

MA'LADIES

WORKSHOP DE R

Isabella Bicalho Frazeto - bicalhoisabella@gmail.com



JÁ TRABALHOU COM O R OU OUTRA LINGUAGEM PARECIDA?

PROGRAMA

- Introdução ao R
- Introdução ao dplyr
- Boas práticas na manipulação e organização de dados
- Importando dados para o R
- Utilização de funções para extrair informação de tabelas no R
- Exportando dados do R
- Elaboração de gráficos usando o ggplot2.

AVISOS GERAIS

**Não é um mini-curso de estatística
Não é um mini-curso de
programação**

O QUE É O R?

R É UMA LINGUAGEM E UM
AMBIENTE PARA
COMPUTAÇÃO ESTATÍSTICA
E GRÁFICOS

BASEADO NA LINGUAGEM S

OPEN SOURCE
é de graça, gente

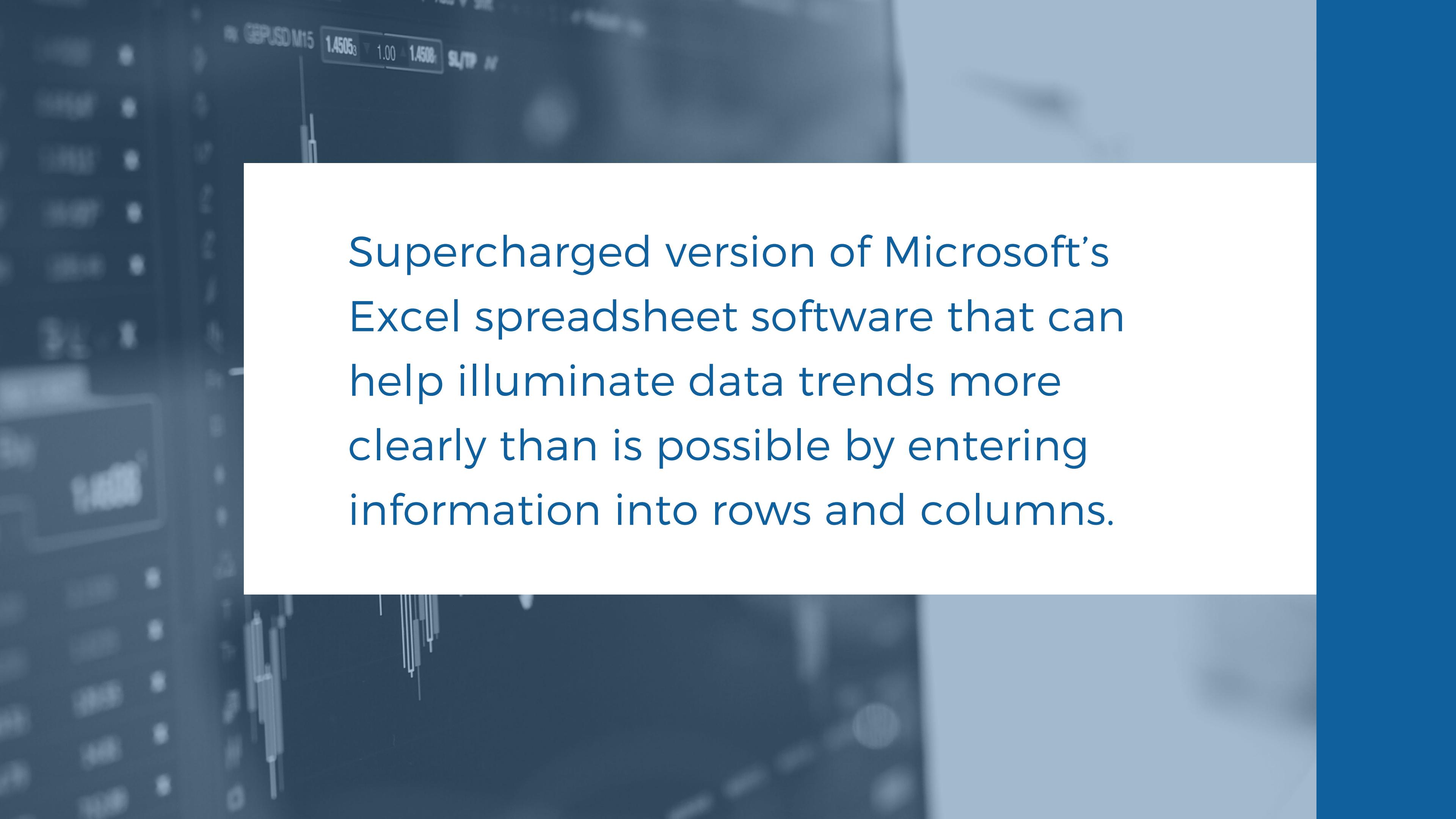
O QUE É O R?

EXTENSÍVEL A PARTIR DE
PACOTES

flexibilidade na hora de quais ferramentas
usar

GRÁFICOS BEM
DESENHADOS PRONTOS
PARA PUBLICAÇÃO

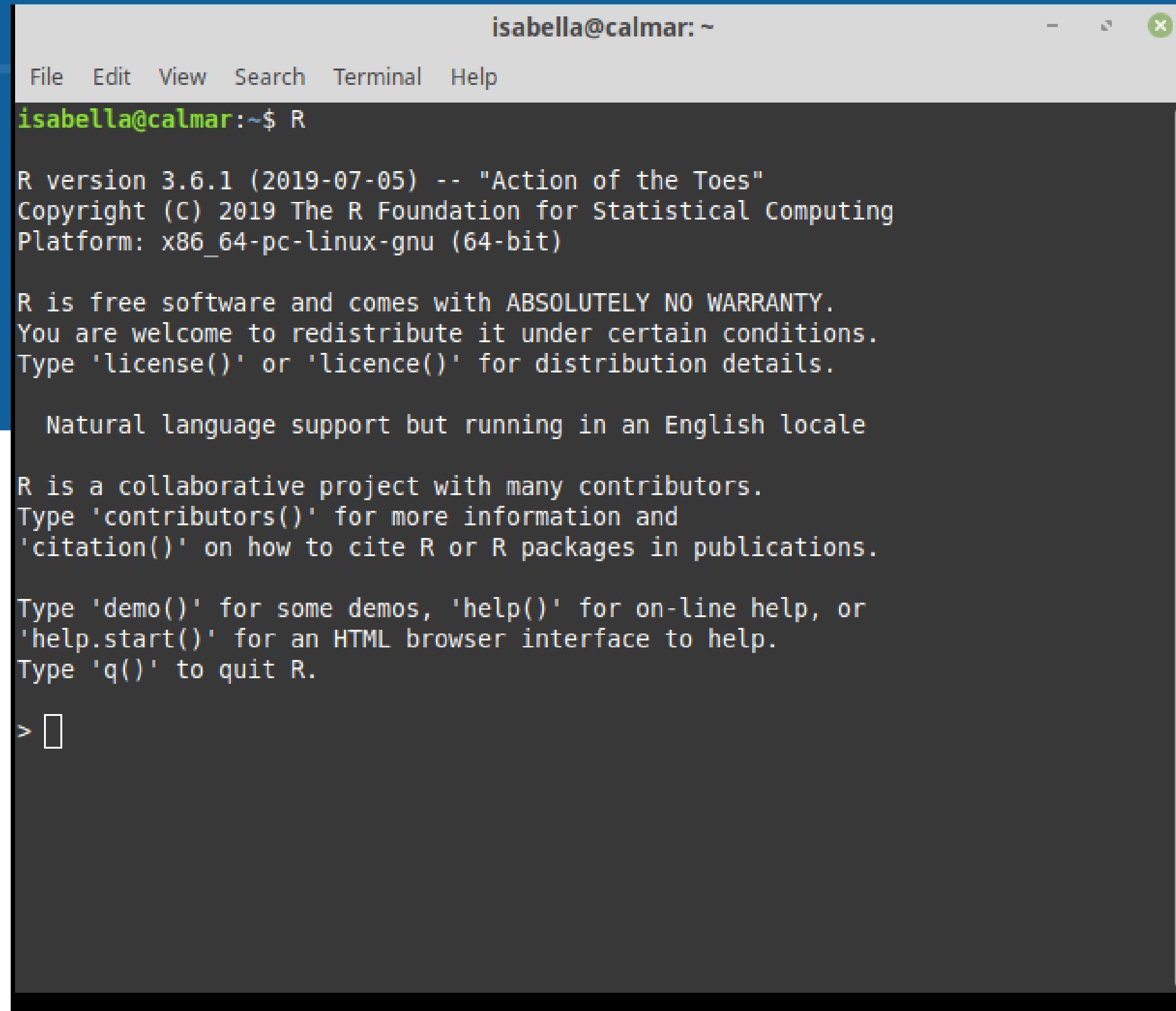
defaults costumam ser bons mas eles
fornecem total liberdade para o usuário



Supercharged version of Microsoft's Excel spreadsheet software that can help illuminate data trends more clearly than is possible by entering information into rows and columns.

COMO USÁ-LO?

PROMPT/TERMINAL



A screenshot of a terminal window titled "isabella@calmar: ~". The window has a standard Linux-style title bar with icons for minimize, maximize, and close. The menu bar includes "File", "Edit", "View", "Search", "Terminal", and "Help". The terminal itself shows the R command-line interface starting up. It displays the R version (3.6.1), copyright information, and platform details. It then provides standard startup messages about warranty, redistribution, and natural language support. It also mentions it's a collaborative project and provides information on how to cite R or packages. Finally, it gives instructions for demos, help, and quitting.

```
isabella@calmar: ~
File Edit View Search Terminal Help
isabella@calmar:~$ R

R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

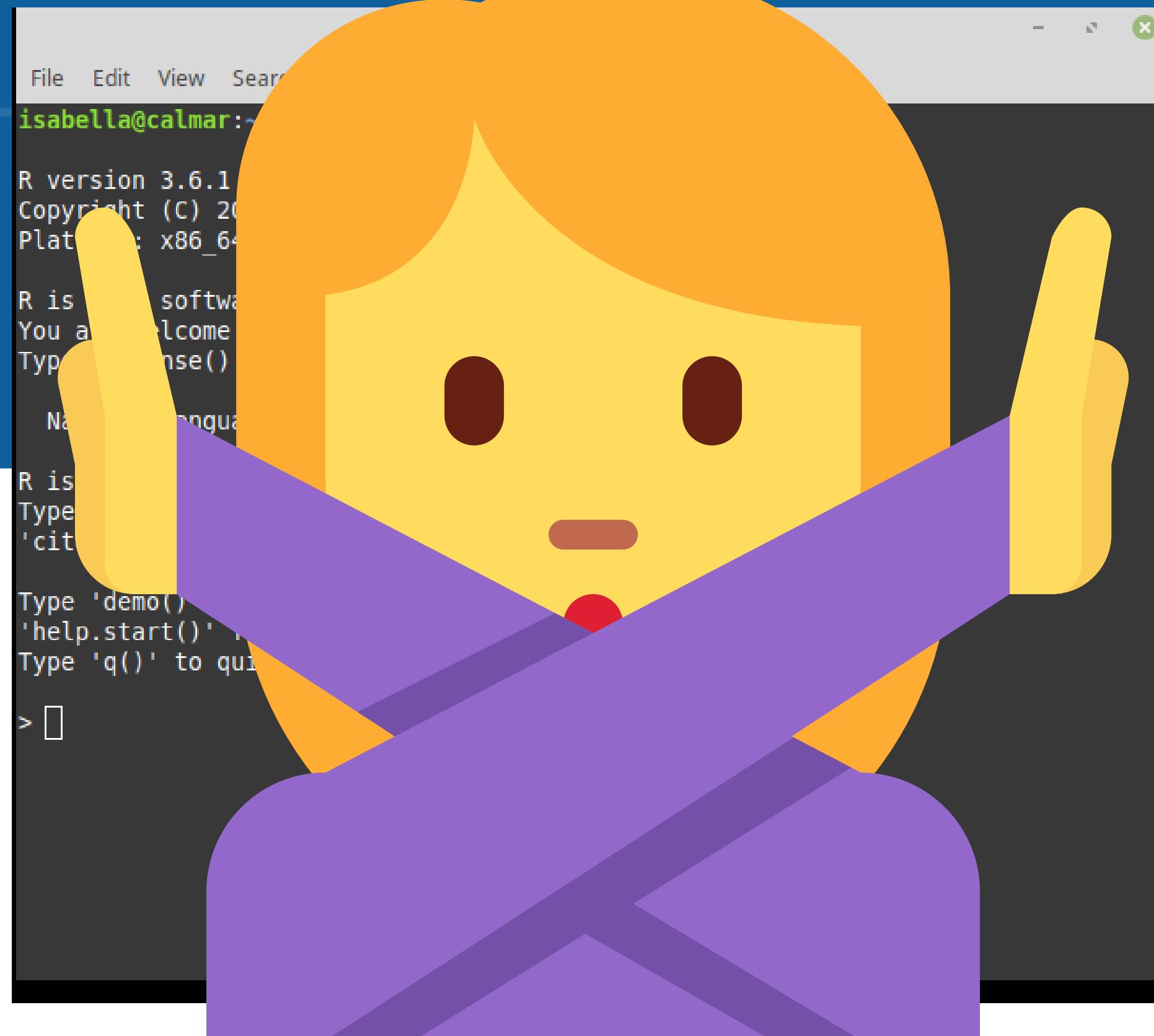
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

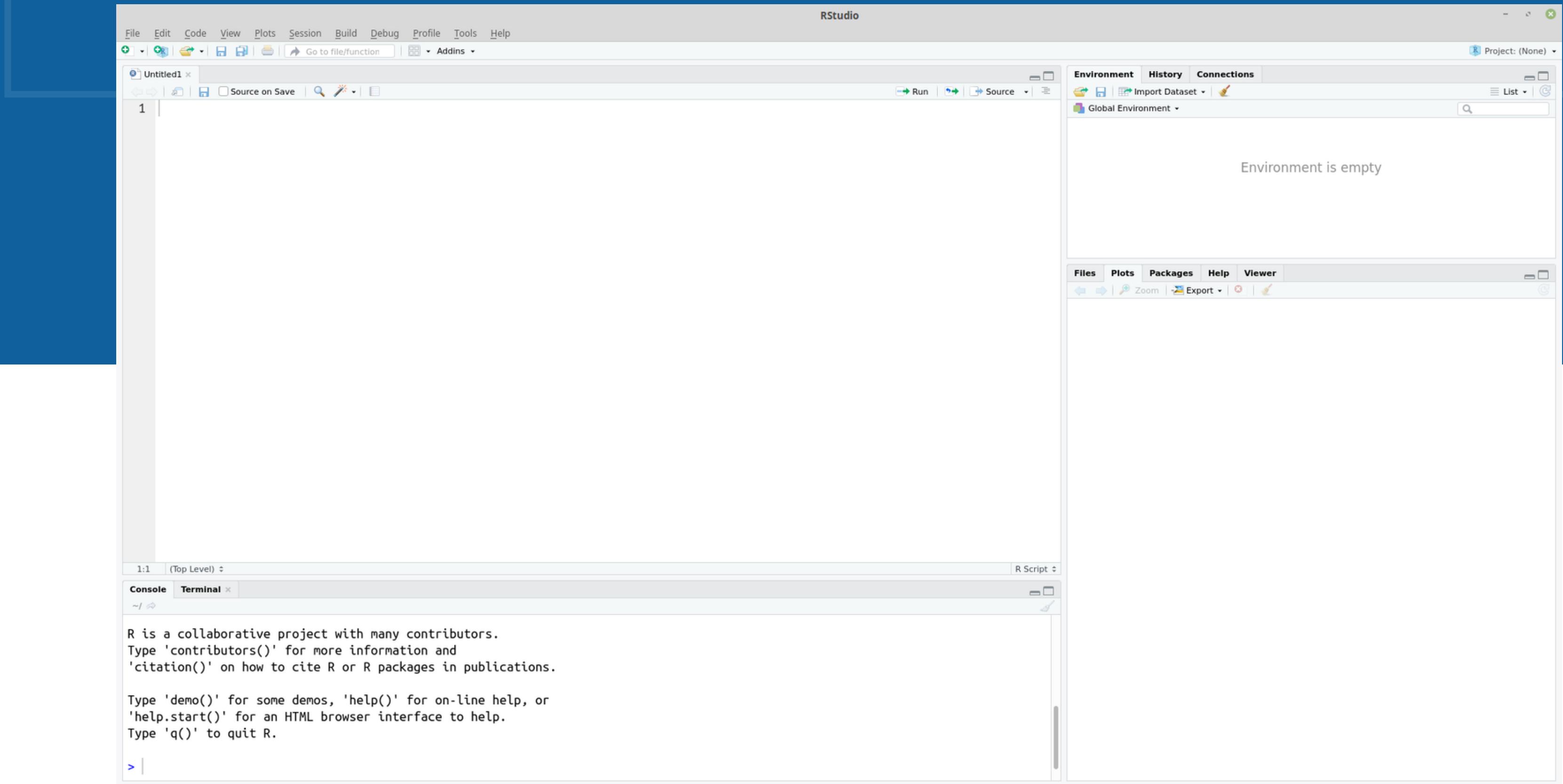
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> □
```

