数据读取实验

周世祥

2021/5/31

```
flights <- read.csv(file = "RawData/flights.csv")
str(object = flights)
```

```
## 'data.frame':    6 obs. of  6 variables:
##  $ carrier : chr  "UA" "UA" "AA" "B6" ...
##  $ flight  : int  1545 1714 1141 725 461 1696
##  $ tailnum : chr  "N14228" "N24211" "N619AA" "N804JB" ...
##  $ origin  : chr  "EWR" "LGA" "JFK" "JFK" ...
##  $ dest    : chr  "IAH" "IAH" "MIA" "BQN" ...
##  $ air_time: int  227 227 160 183 116 150
```

```
flights1 <- read.csv(file = "RawData/flights1.csv")
str(object = flights1)
```

```
## 'data.frame':    6 obs. of  1 variable:
##  $ carrier.flight.tailnum.origin.dest.air_time: chr  "UA\t1545\tN14228\tEWR\tIAH\t227" "UA\t171
```

```
flights3 <- read.csv(file = "RawData/flights1.csv", sep = "\t")
str(flights3)
```

```
## 'data.frame':    6 obs. of  6 variables:
##  $ carrier : chr  "UA" "UA" "AA" "B6" ...
##  $ flight  : int  1545 1714 1141 725 461 1696
##  $ tailnum : chr  "N14228" "N24211" "N619AA" "N804JB" ...
##  $ origin  : chr  "EWR" "LGA" "JFK" "JFK" ...
##  $ dest    : chr  "IAH" "IAH" "MIA" "BQN" ...
##  $ air_time: int  227 227 160 183 116 150
```

```
flights_str <- read.csv(file = "RawData/flightsstrings.csv", sep = "\t", stringsAsFactors = FALSE)
str(object = flights_str)
```

