

Visualização de dados



Pedro C. Junger
Doutorando no LMPB
(pedro.junger@gmail.com)

São Carlos, Janeiro 2018

O que veremos nesta aula:

1. Importância da representação gráfica
2. Introdução a gramática de gráficos no R: usando **ggplot2**
3. Os cinco tipos de gráficos básicos
4. Exercícios

Por quê é importante fazer gráficos?

bio	wei	loc	cat
100.46921	100.42304	A	sim
100.00974	101.75921	A	sim
99.84660	98.18488	A	sim
98.51234	99.02479	A	sim
100.29377	101.94732	A	sim
99.37019	98.78707	A	sim
97.50122	97.73503	A	sim
98.94134	98.25695	A	sim
98.68764	99.91037	A	sim
100.15400	97.58296	A	sim
101.46085	101.27768	B	sim
100.42801	100.43868	B	sim
100.47251	102.39220	B	sim
100.34333	98.79654	B	sim
100.74104	100.46328	B	sim
100.98357	103.12065	B	nao
99.30434	98.70494	B	nao
101.08160	100.77335	B	nao
98.85230	99.78225	B	nao
100.73539	101.17838	B	nao

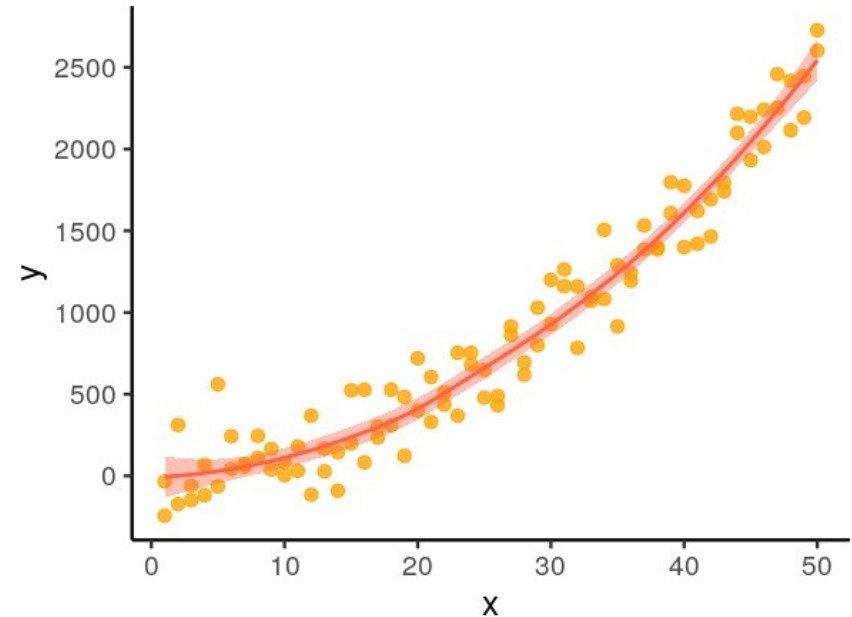
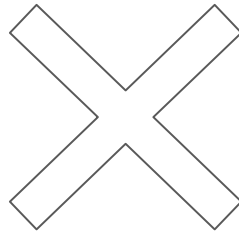
bio	wei	loc	cat
101.46085	101.27768	B	sim
100.42801	100.43868	B	sim
100.47251	102.39220	B	sim
100.34333	98.79654	B	sim
100.74104	100.46328	B	sim
100.98357	103.12065	B	nao
99.30434	98.70494	B	nao
101.08160	100.77335	B	nao
98.85230	99.78225	B	nao
100.73539	101.17838	B	nao
100.57782	101.81497	C	nao
100.03013	100.64929	C	nao
100.14746	101.38060	C	nao
100.46022	99.62030	C	nao
99.59179	99.17260	C	nao
98.67456	99.73205	C	nao
98.82537	95.96297	C	nao
99.74074	101.08266	C	nao
99.00914	97.33448	C	nao
100.76016	101.13691	C	nao

Quais **conclusões**
tiramos ao analisar
esta tabela de
dados?

Por quê é importante fazer gráficos?

Qual você prefere, os dados em uma tabela ou em um gráfico?

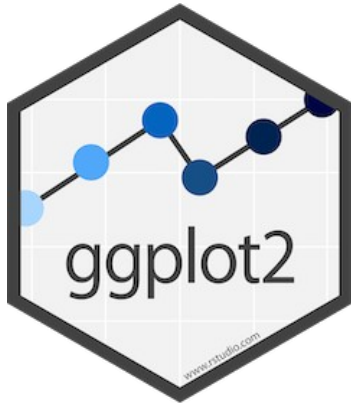
x	y
1	-243.551419
2	-172.228424
3	-59.950273
4	67.909653
5	560.770234
6	42.851166
7	72.787562
8	245.231018
9	38.882495
10	4.698586
11	31.223308
12	368.234669
13	27.771809
14	145.202189
15	198.477820
16	527.238915
17	301.024828
18	527.731958
19	482.842026
20	719.413981
21	604.957292
22	436.225368



Existem diversos pacotes no R-base:



- Built-in package (função `plot()`)
- `htmlwidgets`
- `plot_ly`
- **Ggplot2**
- ...



Introdução ao ggplot2:

ggplot2 = gramática dos gráficos



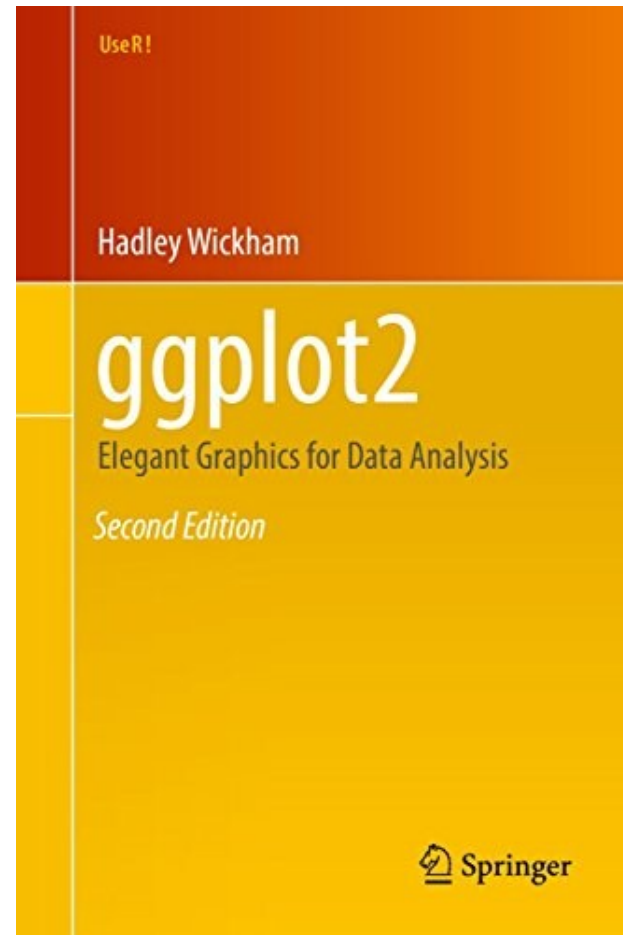
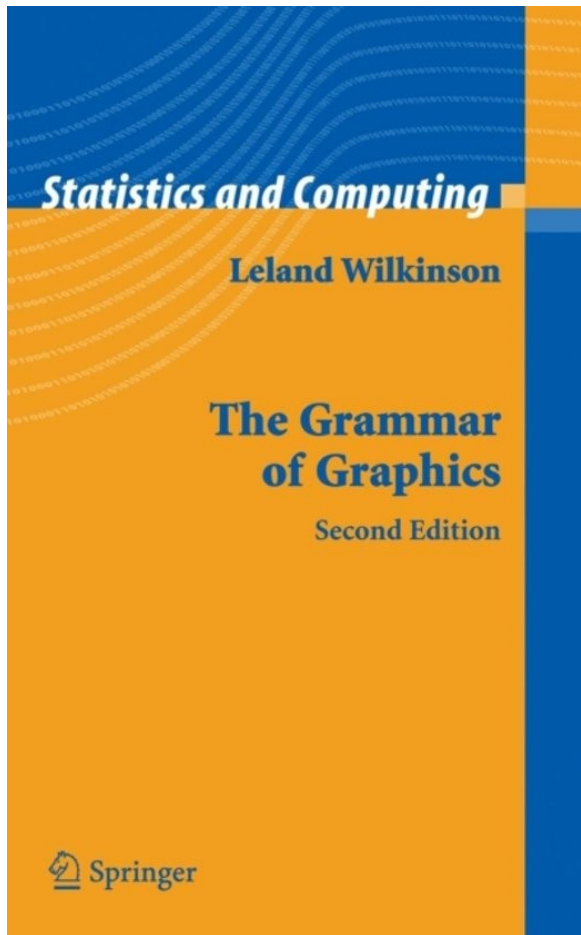
Breve história do ggplot2:



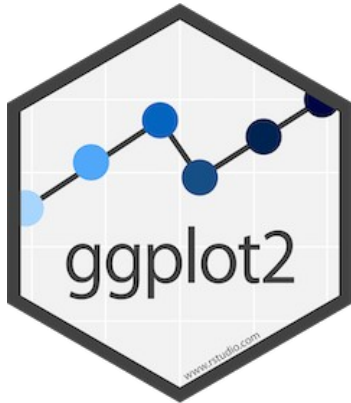
Hadley Wickham
Cientista Chefe do RStudio
Stanford University

Criador do ggplot2

Implementa uma adaptação
da gramática dos gráficos
em "layers".



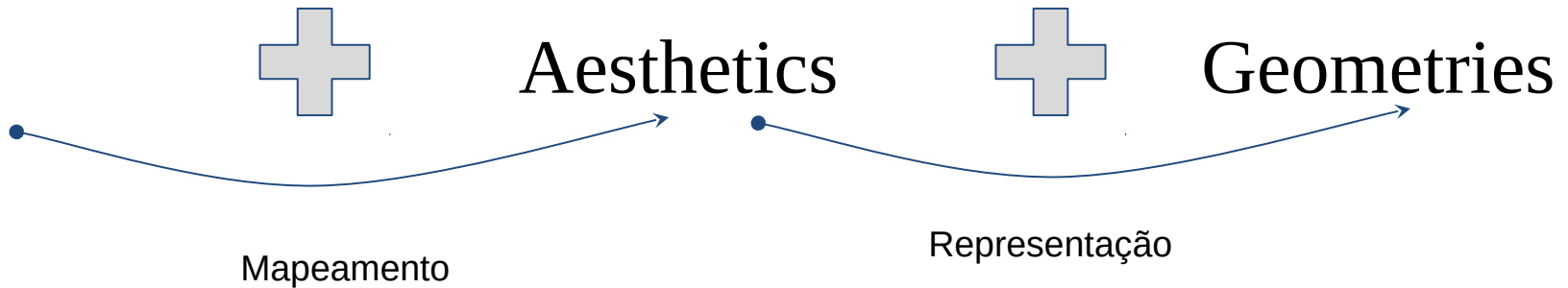
Por quê o ggplot2?



- Código livre e adaptável
- Comunidade forte
- Sintaxe muito mais simples que outros pacotes
- Edição de gráficos mais rápida e fácil
- Gera gráficos de alta qualidade para publicações, sites, apresentações, etc

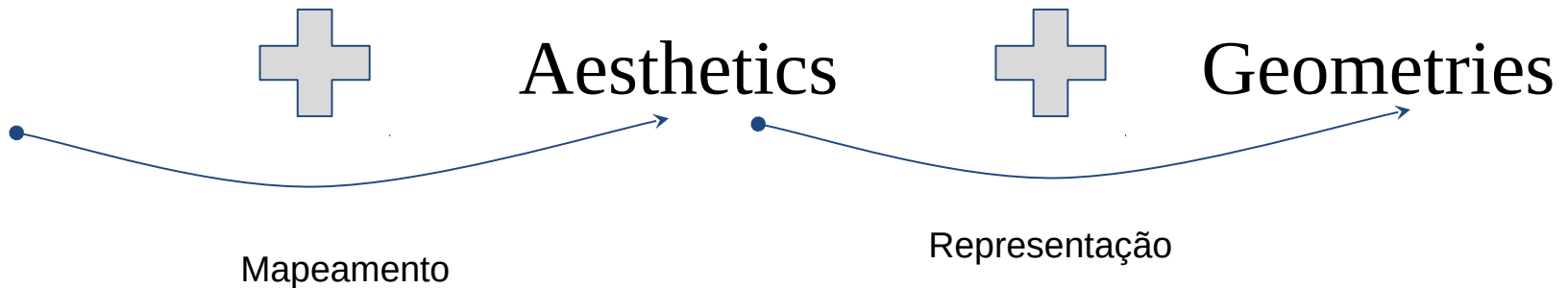
O que é um gráfico?

A representação de variáveis mapeadas em atributos estéticos de objetos geométricos



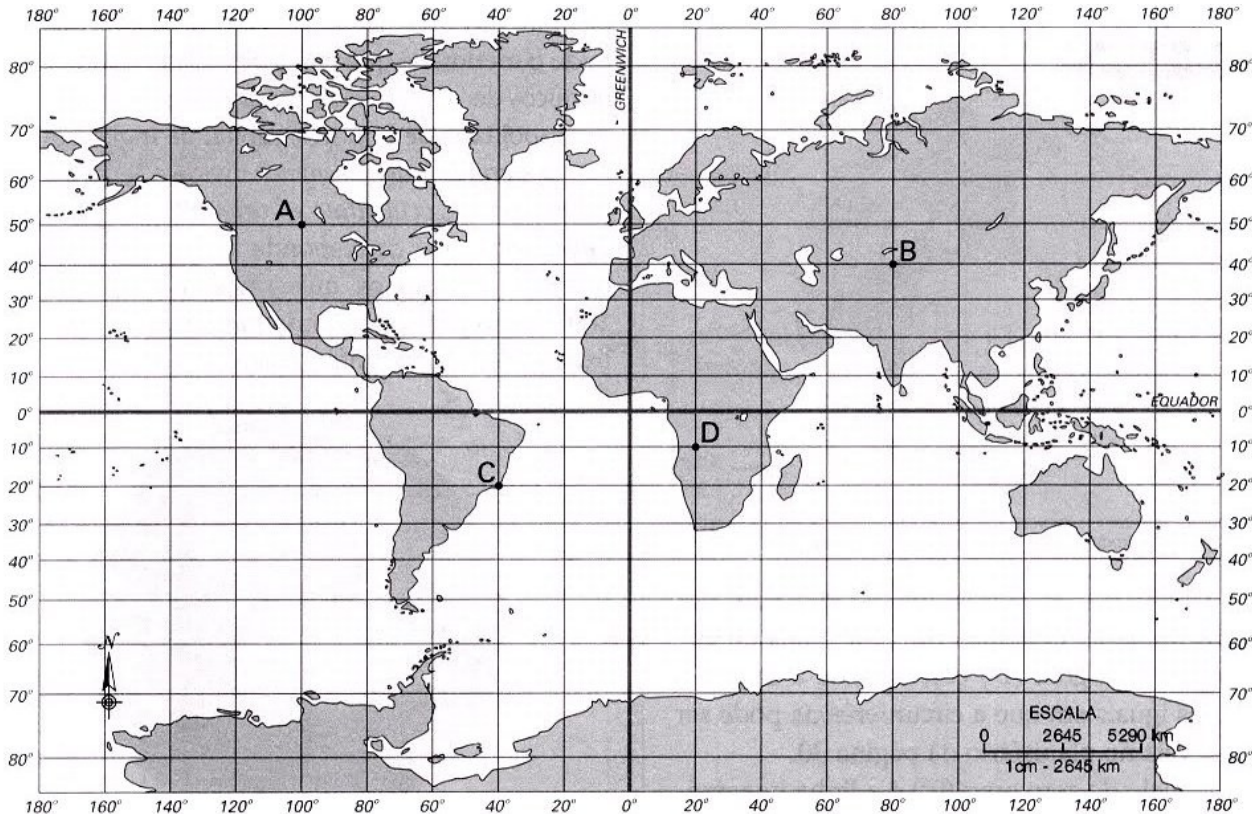
O que é um gráfico?

