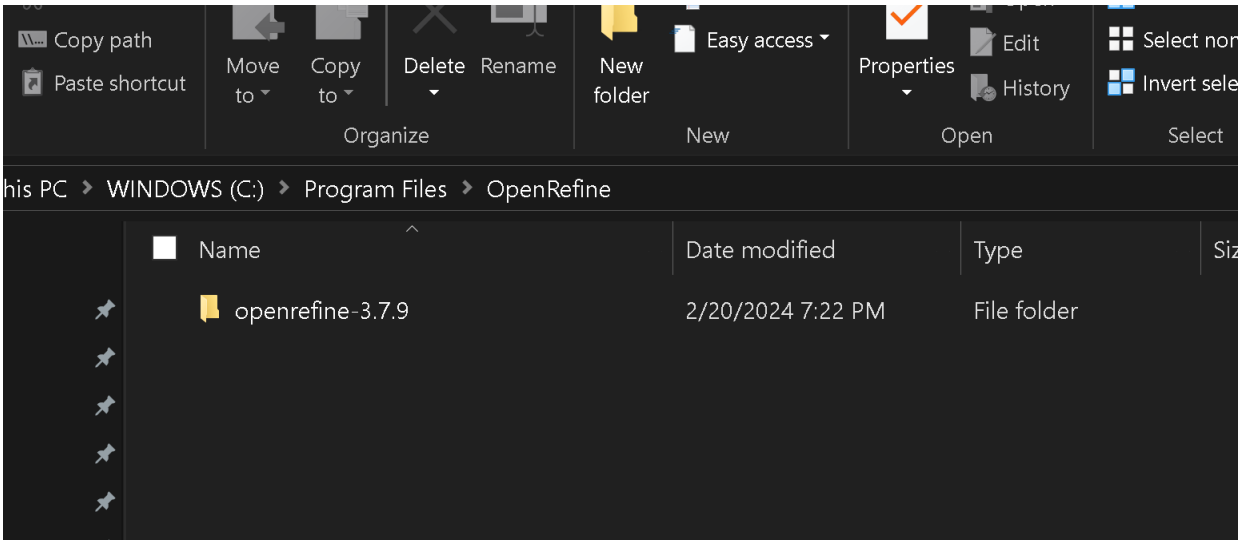
MODULE 6: HOMEWORK 1: ASSIGNMENT: DATA PREP WITH OPENREFINE

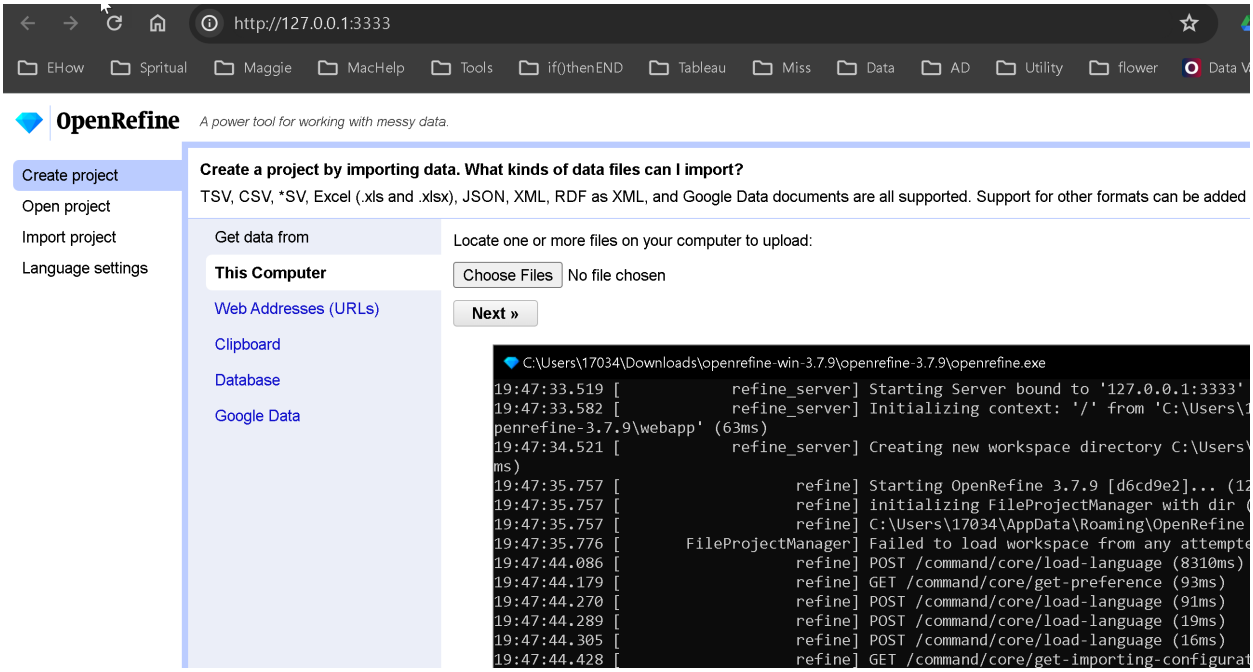
1. DOWNLOAD AND INSTALLATION

- Go to https://openrefine.org/download
- Download the .zip file, and extract it into a folder where you wish to store program files to C:\Program Files\OpenRefine.
- Install Java https://adoptium.net/temurin/releases/?os=windows
- Open the application by right-click on openrefine.exe or refine.bat.



