Data Science

com *software open source*+livre+grátis:
um bom casamento

Ricardo São João
IPSantarém, CEAUL,CIDNUR, CEG-UAb

ricardo.sjoao@esg.ipsantarem.pt

Outubro, 2025

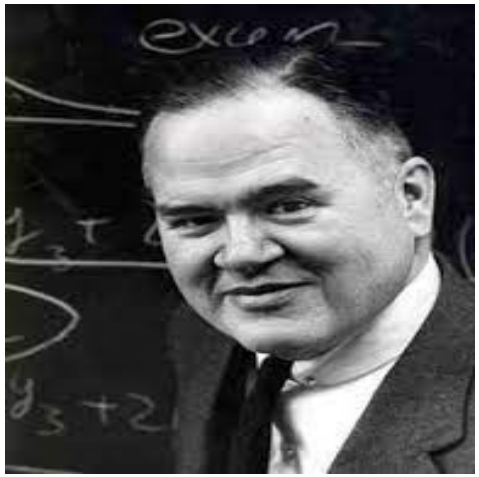
Linhas gerais

Data Science

O que é ?

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

Já em 1962 (há 63 anos) o matemático/estatístico John Tukey aborda a questão na publicação intitulada *The future of data analysis* [1]



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

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

Data Science

✤*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"**.

Data Science

✪ A profissão "Cientista de Dados" é das mais procuradas e melhor remuneradas.
✪ Ocupou o 6° lugar das 10 profissões mais procuradas em 2024. (Fonte: Linkedin)

https://www.linkedin.com/pulse/future-proof-your-career-top-10-in-demand-jobs-xlsvc

Volume de dados na internet[2]

Alguns números que dão que pensar...

- o volume de dados na internet no final de 2020 foi estimado em 44 zettabytes (1 zettabyte= 1.000.000.000.000.000.000.000= 10^{21});
- até 2025, espera-se que a quantidade de dados gerados a cada dia atinja 463 exabytes (1 exabyte=10^{18});

[2] fonte: https://seedscientific.com

Volume de dados na internet[2]

Alguns números que dão que pensar...

- o volume de dados na internet no final de 2020 foi estimado em 44 zettabytes (1 zettabyte= 1.000.000.000.000.000.000.000= 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.

[2] fonte: https://seedscientific.com

Volume de dados na internet[2]

Alguns números que dão que pensar...

- o volume de dados na internet no final de 2020 foi estimado em 44 zettabytes (1 zettabyte= 1.000.000.000.000.000.000.000= 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 mais de 12 mil milhões de dólares em compras na WWW;
- em 2030, nove em cada dez pessoas com idade $\geq$ 6 anos será digitalmente ativa.

[2] fonte: https://seedscientific.com

Software

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

A escolha pode recair em *software* comercial ou . . .

Software

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

A escolha pode recair em *software* comercial ou . . .

- livre;

Software

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

A escolha pode recair em *software* comercial ou . . .

- livre;
- *open source*;

Software

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

A escolha pode recair em *software* comercial ou . . .

- livre;
- *open source*;
- gratuito.

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

Software livre

Não se trata de gratuitidade ("de graça" ou "de borla").

Software livre

Não se trata de gratuitidade ("de graça" ou "de borla").

Estaremos então a falar de que conceito ?

Software livre

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

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

Software livre

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.

Software open source: o quê significa ?

O termo "*open source*" ou "código aberto" é o termo utilizado para *software* que disponibiliza o seu código fonte para edição.

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

https://medium.com/

Origem do *open source*

Origem

Nasce como alternativa ao código proprietário presente em *software* comercial (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.

Características & Vantagens do open source

Características:

- Usualmente comunga dos quatro pilares do *software* livre;
- Não aplica qualquer tipo de discriminação/restrição 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.

