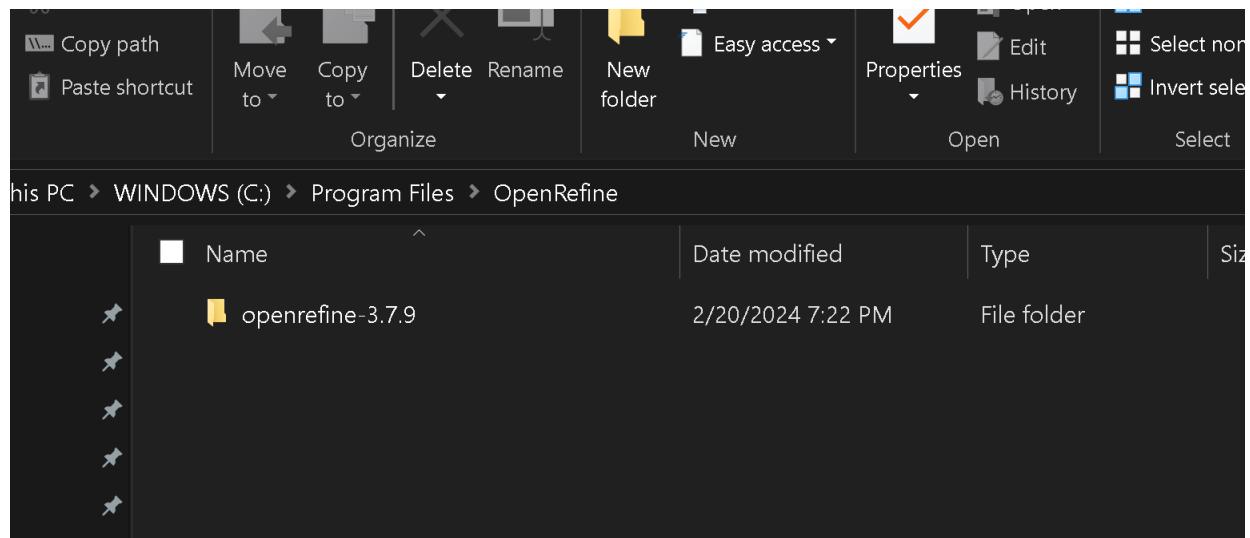


## MODULE 6: HOMEWORK 1: ASSIGNMENT: DATA PREP WITH OPENREFINE

### 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

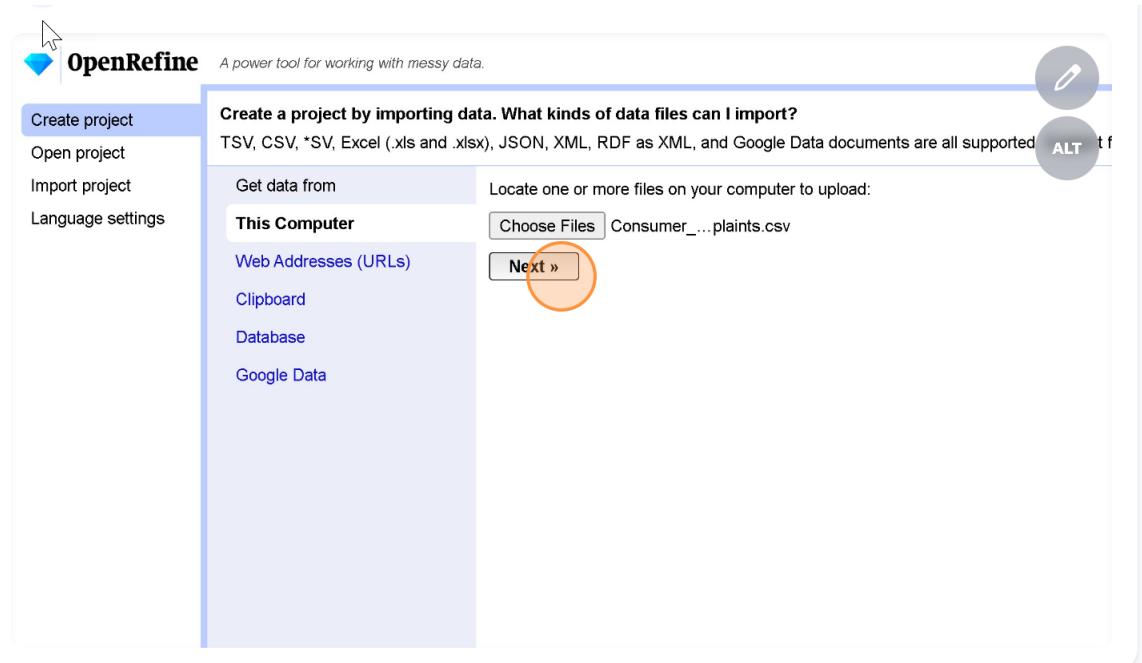
- OpenRefine automatically open to a browser window, or I can manually navigate to the application's URL (using <http://127.0.0.1:3333>).

