MAT4681 - Statistique pour les sciences

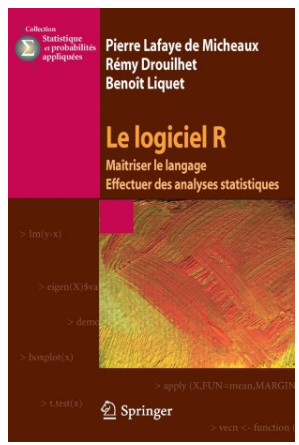
Arthur Charpentier

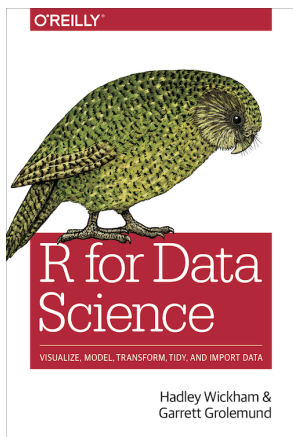
02 - Introduction langage R

été 2022

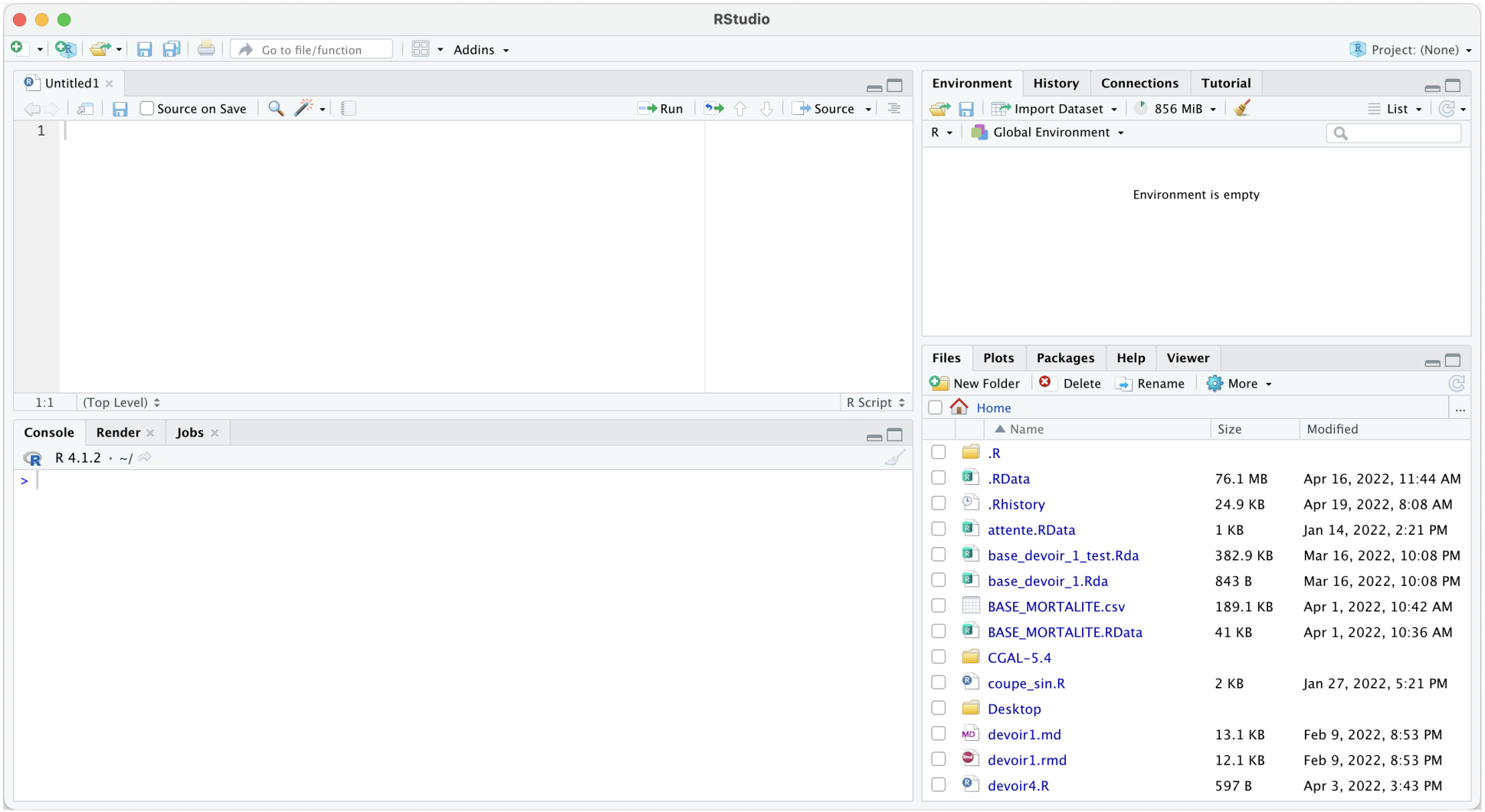
Références

- Introduction à la programmation en R de Vincent Goulet
- Notes de cours de R d'Ewen Gallic
- R pour les débutants d'Emmanuel Paradis
- R pour la statistique et la science des données de Pierre-André Cornillon, Arnaud Guyader, François Husson, Nicolas Jégou, Julie Josse, Nicolas Klutchnikoff, Erwan Le Pennec, Eric Matzner-Lober, Laurent Rouvière, Benoît Thieurmel
- R for data science d'Hadley Wickham et Garrett Grolemund
- YaRrr! The Pirate's Guide to R de Nathaniel Phillips





R Studio


RStudio
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console
Render
Jobs
R 4.1.2 · ~/
>
Environment
History
Connections
Tutorial
Import Dataset
856 MiB
List
R
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Name
Size
Modified
.R
.RData 76.1 MB Apr 16, 2022, 11:44 AM
.Rhistory 24.9 KB Apr 19, 2022, 8:08 AM
attente.RData 1 KB Jan 14, 2022, 2:21 PM
base_devoir_1_test.Rda 382.9 KB Mar 16, 2022, 10:08 PM
base_devoir_1.Rda 843 B Mar 16, 2022, 10:08 PM
BASE_MORTALITE.csv 189.1 KB Apr 1, 2022, 10:42 AM
BASE_MORTALITE.RData 41 KB Apr 1, 2022, 10:36 AM
CGAL-5.4
coupe_sin.R 2 KB Jan 27, 2022, 5:21 PM
Desktop
devoir1.md 13.1 KB Feb 9, 2022, 8:53 PM
