

R para iniciantes

Taís Maria Nunes Carvalho

R-Ladies Fortaleza

27/07/2020

Olá!

- Engenheira Ambiental (UFC)
- Mestre em Eng. Civil (Recursos Hídricos)
- Estudante de doutorado em Eng. Civil (Recursos Hídricos)

Quem somos?

R-Ladies é uma organização mundial que visa promover a diversidade de gênero na comunidade R.

Fundada pela [Gabriela de Queiroz](#) em 2012. O primeiro encontro foi em São Francisco (EUA).



R-Ladies



Esse é o primeiro encontro do capítulo de Fortaleza 🇧🇷

A linguagem R

Por que R?

- Open source
- Comunidade colaborativa
- Grande disponibilidade de pacotes e ferramentas

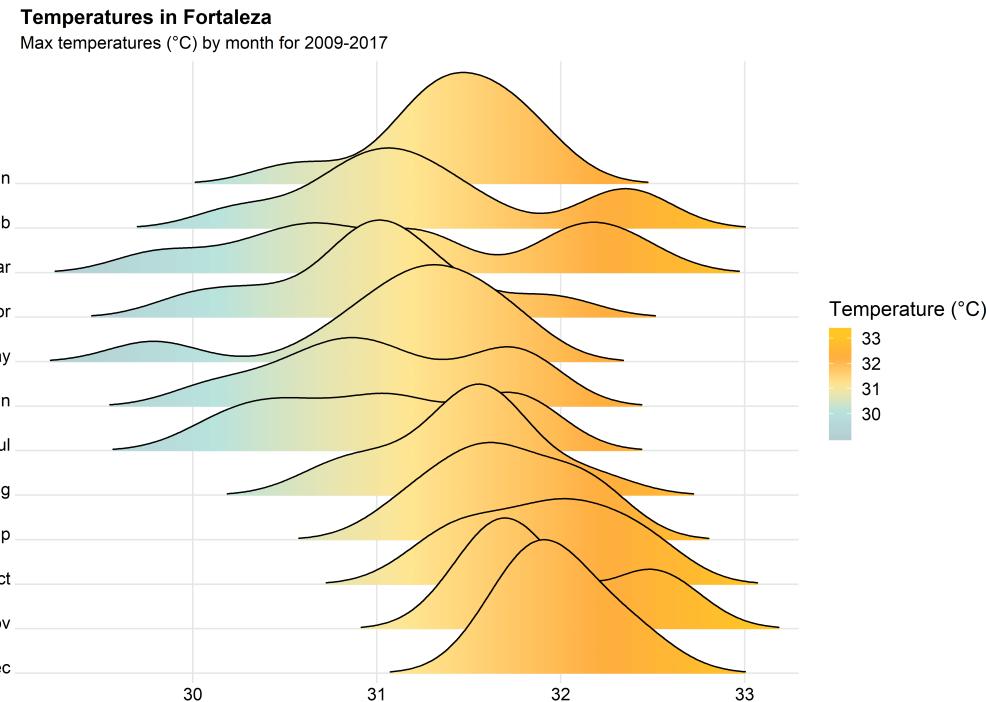
A linguagem R

Comunidade R e onde tirar dúvidas

- Twitter **#rstats**
- Google (sério!)
- **Stack Overflow** (fique atentx às regras da comunidade!)

