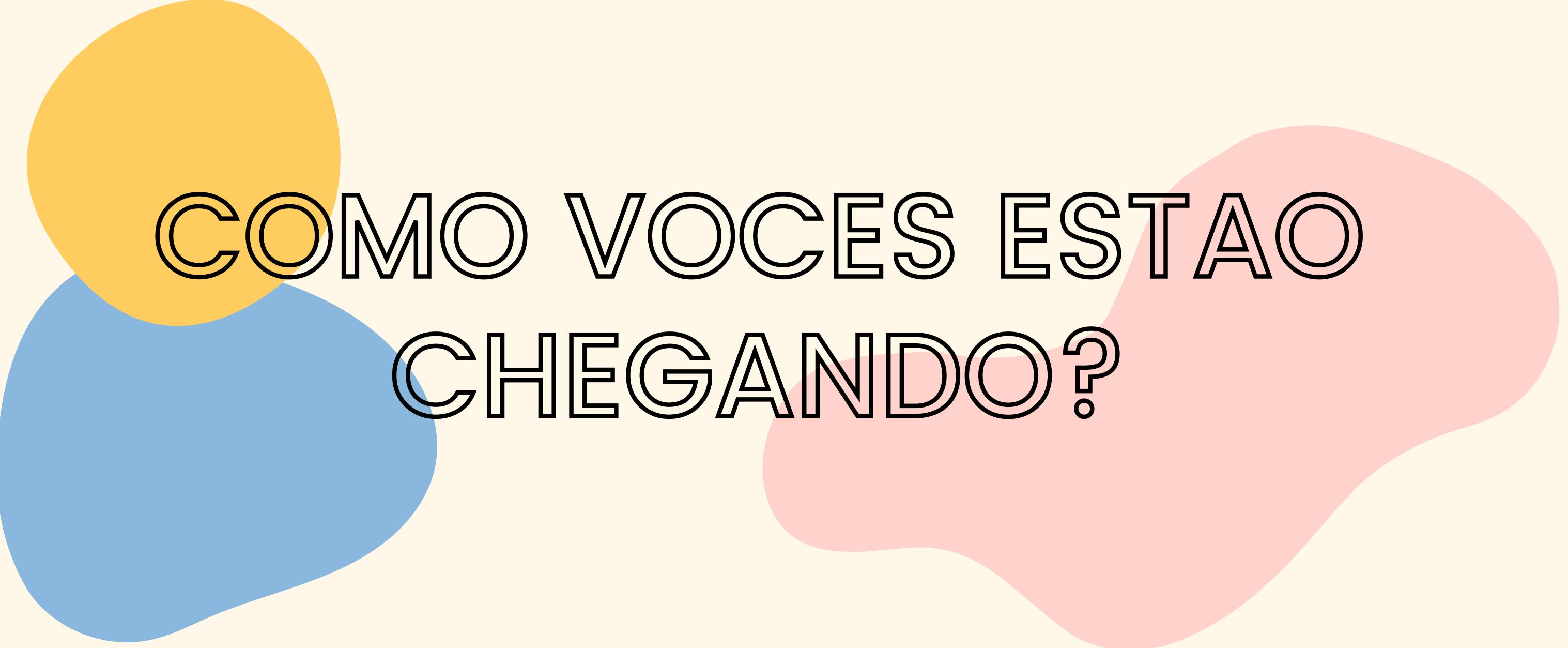


TRABALHANDO COM DADOS NA PRÁTICA

Oficina de R

OPEN DATA DAY 2022
Prefeitura de Mogi das Cruzes
05/03/2022
15-18h

Mônica Rocabado
Mestranda em Adm.Pública
e Pesquisadora



**COMO VOCES ESTAO
CHEGANDO?**

www.menti.com

7217 5852

LEMBRETES DA OFICINA

Enquanto aguarda a entrada dos outros, aqui estão algumas regras e lembretes para ter em mente.

1

Silencie seu microfone

2

Deixem o Rstudio
Cloud logado

3

As perguntas enviadas
através EasyRetro serão
respondidas depois de
cada segmento.

SOBRE MIM



Mestranda em Administracao Pública e Governo pela FGV EAESP e Pesquisadora no CeDHE FGV

Sou Bacharel em Administração Pública pela FGV-EAESP, estou pesquisadora, em constante formação em ciência de dados e design. Sou apaixonada em compreender, descobrir e sintetizar informações complexas em conteúdo visual e agradável ao leitor, para isso utilizando principalmente a linguagem R.

Já trabalhei no governo em um laboratório de inovação, em consultoria política e agora integro um núcleo de pesquisa voltado a empresas e direitos humanos.

Através da programação, do uso de dados e do trabalho em rede vejo o potencial para contribuir com essa mudança.

ESTRUTURA DA OFICINA

Nossa agenda desta manhã

- Importancia da análise de dados – fundamentos
- O que é o R e Rstudio Cloud?
- R 101
- Conhecendo os dados que iremos trabalhar
- R na prática
- Análisando os dados

