Pré-processamento [no R] e Análise Exploratória

Fabrício Jailson Barth

fabricio.barth@gmail.com

Abril de 2015

Sumário

- Projeto R
- O que são **dados**?
- Caracterização dos dados.
- Raw data versus dado tratado.
- Representação de dados no R.
- **Análise Exploratória** de dados [no R].
- Exercícios.

Projeto R

- http://www.r-project.org/
- R Studio - http://www.rstudio.com/
- É free
- É a linguagem de programação mais popular para análise de dados
- Script é melhor que clicar e arastar:
  - ⋆ É mais fácil de comunicar → RMarkdown.
  - ⋆ Reproduzível.
  - ⋆ É necessário pensar mais sobre o problema.
- Existe uma quantia grande de pacotes disponíveis

Definição de dados

"Data are values of qualitative or quantitative variables, belonging to a set of items."

http://en.wikipedia.org/wiki/Data

"Data are values of qualitative or quantitative variables, belonging to a **set of items**."

Set of items: conjunto de itens (objetos) de interesse.

"Data are values of qualitative or quantitative **variables**, belonging to a set of items."

variables: uma medida ou uma característica de um item.

"Data are values of **qualitative** or **quantitative** variables, belonging to a set of items."

qualitative: cidade de origem, sexo, fez ou não tratamento.

quantitative: peso, altura, pressão do sangue.

Caracterização dos dados

- A escala define as operações que podem ser realizadas sobre os valores do atributo.
- Em relação à escala, os atributos podem ser classificados como **nominais**, **ordinais**, **discreto** e **contínuo**.
- Os dois primeiros são do tipo qualitativo e os dois últimos são quantitativos.

- Na escala **nominal**, os valores são apenas **nomes diferentes**, carregando a menor quantidade de informação possível. Não existe uma relação de ordem entre seus valores.
- Os valores em uma escala **ordinal** refletem também uma ordem das categorias representadas. Dessa forma, além dos operadores de igualdade e desigualdade, operadores como $<$, $>$, $\geq$, $\leq$ podem ser utilizados.

- Uma variável quantitativa que pode assumir, teoricamente, qualquer valor entre dois limites recebe o nome de **variável contínua**.
- Uma variável que só pode assumir valores pertencentes a um conjunto enumerável recebe o nome de **variável discreta**.

Raw data versus dados processados

Raw data

- Fonte original dos dados
- Geralmente difícil para fazer algum tipo de análise

http://en.wikipedia.org/wiki/Raw_Data

Dados processados

- Dados que estão prontos para serem analisados
- O processamento pode incluir *merging*, *subsetting*, *transforming*, etc...
- Todas as etapas devem ser registradas

http://en.wikipedia.org/wiki/Compute_data_processing

Exemplo de dados brutos

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html

Exemplo de dados brutos

consideração o projeto da aprendizagem que pensa como didaticamente os cursos devem ser projetados com o uso da tecnologia adequada. Isso inclui levar em conta os aspectos sociais e culturais envolvidos. Deixo abaixo algumas indicações de leitura que tratam isso. Assim, acho que dizer que tecnologia deve ser usada de forma responsável, não é discutir MOOCs. Outro ponto importante é destacar que os MOOCs aparecem no contexto da educação aberta e Ciencia aberta e inclui REAs, que costumavam ser chamados de objetos de aprendizem e agora discutem-se as licenças, as perspectivas de reutilização e de localização; os periódicos abertos que reagem aos altos valores de assinaturas dos periódicos tradicionais, as novas formas de publicação incluindo blogs; a educação hibrida; os ambientes pessoais de aprendizagem, etc. No geral

Exemplo de dado processado

Table 1: Exemplo de tabela com as transações dos usuários

usuário	$categoria_1$	$categoria_2$	$categoria_3$	$\cdots$	$categoria_m$
$user_1$	0	2	0	$\cdots$	1
$user_2$	1	1	0	$\cdots$	0
$user_3$	2	0	1	$\cdots$	0
$user_4$	0	1	0	$\cdots$	0
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$user_n$	1	1	0	$\cdots$	1

Tidy data

- Cada variável (atributo) forma uma coluna.
- Cada observação (exemplo) forma uma linha.
- Cada tabela ou arquivo armazena dados sobre uma observação (i.e., pessoas / hospitais)
- http://vita.had.co.nz/papers/tidy-data.pdf

Representação de dados no R

Tipos de dados importantes no R

- Classes: Character, Numeric, Integer, Logical
- Objetos: Vector, Matrices, Data frames, List, Factors, Missing Values
- Operadores: Subsetting, Logical Subsetting

Character

```
nome = "maria"
class(nome)
```

```
## [1] "character"
```

```
nome
```

```
## [1] "maria"
```

Numeric

```
peso = 76.2
class(peso)
```

```
## [1] "numeric"
```

```
peso
```

```
## [1] 76.2
```

