FACULDADE DE TECNOLOGIA

Prof. Dr. Ivan Luiz Marques Ricarte

FT-UNICAMP

2016

Minicurso: “R - Software Estatístico”

Dia 29/09/2016

Horário	Evento	Local
08:00 às 12:00	Minicurso: R - Software Estatístico - Prof. Dr. Ivan Ricarte (FT-UNICAMP)	PA03

FACULDADE DE TECNOLOGIA

Objetivos

Ao final deste curso, você poderá:

- Instalar um ambiente para desenvolvimento e execução de programas em R
- Encontrar, instalar e utilizar pacotes (bibliotecas) que estendem R
- Realizar análises e manipulações de dados básicas em R
- Gerar gráficos de alta qualidade em R
- Produzir documentos de pesquisa reproduzível

FACULDADE DE TECNOLOGIA

Apresentação

FACULDADE DE TECNOLOGIA

www.ft.unicamp.br/~ricarte/

INÍCIO PESQUISA ENSINO ARTIGOS LIVROS VÍDEOS CONTATO

Professor Titular, Faculdade de Tecnologia, UNICAMP

Pós-doutorado (2011-2012), Faculty of Medicine, McGill University
Livre-docente (2001), Faculdade de Engenharia Elétrica e de Computação, UNICAMP
Ph.D. (1994), A. James Clark School of Engineering, University of Maryland at College Park

Mestre (1987), Faculdade de Engenharia Elétrica, UNICAMP
Engenheiro Eletricista (1984), Faculdade de Engenharia de Campinas, UNICAMP

Membro da Sociedade Brasileira de Computação (SBC),
da Sociedade Brasileira de Informática em Saúde (SBIS),
da Association for Computing Machinery (ACM)

FACULDADE DE TECNOLOGIA

Primeiros passos

FACULDADE DE TECNOLOGIA

http://www.r-project.org/

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

Software livre (ambiente e linguagem)

Análise estatística e apresentação gráfica

Disponível para plataformas Unix, Windows e MacOS

FACULDADE DE TECNOLOGIA

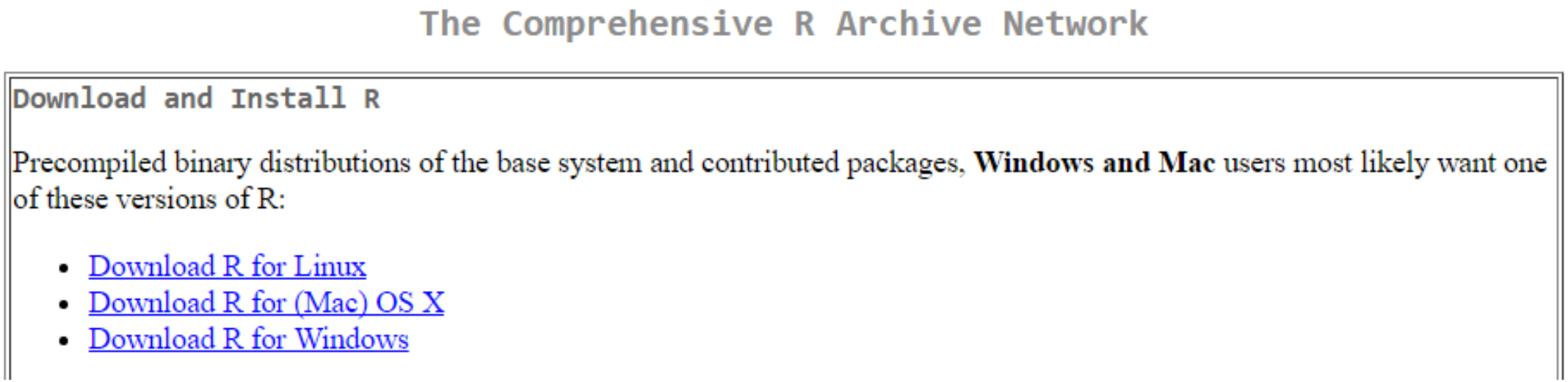
IEEE SPECTRUM

The 2016 Top Programming Languages

C is No. 1, but big data is still the big winner

Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

Instalação



http://cran.r-project.org/

Instalação de R em Windows

R-3.3.1 for Windows (32/64 bit)

Download R 3.3.1 for Windows (70 megabytes, 32/64 bit)

Installation and other instructions

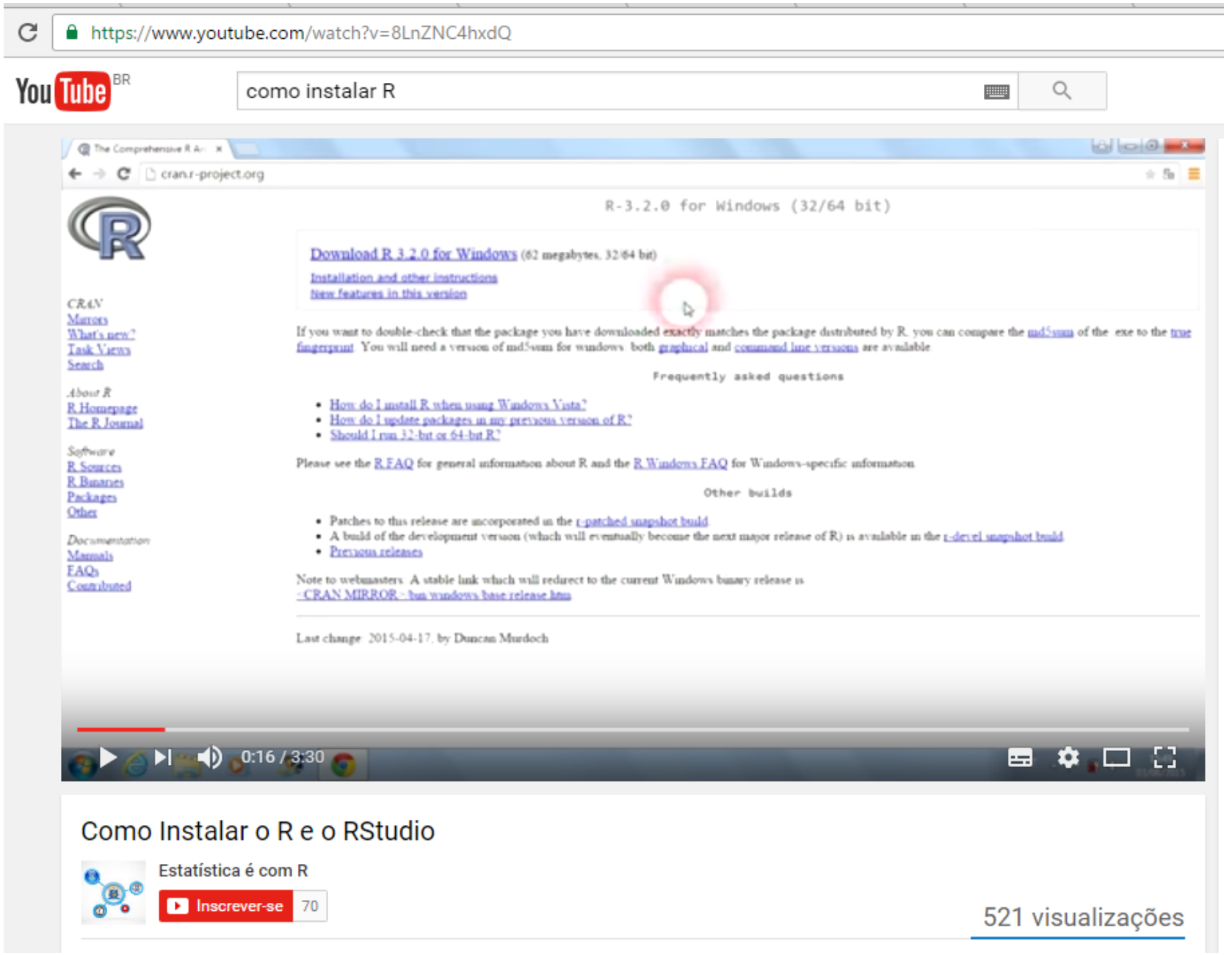
New features in this version

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the md5sum of the .exe to the true fingerprint. You will need a version of md5sum for windows: both graphical and command line versions are available.

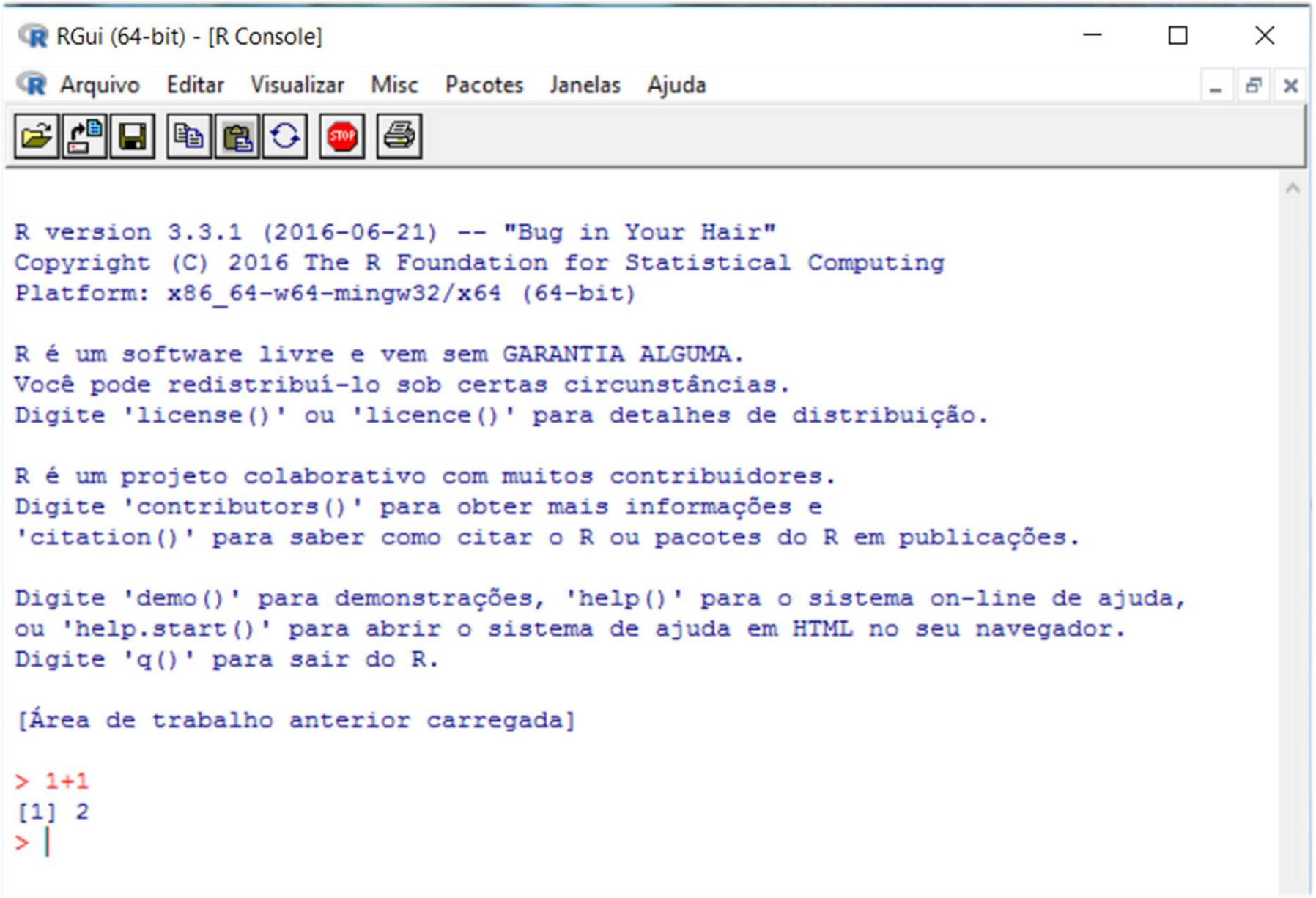
Frequently asked questions

- Does R run under my version of Windows?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

FACULDADE DE TECNOLOGIA



FACULDADE DE TECNOLOGIA



FACULDADE DE TECNOLOGIA

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, as well as tools for plotting, history, debugging and workspace management.

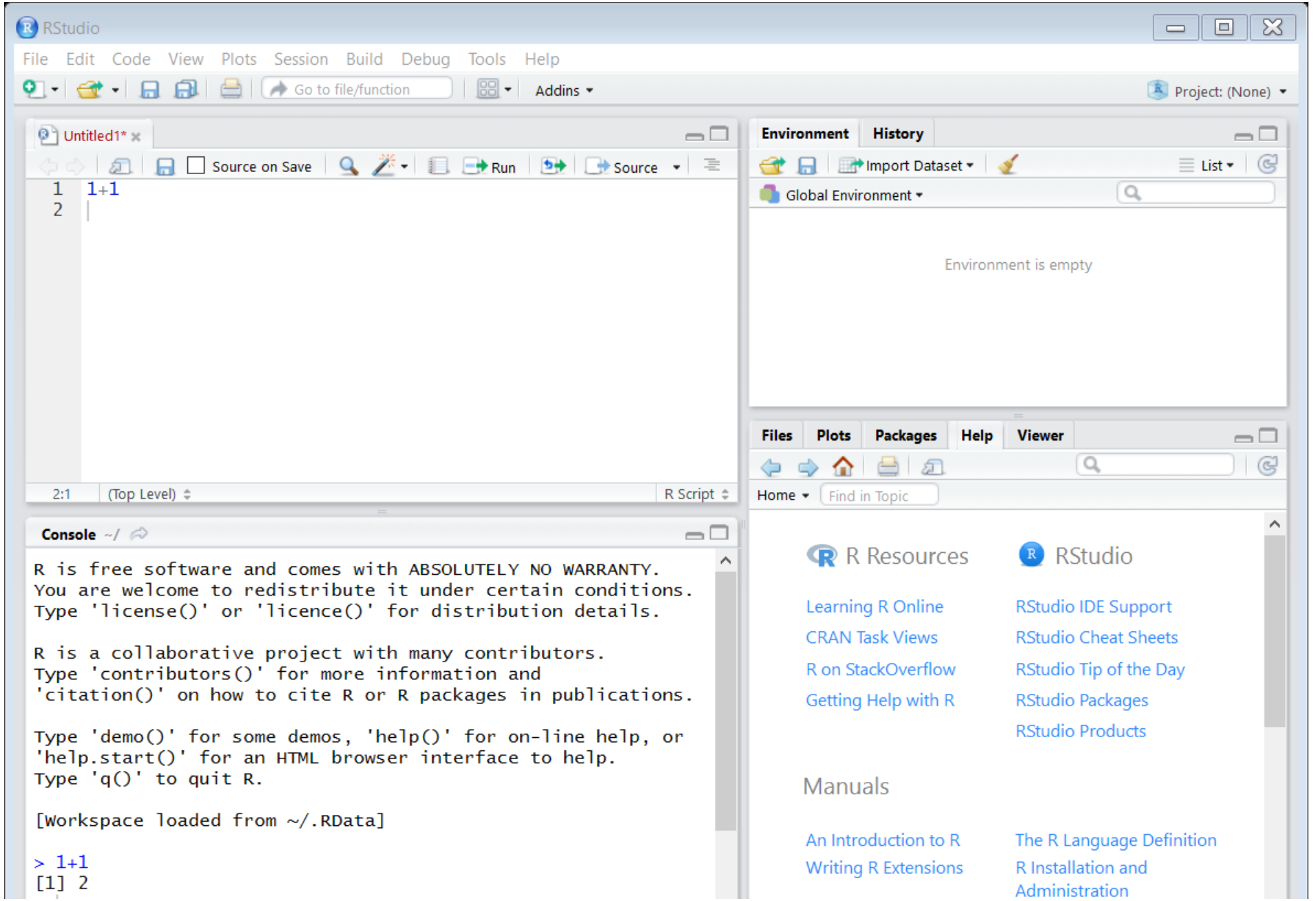
http://www.rstudio.com/products/rstudio/download/

FACULDADE DE TECNOLOGIA

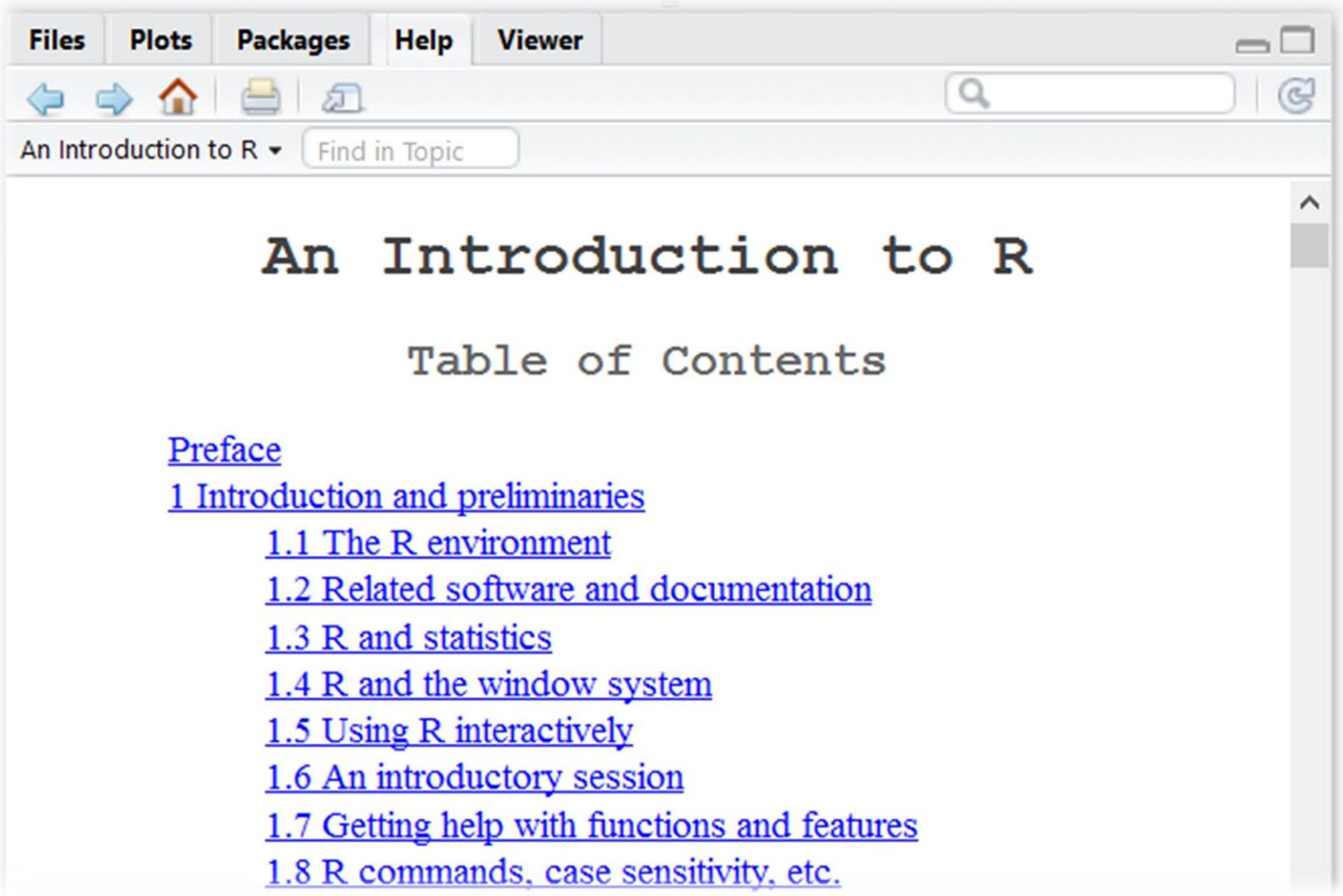
Installers for Supported Platforms

Installers	Size	Date
RStudio 0.99.903 - Windows Vista/7/8/10	77.1 MB	2016-07-18
RStudio 0.99.903 - Mac OS X 10.6+ (64-bit)	60 MB	2016-07-18
RStudio 0.99.903 - Ubuntu 12.04+/Debian 8+ (32-bit)	81.6 MB	2016-07-18
RStudio 0.99.903 - Ubuntu 12.04+/Debian 8+ (64-bit)	88.3 MB	2016-07-18
RStudio 0.99.903 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	81 MB	2016-07-18
RStudio 0.99.903 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	81.9 MB	2016-07-18

FACULDADE DE TECNOLOGIA



FACULDADE DE TECNOLOGIA



FACULDADE DE TECNOLOGIA

Sumário: primeiros passos

R é um ambiente e uma linguagem para computação estatística

R é software livre, disponível para plataformas Unix, Windows e MacOS

R é uma das linguagens mais utilizadas na atualidade

RStudio é um ambiente de desenvolvimento integrado para R

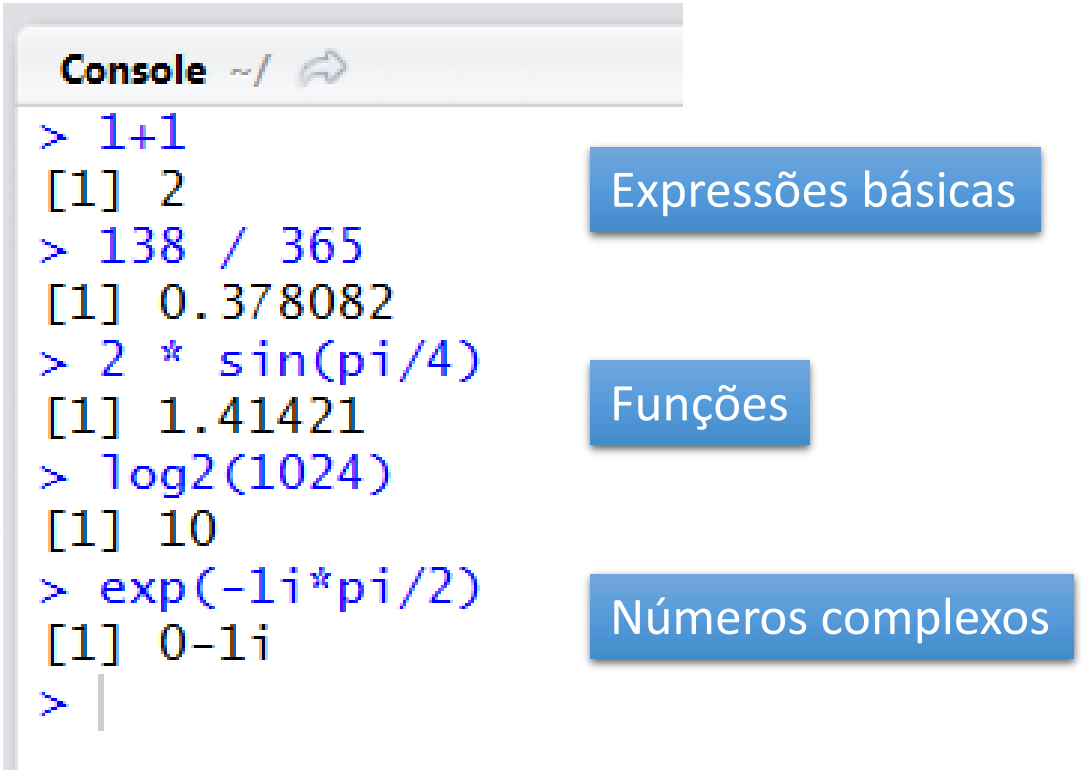
- RStudio também é software livre, disponível para diversas plataformas

FACULDADE DE TECNOLOGIA

Expressões e dados em R

R como uma calculadora

Expressões simples, com funções e constantes pré-definidas



FACULDADE DE TECNOLOGIA

Miscellaneous Mathematical Functions

Description

`abs(x)` computes the absolute value of x, `sqrt(x)` computes the (principal) square root of x, $\sqrt{x}$.

The naming follows the standard for computer languages such as C or Fortran.

Usage

```
abs(x)
sqrt(x)
```

FACULDADE DE TECNOLOGIA

Trigonometric Functions

Description

These functions give the obvious trigonometric functions. They respectively compute the cosine, sine, tangent, arc-cosine, arc-sine, arc-tangent, and the two-argument arc-tangent.

`cospi(x)`, `sinpi(x)`, and `tanpi(x)`, compute `cos(pi*x)`, `sin(pi*x)`, and `tan(pi*x)`.

Usage

```
cos(x)
sin(x)
tan(x)

acos(x)
asin(x)
atan(x)
atan2(y, x)

cospi(x)
sinpi(x)
tanpi(x)
```

FACULDADE DE TECNOLOGIA

Logarithms and Exponentials

Description

`log` computes logarithms, by default natural logarithms, `log10` computes common (i.e., base 10) logarithms, and `log2` computes binary (i.e., base 2) logarithms. The general form `log(x, base)` computes logarithms with base `base`.

`log1p(x)` computes *log(1+x)* accurately also for |*x*| << *1*.

`exp` computes the exponential function.

`expm1(x)` computes *exp(x) - 1* accurately also for |*x*| << *1*.

Usage

```
log(x, base = exp(1))
logb(x, base = exp(1))
log10(x)
log2(x)
```

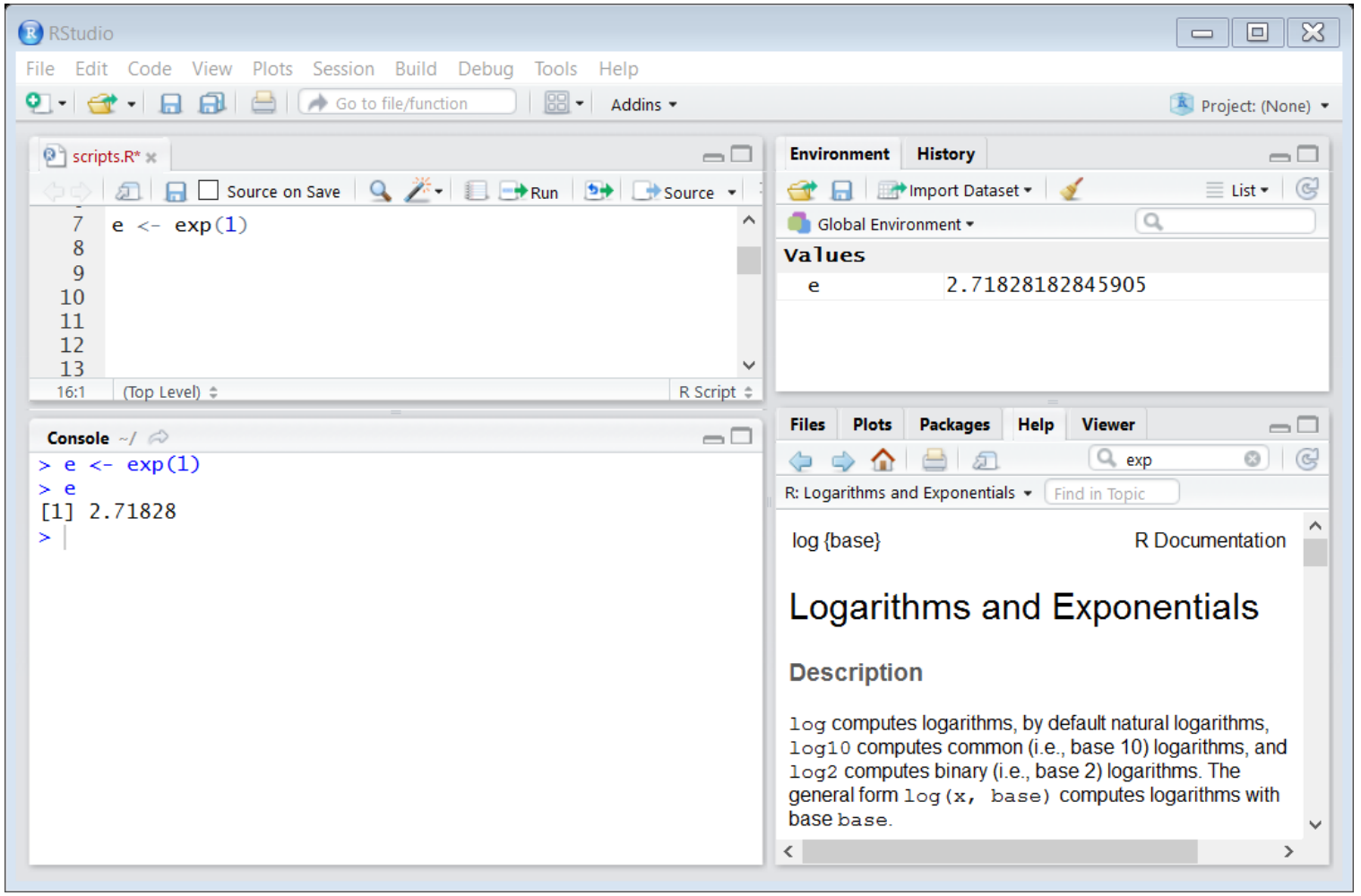
Definição de variáveis

Comando de atribuição

```
e <- exp(1)

e = exp(1)
```

Não é preciso declarar variáveis



Tipos de dados básicos: numéricos

Padrão: numérico (valor decimal)

```
> class(e)
[1] "numeric"
```

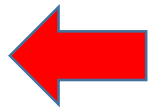
class {base}

Object Classes

Para definir valores inteiros: as.integer(x)

```
> k <- 1
> class(k)
[1] "numeric"
> is.integer(k)
[1] FALSE
>
> k <- as.integer(1)
> class(k)
[1] "integer"
> is.integer(k)
[1] TRUE
```

```
> 1/0
[1] Inf
>
> 1/Inf
[1] 0
```



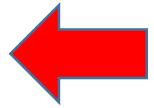
Infinito

Tipos de dados básicos: complexos

Números complexos

```
> z <- 1+ 1i
> class(z)
[1] "complex"
```

```
> sqrt(-1)
[1] NaN
Warning message:
In sqrt(-1) : NaNs produced
> sqrt(-1+0i)
[1] 0+1i
> sqrt(as.complex(-1))
[1] 0+1i
>
```



Not a number

Tipos de dados básicos: lógicos

```
> b <- e > 2
> b
[1] TRUE
> class(b)
[1] "logical"
> f <- FALSE
> b & f
[1] FALSE
> b | f
[1] TRUE
>
```

AND

OR

Tipos de dados básicos: caracteres

```
> nome <- "Ivan"
> class(nome)
[1] "character"
> sobrenome <- "Ricarte"
>
> paste(nome,sobrenome)
[1] "Ivan Ricarte"
>
> paste0(nome,sobrenome)
[1] "IvanRicarte"
>
> sprintf("Valor de e é %.3f", exp(1))
[1] "Valor de e é 2.718"
>
```

paste {base} R Documentation

Concatenate Strings

Description

Concatenate vectors after converting to character.

sprintf {base} R Documentation

Use C-style String Formatting Commands

Description

A wrapper for the C function `sprintf`, that returns a character vector containing a formatted combination of text and variable values.

Data em R: Date

```
> hoje <- Sys.Date()
> hoje
[1] "2016-09-19"
> hoje+60
[1] "2016-11-18"
> hoje - 6*365
[1] "2010-09-21"
>
> class(hoje)
[1] "Date"
>
```

Diferença em dias

Data e hora em R: POSIXct

```
> z <- Sys.time()
> class(z)
[1] "POSIXct" "POSIXt"
> z
[1] "2016-09-19 16:04:00 BRT"
>
> z + 3660
[1] "2016-09-19 17:05:00 BRT"
```

Diferença em segundos

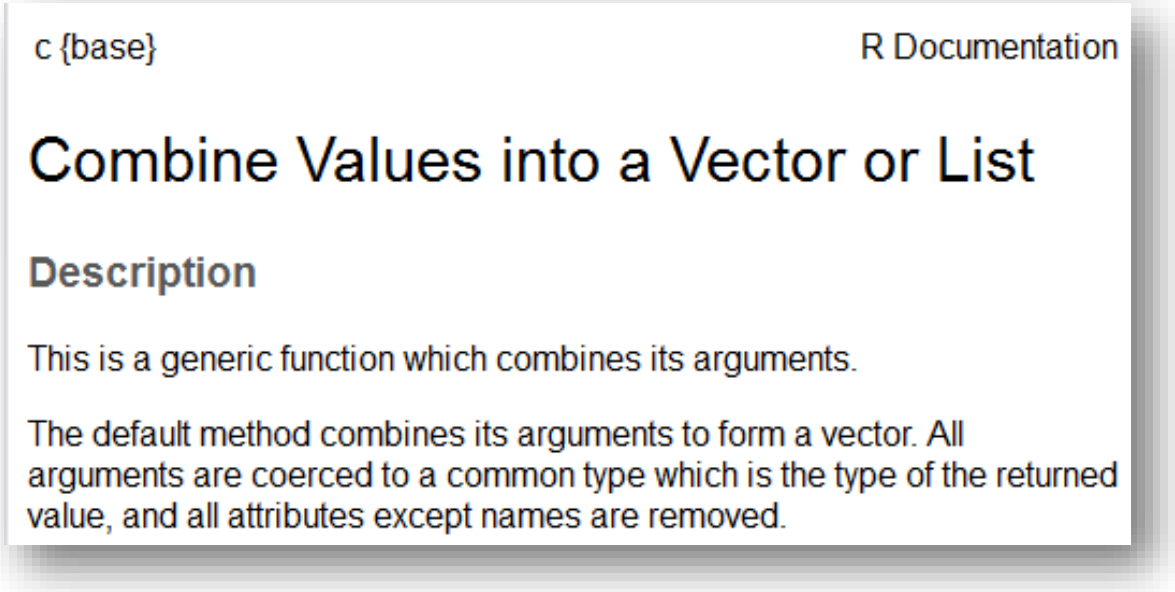
Conversão de string para data: as.Date()

```
> hoje2 <- as.Date("2016-09-29")
> hoje3 <- as.Date("29/09/2016", "%d/%m/%Y")
> hoje3-hoje2
Time difference of 0 days
```

FACULDADE DE TECNOLOGIA

Vetores

Sequência de valores de um mesmo tipo básico



c {base} R Documentation

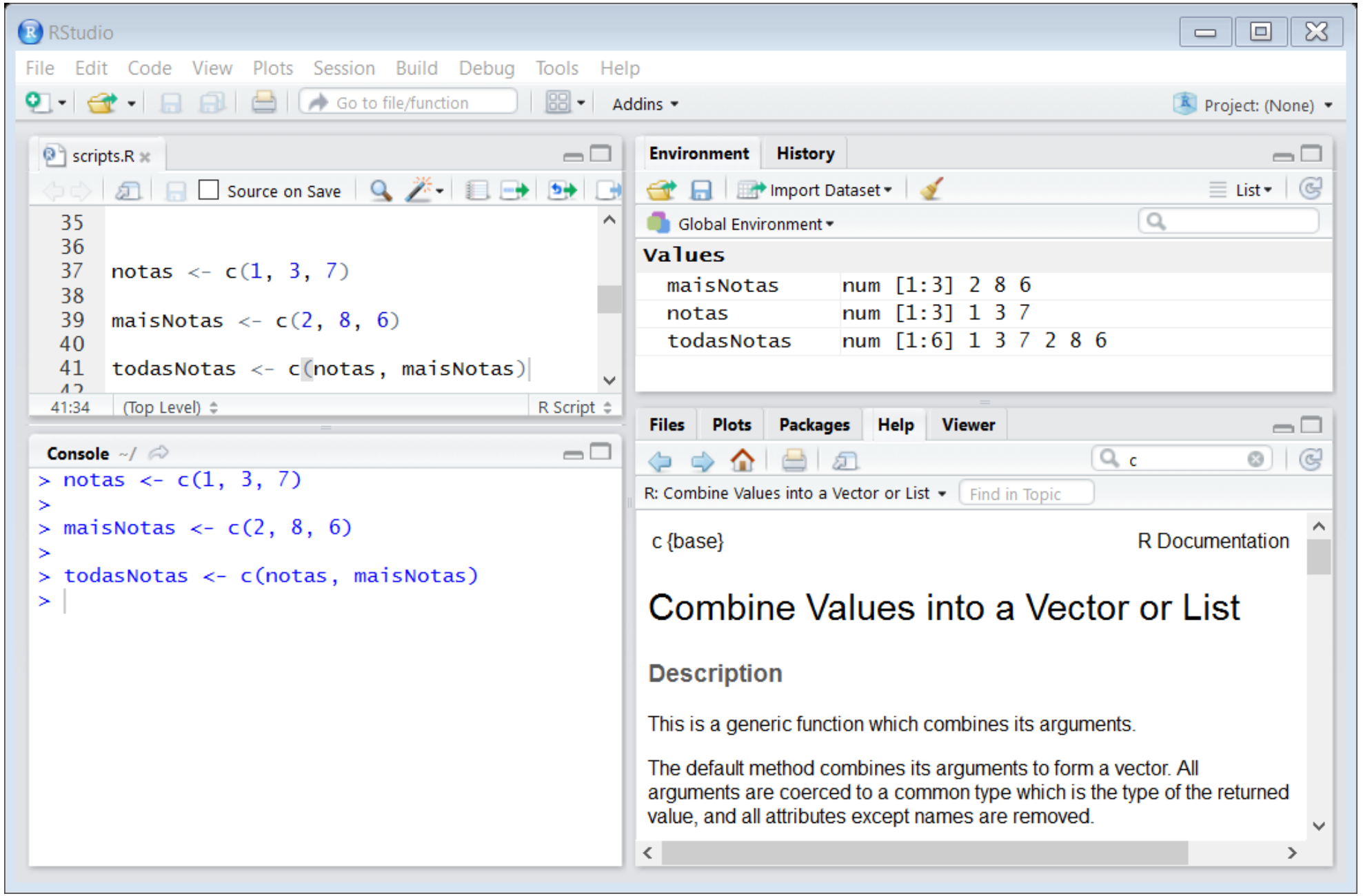
Combine Values into a Vector or List

Description

This is a generic function which combines its arguments.

The default method combines its arguments to form a vector. All arguments are coerced to a common type which is the type of the returned value, and all attributes except names are removed.

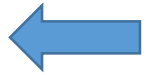
FACULDADE DE TECNOLOGIA



Aritmética de vetores

Values	
maisNotas	num [1:3] 2 8 6
notas	num [1:3] 1 3 7
todasNotas	num [1:6] 1 3 7 2 8 6

```
> 2 * todasNotas
[1]  2  6 14  4 16 12
> sqrt(maisNotas)
[1] 1.41421 2.82843 2.44949
> notas / maisNotas
[1] 0.50000 0.37500 1.16667
> notas + todasNotas
[1]  2  6 14  3 11 13
```



Reciclagem dos membros do vetor

Indexação de elementos do vetor

```
todasNotas      num [1:6] 1 3 7 2 8 6
```

```
> todasNotas[2]
[1] 3
> todasNotas[-2]
[1] 1 7 2 8 6
> todasNotas[3:5]
[1] 7 2 8
> todasNotas[-3:-5]
[1] 1 3 6
```

O elemento indicado

Todos elementos exceto o indicado

Todos os elementos entre os indicados

Todos os elementos exceto entre os indicados

FACULDADE DE TECNOLOGIA

```
todasNotas     num [1:6] 1 3 7 2 8 6
```

```
> todasNotas[7]
[1] NA
```

NA {base}

'Not Available' / Missing Values

```
> todasNotas[c(1, 4, 2)]
[1] 1 2 3
```

FACULDADE DE TECNOLOGIA

```
todasNotas     num [1:6] 1 3 7 2 8 6
```

```
> todasNotas[c(FALSE, TRUE, FALSE, TRUE, TRUE, FALSE)]
[1] 3 2 8
```

```
> todasNotas > 2 & todasNotas < 8
[1] FALSE  TRUE  TRUE FALSE FALSE  TRUE
```

```
> todasNotas[todasNotas > 2 & todasNotas < 8]
[1] 3 7 6
```

Vetores com membros nomeados

```
> ir <- c("Ivan", "Ricarte")
```

ir	chr [1:2] "Ivan" "Ricarte"

```
> names(ir) <- c("nome", "sobrenome")
```

ir	Named chr [1:2] "Ivan" "Ricarte"

```
> ir
     nome sobrenome
   "Ivan" "Ricarte"
```

```
> ir["sobrenome"]
sobrenome
"Ricarte"
```

```
> ir[c("sobrenome", "nome")]
sobrenome      nome
"Ricarte"    "Ivan"
```

Fatores

factor {base} R Documentation

Factors

Description

The function `factor` is used to encode a vector as a factor (the terms 'category' and 'enumerated type' are also used for factors). If argument `ordered` is `TRUE`, the factor levels are assumed to be ordered. For

FACULDADE DE TECNOLOGIA

rep {base} R Documentation

Replicate Elements of Vectors and Lists

```
areaConc <- c(rep("AA", 10), rep("AB", 15), rep("AC", 8))
```

areaConc	chr [1:33] "AA" "AA" "AA" "AA" "AA" ...

```
> table(areaConc)
areaConc
AA AB AC
10 15  8
```

table {base} R Documentation

Cross Tabulation and Table Creation

Description

`table` uses the cross-classifying factors to build a contingency table of the counts at each combination of factor levels.

```
> areaConcFac <- factor(areaConc)
> areaConcFac
 [1] AA AA AA AA AA AA AA AA AA AA AB AB AB AB AB AB AB AB AB AB AB AB AB AB AB
[26] AC AC AC AC AC AC AC AC
Levels: AA AB AC
```

Matrizes

matrix {base} R Documentation

Matrices

Description

`matrix` creates a matrix from the given set of values.

Matrizes: construção e operações

```
todasNotas    num [1:6] 1 3 7 2 8 6
```

```
> m <- matrix (todasNotas, ncol=2, nrow=3)
> m
     [,1] [,2]
[1,]    1    2
[2,]    3    8
[3,]    7    6
> t(m)
     [,1] [,2] [,3]
[1,]    1    3    7
[2,]    2    8    6
> cbind(t(m), c(5,4))
     [,1] [,2] [,3] [,4]
[1,]    1    3    7    5
[2,]    2    8    6    4
```

```
m    num [1:3, 1:2] 1 3 7 2 8 6
```

t {base} R Documentation

Matrix Transpose

cbind {base} R Documentation

Combine R Objects by Rows or Columns

cbind(), rbind()

FACULDADE DE TECNOLOGIA

```
> m <- matrix(c(1,2,3,4), ncol=2)
> m
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> solve(m)
     [,1] [,2]
[1,]   -2  1.5
[2,]    1 -0.5
> m * solve(m)
     [,1] [,2]
[1,]   -2  4.5
[2,]    2 -2.0
> m %*% solve(m)
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

Matriz inversa

Multiplicação elemento a elemento

Multiplicação matricial

FACULDADE DE TECNOLOGIA

Listas

Sequências de objetos de diferentes tipos

```
> lista <- list(notas, m, TRUE)
> lista
[[1]]
[1] 1 3 7

[[2]]
     [,1] [,2]
[1,]    1    2
[2,]    3    8
[3,]    7    6

[[3]]
[1] TRUE
```

list {base} R Documentation

Lists – Generic and Dotted Pairs

Description

Functions to construct, coerce and check for both kinds of R lists.

Listas: acesso pelo índice

```
lista        List of 3
 : num [1:3] 1 3 7
 : num [1:3, 1:2] 1 3 7 2 8 6
 : logi TRUE
```

Fatia de uma lista:

```
> lista[3]
[[1]]
[1] TRUE

> class(lista[3])
[1] "list"
```

Referência a um membro da lista:

```
> lista[[3]]
[1] TRUE
> class(lista[[3]])
[1] "logical"
> lista[[1]][2]
[1] 3
```

Listas: acesso pelo nome

```
> lista2 <- list(p1=notas,p2=maisNotas)
> lista2["p1"]
$p1
[1] 1 3 7
```

Fatia (lista)

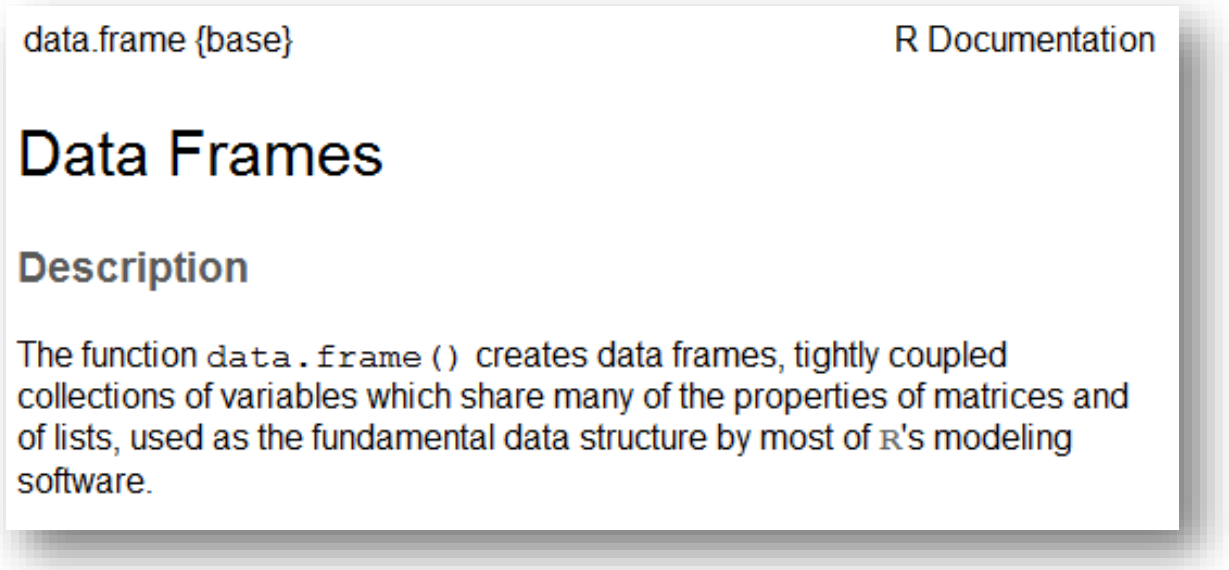
```
> lista2[["p1"]]
[1] 1 3 7
> lista2$p1
[1] 1 3 7
```

Membro da lista

Data frames

Representa uma tabela de dados

Uma lista de vetores de igual tamanho



data.frame {base} R Documentation

Data Frames

Description

The function `data.frame()` creates data frames, tightly coupled collections of variables which share many of the properties of matrices and of lists, used as the fundamental data structure by most of R's modeling software.

Data frames

mtcars: data frame de exemplo, incluído na base de R, com 32 observações (linhas) com 11 variáveis (colunas)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0		
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0		
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0		
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0		
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0		

```
> nrow(mtcars)
[1] 32
> ncol(mtcars)
[1] 11
```

mtcars {datasets} R Documentation

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Data frames: acesso a elementos

	mpg	cyl	disp	hp	drat	wt	qsec
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61

```
> mtcars[1,6]
[1] 2.62
> mtcars["Datsun 710", "hp"]
[1] 93
> nrow(mtcars)
[1] 32
> ncol(mtcars)
[1] 11
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1
```

head {utils} R Documentation

Return the First or Last Part of an Object

Description

Returns the first or last parts of a vector, matrix, table, data frame or function.

Data frames: acesso a colunas

Fatia (data frame):

```
> class(mtcars[1])
[1] "data.frame"
> head(mtcars[1])
                   mpg
Mazda RX4         21.0
Mazda RX4 Wag     21.0
Datsun 710        22.8
Hornet 4 Drive    21.4
Hornet Sportabout 18.7
Valiant           18.1
```

```
> class(mtcars["mpg"])
[1] "data.frame"
```

Vetor:

```
> class(mtcars[[1]])
[1] "numeric"
> head(mtcars[[1]])
[1] 21.0 21.0 22.8 21.4 18.7 18.1
> head(mtcars[,1])
[1] 21.0 21.0 22.8 21.4 18.7 18.1
> head(mtcars[["mpg"]])
[1] 21.0 21.0 22.8 21.4 18.7 18.1
> head(mtcars$mpg)
[1] 21.0 21.0 22.8 21.4 18.7 18.1
```

FACULDADE DE TECNOLOGIA

Data frames: acesso a linhas

```
> mtcars[9,]
          mpg cyl  disp hp drat   wt qsec vs am gear carb
Merc 230 22.8   4 140.8 95 3.92 3.15 22.9  1  0    4    2
> mtcars["Merc 230",]
          mpg cyl  disp hp drat   wt qsec vs am gear carb
Merc 230 22.8   4 140.8 95 3.92 3.15 22.9  1  0    4    2
> class(mtcars[9,])
[1] "data.frame"
```

```
> mtcars[c(3,8,9),]
           mpg cyl  disp hp drat   wt  qsec vs am gear carb
Datsun 710 22.8   4 108.0 93 3.85 2.32 18.61  1  1    4    1
Merc 240D  24.4   4 146.7 62 3.69 3.19 20.00  1  0    4    2
Merc 230   22.8   4 140.8 95 3.92 3.15 22.90  1  0    4    2
> head(mtcars[mtcars$am==1,])
               mpg cyl  disp  hp drat    wt  qsec vs am gear
Mazda RX4     21.0   6 160.0 110 3.90 2.620 16.46  0  1    4
Mazda RX4 Wag 21.0   6 160.0 110 3.90 2.875 17.02  0  1    4
Datsun 710    22.8   4 108.0  93 3.85 2.320 18.61  1  1    4
Fiat 128      32.4   4  78.7  66 4.08 2.200 19.47  1  1    4
```

Importação de dados tabulados

Prova1.csv

```
1  Coluna;Carteira;Entrega;Tempo;Q1;Q2;Q3;Q4;Q5;Q6;Q7;Q8;Q9;Q10;Q11;Q12;Q13;Q14;Q15;Q16;Q17;Q18;Q19;
2  5;2;20:53;01:41;E;E;M;M;E;C;C;C;C;M;E;M;M;C;E;C;C;M;E;C;E;C;E;M;C;C;C;C;E;C;E;E;M;C;E;E;M;E;M;E
3  6;1;19:41;00:29;M;M;M;C;E;E;E;C;E;M;M;E;E;M;C;E;E;M;M;M;E;C;C;E;E;M;E;E;E;M;E;E;E;E;E;E;C;C;M;E
4  1;2;20:20;01:08;E;E;E;C;E;E;E;E;C;C;E;E;C;E;E;E;C;M;E;E;C;C;C;E;C;C;C;E;C;E;E;E;E;C;E;E;E;C;C;E
5  4;1;19:43;00:31;C;C;C;C;C;C;E;M;C;M;M;C;M;E;C;C;M;M;C;C;E;C;C;E;C;M;C;C;M;M;M;M;E;M;E;C;E;C;M;C
6  6;3;20:15;01:03;E;C;E;C;E;E;E;C;E;C;C;C;C;E;C;E;M;C;E;E;E;C;C;C;C;C;C;C;B;E;E;E;C;E;E;E;C;C;E
7  2;2;20:03;00:51;C;C;C;C;M;C;E;E;C;M;M;C;C;M;C;M;C;E;C;E;M;C;C;E;E;C;C;C;C;C;E;C;C;C;E;M;C;C;E;M
```

read.table {utils} R Documentation

Data Input

Description

Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

FACULDADE DE TECNOLOGIA

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
           row.names, col.names, as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(),
           fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)

read.csv(file, header = TRUE, sep = ",", quote = "\"",
         dec = ".", fill = TRUE, comment.char = "", ...)

read.csv2(file, header = TRUE, sep = ";", quote = "\"",
          dec = ",", fill = TRUE, comment.char = "", ...)
```

FACULDADE DE TECNOLOGIA

getwd {base}

Get or Set Working Directory

```
> setwd("C:/Users/pc/OneDrive/Apresentações/Minicurso R/Recursos")
> notasP1 <- read.csv2("Prova1.csv")
```

Prova1.csv

```
1  Coluna;Carteira;Entrega;Tempo;Q1;Q2;Q3;Q4;Q5;Q6;Q7;Q8;Q9;Q10;
2  5;2;20:53;01:41;E;E;M;M;E;C;C;C;C;M;E;M;M;C;E;C;C;M;E;C;E;C;E
3  6;1;19:41;00:29;M;M;M;C;E;E;E;C;E;M;M;E;E;M;C;E;E;M;M;M;E;C;C
4  1;2;20:20;01:08;E;E;E;C;E;E;E;E;C;C;E;E;C;E;E;E;C;M;E;E;C;C;C
5  4;1;19:43;00:31;C;C;C;C;C;C;E;M;C;M;M;C;M;E;C;C;M;M;C;C;E;C;C
```

notasP1 30 obs. of 44 variables

	Coluna	Carteira	Entrega	Tempo	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	5	2	20:53	01:41	E	E	M	M	E	C	C	C	C	M
2	6	1	19:41	00:29	M	M	M	C	E	E	E	C	E	M
3	1	2	20:20	01:08	E	E	E	C	E	E	E	E	C	C
4	4	1	19:43	00:31	C	C	C	C	C	C	E	M	C	M

FACULDADE DE TECNOLOGIA

Sumário: dados em R

Variáveis não precisam ser declaradas

Tipos de dados básicos

- Numérico (real), inteiro, complexo, lógico, caractere, data
- Vetor, fator, matriz, lista, data frame

Valores especiais

- Inf, NaN, NA

Operadores lógicos

- >, >=, <, <=, ==, !=, & (and), | (or), ! (not)

Sumário: algumas funções

sin(), cos(), tan(),..., log(), exp(),..., sqrt()

class(), is.integer(), as.integer(), factor(), matrix(), list(),...

