A Minimal Book Example

John Doe

2023-07-14

Contents

Chapter 1

About

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports; for example, a math equation $a^2 + b^2 = c^2$.

1.1 Usage

Each **bookdown** chapter is an .Rmd file, and each .Rmd file can contain one (and only one) chapter. A chapter *must* start with a first-level heading: `# A good chapter`, and can contain one (and only one) first-level heading.

Use second-level and higher headings within chapters like: `## A short section` or `### An even shorter section`.

The `index.Rmd` file is required, and is also your first book chapter. It will be the homepage when you render the book.

1.2 Render book

You can render the HTML version of this example book without changing anything:

1. Find the **Build** pane in the RStudio IDE, and
2. Click on **Build Book**, then select your output format, or select "All formats" if you'd like to use multiple formats from the same book source files.

Or build the book from the R console:

```
bookdown::render_book()
```

To render this example to PDF as a `bookdown::pdf_book`, you'll need to install XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.org/tinytex/.

1.3 Preview book

As you work, you may start a local server to live preview this HTML book. This preview will update as you edit the book when you save individual .Rmd files. You can start the server in a work session by using the RStudio add-in "Preview book", or from the R console:

```
bookdown::serve_book()
```

Chapter 2

R

2.1 R?

R — . 90- .
. , .
, , .
R — . , ,
, . R
, .

—
R .
, R , ,
. , R .

2.2

R – . –
, . 1 2023 . CRAN 19789
. : GitHub, Dracor,
.

, , – ,
, “ ” - .

, R ,
, . , Shiny
- , Leaflet .
Antibarbari HSE.
. R , .

“ ” , . ,
RPerseus, .
“ ”, , “ ” R. , ,
tidyverse, “ ” . .

2.3

. (SPSS, Minitab),
, “ ” , . R,
, .
– (.R)
. , ,
.

2.4

, (,
GitHub) – , . ,
(Python)

```
#
x <- rnorm(1000)
#
y <- sample(x, 100)
```

, , , ( ), .
, , , . “ ”
, , ,
. , .
, ,
( ). ,
, . !
, , .

2.5 ( ) ?

R , . “ ”
( ) : .
, .
(R).
, . -
.

24 :

- R (1-6)
- text-mining (7-13)
- (14-22)

Plotly Leaflet.

, ,
“ : ?” (, 2023 .).
, . , – .

2.6 RStudio

R RStudio, (IDE)
R.

– R R Studio , ;
, .

2.7

1. R

- R Windows: https://cran.r-project.org/bin/windows/
- R Mac: https://cran.r-project.org/bin/macosx/

2. R Studio

- : https://www.rstudio.com/products/rstudio/download/ (
)

MacOS Stylo XQuartz: https://www.xquartz.org/

2.8

RStudio ( ):

, ( ).

, :

```
sessionInfo()
```

`sessionInfo()` – . ,
. (“ - !”). – -
. (, “ ” tidyverse - , .)
.

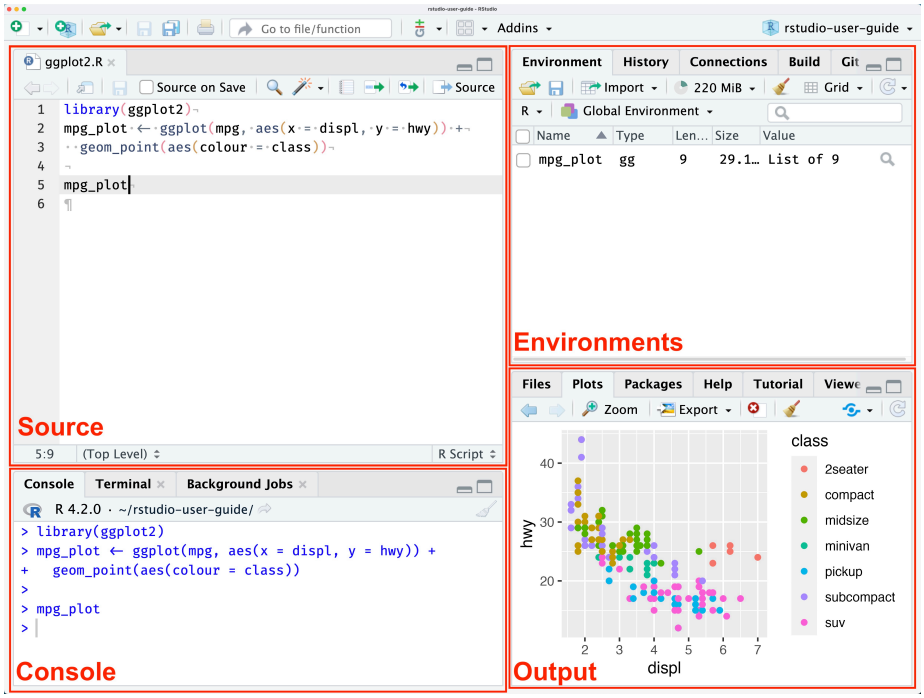


Figure 2.1: RStudio Panes

, , **help**: `?mean()`. ( )
.
(R) `getwd()`
. `setwd()`,
(,). :

```
setwd("/Users/olga/R_Workflow/")
```

R `Session > Set Working Directory`.

, . ,
`Tools > Install Packages`. .
:

```
install.packages("languageR")
```

`library()`, :

```
library(languageR)
```

:

- R [1]
-

2.9 R

. >, **prompt**.

```
sqrt(4) #
```

```
## [1] 2
```

```
2^3 #
```

```
## [1] 8
```

```
log10(100) #
```

```
## [1] 2
```

+, . , .
. `sqrt(2` .

2.10

, -
`<-` (`Alt` + `-` (Windows) `Option` + `-` (Mac)). `=`
, , .

[1] https://intro2r.com/rsprojs.html

```
x <- 2 + 2 #
y <- 0.1 #
x <- y #
x + y
```

```
## [1] 0.2
```

, , , . c() (concatenation) :

```
x <- c(3, 5, 7)
x_mean <- mean(x) #          x.mean    xMean
x_mean
```

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

tidyverse , ; Python, . : R !

, , _ , , , , ,
. _ ,
[?, 24]

, ls(). rm(). :

```
rm(list = ls()) #
```

2.11

R ( ). .

```
x <- 2
class(x) #
```

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

```
length(x) #
```

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

```
y <- c() #
y #                    NULL
```

```
## NULL
```

```
length(y) #        0
```

```
## [1] 0
```

`NULL` , ; `NA` (not available) – , .
`mean(c(1, NA, 2))` , `mean(c(1, NULL, 2))` .
: `mean(c(1, NA, 2), na.rm=T)`. . [?].

, , :

- (integer)
- (numeric, double,)
- (character)
- (logical)
- , (factor)

```
#
x <- sqrt(2)
typeof(x)
```

```
## [1] "double"
```

```
is.integer(x)
```

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

```
is.numeric(x)
```

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

, :

```
x <- c(TRUE, 1, 3, FALSE)
x #
```

```
## [1] 1 1 3 0
```

```
y <- c(1, "a", 2, "     ") #
y #
```

```
## [1] "1"          "a"          "2"          "     "
```

: **R**

(== “ ”, != “ ”, <= “ ”) :

```
x <- c(1:10) #       1   10
y <- x > 5
y #       TRUE             ,
```

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

```
sum(y)
```

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

`all()` `any()` :

```
x <- 10:20
any(x == 15)
```

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

```
all(x > 9)
```

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

:

```
seq(1, 5, 0.5)
```

```
## [1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
```

```
rep("foo", 5)
```

```
## [1] "foo" "foo" "foo" "foo" "foo"
```

, - :

```
x <- seq(1, 5, 0.5)
x[4:5] #          1 (      Python)
```

```
## [1] 2.5 3.0
```

, , :
, ,
.

```
x <- 2; y <- c(10, 20, 30); z <- c(5, 6, 7)
y / x
```

```
## [1]  5 10 15
```

```
x + y
```

```
## [1] 12 22 32
```

```
y + z
```

```
## [1] 15 26 37
```

. ,
. (, 1 0), .

```
t <- factor(c("A", "B", "C"), levels = c("A", "B", "C"))
t
```

```
## [1] A B C
## Levels: A B C
```

2.12

, ( ), .

```
list = list(a = c("a", "b", "c"), b = c(1, 2, 3), c = c(T, F, T))
list
```

```
## $a
## [1] "a" "b" "c"
##
## $b
## [1] 1 2 3
##
## $c
## [1]  TRUE FALSE  TRUE
```

, .

```
list$a #
```

```
## [1] "a" "b" "c"
```

```
list[2] #
```

```
## $b
## [1] 1 2 3
```

```
class(list[2])
```

```
## [1] "list"
```

```
list[[2]] #
```

```
## [1] 1 2 3
```

```
class(list[[2]])
```

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

```
list$c[1]#
```

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

, list[2] list[[2]] .
XML.

: **R**

, .

2.13

– , : . ,
, , . .

```
M = matrix(c(1, 2, 3, 4), nrow = 2)
M #
```

```
##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4
```

```
M = matrix(c(1, 2, 3, "a"), nrow = 2)
M #               !
```

```
##      [,1] [,2]
## [1,] "1"  "3"
## [2,] "2"  "a"
```

. ( ) . 2 x
2. , , . ,
.

```
M = matrix(c(1, 2, 3, 4), nrow = 2)
M[1, ] #
```

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

```
M[,2] #
```

```
## [1] 3 4
```

```
M[1,1] #
```

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

, .

```
M = matrix(c(1, 2, 3, 4), nrow = 2)
class(M)
```

```
## [1] "matrix" "array"
```

```
dim(M) #
```

```
## [1] 2 2
```

```
class(M[1, ]) #
```

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

```
dim(M[1, ])
```

```
## NULL
```

NULL, R
, . , ,
. , . .

```
#    -
C = matrix(c(1, 2, 3), nrow = 1)
C
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
```

```
#    -
D = matrix(c(1, 2, 3), nrow = 3)
D
```

```
##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
```

R ; ,
- .

.

```
#
M = matrix(c(1, 2, 3, 4), nrow = 2) #
diag(M)
```

```
## [1] 1 4
```

```
#                ,
t(M)
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

```
#                ,              .
M * 3
```

```
##      [,1] [,2]
## [1,]    3    9
## [2,]    6   12
```

```
#
M + M
```

```
##      [,1] [,2]
## [1,]    2    6
## [2,]    4    8
```

. , . ,
, .
: , .
R . [?].

2.14

(, data frames) – ( ).
, .

,
.

— [?, 133].

```
#
df <- data.frame(names = c("A", "B"), age = c(10, 11))
df
```

```
##   names age
## 1     A  10
## 2     B  11
```

```
#
df$names #
```

```
## [1] "A" "B"
```

```
df[,"names"] #        ,
```

```
## [1] "A" "B"
```

```
df[1, ] #
```

```
##   names age
## 1     A  10
```

.

– , .

, - .
1867 ., , , “ ”
, , [?, xxxi].

26 , .
, R.

