Introdução à Ciência de Dados com R

Nelson Quesado
Caio Gustavo

A informação no mundo DOBRA a cada 20 meses.

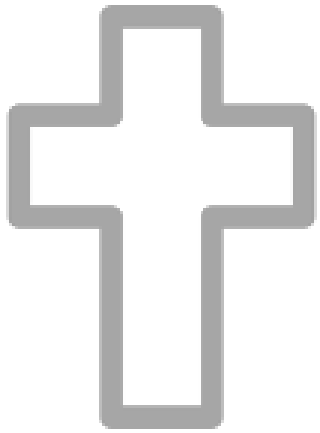
Thomas Runkle, 2012.

Data Analytics.

EMPÍRICO

OBSERVAÇÃO E INTERAÇÃO

RELIGIOSO



FÉ E DOGMA

FILOSÓFICO

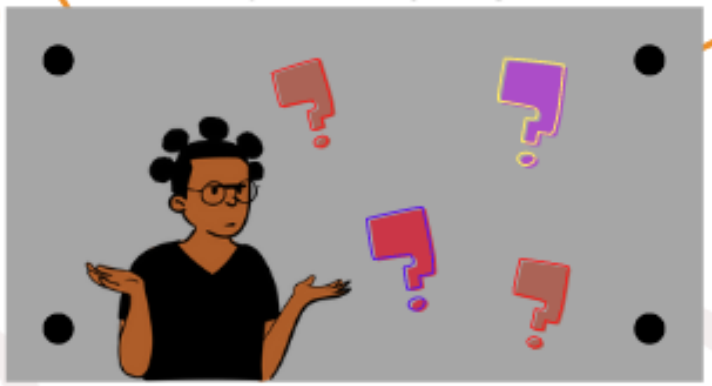
A NATUREZA DAS COISAS

CIENTÍFICO

LÓGICA VERIFICÁVEL

Observação de um fenômeno

Elaboração de perguntas



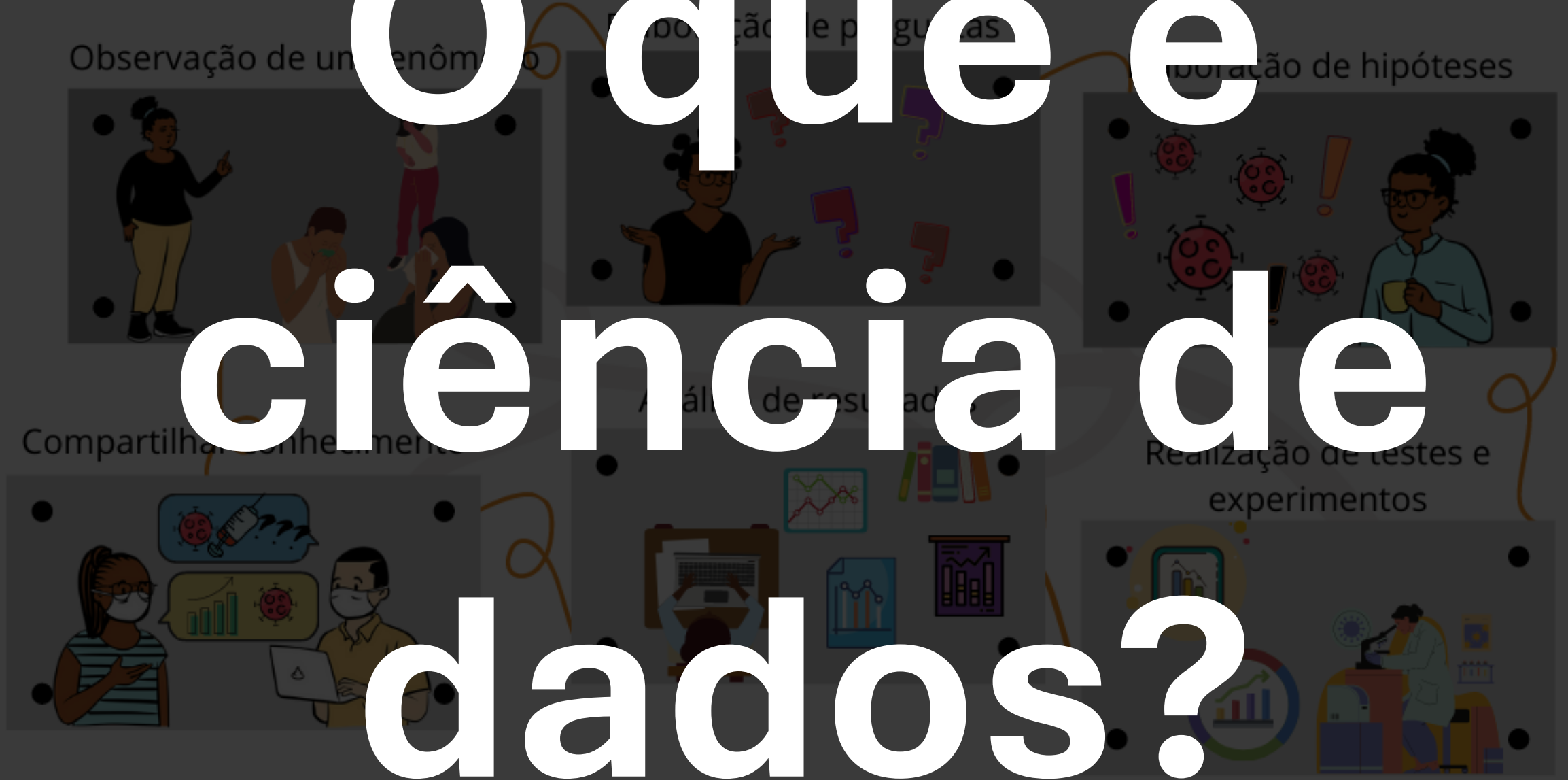
Elaboração de hipóteses

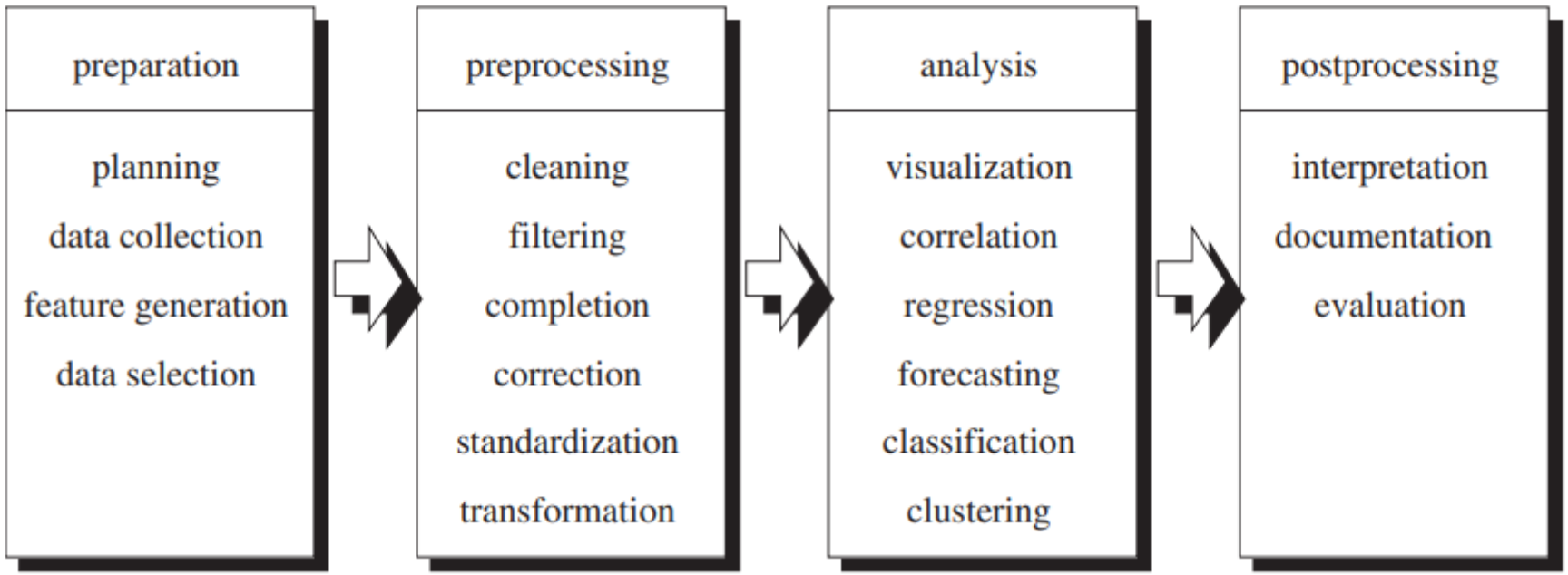
Compartilhar conhecimento

Análise de resultados

Realização de testes e experimentos

O que é ciência de dados?



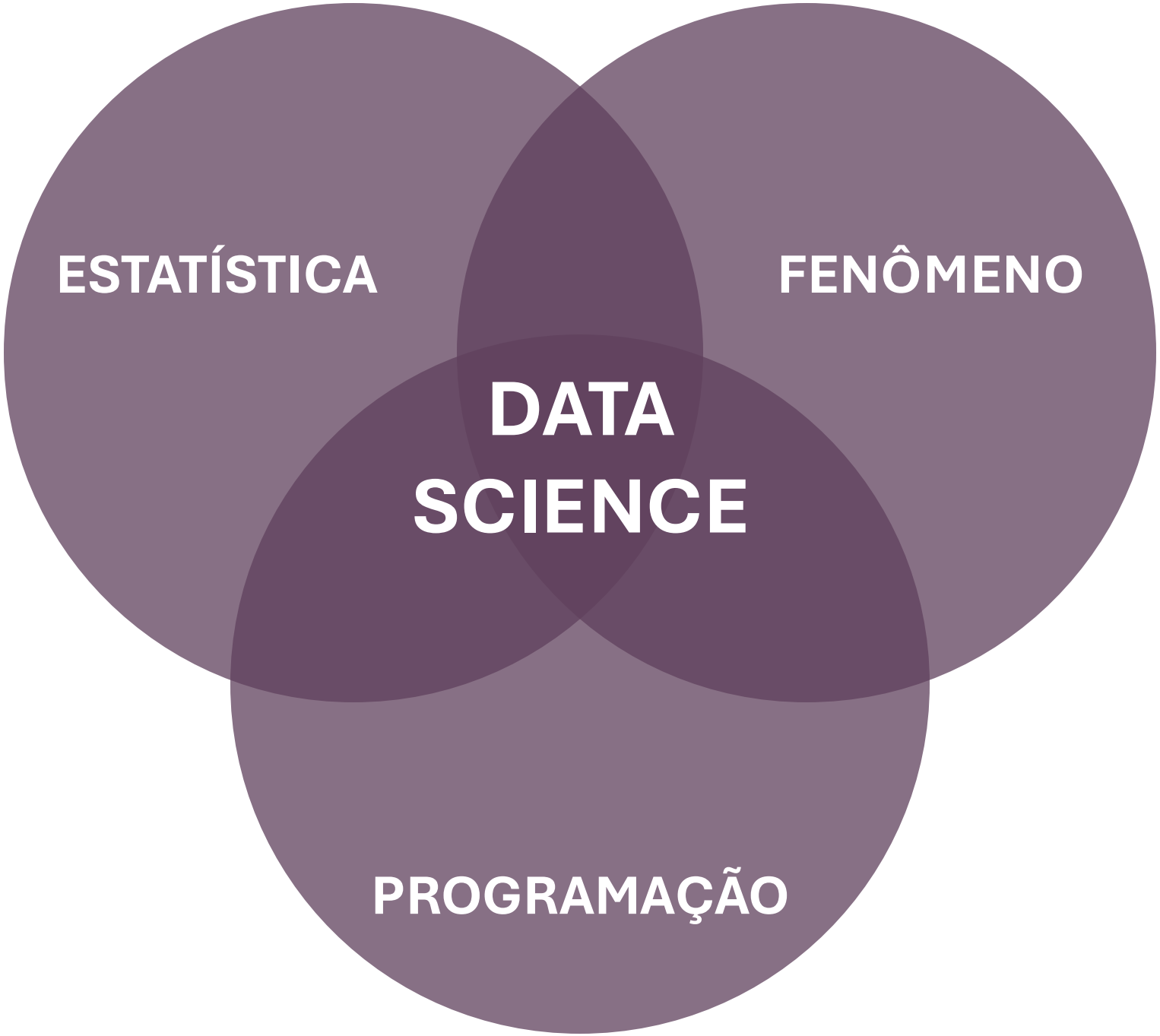


Thomas Runkle, 2012.
Data Analytics.

ESTATÍSTICA


ESTATÍSTICA
PROGRAMAÇÃO


ESTATÍSTICA
FENÔMENO
PROGRAMAÇÃO


ESTATÍSTICA
FENÔMENO
DATA
SCIENCE
PROGRAMAÇÃO

Fraud detection

Speech recognition

Internet search

Finance

Predictive maintenance

Genomics

Search engines

Healthcare

Airline route planning

Logistics

Video game

Supply chain optimization

Manufacturing

Sentiment analysis

Recommendation systems

Pattern recognition

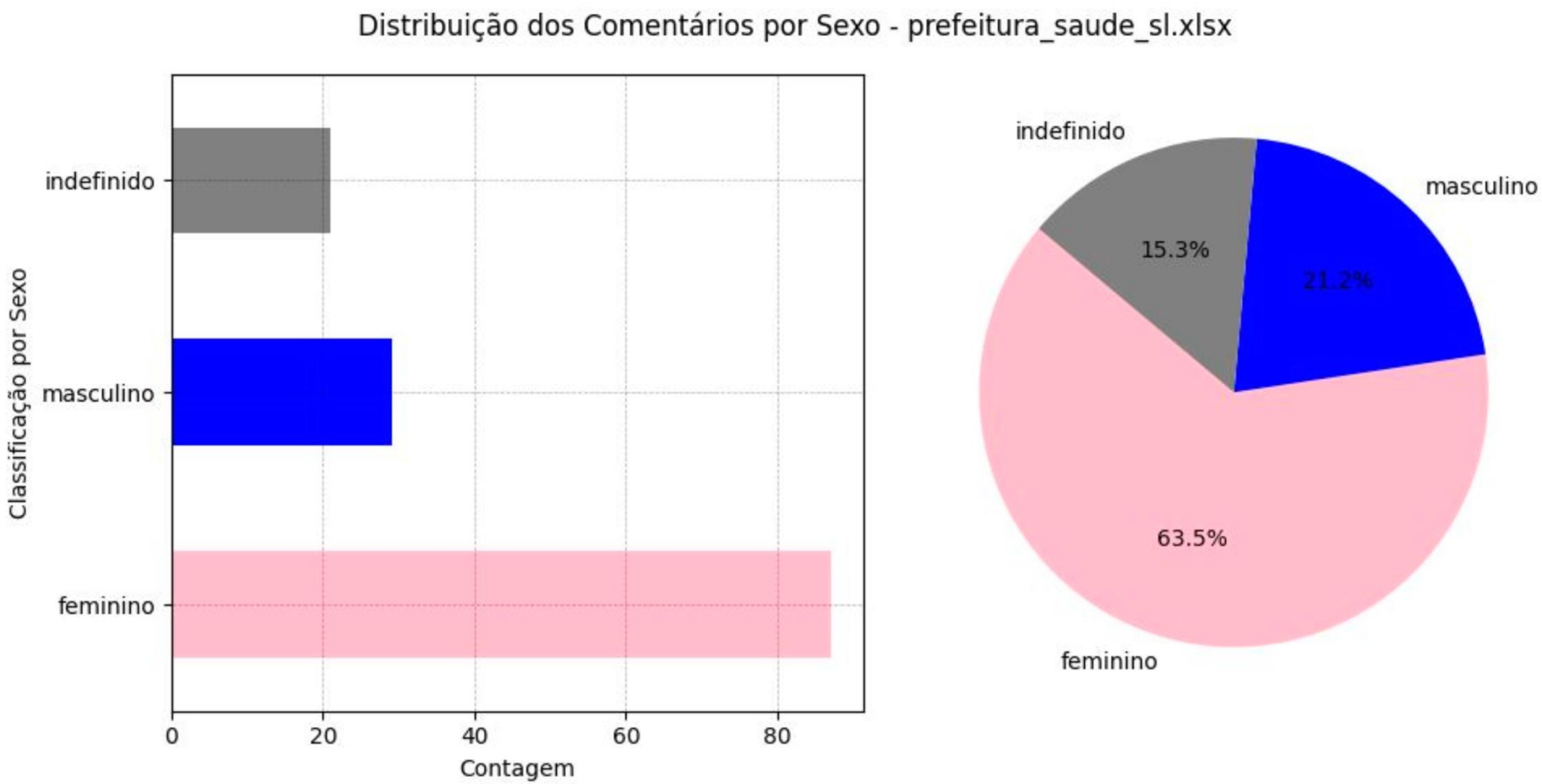
E-commerce

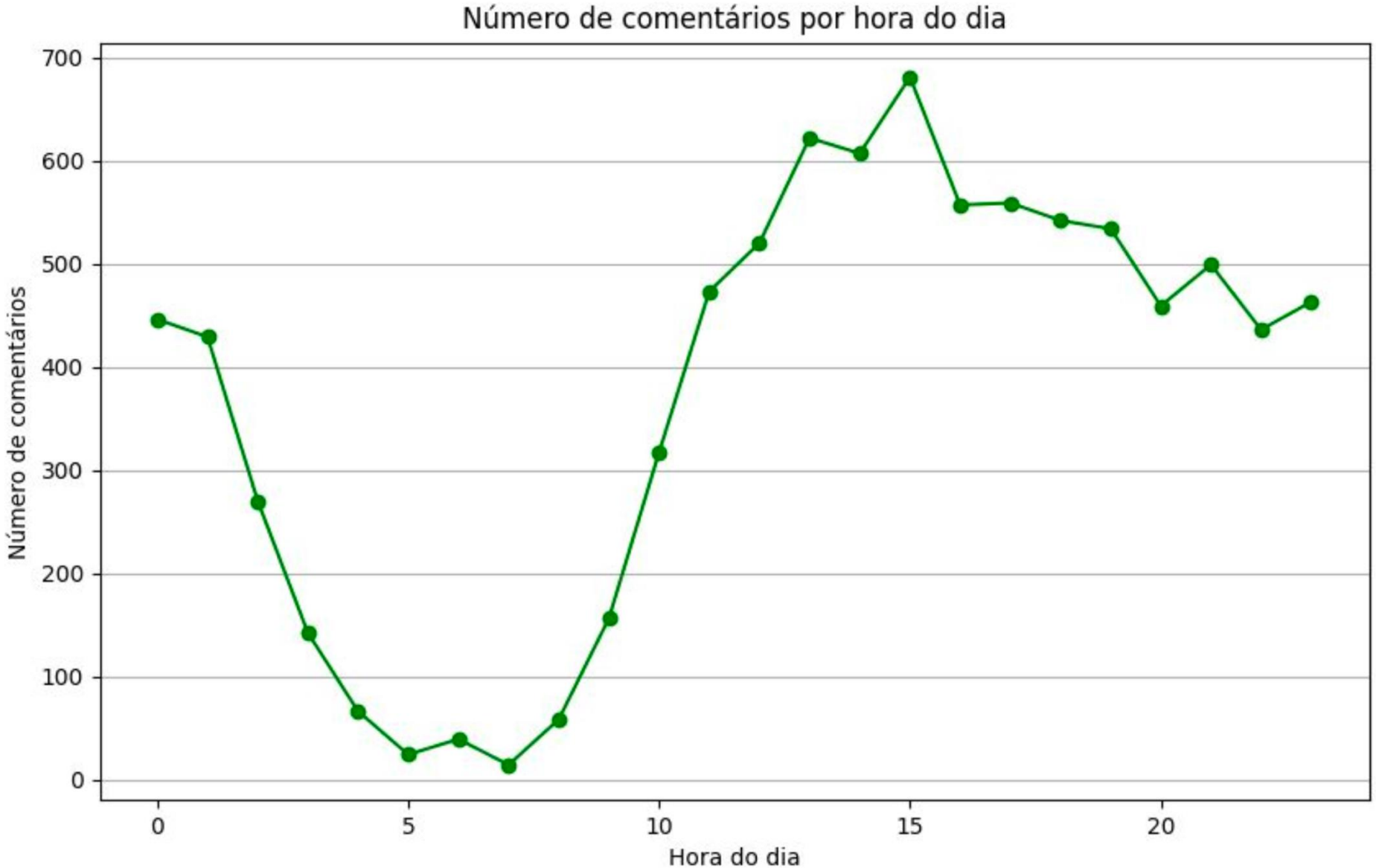
Exploratory data analysis

Targeted advertising

Retail analytics

Augmented reality

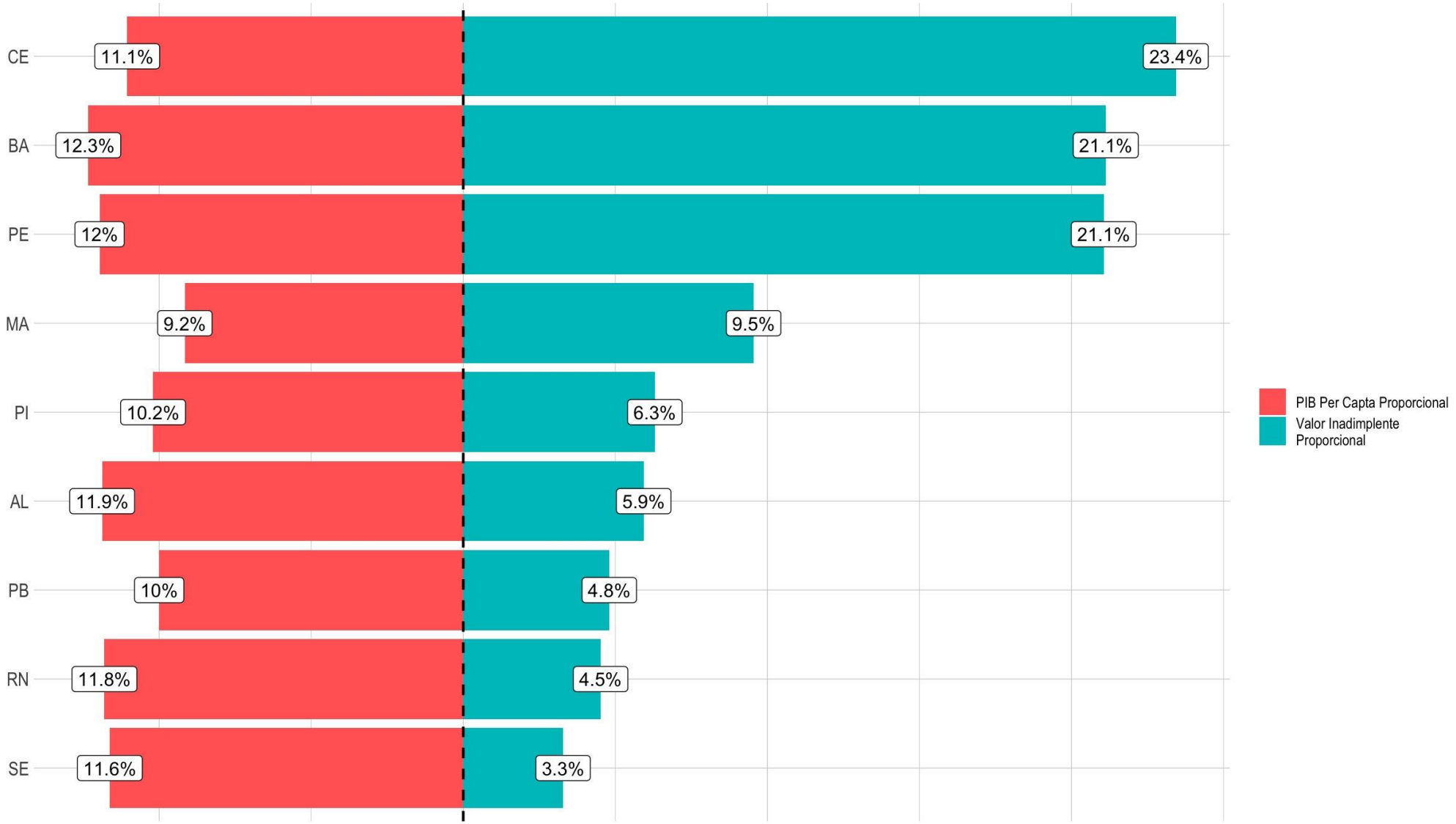

Distribuição dos Comentários por Sexo - prefeitura_saude_sl.xlsx
Classificação por Sexo
indefinido
masculino
feminino
0
20
40
60
80
Contagem
indefinido
masculino
15.3%
21.2%
63.5%
feminino


Número de comentários por hora do dia
Número de comentários
700
600
500
400
300
200
100
0
0
5
10
15
20
Hora do dia


Cadêencomenda
Queroencomenda
Liberemencomenda
Liberaencomenda
Parabéns
Cadeencomenda
celsorussomanno
Quandovãoentregarencomenda
Palhaçada
Porfavor
Respondedirect
Absurdo
Umabsurdo
Respondamdirect
correio
proconsp
pagamentoconfirmado
encomendasparadas
Entreguemencomenda
Eencomenda
Olhemdirect
Descasotota
Descaso
correiosoficialcadêencomenda
proconcariocaoficial procon_rj
CADÊMINHAENCOMENDA
Oacontecendo
Agilidade
Atendimentopéssimo
Issoabsurdo
Queabsurdo
Boatarde
jornaloglobo tvglobo
ConcursoPúblicoUrgente
Mandeidirect
Cadêpedido
PeloamorDeus
jornaloglobo globonews
vergonha
Show
Porque
obrigada
Cadê
Euqueroencomenda
respondam
favor
correiosoficialdirect
Palhaçada
Quemaravilha
SP
Orgulho
Orgulhopertencer
amorDeus
CADEMINHAENCOMENDA
tvglobo globonews
encomenda
Ondeencomenda
Podemresponderdirect
MERESPONDAM
LIBEREMASENCOMENDAS
Porgentileza
Parabénsiniciativa
ParabénsCorreios
Encomendaparadadesdedia12
10agoranada
Liberempedido
receita_federal
correiosoficialrespondedirect
lamentável
olhemdirectfavor
respondedm
LIBEREMMINHAENCOMENDA
tvglobo jornaloglobo
proconbahia proconsp
correiosoficialrespondam
Olhadirect
Entreguempedido
LIBEREMNOSSAENCOMENDA
Ouseja
paga
DESDEODIA20SEMATUALIZAÇÕES
Encomendaparadadesdedia29
jornalhoje
proconcariocaoficial
Serviçohorrível
precisoencomenda
Desrespeitoconsumidor
Oacontecendoencomendas
LIBERENOSSAENCOMENDA
Serviçopéssimo
Mentira
jornaldarecord
Minhaencomendadesdedia29
Desdedia16
gentileza
querodinheirovolta
Desdedia29
carteiroatendido
Olhemdirect
Péssimoserviço
correiosoficialolhadirect
procon_rj proconsp
blá blá
Faltarespeito
09nada
procon_rj
Taxapaga
socorro
Cadêencomendacorreios
Queremosencomendas
Pfv
Fabiano
Desdedia26
liberanossasencomendas
MinhaencomendapagasaiCuritiba
misericórdia
paguei
Piorserviço
Eaí
atendem
precisoajuda
Minhaencomenda
Palhaçad
Querosaberencomenda
Precisosaberondepedido
Queremosrespostas
09Curitiba
respondamdm
Absurdoisso
Encomendaparadadesdedia16
DESDEODIA20SEMATUALIZAÇÕES
NB693164751BR

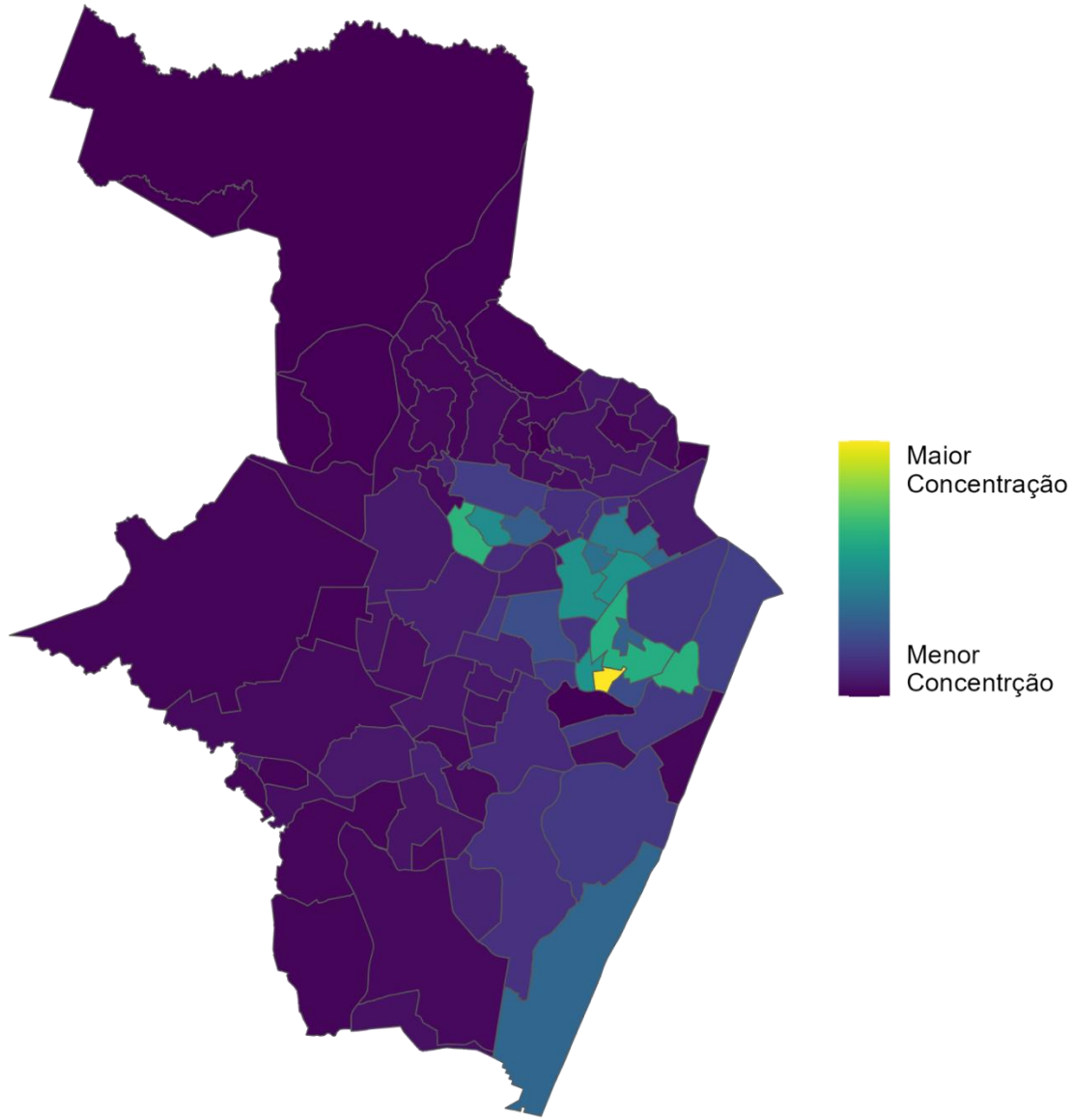
Capacidade de Pagamento de cada Estado
Pib Per Capta Proporcional vs. Valor Inadimplente Proporcional

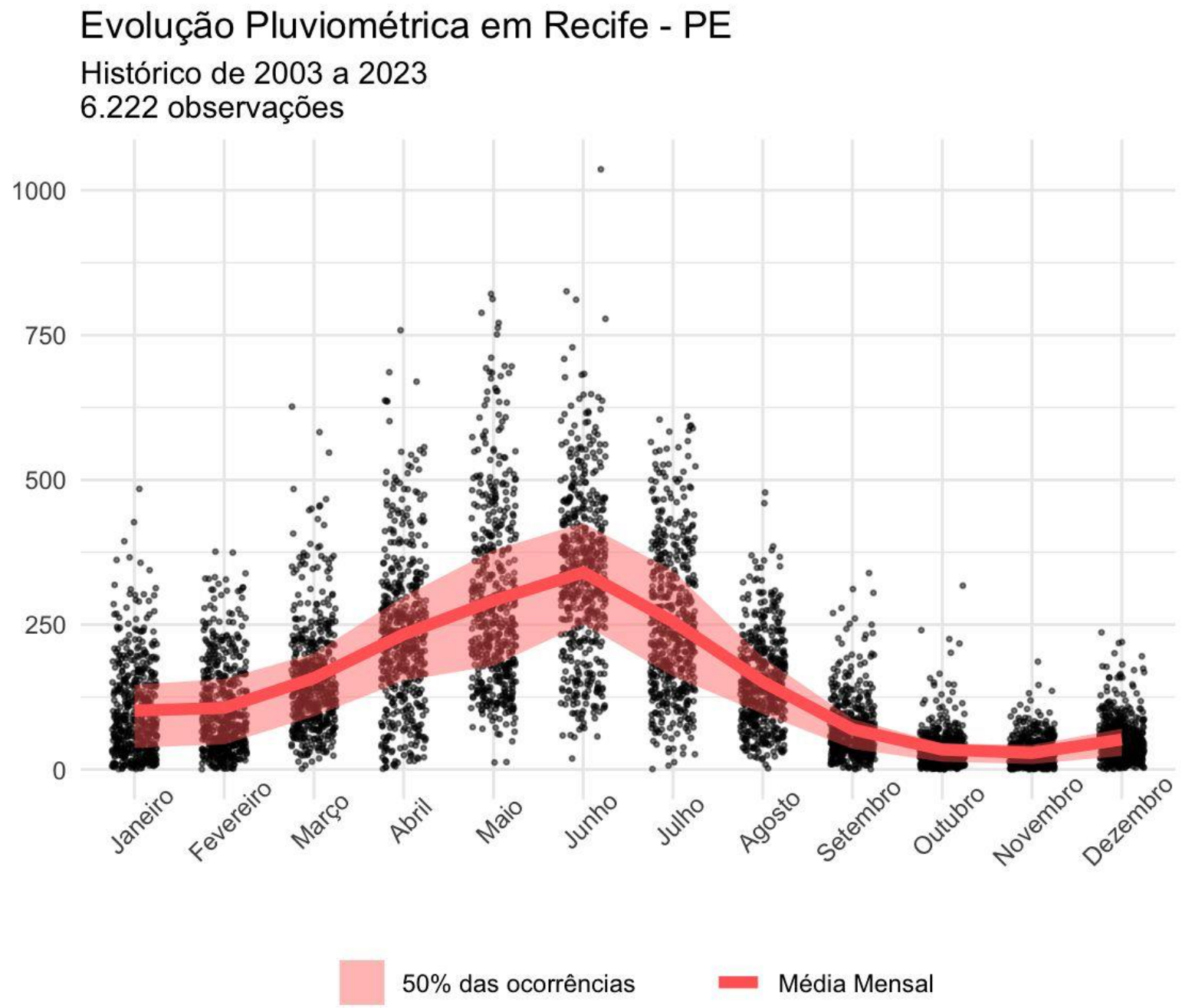
Fonte: Banco do Nordeste e IBGE



Concentração de Firmas por Bairro em Recife

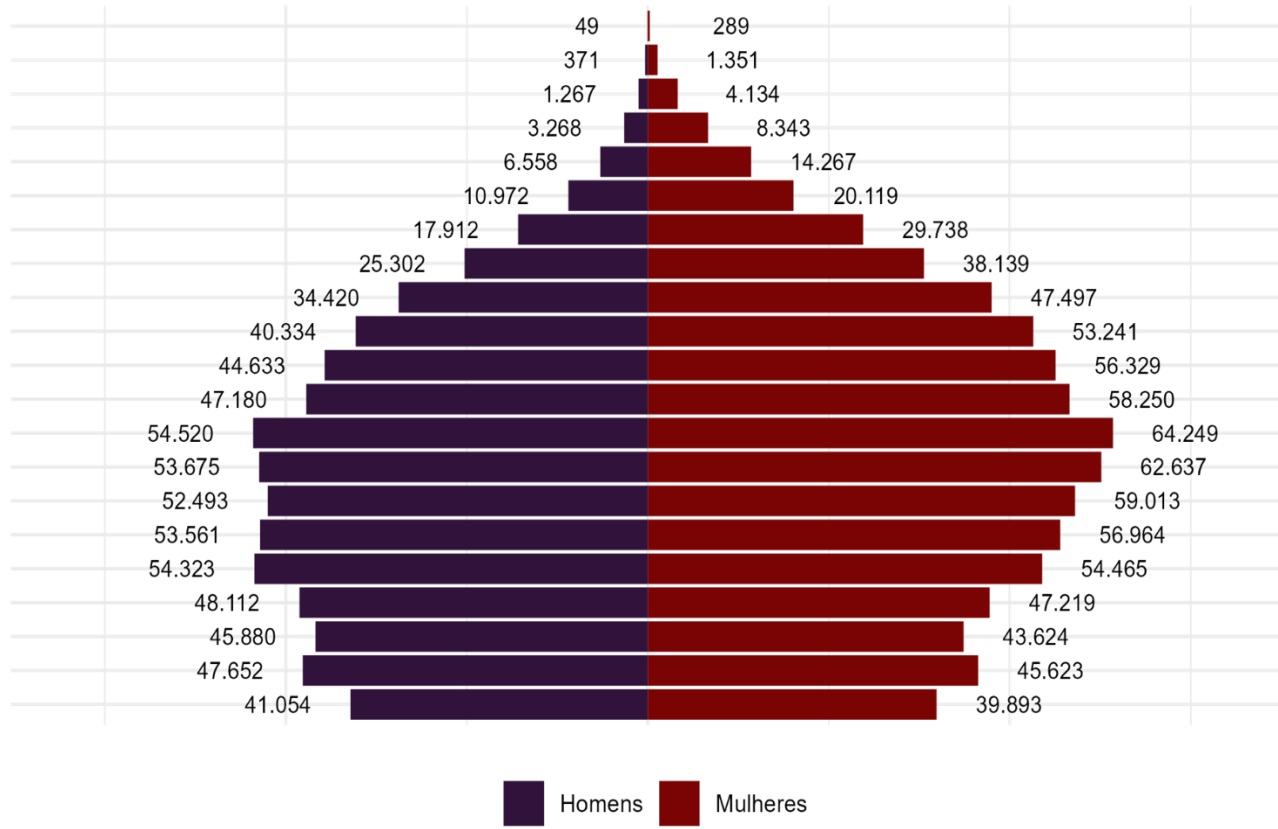
Fonte: Dados Abertos - Prefeitura de Recife

