Atividade 1 - Gabarito

Prof. Leticia Raposo

2019

Doença Cardíaca

O banco de dados disponível no arquivo **ExercicioDoencacardiaca.txt** contém 303 observações e 9 variáveis. As descrições de cada variável são:

- **age**: idade em anos
- **sex**: sexo (1 = masculino; 0 = feminino)
- **cp**: tipo de dor no peito (1 = angina típica; 2 = angina atípica; 3 = sem dor angínica; 4 = assintomático)
- **trstbps**: pressão sanguínea em repouso (em mm/Hg ao ser admitido no hospital)
- **chol**: colesterol em mg/dl
- **fbs**: açúcar no sangue em jejum > 120 mg/dl (1 = V; 0 = F)
- **thalach**: frequência cardíaca máxima alcançada
- **exang**: exercício induziu angina (1 = sim; 0 = não)
- **num**: diagnóstico de doença cardíaca (0 =< 50% estreitamento de diâmetro; 1 => 50% estreitamento de diâmetro)

Antes de iniciar a análise, verifique e responda:

- *(0,3 ponto) As variáveis foram lidas (codificadas) corretamente pelo R? Se não, codifique-as corretamente.*

```
dados <- read.table("C:/Users/Leticia/Google Drive/UNIRIO/Disciplinas Ministradas/2019.2/Biologia - Bio
str(dados)
```

```
## 'data.frame':    303 obs. of  9 variables:
##  $ age    : int  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex    : int  1 1 1 1 0 1 0 0 1 1 ...
##  $ cp     : int  1 4 4 3 2 2 4 4 4 4 ...
##  $ trstbps: int  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol   : int  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs    : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ thalach: int  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang  : int  0 1 1 0 0 0 0 1 0 1 ...
##  $ num    : int  0 1 1 0 0 0 1 0 1 1 ...
```

O aluno precisa mostrar que verificou as variáveis pelo str. Se não tiver isso, retire 0,1.

Não, algumas variáveis estão codificadas como inteiro e deveriam ser fatores.

```
dados$sex <- as.factor(dados$sex)
dados$cp <- as.factor(dados$cp)
dados$sex <- as.factor(dados$sex)
dados$fbs <- as.factor(dados$fbs)
dados$exang <- as.factor(dados$exang)
dados$num <- as.factor(dados$num)
```

- *(0,2 ponto) Há dados ausentes?*

```
summary(dados)
```

```
##       age        sex     cp         trstbps           chol       fbs
##  Min.   :29.00   0: 97   1: 23   Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:48.00   1:206   2: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45
```

```
##  Median :56.00          3: 86   Median :130.0   Median :241.0
##  Mean   :54.44          4:144   Mean   :131.7   Mean   :246.7
##  3rd Qu.:61.00                  3rd Qu.:140.0   3rd Qu.:275.0
##  Max.   :77.00                  Max.   :200.0   Max.   :564.0
##     thalach      exang   num
##  Min.   : 71.0   0:204   0:164
##  1st Qu.:133.5   1: 99   1:139
##  Median :153.0
##  Mean   :149.6
##  3rd Qu.:166.0
##  Max.   :202.0
```

O aluno precisa mostrar que verificou a presença ou não de dados ausentes por algum comando. Se não fez isso, retire 0,1.

Não. Nenhum NA apareceu com o comando summary.

1. Para as variáveis trestbps e chol, pede-se:

i. *(0,3 ponto) Calcule a média aritmética, a mediana e a moda;*
ii. *(0,3 ponto) Calcule o primeiro e o terceiro quartis e também o IQR.*
iii. *(0,8 ponto) Calcule as medidas de dispersão (amplitude, variância, desvio-padrão e coeficiente de variação);*

```
library(summarytools)
descr(dados$trstbps)
```

```
## Descriptive Statistics
## dados$trstbps
## N: 303
##
##                     trstbps
## ----------------- ---------
##              Mean    131.69
##           Std.Dev     17.60
##               Min     94.00
##                Q1    120.00
##            Median    130.00
##                Q3    140.00
##               Max    200.00
##               MAD     14.83
##               IQR     20.00
##                CV      0.13
##          Skewness      0.70
##       SE.Skewness      0.14
##          Kurtosis      0.82
##           N.Valid    303.00
##         Pct.Valid    100.00
```

Para as letras i, ii, iii, aluno pode calcular por qualquer forma, desde que seja por meio do R. Se ele não especificou cada valor, retire 0,1 em cada alternativa.

```
descr(dados$chol)
```

```
## Descriptive Statistics
## dados$chol
## N: 303
##
##                      chol
## ----------------- --------
##              Mean   246.69
##           Std.Dev    51.78
```

```
##                Min   126.00
##                 Q1   211.00
##             Median   241.00
##                 Q3   275.00
##                Max   564.00
##                MAD    47.44
##                IQR    64.00
##                 CV     0.21
##           Skewness     1.12
##        SE.Skewness     0.14
##           Kurtosis     4.35
##            N.Valid   303.00
##          Pct.Valid   100.00
```

```
library(DescTools)
Mode(dados$trstbps, na.rm = T)
```

```
## [1] 120
```

```
Mode(dados$chol, na.rm = T)
```

```
## [1] 197 204 234
```

