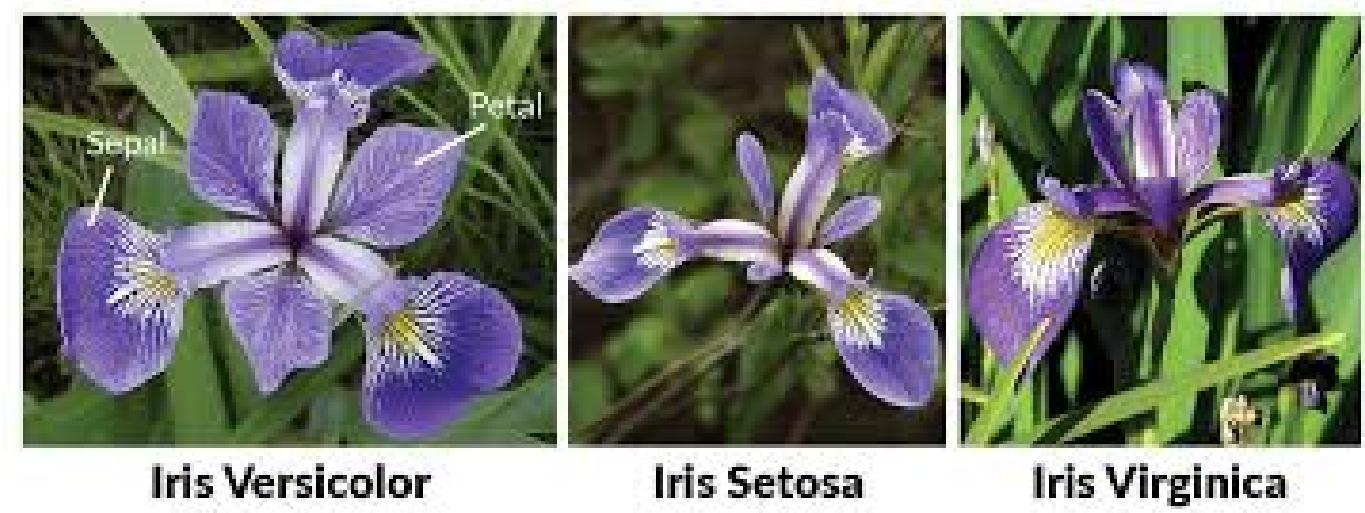
Descrição e visualização de dados multivariados utilizando a base Iris

Table of contents

1 Introdução

A base ***iris*** (FISHER (1936)) é um dos conjuntos de dados mais clássicos e didáticos da estatística. Ela contém **150 observações** (flores) de **três espécies**, *setosa*, *versicolor* e *virginica*, medidas em **quatro variáveis contínuas**:

- `Sepal.Length` (cm)
- `Sepal.Width` (cm)
- `Petal.Length` (cm)
- `Petal.Width` (cm)



Nosso objetivo é revisar **conceitos fundamentais de Estatística Multivariada**:

- Vetor de médias
- Matrizes de covariância e correlação
- Variância total e generalizada
- Matrizes de distância (euclidiana e de Mahalanobis)
- Visualizações (pares, correlograma, heatmaps)

---

2 Preparação e inspeção dos dados

```
# Pacotes necessários
# install.packages(c("tidyverse","GGally","ggcorrplot","pheatmap","factoextra", "MatchIt"))

library(tidyverse)
library(GGally)
library(ggcorrplot)
library(pheatmap)
library(factoextra)
library(MatchIt)

# Carregar a base iris
dados <- iris
glimpse(dados)
```

```
Rows: 150
Columns: 5
$ Sepal.Length <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.~
$ Sepal.Width  <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.~
$ Petal.Length <dbl> 1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5, 1.5, 1.~
$ Petal.Width  <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.2, 0.~
```

```
$ Species      <fct> setosa, setosa, setosa, setosa, setosa, setosa, setosa, s~
```

```r
summary(dados)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