A representação de variáveis mapeadas em atributos estéticos de objetos geométricos

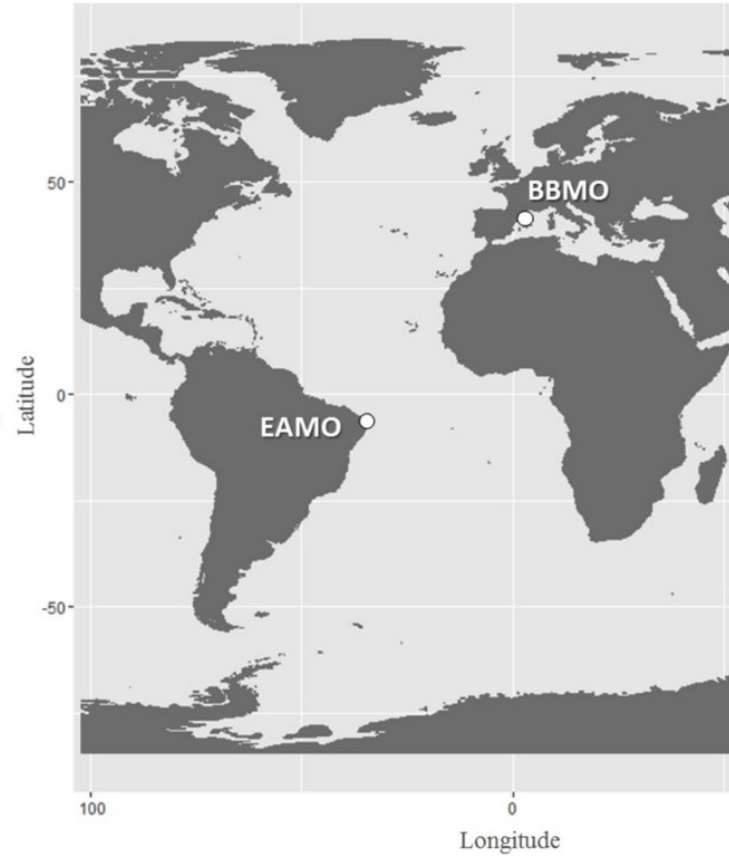
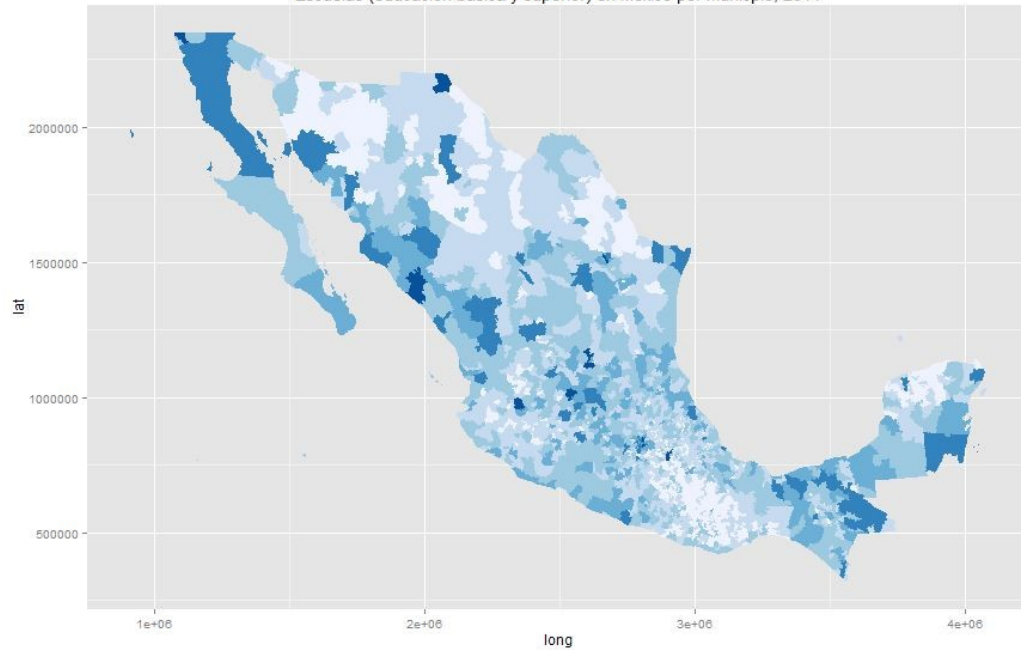


Como funciona o ggplot2?

- Sistema de coordenadas: “mapeamento” dos dados



Escuelas (educación básica y superior) en México por municipio, 2011



Como funciona o ggplot2?

- Duas variáveis contínuas

Dados

TOTAL	HIGH	MEDIUM	LOW
81	2	71	8
44	3	30	11
61	1	30	30
34	1	30	3
34	0	34	0
34	0	34	0
28	5	5	18
32	0	32	0
28	0	28	0
26	1	20	5
22	0	22	0
18	1	14	3
18	0	18	0
9	3	0	6
3	3	0	0

Como funciona o ggplot2?

- Duas variáveis contínuas

Dados

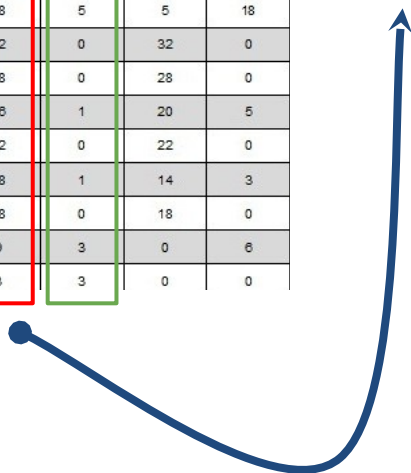
+

Mapeamento

TOTAL	HIGH	MEDIUM	LOW
81	2	71	8
44	3	30	11
61	1	30	30
34	1	30	3
34	0	34	0
34	0	34	0
28	5	5	18
32	0	32	0
28	0	28	0
26	1	20	5
22	0	22	0
18	1	14	3
18	0	18	0
9	3	0	6
3	3	0	0

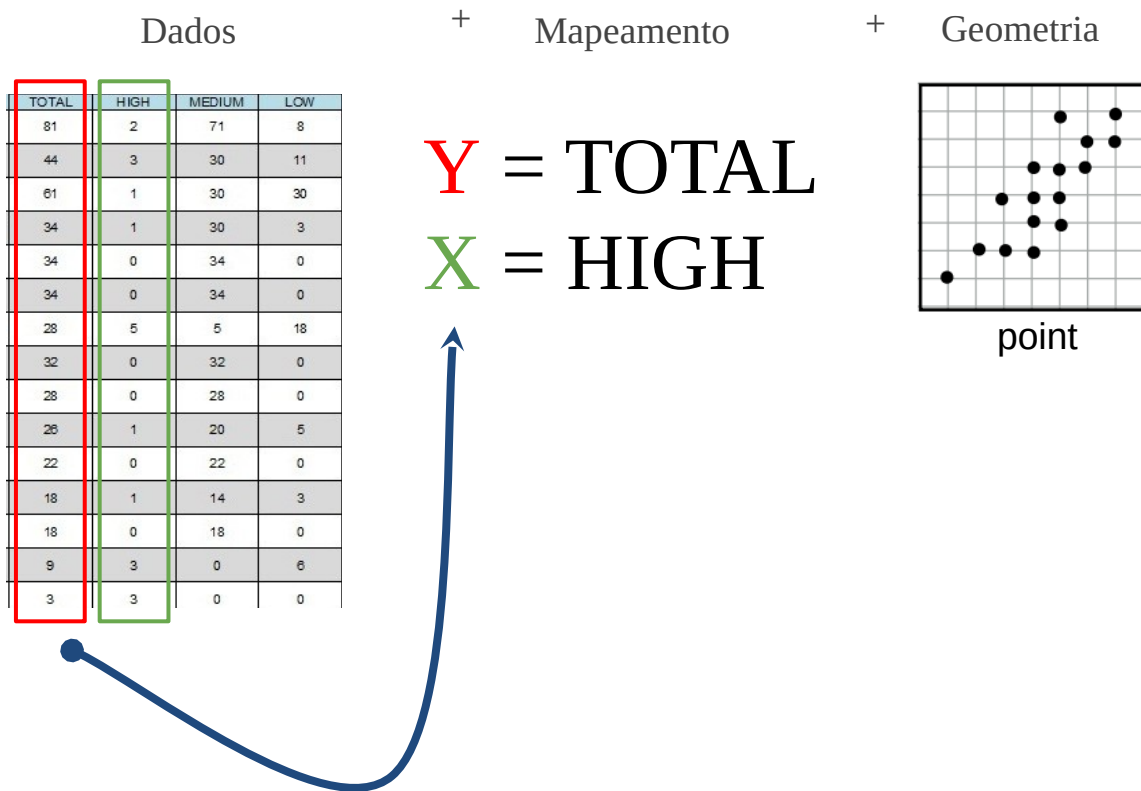
Y = TOTAL

X = HIGH



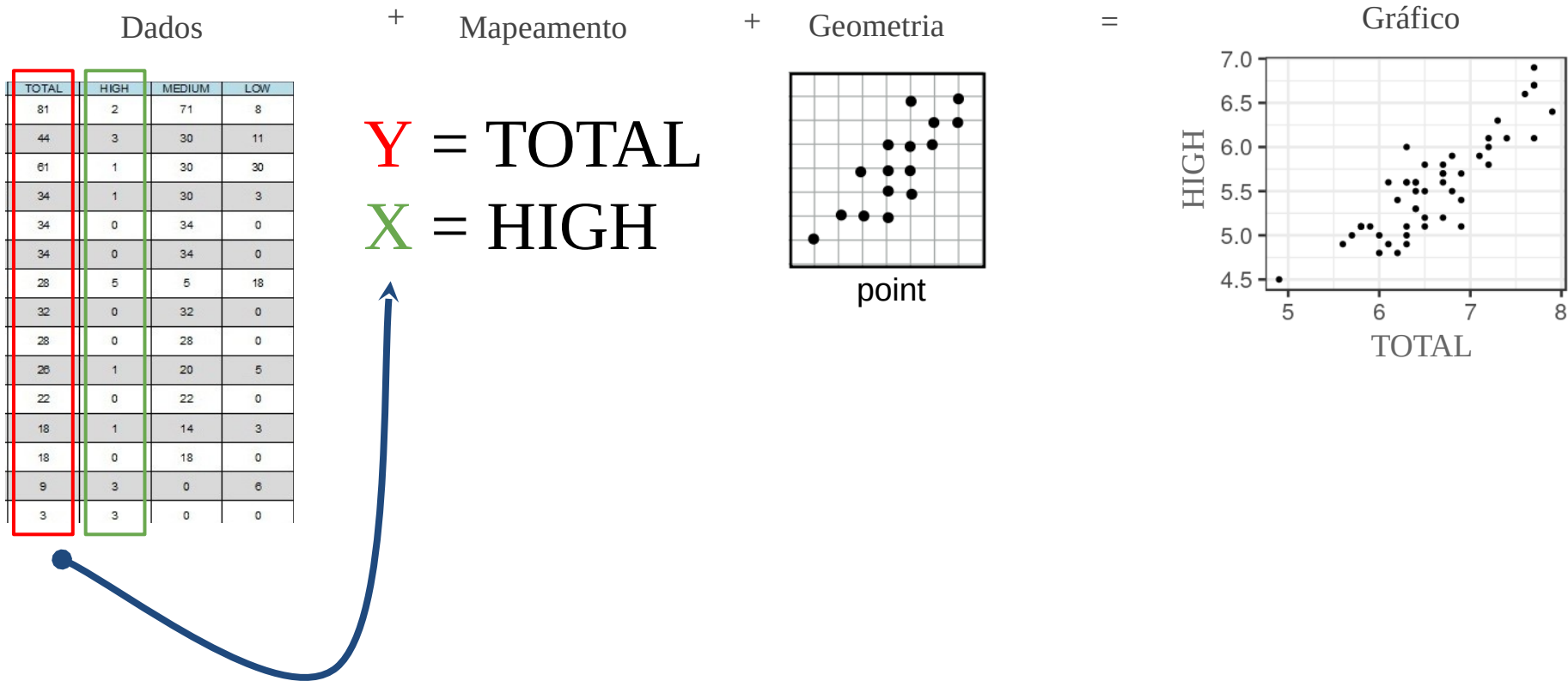
Como funciona o ggplot2?

- Duas variáveis contínuas



Como funciona o ggplot2?

- Duas variáveis contínuas



Como funciona o ggplot2?

- Uma variável contínua

Dados

TOTAL	HIGH	MEDIUM	LOW
81	2	71	8
44	3	30	11
61	1	30	30
34	1	30	3
34	0	34	0
34	0	34	0
28	5	5	18
32	0	32	0
28	0	28	0
26	1	20	5
22	0	22	0
18	1	14	3
18	0	18	0
9	3	0	6
3	3	0	0

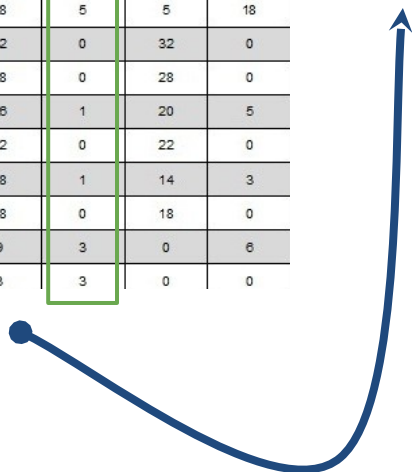
Como funciona o ggplot2?

- Uma variável contínua

Dados + Mapeamento

TOTAL	HIGH	MEDIUM	LOW
81	2	71	8
44	3	30	11
61	1	30	30
34	1	30	3
34	0	34	0
34	0	34	0
28	5	5	18
32	0	32	0
28	0	28	0
26	1	20	5
22	0	22	0
18	1	14	3
18	0	18	0
9	3	0	6
3	3	0	0

X = HIGH



Como funciona o ggplot2?

