

Aula 3 - Visualização de Dados

Introdução ao aprendizado de máquina - UEMA 2025

Thiago S. F .Silva

2025-12-10

Parte I - Considerações gerais

Analisar dados sem uma exploração gráfica é como ir a um encontro às escuras.

Análise visual

Analisar dados sem uma exploração gráfica é como ir a um encontro às escuras.



Analisar dados sem uma exploração gráfica é como ir a um encontro às escuras.

Nós somos seres essencialmente visuais, com uma capacidade incrível de processar imagens.

Por que gráficos?

O “Quarteto de Anscombe”

Anscombe, F.J., 1973. Graphs in Statistical Analysis. The American Statistician 27, 17–21

```
head(anscombe)
```

	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04

Quarteto de Anscombe

```
round(lm(y1 ~ x1, data = anscombe)$coefficients,2)
```

(Intercept)	x1
3.0	0.5

```
round(lm(y2 ~ x2, data = anscombe)$coefficients,2)
```

(Intercept)	x2
3.0	0.5

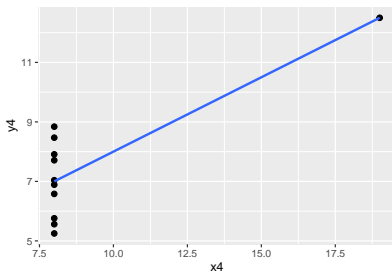
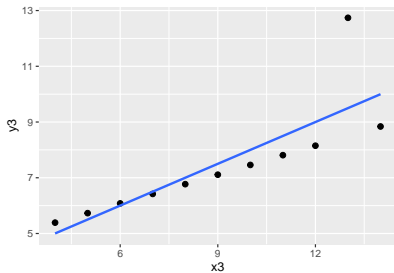
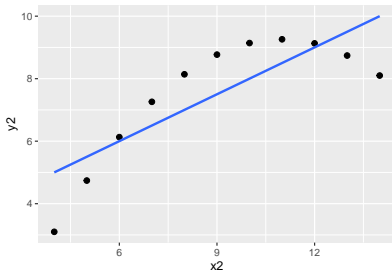
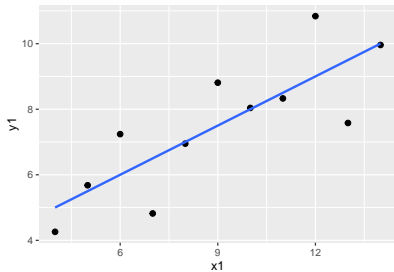
```
round(lm(y3 ~ x3, data = anscombe)$coefficients,2)
```

(Intercept)	x3
3.0	0.5

```
round(lm(y4 ~ x4, data = anscombe)$coefficients,2)
```

(Intercept)	x4
3.0	0.5

Quarteto de Anscombe



Melhor que o quarteto de Anscombe

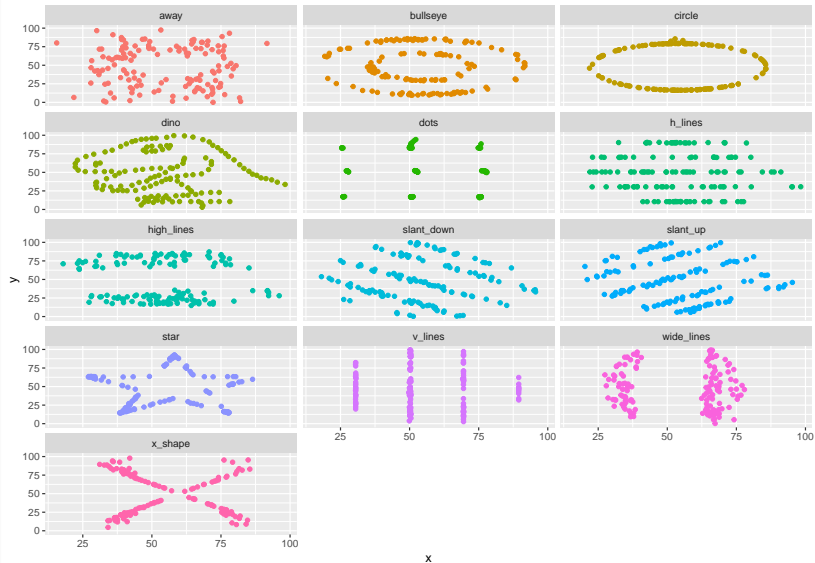
```
library(datasauRus)
library(dplyr)

datasaurus_dozen %>% group_by(dataset) %>%
  summarise(mean_x = mean(x), mean_y = mean(y), sd_x = sd(x),
            sd_y = sd(y))
```

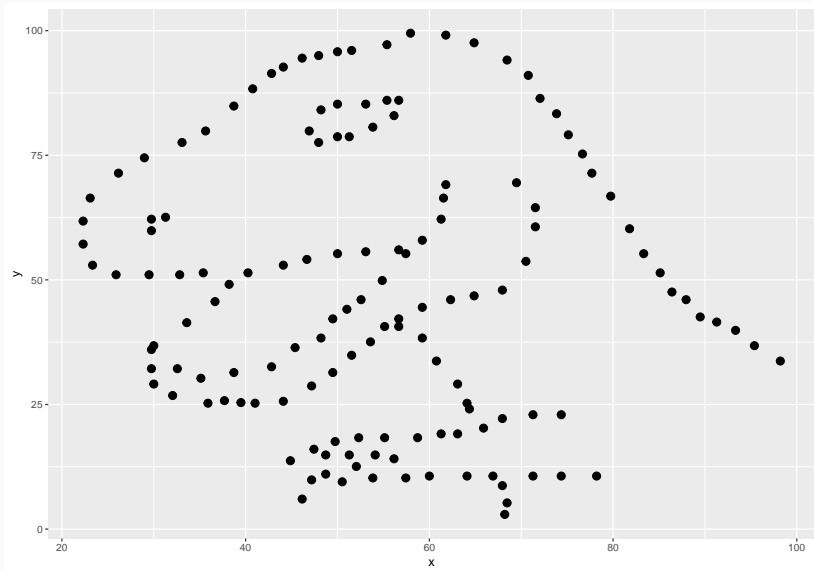
A tibble: 13 x 5

	dataset	mean_x	mean_y	sd_x	sd_y
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	away	54.3	47.8	16.8	26.9
2	bullseye	54.3	47.8	16.8	26.9
3	circle	54.3	47.8	16.8	26.9
4	dino	54.3	47.8	16.8	26.9
5	dots	54.3	47.8	16.8	26.9
6	h_lines	54.3	47.8	16.8	26.9
7	high_lines	54.3	47.8	16.8	26.9
8	slant_down	54.3	47.8	16.8	26.9
9	slant_up	54.3	47.8	16.8	26.9

Melhor que o quarteto de Anscombe

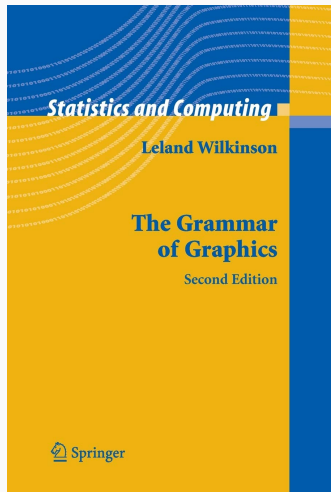


Melhor que o quarteto de Anscombe



A gramática dos gráficos

- Conceito criado por Leland Wilkinson
- Como pensar em gráficos de maneira sistemática



A gramática dos gráficos

Um gráfico é composto pelos seguintes elementos:

- **Dados (*data*):** quais dados se está plotando

A gramática dos gráficos

Um gráfico é composto pelos seguintes elementos:

- **Dados (*data*):** quais dados se está plotando
- **Estética (*aesthetics*):** como estes dados são representados (eixos, forma, - tamanho, cor)

A gramática dos gráficos

Um gráfico é composto pelos seguintes elementos:

- **Dados (*data*):** quais dados se está plotando
- **Estética (*aesthetics*):** como estes dados são representados (eixos, forma, - tamanho, cor)
- **Escala (*Scale*):** qual o intervalo dos dados a ser visualizado?

A gramática dos gráficos

Um gráfico é composto pelos seguintes elementos:

- **Dados (*data*):** quais dados se está plotando
- **Estética (*aesthetics*):** como estes dados são representados (eixos, forma, - tamanho, cor)
- **Escala (*Scale*):** qual o intervalo dos dados a ser visualizado?
- **Geometria (*geometry*):** que tipo de representação gráfica será usada?

A gramática dos gráficos

Um gráfico é composto pelos seguintes elementos:

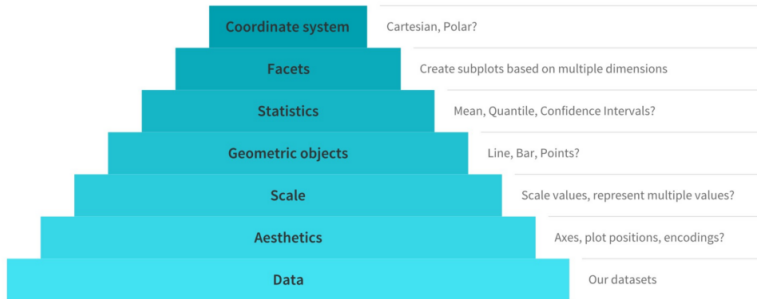
- **Dados (*data*):** quais dados se está plotando
- **Estética (*aesthetics*):** como estes dados são representados (eixos, forma, - tamanho, cor)
- **Escala (*Scale*):** qual o intervalo dos dados a ser visualizado?
- **Geometria (*geometry*):** que tipo de representação gráfica será usada?
- **Estatísticas (*statistics*):** como os dados originais estão sendo resumidos no - gráfico?

A gramática dos gráficos

Um gráfico é composto pelos seguintes elementos:

- **Dados (*data*):** quais dados se está plotando
- **Estética (*aesthetics*):** como estes dados são representados (eixos, forma, - tamanho, cor)
- **Escala (*Scale*):** qual o intervalo dos dados a ser visualizado?
- **Geometria (*geometry*):** que tipo de representação gráfica será usada?
- **Estatísticas (*statistics*):** como os dados originais estão sendo resumidos no - gráfico?
- **Facetas (*facets*):** como os dados devem ser separados em painéis?

Major Components of the Grammar of Graphics



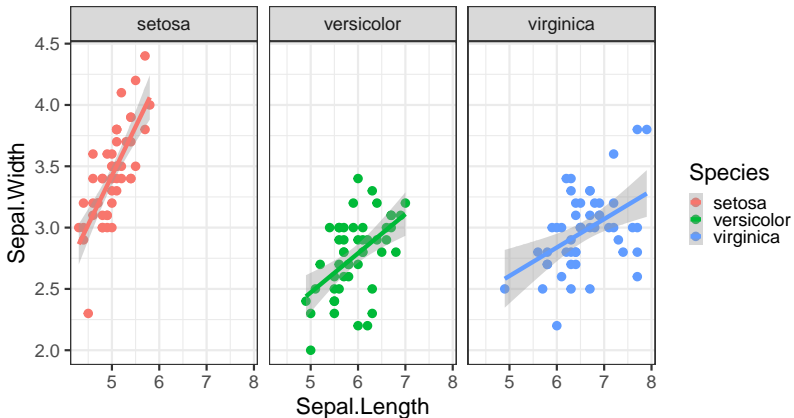
Exemplo

Dados: `iris` , Estética:

`x = Sepal.Length`, `y = Sepal.Width`, `cor = Species`, `valores = círculos` ,

Escala: `x=0:8`, `y=2:4.5` , Geometria = `pontos` , Estatística = `regressão` ,

Facetas = `Species`



Parte II - O pacote `ggplot2`

Uma implementação da gramática dos gráficos que segue a filosofia *tidyverse*.

<https://ggplot2.tidyverse.org/>

ggplot2: Elegant Graphics for Data Analysis (3e)

Funções: `ggplot()` , `geom_*`() , `scale_*`() , `stat_*` ,
`coord_*`() . etc.

Exemplo em R usando o dataset `iris`

Parte III - Principais tipos de gráficos

A escolha do tipo de gráfico é ditada principalmente por:

- número de variáveis (dimensões)
- tipo de variável (categórica vs. contínua)
- relação entre valores (séries vs. comparações)

Quais tipos comuns de gráficos usamos para mostrar apenas uma variável?

- Histograma

Quais tipos comuns de gráficos usamos para mostrar apenas uma variável?

- Histograma
-

Histograma

- Adequado para mostrar distribuições.
- Pode ser usado tanto para dados categóricos quanto contínuos.

É importante otimizar o número de subdivisões (*bins*)

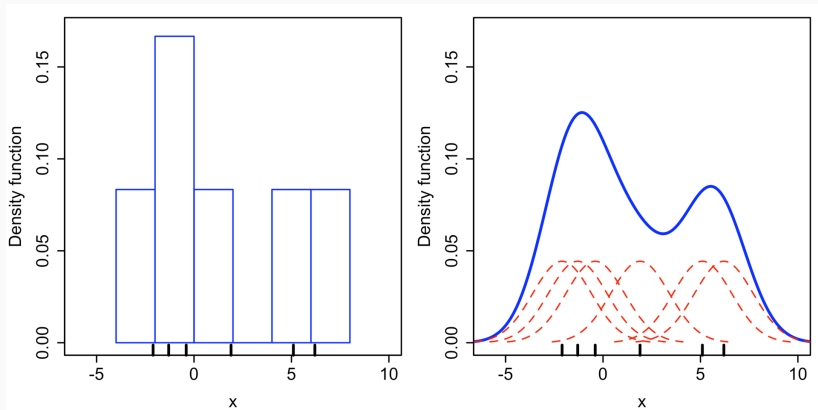
Similar ao histograma, o *dot plot* é adequado para datasets com poucas observações

- Cada observação é um ponto
- A largura do ponto equivale ao *bin* do histograma
- Vantagem: você consegue facilmente contar o numero de observações.

Gráficos de densidade tentam dar uma visão mais contínua da distribuição dos dados

- Utilizam um *kernel* para suavizar os dados
- A largura do *kernel* (*bandwidth*) muda a aparência do gráfico.

Tipos de Gráficos - 1 variável: Densidade



Tipos de Gráficos - 2 variáveis: Gráfico de Barras

- Adequado para mostrar proporções, especialmente apropriado para contagens de variáveis categóricas
- Pode ser mostrado lado a lado ou empilhado -Transmite a impressão de um dado **cumulativo**
- Não é recomendado para valores pontuais (ex: média)

Tipos de Gráficos - 2 variáveis: Gráfico de Pizza

Gráficos de pizza servem para:

Tipos de Gráficos - 2 variáveis: Gráfico de Pizza

Gráficos de pizza servem para:

- Nada!
- Nosso cérebro é muito mais apto em julgar comprimentos (barras) do que áreas ou ângulos.
- Sempre que você pensar em fazer um gráfico de pizza, um gráfico de barras é melhor

Tipos de Gráficos - 2 variáveis: diagrama de dispersão (*scatterplot*)

- Um dos gráficos mais úteis em estatística
- Serve para visualizar duas variáveis contínuas
- Cuidado ao unir os pontos com linhas, pois isso passa uma noção de continuidade!

Tipos de Gráficos - 2 variáveis: intervalos, barras de erro, etc

- Utilizados para mostrar *estatísticas* ao invés de dados brutos
- Podem ser combinados com outros tipos de gráfico para indicar incerteza

Tipos de Gráficos - 2 variáveis: Gráfico de Área

- Pode ser visto como uma versão contínua do gráfico de barras
- Mostra diferenças ponto a ponto e cumulativas
- Não deve ser usado se a área sob a curva não fizer sentido para os dados plotados
- A ordem do empilhamento pode afetar a percepção
- Se a variável x não for contínua, melhor usar barras empilhadas

Tipos de Gráficos - 2 variáveis: Boxplot

- Usado para combinações entre variáveis contínuas e categóricas
- Na opinião de muitos, um dos gráficos mais informativos que existem
- Combina as propriedades de um histograma e de um scatterplot
- Faz uso dos quantis para uma descrição robusta dos dados

Tipos de Gráficos - 2 variáveis: *Violin Plot*

- Tentativa de ir além do boxplot
- Combina as propriedades de um gráfico de densidades e de um scatterplot
- Pode ficar estranho se as distribuições não forem bem-comportadas

Combinando mais de duas variáveis ... sem usar “3D”

- Gráficos ‘3D’ são dependentes de perspectiva, e não enfatizam bem as diferenças
- São uma boa ferramenta de visualização se puderem ser interativos
- Para exibição em papel/tela, dificultam a interpretação
- Ao invés de usar múltiplos eixos, podemos explorar combinações de *estética* e *geometria* (cor, forma, etc.)

Exemplo - macrófitas amazônicas

Dataset coletado durante meu pós-doutorado.

rich_env_june.csv

```
library(readr)
```

```
mac_data <- read_csv('../data/rich_env_jun.csv')
```

```
glimpse(mac_data)
```

Nitrogênio Total vs. Fósforo Total

Nitrogênio Total vs. Fósforo Total vs. riqueza de espécies

Nitrogênio Total vs. Fósforo Total vs. riqueza de espécies
vs. profundidade

Nitrogênio Total vs. Fósforo Total vs. riqueza de espécies
vs. profundidade

Elementos de um bom gráfico}

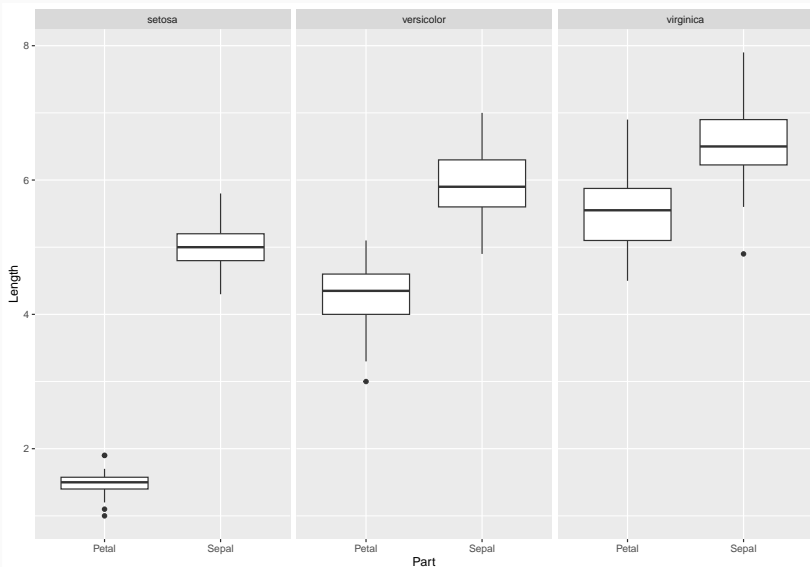
- Foco na informação que se quer enfatizar
- Quanto menor a razão tinta/papel, melhor
- Selecione e ordene suas variáveis de acordo com a pergunta a ser respondida
- Cores e formas só devem ser usadas se também trouxerem informação!

Qual a pergunta a ser respondida?

```
# A tibble: 6 x 3
  Species Part   Length
  <fct>    <fct>   <dbl>
1 setosa  Sepal    5.1
2 setosa  Petal    1.4
3 setosa  Sepal    4.9
4 setosa  Petal    1.4
5 setosa  Sepal    4.7
6 setosa  Petal    1.3
```

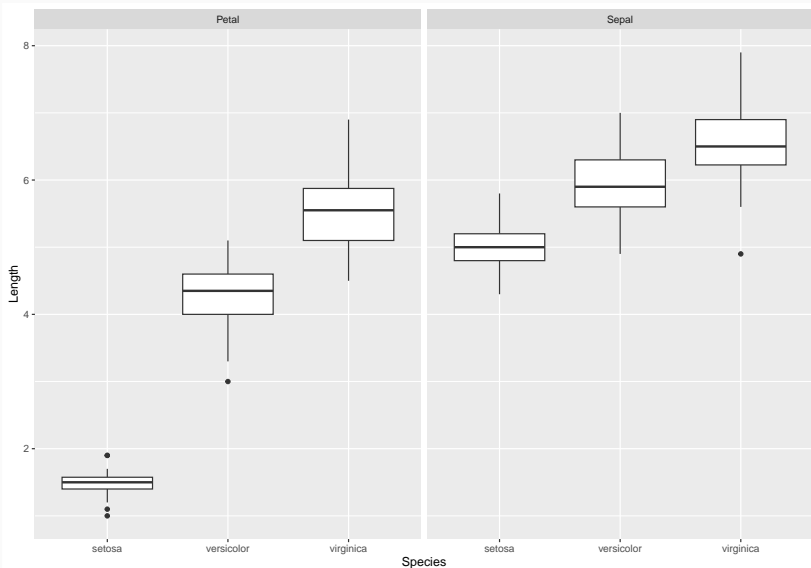
Qual a pergunta a ser respondida?

Diferença entre partes, para cada espécie?



Qual a pergunta a ser respondida?

Ou a diferença entre espécies, para cada parte?



Parte 4 - Praticando Aulas 2 e 3
