05. - Data Collection_CU_18_16_datos_distritos_v_01

June 13, 2023

#

CU18_Infraestructuras_eventos

Citizenlab Data Science Methodology > II - Data Processing Domain *** > # 05.- Data Collection

Data Collection is the process to obtain and generate (if required) necessary data to model the problem.

0.0.1 16. Unión de todos los datos de distritos

- Unir todas los ficheros que contienen datos agregados de distritos

Table of Contents

Settings

Data Load

ETL Processes

Import data from: CSV, Excel, Tab, JSON, SQL, and Parquet files

Synthetic Data Generation

Fake Data Generation

Open Data

Data Save

Main Conclusions

Main Actions

Acciones done

Acctions to perform

0.1 Settings

0.1.1 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")
```

'es_ES.UTF-8/es_ES.UTF-8/es_ES.UTF-8/C/es_ES.UTF-8/C'

0.1.2 Packages to use

- {tcltk} para selección interactiva de archivos locales
- {sf} para trabajar con georeferenciación
- {readr} para leer y escribir archivos csv
- {dplyr} para explorar datos

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

```
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
```

0.1.3 Paths

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

0.2 Data Load

If there are more than one input file, make as many sections as files to import.

Instrucciones - Los ficheros de entrada del proceso están siempre en Data/Input/.
- Si hay más de un fichero de entrada, se crean tantos objetos iFile_xx y file_data_xx como ficheros de entrada (xx número correlativo con dos dígitos, rellenar con ceros a la izquierda)

1. Infraestructuras

OPCION A: Seleccionar fichero en ventana para mayor comodidad

Data load using the {tcltk} package. Ucomment the line if not using this option

[4]:
```
# file_data_01 <- tcltk::tk_choose.files(multi = FALSE)
```

OPCION B: Especificar el nombre de archivo

[5]:
```
iFile_01 <- "CU_18_05_05_01_infraestructuras_distrito.csv"
file_data_01 <- paste0(iPath, iFile_01)

if(file.exists(file_data_01)){
    cat("Se leerán datos del archivo: ", file_data_01)
} else{
    warning("Cuidado: el archivo no existe.")
}
```

```
Se leerán datos del archivo:
Data/Input/CU_18_05_05_01_infraestructuras_distrito.csv
```

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

[6]:
```
data_01 <- read_csv(file_data_01)
```

```
Rows: 239 Columns: 14
  Column specification

Delimiter: ","
chr  (2): cmun, cdis
dbl (12): consultorios_de_salud, helisuperficies,
centros_de_atencion_a_drog…

 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:

[7]:
```
glimpse(data_01)
```

```
Rows: 239
Columns: 14
$ cmun                                  <chr> "001", "002",
"003", "004", "0…
$ cdis                                  <chr> "01", "01",
"01", "01", "01", …
$ consultorios_de_salud                 <dbl> 1, 1, 1, 1,
0, 0, 0, 0, 0, 0, …
$ helisuperficies                       <dbl> 1, 0, 1, 0,
0, 0, 0, 0, 0, 0, …
$ centros_de_atencion_a_drogodependientes <dbl> 0, 0, 0, 0,
1, 0, 0, 0, 0, 1, …
$ centros_de_salud                      <dbl> 0, 0, 0, 0,
2, 3, 2, 1, 2, 5, …
$ estaciones_de_cercanias               <dbl> 0, 0, 0, 0,
```

```
1, 0, 1, 1, 0, 2, …
$ hospitales                          <dbl> 0, 0, 0, 0,
1, 0, 0, 1, 0, 0, …
$ otros_centros_de_salud              <dbl> 0, 0, 0, 0,
1, 0, 0, 0, 0, 0, …
$ centros_de_salud_mental             <dbl> 0, 0, 0, 0,
0, 1, 0, 1, 0, 1, …
$ centros_de_especialidades           <dbl> 0, 0, 0, 0,
0, 0, 0, 1, 0, 1, …
$ bocas_de_metro                      <dbl> 0, 0, 0, 0,
0, 0, 0, 0, 0, 9, …
$ intercambiadores                    <dbl> 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, …
$ aeropuertos                         <dbl> 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, …
```

Muestra de datos:

2. Servicios

OPCION A: Seleccionar fichero en ventana para mayor comodidad

Data load using the {tcltk} package. Ucomment the line if not using this option

[8]:
```
# file_data_02 <- tcltk::tk_choose.files(multi = FALSE)
```

OPCION B: Especificar el nombre de archivo

[9]:
```
iFile_02 <- "CU_18_05_09_01_distritos_adm.csv"
file_data_02 <- paste0(iPath, iFile_02)

if(file.exists(file_data_02)){
    cat("Se leerán datos del archivo: ", file_data_02)
} else{
    warning("Cuidado: el archivo no existe.")
}
```

```
Se leerán datos del archivo:  Data/Input/CU_18_05_09_01_distritos_adm.csv
```

