Dataproofer

Spell check for data

Jonathan Page

UHERO, University of Hawaii

Spell check identifies potential spelling errors.

Dataproofer identifies potential data errors.

Primary Use Case

Quick check of data quality
immediately following data collection

Strengths

- Fast, automated quality checks
- Cross-platform (Window, OS X, Linux)
- Supports XLSX, XLS, TSV, CSV, PSV, and Google Spreadsheets
- Customizable checks

Weaknesses

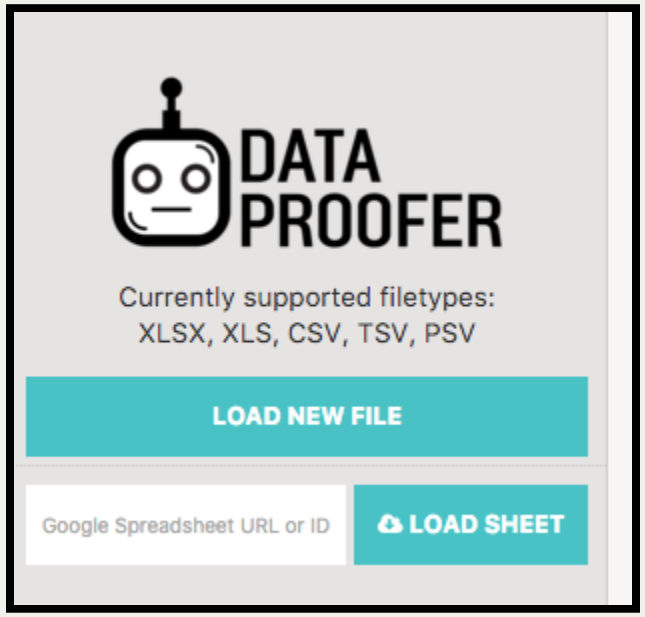
- Creating custom checks requires some programming experience
- Not for editing
- Does not like large data (> 500 KB)
- Does not support statistical software data formats

Using Dataproofer

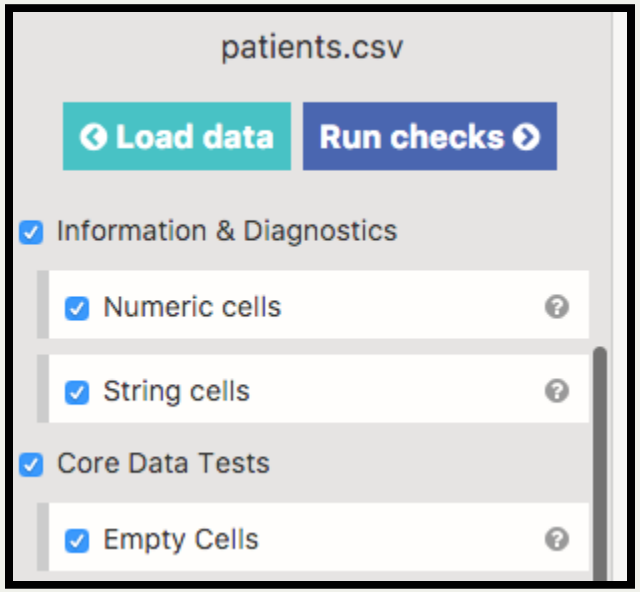
Install Dataproofer

dataproofer.org

Load Raw Data



Run Default Checks



Review Results

patients.csv

Select checks

9 passed out of 15 total

No missing or duplicate column headers

- Numeric cells
- String cells
- Empty Cells
- Duplicate Rows
- Potentially missing rows
- Words at their character limit
- Integer at its SQL upper limit
- Summed integer at its upper limit
- Small integer at its SQL upper limit
- Big integer at its SQL upper limit
- Outliers from the mean

