


DATA CLEANING TUTORIAL

PETRA ISENBERG

Information Visualization

LOADING DATA

 **OpenRefine** *A power tool for working with messy data.*

Create Project

Open Project

Import Project

Language Settings

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Get data from

Locate one or more files on your computer to upload:

This Computer

No files selected.

Web Addresses (URLs)

Clipboard

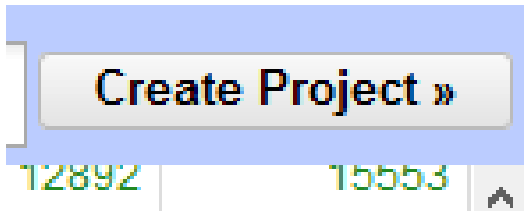
Data Package (JSON URL)

Database

Google Data

CONFIGURE PARSING OPTIONS

Parse cell text into
numbers, dates, ...



Facet / Filter Undo / Redo 0 / 0

75043 rows Extensions: Wikidata

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

	All	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgr
☆ ↻	1.	Paris Universitاس	15	5500	8000	France		2005	
☆ ↻	2.	Paris Universitاس	15	5500	8000	France		2005	
☆ ↻	3.	Lumi%C3%A8re University Lyon 2	121		1355	France		1835	70
☆ ↻	4.	Confederation College	4700000			Canada		1967	not available
☆ ↻	5.	Rocky Mountain College	16586100			United States		1878	
☆ ↻	6.	Rocky Mountain College	16586100			USA		1878	
☆ ↻	7.	Idaho State University	40200750	838		United States	1269	1901	26
☆ ↻	8.	Idaho State University	40200750	838		USA	1269	1901	26
☆ ↻	9.	Idaho State University	40200750	838		United States	1269	1947	26
☆ ↻	10.	Idaho State University	40200750	838		USA	1269	1947	26

CLEAN UP COUNTY NAMES

UniversityData.csv [Permalink](#)

Open... Export Help

75055 rows

Extensions: Freebase RDF

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

« first < previous 1 - 10 next > last »

	All	x	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	
☆	🔊	1.	Paris Universitas	15	5500	8000		2005		
☆	🔊	2.	Paris Universitas	15	5500	8000		2005		
☆	🔊	3.	Lumi%C3%A8re University Lyon 2	121		1355		1835		7046
☆	🔊	4.	Confederation College	4700000						
☆	🔊	5.	Rocky Mountain College	16586100						66
☆	🔊	6.	Rocky Mountain College	16586100						66
☆	🔊	7.	Idaho State University	40200750	838					2661
☆	🔊	8.	Idaho State University	40200750	838					2661
☆	🔊	9.	Idaho State University	40200750	838					2661
☆	🔊	10.	Idaho State University	40200750	838					2661

- Facet
- Text filter
- Edit cells
 - Transform...
 - Common transforms
 - Fill down
 - Blank down
 - Split multi-valued cells...
 - Join multi-valued cells...
 - Cluster and edit...
- Edit column
- Transpose
- Sort...
- View
- Reconcile
 - States
 - USA

Filter:

Cluster & Edit column "country"

This feature helps you find groups of different cell values that might be different presentations 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 ...](#)

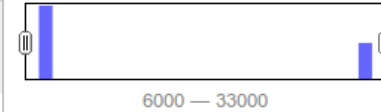
Method **key collision**

Keying Function **fingerprint**

3 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
1	6603	<ul style="list-style-type: none"> US (3004 rows) US (2609 rows) 	<input checked="" type="checkbox"/>	United States
2	32034	<ul style="list-style-type: none"> United States (32033 rows) United States) (1 rows) 	<input checked="" type="checkbox"/>	United States
2	6795	<ul style="list-style-type: none"> USA (6402 rows) U.S.A. (393 rows) 	<input checked="" type="checkbox"/>	United States