O QUE NAO VAMOS VER HOJE

- Não é um curso aprofundado da linguagem R

- Não trataremos de conceitos estatísticos e econométricos

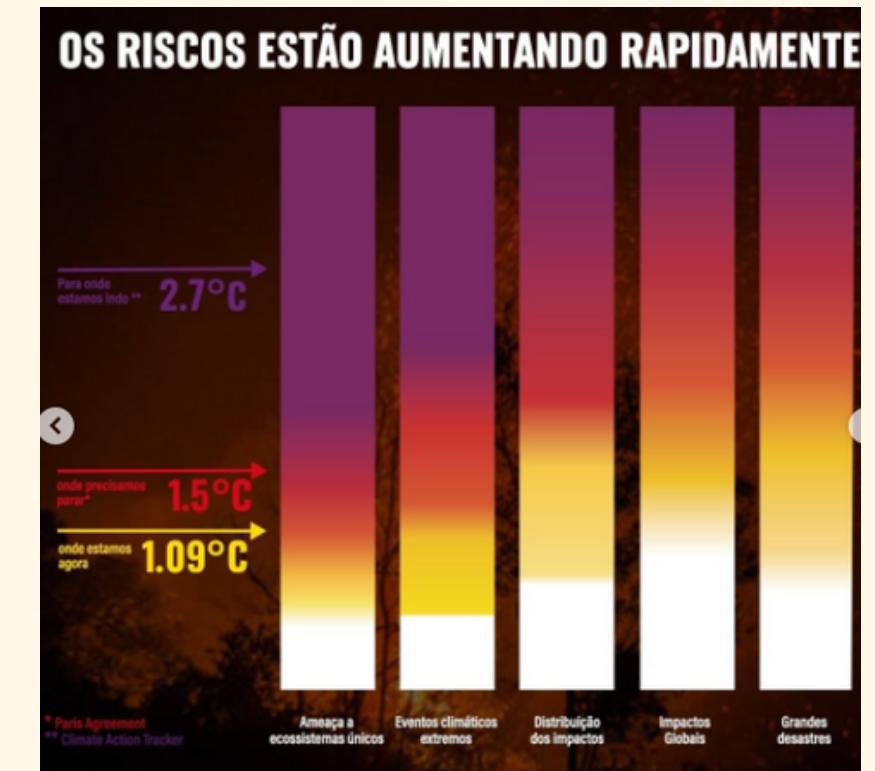
- Não é curso sobre Big Data

IMPORTÂNCIA DA ANÁLISE DE DADOS - FUNDAMENTOS



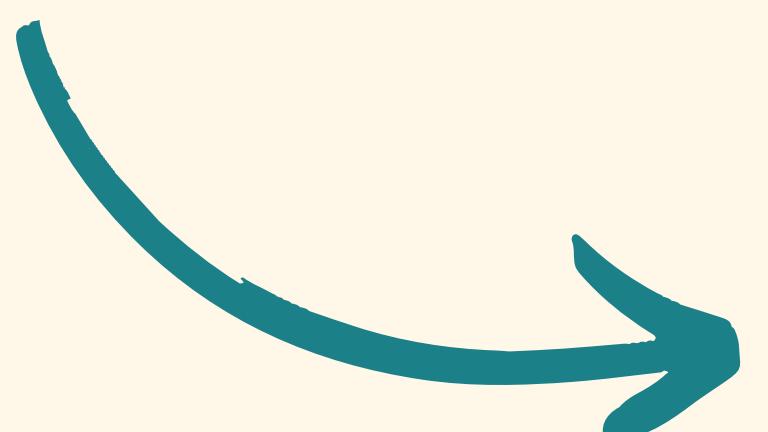
IMPORTANCIA

● Compreensão de informacoes gráficas e lógicas



● Processos e decisoes orientados por dados

● Accountability do poder público
Dados Abertos



IPCC fev/22: os riscos que enfrentaremos nos próximos 20 anos por conta do aquecimento global de $1,5^{\circ}\text{C}$ já são inevitáveis.

FUNDAMENTOS

A importância da base de dados e do contexto

Dados [...] são o resultado, por vezes, de longos processos de construção que envolvem várias decisões metodológicas (p.11) – Guia Brasileiro de Dados

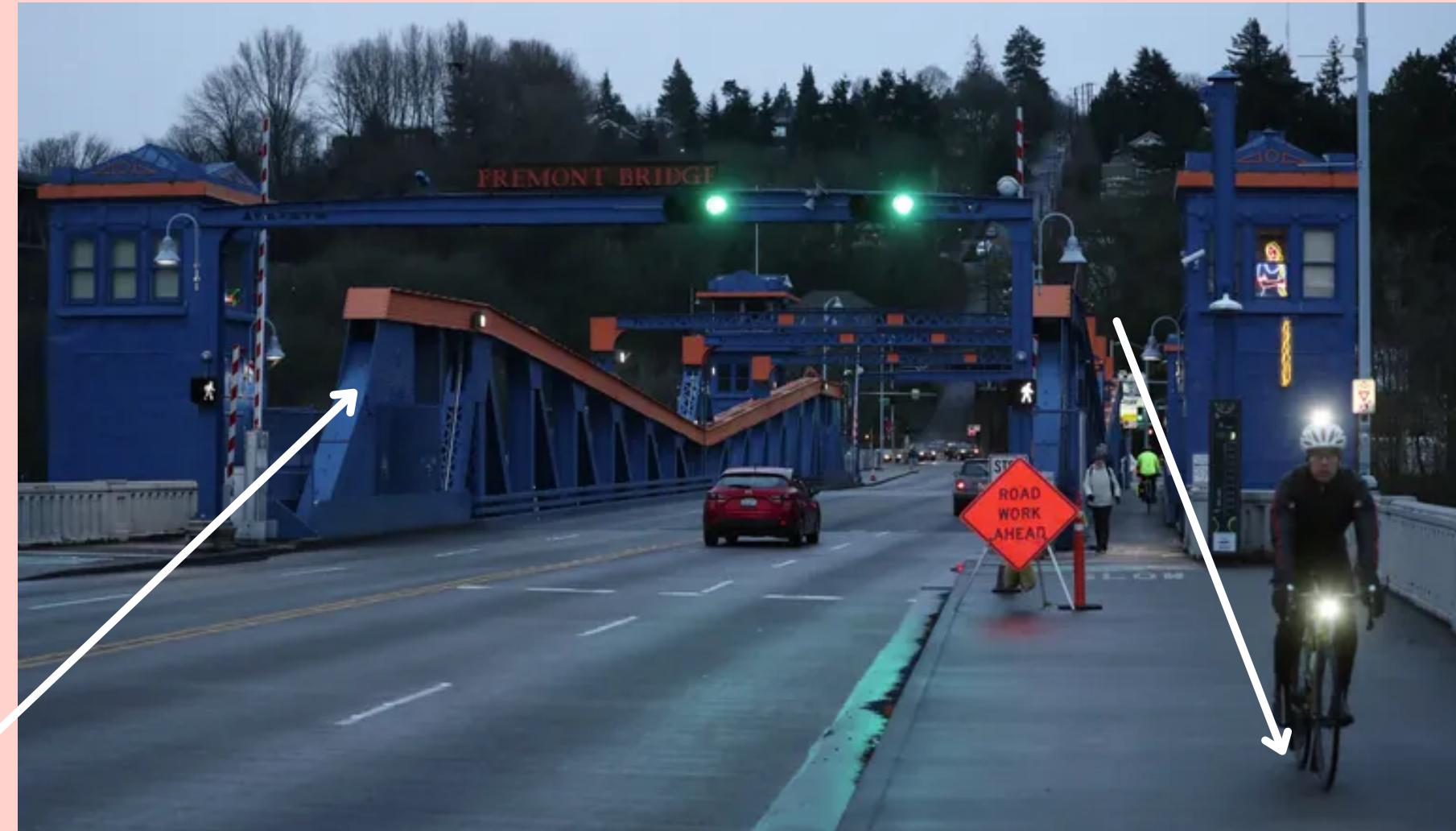
Sempre irá existir uma lacuna entre nossos dados e o mundo real. Os dados não são um reflexo perfeito da realidade –
Avoiding data pitfalls

ATENÇÃO: A seleção da base de dados é um ato de pesquisa.

