AutoClean Report

Name of dataset: Test Data (made up WHASH dataset)
Filepath of messy dataset: Data/Test/Test.csv
Filepath of cleaned dataset: Data/Test/Test_Cleaned.csv
Generated: 25.01.2026, 18:31:49

Summary

- **Original shape:** 52 rows × 18 columns
- **Final shape:** 50 rows × 16 columns
- **Total rows deleted:** 2
- **Total columns deleted:** 2
- **Total values imputed:** 13
- **Total outliers handled:** 3
- **Total semantic outliers detected:** 8
- **Total structural errors fixed:** 352

Preprocessing

- **Completely empty rows removed:** 1
- **Completely empty columns removed:** 1

Duplicates

- **Duplicate rows removed:** 1
- **Duplicate columns removed:** 1

Semantic Outliers

Overview

- **Columns processed:** 2
- **Total outliers detected:** 8
- **Total number of affected rows:** 8

Column: Village

- **Given context:** Location names in Africa
- **Threshold:** 0.5
- **Action:** nan
- **Unique values checked:** 14
- **Outliers detected:** 4

Detected Outliers

Value	Confidence	Number of affected rows
sdflkajsdf	0.0	1
I love studying at ETH!	0.0	1
1.8	0.0	1
unknown	0.0	1

Column: Population served

- **Given context:** Number of people
- **Threshold:** 0.5
- **Action:** nan
- **Unique values checked:** 50
- **Outliers detected:** 4

Detected Outliers

Value	Confidence	Number of affected rows
-250	0.0	1
1234.89	0.0	1
999.99	0.0	1
not so many	0.0	1

Outliers

Lower & Upper Bounds

Column	Lower Bound	Upper Bound
Flow Rate lps	-1.6987	6.7512
well_depth_m	-33.875	167.125
pump_age_years	-6.375	24.625
Water quality score	11.745	98.025
Annual maintenance cost	125.5	633.5

Overview

- **Multiplier:** 1.5
- **Total outliers:** 3
- **Method:** winsorize

Outliers Handled

Column	Original	New Value	Bound
Flow Rate lps	48.7	6.75125	upper
Flow Rate lps	9.2	6.75125	upper
Flow Rate lps	12.8	6.75125	upper

Note: New values shown above are pre-rounding. Final values may be rounded in post-processing to match original column precision.

DateTime Standardization

- **Column:** install_date
- **Format:** European (DD/MM)
- **Invalid handling:** nat
- **Total values:** 50
- **Successfully converted / standardized:** 40
- **Invalid values:** 10

Invalid values handled

Original	Action
15/2020	set to NaT
January 2020	set to NaT
15/25/2020	set to NaT
31/04/2020	set to NaT
29/02/2023	set to NaT
15/05/2200	set to NaT
15/2020/05	set to NaT
01/25/2024	set to NaT
2024/25/01	set to NaT
unknown	set to NaT

Structural Errors

Overview

- **Columns processed:** 12
- **Total values changed:** 352
- **Total unique values before:** 210

- **Total unique values after:** 79

Column: funding organization

- **Similarity method:** embeddings
- **Embedding model:** text-embedding-3-large
- **Clustering method:** connected_components
- **Threshold (connected components):** 0.6
- **Canonical selection:** llm
- **Values changed:** 40
- **Unique values before:** 24
- **Unique values after:** 5

Clustering Results

Original Values	Clustered to Canonical
Red Crss; RED CROSS; ICRC; International Red Cross; red cross; RedCross	International Red Cross
world health org; WHO; World Health Org; WH0; World Health Organization; W.H.O.; W.H.O; WHO.	World Health Organization
WorldBank; WORLD BANK; World Bank; Wrold Bank; world bank	World Bank
UNICEF; unicef; U.N.I.C.E.F.; United Nations Children's Fund	United Nations Children's Fund
WB	WB

Column: funding organization

- **Similarity method:** llm
- **LLM mode:** fast
- **LLM context provided:** Funding organizations
- **Clustering method:** hierarchical
- **Threshold (hierarchical):** 0.9
- **Canonical selection:** llm
- **Values changed:** 2
- **Unique values before:** 5
- **Unique values after:** 4

Clustering Results

Original Values	Clustered to Canonical
International Red Cross	International Red Cross
World Health Organization	World Health Organization
World Bank; WB	World Bank
United Nations Children's Fund	United Nations Children's Fund

