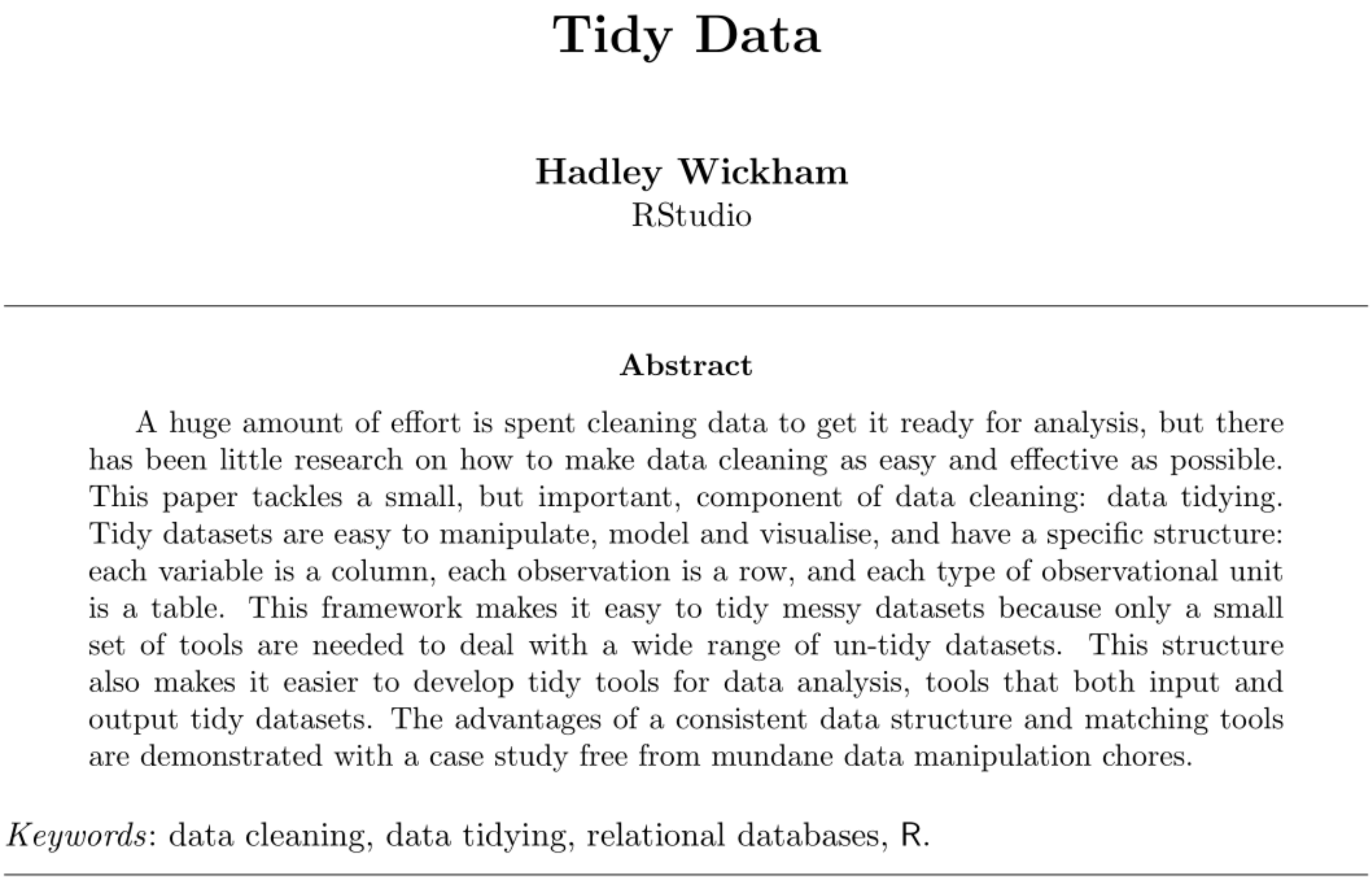
DATA CLEANING TUTORIAL

PETRA ISENBERG

Information

TIDY DATA PRINCIPLES



Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

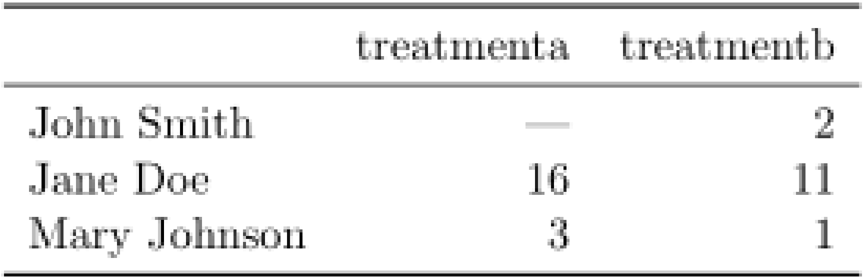
Keywords: data cleaning, data tidying, relational databases, R.

TIDY DATA

= data structured to facilitate analysis

labelled columns

labelled rows



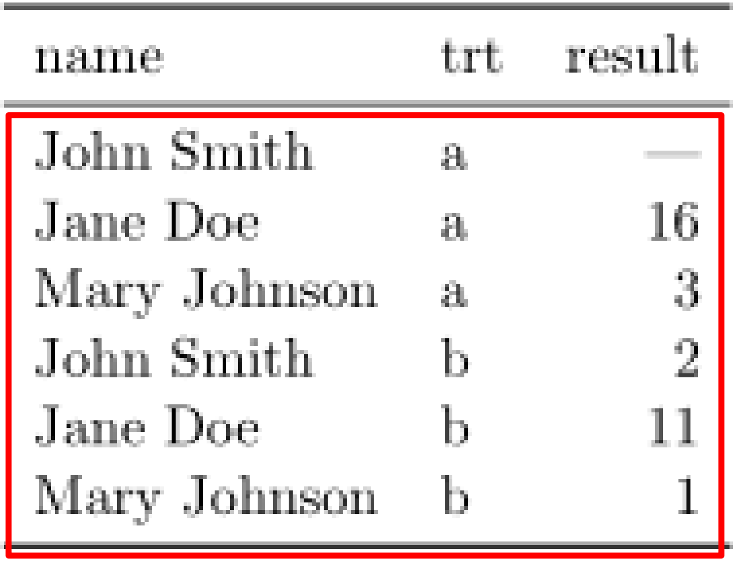
	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

= data structure

TIDY DATA

Data semantics

Attributes, variables
= column names



name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Items,
observations
= rows

values

TIDY DATA

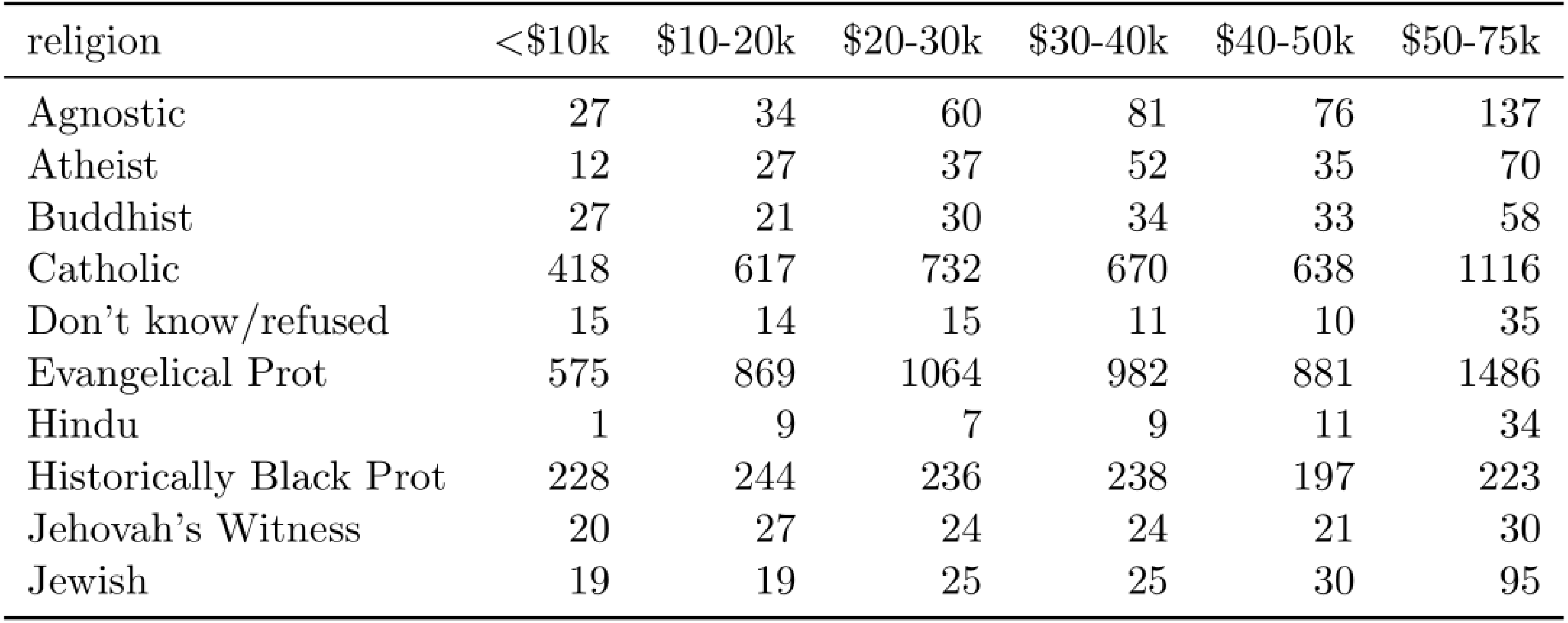
- Variables are columns
- Observations are rows
- Each observational unit in one table

In addition: put fixed variables first and then measured variables last

If you order, do so by the first variable

MESSY DATA - EXAMPLES

Column headers = values, not variables

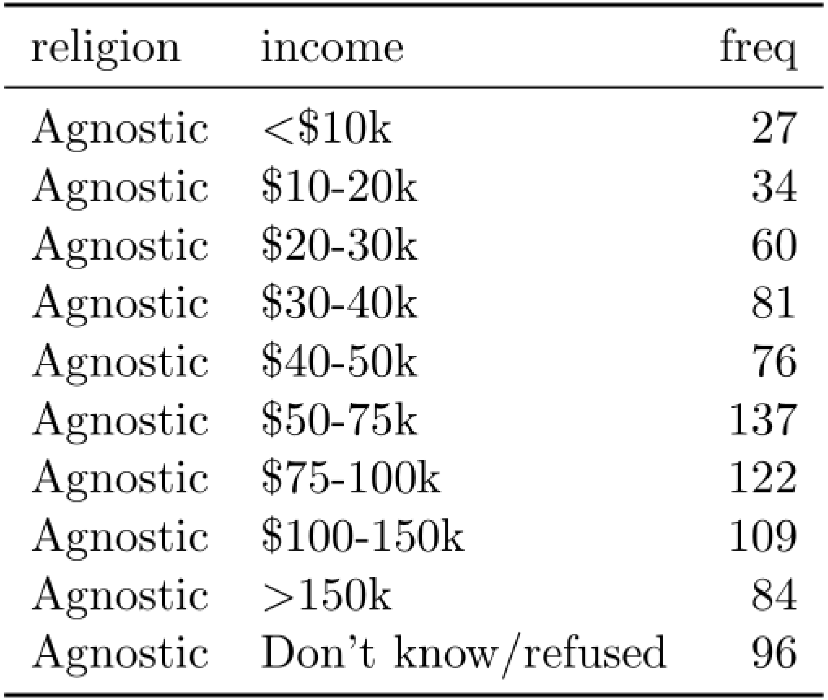


religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

MESSY DATA - EXAMPLES

Better (most of the time)

