CIÊNCIA DE DADOS

INTRODUÇÃO

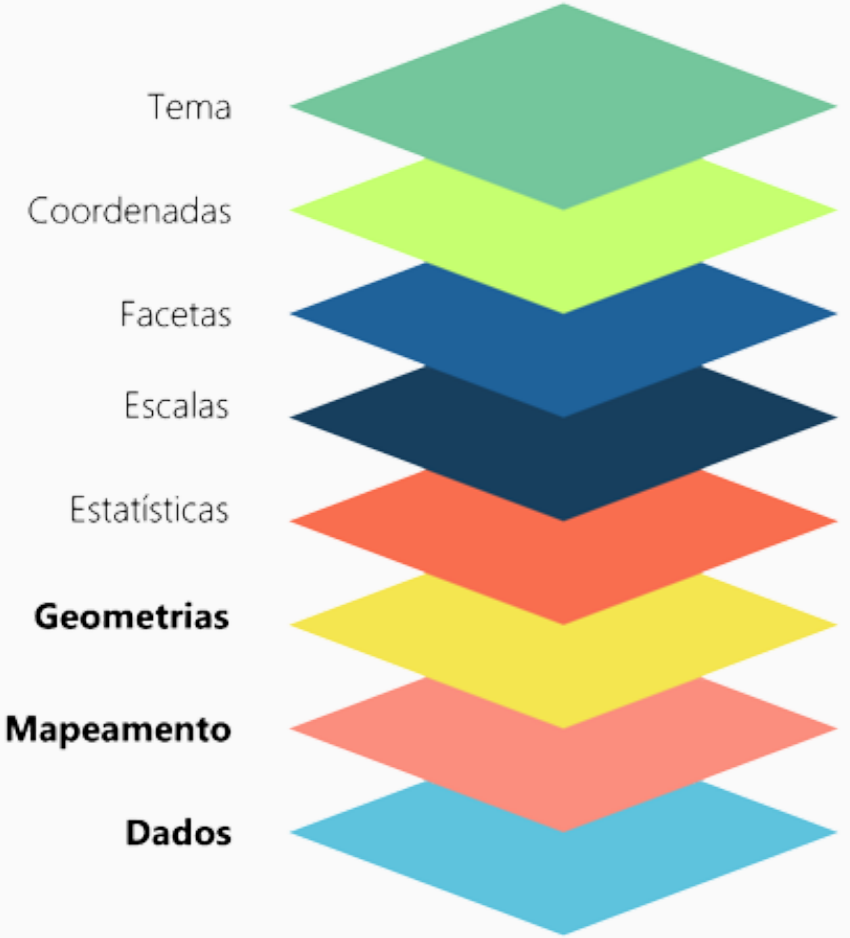
Universidade Federal de Santa Maria - UFSM
Prof. Moizés da Silva Melo

Pacote **ggplot2**

- O **ggplot2** é um pacote do R desenvolvido por Hadley Wickham em 2007 para construção de gráficos.
- Foi baseado no livro **"The Grammar of Graphics"** (Leland Wilkinson, 2005).
- Os gráficos criados com este pacote são, em geral, mais bonitos do que os gráficos do R base.
- É mais simples deixar o gráfico do jeito que desejamos.
- O **ggplot2** faz parte da coleção de pacotes **tidyverse**, que foi projetada para ciência de dados.

Sintaxe Básica

- A idea do **ggplot2** é construir um gráfico camada por camada.
- Cada gráfico é formado por três camadas fundamentais: **dados**, **mapeamento estético** e **geometria**.
- E outras camadas opcionais, que permitem personalizar o gráfico: **estatísticas**, **escalas**, **facetas**, **coordenadas** e **tema.**



Sintaxe Básica

Componente	Função	Explicação
Dados	**ggplot(data)**	**Os dados que serão utilizados (data frame).**
Estética	**aes()**	**Consiste no mapeamento as variáveis que serão utilizadas em Atributos estéticos (posição, cor, forma, tamanho).**
Geometria	**geom_*()**	**Camada onde definimos o formato que os dados serão apresentados (pontos, linhas, barras etc).**
Estatística	stat_*()	Nessa camada é possível aplicar transformações estatísticas aos dados.
Coordenadas	coord_*()	Defini o sistema de coordenadas do gráfico (cartesiano, polar, mapa etc.).
Facetas	facet_*()	Nessa camada podemos dividir o gráfico em facetas a partir das variáveis.
Escala	scale_*()	Permite modificar as propriedades referentes às escalas do gráfico.
Tema	theme() and theme_*()	Permite customizar o gráfico.

Dados

ggplot(data = *)

Dados

- Para utilizar as funções disponíveis no pacote **ggplot2** é necessário que os dados estejam armazenados em um **data frame**.
- Um **data frame** é uma estrutura bidimensional utilizada para armazenar bases de dados, na qual cada coluna corresponde a uma variável e as linhas representam as observações.
- Vamos usar como exemplo o conjunto de dados `iris`, que faz parte do pacote `datasets`.

```
iris %>%
  head(n=9)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
7          4.6         3.4          1.4         0.3  setosa
8          5.0         3.4          1.5         0.2  setosa
9          4.4         2.9          1.4         0.2  setosa
```

Estética - mapeamento da variáveis

aes()

Mapeamento Estético

O mapeamento das variáveis é feito a partir da função aes()

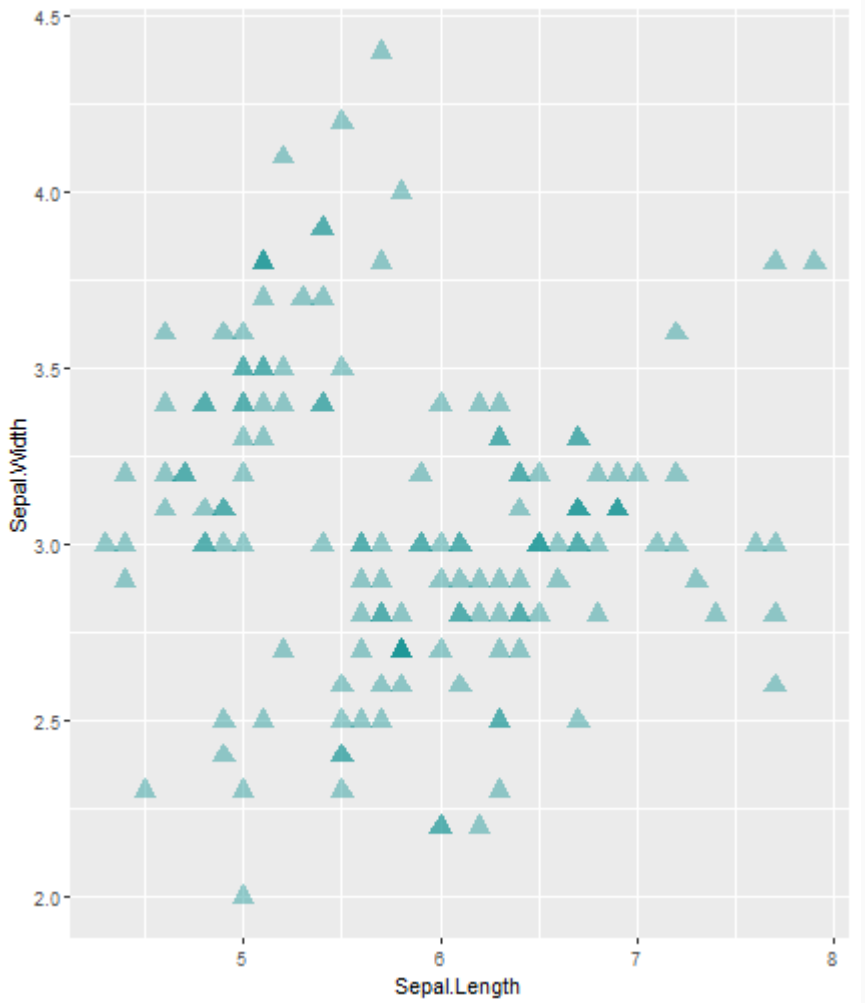
- Posição (x e y)
- Cor (color)
- Preenchimento (fill)
- Tamanho (size)
- Forma (shape)
- Transparência (alpha)
- Grupos (group)
- Texto (text)

- **Link: Especificações dos elementos estéticos**
- **Link: Cores predefinidas no R**
- **Link: Esquema de cores html - site 1**
- **Link: Esquema de cores html - site 2**

Geometria do gráfico

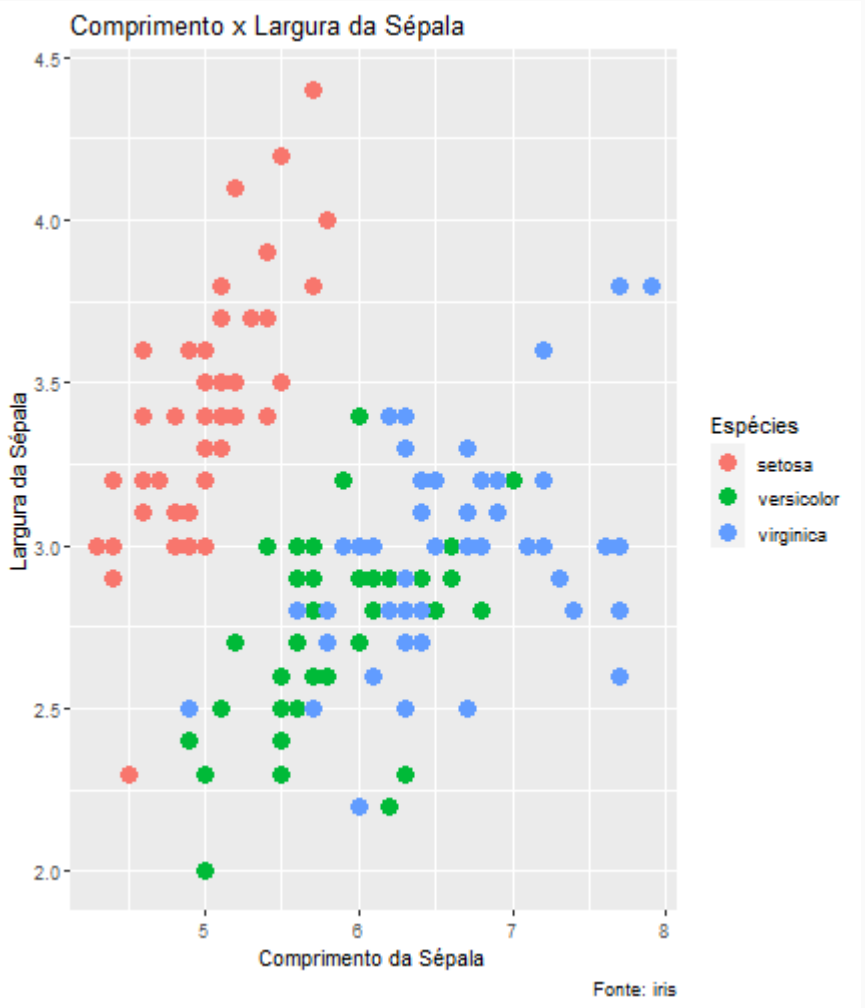
geom_*()

```r
ggplot(data = iris) +
  aes(x = Sepal.Length, y = Sepal.Width) +
  geom_point(
    size = 4,
    color = "darkcyan",
    alpha = .4,
    shape = 17
  )
```



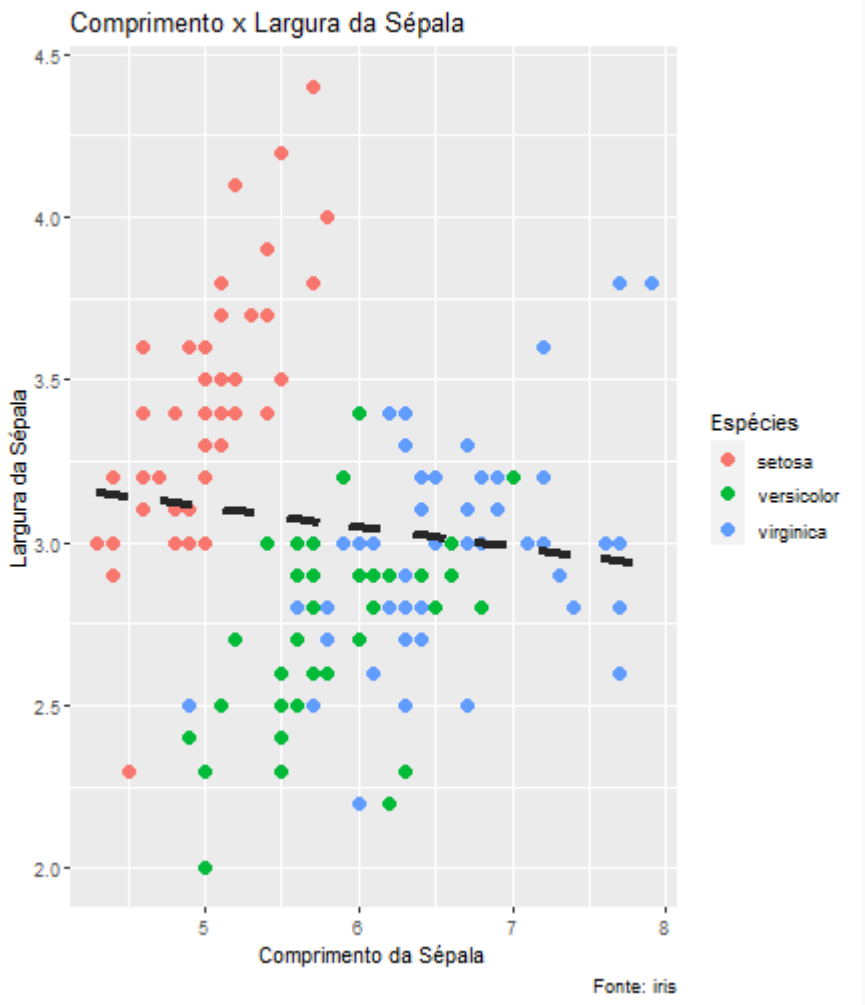
```
ggplot(data = iris,
       aes(x = Sepal.Length, y = Sepal.Width) ) +
  geom_point(size =4) +
  geom_point(aes(color = Sepal.Width > 4), size =4)
```

```r
iris %>%
  ggplot(aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point(
    aes(color = Species),
    size = 4
    ) +
  labs(
    x = "Comprimento da Sépala",
    y = "Largura da Sépala",
    title = "Comprimento x Largura da Sépala",
    color = "Espécies",
    caption = "Fonte: iris"
  )
```



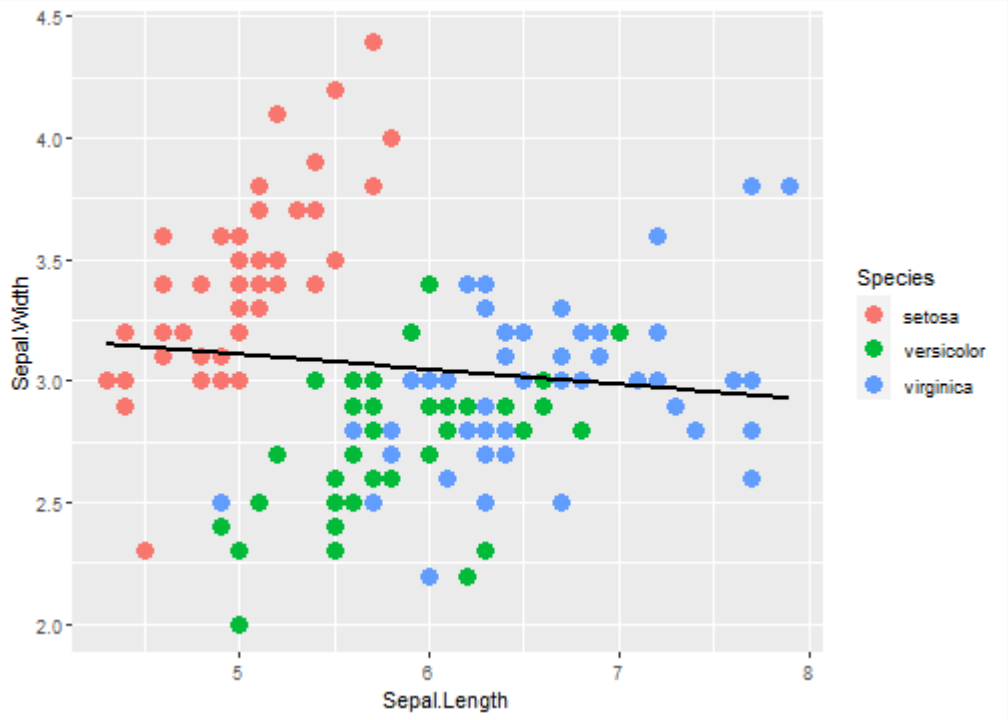
Camada estatística - stat_*()

```r
iris %>%
  ggplot(
    aes(x = Sepal.Length, y = Sepal.Width)
  ) +
  geom_point(
    aes(color = Species),
    size = 3
  ) +
  labs(
    x = "Comprimento da Sépala",
    y = "Largura da Sépala",
    title = "Comprimento x Largura da Sépala",
    color = "Espécies",
    caption = "Fonte: iris"
  ) +
  stat_smooth(
    method = "lm",
    se = FALSE,
    size = 1.5,
    color = "#262626",
    linetype = 2
  )
```

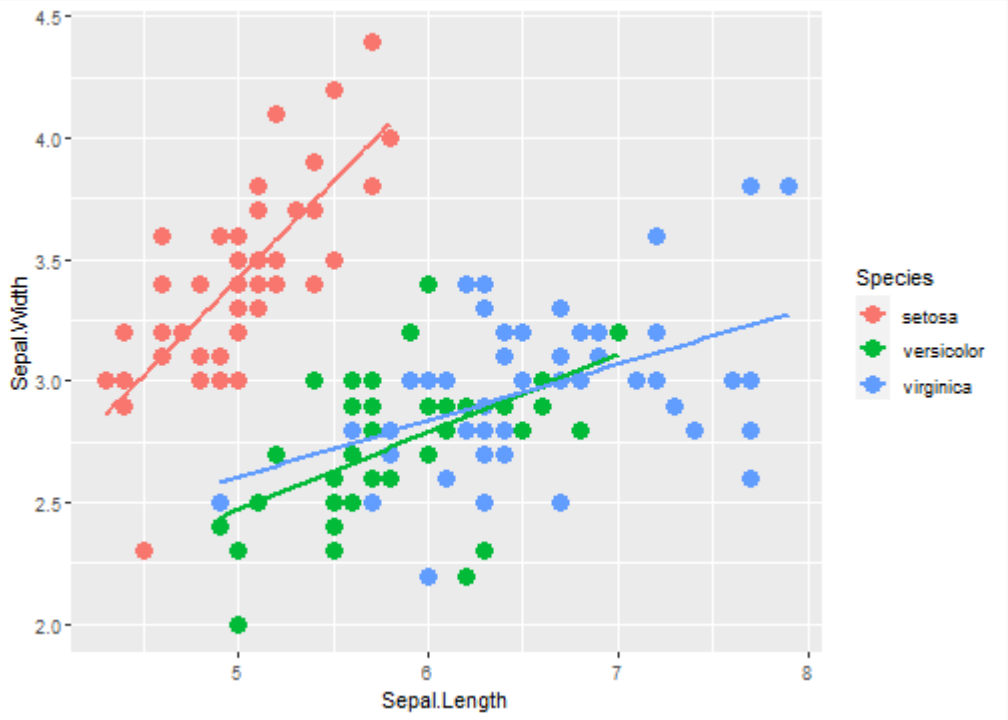


color versus aes(color)

```
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point(aes(color = Species), size = 4) +
  stat_smooth(color = "black", method = "lm",
              se = FALSE)
```




```
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point(aes(color = Species), size = 4) +
  stat_smooth(aes(color = Species), method = "lm",
              se = FALSE)
```



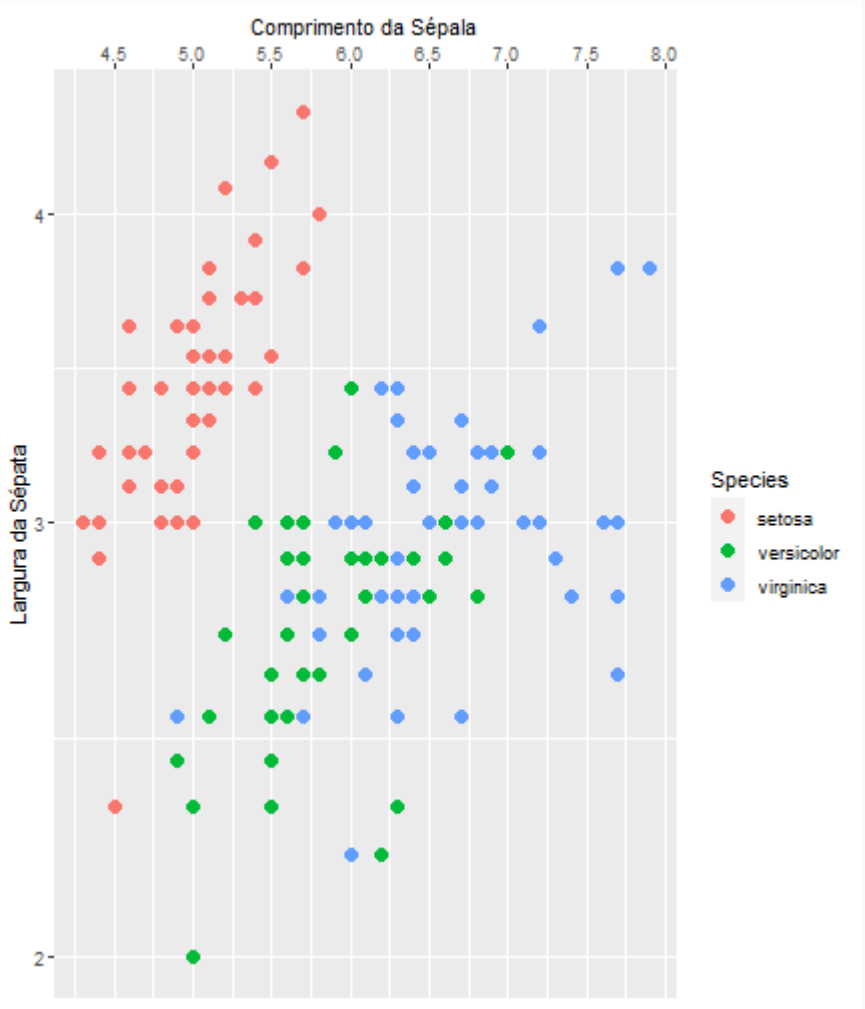
Escala - **scale_*()**

Escalas

- A camada **escala** pode ser utilizada para controlar as propriedades de todos os atributos estéticos mapeados para os dados.
- Esse camada é definida pela função `scale_elemento estético_tipo da escala()`.
- Há uma função diferente para cada elemento estético:

Estética	Função
Posições	`scale_x_*()` e `scale_y_*()`
Cores	`scale_color_*()` e `scale_fill_*()`
Tamanhos	`scale_size_*()` e `scale_radius_*()`
Formas	`scale_shape_*()` e `scale_linetype_*()`
Transparência	`scale_alpha_*()`

```
ggplot(
  data = iris,
  mapping = aes(x = Sepal.Length, y = Sepal.Width)
) +
  geom_point(
    mapping = aes(color = Species),
    size = 3
  ) +
  scale_x_continuous(
    name = "Comprimento da Sépala",
    breaks = seq(4.5, 8, 0.5),
    position = "top"
  ) +
  scale_y_continuous(
    name = "Largura da Sépata",
    trans = "log10"
  )
```

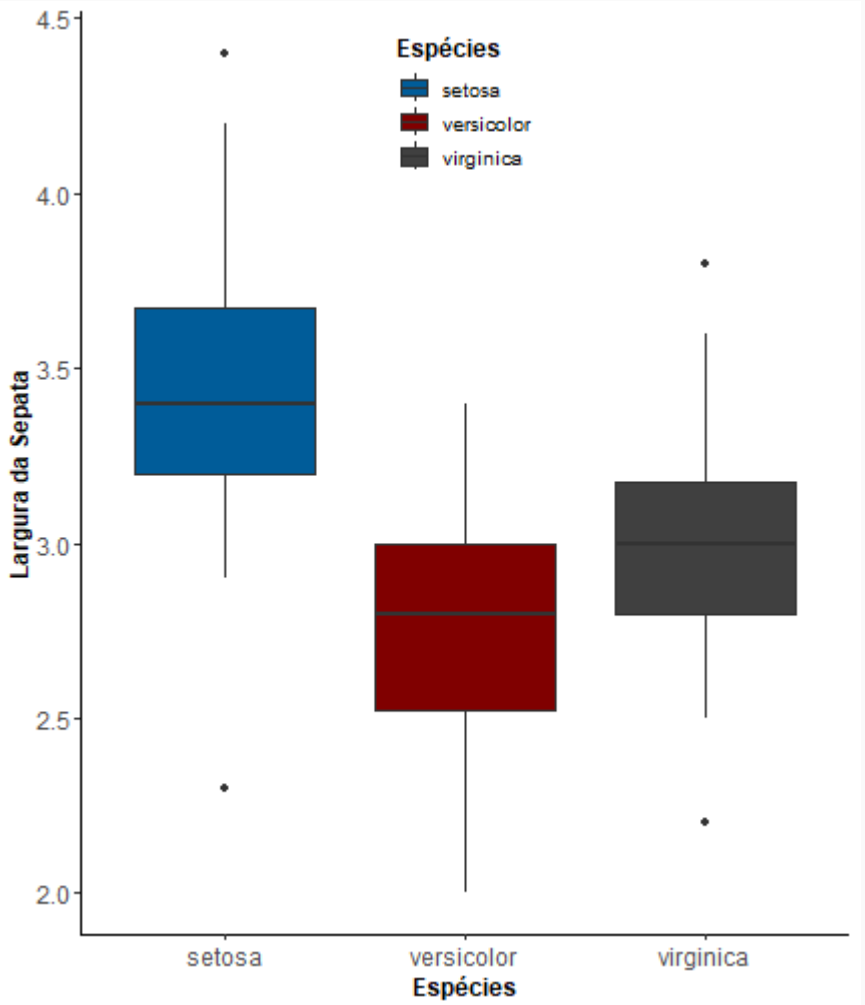


Temas - **theme_*()**

Temas

- Os gráficos criados a partir da função `ggplot()` têm um aspecto visual padrão.
- Podemos mudar esse aspecto utilizando a função `theme()`.
- Funções disponíveis para a camada tema no pacote `ggplot2`:
  - `theme_bw`
  - `theme_classic()`
  - `theme_dark()`
  - `theme_get()`
  - `theme_gray()`
  - `theme_grey()`
  - `theme_light()`
  - `theme_linedraw()`
  - `theme_minimal()`
  - `theme_replace()`
  - `theme_set()`
  - `theme_test()`
  - `theme_update()`
  - `theme_void()`

```r
ggplot(data = iris, aes(x = Species, y = Sepal.Width) )
  geom_boxplot(
    aes(fill = Species)
  ) +
  scale_x_discrete(
    name = "Espécies"
  ) +
  scale_y_continuous(
    name = "Largura da Sepata"
  ) +
  scale_fill_manual(
    values = c("#005c99", "#800000", "#404040"),
    name = "Espécies"
  ) +
  theme_classic() +
  theme(
    axis.title = element_text(size = 12, face = "bold"),
    axis.text = element_text(size = 12),
    legend.title = element_text(size = 12, face = "bold"
    legend.text = element_text(size = 10),
    legend.position = c(0.5, 0.9)
  )
```



Exercício 1

- Usando os microdados do Sistema de Informação sobre Nascidos Vivos (SINASC) referentes ao ano de 2022, construa os gráficos solicitados usando o pacote `ggplot2`

    1. Faça um gráfico de barras para a frequência (quantidade) de nascidos vivos em cada Unidade Federativa. Ordene as barras em ordem decrescente.

    2. Elabore um gráfico de linha para analisar a quantidade de nascidos vivos ao longo dos meses de 2022.

    3. Crie um boxplot para analisar a distribuição dos pesos dos nascidos vivos, distinguindo entre os sexos masculino e feminino.

    4. Recrie a visualização do exercício anterior, substituindo o boxplot por um gráfico de densidade.