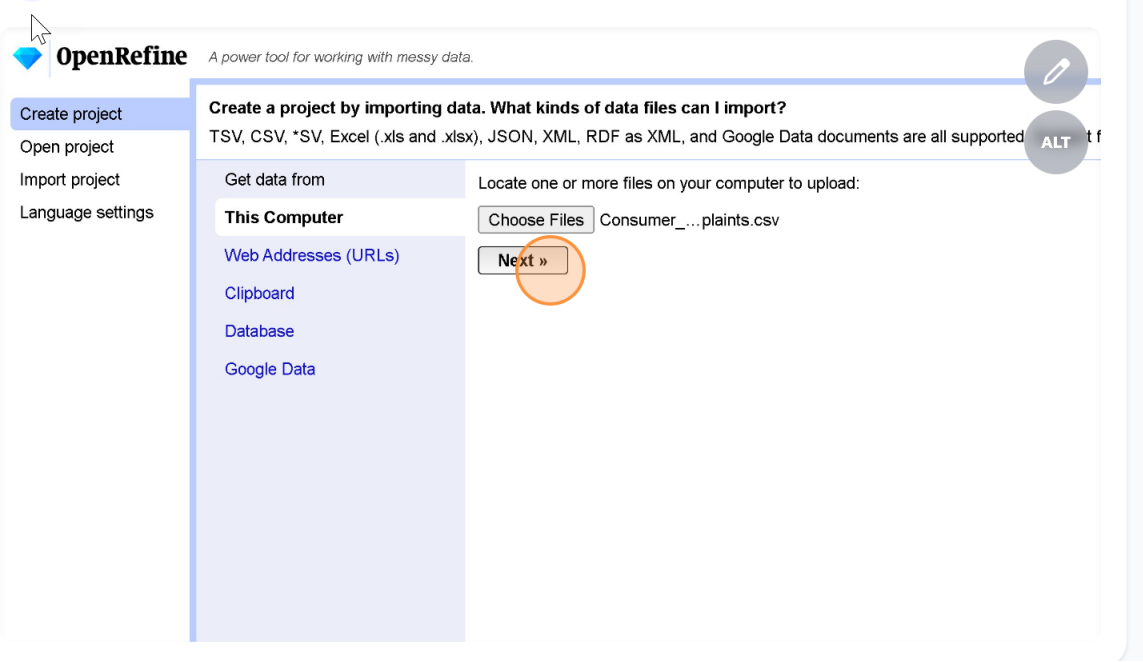
2. OPENREFINE GUI

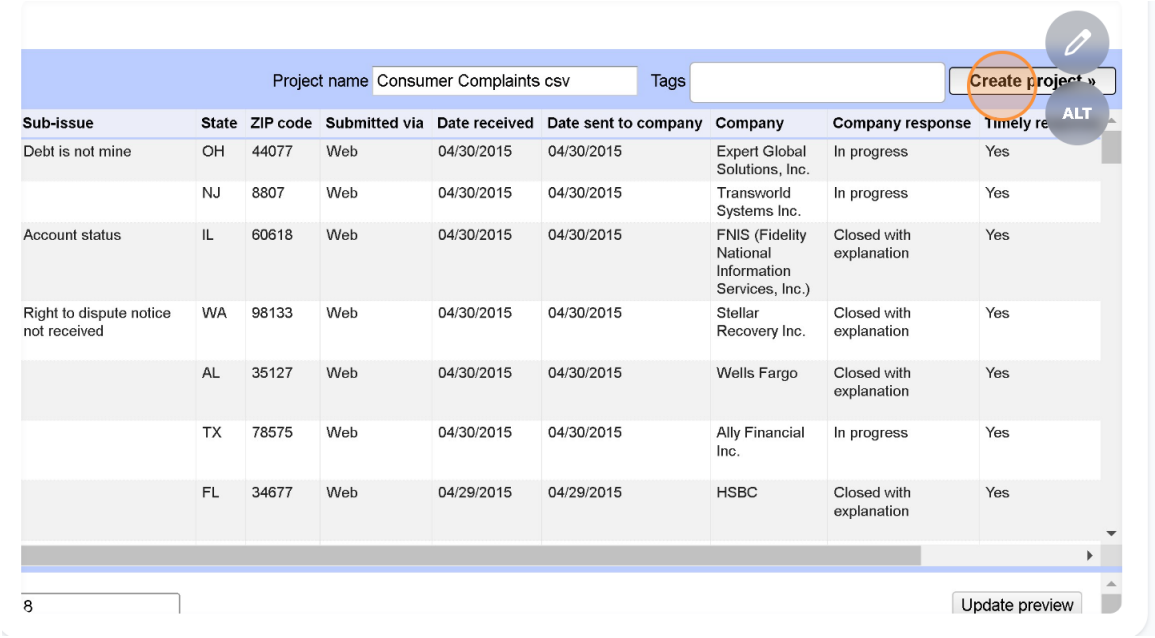
- OpenRefine automatically open to a browser window, or I can manually navigate to the application's URL (using http://127.0.0.1:3333).



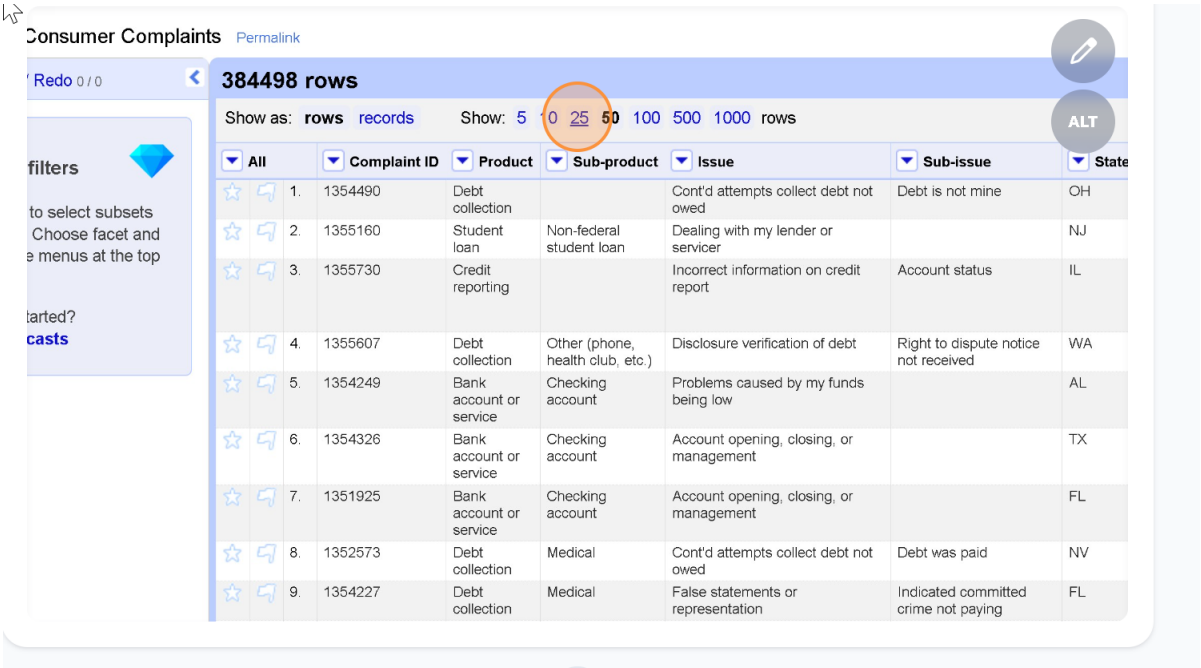
3. CREATE A NEW PROJECT & UPLOAD DATA

- Click > Create Project tab on the left side
- To select the data sources, chose files > Uploaded file, then click Open
- Click next and write your Project name > click Create project



