



**BIE5782**

Unidade 4:

**ANÁLISE EXPLORATÓRIA**

# ROTEIRO

1. Definição e importância de AED
2. Conferência e correção dos dados
3. AED univariadas
4. AED bivariadas e relações entre variáveis
5. AED multivariadas: definição

# AED ou EDA

- Controle de qualidade dos dados
- Sugerir hipóteses para os padrões observados
- Apoia a escolha dos procedimentos estatísticos de testes de hipótese
- Avaliar se os dados atendem às premissas dos procedimentos estatísticos escolhidos
- Indica novos estudos e hipóteses



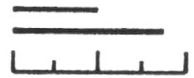
John W. Tukey  
1915-200

# Excelência Gráfica

## Princípios básicos da apresentação gráfica:

- ✓ Ressaltar os padrões de interesse;
- ✓ Manter a estrutura dos dados, de forma que o leitor possa reconstruir os dados a partir da figura;
- ✓ A figura deve ter uma razão **dado:tinta** alta;
- ✓ As figuras **não** devem distorcer, exagerar ou aparar os dados.

## BETTER



1. Position along a common scale



2. Position along identical scales



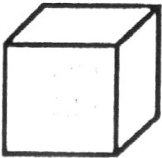
3. Length



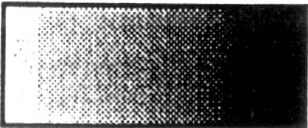
4. Angle/Slope



5. Area



6. Volume



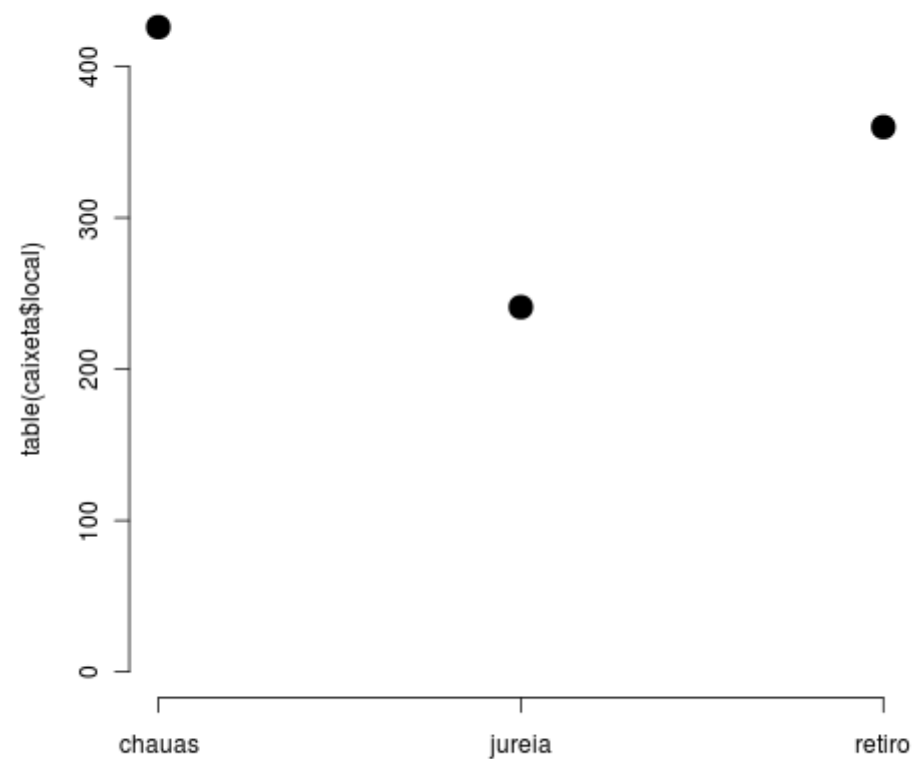
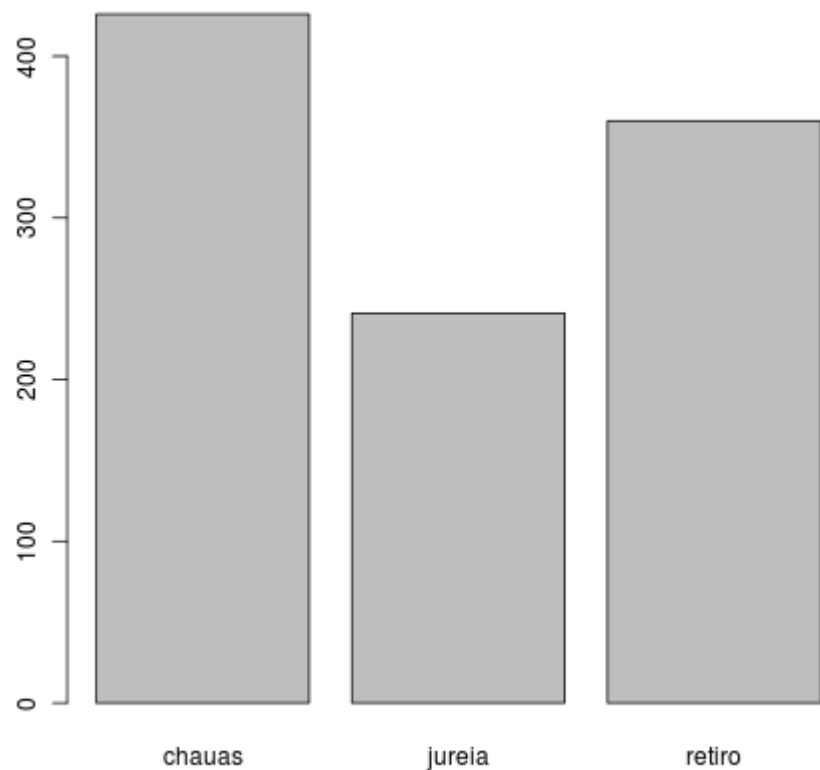
7. Shading: color, saturation, density

## WORSE



William Cleveland

# Razão dado/tinta



**summary()** , **str()** ,  
**head()** , **tail()**

Conferência dos Dados

DEMONSTRAÇÃO NO R



# is.na()

## Teste Lógico para Valores Perdidos

```
> a
[1] 1 2 3 4 5 NA 6 7 8 9 10 NA
> is.na(a)
[1] FALSE FALSE FALSE FALSE FALSE TRUE
[7] FALSE FALSE FALSE FALSE FALSE TRUE
> a[is.na(a)==F]
[1] 1 2 3 4 5 6 7 8 9 10
> a[is.na(a)==T] <- 0
> a
[1] 1 2 3 4 5 0 6 7 8 9 10 0
```



# Uma Variável

- Estatísticas descritivas
- Contagens de valores e tabelas
- Gráficos de distribuição
- Gráfico quantil-quantil



# mean(), median()

## Medidas de Tendência Central

```
> mean( c(0,1,2,3,4,5) )  
[1] 2.5
```

```
> median( c(0,1,2,3,4,5) )  
[1] 2.5
```

```
> mean( c(0,1,2,3,4,100) )  
[1] 18.33333
```

```
> median( c(0,1,2,3,4,100) )  
[1] 2.5
```

`mean(trim=)`, `mean()`,  
`median()`, `quantile()`

Média (normal e truncada) mediana,  
quantis: o pacote básico.

