



GADES

DATA ANALYSIS SOLUTIONS

Data Science com R

em parceria com o Centro de Estatística e Aplicações
Universidade Lisboa - CEAUL

Ricardo São João IPSantarém & CEAUL

ricardo.sjoao@esg.ipsantarem.pt

Abril, 2022

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

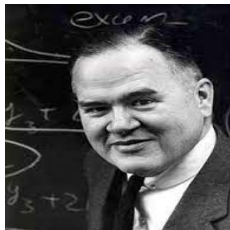
R for Data
Science
Program

- 1 *Data Science*
- 2 *Software livre*
- 3 *Software open source*
- 4 Gráficos
- 5 Documentos dinâmicos
- 6 Requisitos
- 7 Ambiente
- 8 Aplicações

O que é ?

Não existe uma definição formal e consensual relativamente a *Data Science*/Ciência dos Dados.

Já em 1962 (há 60 anos) o matemático/estatístico John Tukey aborda a questão na publicação intitulada *The future of data analysis*¹



fonte: <https://4.bp.blogspot.com>

¹Tukey, J. W. (1962). The future of data analysis. The annals of mathematical statistics, 33(1), 1-67.

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program





❖ *Data Science* é referido como um campo multidisciplinar (matemática, estatística, informática, engenharia) onde são utilizadas metodologias distintas com o objetivo de **extrair valor (informação) dos dados**.

❖ Nas últimas duas décadas a Ciência dos Dados tem assistido a um forte crescimento e protagonismo sendo os dados atualmente apelidados de **“o novo petróleo”**.

★ A profissão “Cientista de Dados” é das mais procuradas e igualmente bem remunerada.

50 Best Jobs in America for 2022

Best Jobs | 2022 | United States

Share |    

	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Enterprise Architect	\$144,997	4.1/5	14,021	View Jobs
#2	Full Stack Engineer	\$101,794	4.3/5	11,252	View Jobs
#3	Data Scientist	\$120,000	4.1/5	10,071	View Jobs
#4	Devops Engineer	\$120,095	4.2/5	8,548	View Jobs

fonte: <https://www.glassdoor.com> acessado em 07/04/2022

Alguns números que dão que pensar...

- o volume de dados na internet no final de 2020 foi estimado em 44 zettabytes (1 zettabyte= 1.000.000.000.000.000.000= 10^{21});
- até 2025, espera-se que a quantidade de dados gerados a cada dia atinja 463 exabytes (1 exabyte= 10^{18});

²fonte: <https://seedscientific.com> acessado em 07/04/2022

Alguns números que dão que pensar...

✿ o volume de dados na internet no final de 2020 foi estimado em 44 zettabytes (1 zettabyte= 1.000.000.000.000.000.000= 10^{21});

✿ até 2025, espera-se que a quantidade de dados gerados a cada dia atinja 463 exabytes (1 exabyte= 10^{18});

entretanto já me “perdi” com tantos zeros

²fonte: <https://seedscientific.com> acessado em 07/04/2022

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Alguns números que dão que pensar...

✿ o volume de dados na internet no final de 2020 foi estimado em 44 zettabytes (1 zettabyte= 1.000.000.000.000.000.000= 10^{21});

✿ até 2025, espera-se que a quantidade de dados gerados a cada dia atinja 463 exabytes (1 exabyte= 10^{18});

entretanto já me “perdi” com tantos zeros

✿ Google, Facebook, Microsoft e Amazon armazenam pelo menos 1.200 petabytes de informação (1 petabyte= 10^{15});

✿ a cada minuto são gastos 1 milhão de dólares em compras na WWW;

✿ em 2030, nove em cada dez pessoas com idade ≥ 6 anos será digitalmente ativa.

²fonte: <https://seedscientific.com> acessado em 07/04/2022

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Porém a análise de tal volume de informação exige recursos/meios, onde o *software* seguramente é um deles.

A escolha pode recair em softwares comerciais ou ...

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Porém a análise de tal volume de informação exige recursos/meios, onde o *software* seguramente é um deles.

A escolha pode recair em softwares comerciais ou ...

- livres;

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Porém a análise de tal volume de informação exige recursos/meios, onde o *software* seguramente é um deles.

A escolha pode recair em softwares comerciais ou ...