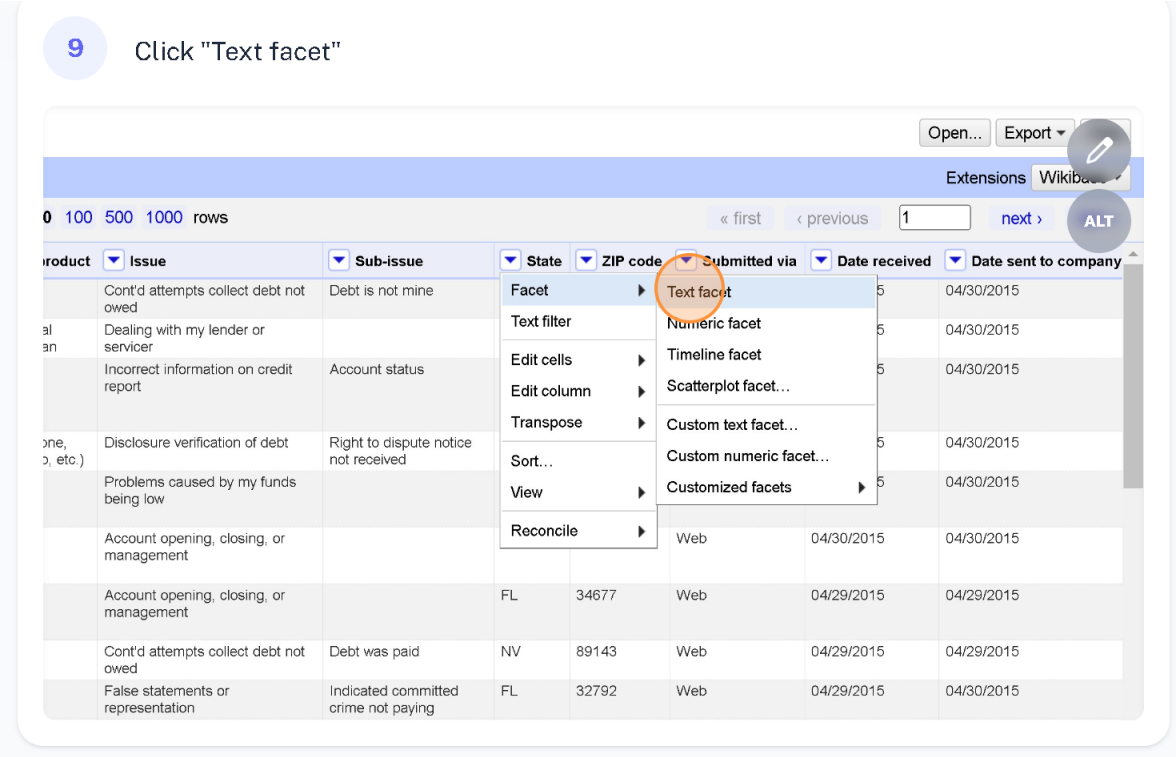


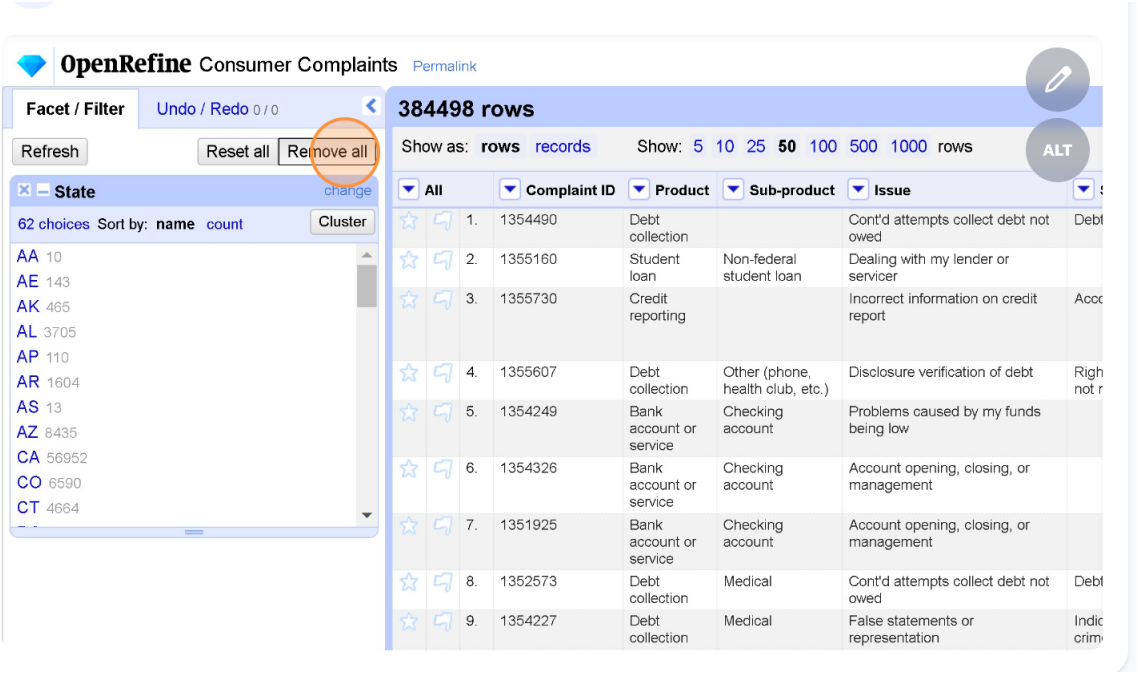
Sub-issue	State	ZIP code	Submitted via	Date received	Date sent to company	Company	Company response	Timely re
Debt is not mine	OH	44077	Web	04/30/2015	04/30/2015	Expert Global Solutions, Inc.	In progress	Yes
	NJ	8807	Web	04/30/2015	04/30/2015	Transworld Systems Inc.	In progress	Yes
Account status	IL	60618	Web	04/30/2015	04/30/2015	FNIS (Fidelity National Information Services, Inc.)	Closed with explanation	Yes
Right to dispute notice not received	WA	98133	Web	04/30/2015	04/30/2015	Stellar Recovery Inc.	Closed with explanation	Yes
	AL	35127	Web	04/30/2015	04/30/2015	Wells Fargo	Closed with explanation	Yes
	TX	78575	Web	04/30/2015	04/30/2015	Ally Financial Inc.	In progress	Yes
	FL	34677	Web	04/29/2015	04/29/2015	HSBC	Closed with explanation	Yes



4. CHECK STATES WITH TEXT FACETS

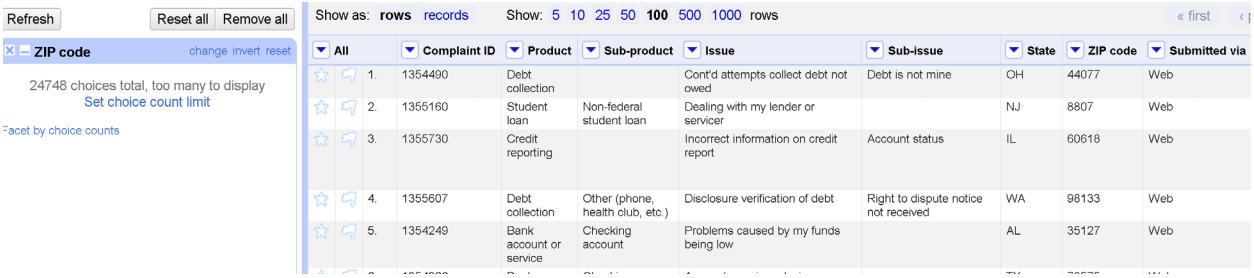
- Select the Column: “State” > click Text facet. I could see all the states in the file as well as their distribution of unique values.



