Open Refine Exercise

Created March 2013

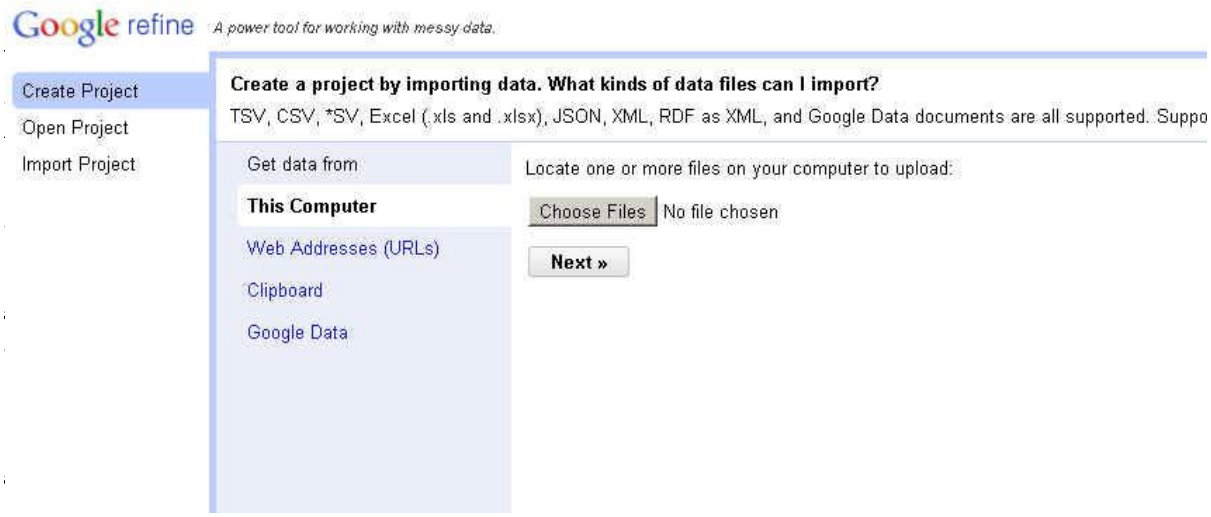
You're going to need a data file called "Franken.xlsx." These are all contributions to the Al Franken for Senate committee during the 2008-09 election cycle.

You'll need to have Open Refine installed on your computer: http://openrefine.org/

1) Launch Refine. This will open up a new window in your web browser
   A little housekeeping, first:
   We need to do one housekeeping task because the data we're working with is rather large. Notice in your address bar it should say "127.0.0.1:3333" or something like that. Edit that by adding "/preferences" to the end of the URL.

   On the Preferences page that pops up, click the Edit button to change the Facet limit. Change it to 5000. Then you can hit the back button to go back to your project.

2) Back to the project…. you'll get the choice to "Create Project", "Open Project" or "Import Project."



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

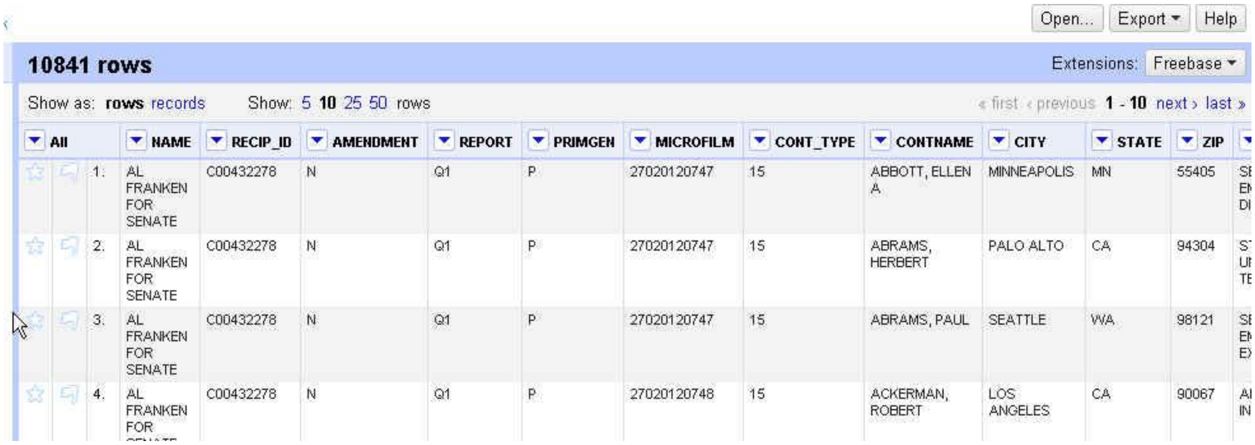
TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Suppo