- livres;
- *open source*;

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Porém a análise de tal volume de informação exige recursos/meios, onde o *software* seguramente é um deles.

A escolha pode recair em softwares comerciais ou ...

- livres;
- *open source*;
- gratuitos

Software livre: o quê significa ?

O termo “livre” não está associado à ideia de não pagamento na aquisição de *software*.



fonte: <https://foreignpolicyi.org>



fonte: <https://poupaeganha.pt>

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Não se trata de gratuidade: “de graça” ou de “borla”

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Não se trata de gratuidade: “de graça” ou de “borla”

Estaremos então a falar de
que conceito ?

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program



fonte:<http://apoesiadaisamar.blogspot.com/>

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program



fonte: <http://apoesiadaisamar.blogspot.com/>

Significa que os seus
utilizadores possuem
“liberdade” (são livres).

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program



fonte:<http://apoesiadaisamar.blogspot.com/>

Significa que os seus
utilizadores possuem
“liberdade” (são livres).

A noção de liberdade assenta em
quatro pilares essenciais:

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program



fonte:<http://apoesiadaisamar.blogspot.com/>

Significa que os seus
utilizadores possuem
“liberdade” (são livres).

A noção de liberdade assenta em
quatro pilares essenciais:

- liberdade na execução do
software/programa;

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

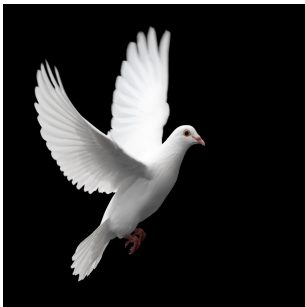
Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program



fonte: <http://apoesiadaisamar.blogspot.com/>

Significa que os seus
utilizadores possuem
“liberdade” (são livres).

A noção de liberdade assenta em
quatro pilares essenciais:

- liberdade na execução do *software*/programa;
- liberdade no desenvolvimento e alteração do código fonte;

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

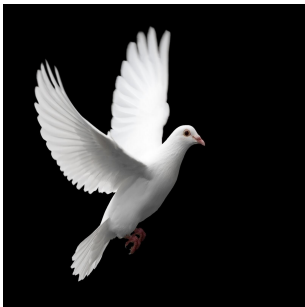
Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program



fonte:<http://apoesiadaisamar.blogspot.com/>

Significa que os seus
utilizadores possuem
“liberdade” (são livres).

A noção de liberdade assenta em
quatro pilares essenciais:

- liberdade na execução do *software*/programa;
- liberdade no desenvolvimento e alteração do código fonte;
- liberdade na redistribuição de cópias;

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

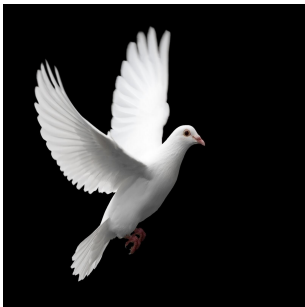
Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program



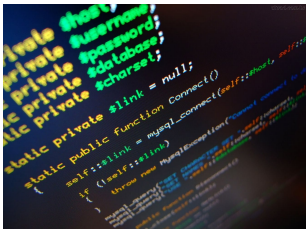
fonte:<http://apoesiadaisamar.blogspot.com/>

Significa que os seus
utilizadores possuem
“liberdade” (são livres).

A noção de liberdade assenta em
quatro pilares essenciais:

- liberdade na execução do *software*/programa;
- liberdade no desenvolvimento e alteração do código fonte;
- liberdade na redistribuição de cópias;
- liberdade na redistribuição de versões modificadas.

O termo *“open source”* ou *“código aberto”* é o termo utilizado para *softwares* que disponibilizam o seu código fonte para edição.



fonte:<https://www.linkoficial.com.br/>



<https://medium.com/>

Ricardo São
João

Data Science

Software livre

**Software open
source**

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Filosofia

A filosofia do *software* de código aberto visa a colaboração entre os seus utilizadores. Adicionalmente grande parte do *software* código aberto disponibiliza o código-fonte de forma gratuita.

Ricardo São
João

Data Science

Software livre

**Software open
source**

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Filosofia

A filosofia do *software* de código aberto visa a colaboração entre os seus utilizadores. Adicionalmente grande parte do *software* código aberto disponibiliza o código-fonte de forma gratuita.

