

## ADTA 5240 – DATA PREP WITH OPENREFINE

### PART I: Cleaning & Wrangling Data

Yog Chaudhary

Data wrangling, cleaning, or munging is about fixing and shaping data to answer questions. Analytics people spend most of their time (80%) fixing data and the rest (20%) analyzing it. It makes data easier to use by changing it in smart ways, like fixing a whole column of data or mixing columns together. If data is collected or shown poorly, we need to wrangle it. Data from people can be mistaken, and data from websites might not be set up well for analysis. This process helps fix these issues and makes the data ready for analysis.

**Remove Irrelevant Values:** The elimination of pointless values is among the most fundamental data-cleaning techniques used in data mining. You should erase all unnecessary data from your machine as soon as possible. You don't need any information that is pointless or unimportant. The context of your problem might not allow for it. This kind of data frequently does not fit the situation you are attempting to evaluate.

**Get rid of Duplicate values:** Getting rid of duplicate values in your data is important because they just take up space and don't help your analysis. These repeats can happen for many reasons, like combining data from different places, mistakes when entering data, or clicking "enter" too many times on a form. Duplicate values are a common issue in datasets. Finding and removing these duplicates, a process known as de-duplication is a key part of cleaning up your data. This helps make your dataset more accurate and easier to work with, saving you time in the long run.

**Uniformity of Language:** Uniformity of language in data is crucial, especially when using Natural Language Processing (NLP) models for analysis. These models usually work with just one language at a time. If your data has multiple languages, it could cause problems because the system might not understand all of them. So, when cleaning your data, make sure that all the information is in the same language. This ensures that your analysis is accurate and that the NLP models can process the data correctly.

**Avoid Typos:** Typos are common and can be fixed using different tools and methods to ensure all words are spelled correctly. It's important to correct them because models see different spellings and cases as different values. For example, "George" with an uppercase "G" is not the same as "George" with a lowercase "g". Fixing these mistakes helps in making your data more accurate and reliable for analysis.

Remember, the goal of data cleaning is to improve the quality and reliability of your data, making it easier to analyze and derive insights.

URL: <https://www.upgrad.com/blog/data-cleaning-techniques/>

## PART II: Exploring, Cleaning, Wrangling Datasets

### Data Prep with Open Refine

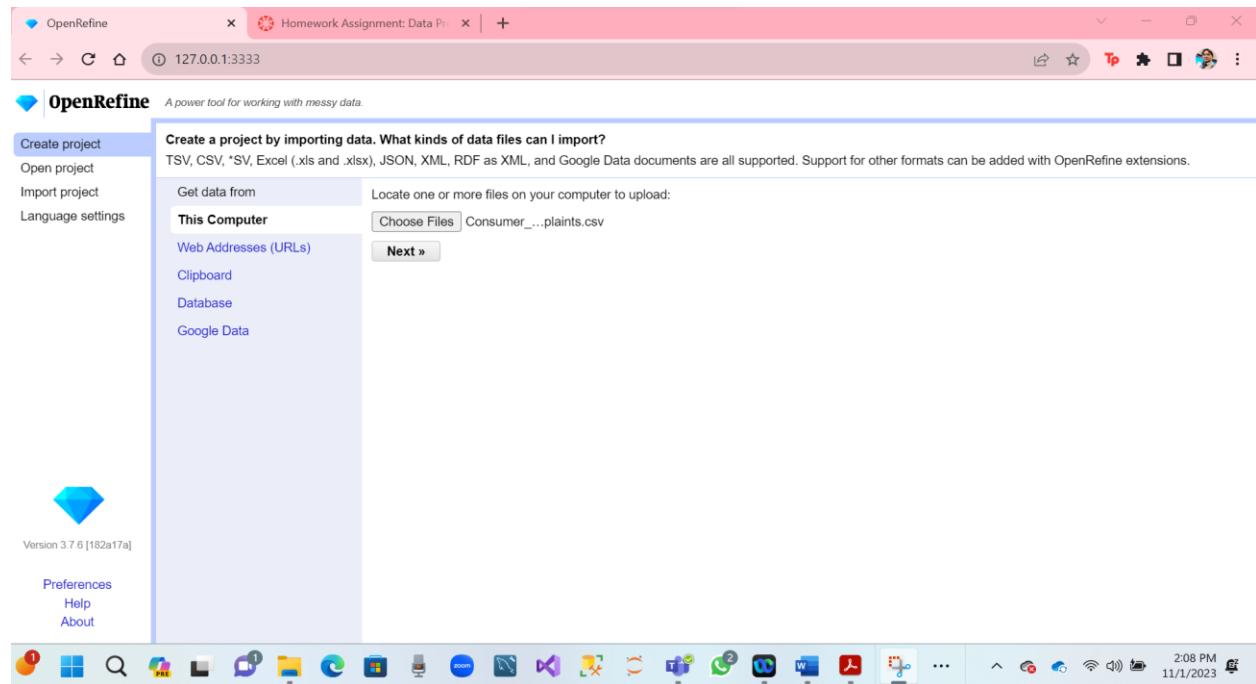
I have downloaded OpenRefine by using this link (<https://openrefine.org/docs/manual/installing>)

After downloading, I was taken to the OpenRefine page and showed a web-browser GUI pops up. (home screen: <http://127.0.0.1:3333/>)

#### Creating a new project and uploading data.

##### Consumer\_Complaints.csv: cleansing & Wrangling Data

- Here, I clicked on choose files.
- Then I browsed for Consumer\_Complaints.csv.
- I clicked on next.



- Hence the Consumer\_Complaints.csv dataset uploaded.
- After this, I was taken to the below page.

OpenRefine | Homework Assignment: Data Project | 127.0.0.1:3333

**OpenRefine** A power tool for working with messy data.

Create project | « start over | Configure parsing options | Project name: Consumer Complaints csv | Tags | Create project »

	Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	Submitted via	Date received	Date sent to company	Company
1.	1354490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH	44077	Web	04/30/2015	04/30/2015	Expert Global Solutions, Inc.
2.	1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ	8807	Web	04/30/2015	04/30/2015	Transworld Systems Inc.
3.	1355730	Credit reporting		Incorrect information on credit report	Account status	IL	60618	Web	04/30/2015	04/30/2015	FNIS (Fidelity National Information Services, Inc.)
4.	1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133	Web	04/30/2015	04/30/2015	Stellar Recovery Inc.

Parse data as: CSV / TSV / separator-based files | Character encoding: UTF-8 | Update preview | Disable auto preview

Version 3.7.6 [182a17a]

Preferences | Help | About

Windows taskbar: File Explorer, Edge, File, Task View, Start, Taskbar settings, Taskbar icons, Taskbar search, Taskbar pinned items, Taskbar notifications, Taskbar status, Taskbar language, Taskbar date/time.

