

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



**DATA
PROOFER**

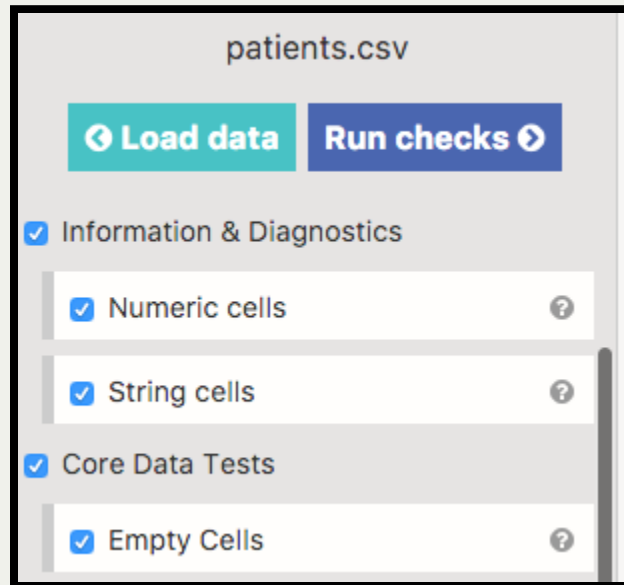
Currently supported filetypes:
XLSX, XLS, CSV, TSV, PSV

LOAD NEW FILE

Google Spreadsheet URL or ID

LOAD SHEET

Run Default Checks



patients.csv

☒ Information & Diagnostics

- ☒ Numeric cells ?
- ☒ String cells ?

☒ Core Data Tests

- ☒ Empty Cells ?

The image shows a web-based interface for running data checks on a file named 'patients.csv'. At the top, there are two buttons: 'Load data' (teal) and 'Run checks' (blue). Below these, there are two main sections of checks. The first section, 'Information & Diagnostics', is expanded and shows two sub-items: 'Numeric cells' and 'String cells', both of which are checked with blue checkmarks and have a question mark icon to their right. The second section, 'Core Data Tests', is also expanded and shows one sub-item: 'Empty Cells', which is also checked with a blue checkmark and has a question mark icon to its right. The interface is clean and modern, with a light gray background and a black border around the main content area.

Review Results

patients.csv		001	M	11/11/1998	88140	80	10	OVERVIEW			
<div>Select checks</div> <div>9 passed out of 15 total</div> <div>No missing or duplicate column headers</div> <div> <div>Numeric cells</div> <div>String cells</div> <div>Empty Cells</div> <div>Duplicate Rows</div> <div>Potentially missing rows</div> <div>Words at their character limit</div> <div>Integer at its SQL upper limit</div> <div>Summed integer at its upper limit</div> <div>Small integer at its SQL upper limit</div> <div>Big integer at its SQL upper limit</div> <div>Outliers from the mean</div> </div>		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			

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		fo	name	contact	contact2	date	location	consent	OVERVIEW
1	3509	8.92540293		fe05932c-14	Mery Njoki M	Simon Maige	704113535	-99	
2	3509	8.92540293		f42d6a5f-46	Mery Njoki M	Muraya Kima	799999999	-99	
3	3509	8.92540293		b6613436-5	Mery Njoki M	Bernard mbu	707026796	-99	
4	3509	8.92540293		f6fbc6c3-e3	Mery Njoki M	Olive Mwihai	728738426	792365406	
5	3509	8.92540293		5e8d7ca6-9	Mery Njoki M	Njoroge ngai	799999999	-99	
6	0569	8.925402100		7cdf2552-35	Maureen Nje	Evelyne Ndu	712665330	-99	
7	0569	8.925402100		acf3caab-e8	Maureen Nje	Francis Kima	799999999	-99	
8	0569	8.925402100		1931e5ac-3e	Maureen Nje	Henry Karanj	720093842	-99	
9	0569	8.925402100		1bc50b60-f7	Maureen Nje	Moffat Ngati	708066916	-99	
10				6a19cf7b-67	Moffat Ndun	Lilly Waringa	723310963	799999999	
11	0569	8.925402100		0fbd189e-c7	Maureen Nje	Ruth Wanjiru	718865716	-99	
12				5069e1e7-2	Moffat Ndun	Eliud Mwauru	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 M	Isabel mwiha	736598202	-99	
19				29cd7d67-1	Dickson Mak	Ann Wambui	710366151	-99	
20				ba1ac56a-1d	Dickson Mak	Eunice Ann M	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				acfeb4e0-6	Maureen Nje	Margaret Nje	723718133	7268182322	
27				f747e864-9	Githu Anne V	James kimar	701828752	720067721	