Locate one or more files on your computer to upload:

Choose "Create Project". Then click "Choose Files" button and navigate to the "Franken.xlsx" data file. Then hit "next"

It will give you a preview of your data and (at the bottom of the page), various options. We're going to leave the defaults. Then click "Create Project" in the upper right corner.

3) Now you’ll see your data – or at least just the first 10 rows.

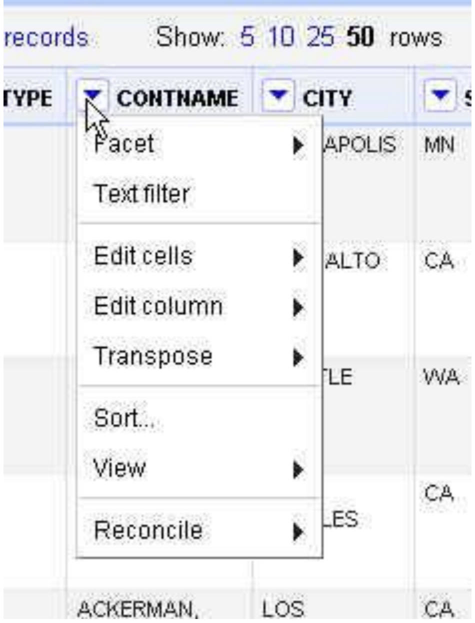


10841 rows

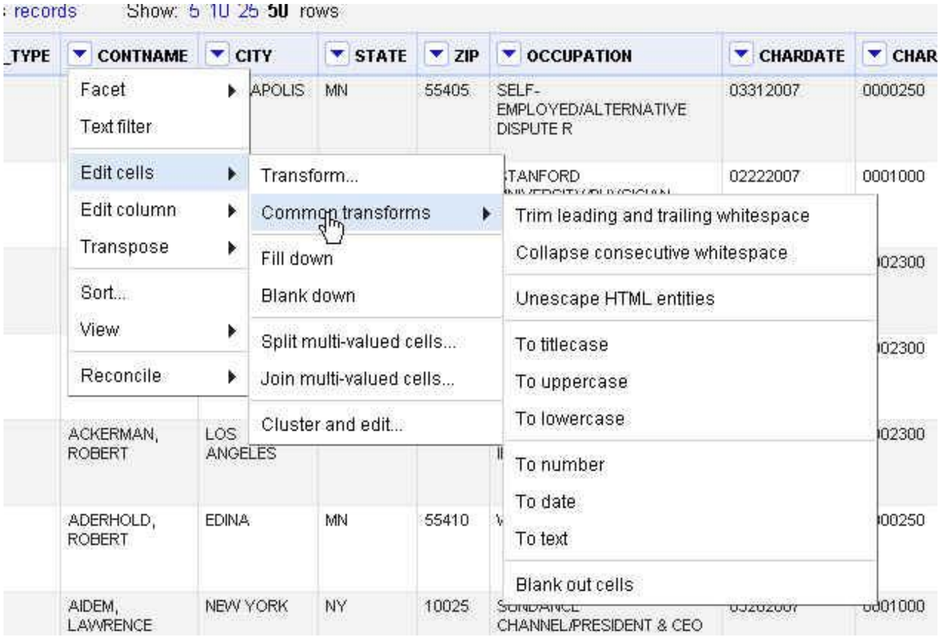
Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 10 next › last »