- Há portanto um ponto de contacto entre *software* livre e de código aberto. Qual é ?

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Filosofia

A filosofia do *software* de código aberto visa a colaboração entre os seus utilizadores. Adicionalmente grande parte do *software* código aberto disponibiliza o código-fonte de forma gratuita.

- Há portanto um ponto de contacto entre *software* livre e de código aberto. Qual é ?

Sinergia: trabalho colaborativo

A liberdade na redistribuição de versões modificadas permite que terceiros (comunidade de utilizadores) possam vir a beneficiar de melhorias e continuem a contribuir para esse desenvolvimento.

Ricardo São
João

Data Science

Software livre

**Software open
source**

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Origem

Nasce como alternativa ao código proprietário presente em *softwares* comerciais (pagamento de licenças).

No *software open source*, o(s) seu(s) autor(es) abdica(m) da propriedade intelectual do código de forma a que outros utilizadores possam tirar benefício.

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Características:

- Usualmente comunga dos quatro pilares do *software* livre;
- Não aplica qualquer tipo de discriminação/restricção aos seus utilizadores.

Vantagens:

- *Download* acessível e gratuito;
- Trabalho em plataformas colaborativas podendo usufruir dos contributos de outros utilizadores;
- Beneficia de uma melhoria constante impulsionada pela comunidade de utilizadores*;
- Redução de custos.

* Tal realidade não é expectável num *software* de código proprietário.

Ricardo São
João

Data Science

Software livre

Software open
source

**Características &
Vantagens**

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Analogia open source + livre com as maçãs:

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

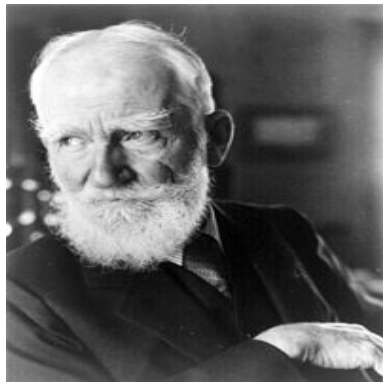
Ambiente

Aplicações

R for Data
Science
Program

Analogia open source + livre com as maçãs:

“Se tu tiveres uma maçã e eu tiver uma maçã, trocando essas maçãs, continuaremos, cada um, a ter uma maçã. Mas se tu tiveres uma ideia e eu tiver uma ideia, trocando essas ideias, cada um de nós passará a ter duas ideias.”



George Bernard Shaw (1856-1950)

R: *open source* + livre + grátis um bom “casamento”!

- O R é um projecto *open source*, que utiliza a linguagem de programação R, linguagem por excelência no tratamento e análise de dados;

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

R: *open source* + livre + grátis um bom “casamento”!

Ricardo São
João

Data Science

Software livre

Software *open
source*

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

- O R é um projecto *open source*, que utiliza a linguagem de programação R, linguagem por excelência no tratamento e análise de dados;
- Surge em 1993 e foi criado originalmente por **Ross Ihaka** e por **Robert Gentleman** no departamento de Estatística da Universidade de Auckland, Nova Zelândia.

R: *open source* + livre + grátis um bom “casamento”!

Ricardo São
João

Data Science

Software livre

Software *open
source*

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

- O R é um projecto *open source*, que utiliza a linguagem de programação R, linguagem por excelência no tratamento e análise de dados;
- Surge em 1993 e foi criado originalmente por **Ross Ihaka** e por **Robert Gentleman** no departamento de Estatística da Universidade de Auckland, Nova Zelândia.
- Funciona em todos os sistemas operativos e tem 19.023 pacotes gratuitos (08/04/2021 **ontem**) transversais a todas as ciências.



