AutoClean Report

Name of dataset: Malawi borehole drilling and construction data
Filepath of messy dataset: Data/Drilling/Drilling.csv
Filepath of cleaned dataset: Data/Drilling/Drilling_Cleaned.csv
Generated: 26.01.2026, 19:03:24

Summary

- **Original shape:** 157 rows × 136 columns
- **Final shape:** 157 rows × 4 columns
- **Total rows deleted:** 0
- **Total columns deleted:** 66
- **Total values imputed:** 0
- **Total outliers handled:** 0
- **Total semantic outliers detected:** 0
- **Total structural errors fixed:** 139

Preprocessing

- **Completely empty columns removed:** 66

Structural Errors

Overview

- **Columns processed:** 5
- **Total values changed:** 139
- **Total unique values before:** 152
- **Total unique values after:** 86

Column: funding_source

- **Similarity method:** rapidfuzz
- **Clustering method:** connected_components
- **Threshold (connected components):** 0.85
- **Canonical selection:** llm
- **Values changed:** 57
- **Unique values before:** 35
- **Unique values after:** 15

Clustering Results

Original Values	**Clustered to Canonical**
The Scottish Government; Scottish Government	Scottish Government
The Scottish Government through the Climate Justice Fund: Water futures Programme	The Scottish Government through the Climate Justice Fund: Water futures Programme
The Scottish Government through the CJf: CJF	The Scottish Government through the CJf: CJF
One foundation; One Foundation; one foundaton; one foundation; One doundation	One Foundation
One Alliance	One Alliance
Onefoundation; OneFoundation	OneFoundation
kindy projects; Kindy projects; Kindy project; Kindy Project	Kindy Project
Habitat for Humanity	Habitat for Humanity
Habitat for humanity,and German government; Habitat for Humanity and Germany government; German government and Habitat for Humanity; Habitat for Humanity and German Government; German Government and Habitat for Humanity; Habitat For Humanity and German Government; Habitat for Humanity and Germany Government; Habitat For humanity and Germany Government; German Government,and Habitat for Humanity; Habitat For Humanity and Germany Government	Habitat for Humanity and German Government
Post Code; Post code	Post code
GALE family	GALE family
Dr Rochelle Holm	Dr Rochelle Holm
SADC-GMI	SADC-GMI
SADC - GMI; SADC -GMI	SADC - GMI
Hilton Foundation	Hilton Foundation

Column: funding_source

- **Similarity method:** llm
- **LLM mode:** fast
- **LLM context provided:** Funding organizations and government bodies for water projects
- **Clustering method:** hierarchical
- **Threshold (hierarchical):** 0.75
- **Canonical selection:** most_frequent
- **Values changed:** 7
- **Unique values before:** 15
- **Unique values after:** 11

Clustering Results

Original Values	Clustered to Canonical
Scottish Government; The Scottish Government through the Climate Justice Fund: Water futures Programme; The Scottish Government through the CJf: CJF	Scottish Government
One Foundation; OneFoundation	One Foundation
One Alliance	One Alliance
Kindy Project	Kindy Project
Habitat for Humanity	Habitat for Humanity
Habitat for Humanity and German Government	Habitat for Humanity and German Government
Post code	Post code
GALE family	GALE family
Dr Rochelle Holm	Dr Rochelle Holm
SADC-GMI; SADC - GMI	SADC - GMI
Hilton Foundation	Hilton Foundation

Column: drilling_contractor

- **Similarity method:** rapidfuzz
- **Clustering method:** connected_components
- **Threshold (connected components):** 0.85
- **Canonical selection:** most_frequent
- **Values changed:** 32
- **Unique values before:** 55
- **Unique values after:** 31

Clustering Results

Original Values	Clustered to Canonical
OG Madzi; OG MADZI	OG Madzi
OG Madzi Drilling Company; OG MADZI drilling comopany; OG MADZI Drilling Company	OG Madzi Drilling Company
Mushtaq Of OG Madzi Drillinv	Mushtaq Of OG Madzi Drillinv
OG Madzi Drilling	OG Madzi Drilling
OG Madzi Construction	OG Madzi Construction
OG Madzi Drillers	OG Madzi Drillers
EAZY Drilling Copmany	EAZY Drilling Copmany
Eazy borehole drillers	Eazy borehole drillers
Eazy borehole drilling Company; EASY BOREHOLE DRILLING COMPANY; Eazy borehole Drilling company	Eazy borehole drilling Company
Saifro Ltd	Saifro Ltd
Saifro Limited	Saifro Limited
Saifro Limited Company	Saifro Limited Company
Saifro Malawi	Saifro Malawi
Saifro Limited Malawi; Saifro Malawi Limited	Saifro Limited Malawi
Saifro	Saifro

Original Values	Clustered to Canonical
Saifro Drilling Company	Saifro Drilling Company
Water Way Malawi; Water Way Malawi; Way Water Malawi	Water Way Malawi
Nditha Drilling and Civil Contractor; Nditha Drilling and Civil Contractors; Nditha Drilling and Civil contractors; Nditha Drilling and Civil Contractors; Nditha drilling and civil contractors; Nditha Civil and Drilling Contractors; Nditha Drilling and Civil contrators; Nditha Drilling and Civil Contractors	Nditha Drilling and Civil contractors
Blue Water Drilling Company; Blue water Drilling company; Blue water Drilling Company	Blue Water Drilling Company
Blue Water Drilling Ltd; Blue water drilling Ltd; Blue Water Drilling Ltd	Blue Water Drilling Ltd
GIMM water experts and drilling; GIMM Water experts and drilling; GIMM water experts and Drilling	GIMM water experts and drilling
GIMM; GIMME	GIMM
China Gansu	China Gansu
China Gansu Engineering Co.	China Gansu Engineering Co.
Nyaungano Drilling company	Nyaungano Drilling company
Blue water	Blue water
Mthunzi Wa Kachere	Mthunzi Wa Kachere
Dec Construction; Dec construction	Dec Construction
Dec construction Limited	Dec construction Limited
Patel and Ghodaniya; Patel and Godhaniya	Patel and Ghodaniya

Original Values	Clustered to Canonical
Rymech	Rymech

Column: drilling_contractor

- **Similarity method:** embeddings
- **Embedding model:** text-embedding-3-large
- **Clustering method:** connected_components
- **Threshold (connected components):** 0.65
- **Canonical selection:** most_frequent
- **Values changed:** 37
- **Unique values before:** 31
- **Unique values after:** 16

Clustering Results

Original Values	Clustered to Canonical
OG Madzi; OG Madzi Drilling Company; OG Madzi Drilling; OG Madzi Construction; OG Madzi Drillers	OG Madzi
Mushtaq Of OG Madzi Drillinv	Mushtaq Of OG Madzi Drillinv
EAZY Drilling Copmany; Eazy borehole drillers; Eazy borehole drilling Company	Eazy borehole drilling Company
Saifro Ltd; Saifro Limited; Saifro Limited Company; Saifro Malawi; Saifro Limited Malawi; Saifro; Saifro Drilling Company	Saifro Limited
Water Way Malawi	Water Way Malawi
Nditha Drilling and Civil contractors	Nditha Drilling and Civil contractors
Blue Water Drilling Company; Blue Water Drilling Ltd	Blue Water Drilling Ltd
GIMM water experts and drilling	GIMM water experts and drilling
GIMM	GIMM
China Gansu; China Gansu Engineering Co.	China Gansu
Nyaungano Drilling company	Nyaungano Drilling company
Blue water	Blue water

Original Values	Clustered to Canonical
Mthunzi Wa Kachere	Mthunzi Wa Kachere
Dec Construction; Dec construction Limited	Dec Construction
Patel and Ghodaniya	Patel and Ghodaniya
Rymech	Rymech

Column: drilling_contractor

- **Similarity method:** llm
- **LLM mode:** fast
- **LLM context provided:** Drilling contractor companies in East Africa
- **Clustering method:** hierarchical
- **Threshold (hierarchical):** 0.7
- **Canonical selection:** most_frequent
- **Values changed:** 6
- **Unique values before:** 16
- **Unique values after:** 13

Clustering Results

Original Values	Clustered to Canonical
OG Madzi; Mushtaq Of OG Madzi Drillinv	OG Madzi
Eazy borehole drilling Company	Eazy borehole drilling Company
Saifro Limited	Saifro Limited
Water Way Malawi	Water Way Malawi
Nditha Drilling and Civil contractors	Nditha Drilling and Civil contractors
Blue Water Drilling Ltd; Blue water	Blue Water Drilling Ltd
GIMM water experts and drilling; GIMM	GIMM water experts and drilling
China Gansu	China Gansu
Nyaungano Drilling company	Nyaungano Drilling company
Mthunzi Wa Kachere	Mthunzi Wa Kachere
Dec Construction	Dec Construction
Patel and Ghodaniya	Patel and Ghodaniya
Rymech	Rymech

Postprocessing

Precision Restoration (rounding)

No precision restoration (rounding) was applied in post-processing.

Renamed Columns

Column renaming was not applied.