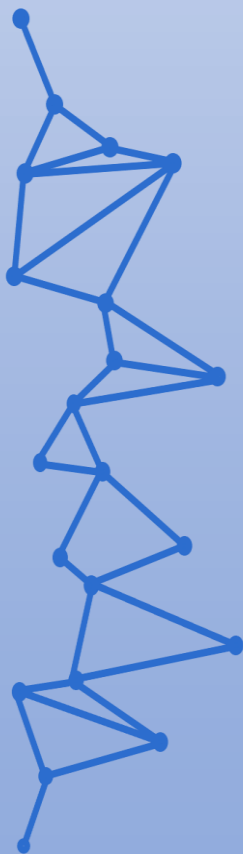




Curso de Especialización de Inteligencia Artificial y Big Data (IABD)



Sistemas de Big Data

UD04. Manejo de Dataframes con
Python/Pandas.
Tarea Online.

JUAN ANTONIO GARCÍA MUELAS

INDICE

	Pag
1. Caso práctico	2
2. Apartado 1: Crear dataframe	2
3. Apartado 2: Añadir columna	3
4. Apartado 3: Dividir columna	3
5. Apartado 4: Añadir nuevas filas	3
6. Apartado 5: Crear columna 'Media'	4
7. Apartado 6: Eliminar fila por índice	5
8. Apartado 7: Mostrar filas Dataframe	5
9. Apartado 8: Ver información estadística	6

Tarea para SBD04

Título de la tarea: Manejo de DataFrames con Python/Pandas

Curso de Especialización: Inteligencia Artificial y Big Data

Módulo profesional: Sistemas de Big Data

¿Qué te pedimos que hagas?

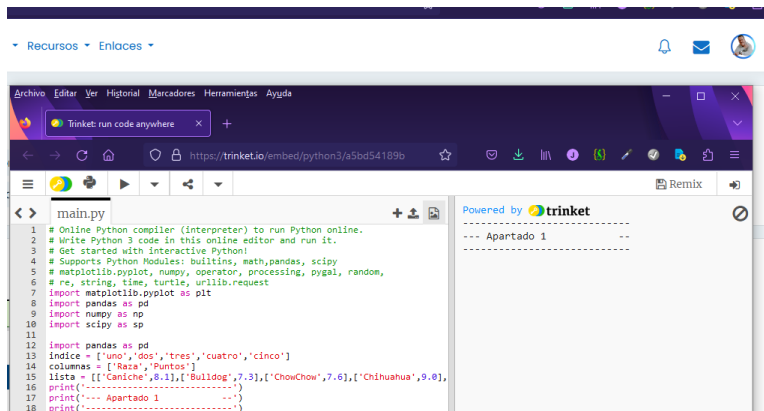
A continuación, tienes la primera parte del programa de Enrique.

Accede al [intérprete de Python en trinket.io](https://trinket.io), pega el código fuente que te entregamos y ejecútalo (con el botón del triángulo negro de la parte superior).

```
import pandas as pd

indice = ['uno', 'dos', 'tres', 'cuatro', 'cinco']
columnas = ['Raza', 'Puntos']
lista = [[ 'Caniche', 8.1], [ 'Bulldog', 7.3], [ 'ChowChow', 7.6], [ 'Chihuahua', 9.0], [ 'Labrador', 9.3]]

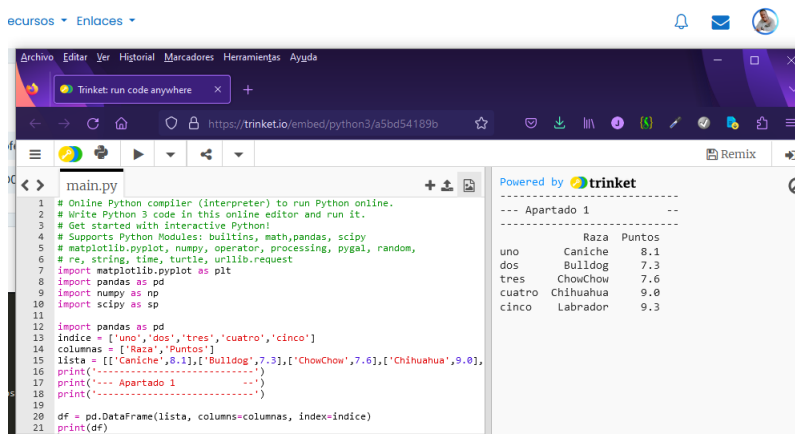
print('-----')
print('--- Apartado 1 ---')
print('-----')
```



