Prof. MSc. Marcos Alexandruk

E-mail: alexandruk@uninove.br

Conteúdo Programático

- Características do R
- Instalação do RStudio
- Comandos Básicos
- Tipos de Dados
- Adicionando Pacotes
- Estruturas Básicas do R
  - Vetores
  - Matrizes
  - Arrays
  - Fatores
  - Data Frames
  - Listas
- Entrada de Dados
- Tipos de Arquivos do R
- Funções: head, tail e summary
- Criando Funções
- Análise Descritiva
  - Medidas de Tendência Central
  - Medidas de Dispersão
  - Separatrizes
- Distribuições Simétricas e Assimétricas
- Curtose
- Gráficos
- ggplot2
- Modelos Probabilísticos Discretos
  - Distribuição binomial
  - Distribuição de Poisson
- Modelos Probabilísticos Contínuos
  - Distribuição Uniforme Contínua
  - Distribuição Normal ou Gaussiana
  - Distribuição Exponencial
- Regressão Linear Simples
- Correlação
- Testes de Hipóteses
  - t de Student
  - Shapiro-Wilk
  - ANOVA
  - Bartlett
  - Tukey HSD

Características do R

R: O que é?

R é uma linguagem e um ambiente de desenvolvimento disponível gratuitamente para computação estatística e gráficos que fornece uma ampla variedade de técnicas estatísticas e gráficas: modelagem linear e não linear, testes estatísticos, análise de séries temporais, classificação, agrupamento, etc.

Principais características

- Open source
- Multi plataforma (Windows, Linux e MacOS)
- Ambiente de linha de comando
- Case sensitive (diferencia maiúsculas de minúsculas)
- Atualizado com frequência
- Extensível através de pacotes
- Milhares de funções de análise de dados
- Gera diferentes tipos de gráficos
- Processamento em memória (melhor performance)
- Integração com: Power BI, Tableau, Oracle, SQL Server, Java, Python, etc.

Instalação do RStudio

Instalação do R

O download do R é realizado através do **CRAN** (Comprehensive R Archive Network), uma rede de servidores ftp e web distribuída mundialmente, que armazena versões idênticas e atualizadas de código e documentação para R. Recomenda-se **utilizar o espelho CRAN mais próximo** para minimizar a carga da rede.

O **download** do R é realizado a partir do seguinte **link**:

https://cran.r-project.org/bin/

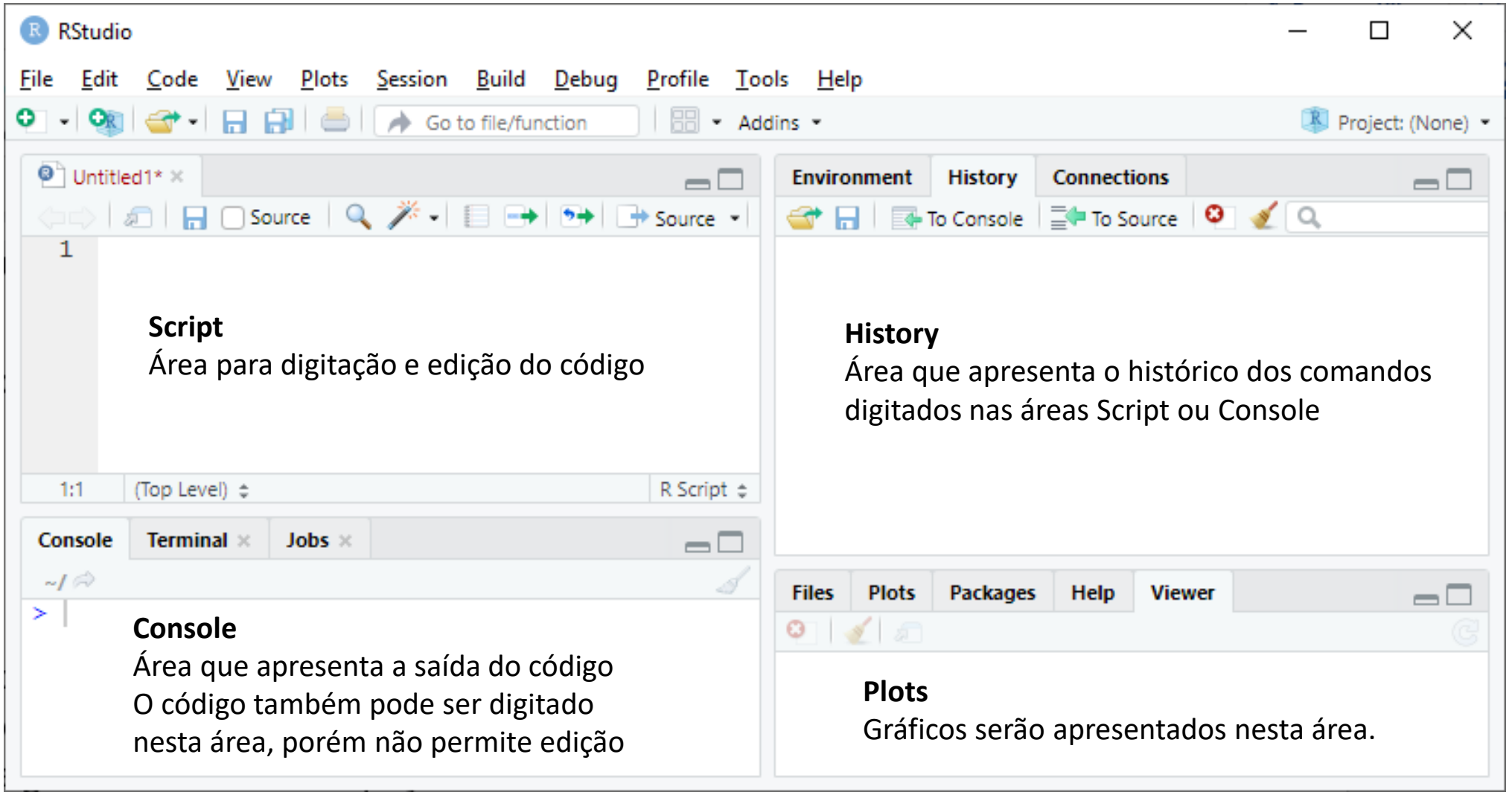
Instalação do R Studio

O RStudio é um **ambiente de desenvolvimento integrado (IDE)** para a linguagem R. Inclui um console, e um editor com destaque de sintaxe que oferece suporte à execução direta de código, bem como ferramentas para plotagem, histórico, depuração e gerenciamento de espaço de trabalho.

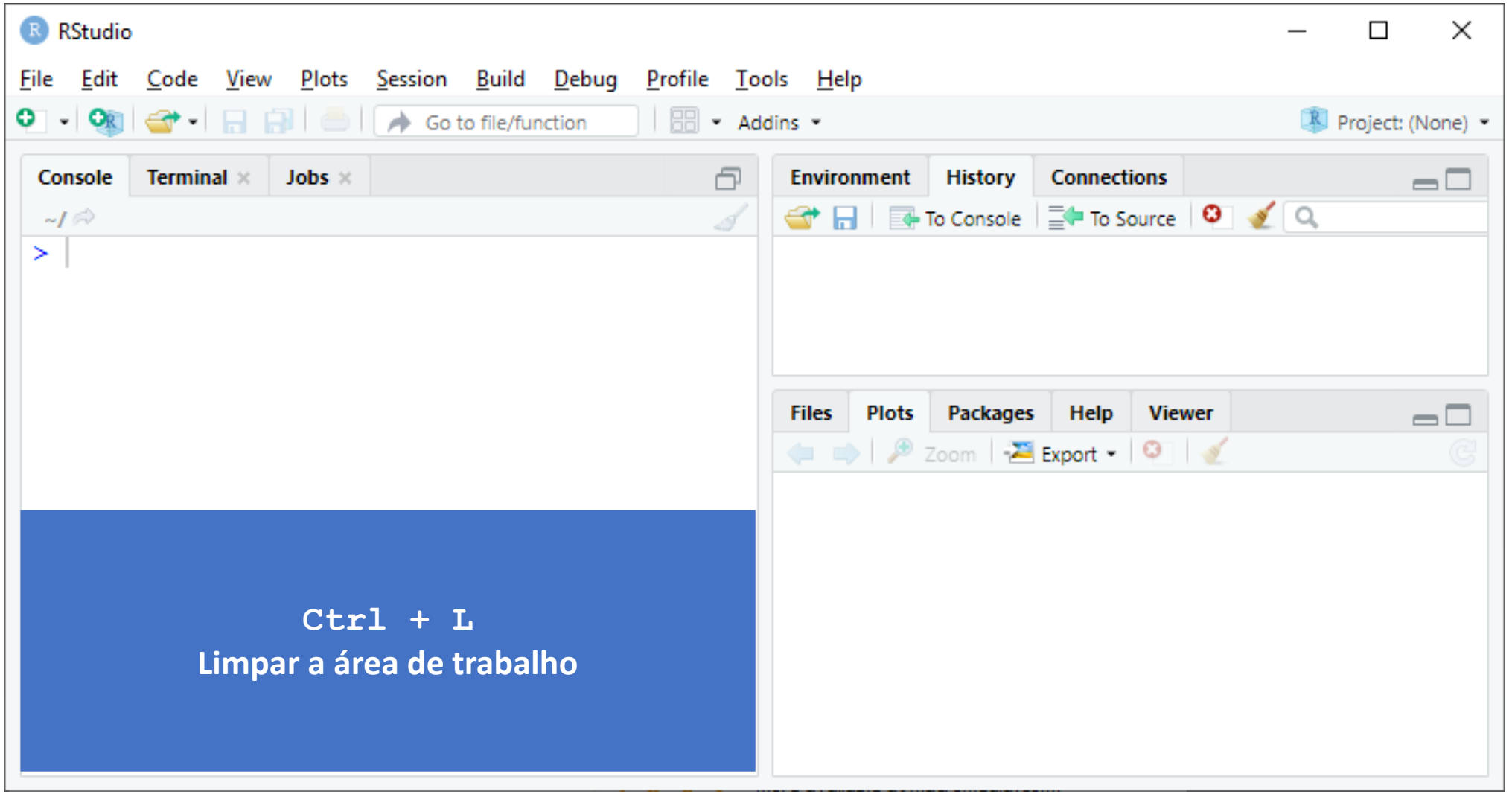
O RStudio está disponível em edições comerciais e de código aberto e pode ser executado nos sistemas operacionais Windows, Mac e Linux ou em um navegador conectado ao RStudio Server ou RStudio Server Pro (Debian, Ubuntu, Red Hat, CentOS e SUSE Linux).

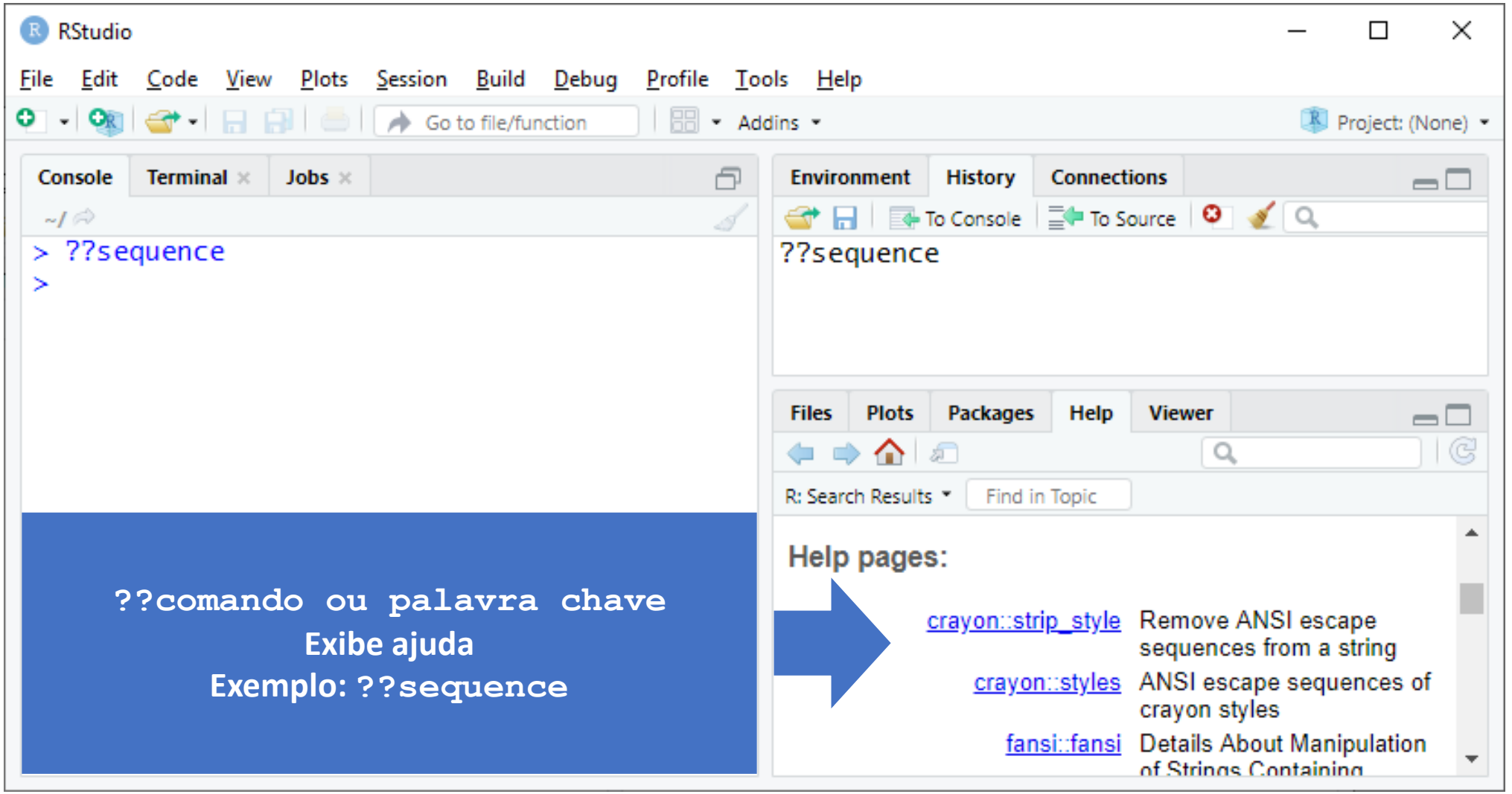
O **download** do RStudio é realizado a partir do seguinte **link**:

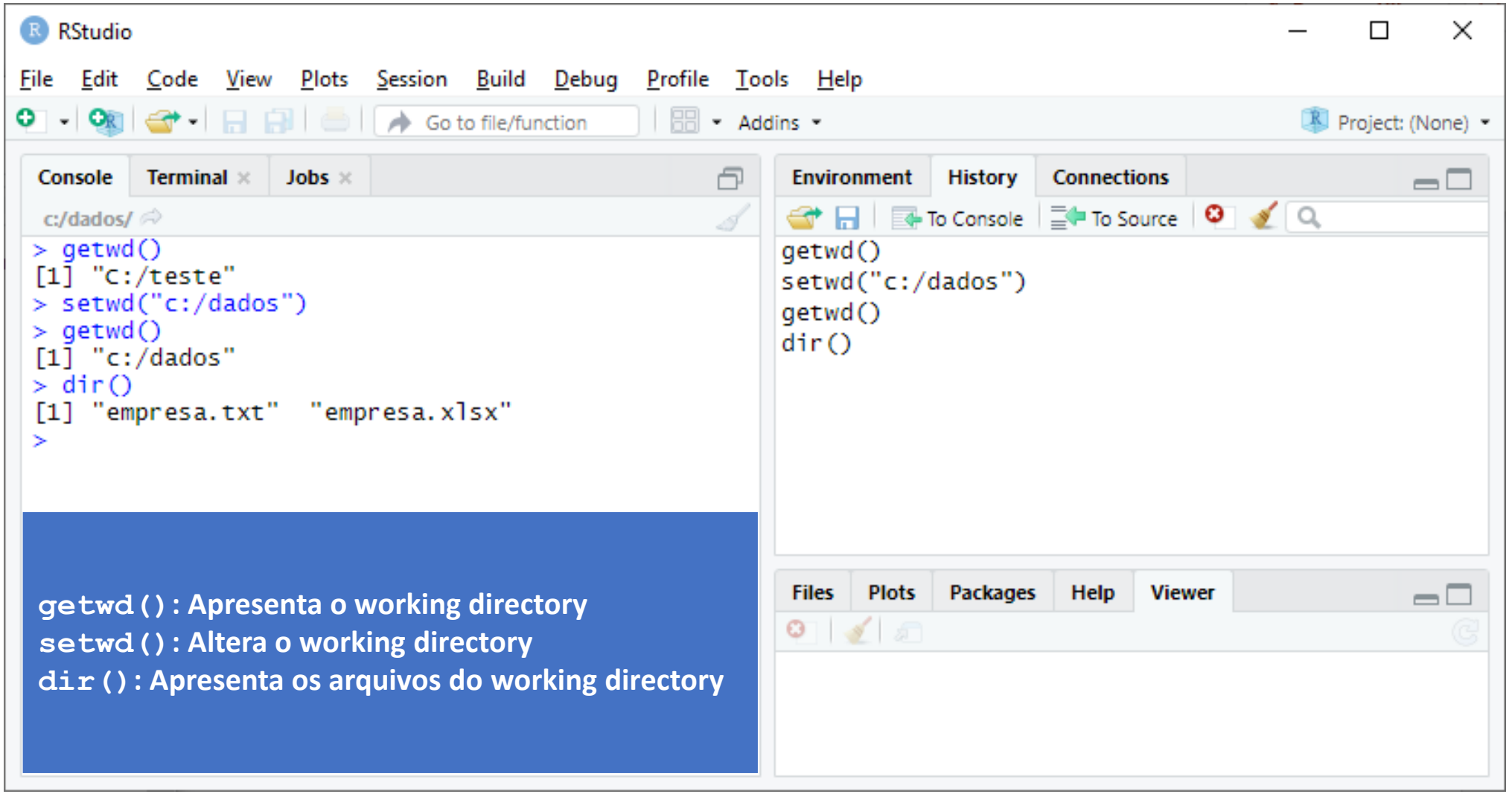
https://rstudio.com/products/rstudio/download/

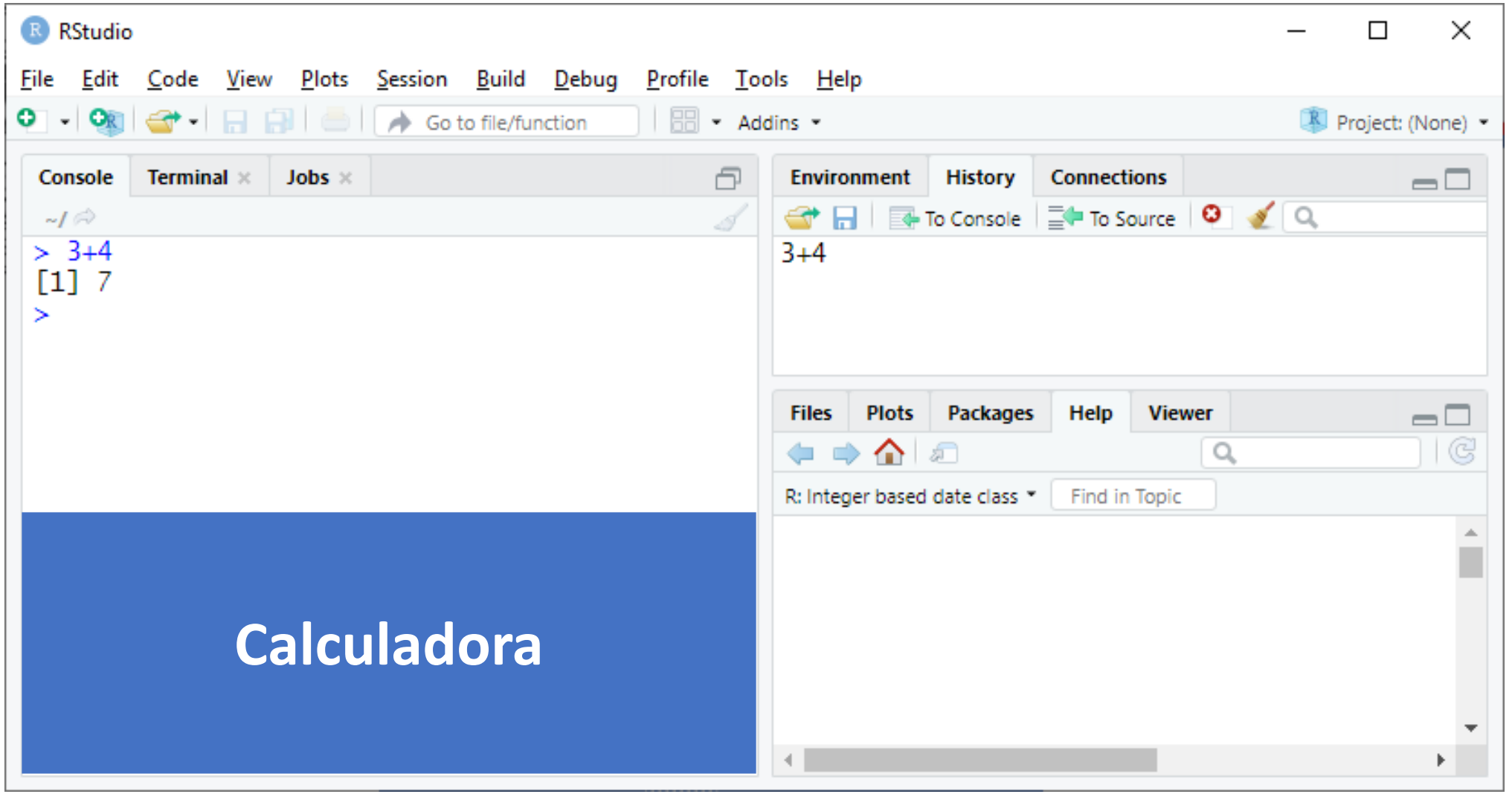


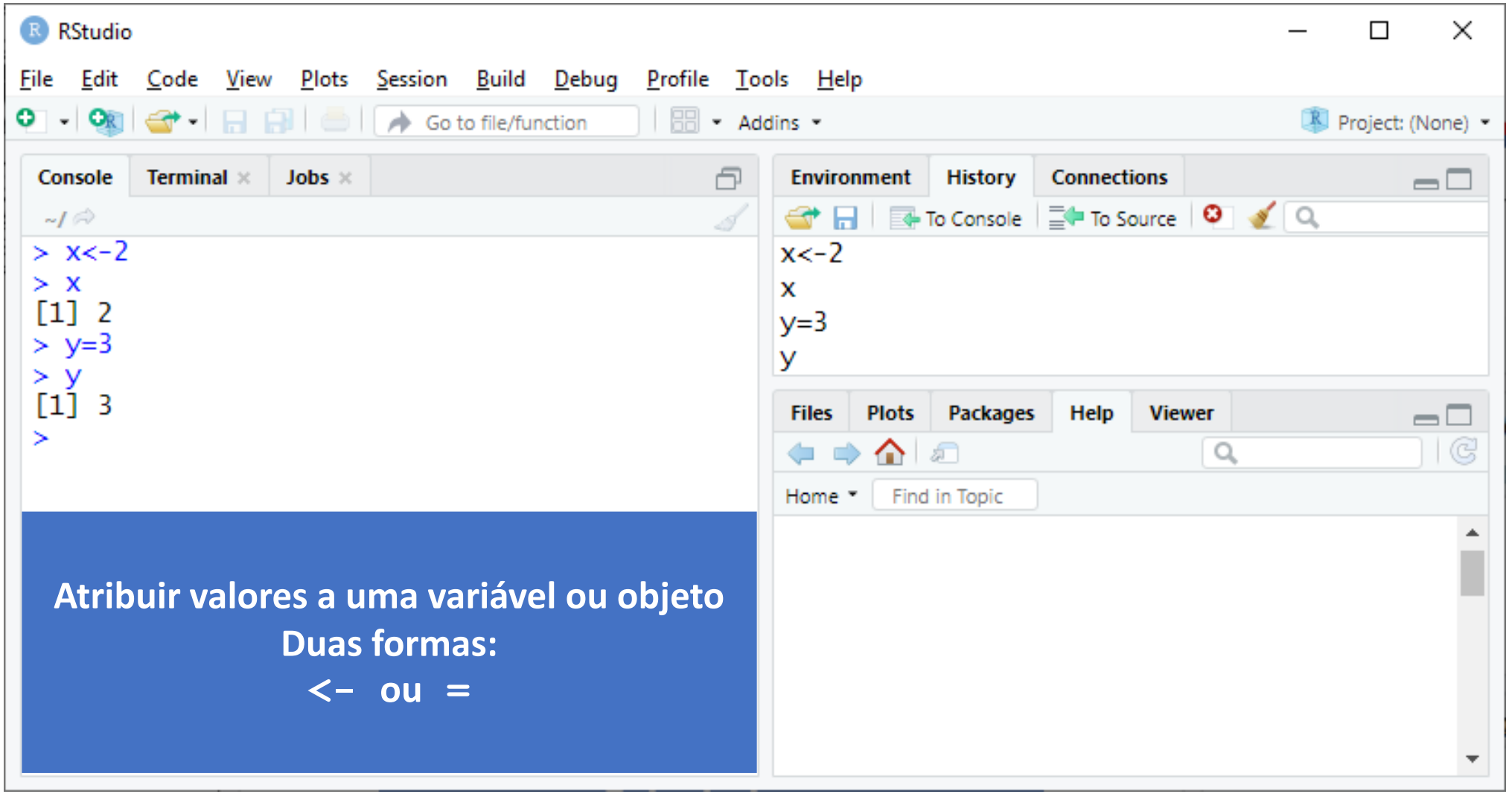
Comandos Básicos





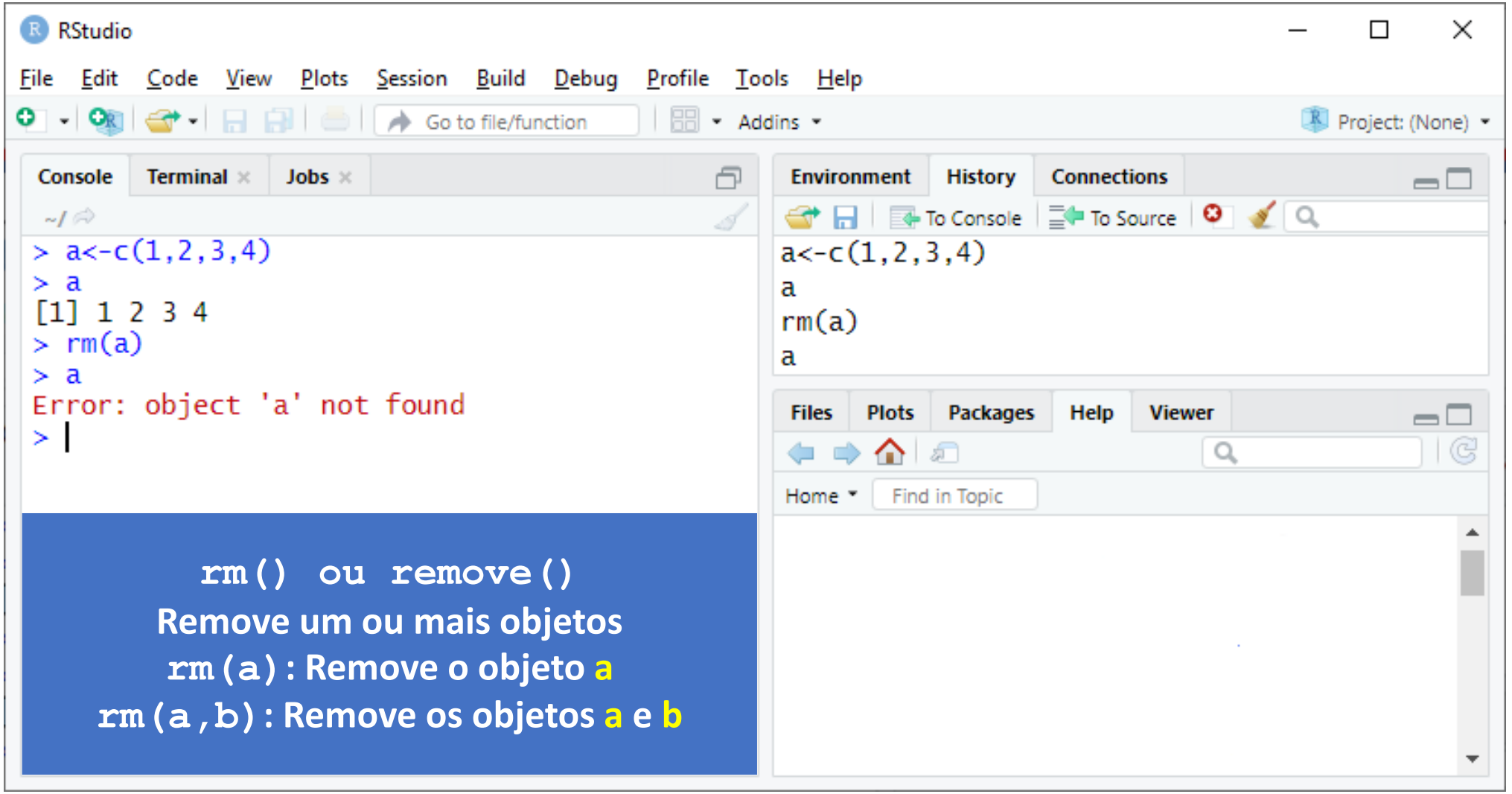


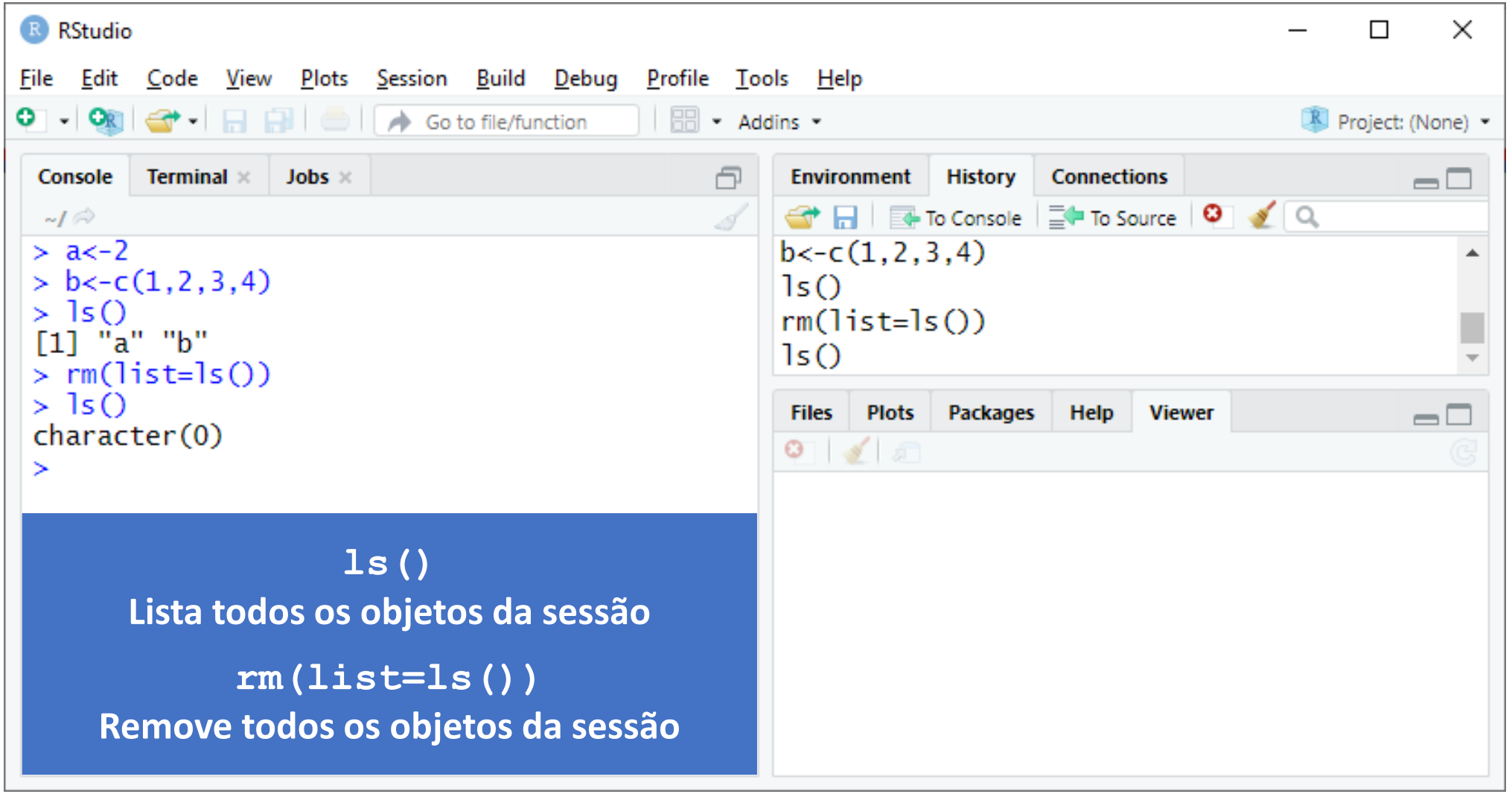




Comandos ou caracteres

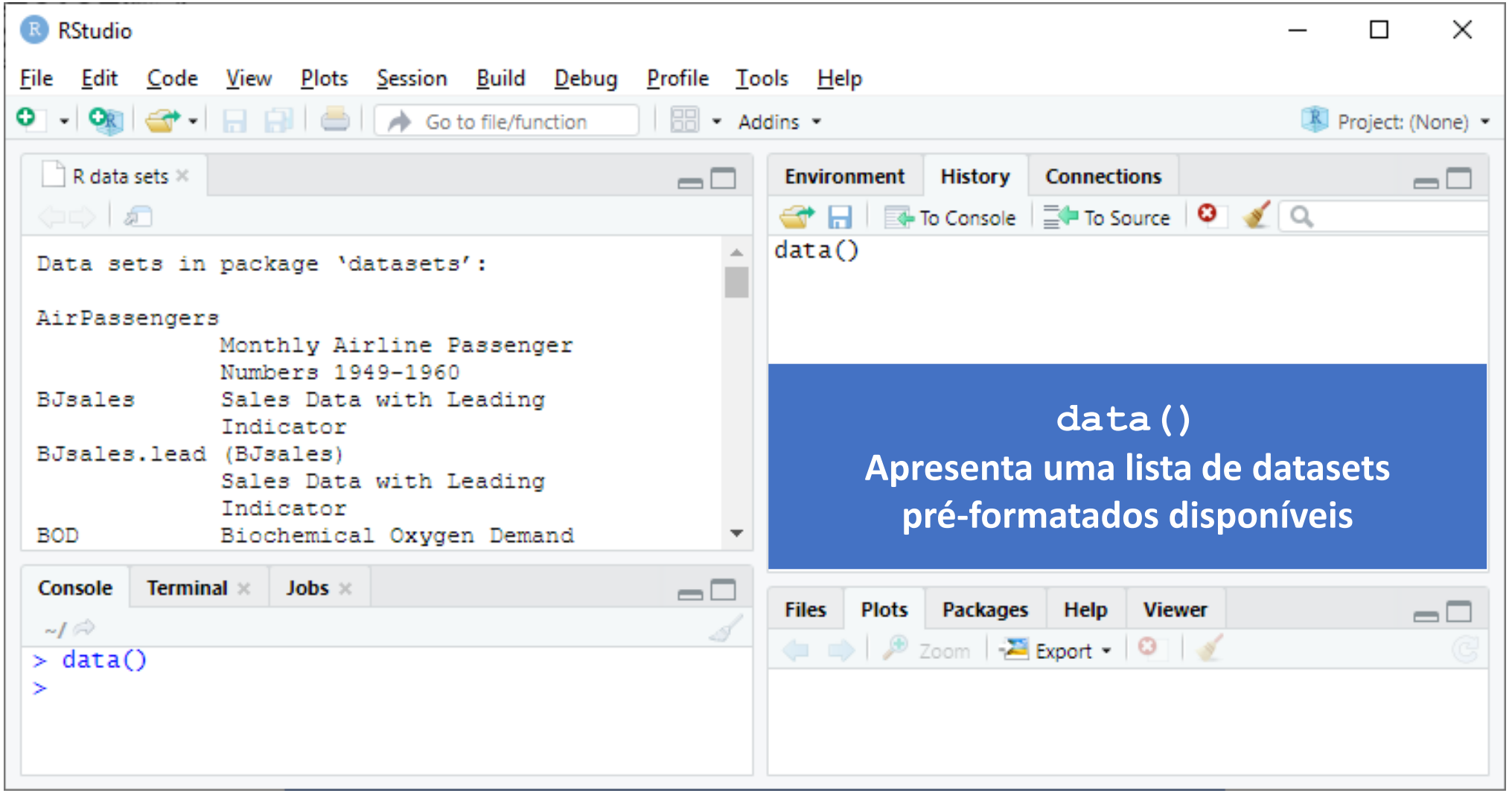
Comandos ou caracteres	Descrição
`#`	Comentário
`;`	Separa dois comando na mesma linha
`NA`	Dado ausente (Not Available)
`NaN`	Não é um Número (Not a Number)
`Inf e -Inf`	Infinito positivo e Infinito negativo
`q()`	Sai do RStudio
`ls()`	Lista todos os objetos da sessão atual
`rm(x)`	Remove o objeto x
`rm(x,y)`	Remove os objetos x e y
`c()`	Concatena valores. Usado para criar vetores. Exemplo: `c(1,2,3,4)`





`ls()`
Lista todos os objetos da sessão

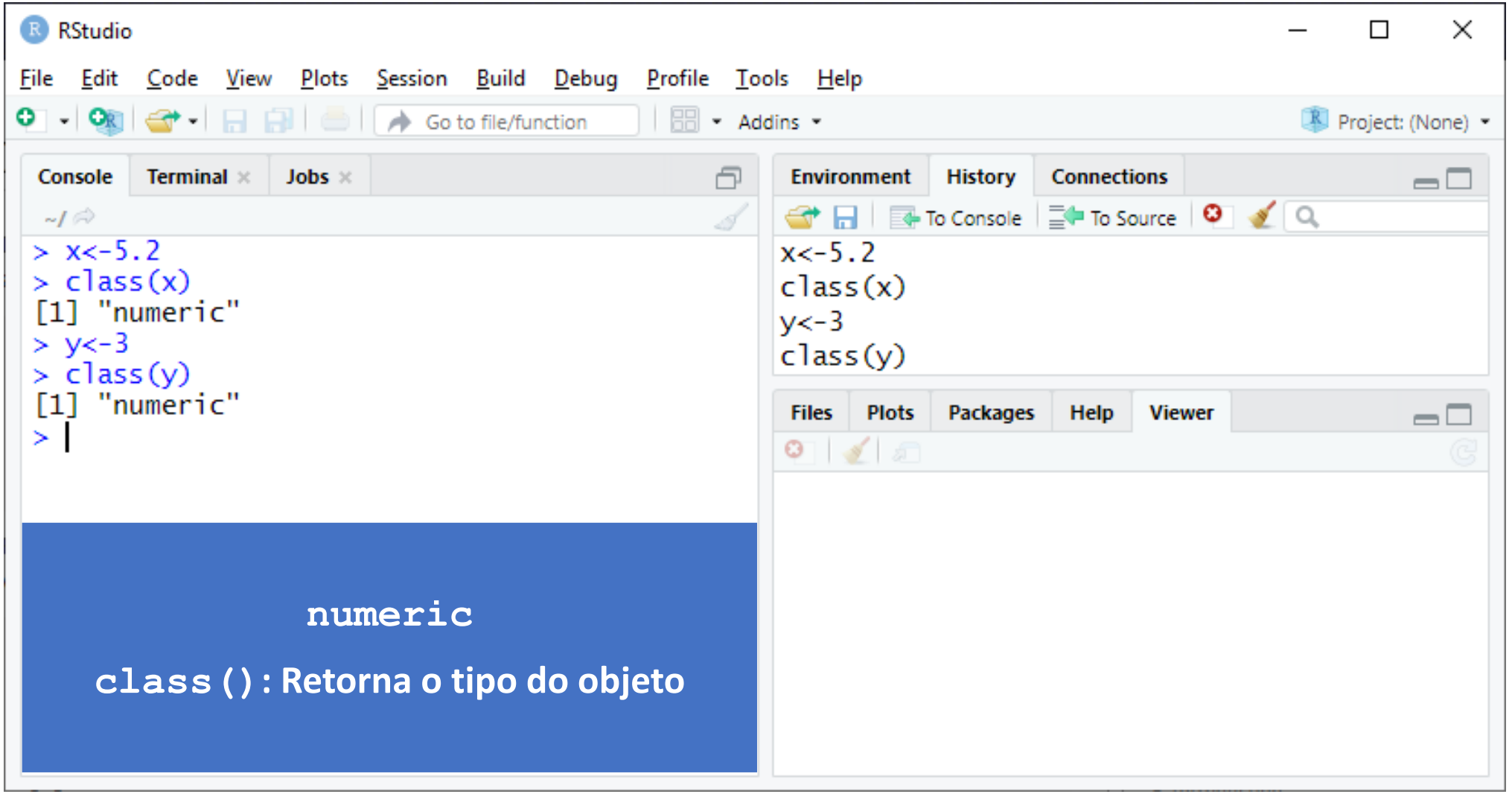
`rm(list=ls())`
Remove todos os objetos da sessão

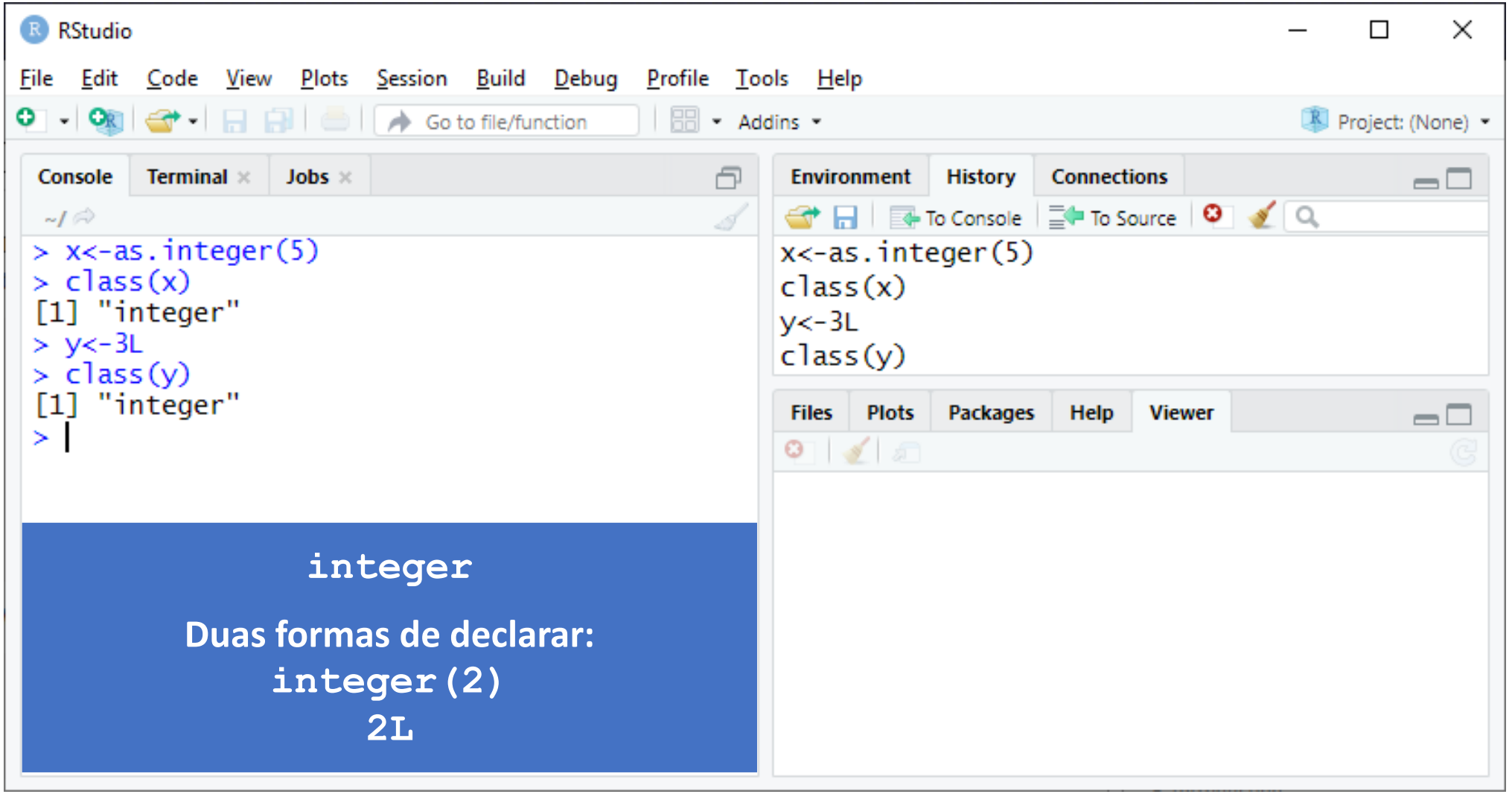

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
R data sets
Data sets in package 'datasets':
AirPassengers
Monthly Airline Passenger
Numbers 1949-1960
BJsales
Sales Data with Leading
Indicator
BJsales.lead (BJsales)
Sales Data with Leading
Indicator
BOD
Biochemical Oxygen Demand
Environment
History
Connections
To Console
To Source
data()
Console
Terminal
Jobs
> data()
>
Files
Plots
Packages
Help
Viewer
Zoom
Export
data()
Apresenta uma lista de datasets
pré-formatados disponíveis

