Estandarizar o escalar, ¿por qué?

Supongamos que tenemos un DataFrame con datos de flores. Hay dos tipos de flores: A y B. Una persona mide los tallos y otra los pétalos.

X

y

df

	largo_tallo	ancho_tallo	largo_pétalo	ancho_petálo	tipo
0	100	10	3	1.5	A
1	120	12	5	1.6	A
2	90	14	4	1.2	B
3	80	15	3	1.4	A
4	110	11	3	1.9	B

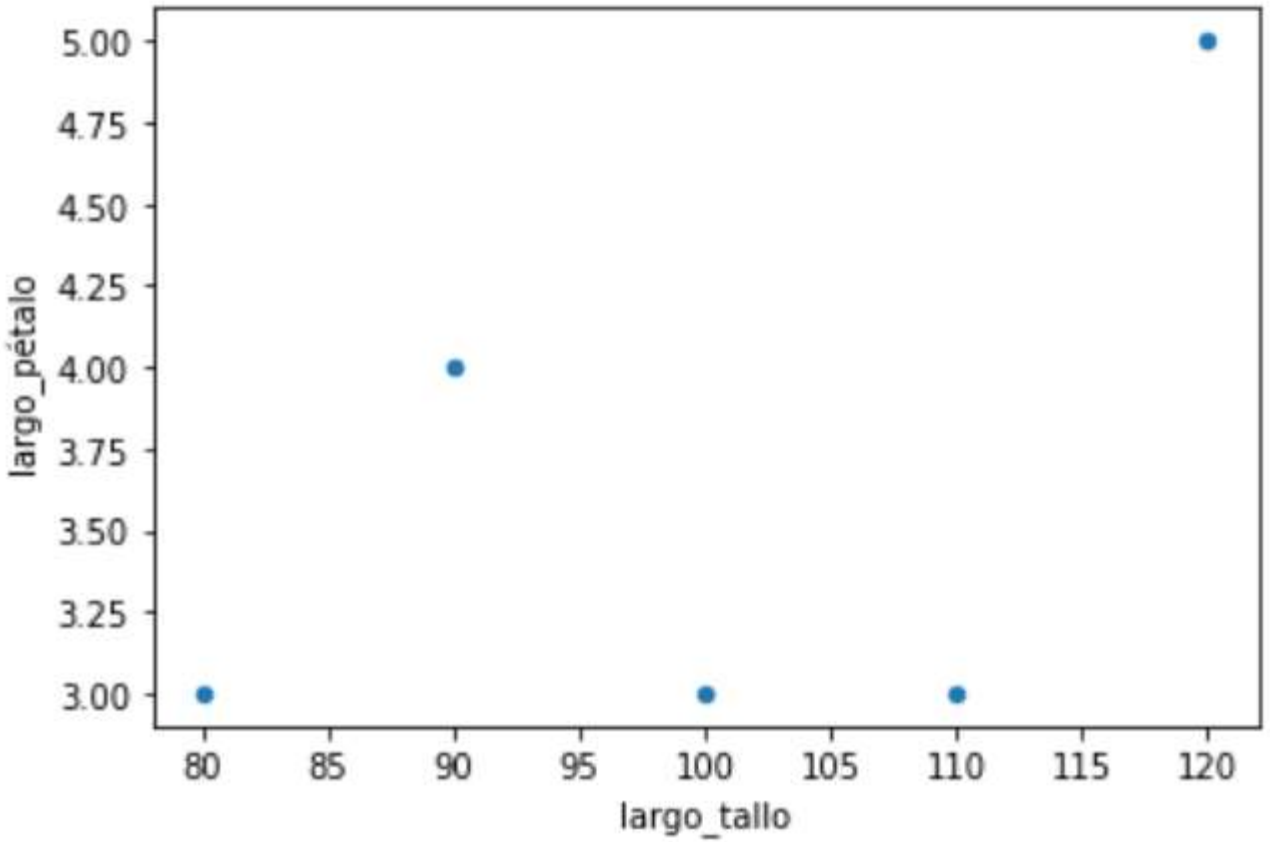
Pero resulta que los tallos están en mm y los pétalos en cm.

```
# largo tallo en mm
# ancho tallo en mm
# largo pétalo en cm
# ancho pétalo en cm
```

df: es un DataFrame
X: es el trozo de df con las variables de entrada
y: es la columna con la variable de salida

X: variables que puedo medir
y: target (objetivo)
Tomo datos X sabiendo y, quiero aprender la relación entre X e y para que luego, cuando no conozca y pueda predecirlo:
y = función(X)

```
df.plot.scatter('largo_tallo', 'largo_pétalo');
```



Dependiendo de las unidades de cada columna, los puntos se van a pintar en sitios distintos.

