

Formação Cientista de Dados

Introdução

R

- <http://www.r-project.org/>
 - Ambiente de Linha de Comando
 - Software **100%** open source
 - Ativo: builds frequentes
 - Extensível: pacotes
 - Multi-plataforma: Windows, Linux, Mac
 - Sensível a maiúsculas e minúsculas

R

- Milhares de funções de análise de dados* no estilo caixa-preta
- Ambiente de produção e visualização de gráficos
- Processamento em memória

Integração “out of box” com quase tudo

- Oracle
- SQL Server
- .NET
- Java
- Python
- Tableau
- Power BI
- Hadoop
- Etc.

RGui

- Ambiente de linha de comando simples, instalado por padrão
- Interface de digitação mais visualização de gráficos



RStudio

- IDE mais avançada
- Possui versão gratuita
- Será utilizada durante o curso

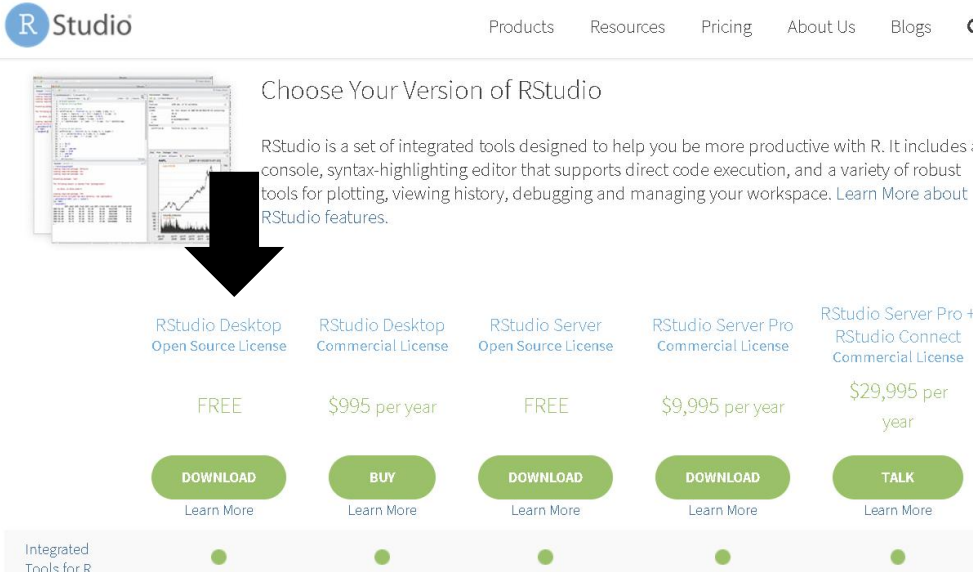


Formação Cientista de Dados

RStudio

R "core" e RStudio

- R: <https://cran.r-project.org/mirrors.html>
- RStudio: <https://www.rstudio.com/products/rstudio/download/>



R Studio Products Resources Pricing About Us Blogs Q

Choose Your Version of RStudio

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace. [Learn More](#) about RStudio features.

RStudio Desktop Open Source License	RStudio Desktop Commercial License	RStudio Server Open Source License	RStudio Server Pro Commercial License	RStudio Server Pro + RStudio Connect Commercial License
FREE	\$995 per year	FREE	\$9,995 per year	\$29,995 per year
DOWNLOAD Learn More	BUY Learn More	DOWNLOAD Learn More	DOWNLOAD Learn More	TALK Learn More