- Here I left defaults as checked i.e.,
- OpenRefine detected that the data file is in CSV format.
- Line (Row) is a header line.
- Quotation marks are used to enclose cells containing column separators.
- Store blank rows.
- Store blank cells as nulls.
- Then I clicked on Create a project.
- After that I was taken to the below page.

Consumer Complaints csv - OpenRefine | Homework Assignment: Data Project | 127.0.0.1:3333/project?project=2646319051008

**OpenRefine** Consumer Complaints csv | Permanent

Facet / Filter | Undo / Redo 0 / 0 | 384498 rows | Extensions: Wikibase | Export | Help

Using facets and filters | Watch these screencasts

Show as: rows records | Show: 5 10 25 50 100 500 1000 rows | first | previous | next | last | 1 |

	Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	Submitted via	Date received	Date sent to company
1.	1354490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH	44077	Web	04/30/2015	04/30/2015
2.	1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ	8807	Web	04/30/2015	04/30/2015
3.	1355730	Credit reporting		Incorrect information on credit report	Account status	IL	60618	Web	04/30/2015	04/30/2015
4.	1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133	Web	04/30/2015	04/30/2015
5.	1354249	Bank account or service	Checking account	Problems caused by my funds being low		AL	35127	Web	04/30/2015	04/30/2015
6.	1354326	Bank account or service	Checking account	Account opening, closing, or management		TX	78575	Web	04/30/2015	04/30/2015
7.	1351925	Bank account or service	Checking account	Account opening, closing, or management		FL	34677	Web	04/29/2015	04/29/2015
8.	1352573	Debt collection	Medical	Cont'd attempts collect debt not owed	Debt was paid	NV	89143	Web	04/29/2015	04/29/2015

Windows taskbar: File Explorer, Edge, File, Task View, Start, Taskbar settings, Taskbar icons, Taskbar search, Taskbar pinned items, Taskbar notifications, Taskbar status, Taskbar language, Taskbar date/time.

- Here I have seen 384,498 rows that are loaded.
- After I clicked on “50” where OpenRefine shows 50 rows at once.
- If we want to see other rows(records) we can click next.

## 2. Check states with Text Facets:

- Here we are analyzing the data: we will use “Text Facet to display the number of occurrences of each unique value in a column.
- It is like a filter in Excel.

The screenshot shows the OpenRefine interface with a large dataset of 384,498 rows. The columns include Complaint ID, Product, Sub-product, Issue, Sub-issue, State, ZIP code, Submitted via, Date received, and Date sent to consumer. A context menu is open over the 'State' column for the entry 'AL'. The menu path 'Facet' -> 'Text facet' is highlighted. Other options in the menu include 'Numeric facet', 'Timeline facet', 'Scatterplot facet...', 'Sort...', 'View', and 'Customized facets'.

Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	Submitted via	Date received	Date sent to consumer
1. 1354490	Debt collection		Cont'd attempts collect debt not owed		Debt is not mine				04/30/2015
2. 1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer						04/30/2015
3. 1355730	Credit reporting		Incorrect information on credit report		Account status				04/30/2015
4. 1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received					04/30/2015
5. 1354249	Bank account or service	Checking account	Problems caused by my funds being low			AL	35127	Web	04/30/2015
6. 1354326	Bank account or service	Checking account	Account opening, closing, or management		TX	78575	Web	04/30/2015	04/30/2015
7. 1351925	Bank account or service	Checking account	Account opening, closing, or management		FL	34677	Web	04/29/2015	04/29/2015
8. 1352573	Debt collection	Medical	Cont'd attempts collect debt not owed	Debt was paid	NV	89143	Web	04/29/2015	04/29/2015

- Here I have clicked on the drop-down menu arrow of state.
- Then I clicked on Facet and Text Facet, after clicking on that I was taken to the below page.

The screenshot shows the OpenRefine interface with the following details:

- Facet / Filter**: A sidebar on the left showing a facet for "State" with 62 choices. Choices include AA 10, AE 143, AK 485, AL 3705, AP 110, AR 1604, AS 13, AZ 8435, CA 56952, CO 6590, CT 4664, DC 2223, DE 2042, FL 36946, FM 21, GA 16616, GU 40, HI 1379, IA 1565, and ID 1275.
- Table View**: The main area displays 384498 rows of data. The columns are: All, Complaint ID, Product, Sub-product, Issue, Sub-issue, State, ZIP code, Submitted via, Date received, and Date sent to consumer.
- Toolbar**: At the top right, there are buttons for Open..., Export, Help, Extensions, and Wikibase.
- Taskbar**: At the bottom, it shows the Windows taskbar with various pinned icons and the system clock indicating 2:33 PM on 11/1/2023.

- Here I have seen 62 states (51 states and other US territories) that are listed in the panel.
- Then I clicked on remove all to clear the result panel.

### 3. Wrangling/Munging – Transforming Data: Checking Zip code.

#### a. Text Facet for Zip code:

- I clicked on drop-down menu arrow of the zip code.
- Then click on Facet and Text Facet.
- After clicking on Text Facet, I was taken to the below page.

Consumer Complaints csv - Open | Homework Assignment: Data Pr... | +

127.0.0.1:3333/project?project=2646319051008

OpenRefine Consumer Complaints csv Permalink

Facet / Filter Undo / Redo 0 / 0 384498 rows

Refresh Reset all Remove all Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first < previous 1 next » last »

	Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	Submitted via	Date received	Date sent to consumer
1.	135490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH				4/30/2015
2.	1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ				4/30/2015
3.	1355730	Credit reporting		Incorrect information on credit report	Account status	IL				4/30/2015
4.	1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA				4/30/2015
5.	1354249	Bank account or service	Checking account	Problems caused by my funds being low		AL	35127	Web	04/30/2015	04/30/2015
6.	1354326	Bank account or service	Checking account	Account opening, closing, or management		TX	78575	Web	04/30/2015	04/30/2015
7.	1351925	Bank account or service	Checking account	Account opening, closing, or management		FL	34677	Web	04/29/2015	04/29/2015
8.	1352573	Debt collection	Medical	Cont'd attempts collect debt not owed	Debt was paid	NV	89143	Web	04/29/2015	04/29/2015

Extensions Wikibase

Open... Export Help

Facet Text facet  
Text filter Numeric facet  
Edit cells Timeline facet  
Edit column Scatterplot facet...  
Transpose Custom text facet...  
Sort... Custom numeric facet...  
View Customized facets  
Reconcile

2:38 PM 11/1/2023

Consumer Complaints csv - Open | Homework Assignment: Data Pr... | +

127.0.0.1:3333/project?project=2646319051008

OpenRefine Consumer Complaints csv Permalink

Facet / Filter Undo / Redo 0 / 0 384498 rows

Refresh Reset all Remove all Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first < previous 1 next » last »

	Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	Submitted via	Date received	Date sent to consumer
1.	135490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH	44077	Web	04/30/2015	04/30/2015
2.	1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ	8807	Web	04/30/2015	04/30/2015
3.	1355730	Credit reporting		Incorrect information on credit report	Account status	IL	60618	Web	04/30/2015	04/30/2015
4.	1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133	Web	04/30/2015	04/30/2015
5.	1354249	Bank account or service	Checking account	Problems caused by my funds being low		AL	35127	Web	04/30/2015	04/30/2015
6.	1354326	Bank account or service	Checking account	Account opening, closing, or management		TX	78575	Web	04/30/2015	04/30/2015
7.	1351925	Bank account or service	Checking account	Account opening, closing, or management		FL	34677	Web	04/29/2015	04/29/2015
8.	1352573	Debt collection	Medical	Cont'd attempts collect debt not owed	Debt was paid	NV	89143	Web	04/29/2015	04/29/2015

Extensions Wikibase

Open... Export Help

Facet Text facet  
Text filter Numeric facet  
Edit cells Timeline facet  
Edit column Scatterplot facet...  
Transpose Custom text facet...  
Sort... Custom numeric facet...  
View Customized facets  
Reconcile

2:40 PM 11/1/2023

- Here the result shows there are 24,748 zip codes in the dataset.
- Some of the zip codes are missing i.e., blank, like in row 57.

## b. Numeric Facet:

- I clicked on the drop-down menu arrow of the Zip code.
- Then click on Facet and Numeric Facet.
- After clicking on that I was taken to the below screenshots.

Consumer Complaints csv - Open | Homework Assignment: Data Pr... | +

127.0.0.1:3333/project?project=2646319051008

OpenRefine Consumer Complaints csv Permalink

Facet / Filter Undo / Redo 0 / 0

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 Submitted via Date received Date sent to

1. 1354490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH	44077	Web	04/30/2015	04/30/2015
2. 1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ	8807	Web	04/30/2015	04/30/2015
3. 1355730	Credit reporting		Incorrect information on credit report	Account status	IL	60618	Web	04/30/2015	04/30/2015
4. 1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133	Web	04/30/2015	04/30/2015
5. 1354249	Bank account or service	Checking account	Problems caused by my funds being low		AL	35127	Web	04/30/2015	04/30/2015
6. 1354326	Bank account or service	Checking account	Account opening, closing, or management		TX	78575	Web	04/30/2015	04/30/2015
7. 1351925	Bank account or service	Checking account	Account opening, closing, or management		FL	34677	Web	04/29/2015	04/29/2015
8. 1352573	Debt collection	Medical	Cont'd attempts collect debt not owed	Debt was paid	NV	89143	Web	04/29/2015	04/29/2015

Extensions Wikibase

Open... Export Help

Facet / Filter Undo / Redo 0 / 0

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 Submitted via Date received Date sent to

1. 1354490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH	44077	Web	04/30/2015	04/30/2015
2. 1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ	8807	Web	04/30/2015	04/30/2015
3. 1355730	Credit reporting		Incorrect information on credit report	Account status	IL	60618	Web	04/30/2015	04/30/2015
4. 1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133	Web	04/30/2015	04/30/2015
5. 1354249	Bank account or service	Checking account	Problems caused by my funds being low		AL	35127	Web	04/30/2015	04/30/2015
6. 1354326	Bank account or service	Checking account	Account opening, closing, or management		TX	78575	Web	04/30/2015	04/30/2015
7. 1351925	Bank account or service	Checking account	Account opening, closing, or management		FL	34677	Web	04/29/2015	04/29/2015
8. 1352573	Debt collection	Medical	Cont'd attempts collect debt not owed	Debt was paid	NV	89143	Web	04/29/2015	04/29/2015

Extensions Wikibase

Open... Export Help

- Here there is no numeric zip code in the data set, which means all of them are included as text values (24748 zip codes).

### C. Transform zip code from Text to Numeric (Wrangling Data).

It is the process of cleaning and unifying messy and complex data sets for easy access and analysis.

- I clicked on the drop-down menu arrow of the zip code.
- Then click to select Edit cells.
- Then click on Common transforms and To number.
- After clicking on that I was taken to the below page.

Consumer Complaints csv - Open | Homework Assignment: Data Pr... | +

127.0.0.1:3333/project?project=2646319051008

OpenRefine Consumer Complaints csv Permalink

Facet / Filter Undo / Redo 0 / 0

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 Submitted via Date received Date sent to

1. 1354490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH	44077	Web	04/30/2015	04/30/2015
2. 1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ	8807	Web	04/30/2015	04/30/2015
3. 1355730	Credit reporting		Incorrect information on credit report	Account status	IL	60618	Web	04/30/2015	04/30/2015
4. 1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133	Web	04/30/2015	04/30/2015
5. 1354249	Bank account or service	Checking account	Problems caused by my funds being low		AL	35127	Web	04/30/2015	04/30/2015
6. 1354326	Bank account or service	Checking account	Account opening, closing, or management		TX	78575	Web	04/30/2015	04/30/2015
7. 1351925	Bank account or service	Checking account	Account opening, closing, or management		FL	34677	Web	04/29/2015	04/29/2015
8. 1352573	Debt collection	Medical	Cont'd attempts collect debt not owed	Debt was paid	NV	89143	Web	04/29/2015	04/29/2015

Extensions Wikibase

Open... Export Help

Facet / Filter Undo / Redo 0 / 0

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 Submitted via Date received Date sent to

1. 1354490	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	OH	44077	Web	04/30/2015	04/30/2015
2. 1355160	Student loan	Non-federal student loan	Dealing with my lender or servicer		NJ	8807	Web	04/30/2015	04/30/2015
3. 1355730	Credit reporting		Incorrect information on credit report	Account status	IL	60618	Web	04/30/2015	04/30/2015
4. 1355607	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Right to dispute notice not received	WA	98133	Web	04/30/2015	04/30/2015
5. 1354249	Bank account or service	Checking account	Problems caused by my funds being low		AL	35127	Web	04/30/2015	04/30/2015
6. 1354326	Bank account or service	Checking account	Account opening, closing, or management		TX	78575	Web	04/30/2015	04/30/2015
7. 1351925	Bank account or service	Checking account	Account opening, closing, or management		FL	34677	Web	04/29/2015	04/29/2015
8. 1352573	Debt collection	Medical	Cont'd attempts collect debt not owed	Debt was paid	NV	89143	Web	04/29/2015	04/29/2015

Extensions Wikibase

Open... Export Help

- In the above screenshot, results show 380136 zip code text values have transformed into numeric values and results show 4362 zip codes are missing.