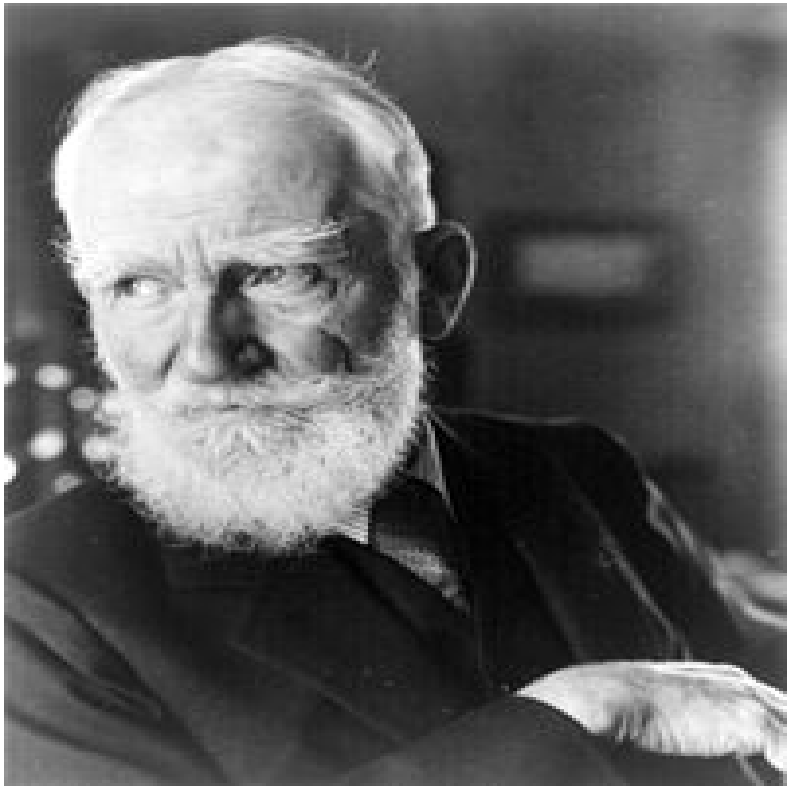
Open source + livre

Analogia *open source* + livre com as maçãs:

Open source + livre

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)
Nóbel da literatura em 1925;
Óscar de melhor argumento adaptado - 1939.

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;

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;
- Surge em 1993 e foi criado originalmente por **R**oss Ihaka e por **R**obert Gentleman no departamento de Estatística da Universidade de Auckland, Nova Zelândia.

R: *open source* + livre + grátis

um bom "casamento"!

