

TAREA DE SISTEMAS DE BIG DATA 04

Caso práctico

Enrique y Ana se conocieron hace 3 años y han decidido irse a vivir juntos.

Les encantan los perros, así que además están pensando en adoptar uno.

Enrique, que es programador, creó hace unos días un sencillo programa en Python para poder dar un vistazo rápido a las razas que más le gustan, a modo de DataFrame de Pandas en el que contempla tanto el nombre de la raza como una puntuación de 0 a 10 indicando cuánto le gusta esa raza.

Ana está de acuerdo en adoptar una de las razas que Enrique tiene contempladas, así que le ha dicho cuánto le gusta cada una de ellas (de 0 a 100) para que Enrique lo integre en su DataFrame. Así podrán calcular medias para decidir qué raza les gusta más a los dos.

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('-----')
```

Utiliza la información que encontrarás en los contenidos de la unidad para crear un documento en el que escribirás el código fuente que Enrique ha necesitado seguir escribiendo para llegar a las soluciones correspondientes a lo que se te dice en los siguientes apartados.

Al final de los apartados te damos toda la salida por pantalla que deberá tener el programa, para que sepas con claridad si estás consiguiendo lo que se te solicita. Tendrás que usar constantemente la función *print* para ello.

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.

The screenshot shows a web browser window with a Trinket.io Python3 environment. The code in the editor is as follows:

```
1 import pandas as pd
2 import numpy as np
3
4 indice = ['uno', 'dos', 'tres', 'cuatro', 'cinco']
5 columnas = ['Raza', 'Puntos']
6 lista = [['Caniche', 8.1], ['Bulldog', 7.3], ['ChowChow', 7.6], ['Chihuahua', 9.0], ['Labrador', 9.3]]
7 print('-----')
8 print('-----Apartado 1-----')
9 print('-----')
10 df = pd.DataFrame(lista, indice, columnas)
11 print(df)
```

The output on the right shows the DataFrame created:

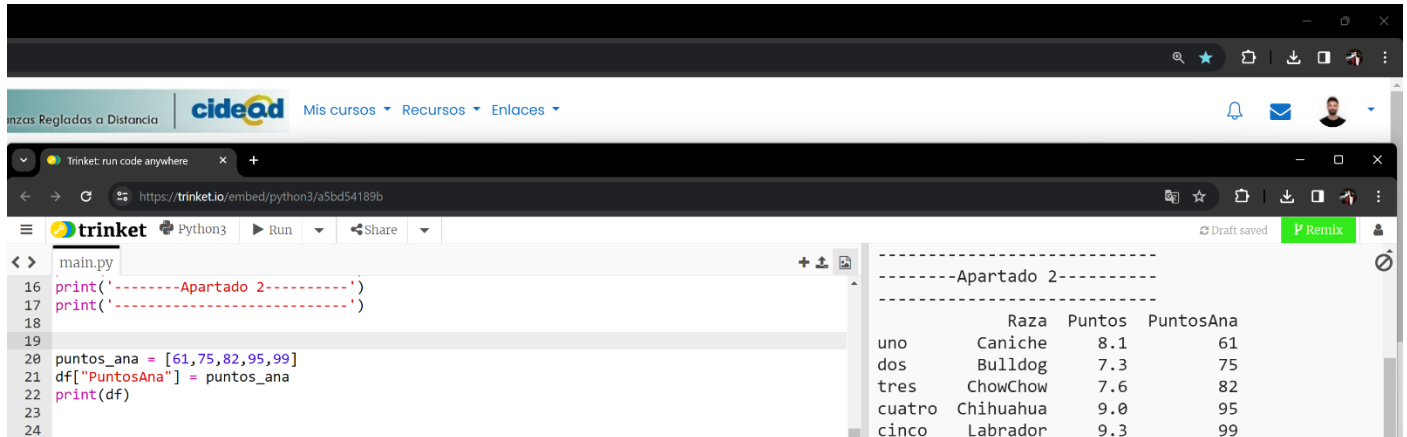
	Raza	Puntos
uno	Caniche	8.1
dos	Bulldog	7.3
tres	ChowChow	7.6
cuatro	Chihuahua	9.0
cinco	Labrador	9.3

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.



The screenshot shows a Trinket.io interface with a Python3 environment. The code in main.py is as follows:

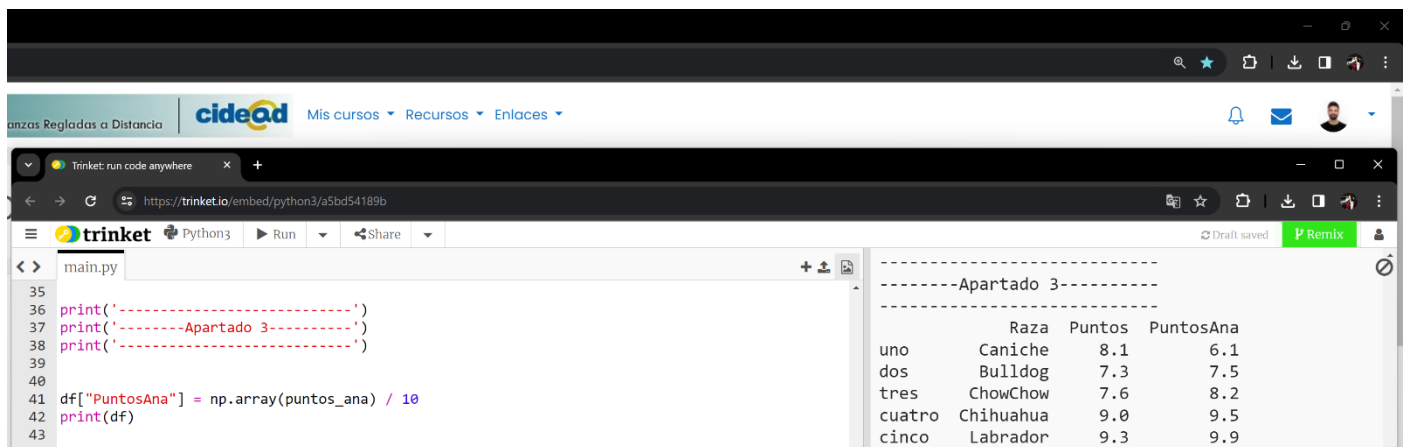
```
16 print('-----Apartado 2-----')
17 print('-----')
18
19
20 puntos_ana = [61,75,82,95,99]
21 df["PuntosAna"] = puntos_ana
22 print(df)
23
24
```

The output on the right shows a DataFrame with the following data:

	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.



The screenshot shows a Trinket.io interface with a Python3 environment. The code in main.py is as follows:

```
35
36 print('-----')
37 print('-----Apartado 3-----')
38 print('-----')
39
40
41 df["PuntosAna"] = np.array(puntos_ana) / 10
42 print(df)
43
44
```

The output on the right shows a DataFrame with the following data:

	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.

Trinket: run code anywhere

https://trinket.io/embed/python3/a5bd54189b

```

main.py
50
51 print('-----')
52 print('-----Apartado 4-----')
53 print('-----')
54
55
56 lista2 = [['Samoyedo',9.2,8.9],['Pinscher',8.1,6.7]]
57 df2 = pd.DataFrame(lista2, columns=['Raza','Puntos','PuntosAna'], index=['nuevo1','nuevo2'])
58 df = df.append(df2)
59 print(df)
60
61

```

	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

Apartado 5:

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

Trinket: run code anywhere

https://trinket.io/embed/python3/a5bd54189b

```

main.py
43
44 print('-----')
45 print('-----Apartado 5-----')
46 print('-----')
47
48
49 media = df[['Puntos', 'PuntosAna']].mean(axis=1)
50 df['Media'] = media.sort_index()
51 print(df)
52
53
54
55

```

	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'.

Trinket: run code anywhere

https://trinket.io/embed/python3/a5bd54189b

```

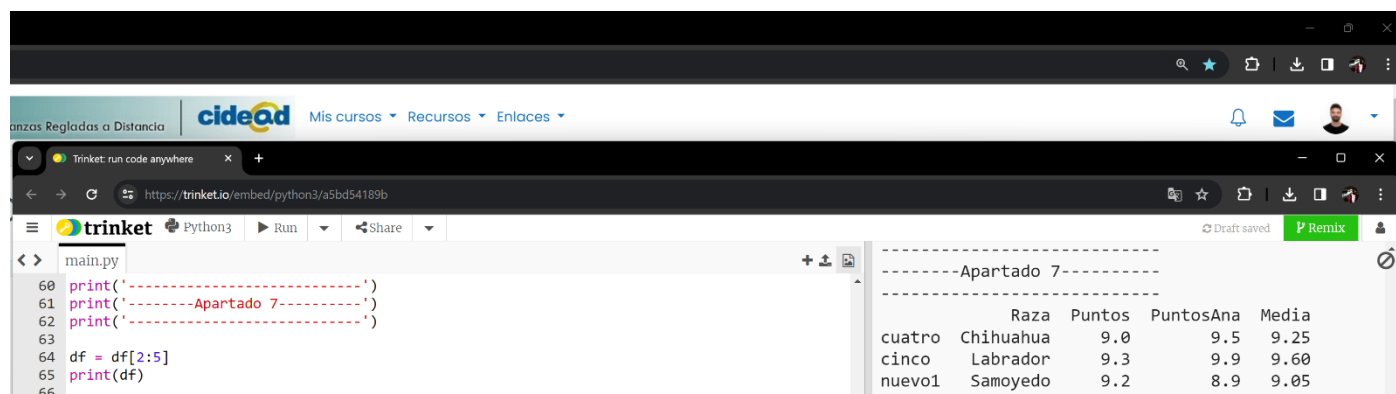
main.py
52
53 print('-----')
54 print('-----Apartado 6-----')
55 print('-----')
56
57 df = df.drop('dos')
58 print(df)
59
60
61
62

```

	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 incluida-).



The screenshot shows a Trinket.io interface with a Python3 environment. The code in `main.py` is as follows:

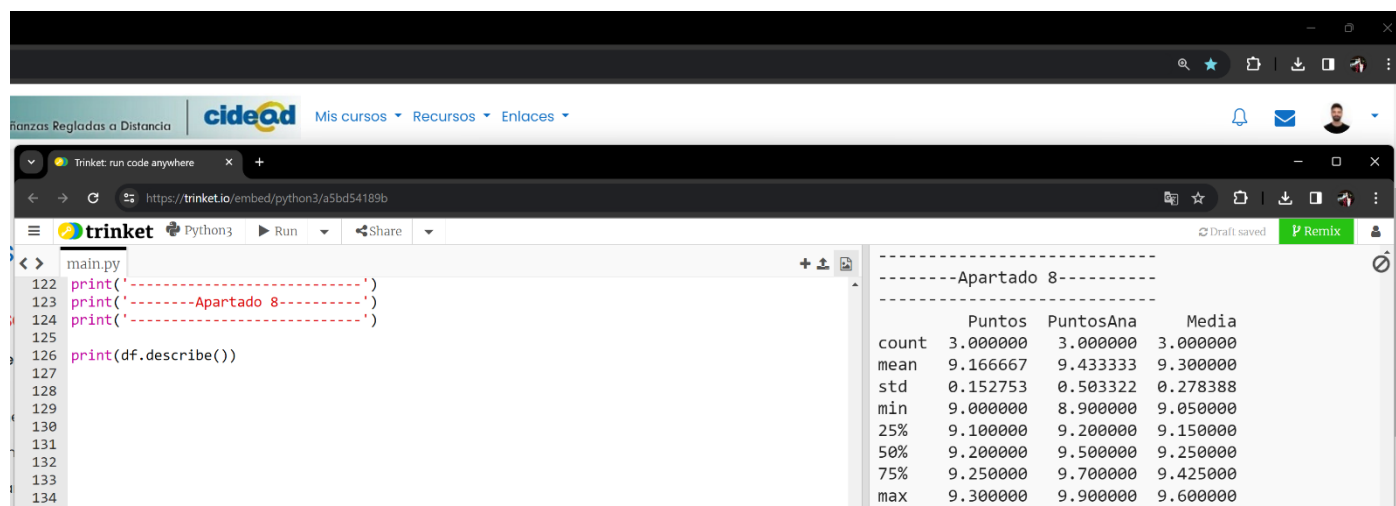
```
60 print('-----Apartado 7-----')
61 print('-----Apartado 7-----')
62 print('-----Apartado 7-----')
63
64 df = df[2:5]
65 print(df)
66
```

The output on the right displays the filtered DataFrame:

	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

Apartado 8:

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



The screenshot shows the same Trinket.io interface. The code in `main.py` is updated to:

```
122 print('-----Apartado 8-----')
123 print('-----Apartado 8-----')
124 print('-----Apartado 8-----')
125
126 print(df.describe())
127
128
129
130
131
132
133
134
```

The output on the right displays the 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

Código completo:

```
import pandas as pd
import numpy as np

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]]
df = pd.DataFrame(lista, indice, columnas)
print(df)

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

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

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)

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

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

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

print(df.describe())
```