#### 4. Cleansing/Wrangling Data: Handling Missing Zip Codes

##### a. Fill down the missing value:

- One way of filling in missing values is taking the previous value and using that to set subsequent empty cells.
- Now looking at row 57, the value of the zip code is missing.
- I clicked on the drop-down menu arrow of the zip code.
- Then click on edit cells and fill down.
- After clicking on that I was taken to the below page.

The screenshot shows the OpenRefine interface with a table titled "384498 rows". The columns include Complaint ID, Product, Sub-product, Issue, Sub-issue, State, ZIP code, Submitted via, and Date. A facet for "ZIP code" is visible on the left, showing a histogram and counts for numeric, non-numeric, and blank values. The table lists several rows of consumer complaints, such as row 51 (Credit reporting, Other (phone, health club, etc.), Incorrect information on credit report, Information is not mine, CA, 92869, Web, 04/28/201), row 52 (Debt collection, Credit card, Conf'd attempts collect debt not owed, Debt is not mine, CA, 92129, Web, 04/28/201), and row 57 (Debt collection, Payday loan, False statements or representation, Attempted to collect wrong amount, CA, 90082, Web, 04/28/201).

- The results show there are no missing zip codes now which means all missing values are filled.
- Now when we look at row 57, here the zip code value is 90082, which was filled from row 56.
- In row 57, the state is NV, not CA.
- Can't be filled down to handle the missing values of zip code.
- Must rescind the wrangling data.

#### Rescind the wrangling data:

- OpenRefine has very advanced UNDO functionalities.
- For undoing the previous step, I clicked on undo as shown in the below page.
- UNDO lists 3 levels of activities 0,1,2 which can be rolled back.
- the last level 'fill down 4362 cells in column zip code' was highlighted.
- Then I clicked on "Text Transform on 380136." as the latest activity.
- By clicking on this the zip code in row 57 was undone as shown below.

Consumer Complaints csv - Open | Homework Assignment: Data Pr... | +

127.0.0.1:3333/project?project=2646319051008

### OpenRefine Consumer Complaints csv Permalink

Facet / Filter Undo / Redo 2 / 2 384498 rows Show as: rows records Show: 5 10 25 50 100 500 1000 rows Extensions Wikibase ▾

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

**ZIP code** change reset  
0 — 100,000  
Numeric 380136 Non-numeric 0 Blank 4362 Error 0

Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	Submitted via	Date received
51. 1349812	Credit reporting		Incorrect information on credit report	Information is not mine	CA	92869	Web	04/28/2015
52. 1351490	Debt collection	Other (phone, health club, etc.)	Cont'd attempts collect debt not owed	Debt is not mine	CA	92129	Web	04/28/2015
53. 1352036	Debt collection	Credit card	Improper contact or sharing of info	Talked to a third party about my debt	PA	15214	Web	04/28/2015
54. 1352128	Debt collection	Credit card	Cont'd attempts collect debt not owed	Debt was paid	CA	93729	Web	04/28/2015
55. 1352057	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	AL	35404	Web	04/28/2015
56. 1352071	Debt collection	Payday loan	False 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		Web	04/28/2015
58. 1352133	Debt collection	Other (phone, health club, etc.)	Communication tactics	Frequent or repeated calls	WA	98498	Web	04/28/2015
59. 1349777	Credit reporting		Incorrect information on credit report	Account terms	OH	44056	Web	04/28/2015
60. 1353626	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Not given enough info to verify debt	WI	54768	Web	04/28/2015
61. 1354096	Debt collection	Medical	Taking/threatening an illegal action	Threatened arrest/jail if do not pay	TX	77385	Web	04/28/2015
62. 1349410	Credit reporting		Incorrect information on credit report	Account status	DE	19968	Web	04/27/2015

3:09 PM 11/1/2023

## b. Handling Missing Zip codes: Create a New Column.

- I clicked on the drop-down menu arrow of the zip code.
- Then click on the edit column and add a column based on this column.
- After clicking on that I was taken to the below page.

Consumer Complaints csv - Open | Homework Assignment: Data Pr... | +

127.0.0.1:3333/project?project=2646319051008

### OpenRefine Consumer Complaints csv Permalink

Facet / Filter Undo / Redo 2 / 2 384498 rows Show as: rows records Show: 5 10 25 50 100 500 1000 rows Extensions Wikibase ▾

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

**ZIP code** change reset  
0 — 100,000  
Numeric 380136 Non-numeric 0 Blank 4362 Error 0

Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	Submitted via	Date received
48. 1351426	Debt collection		Cont'd attempts collect debt not owed	Debt is not mine	ME		Facet	04/28/2015
49. 1352120	Debt collection	Other (phone, health club, etc.)	Communication tactics	Threatened to take legal action	FL		Text filter	04/28/2015
50. 1352031	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Split Into several columns...			Edit cells	04/28/2015
51. 1349812	Credit reporting		Incorrect information on report	Join columns...			Edit column	04/28/2015
52. 1351490	Debt collection	Other (phone, health club, etc.)	Cont'd attempts collect debt not owed	Add column based on this column...			Transpose	04/28/2015
53. 1352036	Debt collection	Credit card	Improper contact or sharing of info	Add column by fetching URLs...			Sort...	04/28/2015
54. 1352128	Debt collection	Credit card	Cont'd attempts collect debt not owed	Add columns from reconciled values...			View	04/28/2015
55. 1352057	Debt collection		Cont'd attempts collect debt not owed	Rename this column...			Reconcile	04/28/2015
56. 1352071	Debt collection	Payday loan	False statements or representation	Remove this column				
57. 1353702	Debt collection		Communication tactics	Move column to beginning				
58. 1352133	Debt collection	Other (phone, health club, etc.)	Communication tactics	Move column to end				
59. 1349777	Credit reporting		Incorrect information on credit report	Move column left				
60. 1353626	Debt collection	Other (phone, health club, etc.)	Disclosure verification of debt	Move column right				