A screenshot of the OpenRefine web interface. The URL in the browser is http://127.0.0.1:3333. On the left, there is a sidebar with links: 'Create project' (which is highlighted), 'Open project', 'Import project', and 'Language settings'. The main content area has a heading 'Create a project by importing data. What kinds of data files can I import?'. Below this, it says 'TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added'. There are two input fields: 'Get data from' (set to 'This Computer') and 'Locate one or more files on your computer to upload:' with a 'Choose Files' button. Below these is a 'Next »' button. To the right, there is a terminal-like window showing the server logs:

```
C:\Users\17034\Downloads\openrefine-win-3.7.9\openrefine-3.7.9\openrefine.exe
19:47:33.519 [           refine_server] Starting Server bound to '127.0.0.1:3333'
19:47:33.582 [           refine_server] Initializing context: '' from 'C:\Users\17034\Downloads\openrefine-win-3.7.9\webapp' (63ms)
19:47:34.521 [           refine_server] Creating new workspace directory C:\Users\17034\Downloads\openrefine-win-3.7.9\webapp\workspaces (1ms)
19:47:35.757 [           refine] Starting OpenRefine 3.7.9 [d6cd9e2]... (12ms)
19:47:35.757 [           refine] initializing FileProjectManager with dir C:\Users\17034\AppData\Roaming\OpenRefine\workspaces
19:47:35.757 [           refine] C:\Users\17034\AppData\Roaming\OpenRefine\workspaces (1ms)
19:47:35.776 [           FileProjectManager] Failed to load workspace from any attempted workspace (1ms)
19:47:44.086 [           refine] POST /command/core/load-language (8310ms)
19:47:44.179 [           refine] GET /command/core/get-preference (93ms)
19:47:44.270 [           refine] POST /command/core/load-language (91ms)
19:47:44.289 [           refine] POST /command/core/load-language (19ms)
19:47:44.305 [           refine] POST /command/core/load-language (16ms)
19:47:44.428 [           refine] GET /command/core/get-importing-configuration (1ms)
```

### 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



The screenshot shows a data grid in OpenRefine. The columns are labeled: Sub-issue, State, ZIP code, Submitted via, Date received, Date sent to company, Company, Company response, and Timely re. The data includes rows for various companies and their responses. A 'Create project »' button is circled in orange at the top right of the grid. At the bottom, there is a page number '8' and a 'Update preview' button.

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

Consumer Complaints [Permalink](#)

Redo 0 / 0 ALT

**filters** ◆

to select subsets  
Choose facet and  
a menu at the top

Started?  
**casts**

**384498 rows**

Show as: rows records Show: 5 0 25 50 100 500 1000 rows ALT

All	Complaint ID	Product	Sub-product	Issue	Sub-issue	State
1.	1354490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH
2.	1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ
3.	1355730	Credit reporting		Incorrect information on credit report	Account status	IL
4.	1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA
5.	1354249	Bank account or service	Checking account	Problems caused by my funds being low		AL
6.	1354326	Bank account or service	Checking account	Account opening, closing, or management		TX
7.	1351925	Bank account or service	Checking account	Account opening, closing, or management		FL
8.	1352573	Debt collection	Medical	Cont'd attempts collect debt not owed	Debt was paid	NV
9.	1354227	Debt collection	Medical	False statements or representation	Indicated committed crime not paying	FL