```
##        dialogue  words hapax ratio group
## 1       Apology   8745    36 0.004     1
## 2     Charmides   8311    31 0.004     1
```

```
## 3      Cratylus  17944   122 0.007     1
## 4       Critias   4950   104 0.021     3
## 5         Crito   4169    19 0.005     1
## 6    Euthydemus  12453    87 0.007     1
## 7     Euthyphro   5181    15 0.003     1
## 8       Gorgias  26337   125 0.005     1
## 9  HippiasMinor   4360    12 0.003     1
## 10          Ion   4024    32 0.008     1
## 11       Laches   7674    27 0.004     1
## 12         Laws 103193   914 0.009     3
## 13        Lysis   6980    49 0.007     1
## 14    Menexenus   4808    43 0.009     1
## 15         Meno   9791    30 0.003     1
## 16   Parmenides  15155    20 0.001     2
## 17       Phaedo  21825   140 0.006     1
## 18     Phaedrus  16645   228 0.014     2
## 19     Philebus  17668    64 0.004     3
## 20   Protagoras  17795   102 0.006     1
## 21     Republic  88878   668 0.008     2
## 22      Sophist  16024   107 0.007     3
## 23    Statesman  16953   180 0.011     3
## 24    Symposium  17461   127 0.007     1
## 25   Theaetetus  22489   162 0.007     2
## 26      Timaeus  23662   370 0.016     3
```

. class() , .

```
## [1] "data.frame"
```

.

```
#
colnames(hapax_plato)
```

```
## [1] "dialogue" "words"    "hapax"    "ratio"    "group"
```

```
#       ()
hapax_plato[hapax_plato$dialogue == "Parmenides", ]
```

```
##      dialogue words hapax ratio group
## 16 Parmenides 15155    20 0.001     2
```

```
#
str(hapax_plato)
```

```
## 'data.frame':    26 obs. of  5 variables:
##  $ dialogue: chr  "Apology" "Charmides" "Cratylus" "Critias" ...
##  $ words   : chr  "8745" "8311" "17944" "4950" ...
##  $ hapax   : chr  "36" "31" "122" "104" ...
```

```
##  $ ratio   : chr  "0.004" "0.004" "0.007" "0.021" ...
##  $ group   : num  1 1 1 3 1 1 1 1 1 1 ...
```

```
#
hapax_plato[hapax_plato$words > 10000, ]
```

```
##        dialogue  words hapax ratio group
## 1       Apology   8745    36 0.004     1
## 2     Charmides   8311    31 0.004     1
## 3      Cratylus  17944   122 0.007     1
## 4       Critias   4950   104 0.021     3
## 5         Crito   4169    19 0.005     1
## 6    Euthydemus  12453    87 0.007     1
## 7     Euthyphro   5181    15 0.003     1
## 8       Gorgias  26337   125 0.005     1
## 9  HippiasMinor   4360    12 0.003     1
## 10          Ion   4024    32 0.008     1
## 11       Laches   7674    27 0.004     1
## 12         Laws 103193   914 0.009     3
## 13        Lysis   6980    49 0.007     1
## 14    Menexenus   4808    43 0.009     1
## 15         Meno   9791    30 0.003     1
## 16   Parmenides  15155    20 0.001     2
## 17       Phaedo  21825   140 0.006     1
## 18     Phaedrus  16645   228 0.014     2
## 19     Philebus  17668    64 0.004     3
## 20   Protagoras  17795   102 0.006     1
## 21     Republic  88878   668 0.008     2
## 22      Sophist  16024   107 0.007     3
## 23    Statesman  16953   180 0.011     3
## 24    Symposium  17461   127 0.007     1
## 25   Theaetetus  22489   162 0.007     2
## 26      Timaeus  23662   370 0.016     3
```

```
#
hapax_plato$group <- as.factor(hapax_plato$group)
hapax_plato[,2:4] <- sapply(hapax_plato[,2:4],as.numeric)
```

summary():

```
summary(hapax_plato)
```

```
##    dialogue             words            hapax            ratio          group
##  Length:26          Min.   :  4024   Min.   : 12.00   Min.   :0.001000   1:16
##  Class :character   1st Qu.:  7154   1st Qu.: 31.25   1st Qu.:0.004000   2: 4
##  Mode  :character   Median : 15590   Median : 94.50   Median :0.007000   3: 6
##                     Mean   : 19364   Mean   :146.69   Mean   :0.007154
##                     3rd Qu.: 17907   3rd Qu.:136.75   3rd Qu.:0.008000
```

```
##                    Max.   :103193   Max.   :914.00   Max.   :0.021000
```

2.15

.

Chapter 3

3.1 R

R , .
R – , [1].

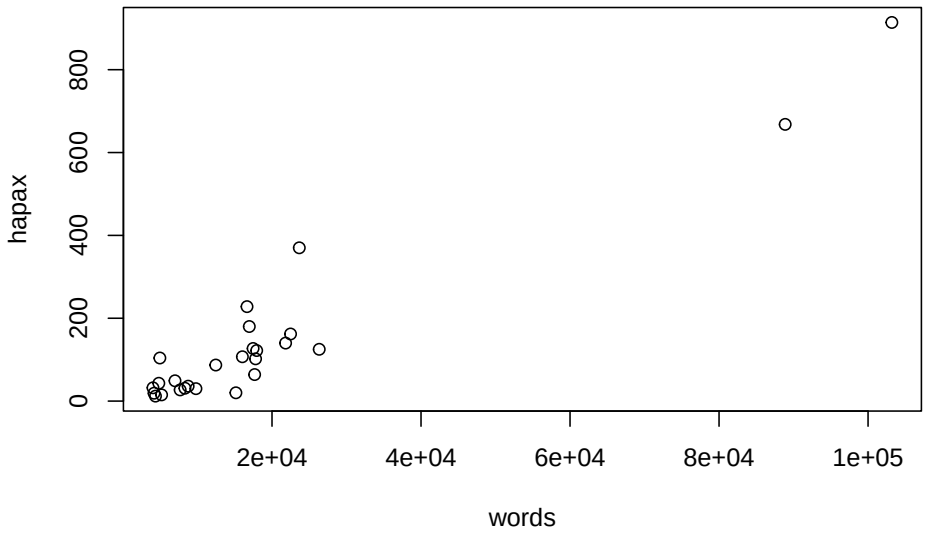
, , - : ,
, , , . .
.

, , , ,
. plot(), , .
, .

(scatter plot), plot()
, x y. : plot(x, y). :
plot(y ~ x). ~ ( ) .

```
attach(hapax_plato)
plot(hapax ~ words)
```

[1] https://youtu.be/a4mvbyNGdBA

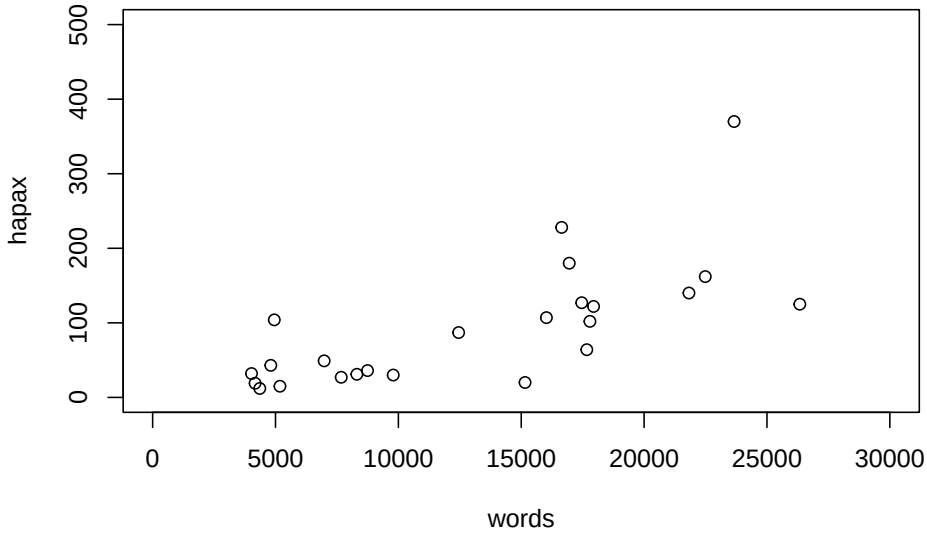


: plot(hapax_plato$hapax ~ hapax_plato$words).

. .

(“ ” “ ”) “ ” , .

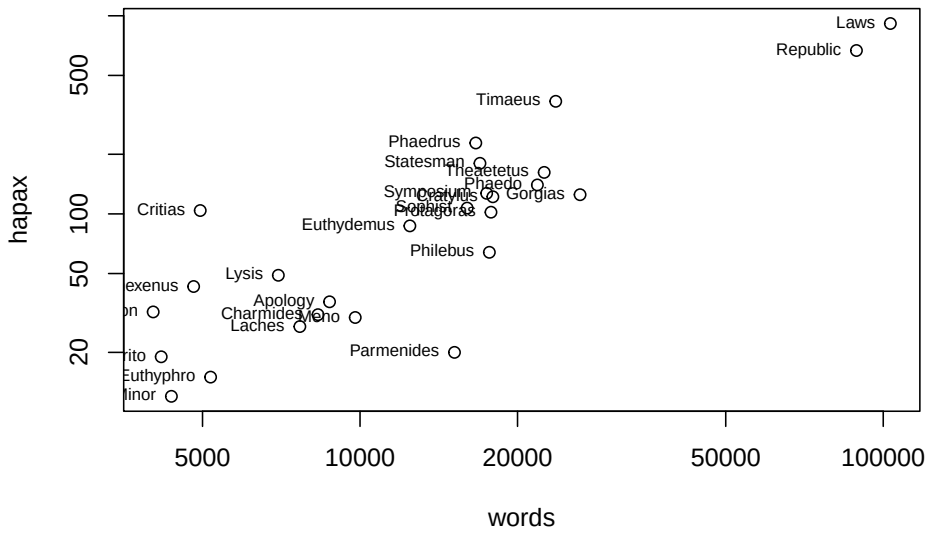
```
attach(hapax_plato)
plot(hapax ~ words, xlim = c(0, 30000), ylim = c(0, 500))
```



- - – ? –

. . , .