Fonte:

<https://cran.r-project.org/>

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

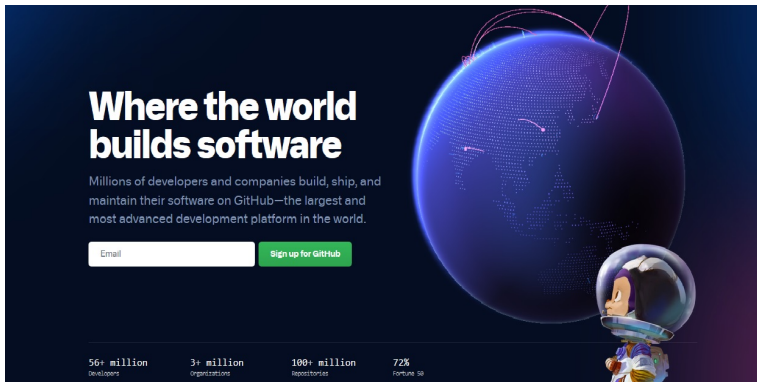
Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program



The image shows a promotional banner for GitHub. On the right, there is a large, glowing blue globe with a grid of dots and red lines connecting various points, symbolizing global connectivity. In the bottom right corner, a small cartoon astronaut in a white and blue suit is looking up at the globe. The background is a dark blue gradient.

Where the world builds software

Millions of developers and companies build, ship, and maintain their software on GitHub—the largest and most advanced development platform in the world.

Email [Sign up for GitHub](#)

56+ million Developers	3+ million Organizations	100+ million Repositories	72% Fortune 50
---------------------------	-----------------------------	------------------------------	-------------------

Fonte: <https://github.com/>

Ricardo São João

Data Science

Software livre

Software open source

Características & Vantagens

Gráficos

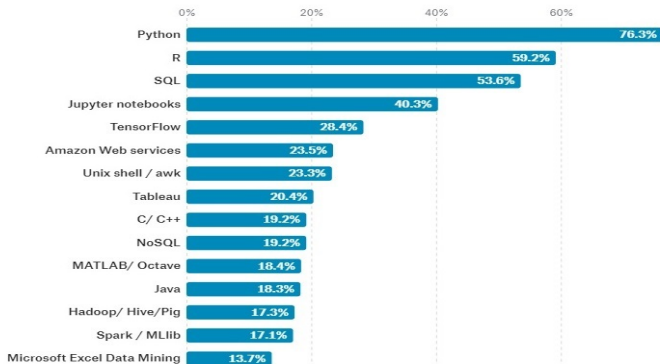
Documentos dinâmicos

Requisitos

Ambiente

Aplicações

R for Data Science Program



Fonte: <https://stackoverflow.blog>

R: Educação, Investigação, Data Science, Business Analyst, Data Analyst, Data Miner, Operations Researcher, Predictive Modeler, Marketing Researcher, Jornalismo, SIG's, . . .

Ricardo São João

Data Science

Software livre

Software open source

Características & Vantagens

Gráficos

Documentos dinâmicos

Requisitos

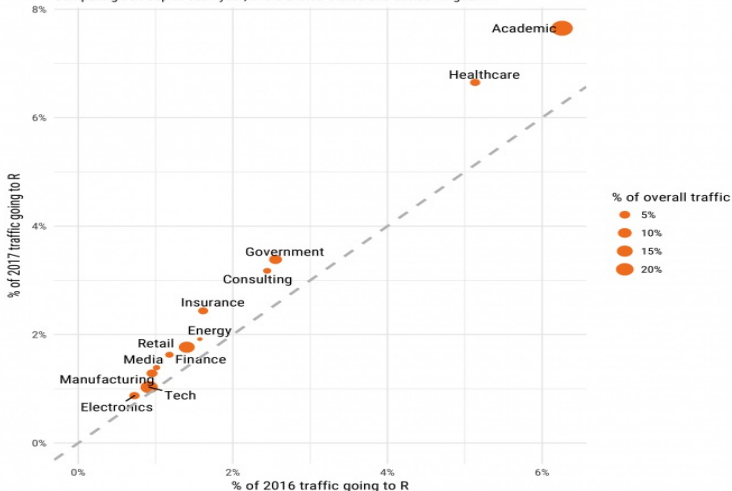
Ambiente

Aplicações

R for Data Science Program

Traffic by industry to R

Comparing Jan-Sep of each year, in the United States and United Kingdom.