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

[10]:
```
data_02 <- read_csv(file_data_02)
```

```
Rows: 246 Columns: 49
  Column specification

Delimiter: ","
chr  (2): cmun, cdis
dbl (47): ayuntamientos_consejerias_ministerios_etc,
```

```
sanidad_y_servicios_soc…

  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:

[11]:
```
glimpse(data_02)
```

```
Rows: 246
Columns: 49
$ cmun
<chr> "001", "002"…
$ cdis
<chr> "01", "01", …
$ ayuntamientos_consejerias_ministerios_etc
<dbl> 1, 1, 1, 1, …
$ sanidad_y_servicios_sociales
<dbl> 1, 8, 0, 10,…
$ actividades_administrativas
<dbl> 0, 13, 0, 2,…
$ actividades_asociativas_y_hogares
<dbl> 0, 2, 0, 2, …
$ actividades_profesionales
<dbl> 0, 11, 1, 3,…
$ administraciones_publicas
<dbl> 0, 2, 2, 2, …
$ alimentacion
<dbl> 0, 2, 0, 0, …
$ centros_educativos_no_universitarios_centros_publicos
<dbl> 0, 3, 0, 4, …
$ comercio_mayorista
<dbl> 0, 34, 0, 3,…
$ comercio_minorista
<dbl> 0, 22, 0, 10…
$ construccion
<dbl> 0, 22, 0, 16…
$ educacion
<dbl> 0, 4, 0, 6, …
$ hosteleria
<dbl> 0, 2, 4, 12,…
$ industria_quimica_y_refino
<dbl> 0, 3, 0, 0, …
$ informacion_y_comunicaciones
<dbl> 0, 1, 0, 0, …
$ maquinaria_industrial
```

```
<dbl> 0, 6, 0, 0, …
$ material_de_transporte
<dbl> 0, 4, 0, 0, …
$ material_electrico_y_electronico
<dbl> 0, 3, 0, 0, …
$ mercadillos
<dbl> 0, 1, 0, 1, …
$ metalicas_basicas_e_intermedias
<dbl> 0, 25, 0, 1,…
$ mineria_y_suministros
<dbl> 0, 1, 0, 0, …
$ otras_manufactureras
<dbl> 0, 14, 0, 2,…
$ otros_servicios_bancos
<dbl> 0, 3, 0, 2, …
$ papel_y_graficas
<dbl> 0, 14, 0, 0,…
$ servicios_financieros
<dbl> 0, 3, 0, 4, …
$ servicios_personales
<dbl> 0, 3, 0, 5, …
$ servicios_recreativos
<dbl> 0, 5, 0, 4, …
$ textil_confeccion_y_calzado
<dbl> 0, 3, 0, 0, …
$ transporte_y_almacenamiento
<dbl> 0, 8, 0, 4, …
$ actividades_inmobiliarias
<dbl> 0, 0, 0, 1, …
$ centros_educativos_no_universitarios_centros_privados
<dbl> 0, 0, 0, 2, …
$ industria_no_metalica
<dbl> 0, 0, 0, 1, …
$ administracion_de_justicia
<dbl> 0, 0, 0, 0, …
$ agencia_tributaria
<dbl> 0, 0, 0, 0, …
$ campus_universitarios
<dbl> 0, 0, 0, 0, …
$ centros_educativos_universitarios
<dbl> 0, 0, 0, 0, …
$ colegios_mayores
<dbl> 0, 0, 0, 0, …
$ galerias_de_alimentacion
<dbl> 0, 0, 0, 0, …
$ grandes_superficies_especializadas
<dbl> 0, 0, 0, 0, …
$ hipermercados
```

```
<dbl> 0, 0, 0, 0, …
$ mercados_de_abastos
<dbl> 0, 0, 0, 0, …
$ agricultura_y_ganaderia
<dbl> 0, 0, 0, 0, …
$ centros_educativos_no_universitarios_servicios_educativos
<dbl> 0, 0, 0, 0, …
$ seguridad_social
<dbl> 0, 0, 0, 0, …
$ centros_comerciales
<dbl> 0, 0, 0, 0, …
$ oficinas_de_empleo
<dbl> 0, 0, 0, 0, …
$ embajadas_y_consulados
<dbl> 0, 0, 0, 0, …
```

Muestra de datos:

[12]:
```
data_02 |> slice_head(n = 5)
```

A spec_tbl_df: 5 x 49

cmun <chr>	cdis <chr>	ayuntamientos_consejerias_ministerios_etc <dbl>	sanidad_y_servicios_so <dbl>
001	01	1	1
002	01	1	8
003	01	1	0
004	01	1	10
005	01	1	42

3. Indicadores

OPCION A: Seleccionar fichero en ventana para mayor comodidad

Data load using the {tcltk} package. Ucomment the line if not using this option

[13]:
```
# file_data_03 <- tcltk::tk_choose.files(multi = FALSE)
```

OPCION B: Especificar el nombre de archivo

[14]:
```
iFile_03 <- "CU_18_05_10_indicadores_distritos.csv"
file_data_03 <- paste0(iPath, iFile_03)

if(file.exists(file_data_03)){
    cat("Se leerán datos del archivo: ", file_data_03)
} else{
    warning("Cuidado: el archivo no existe.")
}
```