```
attach(hapax_plato)
options(scipen=999) #
plot(words, hapax, log = "xy")
#
text(hapax ~ words, labels = dialogue, pos = 2, cex = 0.7)
```



! .

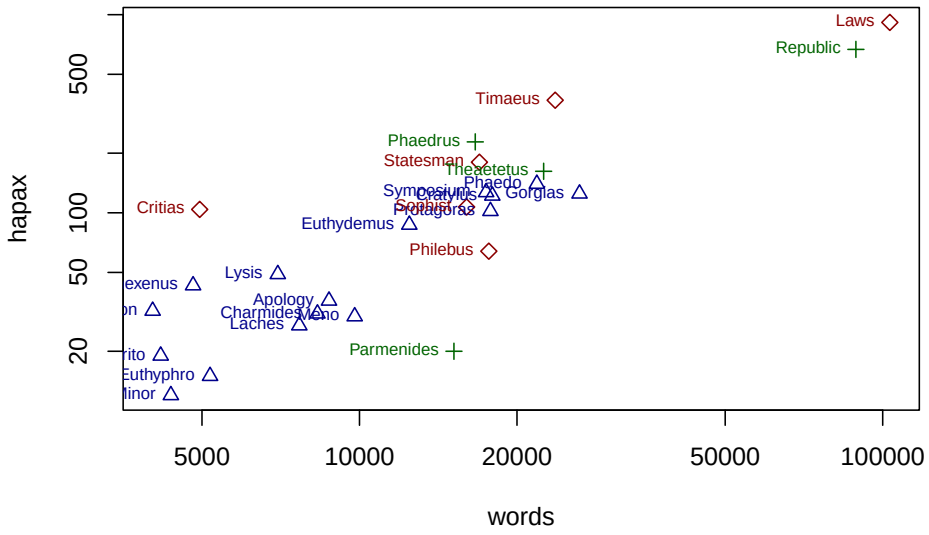
plot() pch. .

2, 3 5.

: **pch**

.

```
attach(hapax_plato)
options(scipen=999) #
plot(words, hapax, log = "xy", col = c("darkblue", "darkgreen", "darkred")[group],
     pch = c(2, 3, 5)[group])
text(hapax ~ words, labels = dialogue,
     pos = 2, cex = 0.7, col = c("darkblue", "darkgreen", "darkred")[group])
```



c , . , ,
(, -).

```
attach(hapax_plato)
options(scipen=999) #
plot(words, hapax, log = "xy", col = c("darkblue", "darkgreen", "darkred")[group], pch
text(hapax ~ words, labels = dialogue,
     pos = 2, cex = 0.7, col = c("darkblue", "darkgreen", "darkred")[group], family = '

#
my_lm <- lm(hapax_plato$hapax ~ hapax_plato$words)
abline(my_lm, lty = "dashed", col = "darkgrey", untf = T)

#
title(main = "                              ")
```

```
## Warning in title(main = "                              "):
## conversion failure on '                              ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                              "):
## conversion failure on '                              ' in
## 'mbcsToSbcs': dot substituted for <a7>

## Warning in title(main = "                              "):
## conversion failure on '                              ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                              "):
## conversion failure on '                              ' in
## 'mbcsToSbcs': dot substituted for <b8>
```

```
## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d1>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <81>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <bb>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <be>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b3>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <bf>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>
```

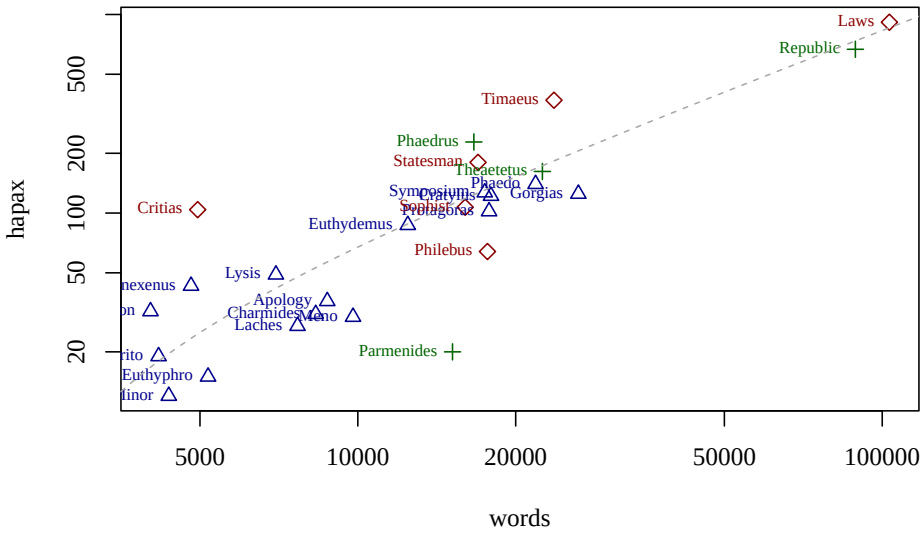
```
## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <ba>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d1>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <81>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <be>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b2>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b2>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b7>
```

```
## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b2>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b8>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d1>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <81>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b8>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <bc>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>
```

```
## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <be>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d1>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <81>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d1>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <82>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b8>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <be>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d1>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <82>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b4>
```

```
## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <bb>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b8>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <bd>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d1>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <8b>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b4>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b8>

## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>
```

```
## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b0>


## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>


## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <bb>


## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>


## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <be>


## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>


## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b3>


## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <d0>


## Warning in title(main = "                          "):
## conversion failure on '                          ' in
## 'mbcsToSbcs': dot substituted for <b0>
```



[2] . .

, .

, R , ,

.

, . , ,

“ ” ? , .

, . ,

.

3.2 Lattice

Lattice (. “ ”) [?].

, -

. (– –). , ,

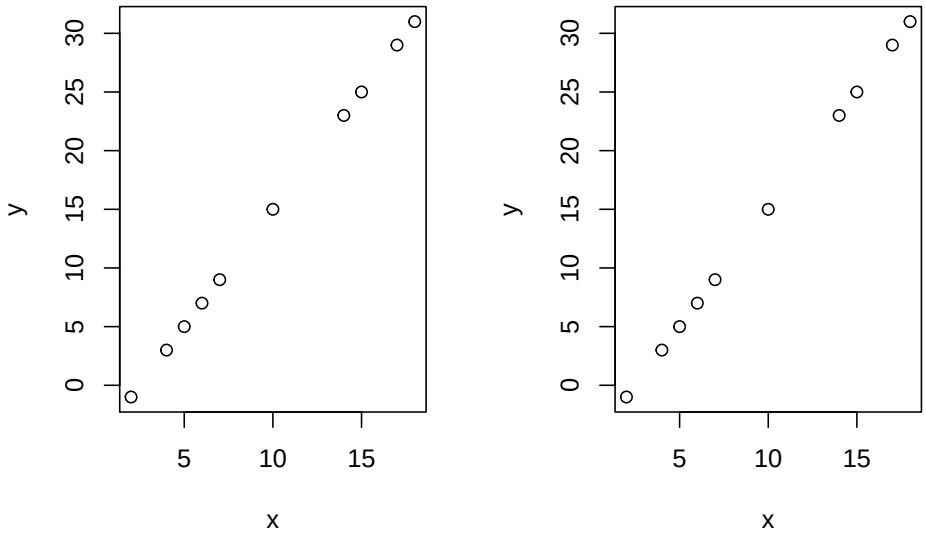
.

R , :

```
x <- sample(1:20, 10)
y <- 2 * x  - 5
par(mfrow = c(1,2)) #
```

[2] https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/par

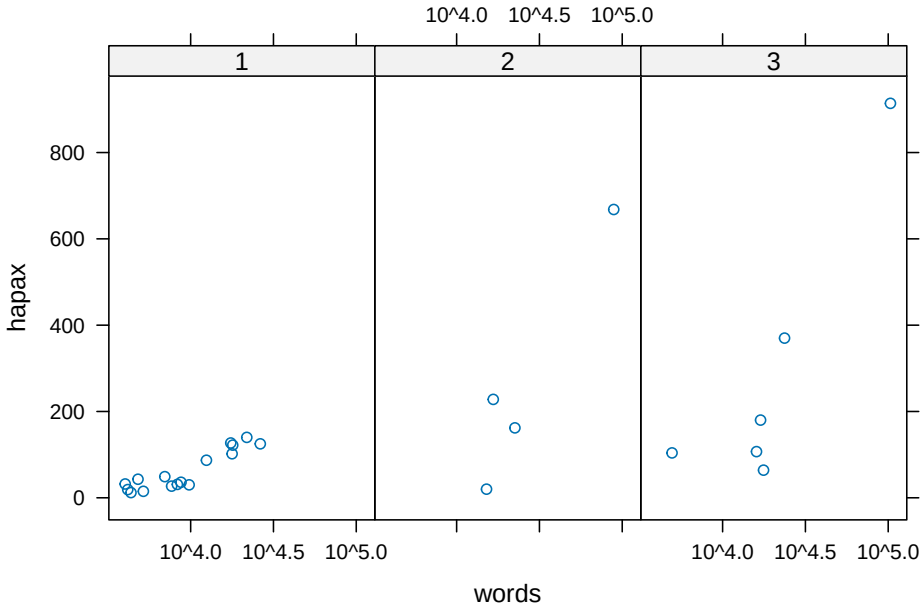
```
plot(x, y)
plot(x, y)
```



, . ,
, . . Lattice.

, .
, .

```
library(lattice)
attach(hapax_plato)

#                          ,                    ;              (      )

xyplot(hapax ~ words | group, data = hapax_plato,
       scales=list(x=list(log=10))) #
```



Lattice, , , ,
. . .

3.3 Ggplot2

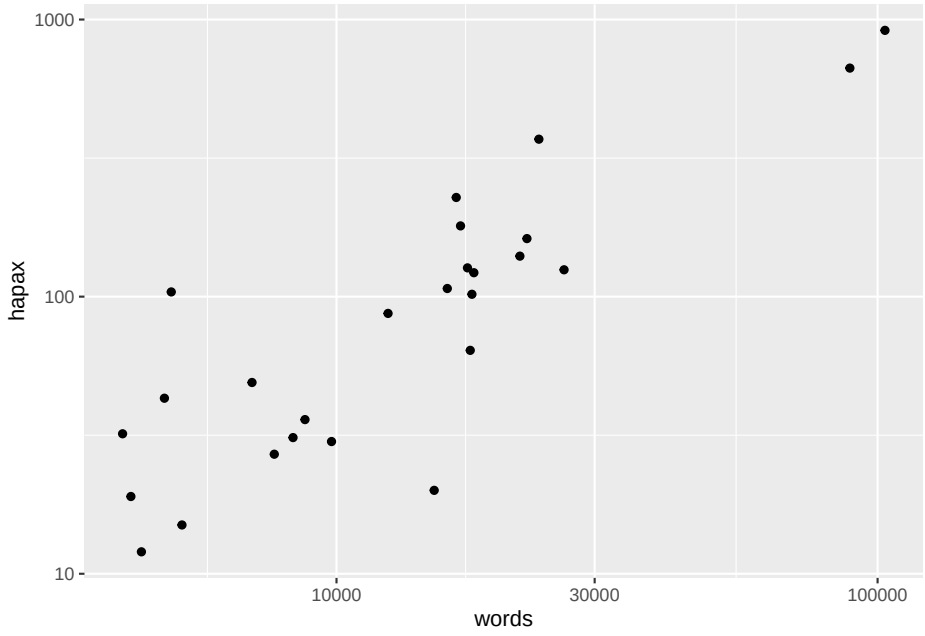
R – ggplot2. “ ”
[?], R Lattice. ,
, ; ,
, Lattice.

3.3.1 : qplot()

; .