Alternative tools

OpenRefine

openrefine.org

Formerly Google Refine

OpenRefine Facets

Refine OPEN field csv Permalink

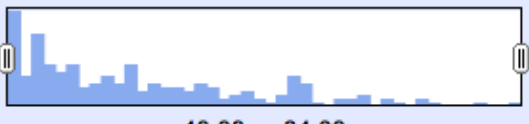
Facet / Filter Undo / Redo 1

Refresh Reset All Remove All

199 rows

Show as: rows records Show: 5 10 25 50 rows

age change reset



40.00 — 84.00

gender change

2 choices Sort by: name count Cluster

Female 124
Male 75
Facet by choice counts

education change

5 choices Sort by: name count

College 9
Form 53
None 18
Standard 116
University 3
Facet by choice counts

intro_location gender age education occupation marital_status occupation_part

1. Facet Text facet
Text filter Numeric facet
Edit cells Timeline facet
Edit column Scatterplot facet
Transpose Custom text facet...
Sort... Custom Numeric Facet...
View Customized facets
Reconcile

2. Teacher Monogamous couple (one man, one wife) Carpenter

3. Electrician Monogamous couple (one man, one wife) Not working

4. Kibera Male 42 None

5. Kibera Female 42 Form 2

6. Kibera Female 41 Form 4

7. Kibera Female 47 Form 4 Own a small business Single

Word facet
Duplicates facet
Numeric log facet
1-bounded numeric log facet
Text length facet
Log of text length facet
Unicode char-code facet
Facet by error
Facet by blank

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

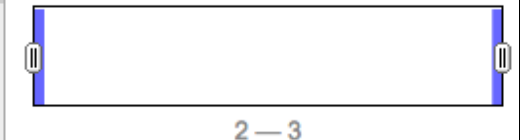
Radius 1.0

Block Chars 1

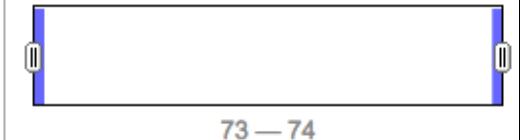
2 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	74	<ul style="list-style-type: none">Kibera (71 rows)kibera (2 rows)Kbera (1 rows)	<input type="checkbox"/>	<input type="text" value="Kibera"/>
2	73	<ul style="list-style-type: none">Kibera (71 rows)kibera (2 rows)	<input type="checkbox"/>	<input type="text" value="Kibera"/>

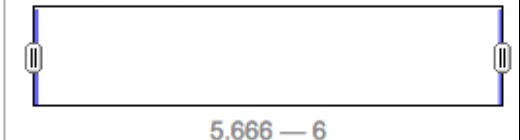
Choices in Cluster



Rows in Cluster



Average Length of Choices



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


 field ▾


Generate Results


Full Datasource - 372.44kB ▾ 561 Columns 199 Rows 5 Data Types Grid ▾


Filter In grid



3 Categories



2 Categories


40 - 83



19 Categories


24 Categories


5 Categories

ABC	intro_location ▾	 gender ▾	#	age ▾	ABC	education ▾	ABC	occupation ▾	ABC	marital_status
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 v
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 v
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		Polvaamous marriade (more than one v

TRANSFORM EDITOR

 Enter transform expression

Add to Script

Trifacta: Filter

The screenshot displays the Trifacta user interface. At the top, a header bar includes a logo, a 'field' dropdown, and a 'Generate Results' button. Below this, a status bar shows 'Full Datasource - 372.44kB', '561 Columns', '199 Rows', '5 Data Types', and a 'Grid' view selector. The main area is divided into a 'Preview' section and a 'SUGGESTIONS' section.

Preview Section: This section shows a data table with columns: 'intro_location', 'gender', '#', 'age', 'education', 'occupation', and 'marital_status'. Each column has a corresponding histogram. The data is filtered to show 110 rows. The 'age' column is highlighted, and the 'SUGGESTIONS' section below it shows filter options for this column.