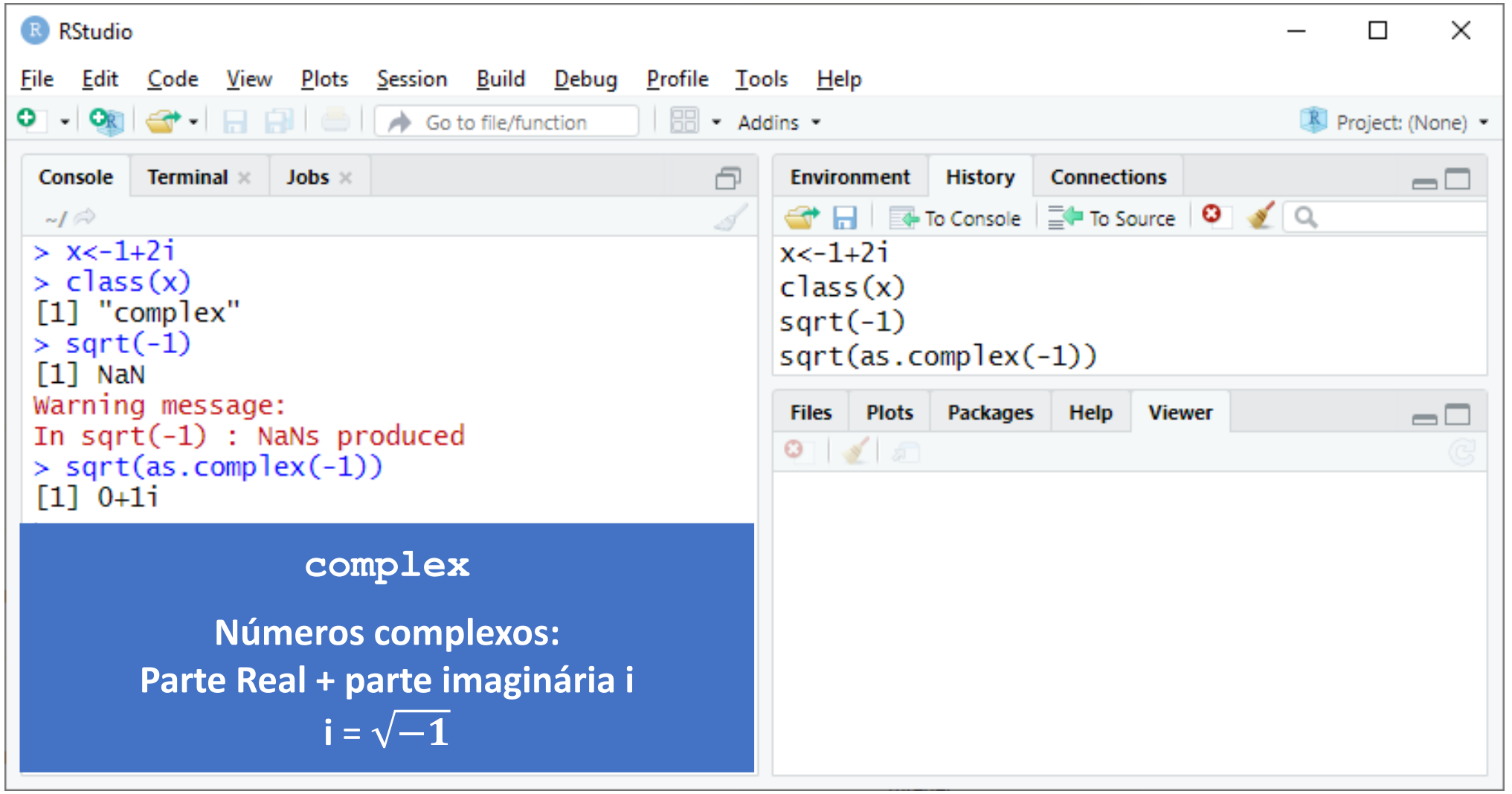
Tipos de Dados

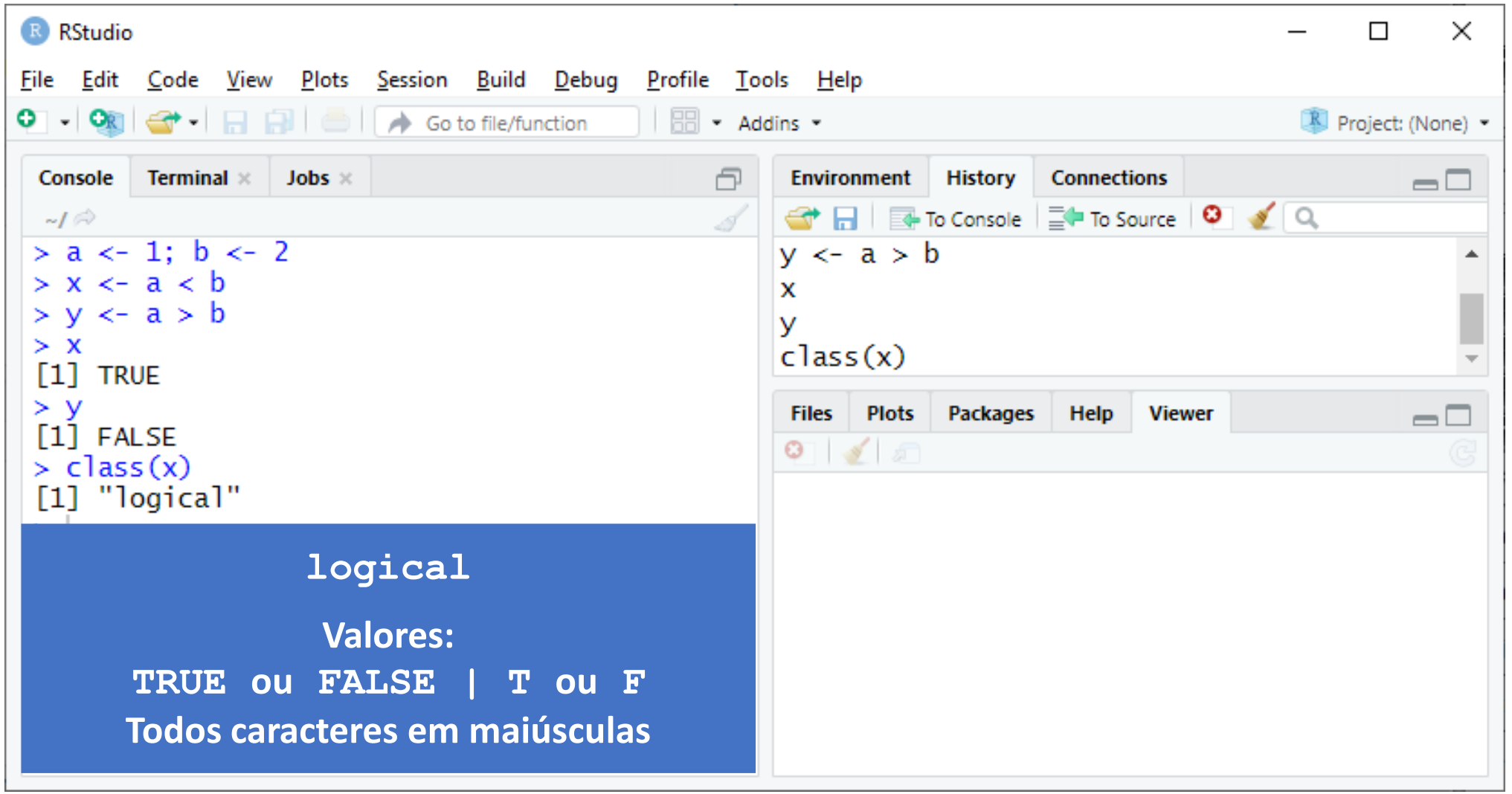
Tipos de Dados

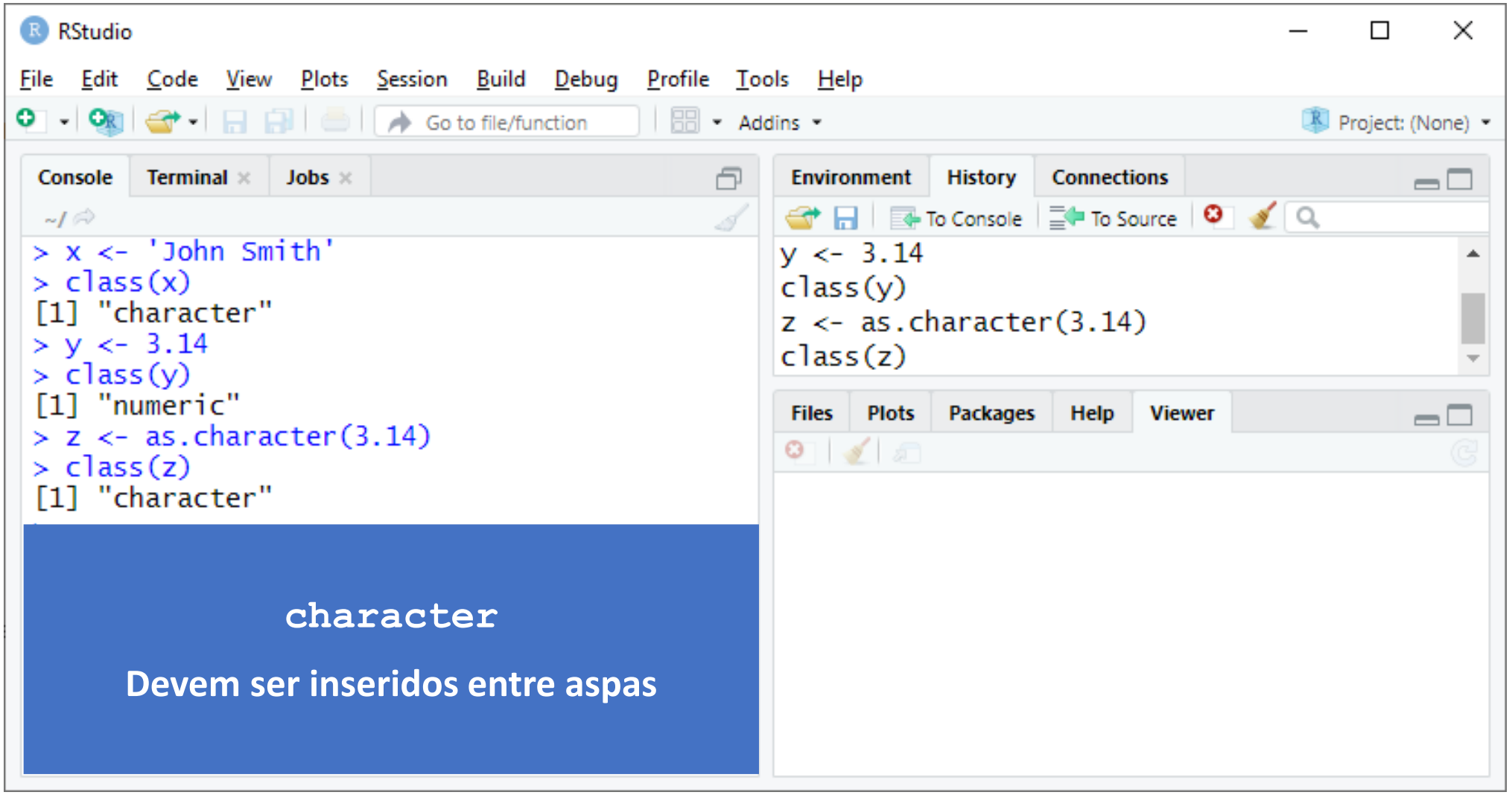
Tipo	Descrição
`numeric`	Números inteiros ou reais
`integer`	Números inteiros
`complex`	Números complexos
`logical`	Booleanos: verdadeiro ou falso (TRUE ou FALSE)
`character`	Caracteres (texto)










RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Console
Terminal
Jobs
> x <- 'John Smith'
> class(x)
[1] "character"
> y <- 3.14
> class(y)
[1] "numeric"
> z <- as.character(3.14)
> class(z)
[1] "character"
Environment
History
Connections
To Console
To Source
y <- 3.14
class(y)
z <- as.character(3.14)
class(z)
Files
Plots
Packages
Help
Viewer
character
Devem ser inseridos entre aspas

Adicionando Pacotes

Adicionando Pacotes (Packages)

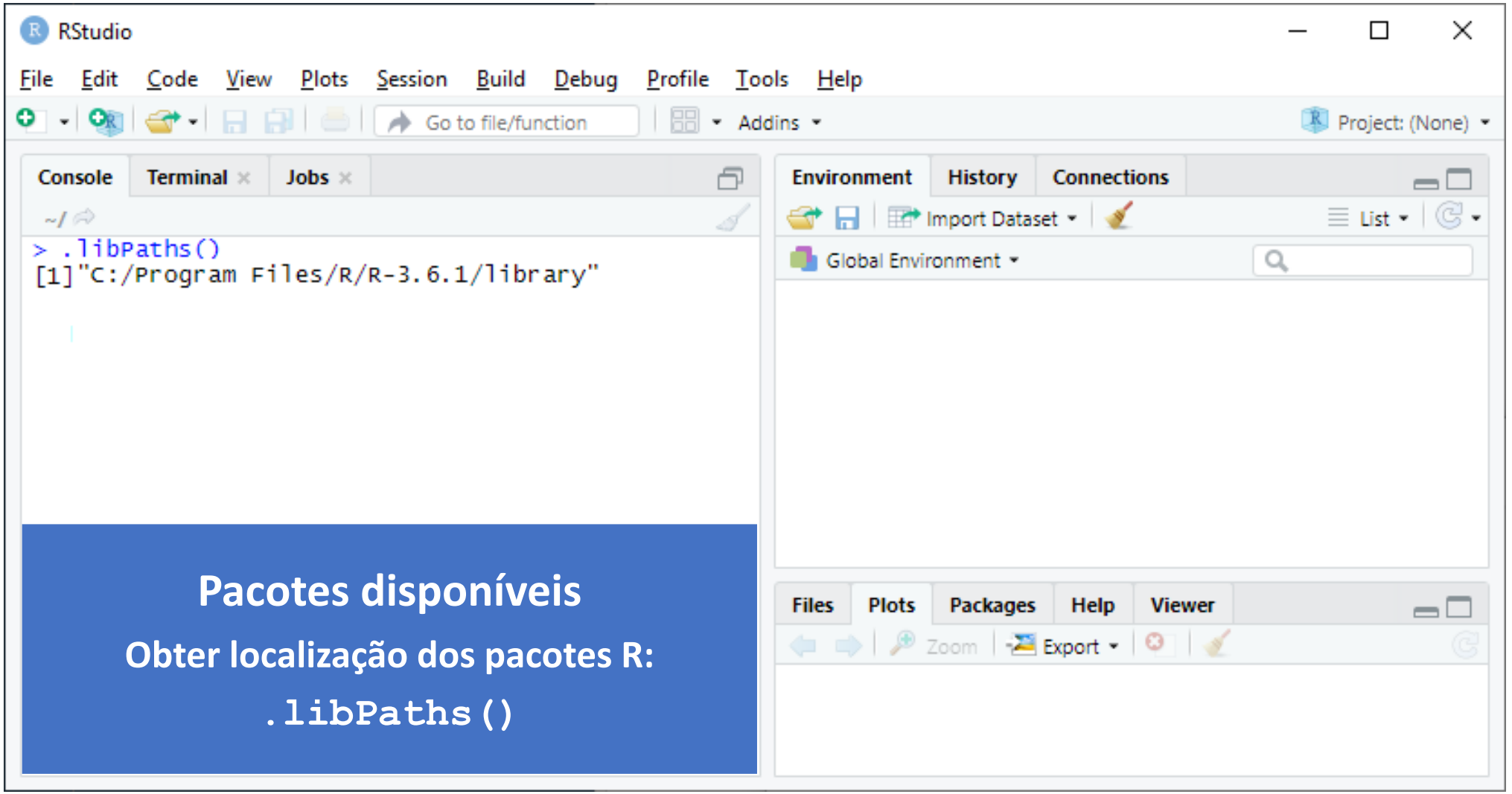
Packages são compostos por uma coleção de funções R, código compilado e dados de amostra, armazenados no diretório "library" no ambiente R.

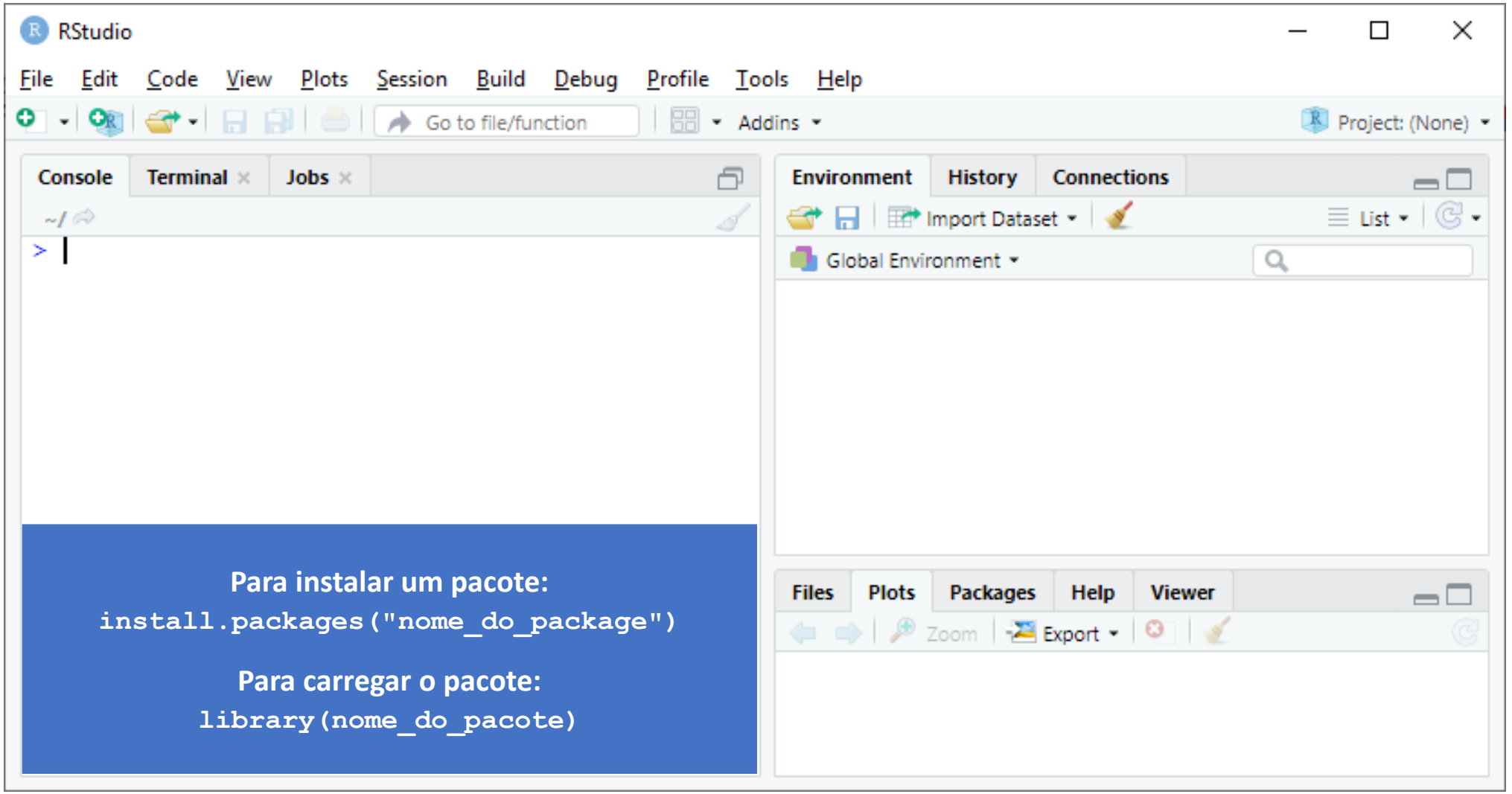
Um conjunto padrão de packages é instalado. Outros pacotes com funções específicas precisam ser instalados, conforme a necessidade.

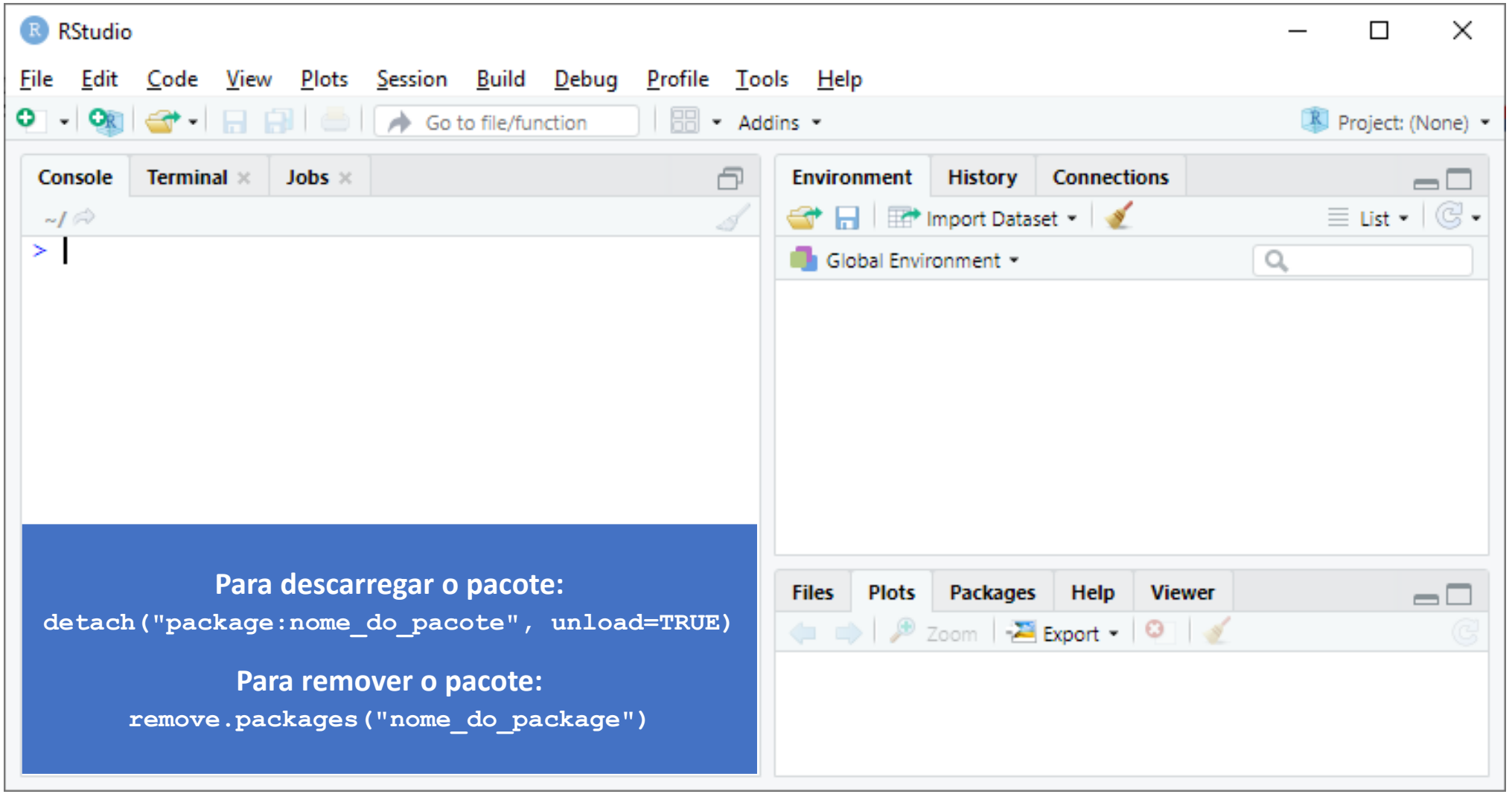
Quando o console R é inicializado, apenas o conjunto padrão de pacotes fica disponível. Outros pacotes, mesmo que já estejam instalados, devem ser carregados explicitamente para serem usados.

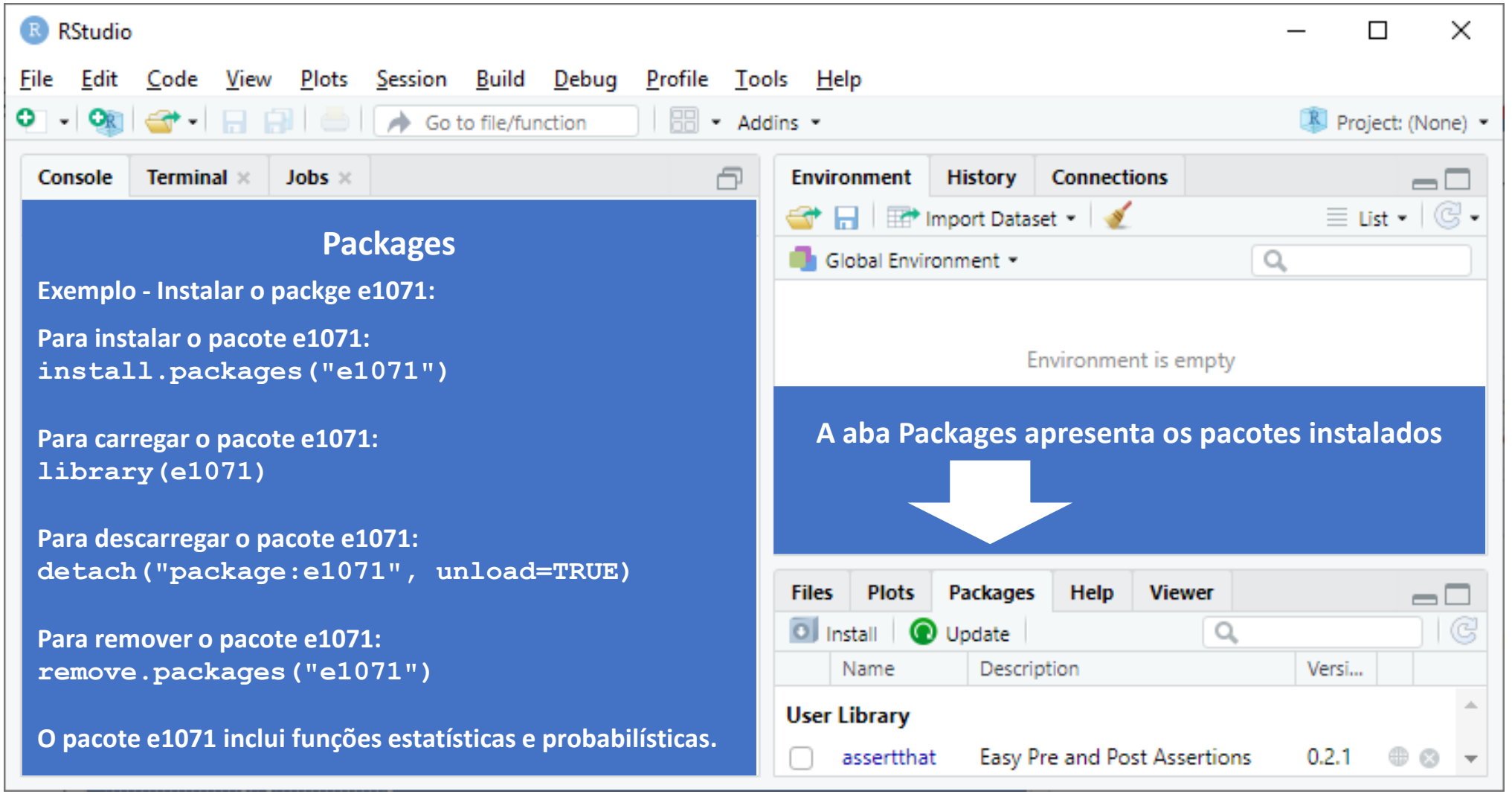
Todos os pacotes disponíveis na linguagem R estão no seguinte link:

https://cran.r-project.org/web/packages/available_packages_by_name.html/






RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Console
Terminal
Jobs
Environment
History
Connections
Import Dataset
List
Global Environment
Files
Plots
Packages
Help
Viewer
Zoom
Export
Para descarregar o pacote:
detach("package:nome_do_pacote", unload=TRUE)
Para remover o pacote:
remove.packages("nome_do_package")

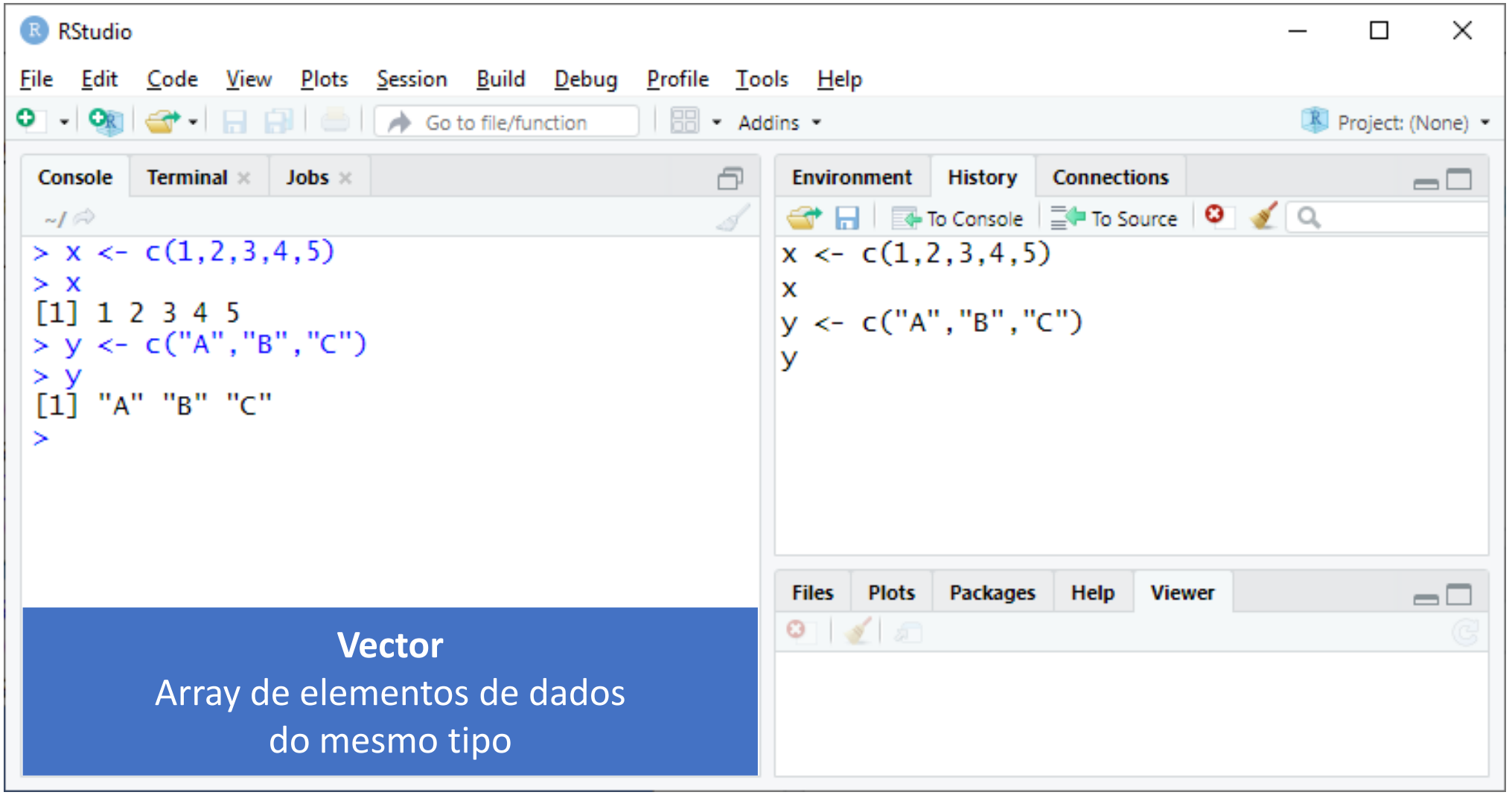


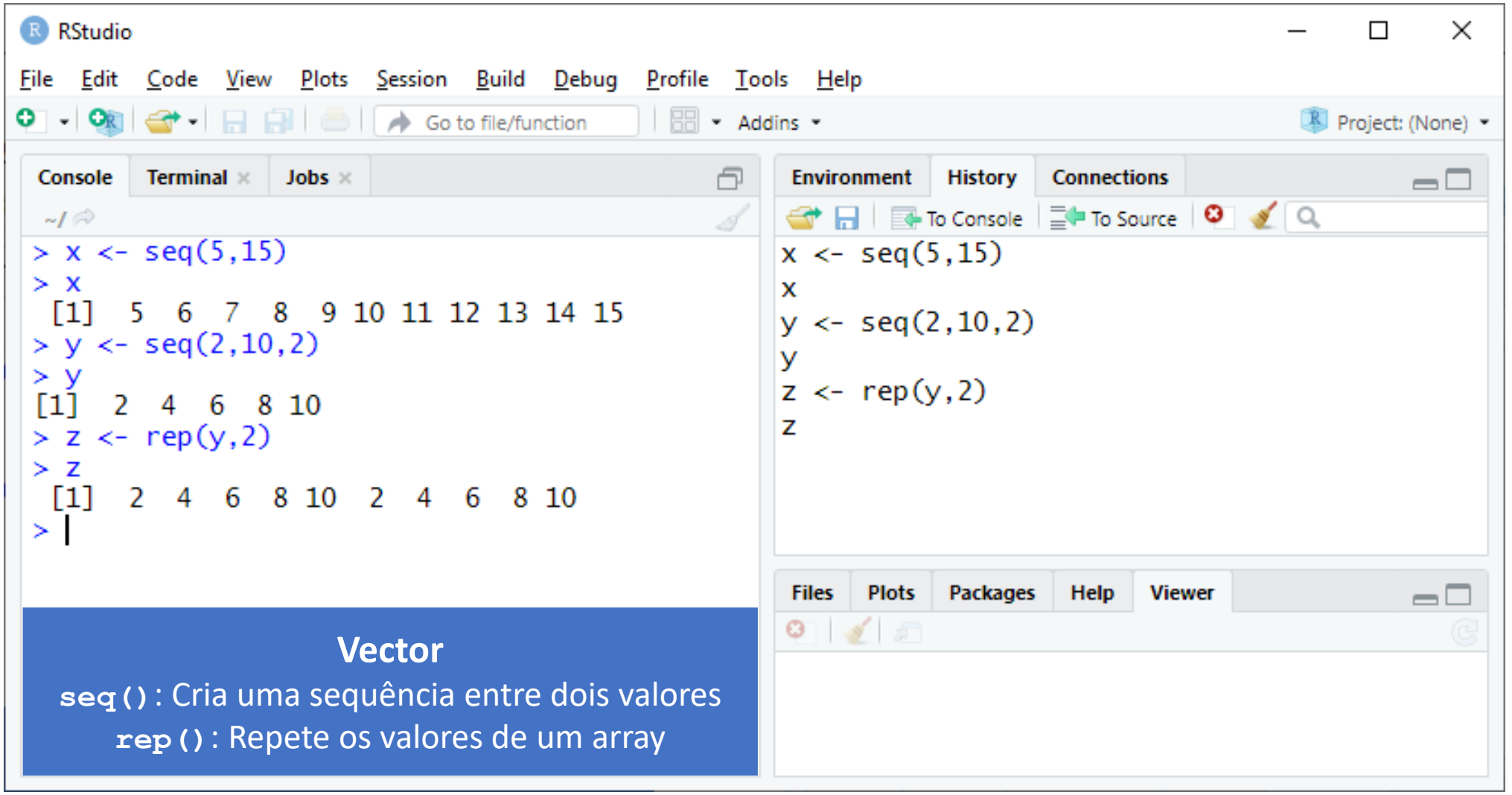
Estruturas Básicas do R

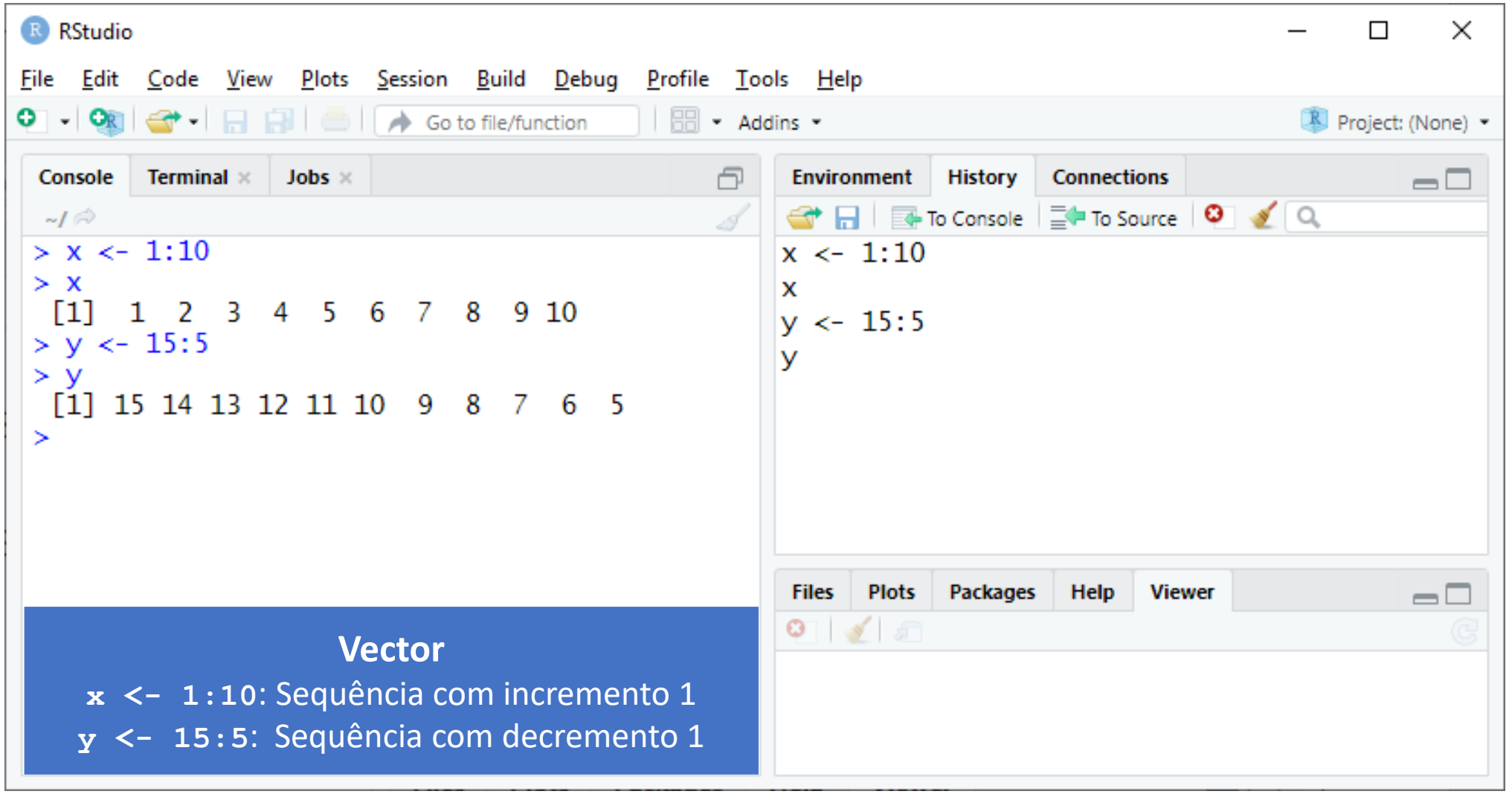
Estruturas Básicas do R

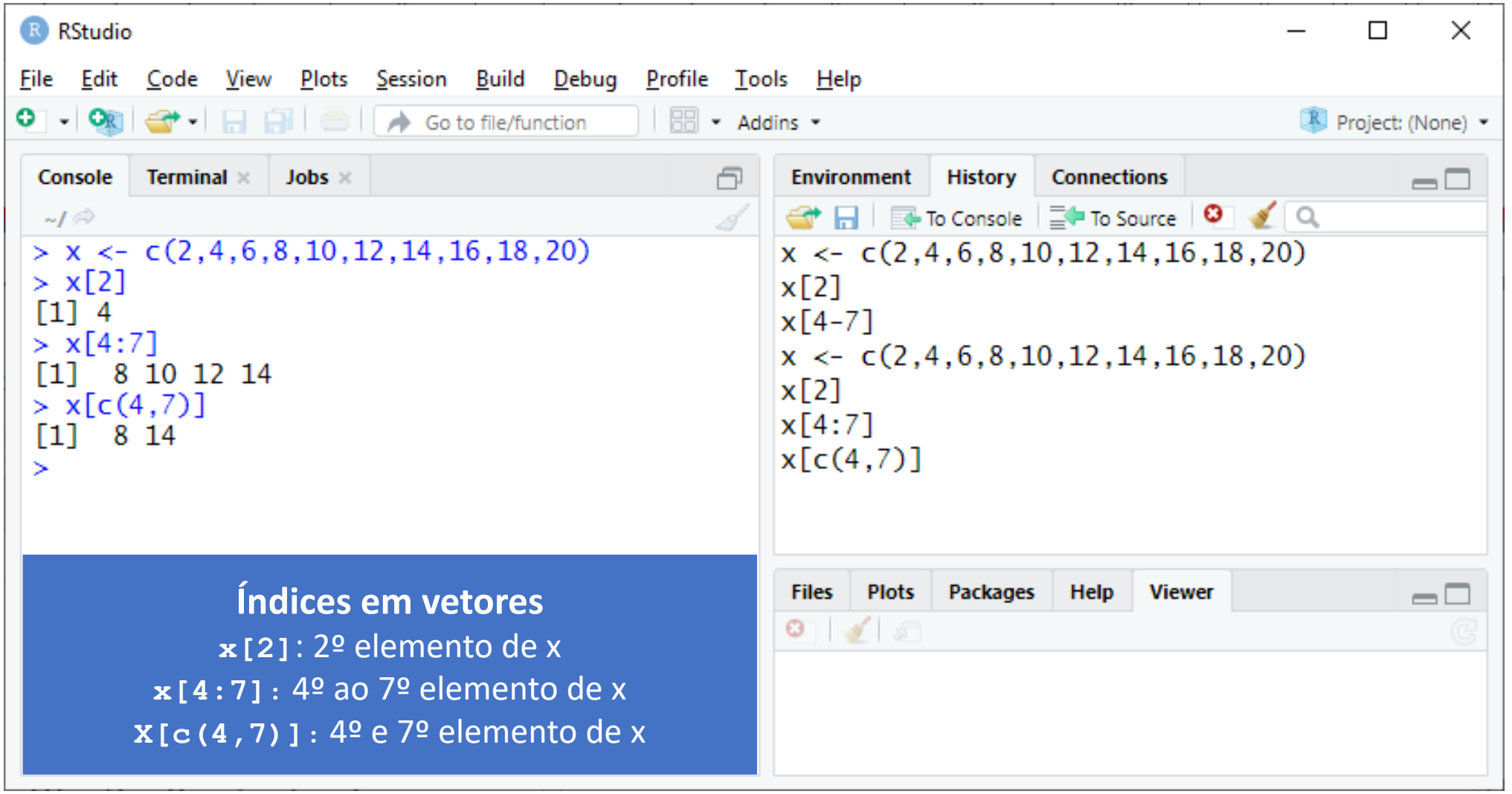
Estrutura	Descrição
`vector`	Array com uma dimensão (uma 'linha')
`matrix`	Array com duas dimensões ('linhas' e 'colunas')
`array`	Objeto que pode conter uma ou mais dimensões
`factor`	Vetor de dados categóricos
`data.frame`	Objeto com duas dimensões (armazena elementos de diferentes tipos)
`list`	Objeto que permite combinar diferentes estruturas