```
library(ggplot2) #              tidyverse
options(scipen = 999)
qplot(words, hapax, data = hapax_plato, log = "xy")
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



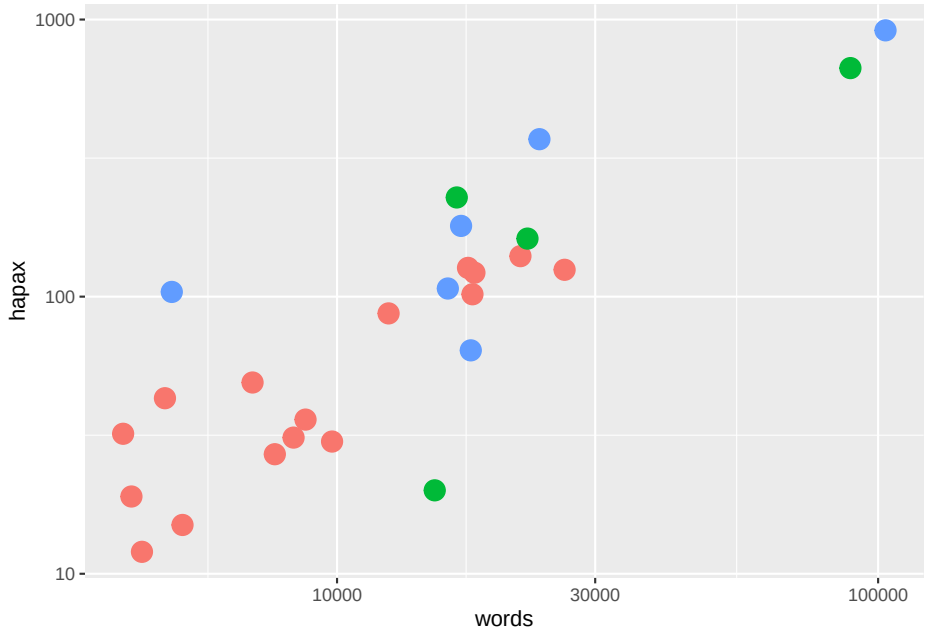
`qplot()` – .

ggplot `qplot()` (deprecated), `ggplot()` .

, .

, , . .

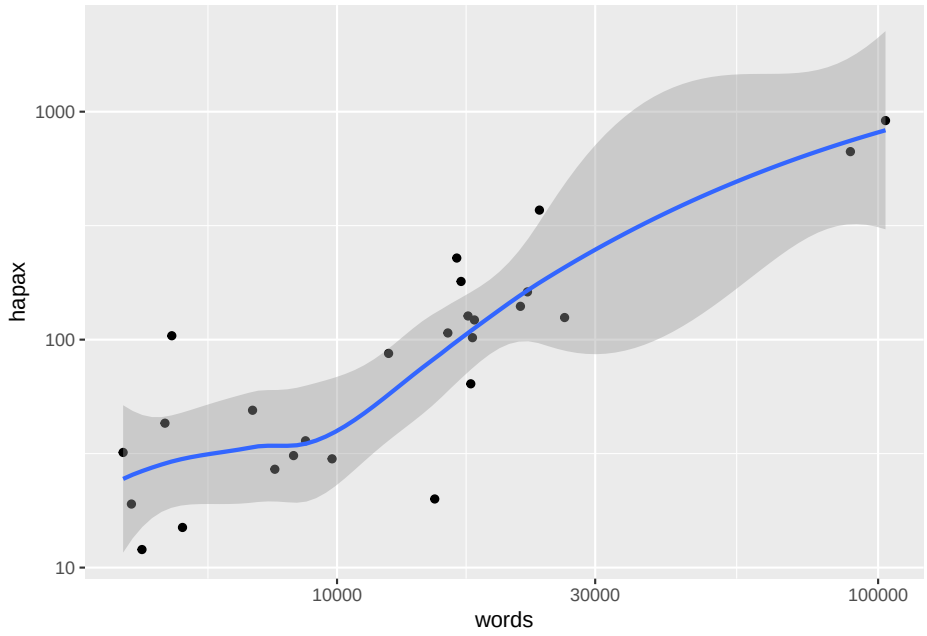
```
qplot(words, hapax, data = hapax_plato, log = "xy", col = group, size = 1.5) + theme(le
```



( ) :

```
qplot(words, hapax, data = hapax_plato, log = "xy", geom = c("point", "smooth"))
```

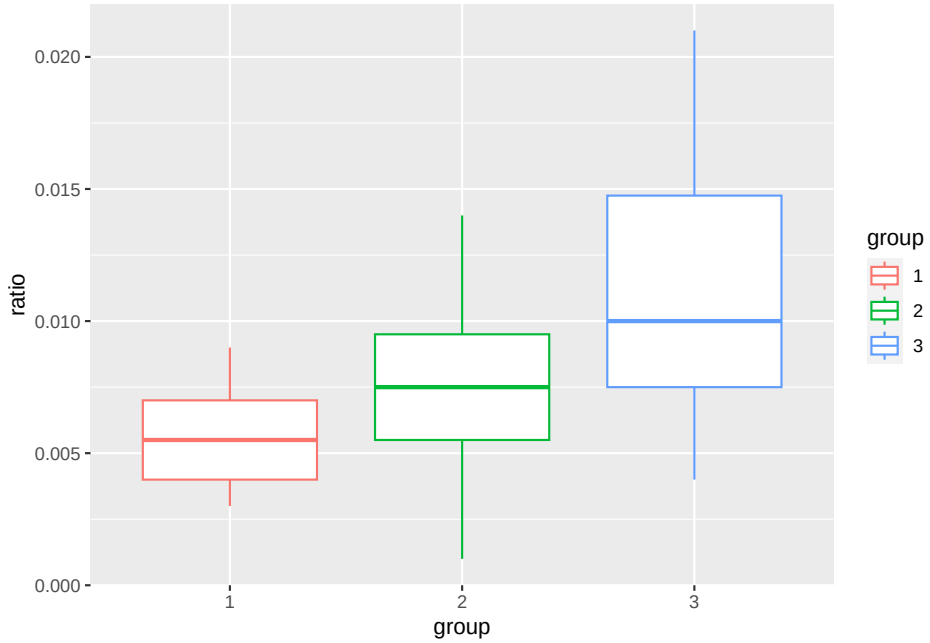
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



( ) ,

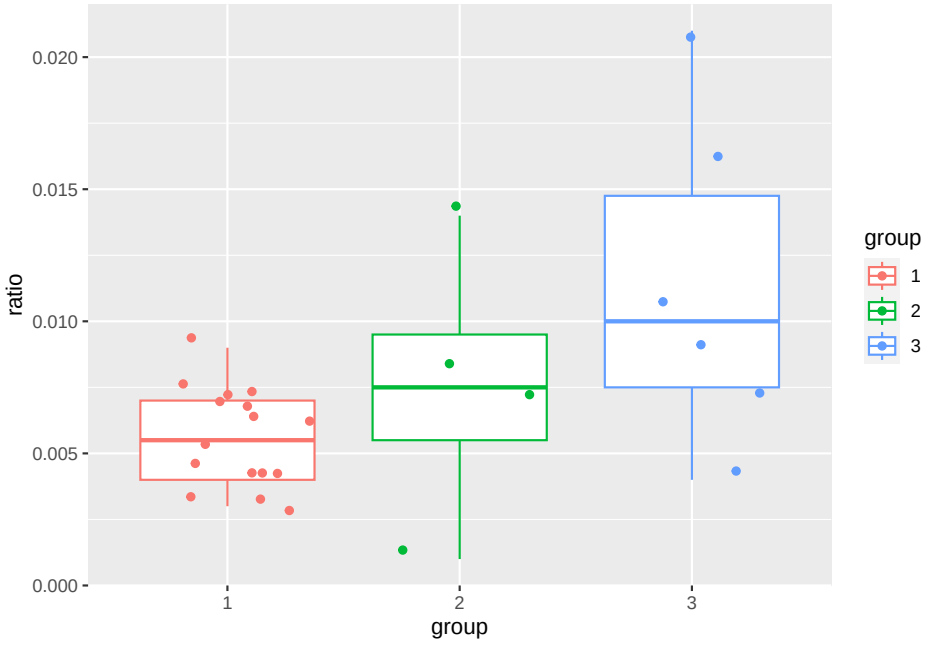
.

```
attach(hapax_plato)
qplot(group, ratio, data = hapax_plato, geom = "boxplot", color = group)
```



.

```
qplot(group, ratio, data = hapax_plato, geom = c("boxplot", "jitter"), color = group)
```



3.3.2 : ggplot()

ggplot(),

: data aes(. *aesthetics*);

.

[?].

, , , .

.

diorisis_meta,

, Diorisis[3].

```
load("./datasets/DiorisisMeta.Rdata")
diorisis_meta
```

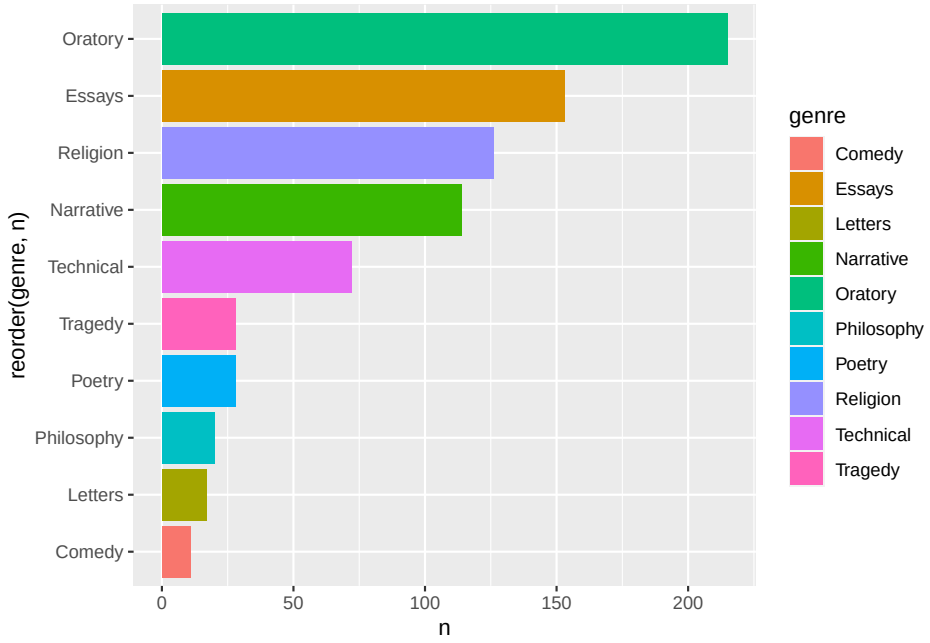
```
## # A tibble: 784 x 5
##    name            title                      date genre     subgenre
##    <chr>           <chr>                     <dbl> <chr>     <chr>
##  1 Achilles Tatius Leucippe and Clitophon      120 Narrative Novel
##  2 Aelian          De Natura Animalium         230 Technical Natural History
##  3 Aelian          Epistulae Rusticae          230 Letters   Letters
##  4 Aelian          Varia Historia              200 Essays    Miscellanea
##  5 Aeneas Tacticus Poliorcetica               -350 Technical Military
##  6 Aeschines       Against Ctesiphon          -330 Oratory   Oratory
##  7 Aeschines       Against Timarchus          -347 Oratory   Oratory
##  8 Aeschines       The Speech on the Embassy  -336 Oratory   Oratory
```