Integrated Tools for R



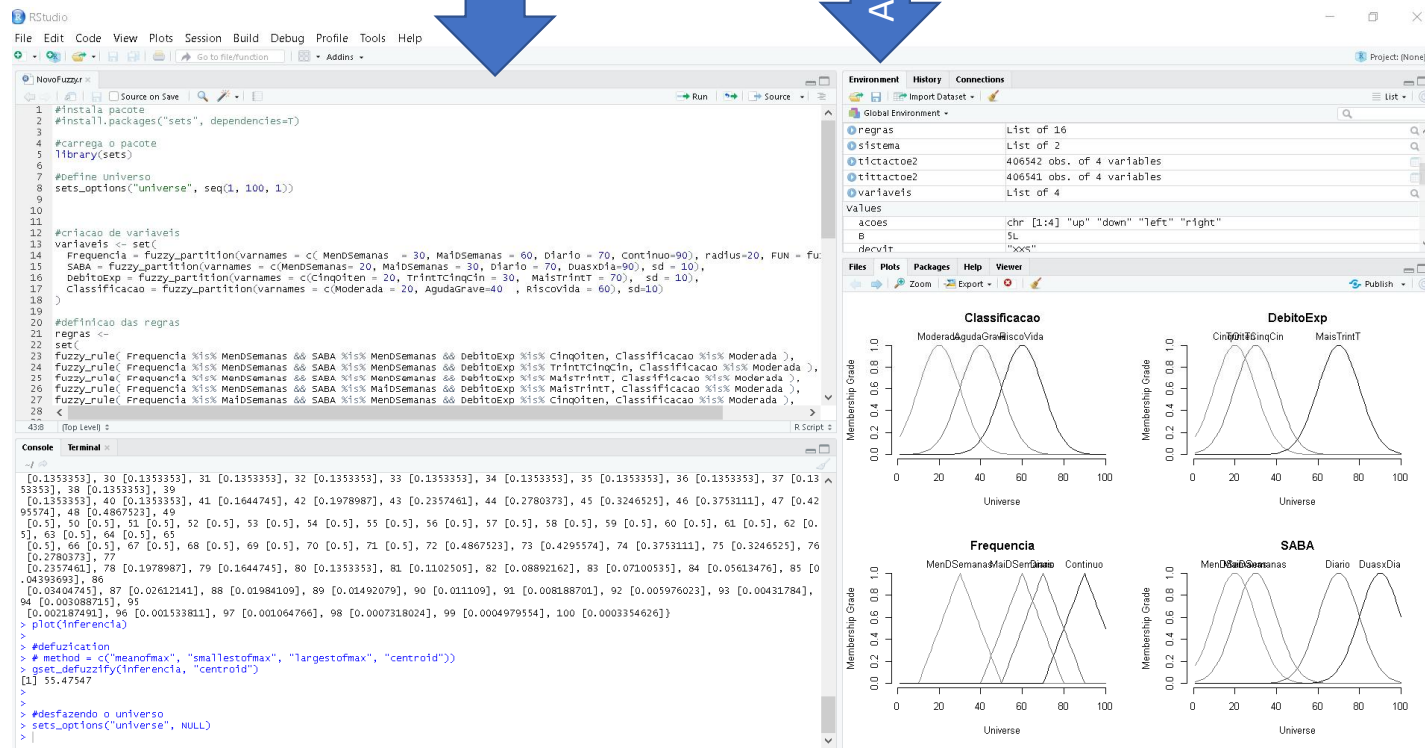
RStudio

Scripts

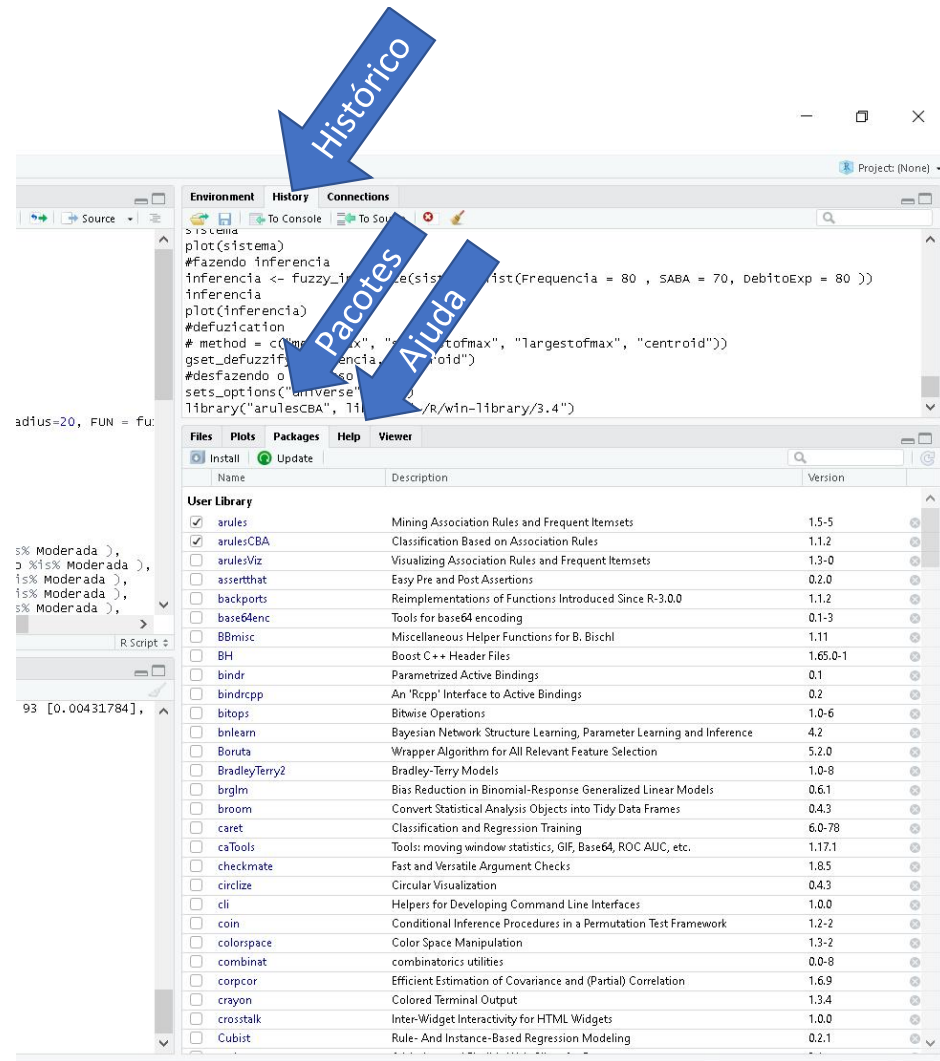
Ambiente

Gráficos

Console



RStudio



RStudio

```
modelling examples from 'An
Introduction to Statistical
Modelling' by Annette Dobson
nlm      Nonlinear least-squares using nlm()
smooth  'Visualize' steps in Tukey's
smoothers
```

Console Terminal x

```
> contour(x, y, volcano, levels = lev, col="yellow", lty="solid", add=TRUE)
> box()
> title("A Topographic Map of Maunga whau", font= 4)
> title(xlab = "Meters North", ylab = "Meters West", font= 3)
> mtext("10 Meter Contour Spacing", side=3, line=0.35, outer=FALSE,
+       at = mean(par("usr")[1:2]), cex=0.7, font=3)
```

◆ identify {graphics} ^
> ◆ identity {base} |
> ◆ if {base} |
> ◆ ifelse {base} |
H ◆ image {graphics} v
◆ image.default {graphics}
> ◆ implicitGeneric {methods}
> ◆ importIntoEnv {base}

