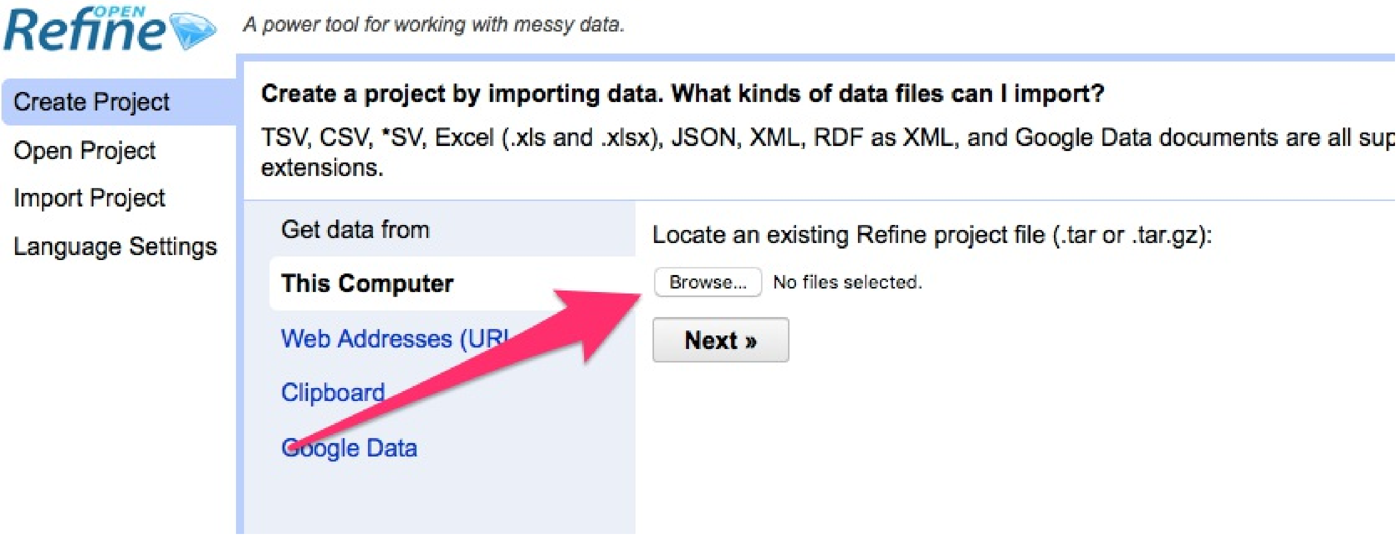
Data normalization walkthrough

Data normalization is where you make sure all things are the same in data you need. In data, St. Petersburg, St Petersburg, St petersburg, ST Petersburg, St Petersberg and St Pete are all different things. When you need to do a count of each city, that's a bad, bad thing.

This begins you on a long road of interrelated problems and solutions that are all meant to get you where you need to go. Some data fixes can be done in a text editor -- one or two simple things. Some data fixes, particularly those you'll do over and over and over every month or every day even, should be programmed and run by a computer. And some fixes are made for Open Refine, a power tool that is very, very good at normalizing data.

Let's do a simple one first. Fire up Open Refine on your computer and then open a browser and go to http://127.0.0.1:3333/. That'll take you to a web application running inside your computer -- Open Refine.

Click browse and find the Mountain Lion data we used in our Agate group-by walkthrough.



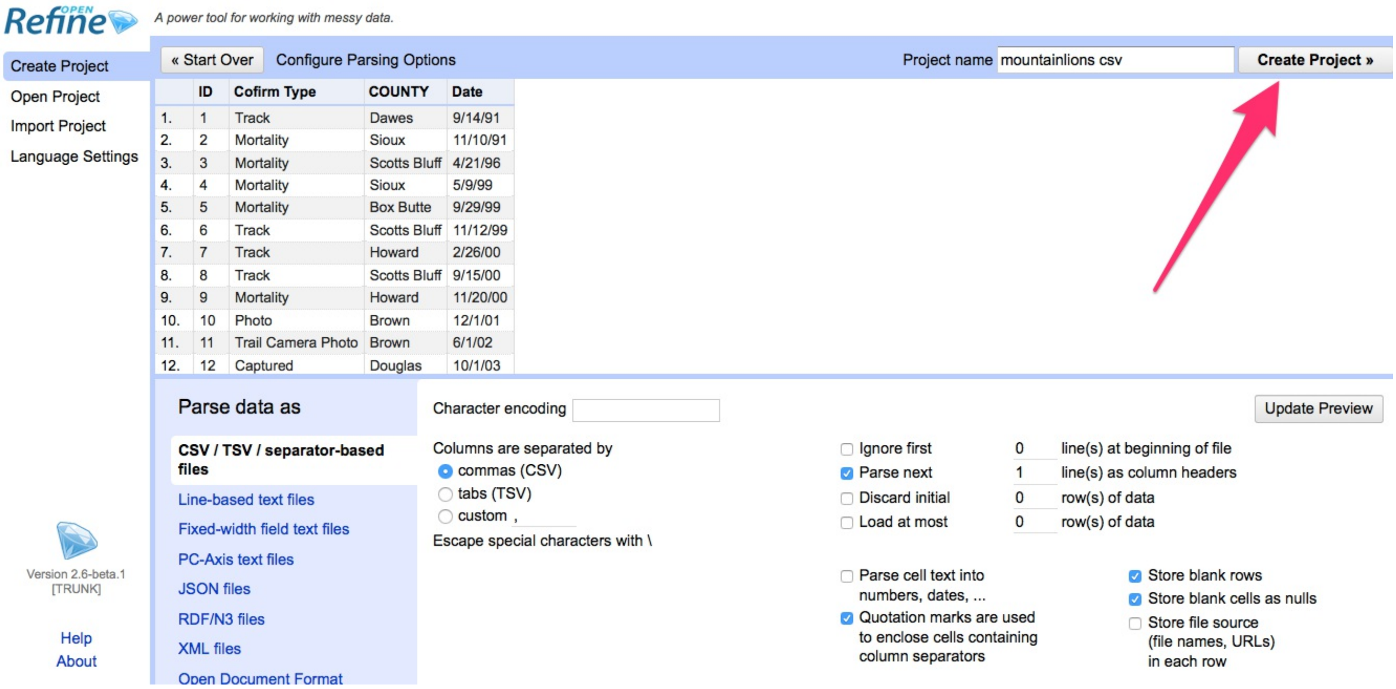
A power tool for working with messy data.

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 sup
extensions.

Locate an existing Refine project file (.tar or .tar.gz):

After we import it, we look at it and see it looks good, so we click Create Project.



Project name mountainlions csv

	ID	Cofirm Type	COUNTY	Date
1.	1	Track	Dawes	9/14/91
2.	2	Mortality	Sioux	11/10/91
3.	3	Mortality	Scotts Bluff	4/21/96
4.	4	Mortality	Sioux	5/9/99
5.	5	Mortality	Box Butte	9/29/99
6.	6	Track	Scotts Bluff	11/12/99
7.	7	Track	Howard	2/26/00
8.	8	Track	Scotts Bluff	9/15/00
9.	9	Mortality	Howard	11/20/00
10.	10	Photo	Brown	12/1/01
11.	11	Trail Camera Photo	Brown	6/1/02
12.	12	Captured	Douglas	10/1/03

Parse data as

- CSV / TSV / separator-based files
- Line-based text files
- Fixed-width field text files
- PC-Axis text files
- JSON files
- RDF/N3 files
- XML files
- Open Document Format

Character encoding

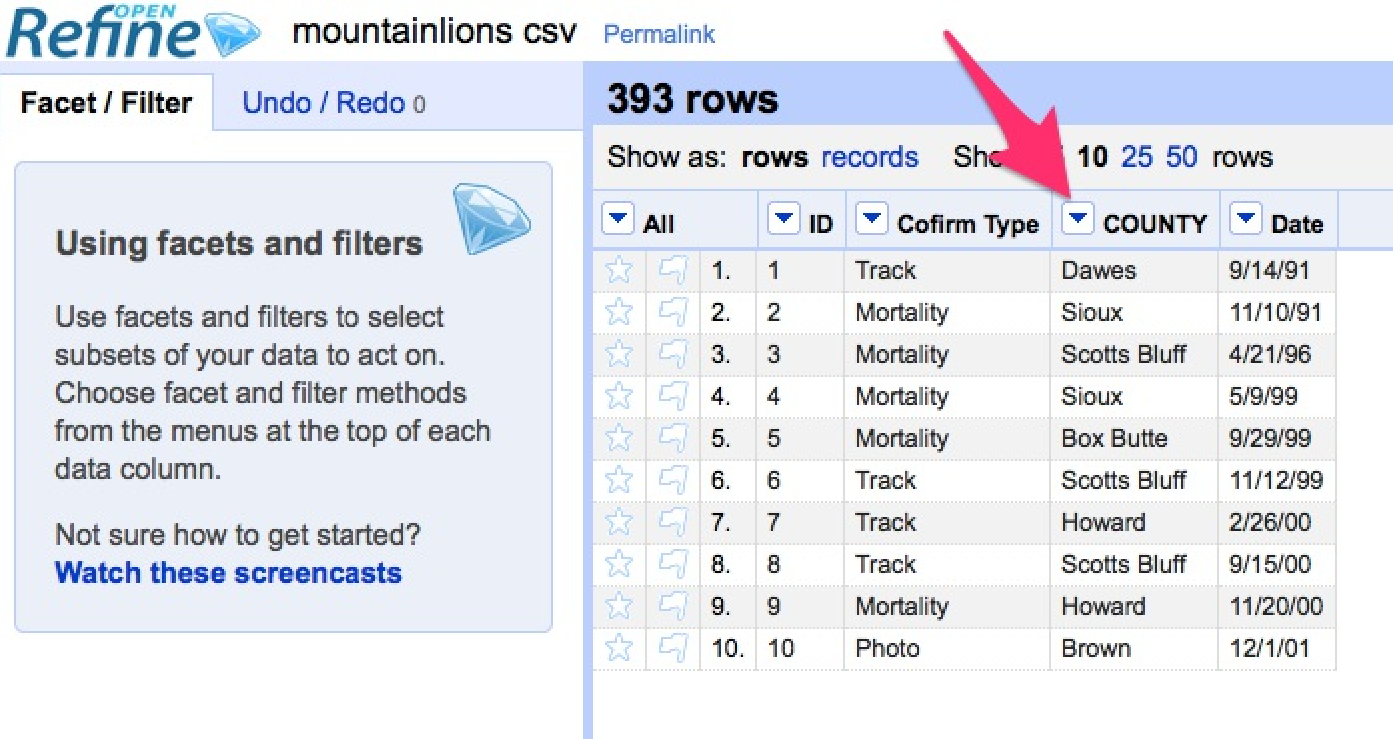
Columns are separated by
● commas (CSV)
○ tabs (TSV)
○ custom ,

Escape special characters with \

□ Ignore first 0 line(s) at beginning of file
☑ Parse next 1 line(s) as column headers
□ Discard initial 0 row(s) of data
□ Load at most 0 row(s) of data

□ Parse cell text into numbers, dates, ...
☑ Quotation marks are used to enclose cells containing column separators
☑ Store blank rows
☑ Store blank cells as nulls
□ Store file source (file names, URLs) in each row

With it imported, we look at the columns, and we see this dropdown menus. Click on the one next to County.



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.

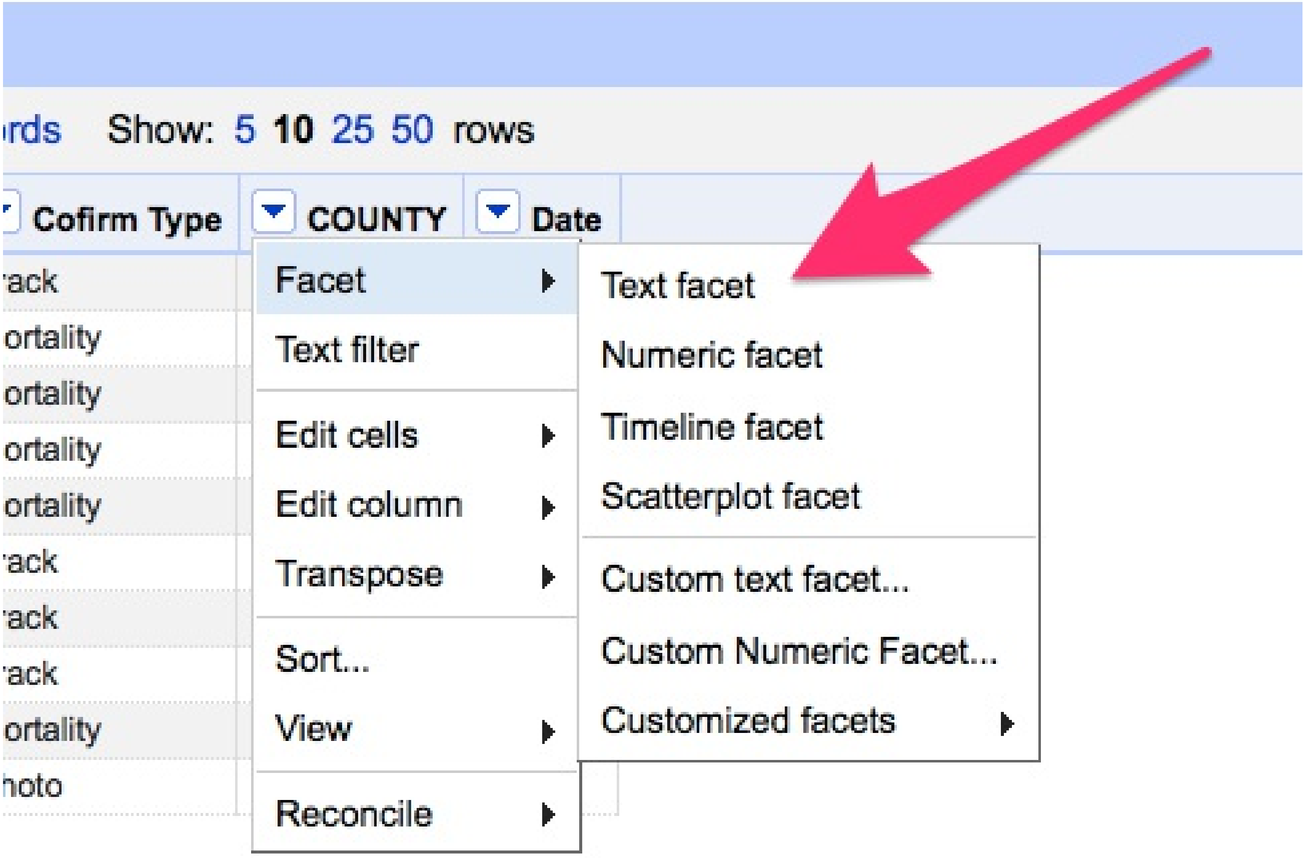
Not sure how to get started?
Watch these screencasts