	001	M	11/11/1998	88140	80	10
1	002	F	11/13/1998	84120	78	X0
2	003	X	10/21/1998	68190	100	31
3	004	F	01/01/1999	101200	120	5A
4	XX5	M	05/07/1998	68120	80	10
5	006		06/15/1999	72102	68	61
6	007	M	08/32/1998	88148	102	0
7	008	F	08/08/1998		210	70
8	009	M	09/25/1999	86240	180	41
9	010	F	10/19/1999	40120	10	
10	011	M	13/13/1998	68300	20	41
11	012	M	10/12/98	60122	74	0
12	013	2	08/23/1999	74108	64	1
13	014	M	02/02/1999	22130	90	1
14	002	F	11/13/1998	84120	78	X0
15	003	M	11/12/1999	58112	74	0
16	015	F		82148	88	31
17	017	F	04/05/1999	208	84	20
18	019	M	06/07/1999	58118	70	0
19	123	M	15/12/1999		60	10
20	321	F		900400	200	51
21	020	F	99/99/9999	10 20	8	0
22	022	M	10/10/1999	48114	82	21
23	023	F	12/31/1998	22 34	78	0
24	024	F	11/09/1998	76 120	80	10
25	025	M	01/01/1999	74102	68	51
26	027	F	NOTAVAIL N	166	106	70
27	028	F	03/28/1998	66150	90	30

OVERVIEW

Demo

patients.csv

Custom Checks

Motivation

Working with survey data, we want to quickly identify the following missing values:

- NA
- N/A
- -99
- -98

Before the Custom Check

field.csv

Anatomy of a Custom Check

```
customDataprooferTest.name("Missing (NA)")
    .description("If a cell contains an NA value.")
    .conclusion("Check for any patterns in missing values.")
    .methodology(cellMethod(isMissingCheck))
```

isMissingCheck Function

```
var isMissingCheck = function(cell) {
    if (cell === 'NA' || cell === 'N/A' ||
        cell  == -99  || cell  == -98) {
        return true;
    }
    return false;
}
```

After Custom Check

fieldv2.csv

Select checks

1 passed out of 2 total

No missing or duplicate column headers

Missing (NA)

		fo	name	contact	contact2	date	location	consent	OVERVIEW
1	3505	8.92540293		fe05932c-14	Mery Njoki M	Simon Maige	704113535	-99	
2	3505	8.92540293		f42d6a5f-46	Mery Njoki M	Muraya Kima	799999999	-99	
3	3505	8.92540293		b6613436-5	Mery Njoki M	Bernard mbu	707026796	-99	
4	3505	8.92540293		f6fbc6c3-e3	Mery Njoki M	Olive Mwihak	728738426	792365406	
5	3505	8.92540293		5e8d7ca6-9	Mery Njoki M	Njoroge ngar	799999999	-99	
6	056	8.92540210		7cdf2552-39	Maureen Nje	Evelyne Ndu	712665330	-99	
7	056	8.92540210		acf3caab-e8	Maureen Nje	Francis Kima	799999999	-99	
8	056	8.92540210		1931e5ac-3a	Maureen Nje	Henry Karanj	720093842	-99	
9	056	8.92540210		1bc50b60-f7	Maureen Nje	Moffat Ngati	708066916	-99	
10				6a19cf7b-67	Moffat Ndun	Lilly Waringa	723310963	799999999	
11	056	8.92540210		0fbd189e-c7	Maureen Nje	Ruth Wanjiru	718865716	-99	
12				5069e1e7-2	Moffat Ndun	Eliud Mwaura	711158253	725522213	
13				080d55c1-4	Moffat Ndun	Lucy Muthor	714170921	712740352	
14				703b2353-e	Moffat Ndun	James Waith	725993888	720475915	
15				ba26a975-d	Mercy Mumb	David ndichu	721987416	-99	
16				554f9b7c-6	Mercy Mumb	Agnes wanjir	714368895	-99	
17				7612ef4c-d4	Mercy Mumb	Samuel wain	999999999	-99	
18				51b282a2-4	Catherine Ml	Isabel mwiha	736598202	-99	
19				29cd7d67-1	Dickson Mak	Ann Wambui	710366151	-99	
20				ba1ac56a-1d	Dickson Mak	Eunice Ann N	726788961	787570073	
21				69d6f86e-8	Dickson Mak	Gladys Waru	727654170	N/A	
22				3c488e02-e	Jackline Kier	James mwau	708663928	-98	
23				3fdff8dc-ce	Jackline Kier	George gikor	721995377	731334166	
24				36701bd5-5	Dickson Mak	Peter Kamau	710652816	-99	
25				2129cd07-9	Jackline Kier	Stephen gita	729781849	-99	
26				acfeb4e0-61	Maureen Nje	Margaret Njo	723718133	7268182322	
27				f747e864-9	Githu Anne V	James kimar	701828752	720067721	