Sys.Date(), Sys.time(), as.Date()

c(), table(), rep(), nrow(), ncol()

t(), cbind(), rbind(), solve()

head(), tail()

setwd(), getwd()

read.table(), read.csv(), read.csv2()

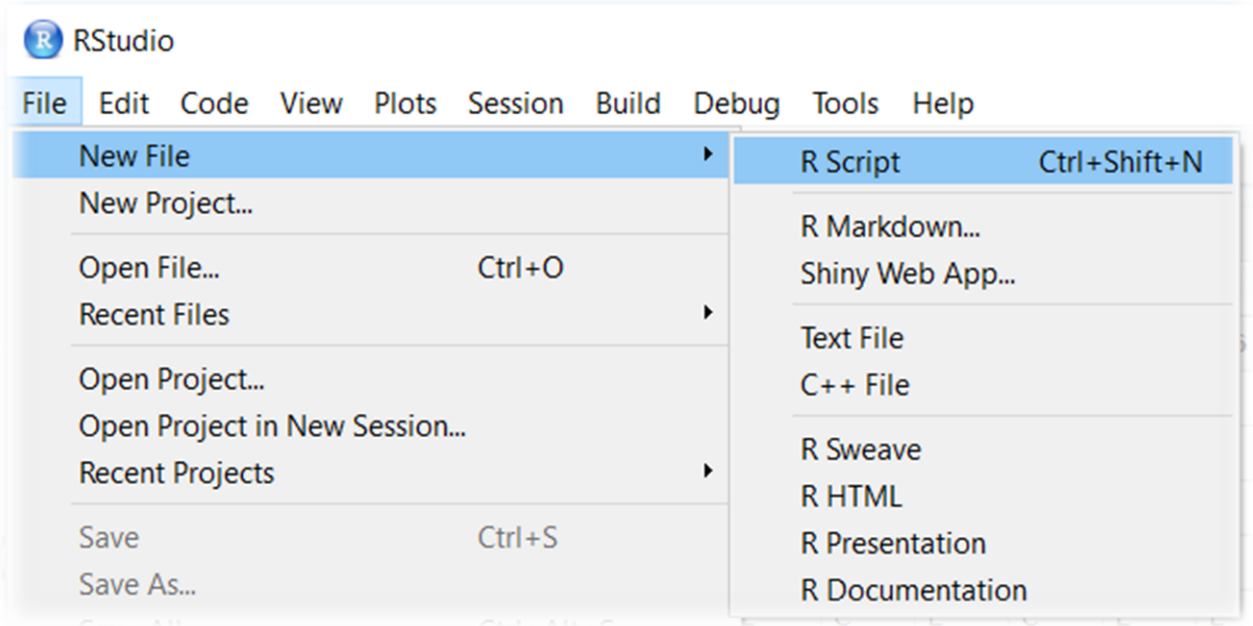
FACULDADE DE TECNOLOGIA

Além do básico: pacotes

FACULDADE DE TECNOLOGIA

Scripts

Uma sequência de comandos R pode ser editada em um arquivo e executada posteriormente



Scripts: sequências de comandos

script1.R

Source on Save | Run

```
1  setwd("C:/Users/pc/OneDrive/Apresentações/Minicurso R/Recursos")
2  notasP1 <- read.csv2("Prova1.csv")
3  nlinhas <- nrow(notasP1)
4  ncolunas <- ncol(notasP1)
5  paste(nlinhas, "x", ncolunas)
6
```

Console C:/Users/pc/OneDrive/Apresentações/Minicurso R/Recursos/

```
> source("script1.R", echo = TRUE)

> setwd("C:/Users/pc/OneDrive/Apresentações/Minicurso R/Recursos")

> notasP1 <- read.csv2("Prova1.csv")

> nlinhas <- nrow(notasP1)

> ncolunas <- ncol(notasP1)

> paste(nlinhas, "x", ncolunas)
[1] "30 x 44"
```

source {base} R Documentation

Read R Code from a File or a Connection

Description

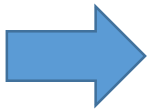
`source` causes `R` to accept its input from the named file or URL or connection. Input is read and `parse`d from that file until the end of the file is reached, then the parsed expressions are evaluated sequentially in the chosen environment.

Scripts: controle de fluxo de execução

```
while (mesN <= ultimoMes) {
    comentMes <- resComent %>%
        filter(mes == mesN, ano = anoN) %>%
        select(record_id, iam_comentario) %>%
        merge(idprof) %>%
        select(profissao, comentario = iam_comentario)
    if (nrow(comentMes) > 0) {
        filename <-
            paste("coment",
                  substr(resumoId, 1, 10),
                  anoN,
                  sprintf("%02d", mesN),
                  sep = "_")
        fileconn <- file(paste0(filename, ".txt"))
```

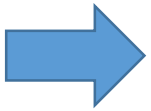
Controle de fluxo: seleção

```
if (a > 10) {
    valor <- 0.9 * a
} else {
    valor <- 1.1 * a
}
```



```
> a <- 5
> if (a > 10) {
+     valor <- 0.9 * a
+ } else {
+     valor <- 1.1 * a
+ }
> valor
[1] 5.5
> a <- 11
> if (a > 10) {
+     valor <- 0.9 * a
+ } else {
+     valor <- 1.1 * a
+ }
> valor
[1] 9.9
```

```
valor <- ifelse(a >10, 0.9*a, 1.1*a)
```



```
> a <- 5
> valor <- ifelse(a >10, 0.9*a, 1.1*a)
> valor
[1] 5.5
>
>
> a <- 11
> valor <- ifelse(a >10, 0.9*a, 1.1*a)
> valor
[1] 9.9
```

Controle de fluxo: repetição

```
n <- 10
f <- c(rep(1, n))
i <- 3
while (i <= n) {
    f[i] <- f[i-2] + f[i-1]
    i <- i + 1
}
```

```
> f
 [1]  1  1  2  3  5  8 13 21 34 55
```

```
for (valor in f) print(valor)
```

```
> for (valor in f) print(valor)
[1] 1
[1] 1
[1] 2
[1] 3
[1] 5
[1] 8
[1] 13
[1] 21
[1] 34
[1] 55
```

Funções

R possibilita que usuários escrevam suas próprias funções

```
myfunction <- function(arg1, arg2, ... ){
statements
return(object)
}
```

```
myObj <- myfunction(param1, param2,...)
```

FACULDADE DE TECNOLOGIA

```r
#' Verifica dígitos de um CPF
#'
#' @param cpf um CPF no formato ###.###.###-##
#' @return TRUE se digitos verificadores estao coerentes com CPF
#' @author Ivan L M Ricarte
verificaCPF <- function(cpf) {
    # digitos verificadores declarados
    d1 <- as.integer(substr(cpf,13,13))
    d2 <- as.integer(substr(cpf,14,14))
    # cálculo dos dígitos verificadores
    rem1 <- (10*as.integer(substr(cpf,1,1)) +
        9*as.integer(substr(cpf,2,2)) +
        8*as.integer(substr(cpf,3,3)) +
        7*as.integer(substr(cpf,5,5)) +
        6*as.integer(substr(cpf,6,6)) +
        5*as.integer(substr(cpf,7,7)) +
        4*as.integer(substr(cpf,9,9)) +
        3*as.integer(substr(cpf,10,10)) +
        2*as.integer(substr(cpf,11,11))) %% 11
    vf1 <- ifelse(rem1 < 2,0,11-rem1)
    rem2 <- (11*as.integer(substr(cpf,1,1)) +
                10*as.integer(substr(cpf,2,2)) +
                9*as.integer(substr(cpf,3,3)) +
                8*as.integer(substr(cpf,5,5)) +
                7*as.integer(substr(cpf,6,6)) +
                6*as.integer(substr(cpf,7,7)) +
                5*as.integer(substr(cpf,9,9)) +
                4*as.integer(substr(cpf,10,10)) +
                3*as.integer(substr(cpf,11,11)) +
                2*vf1) %% 11
    vf2 <- ifelse(rem2 < 2,0,11-rem2)
    # declarados iguais aos calculados?
    d1 == vf1 & d2 == vf2
}
```

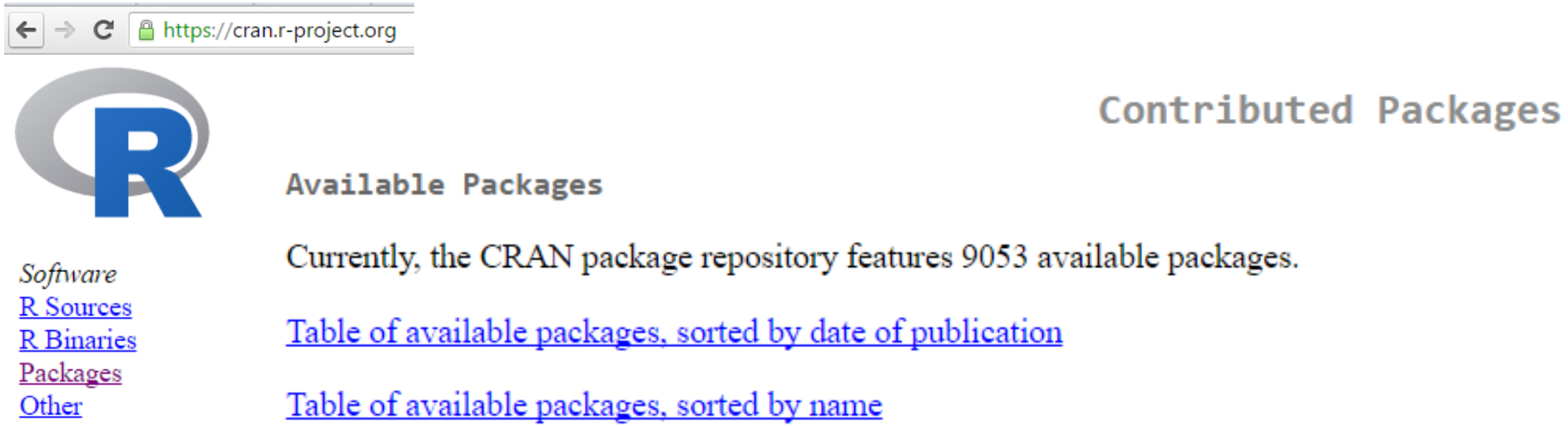
```
> verificaCPF("123.456.789-00")
[1] FALSE

> verificaCPF("123.456.789-09")
[1] TRUE
```

Pacotes

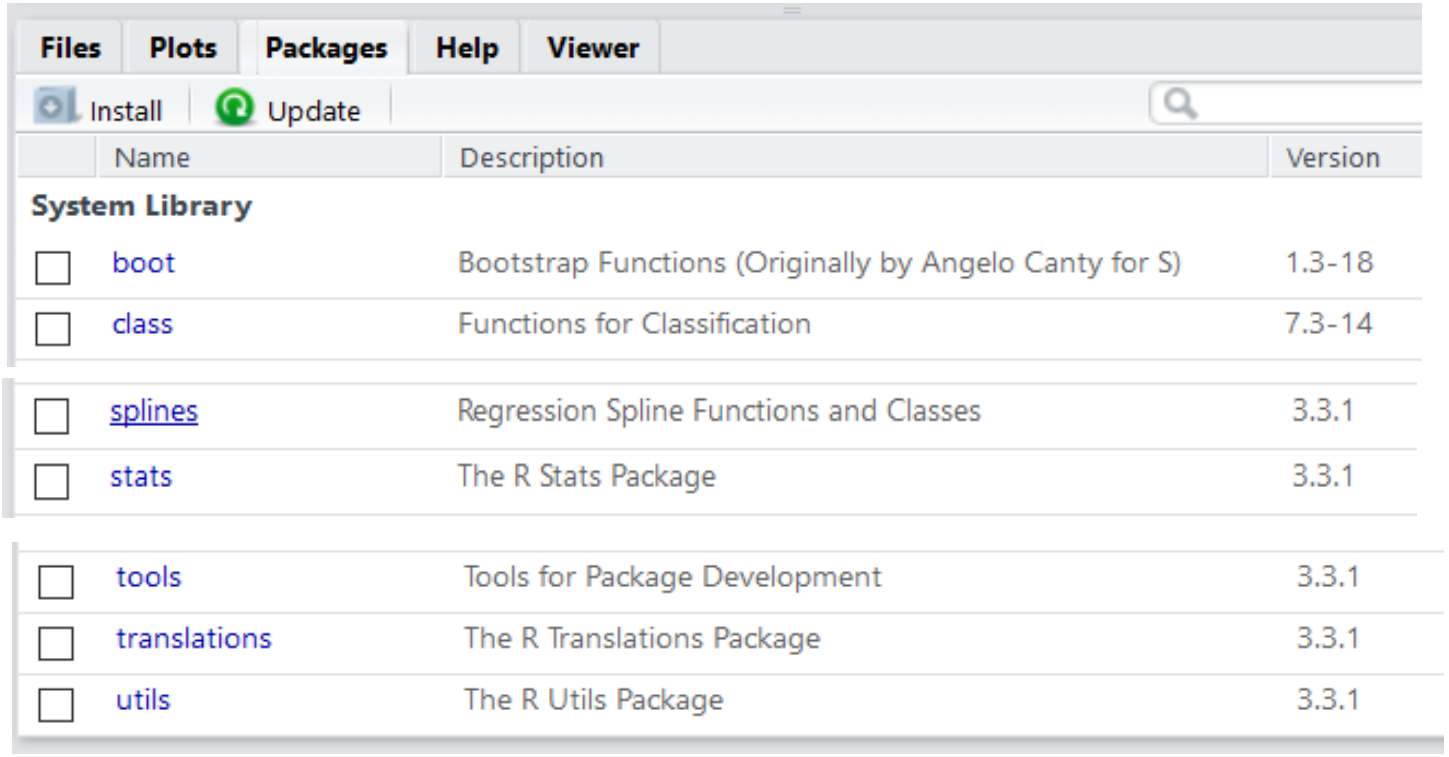
Funções e dados podem ser combinados em pacotes R

Milhares de pacotes estão disponíveis no Repositório R (CRAN)



Pacotes do sistema

Alguns pacotes já fazem parte do núcleo do ambiente de execução R



Documentação do pacote

```
>
> library(help = stats)
```


```
                Information on package 'stats'

Descrição:

Package:        stats
Version:        3.3.1
Priority:       base
Title:          The R Stats Package
Author:         R Core Team and contributors worldwide
Maintainer:     R Core Team <R-core@r-project.org>
Description:    R statistical functions.
License:        Part of R 3.3.1
Imports:        utils, grDevices, graphics
Suggests:       MASS, Matrix, SuppDists, methods, stats4
Built:          R 3.3.1; x86_64-w64-mingw32; 2016-06-21
                12:41:51 UTC; windows
```

Documentação do pacote e suas funções

Em RStudio, com hiperligações

The R Stats Package

Documentation for package ‘stats’ version 3.3.1

- DESCRIPTION file.
- Code demos. Use demo() to run them.

Help Pages

A B C D E F G H I K L M N O P Q R

stats-package The R Stats Package

lm {stats} R Documentation

Fitting Linear Models

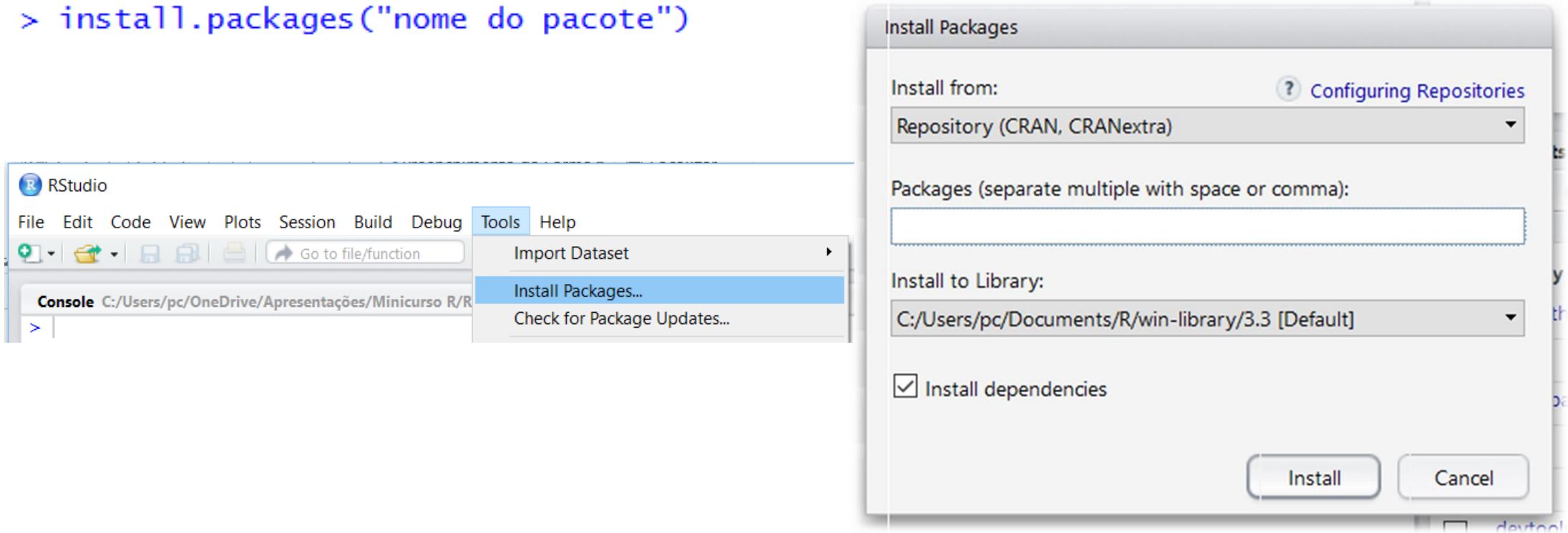
Description

`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although `aov` may provide a more convenient interface for these).

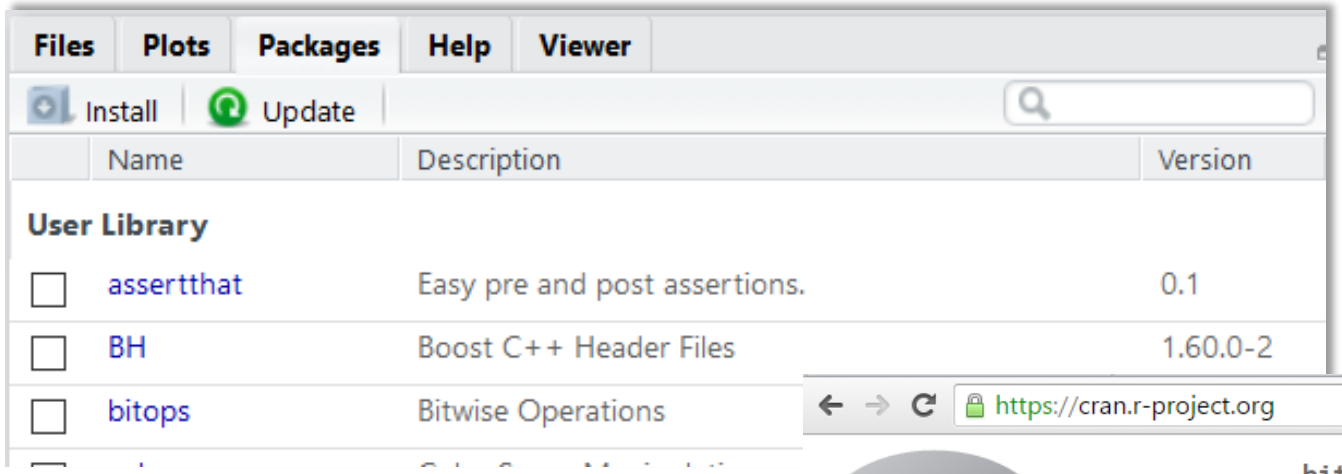
Pacotes de usuários

Outros pacotes (“de usuários”) precisam ser instalados

```
>
> install.packages("nome do pacote")
```



Documentação de pacotes de usuários



Em RStudio

No repositório CRAN

Ativação de pacotes de usuários

Por padrão, funções de pacotes instalados não estão disponíveis

É preciso carregar as funções no ambiente de execução

```
>
> library(nome do pacote)
```

library {base} R Documentation

Loading/Attaching and Listing of Packages

Description

`library` and `require` load and attach add-on packages.

Pacotes em desenvolvimento

Package devtools

install	Install a local development package.
install_bioc	Install a package from a Bioconductor repository
install_bitbucket	Install a package directly from bitbucket
install_cran	Attempts to install a package from CRAN.
install_deps	Install package dependencies if needed.
install_dev_deps	Install package dependencies if needed.
install_git	Install a package from a git repository
install_github	Attempts to install a package directly from GitHub.
install_local	Install a package from a local file
install_svn	Install a package from a SVN repository
install_url	Install a package from a url
install_version	Install specified version of a CRAN package.

FACULDADE DE TECNOLOGIA

No caso de Windows, pode ser necessário instalar as ferramentas auxiliares:

Building R for Windows

This document is a collection of resources for building packages for R under Microsoft Windows, or for building R itself (version 1.9.0 or later). The original collection was put together by Prof. Brian Ripley; it is currently being maintained by Duncan Murdoch.

The authoritative source of information for tools to work with the current release of R is the "R Administration and Installation" manual. In particular, please read the "Windows Toolset" appendix.

Rtools Downloads

Some of the tools are incompatible with obsolete versions of R. We maintain one actively updated version of the tools, and other "frozen" snapshots of them. We recommend that users use the latest release of Rtools with the latest release of R.

The current version of this file is recorded here: VERSION.txt.

Download	R compatibility	Frozen?
Rtools34.exe	R 3.3.x and later	No

FACULDADE DE TECNOLOGIA

Sumário: pacotes

Usuários podem definir scripts e funções em R

Conjunto de funções podem ser organizadas em pacotes

Pacotes com contribuições de usuários podem ser carregados no ambiente de execução

- Inclusive pacotes ainda em desenvolvimento

Algumas funções:

- source(), function()
- ifelse()
- install.packages(), library(), require()

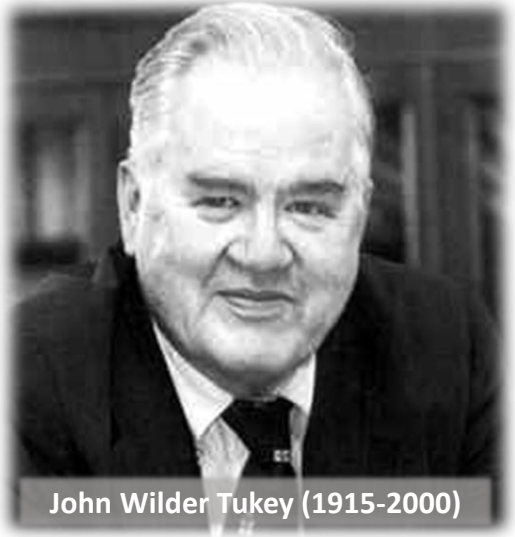
FACULDADE DE TECNOLOGIA

Princípios da análise de dados exploratória

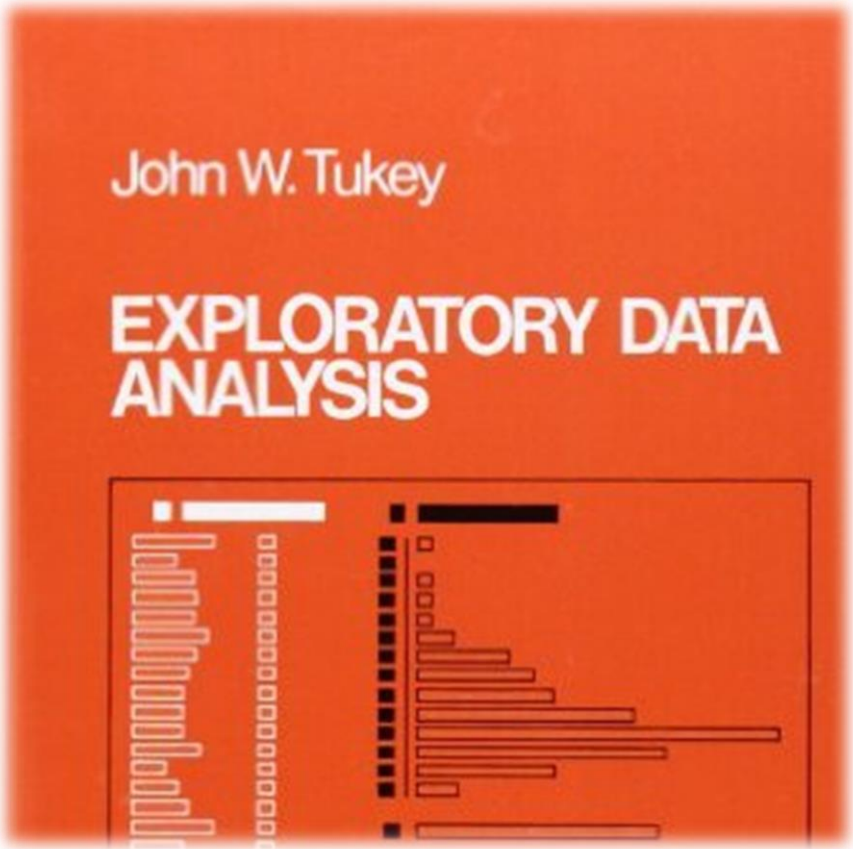
Análise de Dados Exploratória

Abordagem para a análise inicial de um conjunto de dados

- Identificar suas principais características
- Em geral, apoiada por métodos visuais



John Wilder Tukey (1915-2000)



Conjuntos de dados de exemplos

Serão utilizados alguns conjuntos de dados do pacote UsingR

☐	UsingR	Data Sets, Etc. for the Text "Using R for Introductory Statistics", Second Edition	2.0-5

```
> install.packages("UsingR")
```

```
> library(UsingR)
Carregando pacotes exigidos: MASS
Carregando pacotes exigidos: HistData
Carregando pacotes exigidos: Hmisc
```

Altura de pai e de filho

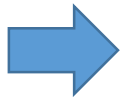
father.son {UsingR} R Documentation

Pearson's data set on heights of fathers and their sons

Description

1078 measurements of a father's height and his son's height.

```
> str(father.son)
'data.frame':   1078 obs. of  2 variables:
 $ fheight: num  65 63.3 65 65.8 61.1 ...
 $ sheight: num  59.8 63.2 63.3 62.8 64.3 ...
```



fheight

Father's height in inches

sheight

Son's height in inches

Ajustes iniciais

Na maior parte dos casos, pequenos ajustes são necessários para facilitar a inteligibilidade de dados e variáveis

```
# conversão de polegadas para centímetros
pai.filho <- 2.54 * father.son
# nomes das variáveis
names(pai.filho) <- c("p_altura", "f_altura")
```

names {base}

The Names of an Object

Description

Functions to get or set the names of an object.

```
> head(pai.filho)
  p_altura f_altura
1 165.2232 151.8368
2 160.6574 160.5637
3 164.9865 160.8897
4 167.0113 159.4926
5 155.2886 163.2741
6 160.0773 163.1752
>
```

Explorações iniciais: summary

```
> summary(pai.filho)
    p_altura        f_altura
 Min.   :149.9   Min.   :148.6
 1st Qu.:167.1   1st Qu.:170.0
 Median :172.1   Median :174.3
 Mean   :171.9   Mean   :174.5
 3rd Qu.:176.8   3rd Qu.:179.0
 Max.   :191.6   Max.   :199.0
>
```

summary {base} R Documentation

Object Summaries

Description

`summary` is a generic function used to produce result summaries of the results of various model fitting functions. The function invokes particular `methods` which depend on the `class` of the first argument.

Explorações iniciais: amostragem

sample {base} R Documentation

Random Samples and Permutations

Description

`sample` takes a sample of the specified size from the elements of `x` using either with or without replacement.

```
> sample(pai.filho$p_altura, 20)
 [1] 163.9947 177.4675 184.4058 169.8413
 [5] 156.8836 177.1446 176.5810 179.0977
 [9] 154.3799 172.9178 163.3682 174.5809
[13] 179.1806 166.6307 175.9183 184.3264
[17] 164.4345 173.3086 172.0276 172.5568
>
```

Funções estatísticas descritivas

```
> min(pai.filho$p_altura)
[1] 149.8803
> max(pai.filho$p_altura)
[1] 191.6022
>
> mean(pai.filho$p_altura)
[1] 171.9252
> median(pai.filho$p_altura)
[1] 172.1272
>
> quantile(pai.filho$p_altura, 1/4)
     25%
167.1008
> quantile(pai.filho$p_altura, 3/4)
     75%
176.7916
>
```

Extremes {base}

Maxima and Minima

mean {base}

Arithmetic Mean

median {stats}

Median Value

quantile {stats}

Sample Quantiles

FACULDADE DE TECNOLOGIA

mean {base} R Documentation

Arithmetic Mean

Description

Generic function for the (trimmed) arithmetic mean.

Usage

```
mean(x, ...)

## Default S3 method:
mean(x, trim = 0, na.rm = FALSE, ...)
```

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

```
> mean(c(1,2,NA))
[1] NA
> mean(c(1,2,NA), na.rm = TRUE)
[1] 1.5
```

FACULDADE DE TECNOLOGIA

Gráficos básicos: hist

Histograms

Description

The generic function `hist` computes a histogram of the given data values. If `plot = TRUE`, the resulting object of class "`histogram`" is plotted by `plot.histogram`, before it is returned.

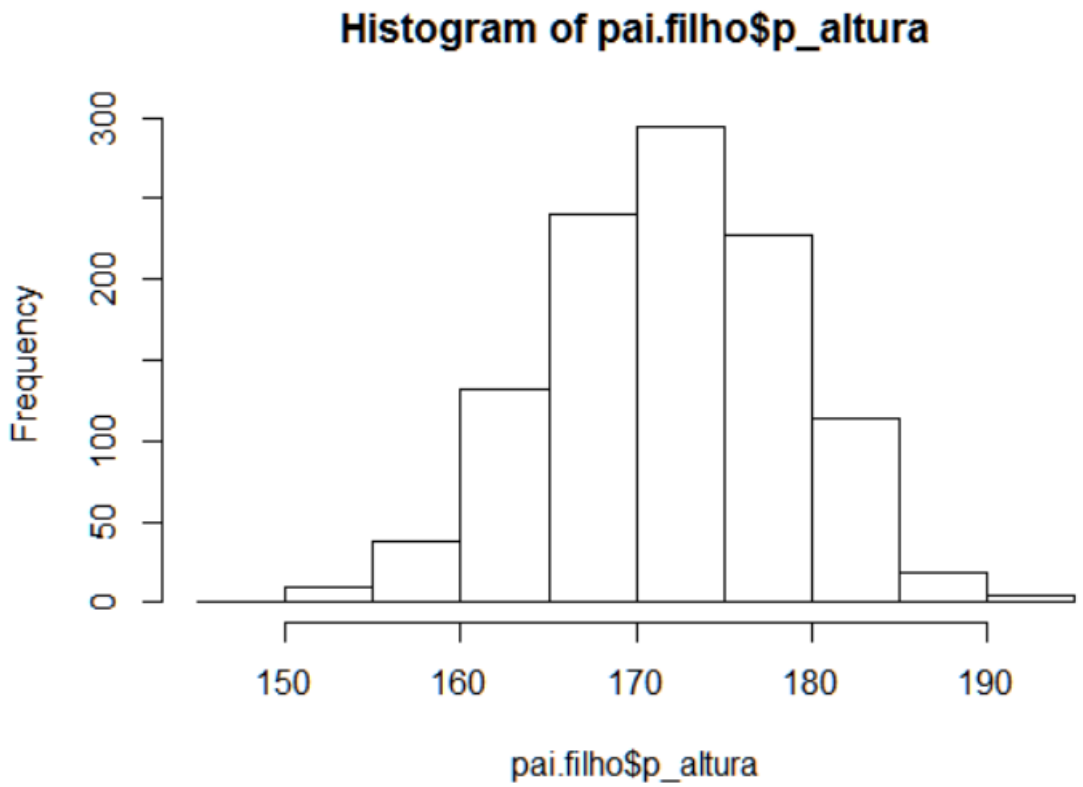
Usage

```
hist(x, ...)

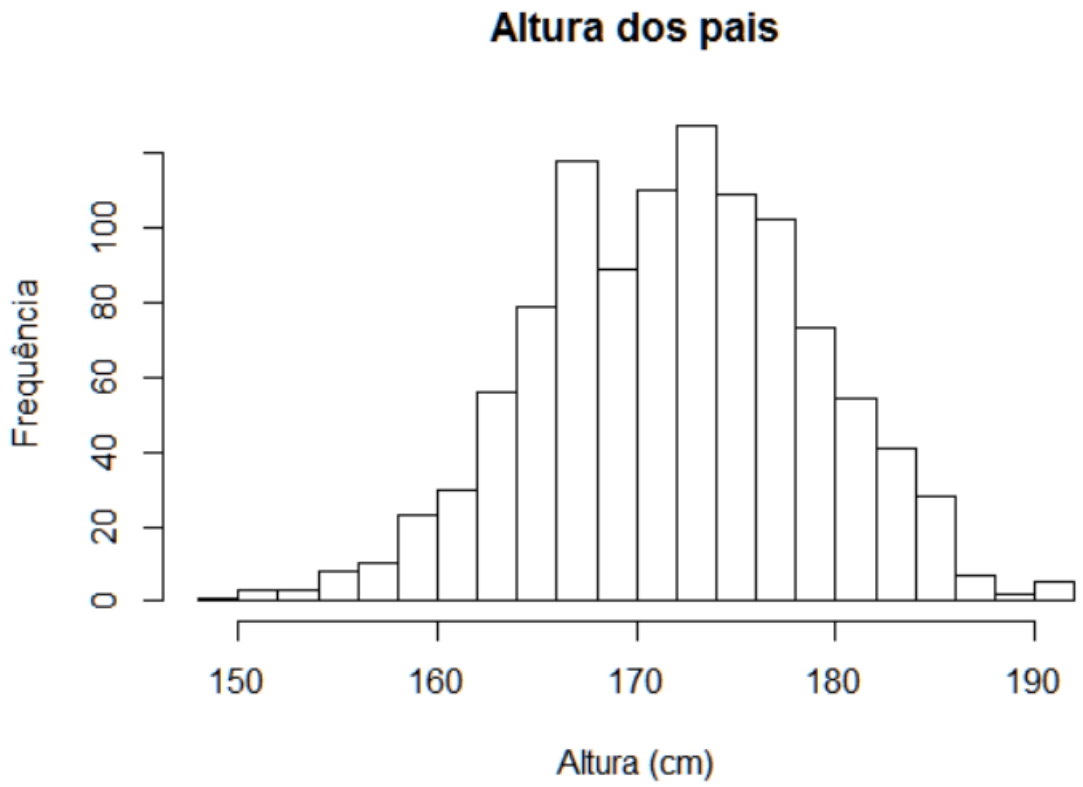
## Default S3 method:
hist(x, breaks = "Sturges",
     freq = NULL, probability = !freq,
     include.lowest = TRUE, right = TRUE,
     density = NULL, angle = 45, col = NULL, border = NULL,
     main = paste("Histogram of" , xname),
     xlim = range(breaks), ylim = NULL,
     xlab = xname, ylab,
     axes = TRUE, plot = TRUE, labels = FALSE,
     nclass = NULL, warn.unused = TRUE, ...)
```

FACULDADE DE TECNOLOGIA

```
hist(pai.filho$p_altura)
```




```
hist(pai.filho$p_altura, breaks=16,
     main="Altura dos pais",
     xlab = "Altura (cm)", ylab = "Frequência")
```



Gráficos básicos: plot

plot {graphics} R Documentation

Generic X-Y Plotting

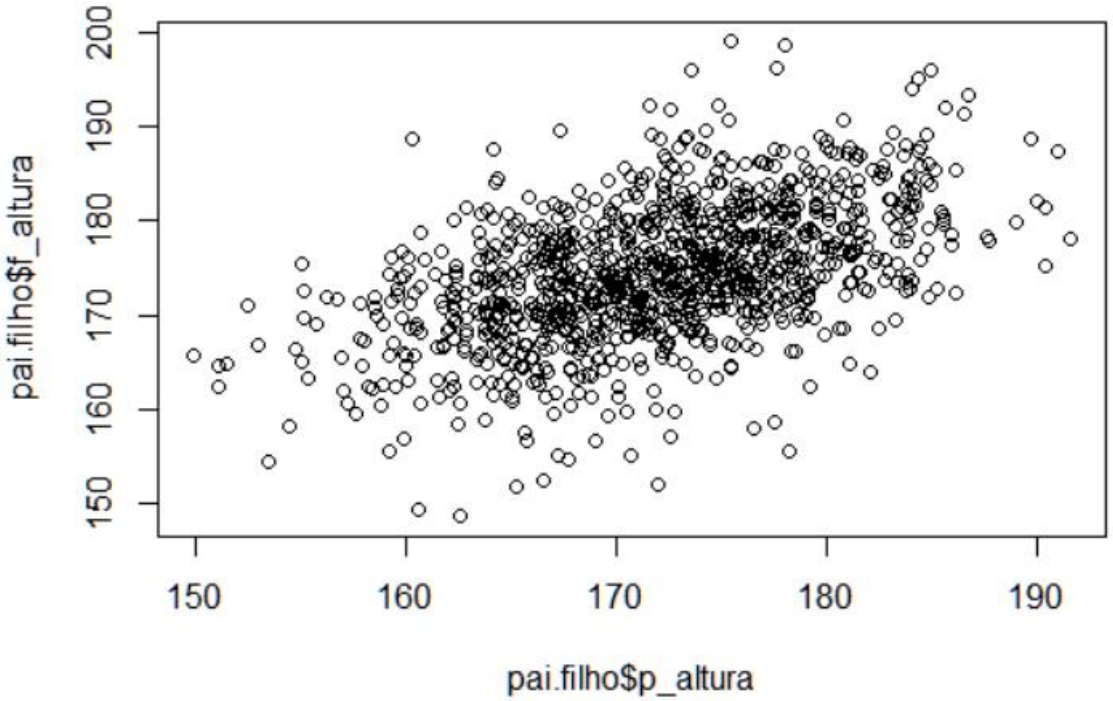
Description

Generic function for plotting of R objects. For more details about the graphical parameter arguments, see `par`.

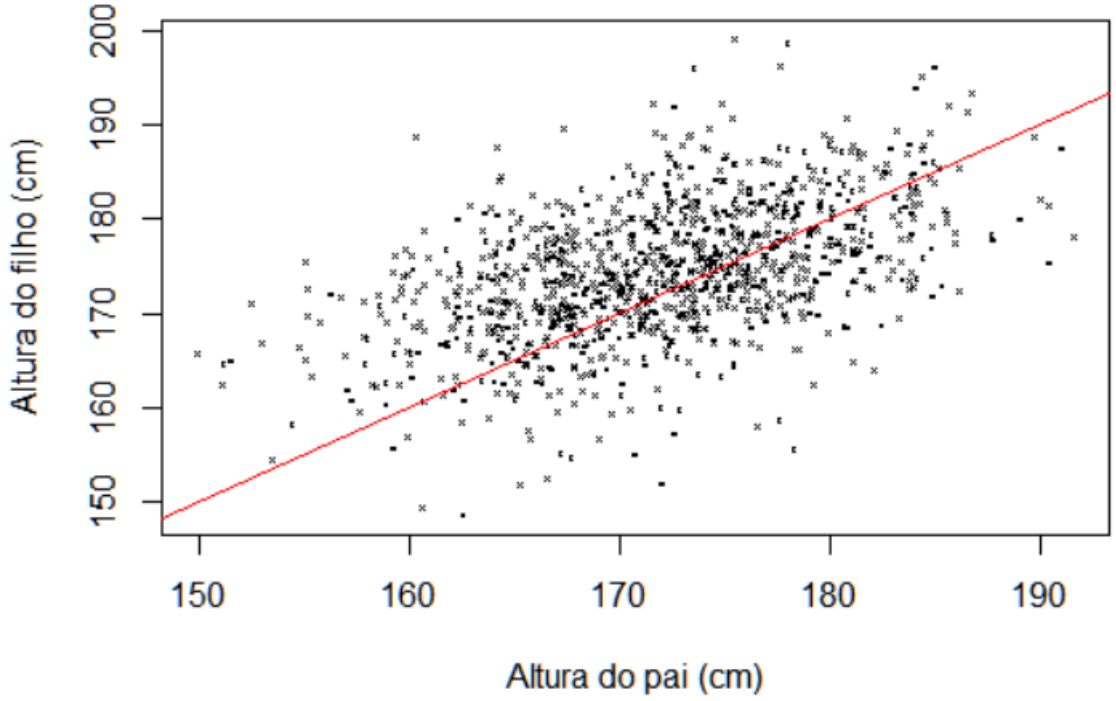
For simple scatter plots, `plot.default` will be used. However, there are `plot` methods for many R objects, including `functions`, `data.frames`, `density` objects, etc. Use `methods(plot)` and the documentation for these.

FACULDADE DE TECNOLOGIA

```
plot(pai.filho$p_altura,pai.filho$f_altura)
```




```
plot(pai.filho$p_altura, pai.filho$f_altura,
     pch = 4, cex = 0.25, xlab = "Altura do pai (cm)",
     ylab = "Altura do filho (cm)")
abline(a=0, b=1, col="red")
```



Gráficos básicos: boxplot

boxplot {graphics} R Documentation

Box Plots

Description

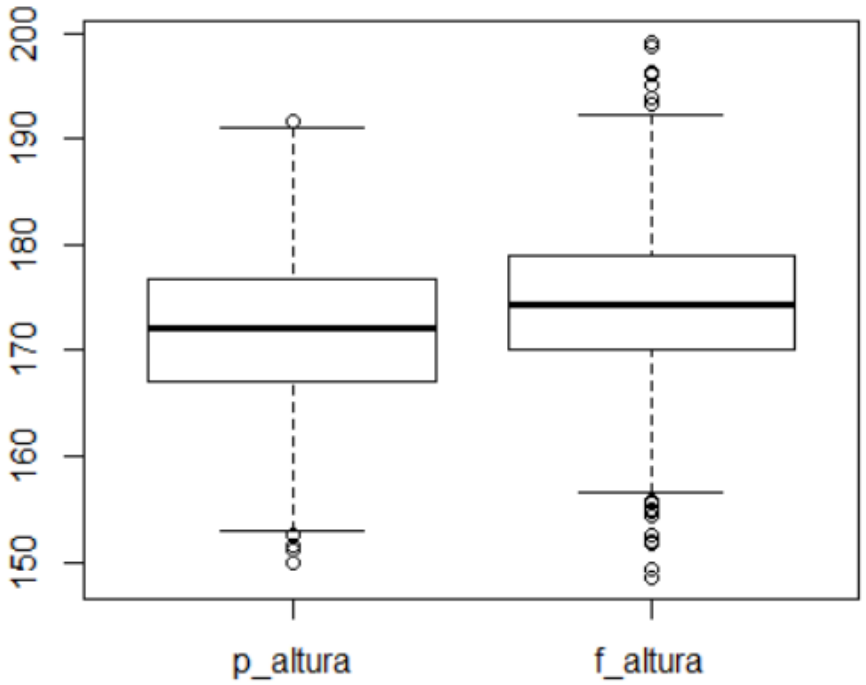
Produce box-and-whisker plot(s) of the given (grouped) values.