```
Se leerán datos del archivo:  Data/Input/CU_18_05_10_indicadores_distritos.csv
```

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

[15]:
```
data_03 <- read_csv(file_data_03)
```

```
Rows: 246 Columns: 22
  Column specification

Delimiter: ","
chr  (2): CMUN, dist
dbl (20): nsec, t3_1, t1_1, t2_1, t2_2, t4_1, t4_2, t4_3, t5_1, t6_1,
t7_1, …

 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:

[16]:
```
glimpse(data_03)
```

```
Rows: 246
Columns: 22
$ CMUN            <chr> "001", "002", "003", "004", "005",
"005", "005", "005"…
$ dist            <chr> "01", "01", "01", "01", "01", "02",
"03", "04", "05", …
$ nsec            <dbl> 1, 2, 1, 4, 23, 36, 16, 18, 32, 67,
19, 37, 29, 35, 1,…
$ t3_1            <dbl> NA, 39.40747, 47.45970, 41.46662,
45.99877, 42.75131, …
$ t1_1            <dbl> 55, 4793, 248, 9946, 31006, 54049,
28117, 38778, 43620…
$ t2_1            <dbl> NA, 0.4777851, 0.3952000, 0.5015109,
0.5298303, 0.5115…
$ t2_2            <dbl> NA, 0.5222149, 0.6048000, 0.4984891,
0.4701697, 0.4884…
$ t4_1            <dbl> NA, 0.1696289, 0.1371000, 0.1810684,
0.1141978, 0.1468…
$ t4_2            <dbl> NA, 0.7218958, 0.6573000, 0.6420633,
0.6427127, 0.6644…
$ t4_3            <dbl> NA, 0.10847530, 0.20560000,
0.17684664, 0.24307467, 0.…
$ t5_1            <dbl> NA, 0.15562992, 0.11290000,
0.15474242, 0.19647849, 0.…
$ t6_1            <dbl> NA, 0.1959363, 0.1411000, 0.1912543,
0.2400069, 0.2147…
$ t7_1            <dbl> NA, 0.41318314, 0.15420000,
0.15732451, 0.07438075, 0.…
$ t8_1            <dbl> NA, 0.40995322, 0.14490000,
```

```
0.15046840, 0.06545227, 0.…
$ t9_1            <dbl> NA, 0.4407990, 0.5280000, 0.2942034,
0.3850913, 0.2235…
$ t10_1           <dbl> NA, 0.10715222, 0.09320000,
0.16675872, 0.14129129, 0.…
$ t11_1           <dbl> NA, 0.6008132, 0.5000000, 0.4777910,
0.4565679, 0.4588…
$ t12_1           <dbl> NA, 0.6730932, 0.5514000, 0.5729139,
0.5316057, 0.5641…
$ X               <dbl> -3.635710, -3.480171, -3.849961,
-3.989259, -3.364638,…
$ Y               <dbl> 41.09315, 40.52396, 40.92033,
40.23240, 40.48268, 40.4…
$ densidad_hab_km <dbl> 2.514002, 242.602224, 9.647117,
451.618447, 17549.3832…
$ area_km2        <dbl> 21.877471, 19.756620, 25.707163,
22.023015, 1.766786, …
```

Muestra de datos:

[17]:
```
data_03 |> slice_head(n = 5)
```

A spec_tbl_df: 5 x 22

CMUN <chr>	dist <chr>	nsec <dbl>	t3_1 <dbl>	t1_1 <dbl>	t2_1 <dbl>	t2_2 <dbl>	t4_1 <dbl>	t4_2 <db
001	01	1	NA	55	NA	NA	NA	NA
002	01	2	39.40747	4793	0.4777851	0.5222149	0.1696289	0.72
003	01	1	47.45970	248	0.3952000	0.6048000	0.1371000	0.65
004	01	4	41.46662	9946	0.5015109	0.4984891	0.1810684	0.64
005	01	23	45.99877	31006	0.5298303	0.4701697	0.1141978	0.64