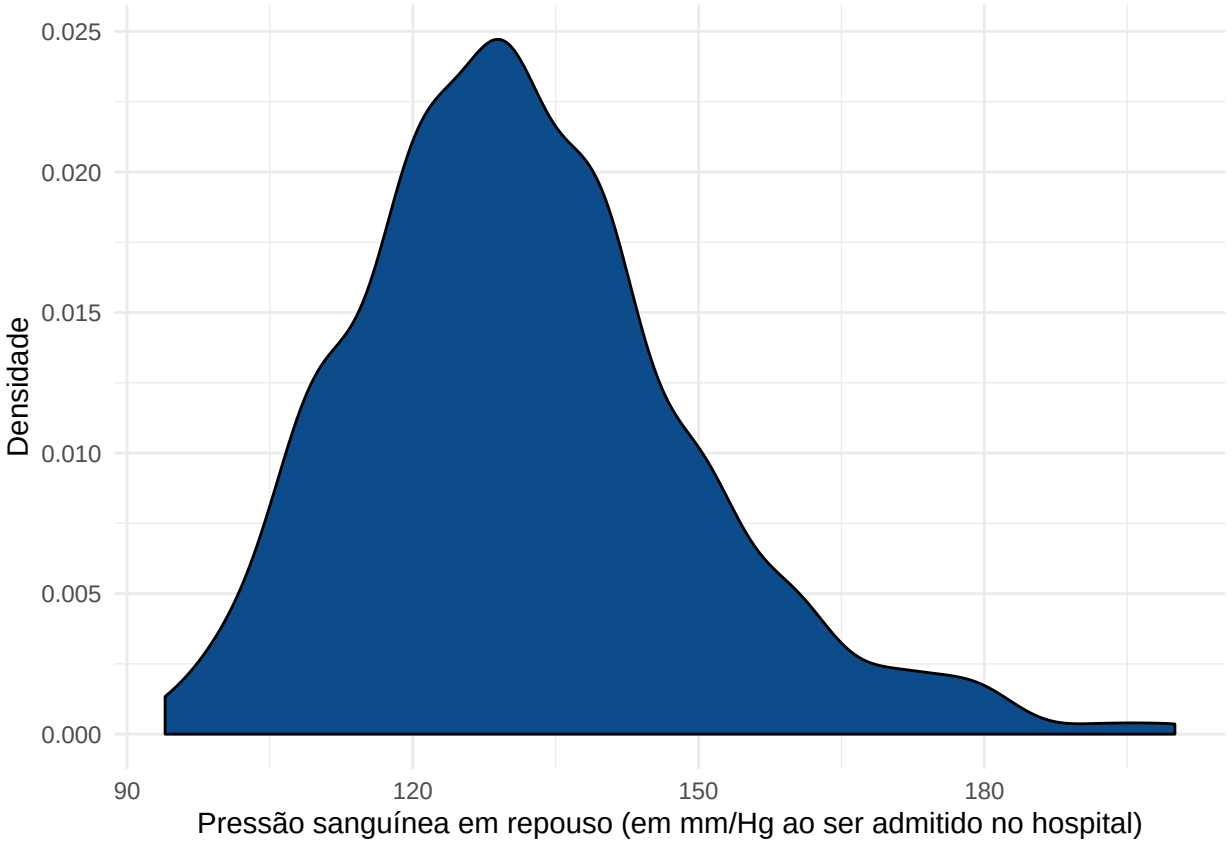
- trestbps: média = 131,69 mm/Hg; mediana = 130 mm/Hg; moda = 120 mm/Hg.
- chol: média = 246,69 mg/dl; mediana = 241 mg/dl; moda = 197, 204 e 234 mg/dl.
- trestbps: Q1 = 120 mm/Hg; Q3 = 140 mm/Hg; IQR = 20 mm/Hg.
- chol: Q1 = 211 mg/dl; Q3 = 275 mg/dl; IQR = 64 mg/dl.
- trestbps: amplitude = 106 mm/Hg; variância = 309,75 mm/Hg^2; dp = 17,60 mm/Hg; CV = 0,13.
- chol: amplitude = 438 mg/dl; variância = 2680,85 mg/dl^2; dp = 51,78 mg/dl; CV = 0,21.

Ambos os CV indicam uma homogeneidade das duas variáveis.

iv. *(0,4 ponto) Verifique se a distribuição é simétrica, assimétrica positiva ou assimétrica negativa;*

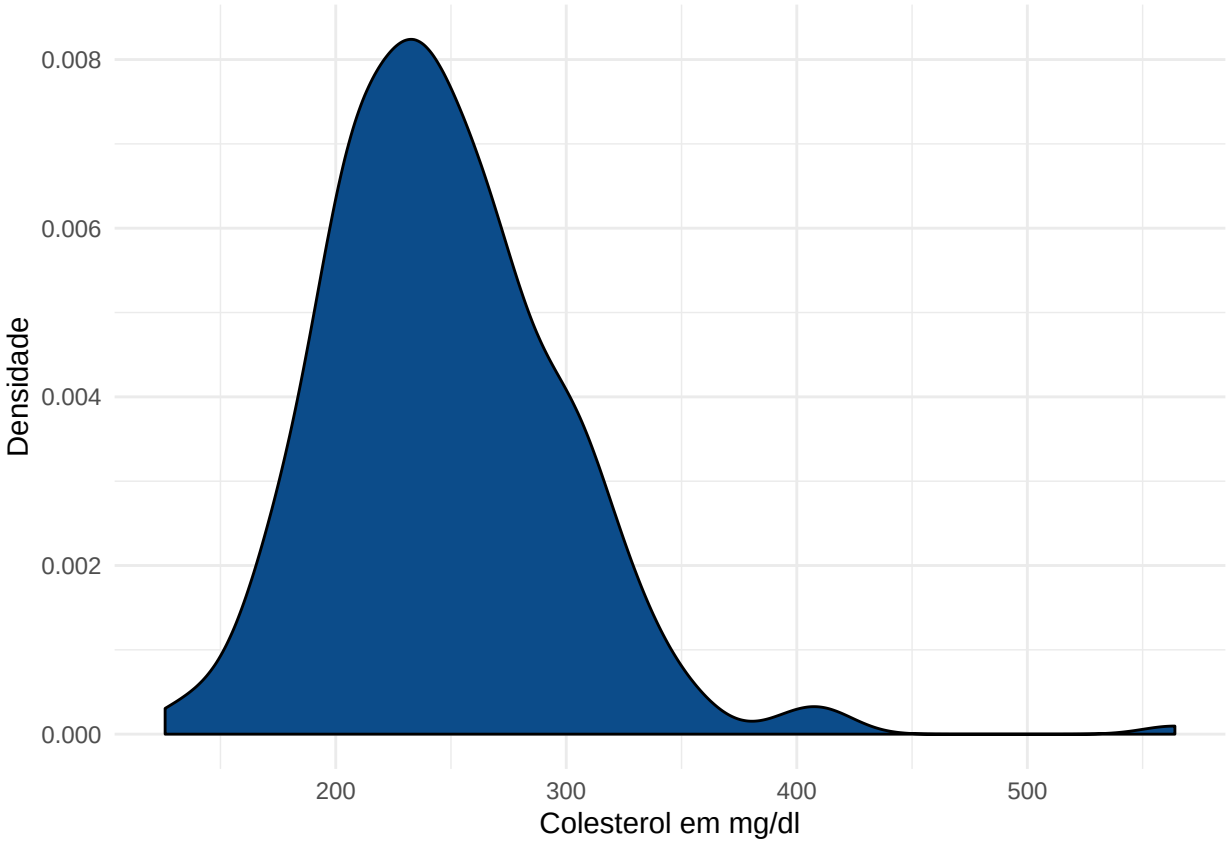
```
library(ggplot2)

ggplot(dados) +
 aes(x = trstbps) +
 geom_density(adjust = 1L, fill = "#0c4c8a") +
 labs(x = "Pressão sanguínea em repouso (em mm/Hg ao ser admitido no hospital)",
      y = "Densidade") +
 theme_minimal()
```

