06.- Data Adequacy_CU_04_19_vacunacion_completo_v_01

June 8, 2023

#

CU04_Optimización de vacunas

Citizenlab Data Science Methodology > II - Data Processing Domain *** > # 06.- Data Adequacy

Data Adequacy is the process to adapt basic and fundamental aspects of the raw data (File Format, Data Separator, Feature Names, Tidy Data, etcetera).

0.1 Tasks

File format
- Verify/Obtain Tabular CSV (row x column) if is appropriate - Change Data Separator Character in CSV files to ","
- Verify decimal point in numeric data is "."

Feature Names
- Verify names are Simple, Usable and Recognizable
- Rename Target column name="Target" and always will be the last column - Verify column names are the first row of csv file - Remove spaces others characters in the column names

Data Types - Verify data types - Change data types - Verify that object type is changed

Tidy Data
- Verify each variable forms a column - Verify each observation forms a row
- Verify each type of observational unit forms a table

0.2 Consideraciones casos CitizenLab programados en R

- La mayoría de las tareas de este proceso se han realizado en los notebooks del proceso 05 Data Collection porque eran necesarias para las tareas ETL. En esos casos, en este notebook se referencia al notebook del proceso 05 correspondiente
- Por tanto en los notebooks de este proceso de manera general se incluyen las comprobaciones necesarias, y comentarios si procede
- Las tareas del proceso se van a aplicar solo a los archivos que forman parte del despliegue, ya que hay muchos archivos intermedios que no procede pasar por este proceso
- El nombre de archivo del notebook hace referencia al nombre de archivo del proceso 05 al que se aplica este proceso, por eso pueden no ser correlativa la numeración
- Las comprobaciones se van a realizar teniendo en cuenta que el lenguaje utilizado en el despliegue de este caso es R

0.3 File

- Input File: CU_04_05_19_vacunacion_gripe_completo.csv
- Output File: CU_04_06_19_vacunacion_gripe_completo.csv

0.4 Settings

0.4.1 Encoding

Con la siguiente expresión se evitan problemas con el encoding al ejecutar el notebook. Es posible que deba ser eliminada o adaptada a la máquina en la que se ejecute el código.

[1]:
```
Sys.setlocale(category = "LC_ALL", locale = "es_ES.UTF-8")
```

```
Warning message in Sys.setlocale(category = "LC_ALL", locale = "es_ES.UTF-8"):
"OS reports request to set locale to "es_ES.UTF-8" cannot be honored"
```

”

0.4.2 Libraries to use

[2]:
```
library(readr)
library(dplyr)
library(tidyr)
library(stringr)
```

```
Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

0.4.3 Paths

[3]:
```
iPath <- "Data/Input/"
oPath <- "Data/Output/"
```

0.5 Data Load

OPCION A: Seleccionar fichero en ventana para mayor comodidad

Data load using the {tcltk} package. Ucomment the line if using this option

[4]:
```
# file_data <- tcltk::tk_choose.files(multi = FALSE)
```

OPCION B: Especificar el nombre de archivo

[5]:
```
iFile <- "CU_04_05_19_vacunacion_gripe_completo.csv"
file_data <- paste0(iPath, iFile)

if(file.exists(file_data)){
    cat("Se leerán datos del archivo: ", file_data)
} else{
    warning("Cuidado: el archivo no existe.")
}
```

```
Se leerán datos del archivo:
Data/Input/CU_04_05_19_vacunacion_gripe_completo.csv
```

0.6 Task

0.6.1 File format

Verify/Obtain Tabular CSV (row x column) if is appropriate

[6]:
```
data <- read.csv(file_data)

    # Verify if the data is appropriate and obtain its dimensions
    if (!is.null(dim(data))) {
      cat("\nEl archivo contiene una representación tabular.")

      # Print the dimensions (rows x columns)
      cat("\nNúmero de filas: ", dim(data)[1])
      cat("\nNúmero de columnas: ", dim(data)[2])

    } else {
      warning("Cuidado: el archivo no contiene una representación tabular.")
    }
```

```
El archivo contiene una representación tabular.
Número de filas:  21736
Número de columnas:  47
```

Remarks

- Se ha verificado que el CSV carga las filas y columnas sin errores

CODE

Data file to dataframe Usar la función adecuada según el formato de entrada (xlsx, csv, json, ...)

[7]:
```
data <- read_csv(file_data)
```

```
Rows: 21736 Columns: 47
  Column specification



Delimiter: ","
chr  (3): GEOCODIGO, DESBDT, nombre_zona
dbl (44): ano, semana, n_vacunas, n_citas, tmed, prec, velmedia,
presMax, be…

  Use `spec()` to retrieve the full column specification for this
data.
  Specify the column types or set `show_col_types = FALSE` to quiet
this message.
```

Estructura de los datos:

[8]:
```
data |> glimpse()
```

