# Histograma y dispersión

**Reparto de precios, y precio frente a metascore** Nombre _____

Superset → **Charts**, dataset `games_clean` (base de datos `Dades classe`, esquema `public`): 26,685 juegos de Steam. Tened **SQL → SQL Lab** abierto en otra pestaña. Pensad en el modo clasificatorio: un **histograma** cuenta los jugadores de cada rango; un **diagrama de dispersión** pone un punto por jugador. Trabajad solos y escribid aquí qué veis.

## Referencia

| | Gráfico de barras | Histograma |
|---|---|---|
| **Eje X** | Categorías que ya existen. | Columna **numérica** cortada en **bins** iguales. |
| **Altura** | Cualquier métrica. | Siempre un recuento: las filas dentro del bin. |
| **Responde** | ¿Qué categoría es mayor? | ¿Qué forma tienen los valores? |

| Nube de puntos | Lectura |
|---|---|
| Sube hacia la derecha | Relación positiva. |
| Baja hacia la derecha | Relación negativa. |
| Mancha redonda | Sin relación clara. |

Una relación nunca dice cuál de las dos cosas causa la otra.

## Parte A · El histograma de precios (18 min)

**Distribución de precios.** ¿Dónde están la mayoría de los juegos, y cómo de larga es la cola?

| Paso | Haced esto | Deberíais ver / escribid |
|---|---|---|
| 1 | Gráfico nuevo sobre `games_clean`, tipo **Histogram**. No está en la primera lista (**Featured**): clic en **All charts**, o escribid `histogram` en el buscador. **Column**: `price`. | **Column**, **Bins** y **Row limit**. |
| 2 | **Row limit**: cambiad `10000` por `50000`. Si no, el gráfico lee solo 10,000 de vuestros juegos. | `50000` en **Row limit**. |
| 3 | Dejad **Bins** en 5. **Create chart**. Pasad el ratón por la primera barra. | Barras que se ven de verdad: _____<br>Rango de la primera barra: _____ |
| 4 | En SQL Lab:<br>`SELECT game, price FROM games_clean`<br>`ORDER BY price DESC LIMIT 6;` | El más caro: _____<br>Su precio: _____ ¿Son juegos? _____ |
| 5 | De vuelta en el gráfico: **Filters** → `price` → operador **Less or equal (<=)** → escribid `60` y elegidlo en la lista → **Save**. | Un chip `price <= 60` bajo **Filters**. |
| 6 | **Bins**: escribid `12` y pulsad Enter. **Update chart**. | Barras: _____ La más alta: la _____<br>La cola va hacia la _____ |
| 7 | **Customize** → **Y Axis Format**: `,d`. Volved a la primera barra. | Recuento exacto, sin redondear: _____ |
| 8 | **Save** → `UT3 · Price distribution`. | Aparece en la lista de **Charts**. |
| 9 | Opcional. Marcad **Cumulative**, **Update chart**: cada barra suma todas las anteriores. No lo guardéis. | Proporción por debajo de 10: _____ |

**A1** Paso 3: los bins son tramos iguales desde la fila más barata hasta la más cara. ¿Por qué una sola barra gigante?

**A2** Leed vuestro histograma final en una frase que contenga un número: «La mayoría de los juegos…».

**A3** Habéis dejado fuera los productos de más de 60. ¿Son datos erróneos? ¿Qué debería decir el título del gráfico?

## Parte B · Precio frente a metascore (16 min)

**Precio frente a metascore.** ¿Reciben mejores reseñas los juegos más caros?

| Paso | Haced esto | Deberíais ver / escribid |
|---|---|---|
| 1 | En SQL Lab:<br>`SELECT has_metascore, COUNT(*) FROM games_clean GROUP BY has_metascore;` | Con metascore: _____ de 26,685<br>Aproximadamente, en %: _____ |
| 2 | Gráfico nuevo sobre `games_clean`, tipo **Scatter Plot**. | Un gráfico vacío en la vista Explore. |
| 3 | **X-axis**: `price`. **Metrics**: pestaña **Simple** → columna `metascore`, agregado `AVG`. | Los dos controles rellenos. |
| 4 | **Filters** → `has_metascore` → operador **Is true** → **Save**. **Create chart**. | Una nube de puntos entre 0 y 60 en el eje X. |
| 5 | Pasad el ratón por tres puntos: el de 9.99, el de 59.99 y el más bajo. | En 9.99: _____<br>En 59.99: _____<br>El más bajo: _____ a precio _____ |
| 6 | **Save** → `UT3 · Price vs metascore`. | Aparece en la lista de **Charts**. |

**B1** Muchos juegos cuestan 9.99, pero solo hay un punto en 9.99. Mirad vuestro control **Metrics**: ¿qué representa un punto?

_____

**B2** ¿La nube sube, baja, o ninguna de las dos cosas? Usad la referencia.

_____

**B3** ¿De qué puntos os fiaríais menos, y por qué?

_____

**B4** Un amigo lee vuestro gráfico y dice: «En Steam, subir el precio de un juego mejora sus reseñas.» Encontrad los dos errores. El paso 1 ayuda con uno.

1. _____

2. _____

**Habéis acabado cuando** Todos los huecos están rellenos · hay dos gráficos guardados: `UT3 · Price distribution` y `UT3 · Price vs metascore`.

**¿Atascados?**

- **Las barras suman 10,000**: **Row limit** sigue en 10,000. Poned 50,000.
- **`price` no aparece en** Column: el control solo lista columnas numéricas. Avisad al profesor.
- **Sigue habiendo una barra gigante después del filtro**: el chip `price <= 60` debe aparecer bajo **Filters**, y tenéis que pulsar **Update chart**.
- **«Must specify a value for filters with comparison operators»**: el chip dice `price <=` sin ningún número. Haced clic en él, escribid `60`, elegidlo en la lista, **Save**.
- **Dispersión con un solo punto**: `price` ha ido a **Metrics** en lugar de **X-axis**.
- **Puntos repartidos por igual, como barras**: **Force categorical** está marcado bajo el eje X. Desmarcadlo.
- **Superset no se abre**: poneos con un compañero y avisad al profesor.

**En casa** Nada obligatorio en vacaciones. Opcional, para el proyecto final: con vuestro conjunto de datos abierto, un histograma de una columna numérica. ¿También tiene valores atípicos? Los gráficos `UT3 ·` guardados son vuestros apuntes para enero.