Lista 02 - MQ

Martha Gaudencio

2026-02-01

```
#MQ101 - Lista 02
#Nome: Martha Gaudencio da Silva
#Data: 22/10/2025
#Descrição: Tipos de variáveis e estatísticas descritivas

#1 - Preparação do ambiente

 library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.6
## ✓ forcats   1.0.1     ✓ stringr   1.6.0
## ✓ ggplot2   4.0.1     ✓ tibble    3.3.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.2
## ✓ purrr     1.2.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
 options(scipen = 999)

#2 - Base de dados

 dados <- read.csv(file.choose())
  glimpse(dados)
```

```
## Rows: 10,000
## Columns: 13
## $ id                <chr> "P00001", "P00002", "P00003", "P00004", "P00005", "P…
## $ sexo              <chr> "M", "F", "F", "F", "F", "F", "F", "F", "F", "F", "M…
## $ escolaridade      <chr> "Medio", "Medio", "Medio", "Medio", "Medio", "Medio"…
## $ anos_estudo       <int> 10, 9, 9, 10, 11, 11, 10, 8, 15, 11, 6, 7, 5, 4, 10,…
## $ rede_escolar      <chr> "privada", "pública", "pública", "pública", "pública…
## $ municipio         <chr> "Santos", "Dourados", "Várzea Grande", "Curitiba", "…
## $ UF                <chr> "SP", "MS", "MT", "PR", "RJ", "SP", "SC", "SP", "SP"…
## $ idade             <int> 39, 30, 31, 44, 26, 51, 47, 48, 25, 53, 23, 27, 35, …
## $ faltas_esc        <int> 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 2, 0, 0, 0…
## $ tempo_estudo_h    <dbl> 2.3, 6.6, 8.8, 4.2, 3.7, 4.9, 0.2, 1.9, 14.4, 3.3, 3…
## $ pressao_sistolica <int> 110, 109, 121, 131, 134, 124, 139, 124, 132, 134, 13…
## $ diagnostico       <chr> "sem", "sem", "sem", "sem", "sem", "HAS", "sem", "se…
## $ plano_saude       <chr> "privado", "privado", "nenhum", "nenhum", "SUS", "pr…
```

```
  summary(dados)
```

```
##       id                sexo           escolaridade        anos_estudo
##  Length:10000       Length:10000       Length:10000       Min.   : 4.00
##  Class :character   Class :character   Class :character   1st Qu.: 8.00
##  Mode  :character   Mode  :character   Mode  :character   Median :11.00
##                                                           Mean   :10.74
##                                                           3rd Qu.:13.00
##                                                           Max.   :20.00
##  rede_escolar        municipio             UF                idade
##  Length:10000       Length:10000       Length:10000       Min.   :18.00
##  Class :character   Class :character   Class :character   1st Qu.:31.00
##  Mode  :character   Mode  :character   Mode  :character   Median :40.00
##                                                           Mean   :40.39
##                                                           3rd Qu.:49.00
##                                                           Max.   :80.00
##    faltas_esc     tempo_estudo_h   pressao_sistolica diagnostico
##  Min.   :0.0000   Min.   : 0.000   Min.   : 85.0     Length:10000
##  1st Qu.:0.0000   1st Qu.: 2.800   1st Qu.:112.0     Class :character
##  Median :0.0000   Median : 5.000   Median :122.0     Mode  :character
##  Mean   :0.2346   Mean   : 6.122   Mean   :122.2
##  3rd Qu.:0.0000   3rd Qu.: 8.200   3rd Qu.:131.0
##  Max.   :3.0000   Max.   :34.300   Max.   :175.0
##  plano_saude
##  Length:10000
##  Class :character
##  Mode  :character
##
##
##
```

```
#Número de colunas: 13; número de linhas: 10.000.
#Classes das variáveis: qualitativas e numéricas
#Não existem NA

# 3 - Classificação de variáveis

# 3.1 - Tipo teórico de cada variável

#Sexo: qualitativa nominal binária
#Escolaridade: qualitativa ordinal
#Anos de estudo: quantitativa discreta
#Rede escolar: qualitativa nominal binária

#3.2 Coerência entre tipo teórico e classe:

dados <- dados |>
  mutate(
    sexo = factor(sexo),
    rede_escolar = factor(rede_escolar),
    plano_saude = factor(plano_saude),
    diagnostico = factor(diagnostico),
    escolaridade = factor(escolaridade,
                          levels = c("Fundamental","Medio","Superior"),
                          ordered = TRUE)
  )
str(dados)
```

```
## 'data.frame':    10000 obs. of  13 variables:
##  $ id               : chr  "P00001" "P00002" "P00003" "P00004" ...
##  $ sexo             : Factor w/ 2 levels "F","M": 2 1 1 1 1 1 1 1 1 1 ...
##  $ escolaridade     : Ord.factor w/ 3 levels "Fundamental"<..: 2 2 2 2 2 2 1 1 3 2 ...
##  $ anos_estudo      : int  10 9 9 10 11 11 10 8 15 11 ...
##  $ rede_escolar     : Factor w/ 2 levels "privada","pública": 1 2 2 2 2 2 2 2 2 2 ...
##  $ municipio        : chr  "Santos" "Dourados" "Várzea Grande" "Curitiba" ...
##  $ UF               : chr  "SP" "MS" "MT" "PR" ...
##  $ idade            : int  39 30 31 44 26 51 47 48 25 53 ...
##  $ faltas_esc       : int  0 0 1 0 1 0 1 0 1 0 ...
##  $ tempo_estudo_h   : num  2.3 6.6 8.8 4.2 3.7 4.9 0.2 1.9 14.4 3.3 ...
##  $ pressao_sistolica: int  110 109 121 131 134 124 139 124 132 134 ...
##  $ diagnostico      : Factor w/ 4 levels "DM","HAS","outros",..: 4 4 4 4 4 2 4 4 4 4 ...
##  $ plano_saude      : Factor w/ 4 levels "ambos","nenhum",..: 3 3 2 2 4 3 4 4 3 2 ...
```

```
#3.3 Registre, em comentários, por que cada variável é daquele tipo

#Sexo é qualitativa nominal porque se trata de uma informação não-numérica e que
#apenas dá nome à categoria, no caso, feminino ou masculino, sendo aqui também
#binária. O mesmo vale para rede escolar. Já escolaridade é também qualitativa,
#mas ordinal, porque existe uma ordem que se segue em continuidade. Por fim,
#anos de estudo é numérica categórica pois a base considera apenas números
#inteiros.

