Visualización de Datos

Pandas: Manejo de *arrays* multidimensionales a través *DataFrames* de la librería *Pandas*

Francisco Javier Mercader Martínez

Ejercicios

Ejercicio 1. Crea un Data con la siguiente información:

Nombre	Grupo Sanguíneo	Edad	Peso	Ritmo Cardiaco	Presión Sistólica	Presión Diastólica
Eduardo	B+	40	70	70	129	-
Ana	O+	35	69	60	133	86
Alejandro	AB+	37	74	68	125	82
Álvaro	O+	24	70	50	110	70
Aitana	A+	48	72	68	-	82
María	A+	53	67	87	130	84
Sofia	B-	67	65	110	155	100
Antonio	A+	25	74	-	126	89
Fernando	AB+	38	75	77	131	90
Laura	O+	21	70	69	127	87

```
import pandas as pd
import numpy as np

data = {
    "Nombre": ["Eduardo", "Ana", "Alejandro", "Álvaro", "Aitana", "María", "Sofia", "Antonio",
    ↪  "Fernando", "Laura"],
    "Grupo Sanguíneo": ["B+", "O+", "AB+", "O+", "A+", "A+", "B-", "A+", "AB+", "O+"],
    "Edad": [40, 35, 37, 24, 48, 53, 67, 25, 38, 21],
    "Peso": [70, 69, 74, 70, 72, 67, 65, 74, 75, 70],
    "Ritmo Cardiaco": [70, 60, 68, 50, 68, 87, 110, None, 77, 69],
    "Presión Sistólica": [129, 133, 125, 110, None, 130, 155, 126, 131, 127],
    "Presión Diastólica": [None, 86, 82, 70, 82, 84, 100, 89, 90, 87]
}

df = pd.DataFrame(data)
df['Altura'] = np.random.uniform(1.55, 1.81, size=len(df)).round(2)
print(df)
```

```
      Nombre Grupo Sanguíneo  Edad  Peso  Ritmo Cardiaco  Presión Sistólica  \
0    Eduardo              B+    40    70            70.0              129.0
1        Ana              O+    35    69            60.0              133.0
2  Alejandro             AB+    37    74            68.0              125.0
3     Álvaro              O+    24    70            50.0              110.0
4     Aitana              A+    48    72            68.0                NaN
5      María              A+    53    67            87.0              130.0
6      Sofia              B-    67    65           110.0              155.0
7    Antonio              A+    25    74             NaN              126.0
8   Fernando             AB+    38    75            77.0              131.0
9      Laura              O+    21    70            69.0              127.0

   Presión Diastólica  Altura
0                 NaN    1.63
1                86.0    1.76
```

```
2               82.0    1.58
3               70.0    1.56
4               82.0    1.67
5               84.0    1.75
6              100.0    1.58
7               89.0    1.64
8               90.0    1.67
9               87.0    1.80
```

Ejercicio 2. Crea una nueva columna en el DataFrame anterior donde aparezca el cálculo del Indice de Masa Corporal (IMC) que se calcula de la siguiente forma:

$$IMC = \frac{\text{peso}}{\text{altura}^2}$$

```
df["IMC"] = (df['Peso'] / (df['Altura']**2)).round(2)
print(df)
```

```
      Nombre Grupo Sanguíneo  Edad  Peso  Ritmo Cardiaco  Presión Sistólica  \
0    Eduardo              B+    40    70            70.0              129.0
1        Ana              O+    35    69            60.0              133.0
2  Alejandro             AB+    37    74            68.0              125.0
3     Álvaro              O+    24    70            50.0              110.0
4     Aitana              A+    48    72            68.0                NaN
5      María              A+    53    67            87.0              130.0
6      Sofia              B-    67    65           110.0              155.0
7    Antonio              A+    25    74             NaN              126.0
8   Fernando             AB+    38    75            77.0              131.0
9      Laura              O+    21    70            69.0              127.0

   Presión Diastólica  Altura    IMC
0                 NaN    1.63  26.35
1                86.0    1.76  22.28
2                82.0    1.58  29.64
3                70.0    1.56  28.76
4                82.0    1.67  25.82
5                84.0    1.75  21.88
6               100.0    1.58  26.04
7                89.0    1.64  27.51
8                90.0    1.67  26.89
9                87.0    1.80  21.60
```

Actividad 3. Convierte la columna Grupo Sanguíneo de 2 formas distintas (sin eliminar la columna original)

- Usando un Label Encoding

```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
# Grupo Sanguíneo Label Encoding (GS_LE)
df['GS_LE'] = label_encoder.fit_transform(df['Grupo Sanguíneo'])
print(df)
```

```
      Nombre Grupo Sanguíneo  Edad  Peso  Ritmo Cardiaco  Presión Sistólica  \
0    Eduardo              B+    40    70            70.0              129.0
1        Ana              O+    35    69            60.0              133.0
2  Alejandro             AB+    37    74            68.0              125.0
3     Álvaro              O+    24    70            50.0              110.0
4     Aitana              A+    48    72            68.0                NaN
5      María              A+    53    67            87.0              130.0
6      Sofia              B-    67    65           110.0              155.0
7    Antonio              A+    25    74             NaN              126.0
8   Fernando             AB+    38    75            77.0              131.0
9      Laura              O+    21    70            69.0              127.0

   Presión Diastólica  Altura    IMC  GS_LE
```

```
   0                NaN    1.63  26.35      2
   1               86.0    1.76  22.28      4
   2               82.0    1.58  29.64      1
   3               70.0    1.56  28.76      4
   4               82.0    1.67  25.82      0
   5               84.0    1.75  21.88      0
   6              100.0    1.58  26.04      3
   7               89.0    1.64  27.51      0
   8               90.0    1.67  26.89      1
   9               87.0    1.80  21.60      4
```