ifelse(test, yes, no)
ifelse returns a value with the same shape as test which is filled with elements selected from either yes or no depending on whether the element of test is TRUE or FALSE.
Press F1 for additional help



Notebooks



Formação Cientista de Dados

Pacotes

Packages

- Implementam funções
- Desenvolvidos no mundo inteiro
- Totalmente open source
- Existem mais de 10 mil !!!

Exemplos

- Machine Learning
- Gráficos
- Series Temporais
- Distribuições de Probabilidade
- Finanças
- Genética
- Etc.

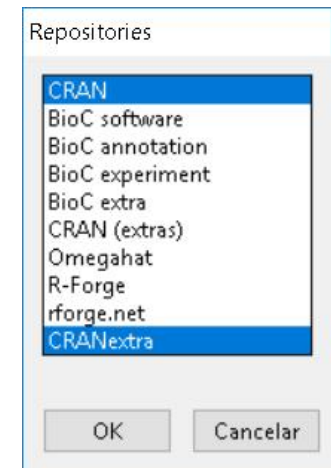
Pacotes populares

- Dplyr: manipulação de dados
- Devtools: desenvolvimento (criação de pacotes)
- Foreign: importar dados de outras ferramentas (SAS, SPSS etc)
- Ggplot2: visualização



Pacotes

- [The Comprehensive R Archive Network:](https://cran.r-project.org/)
<https://cran.r-project.org/>
- Repositórios e Espelhos (Mirrors)