All	NAME	RECIP_ID	AMENDMENT	REPORT	PRIMGEN	MICROFILM	CONT_TYPE	CONTNAME	CITY	STATE	ZIP	
1.	AL FRANKEN FOR SENATE	C00432278	N	Q1	P	27020120747	15	ABBOTT, ELLEN A	MINNEAPOLIS	MN	55405	SE EM DI
2.	AL FRANKEN FOR SENATE	C00432278	N	Q1	P	27020120747	15	ABRAMS, HERBERT	PALO ALTO	CA	94304	ST UN TE
3.	AL FRANKEN FOR SENATE	C00432278	N	Q1	P	27020120747	15	ABRAMS, PAUL	SEATTLE	WA	98121	SE EM EX
4.	AL FRANKEN FOR	C00432278	N	Q1	P	27020120748	15	ACKERMAN, ROBERT	LOS ANGELES	CA	90067	AI IN

4) We’re going to start with some **simple data cleanup** that you would likely need to do on any dataset. Notice that there are pull-down menus next to each column name. Click the one next to CONTNAME and you’ll see the various cleanup tools we have.



Hover over “Edit Cells” and then “Common Transformation”

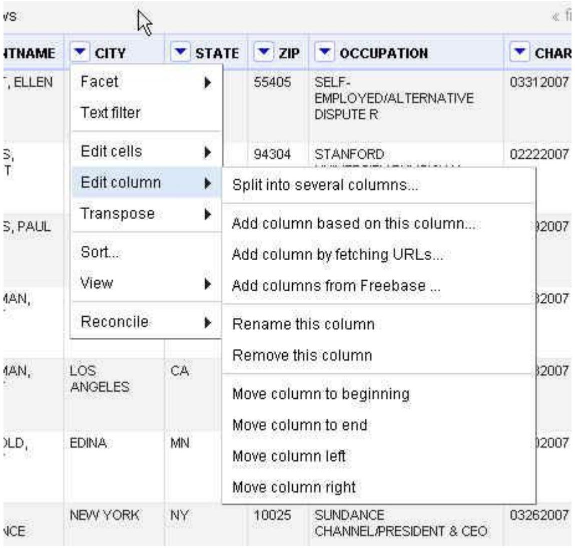


There are lots of great options here, including trimming leading and trailing whitespace, collapsing extra whitespaces in the middle of the data string, turning everything to uppercase. Those are my favorites. Let’s run each of those on CONTNAME.

5) Then repeat those three transformations on CITY, STATE, ZIP, OCCUPATION

6) Next we’re going to start **standardizing data**. The best practice for this is to do the work in a new column, so that you retain the original column of data (in case of errors).

Click on the pull-down for the City field and choose “Edit Column” and then “Add column based on this column”

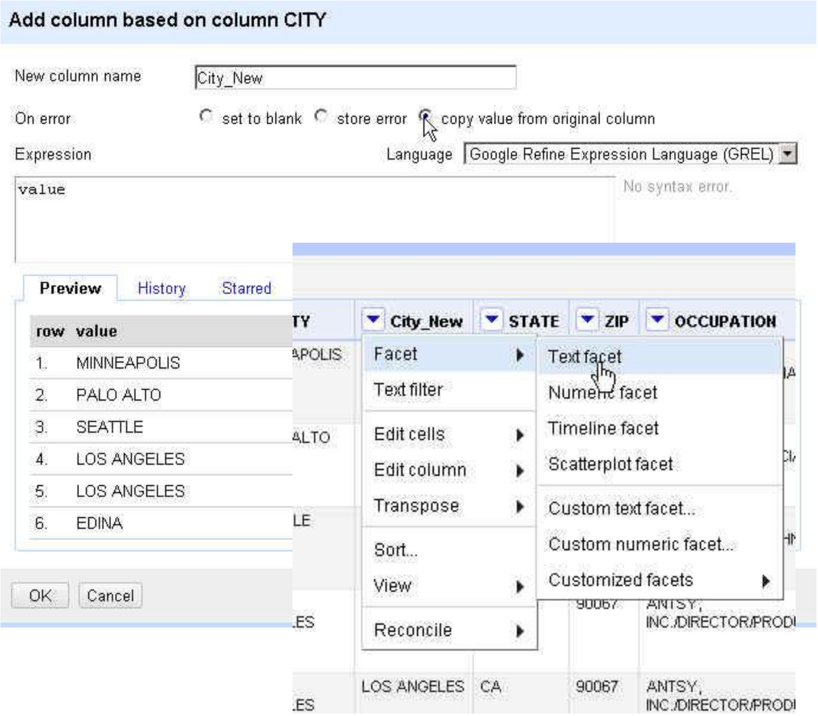


In the pop-up box, give your new column a name…I'm going to call mine "City_New"

And then choose "copy value from original column"

Hit OK.