[3]https://figshare.com/articles/dataset/The_Diorisis_Ancient_Greek_Corpus/6187256

```
##  9 Aeschylus       Agamemnon                  -458 Tragedy   Tragedy
## 10 Aeschylus       Eumenides                  -458 Tragedy   Tragedy
## # i 774 more rows
```

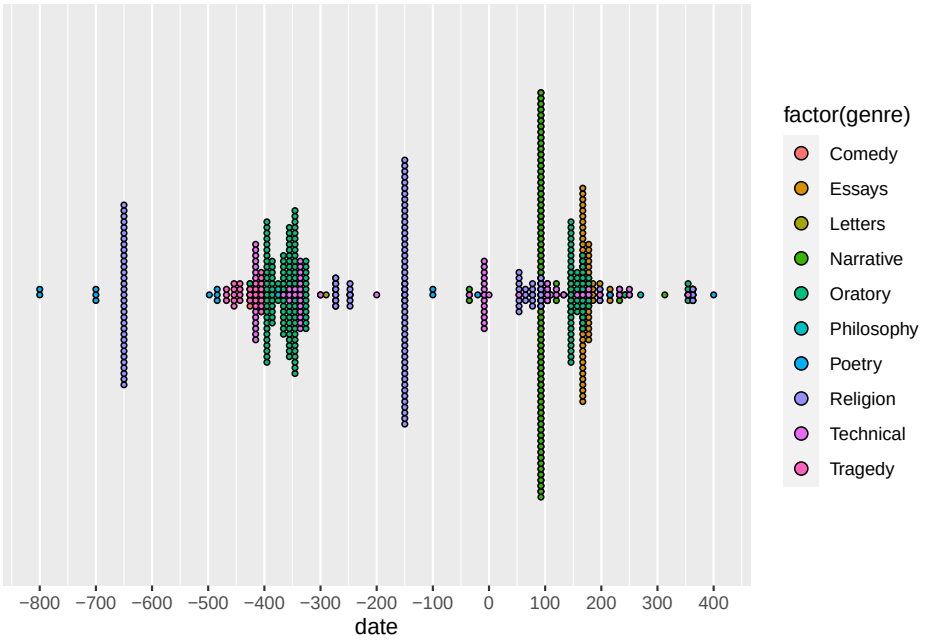
, .

```
library(tidyverse)
diorisis_meta %>%
  group_by(genre) %>%
  count() %>%
  ggplot(aes(reorder(genre, n), n, fill = genre)) +
  geom_bar(stat = "identity") +
  coord_flip()
```



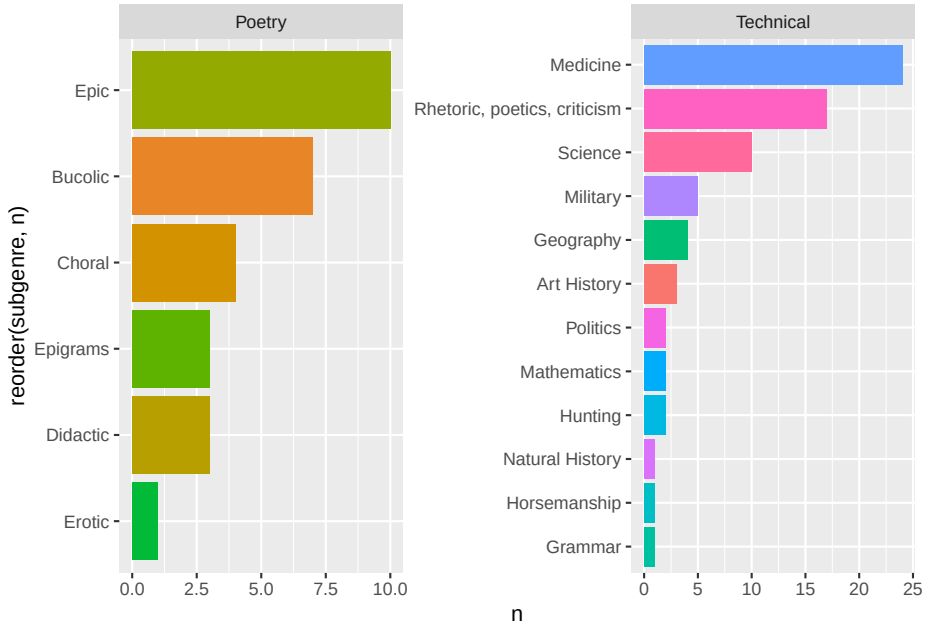
, dotplot, ,
, . , .
(,) ( ).

```
diorisis_meta %>% ggplot(aes(date, fill = factor(genre))) +
  geom_dotplot(binwidth = 10, stackdir = "centerwhole", binpositions = "all") +
  scale_y_continuous(NULL, breaks = NULL) +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10))
```



. : ? ’ –

```
diorisis_meta %>%
  group_by(genre, subgenre) %>%
  count %>%
  filter(genre %in% c("Poetry", "Technical")) %>%
  ggplot(aes(reorder(subgenre, n), n, fill = subgenre)) +
  geom_col(show.legend = F) +
  facet_wrap(~genre, scales = "free") + #
  coord_flip()
```



.

3.4 R

:

- R ;
- ggsave()
- RStudio.

```
#          pdf
pdf(file = "Diorisis.pdf")
diorisis_meta %>%
  group_by(genre, subgenre) %>%
  count %>%
  filter(genre %in% c("Poetry", "Technical")) %>%
  ggplot(aes(reorder(subgenre, n), n, fill = subgenre)) +
  geom_col(show.legend = F) +
  facet_wrap(~genre, scales = "free") +
  coord_flip()
dev.off()

#
ggsave(
  filename = "Diorisis.png",
  plot = last_plot(),
```

```
  device = "png",
  scale = 1,
  width = NA,
  height = 500,
  units = "px",
  dpi = 300
)
```

Chapter 4

> Tidy datasets are all alike, but every messy dataset is messy in its own way.
>
> — Hadley Wickham

4.1 tidyverse

“ ” R, R, tidyverse [?]. Tidyverse – ( ), ., dplyr, ggplot2 .

```
#
library(tidyverse)
```

4.1.1 Tibble

tidyverse – tibble, [1]. , , : [2].

, , , .

:

- , ;[3]
- print(), head();
- ;
- “ ” ;
- .

[1] https://r4ds.had.co.nz/tibbles.html
[2] https://tibble.tidyverse.org/
[3] , : https://simplystatistics.org/posts/2015-07-24-stringsasfactors-an-unauthorized-biography/

```
load("./datasets/DiorisisMeta.Rdata")

#                  10    ,                       ,
as_tibble(diorisis_meta)
```

```
## # A tibble: 784 x 5
##    name            title                       date genre     subgenre
##    <chr>           <chr>                      <dbl> <chr>     <chr>
##  1 Achilles Tatius Leucippe and Clitophon       120 Narrative Novel
##  2 Aelian          De Natura Animalium          230 Technical Natural History
##  3 Aelian          Epistulae Rusticae           230 Letters   Letters
##  4 Aelian          Varia Historia               200 Essays    Miscellanea
##  5 Aeneas Tacticus Poliorcetica                -350 Technical Military
##  6 Aeschines       Against Ctesiphon           -330 Oratory   Oratory
##  7 Aeschines       Against Timarchus           -347 Oratory   Oratory
##  8 Aeschines       The Speech on the Embassy   -336 Oratory   Oratory
##  9 Aeschylus       Agamemnon                   -458 Tragedy   Tragedy
## 10 Aeschylus       Eumenides                   -458 Tragedy   Tragedy
## # i 774 more rows
```

```
#
head(as.data.frame(diorisis_meta)[, 1])  #
```

```
## [1] "Achilles Tatius" "Aelian"          "Aelian"          "Aelian"
## [5] "Aeneas Tacticus" "Aeschines"
```

```
as_tibble(diorisis_meta)[,1] #
```

```
## # A tibble: 784 x 1
##    name
##    <chr>
##  1 Achilles Tatius
##  2 Aelian
##  3 Aelian
##  4 Aelian
##  5 Aeneas Tacticus
##  6 Aeschines
##  7 Aeschines
##  8 Aeschines
##  9 Aeschylus
## 10 Aeschylus
## # i 774 more rows
```

```
#
df <- data.frame('var 1' = 1:2, two = 3:4)
df
```

```
##   var.1 two
```

```
## 1     1   3
## 2     2   4
```

```
tbl <- tibble('var 1' = 1:2, two = 3:4)
tbl
```

```
## # A tibble: 2 x 2
##   `var 1`   two
##     <int> <int>
## 1       1     3
## 2       2     4
```

4.1.2 Dplyr

, tibble “ ”, dplyr[4].

,

:

- mutate() , ;
- select() ;
- filter() ;
- summarise() ;
- arrange() .

group_by(),

“ ”, **pipe %>%** magrittr.

, .

```
diorisis_meta %>%
  select(-subgenre) %>%
  filter(genre == "Narrative") %>%  #          !
  group_by(name) %>%
  count() %>%
  arrange(-n)
```

```
## # A tibble: 20 x 2
## # Groups:   name [20]
##    name                          n
##    <chr>                     <int>
##  1 Plutarch                     71
##  2 Appian                       14
##  3 Flavius Josephus              4
##  4 Xenophon                      4
##  5 Arrian                        3
##  6 Diodorus Siculus              3
##  7 Philostratus the Athenian     2
```

[4] https://dplyr.tidyverse.org/

```
##  8 Achilles Tatius               1
##  9 Cassius Dio                   1
## 10 Chariton                      1
## 11 Diogenes Laertius             1
## 12 Dionysius of Halicarnassus    1
## 13 Eusebius of Caesarea          1
## 14 Herodotus                     1
## 15 Longus                        1
## 16 Lucian                        1
## 17 Polybius                      1
## 18 Pseudo Apollodorus            1
## 19 Thucydides                    1
## 20 Xenophon of Ephesus           1
```

R :

```
diorisis_df <- as.data.frame(diorisis_meta)
diorisis_select <- diorisis_df[,-5] # remove column
diorisis_filter <- diorisis_select[diorisis_select$genre == "Narrative", ]
diorisis_names <- diorisis_filter$name
diorisis_count <- as.data.frame(table(diorisis_names))
diorisis_sort <- diorisis_count[order(diorisis_count$Freq, decreasing =T),]
diorisis_sort
```

```
##                diorisis_names Freq
## 15                   Plutarch   71
## 2                      Appian   14
## 10           Flavius Josephus    4
## 19                   Xenophon    4
## 3                      Arrian    3
## 6            Diodorus Siculus    3
## 14  Philostratus the Athenian    2
## 1              Achilles Tatius    1
## 4                  Cassius Dio    1
## 5                     Chariton    1
## 7           Diogenes Laertius    1
## 8  Dionysius of Halicarnassus    1
## 9        Eusebius of Caesarea    1
## 11                   Herodotus    1
## 12                      Longus    1
## 13                      Lucian    1
## 16                    Polybius    1
## 17         Pseudo Apollodorus    1
## 18                  Thucydides    1
## 20        Xenophon of Ephesus    1
```

- .

4.2

tidyverse – , “ ”. “ ” , , , “ ” [5].

:

- ;
- ;
- ;
- – .

: .

tidyr , .

```
data("table2")
table2
```

```
## # A tibble: 12 x 4
##    country      year type            count
##    <chr>       <dbl> <chr>           <dbl>
##  1 Afghanistan  1999 cases             745
##  2 Afghanistan  1999 population   19987071
##  3 Afghanistan  2000 cases            2666
##  4 Afghanistan  2000 population   20595360
##  5 Brazil       1999 cases           37737
##  6 Brazil       1999 population  172006362
##  7 Brazil       2000 cases           80488
##  8 Brazil       2000 population  174504898
##  9 China        1999 cases          212258
## 10 China        1999 population 1272915272
## 11 China        2000 cases          213766
## 12 China        2000 population 1280428583
```

```
data("table3")
table3
```

```
## # A tibble: 6 x 3
##   country      year rate
##   <chr>       <dbl> <chr>
## 1 Afghanistan  1999 745/19987071
## 2 Afghanistan  2000 2666/20595360
## 3 Brazil       1999 37737/172006362
## 4 Brazil       2000 80488/174504898
## 5 China        1999 212258/1272915272
## 6 China        2000 213766/1280428583
```

[5] https://r4ds.had.co.nz/tidy-data.html

```
data("table4a")
table4a
```

```
## # A tibble: 3 x 3
##   country     `1999` `2000`
##   <chr>        <dbl>  <dbl>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

```
data("table4b")
table4b
```

```
## # A tibble: 3 x 3
##   country         `1999`     `2000`
##   <chr>            <dbl>      <dbl>
## 1 Afghanistan   19987071   20595360
## 2 Brazil       172006362  174504898
## 3 China       1272915272 1280428583
```

tidyr:[6]

- separate() ;
- unite() ;
- pivot_longer() ;
- pivot_wider() ;
- drop_na() replace_na() , NA .

distinct() dplyr, unique() .

, dplyr _join, .[7]

.

4.3 :

4.3.1

. Book-Crossing – 250 .

.

```
load("./datasets/BooksBX.Rdata")
load("./datasets/RatingsBX.Rdata")
load("./datasets/UsersBX.Rdata")
```

[6] https://tidyr.tidyverse.org/reference/index.html
[7] https://r4ds.had.co.nz/relational-data.html

```
ratings
```

```
## # A tibble: 493,813 x 3
##    `User-ID` ISBN       `Book-Rating`
##        <dbl> <chr>              <dbl>
##  1    276725 034545104X             0
##  2    276726 0155061224             5
##  3    276727 0446520802             0
##  4    276729 052165615X             3
##  5    276729 0521795028             6
##  6    276733 2080674722             0
##  7    276736 3257224281             8
##  8    276737 0600570967             6
##  9    276744 038550120X             7
## 10    276745 342310538             10
## # i 493,803 more rows
```

```
users
```

```
## # A tibble: 246,666 x 3
##    `User-ID` Location                           Age
##        <dbl> <chr>                              <chr>
##  1         1 nyc, new york, usa                 NULL
##  2         2 stockton, california, usa          18
##  3         3 moscow, yukon territory, russia    NULL
##  4         4 porto, v.n.gaia, portugal          17
##  5         5 farnborough, hants, united kingdom NULL
##  6         6 santa monica, california, usa      61
##  7         7 washington, dc, usa                NULL
##  8         8 timmins, ontario, canada           NULL
##  9         9 germantown, tennessee, usa         NULL
## 10        10 albacete, wisconsin, spain         26
## # i 246,656 more rows
```

```
books
```

```
## # A tibble: 270,760 x 8
##    ISBN       `Book-Title`         `Book-Author` `Year-Of-Publication` Publisher
##    <chr>      <chr>                <chr>                         <dbl> <chr>
##  1 0195153448 Classical Mythology  Mark P. O. M~                  2002 Oxford U~
##  2 0002005018 Clara Callan         Richard Bruc~                  2001 HarperFl~
##  3 0060973129 Decision in Normandy Carlo D'Este                   1991 HarperPe~
##  4 0374157065 Flu: The Story of t~ Gina Bari Ko~                  1999 Farrar S~
##  5 0393045218 The Mummies of Urum~ E. J. W. Bar~                  1999 W. W. No~
##  6 0399135782 The Kitchen God's W~ Amy Tan                        1991 Putnam P~
##  7 0425176428 What If?: The World~ Robert Cowley                  2000 Berkley ~
##  8 0671870432 PLEADING GUILTY      Scott Turow                    1993 Audiowor~
```

```
##  9 0679425608 Under the Black Fla~ David Cordin~                1996 Random H~
## 10 074322678X Where You'll Find M~ Ann Beattie                  2002 Scribner
## # i 270,750 more rows
## # i 3 more variables: `Image-URL-S` <chr>, `Image-URL-M` <chr>,
## #   `Image-URL-L` <chr>
```

?

- users Location
-
- https://www.bookcrossing.com/ ;
, . . moscow, yukon territory, russia (–
).
- Age .
- , , , .

: - 30 ? -
? - , ? -
? - ? . .

- , .

: ratings books isbn, ratings users
User-ID.

4.3.2

.

```
users_separated <- users %>%
  mutate(Age = as.numeric(Age)) %>%
  filter(!is.na(Age))  %>% # drop_na(Age)
  separate(Location, into = c(NA, NA, "country"), sep = ",")

users_separated #                ,
```

```
## # A tibble: 148,869 x 3
##    `User-ID` country        Age
##        <dbl> <chr>        <dbl>
##  1         2 " usa"          18
##  2         4 " portugal"     17
##  3         6 " usa"          61
##  4        10 " spain"        26
##  5        11 " australia"    14
##  6        13 " spain"        26
##  7        18 " brazil"       25
##  8        19 ""              14
##  9        20 " usa"          19
```

```
## 10        21 " spain"        46
## # i 148,859 more rows
```

, .

```
users_separated %>%
  group_by(country) %>%
  count() %>%
  arrange(-n)
```

```
## # A tibble: 543 x 2
## # Groups:   country [543]
##    country               n
##    <chr>             <int>
##  1 " usa"            67138
##  2 " united kingdom" 10935
##  3 " canada"          9877
##  4 " spain"           9505
##  5 " germany"         8016
##  6 " australia"       7824
##  7  <NA>              5914
##  8 " italy"           4754
##  9 " france"          2395
## 10 " portugal"        2175
## # i 533 more rows
```

:

```
users_separated %>%
  group_by(country) %>%
  count() %>%
  arrange(n)
```

```
## # A tibble: 543 x 2
## # Groups:   country [543]
##    country                    n
##    <chr>                  <int>
##  1 "  pasig city."             1
##  2 " &#20013;&#22269;"         1
##  3 " &#32654;&#22269;"         1
##  4 " 5057chadwick ct."         1
##  5 " 600 083"                  1
##  6 " \\n/a\\\""                1
##  7 " a new year is ahead"      1
##  8 " aberdeenshire"            1
##  9 " agusan del sur"           1
## 10 " alabama"                  1
## # i 533 more rows
```

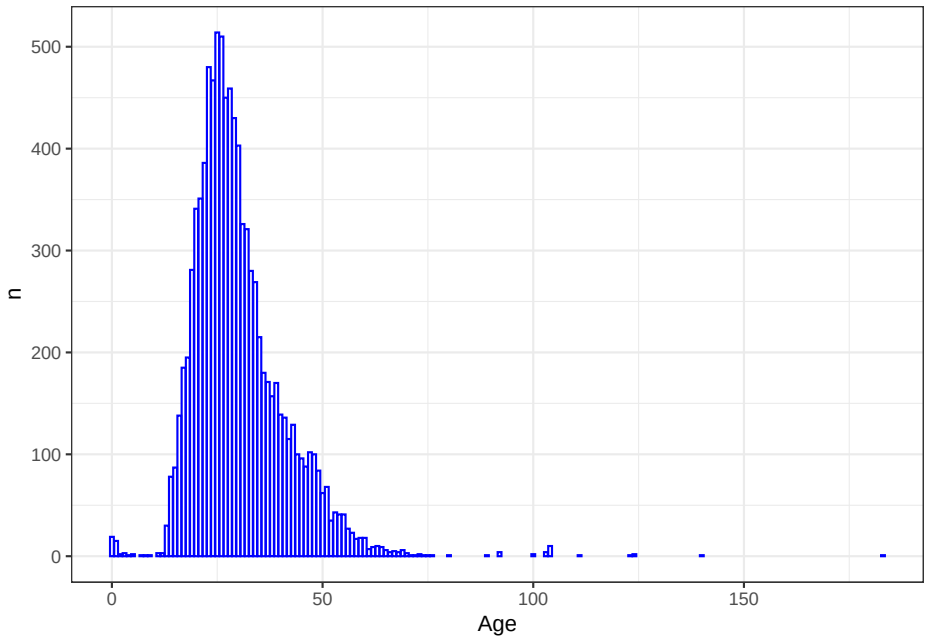
. (
) - . , .
, . ,
separate() , . (
) mutate().

```
spain_data <- users_separated %>%
  mutate(country = str_replace_all(country, pattern = "\\s+", "")) %>% #        ,
  filter(country == "spain") %>%
  group_by(Age) %>%
  count() %>%
  arrange(-n)

spain_data
```

```
## # A tibble: 86 x 2
## # Groups:   Age [86]
##      Age     n
##    <dbl> <int>
##  1    25   514
##  2    26   510
##  3    23   480
##  4    24   467
##  5    28   459
##  6    27   450
##  7    29   430
##  8    30   403
##  9    22   386
## 10    21   351
## # i 76 more rows
```

:

```
spain_data %>%
  ggplot(aes(Age, n)) +
  geom_bar(stat = "identity", col = "blue", fill = "white") +
  theme_bw()
```



! 0 183

, , :

```
spain_id <- users_separated %>%
  mutate(country = str_replace_all(country, pattern = "\\s+", "")) %>%
  filter(country == "spain") #                                    ,          ,
```

4.3.3

, ratings users User-ID, ratings
id, (, . .).
_join[8].

.