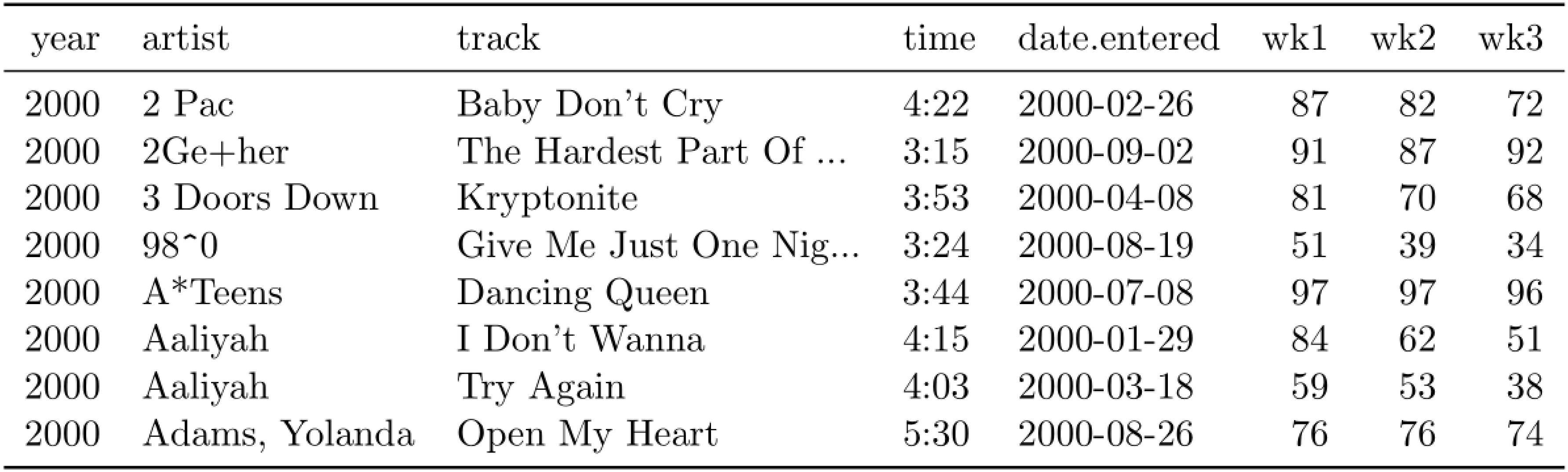
Process to produce this = melting



religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

YOU!

This table is good for data entry but not analysis. How do we tidy it up?

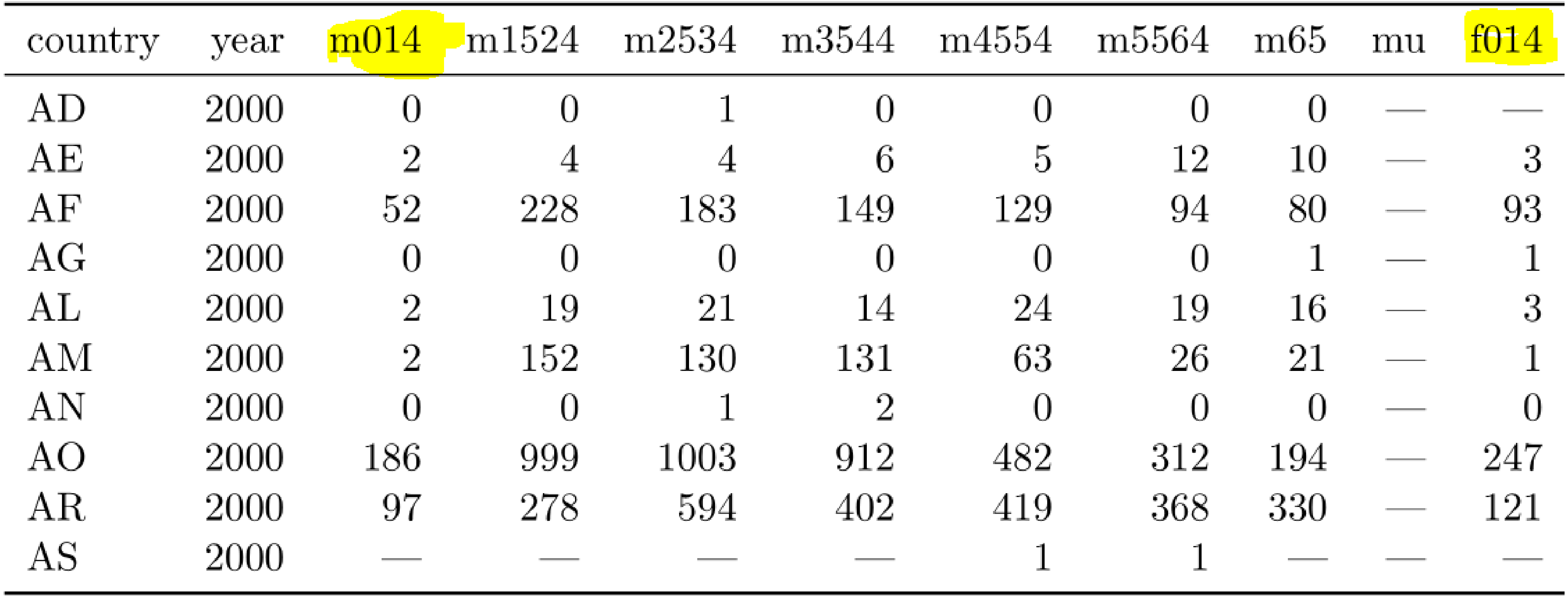


year	artist	track	time	date.entered	wk1	wk2	wk3
2000	2 Pac	Baby Don't Cry	4:22	2000-02-26	87	82	72
2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	91	87	92
2000	3 Doors Down	Kryptonite	3:53	2000-04-08	81	70	68
2000	98^0	Give Me Just One Nig...	3:24	2000-08-19	51	39	34
2000	A*Teens	Dancing Queen	3:44	2000-07-08	97	97	96
2000	Aaliyah	I Don't Wanna	4:15	2000-01-29	84	62	51
2000	Aaliyah	Try Again	4:03	2000-03-18	59	53	38
2000	Adams, Yolanda	Open My Heart	5:30	2000-08-26	76	76	74

year	artist	time	track	date	week	rank
2000	2 Pac	4:22	Baby Don't Cry	2000-02-26	1	87
2000	2 Pac	4:22	Baby Don't Cry	2000-03-04	2	82
2000	2 Pac	4:22	Baby Don't Cry	2000-03-11	3	72
2000	2 Pac	4:22	Baby Don't Cry	2000-03-18	4	77
2000	2 Pac	4:22	Baby Don't Cry	2000-03-25	5	87
2000	2 Pac	4:22	Baby Don't Cry	2000-04-01	6	94
2000	2 Pac	4:22	Baby Don't Cry	2000-04-08	7	99
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-02	1	91
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-09	2	87
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-16	3	92
2000	3 Doors Down	3:53	Kryptonite	2000-04-08	1	81
2000	3 Doors Down	3:53	Kryptonite	2000-04-15	2	70
2000	3 Doors Down	3:53	Kryptonite	2000-04-22	3	68
2000	3 Doors Down	3:53	Kryptonite	2000-04-29	4	67
2000	3 Doors Down	3:53	Kryptonite	2000-05-06	5	66