devoir1.rmd 12.1 KB Feb 9, 2022, 8:53 PM
devoir4.R 597 B Apr 3, 2022, 3:43 PM

Nombres et vecteurs I

```
> set.seed(1)
> U <- runif(20)
> U[1:4]
[1] 0.2655087 0.3721239 0.5728534 0.9082078
> options(digits = 3)
> U[1:4]
[1] 0.266 0.372 0.573 0.908
> options(digits = 22)
> U[1:4]
[1] 0.2655086631420999765396 0.3721238996367901563644
[3] 0.5728533633518964052200 0.9082077899947762489319
> x <- exp(1)
> y <- x
> x <- 2
> y
[1] 2.72
> class(x)
[1] "numeric"
```

Nombres et vecteurs II

```
> (3/10-1/10)
[1] 0.2
> (3/10-1/10)==(7/10-5/10)
[1] FALSE
> (3/10-1/10)-(7/10-5/10)
[1] 2.78e-17
> all.equal((3/10-1/10),(7/10-5/10))
[1] TRUE
> 0/0
[1] NaN
> 1/0
[1] Inf
```

Nombres et vecteurs III

```
> x <- rnorm(8)
> names(x) <- letters[1:6]
> x
      a       b       c       d       e       f
 1.5118  0.3898 -0.6212 -2.2147  1.1249 -0.0449
> x[2:4]
     b       c       d
 0.390 -0.621 -2.215
> x[c("b","c","d")]
     b       c       d
 0.390 -0.621 -2.215
```

Matrices I

```
> M <- 1:24
> dim(M) <- c(6,4)
> M
     [,1] [,2] [,3] [,4]
[1,]    1    7   13   19
[2,]    2    8   14   20
[3,]    3    9   15   21
[4,]    4   10   16   22
[5,]    5   11   17   23
[6,]    6   12   18   24
> str(M)
int [1:6, 1:4] 1 2 3 4 5 6 7 8 9 10 ...
> colnames(M)=letters[1:4]
> rownames(M)=LETTERS[10:15]
> M
   a  b  c  d
J  1  7 13 19
K  2  8 14 20
L  3  9 15 21
M  4 10 16 22
```