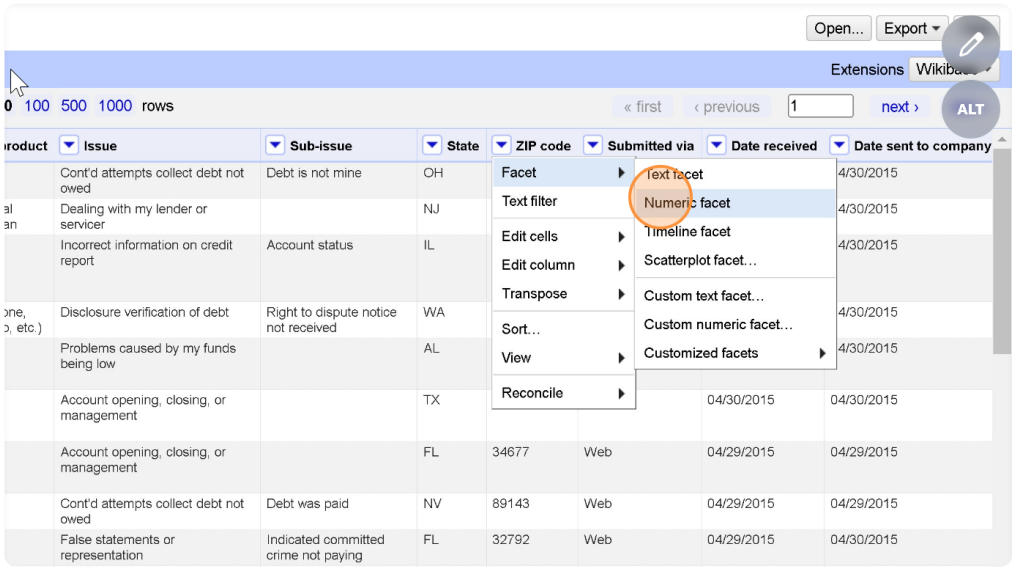


5. WRANGLING/MUNGING – TRANSFORMING DATA: CHECK ZIP CODE

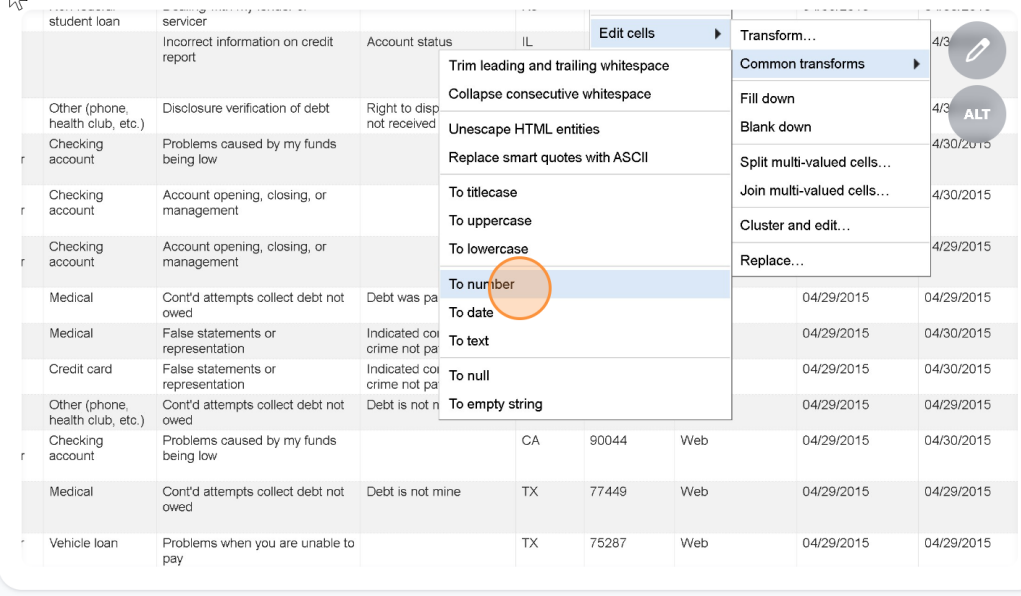
First I was ablet to check the txt facet for the zip code. The result shows 24748 total zipcodes. And it shows all the values are string filed. I follow the step to convert the data to Number.

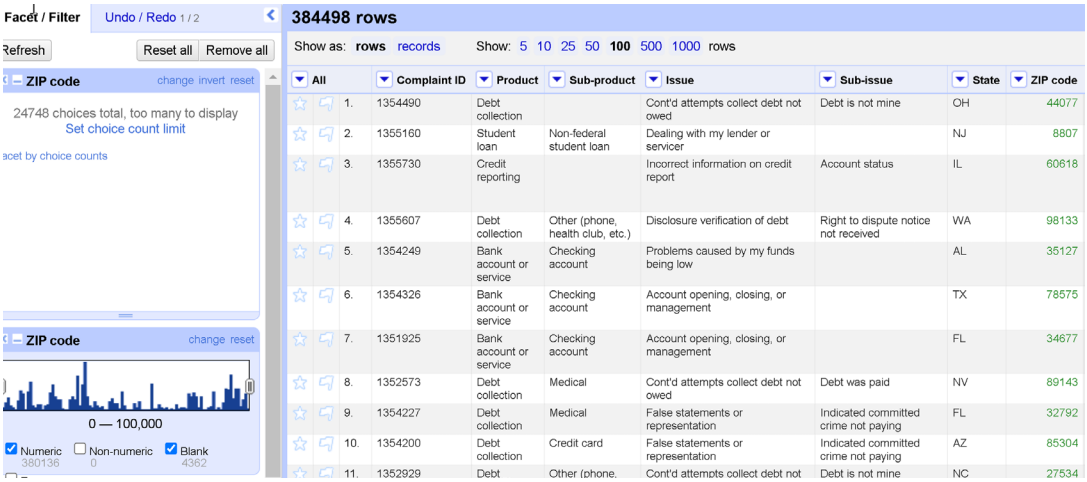


15 Click "Numeric facet"



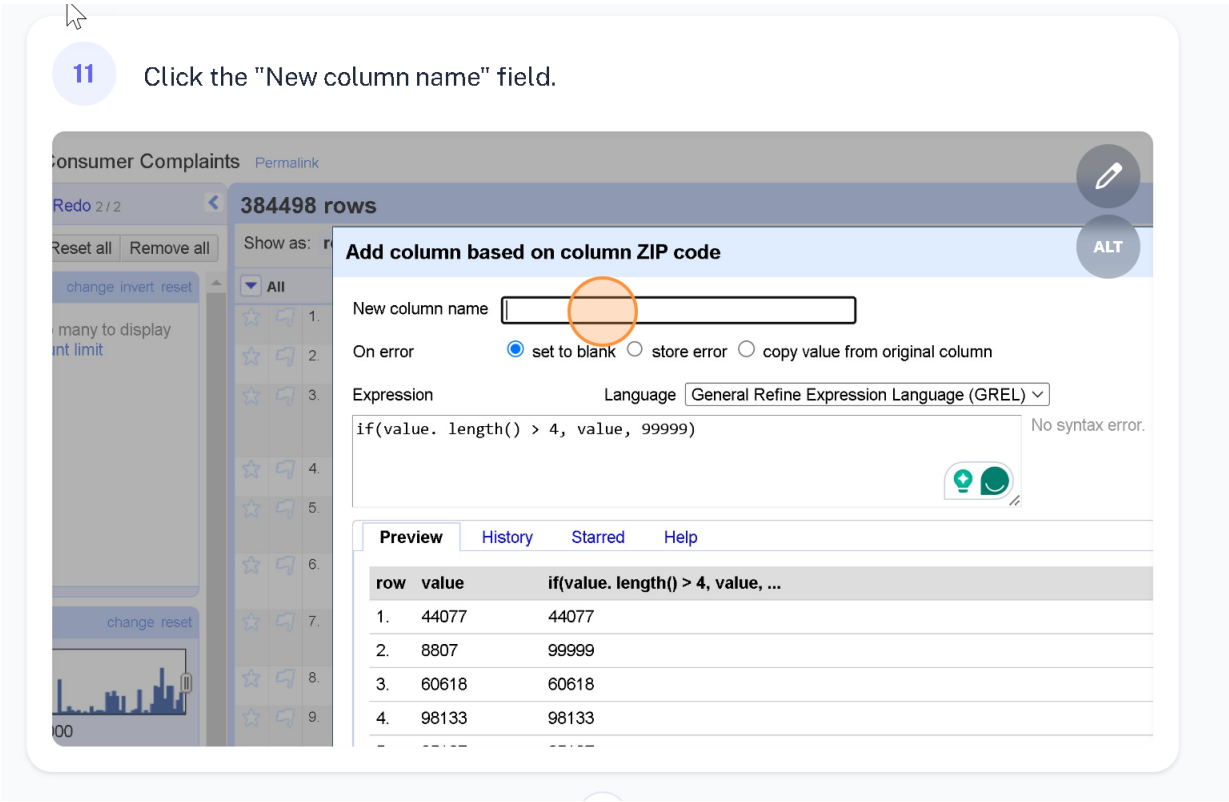
17 Click "To number"