Usage

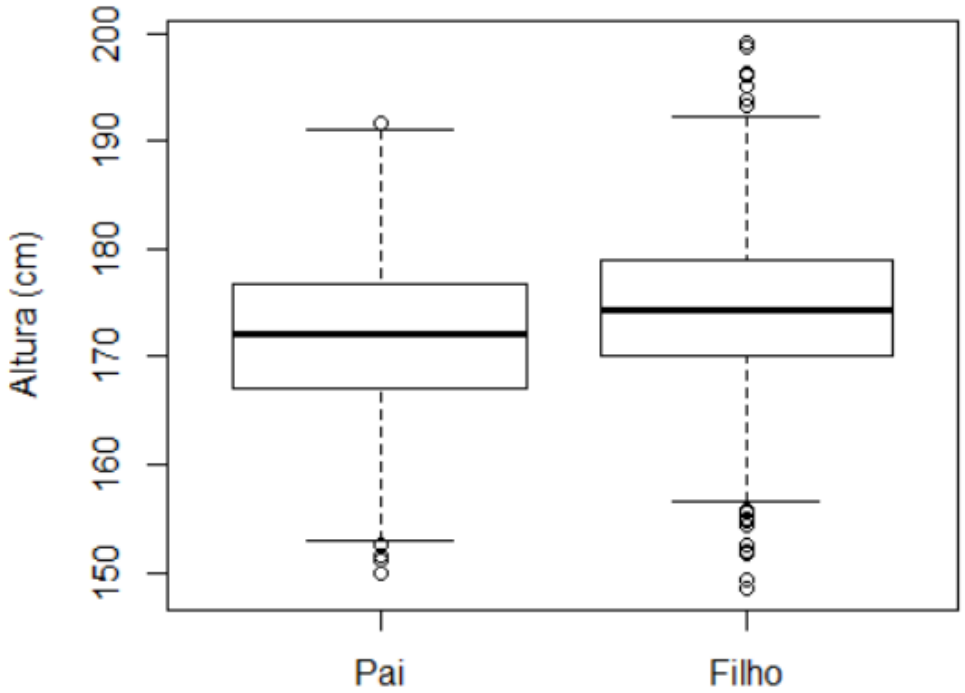
```
boxplot(x, ...)
```

FACULDADE DE TECNOLOGIA

```
boxplot(pai.filho)
```




```
boxplot(pai.filho,
        names = c("Pai", "Filho"),
        ylab = "Altura (cm)")
```



Gráficos básicos: qqnorm, qqplot

qqnorm {stats} R Documentation

Quantile-Quantile Plots

Description

`qqnorm` is a generic function the default method of which produces a normal QQ plot of the values in `y`. `qqline` adds a line to a "theoretical", by default normal, quantile-quantile plot which passes through the `probs` quantiles, by default the first and third quartiles.

`qqplot` produces a QQ plot of two datasets.

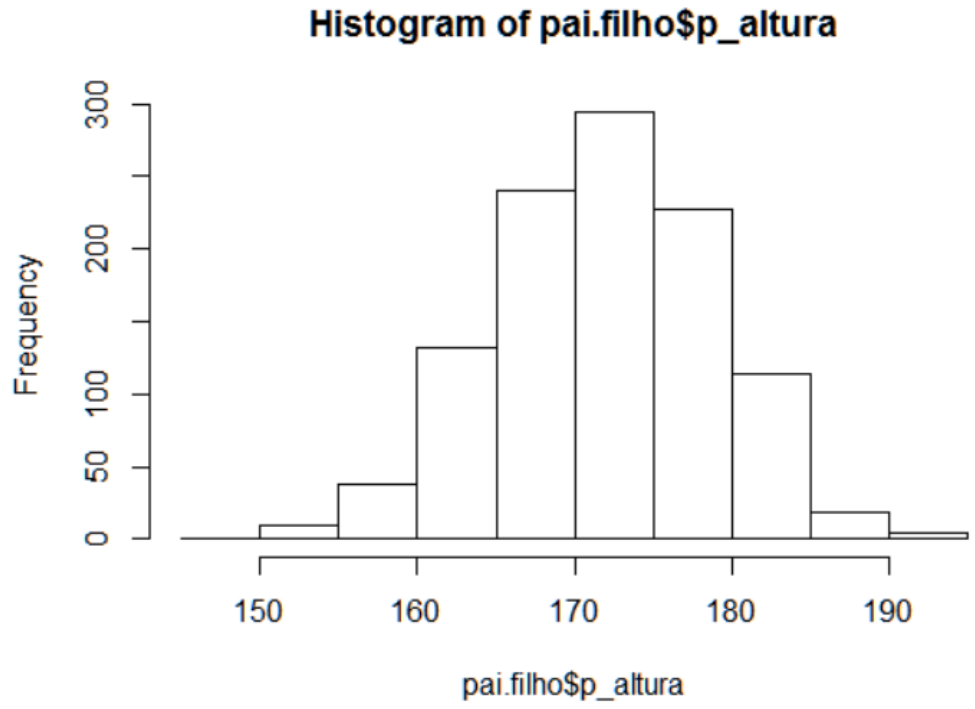
Graphical parameters may be given as arguments to `qqnorm`, `qqplot` and `qqline`.

Usage

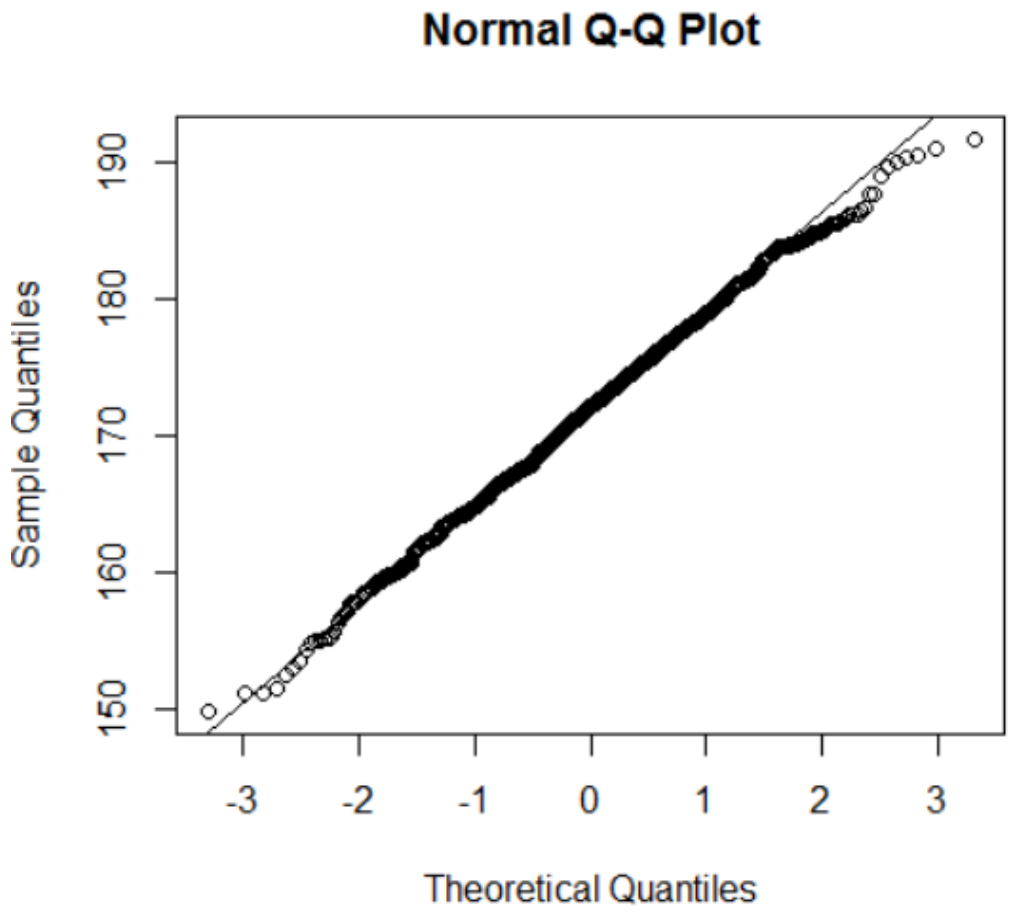
```
qqnorm(y, ...)
```

FACULDADE DE TECNOLOGIA

```
hist(pai.filho$p_altura)
```

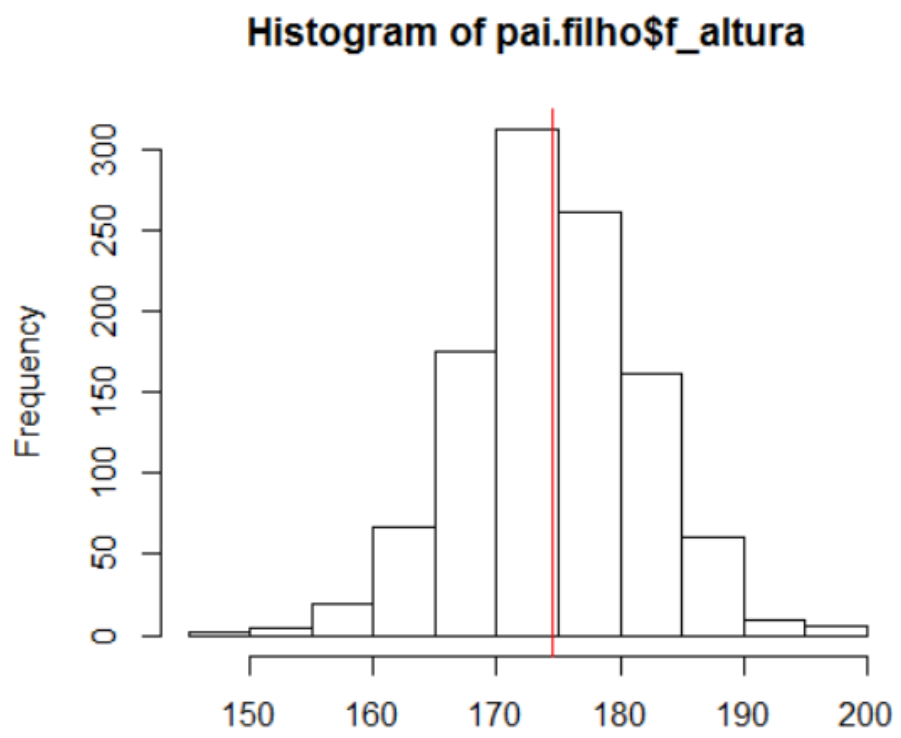



```
qqnorm(pai.filho$p_altura)
qqline(pai.filho$p_altura)
```



Estratificação

Qual a altura de um filho adulto, selecionado aleatoriamente?

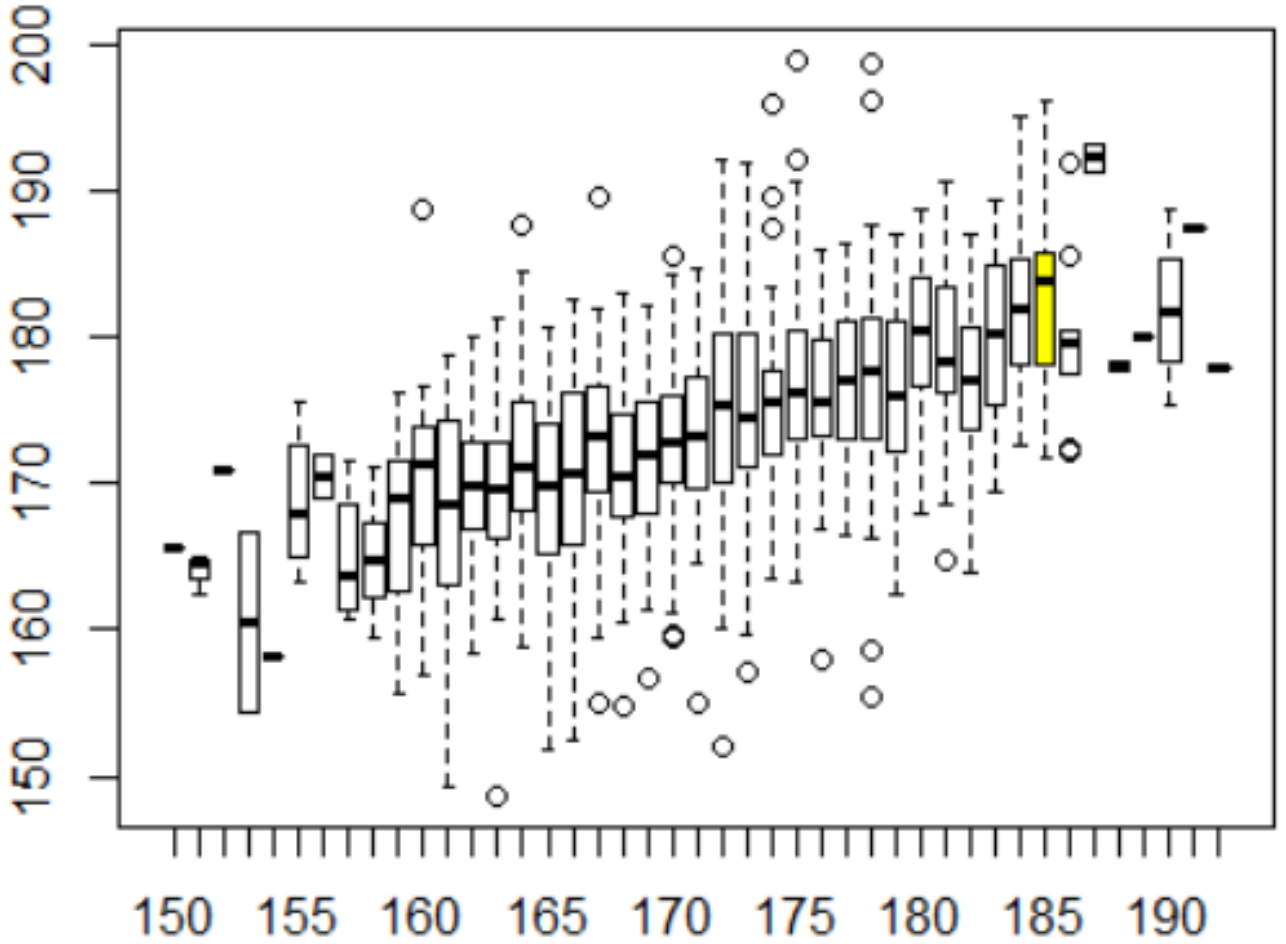



```
> round(mean(pai.filho$f_altura))
[1] 174
```

Qual a altura do filho, dado que o pai mede 185cm?

FACULDADE DE TECNOLOGIA

Altura dos filhos dada a altura dos pais



Estratificação: split

split {base} R Documentation

Divide into Groups and Reassemble

Description

`split` divides the data in the vector `x` into the groups defined by `f`. The replacement forms replace values corresponding to such a division. `unsplit` reverses the effect of `split`.

Usage

```
split(x, f, drop = FALSE, ...)
```

FACULDADE DE TECNOLOGIA

```
f <- pai.filho$f_altura
p <- pai.filho$p_altura

grupos <- split(f, round(p))
```

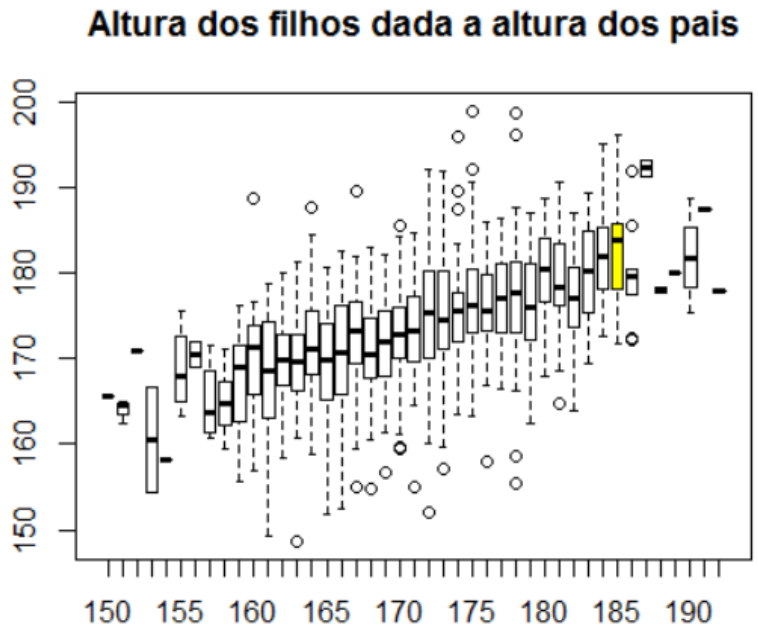
```
grupos          List of 43
 150: num 166
 151: num [1:3] 165 165 162
 152: num 171
 153: num [1:2] 154 167
 154: num 158
 155: num [1:6] 163 165 170 176 166 ...

 184: num [1:28] 181 183 183 187 173 ...
 185: num [1:11] 172 196 181 186 179 ...
 186: num [1:9] 172 180 177 179 180 ...

 190: num [1:4] 175 182 189 181
 191: num 187
 192: num 178
```

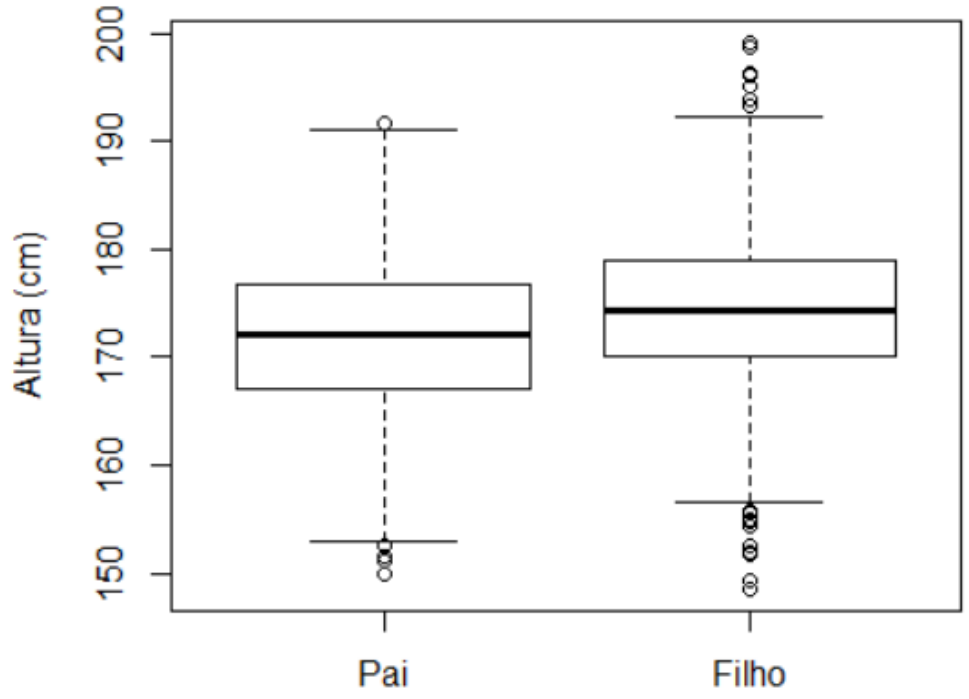
FACULDADE DE TECNOLOGIA

```
boxplot(grupos,
        col = c (rep("white",35), "yellow", rep("white", 6)),
        main = "Altura dos filhos dada a altura dos pais")
```



Altura mais provável do filho
cujo pai mede 185 cm: 182 cm

```
> round(mean(grupos$"185"))
[1] 182
```



Os vários gráficos sugerem que, na média, filhos adultos são mais altos que os pais

Dá para afirmar isso com segurança ou será fruto de um acaso?

```
> mean(pai.filho$f_altura)-mean(pai.filho$p_altura)
[1] 2.532311
```

Teste de hipótese

Welch's *t*-test

From Wikipedia, the free encyclopedia

In statistics, **Welch's *t*-test**, or **unequal variances *t*-test**, is a two-sample location test which is used to test the hypothesis that two populations have equal means. Welch's *t*-test is an adaptation of Student's *t*-test,[1] that is, it has been derived with the help of Student's *t*-test and is more reliable when the two samples have unequal variances and unequal sample sizes.[2]

Refutar hipótese nula com valor $p < \alpha$

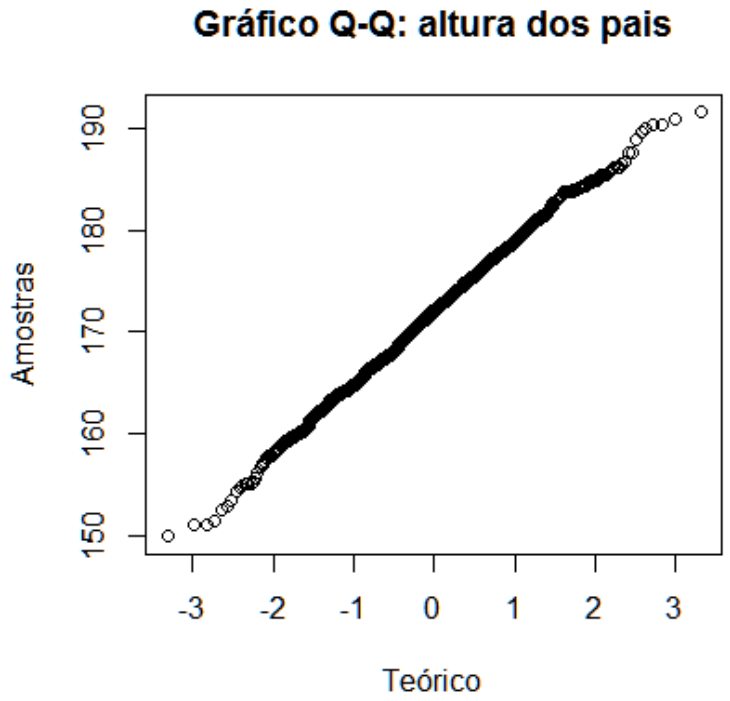
Tipicamente, $\alpha=0.05$ ou $\alpha=0.01$

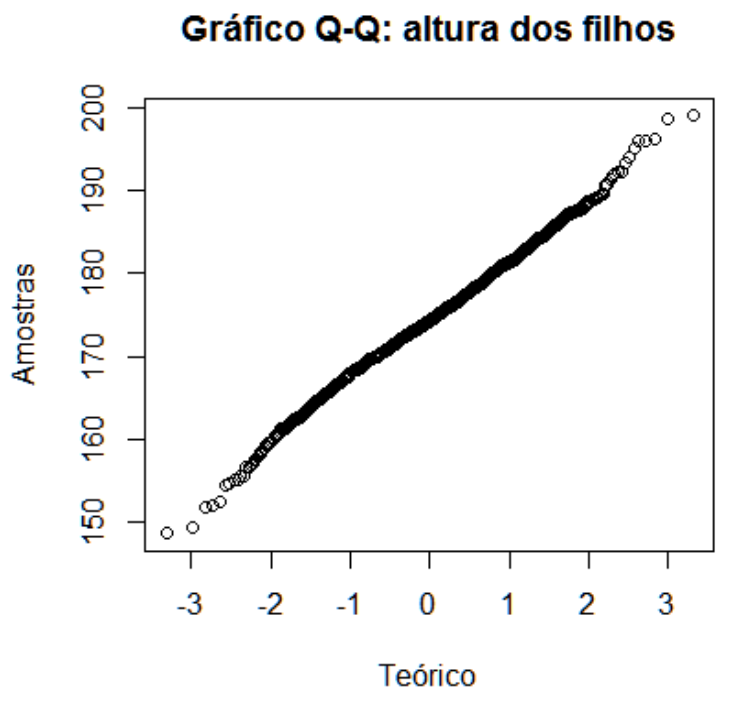
The *p*-value is defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true.[4][5]

Condições para aplicação do teste t

Distribuição das duas populações devem ser (aprox.) normais

Para o Teste de Welch, não é preciso que as variâncias sejam iguais





FACULDADE DE TECNOLOGIA

```
> t.test(pai.filho$p_altura, pai.filho$f_altura)

        Welch Two Sample t-test

data:  pai.filho$p_altura and pai.filho$f_altura
t = -8.3259, df = 2152.6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.128765 -1.935857
sample estimates:
mean of x mean of y
 171.9252  174.4575
```

FACULDADE DE TECNOLOGIA

Sumário: exploração de dados

Funções sobre a estrutura do conjunto de dados

str(), names()

Funções sobre o conjunto de dados

summary(), sample()

Funções estatísticas básicas

min(), max(), mean(), median(), quantile()

Funções auxiliares

round(), split()

FACULDADE DE TECNOLOGIA

Sumário: exploração de dados

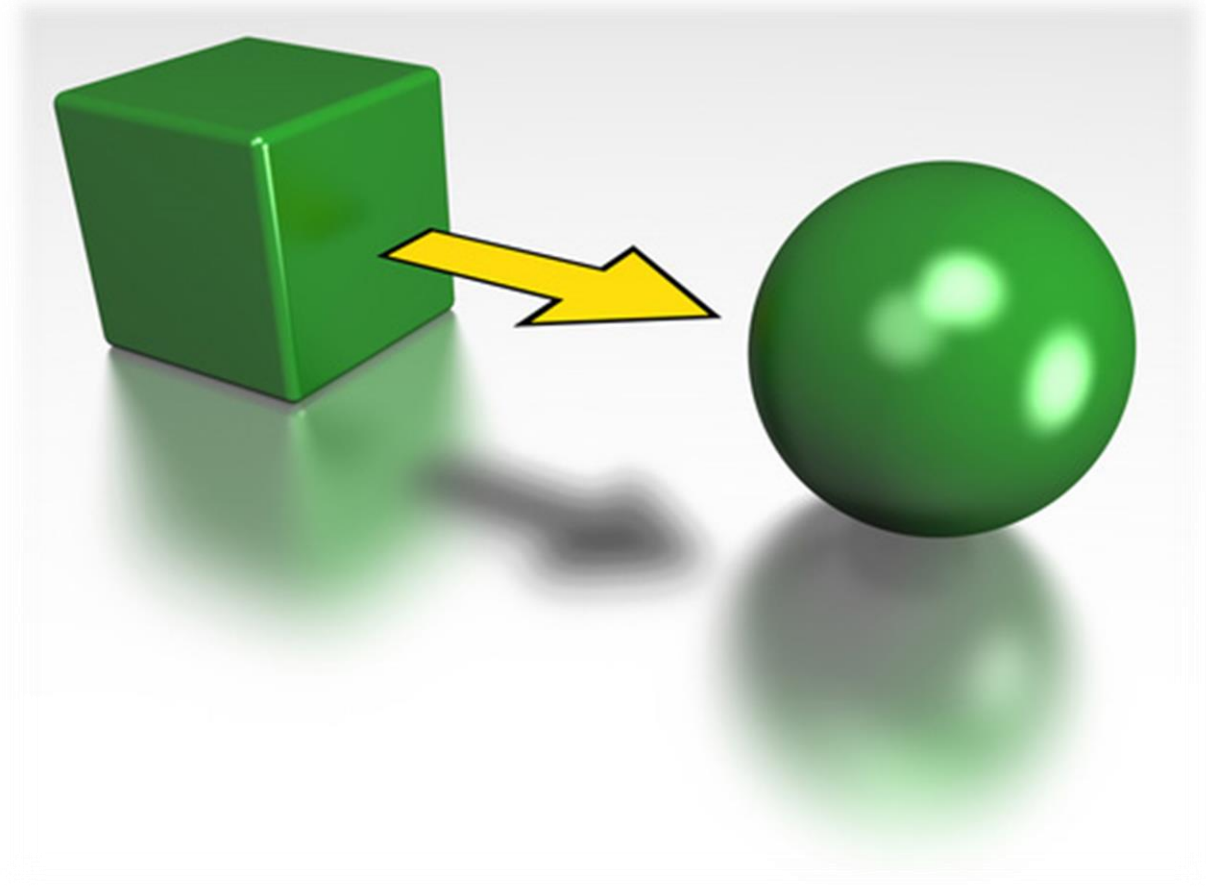
Funções gráficas

- hist(), plot(), boxplot()
- qqnorm()
- abline(), qqline()

Teste de hipóteses

- t.test()

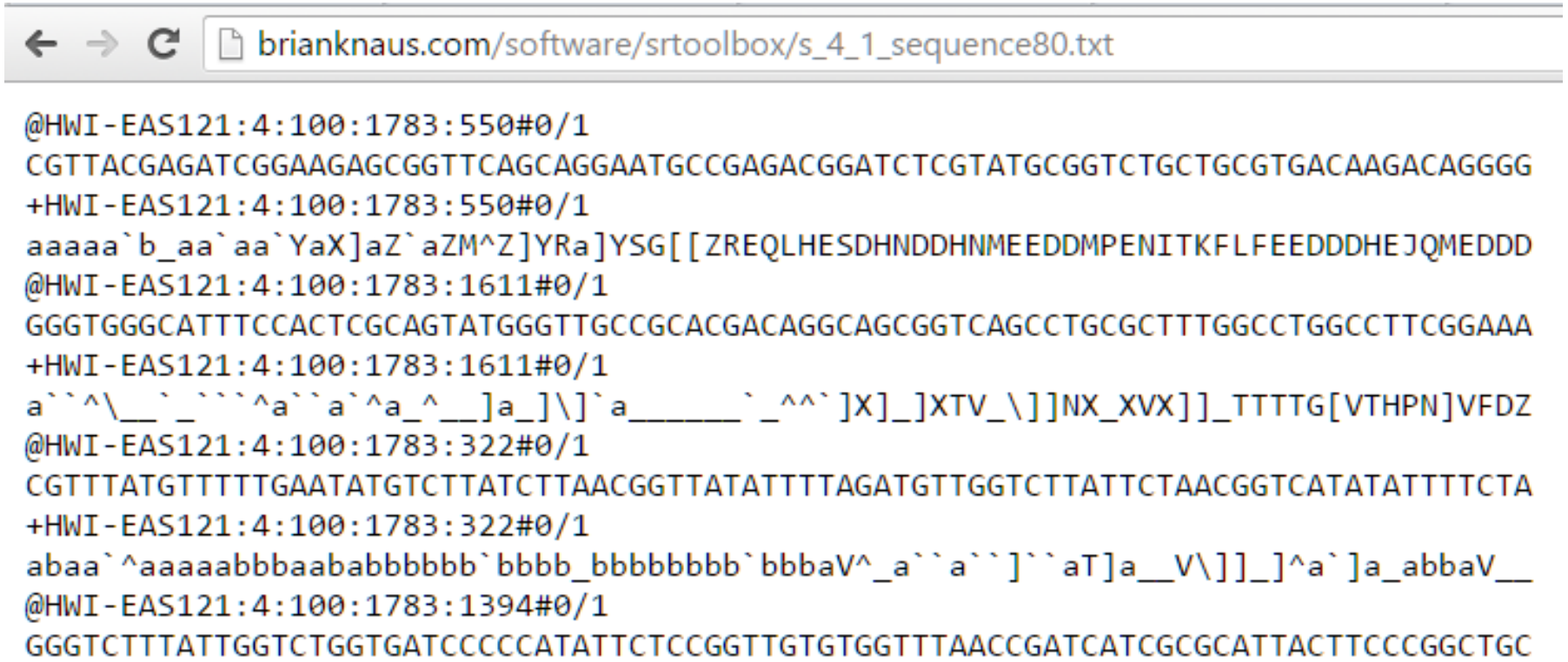
FACULDADE DE TECNOLOGIA



Manipulação de dados

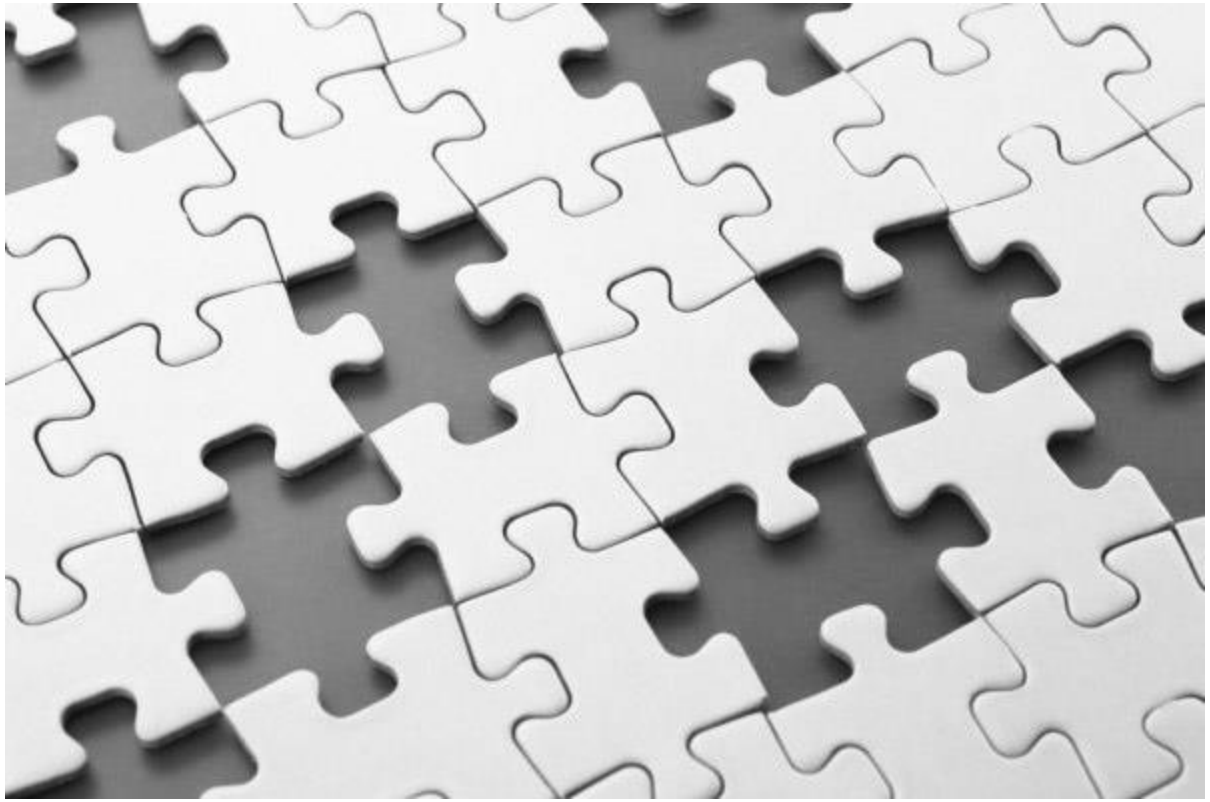
Por que manipular os dados?

Nem sempre a fonte tem dados bem organizados...



Por que manipular os dados?

Nem sempre os dados disponíveis estão completos...

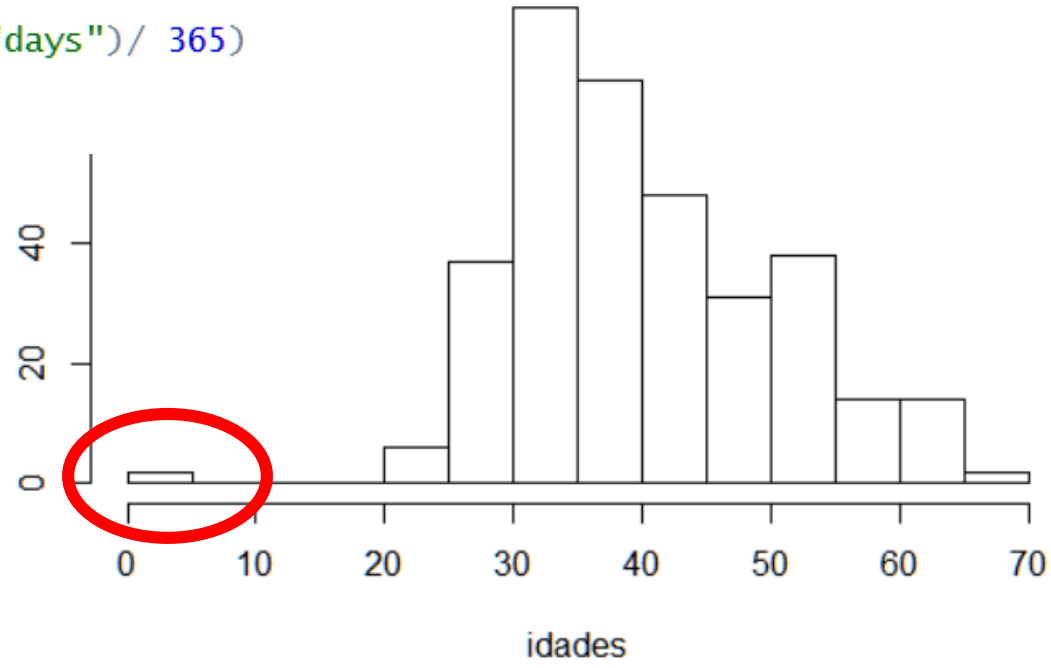


Por que manipular os dados?

Nem sempre os dados disponíveis estão corretos...

```
dataNasc <- as.Date(demog$datanasc)
hoje <- Sys.Date()
idades <- as.vector(difftime(hoje, dataNasc, units = "days")/ 365)

hist(idades)
```



Por que manipular os dados?

Nem sempre os dados estão no formato adequado para o processamento ou para a apresentação gráfica em R

Journal of Statistical Software

August 2014, Volume 59, Issue 10. *http://www.jstatsoft.org/*

Tidy Data

Hadley Wickham
RStudio

FACULDADE DE TECNOLOGIA

Conceito de *Tidy Data*

Dados estão arrumados quando:

1. Cada variável forma uma coluna
2. Cada observação forma uma linha
3. Cada tipo de unidade observacional forma uma tabela

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Manipulação de blocos de dados

Família da função *apply*

apply {base} R Documentation

Apply Functions Over Array Margins

Description

Returns a vector or array or list of values obtained by applying a function to margins of an array or matrix.

Usage

```
apply(X, MARGIN, FUN, ...)
```

FACULDADE DE TECNOLOGIA

```
M <- matrix(seq(1:9), nrow=3)
```

	V1	V2	V3
1	1	4	7
2	2	5	8
3	3	6	9

Dimensão 1: aplicar função por linhas

```
> apply(M,1,sum)
[1] 12 15 18
```

Dimensão 2: aplicar função por colunas

```
> apply(M,2,mean)
[1] 2 5 8
```

Para mais de uma dimensão:

```
> apply(M, c(1,2), log2)
          [,1]     [,2]     [,3]
[1,] 0.000000 2.000000 2.807355
[2,] 1.000000 2.321928 3.000000
[3,] 1.584963 2.584963 3.169925
```

Obs.: especificamente para somas e médias:

colSums {base}

Form Row and Column Sums and Means

Description

Form row and column sums and means for numeric arrays (or data frames).

Família de funções apply

apply

Aplicar função a linhas/colunas de uma matriz

lapply

Aplicar função a uma lista (retorno é uma lista)

sapply

Aplicar função a uma lista (retorno é um vetor)

tapply

Aplicar função a um subconjunto de um vetor, segundo fatores em outro vetor

FACULDADE DE TECNOLOGIA

```
x <- 1:20
```

```
> x
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
```

```
y <- factor(rep(letters[1:5], each = 4))
```

```
> y
 [1] a a a a b b b b c c c c d d d d e e e e
```

```
> tapply(x, y, sum)
 a  b  c  d  e
10 26 42 58 74
```

O pacote dplyr

```
> install.packages("dplyr")
> library(dplyr)
```

Pacote para manipulação de dados em *data frames*

Conjunto de funções que representam ações sobre as tabelas

filter, select, mutate, arrange...

Operador para encadear sequências de ações

%>%

dplyr: exemplo

airquality {datasets}

New York Air Quality Measurements

Description

Daily air quality measurements in New York, May to September 1973.

A data frame with 154 observations on 6 variables.

[,1]	Ozone	numeric Ozone (ppb)
[,2]	Solar.R	numeric Solar R (lang)
[,3]	Wind	numeric Wind (mph)
[,4]	Temp	numeric Temperature (degrees F)
[,5]	Month	numeric Month (1--12)
[,6]	Day	numeric Day of month (1--31)

```
> head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
```

```
> nrow(airquality)
[1] 153
> ncol(airquality)
[1] 6
```

dplyr: mutate

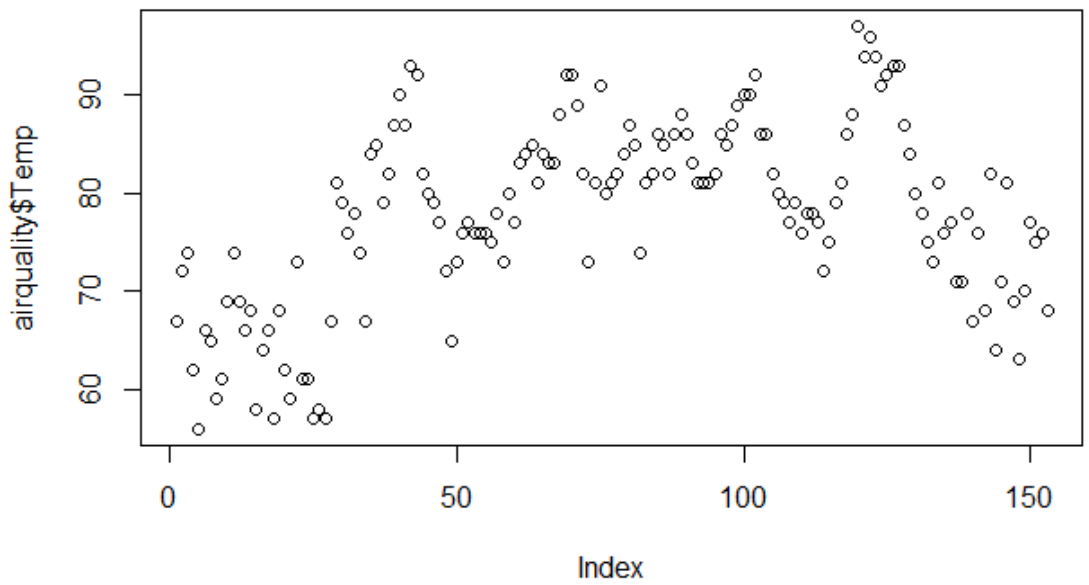
A função *mutate* permite criar novas colunas a partir de valores existentes

```
airQualityC <- airquality %>%
    mutate(TempC = round( (Temp-32)*5/9, 1) )
```

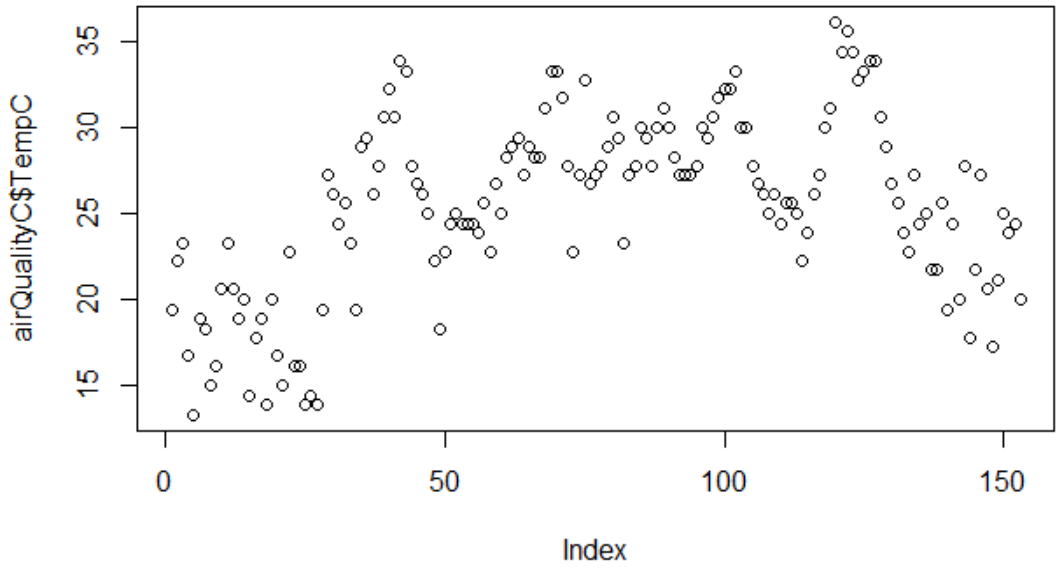
```
> head(airQualityC)
  Ozone Solar.R Wind Temp Month Day TempC
1    41     190  7.4   67     5   1  19.4
2    36     118  8.0   72     5   2  22.2
3    12     149 12.6   74     5   3  23.3
4    18     313 11.5   62     5   4  16.7
5    NA      NA 14.3   56     5   5  13.3
6    28      NA 14.9   66     5   6  18.9
```

FACULDADE DE TECNOLOGIA

```
plot(airquality$Temp)
```




```
plot(airQualityC$TempC)
```



dplyr: filter

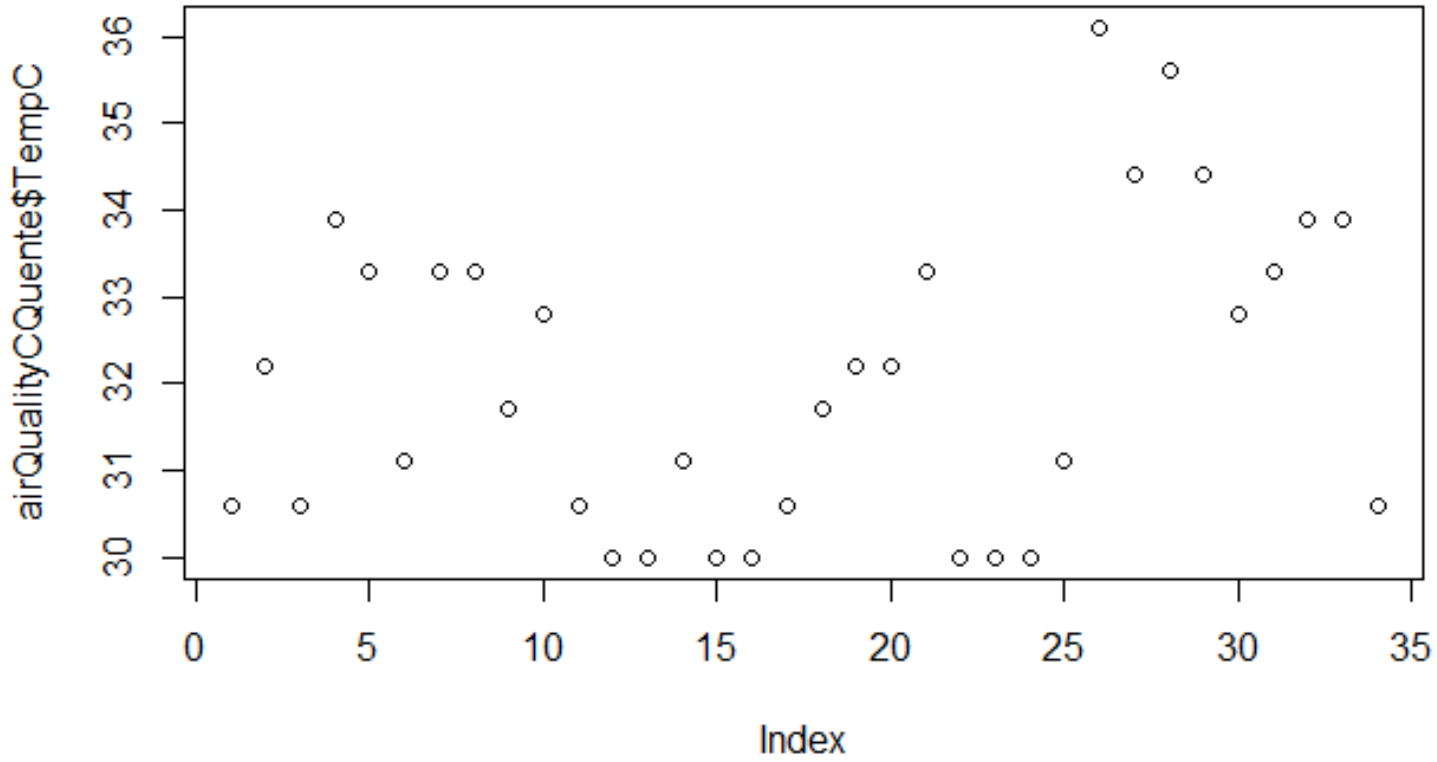
A função *filter* permite selecionar apenas as linhas que satisfazem a uma condição

```
airQualityCQuente <- airQualityC %>%
    filter(TempC >= 30)
```

```
> nrow(airQualityCQuente)
[1] 34
```

```
> head(airQualityCQuente)
  Ozone Solar.R Wind Temp Month Day TempC
1    NA     273  6.9   87     6   8  30.6
2    71     291 13.8   90     6   9  32.2
3    39     323 11.5   87     6  10  30.6
4    NA     259 10.9   93     6  11  33.9
5    NA     250  9.2   92     6  12  33.3
6    77     276  5.1   88     7   7  31.1
```

```
plot(airQualityCQuente$TempC)
```



FACULDADE DE TECNOLOGIA

```
> mean (airquality$Ozone)
[1] NA
```

```
> sum(is.na(airquality$Ozone))
[1] 37
```

```
> mean (airquality$Ozone, na.rm = TRUE)
[1] 42.12931
```

```
airquality_ozone <- airquality %>%
    filter(!is.na(Ozone))
```

```
> nrow(airquality_ozone)
[1] 116
```

```
> mean (airquality_ozone$Ozone)
[1] 42.12931
```

dplyr: select

A função *select* permite escolher as colunas de interesse

```
qualidadeAr <- airQualityC %>%
    select(Ozone, TempC, Month, Day)
```

```
> head(qualidadeAr)
  Ozone TempC Month Day
1    41  19.4     5   1
2    36  22.2     5   2
3    12  23.3     5   3
4    18  16.7     5   4
5    NA  13.3     5   5
6    28  18.9     5   6
```

dplyr: summarise

A função *summarise* combina valores de várias linhas de acordo com alguma função agregadora

Em geral, utilizada juntamente com a função *group_by*

```
avgTemp <- qualidadeAr %>%
    group_by(Month) %>%
    summarise(TempMedia = mean(TempC, na.rm = TRUE), Dias = n())
```

	Month	TempMedia	Dias
1	5	18.63226	31
2	6	26.16333	30
3	7	28.82903	31
4	8	28.86774	31
5	9	24.95000	30

FACULDADE DE TECNOLOGIA

```
diasQuentes <- airQualityCQuente %>%
    group_by(Month) %>%
    summarise(TempMin = min(TempC),
              TempMax = max(TempC),
              TempMedia = round(mean(TempC),1),
              Dias = n())
```

	Month	TempMin	TempMax	TempMedia	Dias
1	6	30.6	33.9	32.1	5
2	7	30.0	33.3	31.4	10
3	8	30.0	36.1	32.3	14
4	9	30.6	33.9	32.9	5

dplyr: arrange

A função *arrange* permite ordenar as linhas pelos valores especificados

```
avgTempDecr <- avgTemp %>%
    arrange(desc(TempMedia))
```

```
> avgTempDecr
# A tibble: 5 x 3
  Month TempMedia  Dias
  <int>     <dbl> <int>
1     8  28.86774    31
2     7  28.82903    31
3     6  26.16333    30
4     9  24.95000    30
5     5  18.63226    31
```

dplyr: encadeamento de funções com %>%

A sequência de funções pode ser combinada num único comando

```
qualidadeAr <- airquality %>%
    mutate(Data = as.Date(paste0("1973", "-", Month, "-", Day)),
           Ozonio = Ozone,
           TempC = round( (Temp-32)*5/9, 1) ) %>%
    select(Data, Ozonio, TempC) %>%
    na.omit()
```

	Data	Ozonio	TempC
1	1973-05-01	41	19.4
2	1973-05-02	36	22.2
3	1973-05-03	12	23.3
4	1973-05-04	18	16.7
6	1973-05-06	28	18.9

O pacote reshape2

O pacote *reshape2* oferece funções para converter formatos de dados em *data frames*

- Aplicação dos princípios de *tidy data*
- Do formato largo (*messy*) para longo (*tidy*): *melt*
- Do formato longo para largo: *dcast*

```
> install.packages("reshape2")
Installing package into 'C:/User
> library(reshape2)
```

reshape2: melt

Melt a data frame into form suitable for easy casting.