SOLUCIÓN: que TODAS las columnas numéricas tengan LAS MISMAS UNIDADES

Seguimos. Ahora además incluyo más datos numéricos: la temperatura de la flor, la densidad de su tallo, la resistencia del pétalo y un índice de viscosidad de la savia.

	largo_tallo	ancho_tallo	largo_pétalo	ancho_petálo	tipo	temperatura	densidad	resistencia	viscosidad
0	100	10	3	1.5	A	15.0	200	0.002	1.8
1	120	12	5	1.6	A	16.0	220	0.004	1.9
2	90	14	4	1.2	B	15.5	250	0.003	1.8
3	80	15	3	1.4	A	14.8	190	0.003	2.0
4	110	11	3	1.9	B	14.9	180	0.002	1.3

Ahora no puedo igualar la escala en las unidades, porque, ¿cómo puedo comparar distintas unidades entre sí?
Simplemente NO PUEDO.

CONCEPTO MUY IMPORTANTE→ <u>Lo que hago es pensar en el histograma de cada variable</u>. Voy a coger cada columna numérica del DataFrame y aplicarle una **transformación REVERSIBLE a toda la columna**.

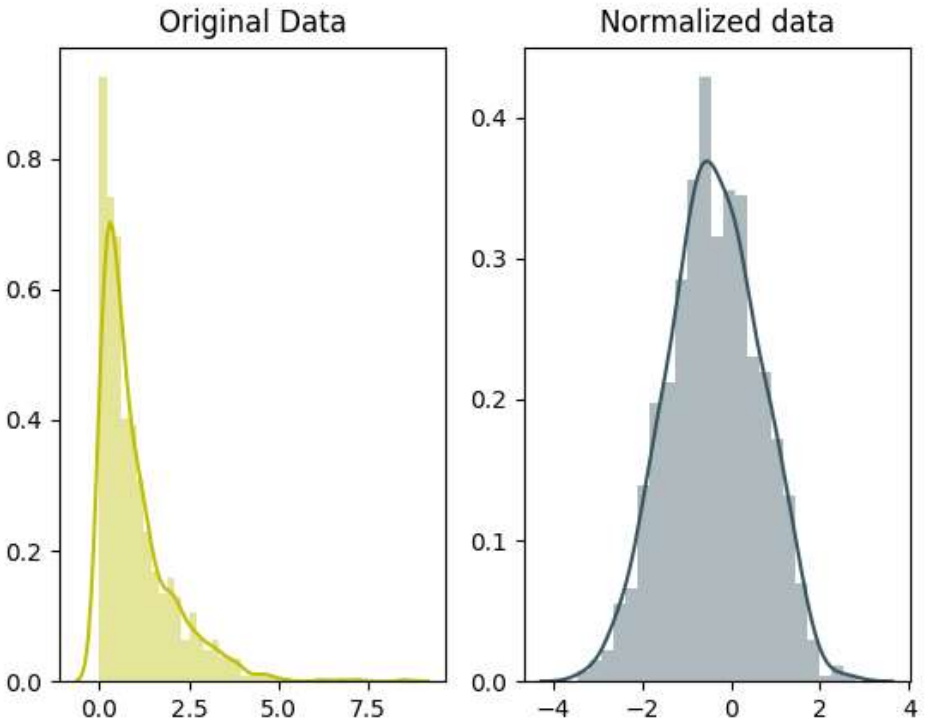
* **Estandarizar**→ Calculo la media de toda la columna, M. Calculo la desviación típica de toda la columna, S.
Ahora dato por dato, a cada uno le resto M y eso lo divido entre S

¿Qué busco? → Que los datos tengan media cero y desviación unitaria
Media cero → Ahora los datos están alrededor de 0 (positivos y negativos, pero en media valen entre todos cero)
Desviación típica unitaria→ Además de estar alrededor de 0, están cerca de 0 (varianza es 1 que es poco)
Conclusión: tengo datos pegados a 0 por arriba y por abajo