```
spain_ratings <- spain_id %>%
  left_join(ratings) %>%
  filter(!is.na(ISBN)) %>%
  filter(`Book-Rating` > 7) %>% #                  ,          "`"
  group_by(ISBN) %>%
  count() %>%
  arrange(-n)
```

```
## Joining with `by = join_by(`User-ID`)`
```

[8]https://r4ds.had.co.nz/relational-data.html

```
spain_ratings
```

```
## # A tibble: 1,281 x 2
## # Groups:   ISBN [1,281]
##    ISBN           n
##    <chr>      <int>
##  1 8432206407     4
##  2 8433969978     4
##  3 846630679X     4
##  4 8472236552     4
##  5 8495501198     4
##  6 840149186X     3
##  7 8401499585     3
##  8 8423310353     3
##  9 8423662152     3
## 10 8432215007     3
## # i 1,271 more rows
```

, . spain_ratings books.

```
spain_books <- spain_ratings %>%
  filter(n > 2) %>%
  left_join(books) %>%
  filter(!is.na(`Book-Title`), !is.na(`Book-Author`)) %>%
  ungroup()
```

```
## Joining with `by = join_by(ISBN)`
```

```
spain_books
```

```
## # A tibble: 15 x 9
##    ISBN           n `Book-Title`   `Book-Author` `Year-Of-Publication` Publisher
##    <chr>      <int> <chr>          <chr>                          <dbl> <chr>
##  1 8432206407     4 Sin Noticias ~ Eduardo Mend~                   1995 Planeta ~
##  2 8433969978     4 El Libro de L~ Paul Auster                     2003 Anagrama
##  3 846630679X     4 La caverna = ~ Jose Saramago                   2002 Punto de~
##  4 8472236552     4 UN Viejo Que ~ Luis Sepulve~                   1993 Tusquets~
##  5 8495501198     4 Memorias de u~ Arthur Golden                   2001 Suma de ~
##  6 840149186X     3 El Club de Lo~ N. H. Kleinb~                   1995 Plaza &a~
##  7 8401499585     3 Los Pilares d~ Ken Follett                     1995 Plaza &a~
##  8 8423310353     3 El Camino (Co~ Miguel Delib~                   1991 Continen~
##  9 8432215007     3 El perfume     Patrick Susk~                   1997 Editoria~
## 10 8445071408     3 El Senor De L~ J. R. R. Tol~                   2001 Minotauro
## 11 8445071416     3 El Hobbit      J. R. R. Tol~                   1991 Minotauro
## 12 8477204055     3 El caballero ~ Robert Fisher                   2000 Obelisco
## 13 8478884459     3 Harry Potter ~ J. K. Rowling                   1999 Lectorum~
## 14 8484602508     3 Diario de Un ~ Antonio Salas                   2003 Temas de~
```

```
## 15 8495501112     3 Son De Mar     Manuel Vicent                  2002 Suma de ~
## # i 3 more variables: `Image-URL-S` <chr>, `Image-URL-M` <chr>,
## #   `Image-URL-L` <chr>
```

, ! (. ,
ISBN, ISBN, ?)

(, ,).
, :

```
spain_books %>%
  select(3:5) %>%
  rename(title = `Book-Title`, author = `Book-Author`)
```

```
## # A tibble: 15 x 3
##    title                                          author `Year-Of-Publication`
##    <chr>                                          <chr>                  <dbl>
##  1 Sin Noticias De Gurb (Biblioteca breve)        Eduar~                  1995
##  2 El Libro de Las Ilusiones                      Paul ~                  2003
##  3 La caverna = A caverna                         Jose ~                  2002
##  4 UN Viejo Que Leia Novelas De Amor/the Old Men W~ Luis ~                1993
##  5 Memorias de una geisha                         Arthu~                  2001
##  6 El Club de Los Poetas Muertos                  N. H.~                  1995
##  7 Los Pilares de La Tierra                       Ken F~                  1995
##  8 El Camino (Coleccion Destinolibro)             Migue~                  1991
##  9 El perfume                                     Patri~                  1997
## 10 El Senor De Los Anillos: LA Comunidad Del Anill~ J. R.~                2001
## 11 El Hobbit                                      J. R.~                  1991
## 12 El caballero de la armadura oxidada            Rober~                  2000
## 13 Harry Potter y la piedra filosofal             J. K.~                  1999
## 14 Diario de Un Skin: Un Topo En El Movimiento Neo~ Anton~                2003
## 15 Son De Mar                                     Manue~                  2002
```

select() - [9], :

- starts_with()
- ends_with()
- contains()
- matches()
- num_range()

```
spain_books %>%
  select(-contains("URL"), -matches("Publisher")) %>% #
  rename(title = `Book-Title`,
         author = `Book-Author`,
         published = `Year-Of-Publication`) #
```

```
## # A tibble: 15 x 5
```

[9] https://r4ds.had.co.nz/transform.html

```
##    ISBN           n title                                    author published
##    <chr>      <int> <chr>                                    <chr>      <dbl>
##  1 8432206407     4 Sin Noticias De Gurb (Biblioteca breve)  Eduar~      1995
##  2 8433969978     4 El Libro de Las Ilusiones                Paul ~      2003
##  3 846630679X     4 La caverna = A caverna                   Jose ~      2002
##  4 8472236552     4 UN Viejo Que Leia Novelas De Amor/the Old ~ Luis ~      1993
##  5 8495501198     4 Memorias de una geisha                   Arthu~      2001
##  6 840149186X     3 El Club de Los Poetas Muertos            N. H.~      1995
##  7 8401499585     3 Los Pilares de La Tierra                 Ken F~      1995
##  8 8423310353     3 El Camino (Coleccion Destinolibro)       Migue~      1991
##  9 8432215007     3 El perfume                               Patri~      1997
## 10 8445071408     3 El Senor De Los Anillos: LA Comunidad Del ~ J. R.~      2001
## 11 8445071416     3 El Hobbit                                J. R.~      1991
## 12 8477204055     3 El caballero de la armadura oxidada      Rober~      2000
## 13 8478884459     3 Harry Potter y la piedra filosofal       J. K.~      1999
## 14 8484602508     3 Diario de Un Skin: Un Topo En El Movimient~ Anton~      2003
## 15 8495501112     3 Son De Mar                               Manue~      2002
```

, - .

Chapter 5

5.1

R . ,
, [?].

, , , , .
. return(); ,
(. ,).

, , :

```
center <- function(x){
  n = x - mean(x)
  return(n)
}

x <- c(5, 10, 15)
center(x) #              ,
```

```
## [1] -5  0  5
```

scale() (scale = F;
,).

n, . .
– . , . :
.

. :

```
center <- function(x, na.rm = F){
  if(na.rm) { x <- x[!is.na(x)]} #
  x - mean(x) #           return()
```

```
}

x <- c(5, 10, NA)
center(x)
```

```
## [1] NA NA NA
```

? ?

```
center(x, na.rm = T)
```

```
## [1] -2.5  2.5
```

R , , . ,
, .

```
center <- function(x, na.rm = F, what_is_your_name){
  if(na.rm) { x <- x[!is.na(x)]} #
  x - mean(x) #           return()
}

center(x, na.rm = T)
```

```
## [1] -2.5  2.5
```

```
center(x, na.rm = T, what_is_your_name = "Locusclassicus")
```

```
## [1] -2.5  2.5
```

, , , , .
(f() –). R !

, .

```
center <- function(x){
  if (length(x) == 1) {stop("            ")}
  x - mean(x) #           return()
}

x <- 10
center(x) #
```

, R . :

```
x <- c(5, 10, 15)
x - mean(x)
```

```
## [1] -5  0  5
```

5.2

f(), x,
f() x. . .

R , [?].

R, . x + 4
+(x, 4):

```
x <- c(1.2, 2.51, 3.8)

#
round(x)
```

```
## [1] 1 3 4
```

```
#
`+`(x, 4)
```

```
## [1] 5.20 6.51 7.80
```

, : ,
.

```
is_article <- function(x){
  x == c("a", "the")
}

x <- "the"
is_article(x) #              ,
```

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

, .

```
x <- c("just", "the")
is_article(x)
```

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

```
x <- c("the", "just")
is_article(x) #
```

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

?

, !

, . , :

```
is_article <- function(x) {
  articles <- c("a", "the")
```

```
  x %in% articles
}

x <- c(rep("the", 5), rep("if", 5))
is_article(x)
```

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

```
sum(is_article(x))
```

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

5.3 ?

:

- , ;
- , ;
- (, ,)

: , !

> Writing good functions is a lifetime journey.
>
> — Hadley Wickham

. , . , ,
.

```
#
x <- c("  ", "   ", "   ")
y <- c(" ", "   ", "  ", " ", "    ", "   ", "    ", "  ")
idx <- which(x %in% y)
x[idx]
```

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

```
#
common_words <- function(x, y){
  idx <- which(x %in% y)
  x[idx]
}

#
x <- c("  ", "   ", "  ", "      ", "    ", "  ")
y <- c("    ", "  ", "  ", " ", "    ", "   ", "    ")
common_words(x, y)
```

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

, !

, ( ). ,
. Stylo;
Delta . , .

```
library(stylo)
dist.delta
```

```
## function (x, scale = TRUE)
## {
##     if (is.matrix(x) == FALSE & is.data.frame(x) == FALSE) {
##         stop("cannot apply a distance measure: wrong data format!")
##     }
##     if (length(x[1, ]) < 2 | length(x[, 1]) < 2) {
##         stop("at least 2 cols and 2 rows are needed to compute a distance!")
##     }
##     if (scale == TRUE) {
##         x = scale(x)
##     }
##     y = dist(x, method = "manhattan")/length(x[1, ])
##     return(y)
## }
## <bytecode: 0x141bb3f58>
## <environment: namespace:stylo>
```

5.4

5.4.1

R, , –
.

R ,
, , .

—

: for while. for,
while .

5.4.1.1 for

.

```
y <- c(" ", "   ", "  ", " ", "    ", "   ", "     ", "  ")

result <- c()
for(i in y) {
  n <- nchar(i)
```

```
  result <- c(result, n)
}

result
```

```
## [1] 1 5 3 2 7 6 9 4
```

, - ; ,
nchar() .

```
nchar(y)
```

```
## [1] 1 5 3 2 7 6 9 4
```

for . , (
,). . ,
.

```
rownames(hapax_plato) <- hapax_plato$dialogue
hapax_plato <- hapax_plato %>% select(-ratio, -group, -dialogue) #      pipe          d
str(hapax_plato)
```

```
## 'data.frame':    26 obs. of  2 variables:
##  $ words: chr  "8745" "8311" "17944" "4950" ...
##  $ hapax: chr  "36" "31" "122" "104" ...
```

chr, , .

```
for (i in seq_along(hapax_plato)) {  # seq_along   1:length(x)
  hapax_plato[,i] <- as.numeric(hapax_plato[,i])
}

str(hapax_plato) #      ,
```

```
## 'data.frame':    26 obs. of  2 variables:
##  $ words: num  8745 8311 17944 4950 4169 ...
##  $ hapax: num  36 31 122 104 19 87 15 125 12 32 ...
```

, . ,
: colSums() (, , rowSums()).
, .

```
medians <- c()
for (i in seq_along(hapax_plato)) {
  m <- median(hapax_plato[,i])
  medians <- c(medians, m)
}

medians
```

```
## [1] 15589.5    94.5
```

, , .
, [1]. –
.

