Open Refine

for efficient data cleaning

Step 1

Open the Georgia Spending file in Google Sheets and take a look. Does everything make sense? We'll be focusing on the recipient_name and recipient_city columns.

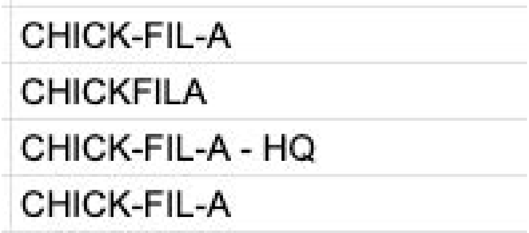
Step 2

Let’s use the skills we learned in Derek’s Evaluating Data lecture to assess this data. What problems do you observe?

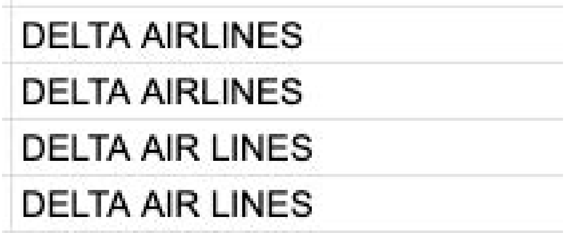
Step 3

Clearly we need to standardize some names and cities.

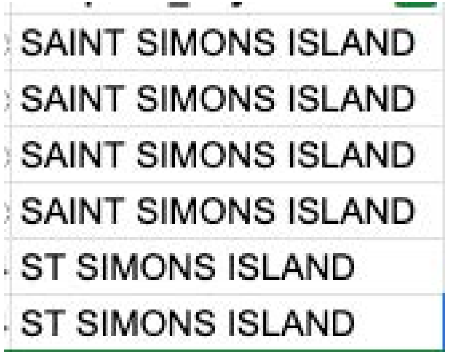
Examples:



CHICK-FIL-A
CHICKFILA
CHICK-FIL-A - HQ
CHICK-FIL-A



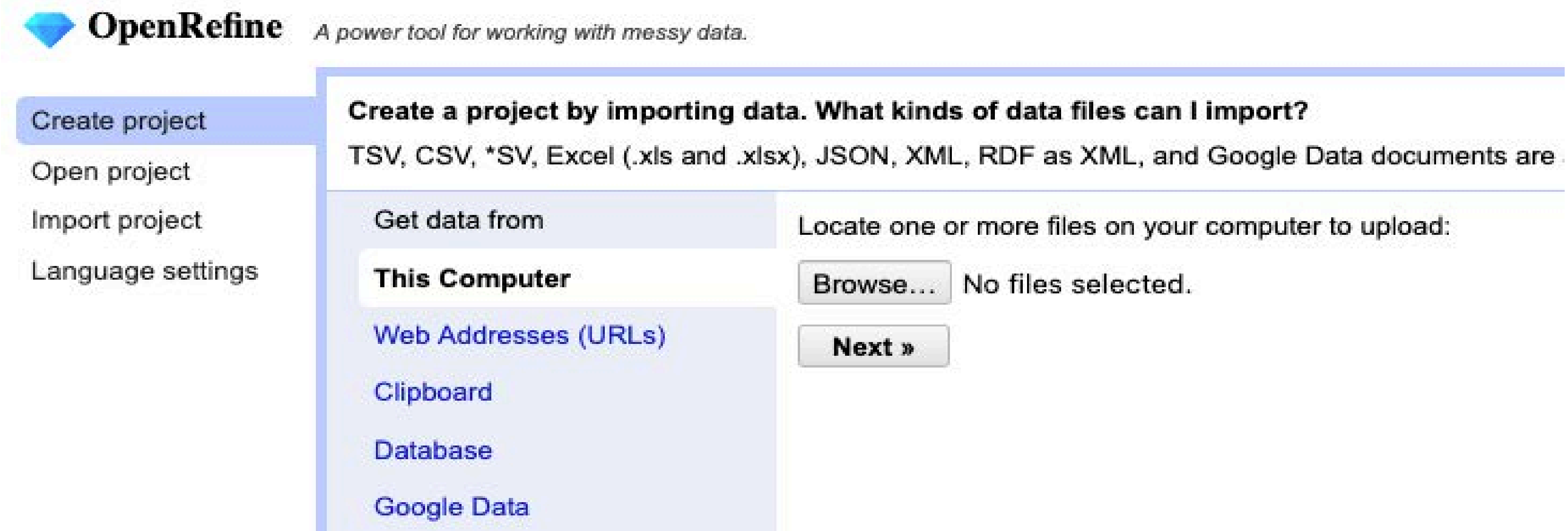
DELTA AIRLINES
DELTA AIRLINES
DELTA AIR LINES
DELTA AIR LINES



SAINT SIMONS ISLAND
SAINT SIMONS ISLAND
SAINT SIMONS ISLAND
SAINT SIMONS ISLAND
ST SIMONS ISLAND
ST SIMONS ISLAND

Step 4

Go to the applications folder and click Open Refine to open. This should automatically launch a browser window. If it doesn't you can go here: http://127.0.0.1:3333/