3:12 PM 11/1/2023

Consumer Complaints csv - Open | Homework Assignment: Data Project | +

127.0.0.1:3333/project?project=2646319051008

**OpenRefine Consumer Complaints csv**

Add column based on column ZIP code

New column name: ZipCode5

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")
```

No syntax error.

Preview:

row	value
1.	44077
2.	8807
3.	60618
4.	98133
5.	35127
6.	78575

OK Cancel

Facet / Filter Undo / Redo 2 / 2

ZIP code change invert reset

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

Facet by choice counts

ZIP code change reset

Numeric 380136 Non-numeric 0 Blank 4362

Error 0

Extensions Wikibase

ZIP code Submitted via Date received

ZIP code	Submitted via	Date received
4210	Web	04/28/2015
33132	Web	04/28/2015
76665	Web	04/28/2015
92869	Web	04/28/2015
92120	Web	04/28/2015
15214	Web	04/28/2015
93729	Web	04/28/2015
35404	Web	04/28/2015
90802	Web	04/28/2015
	Web	04/28/2015
98498	Web	04/28/2015
44056	Web	04/28/2015
54768	Web	04/28/2015

3:16 PM 11/1/2023

- For performing cleansing/wrangling/transforming data we use a scripting language.
- Here I have given the column name ZipCode5 and entered.
- If(value.length()>4, value, "99999").
- Then I clicked on OK.
- After clicking on I was redirected to the below page.

Consumer Complaints csv - Open | Homework Assignment: Data Project | +

127.0.0.1:3333/project?project=2646319051008

**OpenRefine Consumer Complaints csv**

384498 rows

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

« first < previous 1 next » last »

Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	ZipCode5	Submitted via
1. 1354490	Debt collection		Conf'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
6. 1354326	Bank account or service	Checking account	Account opening, closing, or management		TX	78575	78575	Web
7. 1351925	Bank account or service	Checking account	Account opening, closing, or management		FL	34677	34677	Web
8. 1352573	Debt collection	Medical	Conf'd attempts collect debt not owed	Debt was paid	NV	89143	89143	Web
9. 1354227	Debt collection	Medical	False statements or representation	Indicated committed crime not paying	FL	32792	32792	Web
10. 1354200	Debt collection	Credit card	False statements or representation	Indicated committed crime not paying	AZ	85304	85304	Web
11. 1352929	Debt collection	Other (phone, health club, etc.)	Conf'd attempts collect debt not owed	Debt is not mine	NC	27534	27534	Web
12. 1354191	Bank	Checking	Problems caused by my funds		CA	90044	90044	Web

Facet / Filter Undo / Redo 3 / 3

ZIP code change invert reset

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

Facet by choice counts

ZIP code change reset

Numeric 380136 Non-numeric 0 Blank 4362

Error 0

Extensions Wikibase

3:19 PM 11/1/2023

- Here the results show a new column zip code5 has been created.
- When I looked at row 2 the original zip code has 4 digits and a new one has 5 digits (99999), however, it justified in left. But this value is not numeric.
- Then I must undo the operation and start over.
- Again, I clicked on the drop-down menu arrow of the zip code.
- Then click on the edit column and add a column based on this column.
- After clicking on that I was taken to the below page.
- Here I have given column name ZipCode5 and
- If(value.length()>4, value, 99999).
- Then I click on OK.
- After clicking on OK I was taken to the below page.

**OpenRefine Consumer Complaints csv**

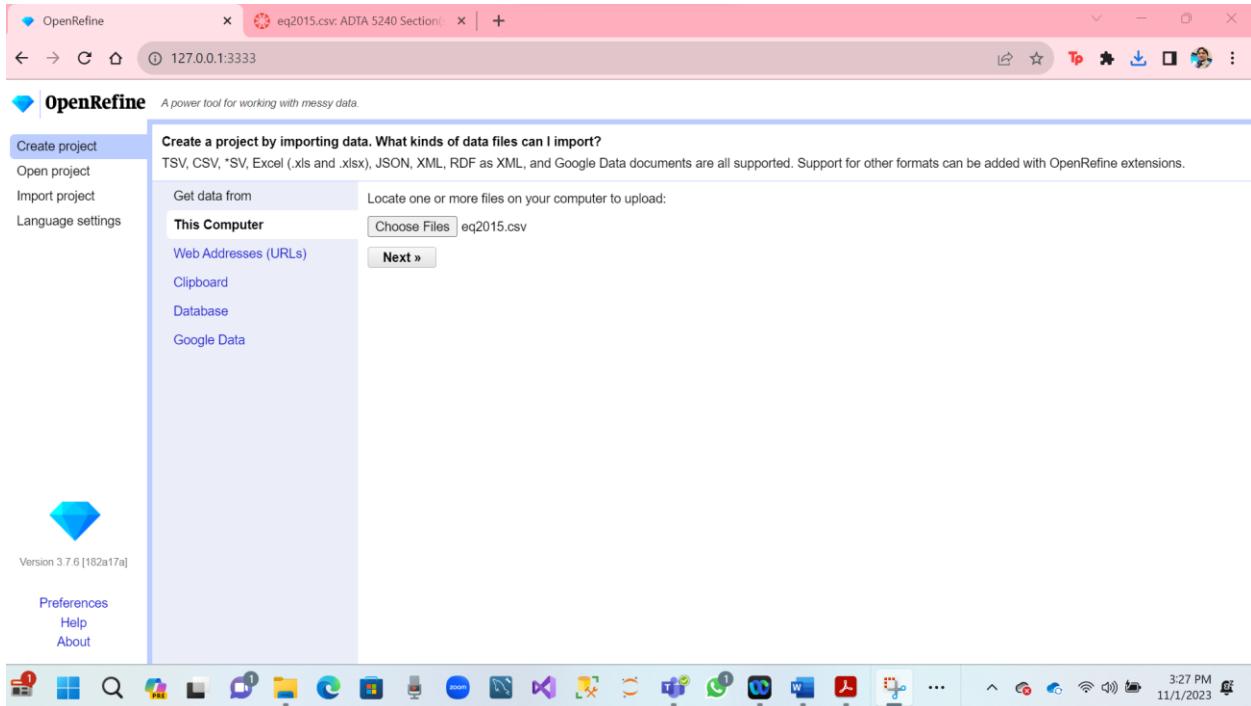
384498 rows

Complaint ID	Product	Sub-product	Issue	Sub-issue	State	ZIP code	ZipCode5	Submitted via
1. 1354490	Debt collection		Conf'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
6. 1354326	Bank account or service	Checking account	Account opening, closing, or management		TX	78575	78575	Web
7. 1351925	Bank account or service	Checking account	Account opening, closing, or management		FL	34677	34677	Web
8. 1352573	Debt collection	Medical	Conf'd attempts collect debt not owed	Debt was paid	NV	89143	89143	Web
9. 1354227	Debt collection	Medical	False statements or representation	Indicated committed crime not paying	FL	32792	32792	Web