Column: water_source

- **Similarity method:** rapidfuzz
- **Clustering method:** hierarchical
- **Threshold (hierarchical):** 0.85
- **Canonical selection:** llm
- **Values changed:** 45
- **Unique values before:** 18
- **Unique values after:** 3

Clustering Results

Original Values	Clustered to Canonical
BOREHOLE; Bore hole; borehole; Borehole; Borehle; Bore Hole; Borhole	Borehole
HAND PUMP; handpump; Hand pump; Hand punp; hand pump	Hand pump
PIPED WATER; piped water; Pipedwater; Piped Water; piped water; Piped water	Piped Water

Column: is_functional

- **Similarity method:** rapidfuzz
- **Clustering method:** hierarchical
- **Threshold (hierarchical):** 0.85
- **Canonical selection:** llm
- **Values changed:** 17
- **Unique values before:** 22
- **Unique values after:** 13

Clustering Results

Original Values	Clustered to Canonical
No; NO; no	No
TRUE; True; true	True
FALSE; false; False	False
broken	broken
0	0
not functional	not functional
working	working
yes; YES	yes
N; n	N

Original Values	Clustered to Canonical
y; Y	y
not working	not working
operational	operational
1	1

Column: is_functional

- **Similarity method:** llm
- **LLM mode:** reliable
- **LLM context provided:** Wether water point is working or not
- **Clustering method:** hierarchical
- **Threshold (hierarchical):** 0.85
- **Canonical selection:** llm
- **Values changed:** 38
- **Unique values before:** 13
- **Unique values after:** 2

Clustering Results

Original Values	Clustered to Canonical
No; False; broken; 0; not functional; N; not working	not functional
True; working; yes; y; operational; 1	True

Column: tank material

- **Similarity method:** rapidfuzz
- **Clustering method:** hierarchical
- **Threshold (hierarchical):** 0.7
- **Canonical selection:** llm
- **Values changed:** 21
- **Unique values before:** 23
- **Unique values after:** 13

Clustering Results

Original Values	Clustered to Canonical
Concrete; concrete; CONCRETE	Concrete
PVC; pvc	PVC
Plastic; Plastic tank; PLASTIC; plastic	Plastic tank
Cement; Cement tank; cement	Cement tank

Original Values	Clustered to Canonical
Poly	Poly
Iron	Iron
Metallic; metal	Metallic
Stainless steel; Stainless	Stainless steel
PE	PE
SS	SS
Reinforced concrete	Reinforced concrete
Polyethylene	Polyethylene
Steel	Steel

Column: tank material

- **Similarity method:** llm
- **LLM mode:** fast
- **LLM context provided:** Material of tank
- **Clustering method:** hierarchical
- **Threshold (hierarchical):** 0.5
- **Canonical selection:** llm
- **Values changed:** 39
- **Unique values before:** 13
- **Unique values after:** 3

Clustering Results

Original Values	Clustered to Canonical
Concrete; Cement tank; Reinforced concrete	Concrete
PVC; Plastic tank; Poly; PE; Polyethylene	Polyethylene
Iron; Metallic; Stainless steel; SS; Steel	Stainless steel

Column: sample Volume

- **Similarity method:** rapidfuzz
- **Clustering method:** hierarchical
- **Threshold (hierarchical):** 0.9
- **Canonical selection:** llm
- **Values changed:** 19
- **Unique values before:** 26
- **Unique values after:** 16

Clustering Results

Original Values	Clustered to Canonical
500l; 500L	500L
1000 l; 1000L; 1000 L	1000 L
250l; 250L	250L
1 kL	1 kL
five hundred liters	five hundred liters
1m³	1m³
250 litres; 250 Liters; 250 liters	250 liters
500 litres; 500 Liters; 500 liters	500 liters
1000000 ml	1000000 ml
1000 Liters; 1000 liters	1000 liters
250 L	250 L
0.25 m³	0.25 m³
1 cubic meter	1 cubic meter
250000ml; 250000 ml	250000 ml
0.5 kL	0.5 kL
500 L	500 L

Column: sample Volume

- **Similarity method:** llm
- **LLM mode:** strict
- **LLM context provided:** Volume measurements
- **Clustering method:** connected_components
- **Threshold (connected components):** 1.0
- **Canonical selection:** llm
- **Values changed:** 42
- **Unique values before:** 16
- **Unique values after:** 3