O aluno pode verificar por qualquer gráfico (histograma, densidade ou boxplot) ou medida de assimetria.
Se não colocar gráfico ou medida, retire 0,1 e se não falar o tipo de assimetria, retire 0,1.




```
ggplot(dados) +
 aes(x = chol) +
 geom_density(adjust = 1L, fill = "#0c4c8a") +
 labs(x = "Colesterol em mg/dl", y = "Densidade") +
 theme_minimal()
```

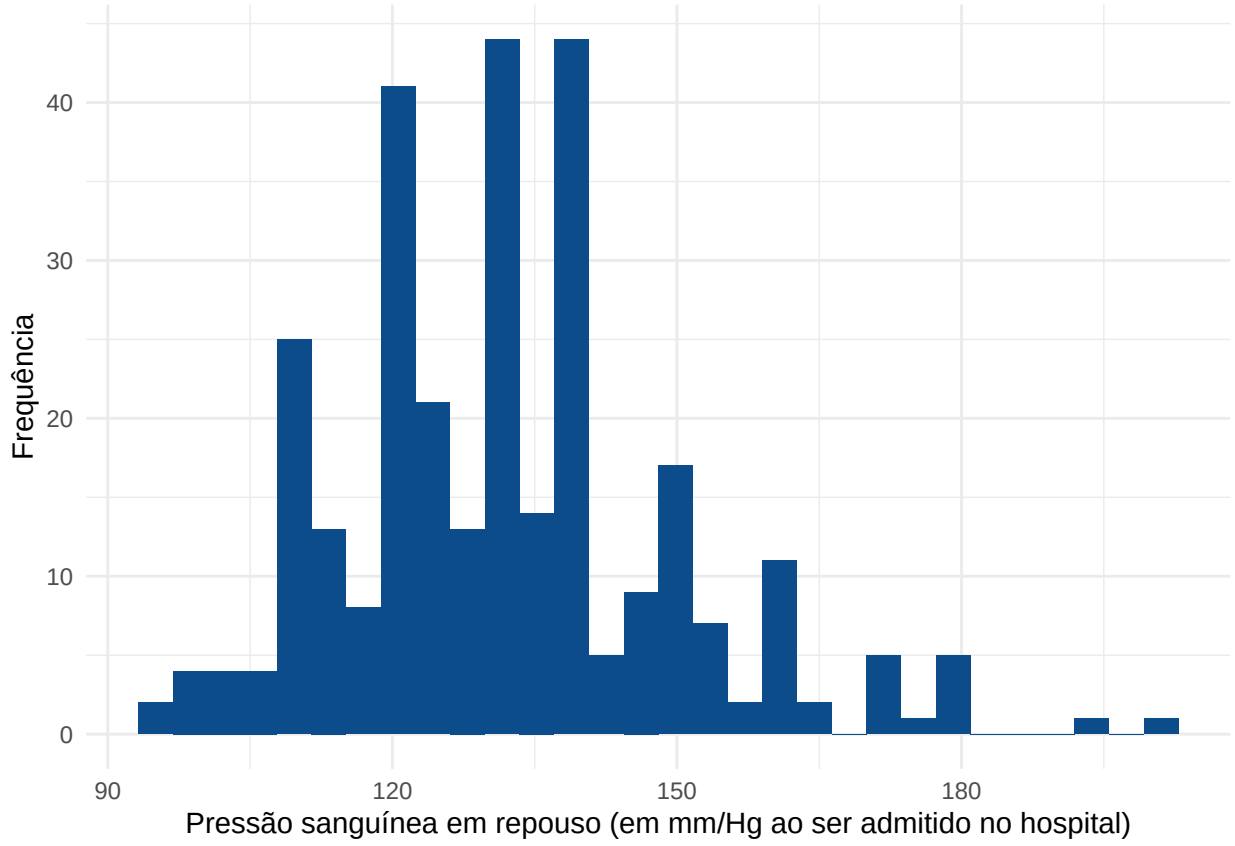
Se os eixos X e Y não forem definidos com textos em português, retire 0,1 de todas as questões que tenham gráficos.