6. CLEANSING/WRANGLING DATA: HANDLING MISSING ZIP CODES

- In this step, I try to fill in the missing values
- Step Add new column using the new Name “ZipCode5” and using Expression if(value.length() > 4, value, 99999)



-

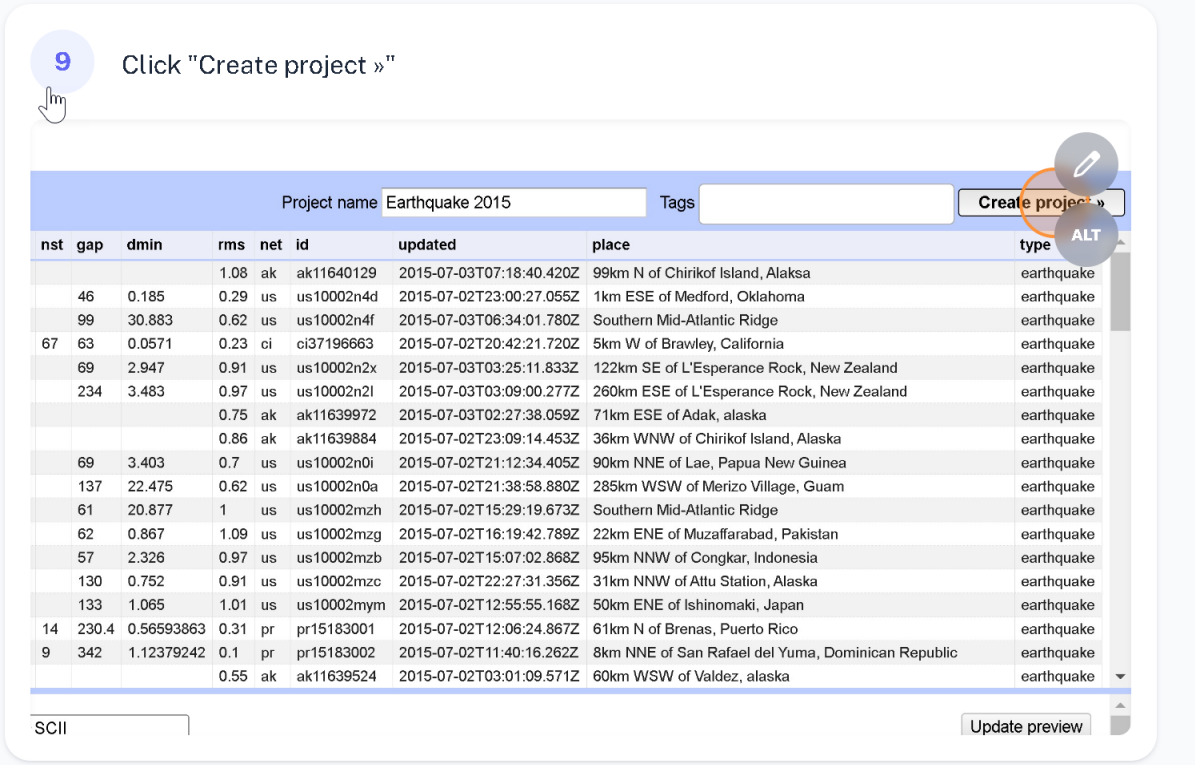
384498 rows

Show as: **rows** records Show: 5 10 25 50 **100** 500 1000 rows « first ‹ previous 1

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	04/
2.	1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ	8807	99999	Web	04/
3.	1355730	Credit reporting		Incorrect information on credit report	Account status	IL	60618	60618	Web	04/
4.	1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133	98133	Web	04/
5.	1354249	Bank account or	Checking account	Problems caused by my funds being low		AL	35127	35127	Web	04/

7. DATA SET EQ2015: CLEANSING & WRANGLING DATA:

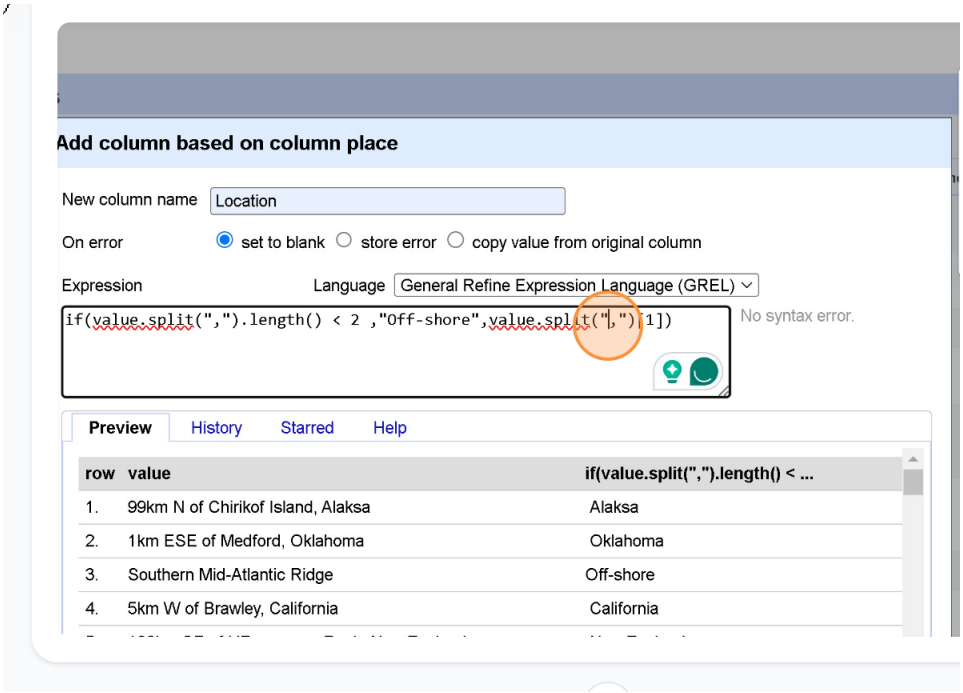
- I opened a new browser and created a new project with the EQ2015 data set.