```
Rows: 21,736
Columns: 47
$ GEOCODIGO         <chr> "001", "001", "001", "001", "001",
"001", "001", "00…
$ DESBDT            <chr> "Abrantes", "Abrantes", "Abrantes",
"Abrantes", "Abr…
$ ano               <dbl> 2021, 2021, 2021, 2021, 2021, 2021,
2021, 2021, 2021…
$ semana            <dbl> 36, 37, 38, 39, 40, 41, 42, 43, 44,
45, 46, 47, 48, …
$ n_vacunas         <dbl> 0, 0, 0, 0, 0, 328, 344, 353, 371,
341, 389, 349, 40…
$ n_citas           <dbl> 0, 0, 0, 0, 0, 305, 327, 328, 341,
334, 373, 332, 37…
$ tmed              <dbl> 23.822231, 20.160014, 18.551058,
18.815387, 17.49447…
$ prec              <dbl> -0.0063023484, 3.5537258042,
3.9769052178, 1.4806472…
$ velmedia          <dbl> 3.573728, 2.494744, 3.316148,
2.384262, 1.850839, 1.…
$ presMax           <dbl> 940.9841, 940.7610, 943.1540,
944.6697, 944.2973, 94…
$ benzene           <dbl> 0.1713567, 0.1573829, 0.1858059,
0.1486437, 0.142803…
$ co                <dbl> 0.1680325, 0.2138607, 0.2034376,
0.2399882, 0.269345…
$ no                <dbl> 4.098371, 6.515572, 5.477654,
9.593391, 18.860535, 1…
$ no2               <dbl> 20.09480, 27.42594, 20.74836,
```

```
37.08524, 40.19475, 44…
$ nox               <dbl> 26.48135, 37.45944, 25.61128,
52.43745, 74.04903, 75…
$ o3                <dbl> 50.03434, 42.41281, 56.29918,
46.79483, 41.06600, 44…
$ pm10              <dbl> 17.447652, 17.658399, 12.844436,
16.395896, 14.90938…
$ pm2.5             <dbl> 3.008675, 10.083070, 7.218588,
9.426029, 8.131753, 1…
$ so2               <dbl> 6.861545, 6.589638, 4.364304,
3.123598, 1.291137, 1.…
$ campana           <dbl> 2021, 2021, 2021, 2021, 2021, 2021,
2021, 2021, 2021…
$ scampana          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,
12, 13, 14, 15, 1…
$ capacidad_zona    <dbl> 7477, 7477, 7477, 7477, 7477, 7477,
7477, 7477, 7477…
$ prop_riesgo       <dbl> 0.1953542, 0.1953542, 0.1953542,
0.1953542, 0.195354…
$ tasa_riesgo       <dbl> 0.006867731, 0.006867731,
0.006867731, 0.006867731, …
$ tasa_mayores      <dbl> 0.03324259, 0.03324259, 0.03324259,
0.03324259, 0.03…
$ poblacion_mayores <dbl> 0.1781619, 0.1781619, 0.1781619,
0.1781619, 0.178161…
$ nombre_zona       <chr> "Abrantes", "Abrantes", "Abrantes",
"Abrantes", "Abr…
$ nsec              <dbl> 23, 23, 23, 23, 23, 23, 23, 23, 23,
23, 23, 23, 23, …
$ t3_1              <dbl> 42.31249, 42.31249, 42.31249,
42.31249, 42.31249, 42…
$ t1_1              <dbl> 29872, 29872, 29872, 29872, 29872,
29872, 29872, 298…
$ t2_1              <dbl> 0.5345094, 0.5345094, 0.5345094,
0.5345094, 0.534509…
$ t2_2              <dbl> 0.4654906, 0.4654906, 0.4654906,
0.4654906, 0.465490…
$ t4_1              <dbl> 0.1561935, 0.1561935, 0.1561935,
0.1561935, 0.156193…
$ t4_2              <dbl> 0.6656459, 0.6656459, 0.6656459,
0.6656459, 0.665645…
$ t4_3              <dbl> 0.1781619, 0.1781619, 0.1781619,
0.1781619, 0.178161…
$ t5_1              <dbl> 0.2183993, 0.2183993, 0.2183993,
0.2183993, 0.218399…
$ t6_1              <dbl> 0.3450855, 0.3450855, 0.3450855,
0.3450855, 0.345085…
$ t7_1              <dbl> 0.04090796, 0.04090796, 0.04090796,
```