```
medians <- vector("double", ncol(hapax_plato))
for (i in seq_along(hapax_plato)) {
  medians[i] <- median(hapax_plato[,i])
}

medians
```

```
## [1] 15589.5    94.5
```

tictoc.

```
library(tictoc)

#
tic()
medians <- c()
for (i in seq_along(hapax_plato)) {
  m <- median(hapax_plato[,i])
  medians <- c(medians, m)
}
toc()
```

```
## 0.003 sec elapsed
```

```
#
tic()
medians <- vector("double", ncol(hapax_plato))
for (i in seq_along(hapax_plato)) {
  medians[i] <- median(hapax_plato[,i])
}
toc()
```

```
## 0.003 sec elapsed
```

, ( ).
, i. !

5.4.1.2 while

, while . , , ,
, 6 . ?

```
k <- 0
n <- 0
```

[1] https://r4ds.had.co.nz/iteration.html

```
while (n != 6) {
  k <- k + 1
  n <- nchar(y[k])
}

y[k]
```

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

. , NA.

```
y[nchar(y) == 6][1]
```

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

5.4.2

, .

```
if(any(nchar(y) > 6)) print("     ")
```

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

| (“ ”) & (“ ”), :

```
y
```

```
## [1] " "          "   "     "  "       " "         "    "    "    "
## [7] "     " "   "
```

```
nchar(y) > 6 | nchar(y) < 2
```

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

|| (“ ”) && (“ ”), , .

```
nchar(y) > 6 || nchar(y) < 2
```

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

:

```
if (sum(nchar(y)) > 10) {
  print("      ")
} else if (sum(nchar(y)) < 5) {
  print("     ")
} else {
  print("     ")
}
```

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

:

```
ifelse((sum(nchar(y)) > 10), "      ", "      ")
```

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

5.5 _apply R

, , R _apply.

5.5.1 tapply()

, ( ) . , .

```
load("./datasets/HapaxPlato.Rdata")
```

```
my_fct <- as.factor(hapax_plato$group)
my_vct <- as.numeric(hapax_plato$ratio)
tapply(my_vct, my_fct, mean)
```

```
##          1          2          3
## 0.00550000 0.00750000 0.01133333
```

tidyverse :

```
hapax_plato %>%
  mutate(ratio = as.numeric(ratio)) %>%
  group_by(group) %>%
  summarise(mean = mean(ratio))
```

```
## # A tibble: 3 x 2
##   group   mean
##   <dbl>  <dbl>
## 1     1 0.0055
## 2     2 0.0075
## 3     3 0.0113
```

5.5.2 apply()

.

```
#
rownames(hapax_plato) <- hapax_plato$dialogue
hapax_plato <- subset(hapax_plato, select = -c(dialogue, group)) #

#
str(hapax_plato)
```

```
## 'data.frame':    26 obs. of  3 variables:
##  $ words: chr  "8745" "8311" "17944" "4950" ...
```

```
##  $ hapax: chr  "36" "31" "122" "104" ...
##  $ ratio: chr  "0.004" "0.004" "0.007" "0.021" ...
```

```
#
hapax_plato<- apply(hapax_plato, 2, as.numeric)

#
round((apply(hapax_plato, 2, sd)), 3)
```

```
##     words     hapax     ratio
## 23640.597   208.856     0.004
```

apply() sd() stats, ( ).

```
hapax_centered <- apply(hapax_plato, 2, function(x) x - mean(x))
head(hapax_centered)
```

```
##           words      hapax         ratio
## [1,] -10619.423 -110.69231 -0.0031538462
## [2,] -11053.423 -115.69231 -0.0031538462
## [3,]  -1420.423  -24.69231 -0.0001538462
## [4,] -14414.423  -42.69231  0.0138461538
## [5,] -15195.423 -127.69231 -0.0021538462
## [6,]  -6911.423  -59.69231 -0.0001538462
```

- , ( ) dplyr:

```
as_tibble(hapax_plato) %>%
  mutate(words = words - mean(words),
         hapax = hapax - mean(hapax),
         ratio = ratio - mean(ratio))
```

! , ! , [2]:

```
as_tibble(hapax_plato) %>%
  mutate_all(function(x) x - mean(x))
```

[3]:

```
fn <- function(x) x - mean(x)
as_tibble(hapax_plato) %>%
  mutate(across(1:3, fn))
```

, , .

[2] https://dplyr.tidyverse.org/reference/mutate_all.html
[3] https://dplyr.tidyverse.org/articles/colwise.html

5.5.3 lapply() sapply()

lapply() sapply() . , ,
.

, , – stylo.corpus,
stylo[4].

.

```
load("./datasets/PlatoStylo.Rdata")
class(corpus)
```

```
## [1] "stylo.corpus"
```

RStudio, , corpus , (,
,):

, – 26 ( ), – .
, , :

```
class(corpus[[1]])
```

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

, 1000 , sample()
. lapply() (l = list) :

```
samples <- lapply(corpus, sample, 1000, replace = T)
```

samples ,
.

sapply() , (s = simplify).

```
s_sample <- sapply(corpus[1:2], sample, 5, replace = F)
s_sample
```

```
##      Apology  Charmides
## [1,] "    " "   "
## [2,] " "     "    "
## [3,] "  "    "     "
## [4,] "  "    " "
## [5,] "   " "    "
```

, dplyr , stylo :

```
corpus_df <- stack(corpus)
head(corpus_df)
```

```
##     values     ind
```

[4] https://rdrr.io/cran/stylo/

```
## 1        Apology
## 2         Apology
## 3          Apology
## 4           Apology
## 5         Apology
## 6       Apology
```

:

```
corpus_tbl <- corpus_df %>%
  as_tibble() %>%
  relocate(ind, .before = values) %>%
  rename(title = ind,
         word = values)

corpus_tbl
```

```
## # A tibble: 503,475 x 2
##    title   word
##    <fct>   <chr>
##  1 Apology
##  2 Apology
##  3 Apology
##  4 Apology
##  5 Apology
##  6 Apology
##  7 Apology
##  8 Apology
##  9 Apology
## 10 Apology
## # i 503,465 more rows
```

:

```
samples <- corpus_tbl %>%
  group_by(title) %>%
  sample_n(size = 1000, replace = T)

dim(samples) #    ,                        1000
```

```
## [1] 26000     2
```

!

5.6 Purrr

– purrr tidyverse[5]. .

[5] https://purrr.tidyverse.org/

> You should never feel bad about using a loop instead of a map function. The map functions are a step up a tower of abstraction, and it can take a long time to get your head around how they work.
>
> — Hadley Wickham & Garrett Grolemund

– `map()` – 23 [6]. `map` – , : [?].

:

- `map(.x, .f, ..., .progress = FALSE)`
- `map_lgl(.x, .f, ..., .progress = FALSE)`
- `map_int(.x, .f, ..., .progress = FALSE)`
- `map_dbl(.x, .f, ..., .progress = FALSE)`
- `map_chr(.x, .f, ..., .progress = FALSE)`

:

- `map_if()`
- `imap()`
- `lmap()`
- `map2()`
- `map_if()`
- `modify()`

. purrr – [7], .

5.6.1 map_df() map_dbl()

, map .

```
hapax_plato <- as_tibble(hapax_plato)
map_df(hapax_plato, center)
```

```
## # A tibble: 26 x 3
##      words  hapax     ratio
##      <dbl>  <dbl>     <dbl>
##  1 -10619. -111.  -0.00315
##  2 -11053. -116.  -0.00315
##  3  -1420.  -24.7 -0.000154
##  4 -14414.  -42.7  0.0138
##  5 -15195. -128.  -0.00215
##  6  -6911.  -59.7 -0.000154
##  7 -14183. -132.  -0.00415
##  8   6973.  -21.7 -0.00215
##  9 -15004. -135.  -0.00415
## 10 -15340. -115.   0.000846
```

[6] https://adv-r.hadley.nz/functionals.html

[7] ., : https://www.emilhvitfeldt.com/post/2018-01-08-purrr-tips-and-tricks/

```
## # i 16 more rows
```

map – pipeable, :

```
hapax_plato %>% map_df(center)
```

. , , dbl:

```
round(map_dbl(hapax_plato, mean), 3)
```

```
##     words     hapax     ratio
## 19364.423   146.692     0.007
```

5.6.2 map2()

map2()
[8].

```
mean = list(1, 10, 100)
sd = list(0.5, 5, 50)
map2(mean, sd, rnorm, n = 3)
```

```
## [[1]]
## [1] 0.9159752 2.0379541 0.6558171
##
## [[2]]
## [1]  7.897317  8.210584 11.209210
##
## [[3]]
## [1]  28.44441 111.72813 165.35726
```

, , ; , , – .

2

, pnorm().

map2() : [9].

```
library(slider)
windows <- slide(corpus_tbl[1:36,], ~.x, .after = 6)
out <- map2(.x = windows, .y = 1:length(windows), ~ mutate(.x, window_id = .y)) # out
out[2]
```

```
## [[1]]
## # A tibble: 7 x 3
##   title   word     window_id
##   <fct>   <chr>        <int>
```

[8] https://adv-r.hadley.nz/functionals.html
[9] https://smltar.com/embeddings.html#understand-word-embeddings-by-finding-themyourself

```
## 1 Apology                2
## 2 Apology                 2
## 3 Apology                  2
## 4 Apology                2
## 5 Apology               2
## 6 Apology                2
## 7 Apology                 2
```

5.7 Furrr

, .

Chapter 6

Blocks

6.1 Equations

Here is an equation.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{6.1}$$

You may refer to using `\@ref(eq:binom)`, like see Equation (6.1).

6.2 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 6.1.

Theorem 6.1. *For a right triangle, if c denotes the* length *of the hypotenuse and a and b denote the lengths of the* ***other*** *two sides, we have*

$$a^2 + b^2 = c^2$$

Read more here https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html.

6.3 Callout blocks

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html

Chapter 7

Sharing your book

7.1 Publishing

HTML books can be published online, see: https://bookdown.org/yihui/bookdown/publishing.html

7.2 404 pages

By default, users will be directed to a 404 page if they try to access a webpage that cannot be found. If you'd like to customize your 404 page instead of using the default, you may add either a `_404.Rmd` or `_404.md` file to your project root and use code and/or Markdown syntax.

7.3 Metadata for sharing

Bookdown HTML books will provide HTML metadata for social sharing on platforms like Twitter, Facebook, and LinkedIn, using information you provide in the `index.Rmd` YAML. To setup, set the `url` for your book and the path to your `cover-image` file. Your book's `title` and `description` are also used.

This `gitbook` uses the same social sharing data across all chapters in your book- all links shared will look the same.

Specify your book's source repository on GitHub using the `edit` key under the configuration options in the `_output.yml` file, which allows users to suggest an edit by linking to a chapter's source file.

Read more about the features of this output format here:

https://pkgs.rstudio.com/bookdown/reference/gitbook.html

Or use:

```
?bookdown::gitbook
```

Bibliography