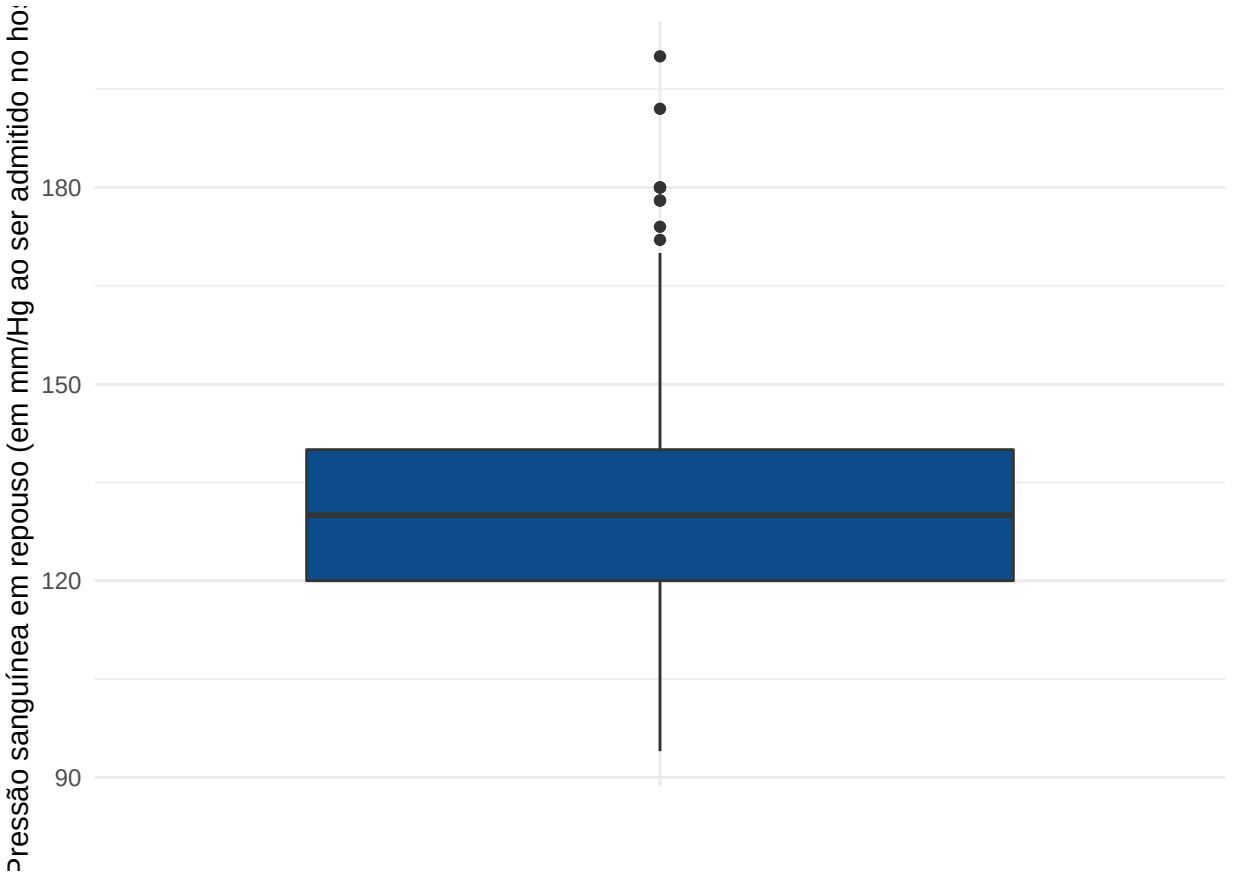
Aparentemente, as duas distribuições apresentam assimetria à direita, pois a cauda da direita é maior.

v. *(0,4 ponto) Construa o histograma e o boxplot para as variáveis em estudo.*

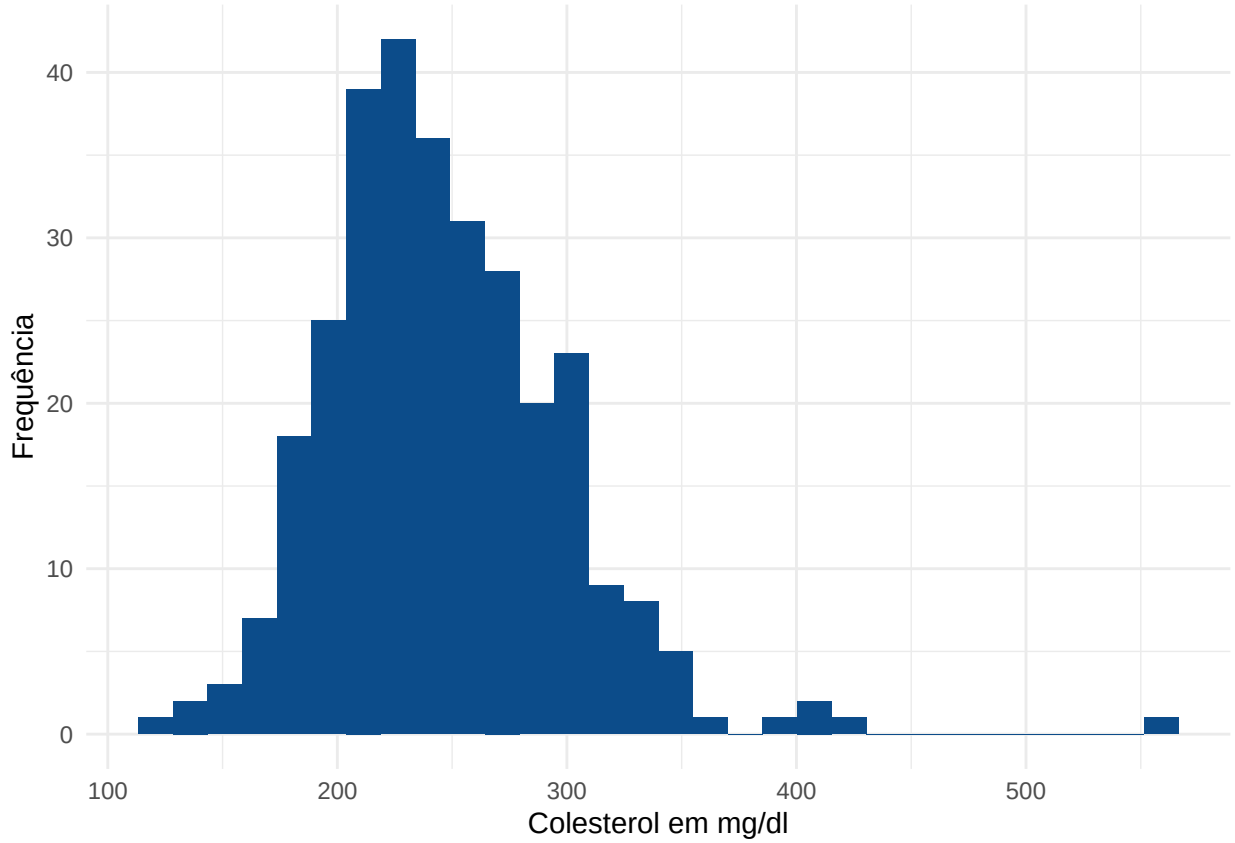
```
library(ggplot2)

# trstbps
ggplot(dados) +
 aes(x = trstbps) +
 geom_histogram(bins = 30L, fill = "#0c4c8a") +
 labs(x = "Pressão sanguínea em repouso (em mm/Hg ao ser admitido no hospital)",
      y = "Frequência") +
 theme_minimal()
```



