The descriptive report

The store page of `games_clean` Names _____

Superset → **SQL → SQL Lab**, database `Dades classe`, schema `public`, table `games_clean`. You write the report in a document (ask the teacher for the template): six headed parts, **250 words maximum**. Pairs: one runs queries, one writes; swap at part 4. This sheet is for your notes and for the review.

Reference · the six parts

A descriptive report is the store page of a dataset: you read it before you "install" the data in a chart.

Part	Question it answers	On the store page
1 · Source	Who collected it, where, when? Who cleaned it? From an API: endpoint, parameters, extraction date.	Developer and release date.
2 · Size	How many rows and columns? What is one row? What period?	Download size.
3 · Variables	What are the important columns, their type and unit?	Features list.
4 · Quality	What is missing, estimated or doubtful? How much?	Known issues.
5 · Key figures	Centre, spread and distribution of what matters.	Review score and tags.
6 · Limits	What can this data **not** tell you?	System requirements: where it will not run.

Task 0 · Two reports, same data (4 min)

Report A	Report B
"The dataset has a lot of games. Most are cheap. The average price is 7.91. Some scores are missing. Games are getting more popular."	"26 685 Steam games released 2004–2018. Half cost 4.99 USD or less (median); the mean is 7.91 USD because of a few items up to 595.99 USD. Only 10.7% have a Metascore."

0.1 Which one would you trust? Underline the exact words that make the difference, then name three things it has that the other lacks.

0.2 One sentence of the other report says something the data cannot prove. Which one, and why?

Task 1 · Parts 1 to 3 (7 min)

Source and variables come from your UT2 notes. Size needs one query:

```sql
SELECT COUNT(*) AS n_games, MIN(release_date) AS first_release,
       MAX(release_date) AS last_release
FROM games_clean;
```

Columns: _____ First release: _____ Last release: _____ One row is: _____

Task 2 · Parts 4 and 5 (11 min)

Reuse last Wednesday's queries (your stat card). The pattern for a distribution with percentages:

```sql
SELECT price_band, COUNT(*) AS n,
       ROUND(100.0 * COUNT(*) / SUM(COUNT(*)) OVER (), 1) AS pct
FROM games_clean GROUP BY price_band ORDER BY n DESC;
```

The minimum, each with its **unit** (USD, games, %):

2.1 One missing-data figure with its percentage: _____

2.2 Centre and spread of `price`: _____

Which centre do you print, and why? _____

2.3 One distribution (`price_band` or `owners_tier`): _____

Task 3 · Part 6, no queries at all (5 min)

What would a journalist wrongly conclude from this table? Think of `owners_min`, `metascore` and `price`. Write two limits.

3.1 This data cannot tell... _____

3.2 Careful: _____

Peer review (11 min)

Give this sheet and your report to another pair. **Reviewers**: tick yes or no. Every "No" needs the line it refers to. No marks, no adjectives: only yes, no and where. Reviewed by _____

#	Check	Yes / No	If no: which line?
1	The six parts are there, in order, each with its heading.		
2	Source says who collected the data, from where, and which years it covers. For an API: endpoint, parameters and extraction date.		
3	Size gives rows **and** columns, and says what one row is.		
4	Every figure has a unit, and its query is attached.		
5	For `price`, the centre is the median (or both), and the text says why.		
6	Quality names at least two problems, each with a count or a percentage.		
7	Limits has at least two things the data cannot tell.		
8	No claim goes beyond the data. 250 words or fewer.		

R **Authors**, when the sheet comes back: fix **one** "No". What did you change?

_____

Done when Six headed parts, 250 words or fewer · every figure has a unit and its SQL pasted below the report · the checklist is filled in by the other pair · your one change is written here · the report is handed in.

Stuck?

- **`relation "games_clean" does not exist`**: ask the teacher for `games_clean.sql` and run it in SQL Lab, or pair up.
- **`function round(double precision, integer) does not exist`**: cast first, `ROUND(metascore::numeric, 1)`.
- **`SUM(COUNT(*)) OVER ()` gives an error**: the `GROUP BY` is missing.
- **Still running queries when half the time is gone?** Stop querying, start writing.

At home Finish the report if needed and hand it in. For Wednesday: reread your notes on `COALESCE`, `CASE WHEN`, `CREATE TABLE AS` and the six parts. Open the `pokemon` table and look at `type_2` and `generation_id`.