REFLEXÃO

Um contador de bicicletas localizado em uma ponte de duas vias, em um determinado dia contou mais bicicletas de um lado do que de outro

2 mil bicicletas
contadas do lado
esquerdo da
ponte



4 mil bicicletas
contadas do lado
esquerdo da ponte

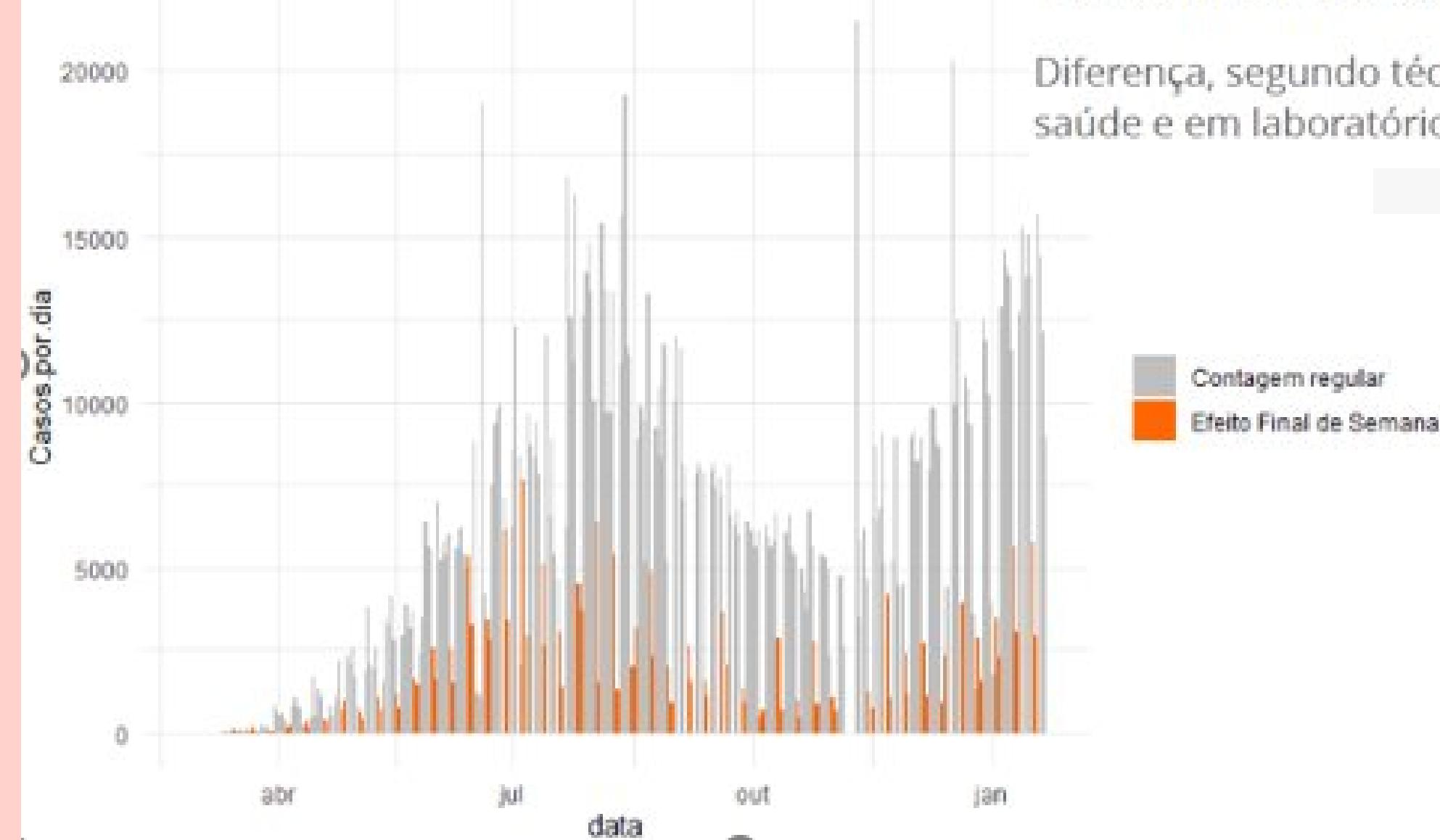
Por que o número de bicicletas que
passou de um dos lados da ponte é
metade do que voltou?

www.menti.com

7217 5852

SAÚDE

Efeito do final de semana na contagem diária de casos
Segunda e Domingo - Contagens no sábado e domingo