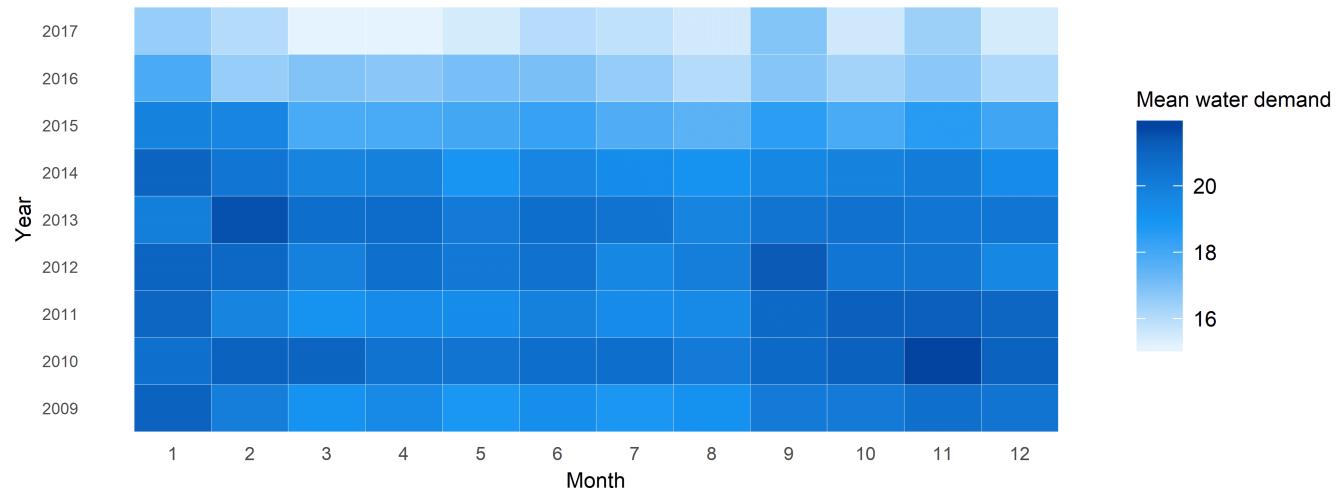
Visualização de dados

- Temperatura máxima em Fortaleza



Visualização de dados

- Demanda residencial de água

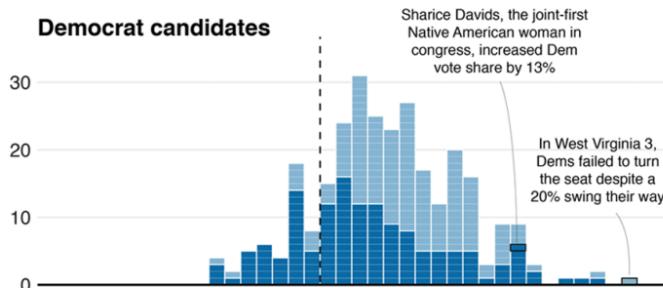


Visualização de dados

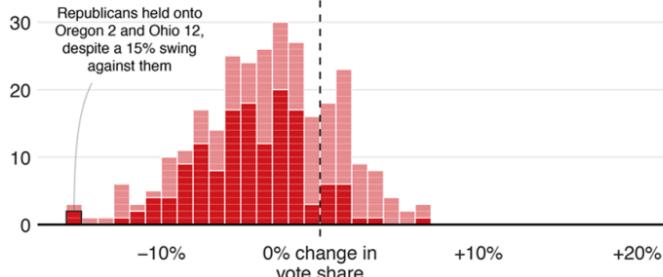
Blue wave

■ Won seat ■ Didn't win

Democrat candidates

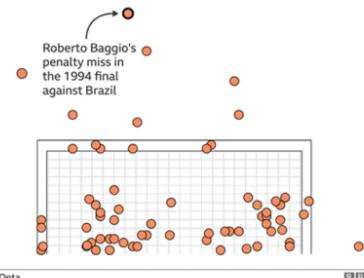


Republican candidates



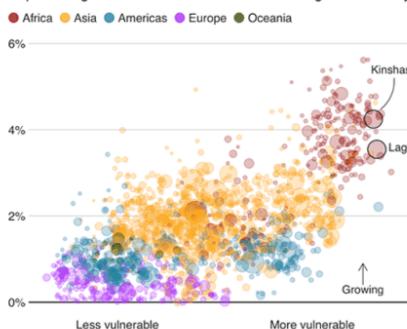
Where penalties are saved

World Cup shootout misses and saves, 1982-2014

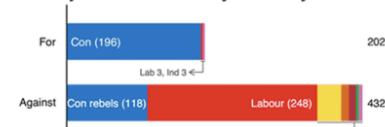


Fast-growing cities face worse climate risks

Population growth 2018-2035 over climate change vulnerability

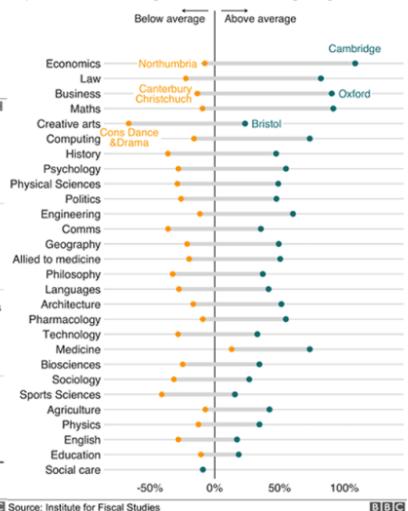


MPs rejected Theresa May's deal by 230 votes



Earnings vary across unis even within subjects

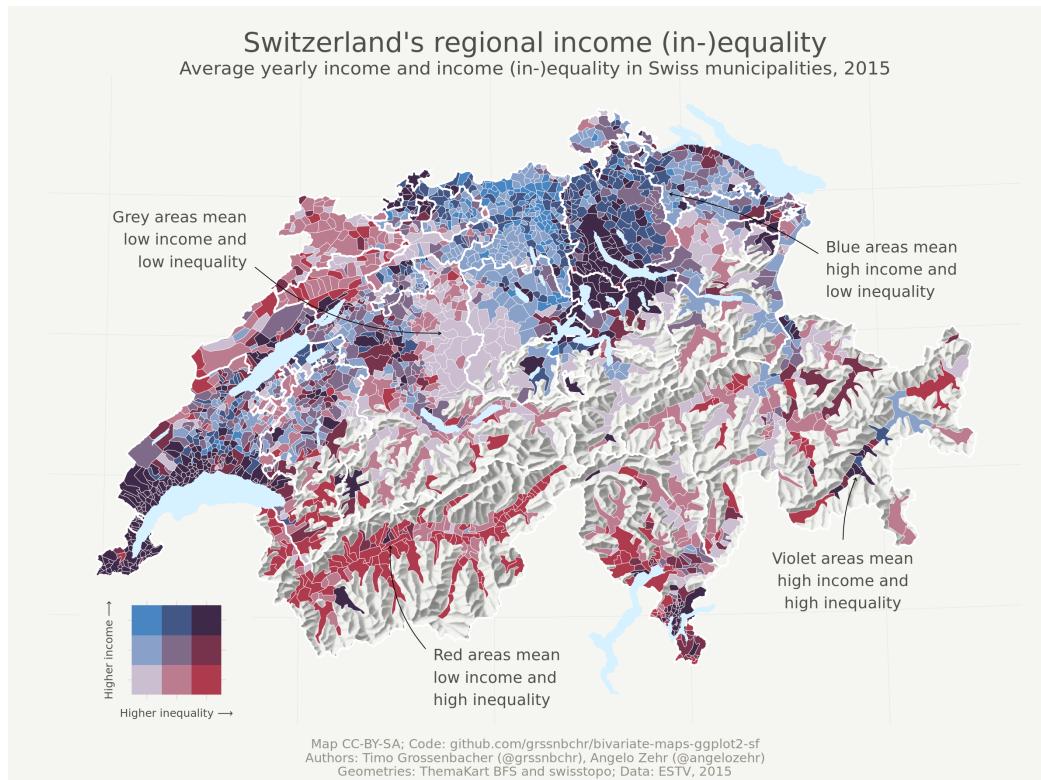
Impact on men's earnings relative to the average degree



Fonte

Visualização de dados

- Mapa temático



Tutorial [aqui](#)

A linguagem R

Algumas aplicações

- Jornalismo de dados
- Ecologia
- Processamento de linguagem natural
 - Análise de clássicos da literatura brasileira - Sillas Gonzaga
- Análise de dados
 - Analisando o seu histórico da Netflix - William Amorim | Curso-R
- Criação de dashboards, relatórios, livros
 - R-Ladies
 - Despesas com educação x IDEB - Fernando Barbalho