DEMONSTRAÇÃO NO R

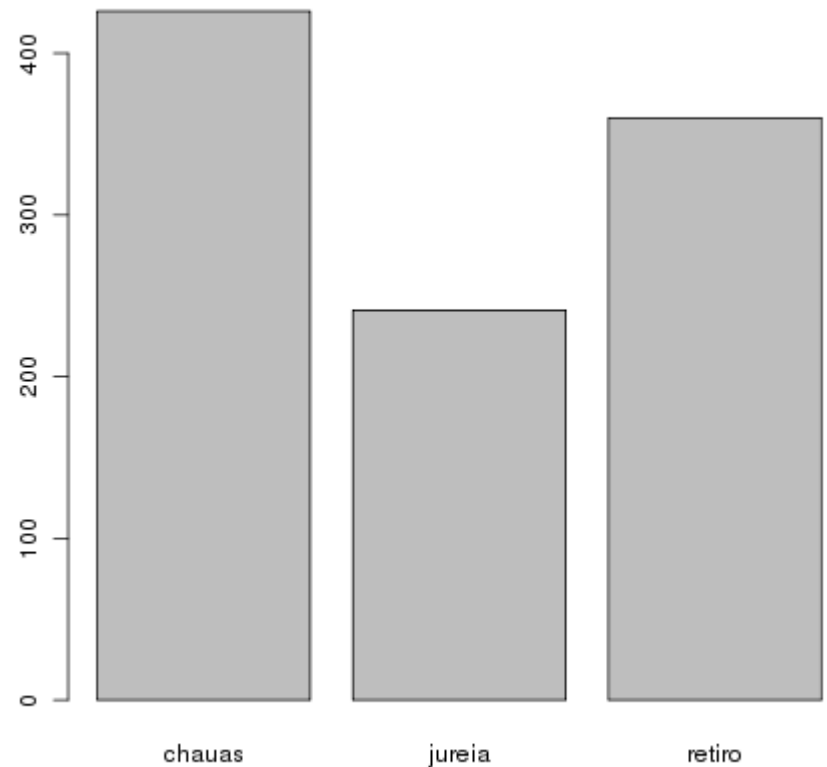


# `table()`, `barplot()`

## Contagens de Fatores

```
> table(caixeta$local)
```

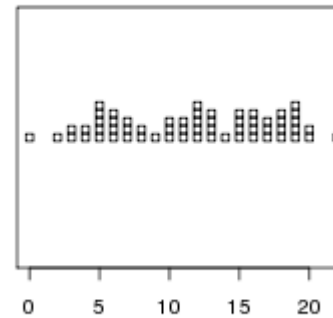
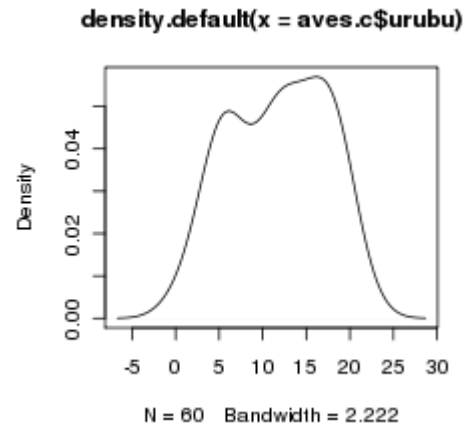
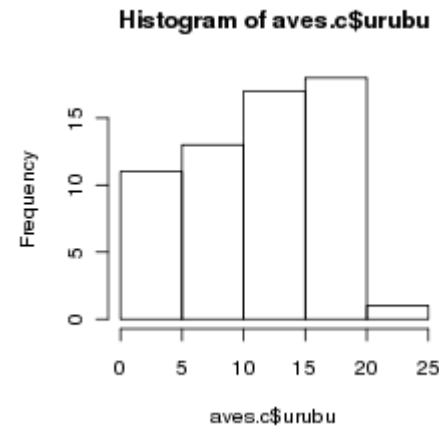
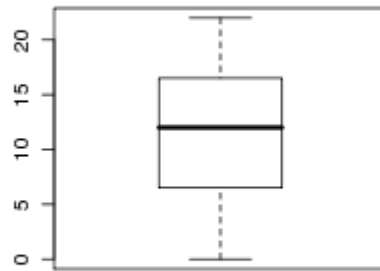
```
chauas  jureia  retiro  
  426    241    360
```



```
> barplot(table(caixeta$local))
```

# boxplot(), hist(), density(), stripchart()

## Gráficos Univariados Básicos



\* ← Valor extremo: + que 1,5 X a distância entre-quartis

← Ultimo ponto até 1,5 X a distância entre-quartis

Distância entre-quartis



← Quartil superior

← Mediana

← Quartil inferior

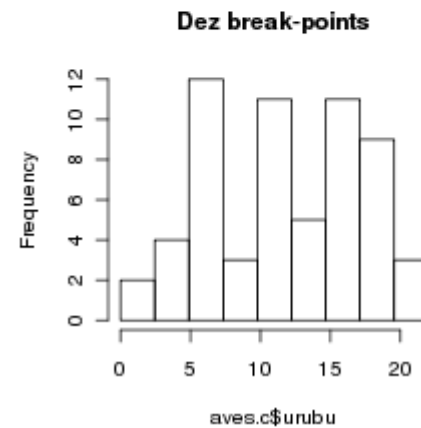
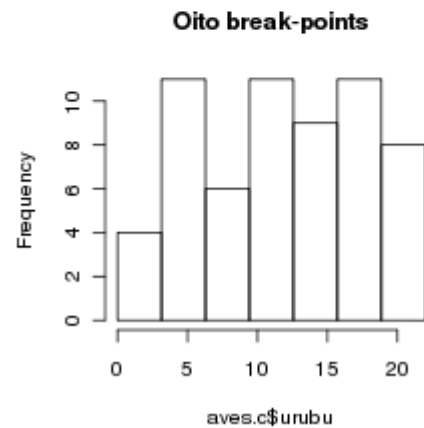
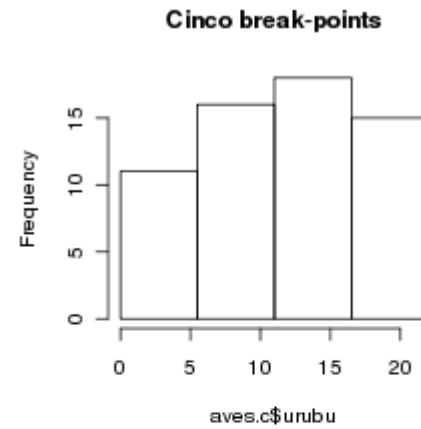
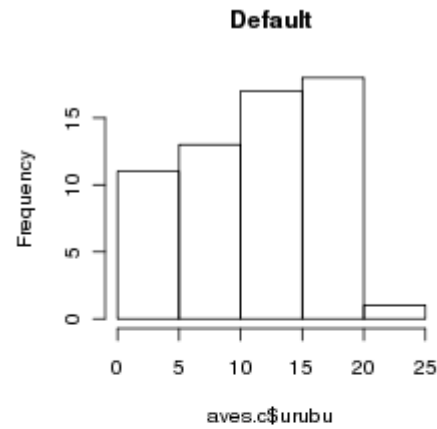
← Ultimo ponto entre 1,5 X a distância entre-quartis