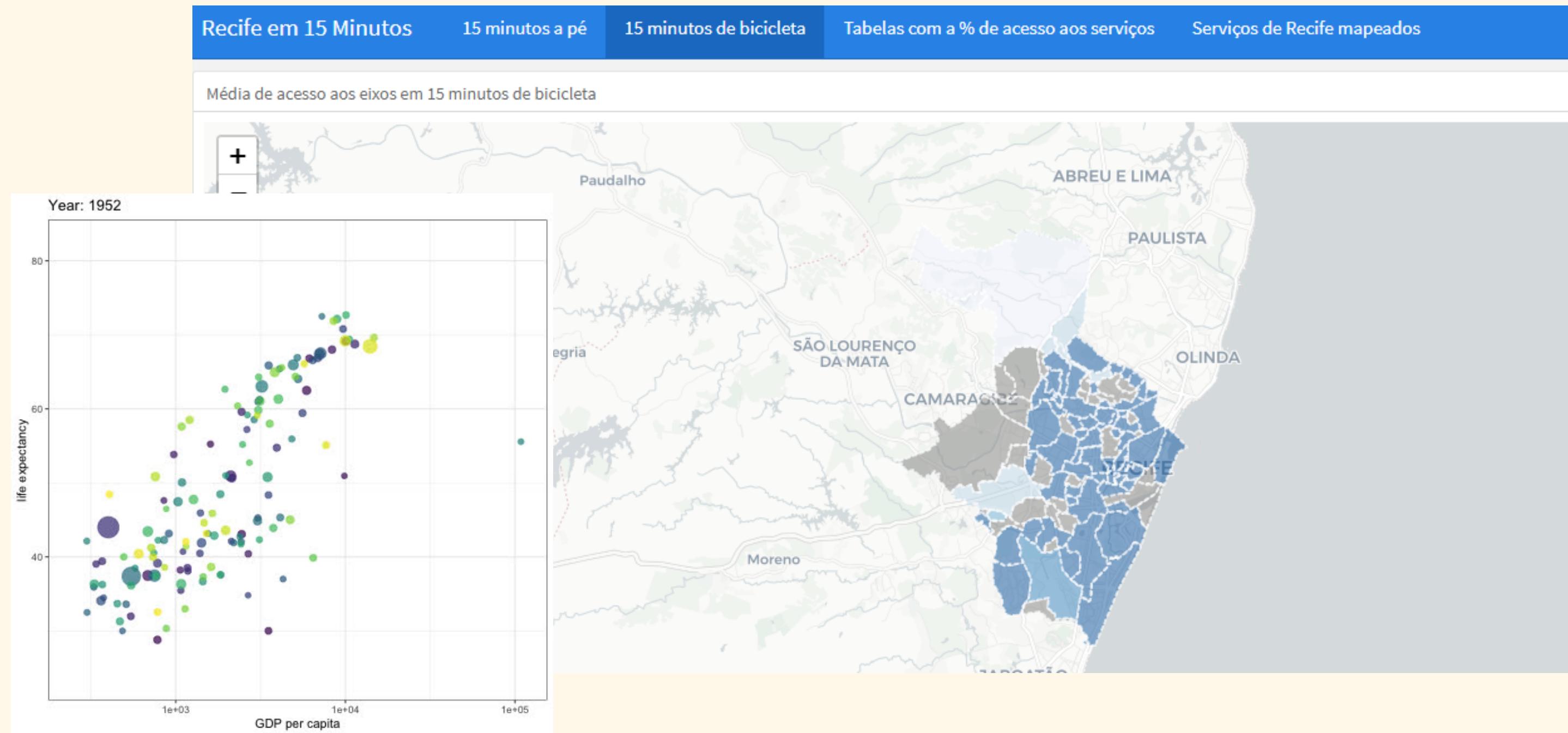
AS LACUNAS SEMPRE VÃO
EXISTIR. O IMPORTANTE É
COMPREENDER QUAL O
SEU TAMANHO



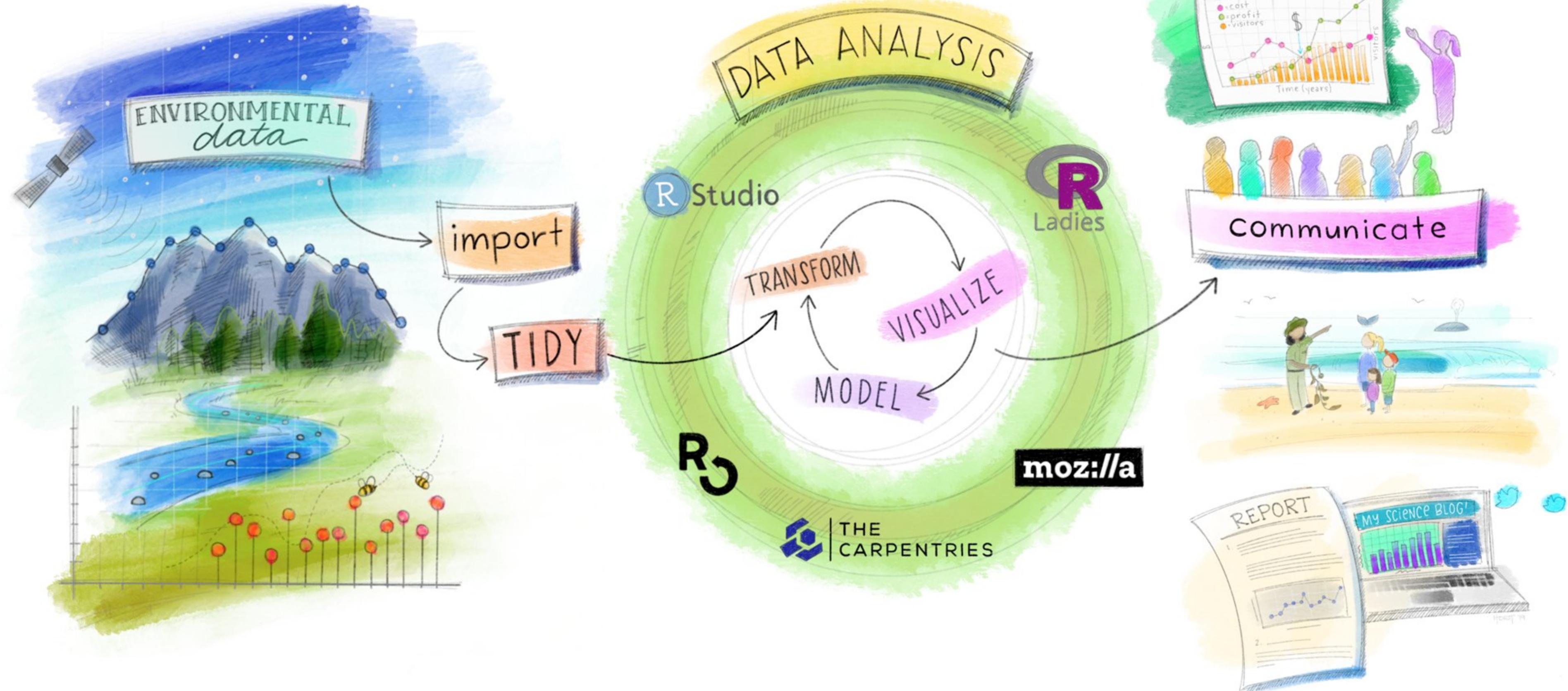
O QUE É A LINGUAGEM R ?

Linguagem de programação gratuita e open-source (funções pré-prontas)

Ambiente de desenvolvimento: análise de dados, Modelos, Dashboards



AUTONOMIA DE PESQUISA





BOM
PROGRAMADOR
= SUPORTE DA
COMUNIDADE

@allison_horst

R: Engine



RStudio: Dashboard



RStudio®

A INTERFACE

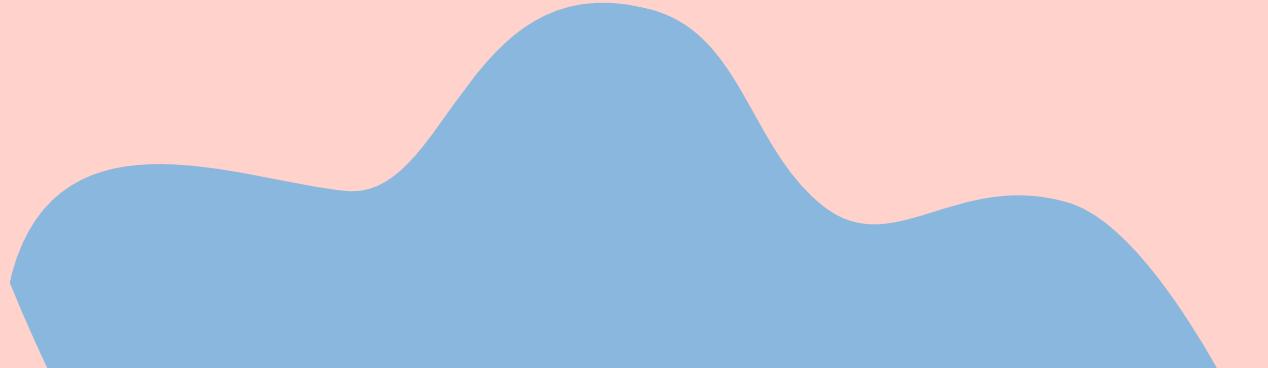
The RStudio interface is organized into several main panels:

- Editor:** Top-left panel showing an R script named "mpg-plot.R" with code to load ggplot2 and create a scatter plot.
- Console:** Bottom-left panel showing the R command-line interface with the same plotting code run.
- Output:** Bottom-right panel displaying the resulting ggplot2 scatter plot of highway fuel economy (hwy) versus engine displacement (displ), where points are colored by vehicle class.
- Environment:** Top-right panel showing the global environment, which is currently empty.

Large blue labels with white outlines identify each panel: "Editor", "Console", "Output", and "Environment".