4. POIS

OPCION A: Seleccionar fichero en ventana para mayor comodidad

Data load using the {tcltk} package. Ucomment the line if not using this option

[18]:
```
# file_data_04 <- tcltk::tk_choose.files(multi = FALSE)
```

OPCION B: Especificar el nombre de archivo

[19]:
```
iFile_04 <- "CU_18_05_13_01_ditrito_pois.csv"
file_data_04 <- paste0(iPath, iFile_04)

if(file.exists(file_data_04)){
    cat("Se leerán datos del archivo: ", file_data_04)
} else{
    warning("Cuidado: el archivo no existe.")
}
```

```
Se leerán datos del archivo:  Data/Input/CU_18_05_13_01_ditrito_pois.csv
```

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

[20]:
```
data_04 <- read_csv(file_data_04)
```

```
Rows: 224 Columns: 60
  Column specification

Delimiter: ","
chr  (2): cmun, cdis
dbl (58): bakery, bar, butcher, cafe, car_dealership, clothes,
convenience, …

 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:

[21]:
```
glimpse(data_04)
```

```
Rows: 224
Columns: 60
$ cmun              <chr> "002", "003", "004", "005", "005",
"005", "005", "00…
$ cdis              <chr> "01", "01", "01", "01", "02", "03",
"04", "05", "01"…
$ bakery            <dbl> 2, 0, 0, 4, 5, 1, 2, 7, 4, 1, 1, 0,
2, 0, 0, 1, 0, 0…
$ bar               <dbl> 2, 0, 1, 22, 10, 2, 3, 11, 7, 1, 1,
2, 6, 0, 2, 17, …
$ butcher           <dbl> 1, 0, 0, 2, 2, 0, 1, 2, 0, 0, 1, 0,
0, 0, 0, 2, 0, 0…
$ cafe              <dbl> 5, 1, 0, 12, 6, 1, 3, 19, 12, 3, 2,
0, 6, 0, 1, 4, 0…
$ car_dealership    <dbl> 1, 0, 0, 1, 1, 3, 0, 3, 5, 1, 0, 0,
0, 0, 0, 0, 0, 0…
$ clothes           <dbl> 1, 0, 0, 13, 1, 0, 1, 28, 0, 0, 0,
0, 0, 0, 1, 1, 0,…
$ convenience       <dbl> 2, 0, 0, 4, 2, 2, 2, 13, 8, 1, 1,
2, 6, 0, 0, 1, 0, …
$ fast_food         <dbl> 1, 0, 0, 8, 1, 2, 5, 11, 10, 7, 4,
0, 6, 0, 2, 5, 0,…
$ food_court        <dbl> 2, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ furniture_shop    <dbl> 2, 0, 0, 3, 0, 0, 1, 8, 2, 1, 0, 0,
0, 0, 0, 0, 0, 0…
$ gift_shop         <dbl> 1, 0, 0, 2, 0, 0, 0, 1, 0, 0, 0, 0,
```

```
0, 0, 0, 1, 0, 0…
$ greengrocer       <dbl> 1, 0, 0, 1, 1, 0, 0, 8, 2, 0, 1, 0,
1, 0, 0, 0, 0, 0…
$ guesthouse        <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 0, 0, 0, 0…
$ hairdresser       <dbl> 2, 0, 0, 16, 4, 0, 4, 21, 7, 1, 0,
1, 5, 0, 0, 1, 0,…
$ hotel             <dbl> 1, 1, 0, 3, 0, 3, 0, 1, 3, 3, 0, 0,
0, 0, 2, 1, 0, 1…
$ pub               <dbl> 2, 0, 0, 24, 4, 2, 2, 31, 7, 0, 0,
0, 4, 0, 0, 2, 1,…
$ restaurant        <dbl> 5, 4, 3, 53, 5, 18, 10, 22, 73, 26,
13, 1, 4, 2, 2, …
$ stationery        <dbl> 1, 0, 0, 1, 3, 0, 1, 2, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1…
$ vending_any       <dbl> 1, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ bicycle_rental    <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 1, 0…
$ hostel            <dbl> 0, 2, 0, 0, 2, 0, 0, 0, 0, 0, 1, 0,
0, 0, 0, 1, 0, 0…
$ supermarket       <dbl> 0, 1, 1, 6, 6, 2, 3, 10, 8, 12, 3,
4, 3, 1, 3, 3, 0,…
$ beauty_shop       <dbl> 0, 0, 0, 3, 0, 0, 2, 10, 6, 0, 0,
0, 0, 0, 1, 0, 0, …
$ beverages         <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ bicycle_shop      <dbl> 0, 0, 0, 3, 1, 1, 0, 2, 4, 0, 1, 1,
1, 0, 0, 2, 0, 0…
$ bookshop          <dbl> 0, 0, 0, 7, 0, 0, 2, 3, 3, 1, 0, 0,
0, 0, 0, 0, 0, 1…
$ chemist           <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ computer_shop     <dbl> 0, 0, 0, 3, 3, 0, 0, 2, 0, 0, 0, 0,
0, 0, 1, 0, 0, 0…
$ doityourself      <dbl> 0, 0, 0, 2, 2, 2, 0, 3, 3, 0, 0, 0,
0, 0, 0, 1, 0, 0…
$ florist           <dbl> 0, 0, 0, 2, 0, 0, 2, 2, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ jeweller          <dbl> 0, 0, 0, 2, 1, 0, 0, 3, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ kiosk             <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ mall              <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ mobile_phone_shop <dbl> 0, 0, 0, 2, 1, 0, 0, 7, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ optician          <dbl> 0, 0, 0, 4, 1, 0, 2, 6, 2, 1, 1, 0,
```

