# El inventario de suciedad

## Mirad antes de limpiar: ¿qué le pasa a `games`?

Nombres _____

Superset → **SQL → SQL Lab**, base de datos `Dades classe`, esquema `public`, tabla `games`: el cofre del botín que subisteis el lunes, un juego por fila. Por parejas, uno teclea y el otro escribe; cambiad en la Tarea 3. Ejecutad **una sentencia cada vez**: seleccionadla y después **Run**. Hoy no arregláis nada: abrís el cofre, contáis lo que está mal y lo apuntáis.

## Referencia · cuatro tipos de suciedad

| Tipo | En el cofre | En una tabla | Cómo encontrarlo |
|---|---|---|---|
| **Valor ausente** | Un hueco vacío. | Un `NULL`: ningún valor. | `COUNT(*) - COUNT(col)` |
| **Duplicado** | El mismo objeto dos veces. | La misma cosa real en dos filas. | `GROUP BY` + `HAVING COUNT(*) > 1` |
| **Tipo o formato incorrecto** | Una espada guardada como foto de una espada. | Una fecha o un número guardado como texto. | Mirad el tipo; ordenadla; `LIKE` |
| **Valor atípico** | Un objeto con 9,999 de daño. | Un valor muy alejado de todos los demás. ¿Un error, o real? | `MIN`, `MAX`, `ORDER BY ... DESC` |

`COUNT(*)` cuenta filas · `COUNT(col)` cuenta las filas donde `col` **no** está vacía · en un patrón `LIKE`, `%` es cualquier texto y `_` es exactamente un carácter.

## Tarea 1 · Tamaño y tipos

```sql
SELECT COUNT(*) FROM games;
SELECT column_name, data_type FROM information_schema.columns
WHERE table_name = 'games' ORDER BY ordinal_position;
```

Filas _____ Columnas _____ Dos columnas son `text` y no deberían serlo: _____ y _____

## Tarea 2 · Valores ausentes

Un `COUNT(*) - COUNT(col)` por columna, las diez en un solo `SELECT`. Empieza así:

```sql
SELECT COUNT(*) - COUNT(price) AS no_price, COUNT(*) - COUNT(game) AS no_game   -- ...and eight more
FROM games;
```

`number` _____ `game` _____ `release_date` _____ `price` _____ `owners` _____

`developer` _____ `publisher` _____ `average_playtime` _____ `median_playtime` _____ `metascore` _____

La peor columna es _____ : vacía en el _____ % de las filas.

Ahora los valores ausentes disfrazados:

```sql
SELECT COUNT(*) FROM games WHERE average_playtime = 0;
SELECT COUNT(*) FROM games WHERE price = 0;
SELECT game, price FROM games WHERE game = 'Dota 2';
```

`average_playtime = 0` en _____ filas. ¿Nadie jugó, o nadie lo midió? _____

`price = 0` en _____ filas. Dota 2 es gratuito. Entonces, ¿qué significa un precio vacío? _____

## Tarea 3 · Formatos incorrectos

Esta consulta debería listar los tres juegos más antiguos. Ejecutadla.

```
SELECT game, release_date FROM games ORDER BY release_date LIMIT 3;
```

Primera fecha _____ Segunda fecha _____ Tercera fecha _____

¿Sale primero el más antiguo? ¿Por qué la base de datos las ha ordenado así? _____

La mayoría de las fechas son como `Nov 16, 2004`. Buscad las que no lo son y después mirad `owners`:

```
SELECT release_date, COUNT(*) AS n FROM games
WHERE release_date NOT LIKE '___ %, ____'       -- 3 letters, space, day, comma, 4 digits
GROUP BY release_date ORDER BY n DESC;
SELECT owners, COUNT(*) AS n FROM games GROUP BY owners ORDER BY n DESC;
SELECT game, owners FROM games ORDER BY owners DESC LIMIT 3;
```

¿Qué les falta a la mayoría de las fechas raras? _____ Una es rara de otra manera: _____

Filas con una fecha rara (`SELECT COUNT(*)` con el mismo `WHERE`) _____ Valores distintos de `owners` _____

La tercera consulta debería dar los juegos con más propietarios. Dice _____

¿Por qué es incorrecto? _____

## Tarea 4 · Duplicados: ¿hay algún juego dos veces en la tabla?

```
SELECT game, COUNT(*) AS n FROM games WHERE game IS NOT NULL
GROUP BY game HAVING COUNT(*) > 1 ORDER BY n DESC, game;
SELECT number, game, release_date, developer FROM games WHERE game = 'Dark Matter';
SELECT COUNT(*) AS all_rows, COUNT(DISTINCT number) AS numbers FROM games;
```

Nombres que aparecen más de una vez _____ Filas que se llaman `Dark Matter` _____

¿Son el mismo juego? ¿Cómo lo sabéis? _____

`all_rows` _____ `numbers` _____ ¿Es `number` un identificador? _____

## Tarea 5 · Valores atípicos (los precios están en dólares de EE. UU.)

```
SELECT MIN(price), MAX(price), ROUND(AVG(price)::numeric, 2) AS mean FROM games;
SELECT game, price, developer FROM games ORDER BY price DESC NULLS LAST LIMIT 8;
```

Mín. _____ Máx. _____ Media _____ La fila más cara: _____

¿Es un error de escritura? Vuestra razón: _____

**Habéis acabado cuando** Todos los huecos están rellenos y el inventario, en una hoja aparte, tiene al menos ocho filas: columna, tipo de suciedad, evidencia (la consulta y su número). Dejad **Decisión** vacía: la rellenaremos juntos. **¿Pareja rápida?** Haced la Tarea 2 con vuestro conjunto de datos abierto y añadid sus tres peores columnas.

**¿Atascados?**

- **`relation "games" does not exist`**: comprobad el esquema, tiene que ser `public`.
- **Solo un resultado para tres consultas**: SQL Lab muestra el último. Ejecutadlas de una en una. **Solo salen 1,000 filas**: es el límite de visualización, no la tabla. Contad con `COUNT(*)`.
- «Hay NULLs» no es evidencia. Sin número, no hay evidencia. ¿Seguís atascados tras dos intentos? Preguntad a los vecinos y después al profesor.

**En casa** Acabad el inventario con su columna **Decisión**. Después, tres líneas de inventario para vuestro conjunto de datos abierto: ¿alguno de sus `NULL` significa algo?