Integer

```
qtdFilhos = 1L
class(qtdFilhos)
```

```
## [1] "integer"
```

```
qtdFilhos
```

```
## [1] 1
```

Logical

```
temCarro = TRUE
class(temCarro)
```

```
## [1] "logical"
```

```
temCarro
```

```
## [1] TRUE
```

Vectors

Um conjunto de valores da mesma classe.

```
pesos = c(76.2, 80.3, 90, 117.4)
pesos
```

```
## [1] 76.2 80.3 90 117.4
```

```
nomes = c("maria", "carlos", "pedro")
nomes
```

```
## [1] "maria" "carlos" "pedro"
```

Lists

Um conjunto de valores que pode ser heterogêneo.

```
pesosV = c(76.2, 80.3, 90, 117.4)
nomesV = c("maria", "carlos", "pedro", "antônio")

myList <- list(pesos = pesosV, nomes = nomesV)
myList

## $pesos
## [1]  76.2  80.3  90.0 117.4
##
## $nomes
## [1] "maria"   "carlos"  "pedro"   "antônio"
```

Matrizes

Vetores com múltiplas dimensões.

```
myMatrix = matrix(c(1, 2, 3, 4), byrow = T, nrow = 2)
myMatrix
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

Data frames

Múltiplos vetores de classes diferentes, mas com o mesmo tamanho.

```
vector1 = c(188.2, 181.3, 193.4)
vector2 = c("jeff", "roger", "andrew", "brian")
myDataFrame = data.frame(heights = vector1,
                         firstNames = vector2)

## Error: arguments imply differing number of rows: 3, 4

myDataFrame

## Error: object 'myDataFrame' not found
```

Data frames

```
> vector1 = c(188.2, 181.3, 193.4)
> vector2 = c("jeff", "roger", "andrew")
> myDataFrame = data.frame(heights = vector1,
                           firstNames = vector2)
> myDataFrame

  heights firstNames
1   188.2       jeff
2   181.3      roger
3   193.4     andrew
```

Factors

Variáveis qualitativas que podem ser incluídas no modelo.

```
smoker = c("yes", "no", "yes", "yes")
smokerFactor = as.factor(smoker)
smokerFactor
```

```
## [1] yes no yes yes
## Levels: no yes
```

Missing values

No R os valores faltantes são codificados como NA

```
vector1 <- c(188.2, 181.3, 193.4, NA)
vector1
```

```
## [1] 188.2 181.3 193.4 NA
```

```
is.na(vector1)
```

```
## [1] FALSE FALSE FALSE TRUE
```

Subsetting

```
vector1 = c(188.2, 181.3, 193.4, 192.3)
vector2 = c("jeff", "roger", "andrew", "brian")
myDataFrame = data.frame(heights = vector1,
                         firstNames = vector2)
vector1[1]
```

```
## [1] 188.2
```

```
vector1[c(1, 2, 4)]
```

```
## [1] 188.2 181.3 192.3
```

Subsetting

```
myDataFrame[1, 1:2]
```

```
## heights firstNames
## 1 188.2 jeff
```

```
myDataFrame$firstNames
```

```
## [1] jeff roger andrew brian
## Levels: andrew brian jeff roger
```

Logical subsetting

```
myDataFrame[myDataFrame$firstNames == "jeff", ]
```

```
## heights firstNames
## 1 188.2 jeff
```

```
myDataFrame[heights < 190, ]
```

```
## heights firstNames
## 1 188.2 jeff
## 2 181.3 roger
## 4 192.3 brian
```

Análise Exploratória de Dados

Dados utiizados

Os exemplos a seguir fazem uso de dois datasets distintos:

- **Survey** sobre dados de alunos de uma turma de estatística.

```
library(UsingR)
data(survey)
names(survey)
sapply(survey, class)
```

- Dados de flores do gênero **iris**.

```
data(iris)
head(iris)
help(iris)
```

Caracterização dos dados

No R, é possível testar se um atributo é **qualitativo** (factor) ou **quantitativo** (numeric).

```
is.numeric(survey$Pulse)
is.factor(survey$Sex)
is.numeric(survey$Smoke)
is.factor(survey$Height)

is.numeric(iris$Sepal.Length)
is.factor(iris$Species)
```

Caracterização dos dados

Os atribudos dos datasets IRIS e SURVEY podem ser classificados como indicado nas tabelas abaixo:

```
class(survey$Pulse) = integer (quantitativo discreto)
class(survey$Sex) = factor (qualitativo)
class(survey$Smoke) = factor (ordinal - qualitativo)
class(survey$Height) = numeric (quantitativo contínuo)

class(iris$Sepal.Length) = numeric (quantitativo contínuo)
class(iris$Species) = factor (qualitativo)
```

Exploração de dados

Uma das formas mais simples de explorar um conjunto de dados é a extração de medidas de uma área da estatística denominada **estatística descritiva**. A estatística descritiva resume de forma quantitativa as principais características de um conjunto de dados.

Tais características podem ser:

- Frequência;
- Localização ou tendência central (por exemplo, a média);
- Dispersão ou espalhamento (por exemplo, o desvio padrão);
- Distribuição ou formato.

No R é trivial identificar a média e mediana de um dado conjunto de valores para um atributo qualquer, como apresentado abaixo:

```
mean(survey$Pulse)
median(survey$Pulse)
```

Ou sumarizar todos estes valores através de um único comando:

```
summary(survey$Pulse)
```

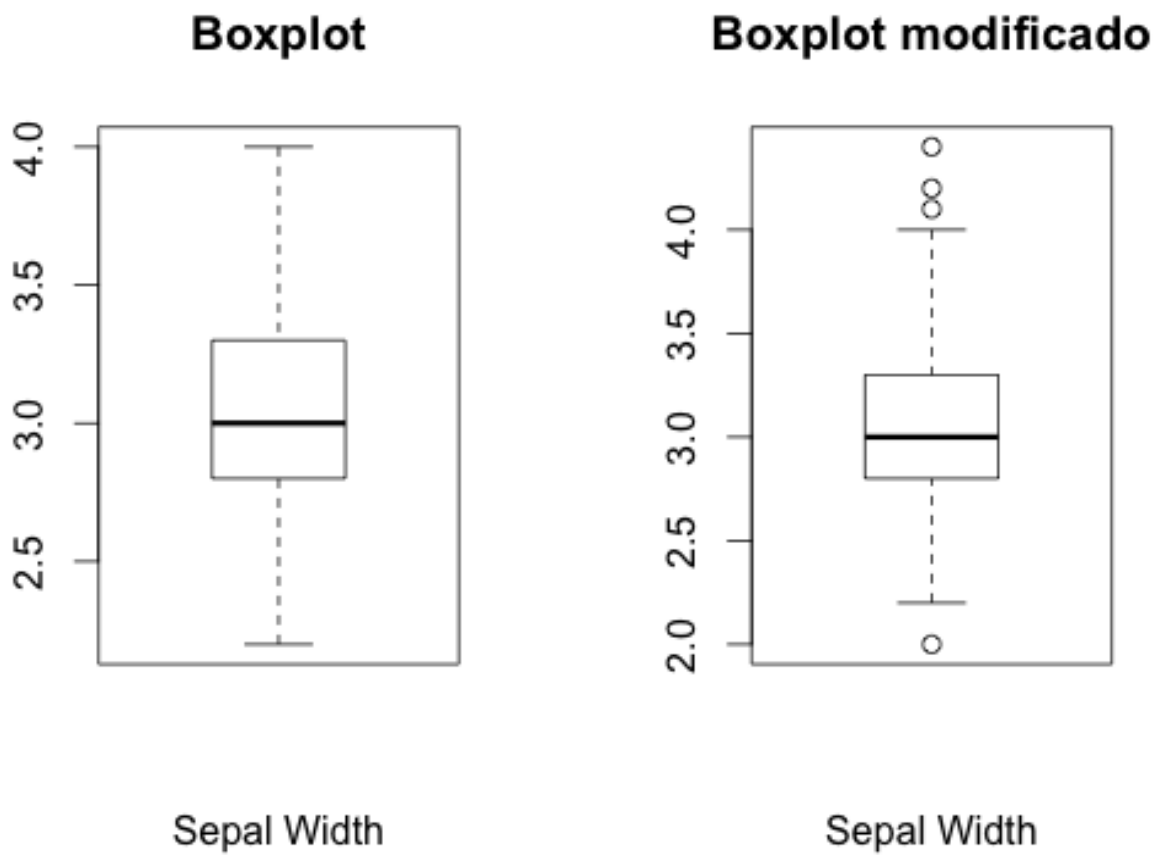
Além das informações textuais obtidas por

```
summary(iris$Sepal.Width)
```

É possível obter um resumo visual da centralidade dos dados através do gráfico *boxplot*. No R é simples gerar este tipo de gráfico.