MESSY DATA - EXAMPLES

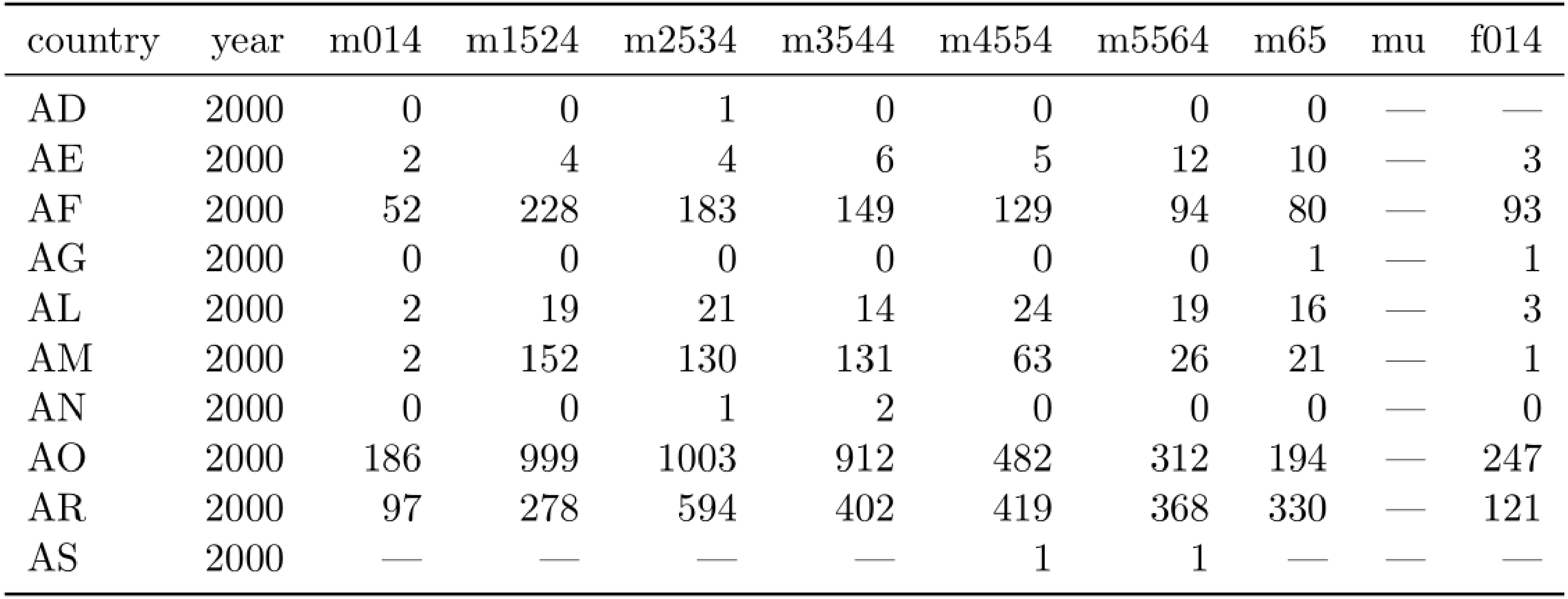
Multiple variables in one column



country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

FIRST WE MELT

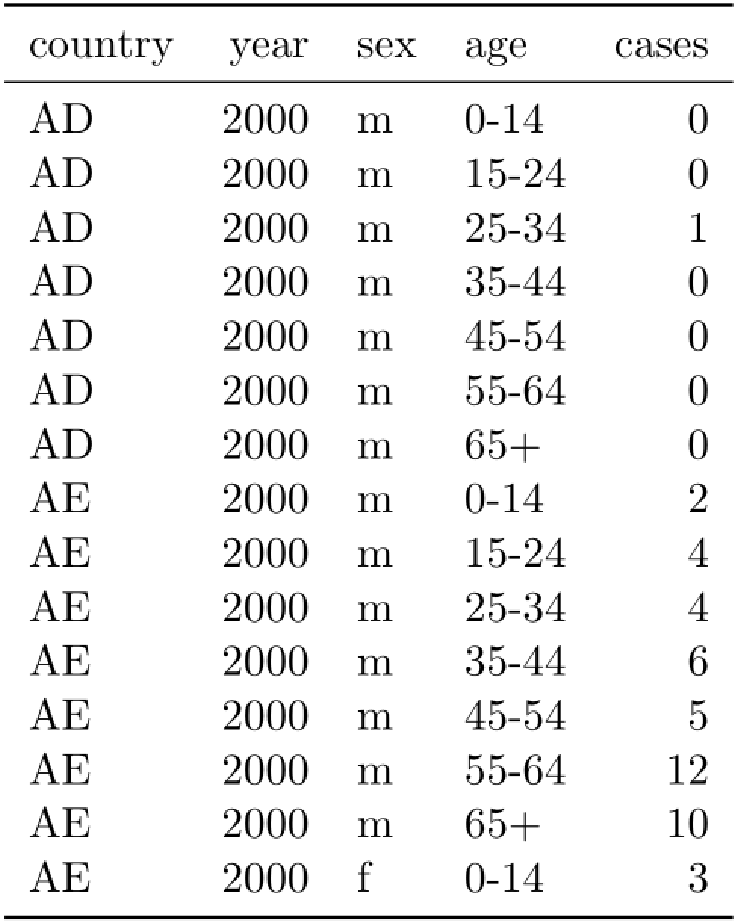
How do we do this…?



country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

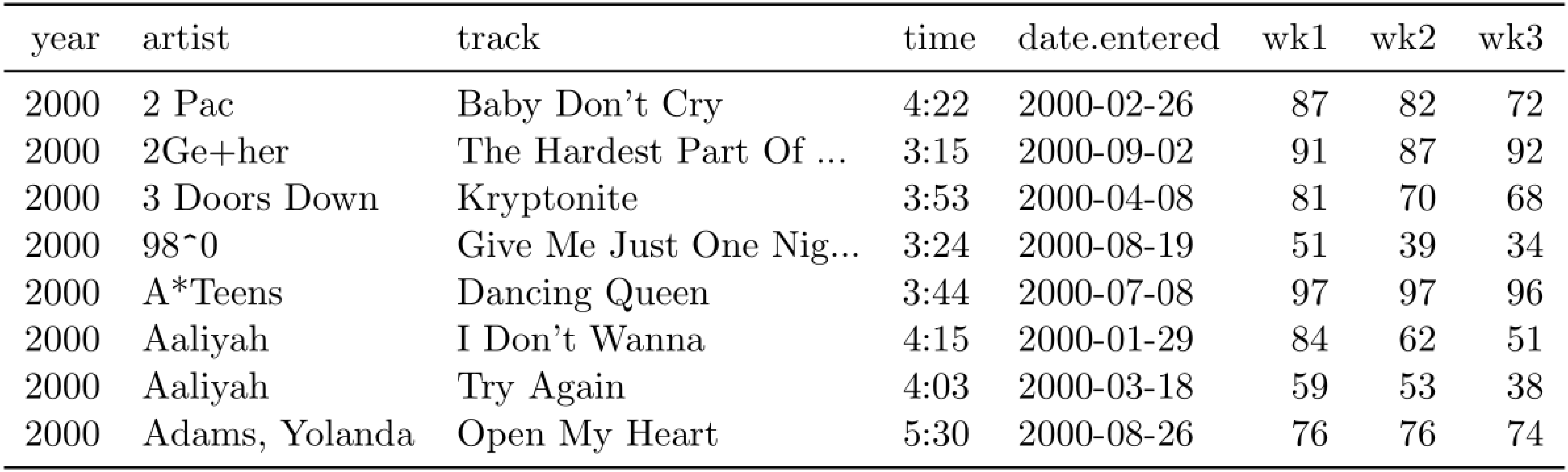
NEXT: SPLIT COLUMNS



country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

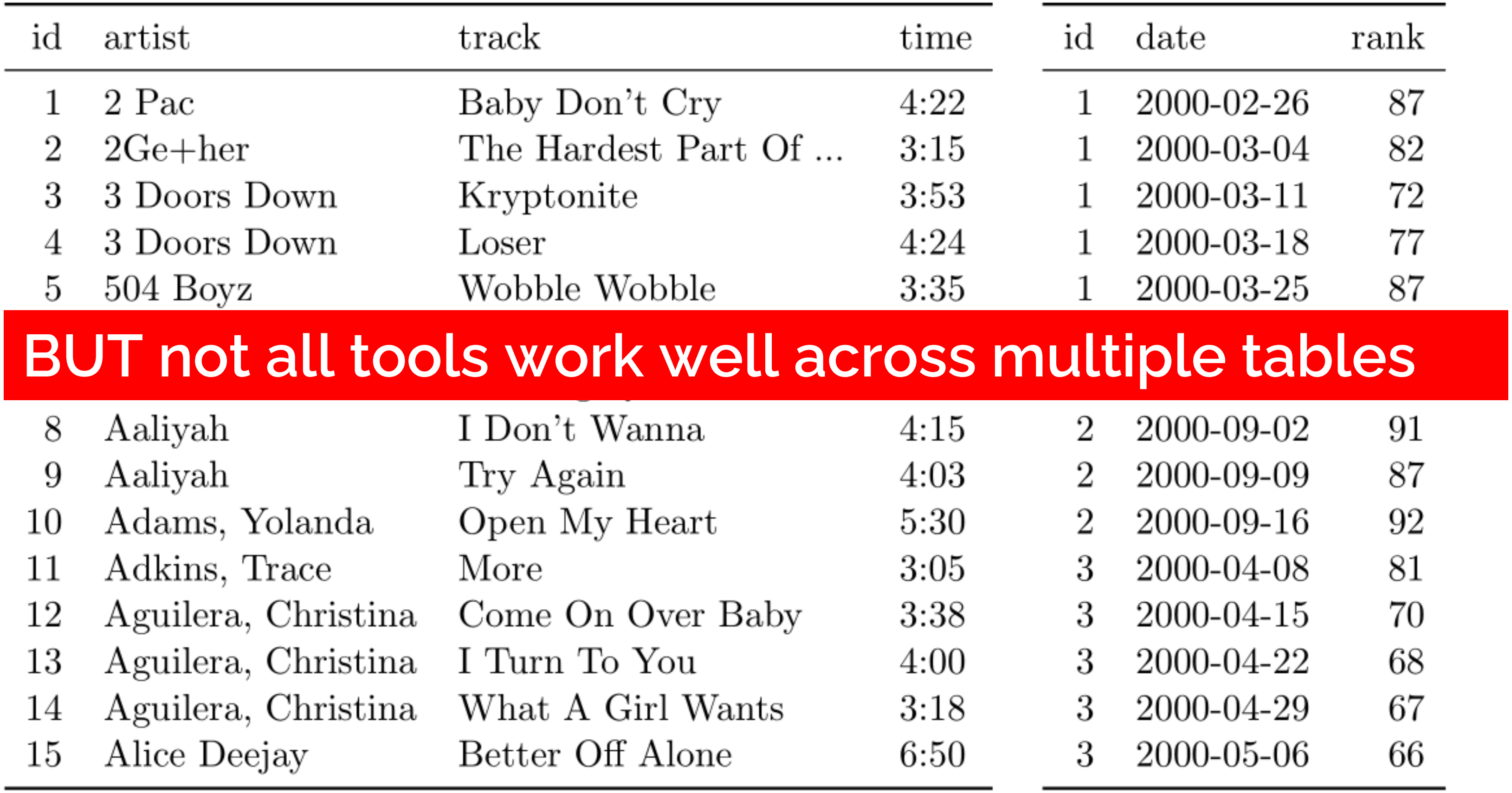
MESSY DATA - EXAMPLES

Multi observational units in the same table



year	artist	track	time	date.entered	wk1	wk2	wk3
2000	2 Pac	Baby Don't Cry	4:22	2000-02-26	87	82	72
2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	91	87	92
2000	3 Doors Down	Kryptonite	3:53	2000-04-08	81	70	68
2000	98^0	Give Me Just One Nig...	3:24	2000-08-19	51	39	34
2000	A*Teens	Dancing Queen	3:44	2000-07-08	97	97	96
2000	Aaliyah	I Don't Wanna	4:15	2000-01-29	84	62	51
2000	Aaliyah	Try Again	4:03	2000-03-18	59	53	38
2000	Adams, Yolanda	Open My Heart	5:30	2000-08-26	76	76	74