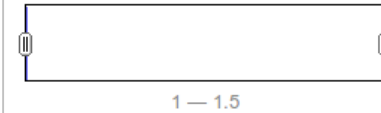
Rows in Cluster



Average Length of Choices



Length Variance of Choices



Select All Deselect All

Merge Selected & Re-Cluster

Merge Selected & Close

Close

OF STUDENTS

universityData.csv [Permalink](#) Open... Export ▾ Help

Undo / Redo 7 **75055 rows** Extensions: Freebase ▾ RDF ▾

Extract... Apply... Show as: **rows** records Show: 5 10 25 50 rows « first ‹ previous 1 - 10 next › last »

	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
	15	5500	8000	France					70000
	15	5500	8000	France					70000
University Lyon 2	121		1355	France					27393
	4700000			Canada					21160
	16586100			United States					894
	16586100			United States					894
	40200750	838		United States	1269				15553
	40200750	838		United States	1269	1901	2661		15553
	40200750	838		United States	1269	1947	2661	12892	15553
	40200750	838		United States	1269	1947	2661	12892	15553

Facet menu options: Text facet, Numeric facet, Timeline facet, Scatterplot facet, Custom text facet..., Custom numeric facet..., Customized facets

Facet sub-menu options: Facet, Text filter, Edit cells, Edit column, Transpose, Sort..., View, Reconcile

What do you notice?

OF STUDENTS

Google refine universityData.csv Permalink

Open... Export ▾ Help

Facet / Filter Undo / Redo 7

4702 matching rows (75055 total)

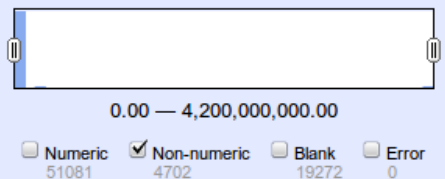
Extensions: Freebase ▾ RDF ▾

Refresh Reset All Remove All

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

« first < previous 1 - 10 next > last »

numStudents change reset



endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
42000000			Honduras	?	1941	Does not offer postgraduate studies	900+	900+
7920	4135		Philippines	6491	18		41991	~50,000
7920	4571		Philippines	6491	18		41991	~50,000
7920	4135		Philippines	213180	18		41991	~50,000
7920	4571		Philippines	213180	18		41991	~50,000
7920	4135		Philippines	6491	18		41991	~50,000
7920	4571		Philippines	6491	18		41991	~50,000
7920	4135		Philippines	213180	18		41991	~50,000
7920	4571		Philippines	213180	18		41991	~50,000
41600000	596		Canada	915	1964	621	2868	http://www.brocku.ca/athle

OF STUDENTS

Google refine universityData.csv Permalink

Open... Export Help

Facet / Filter Undo / Redo 7

4702 matching rows (75055 total)

Extensions: Freebase RDF

Refresh Reset All Remove All

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

« first < previous 1 - 10 next > last »

numStudents change reset

0.00 — 4,200,000,000.00

Numeric 51081 Non-numeric 4702 Blank 19272 Error 0

endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
42000000			Honduras	?	1941	Does not offer postgraduate studies		
7920	4135		Philippines	6491				
7920	4571		Philippines	6491				
7920	4135		Philippines	213180				
7920	4571		Philippines	213180				
7920	4135		Philippines	6491				
7920	4571		Philippines	6491				
7920	4135		Philippines	213180			41991	~50,000
7920	4571		Philippines	213180	18		41991	~50,000
41600000	596		Canada	915	1964	621	2868	http://www.brocku.ca/athle

- Transform...
- Common transforms
- Fill down
- Blank down
- Split multi-valued cells...
- Join multi-valued cells...
- Cluster and edit...

- Facet
- Text filter
- Edit cells
- Edit column
- Transpose
- Sort...
- View
- Reconcile

OF STUDENTS

Custom text transform on column numStudents

Expression `value.replace("+", "")` Language **Google Refine Expression Language (GREL)** No syntax error.