- Uma variável contínua

Dados

+

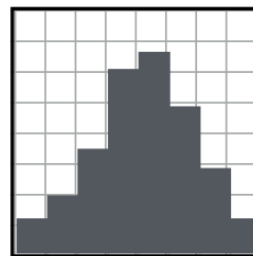
Mapeamento

+

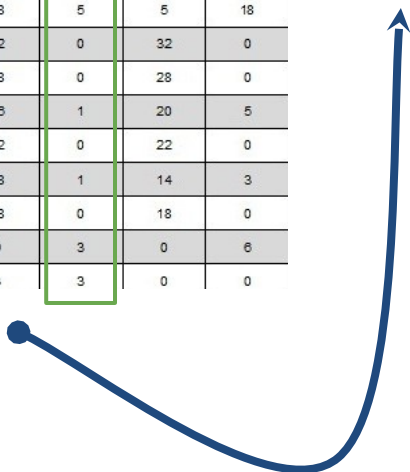
Geometria

TOTAL	HIGH	MEDIUM	LOW
81	2	71	8
44	3	30	11
61	1	30	30
34	1	30	3
34	0	34	0
34	0	34	0
28	5	5	18
32	0	32	0
28	0	28	0
26	1	20	5
22	0	22	0
18	1	14	3
18	0	18	0
9	3	0	6
3	3	0	0

$X = \text{HIGH}$



histogram



Como funciona o ggplot2?

- Uma variável contínua

Dados

+

Mapeamento

+

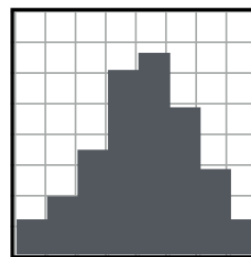
Geometria

=

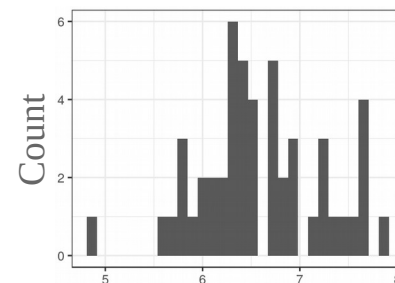
Gráfico

TOTAL	HIGH	MEDIUM	LOW
81	2	71	8
44	3	30	11
61	1	30	30
34	1	30	3
34	0	34	0
34	0	34	0
28	5	5	18
32	0	32	0
28	0	28	0
26	1	20	5
22	0	22	0
18	1	14	3
18	0	18	0
9	3	0	6
3	3	0	0

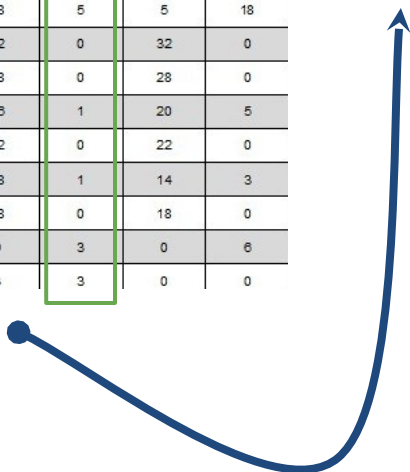
X = HIGH



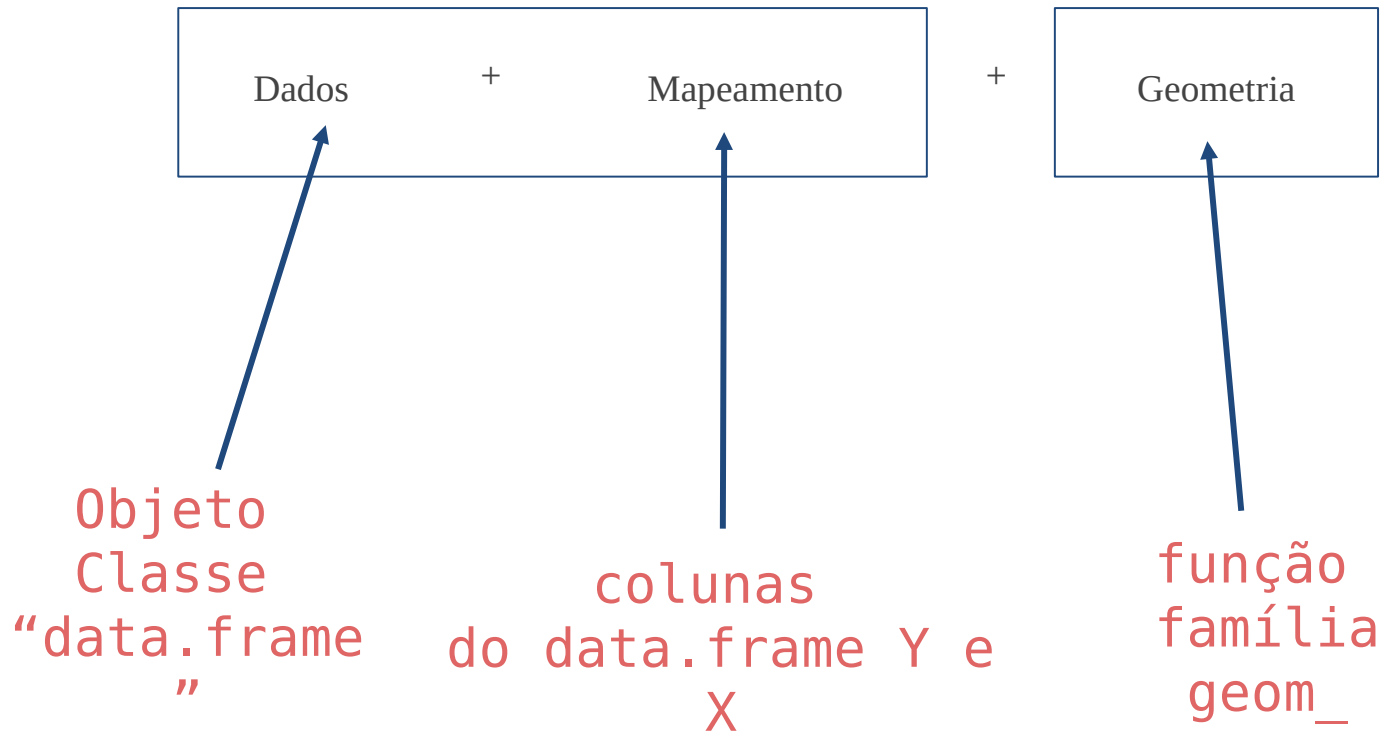
histogram



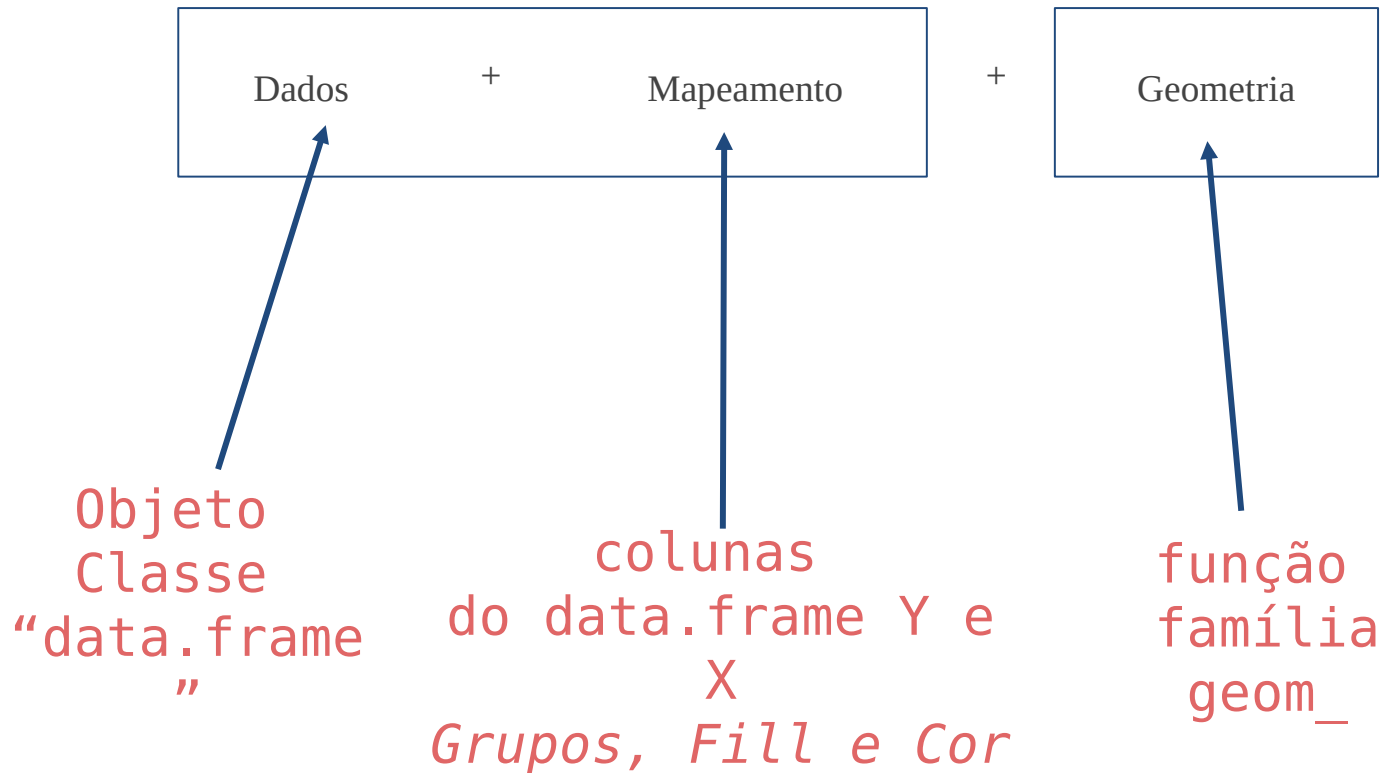
HIGH

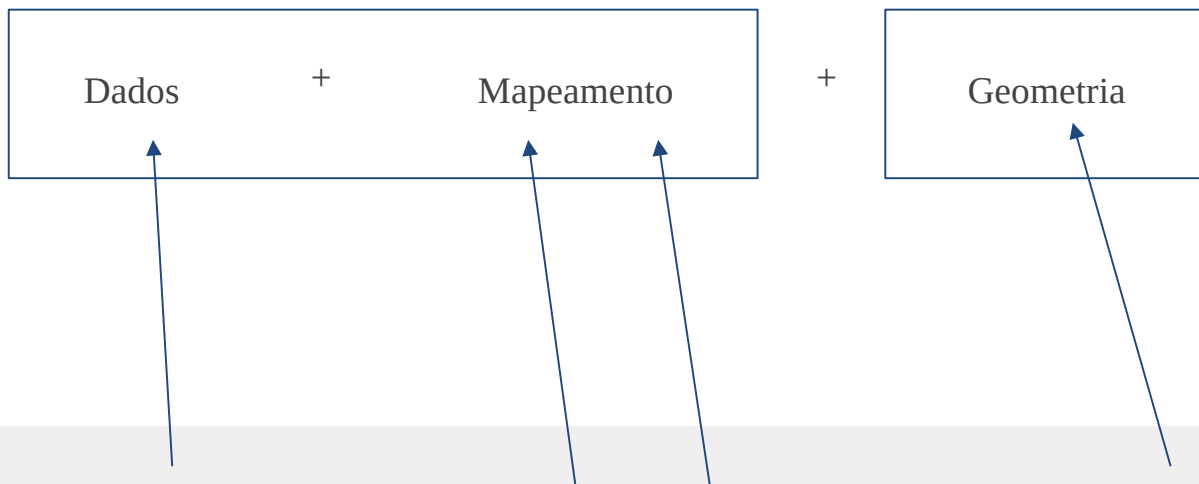


Sintaxe do ggplot2 no R:



Sintaxe do ggplot2 no R:






```
ggplot(data, aes(x, y)) + geom_point()
```

data = Define o objeto com a tabela de dados

aes = Define qual coluna é X e qual é Y

geom_ = Define qual geometria plotar

```
ggplot(data, aes(x, y)) +  
  geom_point() +  
  etc
```



O sinal de mais é utilizado para **adicionar** mais camadas.

Recomendação: colocar apenas uma camada por linha para facilitar a leitura e o pensamento (principalmente quando se tem muitas camadas adicionadas, o que é comum).

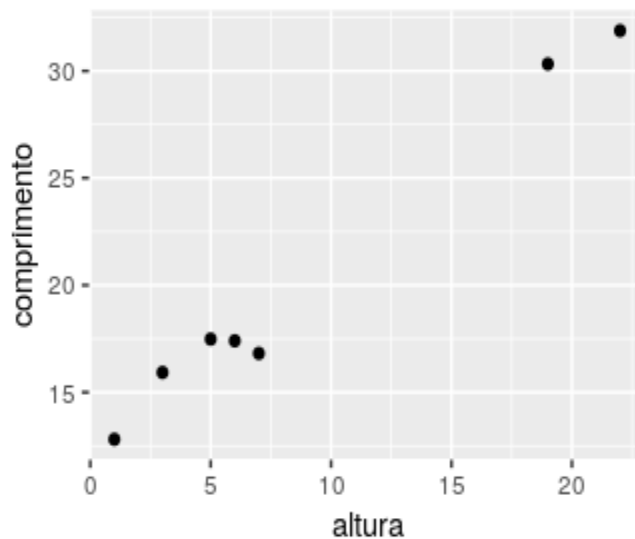
Alguns exemplos

```
ggplot(data = dados, aes(x = altura, y =  
comprimento)) + geom_point()
```

altura	comprimento
1	12.80485
3	15.92552
5	17.48298
6	17.40369
7	16.81471
19	30.32513
22	31.88094

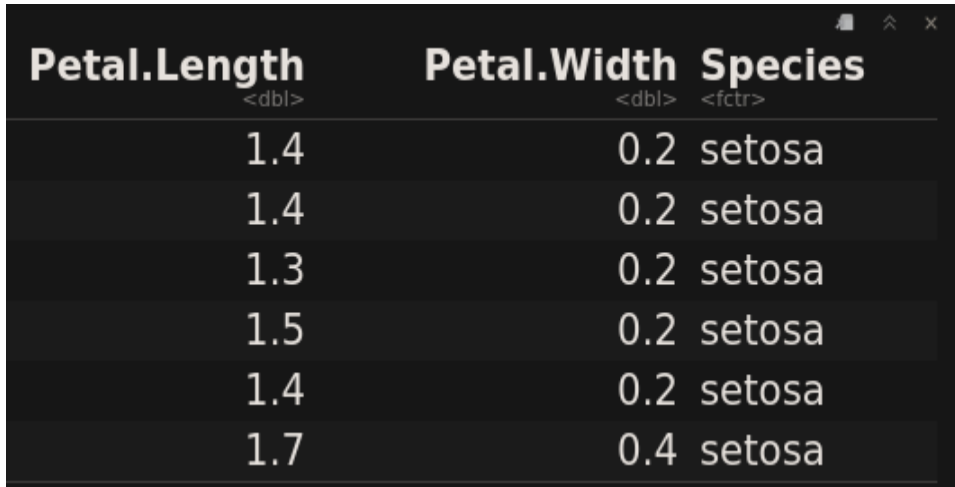
X

Y



Mapeando das variáveis no ggplot

Tabela de dados



Petal.Length <dbl>	Petal.Width <dbl>	Species <fctr>
1.4	0.2	setosa
1.4	0.2	setosa
1.3	0.2	setosa
1.5	0.2	setosa
1.4	0.2	setosa
1.7	0.4	setosa

