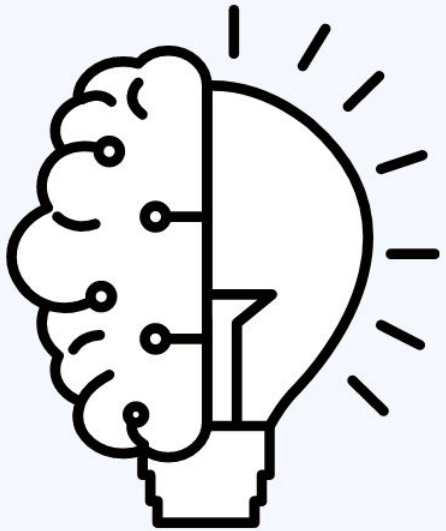
Ciencia de datos

Para el sector público de salud



Importación de datos

En colaboración con...

@RLadiesConce

DATA UC
Estudios y Servicios Estadísticos

MÓDULO 1: Nivelación y conceptos básicos

Actividades sincrónicas (2 hrs cada uno)

Fecha	Hora	Tema
09/12/2020	18.45h	Aspectos generales del curso
15/12/2020	18.45h	Introducción a R y RStudio
17/12/2020	18.45h	Estructura de datos y operadores
22/12/2020	**18.45h**	**Importación de datos**
29/12/2020	18.45h	Análisis prefactibilidad y valor público

¿Qué datos se pueden importar a R?

… Muchos!

Archivos txt o csv

Los datos planos o ASCII, en general, tienen dos características:

- **El header** o encabezamiento
- **La separación de caracteres que indican la separación de columnas**: pueden estar separadas por comas, puntos y comas, por tabulación, etc.

Sin encabezado

```
0.2593466,0,33.25119781,1105.642156,11.25,1,0,16,8.68,1,0,16,0,1.333,0,0
.,-1,54.05338809,2921.768764,.,1,0,9,7.85,1,0,10,1,8,1,0
0.721318058,7,43.57015743,1898.358618,18,1,0,19,8.75,1,0,12,4,3,-1,0
0.011581964,0,30.96783025,959.0065106,16.5,1,1,12,16.31,1,1,12,0,-2,0,1
-0.560984677,0,34.63381246,1199.500965,9.6154,1,1,14,16.85,1,1,14,1,2.917,0,-1
.,2,71.60301164,5126.991275,.,1,0,16,.,1,0,14,-2,24,1,0
1.523260216,-2,34.97878166,1223.515166,35,1,0,13,7.63,1,0,15,-2,3,1,0
.,-1,61.45106092,3776.232888,35,1,0,13,.,1,0,14,-2,25.5,0,0
-0.223143551,-2,29.33880904,860.7657156,12,1,1,12,15,1,1,14,1,1,-1,0
-0.470003629,0,47.60574949,2266.307384,6.25,1,0,12,10,1,0,12,0,-5,0,0
0.051751065,-1,51.90143737,2693.759201,17.25,1,1,12,16.38,1,1,13,-2,9,0,0
0.287682073,0,36.07665982,1301.525384,16,1,1,12,12,1,1,12,0,5,0,0
```

Con encabezado

```
DLHRWAGE,DEDUC1,AGE,AGESQ,HRWAGEH,WHITEH,MALEH,EDUCH,HRWAGEL,WHITEL,MALEL,EDUCL,DEDUC2,DTEN,DMARRIED,DUNCOV
0.2593466,0,33.25119781,1105.642156,11.25,1,0,16,8.68,1,0,16,0,1.333,0,0
.,-1,54.05338809,2921.768764,.,1,0,9,7.85,1,0,10,1,8,1,0
0.721318058,7,43.57015743,1898.358618,18,1,0,19,8.75,1,0,12,4,3,-1,0
0.011581964,0,30.96783025,959.0065106,16.5,1,1,12,16.31,1,1,12,0,-2,0,1
-0.560984677,0,34.63381246,1199.500965,9.6154,1,1,14,16.85,1,1,14,1,2.917,0,-1
.,2,71.60301164,5126.991275,.,1,0,16,.,1,0,14,-2,24,1,0
1.523260216,-2,34.97878166,1223.515166,35,1,0,13,7.63,1,0,15,-2,3,1,0
.,-1,61.45106092,3776.232888,35,1,0,13,.,1,0,14,-2,25.5,0,0
-0.223143551,-2,29.33880904,860.7657156,12,1,1,12,15,1,1,14,1,1,-1,0
-0.470003629,0,47.60574949,2266.307384,6.25,1,0,12,10,1,0,12,0,-5,0,0
0.051751065,-1,51.90143737,2693.759201,17.25,1,1,12,16.38,1,1,13,-2,9,0,0
0.287682073,0,36.07665982,1301.525384,16,1,1,12,12,1,1,12,0,5,0,0
.,4,32.08213552,1029.26342,.,1,0,16,31.25,1,0,12,4,-9.5,0,0
```