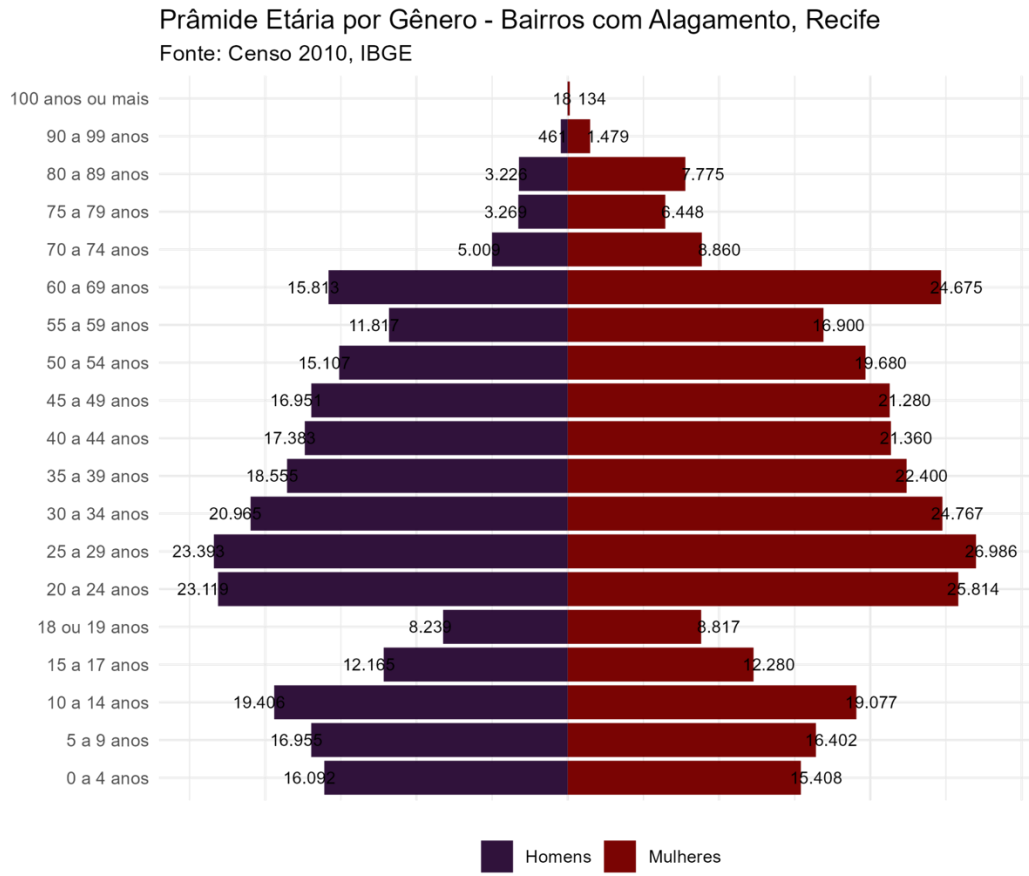


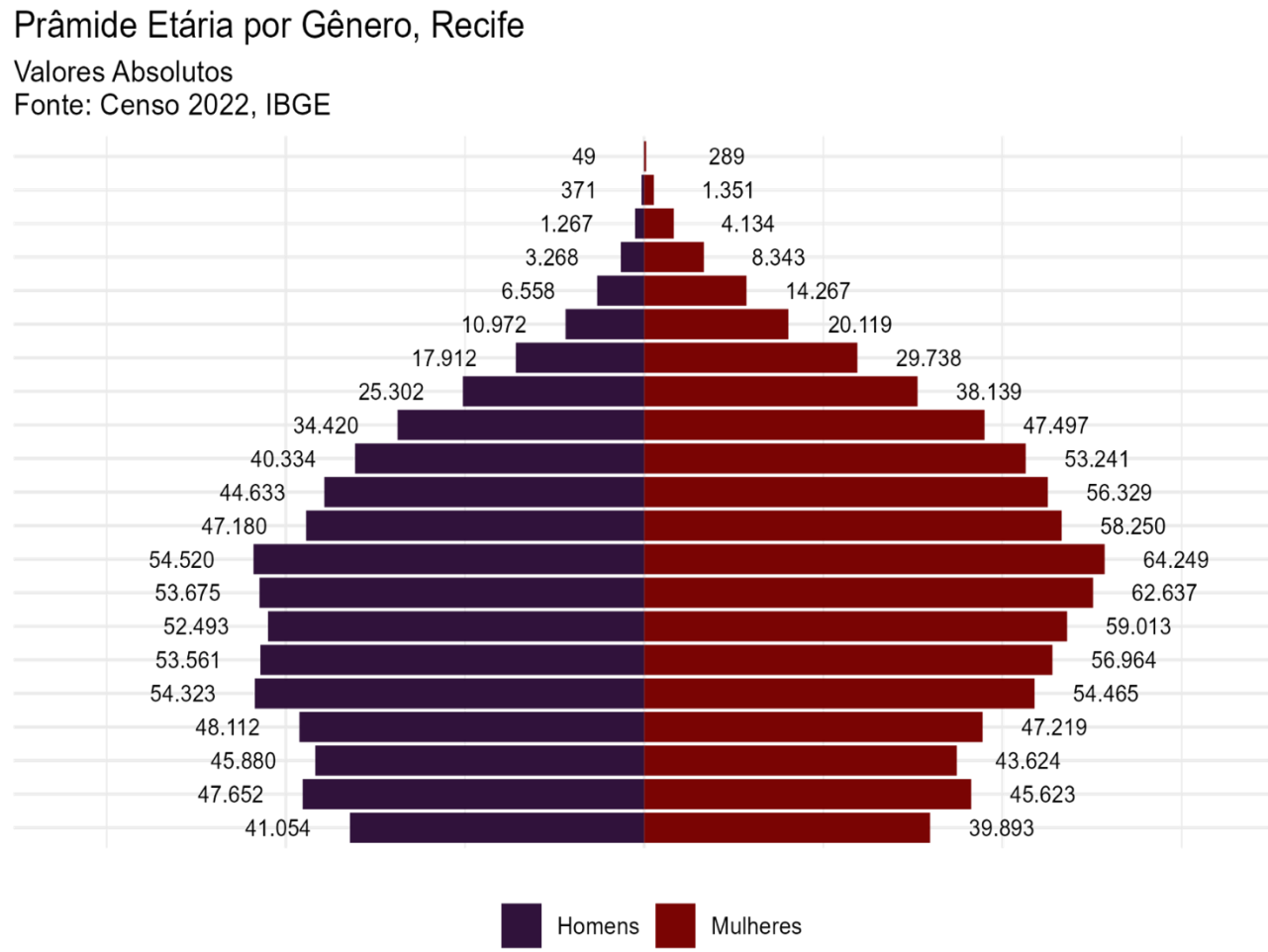

Evolução Pluviométrica em Recife - PE
Histórico de 2003 a 2023
6.222 observações
1000
750
500
250
0
Janeiro
Fevereiro
Março
Abril
Maio
Junho
Julho
Agosto
Setembro
Outubro
Novembro
Dezembro
50% das ocorrências
Média Mensal

Prâmide Etária por Gênero, Recife

Valores Absolutos
Fonte: Censo 2022, IBGE

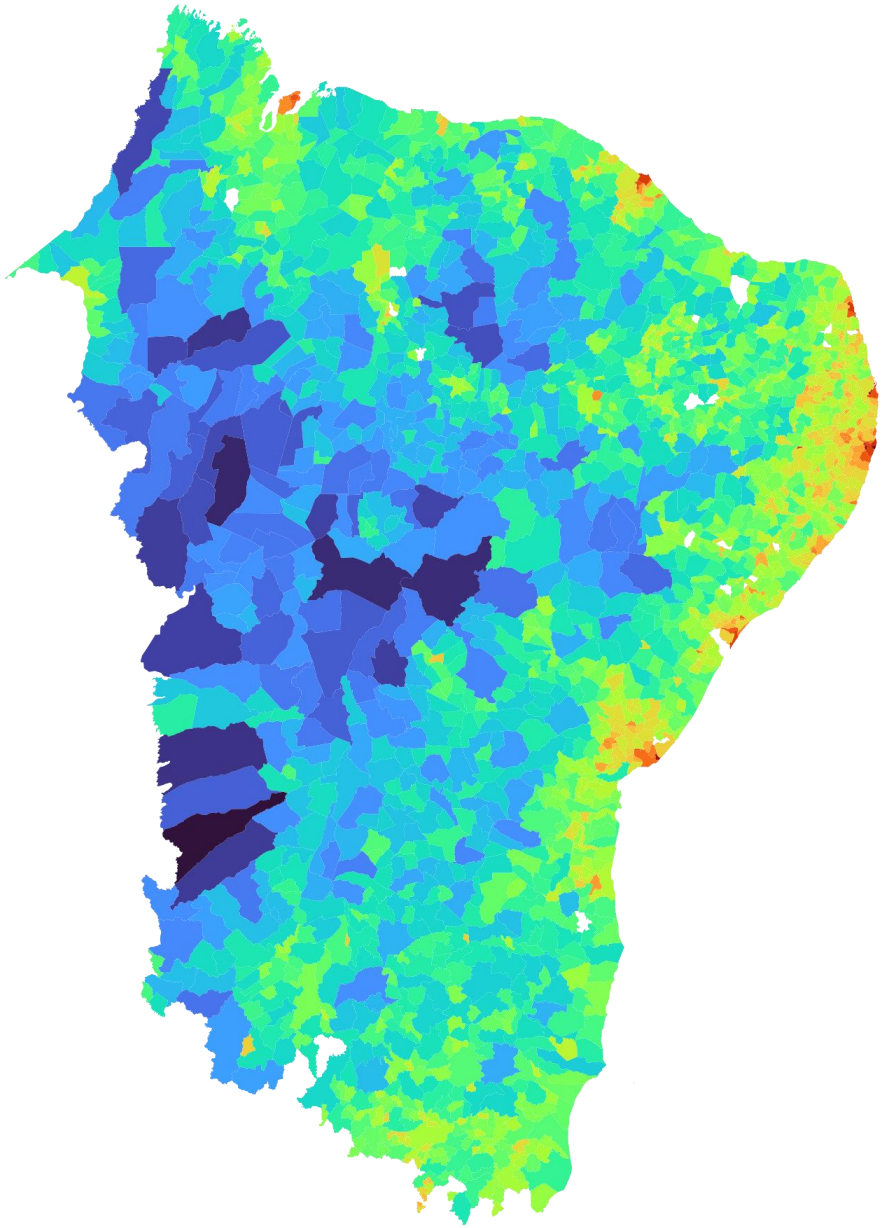


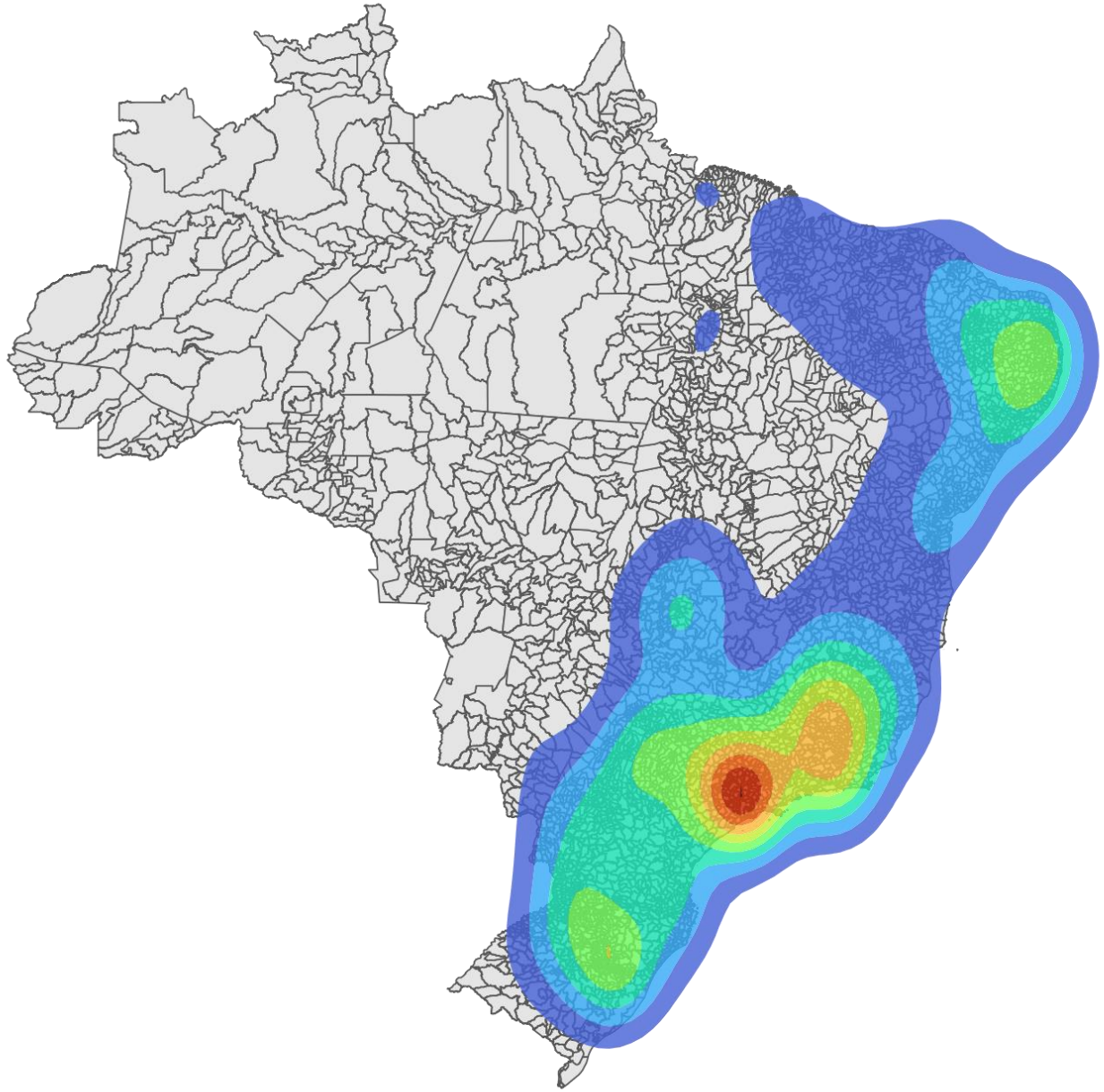

Prâmide Etária por Gênero - Bairros com Alagamento, Recife
Fonte: Censo 2010, IBGE
100 anos ou mais
90 a 99 anos
80 a 89 anos
75 a 79 anos
70 a 74 anos
60 a 69 anos
55 a 59 anos
50 a 54 anos
45 a 49 anos
40 a 44 anos
35 a 39 anos
30 a 34 anos
25 a 29 anos
20 a 24 anos
18 ou 19 anos
15 a 17 anos
10 a 14 anos
5 a 9 anos
0 a 4 anos
15.813
24.675
11.817
16.900
15.107
19.680
16.951
21.280
17.385
21.360
18.555
22.400
20.965
24.767
23.393
26.986
23.119
25.814
8.239
8.817
12.165
12.280
19.406
19.077
16.955
16.402
16.092
15.408
Homens
Mulheres


Prâmide Etária por Gênero, Recife
Valores Absolutos
Fonte: Censo 2022, IBGE
49
289
371
1.351
1.267
4.134
3.268
8.343
6.558
14.267
10.972
20.119
17.912
29.738
25.302
38.139
34.420
47.497
40.334
53.241
44.633
56.329
47.180
58.250
54.520
64.249
53.675
62.637
52.493
59.013
53.561
56.964
54.323
54.465
48.112
47.219
45.880
43.624
47.652
45.623
41.054
39.893
Homens
Mulheres

Densidade de Candidaturas a Vereador

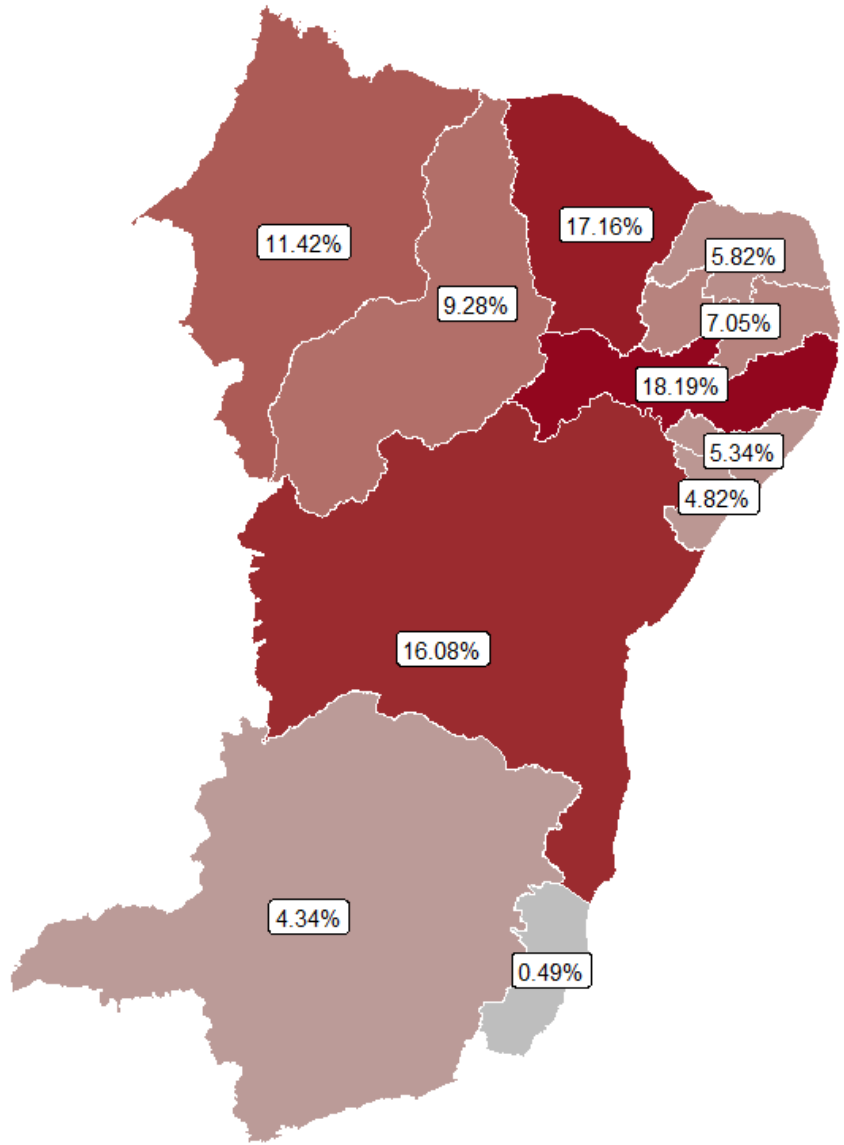
Fonte: Estatísticas Eleitorais Candidatos 2020 - TSE 2024



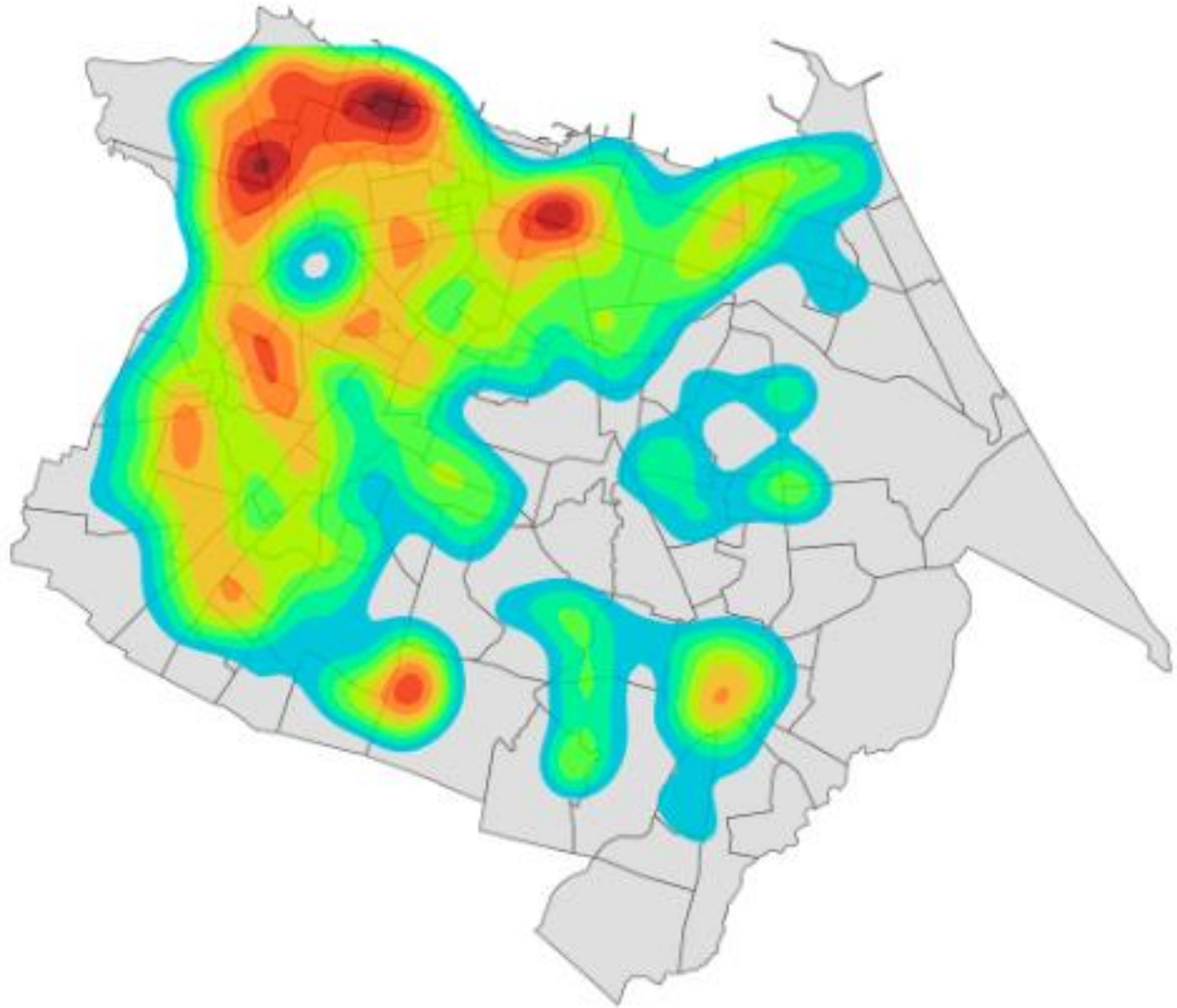


Recomendação Técnica de Investimento por Estado

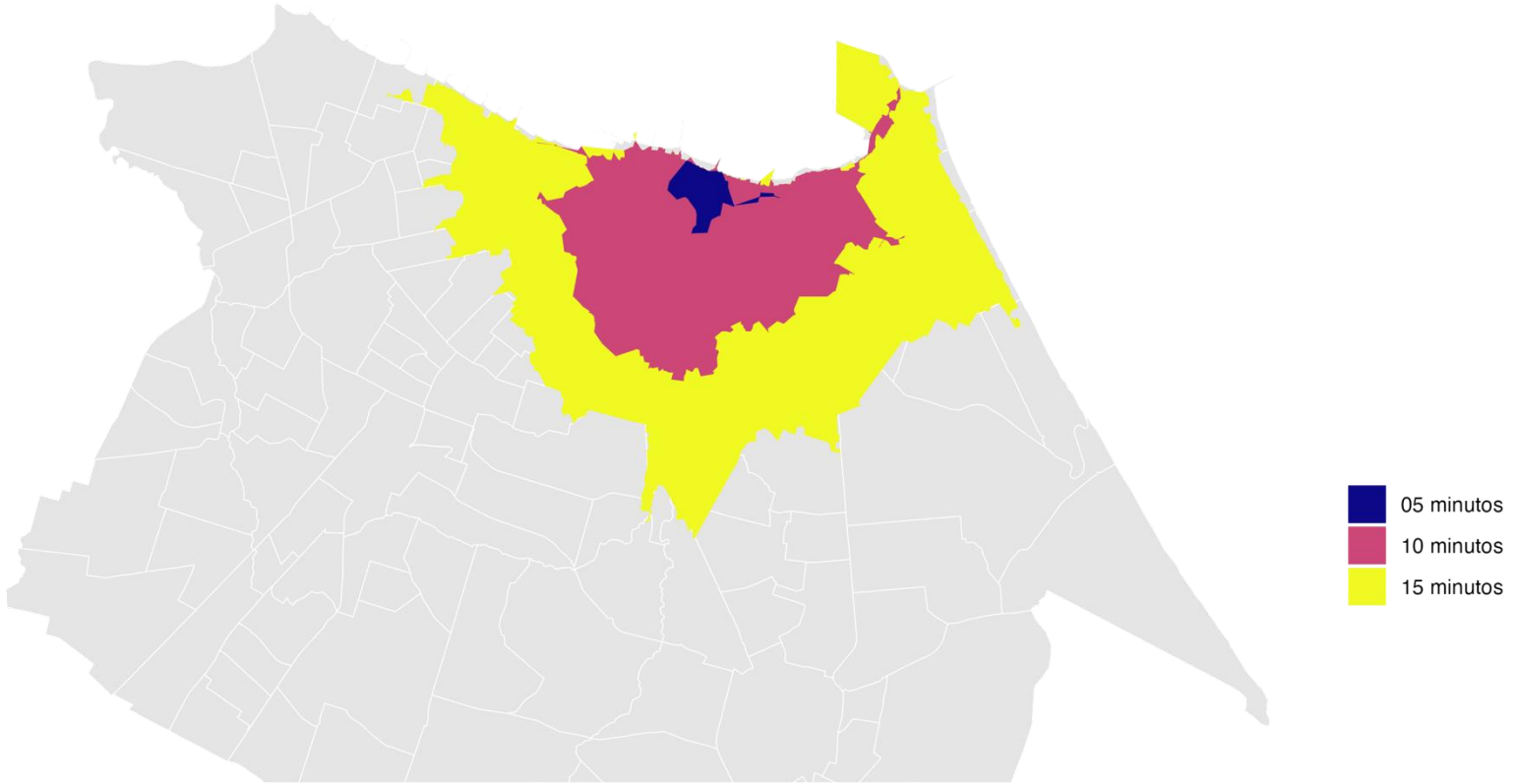
Fonte: Banco do Nordeste



Votos Elmano



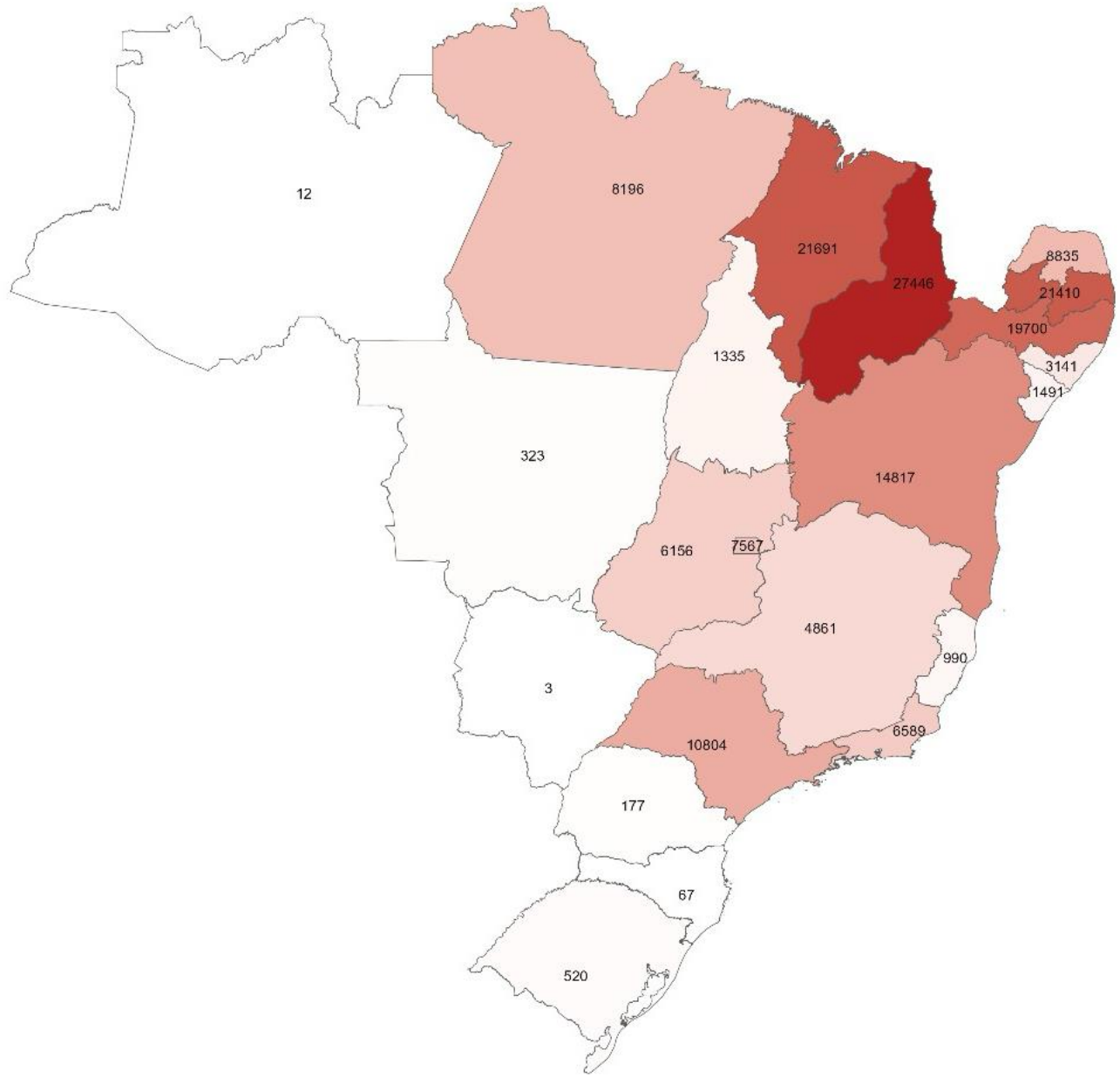
Zonas de Interesse

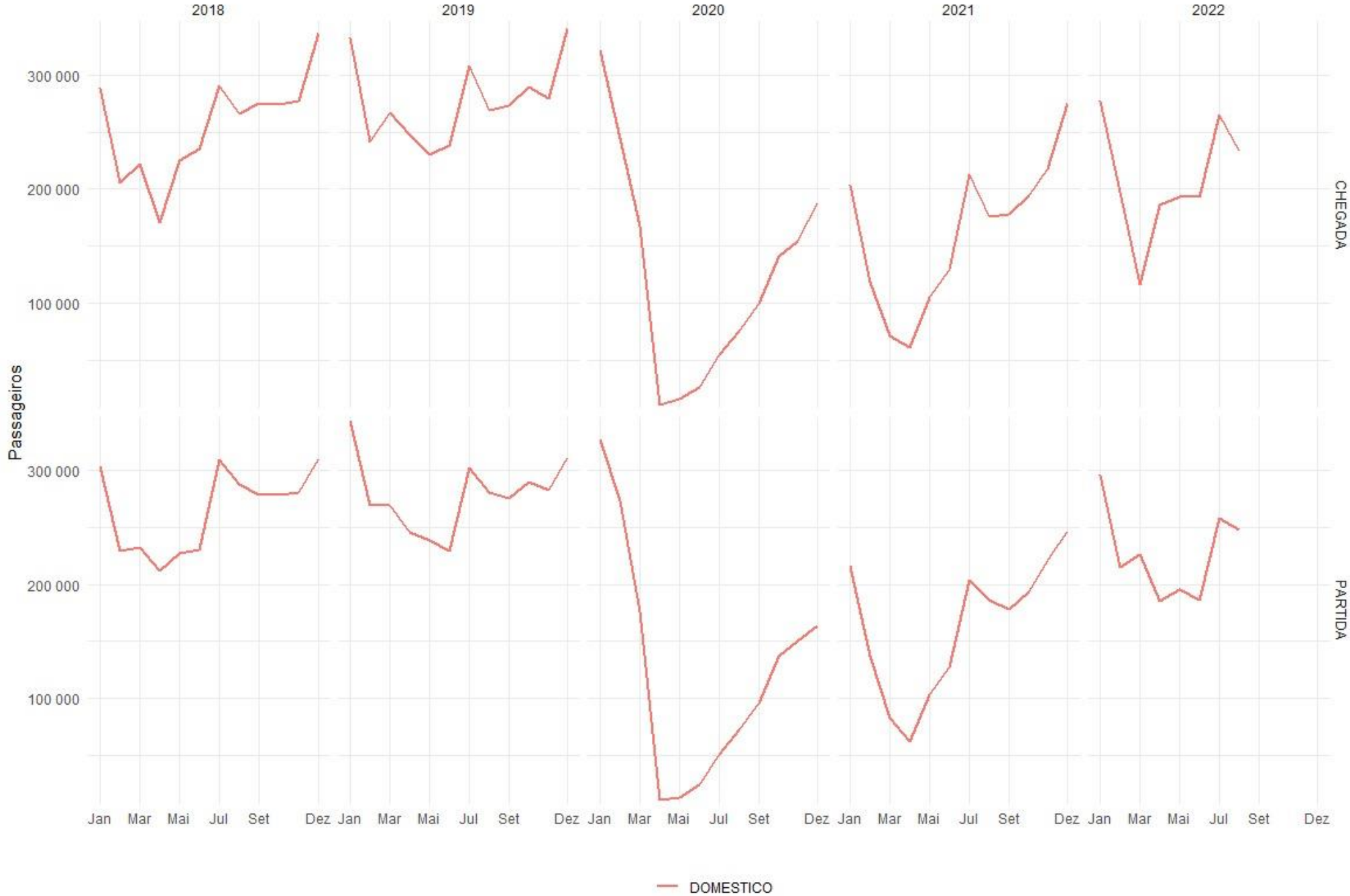


Zona de Interesse	Condomínio	Total de Unidades	Área Média (m²)	Unidades/Lote
05 minutos	Sim	301	88,8	21,5
	Não	70	83,8	1,3
10 minutos	Sim	5.840	88,3	28,8
	Não	3.830	85,3	1,44
15 minutos	Sim	11.527	86,4	31,3
	Não	12.166	85,0	1,27

Mapa de Estados de Destino de Viagens de Ônibus Saindo do Ceará

: ANTT



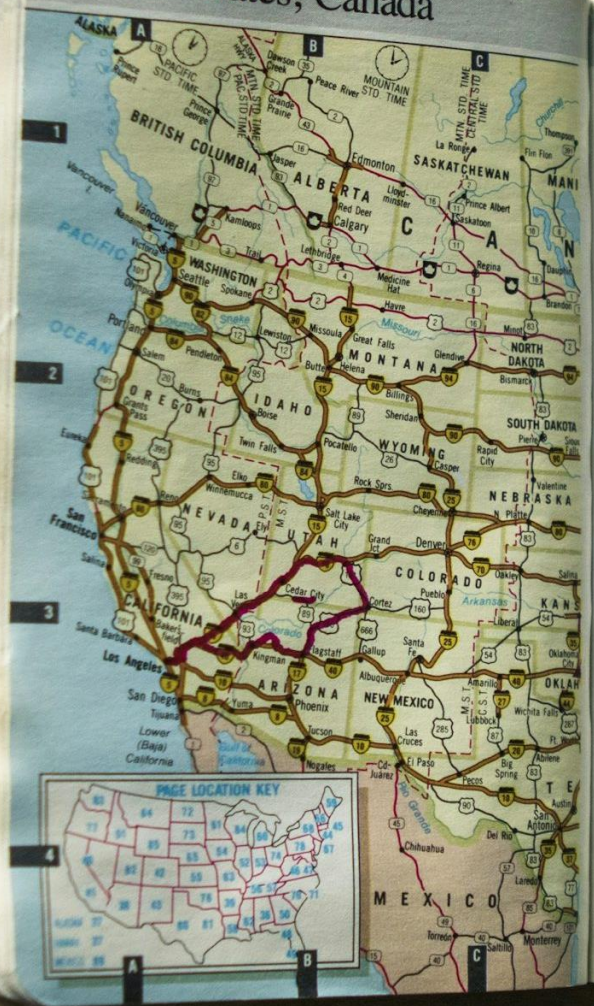

2018
2019
2020
2021
2022
300 000
200 000
100 000
Passageiros
CHEGADA
PARTIDA
Jan
Mar
Mai
Jul
Set
Dez
DOMESTICO

Estudo de caso motivador:

Mudança de carreira!