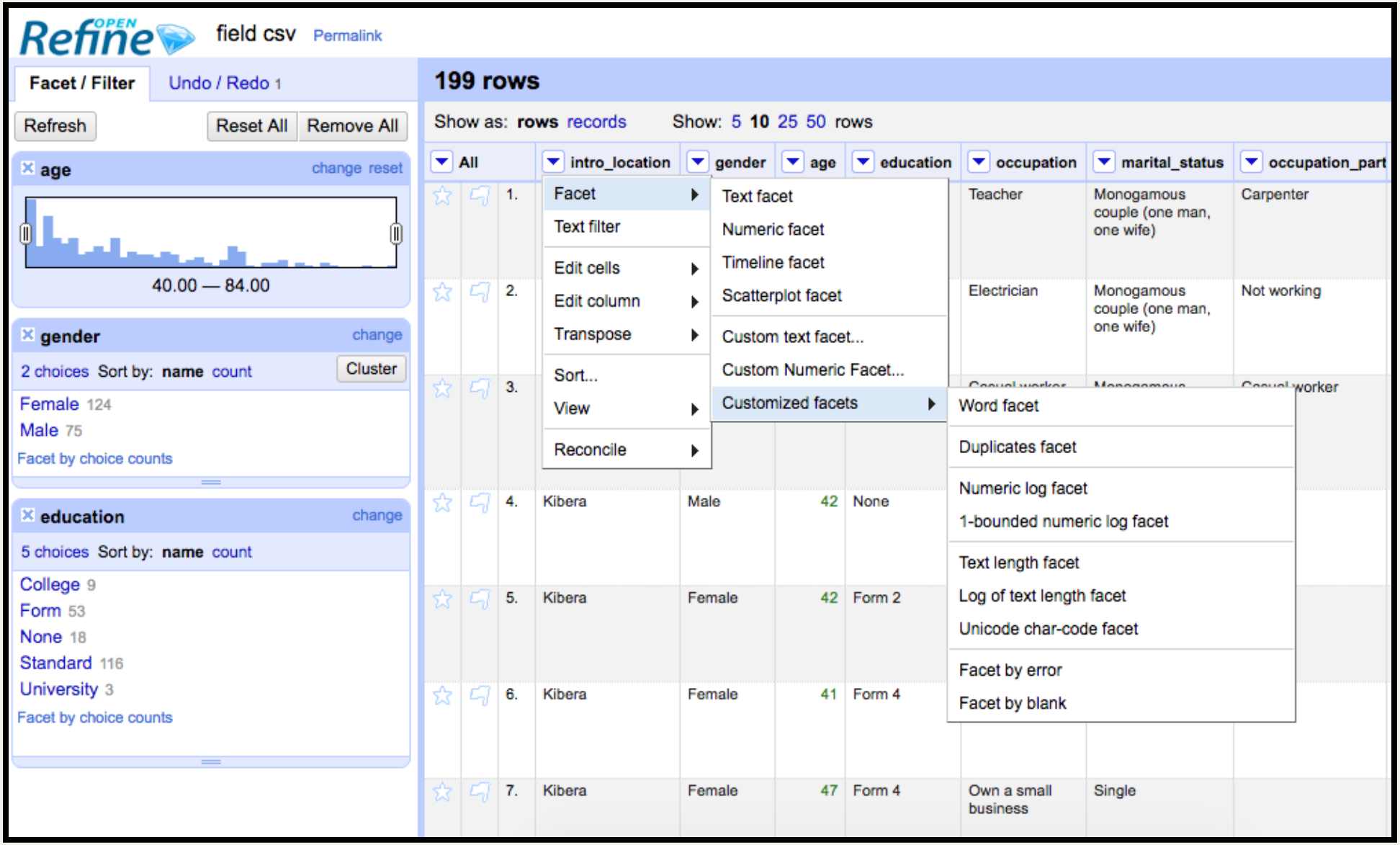
Alternative tools

OpenRefine

openrefine.org

Formerly Google Refine

OpenRefine Facets



OpenRefine Clusters

Cluster & Edit column "intro_location"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method nearest neighbor Distance Function levenshtein Radius 1.0 Block Chars 1 **2** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	74	• Kibera (71 rows) • kibera (2 rows) • Kbera (1 rows)	☐	Kibera
2	73	• Kibera (71 rows) • kibera (2 rows)	☐	Kibera

 # Choices in Cluster

2 — 3

 # Rows in Cluster

73 — 74

Average Length of Choices

5.666 — 6

OpenRefine: Strengths

- Open source software
- Extensible
  custom checks/filters can be written in Clojure, Python/Jython, or Refine's own expression language

OpenRefine: Weaknesses

- Can't handle large files (> 500 KB)
- Feels slow
- No easy way to track changes (not very reproducible)

