Ammanamanchi Sai Karthik

B150310CS

Format of Submission

The initial data set used is included as a .csv file and the final modified data set is availiable in .xls format.

Raw Dataset: universityData.csv

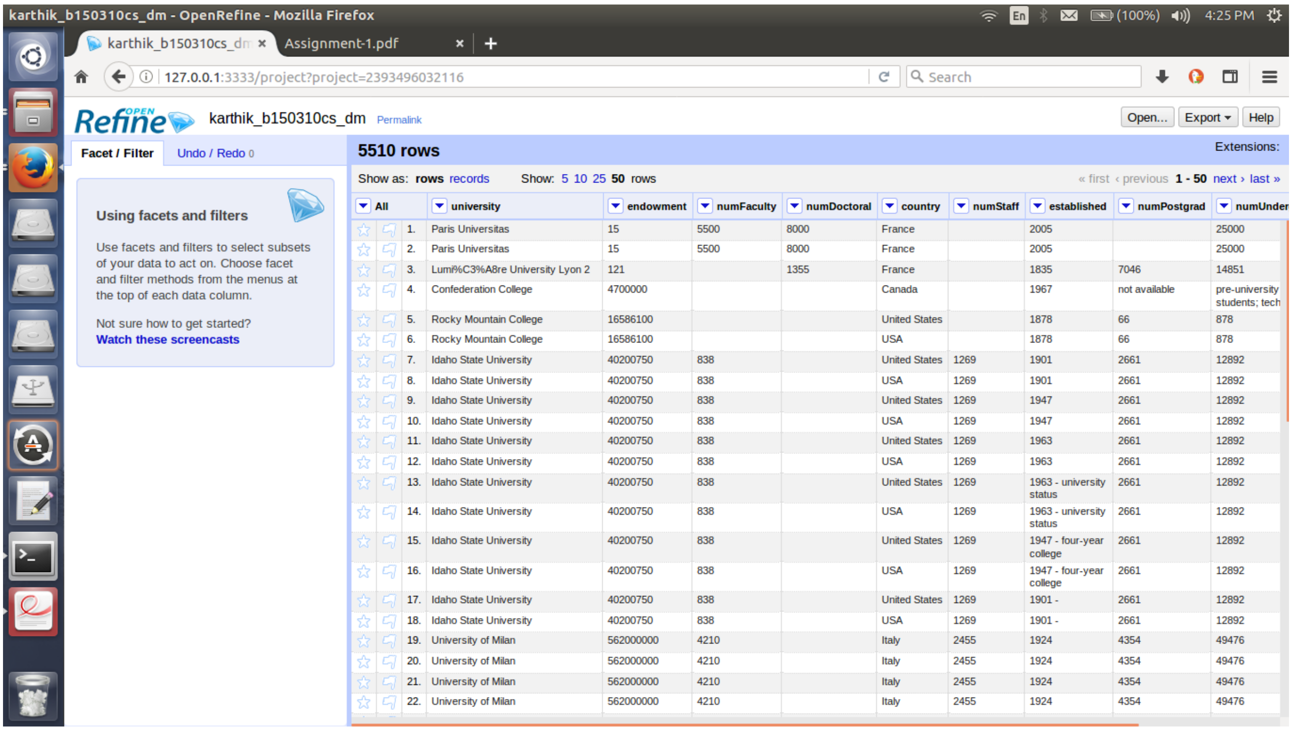
Link: https://drive.google.com/open?id=0B5bh5JNC6R13SWlRS2puMWZHZFk

Pre-Processed Dataset: karthik_b150310cs_dm.xls

Link: https://drive.google.com/open?id=0B5bh5JNC6R13NGw4WFlfMldjQ00

DataSet

The dataset consists of University data with over 5000 rows and 10 attributes.



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

5510 rows

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

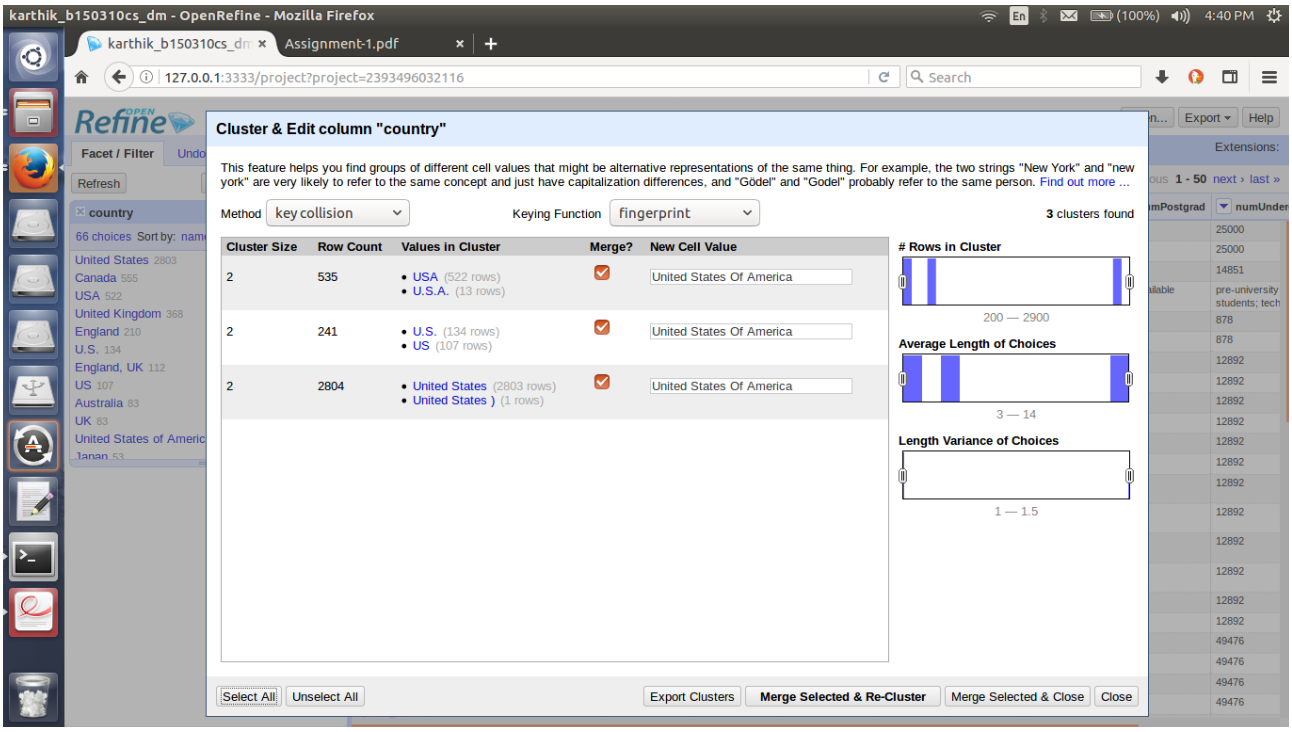
All	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUnder
1.	Paris Universitas	15	5500	8000	France		2005		25000
2.	Paris Universitas	15	5500	8000	France		2005		25000
3.	Lumi%C3%A8re University Lyon 2	121		1355	France		1835	7046	14851
4.	Confederation College	4700000			Canada		1967	not available	pre-university students; tech
5.	Rocky Mountain College	16586100			United States		1878	66	878
6.	Rocky Mountain College	16586100			USA		1878	66	878
7.	Idaho State University	40200750	838		United States	1269	1901	2661	12892
8.	Idaho State University	40200750	838		USA	1269	1901	2661	12892
9.	Idaho State University	40200750	838		United States	1269	1947	2661	12892
10.	Idaho State University	40200750	838		USA	1269	1947	2661	12892
11.	Idaho State University	40200750	838		United States	1269	1963	2661	12892
12.	Idaho State University	40200750	838		USA	1269	1963	2661	12892
13.	Idaho State University	40200750	838		United States	1269	1963 - university status	2661	12892
14.	Idaho State University	40200750	838		USA	1269	1963 - university status	2661	12892
15.	Idaho State University	40200750	838		United States	1269	1947 - four-year college	2661	12892
16.	Idaho State University	40200750	838		USA	1269	1947 - four-year college	2661	12892
17.	Idaho State University	40200750	838		United States	1269	1901 -	2661	12892
18.	Idaho State University	40200750	838		USA	1269	1901 -	2661	12892
19.	University of Milan	562000000	4210		Italy	2455	1924	4354	49476
20.	University of Milan	562000000	4210		Italy	2455	1924	4354	49476
21.	University of Milan	562000000	4210		Italy	2455	1924	4354	49476
22.	University of Milan	562000000	4210		Italy	2455	1924	4354	49476

Attributes

University(Nominal), Endowment(Numeric), numFaculty(Numeric), numDoctoral(Numeric), Country(Ordinal),numStaff(Numeric),Established(Ordinal),numPostgrad(Numeric), numUndergrad(Numeric).

Clean up country names

To sort out the issue of country names like USA, U.S.A, U.S we can use **Edit cells->Cluster and edit** on the country column.



karthik_b150310cs_dm - OpenRefine - Mozilla Firefox
4:40 PM
karthik_b150310cs_dm
Assignment-1.pdf
127.0.0.1:3333/project?project=2393496032116
Search
Facet / Filter
Undo
Refresh
Export
Help
Extensions:

country

66 choices Sort by: name

United States	2803
Canada	555
USA	522
United Kingdom	368
England	210
U.S.	134
England, UK	112
US	107
Australia	83
UK	83
United States of Americ	
Japan	53

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

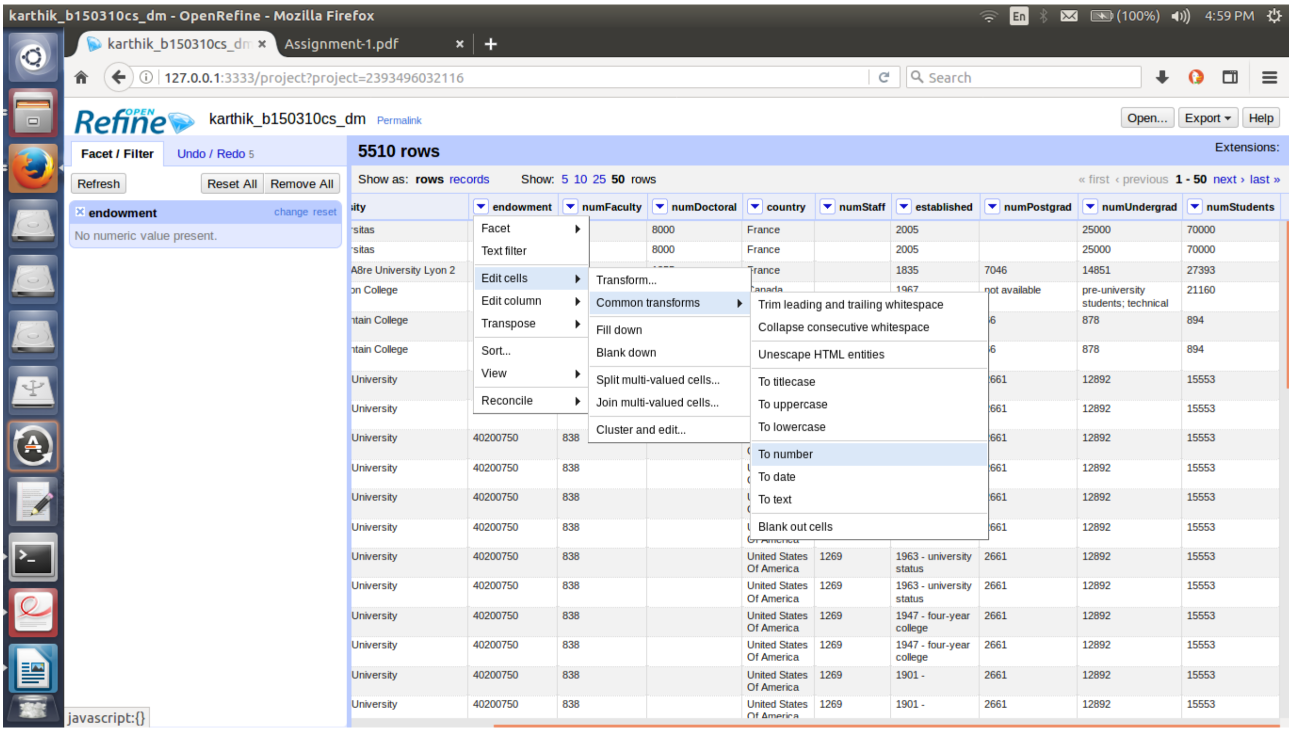
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	535	• USA (522 rows) • U.S.A. (13 rows)	☑	United States Of America
2	241	• U.S. (134 rows) • US (107 rows)	☑	United States Of America
2	2804	• United States (2803 rows) • United States) (1 rows)	☑	United States Of America

We use an inbuilt clustering algorithm with key function fingerprint and rename the country name to United States Of America.

Converting Values such as $123 million to 123000000

We will focus on the attribute endowments.

Firstly, we will have to convert the strings to numbers using the common transformation tonumber.



endowment change reset

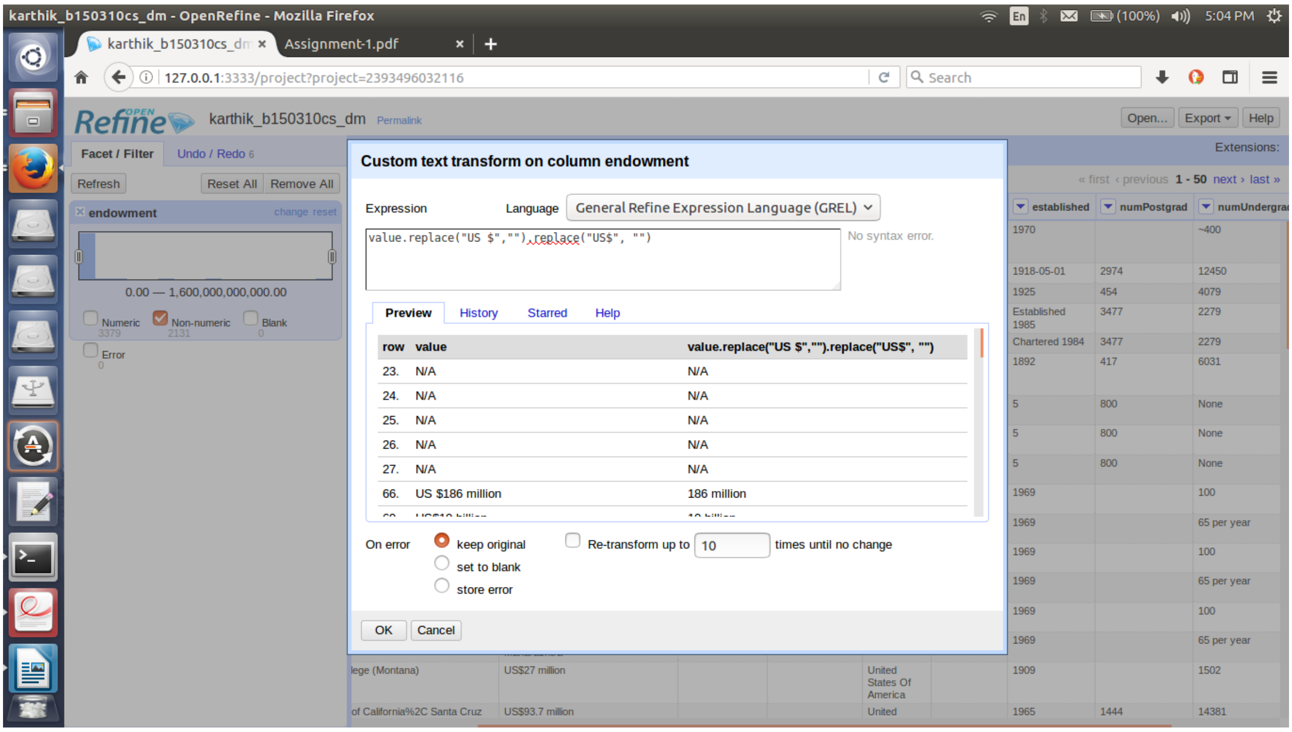
No numeric value present.

5510 rows

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

ity	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
sitas			8000	France		2005		25000	70000
sitas			8000	France		2005		25000	70000
A8re University Lyon 2				rance		1835	7046	14851	27393
on College				anada		1967	not available	pre-university students; technical	21160
ntain College							6	878	894
ntain College							6	878	894
University							661	12892	15553
University							661	12892	15553
University	40200750	838					661	12892	15553
University	40200750	838					661	12892	15553
University	40200750	838					661	12892	15553
University	40200750	838					661	12892	15553
University	40200750	838		United States Of America	1269	1963 - university status	2661	12892	15553
University	40200750	838		United States Of America	1269	1963 - university status	2661	12892	15553
University	40200750	838		United States Of America	1269	1947 - four-year college	2661	12892	15553
University	40200750	838		United States Of America	1269	1947 - four-year college	2661	12892	15553
University	40200750	838		United States Of America	1269	1901 -	2661	12892	15553
University	40200750	838		United States Of America	1269	1901 -	2661	12892	15553

We now change the US$ and US $ to “”(empty string) using value.replace("US $","").replace("US$", "") in tthe transform section of endowments attribute.



Custom text transform on column endowment

Expression Language: General Refine Expression Language (GREL)

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

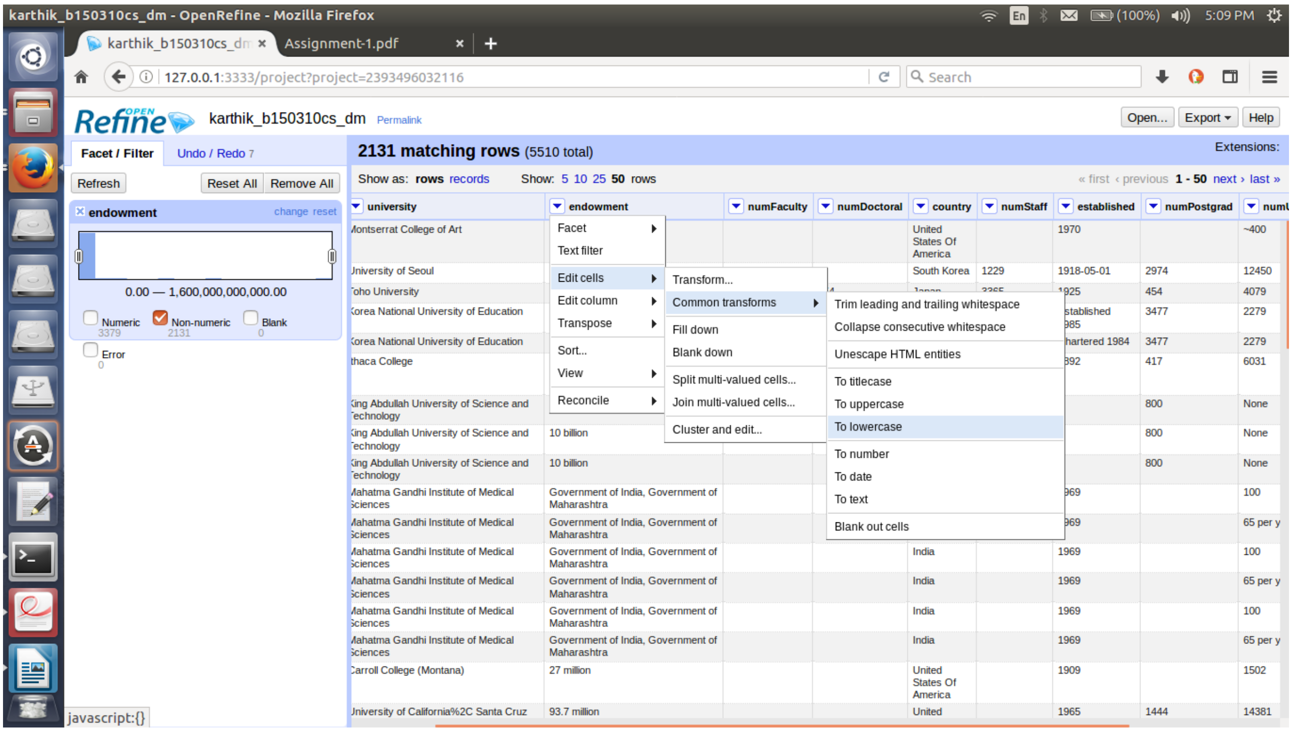
No syntax error.

row	value	value.replace("US $","").replace("US$", "")
23.	N/A	N/A
24.	N/A	N/A
25.	N/A	N/A
26.	N/A	N/A
27.	N/A	N/A
66.	US $186 million	186 million

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

□ Re-transform up to 10 times until no change

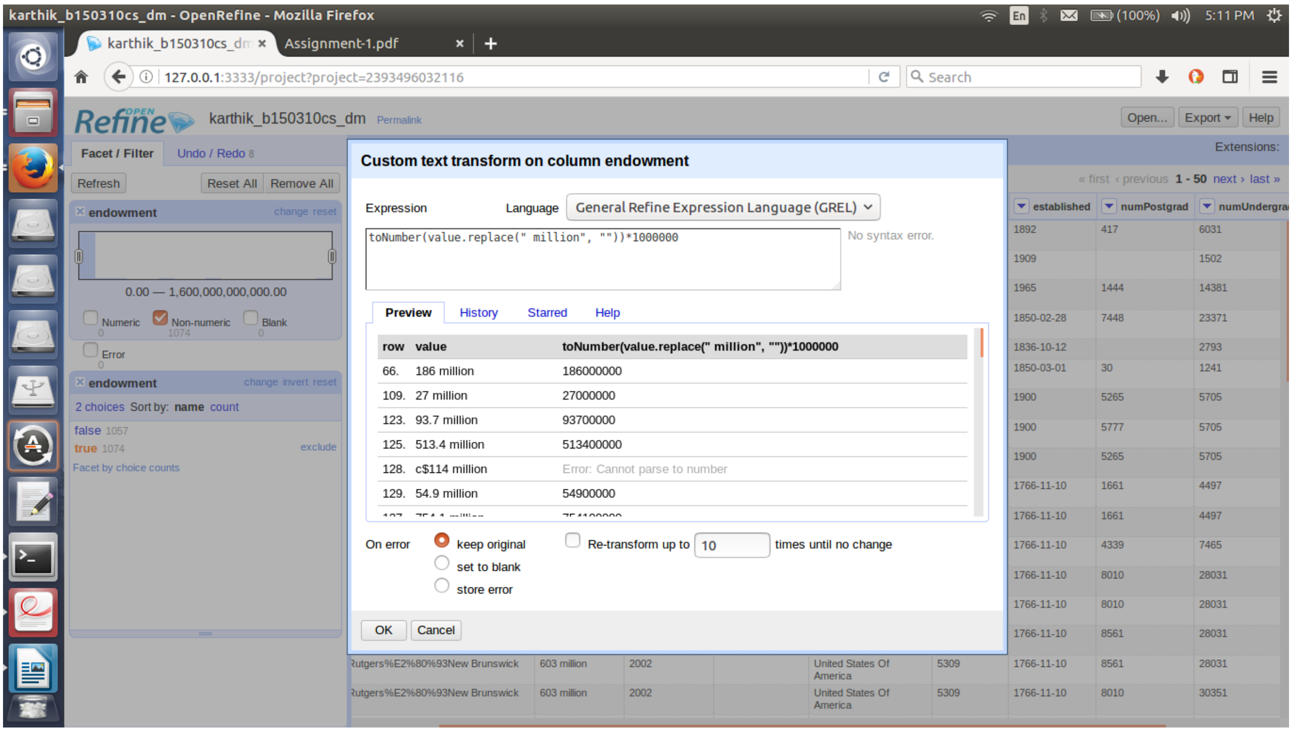
Now we convert the “millions” and “Millions” to the lowercase “millions”.



2131 matching rows (5510 total)

Create a custom text facet to select the rows containing “million”.

```
value.contains("million")
```



karthik_b150310cs_dm - OpenRefine - Mozilla Firefox
127.0.0.1:3333/project?project=2393496032116
Refine karthik_b150310cs_dm Permalink
Open... Export Help
Facet / Filter Undo / Redo 8
Refresh Reset All Remove All
endowment change reset
0.00 — 1,600,000,000,000.00
Numeric Non-numeric Blank Error
endowment change invert reset
2 choices Sort by: name count
false 1057
true 1074 exclude
Facet by choice counts
established numPostgrad numUndergra

Custom text transform on column endowment

Expression Language General Refine Expression Language (GREL)

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

No syntax error.

Preview History Starred Help

row	value	toNumber(value.replace(" million", ""))*1000000
66.	186 million	186000000
109.	27 million	27000000
123.	93.7 million	93700000
125.	513.4 million	513400000
128.	c$114 million	Error: Cannot parse to number
129.	54.9 million	54900000

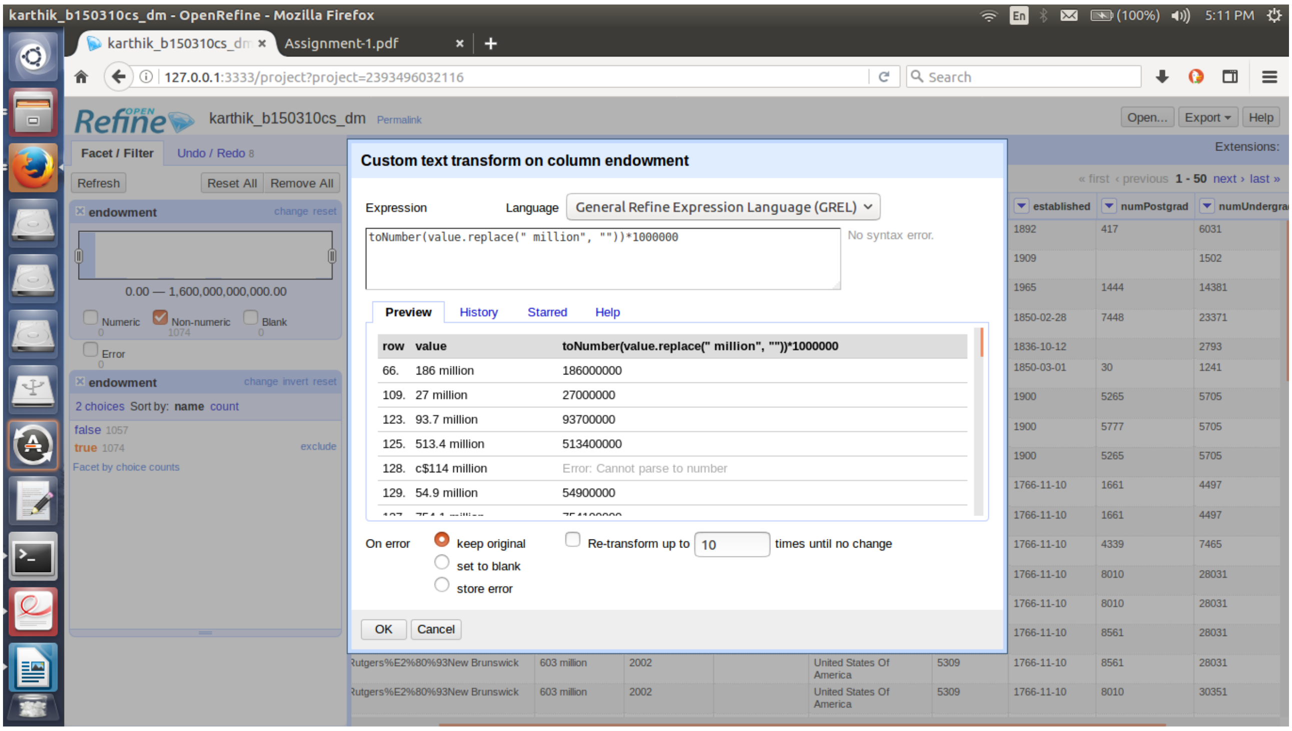
On error ● keep original ○ set to blank ○ store error

□ Re-transform up to 10 times until no change

OK Cancel

We now replace the million with a “”(empty string), convert the resulting string to number and multiply by 1000000.

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



Custom text transform on column endowment

Expression Language General Refine Expression Language (GREL)

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

No syntax error.

Preview History Starred Help

row	value	toNumber(value.replace(" million", ""))*1000000
66.	186 million	186000000
109.	27 million	27000000
123.	93.7 million	93700000
125.	513.4 million	513400000
128.	c$114 million	Error: Cannot parse to number
129.	54.9 million	54900000

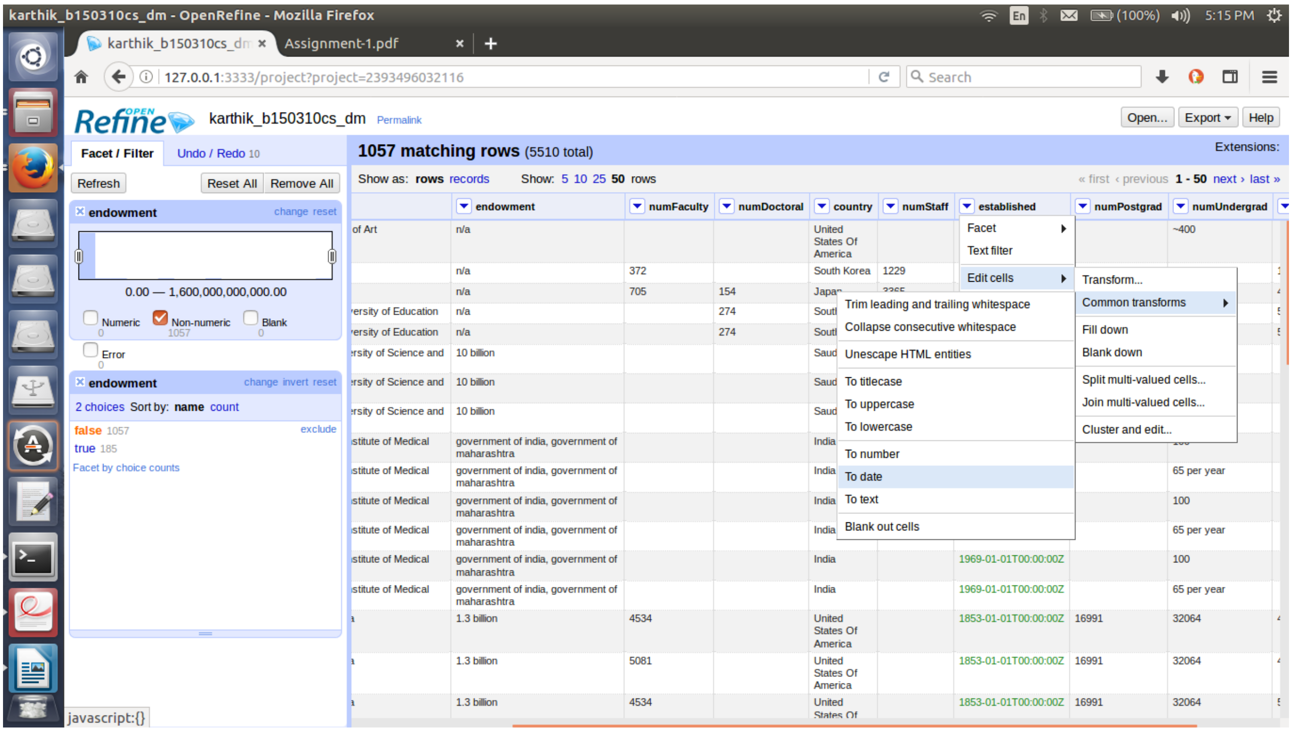
On error ● keep original ○ set to blank ○ store error

□ Re-transform up to 10 times until no change

OK Cancel

Cleaning up dates

First we will use a common transformation to Date on established column.

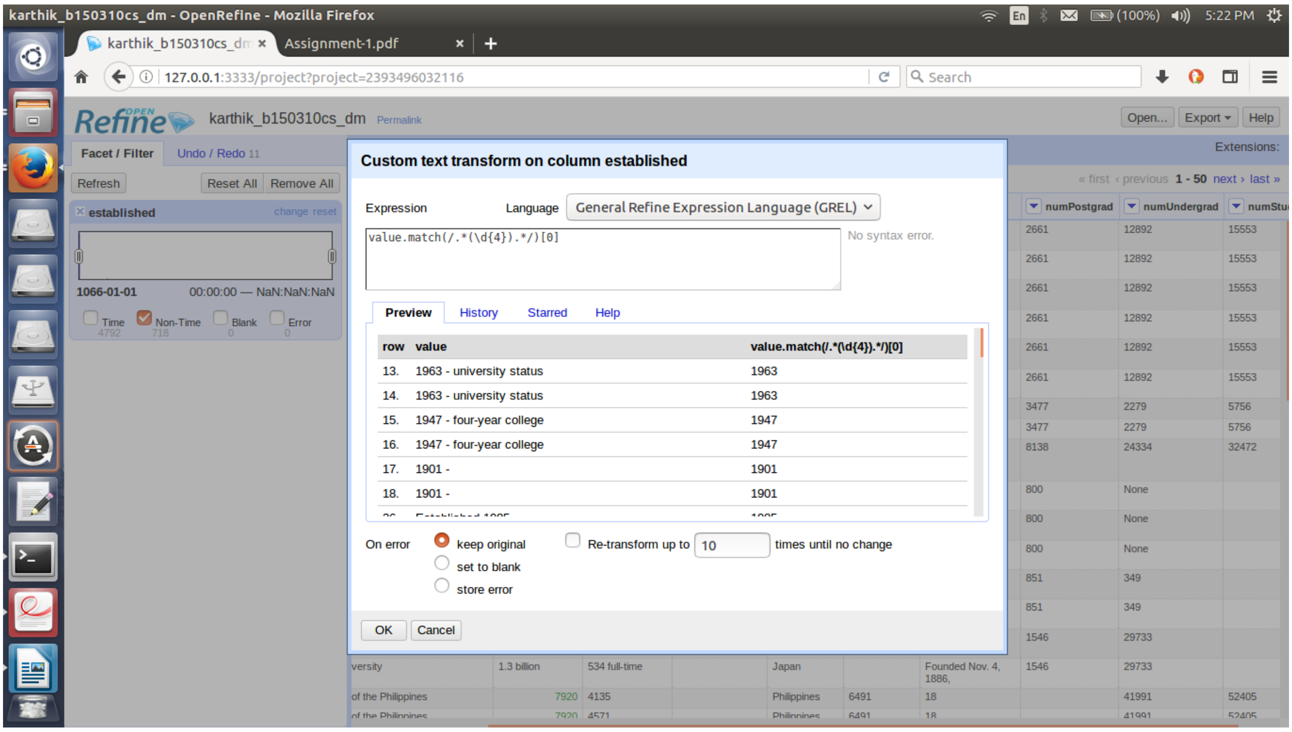


Now we select the non date tuples in the column by using the timeline facet and selecting the non-time dialog box.

Facet -> Timeline facet

Now we use a regular expression to select the first four characters in the string and adain use the to Date common transformation to convert them dates.

```
value.match(/.*(\d{4}).*/)[0]
```



Custom text transform on column established

Expression Language General Refine Expression Language (GREL)

```
value.match(/.*(\d{4}).*/)[0]
```

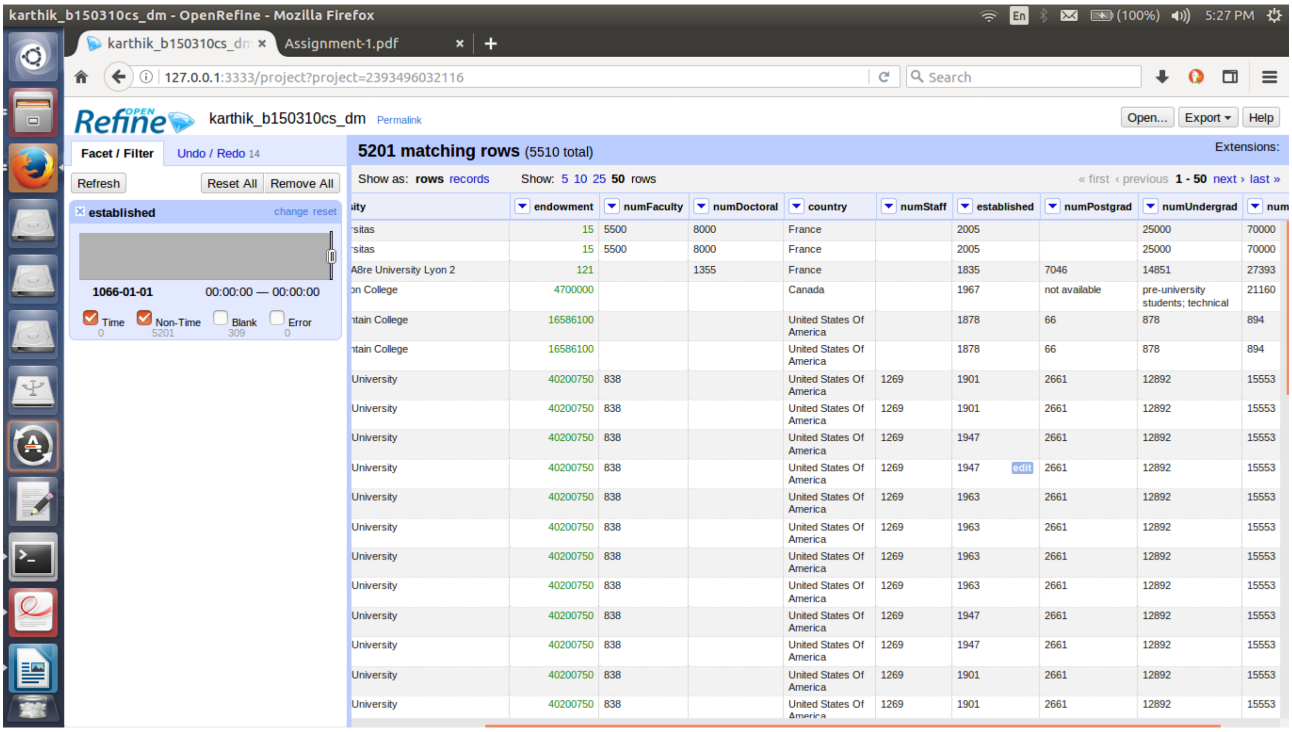
No syntax error.

row	value	value.match(/.*(\d{4}).*/)[0]
13.	1963 - university status	1963
14.	1963 - university status	1963
15.	1947 - four-year college	1947
16.	1947 - four-year college	1947
17.	1901 -	1901
18.	1901 -	1901

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

□ Re-transform up to 10 times until no change

Use transform on the column and apply value.toString('yyyy').



1066-01-01 00:00:00 — 00:00:00

☑ Time 0 ☑ Non-Time 5201 □ Blank 309 □ Error 0

5201 matching rows (5510 total)

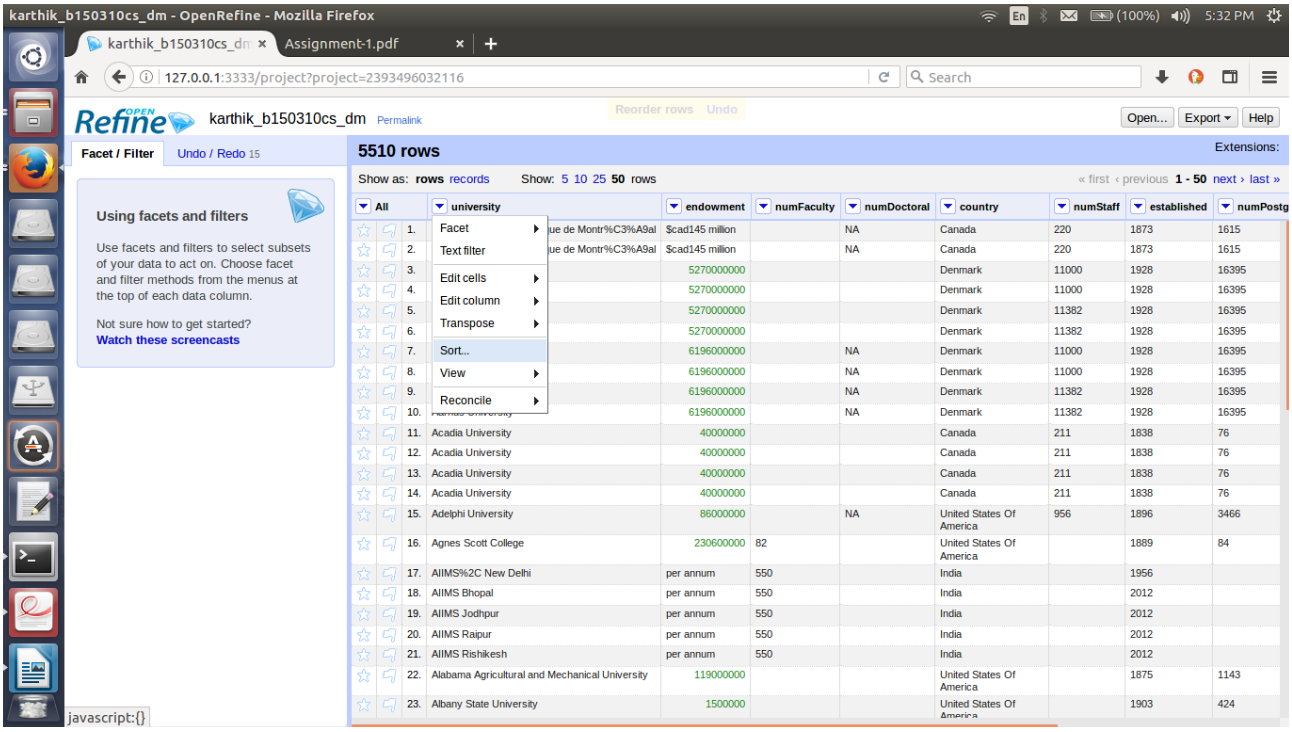
Show as: **rows** records Show: 5 10 25 **50** rows « first ‹ previous **1 - 50** next › last »

ity	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	num
sitas	15	5500	8000	France		2005		25000	70000
sitas	15	5500	8000	France		2005		25000	70000
A8re University Lyon 2	121		1355	France		1835	7046	14851	27393
on College	4700000			Canada		1967	not available	pre-university students; technical	21160
ntain College	16586100			United States Of America		1878	66	878	894
ntain College	16586100			United States Of America		1878	66	878	894
University	40200750	838		United States Of America	1269	1901	2661	12892	15553
University	40200750	838		United States Of America	1269	1901	2661	12892	15553
University	40200750	838		United States Of America	1269	1947	2661	12892	15553
University	40200750	838		United States Of America	1269	1947 edit	2661	12892	15553
University	40200750	838		United States Of America	1269	1963	2661	12892	15553
University	40200750	838		United States Of America	1269	1963	2661	12892	15553
University	40200750	838		United States Of America	1269	1963	2661	12892	15553
University	40200750	838		United States Of America	1269	1963	2661	12892	15553
University	40200750	838		United States Of America	1269	1947	2661	12892	15553
University	40200750	838		United States Of America	1269	1947	2661	12892	15553
University	40200750	838		United States Of America	1269	1901	2661	12892	15553
University	40200750	838		United States Of America	1269	1901	2661	12892	15553

Deduplicate Entries

We will try to deduplicate entries by tackling the university names column.

We first sort the rows and select reorder rows permanently.



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.

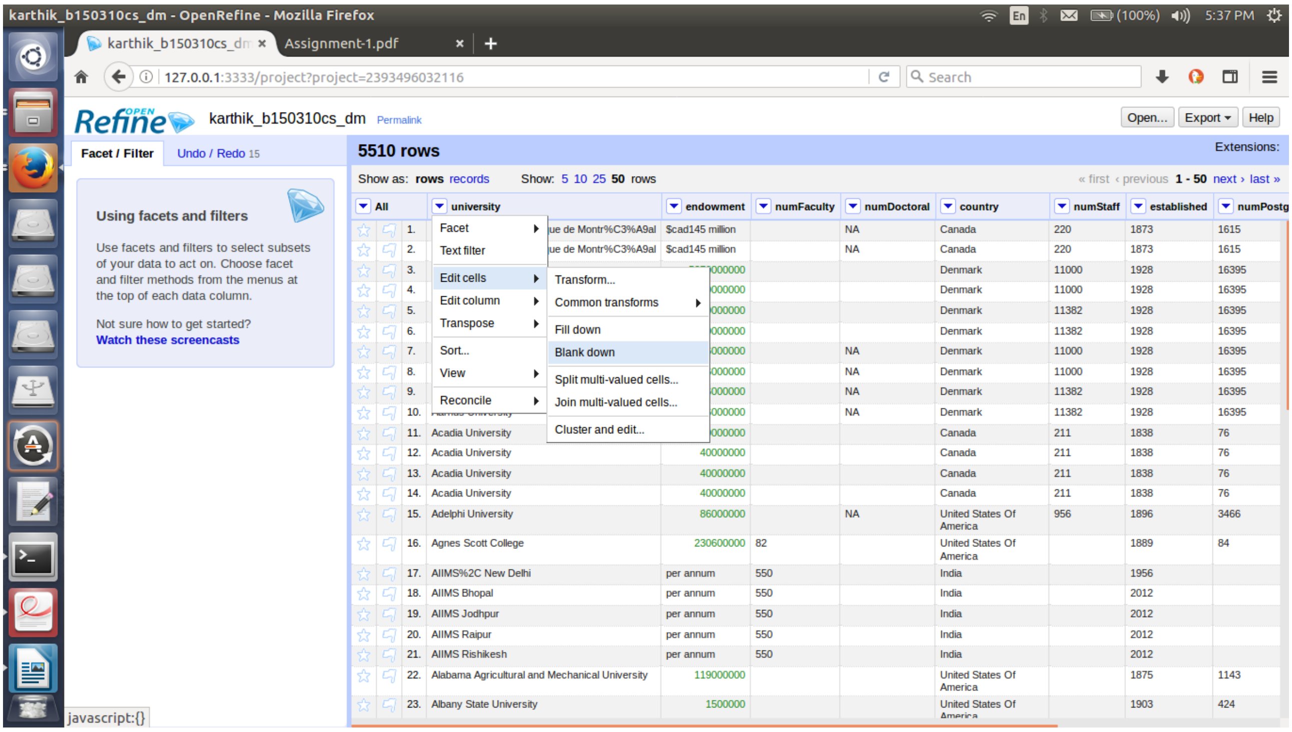
Not sure how to get started?
Watch these screencasts

5510 rows

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

All	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostg
1.	ue de Montr%C3%A9al	$cad145 million		NA	Canada	220	1873	1615
2.	ue de Montr%C3%A9al	$cad145 million		NA	Canada	220	1873	1615
3.		5270000000			Denmark	11000	1928	16395
4.		5270000000			Denmark	11000	1928	16395
5.		5270000000			Denmark	11382	1928	16395
6.		5270000000			Denmark	11382	1928	16395
7.		6196000000		NA	Denmark	11000	1928	16395
8.		6196000000		NA	Denmark	11000	1928	16395
9.		6196000000		NA	Denmark	11382	1928	16395
10.		6196000000		NA	Denmark	11382	1928	16395
11.	Acadia University	40000000			Canada	211	1838	76
12.	Acadia University	40000000			Canada	211	1838	76
13.	Acadia University	40000000			Canada	211	1838	76
14.	Acadia University	40000000			Canada	211	1838	76
15.	Adelphi University	86000000		NA	United States Of America	956	1896	3466
16.	Agnes Scott College	230600000	82		United States Of America		1889	84
17.	AIIMS%2C New Delhi	per annum	550		India		1956	
18.	AIIMS Bhopal	per annum	550		India		2012	
19.	AIIMS Jodhpur	per annum	550		India		2012	
20.	AIIMS Raipur	per annum	550		India		2012	
21.	AIIMS Rishikesh	per annum	550		India		2012	
22.	Alabama Agricultural and Mechanical University	119000000			United States Of America		1875	1143
23.	Albany State University	1500000			United States Of America		1903	424

We now edit the cells by **Edit cells -> Blank down** property which blanks out the row below another row if they are equal.



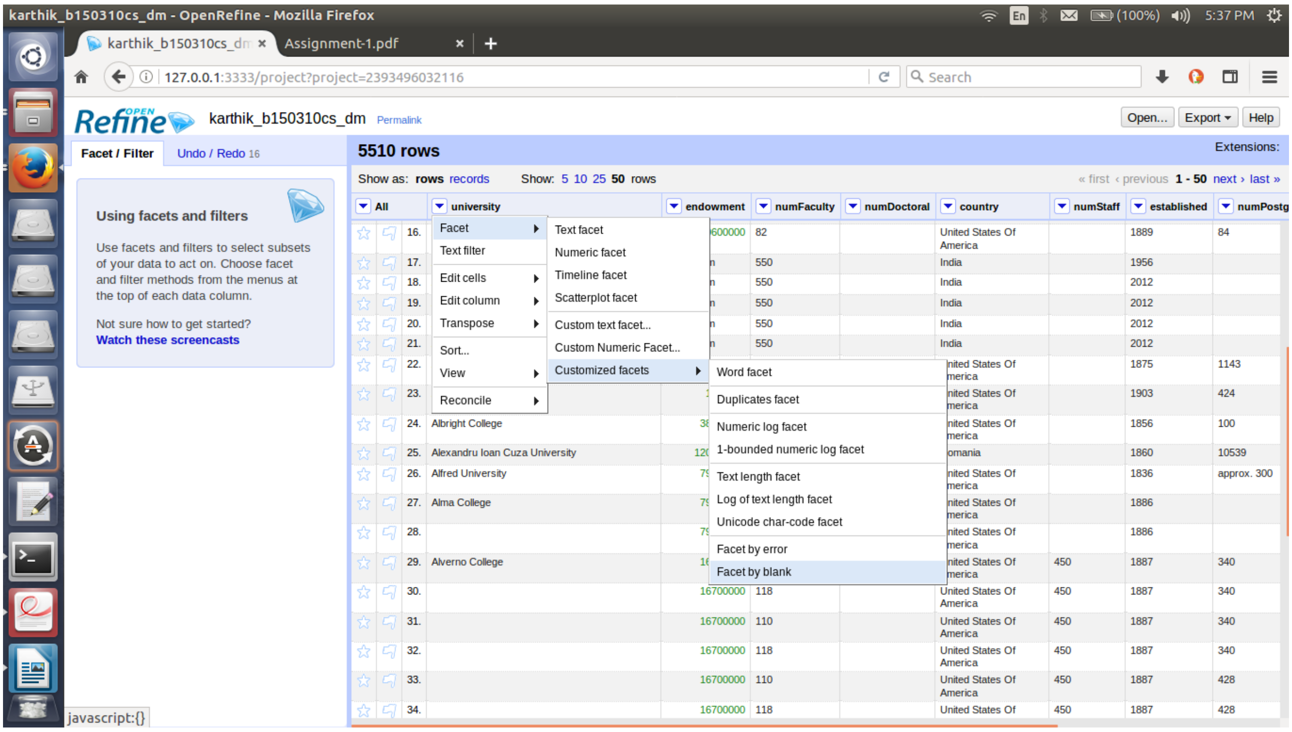
Using facets and filters

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

	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostg
1.	ue de Montr%C3%A9al	$cad145 million		NA	Canada	220	1873	1615
2.	ue de Montr%C3%A9al	$cad145 million		NA	Canada	220	1873	1615
3.		000000			Denmark	11000	1928	16395
4.		000000			Denmark	11000	1928	16395
5.		000000			Denmark	11382	1928	16395
6.		000000			Denmark	11382	1928	16395
7.		000000		NA	Denmark	11000	1928	16395
8.		000000		NA	Denmark	11000	1928	16395
9.		000000		NA	Denmark	11382	1928	16395
10.		000000		NA	Denmark	11382	1928	16395
11.	Acadia University	000000			Canada	211	1838	76
12.	Acadia University	40000000			Canada	211	1838	76
13.	Acadia University	40000000			Canada	211	1838	76
14.	Acadia University	40000000			Canada	211	1838	76
15.	Adelphi University	86000000		NA	United States Of America	956	1896	3466
16.	Agnes Scott College	230600000	82		United States Of America		1889	84
17.	AIIMS%2C New Delhi	per annum	550		India		1956	
18.	AIIMS Bhopal	per annum	550		India		2012	
19.	AIIMS Jodhpur	per annum	550		India		2012	
20.	AIIMS Raipur	per annum	550		India		2012	
21.	AIIMS Rishikesh	per annum	550		India		2012	
22.	Alabama Agricultural and Mechanical University	119000000			United States Of America		1875	1143
23.	Albany State University	1500000			United States Of America		1903	424

Now we use a customized facet Facet By Blank proprty(**Facet -> Customized facets -> Facet by blank)** and select all the blank rows and remove all the blank rows.

