El registro de imputación

¿Qué escribimos en las celdas vacías de `games`?

Nombres _____

Superset → **SQL → SQL Lab**, base de datos `Dades classe`, esquema `public`, tabla `games`. Una celda vacía no se puede limpiar: solo podéis decidir qué hacer con ella. Eso es la **imputación**. En parejas, uno teclea y el otro escribe; cambiad en la Tarea 3. Hoy no se modifica nada: todo son `SELECT`. Ejecutad **una sentencia cada vez**: seleccionadla y **Run**.

Referencia · el cuaderno de calificaciones

Opción	En el cuaderno	Es razonable cuando...
Una constante	Escribir un 0.	**Sabéis** qué significa el vacío: el trabajo no se entregó.
Un valor central	Escribir la media de la clase (o la mediana).	Hay pocos vacíos, y quienes los tienen no tienen nada de especial.
Dejarlo vacío + indicador	No escribir nada, añadir una nota «no evaluado».	No lo sabéis, o la mayor parte de la columna está vacía.

`COUNT(*)` cuenta filas · `COUNT(col)` cuenta las filas donde `col` **no** está vacía · `COALESCE(a, b)` devuelve el primer argumento que no es `NULL`: «a, y si está vacío, b».

Tarea 1 · Medid el hueco

```sql
SELECT COUNT(*)                 AS total_rows,
       COUNT(price)             AS with_value,
       COUNT(*) - COUNT(price)  AS missing,
       ROUND(100.0 * (COUNT(*) - COUNT(price)) / COUNT(*), 1) AS pct_missing
FROM games;
```

Repetidla con las otras tres columnas. Es vuestro **registro**: ahora dos columnas; las otras dos, al acabar cada tarea.

Columna	Vacías	% de filas	Opción (fila del cuaderno)	Por qué, en una frase
`price`				
`metascore`				
`developer`				
`publisher`				

Tarea 2 · `price`: ¿vacío significa gratis? Primero, las pruebas

```sql
SELECT COUNT(*) FROM games WHERE price = 0;        -- how many games cost exactly 0?
SELECT MIN(price) FROM games;                      -- the cheapest price on record
SELECT game, owners FROM games
WHERE price IS NULL ORDER BY number LIMIT 10;      -- who are they?
```

Juegos que cuestan exactamente 0 _____ Precio mínimo _____ Dos que conocéis de la lista: _____

Entonces, ¿qué significa un precio vacío en esta tabla? _____

Ahora rellenadlo y dejad una nota al margen: el **indicador** recuerda qué valores estaban vacíos.

```sql
SELECT game, COALESCE(price, 0) AS price, (price IS NULL) AS is_free
FROM games WHERE game IN ('Half-Life 2', 'Dota 2', 'Portal') ORDER BY game;
```

`Dota 2`: precio _____ `is_free` _____ `Half-Life 2`: precio _____ `is_free` _____

Sin `is_free`, ¿cómo distinguiríais un 0 real de uno rellenado? _____

Dos medias de la misma columna:

```sql
SELECT ROUND(AVG(price)::numeric, 2)              AS avg_ignoring_null,
       ROUND(AVG(COALESCE(price, 0))::numeric, 2) AS avg_free_as_zero
FROM games;
```

`avg_ignoring_null` _____ `avg_free_as_zero` _____ ¿Entre cuántas filas divide cada una? _____

¿Cuál publicaríais, y para qué pregunta? _____

Tarea 3 · `metascore`: intentad rellenarla y mirad los daños

La puntuación media es 71.9. Rellenad los vacíos con 72 y contad.

```sql
SELECT COUNT(*) FROM games WHERE metascore = 72;                -- real 72s
SELECT COUNT(*) FROM games WHERE COALESCE(metascore, 72) = 72;  -- after filling
```

72 reales _____ 72 tras rellenar _____ ¿Cómo sería un gráfico de la columna rellenada? _____

Ahora mirad **a quién** le falta. Primero, los juegos pequeños:

```sql
SELECT (owners LIKE '0%')  AS under_20k_owners,
       COUNT(*)            AS games,
       COUNT(metascore)    AS with_score,
       ROUND(100.0 * COUNT(metascore) / COUNT(*), 1) AS pct_with_score
FROM games GROUP BY 1;
```

Juegos pequeños con puntuación _____ % El resto _____ % ¿Los vacíos se reparten al azar? _____

¿Por qué 72 es una mentira para un juego que nadie ha reseñado? _____

La columna que os quedáis en su lugar: `(metascore IS NOT NULL) AS has_metascore`.

Tarea 4 · `developer` y `publisher`: una etiqueta para el vacío

De una columna de texto no se puede hacer la media. Dad al vacío una categoría propia, para que `GROUP BY` lo muestre en lugar de esconderlo.

```sql
SELECT COALESCE(developer, 'Unknown') AS developer, COUNT(*) AS games
FROM games GROUP BY 1 ORDER BY games DESC LIMIT 3;
```

Primera fila _____ con _____ juegos. ¿Es el desarrollador más grande de la tabla?

¿Por qué es una trampa? _____

Habéis acabado cuando Todos los espacios en blanco están rellenados · el registro de la Tarea 1 tiene sus cuatro líneas. **¿Pareja rápida?** El problema contrario: `average_playtime` es 0 en la mayoría de los juegos, porque nadie lo midió. `NULLIF(x, 0)` vuelve a convertir un 0 en `NULL`: comparad `AVG(average_playtime)` con `AVG(NULLIF(average_playtime, 0))`.

¿Atascados?

- **`pct_missing` muestra 0**: división entera. Escribid `100.0 *`, con el decimal, antes de dividir.
- **`COALESCE types text and integer cannot be matched`**: el sustituto debe tener el tipo de la columna.
- **`function round(double precision, integer) does not exist`**: añadid `::numeric` antes de redondear, como en la Tarea 2.

En casa En vuestro conjunto de datos abierto, elegid la columna con más celdas vacías. Medidla (filas y %), elegid una de las tres opciones y escribid el `COALESCE`, o explicad por qué la dejáis vacía.