ADTA 5240: Data Harvesting
Cleaning & Wrangling Data with OpenRefine Tutorial

1. Download and Installation

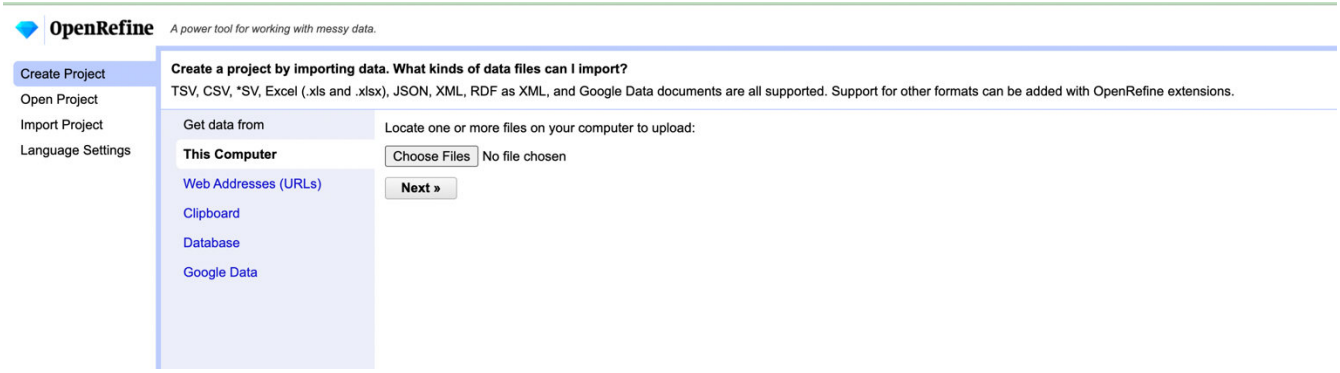
A. You may download OpenRefine to your specific computer. Follow the directions for the download process here: **https://docs.openrefine.org/manual/installing**

B. If you would like to use another browser instead, start OpenRefine and then point your chosen browser at the home screen: http://127.0.0.1:3333/.

OpenRefine works best on browsers based on Webkit, such as:

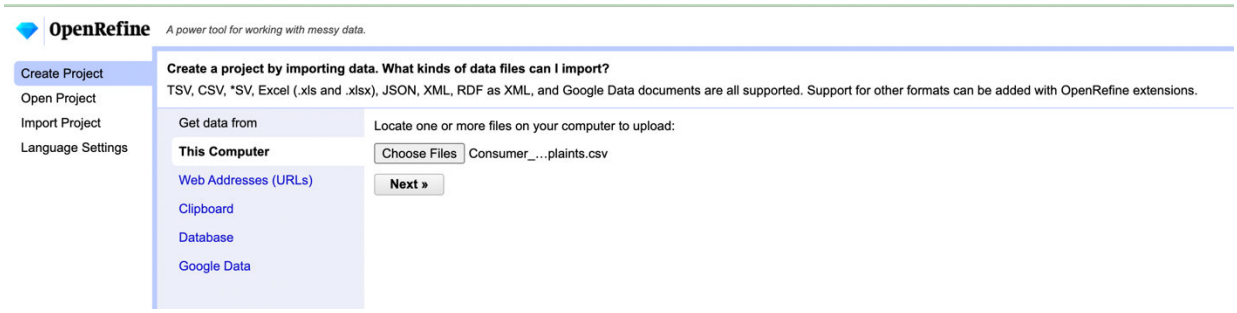
- Google Chrome
- Chromium
- Opera
- Microsoft Edge

2. OpenRefine will open, and a web-browser based GUI pops up



3. Create a New Project & Upload Data

To **create** a new project, you need to **upload** a file. **Download** the Consumer_Complaints.csv file from Canvas. **Click** Choose Files →**Browse** for the file: Consumer_Complaints.csv

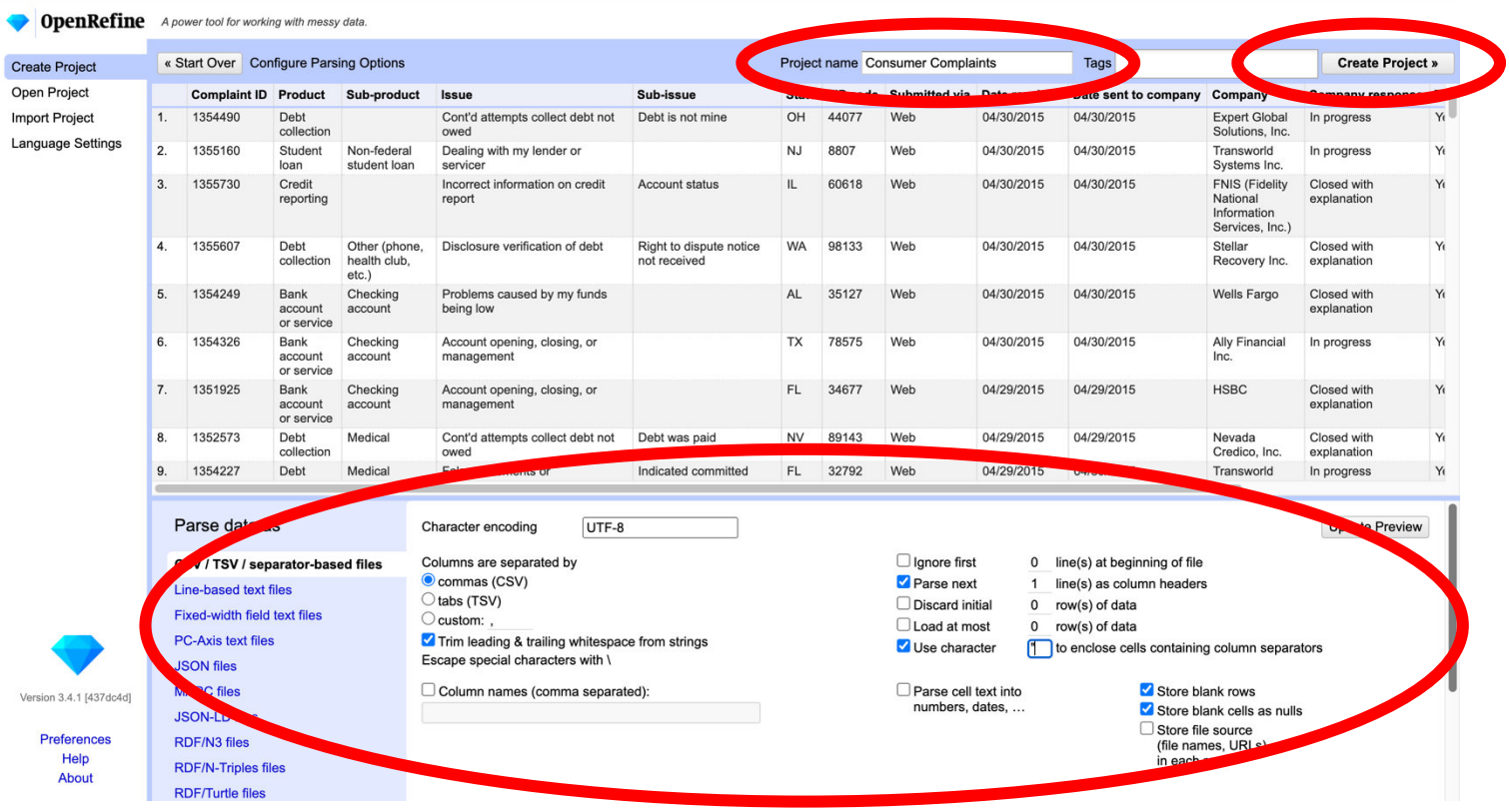


Click Next → the data set is uploaded
Enter the project name: Consumer Complaints (in Project name text field: on the left top corner)

The following details are checked (these should be the default settings but double check):

- OpenRefine detected that the data file is in CSV format
- Line (Row) 1 is a header line
- Quotation marks are used to enclose cells containing column separators
- Store blank rows
- Store blank cells as nulls

Click Create Project

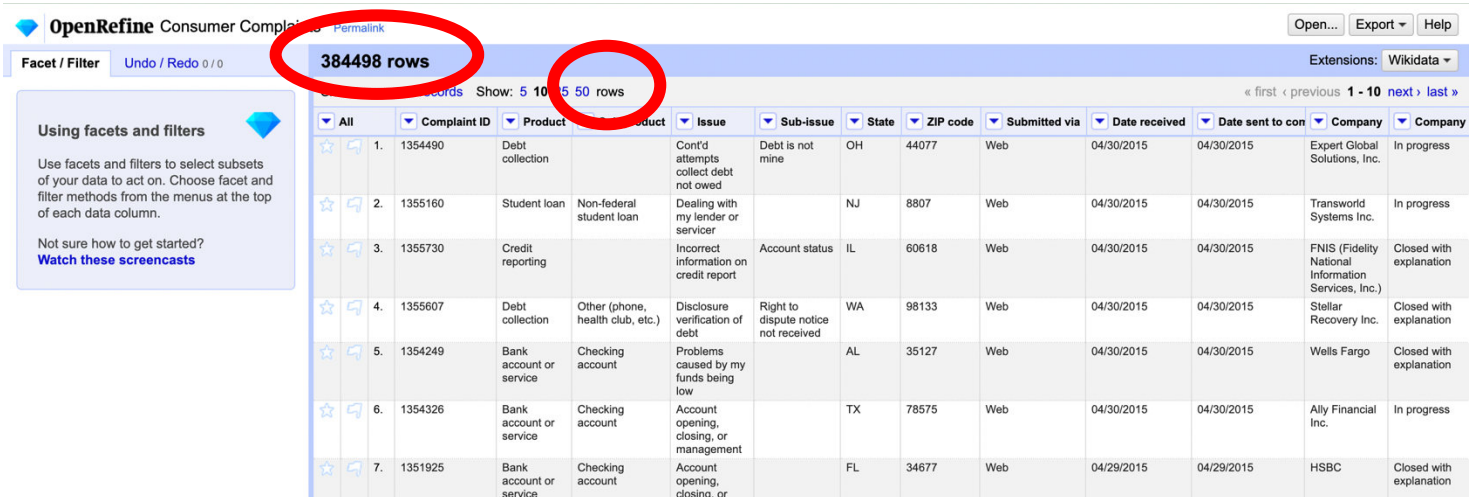


You see 384,498 rows are loaded.

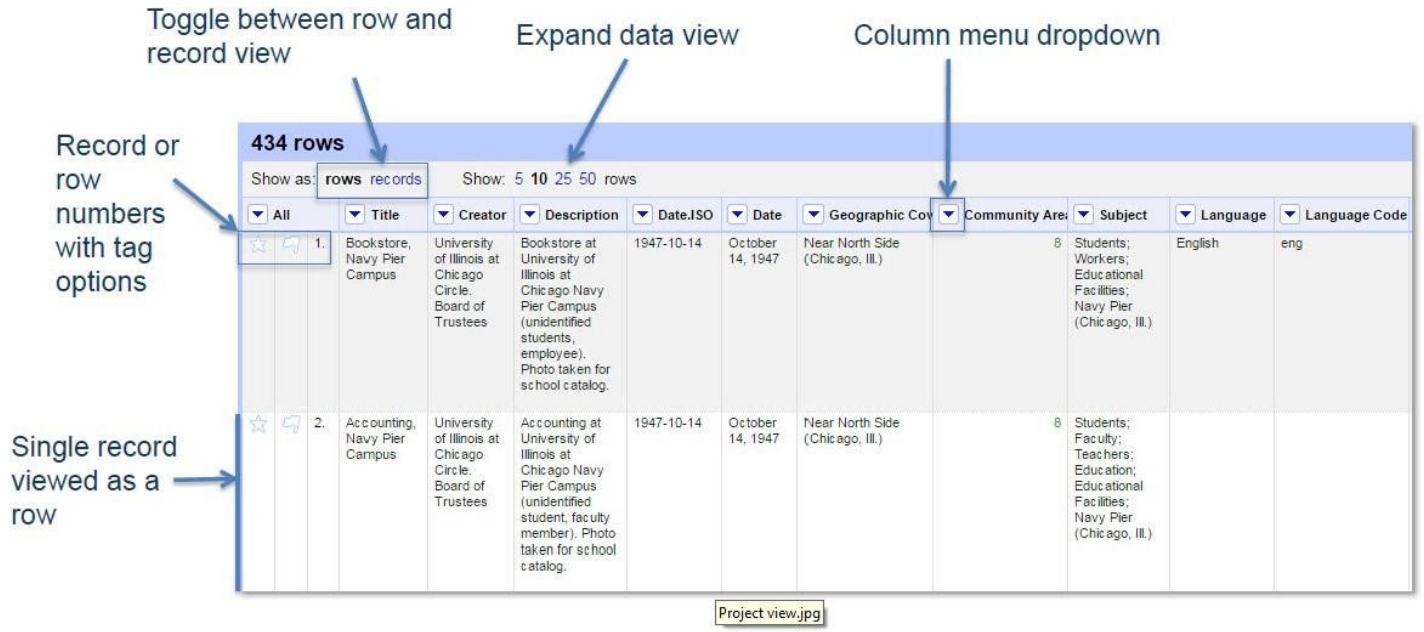
Click on "50" → OpenRefine shows 50 rows at once.

To view rows (records), i.e. 51 and further:

Click Next. You can also click on Previous to go backward.