R 101



A lógica do R

- Computador precisa de contexto
- E se um computador fosse fazer um sanduíche?
- Declarar e informar os ingredientes
- Operacionar as instruções do preparo



```
#Declarar e atribuir valor  
oi <- "Olá mundo"  
# Instrução de imprimir valor  
print(oi)
```

R base: instruções pré-prontas

```
numeros <- c(100 , 50 , 20)  
print(numeros)  
mean(numeros)
```

Formato de leitura de arquivo

- pdf
- excel
- csv
- rds
- etc

Tipo de dados

- data
- integer: números inteiros
- numeric: números decimais
- complex: números complexos
- logical: FALSE, TRUE, NA
- factor: categórica ex: (“ótimo”, “bom”, “médio”, “ruim”)
- character: texto ex: (“eu amo bolo de chocolate”)

tidyverse

tidy

UnitTests

lubridate

gridSVG

dplyr

stringr

rcats

Packages are the fundamental units of reproducible R code.

```
say('time')
#>
#> -----
#> 2020-02-06 10:59:19
#> -----
#> \ \
#>   \ \
#>     \
#>       \|_|
#>       ==) ^Y^ (==
#>       \ ^ /
#>       )=*=(
#>       /   \
#>       |   |
#>       /|_|_|\
#>       \|_|_|/\_
#>       jgs //__\_\_
#>                   \_
```

PACOTES OU NOSSA CAIXA DE FERRAMENTAS ☐

- Pacotes = Caixa de ferramentas
- Open source e a infinidade de pacotes
- Coleção de funções, dados, documentação, etc

QUAIS PACOTES VAMOS USAR?

- Subir os dados
- Filtrar e organizar os dados
- Criar visualizações

QUAIS DADOS VAMOS USAR?

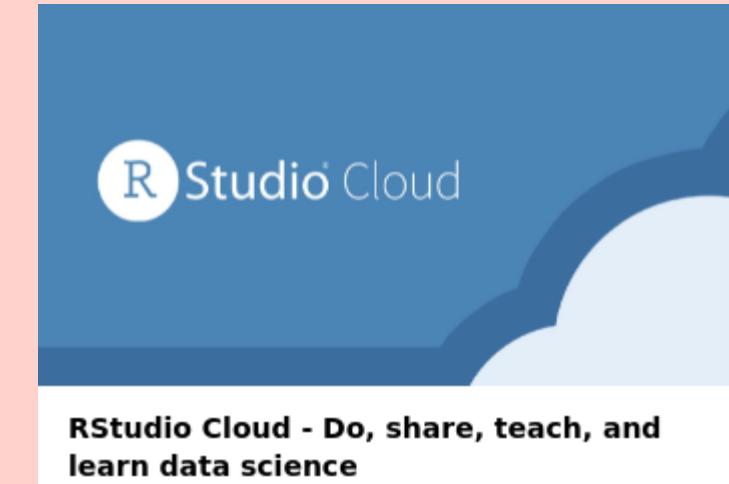
Índice de Desenvolvimento da Educação do Estado de São Paulo – IDESP (2007 a 2019)

- indicador que mede a qualidade das escolas
- O IDESP é composto por dois critérios: o desempenho dos alunos nos exames de proficiência do SARESP (o quanto aprenderam) e o fluxo escolar (em quanto tempo aprenderam). Estes dois critérios se complementam na avaliação da qualidade de ensino oferecido pela escola e permite o acompanhamento de sua evolução de ano para ano.

Índice de Nível Socioeconômico (INSE) por escola

- Calculado a partir dos questionários do SARESP, o INSE considera as seguintes variáveis:
 - Grau de escolaridade dos pais;
 - Posse de bens de consumo duráveis na residência;
 - Renda familiar.
- O INSE varia de 0 a 10, sendo 10 a escola com o nível socioeconômico mais baixo e 0 a escola com nível mais alto.

NA PRÁTICA



O pacote dplyr
Pacote que permite manipulação de dados
filtragem de linhas com condicionais
selecionar colunas
mudar valores das colunas
Resumir valores
Agrupar valores

dplyr::filter()

KEEP ROWS THAT
satisfy
your CONDITIONS

keep rows from... this data... ONLY IF... type is "otter" AND site is "bay"
filter(df, type == "otter" & site == "bay")

type	food	site
otter	urchin	bay
Shark	seal	channel
otter	abalone	bay
otter	crab	wharf

@allison_horst

AS CONDICIONAIS

- Maior ou igual a: \geq
- Igual a: $=$
- Menor ou igual a: \leq
- Diferente de: \neq

`dplyr::mutate`
add column(s),
keep existing:



Horst '18

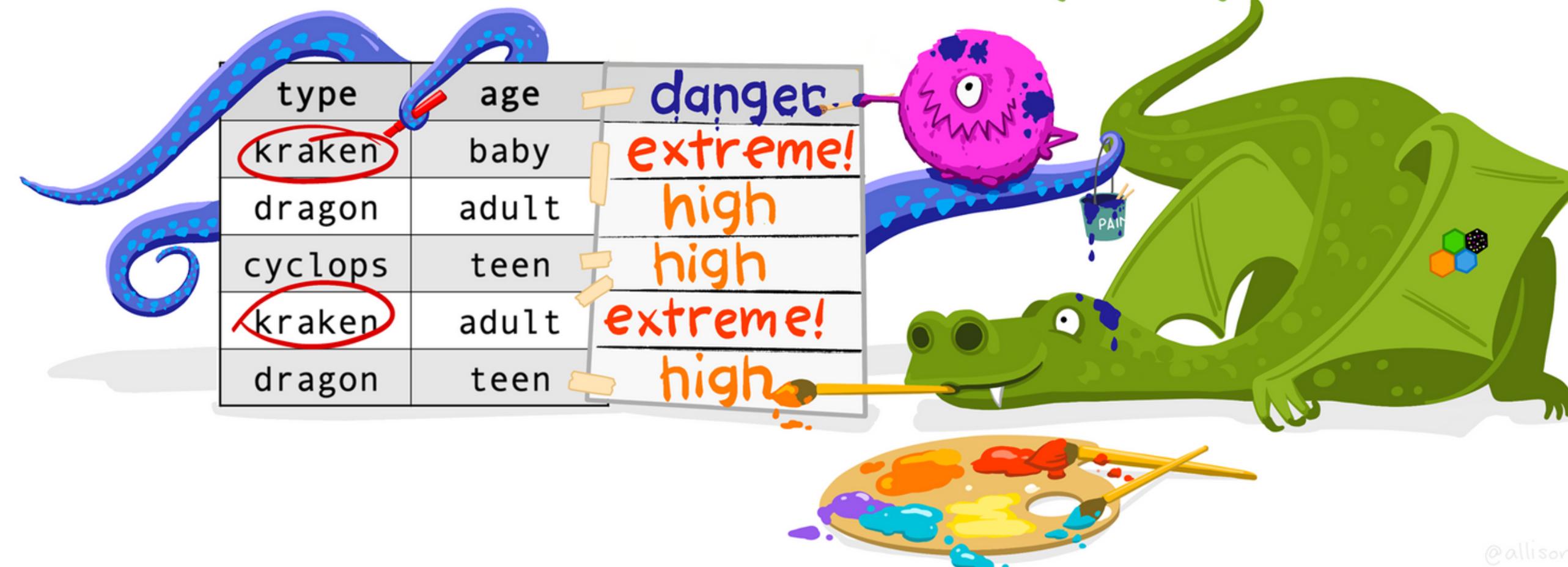
dplyr::case_when()

IF ELSE...
(but you love it?)

df %>% ADD COLUMN
mutate(danger)

IF type is kraken THEN ↓
TRUE ~ "high")
OTHERWISE, danger is high.

danger is
extreme!



@allison_horst

INTERVALO 10 MIN

#

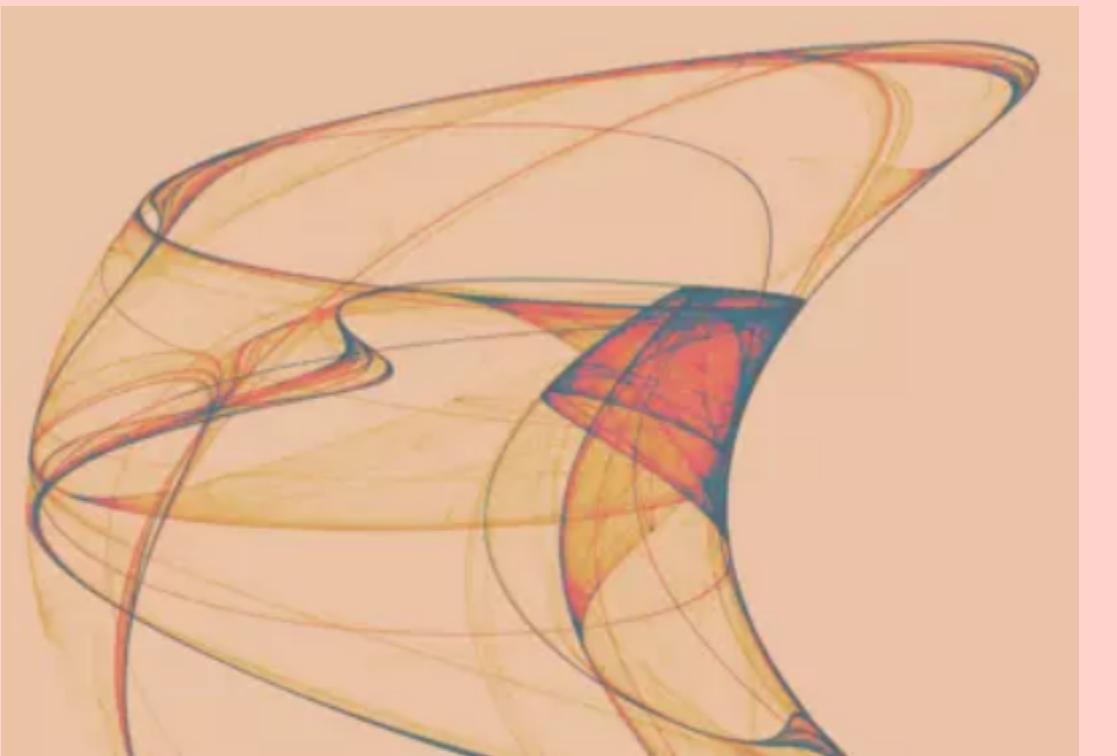
Qualquer número de 0-9 para iniciar um timer



ggplot2:

crie uma OBRA PRIMA
com dados





12 Months of aRt - Will Chase Designs

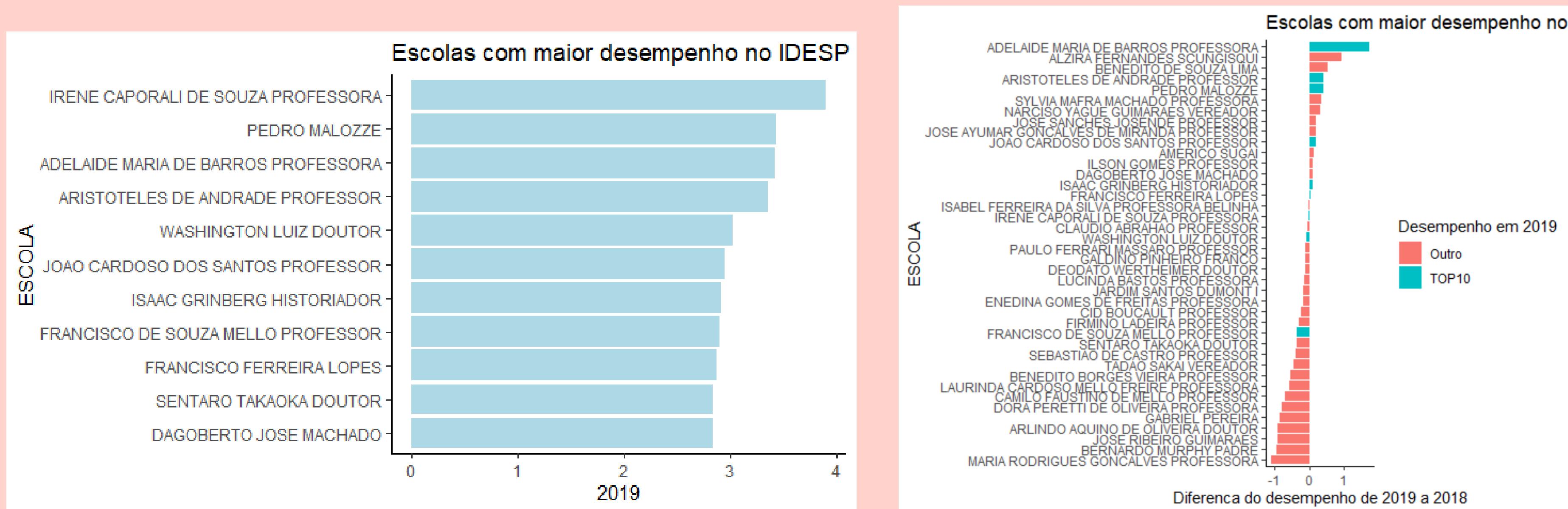
In 2019 I began a project to make a new series of artwork every month made entirely with R. I explore different techniques, develop algorithms, and blog about it each month. All of the code is open sourced, and I...