Vetores

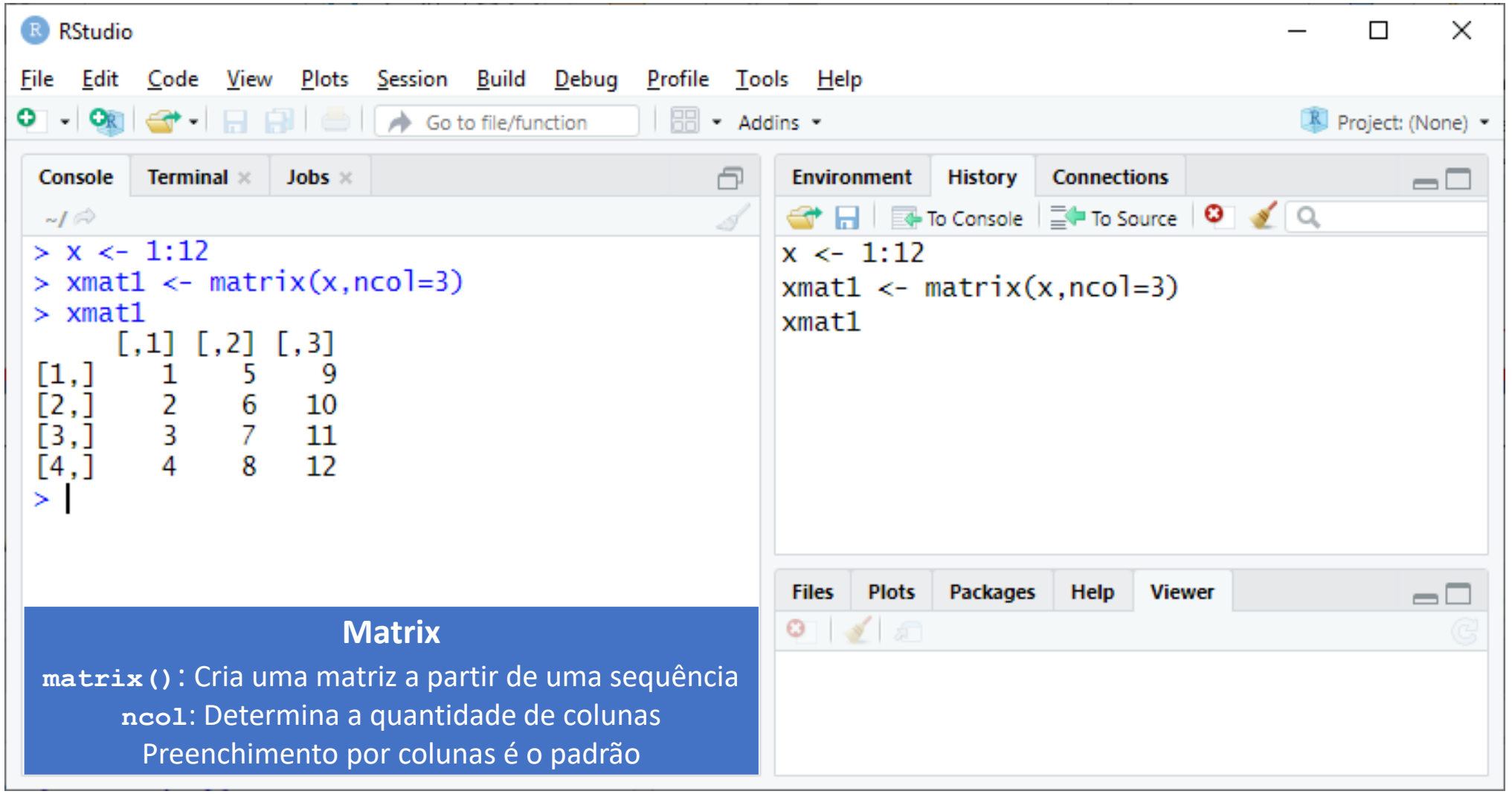


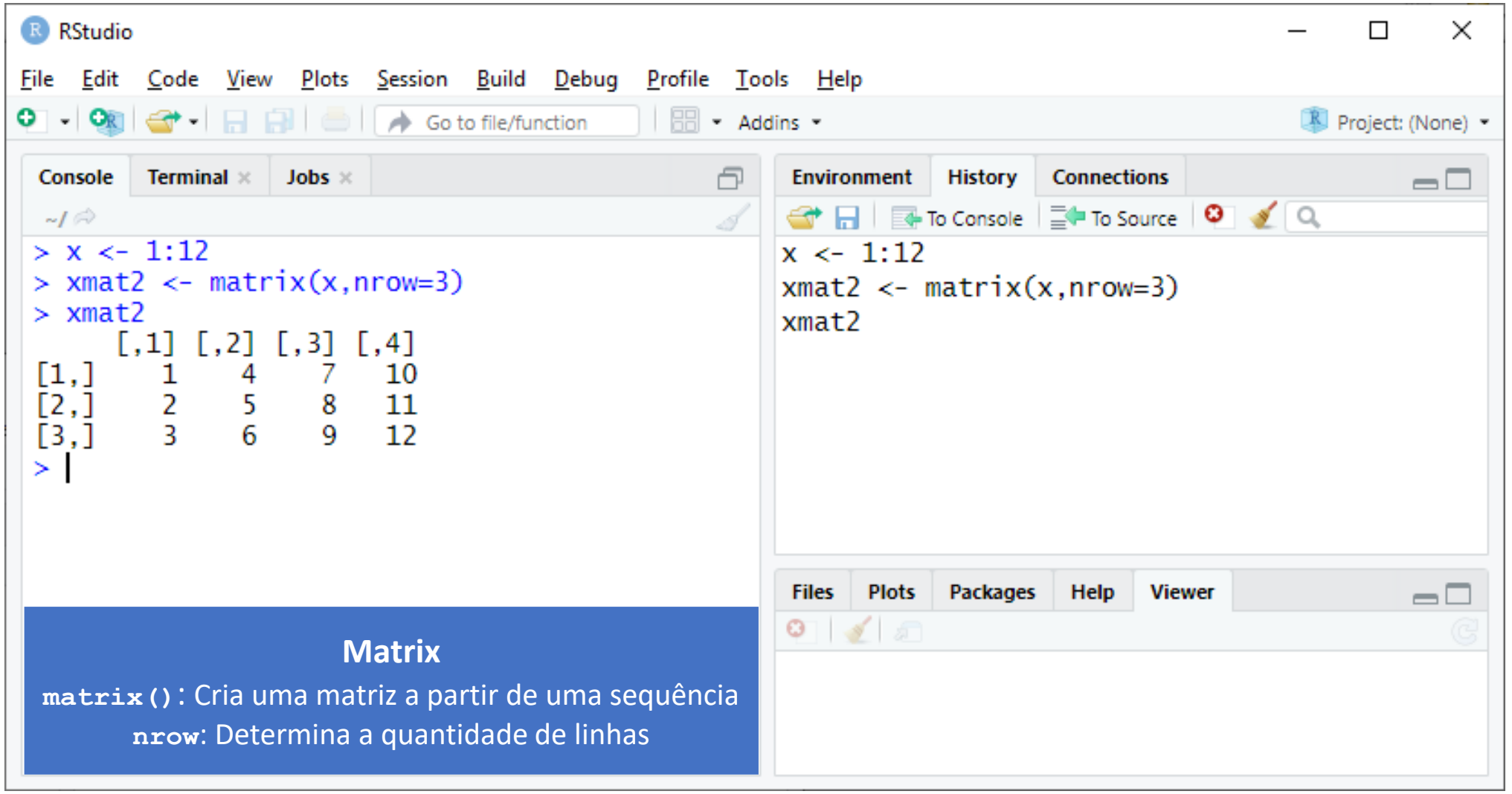


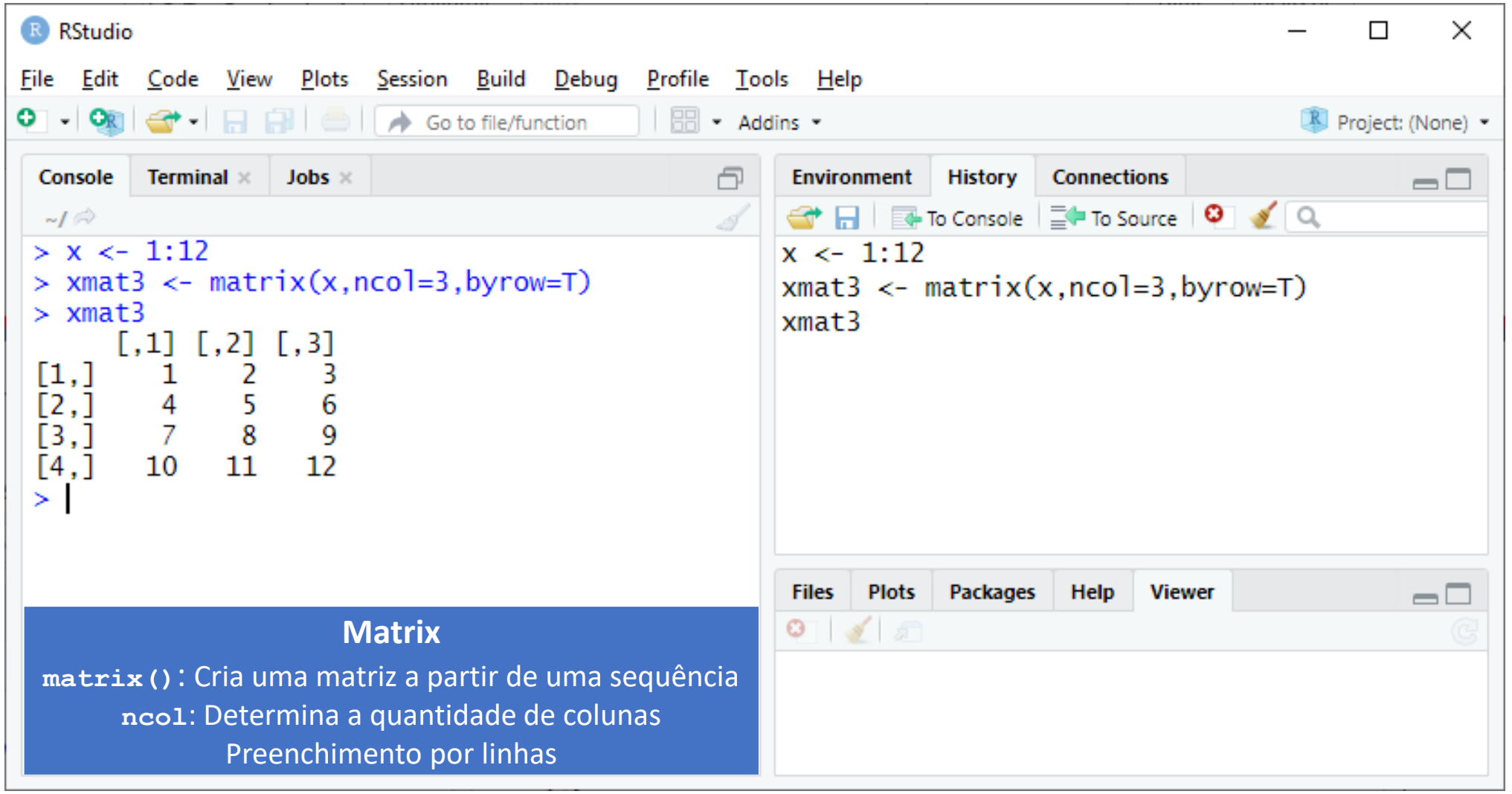




Matrizes





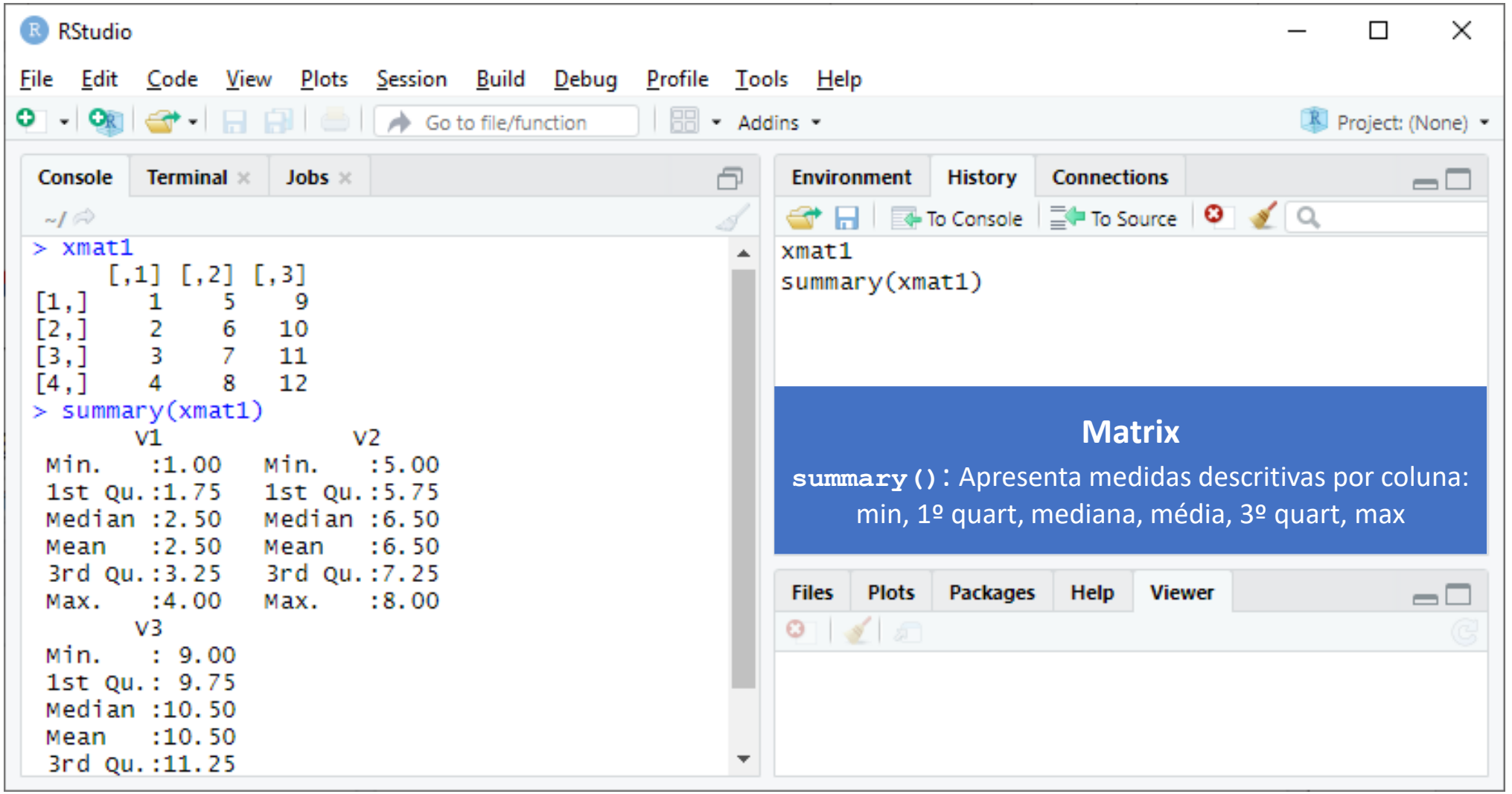


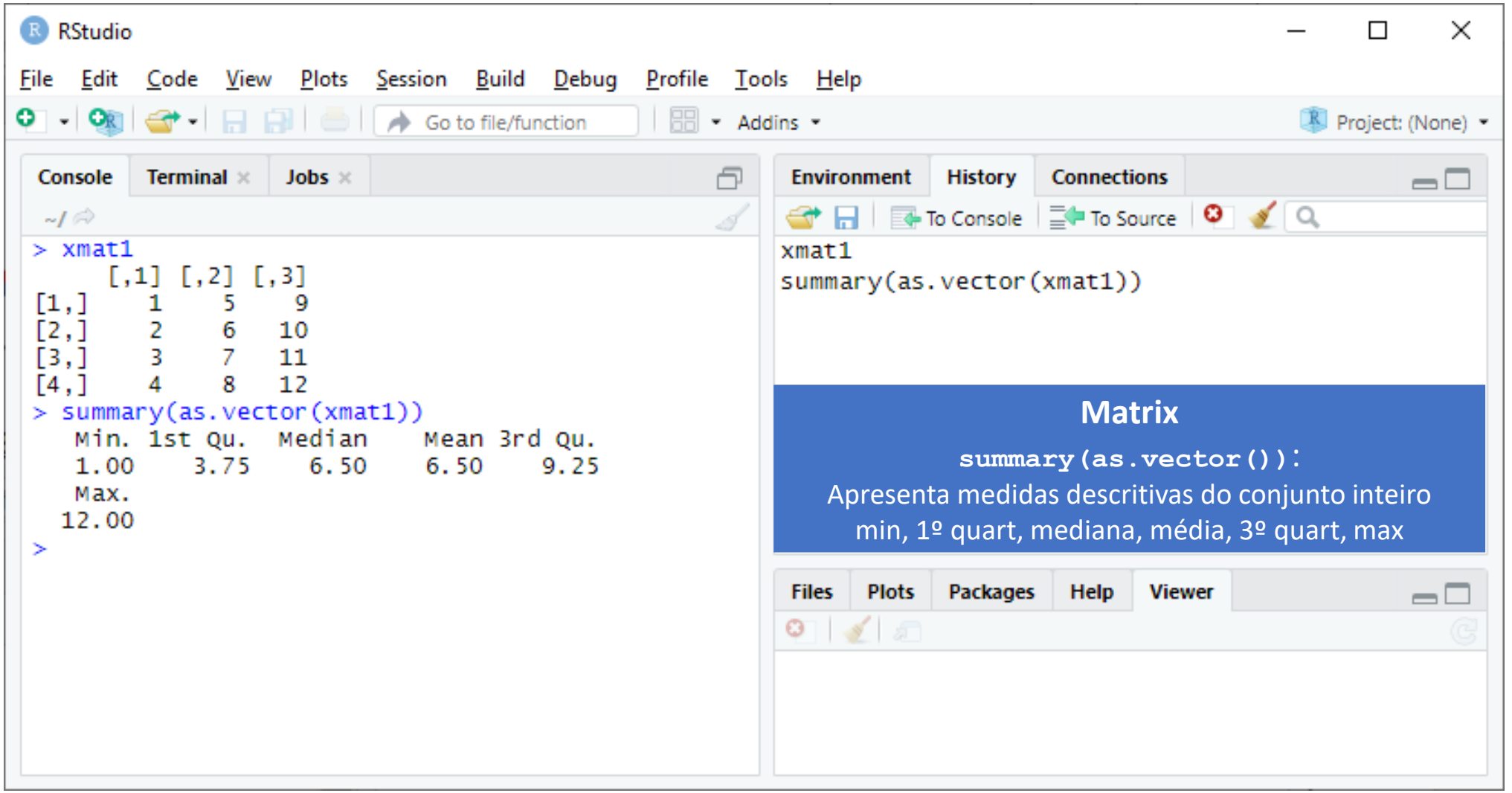
Matrix

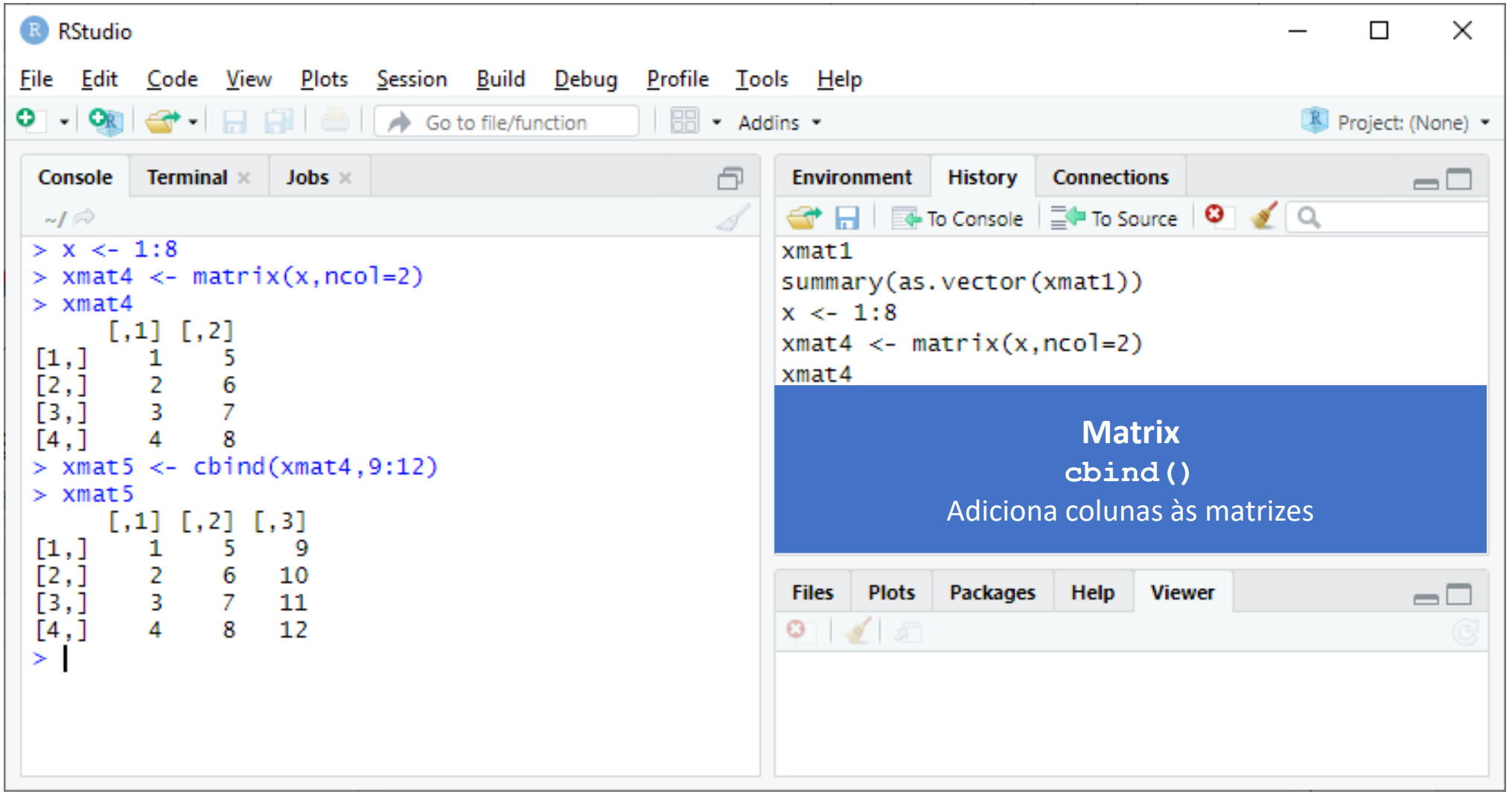
`matrix()`: Cria uma matriz a partir de uma sequência

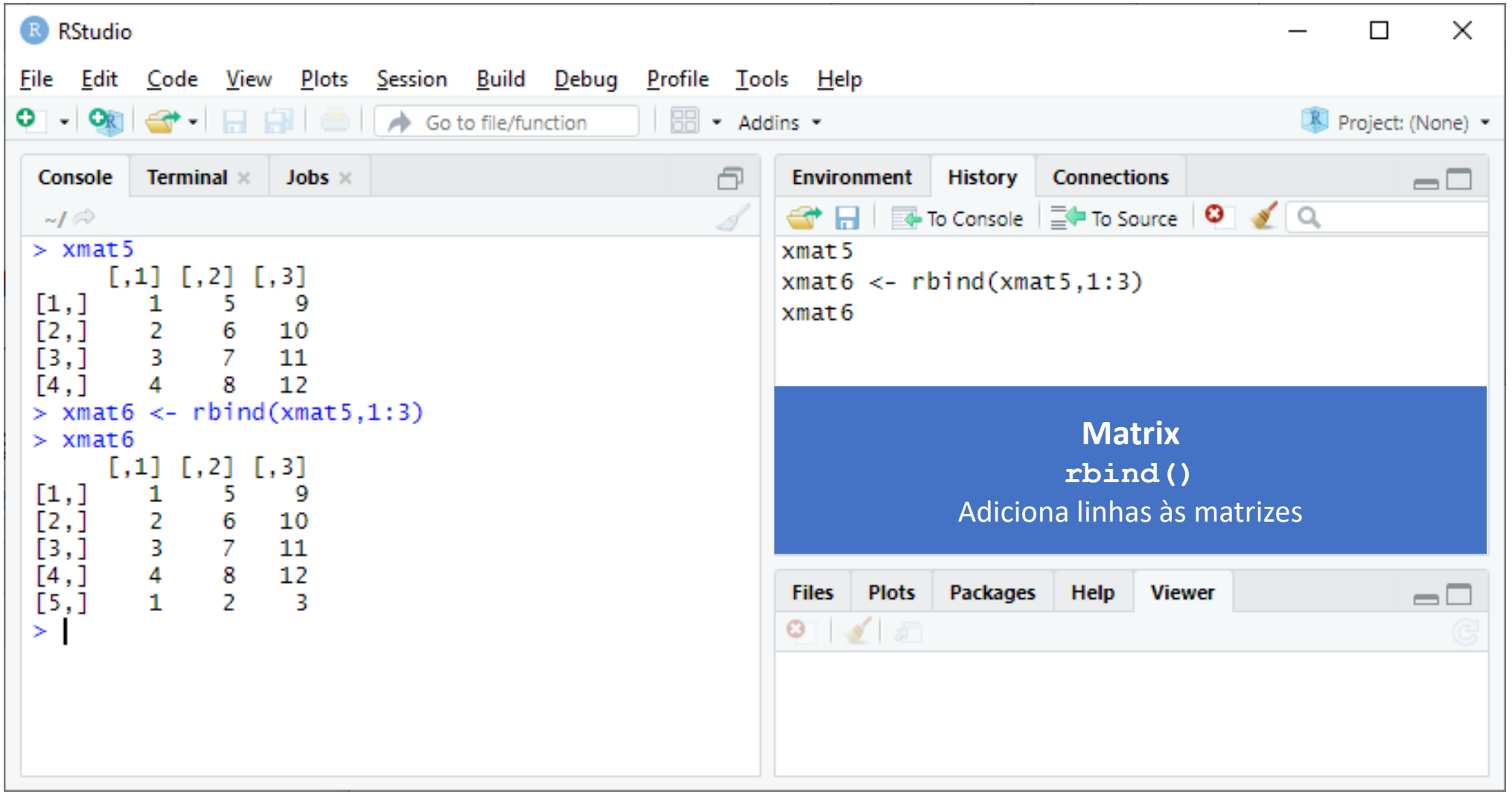
`ncol`: Determina a quantidade de colunas

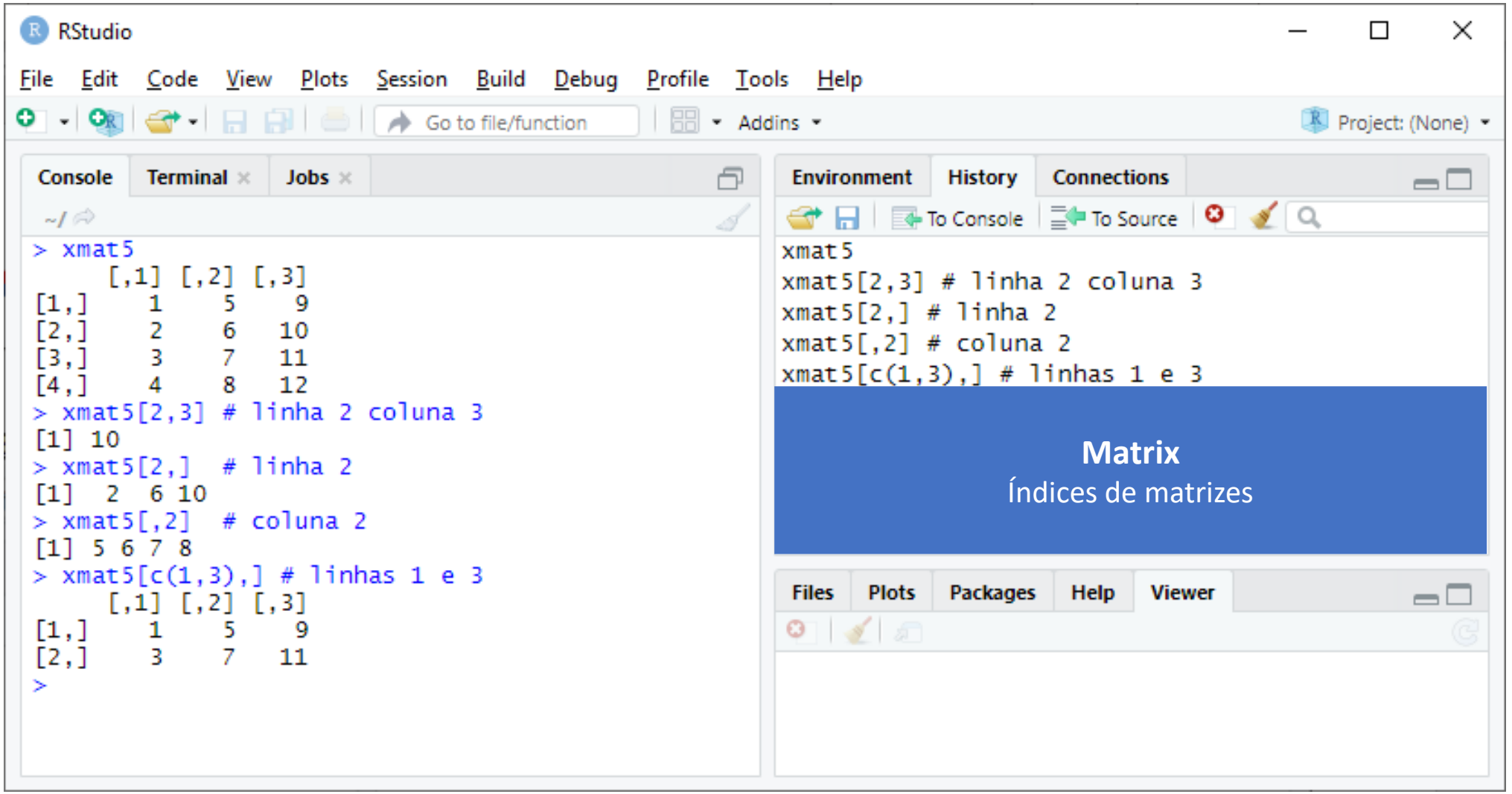
Preenchimento por linhas

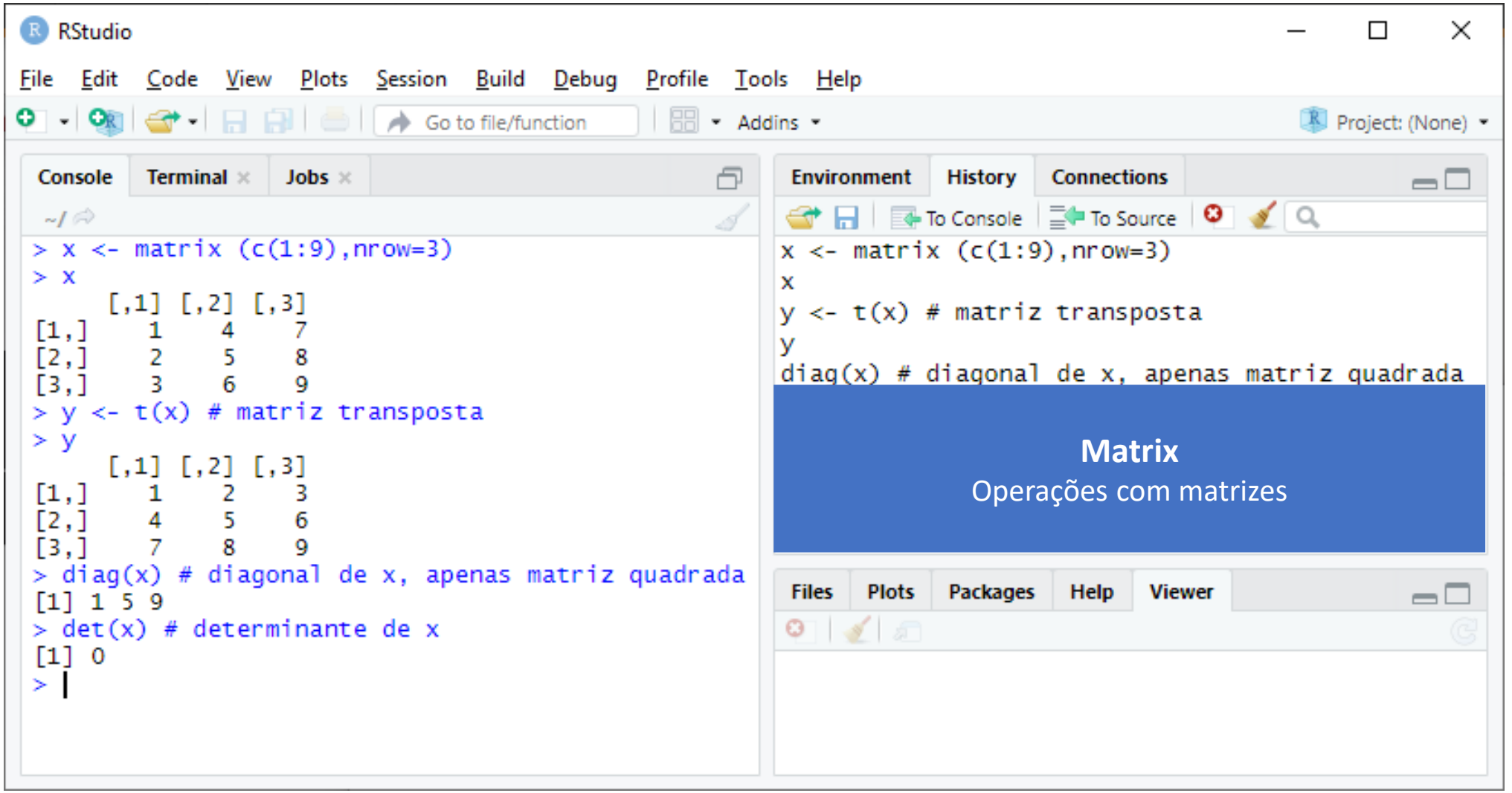




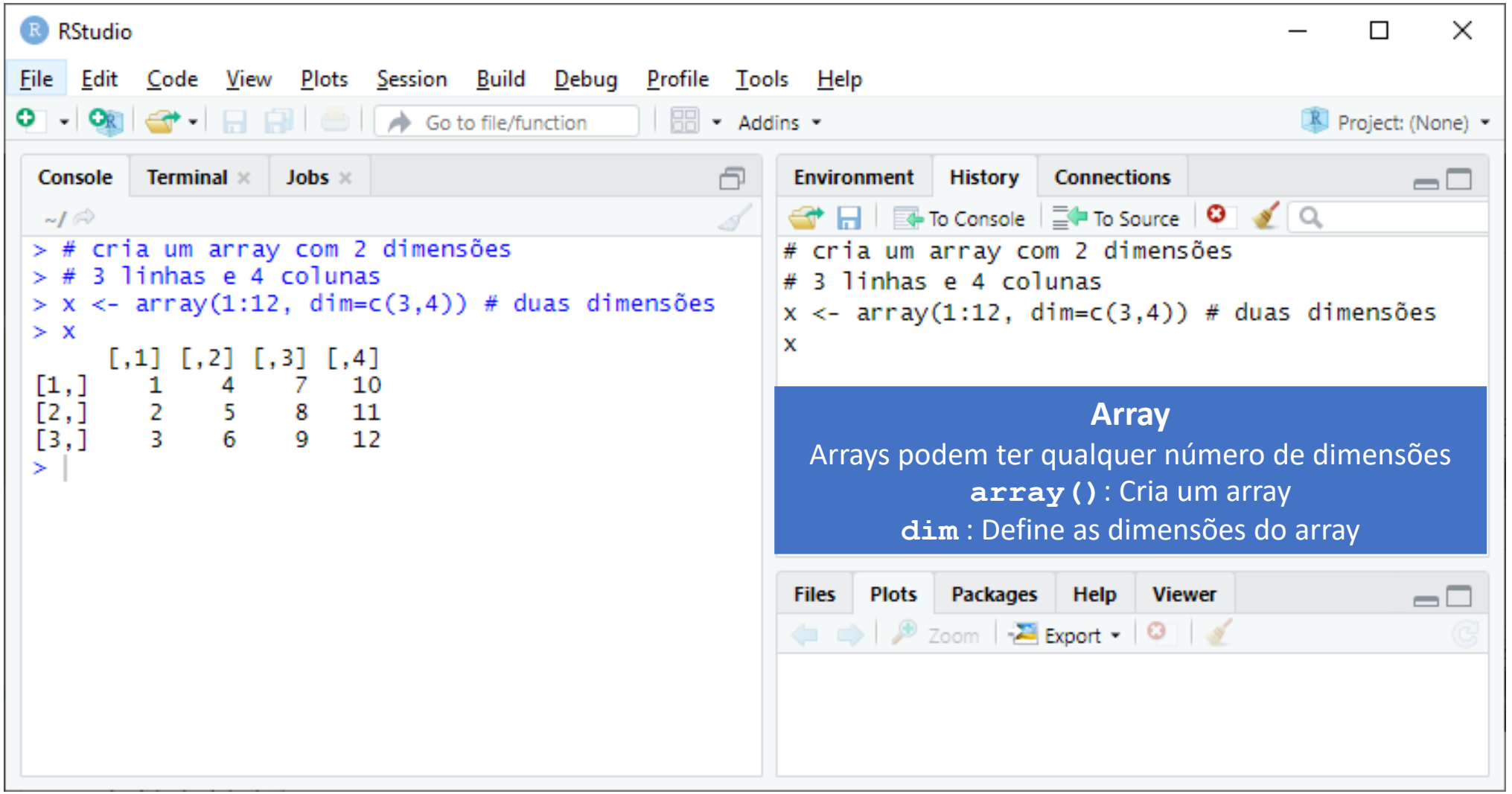



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Console
Terminal
Jobs
> xmat5
[,1] [,2] [,3]
[1,] 1 5 9
[2,] 2 6 10
[3,] 3 7 11
[4,] 4 8 12
> xmat6 <- rbind(xmat5,1:3)
> xmat6
[,1] [,2] [,3]
[1,] 1 5 9
[2,] 2 6 10
[3,] 3 7 11
[4,] 4 8 12
[5,] 1 2 3
>
Environment
History
Connections
To Console
To Source
xmat5
xmat6 <- rbind(xmat5,1:3)
xmat6
Matrix
rbind()
Adiciona linhas às matrizes
Files
Plots
Packages
Help
Viewer





Arrays

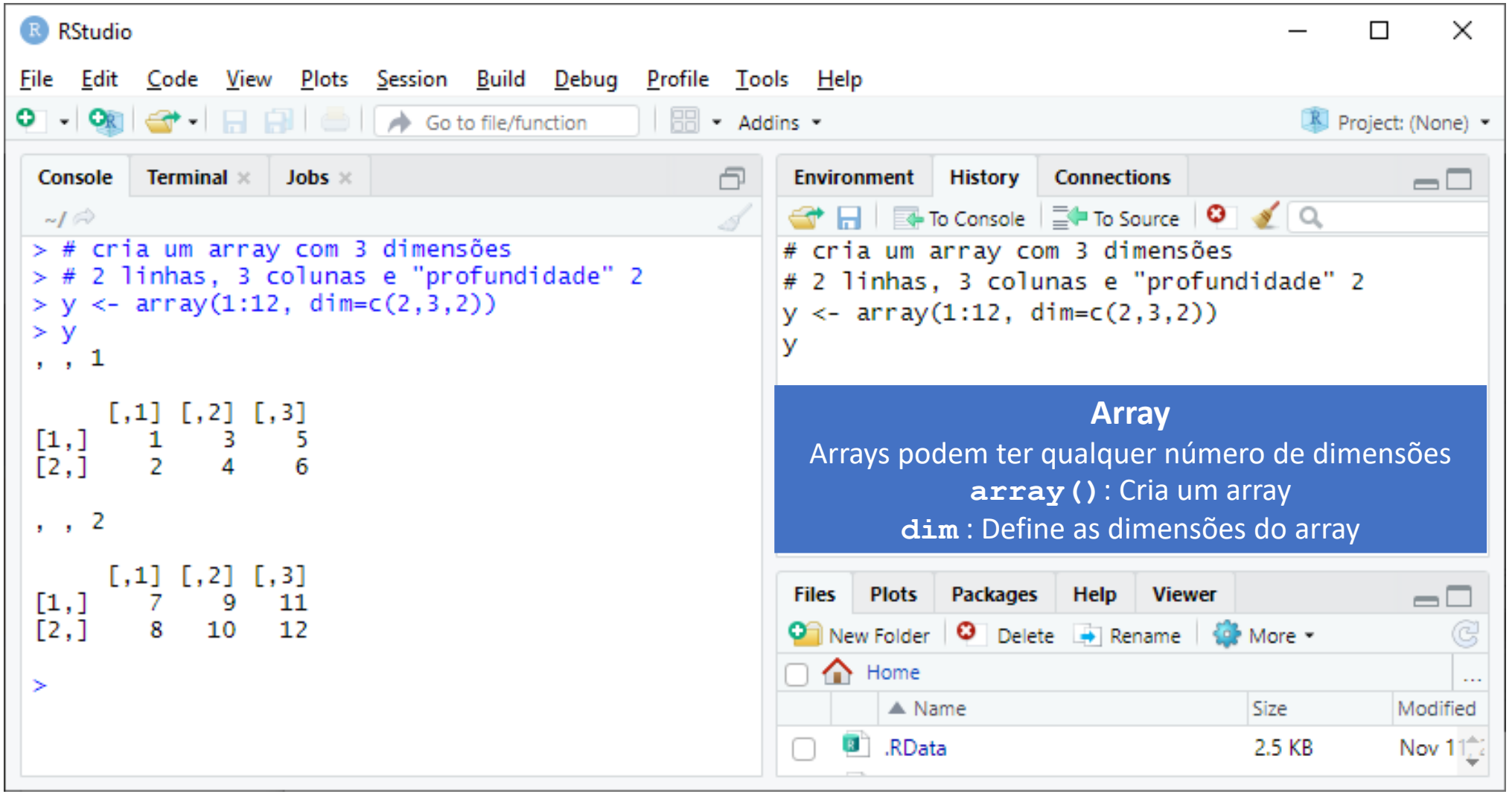


Array

Arrays podem ter qualquer número de dimensões

`array()` : Cria um array

`dim` : Define as dimensões do array



Fatores

Variáveis categóricas e contínuas

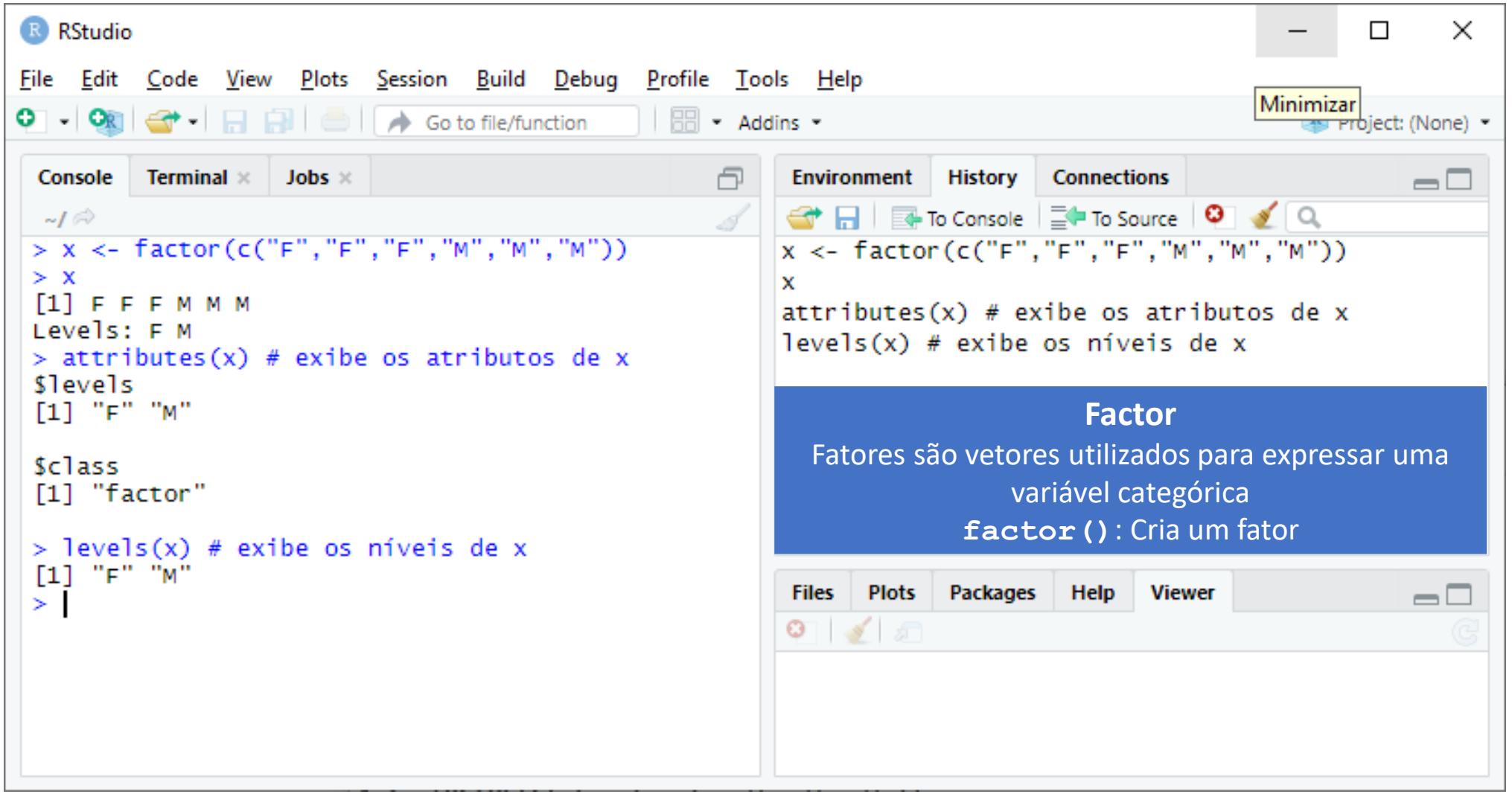
Em um conjunto de dados, podemos distinguir dois tipos de variáveis: categóricas e contínuas.

Variáveis categóricas apresentam um número limitado de valores diferentes geralmente baseados em um determinado grupo finito. Alguns exemplos de variáveis categóricas em R: gênero, ocupação, país, ano, etc.

Variáveis contínuas, por outro lado, podem assumir qualquer valor, de inteiro a decimal. Por exemplo, podemos ter a receita de uma empresa, o preço de uma ação, etc.

Variáveis categóricas no R são armazenadas em um fator.

Os fatores armazenam dados de string e inteiros como "levels" ou níveis.

