12.- Exploratory Data Analysis_25_01_listas_espera_v_01

June 10, 2023

#

CU25_Modelo de gestión de Lista de Espera Quirúrgica

Citizenlab Data Science Methodology > II - Data Processing Domain *** > # 12.- EDA - Exploratory Data Analysis Analysis

0.1 Tasks

0.2 File

- Input File: CU_25_09.2_01_lista_espera_completo_clean_v_01.csv
- Output File: No aplica

0.2.1 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.

[57]:
```
Sys.setlocale(category = "LC_ALL", locale = "es_ES.UTF-8")
```

'LC_COLLATE=es_ES.UTF-8;LC_CTYPE=es_ES.UTF-8;LC_MONETARY=es_ES.UTF-8;LC_NUMERIC=C;LC_TIME=es_ES.UTF-8'

0.3 Settings

0.3.1 Libraries to use

[58]:
```
library(readr)
library(dplyr)
library(sf)
library(tidyr)
library(ggplot2)
library(summarytools)
library(GGally)
library(nortest)
library(lubridate)
```

0.3.2 Paths

[59]:
```
iPath <- "Data/Input/"
oPath <- "Data/Output/"
```

0.4 Data Load

OPCION A: Seleccionar fichero en ventana para mayor comodidad

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

[60]:
```
# file_data <- tcltk::tk_choose.files(multi = FALSE)
```

OPCION B: Especificar el nombre de archivo

[61]:
```
iFile <- "CU_25_09.2_01_lista_espera_completo_clean_v_01.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_25_09.2_01_lista_espera_completo_clean_v_01.csv
```

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

[62]:
```
data <- read_csv(file_data)
```

```
Rows: 55216 Columns: 46
  Column specification

Delimiter: ","
chr  (6): Hospital, Especialidad, nombre_area, Municipio, Clase,
Dependencia
dbl (39): total_pacientes, ano, semana, CODCNH, id_area, cmunicipio,
CAMAS, …
lgl  (1): is_train

  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.
```

0.5 Data Structure

Estructura de los datos:

[63]:
```
data |> glimpse()
```

```
Rows: 55,216
Columns: 46
$ Hospital          <chr> "HOSPITAL REY JUAN CARLOS",
"HOSPITAL CENTRAL DE LA …
$ Especialidad      <chr> "UROLOGÍA", "ODONTOESTOMATOLOGÍA",
"GINECOLOGÍA", "D…
$ total_pacientes   <dbl> 344, 0, 52, 37, 0, 4, 0, 718, 0,
271, 108, 0, 34, 86…
$ ano               <dbl> 2021, 2020, 2021, 2021, 2021, 2020,
2021, 2020, 2021…
$ semana            <dbl> 30, 36, 49, 23, 3, 5, 50, 7, 35, 1,
42, 10, 21, 33, …
$ CODCNH            <dbl> 281348, 280724, 281292, 281292,
281236, 280724, 2807…
$ id_area           <dbl> 8, 7, 11, 11, 11, 7, 3, 6, 1, 2, 2,
8, 11, 11, 1, 3,…
$ nombre_area       <chr> "SUR-OESTE I", "CENTRO-OESTE", "SUR
II", "SUR II", "…
$ cmunicipio        <dbl> 280920, 280796, 280133, 280133,
281610, 280796, 2800…
$ Municipio         <chr> "MÓSTOLES", "MADRID", "ARANJUEZ",
"ARANJUEZ", "VALDE…
$ CAMAS             <dbl> 382, 475, 98, 98, 182, 475, 507,
613, 269, 1143, 156…
$ Clase             <chr> "HOSPITALES GENERALES", "HOSPITALES
GENERALES", "HOS…
$ Dependencia       <chr> "SERVICIOS E INSTITUTOS DE SALUD DE
LAS COMUNIDADES …
$ TAC               <dbl> 2, 2, 1, 1, 1, 2, 3, 3, 0, 0, 1, 2,
6, 6, 1, 3, 4, 1…
$ RM                <dbl> 3, 2, 1, 1, 2, 2, 2, 3, 0, 0, 0, 2,
5, 5, 1, 2, 4, 1…
$ GAM               <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
2, 2, 0, 0, 2, 0…
$ HEM               <dbl> 1, 2, 0, 0, 1, 2, 1, 2, 0, 0, 0, 1,
3, 3, 0, 1, 1, 0…
$ ASD               <dbl> 2, 1, 1, 1, 1, 1, 1, 3, 0, 0, 0, 1,
2, 2, 0, 1, 2, 1…
$ ALI               <dbl> 1, 2, 0, 0, 0, 2, 0, 4, 0, 0, 0, 0,
3, 3, 0, 2, 2, 0…
$ SPECT             <dbl> 1, 1, 0, 0, 0, 1, 0, 4, 0, 0, 0, 0,
3, 3, 0, 0, 0, 0…
$ MAMOS             <dbl> 2, 1, 1, 1, 1, 1, 2, 2, 0, 0, 1, 2,
3, 3, 1, 1, 3, 1…
$ DO                <dbl> 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1,
2, 2, 0, 1, 2, 0…
```

```
$ DIAL              <dbl> 20, 24, 13, 13, 17, 24, 28, 31, 0,
0, 0, 28, 43, 43,…
$ X                 <dbl> -3.870412, -3.745529, -3.610795,
-3.610795, -3.69744…
$ Y                 <dbl> 40.33920, 40.38791, 40.05726,
40.05726, 40.19884, 40…
$ t3_1              <dbl> 42.34715, 45.37878, 42.06149,
42.06149, 42.06149, 45…
$ t1_1              <dbl> 532487, 511605, 899702, 899702,
899702, 511605, 3830…
$ t2_1              <dbl> 0.5122493, 0.5296804, 0.5240445,
0.5240445, 0.524044…
$ t2_2              <dbl> 0.4877507, 0.4703198, 0.4759555,
0.4759555, 0.475955…
$ t4_1              <dbl> 0.1659665, 0.1054260, 0.1540793,
0.1540793, 0.154079…
$ t4_2              <dbl> 0.6371549, 0.6742432, 0.6753787,
0.6753787, 0.675378…
$ t4_3              <dbl> 0.1968769, 0.2203341, 0.1705449,
0.1705449, 0.170544…
$ t5_1              <dbl> 0.1137647, 0.1744493, 0.1747059,
0.1747059, 0.174705…
$ t6_1              <dbl> 0.1604646, 0.2629599, 0.2641879,
0.2641879, 0.264187…
$ t7_1              <dbl> 0.05422176, 0.05481008, 0.04898547,
0.04898547, 0.04…
$ t8_1              <dbl> 0.04120012, 0.04653221, 0.03679912,
0.03679912, 0.03…
$ t9_1              <dbl> 0.3348780, 0.4914365, 0.3346063,
0.3346063, 0.334606…
$ t10_1             <dbl> 0.13692541, 0.12170996, 0.15173209,
0.15173209, 0.15…
$ t11_1             <dbl> 0.5072726, 0.4915713, 0.5024130,
0.5024130, 0.502413…
$ t12_1             <dbl> 0.5849309, 0.5597213, 0.5900028,
0.5900028, 0.590002…
$ capacidad         <dbl> 17, 0, 8, 5, 0, 5, 1, 24, 6, 6, 30,
4, 2, 15, 20, 6,…
$ pacientes         <dbl> 1447, 1211, 1293, 1501, 1240, 1504,
1502, 1533, 1463…
$ consultas         <dbl> 573, 45, 108, 103, 44, 42, 36,
1119, 34, 466, 220, 6…
$ hospitalizaciones <dbl> 12, 0, 2, 2, 0, 1, 0, 4, 0, 12, 3,
0, 2, 4, 1, 2, 15…
$ Target            <dbl> 54.45, 0.00, 37.96, 23.14, 0.00,
6.25, 0.00, 78.20, …
$ is_train          <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE,
TRUE, TRUE, TRUE…
```

Muestra de los primeros datos:

[64]:
```
data |> slice_head(n = 5)
```

A spec_tbl_df: 5 × 46

Hospital <chr>	Especialidad <chr>
HOSPITAL REY JUAN CARLOS	UROLOGÍA
HOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA	ODONTOESTOMATOLO
HOSPITAL UNIVERSITARIO DEL TAJO	GINECOLOGÍA
HOSPITAL UNIVERSITARIO DEL TAJO	DERMATOLOGÍA
HOSPITAL UNIVERSITARIO INFANTA ELENA	ODONTOESTOMATOLO

Tamaño de Memoria de los datos

[65]:
```
object.size(data)
```

```
20133120 bytes
```

Structure of non-numerical features

[66]:
```
# Display non-numerical features
data |> select(where(~ !is.numeric(.x))) |> freq()
```

HOSPITAL CENTRAL DE LA CRUZ ROJA SAN JOSE Y SANTA A
HOSPITAL CENTRAL DE LA DEFENSA GOMEZ
HOSPITAL CLINICO SAN CA
HOSPITAL EL ESC
HOSPITAL GENERAL DE VILL
HOSPITAL GENERAL UNIVERSITARIO GREGORIO MAR
HOSPITAL INFANTIL UNIVERSITARIO NIÑO
HOSPITAL RAMON Y C
HOSPITAL REY JUAN CA
HOSPITAL UNIVERSITARIO 12 DE OCT
HOSPITAL UNIVERSITARIO DE FUENLAB
HOSPITAL UNIVERSITARIO DE GE
HOSPITAL UNIVERSITARIO DE LA PRIN
HOSPITAL UNIVERSITARIO DE MOST
HOSPITAL UNIVERSITARIO DE TORR
HOSPITAL UNIVERSITARIO DEL HEN
HOSPITAL UNIVERSITARIO DEL SUR
HOSPITAL UNIVERSITARIO DEL
HOSPITAL UNIVERSITARIO FUNDACION ALCO
HOSPITAL UNIVERSITARIO FUNDACION JIMENEZ
HOSPITAL UNIVERSITARIO INFANTA CRI
HOSPITAL UNIVERSITARIO INFANTA E
HOSPITAL UNIVERSITARIO INFANTA LE
HOSPITAL UNIVERSITARIO INFANTA
HOSPITAL UNIVERSITARIO L
HOSPITAL UNIVERSITARIO PRINCIPE DE AST
HOSPITAL UNIVERSITARIO PUERTA DE HIERRO MAJADAH
HOSPITAL UNIVERSITARIO SANTA CRI
HOSPITAL UNIVERSITARIO SEVERO C

1. A summarytools: 31×5 of type dbl

	Freq	% V
ANGIOLOGÍA Y CIRUGÍA VASCULAR	3451	6.25
CIRUGÍA CARDIACA	3451	6.25
CIRUGÍA GENERAL Y DEL APARATO DIGESTIVO	3451	6.25
CIRUGÍA ORAL Y MAXILOFACIAL	3451	6.25
CIRUGÍA PEDIÁTRICA GENERAL	3451	6.25
CIRUGÍA PLÁSTICA Y REPARADORA	3451	6.25
CIRUGÍA TORÁCICA	3451	6.25
DERMATOLOGÍA	3451	6.25
GINECOLOGÍA	3451	6.25
NEUROCIRUGÍA	3451	6.25
ODONTOESTOMATOLOGÍA	3451	6.25
OFTALMOLOGÍA	3451	6.25
OTORRINOLARINGOLOGÍA	3451	6.25
TOTAL	3451	6.25
TRAUMATOLOGÍA	3451	6.25
UROLOGÍA	3451	6.25
<NA>	0	NA
Total	55216	100