Primeiros passos: instalação

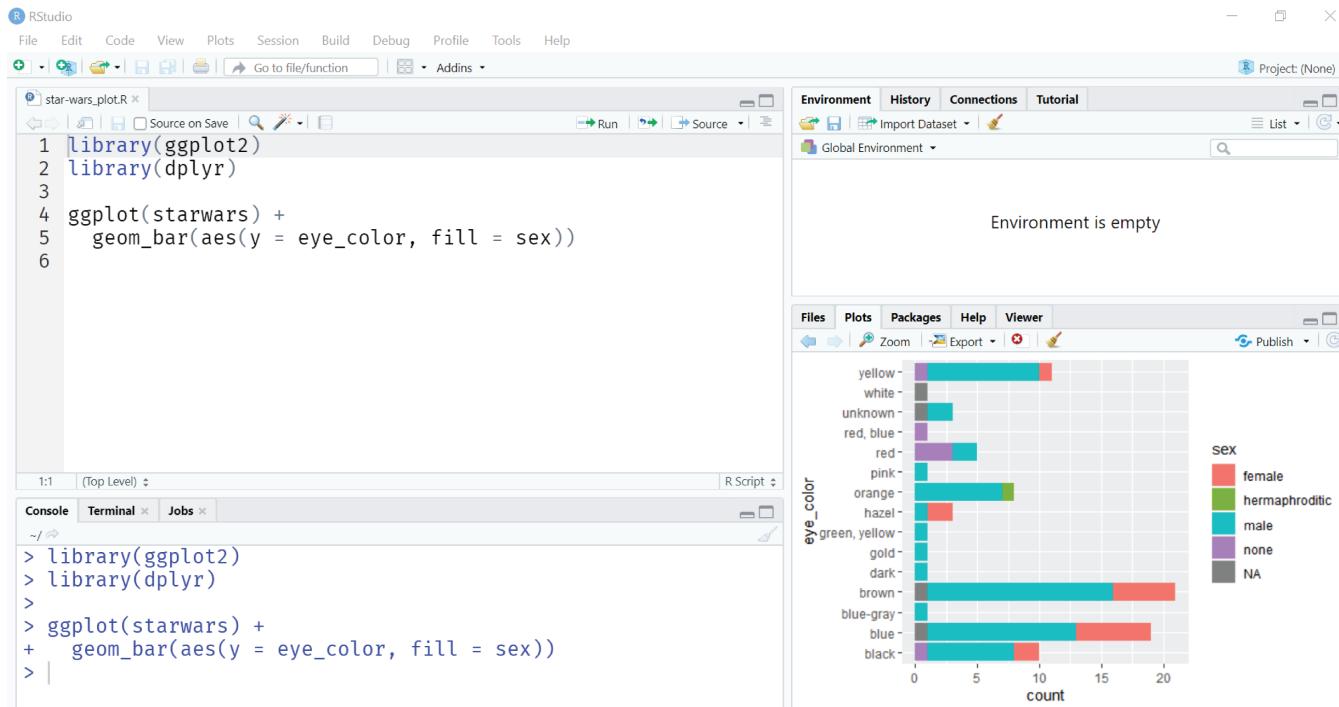
- Instale o R a partir do CRAN
- Baixe o RStudio [aqui](#)

O RStudio é o IDE (Integrated Development Environment) da linguagem R.