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

```
##  $ carrier : chr  "UA" "UA" "AA" "B6" ...
##  $ flight  : int  1545 1714 1141 725 461 1696
##  $ tailnum : chr  "N14228" "N24211" "N619AA" "N804JB" ...
##  $ origin  : chr  "EWR" "LGA" "JFK" "JFK" ...
##  $ dest    : chr  "IAH" "IAH" "MIA" "BQN" ...
##  $ air_time: int  227 227 160 183 116 150
```

```
flights <- read.table(file = "RawData/flights.csv")
head(x = flights)
```

```
##                                          V1
## 1 carrier,flight,tailnum,origin,dest,air_time
## 2                  UA,1545,N14228,EWR,IAH,227
## 3                  UA,1714,N24211,LGA,IAH,227
## 4                  AA,1141,N619AA,JFK,MIA,160
## 5                   B6,725,N804JB,JFK,BQN,183
## 6                   DL,461,N668DN,LGA,ATL,116
```

```
flights <- read.table(file = "RawData/flights.csv",header = TRUE)
head(x = flights)
```

```
##   carrier.flight.tailnum.origin.dest.air_time
## 1                  UA,1545,N14228,EWR,IAH,227
## 2                  UA,1714,N24211,LGA,IAH,227
## 3                  AA,1141,N619AA,JFK,MIA,160
## 4                   B6,725,N804JB,JFK,BQN,183
## 5                   DL,461,N668DN,LGA,ATL,116
## 6                  UA,1696,N39463,EWR,ORD,150
```

```
flights <- read.table(file = "RawData/flights.csv",header = TRUE,sep = ",")
head(flights)
```

```
##   carrier flight tailnum origin dest air_time
## 1      UA   1545  N14228    EWR  IAH      227
## 2      UA   1714  N24211    LGA  IAH      227
## 3      AA   1141  N619AA    JFK  MIA      160
## 4      B6    725  N804JB    JFK  BQN      183
## 5      DL    461  N668DN    LGA  ATL      116
## 6      UA   1696  N39463    EWR  ORD      150
```

```
# airlines <- read.table(file = "RawData/airlines.csv", header = TRUE, sep = "\t", blank.lines.ski
# head(airlines, n = 8)
```

```
airlines <- read.table(file = "RawData/airlines.csv", header = FALSE, sep = "\t", stringsAsFactors
head(airlines)
```

```
##           V1                       V2                      V3
## 1    c!rrier                     n`me
## 2         AA   Aoer)can Airlines Inc.
## 3         B7 Je4Blue Air\177ays\002DL D網秤a Ahr Liles Inc.J
## 4    carrier                   flight                 tailnum
## 5 AA-114\021                   N619AA                     JFK
## 6         B6                       ??                  N804JB
##                        V4  V5  V6
## 1
## 2
## 3
## 4 origin濉狒楂忏¨ir_time
## 5                    MIA 160
## 6                    JFK BQN 18;
```

```
number_of_col <- max(count.fields("RawData/airlines.csv",sep = "\t"))
airlines <- read.table(file = "RawData/airlines.csv", header = FALSE, sep = "\t", stringsAsFactors
head(airlines)
```

```
##           V1                       V2                      V3
## 1    c!rrier                     n`me
## 2         AA   Aoer)can Airlines Inc.
## 3         B7 Je4Blue Air\177ays\002DL D網秤a Ahr Liles Inc.J
## 4    carrier                   flight                 tailnum
## 5 AA-114\021                   N619AA                     JFK
## 6         B6                       ??                  N804JB
##                        V4  V5  V6
## 1
## 2
## 3
## 4 origin濉狒楂忏¨ir_time
## 5                    MIA 160
## 6                    JFK BQN 18;
```

```
flights_uneven <- read.table("RawData/airlines.csv", header = FALSE, sep = "\t", stringsAsFactors
head(flights_uneven)
```

```
##           V1                       V2                      V3
```

```
## 1    c!rrier                     n`me
## 2         AA   Aoer)can Airlines Inc.
## 3         B7 Je4Blue Air\177ays\002DL D鍚秤a Ahr Liles Inc.J
## 4    carrier                   flight                tailnum
## 5 AA-114\021                   N619AA                    JFK
## 6         B6                       ??                N804JB
##                        V4  V5
## 1
## 2
## 3
## 4 origin濉猟楂忏¨ir_time
## 5                    MIA 160
## 6                    JFK BQN
```

```
flights_uneven <- read.table(file = "RawData/flights_uneven.csv", header = FALSE, sep = "\t", stri
head(flights_uneven)
```

```
##        V1     V2      V3     V4   V5       V6    V7
## 1 carrier flight tailnum origin dest air_time  <NA>
## 2      UA   1545  N14228    EWR  IAH      227  <NA>
## 3      UA   1714  N24211    LGA  IAH      227 测试1
## 4      AA   1141  N619AA    JFK  MIA      160 测试2
## 5      B6    725  N804JB    JFK  BQN      183 测试3
## 6      DL    461  N668DN    LGA  ATL      116  <NA>
```

```
flights_uneven <- read.table("RawData/flights_uneven.csv",sep = "\t", stringsAsFactors = FALSE, fi
head(flights_uneven)
```

```
##        V1     V2      V3     V4   V5       V6 V7
## 1 carrier flight tailnum origin dest air_time NA
## 2      UA   1545  N14228    EWR  IAH      227 NA
## 3      UA   1714  N24211    LGA  IAH      227 NA
## 4      AA   1141  N619AA    JFK  MIA      160 NA
## 5      B6    725  N804JB    JFK  BQN      183 NA
## 6      DL    461  N668DN    LGA  ATL      116 NA
```

```
flights_uneven <- read.table("RawData/flights_uneven.csv",sep = "\t", stringsAsFactors = FALSE, fi
replace <- unique(flights_uneven$V7)
replace
```

```
## [1] ""      "测试1" "测试2" "测试3"
```

```
flights_uneven <- read.table("RawData/flights_uneven.csv",sep = "\t", stringsAsFactors = FALSE, fi
head(flights_uneven)
```

```
##        V1     V2      V3     V4   V5       V6    V7
## 1 carrier flight tailnum origin dest air_time  <NA>
## 2      UA   1545  N14228    EWR  IAH      227  <NA>
## 3      UA   1714  N24211    LGA  IAH      227 测试1
## 4      AA   1141  N619AA    JFK  MIA      160  <NA>
## 5      B6    725  N804JB    JFK  BQN      183 测试3
## 6      DL    461  N668DN    LGA  ATL      116  <NA>
```

```
library(tidyverse)
```

```
read_csv("RawData/flights_large.csv")
```

```
flights <- read_csv("RawData/flights_large.csv")
str(flights)
```

```
system.time(read.csv("RawData/flights_large.csv",stringsAsFactors = FALSE))
system.time(read_csv("RawData/flights_large.csv"))
```

```
flights <- read_delim("RawData/flights_large2.csv",delim = "_")
problems(flights)
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.0.5
```

```
readxl_example()[4]
```

```
## [1] "datasets.xlsx"
```

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.0.3
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
fread(file = "./RawData/airlines.csv", header = F, sep = "\t", blank.lines.skip = FALSE)
```

```
## Warning in fread(file = "./RawData/airlines.csv", header = F, sep = "\t", :
## Stopped early on line 3. Expected 2 fields but found 3. Consider fill=TRUE and
```

```
## Liles Inc.J>>
```

```
##         V1                    V2
## 1: c!rrier                  n`me
## 2:      AA Aoer)can Airlines Inc.
```

```
library(tidyverse)
readxl_example(path = "datasets.xlsx")
```

```
## [1] "C:/Users/zhoushixiang/Documents/R/win-library/4.0/readxl/extdata/datasets.xlsx"
```

```
iris <- read_excel(path = readxl_example(path = "datasets.xlsx"))
str(iris)
```

```
## tbl_df [150 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Sepal.Length: num [1:150] 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num [1:150] 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num [1:150] 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num [1:150] 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : chr [1:150] "setosa" "setosa" "setosa" "setosa" ...
```