#### 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.

9 Click "Text facet"

Open... Export ALT

Extensions Wikipedia

0 100 500 1000 rows first < previous 1 next >

Product	Issue	Sub-issue	State	ZIP code	Submitted via	Date received	Date sent to company
Conf'd attempts collect debt not owed	Debt is not mine					5	04/30/2015
al	Dealing with my lender or servicer					5	04/30/2015
an	Incorrect information on credit report	Account status				5	04/30/2015
one,	Disclosure verification of debt	Right to dispute notice not received				5	04/30/2015
o, etc.)	Problems caused by my funds being low					5	04/30/2015
	Account opening, closing, or management					5	04/30/2015
	Account opening, closing, or management		FL	34677	Web	04/29/2015	04/29/2015
	Cont'd attempts collect debt not owed	Debt was paid	NV	89143	Web	04/29/2015	04/29/2015
	False statements or representation	Indicated committed crime not paying	FL	32792	Web	04/29/2015	04/30/2015

Facet ▶ Text facet ALT

Text filter Numeric facet

Edit cells Timeline facet

Edit column Scatterplot facet...

Transpose Custom text facet...

Sort... Custom numeric facet...

View Customized facets

Reconcile Web 04/30/2015 04/30/2015

**OpenRefine Consumer Complaints** [Permalink](#)

Facet / Filter Undo / Redo 0 / 0

Refresh Reset all Remove all

**384498 rows**

Show as: rows records Show: 5 10 25 50 100 500 1000 rows ALT

**State** change

62 choices Sort by: name count Cluster

AA	10
AE	143
AK	465
AL	3705
AP	110
AR	1604
AS	13
AZ	8435
CA	56952
CO	6590
CT	4664

All Complaint ID Product Sub-product Issue

1. 1354490 Debt collection Cont'd attempts collect debt not owed

2. 1355160 Student loan Non-federal student loan Dealing with my lender or servicer

3. 1355730 Credit reporting Incorrect information on credit report

4. 1355607 Debt collection Other (phone, health club, etc.) Disclosure verification of debt Right to dispute notice not received

5. 1354249 Bank account or service Checking account Problems caused by my funds being low

6. 1354326 Bank account or service Checking account Account opening, closing, or management

7. 1351925 Bank account or service Checking account Account opening, closing, or management

8. 1352573 Debt collection Medical Cont'd attempts collect debt not owed

9. 1354227 Debt collection Medical False statements or representation

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

First I was able 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.

Refresh Reset all Remove all

ZIP code change invert reset

24748 choices total, too many to display Set choice count limit

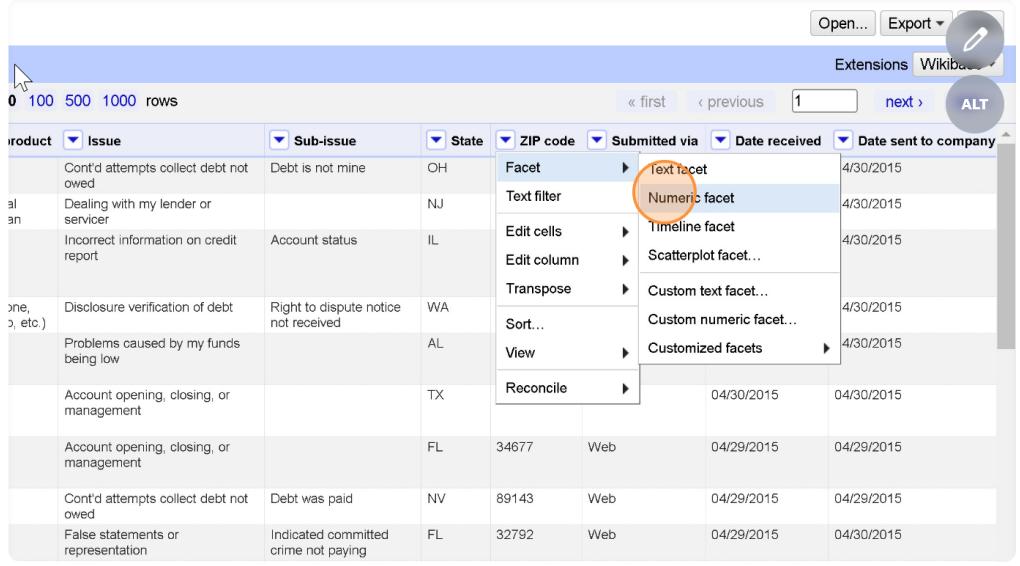
Facet by choice counts

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

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	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