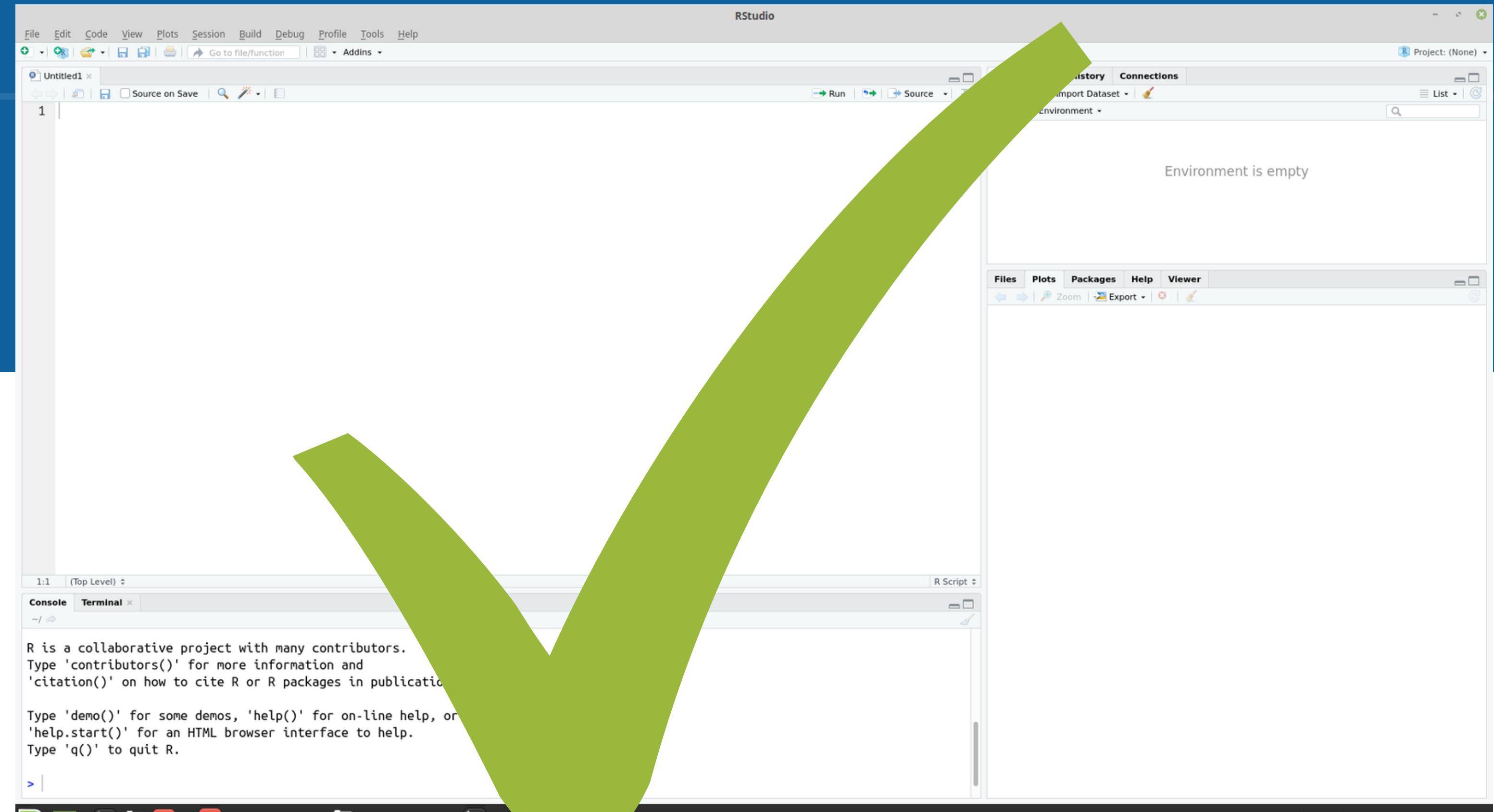
PROMPT/TERMINAL



RSTUDIO



RSTUDIO



RSTUDIO



AVON



RSTUDIO



The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying philosophy and common APIs.

[Project Site Link >](#)



ggplot 2 is an enhanced data visualization package for R. Create stunning multi-layered graphics with ease.

[Project Site Link >](#)



dplyr is the next iteration of plyr, focussing on only data frames. dplyr is faster and has a more consistent API.

[Project GitHub Link >](#)



tidyr makes it easy to “tidy” your data. Tidy data is data that’s easy to work with: it’s easy to munge (with dplyr), visualise (with ggplot2 or ggvis) and model (with R’s hundreds of modelling packages).

[Project Paper Link >](#)



purrr enhances R’s functional programming (FP) toolkit by providing a complete and consistent set of tools for working with functions and vectors.

[Project Site Link >](#)

RSTUDIO



The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying philosophy and common APIs.

[Project Site Link >](#)



ggplot 2 is an enhanced data visualization package for R. Create stunning multi-layered graphics with ease.

[Project Site Link >](#)



dplyr is the next iteration of plyr, focussing on only data frames. dplyr is faster and has a more consistent API.

[Project GitHub Link >](#)



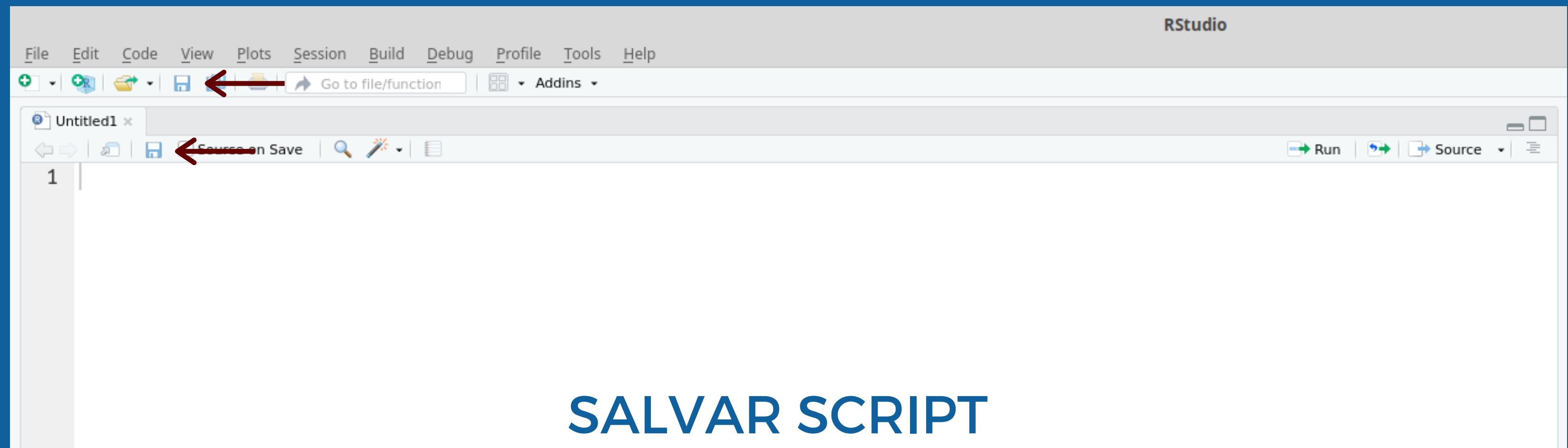
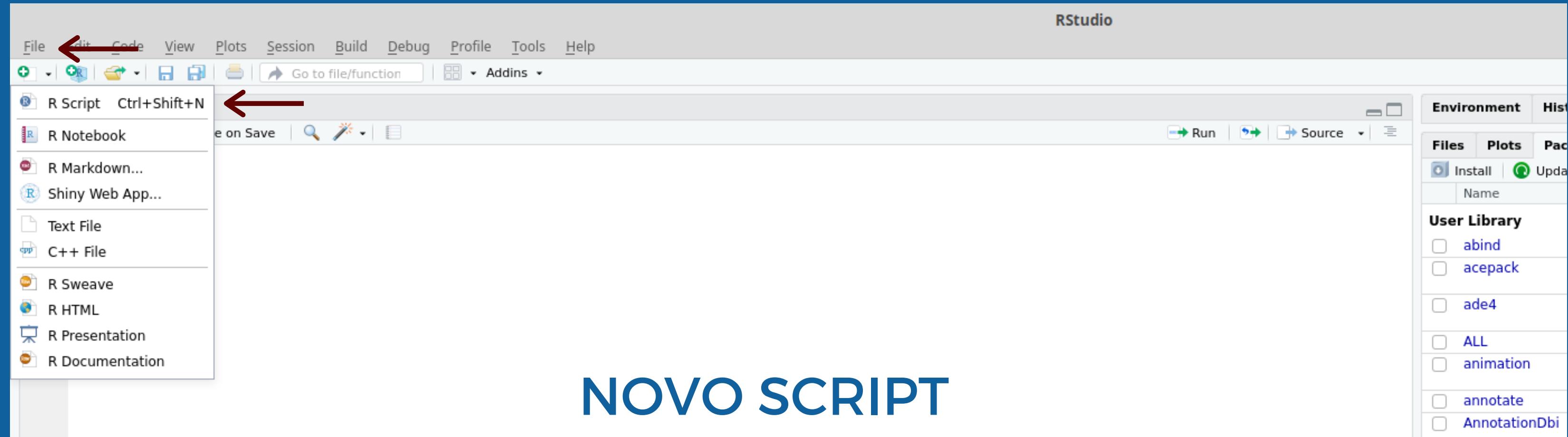
tidyr makes it easy to “tidy” your data. Tidy data is data that’s easy to work with: it’s easy to munge (with dplyr), visualise (with ggplot2 or ggvis) and model (with R’s hundreds of modelling packages).

[Project Paper Link >](#)



purrr enhances R’s functional programming (FP) toolkit by providing a complete and consistent set of tools for working with functions and vectors.

[Project Site Link >](#)





NÃO SALVA AUTOMATICAMENTE
CTRL + S = ATALHO PARA SALVAR



Variável

SÃO USADAS PARA
PARA ARMAZENAR OU
REPRESENTAR
NA MEMÓRIA UM
VALOR OU EXPRESSÃO

NO R

USAMOS O "<-"
PARA
REPRESENTAR A
DESIGNAÇÃO DE UM
VALOR PARA
IDENTIFICAÇÃO

IDENTIFICADOR <- VALOR

O RÉ SENSÍVEL AS DIFERENÇAS
ENTRE MAIÚSCULAS E MINÚSCULAS



x é diferente de x



OUTROS SÍMBOLOS IMPORTANTES

+ -

mais/menos

* /

multiplicação/divisão

== !=

igualdade/diferença

> >= <= <

Maior que/ Menor que

!

negação

#

comentário

OUTROS SÍMBOLOS IMPORTANTES

& &&

e

| ||

ou

identificador\$componente

extrair um componente

[índice] [[índice]]

acessar um índice

:

sequência

<- -> =

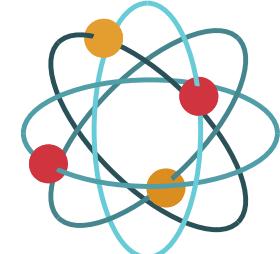
atribuição

Tipos de dados

aka valores

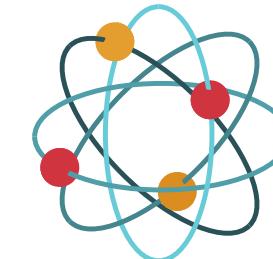
Lógicos/Booleanos

TRUE, FALSE



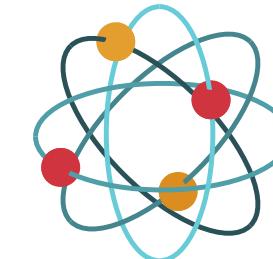
Númericos/Double

12.3, 5, 999



Caracteres/String

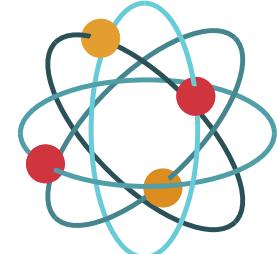
'a' , "good", "TRUE",
'23.4'



Tipos de dados

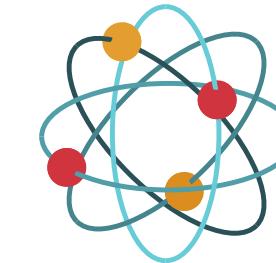
Lógicos/Booleanos

TRUE, FALSE



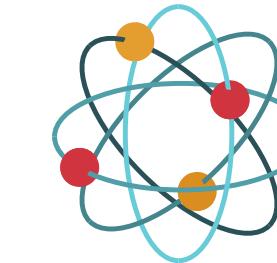
Númericos/Double

12.3, 5, 999



Caracteres/String

→ 'a' , "good", "TRUE", ←
'23.4'



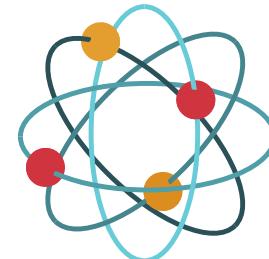
Tipo de dados

Vetores

```
('red','green',"yellow")
```

```
(1, 2, 3)
```

```
c(1:8)
```



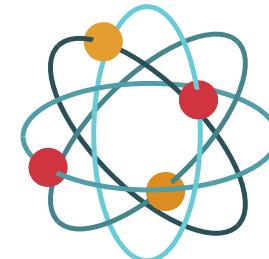
Listas

```
list(c(2,5,3),21.3,sin)
```

```
=/=
```

Dataframes

	gender	height	weight	Age
1	Male	152.0	81	42
2	Male	171.5	93	38
3	Female	165.0	78	26



```
=/=
```

TESTE

X <- 3
Y <- 5

QUE TIPO DE DADOS É X?

QUAL O VALOR DE Y-X?

X < Y

X <- 3
Y <- 5

QUE TIPO DE DADOS É X?
númerico

QUAL O VALOR DE Y-X?

2

X < Y
TRUE

W <- 10 <= 10
K <- FALSE

QUE TIPO DE DADOS É W?

QUAL O VALOR DE W==K?

E DE W != K?

W <- 10 <= 10
K <- False

QUE TIPO DE DADOS É W?
LÓGICO

QUAL O VALOR DE W==K?
FALSE

E DE W != K?
TRUE

J <- c(1:10)

QUE TIPO DE DADOS É J?
QUAL O VALOR DE J[2]?

ESCREVA TODOS OS VALORES
DE J

J É ATÔMICO?

J <- c(1:10)

QUE TIPO DE DADOS É J?
QUAL O VALOR DE J[2]?
VETOR, 2

ESCREVA TODOS OS VALORES
DE J

1, 2, 3, 4, 5, 6, 7, 8, 9, 10

J É ATÔMICO?
SIM

```
T <- list(c(2,5,3),21.3, "k",  
         37)
```

QUE TIPO DE DADOS É T?
QUAL O VALOR DE T[3]?
QUAL SEU TIPO DE OBJETO?

T[[2]] < T[[4]]? E QUAL O VALOR
DE T[[1]][1]?

T É ATÔMICO?

```
T <- list(c(2,5,3),21.3, "k",  
        37)
```

QUE TIPO DE DADOS É T?
QUAL O VALOR DE T[3]?
QUAL SEU TIPO DE OBJETO?
LISTA, 21.3, NÚMERO

T[2] < T[4]? E QUAL O VALOR DE
T[[1]][1]?
TRUE, 2

T É ATÔMICO?
NÃO, ELE É RECURSIVO

**VAMOS FAZER NOSSO PRIMEIRO
SCRIPT**

ORGANIZAÇÃO DE DADOS

Organização dos dados

EVITE CARACTÉRES ESPECIAIS E
ESPAÇOS NO NOME DE
COLUNAS

SALVE OS DADOS EM .CSV, .TSV
OU .TXT

NÃO PREENCHA VALORES QUE
NÃO FORAM REGISTRADOS COM
ZERO

Organização dos dados

NÃO JUNTE COLUNAS OU
LINHAS

NÃO COLOQUE COMENTÁRIOS
OU UNIDADES MÉTRICAS NA
TABELA

NÃO TENHA MAIS DE UMA
INFORMAÇÃO POR CÉLULA

Organização dos dados

VERIFIQUE SE OS NOMES DE
FATORES ESTÃO IDENTICOS
ENTRE SI

NÃO TENHA MAIS DE UMA ABA
POR PLANILHA

NÃO TENHA CORES,
SUBLINHADOS E AFINS

SHOW DE HORRORES

	A	B	C	D	E	F	G
1		SCREEN 1					
2					MULTIPLE SHIFT ASSIGNMENT MO		
3							
4		Preference Total:			0		
5							
6			Mon	Tue	Wed	Thu	
7	DAYS REQD:		2	1	1	3	
8			-2	-1	-1	-3	
9							
10							
11		Mon	Tue	Wed	Thu		
12	EVES REQD:	1	1	0	0		
13		-1	-1	0	0		
14							
15							
16	Mon	Tue	Wed	Thu			
17	NITES REQD:	1	0	0	0		
18		-1	0	0	0		

	A	B	C	D	E	F
1	<i>OTBT gmtpm& mnpm</i>					
2		OTBT Rzpsmr	OTBT N/B	Tsct OTBT NB		
3		11434	64037	200000		
4	<i>Osrkzt vgImOzs mnpm</i>					
5		Lsst 12 Ognths sslzs snd rzpsmr zxpgsmrz pzrmgd cslcmstmgn				
6		Tgtsl bssz	svzrsgz rmntmOz	BzTgrz 12 O	R/Onth -1	R/Onth -2
7	yslzs bmmlt:	1412	11.97	1400	0	0
8	Tgtsl rzpsmr bmmlt:	306	4.36	0	2	23
9	ATT bmmlt mnpm	272	4.01	0	1	15
10	Bzndgr 2 bmmlt mnpm	0	#DIV/0!	0	0	0
11	Bzndgr 3 bmmlt mnpm	34	7.18	0	1	8
12	<i>RzAsmR mNAUT</i>					
13		Rzpsmr mssgz			Rzpsmr pgpmstmgn	
14		Rzpsmrzd DOs		Rzpsmrzd Wsrr	NgrOsl Rzpsmr	LgcsL Rzpsmr
15	306	2		40	264	0
16	<i>Argdmct mnTgrOstmgm</i>				% cmrrznt mnstsllzd bssz sTtzr Tmtmrz ssl	
17	Argdmct NsOz	ytylz NsOz		Rzvmzw Dstz	Ognths sctmvz	svzr Tmtmrz rmntmOz
18		DzhmOmdm1sB6065903CT		25/06/1998	76.8	27
19	<i>Asrt NmObzr mnTgrOstmgm</i>				Ognth/dsy/yzsr	
20	sDy Asrt NmObz	yzt-mp Dstz		ytsndsrD cgst	Ognthly mssgz	12 Ognth's mssgz
21		43588136-00	1/02/1992	675.92	9.7	116.4
22				Ognth/dsy/yzsr		
23						

	A	B	C
1	group	details	
2		1 1M, 1F	
3		2 3F	
4		3 2M, 2F	
5		4 1M	
6		5 1M, 3F	
7		6 2M, 1F	
8		7 1F	

	A	B	C
1	group	female	male
2		1	1
3		2	3
4		3	2
5		4	0
6		5	3
7		6	1
8		7	0

	A	B	C
1	Year	Month	Rate
2	2006	Jan	4.7
3	2006	Feb	4.8
4	2006	Mar	4.7
5	2006	Apr	4.7
6	2006	May	4.6
7	2006	Jun	4.6
8	2006	Jul	4.7
9	2006	Aug	4.7
10	2006	Sep	4.5
11	2006	Oct	4.4
12	2006	Nov	4.5
13	2006	Dec	4.4
14	2007	Jan	4.6
15	2007	Feb	4.5
16	2007	Mar	4.4

IMPORTANDO DADOS DE TABELAS PARA O R

**criar uma pasta como o nome em
documentos "minicurso_r"**

getwd()

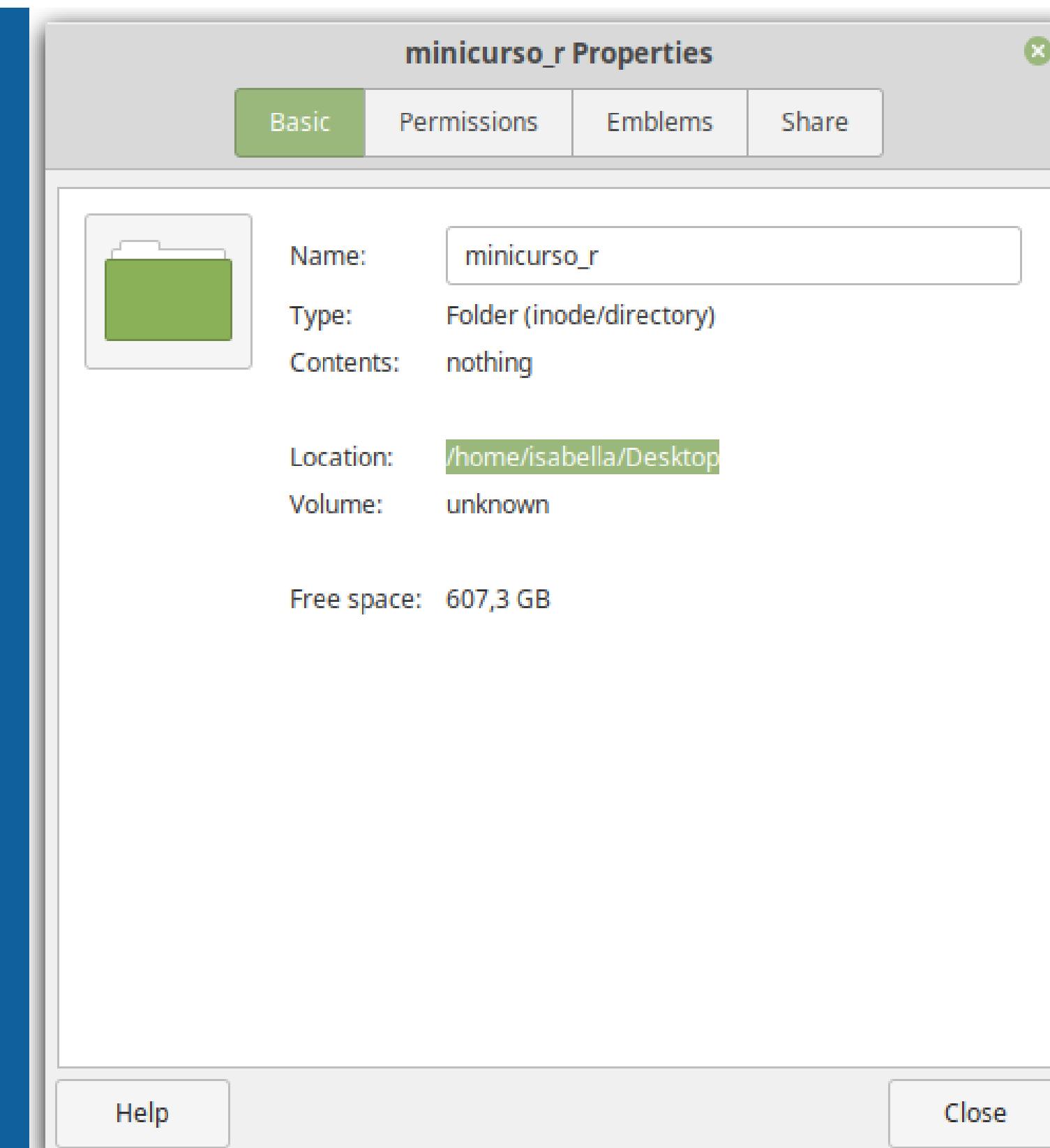
**criar uma pasta como o nome em
documentos "minicurso_r"**

getwd()

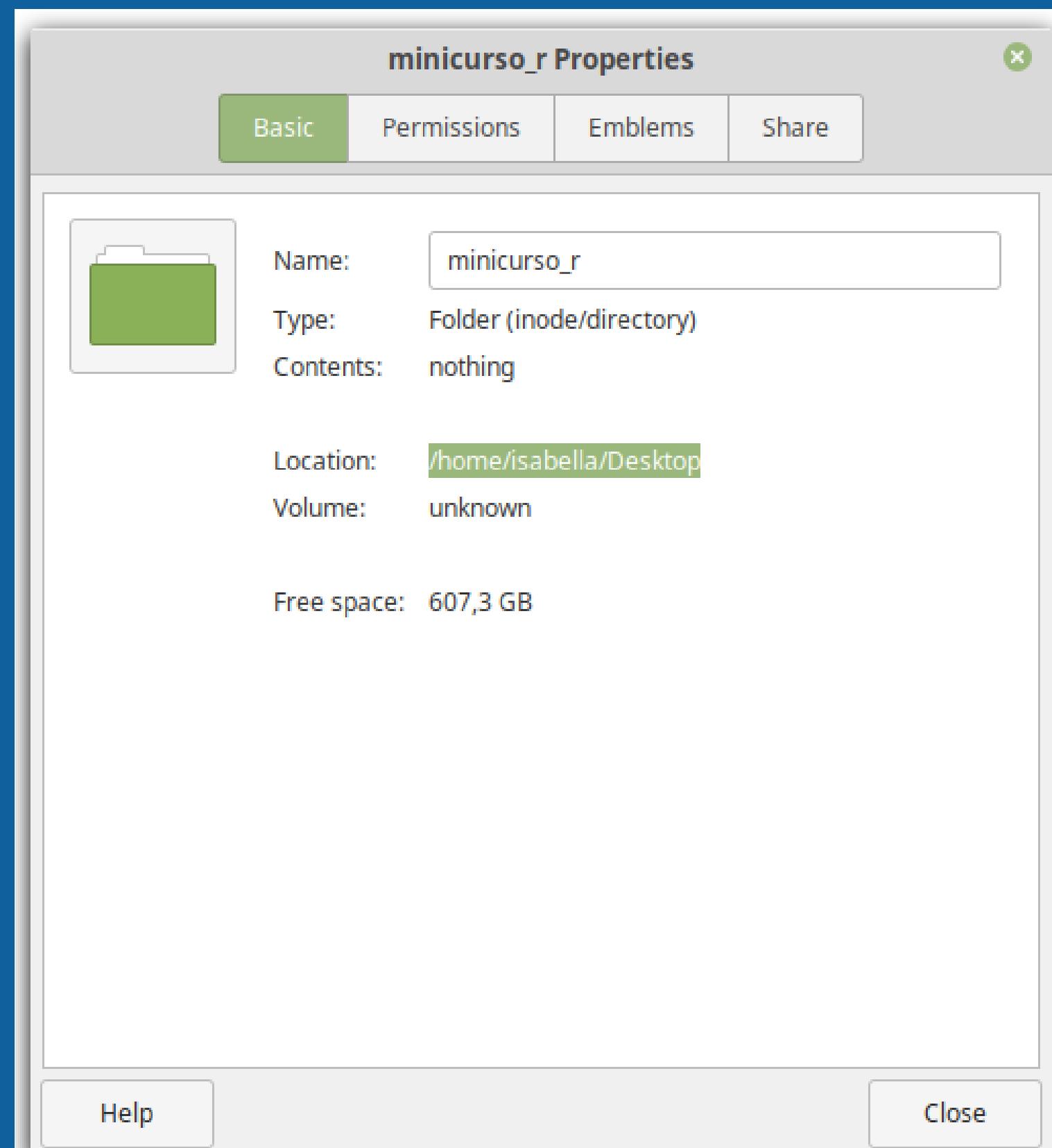
A localização da seu arquivo provavelmente
não é o mesmo do R

Vamos mudar de diretório

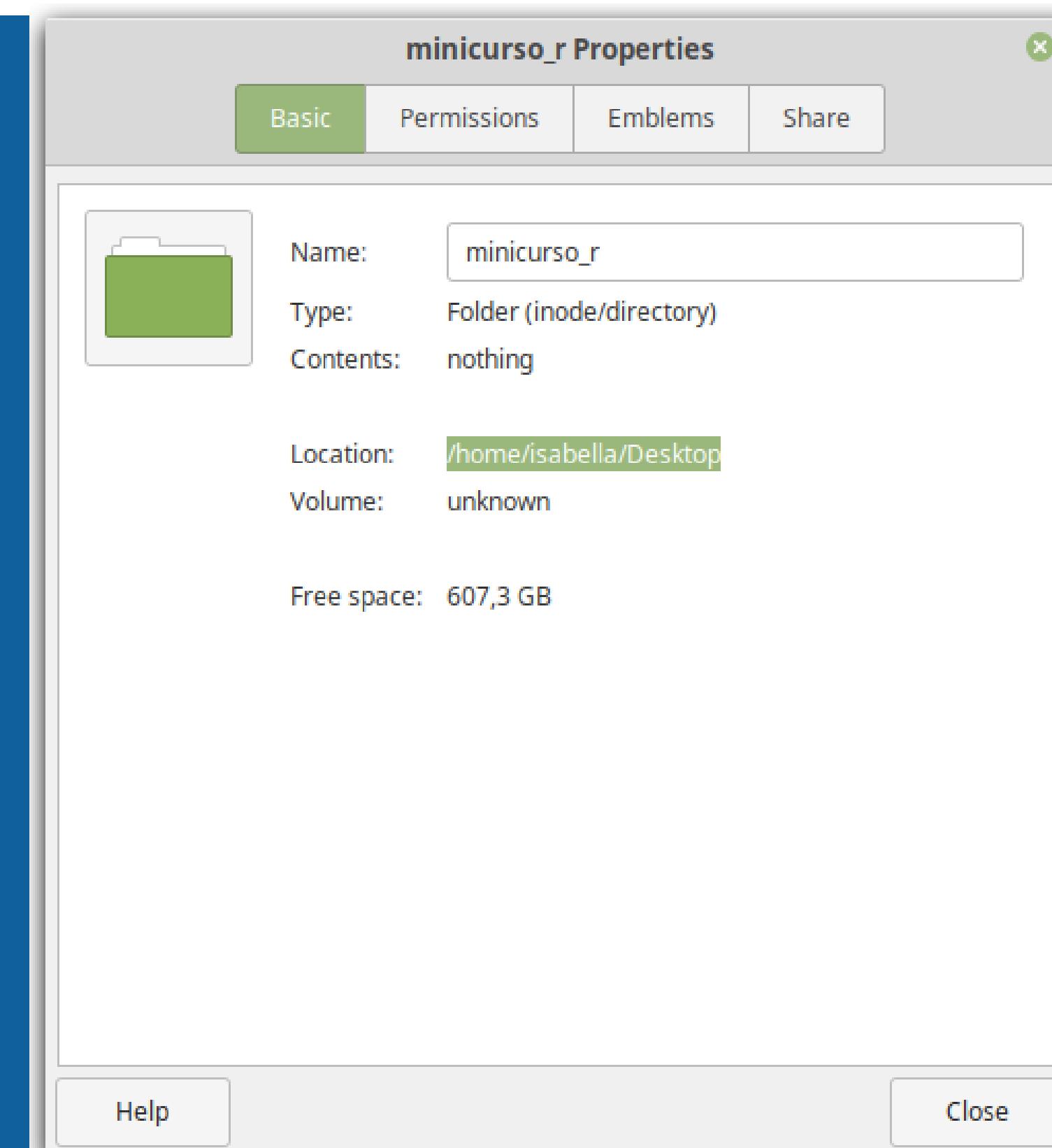
como conseguir a localização da sua pasta



```
setwd("home/isabella/Desktop/minicurso_r")
```



getwd() agora deve retornar igual



```
dados <- read.csv("iris.csv")  
View(dados)
```

MANIPULAÇÃO DE DADOS



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



HADLEY WICKMAN

Vencedor do "Premio Nobel da estatística " de 2019 pela influência do seu trabalho em estatística computacional, visualização, gráficos, análise de dados e por fazer estatística computacional acessível para uma audiência mais ampla

Vantagens

DPLYR

- Sintaxe é mais simples comparado a base do R
- Abstrai como de fato seus dados são armazenados
- A base do R trabalha com vetores enquanto o dplyr trabalha com dataframes
- Todas as funções tem uma sintaxe semelhante, isto é, ele é consistente
- Rápido
- Parecido com o SQL

MAS O QUE É ISSO?

DPLYR É UMA GRAMÁTICA DE MANIPULAÇÃO DE DADOS

DPLYR É UMA GRAMÁTICA DE MANIPULAÇÃO DE DADOS

ELA USA VERBOS PARA FAZER A MANIPULAÇÃO
DE DADOS

VERBOS

- **MUTATE**
adiciona novas variáveis em função de outras
- **SELECT**
seleciona variáveis baseada no nome
- **FILTER**
seleciona variáveis baseada em valores
- **SUMMARIZE**
resume variáveis em valores únicos
- **ARRANGE**
ordena as linhas da tabela

VERBOS

- **MUTATE** ex: % de um valor
adiciona novas variáveis em função de outras
- **SELECT**
seleciona variáveis baseada no nome
- **FILTER**
seleciona variáveis baseada em valores
- **SUMMARIZE** ex: medidas de tendência central
resume variáveis em valores únicos
- **ARRANGE**
ordena as linhas da tabela

PIPE

O tidyverse tem mais um símbolo "%>%" o pipe. Ele permite que agrupar várias operações que seriam realizadas separadamente em um única operação. Ajuda a salvar memória.



PIPE

O tidyverse tem mais um símbolo "%>%" o pipe. Ele permite que a gente junte várias operações que seriam realizadas separadamente em um única operação. Ajuda a salvar memória.

EXEMPLO

```
the_data <- read.csv('/path/to/data/file.csv') %>%
  subset(variable_a > x) %>%
  transform(variable_c = variable_a/variable_b) %>%
  head(100)

iris %>%
  subset(Sepal.Length > mean(Sepal.Length)) %$%
  cor(Sepal.Length, Sepal.Width)
```

TESTE

**QUAL A FUNÇÃO
ADEQUADA
PARA CALCULAR
O DESVIO
PADRÃO?**

SUMMARISE

ARRANGE

MUTATE

**QUAL A FUNÇÃO
ADEQUADA
PARA CALCULAR
O DESVIO
PADRÃO?**

SUMMARISE



ARRANGE

MUTATE

**QUAL A FUNÇÃO
ADEQUADA
PARA CALCULAR
O VOLUME DO
CORPO?**

SUMMARISE

ARRANGE

MUTATE

**QUAL A FUNÇÃO
ADEQUADA
PARA CALCULAR
O VOLUME DO
CORPO?**

SUMMARISE

ARRANGE

MUTATE



**COMO EU
FILTRARIA
APENAS UM
GRUPO DE
ESPÉCIES QUE
EU QUERO?**

FILTER

ARRANGE

SELECT

**COMO EU
FILTRARIA
APENAS UM
GRUPO DE
ESPÉCIES QUE
EU QUERO?**

FILTER

ARRANGE

SELECT



**COMO EU
FILTRARIA
APENAS
INDIVIDUOS
MAIORES QUE
10CM?**

FILTER

ARRANGE

SELECT

**COMO EU
FILTRARIA
APENAS
INDIVIDUOS
MAIORES QUE
10CM?**

FILTER



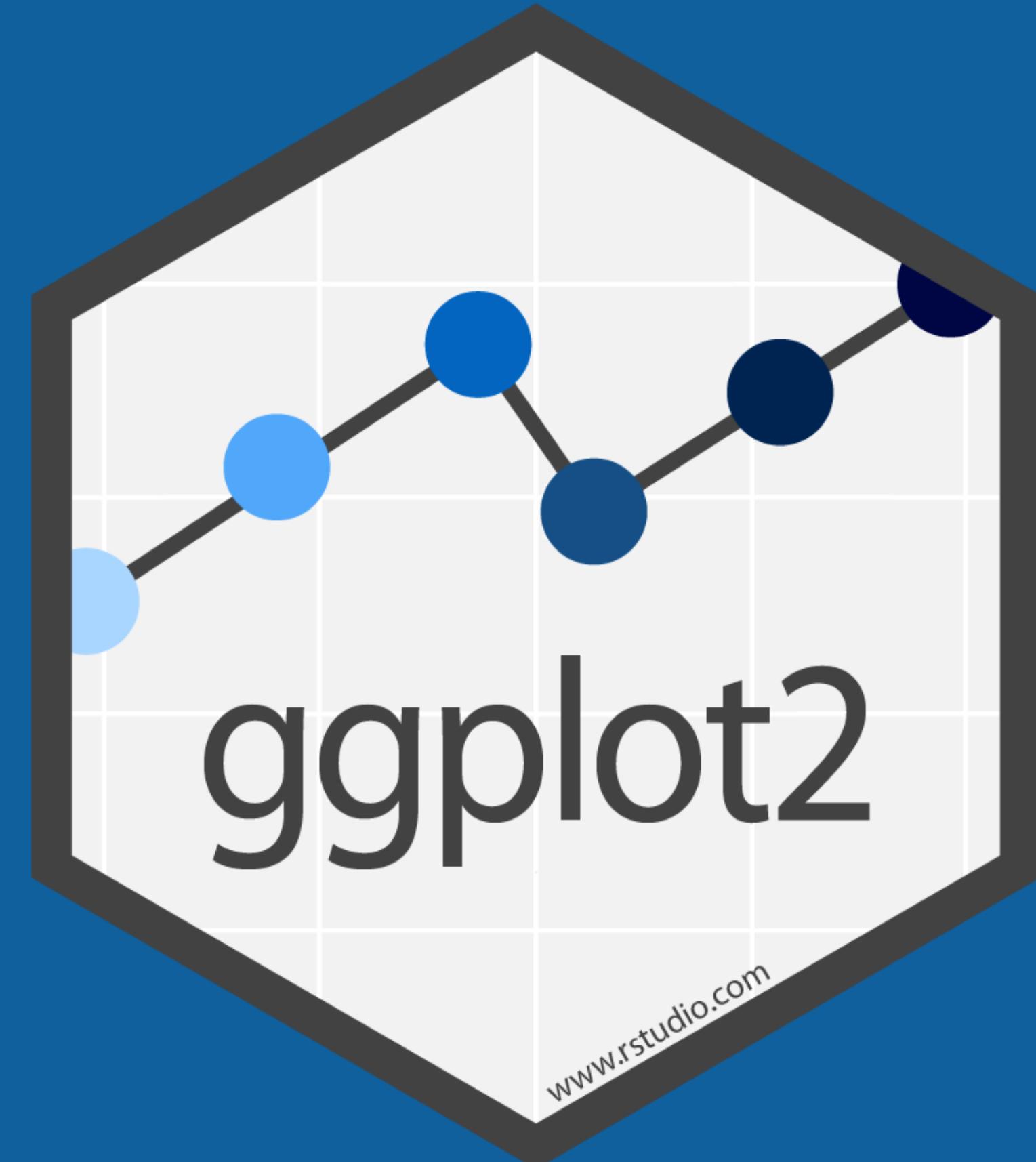
ARRANGE

SELECT

EXPORTANDO DADOS

```
WRITE.CSV(MEDIAS_ESPECIES, "MEDIAS_ESPECIES.CSV")
```

Visualização de dados



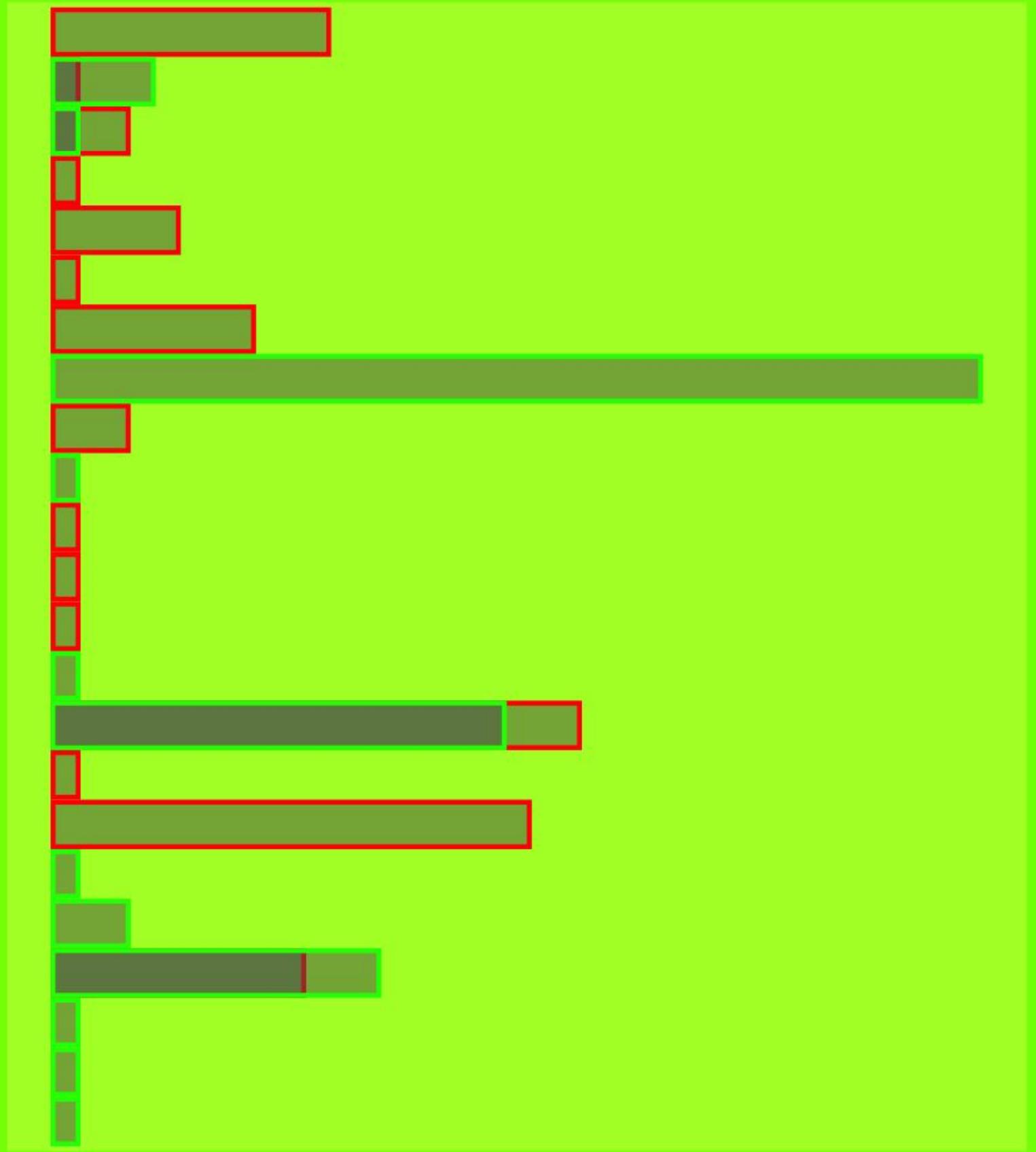
```
install.packages("ggplot2")  
library(ggplot)
```

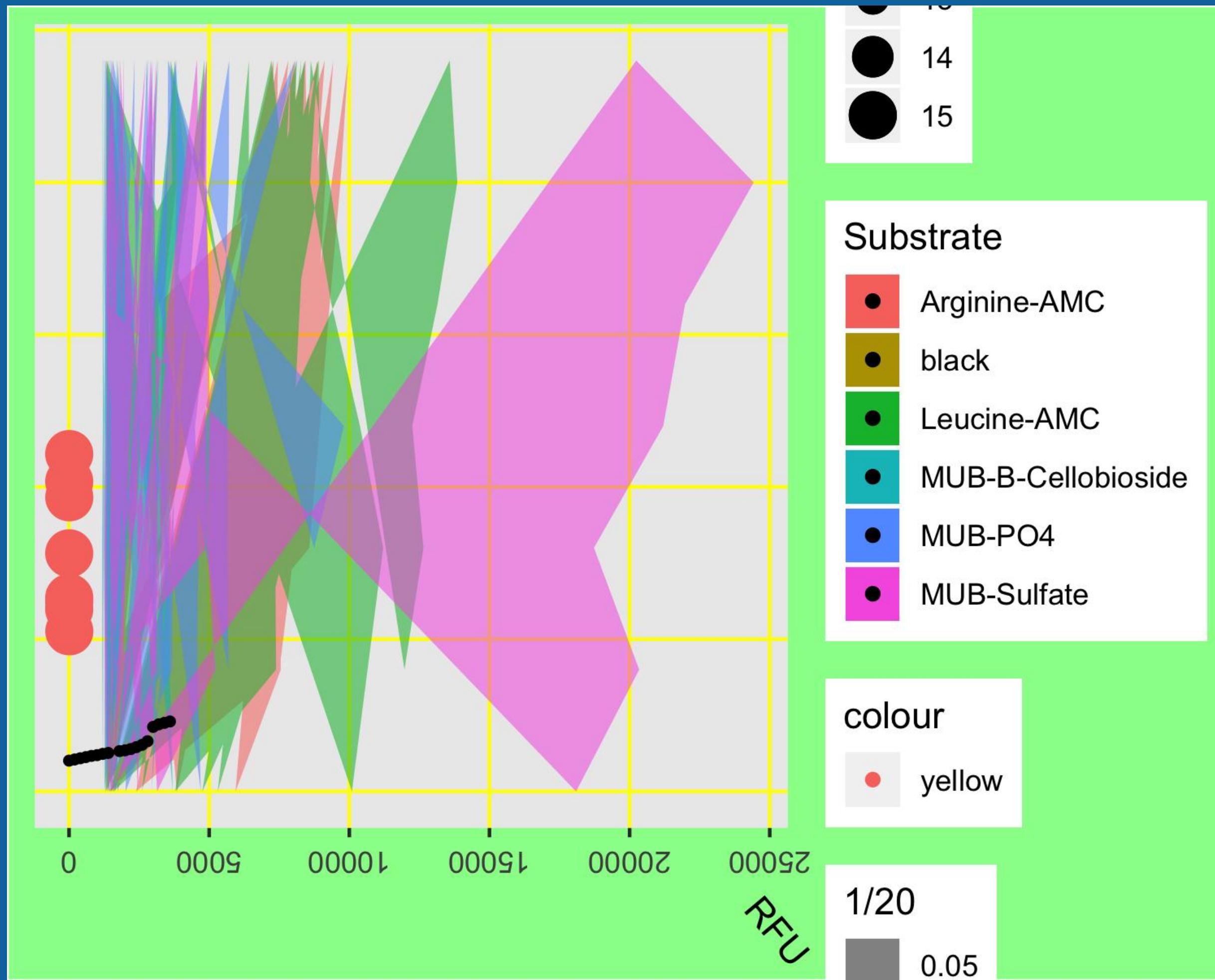
GGPLOT É UMA GRAMÁTICA DE GRÁFICOS

SHOW DE HORRORES

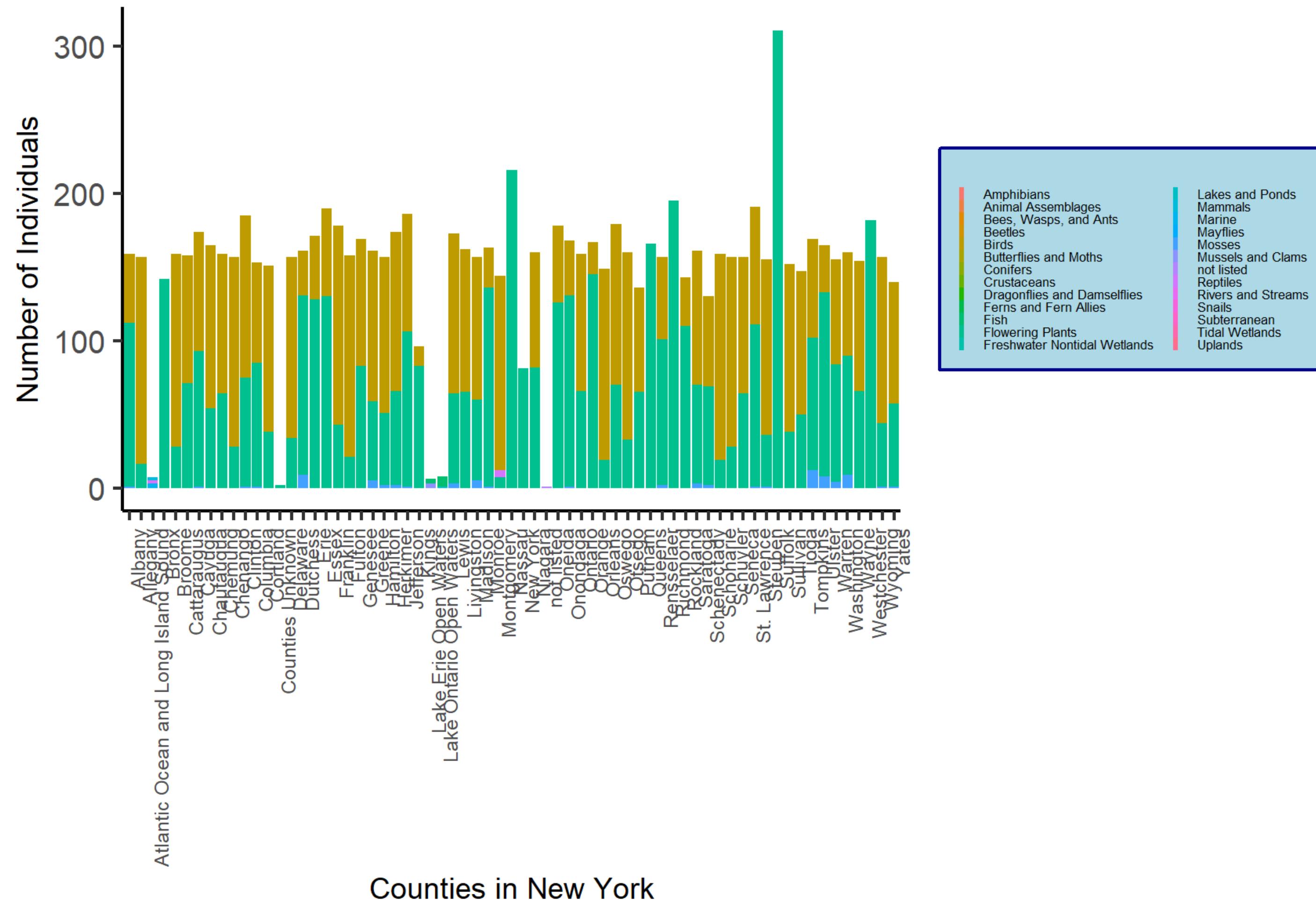
them star wars bois

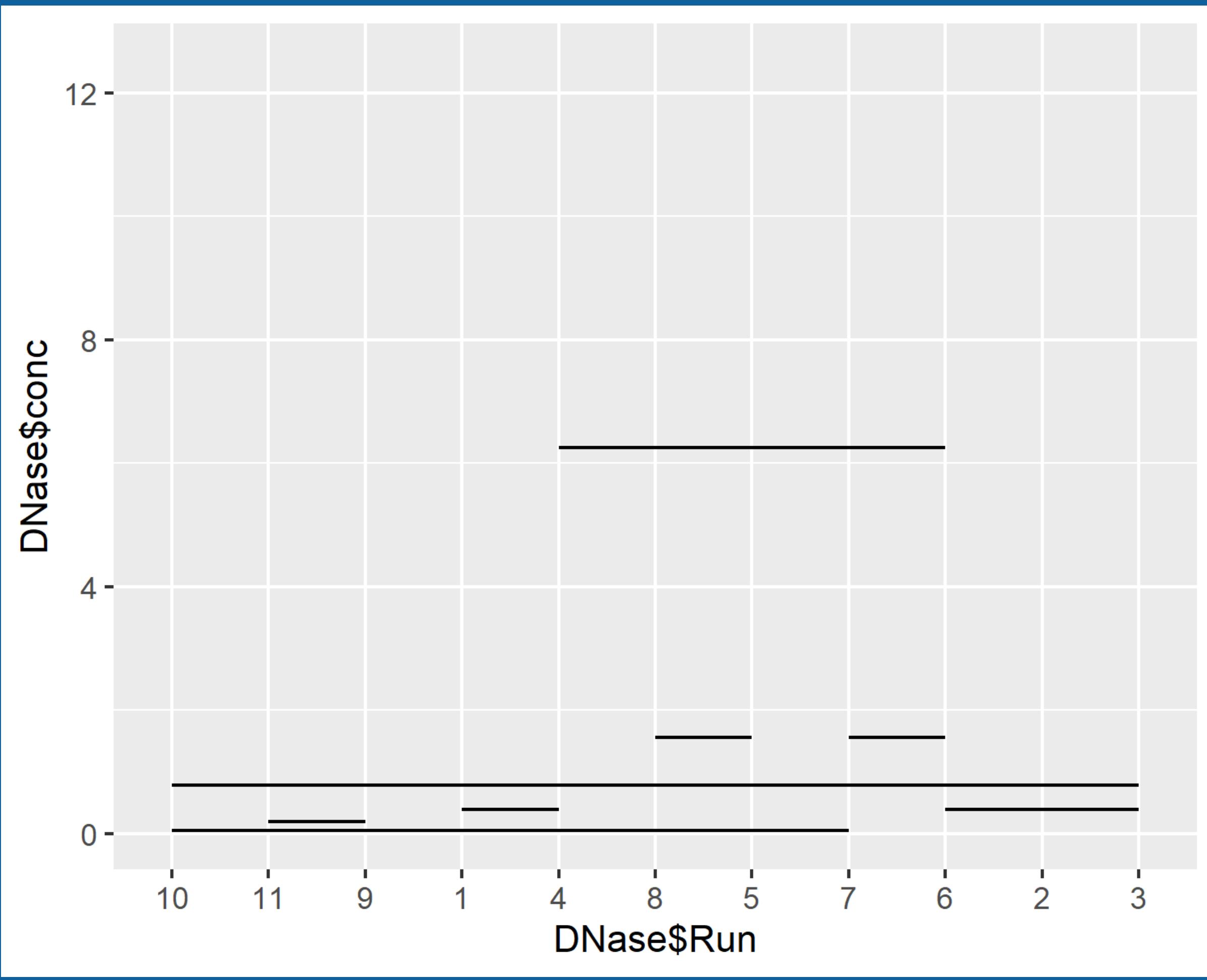
yellow
white
unknown
red, blue
red
pink
orange
none
hazel
grey
green, yellow
gold
dark
brown, grey
brown
blue-gray
blue
blonde
blond
black
auburn, white
auburn, grey
auburn





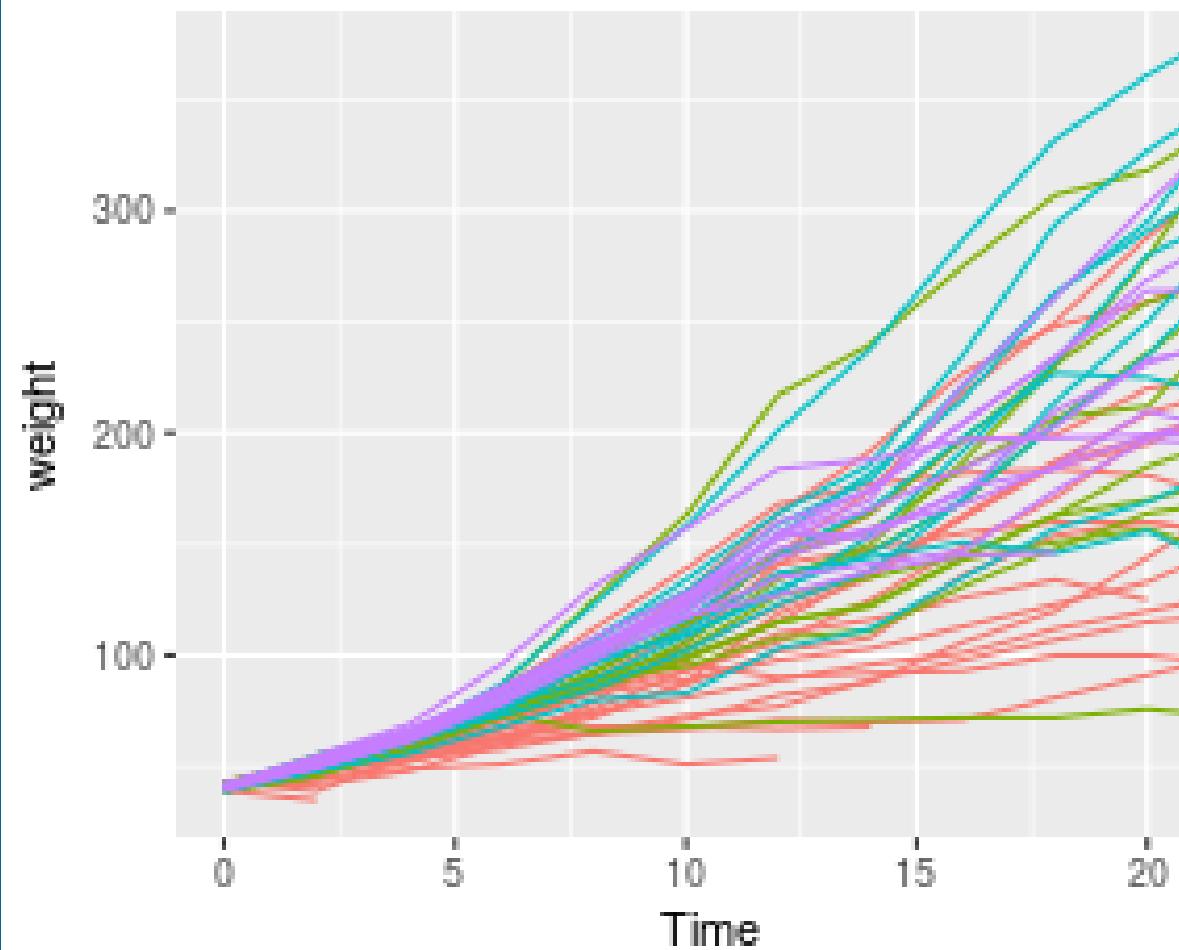
Biodiversity in New York



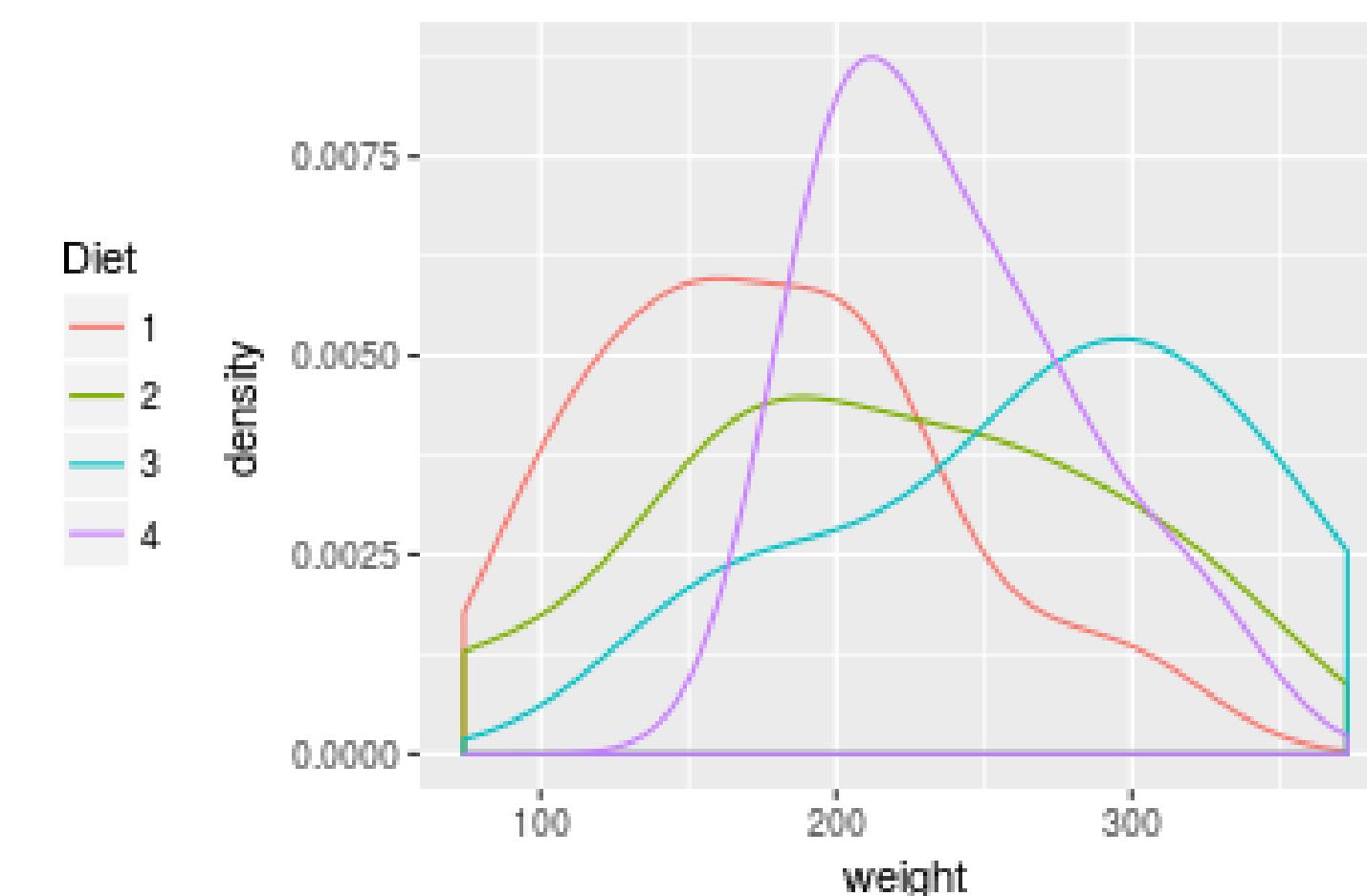


EXEMPLOS DE BONS GRÁFICOS

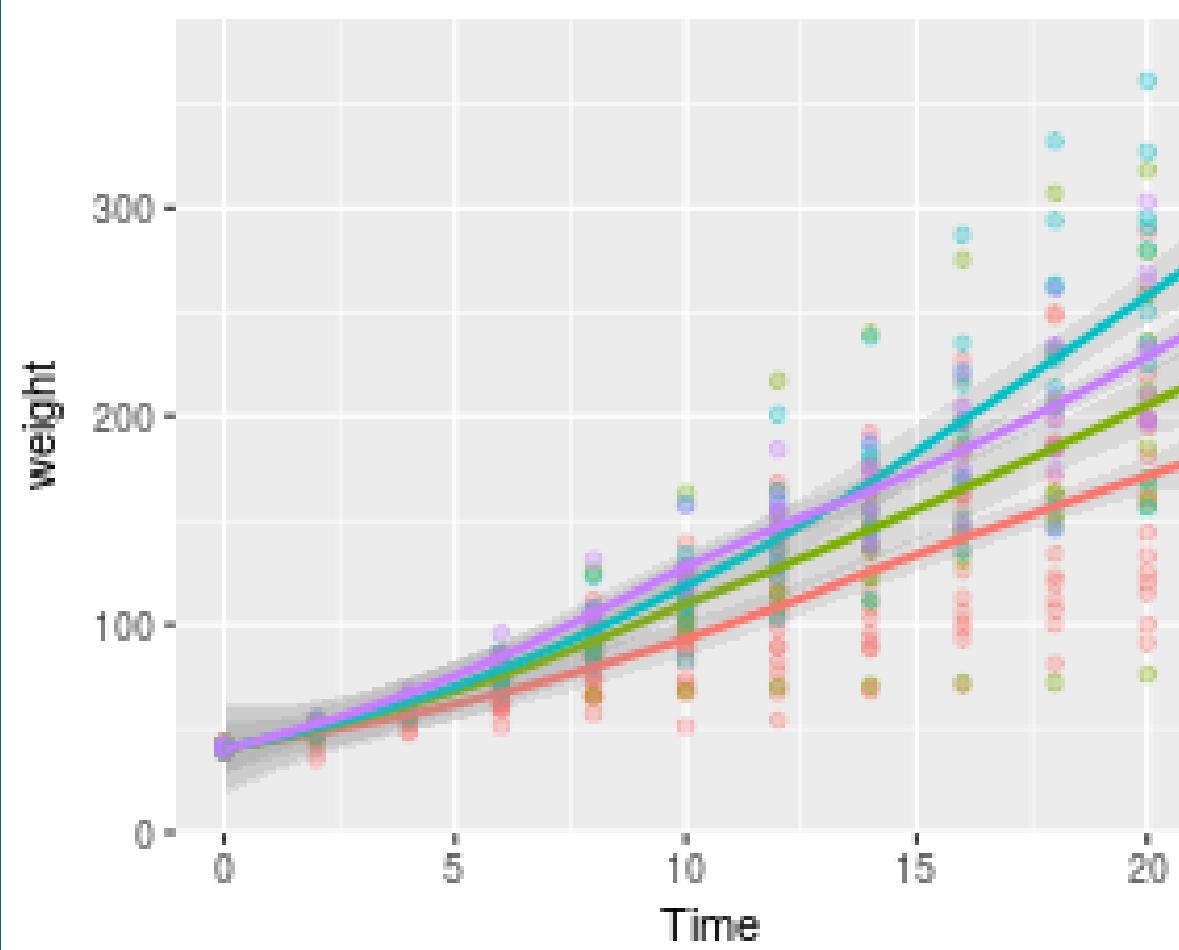
Growth curve for individual chicks



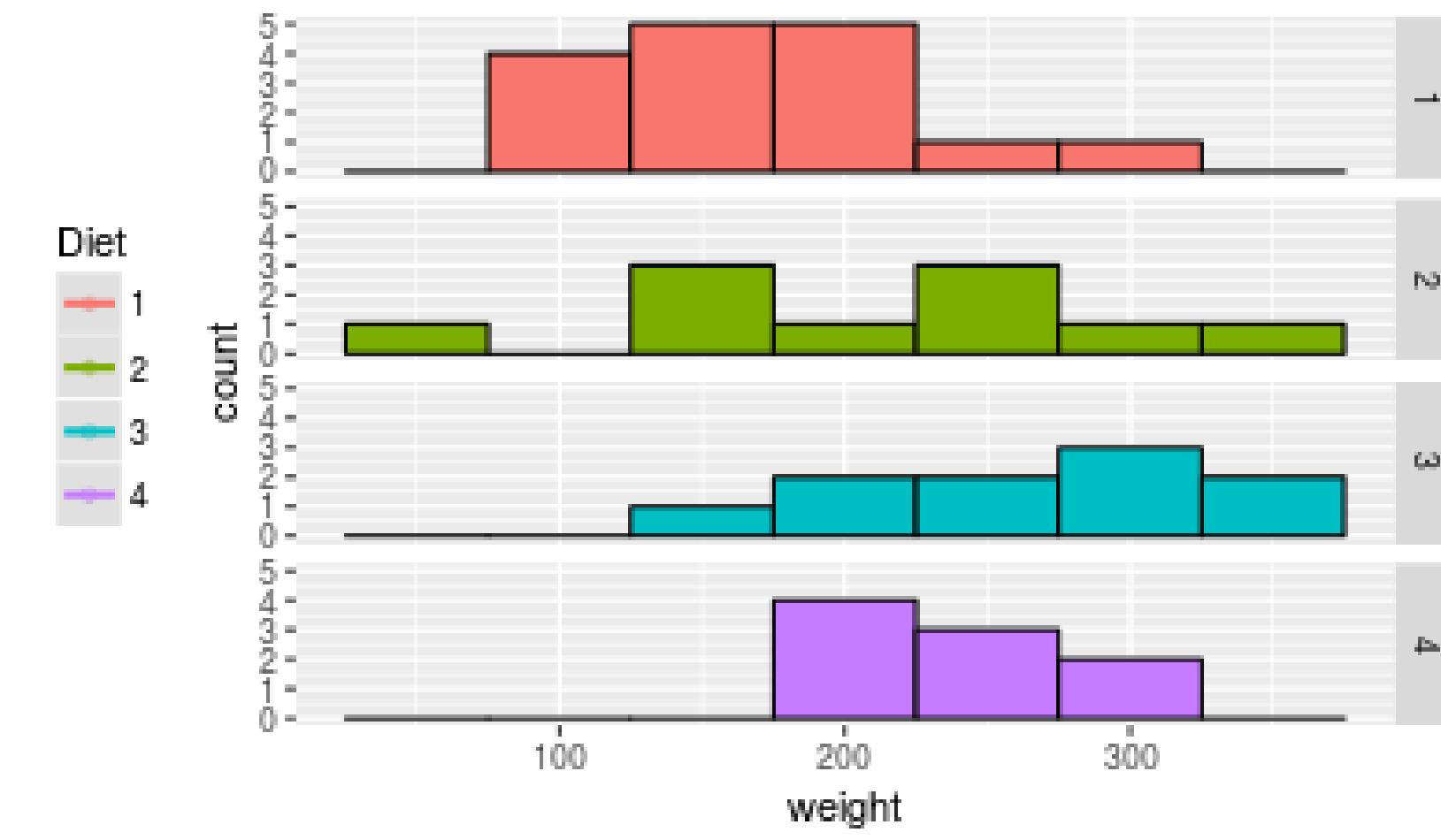
Final weight, by diet



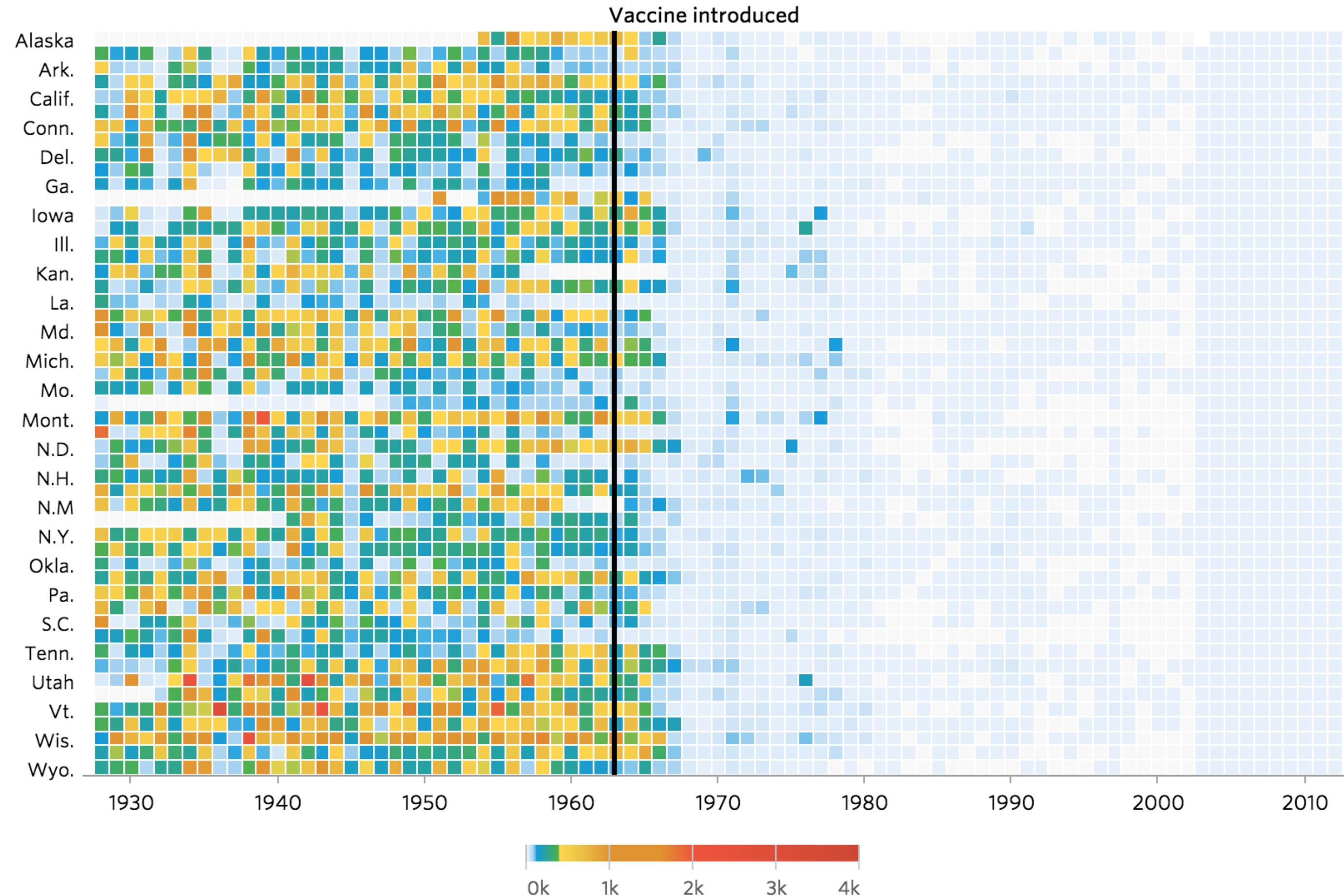
Fitted growth curve per diet



Final weight, by diet



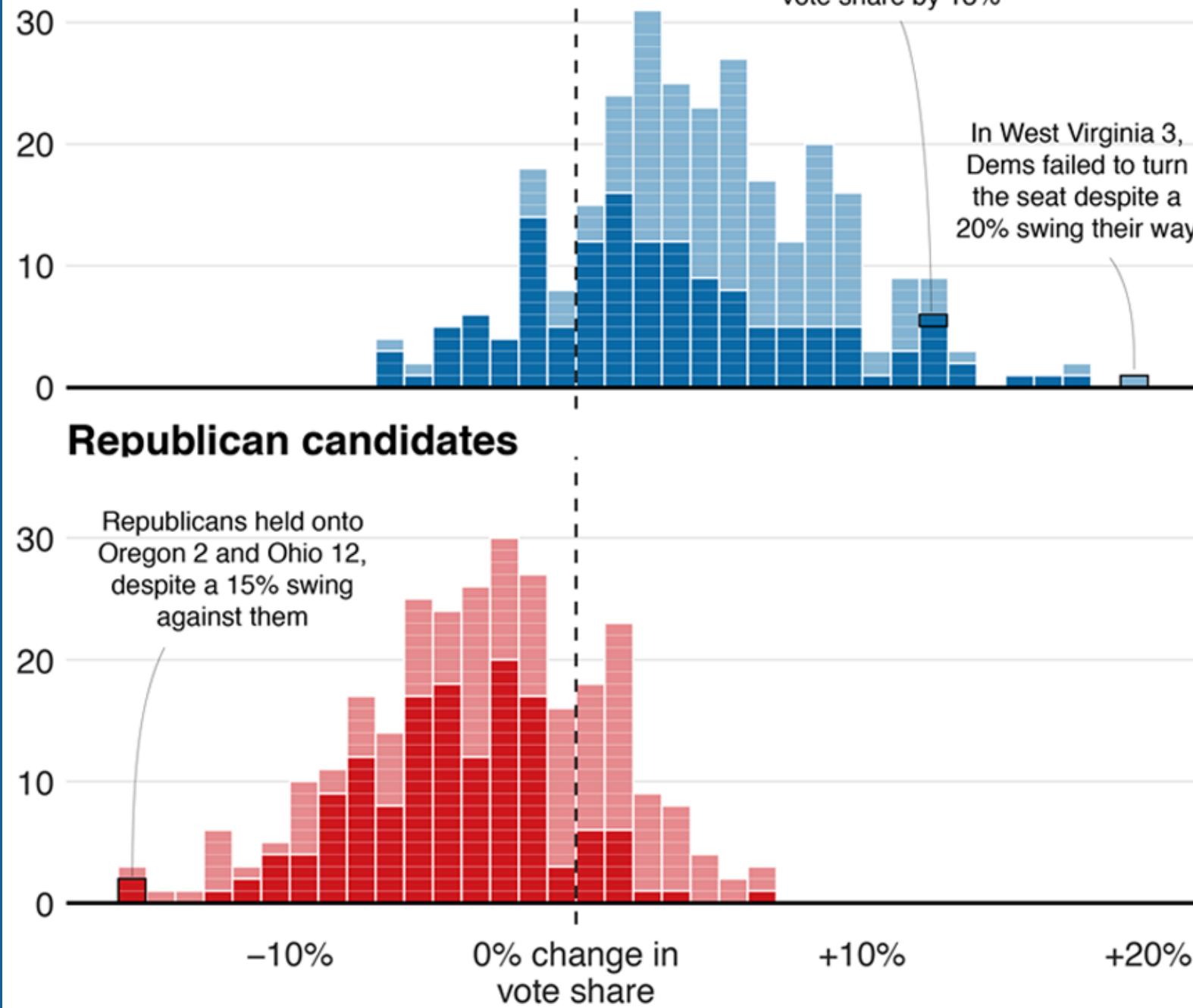
Measles



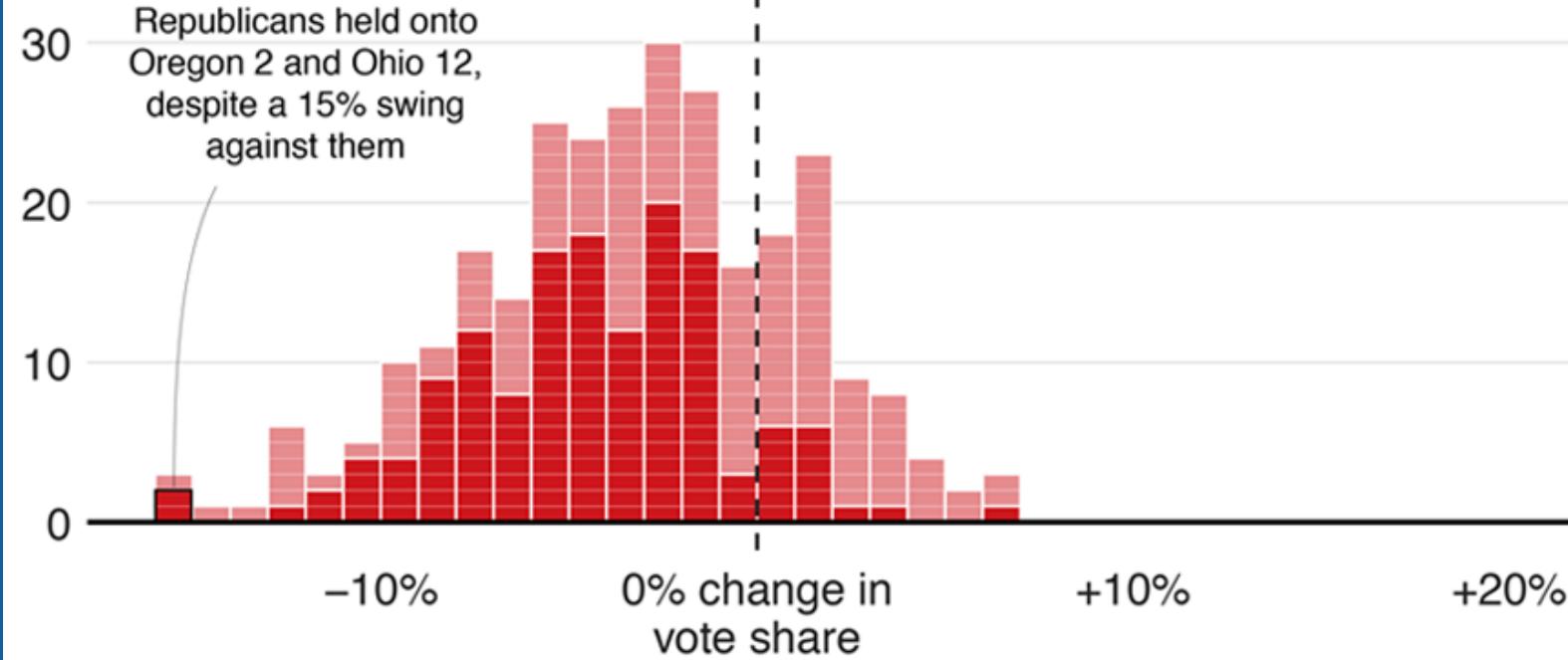
Blue wave

■ Won seat ■ Didn't win

Democrat candidates



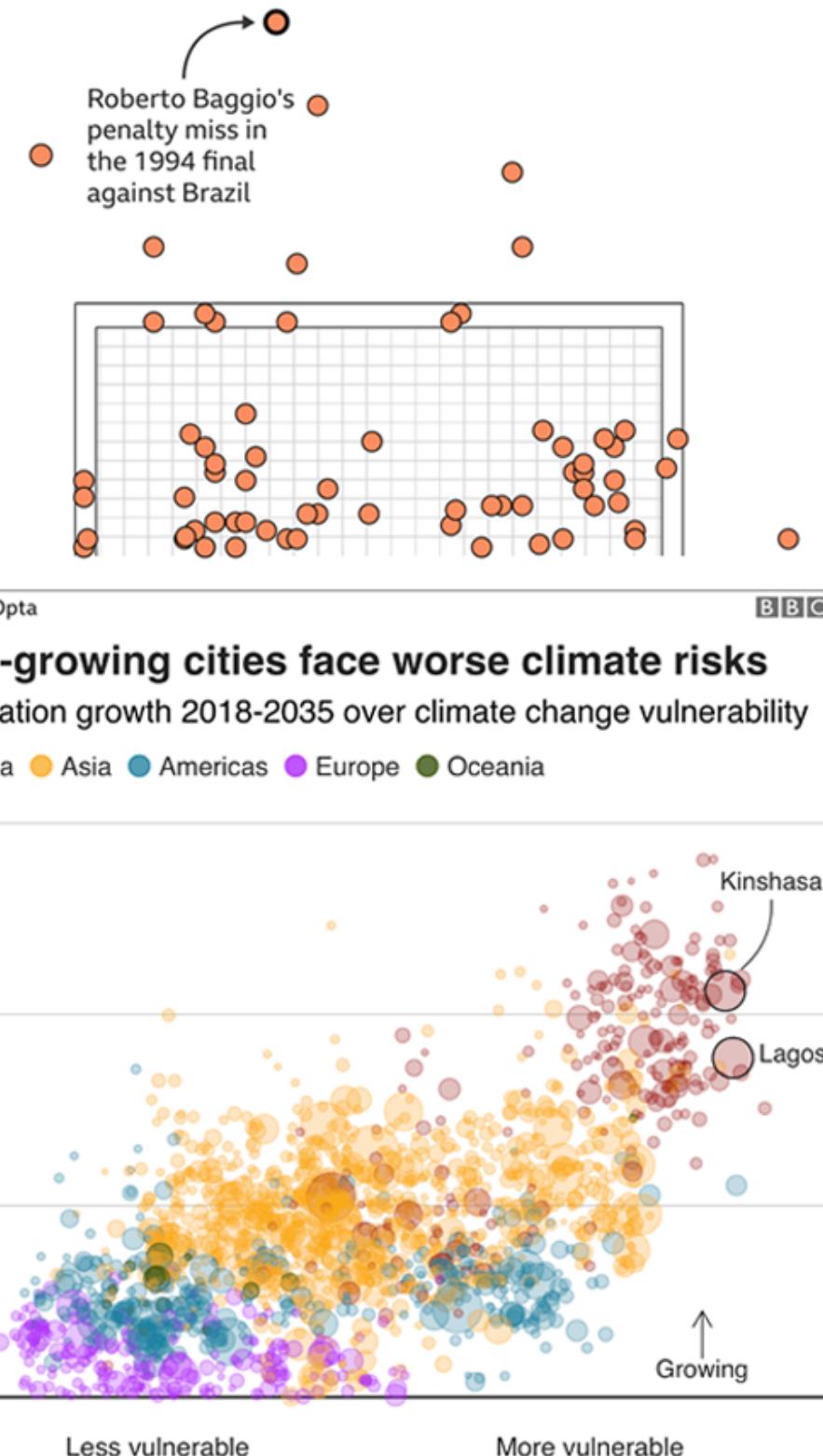
Republican candidates



Source: AP, 19:01 ET

Where penalties are saved

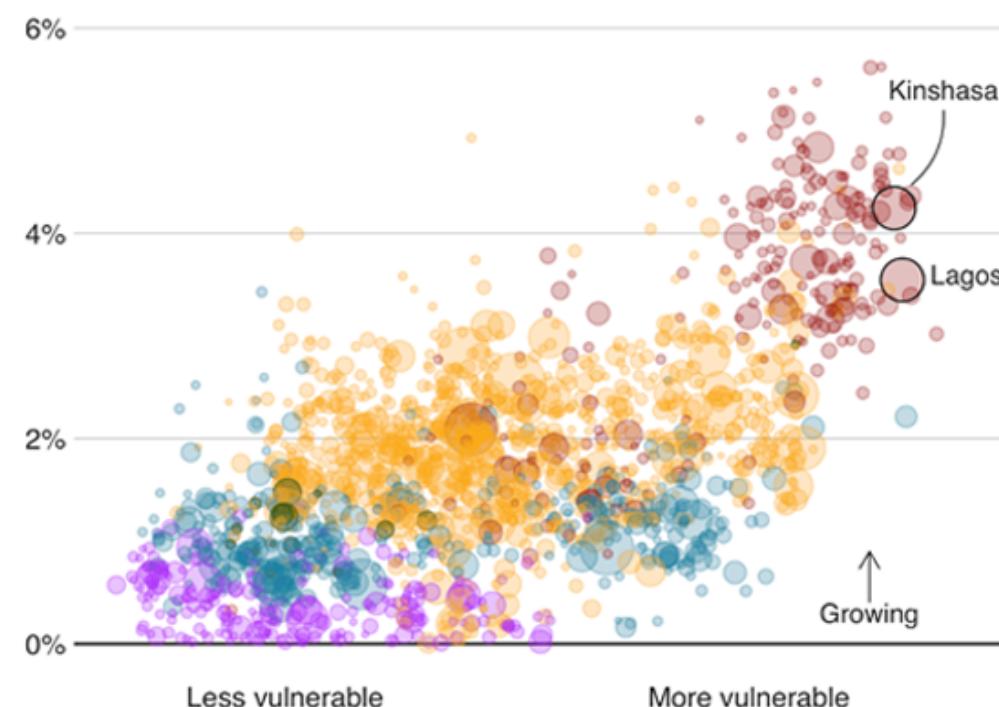
World Cup shootout misses and saves, 1982-2014



Fast-growing cities face worse climate risks

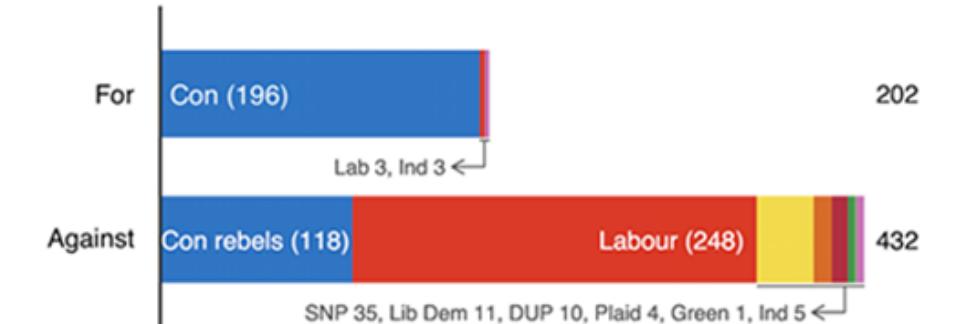
Population growth 2018-2035 over climate change vulnerability

■ Africa ■ Asia ■ Americas ■ Europe ■ Oceania



Source: Verisk Maplecroft. Circle size represents current population.

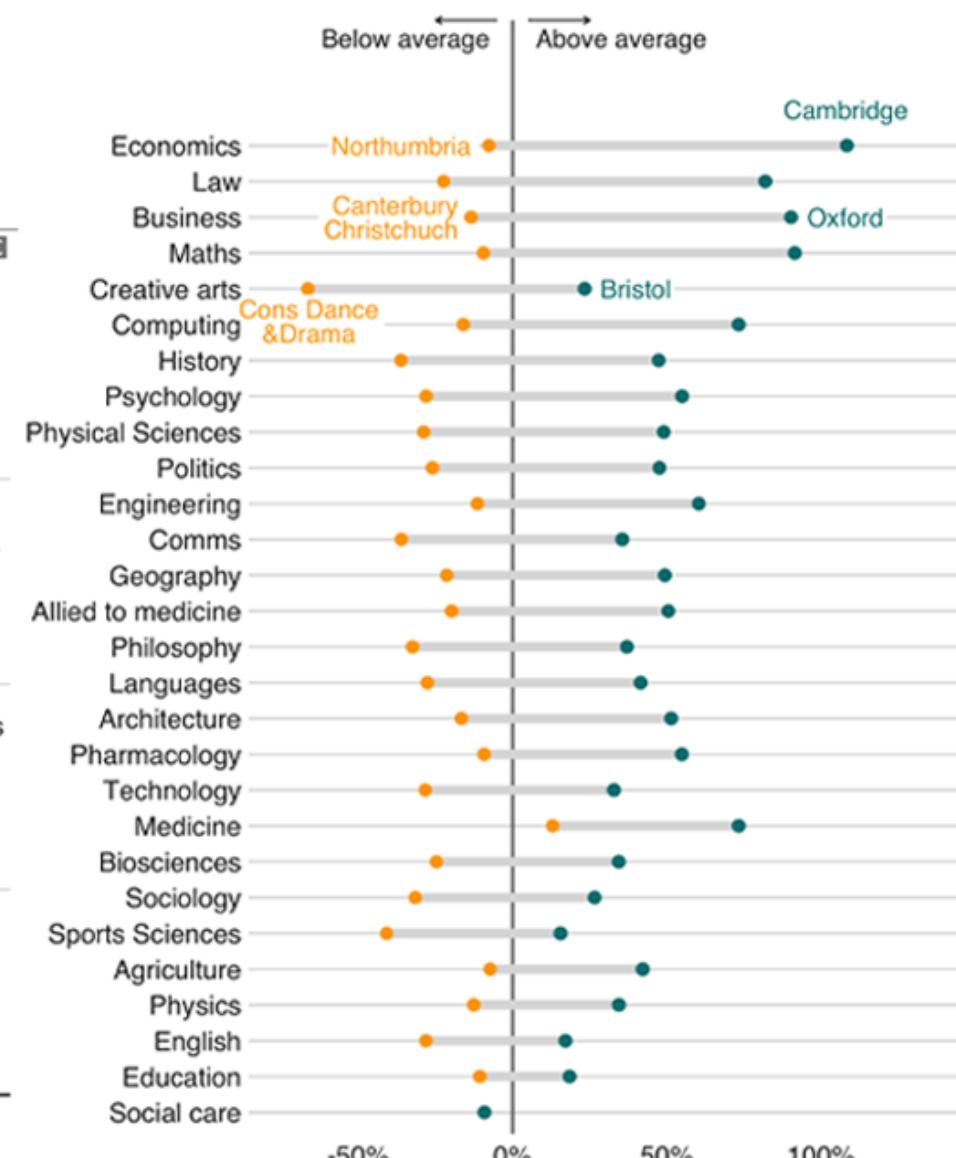
MPs rejected Theresa May's deal by 230 votes



Source: Commons Votes Services. Excludes 'tellers', the Speaker and deputies

Earnings vary across unis even within subjects

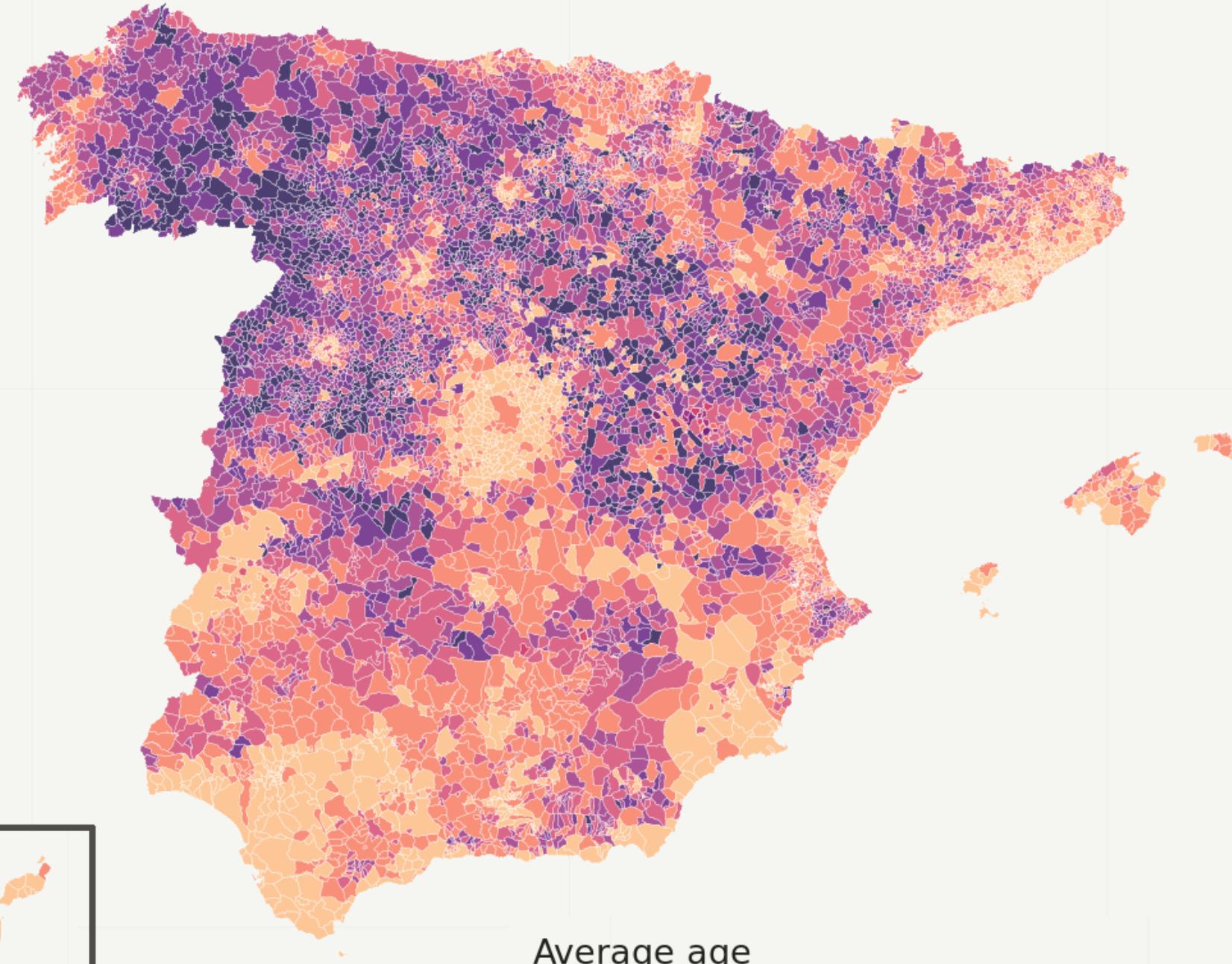
Impact on men's earnings relative to the average degree



Source: Institute for Fiscal Studies

Spain's regional demographics

Average age in Spanish municipalities, 2011



Author: Manuel Garrido (@manugarri) Original Idea: Timo Grossenbacher (@grssnbchr), Geometries: ArcGis Data: INE, 2011;

*thank
you*

bicalhoisabella@gmail.com
@bisnotforbella