#4 - Variáveis Qualitativas

#4.1 - Frequências absolutas e relativas

tab_plano <- table(dados$plano_saude)
prop_plano <- prop.table(tab_plano)
cbind(FA = tab_plano, FR = round(100 * prop_plano, 1))
```

```
##           FA   FR
## ambos    634  6.3
## nenhum  1868 18.7
## privado 2474 24.7
## SUS     5024 50.2
```

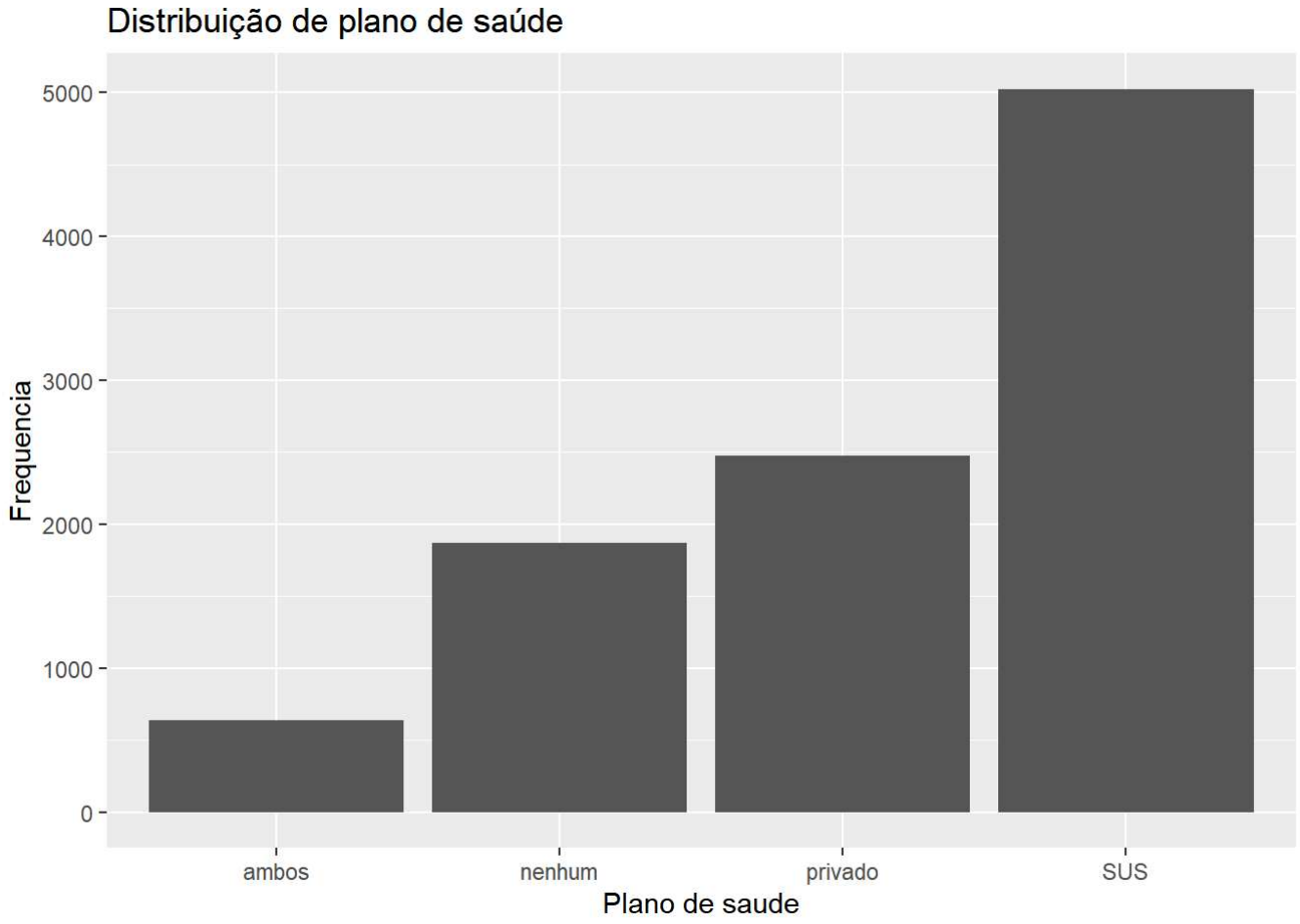
```
#SUS é a moda, porque é a que mais se repete. E a proporção dominante é de
#50,2% do SUS.

#4.2 - Gráfico de barras

install.packages("ggplot2")
```

```
## Warning: o pacote 'ggplot2' está em uso e não será instalado
```

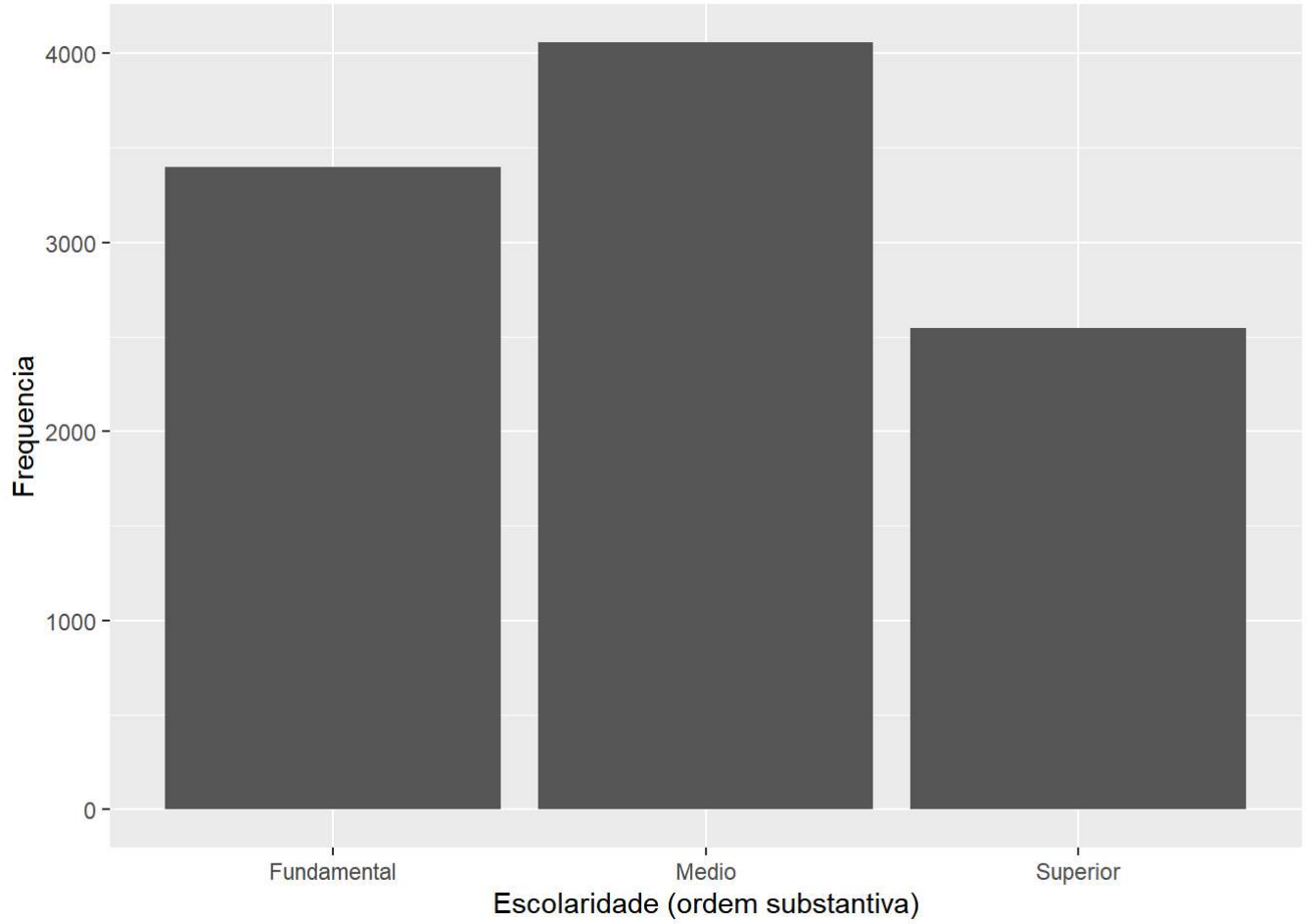
```
library(ggplot2)
install.packages("dplyr")
```

```
## Warning: o pacote 'dplyr' está em uso e não será instalado
```

```
library(dplyr)

dados |>
  count(plano_saude) |>
  ggplot(aes(x = plano_saude, y = n)) +
