     13,75     19,90
```

5.2.3

Média × mediana

- Vamos ver um exemplo simples para entender a diferença entre a média e a mediana.
- Imagine o seguinte vetor com as receitas mensais de algumas pessoas (em milhares de reais:)

```
receitas <- c(1, 2, 2, 3.5, 1, 4, 1)
```

- Eis a mediana e a média deste vetor:

```
summary(receitas)[c('Median', 'Mean')]
##      Median      Mean
## 2,000000 2,071429
```

- A mediana e a média são bem próximas.
- Imagine, agora, que adicionamos ao vetor um sujeito com receita mensal de 100 mil reais:

```
receitas <- c(1, 2, 2, 3.5, 1, 4, 1, 100)
```

- Eis a nova mediana e a nova média:

```
summary(receitas)[c('Median', 'Mean')]
##      Median      Mean
##   2,0000 14,3125
```

- O sujeito com a receita de 2 mil reais continua no meio da fila, mas a média (que é a soma de todas as receitas, dividida pelo número de indivíduos) ficou muito diferente.
- A receita do novo sujeito é um **valor discrepante**, ou, em inglês, um **outlier**.

Conclusão:



A **mediana é robusta**, pouco afetada por *outliers*.

A **média é pouco robusta**, muito sensível a *outliers*.

5.2.4

Intervalo interquartil (IQR) e *outliers*

- Qual fração dos elementos está entre o primeiro e o terceiro quartis?

```
length(
  horas[between(horas, quantile(horas, .25), quantile(horas, .75))]
) /
length(
  horas
)
## [1] 0,4939759
```

- Metade do total de elementos está entre o primeiro e o terceiro quartis.
- Este é o chamado intervalo interquartil (*interquartile range*, em inglês).
- No nosso vetor `horas`, os limites do IQR são

```
quantile(horas, c(.25, .75))
##      25%      75%
##  7,85 13,75
```

- O comprimento deste intervalo é calculado pela função `IQR`:

```
IQR(horas)
## [1] 5,9
```

- Valores muito abaixo do primeiro quartil podem ser considerados discrepantes (*outliers*), mas quão abaixo?
- A resposta (puramente convencional) é $1,5 \times \text{IQR}$ abaixo do primeiro quartil.
- No nosso vetor `horas`, isto significa valores abaixo de

```
limite_inferior <- quantile(horas, .25) - 1.5 * IQR(horas)

unnamed(limite_inferior)
## [1] -1
```

- Neste caso, não há *outliers*:

```
horas[horas < limite_inferior]
## numeric(0)
```

- Da mesma forma, valores muito acima do terceiro quartil podem ser considerados discrepantes (*outliers*), mas quão acima?
- De novo, a resposta (puramente convencional) é $1,5 \times \text{IQR}$ acima do terceiro quartil.

- No nosso vetor `horas`, isto significa valores acima de

```
limite_superior <- quantile(horas, .75) + 1.5 * IQR(horas)

unnname(limite_superior)
## [1] 22,6
```

- Neste caso, também não há *outliers*:

```
horas[horas > limite_superior]
## numeric(0)
```

- Outro exemplo: vamos tomar apenas os mamíferos onívoros:

```
onivoros <- sono %>%
  filter(vore == 'omni')

onivoros
## # A tibble: 20 x 4
##   name                vore order      sleep_total
##   <chr>              <chr> <chr>          <dbl>
## 1 Owl monkey        omni  Primates        17
## 2 Greater short-tailed shrew omni  Soricomorpha    14.9
## 3 Grivet            omni  Primates        10
## 4 Star-nosed mole    omni  Soricomorpha    10.3
## 5 African giant pouched rat omni  Rodentia         8.3
## 6 Lesser short-tailed shrew omni  Soricomorpha     9.1
## # ... with 14 more rows
```

- Vamos extrair o vetor de horas de sono:

```
horas <- onivoros %>%
  pull(sleep_total)

horas
## [1] 17,0 14,9 10,0 10,3 8,3 9,1 18,0 10,1 10,9 9,8 8,0 10,1 9,7
## [14] 9,4 11,0 8,7 9,6 9,1 15,6 8,9
```

- Vamos calcular o primeiro e terceiro quartis:

```
quartis <- horas %>%
  quantile(c(.25, .75))

quartis
##   25%   75%
## 9,100 10,925
```

- Vamos achar o IQR:

```
IQR(horas)
## [1] 1,825
```

- E os limites a partir dos quais os valores são *outliers*:

```
limites <- quartis + c(-1, 1) * 1.5 * IQR(horas)

unnname(limites)
## [1] 6,3625 13,6625
```

- Existem *outliers* inferiores?

```
onivoros %>%
  filter(sleep_total < limites[1])
## # A tibble: 0 x 4
## # ... with 4 variables: name <chr>, vore <chr>, order <chr>,
## #   sleep_total <dbl>
```

Não.

- Existem *outliers* superiores?

```
onivoros %>%
  filter(sleep_total > limites[2])
## # A tibble: 4 x 4
##   name                vore order      sleep_total
##   <chr>               <chr> <chr>         <dbl>
## 1 Owl monkey          omni  Primates        17
## 2 Greater short-tailed shrew omni  Soricomorpha   14.9
## 3 North American Opossum omni  Didelphimorphia 18
## 4 Tenrec              omni  Afrosoricida   15.6
```

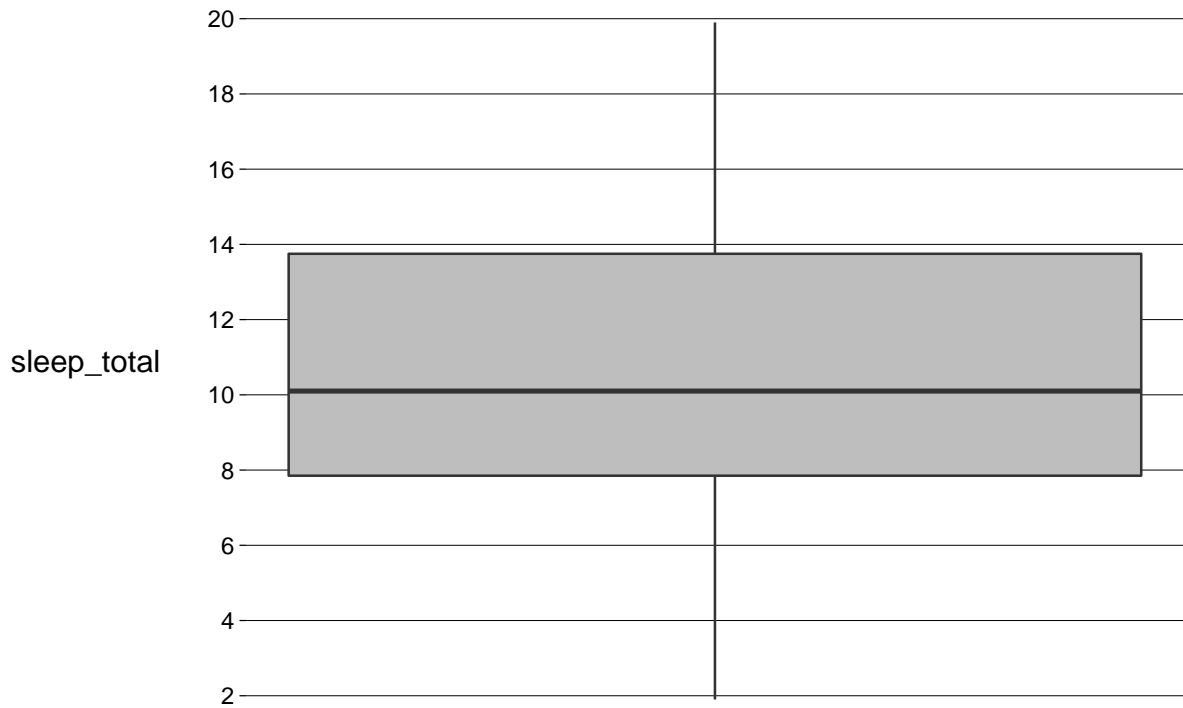
Sim! Estes animais dormem demais em comparação com os outros onívoros.

5.2.5

Gerando boxplots

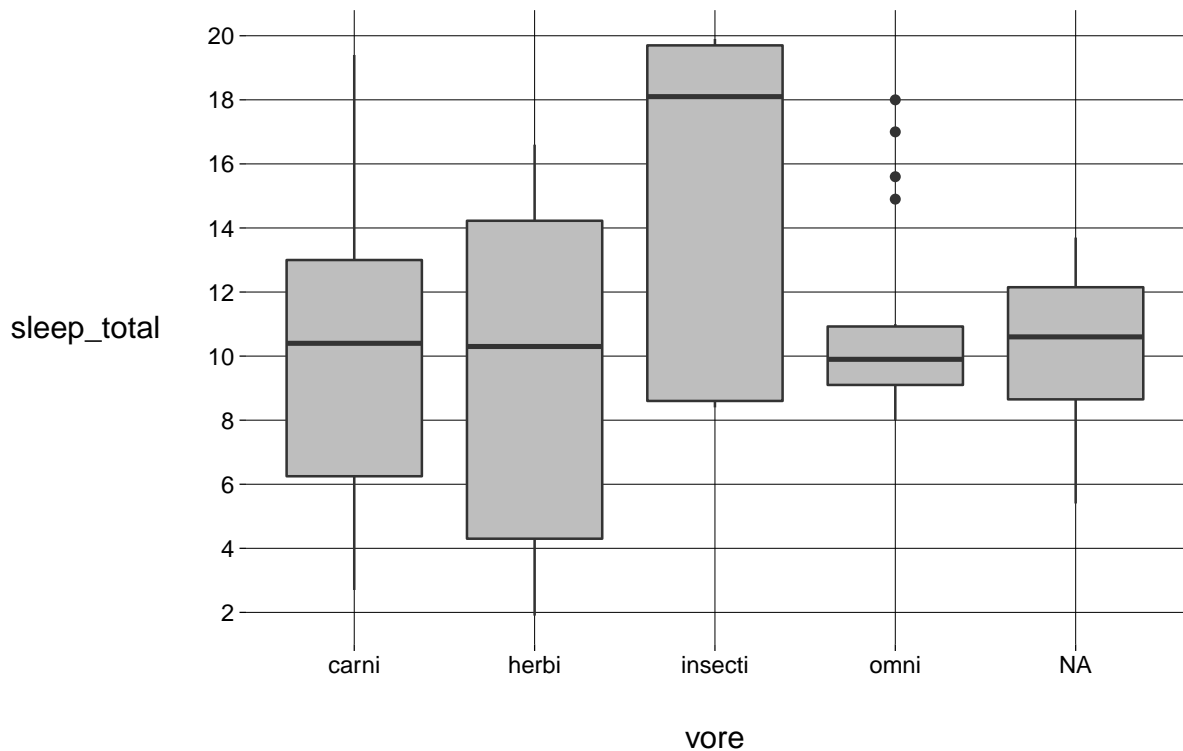
- Um *boxplot* é uma representação visual dos valores que calculamos acima.
- No ggplot2, a geometria `geom_boxplot` constrói *boxplots*:

```
sono %>%
  ggplot(aes(y = sleep_total)) +
  geom_boxplot(fill = 'gray') +
  scale_x_continuous(breaks = NULL) +
  scale_y_continuous(breaks = seq(0, 20, 2))
```

- A **caixa** vai do valor do **primeiro quartil** (embaixo) até o **terceiro quartil** (em cima).
- A **linha horizontal dentro da caixa** representa o valor da **mediana**.
- As **linhas verticais** acima e abaixo da caixa (pitorescamente chamadas de “bigodes”) vão até o **limite inferior** (primeiro quartil $- 1,5 \times \text{IQR}$) e até o **limite superior** (terceiro quartil $+ 1,5 \times \text{IQR}$).
- Neste *boxplot*, não há *outliers*.
- Podemos usar a posição *x* para desenhar vários *boxplots*, um para cada dieta:

```
sono %>%  
  ggplot(aes(x = vore, y = sleep_total)) +  
    geom_boxplot(fill = 'gray') +  
    scale_y_continuous(breaks = seq(0, 20, 2))
```



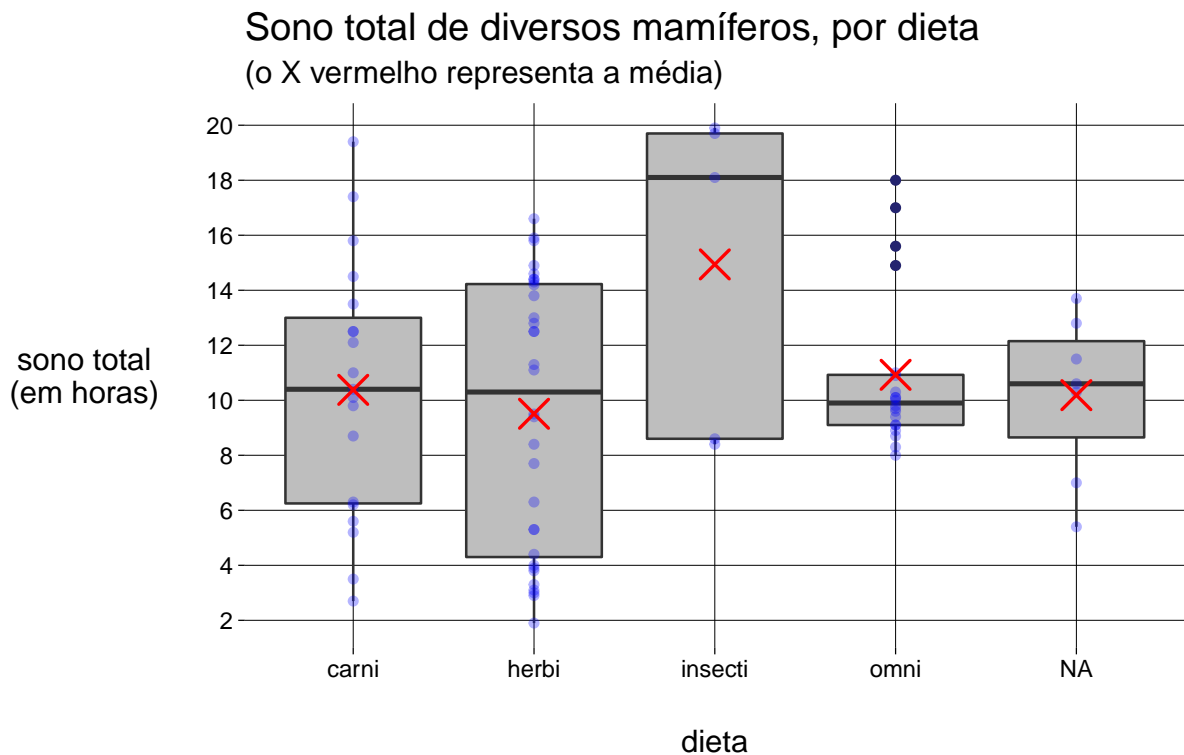
- No *boxplot* de onívoros, **os outliers aparecem como pontos isolados**, acima da caixa, além dos alcances do bigode superior (aliás, onde está bigode superior?).
- *Boxplots* lado a lado são úteis para compararmos grupos diferentes de dados.
- Veja como, com exceção dos insetívoros, as medianas dos grupos são parecidas.
- Veja como carnívoros, insetívoros e herbívoros apresentam maior variação, enquanto onívoros e animais sem dieta registrada apresentam menor variação.
- Vamos combinar, em um só gráfico
 - Os pontos representando os animais,
 - Os *boxplots*,
 - As médias (que podem estar próximas ou distantes das medianas).

```
sono %>%
  ggplot(aes(x = vore, y = sleep_total)) +
    geom_boxplot(fill = 'gray') +
    scale_y_continuous(breaks = seq(0, 20, 2)) +
    geom_point(
      color = 'blue',
      alpha = .3
    ) +
    stat_summary(
      fun = mean,
      geom = 'point',
```

```

color = 'red',
shape = 'cross',
size = 5,
stroke = 1
) +
labs(
  title = 'Sono total de diversos mamíferos, por dieta',
  subtitle = '(o X vermelho representa a média)',
  x = 'dieta',
  y = 'sono total\n(em horas)'
)

```



- Quando a caixa é longa, o IQR é grande, e os valores estão muito espalhados; é o caso dos herbívoros e insetívoros.
- Quando a caixa é curta, o IQR é pequeno, e os valores estão pouco espalhados; é o caso dos onívoros. Como o IQR é pequeno, os 4 mamíferos com mais de 14 horas de sono são *outliers*.
- Observe, ainda, como os *outliers* “puxam” a média dos onívoros para cima.

5.3

Vídeo 2

<https://youtu.be/QqnOvgBXJ-s>

5.4

Gráficos de barras e de colunas

5.4.1

Conjunto de dados

- O R tem um *array* de 3 dimensões com dados sobre as cores dos cabelos e dos olhos de 592 alunos e alunas de uma universidade americana em 1974.
- Se pedirmos para o R exibir os dados, veremos **duas matrizes**, uma para cada sexo:

```
HairEyeColor
## , , Sex = Male
##
##      Eye
## Hair   Brown Blue Hazel Green
## Black   32   11   10    3
## Brown   53   50   25   15
## Red     10   10    7    7
## Blond    3   30    5    8
##
## , , Sex = Female
##
##      Eye
## Hair   Brown Blue Hazel Green
## Black   36    9    5    2
## Brown   66   34   29   14
## Red     16    7    7    7
## Blond    4   64    5    8
```

- Vamos transformar este *array* em um *data frame*.
- O *array* contém apenas os totais de cada classe. Vamos usar a função `uncount` para gerar uma linha para cada aluno:

```
df_orig <- as.data.frame(HairEyeColor) %>%
  uncount(Freq) %>%
  as_tibble()

df_orig
```

```
## # A tibble: 592 x 3
##   Hair Eye Sex
##   <fct> <fct> <fct>
## 1 Black Brown Male
## 2 Black Brown Male
## 3 Black Brown Male
## 4 Black Brown Male
## 5 Black Brown Male
## 6 Black Brown Male
## # ... with 586 more rows
```

- O `ggplot2` e os outros pacotes do `tidyverse` foram projetados para trabalhar com *data frames* neste formato, com uma observação (um indivíduo, um elemento) por linha. É o chamado *formato tidy*.
- Usando vetores com elementos nomeados, podemos traduzir o conteúdo do *data frame* para português:

```
cabelo <- c(
  'Brown' = 'castanhos',
  'Blond' = 'louros',
  'Black' = 'pretos',
  'Red' = 'ruivos'
)

olhos <- c(
  'Brown' = 'castanhos',
  'Blue' = 'azuis',
  'Hazel' = 'avelã',
  'Green' = 'verdes'
)

sexo <- c(
  'Male' = 'homem',
  'Female' = 'mulher'
)

df <- df_orig %>%
  transmute(
    cabelos = cabelo[Hair],
    olhos = olhos[Eye],
    sexo = sexo[Sex]
  )
```

- Um sumário:

```
df %>% dfSummary() %>% print()
```

Variável	Estatísticas / Valores	Freqs (% de Válidos)	Faltante
cabelos [character]	1. castanhos	108 (18,2%)	0
	2. louros	286 (48,3%)	(0,0%)
	3. pretos	71 (12,0%)	
	4. ruivos	127 (21,5%)	
olhos [character]	1. avelã	93 (15,7%)	0
	2. azuis	215 (36,3%)	(0,0%)
	3. castanhos	220 (37,2%)	
	4. verdes	64 (10,8%)	
sexo [character]	1. homem	279 (47,1%)	0
	2. mulher	313 (52,9%)	(0,0%)

5.4.2

Gerando gráficos de barras

- Um **gráfico de barras** contém uma barra para cada valor de uma **variável categórica**.
- Usamos `geom_bar` para gerar um gráfico de barras de cores de cabelo:

```
df %>%
  ggplot(aes(x = cabelos)) +
  geom_bar() +
  labs(y = NULL)
```

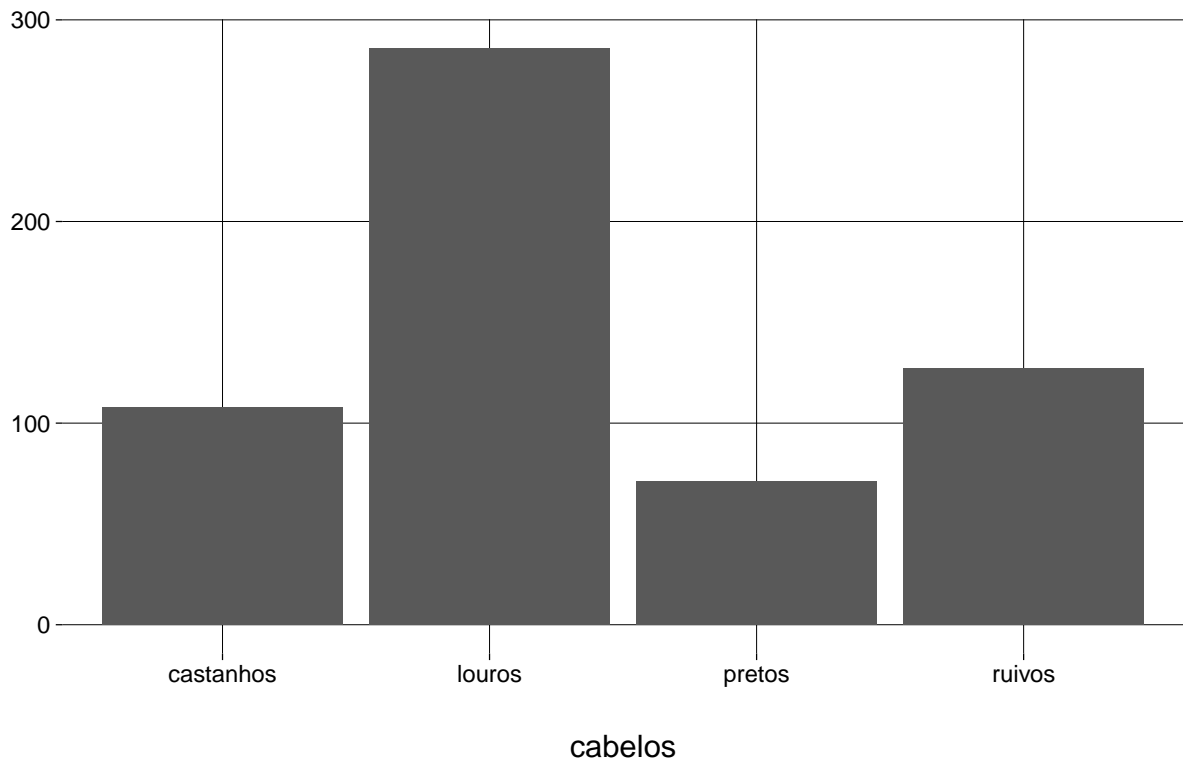


Gráfico de barras × histograma:

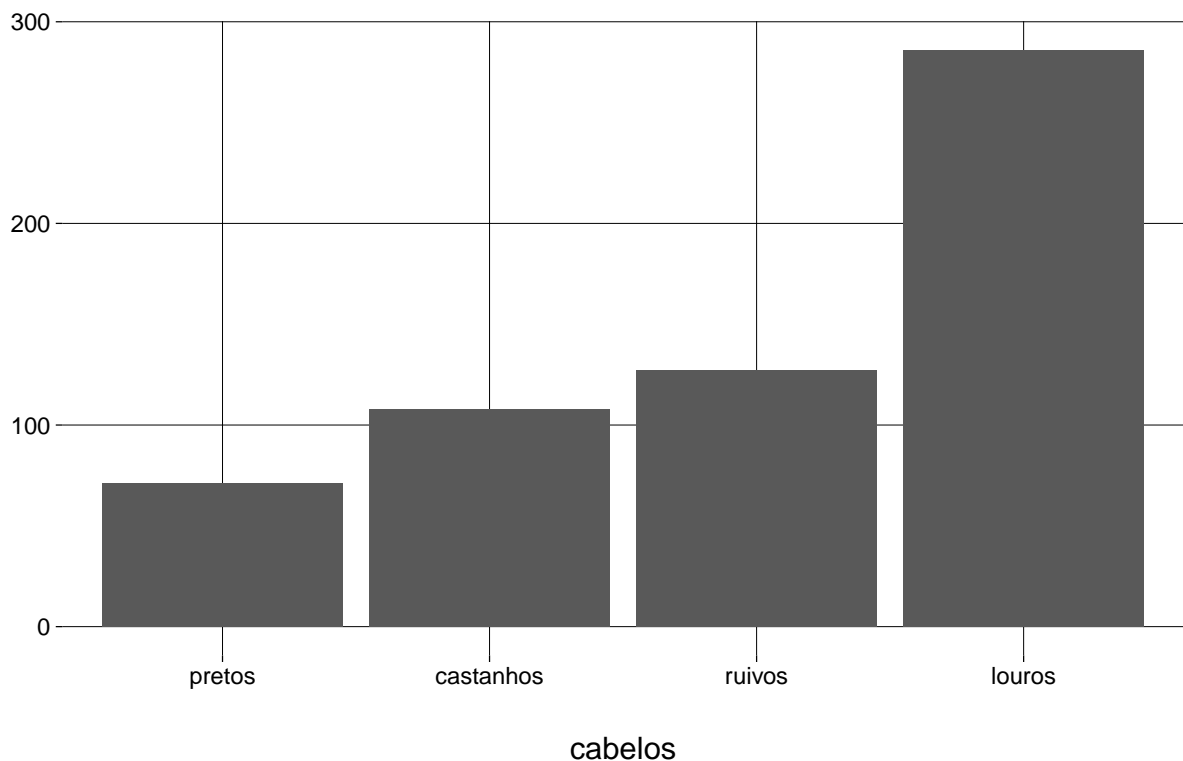
- Os dois tipos de gráficos mostram a frequência (quantidade de elementos) no eixo vertical.
- No gráfico de barras:
 - * A variável é categórica (nominal).
 - * Cada barra corresponde a um valor da variável.
 - * As barras não se tocam, enfatizando o fato de que a variável é categórica.
- No histograma (veja o exemplo):
 - * A variável é quantitativa (intervalar ou racional).
 - * Cada barra corresponde a uma classe de valores da variável.
 - * As barras se tocam, para enfatizar que as classes são contíguas.

- Um gráfico de barras é mais legível quando as barras são mostradas em ordem crescente ou decrescente.
- Embora os valores da variável `cabelos` sejam *strings*, podemos aplicar a eles funções que manipulam fatores.
- A função `fct_infreq`, do pacote `forcats`, ordena os valores em ordem decrescente

de frequência.

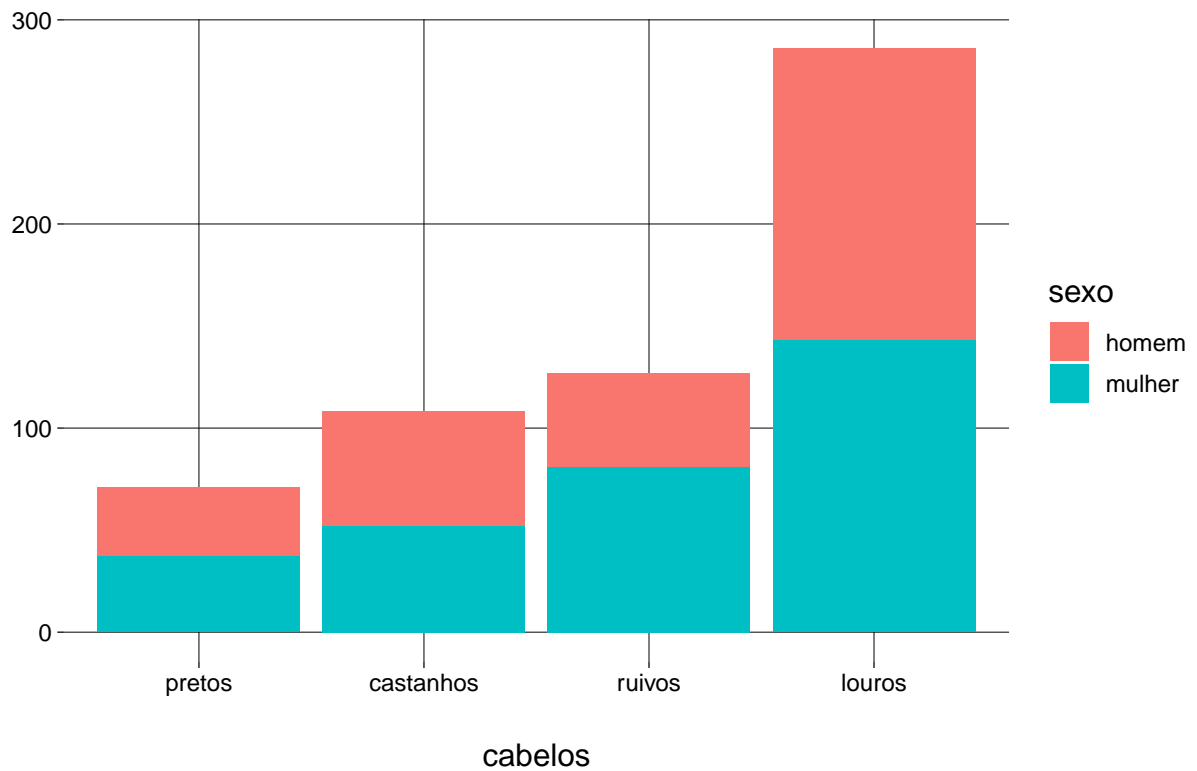
- A função `fct_rev`, também do pacote `forcats`, **inverte a ordenação**.