TIDYER & MORE SPACE EFFICIENT

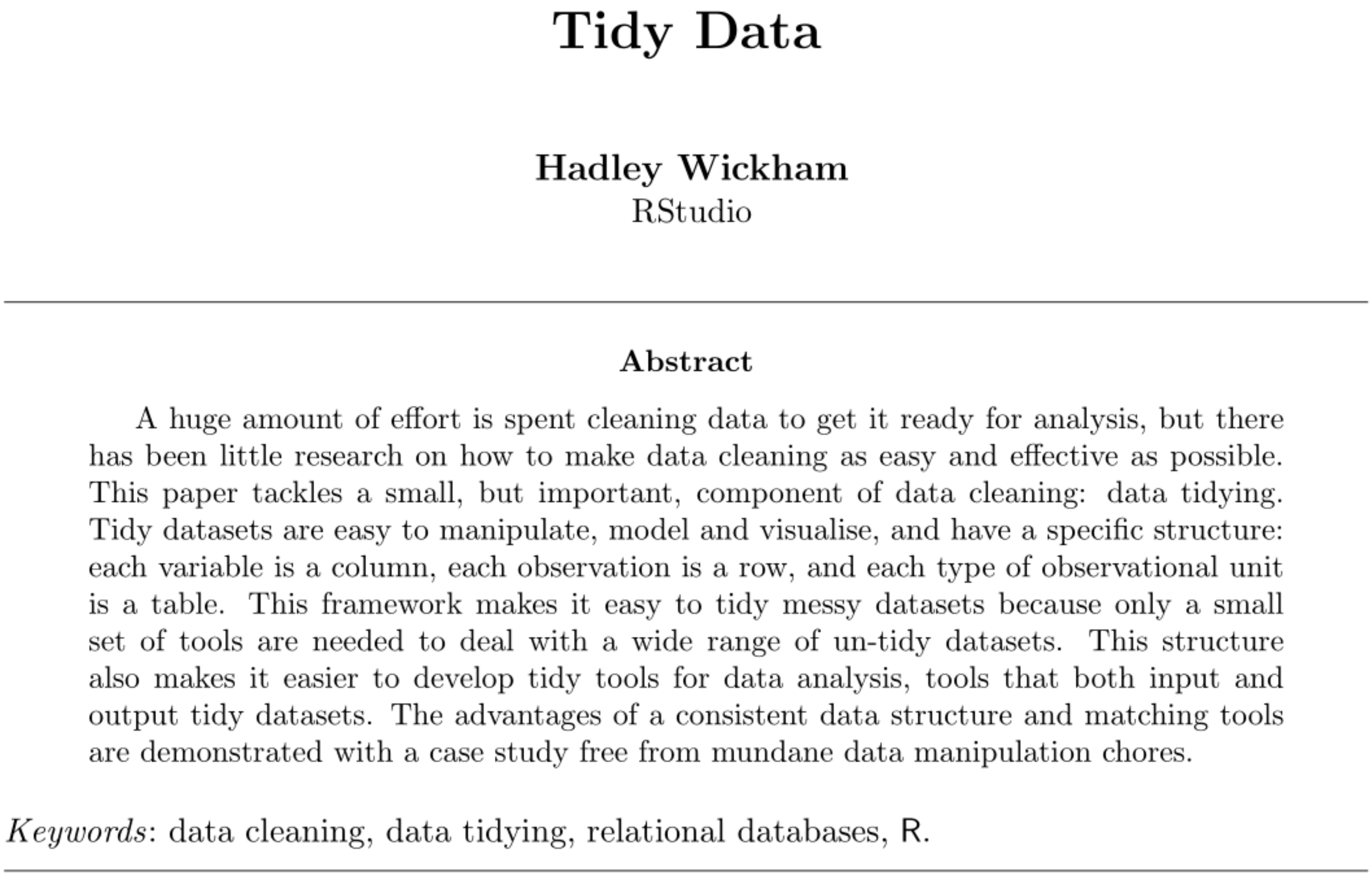


id	artist	track	time
1	2 Pac	Baby Don't Cry	4:22
2	2Ge+her	The Hardest Part Of ...	3:15
3	3 Doors Down	Kryptonite	3:53
4	3 Doors Down	Loser	4:24
5	504 Boyz	Wobble Wobble	3:35
8	Aaliyah	I Don't Wanna	4:15
9	Aaliyah	Try Again	4:03
10	Adams, Yolanda	Open My Heart	5:30
11	Adkins, Trace	More	3:05
12	Aguilera, Christina	Come On Over Baby	3:38
13	Aguilera, Christina	I Turn To You	4:00
14	Aguilera, Christina	What A Girl Wants	3:18
15	Alice Deejay	Better Off Alone	6:50

id	date	rank
1	2000-02-26	87
1	2000-03-04	82
1	2000-03-11	72
1	2000-03-18	77
1	2000-03-25	87
2	2000-09-02	91
2	2000-09-09	87
2	2000-09-16	92
3	2000-04-08	81
3	2000-04-15	70
3	2000-04-22	68
3	2000-04-29	67
3	2000-05-06	66

BUT not all tools work well across multiple tables

MORE EXAMPLES HERE



Tidy Data

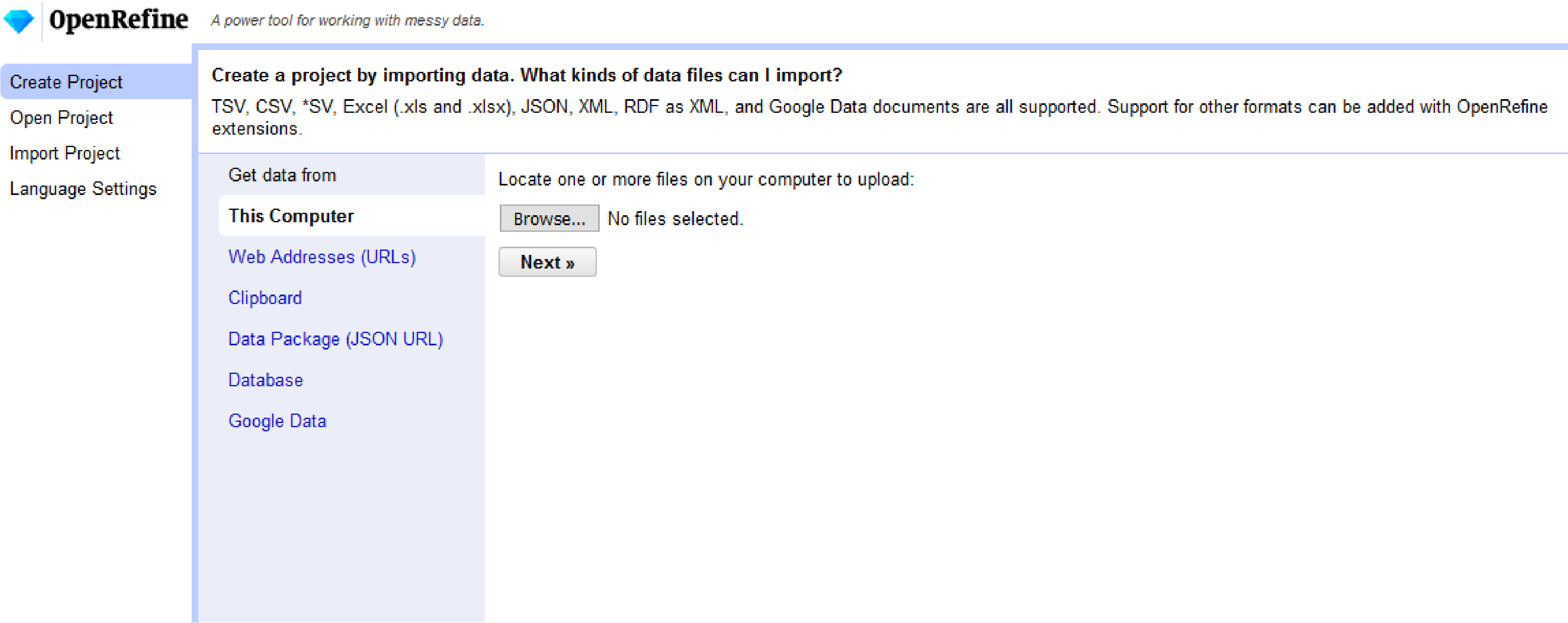
Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

LOADING DATA



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

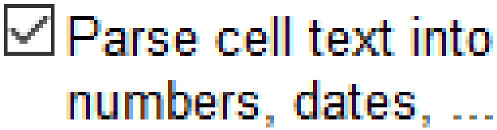
TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Locate one or more files on your computer to upload:

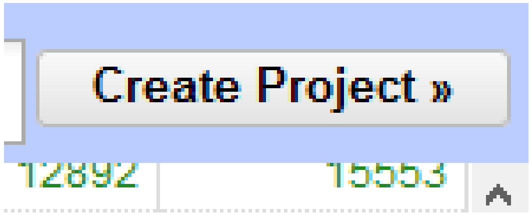
No files selected.

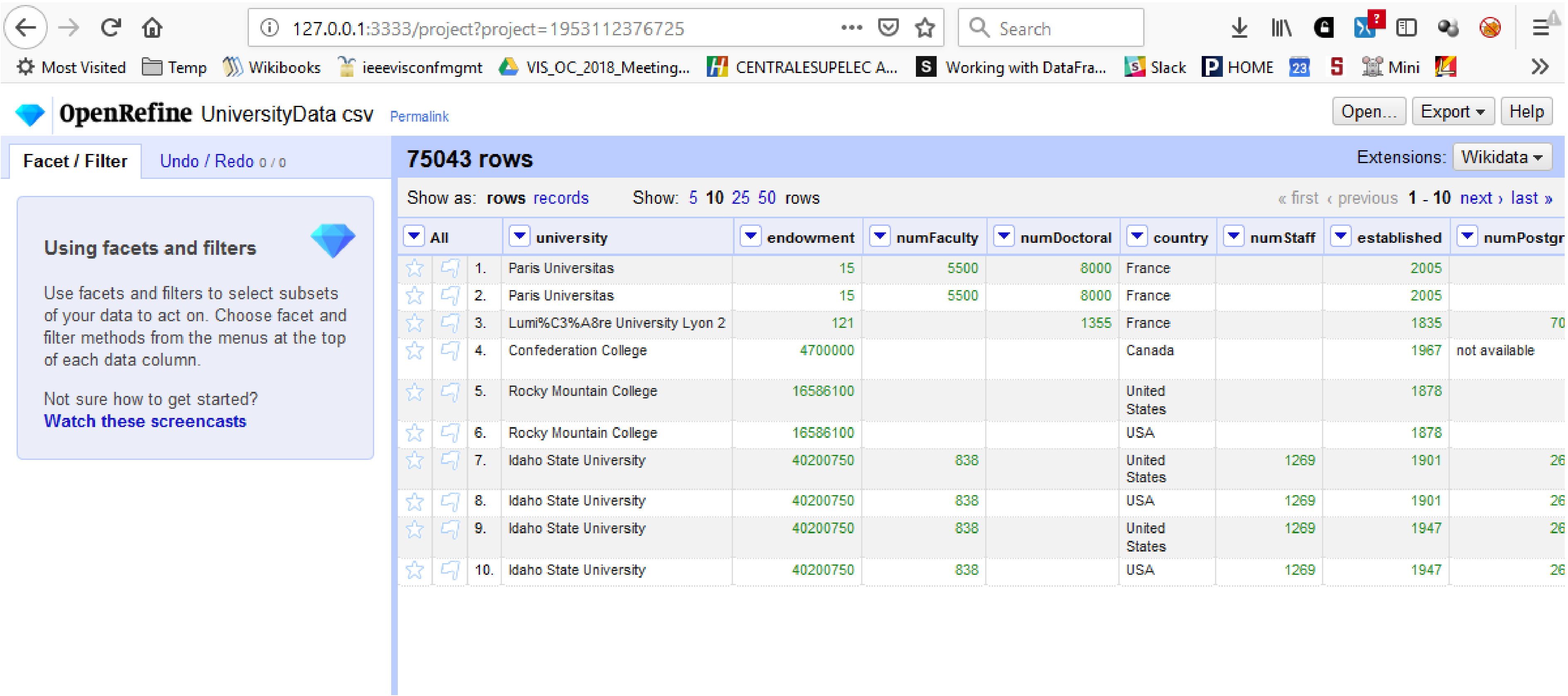
Tutorial from: http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial & https://github.com/gmuggs/teaching/wiki/Openrefine

CONFIGURE PARSING OPTIONS



☑ Parse cell text into numbers, dates, ...





OpenRefine UniversityData csv

Facet / Filter   Undo / Redo 0 / 0

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.