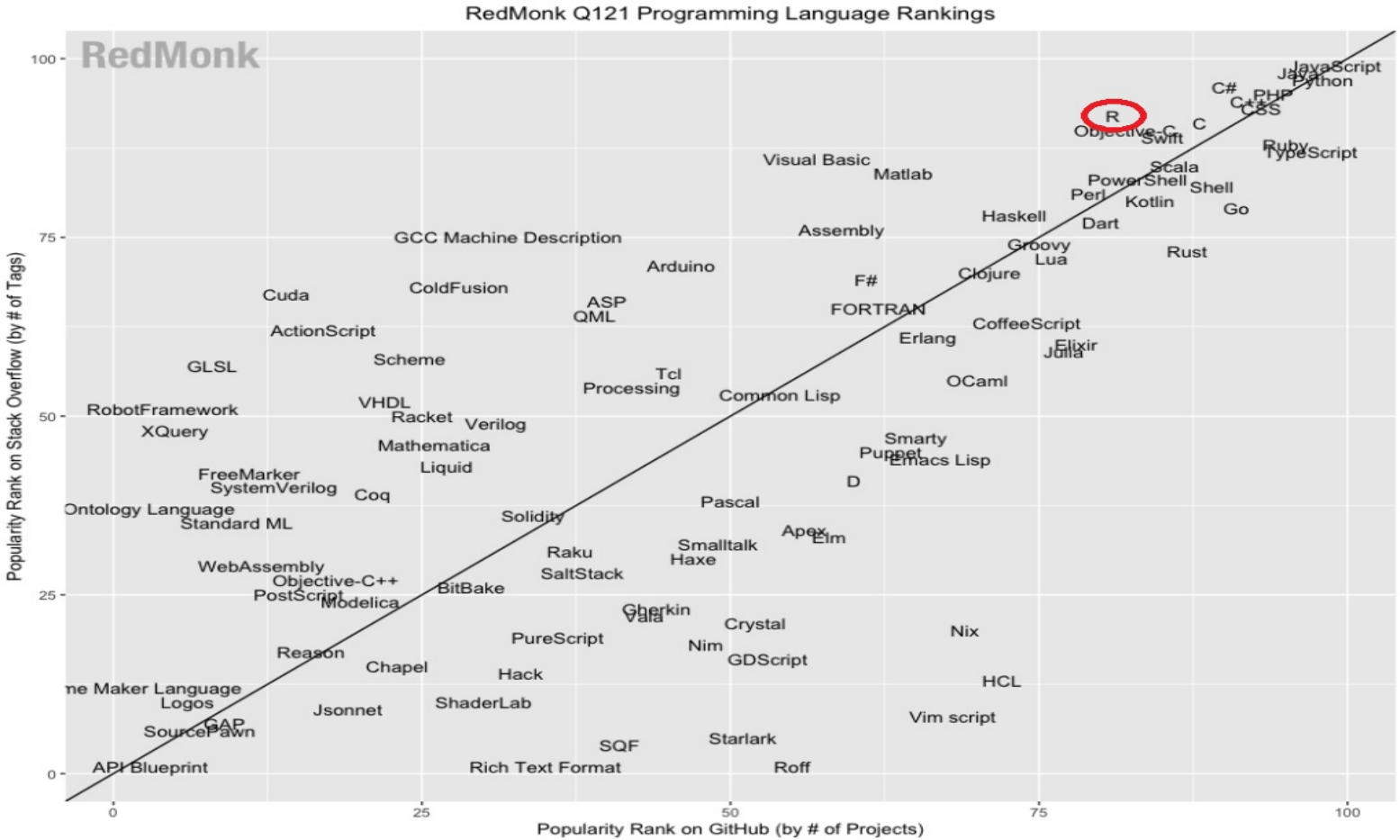
- O R é um projecto *open source*, que utiliza a linguagem de programação R, linguagem por excelência no tratamento e análise de dados;
- Surge em 1993 e foi criado originalmente por **R**oss Ihaka e por **R**obert Gentleman no departamento de Estatística da Universidade de Auckland, Nova Zelândia.
- Funciona em todos os sistemas operativos e tem mais de 22 mil pacotes gratuitos transversais a todas as ciências.

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

Popularidade: presente em plataformas colaborativas

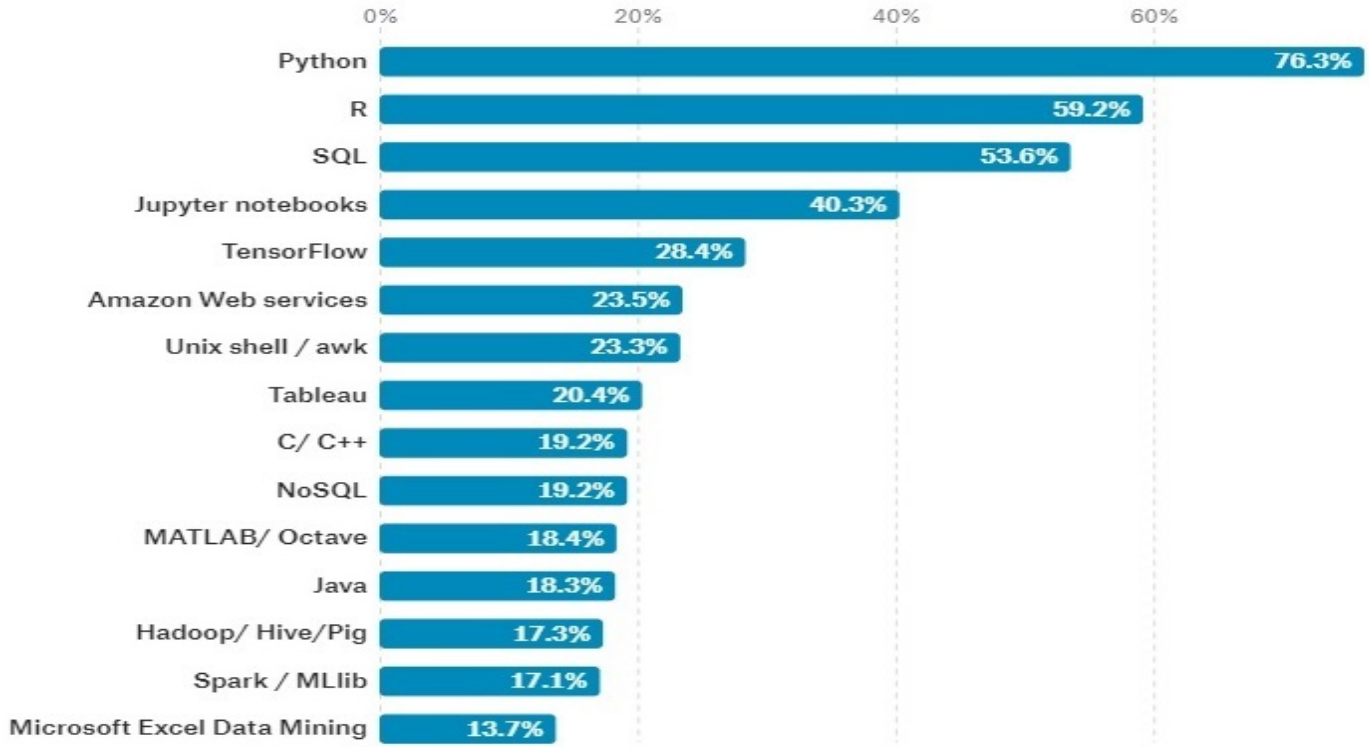
Fonte: https://github.com/

Popularidade: linguagem de programação



Programme Language Rankings (2021) Fonte: https://redmonk.com

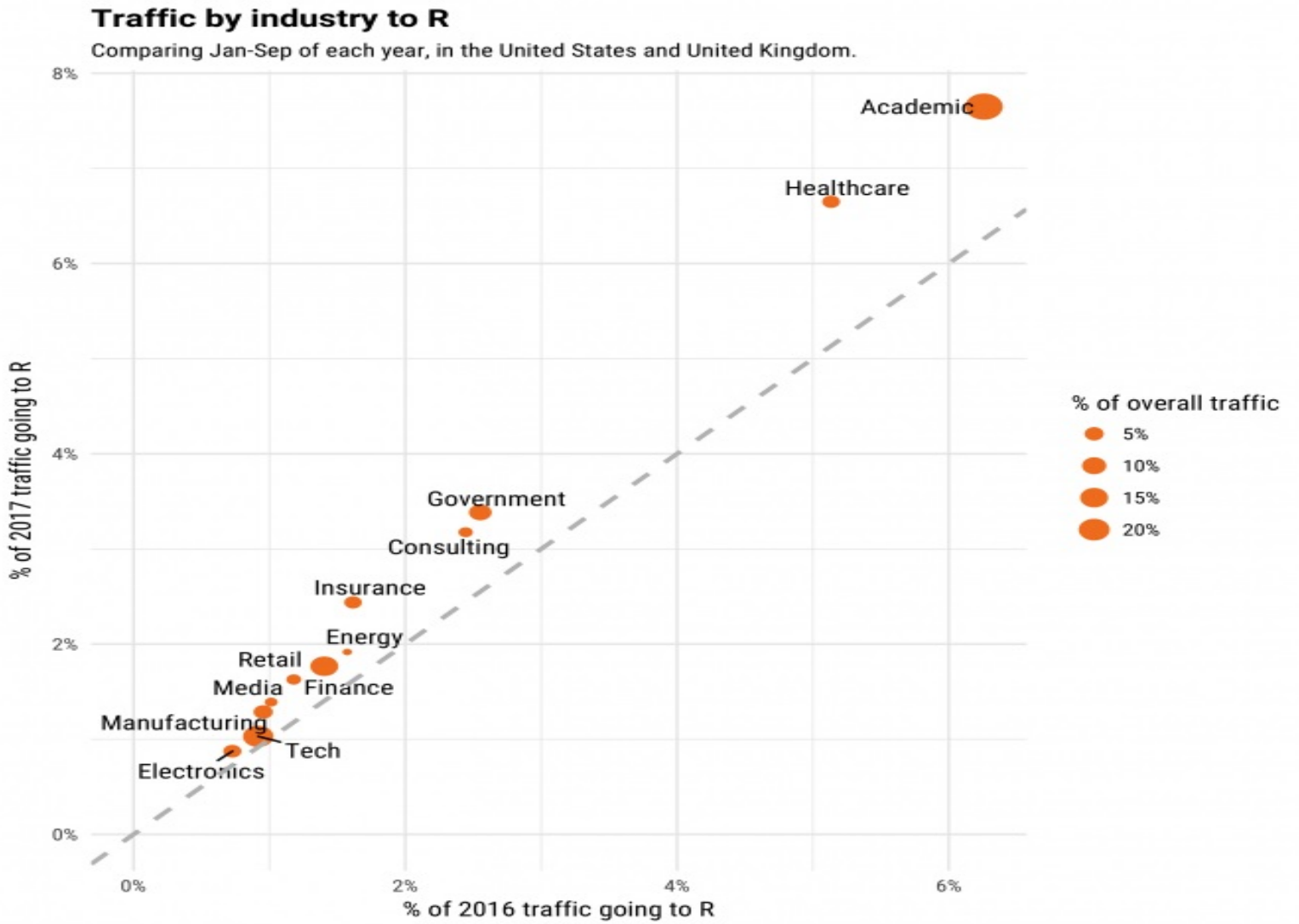
Work tool



Fonte: https://stackoverflow.blog

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

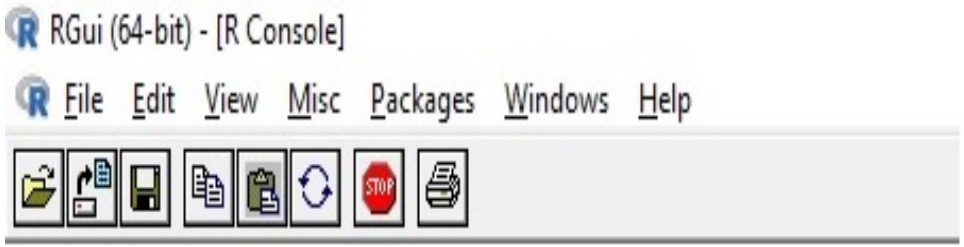
Áreas de atuação



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

Não há bela sem senão . . .

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




```
> x<-c(1,2,3)
> mean(x)
[1] 2
>
```

no entanto . . .

GUI

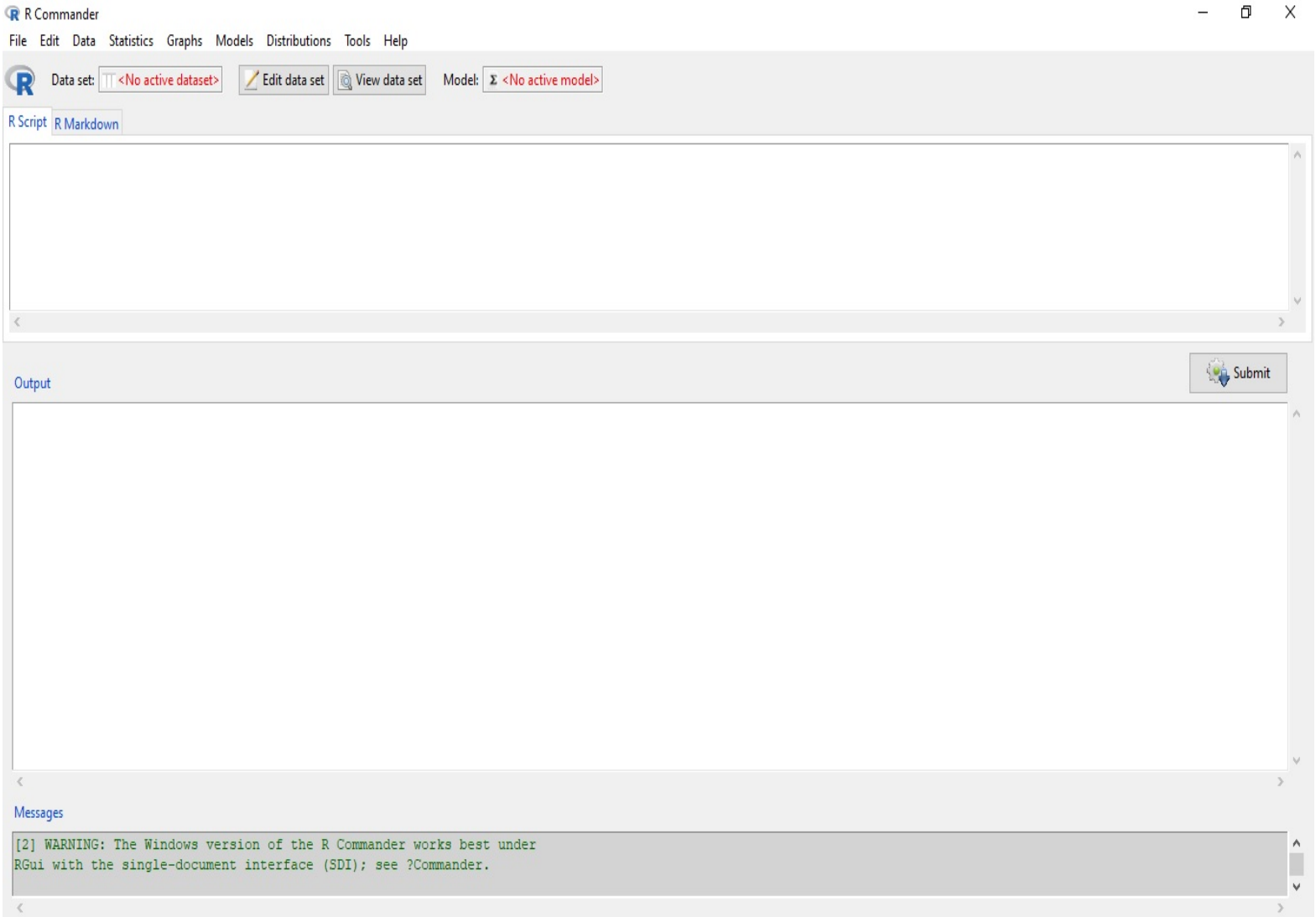
. . .existem interfaces gráficas (*Graphical User Interface - GUI*) que permitem uma utilização mais fácil e intuitiva do R.

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

- R Commander;
- RStudio.

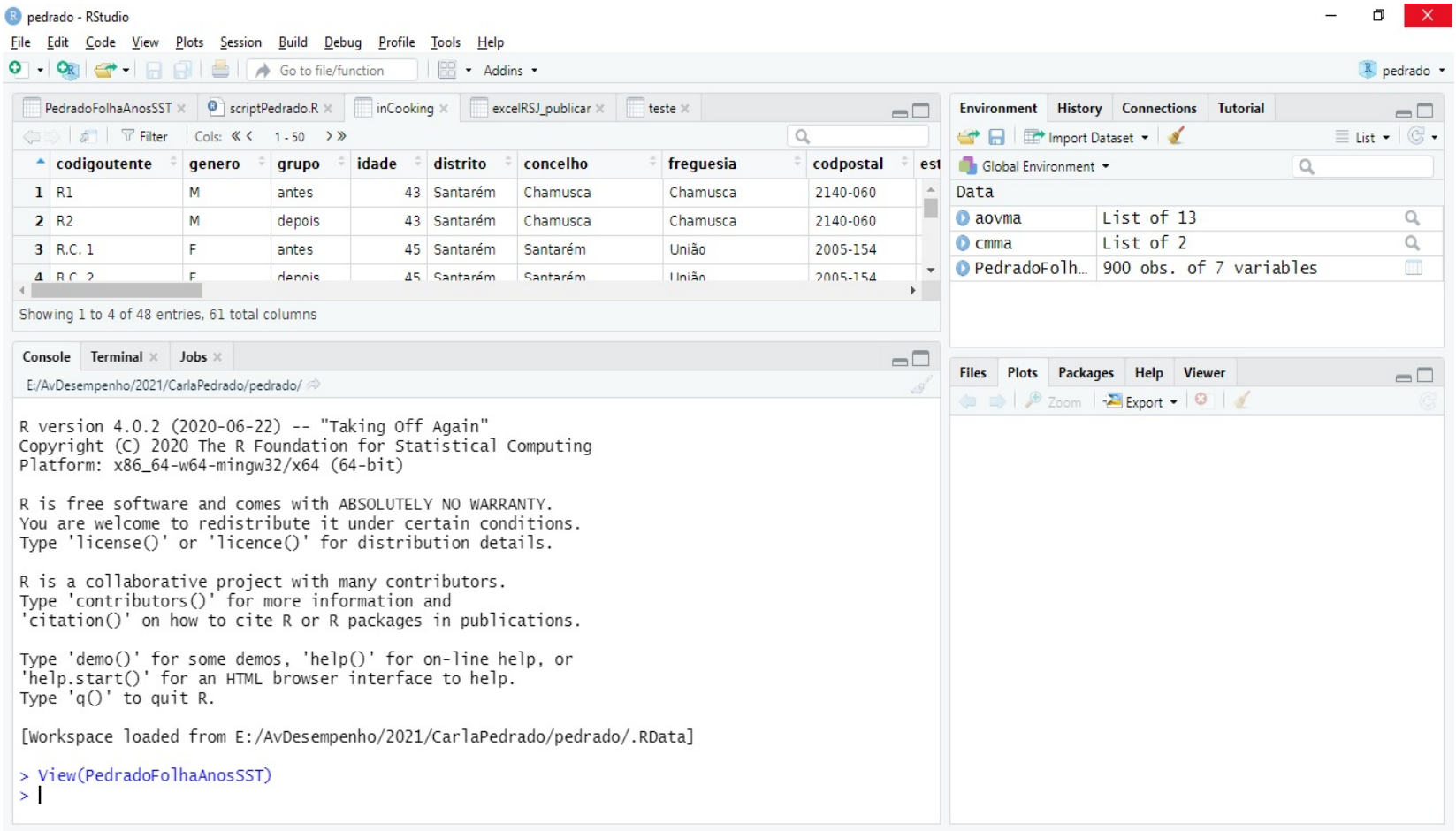
Ambas as GUI necessitam ter o R instalado acessível em https://cran.r-project.org/

R Commander



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

RStudio



https://rstudio.com/

Plug-ins

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

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

Sempre que necessário o **R** permite o *download* de forma gratuita de mais de 22 mil pacotes !

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

Gráficos - *The R graph gallery*

https://r-graph-gallery.com/index.html

✿ Vamos ver/implementar agora alguns . . .

https://r-graph-gallery.com/9-ordered-boxplot.html#grouped

https://r-graph-gallery.com/199-correlation-matrix-with-ggally.html

https://r-graph-gallery.com/44-polynomial-curve-fitting.html

Documentos dinâmicos e reproduzíveis

Quantas vezes . . .

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

Documentos dinâmicos e reproduzíveis

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?

Documentos dinâmicos e reproduzíveis

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 ?