**Box-and-whisker plot**

**OU**

**Diagrama de caixa e  
bigode de gato**

# O problema do n de classes



```
hist(aves.c$urubu)
```

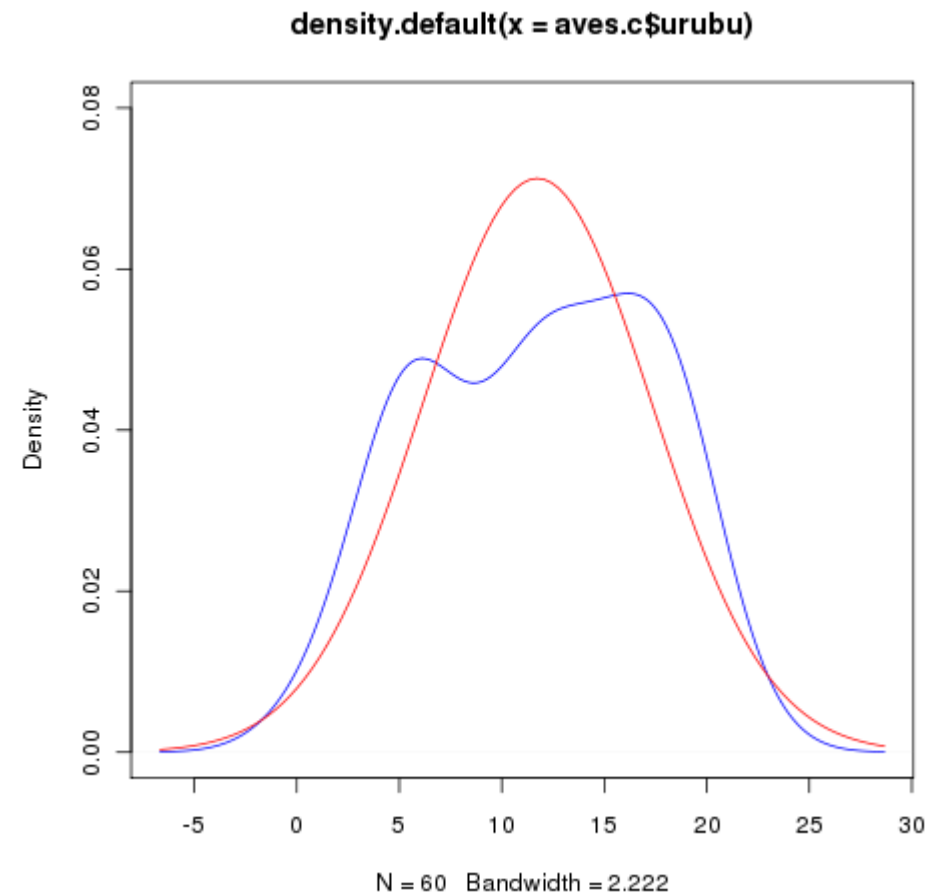
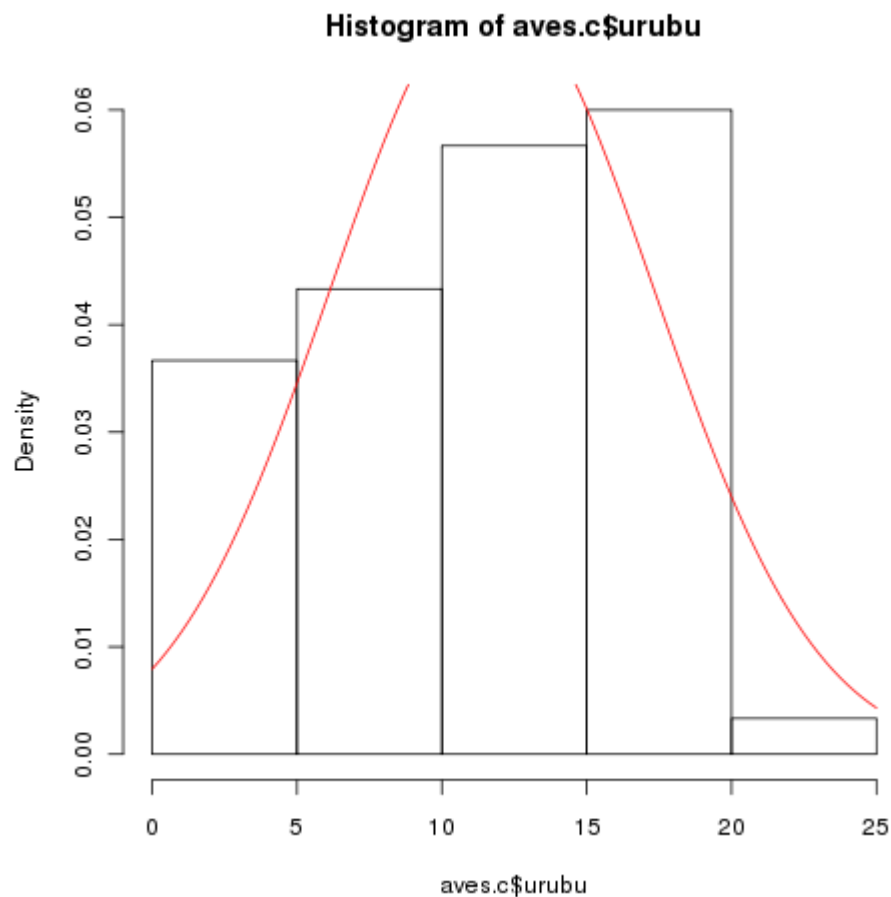
```
hist(aves.c$urubu,breaks=seq(0,max(aves.c$urubu),length=5))
```

```
hist(aves.c$urubu,breaks=seq(0,max(aves.c$urubu),length=8))
```

```
hist(aves.c$urubu,breaks=seq(0,max(aves.c$urubu),length=10))
```

**hist(prob=T)** , **density()** , **curve()**

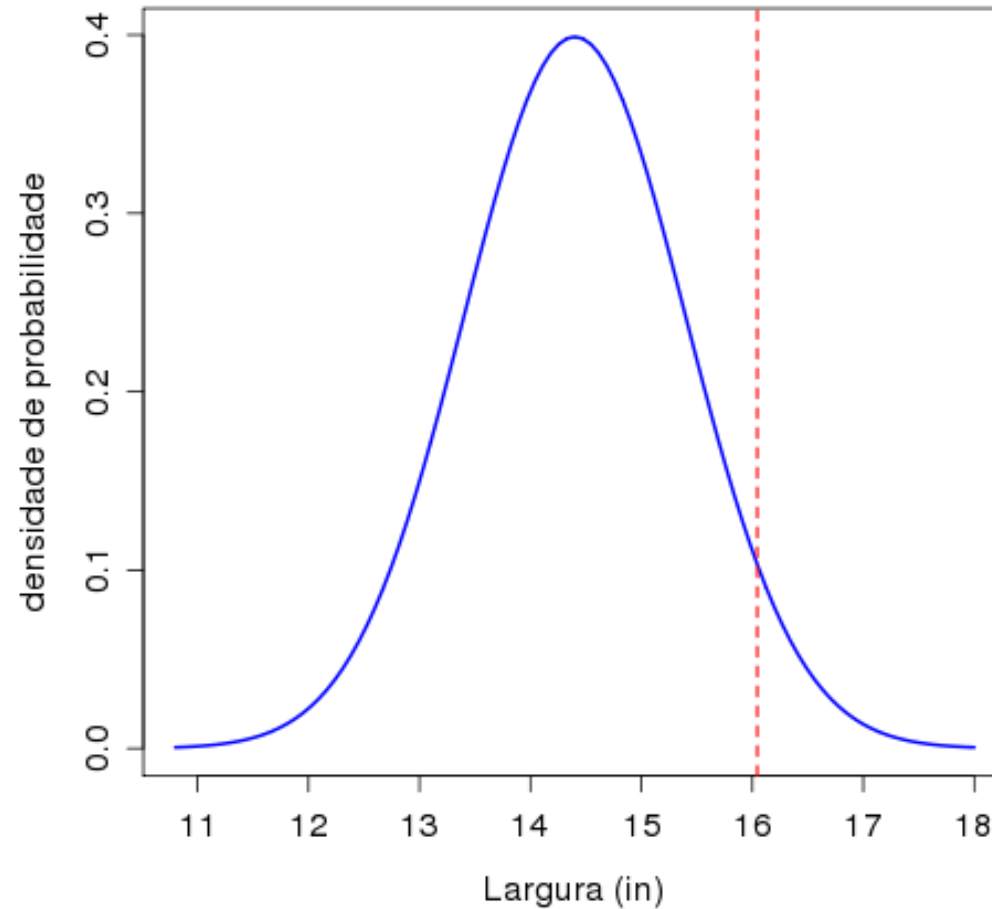
## Curvas Empíricas e Teóricas de Densidade





# Gráficos

## quantil (empírico) x quantil(teórico)



```
> qnorm(p=0.95, mean=14.4, sd=1)  
[1] 16.04485
```



Vai para o R!