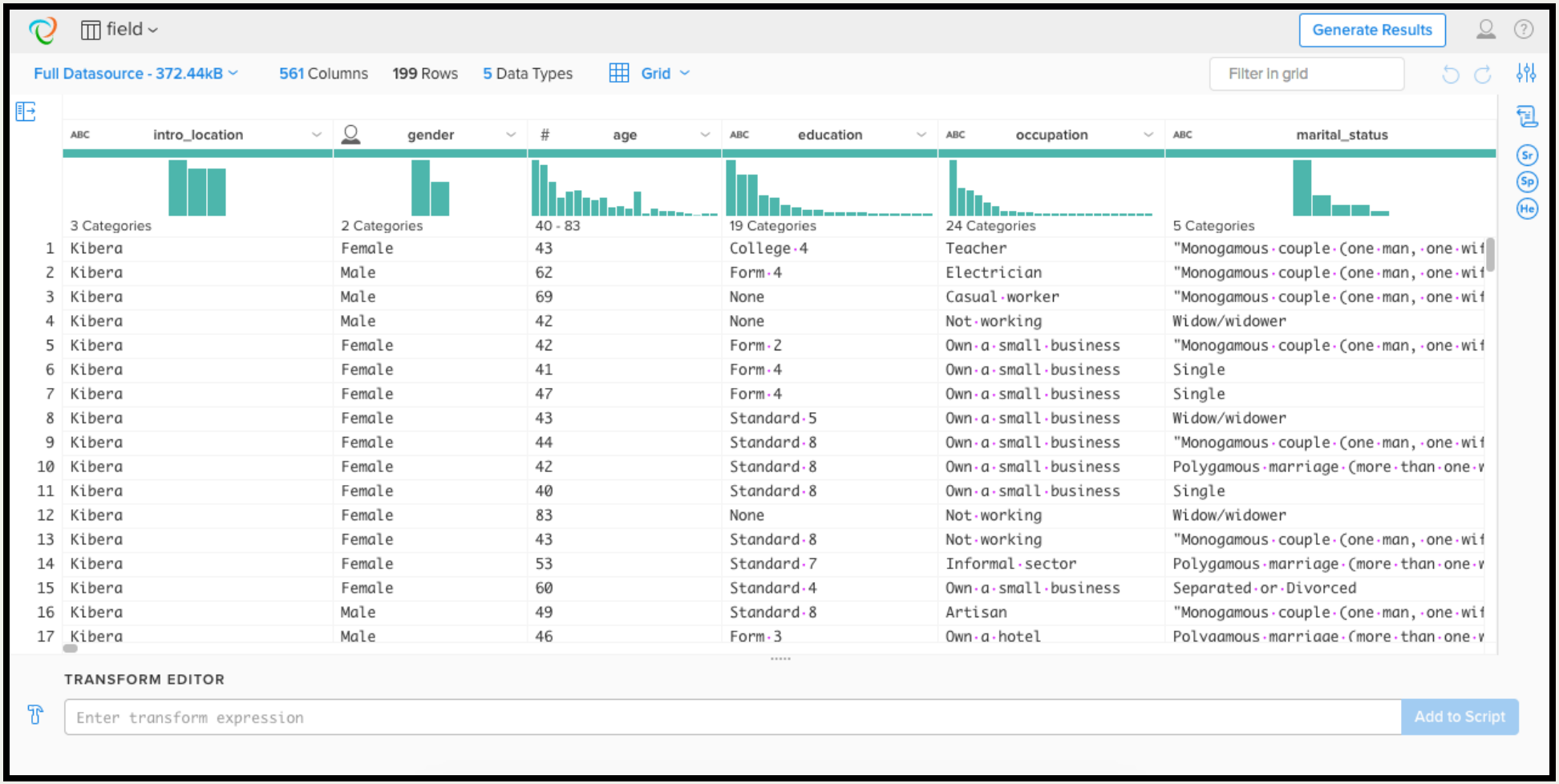
OpenRefine Demo

Trifacta Wrangler

trifacta.com

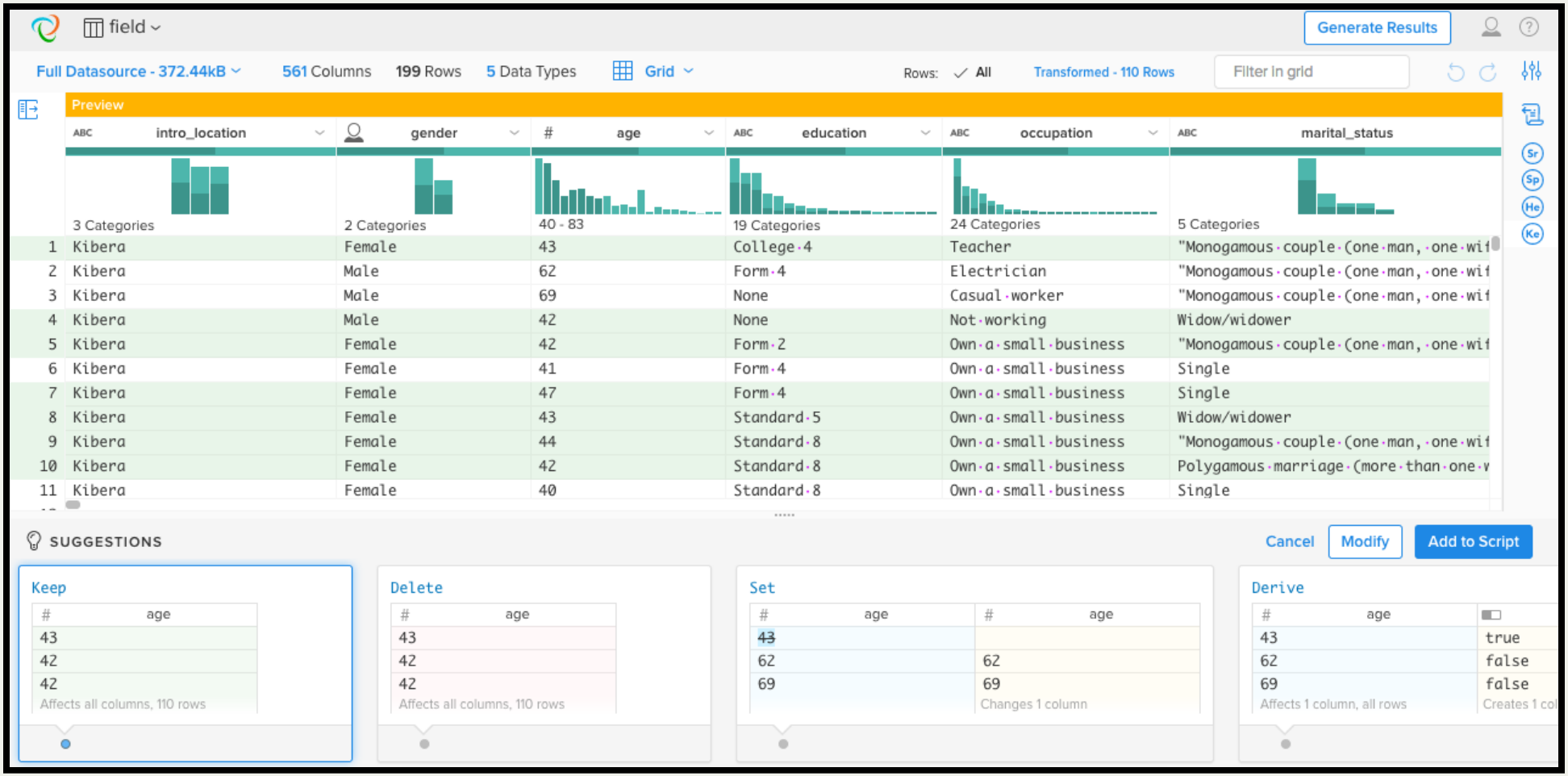
Formerly Stanford Data Wrangler

Trifacta: Overview

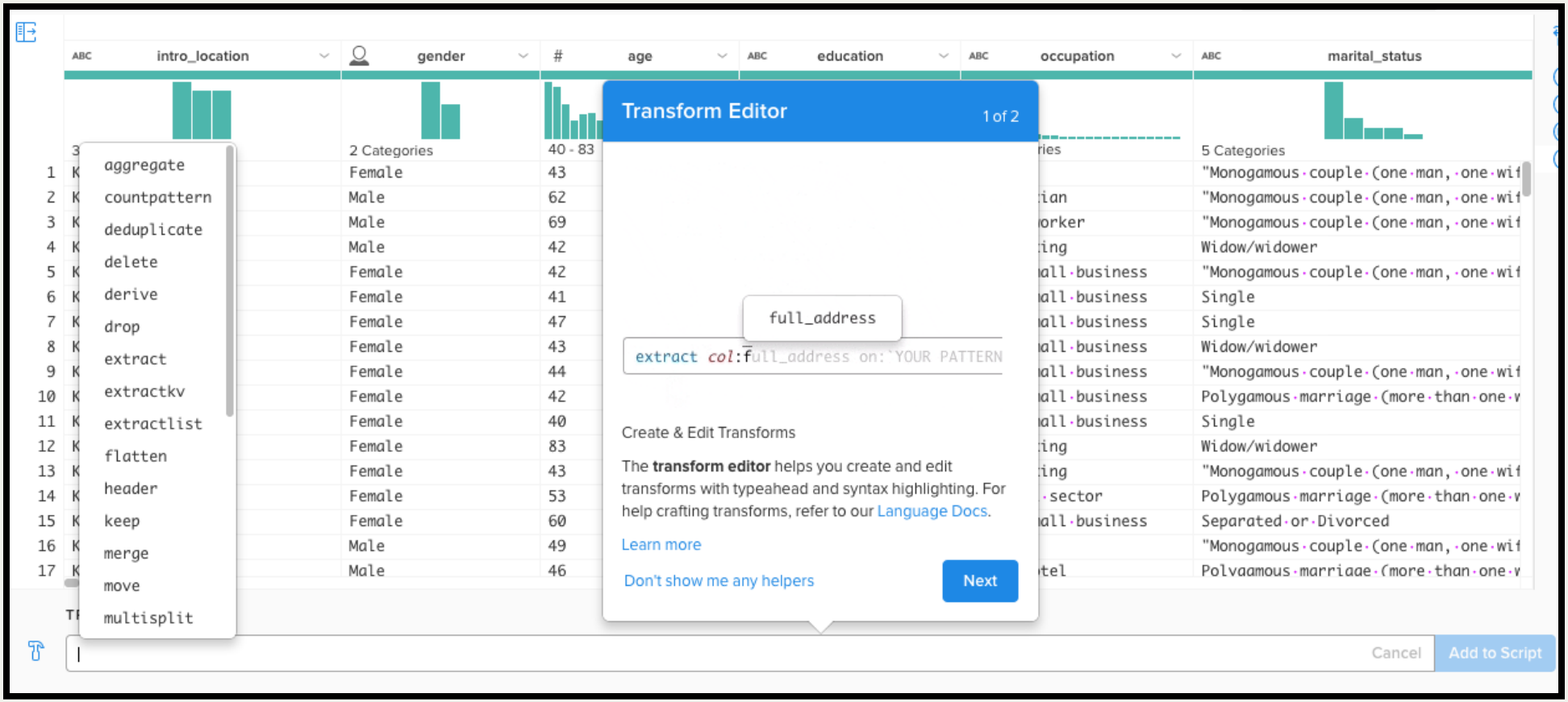


	intro_location	gender	age	education	occupation	marital_status
	3 Categories	2 Categories	40 - 83	19 Categories	24 Categories	5 Categories