Preview History Starred Help

row	value	value.replace("+", "")
155.	900+	900
343.	~50,000	~50,000
344.	~50,000	~50,000
347.	~50,000	~50,000
348.	~50,000	~50,000
351.	~50,000	~50,000

On error keep original Re-transform up to times until no change
 set to blank
 store error

OK Cancel

`value.replace("+", "")`

OF STUDENTS

Custom text transform on column numStudents

Expression Language Google Refine Expression Language (GREL) ▼

`value.replace("+", "")` No syntax error.

Preview History Starred Help

row	value	value.replace("+", "")
155.	900+	900
343.	~50,000	~50,000
344.	~50,000	~50,000
347.	~50,000	~50,000
348.	~50,000	~50,000
351.	~50,000	~50,000

On error keep original Re-transform up to times until no change
 set to blank
 store error

OK Cancel

"Lumi%A8re University Lyon 2"
value.unescape('url')

OF STUDENTS

Google refine universityData.csv Permalink

Open... Export Help

Facet / Filter Undo / Redo 8

4702 matching rows (75055 total)

Extensions: Freebase RDF

Refresh

Reset All Remove All

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

« first < previous 1 - 10 next > last »

numStudents change reset

0.00 — 4,200,000,000.00

Numeric 51081 Non-numeric 4702 Blank 19272 Error 0

numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
00000		Honduras	?	1941	Does not offer postgraduate studies	900+	
7920	4135	Philippines	6491	18			
7920	4571	Philippin					
7920	4135	Philippir					
7920	4571	Philippir					
7920	4135	Philippir					
7920	4571	Philippir					
7920	4135	Philippir					
7920	4571	Philippir				41991	50000
00000	596	Canada			621	2868	http://www.brocku.ca/athletics/quickfacts.php

Transform...

- Common transforms
- Fill down
- Blank down
- Split multi-valued cells...
- Join multi-valued cells...
- Cluster and edit...

- Trim leading and trailing whitespace
- Collapse consecutive whitespace
- Unescape HTML entities
- To titlecase
- To uppercase
- To lowercase
- To number
- To date
- To text
- Blank out cells

- Facet
- Text filter
- Edit cells
- Edit column
- Transpose
- Sort...
- View
- Reconcile

REMOVING UNWANTED ROWS

Facet / Filter Undo / Redo 0

Refresh Reset All Remove All

numStudents change reset

0.00 — 4,200,000,000.00

Numeric 51079 Non-numeric 4695 Blank 19269 Error 0

4695 matching rows (75043 total)

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

All	university	endowment
Facet	iano	42000000
Edit rows		
Edit columns		
View		
☆	347.	Unive Philipp
☆	348.	Unive Philipp
☆	351.	University of the Philippines

Remove all matching rows

ENDOWMENT

Google refine universityData.csv Permalink

Open... Export Help

Facet / Filter Undo / Redo 16

Refresh Reset All Remove All

21591 matching rows (51826 total) Extensions: Freebase RDF

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 50 next > last »

endowment change reset

0.00 — 860,000,000,000.00

Numeric 30235 Non-numeric 21591 Blank 0 Error 0

All	x	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	
☆	23.	University of Seoul	N/A	372	South Korea	1229	1918-05-01	29	
☆	24.	Toho University	N/A	705	154	Japan	3365	1925	4
☆	25.	Korea National University of Education	N/A	274	South Korea	508	Established 1985	34	
☆	26.	Korea National University of Education	N/A	274	South Korea	508	Chartered 1984	34	
☆	128.	Ithaca College	US \$186 million	673	United States	989	1892	4	
☆	157.	University of Utah	US\$513.4 million	2687	United States	14362	1850-02-28	74	
☆	166.	University of Florida	US\$1.3 billion	4534	United States		1853	169	
☆	167.	University of Florida	US\$1.3 billion	5081	United States		1853	169	

What do you notice?

