

How good the data is

fraud_200k.csv - run run_1037d177a96f - part of [REPORT.pdf](#)

The data as it arrived, what was wrong with it, every change made (with any code the AI wrote), the checks on the clean data and the review. The short version is in [REPORT.pdf](#), section 2.

1. The data as received

The file is loaded exactly as it is, every value as text. Nothing is changed at this step; it only describes what is there.

File	fraud_200k.csv
Format	csv, encoding utf-8, delimiter ','
Rows	200,000 in the file, 200,000 loaded, 0 rejected
Fingerprint	file d786623d69718fa8..., data a5c30c71c3f8d6cc...

How the file was interpreted

- Decoded as UTF-8 without errors.
- Delimiter ',' chosen: 11 fields on 100% of the first 50 records.
- Parsed 200000 records with delimiter ','.

Lines that could not be read

None: every line was read.

Column by column (dirty data spotted at load)

Column	Stored as	Read as	Fits	Missing	Distinct	What stands out
step	text	integer	100%	0.0%	501	e.g. 18, 19, 187
type	text	categorical	100%	0.0%	5	e.g. CASH_OUT, PAYMENT, CASH_IN
amount	text	float	100%	0.0%	198,490	2799 far outlier(s); e.g. 1.0E7, 15000.0, 10000.0
nameOrig	text	text	100%	0.0%	199,992	personal data (name), masked; e.g. A9999999999, A9999999999, A9999999999
oldbalanceOrig	text	float	100%	0.0%	98,147	14594 far outlier(s); e.g. 0.0, 186.0, 130.0
newbalanceOrig	text	float	100%	0.0%	86,462	13559 far outlier(s); e.g. 0.0, 277771.86, 12173.96
nameDest	text	text	100%	0.0%	174,594	personal data (name), masked; e.g. A9999999999, A9999999999, A9999999999
oldbalanceDest	text	float	100%	0.0%	114,756	7505 far outlier(s); e.g. 0.0, 1.0E7, 2.0E7
newbalanceDest	text	float	100%	0.0%	122,830	6705 far outlier(s); e.g. 0.0, 207142.75, 3158225.46
isFraud	text	boolean	100%	0.0%	2	e.g. 0, 1
isFlaggedFraud	text	boolean	100%	0.0%	2	e.g. 0, 1

What your brief claimed

- the target column 'isFraud' exists: holds. Column 'isFraud' is present.

Gate 1 checks

Check	Result	Details
unique_column_names	pass	11 column names are unique.
provenance_hashes_present	pass	File and data hashes recorded.
profile_matches_loaded_rows	pass	Profile describes 200,000 rows; 200,000 were loaded.
rows_accounted_for	pass	200,000 loaded + 0 rejected = 200,000 of 200,000 source rows.
reject_rate_within_threshold	pass	0 of 200,000 lines could not be parsed (0.00%); warn above 0.1%, block above 1%.
type_coverage_or_flagged	pass	11/11 columns have a type covering 95% of values; the rest are flagged.
brief_claims_hold	pass	All 1 checkable claims in the brief hold.

2. What was wrong

AI-written: The 200k-row PaySim sample is structurally clean: no missing values, no null tokens and no duplicate rows. Every column is stored as text, so the numeric and flag columns need casts. Following the brief, the four balance columns are marked excluded because fraudulent transactions were annulled. The main open questions are whether to keep extreme amounts and how to treat isFlaggedFraud.