1	Kibera	Female	43	College 4	Teacher	"Monogamous couple (one man, one wif
2	Kibera	Male	62	Form 4	Electrician	"Monogamous couple (one man, one wif
3	Kibera	Male	69	None	Casual worker	"Monogamous couple (one man, one wif
4	Kibera	Male	42	None	Not working	Widow/widower
5	Kibera	Female	42	Form 2	Own a small business	"Monogamous couple (one man, one wif
6	Kibera	Female	41	Form 4	Own a small business	Single
7	Kibera	Female	47	Form 4	Own a small business	Single
8	Kibera	Female	43	Standard 5	Own a small business	Widow/widower
9	Kibera	Female	44	Standard 8	Own a small business	"Monogamous couple (one man, one wif
10	Kibera	Female	42	Standard 8	Own a small business	Polygamous marriage (more than one w
11	Kibera	Female	40	Standard 8	Own a small business	Single
12	Kibera	Female	83	None	Not working	Widow/widower
13	Kibera	Female	43	Standard 8	Not working	"Monogamous couple (one man, one wif
14	Kibera	Female	53	Standard 7	Informal sector	Polygamous marriage (more than one w
15	Kibera	Female	60	Standard 4	Own a small business	Separated or Divorced
16	Kibera	Male	49	Standard 8	Artisan	"Monogamous couple (one man, one wif
17	Kibera	Male	46	Form 3	Own a hotel	Polygamous marriage (more than one w