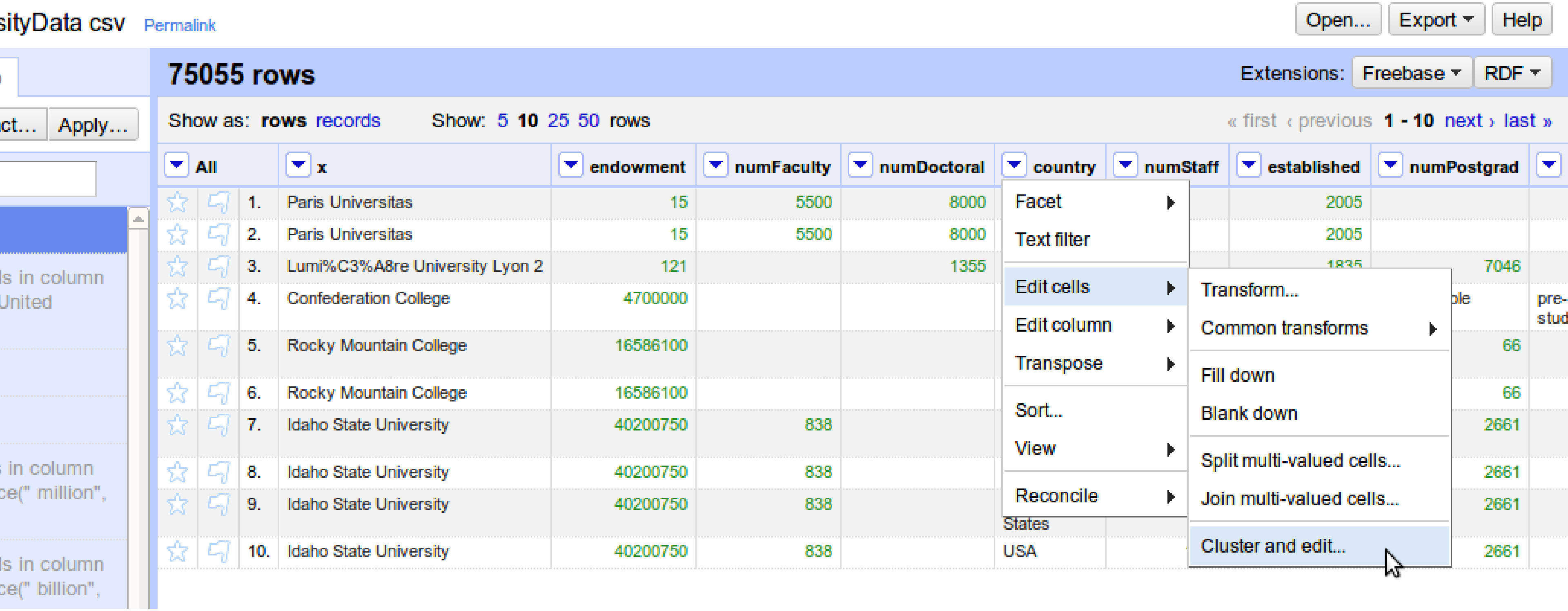
Not sure how to get started?
Watch these screencasts

75043 rows

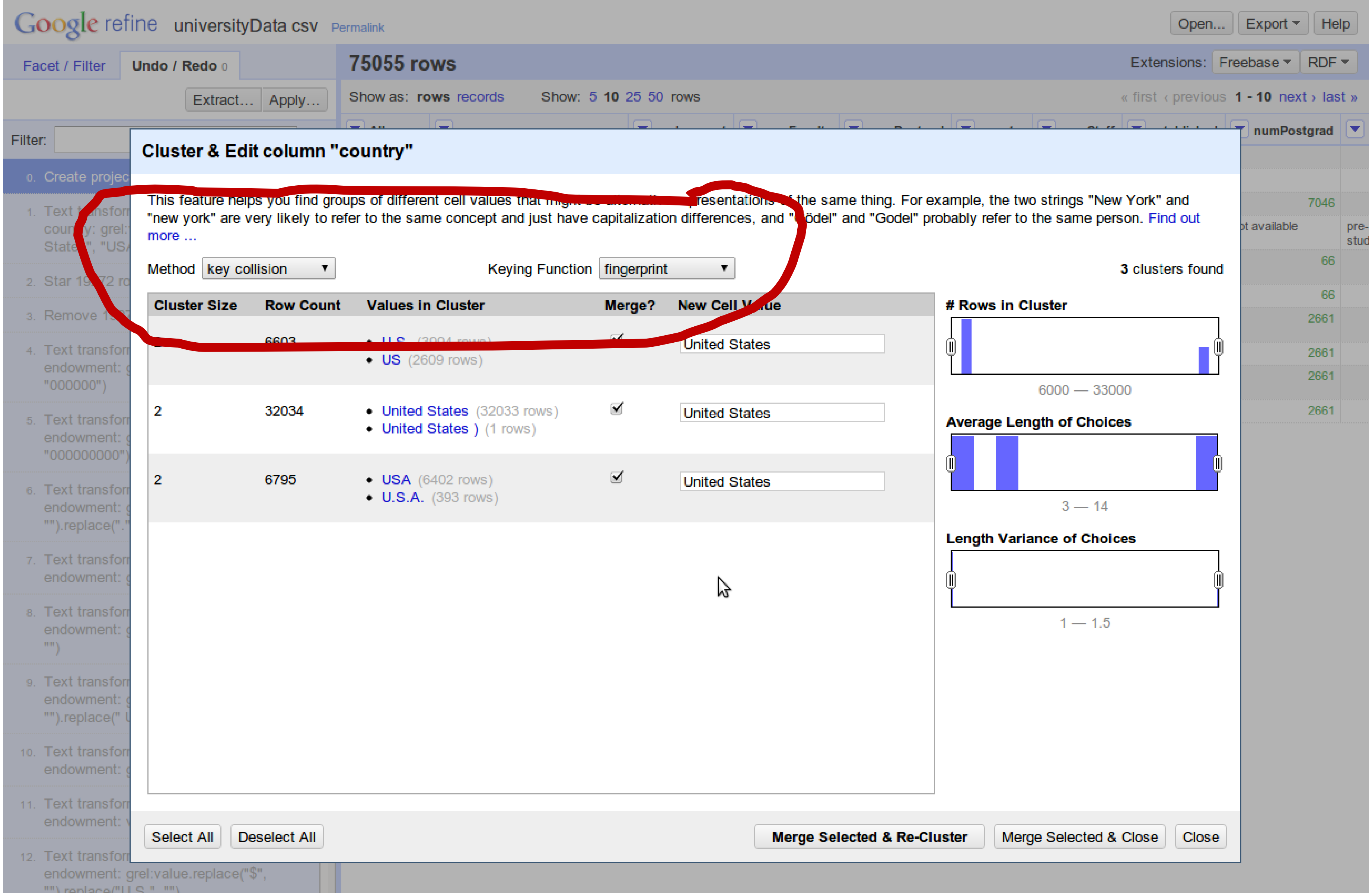
Show as: **rows** records   Show: 5 **10** 25 50 rows   « first ‹ previous **1 - 10** next › last »

All	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgr
1.	Paris Universitas	15	5500	8000	France		2005	
2.	Paris Universitas	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



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						ble	pre- stud
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		States			2661	
10.	Idaho State University	40200750	838		USA			2661	



0. Create projec

1. Text transfor… country: grel:… State", "USA

2. Star 1972 ro

3. Remove 1…

4. Text transfor… endowment: g… "000000")

5. Text transfor… endowment: g… "000000000")

6. Text transfor… endowment: g… "").replace(".

7. Text transfor… endowment: g

8. Text transfor… endowment: g… "")

9. Text transfor… endowment: g… "").replace(" U

10. Text transfor… endowment: g

11. Text transfor… endowment: v

12. Text transfor… endowment: grel:value.replace("$", "").replace("U.S.", "")

Cluster & Edit column "country"

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 ...

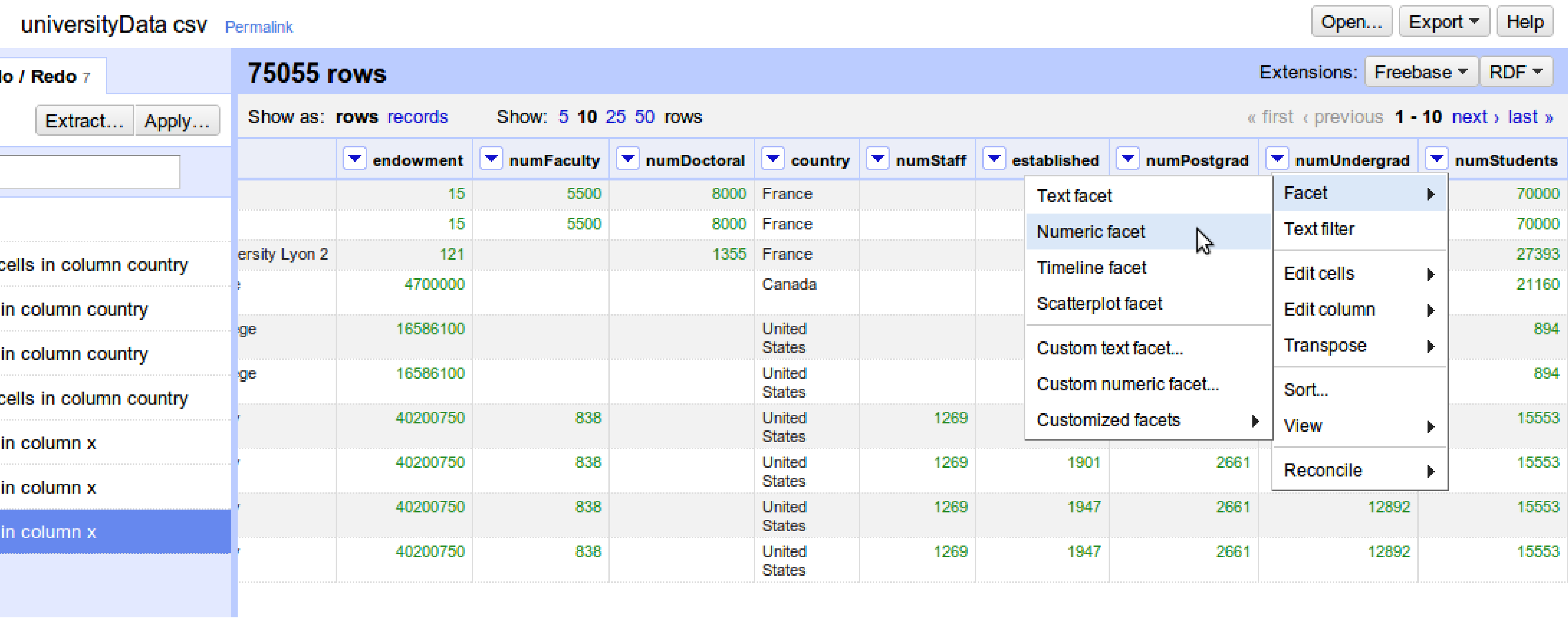
Method: key collision

Keying Function: fingerprint

3 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	6603	- U.S. (3994 rows) - US (2609 rows)	☑	United States
2	32034	- United States (32033 rows) - United States) (1 rows)	☑	United States
2	6795	- USA (6402 rows) - U.S.A. (393 rows)	☑	United States

# OF STUDENTS



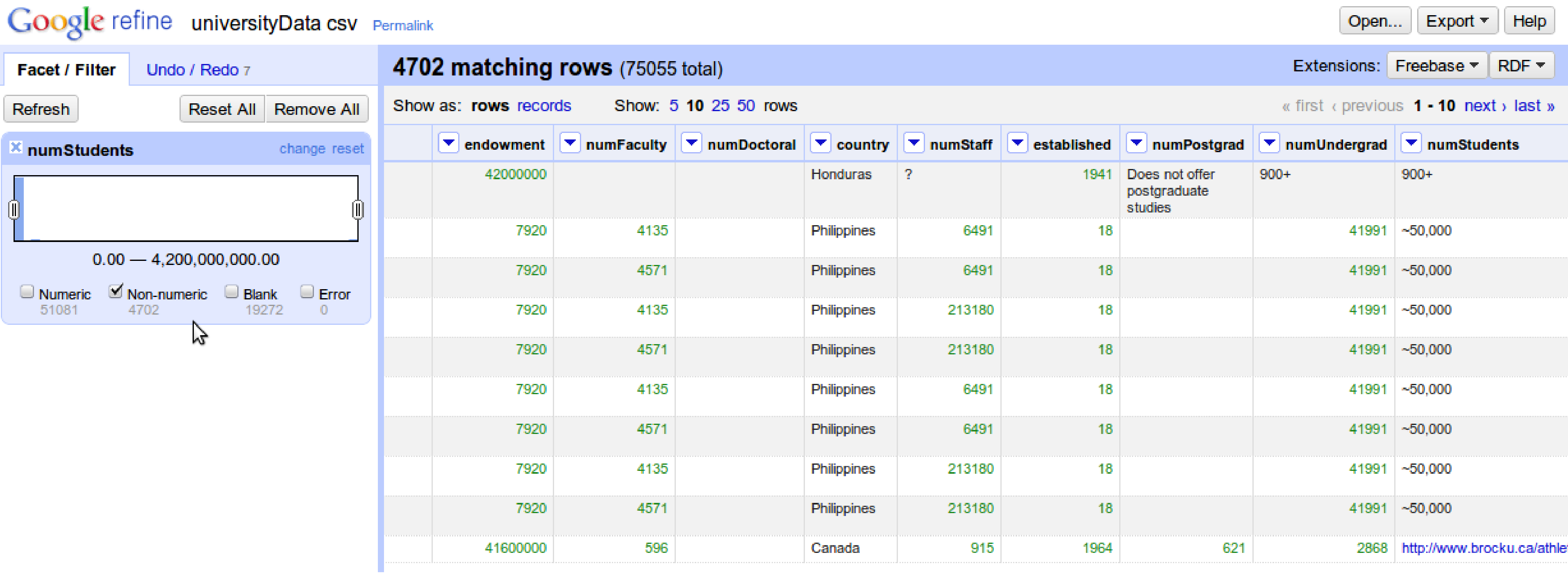
75055 rows

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
ersity Lyon 2	121		1355	France					27393
	4700000			Canada					21160
ge	16586100			United States					894
ge	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

What do you notice?

# OF STUDENTS



Google refine universityData csv Permalink

Open... Export ▾ Help

Facet / Filter Undo / Redo 7

Refresh Reset All Remove All

numStudents change reset

0.00 — 4,200,000,000.00

☐ Numeric 51081 ☑ Non-numeric 4702 ☐ Blank 19272 ☐ Error 0

4702 matching rows (75055 total)

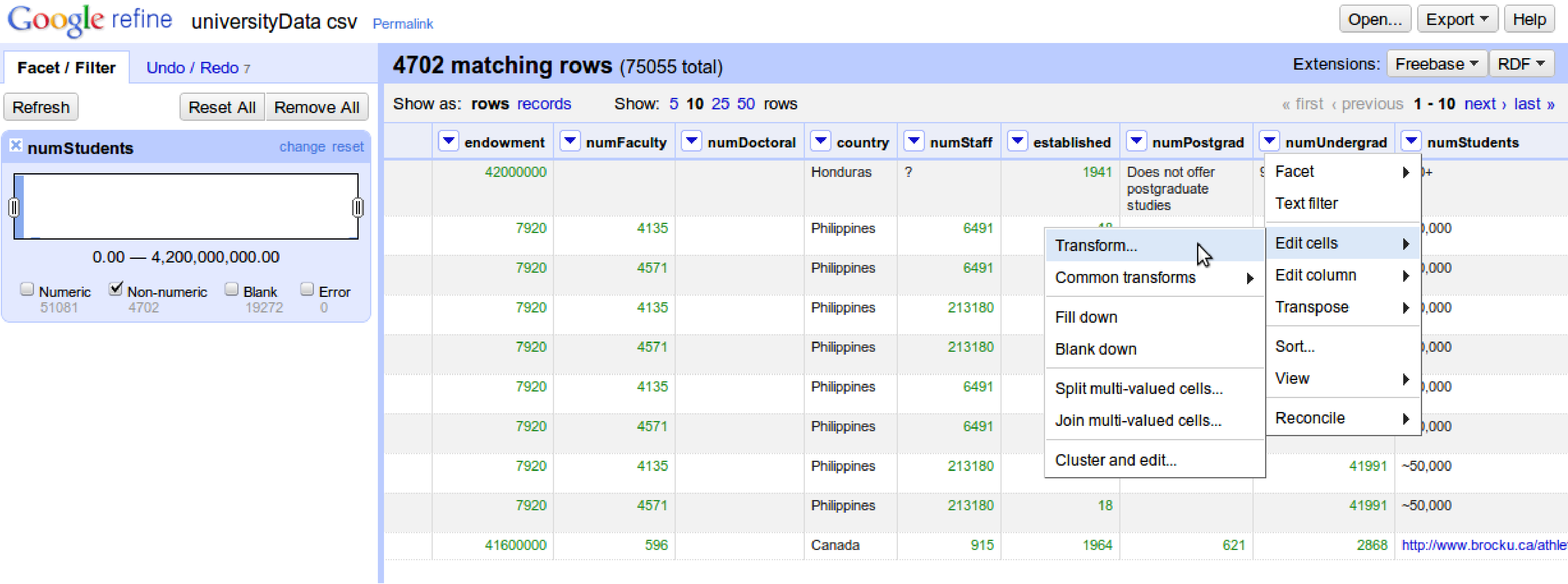
Extensions: Freebase ▾ RDF ▾

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
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



universityData csv

4702 matching rows (75055 total)

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

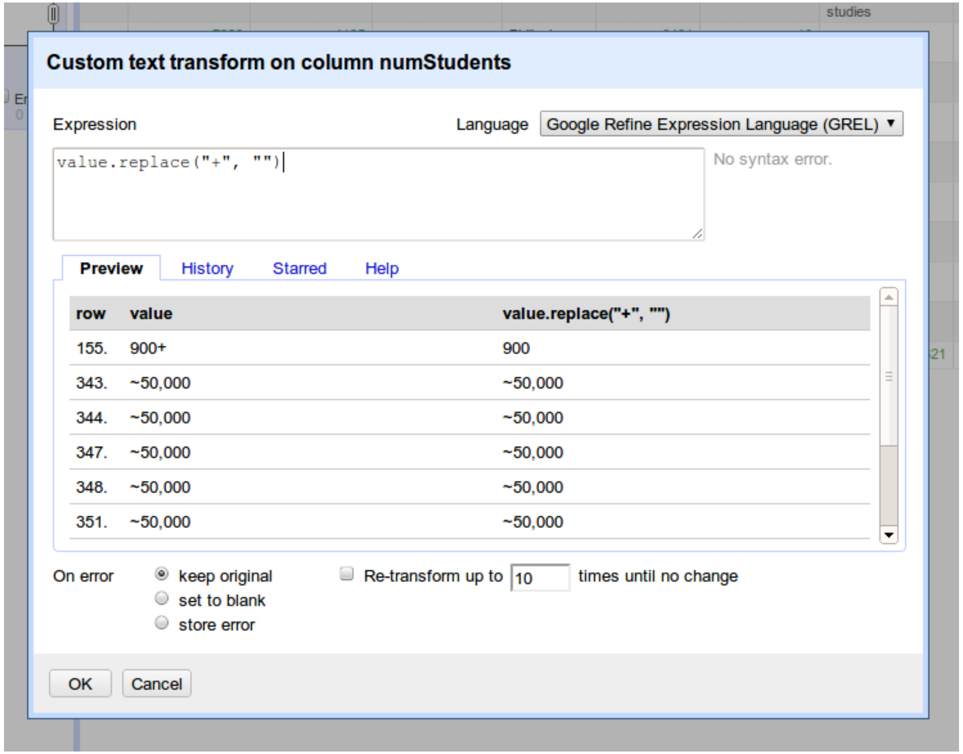
numStudents

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	9	0+
7920	4135		Philippines	6491	18			,000
7920	4571		Philippines	6491				,000
7920	4135		Philippines	213180				,000
7920	4571		Philippines	213180				,000
7920	4135		Philippines	6491				,000
7920	4571		Philippines	6491				,000
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

# OF STUDENTS



Custom text transform on column numStudents

Expression

Language: Google Refine Expression Language (GREL)

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

No syntax error.

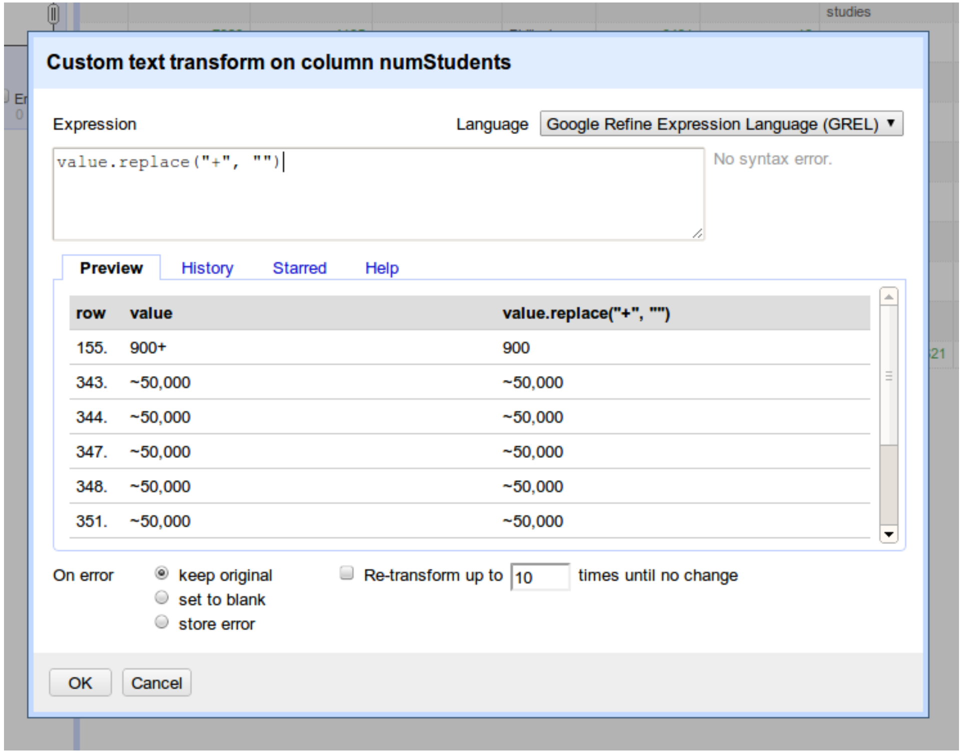
row	value	value.replace("+", "")
155.	900+	900
343.	~50,000	~50,000
344.	~50,000	~50,000
347.	~50,000	~50,000
348.	~50,000	~50,000
351.	~50,000	~50,000

On error ● keep original ○ set to blank ○ store error

□ Re-transform up to 10 times until no change

value.replace("+", "")

# OF STUDENTS



Custom text transform on column numStudents

Expression

Language: Google Refine Expression Language (GREL)

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

No syntax error.

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

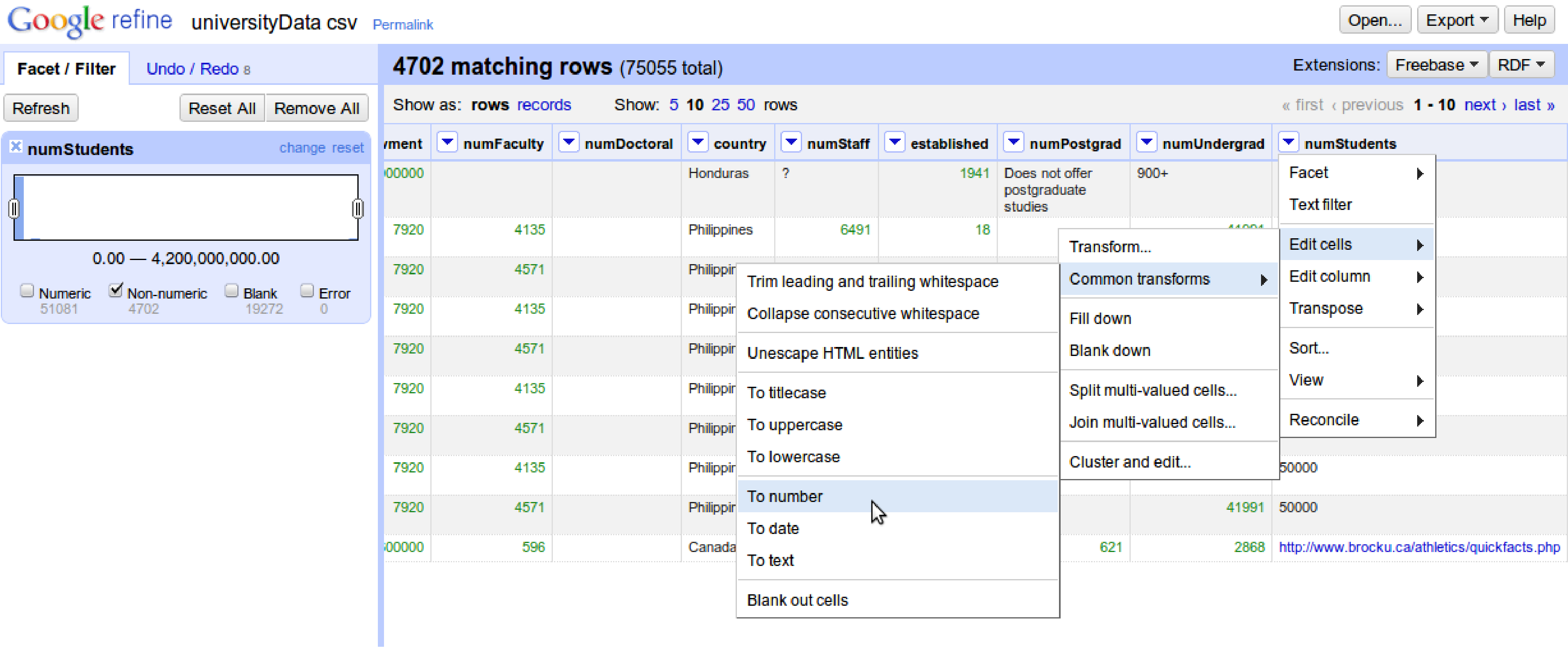
On error: ● keep original ○ set to blank ○ store error

□ Re-transform up to 10 times until no change

"Lumi%C3%A8re University Lyon 2"

value.unescape('url')

# OF STUDENTS



4702 matching rows (75055 total)

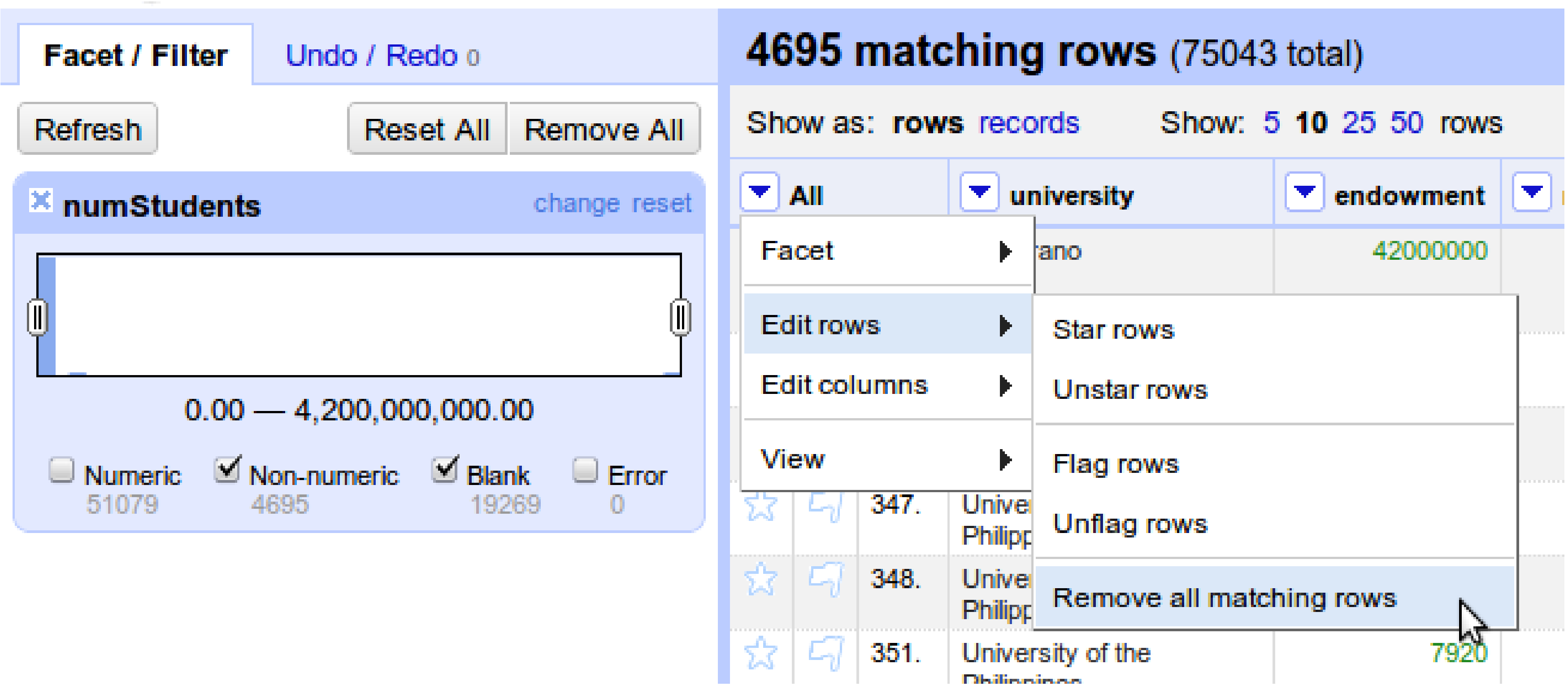
numStudents

0.00 — 4,200,000,000.00

☐ Numeric 51081 ☑ Non-numeric 4702 ☐ Blank 19272 ☐ Error 0

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

REMOVING UNWANTED ROWS

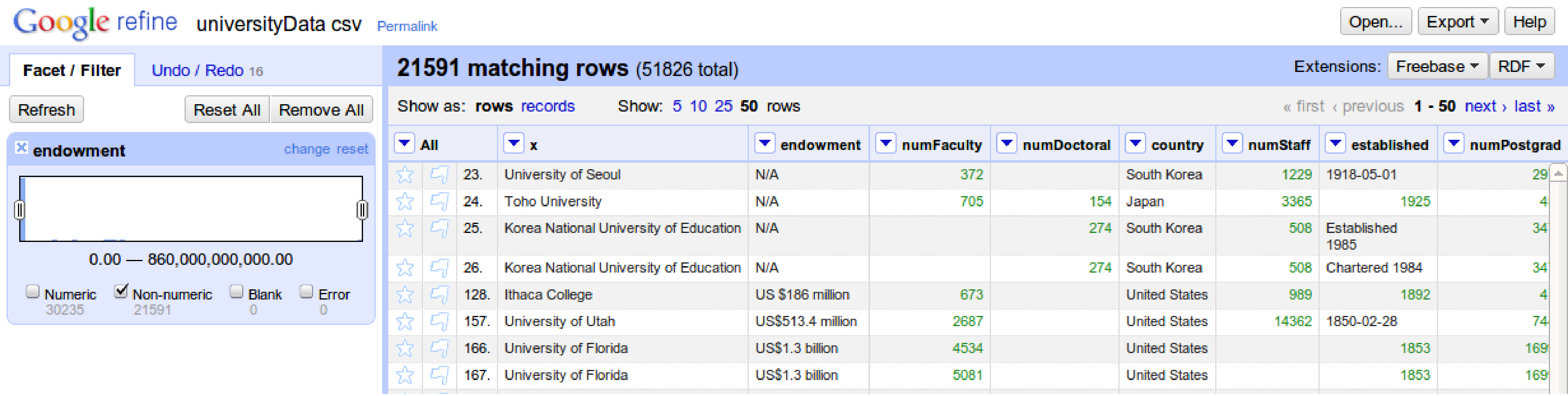


4695 matching rows (75043 total)

numStudents: 0.00 — 4,200,000,000.00

- ☐ Numeric 51079
- ☑ Non-numeric 4695
- ☑ Blank 19269
- ☐ Error 0

ENDOWMENT



Google refine universityData csv Permalink

Open... Export Help

Facet / Filter Undo / Redo 16

Refresh Reset All Remove All

endowment change reset

0.00 — 860,000,000,000.00

□ Numeric 30235 ☑ Non-numeric 21591 □ Blank 0 □ Error 0

21591 matching rows (51826 total)

Extensions: Freebase RDF

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

« first ‹ previous **1 - 50** next › last »

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?