Analytics
Copie de Toutes les données du site Web
Accueil
Personnalisation
Temps réel
Audience
Acquisition
Vue d'ensemble
Tout le trafic
Google Ads
Search Console
Réseaux sociaux
Campagnes
Comportement
Conversions
Attribution
Découvrir
Administration
Principaux canaux
Organic Search
Social
Direct
Referral
(Other)
Utilisateurs
Conversions
Acquisition
Comportement
MacBook Air

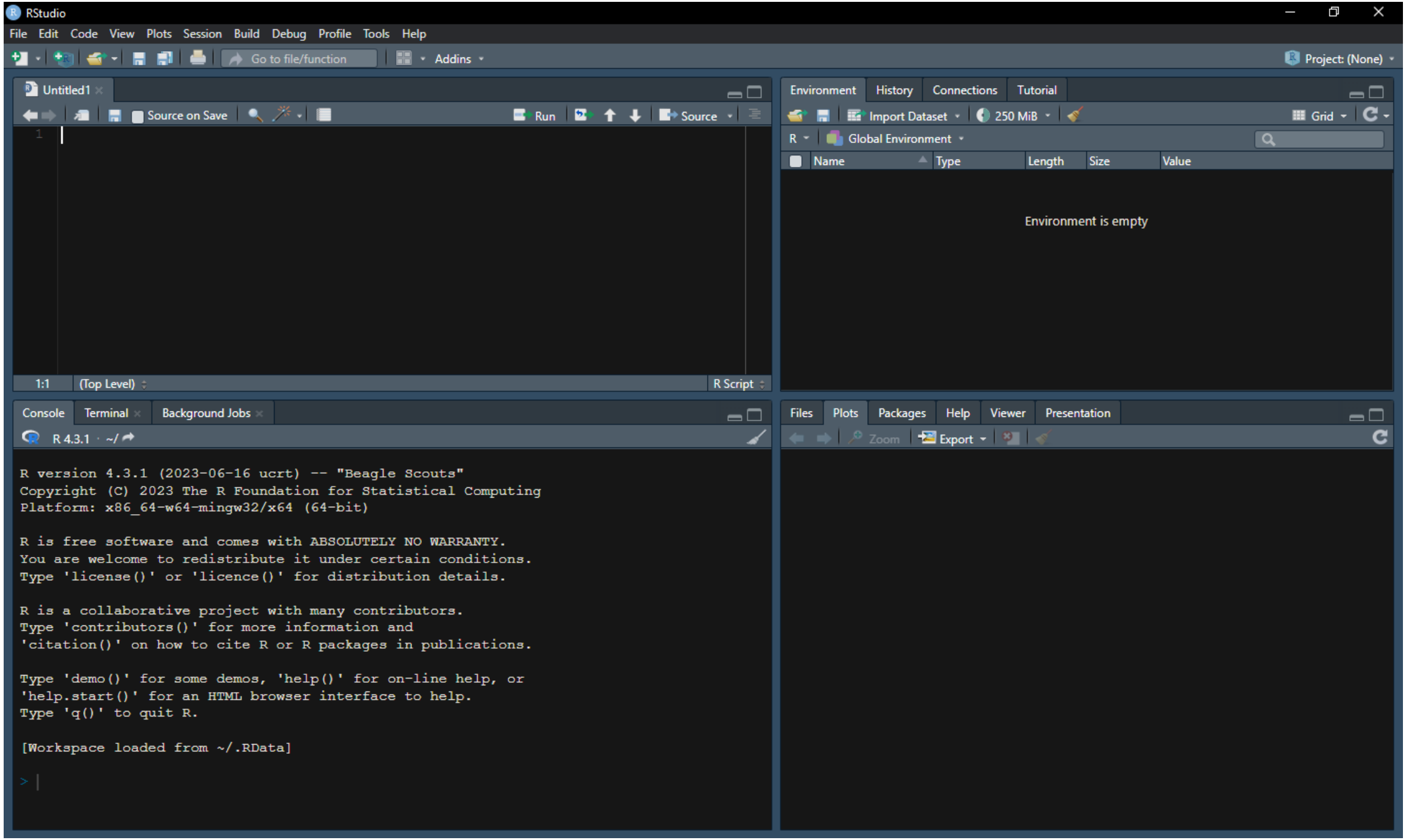
United States, Canada

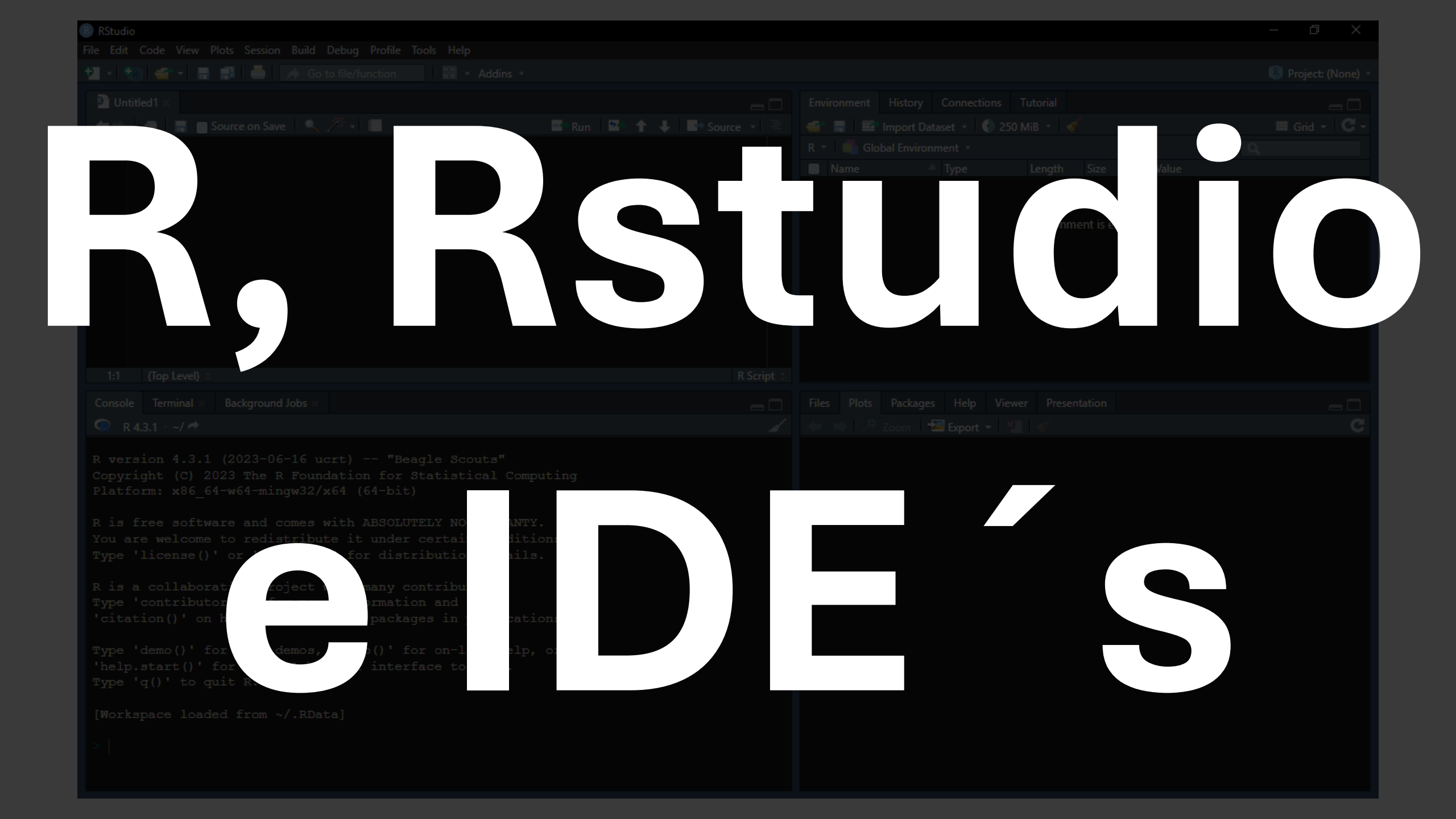


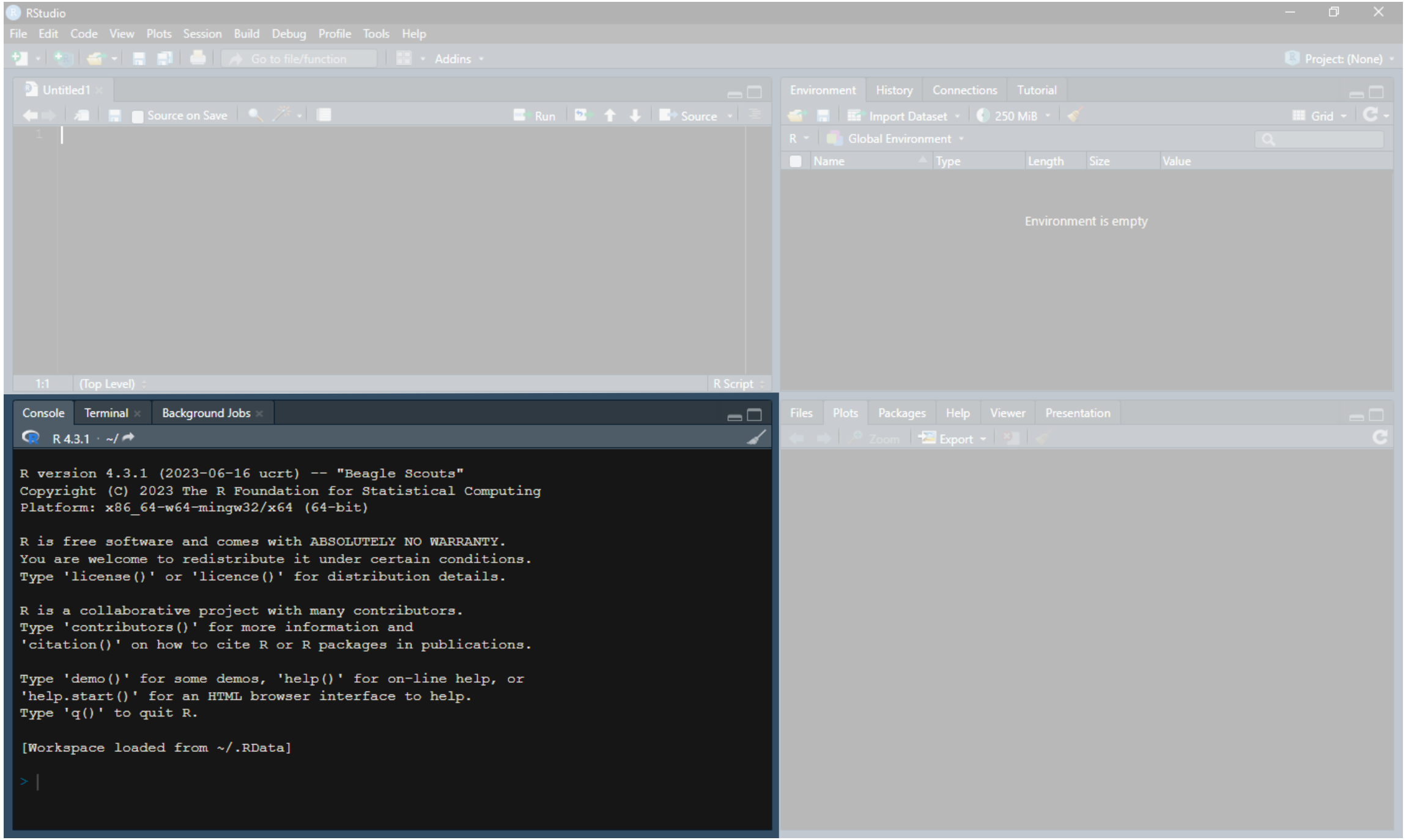

United States, Canada
35
BRITISH COLUMBIA
ALBERTA
SASKATCHEWAN
WASHINGTON
MONTANA
NORTH DAKOTA
MINNESOTA
ONTARIO
QUÉBEC
MAINE
INDIANA
KENTUCKY
ALABAMA
FLORIDA
GULF OF MEXICO
BAHAMAS
MEXICO
PAGE LOCATION KEY
© 1989 HAMMOND INCORPORATED

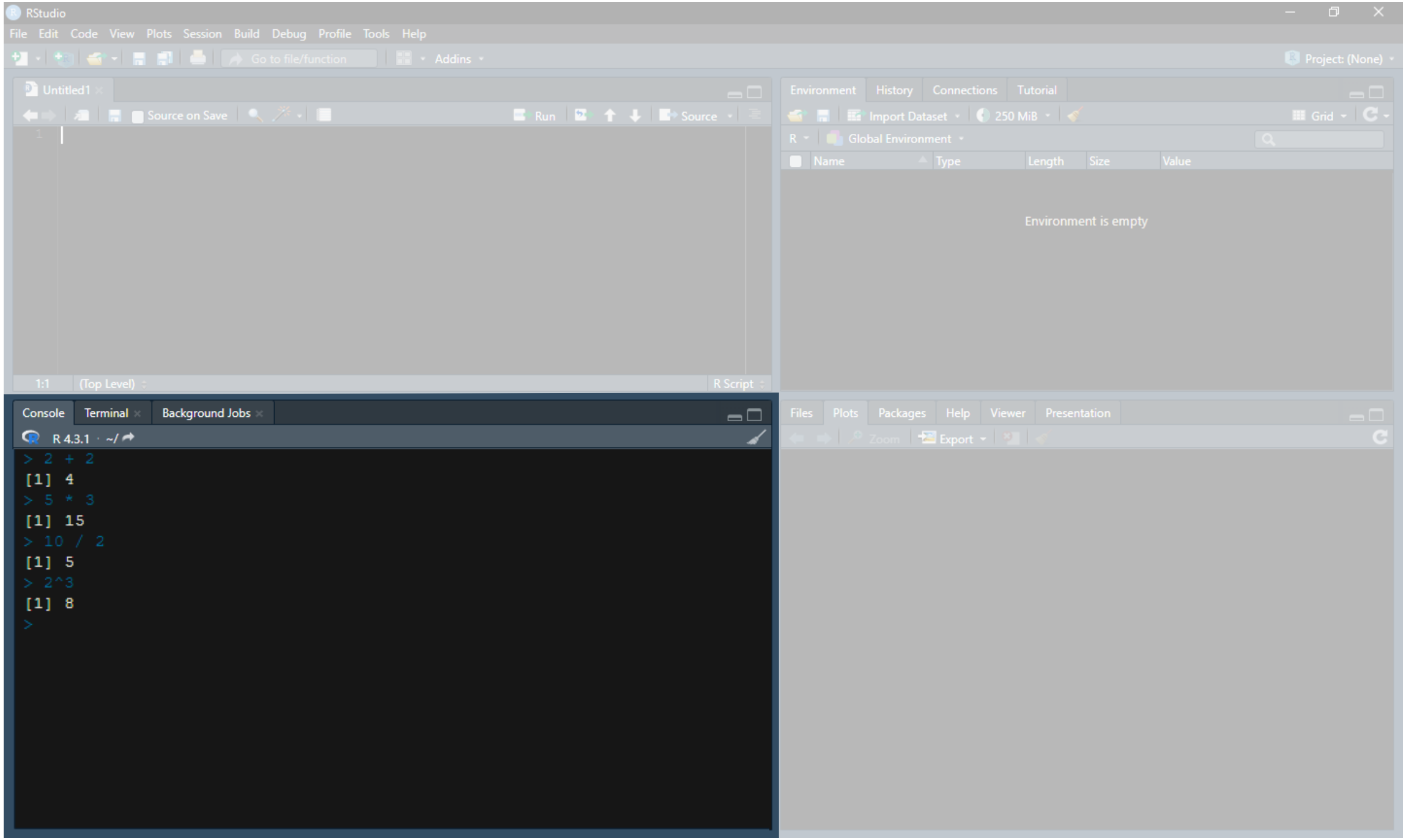
Aonde?

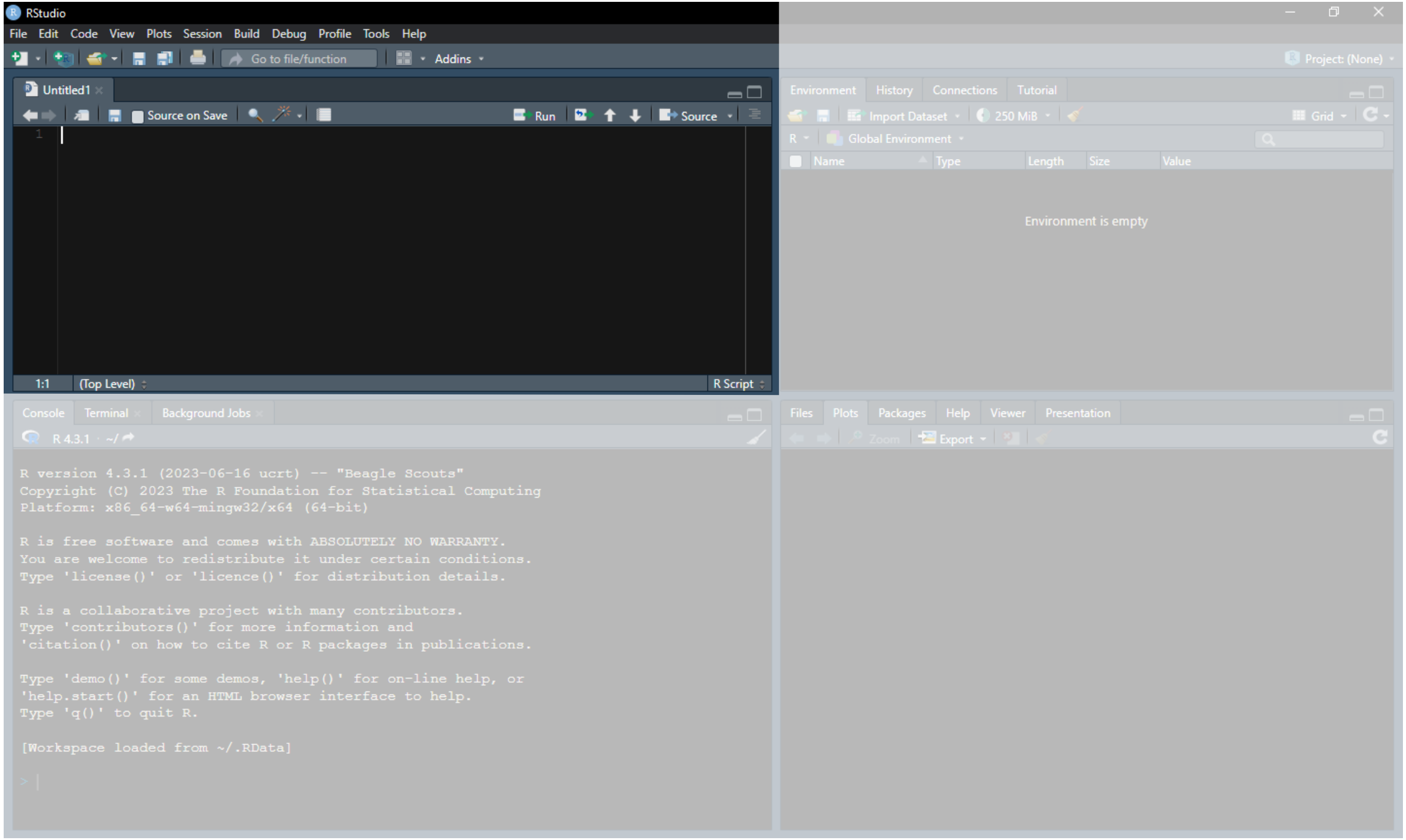
1. Conceitos Básicos

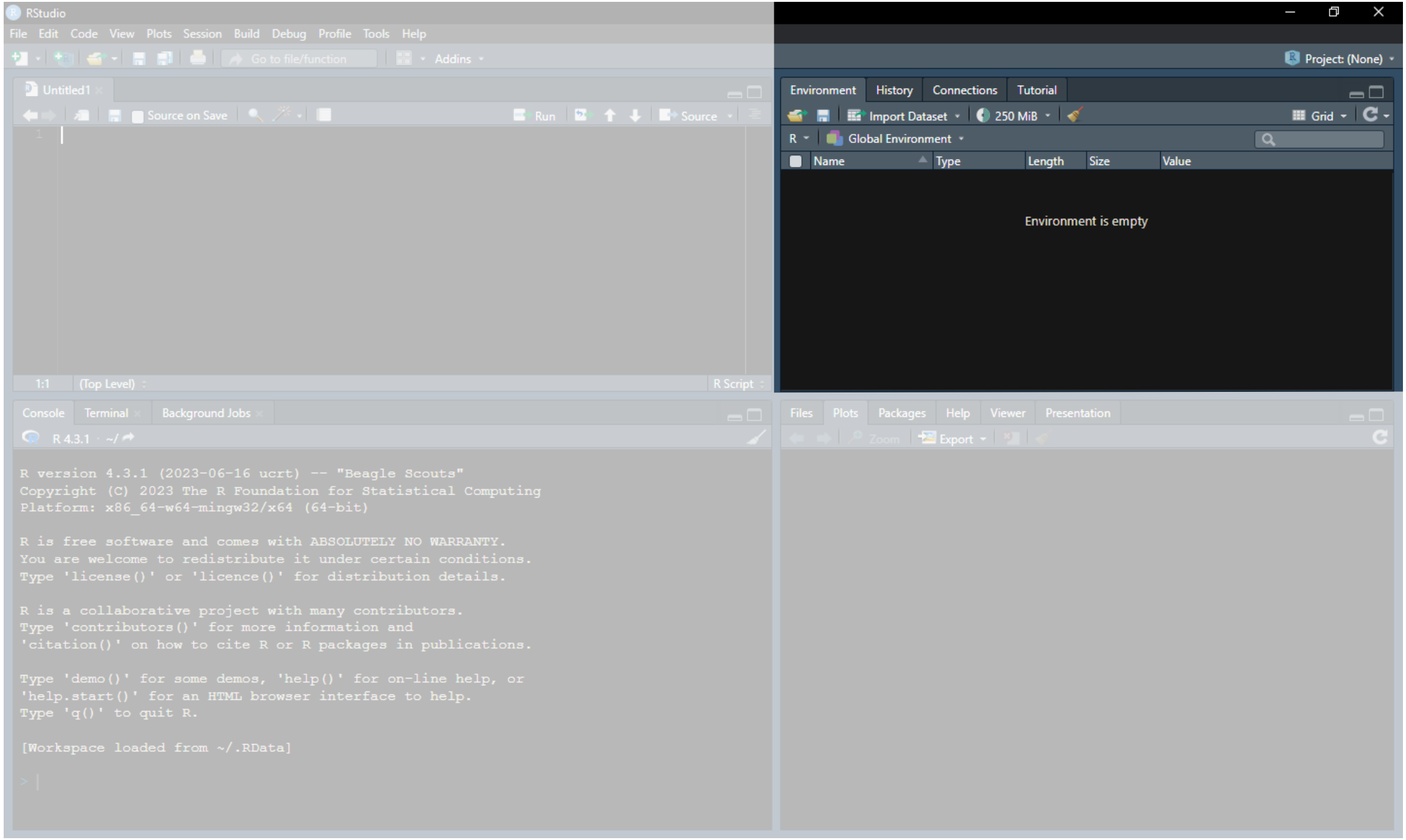

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Environment
History
Connections
Tutorial
Import Dataset
250 MiB
Grid
R
Global Environment
Name
Type
Length
Size
Value
Environment is empty
Console
Terminal
Background Jobs
R 4.3.1 · ~/
R version 4.3.1 (2023-06-16 ucrt) -- "Beagle Scouts"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (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.
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.
[Workspace loaded from ~/.RData]
>
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export

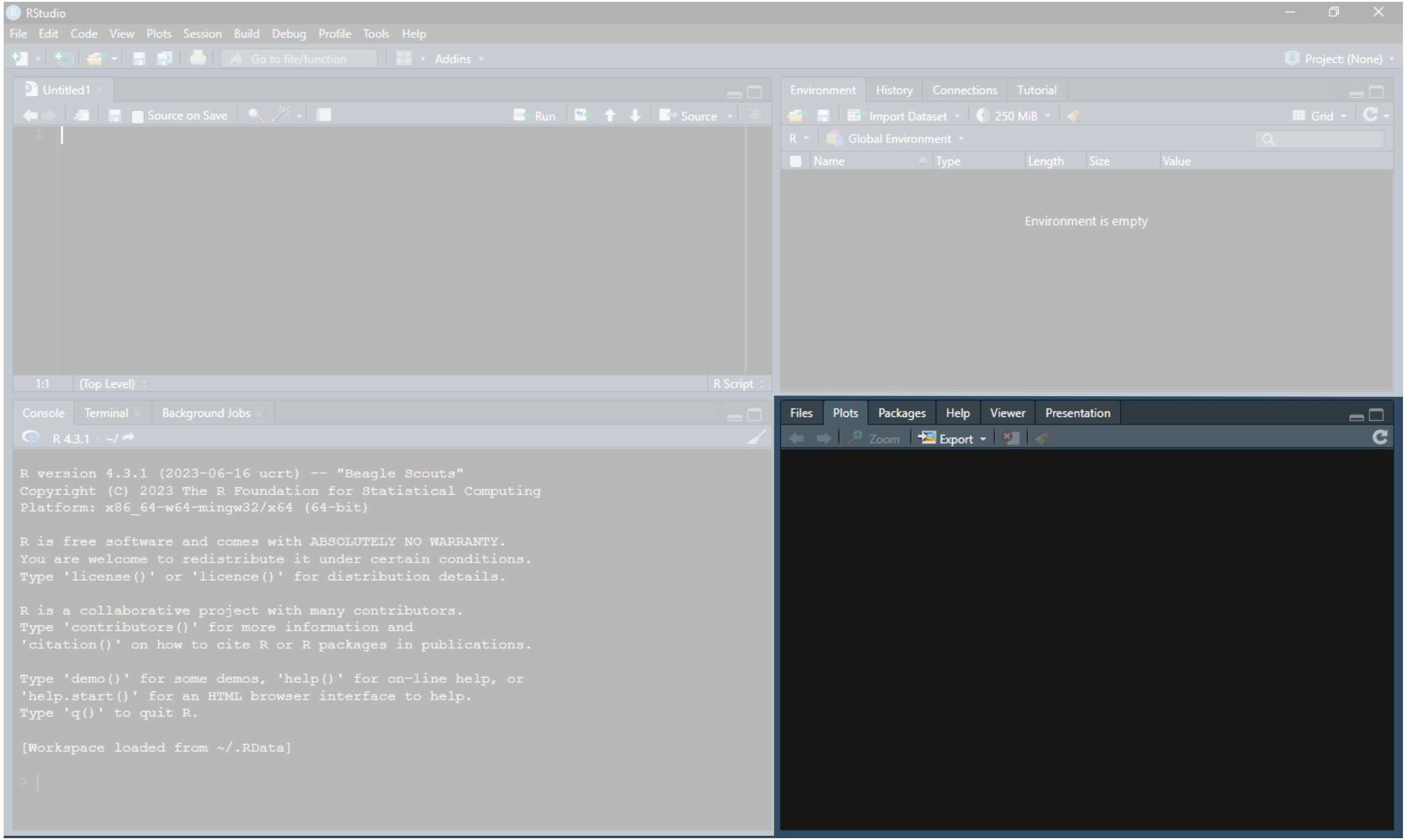

R, Rstudio
eIDE´s

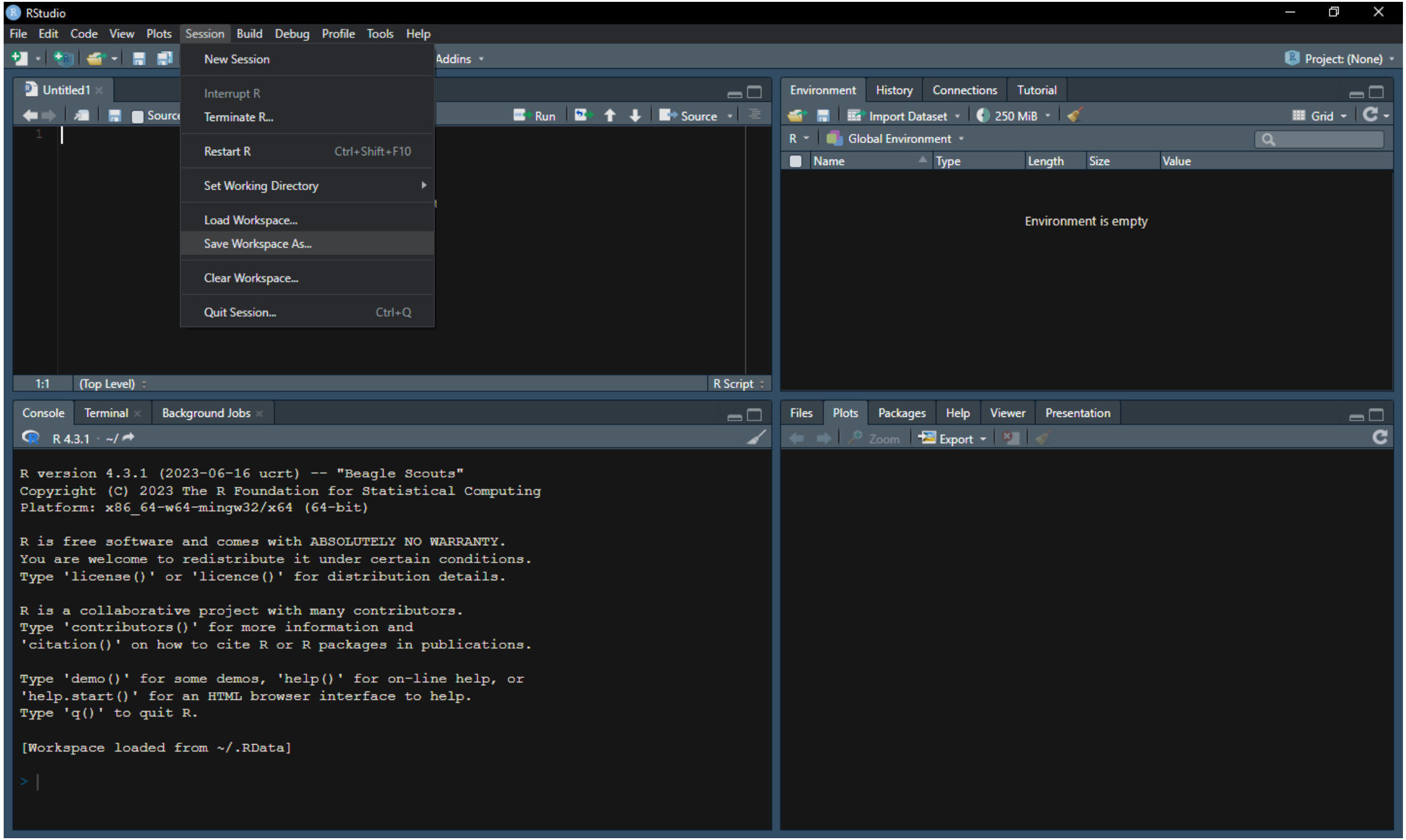

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1:1
(Top Level)
R Script
Environment
History
Connections
Tutorial
Import Dataset
250 MiB
Grid
R
Global Environment
Name
Type
Length
Size
Value
Environment is empty
Console
Terminal
Background Jobs
R 4.3.1 · ~/
R version 4.3.1 (2023-06-16 ucrt) -- "Beagle Scouts"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (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.
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.
[Workspace loaded from ~/.RData]
>
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export

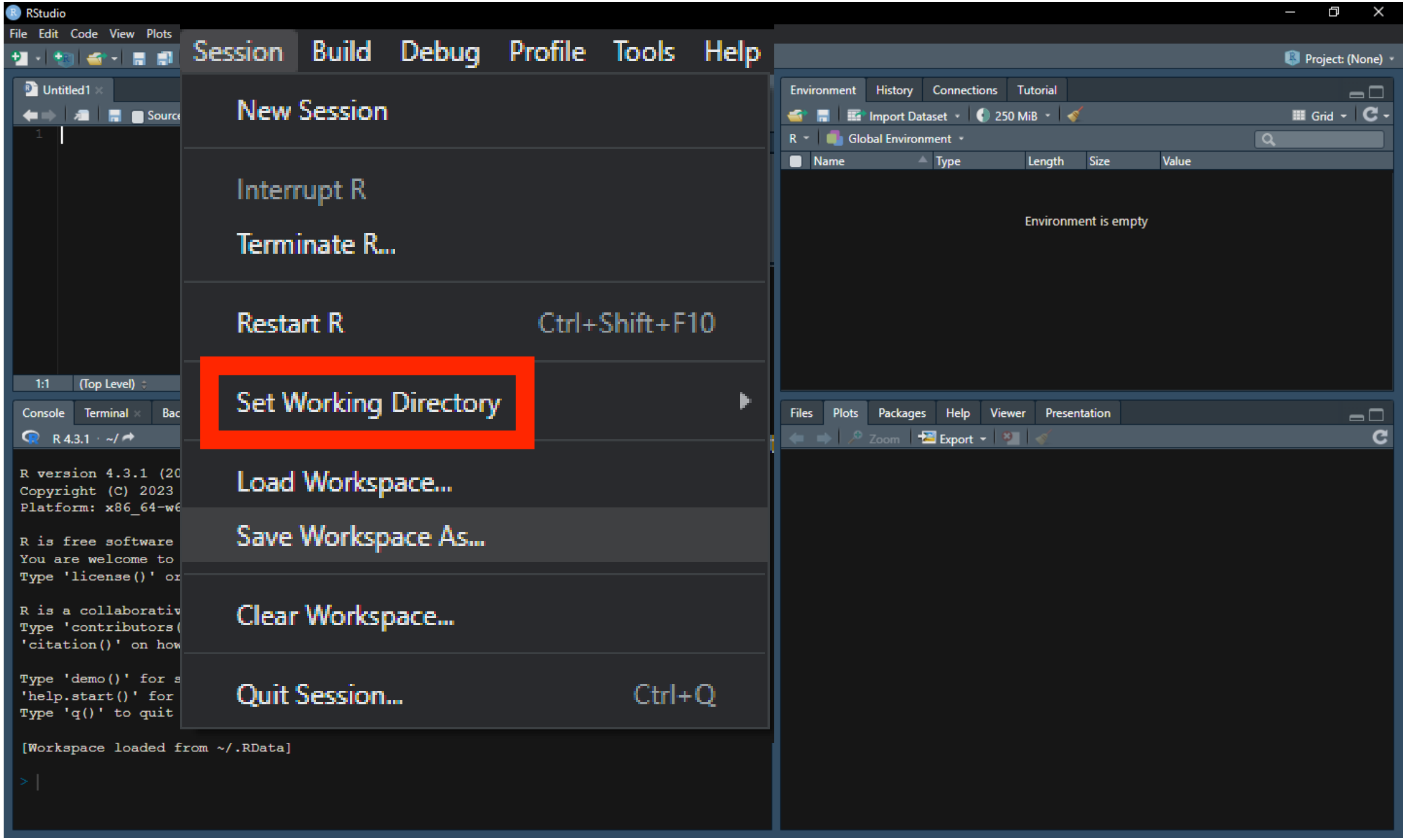

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1:1
(Top Level)
R Script
Environment
History
Connections
Tutorial
Import Dataset
250 MiB
Grid
R
Global Environment
Name
Type
Length
Size
Value
Environment is empty
Console
Terminal
Background Jobs
R 4.3.1 · ~/
> 2 + 2
[1] 4
> 5 * 3
[1] 15
> 10 / 2
[1] 5
> 2^3
[1] 8
>
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Environment History Connections Tutorial
Import Dataset
250 MiB
Grid
R
Global Environment
Name Type Length Size Value
Environment is empty
Console Terminal Background Jobs
R 4.3.1 · ~/
R version 4.3.1 (2023-06-16 ucrt) -- "Beagle Scouts"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (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.
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.
[Workspace loaded from ~/.RData]
>
Files Plots Packages Help Viewer Presentation
Zoom
Export


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1:1
(Top Level)
R Script
Environment
History
Connections
Tutorial
Import Dataset
250 MiB
Grid
R
Global Environment
Name
Type
Length
Size
Value
Environment is empty
Console
Terminal
Background Jobs
R 4.3.1 · ~/
R version 4.3.1 (2023-06-16 ucrt) -- "Beagle Scouts"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (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.
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.
[Workspace loaded from ~/.RData]
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1:1
(Top Level)
R Script
Environment
History
Connections
Tutorial
Import Dataset
250 MiB
Grid
R
Global Environment
Name
Type
Length
Size
Value
Environment is empty
Console
Terminal
Background Jobs
R 4.3.1 · ~/
R version 4.3.1 (2023-06-16 ucrt) -- "Beagle Scouts"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (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.
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.
[Workspace loaded from ~/.RData]
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
New Session
Interrupt R
Terminate R...
Restart R Ctrl+Shift+F10
Set Working Directory
Load Workspace...
Save Workspace As...
Clear Workspace...
Quit Session... Ctrl+Q
Addins
Project: (None)
Untitled1
Run
Source
1:1
(Top Level)
R Script
Environment History Connections Tutorial
Import Dataset
250 MiB
Grid
R
Global Environment
Name Type Length Size Value
Environment is empty
Console Terminal Background Jobs
R 4.3.1 · ~/
R version 4.3.1 (2023-06-16 ucrt) -- "Beagle Scouts"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (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.
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.
[Workspace loaded from ~/.RData]
>
Files Plots Packages Help Viewer Presentation
Zoom
Export


RStudio
File Edit Code View Plots
Session Build Debug Profile Tools Help
New Session
Interrupt R
Terminate R...
Restart R Ctrl+Shift+F10
Set Working Directory
Load Workspace...
Save Workspace As...
Clear Workspace...
Quit Session... Ctrl+Q
Project: (None)
Environment History Connections Tutorial
Import Dataset 250 MiB Grid
Global Environment
Name Type Length Size Value
Environment is empty
Untitled1
1:1 (Top Level)
Console Terminal
R 4.3.1 · ~/
R version 4.3.1 (20
Copyright (C) 2023
Platform: x86_64-w6
R is free software
You are welcome to
Type 'license()' or
R is a collaborativ
Type 'contributors(
'citation()' on how
Type 'demo()' for s
'help.start()' for
Type 'q()' to quit
[Workspace loaded from ~/.RData]
Files Plots Packages Help Viewer Presentation
Zoom Export

```
read.csv(file = 'data.csv')
```

```
> read.csv('data.csv')
                   state abb          region population total
1                Alabama  AL           South    4779736   135
2                 Alaska  AK            West     710231    19
3                Arizona  AZ            West    6392017   232
4               Arkansas  AR           South    2915918    93
5             California  CA            West   37253956  1257
6               Colorado  CO            West    5029196    65
7            Connecticut  CT       Northeast    3574097    97
8               Delaware  DE           South     897934    38
9  District of Columbia  DC           South     601723    99
10               Florida  FL           South   19687653   669
11               Georgia  GA           South    9920000   376
12                Hawaii  HI            West    1360301     7
13                 Idaho  ID            West    1567582    12
14              Illinois  IL North Central   12830632   364
15               Indiana  IN North Central    6483802   142
```