Description

You need to tell melt which of your variables are id variables, and which are measured variables. If you only supply one of `id.vars` and `measure.vars`, melt will assume the remainder of the variables in the data set belong to the other. If you supply neither, melt will assume factor and character variables are id variables, and all others are measured.

Usage

```
## S3 method for class 'data.frame'
melt(data, id.vars, measure.vars,
  variable.name = "variable", ..., na.rm = FALSE, value.name = "value",
  factorsAsStrings = TRUE)
```

FACULDADE DE TECNOLOGIA

```
> head(qualidadeAr)
  Ozone TempC Month Day
1    41  19.4     5   1
2    36  22.2     5   2
3    12  23.3     5   3
4    18  16.7     5   4
5    NA  13.3     5   5
6    28  18.9     5   6
```

```
> tail(qualidadeAr)
    Ozone TempC Month Day
148    14  17.2     9  25
149    30  21.1     9  26
150    NA  25.0     9  27
151    14  23.9     9  28
152    18  24.4     9  29
153    20  20.0     9  30
```

```
tidyQualAr <- melt(qualidadeAr, id.vars = c("Month", "Day"))
```

```
> head(tidyQualAr) ; tail(tidyQualAr)
  Month Day variable value
1     5   1    Ozone    41
2     5   2    Ozone    36
3     5   3    Ozone    12
4     5   4    Ozone    18
5     5   5    Ozone    NA
6     5   6    Ozone    28
      Month Day variable value
301       9  25    TempC  17.2
302       9  26    TempC  21.1
303       9  27    TempC  25.0
304       9  28    TempC  23.9
305       9  29    TempC  24.4
306       9  30    TempC  20.0
```

FACULDADE DE TECNOLOGIA

reshape2: dcast

Cast functions Cast a molten data frame into an array or data frame.

Description

Use `acast` or `dcast` depending on whether you want vector/matrix/array output or data frame output. Data frames can have at most two dimensions.

Usage

```
dcast(data, formula, fun.aggregate = NULL, ..., margins = NULL,
  subset = NULL, fill = NULL, drop = TRUE,
  value.var = guess_value(data))
```

FACULDADE DE TECNOLOGIA

```
> head(tidyQualAr)
  Month Day variable value
1     5   1    Ozone    41
2     5   2    Ozone    36
3     5   3    Ozone    12
```

Notação de formula: Y ~ X

```
backQualAr <- dcast(tidyQualAr, Month + Day ~ variable)
```

```
> head(backQualAr)
  Month Day Ozone TempC
1     5   1    41  19.4
2     5   2    36  22.2
3     5   3    12  23.3
4     5   4    18  16.7
5     5   5    NA  13.3
6     5   6    28  18.9
```

tilde {base} R Documentation

Tilde Operator

Description

Tilde is used to separate the left- and right-hand sides in a model formula.

Usage

```
y ~ model
```

Arguments

`y, model` symbolic expressions.

FACULDADE DE TECNOLOGIA

Sumário: manipulação de dados

Família de funções *apply

apply(), lapply(), sapply(), tapply(),...

Pacote dplyr

mutate(), filter(), select(), summarise(), group_by(), n(), arrange(), desc()

Pacote reshape2

melt(), dcast()

FACULDADE DE TECNOLOGIA

Mais gráficos

FACULDADE DE TECNOLOGIA

Sistemas gráficos em R

Sistema gráfico básico

- Bom para explorações iniciais de dados
- Pouca flexibilidade

Outros sistemas gráficos

- Pacotes R com funções alternativas para a criação de gráficos
  - *Trellis Graphics for R* (lattice)
  - *An Implementation of the Grammar of Graphics* (ggplot2)
  - *The Grid Graphics Package* (grid)


ggplot2.org
ggplot2
ggplot2 is a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts. It takes care of many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics.
Documentation
ggplot2 documentation is now available at docs.ggplot2.org.
> install.packages("ggplot2")

Gráfico rápido em ggplot2: qplot

qplot {ggplot2} R Documentation

Quick plot

Description

`qplot` is the basic plotting function in the ggplot2 package, designed to be familiar if you're used to base `plot()`. It's a convenient wrapper for creating a number of different types of plots using a consistent calling scheme.

```
library(ggplot2)

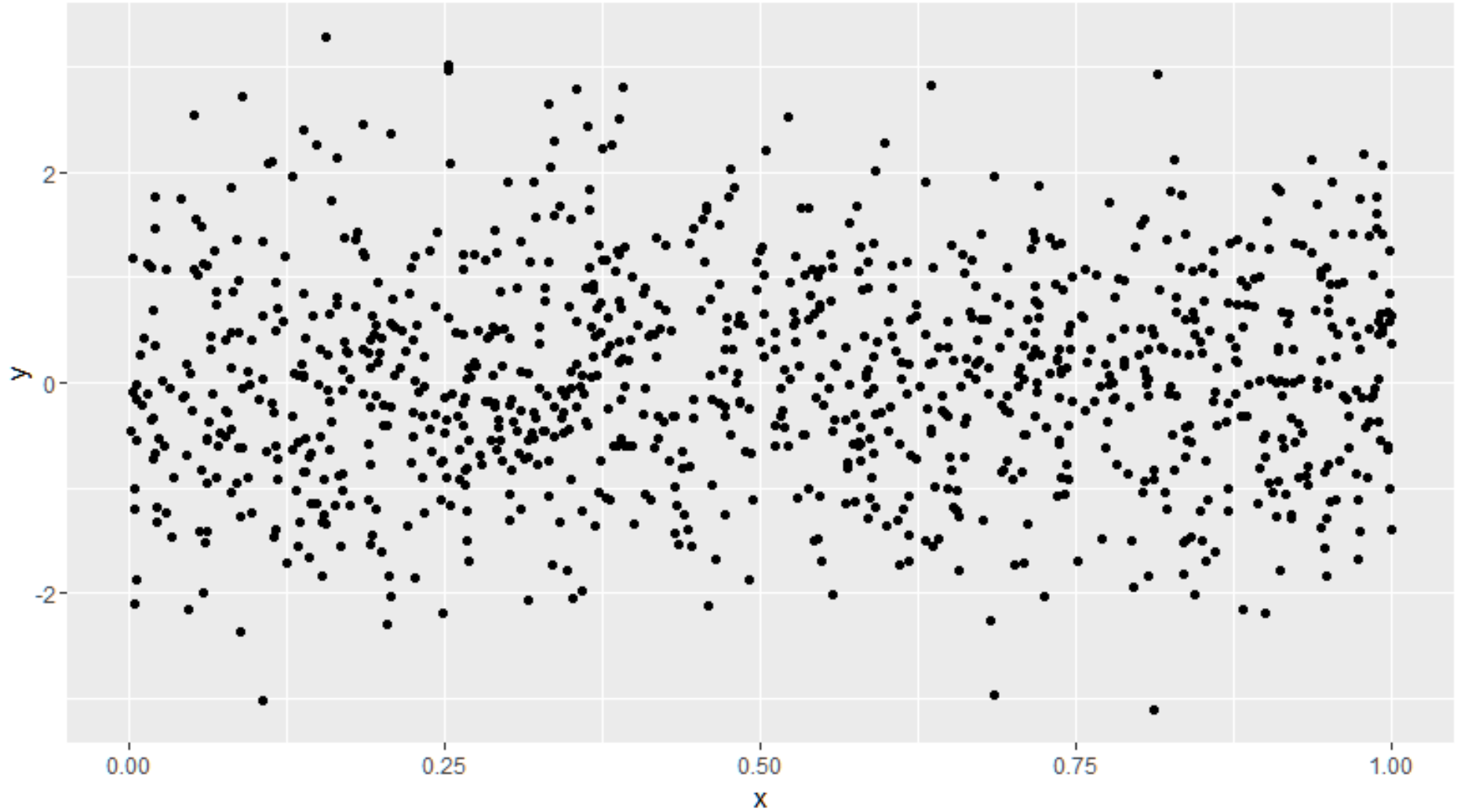
dados <- data.frame(
  x = runif(1000),
  y = rnorm(1000),
  c = sample(
    c("ABC", "DEF", "GHI", "JKL", "MNO", "PQR", "STU", "VW", "XYZ"),
    1000, replace = TRUE
  )
)
```

	x	y	c
1	0.095652930	-1.7722397822	GHI
2	0.140313923	-0.5706917340	MNO
3	0.582859153	-0.0009378132	STU
4	0.869538549	0.6403524729	XYZ
5	0.466994045	-0.3218034604	ABC
6	0.859569707	-0.8537775899	DEF
7	0.920729965	0.9290534527	PQR

FACULDADE DE TECNOLOGIA

Padrão para dados 2d

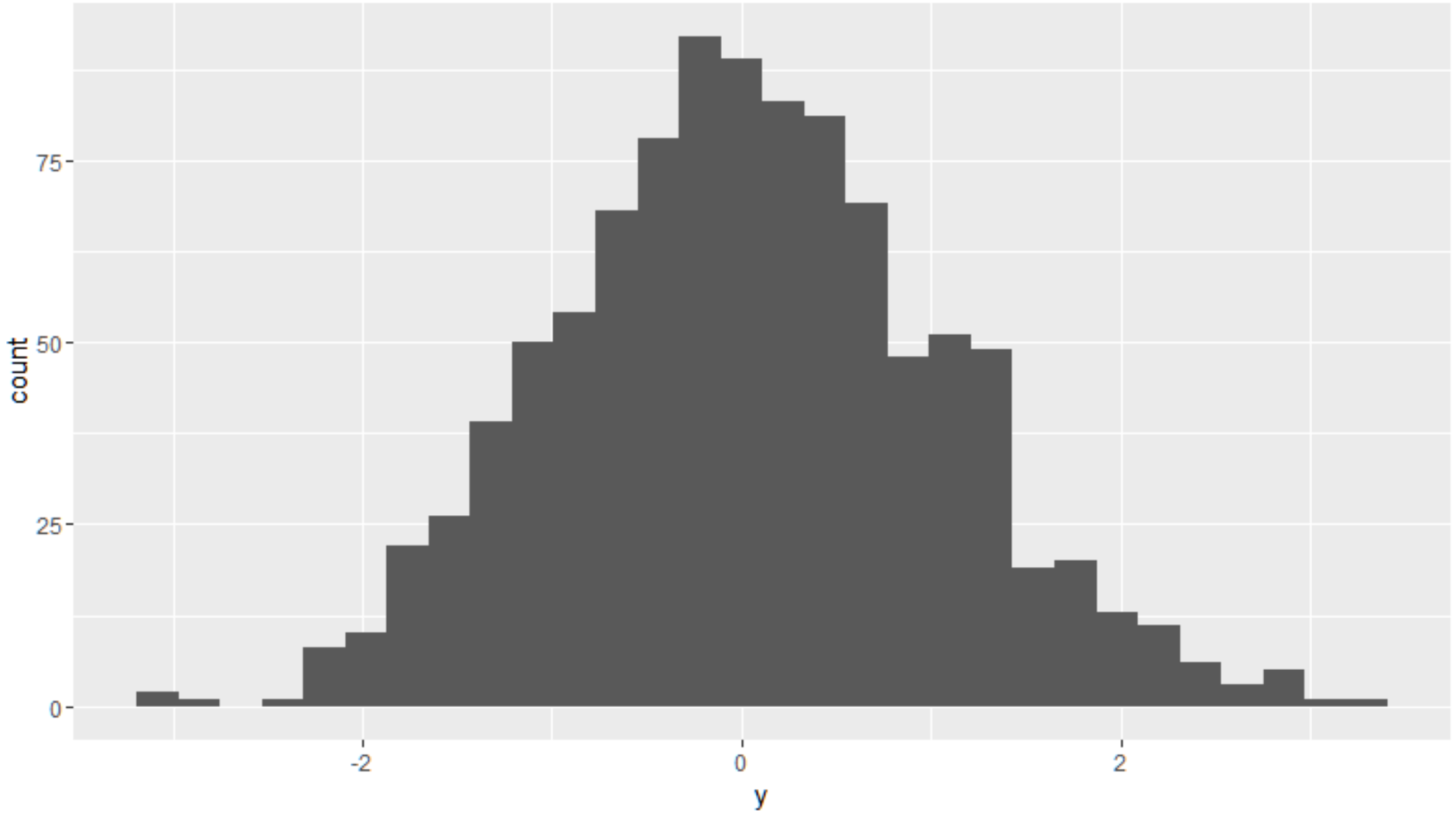
```
qplot(x, y, data = dados)
```



FACULDADE DE TECNOLOGIA

Padrão para dados contínuos 1d

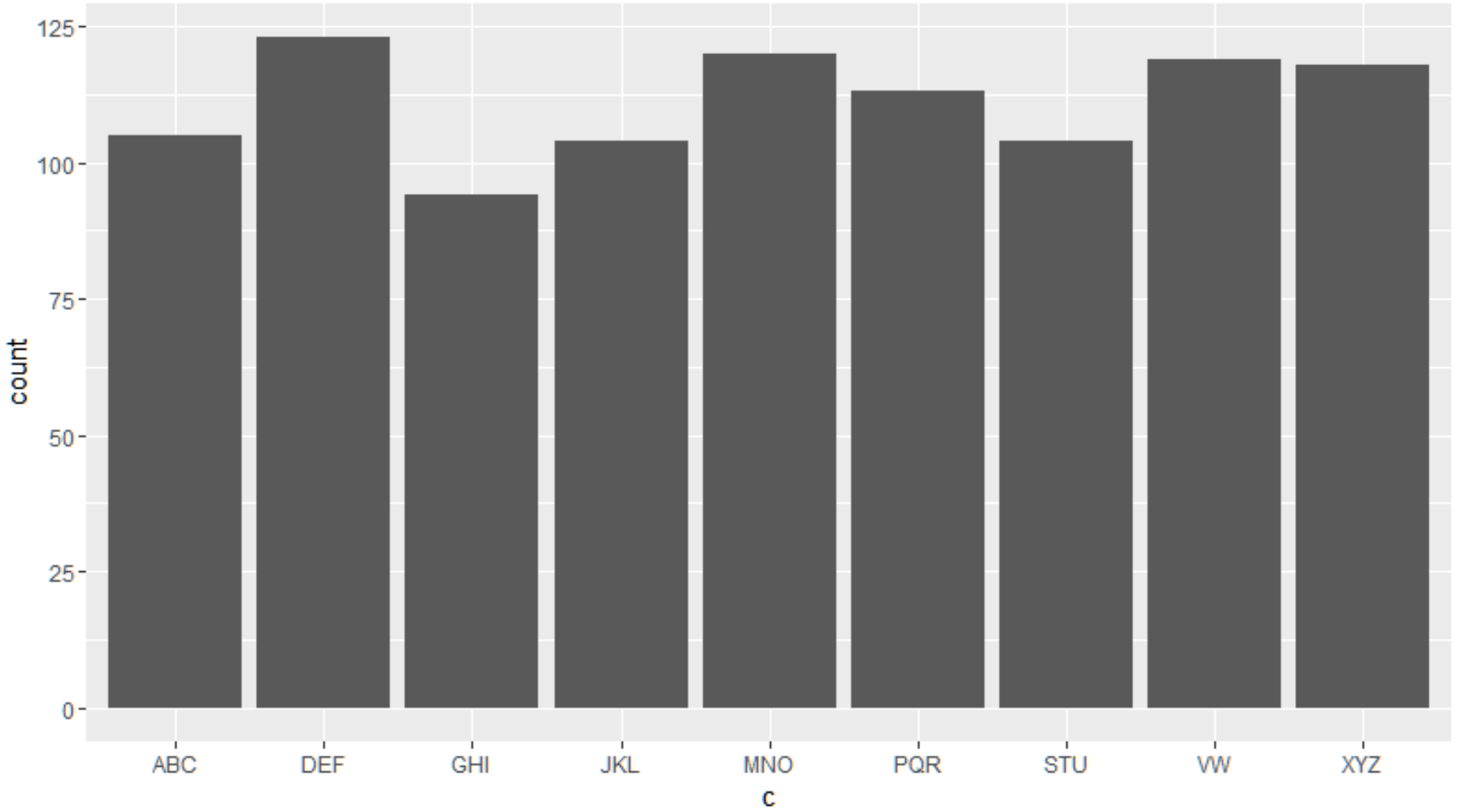
```
qplot(y, data = dados)
```



FACULDADE DE TECNOLOGIA

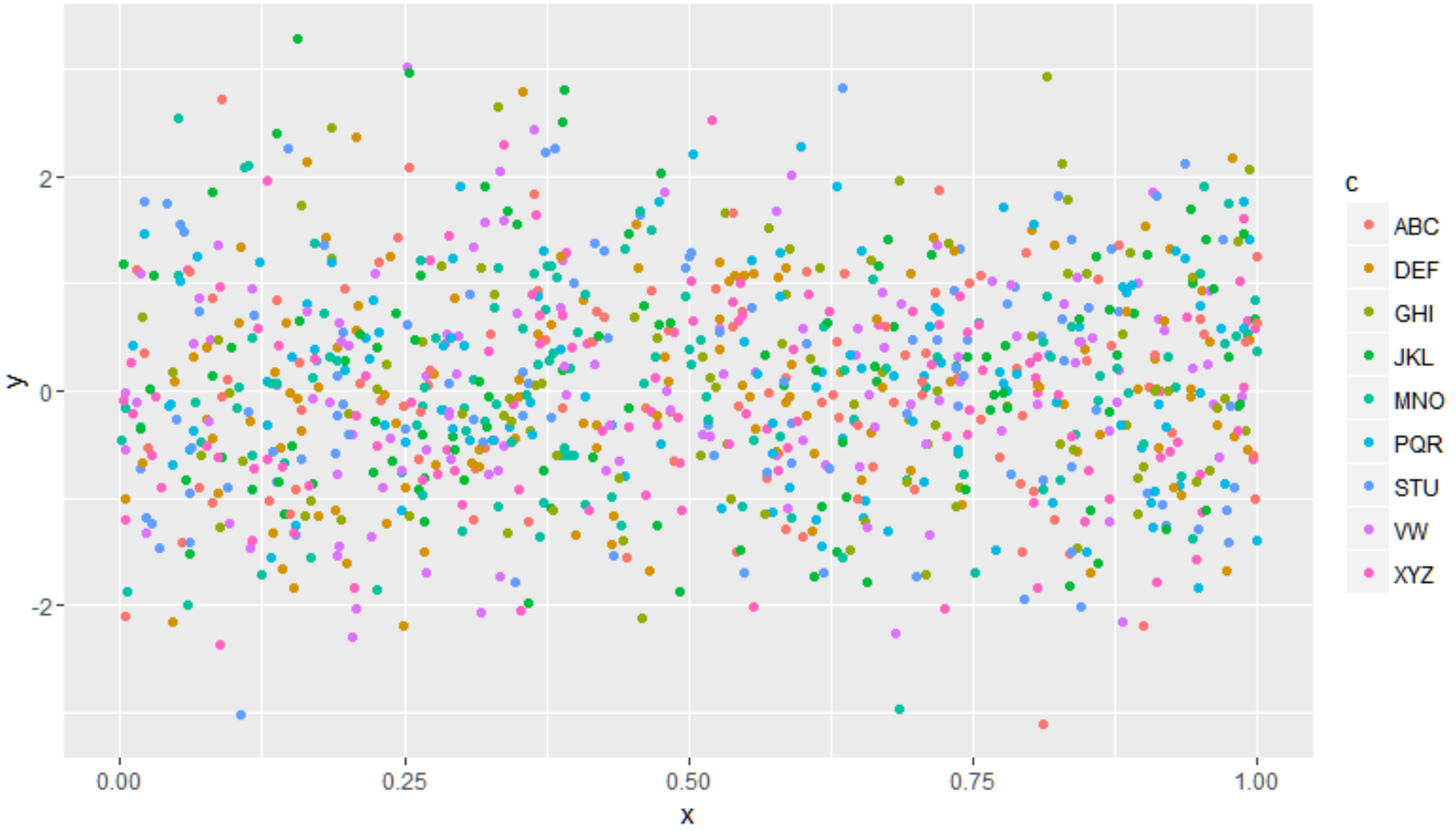
Padrão para dados categóricos 1d

```
qplot(c, data = dados)
```



FACULDADE DE TECNOLOGIA

```
qplot(x, y, data = dados, color = c)
```



Explorando o potencial completo de ggplot2: ggplot

ggplot {ggplot2} R Documentation

Create a new ggplot plot.

Description

`ggplot()` initializes a ggplot object. It can be used to declare the input data frame for a graphic and to specify the set of plot aesthetics intended to be common throughout all subsequent layers unless specifically overridden.

Dados

Estética básica

Camadas

```
ggplot(df, aes(x, y, <other aesthetics>))
```

FACULDADE DE TECNOLOGIA

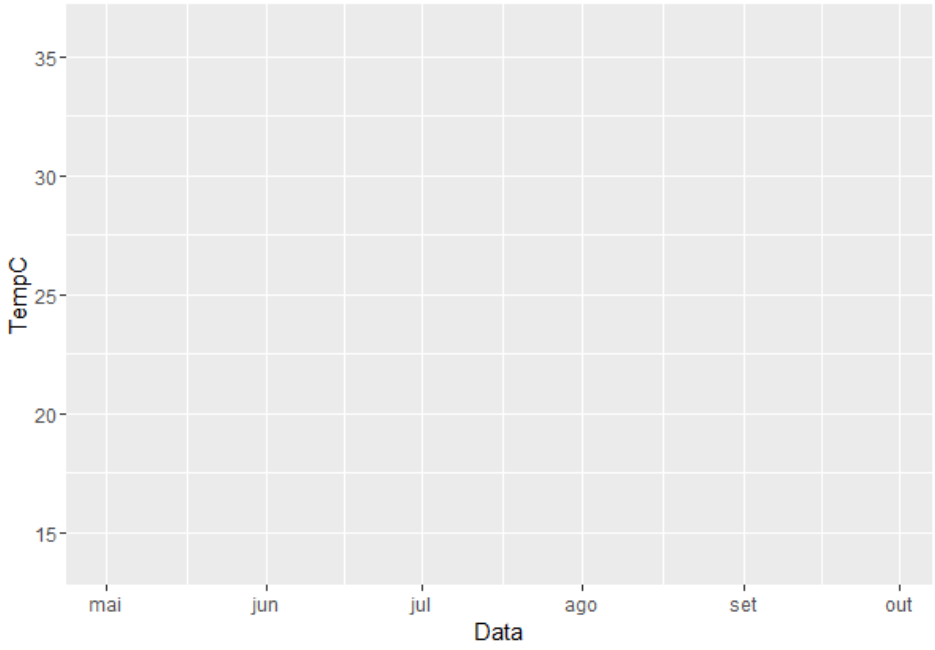
```
library(dplyr)

qualidadeAr <- airquality %>%
    mutate(Data = as.Date(paste0("1973", "-", Month, "-", Day)),
           Ozonio = Ozone,
           TempC = round( (Temp-32)*5/9, 1) ) %>%
    select(Data, Ozonio, TempC) %>%
    na.omit()

library(ggplot2)

p <- ggplot(qualidadeAr, aes(Data, TempC))
```

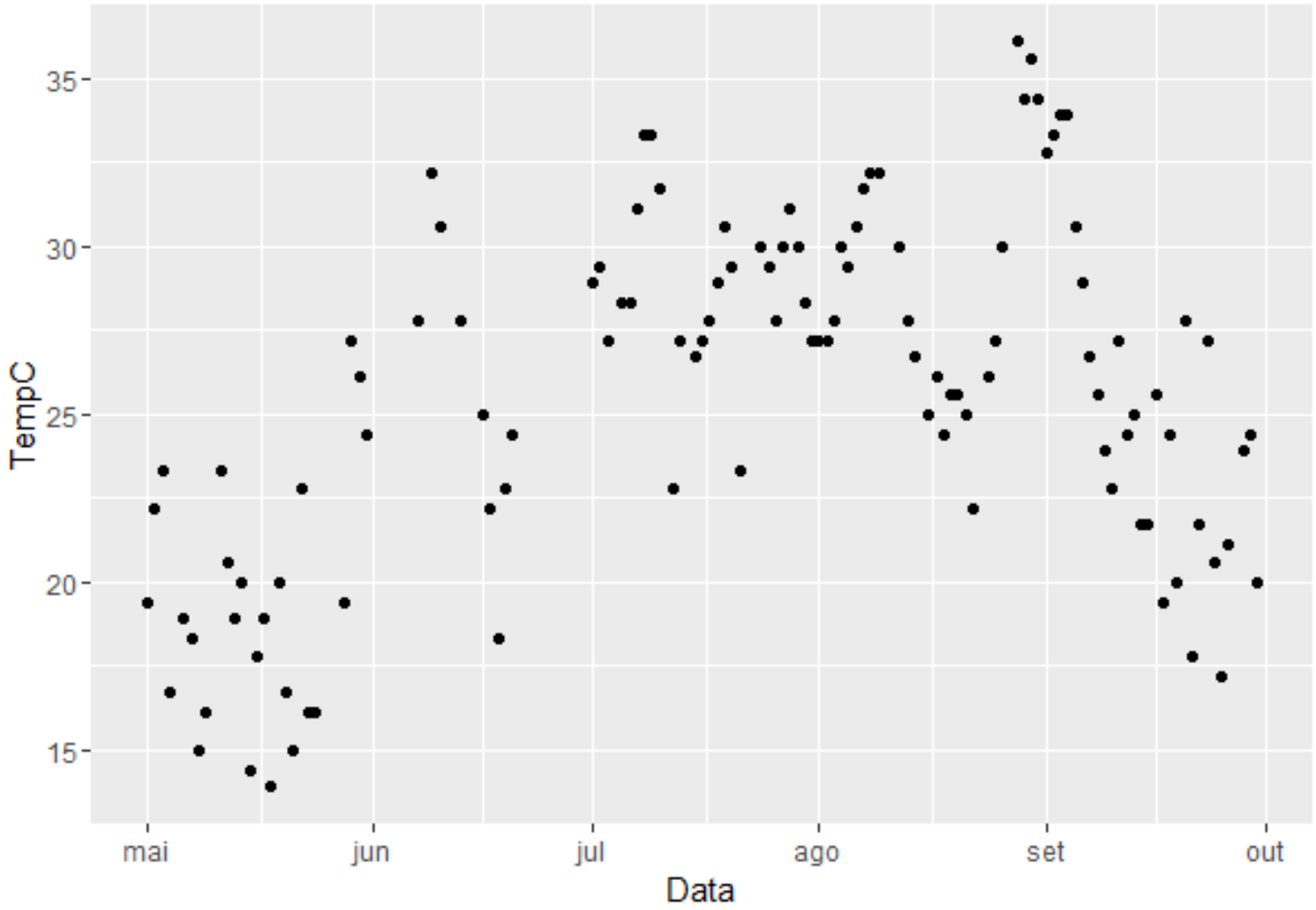
```
> p
```



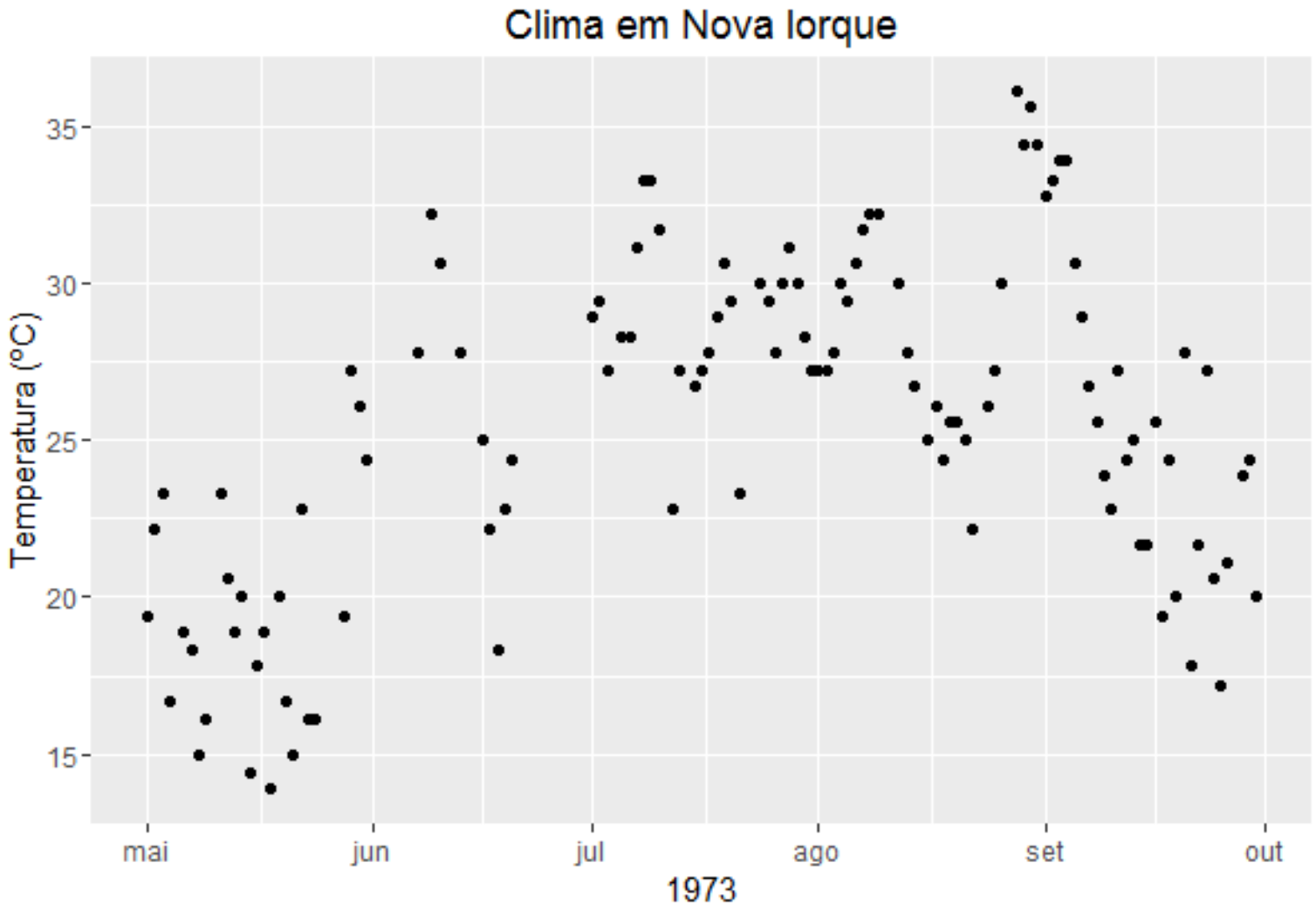
FACULDADE DE TECNOLOGIA

```
p <- ggplot(qualidadeAr, aes(Data, TempC))

p + geom_point()
```



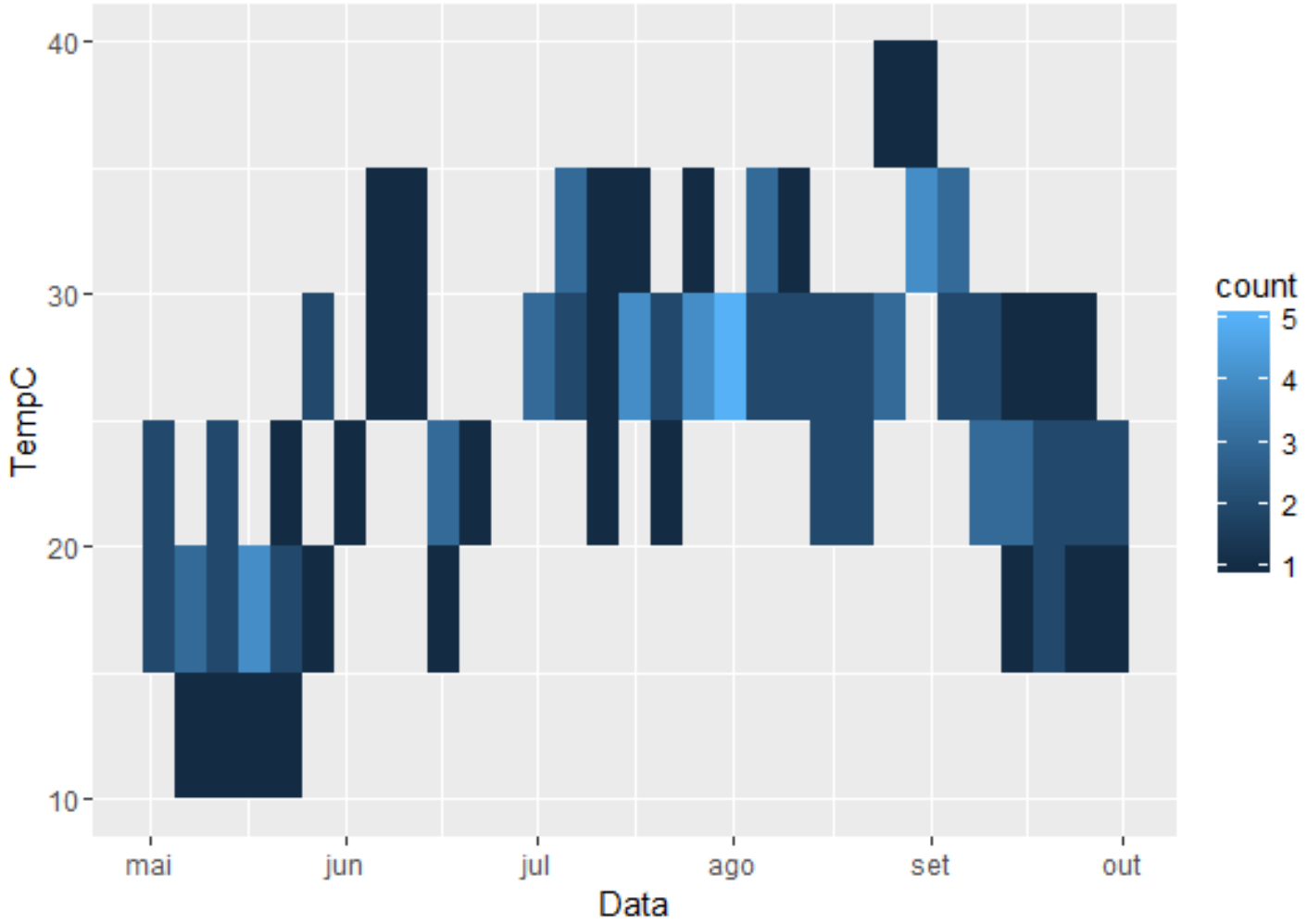
```
p + geom_point() + ggtitle("Clima em Nova Iorque") +
    xlab("1973") + ylab("Temperatura (ºC)")
```



FACULDADE DE TECNOLOGIA

Mesmos dados, *heatmap*

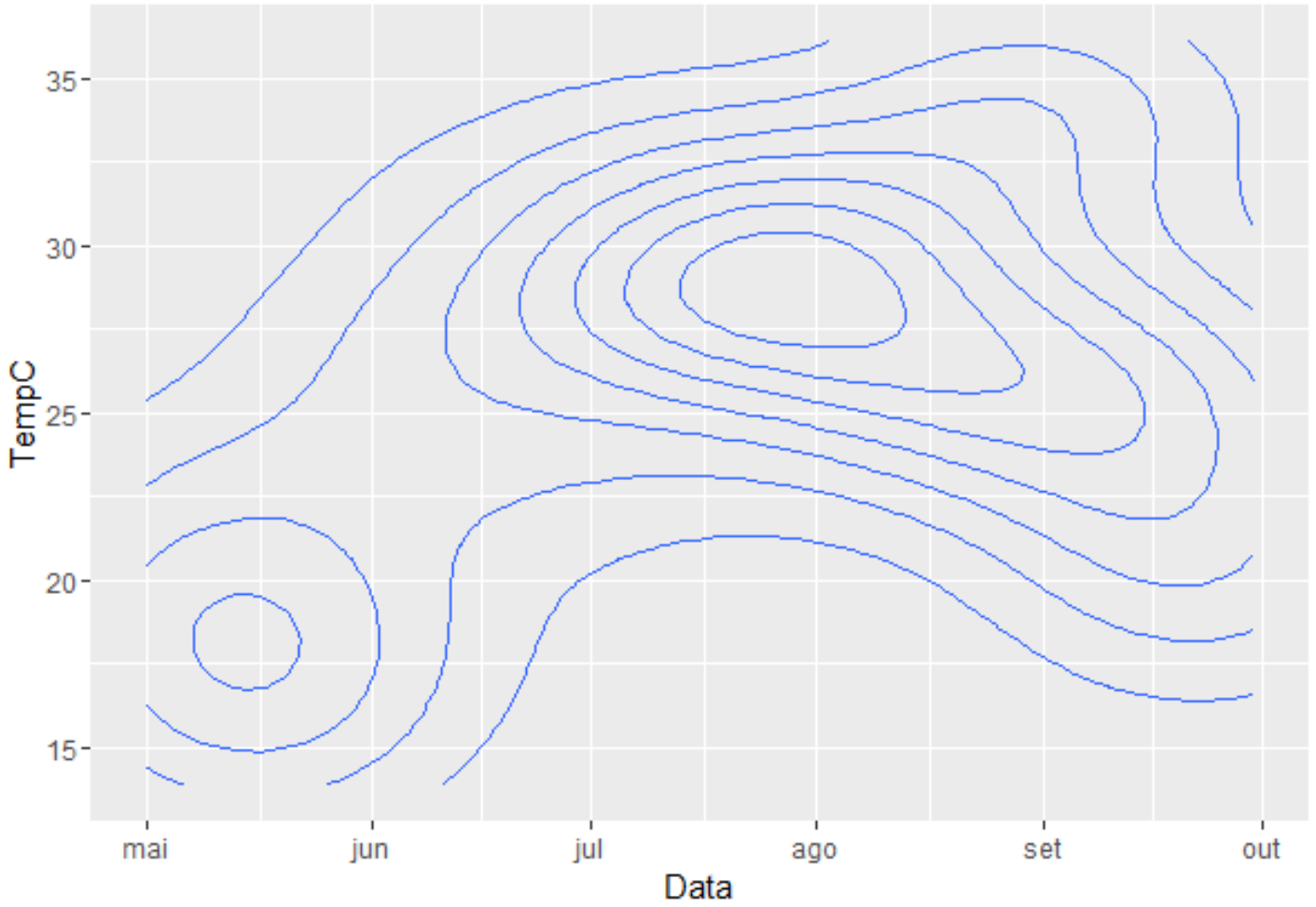
```
p + geom_bin2d(binwidth=5)
```

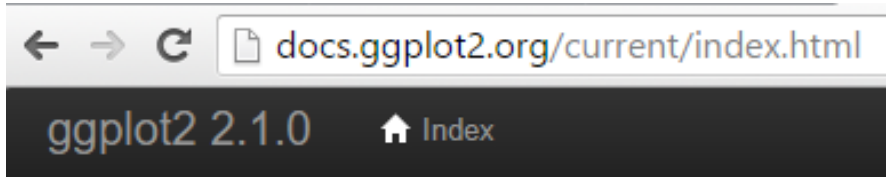


FACULDADE DE TECNOLOGIA

Mesmos dados, densidade

```
p + geom_density2d()
```





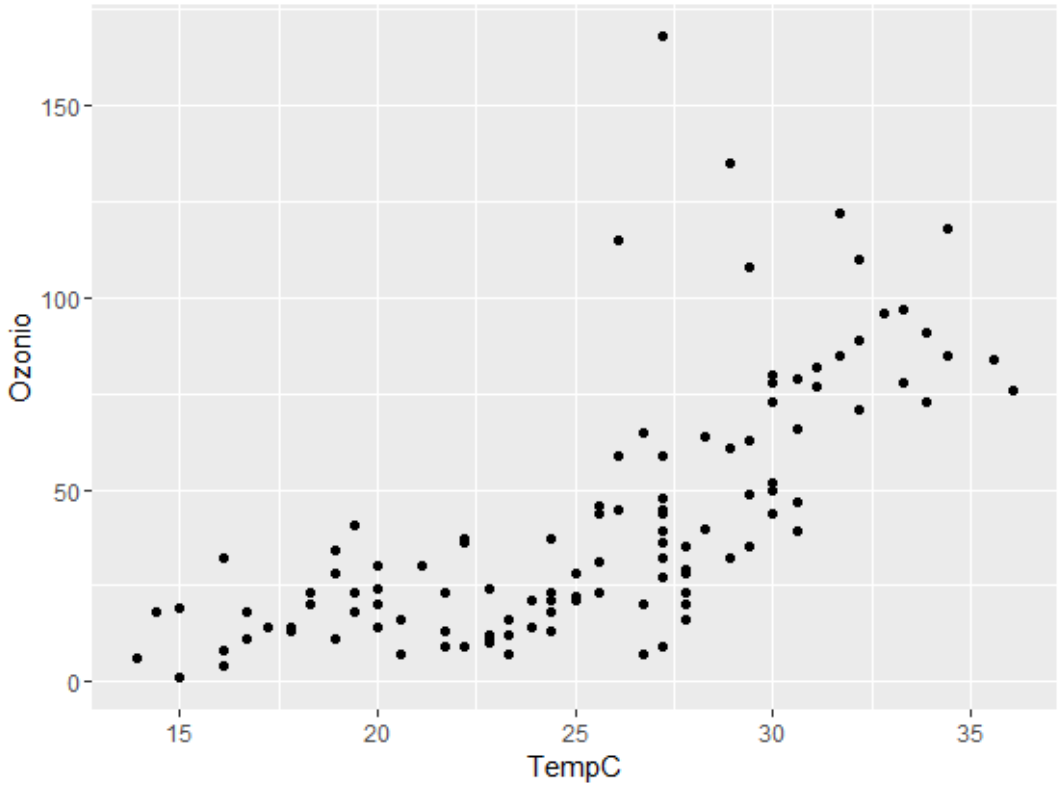
Geoms

Geoms, short for geometric objects, describe the type of plot you will produce.