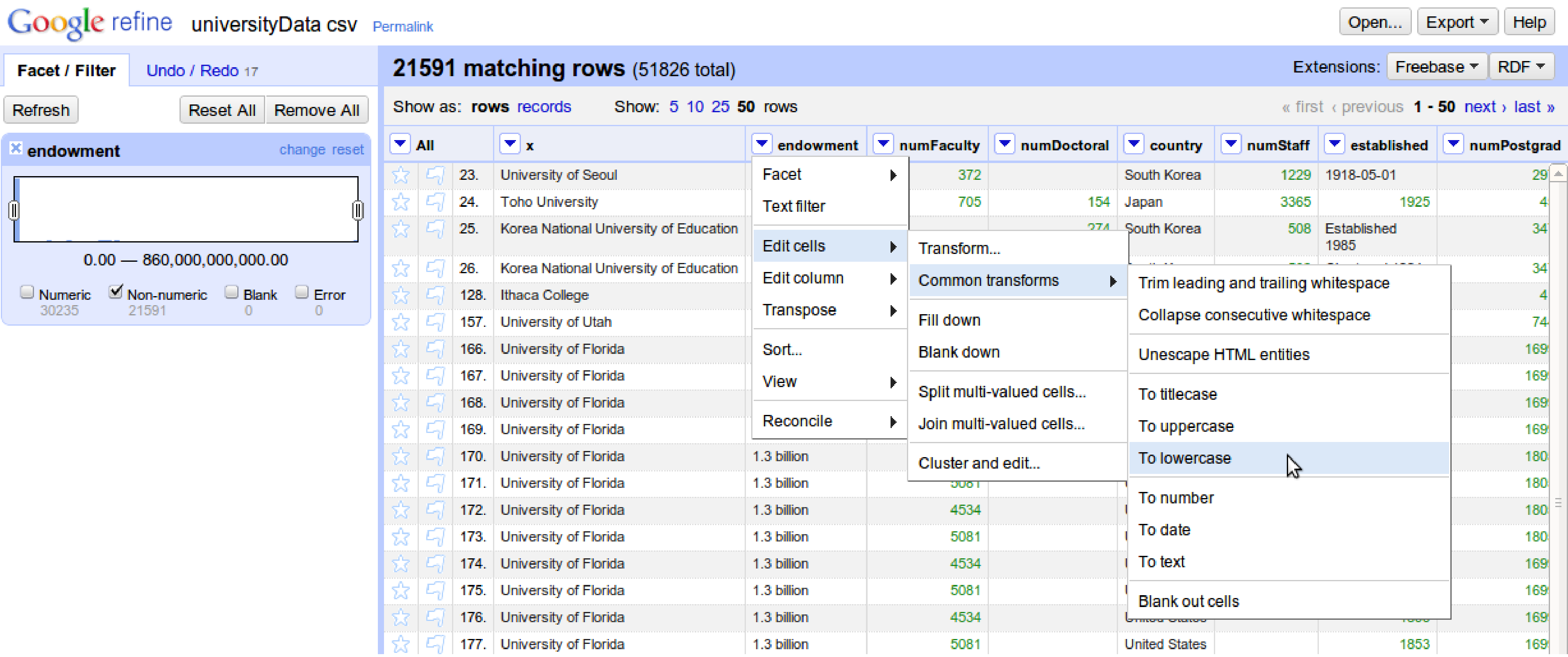
ENDOWMENT

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

-> **Edit cells -> Transform**

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

CONVERT TO LC



0.00 — 860,000,000,000.00

☐ Numeric 30235 ☑ Non-numeric 21591 ☐ Blank 0 ☐ Error 0

21591 matching rows (51826 total)

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

CONVERT TO NUMBERS

$13.8 million

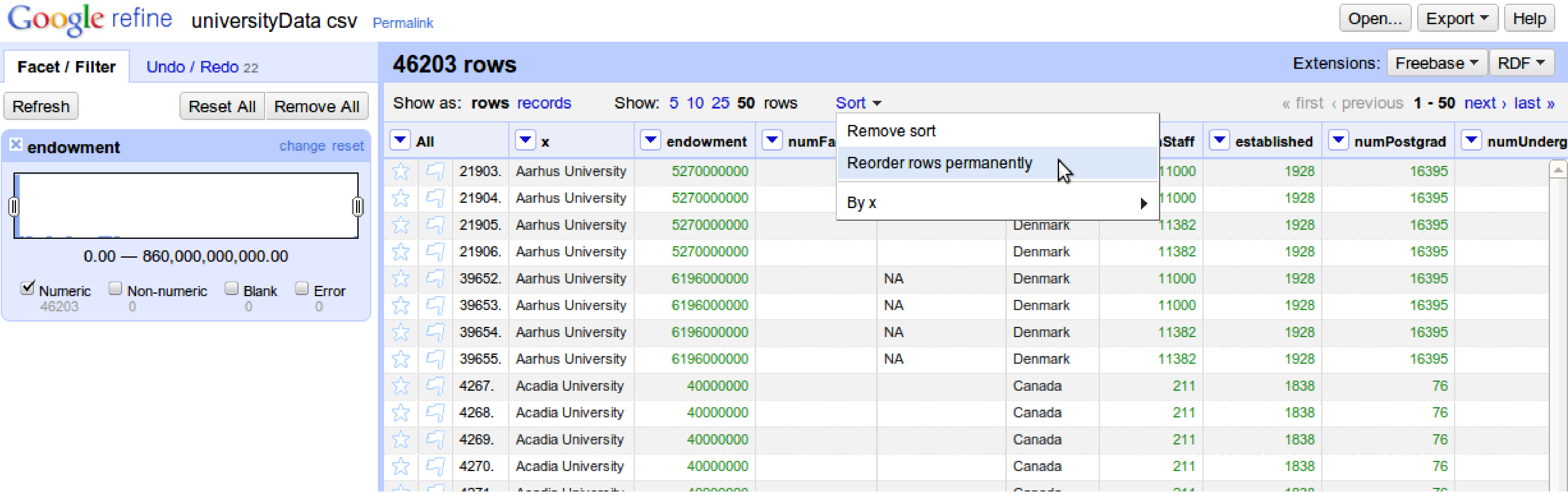
What could we do here?

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

DEDUPLICATION

Dataset has a lot of duplicate rows

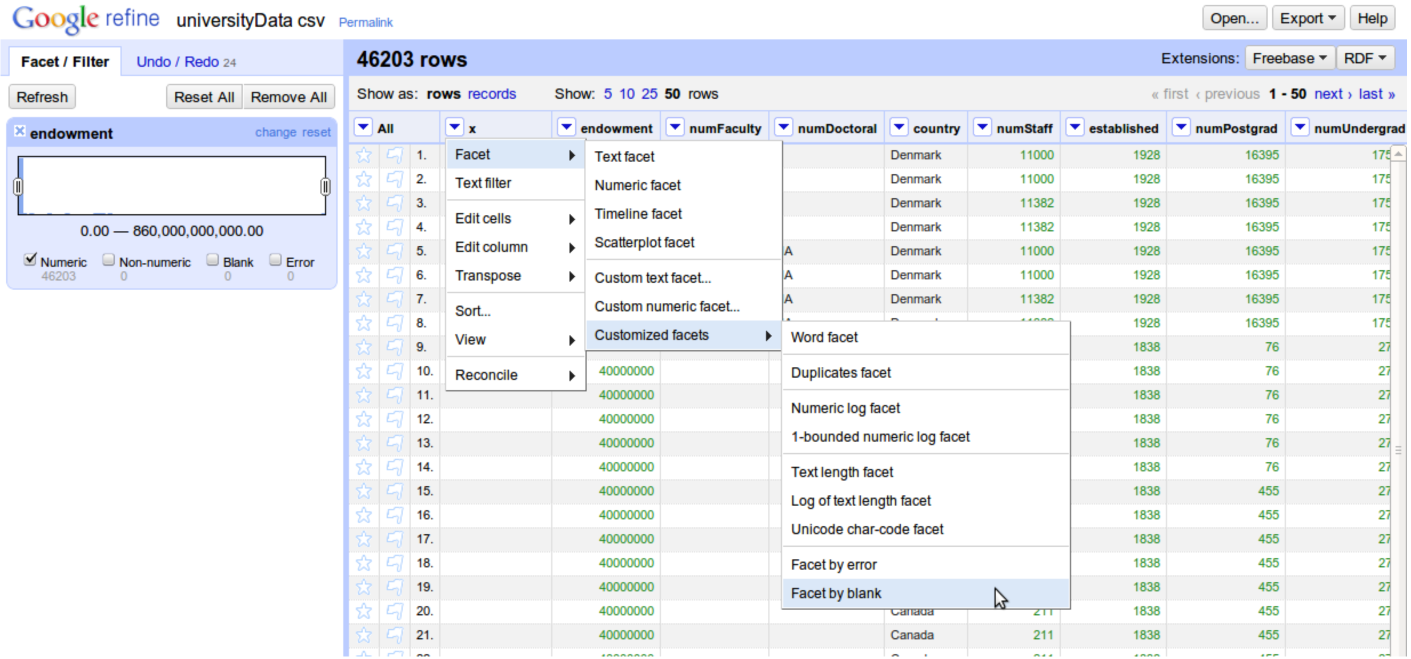
-> university names -> sort -> (image below)



All	x	endowment	numFa			Staff	established	numPostgrad	numUnderg
21903.	Aarhus University	5270000000				11000	1928	16395	
21904.	Aarhus University	5270000000				11000	1928	16395	
21905.	Aarhus University	5270000000			Denmark	11382	1928	16395	
21906.	Aarhus University	5270000000			Denmark	11382	1928	16395	
39652.	Aarhus University	6196000000		NA	Denmark	11000	1928	16395	
39653.	Aarhus University	6196000000		NA	Denmark	11000	1928	16395	
39654.	Aarhus University	6196000000		NA	Denmark	11382	1928	16395	
39655.	Aarhus University	6196000000		NA	Denmark	11382	1928	16395	
4267.	Acadia University	40000000			Canada	211	1838	76	
4268.	Acadia University	40000000			Canada	211	1838	76	
4269.	Acadia University	40000000			Canada	211	1838	76	
4270.	Acadia University	40000000			Canada	211	1838	76	
4271.	Acadia University	40000000			Canada	211	1838	76	

DEDUPLICATION

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



All	x	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad
1.					Denmark	11000	1928	16395	175
2.					Denmark	11000	1928	16395	175
3.					Denmark	11382	1928	16395	175
4.					Denmark	11382	1928	16395	175
5.				A	Denmark	11000	1928	16395	175
6.				A	Denmark	11000	1928	16395	175
7.				A	Denmark	11382	1928	16395	175
8.							1928	16395	175
9.							1838	76	27
10.		40000000					1838	76	27
11.		40000000					1838	76	27
12.		40000000					1838	76	27
13.		40000000					1838	76	27
14.		40000000					1838	76	27
15.		40000000					1838	455	27
16.		40000000					1838	455	27
17.		40000000					1838	455	27
18.		40000000					1838	455	27
19.		40000000					1838	455	27
20.		40000000			Canada	211	1838	455	27
21.		40000000			Canada	211	1838	455	27

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