

¿Cambió la medición o se repitió el envío?

Un sensor ficticio mide CO2 cada diez segundos. La red puede entregar una misma lectura más de una vez. Cada fila de la tabla representa una entrega; `reading_id` identifica la medición original.

<code>reading_id</code>	Medición (s)	CO2 (ppm)	Entrega (s)
r1	0	600	1
r2	10	650	11
r3	20	700	21
r4	30	750	31
r4	30	750	32
r4	30	750	33

Trabaja con la tabla

1. Marca las entregas que repiten un `reading_id` anterior. ¿Cuántas entregas hay? ¿Cuántas mediciones distintas?
2. Calcula la media de CO2 con las seis filas y luego con una fila por medición. Muestra las dos cuentas.
3. ¿Cuál de esas medias responde a «promedio de las mediciones»? Explica por qué la otra cambia.
4. Dibuja las cuatro mediciones distintas frente al segundo en que se tomaron. ¿Qué comprobarías antes de tomar una decisión real sobre el aire de una sala?

Para quienes ya usan Pandas

Si las seis filas tienen una hora de entrega diferente, ¿por qué `drop_duplicates()` sobre la fila completa no elimina los reenvíos? ¿Qué columna usarías como clave?

Todas las cifras son inventadas para aprender. No describen un aula real ni permiten evaluar riesgos para la salud.

Qué debe distinguir el grupo

El objetivo es separar una medición de sus entregas. Se puede resolver en papel antes de usar CSV, Pandas o SQL. El ejemplo está pensado como calentamiento breve, no como sustituto de un conjunto de datos real.

Clave de respuestas

1. Hay seis entregas y cuatro mediciones. Las filas quinta y sexta repiten r4.
2. Media por entrega: $(600 + 650 + 700 + 750 + 750 + 750) / 6 = 700$ ppm. Media por medición: $(600 + 650 + 700 + 750) / 4 = 675$ ppm.
3. Para «promedio de las mediciones», la segunda responde a la pregunta. Contar reenvíos como lecturas nuevas da peso extra a la medición más alta.
4. Los puntos distintos son (0, 600), (10, 650), (20, 700) y (30, 750). Antes de interpretar datos reales habría que revisar calibración, ubicación, intervalos, lecturas ausentes y qué puede indicar el CO2. Cuatro puntos no establecen el promedio temporal de toda una sala.

Extensión opcional con Pandas

Si df contiene las seis filas y la hora de entrega cambia en cada una:

```
df.drop_duplicates()["co2_ppm"].mean() # 700
```

```
df.drop_duplicates(subset=["reading_id"])[ "co2_ppm"].mean() # 675
```

La clave estable `reading_id` permite identificar reenvíos en este ejemplo de un solo sensor. Con varios sensores, la clave tendría que distinguir también el dispositivo. Pide al grupo que justifique la clave antes de escribir la operación de limpieza.

Uso y autoría

JR creó Telemetry. La versión web en telemetry.sh/learn/sensor-retries?lang=es incluye el mismo CSV y una consulta SQL en el navegador sin cuenta. La actividad en papel no necesita la nube. Si se plantea usarla con alumnado, el profesorado debe revisar primero la política del centro y emplear solo datos ficticios.

Preparé este borrador con ayuda de IA. Aún no se ha probado en clase; revisa y adapta el material antes de enseñarlo.