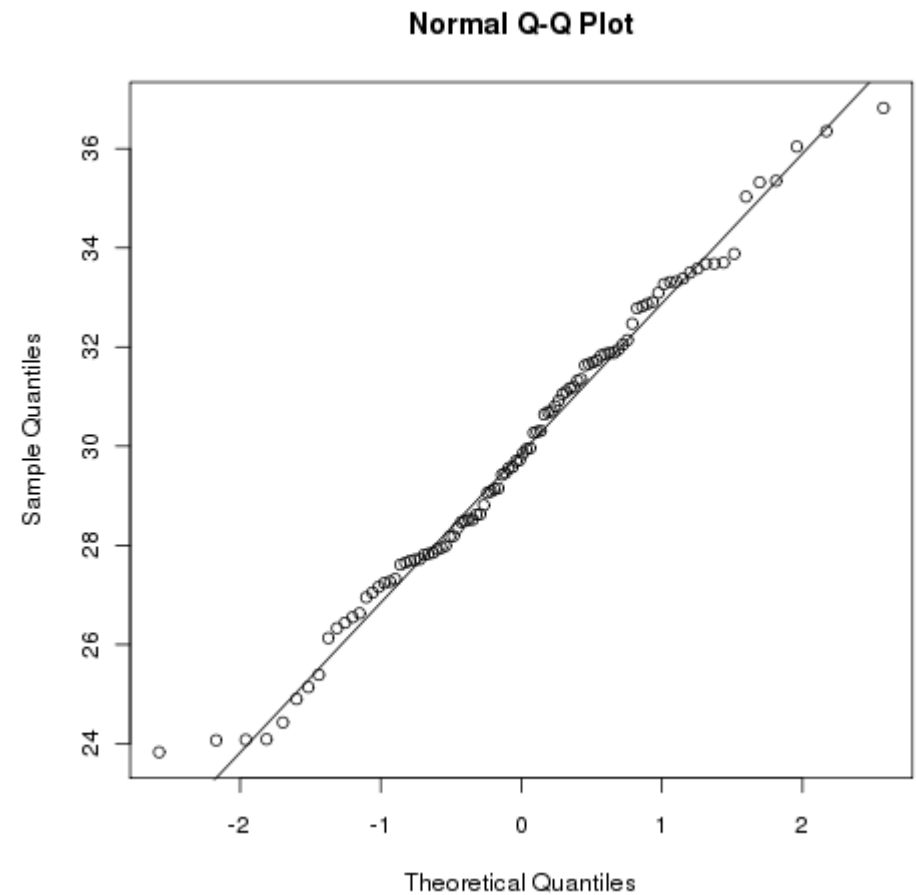
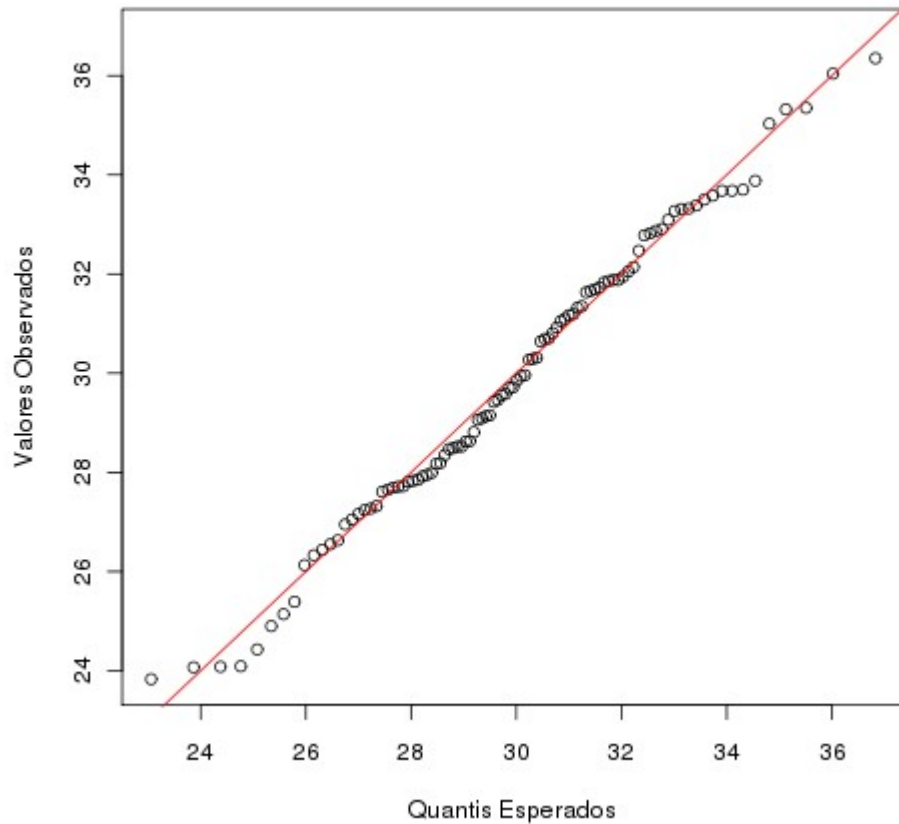
**qqnorm()** , **qqline()**

O melhor teste de normalidade

	<b>x</b>	<b>percentil</b>	<b>q.norm</b>
<b>1</b>	<b>23.83</b>	<b>0.01</b>	<b>23.05859</b>
<b>2</b>	<b>24.07</b>	<b>0.02</b>	<b>23.86540</b>
<b>3</b>	<b>24.08</b>	<b>0.03</b>	<b>24.37730</b>
<b>4</b>	<b>24.09</b>	<b>0.04</b>	<b>24.76238</b>
<b>5</b>	<b>24.43</b>	<b>0.05</b>	<b>25.07561</b>
<b>...</b>			
<b>95</b>	<b>35.03</b>	<b>0.95</b>	<b>34.81219</b>
<b>96</b>	<b>35.32</b>	<b>0.96</b>	<b>35.12542</b>
<b>97</b>	<b>35.35</b>	<b>0.97</b>	<b>35.51050</b>
<b>98</b>	<b>36.04</b>	<b>0.98</b>	<b>36.02240</b>
<b>99</b>	<b>36.35</b>	<b>0.99</b>	<b>36.82921</b>
<b>100</b>	<b>36.82</b>	<b>1.00</b>	<b>Inf</b>

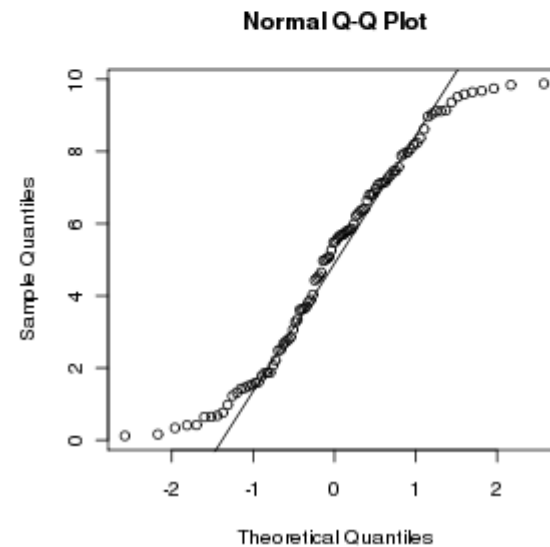
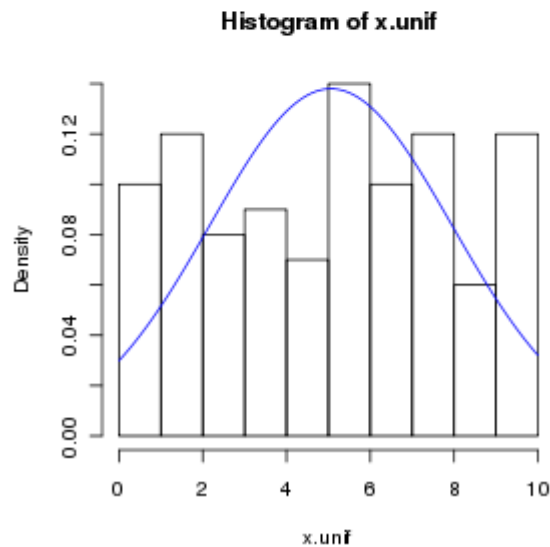
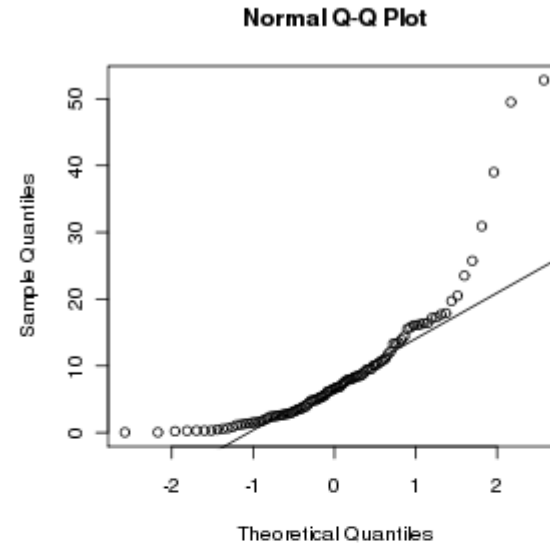
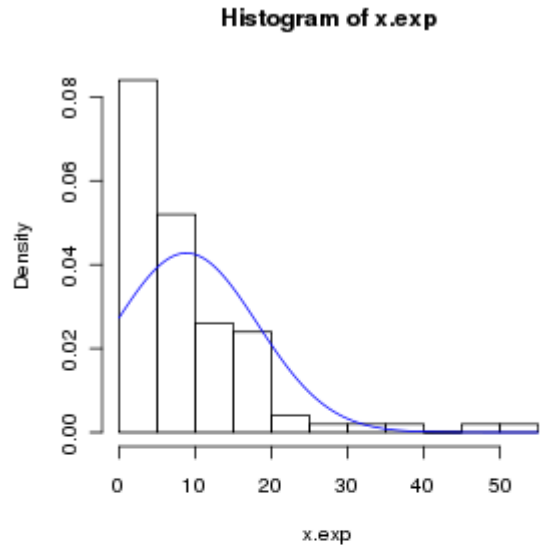
**qqnorm()** , **qqline()**

O melhor teste de normalidade



**qqnorm()** , **qqline()**

O melhor teste de normalidade



# DUAS VARIÁVEIS

- Fatores e contagens:
  - Tabelas de contingência
  - Tabelas de frequência
  - Estatísticas agregadas por fatores
- Gráficos
  - Dispersão
  - Linhas de tendência
  - Box-plot por classes
  - Gráficos condicionais



