

Análise descritiva dos dados: funções, tabelas e gráficos

Curso de Bioestatística com R

Ornella Scardua Ferreira

 ornscar@gmail.com  [@ornscar](https://twitter.com/ornscar)  [@ornscar](https://twitter.com/ornscar)

Sobre mim



Amo gráficos como amo cavalos. Gosto de música ruim e de cinema (bom). Sou apaixonada pelo Botafogo e pelo Bayern de Munique. Prefiro Vila Velha a qualquer lugar no mundo. Não tenho sonhos, mas um dia espero ver a Palestina livre.

Cronograma

1. Dados

- Base 1: dados sobre gestantes diagnosticadas com diabetes gestacional.
- Base 2: dados do Programa das Nações Unidas para o Desenvolvimento (PNUD).

2. Funções bases do

- Frequência relativa e absoluta.
- Medidas-resumo: medidas de posição e de dispersão.

3. Tabelas descritivas

- O pacote `{gtsummary}`.
- Tabela com frequências relativas e absolutas e medidas-resumos.

4. Gráficos

- O pacote `{ggplot2}`.
- Gráficos univariados: barras e histogramas.
- Gráficos bivariados:
 - Variáveis qualitativas x qualitativas: barras agrupadas e empilhadas.
 - Variáveis quantitativas x quantitativas: dispersão.
 - Variáveis qualitativas x quantitativas: linhas e *boxplot*.

Os dados

Base 1

- A base de dados é sobre **gestantes diagnosticadas com diabetes gestacional** que realizaram o pré-natal entre os anos de 2012 a 2015 no Hospital das Clínicas da Universidade de São Paulo.
- Contém **408 observações** e **10 variáveis**, a saber:

idade: idade da gestante;

imc_classe: IMC categórico;

n_gestacoes: número de gestações anteriores;

hist_diab_fam: histórico de diabetes na família;

hist_diab_gest: histórico de diabetes gestacional;

macrossomia_fetal: antecedente de macrossomia fetal;

tabagista: indicador de tabagista;

hac: indicador de hipertensão;

glicemia_jejum: valor do exame de glicemia de jejum (em mg/dL);

insulina: se a gestante precisou usar insulina antes do parto.

- No :

```
# carregando os dados
dados1 <- readxl::read_xlsx("dados/diabetes.xlsx")

# panorama da base de dados
dplyr::glimpse(dados1)
```

```
## Rows: 408
## Columns: 10
## $ idade              <dbl> 24, 42, 35, 34, 42, 41, 41, 37, 31, 39, 28, 37, 42, ...
## $ imc_classe         <chr> "Até normal", "Obeso", "Sobre peso", "Obeso", "Obeso"...
## $ n_gestacoes        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ hist_diab_fam      <chr> "Sim", "Não", "Sim", "Sim", "Sim", "Não", "Sim", "Si...
## $ hist_diab_gest      <chr> "Não", "Não", "Não", "Não", "Não", "Não", "Não", "Nã...
## $ macrossomia_fetal   <chr> "Não", "Não", "Não", "Não", "Não", "Não", "Não", "Nã...
## $ tabagista           <chr> "Não", "Não", "Não", "Não", "Não", "Sim", "Não", "Nã...
## $ hac                 <chr> "Não", "Não", "Não", "Não", "Sim", "Não", "Não", "Nã...
## $ glicemia_jejum       <dbl> 95, 110, 114, 92, 114, 102, 96, 92, 92, 102, 100, 10...
## $ insulina            <chr> "Não", "Não", "Sim", "Não", "Sim", "Sim", "Não", "Nã...
```

Base 2

- Base de dados do **PNUD** cujas informações socioeconômicas são dos anos de 2012 a 2021 a nível Brasil, Unidade Federativa (UF) e região metropolitana.
- Contém **490 observações** e **5 variáveis**, a saber:
 - ano: ano de análise dos indicadores;
 - agregacao: nível nacional, estadual e região metropolitana;
 - nome: nome da UF e região metropolitana;
 - gini: Índice de Gini;
 - espvida: expectativa de vida, em anos.

- No :

Funções bases do

Frequência absoluta e relativa

Frequência absoluta

```
# frequencia absoluta da va 'historico de diabetes na familia' sem considerar na
freq_abs <- table(dados1$hist_diab_fam); freq_abs
```

```
##
## Não Sim
## 150 257
```

```
# frequencia absoluta da va 'historico de diabetes na familia' considerando na
freq_abs_na <- table(dados1$hist_diab_fam, useNA = "always"); freq_abs_na
```

```
##
## Não Sim <NA>
## 150 257     1
```

Frequênci a relativa

```
# frequênci a relativa da va 'hac' sem considerar na  
prop.table(freq_abs)
```

```
##  
##      Não      Sim  
## 0.3685504 0.6314496
```

```
# frequênci a relativa da va 'hac' considerando na  
prop.table(freq_abs_na)
```

```
##  
##      Não      Sim      <NA>  
## 0.36764706 0.62990196 0.00245098
```

Dica!

Use a função `round()` para arredondar os valores. Por exemplo, `round(prop.table(freq_abs), 2)`.

Medidas de posição

Valores mínimo e máximo

```
# valor minimo das vars 'idade' e 'expectativa de vida'  
min(dados1$idade); min(dados2$espvida)
```

```
## [1] 16
```

```
## [1] 66.34
```

```
# valor maximo da var 'expectativa de vida'  
max(dados1$idade); max(dados2$espvida)
```

```
## [1] 47
```

```
## [1] 81.22
```

Observação!

Se a variável tem NA, é necessário incluir o argumento na.rm = TRUE. Por exemplo, min(dados1\$idade, na.rm = TRUE).

Moda

```
# função para calcular moda
calc_moda <- function(x, na.rm = TRUE) {
  if (na.rm) {
    x <- x[!is.na(x)]
  }
  freq <- table(x)
  moda <- names(freq)[freq == max(freq)]
  return(type.convert(moda, as.is = TRUE))
}

# moda da va 'expectativa de vida'
calc_moda(dados2$espvida)

## [1] 72.03 74.16 74.20
```

Observação!

Nesse caso, não é necessário usar o argumento `na.rm = TRUE` porque os NAs da variável já são desconsiderados.

Mediana

```
# mediana das vas 'valor do exame de glicemia de jejum' e 'indice de gini'  
median(dados1$glicemia_jejum); median(dados2$gini)  
  
## [1] 96  
  
## [1] 0.525
```

Média

```
# media das vas 'idade' e 'indice de gini'  
mean(dados1$idade); mean(dados2$gini)  
  
## [1] 32.71814  
  
## [1] 0.5204918
```

Observação!

Se necessário, utilize o argumento na.rm = TRUE.

Quartis

```
# quartis padrao da va 'valor do exame de glicemia de jejum'  
quantile(dados1$glicemia_jejum)  
  
##    0%  25%  50%  75% 100%  
##    92    94    96   101   124
```

Percentis

```
# percentis 10, 20 e 90 da va 'valor do exame de glicemia de jejum'  
quantile(dados1$glicemia_jejum, probs = c(0.1, 0.2, 0.9))  
  
## 10% 20% 90%  
##  92  93 107
```

Observação!

Se necessário, utilize o argumento na.rm = TRUE.

Medidas de dispersão

Amplitude

```
# amplitude da va 'idade'  
max(dados1$idade) - min(dados1$idade)  
  
## [1] 31
```

Intervalo interquartil

```
# intervalo interquartil da va 'idade'  
IQR(dados1$idade)  
  
## [1] 8.25
```

Observação!

Se necessário, utilize o argumento na.rm = TRUE na função IQR().

Variância

```
# variância da va 'indice de gini'  
var(dados2$gini)  
  
## [1] 0.001634909
```

Desvio-padrão

```
# desvio-padrão da va 'indice de gini'  
sd(dados2$gini)  
  
## [1] 0.04043401
```

Coeficiente de variação (CV)

```
(sd(dados2$gini) / mean(dados2$gini)) * 100  
  
## [1] 7.768423
```

Observação!

Nas funções de variância e desvio-padrão, pode-se utilizar o argumento na.rm = TRUE.

Medidas-resumo

```
# medidas-resumo de todas as vars
summary(dados1[, c("idade", "n_gestacoes", "glicemia_jejum")])
```

```
##      idade      n_gestacoes      glicemia_jejum
##  Min.   :16.00   Min.   : 1.000   Min.   : 92.00
##  1st Qu.:28.75   1st Qu.: 1.000   1st Qu.: 94.00
##  Median :34.00   Median : 2.000   Median : 96.00
##  Mean   :32.72   Mean   : 2.799   Mean   : 98.25
##  3rd Qu.:37.00   3rd Qu.: 4.000   3rd Qu.:101.00
##  Max.   :47.00   Max.   :10.000   Max.   :124.00
```

```
# medidas-resumo da var 'idade'
summary(dados1$idade)
```

```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##  16.00  28.75  34.00  32.72  37.00  47.00
```

Tabelas descritivas

O pacote `{gtsummary}`

- Criado pelo [Daniel D. Sjoberg](#), bioestatístico do Memorial Sloan Kettering Cancer Center (EUA).
- Gera automaticamente tabelas com formatação elegante e profissional, prontas para relatórios, artigos científicos e apresentações.
- Infinitas possibilidades de customização!
- Repositório oficial: <https://github.com/ddsjoerg/gtsummary>.

Outros pacotes

- `{modelsummary}`: <https://modelsummary.com>.
- `{summarytools}`: <https://cran.r-project.org/web/packages/summarytools/vignettes/introduction.html>.

Tabela descritiva da Base 1

```
library(gtsummary)

# criando tabela gtsummary
dados1 |>
 tbl_summary() |>
  # ignorar (apenas para efeitos de visu
  as_gt() |>
  compacta_tabela()
```

Characteristic	N = 408 ¹
idade	34 (29, 37)
imc_classe	
Até normal	87 (21%)
Obeso	194 (48%)
Sobrepeso	127 (31%)
n_gestacoes	2.00 (1.00, 4.00)
hist_diab_fam	
Não	150 (37%)
Sim	257 (63%)
Unknown	1
hist_diab_gest	
Não	359 (88%)
Sim	49 (12%)
macrossomia_fetal	
Não	373 (91%)
Sim	35 (8.6%)
tabagista	
Não	371 (92%)
Sim	34 (8.4%)
Unknown	3
hac	
Não	295 (72%)
Sim	113 (28%)
glicemia_jejum	96 (94, 101)
insulina	
Não	273 (67%)
Sim	135 (33%)

¹ Median (Q1, Q3); n (%)

```
# incluindo tema a tabela
theme_gtsummary_journal(journal = "lancet")

dados1 |>
 tbl_summary(
  # incluindo apenas as variaveis de interesse
  include = c(idade, imc_classe, n_gestacoes, hist_diab_fam, insulina, glicemia_jejum)
  )

```

Characteristic	N = 408 ¹
idade	34 (29 – 37)
imc_classe	
Até normal	87 (21%)
Obeso	194 (48%)
Sobrepeso	127 (31%)
n_gestacoes	2·00 (1·00 – 4·00)
hist_diab_fam	
Não	150 (37%)
Sim	257 (63%)
Unknown	1
glicemia_jejum	96 (94 – 101)
insulina	
Não	273 (67%)
Sim	135 (33%)

¹ Median (IQR); n (%)

```

dados1 |>
 tbl_summary(
  # incluindo apenas as variaveis de interesse
  include = c(idade, imc_classe, n_gestacoes,
  # rotulando as variaveis
  label = list(
    idade ~ "Idade (anos)",
    imc_classe ~ "IMC",
    n_gestacoes ~ "Nº de gestações anteriores",
    hist_diab_fam ~ "Histórico de diabetes na família",
    glicemia_jejun ~ "Valor do exame de glicemia (mg/dL)",
    insulina ~ "Usou insulina antes do parto"
  )))

```

Characteristic	N = 408 ¹
Idade (anos)	34 (29 – 37)
IMC	
Até normal	87 (21%)
Obeso	194 (48%)
Sobrepeso	127 (31%)
Nº de gestações anteriores	2·00 (1·00 – 4·00)
Histórico de diabetes na família	
Não	150 (37%)
Sim	257 (63%)
Unknown	1
Valor do exame de glicemia (mg/dL)	96 (94 – 101)
Usou insulina antes do parto	
Não	273 (67%)
Sim	135 (33%)

¹ Median (IQR); n (%)

```

dados1 |>
 tbl_summary(
  # incluindo apenas as variáveis de interesse
  include = c(idade, imc_classe, n_gestacoes,
  # rotulando as variáveis
  label = list(
    idade ~ "Idade (anos)",
    imc_classe ~ "IMC",
    n_gestacoes ~ "Nº de gestações anteriores",
    hist_diab_fam ~ "Histórico de diabetes na família",
    glicemia_jejun ~ "Valor do exame de glicemia (mg/dL)",
    insulina ~ "Usou insulina antes do parto"),
  # calculando as medidas de interesse
  statistic = list(
    all_continuous() ~ "{mean} ± {sd}",
    all_categorical() ~ "{n} ({p}%)"
  )
)

```

Characteristic	N = 408 ¹
Idade (anos)	33 ± 6
IMC	
Até normal	87 (21%)
Obeso	194 (48%)
Sobrepeso	127 (31%)
Nº de gestações anteriores	2.80 ± 1.69
Histórico de diabetes na família	
Não	150 (37%)
Sim	257 (63%)
Unknown	1
Valor do exame de glicemia (mg/dL)	98 ± 7
Usou insulina antes do parto	
Não	273 (67%)
Sim	135 (33%)

¹ Mean ± SD; n (%)

```

dados1 |>
 tbl_summary(
  # incluindo apenas as variaveis de interesse
  include = c(idade, imc_classe, n_gestacoes,
  # rotulando as variaveis
  label = list(
    idade ~ "Idade (anos)",
    imc_classe ~ "IMC",
    n_gestacoes ~ "Nº de gestações anteriores",
    hist_diab_fam ~ "Histórico de diabetes na família",
    glicemia_jejun ~ "Valor do exame de glicemia",
    insulina ~ "Usou insulina antes do parto"),
  # dispondo cada medida em uma linha
  type = all_continuous() ~ "continuous",
  # calculando as medidas de interesse
  statistic = list(
    all_continuous() ~ c(
      "{mean} ± {sd}", # media e dp para medidas continuas
      "{median} ({p25}, {p75})", # mediana e quartis para medidas continuas
      "{min}, {max}" # minimo e maximo para medidas continuas
    ),
    all_categorical() ~ "{n} ({p}%)"
  )
) |>
# ignorar (apenas para efeitos de visualização)
as_gt() |>

```

Characteristic	N = 408 ¹
Idade (anos)	
Mean ± SD	33 ± 6
Median (Q1, Q3)	34 (29, 37)
Min, Max	16, 47
IMC	
Até normal	87 (21%)
Obeso	194 (48%)
Sobrepeso	127 (31%)
Nº de gestações anteriores	
Mean ± SD	2.80 ± 1.69
Median (Q1, Q3)	2.00 (1.00, 4.00)
Min, Max	1.00, 10.00
Histórico de diabetes na família	
Não	150 (37%)
Sim	257 (63%)
Unknown	1
Valor do exame de glicemia (mg/dL)	
Mean ± SD	98 ± 7
Median (Q1, Q3)	96 (94, 101)
Min, Max	92, 124
Usou insulina antes do parto	
Não	273 (67%)
Sim	135 (33%)

¹ n (%)

```

dados1 |>
 tbl_summary(
  # incluindo apenas as variaveis de interesse
  include = c(idade, imc_classe, n_gestacoes,
  # rotulando as variaveis
  label = list(
    idade ~ "Idade (anos)",
    imc_classe ~ "IMC",
    n_gestacoes ~ "Nº de gestações anteriores",
    hist_diab_fam ~ "Histórico de diabetes na família",
    glicemia_jejun ~ "Valor do exame de glicemia",
    insulina ~ "Usou insulina antes do parto"),
  # dispondo cada medida em uma linha
  type = all_continuous() ~ "continuous",
  # calculando as medidas de interesse
  statistic = list(
    all_continuous() ~ c(
      "{mean} ± {sd}", # media e dp p/ contínuas
      "{median} ({p25}, {p75})", # mediana e quartis
      "{min}, {max}" # minimo e maximo
    ),
    all_categorical() ~ "{n} ({p}%)",
    # formatando numero de digitos
    digits = list(
      all_continuous() ~ 2, # 2 digitos
      all_categorical() ~ 0
    )
  )
)
# formatando numero de digitos

```

Characteristic	N = 408 ¹
Idade (anos)	
Mean ± SD	32.72 ± 6.06
Median (Q1, Q3)	34.00 (28.50, 37.00)
Min, Max	16.00, 47.00
IMC	
Até normal	87 (21.3%)
Obeso	194 (47.5%)
Sobrepeso	127 (31.1%)
Nº de gestações anteriores	
Mean ± SD	2.80 ± 1.69
Median (Q1, Q3)	2.00 (1.00, 4.00)
Min, Max	1.00, 10.00
Histórico de diabetes na família	
Não	150 (36.9%)
Sim	257 (63.1%)
Unknown	1
Valor do exame de glicemia (mg/dL)	
Mean ± SD	98.25 ± 6.53
Median (Q1, Q3)	96.00 (94.00, 101.00)
Min, Max	92.00, 124.00
Usou insulina antes do parto	
Não	273 (66.9%)
Sim	135 (33.1%)

¹ n (%)

```

dados1 |>
 tbl_summary(
  # incluindo apenas as variaveis de interesse
  include = c(idade, imc_classe, n_gestacoes,
  # rotulando as variaveis
  label = list(
    idade ~ "Idade (anos)",
    imc_classe ~ "IMC",
    n_gestacoes ~ "Nº de gestações anteriores",
    hist_diab_fam ~ "Histórico de diabetes na família",
    glicemia_jejun ~ "Valor do exame de glicemia",
    insulina ~ "Usou insulina antes do parto"),
  # dispondo cada medida em uma linha
  type = all_continuous() ~ "continuous",
  # calculando as medidas de interesse
  statistic = list(
    all_continuous() ~ c(
      "{mean} ± {sd}", # media e dp p/ medidas continuas
      "{median} ({p25}, {p75})", # mediana e quartis
      "{min}, {max}" # minimo e maximo
    ),
    all_categorical() ~ "{n} ({p}%)",
    # formatando numero de digitos
    digits = list(
      all_continuous() ~ 2, # 2 digitos
      all_categorical() ~ 0
    )
  )
)
# formatando numero de digitos

```

Characteristic	N = 408 ¹
Idade (anos)	
Mean ± SD	32.72 ± 6.06
Median (Q1, Q3)	34.00 (28.50, 37.00)
Min, Max	16.00, 47.00
IMC	
Até normal	87 (21.3%)
Obeso	194 (47.5%)
Sobrepeso	127 (31.1%)
Nº de gestações anteriores	
Mean ± SD	2.80 ± 1.69
Median (Q1, Q3)	2.00 (1.00, 4.00)
Min, Max	1.00, 10.00
Histórico de diabetes na família	
Não	150 (36.9%)
Sim	257 (63.1%)
NA	1
Valor do exame de glicemia (mg/dL)	
Mean ± SD	98.25 ± 6.53
Median (Q1, Q3)	96.00 (94.00, 101.00)
Min, Max	92.00, 124.00
Usou insulina antes do parto	
Não	273 (66.9%)
Sim	135 (33.1%)

¹ n (%)

```

dados1 |>
 tbl_summary(
  # incluindo apenas as variaveis de interesse
  include = c(idade, imc_classe, n_gestacoes,
  # rotulando as variaveis
  label = list(
    idade ~ "Idade (anos)",
    imc_classe ~ "IMC",
    n_gestacoes ~ "Nº de gestações anteriores",
    hist_diab_fam ~ "Histórico de diabetes na família",
    glicemia_jejun ~ "Valor do exame de glicemia",
    insulina ~ "Usou insulina antes do parto"),
  # dispondo cada medida em uma linha
  type = all_continuous() ~ "continuous",
  # calculando as medidas de interesse
  statistic = list(
    all_continuous() ~ c(
      "{mean} ± {sd}", # media e dp para medidas continuas
      "{median} ({p25}, {p75})", # mediana e quartis para medidas continuas
      "{min}, {max}" # minimo e maximo para medidas continuas
    ),
    all_categorical() ~ "{n} ({p}%)",
  ),
  # formatando numero de digitos
  digits = list(
    all_continuous() ~ 2, # 2 digitos
  )
)

```

Characteristic	N = 408 [95% CI] ¹
Idade (anos)	[32, 33]
Mean ± SD	32.72 ± 6.06
Median (Q1, Q3)	34.00 (28.50, 37.00)
Min, Max	16.00, 47.00
IMC	
Até normal	87 (21.3%) [18%, 26%]
Obeso	194 (47.5%) [43%, 53%]
Sobrepeso	127 (31.1%) [27%, 36%]
Nº de gestações anteriores	[2.6, 3.0]
Mean ± SD	2.80 ± 1.69
Median (Q1, Q3)	2.00 (1.00, 4.00)
Min, Max	1.00, 10.00
Histórico de diabetes na família	
Não	150 (36.9%) [32%, 42%]
Sim	257 (63.1%) [58%, 68%]
NA	1
Valor do exame de glicemia (mg/dL)	[98, 99]
Mean ± SD	98.25 ± 6.53
Median (Q1, Q3)	96.00 (94.00, 101.00)
Min, Max	92.00, 124.00
Usou insulina antes do parto	
Não	273 (66.9%) [62%, 71%]
Sim	135 (33.1%) [29%, 38%]

¹ n (%)
Abbreviation: CI = Confidence Interval

```

dados1 |>
 tbl_summary(
  # incluindo apenas as variaveis de interesse
  include = c(idade, imc_classe, n_gestacoes,
  # rotulando as variaveis
  label = list(
    idade ~ "Idade (anos)",
    imc_classe ~ "IMC",
    n_gestacoes ~ "Nº de gestações anteriores",
    hist_diab_fam ~ "Histórico de diabetes na família",
    glicemia_jejun ~ "Valor do exame de glicemia",
    insulina ~ "Usou insulina antes do parto"),
  # dispondo cada medida em uma linha
  type = all_continuous() ~ "continuous",
  # calculando as medidas de interesse
  statistic = list(
    all_continuous() ~ c(
      "{mean} ± {sd}", # media e dp para medidas contínuas
      "{median} ({p25}, {p75})", # mediana e quartis para medidas contínuas
      "{min}, {max}" # minimo e maximo para medidas contínuas
    ),
    all_categorical() ~ "{n} ({p}%)",
  ),
  # formatando numero de dígitos
  digits = list(
    all_continuous() ~ 2, # 2 dígitos
  )
)

```

Characteristic	N	N = 408 [95% CI] ¹
Idade (anos)	408	[32, 33]
Mean ± SD		32.72 ± 6.06
Median (Q1, Q3)		34.00 (28.50, 37.00)
Min, Max		16.00, 47.00
IMC	408	
Até normal		87 (21.3%) [18%, 26%]
Obeso		194 (47.5%) [43%, 53%]
Sobrepeso		127 (31.1%) [27%, 36%]
Nº de gestações anteriores	408	[2.6, 3.0]
Mean ± SD		2.80 ± 1.69
Median (Q1, Q3)		2.00 (1.00, 4.00)
Min, Max		1.00, 10.00
Histórico de diabetes na família	407	
Não		150 (36.9%) [32%, 42%]
Sim		257 (63.1%) [58%, 68%]
NA		1
Valor do exame de glicemia (mg/dL)	408	[98, 99]
Mean ± SD		98.25 ± 6.53
Median (Q1, Q3)		96.00 (94.00, 101.00)
Min, Max		92.00, 124.00
Usou insulina antes do parto	408	
Não		273 (66.9%) [62%, 71%]
Sim		135 (33.1%) [29%, 38%]

¹ n (%)

Abbreviation: CI = Confidence Interval

```

dados1 |>
 tbl_summary(
  # incluindo apenas as variaveis de interesse
  include = c(idade, imc_classe, n_gestacoes,
  # rotulando as variaveis
  label = list(
    idade ~ "Idade (anos)",
    imc_classe ~ "IMC",
    n_gestacoes ~ "Nº de gestações anteriores",
    hist_diab_fam ~ "Histórico de diabetes na família",
    glicemia_jejun ~ "Valor do exame de glicemia",
    insulina ~ "Usou insulina antes do parto"),
  # dispondo cada medida em uma linha
  type = all_continuous() ~ "continuous",
  # calculando as medidas de interesse
  statistic = list(
    all_continuous() ~ c(
      "{mean} ± {sd}", # media e dp para medidas continuas
      "{median} ({p25}, {p75})", # mediana e quartis para medidas continuas
      "{min}, {max}" # minimo e maximo para medidas continuas
    ),
    all_categorical() ~ "{n} ({p}%)",
  ),
  # formatando numero de digitos
  digits = list(
    all_continuous() ~ 2, # 2 digitos
  )
)

```

Variável	N	N = 408 [95% CI] ¹
Idade (anos)	408	[32, 33]
Mean ± SD		32.72 ± 6.06
Median (Q1, Q3)		34.00 (28.50, 37.00)
Min, Max		16.00, 47.00
IMC	408	
Até normal		87 (21.3%) [18%, 26%]
Obeso		194 (47.5%) [43%, 53%]
Sobrepeso		127 (31.1%) [27%, 36%]
Nº de gestações anteriores	408	[2.6, 3.0]
Mean ± SD		2.80 ± 1.69
Median (Q1, Q3)		2.00 (1.00, 4.00)
Min, Max		1.00, 10.00
Histórico de diabetes na família	407	
Não		150 (36.9%) [32%, 42%]
Sim		257 (63.1%) [58%, 68%]
NA		1
Valor do exame de glicemia (mg/dL)	408	[98, 99]
Mean ± SD		98.25 ± 6.53
Median (Q1, Q3)		96.00 (94.00, 101.00)
Min, Max		92.00, 124.00
Usou insulina antes do parto	408	
Não		273 (66.9%) [62%, 71%]
Sim		135 (33.1%) [29%, 38%]