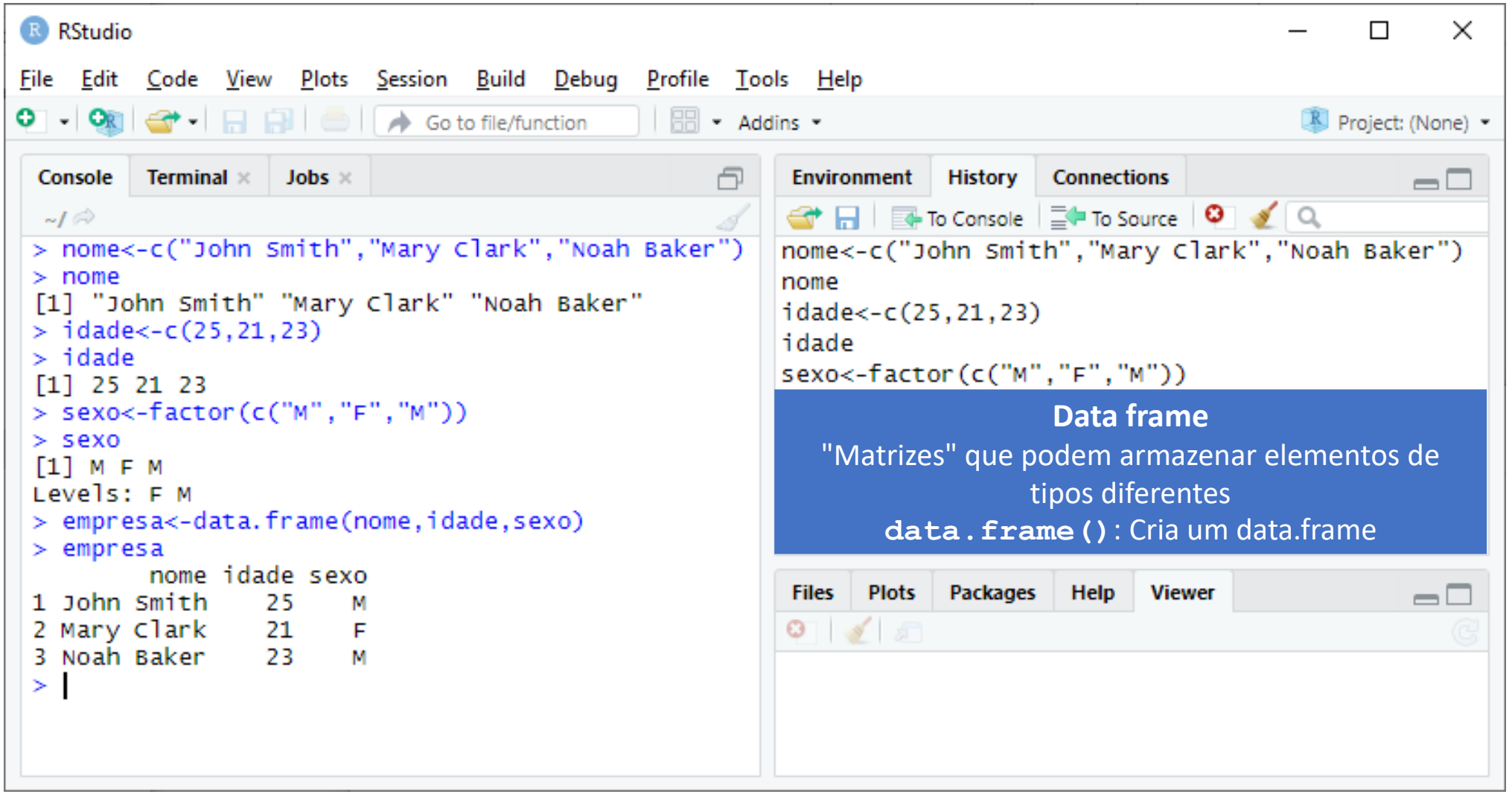


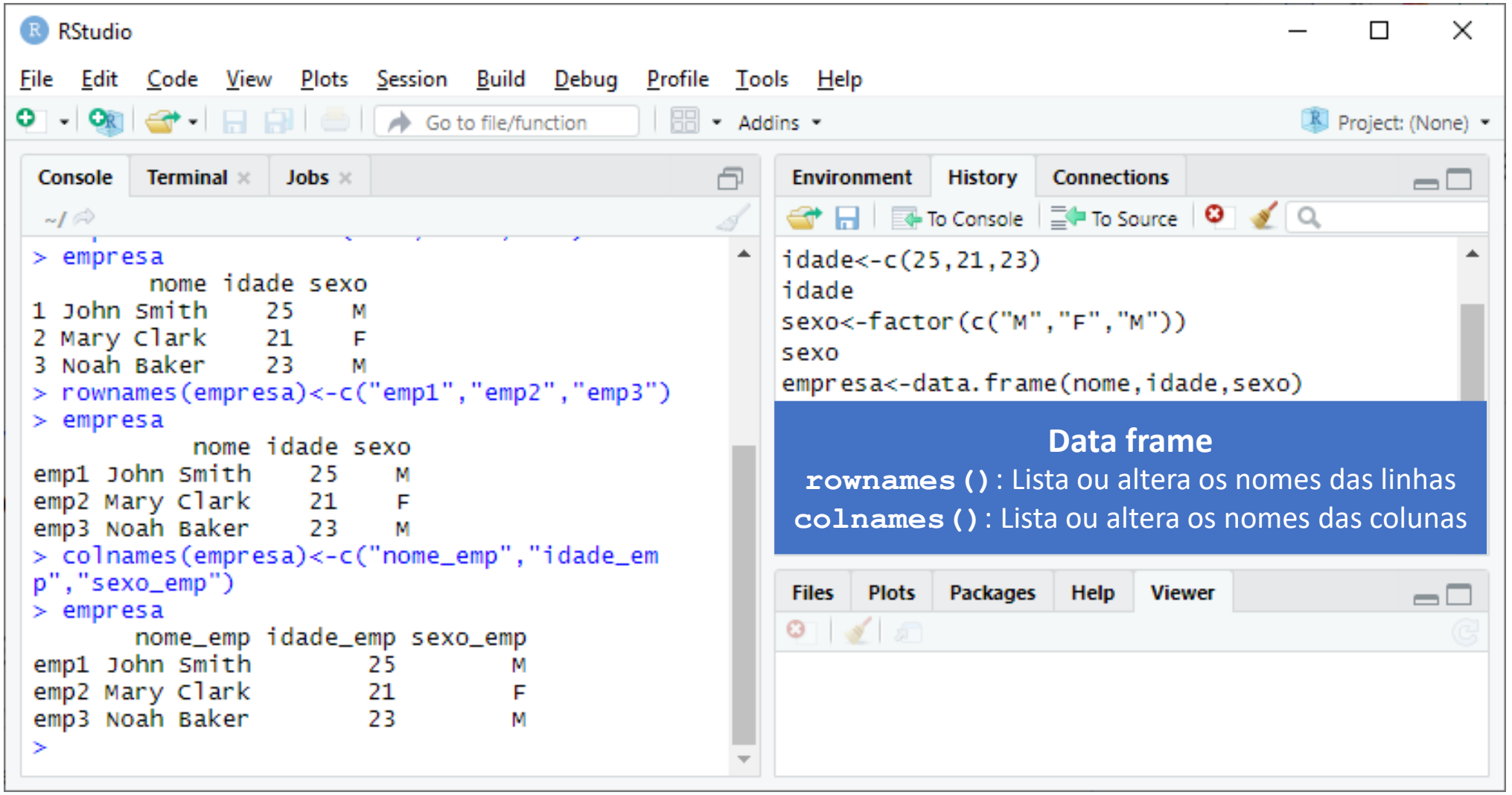
Factor

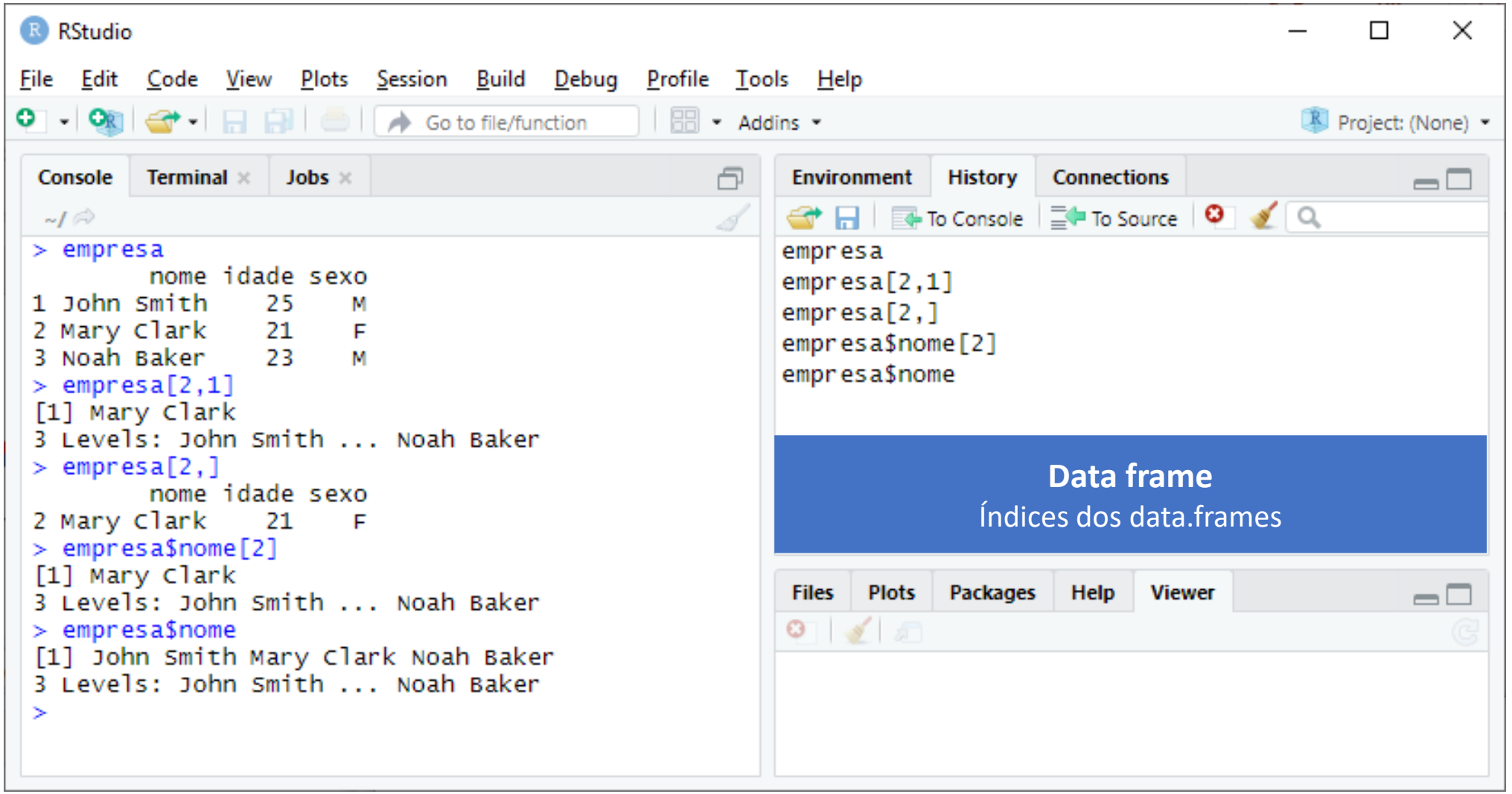
Fatores são vetores utilizados para expressar uma variável categórica

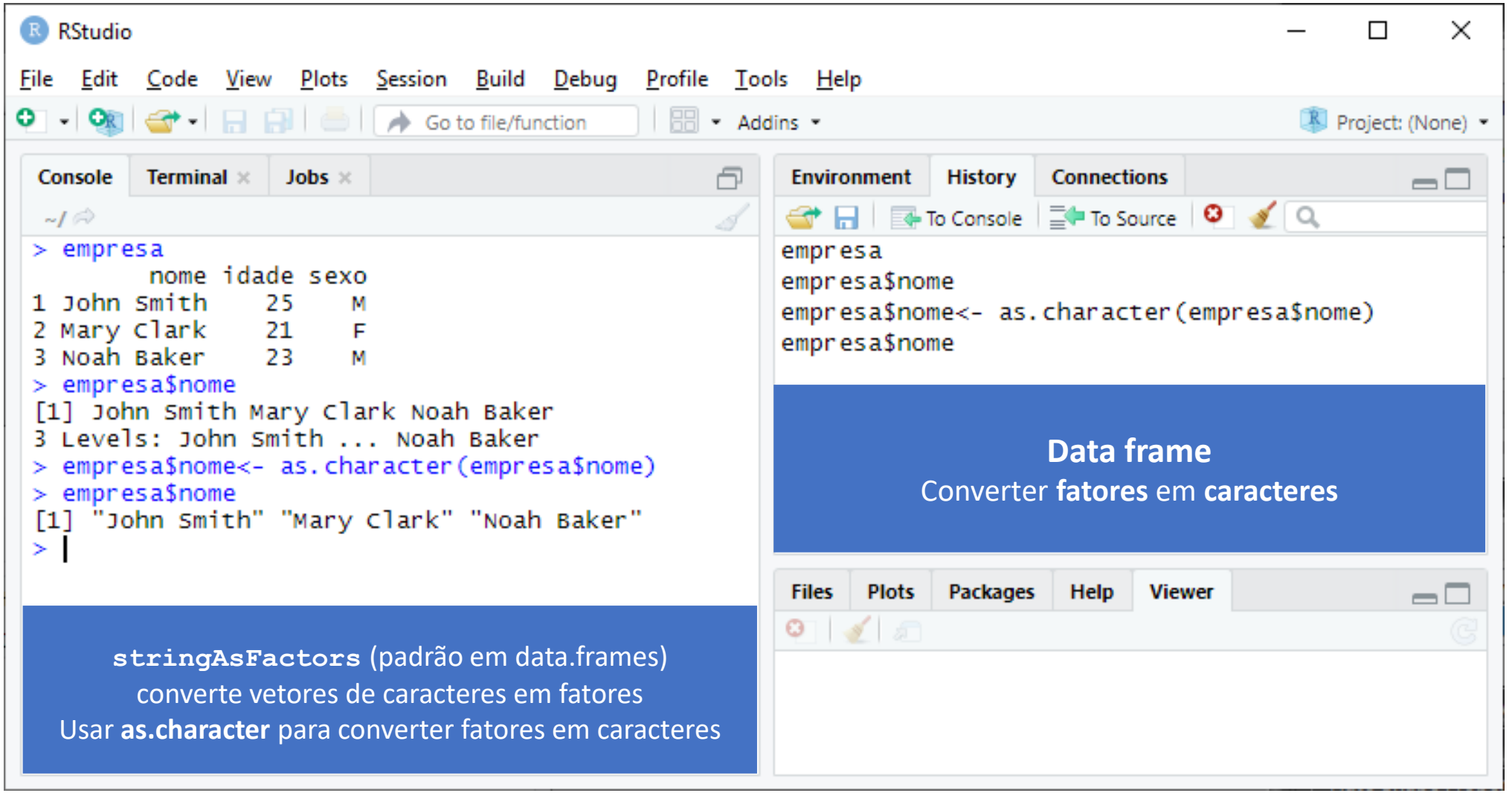
`factor()` : Cria um fator

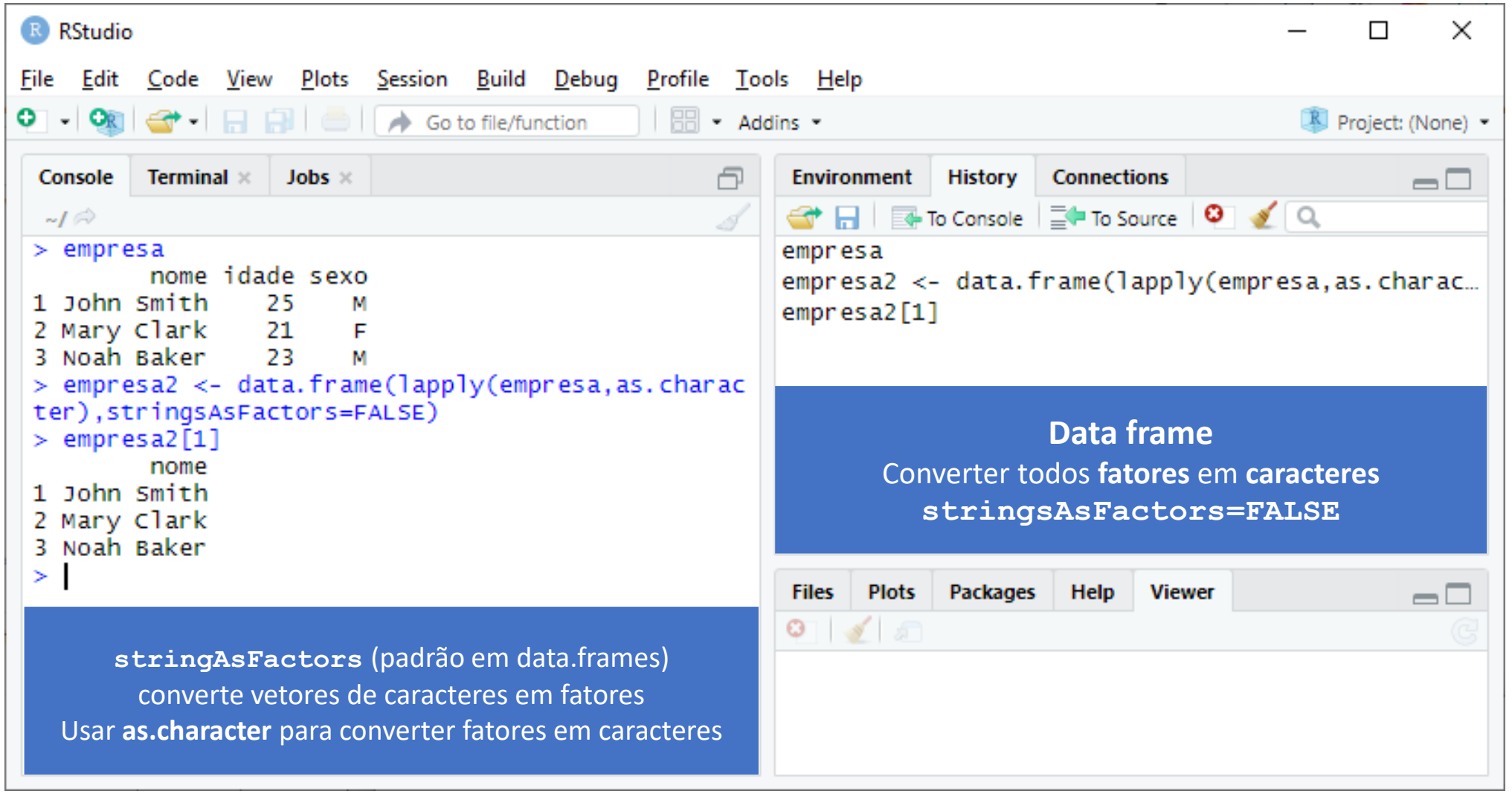
Data Frames

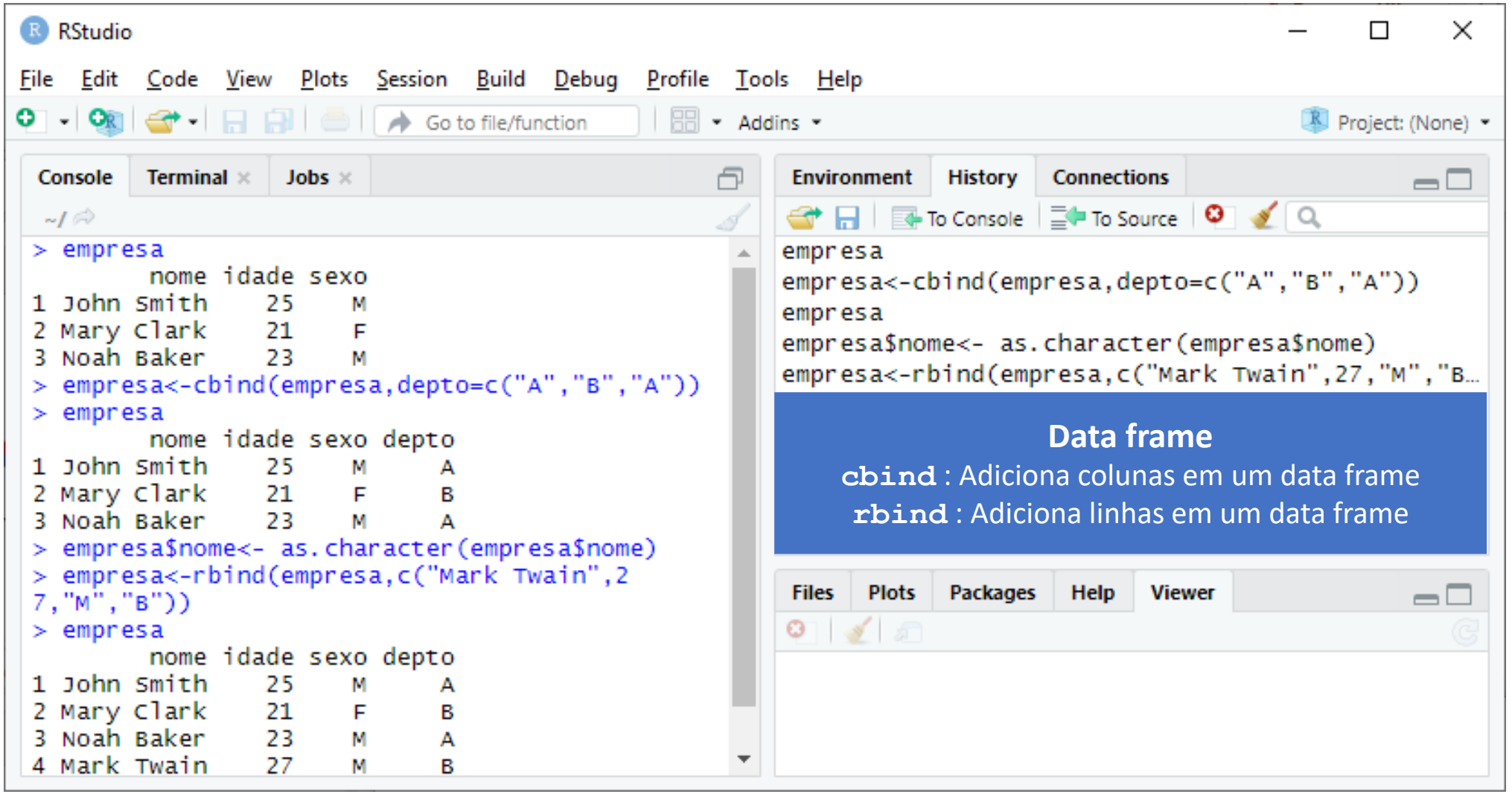


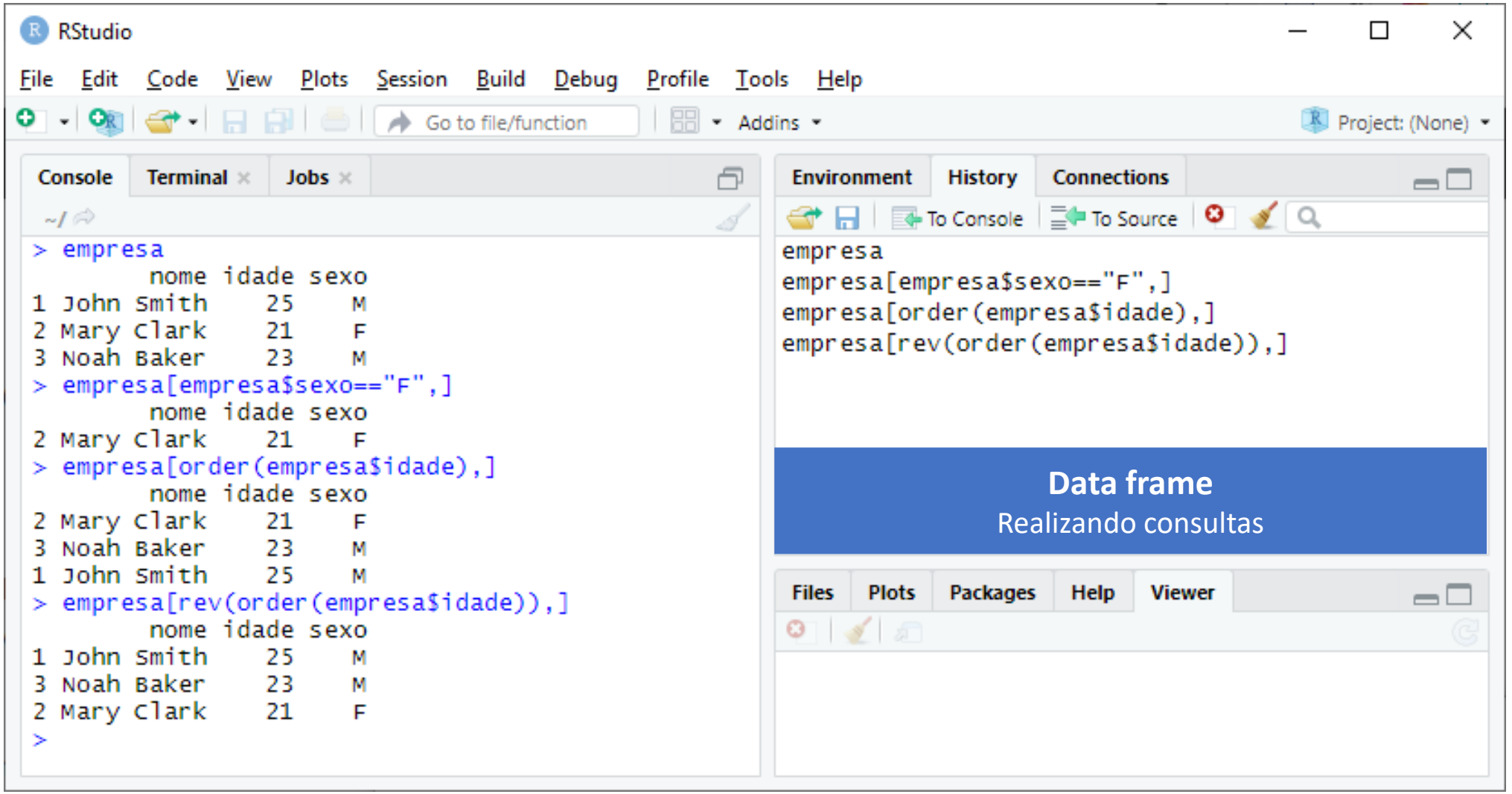


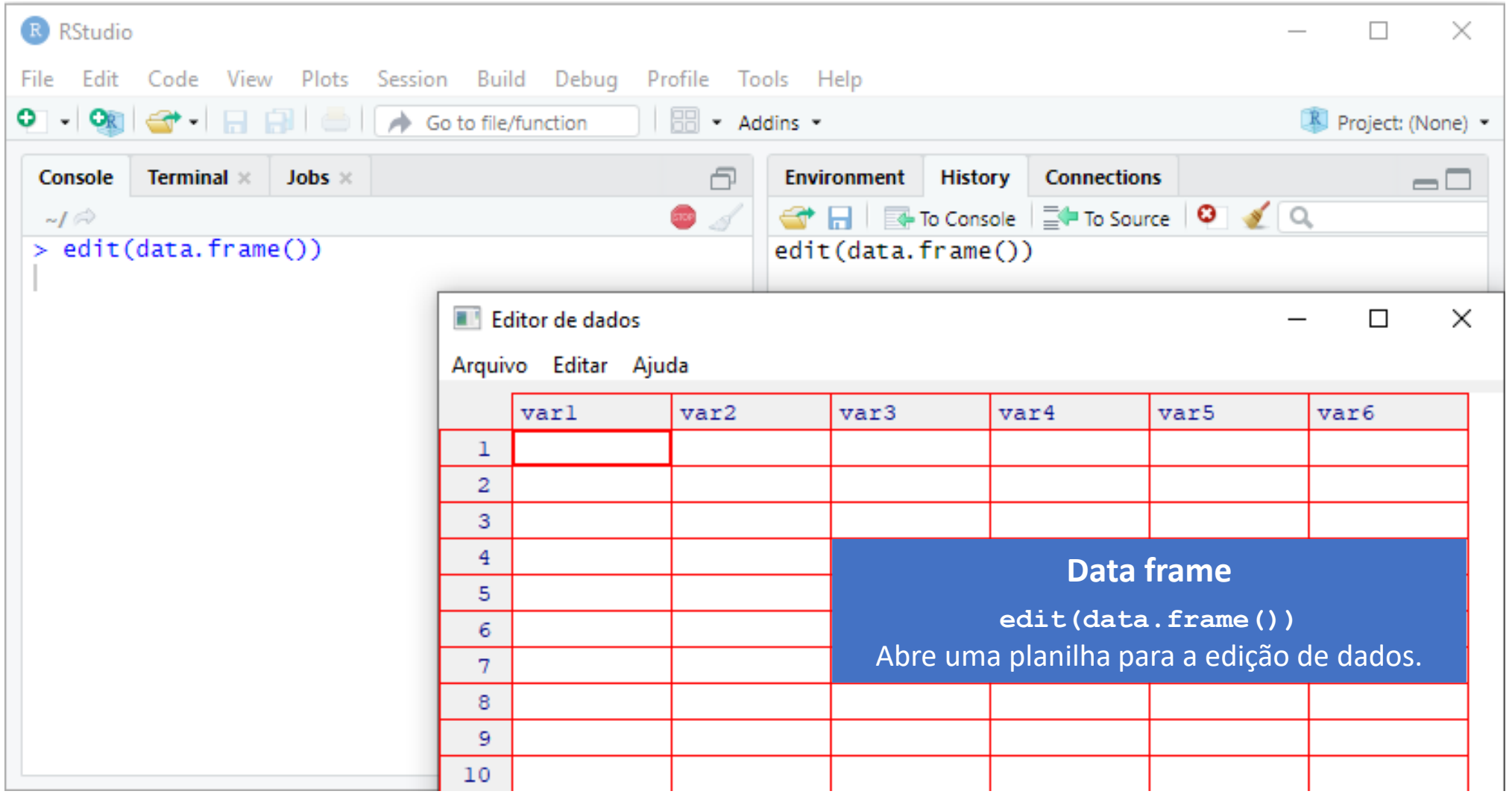




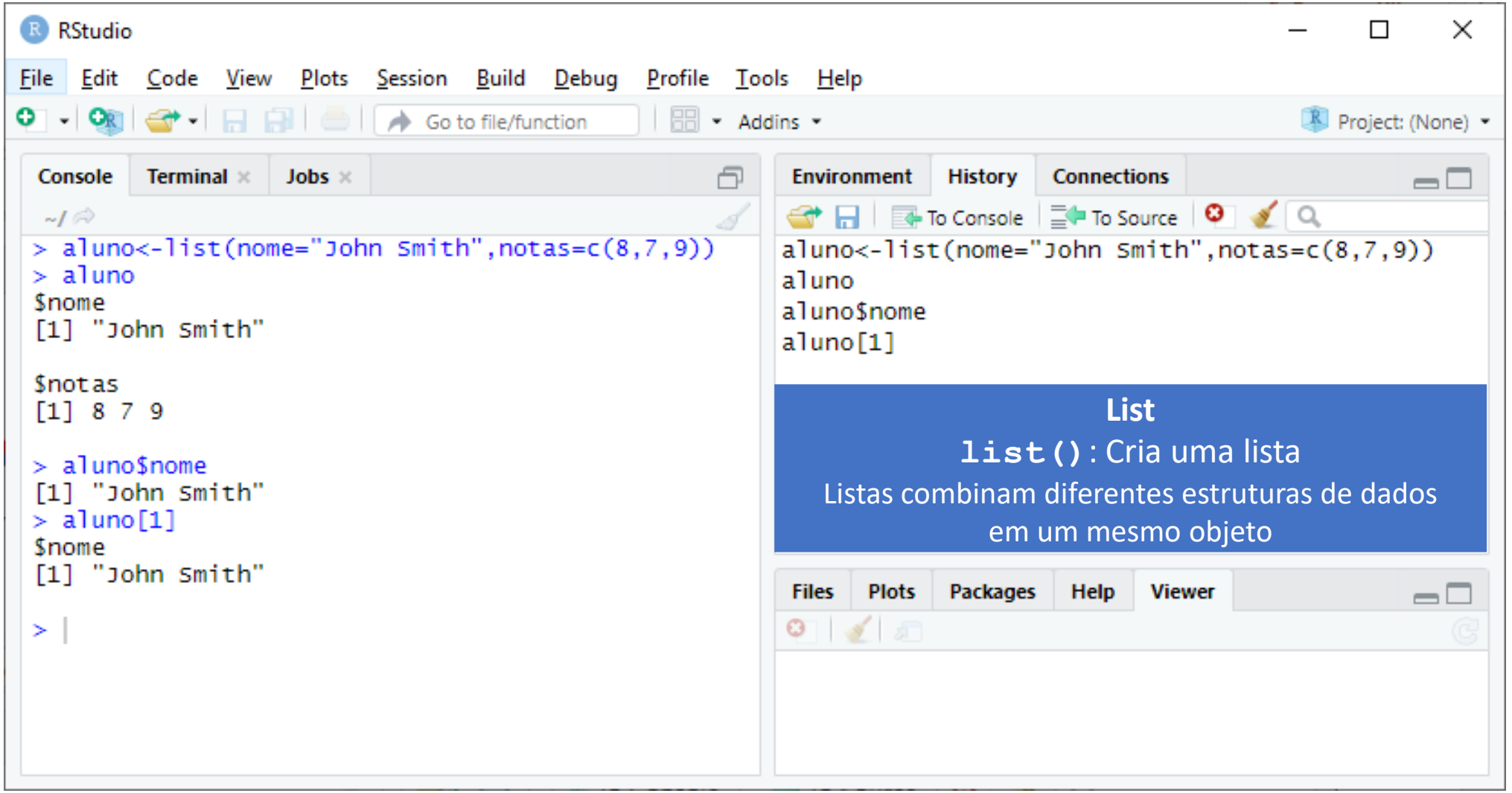

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Console
Terminal
Jobs
> empresa
nome idade sexo
1 John Smith 25 M
2 Mary Clark 21 F
3 Noah Baker 23 M
> empresa2 <- data.frame(lapply(empresa,as.charac
ter),stringsAsFactors=FALSE)
> empresa2[1]
nome
1 John Smith
2 Mary Clark
3 Noah Baker
>
stringAsFactors (padrão em data.frames)
converte vetores de caracteres em fatores
Usar as.character para converter fatores em caracteres
Environment
History
Connections
To Console
To Source
empresa
empresa2 <- data.frame(lapply(empresa,as.charac…
empresa2[1]
Data frame
Converter todos fatores em caracteres
stringsAsFactors=FALSE
Files
Plots
Packages
Help
Viewer







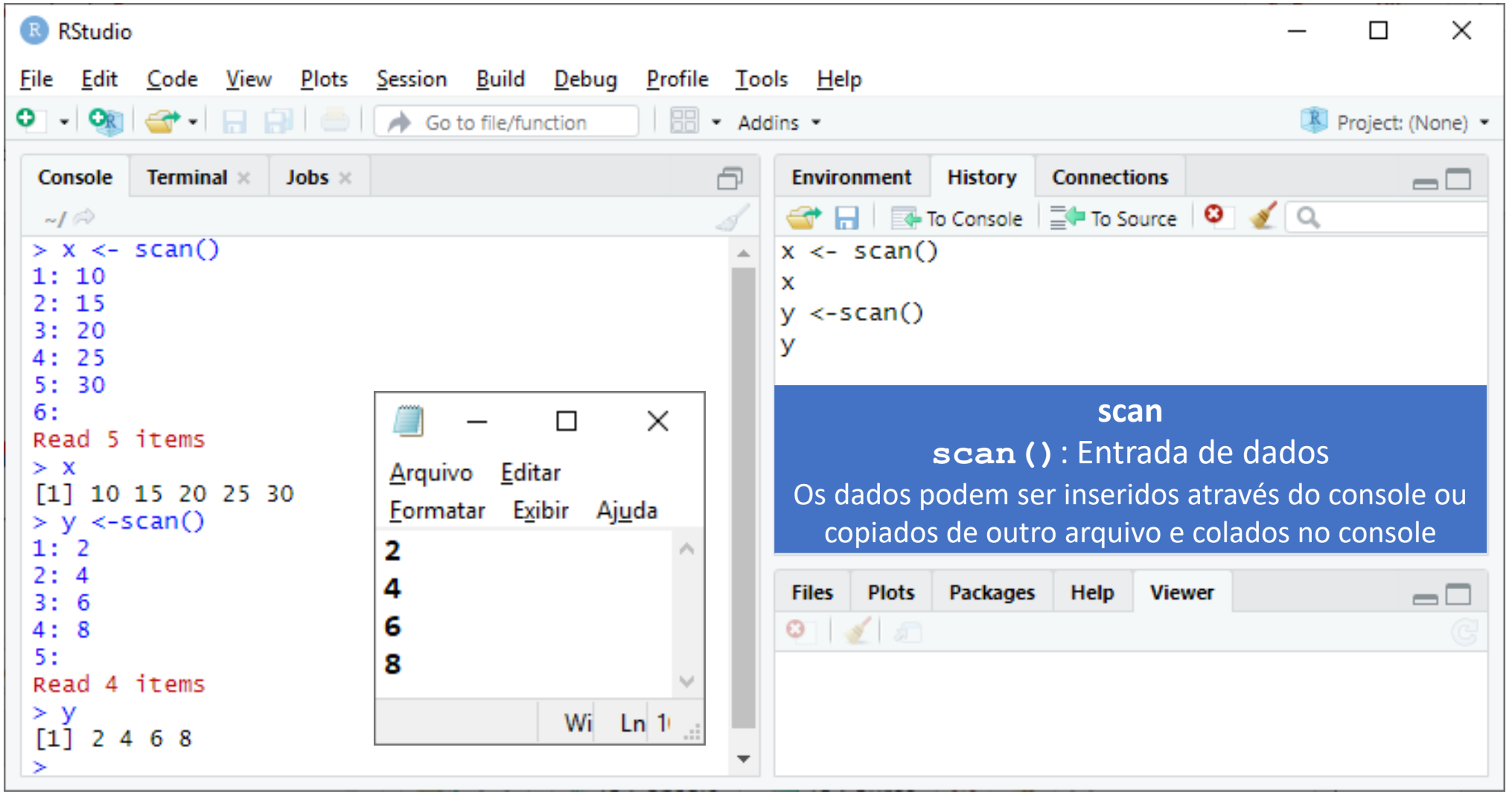
Listas



Entrada de Dados

Entrada de Dados

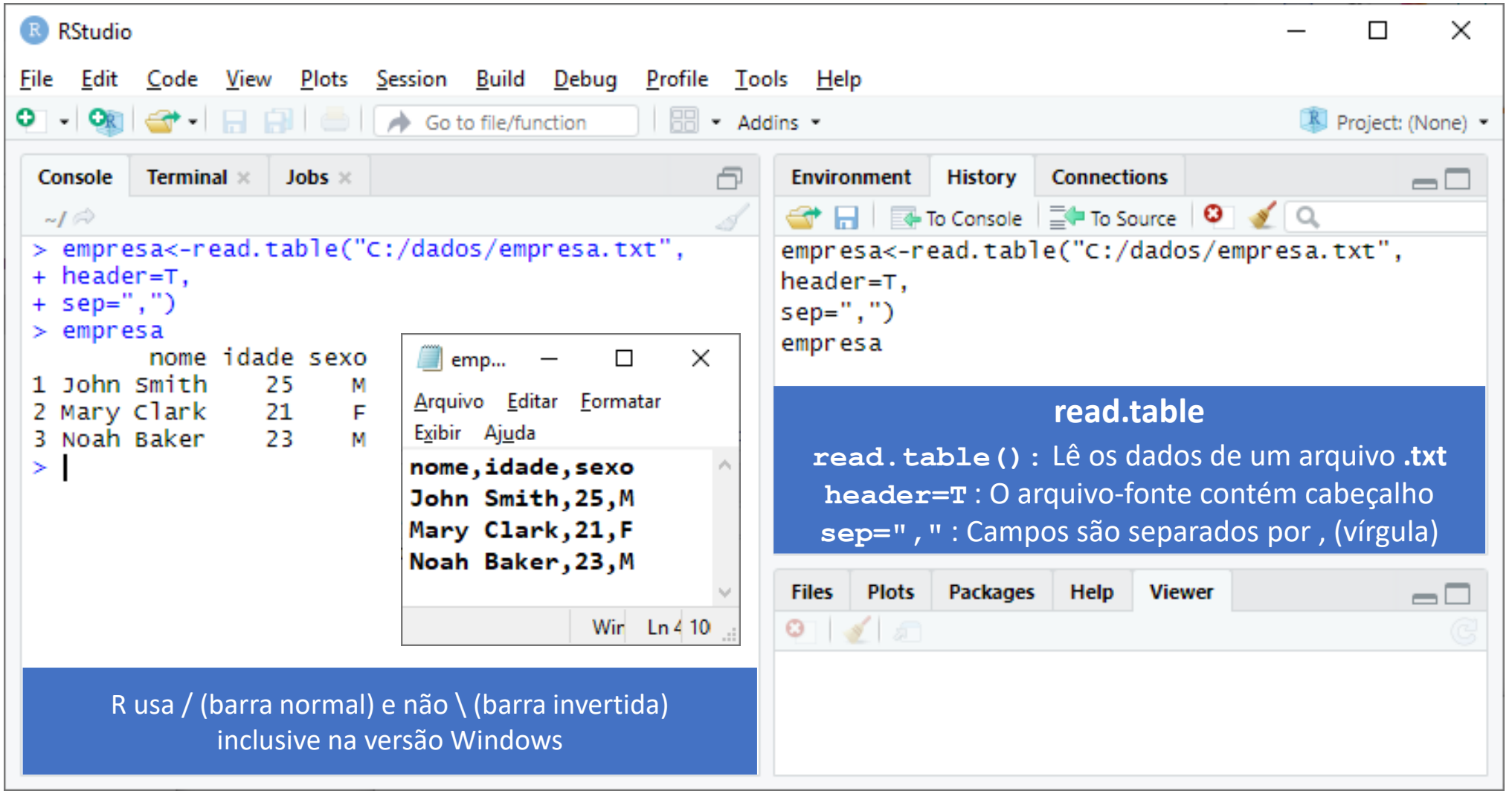
Método	Descrição
`scan()`	Entrada através do console ou de arquivo texto (.txt)
`read.table()`	Entrada através de arquivo texto com separadores (.csv)
`read.xlsx()`	Entrada através de arquivo MS Excel (.xlsx)
`data()`	Entrada através de datasets

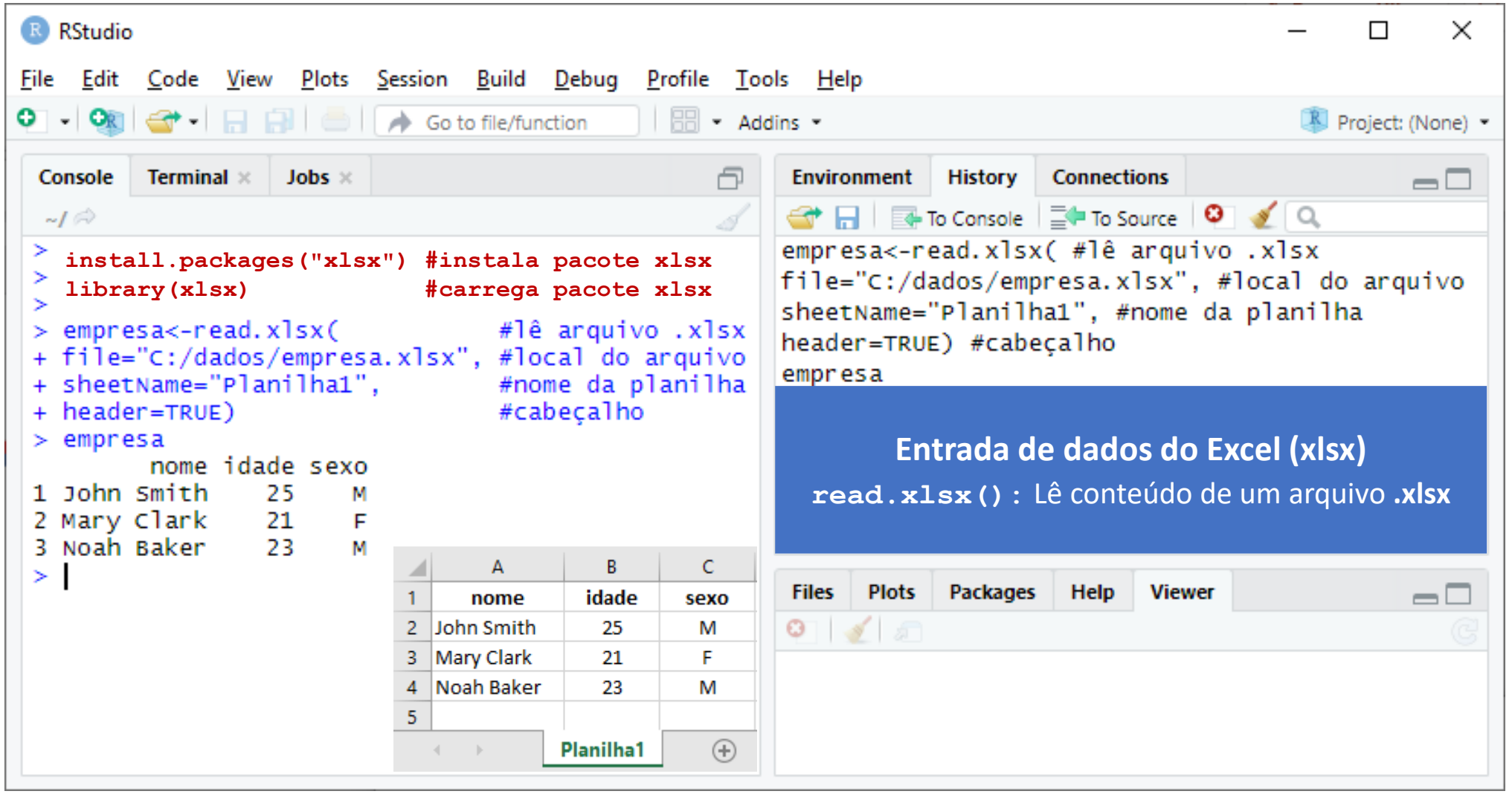


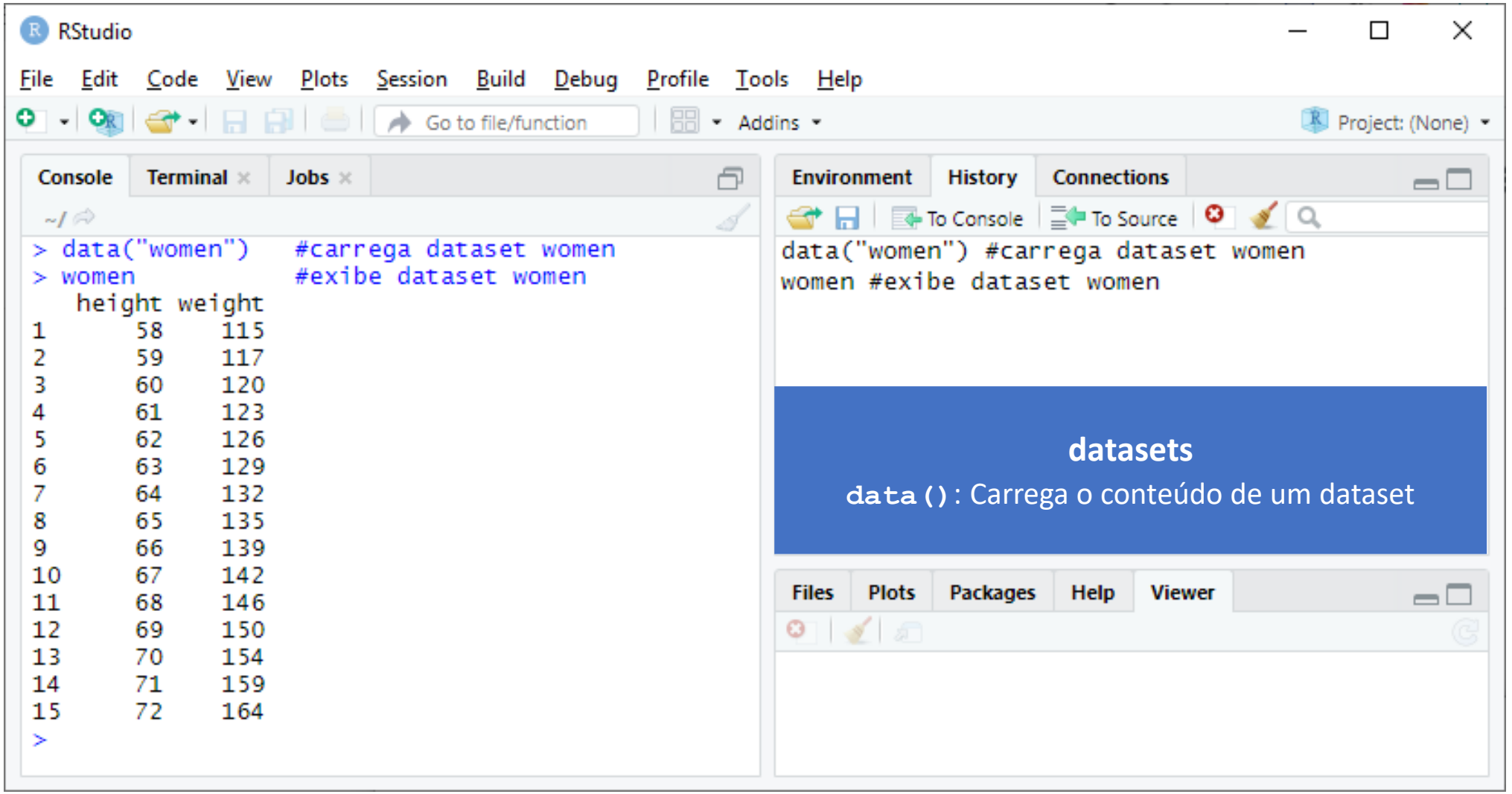
scan

`scan()` : Entrada de dados

Os dados podem ser inseridos através do console ou copiados de outro arquivo e colados no console


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Console
Terminal
Jobs
> empresa<-read.table("C:/dados/empresa.txt",
+ header=T,
+ sep=",")
> empresa
nome idade sexo
1 John Smith 25 M
2 Mary Clark 21 F
3 Noah Baker 23 M
>
emp...
Arquivo Editar Formatar
Exibir Ajuda
nome,idade,sexo
John Smith,25,M
Mary Clark,21,F
Noah Baker,23,M
Wir Ln 4 10
Environment
History
Connections
To Console
To Source
empresa<-read.table("C:/dados/empresa.txt",
header=T,
sep=",")
empresa
read.table
read.table() : Lê os dados de um arquivo .txt
header=T : O arquivo-fonte contém cabeçalho
sep=" , " : Campos são separados por , (vírgula)
Files
Plots
Packages
Help
Viewer
R usa / (barra normal) e não \ (barra invertida)
inclusive na versão Windows

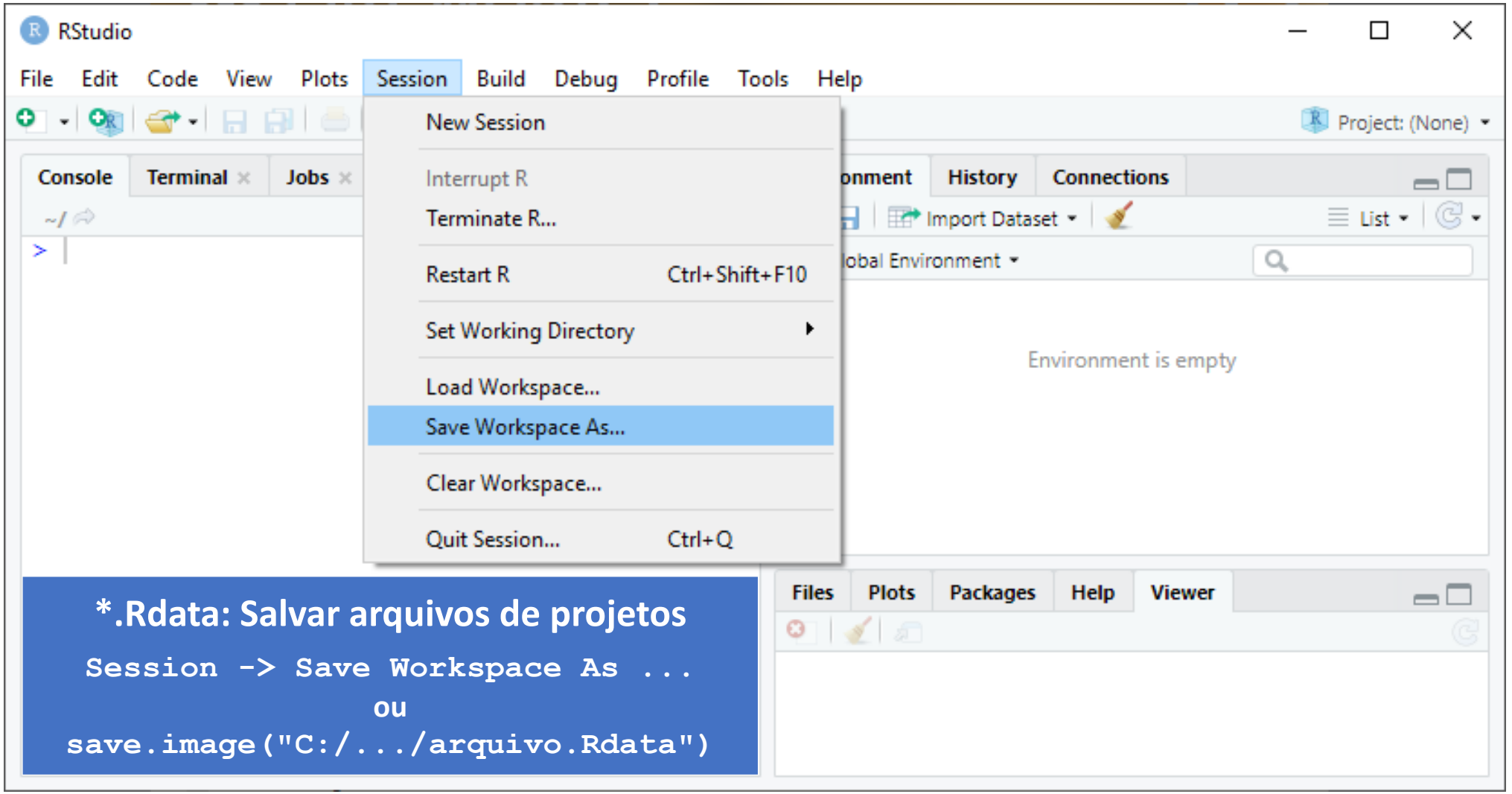


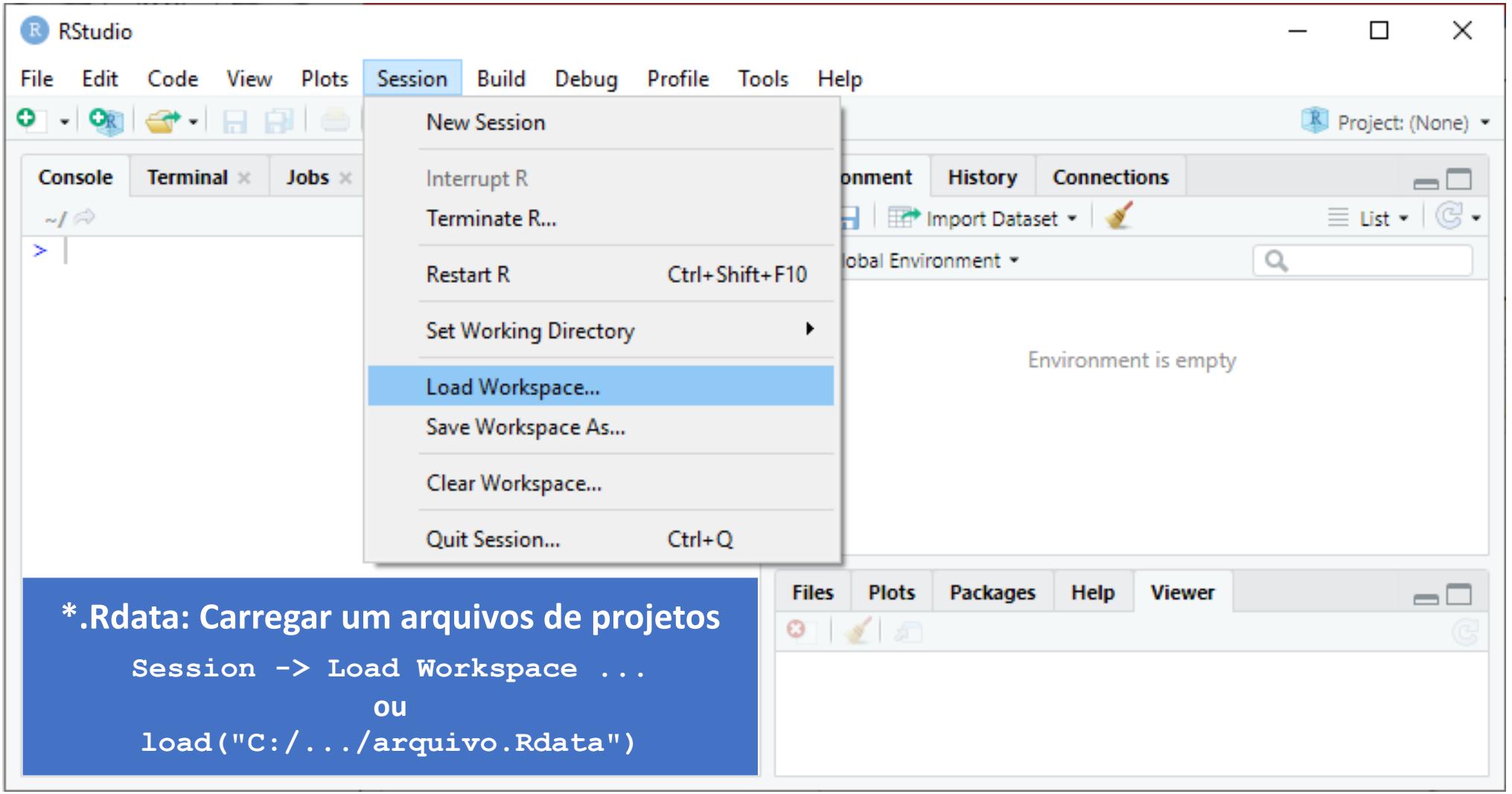


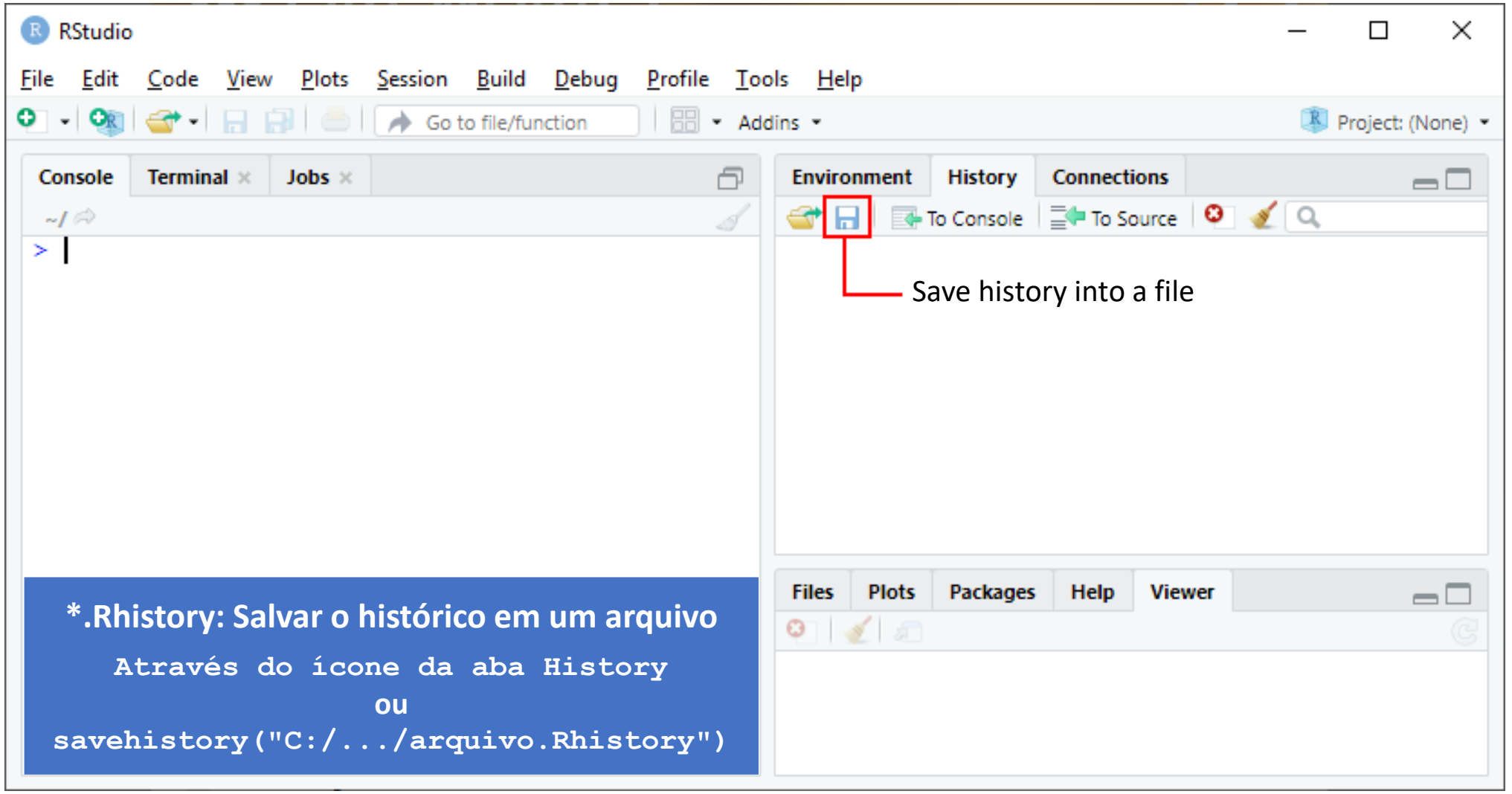
Tipos de arquivos do R

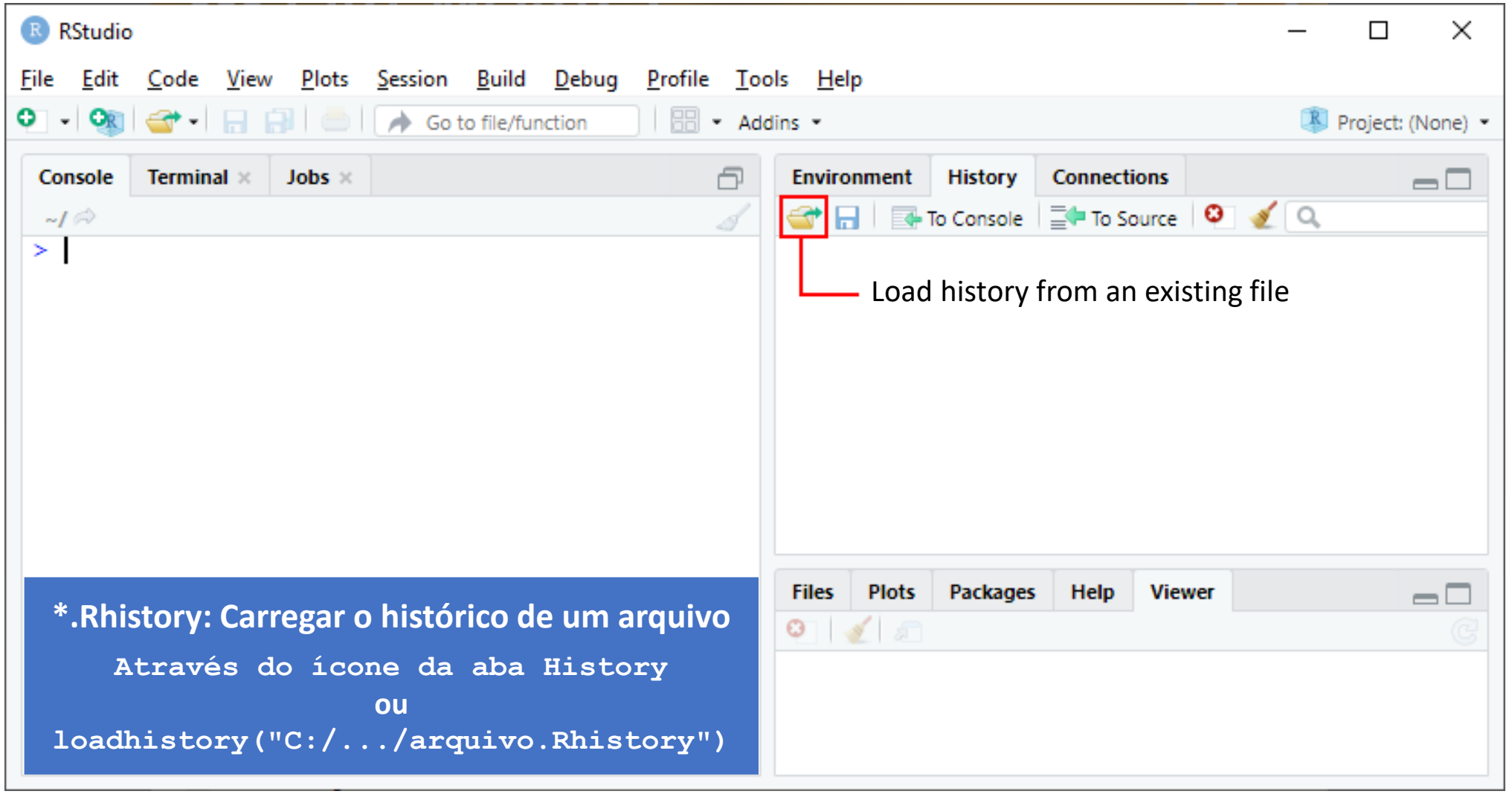
Tipos de arquivos do R

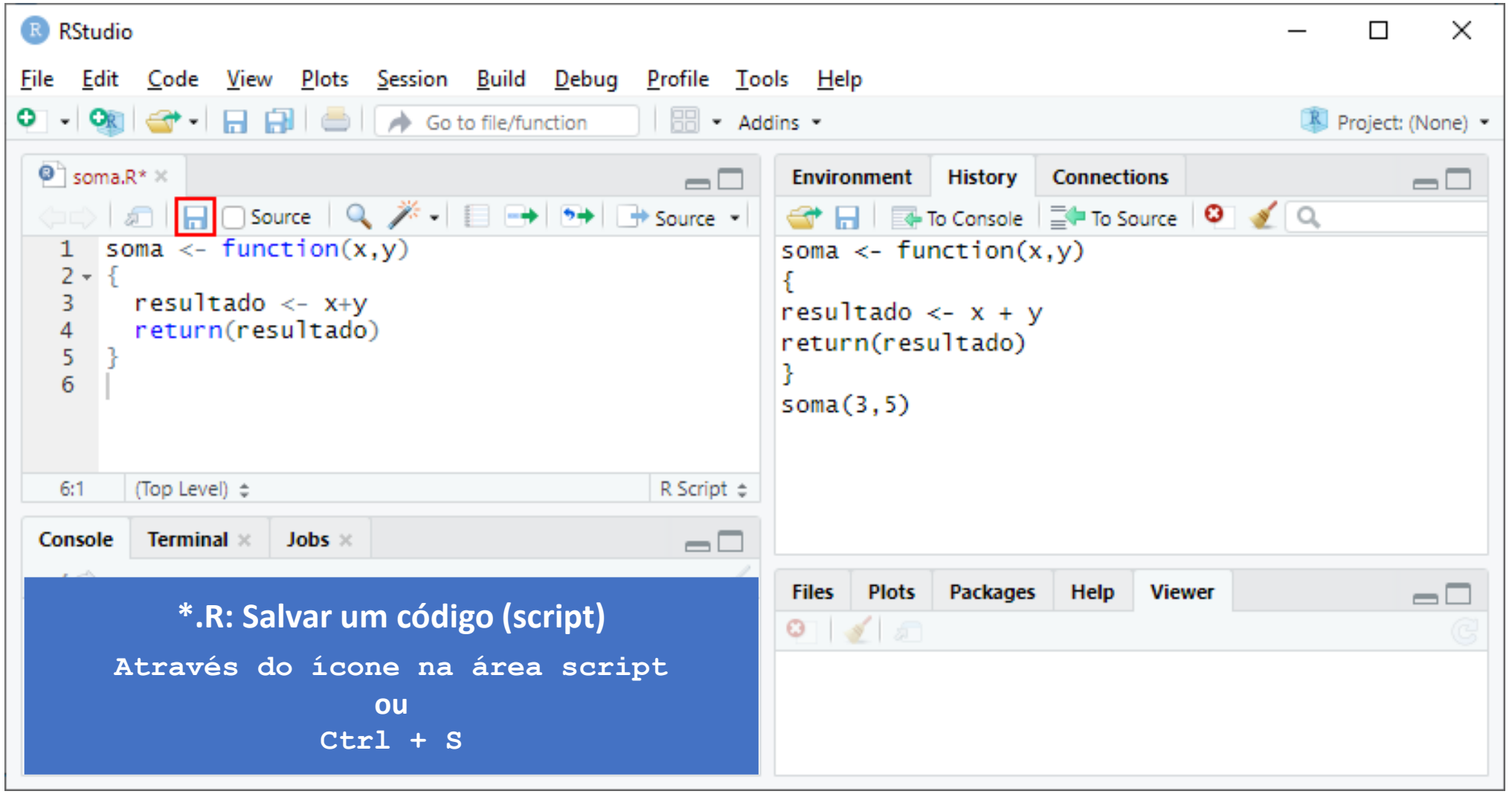
Tipo de Arquivo	Descrição
`*.RData`	Arquivos de projetos. Salvam os objetos criados em uma sessão.
`*.Rhistory`	Contêm o histórico de todos os comandos digitados em uma sessão.
`*.R`	Códigos criados pelos usuários (funções, rotinas de análise, etc.)
`*.Rmd`	Arquivos Markdown usados para criação de conteúdo

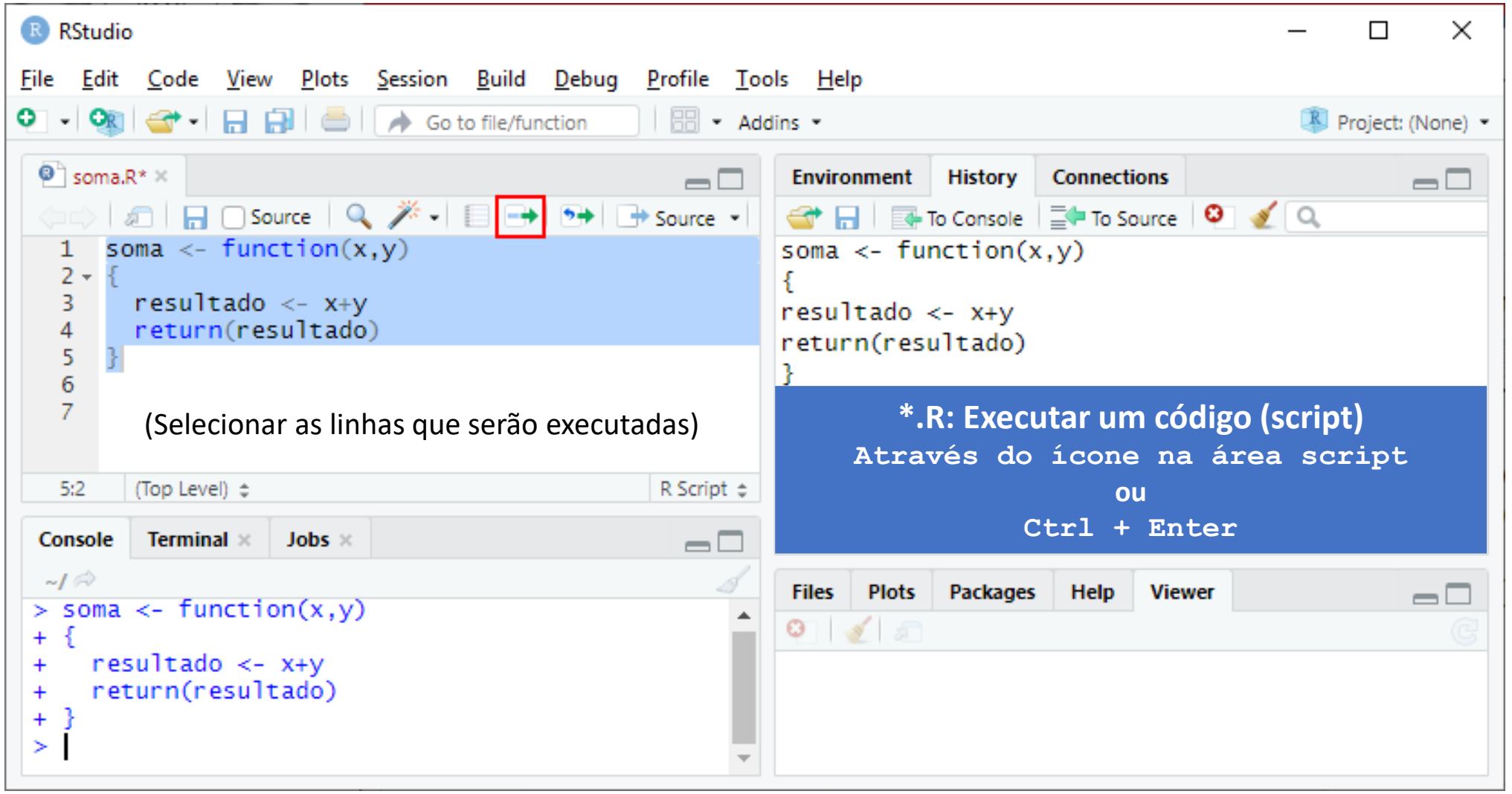


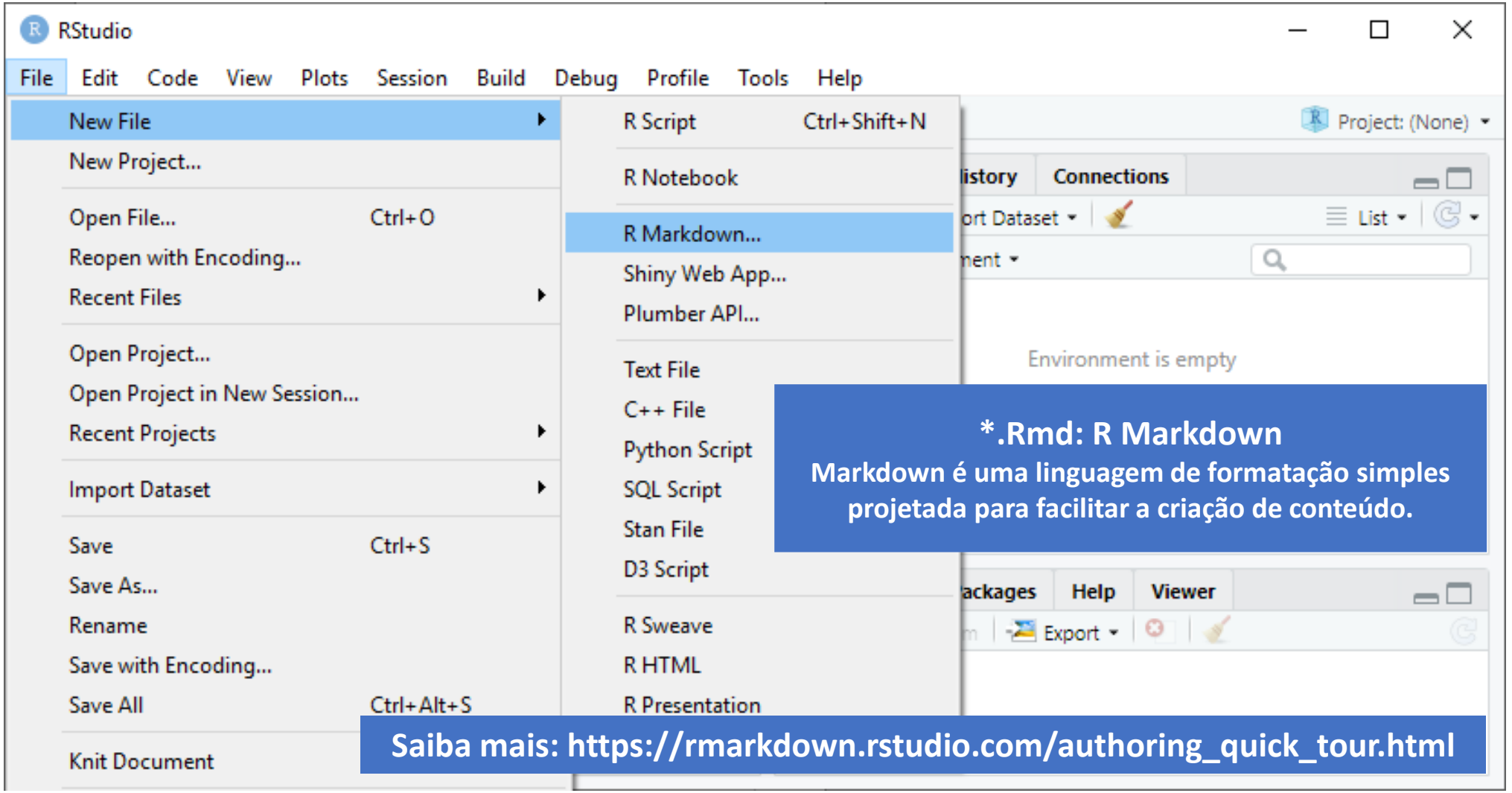


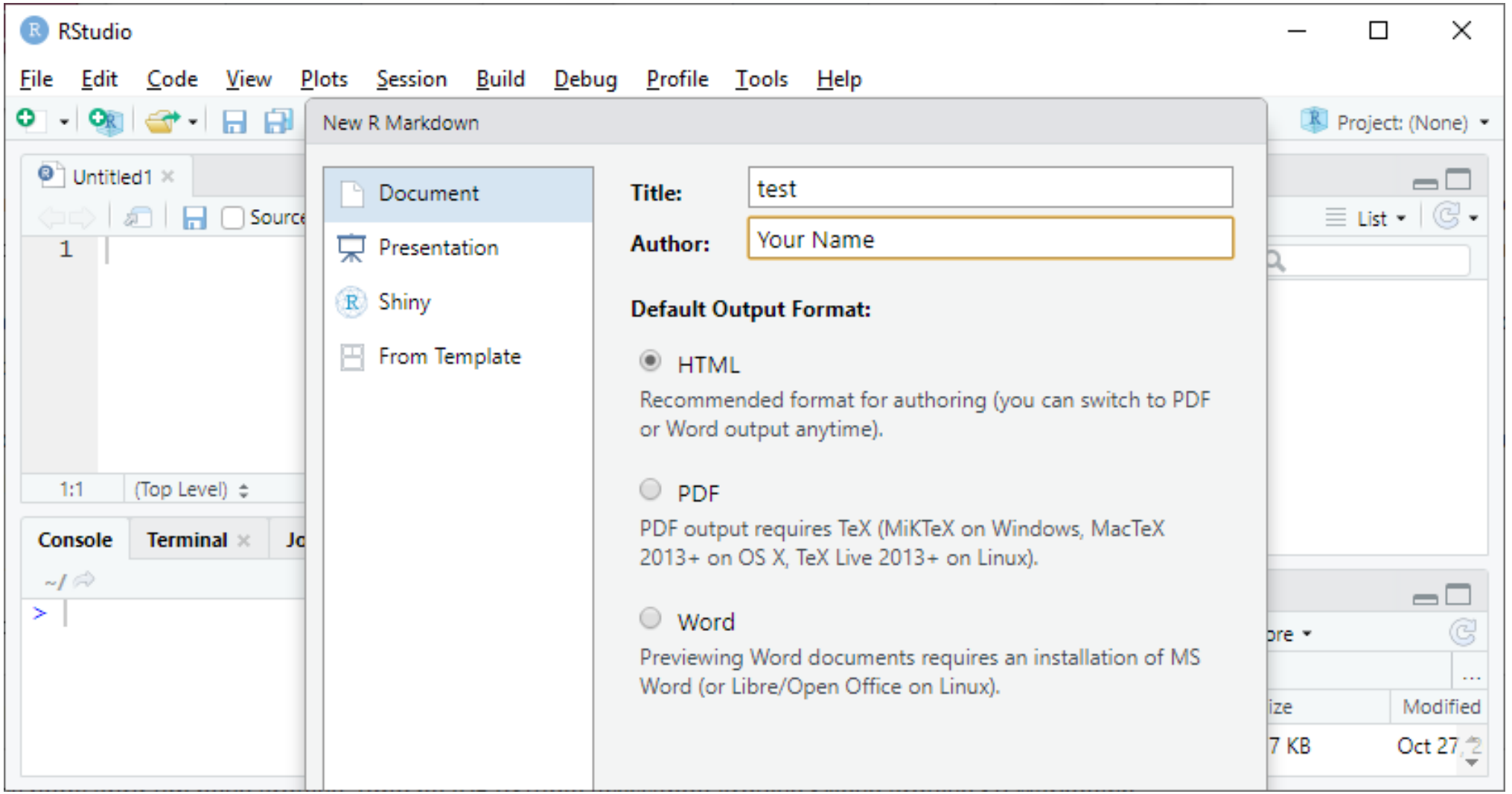


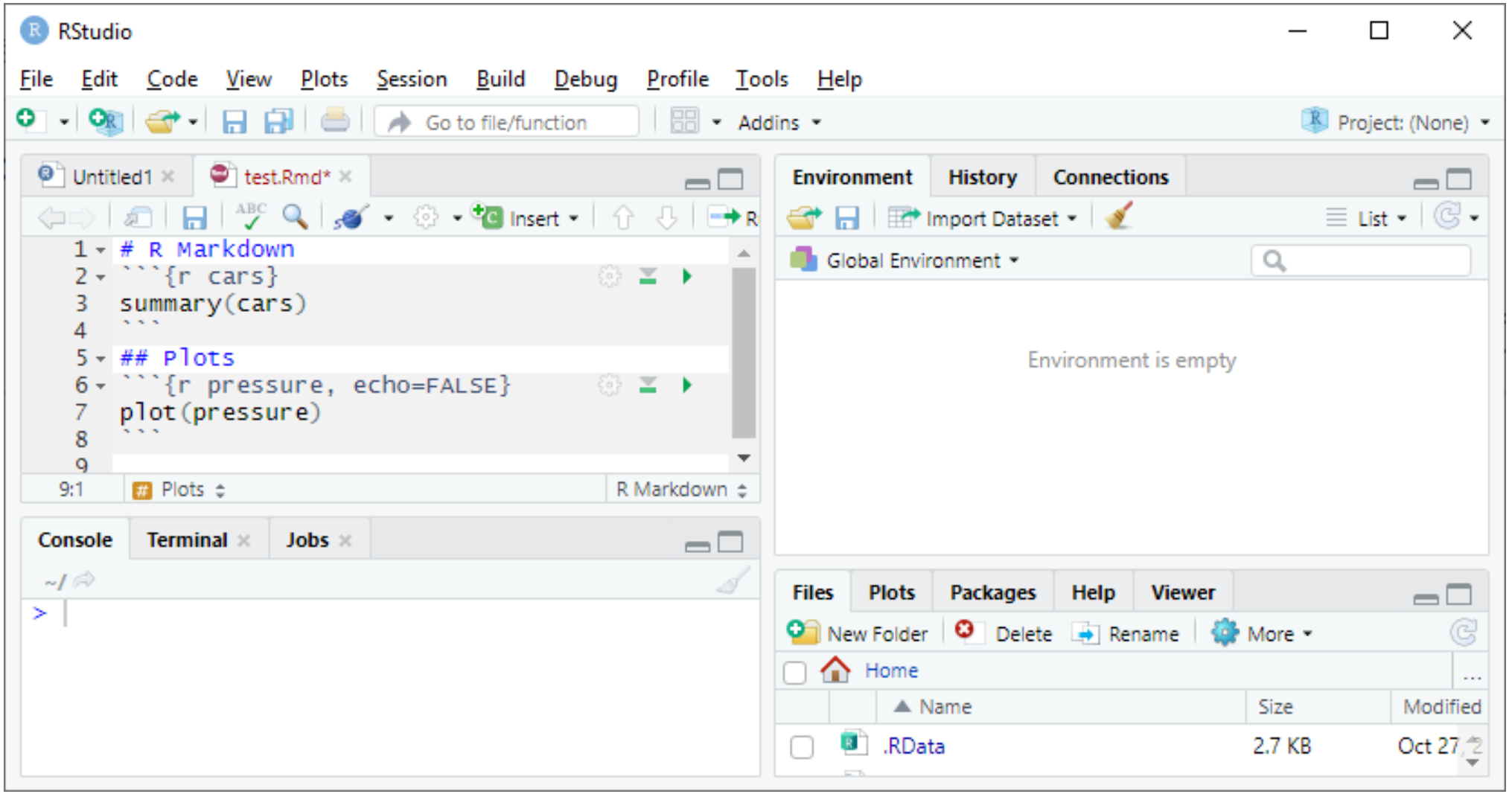


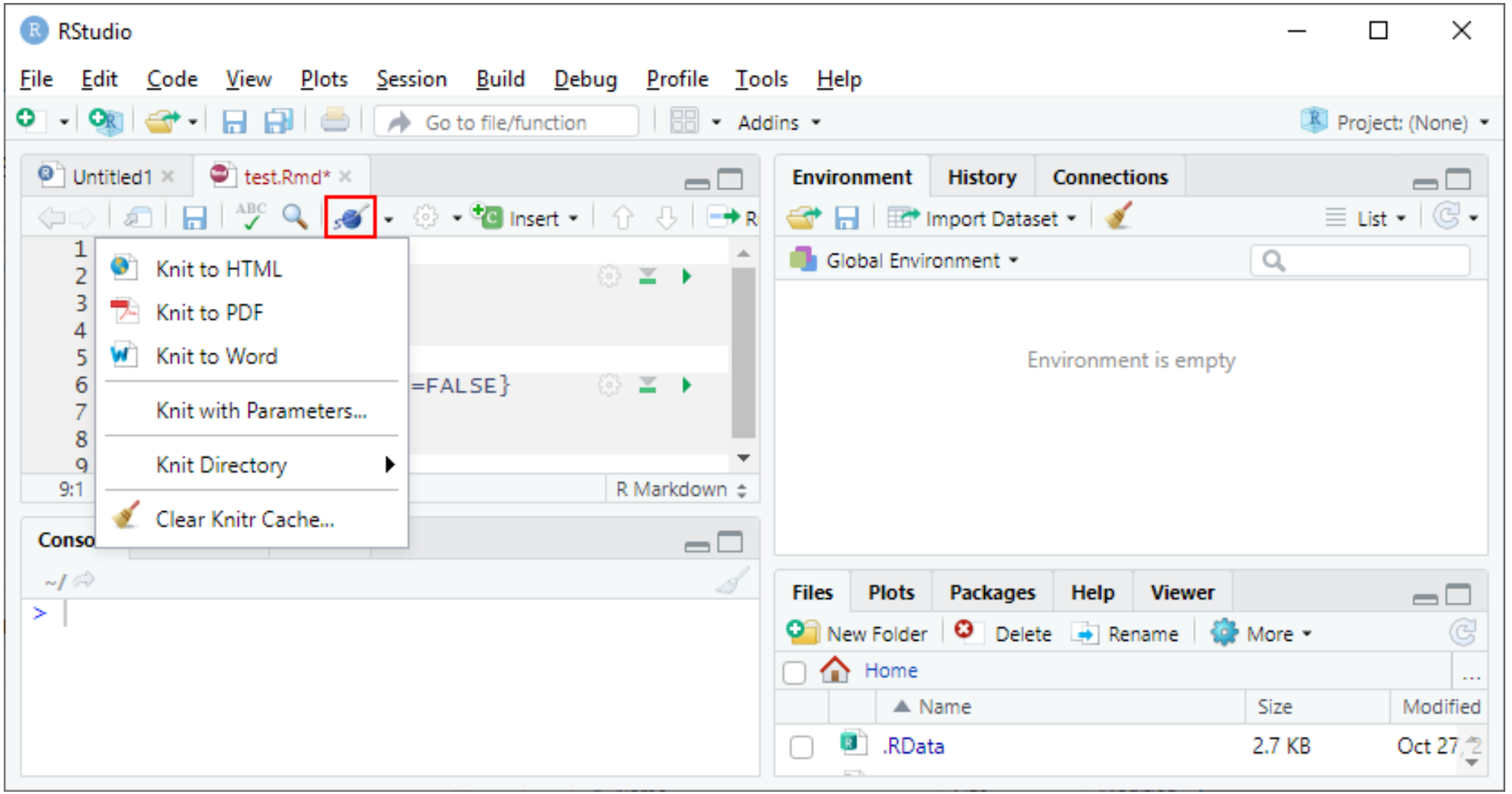


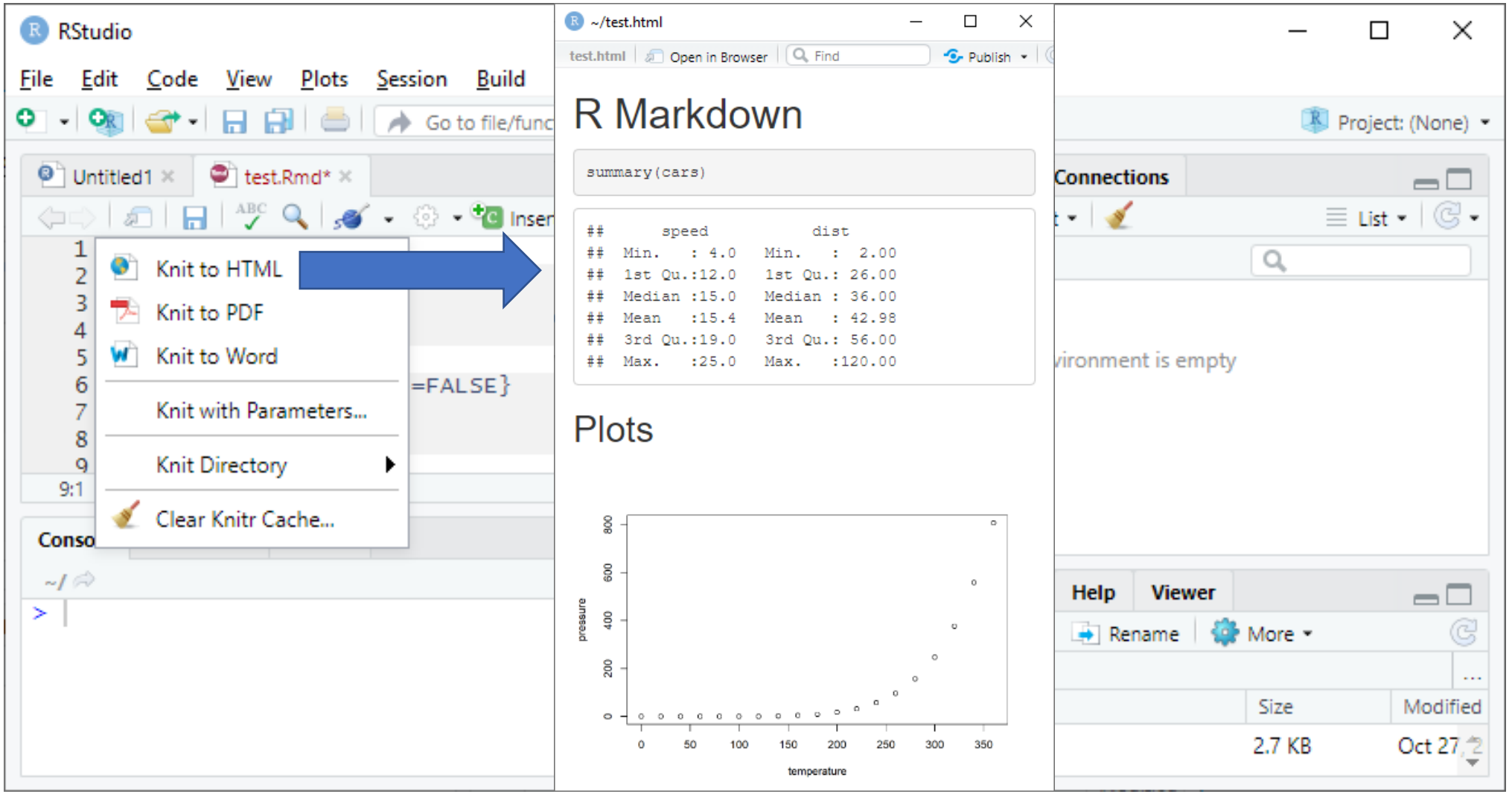



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
New File
New Project...
Open File... Ctrl+O
Reopen with Encoding...
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save Ctrl+S
Save As...
Rename
Save with Encoding...
Save All Ctrl+Alt+S
Knit Document
R Script Ctrl+Shift+N
R Notebook
R Markdown...
Shiny Web App...
Plumber API...
Text File
C++ File
Python Script
SQL Script
Stan File
D3 Script
R Sweave
R HTML
R Presentation
Project: (None)
Connections
List
Environment is empty
Help
Viewer
Export
*.Rmd: R Markdown
Markdown é uma linguagem de formatação simples projetada para facilitar a criação de conteúdo.
Saiba mais: https://rmarkdown.rstudio.com/authoring_quick_tour.html



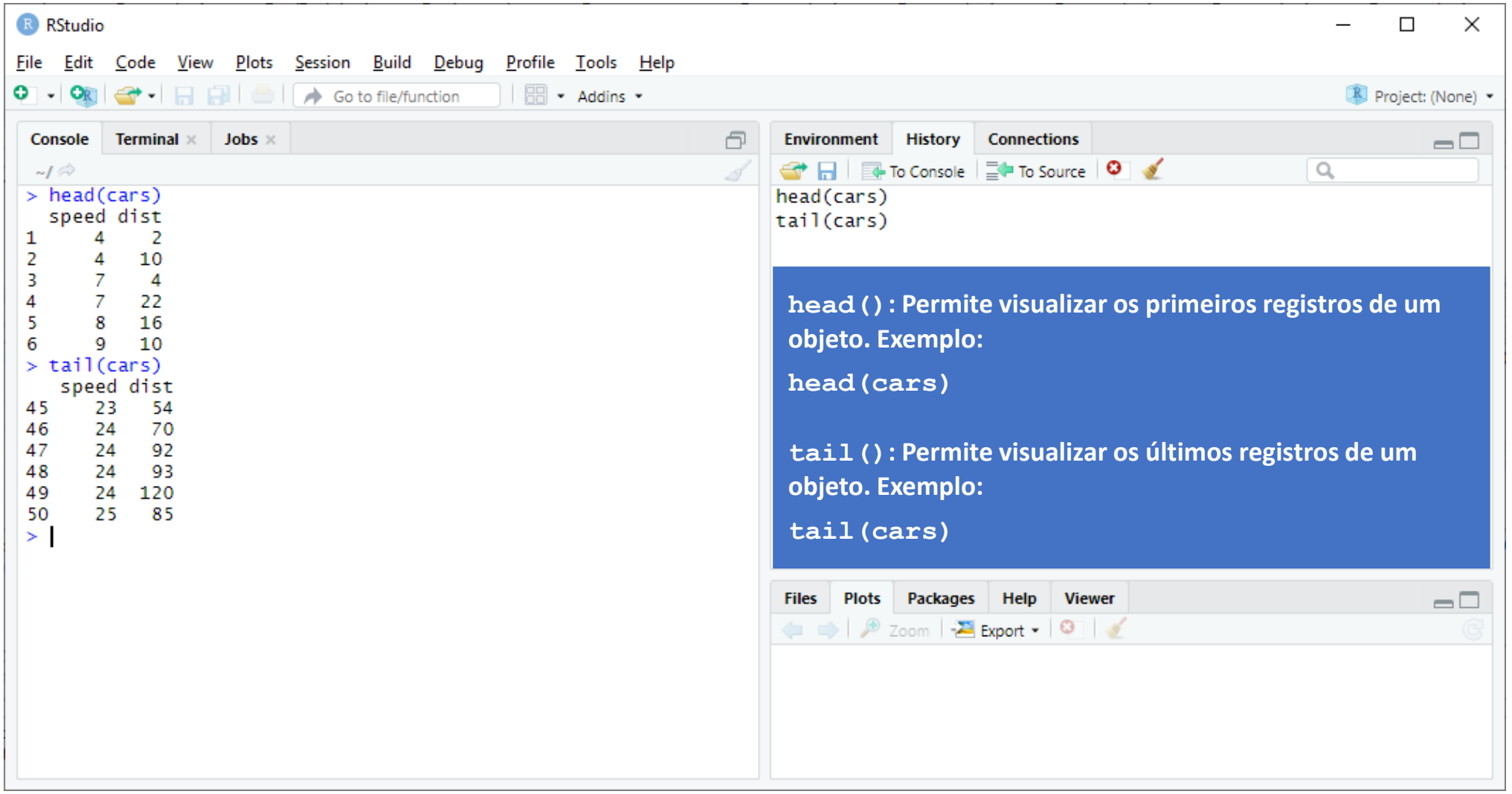


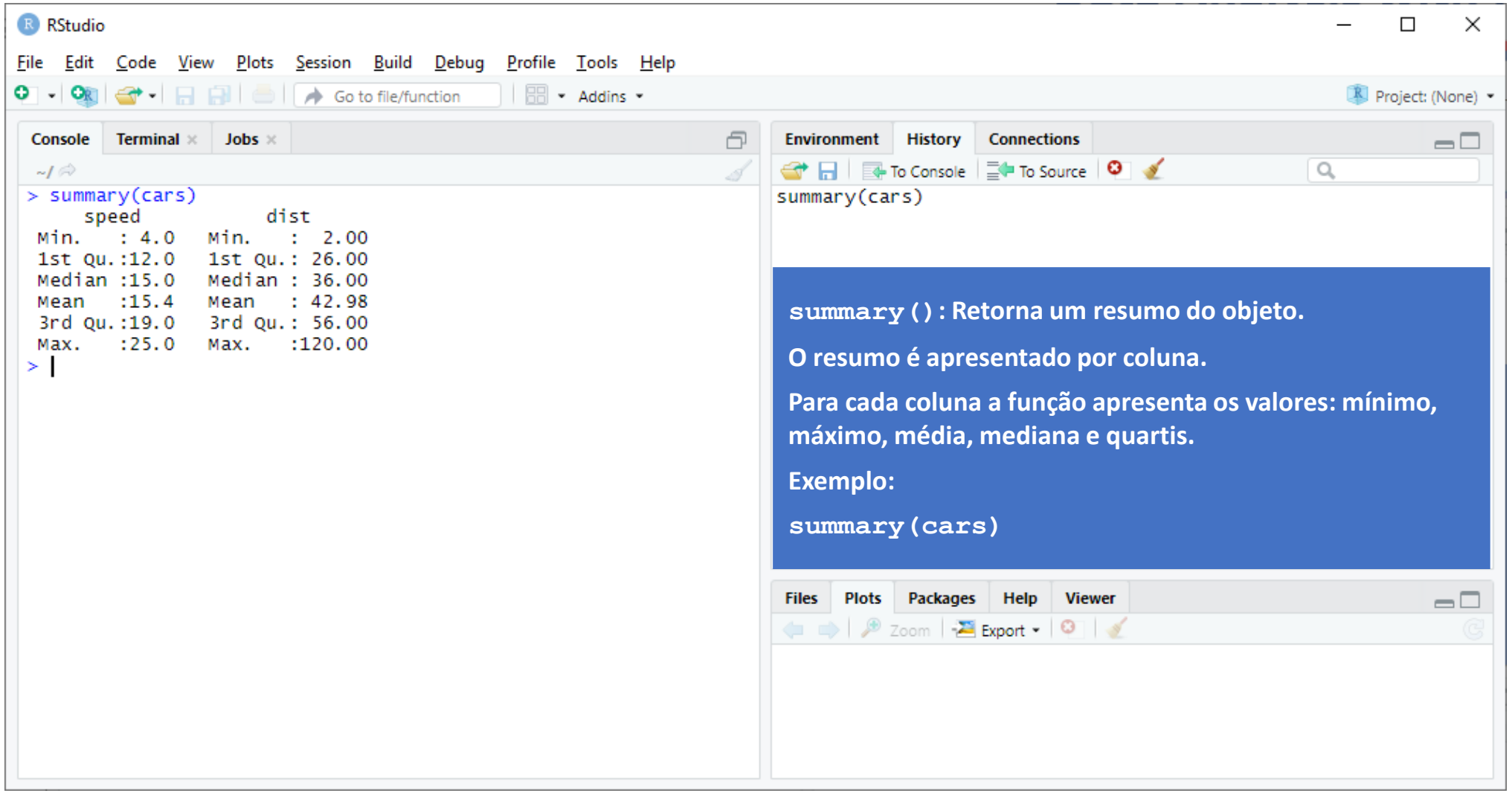




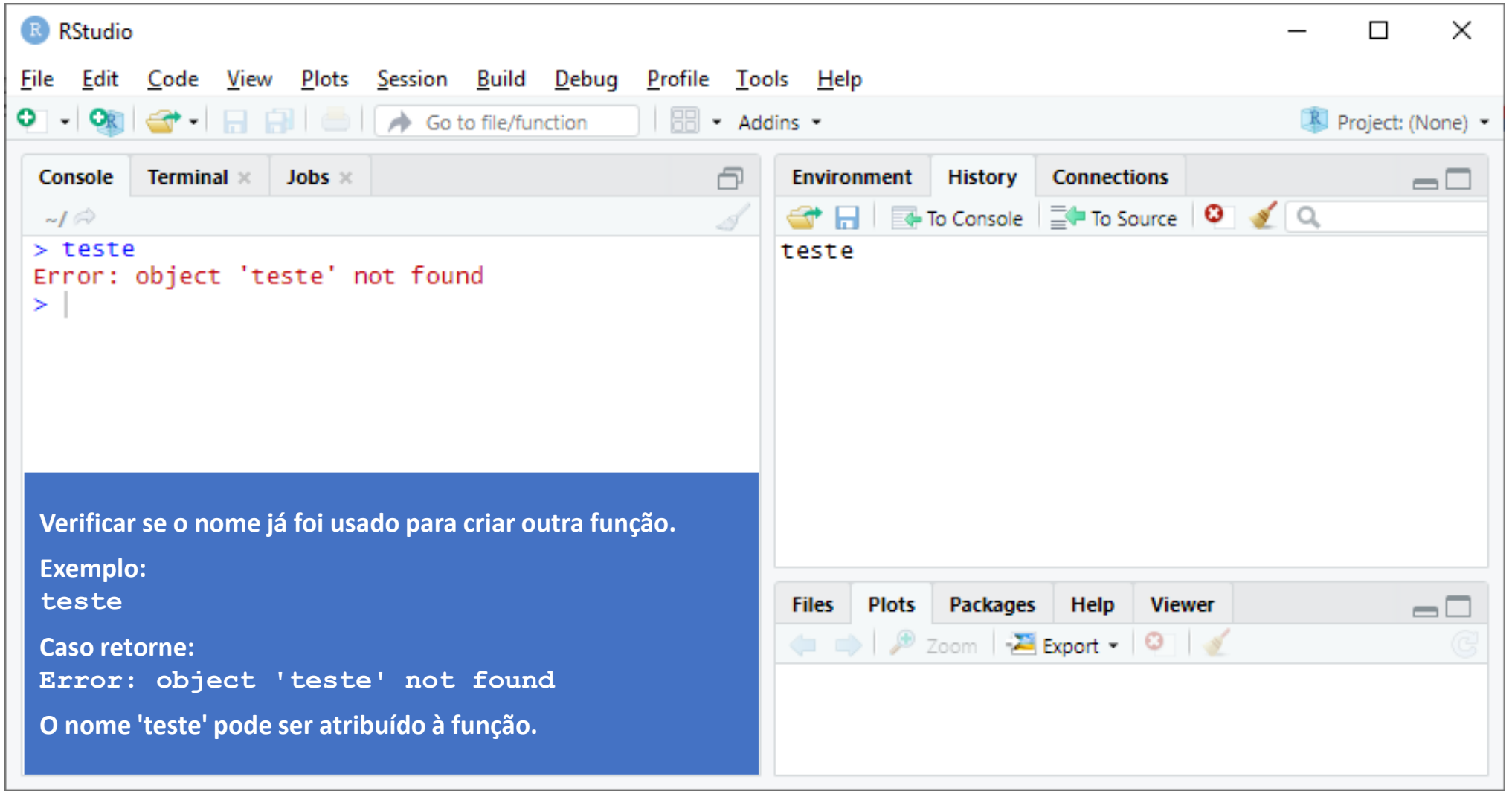
Funções:

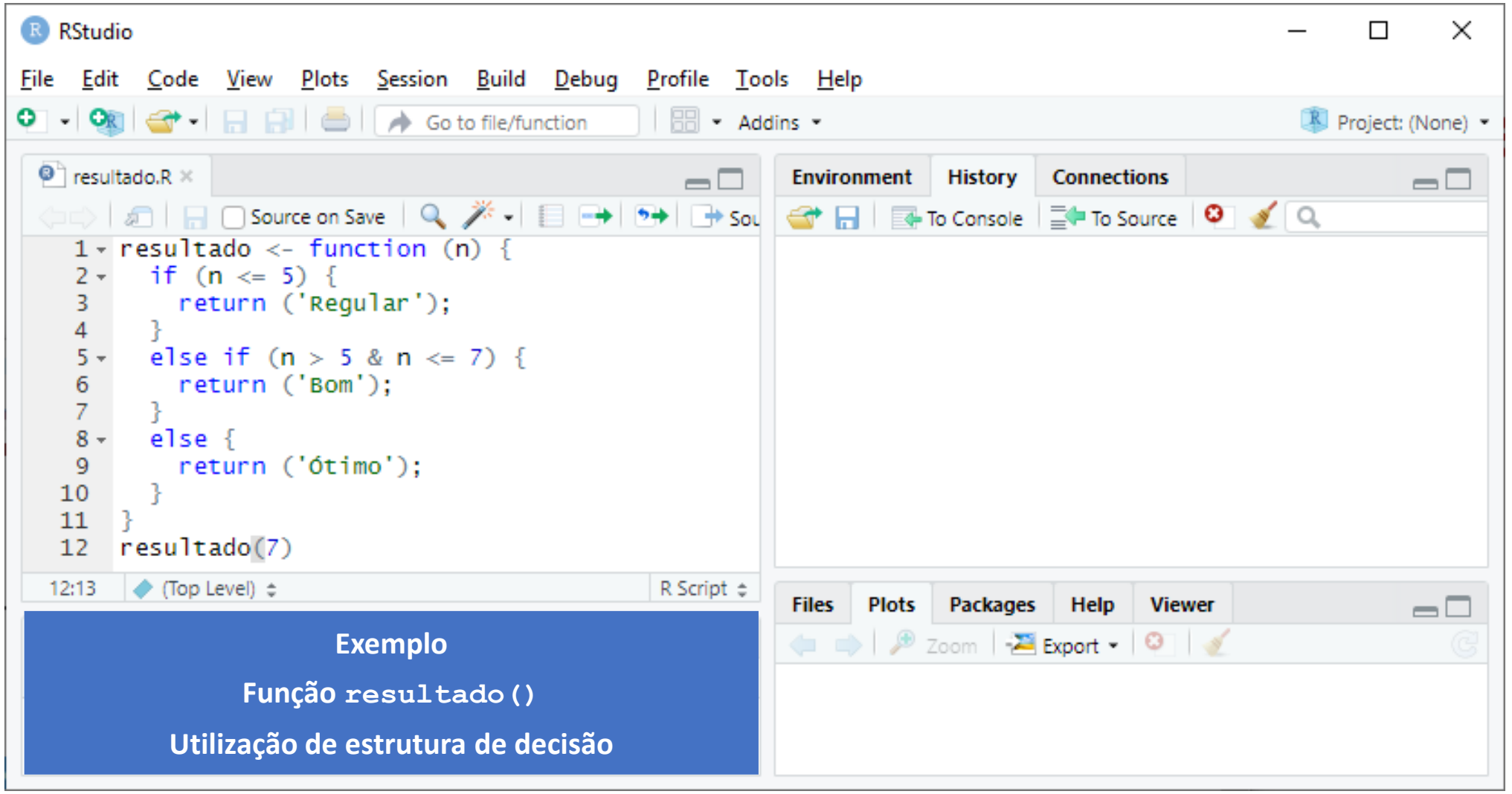
head, tail e summary





Criando Funções

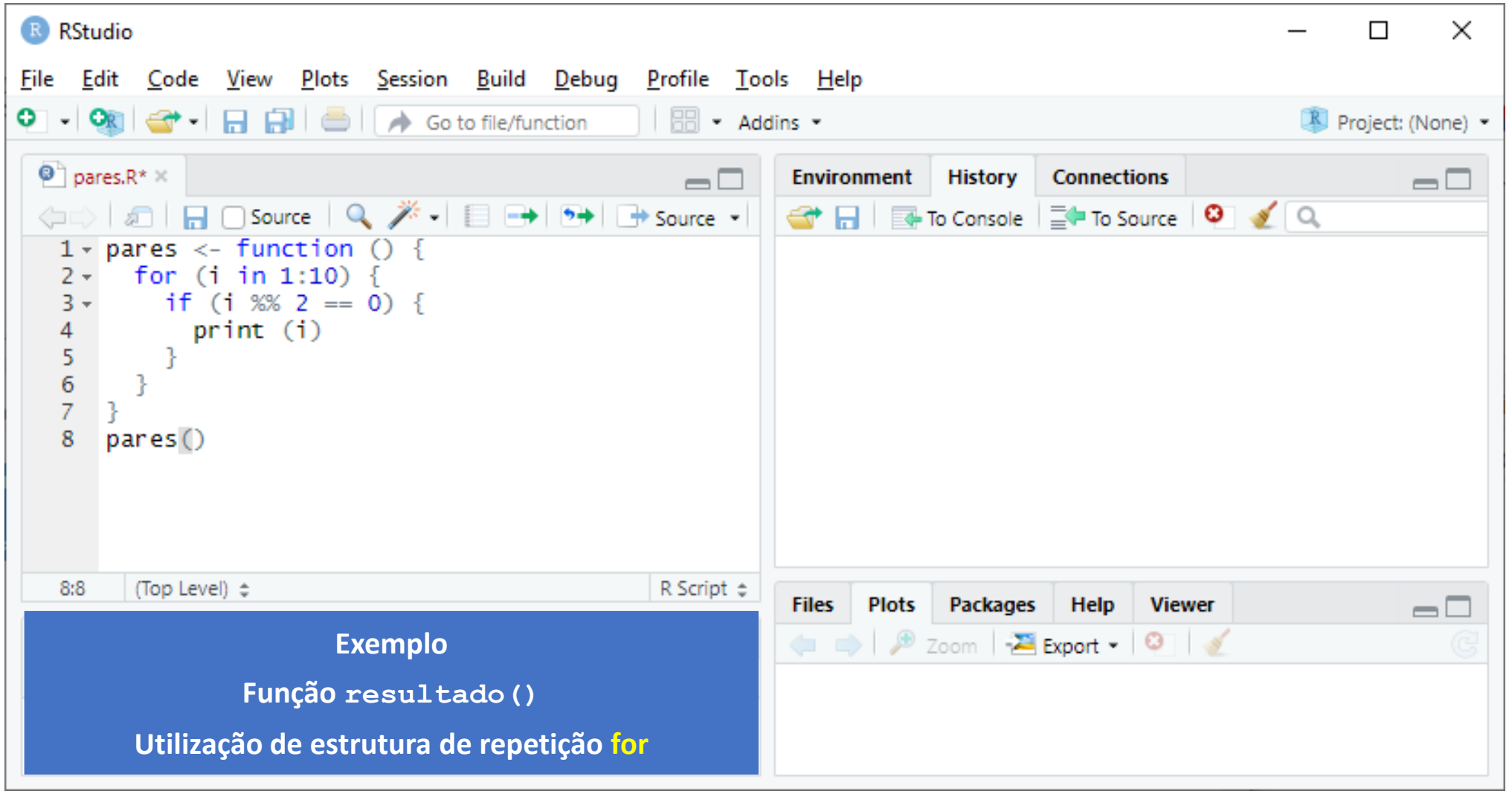


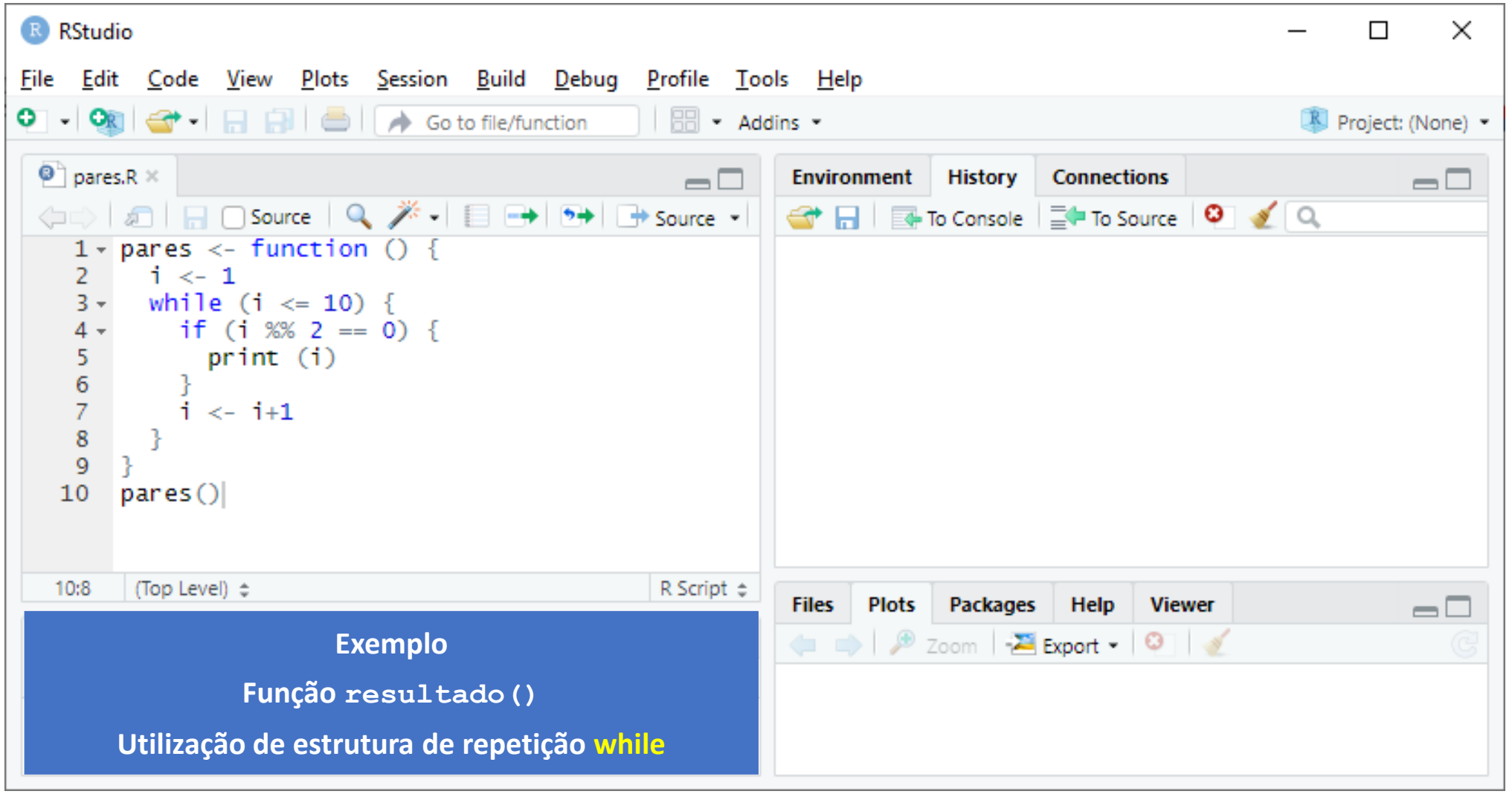


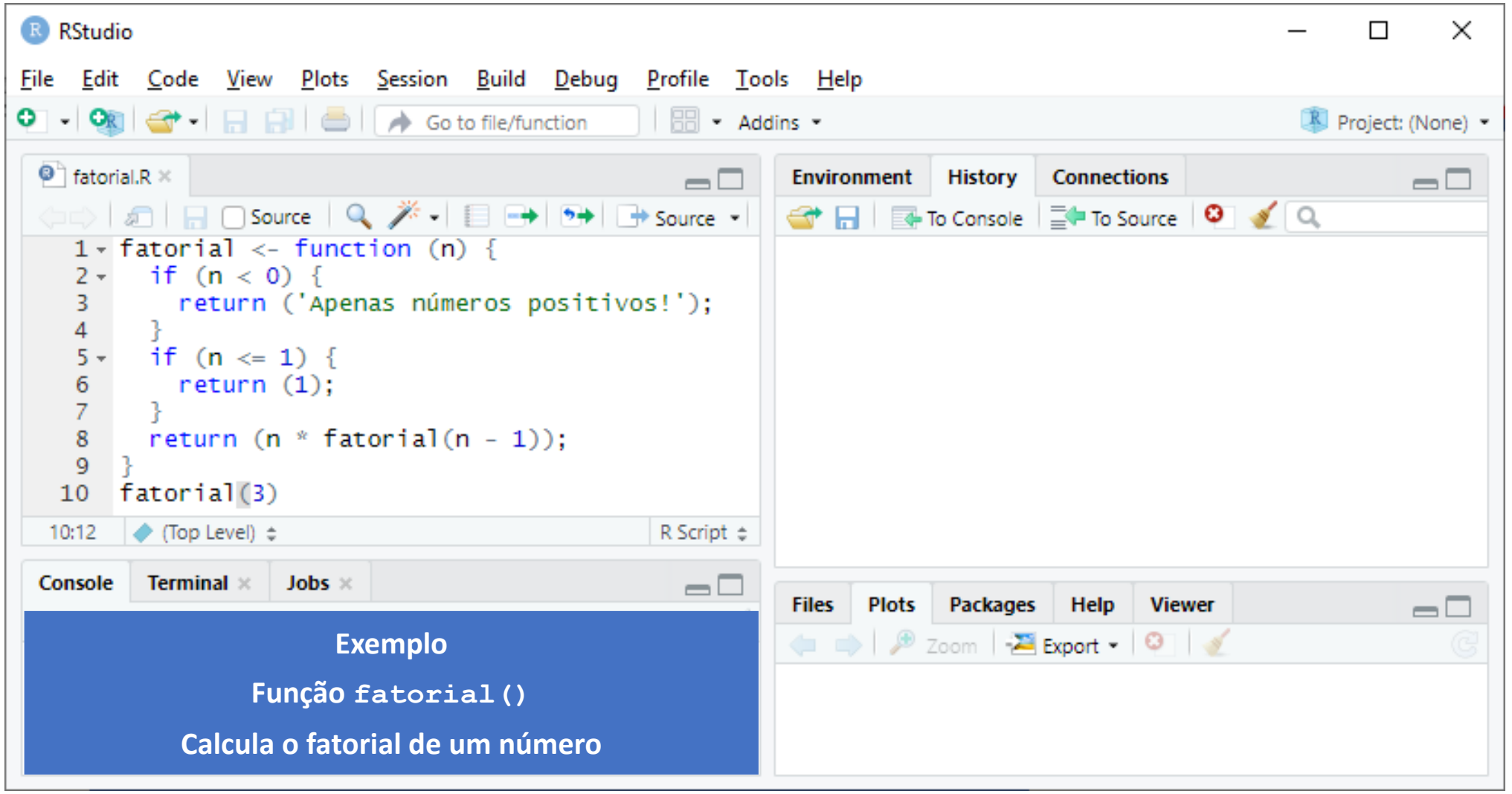
Exemplo

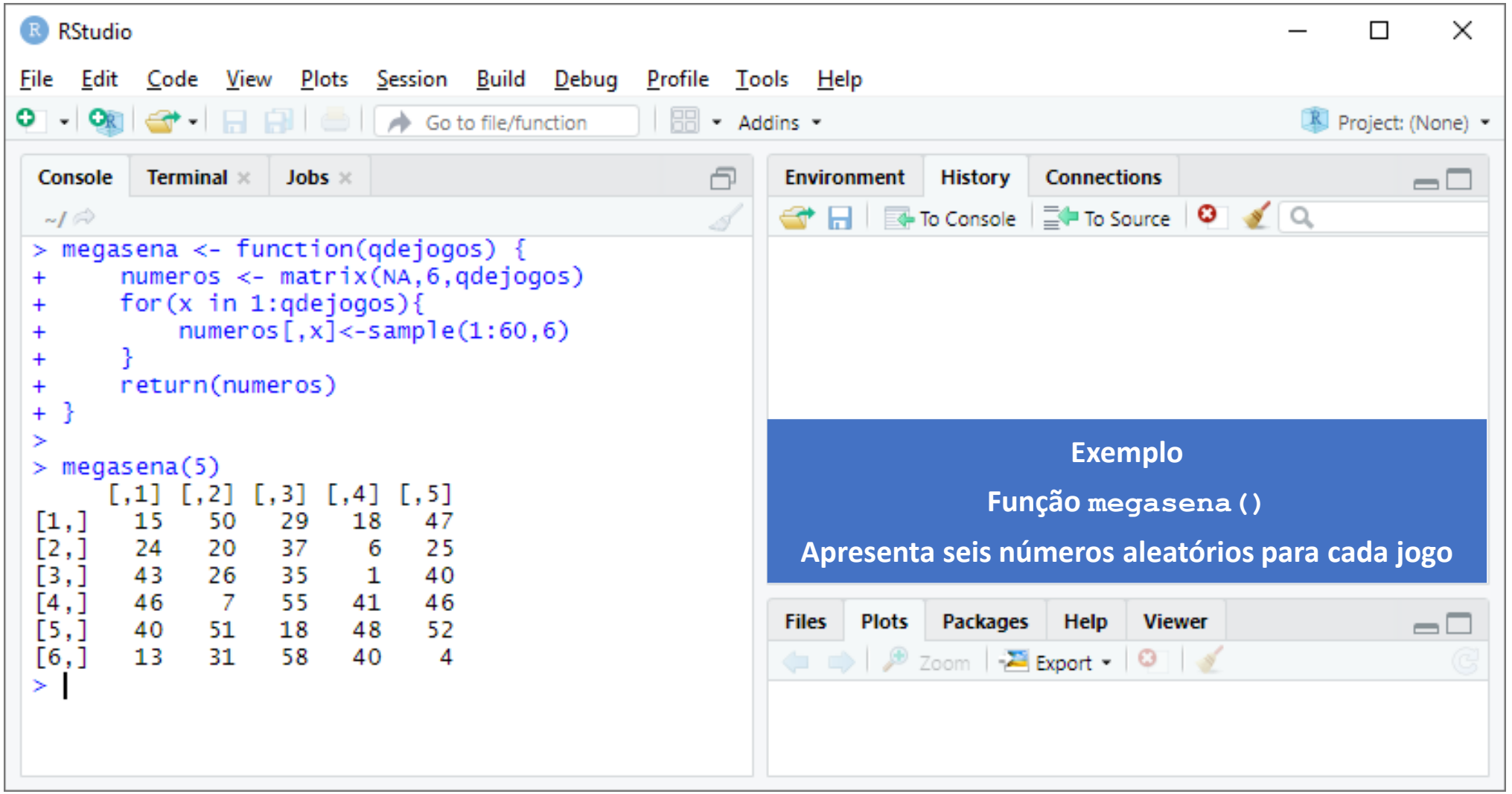
Função `resultado()`

Utilização de estrutura de decisão









Análise Descritiva

Medidas de Resumo Numérico

Medidas de resumo numérico

Uma forma de resumir os dados é através das medidas-resumo. A natureza da variável vai orientar qual tipo de medida deve ser utilizada.

Há três tipos de medidas-resumo:

- Tendência central
- Dispersão(variabilidade)
- Separatrizes

Medidas de resumo numérico

Medidas de Tendência Central

Informam o valor do ponto em torno do qual os dados se distribuem.

Representam um conjunto de valores ou dados através de um único número.

As medidas de tendência central mais utilizadas são:

- **Média aritmética:** soma de todos os valores e dividido pelo número de dados do conjunto;
- **Moda:** valor com maior frequência no conjunto de dados;
- **Mediana:** valor central (quando o número de elementos do conjunto é ímpar) ou média aritmética dos dois valores mais centrais (quando o número de elementos do conjunto é par).

Medidas de resumo numérico

Medidas de Dispersão ou Variabilidade

Usadas para verificar a regularidade de um conjunto de dados, quanto menor o grau de dispersão, maior é a regularidade.

As principais medidas de dispersão são:

- **Amplitude:** Apresenta a diferença entre o maior e o menor valor de um conjunto de dados.
- **Desvio absoluto médio:** Apresenta o afastamento médio dos elementos em relação à média aritmética ou em relação à mediana.
- **Variância:** Mede a variabilidade do conjunto de dados em termos de desvios quadrados em relação à média aritmética.
- **Desvio padrão:** É a raiz quadrada a variância. Indica o quanto um conjunto de dados é uniforme. Quanto mais próximo de 0 (zero) for o desvio padrão, mais homogêneos são os dados.
- **Coeficiente de variação:** É a razão do desvio padrão pela média. É frequentemente expresso como uma porcentagem.

Medidas de resumo numérico

Medidas de Separatrizes

Dividem a distribuição em um certo número de partes iguais. O objetivo das separatrizes é proporcionar uma melhor ideia da dispersão do conjunto de dados, principalmente da simetria ou assimetria da distribuição.

As medidas de separatrizes são:

- **Quartis:** Dividem a distribuição em quatro partes iguais.
    - O primeiro quartil ou quartil inferior delimita os 25% menores valores.
    - O segundo quartil ou quartil central é a mediana, que separa os 50% menores dos 50% maiores valores.
    - O terceiro quartil ou quartil superior delimita os 25% maiores valores.
- **Decis:** Dividem a distribuição em 10 partes iguais.
- **Percentis:** Dividem a distribuição em 100 partes iguais.

Tendência Central

Tendência Central

Indicam, em geral, um valor central em torno do qual os dados estão distribuídos.

As principais medidas de tendência central são:

- média aritmética
- mediana
- moda

Média aritmética

A média aritmética indica o valor em torno do qual há um equilíbrio na distribuição dos dados e pode ser expressa da seguinte forma:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Isto significa que deve-se somar todos os valores observados e dividir o resultado pelo tamanho da amostra.

A função **`mean()`** do R calcula a **média**:

```
> a <- c(1,2,2,2,2,3,4,4,5,6,6,7,8,8,8,9)
> mean(a)
[1] 4.8125
```

Mediana

A mediana é o valor central em um conjunto de dados após ordenado. Esta medida divide os dados em duas partes iguais, sendo metade dos valores abaixo da mediana e, a outra metade, acima.

Se um conjunto de dados contém um total de n observações, no qual n é ímpar, a mediana é o valor do meio. Se n for par, a mediana é usualmente tomada como média dos dois valores mais centrais do intervalo.

A função **`median()`** do R calcula a **mediana**:

```
> a <- c(1,2,2,2,2,3,4,4,5,6,6,7,8,8,8,9)
> median(a)
[1] 4.5
```

Moda

A moda é o valor mais frequente no conjunto de dados.

Esta medida-resumo também pode ser aplicada a variáveis qualitativas.

R não tem uma função in-built padrão para calcular a moda. Portanto, para calcular a moda, é preciso instalar o package **modeest**. Após a instalação, a função **`mfv()`** "most frequent value" deve ser usada para obter-se a **moda**.

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

> a <- c(1,2,2,2,2,3,4,4,5,6,6,7,8,8,8,9)
> mfv(a)
[1] 2
```

Dispersão

Dispersão

Medidas de dispersão têm o objetivo de quantificar a **variabilidade** dos valores em um conjunto de dados. Há várias medidas de dispersão:

- Amplitude total
- Desvio médio absoluto
- Variância
- Desvio padrão
- Coeficiente de variação

Amplitude total

A amplitude total é obtida a partir da diferença entre o máximo($x_{(n)}$) e mínimo($x_{(1)}$) em um conjunto de dados ordenados.

Esta medida possui o valor 0 (zero) como limite inferior e é altamente sensível à valores extremos.

$$\Delta = x_{(n)} - x_{(1)}$$

A diferença entre dos valores retornados pelas funções **`max()`** e **`min()`** do R apresenta a **amplitude** do conjunto de dados:

```
> a <- c(1,2,3,4,5,6,7,8,9)
> amplitude <- (max(a)-min(a))
> amplitude
[1] 8
```

Saiba mais:
A função `range()` retorna os valores max e min:

```
> range(a)
[1] 1 9
```

Desvio mediano absoluto

O desvio mediano absoluto (MAD) é uma medida robusta da tendência central. Apresenta, portanto, bom desempenho para dados extraídos de uma ampla gama de distribuições de probabilidade não normalmente distribuídas.

Diferente da combinação de média e desvio padrão, o MAD não é sensível à presença de outliers.

```
> a <- c(1,2,3,4,5,6,7,8,9)
> mad(a)
[1] 2.9652
```

Desvio médio absoluto

É o cálculo da média dos desvios absolutos. Para o seu cálculo, primeiramente deve ser calculada a média, posteriormente os desvios das observações em relação a média em módulo, e por último a média aritmética destes desvios.

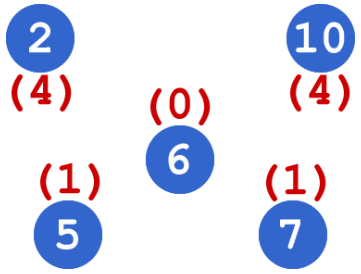
$$dma = \frac{\sum|x_i - \bar{x}|}{n}$$

Lembre-se:

$\bar{x}$ → média

Calcular o **desvio médio absoluto** no R:

```
> x <- c(2,5,6,7,10)
> dma <- sum(abs(x-mean(x)))/length(x)
[1] 2
```



Desvio médio absoluto:

x = (2, 5, 6, 7, 10)
n = 5
média = (2+5+6+7+10) / 5 = 6
| 2-6 | = 4
| 5-6 | = 1
| 6-6 | = 0
| 7-6 | = 1
| 10-6 | = 4
(4+1+0+1+4) = 10
10 / n = 10 / 5 = 2 **dma = 2**

Variância

Variância é a medida de **dispersão** mais usada e conhecida na Estatística. Ela **quantifica a variabilidade dos dados em torno da média**, ou seja, mede o quão longe do valor esperado os dados se encontram. O seu cálculo é feito da seguinte maneira:

$$Var = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Lembre-se:

$\bar{x}$ -> média

A função **`var()`** do R calcula a **variância amostral**:

```
> x <- c(10,12,16,17,20)
> var(x)
[1] 16
```

Cálculo da variância:

n = 5
média = (10+12+16+17+20)/5 = 15
10-15 = -5 => $-5^2 = 25$
12-15 = -3 => $-3^2 = 9$
16-15 = 1 => $1^2 = 1$
17-15 = 2 => $2^2 = 4$
20-15 = 5 => $5^2 = 25$
25+9+1+4+25 = 64
variância = 64 / (5-1) = 16

Desvio padrão

O cálculo da variância eleva ao quadrado a soma da diferença dos desvios. Isso faz com a variância não tenha a interpretação na mesma escala em os dados foram medidos. Utiliza-se, portanto, o desvio padrão, obtido através do cálculo da raiz quadrada da variância. Este novo valor possibilita analisar a variabilidade dos dados na mesma escala em que eles foram medidos.

$$sd = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

A função **`sd()`** do R calcula o **desvio padrão**:

```
> x <- c(10,12,16,17,20)
> sd(x)
[1] 4
```

Desvio padrão é frequentemente representado pela letra grega σ (sigma)

Coeficiente de variação

O coeficiente de variação é a razão entre o desvio padrão e a média. É uma medida relativa que avalia o percentual de variabilidade em relação a média observada.

$$cv = \frac{sd}{\bar{x}} 100$$

O **coeficiente de variação** é calculado no R da seguinte forma:

```
x <- c(10,12,16,17,20)
> cv = (sd(x)/mean(x))*100
> cv
[1] 26.66667 # Aproximadamente 26%
```

Separatrizes

Separatrizes

Separatrizes são medidas que **dividem os dados em um número de partes iguais**. Para isso a amostra deve estar ordenada, portanto as separatrizes são aplicadas a variáveis quantitativas e qualitativas ordinais.

As principais separatrizes são:

- **percentis** (dividem os dados em 100 partes iguais)
- **decis** (dividem os dados em 10 partes iguais)
- **quartis** (dividem os dados em quatro partes iguais)

Os quartis são utilizado com mais frequência.

Quartis

Os quartis dividem a amostra em quatro partes iguais. O primeiro quartil (Q1) estabelece o limite entre as 25% menores observações e as 75% maiores. O segundo quartil (Q2) é igual a mediana e o terceiro quartil (Q3) separa as 75% menores observações das 25% maiores.

Cálculo do IQR (Inter Quarter Range): Q3 – Q1

```
> summary(women$height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   58.0    61.5    65.0    65.0    68.5    72.0
> quantile(women$height)
  0%  25%  50%  75% 100%
58.0 61.5 65.0 68.5 72.0
> IQR(women$height)
[1] 7
```

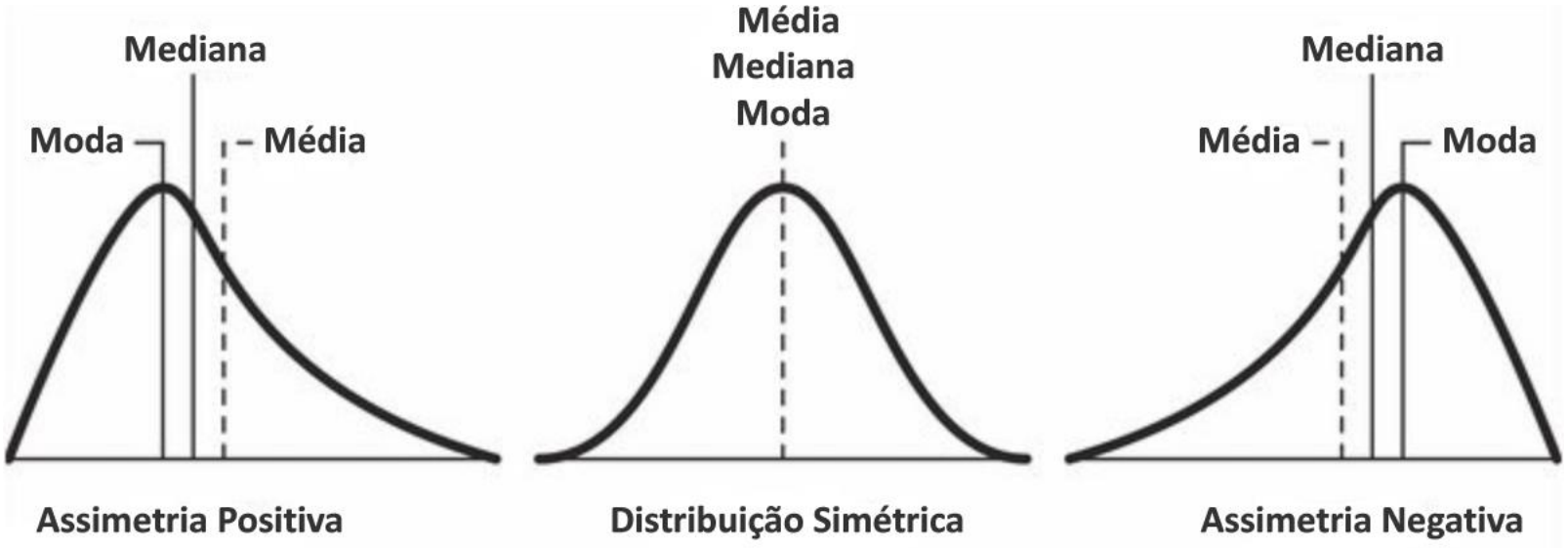
```
> women
   height weight
1      58    115
2      59    117
3      60    120
4      61    123
5      62    126
6      63    129
7      64    132
8      65    135
9      66    139
10     67    142
11     68    146
12     69    150
13     70    154
14     71    159
15     72    164
```

Distribuições
Simétricas e Assimétricas

Distribuições Simétricas e Assimétricas

Distribuições Simétricas: Apresentam uma curva de frequências **unimodal** com duas "caudas" simétricas em relação a uma linha vertical que passa por seu ponto mais alto (eixo de simetria).

Distribuições Assimétricas: Apresentam uma curva de frequências **unimodal** com uma "cauda" mais longa para a direita (assimetria positiva) ou para a esquerda (assimetria negativa) em relação ao seu ponto mais alto.



Coeficiente de Assimetria de Pearson

O Coeficiente de Assimetria de Pearson, A_p, baseia-se na posição relativa das medidas de tendência central de acordo com o tipo de assimetria dos dados.

$$Ap = \frac{3.(\bar{x} - md)}{sd}$$

$\bar{x}$ => média
md => mediana
sd => desvio padrão

Ap = 0 => Simetria
0.15 < | Ap | < 1 => Assimetria moderada
| Ap | > 1 => Assimetria forte

```
x <- c(10,20,20,30)
xa <- mean(x)    # 20
xb <- median(x) # 20
xc <- sd(x)      # 8.16
xAp <- 3*(xa-xb)/xc
xAp
[1] 0 #simetria
```

```
y <- c(10,20,20,50)
ya <- mean(y)    # 25
yb <- median(y) # 20
yc <- sd(y)      # 17.32
yAp <- 3*(ya-yb)/yc
yAp
[1] 0.8660 #positiva
```

```
z <- c(10,20,30,30)
za <- mean(z)    # 22.5
zb <- median(z) # 25
zc <- sd(z)      # 9.57
zAp <- 3*(za-zb)/zc
zAp
[1] -0.7836 #negativa
```

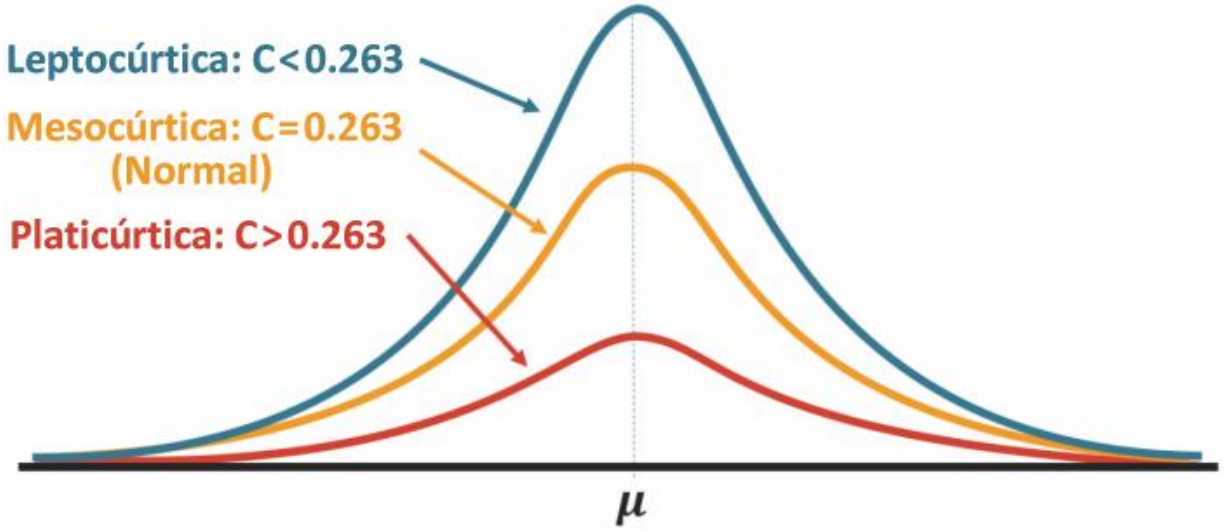
Curtose

Curtose

Curtose é uma medida de dispersão que caracteriza o "achatamento" da curva da função de distribuição.

Analisar um conjunto quanto à curtose significa verificar o "grau de achatamento da curva', ou seja, saber se a Curva de Frequência que representa o conjunto é mais "afilada" (leptocúrtica) ou mais "achatada" (platicúrtica) em relação a uma Curva Padrão, chamada de Curva Normal (mesocúrtica).

Assimetria e Curtose são medidas independentes e que não se influenciam mutuamente.



Curtose

O **Índice Percentílico de Curtose (C)** é calculado conforme a seguinte fórmula:

$$C = \frac{(Q3 - Q1)}{2.(D9 - D1)}$$

Q3: 3º Quartil (75%)
Q1: 1º Quartil (25%)
D9: 9º Decil (90%)
D1: 1º Decil (10%)

```
> x <- c(10,15,20,25,30,35,40)

> quantile(x,c(0.75,0.25,0.9,0.1))
 75%  25%  90%  10%
32.5 17.5 37.0 13.0

> C <- (32.5-17.5)/(2*(37.0-13.0))
> C
[1] 0.3125 # C > 0.263 -> Platicúrtica
```

Gráficos

Gráficos

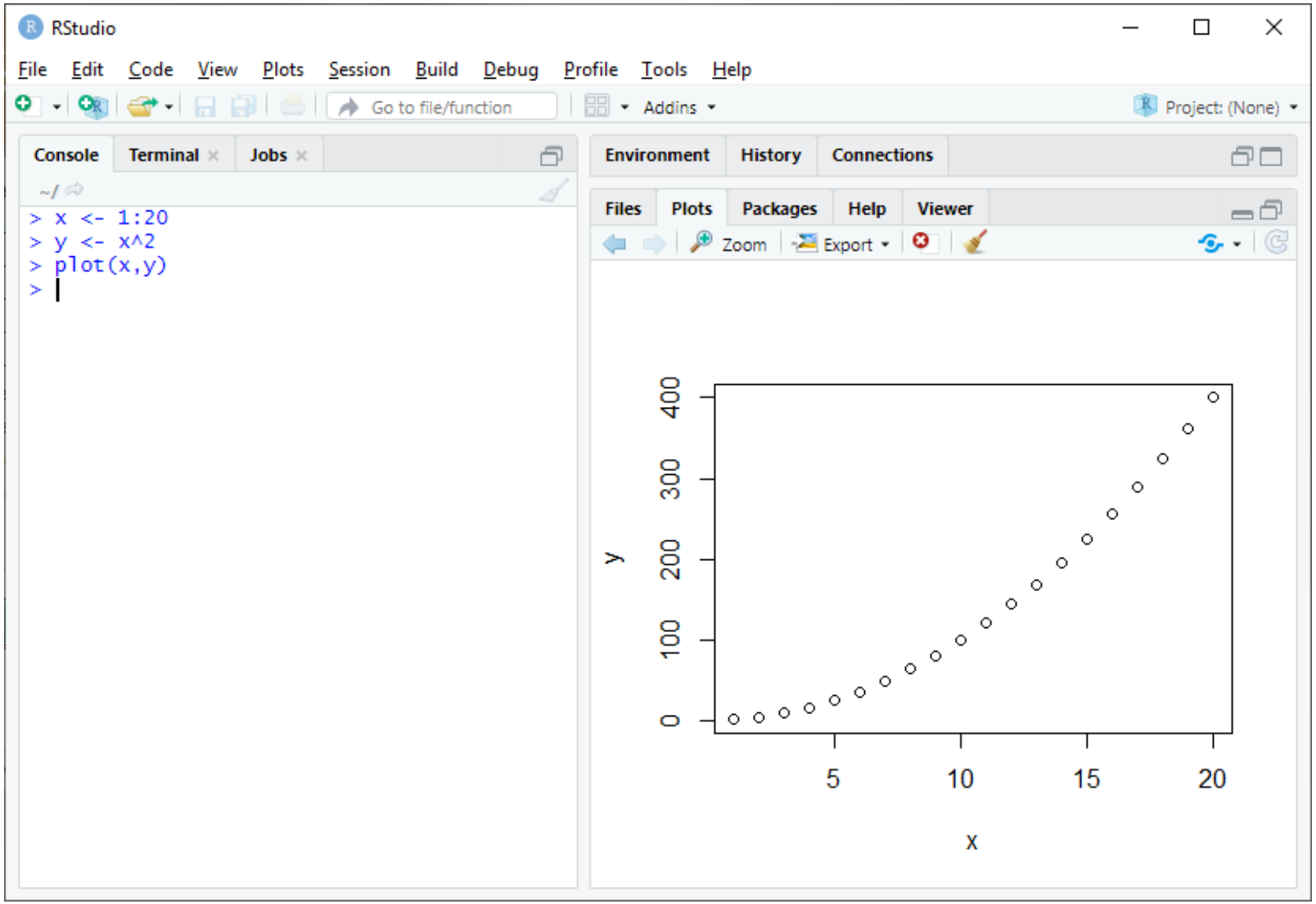
A visualização dos dados é uma etapa muito importante da análise estatística.

Visualizações podem ser uma simples medida resumo (frequência, média, variância, mínimo, máximo, etc.), um conjunto dessas medidas organizadas em uma tabela ou a representação (de uma parte) dos dados em um gráfico.

As ferramentas disponíveis no R facilitam muito a criação de gráficos dos mais diversos tipos.

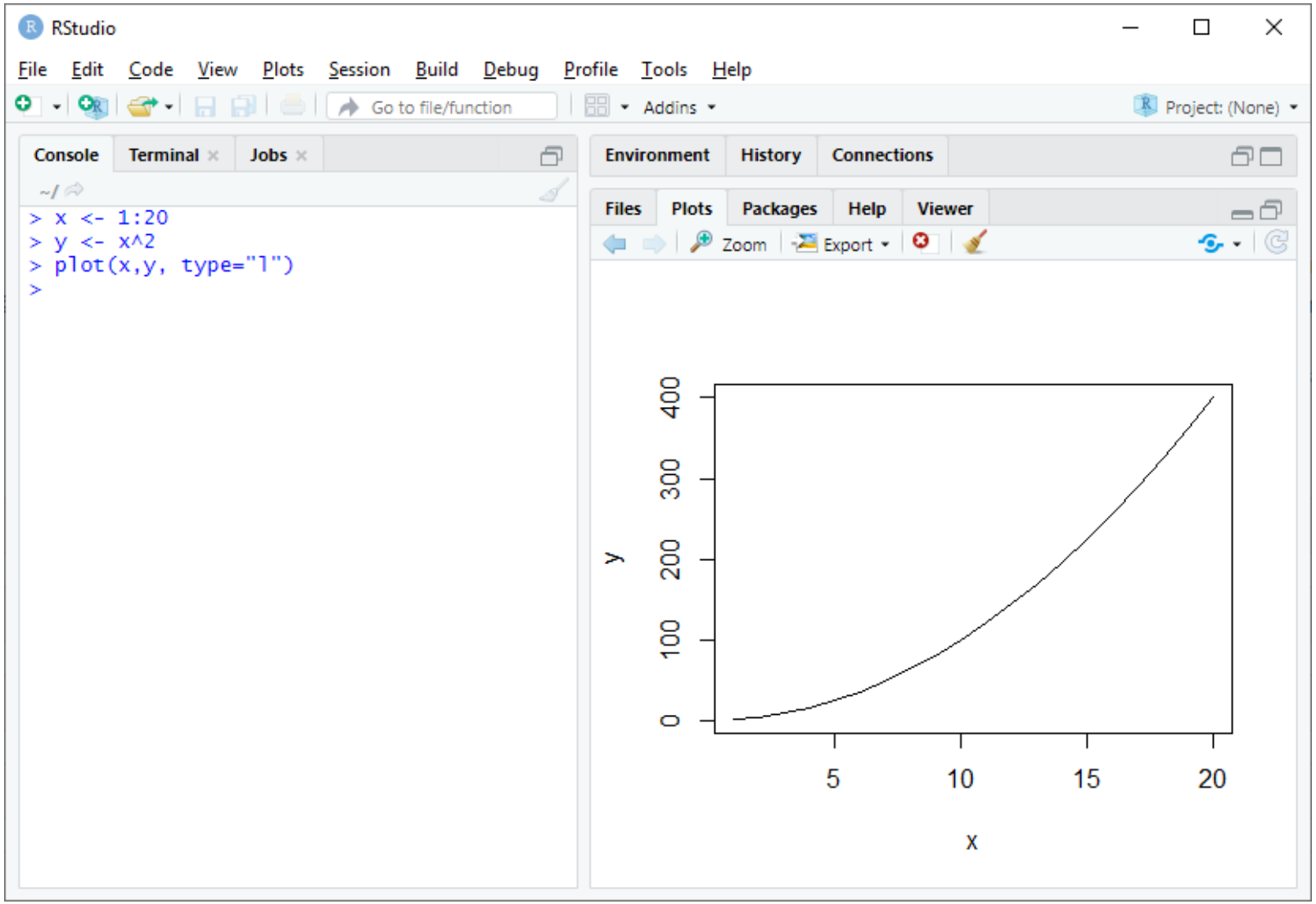
`plot()`

Gera um gráfico simples, atribuindo pontos em coordenadas cartesianas.



`plot()`

Acrescentando o argumento `type='l'` substitui os pontos por uma linha.



`plot()`

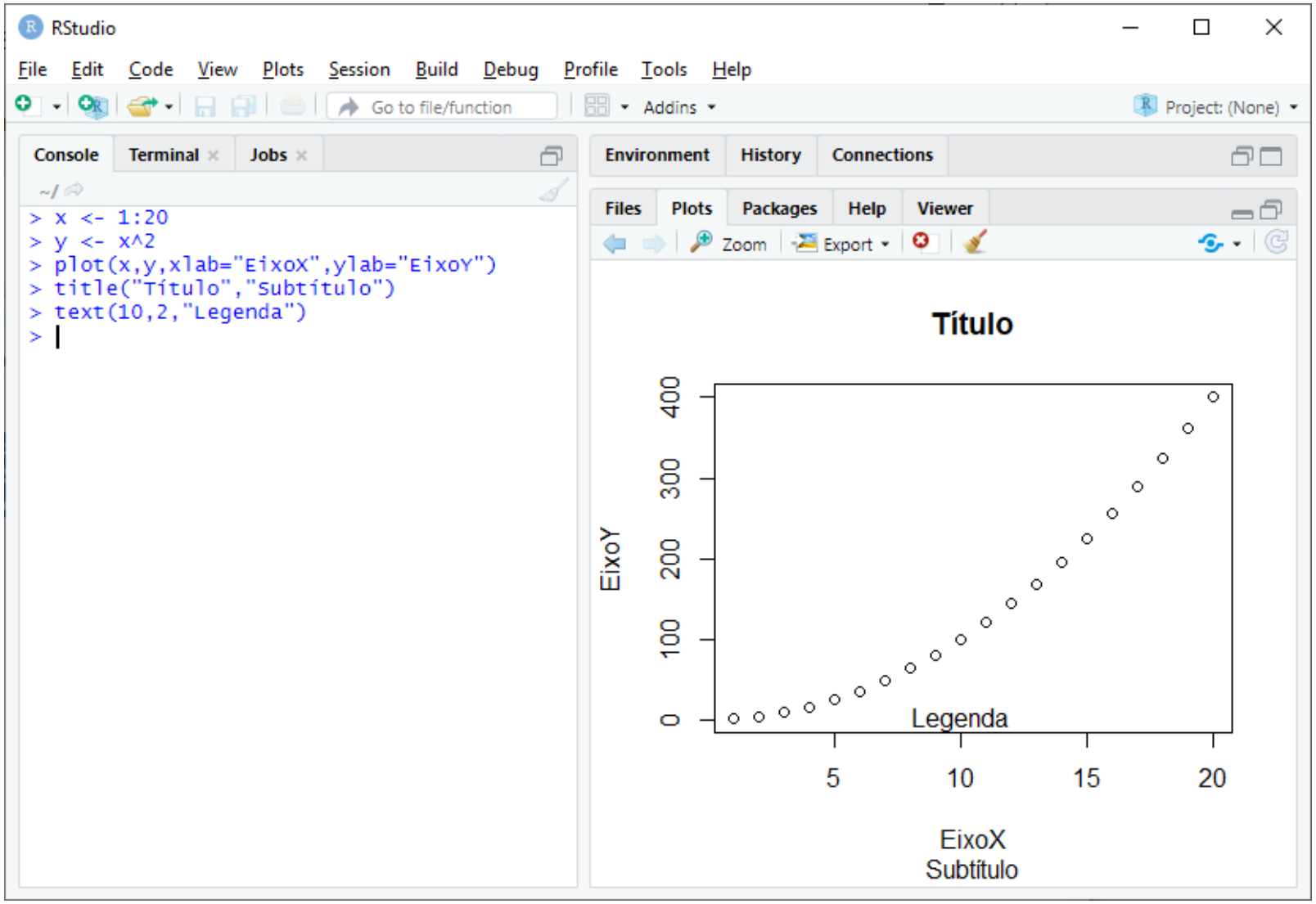
Outros parâmetros:

`xlab`: rótulo do eixo x (horizontal)

`ylab`: rótulo do eixo y (vertical)

`title`("Título","Subtítulo")

`text`(posX,poxY,"Legenda")



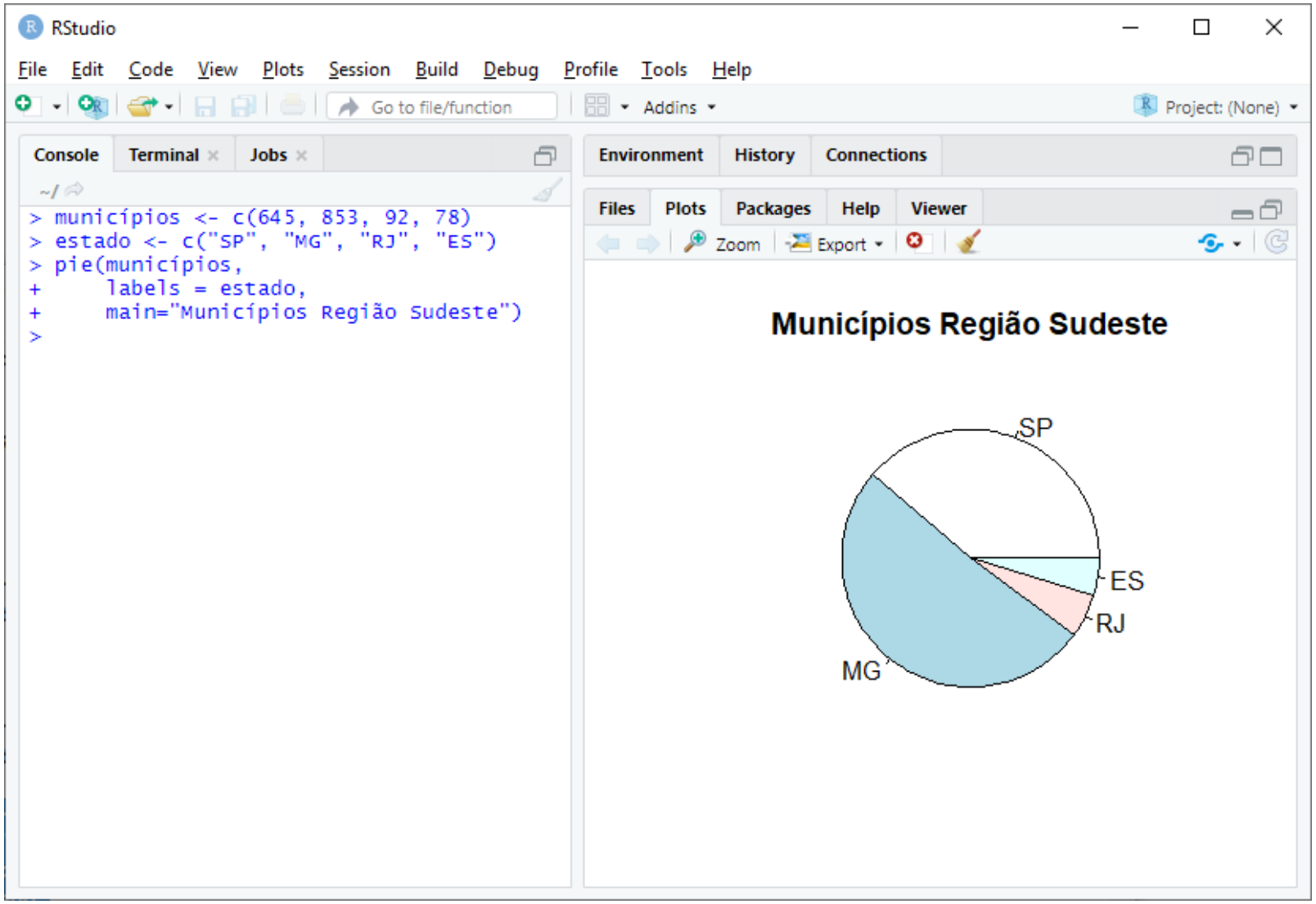
`pie()`

Gráficos de pizza ou de setor

Os gráficos de pizza exibem os dados como proporção de um todo.

São especialmente úteis para fazer comparações entre grupos.

Exemplo: Proporção de munícipios de cada estado na Região Sudeste do Brasil.



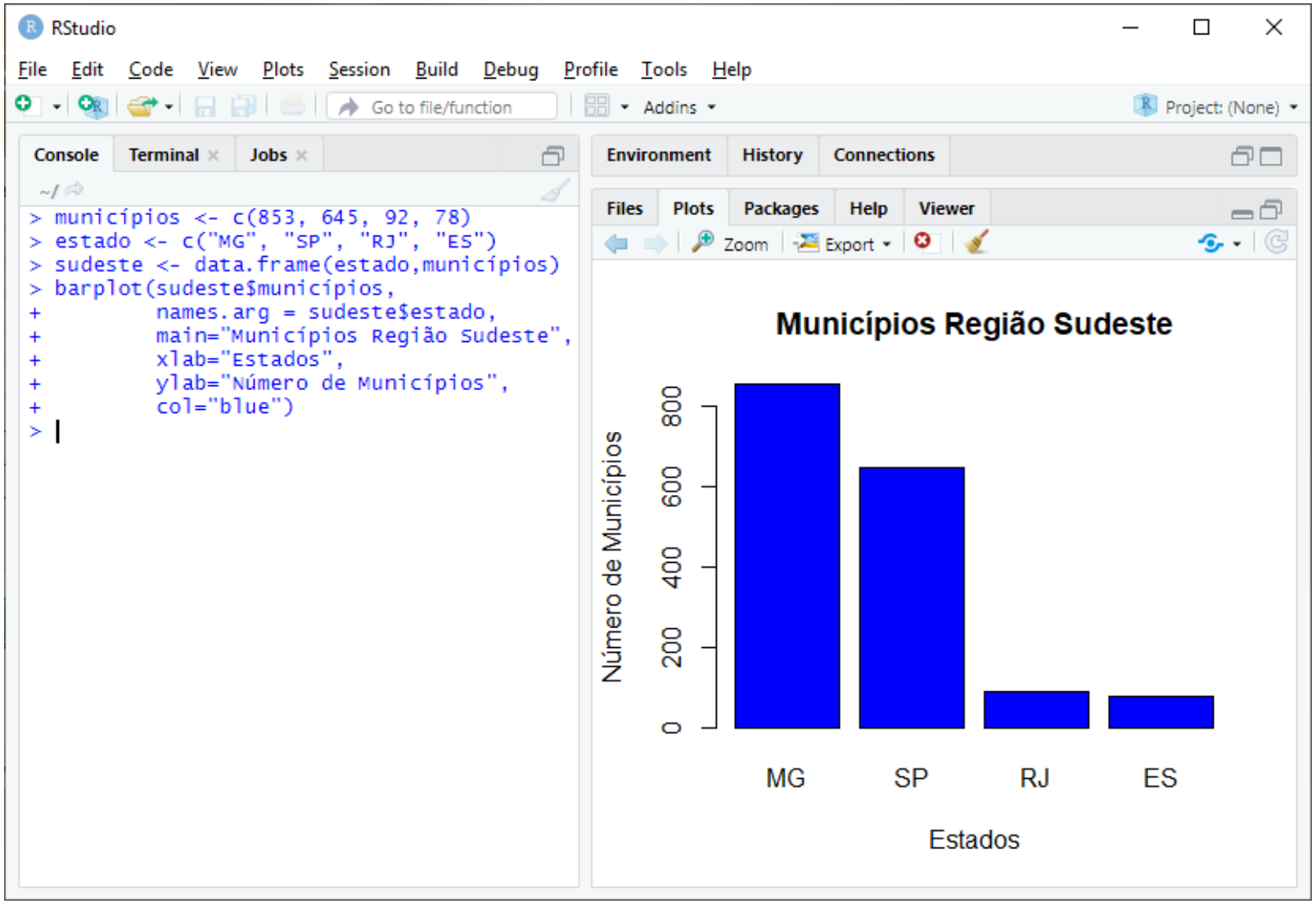
`barplot()`

Gráficos de Barras

No gráfico de barras cada barra representa a medida de cada elemento do vetor.

Cada barra é proporcional à "dimensão" do elemento.

Exemplo: Quantidade de municípios de cada estado da Região Sudeste do Brasil.

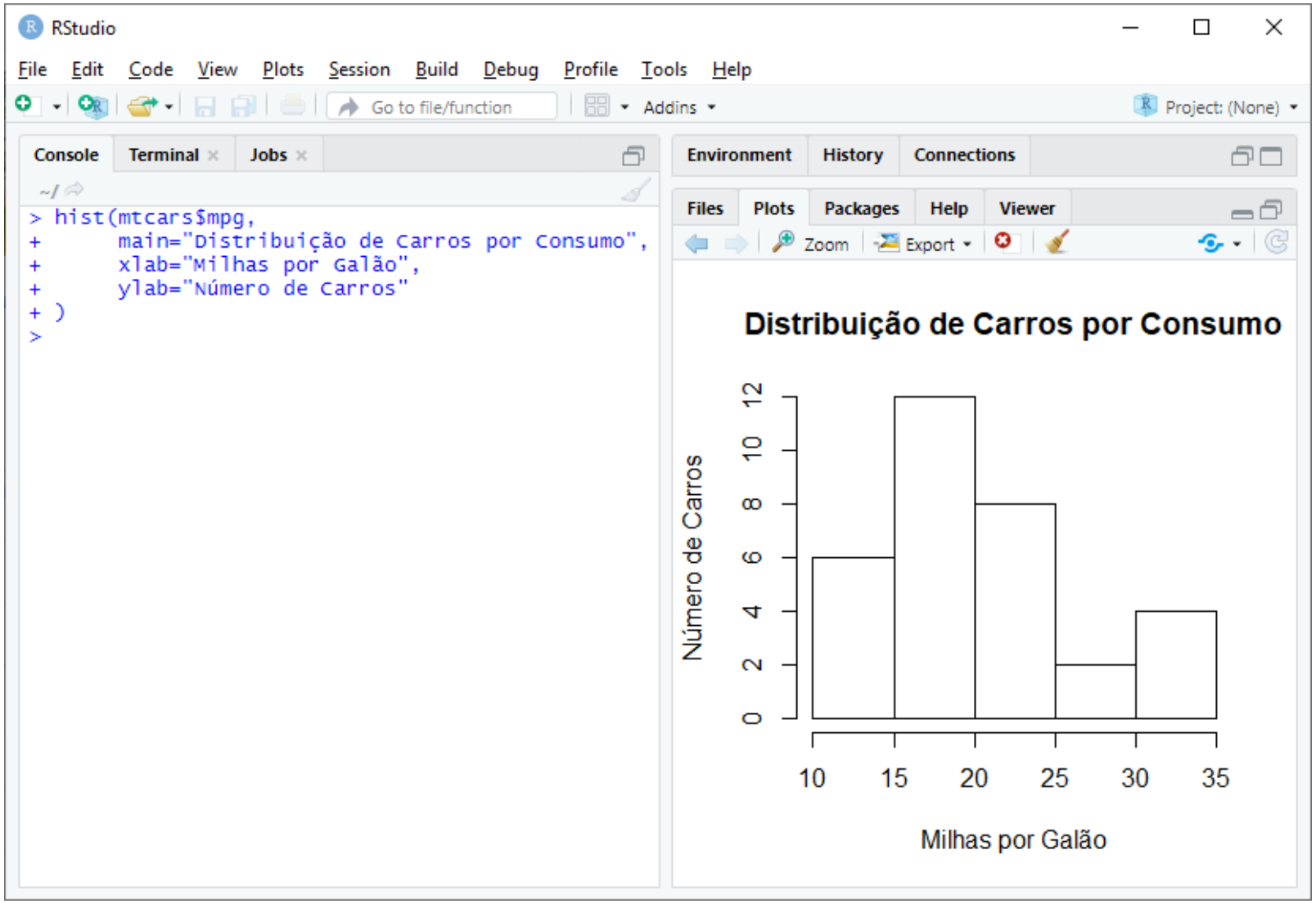


`hist()`

Histogramas

Um histograma divide uma série de dados em diferentes classes com igual espaçamento e mostra a frequência de valores em cada classe.

O eixo vertical (ordenadas) mostra a frequência relativa de cada classe e o eixo horizontal (abscissas) mostra os valores e intervalos das classes.



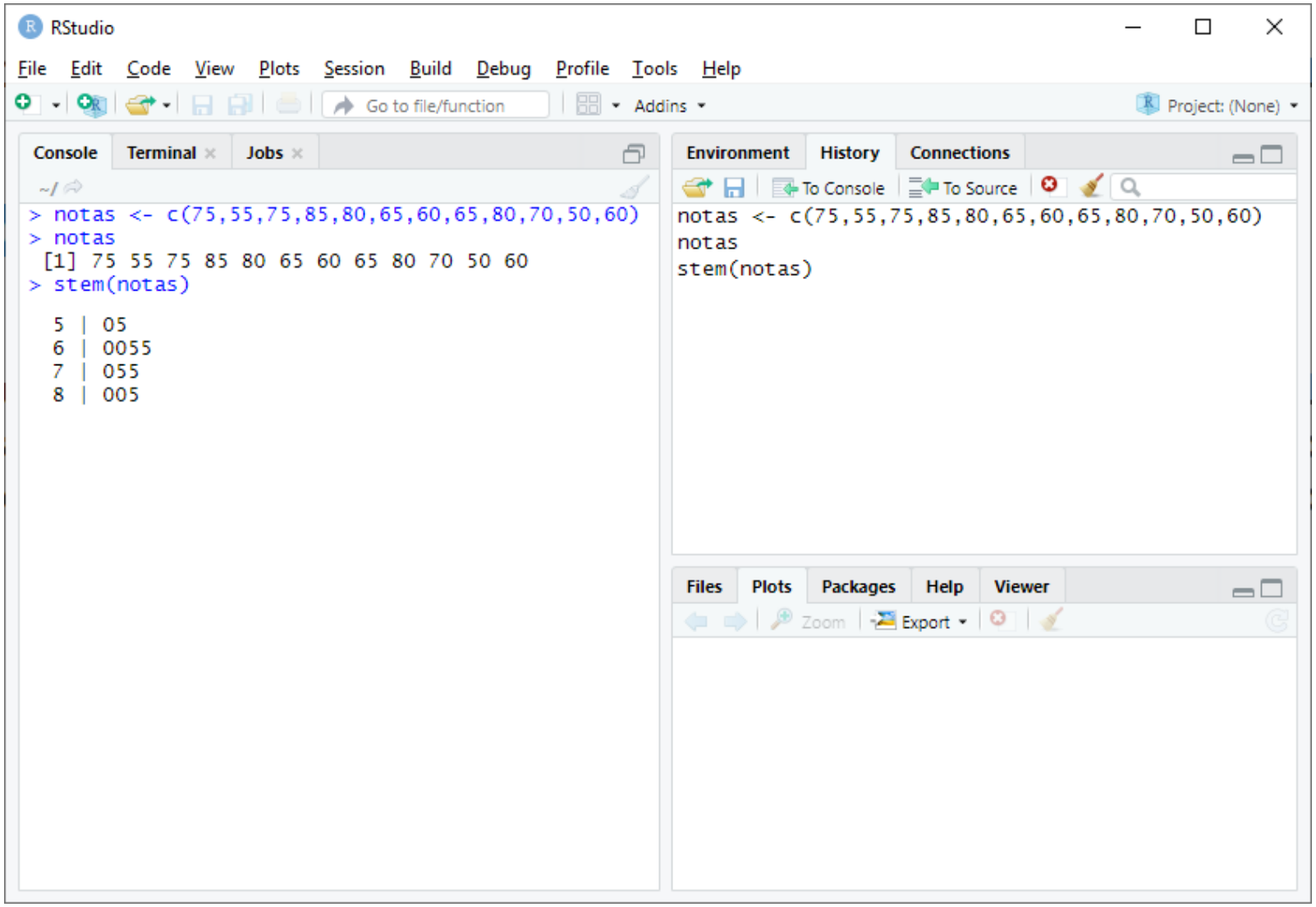
`stem()`

Ramos e Folhas

Stem and Leaf

É um tipo de representação entre a tabela e o gráfico, uma vez que, de um modo geral, são apresentados os verdadeiros valores dos dados, mas numa representação, que faz lembrar um histograma.

Os ramos são os dígitos dominantes colocados do lado esquerdo de um eixo vertical. Para cada dado toma-se o dígito que se segue imediatamente aos dígitos dominantes e coloca-se do lado direito do eixo, em frente ao respectivo ramo. Estes dígitos são as folhas.



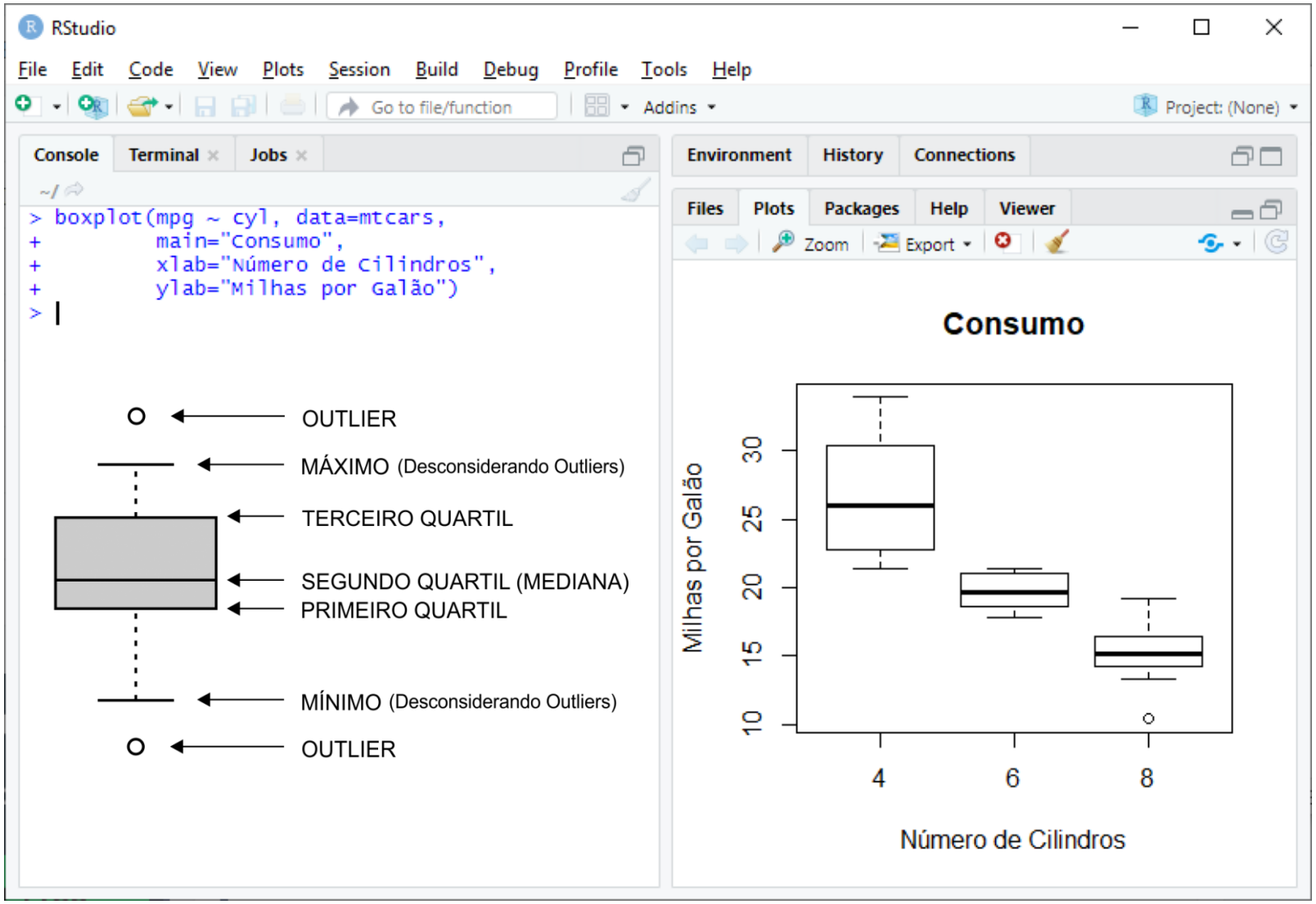
`boxplot()`

Gráficos de Caixa

Apresenta a distribuição de um conjunto de dados com base em alguns de seus parâmetros descritivos: mediana e quartis.

Permite avaliar a simetria e a dispersão do dados.

É recomendado para comparação de dois ou mais conjuntos de dados correspondentes às categorias de uma variável qualitativa.

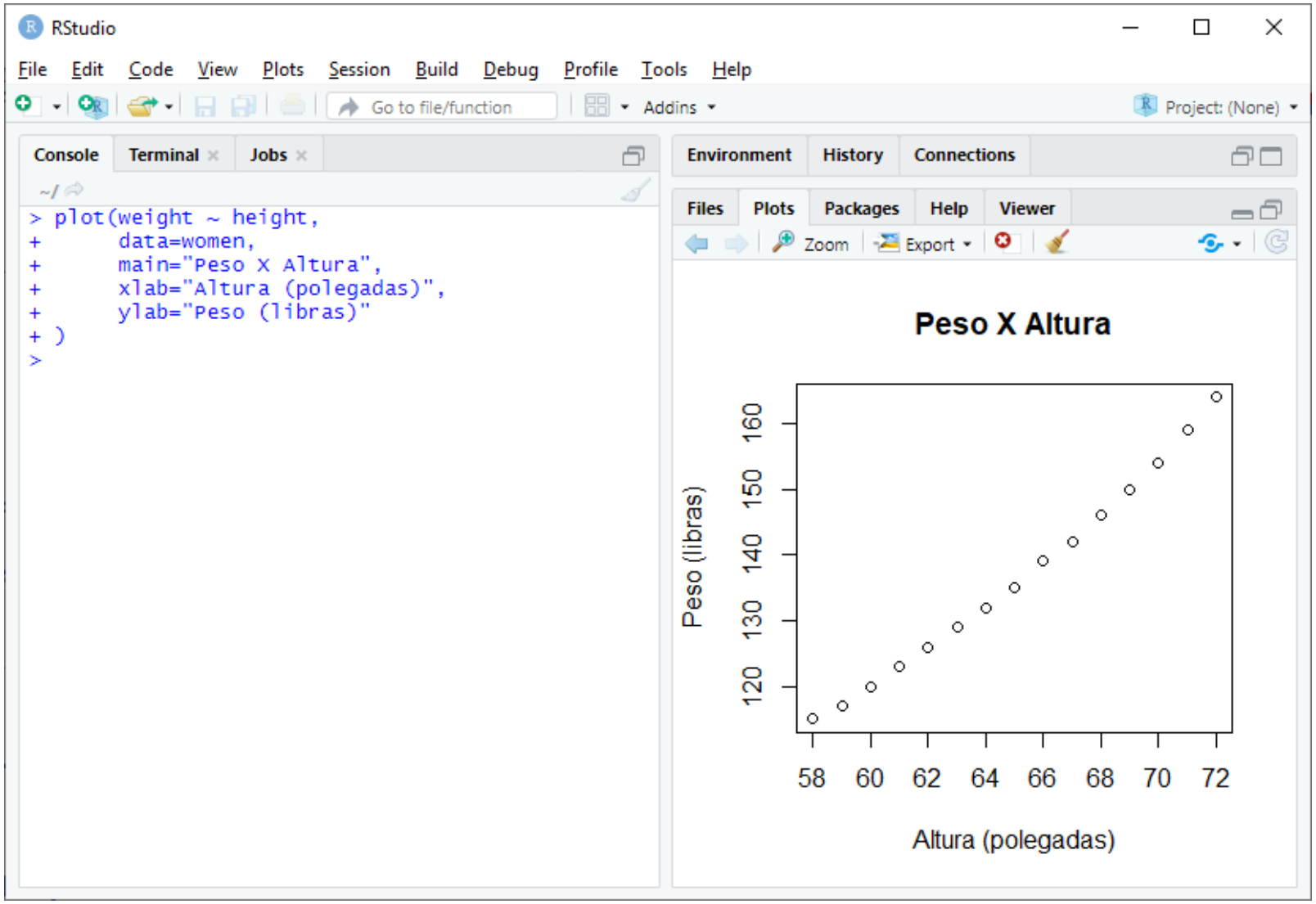


`plot()`

Gráficos de dispersão

O gráfico dispersão é usado para correlacionar duas variáveis quantitativas.

Importante: Correlação não implica em Causalidade. (Duas coisas correlacionadas não implicam, necessariamente, no fato de uma ser causa da outra.)

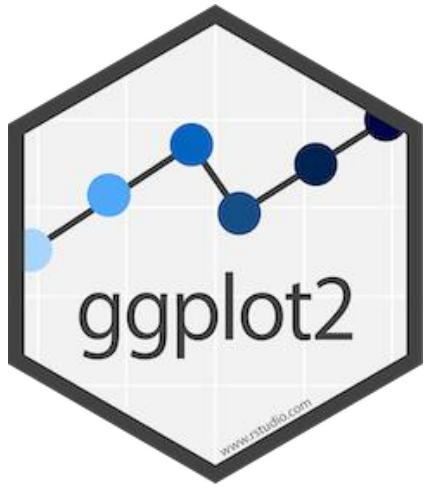


ggplot2

ggplot2

Em 2005, Leland Wilkinson publicou o livro *The Grammar of Graphics*, uma fonte de princípios fundamentais para a construção de gráficos estatísticos. O autor defende que um gráfico é o mapeamento dos dados a partir de atributos estéticos – posição, cor, forma e tamanho – e de objetos geométricos – pontos, linhas, barras e caixas.

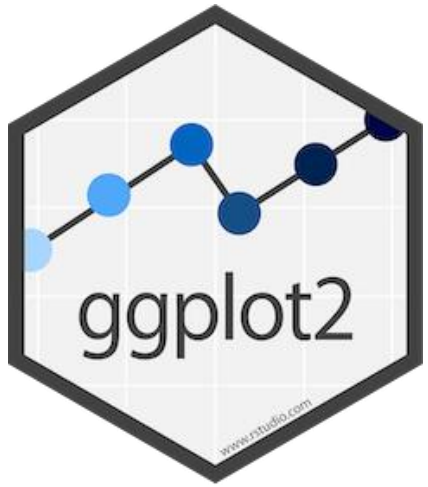
A partir destes princípios, Hadley Wickham escreveu *A Layered Grammar of Graphics*, sugerindo que os principais aspectos de um gráfico – dados, sistema de coordenadas, rótulos e anotações – podiam ser divididos em camadas, construídas uma a uma na elaboração do gráfico. Essa é a essência do **ggplot2**.



ggplot2

Vantagens do ggplot2 em relação aos gráficos básicos do R:

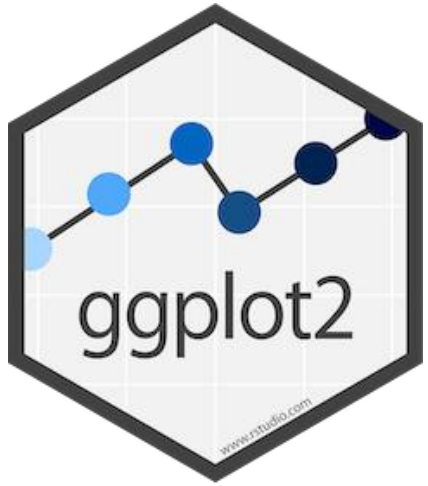
- produz gráficos de melhor aparência;
- facilidade para ajustar o gráfico da forma desejada;
- estrutura padronizada das funções (aprendizado mais intuitivo);
- permite criar uma imensa gama de gráficos com poucas linhas de código.



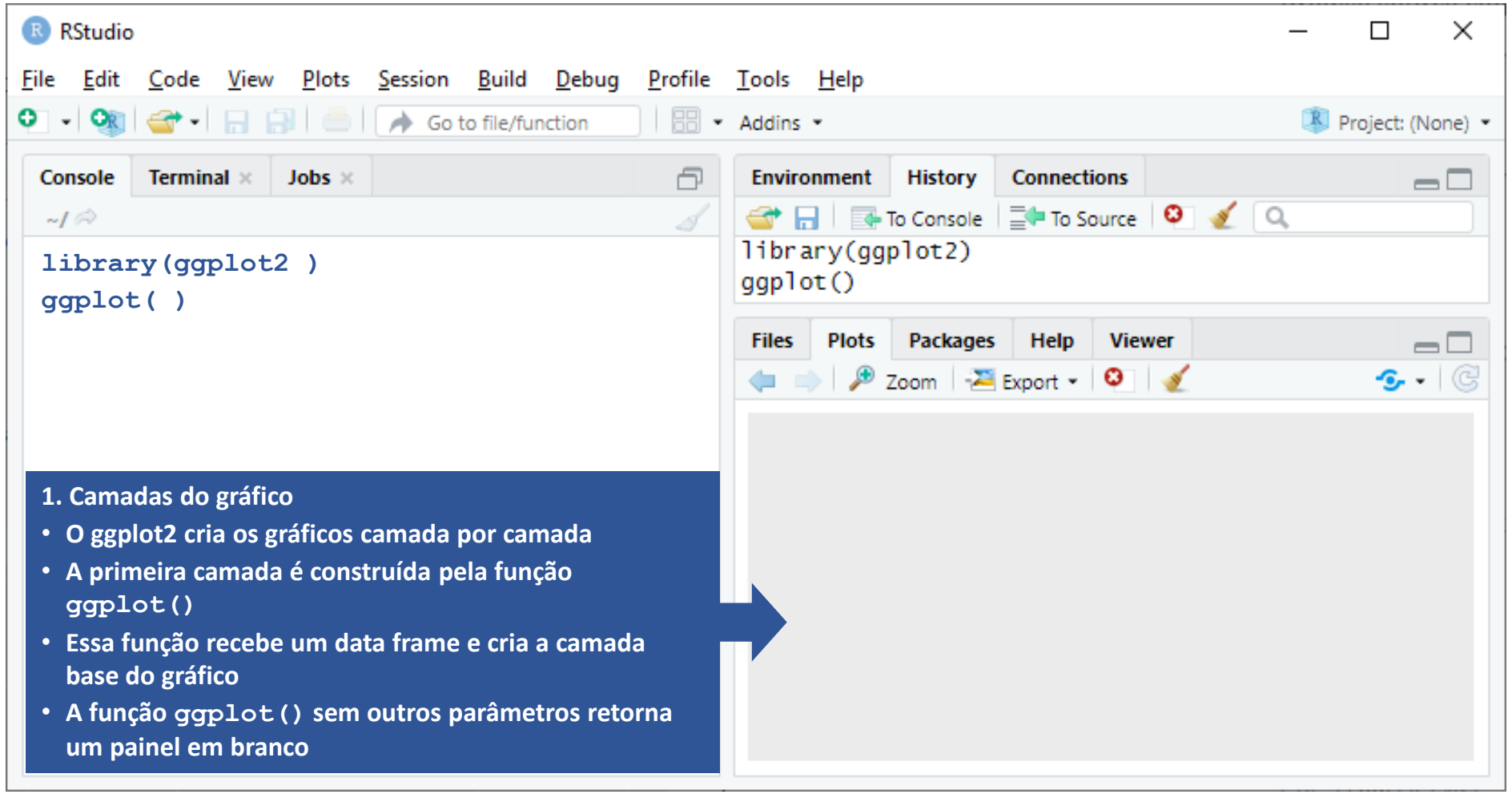
ggplot2

Instalar e carregar o pacote ggplot2:

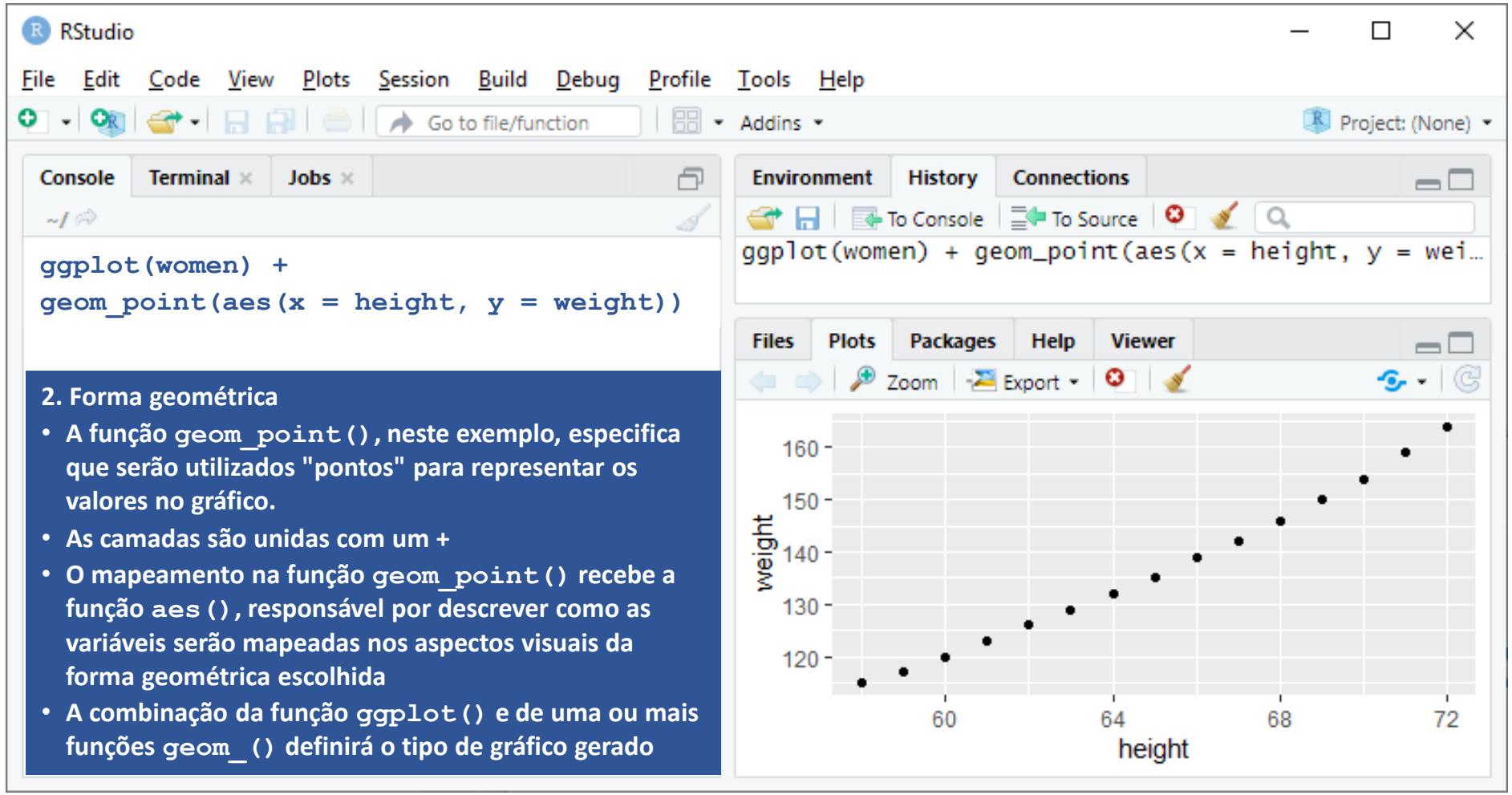
```
install.packages("ggplot2")
library(ggplot2)
```



ggplot2



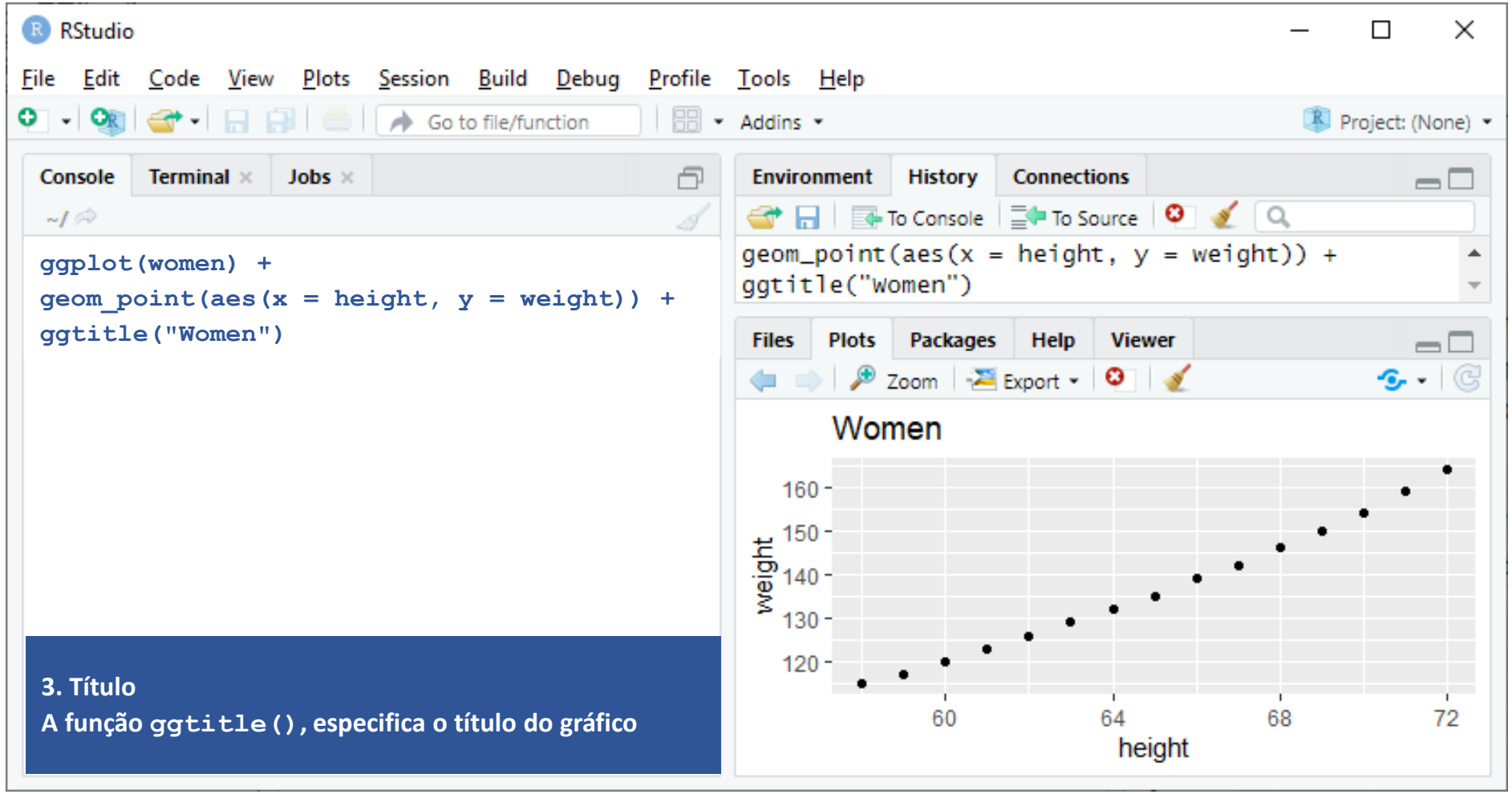
ggplot2



2. Forma geométrica

- A função `geom_point()`, neste exemplo, especifica que serão utilizados "pontos" para representar os valores no gráfico.
- As camadas são unidas com um +
- O mapeamento na função `geom_point()` recebe a função `aes()`, responsável por descrever como as variáveis serão mapeadas nos aspectos visuais da forma geométrica escolhida
- A combinação da função `ggplot()` e de uma ou mais funções `geom_()` definirá o tipo de gráfico gerado

ggplot2



3. Título

A função `ggtitle()`, especifica o título do gráfico

ggplot2

Atributos visuais

A função `aes()` - **aesthetics** - indica a relação entre os dados e cada aspecto visual do gráfico, como qual variável será representada no eixo x, qual será representada no eixo y, a cor e o tamanho dos componentes geométricos, etc.

Os aspectos visuais mais utilizados são:

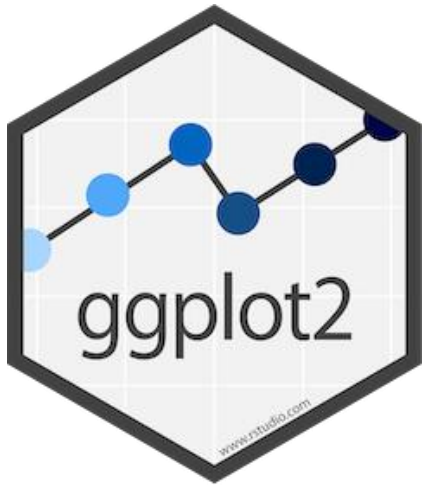
color: altera a cor de formas que não têm área (pontos e retas);

fill: altera a cor de formas com área (barras, caixas, densidades, áreas);

size: altera o tamanho de formas;

type: altera o tipo da forma, geralmente usada para pontos;

linetype: altera o tipo da linha.



ggplot2

Formas geométricas

Os atributos que iniciam com `geom_` definem qual forma geométrica será utilizada para a visualização das observações.

As formas geométrica mais utilizadas são:

geom_point: para gráficos de dispersão transformando pares (x,y) em pontos;

geom_line: para linhas definidas por pares (x,y);

geom_abline: para retas definidas por um intercepto e uma inclinação;

geom_hline: para retas horizontais;

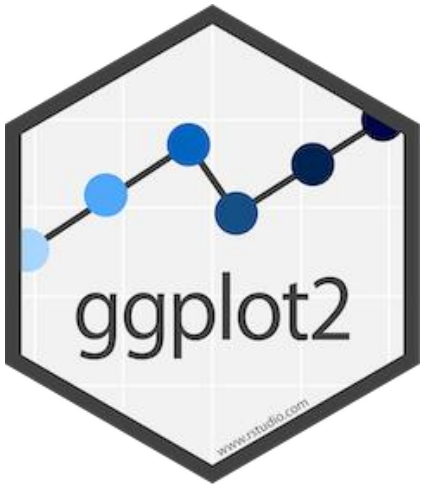
geom_bar: para barras;

geom_histogram: para histogramas;

geom_boxplot: para boxplots;

geom_density: para densidades;

geom_area: para áreas.



Modelos Probabilísticos Discretos

Modelos Probabilísticos

Os **modelos de probabilidade** são construídos para **entender melhor fenômenos aleatórios**.

Conceitos importantes:

- **Espaço amostral** é o conjunto de todos os possíveis resultados do experimento (denotado pela letra grega Ω (ômega).
- Um espaço amostral é denominado **discreto** quando for **finito ou infinito enumerável** e é denominado **contínuo** quando for **infinito, formado por intervalos de números reais**.
- É denominado **evento** qualquer **subconjunto do espaço amostral**.

Modelos Probabilísticos Discretos

Os principais modelos probabilísticos discretos são:

- Uniforme discreto
- Binomial
- Poisson
- Geométrico
- Hipergeométrico

Nas próximas aulas abordaremos dois modelos utilizados com muita frequência:

- Distribuição Binomial
- Distribuição de Poisson

Modelos Probabilísticos Discretos:
Distribuição Binomial

Distribuição binomial

A distribuição binomial é a distribuição de probabilidade discreta do número de sucessos numa sequência de n tentativas que atende aos seguintes requisitos:

- Um número fixo (n) de tentativas;
- Cada tentativa deve ser independente das outras;
- Cada tentativa tem apenas dois resultados possíveis, chamados de "sucesso" (o resultado de interesse) ou "fracasso" (distribuição de Bernoulli).

Há uma probabilidade constante (p) de sucesso para cada tentativa, cujo complemento é a probabilidade (1 - p) de fracasso, muitas vezes representada por q = (1 - p).

$$P(X = x) = \frac{n!}{x!\,(n-x)} . p^x . q^{(n-x)}$$

n: número de tentativas
x: número de sucessos
p: probabilidade de sucesso a cada tentativa

Distribuição binomial

Exemplo: Uma moeda é jogada 5 vezes. Qual é a probabilidade de resultar em cara 3 vezes?

X é uma variável binomial pois cada realização do evento resulta em duas possibilidades: cara ou coroa (sucesso ou fracasso);

Cada lançamento da moeda é independente do lançamento anterior;

Sair cara tem a mesma probabilidade 0.5 (50%) para todos lançamentos.

Identificando os parâmetros na fórmula:

n = 5 - número de tentativas

p = 0.5 - probabilidade de resultar cara

x = 3 - contagem dos sucessos.

$$P(X = x) = \frac{n!}{x!\,(n-x)!} . p^x . q^{(n-x)}$$

Substituindo na fórmula:

$$P(X = x) = \frac{5!}{3!\,(5-3)!} . 0.5^3 . 0.5^{(5-3)} \quad = 0.3125 = 31.25\%$$

Resolvendo no R:

```
> dbinom(3,5,0.5)
[1] 0.3125
```

Modelos Probabilísticos Discretos:
Distribuição de Poisson

Distribuição de Poisson

A distribuição de Poisson (publicada por Siméon Denis Poisson, em 1838) é uma distribuição discreta de probabilidade aplicável a ocorrências de um número de eventos em um intervalo específico.

Para aplicar a distribuição de Poisson, três aspectos devem ser observados:

- O experimento calcula quantas vezes um evento ocorre em um determinado intervalo (tempo, área, volume, etc.);
- A probabilidade do evento ocorrer é a mesma para cada intervalo;
- O número de ocorrências de um intervalo é independente do outro.

A distribuição Poisson tem apenas um parâmetro, **λ** (lambda) que é interpretado como uma taxa média de ocorrência do evento, e a probabilidade de ocorrerem exatamente **x** eventos é dada por:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{n!}$$

onde:

e (número de Euler) ≈ 2,7183

λ > 0

Distribuição de Poisson

Exemplo: Um usuário recebe em média 5 mensagens em seu WhatsApp a cada 15 minutos. Qual a probabilidade desse usuários receber 8 mensagens no mesmo intervalo (15 minutos)?

λ = 5

n = 8

***e* = 2.7183**

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{n!} \Rightarrow P(X = x) = \frac{5^8 e^{-5}}{8!} \Rightarrow 0.065278$$

Resolvendo no R:

```
> dpois(8,5)
[1] 0.06527804 #6.5278%
```

Modelos Probabilísticos Contínuos

Modelos Probabilísticos Contínuos

A distribuição contínua descreve as probabilidades dos possíveis valores de uma variável aleatória contínua.

Uma variável aleatória contínua apresenta um conjunto de valores possíveis (conhecidos como intervalos) que é infinito e incontável.

As probabilidades de variáveis aleatórias contínuas são definidas como a área sob a curva da sua distribuição.

Assim, apenas faixas de valores podem ter uma probabilidade diferente de zero. A probabilidade de que uma variável aleatória contínua seja exatamente igual a algum valor é sempre zero.

Portanto, é possível calcular a probabilidade de que um homem pesa entre 70 e 80 quilos. No entanto, a probabilidade de um homem pesar exatamente 70 quilos para a precisão infinita é zero. Porque a área sob a curva em um único ponto (que não tem nenhuma largura) é zero.

Modelos Probabilísticos Contínuos

Alguns dos principais modelos probabilísticos contínuos são:

- Uniforme Contínuo
- Normal ou Gaussiana
- Qui-quadrado
- t de Student
- F de Snedecor
- Gama
- Beta
- Exponencial
- Weibull
- Logística

Nas próximas aulas abordaremos dois modelos utilizados com muita frequência:

- Distribuição Normal ou Gaussiana
- Distribuição Exponencial

Modelos Probabilísticos Contínuos:
Distribuição Normal ou Gaussiana

Distribuição Normal (Gaussiana)

A distribuição normal é um dos principais modelos de probabilidade para **variáveis aleatórias contínuas**.

Exemplos de variáveis aleatórias contínuas: comprimento, peso, temperatura, pressão, tempo, etc.

Em uma coleta aleatória de dados de fontes independentes, geralmente observa-se que a distribuição dos dados é normal. Ou seja, ao traçar um gráfico com o valor da variável no eixo horizontal **x** e a contagem dos valores no eixo vertical **y**, obtemos uma curva em forma de sino. O **centro da curva** representa a **média** do conjunto de dados. No gráfico, cinquenta por cento dos valores ficam à esquerda da média e os outros cinquenta por cento ficam à direita. Isso é conhecido como distribuição normal.

Variáveis aleatórias com diferentes médias μ e variâncias σ^2 podem ser modeladas pelas funções de densidade de probabilidade normal. A média μ determina o centro do gráfico em forma de sino e a variância σ^2 determina a largura da distribuição.

Distribuição Normal (Gaussiana)

Gerando valores aleatórios para uma distribuição normal:

A função **`rnorm()`** abaixo gera **dois conjuntos diferentes** com 10 valores aleatórios de uma distribuição normal:

```
> rnorm(10)
 [1]  0.3599659 -1.5875803 -0.3890161  0.1307562  0.5890599
 [6] -2.1539719  1.4859364 -0.4329219  0.6385853  1.9210244
> rnorm(10)
 [1] -1.15513692 -2.11647932 -1.13449016  1.61930642 -1.02581667
 [6] -0.45670775 -0.05491379 -0.58710610  0.34518977  1.49852742
```

A função **`set.seed()`** é utilizada para reproduzir os resultados dos geradores de números pseudo-aleatórios. Isto é importante, por exemplo, em simulações ou para ajustar um modelo de classificação com dois subconjuntos dos dados, um para treinar o modelo e outro para o testar. O valor do seed (semente) não é importante desde que a sua utilização seja consistente.

A função **`rnorm()`** abaixo gera **dois conjuntos iguais** com 10 valores aleatórios de uma distribuição normal:

```
> set.seed(128); rnorm(10)
 [1]  0.596771824  0.482611999  1.664405236 -0.025995522  1.209026965
 [6]  0.580352168 -0.458345286  0.004552681  0.209954522  1.511140204
> set.seed(128); rnorm(10)
 [1]  0.596771824  0.482611999  1.664405236 -0.025995522  1.209026965
 [6]  0.580352168 -0.458345286  0.004552681  0.209954522  1.511140204
```

Distribuição Normal (Gaussiana)

R tem quatro funções integradas para gerar distribuição normal:

`dnorm(x,mean,sd)` – função de densidade ou probabilidade

`pnorm(x,mean,sd)` – função de densidade cumulativa

`qnorm(p,mean,sd)` – função quantil

`rnorm(n,mean,sd)` – função de desvios aleatórios

Onde:

x é um vetor de números.

p é um vetor de probabilidades.

n é o número de observações (tamanho da amostra).

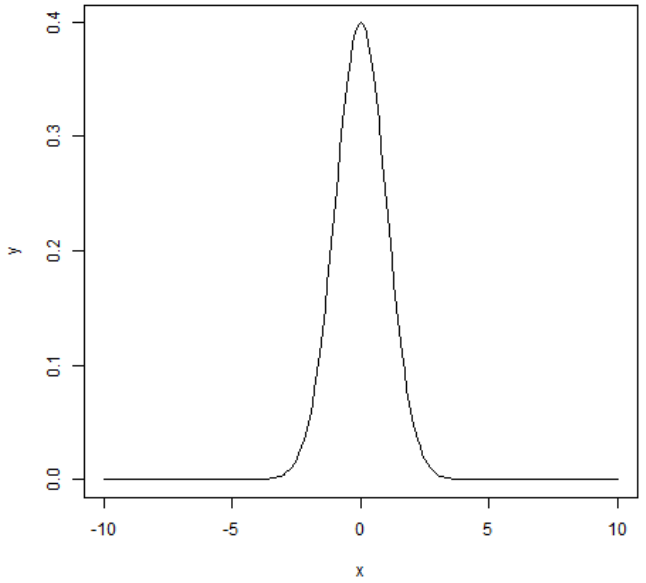
mean é o valor médio dos dados da amostra. O valor padrão é zero.

sd é o desvio padrão. O valor padrão é 1.

Distribuição Normal (Gaussiana)

dnorm() Fornece a altura da distribuição de probabilidade em cada ponto para uma determinada média e desvio padrão.

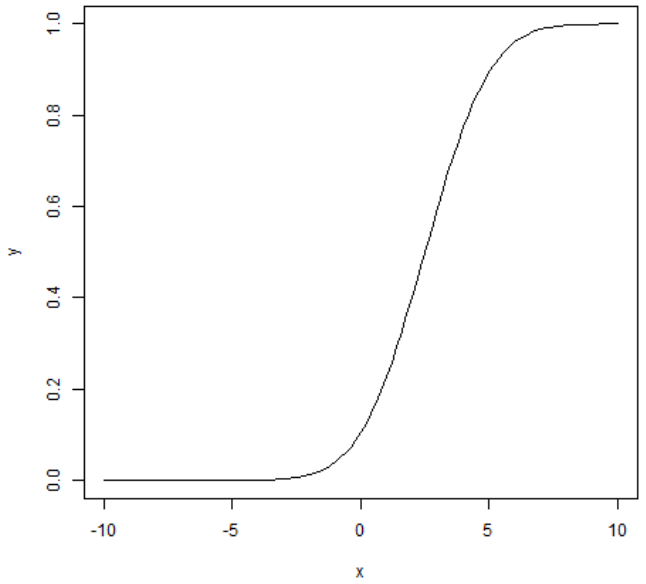
```
# Cria uma sequência de números entre -10 e +10 com incremento 0.1
x <- seq(-10, 10, by = .1)
# Média = 0 e desvio padrão = 1
y <- dnorm(x, mean = 0, sd = 1)
# Gera o gráfico
plot(x,y, type="l")
```



Distribuição Normal (Gaussiana)

pnorm() Fornece a probabilidade de um número aleatório normalmente distribuído ser menor que o valor de um determinado número. (Chamada também de "Função de distribuição cumulativa".)

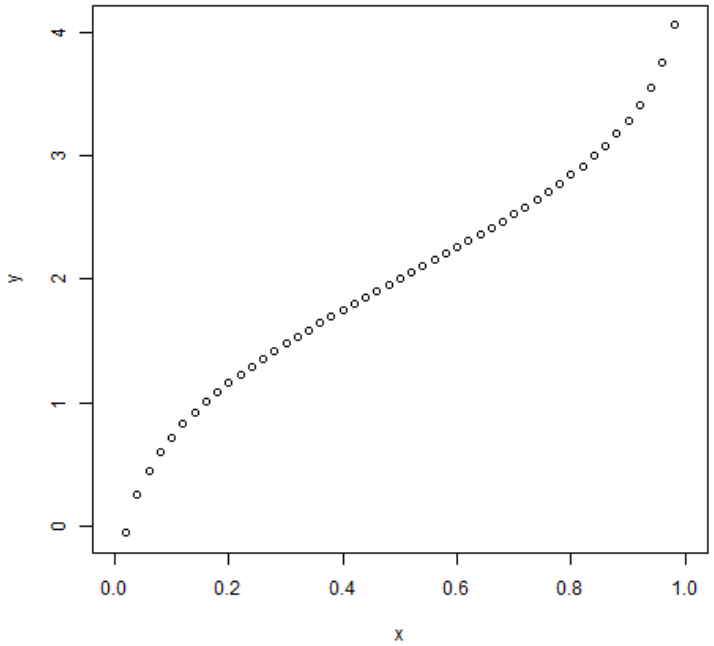
```
# Cria uma sequência de números entre -10 e +10 com incremento 0.2
x <- seq(-10,10,by = .2)
# Média = 2.5 e desvio padrão = 2
y <- pnorm(x, mean = 2.5, sd = 2)
# Gera o gráfico
plot(x,y, type="l")
```



Distribuição Normal (Gaussiana)

qnorm() A função quantil é simplesmente o inverso da função densidade cumulativa (pnorm), isto é, mapeia de probabilidades para valores.

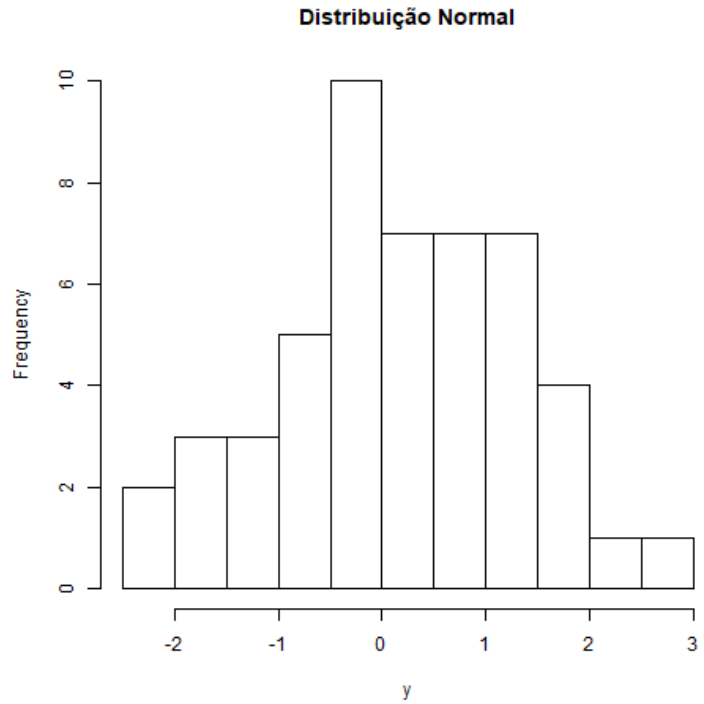
```
# Cria uma sequência de valores de probabilidade com incremento 0.02
x <- seq(0, 1, by = 0.02)
# Média = 2 e desvio padrão = 1
y <- qnorm(x, mean = 2, sd = 1)
# Gera o gráfico
plot(x,y)
```



Distribuição Normal (Gaussiana)

`rnorm()` Gera números aleatórios cuja distribuição é normal. Usa o tamanho da amostra como entrada e gera os números aleatórios. O histograma apresenta a distribuição dos números gerados.

```
# Cria uma amostra de com 50 números normalmente distribuídos

y <- rnorm(50)

# Gera o histograma para amostra criada

hist(y, main = "Distribuição Normal")
```

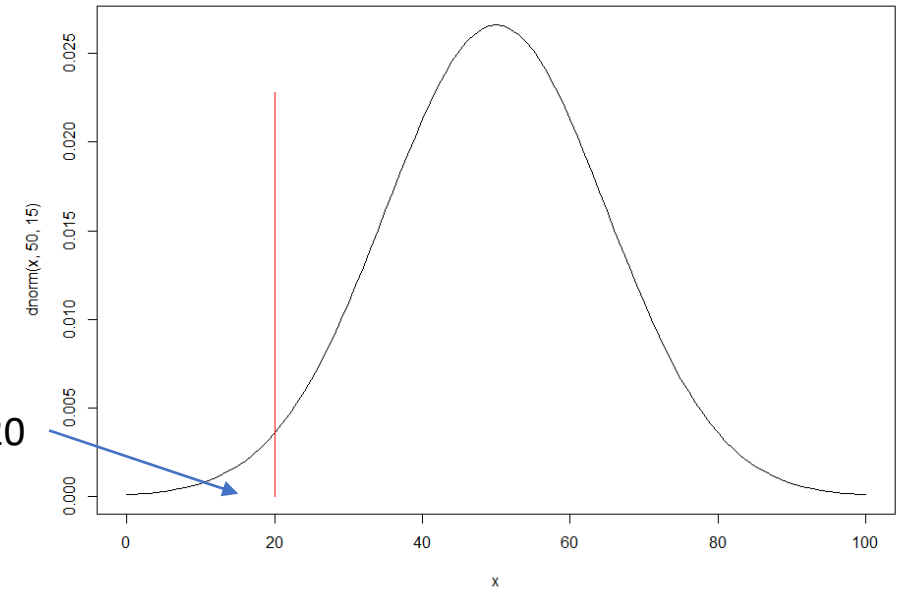


Distribuição Normal (Gaussiana)

Exemplo 1: Qual a probabilidade de ocorrência de um valor **menor** que 20 em uma distribuição normal de média igual a 50 e desvio padrão igual a 15?

```
# x=20; mean=50; sd=15

pnorm(20, 50, 15)

[1] 0.02275013  # ≈ 2.28%

curve(dnorm(x, 50, 15),0,100)

lines(c(20,20),c(0,0.0228),col="red")
```

Probabilidade de valor < 20



Distribuição Normal (Gaussiana)

Exemplo 2: Qual a probabilidade de ocorrência de um valor **maior** que 20 em uma distribuição normal de média igual a 50 e desvio padrão igual a 15?

```
1-pnorm(20, 50, 15)
```

```
[1] 0.9772499 # ≈ 97.72%
```

Exemplo 3: Qual o valor em que a probabilidade de encontrar valores menores que ele seja de 80% em uma distribuição normal de média igual a 50 e desvio padrão igual a 15?

```
qnorm(0.8, 50, 15)
```

```
[1] 62.62432
```

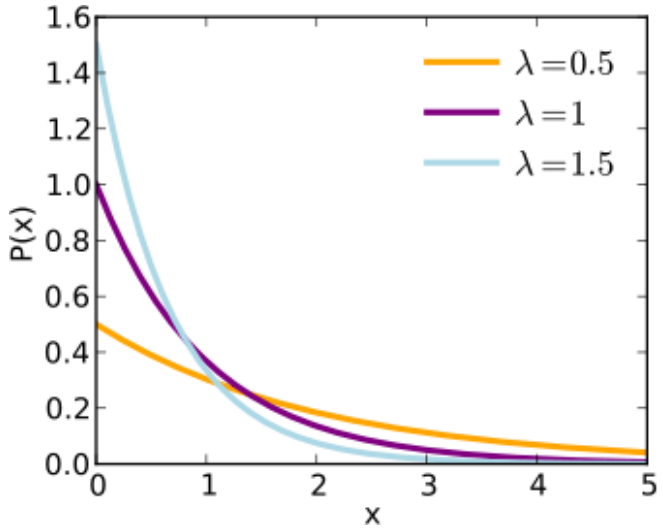
Modelos Probabilísticos Contínuos:
Distribuição Exponencial

Distribuição Exponencial

A distribuição exponencial, muito utilizada em engenharia, descreve o tempo de chegada de uma sequência de eventos independentes recorrentes aleatoriamente.

O modelo ajusta-se bem a dados de tempo para ocorrência de um evento. Exemplo: tempo para atendimento de uma chamada.

É uma distribuição também muito utilizada na prática para modelar tempo de falha de objetos. Por exemplo, pode ser usada para modelar o tempo que demora até uma lâmpada falhar.



Se μ é o tempo médio de espera para a próxima recorrência do evento, sua função de densidade de probabilidade é:

$$f(x) = \begin{cases} \lambda e^{\lambda . x} & se\ x \geq 0 \\ 0 & se\ x < 0 \end{cases}$$

em que λ é o parâmetro de taxa da distribuição e deve satisfazer λ > 0.

Neste caso, λ é o tempo médio de vida e x é um tempo de falha.

O parâmetro deve ter a mesma unidade do tempo da falha x. Isto é, se x é medido em horas, λ também será medido em horas.

Distribuição Exponencial

O tempo médio de checkout de um caixa de supermercado é de três minutos. Qual a probabilidade do checkout de um cliente ser concluído pelo caixa em menos de dois minutos?

Tomando como base o enunciado, verifica-se que a taxa de processamento do checkout é igual a um dividido pelo tempo médio de conclusão do checkout. Portanto, a taxa de processamento é de 1/3 checkouts por minuto. Em seguida, aplica-se a função pexp da distribuição exponencial com taxa = 1/3.

```
> pexp(2, rate=1/3)
[1] 0.4865829
```

A probabilidade de concluir um checkout em menos de dois minutos no caixa é de 48.7%.

A distribuição exponencial é a única distribuição contínua que possui **perda de memória**.

Portanto, suponha que o cliente já esperou um minuto para conclusão do checkout. Isto significa que a probabilidade de ser atendido no próximo minuto seja maior do que no primeiro minuto? A resposta é não. Não importa o quanto tempo o cliente tenha esperado. Esta propriedade da distribuição exponencial é chamada de perda de memória.

Regressão Linear Simples
e
Correlação

Regressão Linear Simples

Regressão linear simples

A regressão linear simples constitui uma tentativa de estabelecer uma equação matemática linear (reta) com apenas uma variável dependente que descreva o relacionamento entre duas variáveis.

A reta de regressão (equação linear) apresenta como principais características:

- **Coeficiente angular** (α) da reta, dado pela tangente da reta;
- **Coeficiente linear** (β), a cota da reta em determinado ponto (o valor de y quando x for igual a zero).

$$y = \alpha + \beta x + \varepsilon$$

Onde:

x: variável independente que busca explicar y

y: variável dependente a ser prevista

ε: erro que corresponde ao desvio entre o valor real e o aproximado (pela reta) de y

Equação linear

A equação linear pode ser obtida no R através da função **`lm()`** que calcula a regressão linear simples.

A regressão linear simples é obtida por meio do seguinte comando:

```
lm(y~x,data)
```

lm: linear model

y~x: y depende de x

data: conjunto de dados (data.frame) valores das "colunas" que correspondem a **x** e **y**

Equação linear

Exemplo:

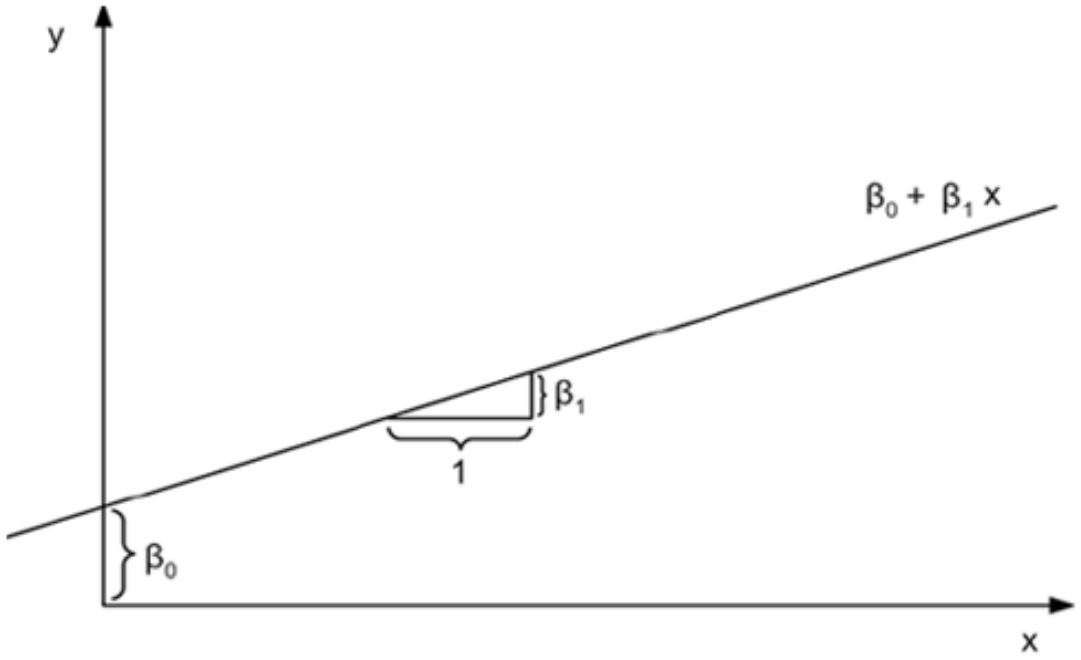
```
> x <- c(205,225,305,380,560,600,685,735)
> y <- c(15,20,21,22,22,25,28,28)
> dados <- data.frame(x,y) #cria um data.frame
> regressão <- lm(y~x,data=dados) #ou regressão <- lm(y~x)
> regressão

Call:
lm(formula = y ~ x, data = dados)

Coefficients:
(Intercept)            x
   13.85362      0.01899
```

Equação da reta: $y = 13.85362 + 0.01899x$

Equação linear

O parâmetro β_0 é chamado intercepto ou coeficiente linear e representa o ponto em que a reta corta o eixo dos y's, quando $x = 0$.

O parâmetro β_1 representa a inclinação da reta e é chamado de coeficiente de regressão ou coeficiente angular.



Equação linear

regressão é uma lista com os seguintes elementos:

```
> names(regressão)
[1] "coefficients"  "residuals"     "effects"     "rank"
[5] "fitted.values" "assign"        "qr"          "df.residual"
[9] "xlevels"       "call"          "terms"       "model"
```

Valores preditos da variável resposta para cada elemento da amostra (previsão):

```
> regressão$fitted.values
       1        2        3        4        5        6        7        8
17.74673 18.12655 19.64582 21.07013 24.48847 25.24811 26.86233 27.81187
```

Erro ou resíduos (valor observado - valor predito) para cada ponto da amostra:

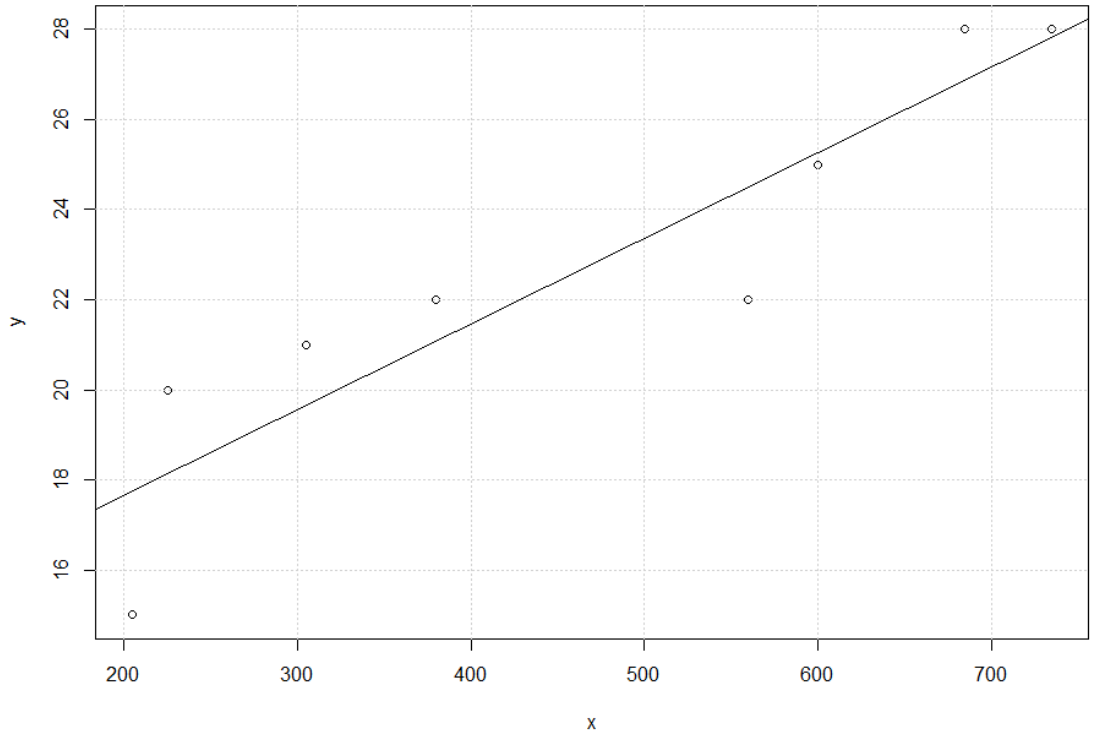
```
> regressão$residuals
 1          2          3          4          5          6          7          8
-2.7467348  1.8734490  1.3541839  0.9298729 -2.4884736 -0.2481061  1.1376747  0.1881341
```

Estimativa dos coeficientes da regressão:

```
> regressão$coefficients
(Intercept)           x
13.85361800  0.01899081
```

Equação linear

```
z = plot(x,y)
grid(z) #aplicando grid ao gráfico
abline(regressão)
```



Correlação

Coeficiente de correlação (r)

O gráfico de dispersão pode indicar se a correlação linear é positiva, negativa ou a inexistência de correlação.

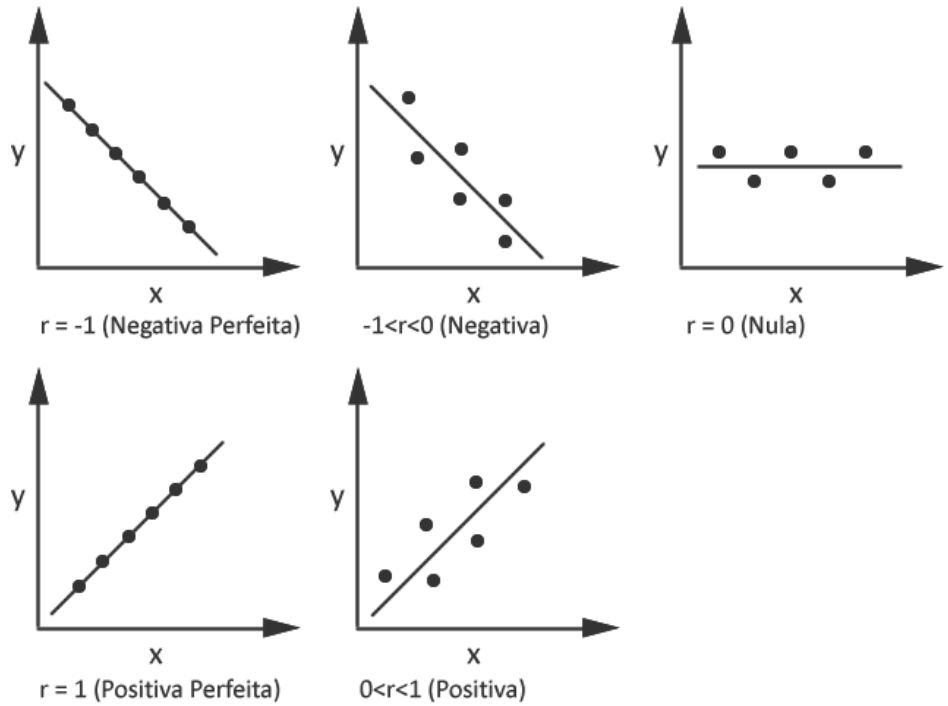
Através do coeficiente de correlação pode-se identificar o tipo de correlação, conforme a seguinte tabela:

r	Tipo	Descrição
r = -1	**Perfeita Negativa**	Os pontos estão perfeitamente alinhados, porém os valores crescentes de x associados a valores decrescentes de y
-1 < r < 0	**Negativa**	Os valores crescentes de x estão associados a valores decrescentes de y
r = 0	**Não há correlação**	Os valores de x e y ocorrem independentemente
0 < r < 1	**Positiva**	Os valores crescentes de x estão associados a valores crescentes de y
r = 1	**Positiva Perfeita**	Os pontos estão perfeitamente alinhados e os valores crescentes de x estão associados a valores crescentes de y

Coeficiente de correlação linear (r)

O coeficiente de correlação demonstra que a correlação será tanto mais forte quanto mais próximo o coeficiente estiver de –1 ou +1, e será tanto mais fraca quanto mais próximo o coeficiente estiver de zero.

r	Tipo
-1	Negativa perfeita
-1 < r < -0.7	Negativa forte
-0.7 < r -0.5	Negativa moderada
-0.5 < r < 0	Negativa fraca
0	Nula
0 < r < 0.5	Positiva fraca
0.5 < r < 0.7	Positiva moderada
0.7 < r < 1	Positiva forte
r = 1	Positiva perfeita



Coeficiente de correlação linear (r)

Exemplo:

```
> x <- c(205,225,305,380,560,600,685,735)
> y <- c(15,20,21,22,22,25,28,28)
> cor(x,y)
[1] 0.9155359
```

r = +0.916 indica uma correlação **positiva forte**, ou seja, a variável dependente **y** cresce quase na mesma proporção que a variável independente **x**.

Coeficiente de determinação (r^2)

O coeficiente de determinação explica o grau de ajuste do modelo, isto é, o percentual de variação de y que é explicado pela variabilidade de x. Seu valor varia de 0 a 1.

O valor do coeficiente de determinação corresponde ao valor do coeficiente de variação elevado ao quadrado.

Exemplo:

```
> x <- c(205,225,305,380,560,600,685,735)
> y <- c(15,20,21,22,22,25,28,28)
> cor(x,y)^2
[1] 0.8382059 #aproximadamente 84%
```

84% da variação da variável dependente **y** é explicada pela variável independente **x**. Os outros 16% possuem causas aleatórias desconhecidas (independentes de **x**).

Coeficiente de correlação linear (r)

```
> x <- c(-1,2,3,4)
> y <- c(-1,5,7,9)
> cor(x,y)
[1] 1
```

r = +1 indica uma **correlação perfeita positiva** entre as duas variáveis **x** e **y**.

$$y = \alpha + \beta x + \varepsilon$$

$$\beta = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\alpha = \frac{\sum y - \beta \sum x}{n}$$

$$n = 4$$

	x	y	xy	x^2	y^2
	-1	-1	1	1	1
	2	5	10	4	25
	3	7	21	9	49
	4	9	36	16	81
Σ	8	20	68	30	156

$$\beta = \frac{4 \cdot 68 - 8 \cdot 20}{4 \cdot 30 - (8)^2} = \frac{112}{56} = 2$$

$$\alpha = \frac{20 - 2 \cdot 8}{4} = \frac{4}{4} = 1$$

$$y = 1 + 2x + 0$$

Testes de Hipóteses

Testes de Hipóteses

Usar dados de uma amostra para tentar dizer algo sobre a população que lhes deu origem é um importante ramo da estatística, chamado **inferência estatística**, e o teste de hipótese é uma das suas principais ferramentas.

Um **teste de hipótese** é uma técnica estatística cujo **objetivo é verificar se uma dada amostra de dados é, ou não, compatível com uma hipótese feita sobre a população que lhe deu origem**.

O teste de hipótese coloca lado-a-lado duas hipóteses sobre a população que deu origem à amostra de dados: uma hipótese inicial, ou hipótese nula, e uma hipótese alternativa, designadas por $\mathbf{H_0}$ e $\mathbf{H_1}$ respectivamente.

É retirada uma amostra da população, cuja informação será tratada para encontrar evidência para se rejeitar, ou não a hipótese nula.

Caso ocorra a rejeição da hipótese nula, deve-se considerar, à partir desta constatação, a hipótese alternativa.

Teste t de Student

Teste t de Student

O test t de Student foi introduzido em 1908 por William Sealy Gosset, matemático e estatístico que trabalhava na cervejaria Guinness, em Dublin, na Irlanda.

Gosset desenvolveu o teste t, a princípio, para monitorar a qualidade da cerveja.

O teste t pode ser aplicado quando o tamanho da amostra é pequeno. Isso permite que se façam inferências usando um menor número de elementos, reduzindo os custos da pesquisa.

O uso de métodos estatísticos na fabricação da cerveja era considerado um segredo industrial. Portanto, quando Gosset publicou o artigo sobre o teste t na revista acadêmica Biometrika em 1908, teve de usar um pseudônimo "Student" e, por isso, o teste t passou a ser conhecido como **teste t de Student**.

Teste t de Student

A pluviosidade média em uma região é: 70.6

O vetor pluv apresenta mês-a-mês a pluviosidade em um determinado ano com relação a uma determinada região:

```
pluv = c(110,100,60,80,70,18,17,17,42,89,108,143)
```

Observando-se que a média mensal de pluviosidade é de 70.6 mm, terá o ano "pluv" sido excecionalmente chuvoso?

A hipótese nula contém sempre uma igualdade e a hipótese alternativa contém sempre uma desigualdade.

Testes com a alternativa diferente são denominados bilaterais. Testes com alternativa menor são designados unilaterais esquerdos, e com alternativa maior são designados unilaterais direitos.

As hipóteses podem ser formalmente descritas assim:

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0 \Rightarrow H_0 : \mu = 70.6 \text{ vs } H_1 : \mu \neq 70.6 \quad \text{(Bilateral)}$$

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu < \mu_0 \Rightarrow H_0 : \mu = 70.6 \text{ vs } H_1 : \mu < 70.6 \quad \text{(Unilateral esquerdo)}$$

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu > \mu_0 \Rightarrow H_0 : \mu = 70.6 \text{ vs } H_1 : \mu > 70.6 \quad \text{(Unilateral direito)}$$

Teste t de Student

Utilizando o R para resolver $H_0 : \mu = 70.6$ vs $H_1 : \mu \neq 70.6$ (teste bilateral)

```
t.test(pluv, mu = 70.6)
```

```
data:  pluv
t = 0.047327, df = 11, p-value = 0.9631
alternative hypothesis: true mean is not equal to 70.6
95 percent confidence interval:
 44.81350 97.51983
sample estimates:
mean of x
 71.16667
```

O **p-value** (valor de prova) de **96.3%** indica "não rejeição" de **H_0**, ou seja, não há razão para rejeitar a hipótese de que o ano "pluv" tenha sido um ano de pluviosidade normal.

Os limites entre não rejeição e rejeição situa-se entre 1% e 10% de **p-value**. Ou seja, para p-values menores que 1% rejeita-se a hipótese nula. Para valores maiores que 10% não se costuma rejeitar. O intervalo entre 1 e 10% é uma "zona cinzenta", pois o julgamento fica à critério do pesquisador.

Teste t de Student

Utilizando o R para resolver $H_0 : \mu = 70.6$ vs $H_1 : \mu < 70.6$ (teste unilateral esquerdo)

```
t.test(pluv, mu = 70.6, alternative = "less")

data:  pluv
t = 0.047327, df = 11, p-value = 0.5184
alternative hypothesis: true mean is less than 70.6
95 percent confidence interval:
     -Inf 92.66942
sample estimates:
mean of x
 71.16667
```

O **p-value** (valor de prova) de **51.8%** indica "não rejeição" de **H_0**.

Teste t de Student

Utilizando o R para resolver $H_0 : \mu = 70.6$ vs $H_1 : \mu > 70.6$ (teste unilateral direito)

```
t.test(pluv, mu = 70.6, alternative = "greater")
```

```
data:  pluv
t = 0.047327, df = 11, p-value = 0.4816
alternative hypothesis: true mean is greater than 70.6
95 percent confidence interval:
 49.66391      Inf
sample estimates:
mean of x
 71.16667
```

O **p-value** (valor de prova) de **48.2%** indica "não rejeição" de **H_0**.

Teste de Shapiro-Wilk

Existem alguns pressupostos teóricos que devem ser observados para que o resultado (p-value) seja confiável.

Para amostras pequenas, $N \leq 30$, a população em estudo deve apresentar uma distribuição normal (gaussiana). Distribuição normal é uma distribuição contínua, frequentemente associada a características físicas (peso, altura, temperatura, etc.).

Para amostras grandes, $N > 30$, não há restrições em relação à população.

A amostra "pluv" é pequena ($N = 12$) e para verificar se é uma distribuição normal deve-se realizar um teste de hipóteses preliminar, o teste de Shapiro-Wilk, disponível no R.

Teste de Shapiro-Wilk

O comando R para executar o teste Shapiro-Wilk é **shapiro.test**:

```
> shapiro.test(pluv)

  Shapiro-Wilk normality test

data:  pluv
W = 0.93775, p-value = 0.4694
```

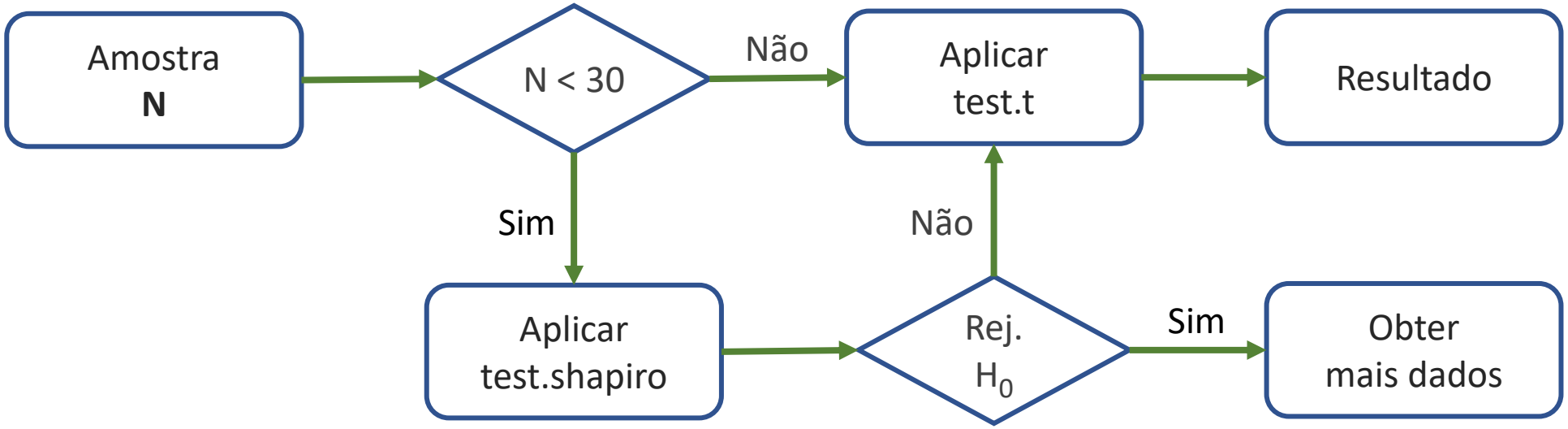
No teste Shapiro-Wilk a hipótese nula é: "a distribuição 'pluv' é normal" e a alternativa é "a distribuição 'pluv' não é normal".

O p-value 46.9% indica que não há evidência para rejeitar a hipótese nula, concluindo-se, portanto, que "pluv" provém de uma distribuição normal.

Desta forma os resultados obtidos anteriormente são válidos.

Caso o p-value do teste Shapiro-Wilk apresentasse valores baixos, (por exemplo, < 1%) seria necessário obter mais dados de "pluv" para chegar a uma amostra com pelo menos 30 valores.

Teste de Shapiro-Wilk



ANOVA - Análise de Variância

A análise de variância (ANOVA - analysis of variance) é um **teste de hipóteses aplicado para comparar médias de mais de duas amostras**. A técnica foi desenvolvida inicialmente para fins agrícolas.

Há diversos tipos de ANOVA. A seguir será aplicado a ANOVA de um fator de efeitos fixos (one-way fixed effects analisys of variance).

Pressupostos a serem observados:

1. Amostras devem ser aleatórias e mutuamente independentes;
2. Populações normalmente distribuída (aplicar o teste de Shapiro-Wilk para $N < 30$);
3. Populações apresentam variância homogênea (aplicar o teste de Bartlett).

ANOVA - Análise de Variância

Três áreas agrícolas foram submetidas a três tratamentos diferentes com adubos identificados por A, B e C. Foram recolhidas 5 colheitas de cada área, obtendo-se as seguintes produtividades por hectare para o mesmo produto:

```
A <- c(14,13,20,15,13)
B <- c(13,14,13,18,15)
C <- c(19,16,17,20,19)
```

Média dos três grupos:

```
> mean(A)
[1] 15
> mean(B)
[1] 14.6
> mean(C)
[1] 18.2
```

A diferença entre as médias amostrais é estatisticamente relevante? ou (em linguagem matemática): Os três grupos têm a mesma média de população, ou pelo menos um grupo tem média de população diferente dos outros?

ANOVA - Análise de Variância

1. **Amostras devem ser aleatórias e mutuamente independentes.**

Deve-se confirmar que não há interferência de um grupo no outro. Por exemplo: se as áreas fossem contíguas e a fertilização ocorresse por polinização, poderia haver mistura genética entre os grupos, invalidando a análise.

O pressuposto de independência não pode ser verificado por testes estatísticos preliminares.

No exemplo, assume-se que não há problemas de contaminação de uma área para outra e, portanto há independência.

ANOVA - Análise de Variância

2. **Populações normalmente distribuída (aplicar o teste de Shapiro-Wilk para N < 30)**

```
> shapiro.test(A)
  Shapiro-Wilk normality test
data:  A
W = 0.77559, p-value = 0.0505

> shapiro.test(B)
  Shapiro-Wilk normality test
data:  B
W = 0.84215, p-value = 0.171

> shapiro.test(C)
  Shapiro-Wilk normality test
data:  C
W = 0.91367, p-value = 0.4899
```

O grupo A (p-value 5.05%) está no borderline da normalidade, mas ainda assim pode ser considerado como seguindo uma distribuição normal.

ANOVA - Análise de Variância

3. Populações têm mesma variância (aplicar o teste de Bartlett).

No teste de Bartlett aceitar a hipótese H_0 indica que os grupos têm variância homogênea. Aceitar H_1 indica que não há variância homogênea (pelo menos um grupo não mantêm variância homogênea em relação aos outros).

```
> bartlett.test(list(A,B,C))

  Bartlett test of homogeneity of variances

data:  list(A, B, C)
Bartlett's K-squared = 1.2051, df = 2, p-value = 0.5474
```

O p-value 54.7% indica não rejeição de H_0, validando o pressuposto da homogeneidade da variância.

Validados os três pressupostos, pode-se preparar e aplicar, a seguir, o **teste ANOVA de um fator**.

ANOVA - Análise de Variância

Formatação dos dados

Para aplicar o teste ANOVA no R é preciso agregar os dados de produtividade das três áreas, A, B e C, em um único vetor.

```
> prod <- c(A,B,C)
> prod
 [1] 14 13 20 15 13 13 14 13 18 15 19 16 17 20 19
```

A seguir, cria-se o vetor com os elementos que correspondem aos grupos (áreas agrícolas).

```
> grupos <- c(rep("A",5),rep("B",5),rep("C",5))
> grupos
 [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "C" "C" "C" "C" "C"
```

O próximo passo será juntar os dois vetores em um único data frame.

ANOVA - Análise de Variância

Formatação dos dados (continuação)

O data frame apresentado a seguir junta os dois vetores: **prod** e **grupos**.

```
> df <- data.frame(prod,grupos)
> df
   prod grupos
1    14      A
2    13      A
3    20      A
4    15      A
5    13      A
6    13      B
7    14      B
8    13      B
9    18      B
10   15      B
11   19      C
12   16      C
13   17      C
14   20      C
15   19      C
```

Esta visualização é importante para verificar se os dados foram inseridos corretamente.

ANOVA - Análise de Variância

Aplicando o teste ANOVA

O comando a seguir executa o ANOVA sobre um modelo linear (**lm**) em que cada valor de produtividade (**prod**) está associado ao respectivo grupo (**grupos**).

```
> anova(lm(prod ~ grupos))
Analysis of Variance Table

Response: prod
          Df Sum Sq Mean Sq F value  Pr(>F)
grupos     2 38.933 19.4667  3.7677 0.05372 .
Residuals 12 62.000  5.1667
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

O p-value do ANOVA é o valor da coluna Pr(>F) = 0.05372, aproximadamente 5.3%.
É um valor no limite da rejeição de H_0, isto é, existe a suspeita de que pelo menos um tratamento tenha gerado diferença de produtividade.
Qual ou quais grupos são responsáveis pelas diferenças?
Um teste de comparação múltipla **post-hoc** tentará identificar o(s) grupo(s) responsáveis pela "quase" rejeição de H_0 no ANOVA.

Tukey HSD - Teste de comparações múltiplas

Aplicando o teste Tukey HSD

O R oferece recursos para realizar a maioria dos testes de comparações múltiplas. Um dos mais utilizados é o Tukey HSD (Honest Significant Difference).

O TukeyHSD atua sobre objetos da classe aov (Analysis Of Variance), portanto é preciso usá-la no comando a seguir:

```
> TukeyHSD(aov(lm(prod ~ grupos)))
 Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = lm(prod ~ grupos))

$grupos
    diff        lwr      upr     p adj
B-A -0.4 -4.2352956 3.435296 0.9583671
C-A  3.2 -0.6352956 7.035296 0.1068512
C-B  3.6 -0.2352956 7.435296 0.0665354
```

O p-value na comparação entre os grupos B-A (95.8%) indica clara não rejeição e H_0.
O p-value na comparação entre os grupos C-A (10.7%) não é suficientemente forte para rejeição de H_0.
O p-value na comparação entre os grupos C-B (6.7%) está no limite para não rejeição de H_0.
Se houve um tratamento que impactou na produtividade este foi aplicado no grupo C. O p-value próximo do limite de rejeição indica que é recomendável recolher mais amostras para testes.

Contato:

Prof. MSc. Marcos Alexandruk

E-mail: alexandruk@uninove.br