Fonte: <https://blog.revolutionanalytics.com/popularity/>

Ricardo São João

Data Science

Software livre

Software open source

Características & Vantagens

Gráficos

Documentos dinâmicos

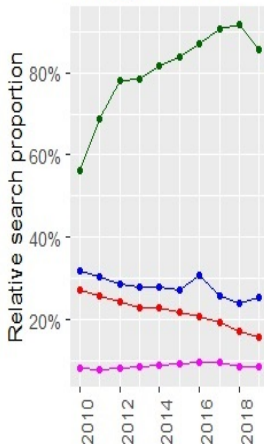
Requisitos

Ambiente

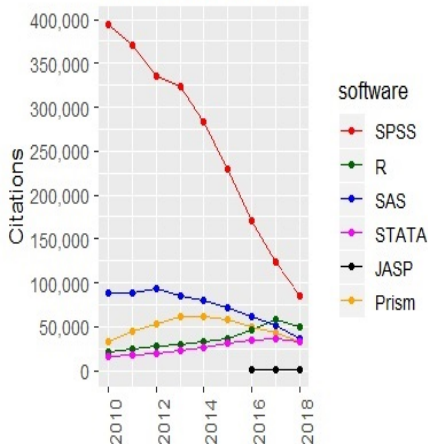
Aplicações

R for Data Science Program

Google Trends

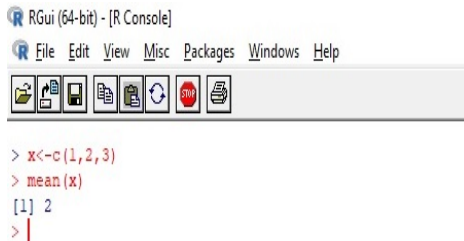


Scholar Citations



Fonte: <https://lindeloev.net/> (13, March 2019)

As instruções no R são dadas por comandos em linhas de código numa consola o que “inibe” à partida os seus (futuros) utilizadores.



no entanto . . .

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

...existem interfaces gráficas (*Graphical User Interface - GUI*) que permitem uma utilização mais fácil e intuitiva do R.

Dentre as várias GUI's disponíveis destacam-se duas:

- R Commander;
- RStudio.

Ambas as GUI necessitam ter o R instalado acessível em
<https://cran.r-project.org/>

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

R Commander

File Edit Data Statistics Graphs Models Distributions Tools Help

Data set: <No active dataset> Edit data set View data set Model: <No active model>

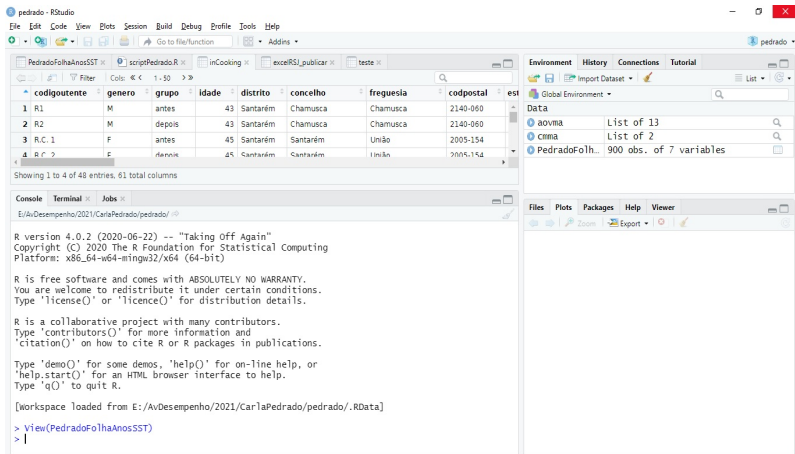
R Script R Markdown

Output Submit

Messages

```
[2] WARNING: The Windows version of the R Commander works best under  
RGui with the single-document interface (SDI); see ?Commander.
```

<https://cran.r-project.org/web/packages/Rcmdr/index.html>



The screenshot shows the RStudio interface with the following components:

- Environment Pane:** Displays a data frame with columns: `codgoutente`, `genero`, `grupo`, `idade`, `distrito`, `concelho`, `freguesia`, `codpostal`, and `est`. The first three rows are visible:

codgoutente	genero	grupo	idade	distrito	concelho	freguesia	codpostal	est
1 R1	M	antes	43	Santarém	Chamusca	Chamusca	2140-060	
2 R2	M	depois	43	Santarém	Chamusca	Chamusca	2140-060	
3 R.C. 1	F	antes	45	Santarém	Santarém	Ulião	2005-154	
- Console:** Shows the R version (4.0.2) and workspace loaded from `E:/AvDesempenho/2021/CarlaPedrado/pedrado/.RData`. It also displays the prompt `> View(PedradoFolhaAnosSST)`.
- Environment:** Lists loaded objects: `aovma` (List of 13), `cnma` (List of 2), and `PedradoFolh...` (900 obs. of 7 variables).

<https://rstudio.com/>

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

O R dispõem de *plug-ins* (módulos) que permitem elaborar análises e gráficos específicos para diferentes áreas de estudo.

Pacotes (*packages*) e Bibliotecas (*libraries*)

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program



Sempre que necessário o **R** permite o *download* de forma gratuita a mais de 19 mil pacotes (08/04/2022)!

```
install.packages("nomepacote");library(nomepacote)
```

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

✿ Vamos implementar agora alguns ...

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

**Documentos
dinâmicos**

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Quantas vezes ...

- não teve de reformular um documento/relatório com base em novas informações/dados ?

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

**Documentos
dinâmicos**

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Quantas vezes ...

- não teve de reformular um documento/relatório com base em novas informações/dados ?
- a deteção de um valor/parâmetro incorreto numa análise, não comprometeu o estudo realizado?

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

**Documentos
dinâmicos**

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Quantas vezes ...

- não teve de reformular um documento/relatório com base em novas informações/dados ?
- a deteção de um valor/parâmetro incorreto numa análise, não comprometeu o estudo realizado?
- não gostaria de poder reproduzir os mesmos valores/resultados com base na mesma informação de um artigo científico ou relatório técnico ?

Ricardo São João

Data Science

Software livre

Software open source

Características & Vantagens

Gráficos

Documentos dinâmicos

Requisitos

Ambiente

Aplicações

R for Data Science Program

Segundo Leek & Jager (2017) são conceitos distintos com uma diferença subtil:

Reprodutibilidade/Reproduzível A reprodutibilidade é a capacidade de com base no código e nos dados de uma publicação/artigo, voltar a executar o código e obter os mesmos resultados;

Replicabilidade/Replicável A replicabilidade é a capacidade em voltar a realizar uma mesma experiência, com novos dados e obter resultados "consistentes" com o estudo original.

Leek, J. T., & Jager, L. R. (2017) Is most published research really false?. *Annual Review of Statistics and Its Application*, 4, 109-122.

✿ Vamos criar um relatório com o rmarkdown.



<https://rmarkdown.rstudio.com/>

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

**Documentos
dinâmicos**

Requisitos

Ambiente

Aplicações

R for Data
Science
Program



<https://shiny.rstudio.com/>
<https://shiny.rstudio.com/gallery/radiant.html>

Ricardo São
João

Data Science

Software livre

Software open
source

Características &
Vantagens

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

- 1 Instale o **R** a partir do CRAN (*The Comprehensive R Archive Network*) acessível em <https://cran.r-project.org/>
- 2 Instale o **RStudio** acessível em <https://www.rstudio.com/products/rstudio/>

Iremos interagir com o **R** via **RStudio**

Graphical User Interface (GUI)

The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains R code:


```
1 x <- seq(from = 1, to = 10, by = 1)
2 |
```
- Console:** Shows the execution of `head(df)` and `df$ssex[1]`, resulting in:


```
# A tibble: 5 x 2
  sex age
<fct> <dbl>
1 male 22
2 female 45
3 male 33
4 male 27
5 female 30
> df$ssex[1]
[1] male
Levels: female male
> df$ssex[4]
[1] male
Levels: female male
> df$ssex[5]
[1] female
Levels: female male
> library(R)
```
- Environment Pane:** Shows a data frame `df` with 5 observations and 2 variables. The values are:

numbers	num
x	[1:6] 12 13 15 11 NA 10
x1	[1:10] 1 2 3 4 5 6 7 8 9 10
x2	[1:10] 11 12 13 14 15 16 17 18 19 20
- Files Pane:** Lists installed and available packages, including `readr`, `readstata13`, `readxl`, `cellranger`, `dplyr`, `foreign`, `geomet`, `ggpubr`, `ini`, `jpeg`, `openxlsx`, `png`, and `prettyunits`.