```
0, 0, 0, 0, 0, 2…
$ shoe_shop         <dbl> 0, 0, 0, 10, 1, 0, 0, 7, 3, 0, 0,
0, 1, 0, 0, 1, 0, …
$ toy_shop          <dbl> 0, 0, 0, 1, 1, 0, 0, 4, 3, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ travel_agent      <dbl> 0, 0, 0, 3, 0, 0, 0, 1, 1, 1, 0, 0,
0, 0, 0, 0, 0, 0…
$ car_rental        <dbl> 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0,
0, 0, 1, 0, 0, 0…
$ car_wash          <dbl> 0, 0, 0, 0, 1, 0, 0, 2, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ outdoor_shop      <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ sports_shop       <dbl> 0, 0, 0, 0, 0, 1, 1, 4, 1, 2, 0, 0,
0, 0, 0, 0, 0, 0…
$ caravan_site      <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ biergarten        <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ garden_centre     <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ laundry           <dbl> 0, 0, 0, 0, 0, 0, 0, 2, 4, 0, 0, 0,
0, 0, 1, 0, 0, 0…
$ department_store  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, 0, 0, 1, 0, 0…
$ newsagent         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0,
0, 0, 0, 0, 0, 1…
$ vending_parking   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ chalet            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ shelter           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ motel             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ camp_site         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ general           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ alpine_hut        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ video_shop        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ vending_machine   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
$ car_sharing       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0…
```

Muestra de datos:

[22]:
```
data_04 |> slice_head(n = 5)
```

A spec_tbl_df: 5 x 60

cmun <chr>	cdis <chr>	bakery <dbl>	bar <dbl>	butcher <dbl>	cafe <dbl>	car_dealership <dbl>	clothes <dbl>	conven <dbl>
002	01	2	2	1	5	1	1	2
003	01	0	0	0	1	0	0	0
004	01	0	1	0	0	0	0	0
005	01	4	22	2	12	1	13	4
005	02	5	10	2	6	1	1	2

5. AEMET

OPCION A: Seleccionar fichero en ventana para mayor comodidad

Data load using the {tcltk} package. Ucomment the line if not using this option

[23]:
```
# file_data_05 <- tcltk::tk_choose.files(multi = FALSE)
```

OPCION B: Especificar el nombre de archivo

[24]:
```
iFile_05 <- "CU_18_05_15_distritos_meteo.csv"
file_data_05 <- paste0(iPath, iFile_05)

if(file.exists(file_data_05)){
    cat("Se leerán datos del archivo: ", file_data_05)
} else{
    warning("Cuidado: el archivo no existe.")
}
```

```
Se leerán datos del archivo:  Data/Input/CU_18_05_15_distritos_meteo.csv
```

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

[25]:
```
data_05 <- read_csv(file_data_05)
```

```
Rows: 246 Columns: 6
  Column specification

Delimiter: ","
chr (2): CMUN, CDIS
dbl (4): tmed, prec, velmedia, presMax

 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:

[26]:
```
glimpse(data_05)
```

```
Rows: 246
Columns: 6
$ CMUN     <chr> "001", "002", "003", "004", "005", "005",
"005", "005", "005"…
$ CDIS     <chr> "01", "01", "01", "01", "01", "02", "03",
"04", "05", "01", "…
$ tmed     <dbl> 11.47600, 15.31806, 12.00935, 15.13839,
15.27439, 15.31156, 1…
$ prec     <dbl> 3.6992270, 0.8713863, 2.0387942, 1.0548851,
0.8788330, 0.8511…
$ velmedia <dbl> 2.090249, 3.352072, 2.986153, 2.919731,
3.331738, 3.372923, 3…
$ presMax  <dbl> 904.2719, 946.6977, 880.7764, 931.9199,
945.8412, 946.8914, 9…
```

Muestra de datos:

[27]:
```
data_05 |> slice_head(n = 5)
```

A spec_tbl_df: 5 x 6

CMUN <chr>	CDIS <chr>	tmed <dbl>	prec <dbl>	velmedia <dbl>	presMax <dbl>
001	01	11.47600	3.6992270	2.090249	904.2719
002	01	15.31806	0.8713863	3.352072	946.6977
003	01	12.00935	2.0387942	2.986153	880.7764
004	01	15.13839	1.0548851	2.919731	931.9199
005	01	15.27439	0.8788330	3.331738	945.8412