¹ n (%)

Abbreviation: CI = Confidence Interval

```

dados1 |>
 tbl_summary(
  # incluindo apenas as variaveis de interesse
  include = c(idade, imc_classe, n_gestacoes,
  # rotulando as variaveis
  label = list(
    idade ~ "Idade (anos)",
    imc_classe ~ "IMC",
    n_gestacoes ~ "Nº de gestações anteriores",
    hist_diab_fam ~ "Histórico de diabetes na família",
    glicemia_jejun ~ "Valor do exame de glicemia",
    insulina ~ "Usou insulina antes do parto"),
  # dispondo cada medida em uma linha
  type = all_continuous() ~ "continuous",
  # calculando as medidas de interesse
  statistic = list(
    all_continuous() ~ c(
      "{mean} ± {sd}", # media e dp p/ medidas continuas
      "{median} ({p25}, {p75})", # mediana e quartis
      "{min}, {max}" # minimo e maximo
    ),
    all_categorical() ~ "{n} ({p}%)",
  ),
  # formatando numero de digitos
  digits = list(
    all_continuous() ~ 2, # 2 digitos
  )
)

```

Variável	N	N = 408 [95% CI] ¹
Idade (anos)	408	[32, 33]
Mean ± SD		32.72 ± 6.06
Median (Q1, Q3)		34.00 (28.50, 37.00)
Min, Max		16.00, 47.00
IMC	408	
Até normal		87 (21.3%) [18%, 26%]
Obeso		194 (47.5%) [43%, 53%]
Sobrepeso		127 (31.1%) [27%, 36%]
Nº de gestações anteriores	408	[2.6, 3.0]
Mean ± SD		2.80 ± 1.69
Median (Q1, Q3)		2.00 (1.00, 4.00)
Min, Max		1.00, 10.00
Histórico de diabetes na família	407	
Não		150 (36.9%) [32%, 42%]
Sim		257 (63.1%) [58%, 68%]
NA		1
Valor do exame de glicemia (mg/dL)	408	[98, 99]
Mean ± SD		98.25 ± 6.53
Median (Q1, Q3)		96.00 (94.00, 101.00)
Min, Max		92.00, 124.00
Usou insulina antes do parto	408	
Não		273 (66.9%) [62%, 71%]
Sim		135 (33.1%) [29%, 38%]

¹ n (%)

Abbreviation: CI = Confidence Interval

```

dados1 |>
  tbl_summary(
    # analisando por grupo de desfecho
    by = insulina,
    # incluindo apenas as variaveis de interesse
    include = c(idade, imc_classe, n_gestacoes,
    # rotulando as variaveis
    label = list(
      idade ~ "Idade (anos)",
      imc_classe ~ "IMC",
      n_gestacoes ~ "Nº de gestações anteriores",
      hist_diab_fam ~ "Histórico de diabetes na família",
      glicemia_jejun ~ "Valor do exame de glicemia (mg/dL)"),
    # dispondo cada medida em uma linha
    type = all_continuous() ~ "continuous",
    # calculando as medidas de interesse
    statistic = list(
      all_continuous() ~ c(
        "{mean} ± {sd}", # media e dp para medidas continuas
        "{median} ({p25}, {p75})", # mediana e quartis para medidas continuas
        "{min}, {max}" # minimo e maximo para medidas continuas),
      all_categorical() ~ "{n} ({p}%)",
      # formatando numero de digitos
      digits = list(

```

Variável	N	Não N = 273 [95% CI] ¹	Sim N = 135 [95% CI] ¹
Idade (anos)	408	[31, 33]	[33, 35]
Mean ± SD		32.03 ± 6.12	34.12 ± 5.72
Median (Q1, Q3)		32.00 (28.00, 37.00)	34.00 (31.00, 38.00)
Min, Max		16.00, 46.00	18.00, 47.00
IMC	408		
Até normal		71 (26.0%) [21%, 32%]	16 (11.9%) [7.1%, 19%]
Obeso		114 (41.8%) [36%, 48%]	80 (59.3%) [50%, 68%]
Sobrepeso		88 (32.2%) [27%, 38%]	39 (28.9%) [22%, 37%]
Nº de gestações anteriores	408	[2.4, 2.8]	[2.9, 3.5]
Mean ± SD		2.61 ± 1.56	3.19 ± 1.87
Median (Q1, Q3)		2.00 (1.00, 3.00)	3.00 (2.00, 4.00)
Min, Max		1.00, 8.00	1.00, 10.00
Histórico de diabetes na família	407		
Não		113 (41.5%) [36%, 48%]	37 (27.4%) [20%, 36%]
Sim		159 (58.5%) [52%, 64%]	98 (72.6%) [64%, 80%]
NA		1	0
Valor do exame de glicemia (mg/dL)	408	[96, 97]	[100, 102]
Mean ± SD		96.81 ± 5.24	101.16 ± 7.81
Median (Q1, Q3)		95.00 (93.00, 99.00)	99.00 (95.00, 106.00)
Min, Max		92.00, 121.00	92.00, 124.00

¹ n (%)

Abbreviation: CI = Confidence Interval

```

dados1 |>
 tbl_summary(
  # analisando por grupo de desfecho
  by = insulina,
  # incluindo apenas as variaveis de interesse
  include = c(idade, imc_classe, n_gestacoes,
  # rotulando as variaveis
  label = list(
    idade ~ "Idade (anos)",
    imc_classe ~ "IMC",
    n_gestacoes ~ "Nº de gestações anteriores",
    hist_diab_fam ~ "Histórico de diabetes na família",
    glicemia_jejun ~ "Valor do exame de glicemia (mg/dL)"),
  # dispondo cada medida em uma linha
  type = all_continuous() ~ "continuous",
  # calculando as medidas de interesse
  statistic = list(
    all_continuous() ~ c(
      "{mean} ± {sd}", # media e dp para medidas continuas
      "{median} ({p25}, {p75})", # mediana e quartis para medidas continuas
      "{min}, {max}" # minimo e maximo para medidas continuas),
    all_categorical() ~ "{n} ({p}%)",
  ),
  # formatando numero de digitos
  digits = list(

```

Variável	N	Usou insulina antes do parto	
		Não N = 273 [95% CI] ¹	Sim N = 135 [95% CI] ¹
Idade (anos)	408	[31, 33]	[33, 35]
Mean ± SD		32.03 ± 6.12	34.12 ± 5.72
Median (Q1, Q3)		32.00 (28.00, 37.00)	34.00 (31.00, 38.00)
Min, Max		16.00, 46.00	18.00, 47.00
IMC	408		
Até normal		71 (26.0%) [21%, 32%]	16 (11.9%) [7.1%, 19%]
Obeso		114 (41.8%) [36%, 48%]	80 (59.3%) [50%, 68%]
Sobrepeso		88 (32.2%) [27%, 38%]	39 (28.9%) [22%, 37%]
Nº de gestações anteriores	408	[2.4, 2.8]	[2.9, 3.5]
Mean ± SD		2.61 ± 1.56	3.19 ± 1.87
Median (Q1, Q3)		2.00 (1.00, 3.00)	3.00 (2.00, 4.00)
Min, Max		1.00, 8.00	1.00, 10.00
Histórico de diabetes na família	407		
Não		113 (41.5%) [36%, 48%]	37 (27.4%) [20%, 36%]
Sim		159 (58.5%) [52%, 64%]	98 (72.6%) [64%, 80%]
NA		1	0
Valor do exame de glicemia (mg/dL)	408	[96, 97]	[100, 102]
Mean ± SD		96.81 ± 5.24	101.16 ± 7.81
Median (Q1, Q3)		95.00 (93.00, 99.00)	99.00 (95.00, 106.00)
Min, Max		92.00, 121.00	92.00, 124.00

¹ n (%)

Abbreviation: CI = Confidence Interval

```

# traduzindo a tabela para pt-br
theme_gtsummary_language("pt", big.mark = ",")

dados1 |>
 tbl_summary(
    # analisando por grupo de desfecho
    by = insulina,
    # incluindo apenas as variaveis de interesse
    include = c(idade, imc_classe, n_gestacoes,
    # rotulando as variaveis
    label = list(
      idade ~ "Idade (anos)",
      imc_classe ~ "IMC",
      n_gestacoes ~ "Nº de gestações anteriores",
      hist_diab_fam ~ "Histórico de diabetes na família",
      glicemia_jejun ~ "Valor do exame de glicemia (mg/dL)"),
    # dispondo cada medida em uma linha
    type = all_continuous() ~ "continuous",
    # calculando as medidas de interesse
    statistic = list(
      all_continuous() ~ c(
        "{mean} ± {sd}", # media e dp para medidas contínuas
        "{median} ({p25}, {p75})", # mediana e quartis
        "{min}, {max}" # minimo e maximo
      ),
      all_categorical() ~ "{n} ({p}%)"
    )
  )

```

Variável	N	Usou insulina antes do parto	
		Não N = 273 [95% CI] ¹	Sim N = 135 [95% CI] ¹
Idade (anos)	408	[31, 33]	[33, 35]
Média ± Desvio Padrão		32,03 ± 6,12	34,12 ± 5,72
Mediana (Q1, Q3)		32,00 (28,00, 37,00)	34,00 (31,00, 38,00)
Min, Max		16,00, 46,00	18,00, 47,00
IMC	408		
Até normal		71 (26,0%) [21%, 32%]	16 (11,9%) [7,1%, 19%]
Obeso		114 (41,8%) [36%, 48%]	80 (59,3%) [50%, 68%]
Sobrepeso		88 (32,2%) [27%, 38%]	39 (28,9%) [22%, 37%]
Nº de gestações anteriores	408	[2,4, 2,8]	[2,9, 3,5]
Média ± Desvio Padrão		2,61 ± 1,56	3,19 ± 1,87
Mediana (Q1, Q3)		2,00 (1,00, 3,00)	3,00 (2,00, 4,00)
Min, Max		1,00, 8,00	1,00, 10,00
Histórico de diabetes na família	407		
Não		113 (41,5%) [36%, 48%]	37 (27,4%) [20%, 36%]
Sim		159 (58,5%) [52%, 64%]	98 (72,6%) [64%, 80%]
NA		1	0
Valor do exame de glicemia (mg/dL)	408	[96, 97]	[100, 102]
Média ± Desvio Padrão		96,81 ± 5,24	101,16 ± 7,81
Mediana (Q1, Q3)		95,00 (93,00, 99,00)	99,00 (95,00, 106,00)
Min, Max		92,00, 121,00	92,00, 124,00

¹ n (%)

Abreviação: IC = Intervalo de Confiança

```

dados1 |>
  tbl_summary(
    # analisando por grupo de desfecho
    by = insulina,
    # incluindo apenas as variaveis de interesse
    include = c(idade, imc_classe, n_gestacoes),
    # rotulando as variaveis
    label = list(
      idade ~ "Idade (anos)",
      imc_classe ~ "IMC",
      n_gestacoes ~ "Nº de gestações anteriores",
      hist_diab_fam ~ "Histórico de diabetes na família",
      glicemia_jejun ~ "Valor do exame de glicemia (mg/dL)"
    ),
    # dispondo cada medida em uma linha
    type = all_continuous() ~ "continuous",
    # calculando as medidas de interesse
    statistic = list(
      all_continuous() ~ c(
        "{mean} ± {sd}", # media e dp para medidas continuas
        "{median} ({p25}, {p75})", # mediana e intervalo interquartil para medidas continuas
        "{min}, {max}" # minimo e maximo para medidas continuas
      ),
      all_categorical() ~ "{n} ({p}%)"
    ),
    # formatando numero de digitos
    digits = list(
      idade = 1,
      imc_classe = 0,
      n_gestacoes = 0,
      hist_diab_fam = 0,
      glicemia_jejun = 2
    )
  )

```

Variável	Usou insulina antes do parto		
	N	Não N = 273 [95% CI] ^{1,2}	Sim N = 135 [95% CI] ^{1,2}
Idade (anos)	408	[31, 33]	[33, 35]
Média ± Desvio Padrão		32,03 ± 6,12	34,12 ± 5,72
Mediana (Q1, Q3)		32,00 (28,00, 37,00)	34,00 (31,00, 38,00)
Min, Max		16,00, 46,00	18,00, 47,00
IMC	408		
Até normal		71 (26,0%) [21%, 32%]	16 (11,9%) [7,1%, 19%]
Obeso		114 (41,8%) [36%, 48%]	80 (59,3%) [50%, 68%]
Sobrepeso		88 (32,2%) [27%, 38%]	39 (28,9%) [22%, 37%]
Nº de gestações anteriores	408	[2,4, 2,8]	[2,9, 3,5]
Média ± Desvio Padrão		2,61 ± 1,56	3,19 ± 1,87
Mediana (Q1, Q3)		2,00 (1,00, 3,00)	3,00 (2,00, 4,00)
Min, Max		1,00, 8,00	1,00, 10,00
Histórico de diabetes na família	407		
Não		113 (41,5%) [36%, 48%]	37 (27,4%) [20%, 36%]
Sim		159 (58,5%) [52%, 64%]	98 (72,6%) [64%, 80%]
NA		1	0
Valor do exame de glicemia (mg/dL)	408	[96, 97]	[100, 102]
Média ± Desvio Padrão		96,81 ± 5,24	101,16 ± 7,81
Mediana (Q1, Q3)		95,00 (93,00, 99,00)	99,00 (95,00, 106,00)
Min, Max		92,00, 121,00	92,00, 124,00

¹ n (%)

² Gestantes que realizaram o pré-natal entre os anos de 2012 e 2015.

Abreviação: IC = Intervalo de Confiança

```

dados1 |>
  tbl_summary(
    # analisando por grupo de desfecho
    by = insulina,
    # incluindo apenas as variaveis de interesse
    include = c(idade, imc_classe, n_gestacoes,
    # rotulando as variaveis
    label = list(
      idade ~ "Idade (anos)",
      imc_classe ~ "IMC",
      n_gestacoes ~ "Nº de gestações anteriores",
      hist_diab_fam ~ "Histórico de diabetes na família",
      glicemia_jejun ~ "Valor do exame de glicemia (mg/dL)"
    )),
    # dispondo cada medida em uma linha
    type = all_continuous() ~ "continuous",
    # calculando as medidas de interesse
    statistic = list(
      all_continuous() ~ c(
        "{mean} ± {sd}", # media e dp para medidas continuas
        "{median} ({p25}, {p75})", # mediana e intervalo interquartil para medidas continuas
        "{min}, {max}" # minimo e maximo para medidas continuas
      ),
      all_categorical() ~ "{n} ({p}%)", # para medidas categoricas
    ),
    # formatando numero de digitos
    digits = list(
  
```

Variável	Usou insulina antes do parto		
	N	Não N = 273 [95% CI] ^{1,2}	Sim N = 135 [95% CI] ^{1,2}
Idade (anos)	408	[31, 33] 32,03 ± 6,12 Mediana (Q1, Q3) 32,00 (28,00, 37,00) Min, Max 16,00, 46,00	[33, 35] 34,12 ± 5,72 34,00 (31,00, 38,00) 18,00, 47,00
IMC	408	Até normal 71 (26,0%) [21%, 32%] Obeso 114 (41,8%) [36%, 48%] Sobrepeso 88 (32,2%) [27%, 38%]	16 (11,9%) [7,1%, 19%] 80 (59,3%) [50%, 68%] 39 (28,9%) [22%, 37%]
Nº de gestações anteriores	408	Média ± Desvio Padrão 2,61 ± 1,56 Mediana (Q1, Q3) 2,00 (1,00, 3,00) Min, Max 1,00, 8,00	[2,4, 2,8] 3,19 ± 1,87 3,00 (2,00, 4,00) 1,00, 10,00
Histórico de diabetes na família³	407	Não 113 (41,5%) [36%, 48%] Sim 159 (58,5%) [52%, 64%] NA 1	37 (27,4%) [20%, 36%] 98 (72,6%) [64%, 80%] 0
Valor do exame de glicemia (mg/dL)	408	Média ± Desvio Padrão 96,81 ± 5,24 Mediana (Q1, Q3) 95,00 (93,00, 99,00) Min, Max 92,00, 121,00	[96, 97] 101,16 ± 7,81 99,00 (95,00, 106,00) 92,00, 124,00

¹ n (%)
² Gestantes que realizaram o pré-natal entre os anos de 2012 e 2015.
³ Parentes de primeiro grau.

Abreviação: IC = Intervalo de Confiança

```

tbl <- dados1 |>
 tbl_summary(
  # analisando por grupo de desfecho
  by = insulina,
  # incluindo apenas as variaveis de interesse
  include = c(idade, imc_classe, n_gestacoes,
  # rotulando as variaveis
  label = list(
    idade ~ "Idade (anos)",
    imc_classe ~ "IMC",
    n_gestacoes ~ "Nº de gestações anteriores",
    hist_diab_fam ~ "Histórico de diabetes na família",
    glicemia_jejun ~ "Valor do exame de glicemia (mg/dL)"
  ),
  # dispondo cada medida em uma linha
  type = all_continuous() ~ "continuous",
  # calculando as medidas de interesse
  statistic = list(
    all_continuous() ~ c(
      "{mean} ± {sd}", # media e dp para medidas continuas
      "{median} ({p25}, {p75})", # mediana e intervalo interquartil
      "{min}, {max}" # minimo e maximo
    ),
    all_categorical() ~ "{n} ({p}%)"
  ),
  # formatando numero de digitos
  digits = list(
    ...
  )
)
  
```

Variável	Usou insulina antes do parto		
	N	Não N = 273 [95% CI] ^{1,2}	Sim N = 135 [95% CI] ^{1,2}
Idade (anos)	408	[31, 33]	[33, 35]
Média ± Desvio Padrão		32,03 ± 6,12	34,12 ± 5,72
Mediana (Q1, Q3)		32,00 (28,00, 37,00)	34,00 (31,00, 38,00)
Min, Max		16,00, 46,00	18,00, 47,00
IMC	408		
Até normal		71 (26,0%) [21%, 32%]	16 (11,9%) [7,1%, 19%]
Obeso		114 (41,8%) [36%, 48%]	80 (59,3%) [50%, 68%]
Sobrepeso		88 (32,2%) [27%, 38%]	39 (28,9%) [22%, 37%]
Nº de gestações anteriores	408	[2,4, 2,8]	[2,9, 3,5]
Média ± Desvio Padrão		2,61 ± 1,56	3,19 ± 1,87
Mediana (Q1, Q3)		2,00 (1,00, 3,00)	3,00 (2,00, 4,00)
Min, Max		1,00, 8,00	1,00, 10,00
Histórico de diabetes na família³	407		
Não		113 (41,5%) [36%, 48%]	37 (27,4%) [20%, 36%]
Sim		159 (58,5%) [52%, 64%]	98 (72,6%) [64%, 80%]
NA		1	0
Valor do exame de glicemia (mg/dL)	408	[96, 97]	[100, 102]
Média ± Desvio Padrão		96,81 ± 5,24	101,16 ± 7,81
Mediana (Q1, Q3)		95,00 (93,00, 99,00)	99,00 (95,00, 106,00)
Min, Max		92,00, 121,00	92,00, 124,00

¹ n (%)

² Gestantes que realizaram o pré-natal entre os anos de 2012 e 2015.

³ Parentes de primeiro grau.

Abreviação: IC = Intervalo de Confiança

Fonte de dados: ambulatório de diabetes gestacional do HCFMUSP.

Salvando uma tabela gtsummary no R

- A tabela precisa ser do tipo gt. Para transformá-la nesse formato, use a função `as_gt()`, do `gtsummary`.
- Para salvar a tabela, execute a função `gtsave()`, do pacote `gt`.

```
# carregando o pacote {gt}
library(gt)

# salvando tabela no formato png
gtsave(tbl, "tabelas/tbl_desc.png")

# salvando tabela no formato docx
gtsave(tbl, "tabelas/tbl_desc.docx")
```

Gráficos

No , é possível fazer...

(outras representações gráficas)

... gráfico animado;

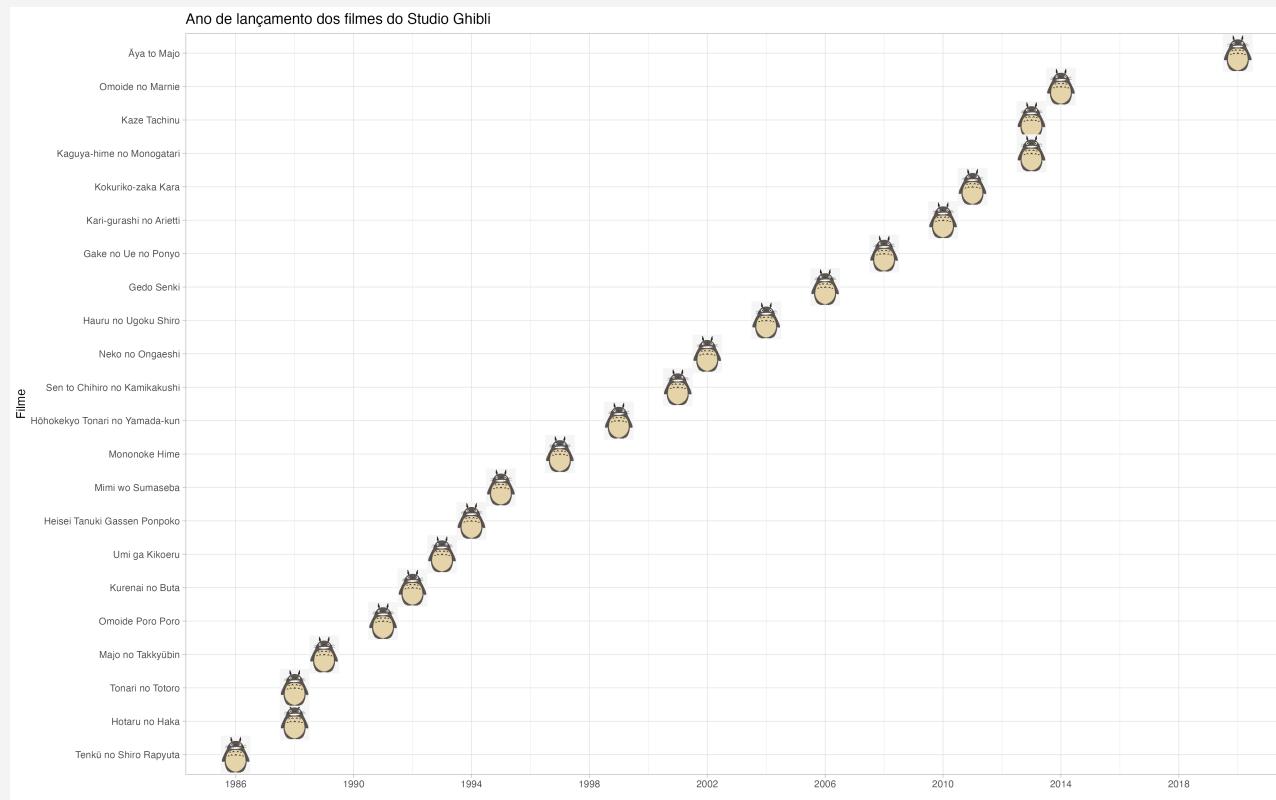


Figura 1: Gráfico animado do ano de lançamento dos filmes do Studio Ghibli.

... gráfico de densidades;

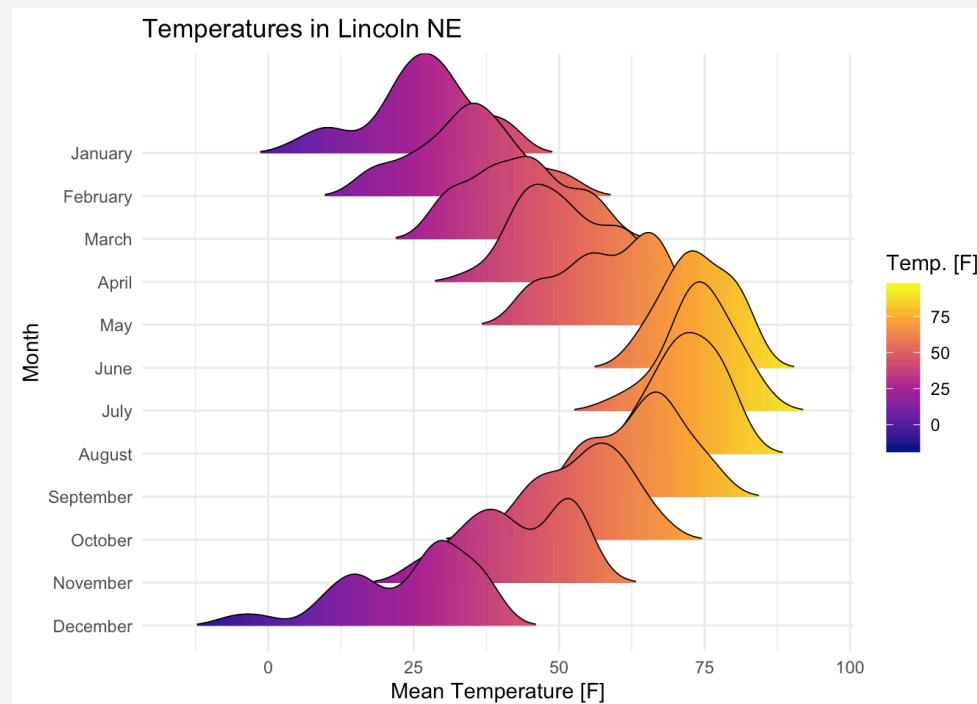


Figura 2: Gráfico das distribuições de densidade da temperatura, por mês do ano de 2016, na cidade de Lincoln, em Nebraska/EUA.

Fonte: [Datanovia | Elegant visualization of density distribution in R using ridgeline.](#)

... nuvem de palavras;



Figura 3: Nuvem de palavras das músicas da Mukeka di Rato.