```
0.04090796, 0.04…
$ t8_1              <dbl> 0.02880326, 0.02880326, 0.02880326,
0.02880326, 0.02…
$ t9_1              <dbl> 0.2685716, 0.2685716, 0.2685716,
0.2685716, 0.268571…
$ t10_1             <dbl> 0.1665607, 0.1665607, 0.1665607,
0.1665607, 0.166560…
$ t11_1             <dbl> 0.4723181, 0.4723181, 0.4723181,
0.4723181, 0.472318…
$ t12_1             <dbl> 0.5637381, 0.5637381, 0.5637381,
0.5637381, 0.563738…
$ area              <dbl> 1571619, 1571619, 1571619, 1571619,
1571619, 1571619…
$ densidad_hab_km   <dbl> 19007.15, 19007.15, 19007.15,
19007.15, 19007.15, 19…
$ tuits_gripe       <dbl> 97, 79, 112, 143, 112, 130, 254,
190, 198, 160, 206,…
$ interes_gripe     <dbl> 13, 15, 19, 29, 38, 65, 100, 94,
63, 70, 64, 64, 57,…
```

Muestra de los primeros datos:

[9]:
```
data |> slice_head(n = 5)
```

A spec_tbl_df: 5 × 47

GEOCODIGO <chr>	DESBDT <chr>	ano <dbl>	semana <dbl>	n_vacunas <dbl>	n_citas <dbl>	tmed <dbl>	prec <dbl>
001	Abrantes	2021	36	0	0	23.82223	-0.006
001	Abrantes	2021	37	0	0	20.16001	3.5537
001	Abrantes	2021	38	0	0	18.55106	3.9769
001	Abrantes	2021	39	0	0	18.81539	1.4806
001	Abrantes	2021	40	0	0	17.49447	-0.033

Change Data Separator Character in CSV files to "," Remarks

- Se ha comprobado que el separador es "," ya que es la opción por defecto de `read_csv()`

CODE

[10]:
```
#
```

Verify decimal point in numeric data is "." Remarks

- Se ha comprobado que el símbolo decimal es "." ya que es la opción por defecto de `read_csv()`

CODE

[11]:
```
#
```

0.6.2 Feature Names

Verify names are Simple, Usable and Recognizable Remarks

- Se ha comprobado que los nombres de columnas son simples, usables y reconocibles

CODE

Visualizo el nombre de las columnas (características)

[12]:
```
colnames(data)
```

1. 'GEOCODIGO' 2. 'DESBDT' 3. 'ano' 4. 'semana' 5. 'n_vacunas' 6. 'n_citas' 7. 'tmed' 8. 'prec' 9. 'velmedia' 10. 'presMax' 11. 'benzene' 12. 'co' 13. 'no' 14. 'no2' 15. 'nox' 16. 'o3' 17. 'pm10' 18. 'pm2.5' 19. 'so2' 20. 'campana' 21. 'scampana' 22. 'capacidad_zona' 23. 'prop_riesgo' 24. 'tasa_riesgo' 25. 'tasa_mayores' 26. 'poblacion_mayores' 27. 'nombre_zona' 28. 'nsec' 29. 't3_1' 30. 't1_1' 31. 't2_1' 32. 't2_2' 33. 't4_1' 34. 't4_2' 35. 't4_3' 36. 't5_1' 37. 't6_1' 38. 't7_1' 39. 't8_1' 40. 't9_1' 41. 't10_1' 42. 't11_1' 43. 't12_1' 44. 'area' 45. 'densidad_hab_km' 46. 'tuits_gripe' 47. 'interes_gripe'

Cambio de nombre a columnas (si es necesario) para mejor uso

[13]:
```
# colnames(data) <- c("")
```

[14]:
```
colnames(data)
```

1. 'GEOCODIGO' 2. 'DESBDT' 3. 'ano' 4. 'semana' 5. 'n_vacunas' 6. 'n_citas' 7. 'tmed' 8. 'prec' 9. 'velmedia' 10. 'presMax' 11. 'benzene' 12. 'co' 13. 'no' 14. 'no2' 15. 'nox' 16. 'o3' 17. 'pm10' 18. 'pm2.5' 19. 'so2' 20. 'campana' 21. 'scampana' 22. 'capacidad_zona' 23. 'prop_riesgo' 24. 'tasa_riesgo' 25. 'tasa_mayores' 26. 'poblacion_mayores' 27. 'nombre_zona' 28. 'nsec' 29. 't3_1' 30. 't1_1' 31. 't2_1' 32. 't2_2' 33. 't4_1' 34. 't4_2' 35. 't4_3' 36. 't5_1' 37. 't6_1' 38. 't7_1' 39. 't8_1' 40. 't9_1' 41. 't10_1' 42. 't11_1' 43. 't12_1' 44. 'area' 45. 'densidad_hab_km' 46. 'tuits_gripe' 47. 'interes_gripe'

Rename Target column name="Target" and always will be the last column Remarks

- La columna a predecir o estimar sería **n_vacunas**
- No procede ya que no es necesario para aplicar los modelos previstos

CODE

[31]:
```
data$Target <- data$n_vacunas
```

Verify column names are the first row of csv file Remarks

- Se ha comprobado que la primera fila del archivo son los nombres de columna, ya que es la opción por defecto de `read_csv()`

CODE

[32]:
```
#
```

Remove spaces and others characters in the column names Remarks

- Se ha comprobado que los nombres de columna solo tienen caracteres ascii y no tienen espacios

CODE

Visualizo el nombre de las columnas (características)

[33]:
```
# Columns names
colnames(data)
```

1. 'GEOCODIGO' 2. 'DESBDT' 3. 'ano' 4. 'semana' 5. 'n_vacunas' 6. 'n_citas' 7. 'tmed' 8. 'prec' 9. 'velmedia' 10. 'presMax' 11. 'benzene' 12. 'co' 13. 'no' 14. 'no2' 15. 'nox' 16. 'o3' 17. 'pm10' 18. 'pm2.5' 19. 'so2' 20. 'campana' 21. 'scampana' 22. 'capacidad_zona' 23. 'prop_riesgo' 24. 'tasa_riesgo' 25. 'tasa_mayores' 26. 'poblacion_mayores' 27. 'nombre_zona' 28. 'nsec' 29. 't3_1' 30. 't1_1' 31. 't2_1' 32. 't2_2' 33. 't4_1' 34. 't4_2' 35. 't4_3' 36. 't5_1' 37. 't6_1' 38. 't7_1' 39. 't8_1' 40. 't9_1' 41. 't10_1' 42. 't11_1' 43. 't12_1' 44. 'area' 45. 'densidad_hab_km' 46. 'tuits_gripe' 47. 'interes_gripe' 48. 'Target'

Cambio espacio por '_' en nombres (si procede)

[34]:
```
colnames(data) <- str_replace_all(colnames(data), " ", "_")
```

[35]:
```
colnames(data)
```

1. 'GEOCODIGO' 2. 'DESBDT' 3. 'ano' 4. 'semana' 5. 'n_vacunas' 6. 'n_citas' 7. 'tmed' 8. 'prec' 9. 'velmedia' 10. 'presMax' 11. 'benzene' 12. 'co' 13. 'no' 14. 'no2' 15. 'nox' 16. 'o3' 17. 'pm10' 18. 'pm2.5' 19. 'so2' 20. 'campana' 21. 'scampana' 22. 'capacidad_zona' 23. 'prop_riesgo' 24. 'tasa_riesgo' 25. 'tasa_mayores' 26. 'poblacion_mayores' 27. 'nombre_zona' 28. 'nsec' 29. 't3_1' 30. 't1_1' 31. 't2_1' 32. 't2_2' 33. 't4_1' 34. 't4_2' 35. 't4_3' 36. 't5_1' 37. 't6_1' 38. 't7_1' 39. 't8_1' 40. 't9_1' 41. 't10_1' 42. 't11_1' 43. 't12_1' 44. 'area' 45. 'densidad_hab_km' 46. 'tuits_gripe' 47. 'interes_gripe' 48. 'Target'

0.6.3 Data Types

Verify data types Remarks

- Se ha comprobado el tipo de datos adecuados al importar los datos con `read_csv()`

CODE

Visualizo el tipo de las columnas (características)

[36]:
```
sapply(data, class)
glimpse(data)
```

GEOCODIGO 'character' **DESBDT** 'character' **ano** 'numeric' **semana** 'numeric' **n_vacunas** 'numeric' **n_citas** 'numeric' **tmed** 'numeric' **prec** 'numeric' **velmedia** 'numeric' **presMax** 'numeric' **benzene** 'numeric' **co** 'numeric' **no** 'numeric' **no2** 'numeric' **nox** 'numeric' **o3** 'numeric' **pm10** 'numeric' **pm2.5** 'numeric' **so2** 'numeric' **campana** 'numeric' **scampana** 'numeric' **capacidad_zona** 'numeric' **prop_riesgo** 'numeric' **tasa_riesgo** 'numeric' **tasa_mayores** 'numeric' **poblacion_mayores** 'numeric' **nombre_zona** 'character' **nsec** 'numeric' **t3_1** 'numeric' **t1_1** 'numeric' **t2_1** 'numeric' **t2_2** 'numeric' **t4_1** 'numeric' **t4_2** 'numeric' **t4_3** 'numeric' **t5_1** 'numeric' **t6_1** 'numeric' **t7_1** 'numeric' **t8_1** 'numeric' **t9_1** 'numeric' **t10_1** 'numeric' **t11_1** 'numeric' **t12_1** 'numeric' **area**

'numeric' **densidad_hab_km** 'numeric' **tuits_gripe** 'numeric' **interes_gripe** 'numeric' **Target** 'numeric'