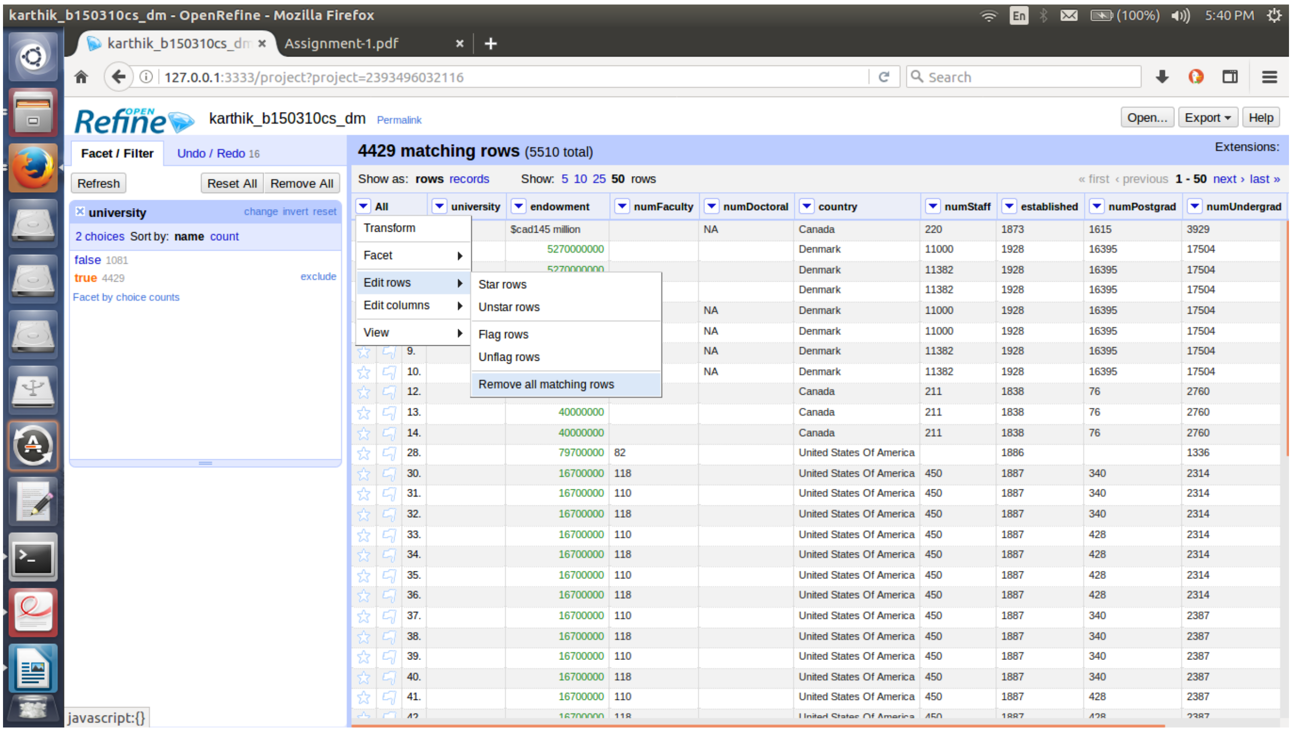


Using facets and filters

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

	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostg
16.		600000	82		United States Of America		1889	84
17.		n	550		India		1956	
18.		n	550		India		2012	
19.		n	550		India		2012	
20.		n	550		India		2012	
21.		n	550		India		2012	
22.					nited States Of merica		1875	1143
23.		1			nited States Of merica		1903	424
24.	Albright College	38			nited States Of merica		1856	100
25.	Alexandru Ioan Cuza University	120			omania		1860	10539
26.	Alfred University	79			nited States Of merica		1836	approx. 300
27.	Alma College	79			nited States Of merica		1886	
28.		79			nited States Of merica		1886	
29.	Alverno College	16			nited States Of merica	450	1887	340
30.		16700000	118		United States Of America	450	1887	340
31.		16700000	110		United States Of America	450	1887	340
32.		16700000	118		United States Of America	450	1887	340
33.		16700000	110		United States Of America	450	1887	428
34.		16700000	118		United States Of	450	1887	428



4429 matching rows (5510 total)

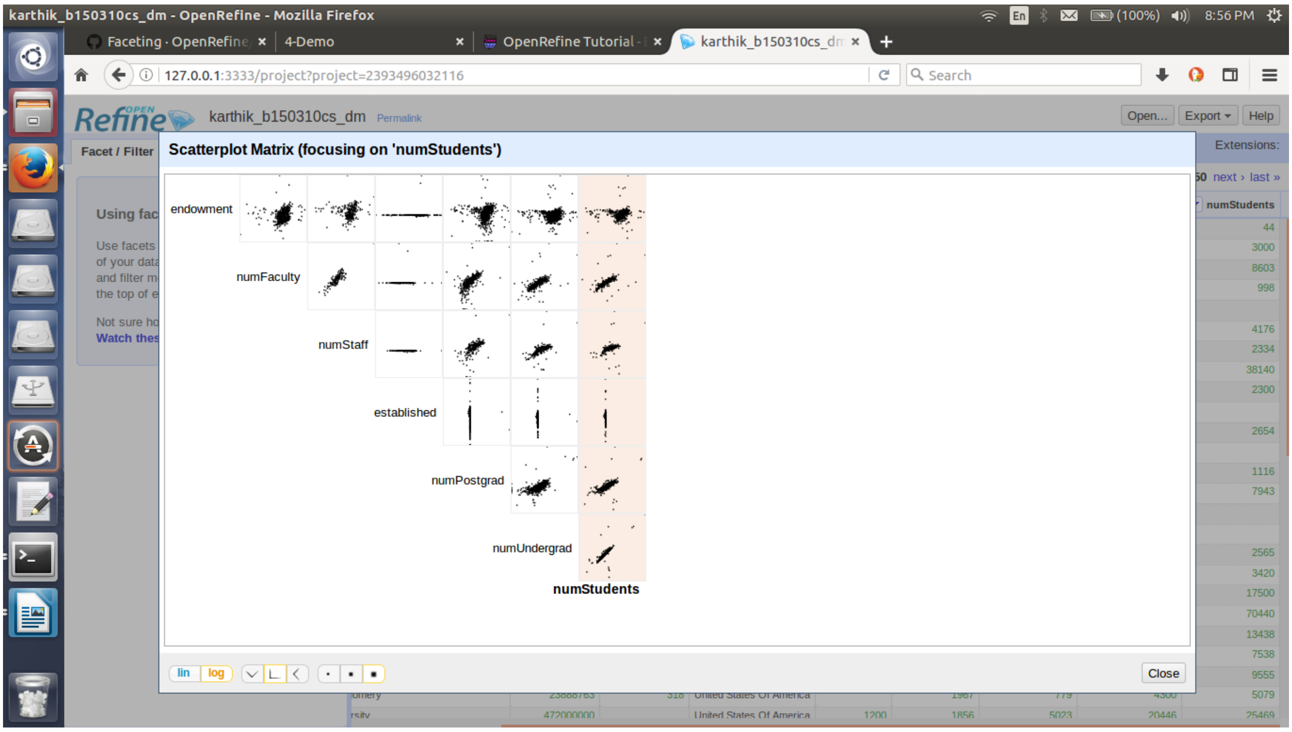
Show as: **rows** records Show: 5 10 25 **50** rows « first ‹ previous **1 - 50** next › last »

All	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad
		$cad145 million		NA	Canada	220	1873	1615	3929
		5270000000			Denmark	11000	1928	16395	17504
		5270000000			Denmark	11382	1928	16395	17504
					Denmark	11382	1928	16395	17504
				NA	Denmark	11000	1928	16395	17504
				NA	Denmark	11000	1928	16395	17504
9.				NA	Denmark	11382	1928	16395	17504
10.				NA	Denmark	11382	1928	16395	17504
12.					Canada	211	1838	76	2760
13.		40000000			Canada	211	1838	76	2760
14.		40000000			Canada	211	1838	76	2760
28.		79700000	82		United States Of America		1886		1336
30.		16700000	118		United States Of America	450	1887	340	2314
31.		16700000	110		United States Of America	450	1887	340	2314
32.		16700000	118		United States Of America	450	1887	340	2314
33.		16700000	110		United States Of America	450	1887	428	2314
34.		16700000	118		United States Of America	450	1887	428	2314
35.		16700000	110		United States Of America	450	1887	428	2314
36.		16700000	118		United States Of America	450	1887	428	2314
37.		16700000	110		United States Of America	450	1887	340	2387
38.		16700000	118		United States Of America	450	1887	340	2387
39.		16700000	110		United States Of America	450	1887	340	2387
40.		16700000	118		United States Of America	450	1887	340	2387
41.		16700000	110		United States Of America	450	1887	428	2387
42.		16700000	118		United States Of America	450	1887	428	2387

Exploring the data with scatter plots

Click on the "endowment" column, **Facet -> Scatterplot facet**.

Use this scatterplot on all the attributes and generate the scatter plot.



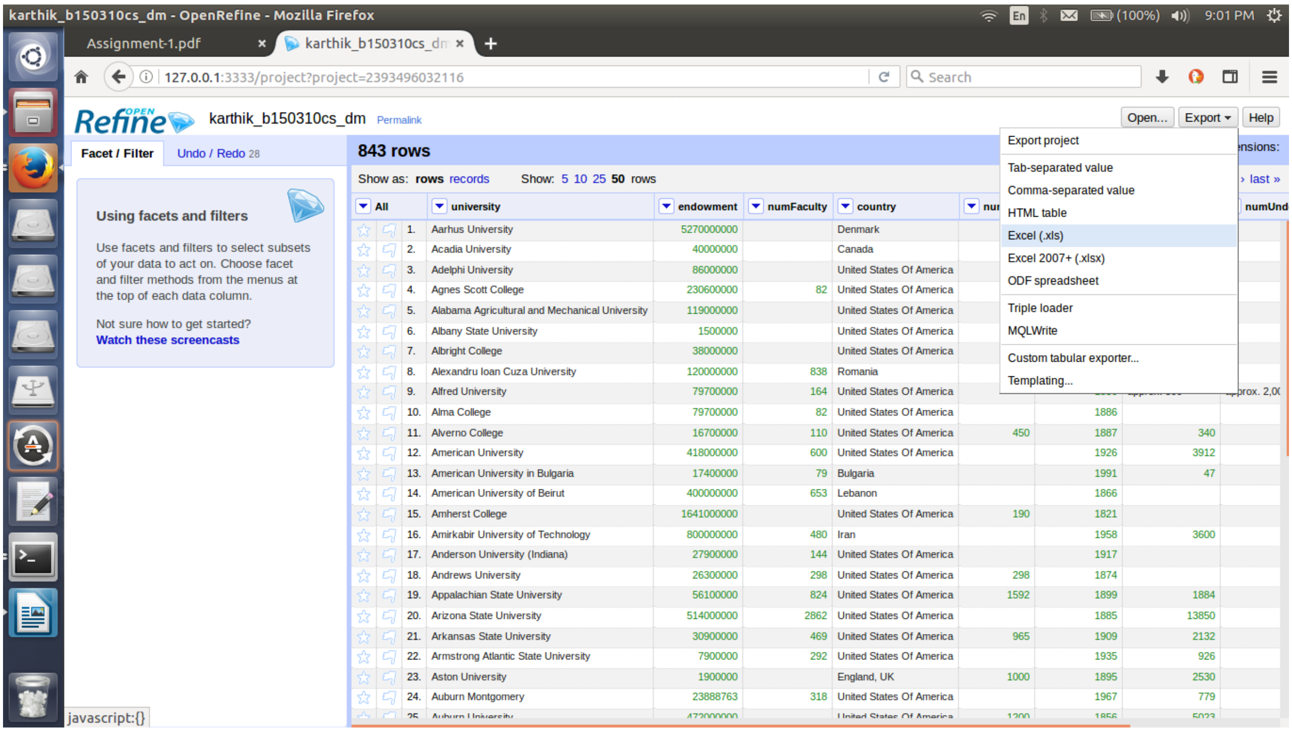
Missing values

Open Refine does not have an option to fill the missing values whereas weka has.

So we have to fill out the missing values manually in OpenRefine.

Exporting data to an Excel sheet

Export the final dataset in .xls format.

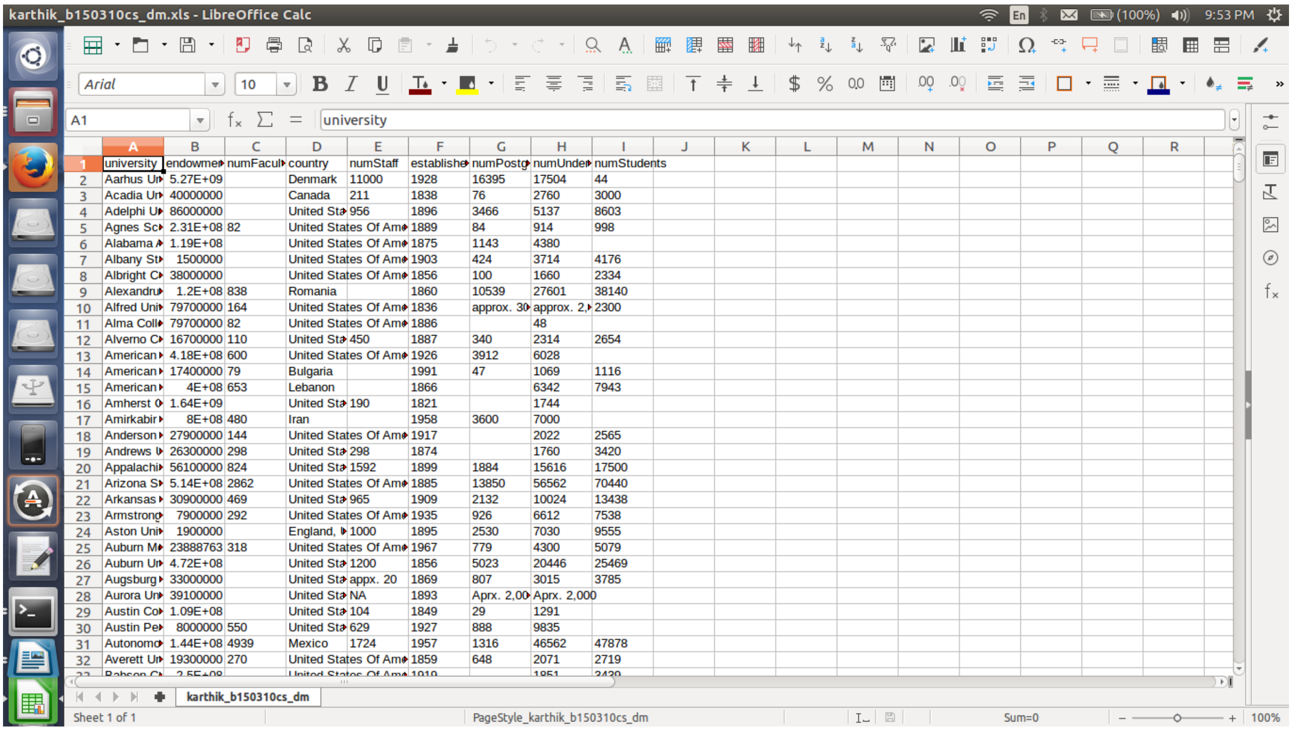


843 rows

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

All	university	endowment	numFaculty	country	nu			numUnd
1.	Aarhus University	5270000000		Denmark				
2.	Acadia University	40000000		Canada				
3.	Adelphi University	86000000		United States Of America				
4.	Agnes Scott College	230600000	82	United States Of America				
5.	Alabama Agricultural and Mechanical University	119000000		United States Of America				
6.	Albany State University	1500000		United States Of America				
7.	Albright College	38000000		United States Of America				
8.	Alexandru Ioan Cuza University	120000000	838	Romania				
9.	Alfred University	79700000	164	United States Of America				approx. 2,0
10.	Alma College	79700000	82	United States Of America		1886		
11.	Alverno College	16700000	110	United States Of America	450	1887	340	
12.	American University	418000000	600	United States Of America		1926	3912	
13.	American University in Bulgaria	17400000	79	Bulgaria		1991	47	
14.	American University of Beirut	400000000	653	Lebanon		1866		
15.	Amherst College	1641000000		United States Of America	190	1821		
16.	Amirkabir University of Technology	800000000	480	Iran		1958	3600	
17.	Anderson University (Indiana)	27900000	144	United States Of America		1917		
18.	Andrews University	26300000	298	United States Of America	298	1874		
19.	Appalachian State University	56100000	824	United States Of America	1592	1899	1884	
20.	Arizona State University	514000000	2862	United States Of America		1885	13850	
21.	Arkansas State University	30900000	469	United States Of America	965	1909	2132	
22.	Armstrong Atlantic State University	7900000	292	United States Of America		1935	926	
23.	Aston University	1900000		England, UK	1000	1895	2530	
24.	Auburn Montgomery	23888763	318	United States Of America		1967	779	
25.	Auburn University	472000000		United States Of America	1200	1856	5023	

Final Data



karthik_b150310cs_dm.xls - LibreOffice Calc
Arial
10
A1
university
karthik_b150310cs_dm
Sheet 1 of 1
PageStyle_karthik_b150310cs_dm
Sum=0
100%

	A	B	C	D	E	F	G	H	I
1	university	endowmen	numFacul	country	numStaff	establishe	numPostg	numUnder	numStudents
2	Aarhus Un	5.27E+09		Denmark	11000	1928	16395	17504	44
3	Acadia Un	40000000		Canada	211	1838	76	2760	3000
4	Adelphi U	86000000		United Sta	956	1896	3466	5137	8603
5	Agnes Sc	2.31E+08	82	United States Of Ame		1889	84	914	998
6	Alabama A	1.19E+08		United States Of Ame		1875	1143	4380	
7	Albany St	1500000		United States Of Ame		1903	424	3714	4176
8	Albright C	38000000		United States Of Ame		1856	100	1660	2334
9	Alexandru	1.2E+08	838	Romania		1860	10539	27601	38140
10	Alfred Uni	79700000	164	United States Of Ame		1836	approx. 30	approx. 2,	2300
11	Alma Coll	79700000	82	United States Of Ame		1886		48	
12	Alverno C	16700000	110	United Sta	450	1887	340	2314	2654
13	American	4.18E+08	600	United States Of Ame		1926	3912	6028	
14	American	17400000	79	Bulgaria		1991	47	1069	1116
15	American	4E+08	653	Lebanon		1866		6342	7943
16	Amherst C	1.64E+09		United Sta	190	1821		1744	
17	Amirkabir	8E+08	480	Iran		1958	3600	7000	
18	Anderson	27900000	144	United States Of Ame		1917		2022	2565
19	Andrews U	26300000	298	United Sta	298	1874		1760	3420
20	Appalachi	56100000	824	United Sta	1592	1899	1884	15616	17500
21	Arizona S	5.14E+08	2862	United States Of Ame		1885	13850	56562	70440
22	Arkansas	30900000	469	United Sta	965	1909	2132	10024	13438
23	Armstrong	7900000	292	United States Of Ame		1935	926	6612	7538
24	Aston Uni	1900000		England,	1000	1895	2530	7030	9555
25	Auburn M	23888763	318	United States Of Ame		1967	779	4300	5079
26	Auburn Un	4.72E+08		United Sta	1200	1856	5023	20446	25469
27	Augsburg	33000000		United Sta	appx. 20	1869	807	3015	3785
28	Aurora Un	39100000		United Sta	NA	1893	Aprx. 2,00	Aprx. 2,000	
29	Austin Co	1.09E+08		United Sta	104	1849	29	1291	
30	Austin Pe	8000000	550	United Sta	629	1927	888	9835	
31	Autonomo	1.44E+08	4939	Mexico	1724	1957	1316	46562	47878
32	Averett Un	19300000	270	United States Of Ame		1859	648	2071	2719
33	Babson C	2.5E+08		United States Of Ame		1919		1951	3439