

# 10ImportarDatos\_cheatsheet.R

moka

2023-04-21

```
# Autor: Monika Avila Marquez, Ph.D.
# Fecha: 12.04.2023
# Objetivo: Importar datos que no estan en R de excel, csv, etc.
# Referencia: Basado en R Programming Fundamentals, StanfordOnline XDFS112

# Limpiar el espacio de trabajo
rm(list=ls())

# Configurar el directorio
midirectorio<-setwd("~/Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/10ImportarDatos")
midirectorio

## [1] "/Users/moka/Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/10ImportarDatos"

# Obetener los archivos disponibles en el directorio
dir()

## [1] "10ImportarDatos_cheatsheet.aux"      "10ImportarDatos_cheatsheet.html"    "10ImportarDatos_cheatsheet.R"
## [4] "10ImportarDatos_cheatsheet.R"        "10ImportarDatos_cheatsheet.spin.R"  "10ImportarDatos_cheatsheetLatex.aux"
## [7] "10ImportarDatos_cheatsheetLatex.aux" "10ImportarDatos_cheatsheetLatex.tex" "10ImportarDatos_conceptos.tex"
## [10] "10ImportarDatos_conceptos.tex"       "10ImportarDatos_lab.aux"            "10ImportarDatos_lab.tex"
## [13] "10ImportarDatos_lab.tex"

list.files()

## [1] "10ImportarDatos_cheatsheet.aux"      "10ImportarDatos_cheatsheet.html"    "10ImportarDatos_cheatsheet.R"
## [4] "10ImportarDatos_cheatsheet.R"        "10ImportarDatos_cheatsheet.spin.R"  "10ImportarDatos_cheatsheetLatex.aux"
## [7] "10ImportarDatos_cheatsheetLatex.aux" "10ImportarDatos_cheatsheetLatex.tex" "10ImportarDatos_conceptos.tex"
## [10] "10ImportarDatos_conceptos.tex"       "10ImportarDatos_lab.aux"            "10ImportarDatos_lab.tex"
## [13] "10ImportarDatos_lab.tex"

# Importar datos en R

# 1. Archivo txt (archivo tipo ascii, con extension .txt): este tipo. de archivo es lisible por humanos
# de archivos son mas dificiles de leer.
# Ver el archivo
#file.show("~/Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/datos/creados/math2.txt")
# Importar los datos
datostxt<-read.table("~/Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/datos/creados/math2.txt")
# Ver los datos
datostxt

##           v1           v2
## I1 0.2740892 0.6317251
## I2 0.5920839 0.1891507
```

```

# Ver que tipo de datos es
class(datostxt)

## [1] "data.frame"

str(datostxt)

## 'data.frame':    2 obs. of  2 variables:
## $ v1: num  0.274 0.592
## $ v2: num  0.632 0.189

is.matrix(datostxt)

## [1] FALSE

# We call this a tidy table because each row is an observation and each column is a variable
# 2. Archivo csv (comma separated variables): la ventaja es que se puede ver
#este archivo en R
?dir
dir("datos/raw")

## character(0)

# Ver los datos
file.show("~/Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/datos/raw/InformalTotal.csv")
# Importar datos
?read.csv # Default para el separadores ,
InformalTotal<-read.csv("~/Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/datos/raw/InformalTotal.csv")
head(InformalTotal)

##      X.1 Reference.area                Source  Sex Time Total Agriculture Non.agriculture
## 1      1  Afghanistan LFS - Labour Force Survey Total 2021  86.4      98.4      82.4
## 2      4    Albania LFS - Labour Force Survey Total 2012  66.6      97.6      40.2
## 3      7    Albania LFS - Labour Force Survey Total 2013  61.0      96.4      33.1
## 4     10    Albania LFS - Labour Force Survey Total 2014  65.6      99.6      41.0
## 5     13    Albania LFS - Labour Force Survey Total 2015  63.4      99.5      38.1
## 6     16    Albania LFS - Labour Force Survey Total 2016  61.6      99.5      36.4
##
## 1 Notes\n\n[8.3.1] Proportion of informal employment in total employment (%) \n-Repository: ILO-STAT
## 2 Notes\n\n[8.3.1] Proportion of informal employment in total employment (%) \n-Repository: ILO-STAT
## 3 Notes\n\n[8.3.1] Proportion of informal employment in total employment (%) \n-Repository: ILO-STAT
## 4 Notes\n\n[8.3.1] Proportion of informal employment in total employment (%) \n-Repository: ILO-STAT
## 5 Notes\n\n[8.3.1] Proportion of informal employment in total employment (%) \n-Repository: ILO-STAT
## 6 Notes\n\n[8.3.1] Proportion of informal employment in total employment (%) \n-Repository: ILO-STAT