```
par(mfrow=c(1,2))
boxplot(iris$Sepal.Width,
        outline= FALSE, main="Boxplot",
        xlab="Sepal Width")
boxplot(iris$Sepal.Width,
        main="Boxplot modificado",
        xlab="Sepal Width")
```

Boxplot


Boxplot
4.0
3.5
3.0
2.5
Sepal Width
Boxplot modificado
4.0
3.5
3.0
2.5
2.0
Sepal Width

Boxplot original

Do lado esquerdo da figura é apresentado o gráfico *boxplot* original. Nele, a linha horizontal mais baixa e a linha horizontal mais alta indicam, respectivamente, os valores mínimo e máximo presentes nos dados. Os lados inferior e superior do retângulo representam o 1o quartil e o 3o quartil, respectivamente. A linha no interior do retângulo é o 2o quartil, ou mediana.

Boxplot modificado

O segundo gráfico ilustra uma variação do gráfico *boxplot*, conhecida como *boxplot* modificado. Neste gráfico, os valores acima do limite superior e abaixo do limite inferior são considerados *outliers*. Neste gráfico, 4 valores *outliers* são representados por círculos, 3 maiores que o 3o quartil + 1,5 x (3o quartil - 1o quartil) e 1 menor que 1o quartil - 1,5 x (3o quartil - 1o quartil).

Espalhamento de valores

As medidas mais utilizadas para avaliar o **espalhamento** de valores é a **variância** (var) e o **desvio padrão** (sd). Sendo que o desvio padrão é dado pela raiz quadrada da variância.

Desvio padrão:

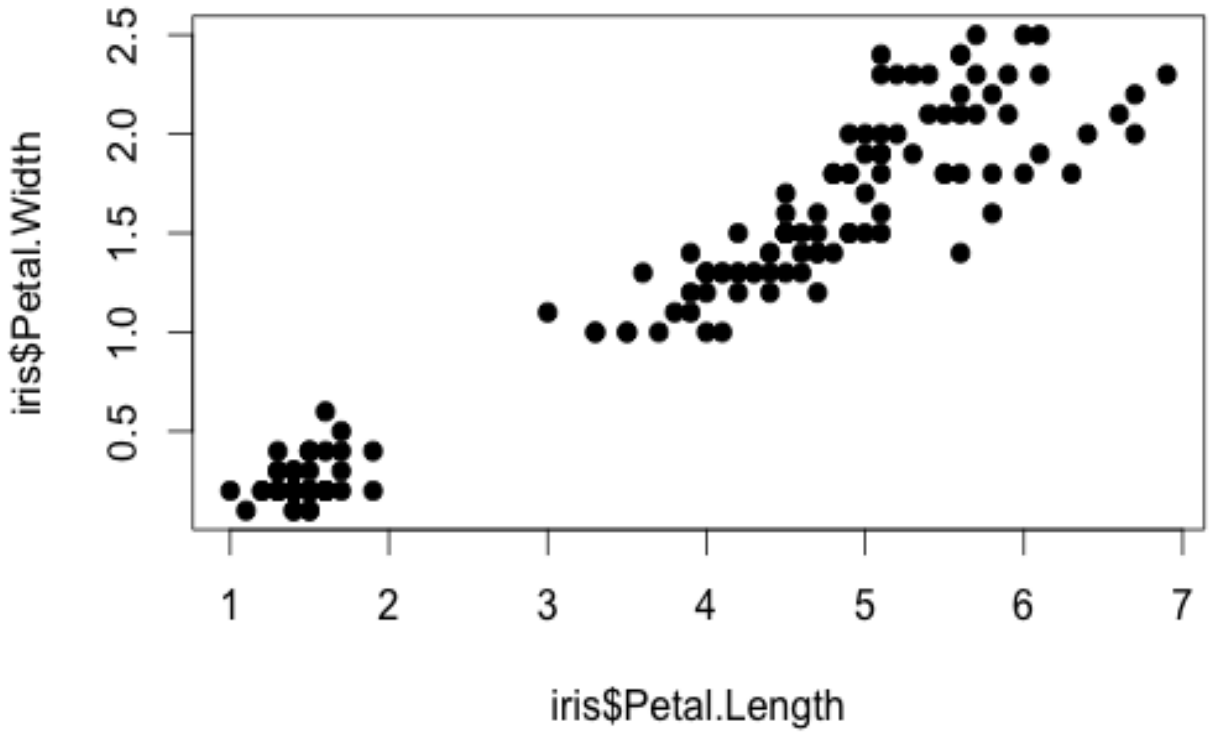
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{1}$$

Variância:

$$s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2}$$

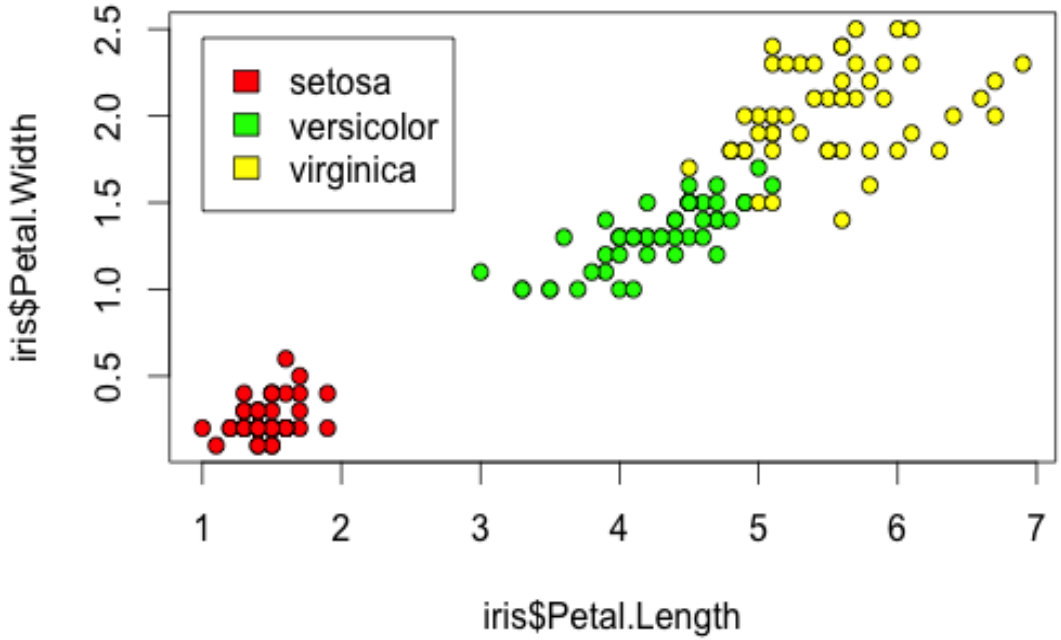
Plot