  geom_col() +
  labs(x = "Plano de saude", y = "Frequencia",
       title = "Distribuição de plano de saúde")
```




```
#4.3 - Barras ordenadas

dados |>
  count(escolaridade) |>
  ggplot(aes(x = escolaridade, y = n)) +
  geom_col() +
  labs(x = "Escolaridade (ordem substantiva)", y = "Frequencia")
```



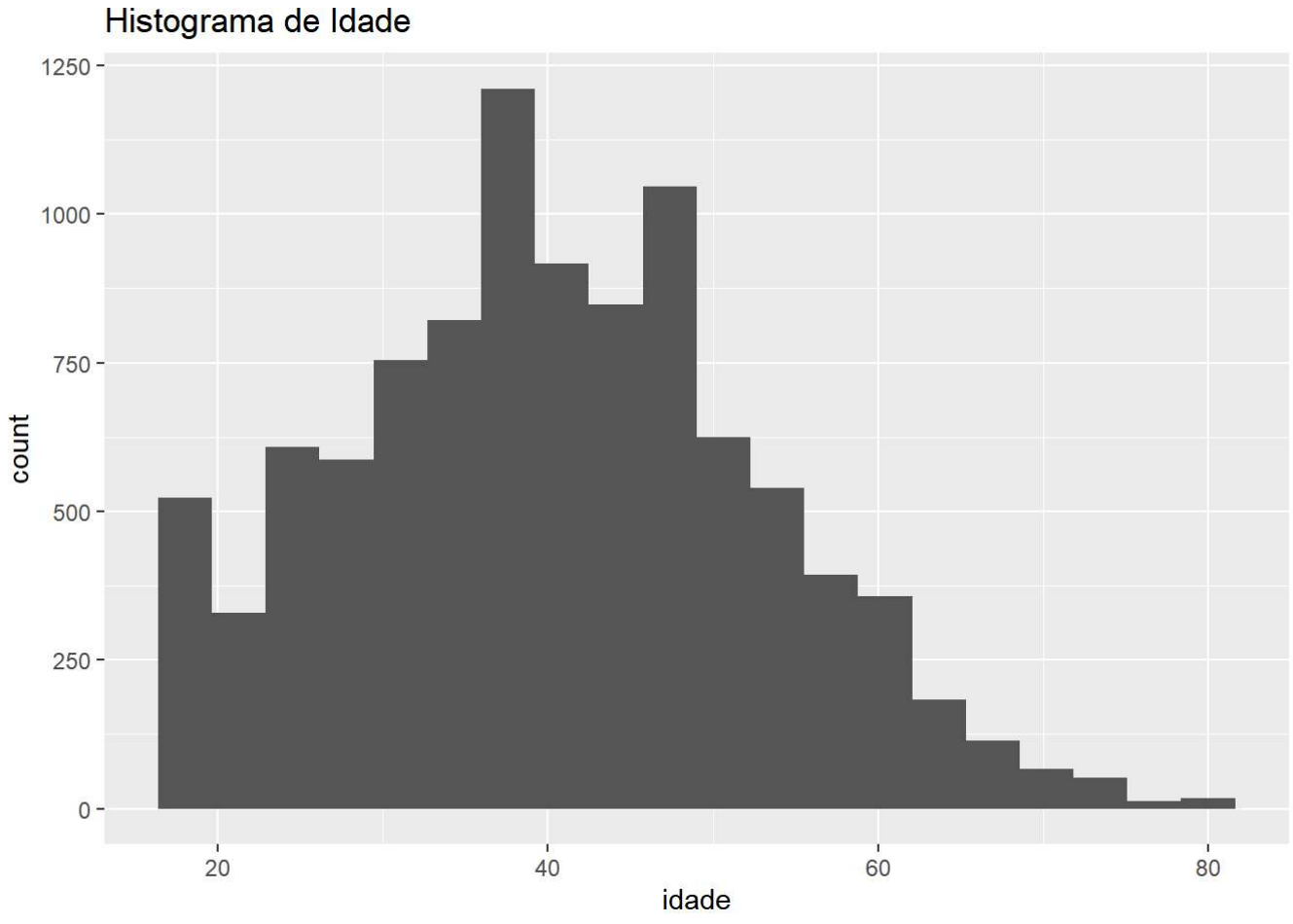

```
#Como trata-se de uma variável ordinal, manter a ordem das variáveis permite
#visualizar a informação com mais sentido, no caso percebendo que o ensino
#superior corresponde ao 3º nível e possui menos membros da amostra com este
#grau em comparação com os que possuem fundamental e médio. Já o gráfico de
#barras em variáveis não ordinais como a do plano de saúde permite ver qual
#possui mais aderência indo do menor para o maior sem que seja uma sequência
#ordinal das variáveis.