0.3 ETL Processes

0.3.1 Import data from: CSV, Excel, Tab, JSON, SQL, and Parquet files

Se han importado en el apartado Data Load anterior:

- Infraestructuras (transporte y sanidad)
- Administración y servicios
- POIS (hostelería, comercio, turismo)
- Indicadores INE
- Datos meteorológicos

Incluir apartados si procede para: Extracción de datos (select, filter), Transformación de datos, (mutate, joins, …). Si es necesario tratar datos perdidos, indicarlo también en NB 09.2

Data transformation

- Nombres de columna para unión consistentes

[28]:
```
tdata_03 <- data_03 |>
    rename(cmun = CMUN,
           cdis = dist)
```

```
tdata_05 <- data_05 |>
    rename(cmun = CMUN,
           cdis = CDIS)
```

- Unión de todas las tablas

Si no aplica: Estos datos no requieren tareas de este tipo.

[29]:
```
data <- data_01 |>
    full_join(data_02, by = c("cmun", "cdis")) |>
    full_join(tdata_03, by = c("cmun", "cdis")) |>
    full_join(data_04, by = c("cmun", "cdis")) |>
    full_join(tdata_05, by = c("cmun", "cdis"))
```

[30]:
```
data |> glimpse()
```

```
Rows: 247
Columns: 143
$ cmun
<chr> "001", "002"…
$ cdis
<chr> "01", "01", …
$ consultorios_de_salud
<dbl> 1, 1, 1, 1, …
$ helisuperficies
<dbl> 1, 0, 1, 0, …
$ centros_de_atencion_a_drogodependientes
<dbl> 0, 0, 0, 0, …
$ centros_de_salud
<dbl> 0, 0, 0, 0, …
$ estaciones_de_cercanias
<dbl> 0, 0, 0, 0, …
$ hospitales
<dbl> 0, 0, 0, 0, …
$ otros_centros_de_salud
<dbl> 0, 0, 0, 0, …
$ centros_de_salud_mental
<dbl> 0, 0, 0, 0, …
$ centros_de_especialidades
<dbl> 0, 0, 0, 0, …
$ bocas_de_metro
<dbl> 0, 0, 0, 0, …
$ intercambiadores
<dbl> 0, 0, 0, 0, …
$ aeropuertos
<dbl> 0, 0, 0, 0, …
$ ayuntamientos_consejerias_ministerios_etc
<dbl> 1, 1, 1, 1, …
```

```
$ sanidad_y_servicios_sociales
<dbl> 1, 8, 0, 10,…
$ actividades_administrativas
<dbl> 0, 13, 0, 2,…
$ actividades_asociativas_y_hogares
<dbl> 0, 2, 0, 2, …
$ actividades_profesionales
<dbl> 0, 11, 1, 3,…
$ administraciones_publicas
<dbl> 0, 2, 2, 2, …
$ alimentacion
<dbl> 0, 2, 0, 0, …
$ centros_educativos_no_universitarios_centros_publicos
<dbl> 0, 3, 0, 4, …
$ comercio_mayorista
<dbl> 0, 34, 0, 3,…
$ comercio_minorista
<dbl> 0, 22, 0, 10…
$ construccion
<dbl> 0, 22, 0, 16…
$ educacion
<dbl> 0, 4, 0, 6, …
$ hosteleria
<dbl> 0, 2, 4, 12,…
$ industria_quimica_y_refino
<dbl> 0, 3, 0, 0, …
$ informacion_y_comunicaciones
<dbl> 0, 1, 0, 0, …
$ maquinaria_industrial
<dbl> 0, 6, 0, 0, …
$ material_de_transporte
<dbl> 0, 4, 0, 0, …
$ material_electrico_y_electronico
<dbl> 0, 3, 0, 0, …
$ mercadillos
<dbl> 0, 1, 0, 1, …
$ metalicas_basicas_e_intermedias
<dbl> 0, 25, 0, 1,…
$ mineria_y_suministros
<dbl> 0, 1, 0, 0, …
$ otras_manufactureras
<dbl> 0, 14, 0, 2,…
$ otros_servicios_bancos
<dbl> 0, 3, 0, 2, …
$ papel_y_graficas
<dbl> 0, 14, 0, 0,…
$ servicios_financieros
<dbl> 0, 3, 0, 4, …
```

```
$ servicios_personales
<dbl> 0, 3, 0, 5, …
$ servicios_recreativos
<dbl> 0, 5, 0, 4, …
$ textil_confeccion_y_calzado
<dbl> 0, 3, 0, 0, …
$ transporte_y_almacenamiento
<dbl> 0, 8, 0, 4, …
$ actividades_inmobiliarias
<dbl> 0, 0, 0, 1, …
$ centros_educativos_no_universitarios_centros_privados
<dbl> 0, 0, 0, 2, …
$ industria_no_metalica
<dbl> 0, 0, 0, 1, …
$ administracion_de_justicia
<dbl> 0, 0, 0, 0, …