```r
# Parte numérica e variável categórica
X <- as_tibble(dados[, 1:4])
y <- dados$Species
```

Interpretação: A base iris possui 150 observações e 5 variáveis, sendo quatro contínuas, `Sepal.Length`, `Sepal.Width`, `Petal.Length` e `Petal.Width` (em centímetros) e uma categórica, `Species`, com três espécies (`setosa`, `versicolor` e `virginica`). O resumo descritivo da base `iris` indica amostra balanceada por espécie (50 `setosa`, 50 `versicolor`, 50 `virginica`) e sugere padrões distintos entre sépalas e pétalas: `Sepal.Length` (4,3–7,9; Q1=5,1; mediana=5,8; média 5,84) e `Sepal.Width` (2,0–4,4; Q1=2,8; mediana=3,0; média 3,06) mostram variação moderada e distribuição quase simétrica (médias próximas às medianas), enquanto `Petal.Length` (1,0–6,9; Q1=1,6; mediana=4,35; média 3,76) e `Petal.Width` (0,1–2,5; Q1=0,3; mediana=1,3; média 1,20) exibem maior dispersão e assimetria à esquerda, refletindo a presença de muitas pétalas pequenas (típicas de `setosa`) e valores maiores nas demais espécies; isso cria um perfil "bimodal" nas pétalas, que costuma discriminar fortemente as espécies, ao passo que as medidas de sépala contribuem de forma mais moderada para a separação.

---

3 Vetor de médias

```r
# Vetor de médias global
media_global <- colMeans(X)
media_global
```

```
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
    5.843333     3.057333     3.758000     1.199333
```

```r
# Vetor de médias por espécie
media_por_especie <- X %>%
  mutate(Species = y) %>%
  group_by(Species) %>%
  summarise(across(everything(), mean), .groups = "drop")
media_por_especie
```

Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
virginica	6.588	2.974	5.552	2.026

Interpretação: o vetor de médias representa o "centro" da nuvem de pontos. Comparar médias por espécie ajuda a perceber separações entre grupos (por exemplo, pétalas maiores em `virginica`). Os valores mostram o padrão clássico da iris: `setosa` tem pétalas bem menores (PL=1,462; PW=0,246) e sépalas relativamente mais largas (SW=3,428), além de menor comprimento de sépala (SL=5,006); `virginica` apresenta as maiores pétalas (PL=5,552; PW=2,026) e o maior comprimento de sépala (SL=6,588), com largura de sépala intermediária (SW=2,974); `versicolor` fica entre as duas em todas as medidas (PL=4,260; PW=1,326) e tem a menor largura de sépala (SW=2,770).

---

4 Matrizes de covariância e correlação

```
# Matriz de covariâncias
S <- cov(X)
S
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063
```

Covariância: mede associação linear nas unidades originais. As variâncias (diagonal) indicam que a maior dispersão está em `Petal.Length` (3.1163), seguida de `Sepal.Length` (0.6857) e `Petal.Width` (0.5810), enquanto `Sepal.Width` varia menos (0.1900). Nos termos cruzados, há covariância positiva forte entre `Petal.Length` e `Petal.Width` (1.2956), e também entre `Sepal.Length` com `Petal.Length` (1.2743) e com `Petal.Width` (0.5163); já `Sepal.Width` apresenta covariâncias negativas com as demais (especialmente com `Petal.Length`: -0.3297).

```
# Matriz de correlações
R <- cor(X)
R
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Correlação: padroniza a covariância (-1 a 1), útil quando variáveis têm escalas diferentes. A matriz de correlações da iris mostra que as medidas de pétala são fortemente associadas: `Petal.Length` e `Petal.Width` têm correlação muito alta (0,963) e ambas se correlacionam bastante com `Sepal.Length` (0,872 e 0,818), indicando forte redundância informacional entre essas variáveis; já `Sepal.Width` se relaciona negativamente com as demais ($-0,118$ com `Sepal.Length`, $-0,428$ com `Petal.Length` e $-0,366$ com `Petal.Width`), sugerindo um padrão em direção oposta. Em síntese, as pétalas dominam a variação e separam melhor as espécies, enquanto `Sepal.Width` adiciona informação complementar (e contrária), com alerta para possível multicolinearidade entre as medidas de pétala.