#5 - Variáveis quantitativas

#5.1 Tendência central e dispersão

sumario_idade <- dados |>
  summarise(
    n = sum(!is.na(idade)),
    media = mean(idade, na.rm = TRUE),
    mediana= median(idade, na.rm = TRUE),
    min = min(idade, na.rm = TRUE),
    max = max(idade, na.rm = TRUE),
    dp = sd(idade, na.rm = TRUE)

  )
sumario_idade
```

```
##       n   media mediana min max       dp
## 1 10000 40.3948      40  18  80 12.48622
```

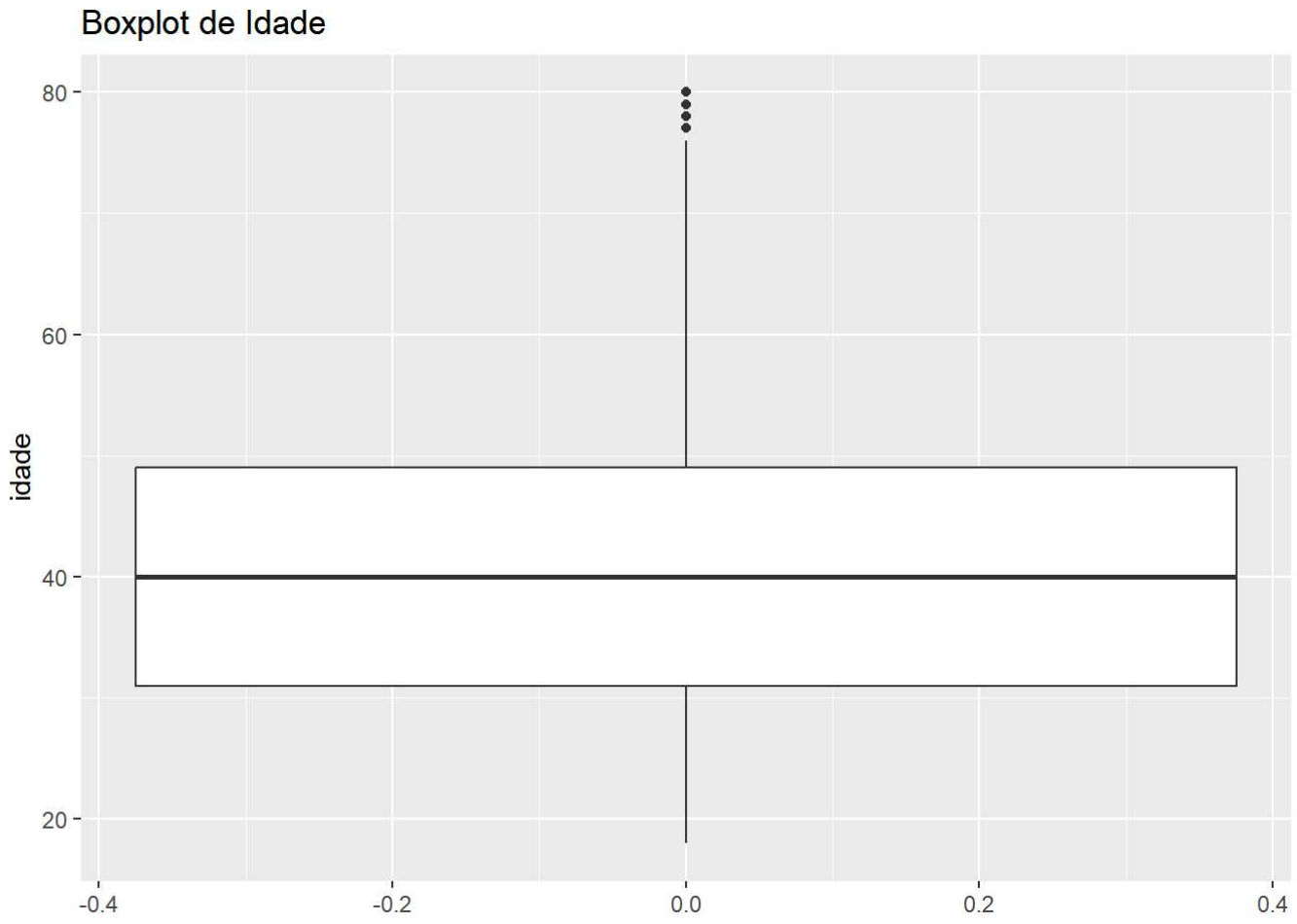
```
sumario_pressao_sistolica <- dados |>
  summarise(
    n = sum(!is.na(pressao_sistolica)),
    media = mean(pressao_sistolica, na.rm = TRUE),
    mediana= median(pressao_sistolica, na.rm = TRUE),
    min = min(pressao_sistolica, na.rm = TRUE),
    max = max(pressao_sistolica, na.rm = TRUE),
    dp = sd(pressao_sistolica, na.rm = TRUE)

  )
sumario_pressao_sistolica
```

```
##       n    media mediana min max       dp
## 1 10000 122.1863     122  85 175 14.05137
```