2. A summarytools: 18×5 of type dbl

	Freq	% Valid	% Valid Cum.	% Total
CENTRO-NORTE	7616	13.793103	13.79310	13.793103
CENTRO-OESTE	5712	10.344828	24.13793	10.344828
ESTE	3808	6.896552	31.03448	6.896552
NORTE	7616	13.793103	44.82759	13.793103
OESTE	5712	10.344828	55.17241	10.344828
SUR-ESTE	5712	10.344828	65.51724	10.344828
SUR-OESTE I	5712	10.344828	75.86207	10.344828
SUR-OESTE II	3808	6.896552	82.75862	6.896552
SUR I	3808	6.896552	89.65517	6.896552
SUR II	5712	10.344828	100.00000	10.344828
<NA>	0	NA	NA	0.000000
Total	55216	100.000000	100.00000	100.000000

3. A summarytools: 12×5 of type dbl

4. A summarytools: 18 × 5 of type dbl

	Freq	% Valid	% Valid C
ALCALÁ DE HENARES	1904	3.448276	3.448276
ALCORCÓN	1904	3.448276	6.896552
ARANJUEZ	1904	3.448276	10.344828
ARGANDA DEL REY	1904	3.448276	13.793103
COLLADO VILLALBA	1904	3.448276	17.241379
COSLADA	1904	3.448276	20.689655
FUENLABRADA	1904	3.448276	24.137931
GETAFE	1904	3.448276	27.586207
LEGANÉS	1904	3.448276	31.034483
MADRID	24752	44.827586	75.862069
MAJADAHONDA	1904	3.448276	79.310345
MÓSTOLES	3808	6.896552	86.206897
SAN LORENZO DE EL ESCORIAL	1904	3.448276	89.655172
SAN SEBASTIÁN DE LOS REYES	1904	3.448276	93.103448
TORREJÓN DE ARDOZ	1904	3.448276	96.551724
VALDEMORO	1904	3.448276	100.00000
<NA>	0	NA	NA
Total	55216	100.000000	100.00000

5. A summarytools: 5 × 5 of type dbl

	Freq	% Valid	%
HOSPITALES ESPECIALIZADOS	1904	3.448276	3
HOSPITALES GENERALES	51408	93.103448	9
OTROS CENTROS CON INTERNAMIENTO	1904	3.448276	1
<NA>	0	NA	N
Total	55216	100.000000	1

6. A summarytools: 5 × 5 of type dbl

	MINISTERIO I
SERVICIOS E INSTITUTOS DE SALUD DE LAS COMUNIDADES A	

7. A summarytools: 4 × 5 of type dbl

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum
FALSE	11046	20.00507	20.00507	20.00507	20.00507
TRUE	44170	79.99493	100.00000	79.99493	100.00000
<NA>	0	NA	NA	0.00000	100.00000
Total	55216	100.00000	100.00000	100.00000	100.00000

Structure of numerical features

[67]:
```
data |> select(where(is.numeric)) |> descr()
```

A summarytools: 15 × 39 of type dbl

	ALI	ano	ASD	CAMAS
Mean	7.586207e-01	2.021101e+03	8.965517e-01	4.397241e+
Std.Dev	1.134278e+00	7.603137e-01	9.227041e-01	3.094683e+
Min	0.000000e+00	2.020000e+03	0.000000e+00	9.100000e+
Q1	0.000000e+00	2.021000e+03	0.000000e+00	1.880000e+
Median	0.000000e+00	2.021000e+03	1.000000e+00	3.820000e+
Q3	2.000000e+00	2.022000e+03	1.000000e+00	5.430000e+
Max	4.000000e+00	2.022000e+03	3.000000e+00	1.196000e+
MAD	0.000000e+00	1.482600e+00	1.482600e+00	2.876244e+
IQR	2.000000e+00	1.000000e+00	1.000000e+00	3.550000e+
CV	1.495185e+00	3.761879e-04	1.029170e+00	7.037783e-(
Skewness	1.191503e+00	-1.707826e-01	7.328005e-01	1.069584e+
SE.Skewness	1.042393e-02	1.042393e-02	1.042393e-02	1.042393e-(
Kurtosis	3.027613e-01	-1.254976e+00	-4.159124e-01	1.613542e-(
N.Valid	5.521600e+04	5.521600e+04	5.521600e+04	5.521600e+
Pct.Valid	1.000000e+02	1.000000e+02	1.000000e+02	1.000000e+

0.6 Data Types

Tipo de datos

[68]:
```
sapply(data, class)
glimpse(data)
```

Hospital 'character' **Especialidad** 'character' **total_pacientes** 'numeric' **ano** 'numeric' **semana** 'numeric' **CODCNH** 'numeric' **id_area** 'numeric' **nombre_area** 'character' **cmunicipio** 'numeric' **Municipio** 'character' **CAMAS** 'numeric' **Clase** 'character' **Dependencia** 'character' **TAC** 'numeric' **RM** 'numeric' **GAM** 'numeric' **HEM** 'numeric' **ASD** 'numeric' **ALI** 'numeric' **SPECT** 'numeric' **MAMOS** 'numeric' **DO** 'numeric' **DIAL** 'numeric' **X** 'numeric' **Y** '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' **capacidad** 'numeric' **pacientes** 'numeric' **consultas** 'numeric' **hospitalizaciones** 'numeric' **Target** 'numeric' **is_train** 'logical'

```
Rows: 55,216
Columns: 46
$ Hospital          <chr> "HOSPITAL REY JUAN CARLOS",
"HOSPITAL CENTRAL DE LA …
$ Especialidad      <chr> "UROLOGÍA", "ODONTOESTOMATOLOGÍA",
"GINECOLOGÍA", "D…
$ total_pacientes   <dbl> 344, 0, 52, 37, 0, 4, 0, 718, 0,
271, 108, 0, 34, 86…
$ ano               <dbl> 2021, 2020, 2021, 2021, 2021, 2020,
2021, 2020, 2021…
$ semana            <dbl> 30, 36, 49, 23, 3, 5, 50, 7, 35, 1,
42, 10, 21, 33, …
$ CODCNH            <dbl> 281348, 280724, 281292, 281292,
281236, 280724, 2807…
```

```
$ id_area           <dbl> 8, 7, 11, 11, 11, 7, 3, 6, 1, 2, 2,
8, 11, 11, 1, 3,…
$ nombre_area       <chr> "SUR-OESTE I", "CENTRO-OESTE", "SUR
II", "SUR II", "…
$ cmunicipio        <dbl> 280920, 280796, 280133, 280133,
281610, 280796, 2800…
$ Municipio         <chr> "MÓSTOLES", "MADRID", "ARANJUEZ",
"ARANJUEZ", "VALDE…
$ CAMAS             <dbl> 382, 475, 98, 98, 182, 475, 507,
613, 269, 1143, 156…
$ Clase             <chr> "HOSPITALES GENERALES", "HOSPITALES
GENERALES", "HOS…
$ Dependencia       <chr> "SERVICIOS E INSTITUTOS DE SALUD DE
LAS COMUNIDADES …
$ TAC               <dbl> 2, 2, 1, 1, 1, 2, 3, 3, 0, 0, 1, 2,
6, 6, 1, 3, 4, 1…
$ RM                <dbl> 3, 2, 1, 1, 2, 2, 2, 3, 0, 0, 0, 2,
5, 5, 1, 2, 4, 1…
$ GAM               <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
2, 2, 0, 0, 2, 0…
$ HEM               <dbl> 1, 2, 0, 0, 1, 2, 1, 2, 0, 0, 0, 1,
3, 3, 0, 1, 1, 0…
$ ASD               <dbl> 2, 1, 1, 1, 1, 1, 1, 3, 0, 0, 0, 1,
2, 2, 0, 1, 2, 1…
$ ALI               <dbl> 1, 2, 0, 0, 0, 2, 0, 4, 0, 0, 0, 0,
3, 3, 0, 2, 2, 0…
$ SPECT             <dbl> 1, 1, 0, 0, 0, 1, 0, 4, 0, 0, 0, 0,
3, 3, 0, 0, 0, 0…
$ MAMOS             <dbl> 2, 1, 1, 1, 1, 1, 2, 2, 0, 0, 1, 2,
3, 3, 1, 1, 3, 1…
$ DO                <dbl> 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1,
2, 2, 0, 1, 2, 0…
$ DIAL              <dbl> 20, 24, 13, 13, 17, 24, 28, 31, 0,
0, 0, 28, 43, 43,…
$ X                 <dbl> -3.870412, -3.745529, -3.610795,
-3.610795, -3.69744…
$ Y                 <dbl> 40.33920, 40.38791, 40.05726,
40.05726, 40.19884, 40…
$ t3_1              <dbl> 42.34715, 45.37878, 42.06149,
42.06149, 42.06149, 45…
$ t1_1              <dbl> 532487, 511605, 899702, 899702,
899702, 511605, 3830…
$ t2_1              <dbl> 0.5122493, 0.5296804, 0.5240445,
0.5240445, 0.524044…
$ t2_2              <dbl> 0.4877507, 0.4703198, 0.4759555,
0.4759555, 0.475955…
$ t4_1              <dbl> 0.1659665, 0.1054260, 0.1540793,
0.1540793, 0.154079…
```

```
$ t4_2              <dbl> 0.6371549, 0.6742432, 0.6753787,
0.6753787, 0.675378…
$ t4_3              <dbl> 0.1968769, 0.2203341, 0.1705449,
0.1705449, 0.170544…
$ t5_1              <dbl> 0.1137647, 0.1744493, 0.1747059,
0.1747059, 0.174705…
$ t6_1              <dbl> 0.1604646, 0.2629599, 0.2641879,
0.2641879, 0.264187…
$ t7_1              <dbl> 0.05422176, 0.05481008, 0.04898547,
0.04898547, 0.04…
$ t8_1              <dbl> 0.04120012, 0.04653221, 0.03679912,
0.03679912, 0.03…
$ t9_1              <dbl> 0.3348780, 0.4914365, 0.3346063,
0.3346063, 0.334606…
$ t10_1             <dbl> 0.13692541, 0.12170996, 0.15173209,
0.15173209, 0.15…
$ t11_1             <dbl> 0.5072726, 0.4915713, 0.5024130,
0.5024130, 0.502413…
$ t12_1             <dbl> 0.5849309, 0.5597213, 0.5900028,
0.5900028, 0.590002…
$ capacidad         <dbl> 17, 0, 8, 5, 0, 5, 1, 24, 6, 6, 30,
4, 2, 15, 20, 6,…
$ pacientes         <dbl> 1447, 1211, 1293, 1501, 1240, 1504,
1502, 1533, 1463…
$ consultas         <dbl> 573, 45, 108, 103, 44, 42, 36,
1119, 34, 466, 220, 6…
$ hospitalizaciones <dbl> 12, 0, 2, 2, 0, 1, 0, 4, 0, 12, 3,
0, 2, 4, 1, 2, 15…
$ Target            <dbl> 54.45, 0.00, 37.96, 23.14, 0.00,
6.25, 0.00, 78.20, …
$ is_train          <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE,
TRUE, TRUE, TRUE…
```