... gráfico de florestas (*forest plot*);

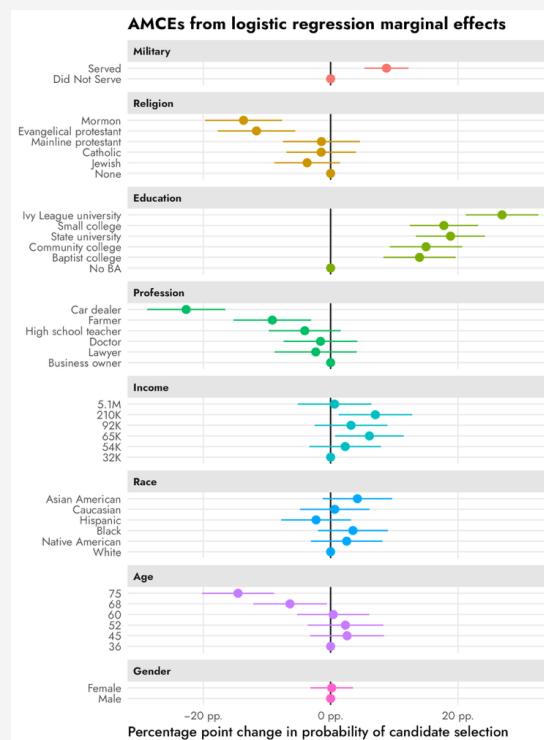


Figura 4: Gráfico de floresta dos efeitos marginais médios entre candidatos, segundo condições sociodemográficas.

Fonte: [Andrew Heiss | The ultimate practical guide to conjoint analysis with R.](#)

... gráfico de radar (ou de aranha);

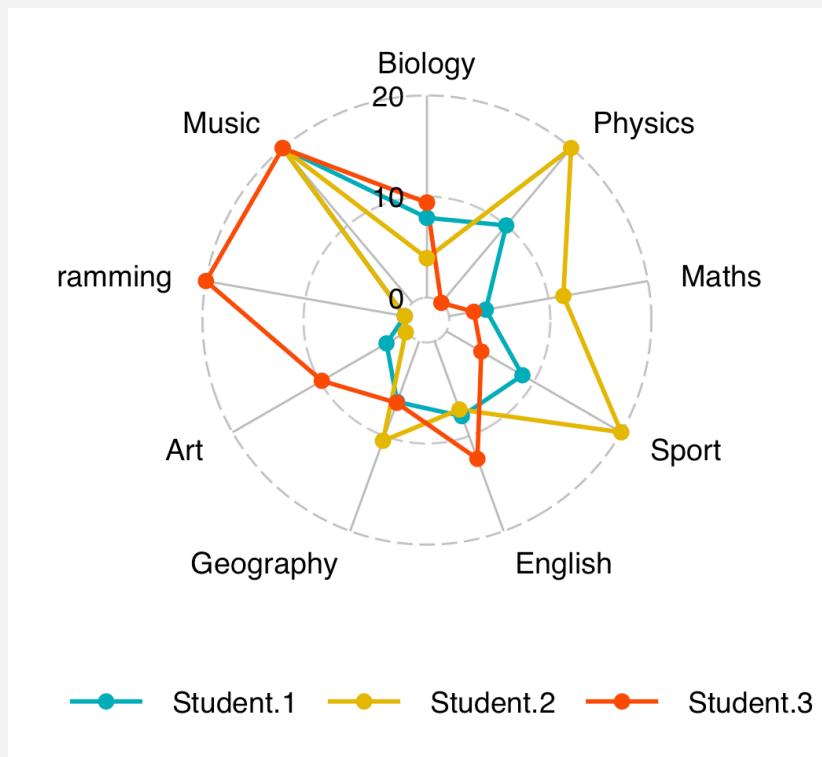


Figura 5: Gráfico de radar das notas de alunos de um determinado colégio.

Fonte: [Datanovia | Beautiful radar chart in R using fmsb and ggplot packages.](#)

... arte;

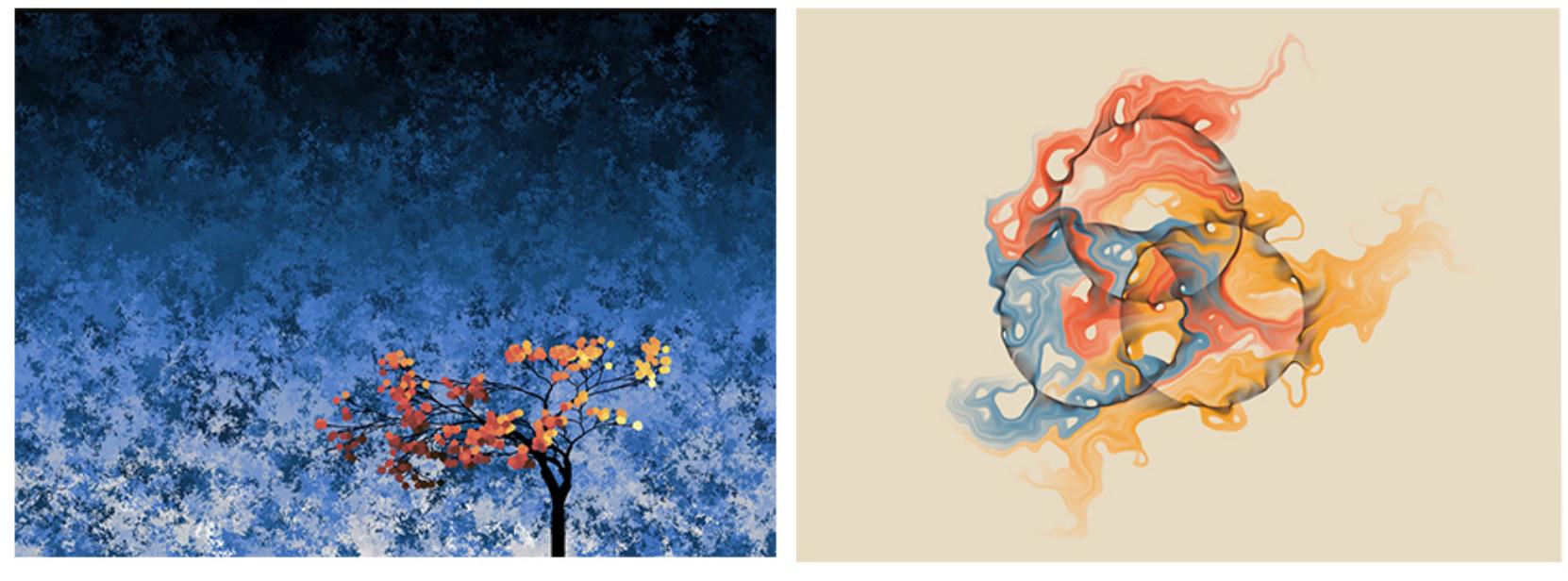


Figura 6: Arte com R por Danielle Navarro (à esquerda) e Thomas Lin Pedersen (à direita).

Fonte: [Art by Danielle Navarro](#) e [Data Imaginist - Visualization and beyond...](#)

... e muito mais!

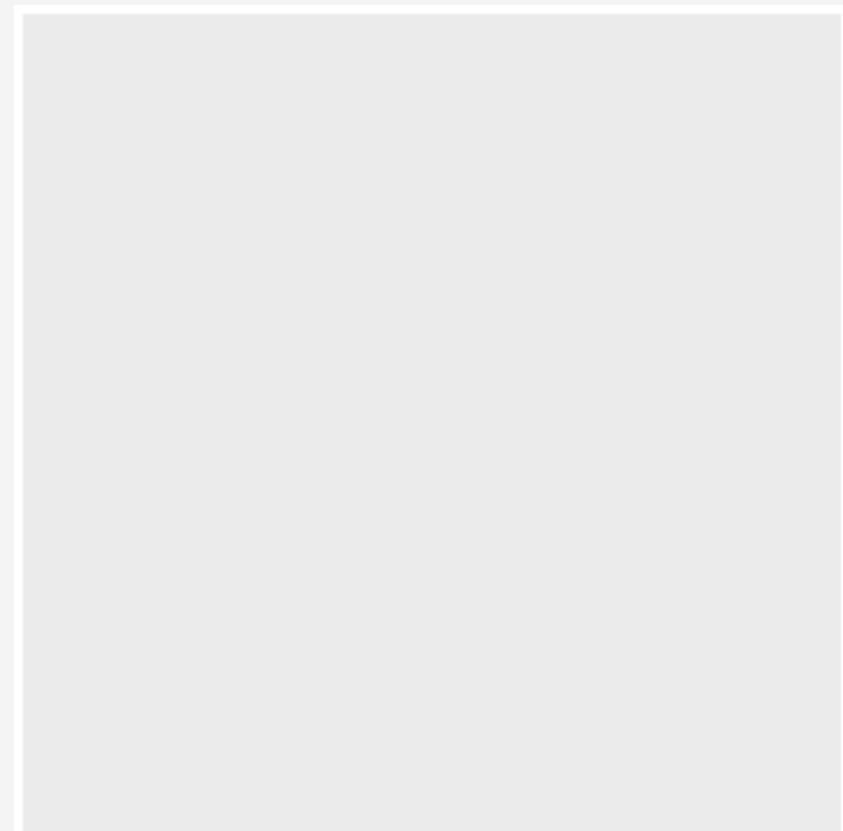
O pacote {ggplot2}

- Criado pelo estatístico neo-zelandês **Hadley Wickham**.
- No livro *A layered grammar of graphics* (em português: "Uma gramática em camadas dos gráficos"), Hadley define que os elementos de um gráfico (dados, cores, formas geométricas, coordenadas, anotações etc) são camadas e que um gráfico é um conjunto de sobreposições de camadas.
- Vantagens:
 - MAIS bonitos;
 - MAIS intuitivos;
 - MAIS customizáveis;
 - sintaxe MAIS padronizada.
- Documentação: <https://ggplot2.tidyverse.org/>.
- Cheatsheet: <https://rstudio.github.io/cheatsheets/data-visualization.pdf>.

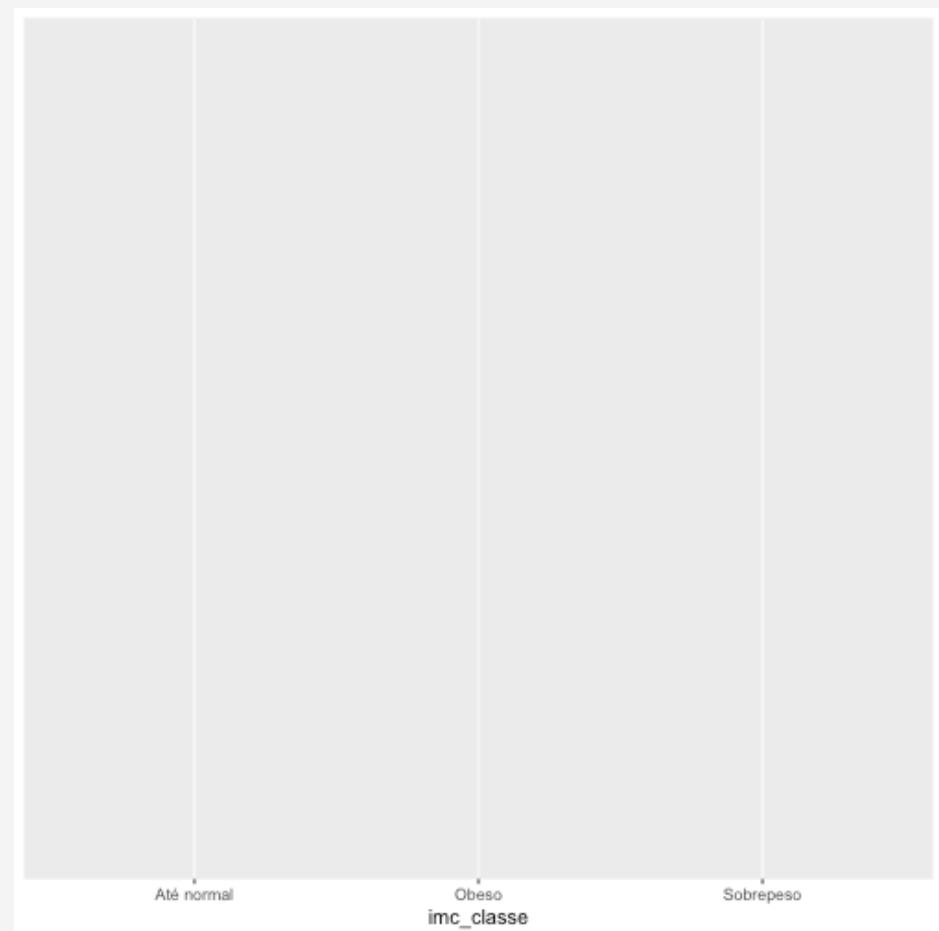
Gráfico univariados

Gráfico de barras

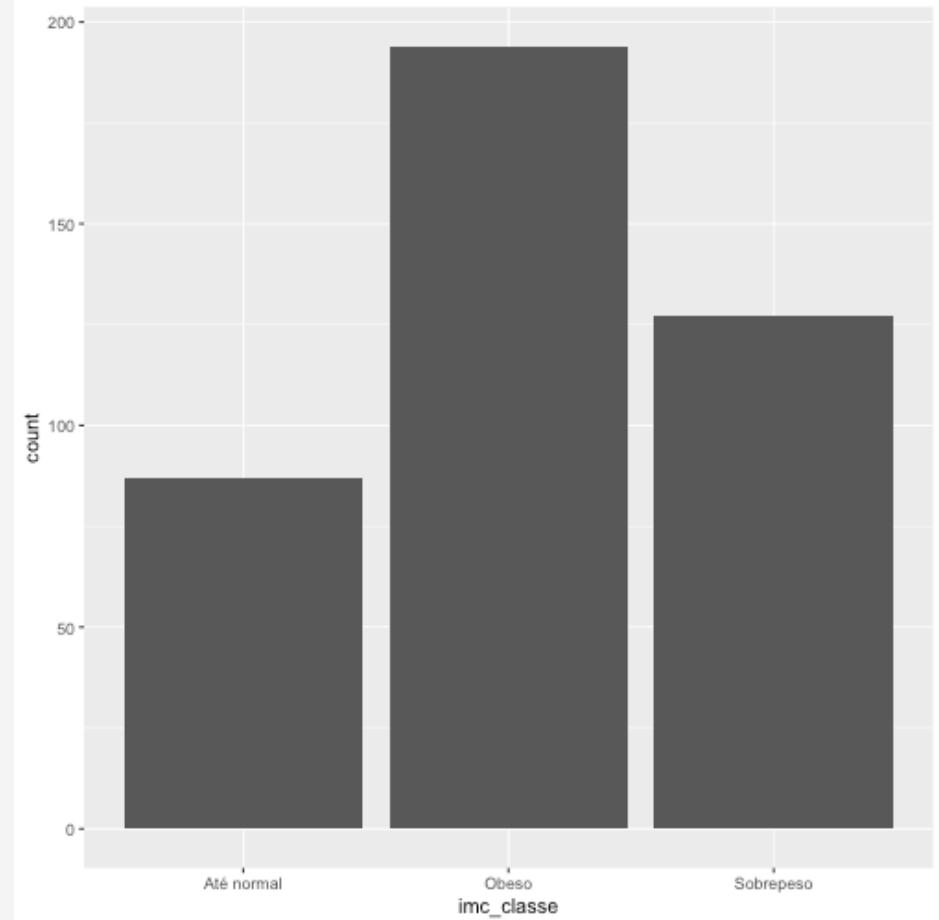
```
library(ggplot2)  
ggplot(dados1)
```



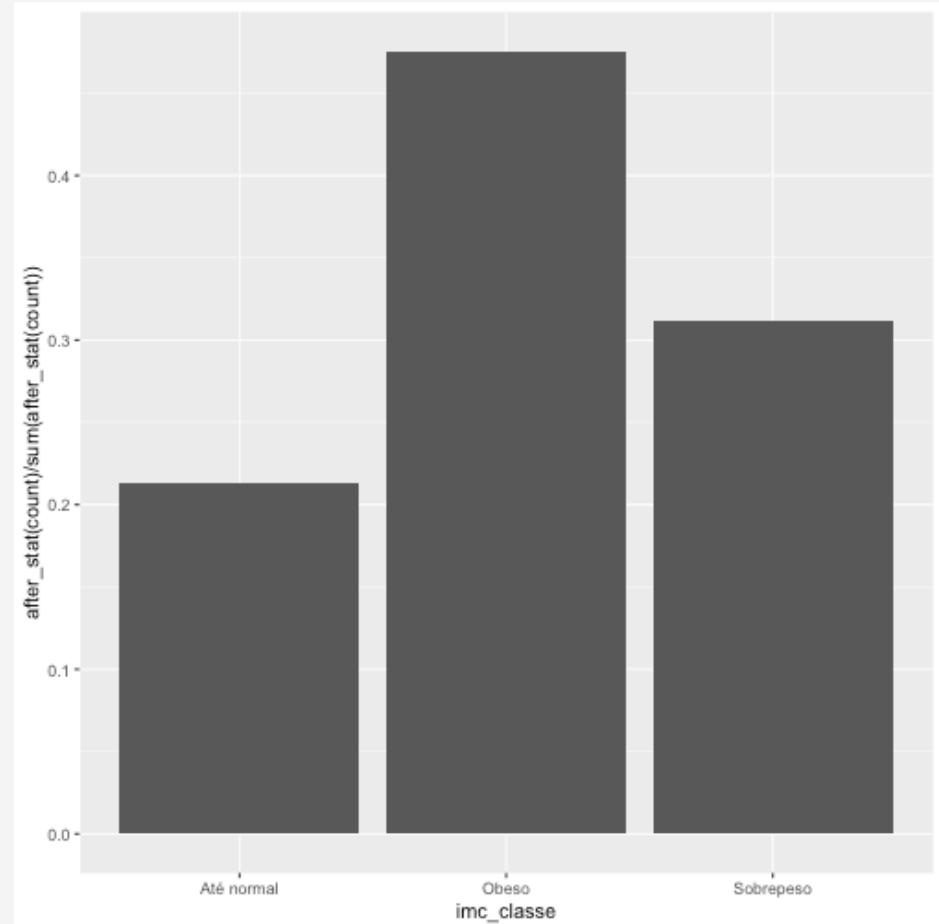
```
ggplot(dados1, aes(x = imc_classe))
```



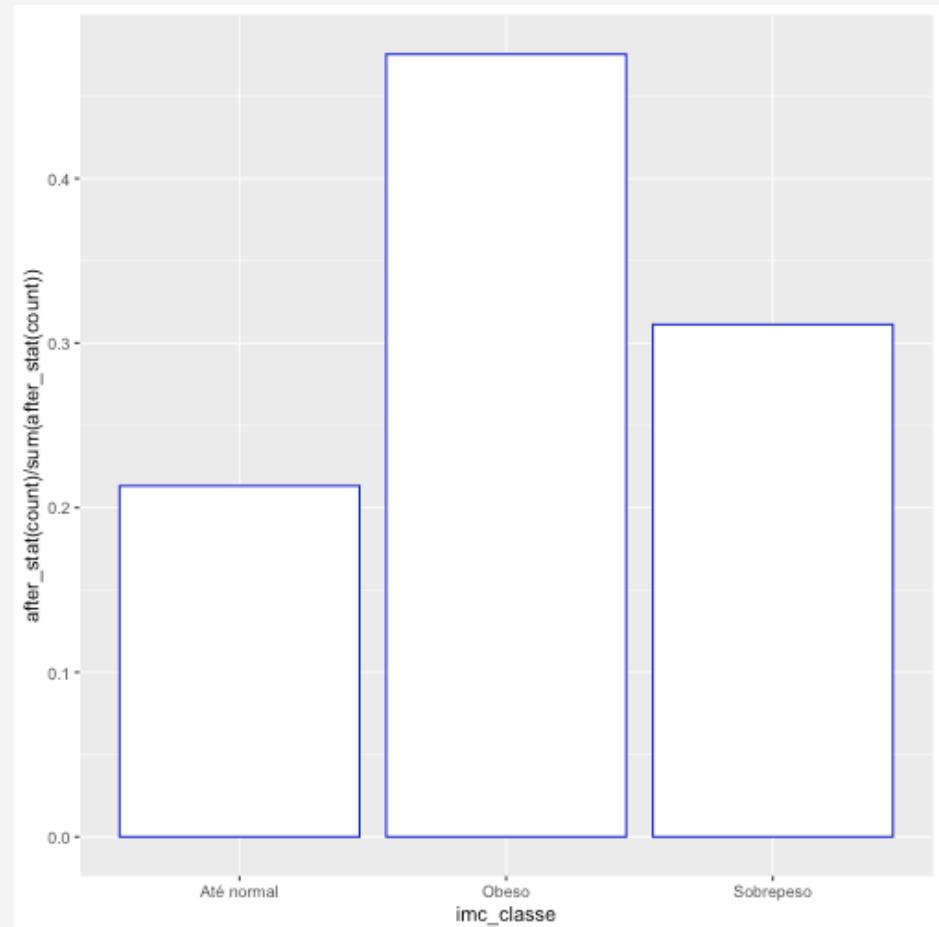
```
ggplot(dados1, aes(x = imc_classe)) +  
  geom_bar()
```



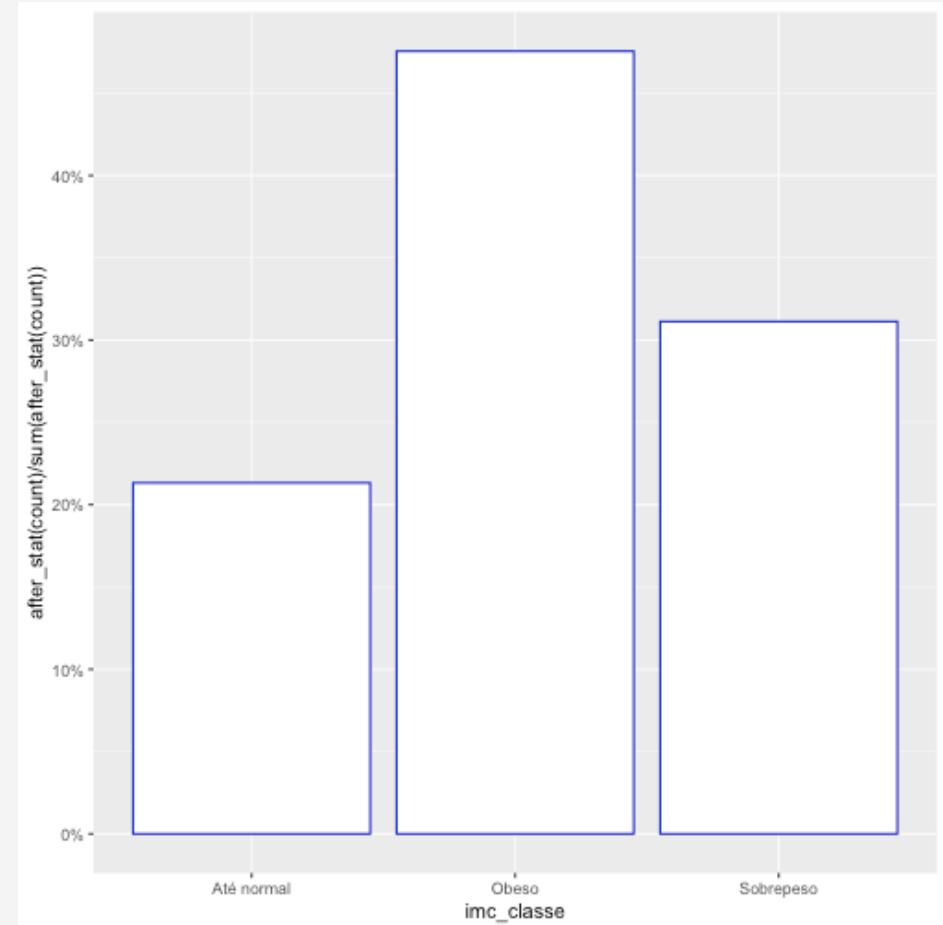
```
ggplot(dados1, aes(x = imc_classe, y = after_stat(count)/sum(after_stat(count)))) +  
  geom_bar()
```



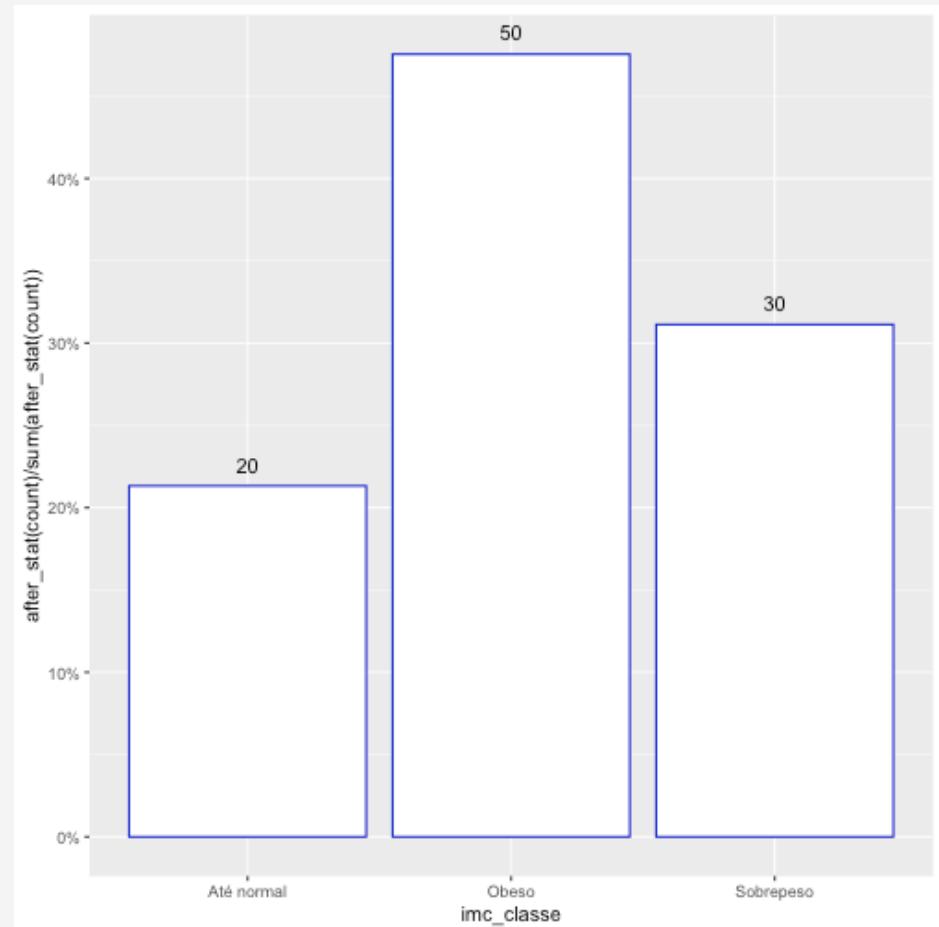
```
ggplot(dados1, aes(x = imc_classe, y = after_stat(count)/sum(after_stat(count))))  
  geom_bar(color = "#0000cd", fill = "#1f77b4")
```