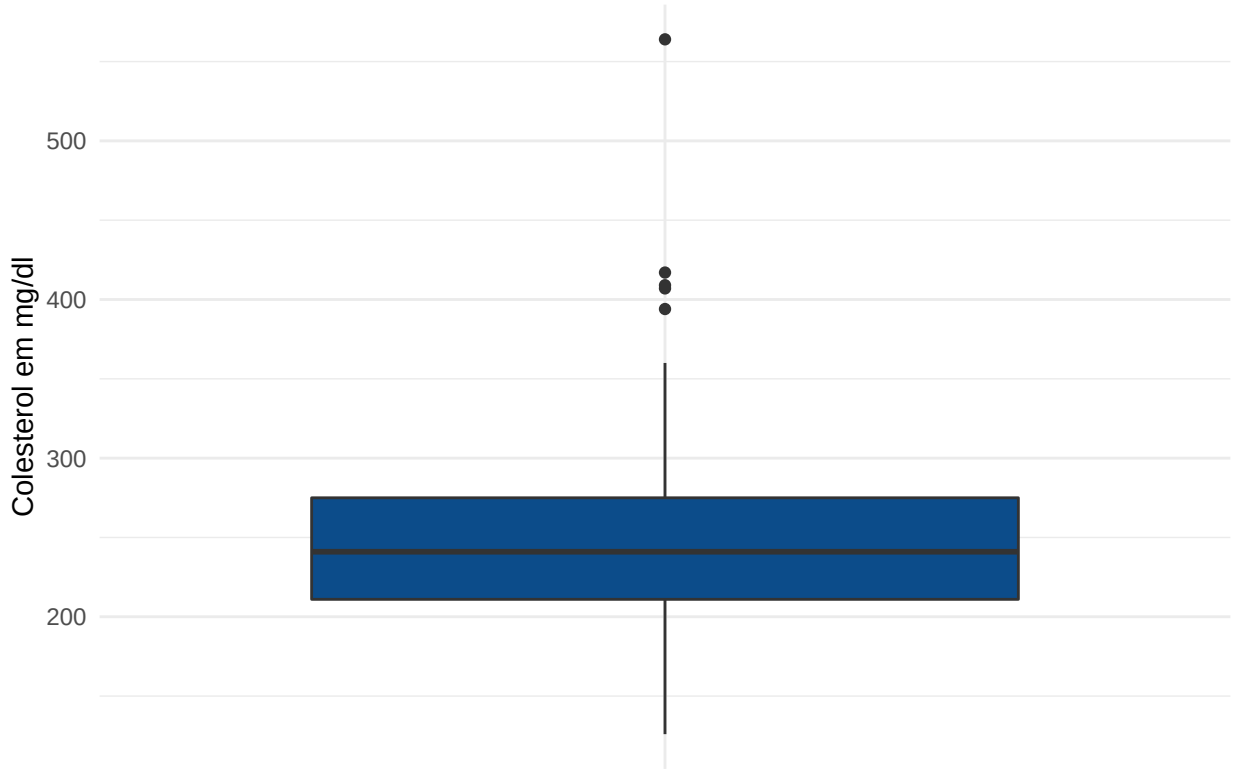

```
ggplot(dados) +
 aes(x = "", y = trstbps) +
 geom_boxplot(fill = "#0c4c8a") +
 labs(x = " ", y = "Pressão sanguínea em repouso (em mm/Hg ao ser admitido no hospital)") +
 theme_minimal()
```




```
# chol
ggplot(dados) +
 aes(x = chol) +
 geom_histogram(bins = 30L, fill = "#0c4c8a") +
 labs(x = "Colesterol em mg/dl", y = "Frequência") +
 theme_minimal()
```




```
ggplot(dados) +
 aes(x = "", y = chol) +
 geom_boxplot(fill = "#0c4c8a") +
 labs(x = " ", y = "Colesterol em mg/dl") +
 theme_minimal()
```



vi. *(0,3 ponto) É possível observar outliers para as duas variáveis analisadas?* Sim, 6 outliers para a variável trstbps e 5 para a variável chol.

Basta uma resposta sim e já é o suficiente para ganhar a questão toda.

2. Para a variável cp, pede-se:

i. *(1,0 ponto) Construa uma tabela de distribuição de frequências. Comente o resultado.*

```
freq(dados$cp)
```

```
## Frequencies
## dados$cp
## Type: Factor
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##           1     23      7.59           7.59      7.59           7.59
##           2     50     16.50          24.09     16.50          24.09
##           3     86     28.38          52.48     28.38          52.48
##           4    144     47.52         100.00     47.52         100.00
##        <NA>      0                                0.00         100.00
##       Total    303    100.00         100.00    100.00         100.00
```