```
Rows: 21,736
Columns: 48
$ GEOCODIGO         <chr> "001", "001", "001", "001", "001",
"001", "001", "00…
$ DESBDT            <chr> "Abrantes", "Abrantes", "Abrantes",
"Abrantes", "Abr…
$ ano               <dbl> 2021, 2021, 2021, 2021, 2021, 2021,
2021, 2021, 2021…
$ semana            <dbl> 36, 37, 38, 39, 40, 41, 42, 43, 44,
45, 46, 47, 48, …
$ n_vacunas         <dbl> 0, 0, 0, 0, 0, 328, 344, 353, 371,
341, 389, 349, 40…
$ n_citas           <dbl> 0, 0, 0, 0, 0, 305, 327, 328, 341,
334, 373, 332, 37…
$ tmed              <dbl> 23.822231, 20.160014, 18.551058,
18.815387, 17.49447…
$ prec              <dbl> -0.0063023484, 3.5537258042,
3.9769052178, 1.4806472…
$ velmedia          <dbl> 3.573728, 2.494744, 3.316148,
2.384262, 1.850839, 1.…
$ presMax           <dbl> 940.9841, 940.7610, 943.1540,
944.6697, 944.2973, 94…
$ benzene           <dbl> 0.1713567, 0.1573829, 0.1858059,
0.1486437, 0.142803…
$ co                <dbl> 0.1680325, 0.2138607, 0.2034376,
0.2399882, 0.269345…
$ no                <dbl> 4.098371, 6.515572, 5.477654,
9.593391, 18.860535, 1…
$ no2               <dbl> 20.09480, 27.42594, 20.74836,
37.08524, 40.19475, 44…
$ nox               <dbl> 26.48135, 37.45944, 25.61128,
52.43745, 74.04903, 75…
$ o3                <dbl> 50.03434, 42.41281, 56.29918,
46.79483, 41.06600, 44…
$ pm10              <dbl> 17.447652, 17.658399, 12.844436,
16.395896, 14.90938…
$ pm2.5             <dbl> 3.008675, 10.083070, 7.218588,
9.426029, 8.131753, 1…
$ so2               <dbl> 6.861545, 6.589638, 4.364304,
3.123598, 1.291137, 1.…
$ campana           <dbl> 2021, 2021, 2021, 2021, 2021, 2021,
2021, 2021, 2021…
$ scampana          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,
12, 13, 14, 15, 1…
$ capacidad_zona    <dbl> 7477, 7477, 7477, 7477, 7477, 7477,
```