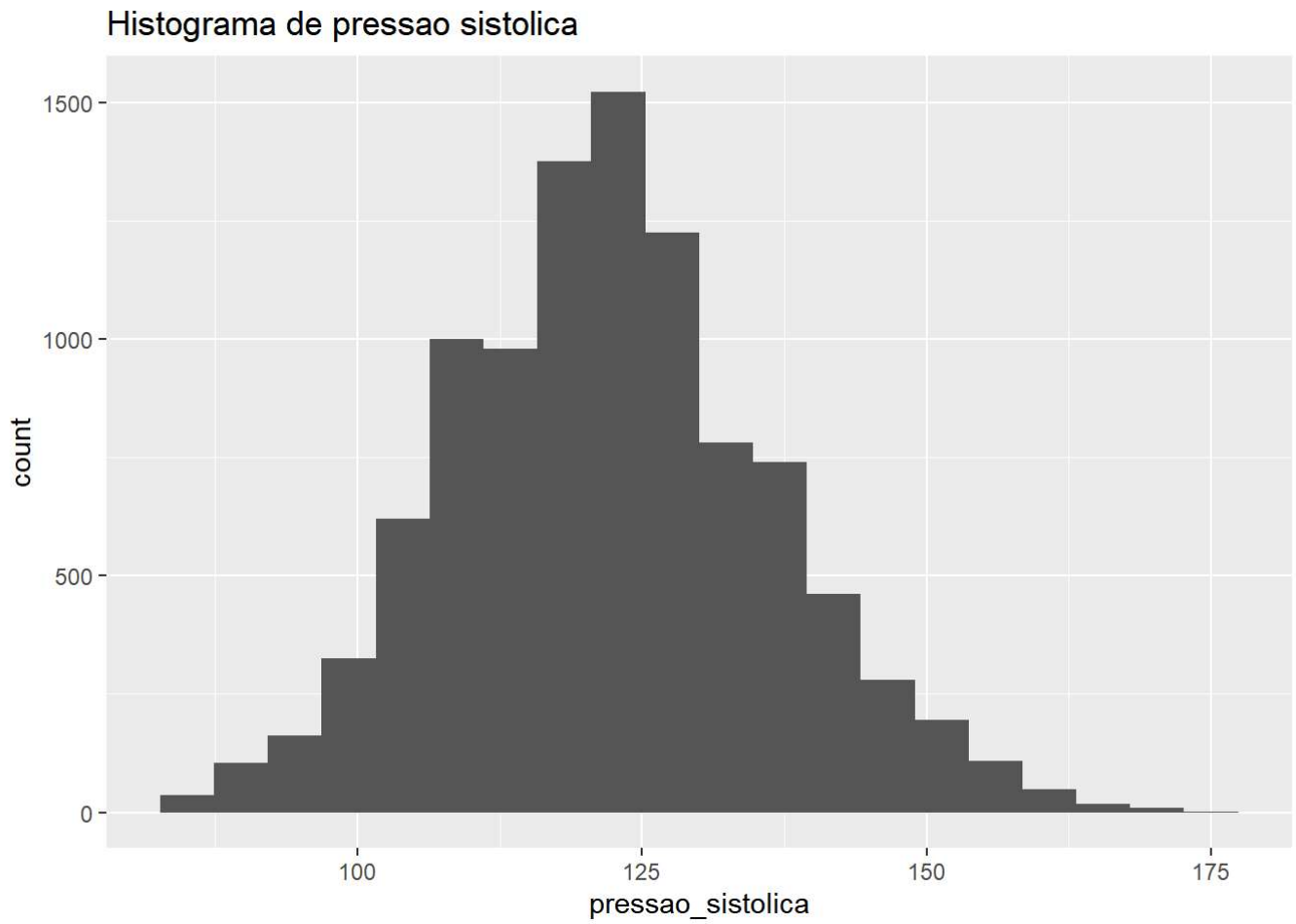
```
#Tanto a média quanto a mediana de idade e de pressão sistólica estão
#simétricas, evidenciando que os dados não possuem valores que destoam
#muito da média para interferir na mediana.

#5.2 Histograma e boxplot

ggplot(dados, aes(x = idade)) +
  geom_histogram(bins = 20) +
  labs(title = "Histograma de Idade")
```



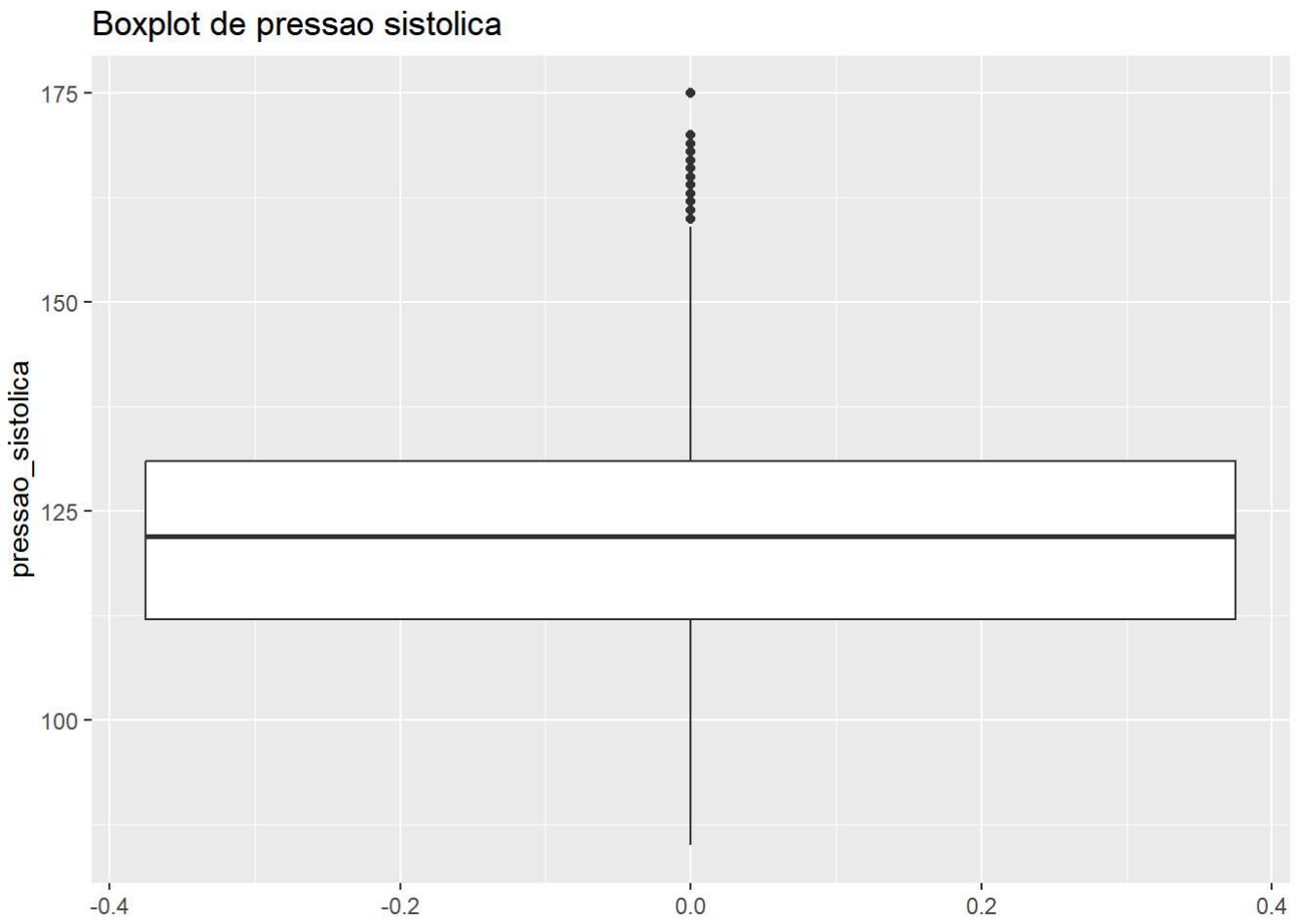
```
ggplot(dados, aes(y = idade)) +
  geom_boxplot() +
  labs(title = "Boxplot de Idade")
```




```
ggplot(dados, aes(x = pressao_sistolica)) +
  geom_histogram(bins = 20) +
  labs(title = "Histograma de pressao sistolica")
```




```
ggplot(dados, aes(y = pressao_sistolica)) +
  geom_boxplot() +
  labs(title = "Boxplot de pressao sistolica")
```




```
#O Histograma de idade tem a cauda ao lado direito, permite visualizar a
#moda na idade 40, o que converge com a média apresentada anteriormente.