```
plot(iris$Petal.Length, iris$Petal.Width, pch=19)
```



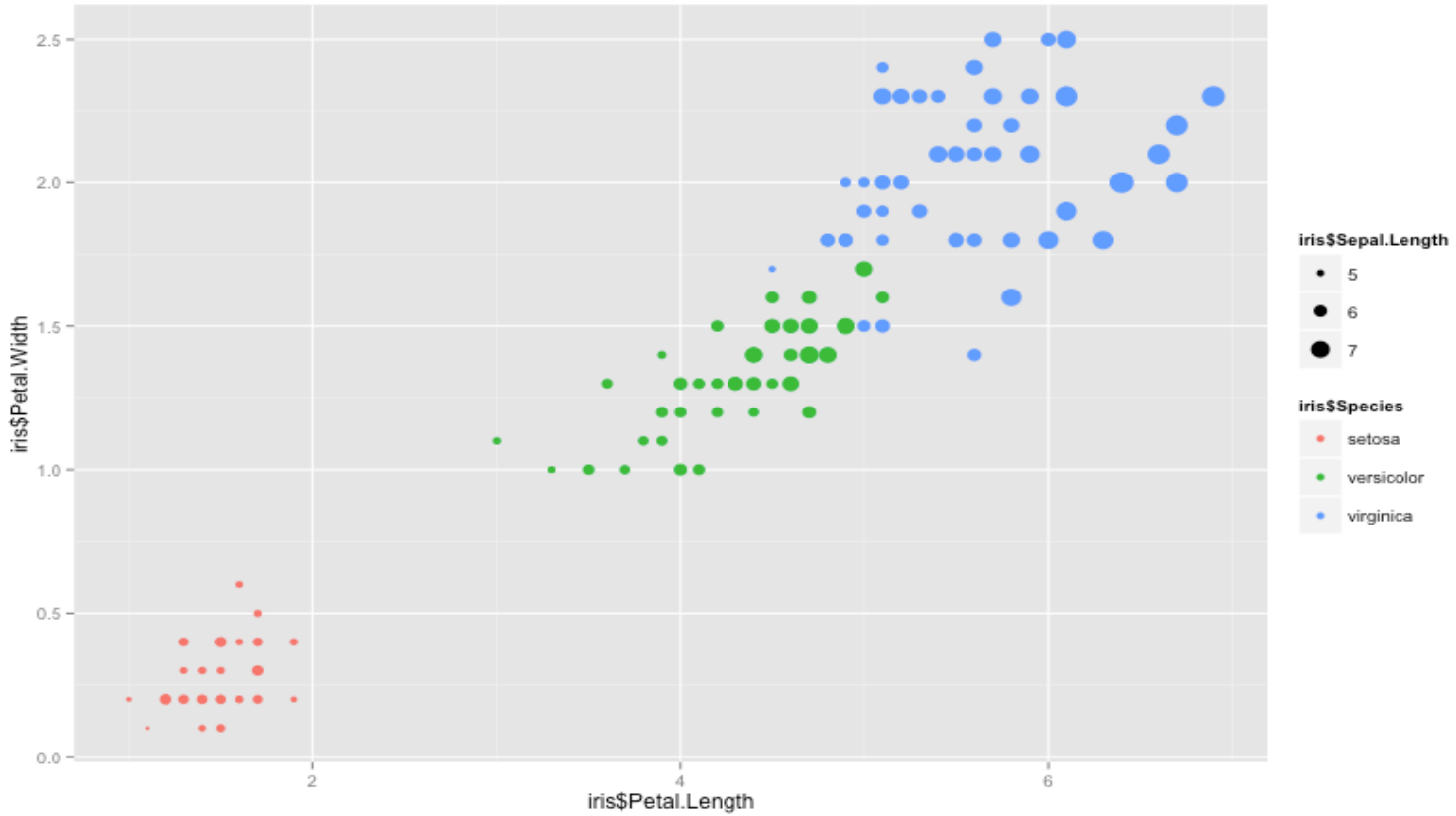
Plot