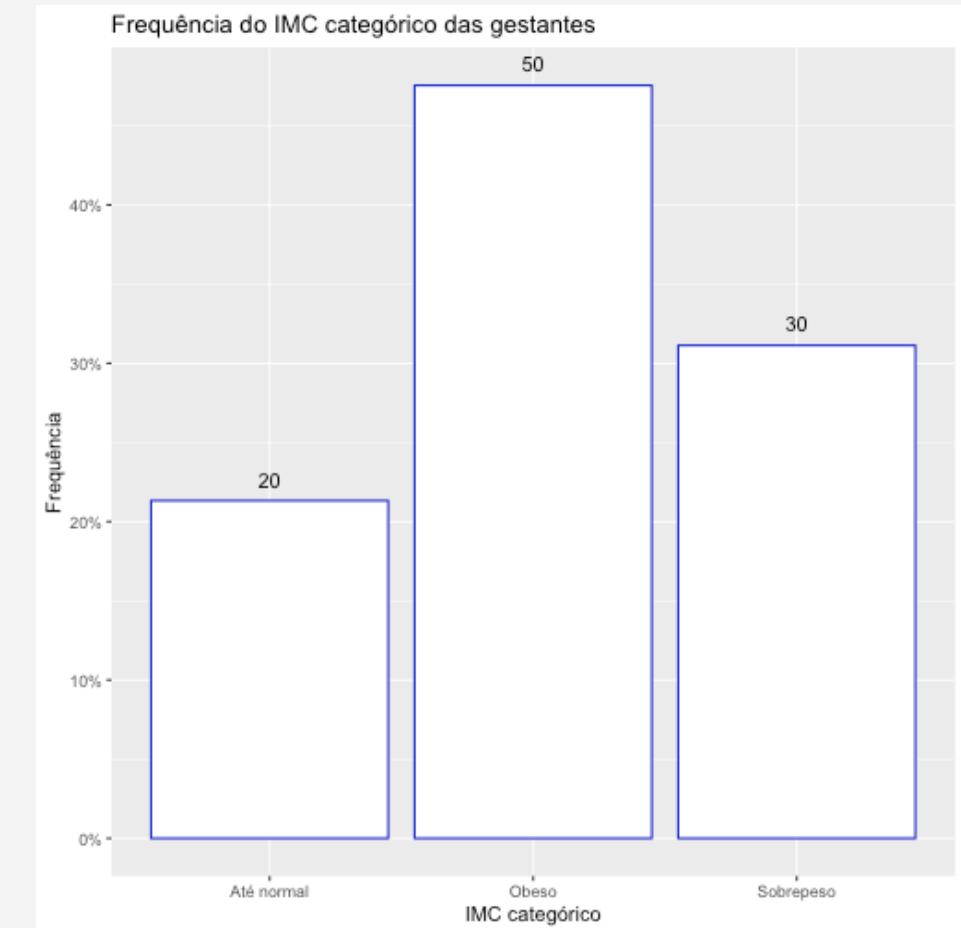
```
ggplot(dados1, aes(x = imc_classe, y = after_stat(count)/sum(after_stat(count)))  
  geom_bar(color = "#0000cd", fill = "#1f77b4")  
  scale_y_continuous(labels = scales::percent))
```



```
ggplot(dados1, aes(x = imc_classe, y = after_stat(count))) +  
  geom_bar(color = "#0000cd", fill = "#1f77b4") +  
  scale_y_continuous(labels = scales::percent) +  
  geom_text(  
    stat = "count",  
    aes(label = round(after_stat(count), 1),  
        vjust = -1))  
)
```

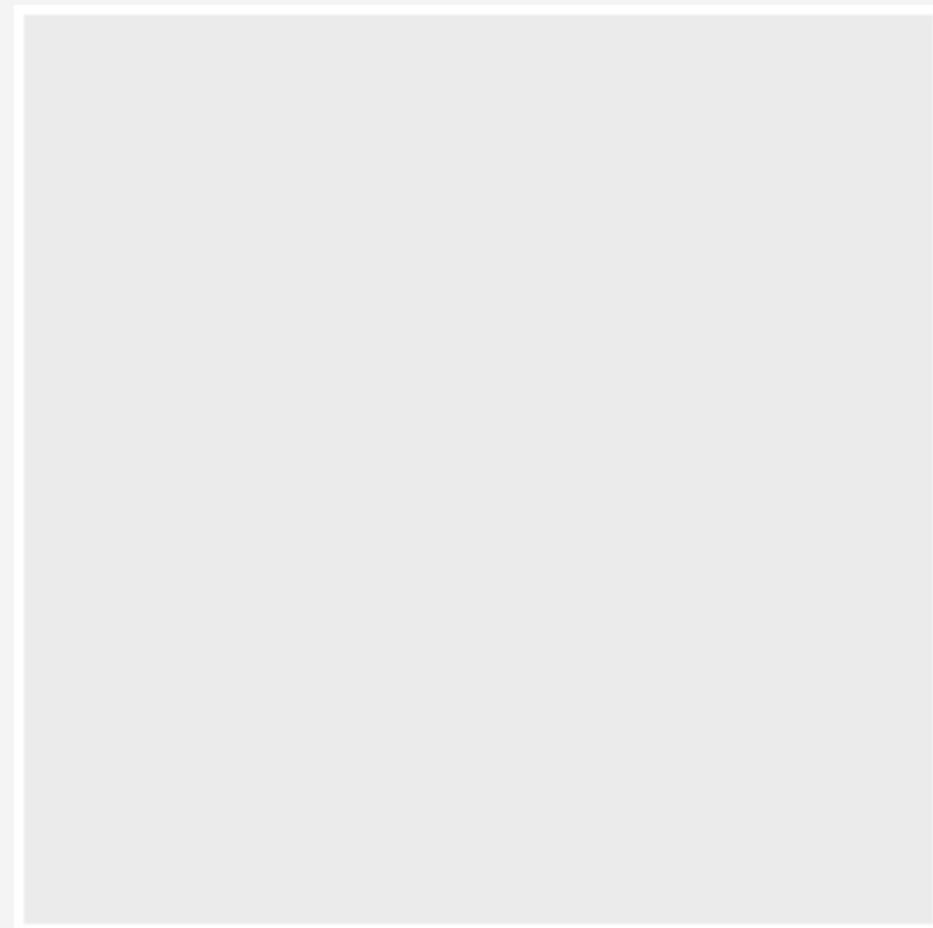


```
ggplot(dados1, aes(x = imc_classe, y = a
  geom_bar(color = "#0000cd", fill = "#1f77b4", stat = "identity")
  scale_y_continuous(labels = scales::percent)
  geom_text(
    stat = "count",
    aes(label = round(after_stat(count), 0), vjust = -1)
  ) +
  labs(
    title = "Frequência do IMC categórico",
    x = "IMC categórico",
    y = "Frequência"
  )
```

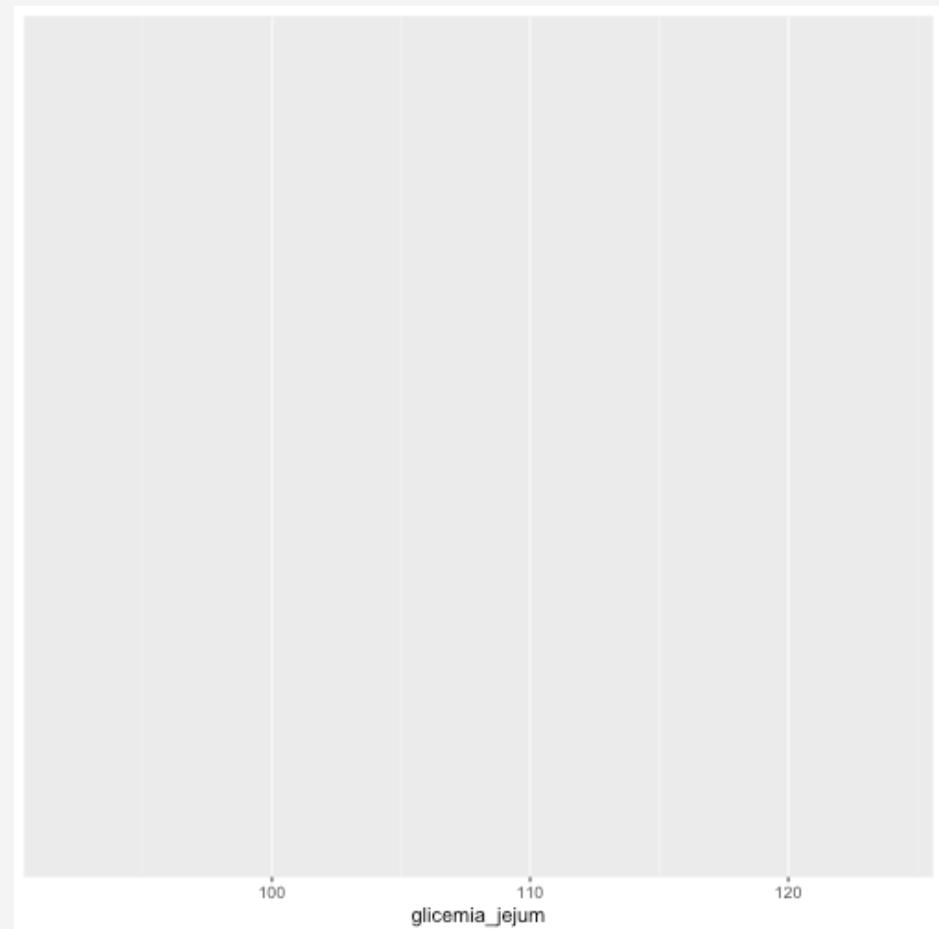


Histograma

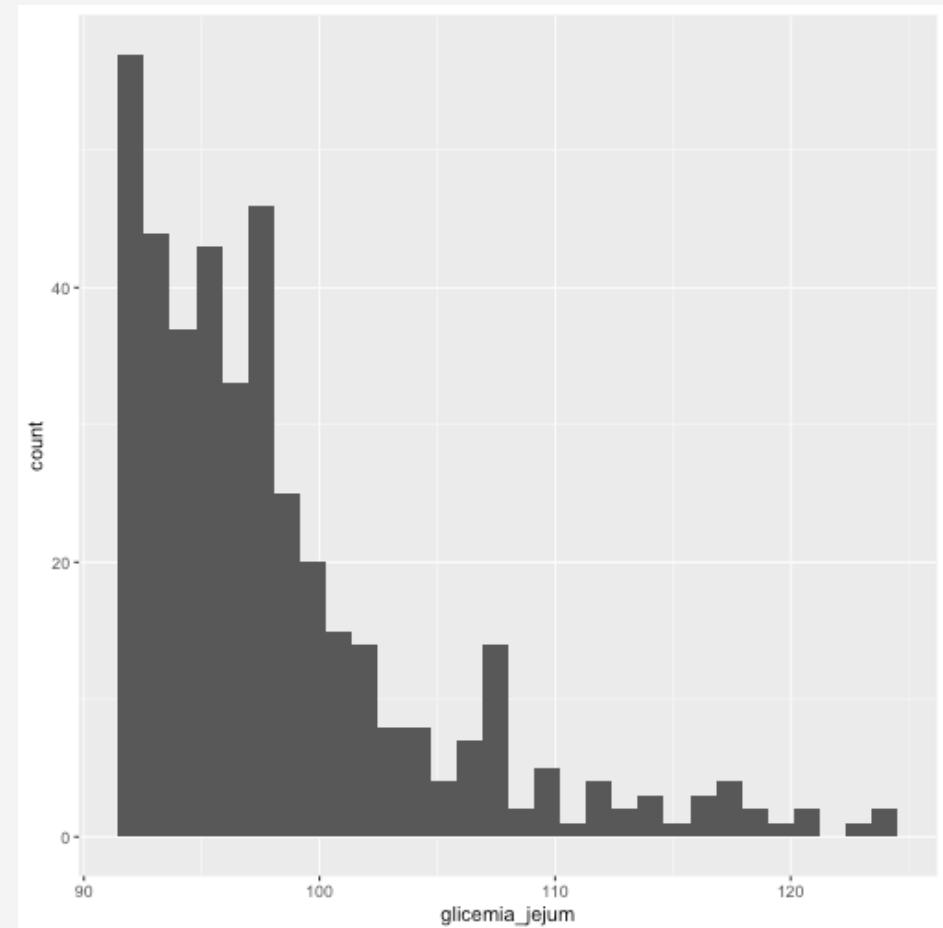
```
ggplot(dados1)
```



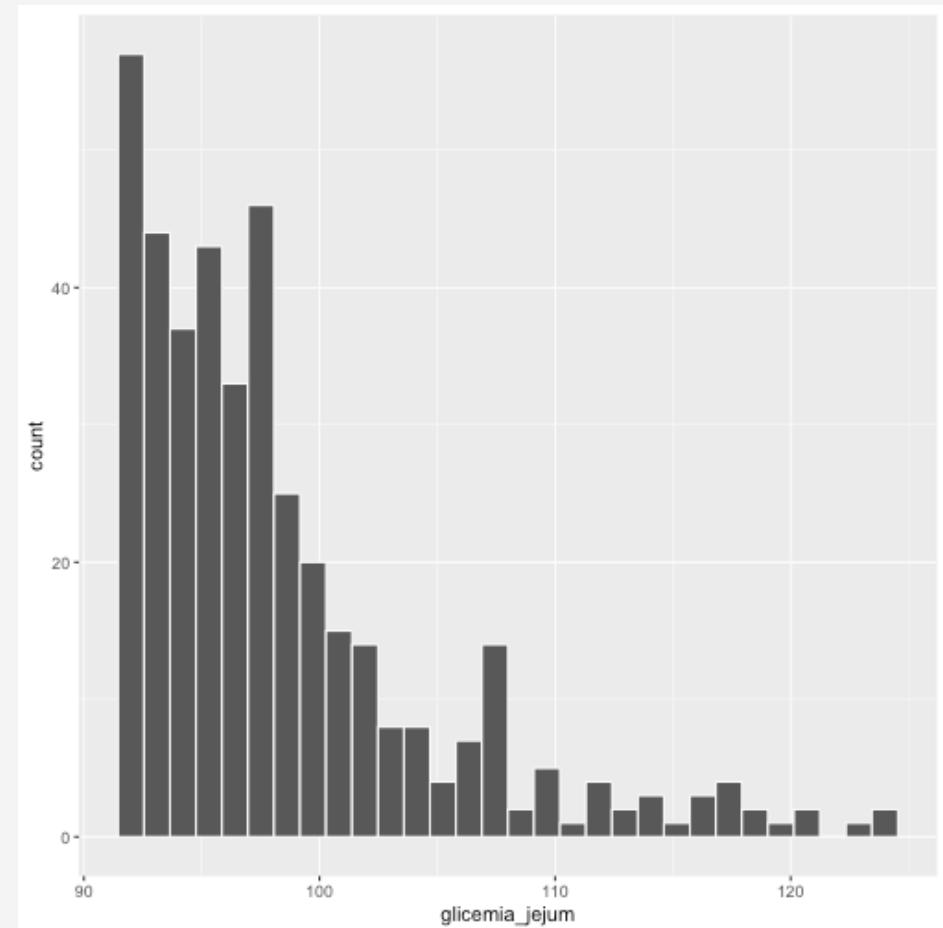
```
ggplot(dados1, aes(x = glicemia_jejum))
```



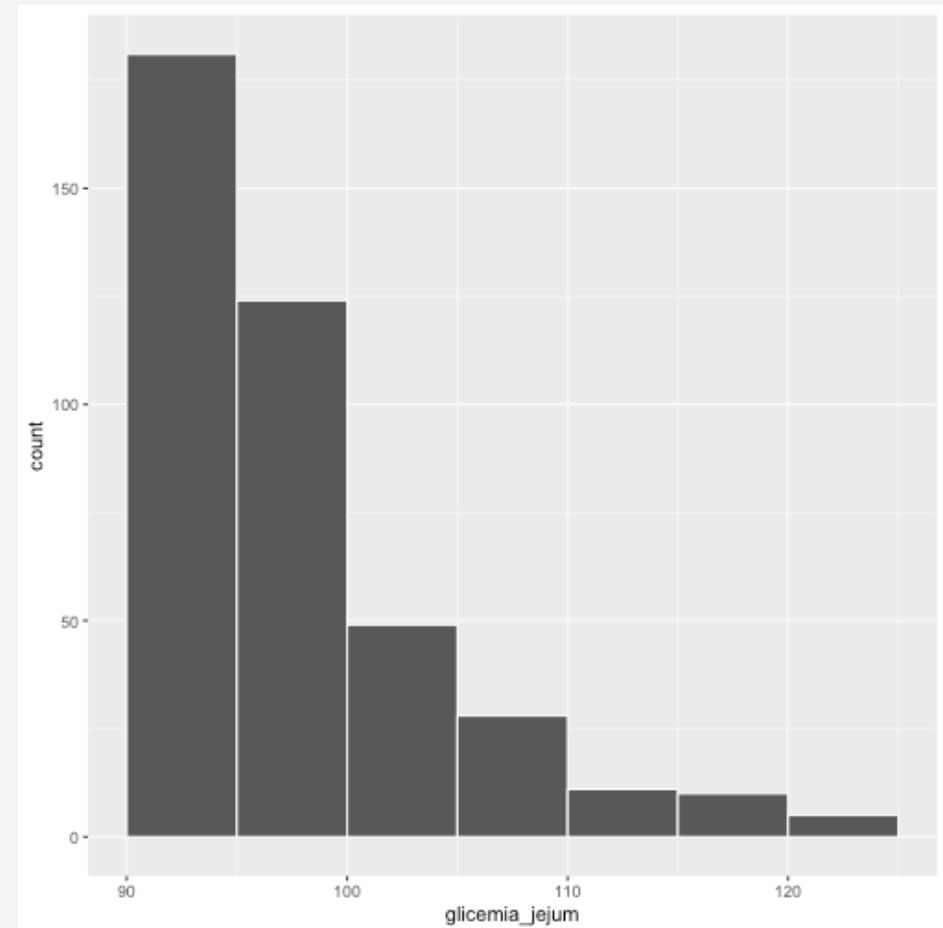
```
ggplot(dados1, aes(x = glicemia_jejum))  
  geom_histogram()
```



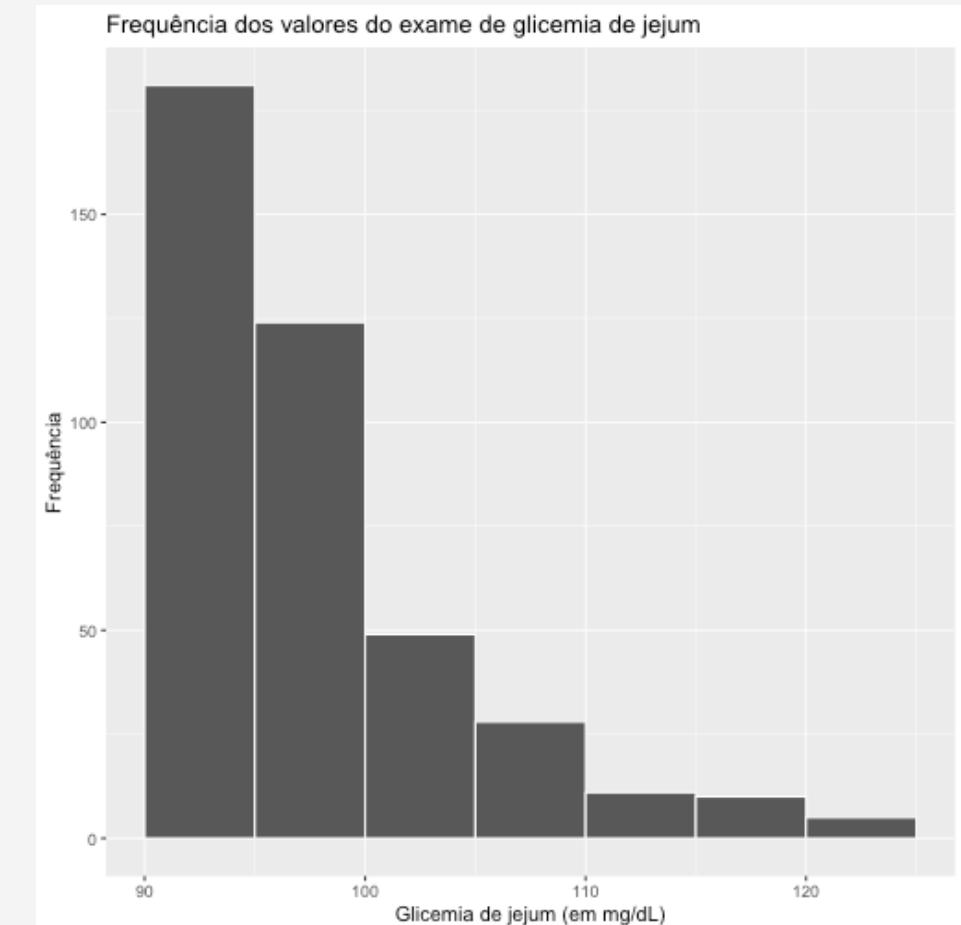
```
ggplot(dados1, aes(x = glicemia_jejum))  
  geom_histogram(color = "#ffffff")
```



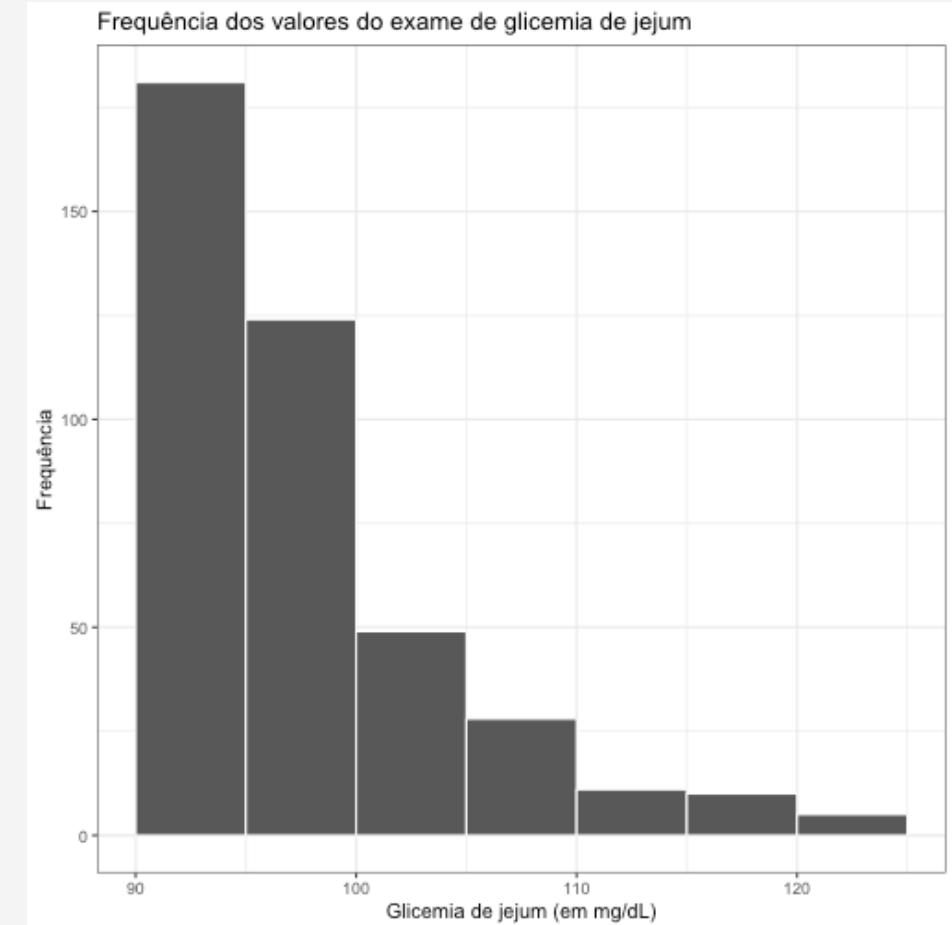
```
ggplot(dados1, aes(x = glicemia_jejum))  
  geom_histogram(color = "#ffffff", breaks = 10)
```



```
ggplot(dados1, aes(x = glicemia_jejun))  
  geom_histogram(color = "#ffffff", breaks = 10)  
  labs(  
    title = "Frequência dos valores do exame de glicemia de jejum",  
    x = "Glicemia de jejum (em mg/dL)",  
    y = "Frequência"  
)
```