 williamrchase.com

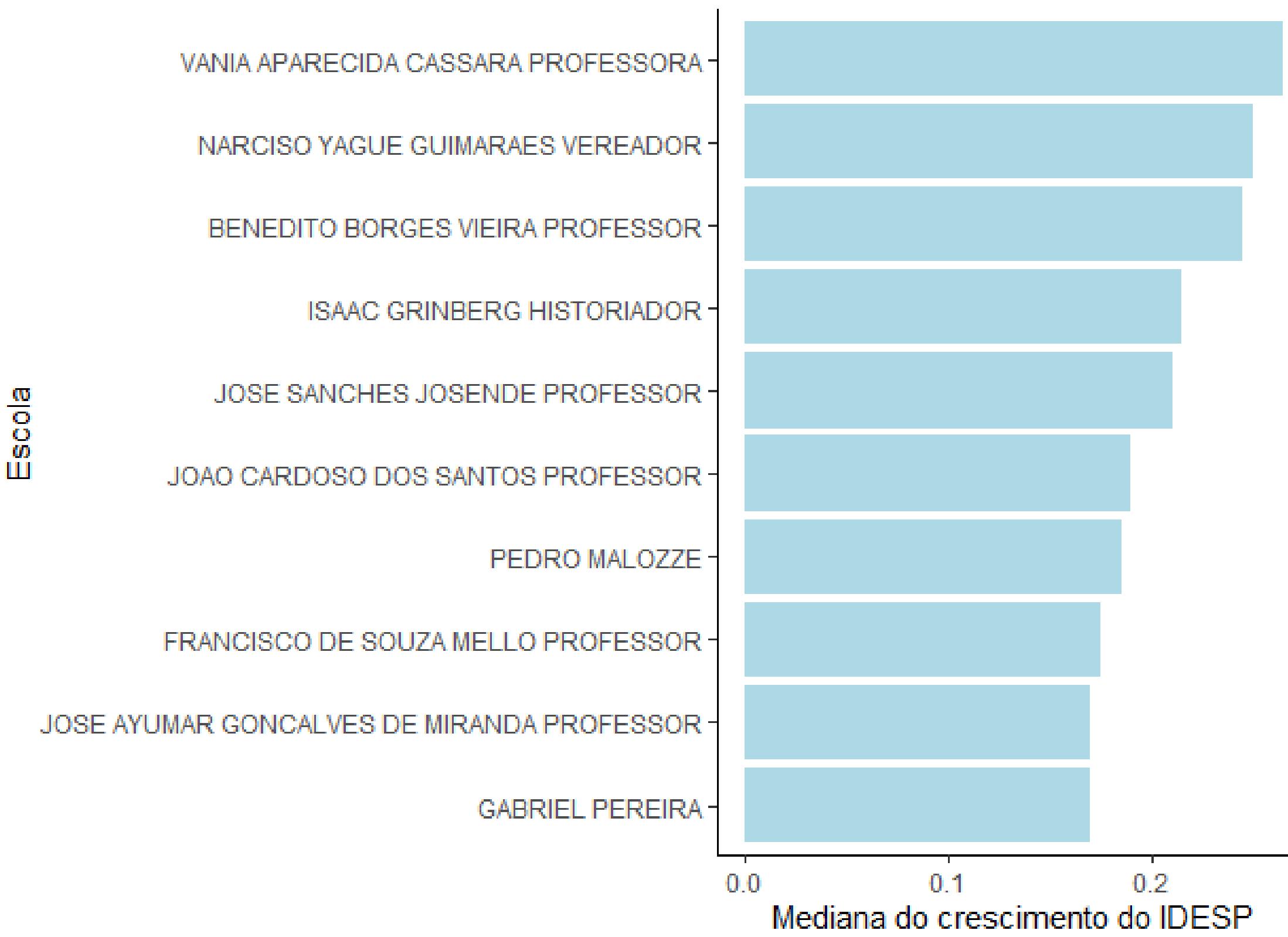
GGPLOT E AS
CAMADAS

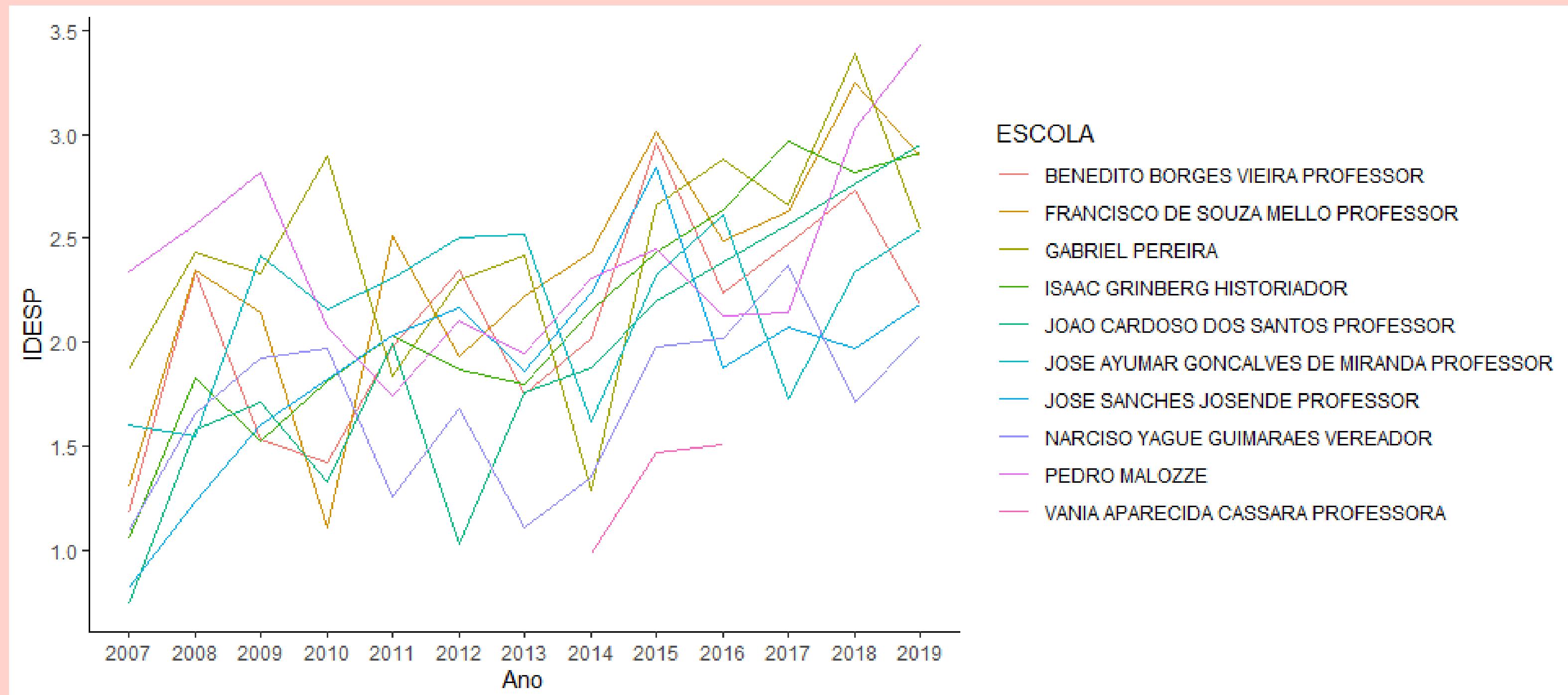


ANALISANDO OS DADOS

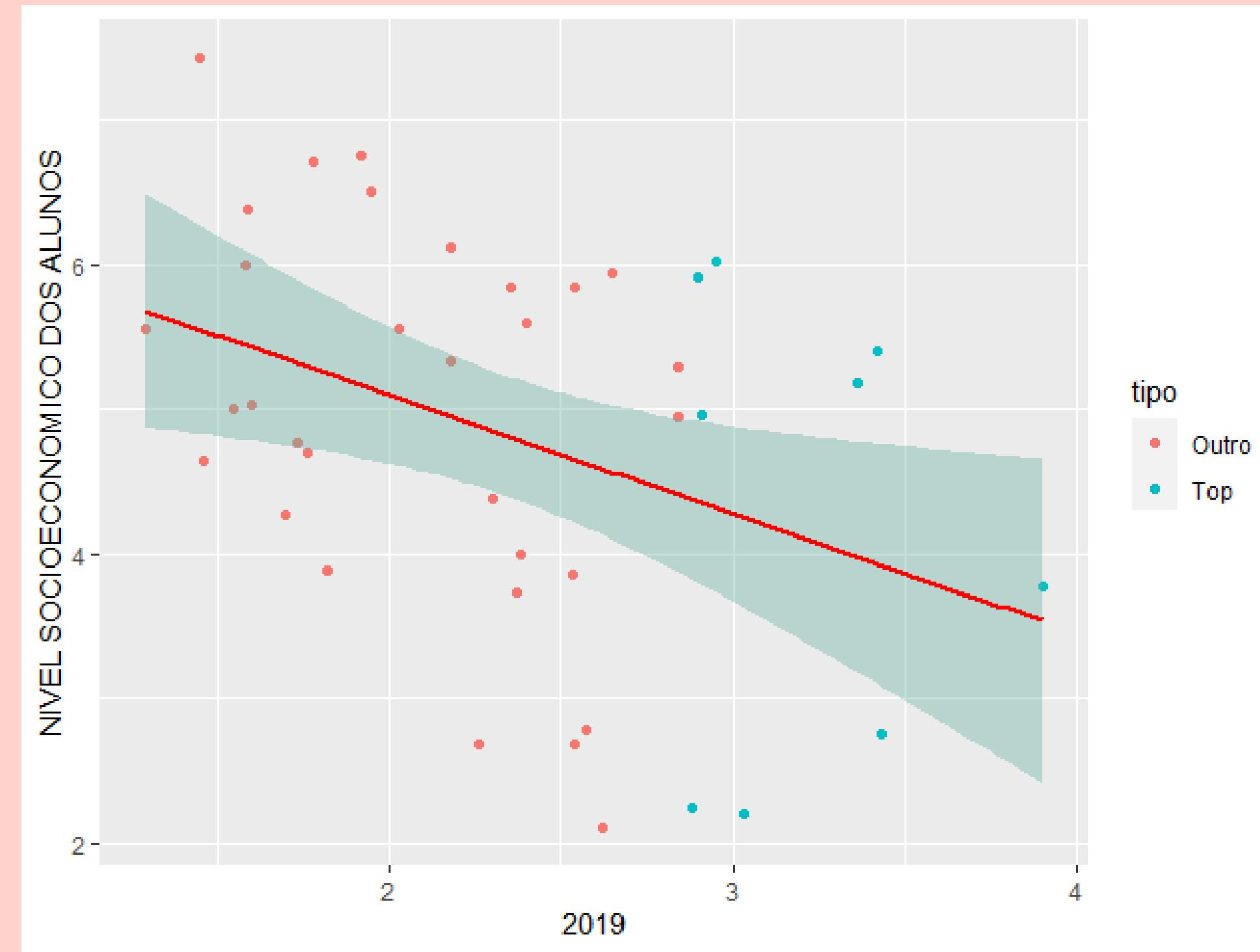


Mediana de crescimento total





ANALISANDO OS DADOS



