



Lectura

Creando columnas derivadas

No siempre la información que almacenamos es suficiente para obtener resultados. Algunas veces nos vemos en la necesidad de generar nueva información a partir de la que tenemos. Por ejemplo, si tenemos la altura y el peso de una persona, podemos calcular su índice de masa corporal en una nueva columna. Este cálculo es muy sencillo, pero imagine el tiempo que tomaría y los recursos que se consumen en obtener los resultados cada vez que los ocupemos.

Realizar la creación de nuevas columnas en el data frame es tan sencillo como asignarle valor a una variable. Los procedimientos de pandas son tan avanzados que entienden que los cálculos se deben realizar utilizando elemento por elemento de las columnas a considerar. A continuación abordaremos dos ejemplos para crear una columna derivada: uno utilizando solamente una columna y otro usando dos columnas de nuestros datos.



Para generar una nueva columna creamos una instrucción de asignación como si se tratará de un arreglo (en realidad una serie) sin la necesidad de ciclos. Dentro de los corchetes colocamos el nombre de la nueva columna y le asignamos el valor deseado, que en muchas ocasiones depende de otra columna de nuestro dataframe. Por ejemplo, en el siguiente script se crea la nueva columna height_m para representar la altura del personaje en metros y los valores que se le asignan son aquellos que están en la columna height dividido entre 100.

```
import pandas as pd
personajesSW = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/M2DCData/characters.csv")
personajesSW["height_m"] = personajesSW["height"] / 100
personajesSW.head()
```

Resultado:

	id	name	height	mass	hair_color	skin_color	eye_color	birth_year	gender	homeworld	species	height_m
0	1	Luke Skywalker	172	77.0	blond	fair	blue	19BBY	male	Tatooine	Human	1.72
1	2	C-3PO	167	75.0	NaN	gold	yellow	112BBY	NaN	Tatooine	Droid	1.67
2	3	R2-D2	96	32.0	NaN	white, blue	red	33BBY	NaN	Naboo	Droid	0.96
3	4	Darth Vader	202	136.0	none	white	yellow	41.9BBY	male	Tatooine	Human	2.02
4	5	Leia Organa	150	49.0	brown	light	brown	19BBY	female	Alderaan	Human	1.50

Como puede observar la nueva columna se encuentra al final de nuestro dataframe.



El siguiente ejemplo utiliza dos columnas del dataframe para generar una nueva que contenga el índice de masa corporal de los personajes.

```
import pandas as pd
personajesSW = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/M2DCData/characters.csv")
personajesSW["imc"] = personajesSW["mass"] / (personajesSW["height"] / 100) ** 2
personajesSW.head()
```

Resultado:

	id	name	height	mass	hair_color	skin_color	eye_color	birth_year	gender	homeworld	species	imc
0	1	Luke Skywalker	172	77.0	blond	fair	blue	19BBY	male	Tatooine	Human	26.027582
1	2	C-3PO	167	75.0	NaN	gold	yellow	112BBY	NaN	Tatooine	Droid	26.892323
2	3	R2-D2	96	32.0	NaN	white, blue	red	33BBY	NaN	Naboo	Droid	34.722222
3	4	Darth Vader	202	136.0	none	white	yellow	41.9BBY	male	Tatooine	Human	33.330066
4	5	Leia Organa	150	49.0	brown	light	brown	19BBY	female	Alderaan	Human	21.777778



Reflexiona

Después de comprender la creación de columnas derivadas reflexiona y contesta las siguientes preguntas.

¿Eres capaz de identificar cuándo es conveniente crear columnas derivadas?

Busca un ejemplo en tu información, muestra la estructura e indica cuál sería la columna derivada y cómo obtendrías sus valores.



¡Ahora es tu turno!

Ahora es momento de practicar:

1. Utiliza el mismo archivo de personajes de Star Wars para:

- Crear una columna nueva llamada grams que contenga el peso del personaje en gramos
- Crear una columna nueva llamada vaccine que contenga el valor 'Apply' si hay que aplicar la vacuna y 'No' en caso de que no haya que aplicarla, la única regla es que se aplicará solamente a las mujeres.
- ¿Puede realizar el punto anterior introduciendo una función de NumPy que realice el trabajo equivalente?

Es importante saber cómo crear columnas adicionales para no tener que recalcular en procesos complejos que demanden mucho tiempo y/o recursos. De esta manera, los cálculos se efectúan una vez y los resultados son almacenados para ocuparse tantas veces como se necesite.