ANÁLISE EXPLORATÓRIA DE DADOS

Hugo Carvalho
hugo@dme.ufrj.br
DME/IM/UFRJ

BIBLIOGRAFIA

- Livros & etc.:
  - Pedro A. Morettin & Wilton O. Bussab - **Estatística Básica**, 9ª Ed.
  - Robert Kabacoff - **R in Action: Data Analysis and Graphics with R**, 2ª ed.
  - Hadley Wickham & Garrett Grolemund - **R for Data Science: Import, Tidy, Transform, Visualize, and Model Data**
  - John W. Tukey - **Exploratory Data Analysis**
  - Edward R. Tufte - **The Visual Display of Quantitative Information**, 2ª ed.
  - Artigos, sites, etc., informados no Google Classroom
- Softwares & etc.:
  - R Studio/Google Colab

Toda a parte computacional pode ser feita somente no *browser*!

AVALIAÇÃO

- 04 testes:
  - Datas no Google Classroom
  - Descarta a menor nota e faz a média do restante
  - Vale 30% da nota final
- 02 avaliações presenciais:
  - Datas no Google Classroom
  - Cada um vale 15% da nota final
- 01 projeto:
  - Data no Google Classroom
  - Vale 40% da nota final

Média para aprovação: 5,0

O QUE É AED (ANÁLISE EXPLORATÓRIA DE DADOS)?

- Análise **exploratória**:
  - Tukey, 1961: "*Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.*"

- Análise **confirmatória**:
  - Análise das evidências coletadas por métodos estatísticos tradicionais – possível tomada de decisão
  - Inferência, testes de hipótese, análise de regressão, etc...

UM EXEMPLO (FONTE: WIKIPEDIA)

- Objetivo
  - Entender fatores que influenciam gorjetas para garçons
- A base de dados – 244 observações das variáveis abaixo:
  - gorjeta (U$)
  - valor da conta (U$)
  - Sexo do pagante
  - Mesa com ou sem fumantes
  - Dia da semana
  - Horário da refeição
  - Quantidade de pessoas
- Nova variável: gorjeta %

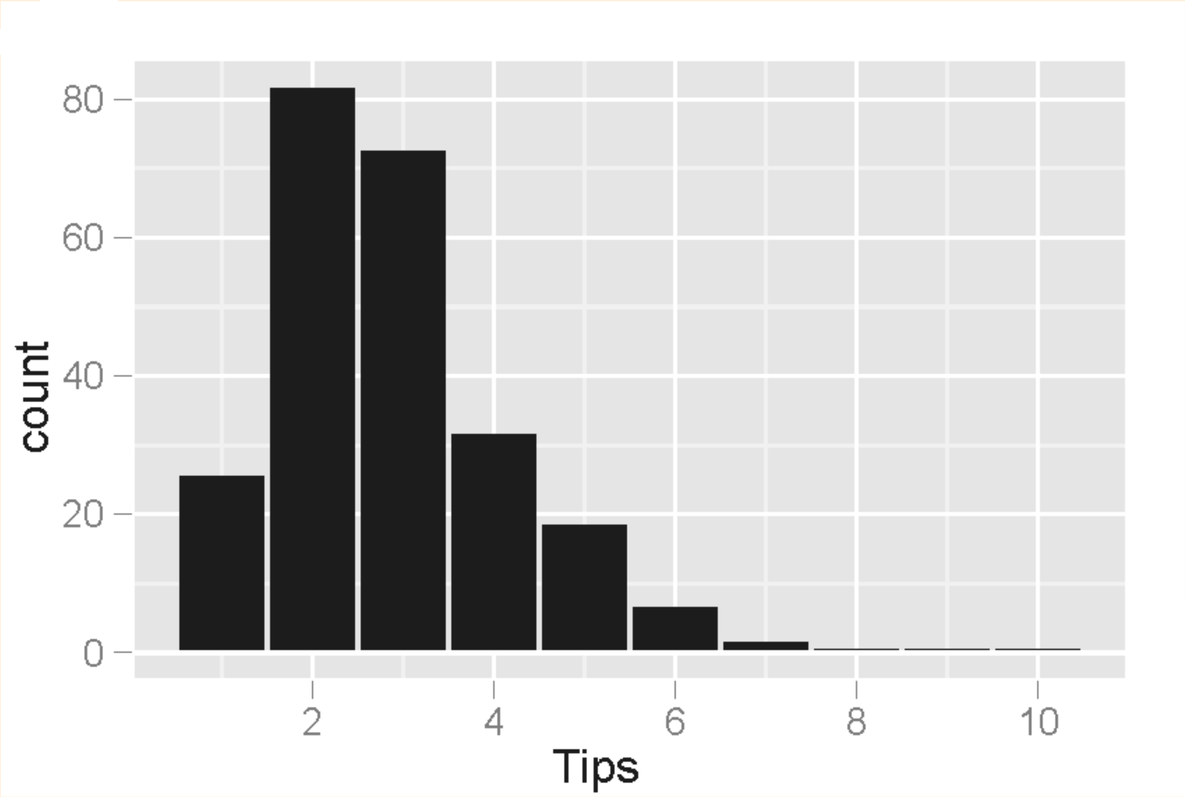
UM EXEMPLO

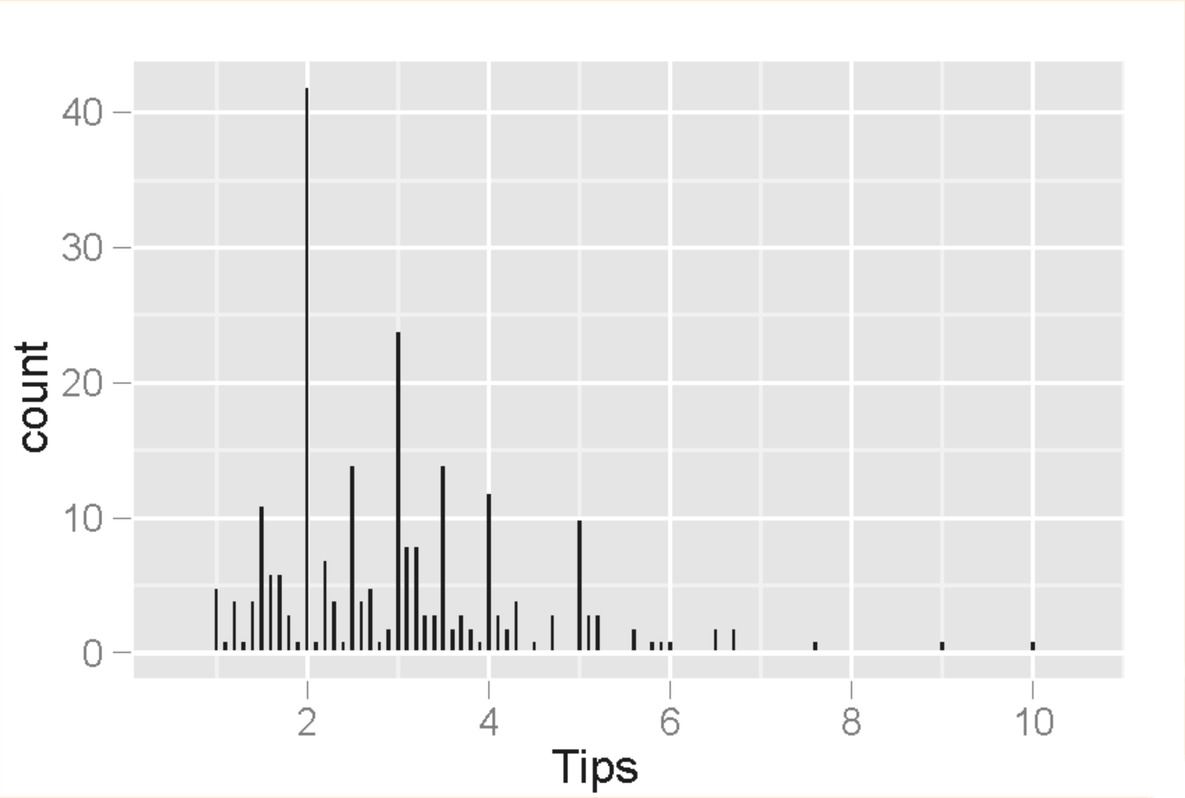
- Uma possível conclusão:
  - "a gorjeta % base é de 18%; uma pessoa a mais na mesa, diminui em 1%"

$$(gorjeta\ \%) = 0{,}18\ - 0{,}01\ * (quantidade\ de\ pessoas)$$

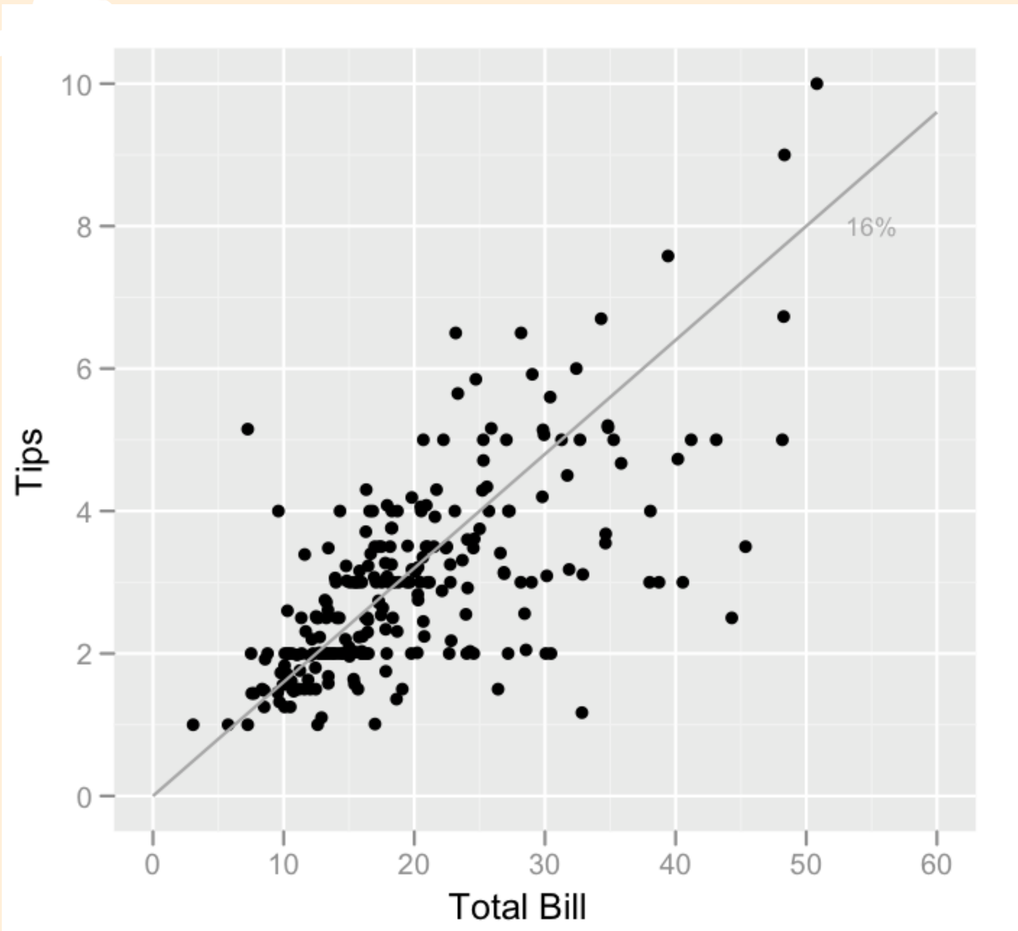
- O que mais esse conjunto de dados pode nos dizer?

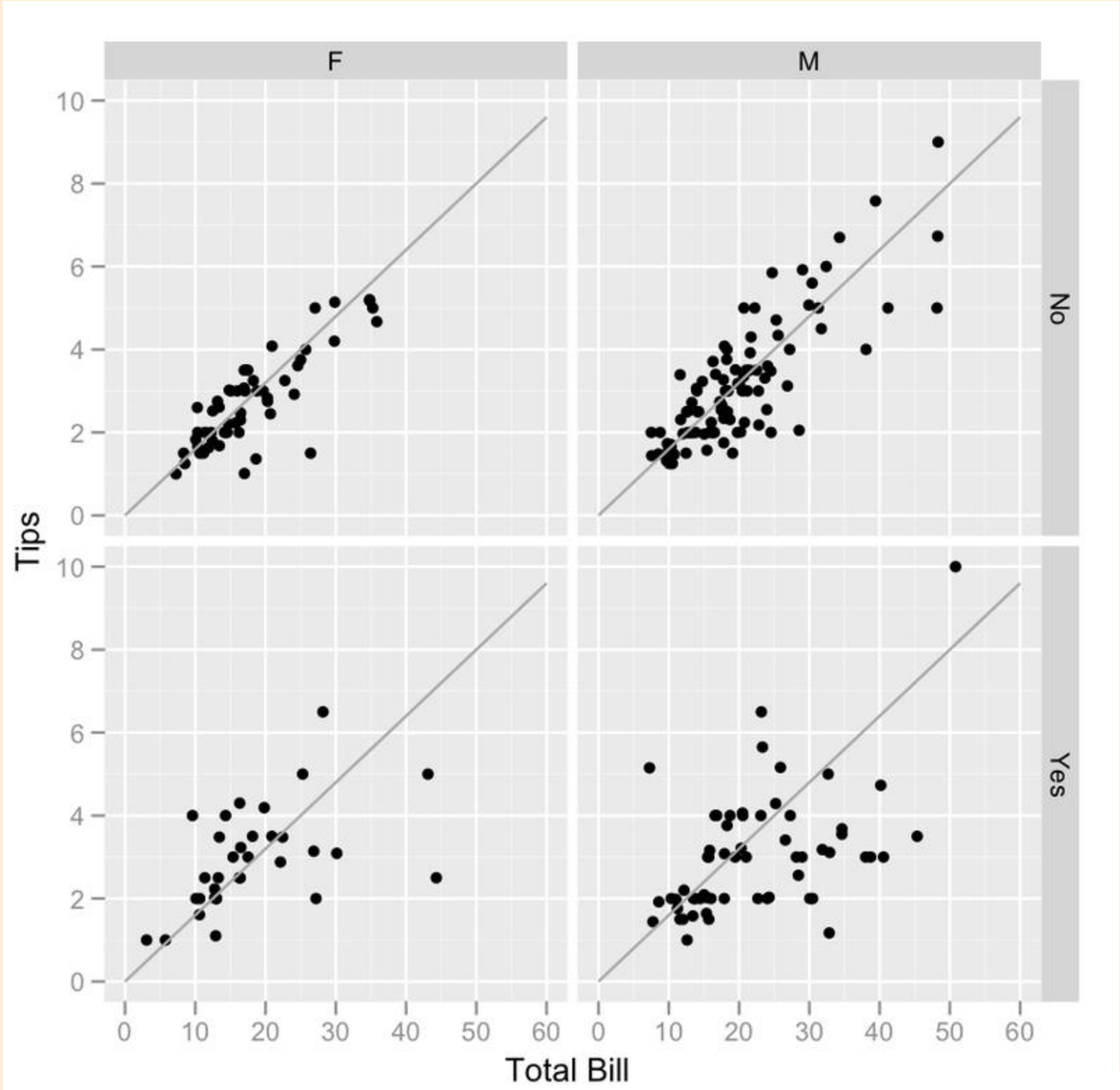
UM EXEMPLO





UM EXEMPLO


Tips
Total Bill
16%
0
2
4
6
8
10
0
10
20
30
40
50
60

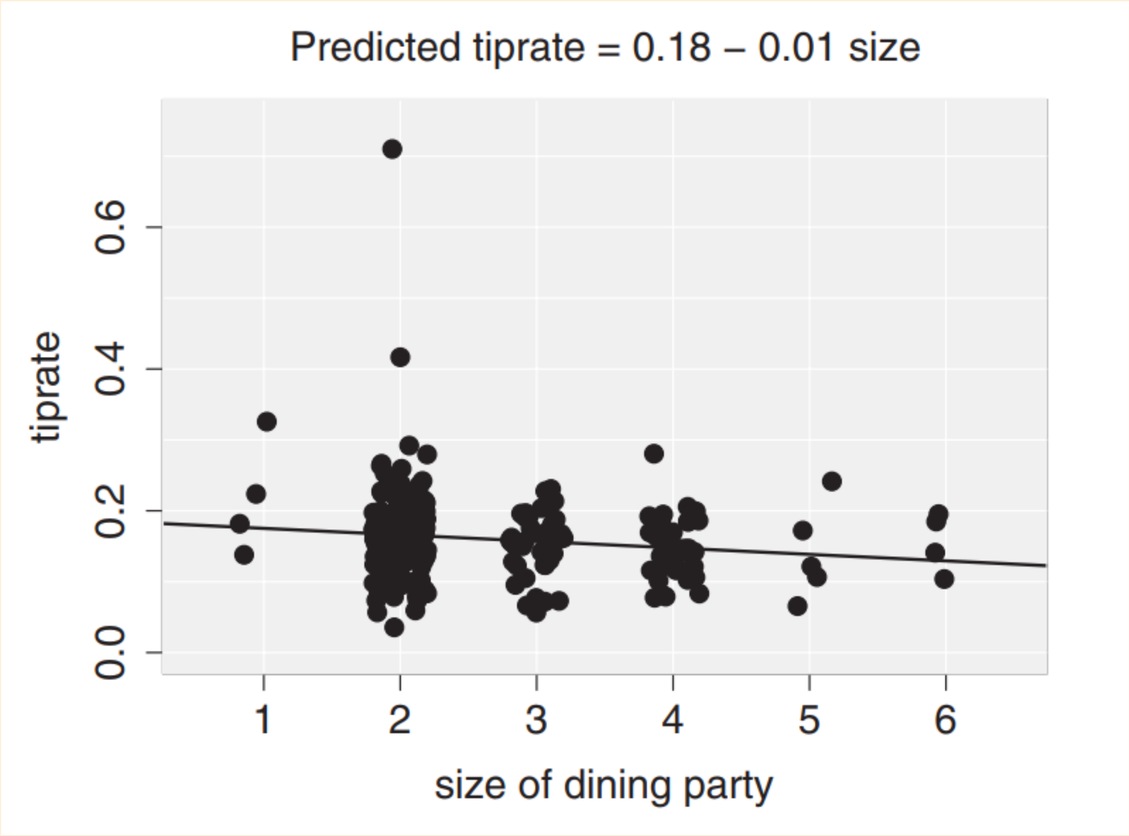

F
M
No
Yes
Tips
Total Bill
0
2
4
6
8
10
0
10
20
30
40
50
60

UM EXEMPLO

- "A gorjeta % base é de 18%; uma pessoa a mais na mesa, diminui em 1%"

$$(gorjeta\ \%) = 0{,}18\ - 0{,}01\ *\ (quantidade\ de\ pessoas)$$

- De onde essa conclusão veio?



O QUE É AED (ANÁLISE EXPLORATÓRIA DE DADOS)?

- Tukey, 1977:
  - *"it is important to understand what you can do before you learn to measure how well you seem to have done it"*
  - *"Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone – as the first step"*
  - *"to learn about data analysis, it is right that each of us try many things that do not work"*

MAIS ALGUMAS FRASES DE SABEDORIA...

- *"if data analysis is to be well done, most of it must be a matter of judgement, and 'theory', whether statistical or not, will have to guide, not to command"*

- *"an approximate answer to the right question is worth a great deal more than a precise answer to a wrong question"*

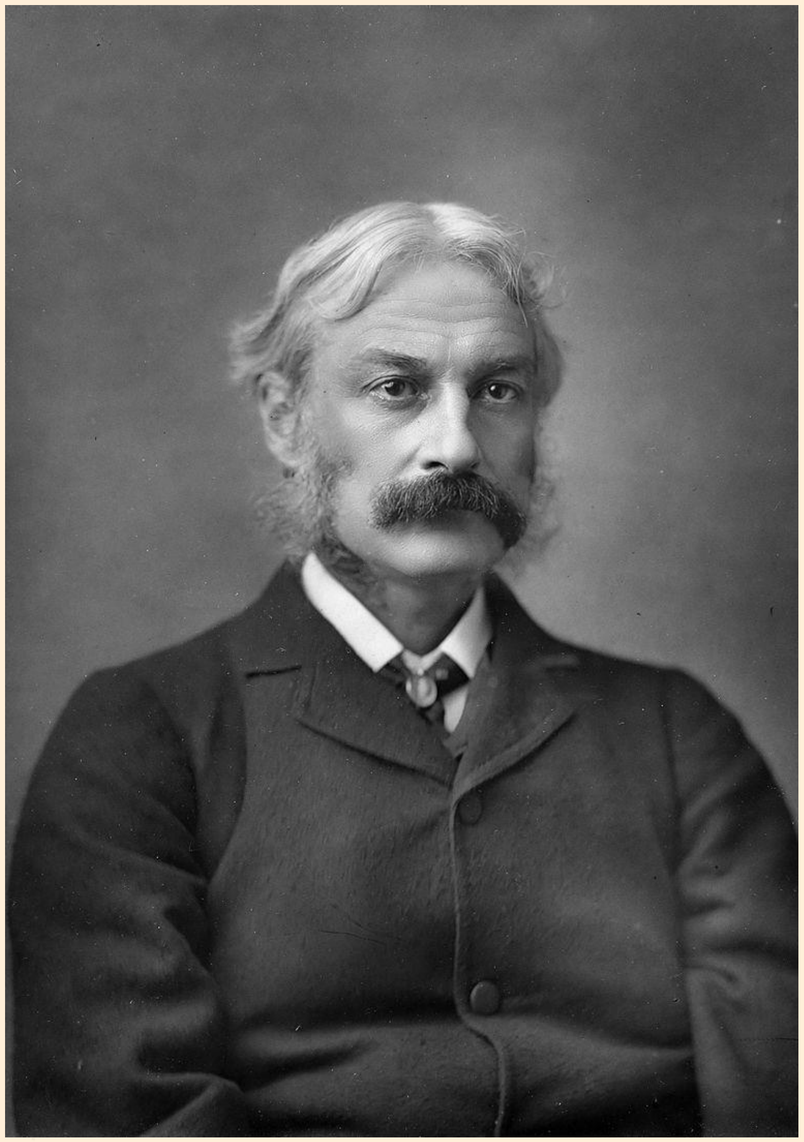
- *"the greatest value of a picture is when it forces us to notice what we never expected to see"*

- *"The best thing about being a statistician is that you get to play in everyone's backyard"*

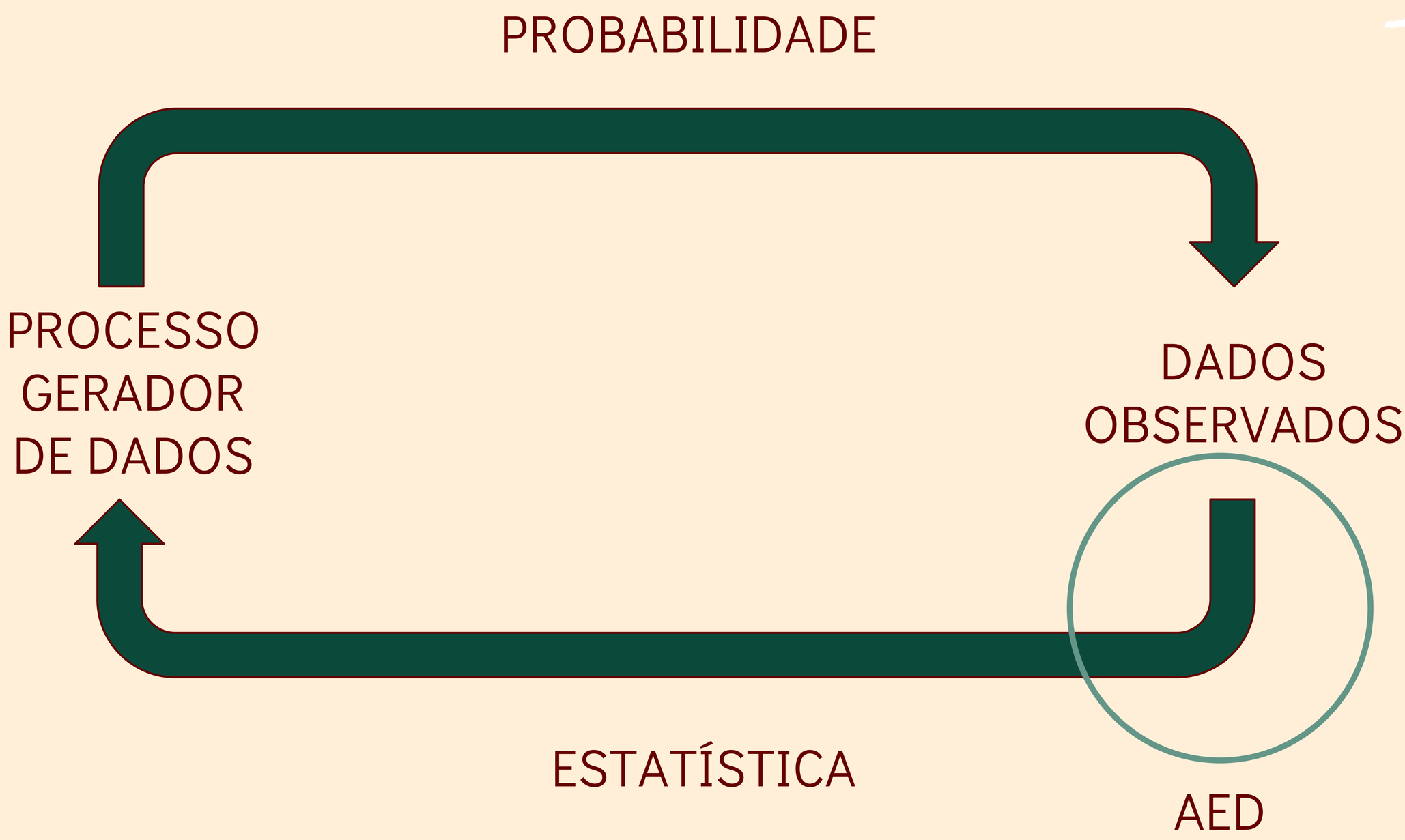
O ENSINAMENTO MAIS IMPORTANTE!