Tabela de dados

Petal.Length <dbl>	Petal.Width <dbl>	Species <fctr>
1.4	0.2	setosa
1.4	0.2	setosa
1.3	0.2	setosa
1.5	0.2	setosa
1.4	0.2	setosa
1.7	0.4	setosa



Y

Mapeamento



X

Tabela de dados

Petal.Length <dbl>	Petal.Width <dbl>	Species <fctr>
1.4	0.2	setosa
1.4	0.2	setosa
1.3	0.2	setosa
1.5	0.2	setosa
1.4	0.2	setosa
1.7	0.4	setosa



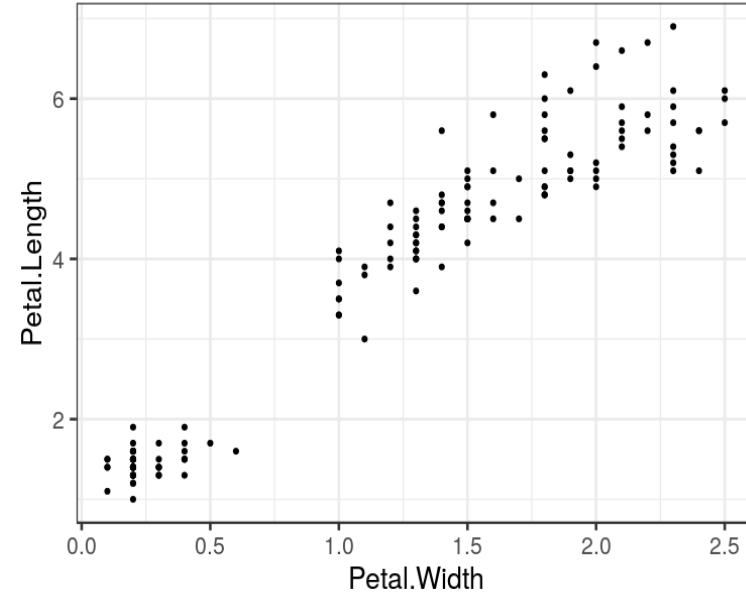
Y

Mapeamento



X

Gráfico



Geometria =



Tabela de dados

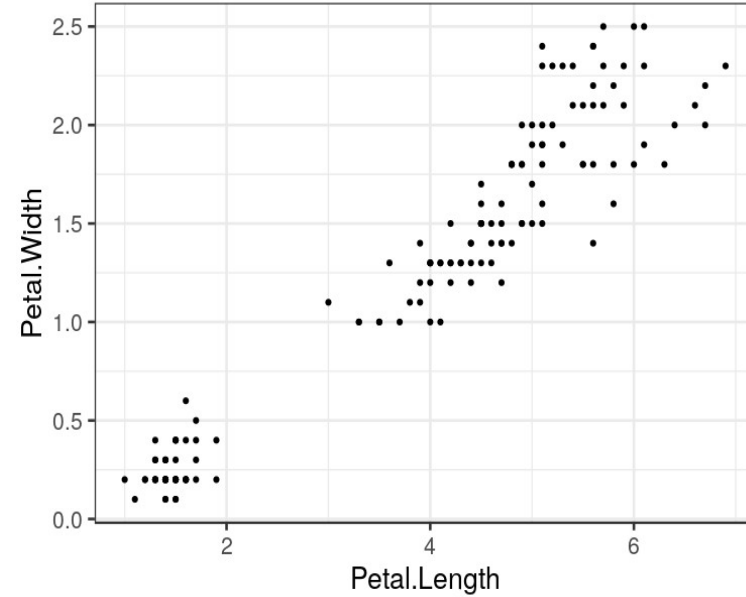
Petal.Length <dbl>	Petal.Width <dbl>	Species <fctr>
1.4	0.2	setosa
1.4	0.2	setosa
1.3	0.2	setosa
1.5	0.2	setosa
1.4	0.2	setosa
1.7	0.4	setosa

Mapeamento

X

Y

Gráfico



Geometria =



Tabela de dados

Petal.Length <dbl>	Petal.Width <dbl>	Species <fctr>
1.4	0.2	setosa
1.4	0.2	setosa
1.3	0.2	setosa
1.5	0.2	setosa
1.4	0.2	setosa
1.7	0.4	setosa

Mapeamento

X

Y

color

Geometria = ●

Gráfico

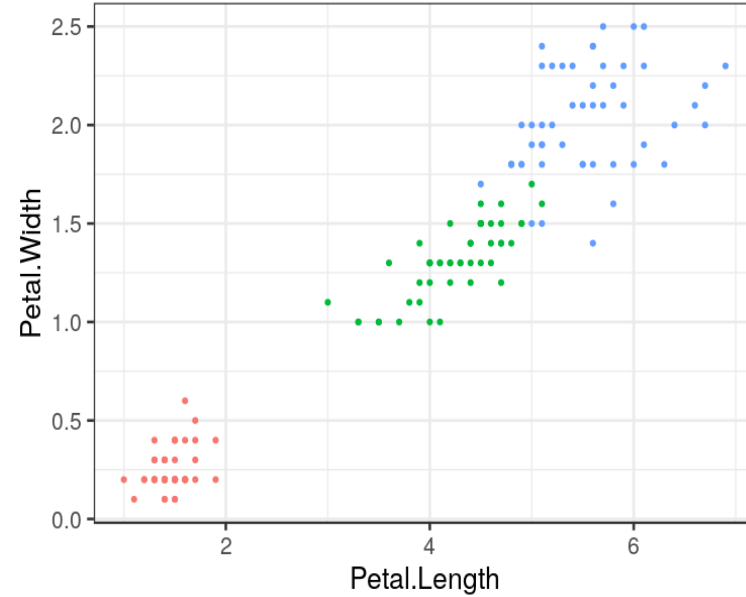


Tabela de dados

Petal.Length <dbl>	Petal.Width <dbl>	Species <fctr>
1.4	0.2	setosa
1.4	0.2	setosa
1.3	0.2	setosa
1.5	0.2	setosa
1.4	0.2	setosa
1.7	0.4	setosa



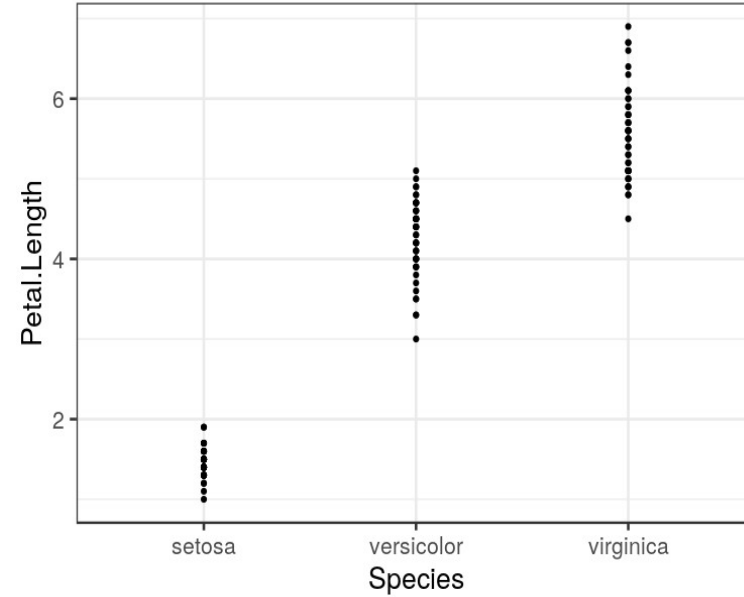
Y

Mapeamento



X

Gráfico



Geometria = ●

Tabela de dados

Petal.Length <dbl>	Petal.Width <dbl>	Species <fctr>
1.4	0.2	setosa
1.4	0.2	setosa
1.3	0.2	setosa
1.5	0.2	setosa
1.4	0.2	setosa
1.7	0.4	setosa



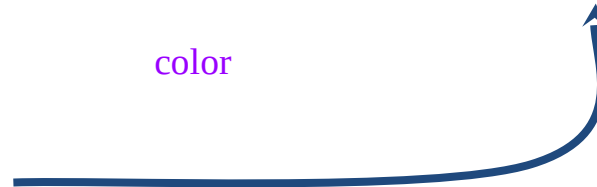
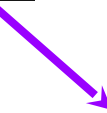
Y

Mapeamento



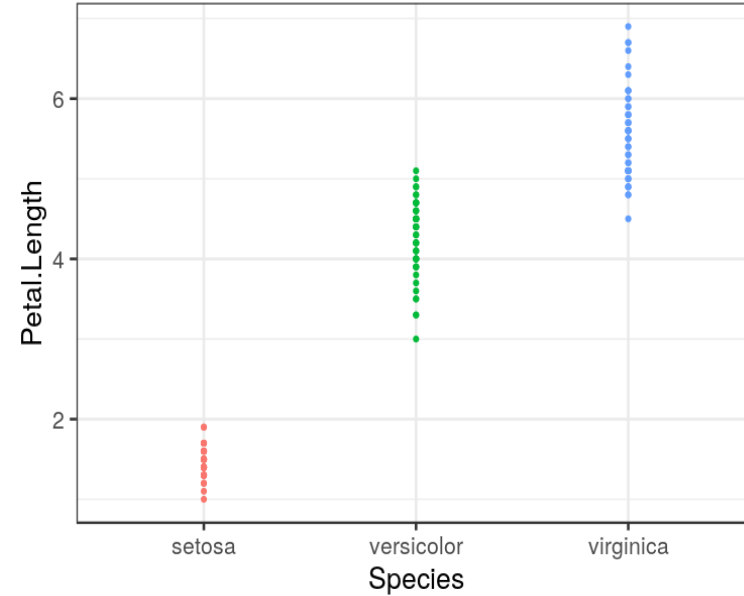
X

color



Geometria = ●

Gráfico



DESAFIO!

Desenhe no papel um gráfico a partir:

❖ Dados

	A	B	C	D
1980	1	3	hot	
1990	2	3	hot	
2000	3	2	cold	
2010	4	1	cold	

❖ Mapeamento

- Eixo **X** é a variável **A**
- Eixo **Y** é a variável **B**

❖ Geometria = ●

DESAFIO!

Desenhe no papel um gráfico a partir:

❖ Dados

A	B	C	D
1980	1	3	hot
1990	2	3	hot
2000	3	2	cold
2010	4	1	cold

❖ Mapeamento

- Eixo **X** é a variável **A**
- Eixo **Y** é a variável **B**
- **Cor** é a variável **D**

❖ Geometria = ●

DESAFIO!

Desenhe no papel um gráfico a partir:

❖ Dados

A	B	C	D
1980	1	3	hot
1990	2	3	hot
2000	3	2	cold
2010	4	1	cold

❖ Mapeamento

- Eixo **X** é a variável **A**
- Eixo **Y** é a variável **B**
- **Cor** é a variável **D**
- **Tamanho** é a variável **C**

❖ Geometria =



Resultado:

Mapeamento

- Eixo **X** é a variável **A**
- Eixo **Y** é a variável **B**

```
library(ggplot2)
ggplot(data = dados, mapping = aes(x = A, y = B)) +
  geom_point()
```

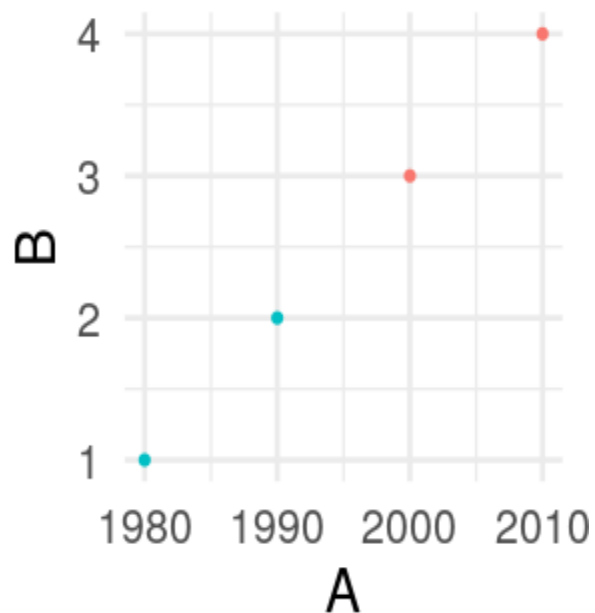


Resultado:

Mapeamento

- Eixo **X** é a variável **A**
- Eixo **Y** é a variável **B**
- **Cor** é a variável **D**

```
library(ggplot2)
ggplot(data = dados, mapping = aes(x = A, y = B, color = D)) +
  geom_point()
```

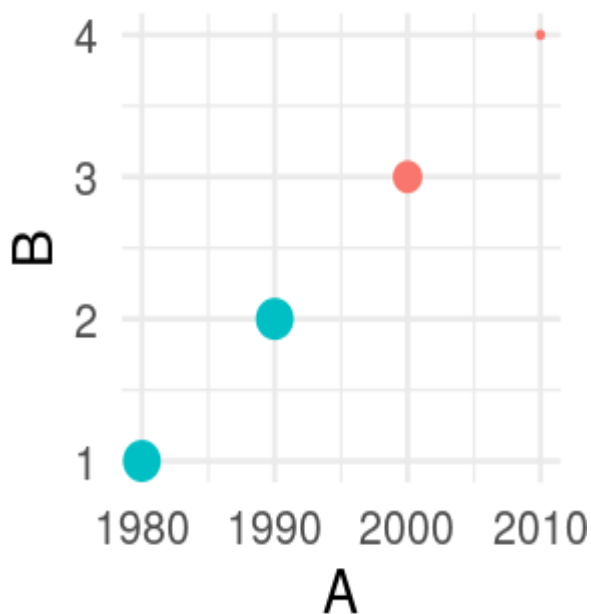


Resultado:

Mapeamento

- Eixo **X** é a variável **A**
- Eixo **Y** é a variável **B**
- **Cor** é a variável **D**
- **Tamanho** é a variável **C**

