Autores: María Inmaculada Rodríguez García, María Gema Carrasco García, Javier González Enrique, Juan Jesús Ruiz Aguilar, Ignacio J. Turias Domínguez. Universidad de Cádiz

2.4. Tratamiento de datos perdidos

Los comandos para Python son:

- Número de filas con NaN: **df.isna().sum().sum()**
- Eliminar filas con NaN: **df.dropna()**
- Reemplazar NaN con un valor: **df.fillna ()**
- Borrar filas duplicadas: **df.drop_duplicates()**

En Matlab son:

- Número de filas con NaN: **num_rows_with_nan = sum(any(isnan(T{:,:}), 2));**
  disp(['Número de filas con NaN: ', num2str(num_rows_with_nan)]);
- Eliminar filas con NaN: **T_cleaned = rmmissing(T, 'row');**
- Reemplazar NaN con un valor: **T_filled = fillmissing(T, 'constant', 0);**
- Borrar filas duplicadas: **T_unique = unique(T);**

Ejercicio 74_01. Tratamiento de datos perdidos en Python

```
# Importamos las librerías necesarias
import numpy as np
import pandas as pd

df = pd.DataFrame({"Est": ["A", "B", "C", "D"], "Contaminante": [np.nan, 'SO2',
'SO2', 'SO2'], "Tipo": ["Train", "Test", "Test", "Train"]})
print('Dataset que se ha cargado: ')
print(df)

print('Número de filas con NaN en el dataset: ')
nulos =  df.isna().sum().sum()
print(nulos)

print('Eliminamos las filas con NaN: ')
df2 =  df.dropna()
print(df2)

print('Reemplazar NaN con un PM10: ')
df3 =  df.fillna('PM10')
print(df3)

print('Borrar filas duplicadas')
df4 = df.drop_duplicates()
print(df4)
```

```
Dataset que se ha cargado:
     Est       Contaminante    Tipo
0    A           NaN          Train
1    B           SO2          Test
2    C           SO2          Test
3    D           SO2          Train
```

Autores: María Inmaculada Rodríguez García, María Gema Carrasco García, Javier González Enrique, Juan Jesús Ruiz Aguilar, Ignacio J. Turias Domínguez. Universidad de Cádiz

```
Número de filas con NaN en el dataset:
1
Eliminamos las filas con NaN:
     Est       Contaminante    Tipo
1    B           SO2          Test
2    C           SO2          Test
3    D           SO2          Train
Reemplazar NaN con un PM10:
     Est       Contaminante    Tipo
0    A           PM10         Train
1    B           SO2          Test
2    C           SO2          Test
3    D           SO2          Train
Borrar filas duplicadas
     Est       Contaminante    Tipo
0    A           NaN          Train
1    B           SO2          Test
3    D           SO2          Train
```

Ejercicio 74_02. Tratamiento de datos perdidos en MATLAB

```
est = {'A';'B';'C';'D'};
contam = {''; 'SO2'; 'SO2'; 'SO2'};
tipo = {'Train';'Test';'Test';'Train'};
df = table(est, contam, tipo)

% Mostramos el DataTable original
disp('DataTable original:');
disp(df);

% Número de filas con missing values en el DataTable
nulos = ismissing(df)
```

```
DataTable original:
     est       contam        tipo
    _____    __________    _________

    {'A'}    {0×0 char}    {'Train'}
    {'B'}    {'SO2'   }    {'Test' }
    {'C'}    {'SO2'   }    {'Test' }
    {'D'}    {'SO2'   }    {'Train'}

nulos =

  4×3 logical array

   0   1   0
   0   0   0
   0   0   0
   0   0   0
```

```
% Reemplazar valores faltantes con 'PM10'
df2 = df;
nanIndices = ismissing(df2.contam);
```

Autores: María Inmaculada Rodríguez García, María Gema Carrasco García, Javier González Enrique, Juan Jesús Ruiz Aguilar, Ignacio J. Turias Domínguez. Universidad de Cádiz

```
df2.contam(nanIndices) = {'PM10'};

% Mostrar el DataFrame con valores faltantes reemplazados
disp('DataTable con valores faltantes reemplazados:');
disp(df2);

df2 = 4×3 table
     est      contam       tipo
    _____    ________    _________

    {'A'}    {'PM10'}    {'Train'}
    {'B'}    {'SO2' }    {'Test' }
    {'C'}    {'SO2' }    {'Test' }
    {'D'}    {'SO2' }    {'Train'}

% Eliminar filas duplicadas basadas en todas las columnas
df3 = unique(df2, 'rows', 'stable');

% Eliminar filas duplicadas sin tener en cuenta la columna 'estación'
colsToConsider = {'contam', 'tipo'};
df3 = unique(df2(:, colsToConsider), 'rows', 'stable');

% Mostrar el DataFrame sin filas duplicadas
disp('DataTable sin filas duplicadas:');
disp(df3);

DataTable sin filas duplicadas:
     contam       tipo
    ________    _________

    {'PM10'}    {'Train'}
    {'SO2' }    {'Test' }
    {'SO2' }    {'Train'}
```