- "*[people] use statistics in the same way that a drunk uses lamp-posts – for support rather than illumination.*"

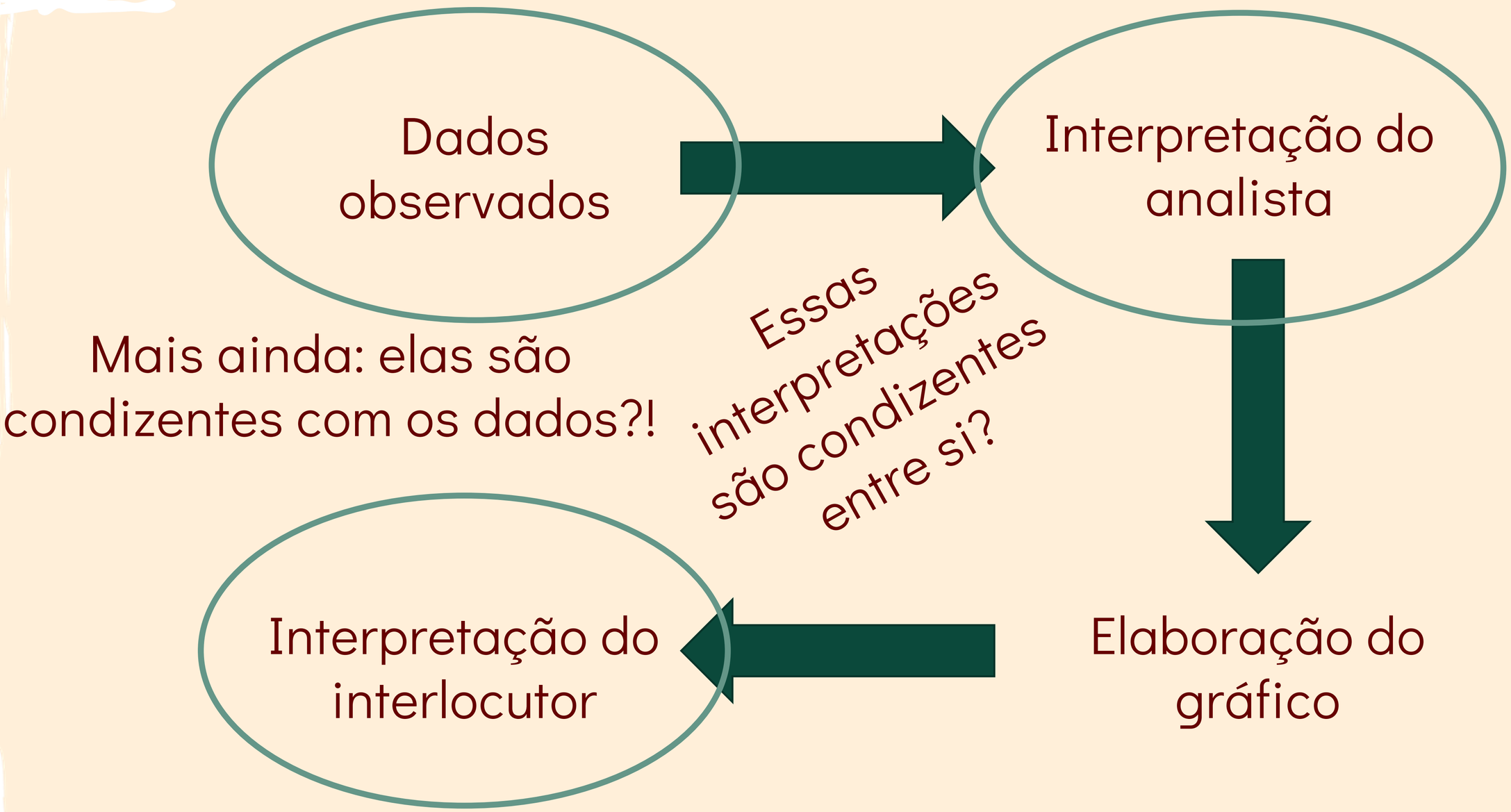
Andrew lang, escritor escocês.



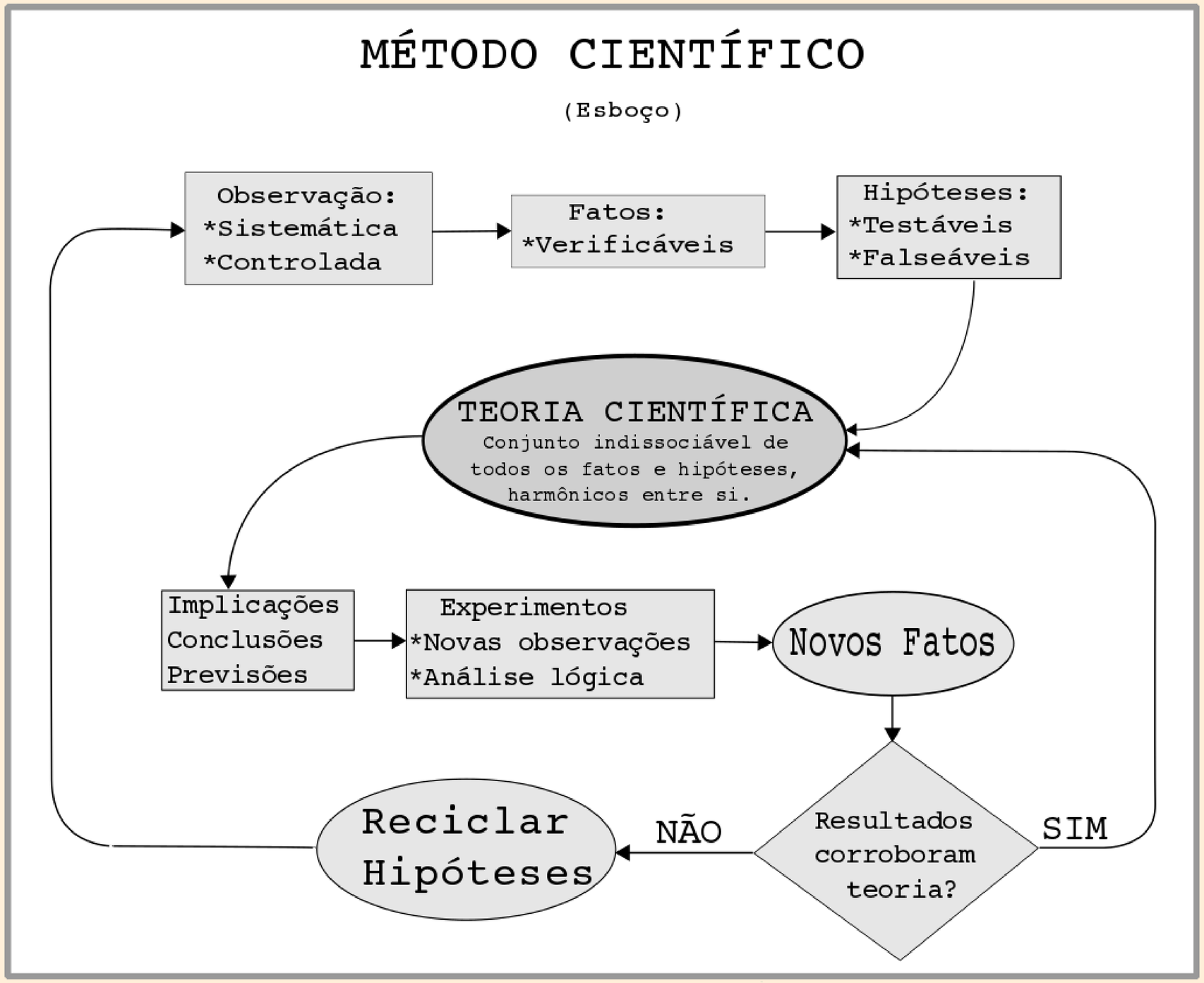
PROBABILIDADE VS. ESTATÍSTICA



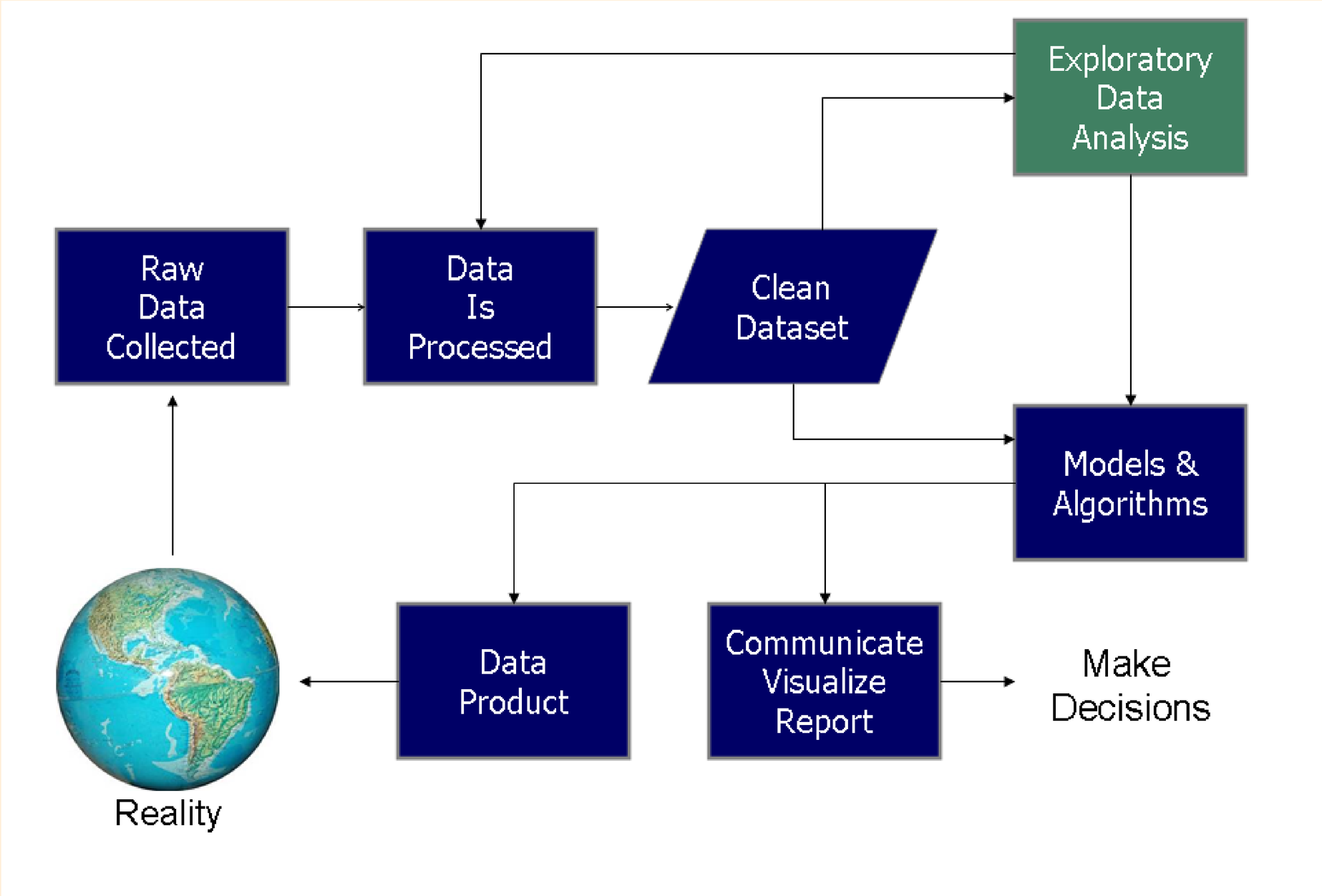
A IMPORTÂNCIA DE SUMÁRIOS VISUAIS DE DADOS



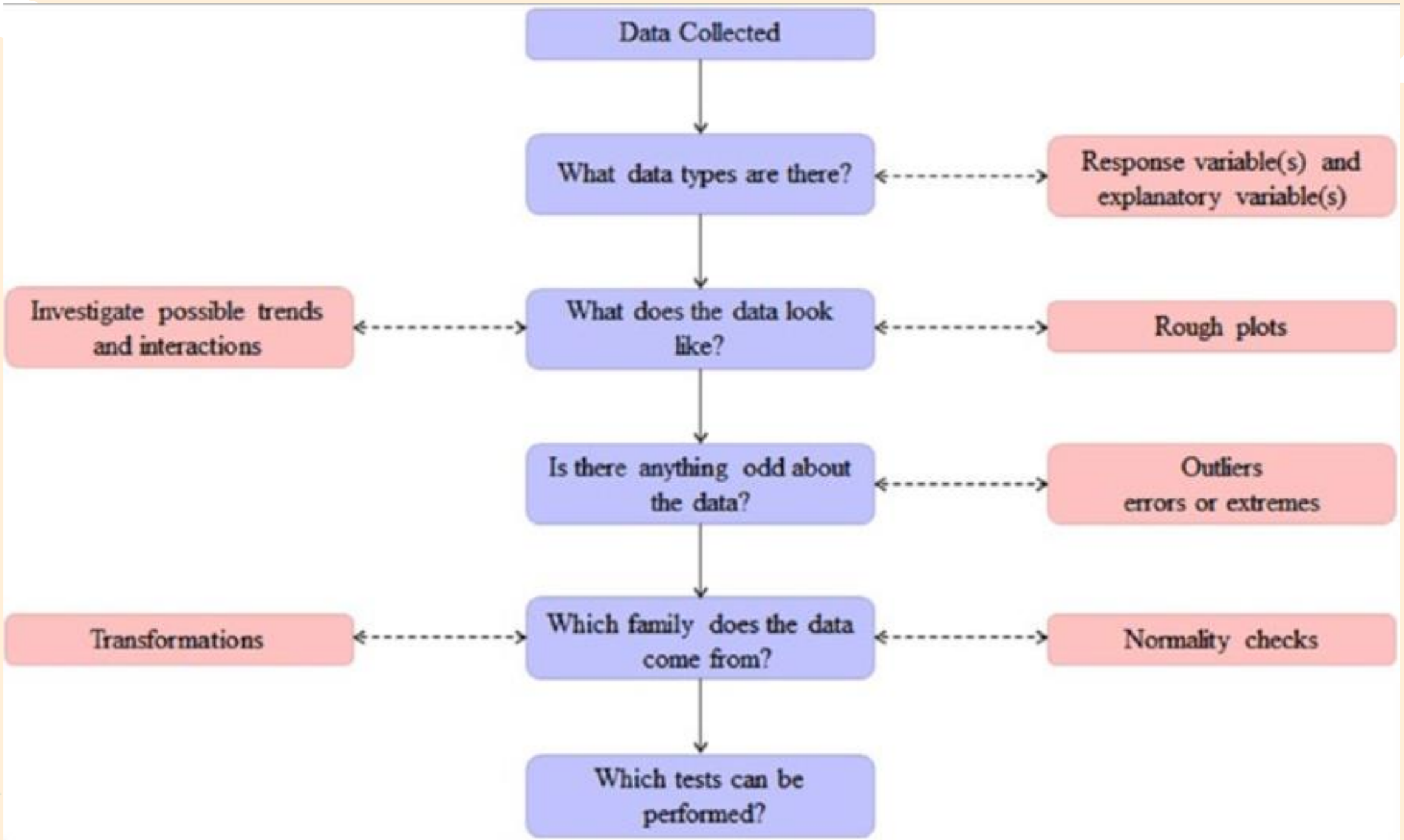
MÉTODO CIENTÍFICO – UM ESBOÇO



MÉTODO CIENTÍFICO – FOCO MAIOR NA ESTATÍSTICA



O PROCESSO DE UMA ANÁLISE EXPLORATÓRIA



FORMALIZANDO UM POUCO...

- Resumidamente:
  - Ciência = observação seguida de inferência/previsão

- Para fazer inferência/previsão, precisamos de um **modelo**

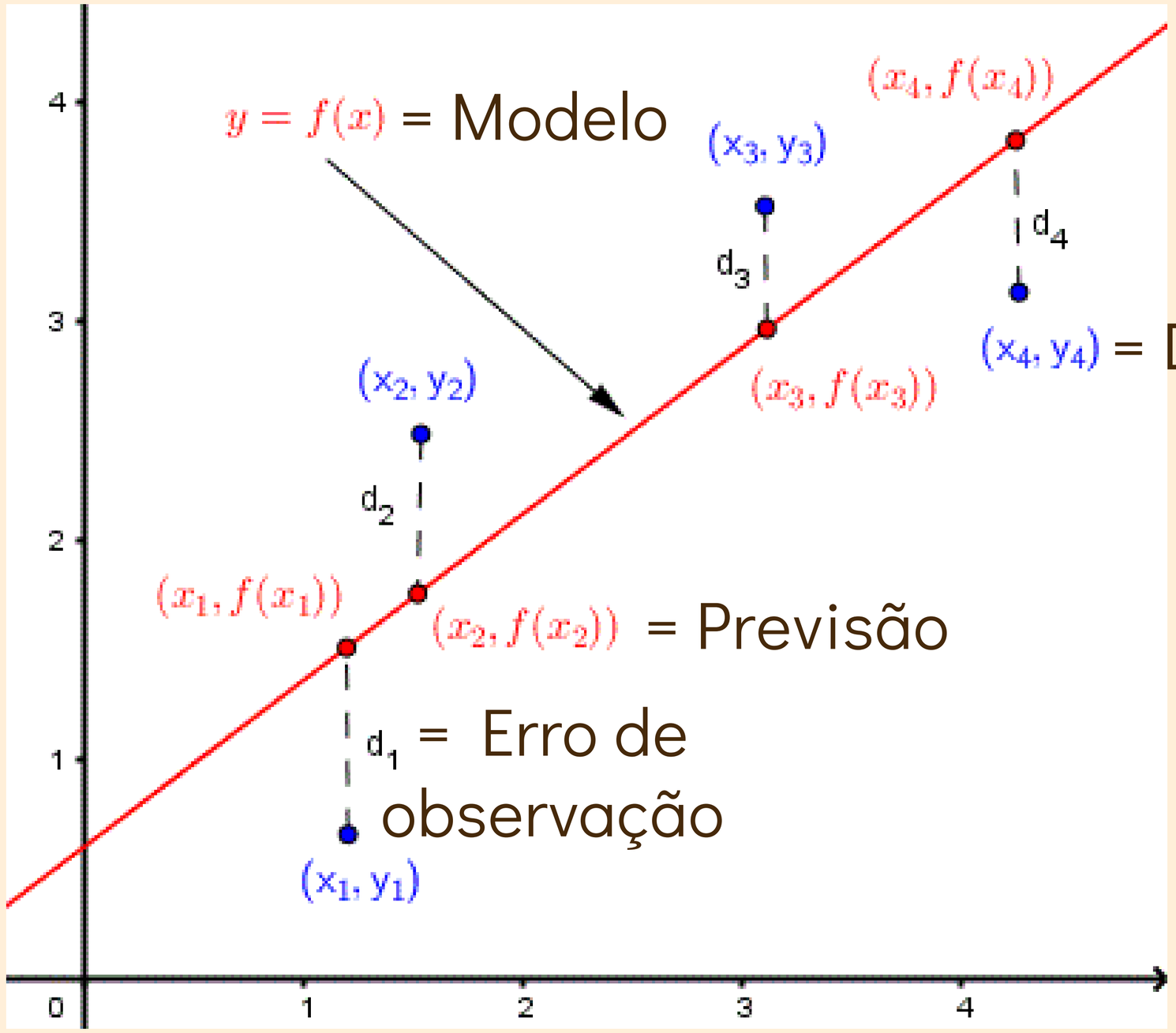
- Modelo = regularidade/padrão nos dados observados