$ agencia_tributaria
<dbl> 0, 0, 0, 0, …
$ campus_universitarios
<dbl> 0, 0, 0, 0, …
$ centros_educativos_universitarios
<dbl> 0, 0, 0, 0, …
$ colegios_mayores
<dbl> 0, 0, 0, 0, …
$ galerias_de_alimentacion
<dbl> 0, 0, 0, 0, …
$ grandes_superficies_especializadas
<dbl> 0, 0, 0, 0, …
$ hipermercados
<dbl> 0, 0, 0, 0, …
$ mercados_de_abastos
<dbl> 0, 0, 0, 0, …
$ agricultura_y_ganaderia
<dbl> 0, 0, 0, 0, …
$ centros_educativos_no_universitarios_servicios_educativos
<dbl> 0, 0, 0, 0, …
$ seguridad_social
<dbl> 0, 0, 0, 0, …
$ centros_comerciales
<dbl> 0, 0, 0, 0, …
$ oficinas_de_empleo
<dbl> 0, 0, 0, 0, …
$ embajadas_y_consulados
<dbl> 0, 0, 0, 0, …
$ nsec
<dbl> 1, 2, 1, 4, …
$ t3_1
<dbl> NA, 39.40747…
```

```
$ t1_1
<dbl> 55, 4793, 24…
$ t2_1
<dbl> NA, 0.477785…
$ t2_2
<dbl> NA, 0.522214…
$ t4_1
<dbl> NA, 0.169628…
$ t4_2
<dbl> NA, 0.721895…
$ t4_3
<dbl> NA, 0.108475…
$ t5_1
<dbl> NA, 0.155629…
$ t6_1
<dbl> NA, 0.195936…
$ t7_1
<dbl> NA, 0.413183…
$ t8_1
<dbl> NA, 0.409953…
$ t9_1
<dbl> NA, 0.440799…
$ t10_1
<dbl> NA, 0.107152…
$ t11_1
<dbl> NA, 0.600813…
$ t12_1
<dbl> NA, 0.673093…
$ X
<dbl> -3.635710, -…
$ Y
<dbl> 41.09315, 40…
$ densidad_hab_km
<dbl> 2.514002, 24…
$ area_km2
<dbl> 21.877471, 1…
$ bakery
<dbl> NA, 2, 0, 0,…
$ bar
<dbl> NA, 2, 0, 1,…
$ butcher
<dbl> NA, 1, 0, 0,…
$ cafe
<dbl> NA, 5, 1, 0,…
$ car_dealership
<dbl> NA, 1, 0, 0,…
$ clothes
<dbl> NA, 1, 0, 0,…
```

```
$ convenience
<dbl> NA, 2, 0, 0,…
$ fast_food
<dbl> NA, 1, 0, 0,…
$ food_court
<dbl> NA, 2, 0, 0,…
$ furniture_shop
<dbl> NA, 2, 0, 0,…
$ gift_shop
<dbl> NA, 1, 0, 0,…
$ greengrocer
<dbl> NA, 1, 0, 0,…
$ guesthouse
<dbl> NA, 1, 0, 0,…
$ hairdresser
<dbl> NA, 2, 0, 0,…
$ hotel
<dbl> NA, 1, 1, 0,…
$ pub
<dbl> NA, 2, 0, 0,…
$ restaurant
<dbl> NA, 5, 4, 3,…
$ stationery
<dbl> NA, 1, 0, 0,…
$ vending_any
<dbl> NA, 1, 0, 0,…
$ bicycle_rental
<dbl> NA, 0, 1, 0,…
$ hostel
<dbl> NA, 0, 2, 0,…
$ supermarket
<dbl> NA, 0, 1, 1,…
$ beauty_shop
<dbl> NA, 0, 0, 0,…
$ beverages
<dbl> NA, 0, 0, 0,…
$ bicycle_shop
<dbl> NA, 0, 0, 0,…
$ bookshop
<dbl> NA, 0, 0, 0,…
$ chemist
<dbl> NA, 0, 0, 0,…
$ computer_shop
<dbl> NA, 0, 0, 0,…
$ doityourself
<dbl> NA, 0, 0, 0,…
$ florist
<dbl> NA, 0, 0, 0,…
```

```
$ jeweller
<dbl> NA, 0, 0, 0,…
$ kiosk
<dbl> NA, 0, 0, 0,…
$ mall
<dbl> NA, 0, 0, 0,…
$ mobile_phone_shop
<dbl> NA, 0, 0, 0,…
$ optician
<dbl> NA, 0, 0, 0,…
$ shoe_shop
<dbl> NA, 0, 0, 0,…
$ toy_shop
<dbl> NA, 0, 0, 0,…
$ travel_agent
<dbl> NA, 0, 0, 0,…
$ car_rental
<dbl> NA, 0, 0, 0,…
$ car_wash
<dbl> NA, 0, 0, 0,…
$ outdoor_shop
<dbl> NA, 0, 0, 0,…
$ sports_shop
<dbl> NA, 0, 0, 0,…
$ caravan_site
<dbl> NA, 0, 0, 0,…
$ biergarten
<dbl> NA, 0, 0, 0,…
$ garden_centre
<dbl> NA, 0, 0, 0,…
$ laundry
<dbl> NA, 0, 0, 0,…
$ department_store
<dbl> NA, 0, 0, 0,…
$ newsagent
<dbl> NA, 0, 0, 0,…
$ vending_parking
<dbl> NA, 0, 0, 0,…
$ chalet
<dbl> NA, 0, 0, 0,…
$ shelter
<dbl> NA, 0, 0, 0,…
$ motel
<dbl> NA, 0, 0, 0,…
$ camp_site
<dbl> NA, 0, 0, 0,…
$ general
<dbl> NA, 0, 0, 0,…
```

```
$ alpine_hut
<dbl> NA, 0, 0, 0,…
$ video_shop
<dbl> NA, 0, 0, 0,…
$ vending_machine
<dbl> NA, 0, 0, 0,…
$ car_sharing
<dbl> NA, 0, 0, 0,…
$ tmed
<dbl> 11.47600, 15…
$ prec
<dbl> 3.6992270, 0…
$ velmedia
<dbl> 2.090249, 3.…
$ presMax
<dbl> 904.2719, 94…
```