```
7477, 7477, 7477…
$ prop_riesgo       <dbl> 0.1953542, 0.1953542, 0.1953542,
0.1953542, 0.195354…
$ tasa_riesgo       <dbl> 0.006867731, 0.006867731,
0.006867731, 0.006867731, …
$ tasa_mayores      <dbl> 0.03324259, 0.03324259, 0.03324259,
0.03324259, 0.03…
$ poblacion_mayores <dbl> 0.1781619, 0.1781619, 0.1781619,
0.1781619, 0.178161…
$ nombre_zona       <chr> "Abrantes", "Abrantes", "Abrantes",
"Abrantes", "Abr…
$ nsec              <dbl> 23, 23, 23, 23, 23, 23, 23, 23, 23,
23, 23, 23, 23, …
$ t3_1              <dbl> 42.31249, 42.31249, 42.31249,
42.31249, 42.31249, 42…
$ t1_1              <dbl> 29872, 29872, 29872, 29872, 29872,
29872, 29872, 298…
$ t2_1              <dbl> 0.5345094, 0.5345094, 0.5345094,
0.5345094, 0.534509…
$ t2_2              <dbl> 0.4654906, 0.4654906, 0.4654906,
0.4654906, 0.465490…
$ t4_1              <dbl> 0.1561935, 0.1561935, 0.1561935,
0.1561935, 0.156193…
$ t4_2              <dbl> 0.6656459, 0.6656459, 0.6656459,
0.6656459, 0.665645…
$ t4_3              <dbl> 0.1781619, 0.1781619, 0.1781619,
0.1781619, 0.178161…
$ t5_1              <dbl> 0.2183993, 0.2183993, 0.2183993,
0.2183993, 0.218399…
$ t6_1              <dbl> 0.3450855, 0.3450855, 0.3450855,
0.3450855, 0.345085…
$ t7_1              <dbl> 0.04090796, 0.04090796, 0.04090796,
0.04090796, 0.04…
$ t8_1              <dbl> 0.02880326, 0.02880326, 0.02880326,
0.02880326, 0.02…
$ t9_1              <dbl> 0.2685716, 0.2685716, 0.2685716,
0.2685716, 0.268571…
$ t10_1             <dbl> 0.1665607, 0.1665607, 0.1665607,
0.1665607, 0.166560…
$ t11_1             <dbl> 0.4723181, 0.4723181, 0.4723181,
0.4723181, 0.472318…
$ t12_1             <dbl> 0.5637381, 0.5637381, 0.5637381,
0.5637381, 0.563738…
$ area              <dbl> 1571619, 1571619, 1571619, 1571619,
1571619, 1571619…
$ densidad_hab_km   <dbl> 19007.15, 19007.15, 19007.15,
19007.15, 19007.15, 19…
$ tuits_gripe       <dbl> 97, 79, 112, 143, 112, 130, 254,
```

```
190, 198, 160, 206,…
$ interes_gripe     <dbl> 13, 15, 19, 29, 38, 65, 100, 94,
63, 70, 64, 64, 57,…
$ Target            <dbl> 0, 0, 0, 0, 0, 328, 344, 353, 371,
341, 389, 349, 40…
```

Change data types Remarks

- No aplica

CODE

[37]:
```
# Changing column types
# data$xx <- as.xx(data$xx)
```

Verify that object type is changed Remarks

- No aplica

CODE

[38]:
```
sapply(data, class)
glimpse(data)
```

GEOCODIGO 'character' **DESBDT** 'character' **ano** 'numeric' **semana** 'numeric' **n__vacunas** 'numeric' **n__citas** 'numeric' **tmed** 'numeric' **prec** 'numeric' **velmedia** 'numeric' **presMax** 'numeric' **benzene** 'numeric' **co** 'numeric' **no** 'numeric' **no2** 'numeric' **nox** 'numeric' **o3** 'numeric' **pm10** 'numeric' **pm2.5** 'numeric' **so2** 'numeric' **campana** 'numeric' **scampana** 'numeric' **capacidad__zona** 'numeric' **prop__riesgo** 'numeric' **tasa__riesgo** 'numeric' **tasa__mayores** 'numeric' **poblacion__mayores** 'numeric' **nombre__zona** 'character' **nsec** 'numeric' **t3__1** 'numeric' **t1__1** 'numeric' **t2__1** 'numeric' **t2__2** 'numeric' **t4__1** 'numeric' **t4__2** 'numeric' **t4__3** 'numeric' **t5__1** 'numeric' **t6__1** 'numeric' **t7__1** 'numeric' **t8__1** 'numeric' **t9__1** 'numeric' **t10__1** 'numeric' **t11__1** 'numeric' **t12__1** 'numeric' **area** 'numeric' **densidad__hab__km** 'numeric' **tuits__gripe** 'numeric' **interes__gripe** 'numeric' **Target** 'numeric'