10. 1354200	Debt collection	Credit card	False statements or representation	Indicated committed crime not paying	AZ	85304	85304	Web
11. 1352929	Debt collection	Other (phone, health club, etc.)	Conf'd attempts collect debt not owed	Debt is not mine	NC	27534	27534	Web
12. 1354191	Bank	Checking	Problems caused by my funds		CA	90044	90044	Web

- Here the results show a new column zip code5 has been created.
- When I looked at row 2 the original zip code has 4 digits and a new one has 5 digits(99999), however, it justified to right.

#### ❖ Data set eq2015: cleansing & Wrangling Data: Getting started.

- Here I clicked on Choose files.
- Then I browsed for eq2015.csv.
- Then I clicked on next



- Hence the eq2015.csv dataset has been uploaded.
- After this I was taken to the below page.

time	latitude	longitude	depth	mag	magType	nst	gap	dmin	rms	net	id	updated	place	type
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, Alaska	earthquake
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
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
2015-07-02T19:38:39.760Z	32.981	-115.5813333	11.718	3.57	ml	67	63	0.0571	0.23	cl	ci37196663	2015-07-02T20:42:21.720Z	5km W of Brawley, California	earthquake
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
2015-07-02T19:22:44.570Z	-32.4952	-176.4412	37.72	4.9	mb		234	3.483	0.97	us	us10002n2l	2015-07-03T03:25:11.833Z	260km ESE of L'Esperance Rock,	earthquake

- Then I left defaults as checked and clicked on create project.
- By clicking on create project I was taken to the below page.

- Then I clicked on 50, and it showed me 50 rows.
- In the above page “nst” column was missing few values.

## 1. Wrangling/Munging Data: Transforming the place column:

Transforming the data:

- Here I am going to extract to approximate area from the “place” column.
- I clicked on the drop-down menu arrow of place.
- Then click on the edit column and add a column based on this column.
- I was redirected to the below page.

Screenshot of OpenRefine interface showing the "Add column based on column place" dialog.

New column name: Location

On error: set to blank

Expression: value.split(',')[1]

Preview:

row	value	value.split(',')[1]
1.	99km N of Chirikof Island, Alaska	Alaska
2.	1km ESE of Medford, Oklahoma	Oklahoma
3.	Southern Mid-Atlantic Ridge	Error: java.lang.ArrayIndexOutOfBoundsException
4.	5km W of Brawley, California	California
5.	122km SE of L'Esperance Rock, New Zealand	New Zealand
6.	260km ESE of L'Esperance Rock, New Zealand	New Zealand

OK Cancel

- I have given “location” for the name of a new column.
- I have given,
- Value. split(',') [1] in expression
- Then click on OK.
- We use “,” for splitting the two substrings.
- After clicking on ok I have taken to the below page

Screenshot of OpenRefine interface showing the main data view.

8708 rows

id	place	Location	type
52	99km N of Chirikof Island, Alaska	Alaska	earthquake
15	1km ESE of Medford, Oklahoma	Oklahoma	earthquake
587	Southern Mid-Atlantic Ridge		earthquake
1	5km W of Brawley, California	California	earthquake
114	122km SE of L'Esperance Rock, New Zealand	New Zealand	earthquake
152	260km ESE of L'Esperance Rock, New Zealand	New Zealand	earthquake
8	71km ESE of Adak, alaska	alaska	earthquake
23	36km WNW of Chirikof Island, Alaska	Alaska	earthquake

- In this we can see a problem in row 3, which means not all the values have 2 strings separated by ",".
- So, to handle this we need to change the expression.
- Then click on edit column and add column based on this column.
- Then I have entered,
- If(value.split(",").length()<2, "offshore", value.split(",")[1]) for expression.
- After clicking I was redirected back to the below page.

The screenshot shows the OpenRefine interface with a dataset titled 'eq2015.csv'. The main workspace displays a table with 8708 rows, showing columns such as latitude, longitude, depth, mag, magType, nst, gap, dmin, rms, net, id, updated, place, Location, and type. A sidebar on the left shows a step-by-step process: 'Create project' and 'Create new column Location based on column place by filling 7,894 rows with gref:value.split(',')[1]'. The bottom of the screen shows a taskbar with various icons and the system clock indicating 3:58 PM on 11/1/2023.