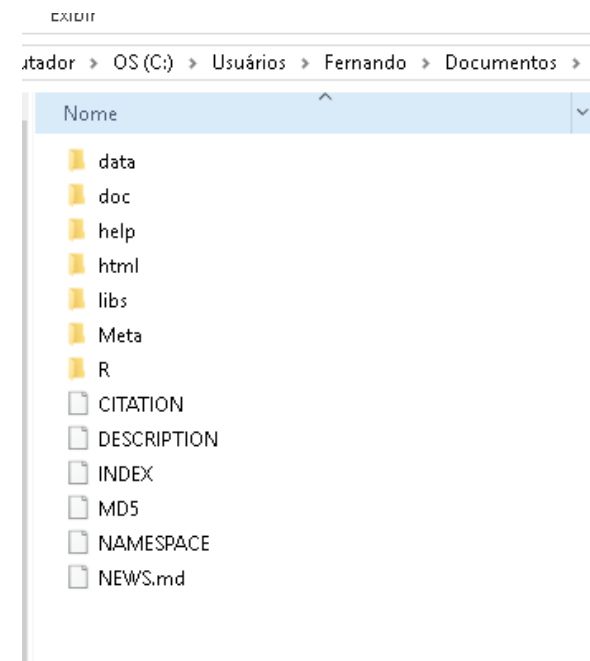
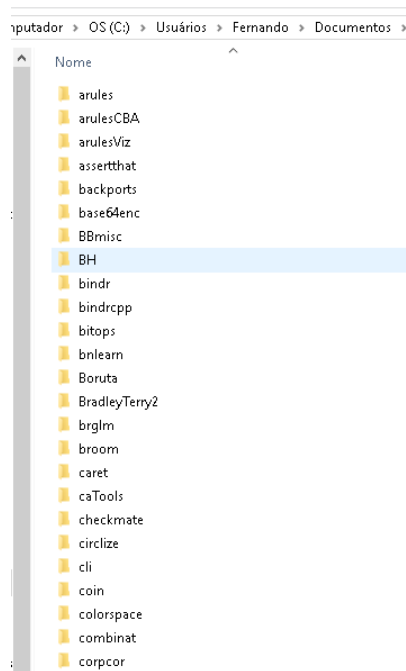
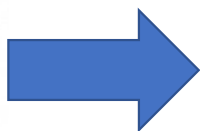


Pacotes



**PROF.
FERNANDO
AMARAL**
www.datascientist.com.br

Pacotes



C:\Users\Fernando\Documents\R\win-library\3.4



Instalação

- Linha de comando
- Manualmente

Instalação – Linha de Comando

- `install.packages("arules", dependencies=TRUE)`
- Seleciona o Espelho do CRAN e aguarda o download
- Verifica a mensagem de instalação ou eventual problema



Instalação Manual: Parte I

- Localiza a página do CRAN do pacote
- Download dos binários conforme SO

arules: Mining Association Rules and Frequent Itemsets

Provides the infrastructure for representing, manipulating and analyzing transaction data and patterns (frequent itemsets and association rules). Also provides C implementations of the association mining algorithms Apriori and Eclat.

Version: 1.5-5
Depends: R ($\geq 3.4.0$), [Matrix](#) ($\geq 1.2-0$)
Imports: stats, methods, graphics, utils
Suggests: [pmmi](#), [XML](#), [arulesViz](#), [testthat](#)
Published: 2018-01-10
Author: Michael Hahsler [aut, cre, cph], Christian Buchta [aut, cph], Bettina Gruen [aut, cph], Kurt Hornik [aut, cph], Ian Johnson [ctb, cph], Christian Borgelt [ctb, cph]
Maintainer: Michael Hahsler <mhahsler@lyle.smu.edu>
BugReports: <https://github.com/mhahsler/arules>
License: [GPL-3](#)
Copyright: The code for apriori and eclat in src/rapriori.c was obtained from <http://www.borgelt.net/> and is Copyright (C) 1996-2003 Christian Borgelt. All other code is Copyright (C) Michael Hahsler, Christian Buchta, Bettina Gruen and Kurt Hornik.
URL: <https://github.com/mhahsler/arules>, <http://lyle.smu.edu/TDA/arules>
NeedsCompilation: yes
Classification/ACM: G.4, H.2.8, I.5.1
Citation: [arules citation info](#)
Materials: [README NEWS](#)
In views: [MachineLearning](#)
CRAN checks: [arules results](#)