✓ Apartado 1:

Crea un DataFrame llamado **df** a partir de la lista de razas y puntuaciones otorgadas por Enrique, usando el índice, las columnas que se te proporcionan.

```
df = pd.DataFrame(lista, columns=columnas, index=indice)
print(df)
```



✓ **Apartado 2:**

Ana le entrega sus puntuaciones a Enrique y éste las transcribe en forma de la siguiente lista:

```
puntos_ana = [61,75,82,95,99]
```

Añade la columna con etiqueta 'PuntosAna' al DataFrame.

```
puntos_ana = [61,75,82,95,99]
df['PuntosAna'] = puntos_ana
print(df)
```

```
1 # Online Python compiler (interpreter) to run Python online.
2 # Write Python 3 code in this online editor and run it.
3 # Get started with interactive Python!
4 # Supports Python Modules: builtins, math, pandas, scipy
5 # matplotlib.pyplot, numpy, operator, processing, pygal, random,
6 # re, string, time, turtle, urllib.request
7 import matplotlib.pyplot as plt
8 import pandas as pd
9 import numpy as np
10 import scipy as sp
11
12 # Apartado 1
13 indice = ['uno','dos','tres','cuatro','cinco']
14 columnas = ['Raza','Puntos']
15 lista = [['Caniche',8.1],['Bulldog',7.3],['ChowChow',7.6],['Chihuahua',9.0],
16         ['Labrador',9.3]]
17 print('--- Apartado 1 ---')
18 print('---')
19
20 df = pd.DataFrame(lista, columns=columnas, index=indice)
21 print(df)
22
23 # Apartado 2
24 print('--- Apartado 2 ---')
25 print('---')
26 print('---')
27
28 # Apartado 3
29 puntos_ana = [61,75,82,95,99]
30 df['PuntosAna'] = puntos_ana
31 print(df)
```

Raza	Puntos	PuntosAna	
uno	Caniche	8.1	61
dos	Bulldog	7.3	75
tres	ChowChow	7.6	82
cuatro	Chihuahua	9.0	95
cinco	Labrador	9.3	99

✓ **Apartado 3:**

Como las puntuaciones de Ana han entrado en valores de 0 a 100 y las de Enrique estaban de 0 a 10, divide la columna 'PuntosAna' entre 10 de modo que todas las puntuaciones queden en la misma escala.

```
df['PuntosAna'] = np.array(puntos_ana)/10
print(df)
```

```
32 # Apartado 3
33 print('--- Apartado 3 ---')
34 print('---')
35 print('---')
36
37 df['PuntosAna'] = np.array(puntos_ana)/10
38 print(df)
```

Raza	Puntos	PuntosAna	
uno	Caniche	8.1	6.1
dos	Bulldog	7.3	7.5
tres	ChowChow	7.6	8.2
cuatro	Chihuahua	9.0	9.5
cinco	Labrador	9.3	9.9

✓ **Apartado 4:**

Enrique y Ana han estado viendo una revista de animales y han decidido añadir otras dos razas más, por lo que han creado la siguiente lista ya con sus puntuaciones integradas.

```
lista2 = [['Samoyedo',9.2,8.9],['Pinscher',8.1,6.7]]
```

Añade esas dos nuevas filas al DataFrame. Fíjate en que se han quedado sin ideas para los índices y han decidido usar 'nuevo1' y 'nuevo2' para las nuevas filas.

```
lista2 = [['Samoyedo',9.2,8.9],['Pinscher',8.1,6.7]]
df2 = pd.DataFrame(lista2, columns=['Raza','Puntos','PuntosAna'],
index=['nuevo1','nuevo2'])
df = df.append(df2)
print(df)
```

```

39 print(df)
40
41 #apartado 4
42 print('-----')
43 print('--- Apartado 4 ---')
44 print('-----')
45 lista2 = [['Samoyedo',9.2,8.9],['Pinscher',8.1,6.7]]
46 df2 = pd.DataFrame(lista2, columns=['Raza','Puntos','PuntosAna'], index=['nuevo1','nuevo2'])
47 df = df.append(df2)
48 print(df)
49
50
51
52
53
54
55

```

	Raza	Puntos	PuntosAna
uno	Caniche	8.1	6.1
dos	Bulldog	7.3	7.5
tres	ChowChow	7.6	8.2
cuatro	Chihuahua	9.0	9.5
cinco	Labrador	9.3	9.9
nuevo1	Samoyedo	9.2	8.9
nuevo2	Pinscher	8.1	6.7

En este ejercicio, el editor avisa de un futuro paso a desuso en la función `append()`.