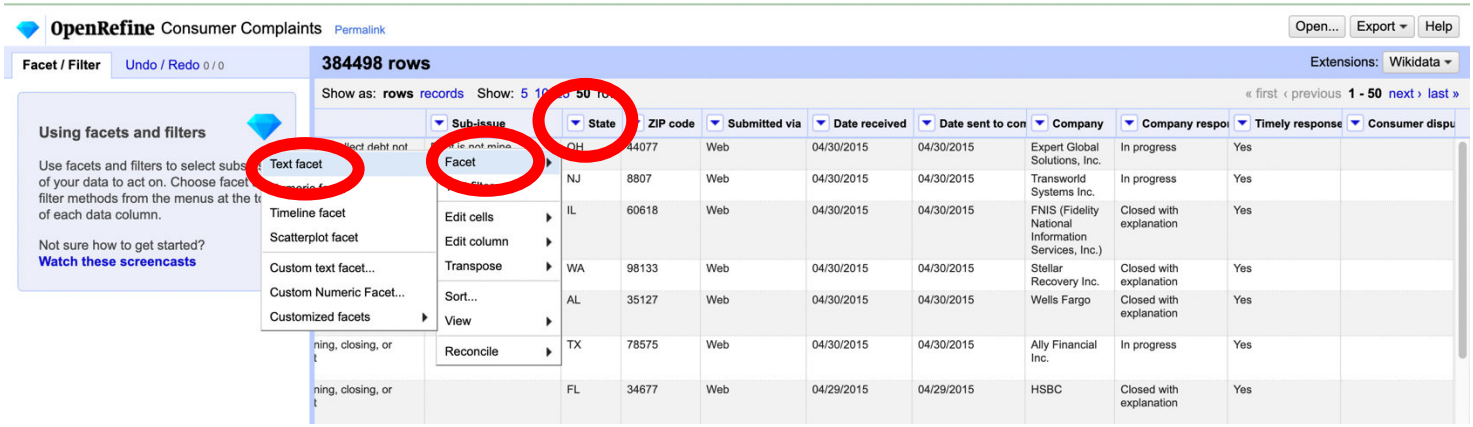
This is a screenshot of another project but gives you a good explanation of the interface.



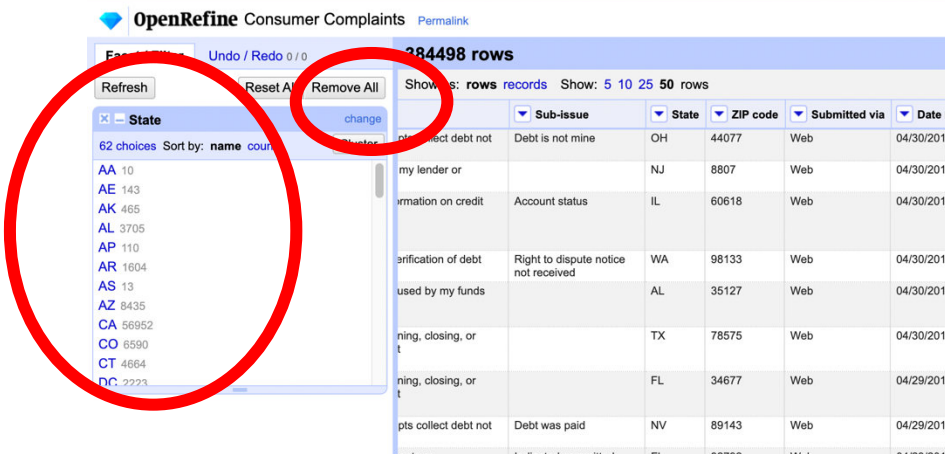
4. Check States with Text Facets

Analyzing data: We will use “Text Facet to display the number of occurrences of each unique value in a column. This is like filter in Excel.

Click on the drop-down menu arrow of State to open the menu. **Click** on Facet. **Click** on Text Facet.



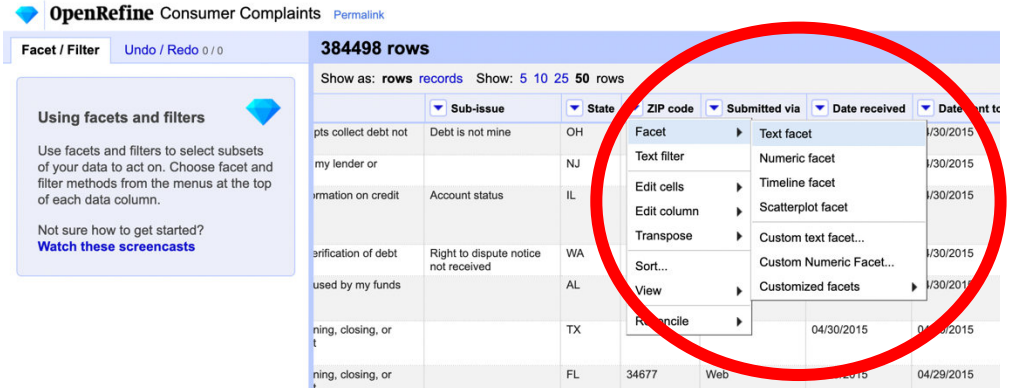
You see 62 states (51 states and other US territories) are listed in the panel to the left.
Click Remove All to clear the result panel to the left.

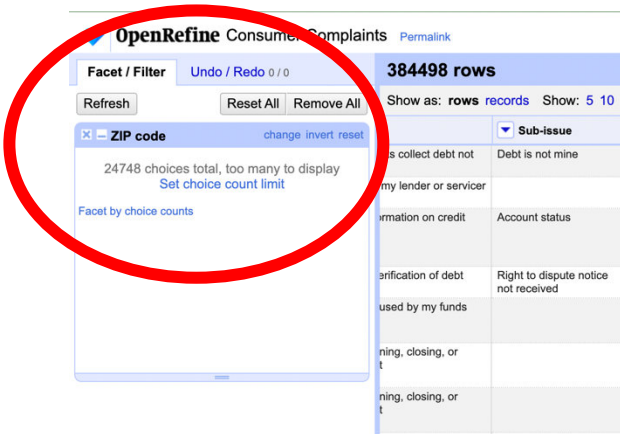


5. Wrangling/Munging – Transforming Data: Check Zip code

5.1 Text Facet for Zip Code

Click on the drop-down menu arrow of Zip code to open the menu. **Click** to select Text facet



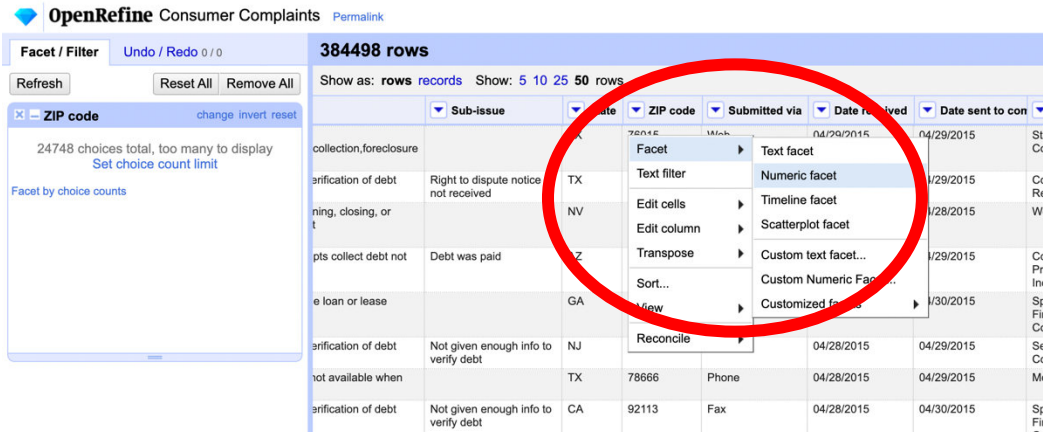


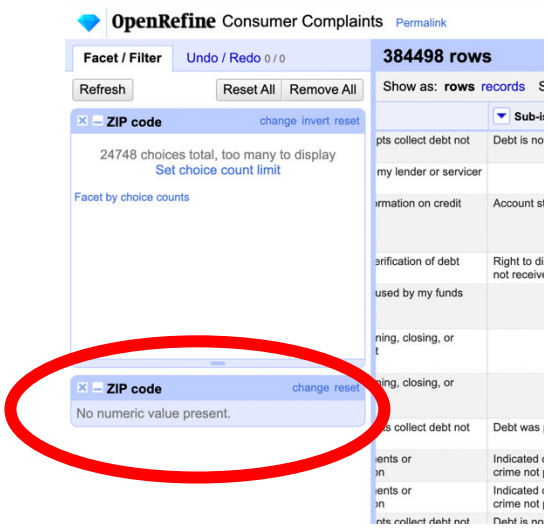
The result shows that there are 24,748 zip codes in the dataset, and this is too many zip codes to be listed. Let's skim over the values of the zip codes, some of them seem:

- Very short 1, 2, and 3 digits
- Short of 5 digits, like in Row 2: 8807, Row 70: 5468.
- Some of them are missing, i.e., blank, like in Row 57

5.2 Numeric Facet

Click on the drop-down menu arrow of Zip code > **Click** to select Numeric facet





There is no numeric zip code in the data set, i.e., all of them are included as text values (24748 zip codes)

5.3 Transform Zip code from Text to Numeric (Wrangling Data)

5.3.1 Overview: What is Wrangling/Munging Data

Definition #1:
Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. https://www.datawatch.com/what-is-data-wrangling/

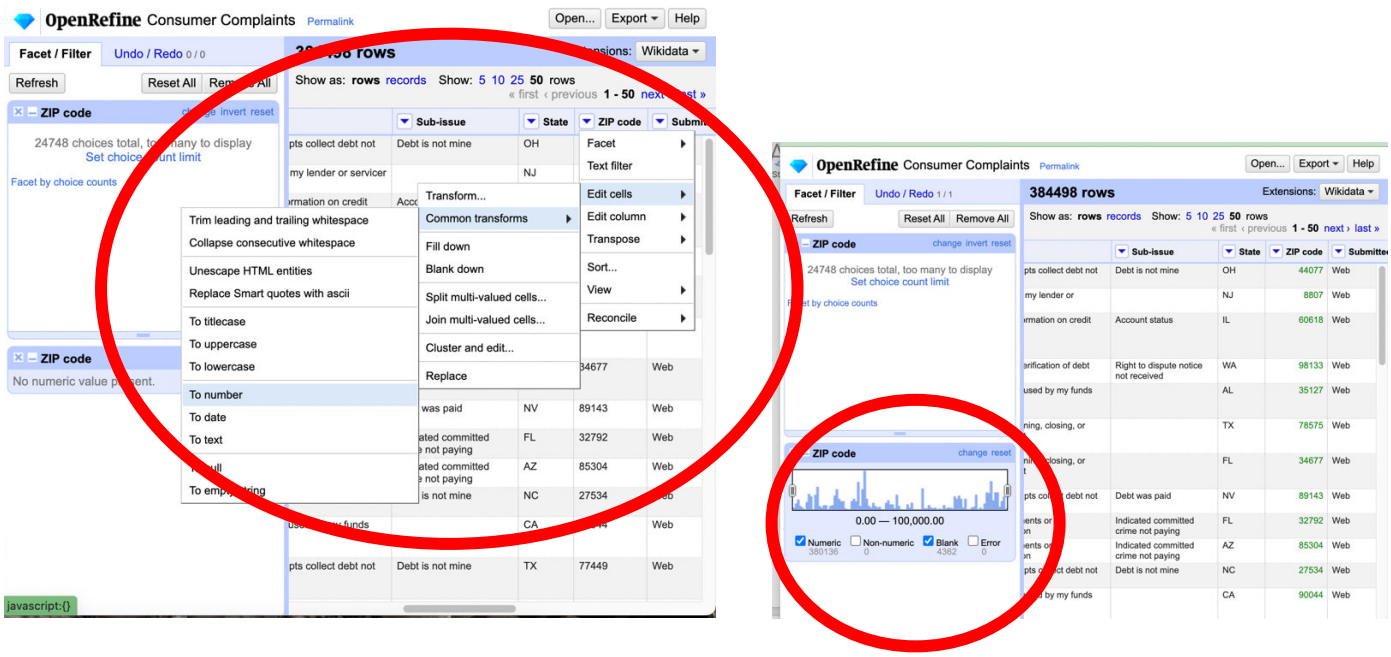
Definition #2:
Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.
https://en.wikipedia.org/wiki/Data_wrangling

Definition #3:
Over the past few years, data wrangling (also known as data preparation) has emerged as a fast-growing space within the analytics industry.
Data wrangling (NEW) =/= ETL (Extract, Transform, Load) (TRADITIONAL) for 3 reasons:

- The users are different
- The data are different
- The user case is different

5.3.2 How To

Click on the drop-down menu arrow of Zip code to open the menu
Click to select Edit cells → **Click** → Common transforms → **Click** To number



The results show that 380136 zip codes text values have been transformed into numeric values
The results also show that 4362 zip codes are missing, i.e., blank

6. Cleansing/Wrangling Data: Handling Missing Zip codes

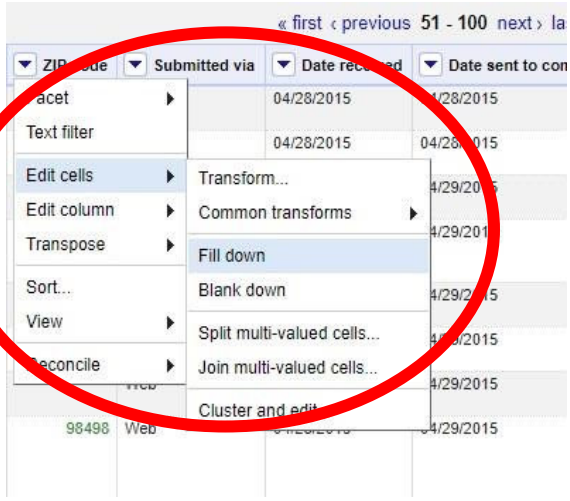
6.1 Fill down the missing value

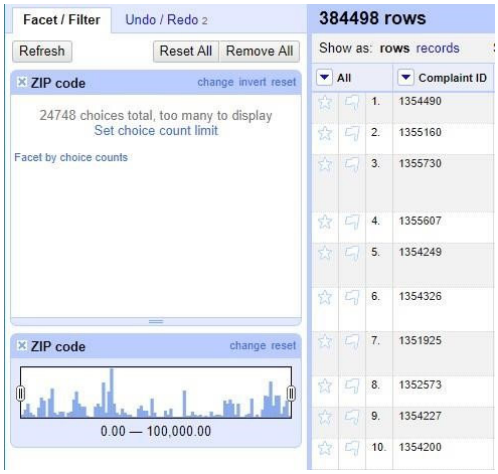
Fill down: One way of filling in missing values is to take the previous value, and use that to set subsequent empty cells (of the same field/column)

Let's pay attention to Row 57 with a missing zip code value:

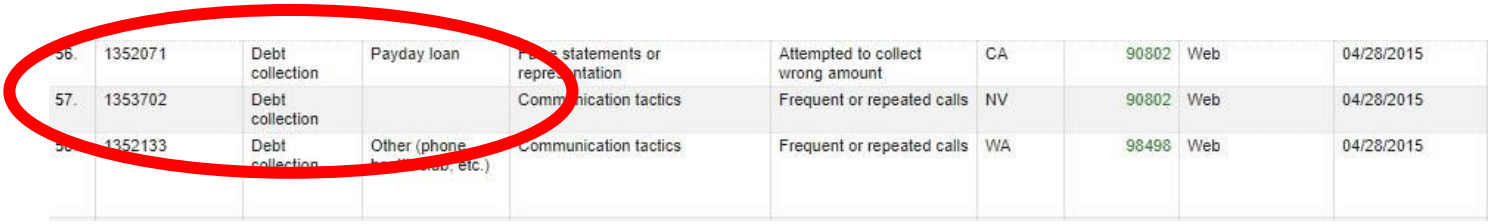
56.	1352071	Debt collection	Payday loan	False statements or representation	Attempted to collect wrong amount	CA	90802	Web
57.	1353702	Debt collection		Communication tactics	Frequent or repeated calls	NV		Web
58.	1352133	Debt collection	Other (phone, health club, etc.)	Communication tactics	Frequent or repeated calls	WA	98498	Web

Click on the drop-down menu arrow of Zip code to open the menu. **Click** to select Edit cells → **Fill** down





The results show that there are no missing zip codes now, i.e., all the missing values have been filled down. Let's pay attention to the zip code value in Row 57 now:



56.	1352071	Debt collection	Payday loan	statements or representation	Attempted to collect wrong amount	CA	90802	Web	04/28/2015
57.	1353702	Debt collection		Communication tactics	Frequent or repeated calls	NV	90802	Web	04/28/2015
58.	1352133	Debt collection	Other (phone, etc.)	Communication tactics	Frequent or repeated calls	WA	98498	Web	04/28/2015

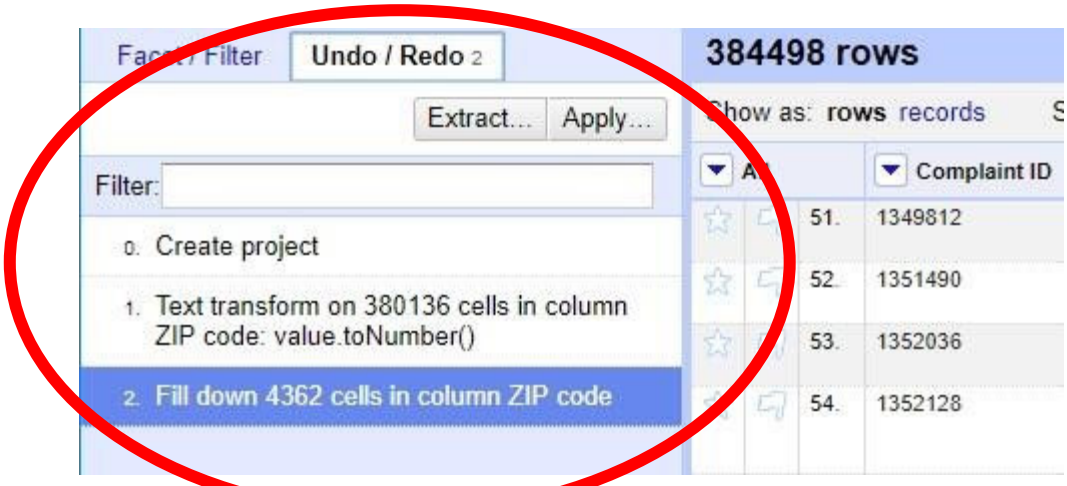
In Row 57, the zip code value is 90082, filled down from Row 56. But

- In Row 56, the state is CA → the zip code 90082 is in CA
- In Row 57, the state is NV, not CA
- Cannot fill down to handle the missing values of zip code
- MUST rescind the wrangling data

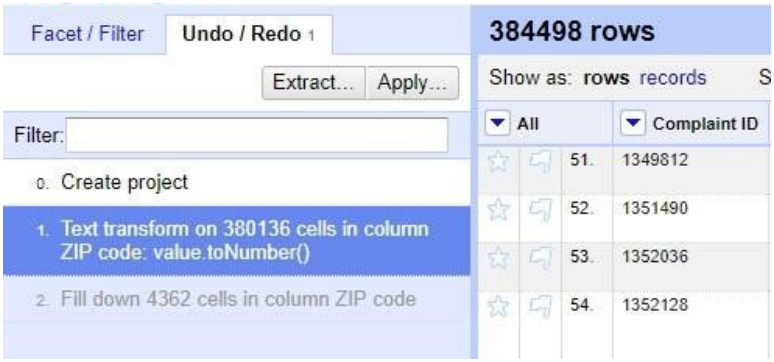
6.1.1 Rescind the wrangling data

OpenRefine has very advanced UNDO functionalities

Click Undo



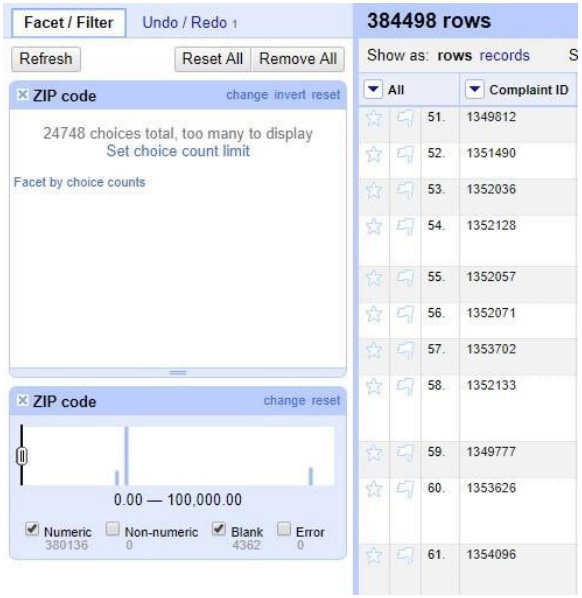
UNDO lists three level of activities – 0, 1, 2 – that can be rolled back.
The last level "Fill down 4362 cells in column ZIP code" is highlighted.
Click to select "Text transform on …" as the latest activity,
(i.e., roll back from level 2 to level 1 & remove all the impacts of level 2 action)



Let's pay attention to the zip code value in Row 57 again:

56.		Debt collection	Payday loan	False statements or representation	Attempted to collect wrong amount	CA	90802	Web
57.	1353702	Debt collection		Communication tactics	Frequent or repeated calls	NV		Web
58.		Debt collection	Other (phone, health club, etc.)	Communication tactics	Frequent or repeated calls	WA	98498	Web

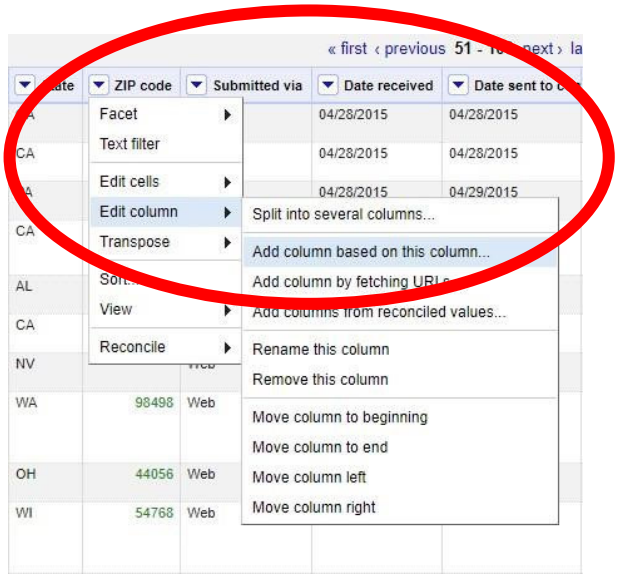
The filled down value has been undone → the zip code value is missing again now.
Also check the facet again to be sure that all the filled down zip code values have been undone.

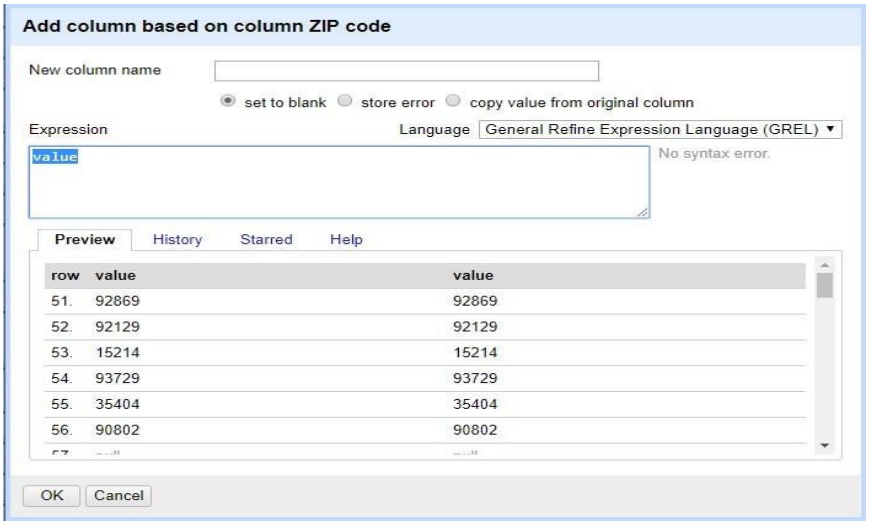


The numeric facet shows 4362 missing zip code values again.

6.2 Handling Missing Zip codes: Create a New Column

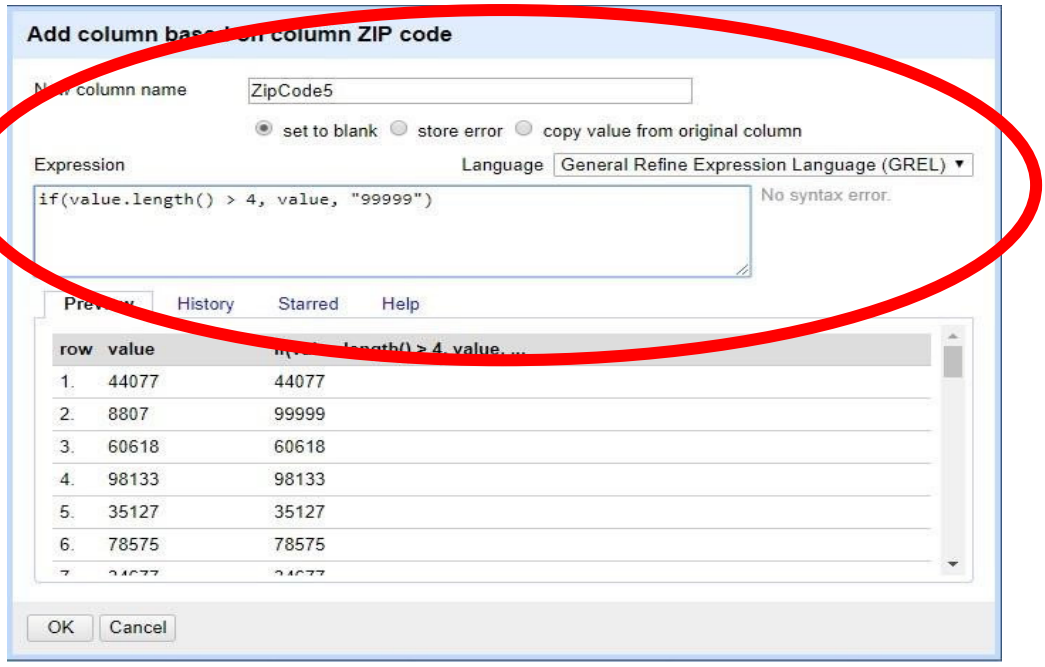
Click on the drop-down menu arrow of Zip code to open the menu.
Click to select Edit column → **Add** column based on this column





- It is assumed that a valid zip code must have 5 digits.
- We want to create a new column based on the ZIP code column: if the current zip code value is valid, keep it; otherwise – either missing or not 5 digits – the cells are filled with the string "99999". Technically, a zip code with 4 digits may be valid
- To perform this cleansing/wrangling/transforming data, i.e., creating new column, we use General Refine Expression Language, a scripting language that can be used with OpenRefine

- **Enter** the new column name: ZipCode5
- **Enter**: if(value.length() > 4, value, "99999") for expression. Sometimes when copying and pasting the quotation marks do not copy well. If you get a parsing error, retype the quotation marks.
- Then **click** OK



All		Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	ZipCode5	Submitted via
	1.	1354490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH	44077	44077	Web
	2.	1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ	8807	99999	Web
	3.	1355730	Credit reporting		Incorrect information on credit report	Account status	IL	60618	60618	Web
	4.	1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133	98133	Web
	5.	1354249	Bank account or service	Checking account	Problems caused by my funds being low		AL	35127	35127	Web

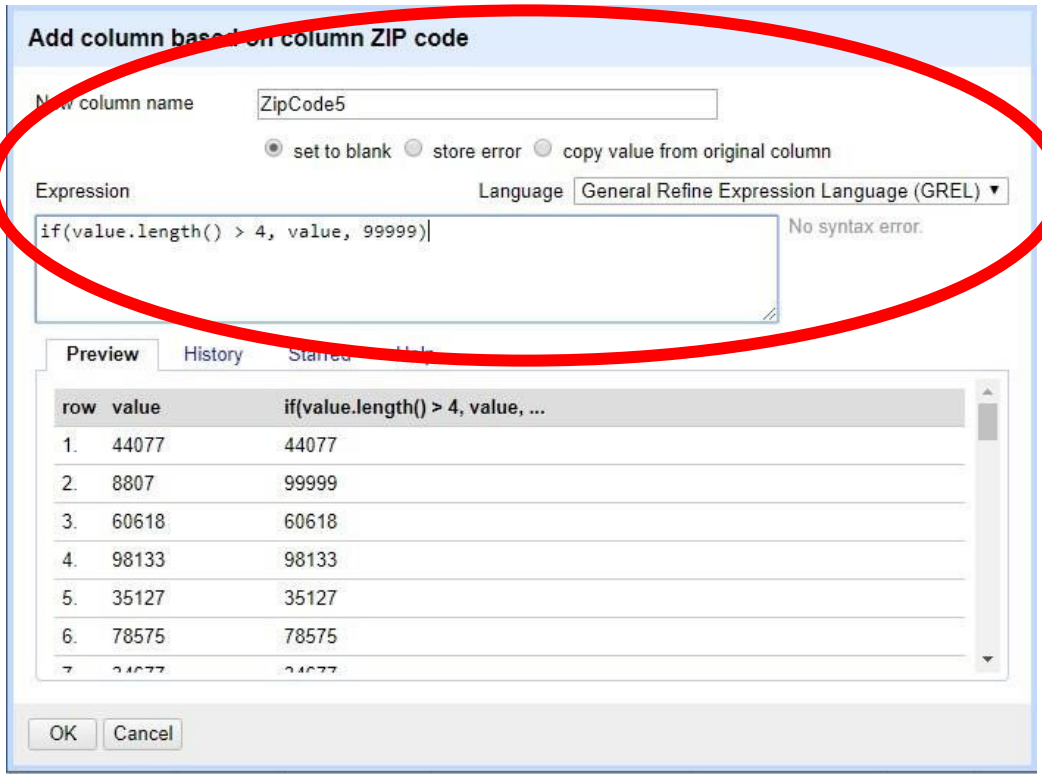
The results have shown that a new column Zip code5 has been created.

Let's pay attention to Row 2:

- The original zip code has 4 digits; the new one has 5 digits 99999. However, it is justified to the left.
- This value is not numeric. It is a string because we enter "99999" in the expression to transform the data
- Let's undo this operation and start over. Do the "**UNDO**" operation to delete the newly created column.

All		Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	Submitted via
	1.	1354490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH	44077	Web
	2.	1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ	8807	Web
	3.	1355730 edit	Credit reporting		Incorrect information on credit report	Account status	IL	60618	Web
	4.	1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133	Web
	5.	1354249	Bank account or service	Checking account	Problems caused by my funds being low		AL	35127	Web

- **Click** on the drop-down menu arrow of Zip code to open the menu. **Click** to select Edit column → **Add** column based on this column
- **Enter**: ZipCode5 for New column name
- **Enter**: if(value.length() > 4, value, 99999) for expression. Then **click** OK



All		Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	ZipCode5	Submitted via
	1.	1354490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH	44077	44077	Web
	2.	1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ	8807	99999	Web
	3.	1355730	Credit reporting		Incorrect information on credit report	Account status	IL	60618	60618	Web
	4.	1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133	98133	Web
	5.	1354249	Bank account or service	Checking account	Problems caused by my funds being low		AL	35127	35127	Web

The results have shown that a new column ZipCode5 has been created

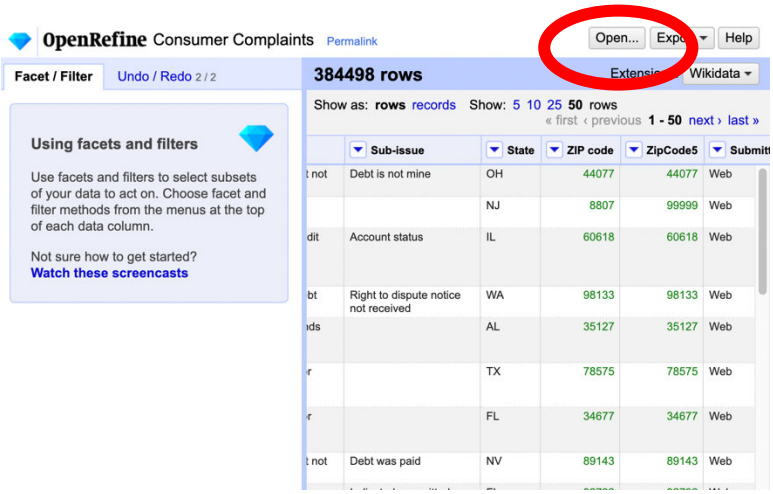
Let's pay attention to Row 2:

- The original zip code has 4 digits; the new one has 5 digits 99999. Now, it is justified to the right. Now, the missing zip code values have been handled by cleansing/wrangling/munging/transforming data.

7. Data Set eq2015: Cleansing & Wrangling Data: Getting Started

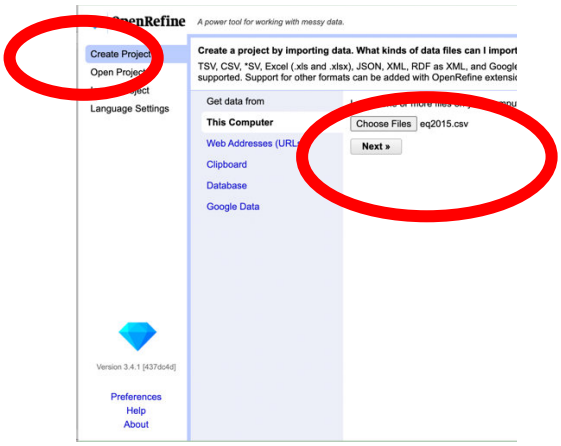
7.1 Start a new project

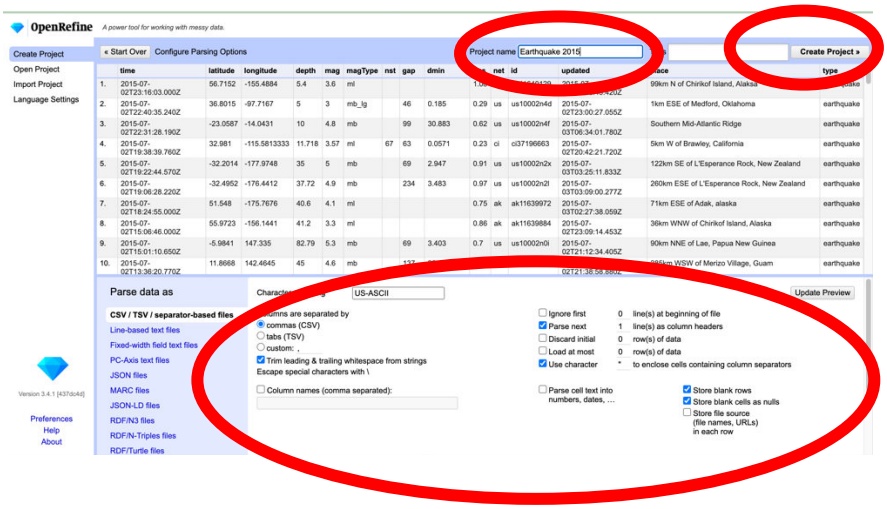
- **Click** Open on the top right corner to open a new web-browser GUI for a new project - **Create** Project



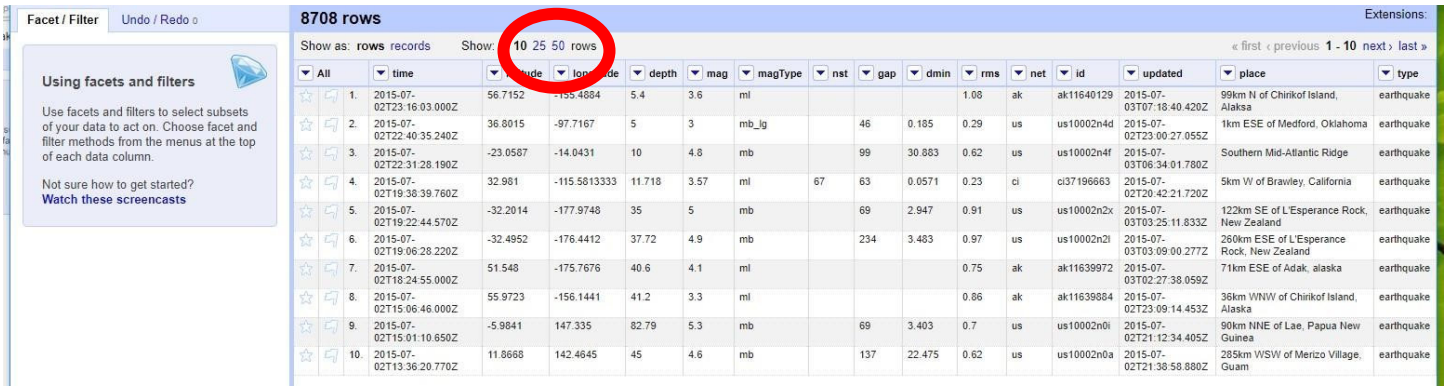
- (Left side below) **Click** Choose Files and browse to the file eq2015.csv and upload the file)
- **Click** Next

- (Right side below) **Name** the project: Earthquake 2015
- **Click** Create Project
- **Leave** defaults as checked





- **Click** “50” to show 50 rows



All	time	depth	mag	magType	nst	gap	dmin	rms	net	id	updated	place	type
1.	2015-07-02T23:16:03.000Z	56.7152	-155.4884	5.4	3.6	ml				1.08	ak	ak11640129	2015-07-03T07:18:40.420Z	99km N of Chirikof Island, Alaksa	earthquake
2.	2015-07-02T22:40:35.240Z	36.8015	-97.7167	5	3	mb_lg		46	0.185	0.29	us	us10002n4d	2015-07-02T23:00:27.055Z	1km ESE of Medford, Oklahoma	earthquake
3.	2015-07-02T22:31:28.190Z	-23.0587	-14.0431	10	4.8	mb		99	30.883	0.62	us	us10002n4f	2015-07-03T06:34:01.780Z	Southern Mid-Atlantic Ridge	earthquake
4.	2015-07-02T19:38:39.760Z	32.981	-115.5813333	11.718	3.57	ml	67	63	0.0571	0.23	ci	ci37196663	2015-07-02T20:42:21.720Z	5km W of Brawley, California	earthquake
5.	2015-07-02T19:22:44.570Z	-32.2014	-177.9748	35	5	mb		69	2.947	0.91	us	us10002n2x	2015-07-03T03:25:11.833Z	122km SE of L'Esperance Rock, New Zealand	earthquake
6.	2015-07-02T19:06:28.220Z	-32.4952	-176.4412	37.72	4.9	mb		234	3.483	0.97	us	us10002n2l	2015-07-03T03:09:00.277Z	260km ESE of L'Esperance Rock, New Zealand	earthquake
7.	2015-07-02T18:24:55.000Z	51.548	-175.7676	40.6	4.1	ml				0.75	ak	ak11639972	2015-07-03T02:27:38.059Z	71km ESE of Adak, alaska	earthquake
8.	2015-07-02T15:06:46.000Z	55.9723	-156.1441	41.2	3.3	ml				0.86	ak	ak11639884	2015-07-02T23:09:14.453Z	36km WNW of Chirikof Island, Alaska	earthquake
9.	2015-07-02T15:01:10.650Z	-5.9841	147.335	82.79	5.3	mb		69	3.403	0.7	us	us10002n0i	2015-07-02T21:12:34.405Z	90km NNE of Lae, Papua New Guinea	earthquake
10.	2015-07-02T13:36:20.770Z	11.8668	142.4645	45	4.6	mb		137	22.475	0.62	us	us10002n0a	2015-07-02T21:38:58.880Z	285km WSW of Merizo Village, Guam	earthquake

7.2 Have a glance at the column nst. *Nst: is the total number of seismic stations used to determine earthquake location.*

- The "nst" column is missing quite a few values.
- Look up the nst attribute in the glossary.
- What would happen if we just ignored a row with missing values?
- Is there an obvious strategy for filling in the missing values?
- What would you suggest we do with the column?

Additional Information

- *Number of seismic stations which reported P- and S-arrival times for this earthquake.*
- *This number may be larger than the Number of Phases.*
- *Used if arrival times are rejected because the distance to a seismic station exceeds the maximum allowable distance or because the arrival-time observation is inconsistent with the solution.*
- We don’t have much knowledge about this attribute, **leave** it as is.

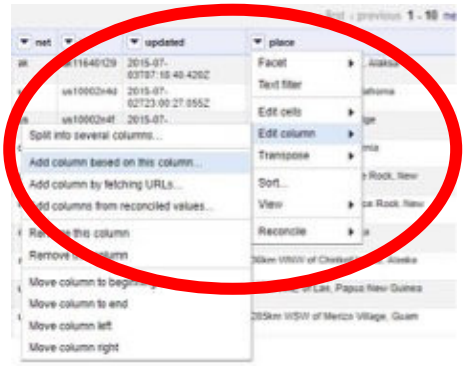
8. Wrangling/Munging Data: Transforming the Place Column

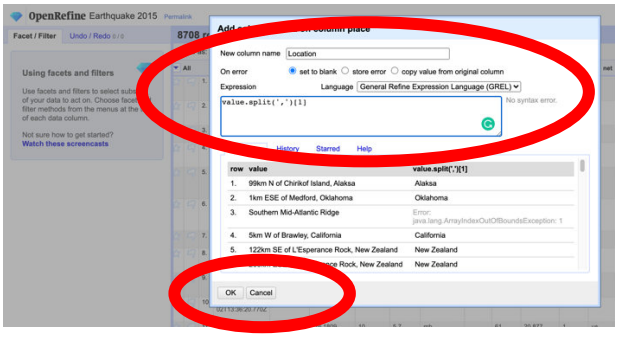
8.1.1 Overview

- We want to extract an approximate area from the "place" column.
- We would like to have a state or country, and to store that information in a separate column, called "location"
- As we review the "place" column, we notice that the cell seems to consist of two comma separated components. The components are a distance and direction, and a general location.

8.1.2 Transforming the data

- **Click** on the drop-down menu arrow of "place" to open the menu
- **Click** to select **Edit** column → **Add** column based on this column





- **Enter**: "Location" for the name of the new column
- **Enter**: value.split('‚')[1] for expression
- **Click** OK

This expression:

- It is assumed that each value contains two substrings separated by a ','
- Split the value into two substrings, using ',' as the separator.
- Store the resulted substrings into some data structure similar to arrays
- The array index starts with 0
- Get the 2nd element at the index 1

Let's preview the results

ermalink

Open... Export Help

8708 rows

Extensions: Wikidata

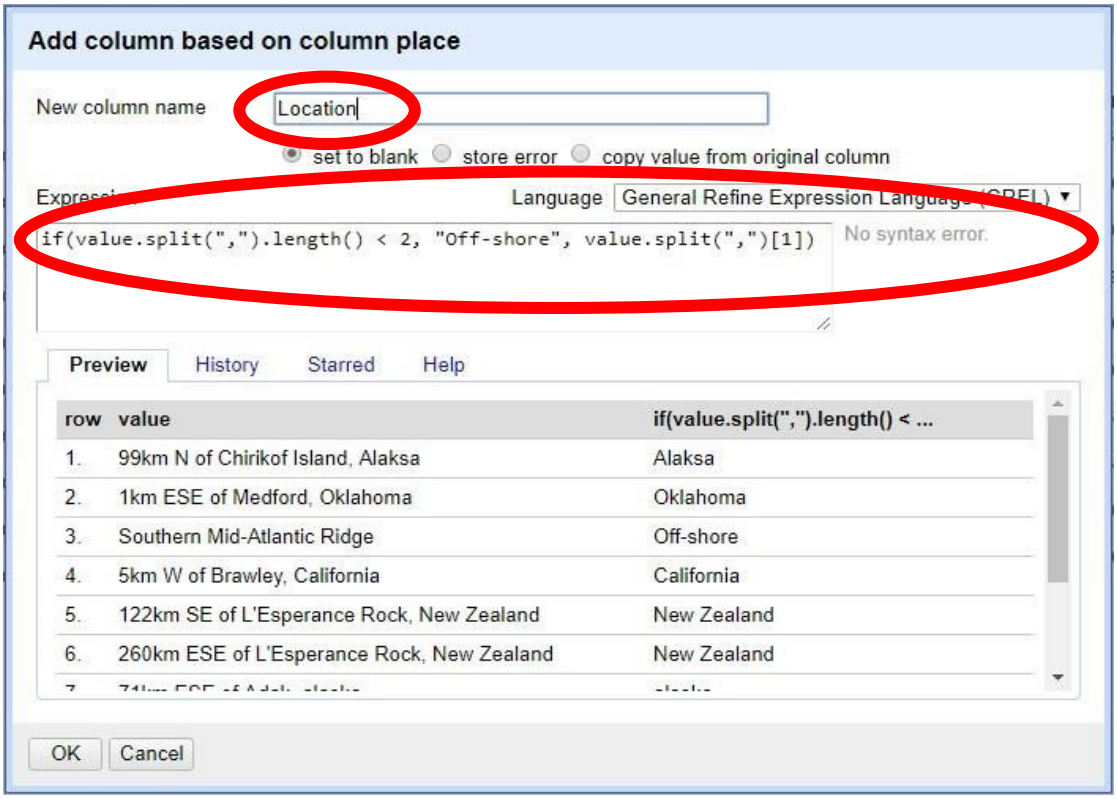
Show as: **rows** records Show: 5 10 25 **50** rows « first ‹ previous **1 - 50** next › last »

All	time	latitude	longitude	depth	mag	magType	nst	gap	dmin	rms	net	id	updated	place	Location
1.	2015-07-02T23:16:03.000Z	56.7152	-155.4884	5.4	3.6	ml				1.08	ak	ak11640129	2015-07-03T07:18:40.420Z	99km N of Chirikof Island, Alaksa	Alaksa
2.	2015-07-02T22:40:35.240Z	36.8015	-97.7167	5	3	mb_lg		46	0.185	0.29	us	us10002n4d	2015-07-02T23:00:27.055Z	1km ESE of Medford, Oklahoma	Oklahoma
3.	2015-07-02T22:31:28.190Z	-23.0587	-14.0431	10	4.8	mb		99	30.883	0.62	us	us10002n4f	2015-07-03T06:34:01.780Z	Southern Mid-Atlantic Ridge	
4.	2015-07-02T19:38:39.760Z	32.981	-115.5813333	11.718	3.57	ml	67	63	0.0571	0.23	ci	ci37196663	2015-07-02T20:42:21.720Z	5km W of Brawley, California	California
5.	2015-07-02T19:22:44.570Z	-32.2014	-177.9748	35	5	mb		69	2.947	0.91	us	us10002n2x	2015-07-03T03:25:11.833Z	122km SE of L'Esperance Rock, New Zealand	New Zealand
6.	2015-07-	-32.4952	-176.4412	37.72	4.9	mb		234	3.483	0.97	us	us10002n2l	2015-07-	260km ESE of	New

- There is some problem in Row #3
- It means that not all the values have two substrings separated by a '‚' --)
- There exist rows in which the value is the name of some place in off-shore territories that are not identified with any state or country. We need to change the expression to handle this

To Do:

Enter: if (value.split(“,”).length() < 2, “Offshore”, value.split(“,”)[1]) for expression



Everything looks good in the preview now.

- **Click** OK

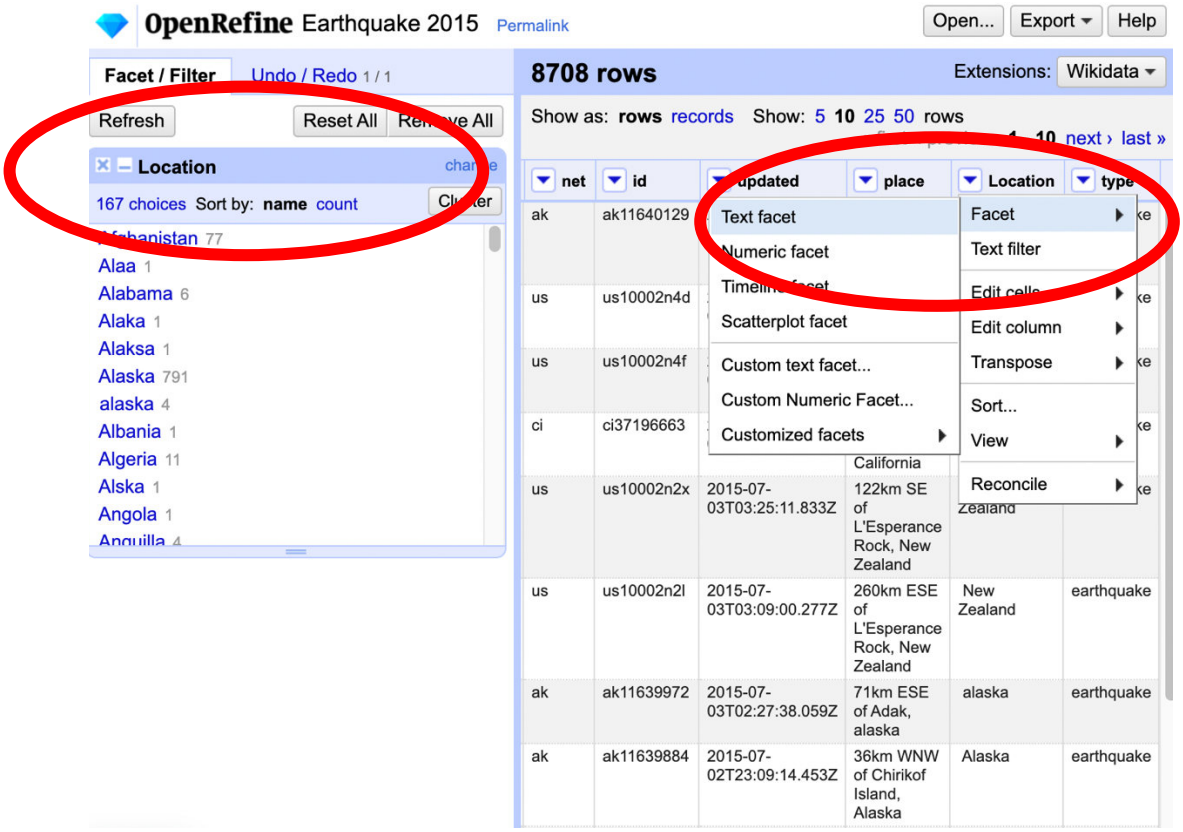
9. Data Set eq2015: Cleansing & Wrangling Data – Location Column

9.1 Overview: Exploring the Location data set

- There are multiple strings that look like "Alaska," but they appear to be misspelled. Let’s look at Row 14 (Alaska), Row 18 (alaska), and Row 21 (alaska)
- It is the same for California
- We need to do further exploration to know the data before we need to fix them

14.	2015-07-02T07:03:23.790Z	53.1083	173.0287	23.49	4.8	mb		130	0.752	0.91	us	us10002mzc	2015-07-02T22:27:31.356Z	31km NNW of Attu Station, Alaska	Alaska
15.	2015-07-02T06:05:46.840Z	38.5198	141.8613	61.57	4.6	mb		133	1.065	1.01	us	us10002mym	2015-07-02T12:55:55.168Z	50km ENE of Ishinomaki, Japa	Japan
16.	2015-07-02T04:03:45.700Z	19.0249	-66.365	40	3.2	Md	14	230.4	0.56593863	0.31	pr	pr15183001	2015-07-02T12:06:24.867Z	61km N of Brenas, Puerto Ric	Puerto Rico
17.	2015-07-02T03:37:41.200Z	18.5076	-68.657	135	3.4	Md	9	342	1.12379242	0.1	pr	pr15183002	2015-07-02T11:40:16.262Z	8km NNE of San Rafael del ama, Dominican Republic	Dominican Republic
18.	2015-07-02T02:18:27.000Z	60.9657	-147.4064	23.8	3.7	ml				0.55	ak	ak11639524	2015-07-02T03:01:09.571Z	60km WSW of Valdez, alask	alaska
19.	2015-07-02T01:56:11.220Z	27.749	85.2992	10	4.2	mb		120	0.039	0.96	us	us10002mxq	2015-07-02T09:42:31.008Z	5km NNW of Kathmandu, Ne al	Nepal
20.	2015-07-02T01:35:49.100Z	19.6	-67.8811	44	3.4	Md	7	327.6	1.34118472	0.18	pr	pr15183000	2015-07-02T09:38:31.557Z	125km NNE of Punta Cana, Dominican Republic	Dominican Republic
21.	2015-07-01T23:20:13.000Z	56.1872	-153.5558	12	4.3	ml				0.71	ak	ak11639423	2015-07-02T21:15:08.040Z	118km S of Larsen Bay, alaska	alaska
22.	2015-07-01T21:59:53.860Z	36.0803	-97.4545	2.23	3	mb_lg		65	0.283	0.29	us	us10002mx3	2015-07-02T02:21:52.713Z	18km NE of Crescent, Oklahoma	Oklahoma

- **Click** on the drop-down menu arrow of Location to open the menu
- **Click** to select Facets → **Text** facets



10. Exploring/Cleansing/Wrangling Data Set by Clustering

10.1 Overview

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.

OpenRefine's text facets are a great mechanism for surfacing patterns from your data. Consider a data set that contains people's names entered in two different ways: "first_name middle_initial last_name", and "last_name, first_name". A text facet on that column might reveal

choice	count
Andy Anderson	79
Andy R. Anderson	9
Anderson, Andy	57

Beatrice Beaufort	28
Beatrice Mansfield	67
Beaufort, Beatrice	19
...	...

Say you want to change every name to "first_name last_name", then using that facet, you would need to edit "Anderson, Andy" to "Andy Anderson", "Beaufort, Beatrice" to "Beatrice Beaufort", and so forth. And of course there are the occasional middle initials to worry about. While editing through the text facet is already orders of magnitude easier than editing the same data in a spreadsheet, there is a more automated way to do this: the Clustering feature.

The Clustering feature can be accessed in 2 different ways.

If you have already created a default text facet on a column, the text facet will show a "Cluster" button near its top right corner. If you haven't, you can invoke the column's drop-down menu and pick Edit cells > Cluster and edit...

The clustering feature works by trying to group the choices in the text facet, so that choices that "look similar" get grouped together.

For instance, operating on our example data above, the clustering feature would generate the group

Andy Anderson (79)
Andy R. Anderson (9)
Anderson, Andy (57) and
the group

Beatrice Beaufort (28)
Beaufort, Beatrice (19) and so
forth.

You can select each group whose choices you want to "merge" and enter the text that all choices in that group will be replaced with. If we select the first group to be merged into "Andy Anderson", then 79 + 9 + 57 = 145 cells will contain the text "Andy Anderson".

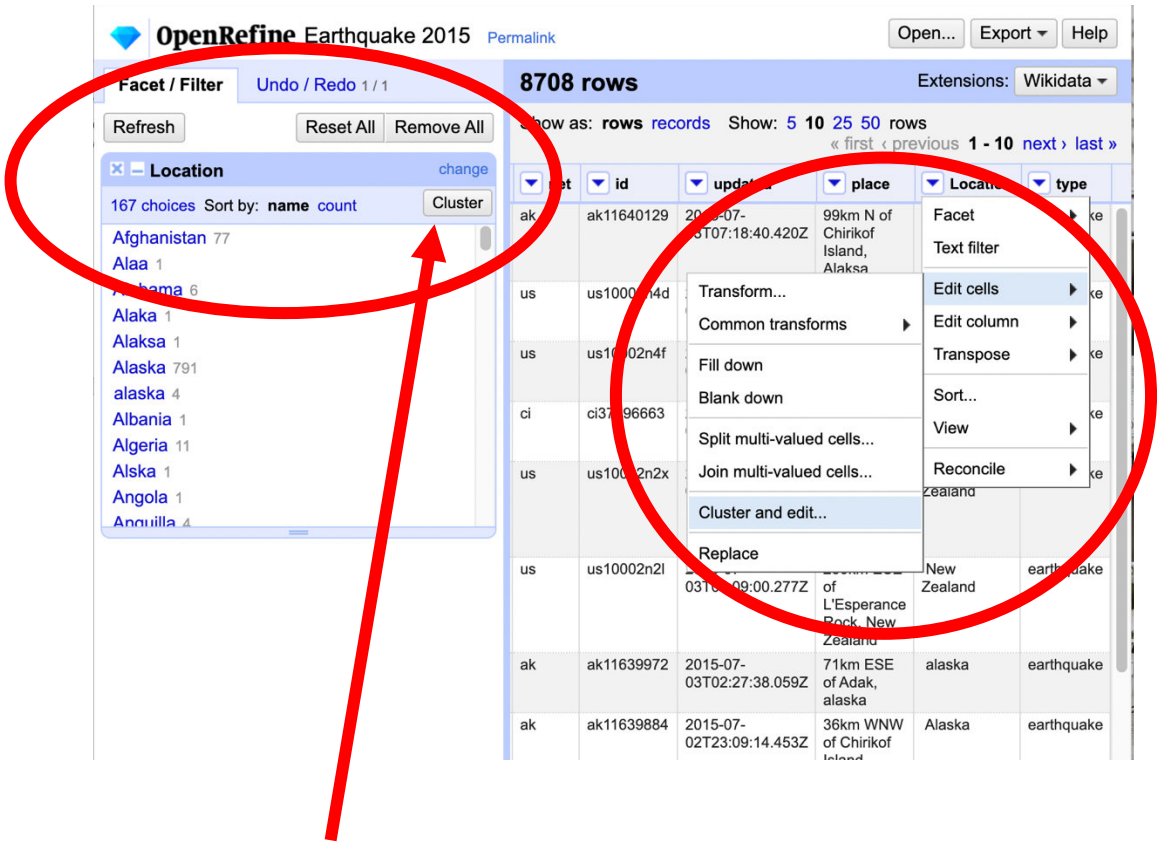
You can adjust the way the clustering feature groups the choices--that is, how the feature determines that choices "look similar". When the feature is "conservative", it might consider "Andy Anderson" and "Anderson, Andy" to be similar, but not "Andy R. Anderson" and "Andy Anderson". When the feature is "liberal" or "aggressive", it might consider "Andy Anderson", "Sandy Anderson", and "Landy Sanderson" to all be similar. How conservative or liberal the feature should work depends on your data, but it is safer to be conservative and cautious in merging. Of course, you can always undo each cluster and edit operation.

10.2 Using the Earthquake 2015 project

You can run a cluster by

- Click on Cluster on the facet
- OR: Click on the Column

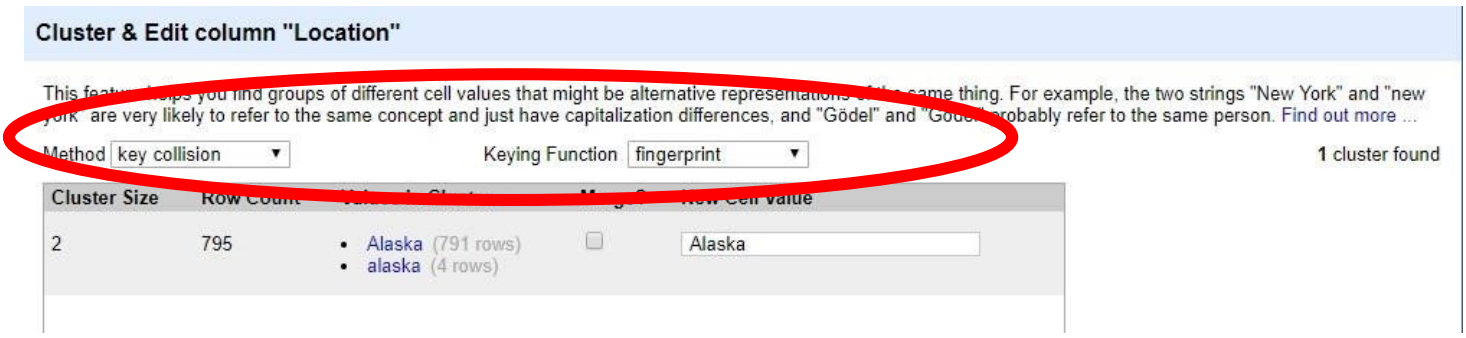
  Example: Location → Edit cells → Cluster and edit …



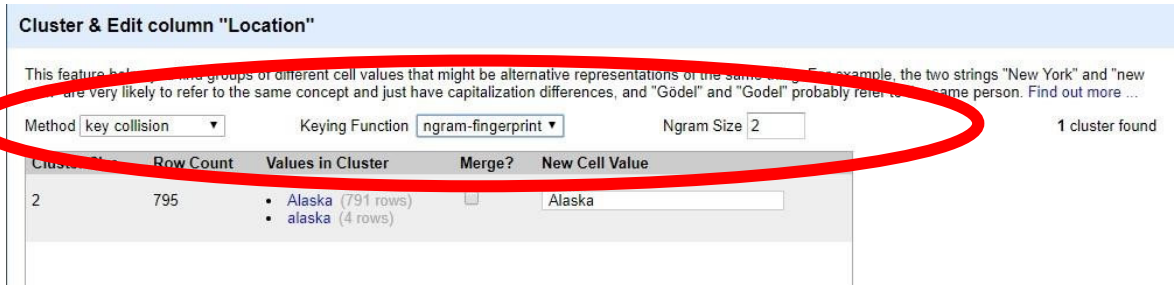
We will **Click** Cluster on the facet

10.2.1 : Different ways to cluster

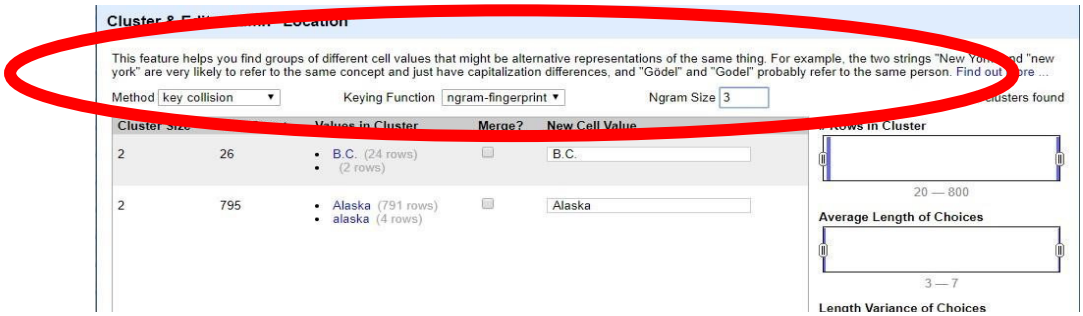
A. OpenRefine: Clustering with the **key collision** method - **fingerprint**



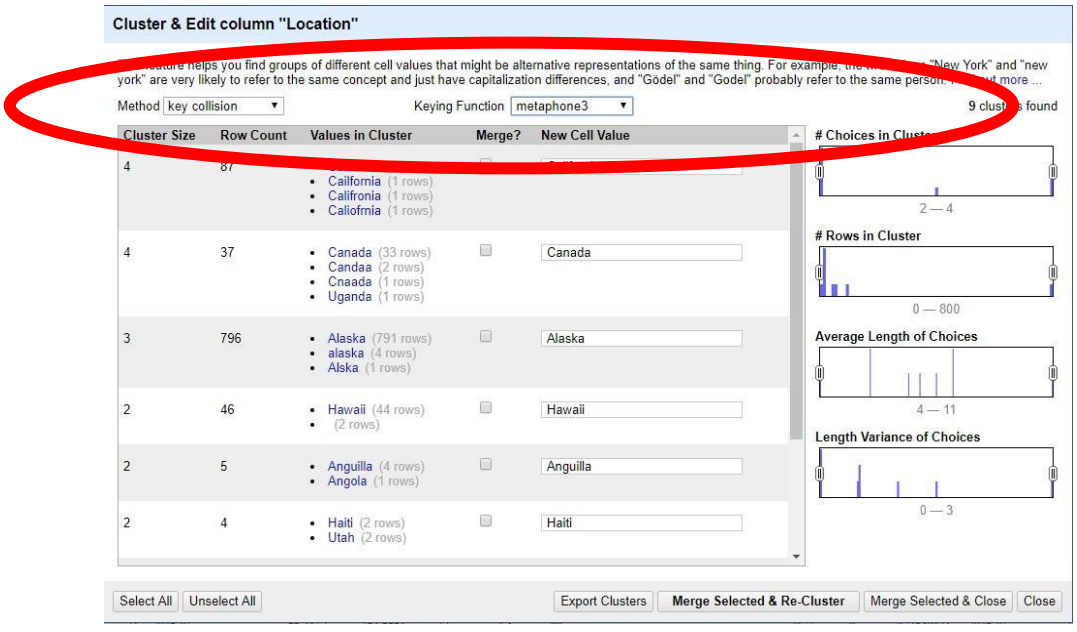
B. OpenRefine: Clustering with the **key collision** method **- ngram-fingerprint with 2 for the size**



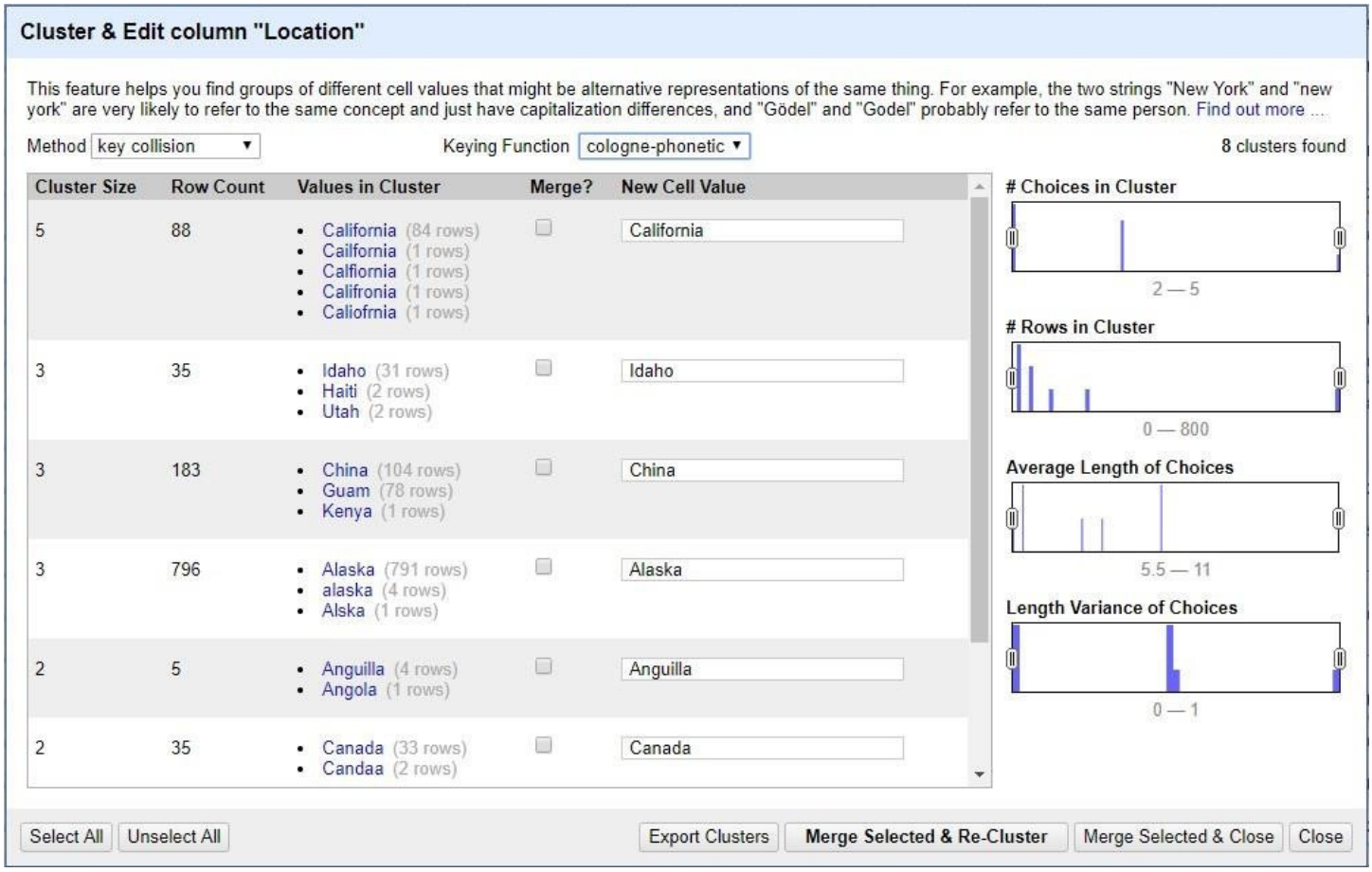
C. OpenRefine: Clustering with the **key collision** method - **ngram-fingerprint with 3for the size**



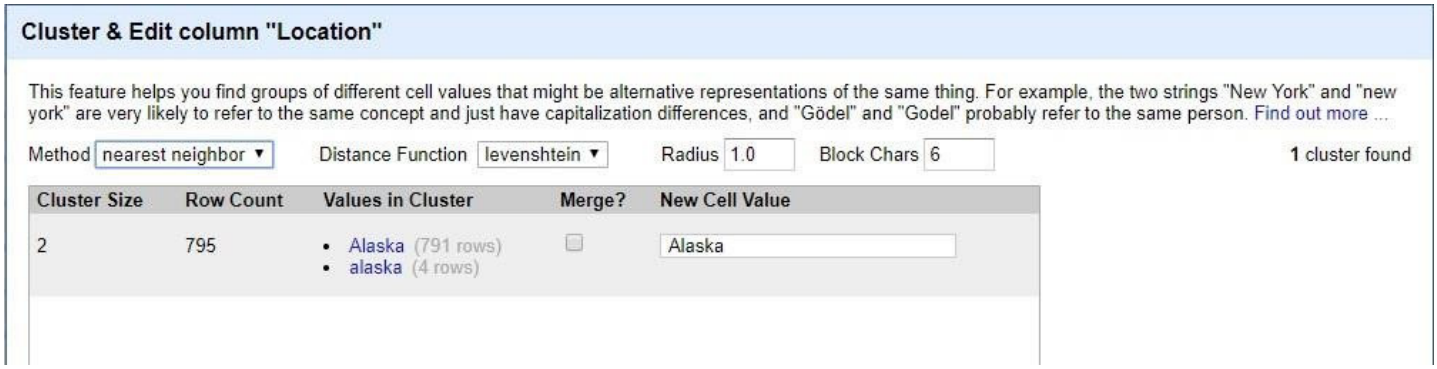
D. OpenRefine: Clustering with the **key collision** method - **metaphone3**



E. OpenRefine: Clustering with the **key collision** method - **cologne-phonetic**



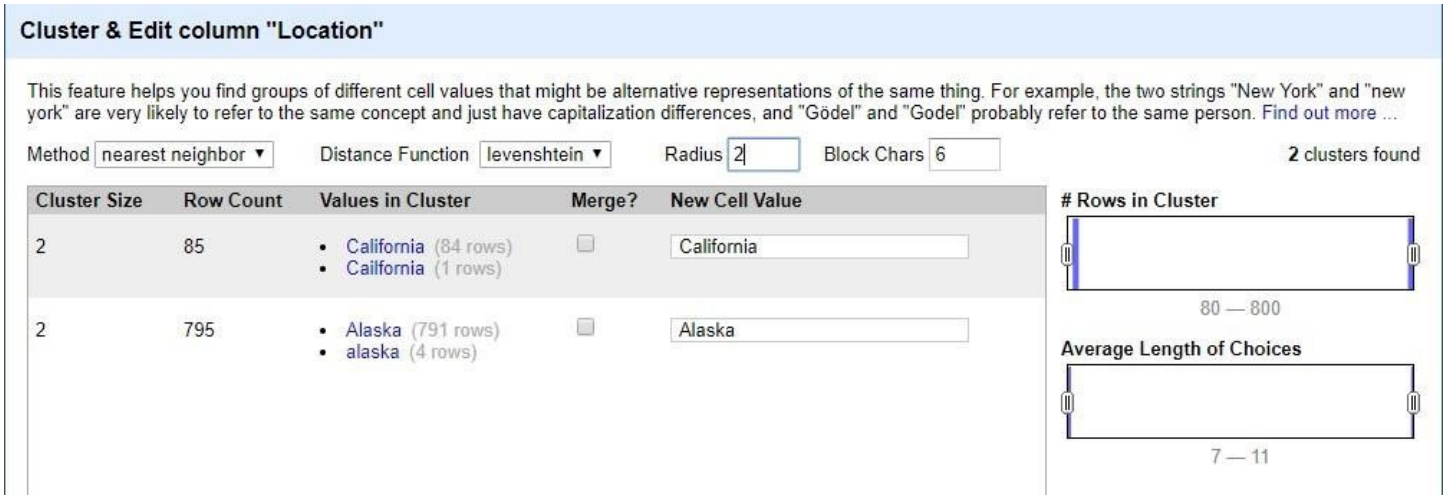
F. OpenRefine: Clustering with the **nearest neighbor** method: **Levenshtein with radius of 1.0 and Block Chars of 6**



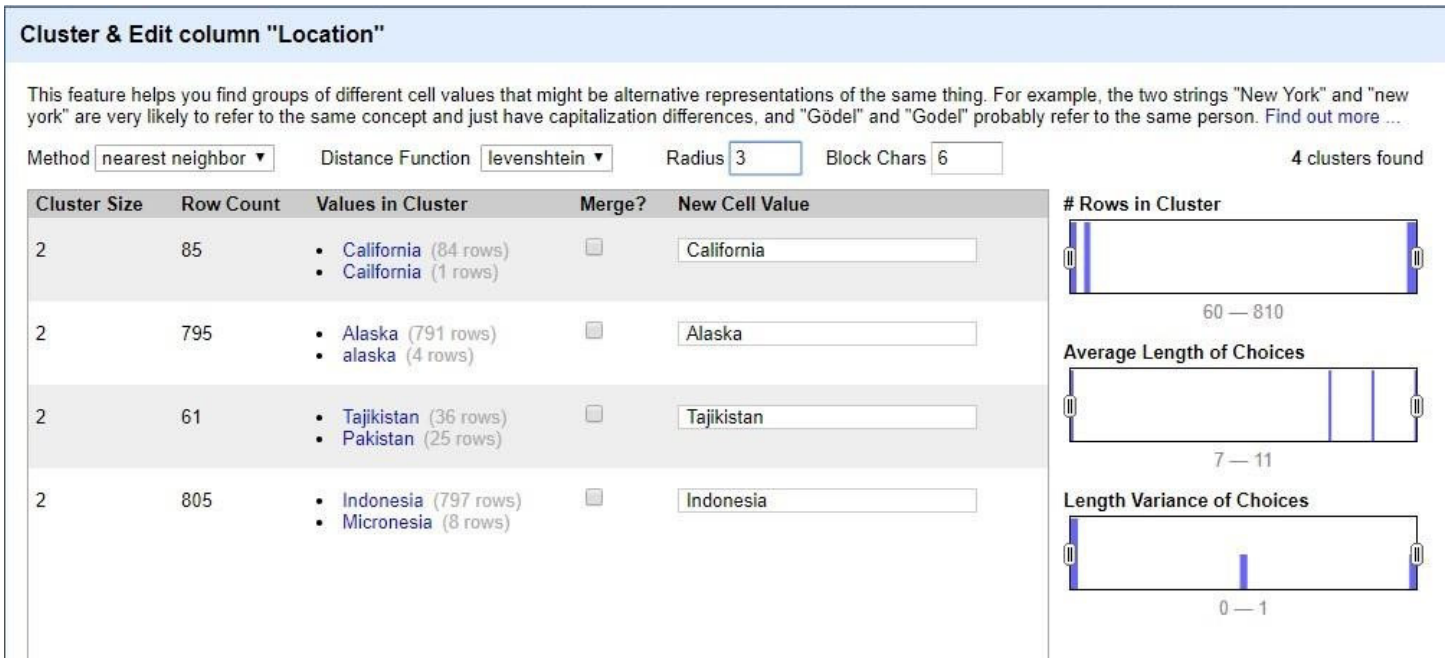
Greater Radius: more liberal in clustering

(". When the feature is "conservative", it might consider "Andy Anderson" and "Anderson, Andy" to be similar, but not "Andy R. Anderson" and "Andy Anderson". When the feature is "liberal" or "aggressive", it might consider "Andy Anderson", "Sandy Anderson", and "Landy Sanderson" to all be similar.)

G. OpenRefine: Clustering with the **nearest neighbor** method: **Levenshtein with radius of 2.0 and Block Chars of 6**

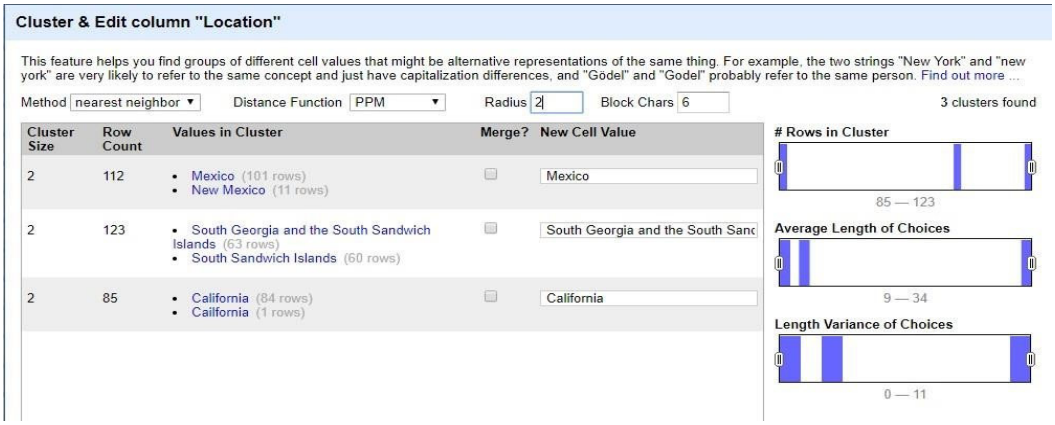


H. OpenRefine: Clustering with the **nearest neighbor** method: **Levenshtein with radius of 3.0 and Block Chars of 6**



I. OpenRefine: Clustering with the **nearest neighbor** method: **PPM with radius of 1.0 and Block Chars of 6**

J. OpenRefine: Clustering with the **nearest neighbor** method: **PPM with radius of 2.0 and Block Chars of 6**



K. OpenRefine: Clustering with the **nearest neighbor** method: **PPM with radius of 3.0 and Block Chars of 6**

To understand the differences between each keying function, read the clustering in-depth tutorial

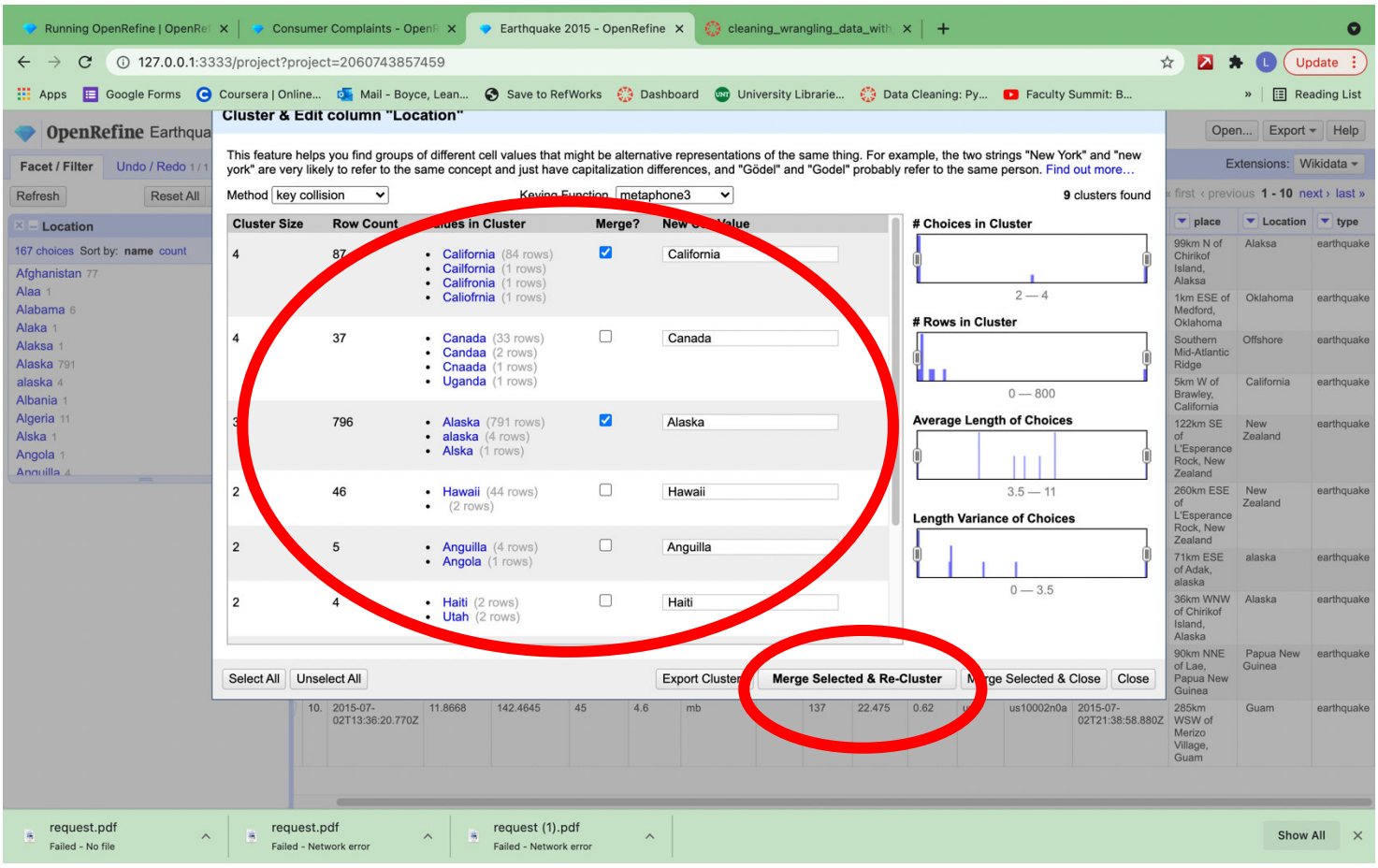
11. Cleansing/Wrangling Data: Editing Cells

11.1 Overview

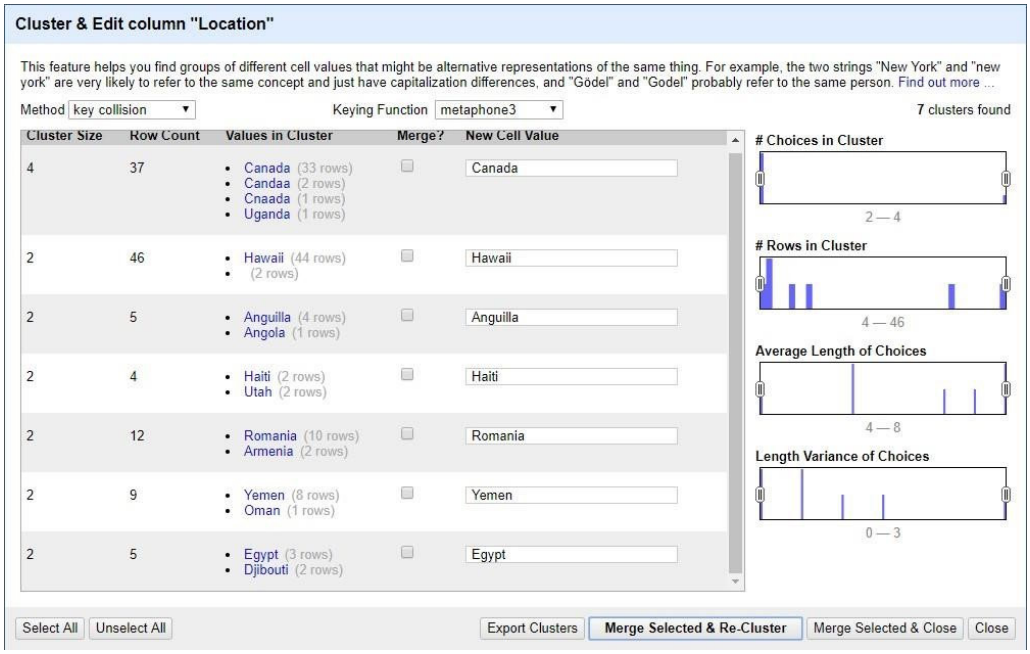
- Having used the clustering feature of OpenRefine to detect the group of cells that may have variants of contents, potential sources of typos or errors
- We need to fix these typos and errors using the clustering feature of OpenRefine
-

11.2 Editing Cells Using Key Collision Method & Metaphone3 Keying Function

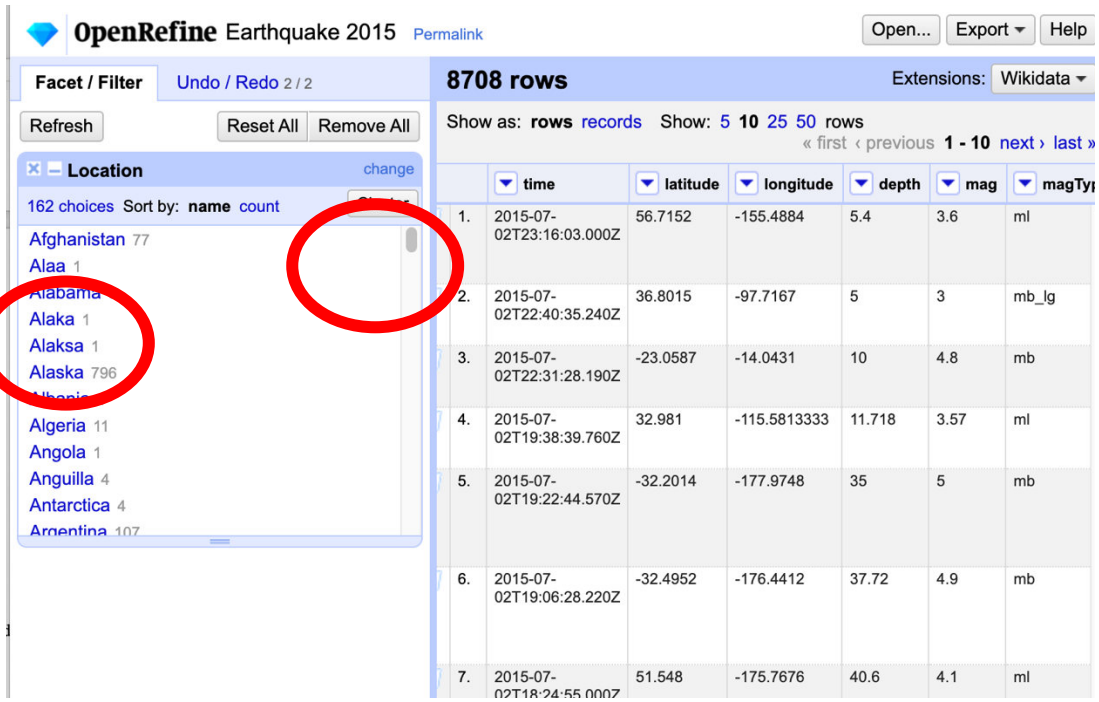
- **Click** on the drop-down menu arrow of Location to open the menu
- **Click** to select Facets → Text facets
- **Click** Cluster in the facet



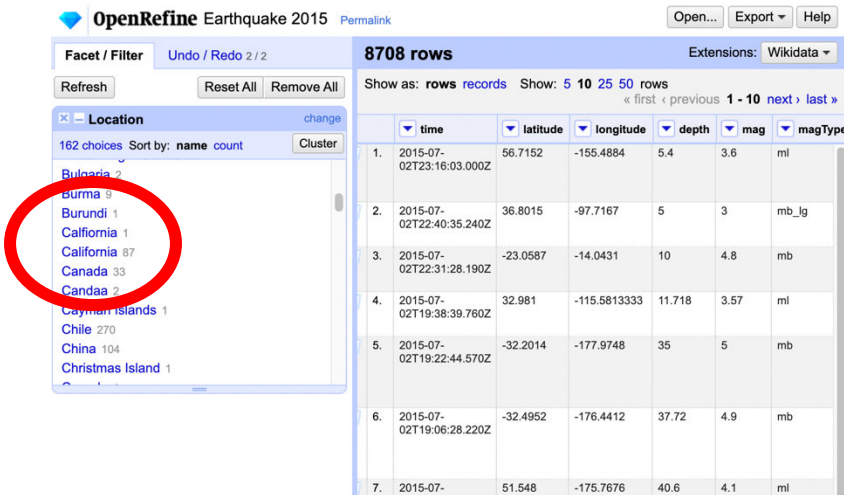
- **Check** boxes under Merge for the clusters California and Alaska
- BE SURE! The new cell value is correct (California and Alaska)
- **Click** Merge Selected and Re-Cluster



- The results show that the two clusters of California and Alaska have gone, i.e., these typos and errors have been fixed.
- **Click** Close
- Let's check the facet again to see if the re-clustering is complete.



- The results show that all the "Alaska" have the capitalized 'A' now.
- However, there are still two potential typos: Alaka and Alaksa (not included in the cluster) → We need to handle manually
- Let's check California (scroll down under "Cluster")



- The results show that all the “California” have the capitalized ‘C’ now.
- However, there are still one potential typo: Calfliornia (not included in the cluster) → Handle manually

12. Cleansing/Wrangling Data: Manually Editing Cell

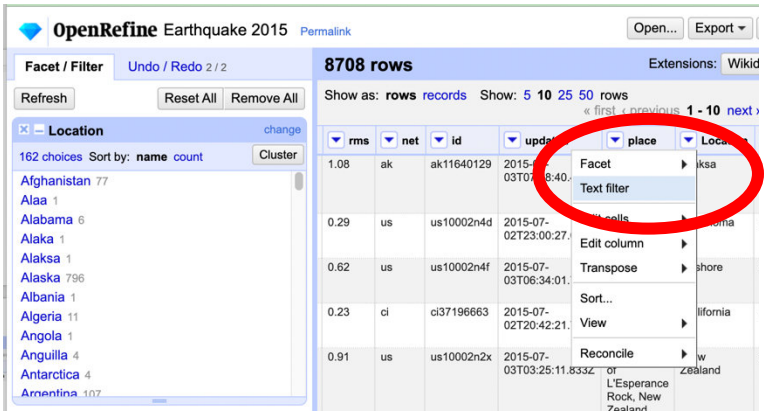
12.1 Overview

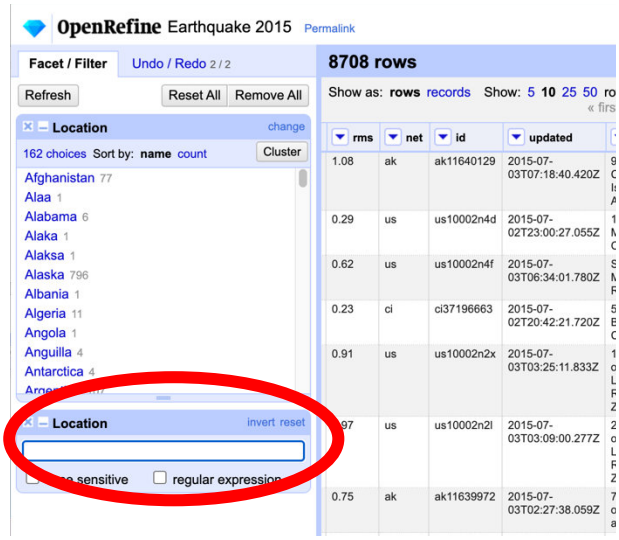
- In some cases, we need to manually edit cells to correct typos or errors
- First, use OpenRefine’s Text filter feature to locate the row/record in which a potential typo or error is found.
- Second, we consider what should be the correct form of this potential typo or error.
- Next, we can use related information found in the same record, or the previous, or the following one, to validate our suggestion.
- If it is verified, use “edit” feature of a cell to edit the text of the value.

12.2 Edit Cell to Correct “Alaka” → “Alaska”

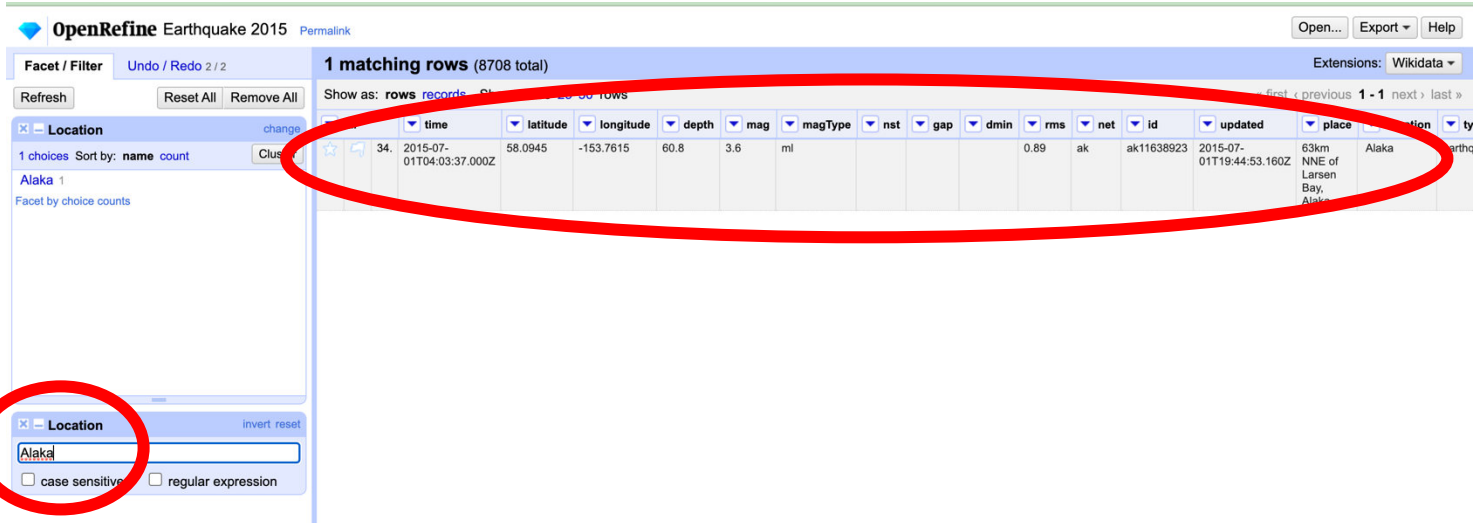
12.2.1 Locate the row/record

- **Click** on the drop-down menu arrow of Location to open the menu
- **Click** → Text filter





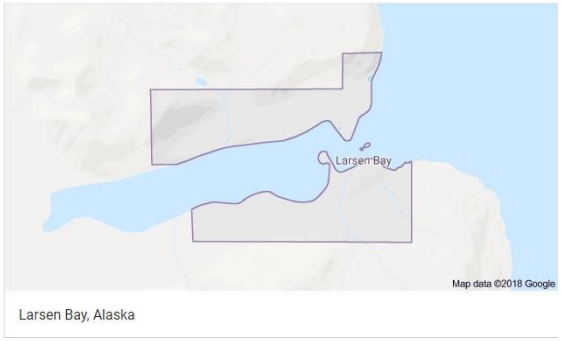
- Enter: Alaka



The results show that this potential typo, Alaka, is found in Row 34

12.2.2 Verified the suggestion: Alaka → Alaska

- In this row, the "place" shows that this location is close to Larsen Bay
- We need to verify that Larsen Bay need to be in Alaska - Using Google Map, we can confirm it:



12.2.3 Edit the cell

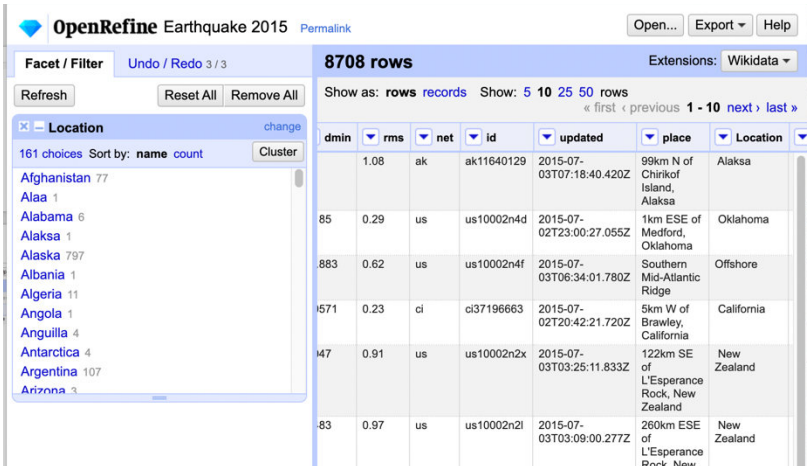
- **Move** the mouse cursor over the cell of "Alaka" ("edit" button will show up)

- **Click** on the "edit" button

- **Type**: Alaska into the editor

- **Click** Apply

The results show that the typo, Alaka, has been fixed.
Let's check again by redoing text facet of this data set again



The results show that the typo, Alaka, has gone.
There is still another potential typo, Alaksa, that may need fixing manually, too. Try it out on your own.