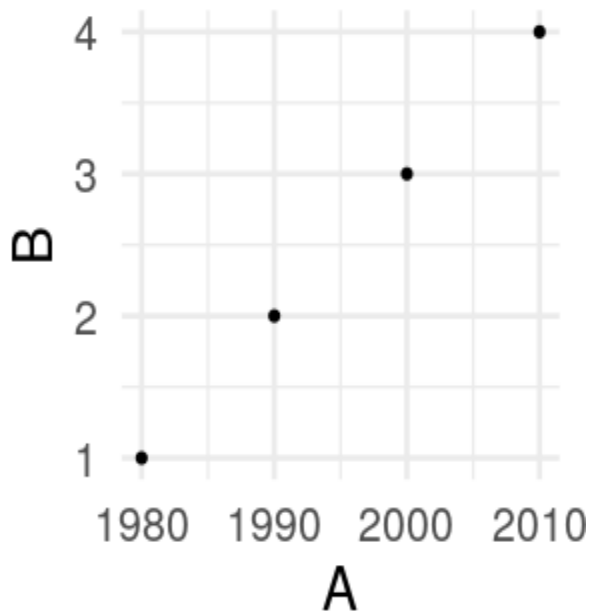
```
library(ggplot2)
ggplot(data = dados, mapping = aes(x = A, y = B, color = D, size = C)) +
  geom_point()
```



Resultados:

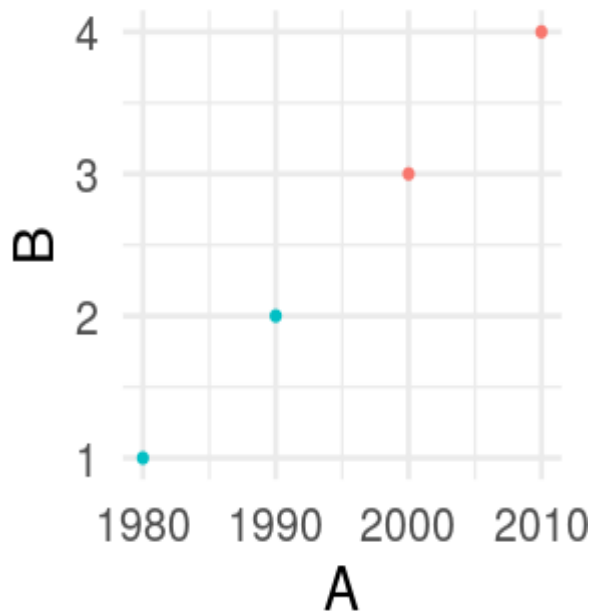
Mapeamento

- Eixo **X** é a variável **A**
- Eixo **Y** é a variável **B**



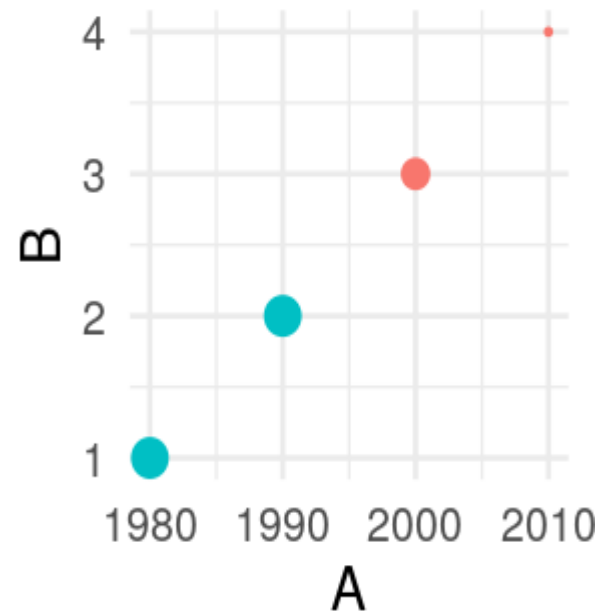
Mapeamento

- Eixo **X** é a variável **A**
- Eixo **Y** é a variável **B**
- **Cor** é a variável **D**



Mapeamento

- Eixo **X** é a variável **A**
- Eixo **Y** é a variável **B**
- **Cor** é a variável **D**
- **Tamanho** é a variável **C**



DESAFIO 2!

O que acontece se colocarmos cor na variável C?

❖ Dados

A	B	C	D
1980	1	3	hot
1990	2	3	hot
2000	3	2	cold
2010	4	1	cold

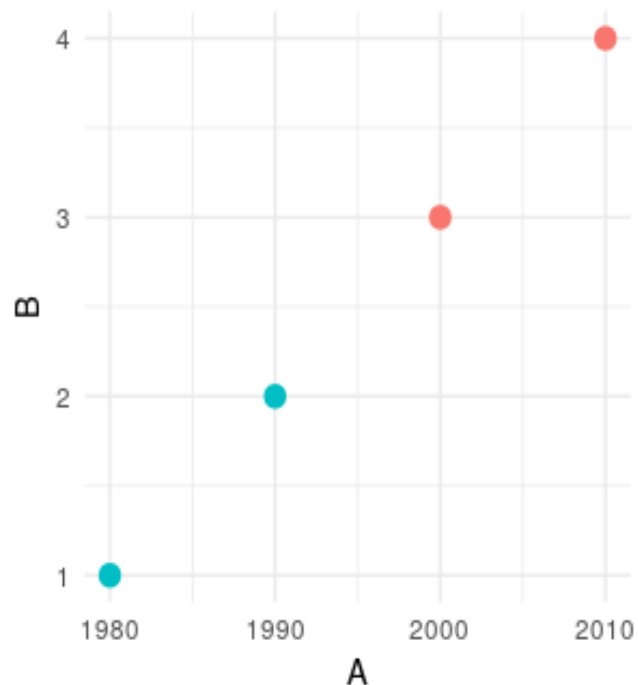
❖ Mapeamento

- Eixo **X** é a variável **A**
- Eixo **Y** é a variável **B**
- **Cor** é a variável **C**

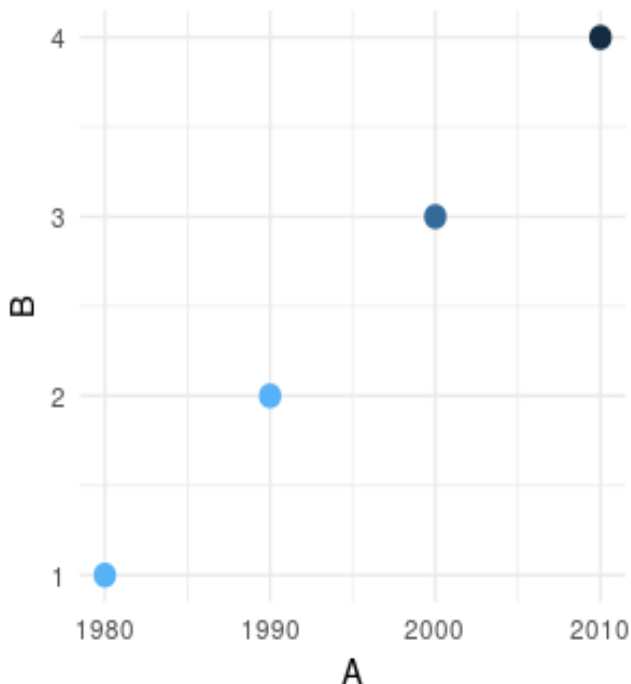
❖ Geometria = ●

Resultado:

Cor é a variável **D**



Cor é a variável **C**



A	B	C	D
1980	1	3	hot
1990	2	3	hot
2000	3	2	cold
2010	4	1	cold

C é numérica

D é texto (fator)

O ggplot compreende variáveis discretas e contínuas de forma distinta

Lógica das variáveis discretas:

O ggplot interpreta qualquer variável **categórica como discreta**



variáveis de texto geralmente são convertidas em fatores pelo R



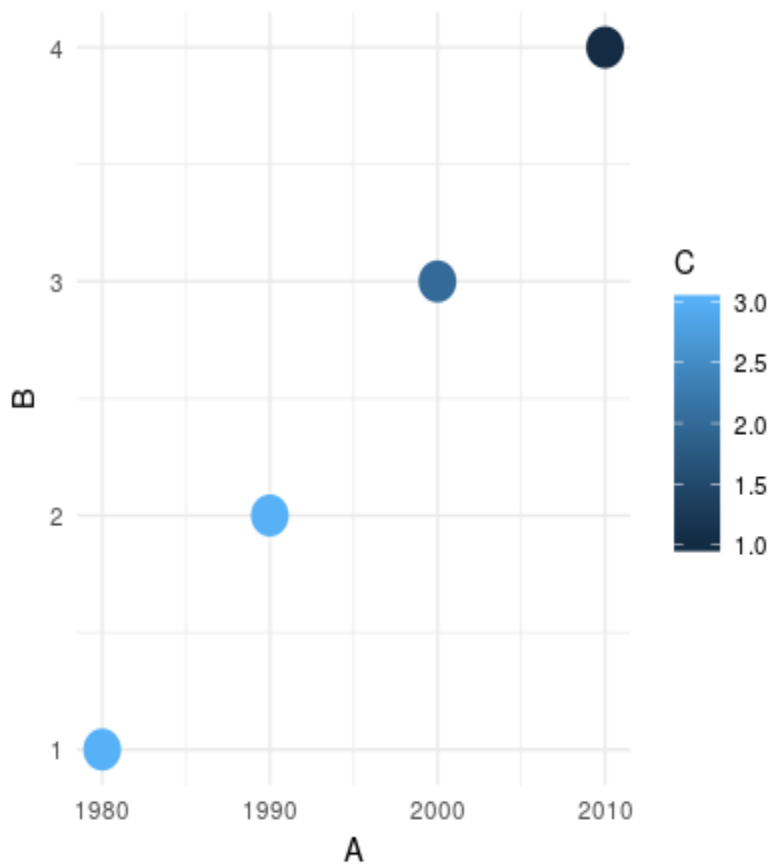
Fatores são considerados variáveis discretas pelo ggplot



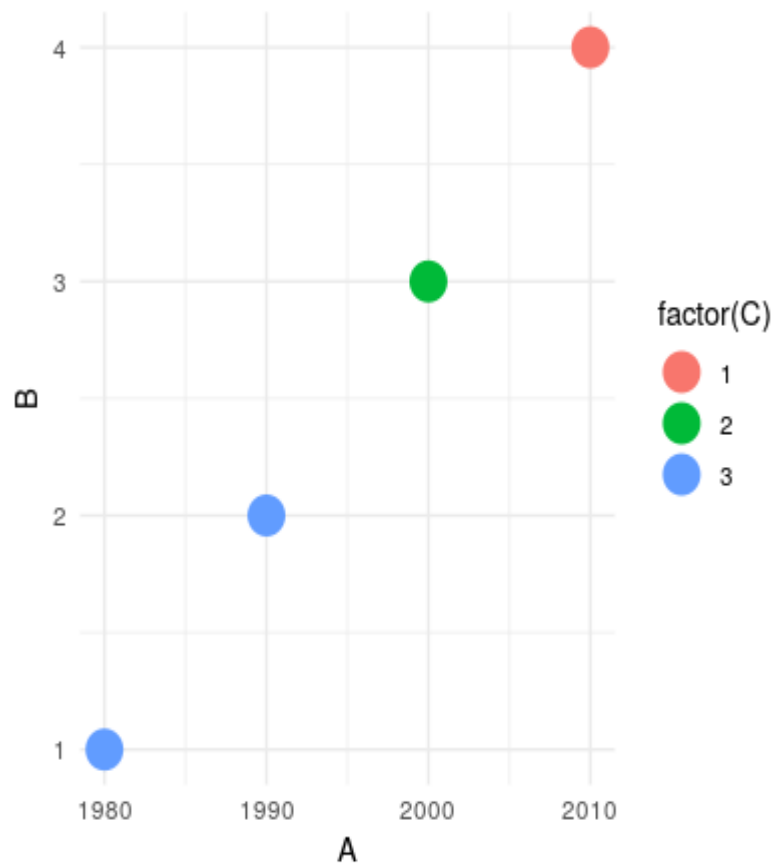
Variáveis discretas são mapeadas por níveis e não de forma contínua

Variáveis discretas vs Variáveis Contínuas

C é uma variável contínua
color = C

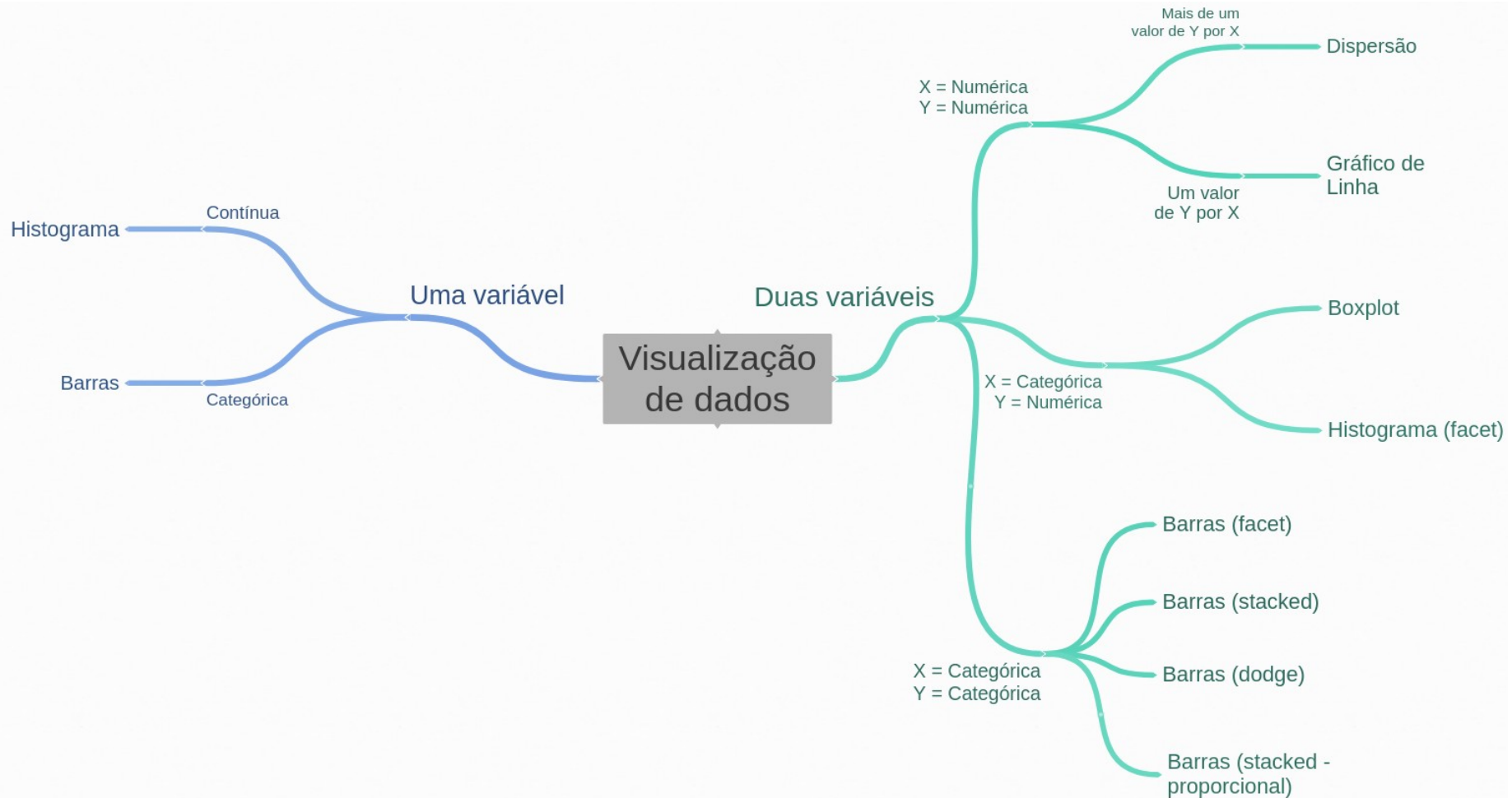


C é uma variável discreta
color = factor(C)



