Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
IMD0601 - Bioestatística

Visualização dos dados em R

Prof. Dr. Tetsu Sakamoto
Instituto Metrópole Digital - UFRN
Sala A224, ramal 182
Email: tetsu@imd.ufrn.br

Baixe a aula (e os arquivos)

- Para aqueles que não clonaram o repositório:

```
> git clone https://github.com/tetsufmbio/IMD0601.git
```

- Para aqueles que já tem o repositório local:

```
> cd /path/to/IMD0601
```

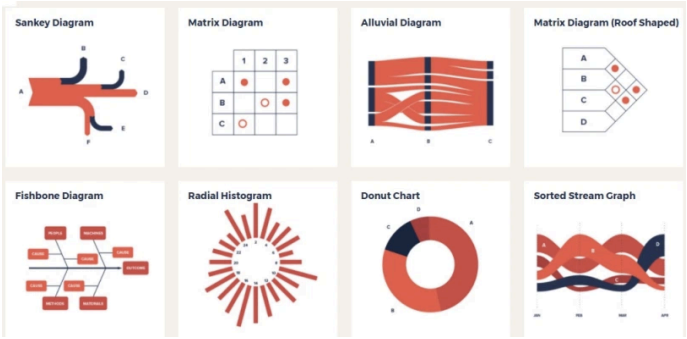
```
> git pull
```

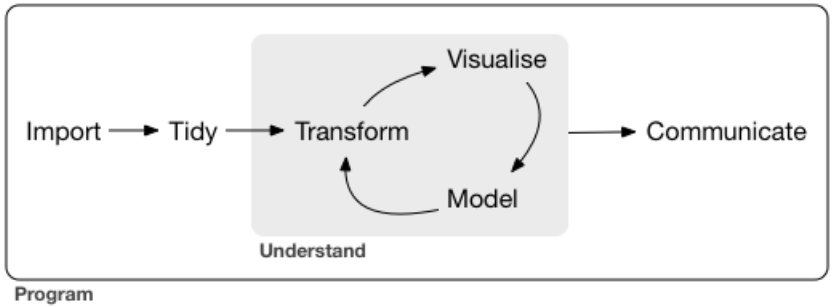
Visualização dos dados

Cientistas de dados;

Combina as áreas de:

- **Estatística**: análise gráfica dos dados (Representação e interpretação dos dados);
- **Design**: Princípios de design (gráficos atrativos e que promove o melhor entendimento e comunicação);

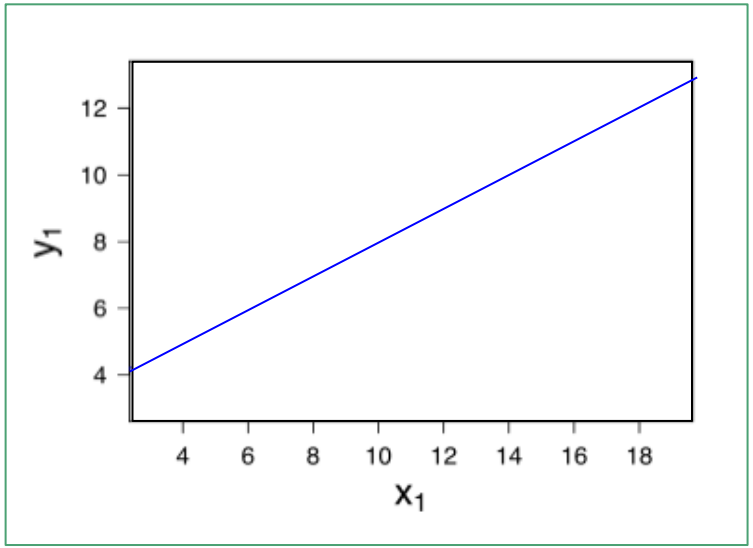


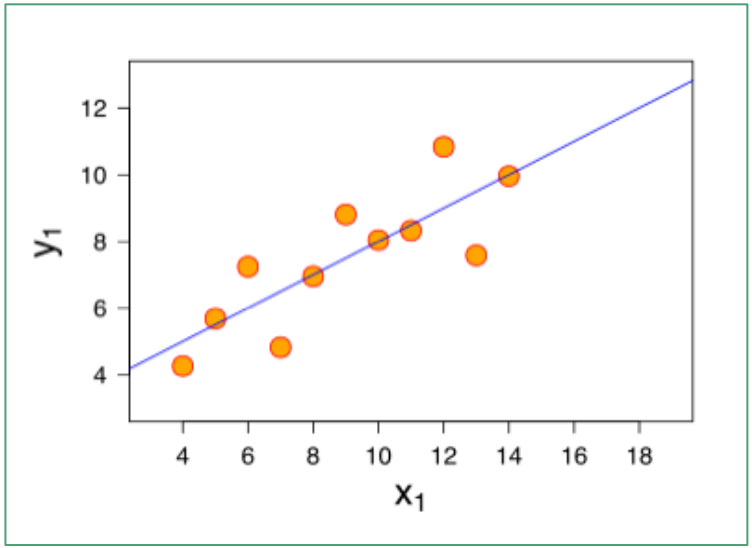


https://r4ds.had.co.nz/introduction.html

Visualização dos dados

Quarteto de Anscombe

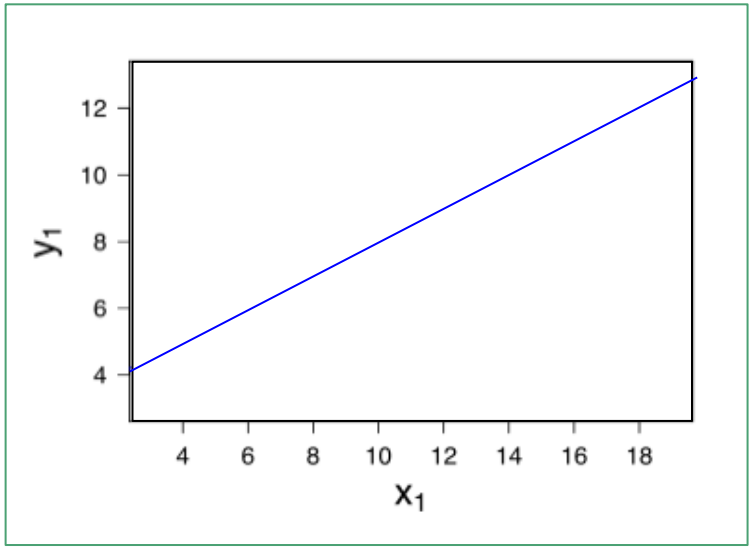


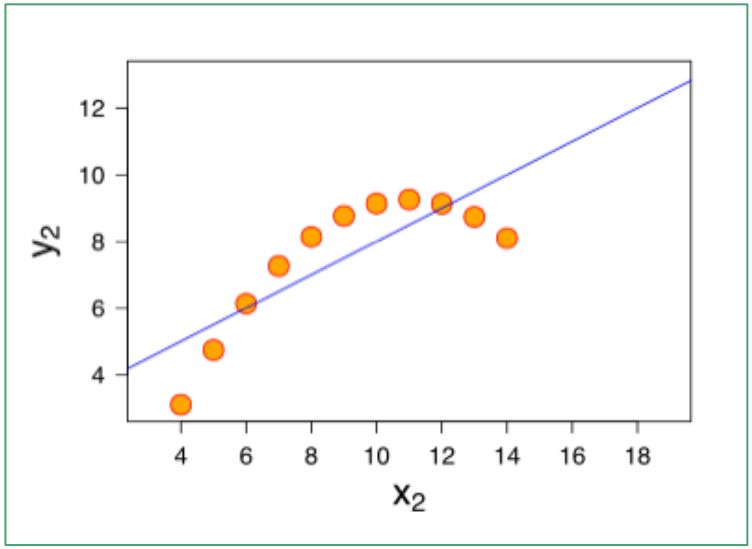


Anscombe, FJ. Graphs in Statistical Analysis. American Statistician, 1973.