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

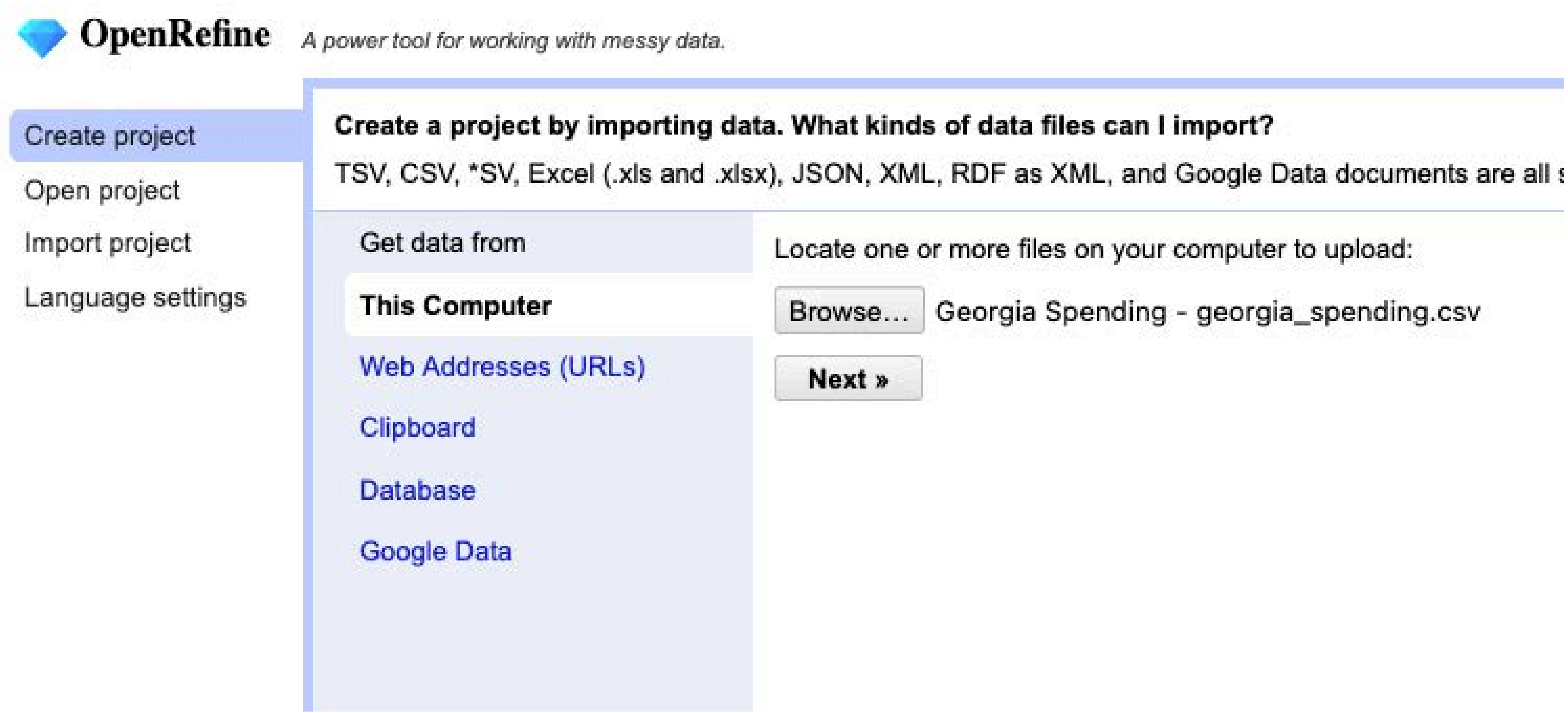
TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are

Locate one or more files on your computer to upload:

No files selected.

Step 5

Export the sheet as CSV and import that file in OpenRefine



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 s

Locate one or more files on your computer to upload:

Georgia Spending - georgia_spending.csv

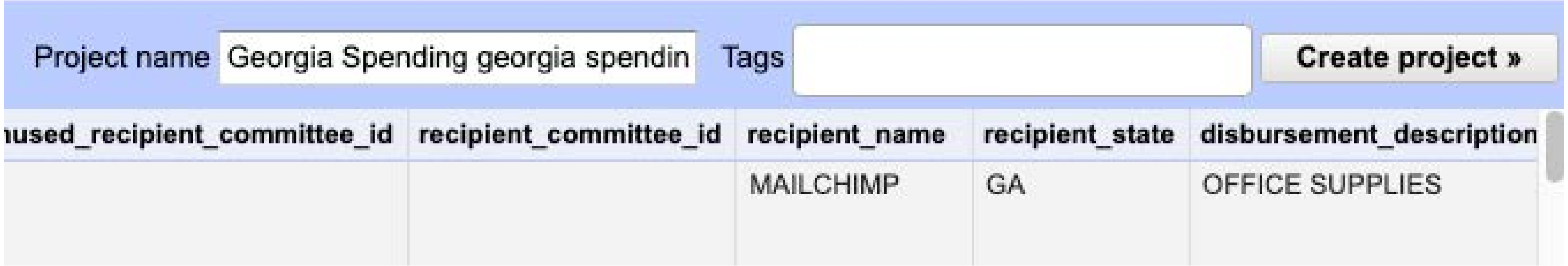
Step 6

Once uploaded, you should see a preview. Check it out: Does it look right? Are there rows and columns like in the Google Sheet?

Note the options in the lower part of the screen. Has Open Refine selected the right ones?

Step 7

If everything looks right, click ‘Create Project’ in the upper right corner to continue.



Project name: Georgia Spending georgia spendin

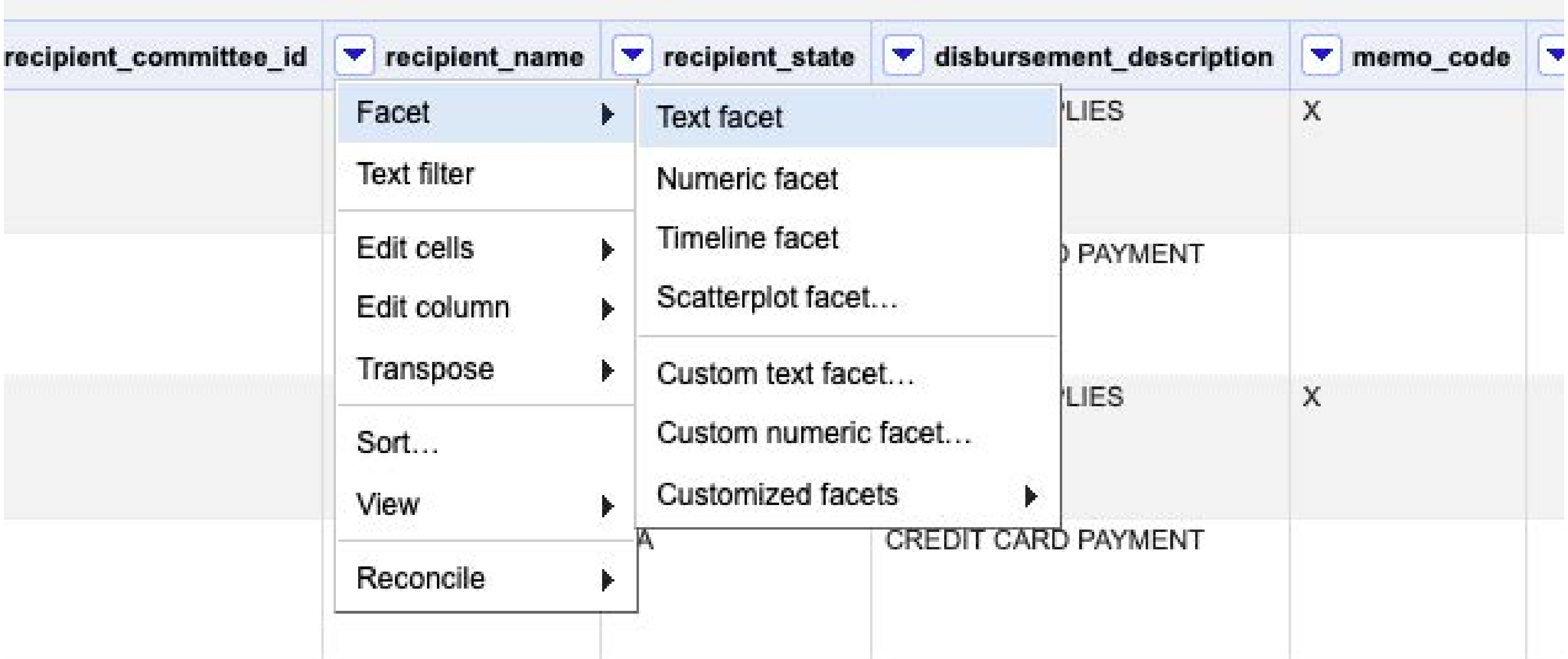
Tags:

used_recipient_committee_id	recipient_committee_id	recipient_name	recipient_state	disbursement_description
		MAILCHIMP	GA	OFFICE SUPPLIES

Step 8

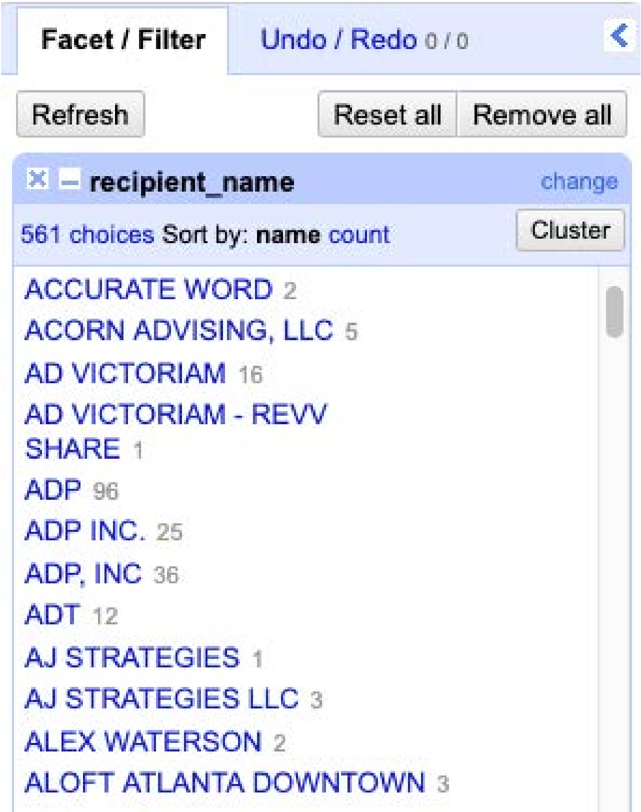
Open Refine relies on things called ‘facets’ to help us clean up data.

Click on the triangle next to the column header ‘recipient_name’ and select ‘Text Facet’ from the dropdown menu.



Step 9

A ‘facet box’ will appear on the left side of your workspace. Click the cluster button to use one of Refine’s most powerful features.



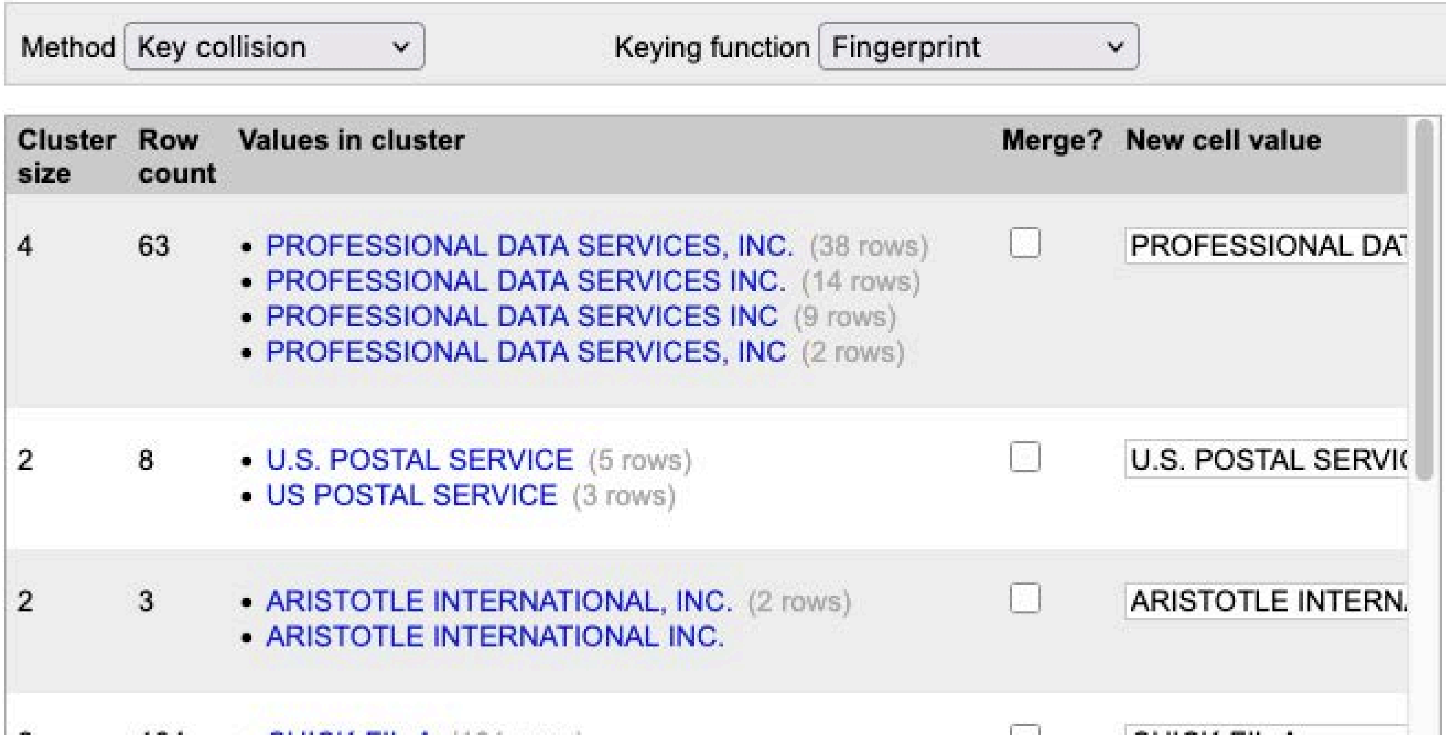
ACCURATE WORD 2
ACORN ADVISING, LLC 5
AD VICTORIAM 16
AD VICTORIAM - REVV SHARE 1
ADP 96
ADP INC. 25
ADP, INC 36
ADT 12
AJ STRATEGIES 1
AJ STRATEGIES LLC 3
ALEX WATERSON 2
ALOFT ATLANTA DOWNTOWN 3

Step 10

Clustering involves using different characteristics of words to group likely identical ones together. Some clustering techniques rely on having letters in common. Others group together words that sound alike even if they are spelled differently. Each method has strengths and weaknesses, so it’s useful to try more than one.

Step 10a

The default clustering method finds some right away



Method: Key collision

Keying function: Fingerprint

Cluster size	Row count	Values in cluster	Merge?	New cell value
4	63	• PROFESSIONAL DATA SERVICES, INC. (38 rows) • PROFESSIONAL DATA SERVICES INC. (14 rows) • PROFESSIONAL DATA SERVICES INC (9 rows) • PROFESSIONAL DATA SERVICES, INC (2 rows)	□	PROFESSIONAL DAT
2	8	• U.S. POSTAL SERVICE (5 rows) • US POSTAL SERVICE (3 rows)	□	U.S. POSTAL SERVIC
2	3	• ARISTOTLE INTERNATIONAL, INC. (2 rows) • ARISTOTLE INTERNATIONAL INC.	□	ARISTOTLE INTERN.

Step 10b

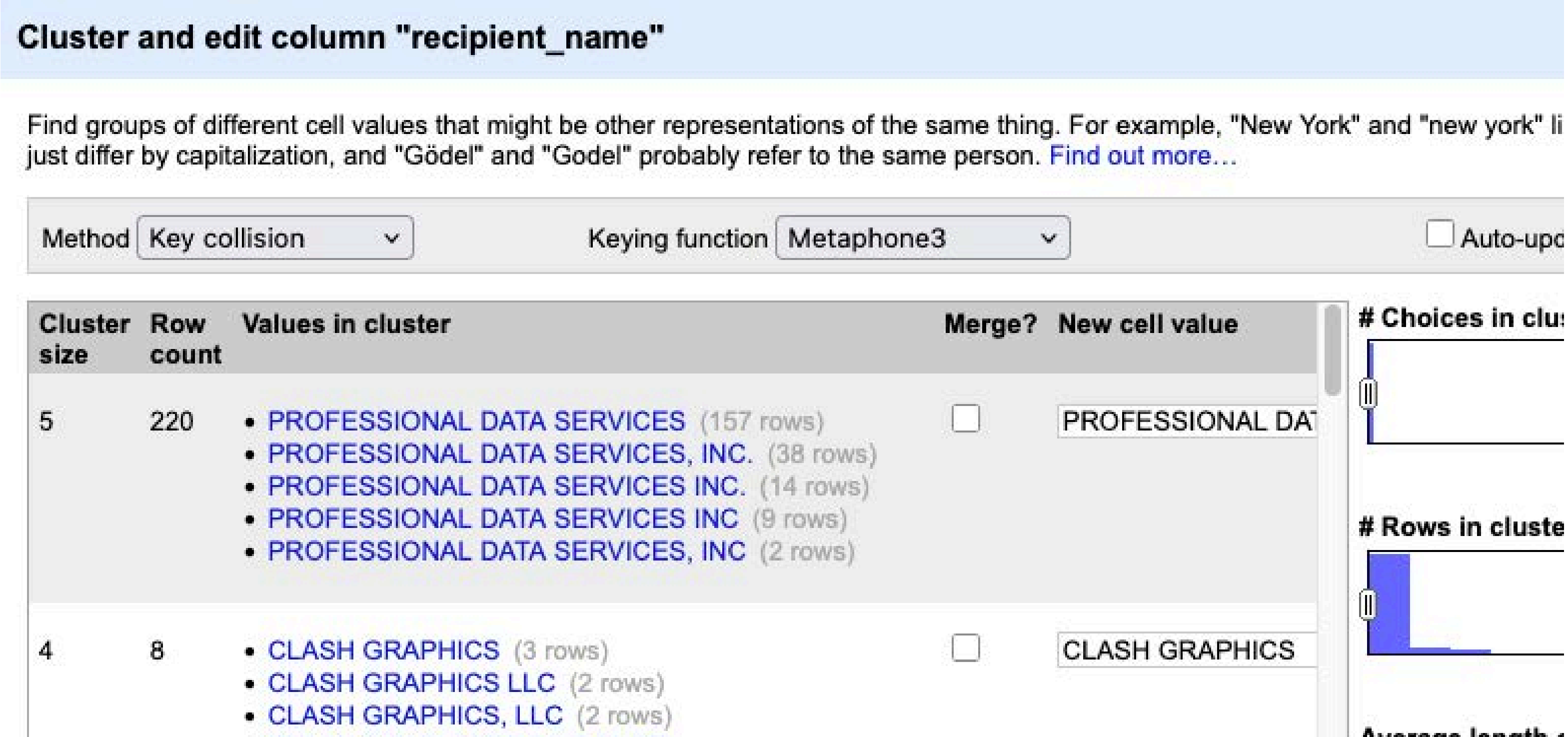
Is this a match we want? If you hover, you get the option to ‘Browse this Cluster’. Click that. A new window will pop open showing just the rows that would be included in that potential cluster.

In this case we can see that the two cities have different zip-codes. They’re probably not actually a match.

Close the additional window to return to the complete data.

Step 11

Let’s click ‘Cluster’ again, and try a different method:



Cluster and edit column "recipient_name"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" li
just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. Find out more…

Method: Key collision ⌄ Keying function: Metaphone3 ⌄ □ Auto-upd

Cluster size	Row count	Values in cluster	Merge?	New cell value
5	220	- PROFESSIONAL DATA SERVICES (157 rows) - PROFESSIONAL DATA SERVICES, INC. (38 rows) - PROFESSIONAL DATA SERVICES INC. (14 rows) - PROFESSIONAL DATA SERVICES INC (9 rows) - PROFESSIONAL DATA SERVICES, INC (2 rows)	□	PROFESSIONAL DAT
4	8	- CLASH GRAPHICS (3 rows) - CLASH GRAPHICS LLC (2 rows) - CLASH GRAPHICS, LLC (2 rows)	□	CLASH GRAPHICS

Step 12

Browse the suggested clusters. Some will look good. Some won’t. If you see one that makes sense, check the ‘merge’ checkbox. Then make sure the ‘New cell value’ is appropriate. If not, you can edit it.

When you’re finished with a cluster, or clusters, you can click ‘Merge selected and re-cluster’ or ‘Merge selected and close.’

It’s a good idea to work your way through the different clustering options.

Step 13

OpenRefine also has some other useful functions you can explore.

- Splitting cells
- Trimming white space
- Changing field data types.

When you are finished, you can export your cleaned file to a new CSV.