**table()**

# Tabelas de Contingência

> **table(caixeta\$especie, caixeta\$local)**

	<b>chauas</b>	<b>jureia</b>	<b>retiro</b>
<b>Alchornea triplinervia</b>	<b>0</b>	<b>3</b>	<b>12</b>
<b>Andira fraxinifolia</b>	<b>0</b>	<b>4</b>	<b>0</b>
<b>bombacaceae</b>	<b>0</b>	<b>1</b>	<b>0</b>
<b>Cabralea canjerana</b>	<b>0</b>	<b>4</b>	<b>0</b>
<b>Callophyllum brasiliensis</b>	<b>7</b>	<b>0</b>	<b>0</b>
<b>Callophyllum brasiliensis</b>	<b>0</b>	<b>4</b>	<b>0</b>
<b>Cecropia sp</b>	<b>0</b>	<b>0</b>	<b>1</b>
<b>Coussapoa macrocarpa</b>	<b>0</b>	<b>3</b>	<b>0</b>
<b>Coussapoa micropoda</b>	<b>2</b>	<b>0</b>	<b>7</b>
<b>Cryptocaria moschata</b>	<b>0</b>	<b>2</b>	<b>0</b>
<b>Cyathea sp</b>	<b>0</b>	<b>0</b>	<b>2</b>

# xtabs()

## Tabulação de Frequências

```
> head(Titanic.df)
  Class      Sex   Age Survived Freq
1   1st    Male Child       No     0
2   2nd    Male Child       No     0
3   3rd    Male Child       No    35
4  Crew    Male Child       No     0
5   1st Female Child       No     0
6   2nd Female Child       No     0
> xtabs(Freq~Sex+Survived, data=Titanic.df)
```

	Survived	
Sex	No	Yes
Male	1364	367
Female	126	344



# aggregate()

## "Tabelas Dinâmicas"

```
> names(caixeta)
[1] "local"      "parcela" "arvore"   "fuste"    "cap"
[5] "h"         "especie" "ab"
```

```
> caixeta.alt <- aggregate(caixeta$h,
+ by=list(local=caixeta$local,
+ especie=caixeta$especie), FUN=max)
```

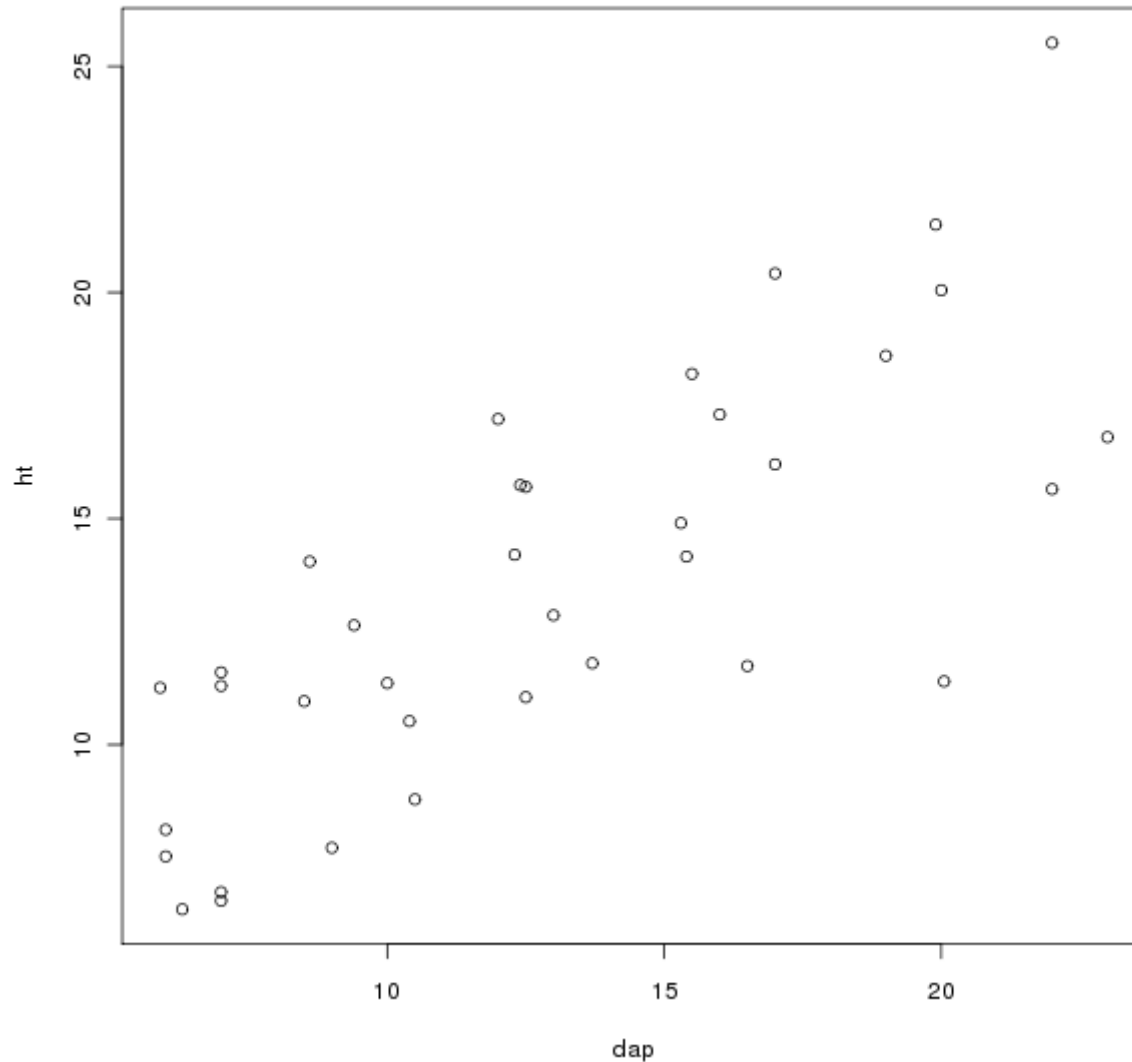
```
> head(caixeta.alt)
  local          especie      x
1 jureia  Alchornea triplinervia 140
2 retiro  Alchornea triplinervia 100
3 jureia      Andira fraxinifolia  90
4 jureia          bombacaceae 150
5 jureia      Cabralea canjerana 150
6 chauas Callophyllum brasiliensis 200
```



Vai para o R!

# plot(y~x)

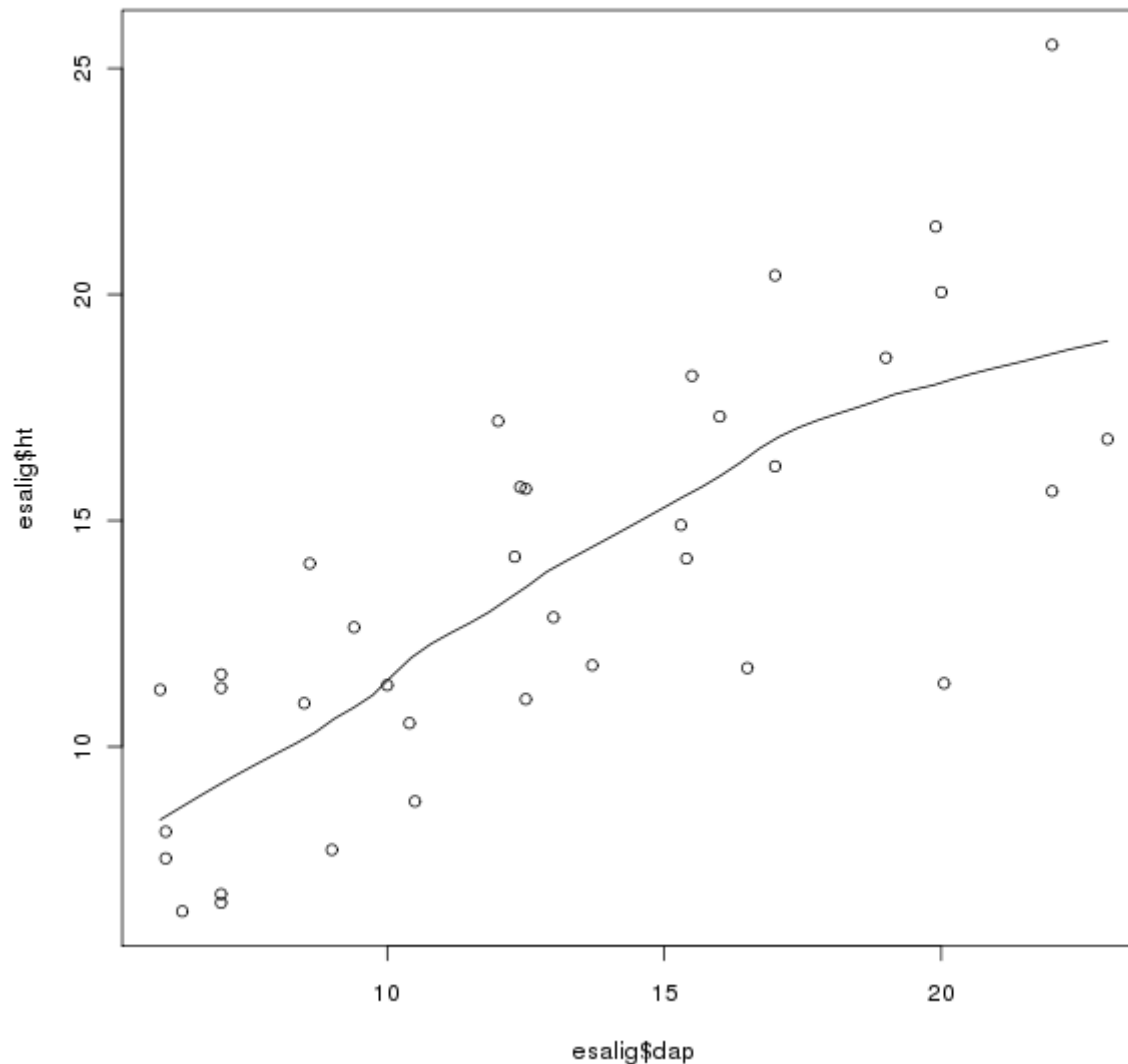
## Espalhogramas



```
> plot(ht~dap, data=esalig)
```