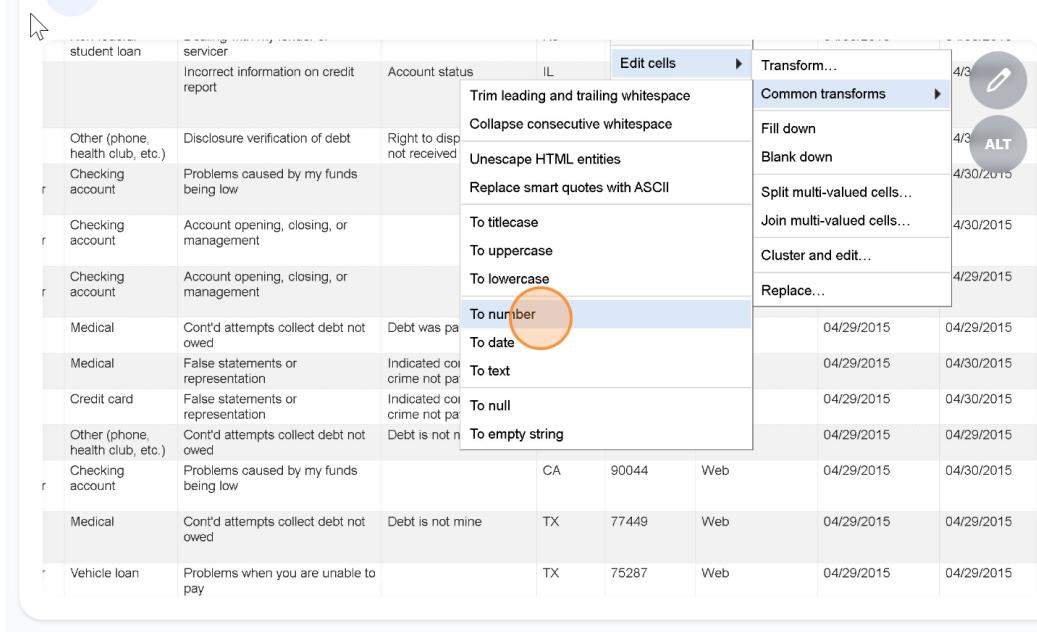
15 Click "Numeric facet"



A screenshot of a data grid interface. At the top, there are buttons for 'Open...', 'Export...', 'Extensions', 'Wikidata', and a pencil icon. Below the header, there are buttons for selecting row counts (0, 100, 500, 1000) and navigating through the data ('first', 'previous', 'next', 'last'). A search bar shows the number '1'. To the right of the search bar is a button labeled 'ALT' with a pencil icon. The main area contains a table with columns: Product, Issue, Sub-issue, State, ZIP code, Submitted via, Date received, and Date sent to company. A context menu is open over a cell in the 'Submitted via' column. The menu items are: Facet, Text facet, Numeric facet, Text filter, Timeline facet, Edit cells, Scatterplot facet..., Edit column, Sort..., Custom text facet..., Transpose, Custom numeric facet..., View, Customized facets, Reconcile, and Reconcile again. The 'Facet' item is the top item, and 'Numeric facet' is the second item under 'Facet', both are highlighted with a red circle.

Product	Issue	Sub-issue	State	ZIP code	Submitted via	Date received	Date sent to company
al	Cont'd attempts collect debt not owed	Debt is not mine	OH		Facet	Text facet	4/30/2015
an	Dealing with my lender or servicer		NJ			Numeric facet	4/30/2015
	Incorrect information on credit report	Account status	IL			Timeline facet	4/30/2015
one, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA			Scatterplot facet...	
	Problems caused by my funds being low		AL			Custom text facet...	4/30/2015
	Account opening, closing, or management		TX			Custom numeric facet...	
	Account opening, closing, or management		FL	34677	Web	04/29/2015	04/29/2015
	Cont'd attempts collect debt not owed	Debt was paid	NV	89143	Web	04/29/2015	04/29/2015
	False statements or representation	Indicated committed crime not paying	FL	32792	Web	04/29/2015	04/30/2015