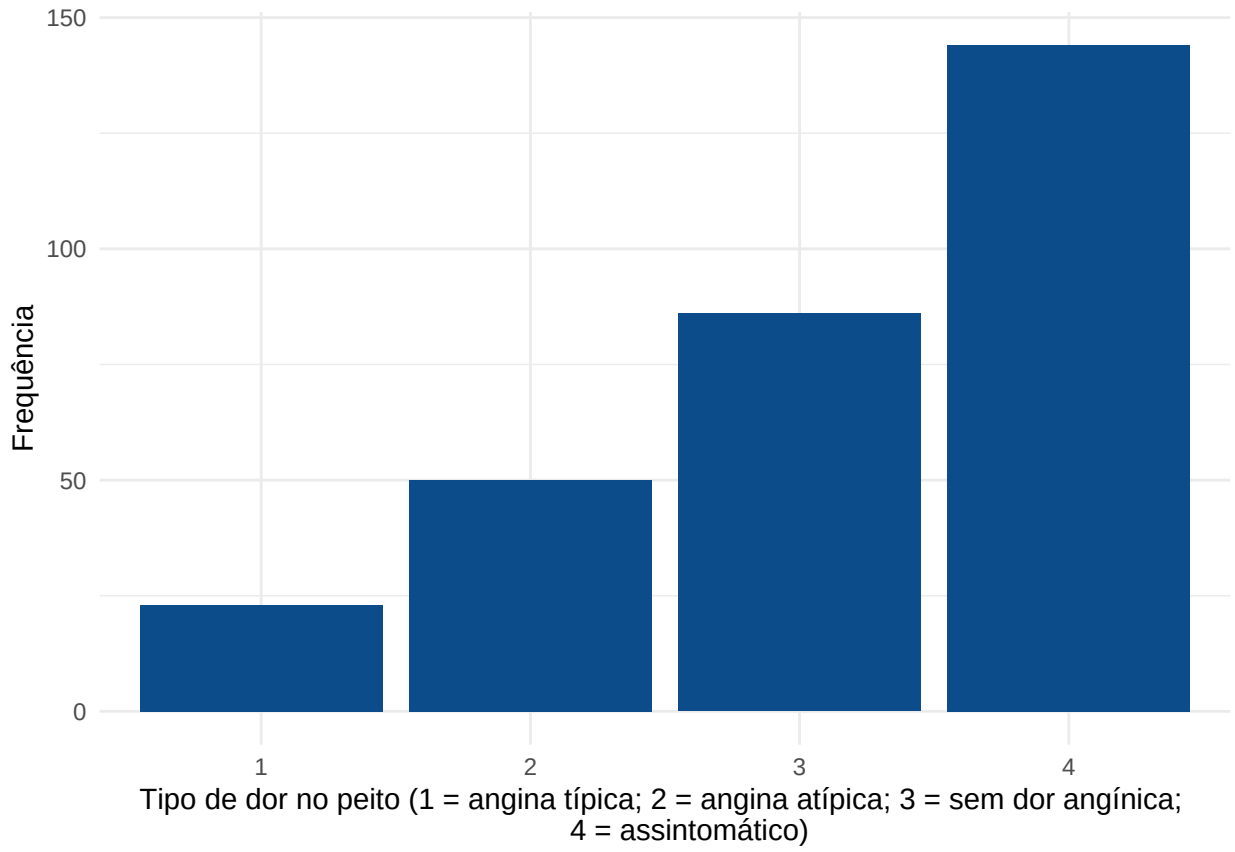
Pode ser qualquer tipo de tabela de distribuição de frequências com frequência absoluta e relativa. Se não tiver um comentário sobre a tabela, retire 0,4. Se não tiver a frequência relativa, retire 0,1.

É possível observar que a maior parte (47,52%) dos pacientes era assintomática em relação à dor no peito. Apenas 7,59% apresentaram angina típica.

ii. *(0,5 ponto) Elabore um gráfico de barras.*

```
library(ggplot2)
```

```r
ggplot(dados) +
 aes(x = cp) +
 geom_bar(fill = "#0c4c8a") +
 labs(x = "Tipo de dor no peito (1 = angina típica; 2 = angina atípica; 3 = sem dor angínica;
      4 = assintomático)", y = "Frequência") +
 theme_minimal()
```



3. Para as variáveis fbs e num, pede-se:

i. *(1,0 ponto) Construa tabelas de contingência com os perfis linha e coluna.*
ii. *(0,5 ponto) Calcule o valor da estatística qui-quadrado. O que podemos interpretar com esse resultado?*

```r
ctable(dados$fbs, dados$num, prop = "r", chisq = T)
```

```
## Cross-Tabulation, Row Proportions
## fbs * num
## Data Frame: dados
##
##
## ------- ----- ------------- ------------- --------------
##           num             0             1          Total
##     fbs
##       0         141 (54.7%)   117 (45.3%)   258 (100.0%)
##       1          23 (51.1%)    22 (48.9%)    45 (100.0%)
##   Total         164 (54.1%)   139 (45.9%)   303 (100.0%)
## ------- ----- ------------- ------------- --------------
##
```

Se apenas uma das tabelas tiver sido feita, dê apenas a metade da questão.

```
## ----------------------------
##  Chi.squared   df   p.value
## ------------- ---- ---------
##    0.0771      1    0.7813
## ----------------------------
```

```
ctable(dados$fbs, dados$num, prop = "c")
```