R markdown

- Vamos criar um relatório com o rmarkdown.

https://rmarkdown.rstudio.com/

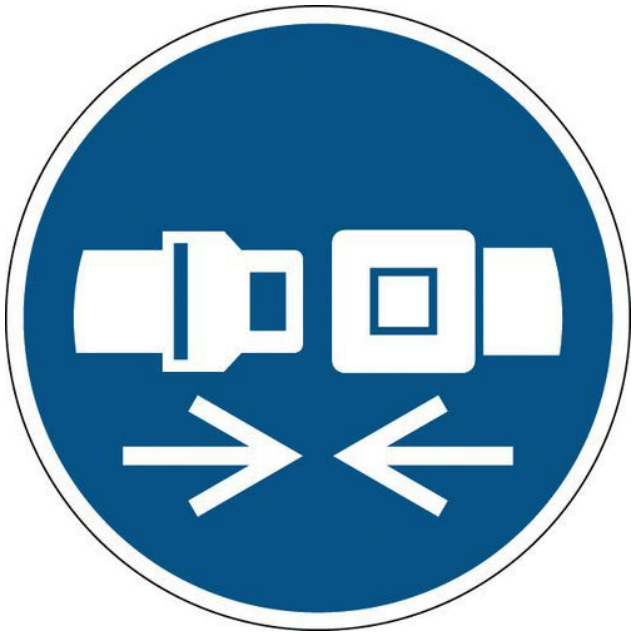
Shiny

https://shiny.rstudio.com/

Requisitos

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://posit.co/download/rstudio-desktop/

✿ Vamos começar o nosso *tour* com alguns exemplos.
Preparados . . . apertem os vossos cintos de segurança !



fonte imagem:https://www.kaiserkraft.pt

Políticos na II Guerra Mundial - MDS não métrico

Hitler

Mussolini

Churchill

Eisenhower

Stalin

Franco

Charles de Gaulle

Tito

imagens retiradas de https://pt.wikipedia.org

Políticos na II Guerra Mundial - MDS não métrico

O modelo MDS permite a representação espacial de **perceções** e **preferências** de consumidores/inquiridos ao nível da (dis)similaridade entre objetos

Foram avaliadas as parecenças/semelhanças [3] entre alguns políticos na II Grande Guerra Mundial com base numa escala ordinal, cuja informação apresenta-se na seguinte matriz:

```
> politicosIIguerramundial
            Hitler Mussolini Churchill Eisenhower Stalin Attlee Franco De_Gaulle Mao_Tse Truman Chamberlain Tiro
Hitler           0
Mussolini        5         0
Churchill       11        14         0
Eisenhower      15        16         7          0
Stalin           8        13        11         16      0
Attlee          17        18        11         16     15      0
Franco           5         3        12         14     13     16      0
De_Gaulle       10        11         5          8     11     12      9         0
Mao_Tse         16        18        16         17     12     16     17        13       0
Truman          17        18         8          6     14     12     16         9      12       0
Chamberlain     12        14        10          7     16      9     10        11      17       9           0
Tiro            16        17         8         12     12     13     12         7      10      11          15   0
```

O modelo MDS não métrico irá ajudar a interpretar a informação.

[3] adaptado de https://rpubs.com/Hasantha_APS_1701/

Interpretação

No eixo das abcissas podemos ter em conta a dimensão ideologia. À medida que caminhamos para a esquerda do eixo das abcissas a ideologia nazi ganha relevo. No eixo y podemos ter a dimensão localização geográfica. Políticos posicionados mais acima do eixo correspondem a países orientais.

Escalas tipo Likert

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 openning 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.

	Sexo do leitor	
Poeta	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** . . .

Testes de hipóteses 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

Regressão Linear

O presente exemplo consiste na informação recolhida em 1974 pela revista *Motor Trend US magazine*[a] e diz respeito ao ensaio de 32 modelos de automóveis, comercializados entre os anos de 1973 e 1974, onde são avaliadas características técnicas e o consumo.
A base de dados está presente no R acessível com a instrução `data(mtcars)`. Para maiores detalhes faça `??mtcars`

[a] ainda comercializada https://www.motortrend.com/

Objetivo

Com base num modelo de RL procure explicar o consumo de um automóvel com base em algumas especificações técnicas.

Obrigado pela Vossa PPA !

Presença, Paciência e Atenção.

Ricardo São João

ricardo.sjoao@esg.ipsantarem.pt