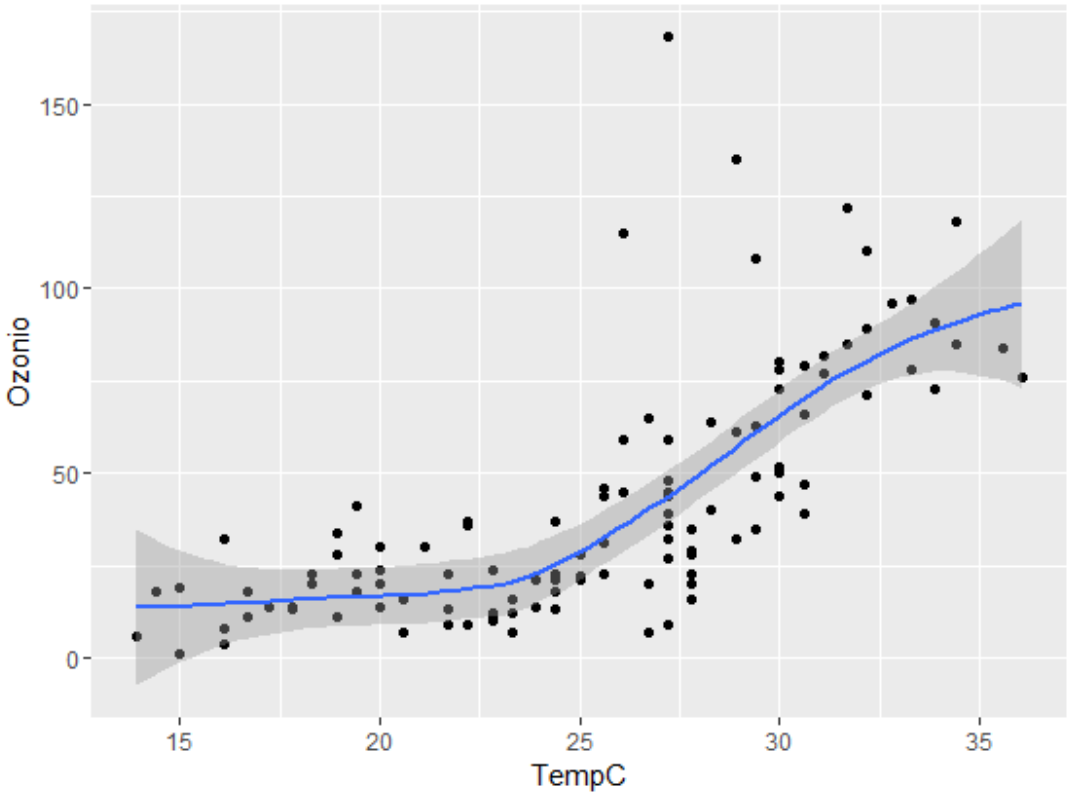
- `geom_abline` (geom_hline, geom_vline)
  Lines: horizontal, vertical, and specified by slope and intercept.
- `geom_bar` (stat_count)
  Bars, rectangles with bases on x-axis
- `geom_bin2d` (stat_bin2d, stat_bin_2d)
  Add heatmap of 2d bin counts.
- `geom_blank`
  Blank, draws nothing.
- `geom_boxplot` (stat_boxplot)
  Box and whiskers plot.
- `geom_contour` (stat_contour)
  Display contours of a 3d surface in 2d.
- `geom_count` (stat_sum)
  Count the number of observations at each location.
- `geom_crossbar` (geom_errorbar, geom_linerange, geom_pointrange)
  Vertical intervals: lines, crossbars & errorbars.
- `geom_density` (stat_density)
  Display a smooth density estimate.
- `geom_density_2d` (geom_density2d, stat_density2d, stat_density_2d)
  Contours from a 2d density estimate.
- `geom_dotplot`
  Dot plot
- `geom_errorbarh`
  Horizontal error bars
- `geom_freqpoly` (geom_histogram, stat_bin)
  Histograms and frequency polygons.
- `geom_hex` (stat_bin_hex, stat_binhex)
  Hexagon binning.
- `geom_jitter`
  Points, jittered to reduce overplotting.
- `geom_label` (geom_text)
  Textual annotations.
- `geom_map`
  Polygons from a reference map.
- `geom_path` (geom_line, geom_step)
  Connect observations.
- `geom_point`
  Points, as for a scatterplot
- `geom_polygon`
  Polygon, a filled path.
- `geom_quantile` (stat_quantile)
  Add quantile lines from a quantile regression.
- `geom_raster` (geom_rect, geom_tile)
  Draw rectangles.
- `geom_ribbon` (geom_area)
  Ribbons and area plots.
- `geom_rug`
  Marginal rug plots.
- `geom_segment` (geom_curve)
  Line segments and curves.
- `geom_smooth` (stat_smooth)
  Add a smoothed conditional mean.
- `geom_violin` (stat_ydensity)
  Violin plot.

geom_smooth()

```
ggplot(qualidadeArMes, aes(TempC, Ozonio)) +
    geom_point()
```

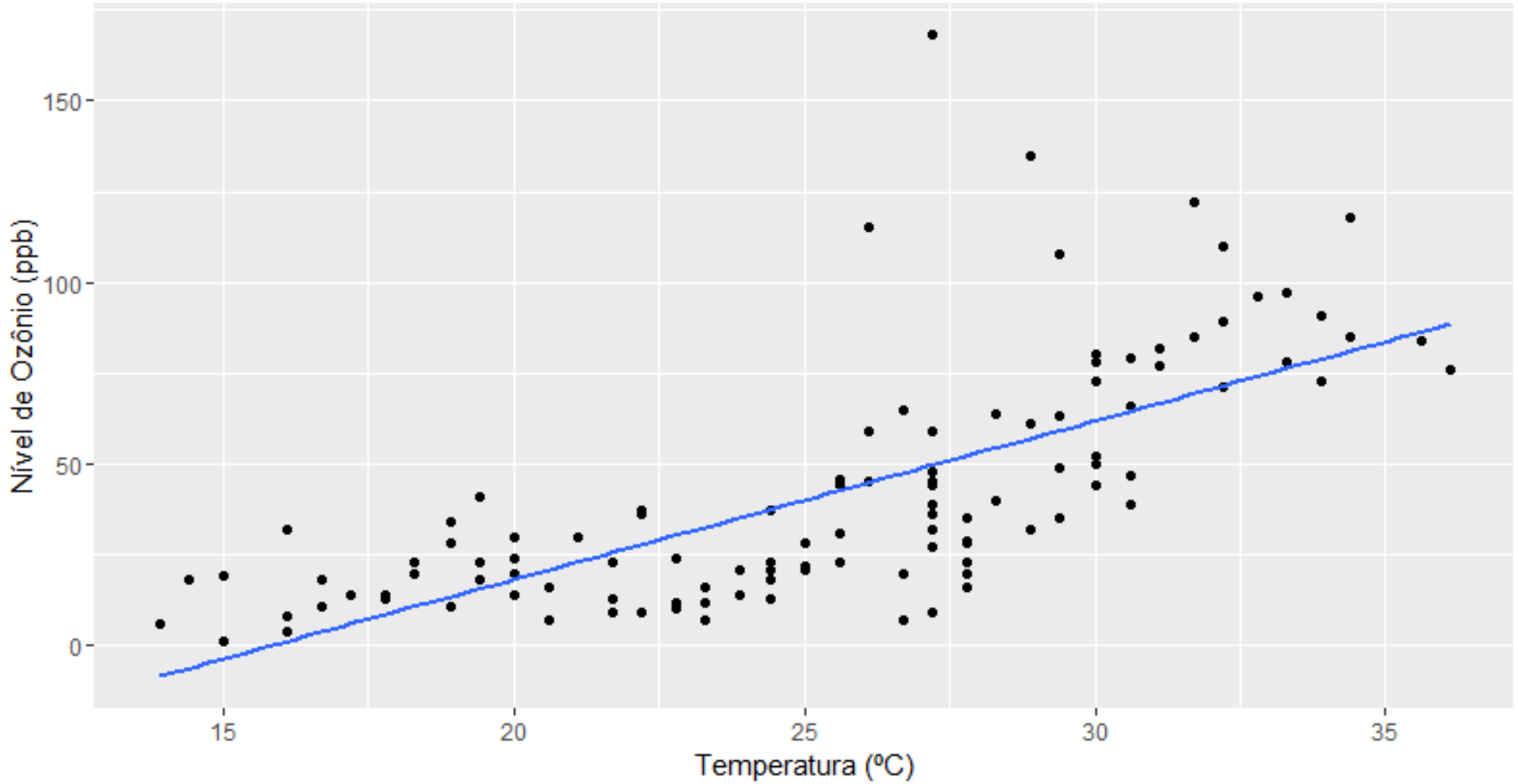



```
ggplot(qualidadeArMes, aes(TempC, Ozonio)) +
    geom_point() + geom_smooth()
```



FACULDADE DE TECNOLOGIA

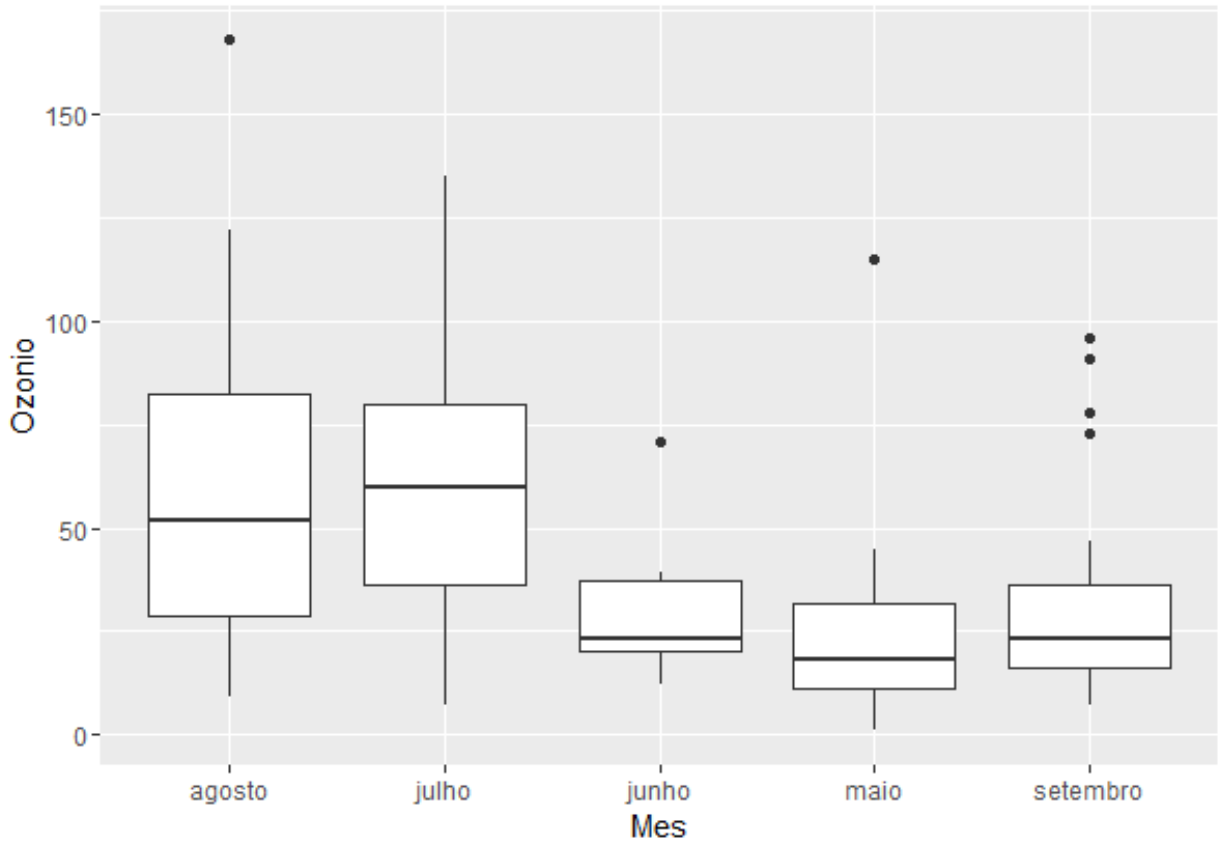
```
ggplot(qualidadeArMes, aes(TempC, Ozonio)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE) +
    xlab("Temperatura (ºC)") + ylab("Nível de Ozônio (ppb)")
```



geom_boxplot()

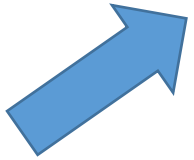
```
qualidadeArMes <- qualidadeAr %>%
    mutate(Mes = months(Data))

p3 <- ggplot(qualidadeArMes, aes(Mes, Ozonio)) +
    geom_boxplot()
```



FACULDADE DE TECNOLOGIA

Estabelece a ordem que deve ser preservada na apresentação

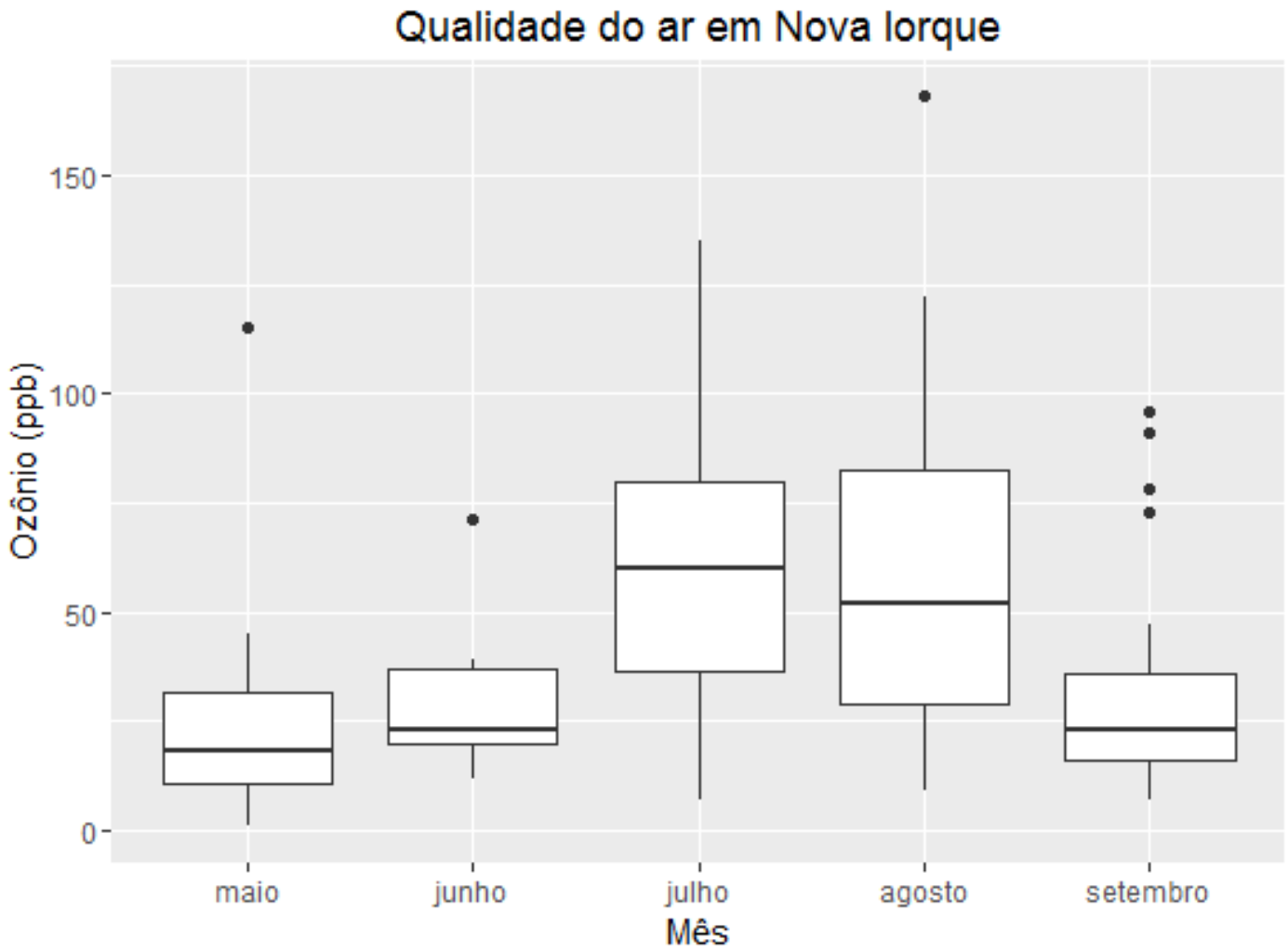


```
mf <- factor(
    qualidadeArMes$Mes,
    levels = c("maio", "junho", "julho", "agosto", "setembro"),
    ordered = TRUE
)
```

```
qualidadeArMes$Mes <- mf

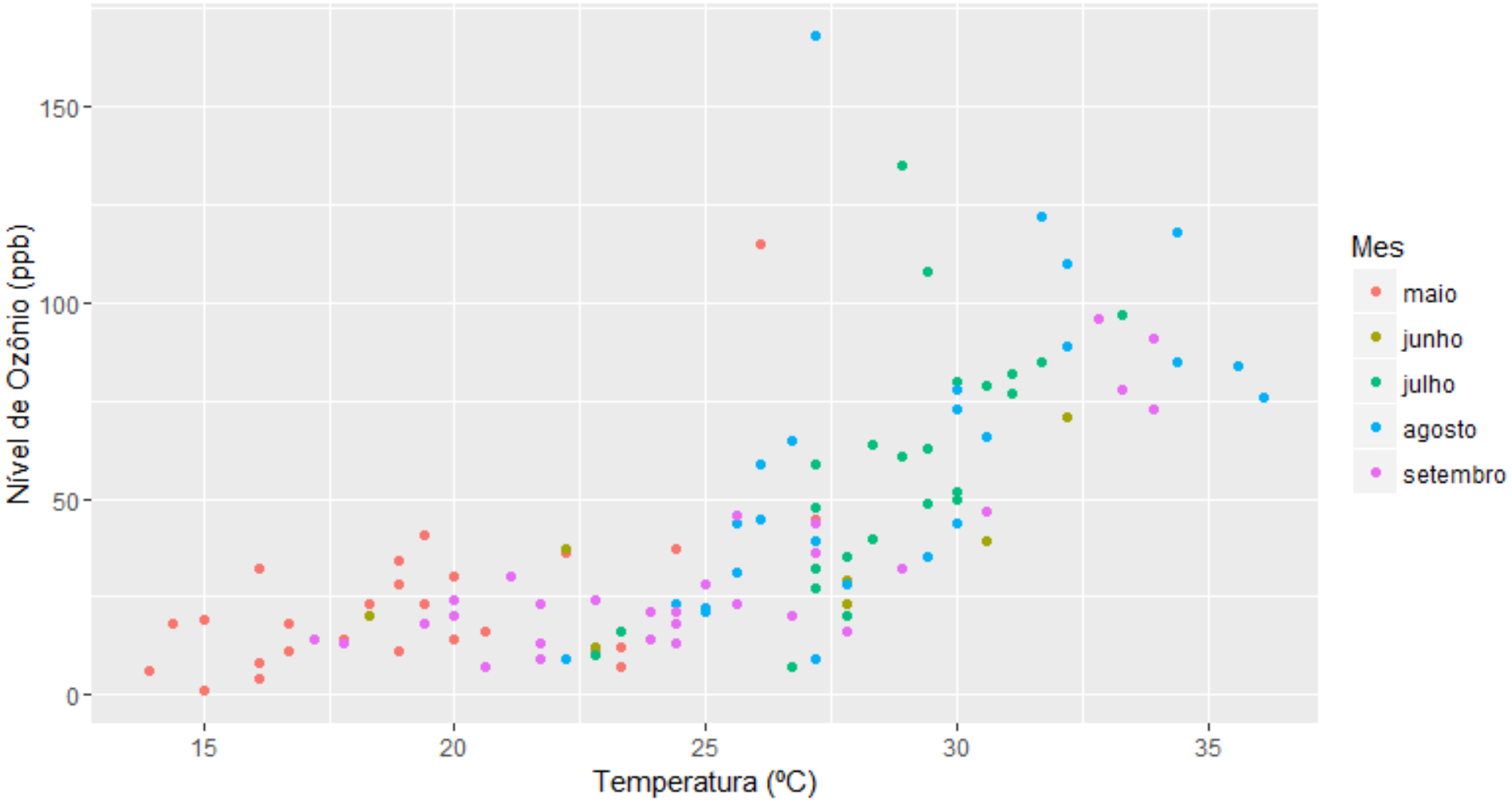
ggplot(qualidadeArMes, aes(Mes, Ozonio)) +
    geom_boxplot() + ggtitle("Qualidade do ar em Nova Iorque") +
    xlab("Mês") + ylab("Ozônio (ppb)")
```

```
ggplot(qualidadeArMes, aes(Mes, Ozonio)) +
    geom_boxplot() + ggtitle("Qualidade do ar em Nova Iorque") +
    xlab("Mês") + ylab("Ozônio (ppb)")
```



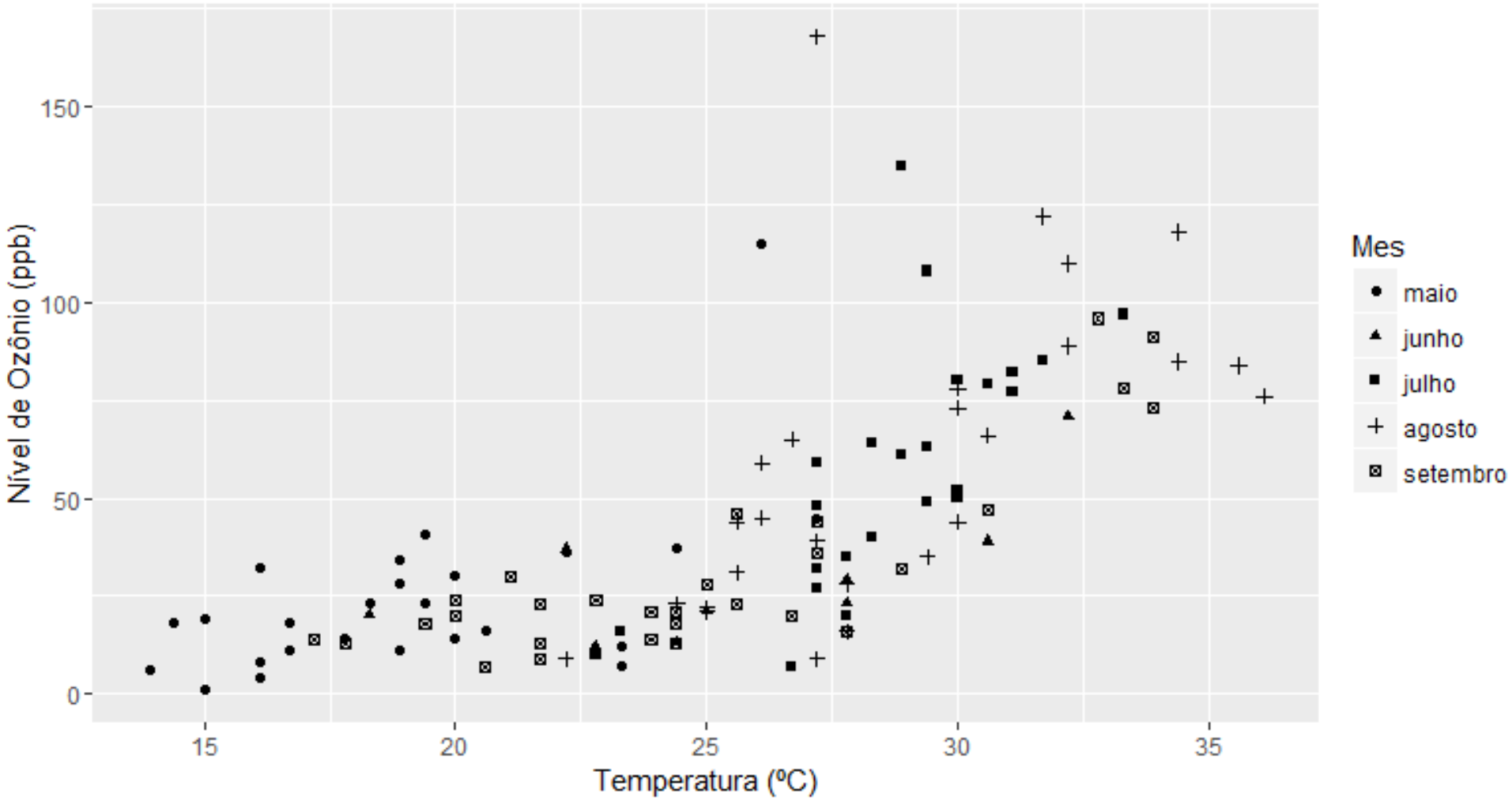
FACULDADE DE TECNOLOGIA

```
ggplot(qualidadeArMes, aes(TempC, Ozonio, colour = Mes)) +
    geom_point() +
    xlab("Temperatura (ºC)") + ylab("Nível de Ozônio (ppb)")
```



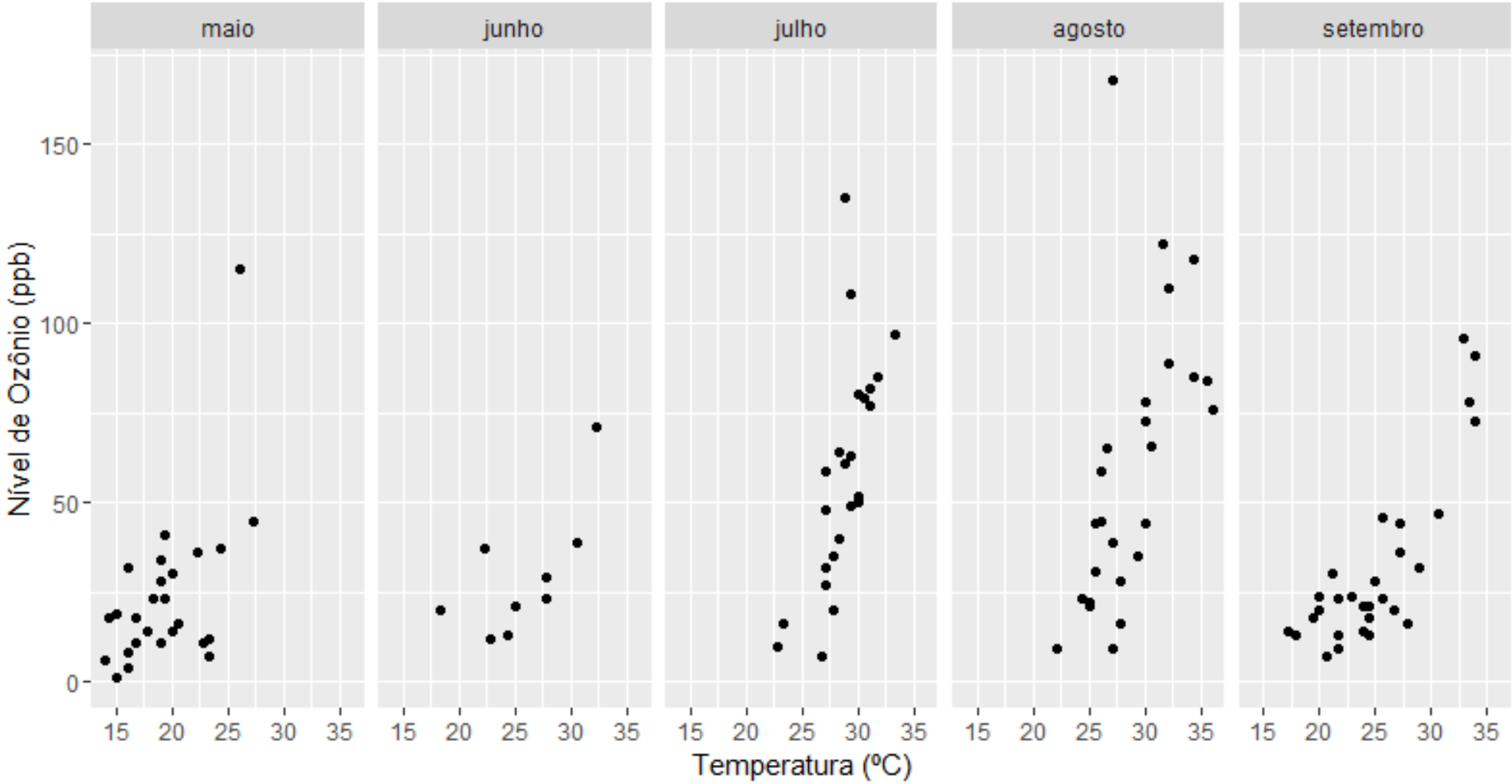
FACULDADE DE TECNOLOGIA

```
ggplot(qualidadeArMes, aes(TempC, Ozonio, shape = Mes)) +
    geom_point() +
    xlab("Temperatura (ºC)") + ylab("Nível de Ozônio (ppb)")
```



FACULDADE DE TECNOLOGIA

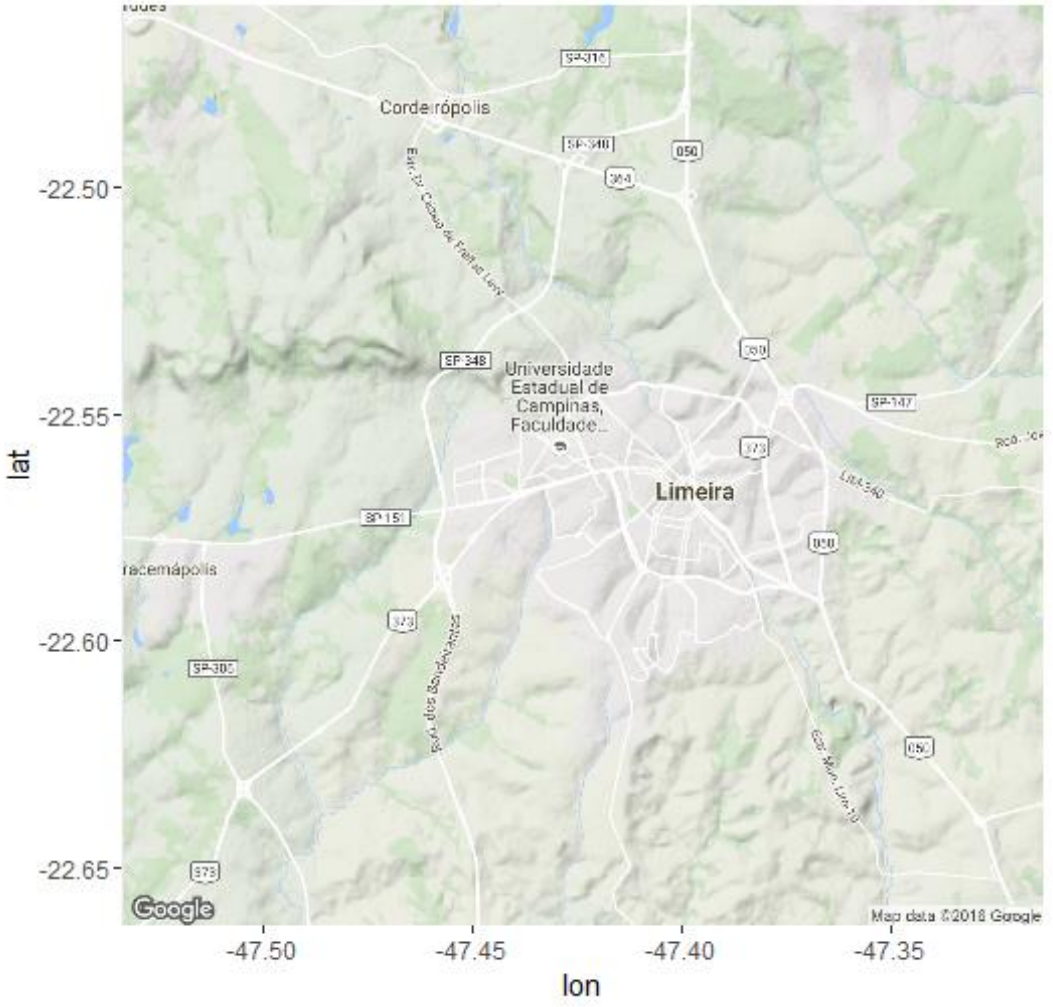
```
ggplot(qualidadeArMes, aes(TempC, Ozonio)) +
    geom_point() +
    xlab("Temperatura (ºC)") + ylab("Nível de Ozônio (ppb)") +
    facet_grid(. ~ Mes)
```



Mapas

```
library(ggmap)
library(mapproj)

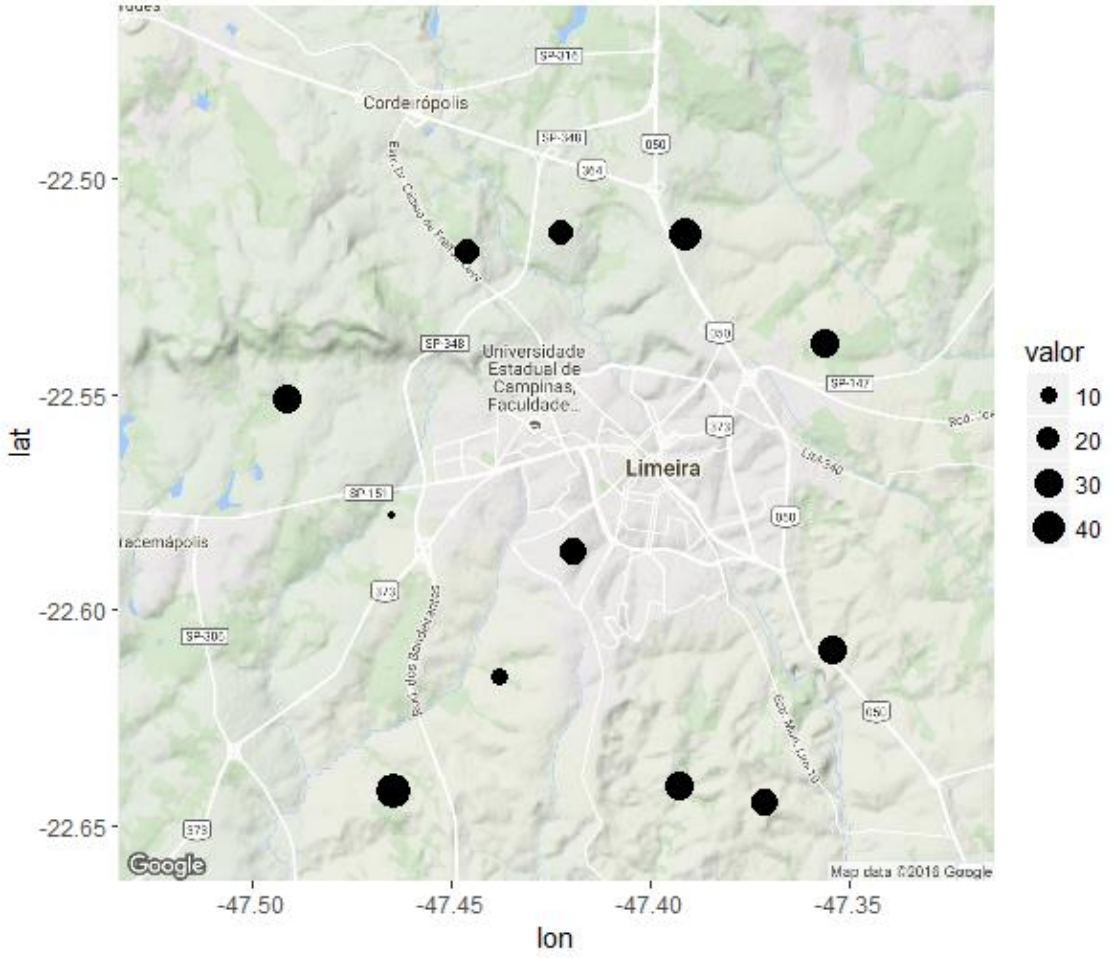
map <- get_map(location = c(lon = -47.42392, lat =-22.56127),
               maptype = "terrain", zoom = 12,
               language = "pt-BR", scale = 2)
ggmap(map)
```



FACULDADE DE TECNOLOGIA

```
ggmap(map) + geom_point(data = dados, aes(lon, lat, size = valor))
```

	lat	lon	valor
1	-22.51295	-47.39125	42.636056
2	-22.58680	-47.41915	25.998992
3	-22.64494	-47.37110	27.555489
4	-22.57798	-47.46489	4.471797
5	-22.61540	-47.43786	11.246833
6	-22.60929	-47.35441	33.873861
7	-22.51700	-47.44640	24.961033
8	-22.64081	-47.39284	32.125557
9	-22.51261	-47.42283	23.198579
10	-22.53811	-47.35632	33.214911
11	-22.64202	-47.46443	48.854826
12	-22.55107	-47.49127	32.966208



Nuvens de palavras

wordcloud {wordcloud} R Documentation

Plot a word cloud

Description

Plot a word cloud

Usage

```
wordcloud(words,freq,scale=c(4,.5),min.freq=3,max.words=Inf,
        random.order=TRUE, random.color=FALSE, rot.per=.1,
        colors="black",ordered.colors=FALSE,use.r.layout=FALSE,
        fixed.asp=TRUE, ...)
```

FACULDADE DE TECNOLOGIA

Processamento dos textos (usando pacote tm, *text mining*)

```
library(tm)
library(wordcloud)

# Identificar diretorio com arquivos com texto
corpus <- Corpus(DirSource("noticias"))

# Aplicar funções de conversão do texto
corpus <- tm_map(corpus, PlainTextDocument)
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeWords, stopwords("pt"))

wordcloud(corpus, max.words = 60, colors = c("blue", "red"))
```

Criação da nuvem de palavras

Tipos de gráficos a evitar

How to Display Data Badly

HOWARD WAINER*

Methods for displaying data badly have been developing for many years, and a wide variety of interesting and inventive schemes have emerged. Presented here is a synthesis yielding the 12 most powerful techniques that seem to underlie many of the realizations found in practice. These 12 (the dirty dozen) are identified and illustrated.

KEY WORDS: Graphics; Data display; Data density; Data-ink ratio.

categorized. This article is the beginning of such a compendium.

The aim of good data graphics is to display data accurately and clearly. Let us use this definition as a starting point for ca
definition h
(b) showin
clearly. Thu
three avenu
quence, parse them into some of their component parts, and see if we can identify means for measuring the success of each strategy.

WAINER, H. How to display data badly. **The American Statistician**, v. 38, n. 2, p. 137–147, 1984.

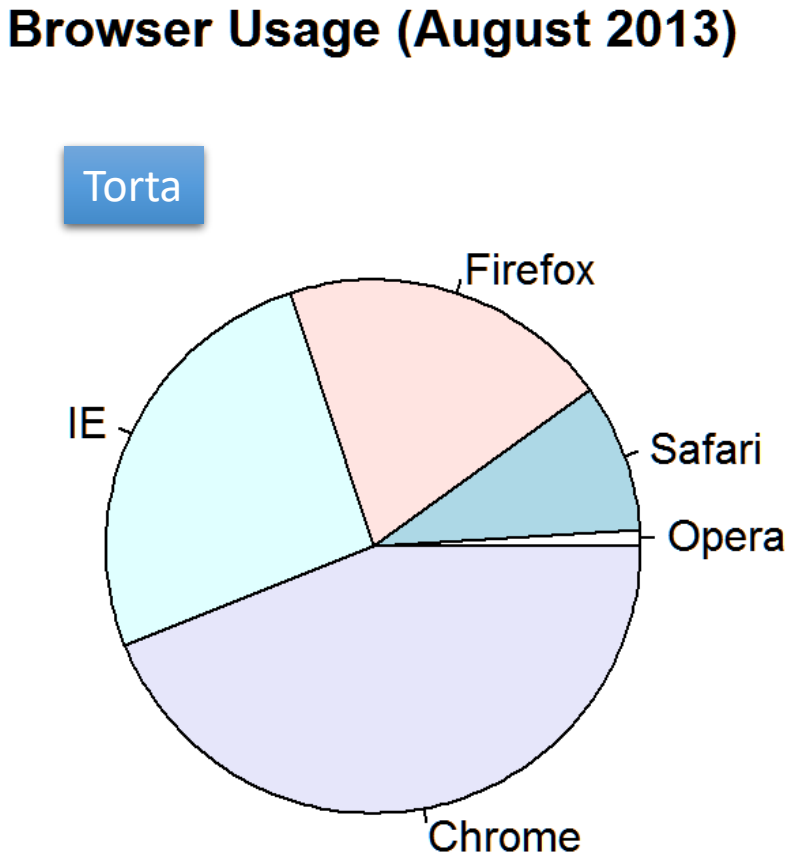
Notas de aula: Statistics and R for the Life Sciences, Rafael Irizarry

FACULDADE DE TECNOLOGIA

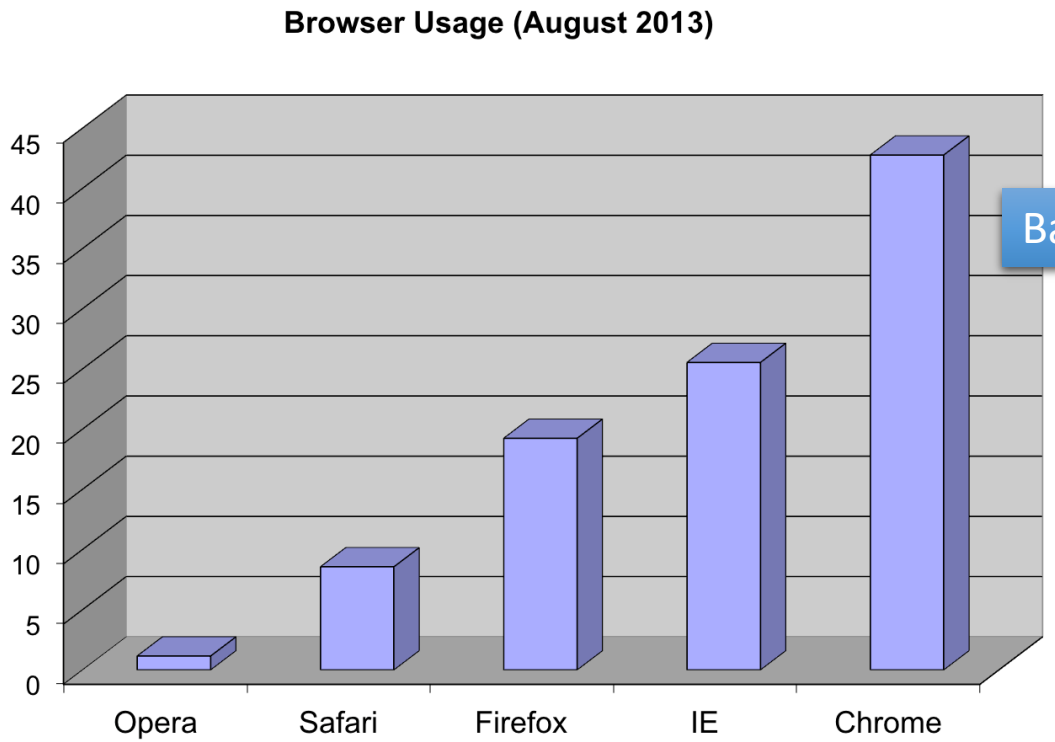
Princípios gerais de gráficos ruins

1. Mostre o mínimo de informação
2. Obscureça o que você apresenta
3. Use cores e pseudo-3D gratuitamente
4. Faça um gráfico de torta (de preferência, colorido e em 3D)
5. Use uma escala escolhida pobremente
6. Ignore algarismos significativos

Problema: gráficos com informação visual pouco adequada à percepção/ comparação pelo olho humano