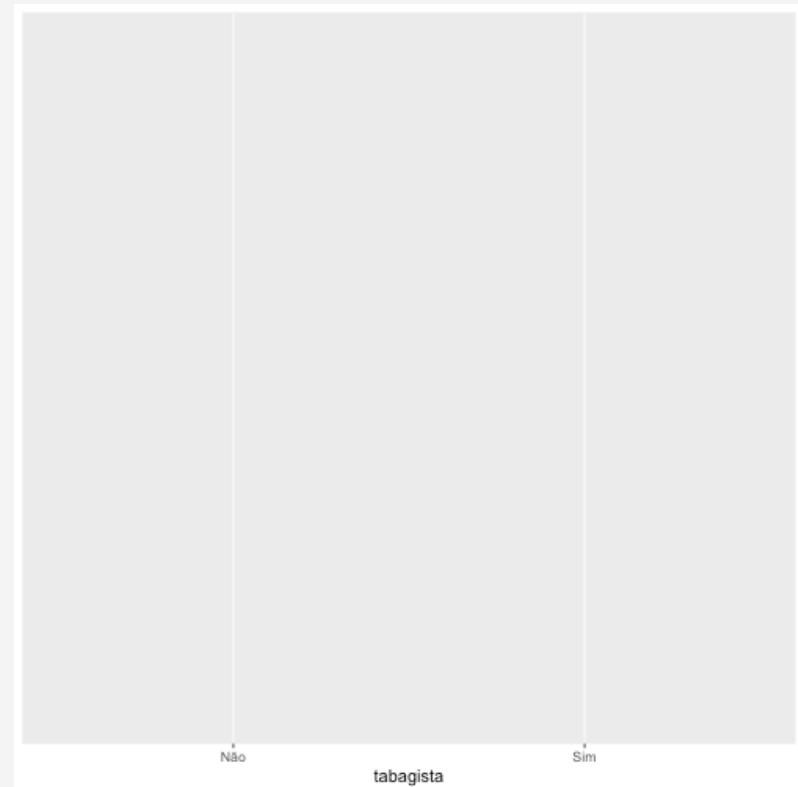
```
ggplot(dados1, aes(x = glicemia_jejum))  
  geom_histogram(color = "#ffffff", breaks = 10)  
  labs(  
    title = "Frequência dos valores do exame de glicemia de jejum",  
    x = "Glicemia de jejum (em mg/dL)",  
    y = "Frequência"  
) +  
  theme_bw()
```



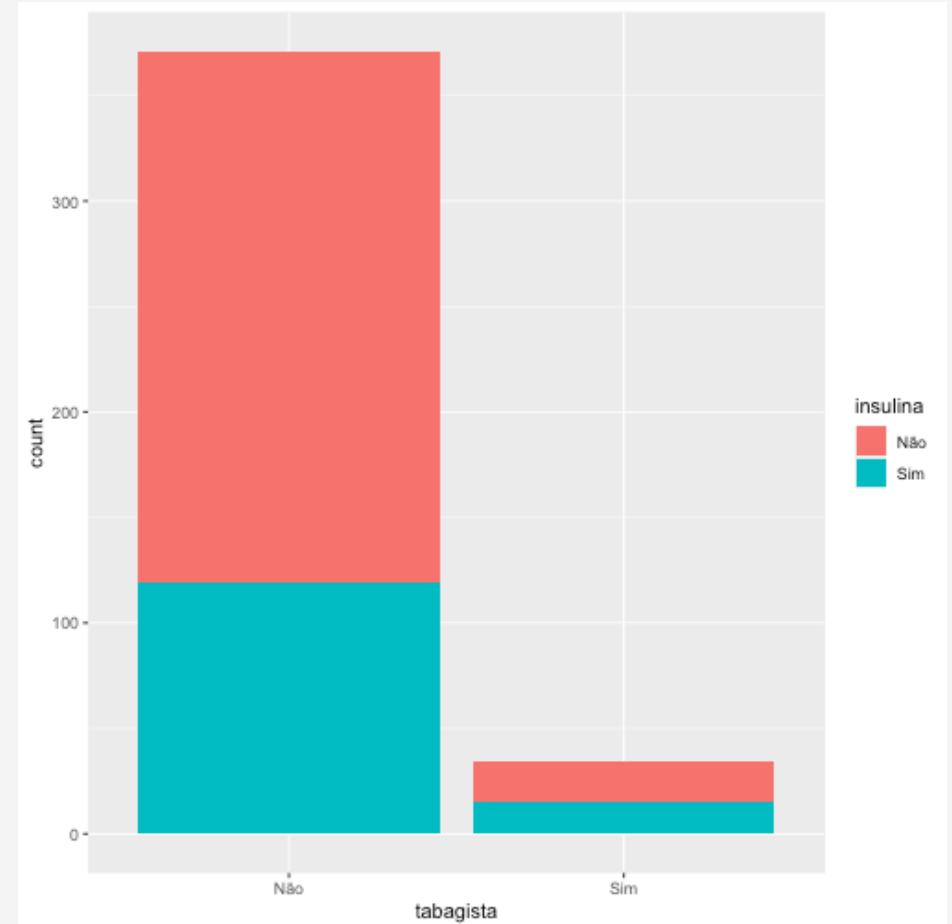
Gráficos bivariados (qualitativas x qualitativas)

Gráfico de barras agrupadas

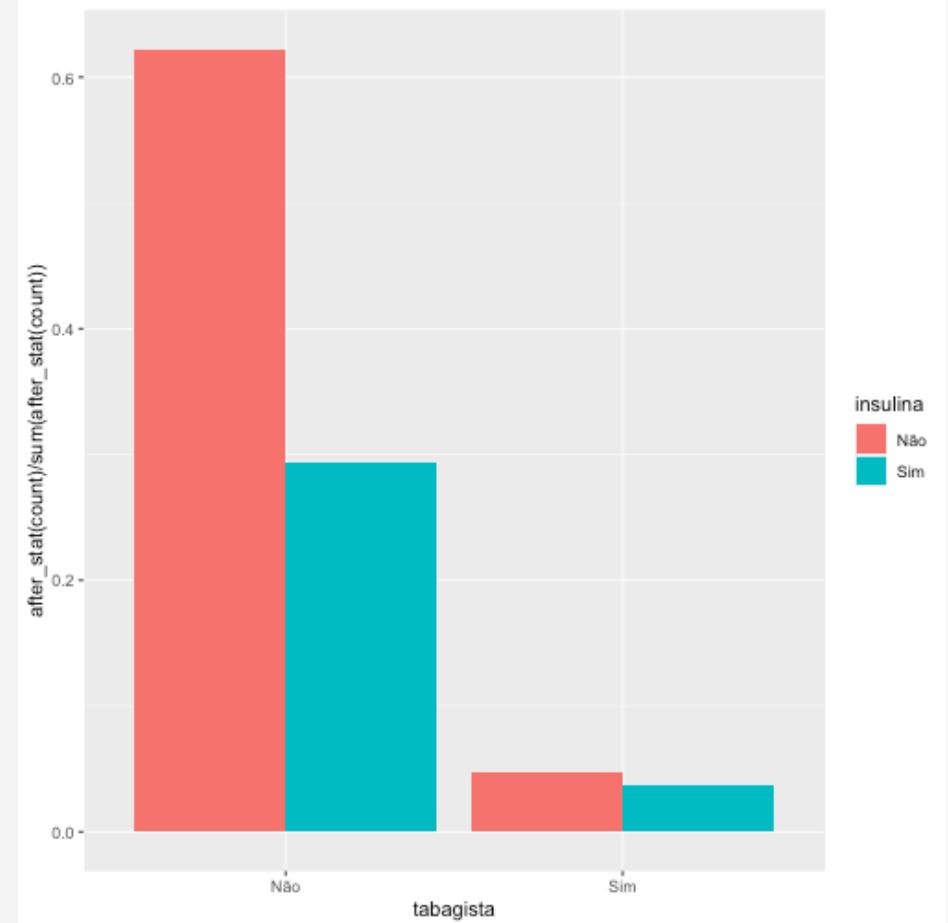
```
dados1 |>  
  dplyr::filter(!is.na(tabagista)) |>  
  ggplot(aes(x = tabagista, fill = insu
```



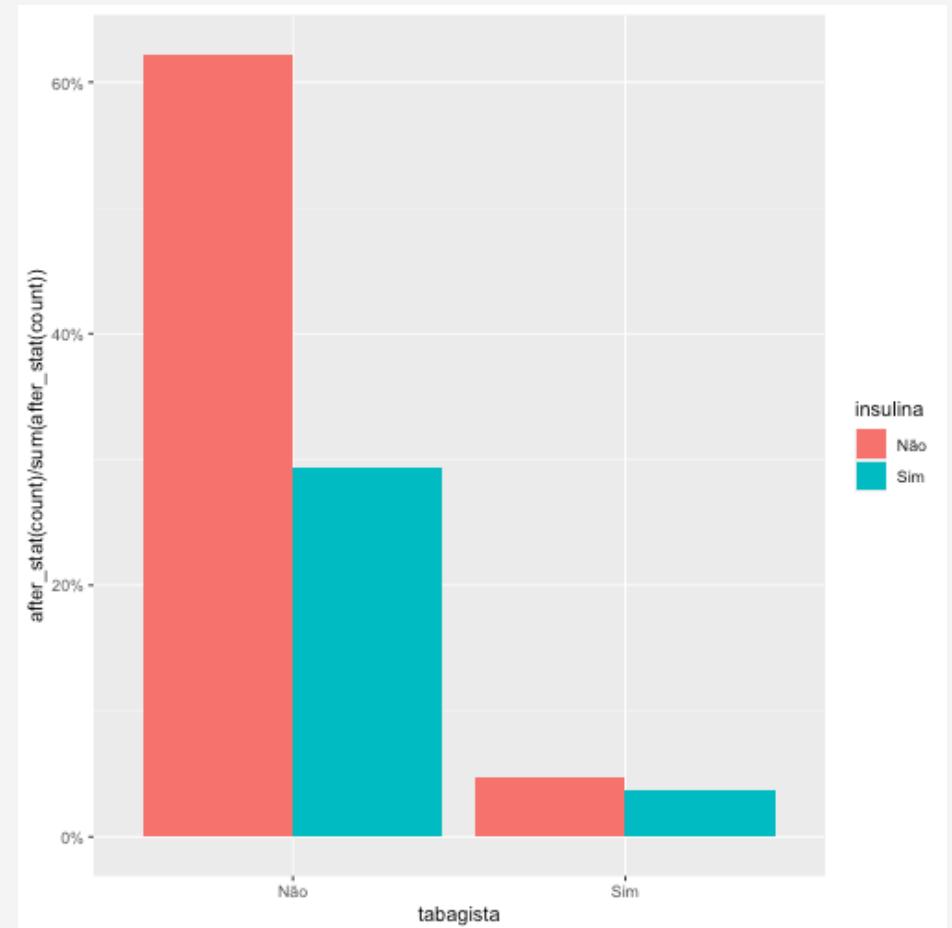
```
dados1 |>  
  dplyr::filter(!is.na(tabagista)) |>  
  ggplot(aes(x = tabagista, fill = insulina)) |>  
  geom_bar()
```



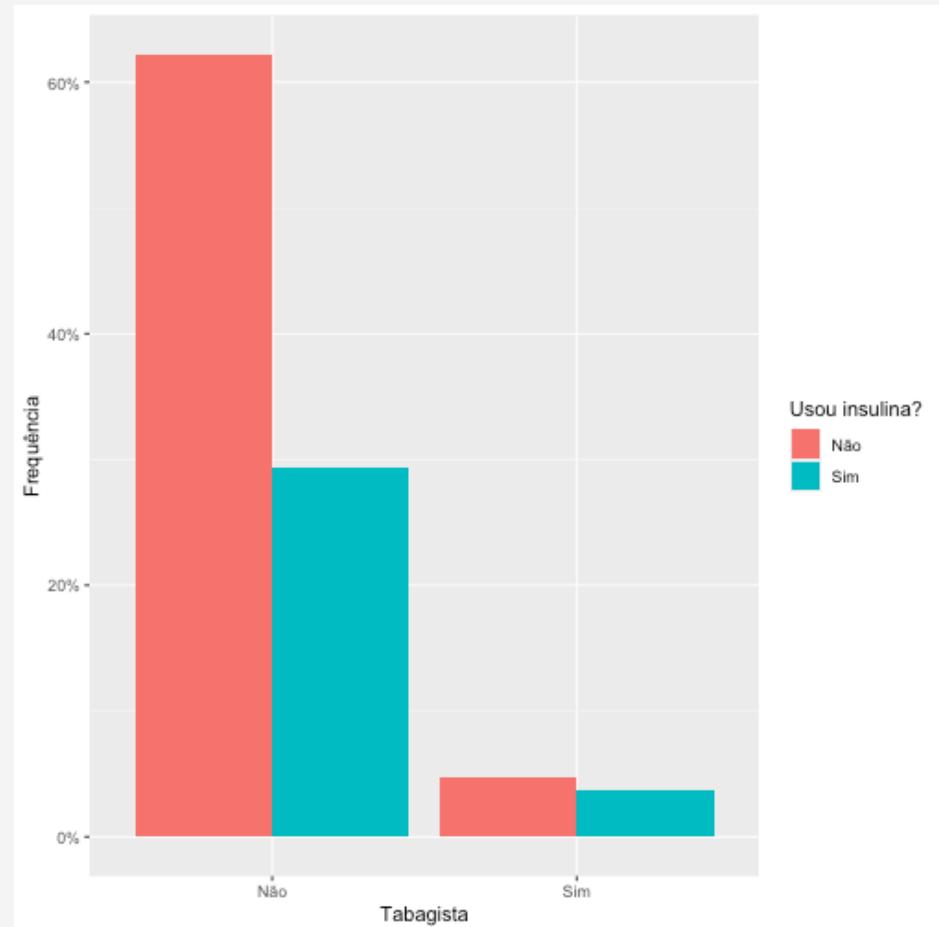
```
dados1 |>  
  dplyr::filter(!is.na(tabagista)) |>  
  ggplot(aes(x = tabagista, fill = insulina)) +  
    geom_bar(aes(y = after_stat(count)/sum(after_stat(count))))
```



```
dados1 |>  
  dplyr::filter(!is.na(tabagista)) |>  
  ggplot(aes(x = tabagista, fill = insulina)) +  
    geom_bar(aes(y = after_stat(count)/sum(after_stat(count)))) +  
    scale_y_continuous(labels = scales::percent_format(0, 2))
```



```
dados1 |>  
  dplyr::filter(!is.na(tabagista)) |>  
  ggplot(aes(x = tabagista, fill = insulin  
  geom_bar(aes(y = after_stat(count)/s  
  scale_y_continuous(labels = scales::  
  labs(  
    x = "Tabagista",  
    y = "Frequência",  
    fill = "Usou insulina?"  
  )
```



```
dados1 |>  
  dplyr::filter(!is.na(tabagista)) |>  
  ggplot(aes(x = tabagista, fill = insulin  
  geom_bar(aes(y = after_stat(count)/s  
  scale_y_continuous(labels = scales::  
  labs(  
    x = "Tabagista",  
    y = "Frequência",  
    fill = "Usou insulina?"  
  ) +  
  theme_bw()
```

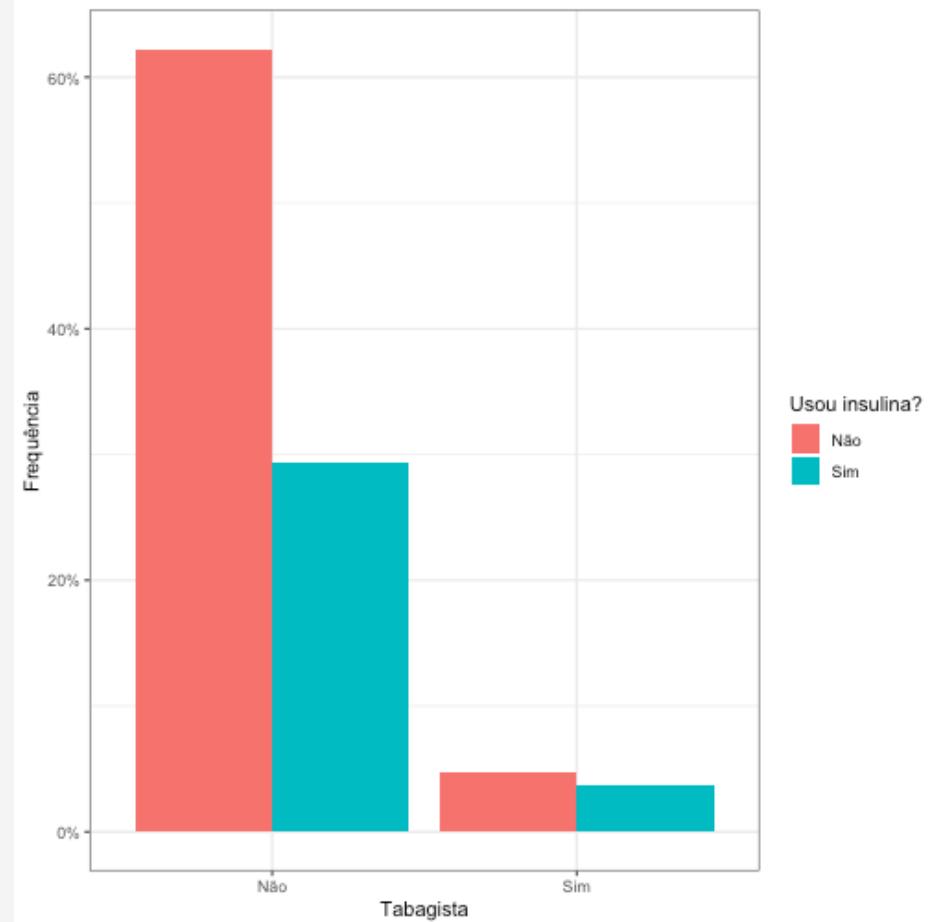
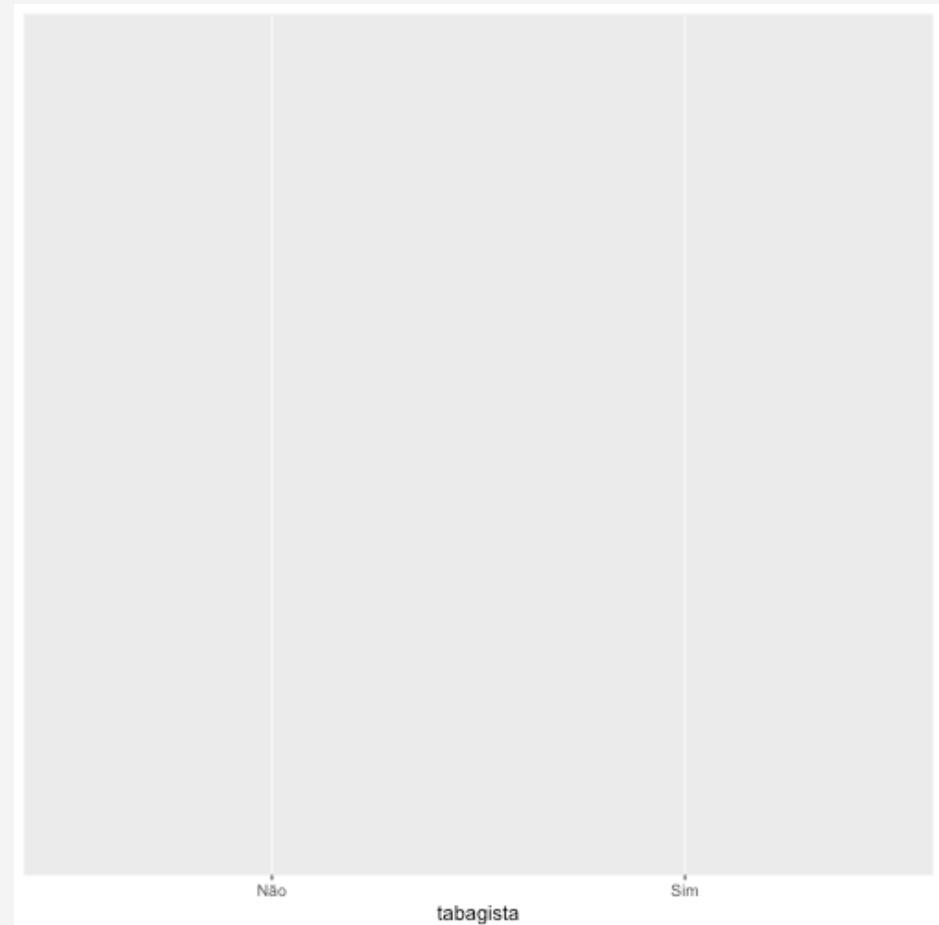
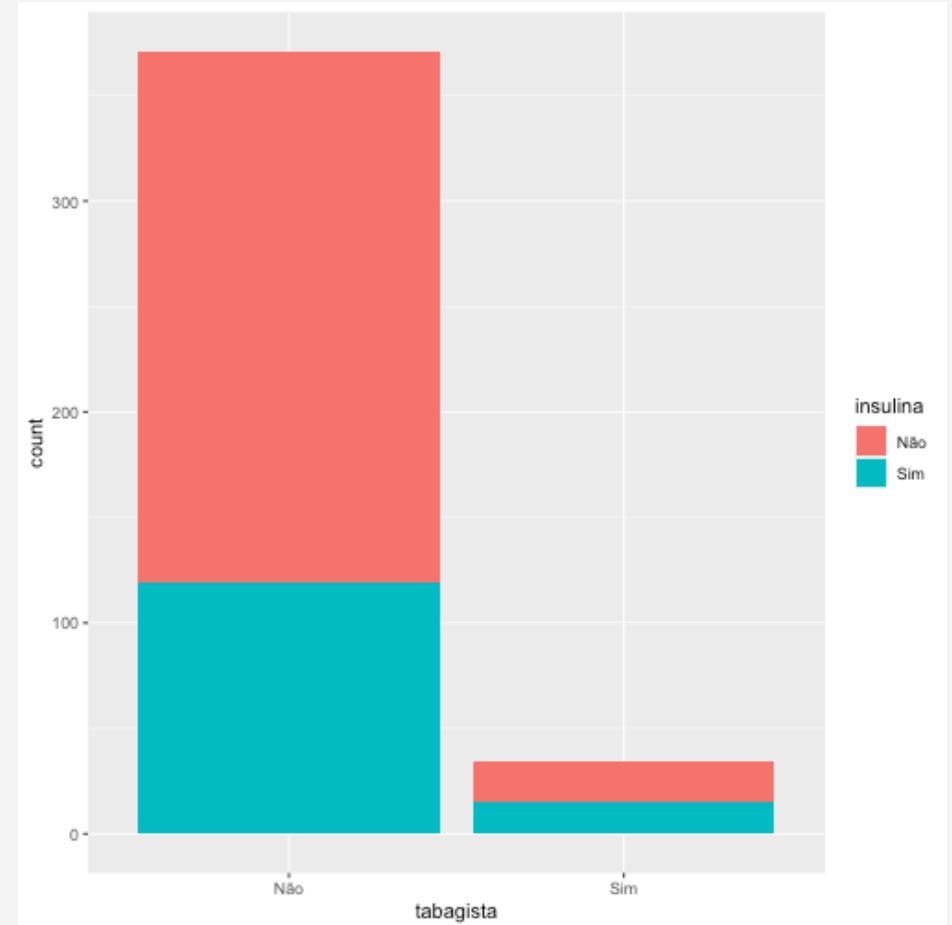


Gráfico de barras empilhadas

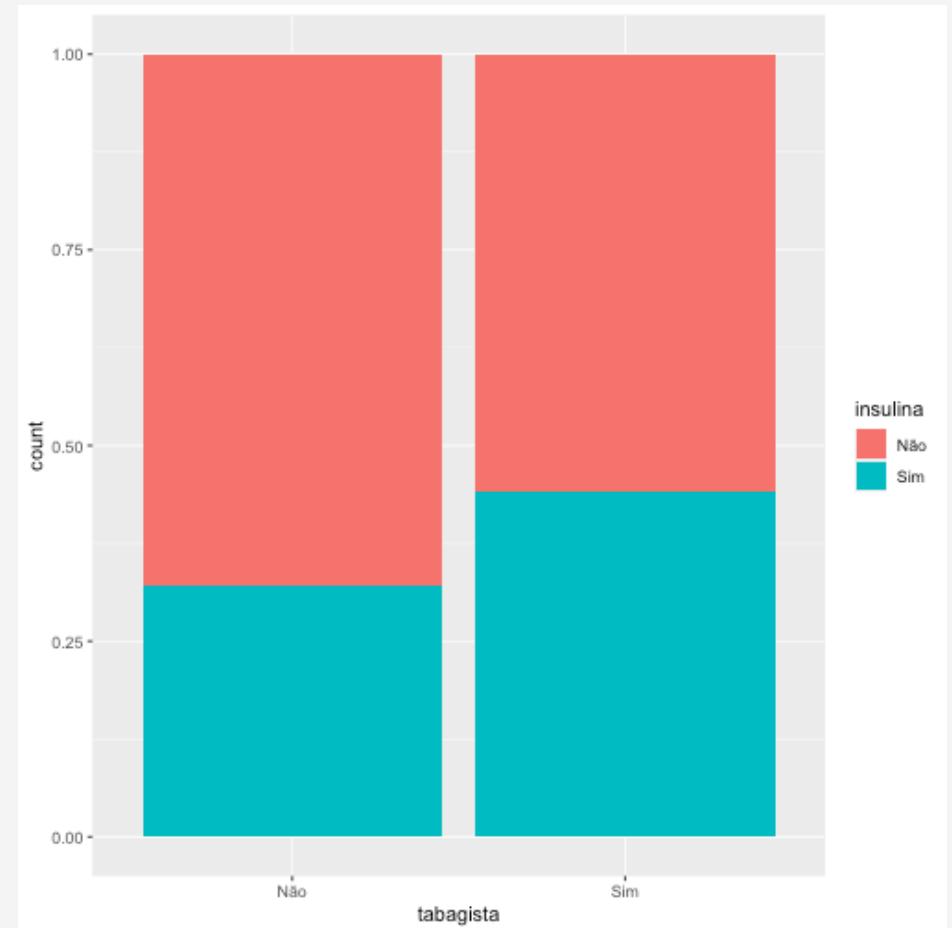
```
dados1 |>  
  dplyr::filter(!is.na(tabagista)) |>  
  ggplot(aes(x = tabagista, fill = insu
```