Con encabezado falso

```
1  ===================================================
2  KEYWORDS FOR DATASET: Income, Education Level, Twins
3  ===================================================
4
5  ================================================================
6  ACCOMPANYING DATA PROVIDED BY: Guido Imbens, PhD
7                                 UCLA, Department of Economics
8  ================================================================
9  DLHRWAGE,DEDUC1,AGE,AGESQ,HRWAGEH,WHITEH,MALEH,EDUCH,HRWAGEL,WHITEL,MALEL,EDUCL,DEDUC2,DTEN,DMARRIED,DUNCOV
10 0.2593466,0,33.25119781,1105.642156,11.25,1,0,16,8.68,1,0,16,0,1.333,0,0
11 .,-1,54.05338809,2921.768764,.,1,0,9,7.85,1,0,10,1,8,1,0
12 0.721318058,7,43.57015743,1898.358618,18,1,0,19,8.75,1,0,12,4,3,-1,0
13 0.011581964,0,30.96783025,959.0065106,16.5,1,1,12,16.31,1,1,12,0,-2,0,1
14 -0.560984677,0,34.63381246,1199.500965,9.6154,1,1,14,16.85,1,1,14,1,2.917,0,-1
15 .,2,71.60301164,5126.991275,.,1,0,16,.,1,0,14,-2,24,1,0
16 1.523260216,-2,34.97878166,1223.515166,35,1,0,13,7.63,1,0,15,-2,3,1,0
17 .,-1,61.45106092,3776.232888,35,1,0,13,.,1,0,14,-2,25.5,0,0
```

¿Cómo importar txt o csv?

library(readr) <- Recomendada

Función R base read.table()

Librería readr (tidyverse)

- read_csv(): para leer archivos con coma (“,”) como separador
- read_csv2(): para leer archivos con punto y coma (“;”) como separador
- read_tsv(): para leer archivos con tabulador (“\t”) como separador
- read_delim(, sep = “|”): para leer archivos con separador distintos como puede ser el símbolo |

read_csv(“la_ruta_del_archivo”)

read_csv(“nombre_archivo.csv”) <- Recomendado

read_csv(file.choose))

Sin encabezado

```
0.2593466,0,33.25119781,1105.642156,11.25,1,0,16,8.68,1,0,16,0,1.333,0,0
.,-1,54.05338809,2921.768764,.,1,0,9,7.85,1,0,10,1,8,1,0
0.721318058,7,43.57015743,1898.358618,18,1,0,19,8.75,1,0,12,4,3,-1,0
0.011581964,0,30.96783025,959.0065106,16.5,1,1,12,16.31,1,1,12,0,-2,0,1
-0.560984677,0,34.63381246,1199.500965,9.6154,1,1,14,16.85,1,1,14,1,2.917,0,-1
.,2,71.60301164,5126.991275,.,1,0,16,.,1,0,14,-2,24,1,0
1.523260216,-2,34.97878166,1223.515166,35,1,0,13,7.63,1,0,15,-2,3,1,0
.,-1,61.45106092,3776.232888,35,1,0,13,.,1,0,14,-2,25.5,0,0
-0.223143551,-2,29.33880904,860.7657156,12,1,1,12,15,1,1,14,1,1,-1,0
-0.470003629,0,47.60574949,2266.307384,6.25,1,0,12,10,1,0,12,0,-5,0,0
0.051751065,-1,51.90143737,2693.759201,17.25,1,1,12,16.38,1,1,13,-2,9,0,0
0.287682073,0,36.07665982,1301.525384,16,1,1,12,12,1,1,12,0,5,0,0
```

read_csv(“archivo.csv”, skip = 0, col_names = FALSE)
read.table(“archivo.csv”, skip = 0, header = FALSE, sep = “,”)