Gráficos: os cinco tipos básicos

Os cinco Gráficos Básicos



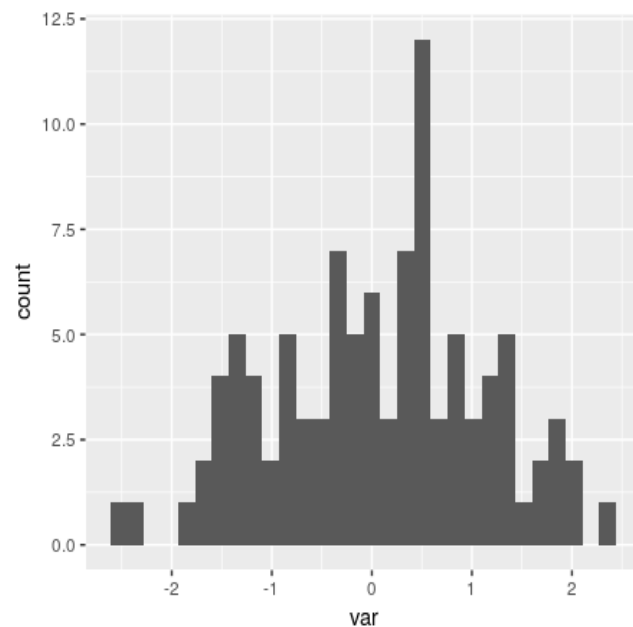
Uma variável - numérica

Histograma é uma representação da distribuição de uma variável numérica

Histograma → **geom_histogram()**

```
ggplot(dados, aes(x = var)) +  
  geom_histogram()
```

Variável
numérica



Uma variável - numérica

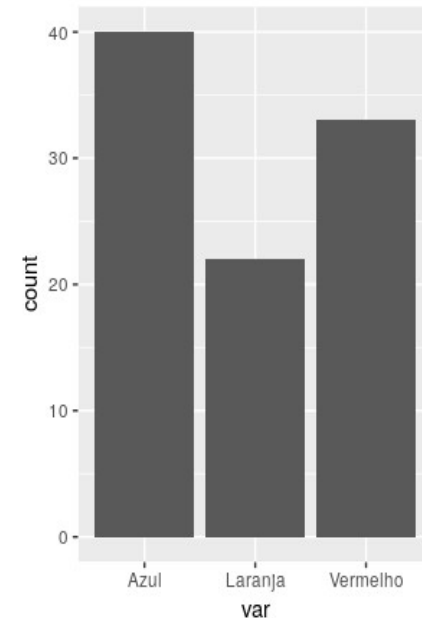
Gráfico de Barras é um gráfico que representa variáveis categóricas com barras retangulares e comprimento proporcional aos valores que ele representa

Barras

```
ggplot(dados, aes(x = var)) +  
  geom_bar()
```

Variável
categórica

geom_bar()



Uma variável - numérica

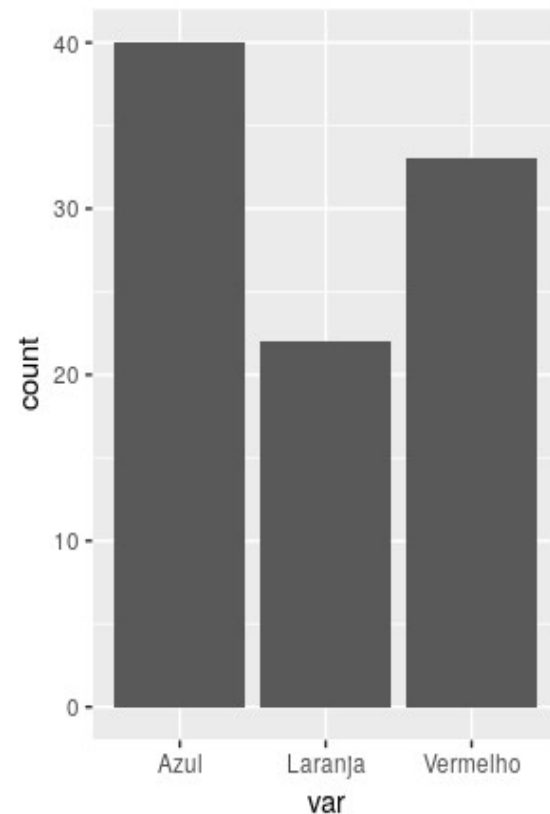
Dados

```
var  
Laranja  
Azul  
Azul  
Azul  
Laranja  
⋮
```

Número de
ocorrências de cada
Nível da variável

Cada nível do fator
da variável

geom_bar()



Duas variáveis - [Y = numérica | X = categórica]

Boxplot é um gráfico que representa quartis de uma variável numérica em função de uma variável categórica (grupos)

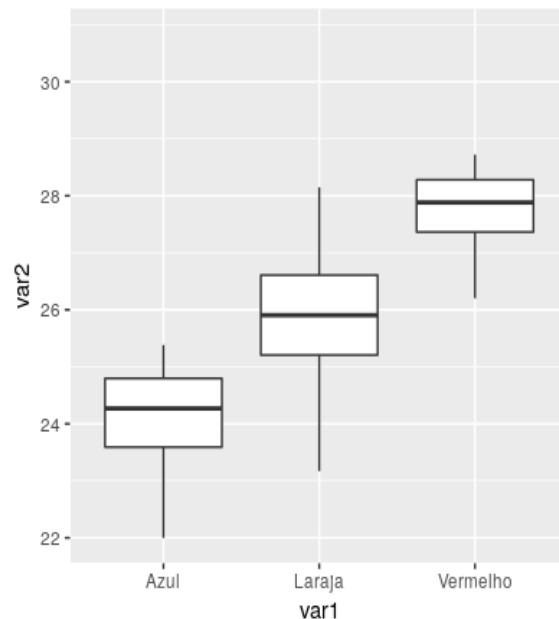
Boxplot

```
ggplot(dados, aes(x = var1, y = var2)) +  
  geom_boxplot()
```

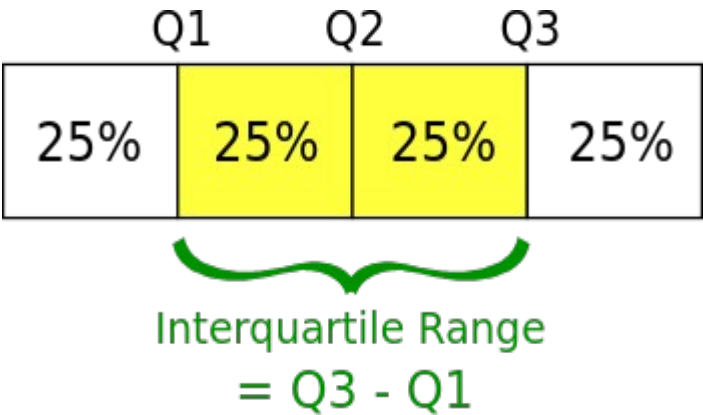
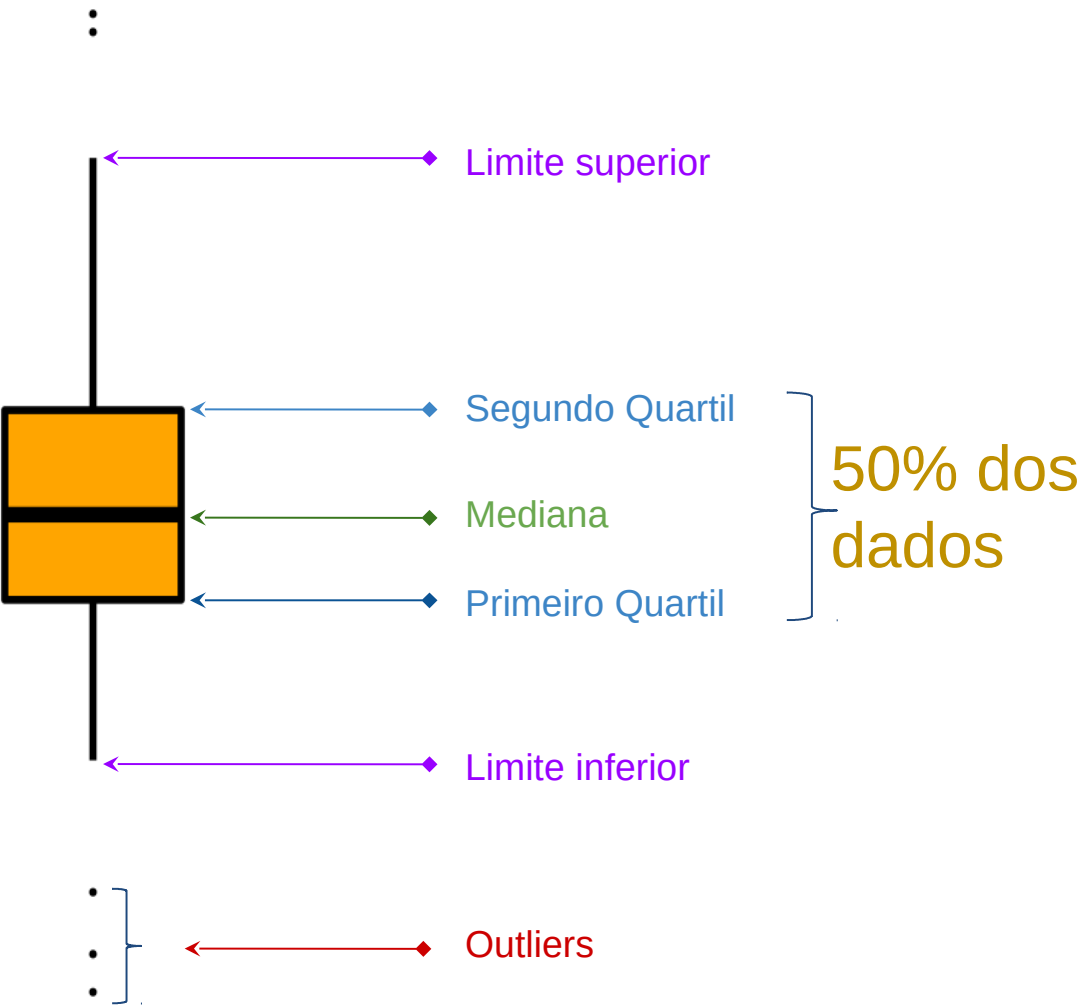
Variável
categórica

Variável
numérica

geom_boxplot()



Anatomia do boxplot



Duas variáveis - [Y = numérica | X = numérica]

Gráfico de dispersão é um gráfico que representa a relação entre duas variáveis numéricas

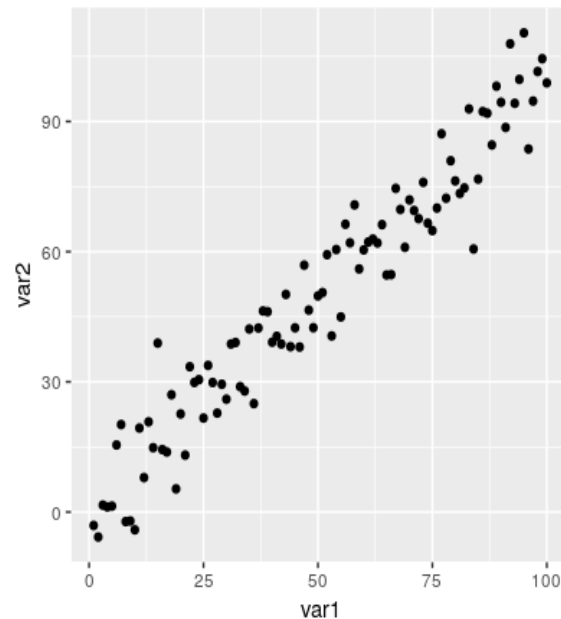
Dispersão

geom_point()

```
gplot(dados, aes(x = var1, y = var2))  
+  
  geom_point()
```

Variável
numérica

Variável
numérica



Duas variáveis - [Y = numérica | X = numérica] | **Um Y por X**

Gráfico de Linha representa informação através de dados em série, conectados por segmentos de linha reta

Linha

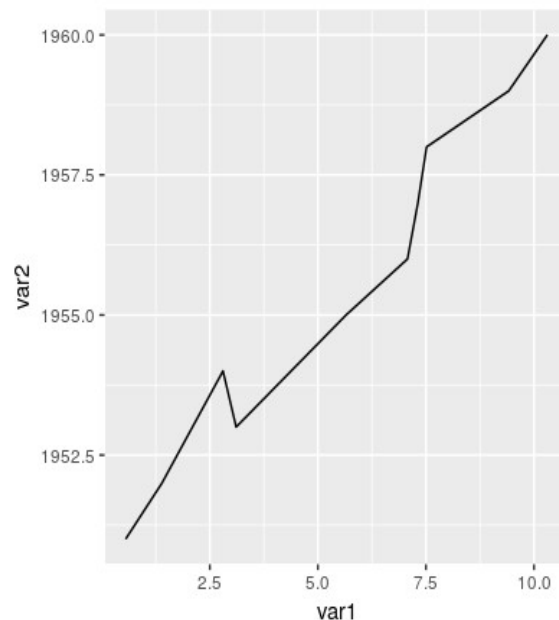


geom_line()