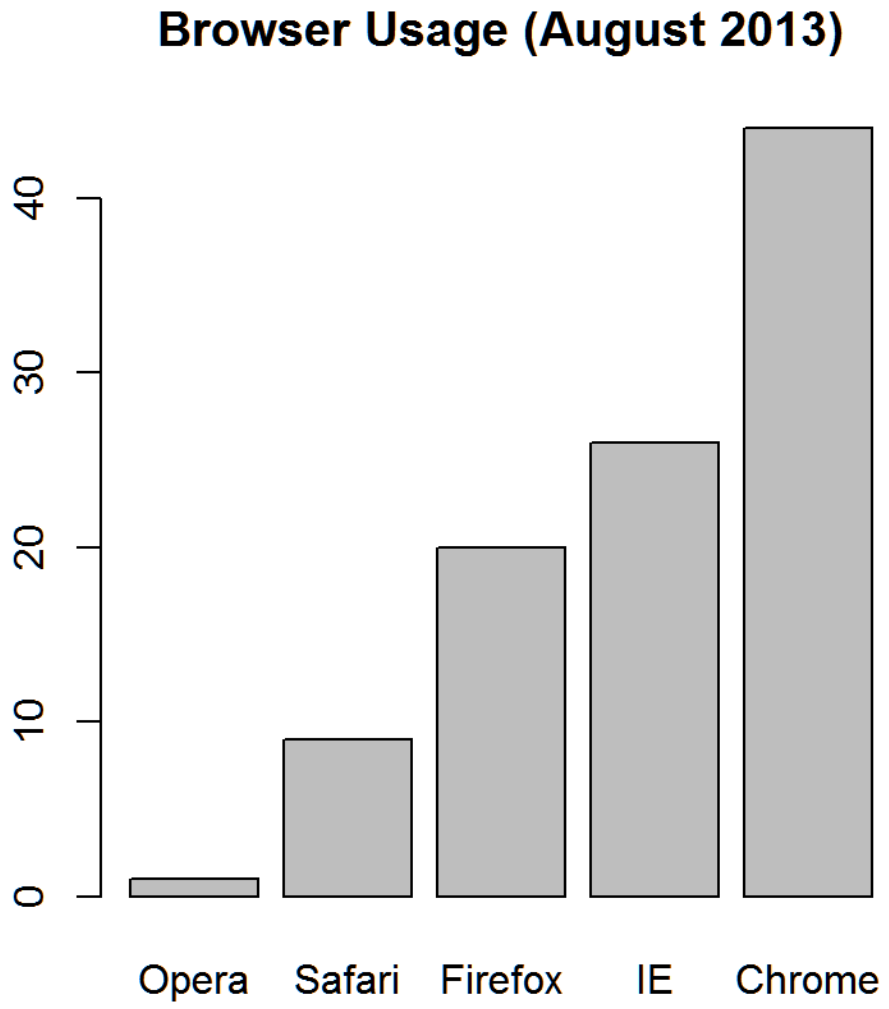


Torta



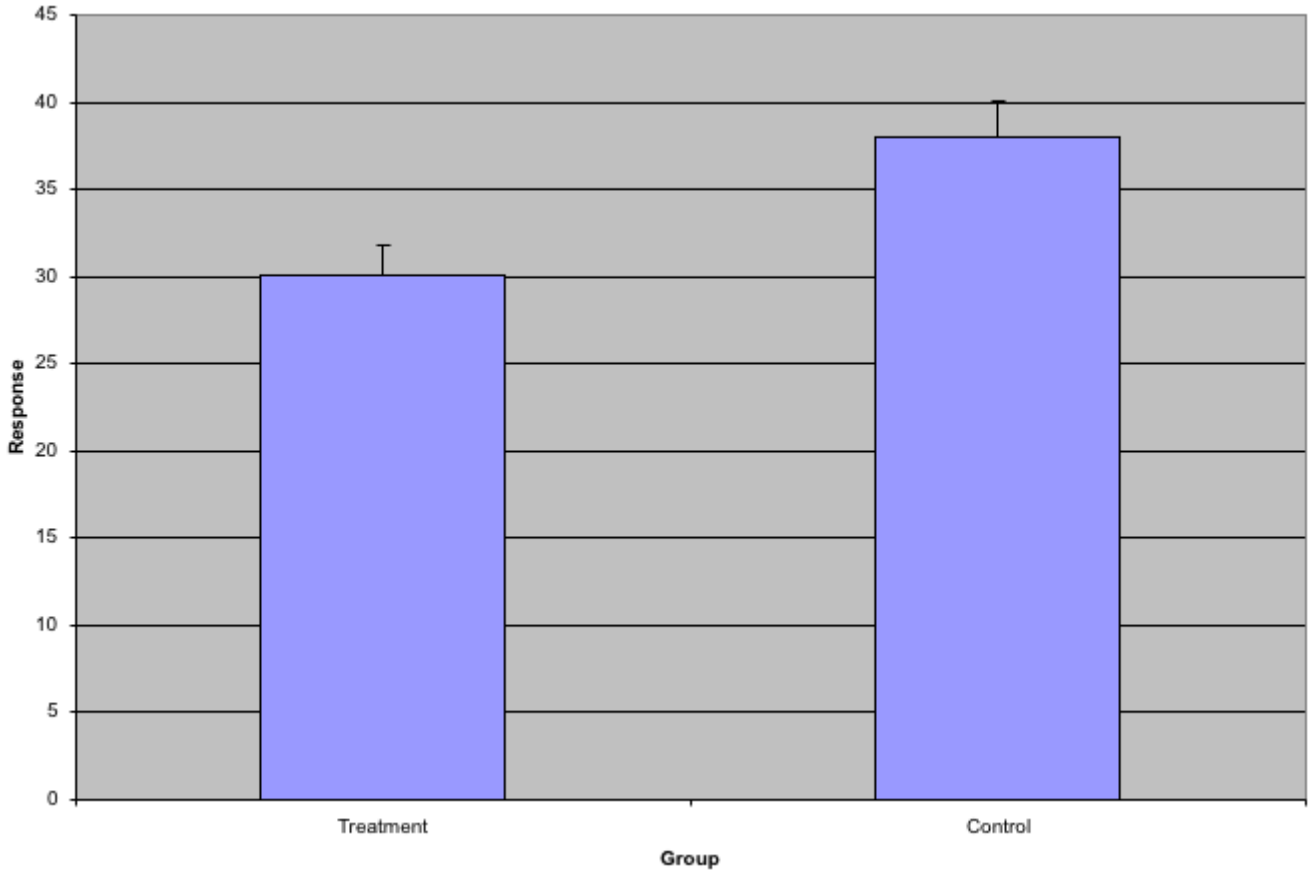
Barras 3D

Donut


Browser Usage (August 2013)
0
10
20
30
40
Opera
Safari
Firefox
IE
Chrome

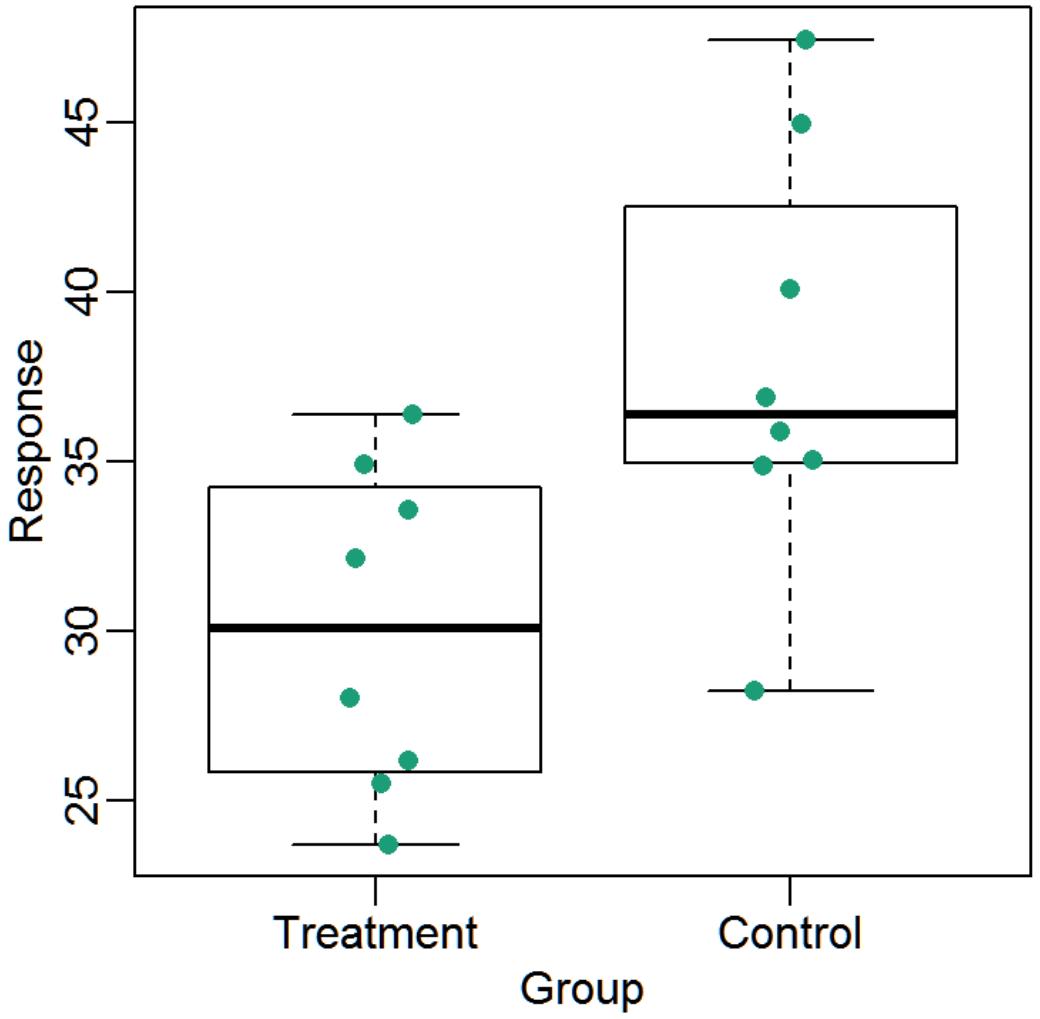
FACULDADE DE TECNOLOGIA

Problema: gráfico de barras pouco informativo



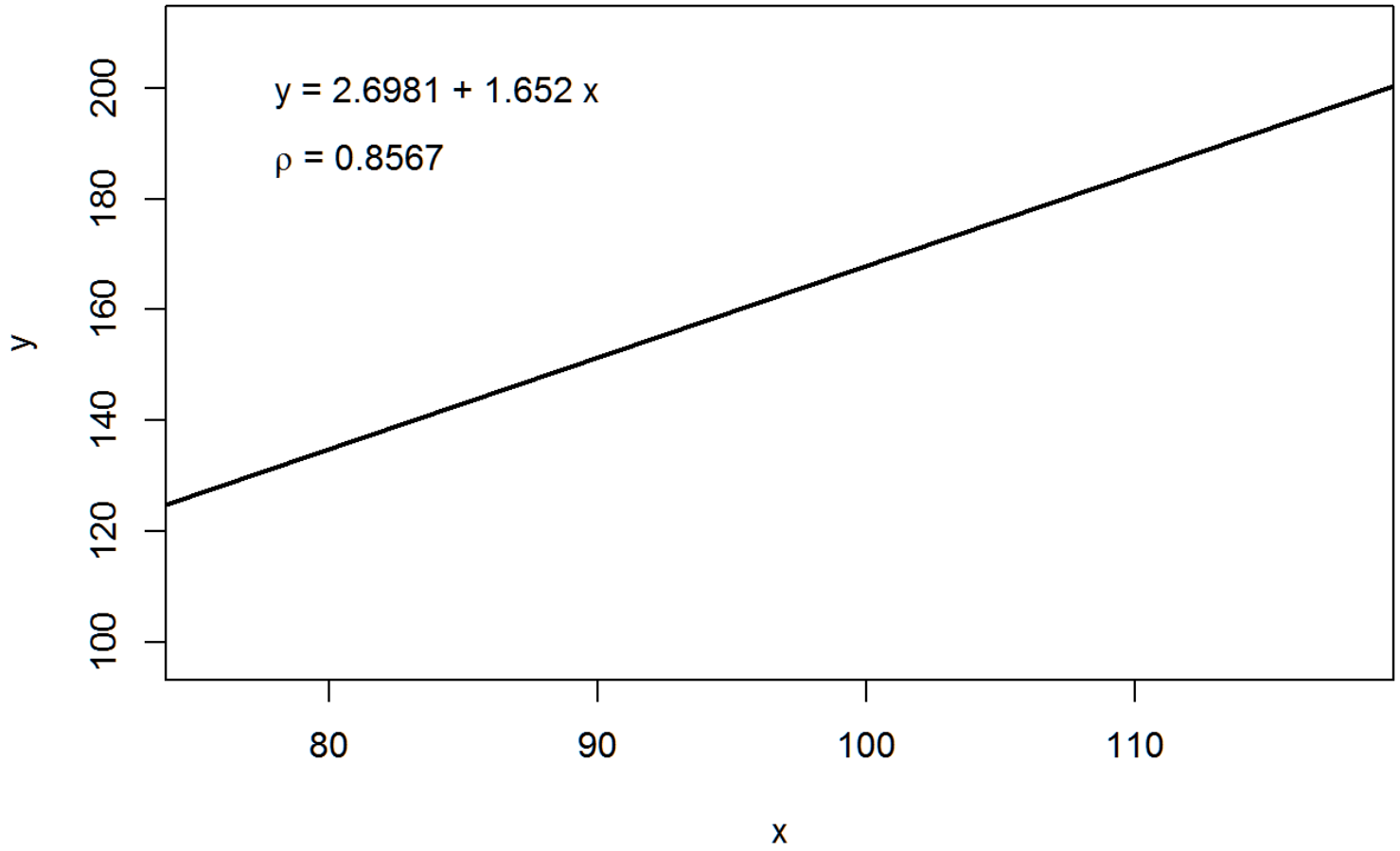
FACULDADE DE TECNOLOGIA

Mesmos dados,
mais informação



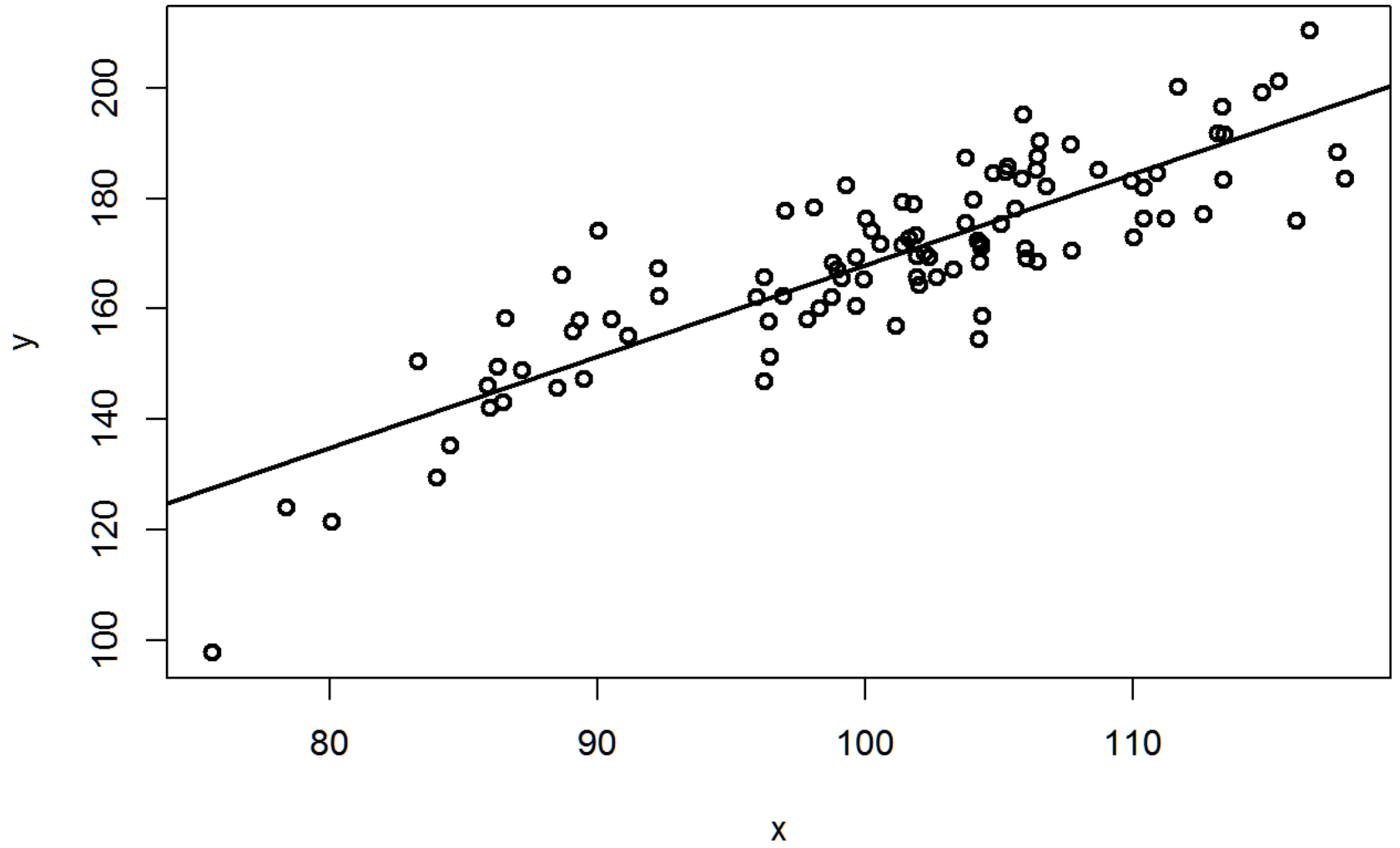
FACULDADE DE TECNOLOGIA

Problema: mostrar apenas a regressão



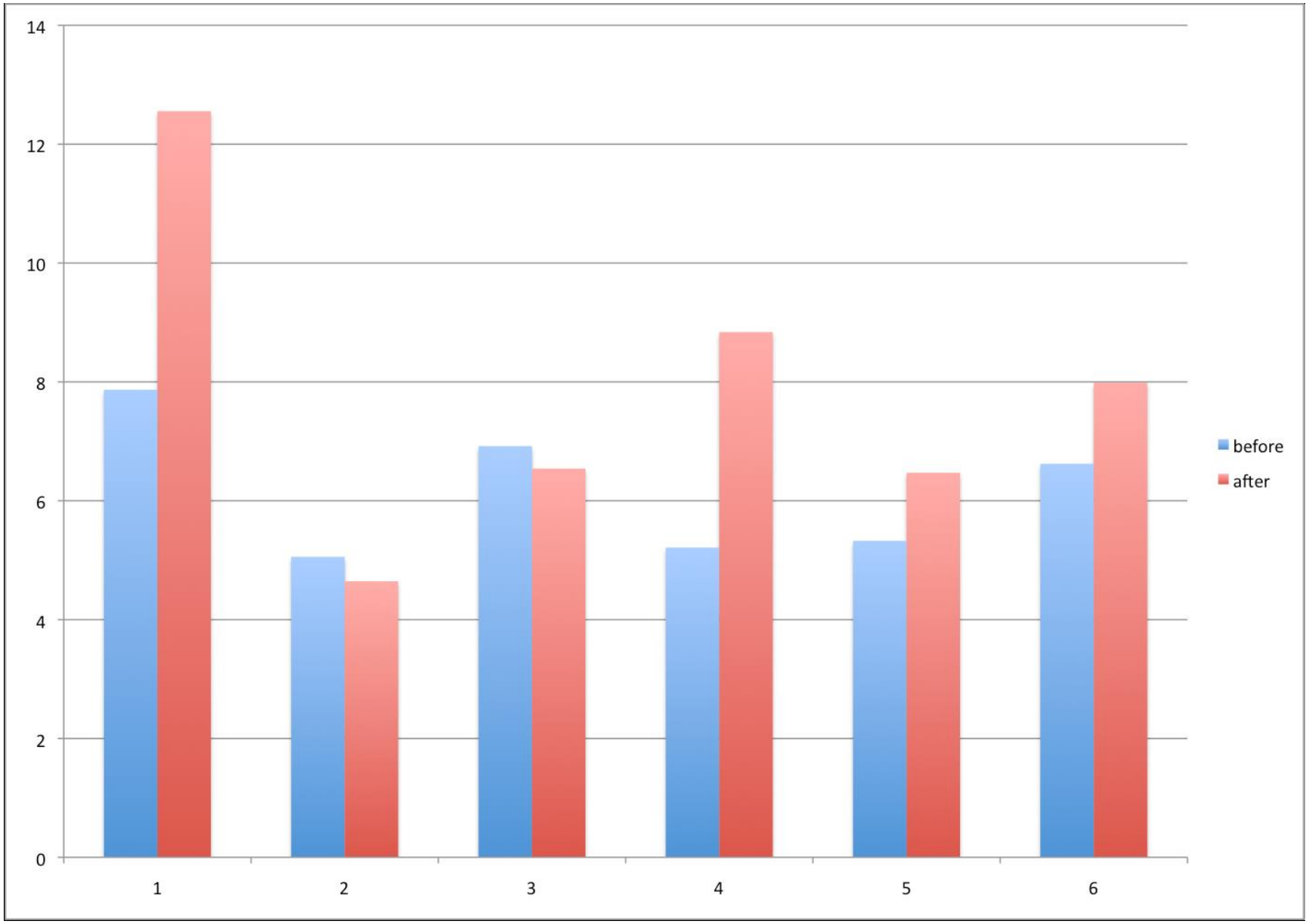
FACULDADE DE TECNOLOGIA

Mostrar os pontos originais



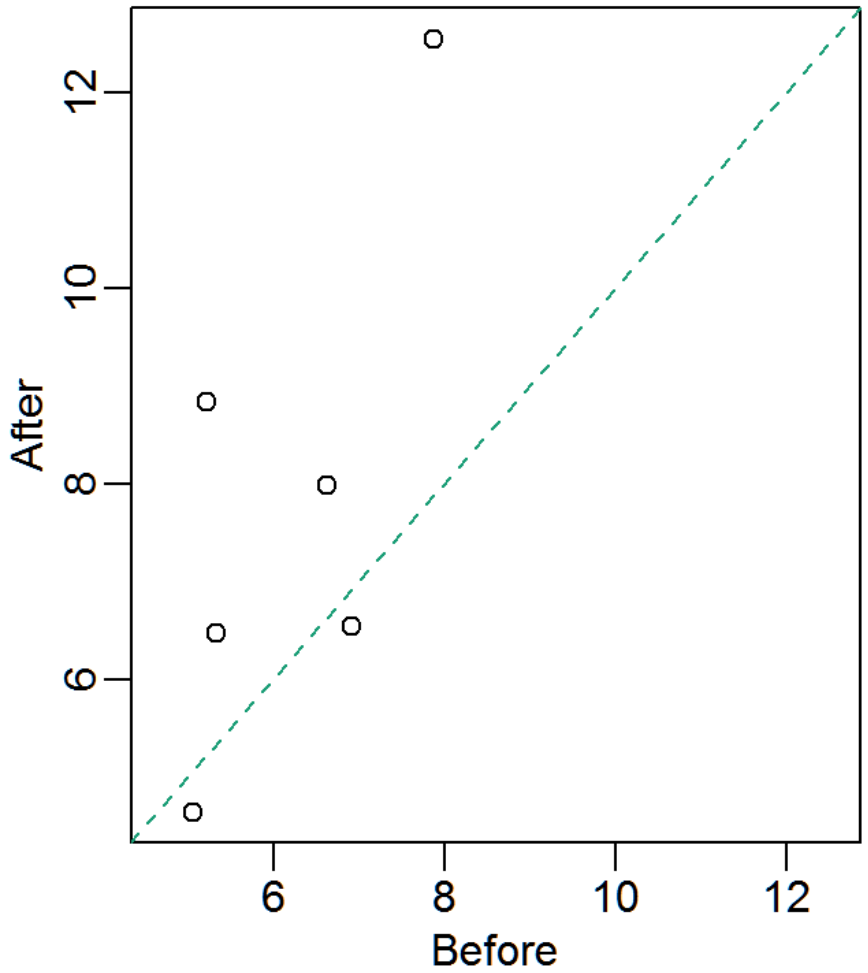
FACULDADE DE TECNOLOGIA

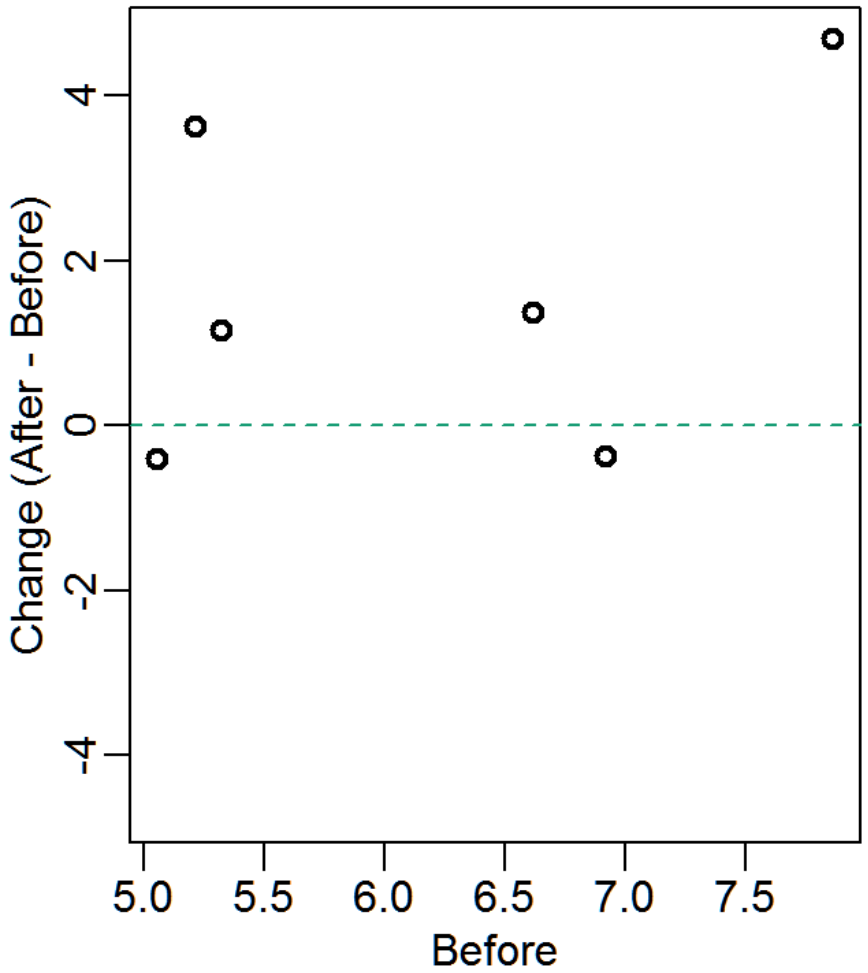
Problema: comparação entre dois grupos (antes & depois, controle & tratamento) usando gráficos de barras



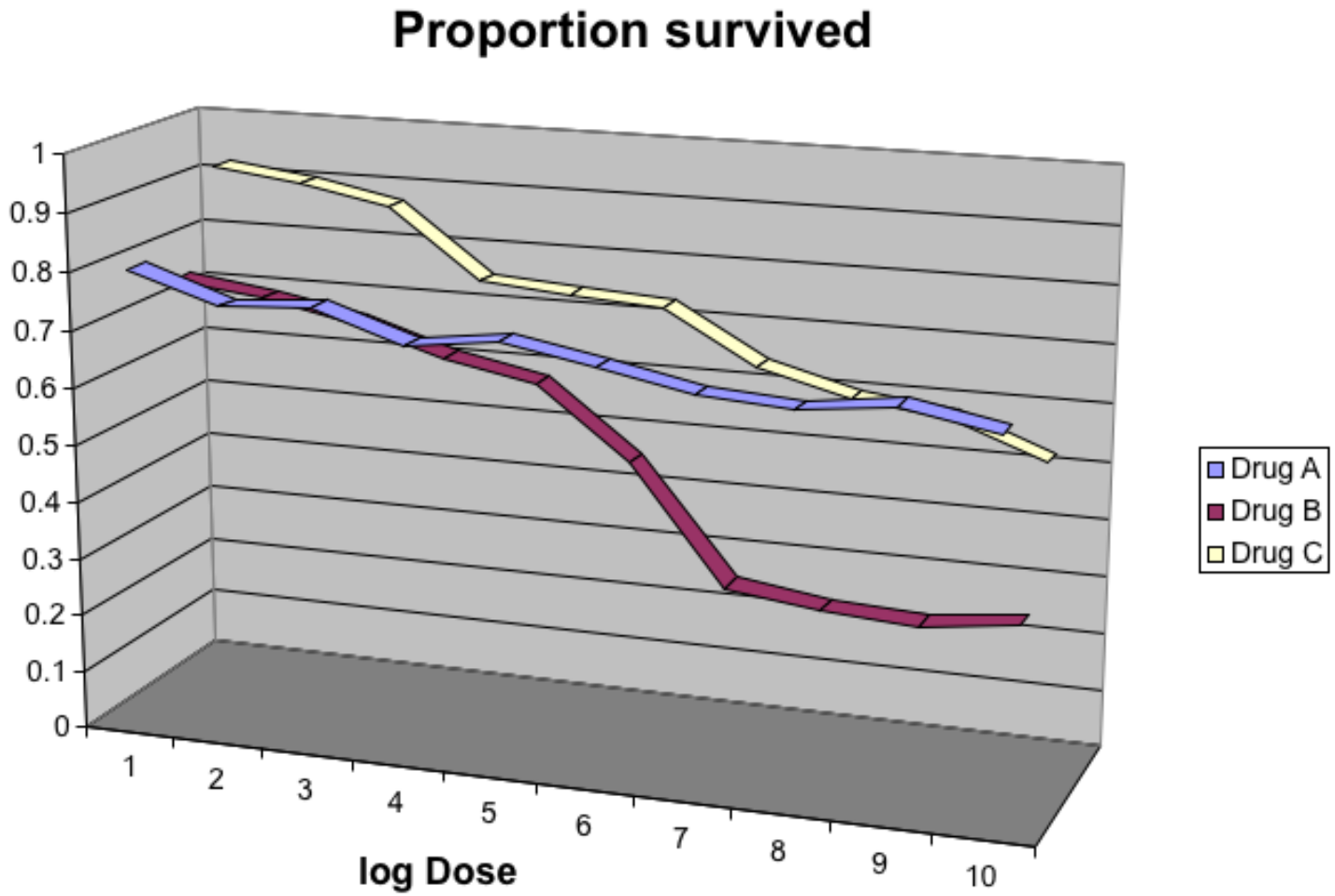
FACULDADE DE TECNOLOGIA

Alternativas melhores



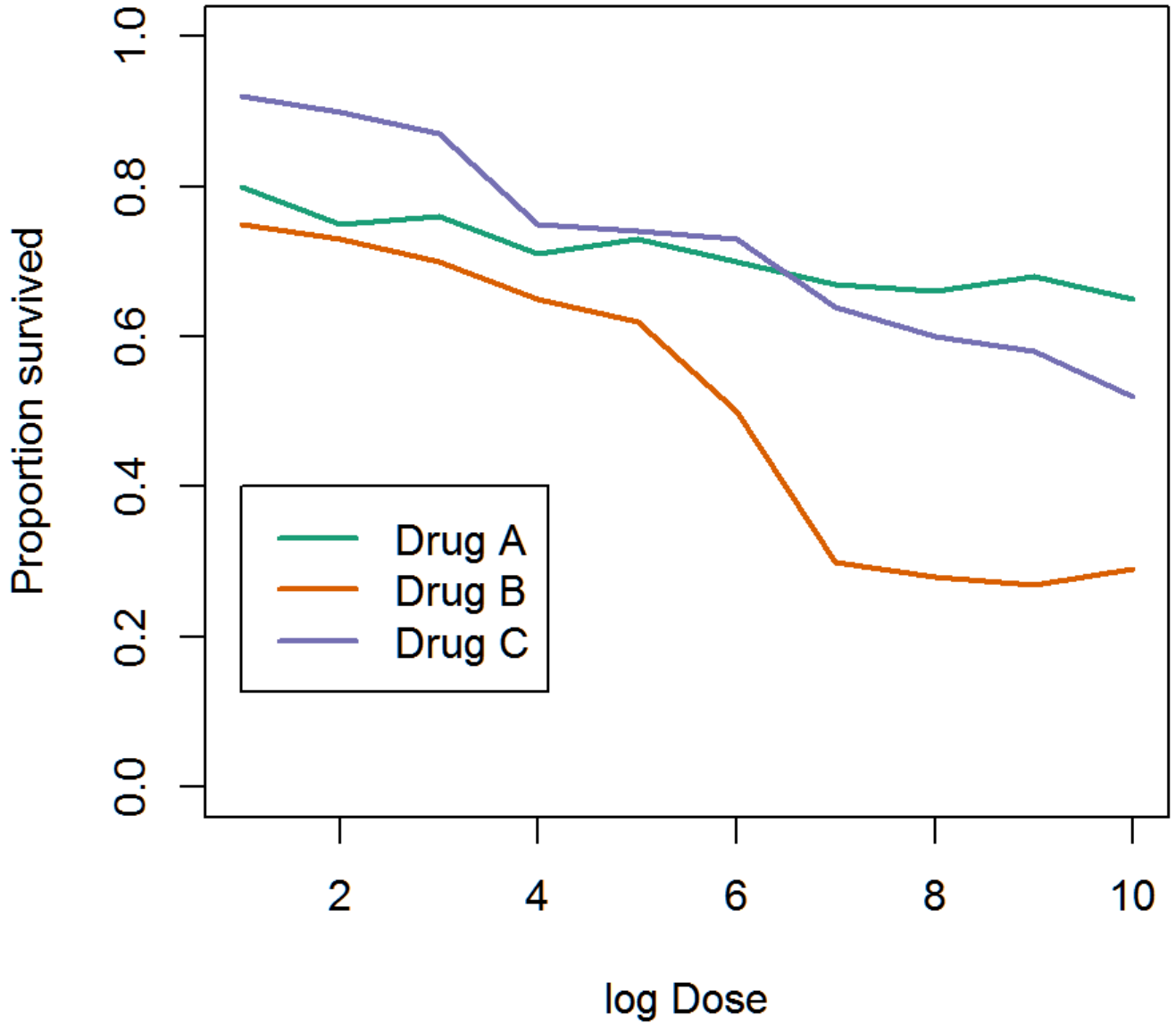


Problema: 3D usado gratuitamente



FACULDADE DE TECNOLOGIA

Alternativa



FACULDADE DE TECNOLOGIA

Problema: dígitos significativos em excesso (com pouco significado)

Tabela 1. Altura de jogadores de basquetebol

```
##                SG       PG       C       PF       SF
## team 1 76.39843 76.21026 81.68291 75.32815 77.18792
## team 2 74.14399 71.10380 80.29749 81.58405 73.01144
## team 3 71.51120 69.02173 85.80092 80.08623 72.80317
## team 4 78.71579 72.80641 81.33673 76.30461 82.93404
## team 5 73.42427 73.27942 79.20283 79.71137 80.30497
## team 6 72.93721 71.81364 77.35770 81.69410 80.39703
## team 7 68.37715 73.01345 79.10755 71.24982 77.19851
## team 8 73.77538 75.59278 82.99395 75.57702 87.68162
```

FACULDADE DE TECNOLOGIA

Alternativa:

```
round(heights,1)
```

```
##             SG   PG    C   PF   SF
## team 1 76.4 76.2 81.7 75.3 77.2
## team 2 74.1 71.1 80.3 81.6 73.0
## team 3 71.5 69.0 85.8 80.1 72.8
## team 4 78.7 72.8 81.3 76.3 82.9
## team 5 73.4 73.3 79.2 79.7 80.3
## team 6 72.9 71.8 77.4 81.7 80.4
## team 7 68.4 73.0 79.1 71.2 77.2
## team 8 73.8 75.6 83.0 75.6 87.7
```

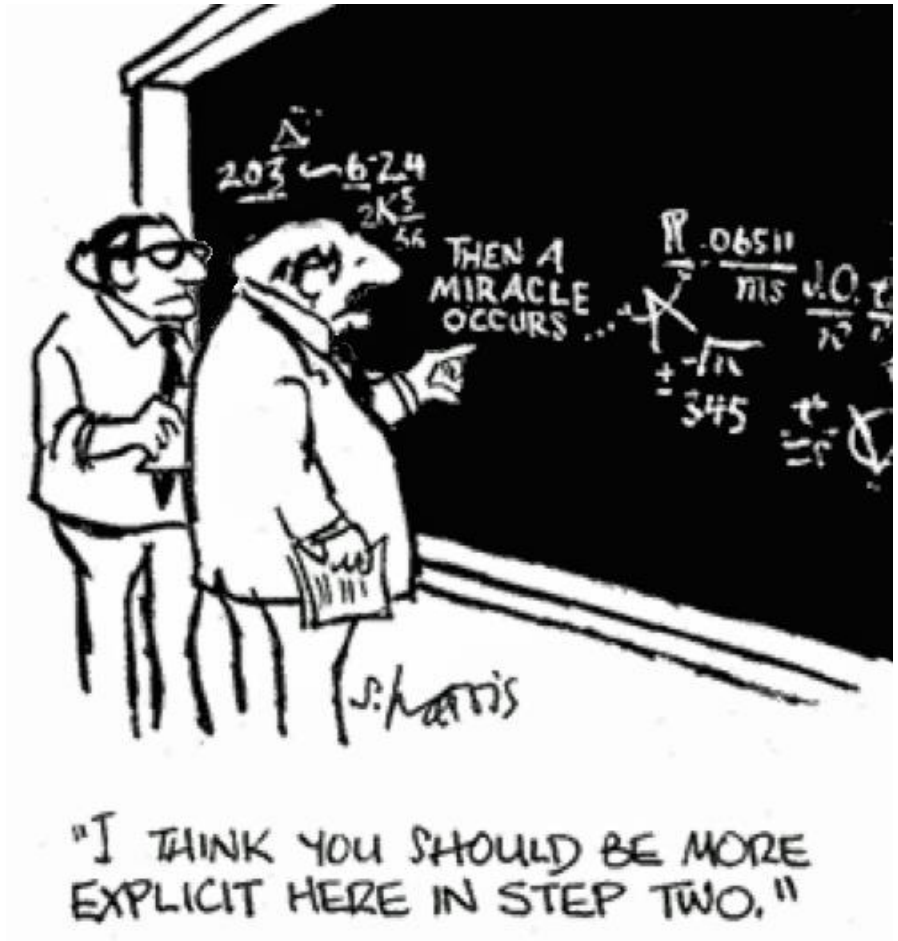
Princípios a serem seguidos

1. Seja preciso e claro
2. Deixe que os dados falem
3. Mostre tanta informação quanto possível, sem obscurecer a mensagem
4. Ciência não é marketing, evite firulas desnecessárias
5. Em tabelas, todos os dígitos devem ser significativos; não deixe de apresentar os 0's finais

FACULDADE DE TECNOLOGIA

Sumário: mais gráficos

- Criação de gráficos com ggplot2
  - Gráficos rápidos com qplot()
  - Gráficos em camadas com ggplot()
    - geom_point(), geom_bin2d(), geom_density_2d(), geom_boxplot()
    - geom_smooth(), ggtitle(), xlab(), ylab(), facet_grid()
  - Mapas com ggmap
- Criação de nuvens de palavras
  - wordcloud
- Recomendações gerais sobre apresentação gráfica



Pesquisa reproduzível em R

O que é pesquisa

Research

From Wikipedia, the free encyclopedia

Research comprises "creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of humans, culture and society, and the use of this stock of knowledge to devise new applications."[1] It is used to establish or confirm facts, reaffirm the results of previous work, solve new or existing problems, support theorems, or develop new theories. A research project may also be an expansion on past work in the field. To test the validity of instruments, procedures, or experiments, research may replicate elements of prior projects, or the project as a whole. The primary

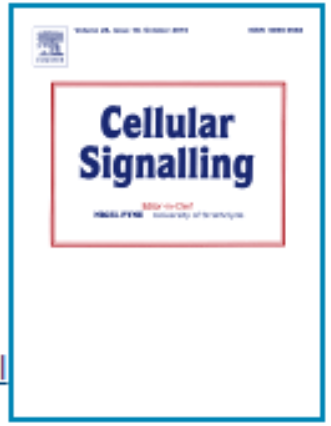
http://phdcomics.com/

FACULDADE DE TECNOLOGIA

Biologist loses second paper — again, for unvalidated figures

without comments

A researcher at Case Western Reserve University in Ohio has retracted a second paper after a review found the figures didn't match the original data.



Last year, we reported on a previous retraction of a paper co-authored by biologist Alan Levine in *Inflammatory Bowel Diseases*, which was pulled for the exact same reason; even the retraction notices use similar language. The first author on both papers is Debasmita Mandal, also listed at Case Western Reserve University.

Here's the retraction notice for "REDOX regulation of IL-13 signaling in intestinal epithelial cells: usage of alternate pathways mediates distinct gene expression patterns," published by *Cellular Signalling*:

> This article has been retracted at the request of the corresponding author and the authors' institution. Case Western Reserve University. In a formal university process, the institution reviewed the data and figures associated with this article and concluded that the figures cannot be validated by original data.

http://retractionwatch.com/2016/09/08/biologist-loses-second-paper-again-for-unvalidated-figure

FACULDADE DE TECNOLOGIA

7th retraction for Ohio researcher who manipulated dozens of figures

with 3 comments

A pharmacology researcher at Ohio State University has added his seventh retraction, four years after a finding of misconduct by the U.S. Office of Integrity (ORI).

An analysis of the work of Terry Elton determined that he had

> falsified and/or fabricated Western blots in eighteen (18) figures and in six (6) published papers.

Terry Elton

In 2012, the ORI finding, which resulted in a three-year funding ban (that is now complete), recommended that Elton retract all six papers, one of which had already been retracted at the time of the report.

Four years later, the last of the six papers flagged by the ORI has finally been retracted by *Molecular and Cellular Endocrinology*.

http://retractionwatch.com/2016/08/24/7th-retraction-for-ohio-researcher-who-manipulated-dozens-of-figures

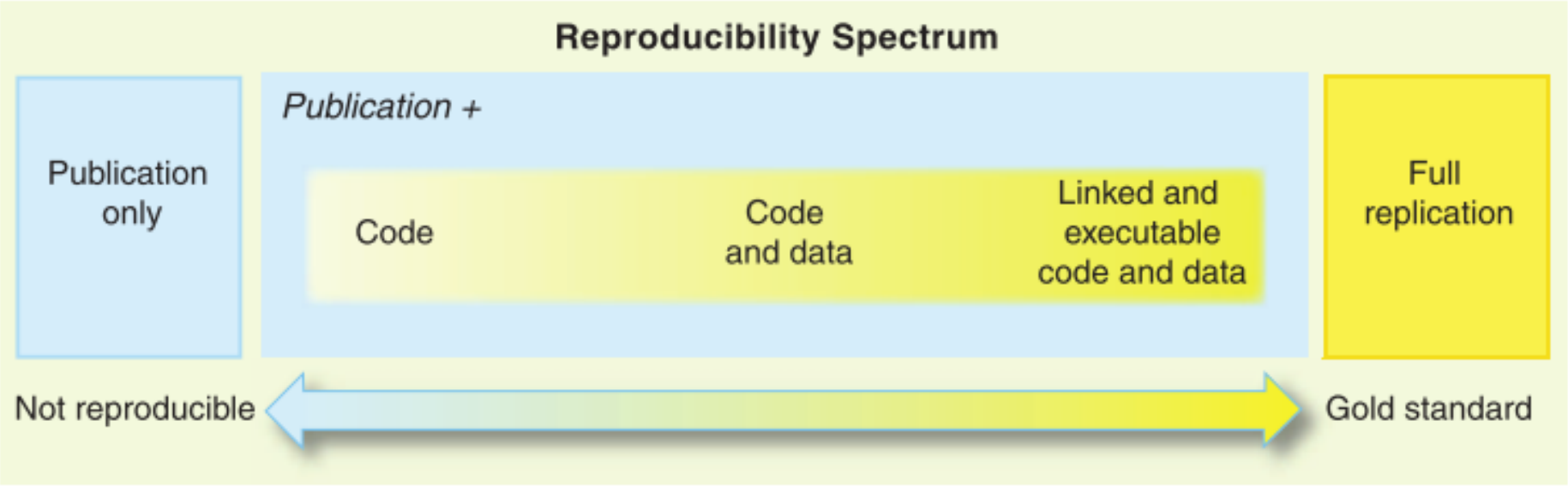
A pesquisa reproduzível

Reproducible research [edit]

See also: Open research computation

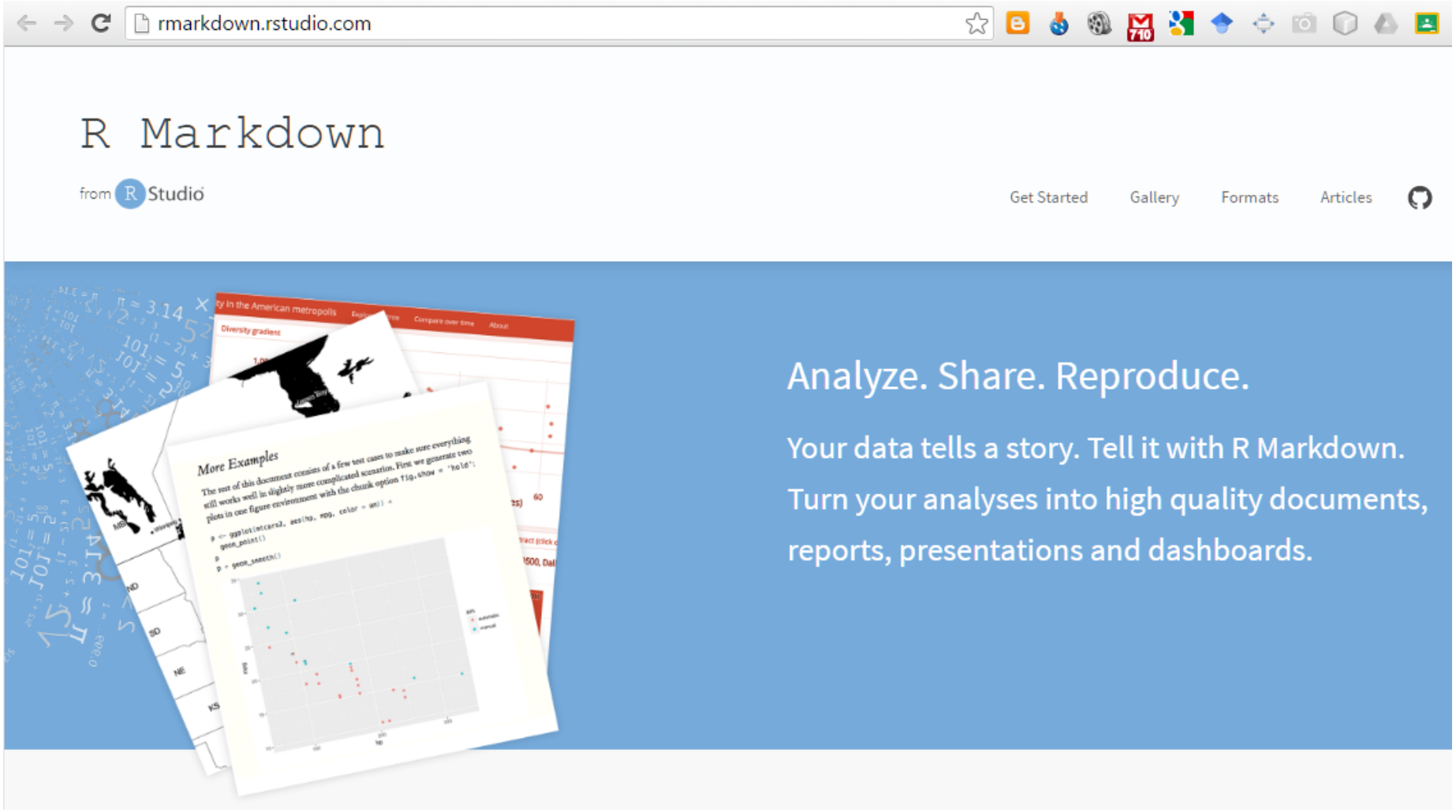
The term *reproducible research* refers to the idea that the ultimate product of academic research is the paper along with the full computational environment used to produce the results in the paper such as the code, data, etc. that can be used to reproduce the results and create new work based on the research.[8][9][10][11]

https://en.wikipedia.org/wiki/Reproducibility#Reproducible_research



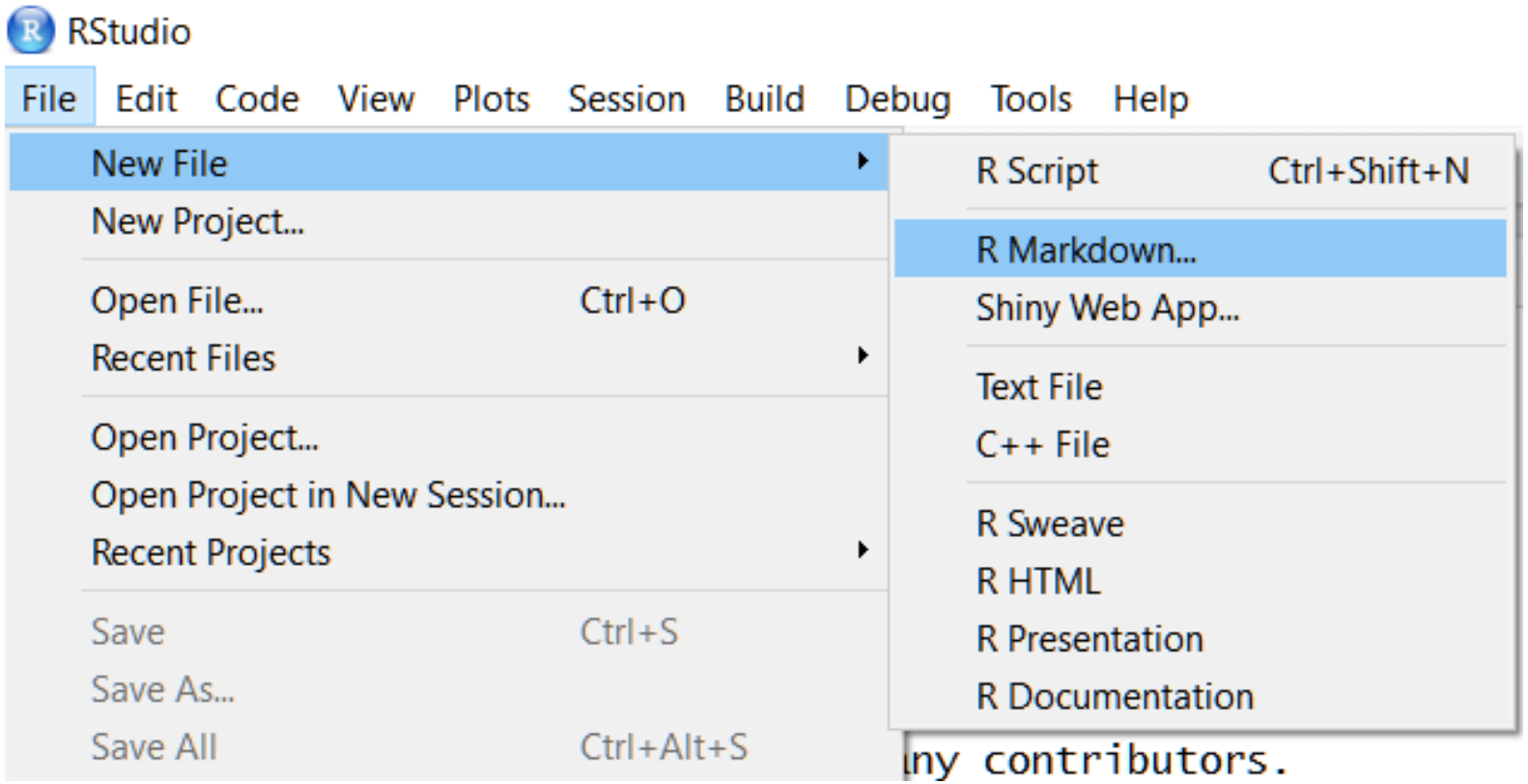
Peng RD. Reproducible research in computational science. Science (80-) [Internet]. 2011;334(6060):1226–7. Available from: http://arxiv.org/abs/0901.4552

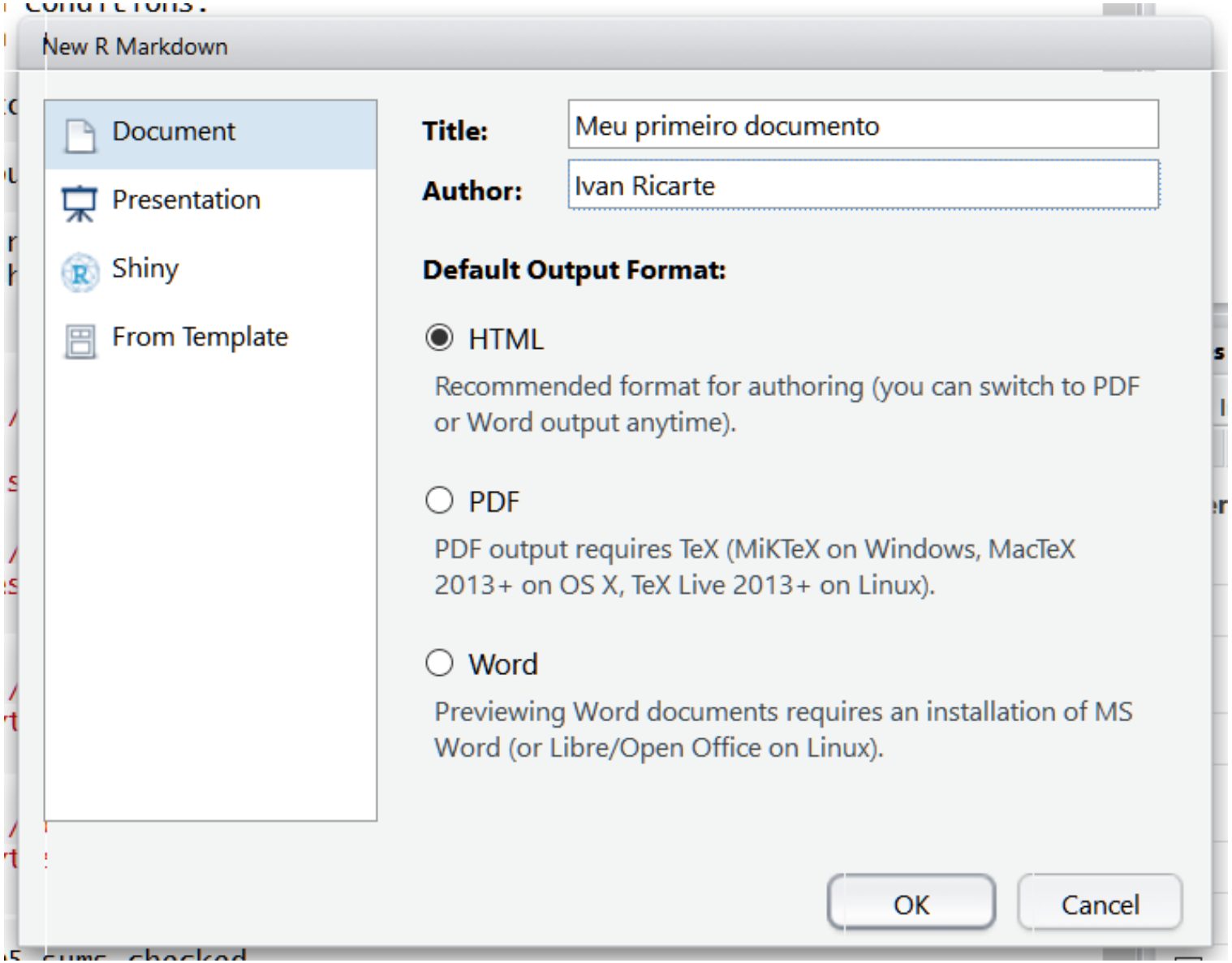
Divulgando pesquisa reproduzível em R



FACULDADE DE TECNOLOGIA

```
> install.packages("rmarkdown")
```




New R Markdown
Document
Presentation
Shiny
From Template
Title:
Meu primeiro documento
Author:
Ivan Ricarte
Default Output Format:
HTML
Recommended format for authoring (you can switch to PDF or Word output anytime).
PDF
PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).
Word
Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).
OK
Cancel