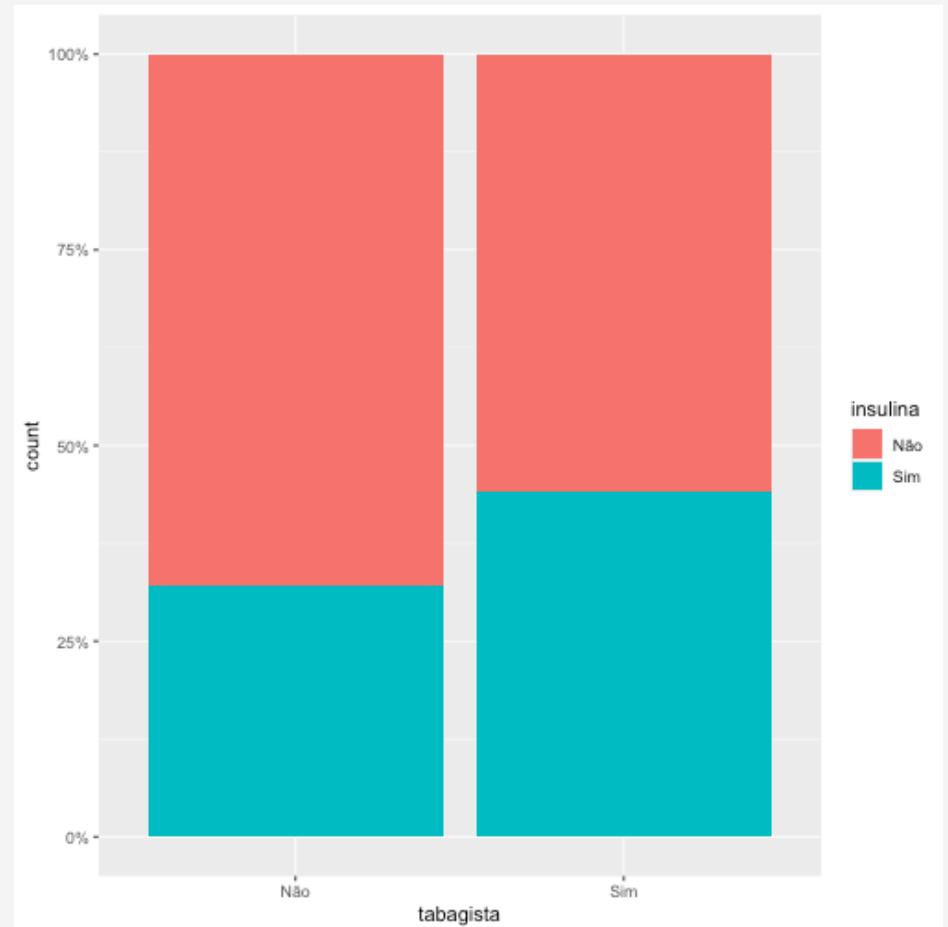
```
dados1 |>  
  dplyr::filter(!is.na(tabagista)) |>  
  ggplot(aes(x = tabagista, fill = insulina)) +  
    geom_bar(position = "stack")
```



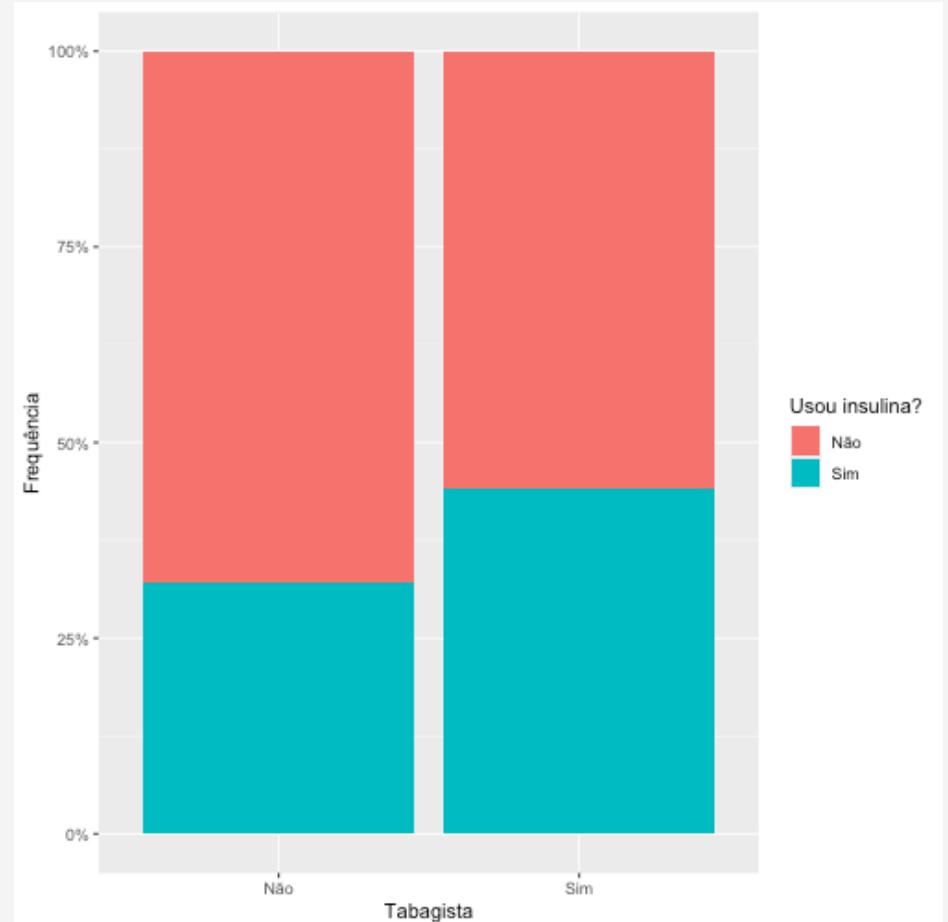
```
dados1 |>  
  dplyr::filter(!is.na(tabagista)) |>  
  ggplot(aes(x = tabagista, fill = insulina)) +  
    geom_bar(position = "fill")
```



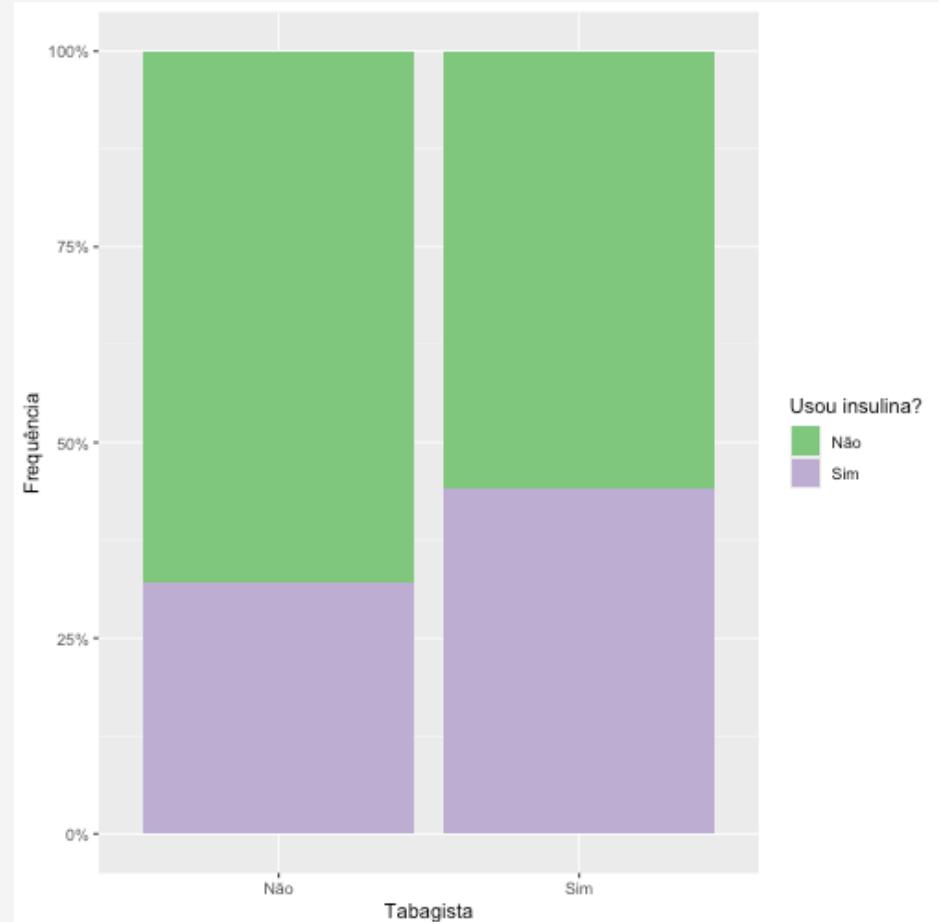
```
dados1 |>  
  dplyr::filter(!is.na(tabagista)) |>  
  ggplot(aes(x = tabagista, fill = insulina)) +  
  geom_bar(position = "fill") +  
  scale_y_continuous(labels = scales::percent)
```



```
dados1 |>  
  dplyr::filter(!is.na(tabagista)) |>  
  ggplot(aes(x = tabagista, fill = insulin  
    geom_bar(position = "fill") +  
    scale_y_continuous(labels = scales::  
    labs(  
      x = "Tabagista",  
      y = "Frequência",  
      fill = "Usou insulina?"  
    )
```



```
dados1 |>  
  dplyr::filter(!is.na(tabagista)) |>  
  ggplot(aes(x = tabagista, fill = insulin  
    geom_bar(position = "fill") +  
    scale_y_continuous(labels = scales::  
    labs(  
      x = "Tabagista",  
      y = "Frequência",  
      fill = "Usou insulina?"  
    ) +  
    scale_fill_brewer(palette = "Accent"
```



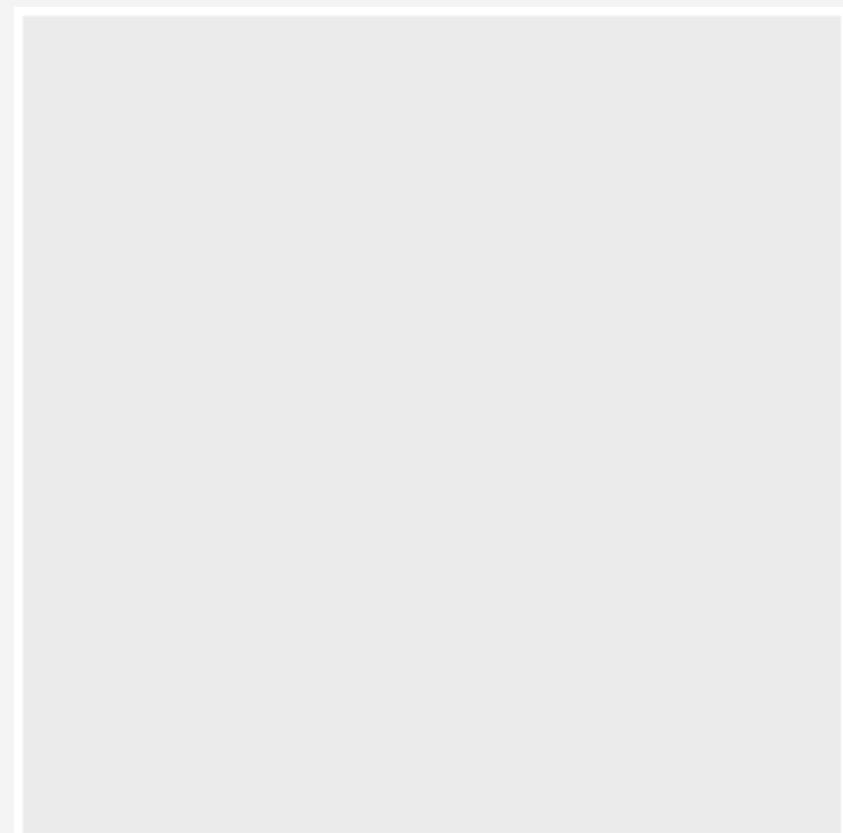
Dica!

Para ver outras paletas R Color Brewer, [clique aqui!](#)

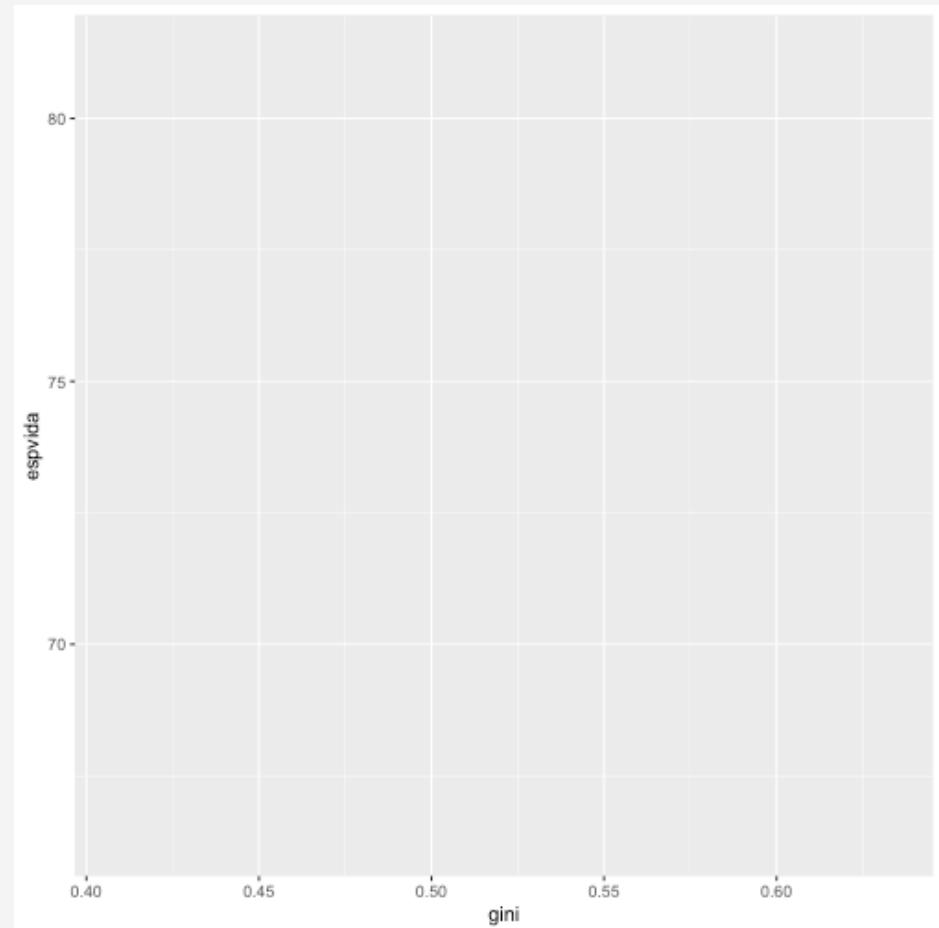
Gráficos bivariados (quantitativas x quantitativas)

Gráfico de dispersão

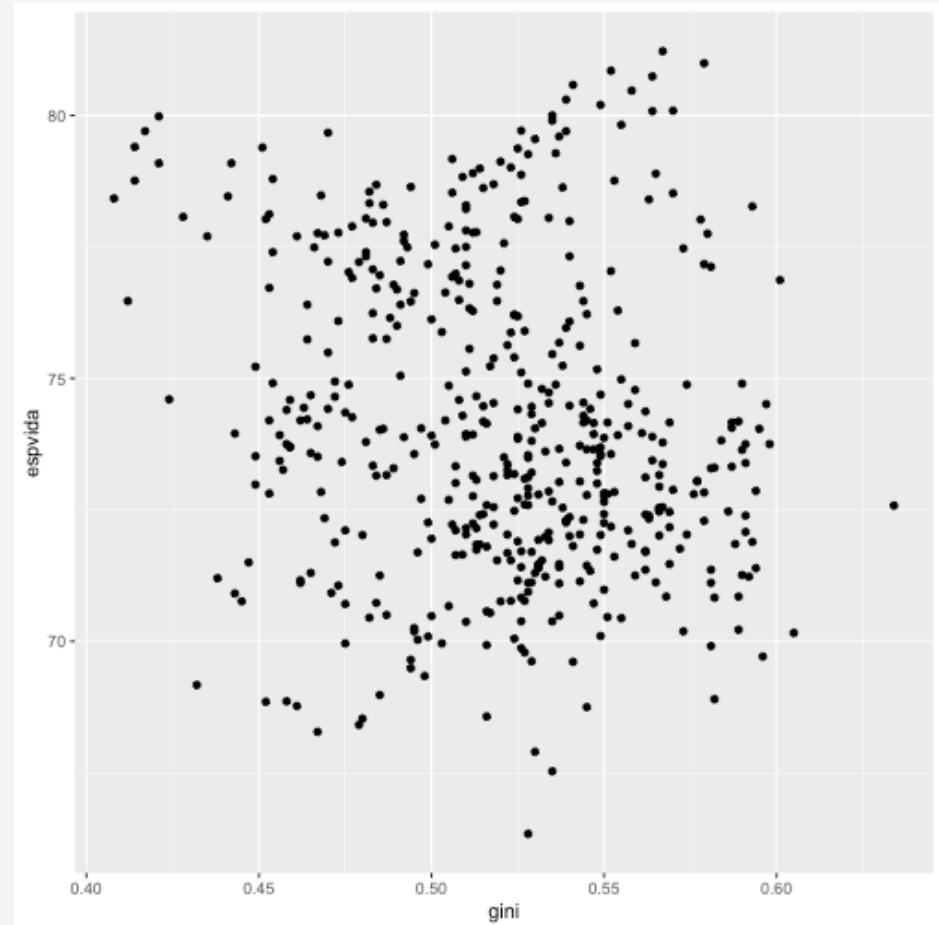
```
ggplot(dados2)
```



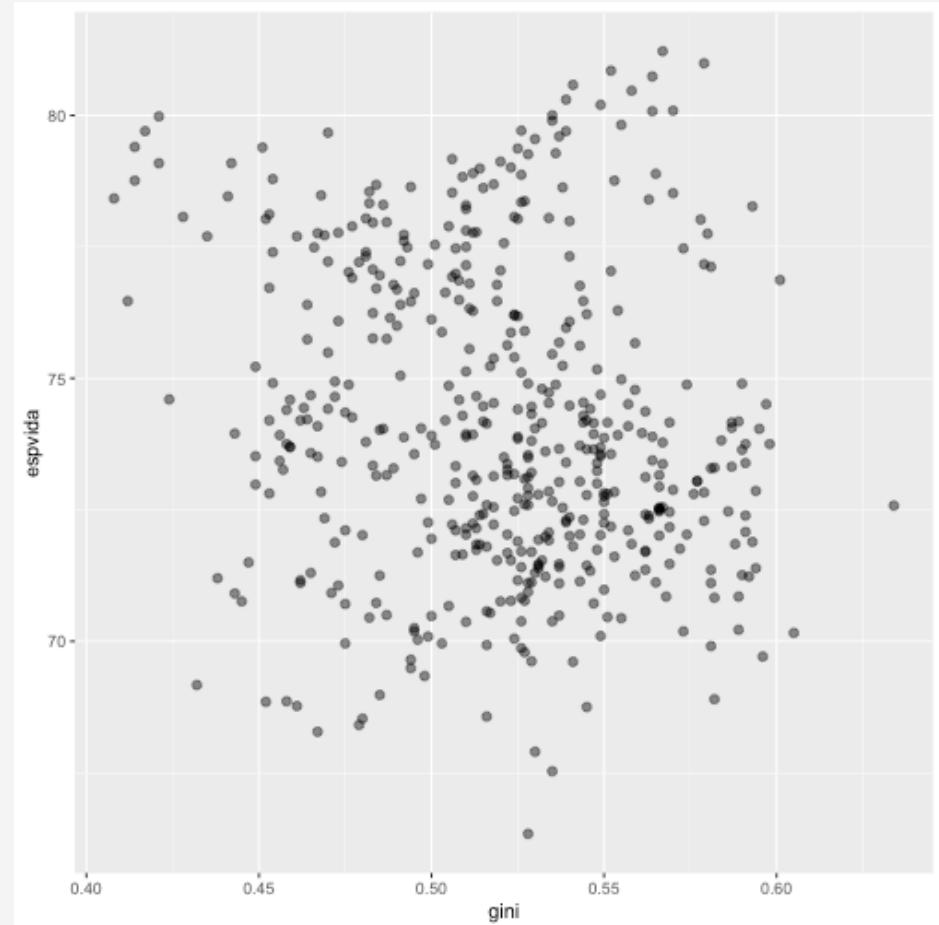
```
ggplot(dados2, aes(x = gini, y = espvida))
```



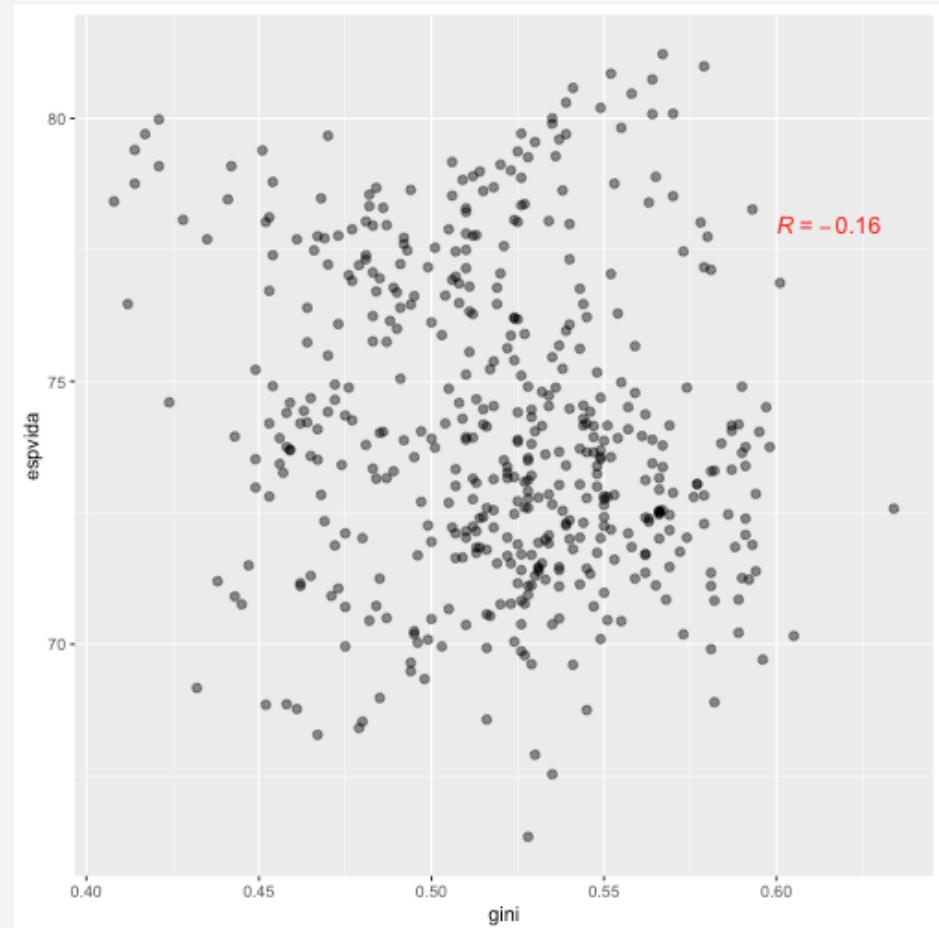
```
ggplot(dados2, aes(x = gini, y = espvida))  
  geom_point()
```



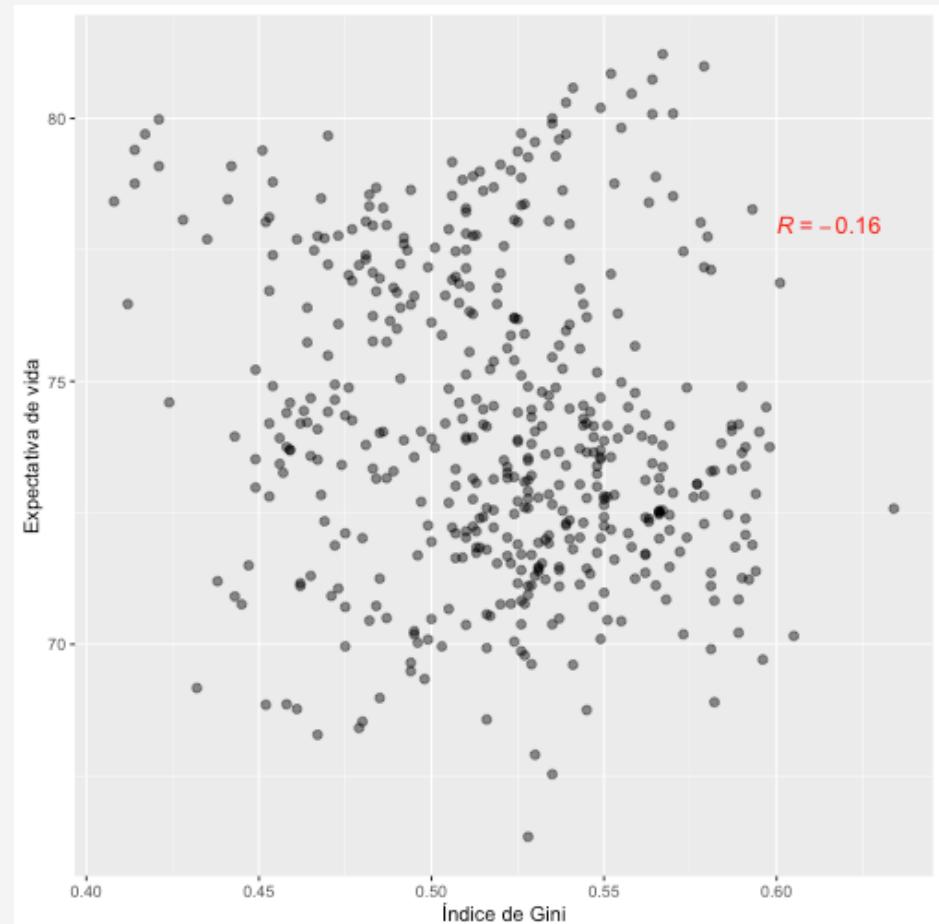
```
ggplot(dados2, aes(x = gini, y = espvida  
geom_point(size = 2, alpha = .5)
```



```
ggplot(dados2, aes(x = gini, y = espvida)) +  
  geom_point(size = 2, alpha = .5) +  
  ggpubr::stat_cor(  
    aes(label = after_stat(r.label)),  
    method = "pearson",  
    label.x = 0.6, label.y = 78, size = 3)  
)
```