17 Click "To number"



A screenshot of a data grid interface. At the top, there are buttons for 'Open...', 'Export...', 'Extensions', 'Wikidata', and a pencil icon. Below the header, there are buttons for selecting row counts (0, 100, 500, 1000) and navigating through the data ('first', 'previous', 'next', 'last'). A search bar shows the number '1'. To the right of the search bar is a button labeled 'ALT' with a pencil icon. The main area contains a table with columns: Product, Issue, Sub-issue, State, ZIP code, Submitted via, Date received, and Date sent to company. A context menu is open over a cell in the 'Submitted via' column. The menu items are: Edit cells, Transform..., Common transforms, Fill down, Blank down, Split multi-valued cells..., Join multi-valued cells..., Cluster and edit..., Replace..., Trim leading and trailing whitespace, Collapse consecutive whitespace, Unescape HTML entities, Replace smart quotes with ASCII, To titlecase, To uppercase, To lowercase, To number, To date, To text, To null, To null, To empty string, and To empty string. The 'Edit cells' item is the top item, and 'To number' is the second item under 'Edit cells', both are highlighted with a red circle.

Product	Issue	Sub-issue	State	ZIP code	Submitted via	Date received	Date sent to company
student loan	servicer	Incorrect information on credit report	Account status	IL	Edit cells	Transform...	4/30/2015
	Other (phone, health club, etc.)	Disclosure verification of debt	Right to disp not received			Common transforms	
Checking account	Problems caused by my funds being low					Fill down	4/30/2015
Checking account	Account opening, closing, or management					Blank down	
Checking account	Account opening, closing, or management					Split multi-valued cells...	4/30/2015
Medical	Cont'd attempts collect debt not owed	Debt was pa				Join multi-valued cells...	
Medical	False statements or representation	Indicated co crime not pa				Cluster and edit...	4/30/2015
Credit card	False statements or representation	Indicated co crime not pa				Replace...	
Other (phone, health club, etc.)	Cont'd attempts collect debt not owed	Debt is not n				04/29/2015	04/29/2015
Checking account	Problems caused by my funds being low		CA	90044	Web	04/29/2015	04/30/2015
Medical	Cont'd attempts collect debt not owed	Debt is not mine	TX	77449	Web	04/29/2015	04/29/2015
Vehicle loan	Problems when you are unable to pay		TX	75287	Web	04/29/2015	04/29/2015

Facet / Filter Undo / Redo 1 / 2

Refresh Reset all Remove all

**ZIP code** change invert reset

24748 choices total, too many to display  
Set choice count limit

acet by choice counts

**ZIP code** change reset

0 — 100,000

Numeric 380136  Non-numeric 0  Blank 4362

384498 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

All	Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code
1.	1354490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH	44077
2.	1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ	8807
3.	1355730	Credit reporting		Incorrect information on credit report	Account status	IL	60618
4.	1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133
5.	1354249	Bank account or service	Checking account	Problems caused by my funds being low		AL	36127
6.	1354326	Bank account or service	Checking account	Account opening, closing, or management		TX	78575
7.	1351925	Bank account or service	Checking account	Account opening, closing, or management		FL	34677
8.	1352573	Debt collection	Medical	Cont'd attempts collect debt not owed	Debt was paid	NV	89143
9.	1354227	Debt collection	Medical	False statements or representation	Indicated committed crime not paying	FL	32792
10.	1354200	Debt collection	Credit card	False statements or representation	Indicated committed crime not paying	AZ	85304
11.	1352929	Debt	Other (phone,...	Cont'd attempts collect debt not owed	Debt is not mine	NC	27534

## 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)

11 Click the "New column name" field.

Consumer Complaints Permalink

Redo 2 / 2

Reset all Remove all

many to display int limit

change invert reset

0 — 100,000

Numeric 380136  Non-numeric 0  Blank 4362

384498 rows

Add column based on column ZIP code

New column name

On error  set to blank  store error  copy value from original column

Expression Language General Refine Expression Language (GREL)

```
if(value.length() > 4, value, 99999)
```

Preview History Starred Help

row	value	if(value.length() > 4, value, ...)
1.	44077	44077
2.	8807	99999
3.	60618	60618
4.	98133	98133
-	- - - - -	- - - - -

384498 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

« first < previous 1 > next » last

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 checking account		Problems caused by my funds being low		AL	35127	35127	Web