FACULDADE DE TECNOLOGIA

Untitled1* ×

Knit HTML | Run

```
---
title: "Meu primeiro documento"
author: "Ivan Ricarte"
date: "29 de setembro de 2016"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

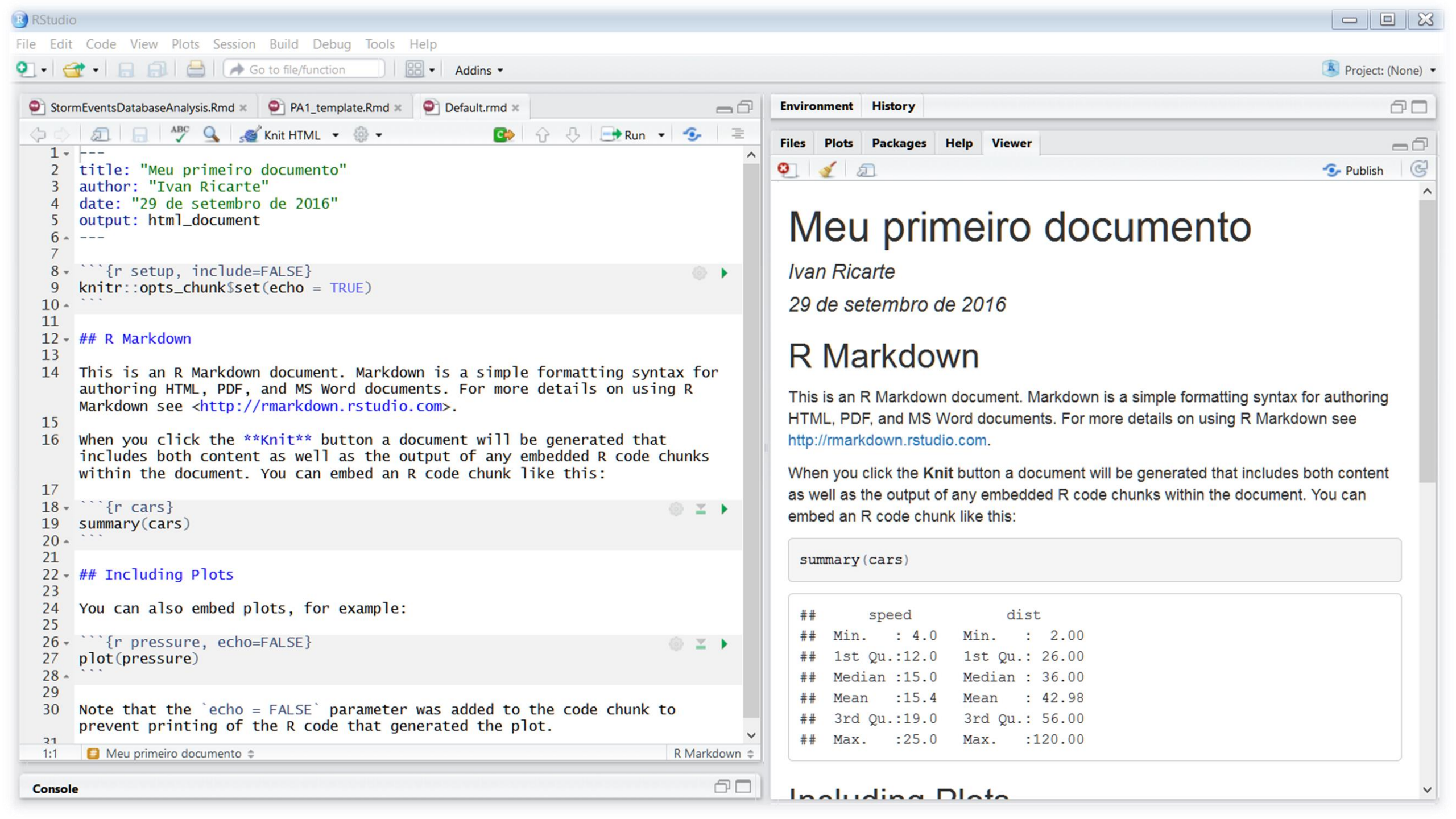
```{r cars}
summary(cars)
```

Including Plots

You can also embed plots, for example:

```{r pressure, echo=FALSE}
plot(pressure)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
```


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Addins
Project: (None)
StormEventsDatabaseAnalysis.Rmd
PA1_template.Rmd
Default.rmd
Knit HTML
Run
1 ---
2 title: "Meu primeiro documento"
3 author: "Ivan Ricarte"
4 date: "29 de setembro de 2016"
5 output: html_document
6 ---
7
8 ```{r setup, include=FALSE}
9 knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
15
16 When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
17
18 ```{r cars}
19 summary(cars)
20 ```
21
22 ## Including Plots
23
24 You can also embed plots, for example:
25
26 ```{r pressure, echo=FALSE}
27 plot(pressure)
28 ```
29
30 Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
31
1:1
Meu primeiro documento
R Markdown
Console
Environment
History
Files
Plots
Packages
Help
Viewer
Publish
Meu primeiro documento
Ivan Ricarte
29 de setembro de 2016
R Markdown
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.
When you click the Knit button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
summary(cars)
speed dist
Min. : 4.0 Min. : 2.00
1st Qu.:12.0 1st Qu.: 26.00
Median :15.0 Median : 36.00
Mean :15.4 Mean : 42.98
3rd Qu.:19.0 3rd Qu.: 56.00
Max. :25.0 Max. :120.00

FACULDADE DE TECNOLOGIA

Default.rmd

```
1  ---
2  title: "Meu primeiro documento"
3  author: "Ivan Ricarte"
4  date: "29 de setembro de 2016"
5  output: html_document
6  ---
```

C:/Users/Ivan/OneDrive/Apresentações/Minicurso R/Recursos/Default.html

Default.html | Open in Browser | Find

Meu primeiro documento

Ivan Ricarte

29 de setembro de 2016

FACULDADE DE TECNOLOGIA

```
7
8  ```{r setup, include=FALSE}
9  knitr::opts_chunk$set(echo = TRUE)
10 ```
11
12 ## R Markdown
13
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word
   documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
15
16 When you click the **Knit** button a document will be generated that includes both content as well as the
   output of any embedded R code chunks within the document. You can embed an R code chunk like this:
17
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

FACULDADE DE TECNOLOGIA

````
output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```
````

embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

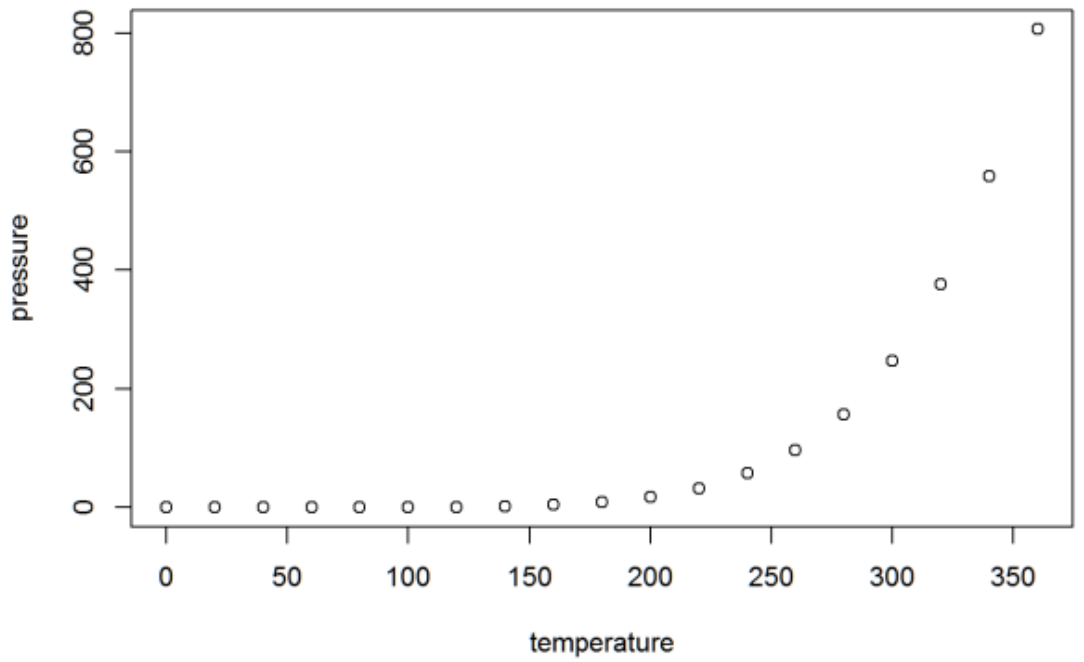
```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

FACULDADE DE TECNOLOGIA

```
21
22  ## Including Plots
23
24  You can also embed plots, for example:
25
26  ```{r pressure, echo=FALSE}
27  plot(pressure)
28  ```
29
30  Note that the `echo = FALSE` parameter was ad
    generated the plot.
31
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

FACULDADE DE TECNOLOGIA

Syntax	Becomes
`Plain text`	Plain text
`End a line with two spaces` `to start a new paragraph.`	End a line with two spaces to start a new paragraph.
`*italics* and _italics_`	*italics* and *italics*
`**bold** and __bold__`	**bold** and **bold**
`superscript^2^`	superscript^{2}
`~~strikethrough~~`	~~strikethrough~~
`[link](www.rstudio.com)`	link

https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf

FACULDADE DE TECNOLOGIA

Syntax	Becomes
`# Header 1`	# Header 1
`## Header 2`	## Header 2
`### Header 3`	### Header 3
`#### Header 4`	#### Header 4
`##### Header 5`	##### Header 5
`###### Header 6`	###### Header 6
`endash: --`	endash: –
`emdash: ---`	emdash: —
`ellipsis: ...`	ellipsis: …

FACULDADE DE TECNOLOGIA

Syntax

```
inline equation: $A = \pi*r^{2}$

image: ![](path/to/smallorb.png)

horizontal rule (or slide break):

***

> block quote

 * unordered list
 * item 2
     + sub-item 1
     + sub-item 2

 1. ordered list
 2. item 2
     + sub-item 1
     + sub-item 2

 Table Header   | Second Header
 -------------- | -------------
 Table Cell     | Cell 2
 Cell 3         | Cell 4
```

Becomes

inline equation: $A = \pi * r^2$

image: 

horizontal rule (or slide break):

> block quote

- unordered list
- item 2
  - sub-item 1
  - sub-item 2

1. ordered list
2. item 2
   - sub-item 1
   - sub-item 2

Table Header	Second Header
Table Cell	Cell 2
Cell 3	Cell 4

FACULDADE DE TECNOLOGIA

Syntax	Becomes
Make a code chunk with three back ticks followed by an r in braces. End the chunk with three back ticks: ```{r} paste("Hello", "World!") ```	Make a code chunk with three back ticks followed by an r in braces. End the chunk with three back ticks: `paste("Hello", "World!")` `## [1] "Hello World!"`
Place code inline with a single back ticks. The first back tick must be followed by an R, like this `r paste("Hello", "World!")`.	Place code inline with a single back ticks. The first back tick must be followed by an R, like this Hello World!.
Add chunk options within braces. For example, `echo=FALSE` will prevent source code from being displayed: ```{r eval=TRUE, echo=FALSE} paste("Hello", "World!") ```	Add chunk options within braces. For example, `echo=FALSE` will prevent source code from being displayed: `## [1] "Hello World!"`

FACULDADE DE TECNOLOGIA

Impact of severe weather events: Analysis of the NOAA Storm Events Database

Ivan L M Ricarte

April 21, 2015

This document presents an analysis on the impact of severe weather events based on the U.S. National Oceanic and Atmospheric Administration (NOAA) Storm Events Database. The current version of this database, as well as its documentation, can be downloaded from the NOAA Storm Events database site. This analysis was performed based on a mirror of this database, made available by instructors of the Reproducible Research course, with data from January 1950 to November 2011. This mirror dataset can be downloaded here.

The goal of this analysis is to answer two basic questions on the impact of severe weather events across the United States. The first part of the analysis aims to evaluate which type of events are most harmful with respect to population health. The second part addresses the question of which types of events have the greatest economic consequences.

Data processing

The database mirror is provided as a compressed single comma-separated value (CSV) file, named *redpdata-data-StormData.csv.bz2*. In order to process this data, the R language was used. The first step is to get the data file, if necessary, and then read the data to a local data frame:

```
sourcefile <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
zipfile    <- "repdata-data-StormData.csv.bz2"
csvfile    <- "repdata-data-StormData.csv"
if(!file.exists(zipfile)) {
        download.file(sourcefile, zipfile, mode="wb", method="curl")
        }
if(!file.exists(csvfile)) {
        unzip(zipfile, csvfile)
        }
storms <- read.csv(csvfile)
```

http://rpubs.com/ricarte/StormEventsDatabaseAnalysis

FACULDADE DE TECNOLOGIA

Data cleaning

An initial exploration of the dataset reveals that there some typographical r
alternative forms for the same events. This leads to a great number of cat

```
length(unique(storms$EVTYPE))
```

```
## [1] 985
```

Data selection and formatting

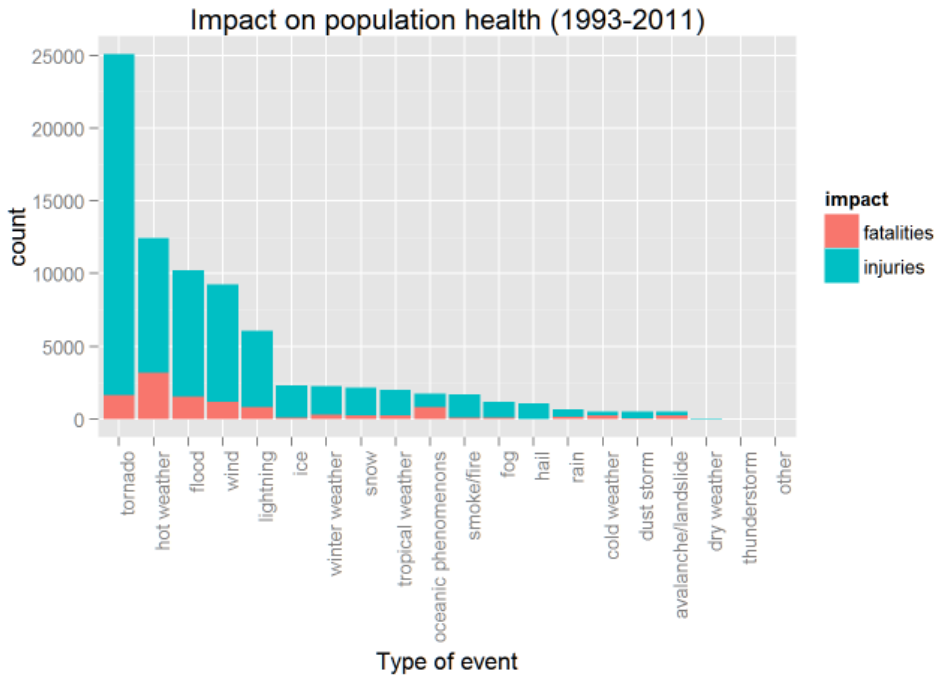
Finally, it is important to evaluate for which years there are enough data
created, only counting the number of distinct events recorded by year:

```
distinctevents <- eventsbyyear %>%
        select(EVTYPE,year) %>%
        group_by(year) %>%
        summarise(count=n())
```

Results

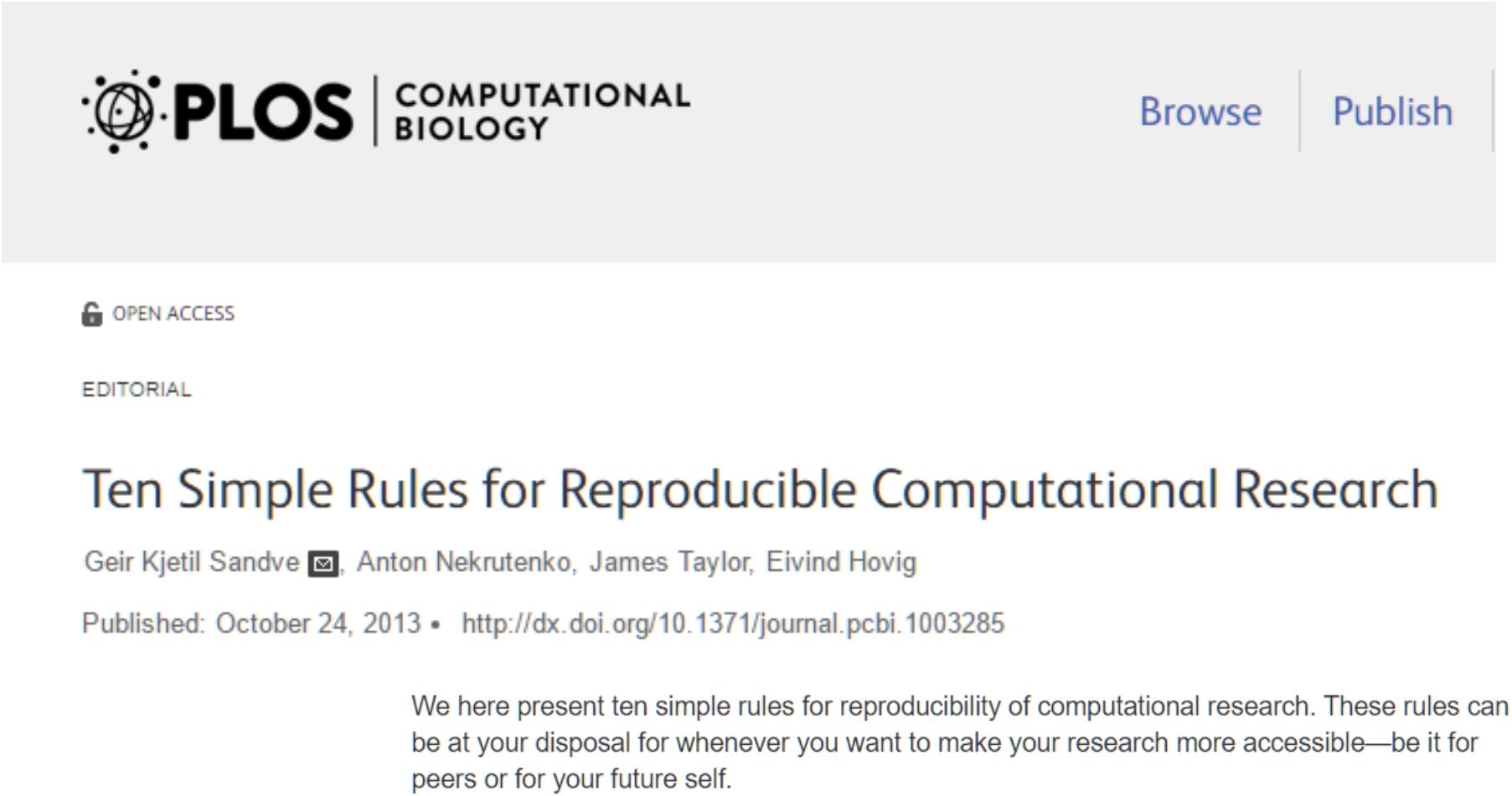
For the period 1993-2011, the weather events that are most harmful with respect to
and wind:

```
ggplot(imp_pop,aes(x=reorder(EVTYPE,-count), y=count, fill=impact)) +
        geom_bar(stat="identity") +
        theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
        ggtitle("Impact on population health (1993-2011)") +
        xlab("Type of event")
```



http://rpubs.com/ricarte/StormEventsDatabaseAnalysis

Dez regras da pesquisa reproduzível



PLOS | COMPUTATIONAL BIOLOGY

Browse | Publish

OPEN ACCESS

EDITORIAL

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, Eivind Hovig

Published: October 24, 2013 • http://dx.doi.org/10.1371/journal.pcbi.1003285

We here present ten simple rules for reproducibility of computational research. These rules can be at your disposal for whenever you want to make your research more accessible—be it for peers or for your future self.

FACULDADE DE TECNOLOGIA

Rule 1: For Every Result, Keep Track of How It Was Produced

Rule 2: Avoid Manual Data Manipulation Steps

Rule 3: Archive the Exact Versions of All External Programs Used

FACULDADE DE TECNOLOGIA

Rule 4: Version Control All Custom Scripts

Rule 5: Record All Intermediate Results, When Possible in Standardized Formats

Rule 6: For Analyses That Include Randomness, Note Underlying Random Seeds

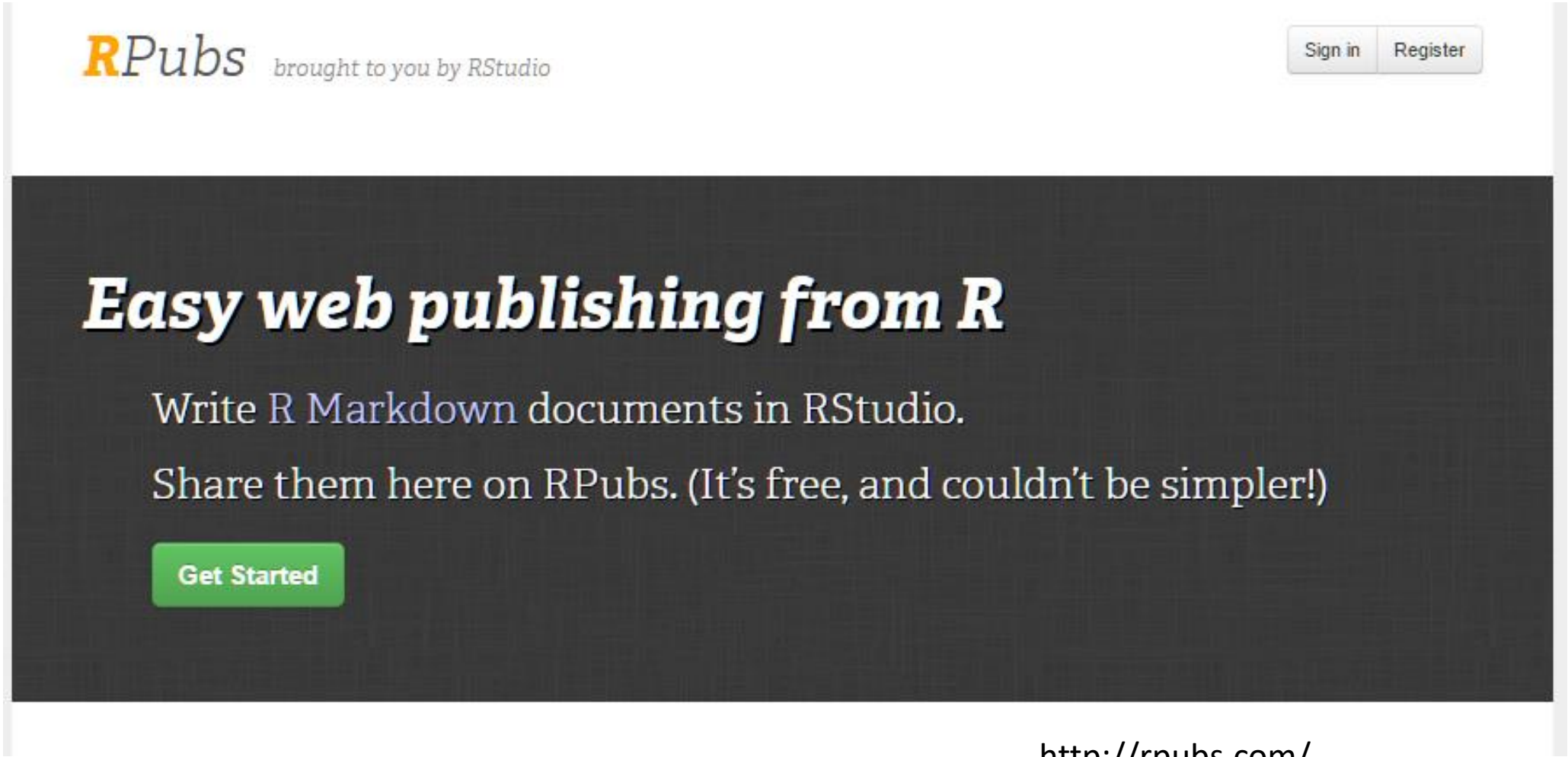
Rule 7: Always Store Raw Data behind Plots

FACULDADE DE TECNOLOGIA

Rule 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected

Rule 9: Connect Textual Statements to Underlying Results

Rule 10: Provide Public Access to Scripts, Runs, and Results



http://rpubs.com/

FACULDADE DE TECNOLOGIA

RPubs brought to you by RStudio

Reproducible Research Using RMarkdown and Git through RStudio

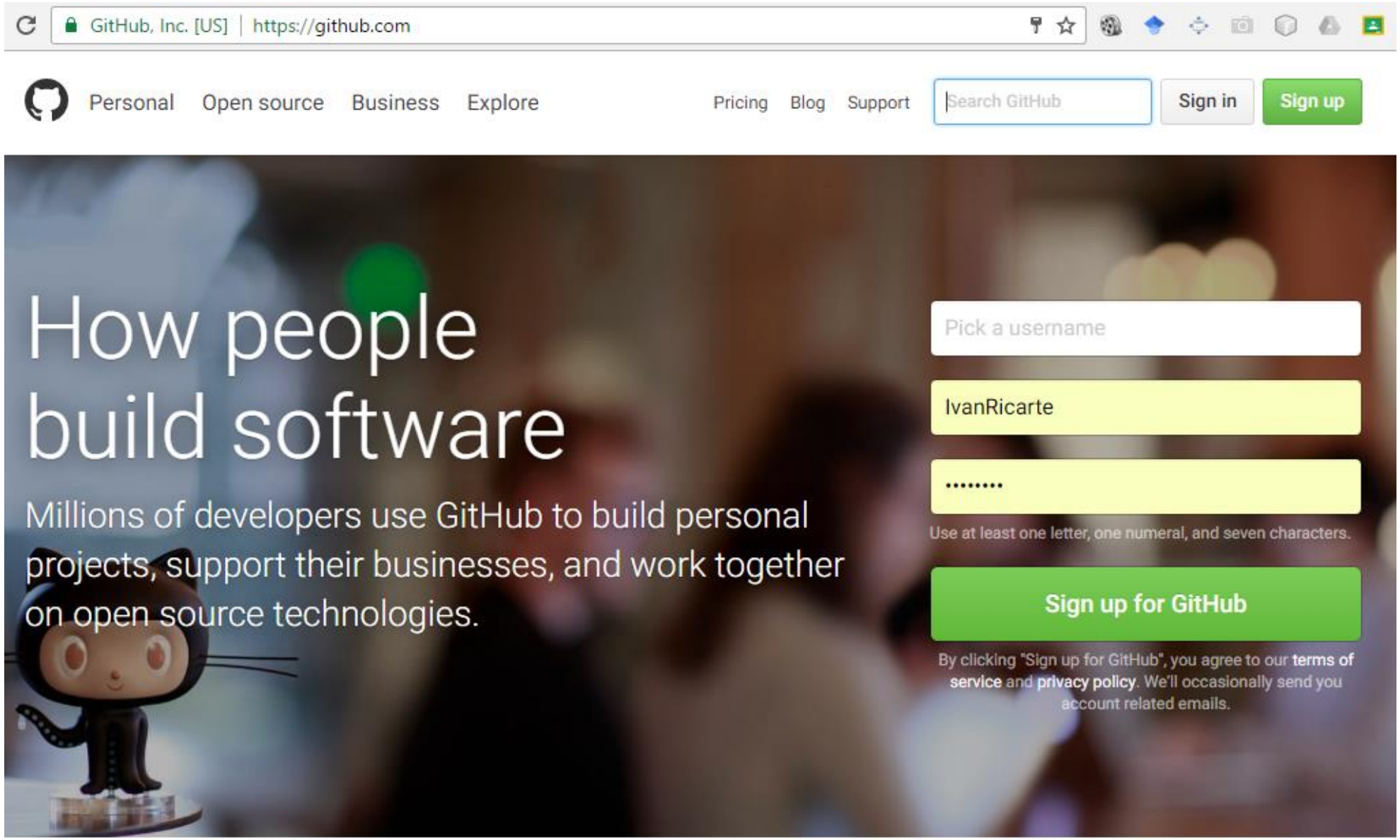
Tutorial for NGS Workshop Week 3

Marian L. Schmidt, @micro_marian, marschmi@umich.edu

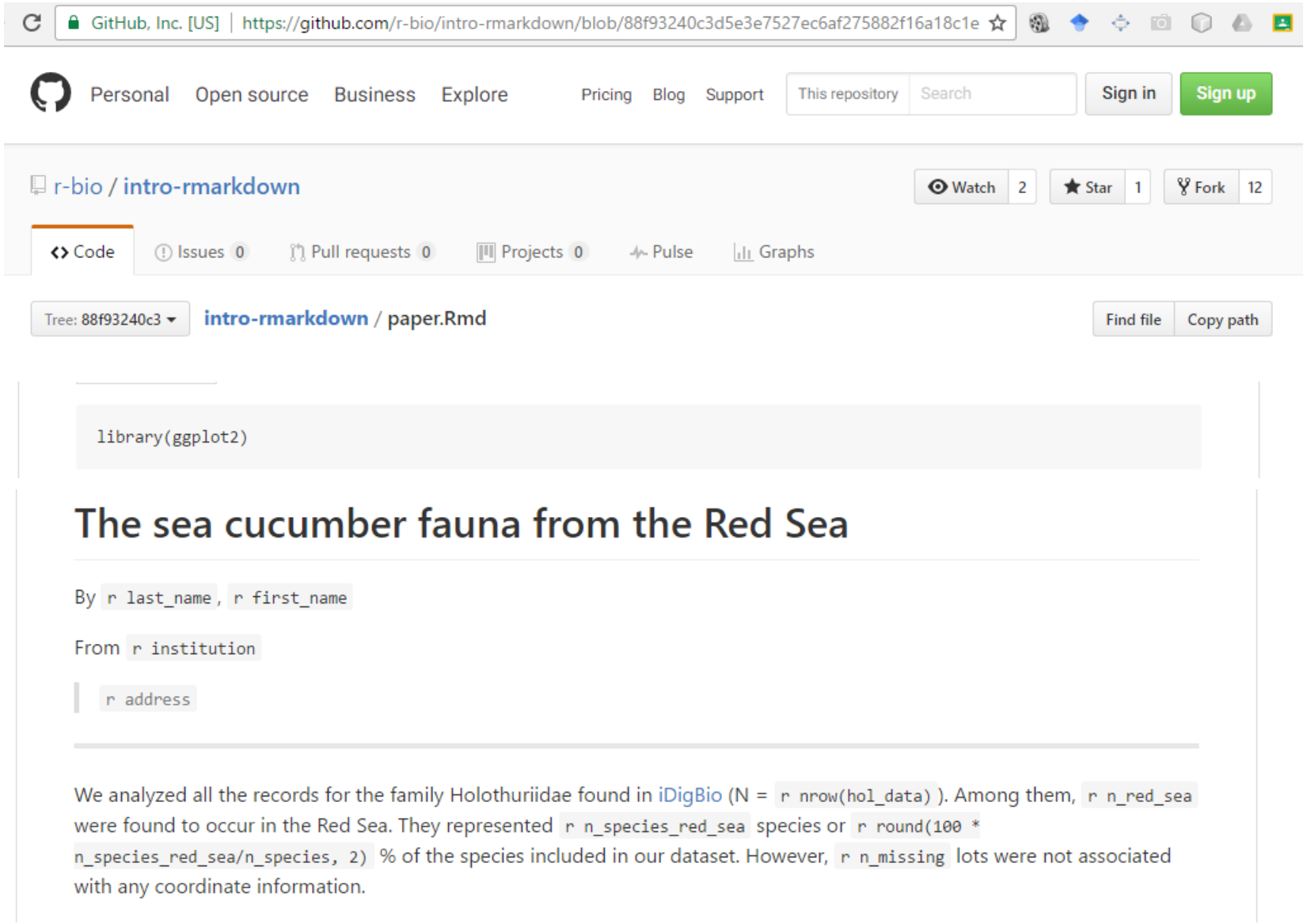
August 25, 2015

- Information that is Relevant to this Tutorial
- The Agenda and Learning Goals
- Key Resources Used to Build this lesson
- We scientists have a few problems
- How to replicate this Figure 1?
    - Could I replicate Figure 1 from your last publication?
- As scientists, it should be our goal to perform **robust** and **reproducible** research.

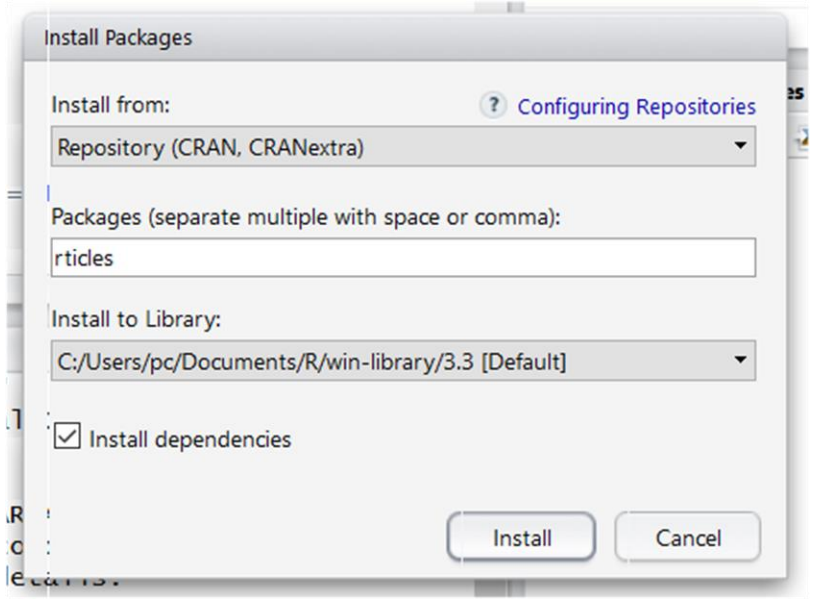
http://rpubs.com/marschmi/105639

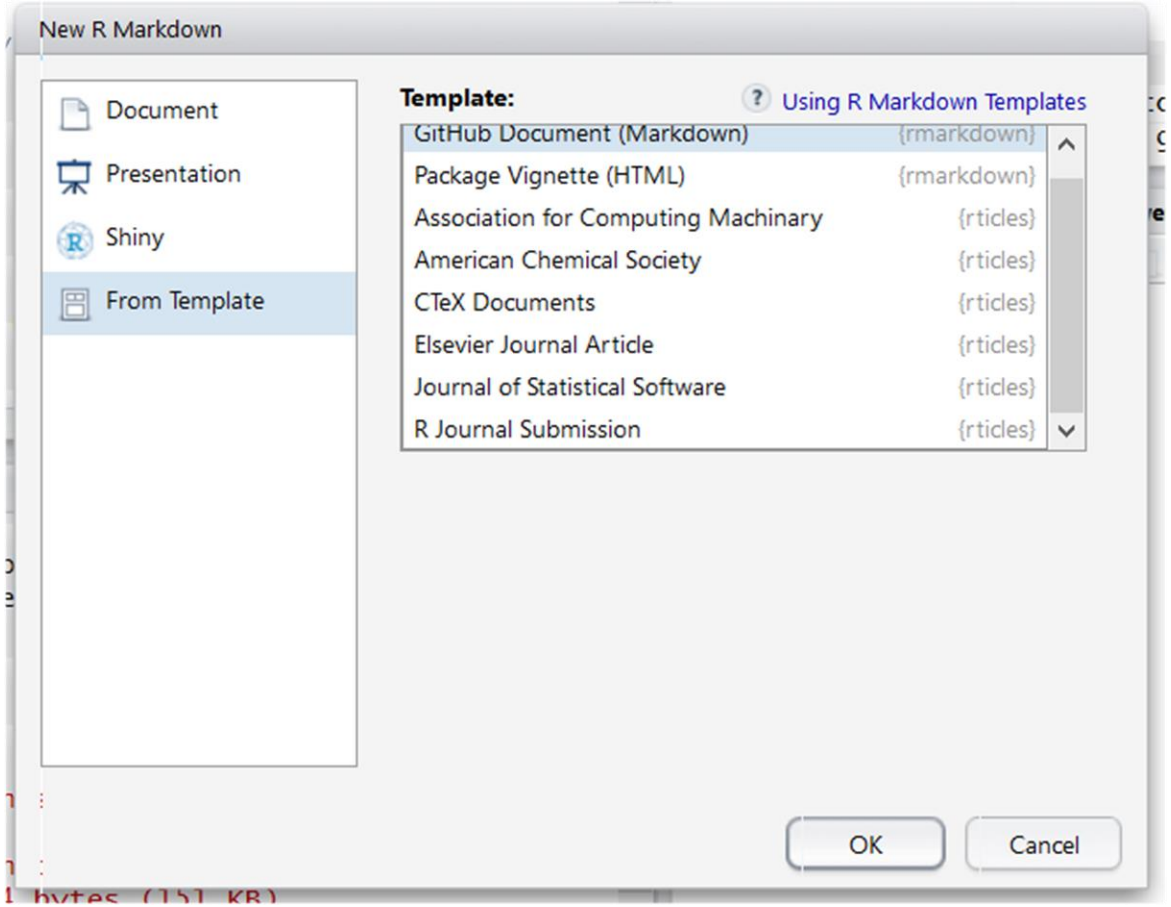

GitHub, Inc. [US] | https://github.com
Personal
Open source
Business
Explore
Pricing
Blog
Support
Search GitHub
Sign in
Sign up
How people build software
Millions of developers use GitHub to build personal projects, support their businesses, and work together on open source technologies.
Pick a username
IvanRicarte
Use at least one letter, one numeral, and seven characters.
Sign up for GitHub
By clicking "Sign up for GitHub", you agree to our terms of service and privacy policy. We'll occasionally send you account related emails.

FACULDADE DE TECNOLOGIA



Escrita de artigos acadêmicos em R



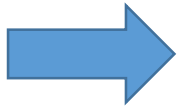


FACULDADE DE TECNOLOGIA

```yaml
---
title: Short Paper
author:
  - name: Alice Anonymous
    email: alice@example.com
    affiliation: Some Institute of Technology
    footnote: Corresponding Author
  - name: Bob Security
    email: bob@example.com
    affiliation: Another University
address:
  - code: Some Institute of Technology
    address: Department, Street, City, State, Zip
  - code: Another University
    address: Department, Street, City, State, Zip
abstract: |
  This is the abstract.

  It consists of two paragraphs.

bibliography: mybibfile.bib
output: rticles::elsevier_article
---
```



Short Paper

Alice Anonymous[*,a], Bob Security[b]

[a] *Department, Street, City, State, Zip*

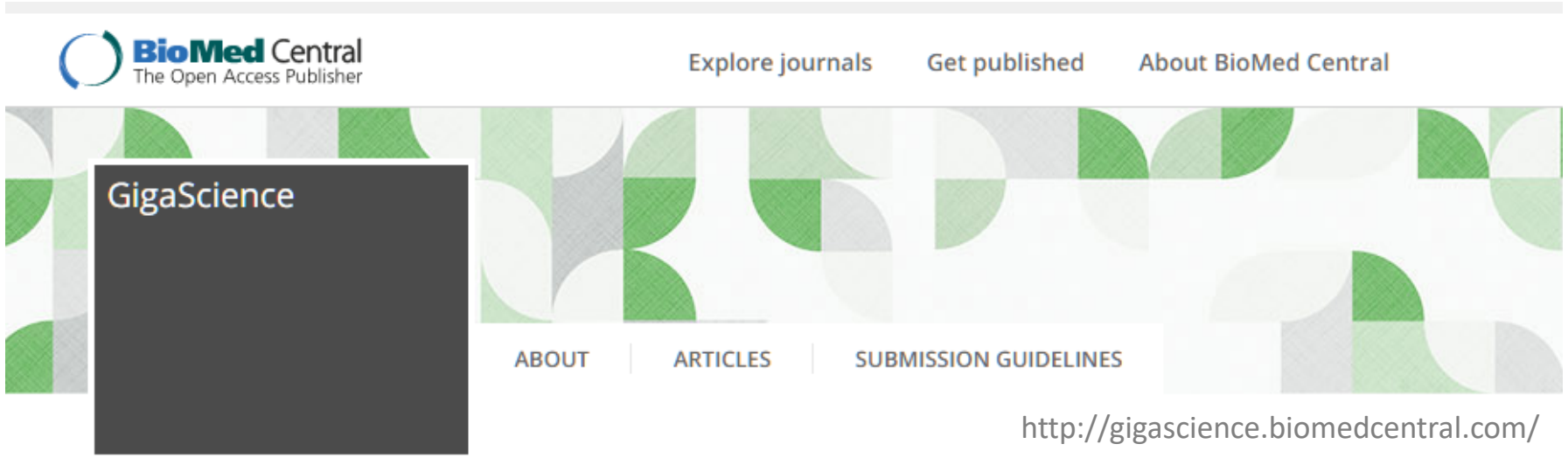
[b] *Department, Street, City, State, Zip*

Abstract

This is the abstract.

It consists of two paragraphs.

FACULDADE DE TECNOLOGIA



http://gigascience.biomedcentral.com/

Contact

Editorial Board

About

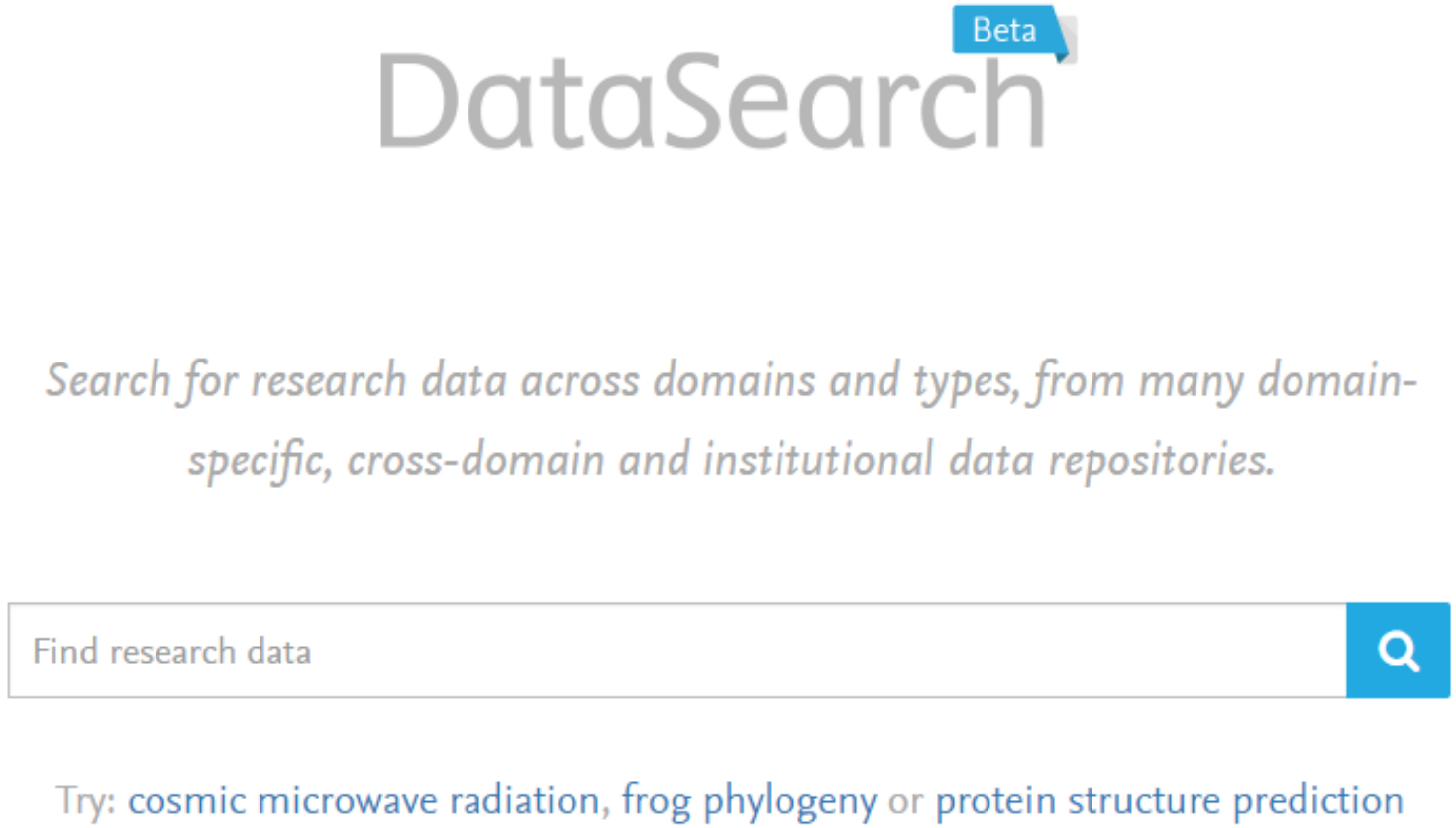
Aims and scope

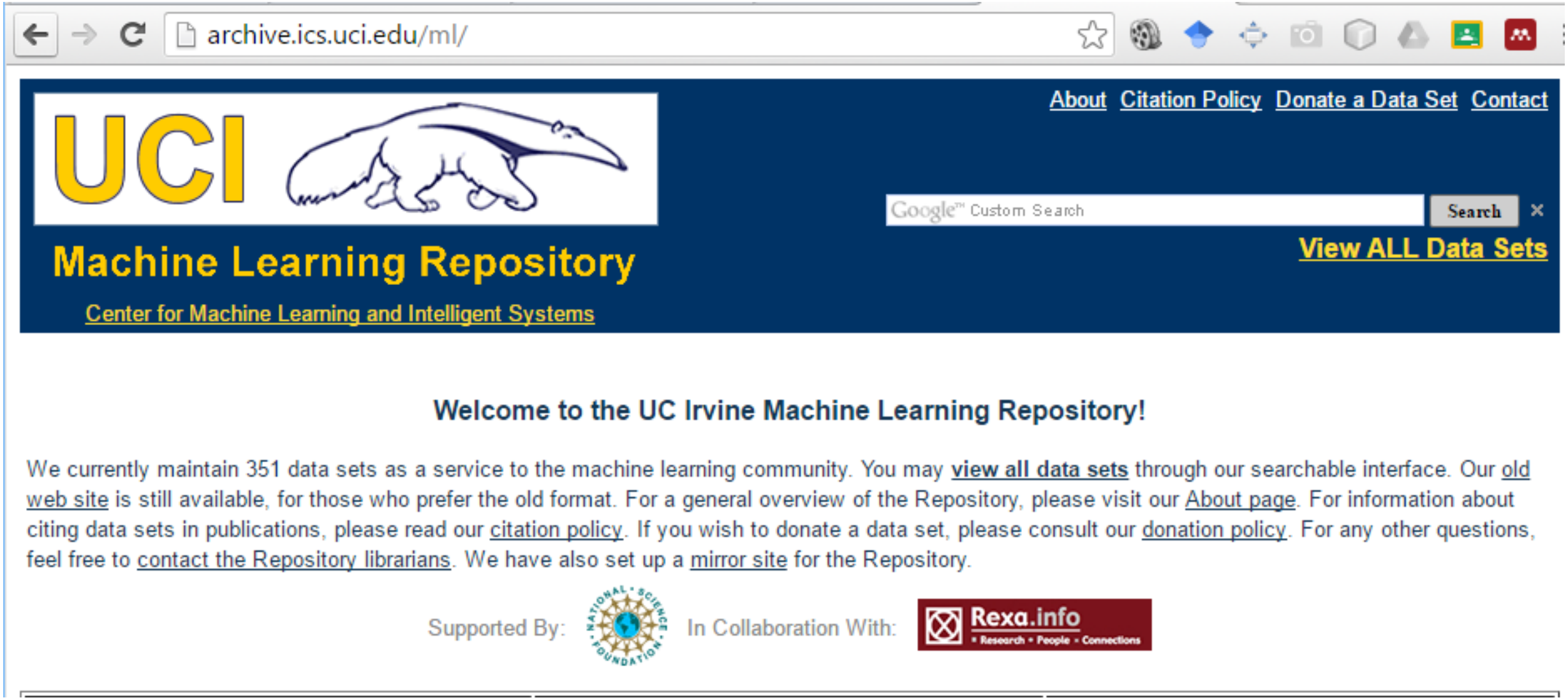
GigaScience aims to revolutionize reproducibility of analyses, data dissemination, organization, understanding, and use. As an open access and open-data journal, we publish all research objects (data, software tools and workflows) from 'big data' studies across the entire spectrum of life and biomedical sciences.

To achieve our goals, the journal has a novel publication format: one that links standard manuscript publication with an extensive database that hosts all associated data and provides data analysis tools and cloud-computing resources. GigaDB provides a direct link between the published manuscript and the relevant supporting data.

