05. - Data Collection_CU_53_04_inversiones_raw_v_01

June 13, 2023

#

CU53_impacto de las políticas de inversión en sanidad, infraestructuras y promoción turística en el SPI

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 04. Obtener datos de inversiones de la Comunidad de Madrid

- Obtener datos de inversiones a partir de los publicados de ejecución de presupuestos
- Están publicados en el siguiente enlace: http://www.madrid.org/presupuestos/index.php/ejecucion-presupuestaria

NOTA: se descargan directamente a la carpeta `Input`

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

ELIMINAR O AÑADIR LO QUE TOQUE. COPIAR VERSIONES AL FINAL Y QUITAR CÓDIGO DE VERSIONES

- {tcltk} para selección interactiva de archivos locales
- {sf} para trabajar con georeferenciación
- {readr} para leer y escribir archivos csv
- {dplyr} para explorar datos
- {stringr} para manipulación de cadenas de caracteres
- {tidyr} para organización de datos

[ ]:
```
# library(sf)
# library(readr)
# library(dplyr)
# library(stringr)
# library(tidyr)

p <- c("tcltk", "sf", "readr", "dplyr", "stringr", "tidyr")
```

0.1.3 Paths

[5]:
```
iPath <- "Data/Input/"
oPath <- "Data/Output/"

base_url <- "http://www.madrid.org/presupuestos/attachments/category/119/"
```

0.2 Data Load

No aplica

0.3 ETL Processes

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

No aplica

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

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

0.4 Synthetic Data Generation

No aplica

0.5 Fake Data Generation

No aplica

0.6 Open Data

Los datos están publicados en abierto, aunque no hay API ni nada parecido, es descarga directa

- Descarga de los ficheros disponibles
- En 2017 el nombre de archivo no está el mes en dos cifras

[ ]:
```
primer_ano <- 2017
ultimo_ano <- 2022
ultimo_mes <- 11
```

[ ]:
```
anos <- primer_ano:ultimo_ano

for(a in anos){
  if(a == ultimo_ano){
    meses <- 1:ultimo_mes
  } else{
    meses <- 1:11
  }
  for (m in meses){
    mm <- ifelse(a == 2017, m, sprintf("%02.0f", m))
    f <- paste0("G-CAPITULO-M",
                mm, "-", a, ".xls")
    url <- paste0(base_url, f)
    download.file(url, destfile = paste0(iPath, "xls/EJECUCION_", a, "-",
                                         sprintf("%02.0f", m), ".xls"))
  }
}
```

```
Warning message in download.file(url, destfile = paste0(iPath, "EJECUCION_", a,
:
"downloaded length 0 != reported length 259"
Warning message in download.file(url, destfile = paste0(iPath, "EJECUCION_", a,
:
"cannot open URL 'http://www.madrid.org/presupuestos/attachments/category/119/G-
CAPITULO-M12-2022.xls': HTTP status was '404 Not Found'"
```

```
Error in download.file(url, destfile = paste0(iPath, "EJECUCION_", a, : no fue␣
 ↪posible abrir la URL 'http://www.madrid.org/presupuestos/attachments/category
 ↪119/G-CAPITULO-M12-2022.xls'

Traceback:
```

```
1. download.file(url, destfile = paste0(iPath, "EJECUCION_", a,

 .     "-", sprintf("%02.0f", m), ".xls"))
```

0.7 Data Save

No aplica: se han guardado directamente sin transformación

0.8 Main Conclusions

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

0.8.1 Prerequisites

Para que funcione este código se necesita:

- Las rutas de archivos `Data/Input` y `Data/Output` deben existir (relativas a la ruta del *notebook*)
- No se requiere ningún paquete específico ya que se usan solo funciones de R-base

0.8.2 Configuration Management

This notebook has been tested with the following versions of R. It cannot be assured that later versions work in the same way: * R 4.2.2

0.8.3 Data structures

- Son archivos Excel que se consolidan en el siguiente Notebook

Observaciones generales sobre los datos

- Los archivos están disponibles con los siguientes nombres:
  - EJECUCION_YYYY-MM.xls
- Hay un archivo por mes y año disponibles.
- Disponibles desde 2017. Diciembre solo algunos años provisional, por lo que para que sean homogéneos se toman solo hasta noviembre

0.8.4 Consideraciones para despliegue en piloto

- No aplica

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

- Si se dispusiera de datos de SPI e inversiones a nivel municipal, se podría afinar muchísimo más

0.9 Main Actions

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

- Se han descargado los archivos de ejecución presupuestaria de la Comunidad de Madrid

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

- Se deben consolidar los archivos Excel en un único csv

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

[ ]:
```
# incluir código
```