Ricardo São João

Data Science

Software livre

Software open source

Características & Vantagens

Gráficos

Documentos dinâmicos

Requisitos

Ambiente

Aplicações

R for Data Science Program

A título ilustrativo vamos considerar um índice constituído por 10 items avaliados numa escala tipo likert com as seguintes pontuações: 0 (no); 5 (sometimes) e 10 (yes).

Item	Description
1	Do you have difficult opening your mouth wide?
2	Do you have difficulty moving your jaw to the sides?
3	Do you feel fatigue or muscle pain when you chew?
4	Do you have frequent headaches?
5	Do you have neck pain or a stiff neck?
6	Do you have ear aches or pain in that area?
7	Have you ever noticed any noise while chewing or opening your mouth?
8	Do you have any habits such as clenching or grinding your teeth?
9	Do you feel that your teeth do not come together well?
10	Do you consider yourself a nervous person?

No **R** é possível realizar uma análise detalhada. Ora vejamos ...

Análise bivariada em tabelas de contingência

Considere um estudo constituído por 100 leitores de poetas portugueses. A seguinte tabela de contingência mostra a frequência dos leitores tendo em conta o seu sexo e o poeta presentemente lido.

Poeta	Sexo do leitor	
	M	F
Luís de Camões (1524-1580)	12	26
Fernando Pessoa (1888-1935)	8	10
Sophia Andresen (1919-2004)	30	14



Existirá alguma relação entre o sexo do leitor e a escolha do poeta ?
Se sim, com que intensidade ? Veja uma abordagem no **R** ...

Exemplo

A base de dados penguins disponível no pacote palmerpenguins do R retrata as medidas de um conjunto de 344 pinguins de três espécies (Adelie, Chinstrap and Gentoo) que passam pelo Arquipélago Palmer. Através da ANOVA procure dar resposta à seguinte questão: “O comprimento médio das barbatanas apresenta diferenças estatisticamente significativas nas 3 espécies de pinguins?”^a

^a adaptado de <https://statsandr.com>. Imagens retiradas de: penguinlovers.de; 4.bp.blogspot.com; images.fineartamerica.com



Adelie



Chinstrap



Gentoo

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

Documentos
dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
Science
Program

Foi realizado um estudo com base em 40 utilizadores de postos de lavagem auto. De acordo com uma teoria no campo da psicanálise, espera-se que pessoas com forte extroversão gastem mais tempo na limpeza do seu automóvel, uma vez que **procuram projetar a sua imagem** através deste bem (e de outros). Por outro lado, o comportamento associado à limpeza de um automóvel poderá estar relacionado com aspetos demográficos tais como a idade e o sexo. A cada utilizador foi registada a seguinte informação: sexo, idade, pontuação na escala extroversão e tempo gasto em minutos, dedicado semanalmente à limpeza do automóvel.

Ricardo São
João

Data Science

Software livre

*Software open
source*

*Características &
Vantagens*

Gráficos

Documentos
dinâmicos

Requisitos

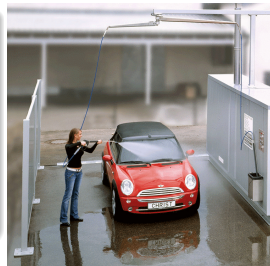
Ambiente

Aplicações

R for Data
Science
Program

Dar resposta à questão: em que medida as variáveis sexo, idade e extroversão podem explicar o tempo dispendido na limpeza do automóvel (variável carro) ?

- ✦ variável resposta? **carro**
- ✦ variáveis explicativas? **sexo, idade e extroversão**



Este seminário está integrado num programa, constituído por um conjunto de 5 cursos (independentes e não sobreponíveis) que pretendem disponibilizar um leque alargado de temáticas de *Data Science* com utilização da linguagem de programação R.



:

O conjunto dos 5 cursos está acessível em
<https://gades-solutions.com/data-science-com-r/> que
 passo a apresentar ...

Ricardo São
 João

Data Science

Software livre

Software open
 source

Características &
 Vantagens

Gráficos

Documentos
 dinâmicos

Requisitos

Ambiente

Aplicações

R for Data
 Science
 Program

Obrigado pela Vossa **PPA** !
Presença, **P**aciência e **A**tenção.

Ricardo São João

ricardo.sjoao@esg.ipsantarem.pt