```
Rows: 21,736
Columns: 48
$ GEOCODIGO         <chr> "001", "001", "001", "001", "001",
"001", "001", "00…
$ DESBDT            <chr> "Abrantes", "Abrantes", "Abrantes",
"Abrantes", "Abr…
$ ano               <dbl> 2021, 2021, 2021, 2021, 2021, 2021,
2021, 2021, 2021…
$ semana            <dbl> 36, 37, 38, 39, 40, 41, 42, 43, 44,
45, 46, 47, 48, …
$ n_vacunas         <dbl> 0, 0, 0, 0, 0, 328, 344, 353, 371,
341, 389, 349, 40…
$ n_citas           <dbl> 0, 0, 0, 0, 0, 305, 327, 328, 341,
334, 373, 332, 37…
```

```
$ tmed              <dbl> 23.822231, 20.160014, 18.551058,
18.815387, 17.49447…
$ prec              <dbl> -0.0063023484, 3.5537258042,
3.9769052178, 1.4806472…
$ velmedia          <dbl> 3.573728, 2.494744, 3.316148,
2.384262, 1.850839, 1.…
$ presMax           <dbl> 940.9841, 940.7610, 943.1540,
944.6697, 944.2973, 94…
$ benzene           <dbl> 0.1713567, 0.1573829, 0.1858059,
0.1486437, 0.142803…
$ co                <dbl> 0.1680325, 0.2138607, 0.2034376,
0.2399882, 0.269345…
$ no                <dbl> 4.098371, 6.515572, 5.477654,
9.593391, 18.860535, 1…
$ no2               <dbl> 20.09480, 27.42594, 20.74836,
37.08524, 40.19475, 44…
$ nox               <dbl> 26.48135, 37.45944, 25.61128,
52.43745, 74.04903, 75…
$ o3                <dbl> 50.03434, 42.41281, 56.29918,
46.79483, 41.06600, 44…
$ pm10              <dbl> 17.447652, 17.658399, 12.844436,
16.395896, 14.90938…
$ pm2.5             <dbl> 3.008675, 10.083070, 7.218588,
9.426029, 8.131753, 1…
$ so2               <dbl> 6.861545, 6.589638, 4.364304,
3.123598, 1.291137, 1.…
$ campana           <dbl> 2021, 2021, 2021, 2021, 2021, 2021,
2021, 2021, 2021…
$ scampana          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,
12, 13, 14, 15, 1…
$ capacidad_zona    <dbl> 7477, 7477, 7477, 7477, 7477, 7477,
7477, 7477, 7477…
$ prop_riesgo       <dbl> 0.1953542, 0.1953542, 0.1953542,
0.1953542, 0.195354…
$ tasa_riesgo       <dbl> 0.006867731, 0.006867731,
0.006867731, 0.006867731, …
$ tasa_mayores      <dbl> 0.03324259, 0.03324259, 0.03324259,
0.03324259, 0.03…
$ poblacion_mayores <dbl> 0.1781619, 0.1781619, 0.1781619,
0.1781619, 0.178161…
$ nombre_zona       <chr> "Abrantes", "Abrantes", "Abrantes",
"Abrantes", "Abr…
$ nsec              <dbl> 23, 23, 23, 23, 23, 23, 23, 23, 23,
23, 23, 23, 23, …
$ t3_1              <dbl> 42.31249, 42.31249, 42.31249,
42.31249, 42.31249, 42…
$ t1_1              <dbl> 29872, 29872, 29872, 29872, 29872,
29872, 29872, 298…
```