$$dados = modelo + erro\ de\ observação$$

$$Y \text{ "=" } f(X) \text{ "+" } E$$

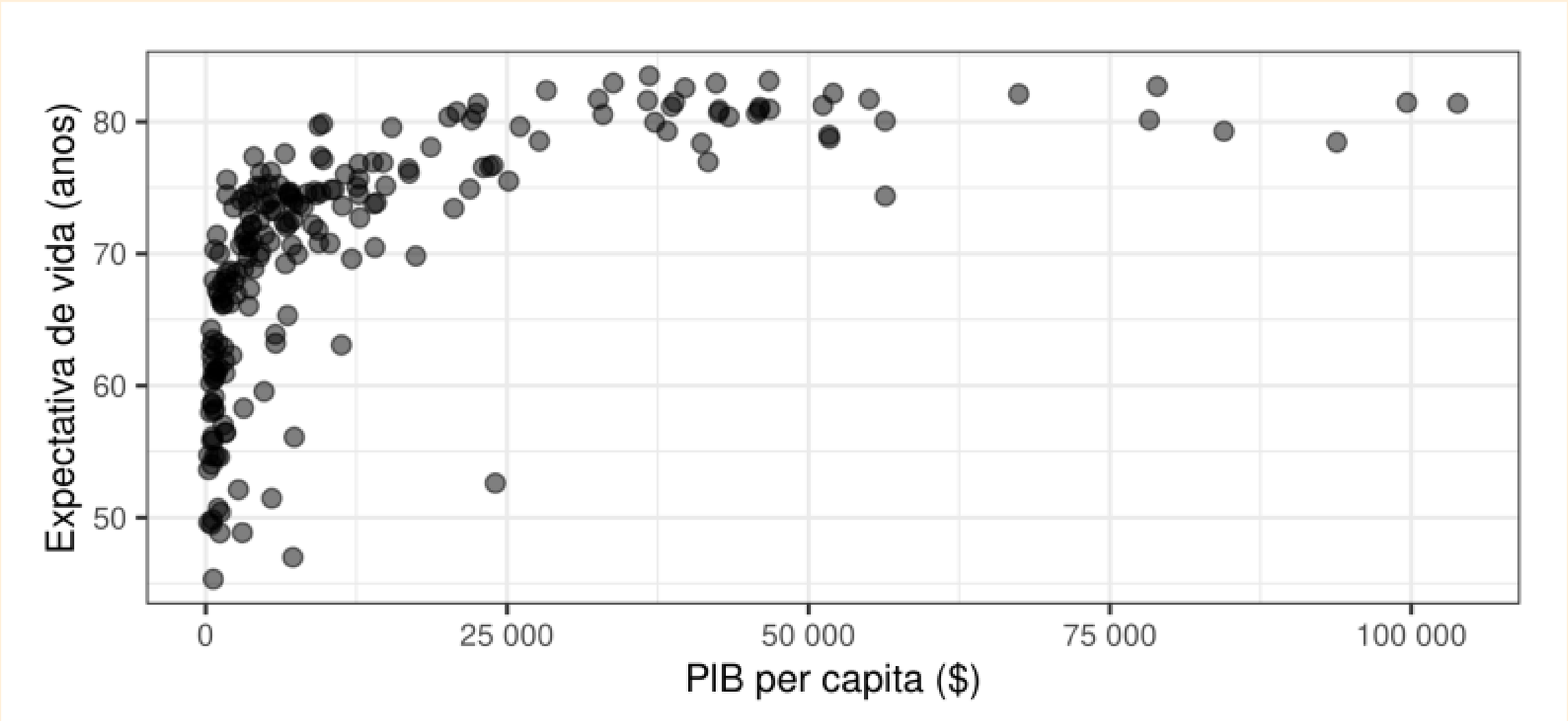
DADOS OBSERVADOS: APRENDIZADO SUPERVISIONADO

Variável resposta



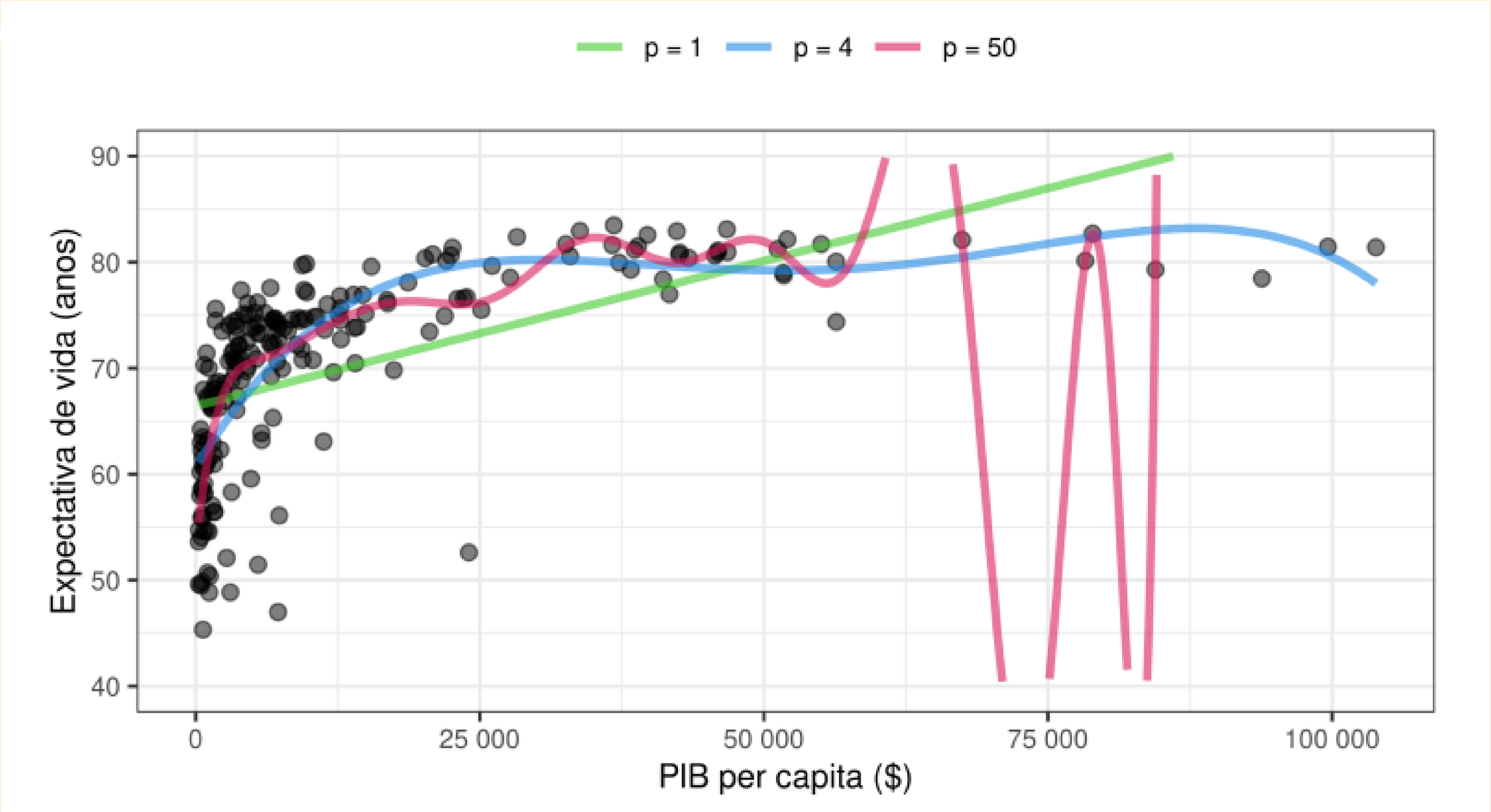
Variável preditora/atributo

⚠️ O ERRO DE OBSERVAÇÃO É FUNDAMENTAL!



Quão complexo deve ser o modelo (relação sistemática entre PIB per capita e expectativa de vida)?

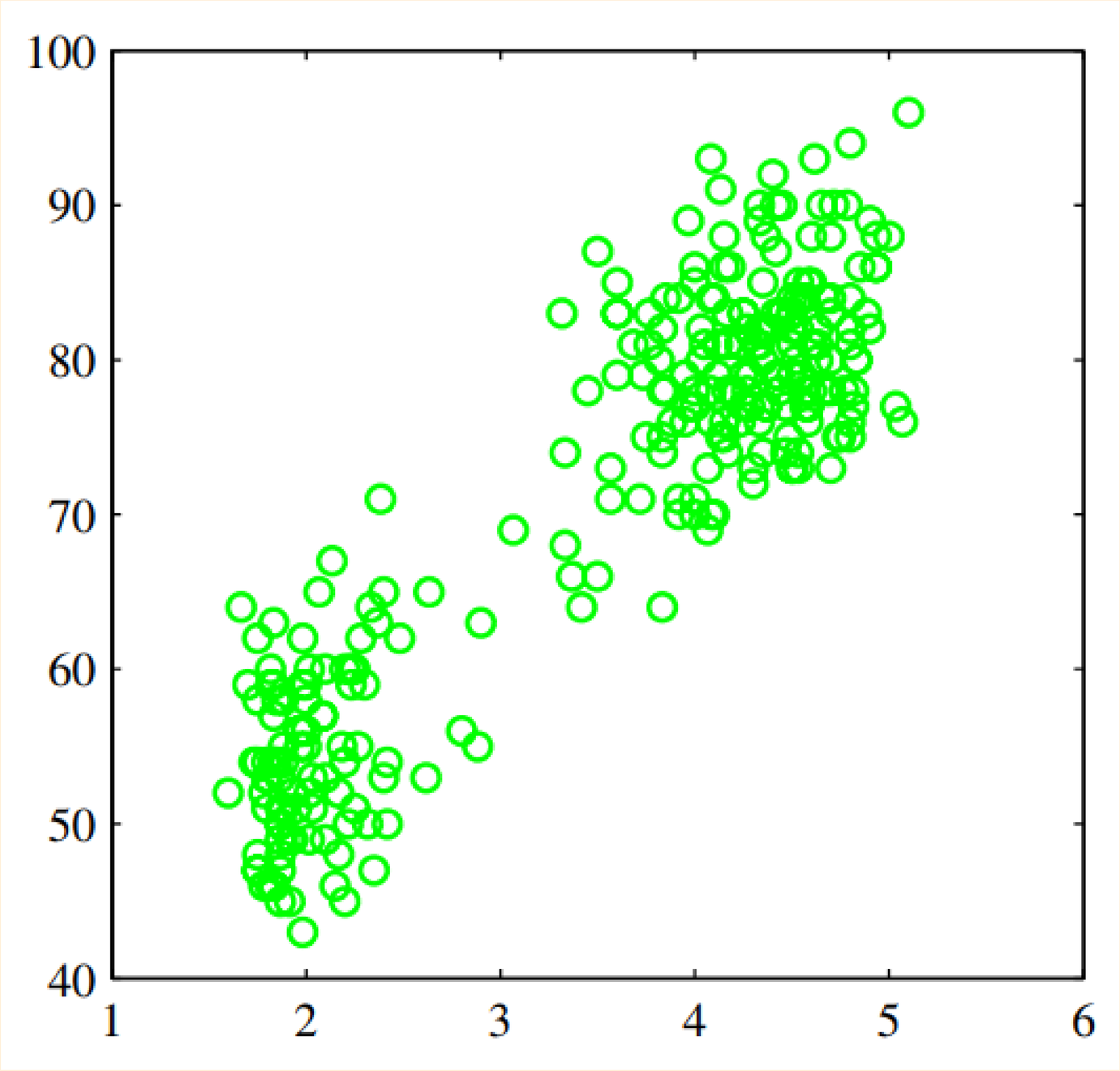
⚠️ O ERRO DE OBSERVAÇÃO É FUNDAMENTAL!



Modelos: polinômios de graus 1, 4 e 50.
Qual é o “melhor”?

DADOS OBSERVADOS: APRENDIZADO NÃO-SUPERVISIONADO

Intervalo até a próxima erupção (min)

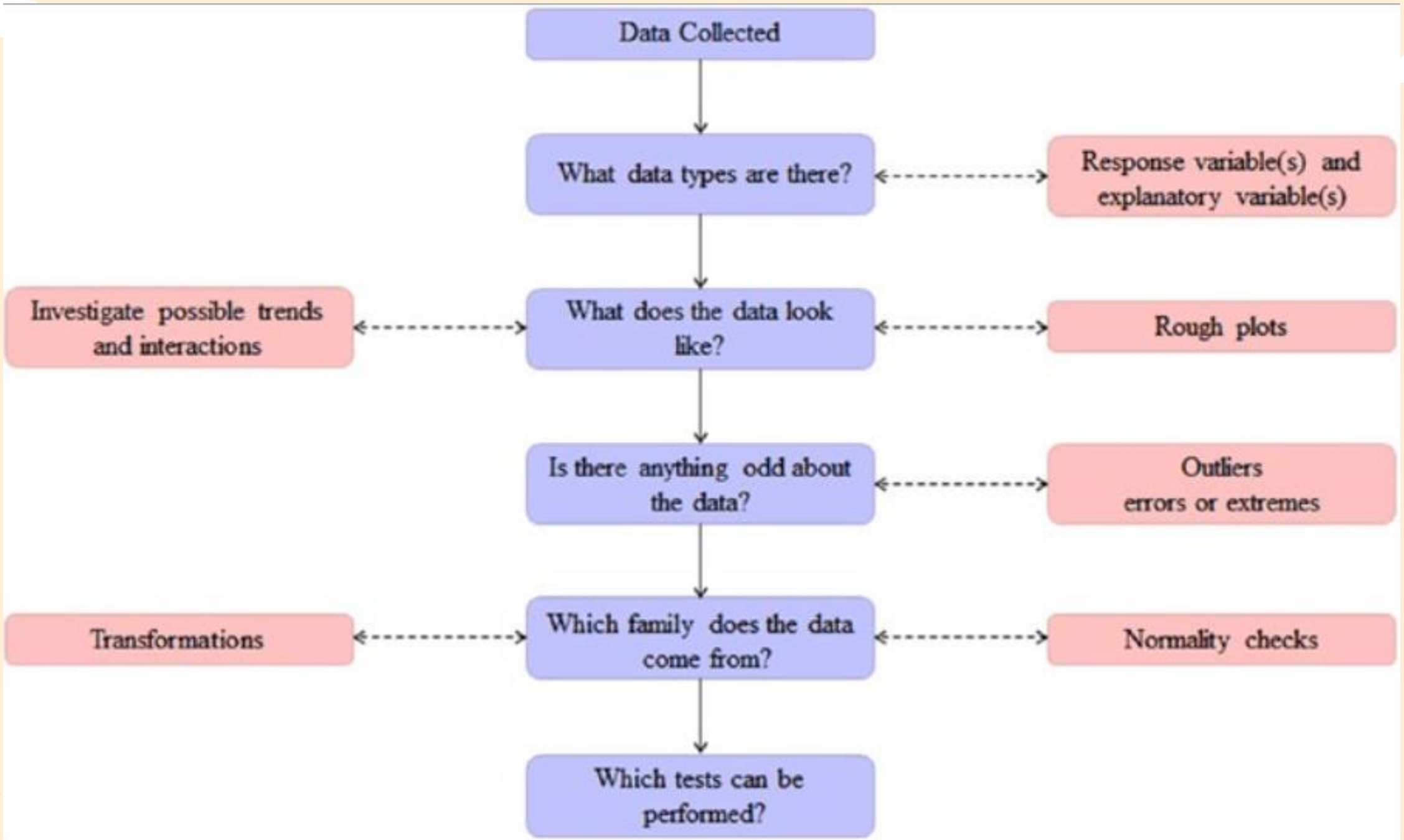


Quem é a variável resposta? Não é claro...

Objetivo: Extrair padrões de dados não-anotados (sem variável resposta clara)

Duração da erupção (min)

O PROCESSO DE UMA ANÁLISE EXPLORATÓRIA



TIPOLOGIA DE VARIÁVEIS

- QUALITATIVAS – qualidades associadas à observação
  - Nacionalidade
  - Sexo

  } Nominal

  - Grau de instrução
  - Classe social

  } Ordinal

⚠️ Estado civil: nominal ou ordinal?

TIPOLOGIA DE VARIÁVEIS

- QUANTITATIVAS – números/contagens associadas a medições
  - Salário
  - Massa
  - Altura

  } Contínua? 🤔

  - Número de filhos
  - Idade

  } Discreta? 🤔

"O contínuo é uma abstração do mundo real, que na verdade é discreto"

UM EXEMPLO

- PESQUISA: investigar aspectos socioeconômicos sobre funcionários de determinada empresa

Variável	Representação	
Estado civil	X	Qualitativa; nominal
Grau de instrução	Y	Qualitativa; ordinal
Número de filhos	Z	Quantitativa; discreta
Salário	S	Quantitativa; contínua
Idade	U	Quantitativa; discreta
Região de procedência	V	Qualitativa; nominal

Atributos

UM EXEMPLO

- PESQUISA: investigar aspectos socioeconômicos sobre funcionários de determinada empresa
  - População: conjunto de itens semelhantes, de interesse para determinado experimento ou análise
  - Amostragem?
  - Forma de coleta de dados?
  - Dados faltantes?

Variável	Representação
Estado civil	X
Grau de instrução	Y
Número de filhos	Z
Salário	S
Idade	U
Região de procedência	V

UM EXEMPLO – OS DADOS COLETADOS

Nº	Estado civil	Grau de instrução	Nº de filhos	Salário (× sal. mín.)	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	—	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	capital
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	—	5,73	20	10	outra
5	solteiro	ensino fundamental	—	6,26	40	07	outra
6	casado	ensino fundamental	0	6,66	28	00	interior
7	solteiro	ensino fundamental	—	6,86	41	00	interior
8	solteiro	ensino fundamental	—	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	capital
10	solteiro	ensino médio	—	7,44	23	06	outra

UM EXEMPLO – OS DADOS COLETADOS... MAS COM QUAL OBJETIVO?

Nº	Estado civil	Grau de instrução	Nº de filhos	Salário (× sal. mín.)	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	—	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	capital
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	—	5,73	20	10	outra

- Investigar "perfil médio" da empresa
- Estudar relações entre pares de variáveis (Nº de filhos e salário, p. ex.)
- Fazer projeções para plano de carreira (salário vs. idade)

⇒ Comportamento geral (individual ou conjunto) dos atributos

VAMOS FAZER NOSSAS ANÁLISES NESSES DADOS! PARA ISSO, PRECISAMOS DO R

R (programming language)

Article Talk

From Wikipedia, the free encyclopedia

R is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. Created by statisticians Ross Ihaka and Robert Gentleman, R is used among data miners, bioinformaticians and statisticians for data analysis and developing statistical software.[7] Users have created packages to augment the functions of the R language.

According to user surveys and studies of scholarly literature databases, R is one of the most commonly used programming languages in data mining.[8] As of December 2022, R ranks 11th in the TIOBE index, a measure of programming language popularity, in which the language peaked in 8th place in August 2020.[9][10]

The official R software environment is an open-source free software environment within the GNU package, available under the GNU General Public License. It is written primarily in C, Fortran, and R itself (partially self-hosting). Precompiled executables are provided for various operating systems. R has a command line interface.[11] Multiple third-party graphical user interfaces are also available, such as RStudio, an integrated development environment, and Jupyter, a notebook interface.

BAIXANDO O R

https://www.r-project.org

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Conferences
Search
Get Involved: Mailing Lists
Get Involved: Contributing
Developer Pages
R Blog

The R Project for Statistical Computing

Getting Started

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

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **R version 4.3.0 (Already Tomorrow) prerelease versions** will appear starting Tuesday 2023-03-21. Final release is scheduled for Friday 2023-04-21.
- **R version 4.2.3 (Shortstop Beagle)** has been released on 2023-03-15.
- **R version 4.1.3 (One Push-Up)** was released on 2022-03-10.
- Thanks to the organisers of useR! 2020 for a successful online conference. Recorded tutorials and talks from the conference are available on the R Consortium YouTube channel.
- You can support the R Foundation with a renewable subscription as a supporting member

BAIXANDO O R

CRAN Mirrors

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: main page, windows release, windows old release.

If you want to host a new mirror at your institution, please have a look at the CRAN Mirror HOWTO.

0-Cloud	
https://cloud.r-project.org/	Automatic redirection to servers worldwide, currently sponsored by Rstudio
Argentina	
http://mirror.fcaglp.unlp.edu.ar/CRAN/	Universidad Nacional de La Plata
Australia	
https://cran.csiro.au/	CSIRO
https://mirror.aarnet.edu.au/pub/CRAN/	AARNET
https://cran.ms.unimelb.edu.au/	School of Mathematics and Statistics, University of Melbourne
https://cran.curtin.edu.au/	Curtin University
Austria	
https://cran.wu.ac.at/	Wirtschaftsuniversität Wien
Belgium	
https://www.freestatistics.org/cran/	Patrick Wessa
https://ftp.belnet.be/mirror/CRAN/	Belnet, the Belgian research and education network
Brazil	
https://cran-r.c3sl.ufpr.br/	Universidade Federal do Parana
https://cran.fiocruz.br/	Oswaldo Cruz Foundation, Rio de Janeiro
https://vps.fmvz.usp.br/CRAN/	University of Sao Paulo, Sao Paulo
https://brieger.esalq.usp.br/CRAN/	University of Sao Paulo, Piracicaba

Você pode selecionar qualquer *mirror*,
mas os do Brasil serão mais rápidos.

BAIXANDO O R

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

Escolha o seu sistema operacional

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2023-03-15, Shortstop Beagle) R-4.2.3.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

BAIXANDO O R (PARA WINDOWS)

R for Windows

Subdirectories:

base	Binaries for base distribution. This is what you want to **install R for the first time**.
contrib	Binaries of contributed CRAN packages (for R >= 3.4.x).
old contrib	Binaries of contributed CRAN packages for outdated versions of R (for R < 3.4.x).
Rtools	Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

You may also want to read the R FAQ and R for Windows FAQ.

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

BAIXANDO O R (PARA WINDOWS)

R-4.2.3 for Windows

Download R-4.2.3 for Windows (77 megabytes, 64 bit)

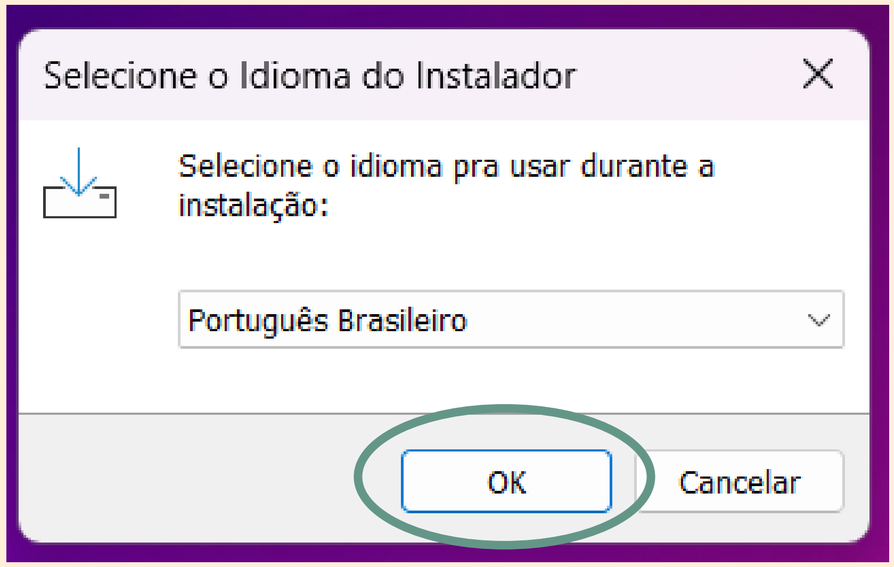
README on the Windows binary distribution

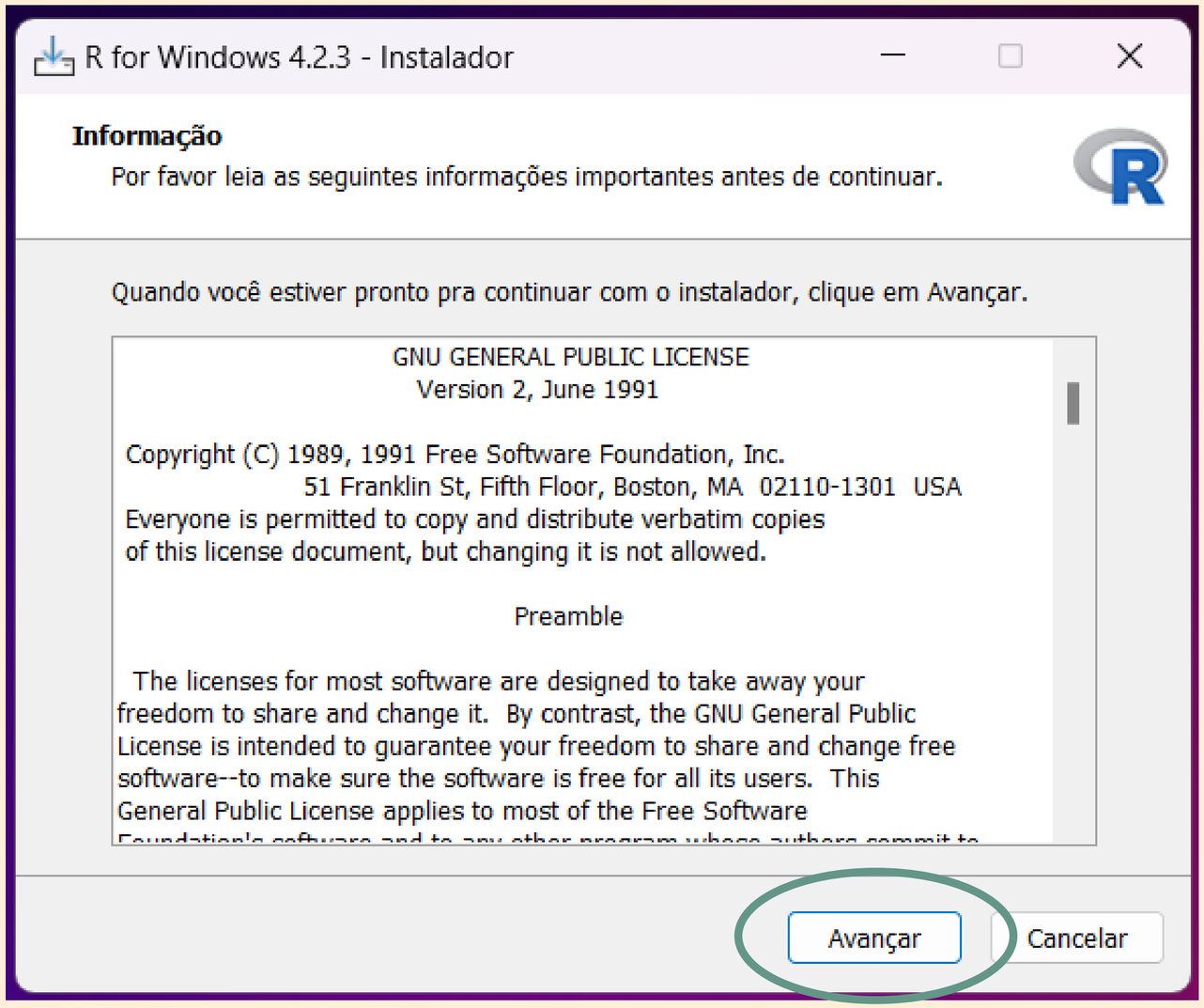
New features in this version

This build requires UCRT, which is part of Windows since Windows 10 and Windows Server 2016. On older systems, UCRT has to be installed manually from here.

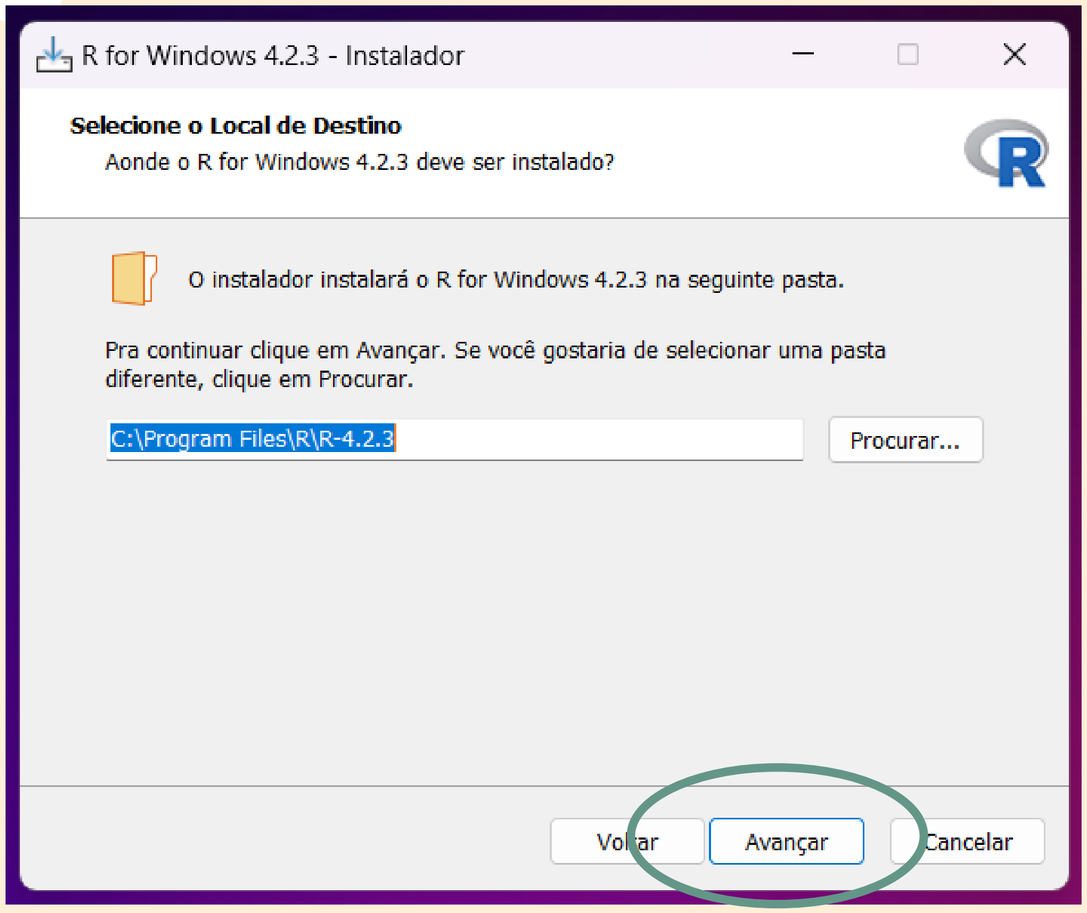
If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server.

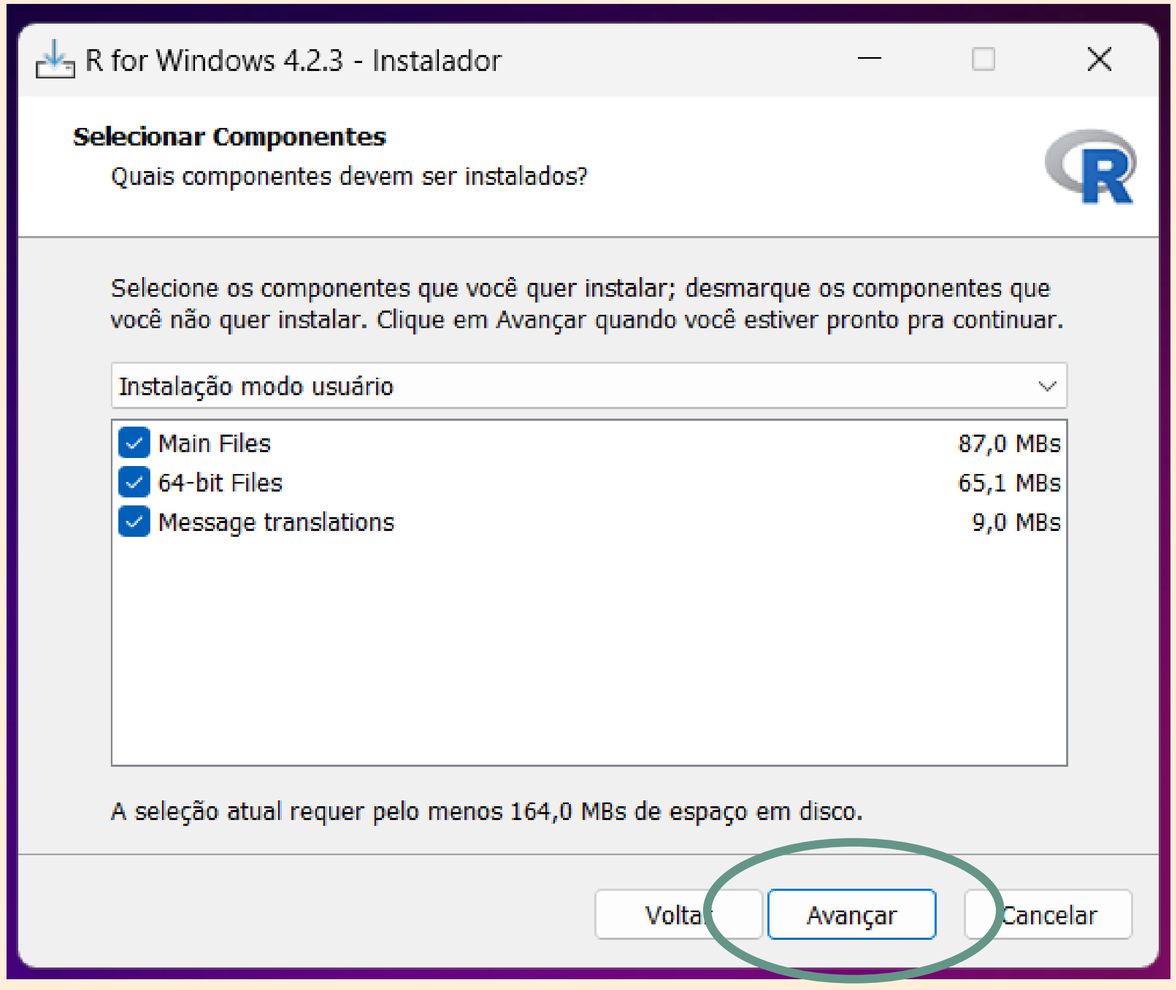
INSTALANDO O R (NO WINDOWS)



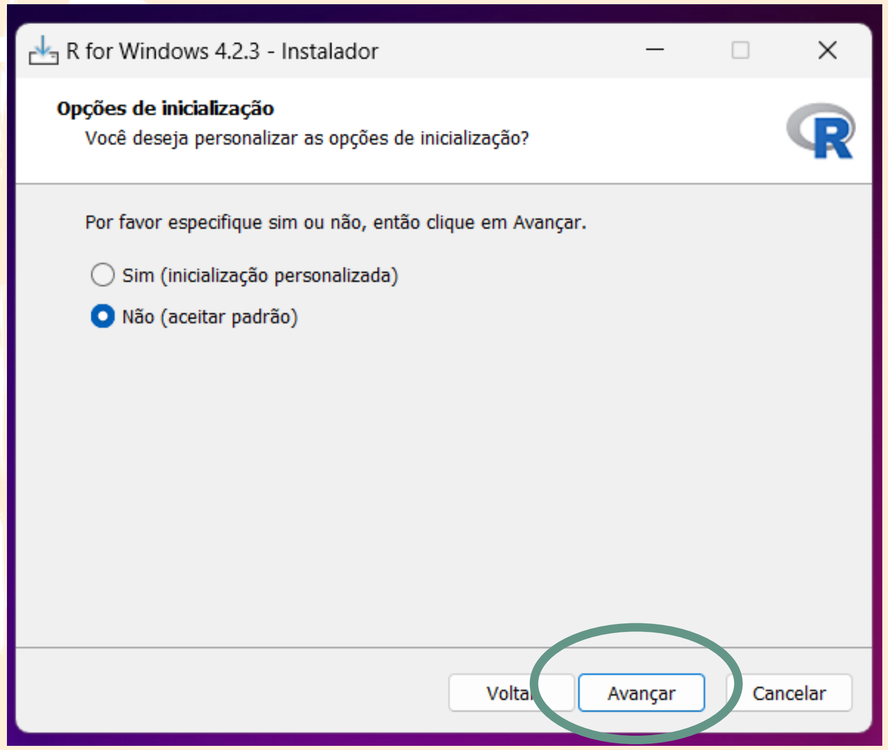


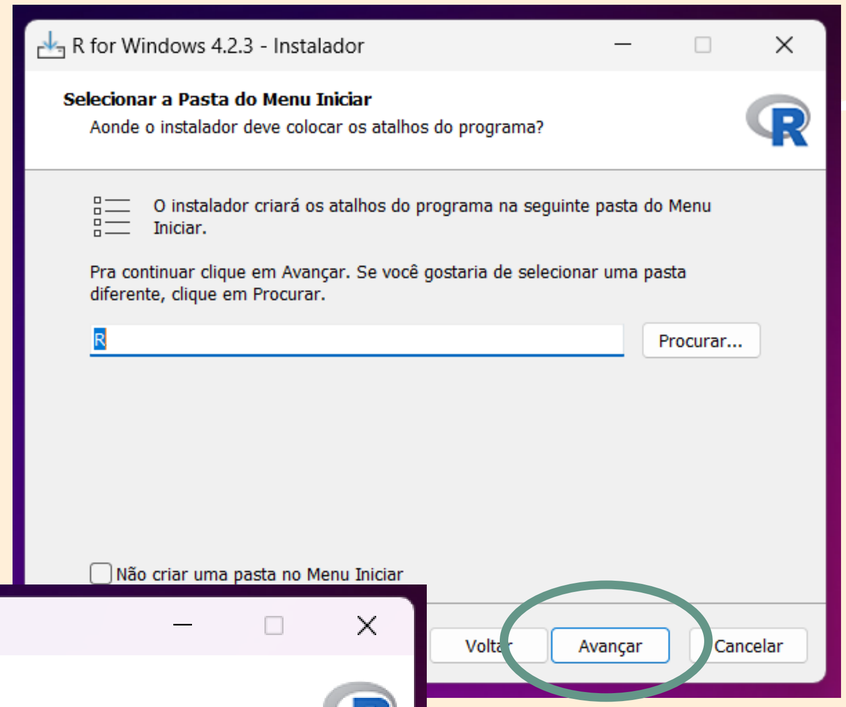
INSTALANDO O R (NO WINDOWS)

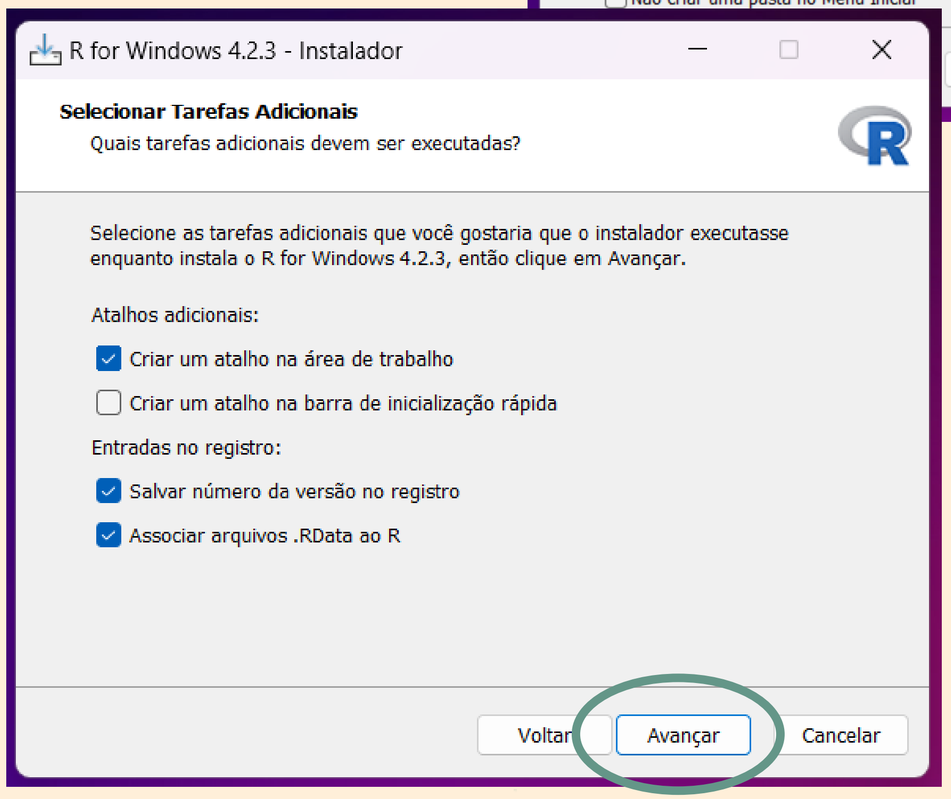




INSTALANDO O R (NO WINDOWS)

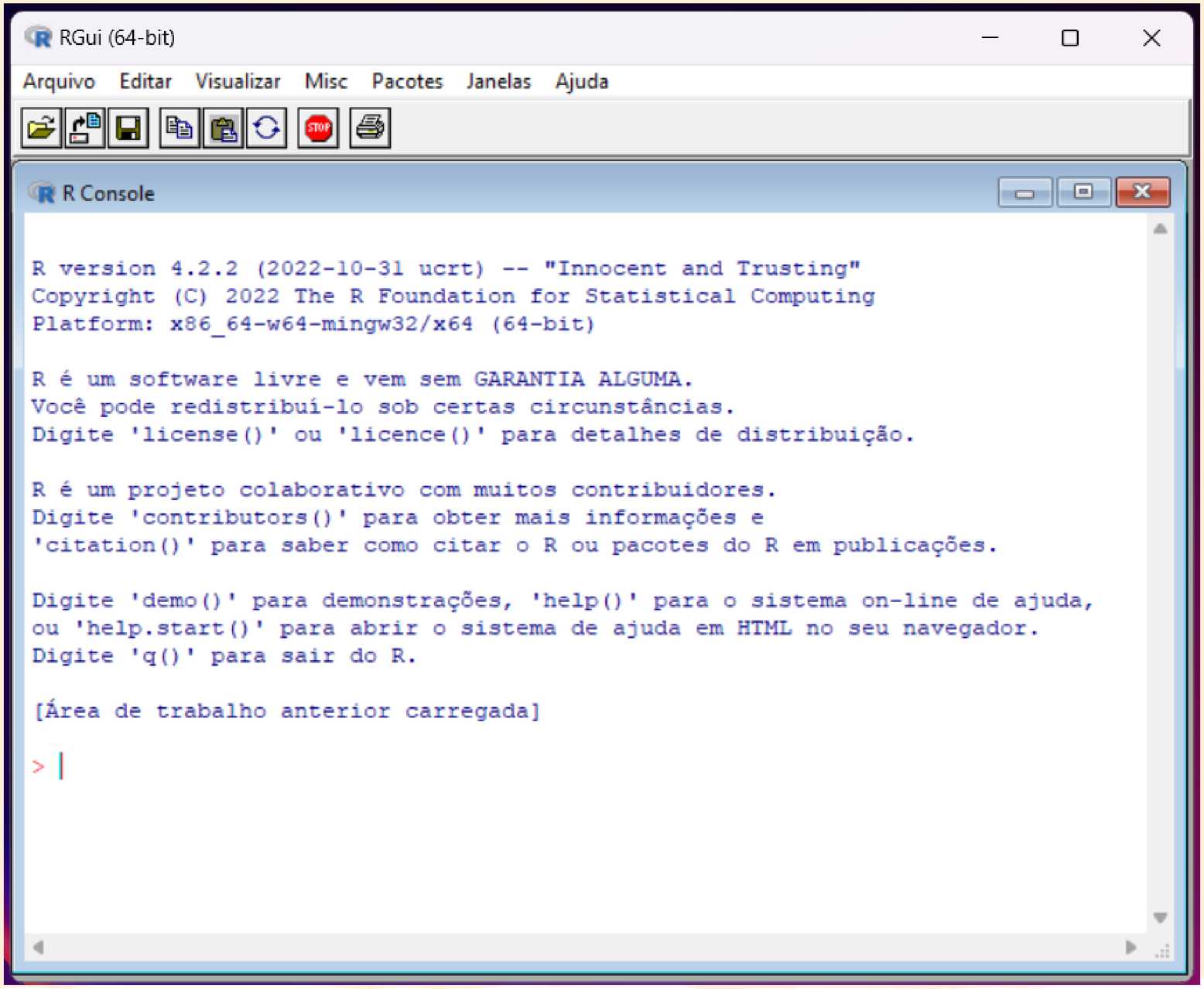






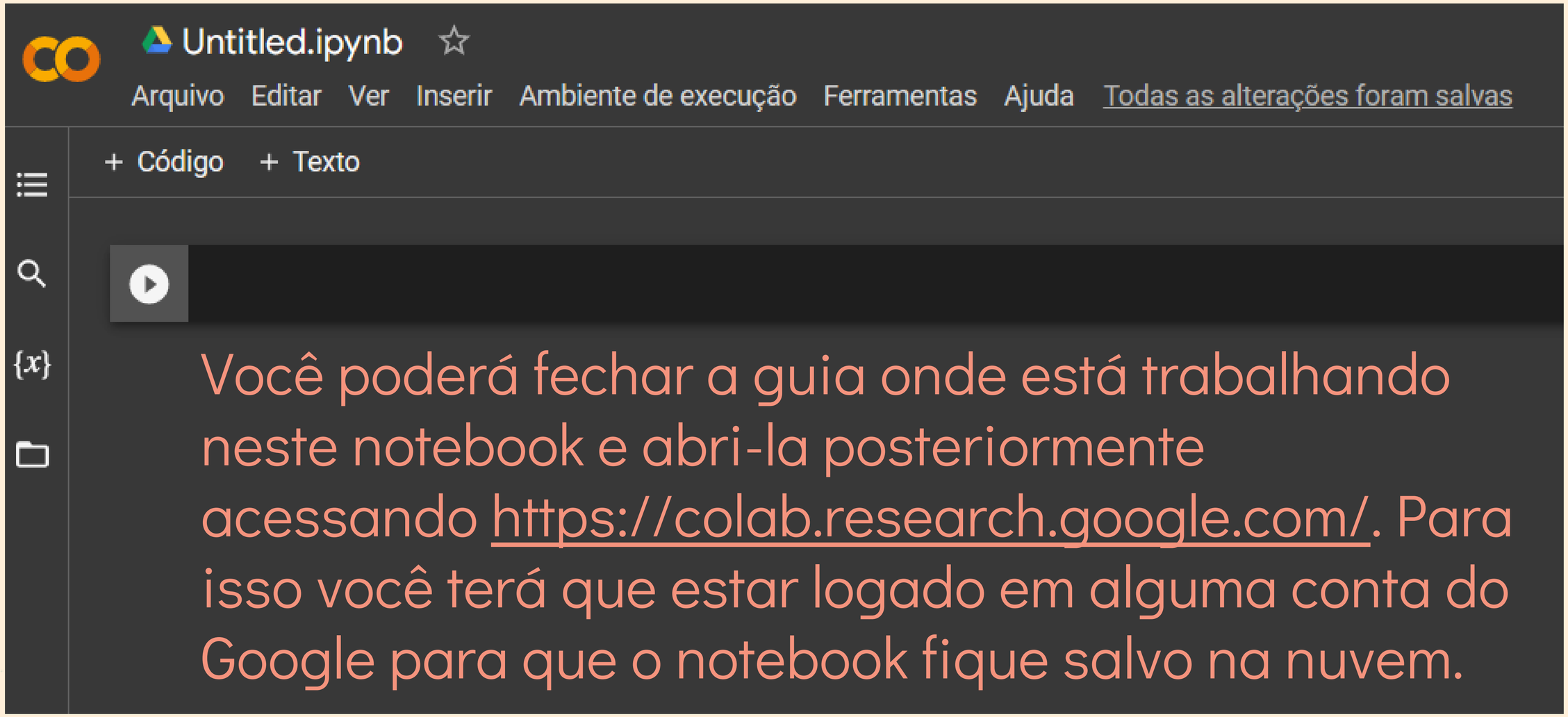
RODANDO O R (LOCALMENTE)

No Windows: Iniciar > Programas > R > R 4.2.3



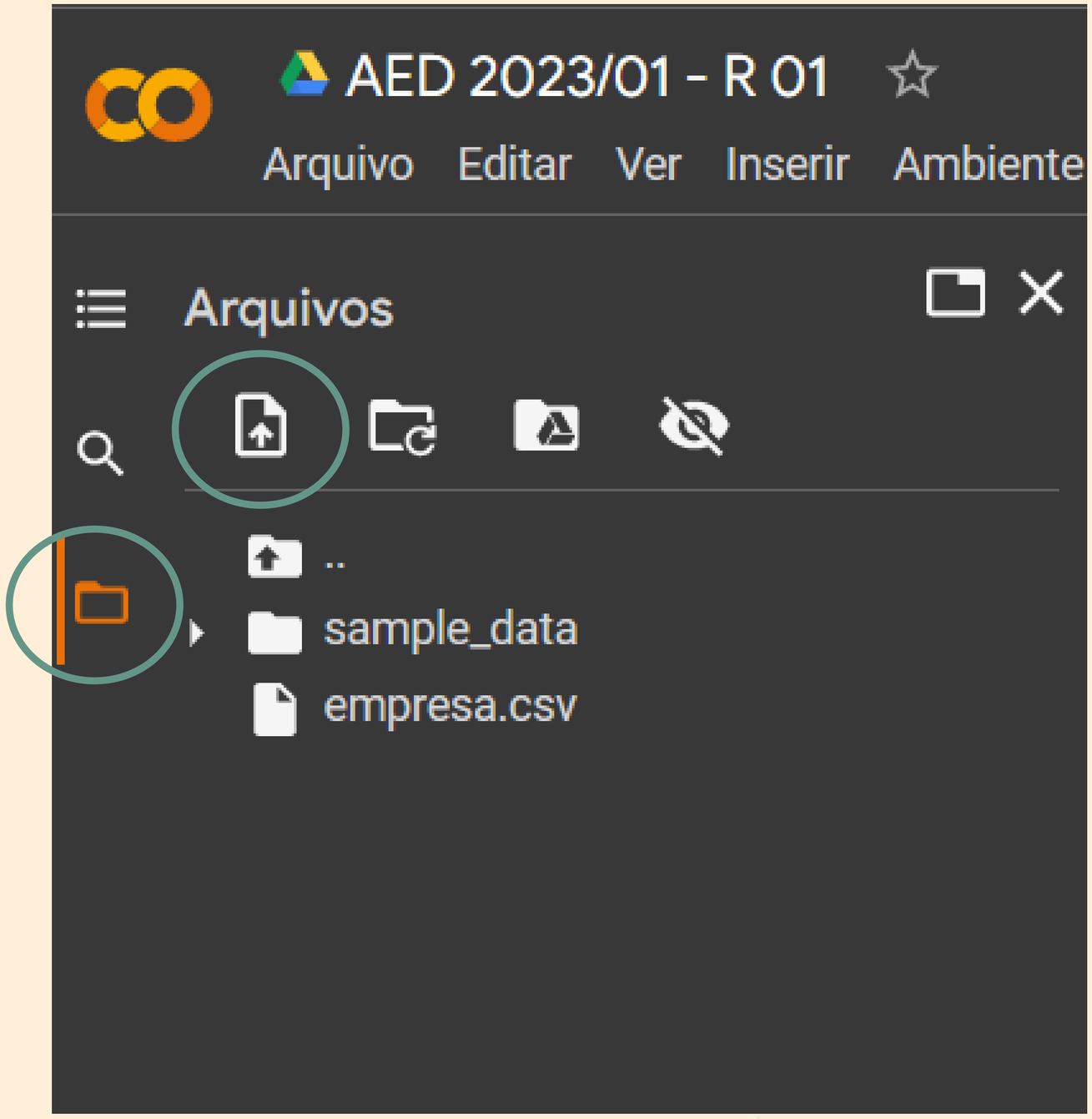
RODANDO O R (NA NUVEM)

Entre no link:
https://colab.research.google.com/notebook#create=true&language=r



Você poderá fechar a guia onde está trabalhando neste notebook e abri-la posteriormente acessando https://colab.research.google.com/. Para isso você terá que estar logado em alguma conta do Google para que o notebook fique salvo na nuvem.

ENVIANDO ARQUIVO PARA A NUVEM



INTRODUÇÃO AO R E RESUMOS EM TABELAS

O *notebook* de nome "**AED_R-01.ipynb**" traz esse conteúdo.

GRÁFICOS PARA VARIÁVEIS QUALITATIVAS

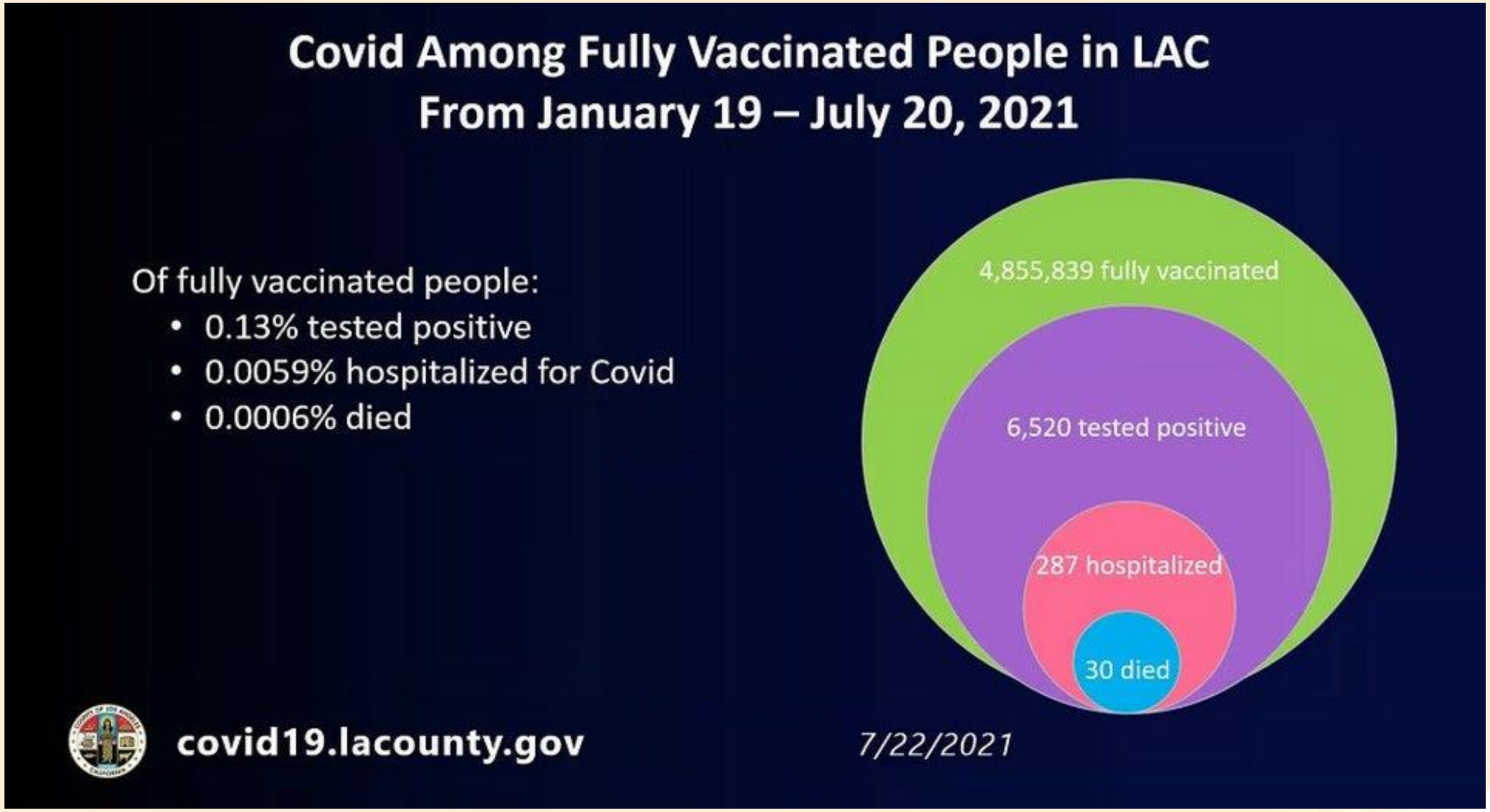
O *notebook* de nome "**AED_R-02.ipynb**" traz esse conteúdo.

⚠️ 🚨 ALERTA 🚨 ⚠️

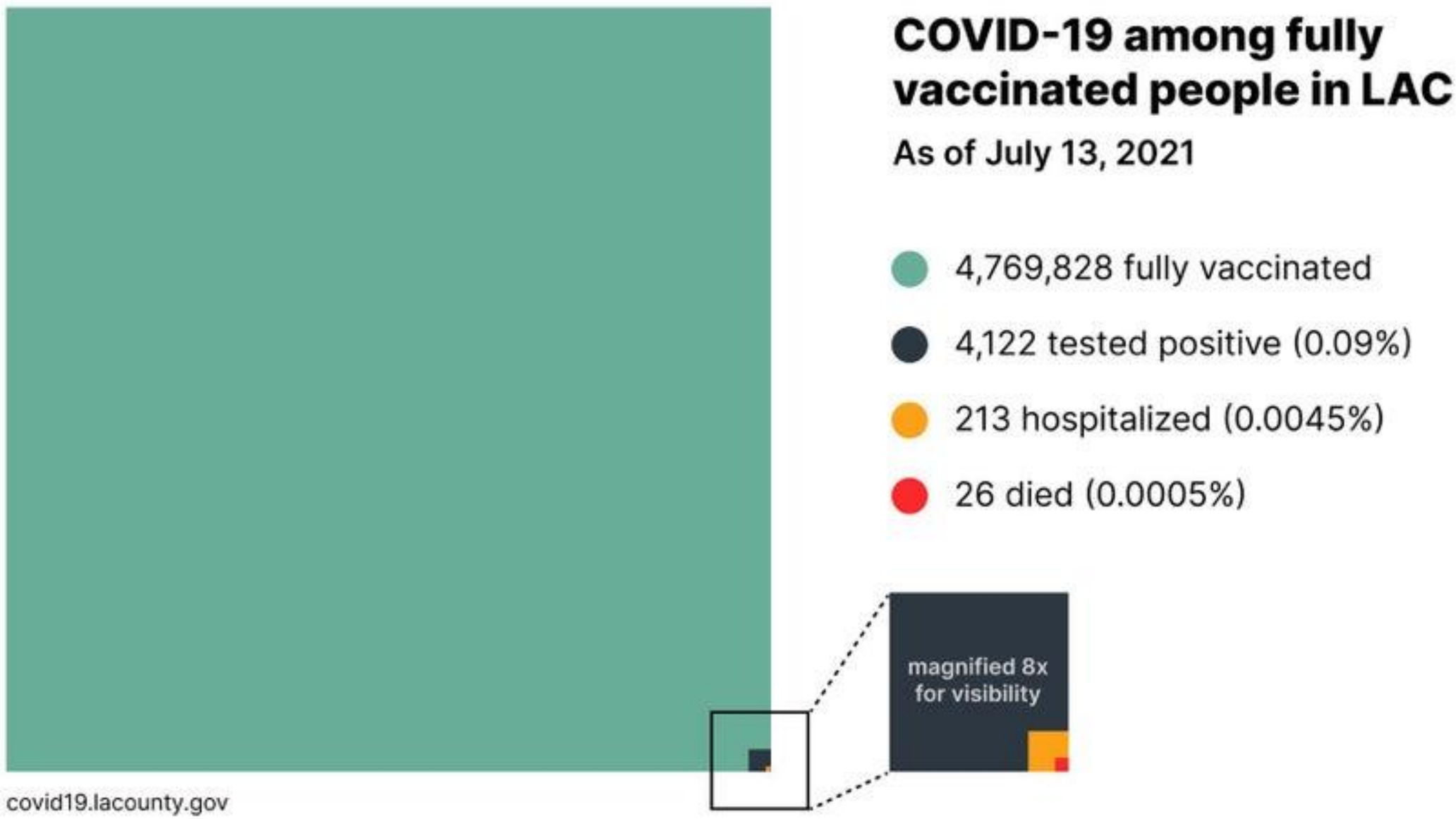
Qualquer tipo de sumário numérico ou visual pode ser (e provavelmente será...) extremamente mal utilizado!

Vejamos alguns exemplos:

UM "GRÁFICO DE BARRAS" ESTRANHO...
(E NÃO É POR NÃO TER BARRAS!)

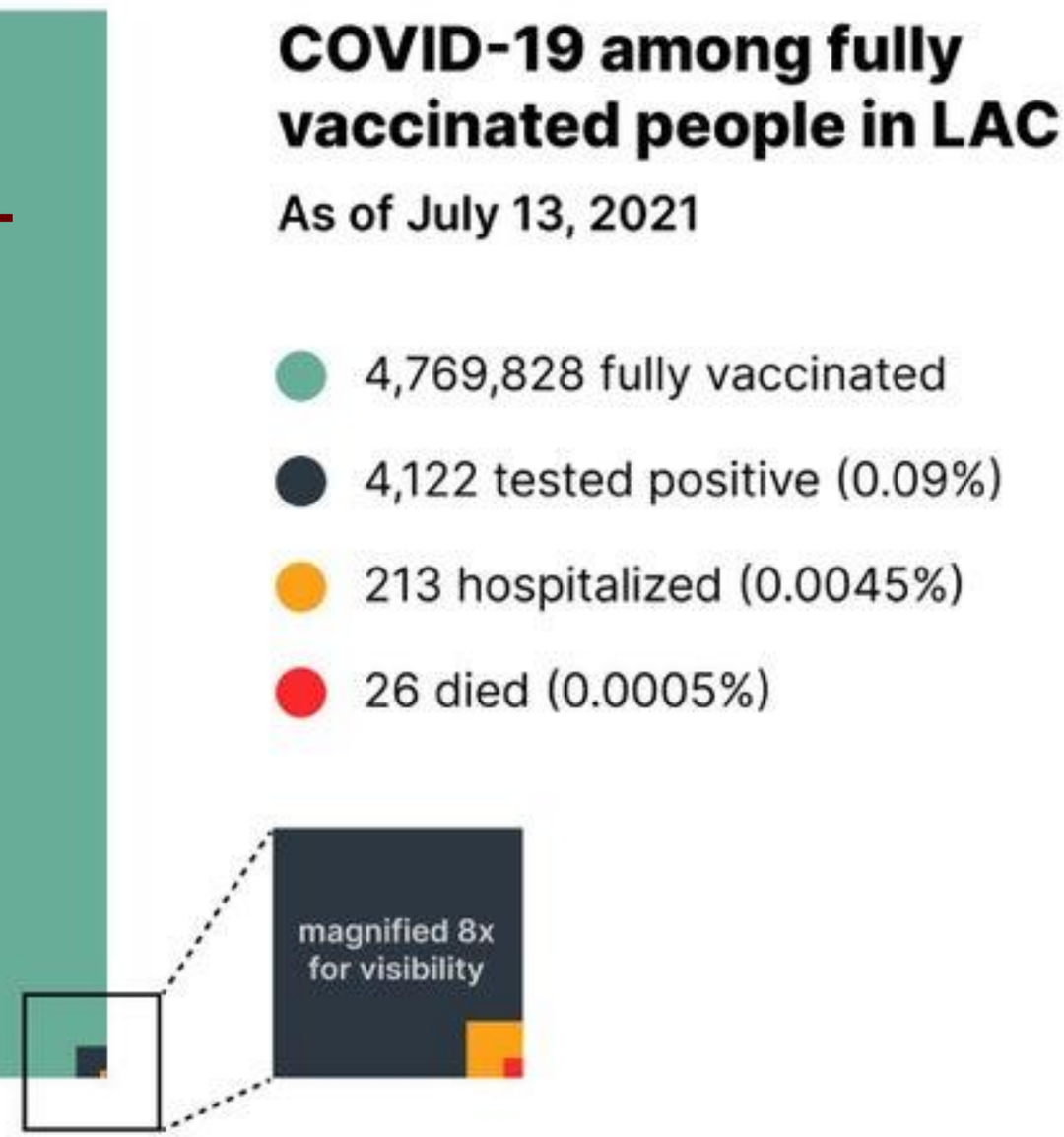


...E UMA MELHOR VISUALIZAÇÃO DESSA INFORMAÇÃO



...E UMA MELHOR VISUALIZAÇÃO DESSA INFORMAÇÃO

Atenção! Uma comparação com não-vacinados é importante, e crucial para inferir relações causais! E principalmente para convencer negacionistas! :-)

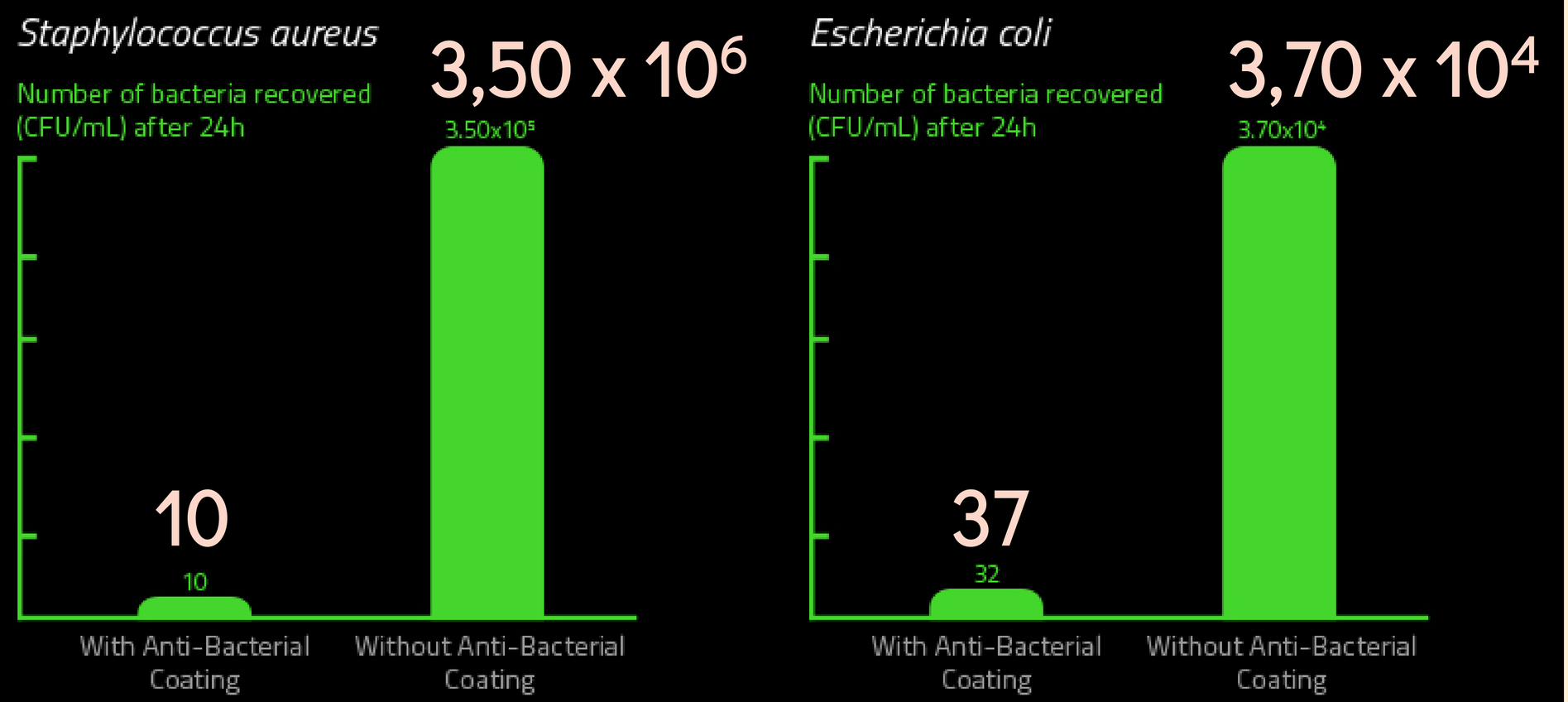


covid19.lacounty.gov

CUIDADO COM A ESCALA!

ANTI-BACTERIAL COATING

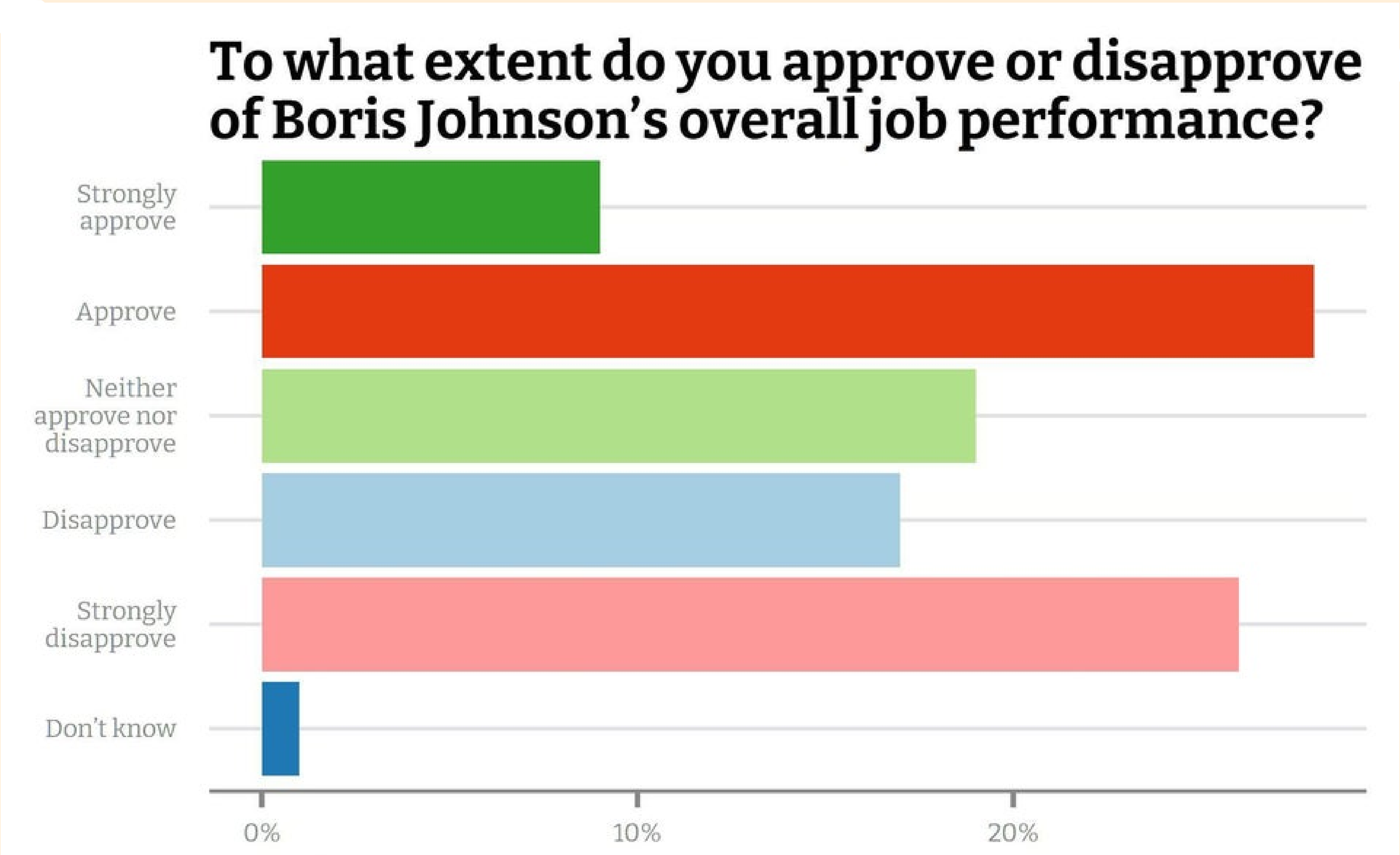
As your phone is a commonly used device that can come into contact with various surfaces or hands during daily use, the case comes with a layer of protection that prevents the growth of bacteria.



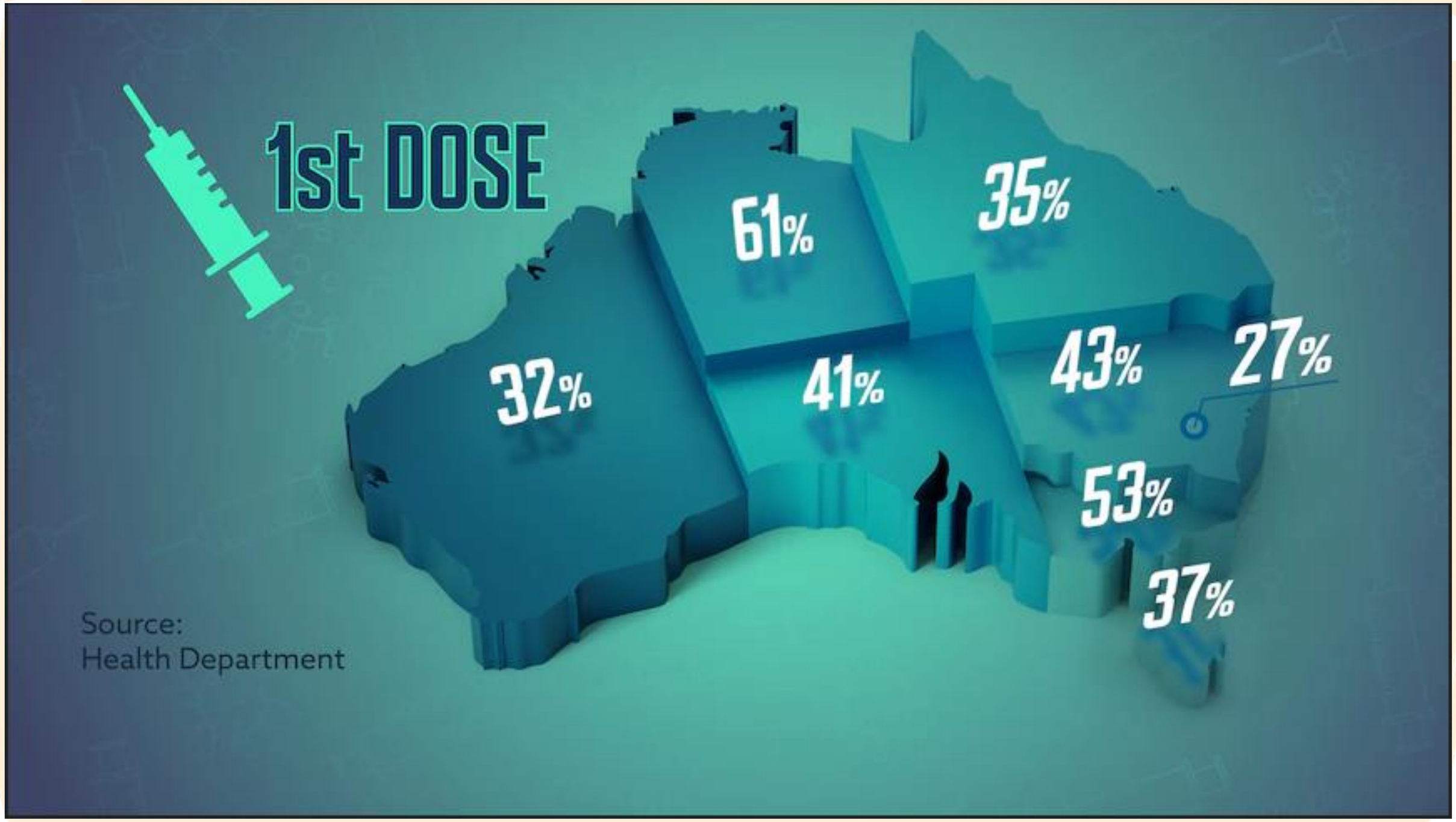
Barras não estão em escala 🙈

**Graph bars not drawn to scale*

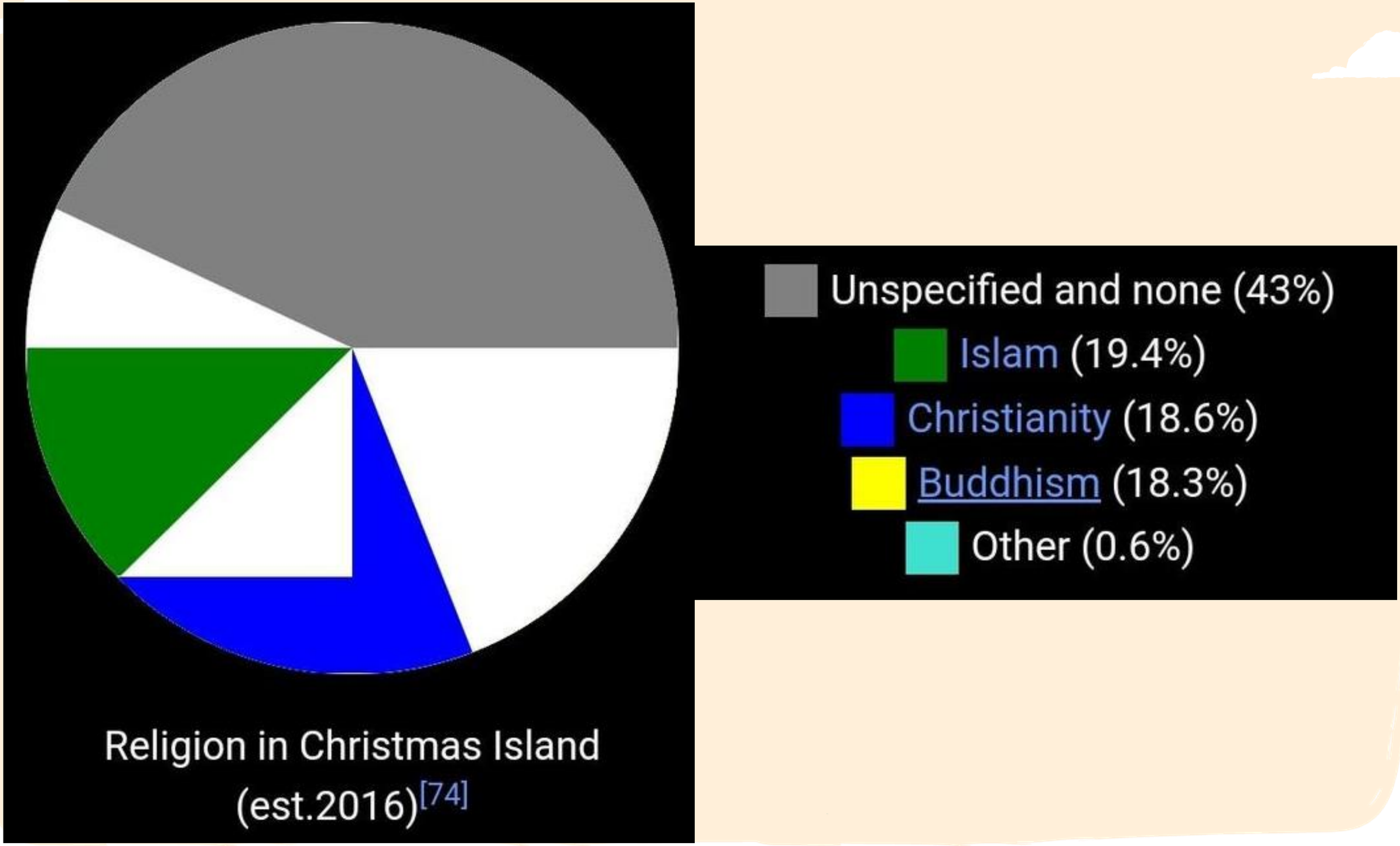
UMA BOA INTENÇÃO COM CORES PODE DAR MUITO ERRADO



MAS... E O TAMANHO DESSAS BARRAS?! 🤯



PIZZA CORTADA À MODA PSICOPATA

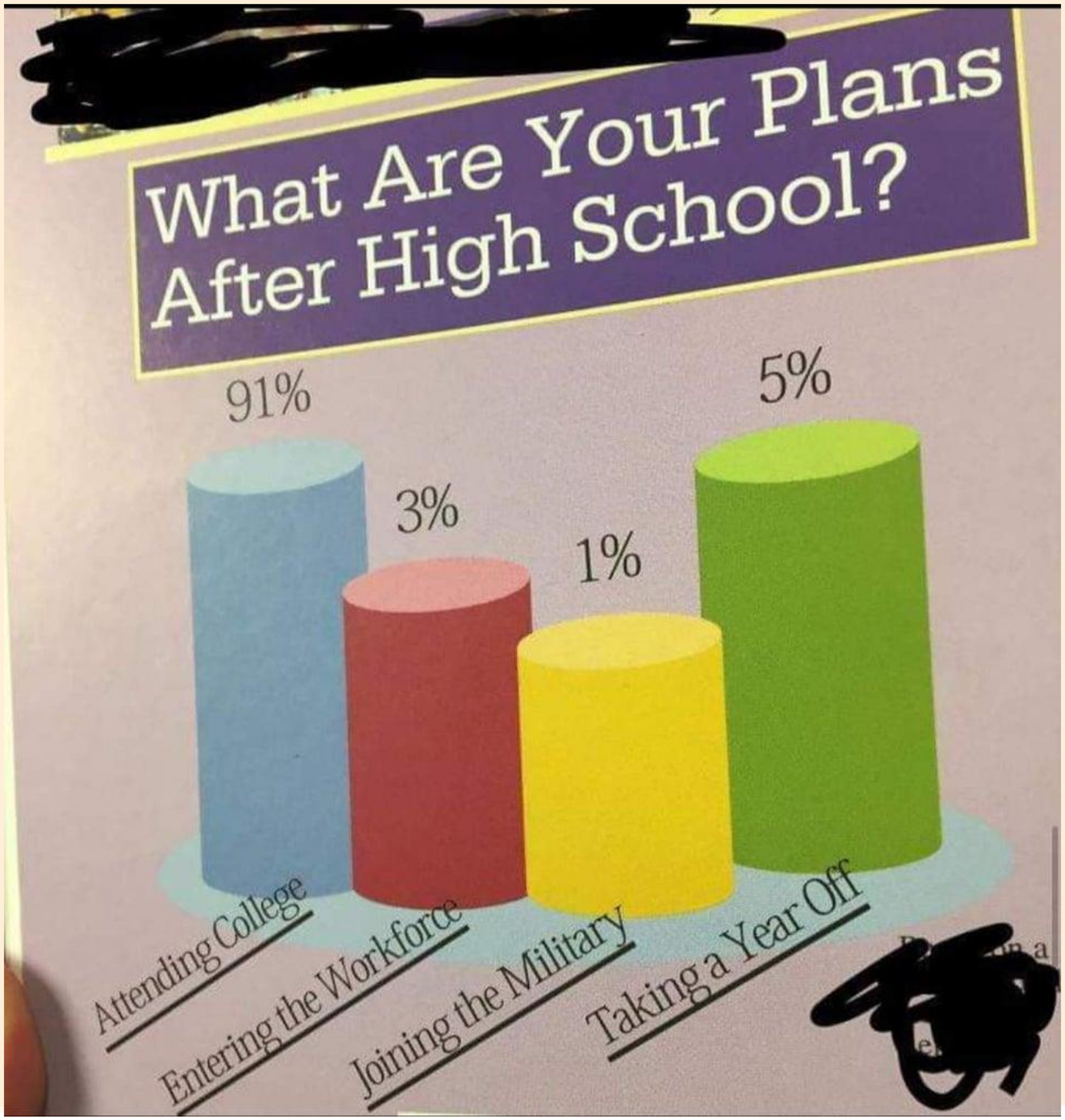


O FAMOSO "GRÁFICO DE CARRINHO DE MERCADO"

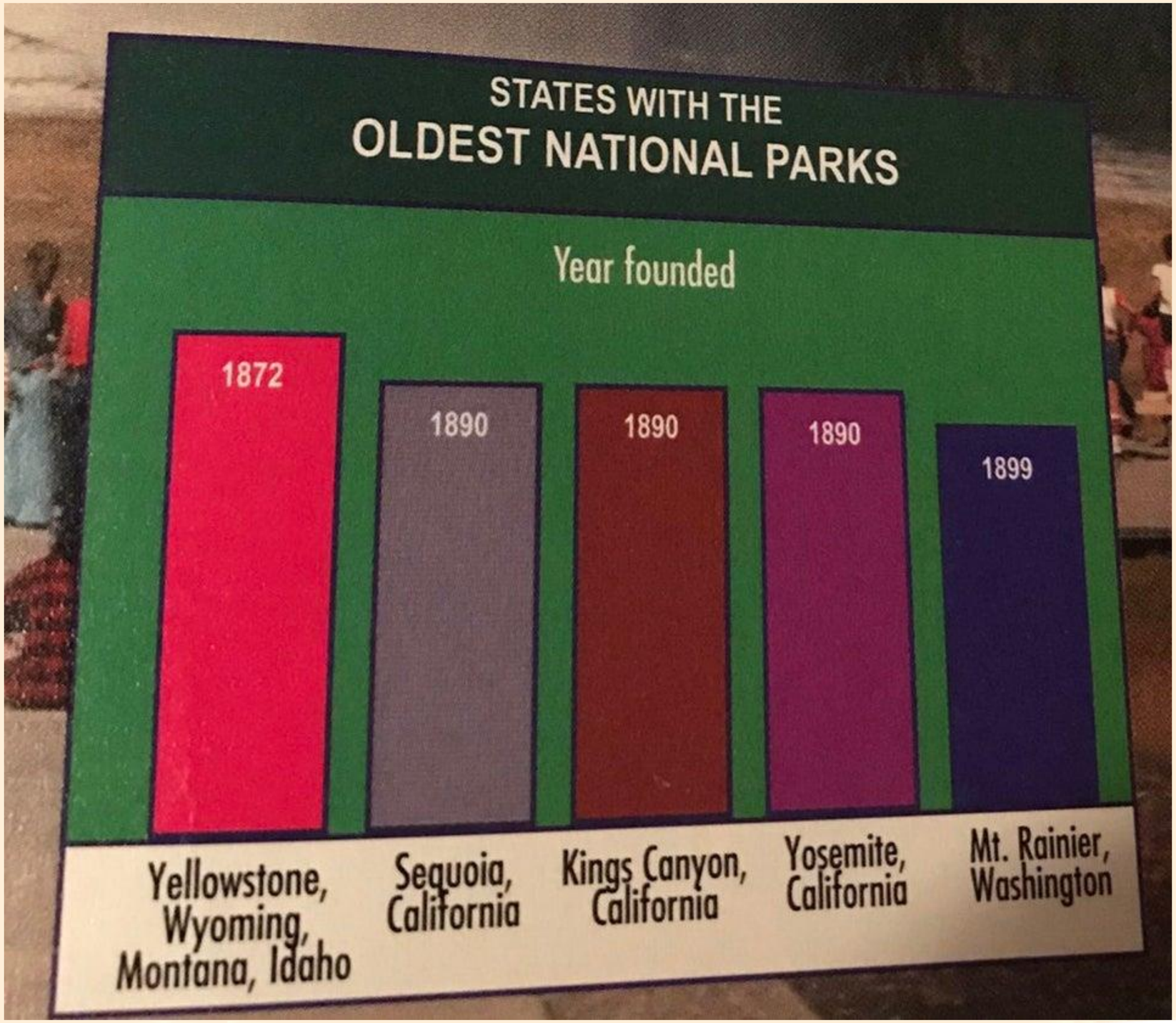
ELEIÇÕES

VOTO 2013
El pueblo decide
ELECCIÓN PRESIDENCIAL 2013
PORCENTAJES
50,66
%
49,07
Nicolás Maduro Moros
Henrique Capriles Rad

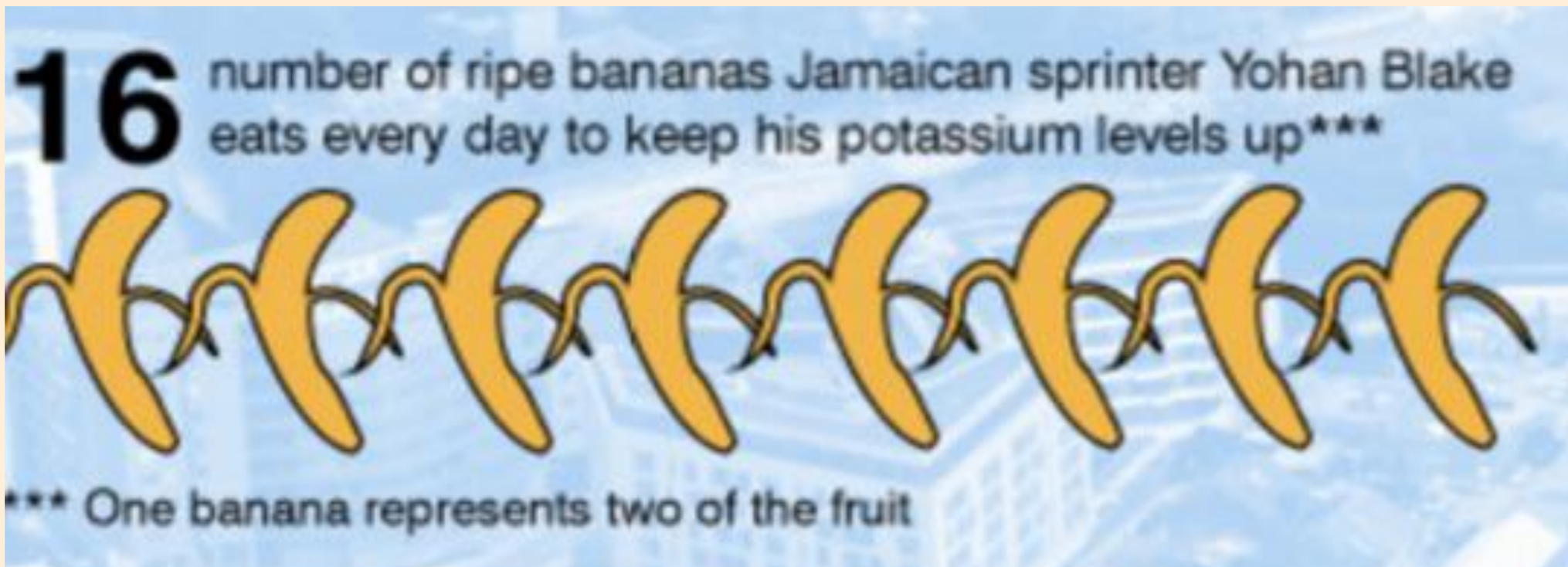
NÃO... APENAS NÃO... 🤦



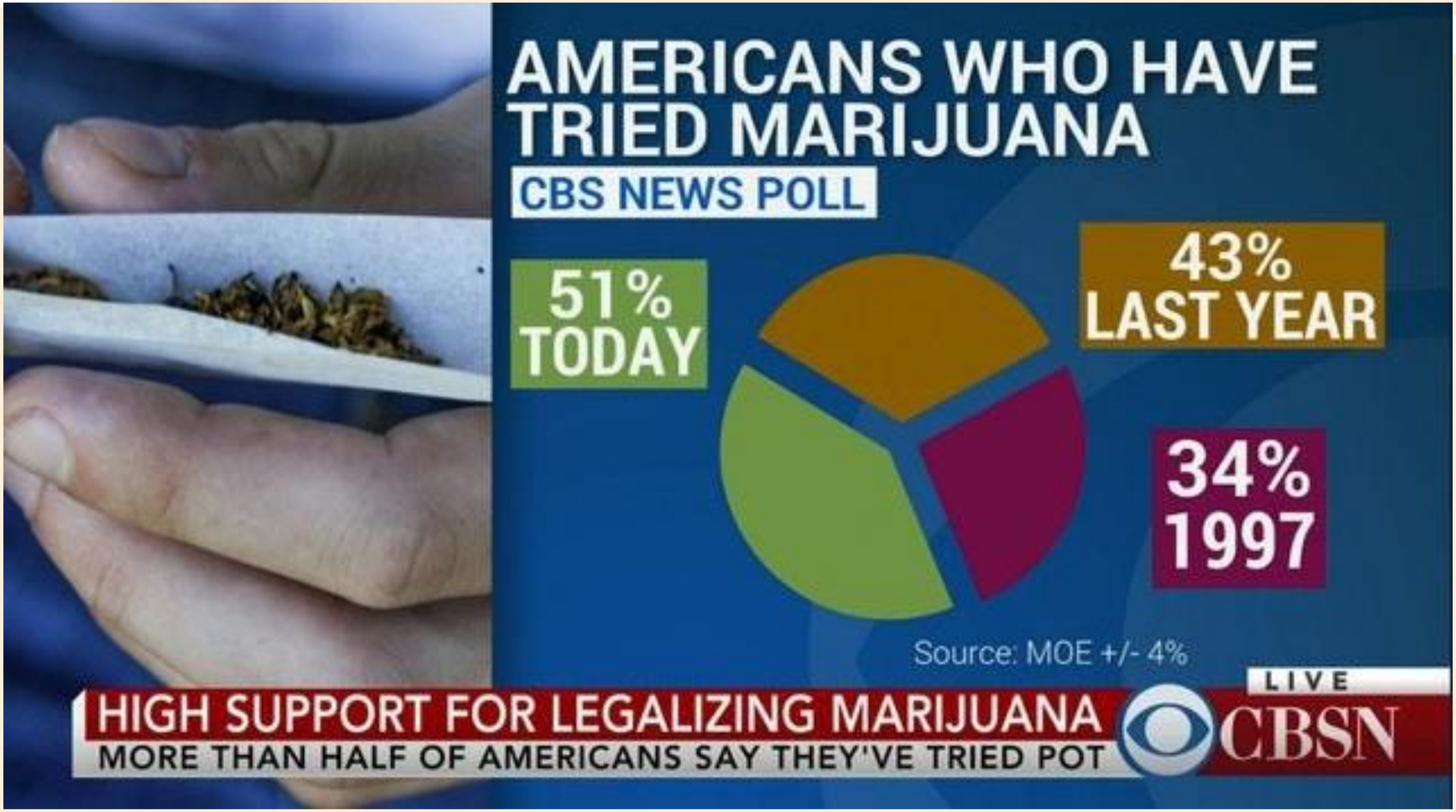
NÃO FAÇA ISSO...



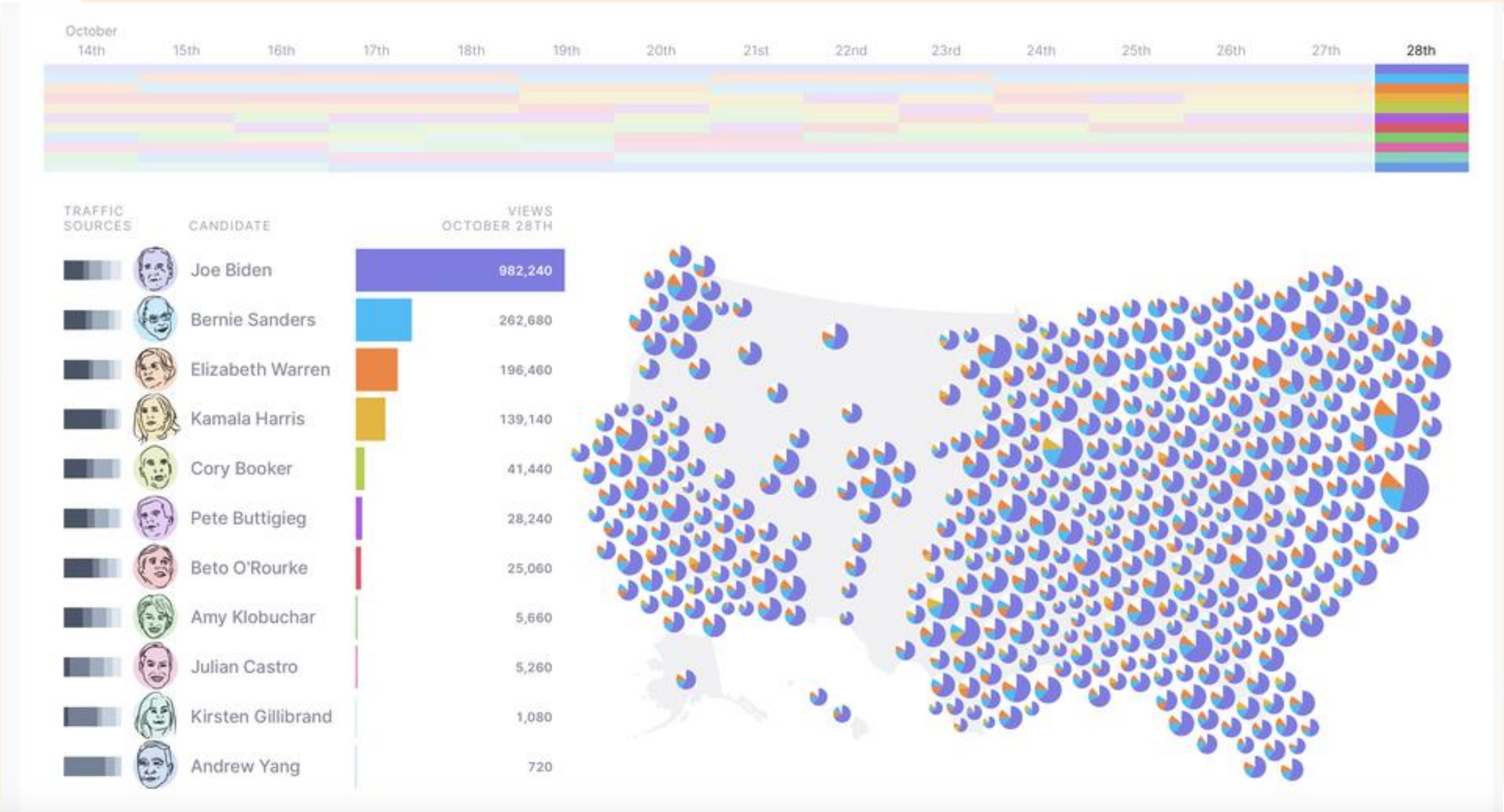
NÃO FAÇA ISSO...



UMA ASSOMBRAÇÃO

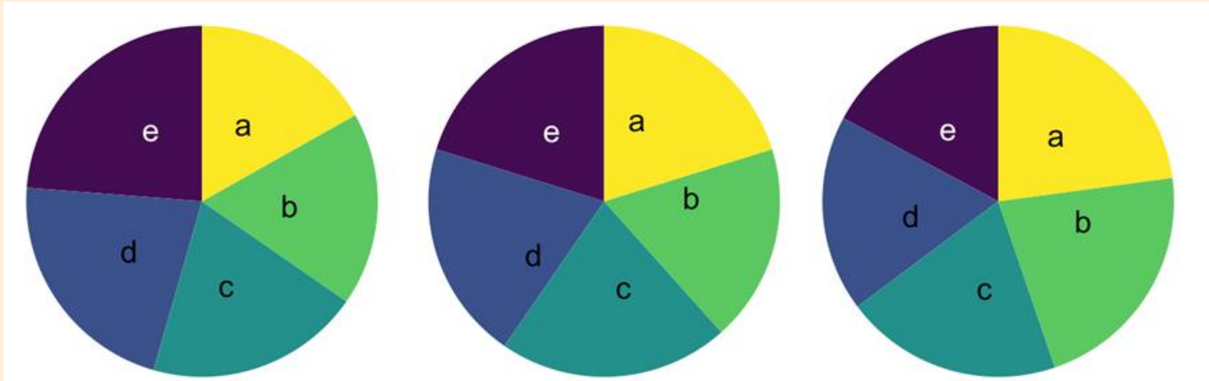


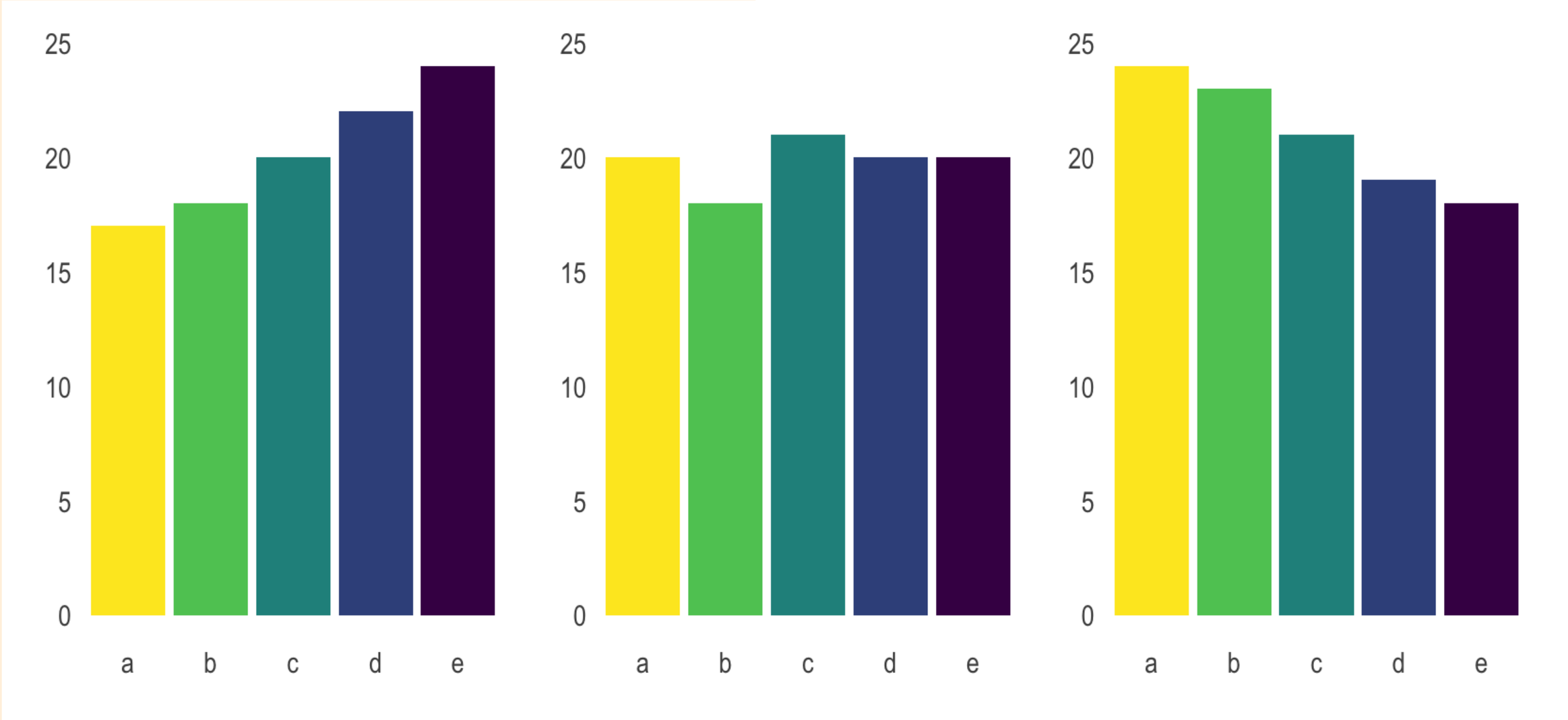
PAC-MAN PARA PRESIDENTE!



PORQUE EU RECLAMO TANTO DE GRÁFICOS DE PIZZA?

PORQUE EU RECLAMO TANTO DE GRÁFICOS DE PIZZA?





Fonte

GRÁFICOS PARA VARIÁVEIS QUANTITATIVAS

O *notebook* de nome "**AED R-03.ipynb**" traz esse conteúdo.

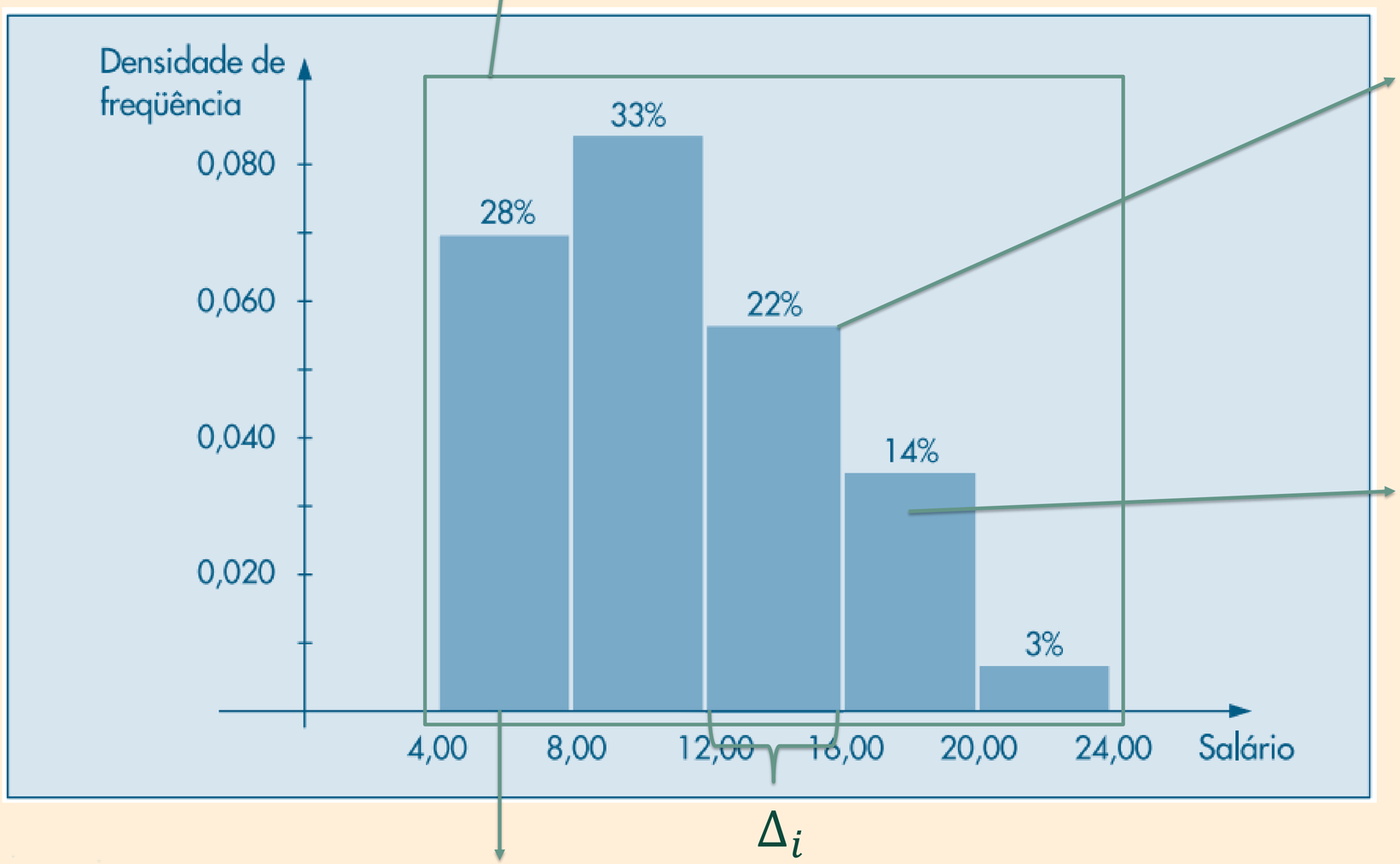
MAIS SOBRE O HISTOGRAMA

Classe de salários	Freqüência n_i	Porcentagem $100\ f_i$
4,00 ⊢ 8,00	10	27,78
8,00 ⊢ 12,00	12	33,33
12,00 ⊢ 16,00	8	22,22
16,00 ⊢ 20,00	5	13,89
20,00 ⊢ 24,00	1	2,78
Total	36	100,00

MAIS SOBRE O HISTOGRAMA

Área do histograma: $\sum_i \Delta_i \frac{f_i}{\Delta_i} = 1$

Classe de salários	Freqüência n_i	Porcentagem $100\ f_i$
4,00 ⊢ 8,00	10	27,78
8,00 ⊢ 12,00	12	33,33
12,00 ⊢ 16,00	8	22,22
16,00 ⊢ 20,00	5	13,89
20,00 ⊢ 24,00	1	2,78
Total	36	100,00



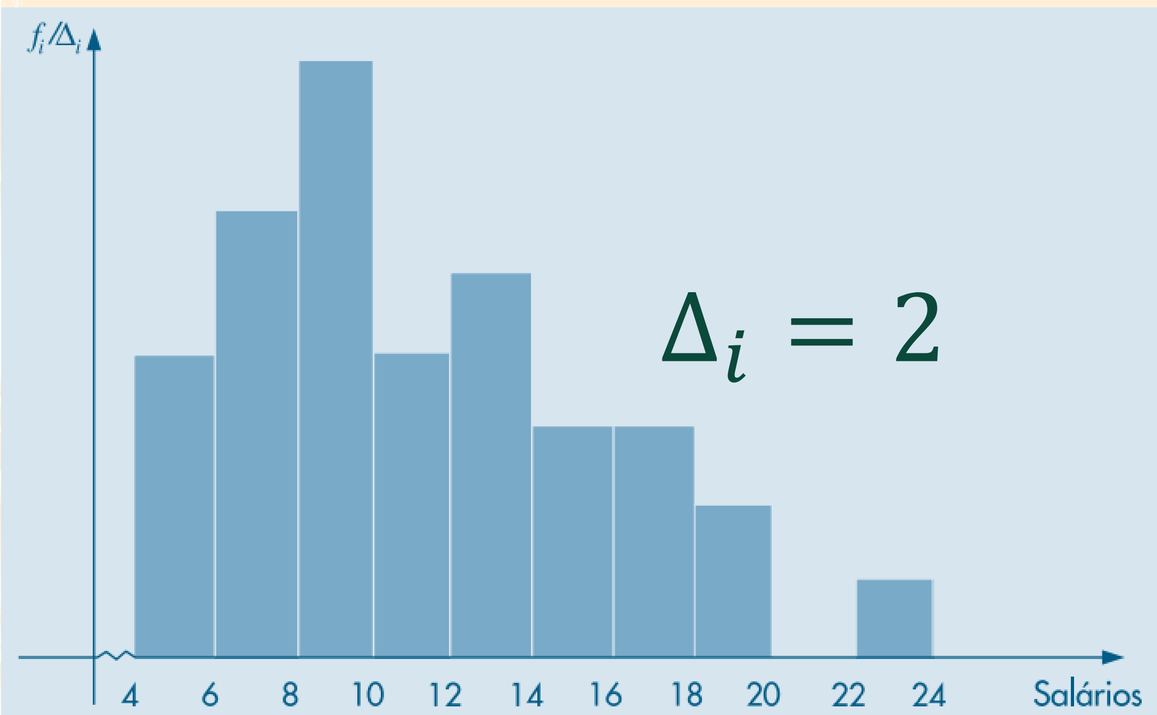
Alturas equivalentes a $\frac{f_i}{\Delta_i}$ (dens. de frequência)

Áreas equivalentes às frequências relativas f_i

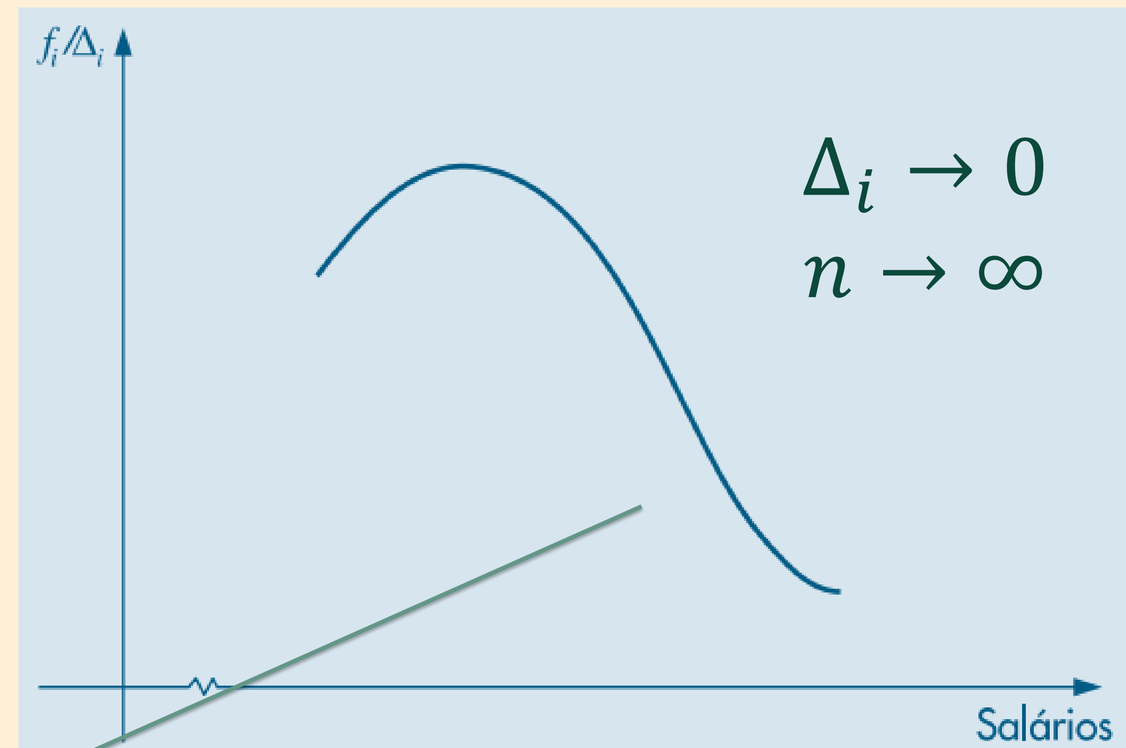
Bases proporcionais aos intervalos da classe

MAIS SOBRE O HISTOGRAMA (GEOGEBRA)

- Aqui tivemos $\Delta_i = \Delta$ valor constante
- Pode ser construído com bases não-constantes



$\Rightarrow$



- Histograma suavizado
- Aproxima uma *função densidade de probabilidade*

FREQUÊNCIAS ACUMULADAS

Definição: Dadas n observações de uma variável quantitativa (discretra ou contínua) e um número real x qualquer, indica-se por $N(x)$ o número de observações menores que ou iguais a x. A **função de frequências acumuladas empírica** é a função $F_n(x)$, dada por

$$F_n(x) = \frac{N(x)}{n}.$$

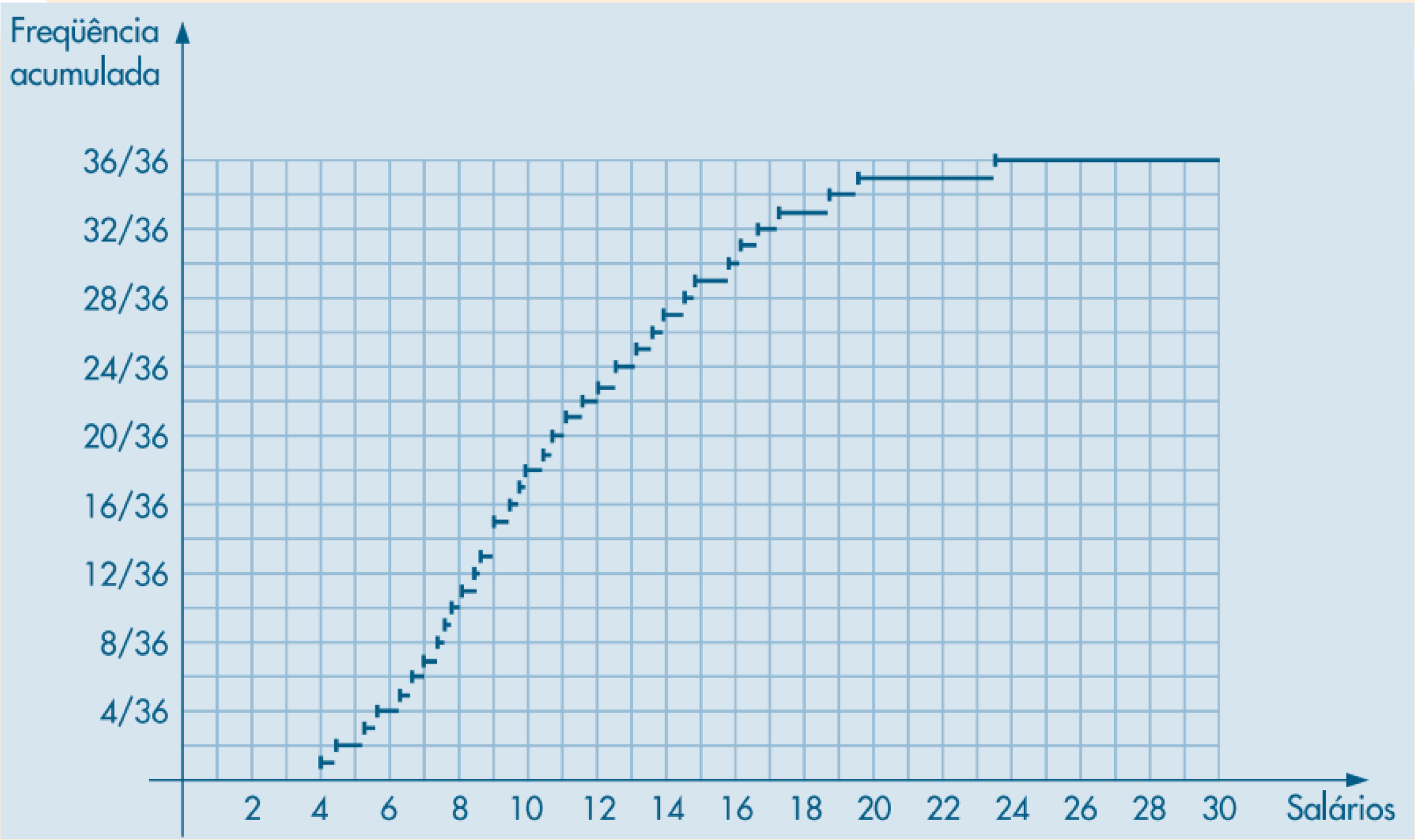
FREQUÊNCIAS ACUMULADAS

Classe de salários	Freqüência n_i	Porcentagem $100\ f_i$
4,00 ⊢ 8,00	10	27,78
8,00 ⊢ 12,00	12	33,33
12,00 ⊢ 16,00	8	22,22
16,00 ⊢ 20,00	5	13,89
20,00 ⊢ 24,00	1	2,78
Total	36	100,00

Salário (× sal. mín.)			
4,00	8,12	11,06	16,22
4,56	8,46	11,59	16,61
5,25	8,74	12,00	17,26
5,73	8,95	12,79	18,75
6,26	9,13	13,23	19,40
6,66	9,35	13,60	23,30
6,86	9,77	13,85	
7,39	9,80	14,69	
7,59	10,53	14,71	
7,44	10,76	15,99	

$$F_{36}(s) = \begin{cases} 0, & \text{se } s < 4{,}00 \\ 1/36\ , & \text{se } 4{,}00 \leqslant s < 4{,}56 \\ 2/36\ , & \text{se } 4{,}56 \leqslant s < 5{,}25 \\ \vdots & \vdots \\ 1, & \text{se } s \geqslant 23{,}30 \end{cases}$$

FREQUÊNCIAS ACUMULADAS



FREQUÊNCIAS ACUMULADAS

Nº de filhos
—
1
2
—
—
0
—
1
—
2
—
—
3
0
—
1
2

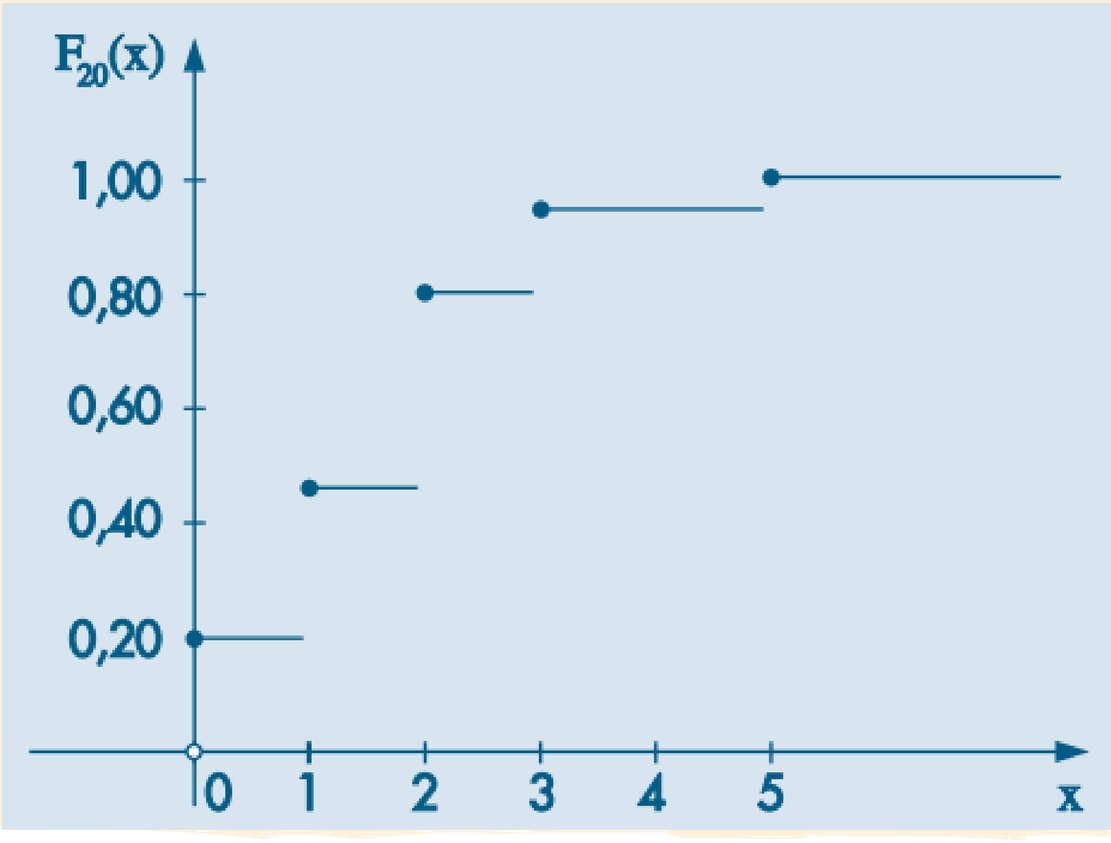
—
—
1
—
—
0
2
2
—
0
5
2
—
1
3
—
2
3

⇒

Nº de filhos z_i	Freqüência n_i	Porcentagem $100\ f_i$
0	4	20
1	5	25
2	7	35
3	3	15
5	1	5
Total	20	100

⇒

$$F_{20}(x) = \begin{cases} 0{,}00, & \text{se } x < 0 \\ 0{,}20, & \text{se } 0 \leqslant x < 1 \\ 0{,}45, & \text{se } 1 \leqslant x < 2 \\ 0{,}80, & \text{se } 2 \leqslant x < 3 \\ 0{,}95, & \text{se } 3 \leqslant x < 5 \\ 1{,}00, & \text{se } x \geqslant 5 \end{cases}$$



SUMÁRIOS NUMÉRICOS

Conteúdo referente às seções 3.1 e 3.2 de [MB] apresentadas no quadro e no *notebook* de nome "**AED R-04.ipynb**".

SUMÁRIOS NUMÉRICOS

Conteúdo referente às seções 3.3 e 3.4 de [MB] apresentadas no quadro e no *notebook* de nome "**AED R-05.ipynb**".

PROBLEMA 2, CAP. 2

Arquivo `empresa.csv`

7. Você foi convidado para chefiar a seção de orçamentos ou a seção técnica da Companhia MB. Após analisar o tipo de serviço que cada seção executa, você ficou indeciso e resolveu tomar a decisão baseado em dados fornecidos para as duas seções. O departamento pessoal forneceu os dados da Tabela 2.1 para os funcionários da seção de orçamentos, ao passo que, para a seção técnica, os dados vieram agrupados segundo as tabelas abaixo, que apresentam as frequências dos 50 empregados dessa seção, segundo as variáveis grau de instrução e salário. Baseado nesses dados, qual seria a sua decisão? Justifique.

Instrução	Frequência
Fundamental	15
Médio	30
Superior	5
Total	50

Classe de Salários	Frequência
7,50 ⊢ 10,50	14
10,50 ⊢ 13,50	17
13,50 ⊢ 16,50	11
16,50 ⊢ 19,50	8
Total	50

PROBLEMA I, CAP. 3

1. Quer se estudar o número de erros de impressão de um livro. Para isso escolheu-se uma amostra de 50 páginas, encontrando-se o número de erros por página da tabela abaixo.
   (a) Qual o número médio de erros por página?
   (b) E o número mediano?
   (c) Qual é o desvio padrão?
   (d) Faça uma representação gráfica para a distribuição.
   (e) Se o livro tem 500 páginas, qual é o número total de erros esperado no livro?

Erros	Frequência
0	25
1	20
2	3
3	1
4	1

PROBLEMA 4, CAP. 3

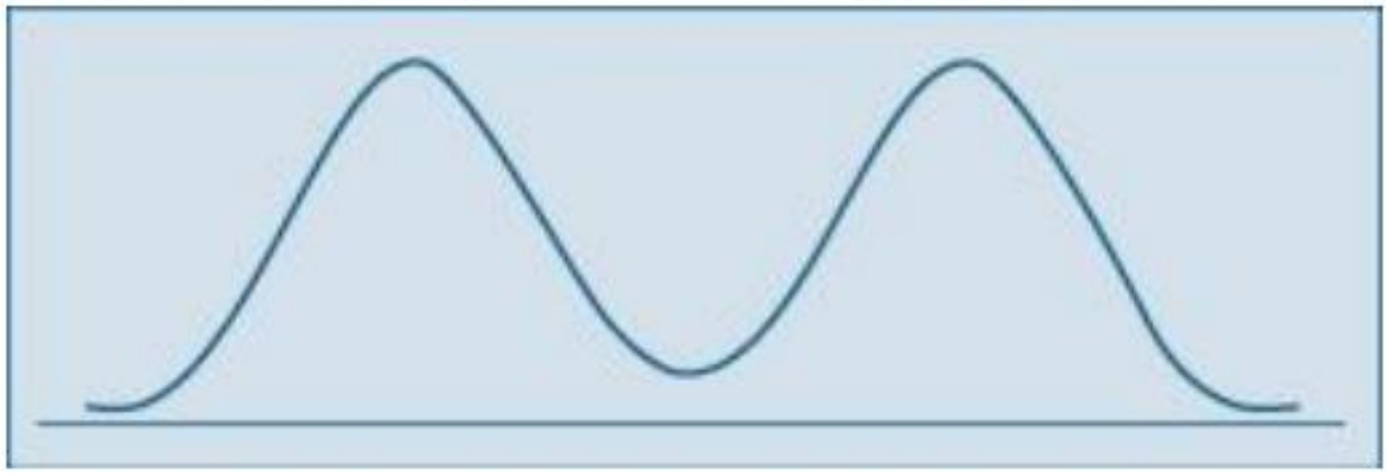
4. (a) Dê uma situação prática em que você acha que a mediana é uma medida mais apropriada do que a média.

(b) Esboce um histograma em que a média e a mediana coincidem. Existe alguma classe de histogramas em que isso sempre acontece?

(c) Esboce os histogramas de três variáveis (X, Y e Z) com a mesma média aritmética, mas com as variâncias ordenadas em ordem crescente.

PROBLEMA 5, CAP. 3

5. Suponha que a variável de interesse tenha a distribuição como na figura abaixo.



Você acha que a média é uma boa medida de posição? E a mediana? Justifique.

PROBLEMA 6, CAP. 3

6. Numa pesquisa realizada com 100 famílias, levantaram-se as seguintes informações:

Número de filhos	0	1	2	3	4	5	mais que 5
Frequência de famílias	17	20	28	19	7	4	5

(a) Qual a mediana do número de filhos?
(b) E a moda?
(c) Que problemas você enfrentaria para calcular a média? Faça alguma suposição e encontre-a.

PROBLEMA 14, CAP. 3

14. Mostre que:

(a) $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$

(b) $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum x_i)^2}{n}$

PROBLEMA 20, CAP. 3

20. O que acontece com a mediana, a média e o desvio padrão de uma série de dados quando:

(a) cada observação é multiplicada por 2?

(b) soma-se 10 a cada observação?

(c) subtrai-se a média geral $\bar{x}$ de cada observação?

(d) de cada observação subtrai-se $\bar{x}$ e divide-se pelo desvio padrão dp(x)?

PROBLEMA 21, CAP. 3

21. Na companhia A, a média dos salários é 10.000 unidades e o 3º quartil é 5.000.

(a) Se você se apresentasse como candidato a funcionário nessa firma e se o seu salário fosse escolhido ao acaso entre todos os possíveis salários, o que seria mais provável: ganhar mais ou menos que 5.000 unidades?

(b) Suponha que, na companhia B, a média dos salários seja 7.000 unidades, a variância praticamente zero e o salário também seja escolhido ao acaso. Em qual companhia você se apresentaria para procurar emprego?

PROBLEMA 22, CAP. 3

22. Estamos interessados em estudar a idade dos 12.325 funcionários da Cia. Distribuidora de Leite Teco, e isso será feito por meio de uma amostra. Para determinar que tamanho deverá ter essa amostra, foi colhida uma amostra-piloto. As idades observadas foram: 42, 35, 27, 21, 55, 18, 27, 30, 21, 24.

(a) Determine as medidas descritivas dos dados que você conhece.

(b) Qual dessas medidas você acredita que será a mais importante para julgar o tamanho final da amostra? Por quê?

PROBLEMA 28 – 3I, CAP. 3

28. Para se estudar o desempenho de duas corretoras de ações, selecionou-se de cada uma delas amostras aleatórias das ações negociadas. Para cada ação selecionada, computou-se a porcentagem de lucro apresentada durante um período fixado de tempo. Os dados estão a seguir.

Corretora A		
45	60	54
62	55	70
38	48	64
55	56	55
54	59	48
65	55	60

Corretora B		
57	55	58
50	52	59
59	55	56
61	52	53
57	57	50
55	58	54
59	51	56

PROBLEMA 28 – 3I, CAP. 3

Que tipo de informação revelam esses dados? (Sugestão: use a análise proposta nas Seções 3.3 e 3.4.)

29. Para verificar a homogeneidade das duas populações do problema anterior, um estatístico sugeriu que se usasse o quociente $F = \frac{\text{var}(X/A)}{\text{var}(X/B)}$, mas não disse qual decisão tomar baseado nesse valor. Que regra de decisão você adotaria para dizer se são homogêneas ou não (var(X/A) = variância de X, para a corretora A; X = % de lucro)?

30. Faça um *box plot* para os dados da corretora A e um para os dados da corretora B. Compare os dois conjuntos de dados por meio desses desenhos.

31. Para decidir se o desempenho das duas corretoras do exercício 29 são semelhantes ou não, adotou-se o seguinte teste: sejam

$$t = \frac{\bar{x}_A - \bar{x}_B}{S_* \sqrt{1/n_A + 1/n_B}}, S_*^2 = \frac{(n_A - 1)\,\text{var}(X/A) + (n_B - 1)\,\text{var}(X/B)}{n_A + n_B - 2}$$

Caso $|t| < 2$, os desempenhos são semelhantes, caso contrário, são diferentes. Qual seria a sua conclusão? Aqui, n_A é o número de ações selecionadas da corretora A e nomenclatura análoga para n_B.

PROBLEMA 36, CAP. 3

36. Usando os dados da variável qualitativa região de procedência, da Tabela 2.1, transforme-a na variável quantitativa X, definida do seguinte modo:

$$X = \begin{cases} 1, \text{ se a regi„o de procedÍncia for capital;} \\ 0, \text{ se a regi„o de procedÍncia for interior ou outra.} \end{cases}$$

(a) Calcule $\bar{x}$ e var(X).
(b) Qual a interpretação de $\bar{x}$?
(c) Construa um histograma para X.

ALGUMAS JUSTIFICATIVAS TEÓRICAS

Conteúdo apresentado no quadro e no *notebook* de nome "**AED R-06.ipynb**".

ANÁLISE BIDIMENSIONAL

Indivíduo	Variável					
	X_1	X_2	...	X_j	...	X_p
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
⋮	⋮	⋮		⋮		⋮
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
⋮	⋮	⋮		⋮		⋮
n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

- Explorar relações entre as variáveis (colunas) ou entre as linhas (indivíduos)
- n observações dos atributos $X_1, \ldots, X_p$
- Comparações possíveis:
  - Quali x quali
  - Quanti x quanti
  - Quali x quanti

DISTRIBUIÇÃO CONJUNTA DE FREQUÊNCIAS

Nº	Estado civil	Y Grau de instrução	Nº de filhos	Salário (× sal. mín.)	Idade		Região de procedência V
					anos	meses	
1	solteiro	ensino fundamental	—	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	capital
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	—	5,73	20	10	outra
5	solteiro	ensino fundamental	—	6,26	40	07	outra
6	casado	ensino fundamental	0	6,66	28	00	interior
7	solteiro	ensino fundamental	—	6,86	41	00	interior
8	solteiro	ensino fundamental	—	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	capital
10	solteiro	ensino médio	—	7,44	23	06	outra

DISTRIBUIÇÃO CONJUNTA DE FREQUÊNCIAS

V \ Y	Ensino Fundamental	Ensino Médio	Superior	Total
Capital	4	5	2	11
Interior	3	7	2	12
Outra	5	6	2	13
Total	12	18	6	36

Frequência observada das realizações simultâneas de Y e V

Distribuição marginal de Y

Distribuição marginal de Z

DISTRIBUIÇÃO CONJUNTA DE FREQUÊNCIAS

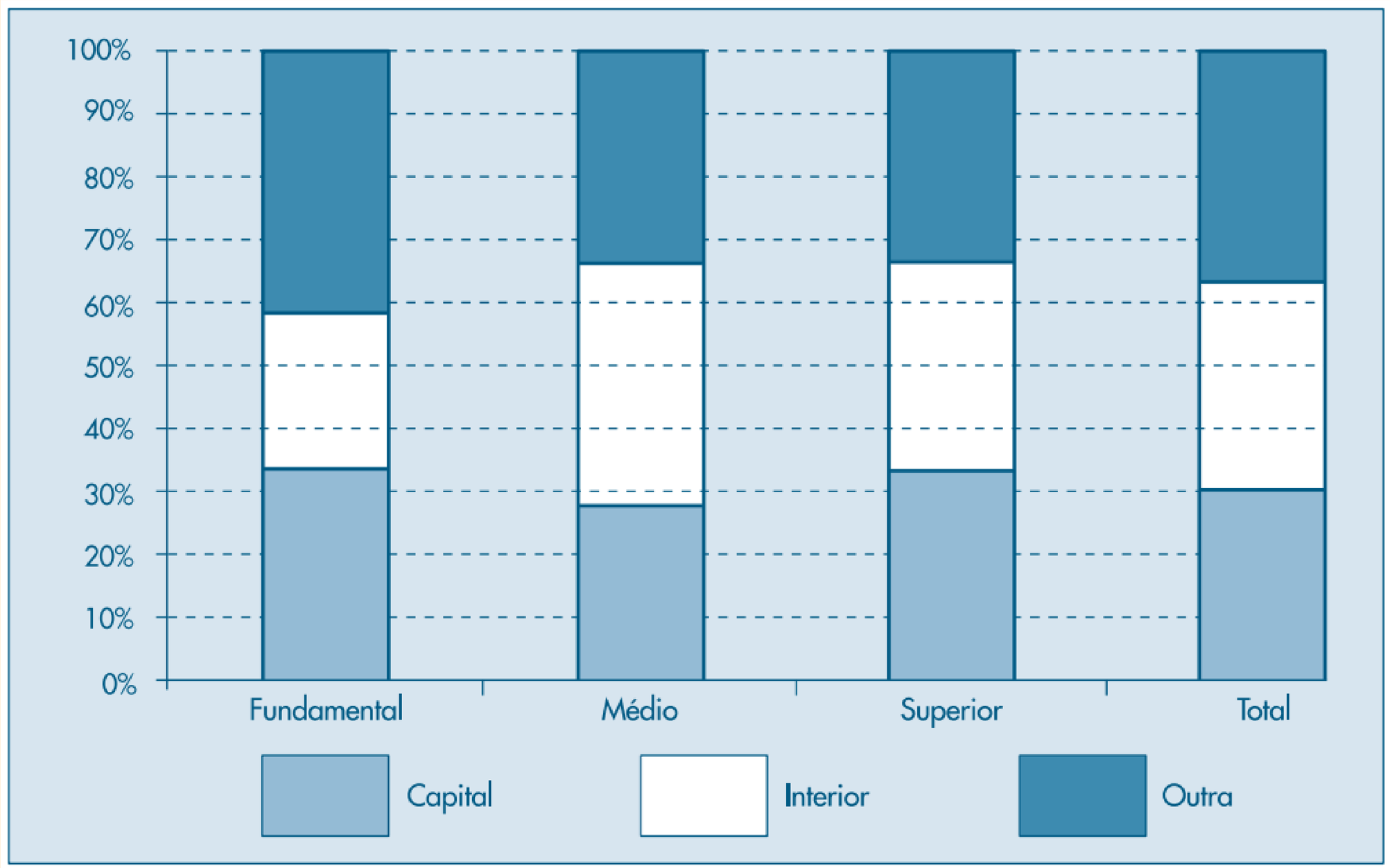
V \ Y	Fundamental	Médio	Superior	Total
Capital	11%	14%	6%	31%
Interior	8%	19%	6%	33%
Outra	14%	17%	5%	36%
Total	33%	50%	17%	100%

- Distribuição conjunta de frequências relativas (em relação ao total)
- Cada bloco soma 100%

V \ Y	Fundamental	Médio	Superior	Total
Capital	33%	28%	33%	31%
Interior	25%	39%	33%	33%
Outra	42%	33%	34%	36%
Total	100%	100%	100%	100%

- Distribuição condicional de frequências relativas (em relação ao total das colunas)
- Cada bloco soma 100%

GRÁFICO DE BARRAS ADAPTADO A ESSE CENÁRIO



UM EXERCÍCIO

V \ Y	Fundamental	Médio	Superior	Total
Capital	11%	14%	6%	31%
Interior	8%	19%	6%	33%
Outra	14%	17%	5%	36%
Total	33%	50%	17%	100%

Sortear um dos 36 funcionários ao acaso. Responda às perguntas abaixo.

- a) Qual será mais provavelmente o seu grau máximo de instrução?
- b) E sua região de procedência?
- c) Qual a probabilidade do sorteado ter nível superior?
- d) Sabendo que o sorteado é do interior, qual a probabilidade de ter nível superior?
- e) Sabendo que o escolhido é da capital, qual a probabilidade de ter nível superior?

ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Y \ X	Masculino	Feminino	Total
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

Estudar relação de dependência entre variáveis qualitativas

Por exemplo, há associação entre sexo e carreira no exemplo acima?

Difícil de inferir diretamente da tabela de frequências!

ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Y \ X	Masculino	Feminino	Total
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

Distribuição de frequências relativas (em relação às colunas)

Y \ X	Masculino	Feminino	Total
Economia	61%	58%	60%
Administração	39%	42%	40%
Total	100%	100%	100%

Preferência entre sexo masculino

Preferência entre sexo masculino

Preferência independente do sexo

Parece não haver associação entre sexo e tais escolhas de carreira!

ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Y \ X	Masculino	Feminino	Total
Física	100 (71%)	20 (33%)	120 (60%)
Ciências Sociais	40 (29%)	40 (67%)	80 (40%)
Total	140 (100%)	60 (100%)	200 (100%)

Agora parece haver associação...

Como quantificar associação?

Como aferir estatisticamente se de fato há associação?

MEDIDAS DE ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

- Obter estatísticas que indiquem associação ou não
- Quantidades adimensionais entre 0 e 1 (ou -1 e 1)
- Valores próximos a zero indicam falta de associação
- Coeficiente de contingência (K. Pearson)
- Em geral, serão chamados “coeficientes de associação” ou “correlação”
- Aprendemos análogos no caso de variáveis quantitativas

MEDIDAS DE ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

A criação de determinado tipo de cooperativa está associada com algum fator regional?

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	214 (33%)	237 (37%)	78 (12%)	119 (18%)	648 (100%)
Paraná	51 (17%)	102 (34%)	126 (42%)	22 (7%)	301 (100%)
Rio G. do Sul	111 (18%)	304 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Fonte: Sinopse Estatística da Brasil — IBGE, 1977.

Valores observados

Talvez sim, pois caso contrário...

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157 (24%)	269 (42%)	143 (22%)	79 (12%)	648 (100%)
Paraná	73 (24%)	124 (42%)	67 (22%)	37 (12%)	301 (100%)
Rio G. do Sul	146 (24%)	250 (42%)	133 (22%)	73 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Valores esperados

MEDIDAS DE ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157 (24%)	269 (42%)	143 (22%)	79 (12%)	648 (100%)
Paraná	73 (24%)	124 (42%)	67 (22%)	37 (12%)	301 (100%)
Rio G. do Sul	146 (24%)	250 (42%)	133 (22%)	73 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

...de acordo com essas porcentagens

Por exemplo, $67 \approx 22\%$ de 301

Quantidades distribuídas nas colunas...

MEDIDAS DE ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Desvios entre observados e esperados

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	214 (33%)	237 (37%)	78 (12%)	119 (18%)	648 (100%)
Paraná	51 (17%)	102 (34%)	126 (42%)	22 (7%)	301 (100%)
Rio G. do Sul	111 (18%)	304 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Fonte: Sinopse Estatística da Brasil — IBGE, 1977.

—

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157 (24%)	269 (42%)	143 (22%)	79 (12%)	648 (100%)
Paraná	73 (24%)	124 (42%)	67 (22%)	37 (12%)	301 (100%)
Rio G. do Sul	146 (24%)	250 (42%)	133 (22%)	73 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

=

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	57 (20,69)	−32 (3,81)	−65 (29,55)	40 (20,25)
Paraná	−22 (6,63)	−22 (3,90)	59 (51,96)	−15 (6,08)
Rio G. do Sul	−35 (8,39)	54 (11,66)	6 (0,27)	−25 (8,56)

"Grandes" desvios

MEDIDAS DE ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Chi-quadrado de Pearson

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	214 (33%)	237 (37%)	78 (12%)	119 (18%)	648 (100%)
Paraná	51 (17%)	102 (34%)	126 (42%)	22 (7%)	301 (100%)
Rio G. do Sul	111 (18%)	304 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Fonte: Sinopse Estatística da Brasil — IBGE, 1977.

o_i valor observado

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157 (24%)	269 (42%)	143 (22%)	79 (12%)	648 (100%)
Paraná	73 (24%)	124 (42%)	67 (22%)	37 (12%)	301 (100%)
Rio G. do Sul	146 (24%)	250 (42%)	133 (22%)	73 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

e_i valor esperado

$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

$= 171{,}76,$

nesse exemplo

- Valor "grande" indica associação
- Mas quão "grande"?

MEDIDAS DE ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Veja o *notebook* de nome "**AED 2023-01 – R-07.ipynb**" anexado no material "**Aula de R 07: Medidas de associação entre variáveis qualitativas**", no tema "**Material**" na aba "**Atividades**".