```
plot(iris$Petal.Length, iris$Petal.Width, pch=21,
     bg=c("red","green","yellow")[as.numeric(iris$Species)])
legend(locator(1), levels(iris$Species),
     fill=c("red","green","yellow"))
```



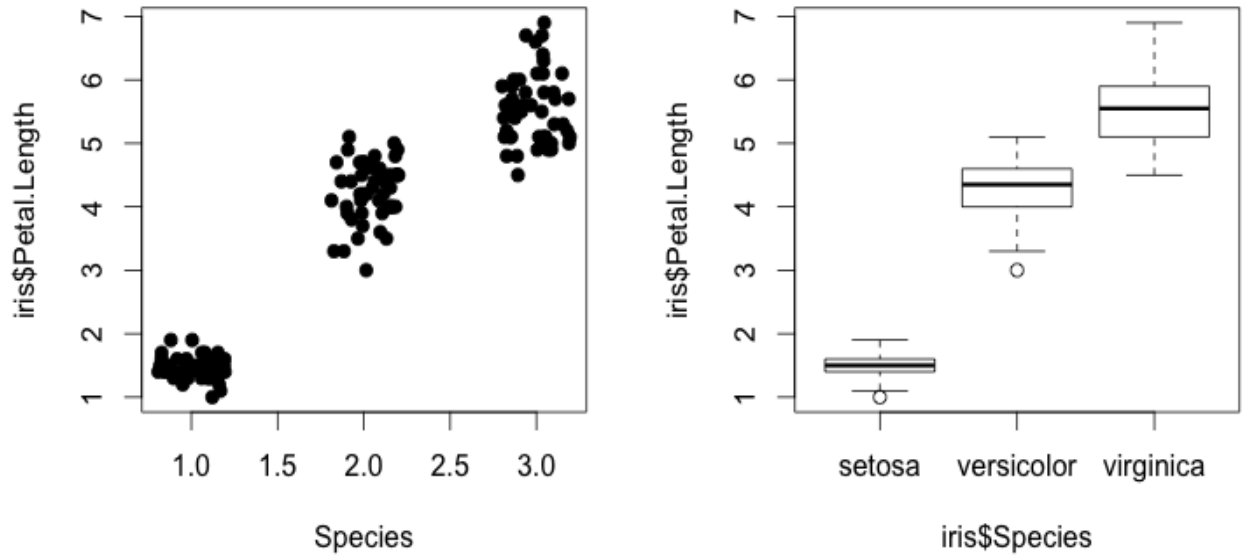
Outras bibliotecas para Plot

```
library(ggplot2)
qplot(iris$Petal.Length, iris$Petal.Width, col=iris$Species, size=iris$Sepal.Length)
```



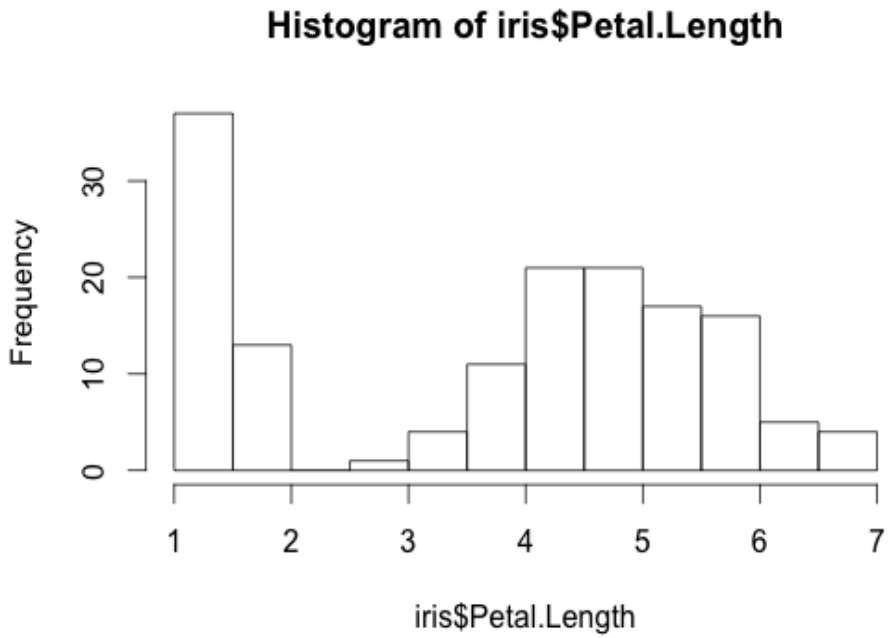
Comparando valores