Clustering Results

Original Values	Clustered to Canonical
500L; five hundred liters; 500 liters; 0.5 kL; 500 L	500 L
1000 L; 1 kL; 1m³; 1000000 ml; 1000 liters; 1 cubic meter	1000 liters
250L; 250 liters; 250 L; 0.25 m³; 250000 ml	250 L

Column: country

- **Similarity method:** embeddings
- **Embedding model:** text-embedding-3-large
- **Clustering method:** hierarchical
- **Threshold (hierarchical):** 0.6
- **Canonical selection:** llm
- **Values changed:** 35
- **Unique values before:** 20
- **Unique values after:** 8

Clustering Results

Original Values	Clustered to Canonical
uganda; Ugandaa; Ugand; Republic of Uganda	Republic of Uganda
United Republic of Tanzania; TANZANIA; tanzania; Tanznia; Tanzania	Tanzania
UG	UG
Kenia; Keyna; KENYA; Republic of Kenya; Kenya	Kenya
Tanz.	Tanz.
TZA; TZ	TZ
KEN	KEN
KE	KE

Column: country

- **Similarity method:** llm
- **LLM mode:** fast
- **LLM context provided:** African countries
- **Clustering method:** hierarchical
- **Threshold (hierarchical):** 0.8
- **Canonical selection:** llm
- **Values changed:** 15
- **Unique values before:** 8
- **Unique values after:** 3

Clustering Results

Original Values	Clustered to Canonical
Republic of Uganda; UG	Republic of Uganda
Tanzania; Tanz.; TZ	Tanzania
Kenya; KEN; KE	Kenya

Column: staff_count

- **Similarity method:** llm
- **LLM mode:** fast
- **LLM context provided:** Number of staff
- **Clustering method:** hierarchical
- **Threshold (hierarchical):** 0.8
- **Canonical selection:** llm
- **Values changed:** 39
- **Unique values before:** 22
- **Unique values after:** 6

Clustering Results

Original Values	Clustered to Canonical
twenty-five; Twenty-Five; 25; twenty five; Twenty Five	25
Fifteen; fifteen; 15; FIFTEEN	15
ten; 10; Ten; TEN	10
20; twenty; Twenty	20
eight; Eight; 8	8
12; twelve; Twelve	12

Missing Values

Overview

- **Columns processed:** 3
- **Total values imputed:** 13
- **Total rows deleted:** 0

Column: Water quality score

- **Method:** missforest
- **Features used:** well_depth_m; pump_age_years
- **n_estimators:** 10
- **max_iter:** 5
- **max_depth:** 3
- **min_samples_leaf:** 3
- **Missing values before imputation:** 5
- **Values imputed:** 5

Imputations

Row	New imputed Value

Row	New imputed Value
3	41.95842004662004
8	60.66916903958615
19	42.199022222222226
31	66.97927571208032
45	51.935611732711735

Note: Imputed values shown above are pre-rounding. Final values may be rounded in post-processing.

Column: Annual maintenance cost

- **Method:** missforest
- **Features used:** well_depth_m; pump_age_years
- **n_estimators:** 10
- **max_iter:** 1
- **max_depth:** 3
- **min_samples_leaf:** 3
- **Missing values before imputation:** 5
- **Values imputed:** 5

Imputations

Row	New imputed Value
5	294.60238095238094
16	421.5731746031746
27	353.9897113997114
36	456.8509523809524
48	434.30095238095237

Note: Imputed values shown above are pre-rounding. Final values may be rounded in post-processing.

Column: System condition

- **Method:** knn
- **Features used:** well_depth_m; pump_age_years; Water quality score
- **n_neighbors:** 3
- **Missing values before imputation:** 3
- **Values imputed:** 3

Imputations

Row	New imputed Value

Row	New imputed Value
12	Fair
25	Fair
42	Fair

Note: Imputed values shown above are pre-rounding. Final values may be rounded in post-processing.

Postprocessing

Precision Restoration (rounding)

Column	Action
Flow Rate lps	Rounded to 2 decimals
well_depth_m	Restored to integer
pump_age_years	Restored to integer
Water quality score	Rounded to 2 decimals
Annual maintenance cost	Restored to integer

Renamed Columns

Original Column Name	New Column Name
Village	village
Population served	population_served
Flow Rate lps	flow_rate_lps
funding organization	funding_organization
sample Volume	sample_volume
tank material	tank_material
Water quality score	water_quality_score
Annual maintenance cost	annual_maintenance_cost
System condition	system_condition