Tipos de banco de dados

```
read.csv('data.csv’)

read.csv(path = 'data.csv’)

?read.csv()
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Files Plots Packages Help Viewer Presentation

R: Data Input | Find in Topic

read.table {utils} R Documentation

Data Input

Description

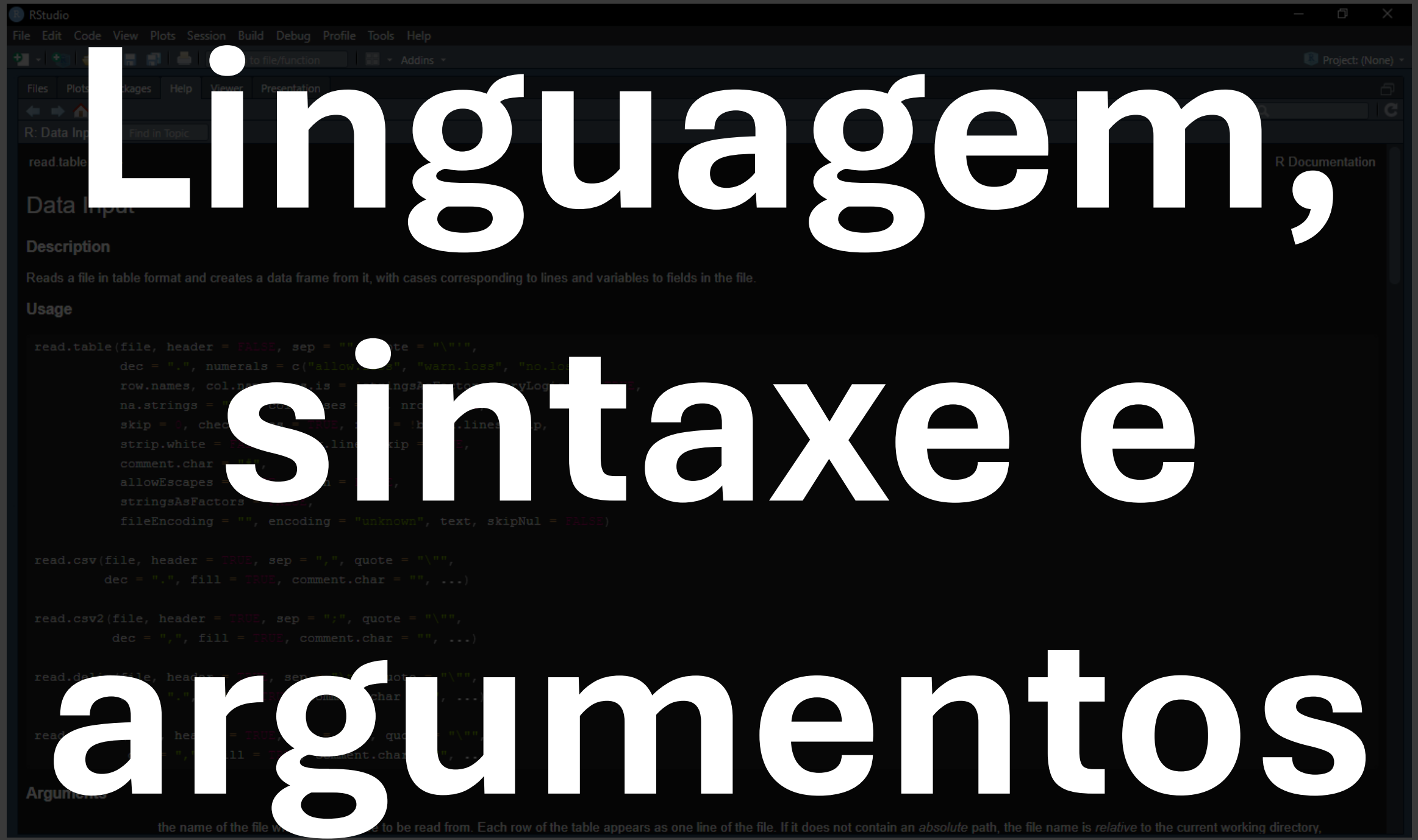
Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

Usage

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
           row.names, col.names, as.is = !stringsAsFactors, tryLogical = TRUE,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = FALSE,
           fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)

read.csv(file, header = TRUE, sep = ",", quote = "\"",
         dec = ".", fill = TRUE, comment.char = "", ...)

read.csv2(file, header = TRUE, sep = ";", quote = "\"",
          dec = ",", fill = TRUE, comment.char = "", ...)

read.delim(file, header = TRUE, sep = "\t", quote = "\"",
           dec = ".", fill = TRUE, comment.char = "", ...)

read.delim2(file, header = TRUE, sep = "\t", quote = "\"",
            dec = ",", fill = TRUE, comment.char = "", ...)
```

Arguments

the name of the file which the data are to be read from. Each row of the table appears as one line of the file. If it does not contain an *absolute* path, the file name is *relative* to the current working directory,



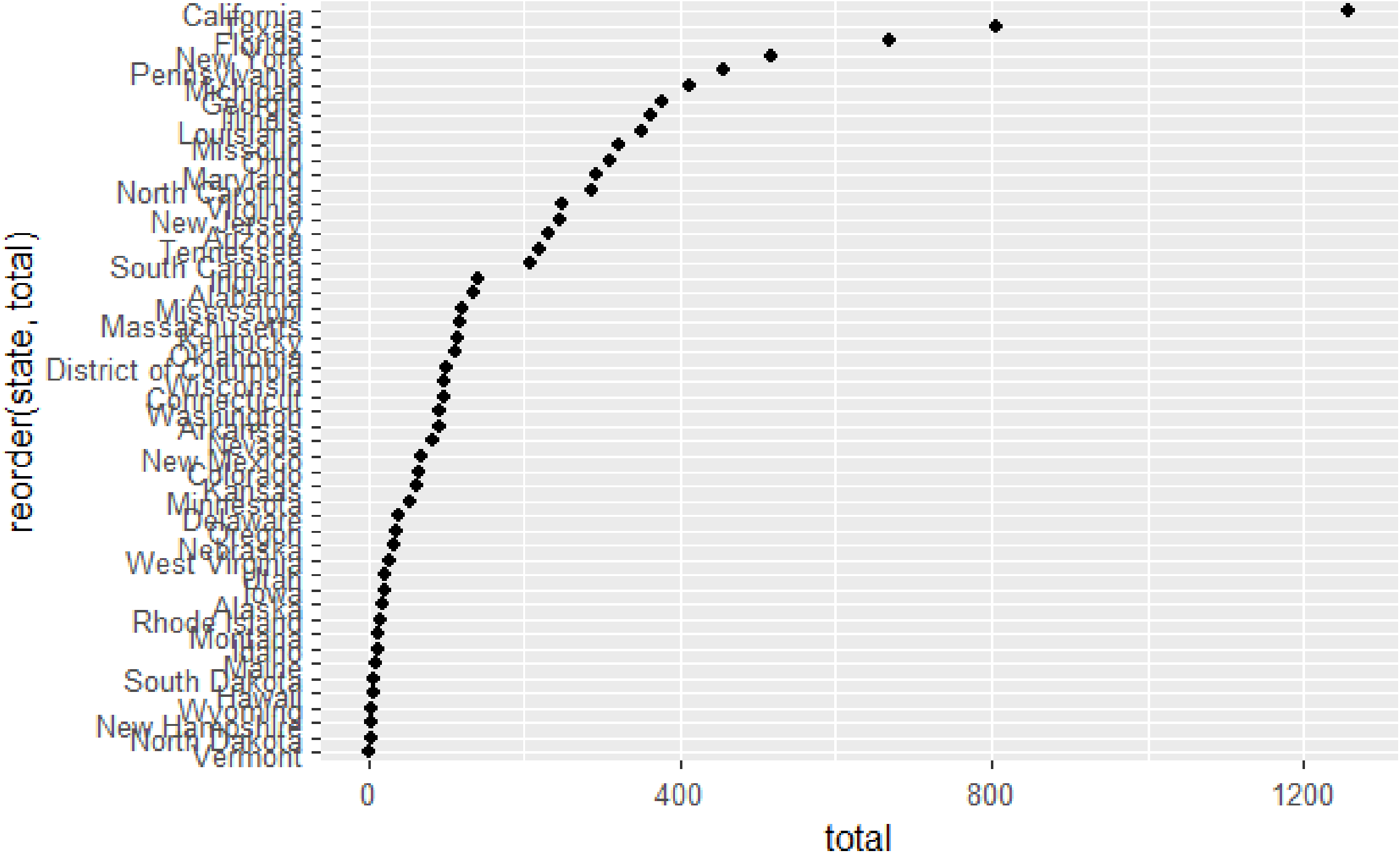
Linguagem, sintaxe e argumentos

funcao(arg_1 = var_1, arg_2 = var_2, ...)

```
install.packages('tidyverse')

library(tidyverse)
```


Pacotes

```
read.csv('data.csv’) %>%
        ggplot() +
        geom_point(aes(total, reorder(state,
total)))
```


reorder(state, total)
California
Texas
Florida
New York
Pennsylvania
Michigan
Georgia
Illinois
Louisiana
Missouri
Ohio
Maryland
North Carolina
Virginia
New Jersey
Arizona
Tennessee
South Carolina
Indiana
Alabama
Mississippi
Massachusetts
Kentucky
Oklahoma
District of Columbia
Wisconsin
Connecticut
Washington
Arkansas
Kansas
Nevada
New Mexico
Colorado
Minnesota
Delaware
Oregon
Nebraska
West Virginia
Utah
Iowa
Alaska
Rhode Island
Montana
Idaho
Maine
South Dakota
Hawaii
Wyoming
New Hampshire
North Dakota
Vermont
0
400
800
1200
total

2. Criando, carregando e inspecionando dados

```
a <- 2
a
print(a)
class(a)
str(a)
a = 2
a == 2
a == a
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

aula 01 - introducao.R

Source on Save | Run | Source

```
31  # plotar um grafico
32  read.csv(file = 'data.csv') |>
33    ggplot() +
34    geom_point(aes(x = total, y = state))
35
36  # plotar um grafico ordenando a variavel y em relação a variável x
37  read.csv(file = 'data.csv') |>
38    ggplot() +
39    geom_point(aes(x = total, y = reorder(state, total)))
40
41  # carregando e manipulando dados ----------------------------------------
42  # salvando objetos e realizando operações parte 1
43  a <- 2
44  a
45  print(a)
46  class(a)
47  str(a)
48
49  a = 2
50  a == 2
51  a == a
52
53
```

52:1 | carregando e manipulando dados | R Script

Console | Terminal | Background Jobs

R 4.3.1 · C:/Users/NELSON QUESADO/Downloads/

```
+   geom_point(aes(x = total, y = reorder(state, total)))
> # carregando e manipulando dados ----------------------------------------
> # salvando objetos e realizando operações parte 1
> a <- 2
> a
[1] 2
> print(a)
[1] 2
> class(a)
[1] "numeric"
> str(a)
 num 2
> a = 2
> a == 2
[1] TRUE
> a == a
[1] TRUE
>
```

Environment | History | Connections | Tutorial

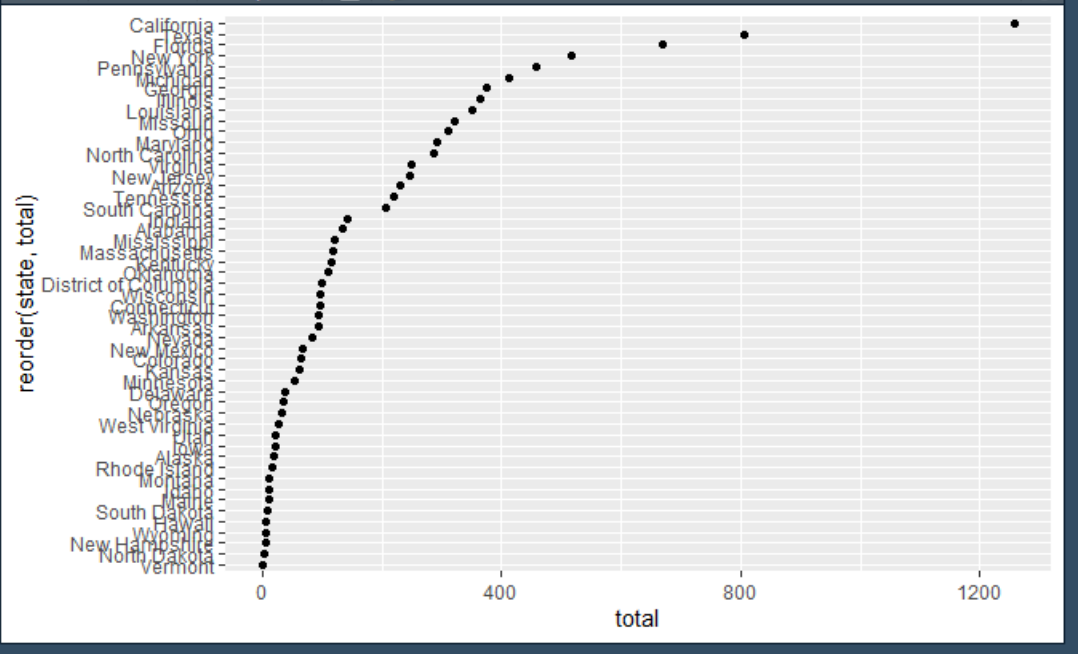
Import Dataset | 239 MiB | Grid

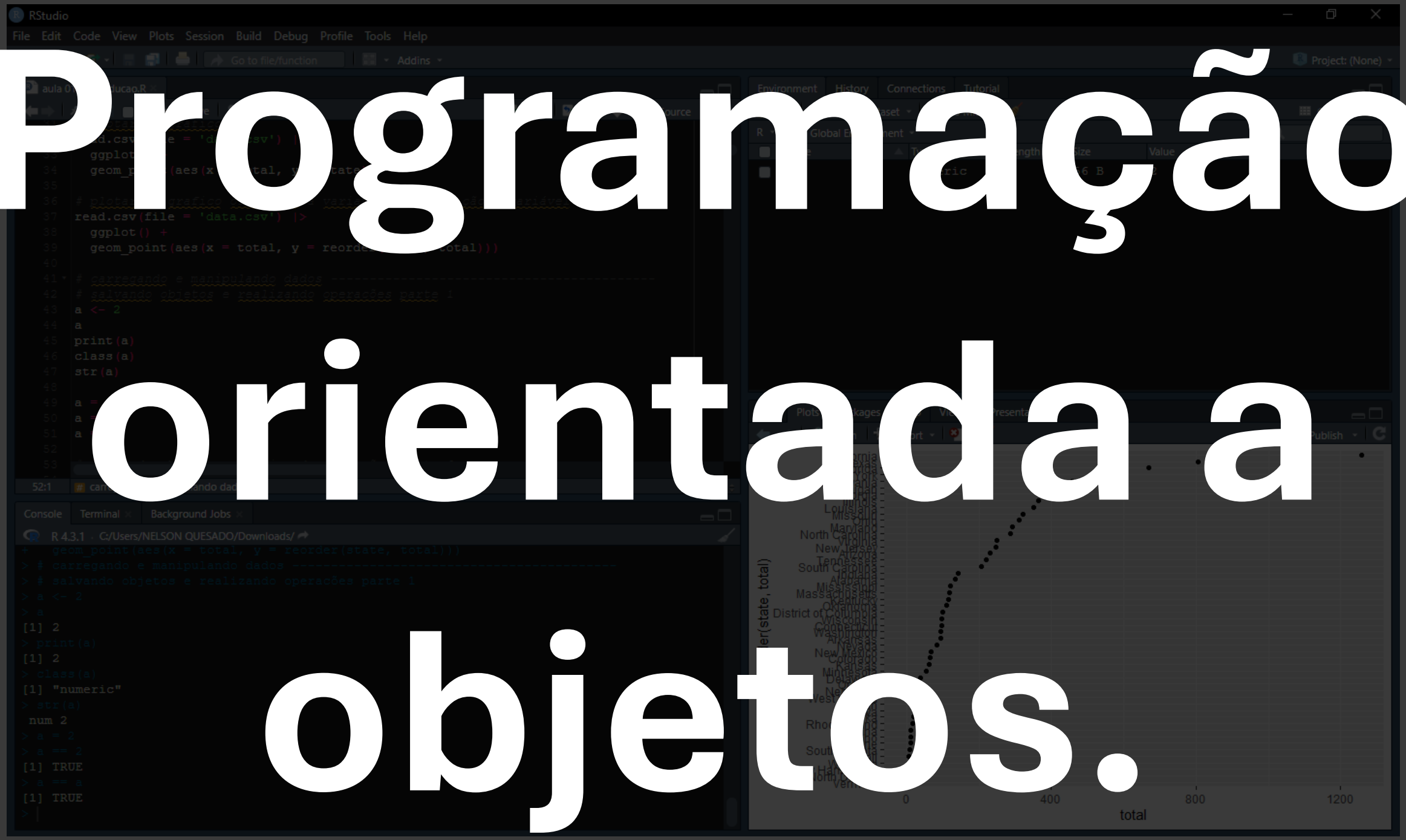
R | Global Environment

Name	Type	Length	Size	Value
a	numeric	1	56 B	2

Files | Plots | Packages | Help | Viewer | Presentation

Zoom | Export | Publish





Programação orientada a objetos.

```
Objeto <- funcao(arg_1 = var_1)
```

```
b <- 5
c <- -4
a == b
c == -4
class(a) == class(b)
a + b * c
(a + b) * c
```

```
d <- (a + b) * c
d
class(d)
A <- '2’
print(A)
class(A)
str(A)
```

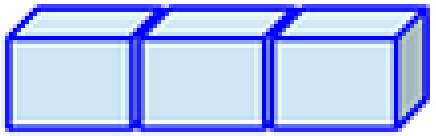
```
class(A) == class(a)

class(a) <- 'character’

class(A) == class(a)

class(a) <- 'numeric’

class(A) == class(a)

ls()

rm(A, b, c, d)
```

```
c(1, 2, 3, 4)
1:4
c(1, 2, 3, 4) == 1:4
x <- 1:4
x
print(x)
class(x)
str(x)
```

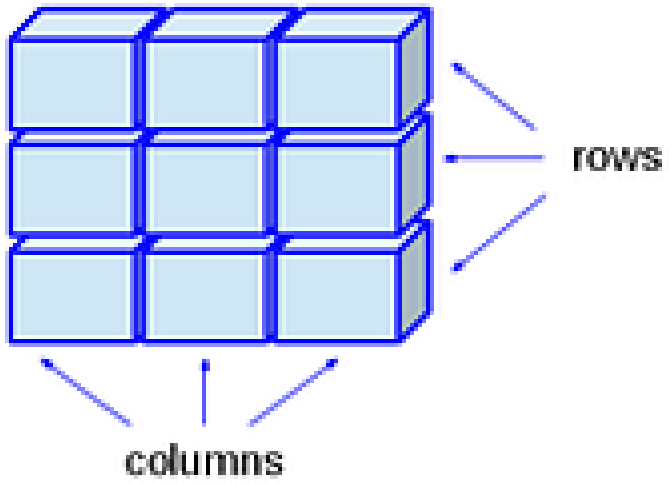
```
x == a
y <- x == a
y
class(y)
as.numeric(y)
sum(as.numeric(y))
length(y)
sum(as.numeric(y)) / length(y)
```

```
x <- c("a", "b", "c", "d", "e")

y <- c("a", "d", "f")

y %in% x
```

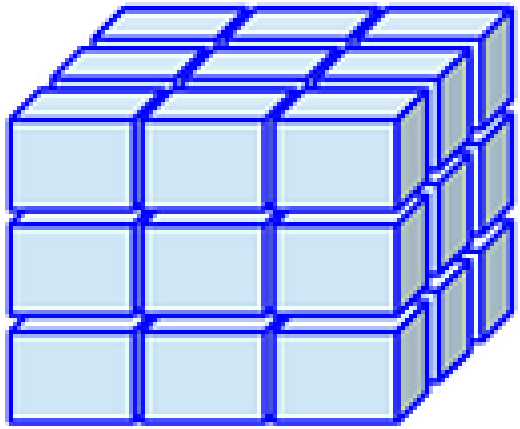
Vector



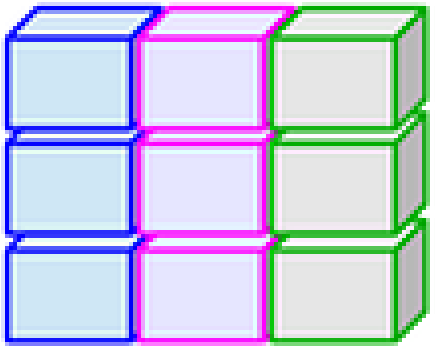
Matrix



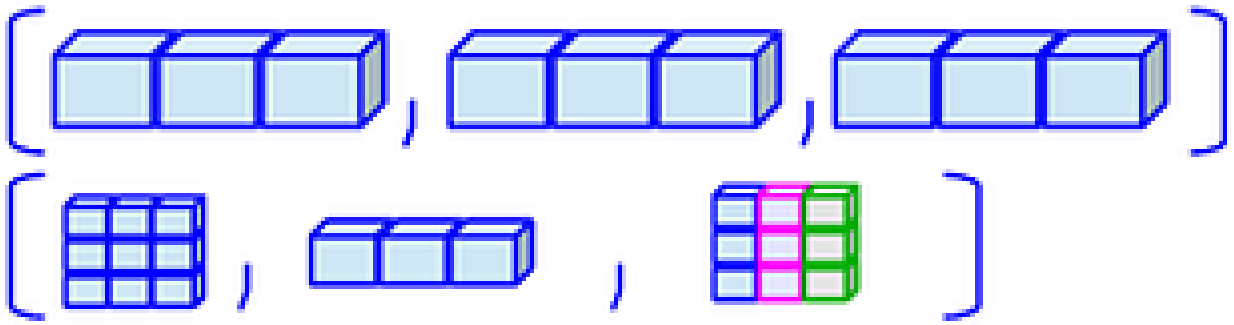
Array



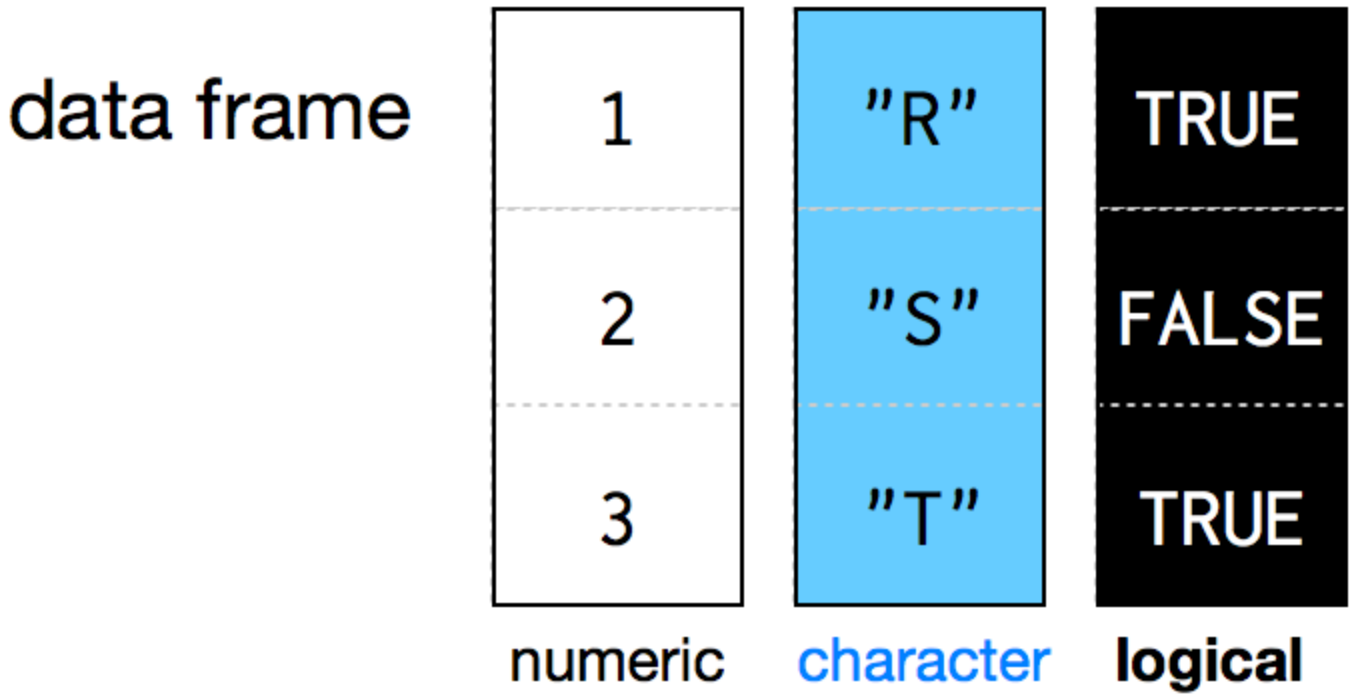
Data Frame
(Table)



Lists



```
data <- read.csv('data.csv’)
data
print(data)
head(data)
names(data)
class(data)
str(data)
summary(data)
c("Boston", "Dakota", "Washington") %in% data$state
```


data frame
1
2
3
"R"
"S"
"T"
TRUE
FALSE
TRUE
numeric
character
logical


United States, Canada
35
BRITISH COLUMBIA
ALBERTA
SASKATCHEWAN
WASHINGTON
MONTANA
NORTH DAKOTA
MINNESOTA
ONTARIO
QUÉBEC
MAINE
INDIANA
KENTUCKY
ALABAMA
FLORIDA
GULF OF MEXICO
BAHAMAS
MEXICO
PAGE LOCATION KEY
© 1989 HAMMOND INCORPORATED

Aonde?

3. Manipulando dados

```
ls()
rm(list = ls())
tibble(data)
data <- tibble(data)
data
data$state
pull(data, state)
data[, 1]
data$state == pull(data, state)
data[, 1] == pull(data, state)
```

```
identical(data$state, pull(data, state))
identical(data[, 1], pull(data, state))
data$state %>% class
pull(data, state) %>% class
data[, 1] %>% class
data[1, ]
data[1, 1]
data[c(1, 2, 3), ]
data[, 2:4]
```

```
data$state %>% class
pull(data, state) %>% class
data[, 1] %>% class

data[1, ]
data[1, 1]
data[c(1, 2, 3), ]
data[, 2:4]
```

%>%

```
object %>% function()
function(object)

object %>% function(arg2)
function(object, arg2)

function(arg1, object)
object %>% function(arg1, .)
```

```
data$region
table(data$region)
data %>% mutate(region = as_factor(region))
data %>% summary
data %>% str
data <- data %>%
    mutate(region = as_factor(region))
data %>% summary
data %>% str
```

```
data %>%
    mutate(dens = total/population)

data %>%
    mutate(dens = total/population * 100000)

data %>%
    mutate(dens = total/population * 100000,
           seguro = dens < 1) %>%
    filter(seguro == TRUE) %>%
    arrange(dens)
```

```
data %>%
    mutate(dens = total/population * 100000,
        seguro = dens < 1) %>%
    arrange(dens) %>%
    select(state, dens, seguro)

data %>%
    reframe(population = sum(population),
    total = sum(total))
```

```
data %>%
    group_by(region) %>%
    reframe(n = n(),
            pop.media = mean(population),
            tot.population = sum(population),
            tot.total = sum(total)) %>%
    mutate(dens = tot.total/tot.population*100000,
            seguro = if_else(dens < 3, 'seguro', 'inseguro'))
```

4. Dataviz

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812 ~1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. ___ Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow en ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

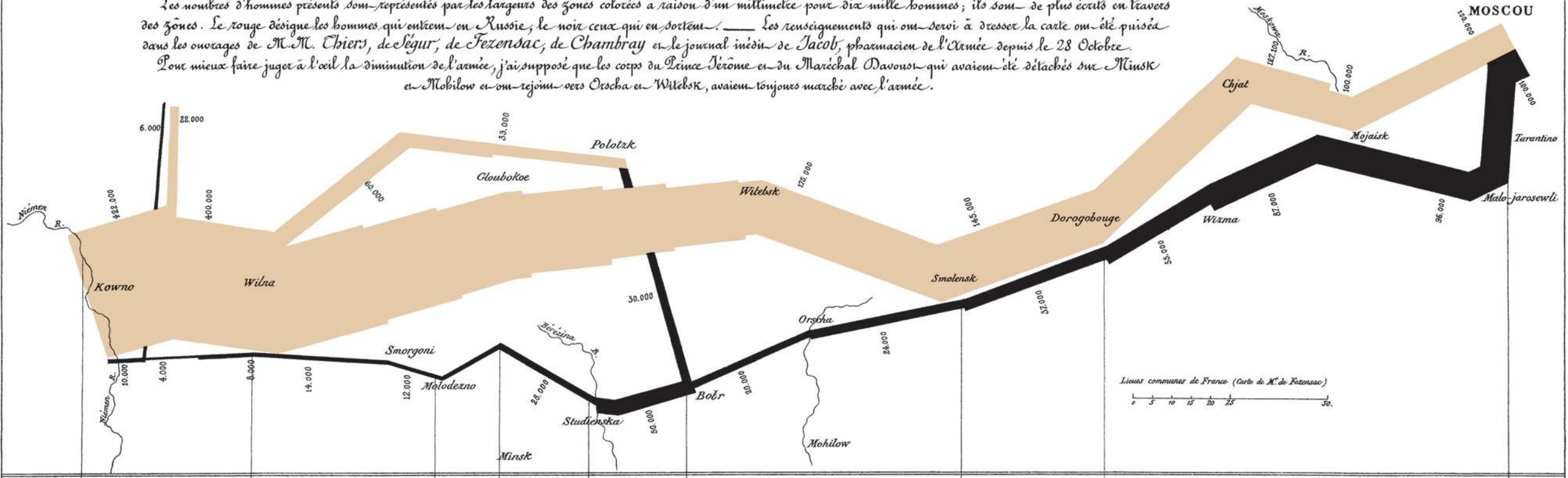
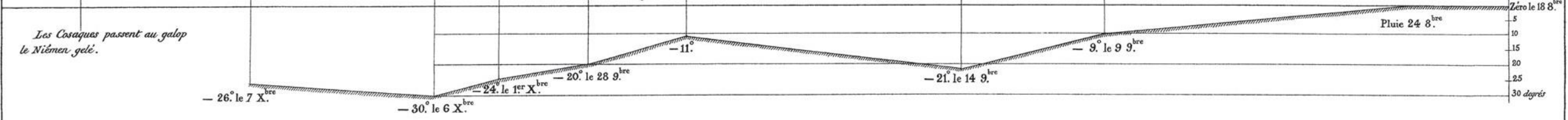


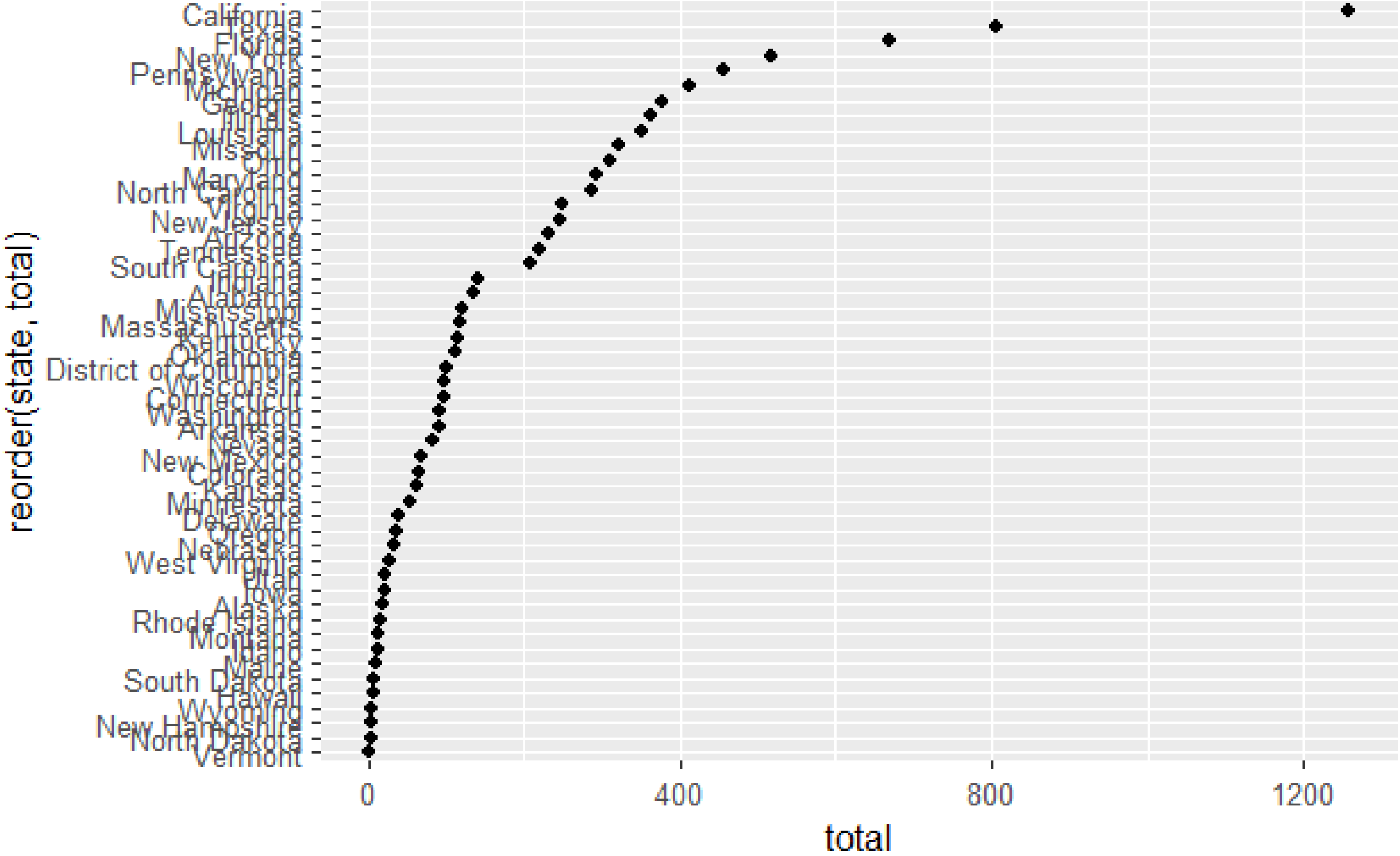
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.

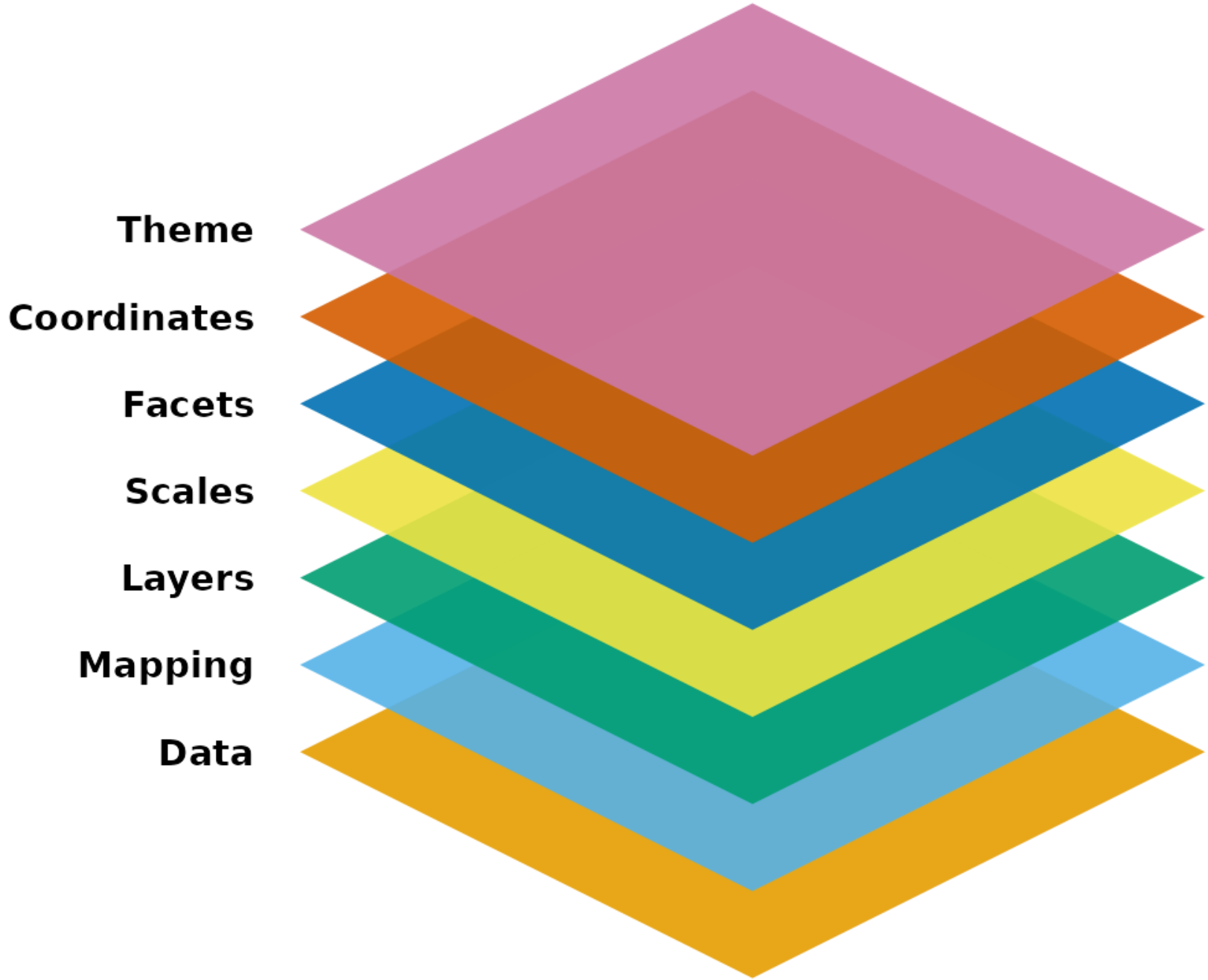
Imp. Lith. Regnier et Dourdet.