- Usando un esquema de One Hot Encoding

```
from sklearn.preprocessing import OneHotEncoder
ohe_encoder = OneHotEncoder()
ohe_df = pd.DataFrame(ohe_encoder.fit_transform(df[["Grupo Sanguíneo"]]).toarray())
df_merged = pd.concat([df, ohe_df], axis=1)
print(df_merged)
```

```
      Nombre Grupo Sanguíneo  Edad  Peso  Ritmo Cardiaco  Presión Sistólica  \
   0    Eduardo              B+    40    70            70.0              129.0
   1        Ana              O+    35    69            60.0              133.0
   2  Alejandro             AB+    37    74            68.0              125.0
   3     Álvaro              O+    24    70            50.0              110.0
   4     Aitana              A+    48    72            68.0                NaN
   5      María              A+    53    67            87.0              130.0
   6      Sofia              B-    67    65           110.0              155.0
   7    Antonio              A+    25    74             NaN              126.0
   8   Fernando             AB+    38    75            77.0              131.0
   9      Laura              O+    21    70            69.0              127.0

      Presión Diastólica  Altura    IMC  GS_LE    0    1    2    3    4
   0                 NaN    1.63  26.35      2  0.0  0.0  1.0  0.0  0.0
   1                86.0    1.76  22.28      4  0.0  0.0  0.0  0.0  1.0
   2                82.0    1.58  29.64      1  0.0  1.0  0.0  0.0  0.0
   3                70.0    1.56  28.76      4  0.0  0.0  0.0  0.0  1.0
   4                82.0    1.67  25.82      0  1.0  0.0  0.0  0.0  0.0
   5                84.0    1.75  21.88      0  1.0  0.0  0.0  0.0  0.0
   6               100.0    1.58  26.04      3  0.0  0.0  0.0  1.0  0.0
   7                89.0    1.64  27.51      0  1.0  0.0  0.0  0.0  0.0
   8                90.0    1.67  26.89      1  0.0  1.0  0.0  0.0  0.0
   9                87.0    1.80  21.60      4  0.0  0.0  0.0  0.0  1.0
```

Ejercicio 4. Elimina la columna original de Grupo Sanguíneo.

```
print(df.drop('Grupo Sanguíneo', axis=1))
```

```
      Nombre  Edad  Peso  Ritmo Cardiaco  Presión Sistólica  \
0    Eduardo    40    70            70.0              129.0
1        Ana    35    69            60.0              133.0
2  Alejandro    37    74            68.0              125.0
3     Álvaro    24    70            50.0              110.0
4     Aitana    48    72            68.0                NaN
5      María    53    67            87.0              130.0
6      Sofia    67    65           110.0              155.0
7    Antonio    25    74             NaN              126.0
8   Fernando    38    75            77.0              131.0
9      Laura    21    70            69.0              127.0

   Presión Diastólica  Altura    IMC  GS_LE
0                 NaN    1.63  26.35      2
1                86.0    1.76  22.28      4
2                82.0    1.58  29.64      1
3                70.0    1.56  28.76      4
4                82.0    1.67  25.82      0
```

```
5          84.0    1.75  21.88      0
6         100.0    1.58  26.04      3
7          89.0    1.64  27.51      0
8          90.0    1.67  26.89      1
9          87.0    1.80  21.60      4
```

Ejercicio 5. Une el DataFrame anterior con el siguiente

Nombre	Cuidad	Nivel de Estudios	Azúcar	Colesterol
Eduardo	Valencia	Primaria	80	156
Ana	Murcia	Primaria	82	167
Manuel	Madrid	Primaria	114	204
Aitana	Barcelona	Primaria	94	226
María	Sevilla	Universidad	130	167
Antonio	Madrid	Formación Profesional	83	190
Fernando	Barcelona	Secundaria	82	199
Angela	Murcia	Secundaria	103	192

```python
# Datos del segundo DataFrame
data2 = {
    "Nombre": ["Eduardo", "Ana", "Manuel", "Aitana", "María", "Antonio", "Fernando", "Angela"],
    "Cuidad": ["Valencia", "Murcia", "Madrid", "Barcelona", "Sevilla", "Madrid", "Barcelona",
    ↪  "Murcia"],
    "Nivel de Estudios": ["Primaria", "Primaria", "Primaria", "Primaria", "Universidad",
    ↪  "Formación Profesional", "Secundaria", "Secundaria"],
    "Azúcar": [80, 82, 114, 94, 130, 83, 82, 103],
    "Colesterol": [156, 167, 204, 226, 167, 190, 199, 192]
}

df2 = pd.DataFrame(data2)

# Unir los DataFrames
df_merged = pd.merge(df, df2, on="Nombre", how="left")
df_merged
```

	Nombre	Grupo Sanguíneo	Edad	Peso	Ritmo Cardiaco	Presión Sistólica	Presión Diastólica	Altura	IMC
0	Eduardo	B+	40	70	70.0	129.0	NaN	1.63	26.35
1	Ana	O+	35	69	60.0	133.0	86.0	1.76	22.28
2	Alejandro	AB+	37	74	68.0	125.0	82.0	1.58	29.64
3	Álvaro	O+	24	70	50.0	110.0	70.0	1.56	28.76
4	Aitana	A+	48	72	68.0	NaN	82.0	1.67	25.82
5	María	A+	53	67	87.0	130.0	84.0	1.75	21.88
6	Sofia	B-	67	65	110.0	155.0	100.0	1.58	26.04
7	Antonio	A+	25	74	NaN	126.0	89.0	1.64	27.51
8	Fernando	AB+	38	75	77.0	131.0	90.0	1.67	26.89
9	Laura	O+	21	70	69.0	127.0	87.0	1.80	21.60