#O boxplot de idade tem a mediana no 40, também conforme mostrado e permite
#visualizar os dados discrepantes das idades em torno de 80 anos, além de
#mostrar que o 1º quartil está em torno de 30 e o 3º quartil está em torno de
#50 anos.

#Já o gráfico de pressão sistólica permite visualizar que os maiores valores
#estão nas proximidades de 125, valor próximo à média e mediana de 122. O
#histograma possui uma simetria, ao contrário da idade. O boxplot apresenta
#também a mediana de 125 e os valores destoantes na faixa de 175, também
#permitindo visualizar os quartis.

#6 - Tabelas cruzadas

#6.1 - Cruzando duas qualitativas

tab_cross <- table(dados$diagnostico, dados$plano_saude)
tab_cross
```

```
##
##          ambos nenhum privado  SUS
##   DM         6     21      18   44
##   HAS       77    227     300  626
##   outros    35    112     163  302
##   sem      516   1508    1993 4052
```

```
round(100 * prop.table(tab_cross, margin = 2), 1)
```

```
##
##          ambos nenhum privado  SUS
##   DM       0.9    1.1     0.7  0.9
##   HAS     12.1   12.2    12.1 12.5
##   outros   5.5    6.0     6.6  6.0
##   sem     81.4   80.7    80.6 80.7
```

```
#No SUS, o mais prevalente é sem diagnóstico (80,7%), seguido de HAS (12,5%).
#Na rede privada, o mais prevalente é sem também (80,6%), seguido de HAS (12,1%).
#A mesma prevalência se repete em quem possui ambos e nenhum.

#6.2 - Resumo de quantitativa por grupo

dados |>
  group_by(sexo) |>
  summarise(
    n = n(),
    media_idade = mean(idade, na.rm = TRUE),
    dp_idade = sd(idade, na.rm = TRUE),
    mediana_idade = median(idade, na.rm = TRUE)
  )
```

```
## # A tibble: 2 × 5
##   sexo      n media_idade dp_idade mediana_idade
##   <fct> <int>       <dbl>    <dbl>         <dbl>
## 1 F      5192        40.3     12.5            40
## 2 M      4808        40.5     12.4            40
```

```
dados |>
  group_by(escolaridade) |>
  summarise(
    n = n(),
    media_idade = mean(idade, na.rm = TRUE),
    dp_idade = sd(idade, na.rm = TRUE),
    mediana_idade = median(idade, na.rm = TRUE)
  )
```

```
## # A tibble: 3 × 5
##   escolaridade     n media_idade dp_idade mediana_idade
##   <ord>        <int>       <dbl>    <dbl>         <dbl>
## 1 Fundamental   3399        40.2     12.4            40
## 2 Medio         4056        40.5     12.6            40
## 3 Superior      2545        40.4     12.5            40
```

```
#Em relação à idade por sexo, há uma semelhança na amostra. A média de homens
#e mulheres é próxima (40,5 e 40,3, respectivamente), com a mesma mediana de 40.
#Na escolaridade, a mediana de 40 se repete pois é a mesma amostra, sendo que
#a média de idade para superior, médio e fundamental também estão na faixa dos
#40 anos. Os desvios-padrão de ambos os grupos também mantém-se próximas pela
#lógica.

#7 - Valores ausentes e outliers

#7.1 - Ausentes

colSums(is.na(dados))
```

```
##                id              sexo      escolaridade       anos_estudo
##                 0                 0                 0                 0
##      rede_escolar         municipio                UF             idade
##                 0                 0                 0                 0
##        faltas_esc    tempo_estudo_h pressao_sistolica       diagnostico
##                 0                 0                 0                 0
##       plano_saude
##                 0
```

```
#A base de dados não retornou valores de células vazias, que seriam os NA.

#7.2 - Outliers

Q <- quantile(dados$tempo_estudo_h, probs = c(.25, .75), na.rm = TRUE)
IQRv <- IQR(dados$tempo_estudo_h, na.rm = TRUE)
lim_inf <- Q[1] - 1.5 * IQRv
lim_sup <- Q[2] + 1.5 * IQRv
subset_out <- dados |>
  filter(tempo_estudo_h < lim_inf | tempo_estudo_h > lim_sup)
nrow(subset_out); head(subset_out)
```

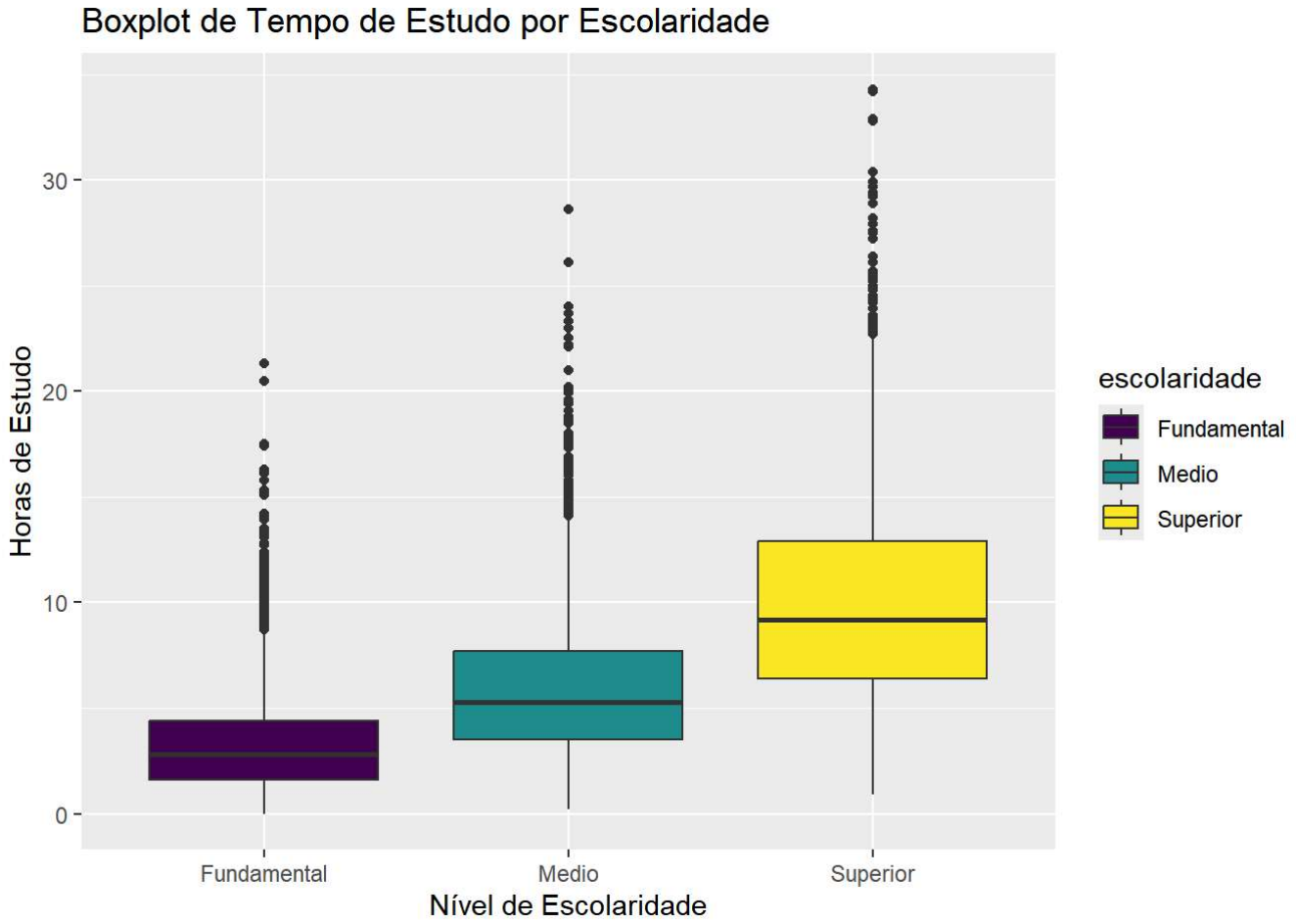
```
## [1] 348
```

```
##       id sexo escolaridade anos_estudo rede_escolar       municipio UF idade
## 1 P00033    F        Medio          11      privada        Blumenau SC    63
## 2 P00037    M     Superior          15      privada         Pelotas RS    44
## 3 P00067    M     Superior          14      pública        São Luís MA    44
## 4 P00086    F     Superior          18      privada      Imperatriz MA    31
## 5 P00095    F     Superior          15      pública  Rio de Janeiro RJ    55
## 6 P00183    M     Superior          16      privada Duque de Caxias RJ    37
##   faltas_esc tempo_estudo_h pressao_sistolica diagnostico plano_saude
## 1          0           19.4              132         sem         SUS
## 2          0           17.6              133         sem         SUS
## 3          0           16.9              111      outros     privado
## 4          1           18.0              126         sem       ambos
## 5          2           21.3              136         HAS         SUS
## 6          1           18.6              116         sem      nenhum
```

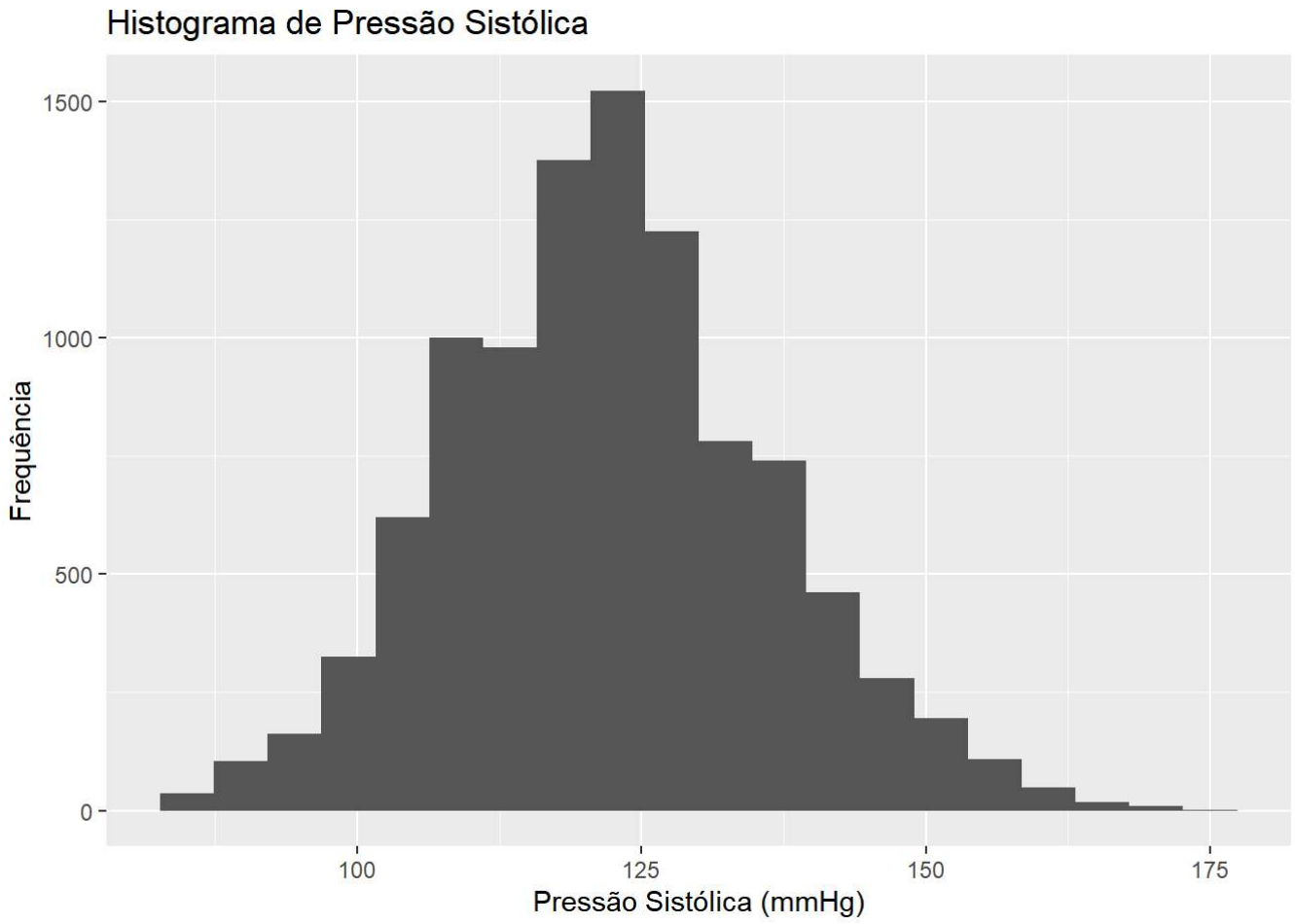
```
#Essa amostra possui 348 outliers. Esses dados são informações válidas, pois
#mesmo que destoem dos demais permitem interpretar o que motiva esses casos
#variados, que podem ser desde preenchimento incorreto até variáveis
#explicativas dentro das políticas públicas.