```
df %>%  
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)))) +  
    geom_bar() +  
    labs(  
      x = 'cabelos',  
      y = NULL  
    )
```



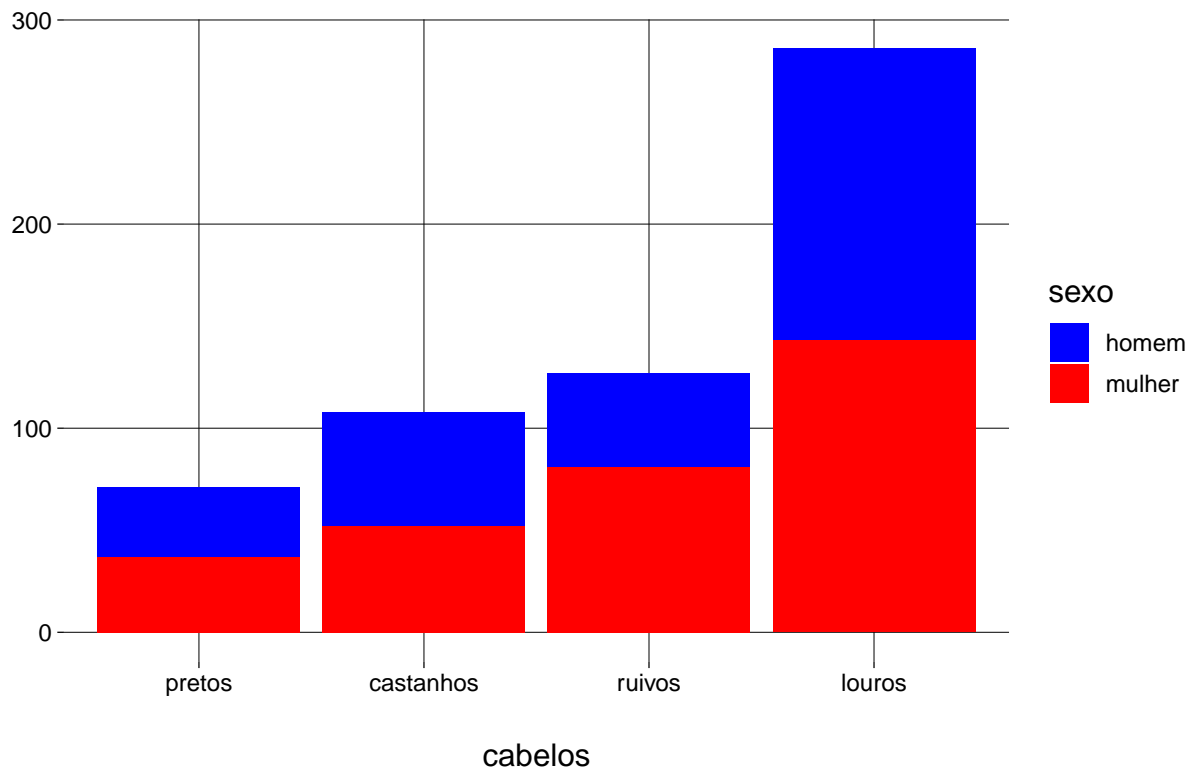
- A posição x e a altura de cada barra são estéticas: **a posição x representa a cor dos cabelos**, e **a altura representa a frequência daquela cor**.
- Vamos acrescentar mais uma estética: **a cor de preenchimento vai representar o sexo**.

```
df %>%  
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = sexo)) +  
    geom_bar() +  
    labs(  
      x = 'cabelos',  
      y = NULL  
    )
```

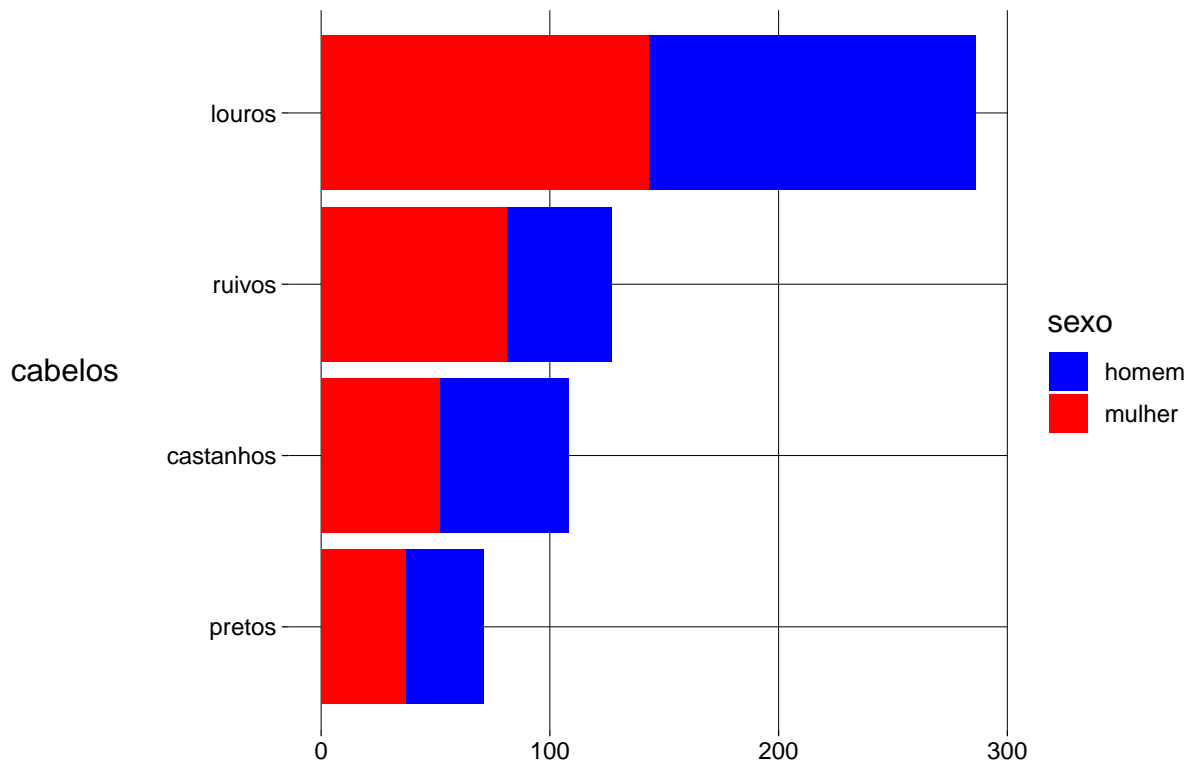
- Se a cor dos homens incomoda você, altere a escala que especifica o preenchimento (scale_fill_discrete):

```
df %>%  
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = sexo)) +  
    geom_bar() +  
    scale_fill_discrete(type = c('blue', 'red')) +  
    labs(  
      x = 'cabelos',  
      y = NULL  
    )
```



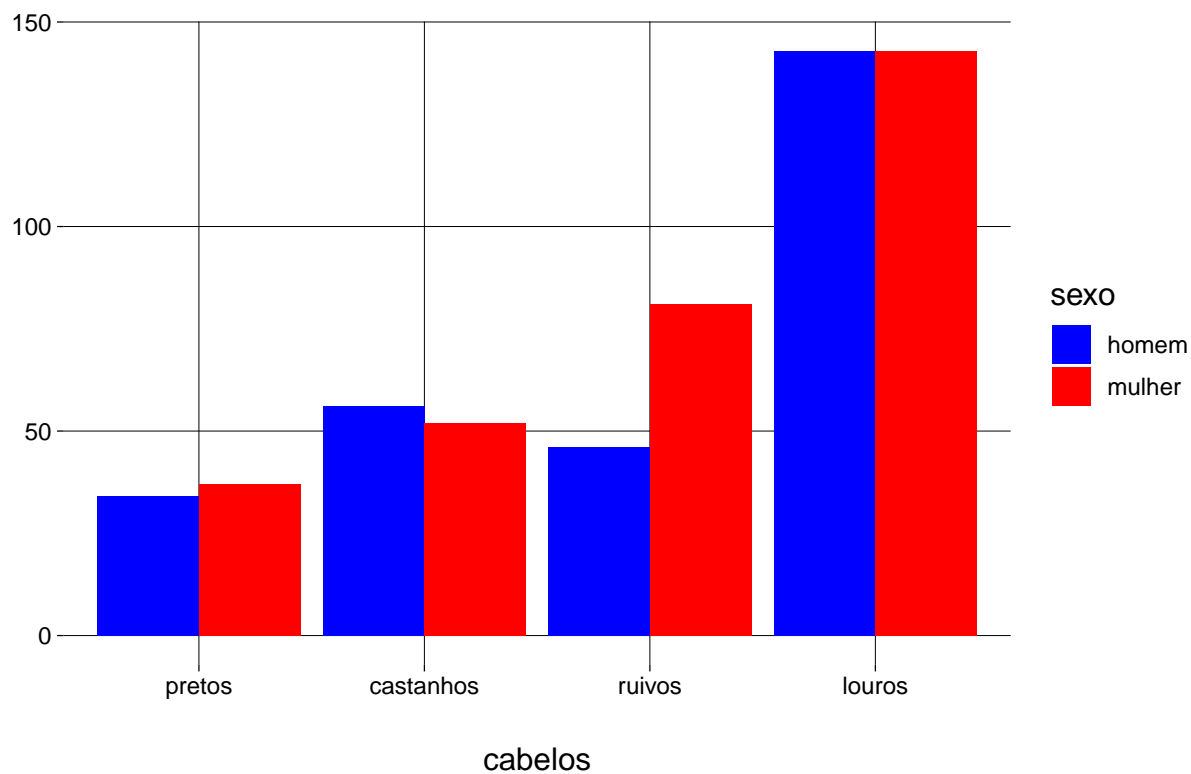
- Podemos fazer um gráfico de barras horizontais com `coord_flip`. Isto geralmente é útil quando os rótulos das barras são longos:

```
df %>%  
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = sexo)) +  
    geom_bar() +  
    scale_fill_discrete(type = c('blue', 'red')) +  
    labs(  
      x = 'cabelos',  
      y = NULL  
    ) +  
    coord_flip()
```



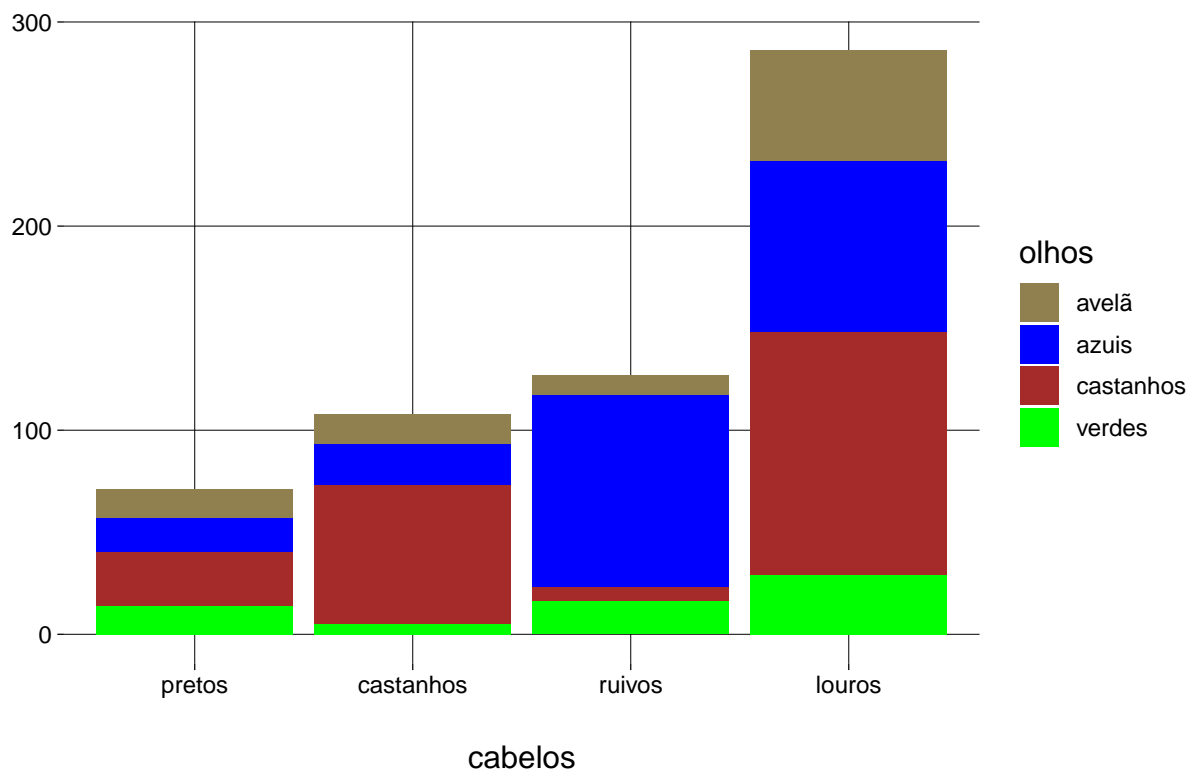
- Você consegue dizer se há mais homens ou mulheres com cabelos pretos? E castanhos? E ruivos?
- Se posicionarmos as barras lado a lado, fica mais fácil responder.
- Usamos o argumento `position = 'dodge'` de `geom_bar`. “*Dodge*” significa “esquivar-se”, em inglês.

```
df %>%
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = sexo)) +
  geom_bar(position = 'dodge') +
  labs(
    x = 'cabelos',
    y = NULL
  ) +
  scale_fill_discrete(type = c('blue', 'red'))
```



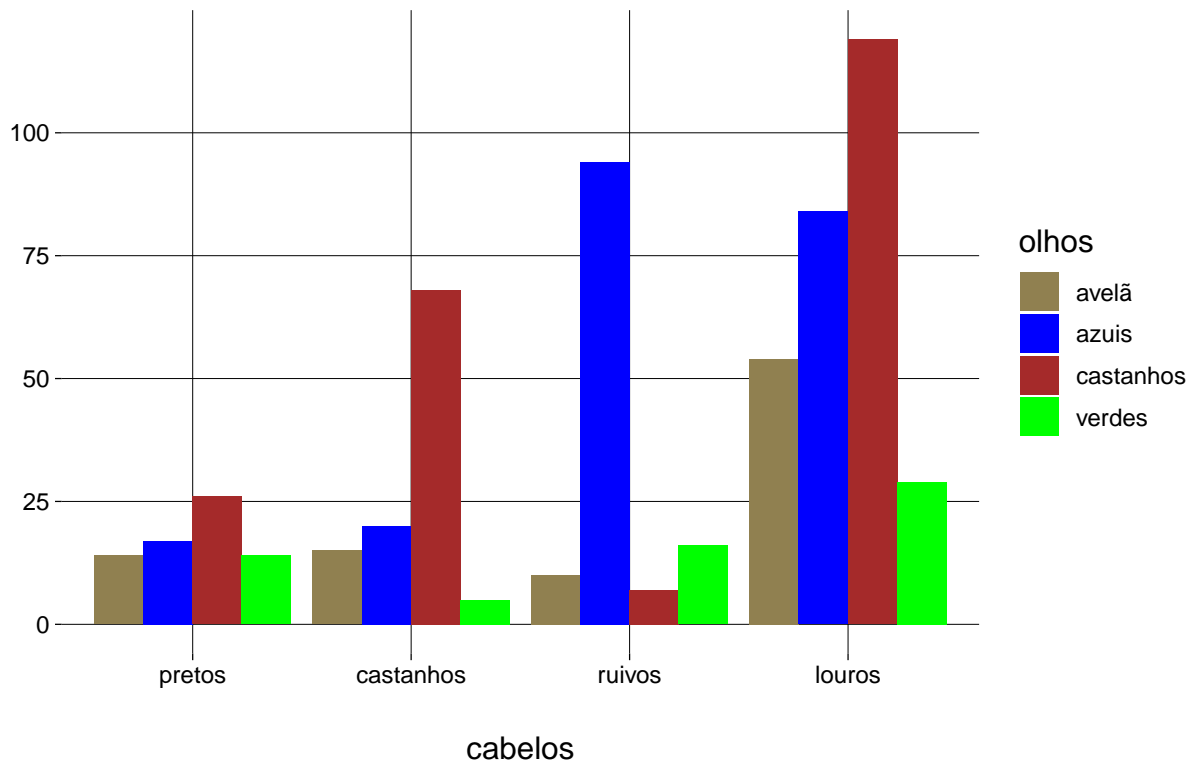
- Agora vamos examinar a relação entre as cores dos olhos e as cores dos cabelos:

```
df %>%
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = olhos)) +
  geom_bar() +
  scale_fill_discrete(
    type = c('#908050', 'blue', 'brown', 'green')
  ) +
  labs(
    x = 'cabelos',
    y = NULL
  )
```



- Ou, com barras lado a lado:

```
df %>%
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = olhos)) +
  geom_bar(position = 'dodge') +
  scale_fill_discrete(
    type = c('#908050', 'blue', 'brown', 'green')
  ) +
  labs(
    x = 'cabelos',
    y = NULL
  )
```



- Observações e perguntas:

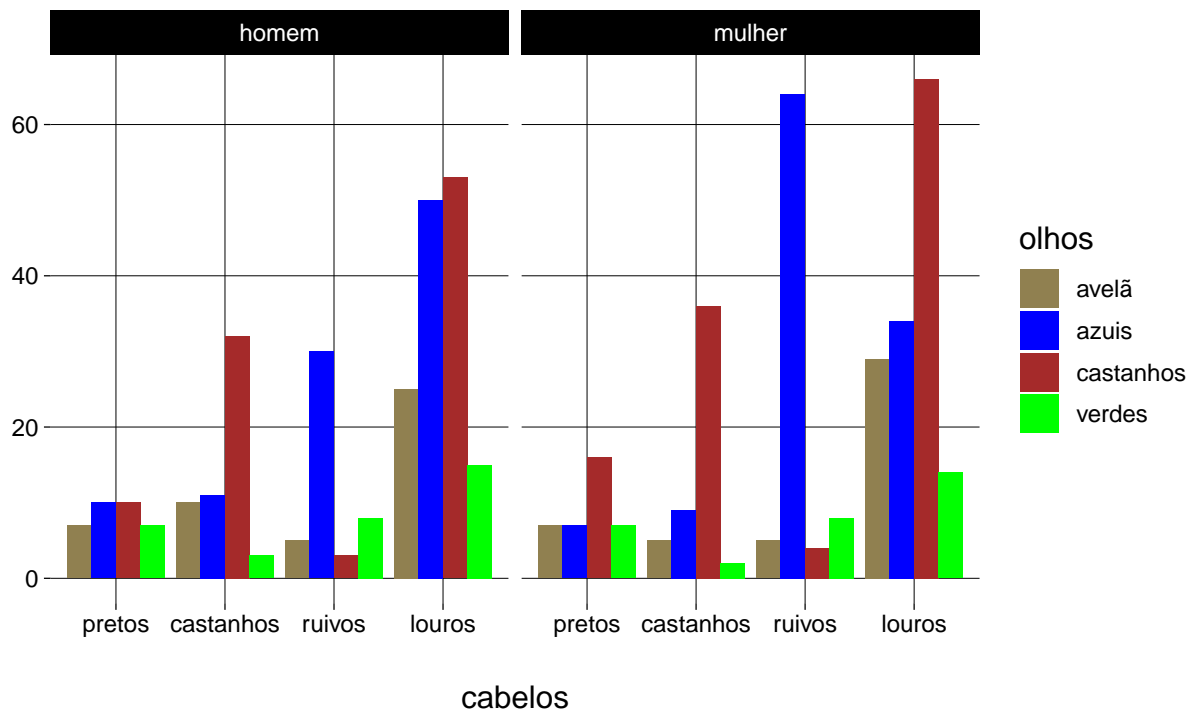
1. Há mais pessoas loiras de olhos castanhos do que loiras de olhos azuis. O esperado não seria mais pessoas loiras de olhos azuis? Pessoas loiras de olhos castanhos pintaram os cabelos?
2. Há muito mais ruivos de olhos azuis do que ruivos de olhos verdes. Não deveria ser o contrário? Também são pessoas que pintaram os cabelos de ruivo? Ou houve erro no registro das cores dos olhos?

- Para incluir o sexo, podemos **facetar** o gráfico. Usando `facet_wrap`¹, geramos dois subgráficos lado a lado:

```
df %>%
  ggplot(aes(x = fct_rev(fct_infreq(cabelos)), fill = olhos)) +
  geom_bar(position = 'dodge') +
  scale_fill_discrete(type = c('#908050', 'blue', 'brown', 'green')) +
  facet_wrap(~sexo) +
  labs(
    title = 'Cores de cabelos e olhos por sexo',
    y = NULL,
    x = 'cabelos'
  )
```

¹O nome da variável segundo a qual facetar deve aparecer depois de um ~.

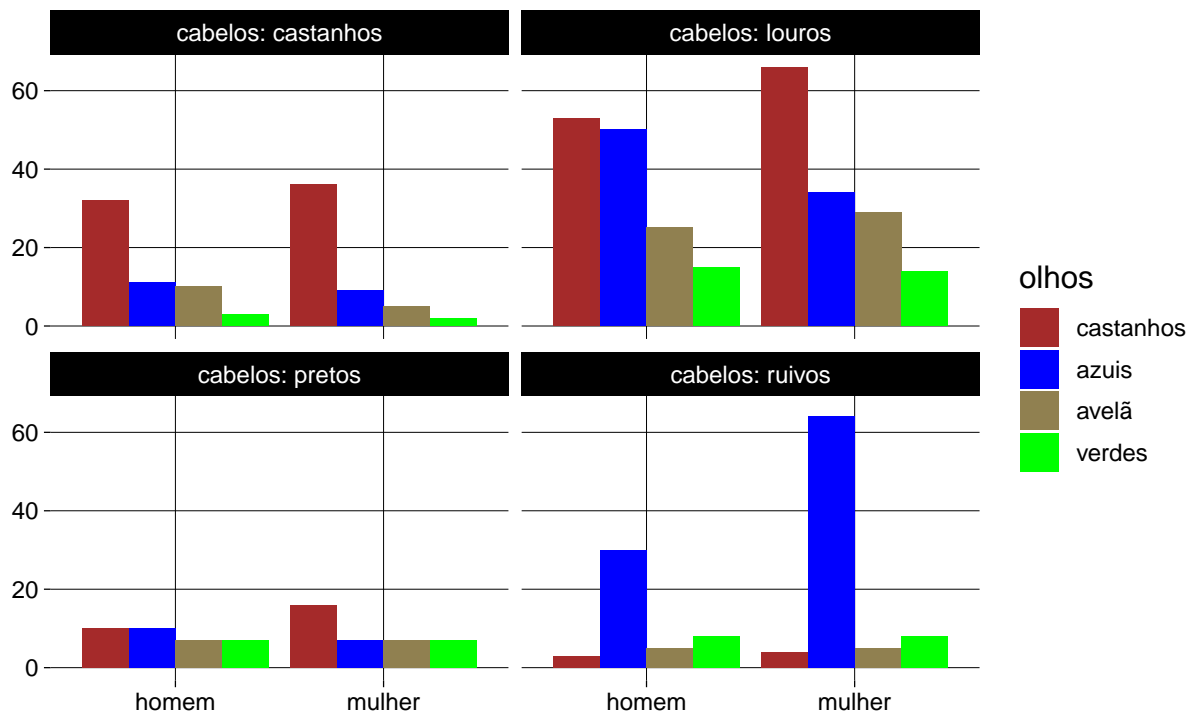
Cores de cabelos e olhos por sexo



- Se a quantidade grande de pessoas louras de olhos castanhos (em comparação com pessoas louras de olhos azuis) for por causa da pintura de cabelos, então o gráfico acima mostra que as mulheres pintam os cabelos de loiro com mais frequência do que os homens.
- Quando facetamos por cor de cabelos, também podemos observar as mesmas diferenças entre homens e mulheres:

```
df %>%
  ggplot(aes(x = sexo, fill = fct_infreq(olhos))) +
  geom_bar(position = 'dodge') +
  facet_wrap(~cabelos, labeller = label_both) +
  scale_fill_discrete(type = c('brown', 'blue', '#908050', 'green')) +
  labs(
    x = NULL,
    y = NULL,
    fill = 'olhos',
    title = 'Cor dos olhos e sexo por cor dos cabelos'
  )
```

Cor dos olhos e sexo por cor dos cabelos



5.4.3

Data frame já contendo os totais

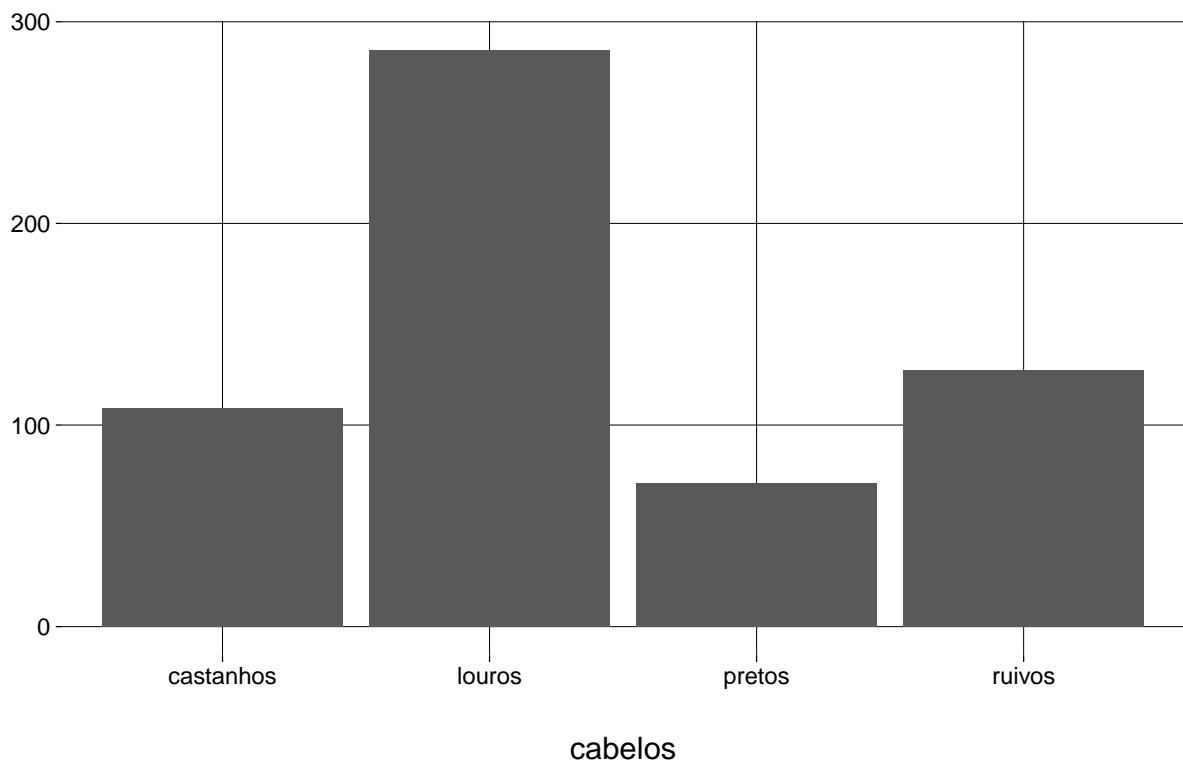
- Você percebeu que `geom_bar` analisa o *data frame* e calcula as frequências necessárias para construir o gráfico.
- Em algumas situações, o *data frame* já contém as frequências (em vez de conter uma linha por indivíduo).
- Vamos usar `count` para criar um *data frame* assim:

```
df_tot <- df %>%
  count(sexo, cabelos, olhos)

df_tot
## # A tibble: 32 x 4
##   sexo  cabelos  olhos      n
##   <chr> <chr>    <chr>   <int>
## 1 homem castanhos avelã     10
## 2 homem castanhos azuis     11
## 3 homem castanhos castanhos  32
## 4 homem castanhos verdes      3
## 5 homem louros   avelã     25
## 6 homem louros   azuis     50
## # ... with 26 more rows
```

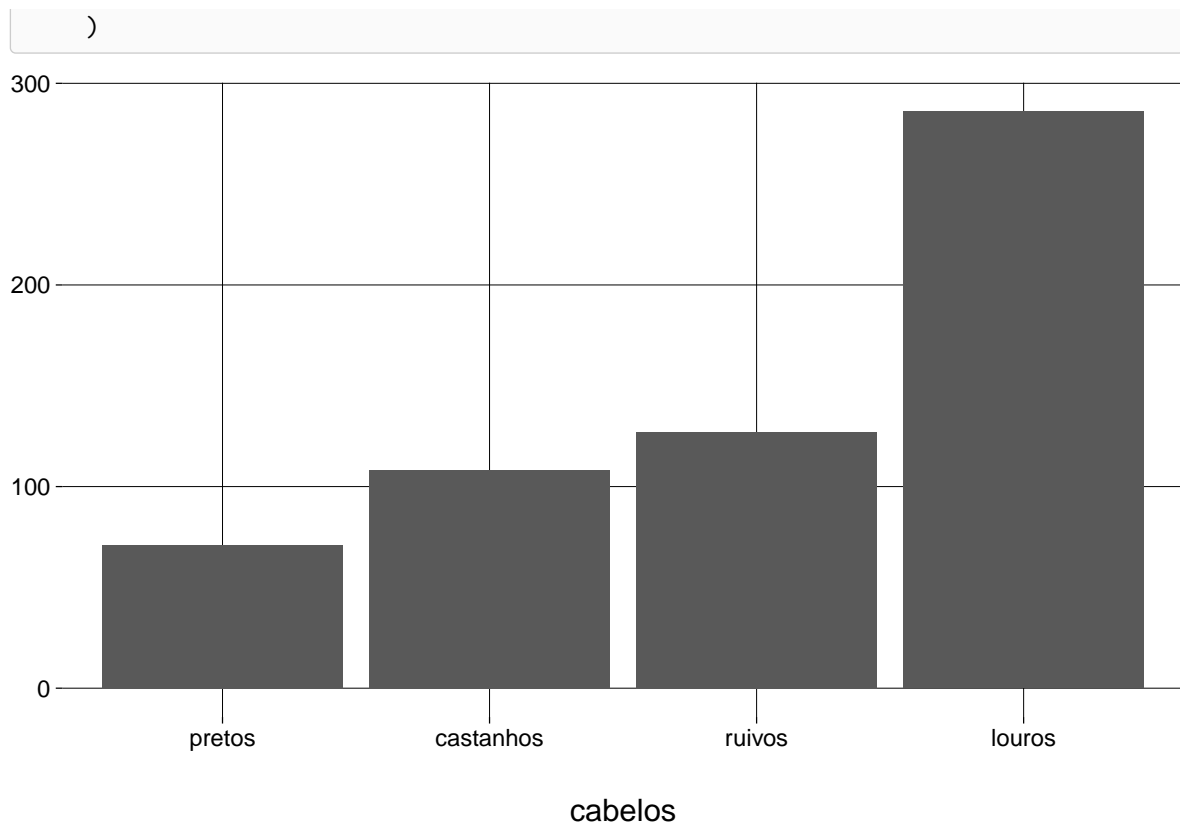

- Para 4 cores de cabelo, 4 cores de olhos, e 2 sexos, são 32 combinações possíveis.
- Com este *data frame*, podemos gerar todos os gráficos anteriores usando `geom_col` no lugar de `geom_bar`. Por exemplo:

```
df_tot %>%
  ggplot(aes(x = cabelos, y = n)) +
    geom_col() +
    labs(
      y = NULL
    )
```



- Com `geom_col`, precisamos passar a estética `y` (no nosso exemplo, a variável `n`, que contém as frequências).
- Para ordenar as barras, usamos a função `fct_reorder`, que ordena os níveis de um fator (`cabelos`) de acordo com o resultado de uma função (`sum`) aplicada sobre os valores de outra variável (`n`):

```
df_tot %>%
  ggplot(aes(x = fct_reorder(cabelos, n, sum), y = n)) +
    geom_col() +
    labs(
      x = 'cabelos',
      y = NULL
    )
```



5.5

Gráficos de linha e séries temporais

5.5.1

Conjunto de dados

- O R tem uma matriz com as quantidades de telefones em várias regiões do mundo ao longo de vários anos:

```
WorldPhones
##      N.Amer Europe Asia S.Amer Oceania Africa Mid.Amer
## 1951  45939  21574 2876   1815   1646    89    555
## 1956  60423  29990 4708   2568   2366   1411   733
## 1957  64721  32510 5230   2695   2526   1546   773
## 1958  68484  35218 6662   2845   2691   1663   836
## 1959  71799  37598 6856   3000   2868   1769   911
## 1960  76036  40341 8220   3145   3054   1905  1008
## 1961  79831  43173 9053   3338   3224   2005  1076
```

- Os números representam milhares.
- Os números dos anos são os nomes das linhas da matriz.

- Vamos transformar esta matriz em uma *tibble*:

```
fones <- WorldPhones %>%
  as_tibble(rownames = 'Ano') %>%
  mutate(Ano = as.numeric(Ano))

fones
## # A tibble: 7 x 8
##   Ano N.Amer Europe Asia S.Amer Oceania Africa Mid.Amer
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 1951 45939 21574 2876 1815 1646 89 555
## 2 1956 60423 29990 4708 2568 2366 1411 733
## 3 1957 64721 32510 5230 2695 2526 1546 773
## 4 1958 68484 35218 6662 2845 2691 1663 836
## 5 1959 71799 37598 6856 3000 2868 1769 911
## 6 1960 76036 40341 8220 3145 3054 1905 1008
## # ... with 1 more row
```

- Esta *tibble* não está no formato *tidy*. Queremos que cada linha corresponda a uma observação, contendo
 - Ano,
 - Região,
 - Quantidade de telefones.
- Usamos a função `pivot_longer` para mudar o formato da *tibble*:

```
fones_long <- fones %>%
  pivot_longer(
    cols = -Ano,
    names_to = 'Região',
    values_to = 'n'
  )

fones_long
## # A tibble: 49 x 3
##   Ano Região      n
##   <dbl> <chr> <dbl>
## 1 1951 N.Amer 45939
## 2 1951 Europe 21574
## 3 1951 Asia 2876
## 4 1951 S.Amer 1815
## 5 1951 Oceania 1646
## 6 1951 Africa 89
## # ... with 43 more rows
```

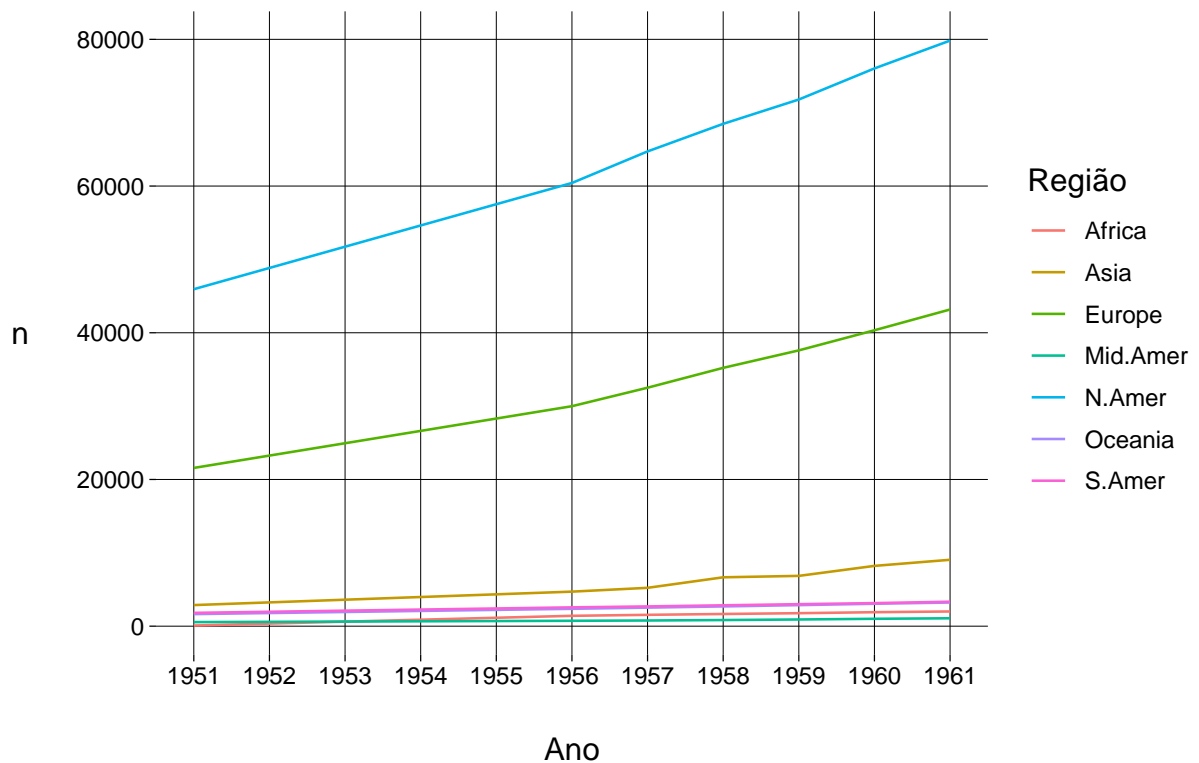
- Confira: antes, tínhamos 7 anos, com 7 quantidades por ano, uma quantidade por região. Eram 49 quantidades. Agora temos uma *tibble* de 49 linhas.

5.5.2

Gerando gráficos de linha

- A geometria `geom_line` gera gráficos de linha. Perceba como geramos uma linha por região:

```
fones_long %>%  
  ggplot(aes(x = Ano, y = n, color = Região)) +  
    geom_line() +  
    scale_x_continuous(breaks = 1951:1961)
```

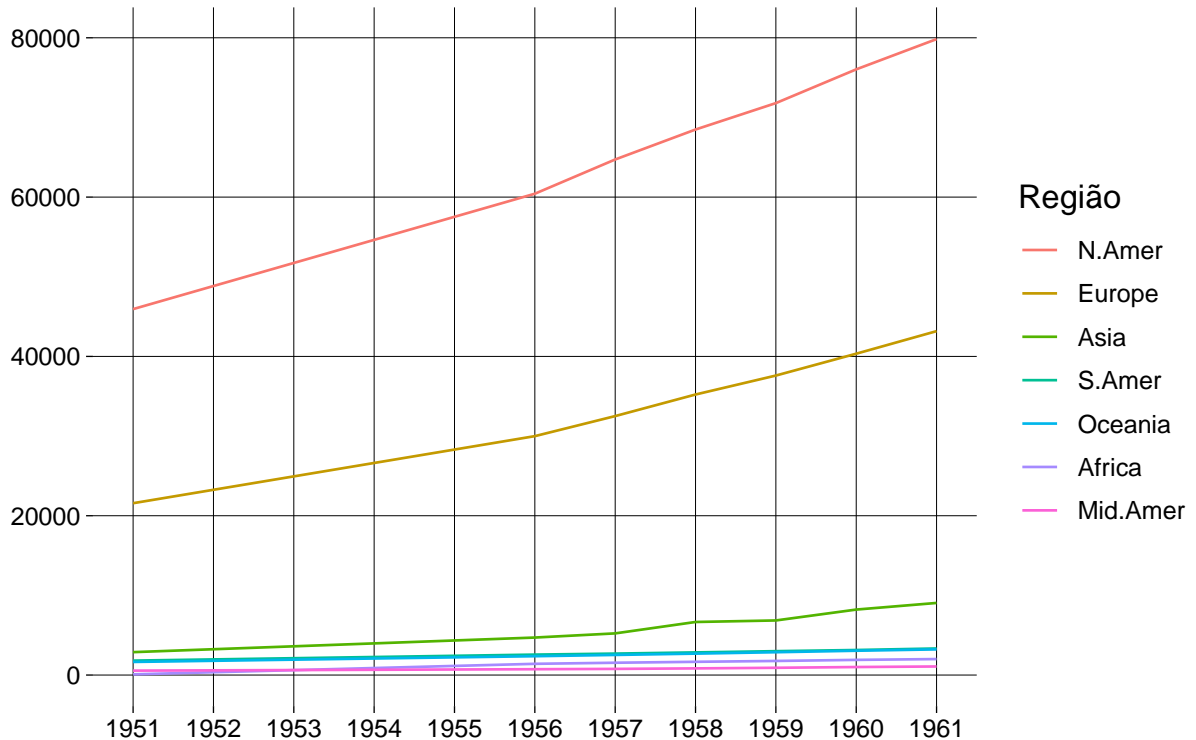


- Embora a legenda associe uma cor a cada região, a leitura seria mais fácil se a ordem das regiões na legenda coincidissem com a posição das linhas na borda direita da grade:

```
fones_long %>%  
  ggplot(  
    aes(  
      x = Ano,  
      y = n,  
      color = fct_rev(fct_reorder(Região, n, max))  
    )  
  ) +
```

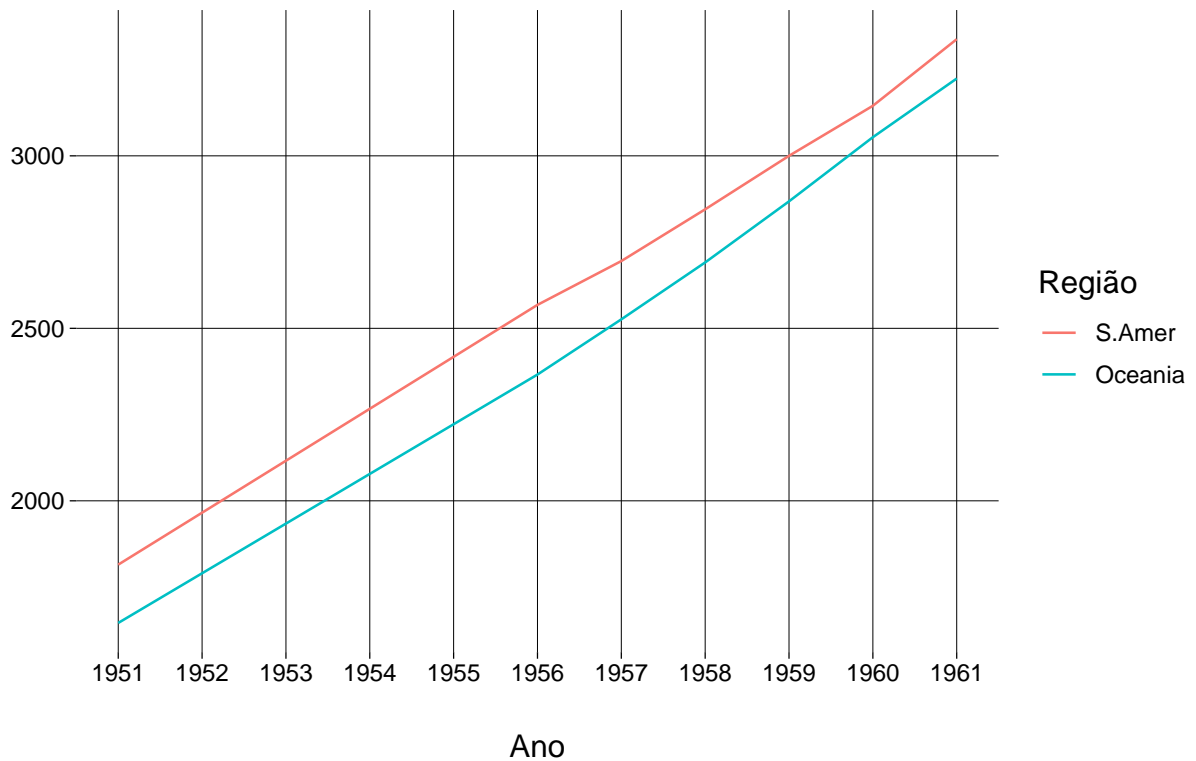
```
geom_line() +
scale_x_continuous(breaks = 1951:1961) +
labs(
  color = 'Região',
  y = '',
  x = NULL,
  title = 'Quantidade de aparelhos de telefone por ano, por região'
)
```

Quantidade de aparelhos de telefone por ano, por região



- Parece que está faltando uma linha, mas o que acontece é que as quantidades da América do Sul e da Oceania são bem parecidas:

```
fones_long %>%
  filter(Região %in% c('S.Amer', 'Oceania')) %>%
  ggplot(
    aes(
      x = Ano,
      y = n,
      color = fct_rev(fct_reorder(Região, n, max))
    )
  ) +
  geom_line() +
  scale_x_continuous(breaks = 1951:1961) +
  labs(y = NULL, color = 'Região')
```



- Estamos tratando estes dados como simples números, mas, na verdade, **este conjunto de dados é uma série temporal (*time series*)**.
- R tem todo um conjunto de funções para tratar séries temporais, calcular tendências, achar padrões cíclicos, fazer estimativas, e gerar gráficos específicos, entre outras coisas.
- Mas não vamos falar mais sobre séries temporais aqui.
- O **pacote tsibble** oferece maneiras de trabalhar com séries temporais de maneira *tidy*. Você pode ler a documentação do pacote entrando

```
library(tsibble)
?`tsibble-package`
```

5.6

Exercícios

5.6.1

O bigode dos onívoros

- Examine o *data frame* `sono` para descobrir o que houve com o bigode superior do *boxplot* dos onívoros **neste gráfico**.

5.6.2

Usando `geom_col`

- Use `geom_col` para reproduzir, a partir do *data frame* `df_tot`, todos os gráficos que foram gerados com `geom_bar` na seção [Gerando gráficos de barras](#).

5.6.3

5.7

Referências sobre visualização e R



Busque mais informações sobre os pacotes `tidyverse` e `ggplot2` [nas referências recomendadas](#).

CAPÍTULO 6

Medidas