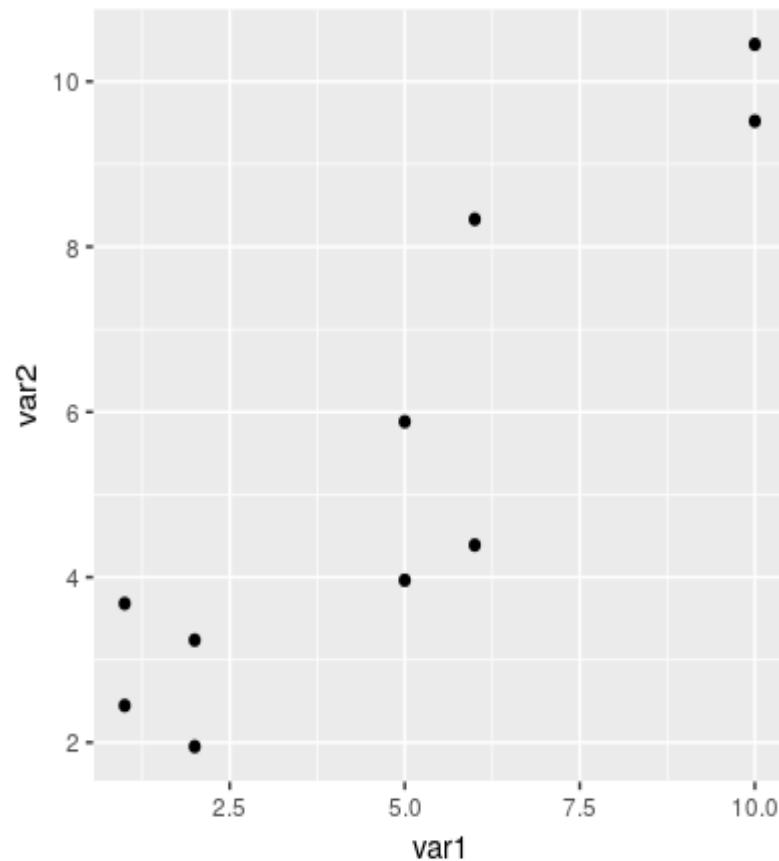
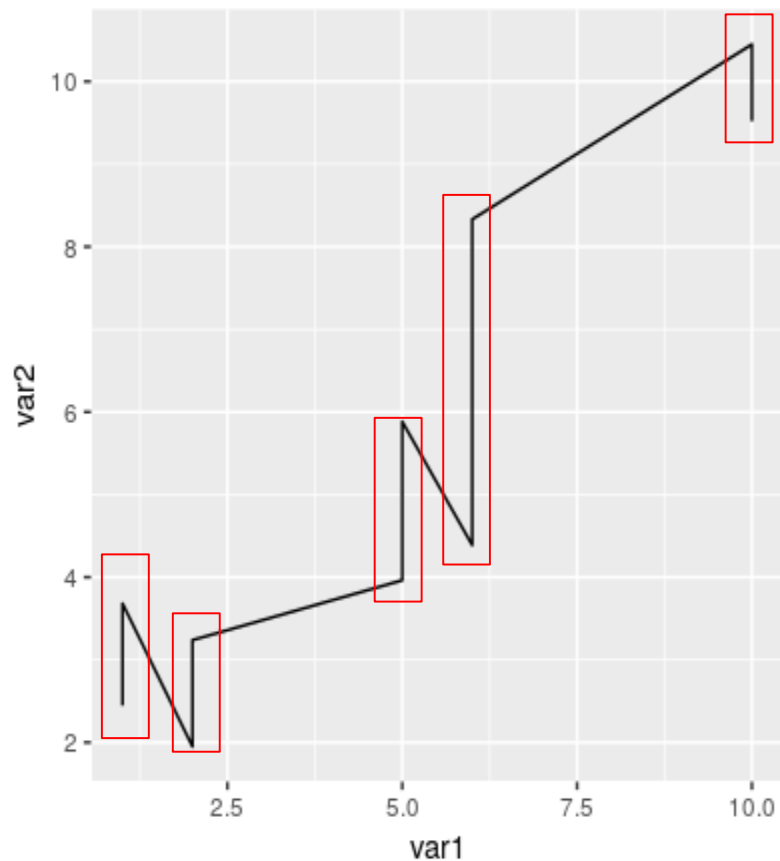
```
ggplot(dados, aes(x = var1, y = var2)) +  
  geom_line()
```

Variável
numérica

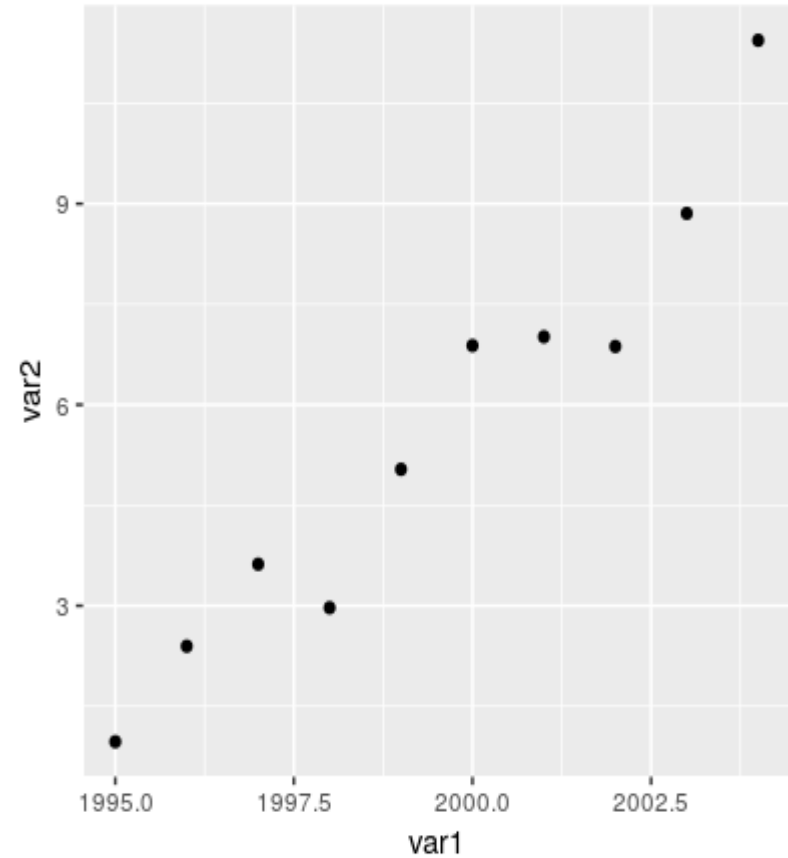
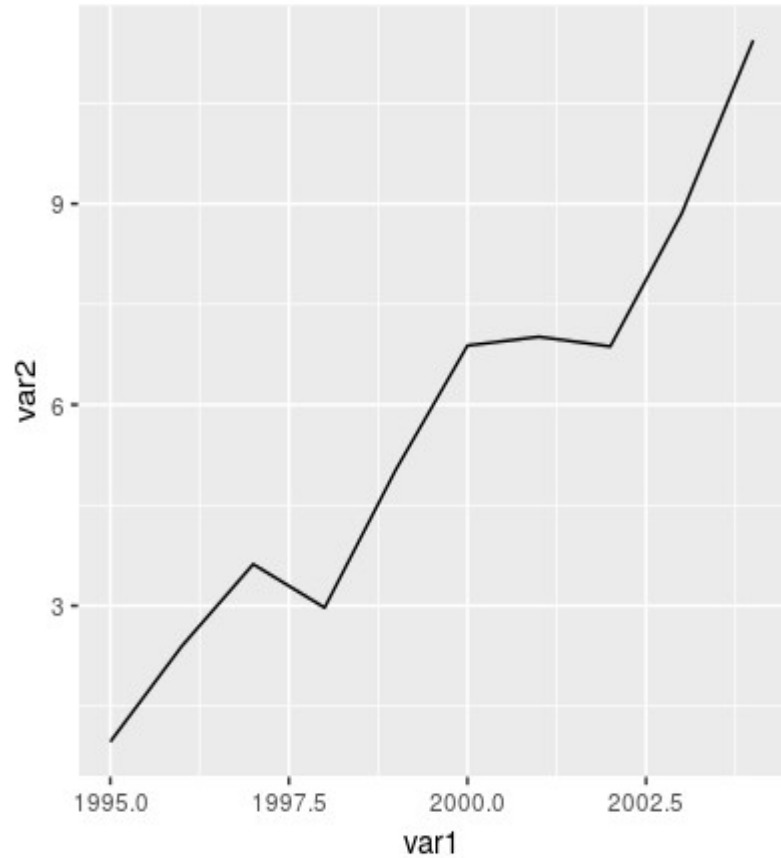
Variável
numérica



Mais de um valor de Y por X **não deve ser usado** no gráfico de linhas

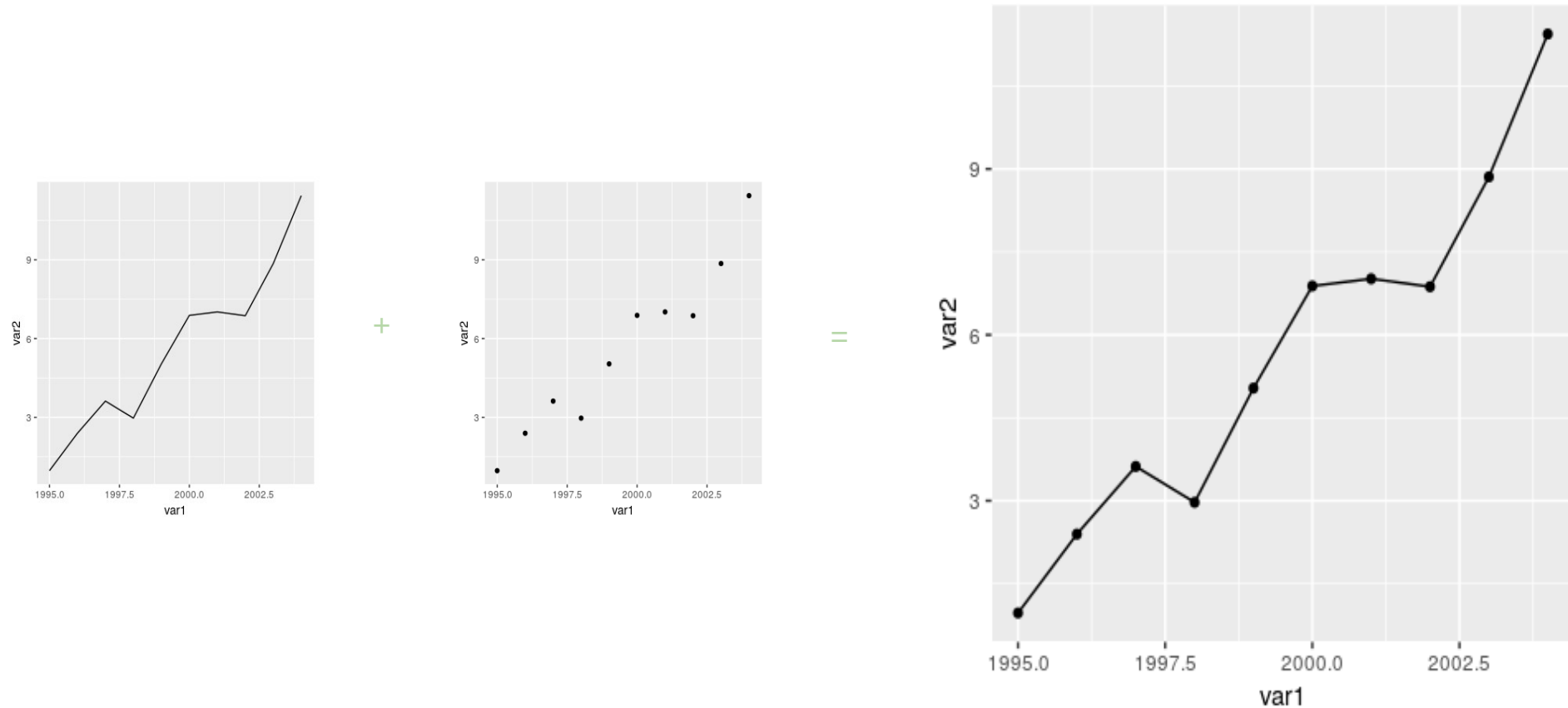


Mas o gráfico de dispersão pode ser utilizado | um valor de Y por X

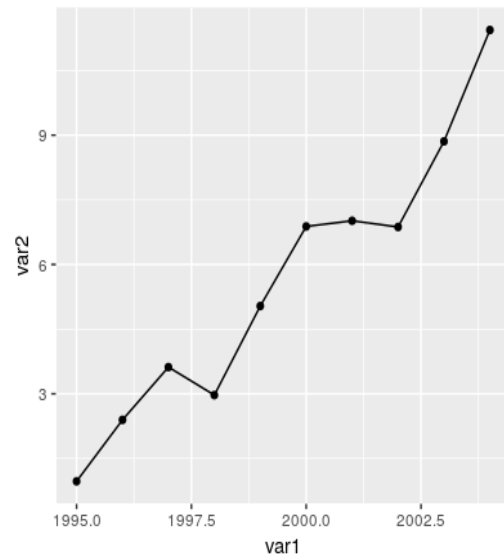


É possível combinar os dois?

Gráfico de Dispersão com Linha

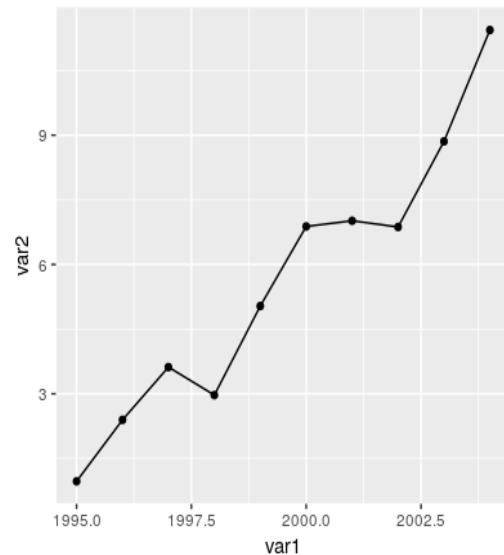
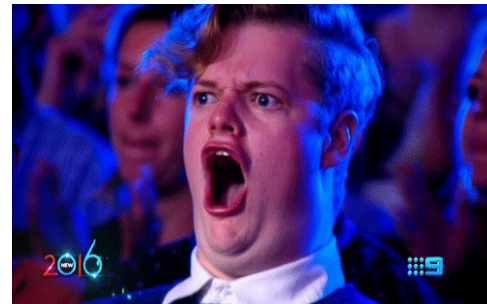


Desafio: Como seria o comando para juntar os dois?

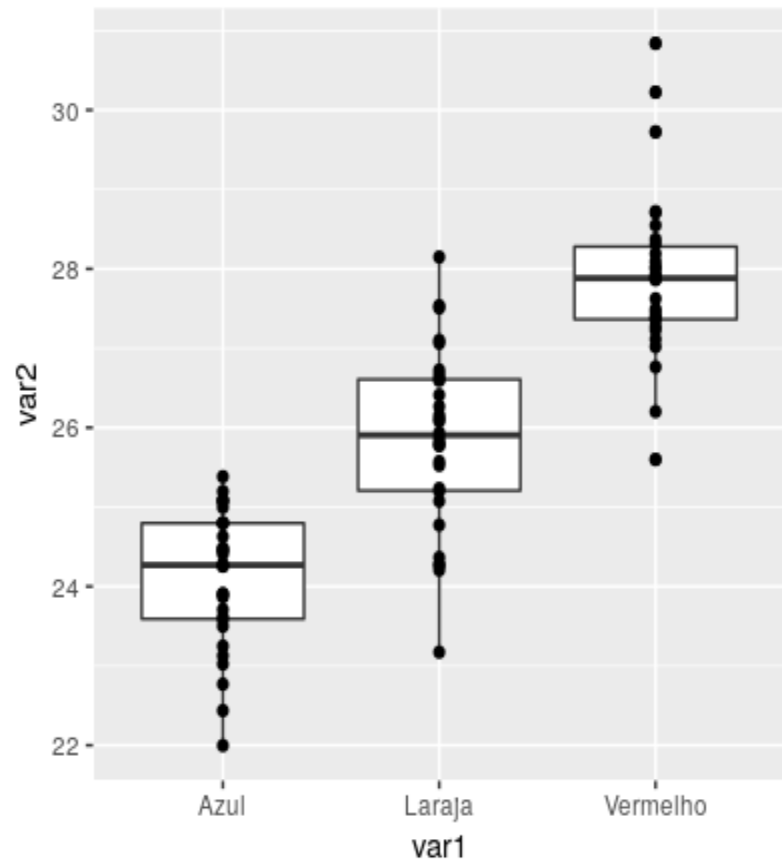


Desafio: Como seria o comando para juntar os dois?