7) Now go to that new column – "City_New" – and from the pull-down menu choose "Facet" and "Text Facet"



Add column based on column CITY

New column name: City_New

On error: ○ set to blank ○ store error ● copy value from original column

Expression — Language: Google Refine Expression Language (GREL)

```
value
```

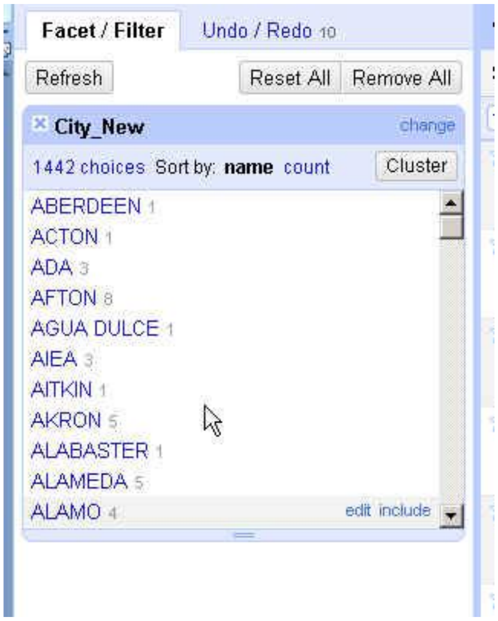
No syntax error.

row	value
1.	MINNEAPOLIS
2.	PALO ALTO
3.	SEATTLE
4.	LOS ANGELES
5.	LOS ANGELES
6.	EDINA

It will add a little box on the left side of the screen showing the various city names that show up in this field.

Go to the bottom of that list of cities and you'll see that we have one blank record. Hover to the right of (blank) and choose "Edit" (you can see it in this picture to the right)

Change that record to "Unknown"



City_New

1442 choices Sort by: name count

ABERDEEN	1
ACTON	1
ADA	3
AFTON	8
AGUA DULCE	1
AIEA	3
AITKIN	1
AKRON	5
ALABASTER	1
ALAMEDA	5
ALAMO	4

8) Close the Facet box for the city_new field. Then let’s repeat all those steps for the State field. First, create a new field called “state_new” (just like in step 5)
Then do a text facet on this new field. This time you’ll see that we have a lot of blank fields. Select “Include” and it will display just those records. You’ll see that we have some records that can obviously be fixed because they have the city name. Fix the ones you can (i.e. Minneapolis, Los Angeles, etc). There are quite a few that are obviously in other countries or you don’t know where they are. Leave those blank.
Then go back to the Facet box and choose to Edit the remaining blank ones. Change them all to “ZZ”

9) In order to standardize the names, I think it would be best to do the city and state together because there are situations where there are cities with the same – or similar – names that are in different states. So first, we need to **merge the city and state together into one field.**

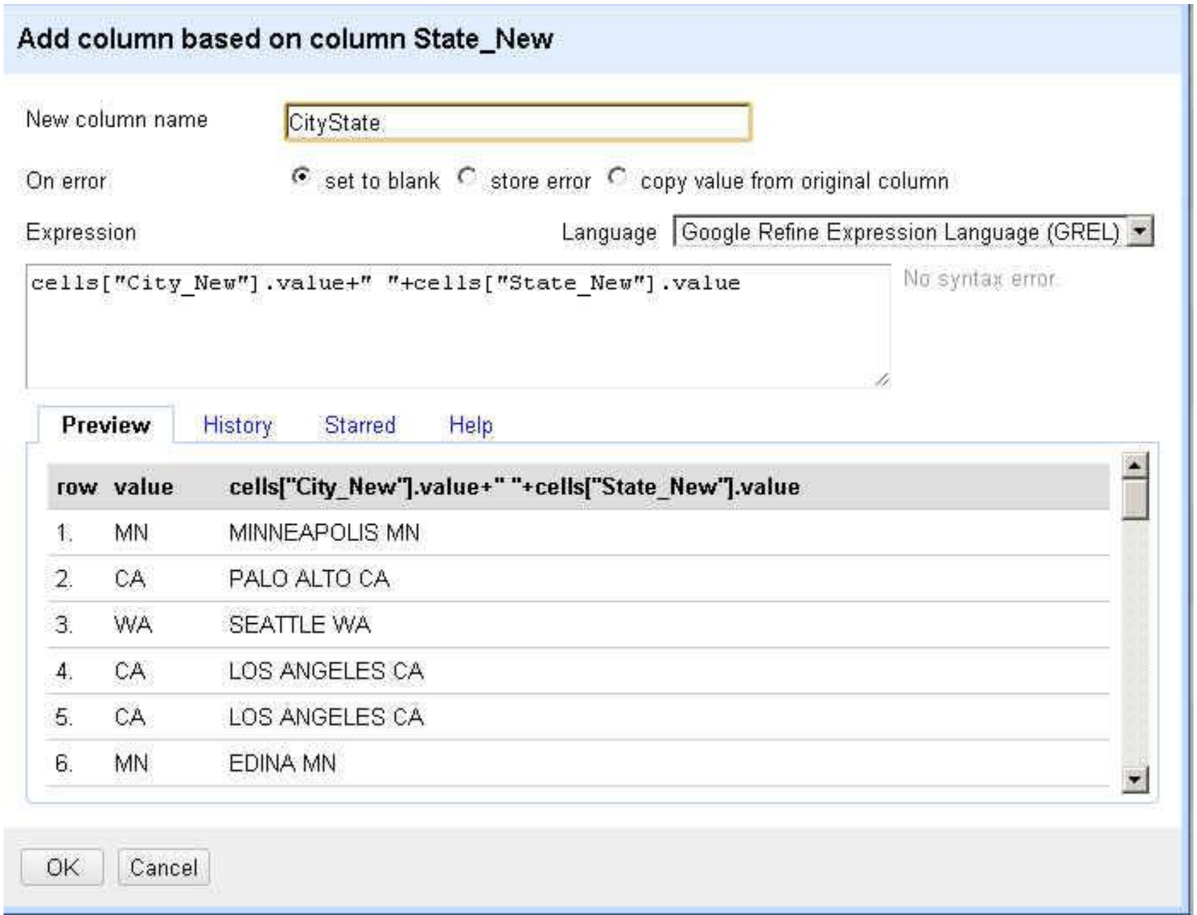
Note: the next step requires that all of your records in those two fields have valid data (no nulls). So that cleanup we did in step 7 was crucial to making this work.

Select either one of your new columns and choose to “add column based on this column”

In the box that comes up (like the one below), give you new column a name – “CityState”
And in the Expression box type the following:

Cells[“City_New”].value + “ “ +cells[“State_New”].value

Note: the names of your city_new and state_new columns (inside those quote marks) are case



Add column based on column State_New

New column name: CityState

On error: ● set to blank ○ store error ○ copy value from original column

Expression — Language: Google Refine Expression Language (GREL)

```
cells["City_New"].value+" "+cells["State_New"].value
```

No syntax error

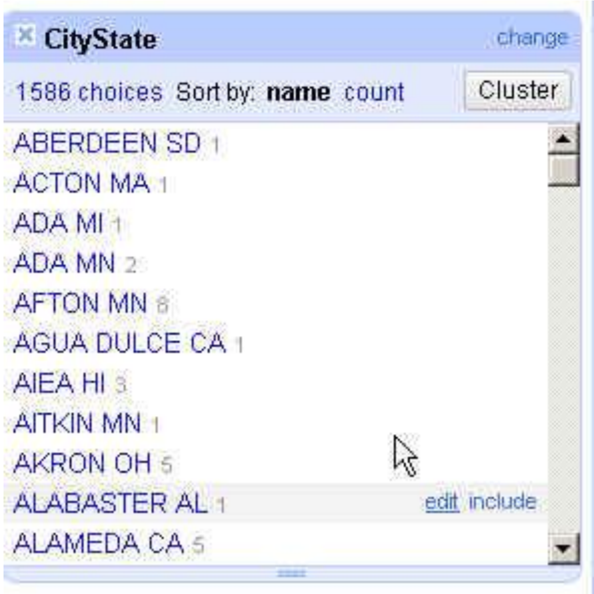
row	value	cells["City_New"].value+" "+cells["State_New"].value
1.	MN	MINNEAPOLIS MN
2.	CA	PALO ALTO CA
3.	WA	SEATTLE WA
4.	CA	LOS ANGELES CA
5.	CA	LOS ANGELES CA
6.	MN	EDINA MN

sensitive. So notice in my code I have a capital letter on the start of each word – because that's how I typed it when I created those fields. If you typed yours differently, you'll need to match to how your fields are listed.

Then push the OK button at the bottom of the box.

10) Now we can do a text facet on that new CityState field and start cleaning up our inconsistent data.

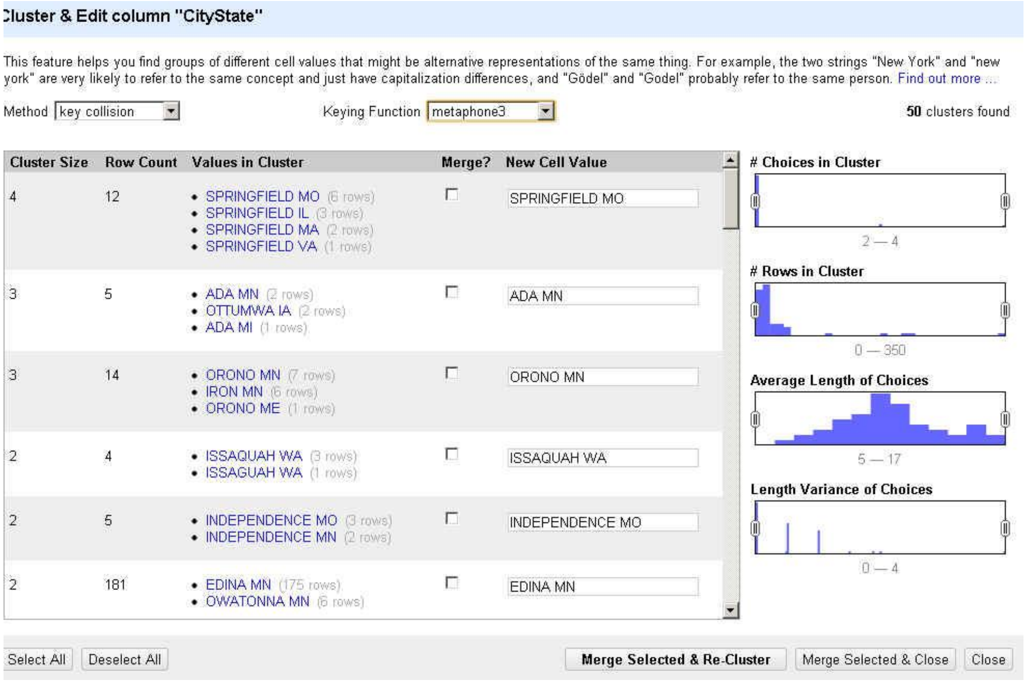
    Once you have this facet box, hit the Cluster button and it will give you several options for grouping data together.



1586 choices Sort by: **name** count

ABERDEEN SD	1
ACTON MA	1
ADA MI	1
ADA MN	2
AFTON MN	8
AGUA DULCE CA	1
AIEA HI	3
AITKIN MN	1
AKRON OH	5
ALABASTER AL	1
ALAMEDA CA	5

Notice that the Cluster & Edit box that comes up has "method" and "keying function" menus – these allow you to change which algorithm Refine will use for trying to find matches. You'll see that some return no results, others return results that seem too drastic (i.e. matching Ada, MN with Ottumwa, IA). You have to try each of the methods until you find something that seems to work on your data.



Cluster & Edit column "CityState"

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 ...

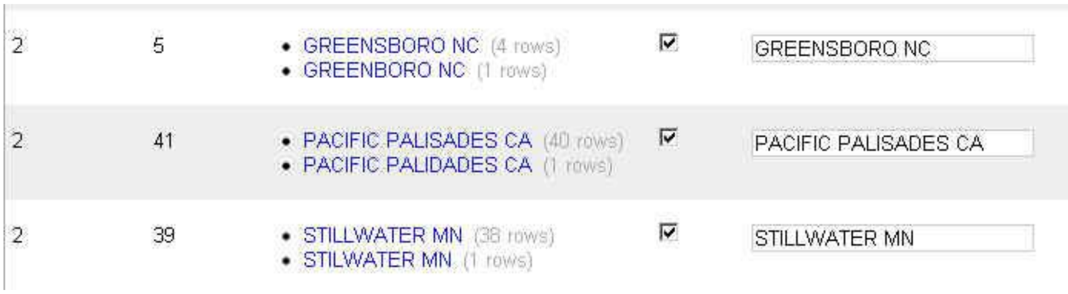
Method key collision   Keying Function metaphone3   **50** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	12	• SPRINGFIELD MO (6 rows) • SPRINGFIELD IL (3 rows) • SPRINGFIELD MA (2 rows) • SPRINGFIELD VA (1 rows)	□	SPRINGFIELD MO
3	5	• ADA MN (2 rows) • OTTUMWA IA (2 rows) • ADA MI (1 rows)	□	ADA MN
3	14	• ORONO MN (7 rows) • IRON MN (6 rows) • ORONO ME (1 rows)	□	ORONO MN
2	4	• ISSAQUAH WA (3 rows) • ISSAGUAH WA (1 rows)	□	ISSAQUAH WA
2	5	• INDEPENDENCE MO (3 rows) • INDEPENDENCE MN (2 rows)	□	INDEPENDENCE MO
2	181	• EDINA MN (175 rows) • OWATONNA MN (6 rows)	□	EDINA MN

For this one, I think the “nearest neighbor” method might be most fruitful. Notice that some of the matches are ones we don’t want to merge – i.e. Andover MN and Andover MA (unless we have reason to question that single one from Massachusetts as possibly being wrong?)

But farther down you’ll see that it found two variations of Greensboro, NC and Pacific Palisades CA.

To fix those, put a checkmark in the box next to the one(s) you want to fix and then make sure the text box on the right side has the correct spelling that you want.



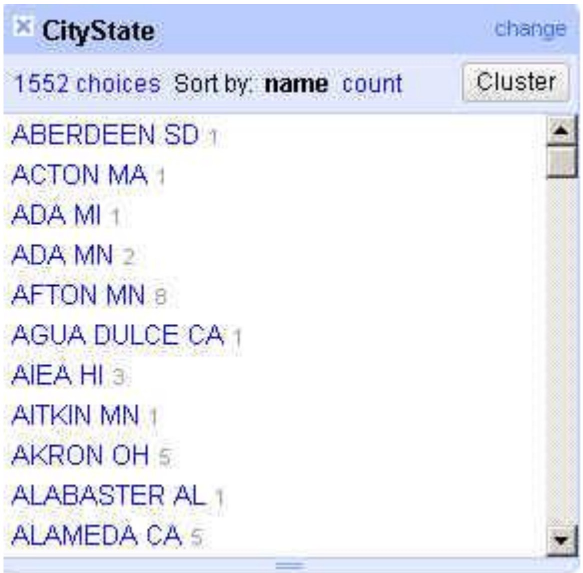
2	5	• GREENSBORO NC (4 rows) • GREENBORO NC (1 rows)	☑	GREENSBORO NC
2	41	• PACIFIC PALISADES CA (40 rows) • PACIFIC PALIDADES CA (1 rows)	☑	PACIFIC PALISADES CA
2	39	• STILLWATER MN (38 rows) • STILWATER MN (1 rows)	☑	STILLWATER MN

You might need to do some outside research or look back at your original data to make sure you’re not screwing something up. For example, the data has “Lincoln NE” (2 records) and “Lincoln NM” (1 record). I know there is a Lincoln NE, but is there a city called Lincoln in New Mexico? Might be worth checking out. Typing NM instead of NE is an easy data entry error.

Once you’ve gone through them all and settled on which ones to merge, push the “Merge Selected & Re-Cluster” button. This will give you a chance to review any of the ones that you did NOT merge. Maybe you’ll find one you missed. When all finished, hit either “Merge Selected & Close” or simply the “Close” button (if you don’t have any others to merge)

You can keep trying the various methods to keep finding matches.

When you’re done, you’ll notice that the original 1,586 variations in the CityState column have been whittled down. I got mine down to 1,552 choices.



1552 choices Sort by: **name** count

ABERDEEN SD 1
ACTON MA 1
ADA MI 1
ADA MN 2
AFTON MN 8
AGUA DULCE CA 1
AIEA HI 3
AITKIN MN 1
AKRON OH 5
ALABASTER AL 1
ALAMEDA CA 5

11) Now let’s repeat all those steps for the Occupation field.

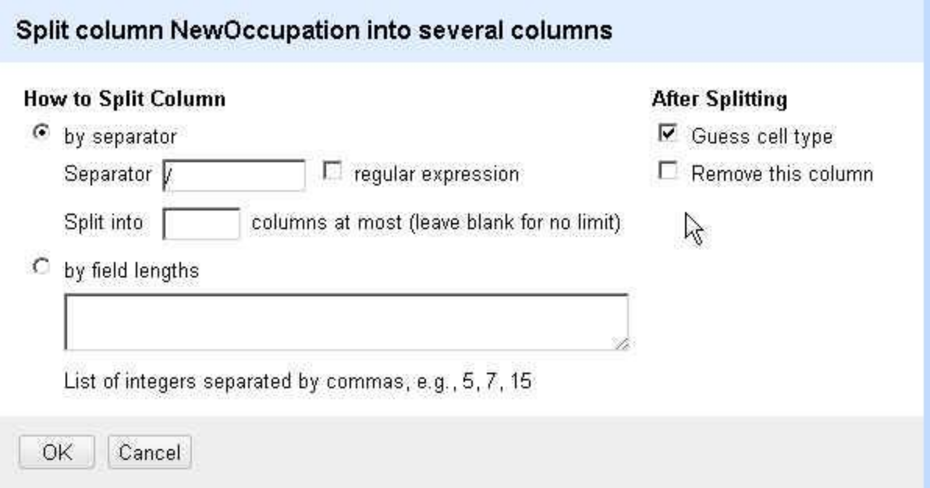
First, create a new column based on that column.

Then, do a text facet and cluster. Try out the various methods and merge and cluster as needed to clean up the data as much as possible.

12) You might find that this field should be split into “employer” and “occupation” (splitting it on the slash), and then cleaning up the two fields separately, might work better.

To do that, select the new occupation column and choose “Edit column” and “split into several columns”

In the box that comes up, put the slash “/” in the separator box and uncheck the box that says “remove this column”



Split column NewOccupation into several columns

How to Split Column

● by separator

Separator / □ regular expression

Split into columns at most (leave blank for no limit)

○ by field lengths

List of integers separated by commas, e.g., 5, 7, 15

After Splitting

☑ Guess cell type

□ Remove this column

More help/tutorials:

Using Refine to Clean Messy Data: http://www.propublica.org/nerds/item/using-google-refine-for-data-cleaning

Open Refine Cheat Sheet: https://docs.google.com/document/d/1kRoK6oDtgRO-g1KAHBMaAFPFOEGYsr5p5kpllRIn_og/edit

Refine Tutorial by David Huynh: http://davidhuynh.net/spaces/nicar2011/tutorial.pdf

MaryJo Webster
St. Paul Pioneer Press
mwebster@pioneerpress.com
651-228-5507
@mndatamine