```
$ t2_1              <dbl> 0.5345094, 0.5345094, 0.5345094,
0.5345094, 0.534509…
$ t2_2              <dbl> 0.4654906, 0.4654906, 0.4654906,
0.4654906, 0.465490…
$ t4_1              <dbl> 0.1561935, 0.1561935, 0.1561935,
0.1561935, 0.156193…
$ t4_2              <dbl> 0.6656459, 0.6656459, 0.6656459,
0.6656459, 0.665645…
$ t4_3              <dbl> 0.1781619, 0.1781619, 0.1781619,
0.1781619, 0.178161…
$ t5_1              <dbl> 0.2183993, 0.2183993, 0.2183993,
0.2183993, 0.218399…
$ t6_1              <dbl> 0.3450855, 0.3450855, 0.3450855,
0.3450855, 0.345085…
$ t7_1              <dbl> 0.04090796, 0.04090796, 0.04090796,
0.04090796, 0.04…
$ t8_1              <dbl> 0.02880326, 0.02880326, 0.02880326,
0.02880326, 0.02…
$ t9_1              <dbl> 0.2685716, 0.2685716, 0.2685716,
0.2685716, 0.268571…
$ t10_1             <dbl> 0.1665607, 0.1665607, 0.1665607,
0.1665607, 0.166560…
$ t11_1             <dbl> 0.4723181, 0.4723181, 0.4723181,
0.4723181, 0.472318…
$ t12_1             <dbl> 0.5637381, 0.5637381, 0.5637381,
0.5637381, 0.563738…
$ area              <dbl> 1571619, 1571619, 1571619, 1571619,
1571619, 1571619…
$ densidad_hab_km   <dbl> 19007.15, 19007.15, 19007.15,
19007.15, 19007.15, 19…
$ tuits_gripe       <dbl> 97, 79, 112, 143, 112, 130, 254,
190, 198, 160, 206,…
$ interes_gripe     <dbl> 13, 15, 19, 29, 38, 65, 100, 94,
63, 70, 64, 64, 57,…
$ Target            <dbl> 0, 0, 0, 0, 0, 328, 344, 353, 371,
341, 389, 349, 40…
```

0.6.4 Tidy Data

Verify each variable forms a column Remarks

- Se ha comprobado que cada columna se refiere a una variable

CODE

[39]:
```
#
ncol(data)
glimpse(data)
```

48

```
Rows: 21,736
Columns: 48
$ GEOCODIGO         <chr> "001", "001", "001", "001", "001",
"001", "001", "00…
$ DESBDT            <chr> "Abrantes", "Abrantes", "Abrantes",
"Abrantes", "Abr…
$ ano               <dbl> 2021, 2021, 2021, 2021, 2021, 2021,
2021, 2021, 2021…
$ semana            <dbl> 36, 37, 38, 39, 40, 41, 42, 43, 44,
45, 46, 47, 48, …
$ n_vacunas         <dbl> 0, 0, 0, 0, 0, 328, 344, 353, 371,
341, 389, 349, 40…
$ n_citas           <dbl> 0, 0, 0, 0, 0, 305, 327, 328, 341,
334, 373, 332, 37…
$ tmed              <dbl> 23.822231, 20.160014, 18.551058,
18.815387, 17.49447…
$ prec              <dbl> -0.0063023484, 3.5537258042,
3.9769052178, 1.4806472…
$ velmedia          <dbl> 3.573728, 2.494744, 3.316148,
2.384262, 1.850839, 1.…
$ presMax           <dbl> 940.9841, 940.7610, 943.1540,
944.6697, 944.2973, 94…
$ benzene           <dbl> 0.1713567, 0.1573829, 0.1858059,
0.1486437, 0.142803…
$ co                <dbl> 0.1680325, 0.2138607, 0.2034376,
0.2399882, 0.269345…
$ no                <dbl> 4.098371, 6.515572, 5.477654,
9.593391, 18.860535, 1…
$ no2               <dbl> 20.09480, 27.42594, 20.74836,
37.08524, 40.19475, 44…
$ nox               <dbl> 26.48135, 37.45944, 25.61128,
52.43745, 74.04903, 75…
$ o3                <dbl> 50.03434, 42.41281, 56.29918,
46.79483, 41.06600, 44…
$ pm10              <dbl> 17.447652, 17.658399, 12.844436,
16.395896, 14.90938…
$ pm2.5             <dbl> 3.008675, 10.083070, 7.218588,
9.426029, 8.131753, 1…
$ so2               <dbl> 6.861545, 6.589638, 4.364304,
3.123598, 1.291137, 1.…
$ campana           <dbl> 2021, 2021, 2021, 2021, 2021, 2021,
2021, 2021, 2021…
$ scampana          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,
12, 13, 14, 15, 1…
$ capacidad_zona    <dbl> 7477, 7477, 7477, 7477, 7477, 7477,
7477, 7477, 7477…
```