```
data %>%
        ggplot() +
        geom_point(aes(total, reorder(state,
total)))
```


reorder(state, total)
California
Texas
Florida
New York
Pennsylvania
Michigan
Georgia
Illinois
Louisiana
Missouri
Ohio
Maryland
North Carolina
Virginia
New Jersey
Arizona
Tennessee
South Carolina
Indiana
Alabama
Mississippi
Massachusetts
Kentucky
Oklahoma
District of Columbia
Wisconsin
Connecticut
Washington
Arkansas
Kansas
Nevada
New Mexico
Colorado
Minnesota
Delaware
Oregon
Nebraska
West Virginia
Utah
Iowa
Alaska
Rhode Island
Montana
Idaho
Maine
South Dakota
Hawaii
Wyoming
New Hampshire
North Dakota
Vermont
0
400
800
1200
total

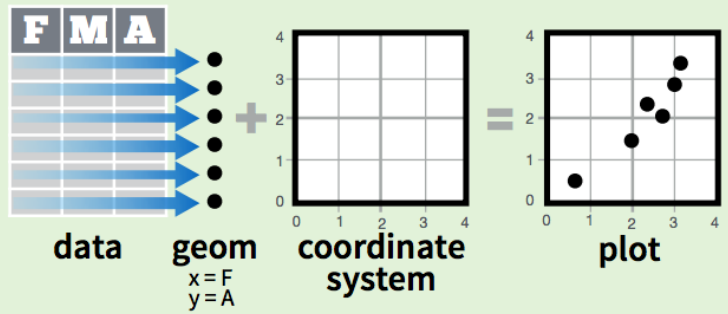
O **ggplot2** é um pacote R que usa uma estrutura conceitual baseada na gramática de gráficos para produzir plots.

Isso permite que você “descreva” um gráfico a partir de elementos combináveis.


Theme
Coordinates
Facets
Scales
Layers
Mapping
Data

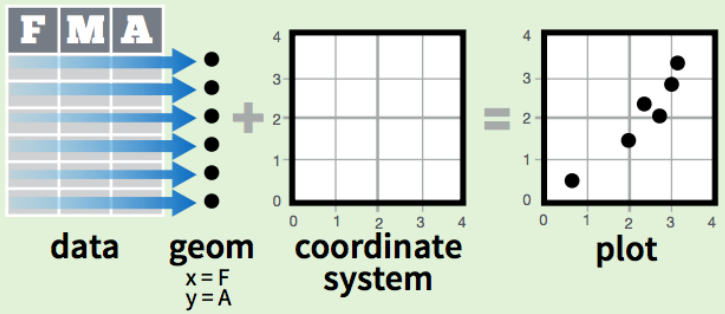
Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same few components: a **data** set, a set of **geoms**—visual marks that represent data points, and a **coordinate system.**

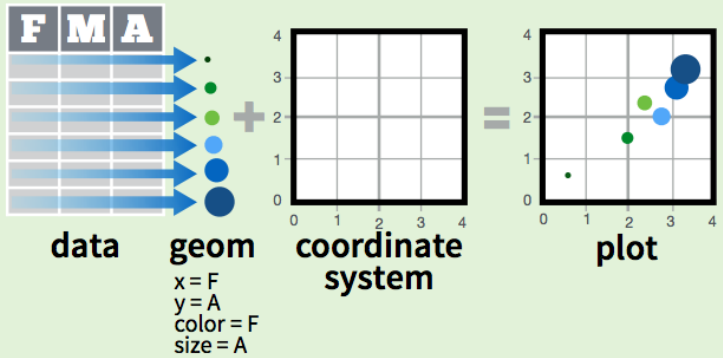


Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same few components: a **data** set, a set of **geoms**—visual marks that represent data points, and a **coordinate system.**

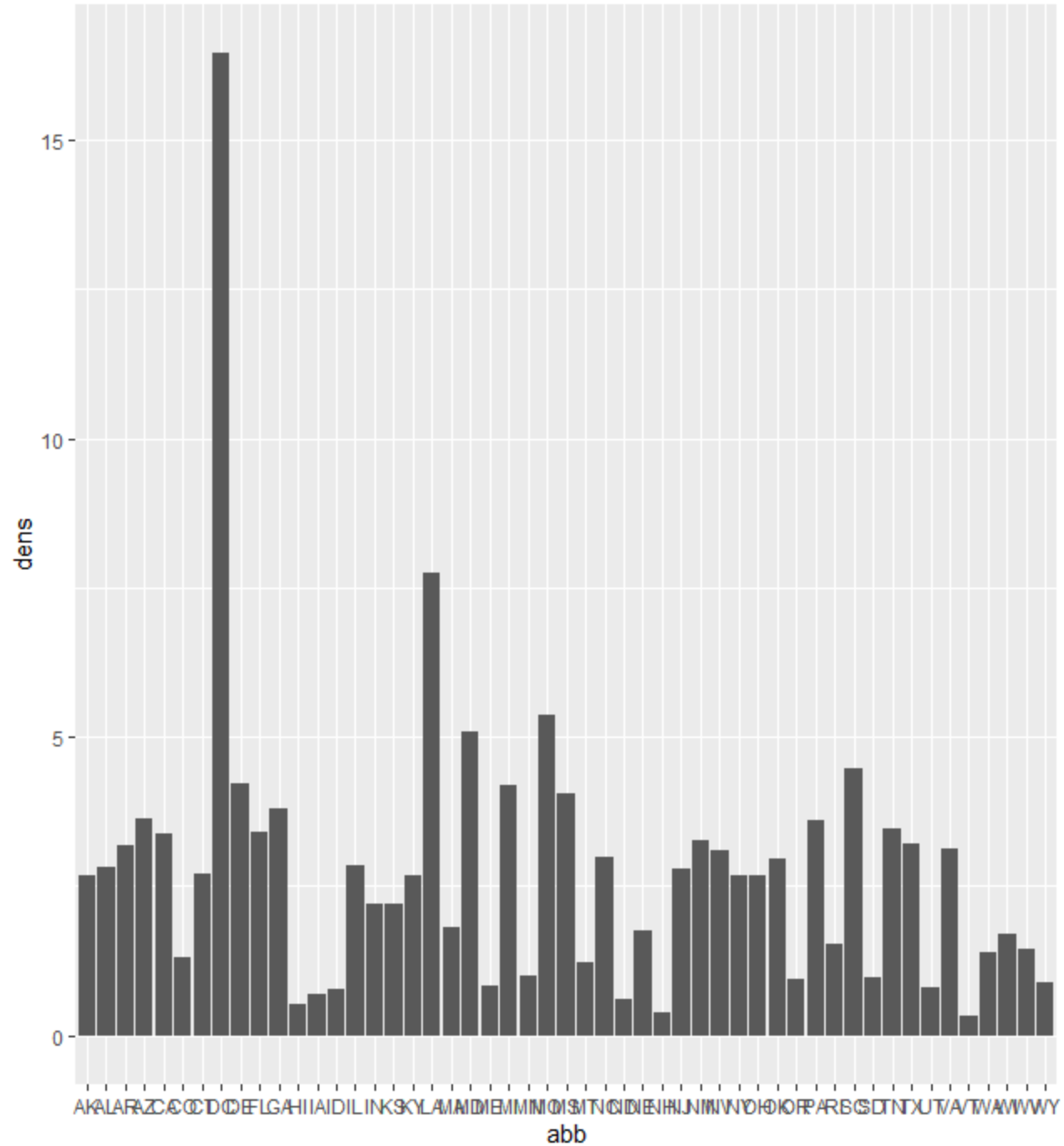


To display data values, map variables in the data set to aesthetic properties of the geom like **size**, **color**, and **x** and **y** locations.

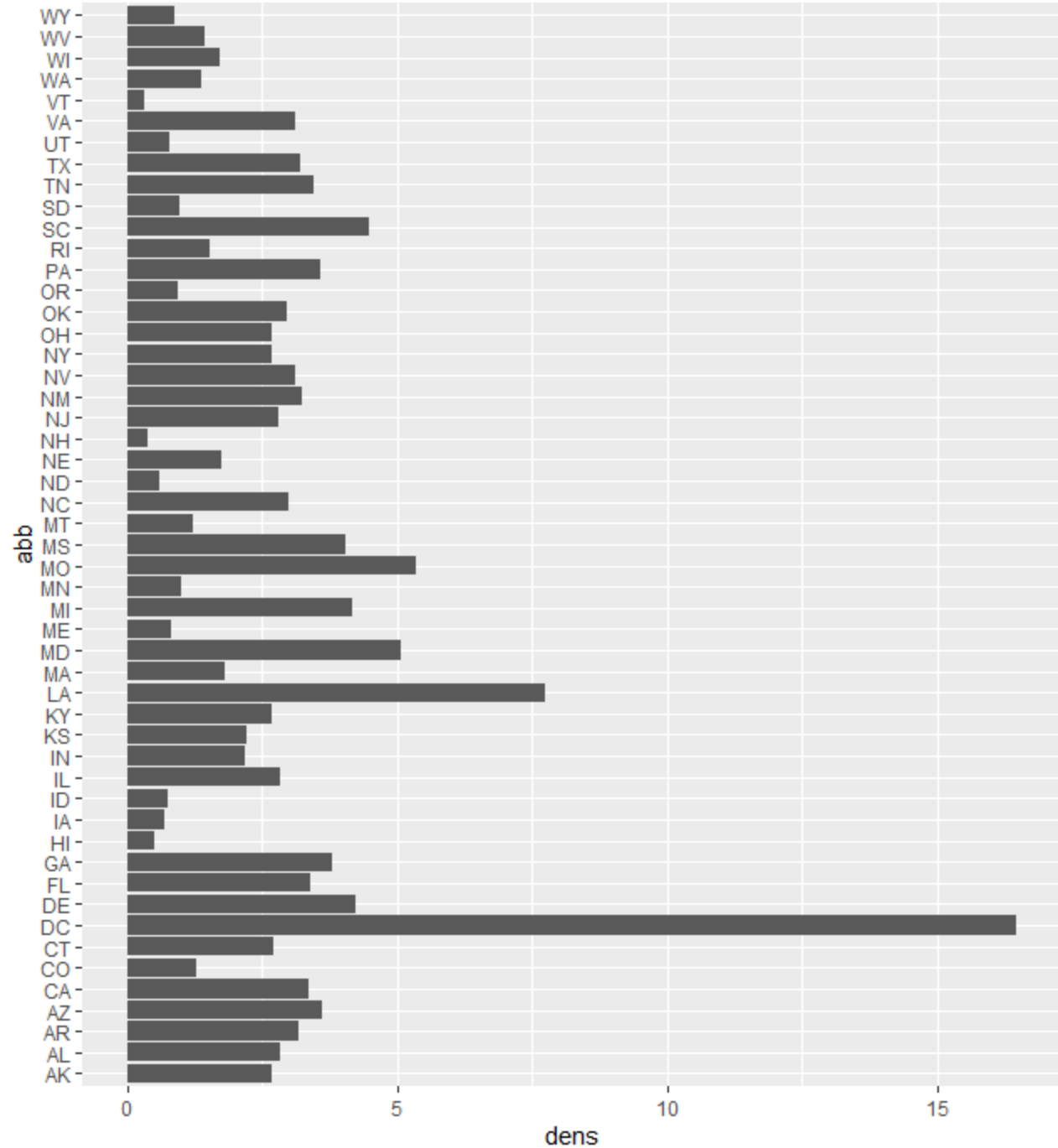


```
data %>%
    ggplot()
```

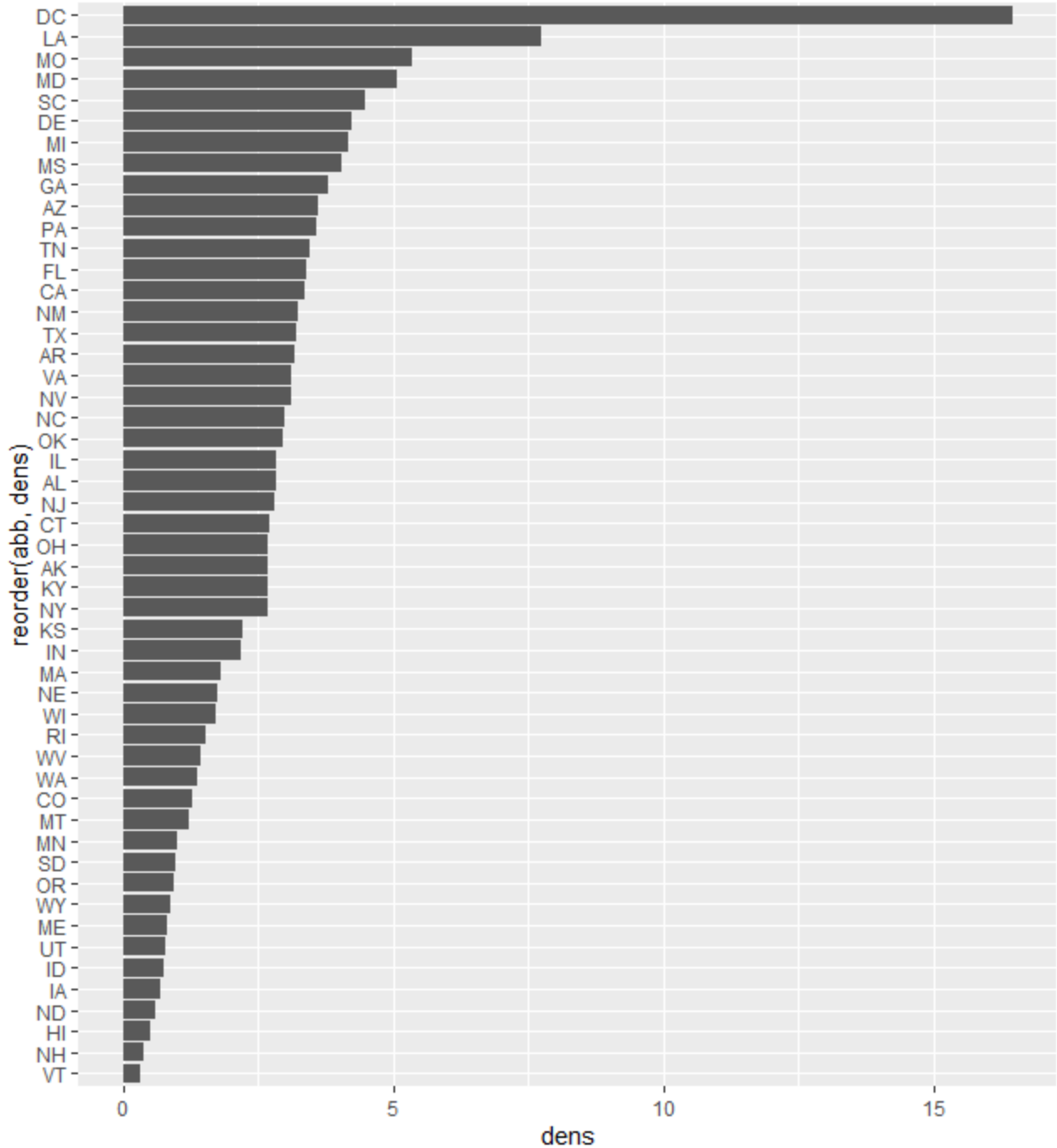
```
data %>%
        ggplot() +
        geom_col(aes(x = abb,
                     y = dens)
```


15
10
5
0
dens
abb

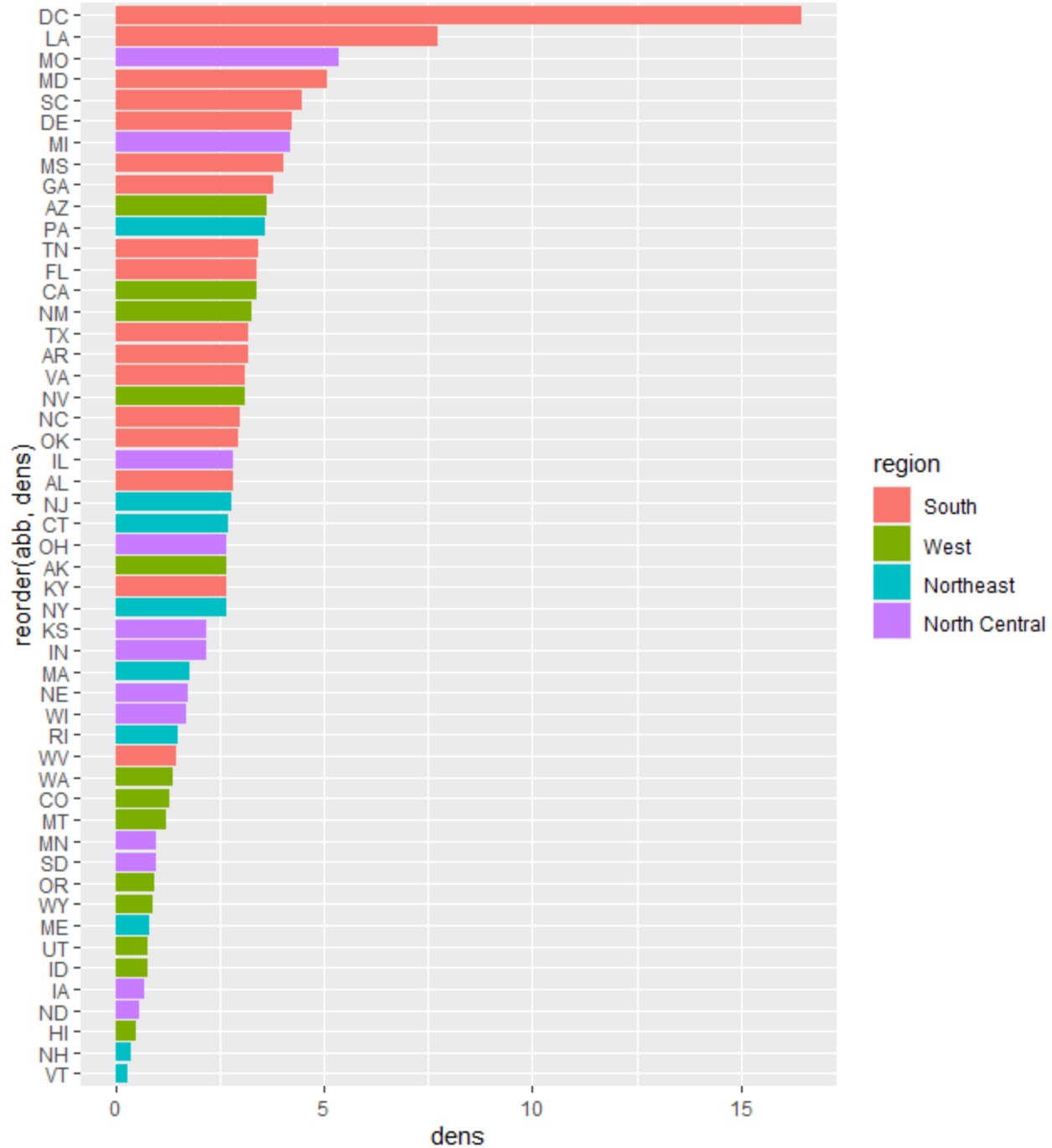
```
data %>%
    ggplot() +
    geom_col(aes(y = abb,
                 x = dens)
```


WY
WV
WI
WA
VT
VA
UT
TX
TN
SD
SC
RI
PA
OR
OK
OH
NY
NV
NM
NJ
NH
NE
ND
NC
MT
MS
MO
MN
MI
ME
MD
MA
LA
KY
KS
IN
IL
ID
IA
HI
GA
FL
DE
DC
CT
CO
CA
AZ
AR
AL
AK
abb
0
5
10
15
dens

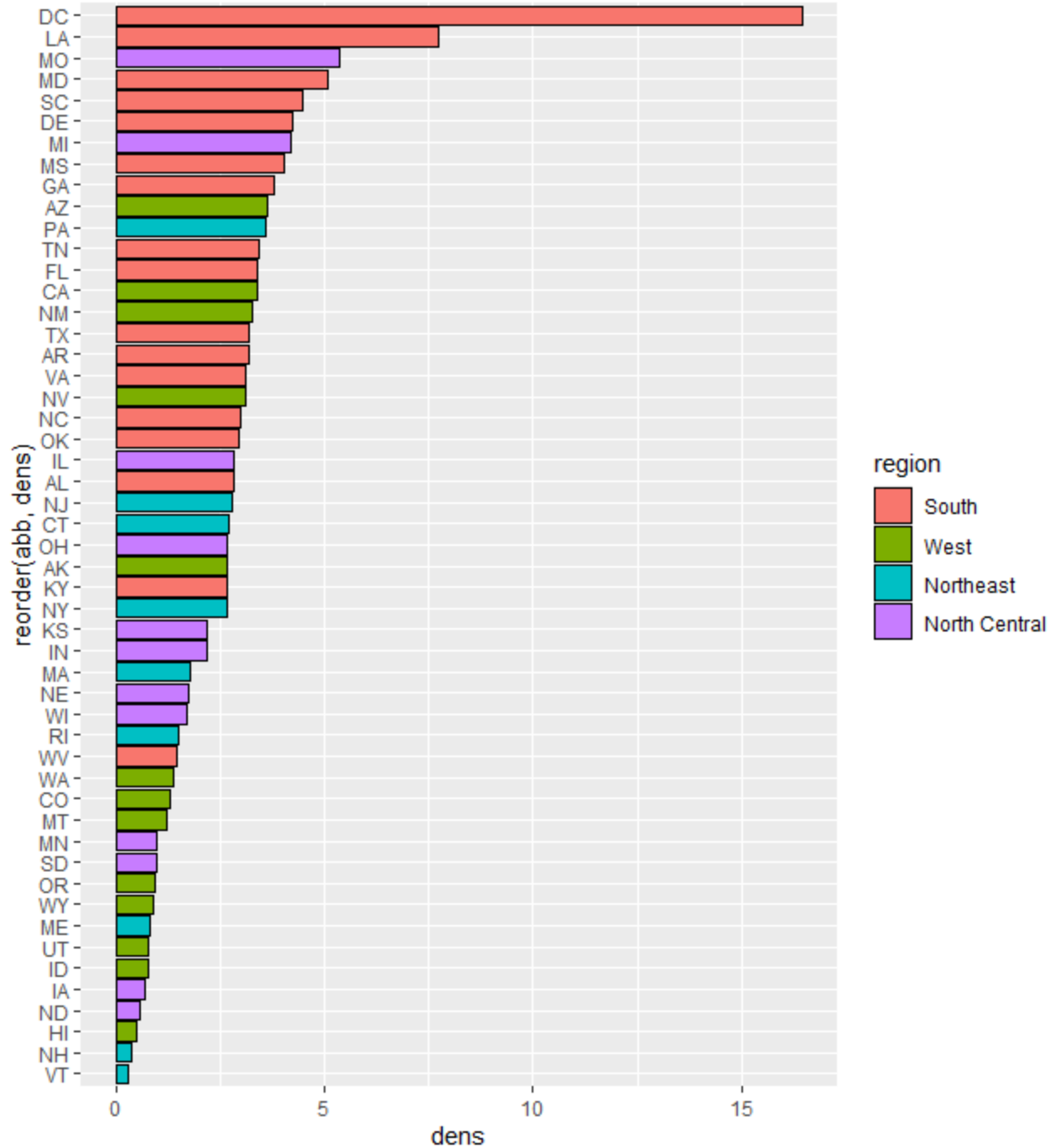
```
data %>%
    ggplot() +
    geom_col(aes(y = reorder(abb, dens),
                 x = dens)
```


DC
LA
MO
MD
SC
DE
MI
MS
GA
AZ
PA
TN
FL
CA
NM
TX
AR
VA
NV
NC
OK
IL
AL
NJ
CT
OH
AK
KY
NY
KS
IN
MA
NE
WI
RI
WV
WA
CO
MT
MN
SD
OR
WY
ME
UT
ID
IA
ND
HI
NH
VT
reorder(abb, dens)
0
5
10
15
dens

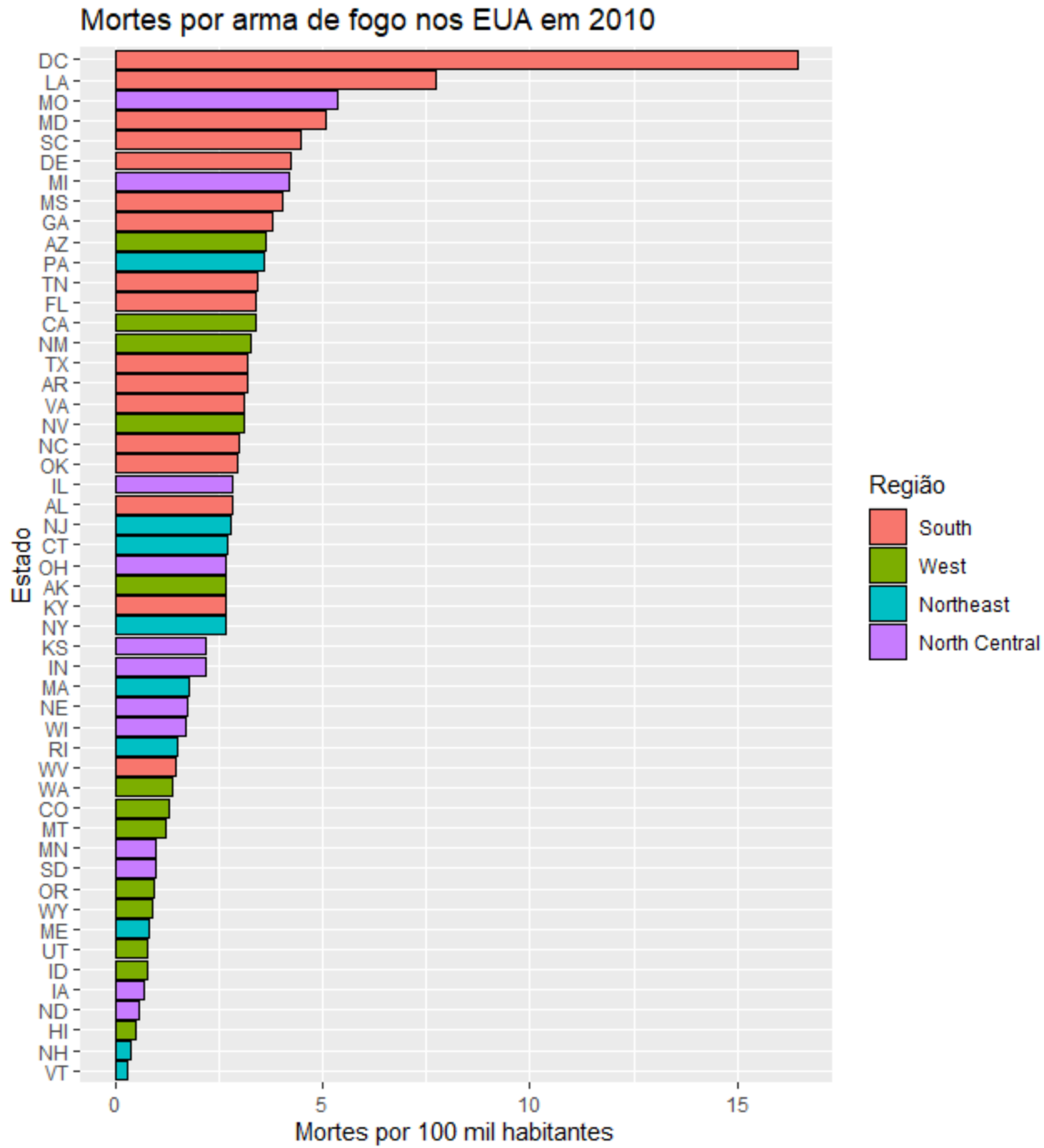
```
data %>%
    ggplot() +
    geom_col(aes(y = reorder(abb, dens),
                 x = dens,
                 fill = region)
```


DC
LA
MO
MD
SC
DE
MI
MS
GA
AZ
PA
TN
FL
CA
NM
TX
AR
VA
NV
NC
OK
IL
AL
NJ
CT
OH
AK
KY
NY
KS
IN
MA
NE
WI
RI
WV
WA
CO
MT
MN
SD
OR
WY
ME
UT
ID
IA
ND
HI
NH
VT
reorder(abb, dens)
0
5
10
15
dens
region
South
West
Northeast
North Central

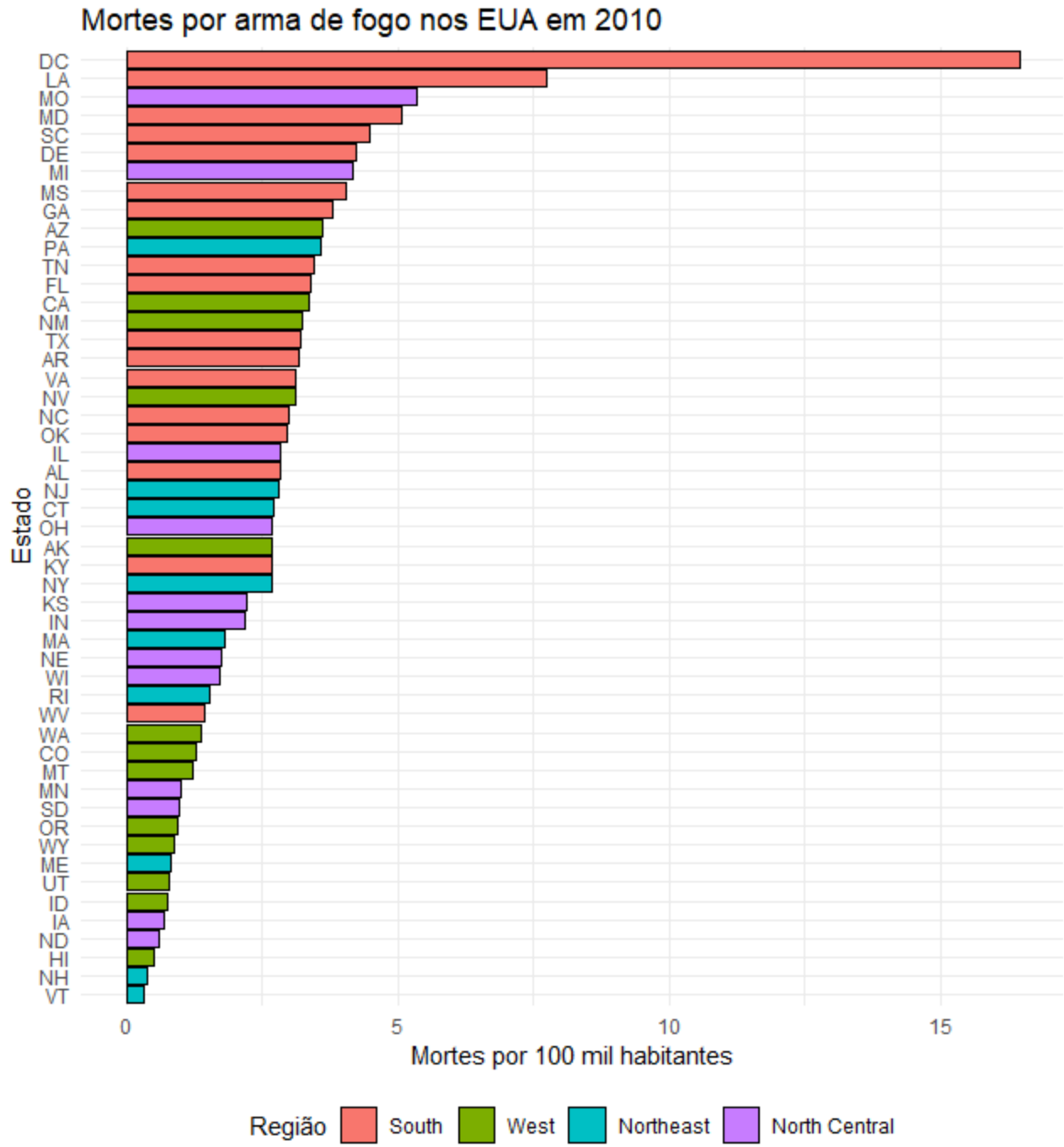
```
data %>%
    ggplot() +
    geom_col(aes(y = reorder(abb, dens),
                 x = dens,
                 fill = region),
    color = 'black')
```

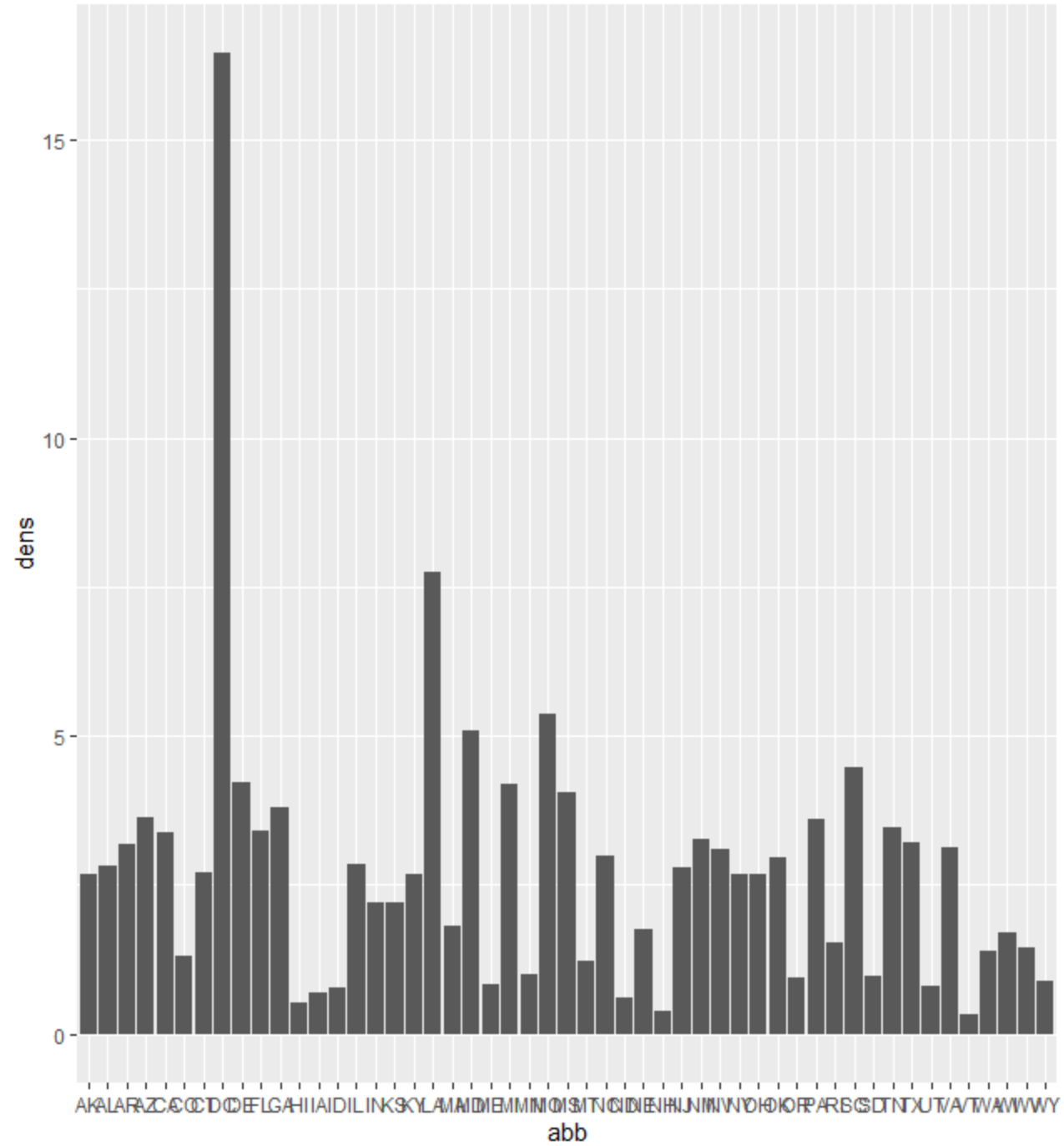

DC
LA
MO
MD
SC
DE
MI
MS
GA
AZ
PA
TN
FL
CA
NM
TX
AR
VA
NV
NC
OK
IL
AL
NJ
CT
OH
AK
KY
NY
KS
IN
MA
NE
WI
RI
WV
WA
CO
MT
MN
SD
OR
WY
ME
UT
ID
IA
ND
HI
NH
VT
reorder(abb, dens)
0
5
10
15
dens
region
South
West
Northeast
North Central

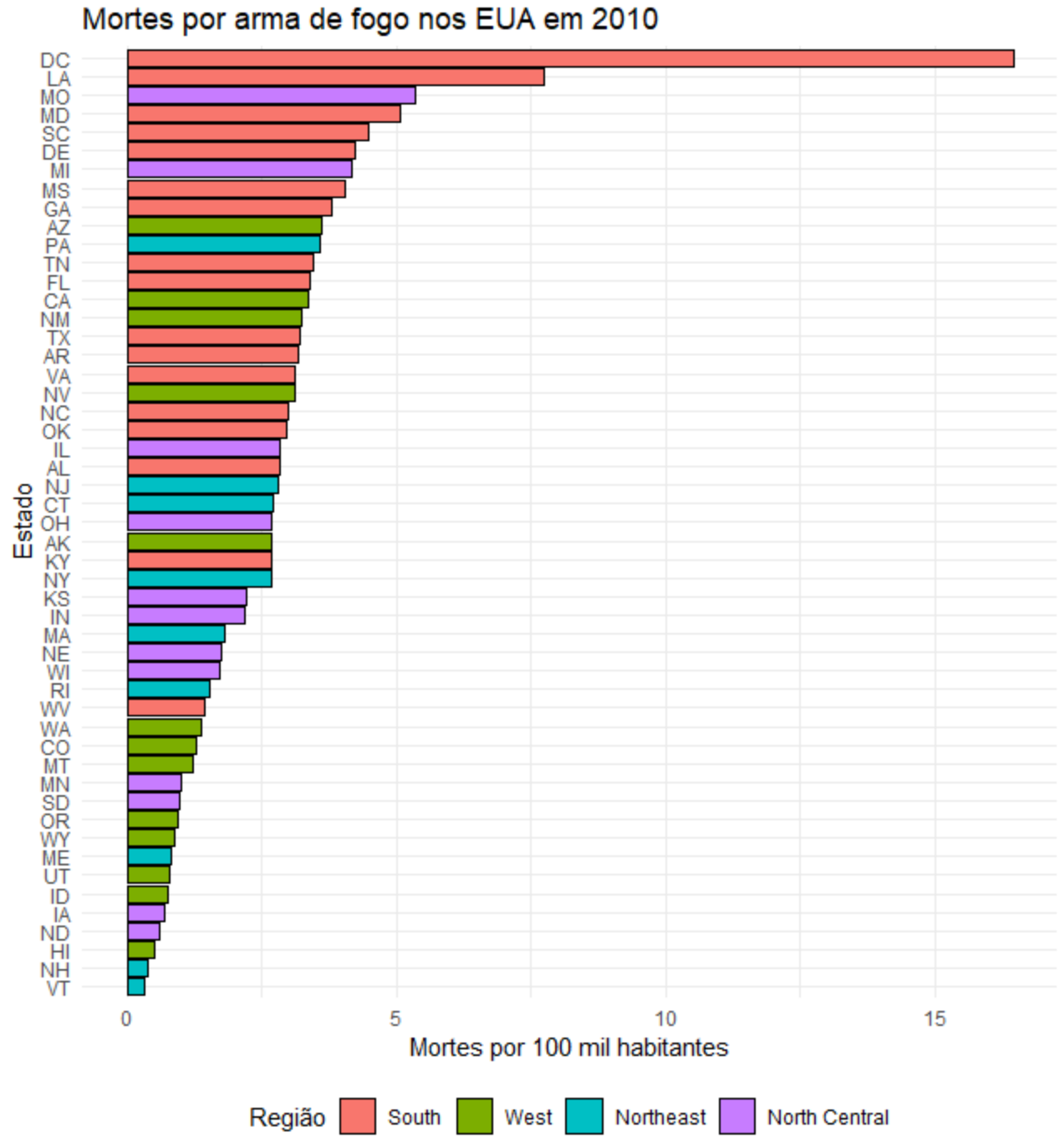
```
data %>%
        ggplot() +
        geom_col(aes(y = reorder(abb, dens),
                                 x = dens,
                                 fill = region),
        color = 'black') +
labs(x = 'Mortes por 100 mil habitantes’,
        y = 'Estado’,
        title = 'Mortes por arma de fogo nos EUA em
2010’,
        fill = 'Região')
```

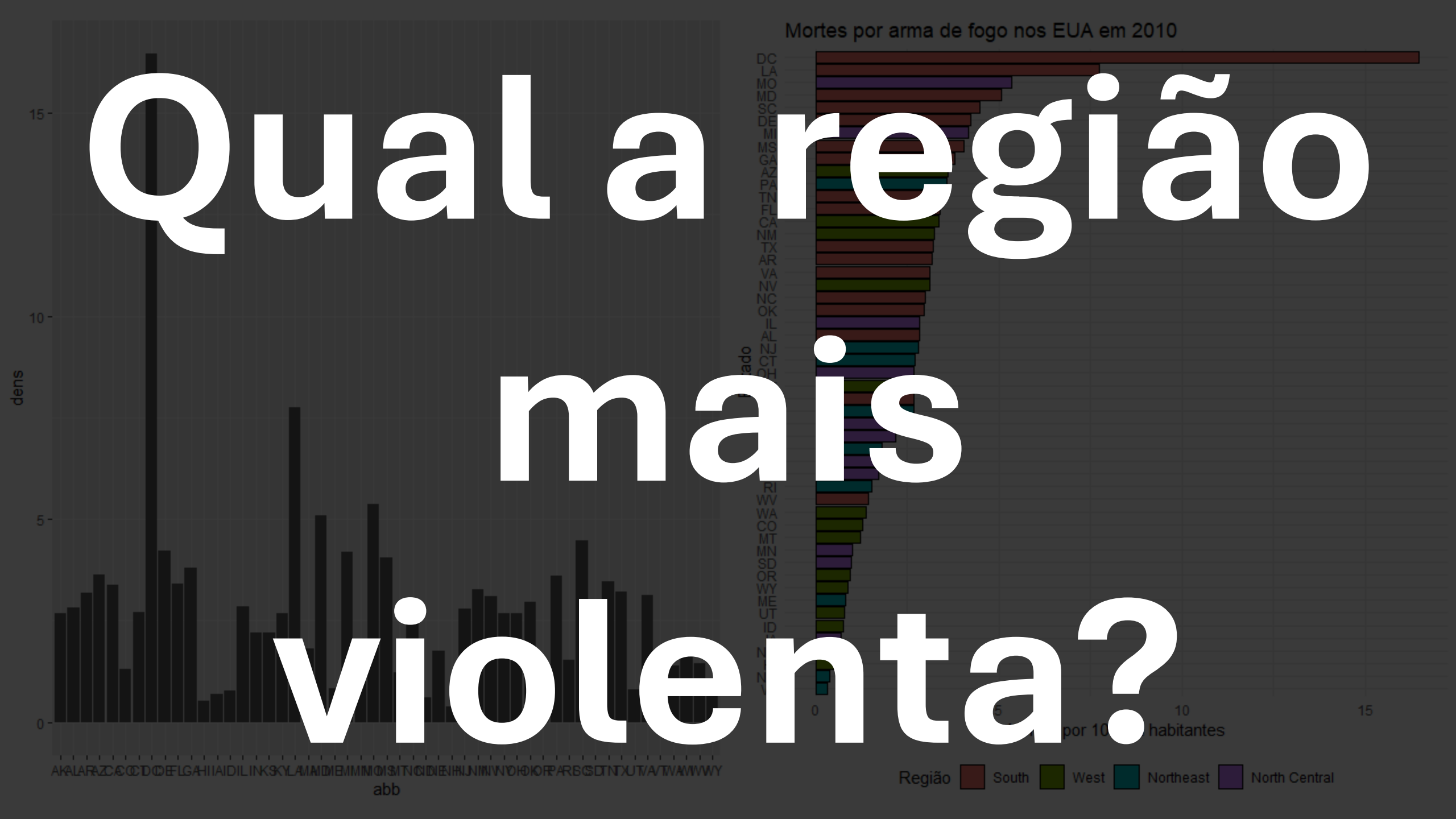

Mortes por arma de fogo nos EUA em 2010
DC
LA
MO
MD
SC
DE
MI
MS
GA
AZ
PA
TN
FL
CA
NM
TX
AR
VA
NV
NC
OK
IL
AL
NJ
CT
OH
AK
KY
NY
KS
IN
MA
NE
WI
RI
WV
WA
CO
MT
MN
SD
OR
WY
ME
UT
ID
IA
ND
HI
NH
VT
Estado
0
5
10
15
Mortes por 100 mil habitantes
Região
South
West
Northeast
North Central