---

5 Variância generalizada e total

```r
# Variância generalizada (determinante de S)
eigen(S)
```

```
eigen() decomposition
$values
[1] 4.22824171 0.24267075 0.07820950 0.02383509

$vectors
           [,1]        [,2]        [,3]       [,4]
[1,]  0.36138659 -0.65658877  0.58202985  0.3154872
[2,] -0.08452251 -0.73016143 -0.59791083 -0.3197231
[3,]  0.85667061  0.17337266 -0.07623608 -0.4798390
[4,]  0.35828920  0.07548102 -0.54583143  0.7536574
```

```r
VG <- det(S)
VG
```

```
[1] 0.00191273
```

VG (determinante): A variância generalizada 0,00191273 é o determinante da matriz de covariâncias **S** e mede o "volume" da dispersão conjunta; um valor tão pequeno indica forte colinearidade/redundância entre as variáveis, especialmente entre `Petal.Length` e `Petal.Width`, de modo que a nuvem de pontos fica "achatada" em algumas direções (um ou mais autovalores de **S** são pequenos, e o produto deles, o determinante, cai). Em termos práticos, isso sugere que as medidas de pétala carregam informação muito semelhante.

```r
# Variância total (traço de S)
VT <- sum(diag(S))
VT
```

```
[1] 4.572957
```

VT (traço): soma das variâncias marginais = dispersão total marginal. Ela mede a dispersão global do conjunto e é dependente de escala; padronizando as variáveis (z-scores), a variância total passaria a ser $p = 4$. Pela decomposição: `Petal.Length` responde por aproximadamente 68,1% da variância total, `Sepal.Length` responde por 15,0%, `Petal.Width` por 12,7% e `Sepal.Width` por 4,2%, confirmando que as medidas de pétala dominam a variabilidade.

---

6 Matrizes de distância

6.1 Distância Euclidiana

```r
D_euclid <- euclidean_dist(data = X)
as.matrix(D_euclid)[1:6, 1:6]
```

```
          1         2        3         4         5         6
1 0.0000000 0.5385165 0.509902 0.6480741 0.1414214 0.6164414
2 0.5385165 0.0000000 0.300000 0.3316625 0.6082763 1.0908712
3 0.5099020 0.3000000 0.000000 0.2449490 0.5099020 1.0862780
4 0.6480741 0.3316625 0.244949 0.0000000 0.6480741 1.1661904
5 0.1414214 0.6082763 0.509902 0.6480741 0.0000000 0.6164414
6 0.6164414 1.0908712 1.086278 1.1661904 0.6164414 0.0000000
```

Distância Euclidiana representa a distância geométrica entre dois pontos.

6.2 Distância de Karl Pearson

```
D_karlP <- scaled_euclidean_dist(data = X)
as.matrix(D_karlP)[1:6, 1:6]
```

```
          1         2         3         4         5         6
1 0.0000000 1.1722914 0.8427840 1.0999999 0.2592702 1.0349769
2 1.1722914 0.0000000 0.5216255 0.4325508 1.3818560 2.1739229
3 0.8427840 0.5216255 0.0000000 0.2829432 0.9882608 1.8477070
4 1.0999999 0.4325508 0.2829432 0.0000000 1.2459861 2.0937597
5 0.2592702 1.3818560 0.9882608 1.2459861 0.0000000 0.8971079
6 1.0349769 2.1739229 1.8477070 2.0937597 0.8971079 0.0000000
```

Distância de Karl Pearson leva em conta as diferenças de escala.

6.3 Distância de Mahalanobis

```
D_mahal <- mahalanobis_dist(data = X)
as.matrix(D_mahal)[1:6, 1:6]
```

```
          1         2         3         4         5         6
1 0.0000000 1.3544572 0.9687298 1.4057253 0.5899110 1.1382566
2 1.3544572 0.0000000 0.9697905 1.4527546 1.8106890 2.4484066
3 0.9687298 0.9697905 0.0000000 0.7170009 1.1253440 1.9227214
4 1.4057253 1.4527546 0.7170009 0.0000000 1.3293220 2.2425573
5 0.5899110 1.8106890 1.1253440 1.3293220 0.0000000 0.9446158
6 1.1382566 2.4484066 1.9227214 2.2425573 0.9446158 0.0000000
```