# Debugging read.csv output
# Si el nombre de las filas es considerado como una variable pero nosotros no queremos esto
# llamada individual, pero inicialmente este es solamente un identificador
# Para evitar esto, podemos poner las opciones de rownames, y colnames.
InformalTotal1<-read.csv("~/Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/datos/raw/InformalTotal.csv")
head(InformalTotal1)

##      Reference.area                Source  Sex Time Total Agriculture Non.agriculture
## 1      Afghanistan LFS - Labour Force Survey Total 2021  86.4      98.4      82.4
## 4      Albania LFS - Labour Force Survey Total 2012  66.6      97.6      40.2
## 7      Albania LFS - Labour Force Survey Total 2013  61.0      96.4      33.1
## 10     Albania LFS - Labour Force Survey Total 2014  65.6      99.6      41.0
## 13     Albania LFS - Labour Force Survey Total 2015  63.4      99.5      38.1

```

```
## 16      Albania LFS - Labour Force Survey Total 2016  61.6      99.5      36.4
##
## 1  Notes\n\n[8.3.1] Proportion of informal employment in total employment (%) \n-Repository: ILO-STA
## 4      Notes\n\n[8.3.1] Proportion of informal employment in tota
## 7      Notes\n\n[8.3.1] Proportion of informal employment in tota
## 10     Notes\n\n[8.3.1] Proportion of informal employment in tota
## 13     Notes\n\n[8.3.1] Proportion of informal employment in tota
## 16     Notes\n\n[8.3.1] Proportion of informal employment in tota
```

```
# Ahora, ya no tenemos la variable individual como variable factor.
```

```
# Utilizar la funcion scan
```

```
scan("~/Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/datos/raw/InformalTotal.csv",nlines=5,what="")
```

```
## [1] ""
## [2] ", \"Reference.area\", \"Source\", \"Sex\", \"Time\", \"Total\", \"Agriculture\", \"Non.agriculture\", \"
## [3] "1"
## [4] ", \"Afghanistan\", \"LFS"
## [5] "- "
## [6] "Labour"
## [7] "Force"
## [8] "Survey\", \"Total\", 2021, 86.4, 98.4, 82.4, \"Notes"
## [9] "[8.3.1]"
## [10] "Proportion"
## [11] "of"
## [12] "informal"
## [13] "employment"
## [14] "in"
## [15] "total"
## [16] "employment"
## [17] "(%)"
## [18] "-Repository:"
## [19] "ILO-STATISTICS"
## [20] "- "
## [21] "Micro"
## [22] "data"
## [23] "processing"
```

```
?scan
```

```
# 3. Archivo Excel con extensiones .xlsx, o .xls
```

```
#install.packages("readxl")
```

```
library(readxl)
```

```
?read_excel
```

```
Informal<-read_excel("~/Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/datos/raw/Informal.xlsx")
```

```
## New names:
```

```
## * `` -> `...8`
```

```
str(Informal)
```