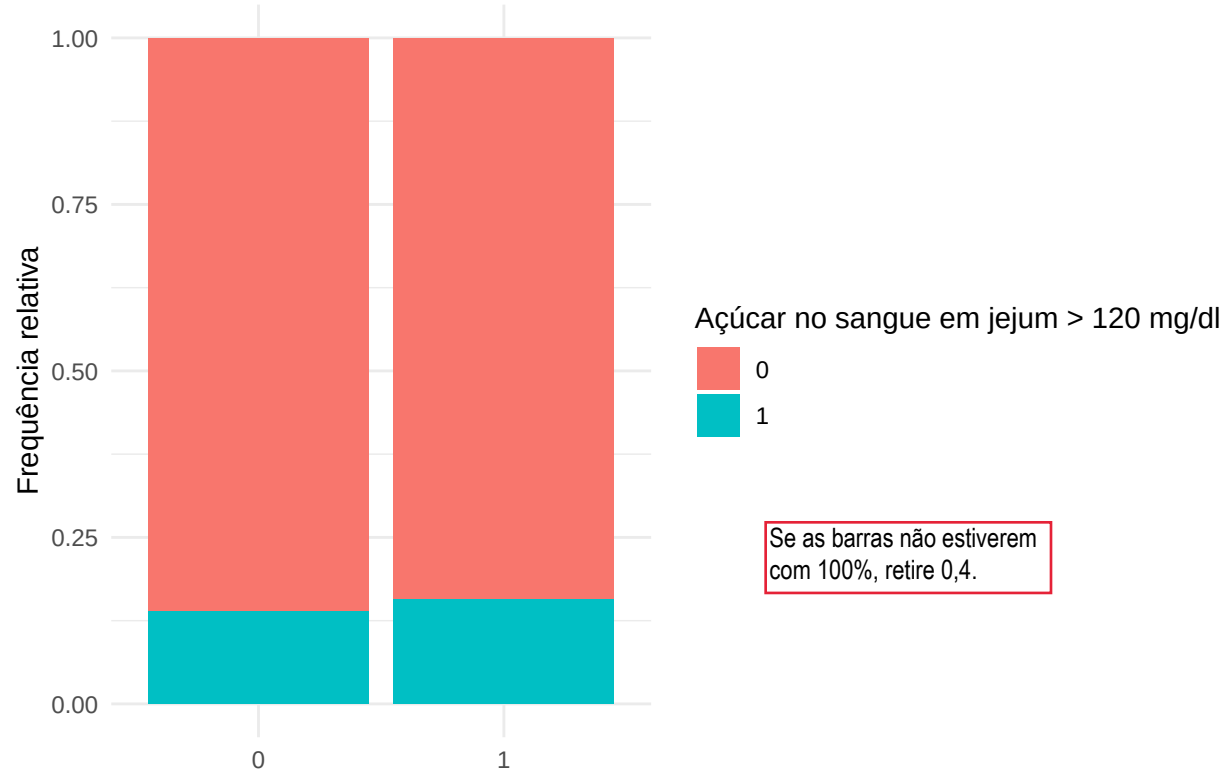
```
## Cross-Tabulation, Column Proportions
## fbs * num
## Data Frame: dados
##
## ------- ----- -------------- -------------- --------------
##           num              0              1          Total
##     fbs
##       0         141 ( 86.0%)   117 ( 84.2%)   258 ( 85.1%)
##       1          23 ( 14.0%)    22 ( 15.8%)    45 ( 14.9%)
##   Total         164 (100.0%)   139 (100.0%)   303 (100.0%)
## ------- ----- -------------- -------------- --------------
```

Veja se foi feita uma interpretação. Se a conclusão estiver errada, retire 0,2.

Como o valor-p foi maior que 0,05, não há uma associação estatisticamente significativa entre ter ou não açúcar no sangue em jejum $>$ 120 mg/dl e o diagnóstico de doença cardíaca.

iii. *(0,5 ponto)Contrua um gráfico com barras empilhadas.*

```
library(ggplot2)

ggplot(dados) +
 aes(x = num, fill = fbs) +
 geom_bar(position = "fill") +
 scale_fill_hue() +
 labs(x = "Diagnóstico de doença cardíaca (0 =< 50% estreitamento de diâmetro;
      1 => 50% estreitamento de diâmetro)", y = "Frequência relativa", fill = "Açúcar no sangue em jejun
 theme_minimal()
```