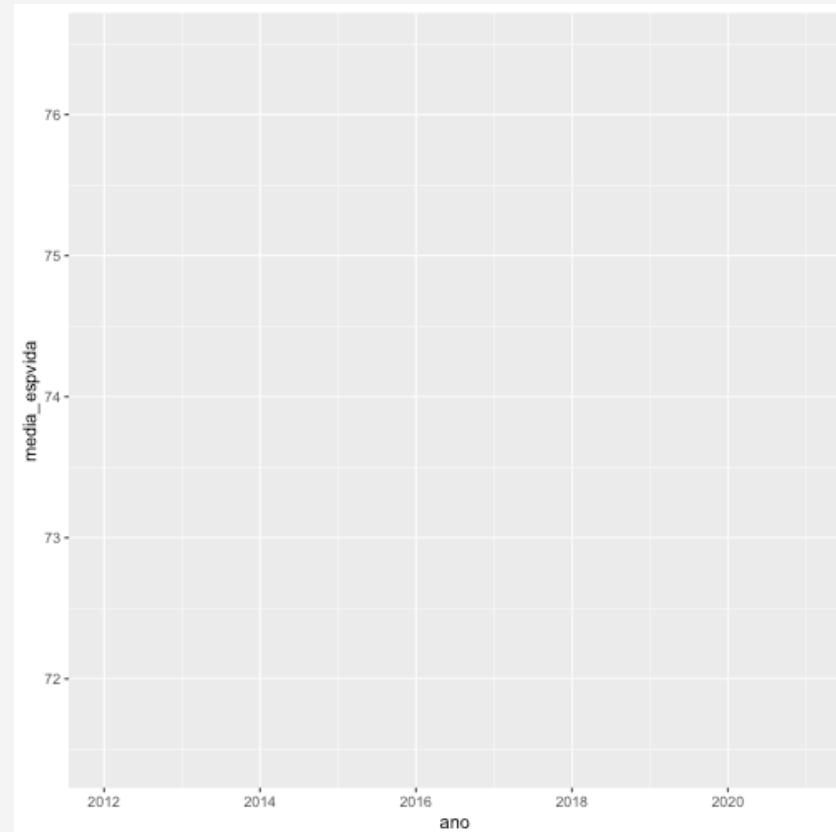
```
ggplot(dados2, aes(x = gini, y = espvida)) +  
  geom_point(size = 2, alpha = .5) +  
  ggpubr::stat_cor(  
    aes(label = after_stat(r.label)),  
    label.x = 0.6, label.y = 78, size = 3  
) +  
  labs(  
    x = "Índice de Gini",  
    y = "Expectativa de vida"  
)
```



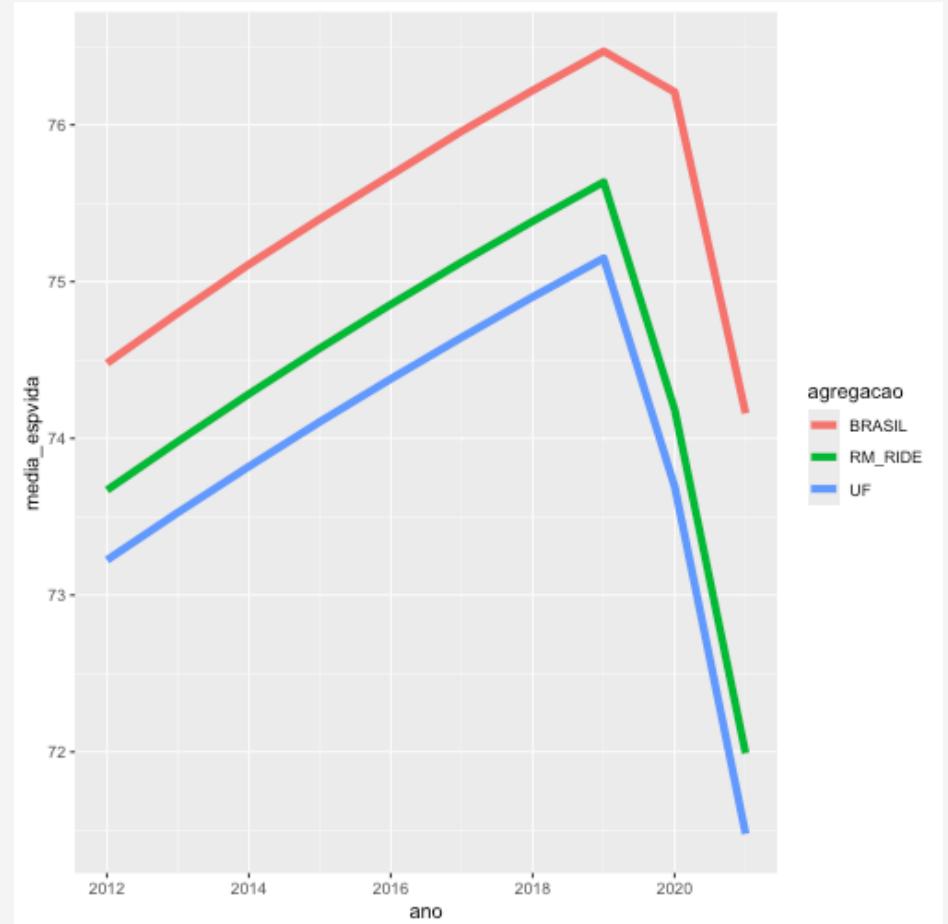
Gráficos bivariados (qualitativas x quantitativas)

Gráfico de linhas

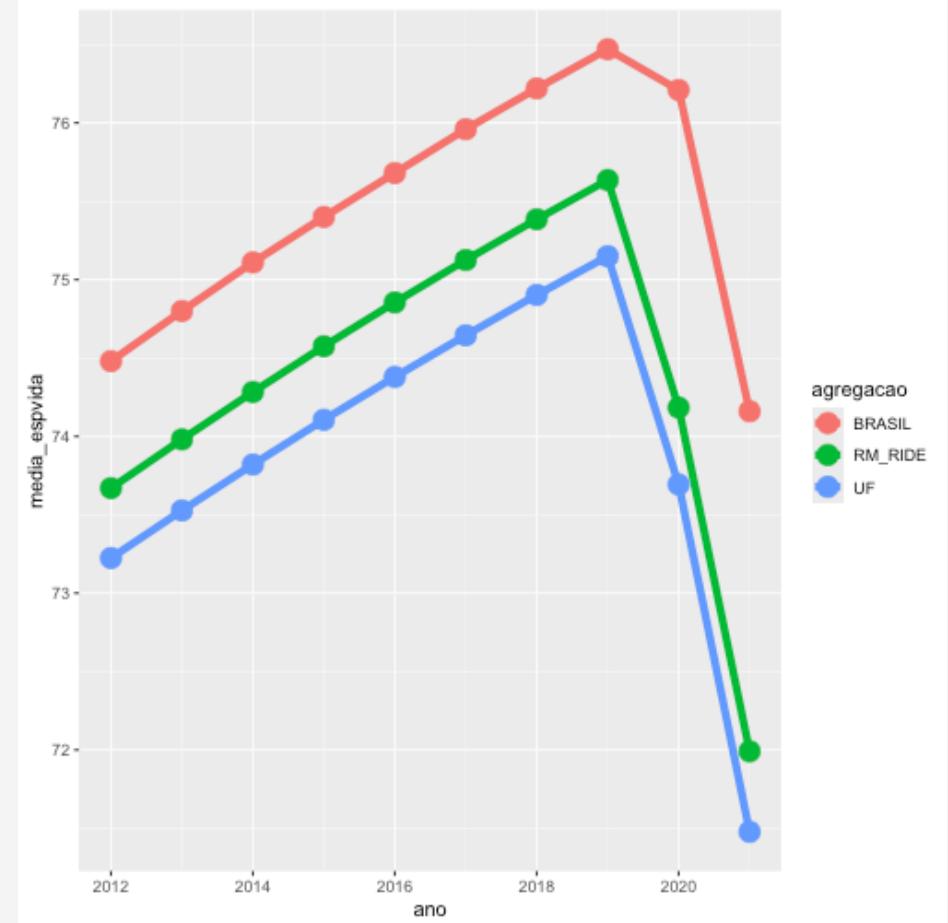
```
dados2 |>  
  dplyr::group_by(ano, agregacao) |>  
  dplyr::summarise(media_espvida = mean(  
    ggplot(aes(x = ano, y = media_espvida,
```



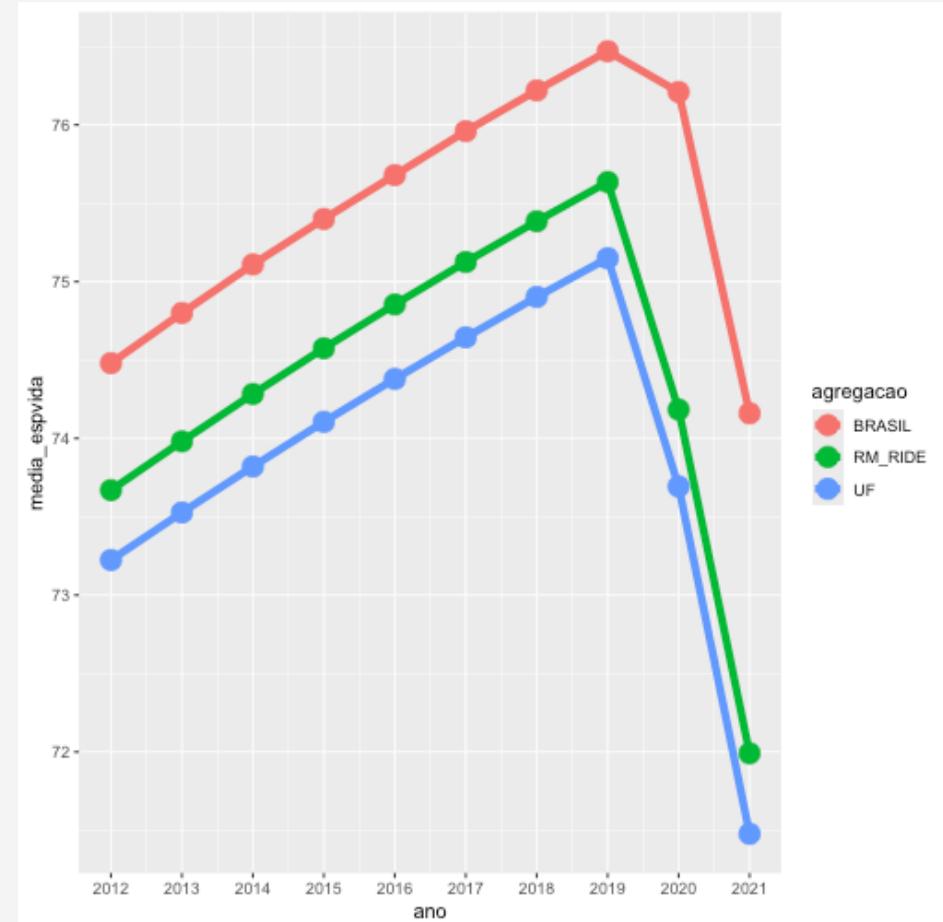
```
dados2 |>  
  dplyr::group_by(ano, agregacao) |>  
  dplyr::summarise(media_espvida = mean(  
  ggplot(aes(x = ano, y = media_espvida,  
    geom_line(linewidth = 2)
```



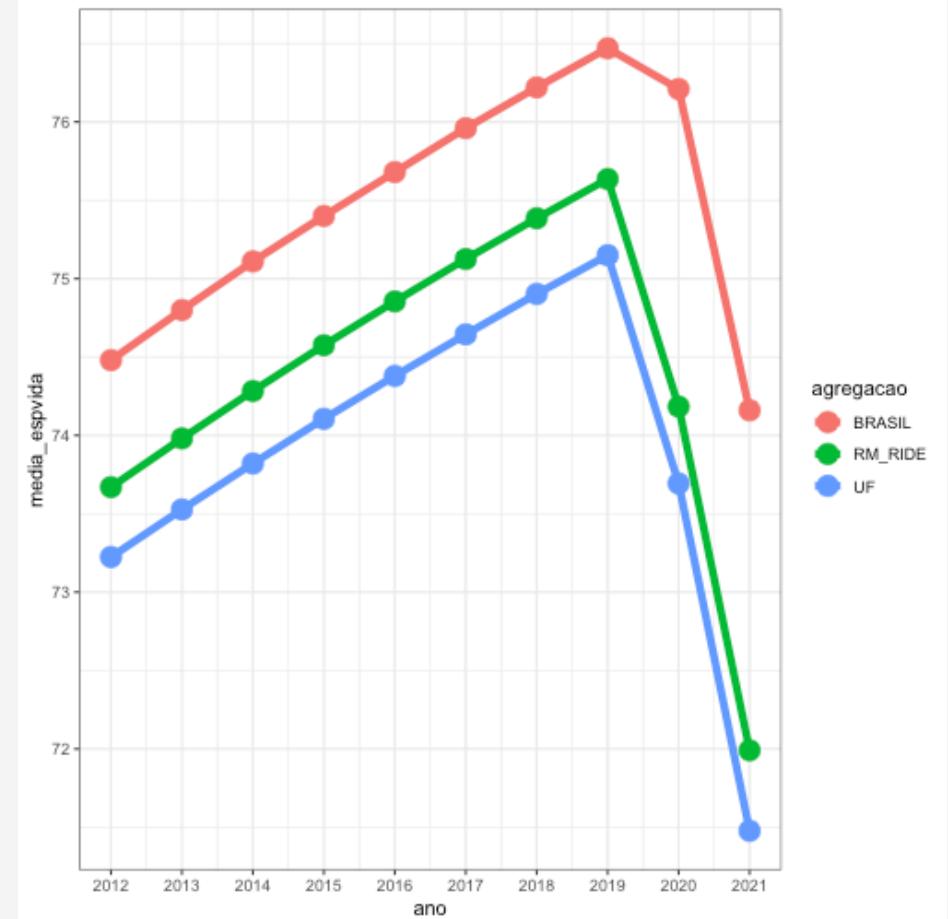
```
dados2 |>  
  dplyr::group_by(ano, agregacao) |>  
  dplyr::summarise(media_espvida = mean(  
    media_espvida),  
    .by = c(ano, agregacao)) |>  
  ggplot(aes(x = ano, y = media_espvida,  
    geom_line(linewidth = 2) +  
    geom_point(size = 5)
```



```
dados2 |>  
  dplyr::group_by(ano, agregacao) |>  
  dplyr::summarise(media_espvida = mean(  
    media_espvida), .by = c(ano, agregacao)) |>  
  ggplot(aes(x = ano, y = media_espvida, group = agregacao)) +  
    geom_line(linewidth = 2) +  
    geom_point(size = 5) +  
    scale_x_continuous(breaks = seq(2012, 2021))
```

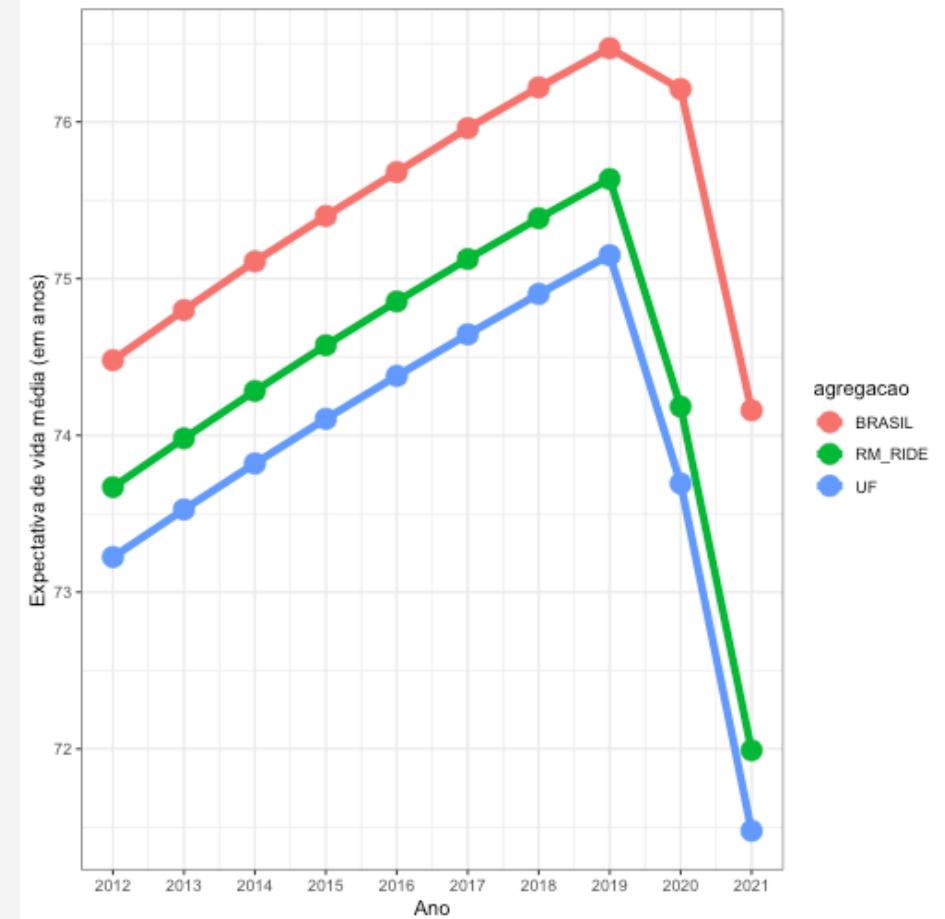


```
dados2 |>  
  dplyr::group_by(ano, agregacao) |>  
  dplyr::summarise(media_espvida = mean(  
    media_espvida, na.rm = TRUE), .by = c(ano, agregacao)) |>  
  ggplot(aes(x = ano, y = media_espvida, group = agregacao)) +  
    geom_line(linewidth = 2) +  
    geom_point(size = 5) +  
    scale_x_continuous(breaks = seq(2012, 2021)) +  
    theme_bw()
```



```

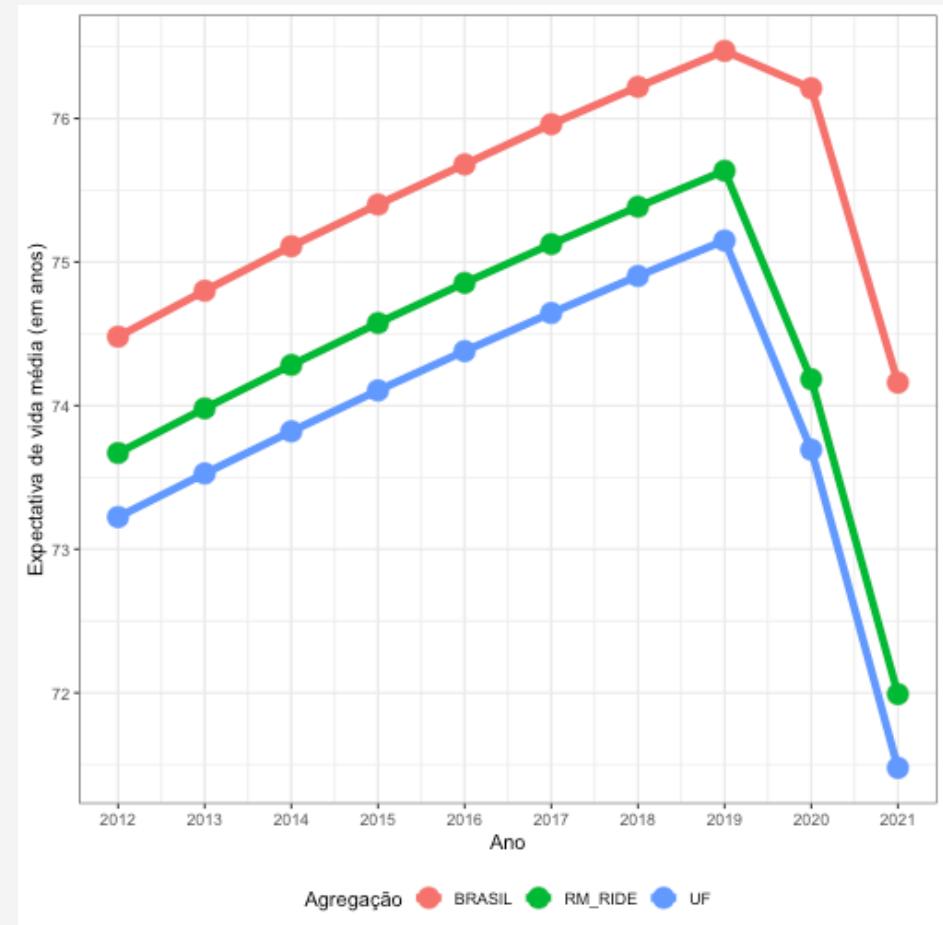
dados2 |>
  dplyr::group_by(ano, agregacao) |>
  dplyr::summarise(media_espvida = mean(
    espvida, na.rm = TRUE), .by_group = TRUE)
  ggplot(aes(x = ano, y = media_espvida,
             group = agregacao)) +
  geom_line(linewidth = 2) +
  geom_point(size = 5) +
  scale_x_continuous(breaks = seq(2012, 2021))
  theme_bw() +
  labs(
    x = "Ano",
    y = "Expectativa de vida média (em anos)",
    fill = "Agregação")
)
  
```



```

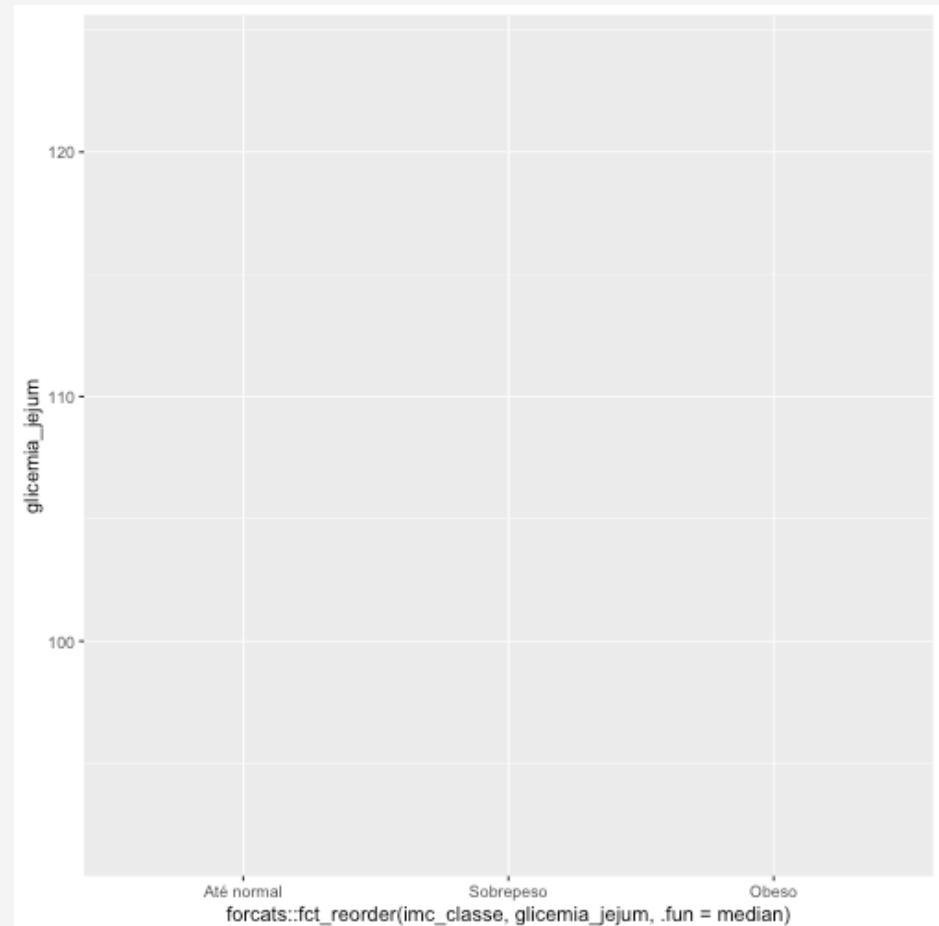
dados2 |>
  dplyr::group_by(ano, agregacao) |>
  dplyr::summarise(media_espvida = mean(
    espvida, na.rm = TRUE), .by_group = TRUE)
  ggplot(aes(x = ano, y = media_espvida, color = agregacao)) +
  geom_line(linewidth = 2) +
  geom_point(size = 5) +
  scale_x_continuous(breaks = seq(2012, 2021, 1))
  theme_bw() +
  labs(
    x = "Ano",
    y = "Expectativa de vida média (em anos)",
    color = "Agregação")
) +
  theme(legend.position = "bottom")