```
## tibble [2,371 x 8] (S3: tbl_df/tbl/data.frame)
## $ Reference area : chr [1:2371] "Afghanistan" "Afghanistan" "Afghanistan" "Albania" ...
## $ Source : chr [1:2371] "LFS - Labour Force Survey" "LFS - Labour Force Survey" "LFS - Labor
## $ Sex : chr [1:2371] "Total" "Male" "Female" "Total" ...
## $ Time : num [1:2371] 2021 2021 2021 2012 2012 ...
## $ Total : num [1:2371] 86.4 83 97.3 66.6 63.4 70.6 61 58.9 63.5 65.6 ...
## $ Agriculture : num [1:2371] 98.4 97.8 99.5 97.6 96.2 98.9 96.4 93.7 98.7 99.6 ...
```

```
## $ Non-agriculture: num [1:2371] 82.4 78.9 96.1 40.2 43.8 33.5 33.1 38.6 24.1 41 ...
## $ ...8 : chr [1:2371] "Notes\r\n\r\n[8.3.1] Proportion of informal employment in total em

class(Informal)

## [1] "tbl_df"      "tbl"          "data.frame"

# read_excel uploads the data in a different format, and we can just convert it to dataframe
# Convert it to a data frame
Informaldf<-as.data.frame(Informal)
str(Informaldf)

## 'data.frame': 2371 obs. of 8 variables:
## $ Reference area : chr "Afghanistan" "Afghanistan" "Afghanistan" "Albania" ...
## $ Source : chr "LFS - Labour Force Survey" "LFS - Labour Force Survey" "LFS - Labour Force
## $ Sex : chr "Total" "Male" "Female" "Total" ...
## $ Time : num 2021 2021 2021 2012 2012 ...
## $ Total : num 86.4 83 97.3 66.6 63.4 70.6 61 58.9 63.5 65.6 ...
## $ Agriculture : num 98.4 97.8 99.5 97.6 96.2 98.9 96.4 93.7 98.7 99.6 ...
## $ Non-agriculture: num 82.4 78.9 96.1 40.2 43.8 33.5 33.1 38.6 24.1 41 ...
## $ ...8 : chr "Notes\r\n\r\n[8.3.1] Proportion of informal employment in total employment

# 4. Importar datos de otros softwares
#install.packages("foreign")
library(foreign)
# spss .sav, read.spss
# stata .dta, read.dta
?read.dta

# 5. Scraping la web y leyendo datos directamente (se necesita conexion al internet)
# Datos que vienen de un Github
trafico<-read.csv(url("https://raw.githubusercontent.com/monikavila/Curso-Introduccion-R/main/datos/
head(trafico)

## year month date_of_month day_of_week autos
## 1 2000 1 1 6 9084.165
## 2 2000 1 2 7 8005.351
## 3 2000 1 3 1 11363.158
## 4 2000 1 4 2 13030.962
## 5 2000 1 5 3 12559.135
## 6 2000 1 6 4 12465.477

# Nota: Scraping datos con paquete rvest
# Bonus: Guardar archivo
write.csv(trafico,file=~ /Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/datos/procesados/trafico1.csv)
file.show(~ /Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/datos/procesados/trafico1.csv")
# Mejores practicas creando archivos de datos: no espacios en nombre de variables
# csv ocupa bastante espacio, se puede crear un archivo binario usando
# save file en binary R formato
save(trafico,file=~ /Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/datos/procesados/trafico1.Rdata")
load(~ /Dropbox/0.POST-PHD/GOALS/2.CODE/R/Ecomienza/datos/procesados/trafico1.Rdata")
# Formato binario de R es mas eficiente y rapido de leer
# Recursos: R import data
# https://cran.r-project.org/doc/manuals/r-devel/R-data.html

# No quitar el comentario de la linea inferior. Solamente copiar en la consola para que ejecute
#rmarkdown::render("10ImportarDatos_cheatsheet.R",c("pdf_document","html_document"))
```