393 rows

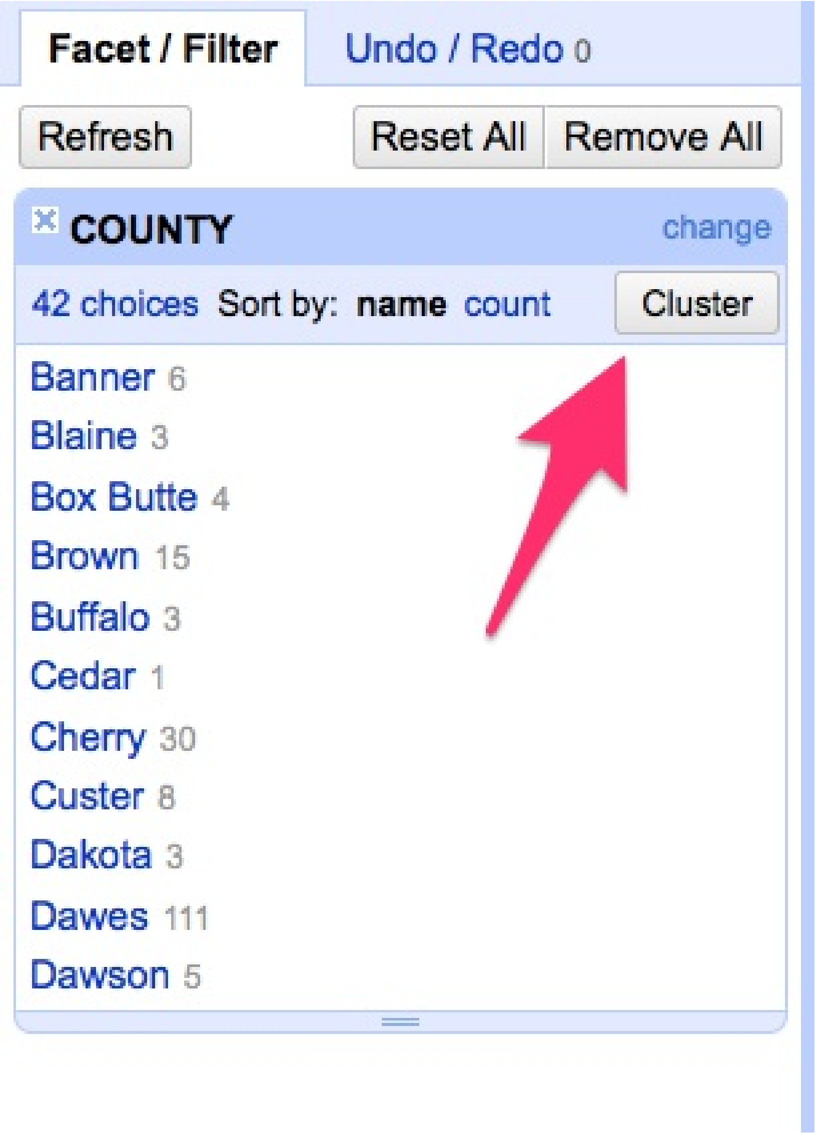
Show as: **rows** records Sh[o]w: **10** 25 50 rows

All			ID	Cofirm Type	COUNTY	Date
☆		1.	1	Track	Dawes	9/14/91
☆		2.	2	Mortality	Sioux	11/10/91
☆		3.	3	Mortality	Scotts Bluff	4/21/96
☆		4.	4	Mortality	Sioux	5/9/99
☆		5.	5	Mortality	Box Butte	9/29/99
☆		6.	6	Track	Scotts Bluff	11/12/99
☆		7.	7	Track	Howard	2/26/00
☆		8.	8	Track	Scotts Bluff	9/15/00
☆		9.	9	Mortality	Howard	11/20/00
☆		10.	10	Photo	Brown	12/1/01

Open Refine works with Facets, which are collections of data -- think of them like Group Bys on acid. Counties are text, so we'll use a Text Facet.

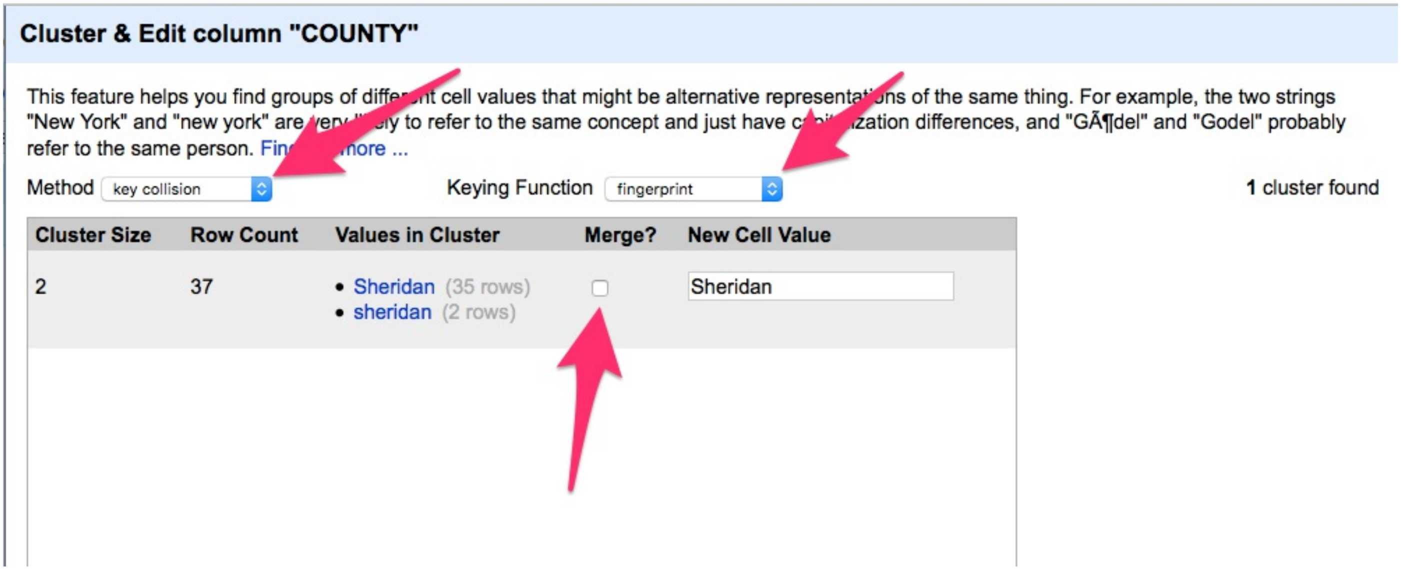


Facets pop a box up on the left, and we want to Cluster them together, so click Cluster.



Banner	6
Blaine	3
Box Butte	4
Brown	15
Buffalo	3
Cedar	1
Cherry	30
Custer	8
Dakota	3
Dawes	111
Dawson	5

Using a variety of algorithms, Open Refines clusters things that appear to be alike together based on the outcome of those algorithms. There isn't a magic bullet here: you've got to try them out, sort through the results and make decisions.



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 "GÃ¶del" and "Godel" probably refer to the same person. Find out more ...

Method key collision

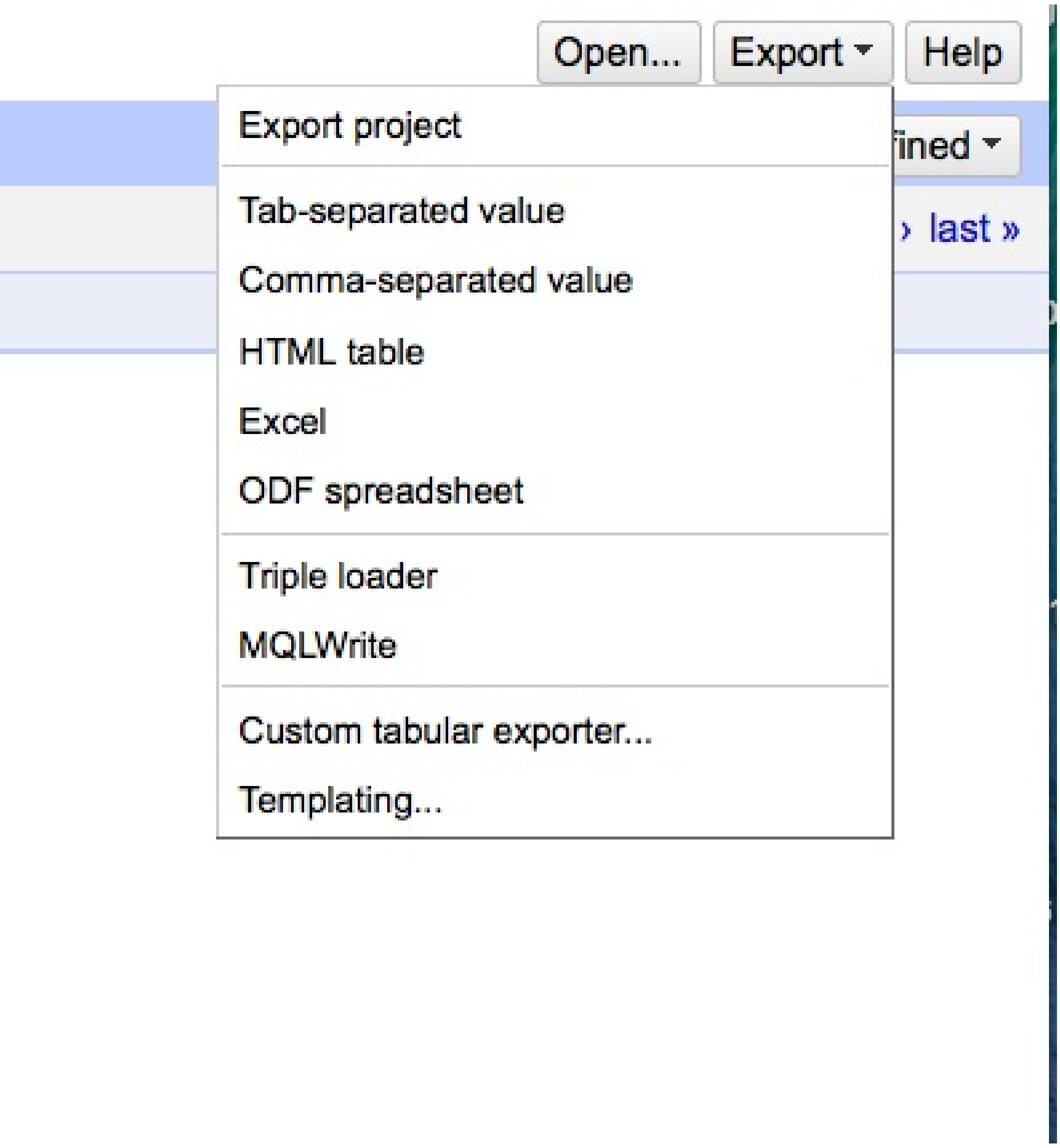
Keying Function fingerprint

1 cluster found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	37	• Sheridan (35 rows) • sheridan (2 rows)	□	Sheridan

This one is easy. Click Merge and then click the Merge checkbox and then Merge Selected and Close. Normally, you'd Merge Selected and Recluster to see if anything new is happening.

Now we export it out.



MAJOR RULE: Work from a copy. Do not alter the original file.

Data normalization assignment

In this assignment, you must clean a file from the United States State Department on refugees resettled in the United States since 2002. We want to group the unique cities together and sum the total number of refugees resettled in them since 2002.

The city field is dirty. It contains mulitple entries for the same city.

1. Download the file. The file you want is called refugees.csv.
2. Normalize the data using Open Refine. Specifically, the field you need to normalize is the city name.
3. Export your newly cleaned data into a new csv file.

4. Import your newly cleaned up data into Agate.
5. Group and sum by unique city name. Remember, there can be more than one city named something. So be sure you are adding the state to it.
6. Turn in your ipynb file and your csv file for credit.