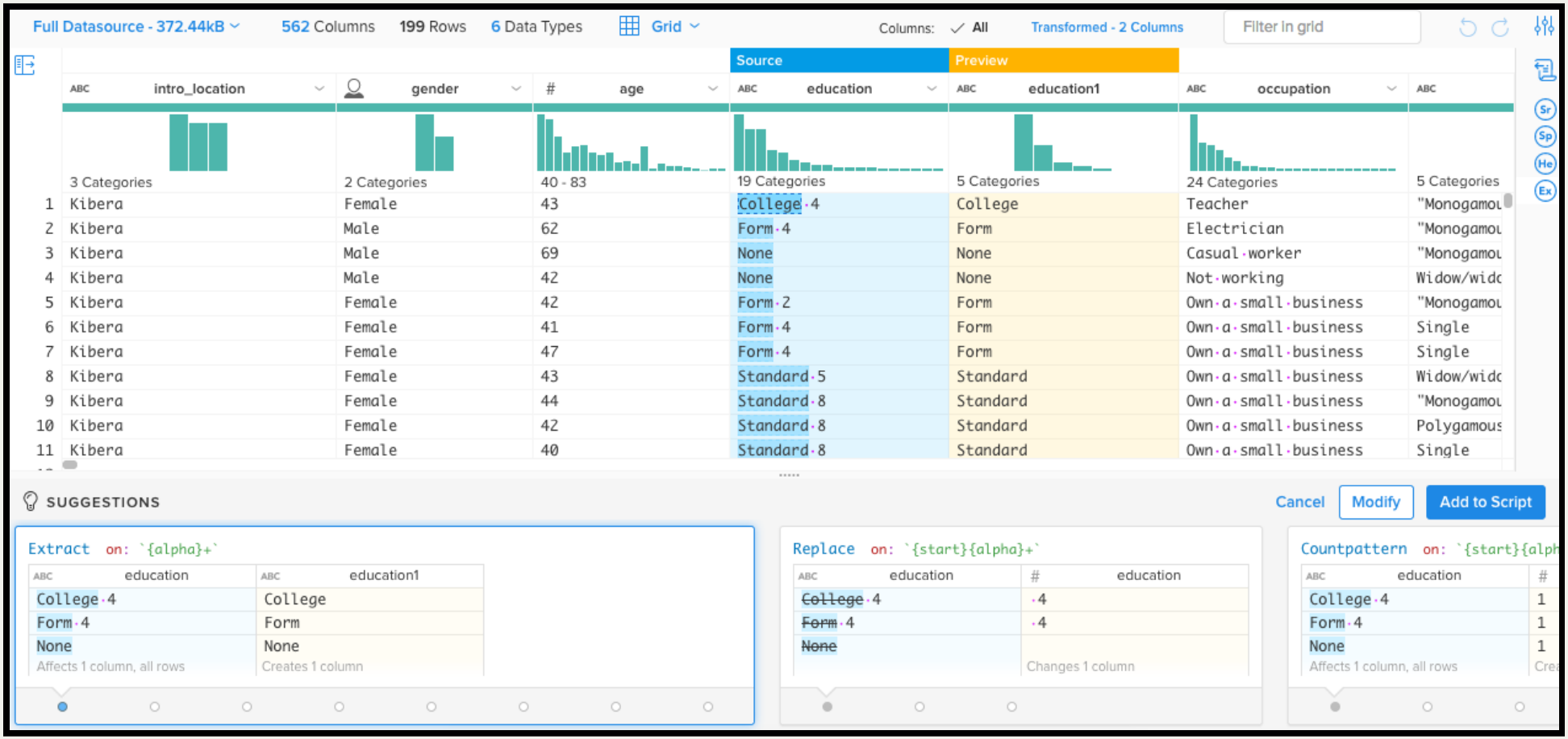
Trifacta: Filter



Trifacta: Scripts



Trifacta: Scripting Without Code



Trifacta: Strengths

- Interactive histograms
- Script builder
- Reproducible data cleaning
- Takes a 500 KB sample from large datasets

Trifacta: Weaknesses

- Requires an internet connection (may have to setup proxy settings)
- May not be appropriate for confidential data

Trifacta Demo

Conclusion

Use Dataproofer for

- Data quality checks for data entry
- Nice high-level overview of potential data errors
- First check after downloading or creating a new dataset

If you need more,
consider OpenRefine and Trifacta Wrangler.