Issues: clear errors with a clear fix

ID	Column	Kind	What was found	What should be done
I1	step	type	type 'integer', stored_as 'text', coverage 1.0, range [1.0, 239.0, 737.0]	Cast step to integer (hours since simulation start, 1-744). (op: cast)
I2	amount	type	type 'float', stored_as 'text', coverage 1.0, examples '1.0E7', '15000.0'; format decimal_separator '.', thousands_separator ','	Cast amount to float. Accept scientific notation such as '1.0E7'. If any values contain ',', thousands separators, remove them before casting so no value is lost. (op: cast)
I3	oldbalanceOrg	type	type 'float', stored_as 'text', coverage 1.0, examples '0.0', '186.0'	Cast to float. The column is excluded from the analysis, but a numeric type keeps it usable for reference. (op: cast)
I4	newbalanceOrig	type	type 'float', stored_as 'text', coverage 1.0, examples '0.0', '277771.86'	Cast to float. The column stays excluded from the analysis. (op: cast)
I5	oldbalanceDest	type	type 'float', stored_as 'text', coverage 1.0, examples '0.0', '1.0E7', '2.0E7'	Cast to float and accept scientific notation. The column stays excluded from the analysis. (op: cast)
I6	newbalanceDest	type	type 'float', stored_as 'text', coverage 1.0, examples '0.0', '207142.75'	Cast to float. The column stays excluded from the analysis. (op: cast)
I7	isFraud	type	type 'boolean', stored_as 'text', values ['0', '1'], missing 0.0%	Cast the target to boolean ('1' to True, '0' to False). Do not fill or alter any values. (op: cast)
I8	isFlaggedFraud	type	type 'boolean', stored_as 'text', values ['0', '1'], missing 0.0%	Cast to boolean ('1' to True, '0' to False). (op: cast)
I9	nameOrig	personal_data	pii level 'possible', kinds ['name'], masked true; distinct 199992 of 200000; shape 'A999999999'	No change. These are synthetic account codes acting as identifiers. Keep them as text, flagged as possible personal data.
I10	nameDest	personal_data	pii level 'possible', kinds ['name'], masked true; distinct 174594; length [4, 11]	No change. Keep as text identifier, flagged as possible personal data. The leading letter (C = customer, M = merchant) may be useful later for feature engineering.

There were also 3 judgment call(s), questions the data cannot answer by itself. They are in 3_decisions.pdf.

What each column should become

Column	Role	Clean type	Notes
step	feature	integer	Hour of the 30-day simulation (1 step = 1 hour).
type	feature	categorical	5 transaction types; fraud occurs in TRANSFER and CASH_OUT.
amount	feature	float	Contains scientific notation (e.g. '1.0E7'); keep outliers.
nameOrig	identifier	id	Possible personal data; synthetic account code.
oldbalanceOrg	excluded	float	Brief: not to be used for fraud analysis (fraud annulled).
newbalanceOrig	excluded	float	Brief: not to be used for fraud analysis.
nameDest	identifier	id	Possible personal data; prefix 'M' marks merchants.
oldbalanceDest	excluded	float	Brief: not to be used; not applicable for merchants.
newbalanceDest	excluded	float	Brief: not to be used; not applicable for merchants.
isFraud	target	boolean	Target; never impute.
isFlaggedFraud	feature	boolean	Rule-based flag for transfers over 200,000; see J3.

3. What was changed

AI-written: Cast the numeric and flag columns from text to their proper types (integer, float, boolean), with no other changes. All three decisions keep the data as it is, so they need no transformation.

Rows: 200,000 before, 200,000 after. Every step below ran on a copy, starting from the original data, and every change it made was measured; the original data is never modified.

Step 1: cast, for I1

AI-written: Hour counter stored as text; cast to integer.

- Changed: no values
- New type: step: VARCHAR -> BIGINT

Parameters: {"column": "step", "to": "integer", "format": null, "on_error": "error"}

Step 2: cast, for I2, J1

AI-written: Cast amount to float (scientific notation like '1.0E7' parses); no clipping, so extreme amounts are kept as decided.

- Changed: amount: 169 cell(s)
- New type: amount: VARCHAR -> DOUBLE

Parameters: {"column": "amount", "to": "float", "format": null, "on_error": "error"}

Step 3: cast, for I3, J2

AI-written: Cast to float; the column is kept (not dropped) and stays excluded from analysis.

- Changed: oldbalanceOrg: 4,467 cell(s)
- New type: oldbalanceOrg: VARCHAR -> DOUBLE

Parameters: {"column": "oldbalanceOrg", "to": "float", "format": null, "on_error": "error"}

Step 4: cast, for I4, J2

AI-written: Cast to float; the column is kept and stays excluded.

- Changed: newbalanceOrig: 4,591 cell(s)
- New type: newbalanceOrig: VARCHAR -> DOUBLE

Parameters: {"column": "newbalanceOrig", "to": "float", "format": null, "on_error": "error"}

Step 5: cast, for I5, J2

AI-written: Cast to float, accepting scientific notation; the column is kept and stays excluded.

- Changed: oldbalanceDest: 3,376 cell(s)
- New type: oldbalanceDest: VARCHAR -> DOUBLE

Parameters: {"column": "oldbalanceDest", "to": "float", "format": null, "on_error": "error"}

Step 6: cast, for I6, J2

AI-written: Cast to float; the column is kept and stays excluded.

- Changed: newbalanceDest: 3,717 cell(s)
- New type: newbalanceDest: VARCHAR -> DOUBLE

Parameters: {"column": "newbalanceDest", "to": "float", "format": null, "on_error": "error"}

Step 7: cast, for I7

AI-written: Cast the target from '0'/'1' to boolean without altering any value.

- Changed: isFraud: 200,000 cell(s)
- New type: isFraud: VARCHAR -> BOOLEAN

Parameters: {"column": "isFraud", "to": "boolean", "format": null, "on_error": "error"}

Step 8: cast, for I8, J3

AI-written: Cast the flag to boolean and keep it as a feature, as decided.

- Changed: isFlaggedFraud: 200,000 cell(s)
- New type: isFlaggedFraud: VARCHAR -> BOOLEAN

Parameters: {"column": "isFlaggedFraud", "to": "boolean", "format": null, "on_error": "error"}

Not done, with reasons

- J1: Decision is to keep all amounts as they are; no clipping step is needed. The amount cast preserves every value.
- J3: Decision is to keep isFlaggedFraud as a feature. The only change is the boolean cast under I8.

4. Checks on the clean data

Gate 2: **PASS**. The clean data matches the diagnosis and every check passed.

Check	Result	Details
schema_types_match_diagnosis	pass	All 11 column(s) have the type the diagnosis expects.
columns_match_diagnosis	pass	The clean columns are exactly the ones the diagnosis expects.
primary_key_intact	pass	No ID column is named in the brief or the diagnosis; nothing to check.
target_not_imputed	pass	None of the 0 missing target value(s) in kept rows was filled in.
row_loss_within_threshold	pass	0 of 200,000 rows dropped (0.0%); under 5% passes, 5-20% is flagged, over 20% blocks.
row_accounting	pass	200,000 raw = 200,000 clean + 0 dropped.
dropped_rows_have_reasons	pass	No rows were dropped.
no_unexplained_new_nulls	pass	No op introduced nulls without declaring it.
untouched_columns_unchanged	pass	3 untouched column(s) are byte-identical to raw.
cell_diff_accounted_for	pass	Every changed cell belongs to a column an op declared.
distribution_shift_within_threshold	pass	No large distribution shift in 8 numeric column(s).
replay_reproduces_output	pass	Replaying 8 op(s) from raw reproduces the output hash.

The clean data, column by column (the schema contract)

Column	Type	Expected	Role	Missing	Distinct	Note
step	BIGINT	integer	feature	0.0%	501	
type	VARCHAR	categorical	feature	0.0%	5	
amount	DOUBLE	float	feature	0.0%	198,490	
nameOrig	VARCHAR	id	identifier	0.0%	199,992	
oldbalanceOrg	DOUBLE	float	excluded	0.0%	98,147	
newbalanceOrig	DOUBLE	float	excluded	0.0%	86,462	
nameDest	VARCHAR	id	identifier	0.0%	174,594	
oldbalanceDest	DOUBLE	float	excluded	0.0%	114,756	
newbalanceDest	DOUBLE	float	excluded	0.0%	122,830	
isFraud	BOOLEAN	boolean	target	0.0%	2	
isFlaggedFraud	BOOLEAN	boolean	feature	0.0%	2	

5. The review

A second AI checked whether the cleaning was right, using its own counting queries on the data (never rows).

Round 1: accepted

AI-written: The cleaning is eight type casts with no rows dropped, no new missing values and no values altered. The target is intact, and all three AI-made decisions were carried out as decided. The only open points are two questions for the user: how useful isFlaggedFraud is, and how to handle the heavy class imbalance. Neither needs a fix.

What it verified, with the numbers that show it:

- The target isFraud was cast without any change to its values or counts. (Raw: '0' = 199760, '1' = 240. Clean: false = 199760, true = 240. Both have 200000 rows and 0 missing, and Gate 2 target_not_imputed passes.)
- No rows were lost and no new nulls appeared. (Every step shows rows 200000 -> 200000 (0 dropped) and missing 0 -> 0. Gate 2 passes row_accounting (200,000 = 200,000 + 0) and no_unexplained_new_nulls.)
- J1 (keep all amounts) was honoured: the amount cast did not clip or truncate anything, and the magnitudes are plausible. (The distinct count is unchanged at 198490, and the clean max is 73823490.36, which is large rather than capped. Quartiles are p25 13350.44, median 74183.06 and p75 208447.15. Scientific notation such as '1.0E7' parses to 10000000.0. The 169 changed cells are consistent with the text form being rewritten.)
- Step values stay inside the 744-hour simulation window. (The count of step > 744 is 0 of 200000. The step log shows min 1 and max 743, and the distinct count is unchanged at 501.)
- J2 (keep the four balance columns, marked excluded) was honoured. (All four columns remain in the clean table as DOUBLE with role 'excluded'. Their missing rate is 0 and the distinct counts are unchanged (98147, 86462, 114756, 122830). Gate 2 columns_match_diagnosis passes.)
- The type categories are clean and need no standardisation. (There are 5 distinct values with 0 missing: CASH_OUT 70084, PAYMENT 67660, CASH_IN 43991, TRANSFER 16898, DEBIT 1367. These sum to 200000.)
- The isFlaggedFraud cast kept the flag counts exactly. (Raw flagged '1' = 1 and clean True = 1, both out of 200000.)

Questions for you:

- Should isFlaggedFraud stay in the analysis as a feature? It is almost constant in this sample, so it adds little for describing fraud.
- How should the very low fraud rate (0.12%) be handled in the analysis? The cleaning left it as is, which is correct, but any later comparison or model needs to account for it.

Queries it ran

Query	What it wanted to know	Answer
q1	target distribution is intact after cast	q1 value_counts of isFraud (clean): 200000 rows, 0 missing, 2 distinct 'false': 199760 'true': 240
q2	compare raw target distribution	q2 value_counts of isFraud (raw): 200000 rows, 0 missing, 2 distinct '0': 199760 '1': 240
q3	zero amount transactions left	q3 count (clean, where amount equals 0): 1 of 200000 rows
q4	steps beyond the 744-hour span	q4 count (clean, where step greater than 744): 0 of 200000 rows
q5	category spellings	q5 value_counts of type (clean): 200000 rows, 0 missing, 5 distinct 'CASH_OUT': 70084 'PAYMENT': 67660 'CASH_IN': 43991 'TRANSFER': 16898 'DEBIT': 1367
q6	number flagged	q6 count (clean, where isFlaggedFraud equals True): 1 of 200000 rows
q7	raw flagged count matches clean	q7 count (raw, where isFlaggedFraud equals '1'): 1 of 200000 rows
q8	full-data amount range	q8 summary of amount (clean, DOUBLE): 200000 rows, missing 0, distinct 198490, min 0, max 73823490.36, mean 179525.1052, p25 13350.4425, median 74183.055, p75 208447.1475