Fontes de dados para uso em pesquisa

http://datasearch.elsevier.com/



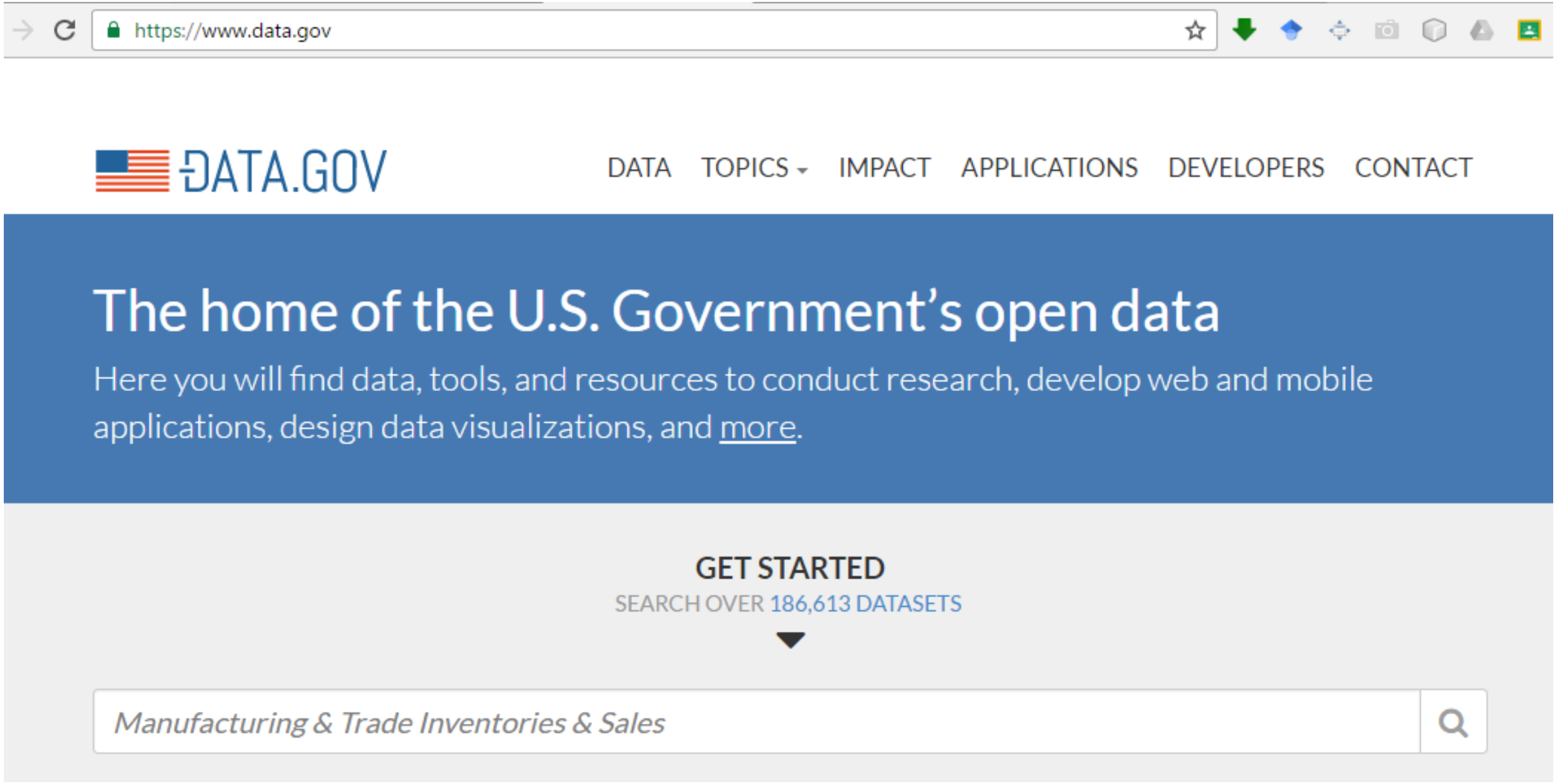

archive.ics.uci.edu/ml/
UCI
About Citation Policy Donate a Data Set Contact
Google™ Custom Search
Search
Machine Learning Repository
View ALL Data Sets
Center for Machine Learning and Intelligent Systems
Welcome to the UC Irvine Machine Learning Repository!
We currently maintain 351 data sets as a service to the machine learning community. You may view all data sets through our searchable interface. Our old web site is still available, for those who prefer the old format. For a general overview of the Repository, please visit our About page. For information about citing data sets in publications, please read our citation policy. If you wish to donate a data set, please consult our donation policy. For any other questions, feel free to contact the Repository librarians. We have also set up a mirror site for the Repository.
Supported By:
In Collaboration With:
Rexa.info
Research People Connections

- SNAP for C++
- SNAP for Python
- SNAP Datasets
- What's new
- People
- Papers
- Projects
- Citing SNAP
- Links
- About
- Contact us

Stanford Large Network Dataset Collection

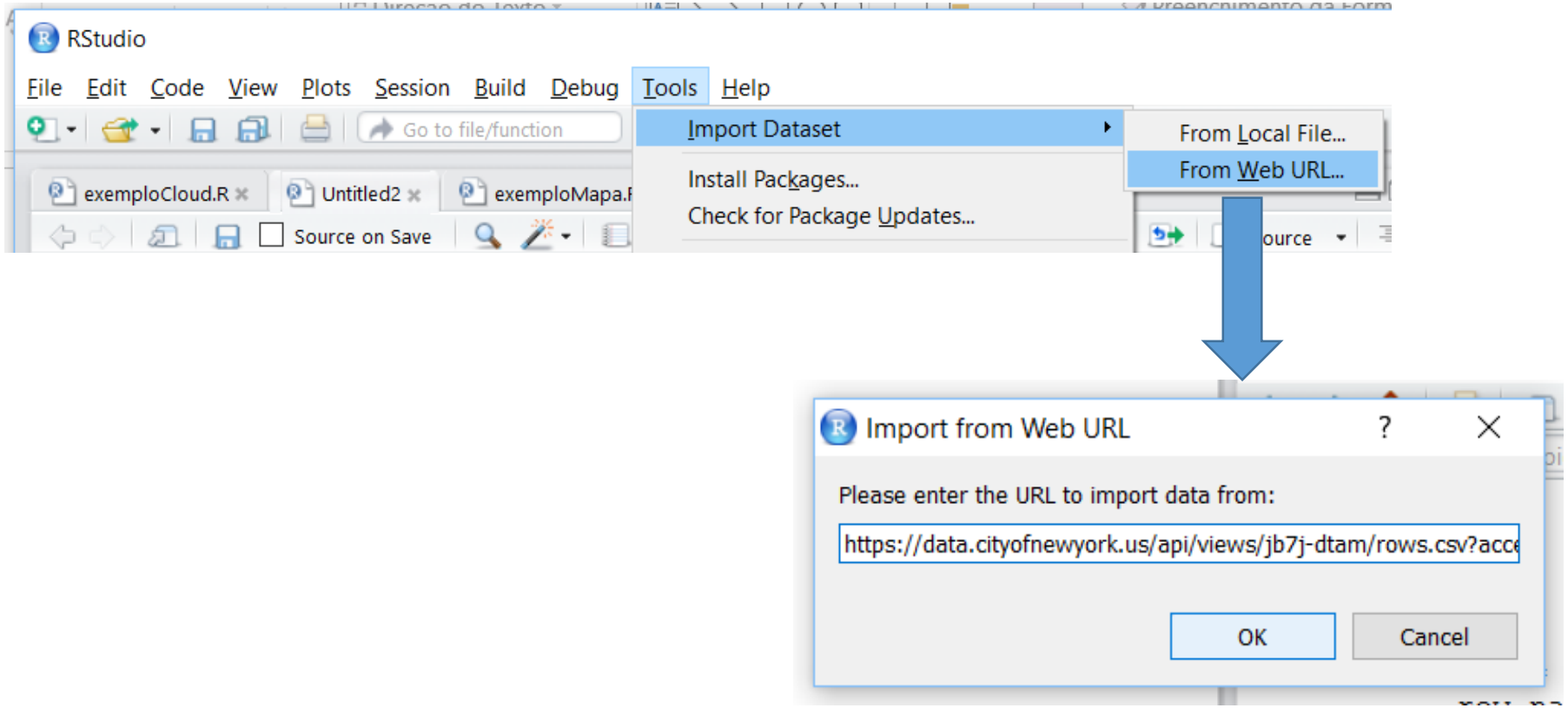
- Social networks : online social networks, edges represent interactions between people
- Networks with ground-truth communities : ground-truth network communities in social and information networks
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Internet networks : nodes represent computers and edges communication
- Road networks : nodes represent intersections and edges roads connecting the intersections
- Autonomous systems : graphs of the internet
- Signed networks : networks with positive and negative edges (friend/foe, trust/distrust)
- Location-based online social networks : Social networks with geographic check-ins
- Wikipedia networks, articles, and metadata : Talk, editing, voting, and article data from Wikipedia
- Twitter and Memetracker : Memetracker phrases, links and 467 million Tweets
- Online communities : Data from online communities such as Reddit and Flickr
- Online reviews : Data from online review systems such as BeerAdvocate and Amazon

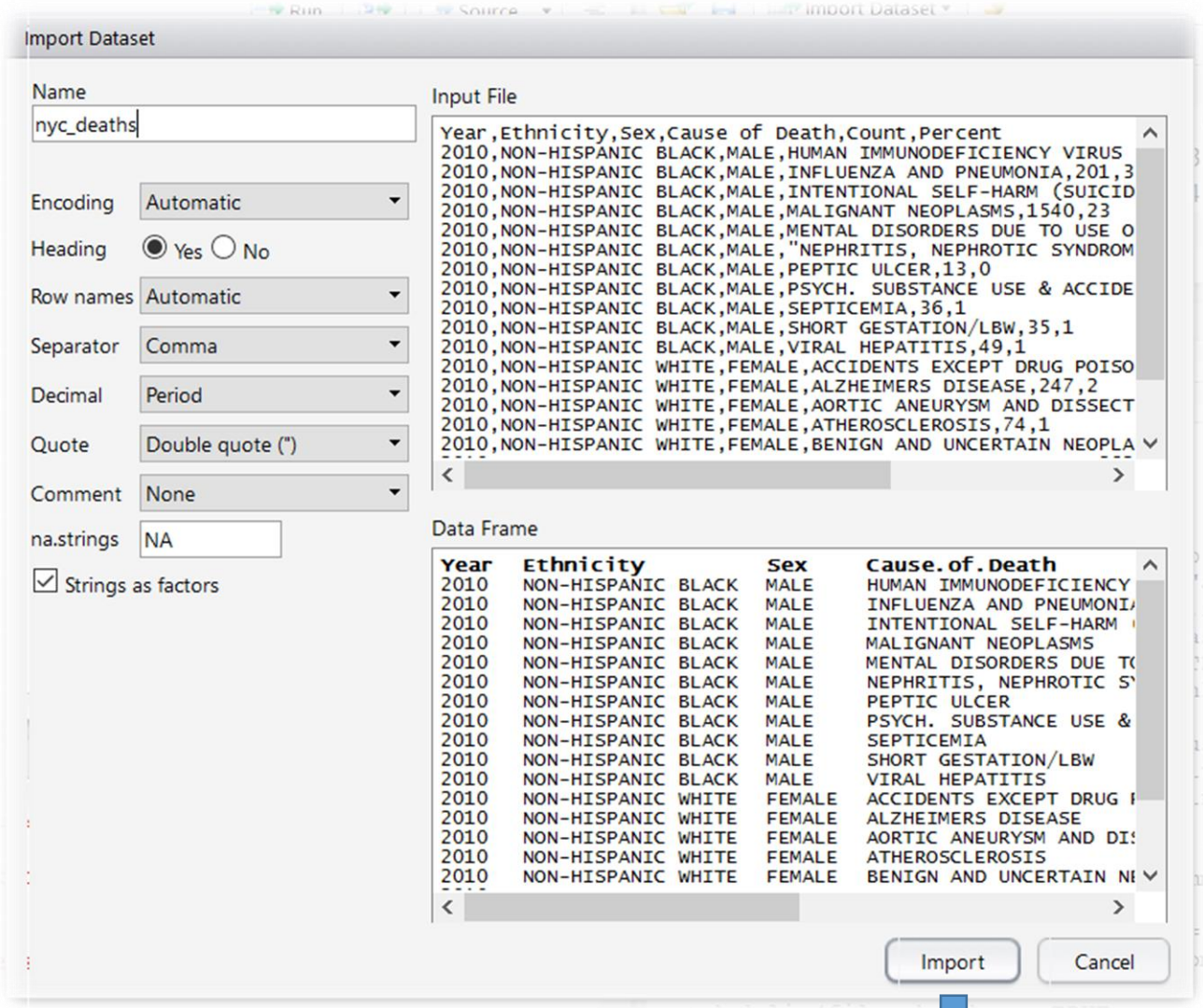
SNAP networks are also availalbe from UF Sparse Matrix collection. **Visualizations of SNAP networks** by Tim Davis.

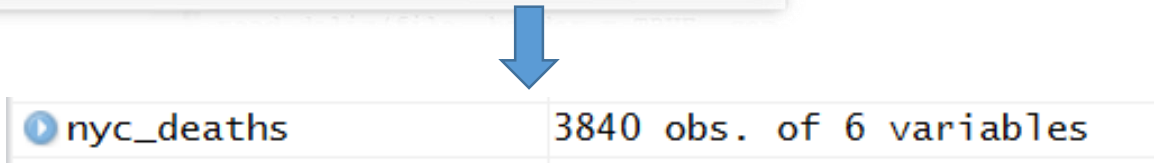

https://www.data.gov
DATA.GOV
DATA
TOPICS
IMPACT
APPLICATIONS
DEVELOPERS
CONTACT
The home of the U.S. Government's open data
Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.
GET STARTED
SEARCH OVER 186,613 DATASETS
Manufacturing & Trade Inventories & Sales

Obter dados da Web

Abordagem exploratória, interativa




Import Dataset
Name
nyc_deaths
Encoding
Automatic
Heading
Yes
No
Row names
Automatic
Separator
Comma
Decimal
Period
Quote
Double quote (")
Comment
None
na.strings
NA
Strings as factors
Input File
Year,Ethnicity,Sex,Cause of Death,Count,Percent
2010,NON-HISPANIC BLACK,MALE,HUMAN IMMUNODEFICIENCY VIRUS
2010,NON-HISPANIC BLACK,MALE,INFLUENZA AND PNEUMONIA,201,3
2010,NON-HISPANIC BLACK,MALE,INTENTIONAL SELF-HARM (SUICID
2010,NON-HISPANIC BLACK,MALE,MALIGNANT NEOPLASMS,1540,23
2010,NON-HISPANIC BLACK,MALE,MENTAL DISORDERS DUE TO USE O
2010,NON-HISPANIC BLACK,MALE,"NEPHRITIS, NEPHROTIC SYNDROM
2010,NON-HISPANIC BLACK,MALE,PEPTIC ULCER,13,0
2010,NON-HISPANIC BLACK,MALE,PSYCH. SUBSTANCE USE & ACCIDE
2010,NON-HISPANIC BLACK,MALE,SEPTICEMIA,36,1
2010,NON-HISPANIC BLACK,MALE,SHORT GESTATION/LBW,35,1
2010,NON-HISPANIC BLACK,MALE,VIRAL HEPATITIS,49,1
2010,NON-HISPANIC WHITE,FEMALE,ACCIDENTS EXCEPT DRUG POISO
2010,NON-HISPANIC WHITE,FEMALE,ALZHEIMERS DISEASE,247,2
2010,NON-HISPANIC WHITE,FEMALE,AORTIC ANEURYSM AND DISSECT
2010,NON-HISPANIC WHITE,FEMALE,ATHEROSCLEROSIS,74,1
2010,NON-HISPANIC WHITE,FEMALE,BENIGN AND UNCERTAIN NEOPLA
Data Frame
Year Ethnicity Sex Cause.of.Death
2010 NON-HISPANIC BLACK MALE HUMAN IMMUNODEFICIENCY
2010 NON-HISPANIC BLACK MALE INFLUENZA AND PNEUMONI
2010 NON-HISPANIC BLACK MALE INTENTIONAL SELF-HARM
2010 NON-HISPANIC BLACK MALE MALIGNANT NEOPLASMS
2010 NON-HISPANIC BLACK MALE MENTAL DISORDERS DUE T
2010 NON-HISPANIC BLACK MALE NEPHRITIS, NEPHROTIC S
2010 NON-HISPANIC BLACK MALE PEPTIC ULCER
2010 NON-HISPANIC BLACK MALE PSYCH. SUBSTANCE USE &
2010 NON-HISPANIC BLACK MALE SEPTICEMIA
2010 NON-HISPANIC BLACK MALE SHORT GESTATION/LBW
2010 NON-HISPANIC BLACK MALE VIRAL HEPATITIS
2010 NON-HISPANIC WHITE FEMALE ACCIDENTS EXCEPT DRUG
2010 NON-HISPANIC WHITE FEMALE ALZHEIMERS DISEASE
2010 NON-HISPANIC WHITE FEMALE AORTIC ANEURYSM AND DI
2010 NON-HISPANIC WHITE FEMALE ATHEROSCLEROSIS
2010 NON-HISPANIC WHITE FEMALE BENIGN AND UNCERTAIN N
Import
Cancel


nyc_deaths
3840 obs. of 6 variables

Abordagem não-interativa: pacote RCurl

```
library(RCurl)

dados <- read.csv(
    textConnection(
        getURLContent("https://data.cityofnewyork.us/api/views/jb7j-dtam/rows.csv?accessType=DOWNLOAD")))
```

	Year	Ethnicity	Sex	Cause.of.Death	Count	Percent
1	2010	NON-HISPANIC BLACK	MALE	HUMAN IMMUNODEFICIENCY VIRUS DISEASE	297	5
2	2010	NON-HISPANIC BLACK	MALE	INFLUENZA AND PNEUMONIA	201	3
3	2010	NON-HISPANIC BLACK	MALE	INTENTIONAL SELF-HARM (SUICIDE)	64	1
4	2010	NON-HISPANIC BLACK	MALE	MALIGNANT NEOPLASMS	1540	23
5	2010	NON-HISPANIC BLACK	MALE	MENTAL DISORDERS DUE TO USE OF ALCOHOL	50	1
6	2010	NON-HISPANIC BLACK	MALE	NEPHRITIS, NEPHROTIC SYNDROME AND NEPHROSIS	70	1
7	2010	NON-HISPANIC BLACK	MALE	PEPTIC ULCER	13	0
8	2010	NON-HISPANIC BLACK	MALE	PSYCH. SUBSTANCE USE & ACCIDENTAL DRUG POISO...	111	2

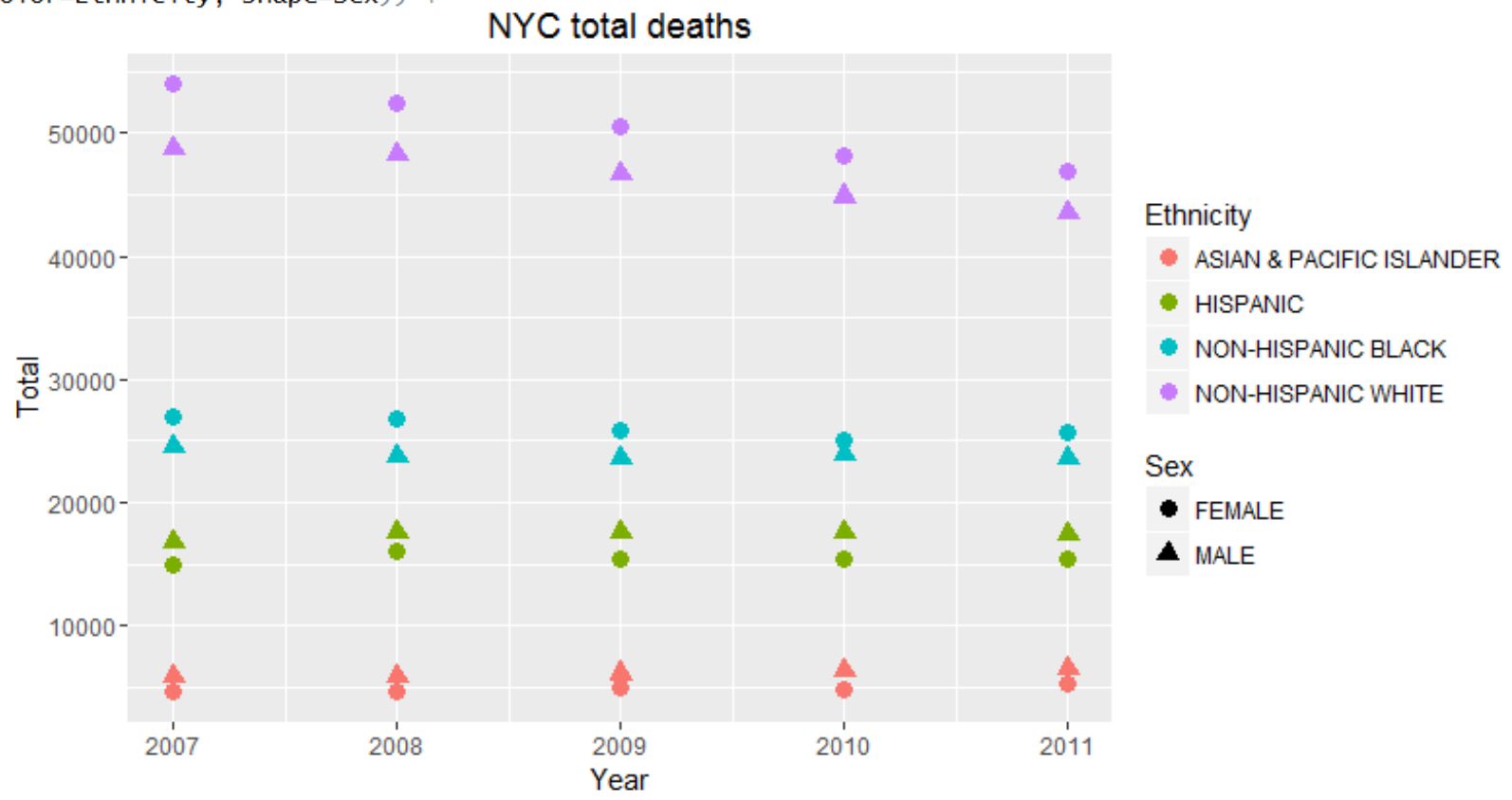
FACULDADE DE TECNOLOGIA

```
library(RCurl)
library(dplyr)
library(ggplot2)

dados <- read.csv(
    textConnection(
        getURLContent("https://data.cityofnewyork.us/api/views/jb7j-dtam/rows.csv?accessType=DOWNLOAD")))

nyc_d <- dados %>%
    group_by(Year, Ethnicity, Sex) %>%
    summarise(Total = sum(Count))

ggplot(nyc_d, aes(Year, Total, color=Ethnicity, shape=Sex)) +
    geom_point(size=3) +
    ggtitle("NYC total deaths")
```



FACULDADE DE TECNOLOGIA

Sumário: pesquisa reproduzível

Criação de pesquisa reproduzível em R: Rmarkdown

- Recomendações para criação de pesquisa reproduzível
- Repositórios de documentos e de códigos

Escrita de artigos científicos em RMarkdown

Veículos acadêmicos que promovem pesquisa reproduzível

Fontes de dados para uso em pesquisa

- *Data sharing, Open data*
- Pacote RCurl

FACULDADE DE TECNOLOGIA

Considerações finais

FACULDADE DE TECNOLOGIA

Em resumo, R...

Ambiente e linguagem para análise de dados

Milhares de pacotes com funções para diferentes funcionalidades

Adequado para tarefas de análise exploratória de dados e de análise estatística de dados experimentais

Funções para limpeza e transformação de dados

Funções gráficas avançadas

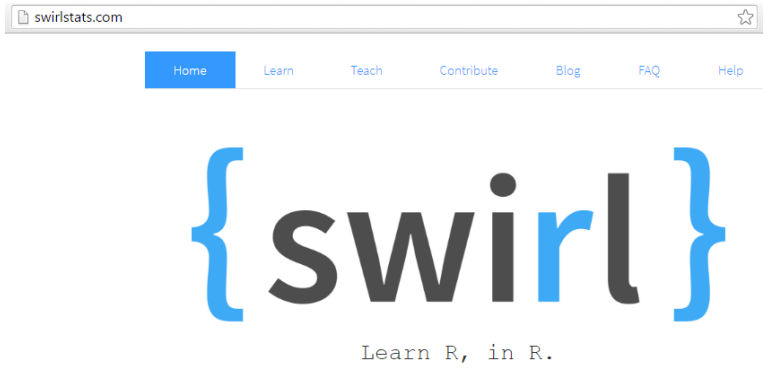
Apoio à pesquisa reproduzível

E agora...

Tutoriais, livros, sites

Galerias de aplicações

Um breve tutorial de R: swirl



http://swirlstats.com/

```
> install.packages("swirl")
Installing package into 'C:/Users/pc/Documents/R,

> library(swirl)

| Hi! I see that you have some variables saved

> swirl()

| Welcome to swirl! Please sign in. If you've been here before,
                                                  re new, call yourself

1: R Programming: The basics of programming in R
2: Regression Models: The basics of regression modeling in R
3: Statistical Inference: The basics of statistical inference in R
4: Exploratory Data Analysis: The basics of exploring data in R
5: Don't install anything for me. I'll do it myself.
```

tryr.codeschool.com

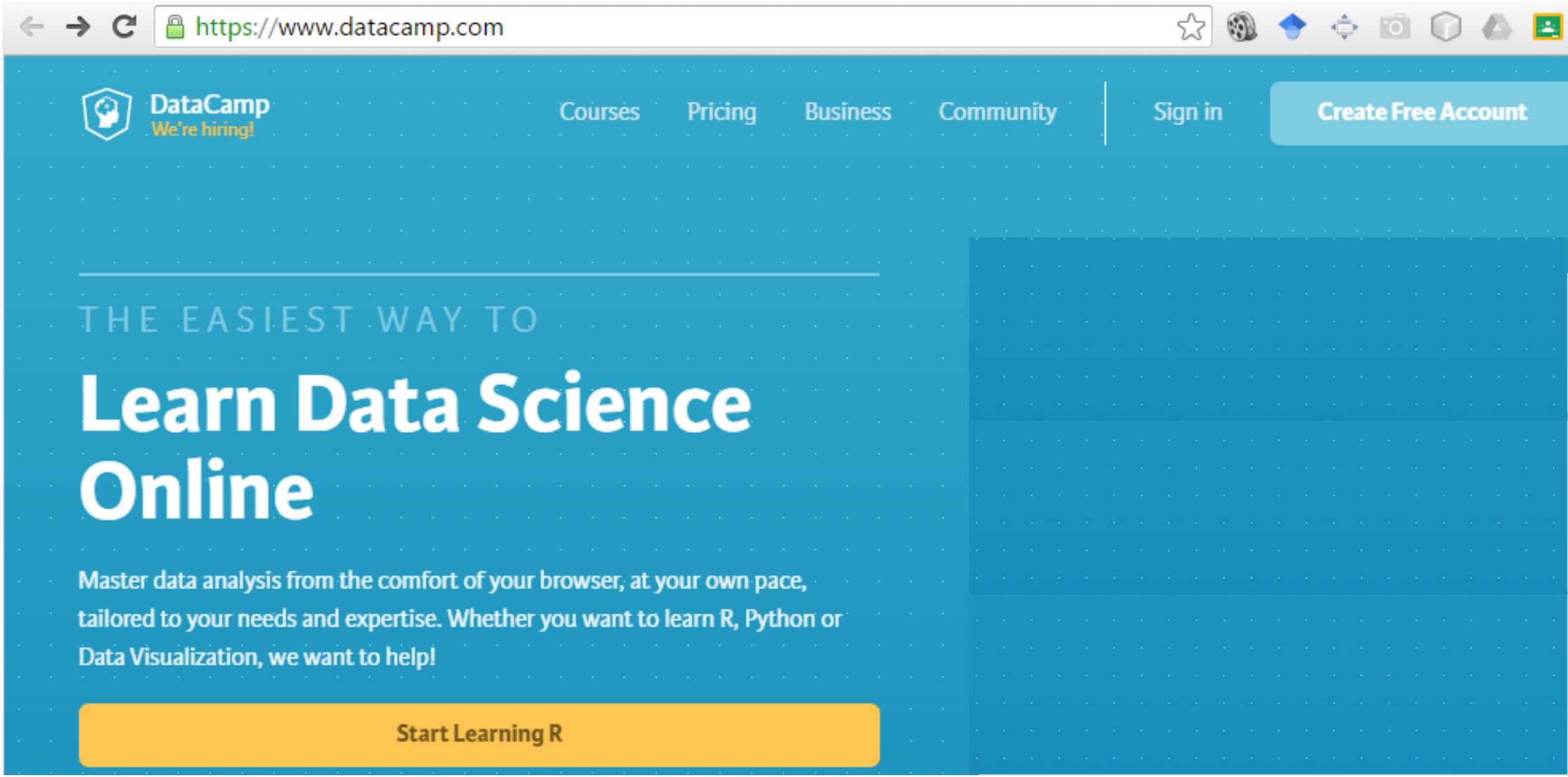
Try R is Sponsored By:

O'REILLY®

Created By:

Code School

R is a tool for statistics and data modeling. The R programming language is elegant, versatile, and has a highly expressive syntax designed around working with data. R is more than that, though — it also includes extremely powerful graphics capabilities. If you want to easily manipulate your data and present it in compelling ways, R is the tool for you.


https://www.datacamp.com
DataCamp
We're hiring!
Courses
Pricing
Business
Community
Sign in
Create Free Account
THE EASIEST WAY TO
Learn Data Science Online
Master data analysis from the comfort of your browser, at your own pace, tailored to your needs and expertise. Whether you want to learn R, Python or Data Visualization, we want to help!
Start Learning R

www.cookbook-r.com

Cookbook for R

Welcome to the Cookbook for R. The goal of the cookbook is to provide solutions to common tasks and problems in analyzing data.

Most of the code in these pages can be copied and pasted into the R command window if you want to see them in action.

1. Basics
2. Numbers
3. Strings
4. Formulas
5. Data input and output
6. Manipulating data
7. Statistical analysis
8. Graphs
9. Scripts and functions
10. Tools for experiments

www.statmethods.net/index.html

Home | Interface | Input | Manage | Stats | Adv Stats | Graphs | Adv Graphs | Blog

Quick-R

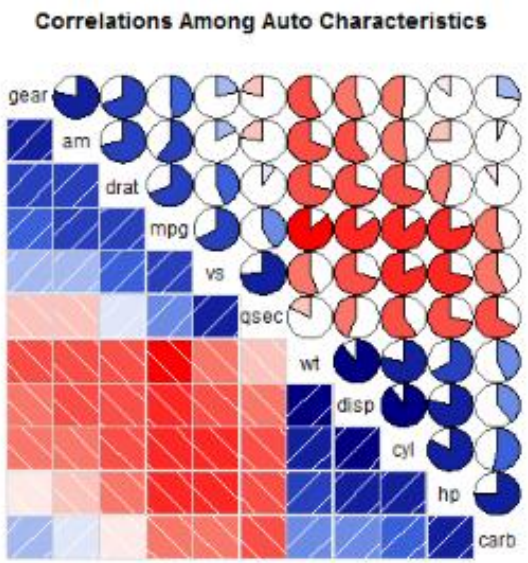
accessing the power of R

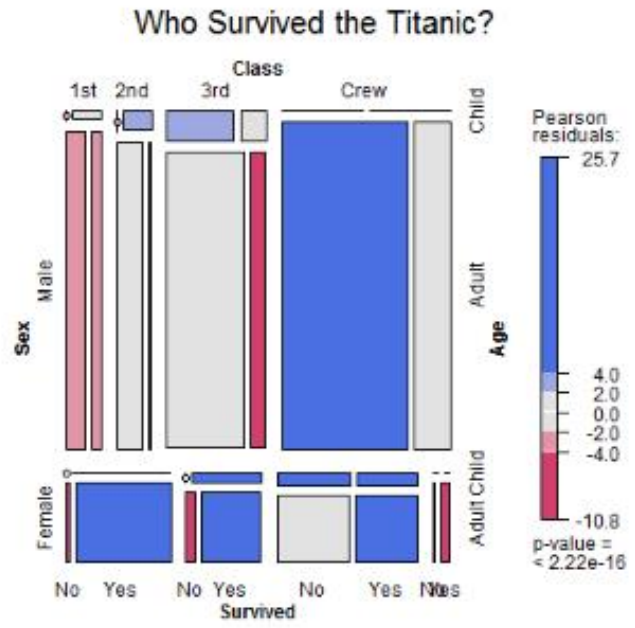
Search

Top Menu

- Home
- The R Interface
- Data Input
- Data Management
- Basic Statistics
- Advanced Statistics
- Basic Graphs
- Advanced Graphs
- Blog

About Quick-R





FACULDADE DE TECNOLOGIA

https://www.r-bloggers.com

R-bloggers

R news and tutorials contributed by (580) R bloggers

Home | About | RSS | add your blog! | Learn R | R jobs | Contact us

WELCOME!

Follow @rbloggers 36K fo

Here you will find daily **news and tutorials about R**, contributed by over 573 bloggers.
There are many ways to **follow us** -
By e-mail:

Your e-mail here
Subscribe
32243 readers
BY FEEDBURNER

On Facebook:

R blog…
39 mil curtidas

Curtir Página

Last few days for EARL London Tickets

September 5, 2016
By Angela Roberts

Last few days to purchase your tickets for the UK's biggest R Conference this year! On September 13th-15th at The Tower Hotel, London, the conference has 1 day of workshops which are selling up fast and 2 days of conference with … Continue reading →

Read more »

MinechaRts #1 (Minecraft + R + Edgar Anderson's

Rcpp 0.12.7: More updates

September 4, 2016

Search & Hit Enter

RECENT POPULAR POSTS

IBM Data Science Experience: First steps with yorkr

MOST VISITED ARTICLES OF THE WEEK

1. How to write the first for loop in R
2. Installing R packages
3. R tutorials
4. Using apply, sapply, lapply in R
5. In-depth introduction to machine learning in 15 hours of expert videos
6. Scatterplots
7. forcats 0.1.0
8. How to perform a Logistic Regression in R
9. How to Make a Histogram with Basic R

Book
SpringerBriefs in Statistics
2011

A Tiny Handbook of R

Authors: Mike Allerhand

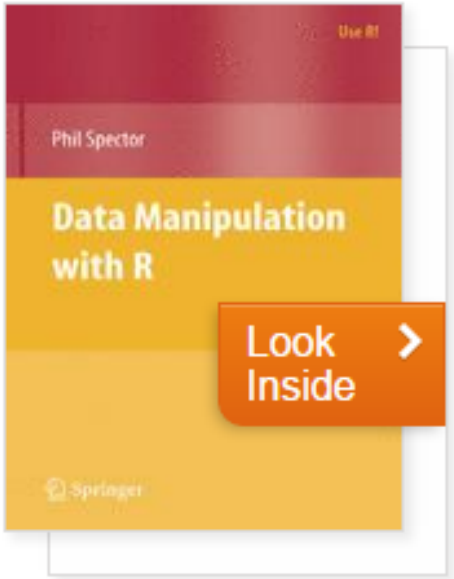
ISBN: 978-3-642-17979-2 (Print) 978-3-642-17980-8 (Online)

Book
Use R!
2008

Data Manipulation with R

Authors: Statistical Computing Facility Phil Spector

ISBN: 978-0-387-74730-9 (Print) 978-0-387-74731-6 (Online)



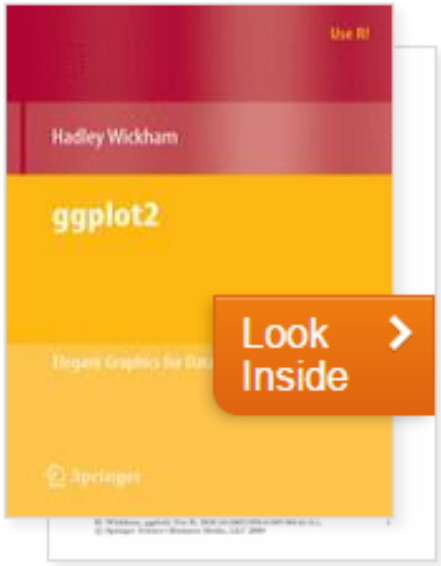
Book
Use R
2009

ggplot2

Elegant Graphics for Data Analysis

Authors: Hadley Wickham

ISBN: 978-0-387-98140-6 (Print) 978-0-387-98141-3 (Online)



Book
SpringerBriefs in Political Science
Volume 5 2013

Qualitative Comparative Analysis with R

A User's Guide

Authors: Alrik Thiem, Adrian Duşa

ISBN: 978-1-4614-4583-8 (Print) 978-1-4614-4584-5 (Online)

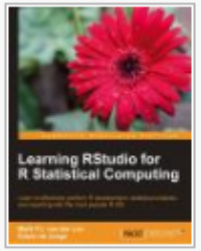
ProQuest ebrary

Início Pesquisar Estante Entrar

Página Intro de 115

SBU
SISTEMA DE BIBLIOTECAS DA UNICAMP

Learning RStudio for R Statistical Computing



por van der Loo, Mark P.J., de Jonge, Edwin

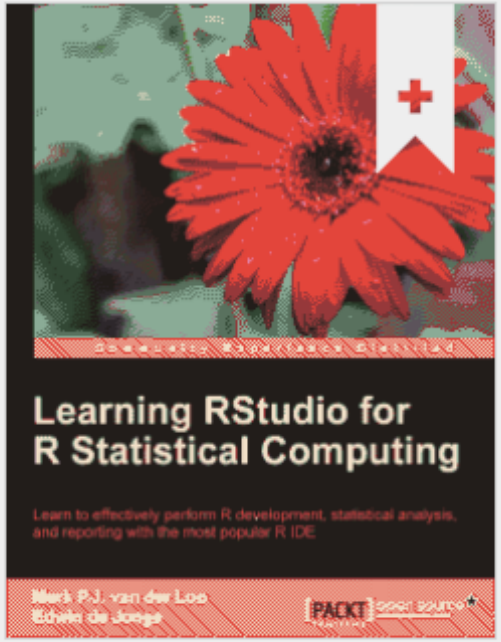
EDITORA
Packt Publishing

DATA
December 2012

Mais...

Pesquisar no livro

SUMÁRIO



4. Managing R Projects

Autor:	P.J, van der Loo Mark
Livro:	Learning RStudio for R Statistical Computing
Editor:	Packt Publishing
ISBN:	1-78216-060-4, 978-1-78216-060-1
Data:	2012
Página:	1-1

Ebrary Academic Complete International Subscription Collection

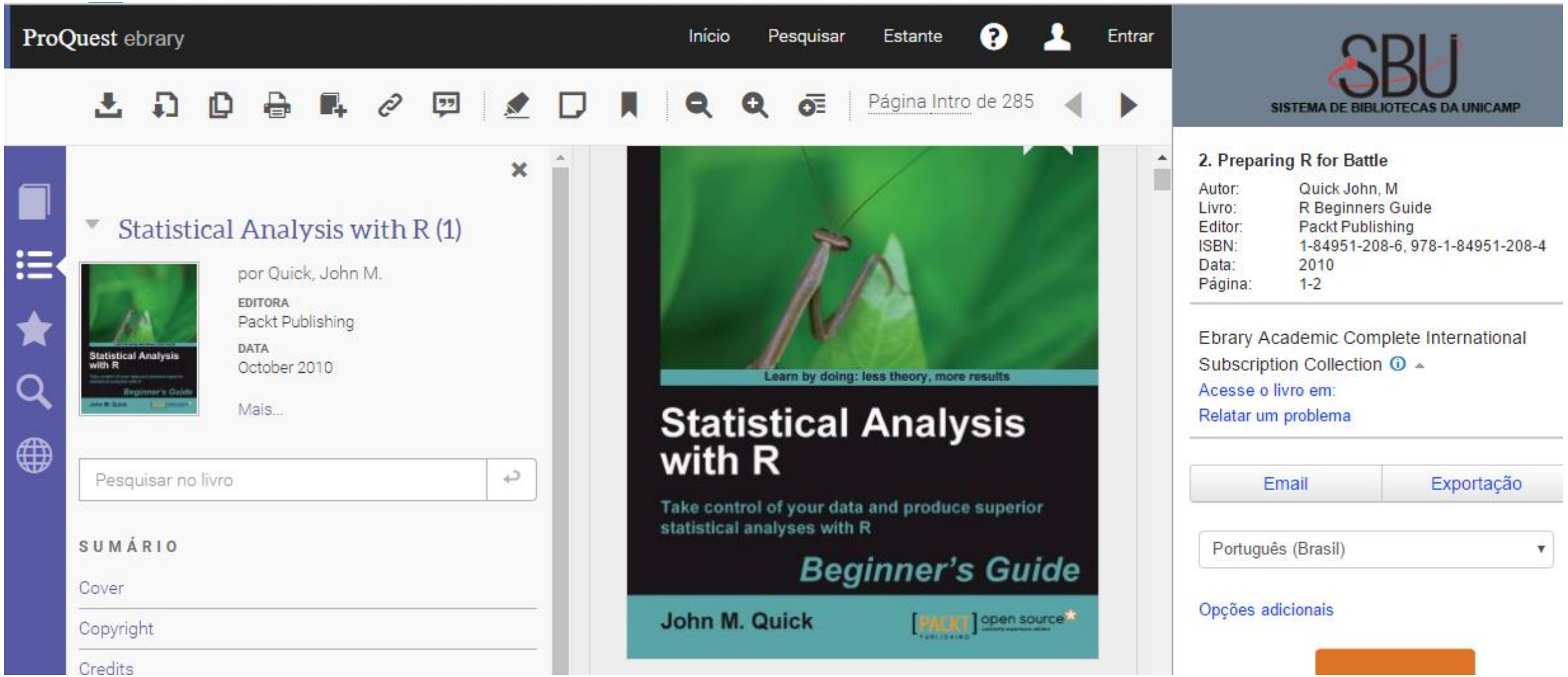
Acesse o livro em:

Relatar um problema

Email | Exportação

Português (Brasil)

Opções adicionais


ProQuest ebrary
Início
Pesquisar
Estante
Entrar
SBU
SISTEMA DE BIBLIOTECAS DA UNICAMP
Página Intro de 285
Statistical Analysis with R (1)
por Quick, John M.
EDITORA
Packt Publishing
DATA
October 2010
Mais...
Pesquisar no livro
SUMÁRIO
Cover
Copyright
Credits
Learn by doing: less theory, more results
Statistical Analysis with R
Take control of your data and produce superior statistical analyses with R
Beginner's Guide
John M. Quick
[PACKT] open source
2. Preparing R for Battle
Autor: Quick John, M
Livro: R Beginners Guide
Editor: Packt Publishing
ISBN: 1-84951-208-6, 978-1-84951-208-4
Data: 2010
Página: 1-2
Ebrary Academic Complete International Subscription Collection
Acesse o livro em:
Relatar um problema
Email
Exportação
Português (Brasil)
Opções adicionais

Book
Use R
2008

Statistical Methods for Environmental Epidemiology with R

A Case Study in Air Pollution and Health

Authors: Francesca Dominici, Roger D. Peng

ISBN: 978-0-387-78166-2 (Print) 978-0-387-78167-9 (Online)

Chapter
Statistical Methods for Environmental Epidemiology with R
Part of the series Use R pp 19-30

3

Reproducible Research Tools

3.1 Introduction

The validity of conclusions from scientific investigations is typically strengthened by the replication of results by independent researchers. Full replication of a study's results using independent methods, data, equipment, and protocols, has long been, and will continue to be, the standard by which scientific claims are evaluated. In many fields of study, there are examples of scientific

Book 2013

R for Business Analytics

Authors: A Ohri

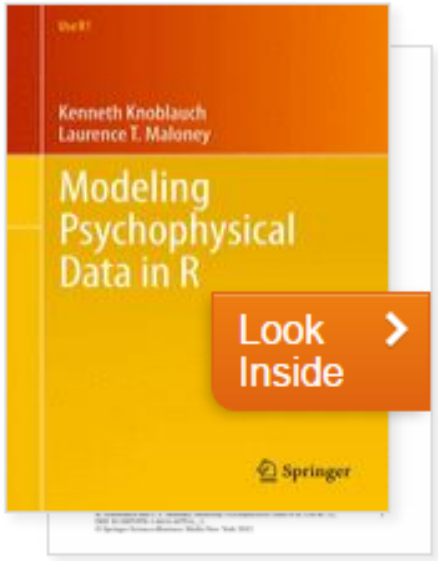
ISBN: 978-1-4614-4342-1 (Print) 978-1-4614-4343-8 (Online)

Book
Use R!
Volume 32 2012

Modeling Psychophysical Data in R

Authors: Kenneth Knoblauch, Laurence T. Maloney

ISBN: 978-1-4614-4474-9 (Print) 978-1-4614-4475-6 (Online)



FACULDADE DE TECNOLOGIA

www.itl.nist.gov/div898/handbook/

NIST
SEMATECH

HANDBOOK CHAPTERS

1. Explore
2. Measure
3. Characterize
4. Model
5. Improve
6. Monitor
7. Compare
8. Reliability

HOW TO USE HANDBOOK

TOOLS & AIDS

SEARCH HANDBOOK

DETAILED CONTENTS

ACKNOWLEDGMENTS

ENGINEERING STATISTICS

H A N D B O O K

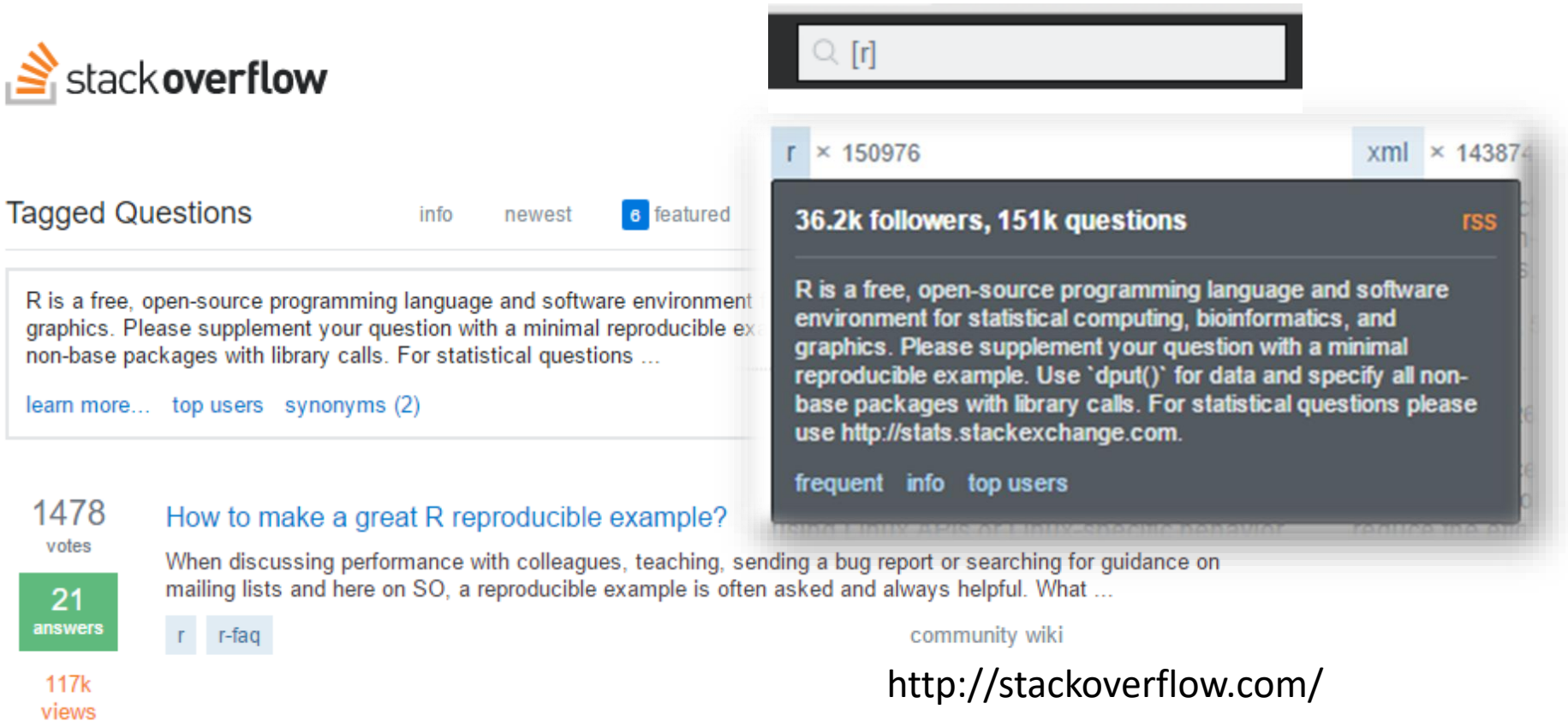
Welcome! The goal of this handbook is to help scientists and engineers incorporate statistical methods in their work as efficiently as possible.

ENGINEERING STATISTICS HANDBOOK

HOME | TOOLS & AIDS | SEARCH | BACK | NEXT

1. Exploratory Data Analysis

This chapter presents the assumptions, principles, and techniques necessary to gain insight into data via EDA--exploratory data analysis.



http://stackoverflow.com/

Prof. Ivan L. M. Ricarte
http://www.ft.unicamp.br/~ricarte
ricarte@unicamp.br