- Crie um novo R script: File -> New File -> R Script

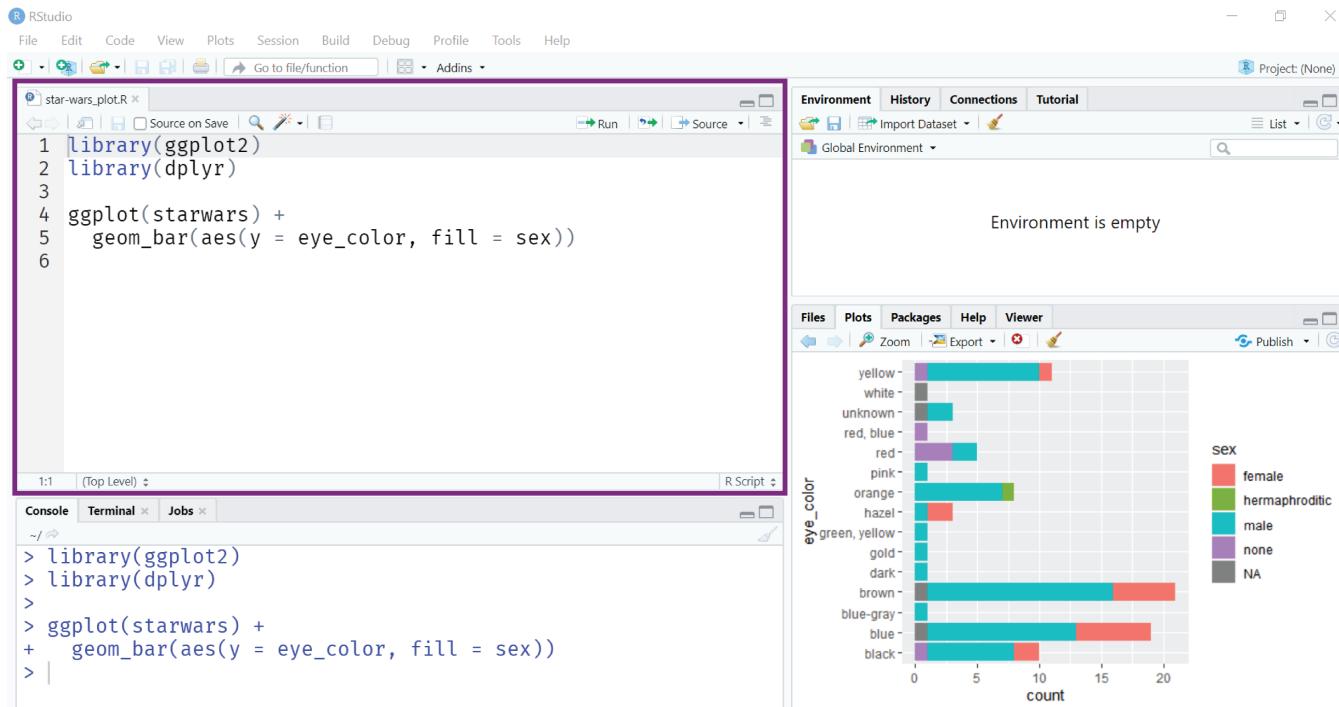
O ambiente do RStudio

File > New File > R Script



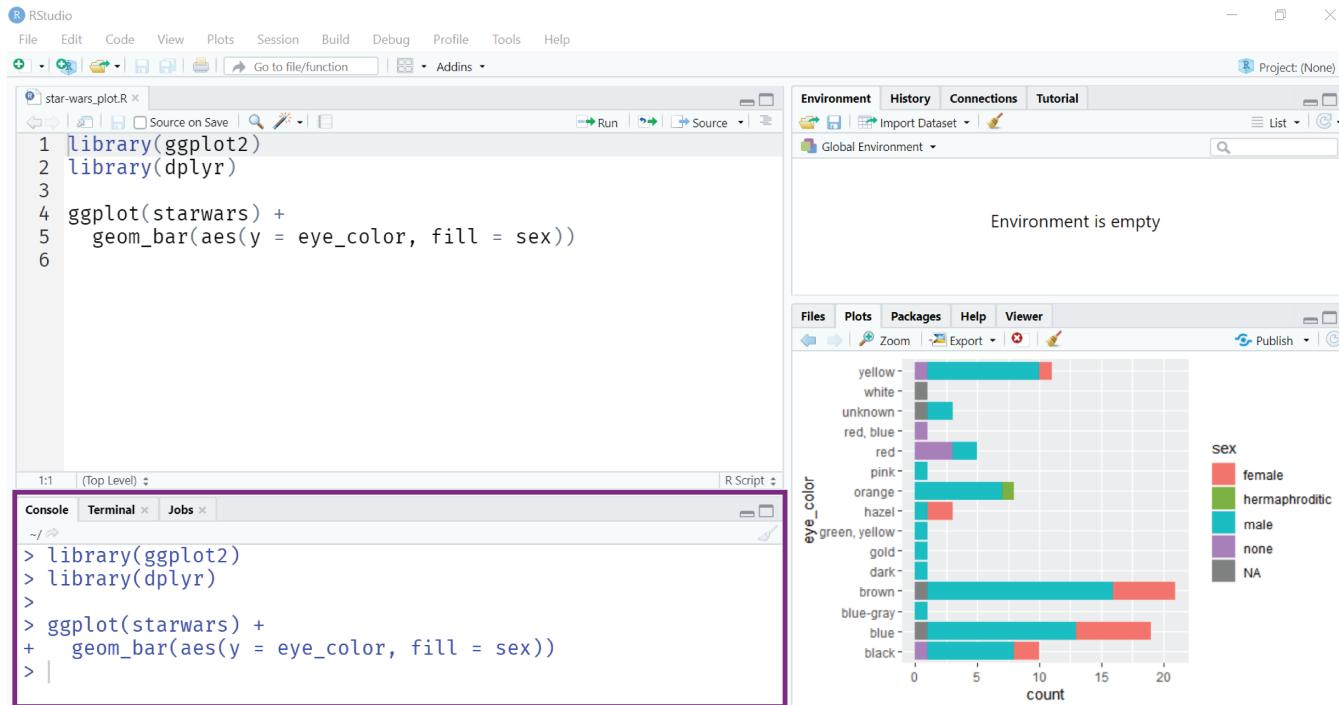
O ambiente do RStudio

Editor



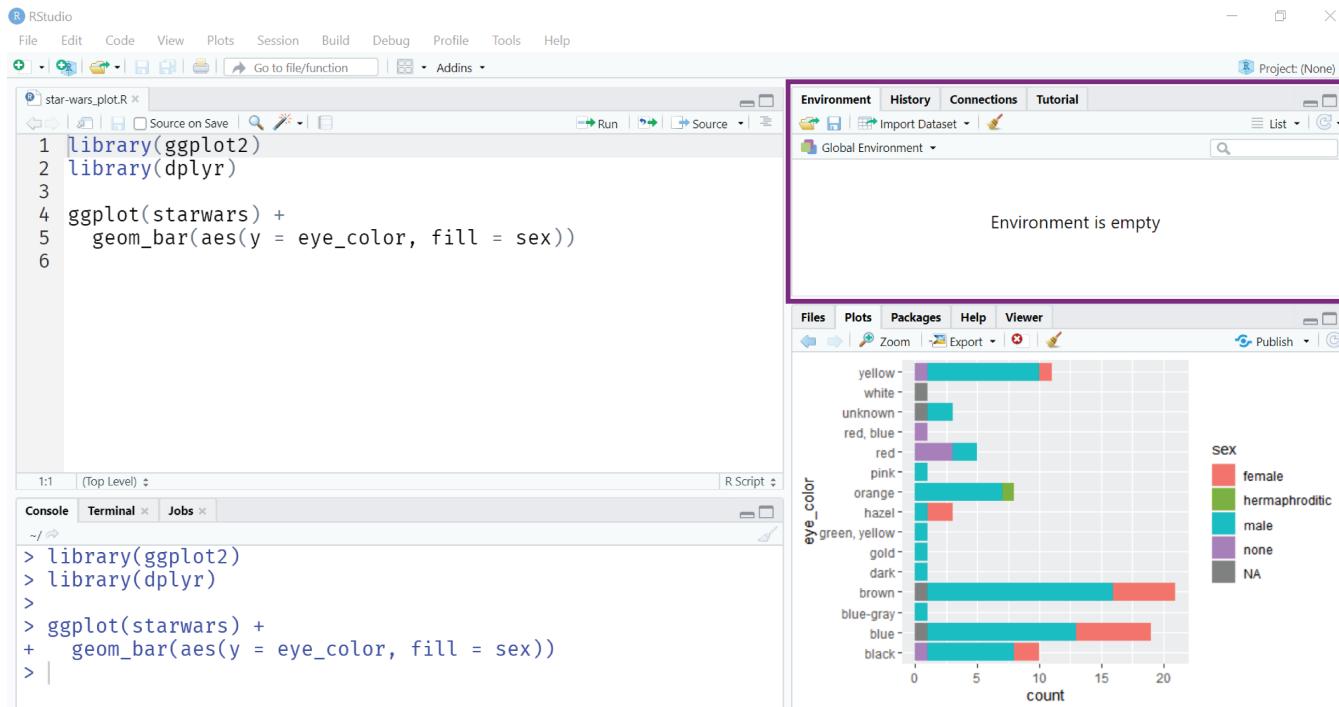
O ambiente do RStudio

Console



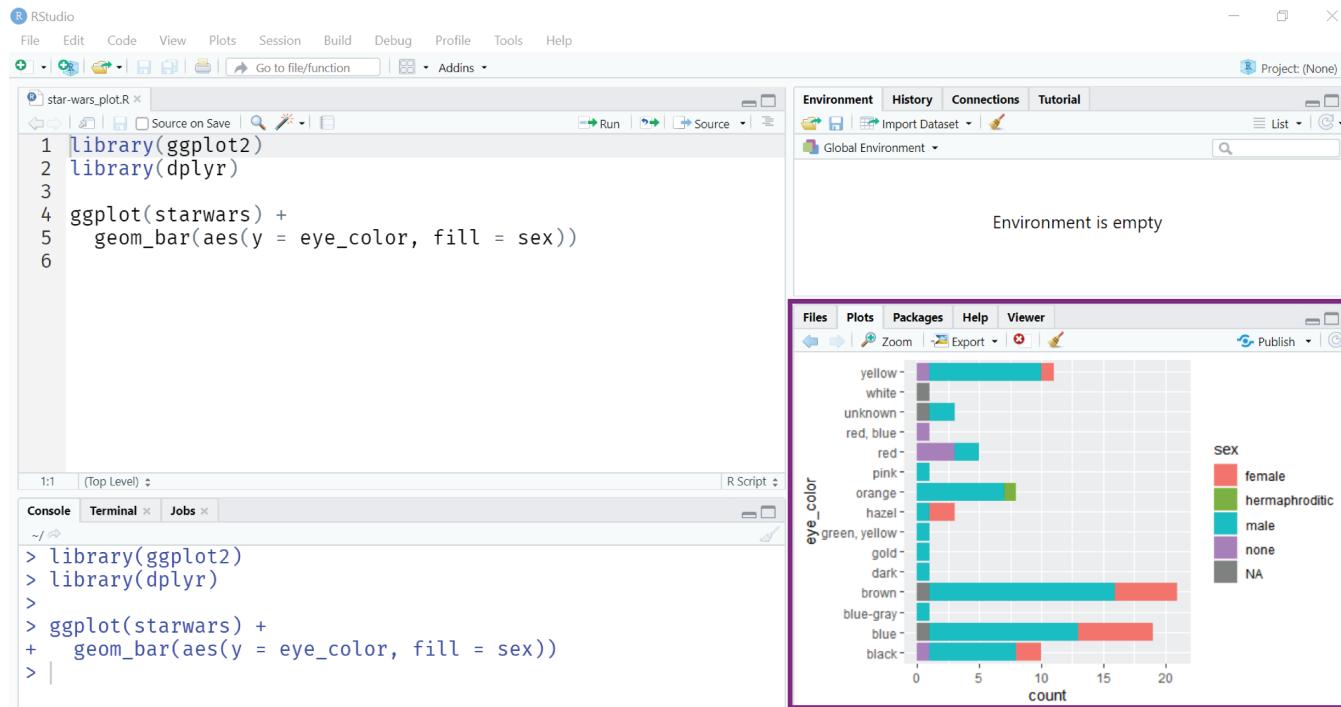
O ambiente do RStudio

Environment



O ambiente do RStudio

Output



Tipos de variáveis

Numéricas e inteiros

```
a <- 2  
b <- 1.5  
c <- 3L  
  
class(a)
```

```
## [1] "numeric"
```

```
class(b)
```

```
## [1] "numeric"
```

```
class(c)
```

```
## [1] "integer"
```

Tipos de variáveis

Caracteres

```
letra <- "R"
mensagem <- "Bem-vindas ao R-Ladies Fortaleza!"
class(letra)
```

```
## [1] "character"
```

```
class(mensagem)
```

```
## [1] "character"
```

Tipos de variáveis

Valores lógicos

```
c <- TRUE  
d <- FALSE  
class(c)  
  
## [1] "logical"
```

```
a <- 2  
b <- 1.5  
b > 6
```

```
## [1] FALSE
```

```
a == b
```

```
## [1] FALSE
```

```
a == b  
  
## [1] FALSE
```

```
a != b #operador "diferente"
```

```
## [1] TRUE
```

```
b > 6 & b <= 2 #operador E
```

```
## [1] FALSE
```

```
b > 6 | a > 1 #operador OU
```

```
## [1] TRUE
```

Operadores aritméticos

```
a + b    #soma
```

```
## [1] 3.5
```

```
a - b    #subtração
```

```
## [1] 0.5
```

```
a / b    #divisão
```

```
## [1] 1.333333
```

```
a * b    #multiplicação
```

```
## [1] 3
```

```
a ^ b    #potenciação
```

```
## [1] 2.828427
```

Estruturas de dados

Vetores

```
custo_produtos <- c(15.50, 20.70, 13.20, 14, 19.30)
class(custo_produtos)
```

```
## [1] "numeric"
```

```
1:6
```

```
## [1] 1 2 3 4 5 6
```

```
cod_produtos <- c("A1", "B2", "A3", "A4", "B5")