0.7 Statistical Measures

[69]:
```
data  |> descr()
```

A summarytools: 15 × 39 of type dbl

	ALI	ano	ASD	CAMAS
Mean	7.586207e-01	2.021101e+03	8.965517e-01	4.397241e+
Std.Dev	1.134278e+00	7.603137e-01	9.227041e-01	3.094683e+
Min	0.000000e+00	2.020000e+03	0.000000e+00	9.100000e+
Q1	0.000000e+00	2.021000e+03	0.000000e+00	1.880000e+
Median	0.000000e+00	2.021000e+03	1.000000e+00	3.820000e+
Q3	2.000000e+00	2.022000e+03	1.000000e+00	5.430000e+
Max	4.000000e+00	2.022000e+03	3.000000e+00	1.196000e+
MAD	0.000000e+00	1.482600e+00	1.482600e+00	2.876244e+
IQR	2.000000e+00	1.000000e+00	1.000000e+00	3.550000e+
CV	1.495185e+00	3.761879e-04	1.029170e+00	7.037783e-(
Skewness	1.191503e+00	-1.707826e-01	7.328005e-01	1.069584e+
SE.Skewness	1.042393e-02	1.042393e-02	1.042393e-02	1.042393e-(
Kurtosis	3.027613e-01	-1.254976e+00	-4.159124e-01	1.613542e-(
N.Valid	5.521600e+04	5.521600e+04	5.521600e+04	5.521600e+
Pct.Valid	1.000000e+02	1.000000e+02	1.000000e+02	1.000000e+

0.8 Uniques values

[70]:
```
# Rthe number of unique values in each column.
data |> summarise(across(everything(), n_distinct))
```

A tibble: 1 × 46

Hospital <int>	Especialidad <int>	total_pacientes <int>	ano <int>	semana <int>	CODCNH <int>	id_area <int>	nombre_a <int>
29	16	3528	3	51	29	10	10

0.9 CrossTab

Select columns

Hacer los cruces que tengan sentido

[71]:
```
data |> select(where(~ !is.numeric(.x))) |> colnames()
Column1 <- "Especialidad"
Column2 <- "Municipio"
```

1. 'Hospital' 2. 'Especialidad' 3. 'nombre_area' 4. 'Municipio' 5. 'Clase' 6. 'Dependencia' 7. 'is_train'

Operation

[72]:
```
# Referencia cruzada de variables
# Create a contingency table
ctable <- table(data[[Column1]], data[[Column2]])

# Print the contingency table
print(ctable)
```

```
                ALCALÁ DE HENARES ALCORCÓN ARANJUEZ
```

ANGIOLOGÍA Y CIRUGÍA VASCULAR	119	119	119
CIRUGÍA CARDIACA	119	119	119
CIRUGÍA GENERAL Y DEL APARATO DIGESTIVO	119	119	119
CIRUGÍA ORAL Y MAXILOFACIAL	119	119	119
CIRUGÍA PEDIÁTRICA GENERAL	119	119	119
CIRUGÍA PLÁSTICA Y REPARADORA	119	119	119
CIRUGÍA TORÁCICA	119	119	119
DERMATOLOGÍA	119	119	119
GINECOLOGÍA	119	119	119
NEUROCIRUGÍA	119	119	119
ODONTOESTOMATOLOGÍA	119	119	119
OFTALMOLOGÍA	119	119	119
OTORRINOLARINGOLOGÍA	119	119	119
TOTAL	119	119	119
TRAUMATOLOGÍA	119	119	119
UROLOGÍA	119	119	119

	ARGANDA DEL REY	COLLADO VILLALBA
ANGIOLOGÍA Y CIRUGÍA VASCULAR	119	119
CIRUGÍA CARDIACA	119	119
CIRUGÍA GENERAL Y DEL APARATO DIGESTIVO	119	119
CIRUGÍA ORAL Y MAXILOFACIAL	119	119
CIRUGÍA PEDIÁTRICA GENERAL	119	119
CIRUGÍA PLÁSTICA Y REPARADORA	119	119
CIRUGÍA TORÁCICA	119	119
DERMATOLOGÍA	119	119
GINECOLOGÍA	119	119
NEUROCIRUGÍA	119	119
ODONTOESTOMATOLOGÍA	119	119
OFTALMOLOGÍA	119	119
OTORRINOLARINGOLOGÍA	119	119
TOTAL	119	119
TRAUMATOLOGÍA	119	119
UROLOGÍA	119	119

	COSLADA	FUENLABRADA	GETAFE	LEGANÉS
ANGIOLOGÍA Y CIRUGÍA VASCULAR	119	119	119	119
CIRUGÍA CARDIACA	119	119	119	119
CIRUGÍA GENERAL Y DEL APARATO DIGESTIVO	119	119	119	119
CIRUGÍA ORAL Y MAXILOFACIAL	119	119	119	119
CIRUGÍA PEDIÁTRICA GENERAL	119	119	119	119
CIRUGÍA PLÁSTICA Y REPARADORA	119	119	119	119
CIRUGÍA TORÁCICA	119	119	119	119
DERMATOLOGÍA	119	119	119	119
GINECOLOGÍA	119	119	119	119
NEUROCIRUGÍA	119	119	119	119
ODONTOESTOMATOLOGÍA	119	119	119	119
OFTALMOLOGÍA	119	119	119	119

OTORRINOLARINGOLOGÍA	119	119	119	119
TOTAL	119	119	119	119
TRAUMATOLOGÍA	119	119	119	119
UROLOGÍA	119	119	119	119

	MADRID	MAJADAHONDA	MÓSTOLES
ANGIOLOGÍA Y CIRUGÍA VASCULAR	1547	119	238
CIRUGÍA CARDIACA	1547	119	238
CIRUGÍA GENERAL Y DEL APARATO DIGESTIVO	1547	119	238
CIRUGÍA ORAL Y MAXILOFACIAL	1547	119	238
CIRUGÍA PEDIÁTRICA GENERAL	1547	119	238
CIRUGÍA PLÁSTICA Y REPARADORA	1547	119	238
CIRUGÍA TORÁCICA	1547	119	238
DERMATOLOGÍA	1547	119	238
GINECOLOGÍA	1547	119	238
NEUROCIRUGÍA	1547	119	238
ODONTOESTOMATOLOGÍA	1547	119	238
OFTALMOLOGÍA	1547	119	238
OTORRINOLARINGOLOGÍA	1547	119	238
TOTAL	1547	119	238
TRAUMATOLOGÍA	1547	119	238
UROLOGÍA	1547	119	238

	SAN LORENZO DE EL ESCORIAL
ANGIOLOGÍA Y CIRUGÍA VASCULAR	119
CIRUGÍA CARDIACA	119
CIRUGÍA GENERAL Y DEL APARATO DIGESTIVO	119
CIRUGÍA ORAL Y MAXILOFACIAL	119
CIRUGÍA PEDIÁTRICA GENERAL	119
CIRUGÍA PLÁSTICA Y REPARADORA	119
CIRUGÍA TORÁCICA	119
DERMATOLOGÍA	119
GINECOLOGÍA	119
NEUROCIRUGÍA	119
ODONTOESTOMATOLOGÍA	119
OFTALMOLOGÍA	119
OTORRINOLARINGOLOGÍA	119
TOTAL	119
TRAUMATOLOGÍA	119
UROLOGÍA	119

	SAN SEBASTIÁN DE LOS REYES
ANGIOLOGÍA Y CIRUGÍA VASCULAR	119
CIRUGÍA CARDIACA	119
CIRUGÍA GENERAL Y DEL APARATO DIGESTIVO	119
CIRUGÍA ORAL Y MAXILOFACIAL	119
CIRUGÍA PEDIÁTRICA GENERAL	119
CIRUGÍA PLÁSTICA Y REPARADORA	119

```
CIRUGÍA TORÁCICA                                    119
DERMATOLOGÍA                                        119
GINECOLOGÍA                                         119
NEUROCIRUGÍA                                        119
ODONTOESTOMATOLOGÍA                                 119
OFTALMOLOGÍA                                        119
OTORRINOLARINGOLOGÍA                                119
TOTAL                                               119
TRAUMATOLOGÍA                                       119
UROLOGÍA                                            119
```

```
                                    TORREJÓN DE ARDOZ VALDEMORO
ANGIOLOGÍA Y CIRUGÍA VASCULAR                     119       119
CIRUGÍA CARDIACA                                  119       119
CIRUGÍA GENERAL Y DEL APARATO DIGESTIVO           119       119
CIRUGÍA ORAL Y MAXILOFACIAL                       119       119
CIRUGÍA PEDIÁTRICA GENERAL                        119       119
CIRUGÍA PLÁSTICA Y REPARADORA                     119       119
CIRUGÍA TORÁCICA                                  119       119
DERMATOLOGÍA                                      119       119
GINECOLOGÍA                                       119       119
NEUROCIRUGÍA                                      119       119
ODONTOESTOMATOLOGÍA                               119       119
OFTALMOLOGÍA                                      119       119
OTORRINOLARINGOLOGÍA                              119       119
TOTAL                                             119       119
TRAUMATOLOGÍA                                     119       119
UROLOGÍA                                          119       119
```

0.10 Analyzing Numerical Variables

0.10.1 Selecting continuous variables

[73]:
```
# Numeric colums
cdata <- data |> select(where(is.numeric))
cdata <- head(cdata, 50)
```

