
**EN ESTE MÓDULO VAS A NECESITAR…**

**Software:**

- Jupyter Notebook - https://jupyter.org/install

**Datos:**

- Módulo 3: http://geonode.mygeoproject.eu/documents/86

**EN ESTE MÓDULO TIENES QUE…**

- Sube una de las databases resultado de la limpieza de datos a GEONODE
- Módulo 3 Exam (5 preguntas, 4 intentos).

## Lectura de datos

### i. Descripción de los datos a manejar

Vamos a practicar la descarga de conjuntos de datos en un *Jupiter Notebook* con la librería *Pandas* de un archivo Excel con información acerca de producción de granjas en Europa.

Archivo local: ***datos/animalEurostatNuts2.xlsx***

### ii. Lectura de archivos de datos

PARA APRENDER MÁS….

Los formatos de datos soportados por las librerias Pandas y como leerlos se describen en: https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html

| Format | Read | Write |
|---|---|---|
| csv | pd.read_csv() | pac.to_csv() |
| json | pd.read_json() | pac.to_json() |
| excel | pd.read_excel() | pac.to_excel() |
| hdf | pd.read_hdf() | pac.to_hdf() |
| sql | pd.read_sql() | pac.to_sql() |
| … | … | … |

Cada operación de lectura de un archivo diferente tiene parámetros distintos para ajustar el proceso.


### a. Lectura de archivos de Excel

Usamos **pandas.read_excel()** para leer el archive Excel y alamacenar los datos en un marco de datos (*dataframe*).

Los parámetros más relevantes son:

*io = la dirección (directorio) del archivo*
*sheet_name = la hoja que se va a leer*

PARA APRENDER MÁS….

Hay muchos parámetros para tener en cuenta

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_excel.html

La carga de la información se puede hacer desde un archivo local o desde una archivo de una web.

**Input [1]:**

```
import pandas as pd
import numpy as np
file = 'datos/animalEurostatNuts2.xlsx'
data = pd.read_excel(file, sheet_name='Data', )
print("Done")
```

**Output [1]:**

```
Done
```

### b. Visualización básica de los datos

Podemos visualizar las primeras y últimas filas de la table para comprobar que se han cargado correctamente.

**Input [2]:**

```
data.head(10)
```

**Output [2]:**


| | Animal populations by NUTS 2 regions [agr_r_animal] | Unnamed: 1 | Unnamed: 2 | Unnamed: 3 | Unnamed: 4 | Unnamed: 5 | Unnamed: 6 | Unnamed: 7 | Unnamed: 8 | Unnamed: 9 | ... | Unnamed: 49 | Unnamed: 50 | Unnamed 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 1 | Last update | 2020-01-28 10:00:49 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 2 | Extracted on | 2020-02-02 22:09:53.978000 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 3 | Source of data | Eurostat | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 4 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 5 | ANIMALS | Live bovine animals | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 6 | UNIT | Thousand heads (animals) | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 7 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 8 | GEO/TIME | 1991 | Flags and footnotes | 1992 | Flags and footnotes | 1993 | Flags and footnotes | 1994 | Flags and footnotes | 1995 | ... | 2015 | Flags and footnotes | 201 |
| 9 | European Union (EU6-1958, EU9-1973, EU10-1981,... | : | NaN | : | NaN | : | NaN | : | NaN | : | ... | : | NaN | |

10 rows × 59 columns

Input [3]:

```
data.tail(14)
```

Output [3]:

| | Animal populations by NUTS 2 regions [agr_r_animal] | Unnamed: 1 | Unnamed: 2 | Unnamed: 3 | Unnamed: 4 | Unnamed: 5 | Unnamed: 6 | Unnamed: 7 | Unnamed: 8 | Unnamed: 9 | ... | Unnamed: 49 | Unnamed: 50 | Unnamed: 51 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 545 | Bosnia and Herzegovina | : | NaN | : | NaN | : | NaN | : | NaN | : | ... | 455 | e | 455 |
| 546 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 547 | Available flags: | NaN | NaN | NaN | Special value: | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 548 | b | break in time series | NaN | NaN | : | not available | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 549 | c | confidential | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 550 | d | definition differs, see metadata | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 551 | e | estimated | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 552 | f | forecast | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 553 | n | not significant | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 554 | p | provisional | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 555 | r | revised | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 556 | s | Eurostat estimate | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 557 | u | low reliability | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |
| 558 | z | not applicable | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | NaN |

14 rows × 59 columns


### c. Ajustes y parámetros de los procesos de carga

Podemos ver que las primeras 9 filas y las últimas 13 no contienen datos válidos, son filas de metadatos de la tabla.

Estas filas deberían ser ignoradas cuando se lee la tabla. La primera fila para leer debe contener los encabezados de la columna.

Además, la primera columna es descriptiva de la información de cada fila.

**Input [4]:**

```
data = pd.read_excel(file, sheet_name='Data', skiprows=9, skipfooter=13, index_col=0
)
print("Done")
```

**Output [4]:**

```
Done
```

**Input [5]:**

```
data.head(2)
```

**Output [5]:**

| GEO/TIME | 1991 | Flags and footnotes | 1992 | Flags and footnotes.1 | 1993 | Flags and footnotes.2 | 1994 | Flags and footnotes.3 | 1995 | Flags and footnotes.4 | ... | 2015 | Flags and footnotes.24 | 2016 | Flags and footnotes.25 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| European Union (EU6-1958, EU9-1973, EU10-1981, EU12-1986, EU15-1995, EU25-2004, EU27-2007, EU28-2013) | : | NaN | : | NaN | : | NaN | : | NaN | : | NaN | ... | : | NaN | : | NaN |
| Belgium | 3105.5 | NaN | 3099.6 | NaN | 3084.2 | NaN | 3161.1 | NaN | 3158.7 | NaN | ... | 2503.26 | NaN | 2501.35 | NaN |

2 rows × 58 columns

**Input [6]:**

```
data.tail(3)
```

**Output [6]:**


| GEO/TIME | 1991 | Flags and footnotes | 1992 | Flags and footnotes.1 | 1993 | Flags and footnotes.2 | 1994 | Flags and footnotes.3 | 1995 | Flags and footnotes.4 | ... | 2015 | Flags and footnotes.24 | 2016 | Flags and footnotes.25 | 2017 | F foot |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sanliurfa, Diyarbakir | : | NaN | : | NaN | : | NaN | : | NaN | 107.4 | NaN | ... | : | NaN | : | NaN | : | |
| Mardin, Batman, Sirnak, Siirt | : | NaN | : | NaN | : | NaN | : | NaN | 44.3 | NaN | ... | : | NaN | : | NaN | : | |
| Bosnia and Herzegovina | : | NaN | : | NaN | : | NaN | : | NaN | : | NaN | ... | 455 | e | 455 | e | 445 | |

3 rows × 58 columns

## Continua… Módulo 3 – Tarea 2