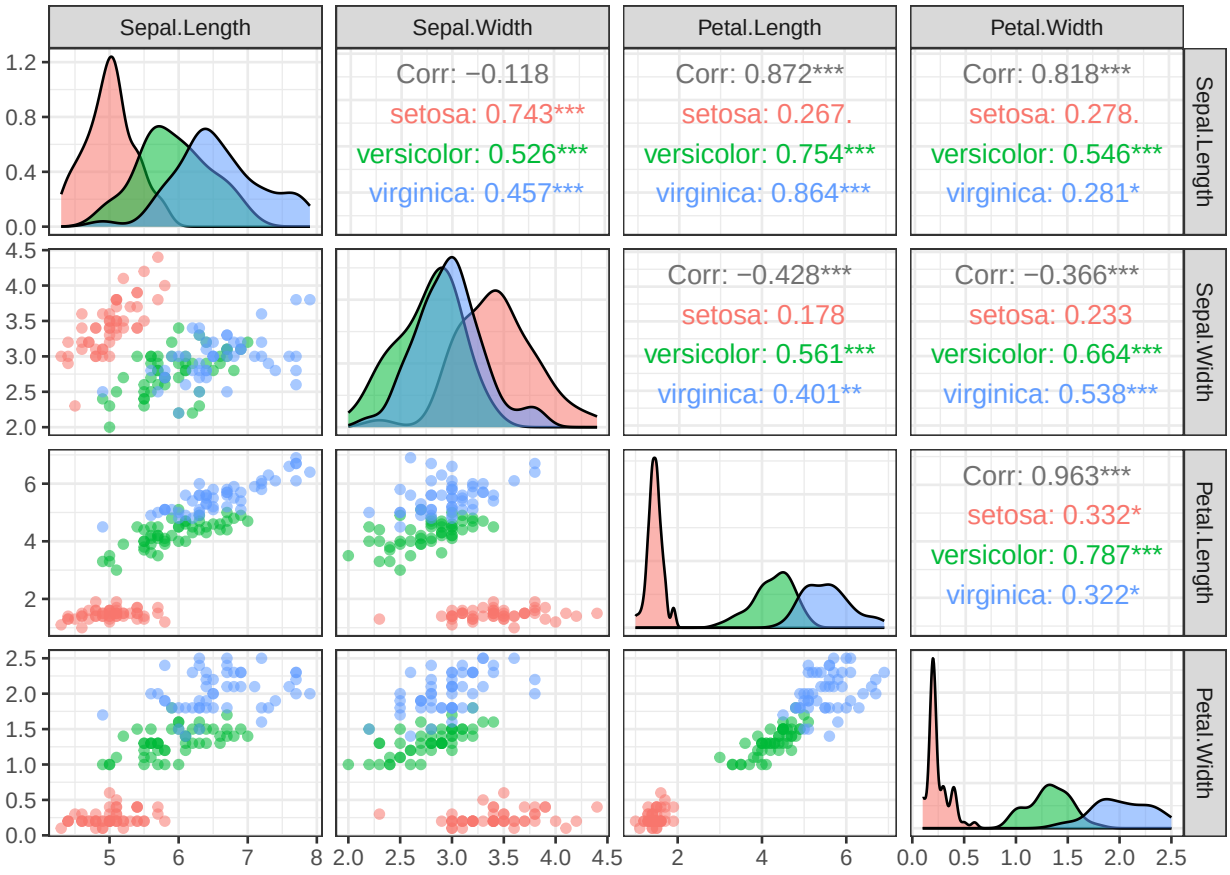
Distância de Mahalanobis leva em conta correlações entre variáveis e diferenças de escala.