```
par(mfrow=c(1,2))
plot(jitter(as.numeric(iris$Species)), iris$Petal.Length, pch=19, xlab="Species")
plot(iris$Petal.Length ~ iris$Species)
```



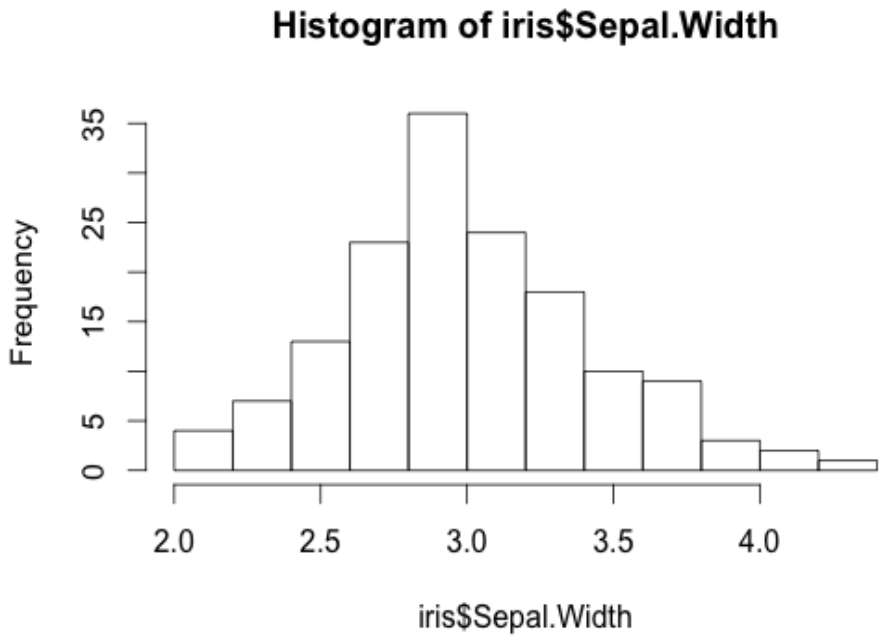
Histograma

```
> hist(iris$Petal.Length)
> summary(iris$Petal.Length)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.600   4.350   3.758   5.100   6.900
> var(iris$Petal.Length)
[1] 3.116278
```



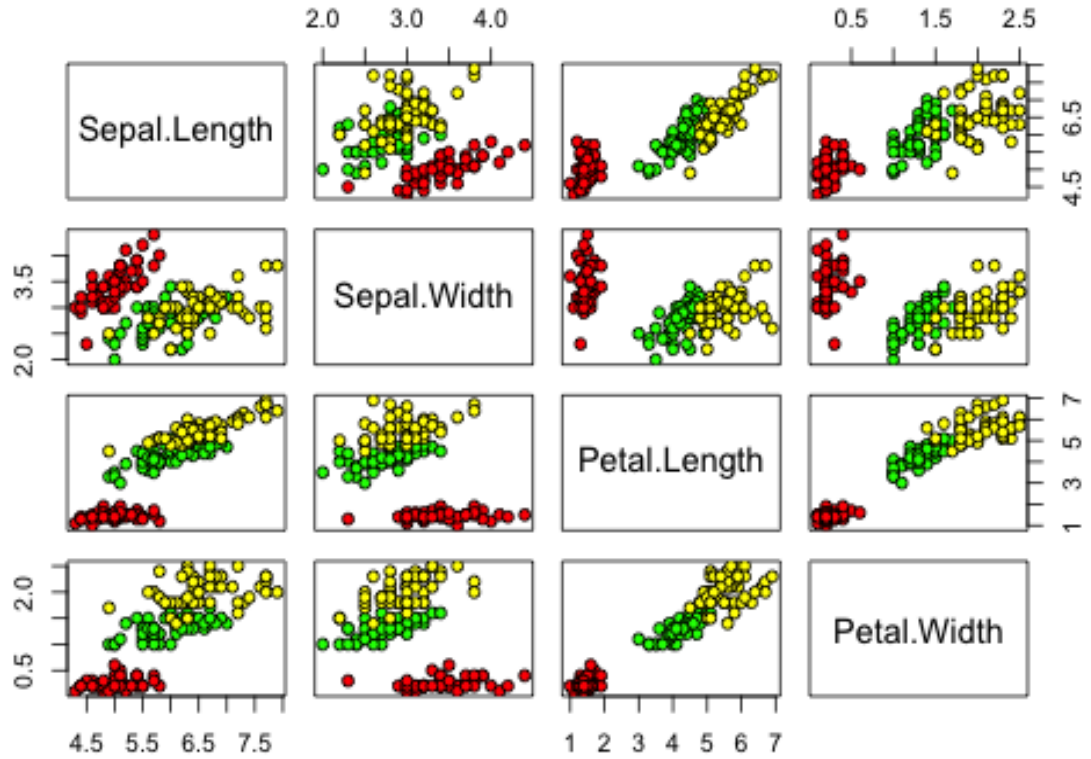
Histograma