```
data %>%
ggplot() +
geom_col(aes(y = reorder(abb, dens),
                        x = dens,
                        fill = region),
color = 'black') +
labs(x = 'Mortes por 100 mil habitantes’,
       y = 'Estado’,
       title = 'Mortes por arma de fogo nos EUA em 2010’,
       fill = 'Região') +
theme_minimal() +
theme(legend.position = 'bottom')
```

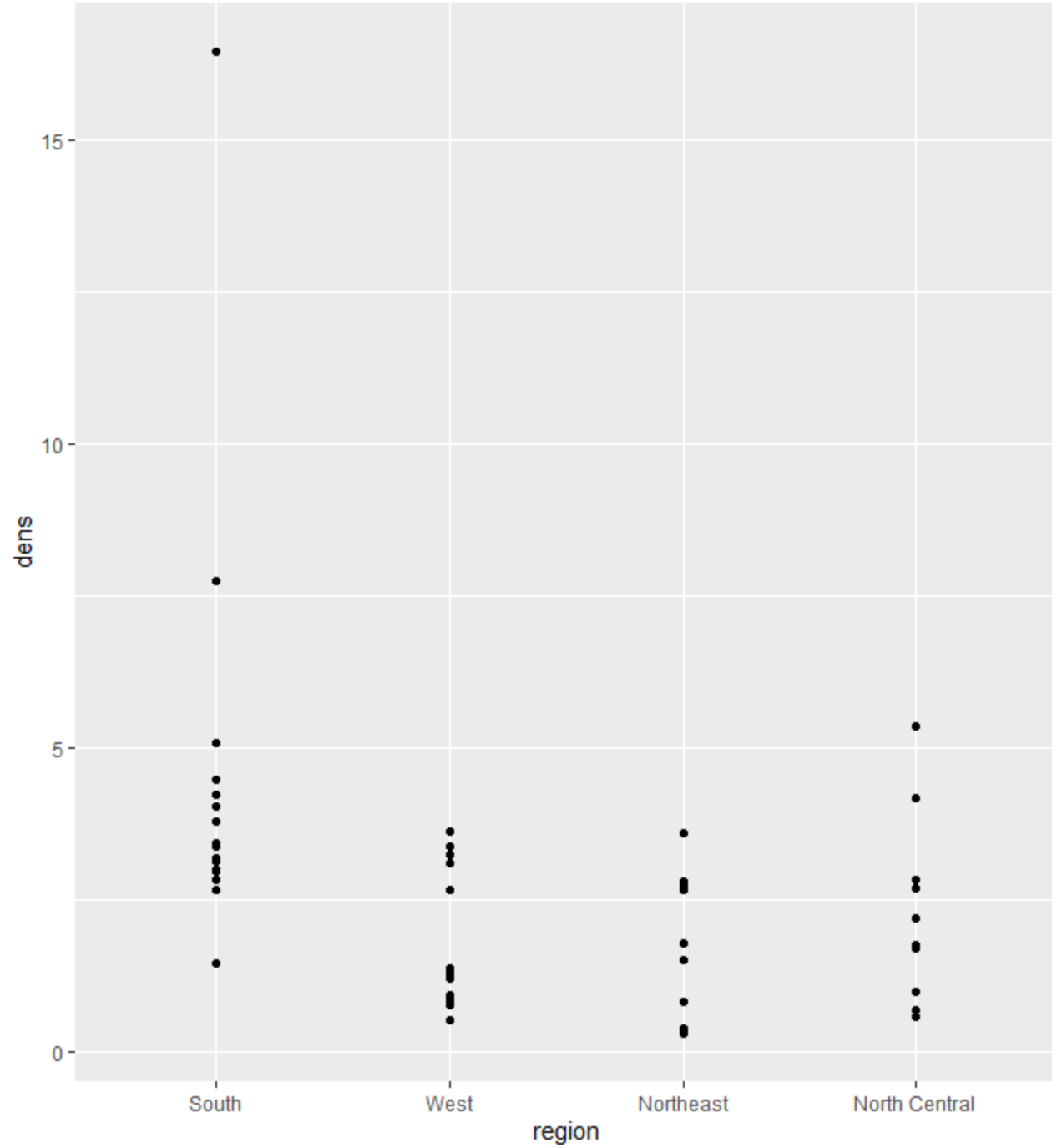

Mortes por arma de fogo nos EUA em 2010
Estado
DC
LA
MO
MD
SC
DE
MI
MS
GA
AZ
PA
TN
FL
CA
NM
TX
AR
VA
NV
NC
OK
IL
AL
NJ
CT
OH
AK
KY
NY
KS
IN
MA
NE
WI
RI
WV
WA
CO
MT
MN
SD
OR
WY
ME
UT
ID
IA
ND
HI
NH
VT
0
5
10
15
Mortes por 100 mil habitantes
Região
South
West
Northeast
North Central


15
10
5
0
dens
abb

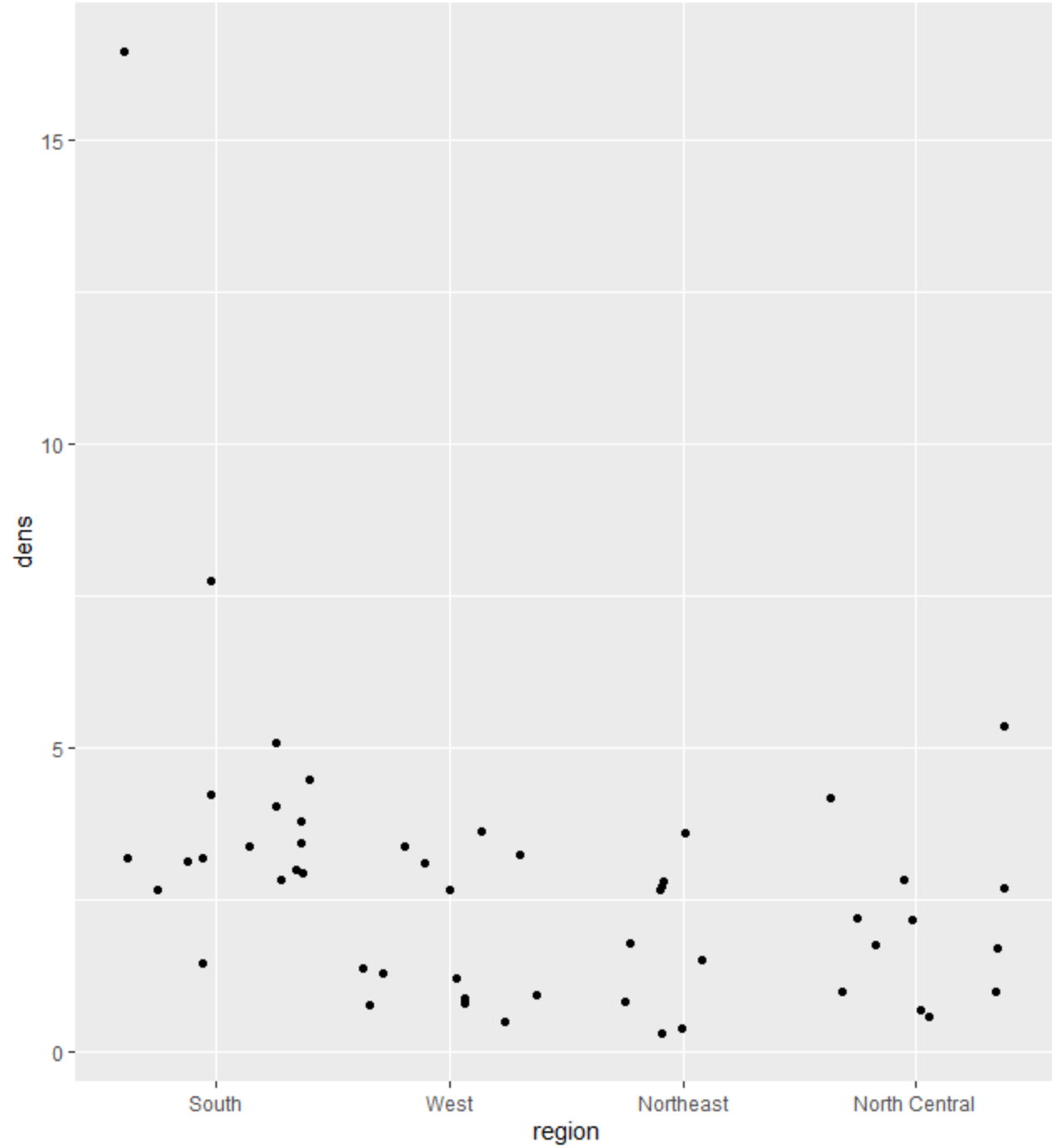

Mortes por arma de fogo nos EUA em 2010
Estado
DC
LA
MO
MD
SC
DE
MI
MS
GA
AZ
PA
TN
FL
CA
NM
TX
AR
VA
NV
NC
OK
IL
AL
NJ
CT
OH
AK
KY
NY
KS
IN
MA
NE
WI
RI
WV
WA
CO
MT
MN
SD
OR
WY
ME
UT
ID
IA
ND
HI
NH
VT
0
5
10
15
Mortes por 100 mil habitantes
Região
South
West
Northeast
North Central


Mortes por arma de fogo nos EUA em 2010
DC
LA
MO
MD
SC
DE
MI
MS
GA
AZ
PA
TN
FL
CA
NM
TX
AR
VA
NV
NC
OK
IL
AL
NJ
CT
OH
RI
WV
WA
CO
MT
MN
SD
OR
WY
ME
UT
ID
0
5
10
15
por 10
habitantes
Região
South
West
Northeast
North Central
dens
0
5
10
15
abb
Qual a região mais violenta?

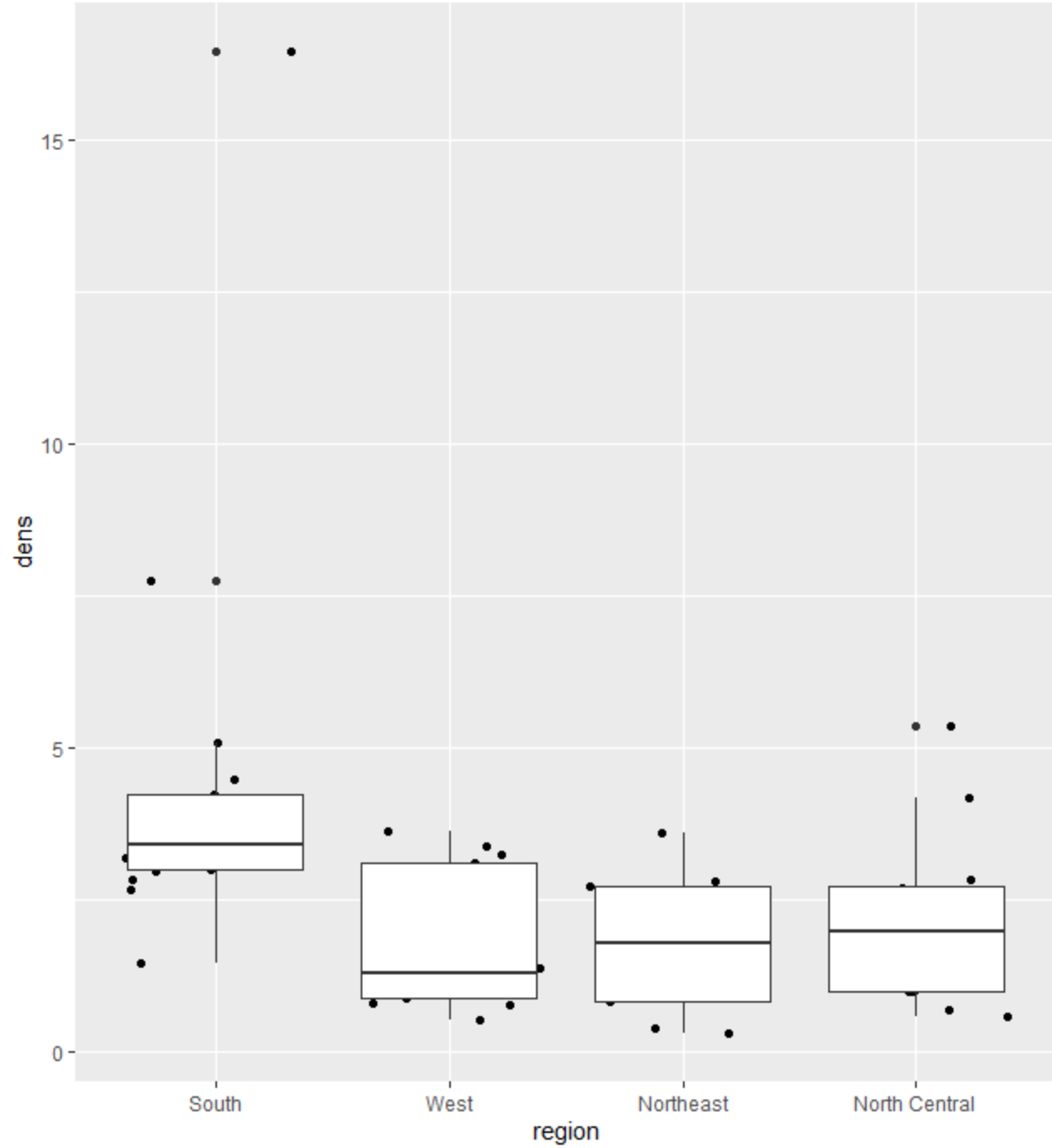
```
data %>%
ggplot() +
geom_point(aes(x = region, y = dens))
```


15
10
5
0
dens
South
West
Northeast
North Central
region

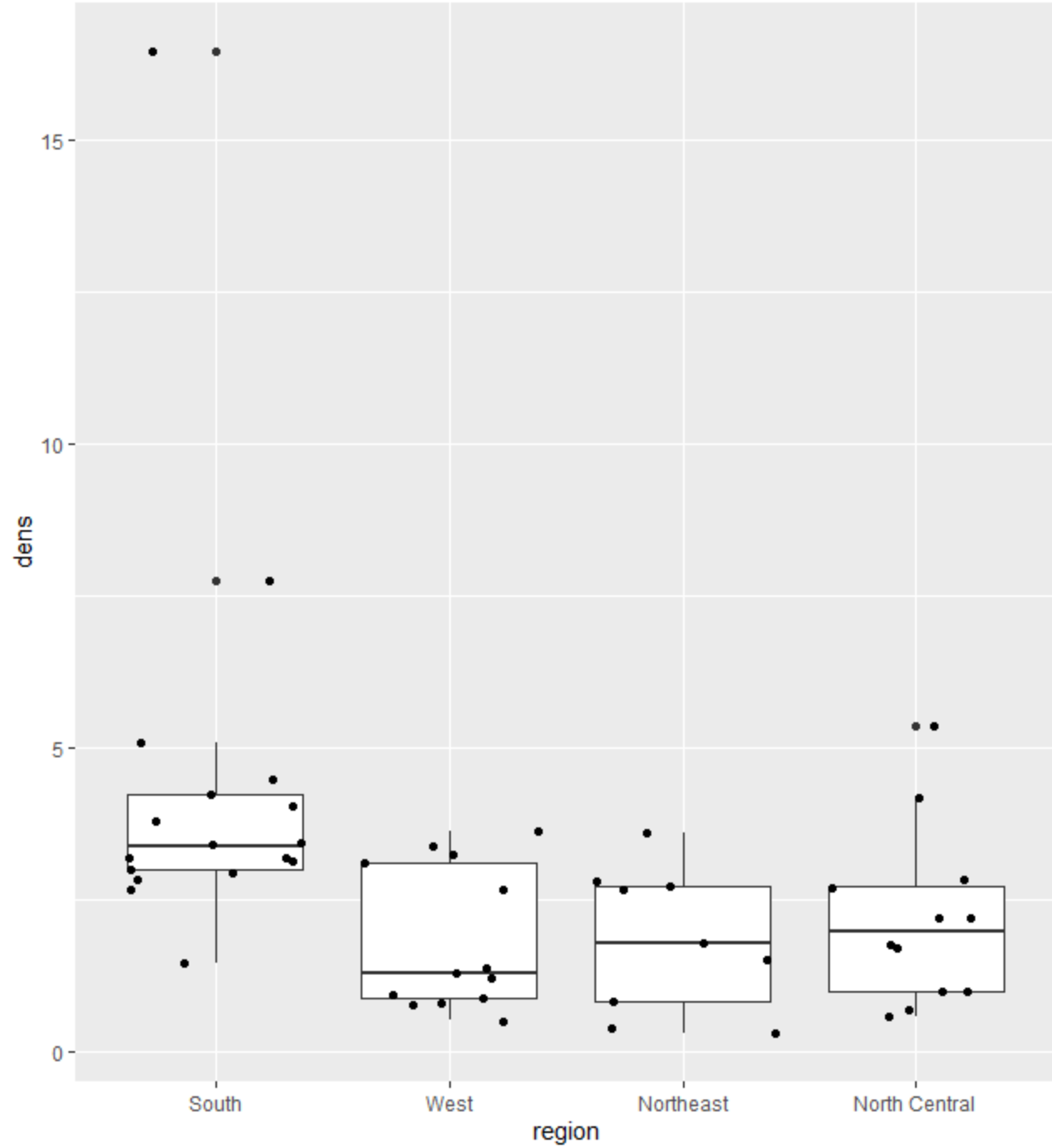
```
data %>%
ggplot() +
geom_jitter(aes(x = region, y = dens))
```


15
10
5
0
dens
South
West
Northeast
North Central
region

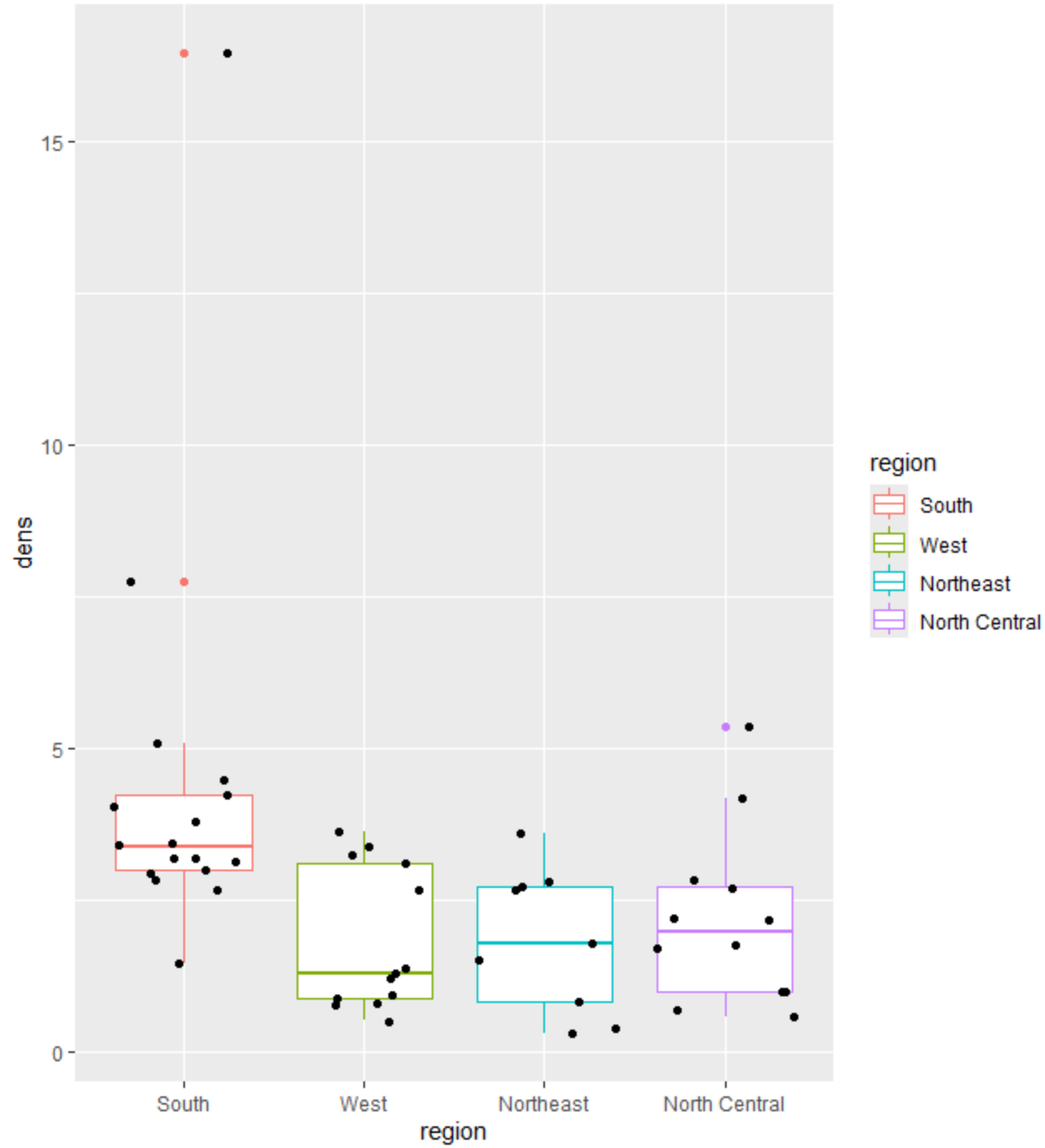
```
data %>%
ggplot() +
geom_jitter(aes(x = region, y = dens)) +
geom_boxplot(aes(y = dens, x = region))
```


15
10
5
0
dens
South
West
Northeast
North Central
region

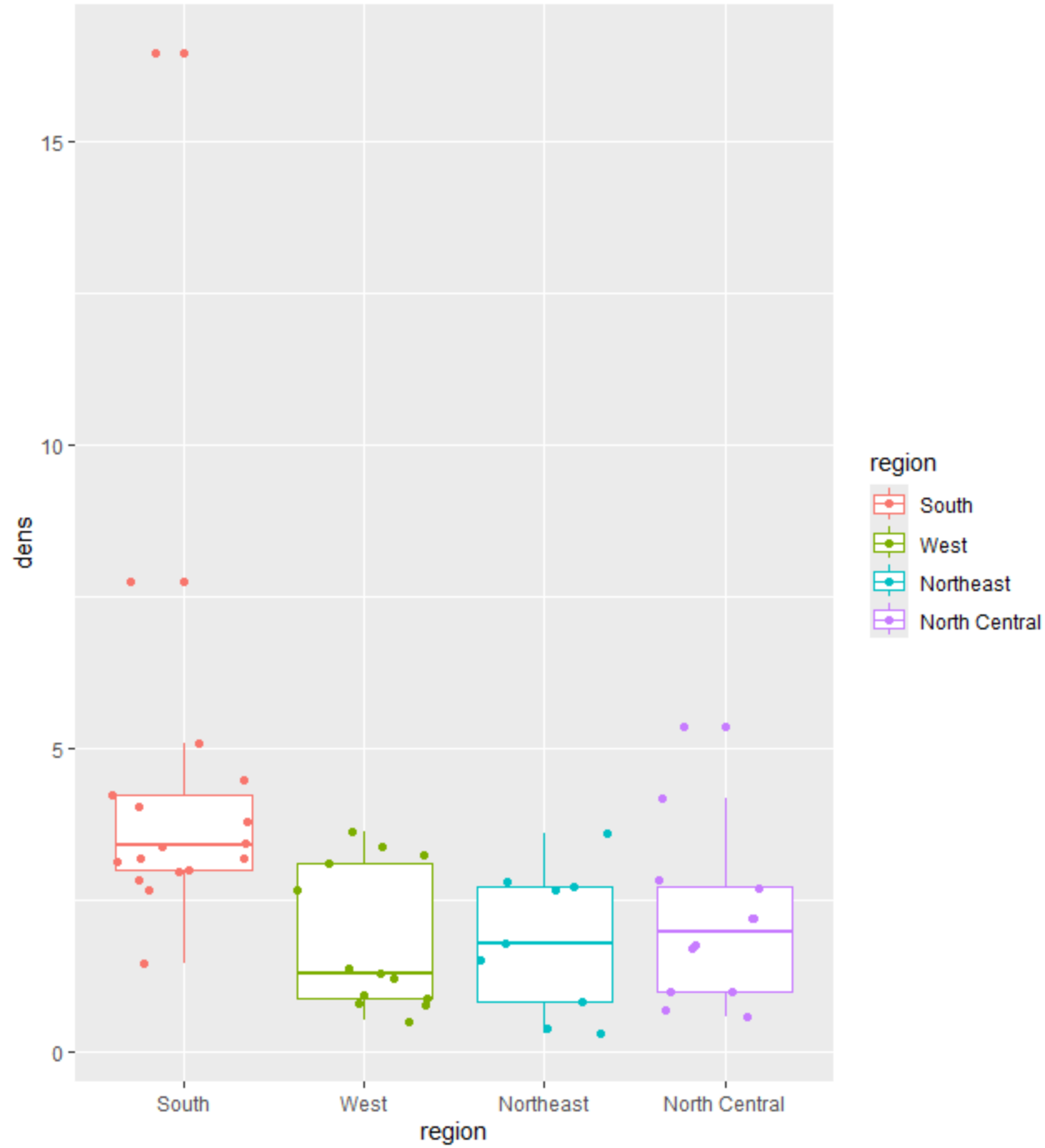
```
data %>%
ggplot() +
geom_boxplot(aes(y = dens, x = region)) +
geom_jitter(aes(x = region, y = dens))
```


15
10
5
0
dens
South
West
Northeast
North Central
region

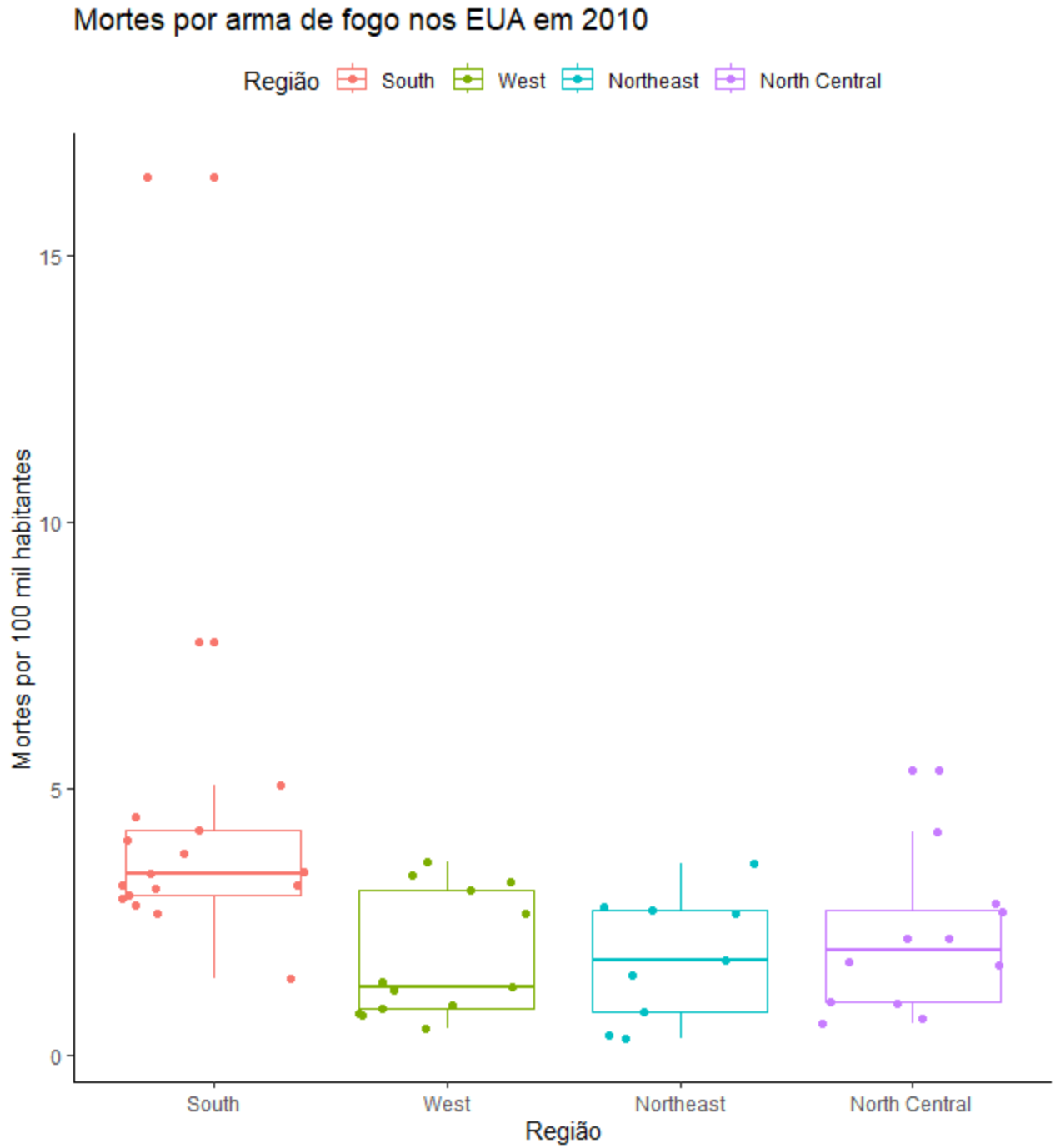
```
data %>%
ggplot() +
geom_boxplot(aes(y = dens, x = region, color = region))
+
geom_jitter(aes(x = region, y = dens))
```

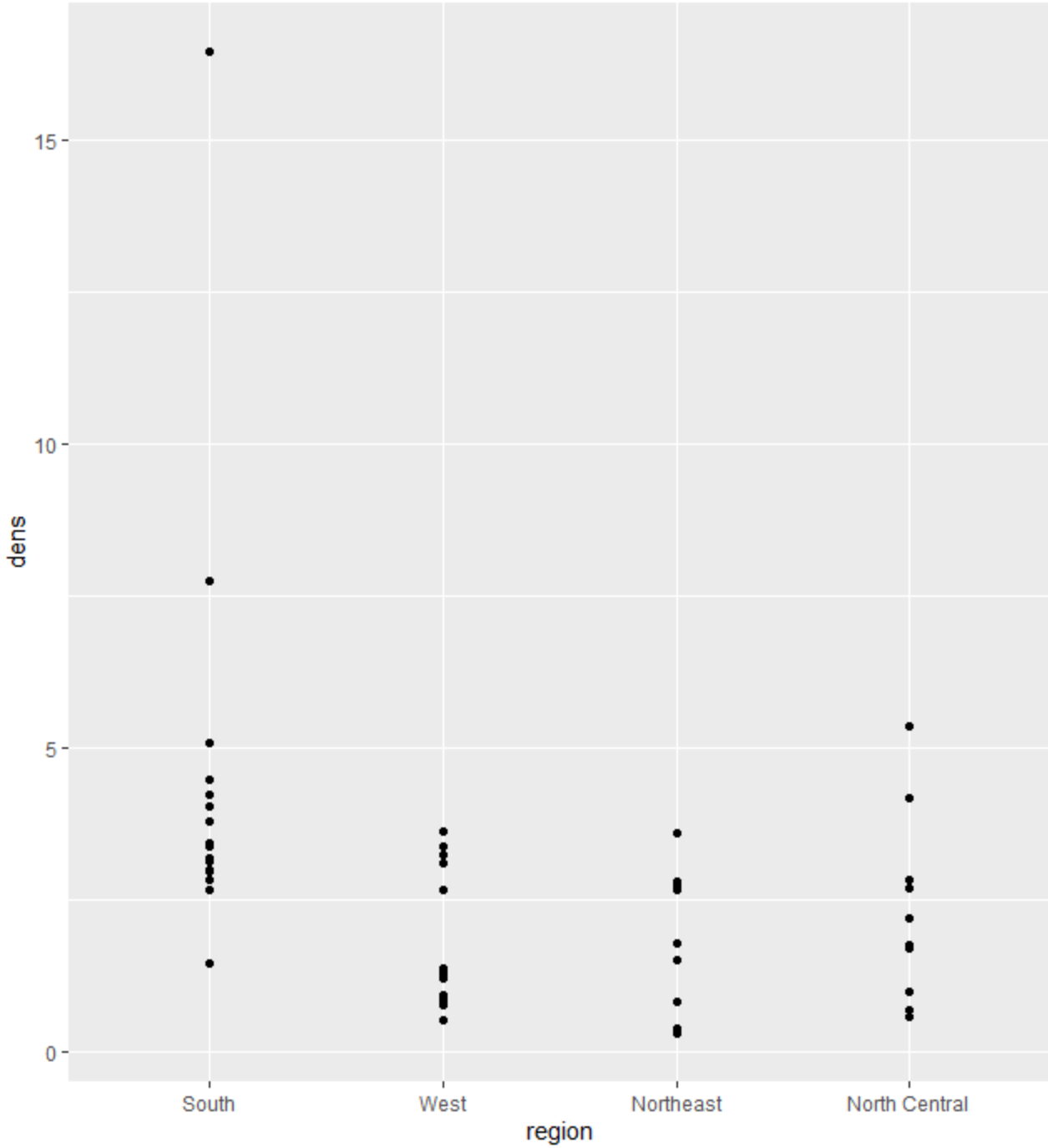

15
10
5
0
dens
South
West
Northeast
North Central
region
region
South
West
Northeast
North Central

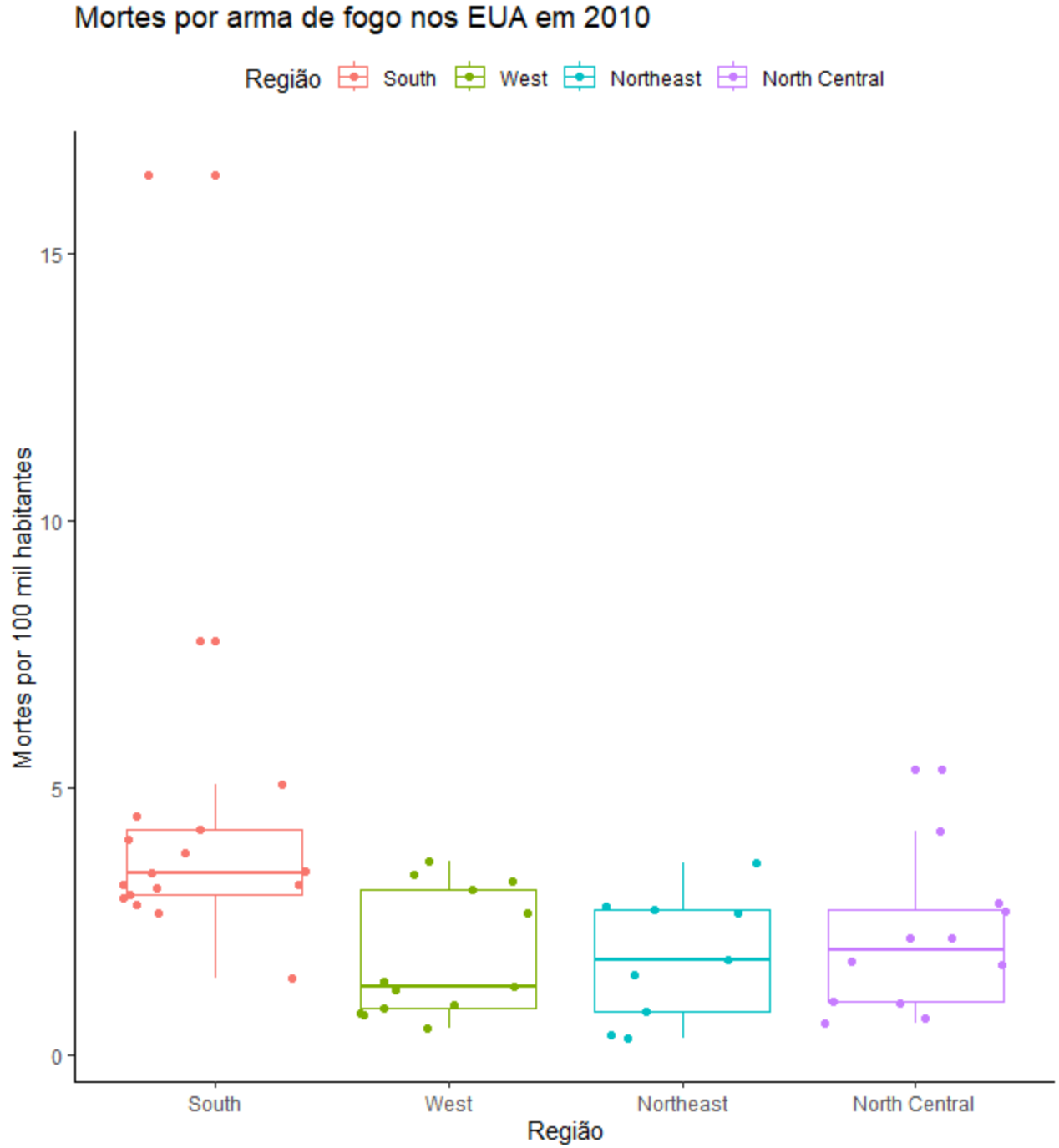
```
data %>%
ggplot() +
geom_boxplot(aes(y = dens, x = region, color = region))
+
geom_jitter(aes(x = region,
                y = dens,
                color = region))
```

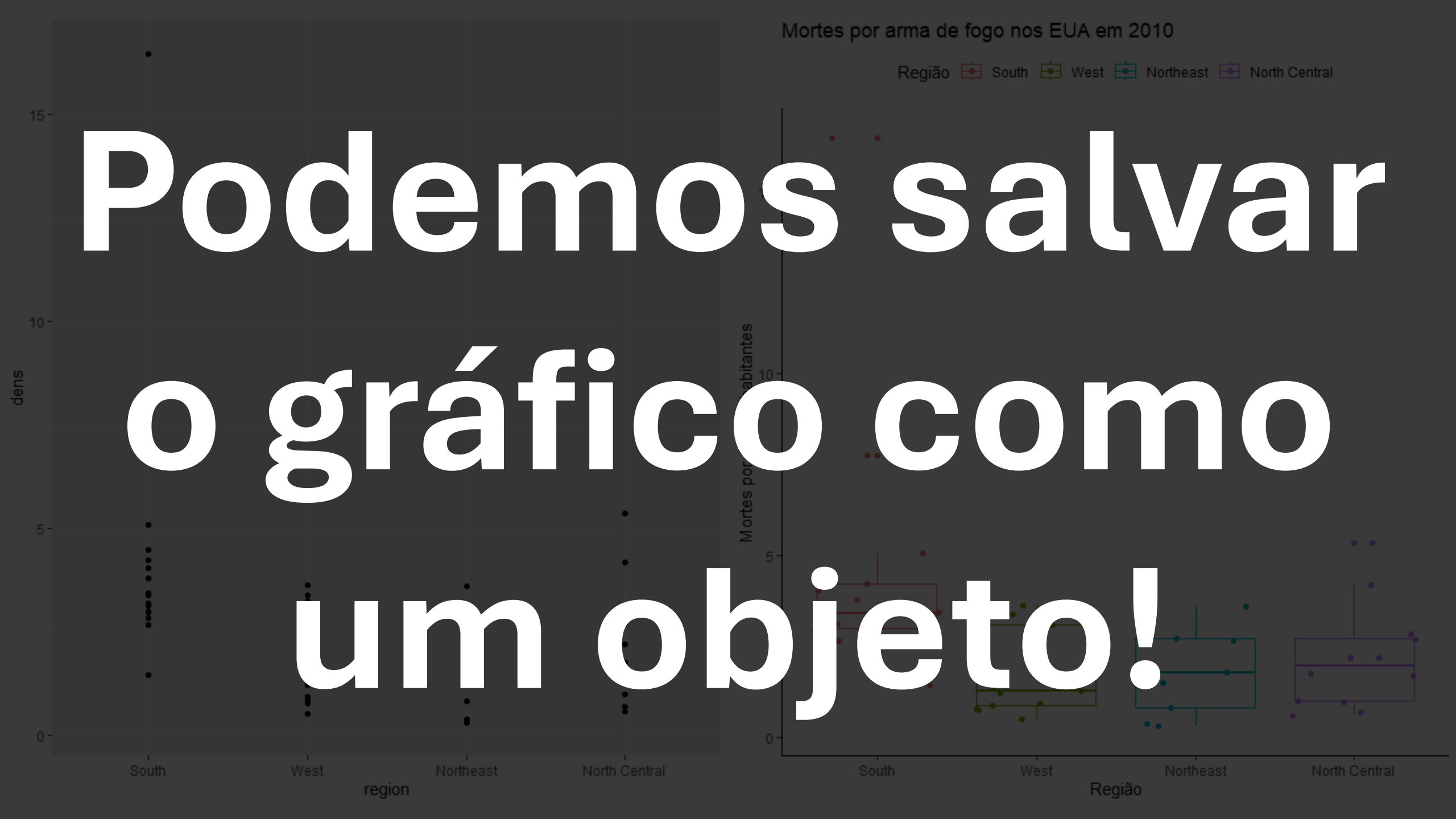

15
10
5
0
dens
South
West
Northeast
North Central
region
region
South
West
Northeast
North Central

```
data %>%
ggplot() +
geom_boxplot(aes(y = dens, x = region, color = region)) +
geom_jitter(aes(x = region,
                y = dens,
                color = region)) +
labs(x = 'Região’,
     y = 'Mortes por 100 mil habitantes’,
     title = 'Mortes por arma de fogo nos EUA em 2010’,
     color = 'Região') +
theme_classic() +
theme(legend.position = 'top')
```


Mortes por arma de fogo nos EUA em 2010
Região
South
West
Northeast
North Central
Mortes por 100 mil habitantes
0
5
10
15
South
West
Northeast
North Central
Região

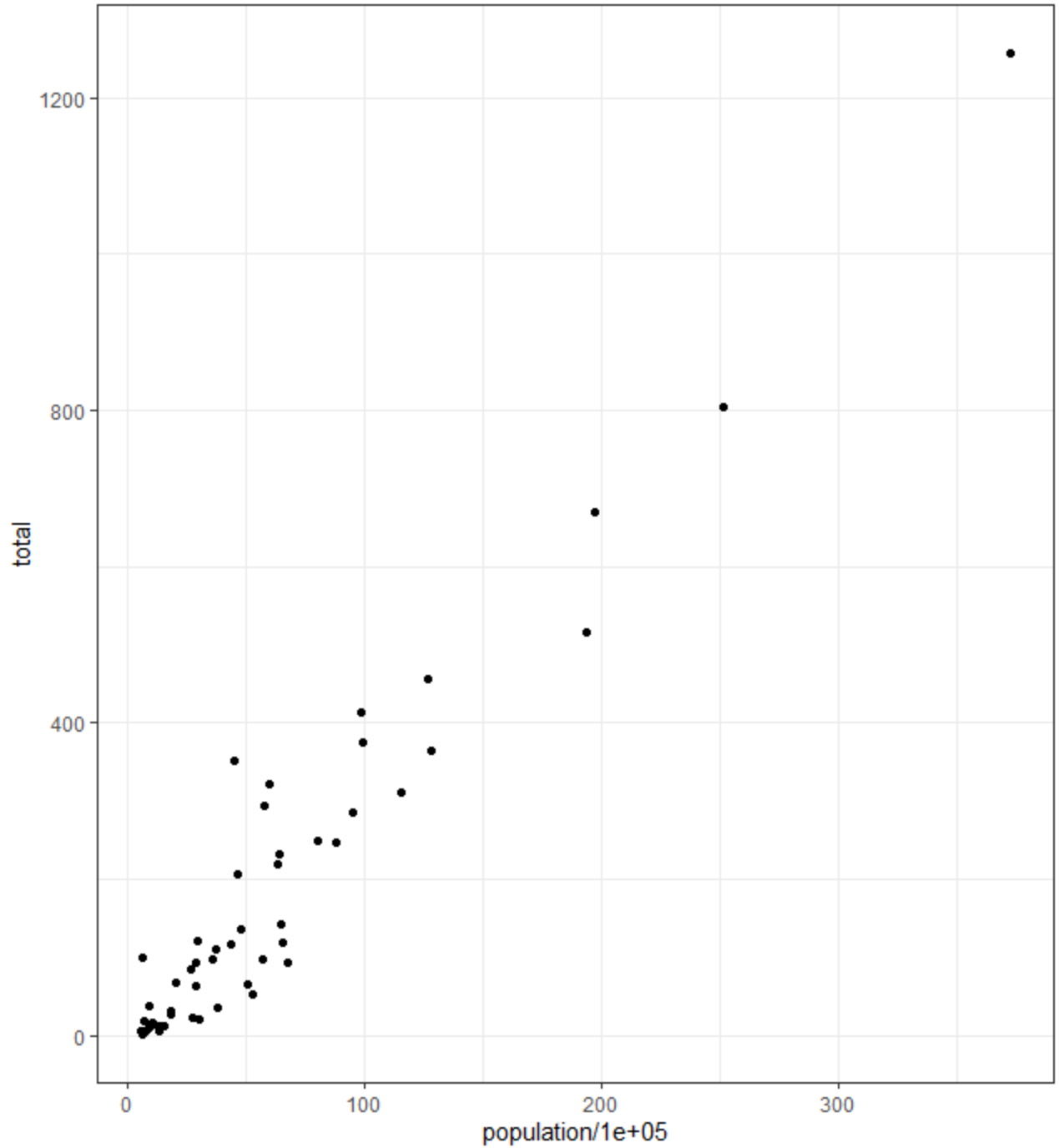

15
10
5
0
dens
South
West
Northeast
North Central
region


Mortes por arma de fogo nos EUA em 2010
Região
South
West
Northeast
North Central
15
10
5
0
Mortes por 100 mil habitantes
South
West
Northeast
North Central
Região

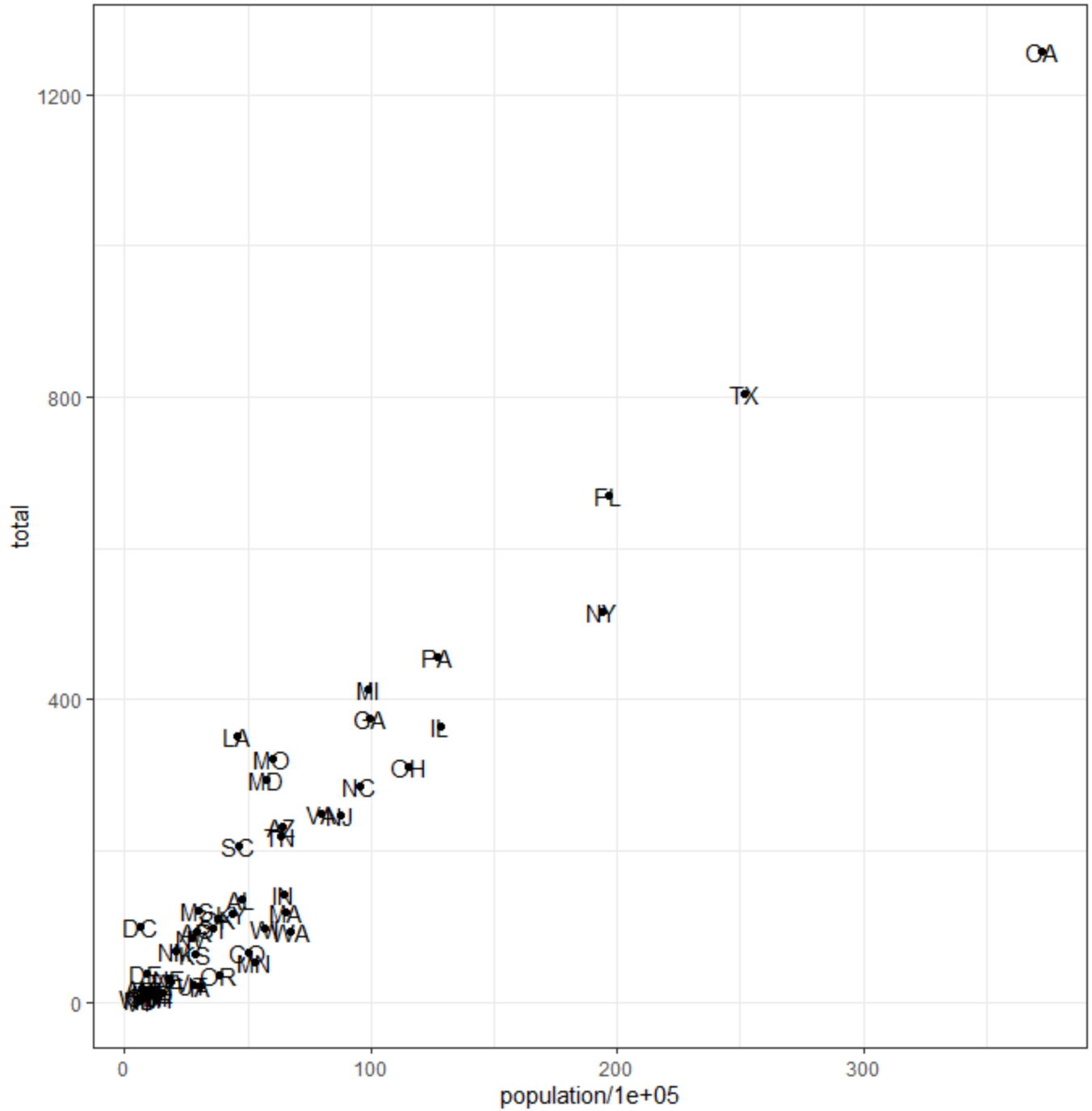


Podemos salvar o gráfico como um objeto!

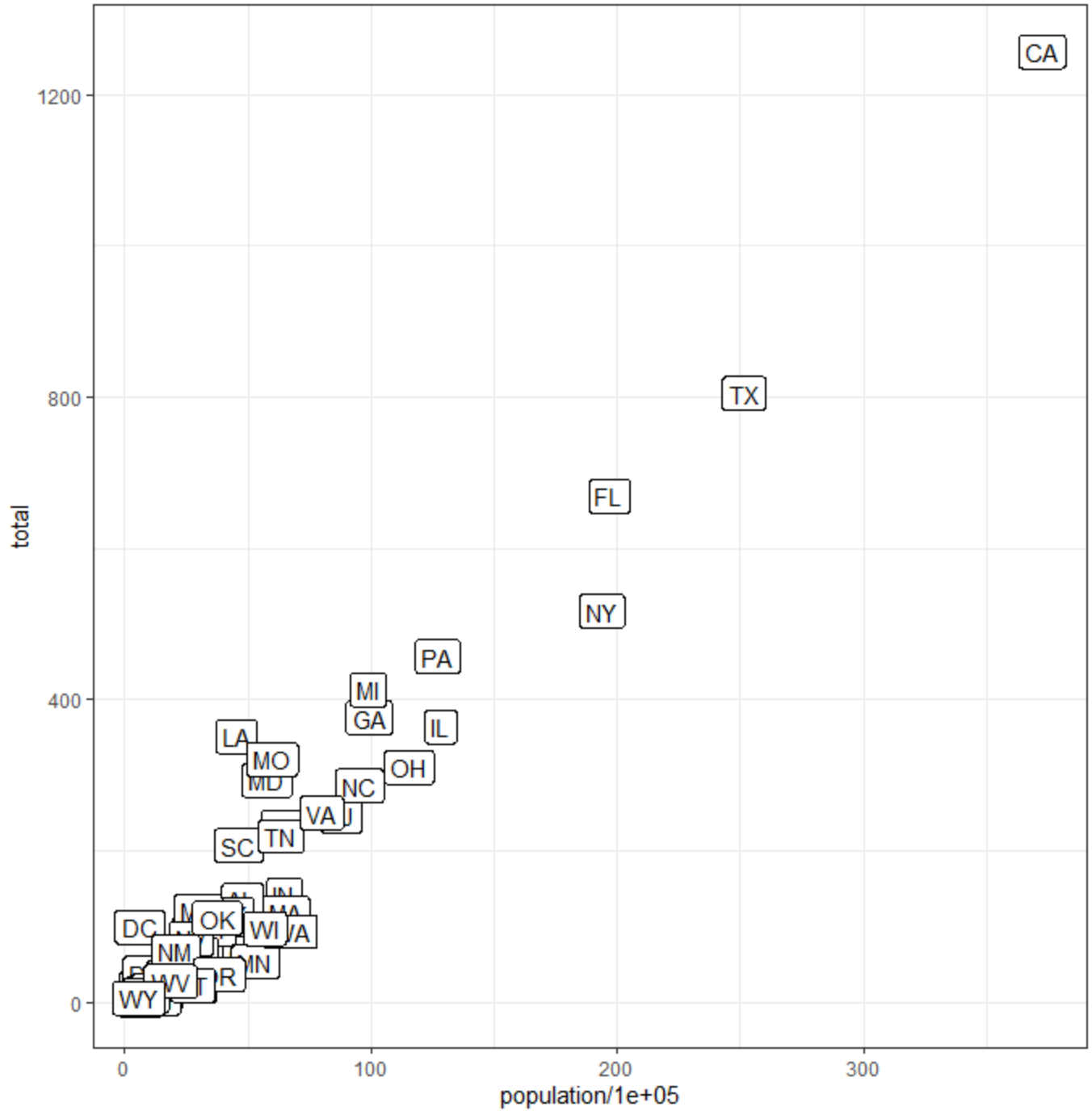
```
p <- ggplot(data = data)
class(p)


p +
  geom_point(aes(x = population/100000, y = total))
```


1200
800
400
0
total
0
100
200
300
population/1e+05

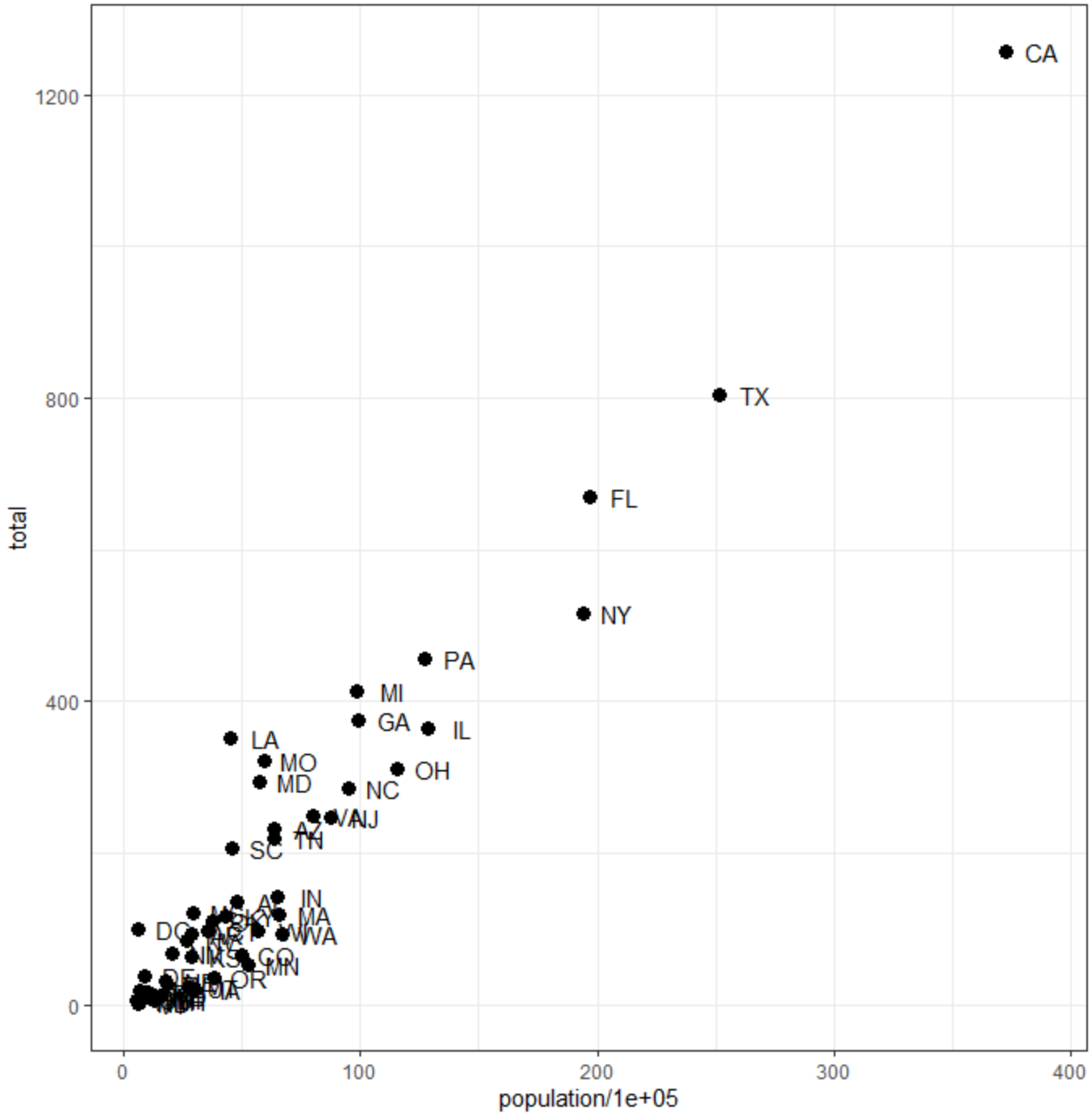
```
p +
geom_point(aes(x = population/100000, y = total), size = 3) +
geom_text(aes(x = population/100000, y = total, label = abb))
```


1200
800
400
0
total
0
100
200
300
population/1e+05
CA
TX
FL
NY
PA
IL
MI
GA
LA
MO
OH
MD
NC
VA
NJ
AZ
TN
SC
IN
AL
MS
KY
MA
WI
WA
DC
MN
OR

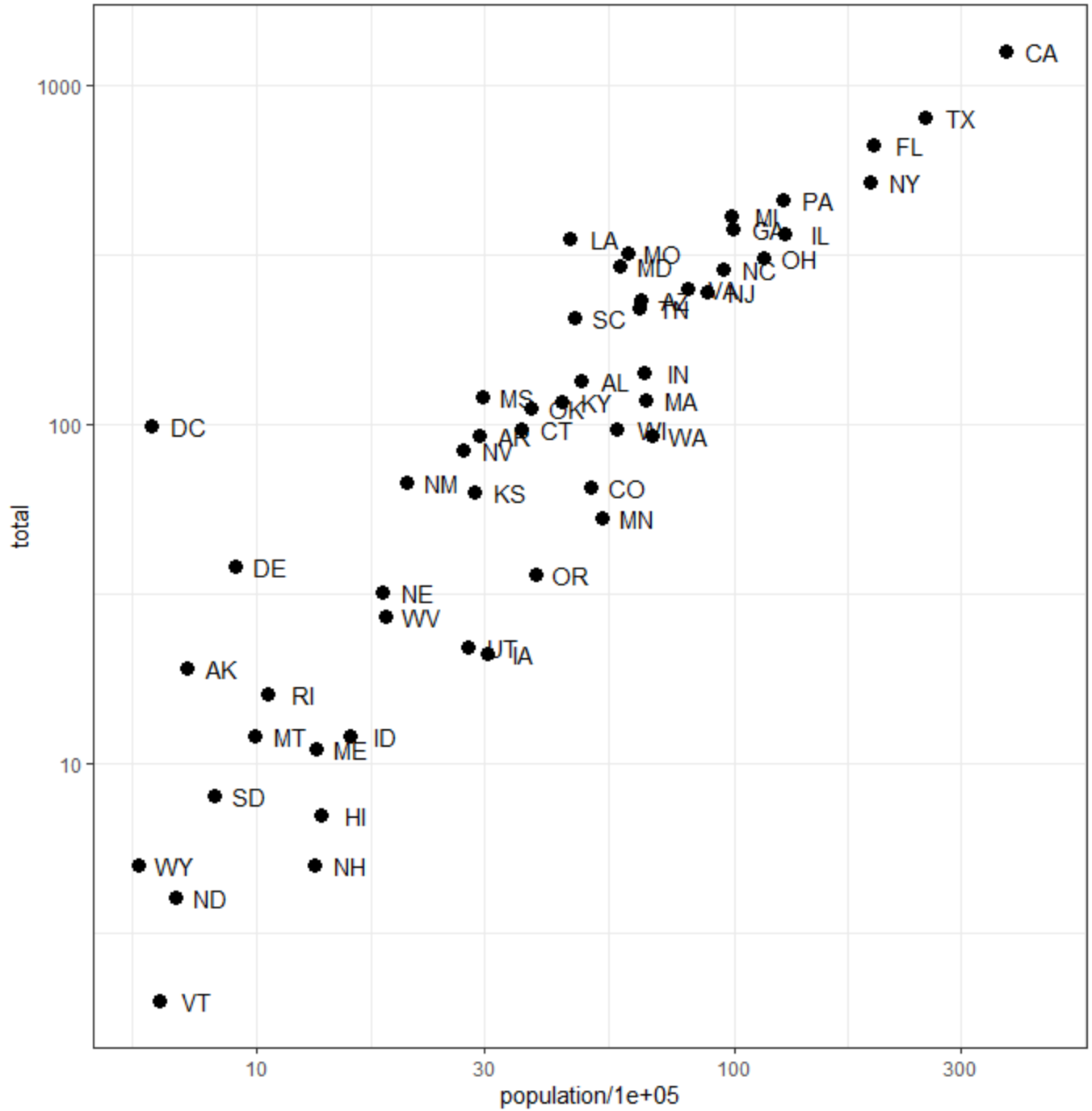
```
p +
geom_point(aes(x = population/100000, y = total), size = 3) +
geom_label(aes(x = population/100000, y = total, label = abb))
```


CA
TX
FL
NY
PA
MI
GA
IL
LA
MO
OH
NC
VA
SC
TN
DC
OK
WI
NM
WY
1200
800
400
0
total
0
100
200
300
population/1e+05

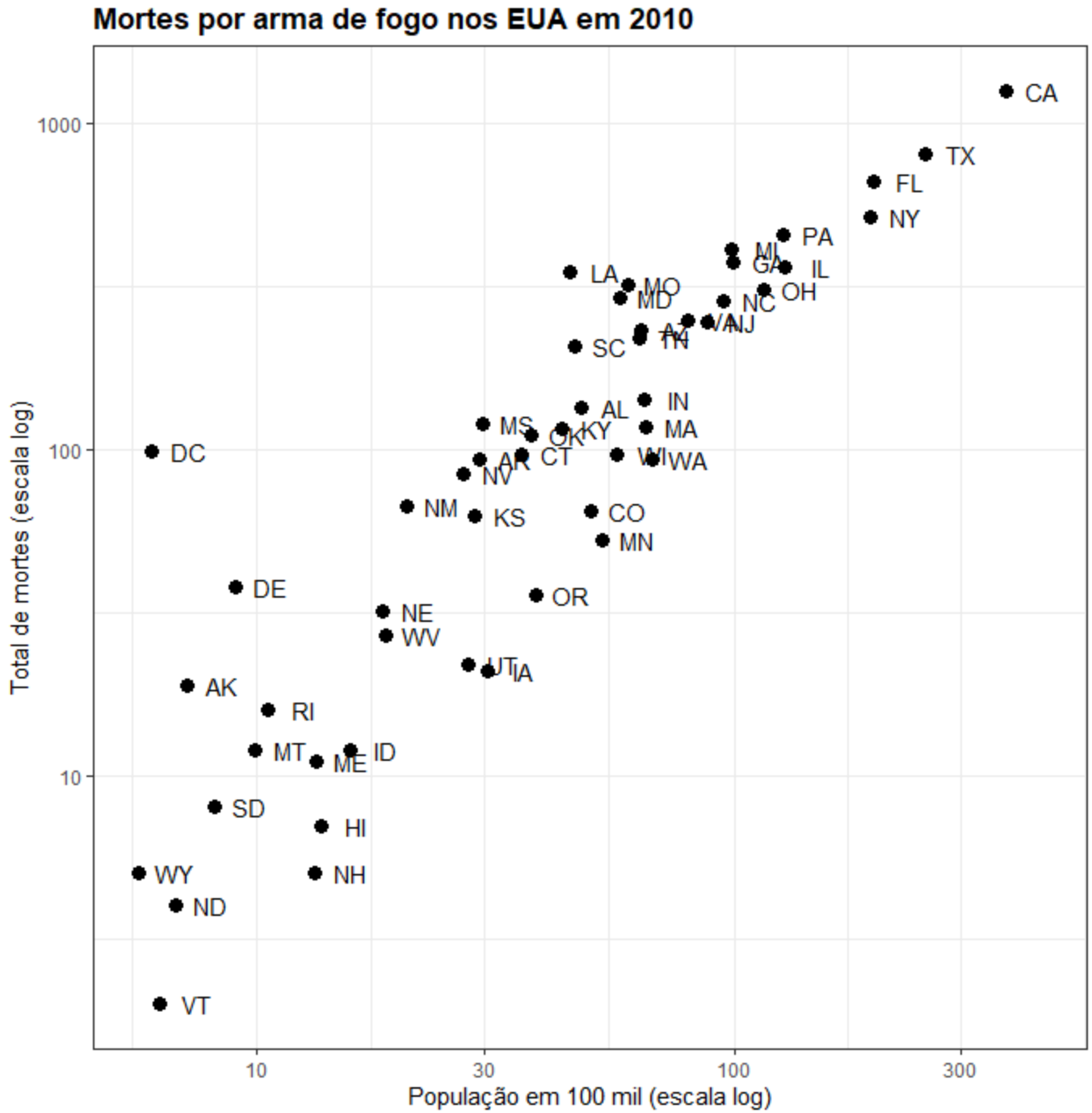
```
p +
geom_point(aes(x = population/100000, y = total), size = 3) +
geom_text(aes(x = population/100000, y = total, label = abb), nudge_x = 15)
```


CA
TX
FL
NY
PA
MI
GA
IL
LA
MO
OH
MD
NC
SC
IN
MA
WA
MN
OR
1200
800
400
0
total
0
100
200
300
400
population/1e+05

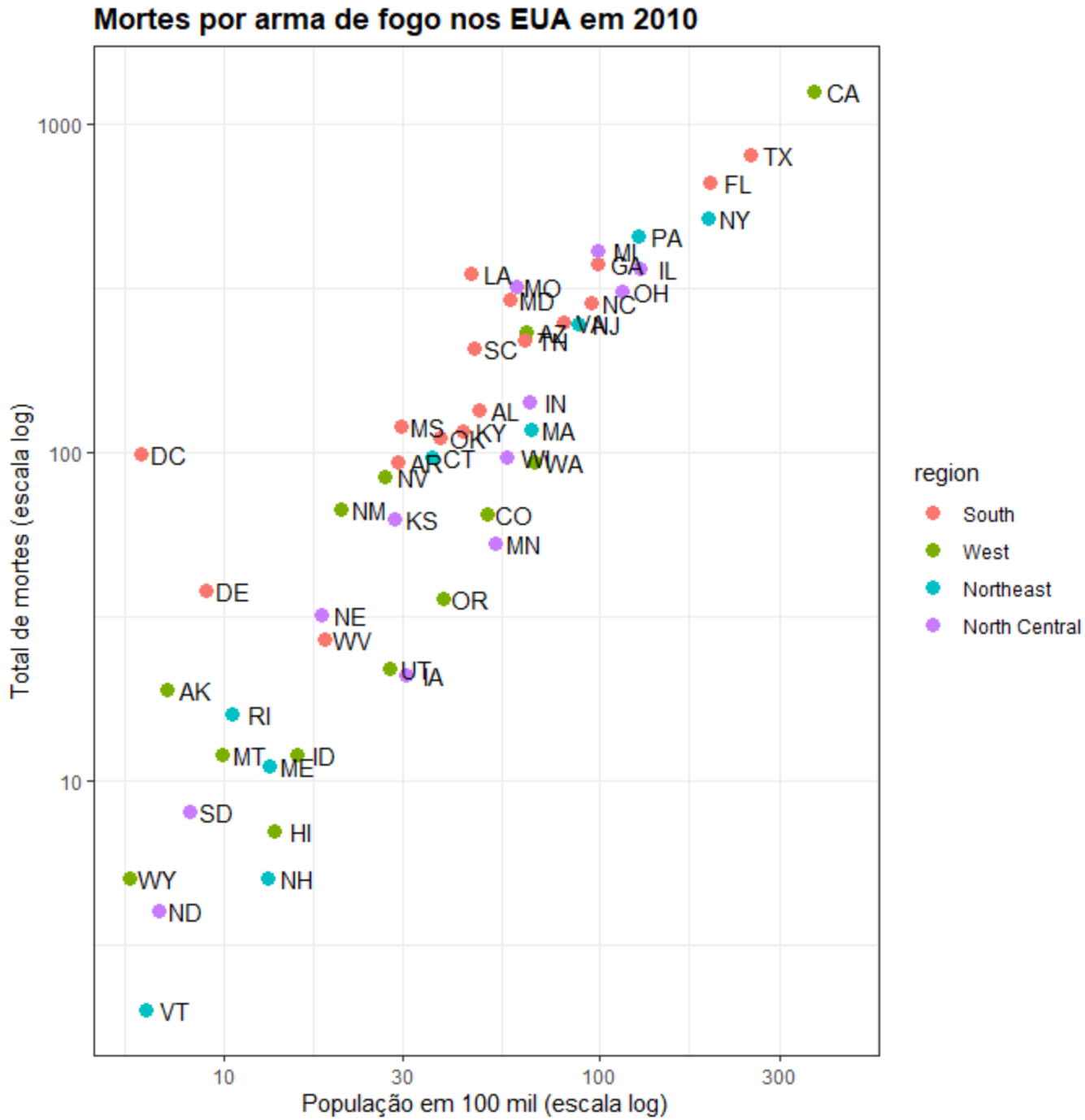
```
p +
geom_point(aes(x = population/100000, y = total), size = 3) +
geom_text(aes(x = population/100000, y = total, label = abb), nudge_x = 0.075) +
scale_x_log10() +
scale_y_log10()
```


CA
TX
FL
NY
PA
MI
GA
IL
LA
MO
OH
MD
NC
VA
NJ
AZ
TN
SC
IN
AL
MS
MA
OK
KY
DC
CT
WI
WA
AR
NV
NM
KS
CO
MN
DE
OR
NE
WV
UT
IA
AK
RI
MT
ID
ME
SD
HI
WY
NH
ND
VT
1000
100
10
total
10
30
100
300
population/1e+05

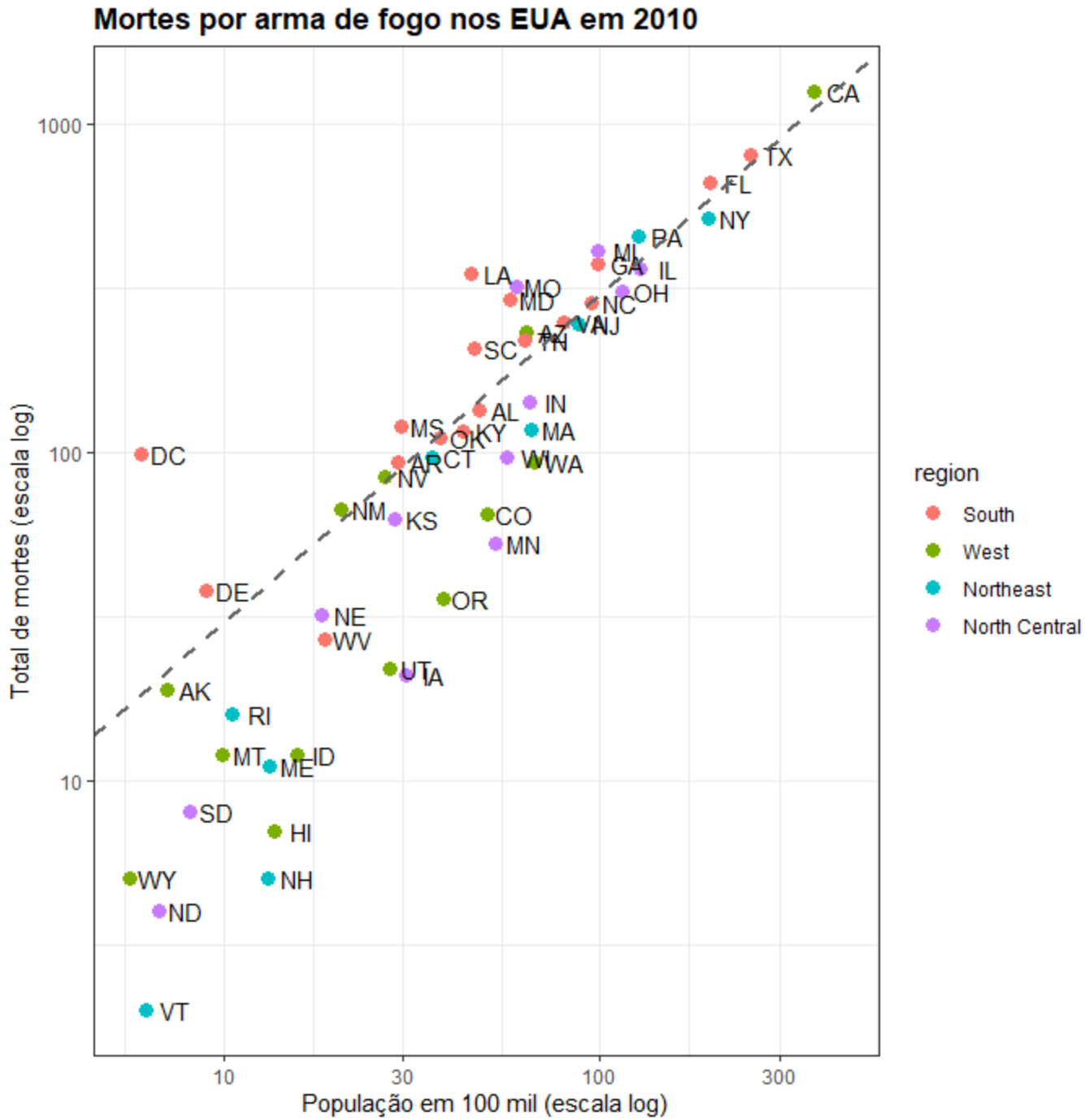
```
p +
geom_point(aes(x = population/100000, y = total), size = 3) +
geom_text(aes(x = population/100000, y = total, label = abb), nudge_x = 0.075) +
scale_x_log10() +
scale_y_log10() +
xlab("População em 100 mil (escala log)") +
ylab("Total de mortes (escala log)") +
ggtitle("Mortes por arma de fogo nos EUA em 2010")
```


Mortes por arma de fogo nos EUA em 2010
Total de mortes (escala log)
1000
100
10
CA
TX
FL
NY
PA
MI
GA
IL
LA
MO
OH
MD
NC
VA
NJ
AZ
TN
SC
AL
IN
MS
KY
MA
OK
DC
AR
CT
WI
WA
NV
NM
KS
CO
MN
DE
OR
NE
WV
UT
IA
AK
RI
MT
ID
ME
SD
HI
WY
NH
ND
VT
10
30
100
300
População em 100 mil (escala log)

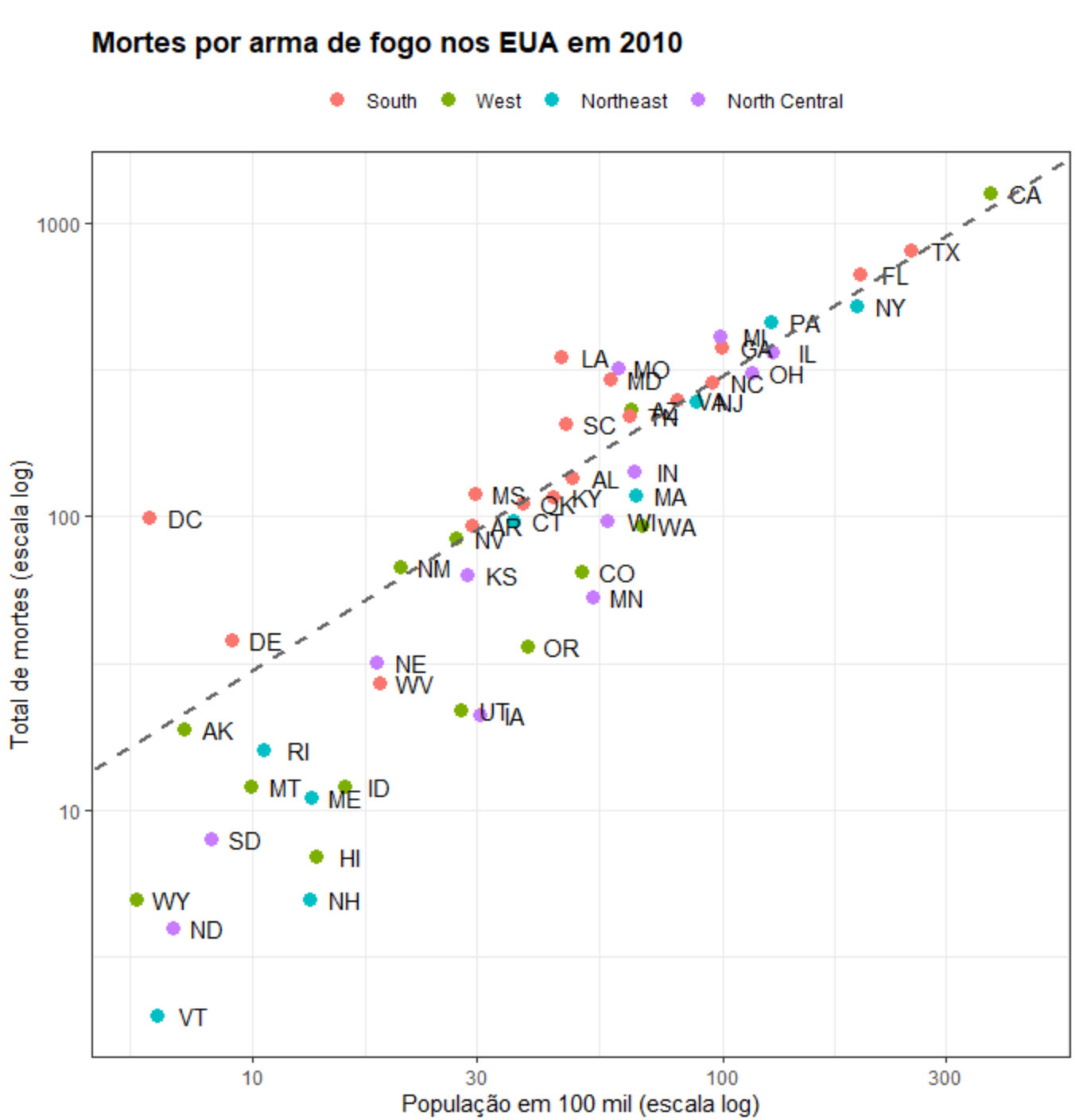
```
p +
geom_point(aes(x = population/100000, y = total, color = region), size = 3) +
geom_text(aes(x = population/100000, y = total, label = abb), nudge_x = 0.075) +
scale_x_log10() +
scale_y_log10() +
xlab("População em 100 mil (escala log)") +
ylab("Total de mortes (escala log)") +
ggtitle("Mortes por arma de fogo nos EUA em 2010")
```


Mortes por arma de fogo nos EUA em 2010
Total de mortes (escala log)
1000
100
10
CA
TX
FL
NY
PA
MI
GA
IL
LA
MO
OH
MD
NC
VA
NJ
AZ
TN
SC
IN
AL
MS
KY
MA
OK
CT
WI
WA
AR
DC
NV
NM
KS
CO
MN
DE
OR
NE
WV
UT
IA
AK
RI
MT
ID
ME
SD
HI
WY
NH
ND
VT
10
30
100
300
População em 100 mil (escala log)
region
South
West
Northeast
North Central