# `scatter.smooth(y~x)`

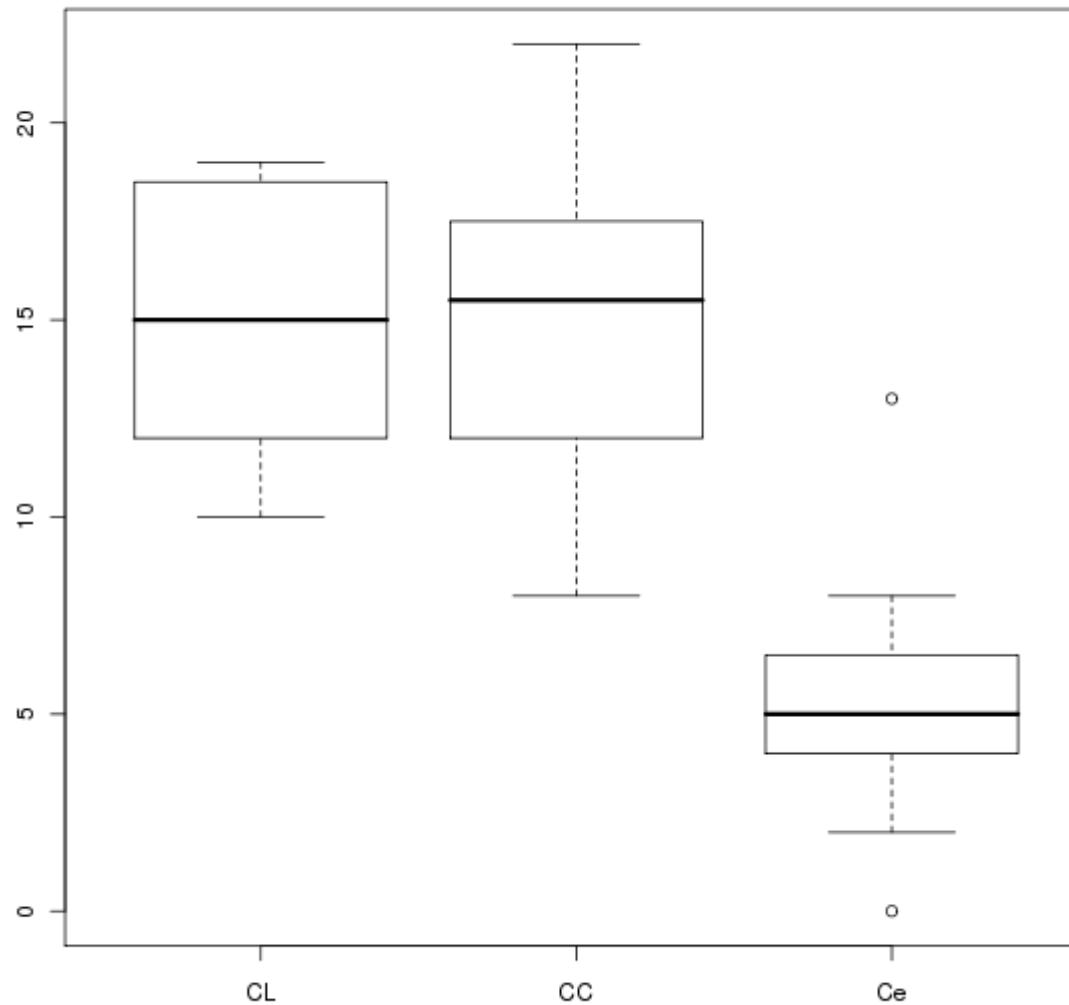
## Espalhogramas com Linha de Tendência



> `scatter.smooth(esalig$ht~esalig$dap, span=1/2)`

# boxplot(y~x)

## Boxplot por Classes



> `boxplot(urubu~fisionomia, data=aves.c)`

# MAIS DE DUAS VARIÁVEIS

- Fatores e contagens:
  - Tabelas multidimensionais
  - Matrizes de correlação e distância
  - Estatísticas agregadas por fatores
- Gráficos
  - Gráficos condicionados
  - Matrizes de gráficos
  - Ordenação e classificação



# Tabelas Multidimensionais

```
> xtabs(Freq~Class+Survived+Sex, data=Titanic.df)
```

```
, , Sex = Male
```

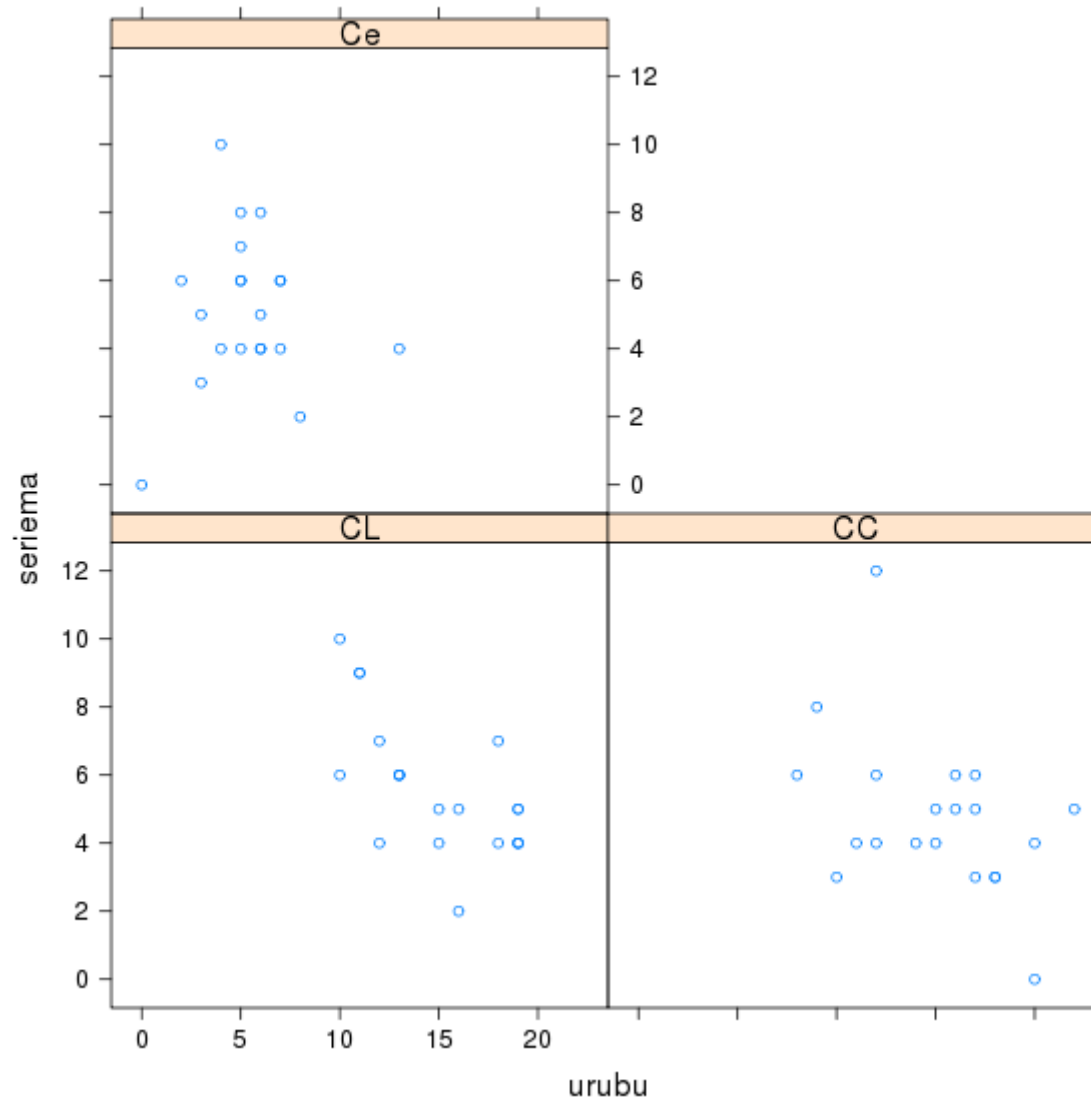
Class	Survived	
	No	Yes
1st	118	62
2nd	154	25
3rd	422	88
Crew	670	192

```
, , Sex = Female
```