---

7 Visualizações básicas

7.1 Gráfico de pares por espécie

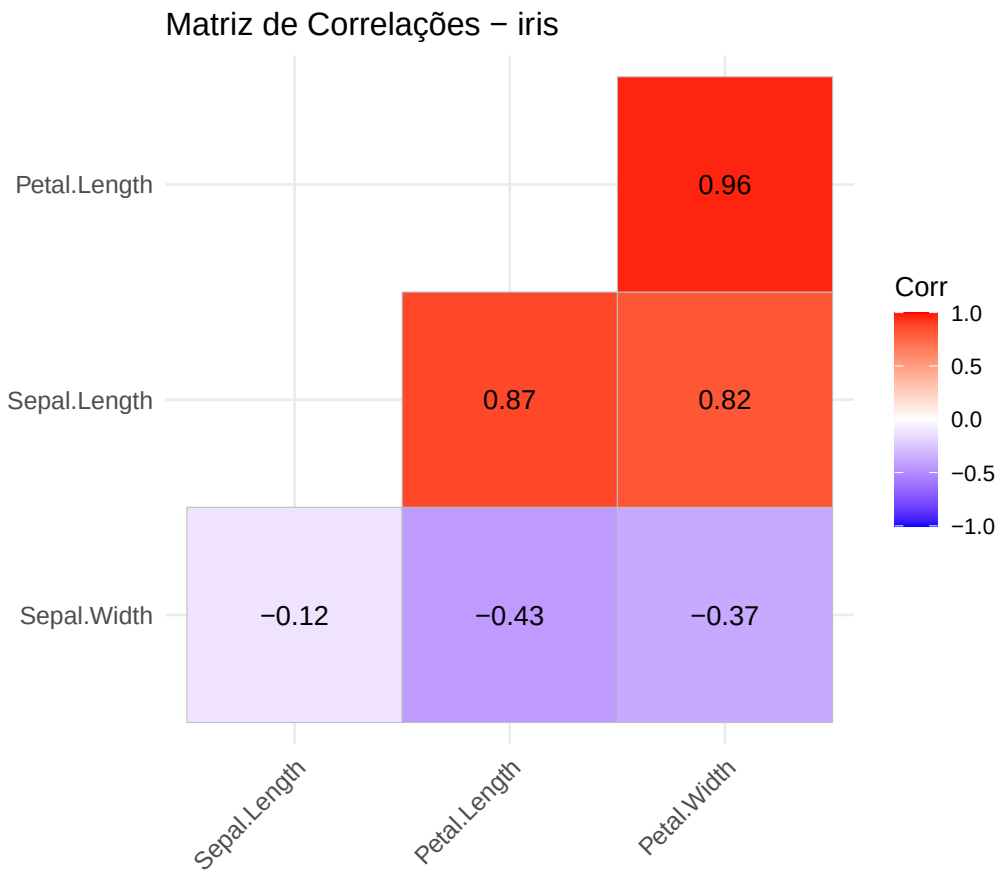
```
GGally::ggpairs(bind_cols(X, Species = y),
                columns = 1:4, aes(color = Species, alpha = 0.8)) + theme_bw()
```



Interpretação: O "pairs plot" da iris mostra que `setosa` tem pétalas muito pequenas e sépalas mais largas, enquanto `versicolor` e `virginica` apresentam pétalas maiores (com alguma sobreposição), e em `Sepal.Width` a espécie `setosa` desloca-se à direita, `versicolor` à esquerda e `virginica` fica intermediária. As correlações globais destacam `Petal.Length` *vs.* `Petal.Width` como fortíssima (0,96) e `Sepal.Length` bem associado às pétalas; já `Sepal.Width` aparece negativamente correlacionado no agregado. Contudo, por espécie as relações com `Sepal.Width` tendem a ser positivas, e os dispersogramas envolvendo medidas de pétala exibem a melhor separação entre espécies. Em síntese, as pétalas dominam a estrutura e a discriminação, com `Sepal.Width` fornecendo informação complementar.