#8. Exercícios aplicados (educação e saúde)

#8.1 - Educação

#Distribuição de rede escolar

tab_rede <- table(dados$rede_escolar)
prop_rede <- prop.table(tab_rede)
cbind(FA = tab_rede, FR = round(100 * prop_rede, 1))
```

```
##           FA FR
## privada 2796 28
## pública 7204 72
```

```
#Gráfico boxplot

ggplot(dados, aes(x = escolaridade, y = tempo_estudo_h, fill = escolaridade)) +
  geom_boxplot() +
  labs(title = "Boxplot de Tempo de Estudo por Escolaridade",
       x = "Nível de Escolaridade",
       y = "Horas de Estudo")
```



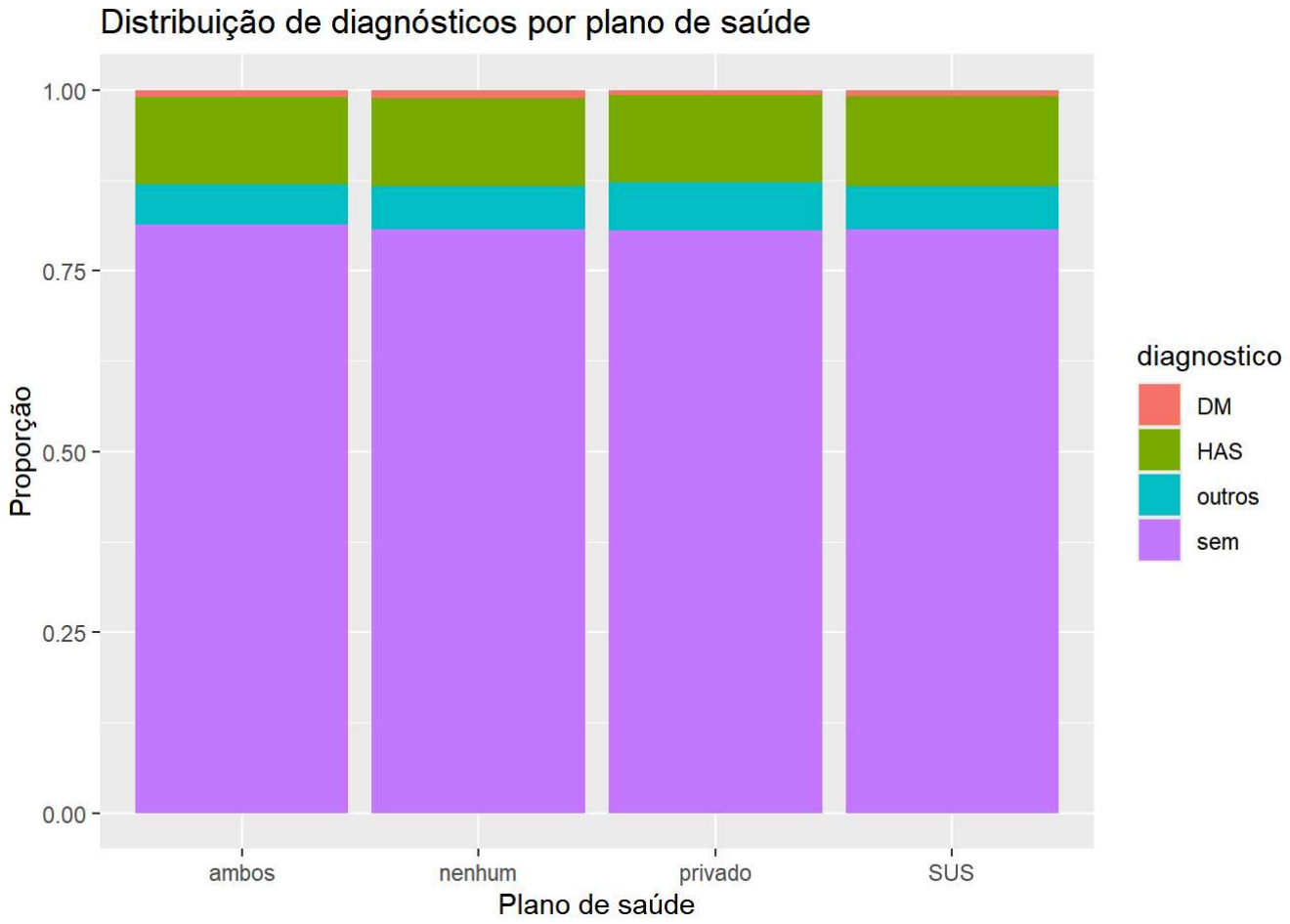
```
#Interprete

#A frequência absoluta é maior na rede pública, representando 72% da amostra
#em contraposição com 28% da privada. Quanto ao boxplot de tempo de estudo
#por escolaridade, este evidencia que há sim um padrão monotônico que é
#crescente: as medianas vão crescendo logicamente, assim como os percentis
#mostrados. Ainda que haja valores discrepantes, estes também sobem, o que é
#lógico visto que uma escolaridade maior exige um tempo de estudo proporcional.

#8.2 - Saúde

#Pressão sistólica - histograma e informações

ggplot(dados, aes(x = pressao_sistolica)) +
  geom_histogram(bins = 20) +
  labs(title = "Histograma de Pressão Sistólica",
       x = "Pressão Sistólica (mmHg)",
       y = "Frequência")
```




```
dados |>
  summarise(
    media = mean(pressao_sistolica, na.rm = TRUE),
    mediana = median(pressao_sistolica, na.rm = TRUE),
    dp = sd(pressao_sistolica, na.rm = TRUE)
  )
```

```
##      media mediana       dp
## 1 122.1863     122 14.05137
```

```
#Cruzamento diagnóstico e plano de saúde

ggplot(dados, aes(x = plano_saude, fill = diagnostico)) +
  geom_bar(position = "fill") +
  labs(x = "Plano de saúde",
       y = "Proporção",
       title = "Distribuição de diagnósticos por plano de saúde")
```




```
#Interprete

#Em todos os planos prevalecem respectivamente sem diagnóstico, HAS, outros e
#DM.
```