Open Refine

for efficient data cleaning

The Goal: Get More Reliable Answers

Standardization is how we do that. Let's take a look at our PPP data (use your own copy) and find some problems of consistency. Use what you've learned from the Evaluating Data presentation.

- Where/how do similar values look different?
- Are there abbreviations or truncated data?
- Weird/missing punctuation or characters?

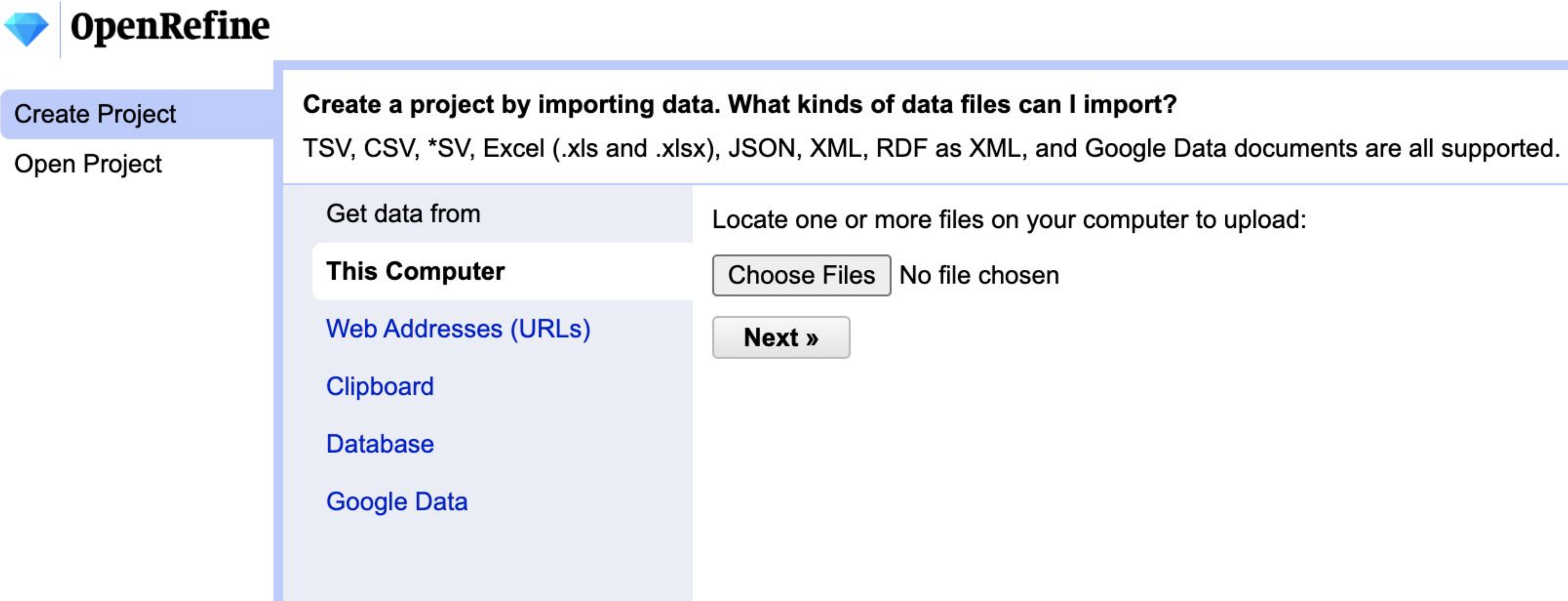
Enter some examples (and describe them) in this spreadsheet.

Get Started with OpenRefine

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/

Then download a CSV version of the PPP data from your sheet.

Import the PPP data



OpenRefine

Create Project

Open 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.

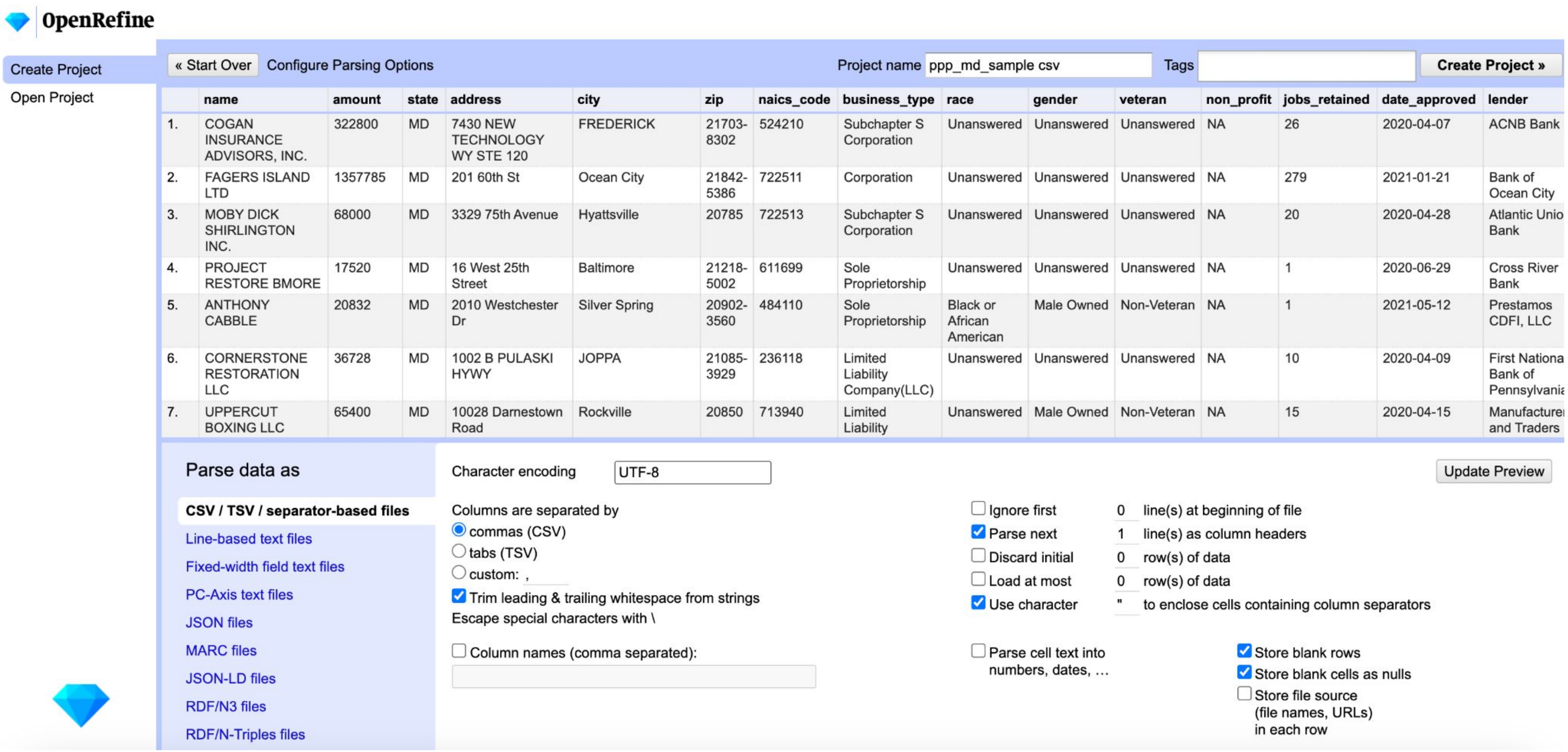
Locate one or more files on your computer to upload:

Check the Results

Once uploaded, you should see a preview. Check it out: Does it look right?

Note the options in the lower right-hand corner of the screen. Has OpenRefine selected the right ones?

How It Should Look



OpenRefine

Create Project

Open Project

« Start Over Configure Parsing Options

Project name: ppp_md_sample csv

Tags:

Create Project »

	name	amount	state	address	city	zip	naics_code	business_type	race	gender	veteran	non_profit	jobs_retained	date_approved	lender
1.	COGAN INSURANCE ADVISORS, INC.	322800	MD	7430 NEW TECHNOLOGY WY STE 120	FREDERICK	21703-8302	524210	Subchapter S Corporation	Unanswered	Unanswered	Unanswered	NA	26	2020-04-07	ACNB Bank
2.	FAGERS ISLAND LTD	1357785	MD	201 60th St	Ocean City	21842-5386	722511	Corporation	Unanswered	Unanswered	Unanswered	NA	279	2021-01-21	Bank of Ocean City
3.	MOBY DICK SHIRLINGTON INC.	68000	MD	3329 75th Avenue	Hyattsville	20785	722513	Subchapter S Corporation	Unanswered	Unanswered	Unanswered	NA	20	2020-04-28	Atlantic Unio Bank
4.	PROJECT RESTORE BMORE	17520	MD	16 West 25th Street	Baltimore	21218-5002	611699	Sole Proprietorship	Unanswered	Unanswered	Unanswered	NA	1	2020-06-29	Cross River Bank
5.	ANTHONY CABBLE	20832	MD	2010 Westchester Dr	Silver Spring	20902-3560	484110	Sole Proprietorship	Black or African American	Male Owned	Non-Veteran	NA	1	2021-05-12	Prestamos CDFI, LLC
6.	CORNERSTONE RESTORATION LLC	36728	MD	1002 B PULASKI HYWY	JOPPA	21085-3929	236118	Limited Liability Company(LLC)	Unanswered	Unanswered	Unanswered	NA	10	2020-04-09	First Nationa Bank of Pennsylvania
7.	UPPERCUT BOXING LLC	65400	MD	10028 Darnestown Road	Rockville	20850	713940	Limited Liability	Unanswered	Male Owned	Non-Veteran	NA	15	2020-04-15	Manufacture and Traders

Parse data as

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

MARC files

JSON-LD files

RDF/N3 files

RDF/N-Triples files

Character encoding: UTF-8

Update Preview

Columns are separated by

● commas (CSV)

○ tabs (TSV)

○ custom: ,

☑ Trim leading & trailing whitespace from strings

Escape special characters with \

□ Column names (comma separated):

□ 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

☑ Use character " to enclose cells containing column separators

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

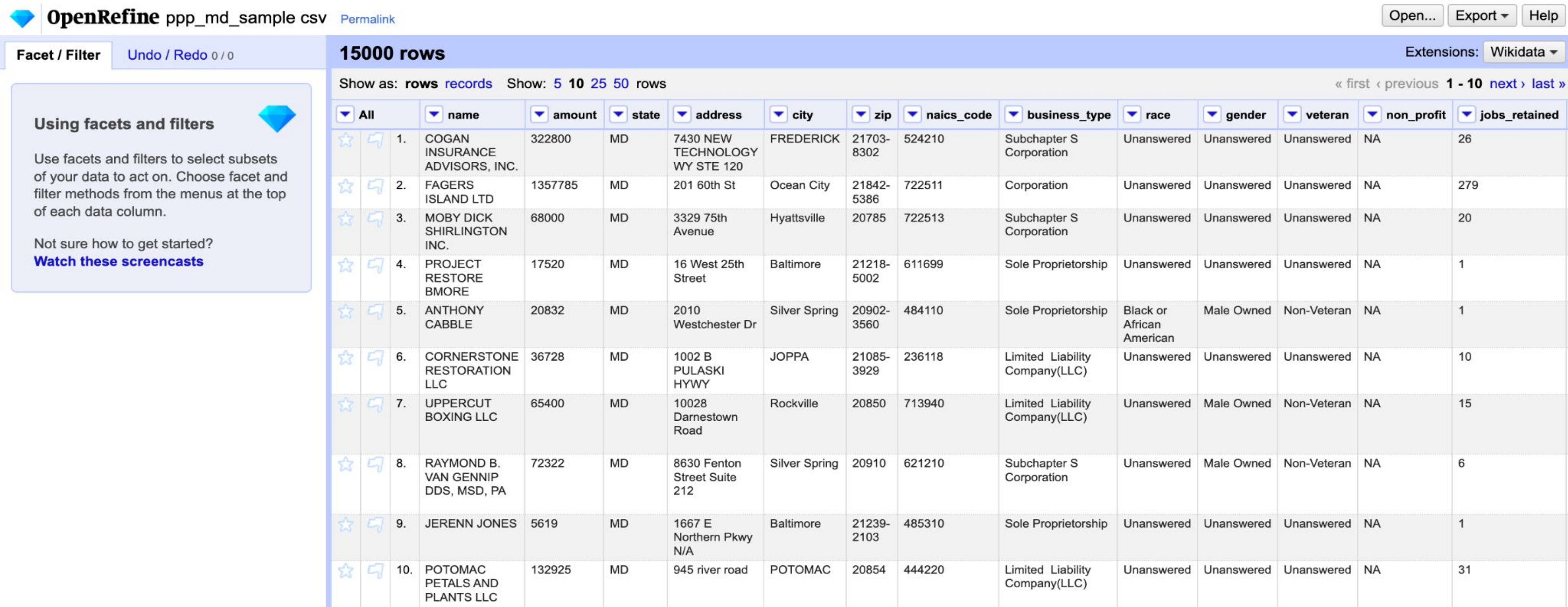
☑ Store blank rows

☑ Store blank cells as nulls

□ Store file source (file names, URLs) in each row

If It Looks Right

Click on “Create Project” and get started. It’ll look like this:



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

15000 rows

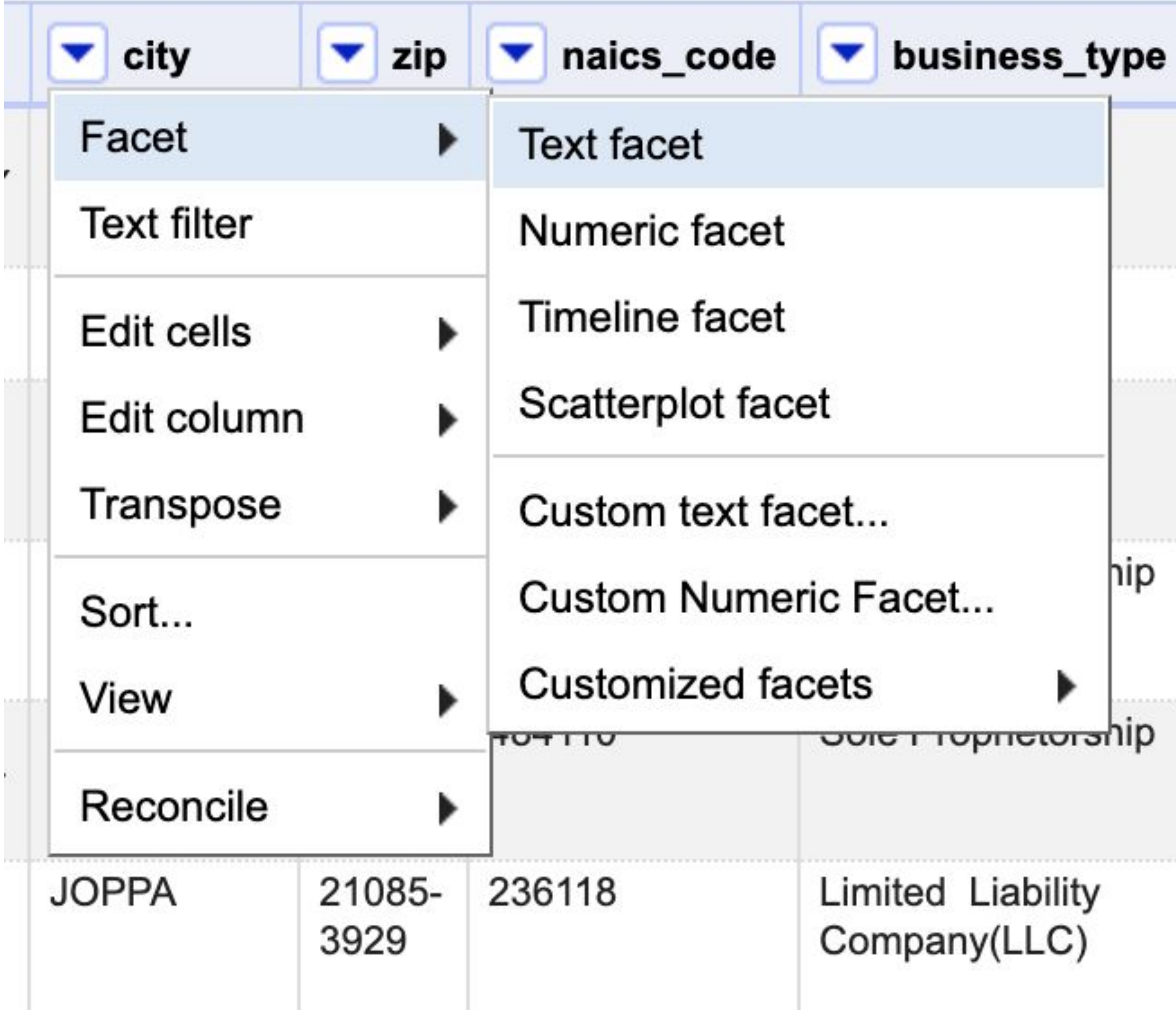
Show as: **rows** records Show: 5 **10** 25 50 rows « first ‹ previous **1 - 10** next › last »

All	name	amount	state	address	city	zip	naics_code	business_type	race	gender	veteran	non_profit	jobs_retained
1.	COGAN INSURANCE ADVISORS, INC.	322800	MD	7430 NEW TECHNOLOGY WY STE 120	FREDERICK	21703-8302	524210	Subchapter S Corporation	Unanswered	Unanswered	Unanswered	NA	26
2.	FAGERS ISLAND LTD	1357785	MD	201 60th St	Ocean City	21842-5386	722511	Corporation	Unanswered	Unanswered	Unanswered	NA	279
3.	MOBY DICK SHIRLINGTON INC.	68000	MD	3329 75th Avenue	Hyattsville	20785	722513	Subchapter S Corporation	Unanswered	Unanswered	Unanswered	NA	20
4.	PROJECT RESTORE BMORE	17520	MD	16 West 25th Street	Baltimore	21218-5002	611699	Sole Proprietorship	Unanswered	Unanswered	Unanswered	NA	1
5.	ANTHONY CABBLE	20832	MD	2010 Westchester Dr	Silver Spring	20902-3560	484110	Sole Proprietorship	Black or African American	Male Owned	Non-Veteran	NA	1
6.	CORNERSTONE RESTORATION LLC	36728	MD	1002 B PULASKI HYWY	JOPPA	21085-3929	236118	Limited Liability Company(LLC)	Unanswered	Unanswered	Unanswered	NA	10
7.	UPPERCUT BOXING LLC	65400	MD	10028 Darnestown Road	Rockville	20850	713940	Limited Liability Company(LLC)	Unanswered	Male Owned	Non-Veteran	NA	15
8.	RAYMOND B. VAN GENNIP DDS, MSD, PA	72322	MD	8630 Fenton Street Suite 212	Silver Spring	20910	621210	Subchapter S Corporation	Unanswered	Male Owned	Non-Veteran	NA	6
9.	JERENN JONES	5619	MD	1667 E Northern Pkwy N/A	Baltimore	21239-2103	485310	Sole Proprietorship	Unanswered	Unanswered	Unanswered	NA	1
10.	POTOMAC PETALS AND PLANTS LLC	132925	MD	945 river road	POTOMAC	20854	444220	Limited Liability Company(LLC)	Unanswered	Unanswered	Unanswered	NA	31

Facets

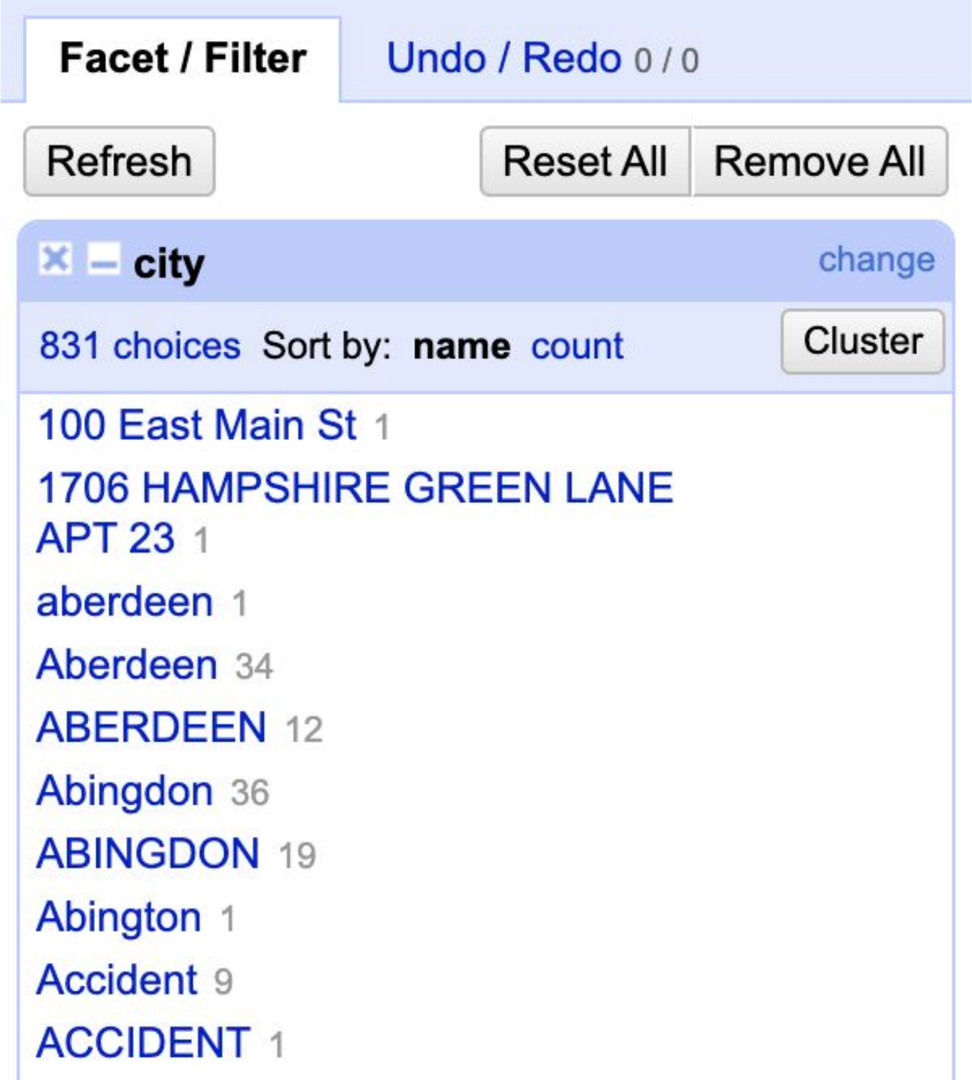
Open Refine relies on things called ‘facets’ to help us clean up data.

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



Facet Box

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.

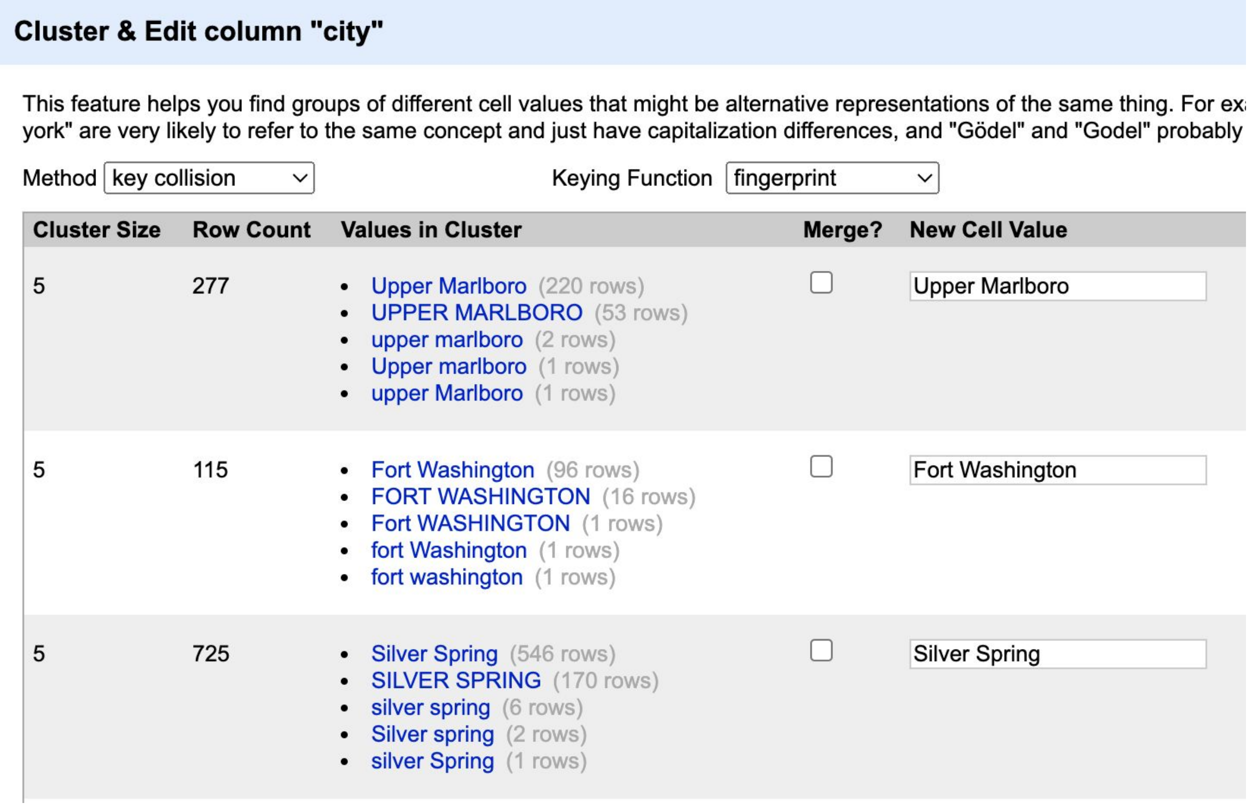


100 East Main St	1
1706 HAMPSHIRE GREEN LANE APT 23	1
aberdeen	1
Aberdeen	34
ABERDEEN	12
Abingdon	36
ABINGDON	19
Abington	1
Accident	9
ACCIDENT	1

What is Clustering?

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.

Putting Like Things Together



Cluster & Edit column "city"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For ex york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably

Method key collision

Keying Function fingerprint

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
5	277	• Upper Marlboro (220 rows) • UPPER MARLBORO (53 rows) • upper marlboro (2 rows) • Upper marlboro (1 rows) • upper Marlboro (1 rows)	□	Upper Marlboro
5	115	• Fort Washington (96 rows) • FORT WASHINGTON (16 rows) • Fort WASHINGTON (1 rows) • fort Washington (1 rows) • fort washington (1 rows)	□	Fort Washington
5	725	• Silver Spring (546 rows) • SILVER SPRING (170 rows) • silver spring (6 rows) • Silver spring (2 rows) • silver Spring (1 rows)	□	Silver Spring

Don’t Trust, Verify

Is this a match we want? Probably, and we can check. 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 cities have very similar zip-codes. They’re probably a match.

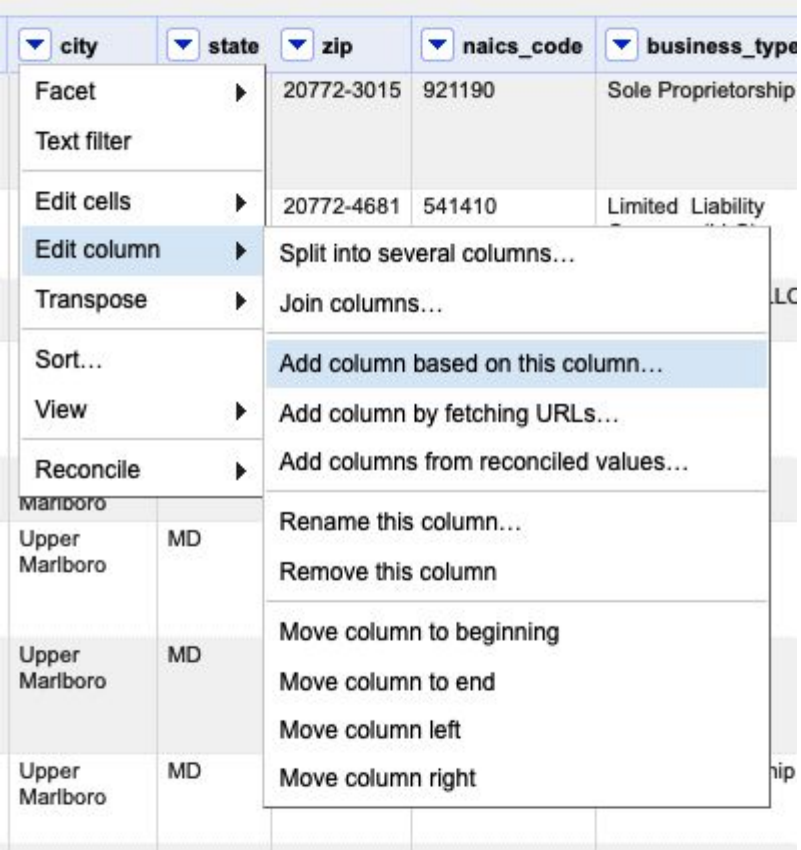
Close the additional window to return to the complete data.

Make a Copy First!

Remember how Sophie said “make a copy of the original data”? We can do that with columns, too, so we can see what edits we make.

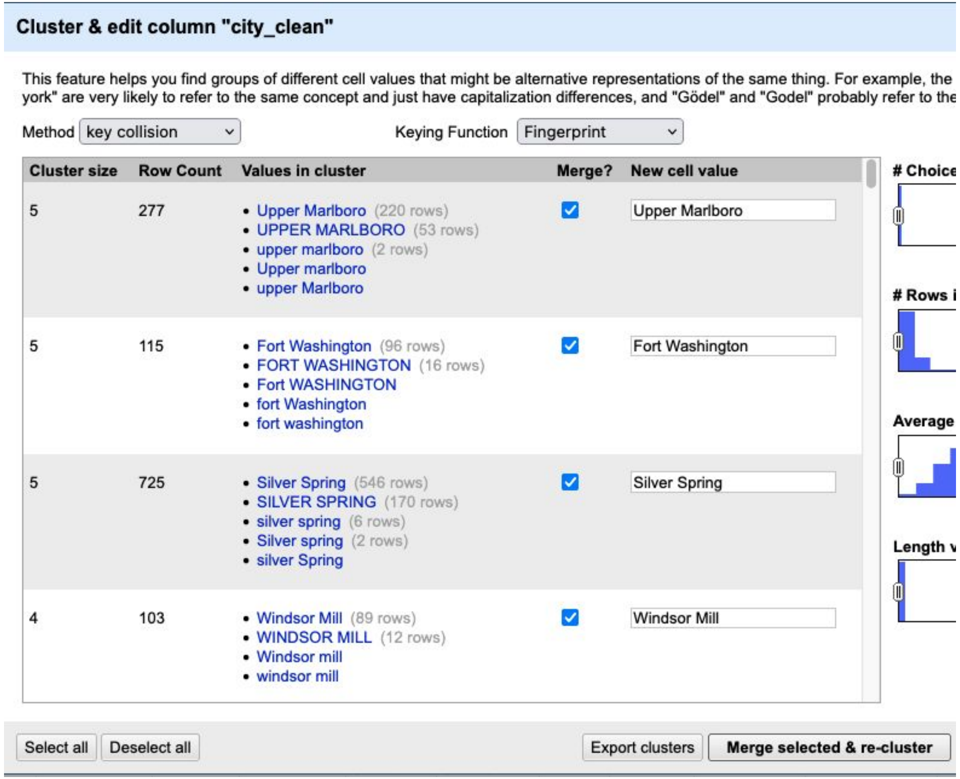
Go to the “city” column and choose Edit column -> Add column based on this column...

Then give the new column a name like “city_clean” and cluster on that column.



Everyday I’m Clustering

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.



Cluster & edit column "city_clean"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the

Method: key collision

Keying Function: Fingerprint

Cluster size	Row Count	Values in cluster	Merge?	New cell value
5	277	• Upper Marlboro (220 rows) • UPPER MARLBORO (53 rows) • upper marlboro (2 rows) • Upper marlboro • upper Marlboro	☑	Upper Marlboro
5	115	• Fort Washington (96 rows) • FORT WASHINGTON (16 rows) • Fort WASHINGTON • fort Washington • fort washington	☑	Fort Washington
5	725	• Silver Spring (546 rows) • SILVER SPRING (170 rows) • silver spring (6 rows) • Silver spring (2 rows) • silver Spring	☑	Silver Spring
4	103	• Windsor Mill (89 rows) • WINDSOR MILL (12 rows) • Windsor mill • windsor mill	☑	Windsor Mill

But Wait, There’s More!

OpenRefine also has some other useful functions you can explore.

- Splitting cells into multiple columns
- Trimming unneeded white space
- Changing column data types
- Changing cases