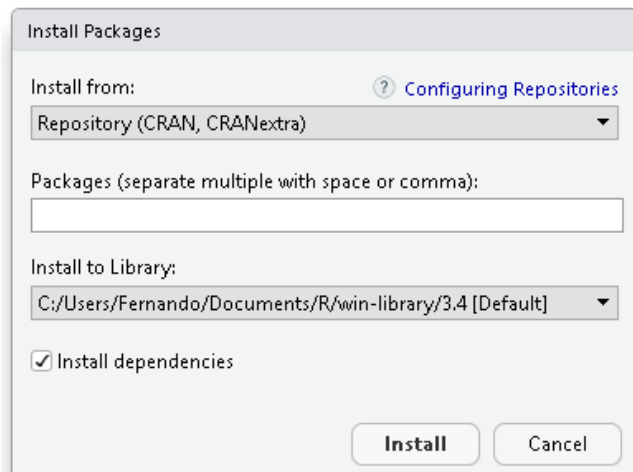
Downloads:

Reference manual: [arules.pdf](#)
Vignettes: [Introduction to arules](#)
Package source: [arules_1.5-5.tar.gz](#)
Windows binaries: r-devel: [arules_1.5-5.zip](#), r-release: [arules_1.5-5.zip](#), r-oldrel: [arules_1.5-4.zip](#)
OS X El Capitan binaries: r-release: [arules_1.5-5.tgz](#)
OS X Mavericks binaries: r-oldrel: [arules_1.5-4.tgz](#)
Old sources: [arules archive](#)



Instalação Manual: Parte II

- RSudio: Acessar menu tools, Install Packages



Install Packages

Install from: [? Configuring Repositories](#)

Repository (CRAN, CRANextra) ▼

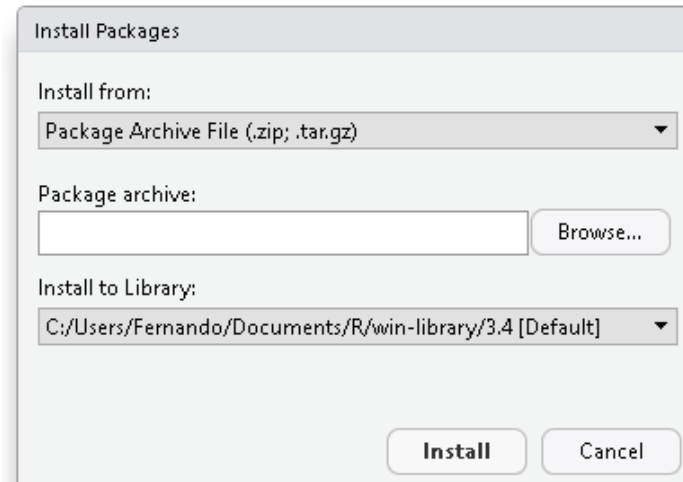
Packages (separate multiple with space or comma):

Install to Library:

C:/Users/Fernando/Documents/R/win-library/3.4 [Default] ▼

☒ Install dependencies

Install Cancel



Install Packages

Install from:

Package Archive File (.zip; .tar.gz) ▼

Package archive:

Browse...

Install to Library:

C:/Users/Fernando/Documents/R/win-library/3.4 [Default] ▼

Install Cancel



Carregar e Descarregar Pacote

```
library(arules)
```

```
detach("package:arules", unload=TRUE)
```