Con encabezado

```
DLHRWAGE,DEDUC1,AGE,AGESQ,HRWAGEH,WHITEH,MALEH,EDUCH,HRWAGEL,WHITEL,MALEL,EDUCL,DEDUC2,DTEN,DMARRIED,DUNCOV
0.2593466,0,33.25119781,1105.642156,11.25,1,0,16,8.68,1,0,16,0,1.333,0,0
.,-1,54.05338809,2921.768764,.,1,0,9,7.85,1,0,10,1,8,1,0
0.721318058,7,43.57015743,1898.358618,18,1,0,19,8.75,1,0,12,4,3,-1,0
0.011581964,0,30.96783025,959.0065106,16.5,1,1,12,16.31,1,1,12,0,-2,0,1
-0.560984677,0,34.63381246,1199.500965,9.6154,1,1,14,16.85,1,1,14,1,2.917,0,-1
.,2,71.60301164,5126.991275,.,1,0,16,.,1,0,14,-2,24,1,0
1.523260216,-2,34.97878166,1223.515166,35,1,0,13,7.63,1,0,15,-2,3,1,0
.,-1,61.45106092,3776.232888,35,1,0,13,.,1,0,14,-2,25.5,0,0
-0.223143551,-2,29.33880904,860.7657156,12,1,1,12,15,1,1,14,1,1,-1,0
-0.470003629,0,47.60574949,2266.307384,6.25,1,0,12,10,1,0,12,0,-5,0,0
0.051751065,-1,51.90143737,2693.759201,17.25,1,1,12,16.38,1,1,13,-2,9,0,0
0.287682073,0,36.07665982,1301.525384,16,1,1,12,12,1,1,12,0,5,0,0
.,4,32.08213552,1029.26342,.,1,0,16,31.25,1,0,12,4,-9.5,0,0
```

read_csv(“archivo.csv”, skip = 0, col_names = TRUE)
read.table(“archivo.csv”, skip = 0, header = TRUE, sep =’,’)

Con encabezado falso

```
1  ===================================================
2  KEYWORDS FOR DATASET: Income, Education Level, Twins
3  ===================================================
4
5  ============================================================
6  ACCOMPANYING DATA PROVIDED BY: Guido Imbens, PhD
7                                 UCLA, Department of Economics
8  ============================================================
9  DLHRWAGE,DEDUC1,AGE,AGESQ,HRWAGEH,WHITEH,MALEH,EDUCH,HRWAGEL,WHITEL,MALEL,EDUCL,DEDUC2,DTEN,DMARRIED,DUNCOV
10 0.2593466,0,33.25119781,1105.642156,11.25,1,0,16,8.68,1,0,16,0,1.333,0,0
11 .,-1,54.05338809,2921.768764,.,1,0,9,7.85,1,0,10,1,8,1,0
12 0.721318058,7,43.57015743,1898.358618,18,1,0,19,8.75,1,0,12,4,3,-1,0
13 0.011581964,0,30.96783025,959.0065106,16.5,1,1,12,16.31,1,1,12,0,-2,0,1
14 -0.560984677,0,34.63381246,1199.500965,9.6154,1,1,14,16.85,1,1,14,1,2.917,0,-1
15 .,2,71.60301164,5126.991275,.,1,0,16,.,1,0,14,-2,24,1,0
16 1.523260216,-2,34.97878166,1223.515166,35,1,0,13,7.63,1,0,15,-2,3,1,0
17 .,-1,61.45106092,3776.232888,35,1,0,13,.,1,0,14,-2,25.5,0,0
```

read_csv(“archivo.csv”, skip = 8, col_names = TRUE)
read.table(“archivo.csv”, skip = 8, header = TRUE, sep =’,’)

¿Cómo importar archivos Excel?

library(readxl) <- Recomendada

Usar la funcionalidad de RStudio “Import Dataset”

Usando un truco poco estético

¿Qué más puedo importar?

library(haven)

https://haven.tidyverse.org/

SPSS: read_sav()

STATA: read_dta()

SAS: read_sas()

… Y algo más??

PDF scraping

Web scraping

Audio / Video

Imágenes

Redes sociales

Datos vectoriales y ruster (QGIS)

Archivos de internet

Desde bases de datos (SQL, MySQL, Oracle, PostgreSQL, MongoDB...)

¿Datasets?

Probar con los propios!!

Kaggle https://www.kaggle.com/tags/healthcare

ODSC https://medium.com/@ODSC/15-open-datasets-for-healthcare-830b19980d9

Data.World https://data.world/datasets/healthcare

Ministerio de Ciencia, Chile https://github.com/MinCiencia/Datos-COVID19

DEIS https://deis.minsal.cl/

MINEDUC http://informacionestadistica.agenciaeducacion.cl/#/bases