7.2 Correlograma

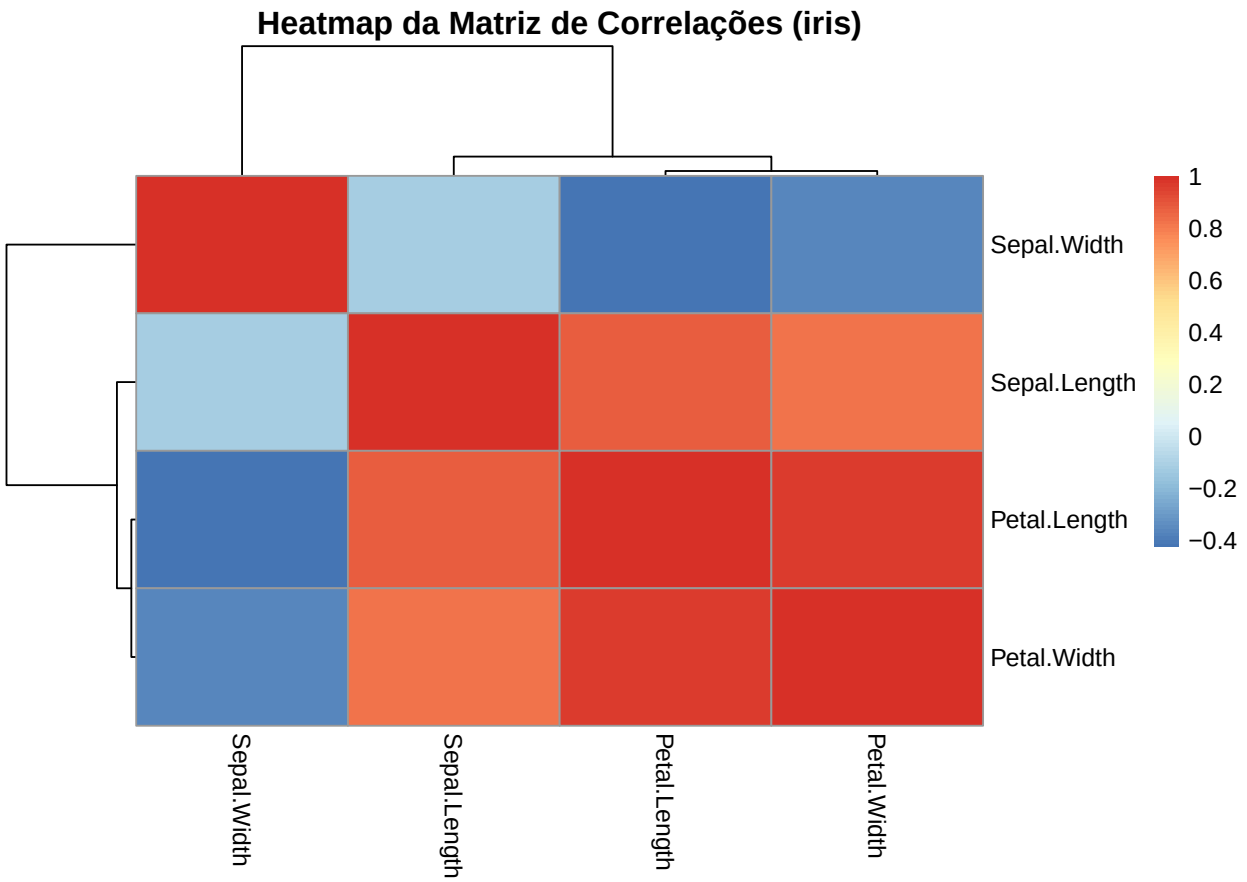
```
ggcorrplot(R, hc.order = TRUE, type = "lower",
           lab = TRUE, tl.cex = 10,
           title = "Matriz de Correlações - iris")
```



Interpretação: O mapa de correlações da iris evidencia três padrões centrais: (1) forte associação positiva entre as medidas de pétala: `Petal.Length` *vs.* `Petal.Width` 0,96 e também de `Sepal.Length` com as pétalas (0,87 e 0,82), indicando redundância informacional e provável multicolinearidade; (2) `Sepal.Width` apresenta correlações negativas com as demais variáveis (−0,12 com `Sepal.Length`, −0,43 com `Petal.Length` e −0,37 com `Petal.Width`), sugerindo um eixo de variação em sentido oposto ao das pétalas; e (3) como consequência, em tarefas de PCA ou classificação, as pétalas tendem a dominar a separação entre espécies, enquanto `Sepal.Width` adiciona sinal complementar.

7.3 Heatmap das correlações

```
pheatmap(R, cluster_rows = TRUE, cluster_cols = TRUE,
         main = "Heatmap da Matriz de Correlações (iris)")
```



Interpretação: O heatmap das correlações da iris confirma dois blocos de variáveis: (i) `Petal.Length` e `Petal.Width` fortemente positivas entre si (vermelho intenso) e também bem alinhadas com `Sepal.Length` (vermelho), formando um grupo altamente correlacionado que indica redundância e tende a dominar a variação; (ii) `Sepal.Width` aparece em azul frente às demais, mostrando correlação negativa moderada, o que a isola no dendrograma e sugere um eixo complementar de informação. Em termos práticos, as pétalas são as melhores para discriminar espécies (e podem sofrer multicolinearidade), enquanto `Sepal.Width` acrescenta sinal em direção oposta.

8 Referências Bibliográficas

FISHER, R. A. 1936. "THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS." *Annals of Eugenics* 7 (2): 179–88. https://doi.org/https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.