CRAN Task Views

CRAN Task Views

Bayesian
ChemPhys
ClinicalTrials
Cluster
DifferentialEquations
Distributions
Econometrics
Environmetrics
ExperimentalDesign
ExtremeValue
Finance
FunctionalData
Genetics
Graphics
HighPerformanceComputing
MachineLearning
MedicalImaging
MetaAnalysis
Multivariate
NaturalLanguageProcessing
NumericalMathematics
OfficialStatistics
Optimization
Pharmacokinetics
Phylogenetics
Psychometrics
ReproducibleResearch
Robust
SocialSciences
Spatial
Bayesian Inference
Chemometrics and Computational Physics
Clinical Trial Design, Monitoring, and Analysis
Cluster Analysis & Finite Mixture Models
Differential Equations
Probability Distributions
Econometrics
Analysis of Ecological and Environmental Data
Design of Experiments (DoE) & Analysis of Experimental Data
Extreme Value Analysis
Empirical Finance
Functional Data Analysis
Statistical Genetics
Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
High-Performance and Parallel Computing with R
Machine Learning & Statistical Learning
Medical Image Analysis
Meta-Analysis
Multivariate Statistics
Natural Language Processing
Numerical Mathematics
Official Statistics & Survey Methodology
Optimization and Mathematical Programming
Analysis of Pharmacokinetic Data
Phylogenetics, Especially Comparative Methods
Psychometric Models and Methods
Reproducible Research
Robust Statistical Methods
Statistics for the Social Sciences
Analysis of Spatial Data

- <https://cran.r-project.org/web/views/>
- Agrupamentos de pacotes e recursos por assuntos



Formação Cientista de Dados

Aspectos Diversos

Diretório de Trabalho

- Local onde o R busca por padrão os arquivos
- Ideal para não precisar informar caminhos absolutos

- Saber o diretório padrão

`getwd()`

- Alterar o diretório padrão

`setwd("c:\\dados")`



Notebook

- O diretório de trabalho é o local do arquivo Rmd
- Se você alterar o diretório de trabalho, só dura durante o bloco

Encerrando o R

`quit()`

Save workspace image to c:/dados/.RData? [y/n]:

Classes

```
class(iris)
```

Salvando e Carregando Objetos

```
save(objetos, file="arquivo.Rdata")
```

```
load(file="arquivo.Rdata")
```

Visualização de Dados

`plot()`: função genérica

`hist()`

`boxplot()`

Formação Cientista de Dados

Importando Dados

Tipos de Dados

- Caractere
- Numérico
- Inteiro
- Fator

Atribuição de Valor

=

<-

delta <- 8

delta = 8

Declaração de Variável

- Implícita

delta <- 8

delta <- 8L

logico <- TRUE

logico <- F

caractere <- "Texto"

Linguagem Vetorial

delta

[1] 8

Principais Operadores

+	Adição
-	Subtração
*	Multiplicação
/	Divisão
^	Potência
%%	Modo
%/%	Divisão de Inteiros

Operadores Lógicos

<	Menor que
>	Maior que
<=	Menor ou igual que
>=	Maior ou igual que
==	Igual
!=	Diferente
!	Not
	Ou
&	E





Funções Matemáticas

abs	Valor absoluto
sqrt	Raiz quadrada
sum	Soma
log	Logaritmo base 10
cos	Cosseno
sin	Seno
tan	Tangente
exp	Exponencial

Formação Cientista de Dados

Estrutura de Dados

Vetores

- Qualquer objeto declarado

```
X <- 8
```

Vetor de uma posição

Vetores

```
X <- c(1,2,3,4,5,6)
```

- Vetor 6 posições

```
X
```

- Le todo o vetor

```
X[1]
```

- Lê a posição 1

```
X[1] <- 10
```

- Altera a posição 1

Matrizes

- Duas dimensões (linhas e colunas)
- Permite um único tipo de dados
- Linhas e colunas podem ter nomes

Matrizes

- Ler ou alterar posição:
- Volcano[linha,coluna]

Data Frame

- Semelhante a Matrizes, porém:
 - Permite diferentes tipos de dados por coluna
- Duas dimensões (linhas e colunas)
- Linhas e colunas podem ter nomes

- Sintaxe para acessar coluna

Dataframe\$coluna

Listas

- N objetos, em sequências, de classes diferentes

Harman23.cor[1]

Harman23.cor[2]



Fatores

- Variáveis Categóricas

```
Dados = c(1,2,3)
```

```
Dados = factor(dados)
```


Formação Cientista de Dados

Funções

Funções

- Semelhantes a functions e procedures de qualquer linguagem
- Podem ou não requer argumentos (parâmetros)

- Exemplo

```
> getwd()
```

```
[1] "C:/Users/Fernando/Documents"
```

```
> sd(x)
```

```
[1] 14.93039
```