Class	Survived	
	No	Yes
1st	4	141
2nd	13	93
3rd	106	90
Crew	3	20

# xyplot(y~x|z)

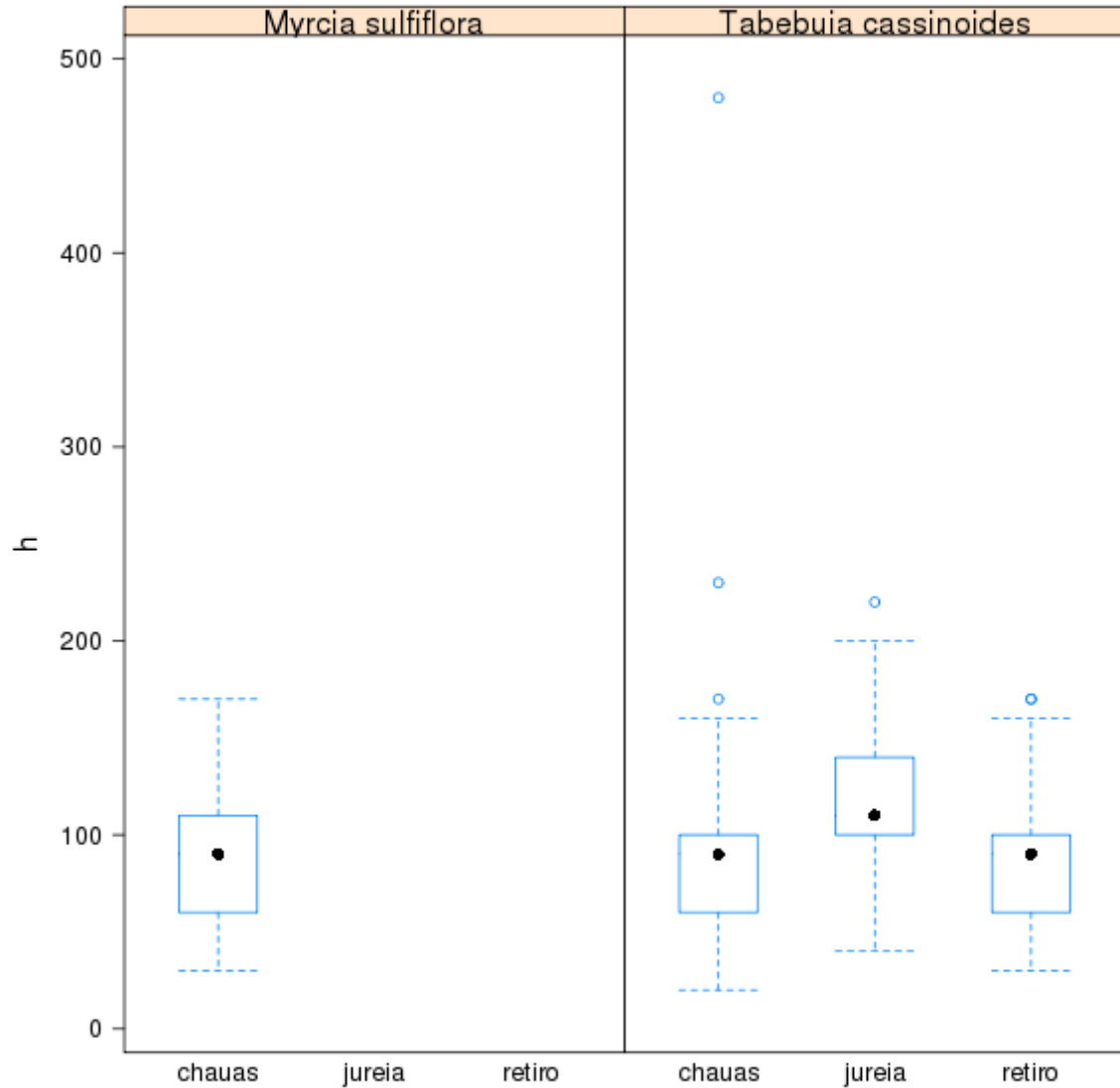
Pacote lattice: gráficos condicionados



> xyplot(seriema~urubu|fisionomia, data= aves.c)



# `bwplot(y~x|z)` Box-plot no lattice



> `bwplot(h~local|especie, data=caixeta.abund)`

# cor()

## Matrizes de correlação

```
> cor(esaligna[,4:7])
```

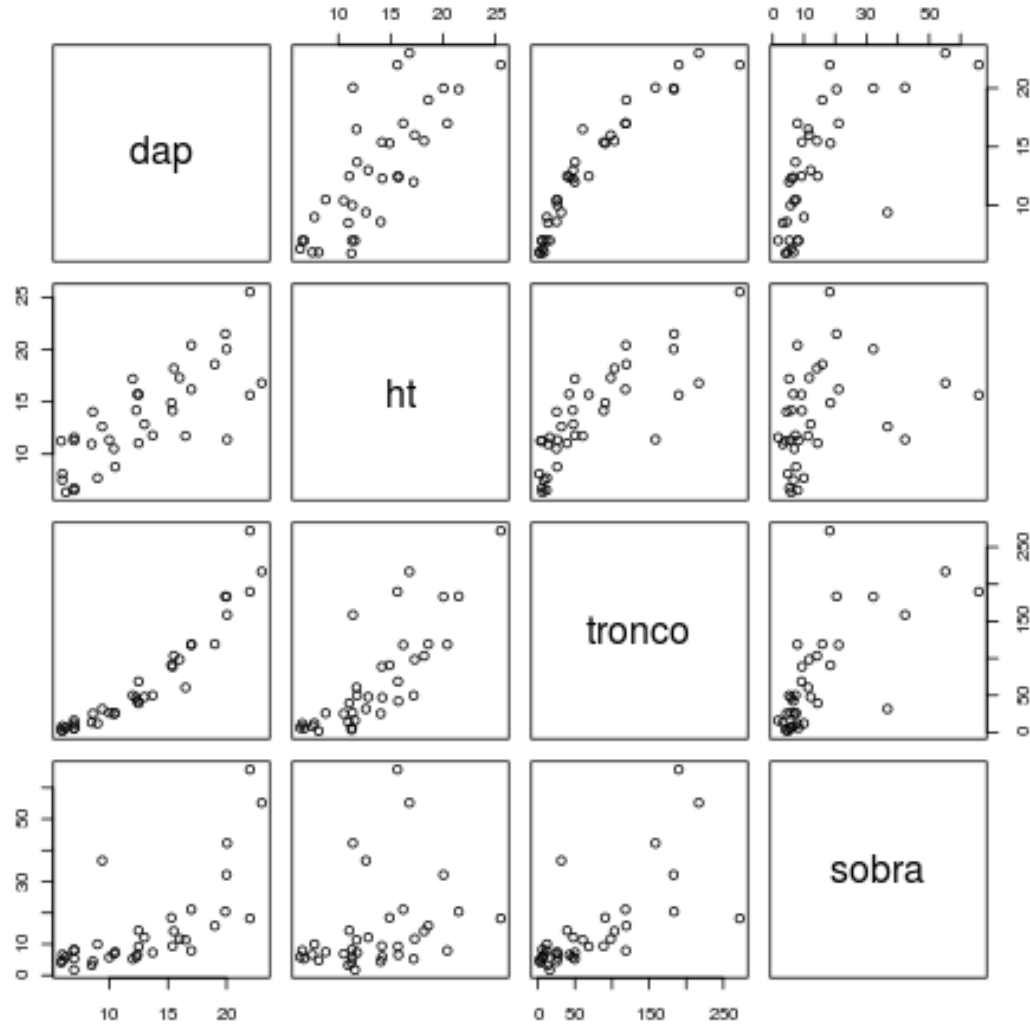
	dap	ht	tronco	sobra
dap	1.0000000	0.7745167	0.9407805	0.6863613
ht	0.7745167	1.0000000	0.8054810	0.3204422
tronco	0.9407805	0.8054810	1.0000000	0.6933458
Sobra	0.6863613	0.3204422	0.6933458	1.0000000

```
> cor(esaligna[,4:7], method="spearman")
```

	dap	ht	tronco	sobra
dap	1.0000000	0.7795958	0.9773287	0.7850061
ht	0.7795958	1.0000000	0.8512227	0.4857143
tronco	0.9773287	0.8512227	1.0000000	0.7534106
sobra	0.7850061	0.4857143	0.7534106	1.0000000