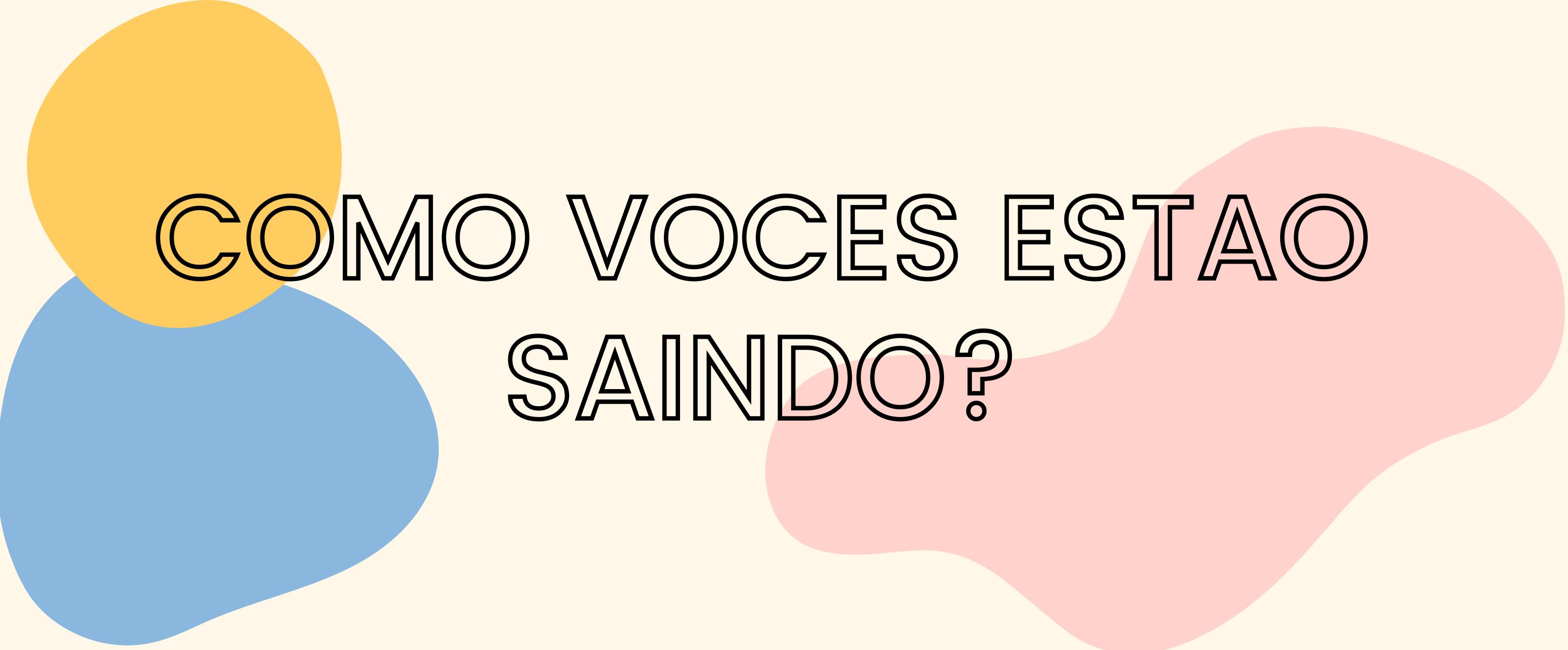
para confete

C



Referencias e próximos passos

- [R4ds](#)
- [RLadies](#)
- [Rstudio Education](#)
- [Txt4cs](#)



**COMO VOCES ESTAO
SAINDO?**

www.menti.com

7217 5852

**A GRADEÇO
SUA ATENÇÃO!**