```

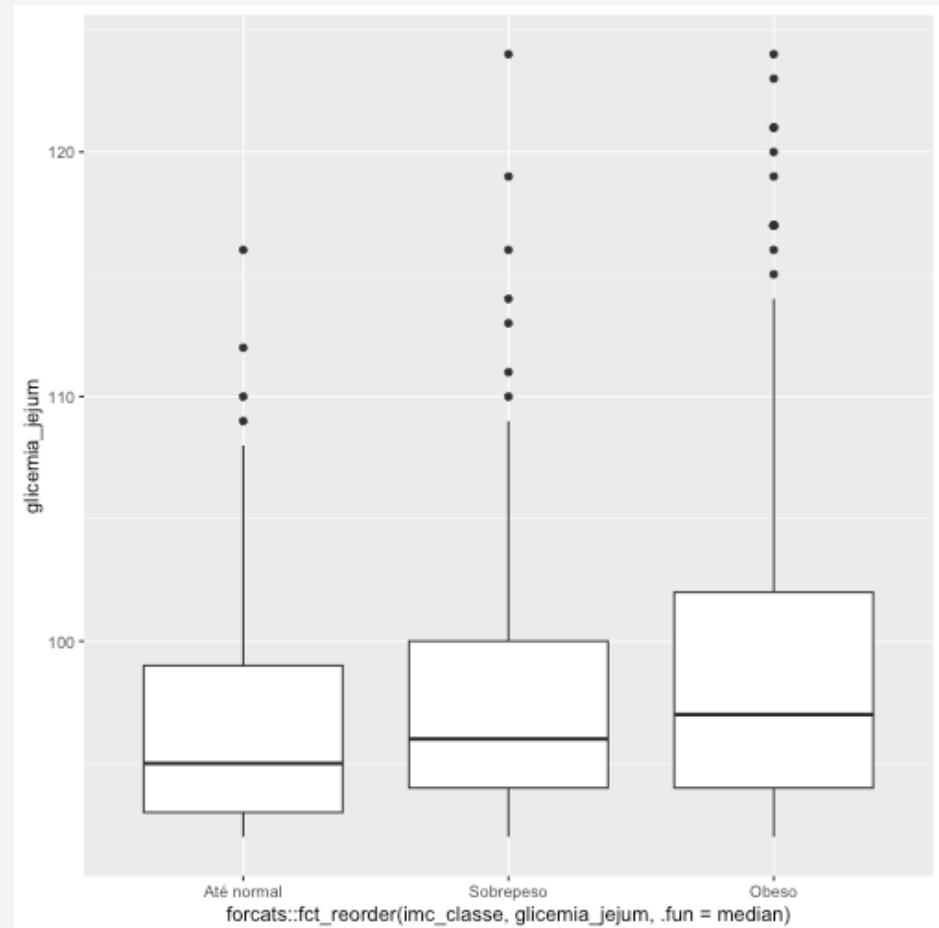


Boxplot

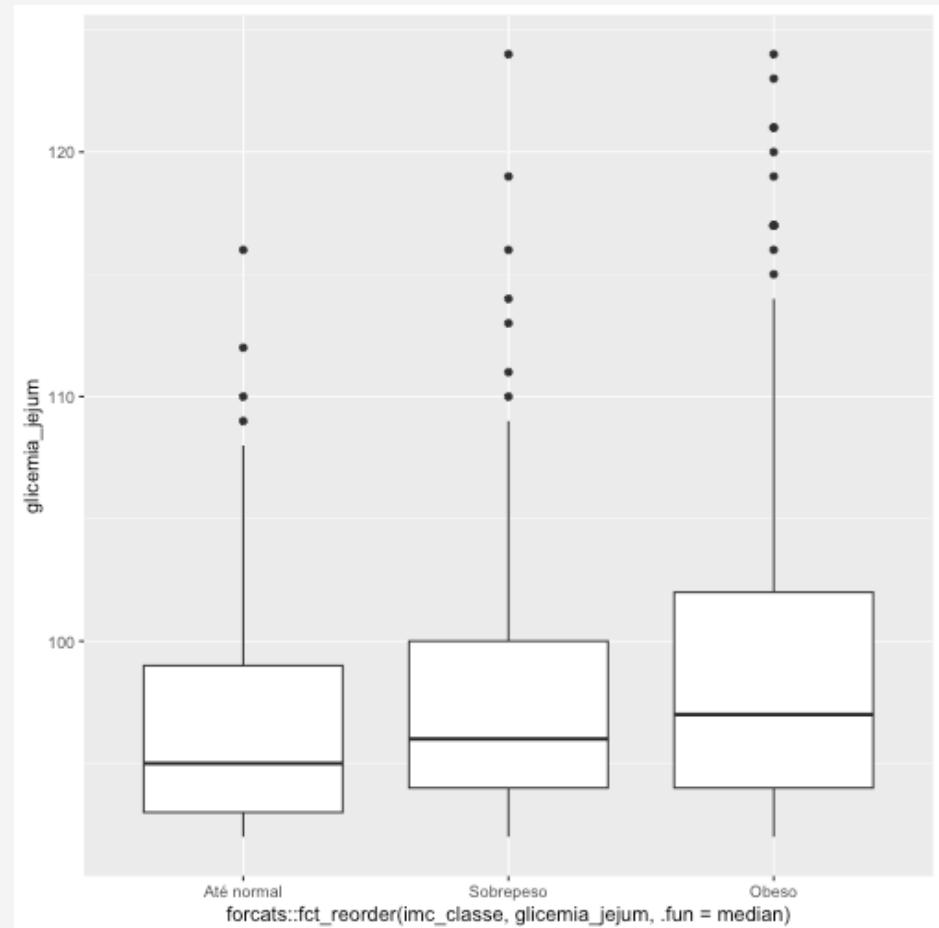
```
dados1 |>  
  ggplot(  
    aes(  
      x =forcats::fct_reorder(imc_classe,  
      y = glicemia_jejum  
    )  
  )
```



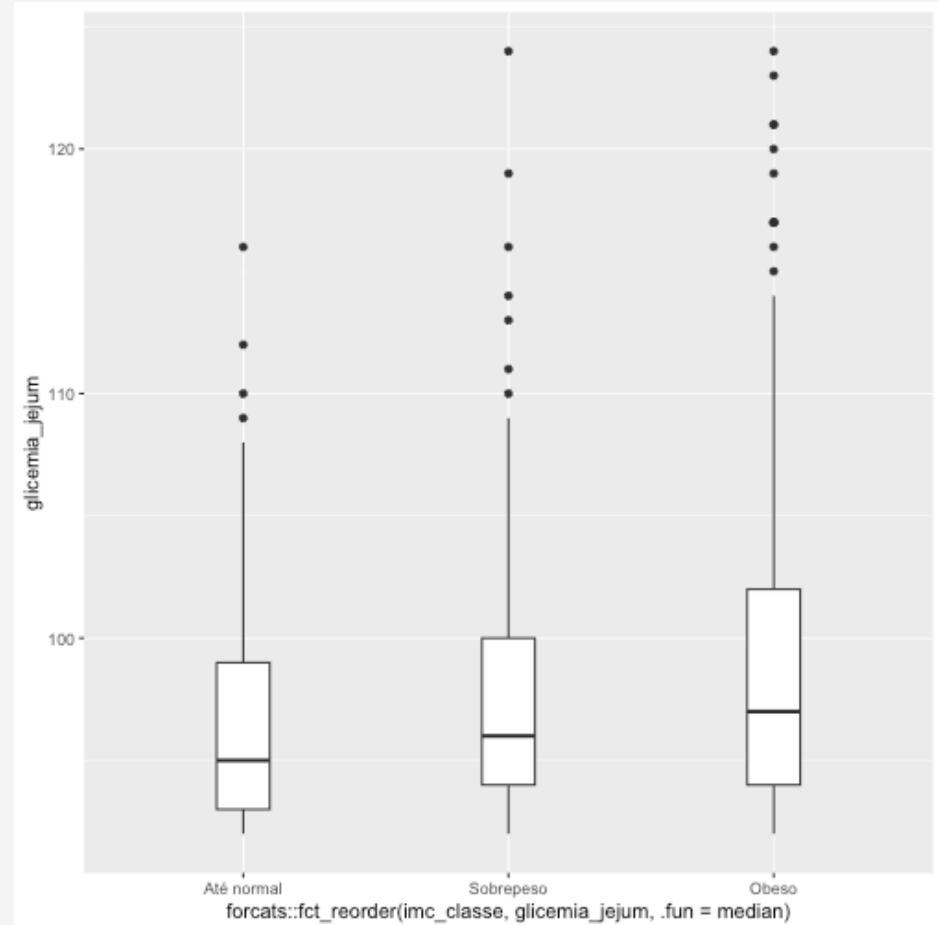
```
dados1 |>  
  ggplot(aes(x =forcats::fct_reorder(imc_classe, glicemia_jejum, .fun = median),  
             y = glicemia_jejum)) +  
  geom_boxplot()
```



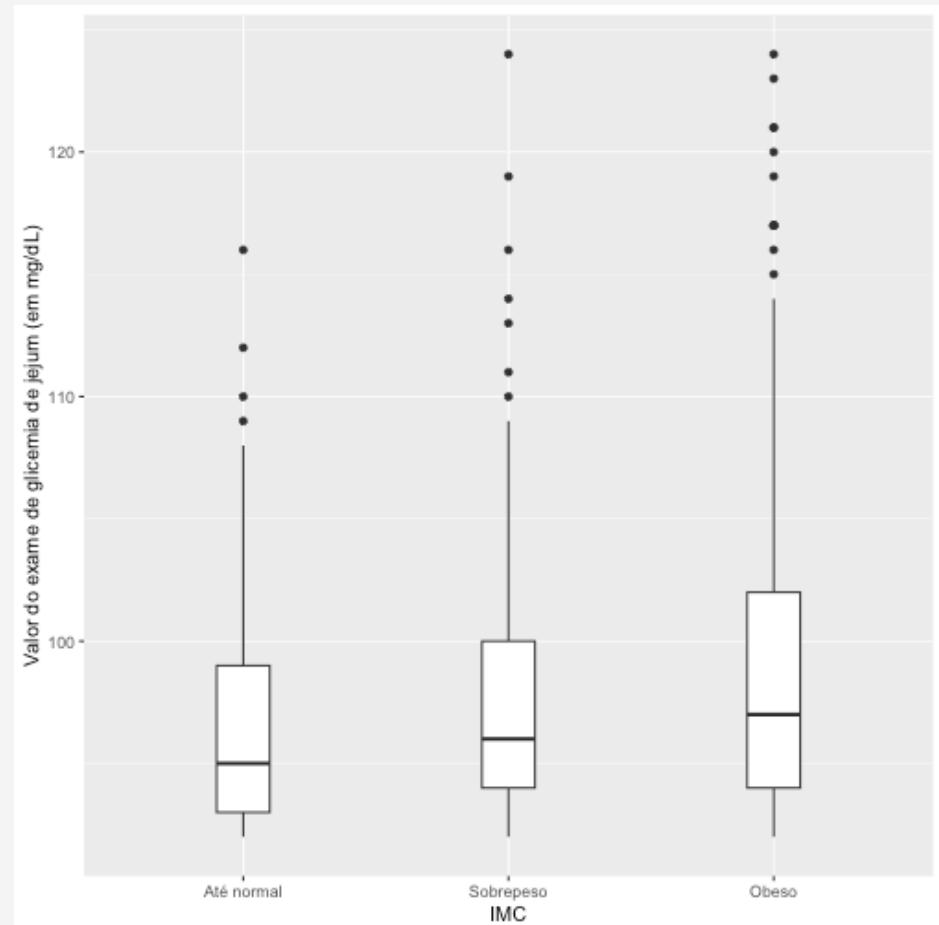
```
dados1 |>  
  ggplot(aes(x =forcats::fct_reorder(imc_classe, glicemia_jejum, .fun = median),  
             y = glicemia_jejum)) +  
  geom_boxplot()
```



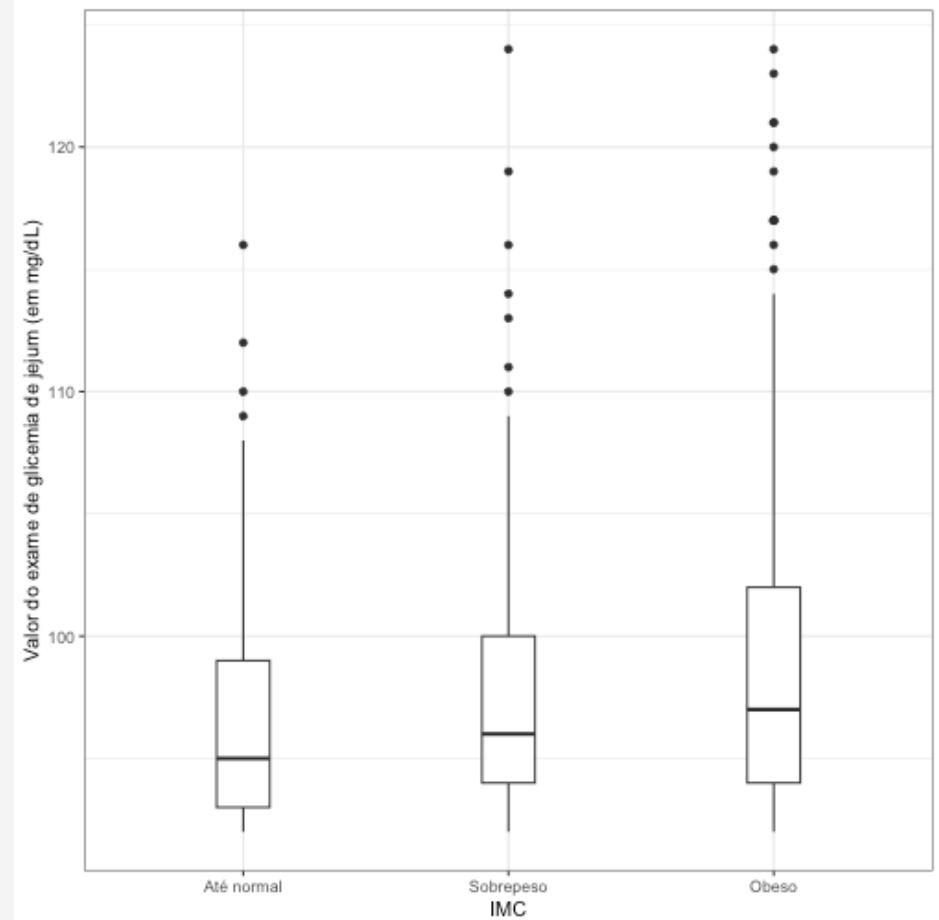
```
dados1 |>  
  ggplot(aes(x =forcats::fct_reorder(ir  
  geom_boxplot(width = .2)
```



```
dados1 |>  
  ggplot(aes(x =forcats::fct_reorder(imc, uso_insulina)),  
  geom_boxplot(width = .2) +  
  labs(  
    x = "IMC",  
    y = "Valor do exame de glicemia de jejum (em mg/dL)",  
    fill = "Usou insulina?",  
    title = "Glicemia de jejum por IMC")
```



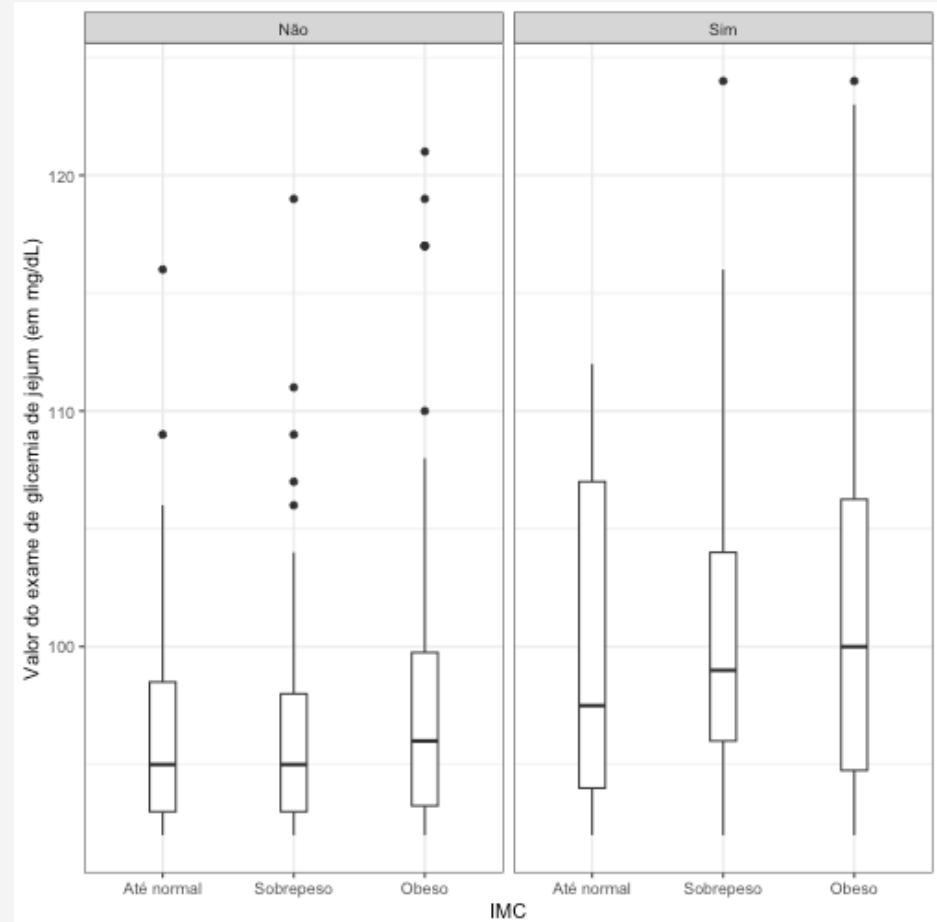
```
dados1 |>  
  ggplot(aes(x =forcats::fct_reorder(imc, glicemia),  
             geom_boxplot(width = .2) +  
             labs(  
               x = "IMC",  
               y = "Valor do exame de glicemia de jejum (em mg/dL)",  
               fill = "Usou insulina?",  
             ) +  
             theme_bw()
```



```

dados1 |>
  ggplot(aes(x =forcats::fct_reorder(imc,
  geom_boxplot(width = .2) +
  labs(
    x = "IMC",
    y = "Valor do exame de glicemia de jejum (em mg/dL)",
    fill = "Usou insulina?"
  ) +
  theme_bw() +
  facet_wrap(. ~ insulina)

```



Salvando um gráfico ggplot no R

- Vamos atribuir o gráfico de barras e o histogramas a objetos:

```
g1 <- ggplot(dados1, aes(x = imc_classe, y = after_stat(count)/sum(after_stat(count)))  
  geom_bar(color = "#0000cd", fill = "#ffffff") +  
  scale_y_continuous(labels = scales::percent) +  
  geom_text(stat = "count", aes(label = round(after_stat(count)/sum(after_stat(count)  
  labs(title = "Frequência do IMC categórico das gestantes", x = "IMC categórico", y
```



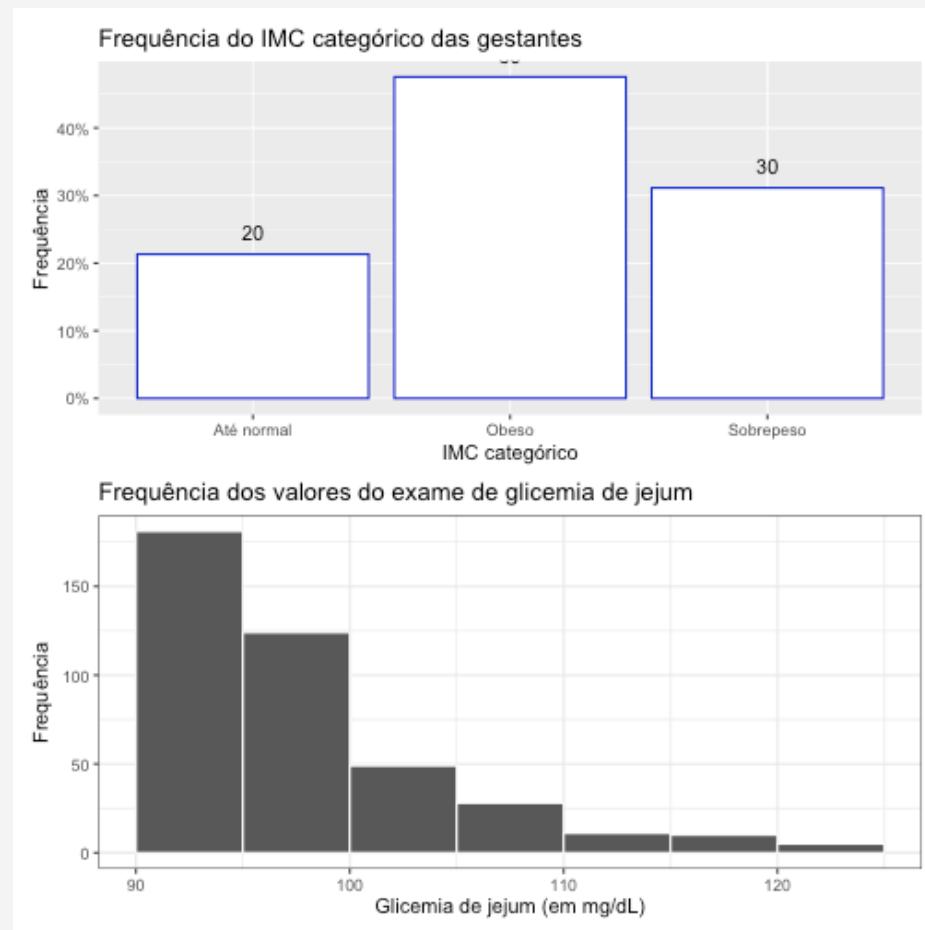
```
g2 <- ggplot(dados1, aes(x = glicemia_jejum)) +  
  geom_histogram(color = "#ffffff", breaks = seq(90, 125, 5)) +  
  labs(title = "Frequência dos valores do exame de glicemia de jejum", x = "Glicemia  
  theme_bw()
```

- Podemos salvá-los em uma única imagem. Fazemos isso com o pacote `{patchwork}`.

```
library(patchwork)
```

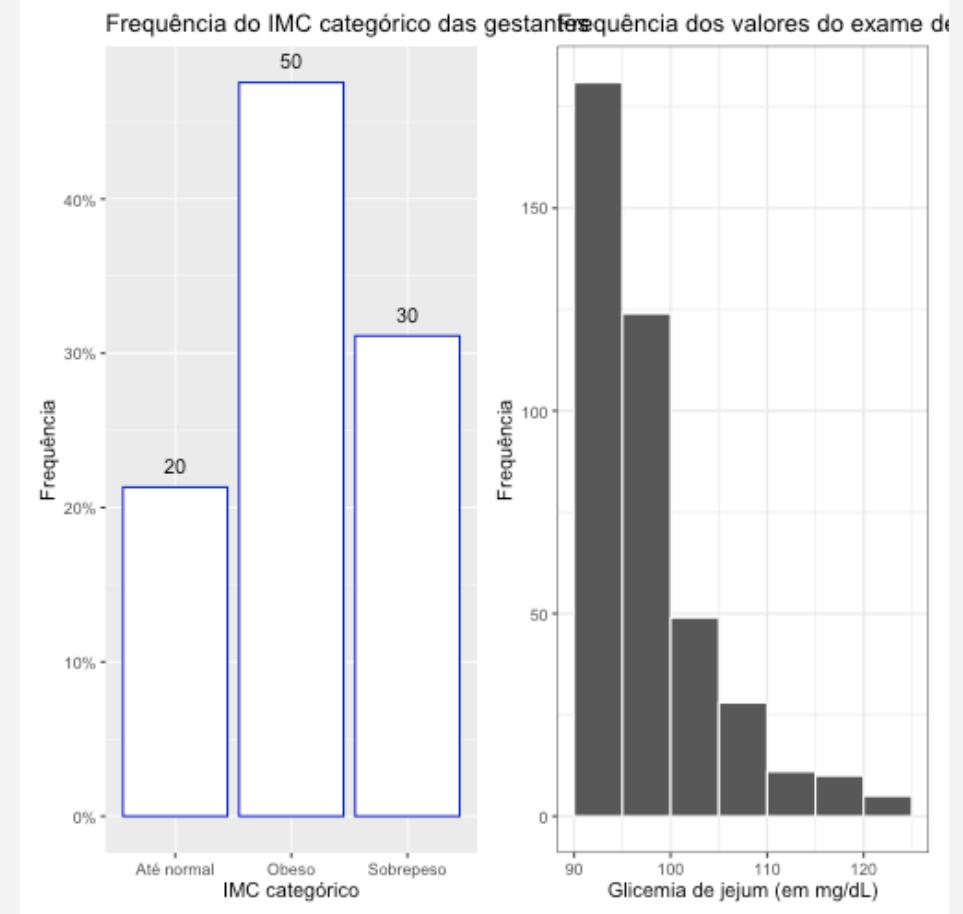
- Para o caso de dispor os gráficos um embaixo do outro, usa-se barra:

g1 / g2



- Já para o caso em que os gráficos fiquem lado a lado, usa-se sinal de adição ou barra vertical:

g1 + g2



- Agora sim, vamos salvar os gráficos! Fazemos isso com a função `ggsave()`:

```
ggsave("graficos/univariados.png", width = 16, height = 10)
```

- **Importante!** Por padrão, a função `ggsave()` salva o último gráfico que foi rodado em seu editor ou console.
- É possível salvar em vários formatos, como TEX, PDF, JPEG, TIFF, PNG e SVG.
- Por padrão, a imagem tem resolução 300dpi. Para alterá-la, use o argumento `dpi`.

Meu obrigada!



 ornscar@gmail.com

 [@ornscar](https://github.com/ornscar)

 [@ornscar](https://twitter.com/ornscar)