```
excel_sheets(path = readxl_example(path = "datasets.xlsx"))
```

```
## [1] "iris"     "mtcars"   "chickwts" "quakes"
```

```
mtcars <- read_excel(path = readxl_example(path = "datasets.xlsx"),sheet = 2)
mtcars <- read_excel(path = readxl_example(path = "datasets.xlsx"),sheet = "mtcars")
```

```
library(pdftools)
```

```
## Warning: package 'pdftools' was built under R version 4.0.5
```

```
## Using poppler version 21.04.0
```

```
text<- pdf_text("./RawData/pdftools.pdf")
length(text)
```

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

```
class(text)
```

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

```
text
```

```
## [1] "                                Package  ‘pdftools’ \n
## [2] "2
## [3] "pdf_render_page
## [4] "4                                                                                pdf_render_pag
## [5] "Index\n\npdf_attachments (pdf_info), 2\npdf_convert (pdf_render_page), 3\npdf_data (pdf_in
```

```
print(text)
```

```
## [1] "                                Package  ‘pdftools’ \n
## [2] "2
## [3] "pdf_render_page
## [4] "4                                                                                pdf_render_pag
## [5] "Index\n\npdf_attachments (pdf_info), 2\npdf_convert (pdf_render_page), 3\npdf_data (pdf_in
```

```
text[1]
```

```
## [1] "                                Package  ‘pdftools’ \n
```

```
pdf_info("./RawData/pdftools.pdf")
```

```
## $version
## [1] "1.5"
##
## $pages
## [1] 5
##
## $encrypted
## [1] FALSE
##
## $linearized
## [1] FALSE
##
## $keys
## $keys$Author
## [1] ""
##
## $keys$Title
## [1] ""
##
```

```
## $keys$Subject
## [1] ""
##
## $keys$Creator
## [1] "LaTeX with hyperref package"
##
## $keys$Producer
## [1] "pdfTeX-1.40.15"
##
## $keys$Keywords
## [1] ""
##
## $keys$Trapped
## [1] ""
##
## $keys$PTEX.Fullbanner
## [1] "This is pdfTeX, Version 3.14159265-2.6-1.40.15 (TeX Live 2015/dev/Debian) kpathsea version
##
##
## $created
## [1] "2018-05-27 21:56:10 CST"
##
## $modified
## [1] "2018-05-27 21:56:10 CST"
##
## $metadata
## [1] ""
##
## $locked
## [1] FALSE
##
## $attachments
## [1] FALSE
##
## $layout
## [1] "no_layout"
```

```
pdf_attachments("./RawData/pdftools.pdf")
```

```
## list()
```

```
pdf_fonts("./RawData/pdftools.pdf")
```

```
## # A tibble: 6 x 4
##   name                                type  embedded file
##   <chr>                               <chr> <lgl>    <chr>
## 1 DSHWTW+NimbusRomNo9L-Medi           type1 TRUE     ""
## 2 UTHPMJ+NimbusRomNo9L-Regu           type1 TRUE     ""
## 3 DSQFGA+Inconsolata-zi4r             type1 TRUE     ""
## 4 LVIJIF+NimbusSanL-Regu              type1 TRUE     ""
## 5 DQRZJT+NimbusRomNo9L-Regu-Slant_167 type1 TRUE     ""
## 6 YIECHJ+NimbusRomNo9L-ReguItal       type1 TRUE     ""
```

```
pdf_toc(pdf = "./RawData/pdftools.pdf")
```

```
## $title
## [1] ""
##
## $children
## $children[[1]]
## $children[[1]]$title
## [1] "pdf_info"
##
## $children[[1]]$children
## list()
##
##
## $children[[2]]
## $children[[2]]$title
## [1] "pdf_render_page"
##
## $children[[2]]$children
## list()
##
##
## $children[[3]]
## $children[[3]]$title
## [1] "Index"
##
## $children[[3]]$children
## list()
```

```
library(jsonlite)
```

```
## Warning: package 'jsonlite' was built under R version 4.0.5
```

```
##
## Attaching package: 'jsonlite'
```

```
## The following object is masked from 'package:purrr':
##
##     flatten
```

```
jsonlite::toJSON(x = pdf_toc(pdf = "./RawData/pdftools.pdf"), pretty = TRUE)
```

```
## {
##   "title": [""],
##   "children": [
##     {
##       "title": ["pdf_info"],
##       "children": []
##     },
##     {
##       "title": ["pdf_render_page"],
##       "children": []
##     },
##     {
##       "title": ["Index"],
##       "children": []
##     }
##   ]
## }
```

```
l <- toJSON(iris,pretty = T)
identical(fromJSON(l,simplifyDataFrame = T),iris)
```

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

```
example <-  '["a", "b", 0, "c"]'
class(example)
```

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

```
example
```

```
## [1] "[\"a\", \"b\", 0, \"c\"]"
```

```
fromJSON(example)
```

```
## [1] "a" "b" "0" "c"
```

```
fromJSON(example,simplifyVector = F)
```

```
## [[1]]
## [1] "a"
##
## [[2]]
## [1] "b"
##
## [[3]]
## [1] 0
##
## [[4]]
## [1] "c"
```