Sembla igual, no ho és

`owners_min` i la caça de duplicats a `games` Noms _____

Superset → **SQL → SQL Lab**, base de dades `Dades classe`, esquema `public`, taula `games`, i la vostra consulta desada `games_step1`. Dos objectes que semblen idèntics i no s'apilen: alguna cosa amagada els fa diferents. En parelles, un sol teclat; canvieu qui tecleja en acabar la Part A. Executeu **una sentència cada vegada**: seleccioneu-la i premeu **Run**.

Referència

Eina	Què fa
`ASCII(c)`	El número d'un caràcter: la seva empremta.
`SPLIT_PART(x, sep, n)`	Talla `x` a cada `sep` i retorna el tros `n`.
`REGEXP_REPLACE(x, p, new, 'g')`	Substitueix **totes** les coincidències del patró `p`. `[^0-9]` vol dir "qualsevol cosa que no sigui una xifra".
`SELECT DISTINCT *`	Conserva una còpia de cada fila totalment idèntica.
`ROW_NUMBER() OVER (PARTITION BY k ...)`	Numera les files 1, 2, 3… dins de cada grup. El grup `k` és la **clau**: les columnes que diuen "això és la mateixa cosa".

Números de caràcter: 32 l'espai que escriu el teclat · 44 la coma · 46 el punt · 160 l'**espai dur**, que sembla exactament un espai. En una pàgina web s'escriu ` ` i vol dir "no trenquis la línia aquí".

Part A · D'un interval en text a un nombre (16 min)

`owners` diu `10,000,000 .. 20,000,000`. Voleu el primer nombre, com a nombre de debò. Tallar a `' .. '` ha fallat.

A1 · L'empremta. Agafeu el caràcter de just abans dels punts i demaneu-ne el número.

```sql
SELECT ASCII(SUBSTR(owners, POSITION('..' IN owners) - 1, 1)) AS mystery,
       ASCII(' ') AS normal_space
FROM games WHERE game = 'Half-Life 2';
```

`mystery` _____ `normal_space` _____ Per què `' .. '` no coincidia mai? _____

A2 · Talleu per on us en podeu fiar. Els dos punts són caràcters normals. Talleu allà i mesureu.

```sql
SELECT SPLIT_PART(owners, '..', 1) AS part,
       LENGTH(SPLIT_PART(owners, '..', 1)) AS len, LENGTH('10,000,000') AS expected
FROM games WHERE game = 'Half-Life 2';
```

`len` _____ `expected` _____ D'on surt la diferència? _____

A3 · Aquesta falla a propòsit. Eines de la setmana passada: `TRIM` l'espai, `REPLACE` les comes, convertir.

```sql
SELECT REPLACE(TRIM(SPLIT_PART(owners, '..', 1)), ',', '')::bigint FROM games;
```

L'error diu: _____

Mireu dins de les cometes, després de les xifres. Què hi queda, i per què `TRIM` no ho ha tret? _____

A4 · Quedeu-vos només amb les xifres. Comes, espais de qualsevol mena, el que vingui l'any que ve: tot fora.

```sql
SELECT REGEXP_REPLACE(SPLIT_PART(owners, '..', 1), '[^0-9]', '', 'g')::bigint AS owners_min,
       COUNT(*) AS n
FROM games GROUP BY owners_min ORDER BY owners_min;
```

Files _____ El més petit _____ El més gran _____ Jocs amb el valor més petit _____
El valor més petit vol dir que ningú té aquests jocs? Què mesura realment `owners_min`?

_____

A5 · Afegiu-ho a la consulta neta. Obriu `games_step1` (**SQL → Saved Queries**, clic al nom). Al bloc `WITH`, substituïu l'`owners` tot sol per l'expressió d'A4 `AS owners_min`, i canvieu el nom del bloc a `step2`. Comproveu-ho amb:

```
SELECT game, owners_min FROM step2 ORDER BY owners_min DESC LIMIT 4;
```

Deseu-la com a `games_step2`: **Save**, canvieu el nom, **Save as new** (**Update** sobreescriuria `games_step1`).

El joc de dalt de tot: _____ Els tres següents: _____

Part B · La caça de duplicats (18 min)

74 noms surten més d'un cop a `games`. Abans d'esborrar res, tres casos. **Duplicat real**: totes les columnes idèntiques, la mateixa fila dos cops (un fitxer carregat dos cops); elimineu les còpies. **Duplicat aparent**: mateix nom, un altre desenvolupador i data, com els tres `Dark Matter`; conserveu-los tots. **Zona grisa**: mateix nom **i** desenvolupador, dues dates, com `Penumbra Overture`; una decisió, per escrit.

B1 · `games` té duplicats reals?

```
SELECT COUNT(*) FROM games;
SELECT COUNT(*) FROM (SELECT DISTINCT * FROM games) AS t;
```

Primer recompte _____ Segon recompte _____ Duplicats reals, sí o no? _____

B2 · La doble càrrega. `UNION ALL` enganxa dos resultats un sota l'altre: prémer Upload dos cops sense tocar la taula.

```
WITH twice AS (SELECT * FROM games UNION ALL SELECT * FROM games)
SELECT COUNT(*) FROM twice;
```

Files a `twice` _____. Conserveu la línia `WITH` i compteu `SELECT DISTINCT * FROM twice`, embolicat com a B1.

Files ara _____

B3 · El mateix amb una clau. Quedar-se només amb `WHERE rn = 1` eliminaria les còpies.

```
WITH twice AS (SELECT * FROM games UNION ALL SELECT * FROM games),
numbered AS (
  SELECT *, ROW_NUMBER() OVER (PARTITION BY game, release_date, developer ORDER BY number) AS rn
  FROM twice
)
SELECT rn, COUNT(*) FROM numbered GROUP BY rn ORDER BY rn;
```

Files amb `rn = 1` _____ Files amb `rn = 2` _____

B4 · La clau descuidada. "Un joc és el seu nom": useu només `PARTITION BY game`, sobre la taula **original** (`FROM games`, sense `twice`).

Files amb `rn = 1` _____ Files que un `WHERE rn = 1` esborraria _____ Un joc que perdríeu: _____

B1 deia que no hi ha duplicats reals. Aleshores, què ha esborrat aquesta neteja? _____

Heu acabat quan Tots els espais plens · `games_step2` és a les consultes desades. **Parella ràpida?** Construïu `owners_max` amb el tros 2 del tall. Al vostre conjunt de dades obert, alguna columna és una clau de debò?

Encallats?

- **`invalid input syntax for type bigint`** a A4, amb una coma dins de les cometes: falta el `'g'` a `REGEXP_REPLACE`.
- **El resultat diu `1k rows`**: és el límit de visualització. Emboliqueu-la amb `SELECT COUNT(*) FROM (...) AS t`.
- **`games_step1` no és a Saved Queries**: demaneu-ne l'SQL al professor.

A casa Al vostre conjunt de dades obert, quines columnes juntes identifiquen una fila? Demostreu-ho amb un `GROUP BY ... HAVING COUNT(*) > 1` que no retorni res.