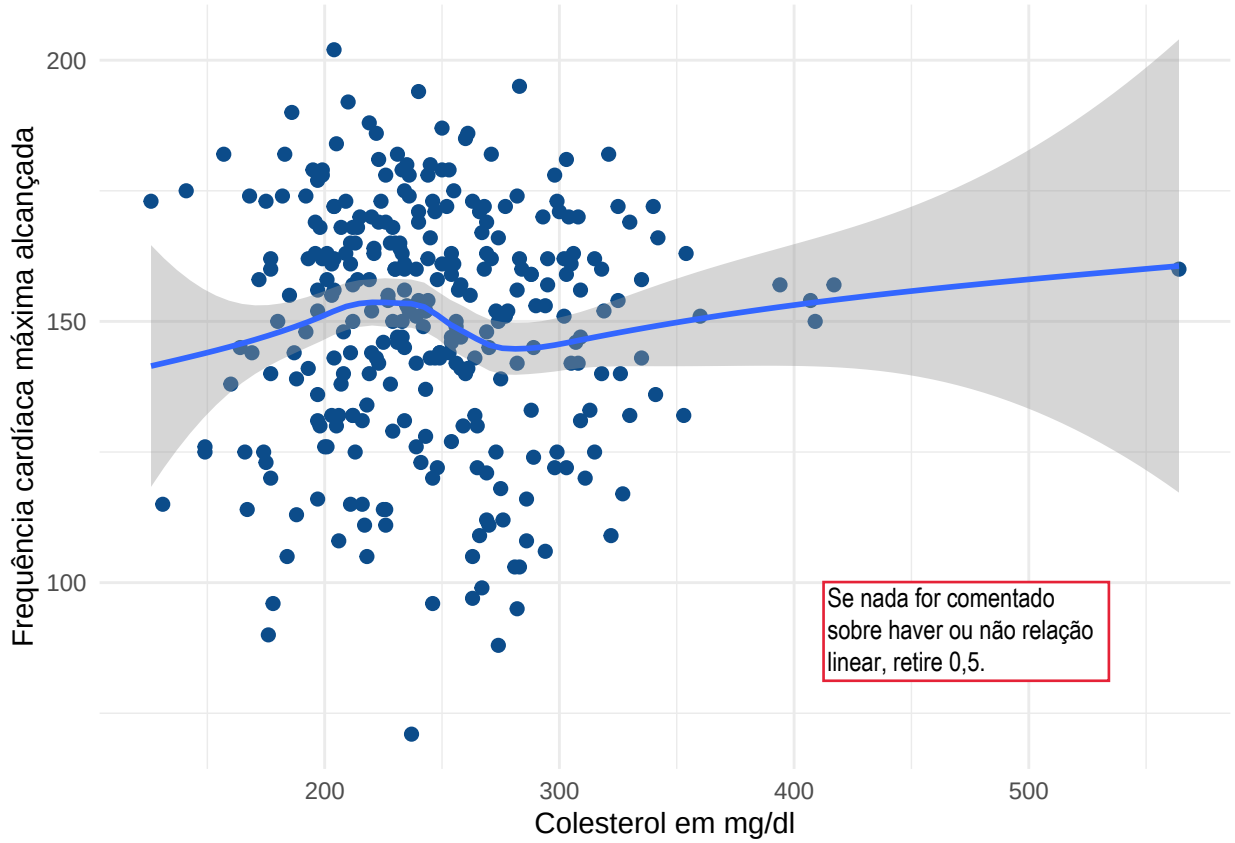


4. Para as variáveis chol e thalach:

i. *(1,0 ponto) Construa um gráfico de dispersão e avalie se há indícios de relação linear entre as variáveis.*

```
library(ggplot2)

ggplot(dados) +
 aes(x = chol, y = thalach) +
 geom_point(size = 2L, colour = "#0c4c8a") +
 geom_smooth(span = 0.75) +
 labs(x = "Colesterol em mg/dl", y = "Frequência cardíaca máxima alcançada") +
 theme_minimal()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



O gráfico de dispersão não dá indícios de relação linear entre as variáveis. Podemos observar que os pontos se assemelham a uma nuvem sem tendência.

ii. *(1,0 ponto)Calcule também o coeficiente de correlação de Pearson e interprete-o.*

Se não tiver uma interpretação do resultado, retire 0,5.

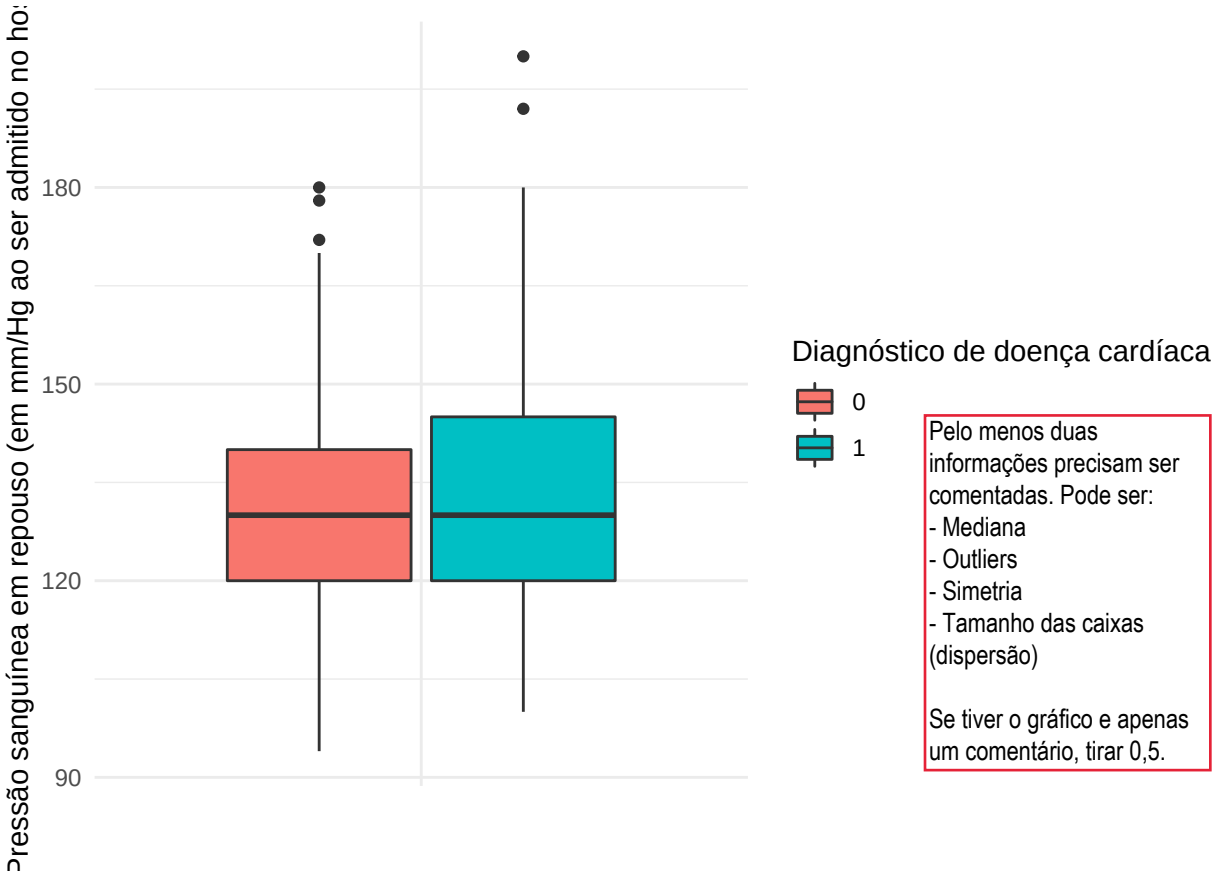
```
cor(dados$chol, dados$thalach)
```

```
## [1] -0.003431832
```

O coeficiente de correlação de Pearson mostra um valor muito próximo de zero, indicando a não existência de relação linear.

5. (1,5 ponto) Construa boxplots para a variável trstbps segundo a variável num. Descreva o que é observado no gráfico.

```
library(ggplot2)

ggplot(dados) +
 aes(x = "", y = trstbps, fill = num) +
 geom_boxplot() +
 scale_fill_hue() +
 labs(x = " ", y = "Pressão sanguínea em repouso (em mm/Hg ao ser admitido no hospital)",
      fill = "Diagnóstico de doença cardíaca") +
 theme_minimal()
```



É possível observar que em ambas as situações, os pacientes apresentaram medianas de pressão sanguínea próximas. O grupo dos pacientes sem doença cardíaca apresentou 3 outliers para a variável trstbps e o grupo com a doença cardíaca apresentou 2 outlier.

DESAFIO! (Bônus) Construa um gráfico não ensinado em sala de aula utilizando alguma das variáveis do banco de dados.