# pairs()

## Matriz de espalhogramas



> pairs(esaligna[,4:7])

# dist()

## Matrizes de distância

```
> aves.cf
```

```
      fisio urubu carcara seriema  
CL    CL    298      88     112  
CC    CC    299     212      96  
Ce    Ce    107     305     102
```

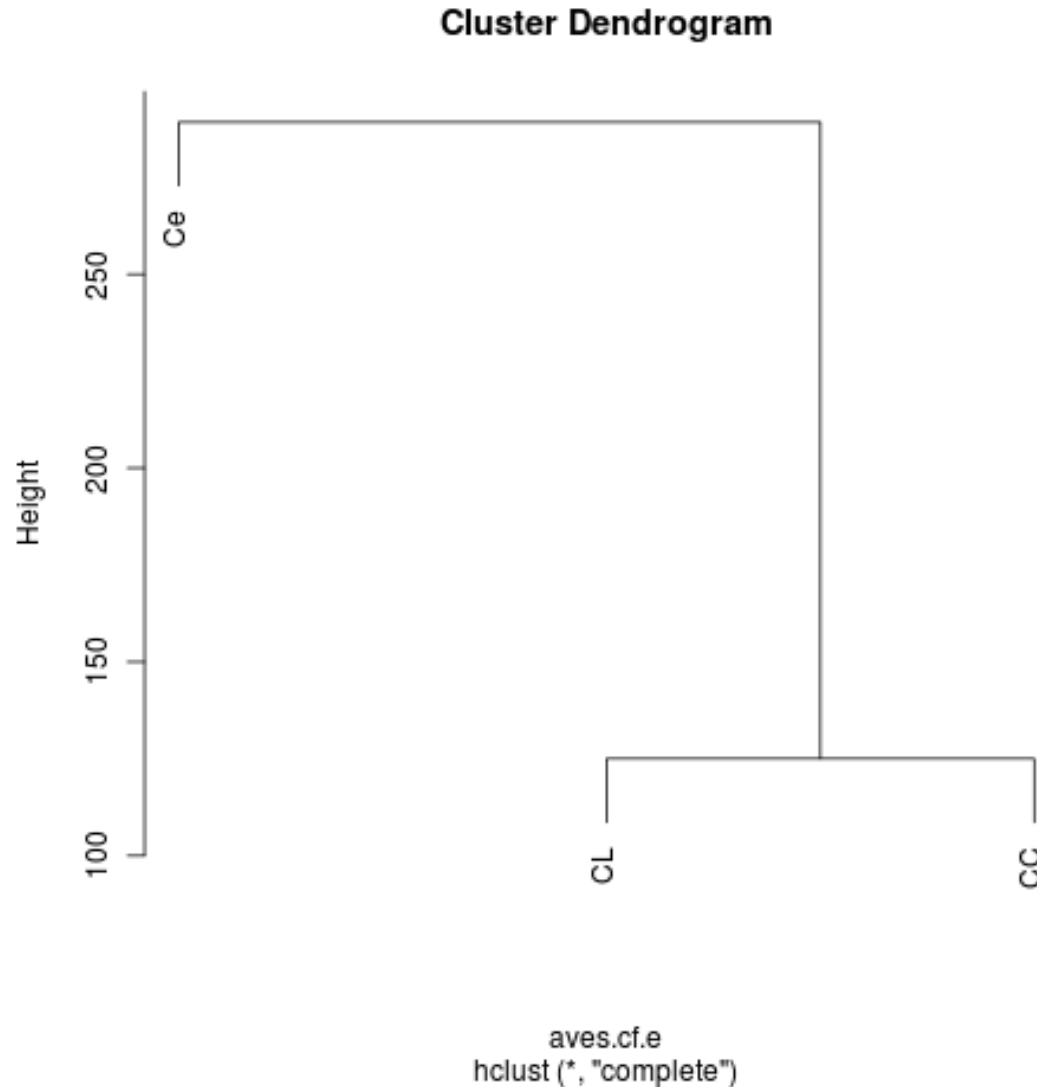
```
> aves.cf.e <- dist(aves.cf[,2:4])
```

```
> aves.cf.e
```

```
      CL      CC  
CC 125.0320  
Ce 289.2577 213.4221
```

# hclust()

## Análise de aglomerados

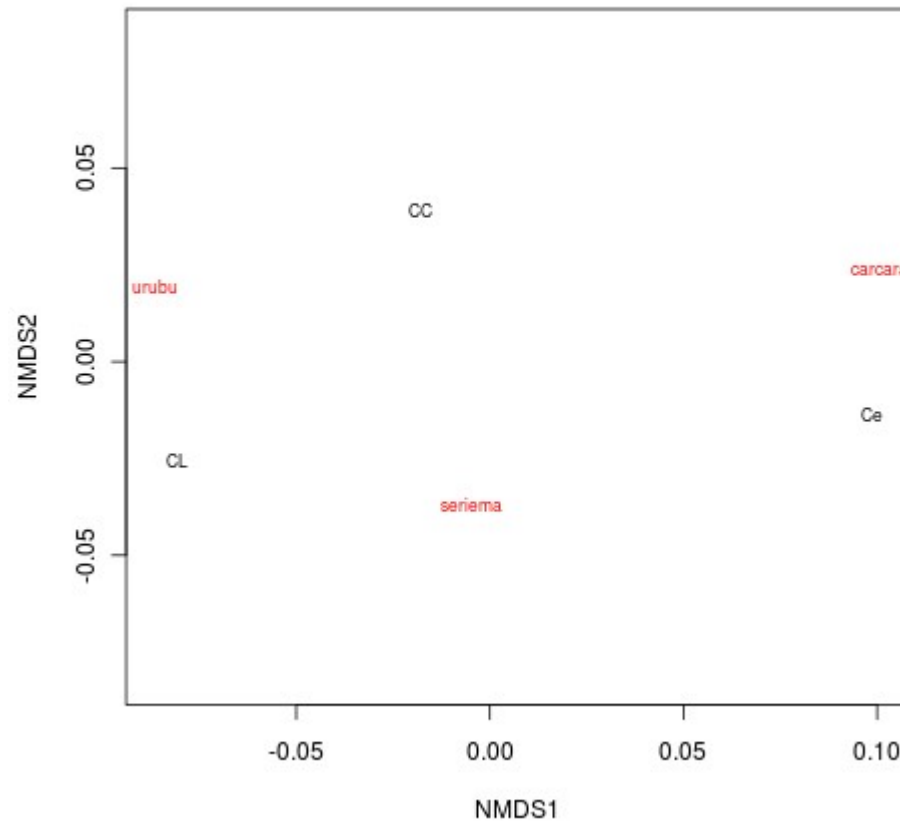


Esta é a função básica,  
ver pacotes vegan e  
ADE4 para análises  
multivariadas em  
Ecologia

> `plot(hclust(aves.cf.e))`

# metaMDS()

## Ordenação (um exemplo)



- > require(vegan)
- > plot(metaMDS(aves.cf[, 2:4]), type="t")

## Sugestão de leitura

**Cleveland, W. 1993. Visualizing data. Hobart Press.**

**Cleveland, W. 1994. Elements of graphing data. Hobart Press.**

**Ellison, A. M. 1993. Exploratory data analysis and graphic display. In: Scheiner, S. M. (ed.), *Design and analysis of ecological experiments*. Chapman & Hall, pp. 14-45.**

**McGill, R., Tukey, J. W. and Larsen, W. A. 1978. Variations of Box Plots. *Am. Statist.* 32: 12-16.**

# FIM DA UNIDADE 4

Para a tarde:

Plantão Tutoriais e exercícios Unidade 4

Até quarta:

Lista 4 de Exercícios: