

# Survey Data Cleaning

Standalone verified Python analysis

Method: Survey Data Cleaning

Method type: data\_cleaning\_workflow

Design: Deterministic audit of duplicate rows, missing cells, categorical domains, numeric ranges, and final retained rows.

Variables: school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, guardian, traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2, G3

Inference scope: No hypothesis test is required; data-quality rules determine whether rows or cells require correction.

Formula: duplicate\_i=I(row\_i appeared earlier); range violation=I(x<min or x>max); clean rows=distinct valid rows.

Source rows: 649

Calculated metrics:

input\_rows: 649

duplicate\_rows: 0

missing\_cells: 0

category\_violations: 0

range\_violations: 0

clean\_rows: 649

rows\_removed: 0

Independent reference values:

input\_rows: 649

duplicate\_rows: 0

missing\_cells: 0

category\_violations: 0

range\_violations: 0

clean\_rows: 649

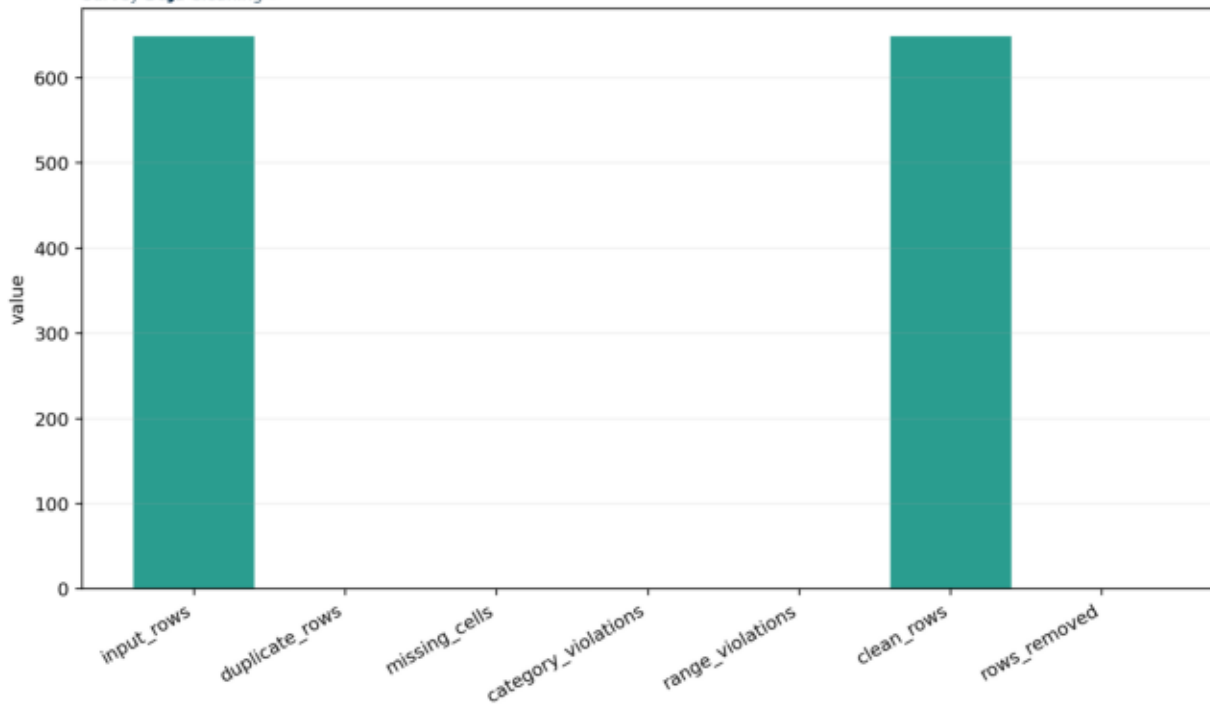
rows\_removed: 0

Interpretation:

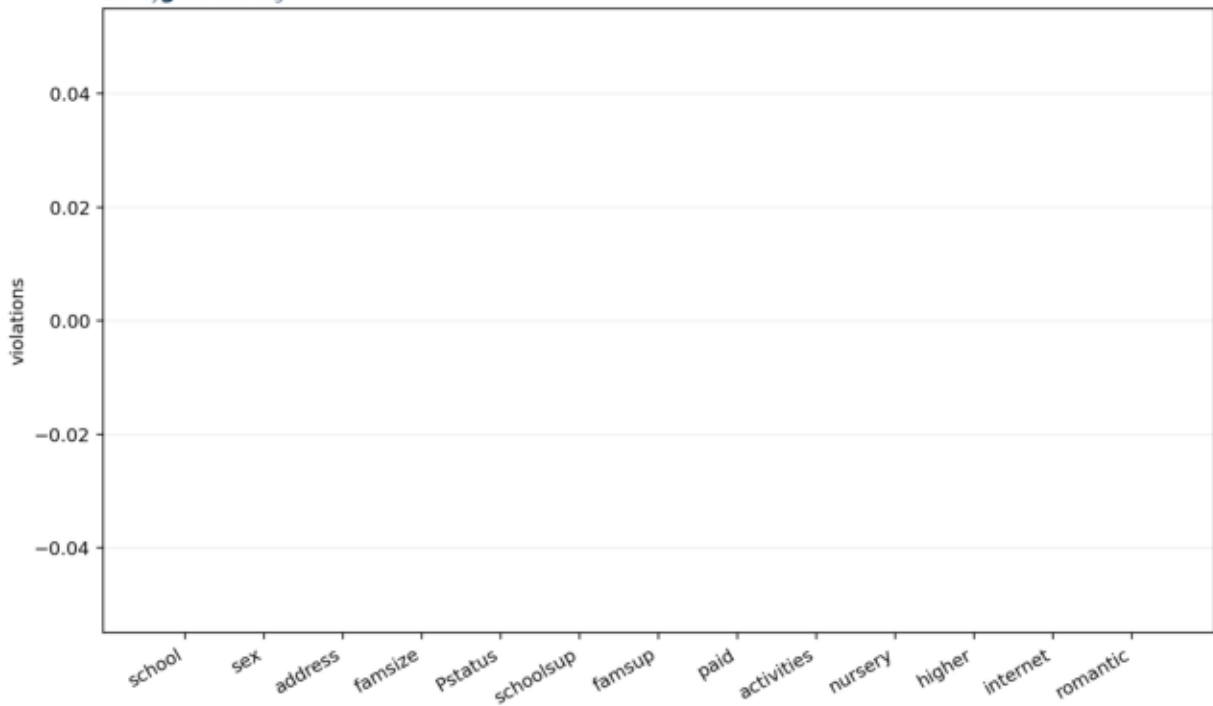
Exact duplicates, missing cells, categorical domains, and numeric ranges are audited separately.

Input rows=649; retained clean rows=649.

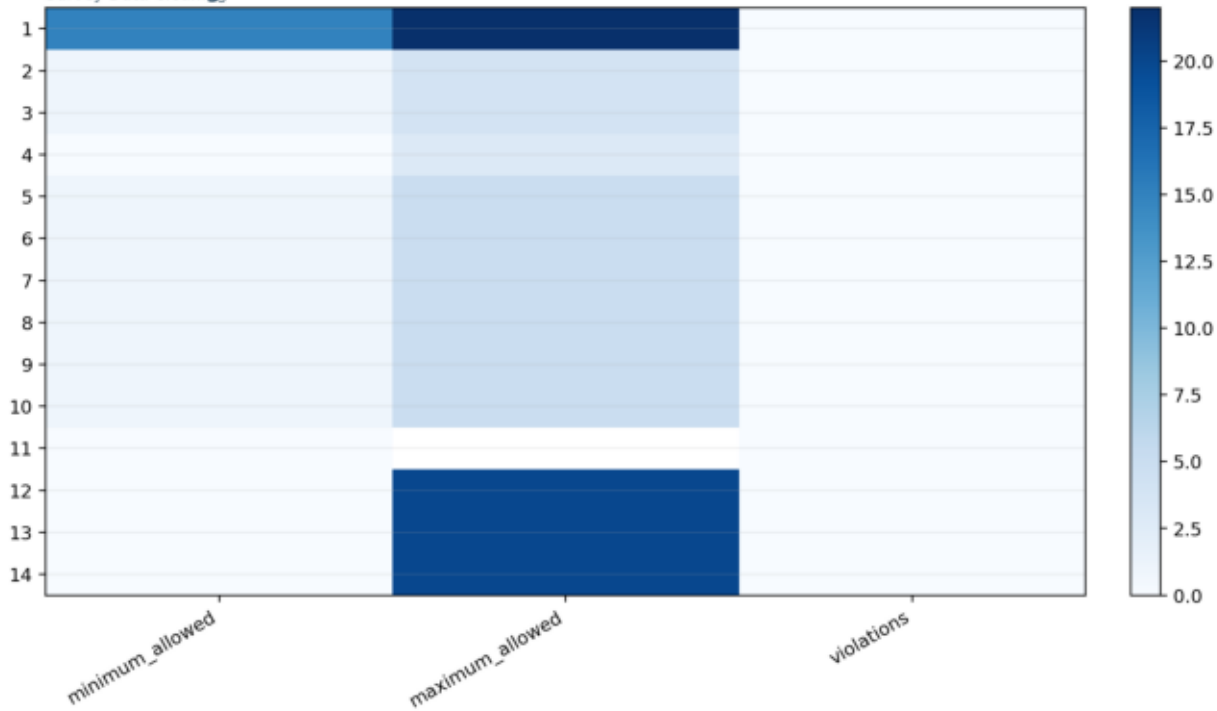
Primary Metrics



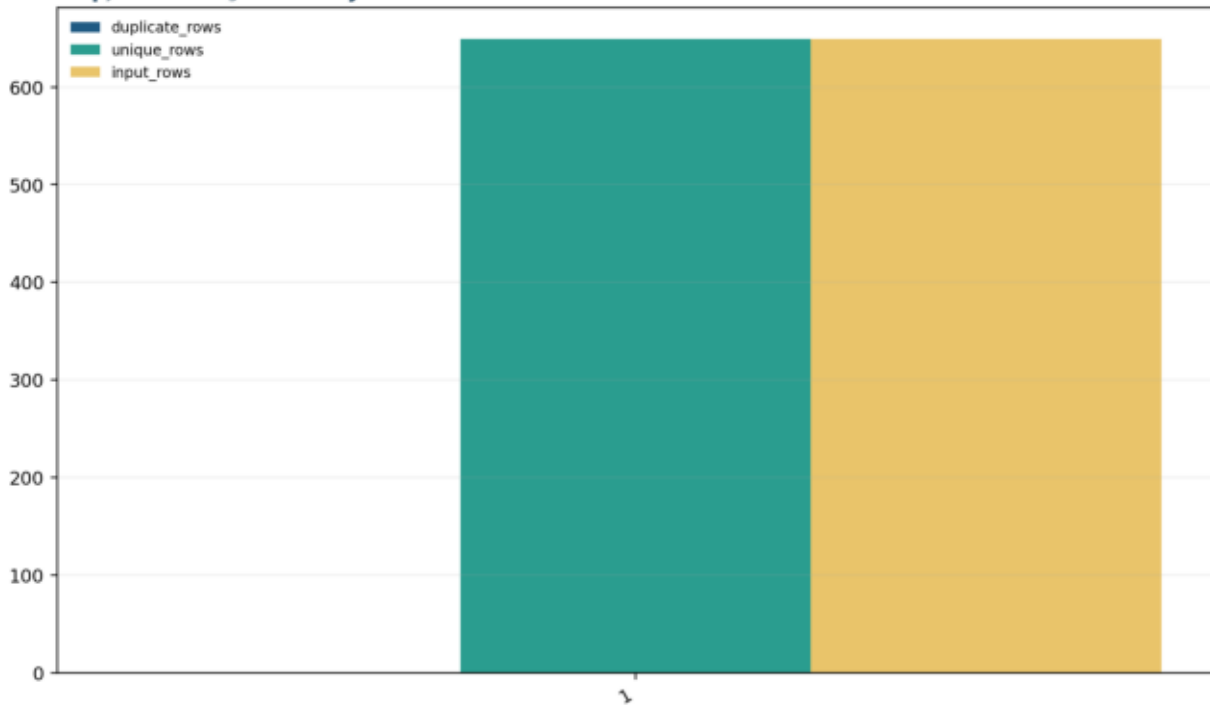
**Categorical Domain Checks**



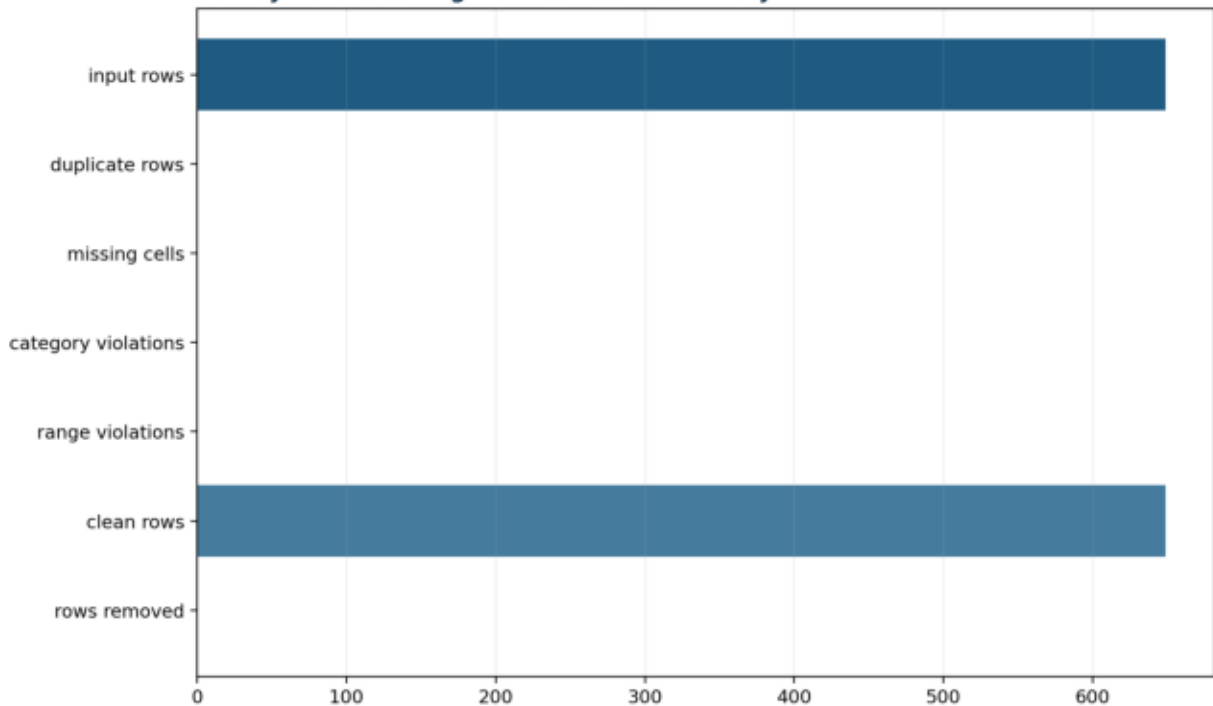
## Numeric Range Checks



## Duplicate Row Summary



## Survey Data Cleaning verified result summary



# Primary Metrics

| metric              | value |
|---------------------|-------|
| input_rows          | 649   |
| duplicate_rows      | 0     |
| missing_cells       | 0     |
| category_violations | 0     |
| range_violations    | 0     |
| clean_rows          | 649   |
| rows_removed        | 0     |

# Categorical Domain Checks

| variable   | violations | permitted_values |
|------------|------------|------------------|
| school     | 0          | GP MS            |
| sex        | 0          | F M              |
| address    | 0          | R U              |
| famsize    | 0          | GT3 LE3          |
| Pstatus    | 0          | A T              |
| schoolsup  | 0          | no yes           |
| famsup     | 0          | no yes           |
| paid       | 0          | no yes           |
| activities | 0          | no yes           |
| nursery    | 0          | no yes           |
| higher     | 0          | no yes           |
| internet   | 0          | no yes           |
| romantic   | 0          | no yes           |

# Numeric Range Checks

| variable   | minimum_allowed | maximum_allowed | violations |
|------------|-----------------|-----------------|------------|
| age        | 15              | 22              | 0          |
| traveltime | 1               | 4               | 0          |
| studytime  | 1               | 4               | 0          |
| failures   | 0               | 3               | 0          |
| famrel     | 1               | 5               | 0          |
| freetime   | 1               | 5               | 0          |
| goout      | 1               | 5               | 0          |
| Dalc       | 1               | 5               | 0          |
| Walc       | 1               | 5               | 0          |
| health     | 1               | 5               | 0          |
| absences   | 0               | nan             | 0          |
| G1         | 0               | 20              | 0          |
| G2         | 0               | 20              | 0          |
| G3         | 0               | 20              | 0          |

# Duplicate Row Summary

| duplicate_rows | unique_rows | input_rows |
|----------------|-------------|------------|
| 0              | 649         | 649        |