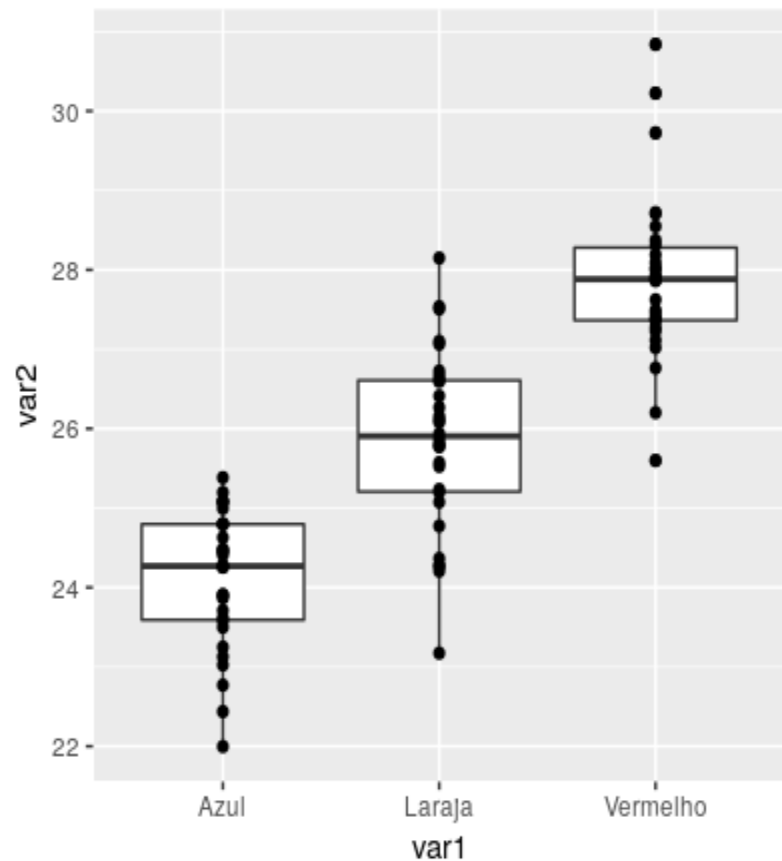
```
ggplot(dados, aes(x = var1, y = var2)) +  
  geom_line() +  
  geom_point()
```



Desafio 2: Como este gráfico foi feito?



Desafio 2: Como este gráfico foi feito?

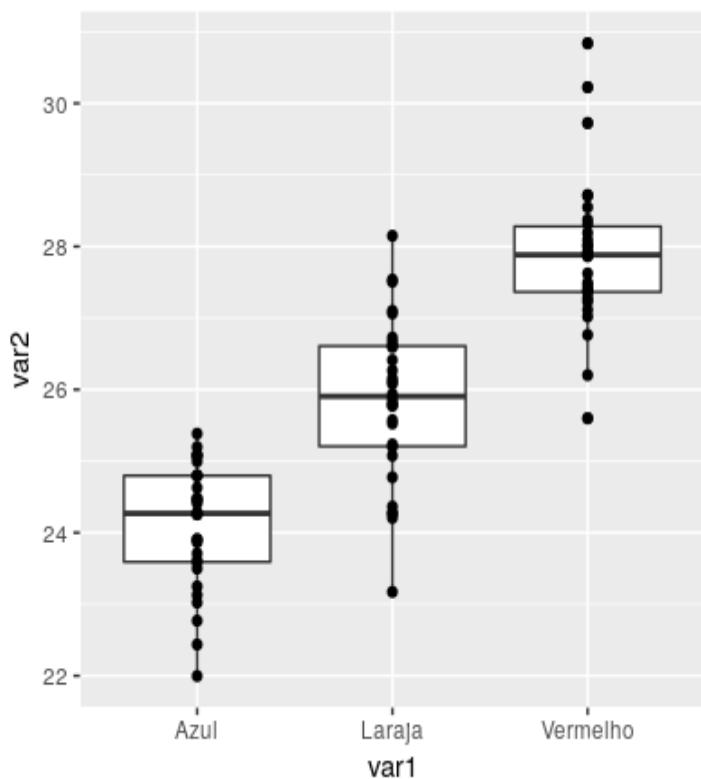


```
ggplot(dados, aes(x = var1, y = var2)) +  
  geom_boxplot() +  
  geom_point()
```

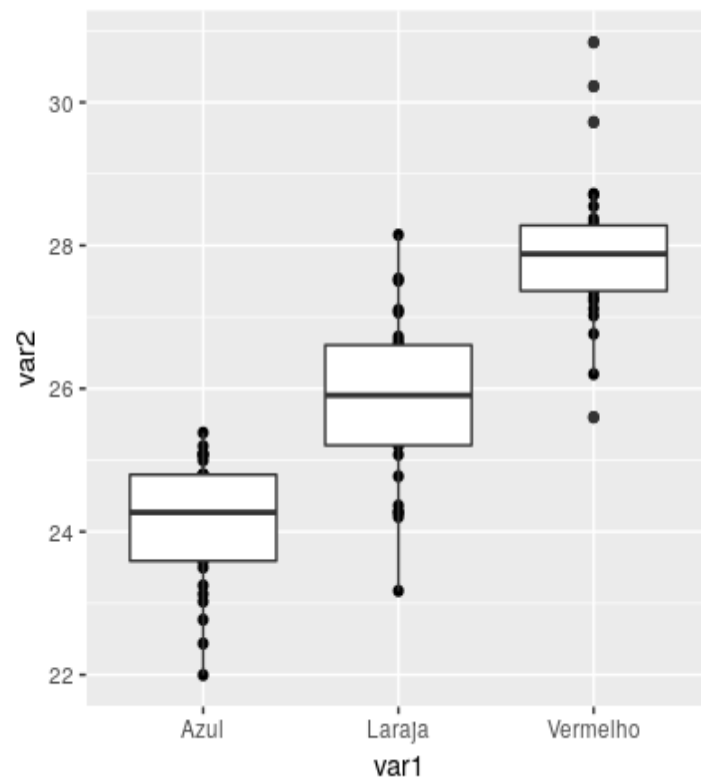


A ordem das camadas importa!

```
ggplot(dados, aes(x = var1, y = var2)) +  
  geom_boxplot() +  
  geom_point()
```



```
ggplot(dados, aes(x = var1, y = var2)) +  
  geom_point() +  
  geom_boxplot()
```



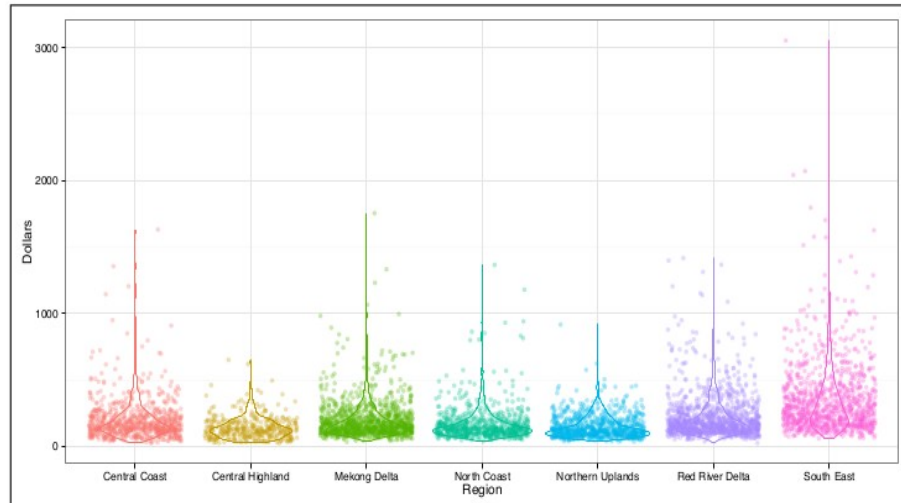
Aesthetic Mappings

- Aesthetic mappings define how graphical elements are visually perceived **when these elements vary**
 - Define x-dimension (predictor), y-dimension (response), size, color, fill, groupings, etc.
 - Each aesthetic can be mapped to a variable or set to a constant value
 - Aesthetics can be set globally (in **ggplot()** function) or locally (in a specific layer)
- Specified with **aes()**

Global Aesthetic Mappings

```
> ggplot(data = vlss, aes(x = Region, y = Dollars, color = Region) ) +  
  geom_violin() +  
  geom_jitter(alpha = 0.3)
```

Aesthetic mappings inside the **ggplot()** layer are applied to every layer.

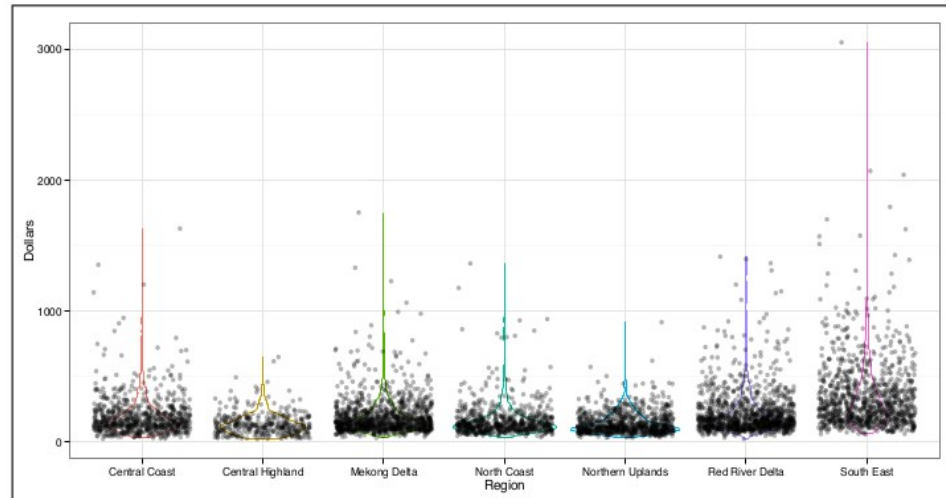




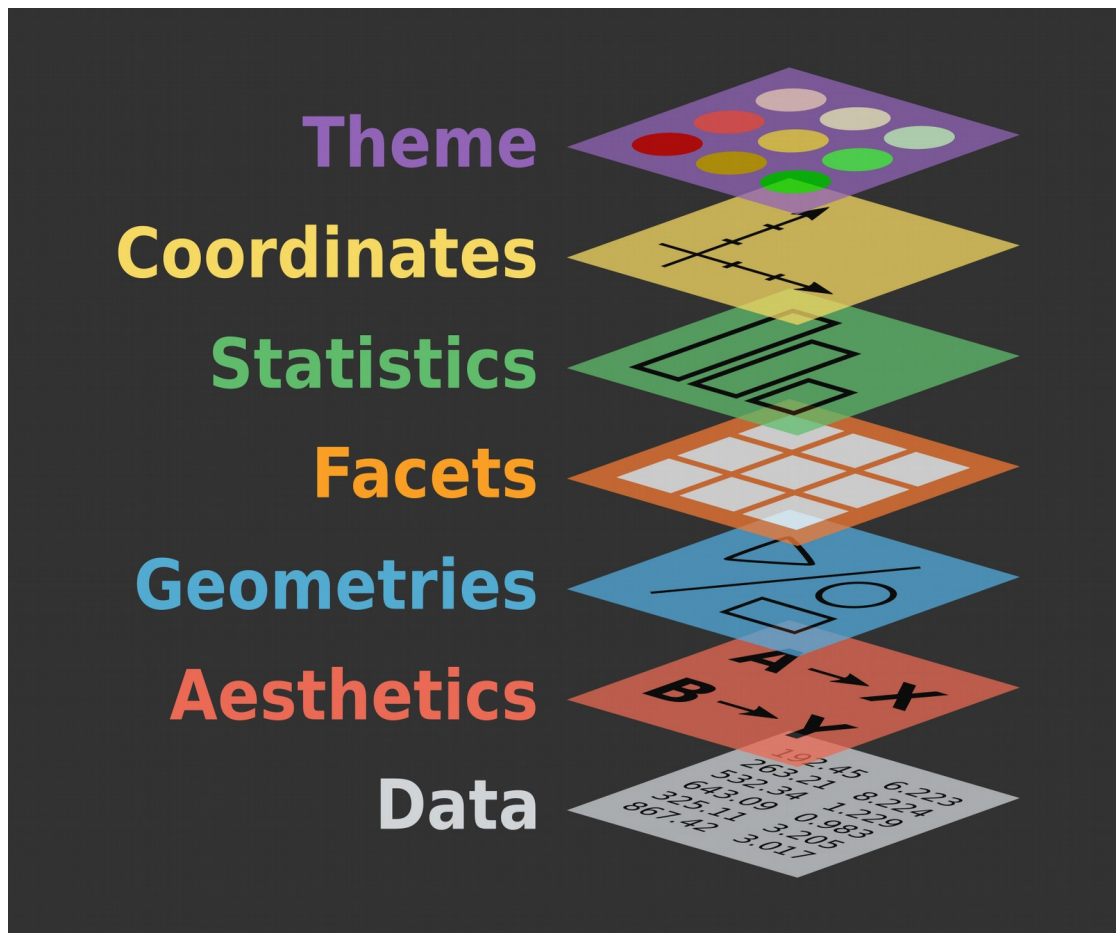
Local Aesthetic Mappings

```
> ggplot(data = vlss, aes(x = Region, y = Dollars) ) +  
  geom_violin( aes(color = Region) ) +  
  geom_jitter(alpha = 0.3)
```

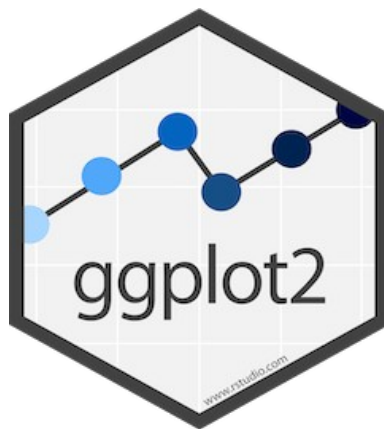
Aesthetic mappings inside a particular layer are only applied to that layer.



Resumindo os layers



Vamos aos exercícios!



Agradecimento



**Gustavo Brant
Paterno**
paternogbc

Unfollow

Block or report user

PhD candidate

👤 Departamento de Ecologia - U...

📍 Natal - Brasil

✉️ paternogbc@gmail.com

🔗 <https://www.researchgate.net/...>

https://github.com/paternogbc/guia_ggplot2

<https://github.com/paternogbc/curso-graficos-ufjf>

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