Argumentos

- O R é flexível com argumentos:
 - Você pode simplesmente passar os argumentos pela ordem esperada, sem nome
 - Você pode nomear os argumentos
 - Você passar os primeiros sem nome e os últimos nomeados, omitindo intermediários

```
head(x=iris, n=2)
```

```
head(iris)
```

```
head(iris,2)
```

```
head(n=22)
```

```
Error in head.default(n = 22) : argumento "x" ausente, sem padrão
```

Formação Cientista de Dados

Ajuda

Ajuda

help(sd)

mean {base}

Biblioteca

Arithmetic Mean

Título

Description

Generic function for the (trimmed) arithmetic mean.

Descrição

Usage

```
mean(x, ...)
```

Formas de uso

```
## Default S3 method:
```

```
mean(x, trim = 0, na.rm = FALSE, ...)
```

Mostra valores padrão dos argumentos

Argumentos

Arguments

x

An R object. Currently there are methods for numeric/logical vectors and [date](#), [date-time](#) and [time interval](#) objects. Complex vectors are allowed for `trim = 0`, only.

trim

the fraction (0 to 0.5) of observations to be trimmed from each end of `x` before the mean is computed. Values of `trim` outside that range are taken as the nearest endpoint.

na.rm

a logical value indicating whether NA values should be stripped before the computation proceeds.

...

further arguments passed to or from other methods.



x
An R object. Currently there are methods for numeric/logical vectors and [date](#), [date-time](#) and [time interval](#) objects. Complex vectors are allowed for `trim = 0`, only.

trim
the fraction (0 to 0.5) of observations to be trimmed from each end of `x` before the mean is computed. Values of `trim` outside that range are taken as the nearest endpoint.

na.rm
a logical value indicating whether NA values should be stripped before the computation proceeds.

...
further arguments passed to or from other methods.

Value

Resultado

If `trim` is zero (the default), the arithmetic mean of the values in `x` is computed, as a numeric or complex vector of length one. If `x` is not logical (coerced to numeric), numeric warning.

If `trim` is non-zero, a symmetrically trimmed mean is computed with a fraction of `trim` observations deleted from each end before the mean is computed.

References

Referências

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.

See Also

[weighted.mean](#), [mean.POSIXct](#), [colMeans](#) for row and column means.

Funções Relacionadas

Examples

```
x <- c(0:10, 50)
xm <- mean(x)
c(xm, mean(x, trim = 0.10))
```

Exemplos



AMARAL
www.datascientist.com.br

Formação Cientista de Dados

Principais Funções

Principais Funções

- Visualizar primeiras linhas de um conjunto de dados

`head()`

- Visualizar últimas linhas de um conjunto de dados

`tail()`

- Resumo estatístico de um conjunto de dados

`summary()`

- Caminho de arquivos

`file.choose()`

Principais Funções

- Dimensões de um conjunto de dados (numero de colunas e número de linhas)

`dim()`

- Comprimento de um vetor

`length()`

- Nomes das colunas de um conjunto de dados

`colnames()`

- Nomes das linhas de um conjunto de dados

`rownames()`

- Adiciona coluna

`colbind()`

Principais Funções

- Funções genéricas de Machine Learning
`predict()`

- Formula

$$VD_1 + VD_2 + VD_n \sim VI_1 + VI_2 + VI_n$$

Vendas \sim Temperatura

Formação Cientista de Dados

Importando Dados

Texto

read.csv()

Banco de Dados

- Pacote RODB

odbcDriverConnect()

sqlQuery()

odbcClose()



Planilha Excel



PACOTE XLSX



READ.XLSX



Outras Ferramentas: Weka

- Pacote foreign

```
read.arff()
```



Formação Cientista de Dados

Programação

If / ifelse

```
if (condicao)
```

```
{
```

```
}
```

```
else
```

```
{
```

```
}
```

ifelse

ifelse (condição, ret T, ret F)

Laços

```
for(var in seq)  
{  
}
```

```
while(condicao)  
{  
}
```

```
break  
next
```

Funções

```
nome <- function(parametros) {
```

```
  return = x
```

```
}
```

Formação Cientista de Dados

Referências Adicionais

Recursos Adicionais



Referências
Adicionais