SUGGESTIONS Section: This section provides four filter suggestions for the 'age' column:

- Keep:** A table with 3 rows (43, 42, 42) and 2 columns (#, age). It affects all columns, 110 rows.
- Delete:** A table with 3 rows (43, 42, 42) and 2 columns (#, age). It affects all columns, 110 rows.
- Set:** A table with 3 rows (43, 62, 69) and 2 columns (#, age). It changes 1 column.
- Derive:** A table with 3 rows (43, 62, 69) and 3 columns (#, age, and a boolean column). It affects 1 column, all rows, and creates 1 column.

intro_location	gender	#	age	education	occupation	marital_status
Kibera	Female	43		College 4	Teacher	"Monogamous couple (one man, one wife)"
Kibera	Male	62		Form 4	Electrician	"Monogamous couple (one man, one wife)"
Kibera	Male	69		None	Casual worker	"Monogamous couple (one man, one wife)"
Kibera	Male	42		None	Not working	Widow/widower
Kibera	Female	42		Form 2	Own a small business	"Monogamous couple (one man, one wife)"
Kibera	Female	41		Form 4	Own a small business	Single
Kibera	Female	47		Form 4	Own a small business	Single
Kibera	Female	43		Standard 5	Own a small business	Widow/widower
Kibera	Female	44		Standard 8	Own a small business	"Monogamous couple (one man, one wife)"
Kibera	Female	42		Standard 8	Own a small business	Polygamous marriage (more than one wife)"
Kibera	Female	40		Standard 8	Own a small business	Single

Trifacta: Scripts

The screenshot displays the Trifacta interface with a data table and a transform editor. The data table has columns: intro_location, gender, #, age, education, occupation, and marital_status. A transform editor is open, showing a transform named 'full_address' with the code: `extract col:full_address on: 'YOUR PATTERN'`. A list of available transforms is shown on the left, including: aggregate, countpattern, deduplicate, delete, derive, drop, extract, extractkv, extractlist, flatten, header, keep, merge, move, and multisplit. The transform editor also includes a 'Next' button and a 'Cancel' button.

intro_location	gender	#	age	education	occupation	marital_status
	2 Categories	40 - 83				5 Categories
	Female	43				"Monogamous · couple · (one · man, · one · wif
	Male	62				"Monogamous · couple · (one · man, · one · wif
	Male	69				"Monogamous · couple · (one · man, · one · wif
	Male	42				Widow/widower
	Female	42				"Monogamous · couple · (one · man, · one · wif
	Female	41				Single
	Female	47				Single
	Female	43				Widow/widower
	Female	44				"Monogamous · couple · (one · man, · one · wif
	Female	42				Polygamous · marriage · (more · than · one · v
	Female	40				Single
	Female	83				Widow/widower
	Female	43				"Monogamous · couple · (one · man, · one · wif
	Female	53				Polygamous · marriage · (more · than · one · v
	Female	60				Separated · or · Divorced
	Male	49				"Monogamous · couple · (one · man, · one · wif
	Male	46				Polvaamous · marriade · (more · than · one · v

Transform Editor 1 of 2

full_address

`extract col:full_address on: 'YOUR PATTERN'`

Create & Edit Transforms

The **transform editor** helps you create and edit transforms with typeahead and syntax highlighting. For help crafting transforms, refer to our [Language Docs](#).

[Learn more](#)

[Don't show me any helpers](#)

[Next](#)

[Cancel](#) [Add to Script](#)

Trifacta: Scripting Without Code

Full Datasource - 372.44kB | 562 Columns | 199 Rows | 6 Data Types | Grid | Columns: All | Transformed - 2 Columns | Filter in grid

intro_location	gender	#	age	education	education1	occupation	
3 Categories	2 Categories	40 - 83	19 Categories	5 Categories	24 Categories	5 Categories	
1 Kibera	Female	43	College · 4	College	Teacher	"Monogamou	
2 Kibera	Male	62	Form · 4	Form	Electrician	"Monogamou	
3 Kibera	Male	69	None	None	Casual · worker	"Monogamou	
4 Kibera	Male	42	None	None	Not · working	Widow/widc	
5 Kibera	Female	42	Form · 2	Form	Own · a · small · business	"Monogamou	
6 Kibera	Female	41	Form · 4	Form	Own · a · small · business	Single	
7 Kibera	Female	47	Form · 4	Form	Own · a · small · business	Single	
8 Kibera	Female	43	Standard · 5	Standard	Own · a · small · business	Widow/widc	
9 Kibera	Female	44	Standard · 8	Standard	Own · a · small · business	"Monogamou	
10 Kibera	Female	42	Standard · 8	Standard	Own · a · small · business	Polygamous	
11 Kibera	Female	40	Standard · 8	Standard	Own · a · small · business	Single	

SUGGESTIONS

Extract on: `{alpha}+`

education	education1
College · 4	College
Form · 4	Form
None	None

Affects 1 column, all rows | Creates 1 column

Replace on: `{start}{alpha}+`

education	#	education
College · 4	· 4	
Form · 4	· 4	
None		

Changes 1 column

Countpattern on: `{start}{alpha}+`

education	#
College · 4	1
Form · 4	1
None	1

Affects 1 column, all rows | Creates 1 column

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.