```
> hist(iris$Sepal.Width)
> summary(iris$Sepal.Width)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.800   3.000   3.057   3.300   4.400
> var(iris$Sepal.Width)
[1] 0.1899794
```



Scatter Plot

```
plot(iris[,1:4], pch=21,
     bg=c("red","green","yellow")[as.numeric(iris$Species)])
```



Correlação

Dados multivariados permitem análises da relação entre dois ou mais atributos. Por exemplo, para atributos quantitativos, pode-se utilizar uma medida de correlação para identificar a relação linear entre dois atributos.

Coeficiente de correlação de Pearson

Este coeficiente, normalmente representado por ρ assume apenas valores entre -1 e 1.

- $\rho = 1$ significa uma correlação perfeita positiva entre as duas variáveis.
- $\rho = -1$ significa uma correlação perfeita negativa entre as duas variáveis.
- $\rho = 0$ significa que as duas variáveis não dependem linearmente uma da outra. No entanto, pode existir uma depência não linear. Assim, o resultado $\rho = 0$ deve ser investigado por outros meios.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

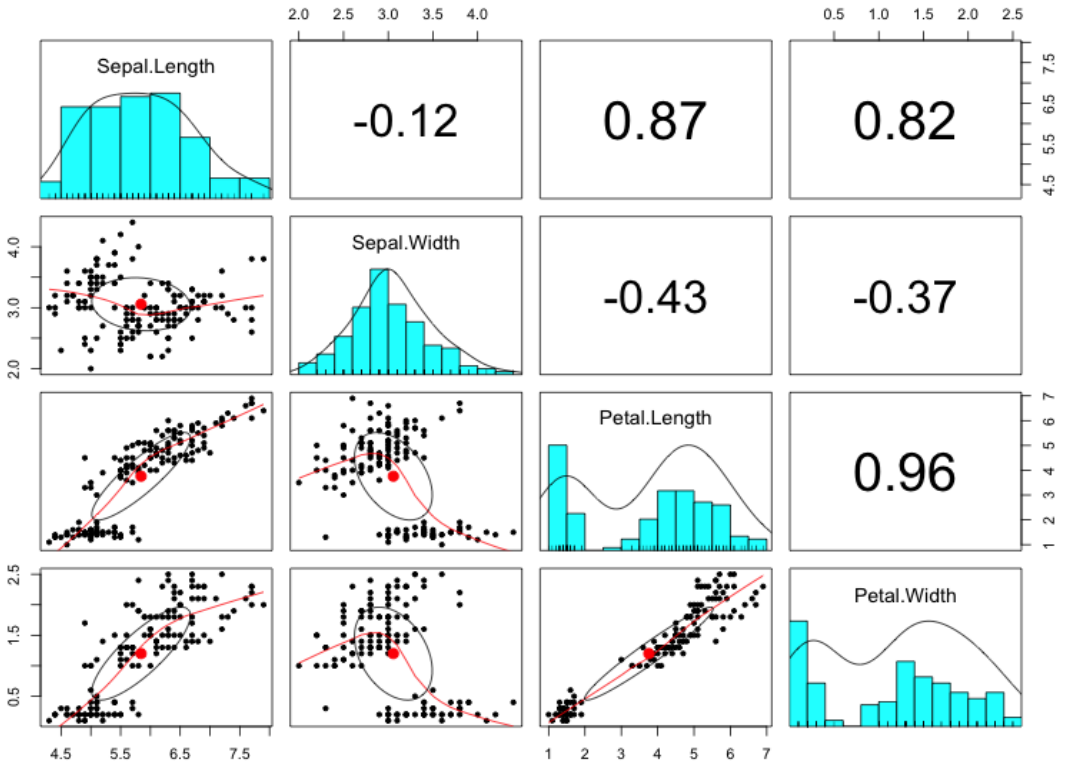
$$\rho = \frac{cov(X, Y)}{\sqrt{var(X) \times var(Y)}} \tag{4}$$

Exemplo de medidas de correlação

```
> cor(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

Resumindo a relação entre dados numéricos

```
library(psych)
pairs.panels(iris[,1:4])
```



Material de **consulta**

- Capítulo 3 do livro EMC Education Services, editor. Data Science and Big Data Analytics: Discovering, Analysing, Visualizing and Presenting Data. John Wiley & Sons, 2015.
- Hadley Wickham. Tidy data. Journal of Statistical Software, 59(10), 2014.

Próximas Atividades: Exercícios