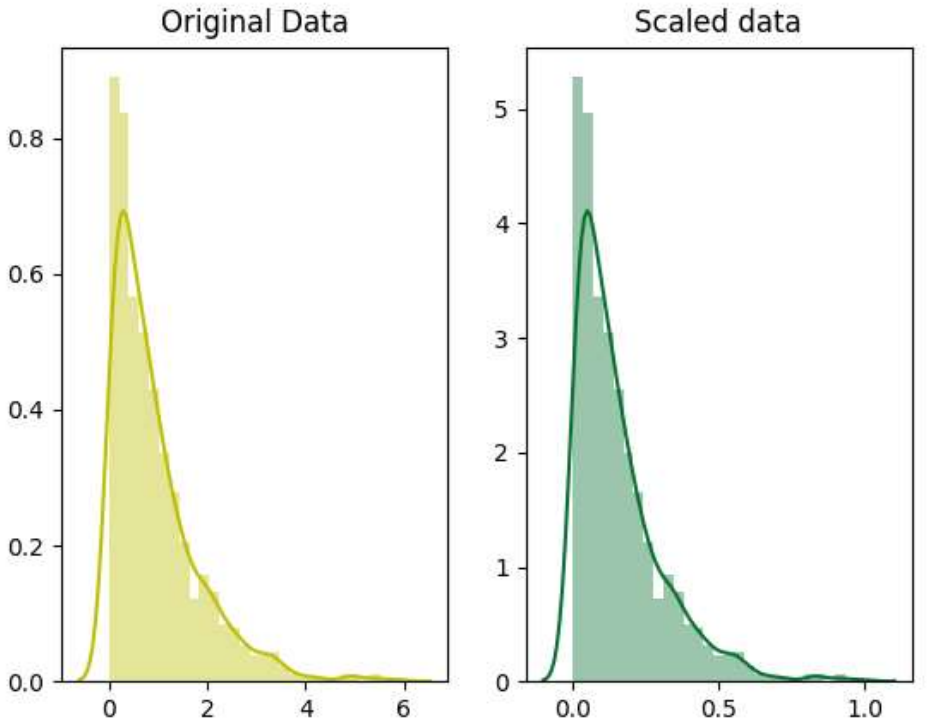
* **Escalar**→ Calculo el mínimo de toda la columna, Mín. Calculo el máximo de toda la columna, Máx. Calculo el rango de la columna, R = Máx - Mín
Ahora dato por dato, a cada uno le resto Mín y eso lo divido entre R

¿Qué busco? → Que los datos estén entre 0 y 1
Conclusión: tengo datos solo dentro del intervalo [0,1]

Estandarizar



Escalar



Ventajas:

- Son transformaciones REVERSIBLES, puedo volver a los datos originales
- Ahora todos los histogramas de todas las columnas tienen los datos
  - Estandarización→ cerca de 0 (positivos y negativos)
  - Escalado→ en el intervalo [0,1]

Ventajas:

- Puedo “comparar” entre columnas, ahora todos los datos tienen el mismo peso en un algoritmo.
- Si no hiciera esto, una columna con **números muy grandes “pesaría” mucho más que una con números pequeños y se llevaría el protagonismo aunque no fuese la más importante**
- La relación interna entre los datos y el tipo de flor se mantiene
- Realmente SOLO me hace falta estandarizar o escalar cuando después vaya a utilizar un modelo que se base EN DISTANCIAS (midiendo distancias entre puntos)
- Estandarizar o escalar entonces o **es NECESARIO** (mi algoritmo comparará distancias entre puntos) o **NO AFECTA** (mi algoritmo no comparará distancias entre puntos), pero no te va a estropear tu modelo si quieres comparar columnas, que es lo más común.
- Hay algoritmos que funcionan mejor cuando las distribuciones son Gaussianas. Ojo, estandarizar NO convierte automáticamente los datos en Gaussianos si en origen NO LO SON. Para esto hay transformaciones un poco más complejas que los mueven para ser gaussianos.
- Por último, “normalizar” se puede entender como un procesado por FILAS y NO COLUMNAS. Depende de la fuente que consultéis. **Es mejor usar siempre las palabras “estandarizar” y “escalar” para las columnas**

Y la pregunta: ¿estandarizar o escalar?

Depende de las operaciones que vayamos a realizar después (no hay una regla mágica)

- Me interesa más que la media sea cero y la desviación típica sea uno → Estandarizo
- Me interesa más no salirme del intervalo [0,1] → Escalo
- Caso frecuente: si en la columna origen la distribución es **Gaussiana**, suelo **estandarizar** y tener en la columna transformada una Gaussiana de media nula y desviación típica unitaria **G(mu=0,sigma=1)**