✓ Apartado 5:

Crea la columna 'Media' para poder ver la puntuación media para cada raza.

```
x1 = df[['Puntos', 'PuntosAna']].mean(axis=1)
df['Media'] = x1.sort_index()
print(df)
```

```

48 print(df)
49
50 #apartado 5
51 print('-----')
52 print('--- Apartado 5 ---')
53 print('-----')
54
55 x1 = df[['Puntos', 'PuntosAna']].mean(axis=1)
56 df['Media'] = x1.sort_index()
57 print(df)
58
59
60
61
62
63
64

```

	Raza	Puntos	PuntosAna	Media
uno	Caniche	8.1	6.1	7.10
dos	Bulldog	7.3	7.5	7.40
tres	ChowChow	7.6	8.2	7.90
cuatro	Chihuahua	9.0	9.5	9.25
cinco	Labrador	9.3	9.9	9.60
nuevo1	Samoyedo	9.2	8.9	9.05
nuevo2	Pinscher	8.1	6.7	7.40

✓ **Apartado 6:**

Enrique y Ana deciden que no tendrán un bulldog. Elimina la fila cuyo índice es 'dos'.

```
df = df.drop('dos')
print(df)
```

The screenshot shows a Trinket.io Python environment. The code in `main.py` calculates the mean of 'PuntosAna' for each 'Raza', sorts the DataFrame by this mean, and then drops the row with index 'dos'. The output shows the resulting DataFrame with 10 rows.

Raza	Puntos	PuntosAna	Media	
uno	Caniche	8.1	6.1	7.10
tres	ChowChow	7.6	8.2	7.90
cuatro	Chihuahua	9.0	9.5	9.25
cinco	Labrador	9.3	9.9	9.60
nuevo1	Samoyedo	9.2	8.9	9.05
nuevo2	Pinscher	8.1	6.7	7.40

✓ **Apartado 7:**

Viendo las puntuaciones medias, Enrique y Ana deciden que la decisión va a estar entre chihuahua, labrador y samoyedo, por lo que deciden obtener sólo esas filas del DataFrame (en concreto desde la posición 2 hasta la 5 -no incluída-).

```
df = df[2:5]
print(df)
```

The screenshot shows the same Trinket.io Python environment. The code now slices the DataFrame to keep only rows with indices 2, 3, 4, and 5. The output shows the resulting DataFrame with 4 rows.

Raza	Puntos	PuntosAna	Media	
cuatro	Chihuahua	9.0	9.5	9.25
cinco	Labrador	9.3	9.9	9.60
nuevo1	Samoyedo	9.2	8.9	9.05
nuevo2	Pinscher	8.1	6.7	7.40

✓ **Apartado 8:**

Por último, sólo por curiosidad, deciden ver información estadística sobre las filas que les han quedado.

```
print(df.describe())
```

The screenshot shows a Trinket.io Python code editor with a file named `main.py`. The code includes a `print(df)` statement followed by a section header `#apartado 8`, a separator line, and a `print(df.describe())` statement. The output of the code is displayed on the right side of the editor.

The output of `print(df)` is a DataFrame with the following data:

	Raza	Puntos	PuntosAna	Media
cuatro	Chihuahua	9.0	9.5	9.25
cinco	Labrador	9.3	9.9	9.60
nuevo1	Samoyedo	9.2	8.9	9.05

The output of `print(df.describe())` is a statistical summary of the DataFrame:

	Puntos	PuntosAna	Media
count	3.000000	3.000000	3.000000
mean	9.166667	9.433333	9.300000
std	0.152753	0.503322	0.278388
min	9.000000	8.900000	9.050000
25%	9.100000	9.200000	9.150000
50%	9.200000	9.500000	9.250000
75%	9.250000	9.700000	9.425000
max	9.300000	9.900000	9.600000