```
$ prop_riesgo       <dbl> 0.1953542, 0.1953542, 0.1953542,
0.1953542, 0.195354…
$ tasa_riesgo       <dbl> 0.006867731, 0.006867731,
0.006867731, 0.006867731, …
$ tasa_mayores      <dbl> 0.03324259, 0.03324259, 0.03324259,
0.03324259, 0.03…
$ poblacion_mayores <dbl> 0.1781619, 0.1781619, 0.1781619,
0.1781619, 0.178161…
$ nombre_zona       <chr> "Abrantes", "Abrantes", "Abrantes",
"Abrantes", "Abr…
$ nsec              <dbl> 23, 23, 23, 23, 23, 23, 23, 23, 23,
23, 23, 23, 23, …
$ t3_1              <dbl> 42.31249, 42.31249, 42.31249,
42.31249, 42.31249, 42…
$ t1_1              <dbl> 29872, 29872, 29872, 29872, 29872,
29872, 29872, 298…
$ t2_1              <dbl> 0.5345094, 0.5345094, 0.5345094,
0.5345094, 0.534509…
$ t2_2              <dbl> 0.4654906, 0.4654906, 0.4654906,
0.4654906, 0.465490…
$ t4_1              <dbl> 0.1561935, 0.1561935, 0.1561935,
0.1561935, 0.156193…
$ t4_2              <dbl> 0.6656459, 0.6656459, 0.6656459,
0.6656459, 0.665645…
$ t4_3              <dbl> 0.1781619, 0.1781619, 0.1781619,
0.1781619, 0.178161…
$ t5_1              <dbl> 0.2183993, 0.2183993, 0.2183993,
0.2183993, 0.218399…
$ t6_1              <dbl> 0.3450855, 0.3450855, 0.3450855,
0.3450855, 0.345085…
$ t7_1              <dbl> 0.04090796, 0.04090796, 0.04090796,
0.04090796, 0.04…
$ t8_1              <dbl> 0.02880326, 0.02880326, 0.02880326,
0.02880326, 0.02…
$ t9_1              <dbl> 0.2685716, 0.2685716, 0.2685716,
0.2685716, 0.268571…
$ t10_1             <dbl> 0.1665607, 0.1665607, 0.1665607,
0.1665607, 0.166560…
$ t11_1             <dbl> 0.4723181, 0.4723181, 0.4723181,
0.4723181, 0.472318…
$ t12_1             <dbl> 0.5637381, 0.5637381, 0.5637381,
0.5637381, 0.563738…
$ area              <dbl> 1571619, 1571619, 1571619, 1571619,
1571619, 1571619…
$ densidad_hab_km   <dbl> 19007.15, 19007.15, 19007.15,
19007.15, 19007.15, 19…
$ tuits_gripe       <dbl> 97, 79, 112, 143, 112, 130, 254,
190, 198, 160, 206,…
```

```
$ interes_gripe     <dbl> 13, 15, 19, 29, 38, 65, 100, 94,
63, 70, 64, 64, 57,…
$ Target            <dbl> 0, 0, 0, 0, 0, 328, 344, 353, 371,
341, 389, 349, 40…
```

Verify each observation forms a row Remarks

- Se ha comprobado que cada fila se refiere a una observación

CODE

[40]:
```
#
nrow(data)
head(data, 1)
```

21736

A tibble: 1 × 48

GEOCODIGO <chr>	DESBDT <chr>	ano <dbl>	semana <dbl>	n_vacunas <dbl>	n_citas <dbl>	tmed <dbl>	prec <dbl>
001	Abrantes	2021	36	0	0	23.82223	-0.006302348

Verify each type of observational unit forms a table Remarks

- No aplica

CODE

[41]:
```
#
```

0.7 Data Save

- Solo si se han hecho cambios
- No aplica

Identificamos los datos a guardar

[42]:
```
data_to_save <- data
```

Estructura de nombre de archivos:

- Código del caso de uso, por ejemplo “CU_04”
- Número del proceso que lo genera, por ejemplo ”_06”.
- Resto del nombre del archivo de entrada
- Extensión del archivo

Ejemplo: ”CU_04_06_01_01_zonasgeo.json, primer fichero que se genera en la tarea 01 del proceso 05 (Data Collection) para el caso de uso 04 (vacunas) y que se ha transformado en el proceso 06

Importante mantener los guiones bajos antes de proceso, tarea, archivo y nombre

0.7.1 Proceso 06

[43]:
```
caso <- "CU_04"
proceso <- '_06'
tarea <- "_19"
archivo <- ""
proper <- "_vacunacion_gripe_completo"
extension <- ".csv"
```

OPCION A: Uso del paquete "tcltk" para mayor comodidad

- Buscar carpeta, escribir nombre de archivo SIN extensión (se especifica en el código)
- Especificar sufijo2 si es necesario
- Cambiar datos por datos_xx si es necesario

[44]:
```
# file_save <- paste0(caso, proceso, tarea, tcltk::tkgetSaveFile(), proper,␣
 ↪extension)
# path_out <- paste0(oPath, file_save)
# write_csv(data_to_save_xxxx, path_out)

# cat('File saved as: ')
# path_out
```

OPCION B: Especificar el nombre de archivo

- Los ficheros de salida del proceso van siempre a Data/Output/.

[45]:
```
file_save <- paste0(caso, proceso, tarea, archivo, proper, extension)
path_out <- paste0(oPath, file_save)
write_csv(data_to_save, path_out)

cat('File saved as: ')
path_out
```

```
File saved as:
```

'Data/Output/CU_04_06_19_vacunacion_gripe_completo.csv'

Copia del fichero a Input Si el archivo se va a usar en otros notebooks, copiar a la carpeta Input

[46]:
```
path_in <- paste0(iPath, file_save)
file.copy(path_out, path_in, overwrite = TRUE)
```

TRUE

0.8 REPORT

A continuación se realizará un informe de las acciones realizadas

0.9 Main Actions Carried Out

Ejemplos - Se han comprobado todas las tareas de Data Adecuacy - No se ha tenido que realizar ninguna acción adicional

0.10 Main Conclusions

- Los datos ya se habían tratado en el proceso 05 para poder hacer las tareas de ETL y no ha sido necesaria ninguna modificación

[ ]:

[ ]: