Open Refine

For efficient data cleaning

Step 1

Open the Orders.csv text file using a text editor and take a look. Does everything make sense? Are there going to be any special formatting issues? What else do you see?

Okay now open the file in Google Sheets.

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 of the fields.

- Names
- Addresses
- Cities

Example:

ABRON CHARLES	317 W 101ST PL
ABRON CHARLES	317 W 101ST PLACE

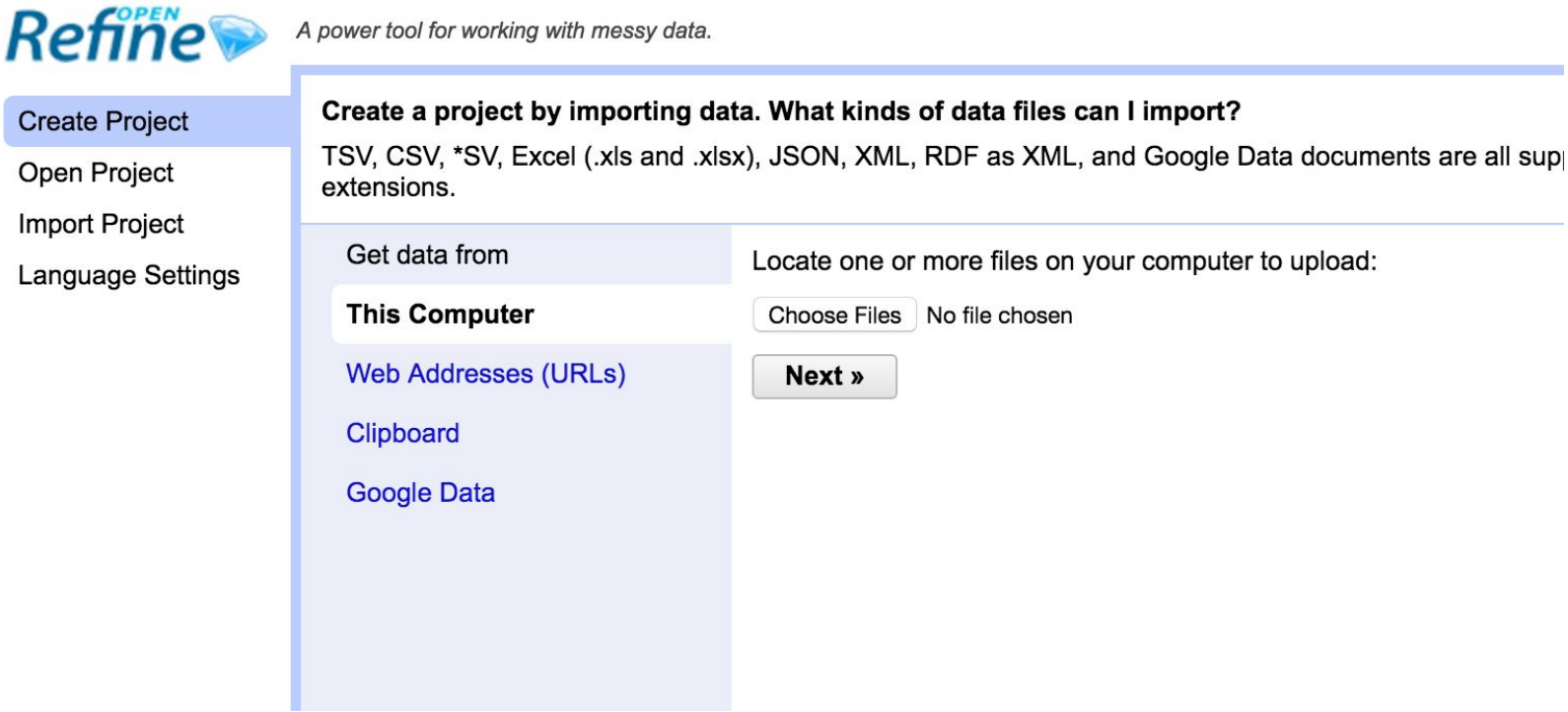
Step 4

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/

(Refine can sometimes be a bit buggy.)

Step 5

Import the Orders.csv dataset.

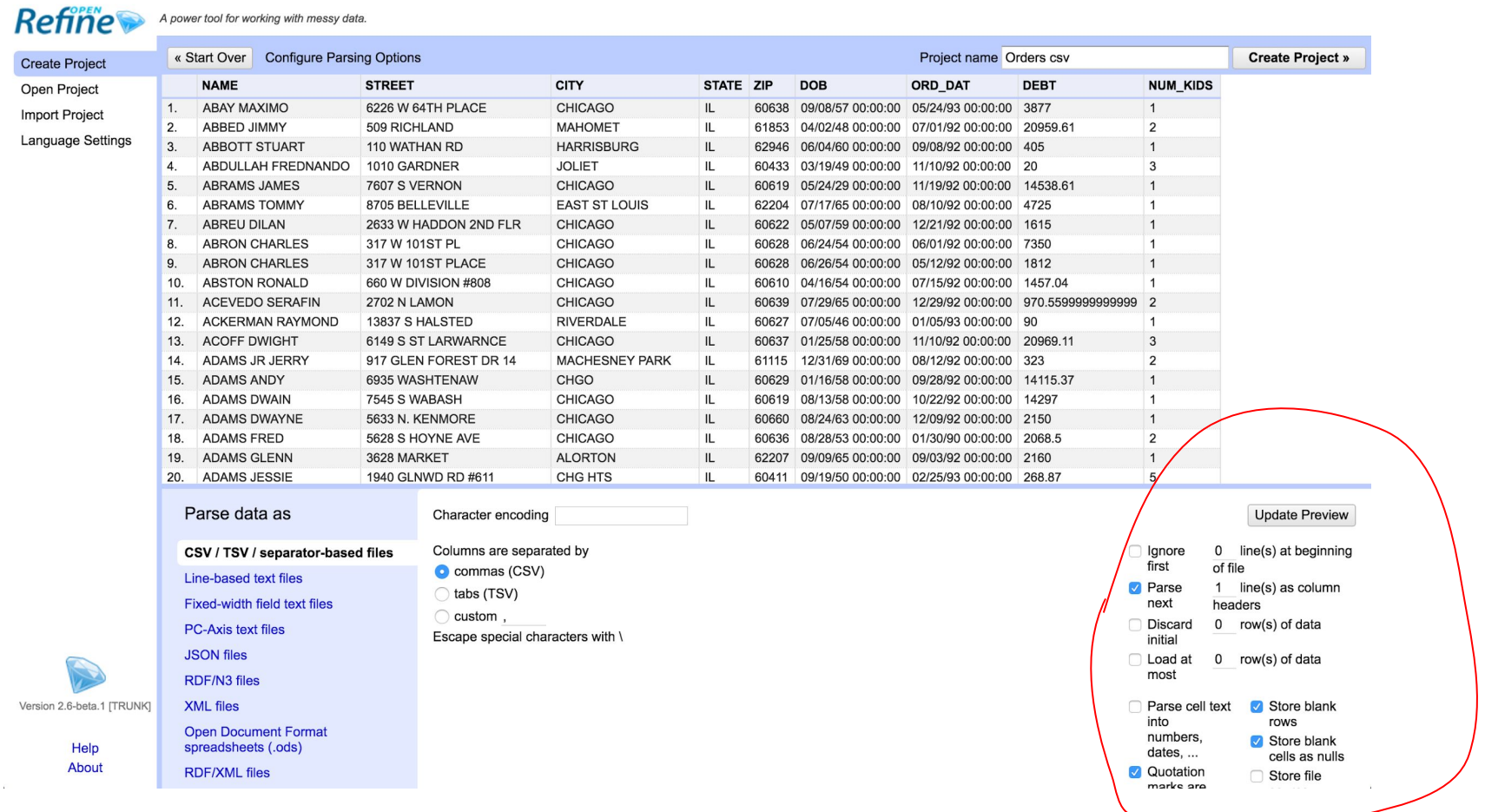


Step 6

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 Open Refine selected the right ones?

Step 6



	NAME	STREET	CITY	STATE	ZIP	DOB	ORD_DAT	DEBT	NUM_KIDS
1.	ABAY MAXIMO	6226 W 64TH PLACE	CHICAGO	IL	60638	09/08/57 00:00:00	05/24/93 00:00:00	3877	1
2.	ABBED JIMMY	509 RICHLAND	MAHOMET	IL	61853	04/02/48 00:00:00	07/01/92 00:00:00	20959.61	2
3.	ABBOTT STUART	110 WATHAN RD	HARRISBURG	IL	62946	06/04/60 00:00:00	09/08/92 00:00:00	405	1
4.	ABDULLAH FREDNANDO	1010 GARDNER	JOLIET	IL	60433	03/19/49 00:00:00	11/10/92 00:00:00	20	3
5.	ABRAMS JAMES	7607 S VERNON	CHICAGO	IL	60619	05/24/29 00:00:00	11/19/92 00:00:00	14538.61	1
6.	ABRAMS TOMMY	8705 BELLEVILLE	EAST ST LOUIS	IL	62204	07/17/65 00:00:00	08/10/92 00:00:00	4725	1
7.	ABREU DILAN	2633 W HADDON 2ND FLR	CHICAGO	IL	60622	05/07/59 00:00:00	12/21/92 00:00:00	1615	1
8.	ABRON CHARLES	317 W 101ST PL	CHICAGO	IL	60628	06/24/54 00:00:00	06/01/92 00:00:00	7350	1
9.	ABRON CHARLES	317 W 101ST PLACE	CHICAGO	IL	60628	06/26/54 00:00:00	05/12/92 00:00:00	1812	1
10.	ABSTON RONALD	660 W DIVISION #808	CHICAGO	IL	60610	04/16/54 00:00:00	07/15/92 00:00:00	1457.04	1
11.	ACEVEDO SERAFIN	2702 N LAMON	CHICAGO	IL	60639	07/29/65 00:00:00	12/29/92 00:00:00	970.5599999999999	2
12.	ACKERMAN RAYMOND	13837 S HALSTED	RIVERDALE	IL	60627	07/05/46 00:00:00	01/05/93 00:00:00	90	1
13.	ACOFF DWIGHT	6149 S ST LARWARNCE	CHICAGO	IL	60637	01/25/58 00:00:00	11/10/92 00:00:00	20969.11	3
14.	ADAMS JR JERRY	917 GLEN FOREST DR 14	MACHESNEY PARK	IL	61115	12/31/69 00:00:00	08/12/92 00:00:00	323	2
15.	ADAMS ANDY	6935 WASHTENAW	CHGO	IL	60629	01/16/58 00:00:00	09/28/92 00:00:00	14115.37	1
16.	ADAMS DWAIN	7545 S WABASH	CHICAGO	IL	60619	08/13/58 00:00:00	10/22/92 00:00:00	14297	1
17.	ADAMS DWAYNE	5633 N. KENMORE	CHICAGO	IL	60660	08/24/63 00:00:00	12/09/92 00:00:00	2150	1
18.	ADAMS FRED	5628 S HOYNE AVE	CHICAGO	IL	60636	08/28/53 00:00:00	01/30/90 00:00:00	2068.5	2
19.	ADAMS GLENN	3628 MARKET	ALORTON	IL	62207	09/09/65 00:00:00	09/03/92 00:00:00	2160	1
20.	ADAMS JESSIE	1940 GLNWD RD #611	CHG HTS	IL	60411	09/19/50 00:00:00	02/25/93 00:00:00	268.87	5

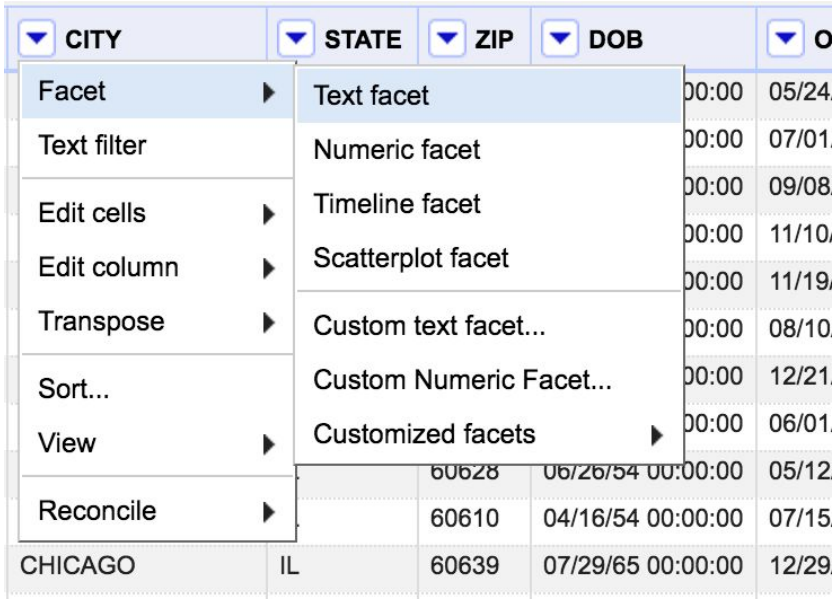
Step 7

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

Step 8

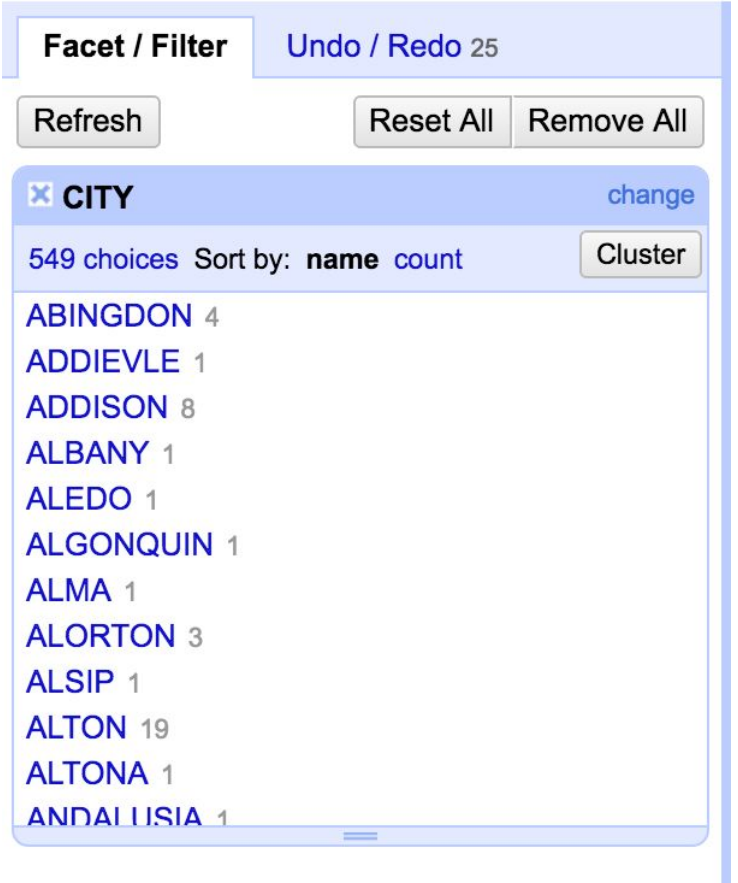
Open Refine relies on things called ‘facets’ to help us clean up data.

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



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.

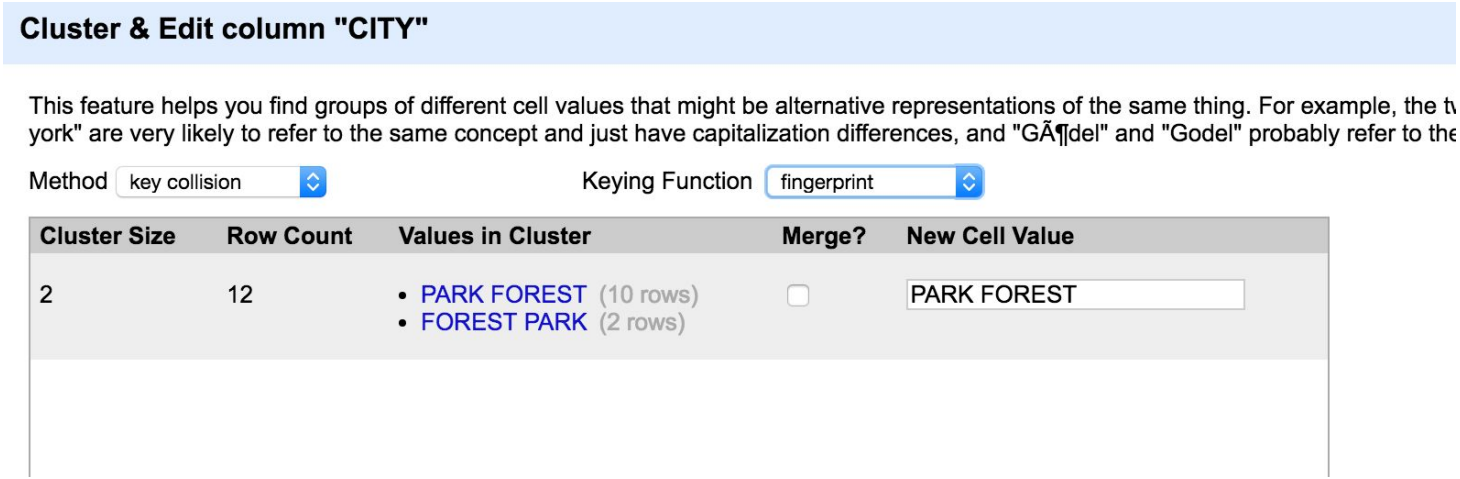


Step 10

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 10

The default clustering method doesn’t get us that many matches.



Step 10

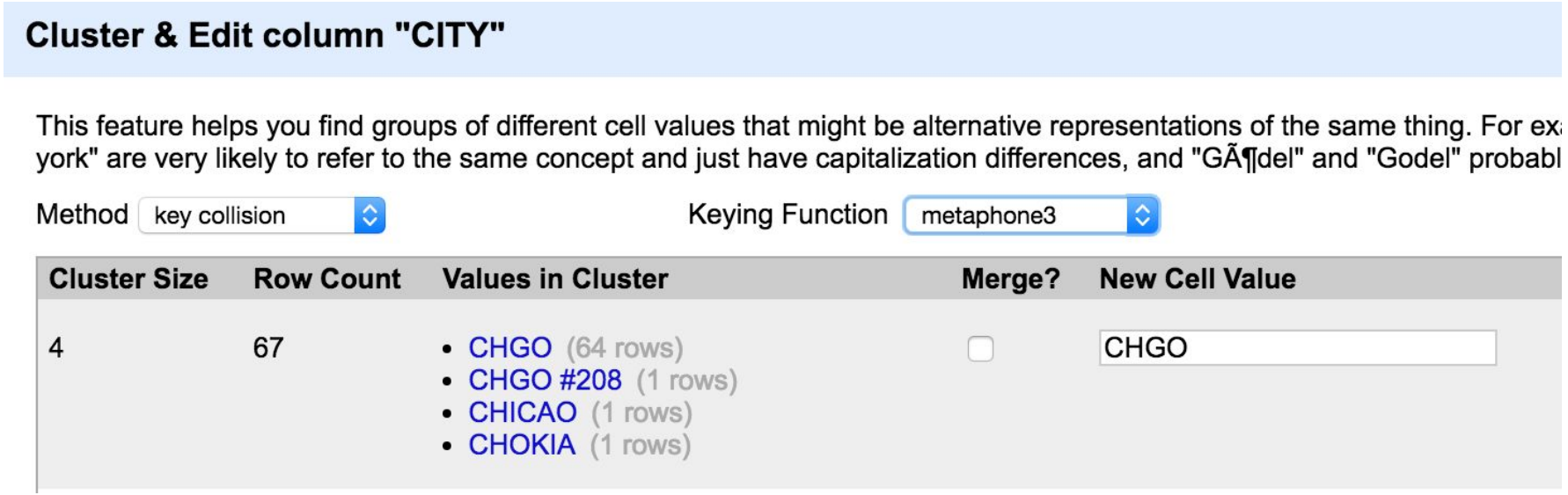
Is this a match we want? Hard to say. If you hover, you get the option to ‘Browse this Cluster’. Click that. A new window will pop open showing just the rows that would be included in that potential cluster.

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 clustering method:



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

Refine also has some other useful functions you can explore.

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