0.10.2 Global view of the numerical variables

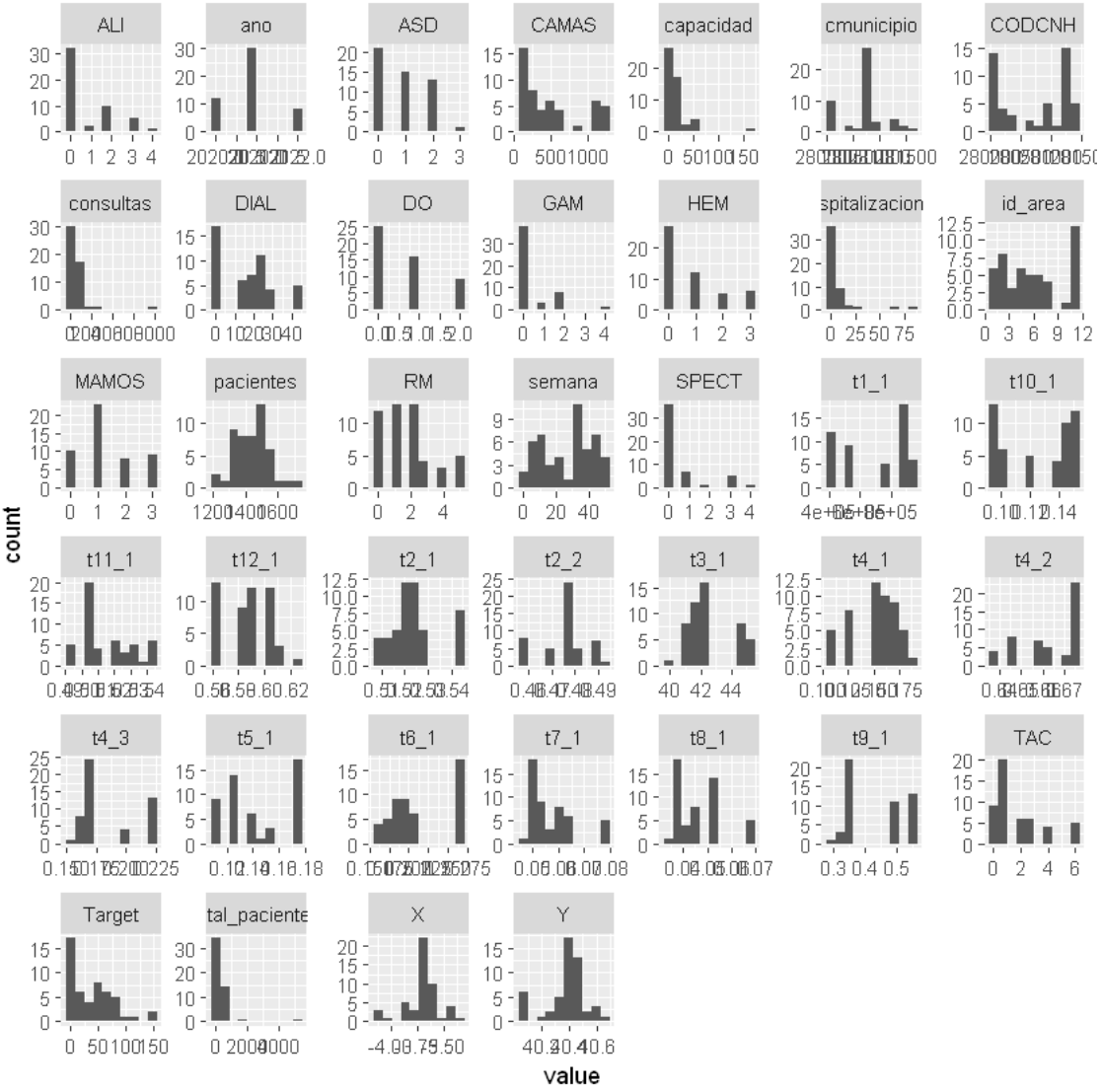
Global view on the dataset to identify some very unusual patterns.

NOTA: Esto puede tardar si hay muchas variables

[74]:
```
# pairs(cdata)
# cdata |> ggpairs()
```

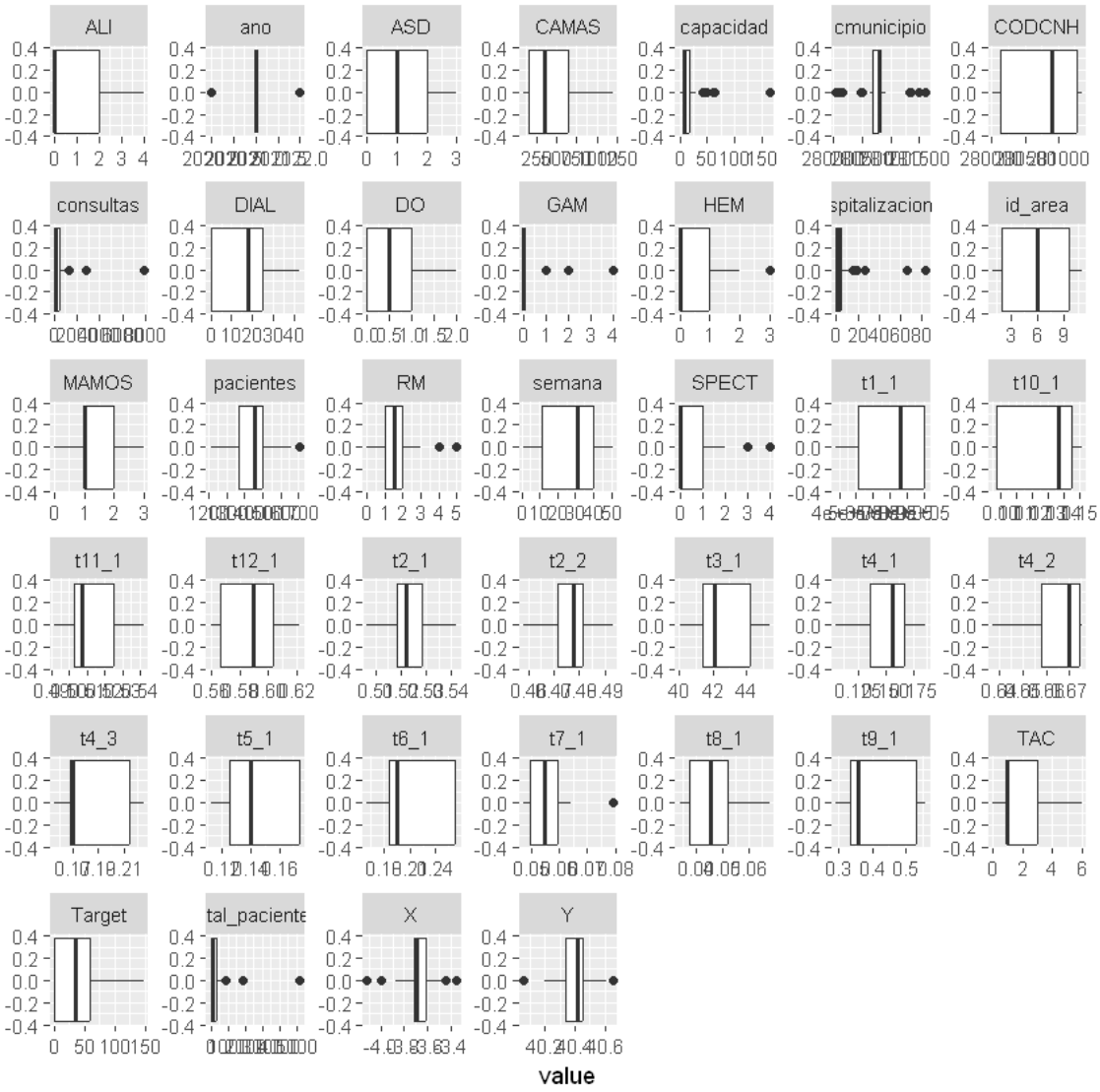
0.10.3 Histograms

[75]:
```
cdata |>
  pivot_longer(cols = everything()) |>
  ggplot(aes(x = value)) +
  geom_histogram(bins = 10) +
  facet_wrap(~name, scales = "free")
```



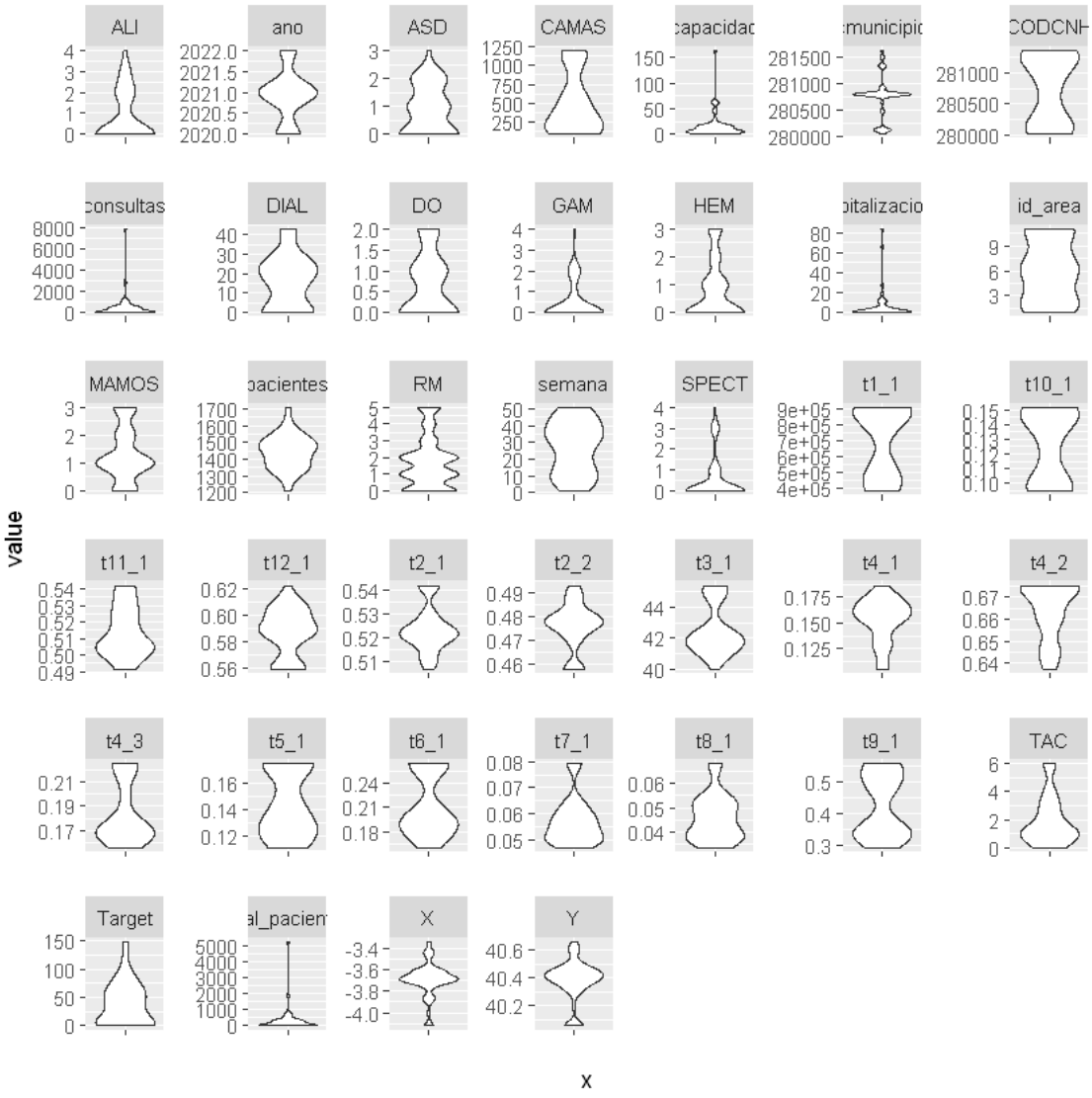
0.10.4 Box plot

[76]:
```
cdata |>
  pivot_longer(cols = everything()) |>
  ggplot(aes(x = value)) +
  geom_boxplot() +
  facet_wrap(~name, scales = "free")
```



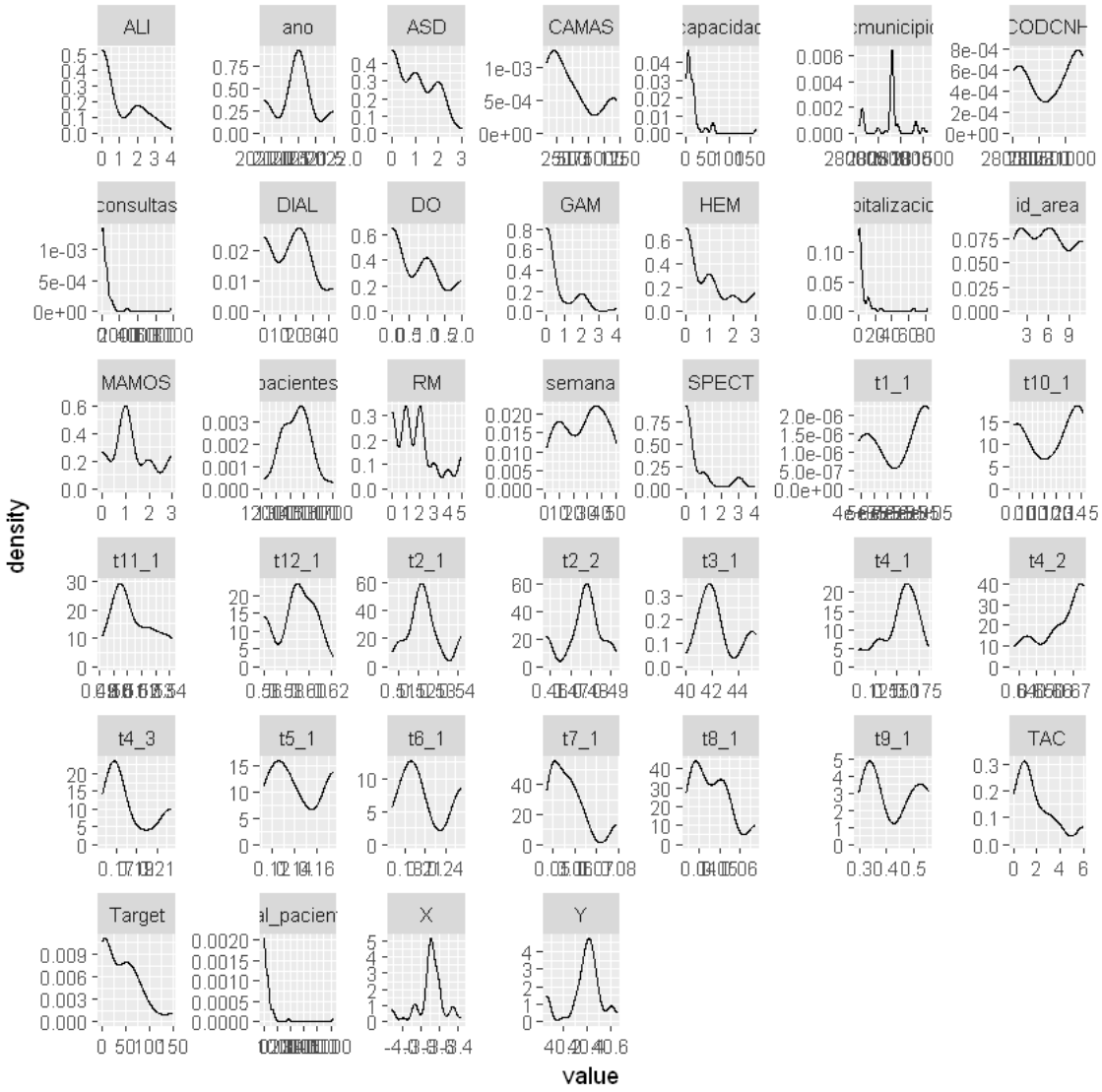
0.10.5 Violin plot

[77]:
```
cdata |>
  pivot_longer(cols = everything()) |>
  ggplot(aes(x = "", y = value)) +
  geom_violin() +
  facet_wrap(~name, scales = "free")
```



0.10.6 Distribution plot

[78]:
```
cdata |>
  pivot_longer(cols = everything()) |>
  ggplot(aes(x = value)) +
  geom_density() +
  facet_wrap(~name, scales = "free")
```



0.11 Analyzing Categorical Variables

0.11.1 Selecting categorical variables

[79]:
```
# Category colums
char_cols <- data |> select(where(~ !is.numeric(.x))) |> colnames()
char_cols
```

1. 'Hospital' 2. 'Especialidad' 3. 'nombre_area' 4. 'Municipio' 5. 'Clase' 6. 'Dependencia' 7. 'is_train'

[80]:
```
# Category colums
char_data <- data |> select(where(~ !is.numeric(.x)))
head(char_data)
```

A tibble: 6 × 7

Hospital <chr>	Especialidad <chr>	r
HOSPITAL REY JUAN CARLOS	UROLOGÍA	S
HOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA	ODONTOESTOMATOLOGÍA	C
HOSPITAL UNIVERSITARIO DEL TAJO	GINECOLOGÍA	S
HOSPITAL UNIVERSITARIO DEL TAJO	DERMATOLOGÍA	S
HOSPITAL UNIVERSITARIO INFANTA ELENA	ODONTOESTOMATOLOGÍA	S
HOSPITAL CENTRAL DE LA DEFENSA GOMEZ ULLA	CIRUGÍA TORÁCICA	C

0.11.2 Visualization of categorical variables

0.12 Statistical Normality Tests

[81]:
```
cdata_long <- cdata |>
  pivot_longer(cols = everything())
```

0.12.1 Test de Shapiro-Wilk

Si hay muchos datos este no se puede hacer

[82]:
```
tapply(cdata_long$value, cdata_long$name, shapiro.test)
```

```
$ALI

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.6947, p-value = 6.657e-09


$ano

        Shapiro-Wilk normality test

data:  X[[i]]
```

```
W = 0.78117, p-value = 3.316e-07


$ASD

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.81024, p-value = 1.522e-06


$CAMAS

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.83471, p-value = 6.115e-06


$capacidad

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.55977, p-value = 5.149e-11


$cmunicipio

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.84604, p-value = 1.211e-05


$CODCNH

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.81409, p-value = 1.88e-06


$consultas

        Shapiro-Wilk normality test

data:  X[[i]]
```

```
W = 0.39785, p-value = 5e-13


$DIAL

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.87571, p-value = 8.314e-05


$DO

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.75776, p-value = 1.06e-07


$GAM

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.5598, p-value = 5.153e-11


$HEM

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.73934, p-value = 4.524e-08


$hospitalizaciones

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.4561, p-value = 2.364e-12


$id_area

        Shapiro-Wilk normality test

data:  X[[i]]
```

```
W = 0.88563, p-value = 0.0001666


$MAMOS

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.8455, p-value = 1.171e-05


$pacientes

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.98843, p-value = 0.9027


$RM

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.87087, p-value = 5.98e-05


$semana

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.93188, p-value = 0.006515


$SPECT

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.58175, p-value = 1.053e-10


$t1_1

        Shapiro-Wilk normality test

data:  X[[i]]
```

```
W = 0.7821, p-value = 3.476e-07


$t10_1

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.80572, p-value = 1.19e-06


$t11_1

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.89374, p-value = 0.0003005


$t12_1

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.87254, p-value = 6.694e-05


$t2_1

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.89198, p-value = 0.0002638


$t2_2

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.89067, p-value = 0.0002398


$t3_1

        Shapiro-Wilk normality test

data:  X[[i]]
```

```
W = 0.81175, p-value = 1.653e-06


$t4_1

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.87105, p-value = 6.054e-05


$t4_2

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.81874, p-value = 2.438e-06


$t4_3

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.75101, p-value = 7.724e-08


$t5_1

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.83106, p-value = 4.934e-06


$t6_1

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.79946, p-value = 8.52e-07


$t7_1

        Shapiro-Wilk normality test

data:  X[[i]]
```

```
W = 0.81847, p-value = 2.401e-06


$t8_1

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.86068, p-value = 3.05e-05


$t9_1

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.79704, p-value = 7.497e-07


$TAC

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.82779, p-value = 4.08e-06


$Target

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.87291, p-value = 6.867e-05


$total_pacientes

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.39408, p-value = 4.539e-13


$X

        Shapiro-Wilk normality test

data:  X[[i]]
```

```
W = 0.89074, p-value = 0.000241


$Y

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.8732, p-value = 7.001e-05
```

0.12.2 Test de Anderson-Darling

[83]:
```
tapply(cdata_long$value, cdata_long$name, ad.test)
```

```
$ALI

        Anderson-Darling normality test

data:  X[[i]]
A = 7.2641, p-value < 2.2e-16


$ano

        Anderson-Darling normality test

data:  X[[i]]
A = 5.2013, p-value = 4.857e-13


$ASD

        Anderson-Darling normality test

data:  X[[i]]
A = 3.9466, p-value = 5.636e-10


$CAMAS

        Anderson-Darling normality test

data:  X[[i]]
A = 2.9171, p-value = 1.928e-07
```

```
$capacidad

        Anderson-Darling normality test

data:  X[[i]]
A = 6.4759, p-value = 3.979e-16


$cmunicipio

        Anderson-Darling normality test

data:  X[[i]]
A = 3.9704, p-value = 4.926e-10


$CODCNH

        Anderson-Darling normality test

data:  X[[i]]
A = 3.5705, p-value = 4.729e-09


$consultas

        Anderson-Darling normality test

data:  X[[i]]
A = 8.9599, p-value < 2.2e-16


$DIAL

        Anderson-Darling normality test

data:  X[[i]]
A = 2.0702, p-value = 2.416e-05


$DO

        Anderson-Darling normality test

data:  X[[i]]
A = 5.0388, p-value = 1.207e-12
```

```
$GAM

        Anderson-Darling normality test

data:  X[[i]]
A = 10.381, p-value < 2.2e-16


$HEM

        Anderson-Darling normality test

data:  X[[i]]
A = 5.3011, p-value = 2.778e-13


$hospitalizaciones

        Anderson-Darling normality test

data:  X[[i]]
A = 9.2752, p-value < 2.2e-16


$id_area

        Anderson-Darling normality test

data:  X[[i]]
A = 1.7151, p-value = 0.0001847


$MAMOS

        Anderson-Darling normality test

data:  X[[i]]
A = 3.1682, p-value = 4.629e-08


$pacientes

        Anderson-Darling normality test

data:  X[[i]]
A = 0.24866, p-value = 0.736
```

```
$RM

        Anderson-Darling normality test

data:  X[[i]]
A = 2.1257, p-value = 1.76e-05


$semana

        Anderson-Darling normality test

data:  X[[i]]
A = 1.0734, p-value = 0.007375


$SPECT

        Anderson-Darling normality test

data:  X[[i]]
A = 9.5767, p-value < 2.2e-16


$t1_1

        Anderson-Darling normality test

data:  X[[i]]
A = 4.5047, p-value = 2.423e-11


$t10_1

        Anderson-Darling normality test

data:  X[[i]]
A = 3.7218, p-value = 2.009e-09


$t11_1

        Anderson-Darling normality test

data:  X[[i]]
A = 2.0742, p-value = 2.362e-05
```

```
$t12_1

        Anderson-Darling normality test

data:  X[[i]]
A = 2.6325, p-value = 9.748e-07


$t2_1

        Anderson-Darling normality test

data:  X[[i]]
A = 2.1094, p-value = 1.93e-05


$t2_2

        Anderson-Darling normality test

data:  X[[i]]
A = 2.1684, p-value = 1.378e-05


$t3_1

        Anderson-Darling normality test

data:  X[[i]]
A = 4.3577, p-value = 5.545e-11


$t4_1

        Anderson-Darling normality test

data:  X[[i]]
A = 2.5265, p-value = 1.783e-06


$t4_2

        Anderson-Darling normality test

data:  X[[i]]
A = 3.4201, p-value = 1.109e-08
```

```
$t4_3

        Anderson-Darling normality test

data:  X[[i]]
A = 5.746, p-value = 2.313e-14


$t5_1

        Anderson-Darling normality test

data:  X[[i]]
A = 3.179, p-value = 4.354e-08


$t6_1

        Anderson-Darling normality test

data:  X[[i]]
A = 4.3746, p-value = 5.043e-11


$t7_1

        Anderson-Darling normality test

data:  X[[i]]
A = 2.777, p-value = 4.279e-07


$t8_1

        Anderson-Darling normality test

data:  X[[i]]
A = 2.2259, p-value = 9.918e-06


$t9_1

        Anderson-Darling normality test

data:  X[[i]]
A = 4.4212, p-value = 3.878e-11
```

```
$TAC

        Anderson-Darling normality test

data:  X[[i]]
A = 3.2061, p-value = 3.732e-08


$Target

        Anderson-Darling normality test

data:  X[[i]]
A = 1.8527, p-value = 8.394e-05


$total_pacientes

        Anderson-Darling normality test

data:  X[[i]]
A = 9.069, p-value < 2.2e-16


$X

        Anderson-Darling normality test

data:  X[[i]]
A = 2.2919, p-value = 6.805e-06


$Y

        Anderson-Darling normality test

data:  X[[i]]
A = 2.4939, p-value = 2.149e-06
```

0.12.3 Test de Lilliefors

[84]:
```
tapply(cdata_long$value, cdata_long$name, lillie.test)
```

```
$ALI

        Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data:  X[[i]]
D = 0.39463, p-value < 2.2e-16


$ano

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.31023, p-value = 3.047e-13


$ASD

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.26349, p-value = 2.251e-09


$CAMAS

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.18643, p-value = 0.0001573


$capacidad

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.30254, p-value = 1.473e-12


$cmunicipio

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.29784, p-value = 3.776e-12


$CODCNH

        Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data:  X[[i]]
D = 0.22159, p-value = 1.672e-06


$consultas

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.33388, p-value = 1.814e-15


$DIAL

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.216, p-value = 3.66e-06


$DO

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.31212, p-value = 2.052e-13


$GAM

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.45374, p-value < 2.2e-16


$HEM

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.31699, p-value = 7.344e-14


$hospitalizaciones

        Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data:  X[[i]]
D = 0.3286, p-value = 5.905e-15


$id_area

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.15542, p-value = 0.004051


$MAMOS

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.28566, p-value = 4.021e-11


$pacientes

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.070472, p-value = 0.7738


$RM

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.19883, p-value = 3.515e-05


$semana

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.14038, p-value = 0.01516


$SPECT

        Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data:  X[[i]]
D = 0.41928, p-value < 2.2e-16


$t1_1

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.26015, p-value = 3.997e-09


$t10_1

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.21451, p-value = 4.49e-06


$t11_1

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.24929, p-value = 2.444e-08


$t12_1

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.19816, p-value = 3.822e-05


$t2_1

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.23069, p-value = 4.443e-07


$t2_2

        Lilliefors (Kolmogorov-Smirnov) normality test
```

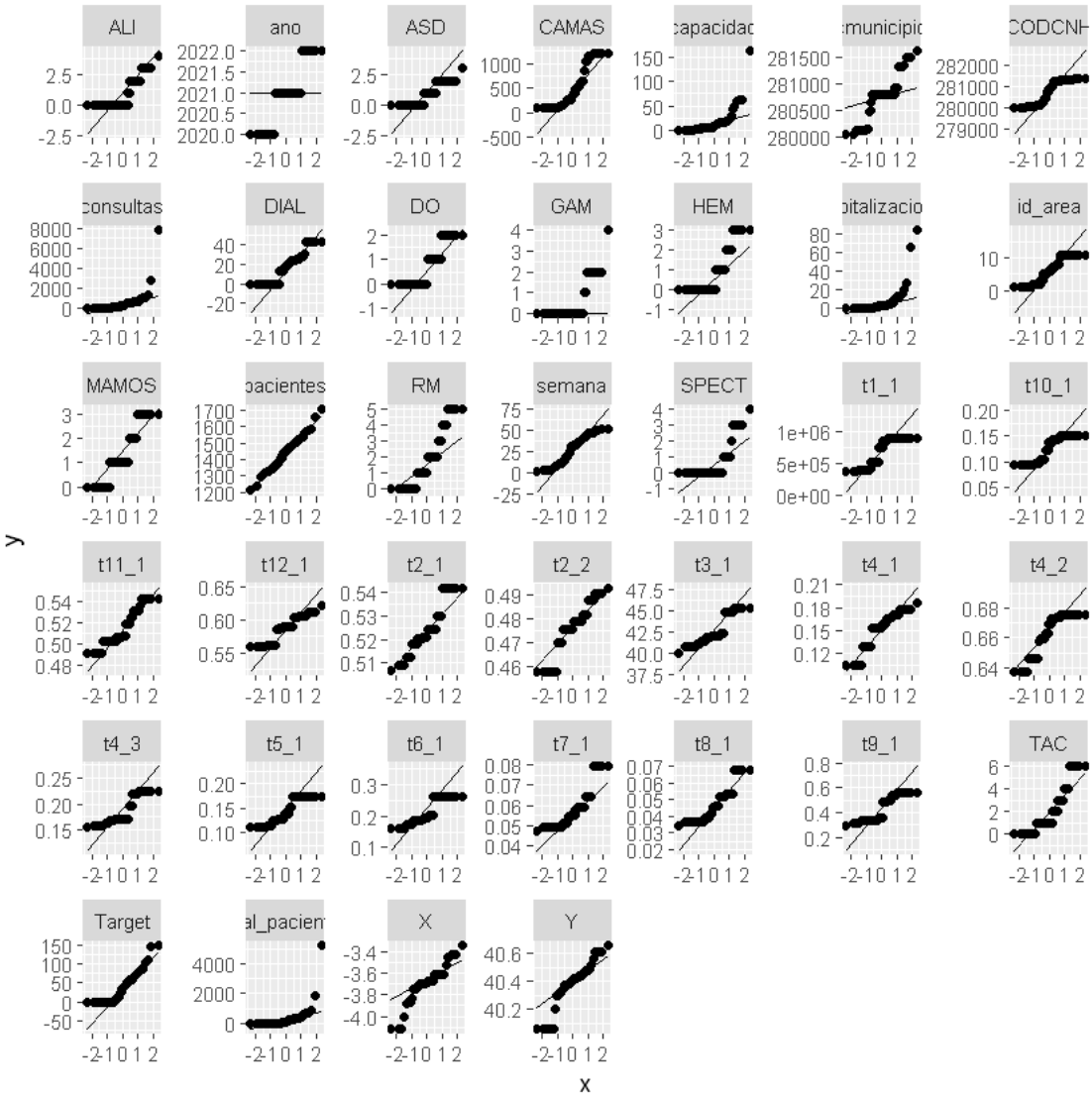
```
data:  X[[i]]
D = 0.23397, p-value = 2.714e-07


$t3_1

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.29141, p-value = 1.335e-11


$t4_1

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.27134, p-value = 5.644e-10


$t4_2

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.23442, p-value = 2.537e-07


$t4_3

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.36784, p-value < 2.2e-16


$t5_1

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.23353, p-value = 2.902e-07


$t6_1

        Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data:  X[[i]]
D = 0.24703, p-value = 3.528e-08


$t7_1

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.18735, p-value = 0.0001413


$t8_1

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.16408, p-value = 0.001757


$t9_1

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.27338, p-value = 3.91e-10


$TAC

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.27597, p-value = 2.441e-10


$Target

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.1656, p-value = 0.001508


$total_pacientes

        Lilliefors (Kolmogorov-Smirnov) normality test
```

```
data:  X[[i]]
D = 0.33907, p-value = 5.572e-16


$X

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.20848, p-value = 1.012e-05


$Y

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.20794, p-value = 1.087e-05
```

0.12.4 QQ-plots

[85]:
```
cdata |>
  pivot_longer(cols = everything()) |>
  ggplot(aes(sample = value)) +
  geom_qq() +
  geom_qq_line() +
  facet_wrap(~name, scales = "free")
```



0.13 Bivariate analysis

- Ver gráficos de dispersión y ggpairs arriba
- Completar si es necesario con alguna comparación específica (gráfico de dispersión o boxplot por grupos)

Correlaciones

[86]:
```
cor(cdata, use = "pairwise.complete.obs")
```

A matrix: 39 × 39 of type dbl

	total_pacientes	ano	semana	CODCNH	i
total_pacientes	1.000000000	0.015072844	-0.065076294	-0.19430715	-
ano	0.015072844	1.000000000	-0.255009696	-0.22373644	-
semana	-0.065076294	-0.255009696	1.000000000	0.14510980	0
CODCNH	-0.194307145	-0.223736437	0.145109802	1.00000000	0
id_area	-0.190014107	-0.044276161	0.039448729	0.07879005	1
cmunicipio	0.066341035	0.067890616	-0.133486833	-0.15296528	-
CAMAS	0.336196650	0.164561885	-0.127972906	-0.71776882	0
TAC	-0.043785077	0.066061277	0.187626371	-0.36870927	0
RM	-0.055798691	-0.019827426	0.195928021	-0.20458234	0
GAM	0.152550212	0.136150902	0.135751701	-0.49746417	0
HEM	0.023236887	0.006135085	0.109571988	-0.38746608	0
ASD	-0.002672427	-0.054640066	0.021111767	-0.10671986	0
ALI	0.031968878	-0.019485107	0.028693289	-0.32913218	0
SPECT	-0.022068399	0.007208873	0.041894755	-0.24995346	0
MAMOS	-0.042390016	-0.055456225	0.211114730	-0.16337675	0
DO	0.016303692	0.030201002	0.099071048	-0.38474474	0
DIAL	-0.063650366	-0.070978287	0.168172754	-0.08955304	0
X	0.048355588	0.104575704	0.062253726	0.04710977	-
Y	0.119979788	0.004305994	0.021592552	-0.34704892	-
t3_1	0.152258631	-0.052174498	-0.235183527	-0.34086859	-
t1_1	-0.112392380	0.179771695	0.006105534	0.05857178	0
t2_1	0.219547701	0.156825981	-0.313710849	-0.44957026	-
t2_2	-0.223372482	-0.158352949	0.318242122	0.45854927	0
t4_1	-0.096533999	0.052702699	0.201839280	0.31045474	0
t4_2	-0.195981379	0.078254485	0.118247162	0.13766028	0
t4_3	0.185366766	-0.087210458	-0.235751711	-0.33859704	-
t5_1	-0.148725821	0.057059643	0.043178830	-0.02231410	0
t6_1	-0.132813418	0.077549617	-0.033666347	-0.09199249	0
t7_1	0.083629261	-0.046867375	-0.085581292	-0.03759493	-
t8_1	0.128467157	-0.042406216	-0.141597770	-0.13150022	-
t9_1	0.230535599	-0.016791543	-0.290853292	-0.37662854	-
t10_1	-0.243382553	0.030782513	0.260240659	0.34470449	0
t11_1	0.069703811	0.023692711	0.048837532	0.01829785	-
t12_1	-0.127024912	0.053516487	0.253982619	0.29170826	0
capacidad	0.875322662	-0.061540286	-0.063136237	-0.22826114	-
pacientes	0.126162282	-0.057162673	0.123566872	-0.24002853	-
consultas	0.999961634	0.015536770	-0.065936204	-0.19367793	-
hospitalizaciones	0.794250190	0.044480511	-0.005287760	-0.24455709	-
Target	0.316682109	-0.062370615	0.014929694	-0.06230871	-

0.14 Regression analysis

0.14.1 Modelo completo regresión lineal simple

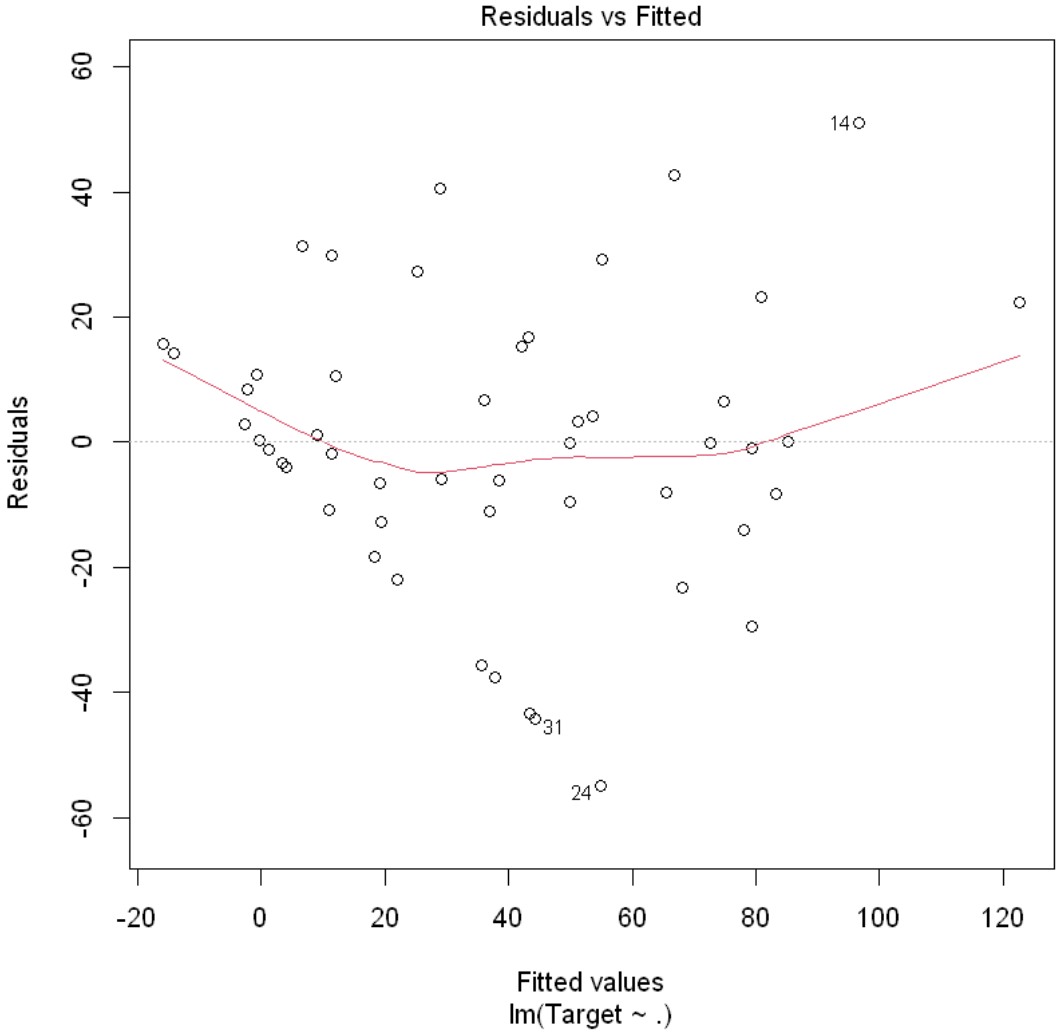
[87]:
```
modelo <- lm(Target ~ ., data = cdata)
summary(modelo)
```

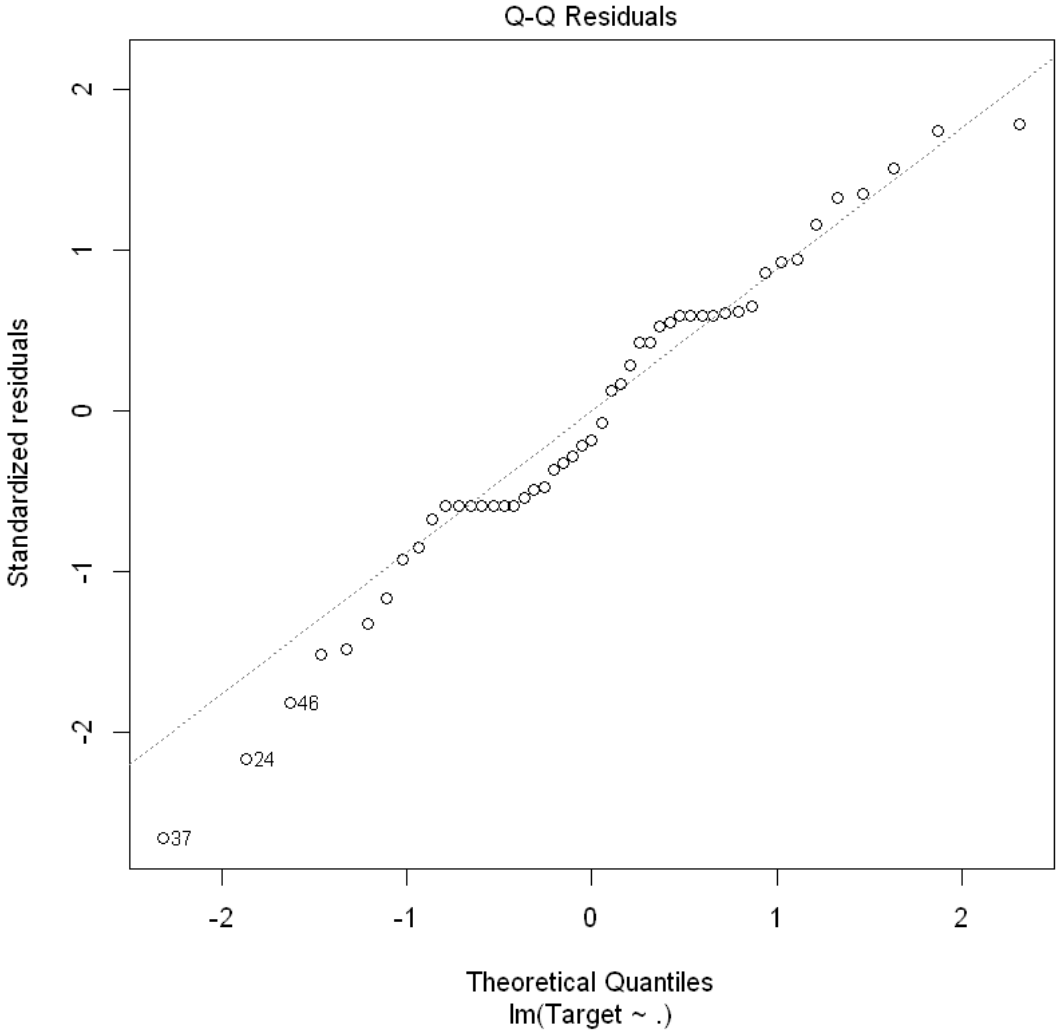
```
Call:
lm(formula = Target ~ ., data = cdata)

Residuals:
    Min      1Q  Median      3Q     Max
-54.882 -10.576  -0.073  13.368  50.968

Coefficients: (8 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        9.661e+05  9.511e+05   1.016   0.3225
total_pacientes   -9.296e-01  1.227e+00  -0.758   0.4578
ano               -4.167e+01  1.401e+01  -2.974   0.0078 **
semana            -3.239e-01  5.438e-01  -0.596   0.5585
CODCNH             2.016e-01  8.553e-02   2.357   0.0293 *
id_area           -1.670e+01  6.839e+01  -0.244   0.8097
cmunicipio         4.052e-02  8.307e-02   0.488   0.6313
CAMAS              1.512e-01  1.020e-01   1.482   0.1546
TAC                2.015e+02  9.082e+01   2.219   0.0389 *
RM                 9.319e+00  4.767e+01   0.196   0.8471
GAM               -3.268e+01  5.014e+01  -0.652   0.5224
HEM               -5.000e+01  4.333e+01  -1.154   0.2628
ASD               -3.217e+01  4.003e+01  -0.804   0.4316
ALI               -1.632e+02  8.023e+01  -2.034   0.0561 .
SPECT              8.931e+01  4.879e+01   1.830   0.0829 .
MAMOS             -6.666e+01  8.683e+01  -0.768   0.4521
DO                 9.177e+01  5.833e+01   1.573   0.1322
DIAL              -4.919e+00  4.962e+00  -0.991   0.3340
X                 -5.747e+02  1.460e+03  -0.394   0.6983
Y                 -1.353e+03  1.033e+03  -1.310   0.2059
t3_1               1.294e+03  1.867e+03   0.693   0.4965
t1_1              -9.413e-04  1.066e-03  -0.883   0.3884
t2_1              -1.005e+06  1.054e+06  -0.953   0.3526
t2_2              -9.962e+05  1.040e+06  -0.958   0.3503
t4_1               8.624e+04  1.346e+05   0.641   0.5293
t4_2               5.470e+04  5.955e+04   0.919   0.3698
t4_3                      NA         NA      NA       NA
t5_1              -5.036e+03  1.594e+04  -0.316   0.7554
t6_1                      NA         NA      NA       NA
t7_1                      NA         NA      NA       NA
t8_1                      NA         NA      NA       NA
t9_1                      NA         NA      NA       NA
```

```
t10_1                       NA         NA      NA       NA
t11_1                       NA         NA      NA       NA
t12_1                       NA         NA      NA       NA
capacidad          -2.562e+00  1.228e+00  -2.087   0.0506 .
pacientes          -2.050e-02  7.233e-02  -0.283   0.7799
consultas           6.627e-01  8.171e-01   0.811   0.4274
hospitalizaciones   1.388e+00  1.951e+00   0.712   0.4853
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.06 on 19 degrees of freedom
Multiple R-squared:  0.6698,        Adjusted R-squared:  0.1484
F-statistic: 1.285 on 30 and 19 DF,  p-value: 0.2877
```

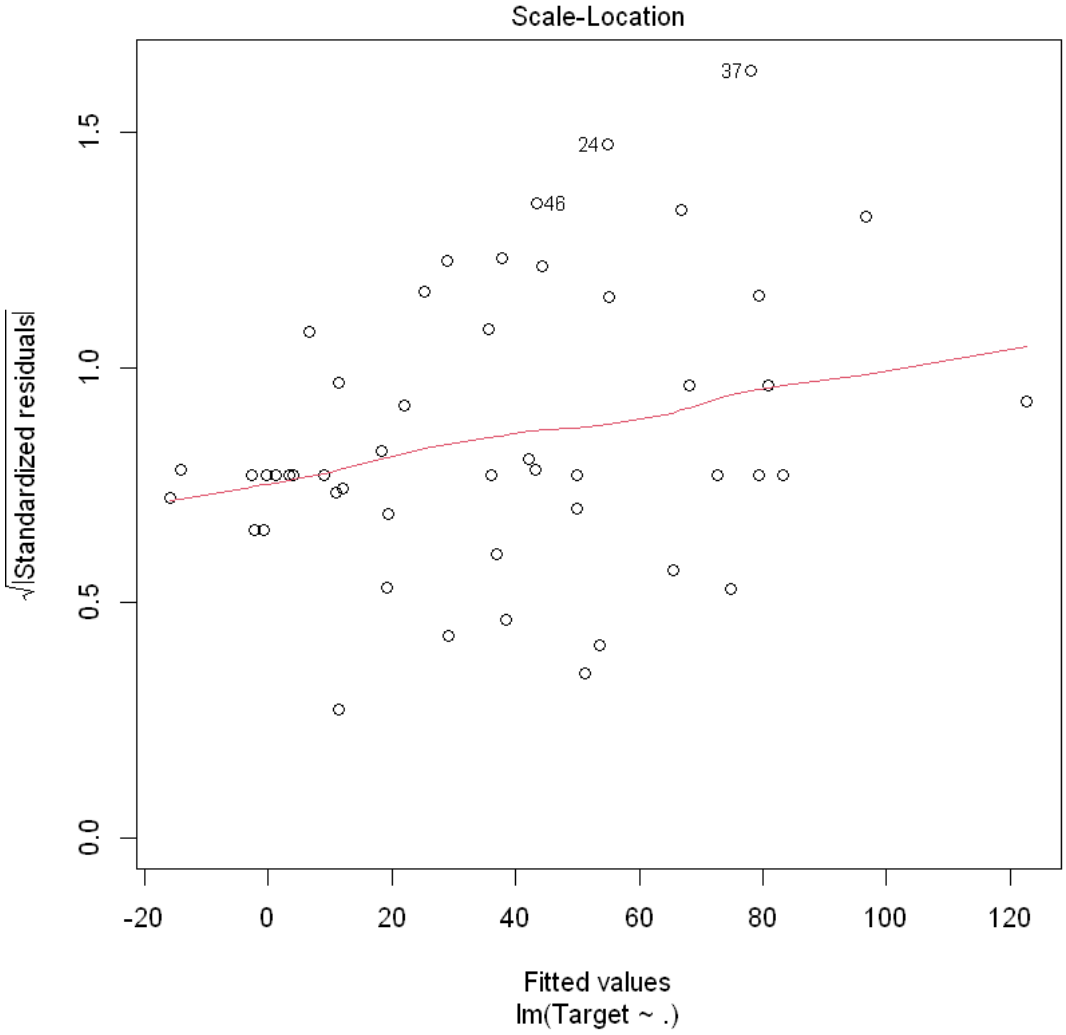
[88]:
```
plot(modelo)
```

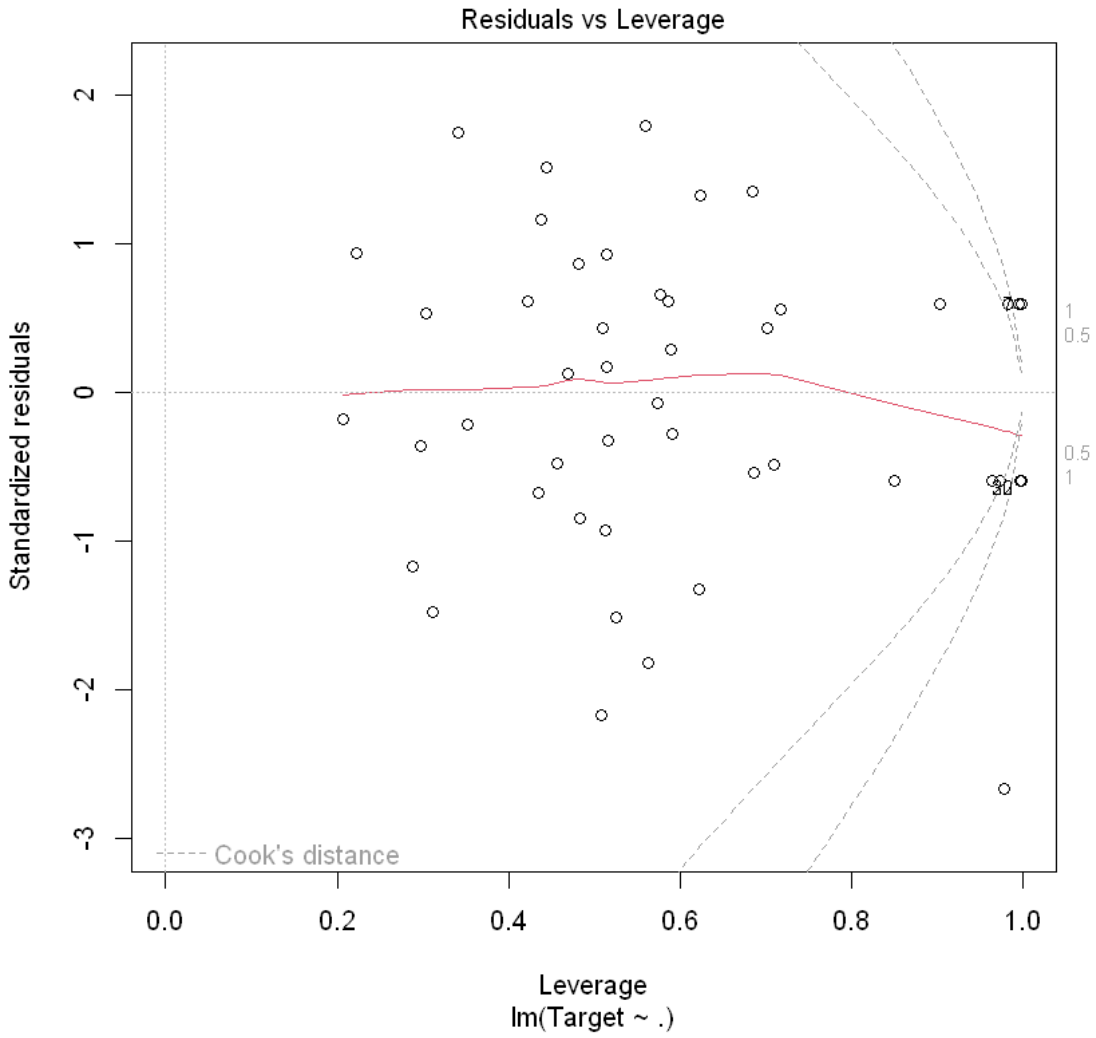
```
Warning message:
"not plotting observations with leverage one:
  25"
```


Residuals vs Fitted
Residuals
Fitted values
lm(Target ~ .)
60
40
20
0
-20
-40
-60
-20
0
20
40
60
80
100
120
14
31
24




```
Warning message in sqrt(crit * p * (1 - hh)/hh):
"NaNs produced"
Warning message in sqrt(crit * p * (1 - hh)/hh):
"NaNs produced"
```


Scale-Location
37
24
46
√|Standardized residuals|
1.5
1.0
0.5
0.0
-20
0
20
40
60
80
100
120
Fitted values
lm(Target ~ .)



0.14.2 Selección de variables

Puede que dé error por la estructura de los datos, en ese caso dejarlo indicado

[89]:
```
modelo2 <- step(modelo, trace = FALSE)
summary(modelo2)
```

```
Call:
lm(formula = Target ~ total_pacientes + ano + CODCNH + cmunicipio +
    CAMAS + TAC + HEM + ALI + SPECT + MAMOS + DO + DIAL + Y +
    t3_1 + t1_1 + t2_1 + t2_2 + t4_2 + t5_1 + capacidad + consultas,
    data = cdata)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-55.927 -12.013   0.651  14.491  52.786

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.280e+05  8.924e+04   3.676 0.000995 ***
total_pacientes -1.355e+00  8.522e-01  -1.590 0.123017
ano             -3.097e+01  9.899e+00  -3.129 0.004077 **
CODCNH           1.410e-01  4.840e-02   2.914 0.006946 **
cmunicipio       3.312e-02  2.725e-02   1.216 0.234321
CAMAS            7.888e-02  3.240e-02   2.434 0.021552 *
TAC              1.337e+02  3.937e+01   3.395 0.002069 **
HEM             -2.572e+01  1.881e+01  -1.368 0.182339
ALI             -8.843e+01  3.475e+01  -2.545 0.016732 *
SPECT            4.331e+01  1.714e+01   2.527 0.017432 *
MAMOS           -5.821e+01  2.144e+01  -2.715 0.011212 *
DO               5.174e+01  2.590e+01   1.998 0.055568 .
DIAL            -5.423e+00  2.216e+00  -2.447 0.020951 *
Y               -5.763e+02  2.097e+02  -2.748 0.010379 *
t3_1             3.877e+01  1.907e+01   2.033 0.051637 .
t1_1            -2.233e-04  9.135e-05  -2.445 0.021044 *
t2_1            -2.965e+05  9.116e+04  -3.253 0.002976 **
t2_2            -2.992e+05  9.126e+04  -3.279 0.002786 **
t4_2             9.175e+03  3.320e+03   2.764 0.009984 **
t5_1            -6.324e+03  2.166e+03  -2.919 0.006857 **
capacidad       -2.211e+00  8.354e-01  -2.647 0.013176 *
consultas        9.517e-01  5.695e-01   1.671 0.105850
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.7 on 28 degrees of freedom
Multiple R-squared:  0.6238,        Adjusted R-squared:  0.3417
F-statistic: 2.211 on 21 and 28 DF,  p-value: 0.02531
```

0.15 Stationary analysis

- Si hay una variable fecha, usarla
- Si hay mes, o semana, convertir a fecha

[90]:
```
tsdata <- data |>
    mutate(fecha = as.Date(parse_date_time(paste(ano, semana, 1, sep="/"),'Y/W/
  ↪W')))
```

0.16 Data Save

- No aplica

Identificamos los datos a guardar

[ ]:
```
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

0.16.1 Proceso 12

[ ]:
```
# caso <- "CU_XX"
# proceso <- '_10'
# tarea <- "_XX"
# archivo <- ""
# proper <- "_xxxxx"
# 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

[ ]:
```
# 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/.

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

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

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

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

0.17 REPORT

A continuación se realizará un informe de las acciones realizadas

0.18 Main Actions Carried Out

- Se ha realizado exploratorio de los datos del caso de uso

0.19 Main Conclusions

- Los datos son adecuados para el caso de uso

0.20 CODE TO DEPLOY (PILOT)

A continuación se incluirá el código que deba ser llevado a despliegue para producción, dado que se entiende efectúa operaciones necesarias sobre los datos en la ejecución del prototipo

Description

- No hay nada que desplegar en el piloto, ya que estos datos son estáticos o en todo caso cambian con muy poca frecuencia, altamente improbable durante el proyecto.

CODE