ENDOWMENT

Probably not a good idea, but for now we assume everything is in \$

-> **Edit cells** -> **Transform**

```
value.replace("US $", "").replace("US$", "")
```

CONVERT TO LC

Google refine universityData.csv Permalink

Open... Export Help

Facet / Filter Undo / Redo 17

21591 matching rows (51826 total)

Extensions: Freebase RDF

Refresh Reset All Remove All

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

« first < previous 1 - 50 next > last »

endowment change reset

0.00 — 860,000,000,000.00

Numeric 30235 Non-numeric 21591 Blank 0 Error 0

All	x	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad
☆	23.	University of Seoul	372		South Korea	1229	1918-05-01	29
☆	24.	Toho University	705	154	Japan	3365	1925	4
☆	25.	Korea National University of Education		274	South Korea	508	Established 1985	34
☆	26.	Korea National University of Education						34
☆	128.	Ithaca College						4
☆	157.	University of Utah						74
☆	166.	University of Florida						169
☆	167.	University of Florida						169
☆	168.	University of Florida						169
☆	169.	University of Florida						169
☆	170.	University of Florida	1.3 billion					180
☆	171.	University of Florida	1.3 billion	5081				180
☆	172.	University of Florida	1.3 billion	4534				180
☆	173.	University of Florida	1.3 billion	5081				180
☆	174.	University of Florida	1.3 billion	4534				169
☆	175.	University of Florida	1.3 billion	5081				169
☆	176.	University of Florida	1.3 billion	4534				169
☆	177.	University of Florida	1.3 billion	5081	United States		1853	169

- Facet
- Text filter
- Edit cells
 - Transform...
- Edit column
 - Common transforms
 - Trim leading and trailing whitespace
 - Collapse consecutive whitespace
 - Unescape HTML entities
 - To titlecase
 - To uppercase
 - To lowercase
 - To number
 - To date
 - To text
 - Blank out cells
- Transpose
 - Fill down
- Sort...
- View
 - Split multi-valued cells...
- Reconcile
 - Join multi-valued cells...
- Cluster and edit...

CONVERT TO NUMBERS

\$13.8 million

What could we do here?

```
toNumber(value.replace(" million", ""))*1000000
```

DEDUPLICATION

Dataset has a lot of duplicate rows

-> university names -> sort -> (image below)

The screenshot shows a data table interface with the following elements:

- Header:** "Google refine universityData.csv Permalink" on the left, and "Open...", "Export", and "Help" buttons on the right.
- Table Status:** "46203 rows" and "Extensions: Freebase, RDF".
- Table Controls:** "Facet / Filter", "Undo / Redo 22", "Refresh", "Reset All", and "Remove All".
- Table Columns:** "All", "x", "endowment", "numFa", "Staff", "established", "numPostgrad", "numUnderg".
- Table Content:** A list of rows with columns for ID, University Name, Endowment, and other metrics. The first few rows are for Aarhus University with an endowment of 5270000000.
- Sort Menu:** A dropdown menu is open over the "endowment" column, showing options: "Remove sort", "Reorder rows permanently" (highlighted), and "By x".
- Facet Panel:** A panel for the "endowment" column with a range of "0.00 — 860,000,000,000.00" and checkboxes for "Numeric" (checked), "Non-numeric", "Blank", and "Error".

DEDUPLICATION

Column with university names, **Edit cells -> Blank down**
Then on the same column, **Facet -> Customized facets -> Facet by blank**

The screenshot shows the Google Refine interface for a dataset named 'universityData.csv'. The main table has 46,203 rows and columns for 'endowment', 'numFaculty', 'numDoctoral', 'country', 'numStaff', 'established', 'numPostgrad', and 'numUndergrad'. A context menu is open over the 'endowment' column, with the 'View' option expanded to show 'Customized facets'. The 'Facet by blank' option is highlighted. On the left, a 'Facet / Filter' panel for 'endowment' is visible, showing a range from 0.00 to 860,000,000,000.00 and checkboxes for 'Numeric', 'Non-numeric', 'Blank', and 'Error'. The 'Numeric' checkbox is checked.

select **true**, then on the "All" column on the left,
Edit rows -> Remove all matching rows