Visualização dos dados

Quarteto de Anscombe

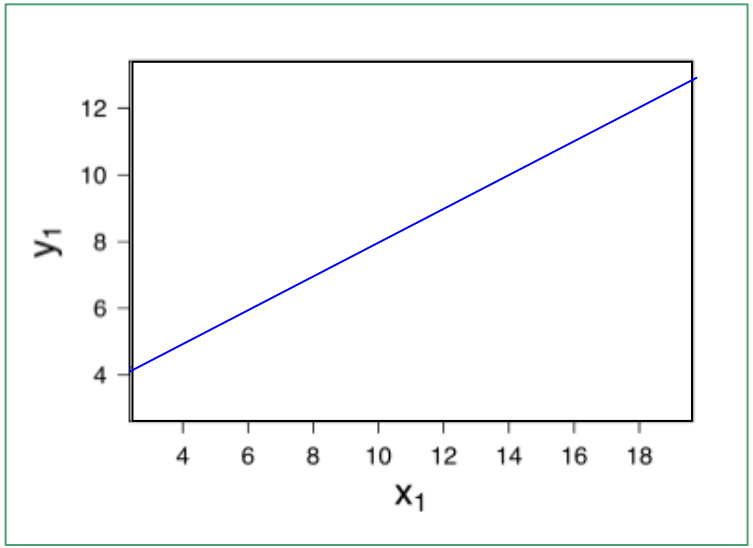


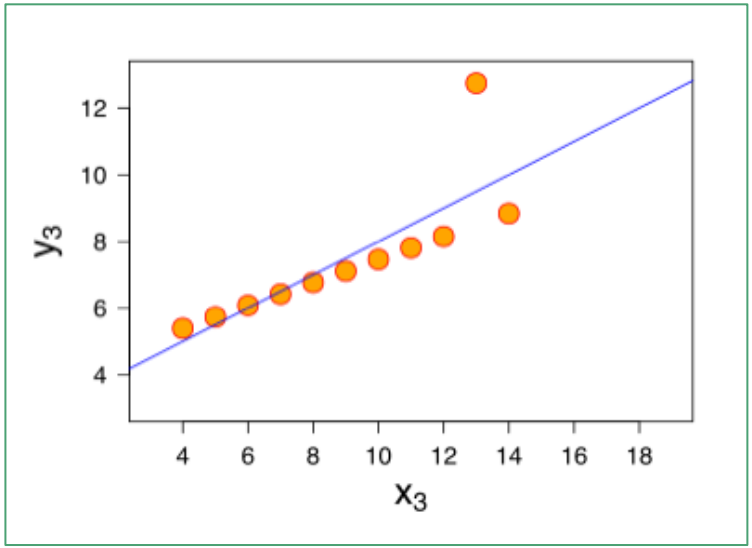


Anscombe, FJ. Graphs in Statistical Analysis. American Statistician, 1973.

Visualização dos dados

Quarteto de Anscombe

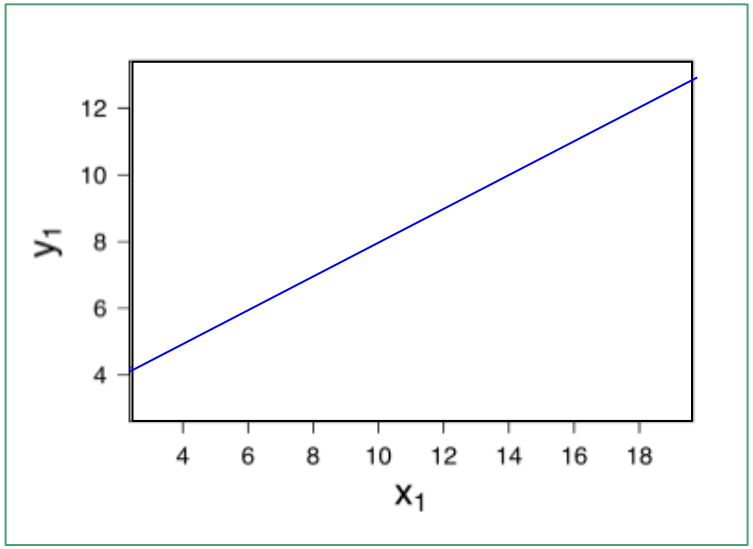


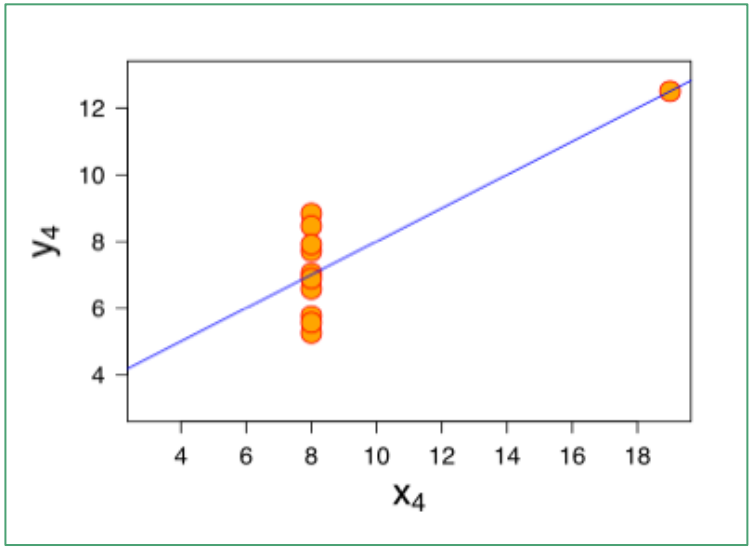


Anscombe, FJ. Graphs in Statistical Analysis. American Statistician, 1973.

Visualização dos dados

Quarteto de Anscombe



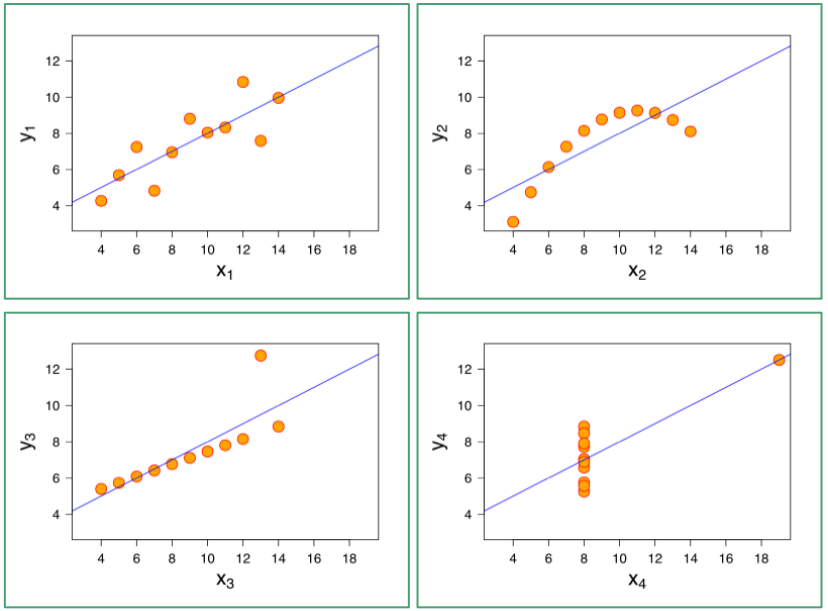


Anscombe, FJ. Graphs in Statistical Analysis. American Statistician, 1973.

Visualização dos dados

Quarteto de Anscombe

Propriedade	Valor	Precisão
Média de *x*	9	exato
Variância de *x*	11	exato
Média de *y*	7,50	até 2 casas decimais
Variância de *y*	4,125	±0,003
Correlação entre *x* e *y*	0,816	até 3 casas decimais
Reta de regressão linear	$y = 3,00 + 0,500x$	até 2 e 3 casas decimais, respectivamente
Coeficiente de determinação da regressão linear: R^2	0,67	até 2 casas decimais



Anscombe, FJ. Graphs in Statistical Analysis. American Statistician, 1973.

Visualização dos dados

Gráficos exploratórios

- Gerado facilmente;
- Dados pesados;
- Para uma audiência específica (você e seus colegas);
- Análise gráfica dos dados;
- Servem para analisar e confirmar.

R Base Graphics

Gráficos explanatórios

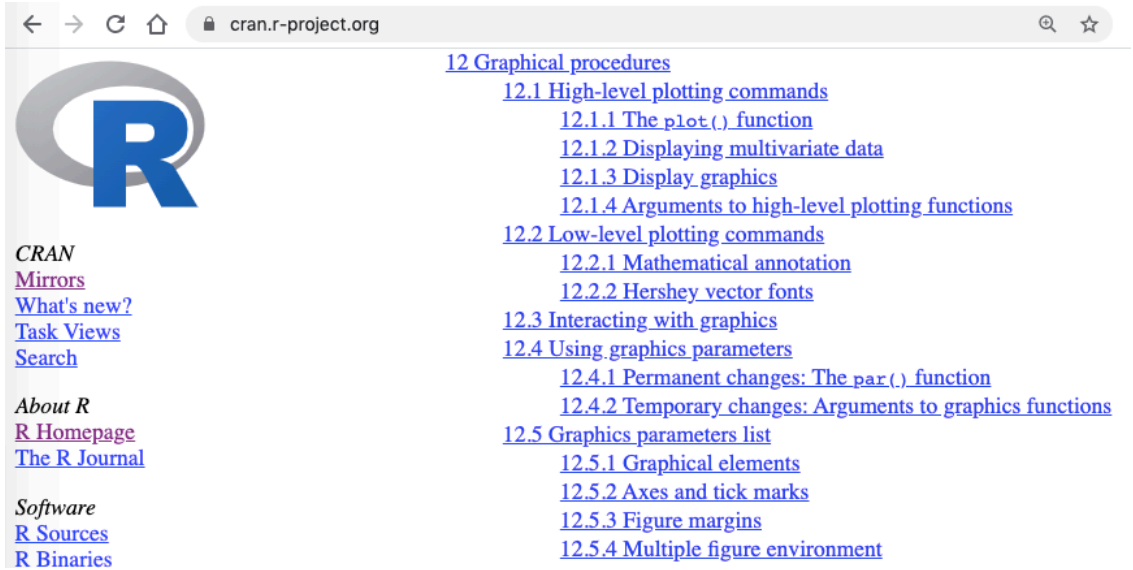
- Laboriosos;
- Específicos para determinados dados;
- Para uma audiência ampla (publicação ou apresentação);
- Parte comunicativa do processo.
- Servem para informar e persuadir.

Ggplot2, ggvis
lattice

R Base Graphics

R Base Graphics:

An Idiot's Guide



PROGRAMMING, R, UDACITY

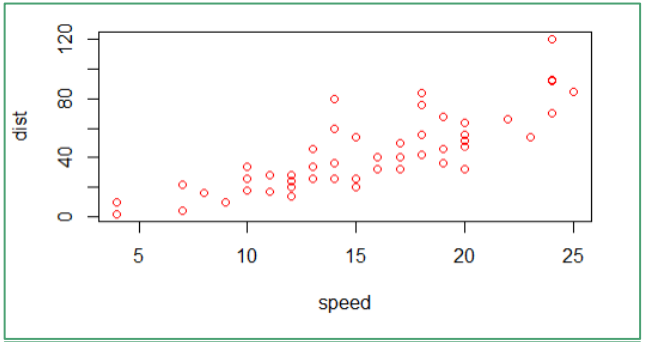
SWIRL – R PROGRAMMING – LESSON 15 – BASE GRAPHICS

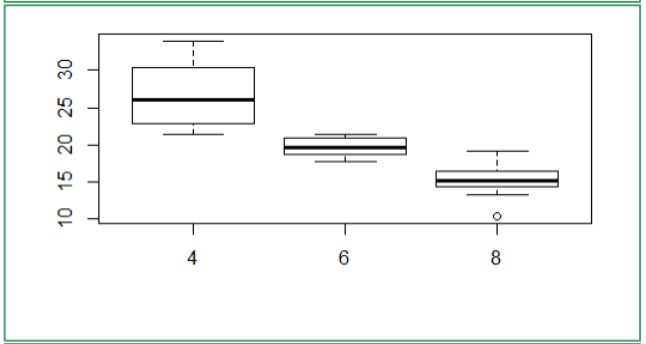
HOW-TO & USEFUL STUFF

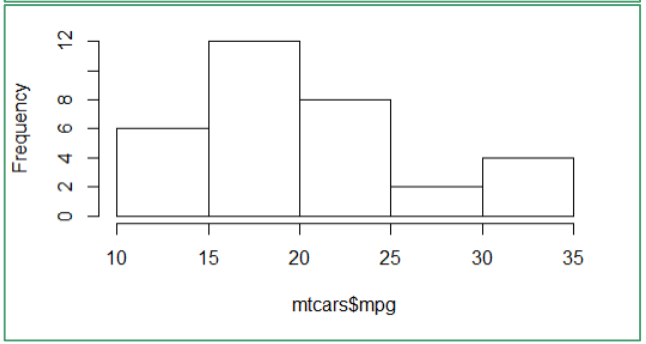
Base Graphics in R: A Detailed Idiot's Guide

R Base Graphics

```
plot(x = cars$speed, y = cars$dist,
xlab = "Speed", ylab = "Stopping
Distance", col = 2)

boxplot(formula = mpg~cyl, data =
mtcars)

hist(mtcars$mpg)
```





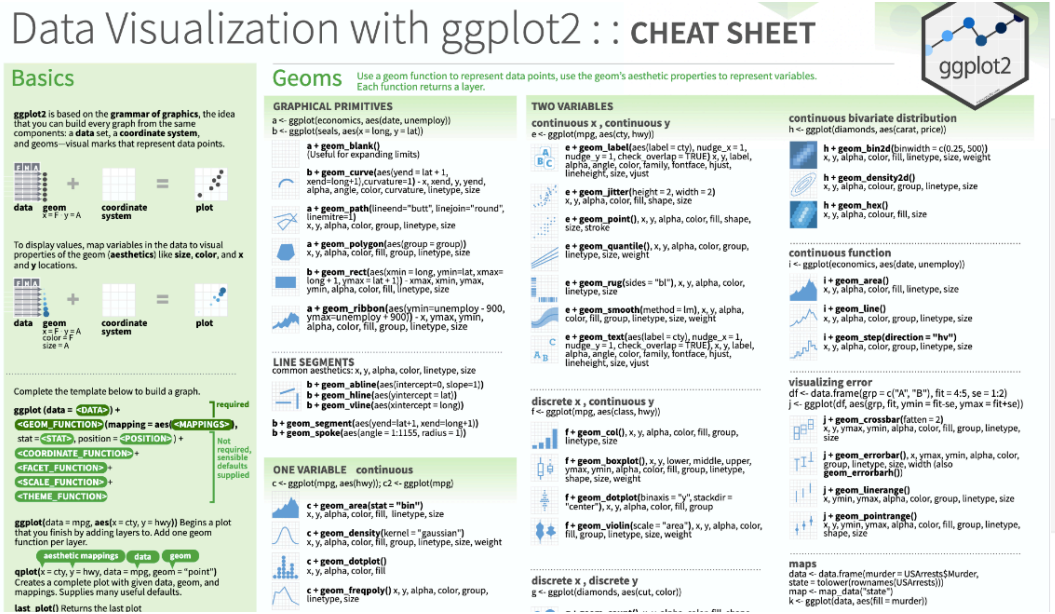


Overview

ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

Installation

```
# The easiest way to
install.packages("ti

# Alternatively, ins
install.packages("gg

# Or the development
# install.packages('
devtools::install_gi
```



Links

Download from CRAN at
https://cloud.r-project.org/package=ggplot2

Browse source code at
https://github.com/tidyverse/ggplot2/

Report a bug at
https://github.com/tidyverse/ggplot2/issues

Learn more at
https://r4ds.had.co.nz/data-visualisation.html

Extensions at
https://exts.ggplot2.tidyverse.org/gallery/

ggplot2

Hadley Wickham

“The Grammar of Graphics”

Adiciona camadas nos gráficos para melhor visualização dos dados;

```
library(ggplot2)
```

ggplot2

Data

Dados (tabela) onde se encontram as variáveis a serem representados graficamente.

```
ggplot(data=mtcars)
```

```
ggplot(mtcars)
```

ggplot2

Aesthetics

Permite especificar as variáveis que queremos utilizar na representação gráfica.

```
ggplot(mtcars, aes(x=mpg, y=wt))
```

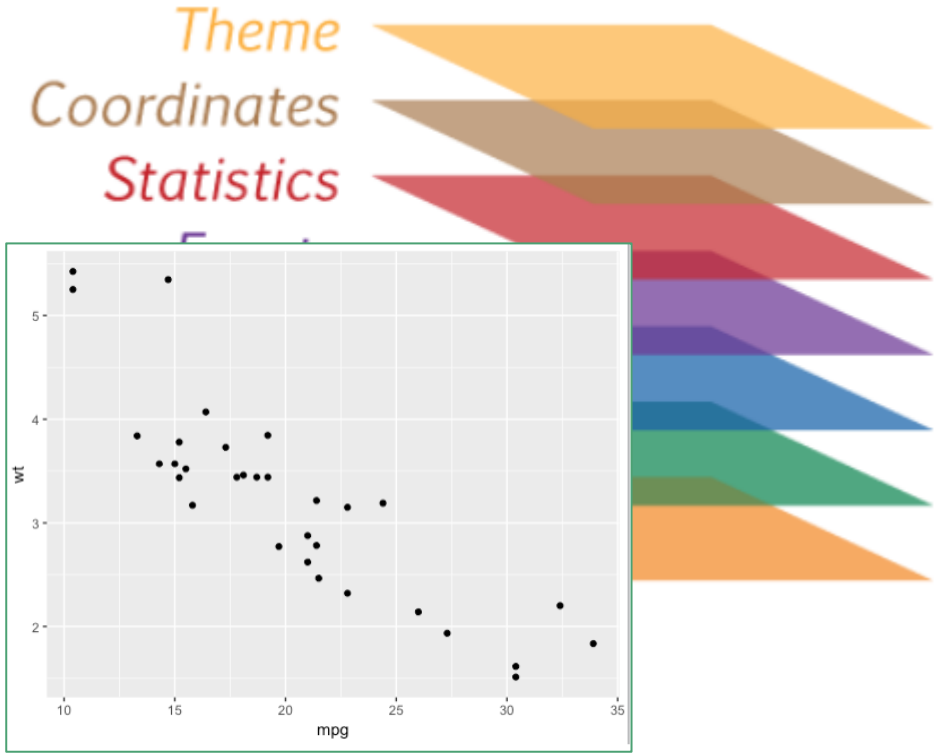
ggplot2

Geometries

Camada que indica a forma como os dados devem ser apresentados no gráfico.

```
ggplot(mtcars, aes(x=mpg, y=wt)) + geom_point()
```

```
g <- ggplot(mtcars, aes(x=mpg, y=wt)) + geom_point()
```



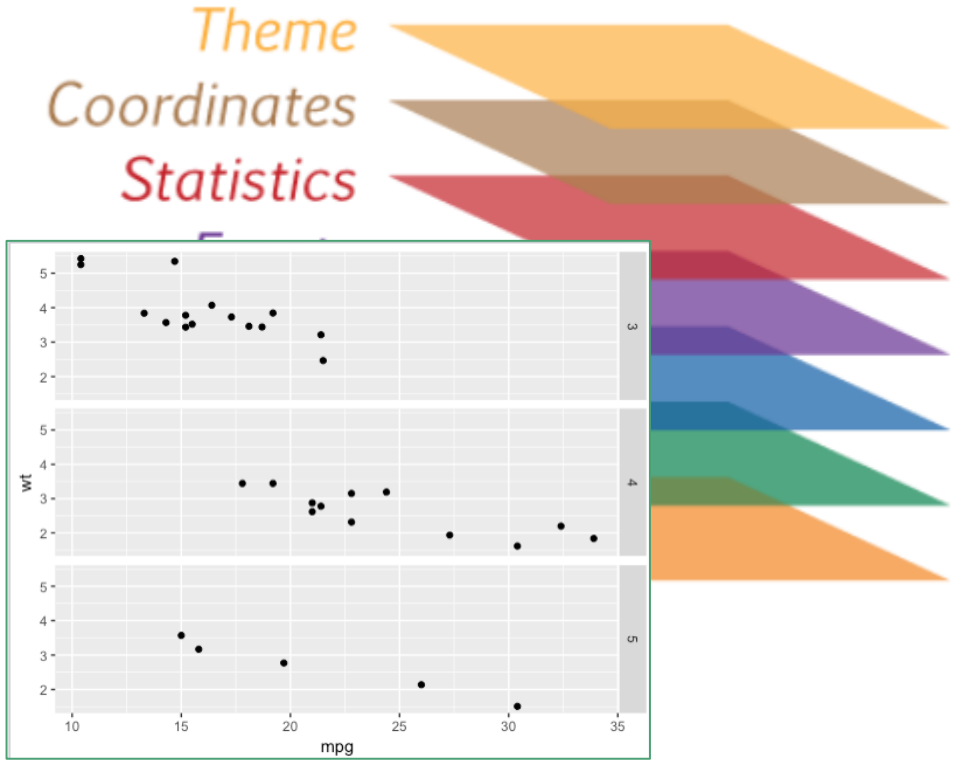
ggplot2

Facets

Permite colocar múltiplos gráficos em um canvas.

```
g + facet_grid(gear~.)
```

```
g <- g + facet_grid(gear~.)
```



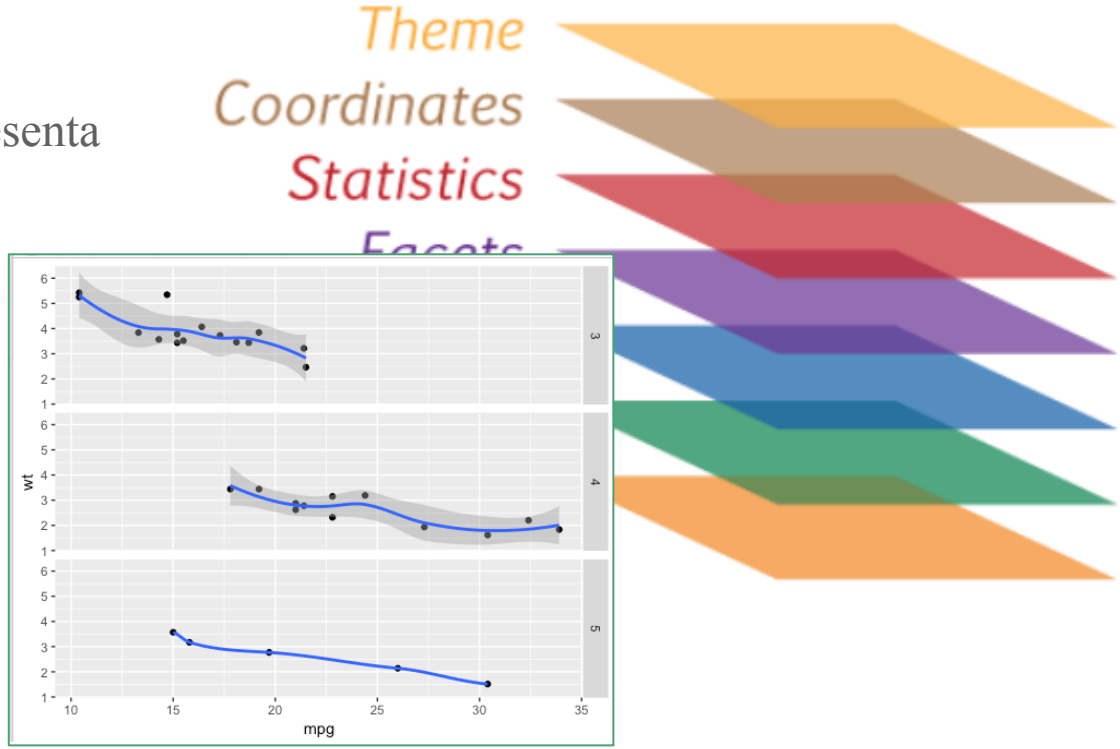
ggplot2

Statistics

Adiciona uma camada que representa uma análise estatística.

```
g + stat_smooth()
```

```
g <- g + stat_smooth()
```



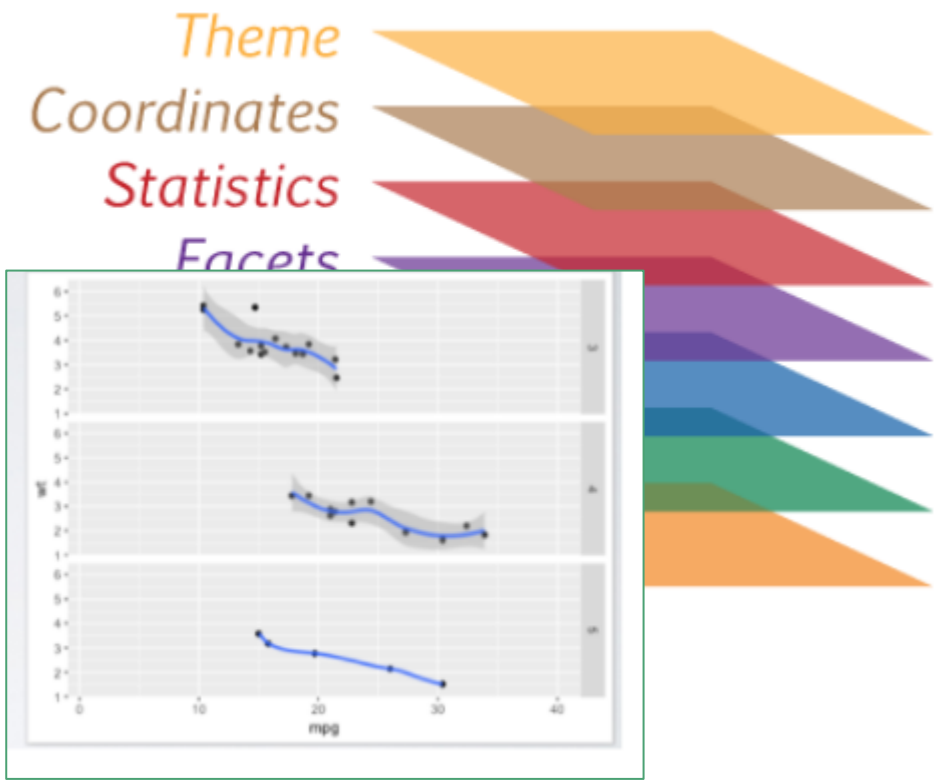
ggplot2

Coordinates

Camada que controla como as posições devem ser mapeadas no gráfico (limites do eixo x e y).

```
g + coord_cartesian(xlim =
c(1, 40))
```

```
g <- g + coord_cartesian(xlim
= c(1, 40))
```

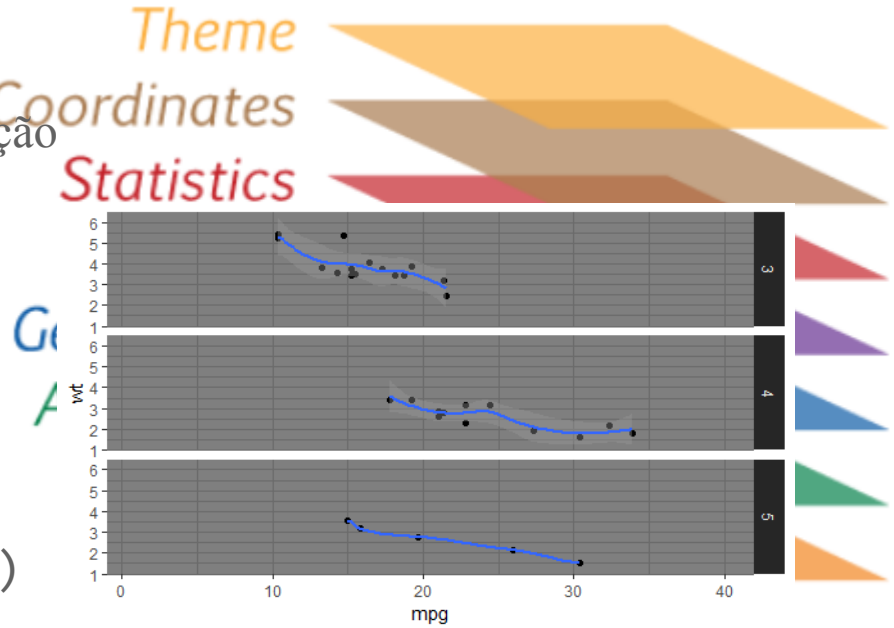


ggplot2

Themes

Camada que permite enriquecer a apresentação do gráfico (rótulos, fonte, cor, etc).

```
g + theme_dark()

g <- g + theme_dark()


Install.packages('ggThemeAssist')
```

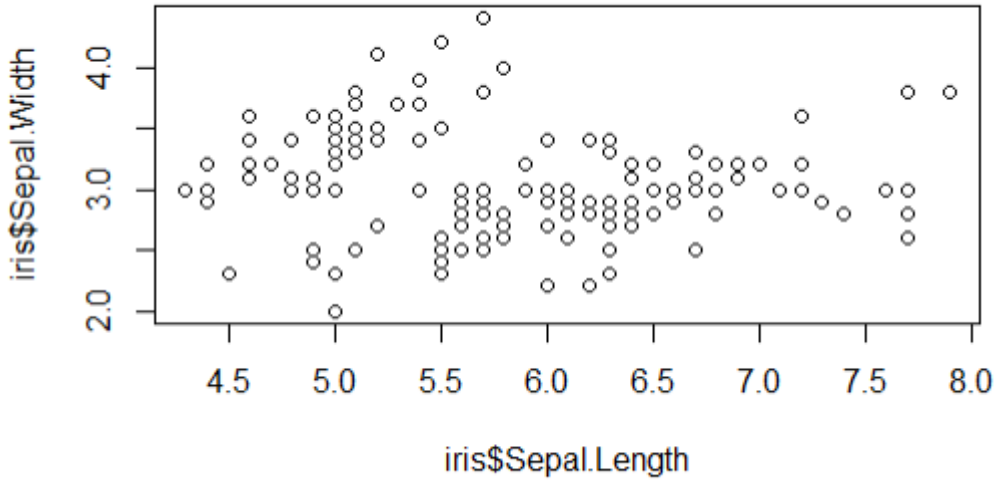


Iris dataset


```
data(iris)

str(iris)

'data.frame':  150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1
1 1 1 1 1 1 1 1 ...
```

Base plot

Sepal length X Sepal width

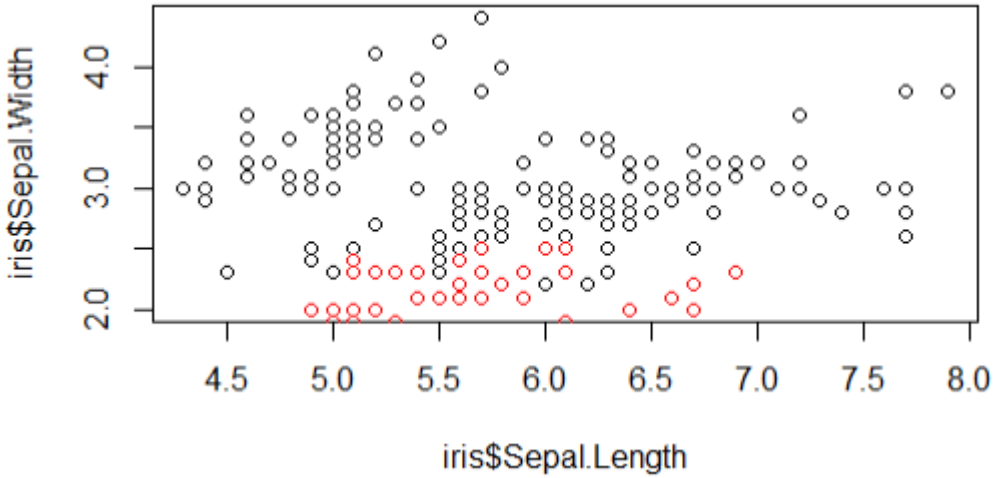



```
> plot(iris$Sepal.Length, iris$Sepal.Width)
```

Base plot

Como adicionar os dados de Petal length e Petal width?

```
> plot(iris$Sepal.Length, iris$Sepal.Width)

> points(iris$Petal.Length, iris$Petal.Width, col = "red")
```



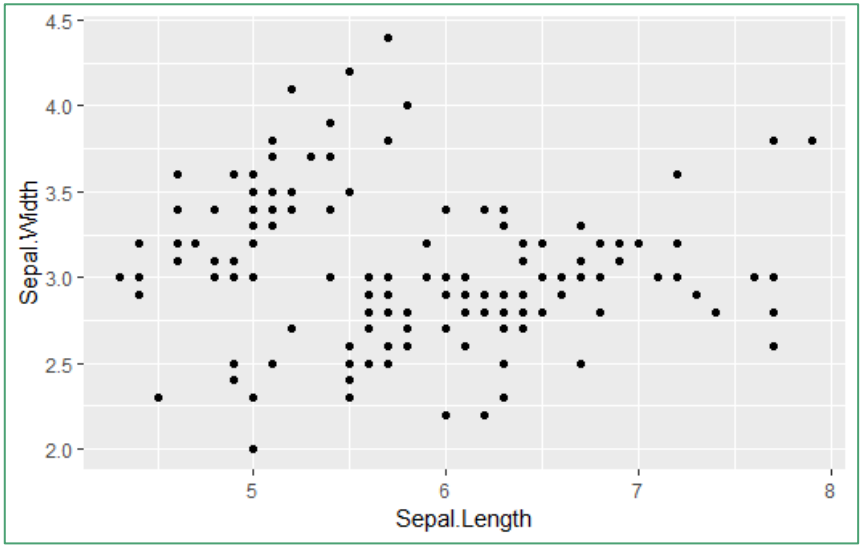
Limitações:

- O gráfico não é redesenhado;
- O gráfico gerado é uma imagem;
- A legenda é adicionada manualmente;
- Não há uma framework unificado para gerar diferentes tipos de gráficos.

ggplot2

Sepal length X Sepal width

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



ggplot2

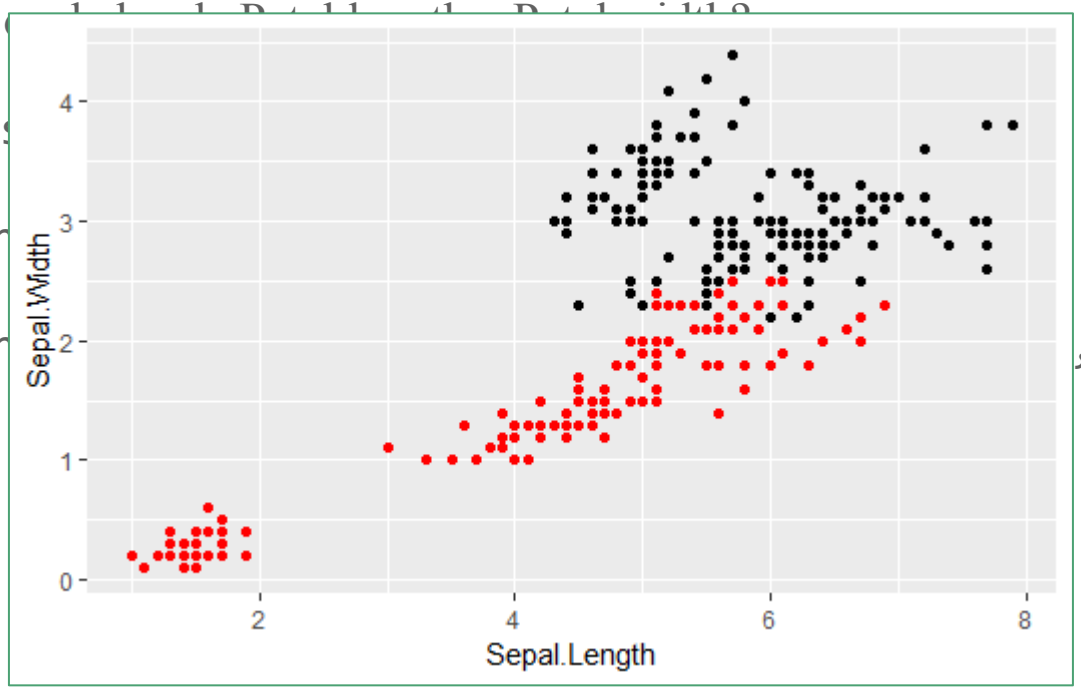
Como adicionar os dados de Petal length e Petal width?

```
> ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +

    geom_point() +

    geom_point(aes(x = Petal.Length, y = Petal.Width),
col = "red")
```

ggplot2

Como adicionar

```
> ggplot(iris
    geom_poir
    geom_poir
col = "red")
```



ggplot2

Exercício

Adicione uma coluna na tabela **iris** que corresponda a um identificador único de cada observação.

```
> iris$Flower <- 1:nrow(iris)
```

Crie uma tabela onde as variáveis Length e Width estejam cada uma em uma coluna, como abaixo.

```
  Species Flower  part Length Width
1  setosa      1 Petal    1.4   0.2
2  setosa      1 Sepal    5.1   3.5
...
```

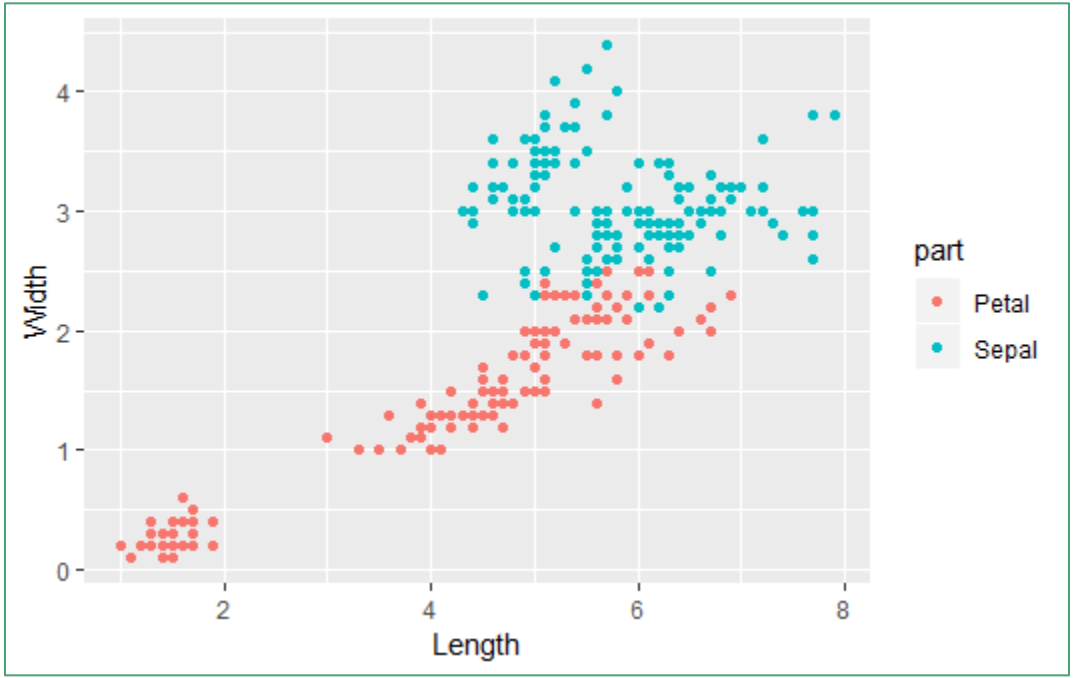
ggplot2

Exercício

```
> library(tidyr)

> iris$Flower <- 1:nrow(iris)

> iris.wide <- gather(iris, part_measure, val, -Species, -Flower )

> iris.wide <- separate(iris.wide, part_measure,
c("part","measure"))

> iris.wide <- spread(iris.wide, measure, val)
```

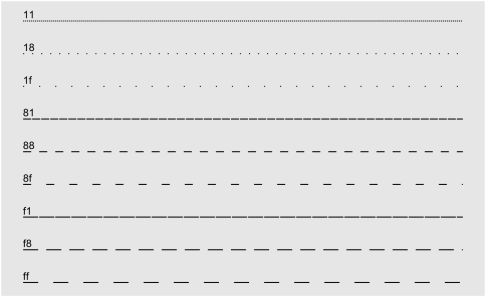
ggplot2

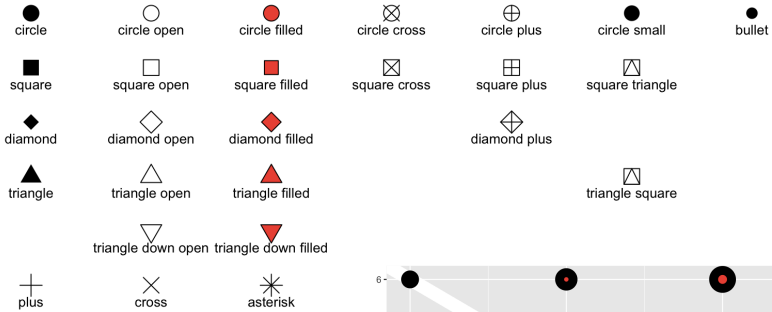
```
ggplot(iris.wide, aes(Length, Width, col = part)) + geom_point()
```

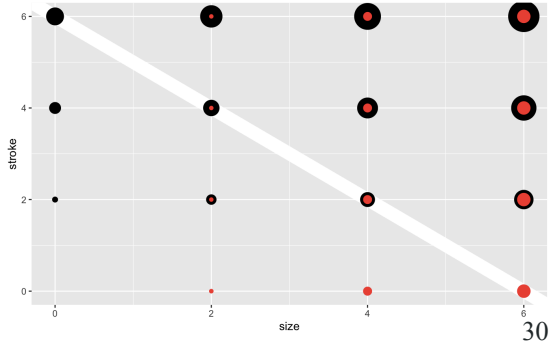


Parâmetros típicos da estética

- **x** → posição no eixo x;
- **y** → posição no eixo y;
- **col** → cor dos pontos ou formas;
- **fill** → cor a ser preenchida;
- **size** → diâmetro do ponto, largura da linha;
- **alpha** → transparência;
- **linetype** → padrão de tracejamento;
- **labels** → texto no gráfico;
- **shape** → formas;

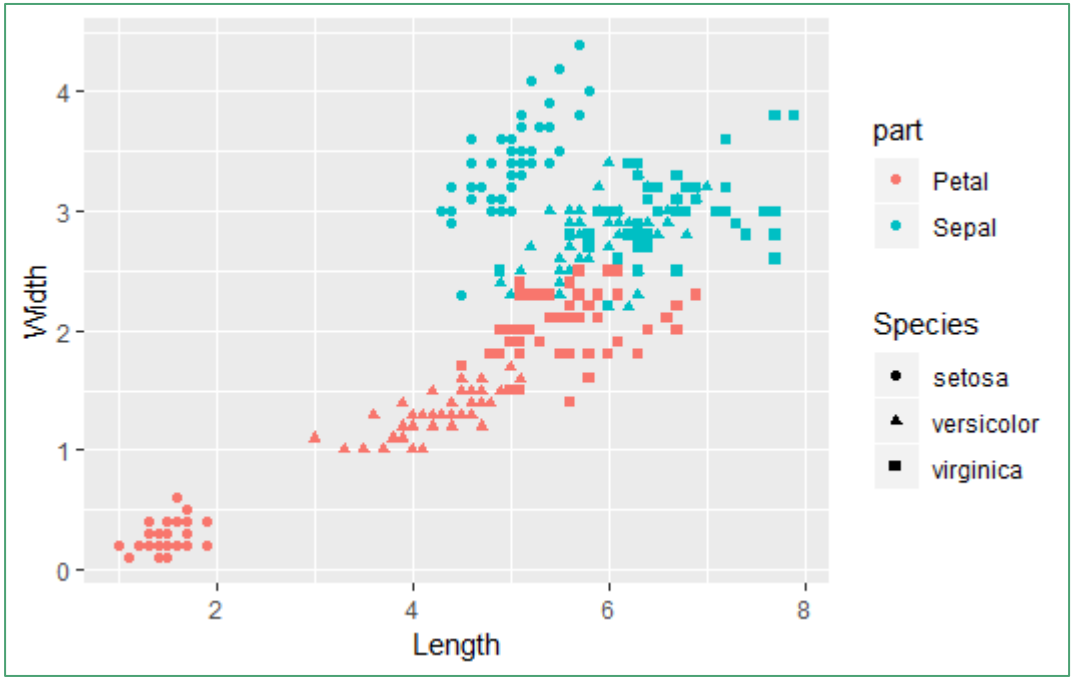






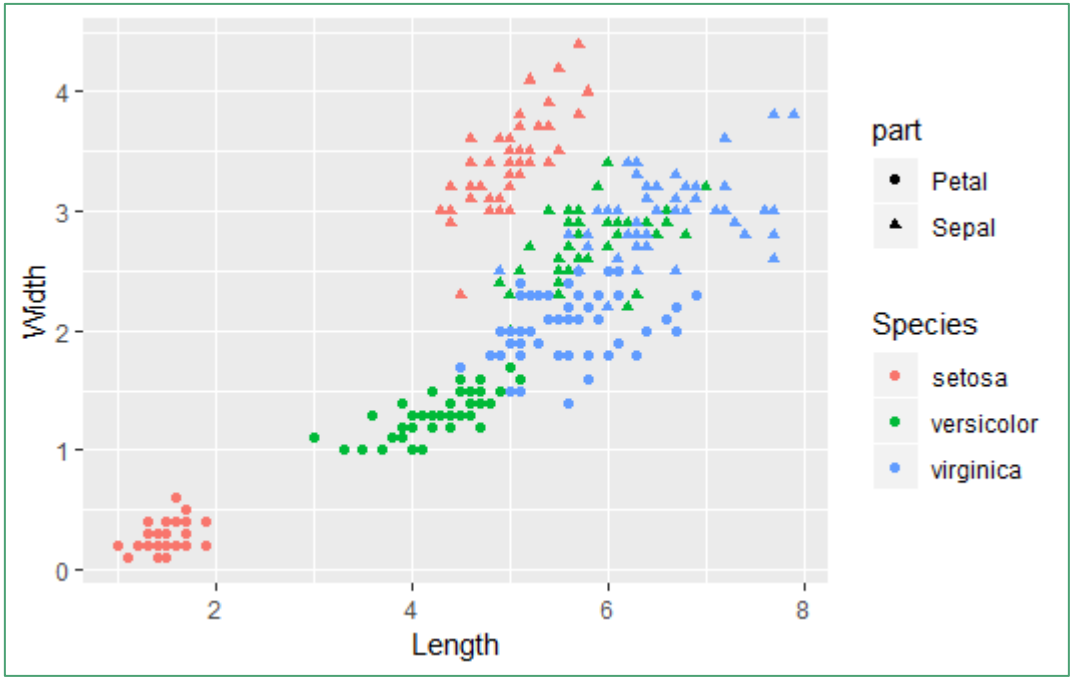
ggplot2

```
ggplot(iris.wide, aes(Length, Width, col = part, shape = Species))
+ geom_point()
```



ggplot2

```
ggplot(iris.wide, aes(Length, Width, col = Species, shape = part))
+ geom_point()
```



A sintaxe do ggplot

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(
     mapping = aes(<MAPPINGS>),
     stat = <STAT>,
     position = <POSITION>
  ) +
  <COORDINATE_FUNCTION> +
  <FACET_FUNCTION>
```

Data	*{variables of interest}*				
Aesthetics	*x-axis* *y-axis*	*colour* *fill*	*size* *labels*	*alpha* *shape*	*line width* *line type*
Geometries	*point*	*line*	*histogram*	*bar*	*boxplot*
Themes	*non-data ink*				
Statistics	*binning*	*smoothing*	*descriptive*	*inferential*	
Coordinates	*cartesian*	*fixed*	*polar*	*limits*	
Facets	*columns*	*rows*			

r4ds.had.co.nz/

Referências

- Aula baseada no curso “**Data Visualization with ggplot2 (Part 1)**” de Rick Scavetta: https://www.datacamp.com/courses/data-visualization-with-ggplot2-1

- Ciência de Dados com R – IBPAD: https://www.ibpad.com.br/o-que-fazemos/publicacoes/introducao-ciencia-de-dados-com-r/

- Tidyverse: https://ggplot2.tidyverse.org/index.html (link para cheat sheet!)

- R for data Science: https://r4ds.had.co.nz/

- https://skillgaze.com/2017/10/31/understanding-different-visualization-layers-of-ggplot/