[31]:
```
data |> slice_head(n = 5)
```

A spec_tbl_df: 5 x 143

cmun <chr>	cdis <chr>	consultorios_de_salud <dbl>	helisuperficies <dbl>	centros_de_atencion_a_dro
001	01	1	1	0
002	01	1	0	0
003	01	1	1	0
004	01	1	0	0
005	01	0	0	1

0.4 Synthetic Data Generation

No aplica

0.5 Fake Data Generation

No aplica

0.6 Open Data

Los datos fueron descargados de datos abiertos (INE, CM, OSM, AEMET)

0.7 Data Save

Este proceso, puede copiarse y repetirse en aquellas partes del notebbok que necesiten guardar datos. Recuerde cambiar las cadenas añadida del fichero para diferenciarlas

Identificamos los datos a guardar

[32]:
```
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 "_05".
- Número de la tarea que lo genera, por ejemplo "_01"
- En caso de generarse varios ficheros en la misma tarea, llevarán _01 _02 … después
- Nombre: identificativo de "properData", por ejemplo "_zonasgeo"
- Extensión del archivo

Ejemplo: "CU_04_05_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)

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

0.7.1 Proceso 05

[33]:
```
caso <- "CU_18"
proceso <- '_05'
tarea <- "_16"
archivo <- ""
proper <- "_distritos_variables"
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

[34]:
```
# file_save <- paste0(caso, proceso, tarea, tcltk::tkgetSaveFile(), proper,␣
 ↪extension)
# path_out <- paste0(oPath, file_save)
# write_csv(data_to_save, 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/.

[35]:
```
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_18_05_16_distritos_variables.csv'

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

[36]:
```
path_in <- paste0(iPath, file_save)
file.copy(path_out, path_in, overwrite = TRUE)
```

TRUE

0.8 Main Conclusions

List and describe the general conclusions of the analysis carried out.

0.8.1 Prerequisites

This working code needs the following conditions:

- For using the interactive selection of file, the {tcltk} package must be installed. It is not needed in production.
- The {readr} and {dplyr} packages must be installed.
- The data paths `Data/Input` and `Data/Output` must exist (relative to the notebook path)

0.8.2 Configuration Management

This notebook has been tested with the following versions of R and packages. It cannot be assured that later versions work in the same way: * R 4.2.2 * tcltk 4.2.2 * readr 2.1.3 * dplyr 1.0.10

0.8.3 Data structures

Objeto `data`

- Hay 248 filas con 143 columnas (ver tareas anteriores para detalle)

Observaciones generales sobre los datos

- Hay valores perdidos ya que algunos de los distritos no tienen datos en todos los archivos de datos

0.8.4 Consideraciones para despliegue en piloto

- Ninguna

0.8.5 Consideraciones para despliegue en producción

- Se deben crear los procesos ETL en producción necesarios para que los datos de entrada estén actualizados

0.9 Main Actions

Acciones done Indicate the actions that have been carried out in this process

- Se han unido las tablas con información por distritos

Acctions to perform Indicate the actions that must be carried out in subsequent processes

- Se debe decidir si imputar valores perdidos u omitirlos en cada modelo

0.10 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

[37]:
```
# incluir código
```