```
tx <- sum(data$total) / sum(data$population) * 100000


p +
geom_point(aes(x = population/100000, y = total, color = region), size = 3) +
geom_text(aes(x = population/100000, y = total, label = abb), nudge_x = 0.075) +
geom_abline(intercept = log10(tx), slope = 1,
            linetype = 'dashed’, linewidth = 1, color = 'gray40’) +
scale_x_log10() +
scale_y_log10() +
xlab("População em 100 mil (escala log)") +
ylab("Total de mortes (escala log)") +
ggtitle("Mortes por arma de fogo nos EUA em 2010")
```

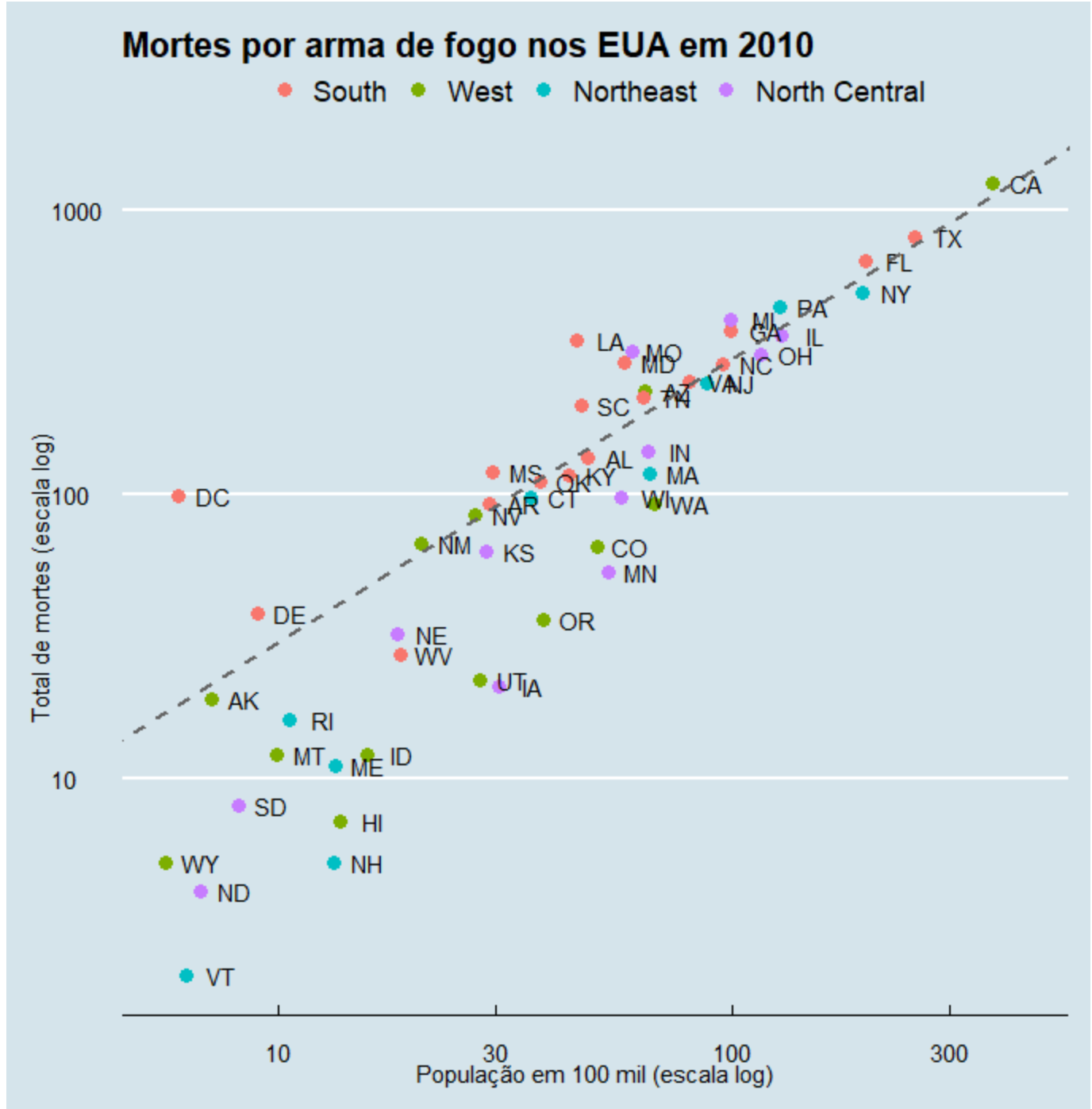

Mortes por arma de fogo nos EUA em 2010
Total de mortes (escala log)
1000
100
10
10
30
100
300
População em 100 mil (escala log)
CA
TX
FL
NY
PA
IL
OH
NC
MI
GA
NJ
VA
TN
AZ
LA
MO
MD
SC
IN
AL
MS
KY
OK
MA
WI
WA
CT
AR
NV
DC
NM
KS
CO
MN
DE
OR
NE
WV
UT
IA
AK
RI
MT
ID
ME
SD
HI
WY
NH
ND
VT
region
South
West
Northeast
North Central

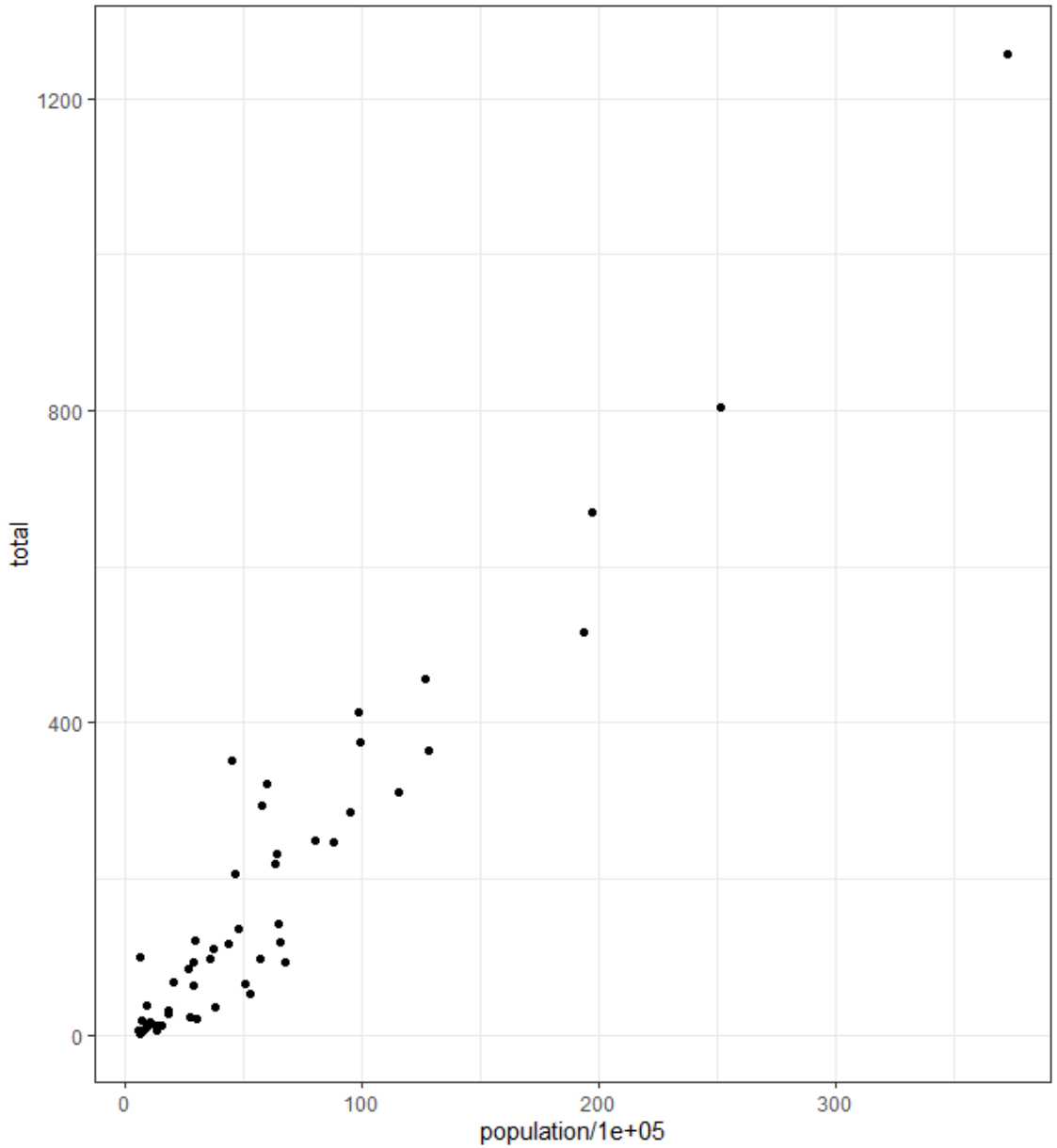
```
p +
geom_point(aes(x = population/100000, y = total, color = region), size = 3) +
geom_text(aes(x = population/100000, y = total, label = abb), nudge_x = 0.075) +
geom_abline(intercept = log10(tx), slope = 1,
            linetype = 'dashed’, linewidth = 1, color = 'gray40’) +
scale_x_log10() +
scale_y_log10() +
xlab("População em 100 mil (escala log)") +
ylab("Total de mortes (escala log)") +
ggtitle("Mortes por arma de fogo nos EUA em 2010") +
theme(legend.title = element_blank(), legend.position = 'top')
```

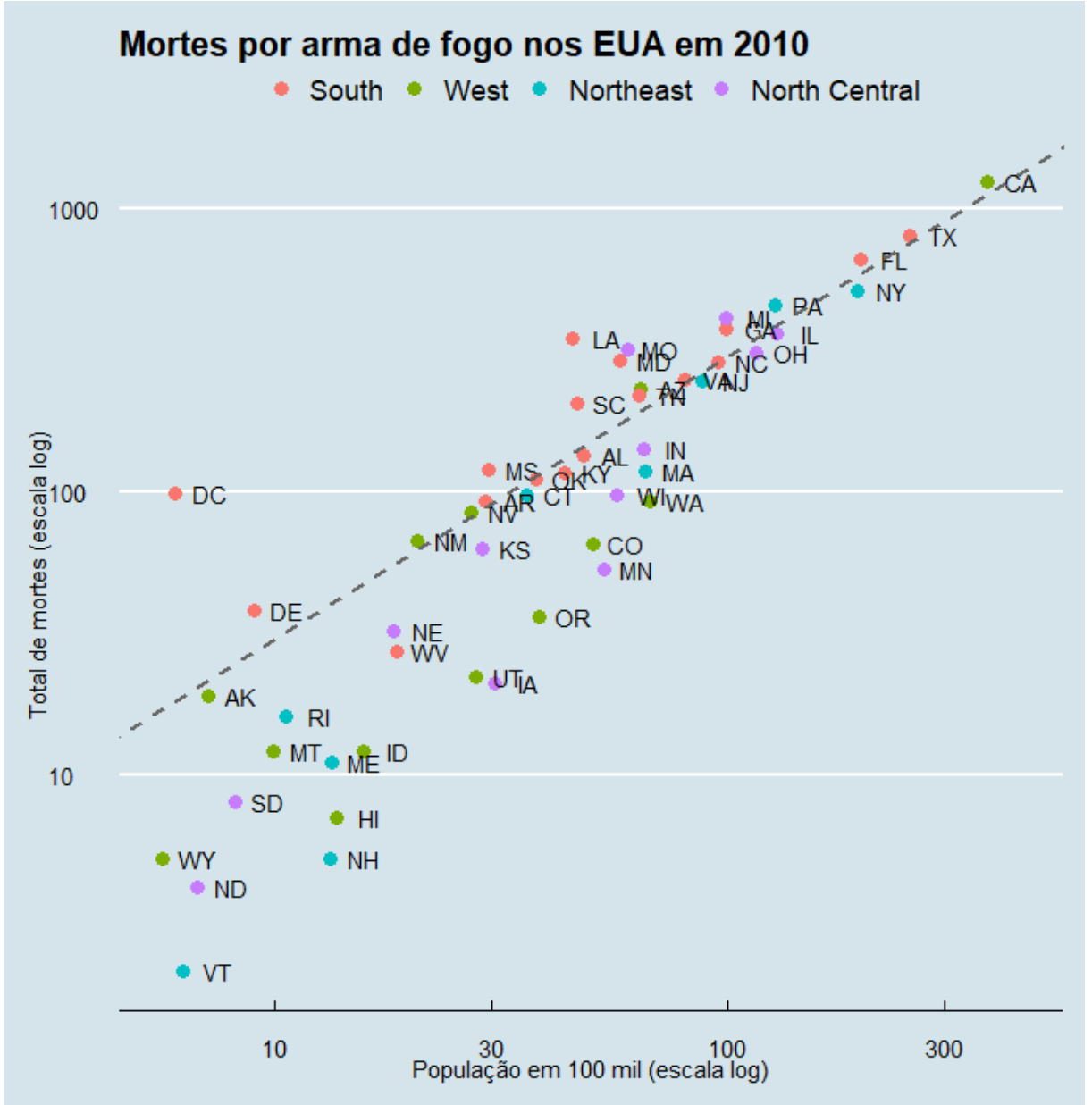

Mortes por arma de fogo nos EUA em 2010
South
West
Northeast
North Central
Total de mortes (escala log)
1000
100
10
DC
DE
AK
RI
MT
ME
ID
SD
HI
WY
NH
ND
VT
NE
WV
UT
IA
OR
NV
NM
KS
CO
MN
MS
OK
KY
MA
AR
CT
WI
WA
AL
IN
SC
LA
MO
MD
NC
OH
GA
IL
MI
PA
NJ
VA
NY
FL
TX
CA
10
30
100
300
População em 100 mil (escala log)

```
install.packages(“ggthemes”)

library(ggthemes)
```

```
p +
geom_point(aes(x = population/100000, y = total, color = region), size = 3) +
geom_text(aes(x = population/100000, y = total, label = abb), nudge_x = 0.075) +
geom_abline(intercept = log10(tx), slope = 1,
            linetype = 'dashed’, linewidth = 1, color = 'gray40’) +
scale_x_log10() +
scale_y_log10() +
xlab("População em 100 mil (escala log)") +
ylab("Total de mortes (escala log)") +
ggtitle("Mortes por arma de fogo nos EUA em 2010") +
theme_economist() +
theme(legend.title = element_blank(), legend.position = 'top')
```

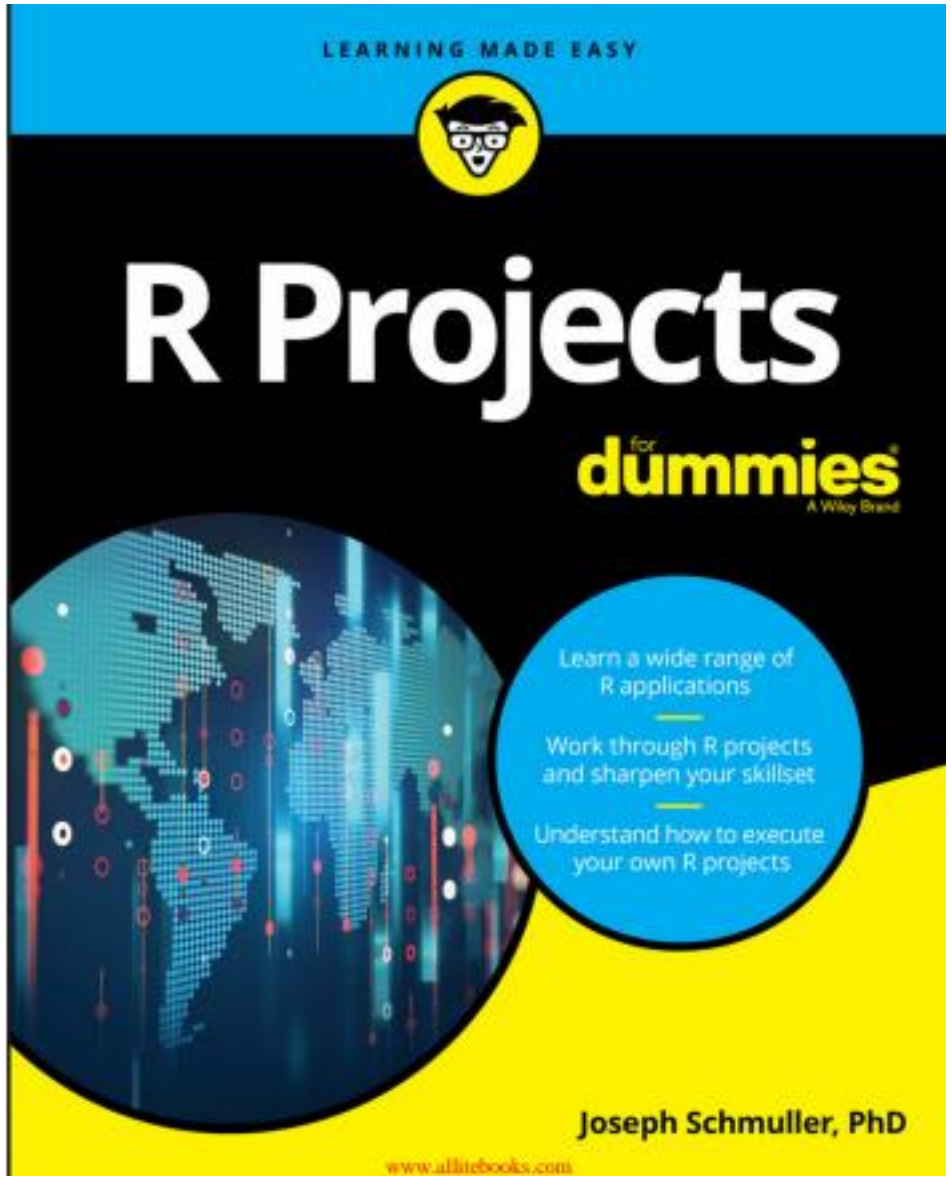

Mortes por arma de fogo nos EUA em 2010
South
West
Northeast
North Central
1000
100
10
Total de mortes (escala log)
CA
TX
FL
NY
PA
IL
OH
NC
MI
GA
LA
MO
MD
VA
NJ
AZ
TN
SC
AL
IN
MS
KY
OK
MA
WI
WA
AR
CT
NV
NM
KS
CO
MN
DC
DE
OR
NE
WV
UT
IA
AK
RI
MT
ID
ME
SD
HI
WY
NH
ND
VT
10
30
100
300
População em 100 mil (escala log)

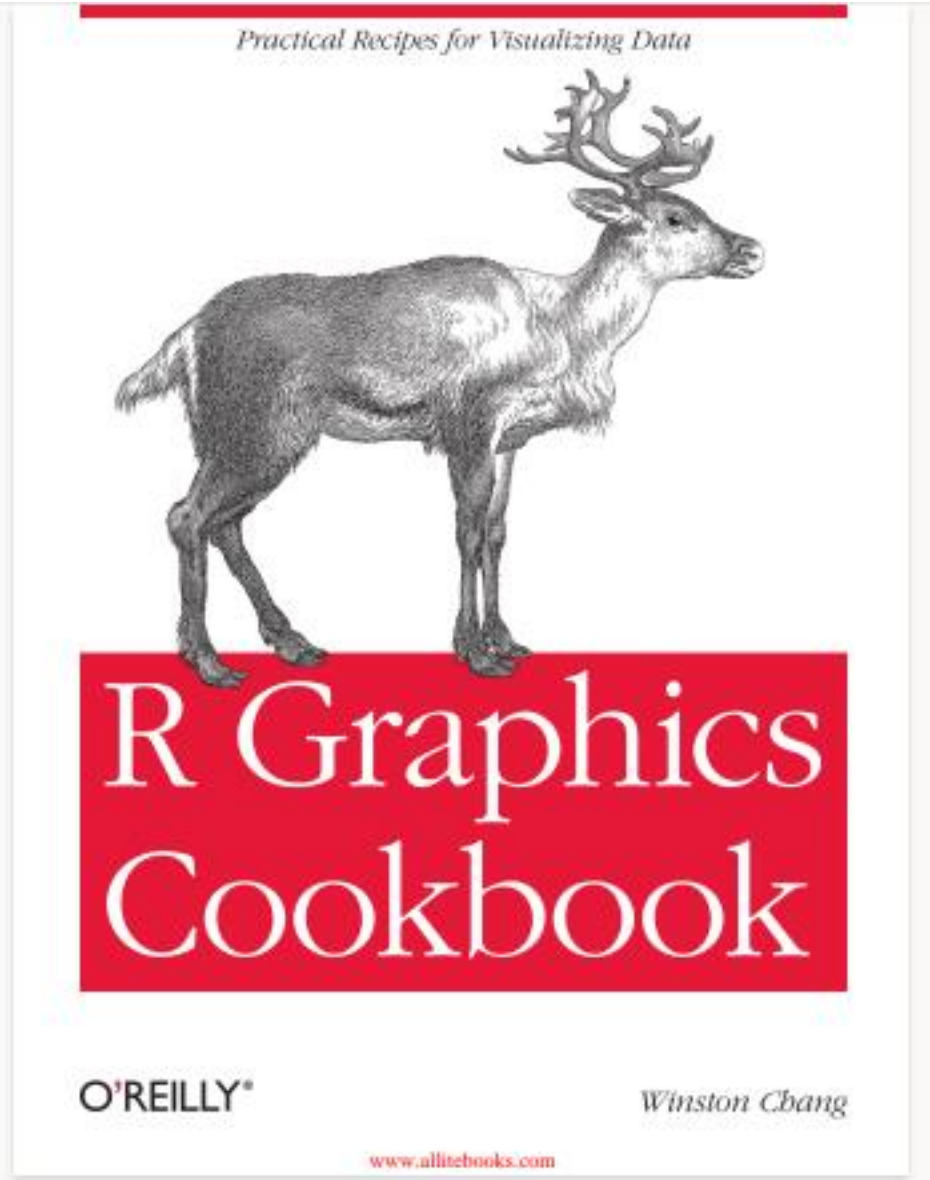

1200
800
400
0
total
0
100
200
300
population/1e+05


Mortes por arma de fogo nos EUA em 2010
South
West
Northeast
North Central
1000
100
10
Total de mortes (escala log)
CA
TX
FL
NY
PA
IL
OH
NC
MI
GA
LA
MO
MD
AZ
TN
VA
NJ
SC
IN
AL
MA
MS
KY
OK
AR
CT
WI
WA
NV
NM
KS
CO
MN
DC
DE
OR
NE
WV
UT
IA
AK
RI
MT
ME
ID
SD
HI
WY
NH
ND
VT
10
30
100
300
População em 100 mil (escala log)

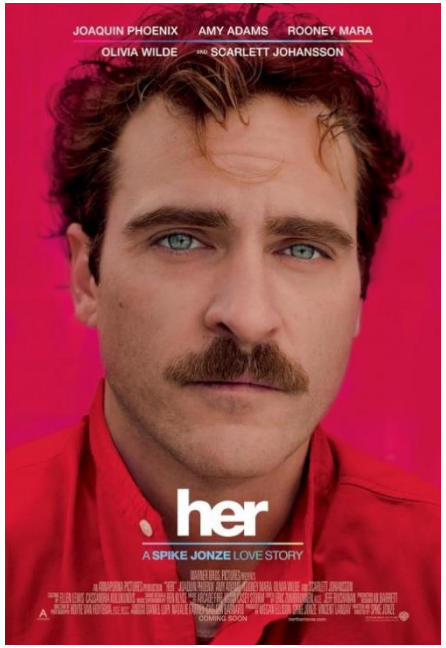
```
ggsave('meu.primeiro.plot.png')
```

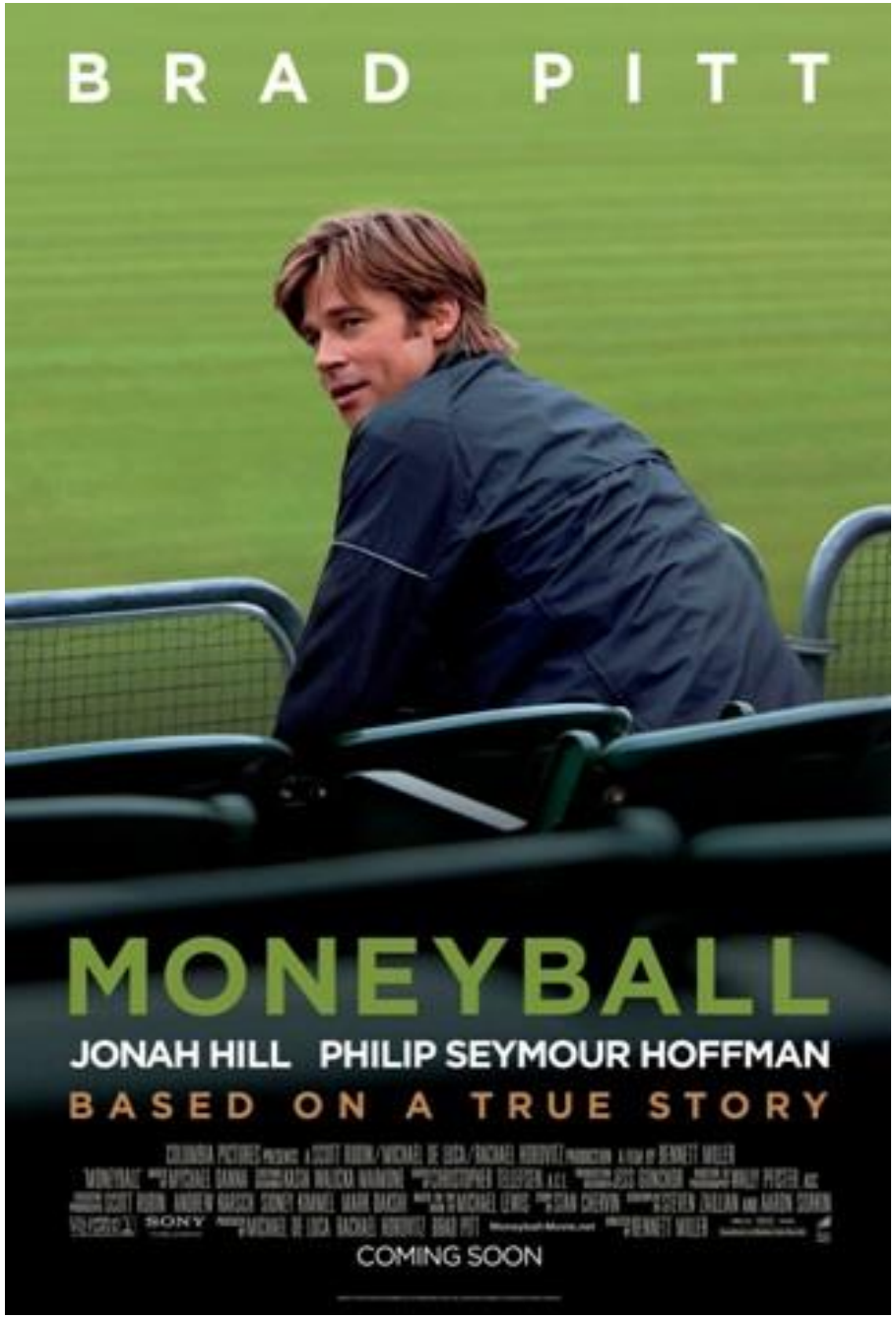
Outras fontes de inspiração...


LEARNING MADE EASY
R Projects
for dummies
A Wiley Brand
Learn a wide range of R applications
Work through R projects and sharpen your skillset
Understand how to execute your own R projects
Joseph Schmuller, PhD
www.allitebooks.com


Practical Recipes for Visualizing Data
R Graphics Cookbook
O'REILLY
Winston Chang
www.allitebooks.com


21
SPRING 2008


JOAQUIN PHOENIX
AMY ADAMS
ROONEY MARA
OLIVIA WILDE
AND SCARLETT JOHANSSON
her
A SPIKE JONZE LOVE STORY


BRAD PITT
MONEYBALL
JONAH HILL
PHILIP SEYMOUR HOFFMAN
BASED ON A TRUE STORY
COMING SOON


THEY TOOK YOUR DATA.
THEN THEY TOOK CONTROL.
A NETFLIX ORIGINAL DOCUMENTARY
THE GREAT HACK
From the Academy Award® nominated filmmakers behind THE SQUARE
JULY 24 | NETFLIX


CHRISTIAN BALE
STEVE CARELL
RYAN GOSLING
BRAD PITT
THE BIG SHORT
THIS IS A TRUE STORY
FROM THE ACTOR OF
"MONEYBALL"
AND
"THE BLIND SIDE"
COMING SOON

OBRIGADO!

Nelson Quesado
Caio Gustavo

Introdução à Ciência de Dados com R