Index	place	Location	type
52	90km N of Chirikof Island, Alaska	Alaska	earthquake
15	1km ESE of Medford, Oklahoma	Oklahoma	earthquake
587	Southern Mid-Atlantic Ridge		earthquake
1	5km W of Bodega Bay, California	California	earthquake
114	122km SE of 'Esperance Rock, New Zealand'	New Zealand	earthquake
352	260km ESE of 'Esperance Rock, New Zealand'	New Zealand	earthquake
8	71km ESE of Adak, Alaska	Alaska	earthquake
23	36km WNW of Chirikof Island, Alaska	Alaska	earthquake

## 2. Data set eq2015: Cleansing & Wrangling Data – Location Column:

- Here we have multiple strings like Alaska, but some of them appear to be misspelled.
- When we look at row 14(Alaska), row 18 (Alaska), etc.
- It was the same for California, so before we need to remedy them, we need to conduct further research to understand the data.
- For this I clicked on the drop-down menu arrow of location.
- Then click on Facets and Text facets.

The screenshot shows the OpenRefine interface with a dataset titled 'eq2015.csv'. The main view displays 8708 rows of earthquake data. A context menu is open over a specific row in the 'place' column, providing options for collapsing or expanding columns, and a 'View' option which is currently selected.

### 3. Exploring/Cleansing/Wrangling Data Set by Clustering

#### a. Using Earthquake 2015 project.

- I clicked on the drop-down menu arrow of location.
- Then I clicked on Edit cells and cluster and edit...
- After clicking on that I was taken to the below page.

#### b. Different ways to cluster

- OpenRefine: Clustering with the key collision method – fingerprint.

The screenshot shows the 'Cluster and edit column "Location"' dialog in OpenRefine. It displays a table with one cluster found, 'Alaska', containing 795 rows. The 'Merge?' checkbox is unchecked. The main pane lists other locations such as 'dahoma', 'california', 'ew island', 'oska', and 'eska'.

- OpenRefine: Clustering with the key collision method – n-Gram fingerprint with 2 for size.

**Cluster and edit column "Location"**

Method: Key collision    Keying function: n-Gram fingerprint    n-Gram size: 2    1 cluster found

Cluster size	Row count	Values in cluster	Merge?	New cell value
2	795	• Alaska (791 rows) • alaska (4 rows)	<input type="checkbox"/>	Alaska

Location type

- Alaska earthquake
- Idaho earthquake
- California earthquake
- New Zealand earthquake
- New Zealand earthquake
- Alaska earthquake
- Alaska earthquake

- OpenRefine: Clustering with the key collision method – n-Gram fingerprint with 3 for size

**Cluster and edit column "Location"**

Method: Key collision    Keying function: n-Gram fingerprint    n-Gram size: 3    2 clusters found

Cluster size	Row count	Values in cluster	Merge?	New cell value
2	26	• B.C. (24 rows) • (2 rows)	<input type="checkbox"/>	B.C.
2	795	• Alaska (791 rows) • alaska (4 rows)	<input type="checkbox"/>	Alaska

# Rows in cluster: 20 — 800

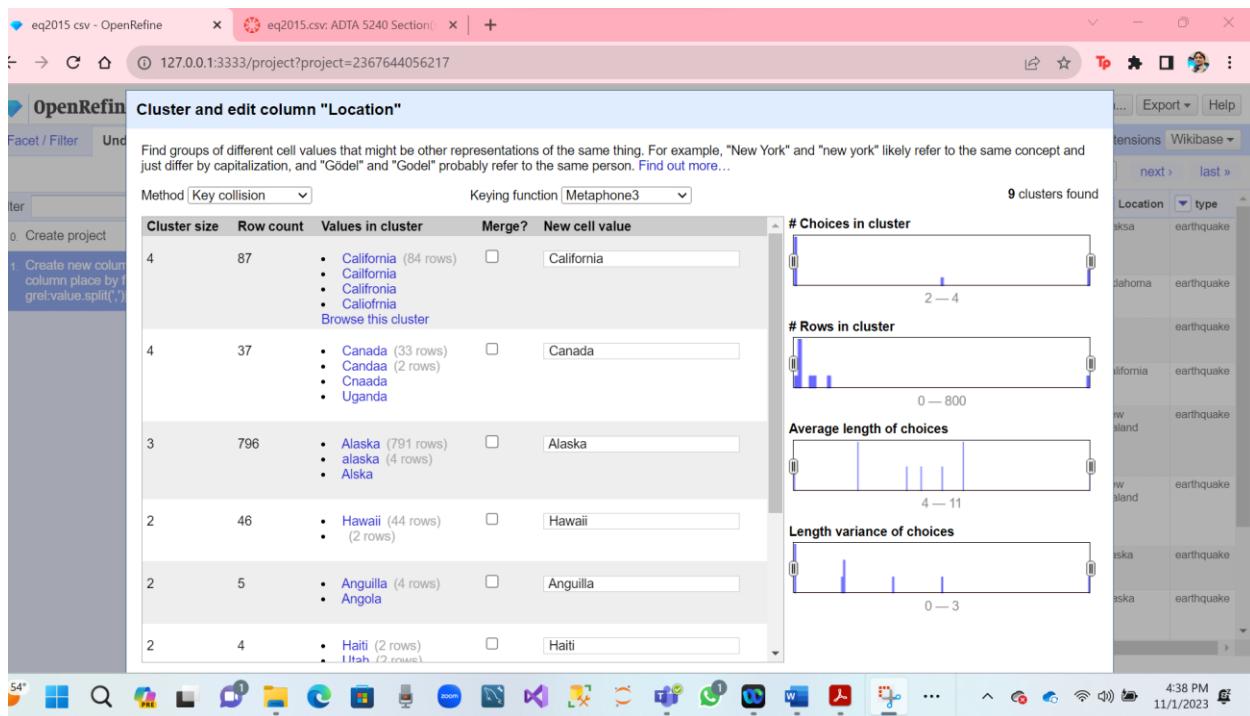
Average length of choices: 3 — 7

Length variance of choices: 0 — 2

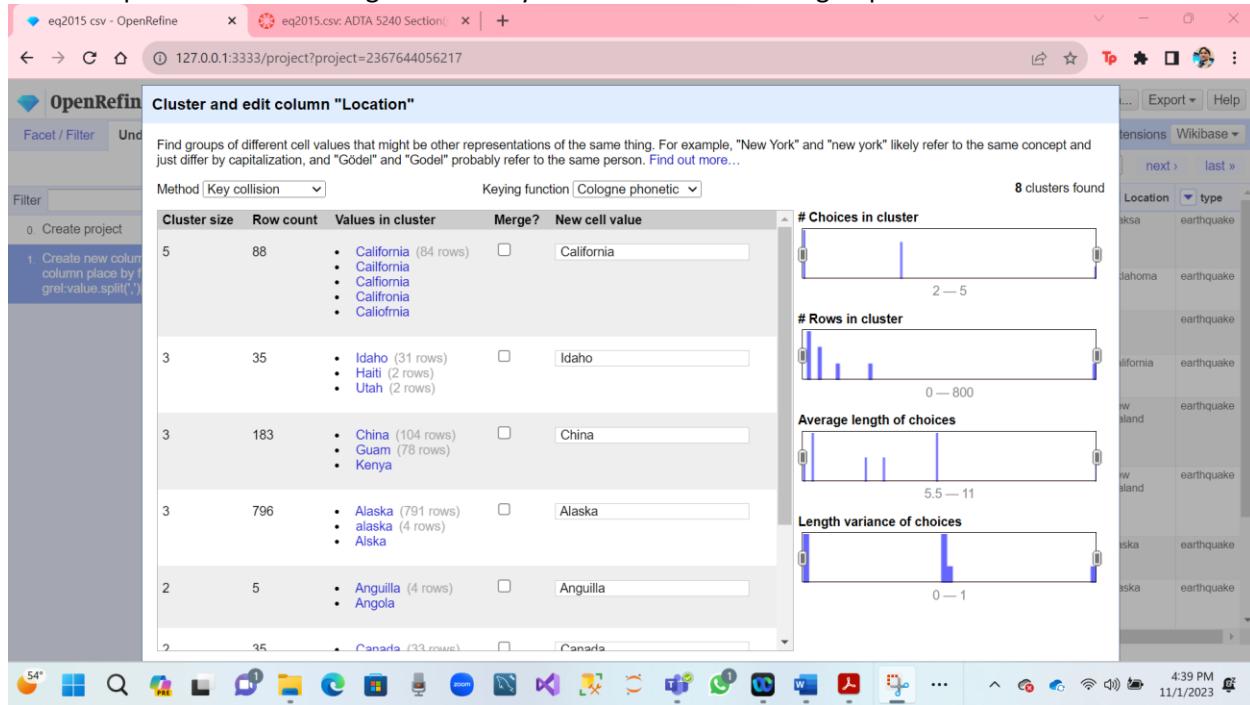
Location type

- Alaska earthquake
- Idaho earthquake
- California earthquake
- New Zealand earthquake
- New Zealand earthquake
- Alaska earthquake
- Alaska earthquake

- OpenRefine: Clustering with the key collision method – metaphone3.



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



- OpenRefine: Clustering with the nearest neighbour method: Levenshtein with radius of 1.0 and block chars of 6.

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 **Levenshtein** Radius **1.0** Block chars **6** 1 cluster found

Cluster size	Row count	Values in cluster	Merge?	New cell value
2	795	• Alaska (791 rows) • alaska (4 rows)	<input type="checkbox"/>	Alaska

Location	type
Alaska	earthquake
Hawaii	earthquake
California	earthquake
New Island	earthquake
New Island	earthquake
Alaska	earthquake
Alaska	earthquake

- OpenRefine: Clustering with the nearest neighbor method: Levenshtein with a radius of 2.0 and block chars of 6.
- OpenRefine: Clustering with the nearest neighbor method: Levenshtein with a radius of 3.0 and block chars of 6.
- OpenRefine: Clustering with the nearest neighbor method: PPM with a radius of 1.0 and block chars of 6.

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 **1.0** Block chars **6**

No clusters were found with the selected method

Try selecting another method above or changing its parameters

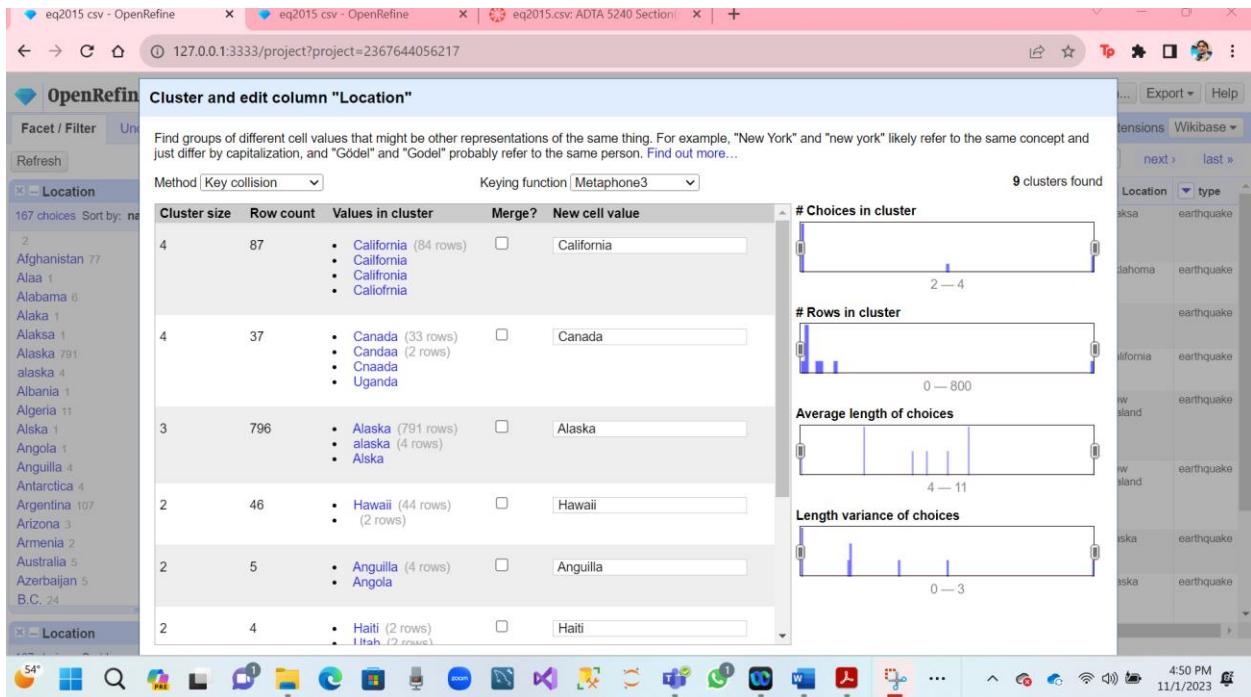
Location	type
Alaska	earthquake
Hawaii	earthquake
California	earthquake
New Island	earthquake
New Island	earthquake
Alaska	earthquake
Alaska	earthquake

#### 4. Cleansing/Wrangling data: Editing Cells

- Here we need to fix these types and errors using the clustering feature of OpenRefine.

##### Editing cells using the key collision method and metaphone3 keying function:

- Here I clicked on check boxes for the clusters California and Alaska.
- Then I clicked on merge.
- Here results show California and Alaska are gone.



- Here the results show “Alaska” was capitalized as “A” now and “California” as “C”.

#### 5. Cleansing/Wrangling Data: Manually editing cell:

- Sometimes we need to edit the cells manually to correct the type of errors.
- Here we use “edit” feature of a cell to edit the text of the value.
- Then I clicked on undo and redo selected mass edit 8803 cells in column location.
- Then I clicked on Alaka.
- We will look at it now because we want to confirm that as Larson Bay is Alaska and not something else.
- I right clicked on Larson Bay in place column and clicked on google search.

eq2015 csv - OpenRefine eq2015 csv - OpenRefine eq2015.csv: ADTA 5240 Section(1) | +

127.0.0.1:3333/project?project=2367644056217

**OpenRefine eq2015 csv** Permalink

Facet / Filter Undo / Redo 2 / 2

Refresh Reset all Remove all

Alaksa 1

Facet by choice counts

**1 matching rows (8708 total)**

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

« first < previous 1 next » last »

latitude	longitude	depth	mag	magType	nst	gap	dmin	rms	net	id	updated	place	Location	type
56.7152	-155.4884	5.4	3.6	ml				1.08	ak	ak11640129	2015-07-03T07:18:40Z	99km N of Chirikof Island, Alaska	Alaksa	earthquake

Location

Alaksa

case sensitive  regular expression

54° 4:58 PM 11/1/2023

- After clicking on that it took me to the below page.



- Then I came back to undo and redo in Open Refine page.
- Here I have undo all the functions previously I did.
- Hence, we have cleansed the data.
- It is OpenRefine with a couple of different data sets.

eq2015 csv - OpenRefine eq2015 csv - OpenRefine eq2015.csv: ADTA 5240 Section| larsen bay location map - Google Chrome

127.0.0.1:3333/project?project=2367644056217

**OpenRefine eq2015 csv Permalink**

**Facet / Filter Undo / Redo 1 / 2**

**8708 records**

Show as: rows records Show: 5 10 25 50 100 500 1000 records « first < previous 1 next » last »

Extensions Wikibase

Facet / Filter Undo / Redo 1 / 2

Refresh Reset all Remove all

**Location** change

167 choices Sort by: name count Cluster

	time	latitude	longitude	depth	mag	magType	nst	gap	dmin	rms	net	id	updated	place
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 Chirikof Island, Alaska
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 Medford, Oklahoma
3.	2015-07-02T22:31:28.190Z	-23.058 <sup>edit</sup>	-14.0431	10	4.8	mb		99	30.883	0.62	us	us10002n4f	2015-07-03T06:34:01.780Z	Southern Mid-Atla 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	122km S of L'Espere 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 E of L'Espere Rock, New Zealand
7.	2015-07-02T18:24:55.000Z	51.548	-175.7676	40.6	4.1	ml				0.75	ak	ak11639972	2015-07-03T02:27:38.059Z	71km E of Adak, Alaska
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 W of Chirikof Island, Alaska

53° 5:09 PM 11/1/2023

- Finally, Successfully Data mine with OpenRefine.