class(cod_produtos)
```

```
## [1] "character"
```

Operações com vetores

```
length(custo_produtos)
```

```
## [1] 5
```

```
custo_produtos[1]
```

```
## [1] 15.5
```

```
custo_produtos[-1]
```

```
## [1] 20.7 13.2 14.0 19.3
```

```
custo_produtos[c(1, 3, 5)]
```

```
## [1] 15.5 13.2 19.3
```

```
aumento <- 1.1  
custo_produtos * aumento
```

```
## [1] 17.05 22.77 14.52 15.40 21.23
```

```
custo_produtos + 5
```

```
## [1] 20.5 25.7 18.2 19.0 24.3
```

```
custo_produtos < 19
```

```
## [1] TRUE FALSE TRUE TRUE FALSE
```

```
custo_produtos[custo_produtos < 19]
```

```
## [1] 15.5 13.2 14.0
```

```
custo_produtos[c(TRUE, FALSE, TRUE, TR
```

```
## [1] 15.5 13.2 14.0
```

Estruturas de dados

Dataframes

```
lista_produtos <- data.frame(custo_produtos, cod_produtos)
lista_produtos
```

```
##   custo_produtos cod_produtos
## 1      15.5        A1
## 2      20.7        B2
## 3      13.2        A3
## 4      14.0        A4
## 5      19.3        B5
```

Estruturas de dados

Dataframes

```
nrow(lista_produtos)  
## [1] 5
```

```
ncol(lista_produtos)  
## [1] 2
```

```
colnames(lista_produtos)  
## [1] "custo_produtos" "cod_produtos"
```

```
rownames(lista_produtos)  
## [1] "1" "2" "3" "4" "5"
```

```
# acessar coluna  
# "custo_produtos"  
lista_produtos["custo_produtos"]
```

```
##   custo_produtos  
## 1      15.5  
## 2      20.7  
## 3      13.2  
## 4      14.0  
## 5      19.3
```

```
# outra forma de acessar  
# a coluna "custo_produtos"  
lista_produtos$custo_produtos
```

```
## [1] 15.5 20.7 13.2 14.0 19.3
```

Funções

```
custo_produtos
```

```
## [1] 15.5 20.7 13.2 14.0 19.3
```

```
max(custo_produtos)
```

```
## [1] 20.7
```