Matrices II

```
 N  5  11  17  23
 O  6  12  18  24
> M["K", ]
  a  b  c  d
  2  8 14 20
> M[c("K","N"), ]
   a  b  c  d
 K 2  8 14 20
 N 5 11 17 23
> M[c(2,5), ]
   a  b  c  d
 K 2  8 14 20
 N 5 11 17 23
```

Facteur - variable catégorielle I

```
> x <- factor(c("b","a","b"))
> levels(x)
[1] "a" "b"
> x[3] <- "c"
Warning in `[<-.factor`(`*tmp*`, 3, value = "c"):
    invalid factor level, NA generated
> x
[1] b    a    <NA>
Levels: a b
> factor(x,levels=c("b","a"))
[1] b    a    <NA>
Levels: b a
> x[1]
[1] b
Levels: a b
```

Facteur - variable catégorielle I

```
1
2
3  U <- runif(20)
4  cut(U,breaks=2)
5
6  ##  [1] (0.145,0.544] (0.145,0.544] (0.145,0.544]
       (0.544,0.944] (0.544,0.944]
7  ##  [6] (0.145,0.544] (0.145,0.544] (0.145,0.544]
       (0.544,0.944] (0.145,0.544]
8  ## [11] (0.145,0.544] (0.145,0.544] (0.145,0.544]
       (0.544,0.944] (0.544,0.944]
9  ## [16] (0.145,0.544] (0.145,0.544] (0.544,0.944]
       (0.145,0.544] (0.544,0.944]
10 ## Levels: (0.145,0.544] (0.544,0.944]
11
12 cut(U,breaks=2,labels=c("small","large"))
13
14 ##  [1] small small small large large small small
       small large small small
```

Facteur - variable catégorielle II

```
## [12] small small large large small small large
    small large
## Levels: small large

cut(U,breaks=c(0,.3,.8,1),labels=c("small","medium","
    large"))

##  [1] medium small  medium large  medium medium
    medium small  large  medium
## [11] medium small  medium medium large  medium
    medium medium medium large
## Levels: small medium large

table(cut(U,breaks=c(0,.3,.8,1),labels=c("small","
    medium","large")))

##
##  small medium  large
##      3     13      4
```

Facteur - variable catégorielle I

```
"Be carefull of 'quotes'"
'Be carefull of "quotes"'

cities <- c("New York, NY", "Los Angeles, CA", "Boston, MA")
substr(cities, nchar(cities)-1, nchar(cities))

## [1] "NY" "CA" "MA"

unlist(strsplit(cities, ", "))[seq(2,6,by=2)]

## [1] "NY" "CA" "MA"

some.dates <- as.Date(c("16/10/12","19/11/12"),format="%d/%m/%y")
diff(some.dates)

## Time difference of 34 days
```

Facteur - variable catégorielle I

```
x <- list(1:5,c(1,2,3,4,5),a="test",
b=c(TRUE,FALSE),rpois(5,8))
x

## [[1]]
## [1] 1 2 3 4 5
##
## [[2]]
## [1] 1 2 3 4 5
##
## $a
## [1] "test"
##
## $b
## [1]  TRUE FALSE
##
## [[5]]
## [1]  6 13  9  4  7

f <- function(x) { return(x*(1-x)) }
```

Facteur - variable catégorielle II

```
optimize(f, interval=c(0, 1), maximum=TRUE)

## $maximum
## [1] 0.5
##
## $objective
## [1] 0.25

set.seed(1)
u <- runif(1)
if(u>.5) {("greater than 50%")} else {("smaller than
    50%")}

## [1] "smaller than 50%"

ifelse(u>.5,("greater than 50%"),("smaller than 50%"))

## [1] "smaller than 50%"

u
```

Facteur - variable catégorielle III

```
40
41 ## [1] 0.266
42
43 df <- data.frame(x=1:3,y=letters[1:3])
44 str(df)
45
46 ## 'data.frame':    3 obs. of  2 variables:
47 ##  $ x: int  1 2 3
48 ##  $ y: Factor w/ 3 levels "a","b","c": 1 2 3
49
50 typeof(df)
51
52 ## [1] "list"
53
54 class(df)
55
56 ## [1] "data.frame"
57
58 df$z<-5:3
59 df
```

Facteur - variable catégorielle IV

```
60
61 ##   x y z
62 ## 1 1 a 5
63 ## 2 2 b 4
64 ## 3 3 c 3
65
66 set.seed(1)
67 df[sample(nrow(df)),]
68
69 ##   x y z
70 ## 1 1 a 5
71 ## 3 3 c 3
72 ## 2 2 b 4
```

Facteur - variable catégorielle I

```
download.file("http://freakonometrics.free.fr/
    superheroes.RData","superheroes.RData")
load("superheroes.RData")
superheroes

##        name alignment gender          publisher
## 1   Magneto       bad   male             Marvel
## 2     Storm      good female             Marvel
## 3 Mystique       bad female             Marvel
## 4    Batman      good   male                 DC
## 5    Joker       bad   male                 DC
## 6 Catwoman       bad female                 DC
## 7   Hellboy      good   male Dark Horse Comics

publishers

##    publisher yr_founded
## 1         DC        1934
## 2     Marvel        1939

```

Facteur - variable catégorielle II

```
library(dplyr, verbose=FALSE)

inner_join(publishers, superheroes)

## Joining by: "publisher"

## Warning in inner_join_impl(x, y, by$x, by$y):
    joining factors with
## different levels, coercing to character vector

##   publisher yr_founded      name alignment gender
## 1        DC        1934    Batman      good   male
## 2        DC        1934     Joker       bad   male
## 3        DC        1934  Catwoman       bad female
## 4    Marvel        1939   Magneto       bad   male
## 5    Marvel        1939     Storm      good female
## 6    Marvel        1939  Mystique       bad female

merge(superheroes, publishers, all = TRUE)

```

Facteur - variable catégorielle III

```
##            publisher      name alignment gender
    yr_founded
## 1 Dark Horse Comics   Hellboy       good   male
       NA
## 2                DC    Batman       good   male
     1934
## 3                DC     Joker        bad   male
     1934
## 4                DC  Catwoman        bad female
     1934
## 5            Marvel   Magneto        bad   male
     1939
## 6            Marvel     Storm       good female
     1939
## 7            Marvel  Mystique        bad female
     1939
## 8             Image      <NA>       <NA>   <NA>
     1992

left_join(superheroes, publishers)
```

Facteur - variable catégorielle IV

```
50
51 ## Joining by: "publisher"
52
53 ## Warning in left_join_impl(x, y, by$x, by$y):
    joining factors with different
54 ## levels, coercing to character vector
55
56 ##       name alignment gender          publisher
    yr_founded
57 ## 1   Magneto       bad   male             Marvel
     1939
58 ## 2     Storm      good female             Marvel
     1939
59 ## 3 Mystique       bad female             Marvel
     1939
60 ## 4    Batman      good   male                DC
     1934
61 ## 5     Joker       bad   male                DC
     1934
```

Facteur - variable catégorielle V

```
## 6 Catwoman        bad female                    DC
     1934
## 7  Hellboy       good   male Dark Horse Comics
       NA

left_join(publishers, superheroes)

## Joining by: "publisher"

## Warning in left_join_impl(x, y, by$x, by$y):
    joining factors with different
## levels, coercing to character vector

##   publisher yr_founded     name alignment gender
## 1        DC       1934   Batman      good   male
## 2        DC       1934    Joker       bad   male
## 3        DC       1934 Catwoman       bad female
## 4    Marvel       1939  Magneto       bad   male
## 5    Marvel       1939    Storm      good female
## 6    Marvel       1939 Mystique       bad female
```

Facteur - variable catégorielle VI

```
79 ## 7       Image          1992       <NA>          <NA>     <NA>
```

Facteur - variable catégorielle I

```
download.file("http://freakonometrics.free.fr/
    gapminderDataFiveYear.txt","gapminderDataFiveYear.
    txt")
gdf <- read.delim("gapminderDataFiveYear.txt")
head(gdf,4)

##       country year      pop continent lifeExp
    gdpPercap
## 1 Afghanistan 1952  8425333      Asia    28.8
     779
## 2 Afghanistan 1957  9240934      Asia    30.3
     821
## 3 Afghanistan 1962 10267083      Asia    32.0
     853
## 4 Afghanistan 1967 11537966      Asia    34.0
     836

str(gdf)

## 'data.frame':    1704 obs. of  6 variables:
```

Facteur - variable catégorielle II

```
##  $ country   : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ year      : int  1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ pop       : num  8425333 9240934 10267083 11537966 13079460 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ lifeExp   : num  28.8 30.3 32 34 36.1 ...
##  $ gdpPercap: num  779 821 853 836 740 ...

subset(gdf, lifeExp < 30)

##           country year     pop continent lifeExp gdpPercap
## 1     Afghanistan 1952 8425333      Asia    28.8       779
## 1293      Rwanda 1992 7290203    Africa    23.6       737
```

Facteur - variable catégorielle III

```
gdf[gdf$lifeExp < 30,]

##           country year      pop continent lifeExp
    gdpPercap
## 1     Afghanistan 1952  8425333      Asia    28.8
       779
## 1293       Rwanda 1992  7290203    Africa    23.6
       737

gdf[gdf$country == "Italy", c("year", "lifeExp")]

##      year lifeExp
## 769  1952    65.9
## 770  1957    67.8
## 771  1962    69.2
## 772  1967    71.1
## 773  1972    72.2
## 774  1977    73.5
## 775  1982    75.0
## 776  1987    76.4
```

Facteur - variable catégorielle IV

```
## 777  1992     77.4
## 778  1997     78.8
## 779  2002     80.2
## 780  2007     80.5

small_df <- df[df$country %in% c("France","Italy","
    Spain"), c("country","year", "lifeExp")]

aggregate(small_df$lifeExp,FUN=max,by=list(
    small_df$country))

##    Group.1    x
## 1   France 80.7
## 2    Italy 80.5
## 3    Spain 80.9
```

Facteur - variable catégorielle I

```
library(gamair)
data(chicago)
head(chicago)

##    death pm10median pm25median o3median so2median
    time tmpd
## 1    130      -7.434          NA     -19.6      1.928
    -2556 31.5
## 2    150          NA          NA     -19.0     -0.986
    -2556 33.0
## 3    101      -0.827          NA     -20.2     -1.891
    -2554 33.0
## 4    135       5.566          NA     -19.7      6.139
    -2554 29.0
## 5    126          NA          NA     -19.2      2.278
    -2552 32.0
## 6    130       6.566          NA     -17.6      9.859
    -2552 40.0

base=data.frame(death=chicago$death,
```

Facteur - variable catégorielle II

```
14                    temp_F=chicago$tmpd,
15                    o3=chicago$o3median,
16                    date=seq(as.Date("1987-01-01"),
17                             as.Date("2000-12-31"),by=1))
18 base$temp_C <- (base$temp_F-32)/1.8
19 base$year <- substring(base$date,1,4)
20
21 date2season <- function(date){
22   m <- as.numeric(format(as.Date(date, format = "%d/%m
       /%Y"), "%m"))
23   d <- as.numeric(format(as.Date(date, format = "%d/%m
       /%Y"), "%d"))
24   s <- NA
25   if(m %in% c(1,2) | ((m==12)&(d>=21)) |  ((m==3)&(d
       <21))) s <- "winter"
26   if(m %in% c(4,5) | ((m==3)&(d>=21)) |  ((m==6)&(d
       <21))) s <- "spring"
27   if(m %in% c(7,8) | ((m==6)&(d>=21)) |  ((m==9)&(d
       <21))) s <- "summer"
```

Facteur - variable catégorielle III

```
28   if(m %in% c(10,11) | ((m==9)&(d>=21)) |  ((m==12)&(d
       <21))) s <- "autumn"
29   return(s)}
30 base$season <- sapply(base$date,date2season)
31 head(base)
32
33 ##    death temp_F     o3        date temp_C year season
34 ## 1    130   31.5 -19.6 1987-01-01 -0.278 1987 winter
35 ## 2    150   33.0 -19.0 1987-01-02  0.556 1987 winter
36 ## 3    101   33.0 -20.2 1987-01-03  0.556 1987 winter
37 ## 4    135   29.0 -19.7 1987-01-04 -1.667 1987 winter
38 ## 5    126   32.0 -19.2 1987-01-05  0.000 1987 winter
39 ## 6    130   40.0 -17.6 1987-01-06  4.444 1987 winter
40
41 plot(base[,c("date", "temp_C")])
```