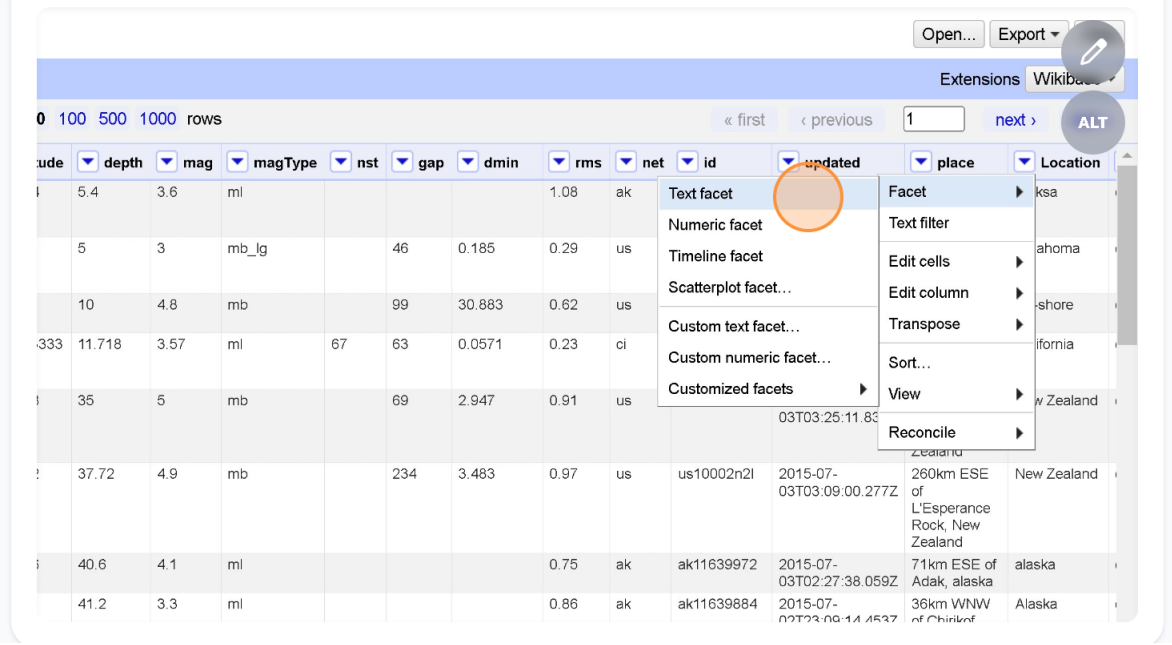
nst	gap	dmin	rms	net	id	updated	place	type
			1.08	ak	ak11640129	2015-07-03T07:18:40.420Z	99km N of Chirikof Island, Alaksa	earthquake
	46	0.185	0.29	us	us10002n4d	2015-07-02T23:00:27.055Z	1km ESE of Medford, Oklahoma	earthquake
	99	30.883	0.62	us	us10002n4f	2015-07-03T06:34:01.780Z	Southern Mid-Atlantic Ridge	earthquake
67	63	0.0571	0.23	ci	ci37196663	2015-07-02T20:42:21.720Z	5km W of Brawley, California	earthquake
	69	2.947	0.91	us	us10002n2x	2015-07-03T03:25:11.833Z	122km SE of L'Esperance Rock, New Zealand	earthquake
	234	3.483	0.97	us	us10002n2l	2015-07-03T03:09:00.277Z	260km ESE of L'Esperance Rock, New Zealand	earthquake
			0.75	ak	ak11639972	2015-07-03T02:27:38.059Z	71km ESE of Adak, alaska	earthquake
			0.86	ak	ak11639884	2015-07-02T23:09:14.453Z	36km WNW of Chirikof Island, Alaska	earthquake
	69	3.403	0.7	us	us10002n0i	2015-07-02T21:12:34.405Z	90km NNE of Lae, Papua New Guinea	earthquake
	137	22.475	0.62	us	us10002n0a	2015-07-02T21:38:58.880Z	285km WSW of Merizo Village, Guam	earthquake
	61	20.877	1	us	us10002mzh	2015-07-02T15:29:19.673Z	Southern Mid-Atlantic Ridge	earthquake
	62	0.867	1.09	us	us10002mzg	2015-07-02T16:19:42.789Z	22km ENE of Muzaffarabad, Pakistan	earthquake
	57	2.326	0.97	us	us10002mzb	2015-07-02T15:07:02.868Z	95km NNW of Congkar, Indonesia	earthquake
	130	0.752	0.91	us	us10002mzc	2015-07-02T22:27:31.356Z	31km NNW of Attu Station, Alaska	earthquake
	133	1.065	1.01	us	us10002mym	2015-07-02T12:55:55.168Z	50km ENE of Ishinomaki, Japan	earthquake
14	230.4	0.56593863	0.31	pr	pr15183001	2015-07-02T12:06:24.867Z	61km N of Brenas, Puerto Rico	earthquake
9	342	1.12379242	0.1	pr	pr15183002	2015-07-02T11:40:16.262Z	8km NNE of San Rafael del Yuma, Dominican Republic	earthquake
			0.55	ak	ak11639524	2015-07-02T03:01:09.571Z	60km WSW of Valdez, alaska	earthquake

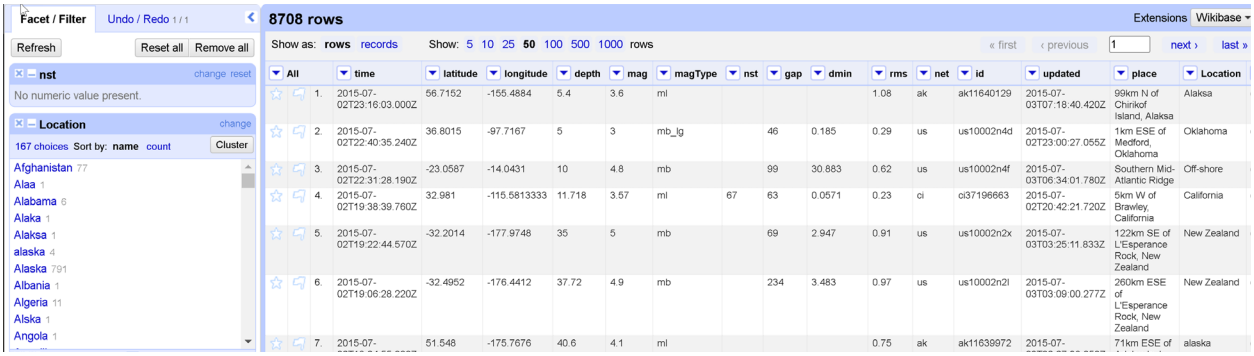
8. STARTED WRANGLING/MUNGING DATA: TRANSFORMING THE PLACE COLUMN

- Extract transforming the Place Column > Click drop-down mean of Place Column > Add Column based on this column. Enter value.split(‘,’)[1] for expression to split substring based on coma.



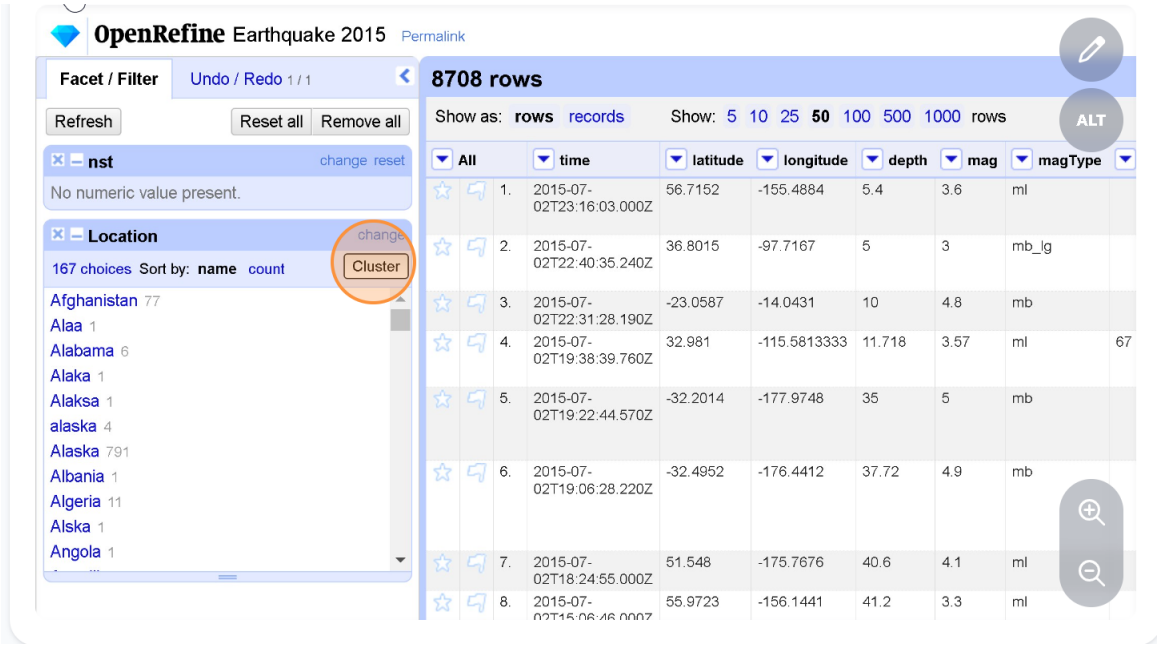
9. DATA SET EQ2015: CLEANSING & WRANGLING DATA – LOCATION COLUMN

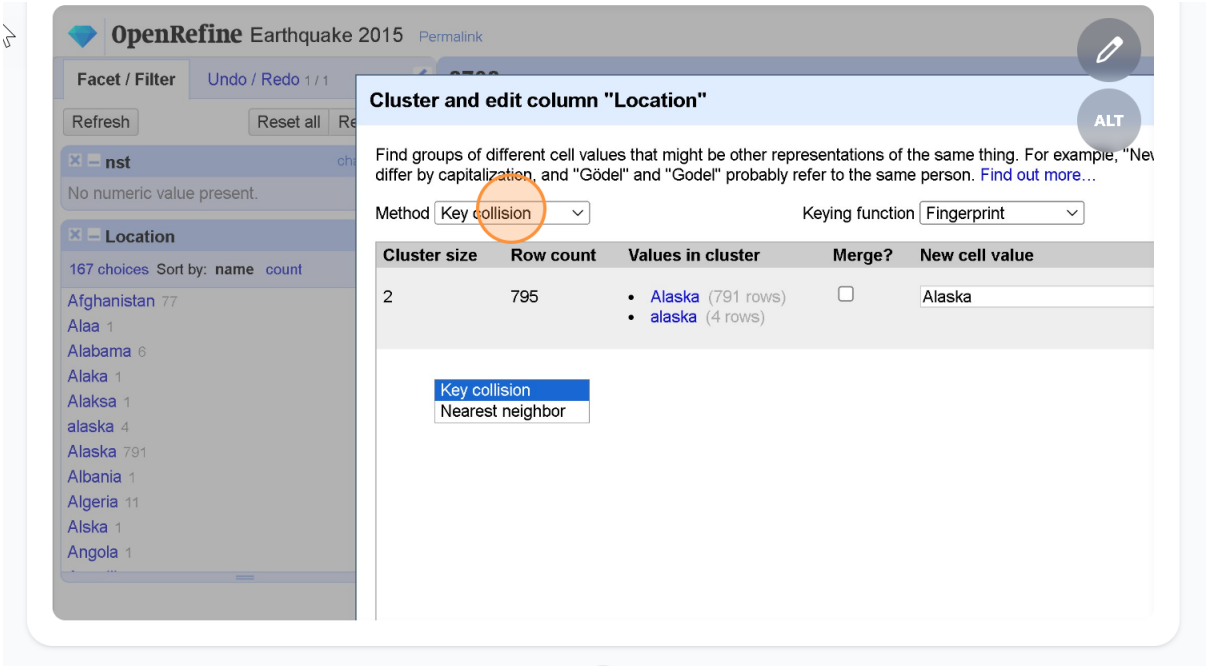




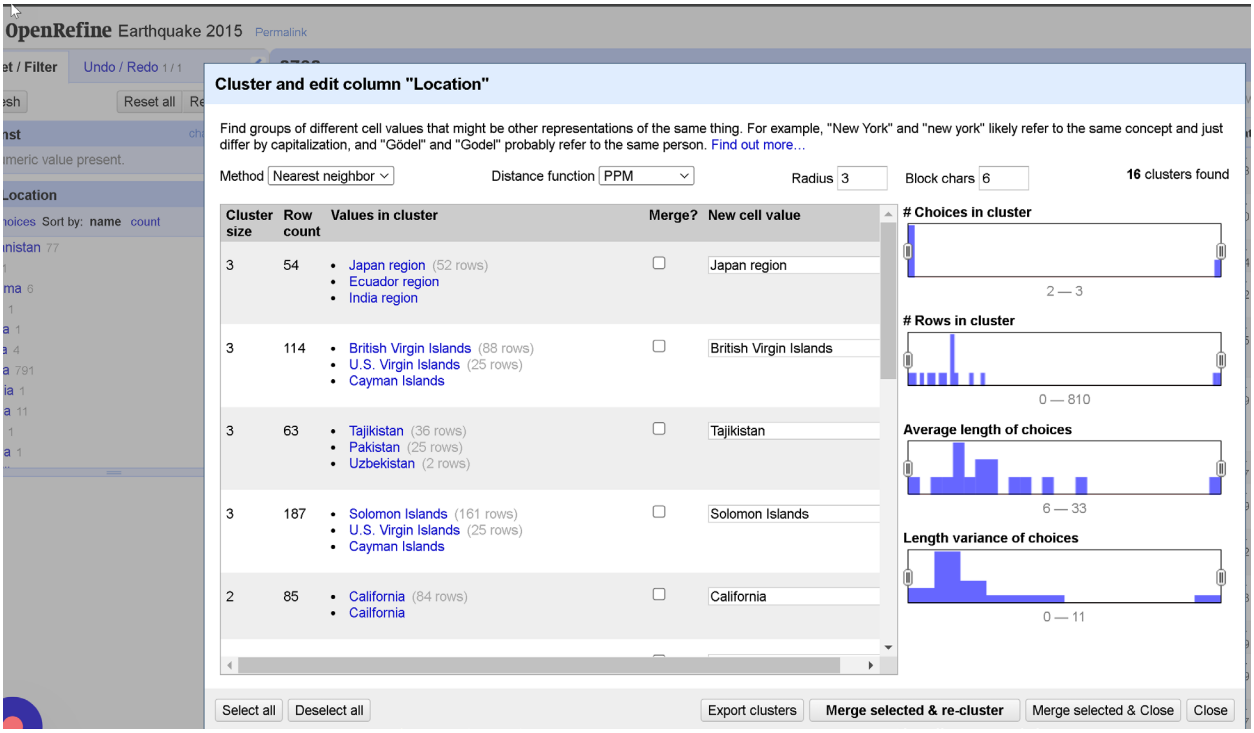
10. EXPLORING/CLEANSING/WRANGLING DATA SET BY CLUSTERING

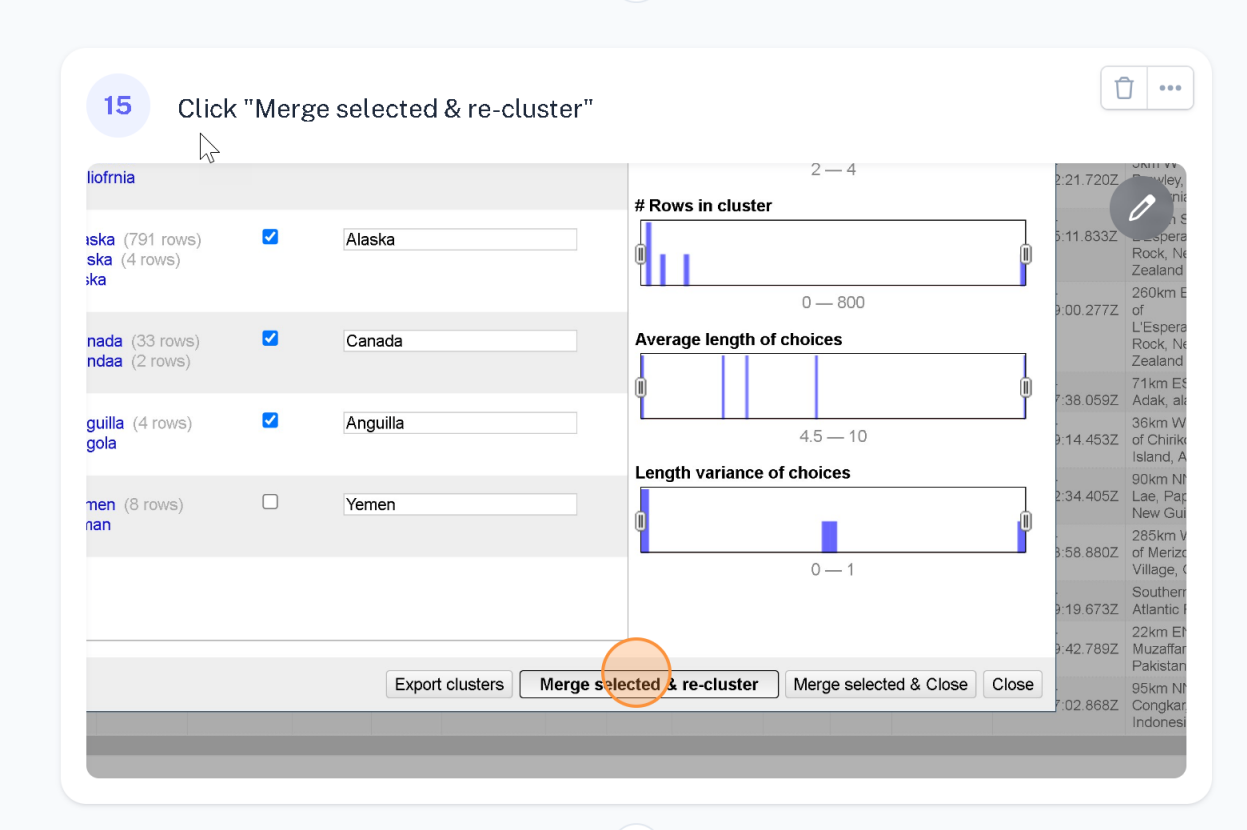
- Clustering: - this feature is designed to identify groups of cell values that may represent the same thing,





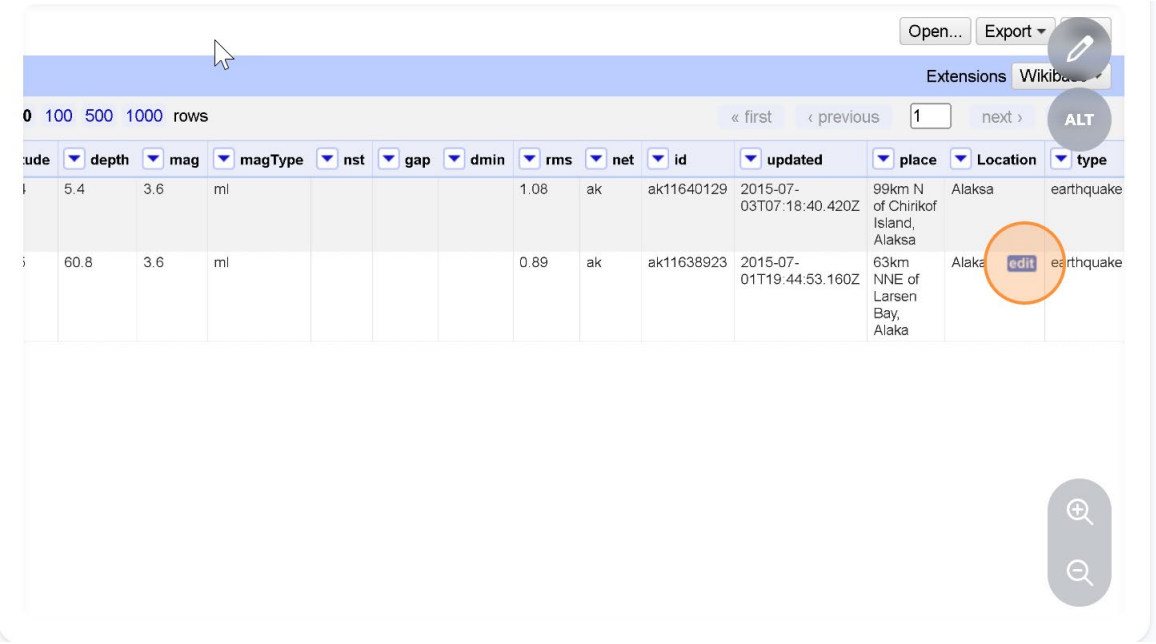
11. CLEANSING/WRANGLING DATA: EDITING CELLS





12. CLEANSING/WRANGLING DATA: MANUALLY EDITING CELL

- Go to Location, Click edit, and update the name.



ıe	latitude	longitude	depth	mag	magType	nst	gap	dmin	rms	net	id	updated	place	Location	
7-24:55.000Z	51.548	-175.7676	40.6	4.1	ml				0.75	ak	ak11639972	2015-07-03T02:27:38.059Z	71km ESE of Adak, alaska	Alaska	e
7-06:46.000Z	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	Alaska	e
7-03:23.790Z	53.1083	173.0287	23.49	4.8	mb edit		130	0.752	0.91	us	us10002mzc	2015-07-02T22:27:31.356Z	31km NNW of Attu Station, Alaska	Alaska	e
7-18:27.000Z	60.9657	-147.4064	23.8	3.7	ml				0.55	ak	ak11639524	2015-07-02T03:01:09.571Z	60km WSW of Valdez, alaska	Alaska	e
7-20:13.000Z	56.1872	-153.5558	12	4.3	ml				0.71	ak	ak11639423	2015-07-02T21:15:08.040Z	118km S of Larsen Bay,	Alaska	e