## 7. DATA SET EQ2015: CLEANSING & WRANGLING DATA:

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

9 Click "Create project »"

Project name Earthquake 2015 Tags Create project »

nst	gap	dmin	rms	net	id	updated	place	type
			1.08	ak	ak11640129	2015-07-03T07:18:40.420Z	99km N of Chirikof Island, Alaska	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
			0.91	us	us10002n2x	2015-07-03T03:25:11.833Z	122km SE of L'Esperance Rock, New Zealand	earthquake
69	2.947		0.97	us	us10002n2l	2015-07-03T03:09:00.277Z	260km ESE of L'Esperance Rock, New Zealand	earthquake
234	3.483		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	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 Bienes, 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

SCII Update preview

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

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

Add column based on column place

New column name: Location

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

Expression: `if(value.split(",").length() < 2 , "Off-shore", value.split(",") [1])`

Language: General Refine Expression Language (GREL)

No syntax error.

Preview:

row	value	Location
1.	99km N of Chirikof Island, Alaska	Alaska
2.	1km ESE of Medford, Oklahoma	Oklahoma
3.	Southern Mid-Atlantic Ridge	Off-shore
4.	5km W of Brawley, California	California

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

0 100 500 1000 rows

« first < previous 1 next » ALT

id	depth	mag	magType	nst	gap	dmin	rms	net	id	updated	place	Location
1	5.4	3.6	ml				1.08	ak		Text facet	Facet	ksa
	5	3	mb_lg		46	0.185	0.29	us		Numeric facet	Text filter	
	10	4.8	mb		99	30.883	0.62	us		Timeline facet	Edit cells	ahoma
333	11.718	3.57	ml	67	63	0.0571	0.23	ci		Scatterplot facet...	Edit column	shore
	35	5	mb		69	2.947	0.91	us		Custom text facet...	Transpose	ifornia
										Custom numeric facet...	Sort...	
										Customized facets	View	ew Zealand
											Reconcile	
											Zearland	
	37.72	4.9	mb		234	3.483	0.97	us	us10002n2l	2015-07-03T03:09:27Z	260km ESE of L'Esperance Rock, New Zealand	New Zealand
	40.6	4.1	ml				0.75	ak	ak11639972	2015-07-03T02:27:38.059Z	71km ESE of Adak, Alaska	alaska
	41.2	3.3	ml				0.86	ak	ak11639884	2015-07-03T23:00:11.453Z	36km WNW of Chirikof	Alaska

Facet / Filter Undo / Redo 1 / 1 < 8708 rows Extensions Wikibase >

Refresh Reset all Remove all

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

All time latitude longitude depth mag magType nst gap dmin rms net id updated place Location

No numeric value present.

**nst** change reset

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 55km N of Onewa Island, Alaska

**Location** change Cluster

167 choices Sort by: name count

Afghanistan 77  
Alaa 1  
Alabama 6  
Alaka 1  
Alaksa 1  
alaska 4  
Alaska 791  
Albania 1  
Algeria 11  
Alksa 1  
Angola 1

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

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

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 112km SE of L'Esperance Rock, New Zealand

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

7. 2015-07-02T18:24:55.000Z 51.548 -175.7676 40.6 4.1 ml 0.75 ak ak11639972 2015-07-02T18:24:55.000Z 71km ESE of alaska

## 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.

OpenRefine Earthquake 2015 Permalink

Facet / Filter Undo / Redo 1 / 1 < 8708 rows ALT

Refresh Reset all Remove all

All time latitude longitude depth mag magType

No numeric value present.

**nst** change reset

**Location** change Cluster

167 choices Sort by: name count

Afghanistan 77  
Alaa 1  
Alabama 6  
Alaka 1  
Alaksa 1  
alaska 4  
Alaska 791  
Albania 1  
Algeria 11  
Alksa 1  
Angola 1

1. 2015-07-02T23:16:03.000Z 56.7152 -155.4884 5.4 3.6 ml

2. 2015-07-02T22:40:35.240Z 36.8015 -97.7167 5 3 mb\_lg

3. 2015-07-02T22:31:28.190Z -23.0587 -14.0431 10 4.8 mb

4. 2015-07-02T19:38:39.760Z 32.981 -115.5813333 11.718 3.57 ml 67

5. 2015-07-02T19:22:44.570Z -32.2014 -177.9748 35 5 mb

6. 2015-07-02T19:06:28.220Z -32.4952 -176.4412 37.72 4.9 mb

7. 2015-07-02T18:24:55.000Z 51.548 -175.7676 40.6 4.1 ml

8. 2015-07-02T18:06:46.000Z 55.9723 -156.1441 41.2 3.3 ml

**OpenRefine Earthquake 2015** Permalink

Facet / Filter Undo / Redo 1 / 1

Refresh Reset all Revert

nst

No numeric value present.

Location

167 choices Sort by: name count

Afghanistan	77
Alaa	1
Alabama	6
Alaka	1
Alaksa	1
alaska	4
Alaska	791
Albania	1
Algeria	11
Alksa	1
Angola	1

Cluster and edit column "Location"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method **Key collision** Keying function **Fingerprint**

Cluster size	Row count	Values in cluster	Merge?	New cell value
2	795	<ul style="list-style-type: none"> <li>Alaska (791 rows)</li> <li>alaska (4 rows)</li> </ul>	<input type="checkbox"/>	Alaska

Key collision Nearest neighbor

## 11. CLEANSING/WRANGLING DATA: EDITING CELLS

**OpenRefine Earthquake 2015** Permalink

Facet / Filter Undo / Redo 1 / 1

Refresh Reset all Revert

nst

No numeric value present.

Location

167 choices Sort by: name count

Afghanistan	77
Alaa	1
Alabama	6
Alaka	1
Alaksa	1
alaska	4
Alaska	791
Albania	1
Algeria	11
Alksa	1
Angola	1

Cluster and edit column "Location"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method **Nearest neighbor** Distance function **PPM** Radius **3** Block chars **6** 16 clusters found

Cluster size	Row count	Values in cluster	Merge?	New cell value
3	54	<ul style="list-style-type: none"> <li>Japan region (52 rows)</li> <li>Ecuador region</li> <li>India region</li> </ul>	<input type="checkbox"/>	Japan region
3	114	<ul style="list-style-type: none"> <li>British Virgin Islands (88 rows)</li> <li>U.S. Virgin Islands (25 rows)</li> <li>Cayman Islands</li> </ul>	<input type="checkbox"/>	British Virgin Islands
3	63	<ul style="list-style-type: none"> <li>Tajikistan (36 rows)</li> <li>Pakistan (25 rows)</li> <li>Uzbekistan (2 rows)</li> </ul>	<input type="checkbox"/>	Tajikistan
3	187	<ul style="list-style-type: none"> <li>Solomon Islands (161 rows)</li> <li>U.S. Virgin Islands (25 rows)</li> <li>Cayman Islands</li> </ul>	<input type="checkbox"/>	Solomon Islands
2	85	<ul style="list-style-type: none"> <li>California (84 rows)</li> <li>Califonia</li> </ul>	<input type="checkbox"/>	California

# Choices in cluster  
2 — 3

# Rows in cluster  
0 — 810

Average length of choices  
6 — 33

Length variance of choices  
0 — 11

Select all Deselect all Export clusters Merge selected & re-cluster Merge selected & Close Close

15

Click "Merge selected &amp; re-cluster"



liofrnia

Alaska

Canada

Anguilla

Yemen

# Rows in cluster

Average length of choices

Length variance of choices

Export clusters **Merge selected & re-cluster** Merge selected & Close Close

2.21.720Z ...  
5.11.833Z ...  
9.00.277Z ...  
7.38.059Z ...  
9.14.453Z ...  
2.34.405Z ...  
3.58.880Z ...  
9.19.673Z ...  
9.42.789Z ...  
7.02.868Z ...

## 12. CLEANSING/WRANGLING DATA: MANUALLY EDITING CELL

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

0 100 500 1000 rows

« first < previous 1 next » ALT

ude	depth	mag	magType	nst	gap	dmin	rms	net	id	updated	place	Location	type
1	5.4	3.6	ml				1.08	ak	ak11640129	2015-07-03T07:18:40.420Z	99km N of Chirikof Island, Alaska	Alaska	earthquake
2	60.8	3.6	ml				0.89	ak	ak11638923	2015-07-01T19:44:53.160Z	63km NNE of Larsen Bay, Alaska	Alaska	<b>edit</b> earthquake

Open... Export Extensions Wikidata

🔍 🔍

ie	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		
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		
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	
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		
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		