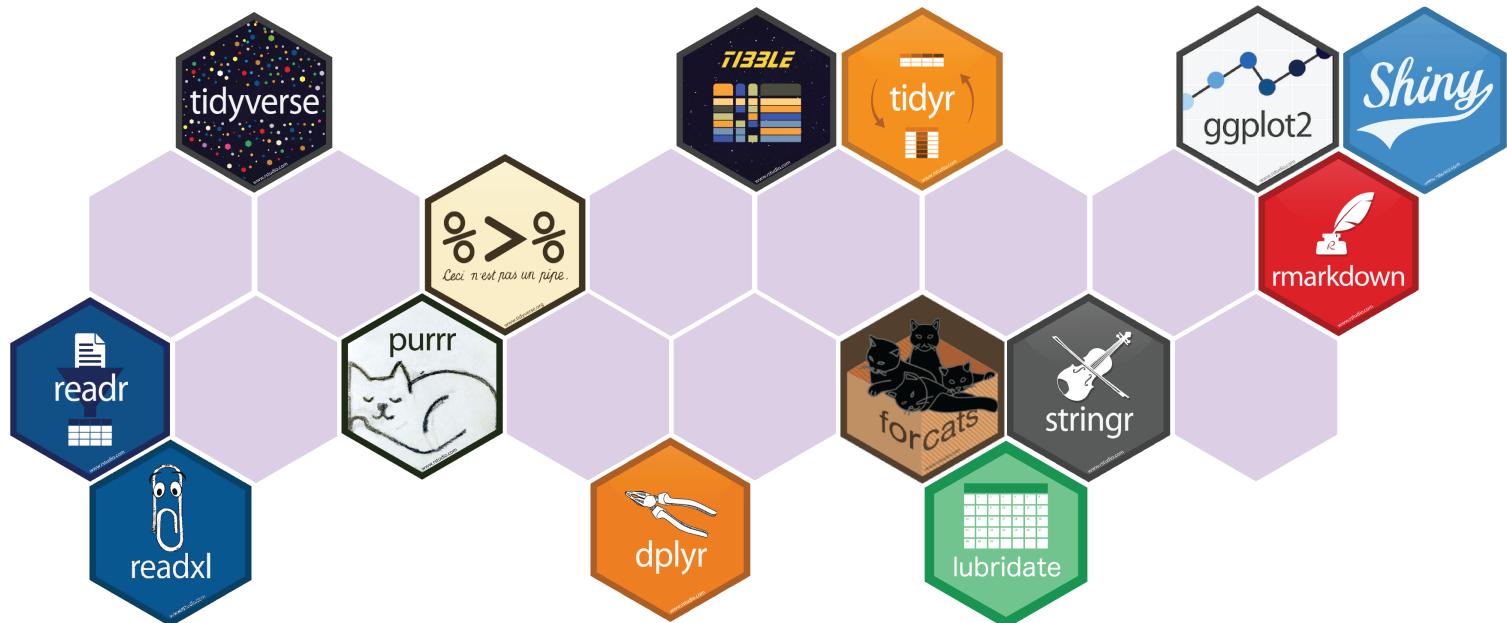
```
mean(custo_produtos)
```

```
## [1] 16.54
```

```
num <- 15.78364  
round(x = num, digits = 2)
```

```
## [1] 15.78
```

Tidyverse



Documentação do Tidyverse

O operador pipe %>%

```
custo_produtos
```

```
## [1] 15.5 20.7 13.2 14.0 19.3
```

```
mean(custo_produtos)
```

```
## [1] 16.54
```

```
custo_produtos %>% mean()
```

```
## [1] 16.54
```

Atalho para usar o %>%:

ctrl + shift + M



Pacote dplyr

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

Base de dados: starwars

```
## # A tibble: 87 x 14
##   name    height  mass hair_color skin_color eye_color birth_year sex gender homeworld
##   <chr>    <int> <dbl> <chr>       <chr>      <chr>          <dbl> <chr> <chr> <chr>
## 1 Luke~     172     77 blond       fair        blue            19   male  masculin~ Tatooine
## 2 C-3PO      167     75 <NA>        gold        yellow         112   none  masculin~ Tatooine
## 3 R2-D2      96      32 <NA>        white, bl~ red             33   none  masculin~ Naboo
## 4 Dart~     202     136 none        white        yellow         41.9  male  masculin~ Tatooine
## 5 Leia~     150      49 brown       light        brown           19   femal feminin~ Alderaan
## 6 Owen~     178     120 brown, gr~ light        blue            52   male  masculin~ Tatooine
## 7 Beru~     165      75 brown       light        blue            47   femal feminin~ Tatooine
## 8 R5-D4      97      32 <NA>        white, red red             NA   none  masculin~ Tatooine
## 9 Bigg~     183      84 black       light        brown           24   male  masculin~ Tatooine
## 10 Obi-~    182      77 auburn, w~ fair        blue-gray        57   male  masculin~ Stewjon
## # ... with 77 more rows, and 4 more variables: species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

Funções do pacote dplyr

Selecionar variáveis: select

```
library(dplyr)
starwars %>%
  select(name, height, mass)

## # A tibble: 87 x 3
##   name           height  mass
##   <chr>        <int> <dbl>
## 1 Luke Skywalker    172    77
## 2 C-3PO             167    75
## 3 R2-D2              96    32
## 4 Darth Vader       202   136
## 5 Leia Organa        150    49
## 6 Owen Lars           178   120
## 7 Beru Whitesun lars  165    75
## 8 R5-D4              97    32
## 9 Biggs Darklighter   183    84
## 10 Obi-Wan Kenobi     182    77
## # ... with 77 more rows
```

Funções do pacote dplyr

Criar, modificar e deletar colunas: mutate

```
starwars %>%  
  select(name, height, mass) %>%  
  mutate(imc = mass/height^2)  
  
## # A tibble: 87 x 4  
##   name           height   mass     imc  
##   <chr>         <int>   <dbl>    <dbl>  
## 1 Luke Skywalker      172     77  0.00260  
## 2 C-3PO              167     75  0.00269  
## 3 R2-D2                96     32  0.00347  
## 4 Darth Vader          202    136  0.00333  
## 5 Leia Organa           150     49  0.00218  
## 6 Owen Lars              96     32  0.00379  
## 7 Beru Whitesun lars     165     75  0.00275  
## 8 R5-D4                97     32  0.00340  
## 9 Biggs Darklighter       183     84  0.00251  
## 10 Obi-Wan Kenobi        182     77  0.00232  
## # ... with 77 more rows
```

Funções do pacote dplyr

Escolher observações com base nos seus valores: filter

```
starwars %>%
  select(name, species, height, mass) %>%
  filter(species == "Human") %>%
  mutate(imc = mass/height^2)

## # A tibble: 35 x 5
##   name      species  height  mass     imc
##   <chr>     <chr>    <int> <dbl>    <dbl>
## 1 Luke Skywalker Human     172    77  0.00260
## 2 Darth Vader  Human     202   136  0.00333
## 3 Leia Organa Human     150     49  0.00218
## 4 Owen Lars   Human     178   120  0.00379
## 5 Beru Whitesun lars Human     165    75  0.00275
## 6 Biggs Darklighter Human     183    84  0.00251
## 7 Obi-Wan Kenobi Human     182    77  0.00232
## 8 Anakin Skywalker Human     188    84  0.00238
## 9 Wilhuff Tarkin Human     180    NA  NA
## 10 Han Solo   Human     180    80  0.00247
## # ... with 25 more rows
```

Funções do pacote dplyr

Escolher observações com base nos seus valores: filter

```
starwars %>%  
  select(name, species, height, mass) %>%  
  filter(mass > 100 & height > 110) %>%  
  mutate(imc = mass/height^2)
```

```
## # A tibble: 10 x 5  
##   name           species   height   mass     imc  
##   <chr>          <chr>     <int>   <dbl>    <dbl>  
## 1 Darth Vader   Human      202     136  0.00333  
## 2 Owen Lars    Human      178     120  0.00379  
## 3 Chewbacca    Wookiee    228     112  0.00215  
## 4 Jabba Desilijic Tiure Hutt       175    1358 0.0443  
## 5 Jek Tono Porkins Human      180     110  0.00340  
## 6 IG-88         Droid      200     140  0.0035  
## 7 Bossk        Trandoshan  190     113  0.00313  
## 8 Dexter Jettster Besalisk    198     102  0.00260  
## 9 Grievous     Kaleesh     216     159  0.00341  
## 10 Tarfful      Wookiee    234     136  0.00248
```

Funções do pacote dplyr

Agrupar e resumir valores: group_by e summarise

```
starwars %>%
  select(name, species, gender, height, mass) %>%
  filter(species == "Human") %>%
  mutate(imc = mass/height^2) %>%
  group_by(gender) %>%
  summarise(mean = mean(imc, na.rm = T))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2
##   gender      mean
##   <chr>      <dbl>
## 1 feminine  0.00220
## 2 masculine 0.00260
```

Funções do pacote dplyr

Mudar a ordem das linhas: arrange

```
starwars %>%  
  select(name, species, gender, height, mass) %>%  
  filter(species == "Human") %>%  
  arrange(desc(height))
```

```
## # A tibble: 35 x 5  
##   name           species gender  height  mass  
##   <chr>          <chr>   <chr>    <int> <dbl>  
## 1 Darth Vader   Human   masculine  202   136  
## 2 Qui-Gon Jinn  Human   masculine  193    89  
## 3 Dooku          Human   masculine  193    80  
## 4 Bail Prestor Organa Human   masculine  191    NA  
## 5 Anakin Skywalker Human   masculine  188    84  
## 6 Mace Windu    Human   masculine  188    84  
## 7 Raymus Antilles Human   masculine  188    79  
## 8 Gregar Typho   Human   masculine  185    85  
## 9 Biggs Darklighter Human   masculine  183    84  
## 10 Boba Fett     Human   masculine  183   78.2  
## # ... with 25 more rows
```

Funções do pacote dplyr

Mudar a ordem das linhas: arrange

```
starwars %>%  
  select(name, species, gender, height, mass) %>%  
  filter(species == "Human") %>%  
  group_by(gender) %>%  
  arrange(desc(height), .by_group = TRUE)
```

```
## # A tibble: 35 x 5  
## # Groups:   gender [2]  
##   name           species gender   height   mass  
##   <chr>          <chr>   <chr>     <int>   <dbl>  
## 1 Jocasta Nu    Human   feminine  167     NA  
## 2 Beru Whitesun lars Human   feminine  165     75  
## 3 Dormé          Human   feminine  165     NA  
## 4 Padmé Amidala Human   feminine  165     45  
## 5 Shmi Skywalker Human   feminine  163     NA  
## 6 Cordé          Human   feminine  157     NA  
## 7 Leia Organa    Human   feminine  150     49  
## 8 Mon Mothma    Human   feminine  150     NA  
## 9 Rey            Human   feminine  NA      NA  
## 10 Darth Vader   Human   masculine 202     136  
## # ... with 25 more rows
```

Tipos de união: inner join

```
band_members
```

```
## # A tibble: 3 x 2
##   name   band
##   <chr> <chr>
## 1 Mick  Stones
## 2 John  Beatles
## 3 Paul  Beatles
```

```
band_members %>%
  inner_join(band_instruments)
```

```
## Joining, by = "name"

## # A tibble: 2 x 3
##   name   band     plays
##   <chr> <chr> <chr>
## 1 John  Beatles guitar
## 2 Paul  Beatles bass
```

```
band_instruments
```

```
## # A tibble: 3 x 2
##   name   plays
##   <chr> <chr>
## 1 John  guitar
## 2 Paul  bass
## 3 Keith guitar
```

```
inner_join(x, y)
```

1	x1	1	y1
2	x2	2	y2
3	x3	4	y4

Tipos de união: left join

```
band_members
```

```
## # A tibble: 3 x 2
##   name   band
##   <chr> <chr>
## 1 Mick  Stones
## 2 John  Beatles
## 3 Paul  Beatles
```

```
band_members %>%
  left_join(band_instruments)
```

```
## Joining, by = "name"

## # A tibble: 3 x 3
##   name   band     plays
##   <chr> <chr>    <chr>
## 1 Mick  Stones  <NA>
## 2 John  Beatles guitar
## 3 Paul  Beatles bass
```

```
band_instruments
```

```
## # A tibble: 3 x 2
##   name   plays
##   <chr> <chr>
## 1 John  guitar
## 2 Paul  bass
## 3 Keith guitar
```

```
left_join(x, y)
```

1	x1	1	y1
2	x2	2	y2
3	x3	4	y4

Tipos de união: right join

```
band_members
```

```
## # A tibble: 3 x 2
##   name   band
##   <chr> <chr>
## 1 Mick  Stones
## 2 John  Beatles
## 3 Paul  Beatles
```

```
band_members %>%
  right_join(band_instruments)
```

```
## Joining, by = "name"

## # A tibble: 3 x 3
##   name   band     plays
##   <chr> <chr>    <chr>
## 1 John  Beatles  guitar
## 2 Paul  Beatles  bass
## 3 Keith <NA>     guitar
```

```
band_instruments
```

```
## # A tibble: 3 x 2
##   name   plays
##   <chr> <chr>
## 1 John  guitar
## 2 Paul  bass
## 3 Keith guitar
```

```
right_join(x, y)
```

1	x1	1	y1
2	x2	2	y2
3	x3	4	y4

Tipos de união: full join

band_members

```
## # A tibble: 3 x 2
##   name   band
##   <chr> <chr>
## 1 Mick  Stones
## 2 John  Beatles
## 3 Paul  Beatles
```

```
band_members %>%
  full_join(band_instruments)
```

```
## Joining, by = "name"

## # A tibble: 4 x 3
##   name   band     plays
##   <chr> <chr>    <chr>
## 1 Mick  Stones  <NA>
## 2 John  Beatles guitar
## 3 Paul  Beatles bass
## 4 Keith <NA>    guitar
```

band_instruments

```
## # A tibble: 3 x 2
##   name   plays
##   <chr> <chr>
## 1 John  guitar
## 2 Paul  bass
## 3 Keith guitar
```

full_join(x, y)

1	x1	1	y1
2	x2	2	y2
3	x3	4	y4

Trabalhando com dados reais

- Base de dados: Educação (Ensino Fundamental)

IPECE - Instituto de pesquisa e estratégia econômica do Ceará

Abrangência Geográfica	Ano	regiao mapa	Nome Curto Indicador	regiao_metropolitana	regiao_planejamento	Tema	Abrangencia Geografica	Ano comple
Abaiara	2008	Abaiara	Abandono no ensino fundamental	Municípios não Metropolitanos	Cariri	Ensino fundamental	NA	01/01/2008
Altaneira	2015	Altaneira	Alunos avaliados - SPAECE 5º ano - Rede estadual	Municípios não Metropolitanos	Cariri	Ensino fundamental	Estado, Municípios	01/01/2015

Preparação dos dados

```
library(readr)

dados_ens_fund <- read_delim("data/data_ens_fundamental.csv",
                               delim = ";")

colnames(dados_ens_fund)

## [1] "Abrangência Geografica"
## [3] "regiao mapa"
## [5] "regiao_metropolitana"
## [7] "Tema"
## [9] "Ano completo"
## [11] "Dimensao"
## [13] "Frequencia"
## [15] "Nome Indicador"
## [17] "nome_tabela"
## [19] "Rede Grupo"
## [21] "Unidade Geográfica"
## [23] "Numero de Municípios"
## [25] "Valor"

                                         "Ano"
                                         "Nome Curto Indicador"
                                         "regiao_planejamento"
                                         "Abrangencia Geografica"
                                         "Cod Ibge"
                                         "Fonte"
                                         "Municipio"
                                         "Nome Indicador + Unidade de Medida "
                                         "Origem"
                                         "regionalizado"
                                         "Unidade Medida"
                                         "Número de registros"
```

Preparação dos dados

```
library(janitor)

dados_ens_fund <- read_delim("data/data_ens_fundamental.csv",
                               delim = ";") %>%
  janitor::clean_names()

colnames(dados_ens_fund)

## [1] "abrangencia_geografica"                 "ano"
## [3] "regiao_mapa"                           "nome_curto_indicador"
## [5] "regiao_metropolitana"                  "regiao_planejamento"
## [7] "tema"                                  "abrangencia_geografica_2"
## [9] "ano_completo"                         "cod_ibge"
## [11] "dimensao"                            "fonte"
## [13] "frequencia"                          "municipio"
## [15] "nome_indicador"                     "nome_indicador_unidade_de_medida"
## [17] "nome_tabela"                         "origem"
## [19] "rede_grupo"                          "regionalizado"
## [21] "unidade_geografica"                 "unidade_medida"
## [23] "numero_de_municipios"               "numero_de_registros"
## [25] "valor"
```

Preparação dos dados

```
library(dplyr)

dados_ens_fund <- read_delim("data/data_ens_fundamental.csv",
                               delim = ";") %>%
  janitor::clean_names() %>%
  select(c("ano", "nome_curto_indicador",
          "regiao_planejamento", "municipio", "valor"))

head(dados_ens_fund, 5)
```

```
## # A tibble: 5 x 5
##       ano nome_curto_indicador      regiao_planejamento municipio  valor
##   <dbl> <chr>                      <chr>                <chr>     <dbl>
## 1  2008 Abandono no ensino fundamental Cariri              Abaiara    5
## 2  2015 Alunos avaliados - SPAECE 5º ano - Re~ Cariri            Altaneira N
## 3  2013 Estabelecimentos com ensino fundament~ Litoral Oeste / Vale d~ Amontada N
## 4  2015 Alunos avaliados - SPAECE 5º ano - Re~ Cariri            Antonina do ~ N
## 5  2015 Alunos avaliados - SPAECE 5º ano - Re~ Cariri            Araripe     N
```

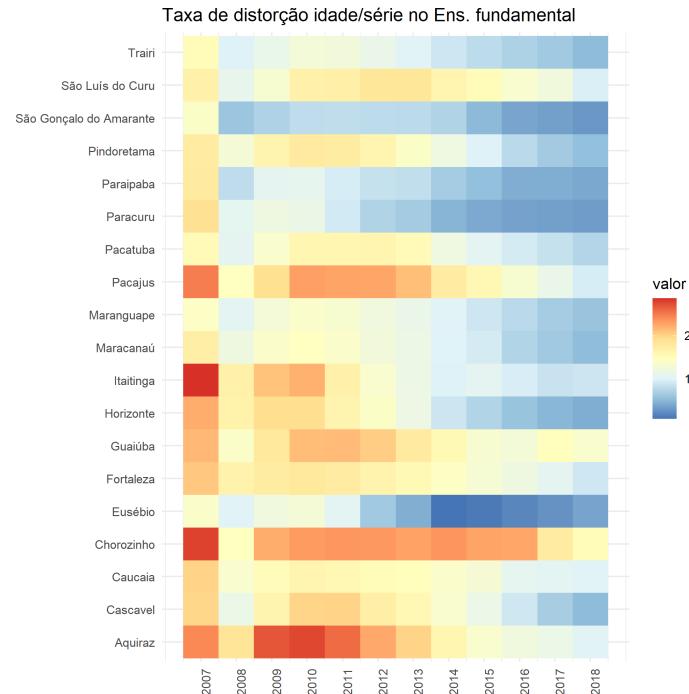
Preparação dos dados

```
dados_ens_fund <- read_delim("data/data_ens_fundamental.csv",
                               delim = ";") %>%
  janitor::clean_names() %>%
  select(c("ano", "nome_curto_indicador",
          "regiao_planejamento", "municipio", "valor")) %>%
  filter(regiao_planejamento == "Grande Fortaleza",
         nome_curto_indicador == "Taxa de distorção idade/série no ensino fundamen
```

Visualização dos dados

```
library(ggplot2)

ggplot(data = dados_ens_fund,
       aes(x = ano, y = municipio, fill =
geom_tile() +
  labs(title = "Taxa de distorção idade/série no Ens. fundamental",
       y = NULL, x = NULL) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90)),
       scale_fill_distiller(palette = "RdYlBu"),
       scale_x_continuous(breaks = c(2007:2018),
                          labels = c(2007:2018)))
```



Para aprender mais

Cursos e tutoriais

- Introduction to R
- R bootcamp
- Pacote `swirl`

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

Livros

- Livro da [Curso-R](#)
- An introduction to statistical learning
- R for Data Science - Hadley Wickham & Garrett Grolemund

Obrigada!

Slides criados com o pacote `xaringan`

Agradecimentos a [Beatriz Milz](#) pelo [curso de xaringan](#)

Gifs do tidyverse criados por [Garrick Aden-Buie](#)

