AutoClean Report

Name of dataset: Ask A Manager Salary Survey 2021 (Sample)
Filepath of messy dataset: Data/Salary/Salary.csv
Filepath of cleaned dataset: Data/Salary/Salary_Cleaned.csv
Generated: 27.01.2026, 01:04:00

Summary

- **Original shape:** 28187 rows × 18 columns
- **Final shape:** 750 rows × 2 columns
- **Total rows deleted:** 0
- **Total columns deleted:** 0
- **Total values imputed:** 0
- **Total outliers handled:** 0
- **Total semantic outliers detected:** 3
- **Total structural errors fixed:** 442

Preprocessing

No completely empty rows or columns found respectfully removed.

Semantic Outliers

Overview

- **Column processed:** What country do you work in?
- **Given context:** Country names
- **Threshold:** 0.5
- **Action:** nan
- **Unique values checked:** 51
- **Outliers detected:** 3

Detected Outliers

Value	Confidence	Number of affected rows
Y	0.0	1
I was brought in on this salary to help with the EHR and very quickly was promoted to current position but compensation was not altered.	0.0	1

Value	Confidence	Number of affected rows
United y	0.4	1

Structural Errors

Overview

- **Columns processed:** 3
- **Total values changed:** 442
- **Total unique values before:** 109
- **Total unique values after:** 84

Column: What country do you work in?

- **Similarity method:** rapidfuzz
- **Clustering method:** connected_components
- **Threshold (connected components):** 0.88
- **Canonical selection:** most_frequent
- **Values changed:** 52
- **Unique values before:** 48
- **Unique values after:** 35

Clustering Results

Original Values	Clustered to Canonical
United States; united states; United states; Unites States; United Sates; United Stated; Unites states	United States
England	England
US; Us; us	US
Canada	Canada
USA; Usa; usa	USA
Australia; australia	Australia
Spain	Spain
Scotland	Scotland
U.S.A.	U.S.A.
Africa	Africa
U.S.	U.S.

Original Values	Clustered to Canonical
United Kingdom; United kingdom	United Kingdom
United States of America	United States of America
Austria	Austria
Uk; UK	UK
Japan	Japan
Thailand	Thailand
France	France
Belgium	Belgium
New Zealand	New Zealand
South africa	South africa
Ireland	Ireland
croatia	croatia
Italy	Italy
Netherlands	Netherlands
Northern Ireland, United Kingdom	Northern Ireland, United Kingdom
The Netherlands	The Netherlands
U. S.	U. S.
U.S	U.S
Sweden	Sweden
Denmark	Denmark
Germany	Germany
Switzerland	Switzerland
Congo	Congo
Kenya	Kenya

Column: What country do you work in?

- **Similarity method:** embeddings
- **Embedding model:** text-embedding-3-large
- **Clustering method:** connected_components
- **Threshold (connected components):** 0.65
- **Canonical selection:** most_frequent

- **Values changed:** 378
- **Unique values before:** 35
- **Unique values after:** 26

Clustering Results

Original Values	Clustered to Canonical
United States; US; USA; U.S.A.; U.S.; United States of America; U. S.; U.S	United States
England	England
Canada	Canada
Australia	Australia
Spain	Spain
Scotland	Scotland
Africa	Africa
United Kingdom; UK	UK
Austria	Austria
Japan	Japan
Thailand	Thailand
France	France
Belgium	Belgium
New Zealand	New Zealand
South africa	South africa
Ireland	Ireland
croatia	croatia
Italy	Italy
Netherlands; The Netherlands	Netherlands
Northern Ireland, United Kingdom	Northern Ireland, United Kingdom
Sweden	Sweden
Denmark	Denmark
Germany	Germany
Switzerland	Switzerland
Congo	Congo

Original Values	Clustered to Canonical
Kenya	Kenya

Column: What country do you work in?

- **Similarity method:** llm
- **LLM mode:** fast
- **LLM context provided:** Answers to question: What country do you work in?
- **Clustering method:** connected_components
- **Threshold (connected components):** 0.7
- **Canonical selection:** most_frequent
- **Values changed:** 12
- **Unique values before:** 26
- **Unique values after:** 23

Clustering Results

Original Values	Clustered to Canonical
United States	United States
England; Scotland; UK; Northern Ireland, United Kingdom	UK
Canada	Canada
Australia	Australia
Spain	Spain
Africa	Africa
Austria	Austria
Japan	Japan
Thailand	Thailand
France	France
Belgium	Belgium
New Zealand	New Zealand
South africa	South africa
Ireland	Ireland
croatia	croatia
Italy	Italy
Netherlands	Netherlands
Sweden	Sweden

Original Values	Clustered to Canonical
Denmark	Denmark
Germany	Germany
Switzerland	Switzerland
Congo	Congo
Kenya	Kenya

Postprocessing

Precision Restoration (rounding)

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

Renamed Columns

Column renaming was not applied.